

Bertram K. C. Chan

Biostatistics for Human Genetic Epidemiology

Advances in Experimental Medicine and Biology

Volume 1082

Editorial Board

IRUN R. COHEN, *The Weizmann Institute of Science, Rehovot, Israel*

ABEL LAJTHA, *N.S. Kline Institute for Psychiatric Research, Orangeburg, NY, USA*

JOHN D. LAMBRIS, *University of Pennsylvania, Philadelphia, PA, USA*

RODOLFO PAOLETTI, *University of Milan, Milan, Italy*

NIMA REZAEI, *Tehran University of Medical Sciences, Children's Medical Center Hospital, Tehran, Iran*

More information about this series at <http://www.springer.com/series/5584>

Bertram K. C. Chan

Biostatistics for Human Genetic Epidemiology

Bertram K. C. Chan
Epidemiology and Biostatistics
Loma Linda University School of Medicine
and Public Health
Sunnyvale, CA, USA

ISSN 0065-2598 ISSN 2214-8019 (electronic)
Advances in Experimental Medicine and Biology
ISBN 978-3-319-93790-8 ISBN 978-3-319-93791-5 (eBook)
<https://doi.org/10.1007/978-3-319-93791-5>

Library of Congress Control Number: 2018953701

© Springer International Publishing AG, part of Springer Nature 2018
This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Dedicated to the glory of God and to my better half Marie Nashed Yacoub Chan

Preface

Some genetic epidemiologic experiences of, and concomitant challenges for, this writer are as follows -

Experience and Challenge #1: Type-1 Diabetes

CASE SUBJECT: A child with Type-1 Diabetes – the case subject was a 14-year-old child who was clinically diagnosed, some 2 years previously, as suffering from Type-1 (juvenile) diabetes (Chan 2015). Now, at the currently accepted level of understanding, only about 5% of people with diabetes have this form of the disease. In a case subject with Type-1 diabetes, the body does **not** produce insulin. Normally, the human body breaks down the starches and sugars, that one eats, first into a simple sugar (called glucose) which it then uses for energy. In this process, insulin is a hormone that the body **needs** to get glucose from the bloodstream into the cells of the body (American Diabetes Association 2017). And with the help of insulin therapy and other treatments, young children may learn to manage their conditions and live long, healthy, and productive lives. In almost all cases of Type-1 diabetes, the medical and health communities focus on the medical engineering aspect of how best to effectively “pump” insulin into the patient’s system. The disease is generally considered as permanently irreversible, viz., “incurable”!

One would certainly like to learn more about the genetic basis of case subjects diagnosed with Type-1 diabetes!

Moreover, this particular 14-year-old case subject was enrolled in a test in which the child orally took a prescribed medication for a period of about 3 months. Interestingly, this special medication was a Traditional Chinese Medicine (TCM) formulation of herbal origin. During this period of special medication, A1C blood tests were taken to monitor the progress of the case subject. The progressive A1C test results were as follows:

$$9+ \rightarrow 8.4 \rightarrow 7.8 \rightarrow 7.45 \rightarrow 6.7(\%)$$

The low/last reading was below the 6.9% level – which may be considered as the A1C reading for a normal and non-diabetic person!

How does this particular test result affect the accepted medical position that Type-1 diabetes is permanently irreversible? Can epidemiologic research help? Clearly much epidemiologic investigation is called for in this situation. Actually, there had been a clinical trial in which the same TCM treatment was given to more than 10,000 case subjects resulted in a positive response (viz., improved stability of blood glucose control **without** insulin) in about 30% of the test population. Such results should be

considered as strong justification for further epidemiologic studies (including genetic-epidemiologic investigations) in this particular area!

Experience and Challenge #2: Autism Spectrum Disorders (ASD) – A Costly Condition!

Recently, this writer experienced a cultural shock: a certain rental property placed on the open market received the highest bid from an organization which provides daily care to autistic children. This serendipitous result came when it was discovered the State Health Department considers it appropriate to heavily support such an organization which provides daily health and educational care to all qualified children with ASD! (Newschaffer et al. 2007) Can one expect some relief for the financial cost for supporting such a societal program? And what about the concomitant social and personal costs for supporting such a program?

Confronting the question: “Is there a relief, or a cure, in sight for the autistic state of human conditions?” A recent report (<http://edition.cnn.com/2017/04/05/health/autism-cord-blood-stem-cells-duke-study/index.html>) points to a study on the safety and effectiveness of infusion of umbilical cord blood into children with autism did yield some promising results! Perhaps one may raise the question: Is genetics still a critical factor involved in such an extraordinary and heroic approach to medicine and health care?

Experience and Challenge #3: Childhood Brain Tumors

The challenge of understanding the genetic epidemiology of fatal childhood brain cancer tumors was experienced by this writer – he was acquainted with a married couple (both of whom came from very close and similar ethnic backgrounds). Later, this couple was blessed with the birth of a child, who, at the age of 12 months, developed fatal brain cancer tumors. The infant spent the next few months in the hospital before passing away! It was, indeed, a very sad occasion at the memorial service of that precious child. Sometime later, the parents decided to forego the conceiving and birthing another offspring, and chose to adopt an infant – across ethnic and racial lines!

This instance seems to call for a much-motivated understanding of the “Genetic Epidemiology of Childhood Brain Cancer Tumors” (Bondy 1990).

Genetic Epidemiology holds a critically important and effective role in understanding the critical factors in the aforementioned diseases and health issues, especially with respect to hereditary and environmental factors (the “Nature vs. Nurture” issues). Starting from population-based methods, the magnitude of genetic effects on health and diseases may be assessed (Austin 2013). To these approaches one may add quantitative methods, including biostatistical analysis. The latter methodology may be efficiently enhanced with the now-popular R programming software (Chan 2015).

To understand, and ultimately to apply the knowledge of, the etiology of a disease, it seems imminently helpful to unravel the relationships (if any) that govern the genetic bases of the disease. While genetics is complicated, it is to be hoped that the use of available biostatistical power, supported by the efficiencies of the computer program (developed largely for biostatistical applications), will go some positive way toward resolving some of the complex relations within genetic epidemiology.

Experience and Challenge #4: CMT (Charcot-Marie-Tooth) Disease^[w]

From a personal friend of the author was learnt that a rather common hereditary disease of genetic origins can cause severe weakness of the limbs that required supporting metallic braces to aid simple daily walking. This disease, known as CMT, was recently diagnosed in a personal acquaintance!

CMT is one of the most common inherited neurological disorders, affecting approximately 1 in 2,500 people in the United States of America. The disease is named after the three physicians who first identified it in 1886 – Jean-Martin **Charcot** and Pierre **Marie** in Paris, France, and Howard Henry **Tooth** in Cambridge, England. CMT, also known as **Hereditary Motor and Sensory Neuropathy (HMSN)** or **Peroneal Muscular Atrophy (PMA)**, comprises a group of disorders that affect peripheral nerves. The peripheral nerves lie outside the brain and spinal cord and supply the muscles and sensory organs in the limbs. Disorders that affect the peripheral nerves are called peripheral neuropathies.

Although there is no known cure for CMT, physical therapy, occupational therapy, braces and other orthopedic devices, and even orthopedic surgery may help individuals deal with the disabling symptoms of the disease. In addition, pain-killing drugs can be prescribed for individuals who have severe pain.

Physical and occupational therapy, the preferred treatment for CMT, involves muscle strength training, muscle and ligament stretching, stamina training, and moderate aerobic exercise. Most therapists recommend a specialized treatment program designed with the approval of the person's physician to fit individual abilities and needs. Therapists also suggest entering into a treatment program *early* as muscle strengthening may delay or reduce muscular atrophy, so strength training is most useful if it begins before nerve degeneration and muscle weakness progresses to the point of disability!

Experience and Challenge #5: Alzheimer Disease

To this author, this experience has a rather personal and emotional background: the beloved pastor of the home church retired, but soon his wife died suddenly owing to the rupture of an abdominal aneurism – and a year later, the pastor himself rapidly lapsed into severe symptoms of Alzheimer Disease (AzD) – unable even to remember the first name of the author who had been his personal friends for years! Thus, the beloved pastor had become a “stranger,” then passed away within a year or so! And, more than that:

On a national, if not worldwide, scale, it has been reported that (<http://www.foxnews.com/health/2017/03/01/could-alzheimers-really-bankrupt-medicare-and-medicaid.html>):

Could Alzheimer's really bankrupt Medicare and Medicaid? By Lindsay Carlton Published March 01, 2017



The disease that could collapse Medicare, Medicaid

The most expensive medical condition in America threatens to bankrupt Medicare, Medicaid and the life savings of millions of Americans. But the perpetrator isn't cancer or heart disease — it's Alzheimer's.

Fox News' Dr. Manny Alvarez sat down with Dr. Rudolph Tanzi, a professor of neurology at Harvard Medical School who participated in PBS' "Alzheimer's: Every Minute Counts" documentary, which takes a closer look at the critical financial problem Americans are facing with the disease, to discuss the issue.

"Because we're living so long, our health span, especially our brain health span, is not keeping up with our life span," Tanzi told Fox News. "All of modern medicine has us living on average till 80 years old, and by 85 years old you have a 40 to 50% chance of having Alzheimer's."

In 2016, total payments for health care, long-term care and hospice were estimated to be \$236 billion for people with Alzheimer's and other dementias, according to the Alzheimer's Association.

Tanzi explained that right now, \$1 of every \$5 (20.0%) in Medicare and Medicaid funding goes toward Alzheimer's patients' care. Given how many more Alzheimer's patients are expected to be diagnosed within the next decade, that number is predicted to increase to every \$1 in \$3 (33.3%). In that case, the program's funding may collapse, which would leave insufficient funds to prevent other age-related disease, he said.

"It hits every sector from the burden on the family: the caregiver taking care of their loved one who they're losing in front of their eyes, and then the government costs, assisted living," Tanzi said.

Much have been achieved in the study of population-based genetics, biostatistical genetics, epidemiology, and hopefully and finally make a significantly useful impact on genetic epidemiology. To that end, the author is prepared to introduce the title

"Biostatistics for Genetic Epidemiology: An Introduction Using R"

in terms of the following chapters:

1. **Introduction to Genetic Epidemiology**
2. **Data Analysis Using R Programming**
3. **Human Genetics and Genetic Epidemiology**
4. **Statistical Human Genetics Using R**
5. **Genetic Epidemiology Using R**

Sunnyvale, CA, USA

Bertram K. C. Chan

References

- American Diabetes Association (2017) <http://diabetes.org>
- Austin MA (2013) Genetic epidemiology: methods and applications (Modular Text Series). CABI Publishing, Wellingford
- Bondy ML (1990) Genetic epidemiology of childhood brain tumors, Texas Medical Center Dissertations (via ProQuest). AA119109972. <http://digitalcommons.library.tmc.edu/dissertations/AA19109972>
- Chan BKC (2015) Biostatistics for epidemiology and public health disorders. Ann Rev Pub Health 28:235–258. 10.1146/annurev.pubhealth.28.021406.144007
- <http://edition.cnn.com/2017/04/05/health/autism-cord-blood-stem-cells-duke-study/index.html>
- <http://www.foxnews.com/health/2017/03/01/could-alzheimers-really-bankrupt-medicare-and-medicaid.html>
- Newschaffer CJ et al (2007) The epidemiology of autism spectrum using R, Springer Publishing Company, New York

Contents

1	Introduction to Human Genetic Epidemiology	1
1.1	Medicine, Preventive Medicine, Public Health, and Epidemiology	1
1.1.1	An Overseas Vacation Tour and Worldwide Infectious Diseases	1
1.1.2	Genetics and Infectious Diseases [<i>infectious diseases and genetics => Genetics of infectious diseases => academic.oup.com</i>]	5
1.2	Human Genetic Epidemiology (HGE)	5
1.2.1	The Human Genome Project (HGP) ^[W]	6
1.2.2	Human Genes, Genetics, and Health	12
1.2.3	A Glossary of Common Terms in Human Genetics	18
1.2.4	Human Genetics in Medicine ^[W]	18
1.2.5	Human Genetic Epidemiology	44
1.2.6	Applied Statistical Human Genetics	46
	References	46
2	Data Analysis Using R Programming	47
2.1	Data and Data Processing	48
2.2	Beginning R	53
2.2.1	A First Session Using R	56
2.2.2	The R Environment	68
2.3	R As a Calculator	70
2.3.1	Mathematical Operations Using R	70
2.3.2	Assignment of Values in R, and Computations Using Vectors and Matrices	72
2.3.3	Computations in Vectors and Simple Graphics	72
2.3.4	Use of Factors in R Programming	73
2.3.5	Simple Graphics	74
2.3.6	x As Vectors and Matrices in Statistics	78
2.3.7	Some Special Functions That Create Vectors	79
2.3.8	Arrays and Matrices	80
2.3.9	Use of the Dimension Function <code>dim()</code> in R	81
2.3.10	Use of the Matrix Function <code>matrix()</code> in R	81
2.3.11	Some Useful Functions Operating on Matrices in R	81
2.3.12	NA ‘Not Available’ for Missing Values in Datasets	82
2.3.13	Special Functions That Create Vectors	83

2.4	Using R in Data Analysis in Human Genetic Epidemiology	87
2.4.1	Entering Data at the R Command Prompt	87
2.4.2	The Function list() and the Construction of data.frame() in R	98
2.5	Univariate, Bivariate, and Multivariate Data Analysis	100
2.5.1	Univariate Data Analysis	101
2.5.2	Bivariate and Multivariate Data Analysis	103
Appendix 1		121
	Documentation for the plot function	121
Special References		122
3	Applied Statistics for Human Genetics Using R	123
3.1	Some Fundamental Concepts in the Theory of Probability and Applied Statistics in Epidemiology	123
3.2	Biostatistical Concepts and Measures in Genetic Association	124
3.2.1	Familial Aggregation Studies	125
3.2.2	Segregation Studies	127
3.2.3	Linkage Studies	128
3.2.4	Association Studies	130
3.3	Genome-wide Association Studies (GWAS)	136
3.3.1	A Worked Example of SNPs-based Whole Genome Association Study	137
3.4	Big Data and Human Genomics	142
3.4.1	What Is Big Data? [W]	142
3.4.2	What Is Genetic Big Data? And Where Is It Taking Genetics?	142
3.4.3	Analysis of Human Genomics	143
References		144
4	Applied Human Genetic Epidemiology	145
4.1	The Study of Human Genetic Epidemiology	146
4.1.1	Family Studies in Genetic Epidemiology	148
4.2	Human Genetic Influences on Diseases	165
4.2.1	Genetic Relationships in a Familial Aggregation ^[*]	165
4.2.2	Familial Risk of Diseases	167
4.2.3	Heritability Analysis ^[G]	174
4.2.4	Molecular Variation Study Methods	181
4.3	Genomics for Human Genetic Epidemiology	183
4.3.1	Complex Traits and Mendelian Inheritance	184
4.4	Factors in Human Genetic Epidemiology	198
4.4.1	Linkage Analysis	199
4.4.2	Family Association Studies	200
4.5	Human Genetic Association	204
4.6	Genetic Epidemiology Owing to Population Stratification	205
4.7	Environmental Effects on Genetic Epidemiology ^[Google]	206
4.7.1	Environmental Factors on Genetic Epidemiology ^[Google]	207
4.8	Genetic Epidemiology and Public Health ^[Google]	209
Special References		216

5 Human Genetic Epidemiology Using R	217
5.1 Biostatistical Human Genetics	219
5.1.1 Some Preliminary Remarks on the <i>T</i> -Test in Statistics	222
5.2 Human Genetic Data Concepts	231
5.2.1 The Study of Human Genetic Variation	231
5.2.2 Manhattan Plots ^[Wikipedia]	252
5.3 Procedures for Multiple Comparison	263
5.3.1 Worked Examples of Statistical Tests and Utilities for Genetic Association	265
5.3.2 Worked Examples of Statistical Tests and Utilities for Genetic Association	281
5.4 Regression Decision Trees and Classifications ^[Google]	308
5.5 Multi-dimensional Analysis in Genetic Epidemiology	311
5.5.1 Biomedical Background Challenges to Genetic Epidemiology	312
5.5.2 Worked Examples in Epidemiology	324
References	341
References	343
Index	353

About the Author

Bertram K. C. Chan PhD, PE, Life Member-IEEE, completed his secondary education in Sydney, Australia, having passed the New South Wales State Leaving Certificate (viz., university matriculation examination) with excellent results in mathematics and in honours physics and in honours chemistry.

He then completed both a Bachelor of Science degree in Chemical Engineering, with First Class Honours (summa cum laude), and a Master of Engineering Science degree in Nuclear Engineering at the University of New South Wales, and a PhD degree in Engineering at the University of Sydney.

This was followed by 2 years of work as a Research Engineering Scientist at the Australian Atomic Energy Commission Research Establishment, and 2 years of a Canadian Atomic Energy Commission postdoctoral fellowship at the University of Waterloo, Canada.

He had undertaken additional graduate studies at the University of New South Wales, at the American University of Beirut, and at Stanford University, in mathematical statistics, computer science, and pure and applied mathematics (abstract algebra, automata theory, numerical analysis, etc.,), and in electronics and electromagnetic engineering.

His professional career includes over 10 years of full-time, and 10 years of part-time, university-level teaching and research experience in several institutions, including an appointment as a research associate in biomedical and statistical analysis, Perinatal Biology Section, ObGyn Department, University of Southern California Medical School, teaching at Loma Linda University, Middle East University, and research engineering staff positions at Lockheed Missile & Space (10 years), Apple (7 years), Hewlett-Packard (3 years), and at a start-up company (Foundry Networks) in the manufacture of Internet hardware and software: gigahertz switches and routers (7 years).

In recent years:

- He supported the biostatistical work of the Adventist Health Studies II research program at the Loma Linda University Health (LLUH) School of Medicine, California, and consulted as a forum Lecturer for several years in the LLUH School of Public Health (biostatistics, epidemiology, and population medicine). The LLUH lectures formed part of this book. In these lectures, Dr. Chan introduced the use of the programming language R and designed these lectures for the biostatistical elements for courses in the MPH, MsPH, DrPH, and PhD programs, with special reference to epidemiology in particular and public health and population medicine in general.
- Dr. Chan has three US patents in electromagnetic engineering, has published over 30 engineering research papers, and authored a 16-book set in educational mathematics (Chan 1978), as well as two

- monograms entitled: *Biostatistics for Epidemiology and Public Health Using R* (Chan 2016) and *Applied Probabilistic Calculus for Financial Engineering: An Introduction Using R* (Chan 2017).
- He is a registered Professional Engineer (**PE**) in the State of California, as well as a life member of the Institute of Electrical and Electronic Engineers (**IEEE**).

References

- Chan BKC (1978) A new school mathematics for Hong Kong, 10 Volumes: 1A, 1B, 2A, 2B, 3A, 3B, 4A, 4B, 5A, 5B, 6 Workbooks: 1A, 1B, 2A, 2B, 3A, 3B. Ling Kee Publishing Co., Hong Kong
- Chan BKC (2016) Biostatistics for epidemiology and public health using R. Springer, New York (with additional materials on the Publisher's website)
- Chan BKC (2017) Applied probability calculus for financial engineering: an introduction using R. Wiley, Hoboken



Introduction to Human Genetic Epidemiology

1

Abstract

Human genetic epidemiology (HGP) is concerned with a knowledge of medicine, preventive medicine, public health, and epidemiology. In the modern era of genetic medicine, HGP must be concerned with human genetic diversity including mutation and polymorphism.

Keywords

Medicine · Preventive medicine · Public health · Epidemiology · Human Genome Project · Human genetics · Human genome · Genetic medicine · Mutation and polymorphism · Clinical cytogenetics · Genome analysis · Chromosomal and genomic bases of diseases · Genetic bases for human diseases · Molecular basis of genetic diseases · The treatment of genetic diseases · Developmental genetics and birth defects · Cancer genetics and genomics · Risk assessment and genetic counseling · Prenatal diagnosis and screening · Genomics for medicine and personal health · Social and ethical issues in genetic medicine · Human genetics and genetic epidemiology · Statistical human genetics and statistical genetic epidemiology · Applied statistical human genetics

1.1 Medicine, Preventive Medicine, Public Health, and Epidemiology

The interactional relationships among these four topics are most interesting and will be thoroughly explored in this book, including quantitative measures for the last of these four, using statistical and computational methodologies now available.

1.1.1 An Overseas Vacation Tour and Worldwide Infectious Diseases

Recently, after having taken permanent retirement, Mr. and Mrs. Smith (not their real names) decided to take an extended vacation tour of Europe (visiting several countries around the Mediterranean Sea) and of Northern Africa, including the Kingdom of Morocco and the culturally-fascinating ancient land of Egypt (now officially known as the Arabic Republic of Egypt, ARE). In preparation for the trip, they consulted with their family physician in California to take care of any anticipated and unanticipated health-related needs, especially with respect to their planned travels.

During their regular wellness examinations, the Smiths disclosed their travel plans to their physician who recommended that, in addition to their annual ‘flu shots’, etc., they may be well-advised to receive the Pneumonia Prevention Vaccination (PPV): it is known that the PPV can lower ones chances of catching the disease. And even if one had the shot *and* later one does get pneumonia, one will most probably have a much milder one! Pneumonia is a pulmonary condition in which there is inflammation of the alveoli, viz., the small air sacs in the lungs. Infection by *micropulmona*, pneumonia is more common among people whose immune systems are milder – weaker one, especially older ones!

The Workings of Public Health- The Public Health Card in Egypt

During the planned vacation tour, as the Smiths reach the ancient city of Cairo, Egypt, they were cordially greeted at the beautiful brand new Cairo International Airport terminal by a special welcome-greeting card from the “Arab Republic of Egypt, Ministry of Health & Population, Preventive Sector, General Administration of Quarantine” which states:

“Dear Passenger

Pay attention to your Health when you come from these countries – (*followed by 4 lists of countries, each list pertaining to a specific transmissible disease!*)

1. Countries with risk of *malaria* transmission:-

- Algeria, Angola, Argentina, Azerbaijan, Afghanistan,
- Botswana, Benin, Burkina Faso, Burundi, Bolivia, Brazil, Belize, Bahamas, Bangladesh, Bhutan,
- Congo, Cape Verde, Cameroon, Central Africa, Cambodia, Chad, China, Colombia, Comoros, Costa Rica, Cote D’ivoir,
- Djibouti, Democratic Republic of the Congo, Democratic Peoples’ Republic of Korea, Dominica,
- Ethiopia, Eritrea, Ecuador, Equatorial Guinea,
- French Guiana,
- Gabon, Gambia, Georgia, Ghana, Greece, Guinea-Bissau, Guyana, Guatemala,
- Haiti, Honduras,
- Iraq, Island of Salomon, India, Iran, Indonesia,
- Jamaica,
- Kenya, Kyrgyzstan,
- Liberia,
- Malawi, Mali, Madagascar, Mayotte, Malaysia, Mauritania, Mozambique, Myanmar,
- Namibia, Nicaragua, Niger, Nigeria, Nepal,
- Oman,
- Pakistan, Panama, Papua New Guinea, Paraguay, Peru, Philippines,
- Republic of Laos, Rwanda,
- Salvador, Sao Tome, Salvador, Saudi Arabia, Senegal, Singapore,
- Sierra Leone, South Africa, South Korea, South Sudan, Sudan, Swaziland,
- Tajikistan, Tanzania, Thailand, Timor, Togo, Turkey,
- Uzbekistan,
- Vanuatu, Venezuela, Vanuatu, Vietnam,
- Yemen,
- Zambia, Zimbabwe, (a list of about 100 countries)

2. Countries with risk of *yellow fever* transmission:-

- Angola, Argentina,
- Benin, Burkina Faso, Brazil, Belize, Burundi, Bolivia,
- Congo, Chad, Central Africa, Colombia, Cote D'ivoir,
- Democratic Republic of the Congo,
- Ecuador, Equatorial Guinea, Ethiopia,
- French Guiana,
- Gabon, Gambia, Georgia, Ghana, Guinea, Guinea-Bissau,
- Kenya,
- Liberia,
- Mali, Mauritania,
- Niger, Nigeria,
- Panama, Paraguay, Peru,
- Rwanda,
- Senegal, Sierra Leone, South Sudan, Sudan, Suriname,
- Togo, Trinidad,
- Uganda,
- Venezuela, (a list of about 43 countries)

3. Countries with risk of *meningitis* transmission:-

- Benin, Burkina Faso,
- Cameroon, Central Africa, Chad, Cote D'ivoir,
- Democratic Republic of the Congo,
- Ethiopia, Eritrea,
- Gambia, Ghana, Greece, Guinea-Bissau,
- Kenya,
- Mali, Mauritania, Mozambique, Myanmar,
- Niger, Nigeria,
- Senegal, Singapore, South Sudan, Sudan,
- Togo,
- Uganda (a list of about 26 countries)

4. Countries with risk of *cholera* transmission:-

- Afghanistan, Angola,
- Benin, Burkina Faso, Burundi,
- Cameroon, Republic of the Central Africa, Chad, China, Democratic Republic of the Congo, Cote D'ivoir, Cuba,
- Dominica,
- Ghana, Guinea-Bissau
- Haiti,
- Iraq, Iran,
- Liberia,
- Malawi, Mali, Malaysia, Mozambique, Myanmar,
- Nepal, Niger, Nigeria,
- Pakistan, Philippines,
- Rwanda,
- Senegal, Sierra Leone, Somalia,
- Tanzania, Thailand, Togo,
- Uganda,
- Zambia, Zimbabwe, (a list of about 39 countries)

Along with these impressive lists is the following medical and public health advice:

“Dear Passenger – when you feel any of the following symptoms within four weeks from the date of arrival from any of the foregoing list of countries:

- ***Rise in body temperature***
- ***Headache***
- ***Profuse sweating***
- ***Chills***
- ***Muscle pain***
- ***Severe fatigue***
- ***Coughs***
- ***Diarrhea, vomiting***
- ***Rash bleeding from the mouth and nose***

Please go to the nearest fever hospital and seek medical advice, informing your dates of arrival and the country visited.

(from the General Administration of Quarantine Preventive Sector)

It is clear that such an overseas vacation tour can indicate much about the state of worldwide infectious diseases!

The War on Cancer

In January 1971, the United States President Richard Nixon made a State of the Union Address that had become known as the declaration of war against cancer. US\$10 Billion was pledged to find a cure for cancer! And now, almost half a century later, the war is still raging on!

Herebelow is a recent report on that “War on Cancer”, still raging on, and on the Genetics front:

Genetics

China Has Already Gene-Edited 86 People With CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats)(by Kristen V. Brown, as reported in the Wall Street Journal, 2018)

CRISPR is now a name attached to the process of Gene DNA Editing. The name was minted at a time when the origin and use of the interspacing subsequences were not known. At that time the CRISPRs were described as segments of ***prokaryotic DNA*** containing short, repetitive base sequences. In a ***palindromic*** repeat, the sequence of ***nucleotides*** is the same in both directions. Each repetition is followed by short segments of ***spacer DNA*** from previous exposures to foreign DNA (e.g., a ***virus*** or ***plasmid***). Small clusters of ***cas*** (CRISPR-associated system) genes are located next to CRISPR sequences.

In the U.S., the first planned clinical trials of CRISPR gene editing in people are about to start. In China, meanwhile, CRISPR has been racing ahead, having already used the gene-altering tool to change the DNA of dozens of people in several clinical trials.

The ***Wall Street Journal*** reports that so far in China, at least 86 people have had their genes edited, and there is evidence of at least 11 Chinese clinical trials using CRISPR. One of those trials began a year earlier than previously reported, putting the start of the first Chinese CRISPR trial in 2015.

China’s rapid advancement is the result of more relaxed regulations, and a willingness to forge ahead with cutting-edge research despite potential unknowns and safety concerns, which are significant. One recent paper, for example, suggested that CRISPR could ***trigger an immune response*** in a

majority of patients, which could render potential treatments either ineffective or dangerous. China's rapid-fire approach has set off a **biomedical duel** between the U.S. and China, and sparked concerns among Western scientists that the Chinese **trials have been irresponsibly premature**.

In China's 2015 CRISPR trial, the *WSJ* reports, 36 patients with cancers of the kidney, lung, liver and throat had cells removed from their bodies, altered with CRISPR, and then infused back into their bodies to fight the cancer. Other Chinese trials have sought to use CRISPR to treat HIV, esophageal cancer, and leukemia. A trial slated for this year in China will enroll 16 patients. Meanwhile, the *first human CRISPR trial in the U.S.*, at University of Pennsylvania, will enroll just 18 people, and is designed primarily to test whether CRISPR is safe.

Chinese scientists may end up being the first to cure cancer using CRISPR, but it's unclear what repercussions may come with rushing through these early safety trials.

1.1.2 Genetics and Infectious Diseases [*infectious diseases and genetics* => *Genetics of infectious diseases* => academic.oup.com]

Both in terms of mortality and morbidity, the foregoing lists of infectious diseases illustrate major health problems worldwide! It is now well recognized that a complex combination of host genetic factors, environmental, and pathogen is fundamental in determining *both* the course of infection *and* the susceptibility to particular microbes as well as the path of infection. Numerous medical and epidemiologic studies have identified and traced the course of infection: these studies have successfully identified and mapped the relevant genes by means of population-based and family-based approaches.

For example, much investigations have been done on human susceptibility to HIV/AIDS, malaria, and mycobacterial infection. By genome scans of multi-case families, some major genes have been positively identified. To define the majority of the relevant polygenes, it is clear that Genome-Wide Association Studies (GWAS) with large sample sizes will be needed.

Generally, using the classical approach of case-control studies in epidemiology, one may discover underlying genetic effects. However, large sample sets are required to detect moderate genetic effects (in order to eliminate the possibility of false positive association). Thus the use of microarray technology has been successful in identifying novel candidate genes.

Family-based approaches have also contributed to an understanding of linkages to infectious diseases. Also, linkage studies may be used to identify genes that cause rare, monogenic susceptibility phenotypes.

1.2 Human Genetic Epidemiology (HGE)

In recent decades, the progress in various aspects of HGE has been well-supported by concomitant development and progress in several areas, including:

- (i) The Human Genome Project (HGP)
- (ii) The Epidemiology of Big Data (EBD), and
- (iii) The availability and the timely development of the open-source statistical software R (Fig. 1.1).

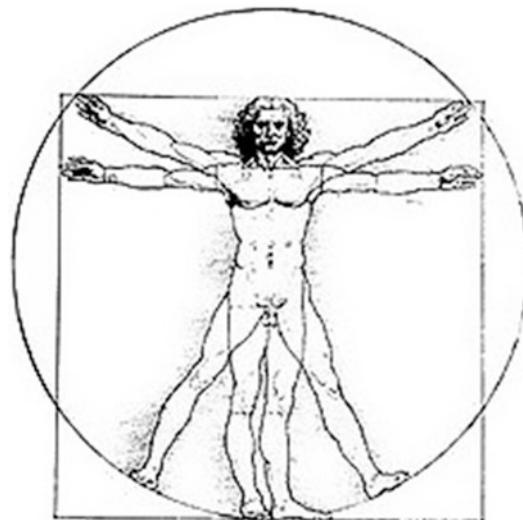


Fig. 1.1 The Symbolic HGP (*Vitruvian Man*, Leonardo da Vinci)

1.2.1 The Human Genome Project (HGP)^[W]

^[W] Wikipedia

- (i) The **Human Genome Project (HGP)** was a collaborative international [scientific research](#) project to determine the sequence of nucleotide [base pairs](#) that make up the human [DNA \(Deoxyribonucleic Acid\)](#) project. After the idea was accepted up in 1984 by the US government when the planning started, the project formally launched in 1990 and was declared complete in 2000. Funding came from the US government through the [National Institutes of Health \(NIH\)](#) as well as numerous other groups from around the world. A parallel project was conducted outside government by the [Celera Corporation](#), or Celera Genomics, which was formally launched in 1998. Most of the government-sponsored sequencing was performed in twenty [universities](#) and research centers in the United States, the United Kingdom, Japan, France, Germany, Canada, and China.

Initially the HGP aimed to map the [nucleotides](#) contained in a human [haploid reference genome](#) (more than three billion). *The “genome” of any given individual is unique*; mapping the “human genome” involved sequencing a small number of individuals and then assembling these together to get a complete sequence for each chromosome. Therefore, the finished human genome is thus a mosaic, not representing any one individual (Fig. 1.2).

Initiated in 1990, the Human Genome Project was a 15-year-long, publicly funded project with the objective of determining the DNA sequence of the entire euchromatic human genome within 15 years. In May 1985, Robert Sinsheimer organized a workshop to discuss sequencing the human genome but for a number of reasons the NIH was uninterested in pursuing the proposal. The following March, the Santa Fe Workshop was organized by [Charles DeLisi](#) and David Smith of the Department of Energy’s Office of Health and Environmental Research (OHER). At the same time [Renato Dulbecco](#) proposed whole genome sequencing in an essay in *Science*. James Watson followed two months later with a workshop held at the Cold Spring Harbor Laboratory.

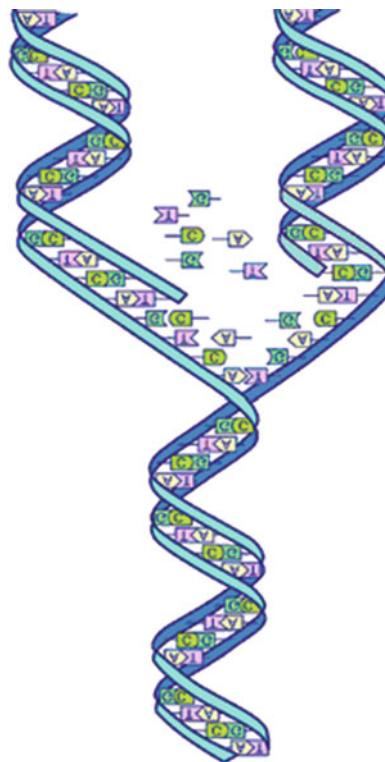


Fig. 1.2 The symbolic HGP

The fact that the Santa Fe workshop was motivated and supported by a Federal Agency opened a path, albeit a difficult and tortuous one, for converting the idea into public policy. Later, Congress added a comparable amount to the NIH budget, thereby beginning official funding by both agencies.

In 1993, Aristides Patrinos succeeded Galas and [Francis Collins](#) succeeded [James Watson](#), assuming the role of overall Project Head as Director of the NIH National Center for Human Genome Research (which later became the [National Human Genome Research Institute](#)). A working draft of the genome was announced in 2000 and the papers describing it were published in February 2001. A more complete draft was published in 2003, and genome “finishing” work continued for more than a decade.

The \$3-billion project was formally founded in 1990 by the US Department of Energy and the National Institutes of Health, and was expected to take 15 years. In addition to the United States, the international [consortium](#) comprised [geneticists](#) in the United Kingdom, France, Australia, China, and others.

Owing to widespread international cooperation and advances in the field of [genomics](#) (especially in [sequence analysis](#)), as well as major advances in computing technology, a draft of the genome was completed in 2000 (announced jointly by U.S. President [Bill Clinton](#) and the British Prime Minister [Tony Blair](#) on June 26, 2000). This first available rough draft [assembly](#) of the genome was completed by the Genome Bioinformatics Group at the [University of California, Santa Cruz](#), primarily led by then graduate student [Jim Kent](#). Ongoing [sequencing](#) led to the announcement of the essentially complete genome on April 14, 2003, two years earlier than planned! In May 2006, another milestone was passed

on the way to completion of the project, when the sequence of the [last chromosome](#) was published in [Nature](#). An initial draft of the human genome was available in June 2000 and by February 2001 a working draft had been completed and published, followed by the final sequencing mapping of the human genome on April 14, 2003. Although this was reported to cover 99% of the euchromatic human genome with 99.99% accuracy, a major quality assessment of the human genome sequence was published on May 27, 2004 indicating over 92% of sampling exceeded 99.99% accuracy which was within the intended goal.

Applications and Proposed Benefits of the HGP

The sequencing of the human genome holds benefits for many fields, from [molecular medicine](#) to [human evolution](#). The Human Genome Project, through its sequencing of the DNA, can help the understanding of diseases including: [genotyping](#) of specific [viruses](#) to direct appropriate treatment; identification of [mutations](#) linked to different forms of [cancer](#); the design of medication and more accurate prediction of their effects; advancement in [forensic](#) applied sciences; [biofuels](#) and other energy applications; [agriculture](#), [animal husbandry](#), [bioprocessing](#); [risk assessment](#); [bioarcheology](#), [anthropology](#) and [evolution](#). Another proposed benefit is the commercial development of [genomics](#) research related to DNA based products, a multibillion-dollar industry.

The sequence of the DNA is stored in [databases](#) available to anyone on the [Internet](#). The U.S. [National Center for Biotechnology Information](#) (and sister organizations in Europe and Japan) house the gene sequence in a database known as *GenBank* with sequences of known and hypothetical genes and proteins. Other organizations presented additional data and annotation and powerful tools for visualizing and searching it. [Computer programs](#) have been developed to analyze the data, because the data itself is difficult to interpret without such programs. Generally, advances in genome sequencing technology have followed Moore's Law, a concept from computer science which states that integrated circuits can increase in complexity at an exponential rate. This means that the speeds at which whole genomes can be sequenced can increase at a similar rate, as was seen during the development of the above-mentioned Human Genome Project.

Techniques and Analysis Associated with the HGP

The process of identifying the boundaries between genes and other features in a raw DNA sequence is called [genome annotation](#), usually studied under the domain of [bioinformatics](#). While expert biologists make the best annotators, their work proceeds slowly, and computer programs are increasingly used to meet the high-throughput demands of genome sequencing projects. Beginning in 2008, a new technology known as [RNA-seq](#) was introduced that allowed scientists to directly sequence the messenger RNA in cells. This replaced previous methods of annotation, which relied on inherent properties of the DNA sequence, with direct measurement, which was much more accurate. Later, annotation of the human genome relies primarily on deep sequencing of the transcripts in every human tissue using RNA-seq. These experiments have revealed that over [90% of genes contain at least one and usually several alternative splice variants](#), in which the [exons](#) are combined in different ways to produce 2 or more gene products from the same location.

The genome published by the HGP does not represent the sequence of every individual's genome. It is the combined mosaic of a small number of anonymous donors, all of European origin. The HGP genome is a [scaffold](#) for future work in identifying differences among individuals. Subsequent projects sequenced the genomes of multiple distinct ethnic groups, though as of today there is still only one "reference genome."

Later Findings and Accomplishment

Key findings of the draft (2001) and complete (2004) genome sequences include:

1. There are approximately 22,300 protein-coding genes in human beings, the *same* range as in other mammals.
2. The human genome has significantly more **segmental duplications** (nearly identical, repeated sections of DNA) than had been previously suspected.
3. At the time when the draft sequence was published fewer than 7% of **protein families** appeared to be vertebrate specific.

The first printout of the human genome to be presented as a series of books, displayed at the [Wellcome Collection](#), London

The Human Genome Project was started in 1990 with the goal of sequencing and identifying all three billion chemical units in the human genetic instruction set, finding the genetic roots of disease and then developing treatments. It is considered a **Mega Project** because the human genome has approximately 3.3 billion base-pairs. **With the sequence in hand, the next step was to identify the genetic variants that increase the risk for common diseases like cancer and diabetes.**

It was far too expensive at that time to think of sequencing patients' whole genomes. So the National Institutes of Health embraced the idea for a "shortcut", which was to look just at sites on the genome where many people have a variant DNA unit. The rationale behind the "shortcut" was that, since the major diseases are common, so too would be the genetic variants that caused them. **Natural selection** keeps the human genome free of variants that damage health before children are grown, the theory held, but fails against variants that strike later in life, allowing them to become quite common.

(For example: in 2002 the National Institutes of Health started a \$138 million project called the [HapMap](#) to catalog the common variants in European, East Asian and African genomes.)

The genome was broken into smaller pieces; approximately 150,000 base pairs in length. These pieces were then ligated into a type of vector known as "[bacterial artificial chromosomes](#)", or BACs, which are derived from bacterial chromosomes which have been genetically engineered. The vectors containing the genes can be inserted into bacteria where they are copied by the bacterial [DNA replication](#) machinery. Each of these pieces was then sequenced separately as a small "[shotgun](#)" project and then assembled. The larger, 150,000 base pairs go together to create chromosomes. This is known as the "hierarchical shotgun" approach, because the genome is first broken into relatively large chunks, which are then mapped to chromosomes before being selected for sequencing.

Funding came from the US government through the National Institutes of Health in the United States, and a UK charity organization, the [Wellcome Trust](#), as well as numerous other groups from around the world. The funding supported a number of large sequencing centers including those at [Whitehead Institute](#), the [Sanger Centre](#), [Washington University in St. Louis](#), and [Baylor College of Medicine](#).

The United Nations Educational, Scientific and Cultural Organization (UNESCO) served as an important channel for the involvement of developing countries in the Human Genome Project.

Public Versus Private Approaches

In 1998, a similar, privately funded quest was launched by the American researcher [Craig Venter](#), and his firm Celera Genomics. Venter was a scientist at the NIH during the early 1990s when the project was initiated. The \$300-million ($\3×10^8) Celera effort was intended to proceed at a faster pace and at a fraction of the cost of the roughly \$3 billion ($\3×10^9) [publicly funded project](#). The Celera approach was able to proceed at a much more rapid rate, and at a lower cost than the public project because it relied upon data made available by the publicly funded project.

Celera used a technique called *whole genome shotgun sequencing*, employing *pairwise end sequencing*, which had been used to sequence bacterial genomes of up to six million base pairs in length, but not for anything nearly as large as the three billion base pair human genome.

Celera initially announced that it would seek patent protection on “only 200–300” genes, but later amended this to seeking “intellectual property protection” on “fully-characterized important structures” amounting to 100–300 targets. The firm eventually filed preliminary (“place-holder”) patent applications on 6,500 whole or partial genes. Celera also promised to publish their findings in accordance with the terms of the 1996 “[Bermuda Statement](#)”, by releasing new data annually (the HGP released its new data daily), although, unlike the publicly funded project, they would not permit free redistribution or scientific use of the data. The publicly funded competitors were compelled to release the first draft of the human genome before Celera for this reason. On July 7, 2000, the UCSC Genome Bioinformatics Group released a first working draft on the web. The scientific community downloaded about 500 GB of information from the UCSC genome server in the first 24 hours of free and unrestricted access. In March 2000, [President Clinton](#) announced that the *genome sequence* could not be patented, and should be made freely available to all researchers. The statement sent Celera’s stock plummeting and dragged down the [biotechnology](#)-heavy Nasdaq. The biotechnology sector lost about \$50 billion in [market capitalization](#) in two days!

Although the working draft was announced in June 2000, it was not until February 2001 that Celera and the HGP scientists published details of their drafts. Special issues of [Nature](#) (which published the publicly funded project’s [scientific paper](#)) and [Science](#) (which published Celera’s paper) described the methods used to produce the draft sequence and offered analysis of the sequence. These drafts covered about 83% of the genome (90% of the euchromatic regions with 150,000 gaps and the order and orientation of many segments not yet established). In February 2001, at the time of the joint publications, [press releases](#) announced that the project had been completed by both groups. Improved drafts were announced in 2003 and 2005, filling in to approximately 92% of the sequence currently.

Genome Donors

In the IHGSC international [public-sector](#) HGP, researchers collected blood (female) or sperm (male) samples from a large number of donors. Only a few of many collected samples were processed as DNA resources. Thus the donor identities were protected so neither donors nor scientists could know whose DNA was sequenced. DNA clones from many different [libraries](#) were used in the overall project, with most of those libraries being created by [Pieter J. de Jong’s](#). Much of the sequence (>70%) of the [reference genome](#) produced by the public HGP came from a single anonymous male donor from Buffalo, New York (code name RP11).

HGP scientists used [white blood cells](#) from the blood of two male and two female donors (randomly selected from 20 of each) – each donor yielding a separate DNA library. One of these libraries (RP11) was used considerably more than others, due to quality considerations. One minor technical issue is that male samples contain just over half as much DNA from the sex chromosomes (one [X chromosome](#) and one [Y chromosome](#)) compared to female samples (which contain two [X chromosomes](#)). The other 22 chromosomes (the autosomes) are the same for both sexes.

Although the main sequencing phase of the HGP has been completed, studies of DNA variation continue in the [International HapMap Project](#), whose goal is to identify patterns of [single-nucleotide polymorphism](#) (SNP) groups (called [haplotypes](#), or “haps”). The DNA samples for the HapMap came from a total of 270 individuals: [Yoruba people](#) in Ibadan, Nigeria; [Japanese people](#) in Tokyo; [Han Chinese](#) in Beijing; and the French [Centre d’Etude du Polymorphisme Humain](#) (CEPH) resource, which consisted of residents of the United States having ancestry from Western and [Northern Europe](#).

In the Celera Genomics [private-sector](#) project, DNA from five different individuals were used for sequencing. The lead scientist of Celera Genomics at that time, Craig Venter, later acknowledged (in a

public letter to the journal *Science*) that his DNA was one of 21 samples in the pool, five of which were selected for use.

In 2007, a team led by [Jonathan Rothberg](#) published [James Watson](#)'s entire genome, unveiling the six-billion-nucleotide genome of a single individual for the first time.

Developments

The work on interpretation and analysis of genome data is still in its initial stages. It is anticipated that detailed knowledge of the human genome will provide new avenues for advances in [medicine](#) and [biotechnology](#). Clear practical results of the project emerged even before the work was finished. For example, a number of companies, such as [Myriad Genetics](#), started offering easy ways to administer genetic tests that can show predisposition to a variety of illnesses, including [breast cancer](#), [hemostasis disorders](#), [cystic fibrosis](#), [liver](#) diseases and many others. Also, the [etiologies for cancers](#), [Alzheimer's disease](#) and other areas of clinical interest are considered likely to benefit from genome information and possibly may lead in the long term to significant advances in their management.

There are also many tangible benefits for biologists. For example, a researcher investigating a certain form of [cancer](#) may have narrowed down their search to a particular gene. By visiting the human genome database on the [World Wide Web](#), this researcher can examine what other scientists have written about this gene, including (potentially) the three-dimensional structure of its product, its function(s), its evolutionary relationships to other human genes, or to genes in mice or yeast or fruit flies, possible detrimental mutations, interactions with other genes, body tissues in which this gene is activated, and diseases associated with this gene or other datatypes. Further, deeper understanding of the disease processes at the level of molecular biology may determine new therapeutic procedures. Given the established importance of DNA in molecular biology and its central role in determining the fundamental operation of [cellular processes](#), it is likely that expanded knowledge in this area will facilitate medical advances in numerous areas of clinical interest that may not have been possible without them.

The analysis of similarities between DNA sequences from different organisms is also opening new avenues in the study of [evolution](#). In many cases, evolutionary questions can now be framed in terms of [molecular biology](#); indeed, many major evolutionary milestones (the emergence of the [ribosome](#) and [organelles](#), the development of [embryos](#) with body plans, the [vertebrate immune system](#)) can be related to the molecular level. Many questions about the similarities and differences between humans and our closest relatives (the [primates](#), and indeed the other [mammals](#)) are expected to be illuminated by the data in this project.

The project inspired and paved the way for genomic work in other fields, such as agriculture. For example, by studying the genetic composition of *Triticum aestivum*, the world's most commonly used bread wheat, great insight has been gained into the ways that domestication has impacted the evolution of the plant. Which loci are most susceptible to manipulation, and how does this play out in evolutionary terms? Genetic sequencing has allowed these questions to be addressed for the first time, as specific loci can be compared in wild and domesticated strains of the plant. This will allow for advances in genetic modification in the future which could yield healthier, more disease-resistant wheat crops.

Ethical, Legal and Social Issues

At the onset of the Human Genome Project several ethical, legal, and social concerns were raised in regards to how increased knowledge of the human genome could be used to discriminate against people. One of the main concerns of most individuals was the fear that both employers and health insurance companies would refuse to hire individuals or refuse to provide insurance to people because of a health concern indicated by someone's genes.

In 1996 the United States passed the [Health Insurance Portability and Accountability Act \(HIPAA\)](#) which protects against the unauthorized and non-consensual release of individually identifiable health information to any entity not actively engaged in the provision of healthcare services to a patient.

Along with identifying all of the approximately 20,000–25,000 genes in the human genome, the Human Genome Project also sought to address the ethical, legal, and social issues that were created by the onset of the project. For that the Ethical, Legal, and Social Implications (ELSI) program was founded in 1990. Five percent of the annual budget was allocated to address the ELSI arising from the project. This budget started at approximately \$1.57 million in the year 1990, but increased to approximately \$18 million in the year 2014.

Whilst the project may offer significant benefits to medicine and scientific research, some authors have emphasized the need to address the potential social consequences of mapping the human genome. *“Molecularising disease and their possible cure will have a profound impact on what patients expect from medical help and the new generation of doctors’ perception of illness.”*

Pursuit of the study of the biostatistics aspects of human genetic epidemiology, being the mathematical and statistical studies of medical genetic diseases, may be considered as running parallel to the great discoveries of the Human Genome Project.

1.2.1.1 Human Genetics vs Biomedical Genetics

It should not escape ones attention that human genetics is *not* identical with medical genetics. It cannot be denied that human genetics and genomics are having an important and a major impact in all aspects of medicine and across all age groups – and this impact will only increase as knowledge expands as the reach and power of genetic sequencing technology grows. The former refers to the genomics of all human genes (including such aspects of human body physical dimensions, individual’s hair and eye colors, etc.), and the latter is concerned with aspects of human genetics that specifically relate to diseases especially those having genetic origins, such as those aspects of human genetic-related diseases mentioned heretofore.

Biomedical Ethics in Medical Genetics

In any discussion of ethical issues in medical practice, four important principles are considered:

1. **Respect for individual autonomy** – respecting and safeguarding the rights of an individual to control his/her medical information and medical care, without coercion
2. **Avoid maleficence** – “*First of all, do no harm*” – from the Latin phrase: “*primum non nocere*”
3. **Beneficence** – doing good
4. **Justice** – Treat all individuals fairly and equally.

1.2.2 Human Genes, Genetics, and Health

<https://www.betterhealth.vic.gov.au/health/conditionsandtreatments/genes-and-genetics> [G]

A useful working model for understanding human genes, genetics, and genetic epidemiology is as follows:

1. Genes are the blue print for the human bodies.
2. A human genetic mutation implies that a certain gene undergoes a change, not unlike a spelling error on a printed page, that may disrupt the message normally borne by the gene – making the gene “faulty”.
3. Human genetic mutations may occur spontaneously.

4. Occasionally, a faulty gene may be inherited, passing on from parent(s) to children.
5. Human genetic changes that result in a faulty gene may cause a wide variety and range of conditions.
6. Although most related parents will have healthy children, these parents are more likely than unrelated parents to have children with genetic disorders or health problems. **Consanguinity:** Close blood relationship, sometimes used to denote human inbreeding. Mating of closely related persons can cause significant human genetic disease in offspring. ***Everyone carries rare recessive genes that, in the company of other genes of the same type, are capable of causing autosomal recessive diseases.***

Parents may pass on distinguishing traits or characteristics such as hair colors and eye colors to their children through their genes. **Many health conditions and diseases are also genetic.** Genes may also influence some behavioral characteristics, such as intelligence and natural talents. Genes may be considered as the blueprint for human bodies. Almost every cell in the human body contains a copy of this blueprint, mostly stored inside a special containment within the cell called the nucleus. Genes are part of chromosomes, which are long strands of a chemical substance called DeoxyriboNucleic Acid (DNA): therefore, genes are made up of DNAs.

A DNA strand looks like a twisted ladder. The genes are like a series of letters strung along each rung. These letters are used like a book of instructions. The letter sequence of each gene contains information on building specific molecules (such as proteins or hormones, both essential to the growth and maintenance of the human body).

The genes are copied 'letter for letter' to a similar substance called RiboNucleic Acid (RNA). The working parts of the cell read the RNA to create the protein or hormone according to the instructions. Each gene codes the instruction for a single protein only, but one protein may have many different roles in the human body. Also, one characteristic, such as eye color, may be influenced by **many** genes.

Sometimes, a gene contains a variation – like a spelling mistake – that disrupts the gene's coded message. A variation may occur spontaneously (causes unknown) or it may be inherited. Variations in the coding that make a gene not work properly are called *mutations* and may, directly or indirectly, lead to a wide range of conditions.

Chromosomes and Sperm and Egg Cells

Humans have 46 paired chromosomes, with about 23,000 genes. The 46 chromosomes in the human cell are made up of 22 paired chromosomes. These are numbered from 1 to 22 according to size, with chromosome number 1 being the biggest. These numbered chromosomes are called *autosomes*. Cells in the body of a woman also contain two sex chromosomes called X chromosomes, in addition to the 44 autosomes. Body cells in men contain an X and a Y chromosome and 44 autosomes.

The 23,000 genes *come in pairs*. One gene in each pair is inherited from the person's father and the other from their mother. A sperm and an egg each contain one copy of every gene needed to make up a person (one set of 23 chromosomes each). When the sperm fertilizes the egg, two copies of each gene are present (46 chromosomes), and so a new life can begin.

The chromosomes that decide the gender of the baby are called sex chromosomes. The mother's egg always contributes an X, while the father's sperm provides either an X or a Y. An XX pairing means a girl, while an XY pairing means a boy. As well as determining gender, these chromosomes carry genes that control other body functions. There are many genes located on the X chromosome, but only a few on the Y chromosome.

Inheritance of Human Characteristics

Human characteristics may be inherited in many different ways: one characteristic can have many different forms – for example, blood type can be A, B, AB or O. Variations in the gene for that characteristic cause these different forms. Each variation of a gene is called an allele (pronounced ‘Allel’). One may inherit different alleles of the gene pair (one from each parent) in different ways:

- (i) **Dominant and recessive genes:** The two copies of the genes contained in each set of chromosomes both send coded messages to influence the way the cell works. The actions of some of these genes, however, appear to be ‘dominant’ over others. Generally, for example, the coded message from the genes that tells the eye cells to make brown color is dominant over blue eye color. However, a number of different genes together determine eye color and so blue-eyed parents may have a child with brown eyes.
- (ii) **Dominant and recessive blood-group inheritance:** Dominant inheritance occurs when one allele of a gene is *dominant* within the pair. For blood groups, the A allele is dominant over the O allele, so a person with one A allele and one O allele has the blood group AO. In other words, the O group is *recessive* – a person needs two O alleles to have the blood group O. Thus a child may have blood group A because the blood group A gene inherited from the mother is dominant over the blood group O gene inherited from the father. If the mother has an A allele and an O allele (AO), her blood group will be A because the A is dominant. The father has two O alleles (OO), so he has the blood group O. Each one of their children has a 50% chance of having blood group A (AO) and a 50% chance of having blood group O (OO), depending on which alleles they inherit.
- (iii) **Co-dominant gene:** Not all genes are either dominant or recessive. Sometimes, each allele in the gene pair carries *equal* weight and will show up as a combined physical characteristic. For example, with blood groups, the A allele is as ‘strong’ as the B allele. So someone with one copy of A and one copy of B has the blood group AB.
- (iv) **Genotype and Phenotype:** *Genotype* and *phenotype* are terms commonly used in human genetics. Thus, a person with the alleles AO will have the blood group A. The observable trait – blood group – is known as the *phenotype*. The genotype is the genes that produce the observable trait. So the person with blood group A and AO alleles has the blood group A phenotype but the AO genotype.
- (v) **Chemical Communication:** Although every cell has two copies of the 23,000 genes, each cell needs only some specific genes to be switched on in order to perform its particular functions. The unnecessary genes are switched off. Genes communicate with the cell in chemical code, known as the *genetic code*. The cell carries out its instructions to the letter. A cell reproduces by copying its genetic information then splitting in half, forming two individual cells. Occasionally, a mistake is made, causing a variation (genetic mutation) and the wrong chemical message is sent to the cell. This spontaneous genetic mutation can cause problems in the way the person’s body functions.
- Genetic mutations are permanent.** Some of the causes of a spontaneous genetic mutation include exposure to radiation, chemicals and cigarette smoke. Genetic mutations also build up in our cells as one ages.
- (vi) **Variations in the Genes in the Cells:** Sperm and egg cells are known as ‘germ’ cells, while every other cell in the body is called ‘somatic’. If a variation in the information in a gene (viz., mutation) happens spontaneously in a person’s somatic cells, they may develop the condition related to that gene change, but will not pass it on to their children. For example, skin cancer can be caused by a build-up of spontaneous mutations in genes in the skin cells caused by damage from UV radiation.

However, if the mutation occurs in a person's germ cells, that person's children each have a 50 per cent chance of inheriting the faulty (mutated) gene. Sometimes, a parent may have one copy of a gene that is faulty and the other copy containing the correct information. They are said to 'carry' the faulty gene although they themselves will not have the condition caused by the faulty gene – they are a ***genetic carrier for the condition***.

The correct copy of a gene overrides the faulty copy. For example, the gene controlling red-green color recognition is located on the X chromosome. A mother who carries the faulty gene causing red-green color blindness on one of her X chromosome copies will have perfectly normal vision, as she still has a functioning gene copy for red-green color recognition on her other X chromosome.

However, her sons have a 50%-chance of being colorblind! This is owing to the condition there is a 50%-chance that they will inherit the X chromosome from their mother that contains the faulty gene. There is also a 50 per cent chance that they will inherit the X chromosome containing the correct copy of the gene and so will have normal vision.

Genetic Conditions

To date, scientists have identified around 1,700 conditions caused directly or indirectly by changes in the genes. Around half of all miscarriages are caused by changes in the total number of genes in the developing baby. Similarly, about half of a country's population will be affected at some point in their life by an illness that is at least partly genetic in origin

The three ways in which genetic conditions can happen are:

- (1) The variation in the gene that makes it faulty (viz., a mutation) happens spontaneously in the formation of the egg or sperm, or at conception.
- (2) The faulty gene is passed from parent to child and may directly cause a problem that affects the child at birth or later in life.
- (3) The faulty gene is passed from parent to child, and may cause a genetic susceptibility. Usually, environmental factors, such as diet and exposure to chemicals, combine with this susceptibility to trigger the onset of the disorder.

Genetic Predisposition (Inherited Susceptibility)

In many cases, being born with a faulty gene associated with a particular disease does not mean one is destined to develop that particular disease. ***It simply means that such a person will likely be at increased risk of developing the condition.*** Many conditions involving genetic susceptibility, such as some types of cancer, need to be ***triggered*** by environmental factors such as diet and lifestyle. For example, prolonged exposure to the sun is linked to melanoma. Avoiding such triggers means significantly reducing the risks.

Indeed: "***Nature loads the gun, and Nurture may pull the trigger.***"

Regarding the mutual dependence and supports between Nature and Nurture, perhaps an additional epistemic and religious viewpoint may shed some insight. From the pen of the following well-known American thought leader in health sciences*:

*White, E. G. (1905).- "The Ministry of Healing": Pages 261–266, <http://whiteestate.org/search/search.asp>

Regarding the principles of healthful living, natural remedies, this respected author admonished:

"Institutions for the care of the sick would be far more successful if they could be established away from the cities. And so far as possible, all who are seeking to recover health should place themselves amid country surroundings where they can have the benefit of outdoor life. Nature is God's physician. The pure air, the glad sunshine, the flowers and trees, the orchards and vineyards, and outdoor exercise amid these surroundings, are health-giving, life-giving."

Physicians and nurses should encourage their patients to be much in the open air. Outdoor life is the only remedy that many invalids need. It has a wonderful power to heal diseases caused by the excitements and excesses of fashionable life, a life that weakens and destroys the powers of body, mind, and soul.

How grateful to the invalids weary of city life, the glare of many lights, and the noise of the streets, are the quiet and freedom of the country! How eagerly do they turn to the scenes of nature! How glad would they be to sit in the open air, rejoice in the sunshine, and breathe the fragrance of tree and flower! There are life-giving properties in the balsam of the pine, in the fragrance of the cedar and the fir, and other trees also have properties that are health restoring.

To the chronic invalid, nothing so tends to restore health and happiness as living amid attractive country surroundings. Here the most helpless ones can sit or lie in the sunshine or in the shade of the trees. They have only to lift their eyes to see above them the beautiful foliage. A sweet sense of restfulness and refreshing comes over them as they listen to the murmuring of the breezes. The drooping spirits revive. The waning strength is recruited. Unconsciously the mind becomes peaceful, the fevered pulse more calm and regular. As the sick grow stronger, they will venture to take a few steps to gather some of the lovely flowers, precious messengers of God's love to His afflicted family here below.

Plans should be devised for keeping patients out of doors. For those who are able to work, let some pleasant, easy employment be provided. Show them how agreeable and helpful this outdoor work is. Encourage them to breathe the fresh air. Teach them to breathe deeply, and in breathing and speaking to exercise the abdominal muscles. This is an education that will be invaluable to them.

Exercise in the open air should be prescribed as a life-giving necessity. And for such exercises there is nothing better than the cultivation of the soil. Let patients have flower beds to care for, or work to do in the orchard or vegetable garden. As they are encouraged to leave their rooms and spend time in the open air, cultivating flowers or doing some other light, pleasant work, their attention will be diverted from themselves and their sufferings.

The more the patient can be kept out of doors, the less care will he require. The more cheerful his surroundings, the more helpful will he be. Shut up in the house, be it ever so elegantly furnished, he will grow fretful and gloomy. Surround him with the beautiful things of nature; place him where he can see the flowers growing and hear the birds singing, and his heart will break into song in harmony with the songs of the birds. Relief will come to body and mind. The intellect will be awakened, the imagination quickened, and the mind prepared to appreciate the beauty of God's word.

In nature may always be found something to divert the attention of the sick from themselves and direct their thoughts to God. Surrounded by His wonderful works, their minds are uplifted from the things that are seen to the things that are unseen. The beauty of nature leads them to think of the heavenly home, where there will be nothing to mar the loveliness, nothing to taint or destroy, nothing to cause disease or death.

Let physicians and nurses draw from the things of nature, lessons teaching of God. Let them point the patients to Him whose hand has made the lofty trees, the grass, and the flowers, encouraging them to see in every bud and flower an expression of His love for His children. He who cares for the birds and the flowers will care for the beings formed in His own image."

Genes and Genetics – Consanguinity Inherited from Related Parents

Many cultures practice marriages between relatives such as first cousins (especially those with the same maternal or paternal grandparents). The objectives of such intermarriages are often to bolster family unity and keep wealth within the family. A relationship between blood-related people is called consanguinity – meaning ‘shared blood’ in Latin.

Consanguinity is Often Associated with Factors Such as:

- cultural and religious practices
- isolated groups (such as migrants) who prefer to marry within their own culture
- low socioeconomic status
- illiteracy
- living in rural areas.

Related parents are more likely than unrelated parents to have children with health problems or genetic disorders. This is owing to the two parents sharing one or more common ancestors and so carry some of the same genetic material. If both partners carry the same inherited altered (mutated) gene, their children are more likely to have a genetic disorder.

Related couples should seek advice from a clinical genetics service if their family has a history of a genetic condition or mental and emotional deviations.

Autosomal Recessive Genetic Disorders

If two parents have a copy of the same altered gene, they may both pass their copy of this altered gene on to a child, so the child receives both altered copies. As the child then does not have a normal, functioning copy of the gene, the child will most likely develop the disorder. This is called **autosomal recessive inheritance**. The parents are ‘carriers’ of the genetic condition but are *unaffected themselves*. Autosomal recessive genetic disorders are more likely if two parents are related, although they are still quite rare.

Examples of autosomal recessive genetic disorders include cystic fibrosis and phenylketonuria (PKU). When both parents are carriers of the same altered gene, there is a one in four (25%) chance that each pregnancy will be affected. Other children of the same parents may also be affected or may be carriers, having only one copy of the altered gene.

A child with only one copy of the altered gene will not be affected, as that child also has a normal copy of that gene – the same as the healthy parents.

Degrees of Relationship

Relatives are described by the closeness of their blood relationship. For example:

- First-degree relatives share half their genetic information. First-degree relatives include a person’s siblings, non-identical twin, parents, and children.
- Second-degree relatives share one-quarter of their genetic information. Second-degree relatives include a person’s half-siblings, uncles and aunts, nephews and nieces, and grandparents.
- Third-degree relatives share one-eighth of their genetic material and include a person’s first cousins, half-uncles, half-aunts, half-nephews and half-nieces.
- Generally speaking, the closer the genetic relationship between the parents, the greater the risk of birth defects for their children.

Incidence of Birth Defects in Children of Related Parents

* A child of unrelated parents has a risk of around 2 to 3% of being born with a serious birth defect or genetic disorder.

This risk is approximately doubled (to between 4% and 6%) for children of first cousins *without* a family history of genetic disorders. The risk of birth defects or death for children of first-degree relatives – for example, parent and child or brother and sister – rises to about 30%.

Genetic Counselling and Testing

Some genetic services may provide information and counselling for couples considering prenatal diagnosis or following diagnosis of fetal abnormalities, and referral to community resources including support groups if needed. A couple who suspect they may be related may seek genetic counselling. If the family has a history of a known autosomal recessive genetic disorder, genetic testing may be possible to see whether the couple are both carriers of the condition.

Points of Critical Human Genetic Issues

- Human genes are the blueprint for our bodies.
- A human genetic mutation means that a gene contains a change – like a spelling mistake – that disrupts the gene message (makes the gene faulty).
- Human genetic mutations can occur spontaneously.
- Sometimes a faulty human gene is inherited, which means it is passed on from parent to child.
- Human genetic changes that make a human gene faulty can cause a wide range of conditions.
- Although most related parents will have healthy children, they are more likely than unrelated parents to have children with health problems or genetic disorders.

1.2.3 A Glossary of Common Terms in Human Genetics

(A collection of common terms often encountered in the study and description of human genetics is included towards the end of this book, designated **Glossary**, to be followed by the sections **References** and the **Index**.)

1.2.4 Human Genetics in Medicine^[W]

^[W] Wikipedia

Human Genetics in Medicine, also known as Medical Genetics or Clinical Genetics, is the branch of **medicine** that includes the diagnosis and management of **hereditary disorders**. **Medical genetics** differs from **human genetics** in that human genetics is a field of scientific research that may or may not apply to medicine, while medical genetics refers to the application of genetics to medical care. For example, research on the causes and inheritance of **genetic disorders** would be considered within both human genetics and medical genetics, while the diagnosis, management, and counselling patients and associated people with genetic disorders would be considered part of medical genetics.

In contrast, the study of typically non-medical **phenotypes** such as the genetics of eye and hair color would be considered part of human genetics, but not necessarily relevant to medical genetics (except in situations such as albinism). **Genetic Medicine** is a newer term for medical genetics and incorporates areas such as **gene therapy**, **personalized medicine**, and the rapidly emerging new medical specialty, **predictive medicine**.

Medical Genetics encompasses many different areas, including the clinical practice of physicians, genetic counselors, and nutritionists, clinical diagnostic laboratory activities, and research into the causes and concomitant results of genetic disorders. Thus, the scope of medical genetics include **autism**, and mitochondrial disorders, birth defects and dysmorphology, mental retardation, skeletal dysplasia, connective tissue disorders, cancer genetics, teratogens, and prenatal diagnosis. This specialty, viz., medical genetics, is increasingly becoming relevant to many common diseases. Overlaps with other medical specialties are beginning to develop, as recent advances in genetics are revealing etiologies for neurologic, endocrine, cardiovascular, pulmonary, ophthalmologic, renal, psychiatric, dermatologic conditions, etc.

Many of the individual fields within medical genetics are hybrids between clinical care and research. This is due in part to recent advances in science and technology (for example, advances in the [Human Genome project](#)) that have enabled an unprecedented understanding of [genetic disorders](#).

Clinical genetics is the practice of clinical medicine with particular attention to [hereditary disorders](#). Referrals are made to genetics clinics for a variety of reasons, including [birth defects](#), [developmental delay](#), [autism](#), [epilepsy](#), [short stature](#), and many others. Examples of genetic syndromes that are commonly seen in the genetics clinic include [chromosomal rearrangements](#), [Down syndrome](#), etc. In the United States, physicians who practice clinical genetics are accredited by the American Board of Medical Genetics and Genomics (ABMGG). To become a board-certified practitioner of Clinical Genetics, a physician must complete a minimum of 24 months of training in a program accredited by the ABMGG. Individuals seeking acceptance into clinical genetics training programs must hold an M.D. or equivalent degree, and have completed a minimum of 24 months of training in an accredited [residency](#) program in [internal medicine](#), [pediatrics](#), [obstetrics and gynecology](#), or other medical specialty.

Sub-Specialties of Medical Genetics Include:

1. Metabolic/Biochemical Genetics

Metabolic (or biochemical) genetics involves the diagnosis and management of [inborn errors of metabolism](#) in which patients have enzymatic deficiencies that perturb [biochemical](#) pathways involved in metabolism of [carbohydrates](#), [amino acids](#), and [lipids](#). Examples of metabolic disorders include:

- [galactosemia](#),
- [glycogen storage disease](#),
- [lysosomal storage disorders](#),
- [metabolic acidosis](#),
- [peroxisomal disorders](#),
- [phenylketonuria](#), and
- [urea cycle disorders](#).

2. Cytogenetics

Cytogenetics is the study of [chromosomes](#) and [chromosome abnormalities](#). While cytogenetics usually relied on [microscopy](#) to analyze chromosomes, new molecular technologies such as [array comparative genomic hybridization](#) are now becoming widely used. Examples of chromosome abnormalities include aneuploidy, chromosomal rearrangements, and genomic deletion/duplication disorders.

3. Molecular Genetics

Molecular genetics involves the discovery of and laboratory testing for **DNA** mutations that underlie many **single gene disorders**. Examples of single gene disorders include **achondroplasia**, **cystic fibrosis**, **Duchenne muscular dystrophy**, hereditary **breast cancer** (BRCA1/2), **Huntington disease**, **Marfan syndrome**, **Noonan syndrome**, and **Rett syndrome**. Molecular tests are used in the diagnosis of syndromes involving **epigenetic** abnormalities, such as **Angelman syndrome**, **Beckwith-Wiedemann syndrome**, **Prader-willi syndrome**, and **uniparental disomy**.

4. Mitochondrial Genetics

Mitochondrial genetics concerns the diagnosis and management of **mitochondrial** disorders, which have a molecular basis but often result in biochemical abnormalities owing to deficient energy production.

Genetic Counseling

Genetic counseling is the process of providing information about genetic conditions, diagnostic testing, and risks in other family members, within the framework of nondirective counseling.

Modern Aspects of Chromosome Studies Include:

- **Chromosome studies** are used in the general genetics clinic to determine a cause for developmental delay/mental retardation, birth defects, dysmorphic features, and/or autism.
- **Chromosome analysis** is also performed in the prenatal setting to determine whether a fetus is affected with aneuploidy or other chromosome rearrangements.
- Finally, **chromosome abnormalities** are often detected in cancer samples. A large number of different methods have been developed for chromosome analysis:
- **Chromosome analysis** using a **karyotype** involves special stains that generate light and dark bands, allowing identification of each chromosome under a microscope.
- **Fluorescence in Situ Hybridization** (FISH) involves fluorescent labeling of probes that bind to specific DNA sequences, used for identifying aneuploidy, genomic deletions or duplications, characterizing chromosomal translocations and determining the origin of **ring chromosomes**.
- **Chromosome Painting** is a technique that uses fluorescent probes specific for each chromosome to differentially label each chromosome. This technique is more often used in cancer cytogenetics, where complex chromosome rearrangements can occur.
- **Array Comparative Genomic Hybridization** is a new molecular technique that involves hybridization of an individual DNA sample to a glass slide or microarray chip containing molecular probes (ranging from large ~200kb **bacterial artificial chromosomes** to small oligonucleotides) that represent unique regions of the genome. This method is particularly sensitive for detection of genomic gains or losses across the genome but does not detect balanced translocations or distinguish the location of duplicated genetic material (for example, a tandem duplication versus an insertional duplication).

Basic Metabolic Studies

Biochemical studies are performed to screen for imbalances of metabolites in the bodily fluid, usually the blood (plasma/serum) or urine, but also in cerebrospinal fluid (CSF). Specific tests of enzyme function (either in leukocytes, skin fibroblasts, liver, or muscle) are also used. In the USA, the **newborn screen** incorporates biochemical tests to screen for treatable conditions such as **galactosemia** and

phenylketonuria (PKU). Patients suspected to have a metabolic condition might undergo the following tests:

- Quantitative amino acid analysis is typically performed using the ninhydrin reaction, followed by *liquid chromatography* to measure the amount of amino acid in the sample (either urine, plasma/serum, or CSF). Measurement of amino acids in plasma or serum is used in the evaluation of *disorders of amino acid metabolism* such as *urea cycle disorders*, *maple syrup urine disease*, and *PKU*. Measurement of amino acids in urine can be useful in the diagnosis of *cystinuria* or renal *Fanconi syndrome* as in *cystinosis*.
- Urine organic acid analysis can be either performed using quantitative or qualitative methods, but in either case the test is used to detect the excretion of abnormal *organic acids*. These compounds are normally produced during bodily metabolism of amino acids and odd-chain fatty acids, but accumulate in patients with certain *metabolic conditions*.
- The acylcarnitine combination profile detects compounds such as organic acids and fatty acids conjugated to carnitine. The test is used for detection of disorders involving fatty acid metabolism, including *MCAD*.
- Pyruvate and lactate are byproducts of normal metabolism, particularly during *anaerobic metabolism*. These compounds normally accumulate during exercise or ischemia, but are also elevated in patients with disorders of pyruvate metabolism or mitochondrial disorders.
- Ammonia** is an end product of amino acid metabolism and is converted in the liver to **urea** through a series of enzymatic reactions termed the *urea cycle*. Elevated ammonia can therefore be detected in patients with *urea cycle disorders*, as well as other conditions involving *liver failure*.
- Enzyme testing is performed for a wide range of metabolic disorders to confirm a diagnosis suspected based on screening tests.

Molecular Studies

- DNA sequencing** is used to directly analyze the genomic DNA sequence of a particular gene. In general, only the parts of the gene that code for the expressed protein (**exons**) and small amounts of the flanking untranslated regions and **introns** are analyzed. Therefore, although these tests are highly specific and sensitive, they do not routinely identify all of the mutations that could cause disease.
- DNA methylation** analysis is used to diagnose certain genetic disorders that are caused by disruptions of *epigenetic* mechanisms such as *genomic imprinting* and *uniparental disomy* (viz., only one, of the two copies, is turned on! This inherited copy may be from either one of the two parents).
- To detect fragments of DNA separated by size, one may use *gel electrophoresis* and detect using radiolabeled probes. This test was routinely used to detect deletions or duplications in conditions such as *Duchenne muscular dystrophy* but is being replaced by high-resolution *array comparative genomic hybridization* techniques. Southern blotting is still useful in the diagnosis of disorders caused by *trinucleotide repeats*.

Treatments

Each cell of the body contains the hereditary information (**DNA**) wrapped up in structures called **chromosomes**. Since genetic syndromes are typically the result of alterations of the chromosomes or genes, there is no treatment currently available that can correct the genetic alterations in every cell of the body. **Therefore, there is currently no “cure” for genetic disorders**. However, for many genetic syndromes there is treatment available to manage the symptoms. In some cases, particularly *inborn errors of metabolism*, the mechanism of disease is well understood and offers the potential for dietary

and medical management to prevent or reduce the long-term complications. In other cases, *infusion therapy* is used to replace the missing enzyme. Current research is actively seeking to use *gene therapy* or other new medications to treat specific genetic disorders.

Management of Metabolic Disorders

In general, metabolic disorders arise from enzyme deficiencies that disrupt normal metabolic pathways.

For instance, in the hypothetical example:



- Compound “A” is metabolized to “B” by enzyme “X”, compound “B” is metabolized to “C” by enzyme “Y”, and compound “C” is metabolized to “D” by enzyme “Z”.
- If enzyme “Z” is missing, compound “D” will be missing, while compounds “A”, “B”, and “C” will build up. The pathogenesis of this particular condition could result from lack of compound “D”, if it is critical for some cellular function, or from toxicity due to excess “A”, “B”, and/or “C”. Treatment of the metabolic disorder could be achieved through dietary supplementation of compound “D” and dietary restriction of compounds “A”, “B”, and/or “C” or by treatment with a medication that promoted disposal of excess “A”, “B”, or “C”. Another approach that can be taken is enzyme replacement therapy, in which a patient is given an infusion of the missing enzyme.

Diet

Dietary restriction and supplementation are key measures taken in several well-known metabolic disorders, including [galactosemia](#), [phenylketonuria \(PKU\)](#), [maple syrup urine disease](#), organic acidurias, and [urea cycle disorders](#).

Such restrictive diets can be difficult for the patient and family to maintain, and require close consultation with a nutritionist who has special experience in metabolic disorders. The composition of the diet will change depending on the caloric needs of the growing child and special attention is needed during a pregnancy if a woman is affected with one of these disorders.

Medication

Medical approaches include enhancement of residual enzyme activity (in cases where the enzyme is made but is not functioning properly), inhibition of other enzymes in the biochemical pathway to prevent buildup of a toxic compound, or diversion of a toxic compound to another form that can be excreted.

Example 1: Medical Treatments

(1) Use of Vitamin B6

Include:

- the use of high doses of [pyridoxine](#) (Vitamin B6) in some patients with [homocystinuria](#) to boost the activity of the residual cystathione synthase enzyme,

- (b) administration of **biotin** to restore activity of several enzymes affected by deficiency of **biotinidase**,
- (c) treatment with **NTBC** in **Tyrosinemia** to inhibit the production of succinylacetone which causes liver toxicity, and
- (d) the use of **sodium benzoate** to decrease **ammonia** build-up in **urea cycle disorders**.

Example 2: Medical Therapies

Enzyme Replacement Therapy

Certain **lysosomal storage diseases** are treated with infusions of a recombinant enzyme (produced in a laboratory), which can reduce the accumulation of the compounds in various tissues.

Gaucher disease, **Fabry disease**, **Mucopolysaccharidoses** and **Glycogen storage disease Type II**. Such treatments are limited by the ability of the enzyme to reach the affected areas (for example: the **blood brain barrier** prevents enzyme from reaching the brain), and sometimes may be associated with allergic reactions.

The long-term clinical effectiveness of enzyme replacement therapies vary widely among different disorders.

Other Examples

- Angiotensin receptor blockers in Marfan syndrome & Loeys-Dietz
- Bone marrow transplantation
- Gene therapy

(4) Some Notable Requirements and Approaches in the Treatment of Genetic Diseases

*** From Mendel's Genetics to the Complete Human Genome [NWM]**

[NWM] Nussbaum, R. L., McInnes, R. R., Willard, H. F., Hamosh, A. (2016).- “Thompson & Thompson: Genetics in Medicine”, 8/e, Elsevier, Philadelphia, PA 19103“Genome – Your Health is Personal”, Fall 2015, ISSN 2374-5800, Vol 2,

One may note that, at the beginning of the twenty-first century, the Human Genome Project provided a complete sequence of human DNA – the human genome (the suffix “-ome” is borrowed from the Greek language, meaning “complete” or “all”) – thus allowing the human genes to be studied in their entirety, **advancing Genetic Medicine to Genomic Medicine!**

**** The Human Genome as the Chromosomal Basis of Heredity, and on to an Understanding of the Impact on Human Genetic Epidemiology**

Central to understanding the role of genetics in medicine is knowing the organization, variations, and functional transmission of the human genome, in addition to the principles of genomics and personalized medicine. From this beginning, the first major contribution from human genomic medicine to human genetic epidemiology is the possible understanding of the impact of human genomics on human health on a broader scale! Thus, one should appreciate that every individual has ones own unique genomic sequence of input of genetic products, resulting in response to the totality of inputs of the genome sequence as well as ones individual set of experiences and environmental exposures – resulting in a very personal “chemical individuality” – a unique assembly!

1.2.4.1 The Human Genome

The Modern Era of Genetic Medicine

Stanford University Biochemistry Professor Paul Berg, PhD, and the winner of the 1980 Nobel Prize in Chemistry, described the new era of genetic medicine, in comparison with classical medicine, in relative terms: whereas a knowledge and practice of traditional medicine depends on an in-depth knowledge of human anatomy, biochemistry, and physiology, dealing with future diseases will require a similar in-depth understanding of the molecular anatomy, physiology, and biochemistry of the human genome! Thus, we need a more such detailed knowledge, and how human genes are organized and regulated, and how they function. Furthermore, we should have physicians who are as conversant with the molecular anatomy and physiology of genes and chromosomes, as the cardio-thoracic surgeons are with the workings and anatomy of the human heart!

Human Genetic Diversity: Mutation and Polymorphism

Between any two unrelated humans, the sequence of nuclear DNA is about 99.5% identical – yet it is precisely such a small fraction of DNA sequence difference among individuals that is responsible for the ***genetically determined variability*** that is evident both in one's daily existence and on outward appearance, whereas other differences are directly responsible for causing diseases! Between these two apparent extremes is the variation responsible for ***genetically-determined variability*** in anatomy, physiology, susceptibility to infection, dietary intolerances, predisposition to many types of cancers, therapeutic responses to adverse reactions to medicines, as well as possibly variability in personality traits, and artistic or athletic aptitudes, musical talents, etc.

The most common and simplest of all polymorphisms in DNA are Single Nucleotide Polymorphisms (SNPs, *pronounced “snips”!*). In a later chapter, statistical computations will be undertaken involving these SNPs.

Clinical Cytogenetics and Genome Analysis

As applied to medical practices, clinical cytogenetics is the study of chromosomes, their structure, and their inheritance. For over 50 years, it is well-known that chromosome abnormalities, viz., microscopically visible changes in the *number* or *structures* of chromosomes, could account for many clinical conditions which may be considered as ***chromosome disorders***! Focusing on the complete set of genetic material, these cytogeneticists then further considered a genome-wide perspective to the practice of medicine! Currently, ***chromosome analysis***, with increasing precision and resolution at both the genomic and cytological levels, has become a critically important diagnostic procedure in clinical medicine – including ***chromosomal microarrays*** and ***whole-genome sequencing*** – which are typically impressive improvements in resolution and capacity.

It is now well-known that chromosome disorders form an important and major category of genetic disease, accounting for a major proportion of all reproductive wastage, congenital malformation, and intellectual disability and is an important factor in the pathogenesis of cancer! Certain specific cytogenetic disorders may well be responsible for hundreds of syndromes that collectively have become more common than all the single-gene diseases together. Cytogenetic abnormalities are found in:

- (i) Nearly 1% of live births
- (ii) about 2% of pregnancies in women older than 35 years, and who undergo prenatal diagnoses,
- (iii) about 50% of all spontaneous, first-trimester abortions!

Chromosomal and Genomic Bases of Diseases: Disorders of Autosomes and Sex

Chromosomes

The most common (and best understood) chromosomal and genomic disorders encountered in clinical practice may be linked to the principles of ***dosage balance and imbalance*** at the level of chromosomes and sub-chromosomal regions of the genome. Overall, there are at least 5 different categories of such abnormalities, each of which may lead to disorders of clinical significance. They are disorders owing to:

Table 1.1 summaries these distinguishes features of the underlying mechanism

Remarks

- (1) **Aneuploidy:** The most common human mutation involved errors in chromosome segregation, leading to the production of an abnormal gamete that has two copies or no copies of the chromosome involved in the non-disjunction events, mainly: trisomy 21 (Down Syndrome), trisomy 18 and 13. Each of these autosomal trisomes as associated with growth retardation, intellectual disability, and multiple congenital anomalies.
- (2) **Down Syndrome^[W]** is the most common and best known of the chromosome disorders, and is the single most common genetic cause of moderate intellectual disability. About 1 child in 850 is born with this abnormality, and among liveborn children or fetuses of mothers 35 years of age or older, the incidence of trisomy 21 is much higher!

(Trisomy 21 is a genetic condition caused by an *extra* chromosome: normally babies inherit 23 chromosomes from each of the 2 parents: for a total of $(23 \times 2 = 46)$ chromosomes. Babies with Down syndrome “Trisomy 21”, however, end up with 3 chromosomes at Position 21, instead of the usual pair! More than 90% of Down syndrome cases are caused by trisome 21!

Patterns of Single-Gene Inheritance

In biology, an ***allosome*** is a sex chromosome.

An **autosome** is a chromosome that is *NOT* an allosome.

A **diploid** cell is a cell that contains 2 sets of chromosomes. Each chromosome pair is considered to be one set of homologous chromosomes. A single chromosome set consists of 2 chromosomes, 1 of which is inherited from each of the 2 parents!

Humans have a diploid genome that usually contains 22 autosome ***pairs*** and 1 allosome ***pair***, making a total of $(22 + 1) = 23$ ***pairs*** for a total of $(23 \times 2) = 46$ chromosomes.

Table 1.1 Mechanisms of chromosome abnormalities and genomic imbalances

Category	Consequences
(Underlying Mechanisms) (Examples)	(Examples)
(1) Abnormal chromosome segregation (non-disjunction)	Aneuploidy (Down syndromes)
(2) Recurrent chromosomal syndromes (Recombination at segmental duplication)	Duplication (Copy Number Variations)
(3) Idiopathic chromosome abnormalities (Sporadic, variable breakpoints) (De novo balanced translocations)	Deletion Syndromes (Gene Disruptions)
(4) Unbalanced familial abnormalities (Unbalanced segregations)	Offsprings of Balanced Translocations (Offsprings of Pericentric Inversions)
(5) Syndromes involving genomic imprinting (Any event that reveals imprinted genes)	Prader Willi/Angelman Syndromes (Offsprings of Willi/Angelman Syndromes)

Autosomal recessive diseases occur only in individuals with 2 mutant alleles and no wild-type allele. Such homozygotes must have inherited a mutant allele from each parent, each of whom is a heterozygote for the allele. When a disorder shows recessive inheritance, the mutant allele responsible generally reduces or eliminates the function of the gene product, a “loss of function mutation.

Complex Inheritance of Common Multifactorial Disorders

During their lifetimes, nearly 2 out of every 3 persons suffer or prematurely die of common disease such as:

- **Alzheimer diseases,**
- **birth defects,**
- **cancer,**
- **diabetes,**
- **myocardial infarction,**
- **neuropsychiatric disorders !**

Many of these diseases ‘run in families’ so that the cases appear to cluster among the relatives of the affected individuals more frequently than in the general population.

Genetic Variations Within Populations

Population genetics is the quantitative study of the distribution of genetic variations in population and of the maintenance or changes over time both within and between populations. It deals both with:

- (i) genetic factors, such as reproduction and mutation, and with
- (ii) societal and environmental factors, such as migration and selection which together determine the distribution and frequency of genotypes and alleles in ov

Example 3: A Common Autosomal Trait Governed by a Single Pair of Alleles

Consider the gene *CCR5* (which encodes a cell surface cytokine receptor that serves as an entry point for certain strains of the *Human Immunodeficiency Virus (HIV)*, which causes the *Acquired Immunodeficiency Syndrome (AIDS)*). A 32-bp deletion on this gene results in an allele ($\Delta CCR5$) that encodes a non-functional protein owing to a premature termination and a frameshift. Individuals homozygous for this allele ($\Delta CCR5$) do not express the receptor on the surface of their immune cells, and therefore are resistant to HIV infection. Moreover, the loss of function of *CCR5* seems to be a benign trait, and its only known phenotypic result is the resistance to HIV infection. Table 1.2 shows a sampling of some 788 case subjects, from Europe, which illustrates the distribution of individuals who were homozygous for the wildtype *CCR5* allele, which is homozygous for the $\Delta CCR5$ allele, or heterozygous.

Table 1.2 Genotype frequencies for the wild type *CCR5* allele and the $\Delta CCR5$ deletion allele

Genotype	Number of case subjects	Observed genotype frequency	Allele	Derived allele frequencies
<i>CCR5/CCR5</i>	647	0.821		
<i>CCR5/ΔCCR5</i>	134	0.168	<i>CCR5</i>	0.906
<i>CR5/ΔCCR5</i>	7	0.011	$\Delta CCR5$	0.094
Total	788	1.000		

Data from Nussbaum, R. L., McInnes, R. R., and Williard, H. F. (2016).- Thompson & Thompson – Genetics in Medicine”, 8/e, p.156, Elsevier, Philadelphia, PA

Based on the observed genotype frequencies, one may determine directly the allele frequencies by counting the alleles. The population frequency of an allele may be obtained by considering a *hypothetical gene pool* as a collection of all the alleles at a specific locus for the whole population. For autosomal loci, the size of the gene pool at one locus is twice the number of individuals in the population because each autosomal genotype consists of 2 alleles, viz.:

- a $\Delta CCR5/\Delta CCR5$ individual has 2 $\Delta CCR5$ alleles, and
- a $CCR5/\Delta CCR5$ individual has one of each.

Thus, in this example, the observed frequency of the *CCR5 allele* is:

$$f_{CCR5} = [(2 \times 647) + (1 \times 134)] / (788 \times 2) = 1,428 / 1,576 = 0.906$$

Similarly, one may compute the frequency of the *$\Delta CCR5$ allele* as 0.094, by adding up how many $\Delta CCR5$ alleles are present:

$$f_{\Delta CCR5} = [(2 \times 7) + (1 \times 134)] / (788 \times 2) = 148 / 1,576 = 0.094$$

Alternately, one may obtain the frequency of the *$\Delta CCR5$ allele* as

$$1 - 0.906 = 0.094,$$

because the frequencies of the *two alleles must add up to 1*:

$$\text{viz., } f_{CCR5} + f_{\Delta CCR5} = 1$$

Harry-Weinberg (Ideal) Equilibrium Model

In human population genetics, as in mathematical anthropology and biology, an important element in such disciplines would be a mathematical description of the behavior of the alleles in a given population. The **Hardy-Weinberg (Ideal) Equilibrium Model** is generally adopted as a useful reference model.

Criteria of the Hardy-Weinberg Law:

- The population (under study) is large,*
- All matings are random with respect to the locus,*
- Allele frequencies remain constant over time, because:*
 - There is no appreciable rate of new mutations.
 - Individuals with all genotypes are equally capable of mating and passing on their genes; viz., there is no selection against any particular genotype.
 - There has been no significant immigration of individuals from a population with allele frequencies significantly different from the endogenous population.

Any population that reasonably and realistically appears to meet the above set of criteria may be considered to be in **Hardy-Weinberg (H-W) Equilibrium**. Moreover, if a population meets these equilibrium criteria, there exists a simple mathematical equation for computing genotype frequencies from allele frequencies! This equilibrium equation is called the **Hardy-Weinberg Law**, which is the cornerstone of population genetics. This law was named after an English pure mathematician, Godfrey Hardy, at Cambridge University, and a German physician Wilhelm Weinberg. This law has two critical components:

Criterion I Under certain idealized, as stated in the *H-W Equilibrium*, a simple relationship exists between allele frequencies and genotype frequencies in a population:

In the gene pool, if p is the frequency of allele A , and q is the frequency of allele a , and assuming the alleles combine into genotypes randomly to get a population: viz., mating in the population is *entirely at random with respect to the genotypes at this locus*, then

- the chance that two A alleles will pair up, to form the AA genotype, is p^2 ;
- the chance that two a alleles will pair up, to form the aa genotype, is q^2 ; and
- the chance that one A allele and one a allele will pair up, to form the Aa genotype, is $2pq$, in which the factor 2 is derived from the fact the A allele could be inherited from the father and the a allele from the mother, or vice versa.

The **Hardy-Weinberg Law** states that the frequency of the 3 genotypes AA , Aa , and aa is given respectively by the 3 terms of the Binomial Expansion of:

$$(p+q)^2 = p^2 + 2pq + q^2$$

This law applies to all autosomal loci and to the X chromosome in females, but **not** to X -linked loci in males who have only a single X chromosome.

REMARKS:

- Applying the Hardy-Weinberg Law to the $CCR5$ system in **Example 3**, with relative frequencies of the two alleles in the population of 0.906, for the wild-type allele $CCR5$, and 0.094, for $\Delta CCR5$, it follows that the relative proportions of the three combinations of the genotype alleles are:

$$\begin{aligned} p^2 &= 0.906 \times 0.906 = 0.821, \text{ for a case subject with 2 wild-type } CCR5 \text{ alleles,} \\ q^2 &= 0.094 \times 0.094 = 0.009, \text{ for 2 } \Delta CCR5 \text{ alleles, and} \\ 2pq &= (0.906 \times 0.094) + (0.094 \times 0.906) \\ &= 0.170, \text{ for one } CCR5 \text{ and one } \Delta CCR5 \text{ allele.} \end{aligned}$$

- Applying the genotype frequencies, computed by the Hardy-Weinberg Law, to a population of 788 case subjects, the derived number of case subjects with the three different genotypes (647:134:7), are identical to the actual observed numbers in Table 1.2. Thus, when the assumptions of the Hardy-Weinberg Law are applicable in a population, one should expect the genotype frequencies (0.821:0.170:0.009) to remain constant in that population for all subsequent generations.
- This law may be applied for genes with more than 2 alleles. For example, if a locus had 3 alleles, with frequencies p , q , and r , then the at

$$(p + q + r)^2$$

Example 4: Genotypes and Phenotypes in Populations Allele and Genotype Frequencies in Populations

In general terms, the genotypic frequencies for any fixed number of alleles and with allele frequencies $p_1, p_2, p_3, \dots, p_n$ may be derived from the terms of the expansion of

$$(p_1 + p_2 + p_3 + \dots + p_n)^2$$

Another characteristic of the Hardy-Weinberg Law is: if allele frequencies do not change from generation to generation, then the proportion of the genotypes will not change either. That is, the

Table 1.3 Frequencies of Parental Mating Types for a Population in Hardy-Weinberg Equilibrium with Parental Genotypes in the Proportion $p^2: 2pq: q^2$

Type of parental matings			
Case	Father	Mother	Frequency
1	AA	AA	$p^2 \times p^2 = p^4$
2	Aa	AA	$2pq \times p^2 = 2p^3q$
3	AA	Aa	$p^2 \times 2pq = 2p^3q$
4	aa	AA	$q^2 \times p^2 = p^2q^2$
5	AA	aa	$p^2 \times q^2 = p^2q^2$
6	Aa	Aa	$2pq \times 2pq = 4p^2q^2$
7	aa	Aa	$q^2 \times 2pq = 2pq^3$
8	Aa	aa	$2pq \times q^2 = 2pq^3$
9	aa	aa	$q^2 \times q^2 = q^4$

Table 1.4 Frequencies of Offsprings for a Population in Hardy-Weinberg Equilibrium with Parental Genotypes in the Proportion $p^2 : 2pq : q^2$

The Offsprings			
Case	AA	Aa	aa
1	p^4		
2	$\frac{1}{2}(2p^3q)$	$\frac{1}{2}(2p^3q)$	
3	$\frac{1}{2}(2p^3q)$	$\frac{1}{2}(2p^3q)$	
4		p^2q^2	
5		p^2q^2	
6	$\frac{1}{4}(4p^2q^2)$	$\frac{1}{4}(4p^2q^2)$	$\frac{1}{4}(4p^2q^2)$
7		$\frac{1}{2}(2pq^3)$	$\frac{1}{2}(2pq^3)$
8		$\frac{1}{2}(2pq^3)$	$\frac{1}{2}(2pq^3)$
9			q^4

population genotype frequencies from generation to generation will remain constant, at equilibrium, if the allele frequencies p and q remain constant. Specifically, when there is random mating in a population that is at equilibrium, and genotypes AA, Aa, and aa are present in the proportions

$$p^2 : 2pq : q^2$$

A proof of this equilibrium is shown in Tables 1.3 and 1.4:

Identifying the Genetic Bases for Human Diseases

To identify genetic contributions to diseases, medical geneticists examine families and populations. The disease may be inherited in a recognizable mendelian pattern, as mentioned in section “[Patterns of single-gene inheritance](#)”. The different genomic and genetic variations, carried by affected family members of the may affect:

- (a) individuals in the population that cause disease directly, or
- (b) their susceptibility to diseases.

Genomic researches have supplied medical geneticists with:

- a list of all known human genes,
- knowledge of their structures and locations, and
- lists of millions of variants in DNA sequence found among individuals in different populations.

As a result of all these researches, several analytical approaches have been developed allowing medical geneticists to relate particular genes associated with specific diseases, as well as the variants that they contain that may contribute, or associated with, specific human diseases. In particular, 3 approaches appear to be relevant:

1. **Linkage Analysis:** This family-based approach considers the explicit advantage of family pedigrees in following the inheritance of any disease among family members, and to test for repeated and consistent coinheritance of any disease associated with a specific genomic region, or with particular variants, whenever the disease is inherited in a family.
2. **Association Analysis:** This population-based approach considers the entire history of a population and seek for decreased or increased frequency of a certain allele or group of alleles in a sample of affected case subjects taken from the population, compared with a control group of unaffected members from the same population. This approach may be particularly effective for complex diseases which do not reveal a Mendelian inheritance pattern.
3. **Direct Genome Sequencing:** This approach considers any sequencing of affected case subject and their parents and/or other people in the family or population. It is useful for rare Mendelian disorders where linkage analysis is performing linkage analysis or because the disorder is a genetic condition that always results from new mutations and is not inherited. In these cases, sequencing the genome, or simply coding the exons of every gene, the exome) of an affected individual used to find the gene responsible for the disorder. This approach takes advantage of newly developed technology that has reduced the cost of DNA sequencing a million fold from previous processes in which the original reference genome was prepared.

These 3 approaches for mapping and identifying diseased genes has had a big impact on the understanding of the pathophysiology and pathogenesis of many diseases. Over time, a knowledge of the genetic contribution to diseases may also suggest novel and effective methods of treatment, management, and prevention!

The Molecular Basis of Genetic Diseases

Molecular Mutations – A Basis of Genetic Diseases

Over 60 years ago, the concept of a *Molecular Disease* was introduced when referring to an illness in which the primary disease-causing event was a change, either acquired or inherited, acting on a gene, its structure, or its expression. First, the basic biochemical and genetic mechanisms, underlying single-gene or monogenetic disorders, are described. This is then illustrated, in terms of their molecular and clinical results, by considering the inherited sicknesses of hemoglobin (the hemoglobinopathies).

Genetic diseases occur when a change in the DNA of an essential gene changes the function and/or amount of the gene products – typically the messenger RNAs (mRNA), protein, or specific non-coding RNAs (ncRNA) with regulatory or structural functions. Most known single-gene disorders come from mutations that affect the function of a protein, there appears to be some exceptions! These exceptions are the diseases resulting from mutations in ncRNA, including microRNA (miRNA) genes

that encode transfer RNAs (tRNA). Thus one need to understand genetic diseases at the molecular and biochemical levels as this forms the foundation of rational therapy. To begin with, one should first understand the causes of diseases owing to defects in protein-coding genes, to be followed by the study of phenotype at the level of proteins, as well as the biochemistry and metabolism – constituting the field of biochemical genetics.

An Overview of a Useful Published Source Mendelian Inheritance in Man (2014)

Reference:

Nucleic Acids Res. 2015 Jan 28; 43(Database issue): D789–D798.

Published online 2014 Nov 26. doi: <https://doi.org/10.1093/nar/gku1205>

OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders, by

[Amberger, J. S.](#), [Bocchini, C.A.](#), [Schiettecatte, F.](#), [Scott, A. F.](#), and [Hamosh, A.](#)

Online Mendelian Inheritance in Man, OMIM, is a comprehensive, authoritative and timely research resource of curated descriptions of human genes and phenotypes and the relationships between them. The new official website for OMIM, OMIM.org (<http://omim.org>), was launched in January 2011. OMIM is based on the published peer-reviewed biomedical literature and is used by overlapping and diverse communities of clinicians, molecular biologists and genome scientists, as well as by students and teachers of these disciplines. Genes and phenotypes are described in separate entries and are given unique, stable six-digit identifiers (MIM numbers). OMIM entries have a structured free-text format that provides the flexibility necessary to describe the complex and nuanced relationships between genes and genetic phenotypes in an efficient manner. OMIM also has a derivative table of genes and genetic phenotypes, the Morbid Map. OMIM.org has enhanced search capabilities such as genome coordinate searching and thesaurus-enhanced search term options. Phenotypic series have been created to facilitate viewing genetic heterogeneity of phenotypes. Clinical synopsis features are enhanced with UMLS, Human Phenotype Ontology and Elements of Morphology terms and image links. All OMIM data are available for FTP download and through an API. MIMmatch is a novel outreach feature to disseminate updates and encourage collaboration.

In the next section, this overview of genetic disease mechanisms will be expanded to include other major genetic diseases in medicine.

An Example of Notable Molecular Mutations with Important Medical Genetic Implications: Fig. 1.3

The Molecular, Biochemical, and Cellular Bases of Genetic Diseases

It is anticipated that in the coming decades, many more of the approximately 25,000 coding genes in the human genome may well be associated with both monogenic and genetically complex but well-known diseases, including:

- Alzheimer's Disease
- Amyotrophic Lateral Sclerosis (ALS), viz., Lou Gehrig's Disease
- Cystic Fibrosis
- Huntington's Disease
- Muscular Dystrophy
- Parkinson's Disease
- Phenyl-Ketonuria

Examples of notable Mutations

2nd base				G
U	C	A		
UUU (Phe/F) Phenylalanine	UCU (Ser/S) Serine	UAU (Tyr/Y) Tyrosine	UGU (Cys/C) Cysteine	
UUC (Phe/F) Phenylalanine	UCC (Ser/S) Serine	UAC (Tyr/Y) Tyrosine	UGC (Cys/C) Cysteine	
UUA (Leu/L) Leucine	UCA (Ser/S) Serine	UAA Ochre (Stop)	UGA Opal (Stop)	
UUG (Leu/L) Leucine	UCG (Ser/S) Serine	UAG Amber (Stop)	UGG (Trp/W) Tryptophan	
β-Thalassemia				
CUU (Leu/L) Leucine	CCU (Pro/P) Proline	CAU (His/H) Histidine	CGU (Arg/R) Arginine	
CUC (Leu/L) Leucine	CCC (Pro/P) Proline	CAC (His/H) Histidine	CGC (Arg/R) Arginine	
CUA (Leu/L) Leucine	CCA (Gln/Q) Glutamine	CAA (Gln/Q) Glutamine	CGA (Arg/R) Arginine	
CUG (Leu/L) Leucine	CCG (Pro/P) Proline	CAG (Gln/Q) Glutamine	CGG (Arg/R) Arginine	
McArdle's disease				
AUU (Ile/I) Isoleucine	ACU (Thr/T) Threonine	AAU (Asn/N) Asparagine	AGU (Ser/S) Serine	
AUC (Ile/I) Isoleucine	ACC (Thr/T) Threonine	AAC (Asn/N) Asparagine	AGC (Ser/S) Serine	
Prostate cancer				
AUA (Ile/I) Isoleucine	ACA (Thr/T) Threonine	AAA (Lys/K) Lysine	AGA (Arg/R) Arginine	
AUG (Met/M) Methionine	ACG (Thr/T) Threonine	AAG (Lys/K) Lysine	AGG (Arg/R) Arginine	
Colorectal cancer				
GUU (Val/V) Valine	GCU (Ala/A) Alanine	GAU (Asp/D) Aspartic acid	GGU (Gly/G) Glycine	
GUC (Val/V) Valine	GCC (Ala/A) Alanine	GAC (Asp/D) Aspartic acid	GGC (Gly/G) Glycine	
GUA (Val/V) Valine	GCA (Ala/A) Alanine	GAA (Glu/E) Glutamic acid	GGA (Gly/G) Glycine	
GUG (Val/V) Valine	GCG (Ala/A) Alanine	GAG (Glu/E) Glutamic acid	GGG (Gly/G) Glycine	

3rd base in each row

ΔF508 deletion in cystic fibrosis

Clinically important missense mutations generally change the properties of the coded amino acid residue between being basic, acidic, polar or nonpolar, while nonsense mutations result in a stop codon.

Amino acids

- Basic
- Acidic
- Polar
- Nonpolar (hydrophobic)

Polyglutamine (PolyQ) Diseases

- Huntington's disease
- Spinocerebellar atrophy (SCA)
- (most types)
- Spinobulbar muscular atrophy (Kennedy disease)
- Dentatorubral-pallidoluysian atrophy

Mutation type

- = Trinucleotide repeat
- = Deletion
- = Missense
- = Nonsense

Friedreich's ataxia

Sickle-cell disease

Reference: Mutation [W]

Fig. 1.3 Examples of notable mutations

Neurodegenerative diseases which are chronic and progressive, are characterized by loss of neurons in motor, sensory, or cognitive systems. Examination of the patterns of cell loss and the identification of disease-specific cellular markers have contributed to the nosologic classification. For example:

- senile plaques, neurofibrillary tangles, neuronal loss, and acetylcholine deficiency, etc. define **Alzheimer's Disease**;
- Cellular inclusions and swollen motor axons are found in **Amyotrophic Lateral Sclerosis (ALS)**;
- Lewy bodies and depletion of dopamine characterize **Parkinson's Disease**;
- γ -aminobutyric acid-containing neurons of the neostriatum are lost in **Huntington's Disease**; etc.

Mendelian inheritance can be found in *all* these disorders . . .

Reference: Martin, J. B. (June 24, 1999).- New England Journal of Medicine, **DOI: 10.1056/NEJM199906243402507**

The Treatment of Genetic Diseases

A rational treatment and therapy for genetic diseases requires the understanding of these diseases *at a molecular level*. Clearly, a significant impact on the treatment of genetic conditions and associated disorders will be, in the coming years, the cataloging of more human genes, RNA, and protein therapies. One may begin by considering new strategies for treating genetic diseases, including the therapies that are supported by the genetic approach to medical practices, while first focusing on single-gene diseases.

Treating genetic diseases may eliminate or reduce the effects of the disorder – or with the therapy that may be lifelong inconvenient. By way of genetic counseling, the family of the case subject will be informed (for several generations) regarding the concomitant risk that the disease may occur in other members.

For *single-gene disorders* owing to loss-of-function mutations, treatment may consist of:

- replacing the defective protein or RNA, say, by direct administration, cell or organ transplantation or gene therapy
- minimizing the consequences of its deficiency, and
- improving its function.

Replacement of the defective gene product (protein or RNA) may be done:

- directly
- by gene therapy, or
- by organ or cell transplant.

Developmental Genetics and Birth Defects

Developmental genetics, at its current stage of development, support a working development which allow the medical practitioners to undertake diagnostic evaluation of a patient with birth defects. In this context, the physicians can is

- predict prognosis
- recommend management options, and
- provide accurate recurrence risks for the parents and other close relatives of the affected babies.

From early understanding of the principles of genetics [Sinnott et al. 1950], breeders of animals and plants had known for centuries that inbreeding, or mating of individuals closely related in descent, often results in reduced size, lessened nyIt has *always* been well-known that, within the Chinese culture, intermarriages between individuals of the same last name, or family name or surname, should be avoided to minimize the possible undesirable effects of consanguinity. Human history, however, testified that some marriages between relatives gave progeny afflicted with hereditary diseases, while other such marriages can produce healthy offsprings: it had been reported that in ancient Egypt, royal families had been “successfully” maintained for generations by preferred-and-selective brother-sister marriages!

From a public health perspective, the medical impact of birth defects is considerable and growing! For the U.S.A., in the most recent year, 2013, for which statistics are available:

- The **Infant Mortality** rate was 6 infant deaths per 1000 live birth: and more than 20% of infant deaths were due to genetic birth defects. Nearly 50% of the death of infants are due to derangements of normal development.
- As of 2010, 1 in 68 births in the U.S.A. are autistic, viz., diagnosed with **Autism Spectrum Disorder** (ASD), with boys affected 4 times more frequently than girls!
- ALS (**Amyotrophic Lateral Sclerosis**)^[W], also known as Motor Neuron Disease (MND) and as Lou Gehrig’s Disease (named after the famous American baseball player Lou Gehrig who was so affected in 1939), is a specific disease which causes the death of neurons controlling voluntary muscles. This disease is characterized by stiff muscles, muscle twitching, and gradually deteriorating weakness owing to muscles decreasing in size, and thus resulting in increasing difficulty in swallowing, speaking, and finally breathing! *About 5 to 10% of the cases are known to have been inherited from the case subject’s parents – about half of these genetic cases are due to one or two specific genes.* The diagnosis of this disease is based on a person’s signs and symptoms, with testing carried out to rule out other potential causes. In the U.S.A. and Europe, this disease affects about 2 people per 100,000 per year. **No cure for ALS is known.** Currently, gene therapy is being experimented – using stem cells!

• **Parkinson’s Disease**

[<https://www.healthline.com/health/what-causes-parkinsons-disease#loss-of-dopamine>]

Parkinson’s disease is a chronic disorder of the nervous system. It affects at least **500,000 people** in the United States, according to the National Institute of Neurological Disorders and Stroke. Approximately 60,000 new cases are reported in the United States each year. This disease is not fatal, but it can cause debilitating symptoms that impact everyday movement and mobility. Hallmark symptoms of this disease include tremors and gait and balance problems. These symptoms develop because the brain’s ability to communicate is damaged. Researchers are not yet certain what causes Parkinson’s. There are several factors that may contribute to the disease.

1. Genetics

- **Some studies** suggest that genes play a role in the development of Parkinson’s. An estimated 15 percent of people with Parkinson’s have a family history of the condition. The **Mayo Clinic** reports that someone with a close relative (e.g., a parent or sibling) who has Parkinson’s is at an increased risk of developing the disease. It also reports that the risk of developing Parkinson’s is low unless you have several family members with the disease.
- How does genetics factor into Parkinson’s in some families? According to **Genetics Home Reference**, one possible way is through the mutation of genes responsible for producing dopamine and certain proteins essential for brain function.

2. Environment

- There is also some evidence that one's environment can play a role. Exposure to certain chemicals has been suggested as a possible link to Parkinson's disease. These include **pesticides** such as insecticides, herbicides, and fungicides. It is also possible that **Agent Orange** exposure may be linked to Parkinson's.
- Parkinson's has also been potentially linked to **drinking well water** and **consuming manganese**.
- Not everyone exposed to these environmental factors develops Parkinson's. **Some researchers suspect** that a combination of genetics and environmental factors cause Parkinson's.

3. Lewy Bodies

- Lewy bodies are abnormal clumps of proteins found in the brain stem of people with Parkinson's disease. These clumps contain a protein that cells are unable to break down. They surround cells in the brain. In the process they interrupt the way the brain functions.
- Clusters of Lewy bodies cause the brain to degenerate over time. This causes problems with motor coordination in people with Parkinson's disease.

4. Loss of Dopamine

Dopamine is a neurotransmitter chemical that aids in passing messages between different sections of the brain. The cells that produce dopamine are damaged in people with Parkinson's disease. Without an adequate supply of dopamine the brain is unable to properly send and receive messages. This disruption affects the body's ability to coordinate movement. It can cause problems with walking and balance.

5. Age and Gender

Aging also plays a role in Parkinson's disease. Advanced age is the most significant risk **factor** for developing Parkinson's disease. Scientists believe that **brain and dopamine function begin to decline** as the body ages, making a person more susceptible to Parkinson's. Gender also plays a role in Parkinson's disease.

6. Occupations

Some research suggests that certain occupations may put a person at greater risk for developing Parkinson's. In particular, Parkinson's disease may be more likely for people who have jobs in welding, agriculture, and industrial work. This may be because individuals in these occupations are exposed to toxic chemicals. However, **study results have been inconsistent** and more research needs to be done.

- Congenital anomalies are a major cause of long-term morbidity, intellectual disability, and other dysfunctions that limit the productivity of affected individuals. For example: 1 in 800 babies are born with the **Down Syndrome**: Fig. 1.4a, b[W]

(An **idiogram** is a diagrammatic representation of chromosome morphology characteristic of a species or population.)

Classification of Birth Defects^[W]

Every year, about 7.9 million infants (6% of worldwide births) are born with serious birth defects. With the causes of over 50% of birth defects unknown, how does one diagnose and prevent them?

Genetic causes of birth defects fall into three general categories:

- chromosomal abnormalities,
- single-gene defects, and
- multifactorial influences.

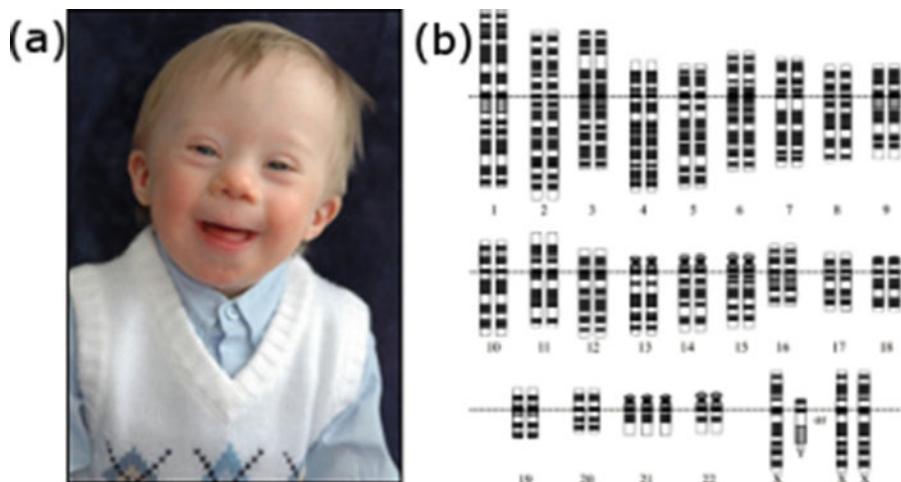


Fig. 1.4 Primary Down Syndrome, Caused by the Presence of 3 copies of Chromosome 21: (a) (Typically) A child who has Down Syndrome. (b) Idiogram of a person who has primary Down Syndrome

Prenatal environments can play a major role in the development of defects in all three categories, especially those linked to multifactorial causes. Medical geneticists classify birth defects into 3 categories:

1. **Disruptions**
2. **Malformations**
3. **Deformations**

Cancer Genetics and Genomics

Cancer describes the more virulent forms of **neoplasia**, which normally is a disease process characterized by uncontrolled cellular proliferation leading to a tumor or mass, viz., a neoplasia. The abnormal gathering of cells in a neoplasm occurs owing to an imbalance between the normal processes of cellular proliferation and cellular attrition: cells proliferate as they pass through the cell cycle and undergo **mitosis**.

(**Mitosis**, a process of cell duplication, or reproduction, during which one cell gives rise to two genetically identical daughter cells. Here, the term *mitosis* is used to describe the duplication and distribution of **chromosomes**, the structures that carry the genetic information, see Fig. 1.5)

Prior to the onset of mitosis, the chromosomes have replicated and the proteins that will form the mitotic spindle have been synthesized. Mitosis begins at prophase at **prophase** with the thickening and coiling of the chromosomes. The nucleus, a rounded structure, shrinks and disappears. The end of prophase is marked by the beginning of the organization of a group of fibers to form a spindle and the disintegration of the nuclear membrane.

The chromosomes, each of which is a double structure consisting of duplicate chromatids, line up along the midline of the cell at **metaphase**. In **anaphase** each chromatid pair separates into two identical chromosomes that are pulled to opposite ends of the cell by the spindle fibers. During **telophase**, the chromosomes begin to **decondense**, the spindle breaks down, and the nuclear membranes and nucleoli re-form. The cytoplasm of the mother cell divides to form two daughter cells, each containing the same number and kind of chromosomes as the mother cell. The stage, or phase, after the completion of mitosis is called **interphase**.

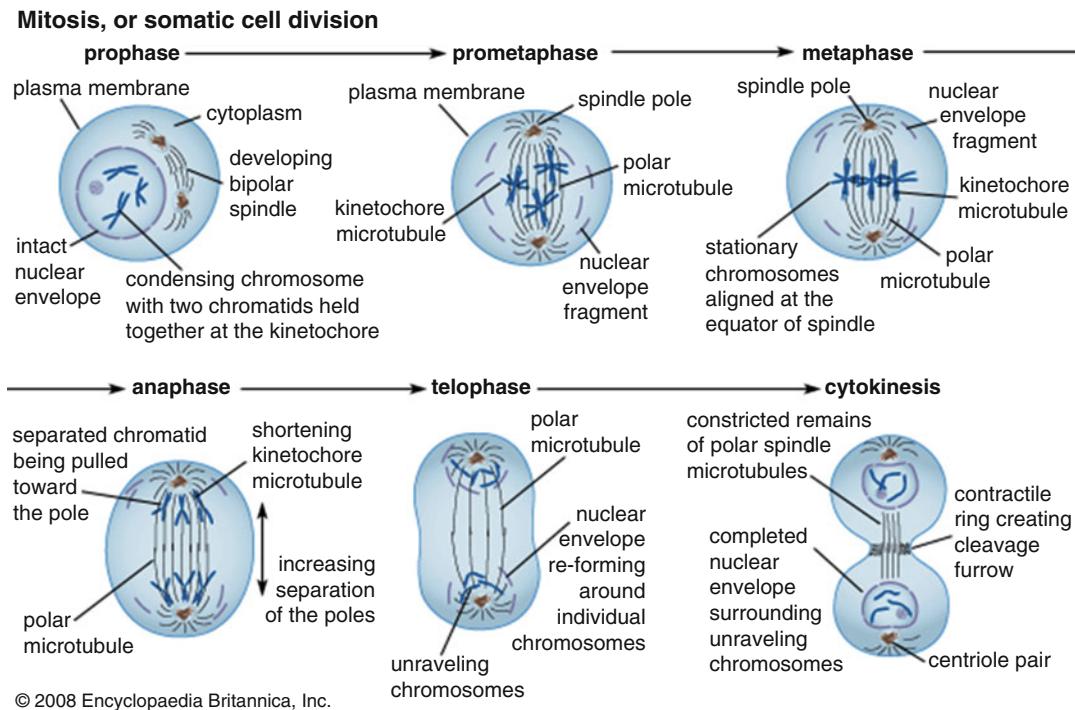


Fig. 1.5 The process of mitosis

Mitosis is essential to life: it provides new cells for growth and for replacement of worn-out cells. This process may take minutes or hours, depending upon the kind of cells and species of organisms. It is influenced by time of day, temperature, and chemicals.

- **Heredity: During mitosis**

When the chromosomes condense during cell division, they have already undergone replication. Each chromosome thus consists of **two identical replicas**, called chromatids, joined at a point called the centromere. During mitosis the sister chromatids separate, one going to each daughter cell.

Genetic Basis of Cancer

To the study of cancer, the application of:

- expression studies (see **1.2.4.1.1 The Modern Era of Genetic Medicine**) and
- sequencing technologies (see section “[Human genetic diversity: mutation and polymorphism](#)”) for genome sequencing and of RNA expression,

results for over 30 types of cancers, published in “**The Cancer Genome Atlas**”, provided a public catalog of human cell mutations, epigenomic modifications, and abnormal gene expression profiles found in a variety of cancers. For example, the

number of mutations in a tumor may vary from a few to tens of thousands! Many mutations found through sequencing of tumor tissues seem to be random, are **not**

recurrent in any particular cancer type – probably occurring **as the cancer develops** rather than directly causing the neoplasia to progress or develop! These are the “passenger” mutations! Nevertheless, there appears to be some subsets, of a few hundred genes each, which will undergo repeated

mutations at high frequencies in many samples of the same type of cancer or in multiple different types of cancers, mutating far too frequently to be considered as passenger mutations. These are the “driver” genes – since they undertake mutations (the driver gene mutations) that are likely to be causing a cancer to progress! For example: although many driver genes are specific to specific tumor types, some are found in the majority of cancers of many different types: such as the TP53 gene encoding the p53 protein.

Currently, this catalog of driver cancer genes is rapidly growing!

Risk Assessment and Genetic Counseling

In the diagnosis and risk assessment of genetic diseases, family history is of great importance, especially in assessing the risk for complex disorders: compare section “[Complex inheritance of common multifactorial disorders](#)” Complex Inheritance of Common Multifactorial Disorders, which allow the geneticist to produce effective evaluation of risks for diseases in the relatives of the affected case subjects. And as mentioned in section “[Genetic variations within populations](#)”, family history is also important when a geneticists evaluate the risk for complex disorders. Since a case subject’s genes are shared with blood relatives, the family history may provide the clinicians with information on the effect that a person’s genetic makeup may have on one’s health, basing upon the medical behavior, and lifestyle, diet; and thus relatives may provide indicators of one’s own genetic susceptibilities. And if some family members do share environmental factors, such as behavior, lifestyle, and diet, etc., they may providing additional critical information regarding both shared genes and shared environmental factors

Risk assessment based upon family history may indicate the following critical levels:

I. High Risk Factors:

- (a) Age at onset of a critical disease in a first-degree relative comparatively early compared to the general population
- (b) Two affected first-degree relatives
- (c) One first-degree relative with unknown or late onset of a disease, and an affected second-degree relative with premature disease from the same lineage
- (d) Two second-degree paternal or maternal relatives with at least one having premature onset of the disease
- (e) Three or more affected paternal or maternal relatives
- (f) Presence of a “moderate risk” family history on both sides of the pedigree.

II. Moderate Risk Factors:

- (a) One first-degree relative with unknown or late onset of the disease
- (b) Two second-degree relatives, from the same lineage, with late or unknown disease onset

III. Average Risk Factors:

- (a) No affected relatives
- (b) Only one affected second-degree relative from one or both sides of the pedigree
- (c) No known family history
- (d) Adopted person with relatively unknown family history!

Clinical Genetic Counselling

The scope of the counselling should cover the associated psychological, social, and medical issues of hereditary diseases. The following procedure is generally followed:

- (a) Diagnose correctly, involving laboratory testing (including generic testing) to ascertain the responsible mutations.

- (b) Ensure the case subject, and the associated family members, understand and appreciate the nature and concomitant consequences and implications of the genetic disorder.
- (c) Provide all appropriate treatment and management, such as referrals to other available professional support as needed.

Just as the unique and major characteristic of any genetic disease is its likelihood to recur within families, the unique feature of genetic counseling is to focus on both the case subject and also on both the present and future members of the associated family members. Thus, the responsibilities of genetic counselors should include the following:

- (a) Work with the case subject, as well as inform other family members of their potential risks.
- (b) Make available mutation tests, as well as other tests, for providing the most accurate assessments for the other family members.
- (c) To both the case subject, as well as the family members, explain what steps are being undertaken to minimize the associated risks.

It should not escape ones attention that genetic counseling includes both identifying and informing individuals at risk for diseases, exploring and communicating the complex psycho-social issues associated with a genetic disease to a family, as well as providing counseling to assist individuals to adapt and adjust to the impact and implications of the genetic disorder within the affected family. This may call for periodic follow-up contacts with the family of the case subject to review and, if necessary, update the process to provide on-going medical support.

Prenatal Diagnosis and Screening

In the practice of medical genetics, prenatal diagnosis refers to the testing of a fetus, *already* known to be at elevated risks owing to some genetic disorder, to verify if the fetus is affected or not in view of the disorder in question. In many situations, an elevated risk is suspected and recognized owing to:

- (a) the birth of a previous child with the disease, coupled with
- (b) a family history of birth disorders,
- (c) a positive parental carrier test, and/or
- (d) when a prenatal screening test indicates an increased risk.

Prenatal diagnosis, such as obtaining fetal cells, or amniotic fluid for analysis, often requires an invasive procedure such as **CVS** (Chorionic Villus Sampling) or **amniocentesis**. Generally speaking, such prenatal diagnosis is undertaken only to provide a definitive result as to whether the impending fetus is positively affected a with specific disorder.

Prenatal screening refers to testing for specific, but common, birth defects (such as neural tube defects, chromosomal aneuploidies, and other structural anomalies in pregnancies that are not known to be increased risks for genetic disorders or birth defects. Concomitantly, a number of screening tests have been developed for testing common birth defects often occurring in not known to be at elevated risks. These would not have been offered for prenatal diagnosis. The preferred screening tests are preferably **noninvasive** – depending on obtaining an imaging, usually by MRI (Magnetic Resonance Imaging) or ultrasonography, or by maternal blood sampling. These tests are suitable for screening all pregnant women.

Genomics for Medicine and Personal Health

At last, a choice!

Since many human illnesses are genetically-induced in normal human life, it begs the question: Is there an optimum environment in which these sicknesses may find its least likely to develop so that living in such an environment will provide the best induced! The next paragraph provides a qualified best suggestion:

Medical Genetic Screening

Medical genetic screening is a population-based approach for finding case subjects who show increased susceptibility with respect to a specific genetic disease. This screening, at the population level, is unlike testing for affected case subjects or carriers within a specific family that has been identified owing to family history -as described in section “[Risk assessment and genetic counseling](#)”. The main objective of medical genetic screening of a population, independent of clinical status and of family history, is to examine all members of a designated population for variants relevant to disease and health. The information so obtained may be applied to the whole population. This information to be demonstrated that subject, as well as its usefulness for guiding health care. As such, genetic

screening constitutes an important public health activity that may become increasingly important as more and improved screening tests are made available for the determination of genetic susceptibilities for various diseases.

The Newborn Screening Program

In genetic screening programs, the known is the public-supported and government-mandated programs which identify pre-symptomatic infants with diseases for which early intervention may prevent, or ameliorate, the consequences if not treated. In screening newborns, the presence of diseases is generally not assessed via direct determining the genotype. In most cases, asymptomatic newborns are screened for abnormalities in the level of various substances in the blood. Such abnormalities, in the metabolites, call for additional evaluation – to confirm or rule out the presence of a disorder.

Some Critical Social and Ethical Issues in Genetic Medicine^[W]

**** In Contact With Nature** Nature chose for our first parents the surroundings best adapted for their health and happiness. They were not placed in a palace or surrounded with the artificial adornments and luxuries that so many today are struggling to obtain. They were placed in close touch with nature and in close communion with the holy ones of heaven.

In the garden that was prepared as a home with graceful shrubs and delicate flowers greeted the eye at every turn. There were trees of every variety, many of them laden with fragrant and delicious fruit. On their branches the birds caroled their songs of praise. Under their shadow the creatures of the earth sported together without a fear. In the beginning, mankind in their untainted purity, delighted in the sights and sounds. Their assigned work was in the garden: “to dress it and to keep it.” Each day’s labor brought them health and gladness. Daily they were great lessons. The plan of life which originally appointed for our humans has lessons for us all.

The more closely this plan of life is followed, the more wonderfully will suffering humanity be restored to health. The sick need to be brought into close touch with nature. An outdoor life amid natural surroundings would work wonders for many a helpless and almost hopeless invalid.

The noise and excitement and confusion of the cities, their constrained and artificial life, are most wearisome and exhausting to the sick. The air, laden with smoke and dust, with poisonous gases, and with germs of disease, is a peril to life. The sick, for the most part shut within four walls, come almost to feel as if they were prisoners in their rooms. They look out on houses and pavements and hurrying crowds, with perhaps not even a glimpse of blue sky or sunshine, of grass or flower or tree. Shut up in this way, they brood over their suffering and sorrow, and become a prey to their own sad thoughts.

And for those who are weak in moral power, the cities abound in dangers. In them, patients who have unnatural appetites to overcome are continually exposed to mental, emotional, and moral challenges. They need to be placed amid new surroundings where the current of their thoughts will be changed; they need to be placed under influences wholly different from those that have wrecked their lives. Let them for a season be removed from these poisonous influences away and into a purer atmosphere. Institutions for the care of the sick would be far more successful if they could be established away from the cities. And so far as possible, all who are seeking to recover health should place themselves amid country surroundings where they can have the benefit of outdoor life. Nature is a helpful physician.

The pure air, the glad sunshine, the flowers and trees, the orchards and vineyards, and outdoor exercise amid these surroundings, are health-giving, life-giving.

Physicians, nurses, and all health-care workers should encourage their patients to be much in the open air. Outdoor life is the only remedy that many invalids need. ***It has a wonderful power to heal diseases caused by the excitements and excesses of fashionable life, a life that weakens and destroys the powers of body, mind, and soul.***

How grateful to the invalids weary of city life, the glare of many lights, and the noise of the streets, are the quiet and freedom of the country! How eagerly do they turn to the scenes of nature! How glad would they be to sit in the open air, rejoice in the sunshine, and breathe the fragrance of tree and flower! There are life-giving properties in the balsam of the pine, in the fragrance of the cedar and the fir, and other trees also have properties that are health restoring.

To the chronic invalid, nothing so tends to restore health and happiness as living amid attractive country surroundings. Here the most helpless ones can sit or lie in the sunshine or in the shade of the trees. They have only to lift their eyes to see above them the beautiful foliage. A sweet sense of restfulness and refreshing comes over them as they listen to the murmuring of the breezes. The drooping spirits revive. The waning strength is recruited. Unconsciously the mind becomes peaceful, the fevered pulse more calm and regular. As the sick grow stronger, they will venture to take a few steps to gather some of the lovely flowers, precious messengers of love to the afflicted family here below.

Plans should be devised for keeping patients out of doors. For those who are able to work, let some pleasant, easy employment (such as gardening) be provided. Show them how agreeable and helpful this outdoor work is. Encourage them to breathe the fresh air. Teach them to breathe deeply, and in breathing and speaking to exercise the abdominal muscles. This is an education that will be invaluable to them.

Exercise in the Open Air Should be Prescribed as a Life-Giving Necessity. And for such exercises there is nothing better than the cultivation of the soil. Let patients have flower beds to care for, or work to do in the orchard or vegetable garden. As they are encouraged to leave their rooms and spend time in the open air, cultivating flowers or doing some other light, pleasant work, their attention will be diverted from themselves and their sufferings.

The more the patient can be kept out of doors, the less care will he require. The more cheerful his surroundings, the more helpful will he be. Shut up in the house, be it ever so elegantly furnished, he will grow fretful and gloomy. Surround him with the beautiful things of nature; place him where he can

see the flowers growing and hear the birds singing, and his heart will break into song in harmony with the songs of the birds. Relief will come to body and mind. The intellect will be awakened, the imagination quickened, and the mind prepared to appreciate the beauty of God's word.

In nature may always be found something to divert the attention of the sick from themselves and direct their thoughts to Nature. Surrounded by Nature's wonderful works, their minds are uplifted from the things that are seen to the things that are unseen. The beauty of nature leads them to think of the heavenly home, where there will be nothing to mar the loveliness, nothing to taint or destroy, nothing to cause disease or death.

Let physicians and nurses draw from the things of nature, lessons teaching of Nature. Let them point the patients to Nature whose hand has made the lofty trees, the grass, and the flowers, encouraging them to see in every bud and flower an expression of Nature's love for His children. He who cares for the birds and the flowers will care for the beings formed in His own image.

Out of doors, amid the things that Nature has made, breathing the fresh, health-giving air, the sick can best be told of the new life in Nature. Here Nature's message may be read: shining into the human hearts. Men and women in need of physical and spiritual healing are to be thus brought into contact with those whose words and acts will draw them to Nature. They are to hear the story of the Nature's love, of the pardon freely provided for all who come to Him confessing their faults.

Under such influences as these, many suffering ones will be guided into the way of life. Nature co-operate with human instrumentalities in bringing encouragement and hope and joy and peace to the hearts of the sick and suffering. Under such conditions the sick are doubly blessed, and many find health. The feeble step recovers its elasticity. The eye regains its brightness. The hopeless become hopeful. The once despondent countenance wears an expression of joy. The complaining tones of the voice give place to tones of cheerfulness and content.

As physical health is regained, men and women are better able to exercise that faith in Nature which secures the health of the soul. In the consciousness of sins forgiven there is inexpressible peace and joy and rest. The clouded hope of is brightened.

The Human Genetics of Autism^[W]

Autism seems to be a complex *Autism Complex Spectrum* (ACS) that can involve many genes – which may be responsible for managing the connections between synapses in the brain. This complex affects more than 1% of the world's population. In California, USA, it has been estimated to be affecting 1 in 68 lives! People with autism have rather atypical communication and social skills and limited interests – often exhibiting repetitive behavior. It may coexist with other medical and psychiatric conditions, such as epilepsy, sleep disorder, intellectual disability, and gastrointestinal problems. Over 100 risk genes for autism have been found. In some cases, a single mutation may cause the development of autism! In other situations, over 1,000 genetic variations, each with some specific effect at a low level, may increase the risk of autism! Many of these are key regulators of brain connectivity regulating contacts among neurons. Current research on autism focuses on the roles of these special genes affecting brain development.

1.2.4.2 Human Genetics and Genetic Epidemiology

For the “post-genomic” era, where large amounts of genetic data are readily available, it is important to design studies and analytical techniques to accurately detect and describe the role genes play in human disease. Genes alone can cause some human diseases, and the public health impact of genetic diseases may best be addressed by formal and disciplined Human Genetic Epidemiology.

Human Genetics concerns the study of genetic forces in man. By studying ones genetic make-up one is enabled to understand more about ones heritage and change. Some of the original, and most

significant research in genetics centered around the study of the genetics of complex diseases – Human Genetic Epidemiology.

The field of Genetic Epidemiology is focused on designs and analytical techniques to identify how genes contribute to risk for diseases. The academic program in the genetic epidemiology track provides a comprehensive introduction to study designs and statistical approaches used in biostatistics as applied to medical genetics.

Biostatistical Human Genetics and Biostatistical Genetic Epidemiology

For a subject that has seen a recent explosion of interest following the completion of the first draft of the Human Genome Mapping Project. This is understandably a growing field of knowledge. The author have strived to give this book a medical and human genetic feel. To suit the biostatistical nature of Genetic Epidemiology, it seems appropriate to include the use of the popular open-sourced computer program R, originally designed for statistical analysis.

1.2.4.3 Molecular Epidemiology and Genetic Epidemiology

Molecular epidemiology is a branch of *epidemiology* and medical science that focuses on the contribution of potential *genetic* and environmental risk factors, identified at the molecular level, to the etiology, distribution and prevention of disease within families and across populations.

1.2.4.4 Human Molecular Epidemiology

The term “molecular epidemiology” was first coined by Kilbourne in a 1973 article entitled “The molecular epidemiology of influenza”. The term became more formalized with the formulation of the first book on *Molecular Epidemiology: Principles and Practice* by Schulte and Perera. This book discusses advances in molecular research that have given rise to and enable the measurement and exploitation of the **biomarker** as a vital tool to link traditional molecular and epidemiological research strategies to understand the underlying mechanisms of disease in populations.

While most molecular epidemiology studies are using conventional disease designation system for an outcome (with the use of exposures at the molecular level), evidence indicates that disease evolution represents inherently heterogeneous process differing from person to person. Since each case subject has a unique disease process different from any other individual (“the unique disease principle”), considering uniqueness of the exposure and its unique influence on **molecular pathologic** process in each individual. Studies to investigate the relationship between an exposure and molecular pathologic signature of disease (particularly, cancer) became increasingly common throughout the 2000s. The use of molecular pathology in epidemiology posed unique issues including lack of standardized methodologies and guidelines as well as the lack of interdisciplinary experts and training programs. The use of “molecular epidemiology” for this type of research masked the presence of these challenges, and hindered the development of methods and guidelines.

The genome of a bacterial species fundamentally determines its identity. Thus, **gel electrophoresis** techniques like **pulsed-field gel electrophoresis** can be used in molecular epidemiology to comparatively analyze patterns of bacterial chromosomal fragments and to elucidate the genomic content of bacterial cells. Owing to its widespread use and ability to analyze epidemiological information about most bacterial pathogens based on their molecular markers, pulsed-field gel electrophoresis is relied upon heavily in molecular epidemiological studies.

Molecular epidemiology depends on the molecular outcomes and implications of diet, lifestyle, and environmental exposure, particularly how these choices and exposures result in acquired genetic mutations and how these mutations are distributed throughout selected populations through the use of biomarkers and genetic information. Molecular epidemiological studies are able to provide

additional understanding of previously-identified [risk factors](#) and disease mechanisms (Slattery 2002). Specific applications include:

- Molecular surveillance of disease risk factors
- Measuring the geographical and temporal distribution of disease risk factors

Characterizing the evolution of pathogens and classifying new pathogen species (Field 2014)

While the use of advanced molecular analysis techniques within the field of molecular epidemiology is providing the larger field of epidemiology with greater means of analysis, [Porta](#) identified several challenges that the field of molecular epidemiology faces, particularly selecting and incorporating requisite applicable data in an unbiased manner. Limitations of molecular epidemiological studies are similar in nature to those of generic epidemiological studies, that is, samples of convenience – both of the target population and genetic information, small sample sizes, inappropriate statistical methods, poor quality control, and poor definition of target populations.

1.2.5 Human Genetic Epidemiology

Human Genetic Epidemiology is the study of the role of human and medical [genetic](#) factors in determining health and disease in families and in populations, and the interplay of such genetic factors with environmental factors. Human genetic epidemiology seeks to derive a statistical and quantitative analysis of how genetics work in large groups.

The use of the term *Genetic epidemiology* emerged in the mid-1980s as a new scientific field. In formal language, genetic epidemiology was defined by [Newton Morton](#), one of the pioneers of the field, as “a science which deals with the [etiology](#), distribution, and control of disease in groups of relatives and with inherited causes of disease in populations” (What is Molecular Epidemiology). It is closely allied to both [molecular epidemiology](#) and [statistical genetics](#), but these overlapping fields each have distinct emphases, societies and journals (What is Molecular Epidemiology).

One definition of the field closely follows that of [behavior genetics](#), defining genetic epidemiology as “the scientific discipline that deals with the analysis of the familial distribution of traits, with a view to understanding any possible genetic basis”, and that seeks to understand both the genetic and environmental factors and how they interact to produce various diseases and traits in humans. The British Medical Journal adopted a similar definition, Genetic epidemiology is the study of the aetiology, distribution, and control of disease in groups of relatives and of inherited causes of disease in populations.

As early as the 4th century BC, [Hippocrates](#) suggested in his essay “On Airs, Waters, and Places” that factors such as behavior and environment may play a role in disease. Epidemiology entered a more systematic phase with the work of [John Graunt](#), who in 1662 tried to quantify mortality in London using a statistical approach, tabulating various factors he thought played a role in mortality rates. [John Snow](#) is considered to be the father of epidemiology, and was the first to use statistics to discover and target the cause of disease, specifically of cholera outbreaks in 1854 in London. He investigated the cases of cholera and plotted them onto a map identifying the most likely cause of cholera, which was shown to be contaminated water wells.

Modern genetics began on the foundation of [Mendel's](#) work. Once this became widely known, it spurred a revolution in studies of heredity throughout the animal kingdom; with studies showing genetic transmission and control over characteristics and traits. As gene variation was shown to affect disease, work began on quantifying factors affecting disease, accelerating in the twentieth century. The period since the Second World War (1939–1945) saw the greatest advancement of the field, with

scientists such as **Newton Morton** helping form the field of genetic epidemiology as it is known today, with the application of modern genetics to the statistical study of disease, as well as the establishment of large-scale epidemiological studies such as the **Framingham Heart Study**.

In the 1960s and 1970s, epidemiology played a part in strategies for the worldwide eradication of naturally occurring smallpox. Traditionally, the study of genetics in disease progresses through the following study designs, each answering a different question: ([Ogino et al. 2013](#))

- **Familial Aggregation studies:** Is there a genetic component to the disease, and what are the relative contributions of genes and environment?
- **Segregation studies:** What is the **pattern of inheritance** of the disease (recessive or dominant)?
- **Linkage studies:** On which part of which **chromosome** is the disease gene located?
- **Association studies:** Which **allele** of which gene is associated with the disease?

This traditional approach has proved highly successful in identifying **monogenic disorders** and locating the genes responsible.

Nowadays, the scope of genetic epidemiology has expanded to include common diseases for which many genes each make a smaller contribution (**polygenic, multifactorial or multi-genetic disorders**). This has developed rapidly in the first decade of the twenty-first century (2001–2010) following completion of the **Human Genome Project**, as advances in **genotyping** technology and associated reductions in cost has made it feasible to conduct large-scale **genome-wide association studies** that genotype many thousands of **single nucleotide polymorphisms** in thousands of individuals. **These have led to the discovery of many genetic polymorphisms that influence the risk of developing many common diseases.**

Modern Approaches in Human Epidemiological Research

Genetic epidemiological research follows 3 discreet steps, as outlined by M.Tevfik Dorak:

1. Establishing that there is a genetic component to the disorder.
2. Establishing the relative size of that genetic effect in relation to other sources of variation in disease risk (environmental effects such as intrauterine environment, physical and chemical effects as well as behavioral and social aspects).
3. Identifying the gene(s) responsible for the genetic component.

These research methodologies can be assessed through either family or population studies.

Is Race real? Dismissal of the “Race Card” in scientific Human Genetic Epidemiology

An interesting question that had arisen in the study of human genetic epidemiology, and it is as follows:

In 1985, nearly two decades before the human genome was decoded, a survey among 1200 scientists who were asked how many would **disagree** with the following proposition:

“There are biological races in the species *Homo Sapiens*.” ?

The responses were:

- **biologists:** 16% disagreed
- **developmental psychologists:** 36% disagreed
- **physical anthropologists:** 41% disagreed
- **cultural anthropologists:** 53% disagreed

This article does not intend to address how many people believed in biological “races” in 1985. It is at best an irrelevant semantic point.

From the viewpoint of scientific human or medical genetic epidemiology, the scientific approach is that the subject is race-neutral!

This is the position taken in this book!

1.2.6 Applied Statistical Human Genetics

The remainder of this book will showcase the use of statistical methods, including using the computer programming language R, to solve critical practical problems in human genetic epidemiology, with particular emphasis on the following four areas:

- **Familial Aggregation studies:** Is there a genetic component to the disease, and what are the relative contributions of genes and environment?
- **Segregation studies:** What is the **pattern of inheritance** of the disease (recessive or dominant)?
- **Linkage studies:** On which part of which **chromosome** is the disease gene located?
- **Association studies:** Which **allele** of which gene is associated with the disease?

In the next chapter, the use of the computer programming language R will be described. Originally, R was written mainly for solving problems in applied statistics and biostatistics.

References

- Field N (2014) Strengthening the REPORTING OF MOLECULAR Epidemiology for Infectious Diseases (STROME-ID): an extension of the STROBE statement. Lancet Infect Dis 14(4):341–352. [https://doi.org/10.1016/S1473-3099\(13\)70324-4](https://doi.org/10.1016/S1473-3099(13)70324-4). PMID 24631223
- Kilbourne ED (1973) The molecular epidemiology of influenza. J Infect Dis 127(4):478–87. <https://doi.org/10.1093/infdis/127.4.478>. PMID 4121053
- Ogino S, Lochhead P, Chan AT, Nishihara R, Cho E, Wolpin BM, Meyerhardt AJ, Meissner A, Schernhammer ES, Fuchs CS, Giovannucci E (2013) Molecular pathological epidemiology of epigenetics: emerging integrative science to analyze environment, host, and disease. Mod Pathol 26:465–484
- Slattery M (2002) The science and art of molecular epidemiology. J Epidemiol Commun Health 56(10):728–729. <https://doi.org/10.1136/jech.56.10.728>. PMID 1732025
- “What is molecular epidemiology?”. Molecular Epidemiology Homepage. University of Pittsburgh. 28 July 1998. Retrieved 15 January 2010
- “What is molecular epidemiology?”. aacr.org. Retrieved 2008-02-19

Special Reference

- Tevfik Dorak M (2008-03-03) Introduction to genetic epidemiology



Data Analysis Using R Programming

2

Abstract

Beginning R

R is an open-source, freely available, integrated software environment for data manipulation, computation, analysis, and graphical display. The R environment consists of

- *a data handling and storage facility,
- *operators for computations on arrays and matrices,
- *a collection of tools for data analysis
- *graphical capabilities for analysis and display, and
- *an efficient, and continuing developing programming algebra-like programming language which consists of loops, conditionals, user-defined functions, and input and output capabilities.

Many R programs are available for biostatistical analysis in Genetic Epidemiology. Typical examples are shown.

Keywords

R environment · R as a calculator · R graphics · R in statistics · R in data analysis in human genetic epidemiology · Function `data.entry()` · Function `source()` · Spreadsheet interface in R · `plot()` function

In an Internet on-line advertisement, a job vacancy advertisement for a Statistician. The complete job description reads as follows:

Job Summary

Statistician I

Salary: Open

Employer: XYZ Research and Statistics

Location: City X, State Y

Type: Full Time – Entry Level

Category: Financial analyst/Statistics,

Data analysis/processing, Statistical organization & administration

Required Education: Masters Degree preferred

XYZ Research and Statistics is a national leader in designing, managing, and analyzing financial data. XYZ partners with other investigators to offer respected statistical expertise supported by sophisticated web-based data management systems. XYZ services assure timely and secure implementation of trials and reliable data analyses.

Job Description

Position Summary: An exciting opportunity is available for a statistician to join a small but growing group focused on financial investment analysis and related translational research. XYZ, which is located in downtown City XX, is responsible for the design, management and analysis of a variety of investment and financial, as well as the analysis of associated market data. The successful candidate will collaborate with fellow statistics staff and financial investigators to design, evaluate, and interpret investment studies.

Primary Duties and Responsibilities: Analyzes investment situations and associated ancillary studies in collaboration with fellow statisticians and other financial engineers. Prepares tables, figures, and written summaries of study results; interprets results in collaboration with other financial; and assists in preparation of manuscripts. Provides statistical consultation with collaborating staff. Performs other job-related duties as assigned.

Requirements

Required Qualifications: Masters Degree in Statistics, Applied Mathematics, or related field. Sound knowledge of applied statistics. Proficiency in statistical computing in R.

Preferred Responsibilities/Qualifications: Statistical consulting experience.

S-Plus or R programming language experience. Experience with analysis of high-dimensional data. Ability to communicate well orally and in writing. Excellent interpersonal/teamwork skills for effective collaboration. Spanish language skills a plus.

*In your cover letter, describe how your skills and experience match the qualifications for the position.

To learn more about XYZ, visit www.XYZ.org.

Clearly, one should be cognizant of the overt requirement of an acceptable level of professional proficiency in data analysis using R programming!

Even if one is not in such a job market, as a statistician working in the fields of Finance, Asset Allocations, Portfolio Optimization, etc., a skill set that would include R programming would be helpful and interesting.

2.1 Data and Data Processing

Data are facts or figures from which conclusions can be drawn. When the data have been recorded, classified, and organized, related or interpreted within a framework so that meaning emerges, they become **information**. There are several steps involved in turning data into information, and these steps are known as **data processing**. This section describes data processing and how computers perform these steps efficiently and effectively.

It will be indicated that many of these processing activities may be undertaken using R programming, or performed in an R environment with the aid of available R packages – where R functions and datasets are stored..

Introduction (Statistics Canada)CodingAutomated coding systems

The simplified flowchart below, shows how raw **data** are transformed into information:

Data → Collection → Processing → Information

Data processing takes place once all of the relevant data have been collected. They are gathered from various sources and entered into a computer where they can be processed to produce **information** – the output.

Data processing includes the following steps:

Data [Coding](#)

Data [Capture](#)

[Editing](#)

[Imputation](#)

[Quality control](#)

Producing results

Data Coding

First, before raw data can be entered into a computer, they must be coded. To do this, survey responses must be labeled, usually with simple, numerical codes. This may be done by the interviewer in the field or by an office employee. The data coding step is important because it makes data entry and data processing easier.

Surveys have two types of questions—closed questions and open questions. The responses to these questions affect the type of coding performed.

A **closed question** means that only a **fixed** number of predetermined survey responses are permitted. These responses will have already been coded.

The following question, in a survey on Sporting activities, is an example of a closed question:

To what degree is sport important in providing you with the following benefits?

<1> Very important

<2> Somewhat important

<3> Not important

An **open question** implies that **any** response is allowed, making subsequent coding more difficult. In order to code an open question, the processor must sample a number of responses, and then design a code structure that includes all possible answers.

The following code structure is an example of an open question:

What sports do you participate in?

Specify (28 characters)_____

In the [Census](#) and almost all other surveys, the codes for each question field are pre-marked on the **questionnaire**. To process the questionnaire, the codes are entered directly into the database and are prepared for data capturing. The following is an example of pre-marked coding:

What language does this person speak most often at home?

- <18> English
- <19> French
- <20> Other—Specify _____

Automated Coding Systems

There are programs in use that will automate repetitive and routine tasks. Some of the advantages of an automated coding system are that the process increasingly becomes

faster,
more consistent, and
more economical.

The next step in data processing is inputting the coded data into a computer database. This method is known as **data capture**.

Data Capture

This is the process by which data are transferred from a paper copy, such as [questionnaires](#) and survey responses, to an electronic file. The responses are then put into a computer. Before this procedure takes place, the questionnaires must be groomed (prepared) for data capture. In this processing step, the questionnaire is reviewed to ensure that all of the minimum required data have been reported, and that they are decipherable. This grooming is usually performed during extensive automated edits.

There are several methods used for capturing data:

Tally charts are used to record data such as the number of occurrences of a particular event and to develop frequency distribution tables.

Batch keying is one of the oldest methods of data capture. It uses a computer keyboard to type in the data. This process is very practical for high-volume entry where fast production is a requirement. No editing procedures are necessary but there must be a high degree of confidence in the editing program.

Interactive capture is often referred to as intelligent keying. Usually, captured data are edited before they are imputed. However, this method combines data capture and data editing in one function.

Optical character readers or bar-code scanners, are able to recognize alpha or numeric characters. These readers scan lines and translate them into the program. These bar-code scanners are quite common and often seen in department stores. They can take the shape of a gun or a wand.

Magnetic recordings allow for both reading and writing capabilities. This method may be used in areas where data security is important. The largest application for this type of data capture is the PIN number found on automatic bank cards. A computer keyboard is one of the best known input (or data entry) devices in current use. In the past, people performed data entry using punch cards or paper tape.

Some modern examples of data input devices are

- optical mark reader
- bar-code reader
- scanner used in desktop publishing
- light pen
- trackball
- mouse

Once data have been entered into a computer database, the next step is ensuring that all of the responses are accurate. This method is known as data editing.

Data Editing

Data should be **edited** before being presented as information. This action ensures that the information provided is accurate, complete and consistent.

There are two levels of data editing—**micro-** and **macro-editing**.

Micro-editing corrects the data at the record level. This process detects errors in data through checks of the individual data records. The intent at this point is to determine the consistency of the data and correct the individual data records.

Macro-editing also detects errors in data, but does this through the analysis of aggregate data (totals). The data are compared with data from other **surveys**, **administrative files**, or earlier versions of the same data. This process determines the compatibility of data.

Imputations

Editing is of little value to the overall improvement of the actual survey results, if no corrective action is taken when items fail to follow the rules set out during the editing process. When all of the data have been edited using the applied rules and a file is found to have missing data, then imputation is usually done as a separate step.

Non-response and invalid data definitely impact the quality of the **survey** results.

Imputation resolves the problems of missing, invalid, or incomplete responses identified during editing, as well as any editing errors that might have occurred.

At this stage, all of the data are screened for errors because respondents are not the only ones capable of making mistakes; errors can also occur during coding and editing.

Some other types of imputation methods include:

hot deck uses other records as 'donors' in order to answer the question (or set of questions) that needs imputation.

substitution relies on the availability of comparable data. Imputed data can be extracted from the respondent's record from a previous cycle of the survey, or the imputed data can be taken from the respondent's alternative source file (e.g. **administrative** files or other survey files for the same respondent).

estimator uses information from other questions or from other answers (from the current cycle or a previous cycle), and through mathematical operations, derives a plausible value for the missing or incorrect field.

cold deck makes use of a fixed set of values, which covers all of the data items. These values can be constructed with the use of historical data, subject-matter expertise, etc.

The donor can also be found through a method called **nearest neighbor imputation**. In this case, some sort of criteria must be developed to determine which responding unit is 'most like' the unit with the missing value in accordance with the predetermined characteristics. The closest unit to the missing value is then used as the donor.

Imputation methods can be performed automatically, manually, or in combination.

Data Quality

- [Quality assurance](#)
- [Quality control](#)
- [Quality management in statistical agencies](#)

Quality is an essential element at all levels of processing. To ensure the quality of a product or service in survey development activities, both quality assurance and quality control methods are used.

Quality Assurance

Quality assurance refers to all planned activities necessary in providing confidence that a product or service will satisfy its purpose and the users' needs. In the context of survey conducting activities, this can take place at any of the major stages of survey development: planning, design, implementation, processing, evaluation and dissemination.

This approach anticipates problems prior to their unexpected occurrences, and uses all available information to generate improvements. It is not restricted to any specific quality the planning stage and is all-encompassing in its activities standards. It is applicable mostly at the planning stage, and is all-encompassing in its activities.

Quality Control

Quality control is a regulatory procedure through which one may measure quality, with pre-set standards, and then act on any differences. Examples of this include controlling the quality of the coding operation, the quality of the survey interviewing, and the quality of the data capture.

Quality control responds to observed problems, using on-going measurements to make decisions on the processes or products. It requires a pre-specified quality for comparability. It is applicable mostly at the processing stage, following a set procedure that is a subset of quality assurance.

Quality Management in Statistical Agencies

The quality of the data must be defined and assured in the context of being "fit for use", which will depend on the intended function of the data and the fundamental characteristics of quality. It also depends on the users' expectations of what is considered to be useful information.

There is no standard definition among statistical agencies for the term "official Statistics". There is a generally accepted, but evolving, range of quality issues underlying the concept of 'fitness for use'. These elements of quality need to be considered and balanced in the design and implementation of an agency's statistical program.

The following is a list of the elements of quality:

Relevance

Accuracy

Timeliness

Accessibility

Interpretability

Coherence

These elements of quality tend to overlap. Just as there is no single measure of accuracy, there is no effective statistical model for bringing together all these characteristics of quality into a single

indicator. Also, except in simple or one-dimensional cases, there is no general statistical model for determining whether one particular set of quality characteristics provides higher overall quality than another.

Producing Results

After [editing](#), data may be processed further to produce a desired output. The computer [software](#) used to process the data will depend on the form of output required. Software applications for word processing, desktop publishing, graphics (including graphing and drawing), [programming](#), [databases](#) and spreadsheets are commonly used. The following are some examples of ways that software can produce data:

Spreadsheets are programs that automatically add columns and rows of figures, calculate means, and perform statistical analyses.

Databases are electronic filing cabinets. They systematically store data for easy access, and produce summaries, aggregates or reports.

Specialized programs can be developed to edit, clean, impute and process the final tabular output.

Review Questions for Sect. 2.1

1. In the Job Description for an entry level Statistician to-day, from the viewpoint of a prospective applicant for that position, what basic statistical computing languages are important in order to meet the requirement? Why?
2. For a typical MBA (Master of Business Administration) program in Business and Finance, should the core curriculum include the development of proficient skill in the use of R programming in Statistics? Why?
3. (a) Contrast the concepts of Data and Information.
(b) How does the process of Data Processing convert Data to Information?
4. In the steps which convert Data into Information, how are statistics and computing applied to the various Data Processing steps.
5. (a) Describe and delineate Quality Assurance and Quality Control in computer Data Processing.
(b) In what way does statistics feature in these phases of Data Processing?

2.2 Beginning R

R is an open-source, freely available, integrated software environment for data manipulation, computation, analysis, and graphical display.

The R environment consists of

- *a data handling and storage facility,
- *operators for computations on arrays and matrices,
- *a collection of tools for data analysis
- *graphical capabilities for analysis and display, and
- *an efficient, and continuing developing programming algebra-like programming language which consists of loops, conditionals, user-defined functions, and input and output capabilities.

The term “environment” is used to show that it is indeed a planned and coherent system. (Venables and Smith 2004; Aragon 2011)

R and Statistics

R was initially written by Robert Gentleman and Ross Ihaka of the Statistics

Department of the University of Auckland, New Zealand, in 1997. Since then there has been the R-development core group of about 20 people with write-access to the R source code.

The original introduction to the R environment, evolved from the S/S-Plus languages, was **not** primarily directed towards statistics. However, since its development in the 1990s, it appeared to have been “hijacked” by many working in the areas of classical and modern statistical techniques, including many applications in financial engineering, econometrics, biostatistics with respect to epidemiology, public health and preventive medicine! These applications have led to the *raison d'etat* for writing this book.

As of this writing, the latest version of R is R-3.3.2, officially released on October 31, 2016. The primary source of R packages is the Comprehensive R Archive Network, CRAN, at <http://cran.r-project.org/>

Another source of R packages may be found in numerous publications, e.g., the *Journal of Statistical Software*, now at its 45th volume, is available at <http://www.jstatsoft.org/v45> .

Let us get started – (the R-3.3.2 version environment is being used here)

Recall in Sect. 2.1, the R environment was obtained as follows:

Here is R:

Let us download the open-source high-level program R from the Internet and take a first look at the R computing environment.

Remark: Access the Internet at the website of CRAN (The Comprehensive

R Archive Network: <http://cran.r-project.org/>

To install R: R-3.3.2-win32.exe

<http://www.r-project.org/>

=> download R

=> Select: USA

[<http://cran.cnr.Berkeley.edu>](http://cran.cnr.Berkeley.edu)

University of California, Berkeley, CA

=> <http://cran.cnr.berkeley.edu/>

=> Windows (95 and later)

=> base

=> R-3.3.2-win32.exe

AFTER the down-loading:

=> Double-click on: R-3.3.2-win32.exe

(on the DeskTop) to un-zip & install R

An icon (Script R 3.3.2) will appear on ones Computer “desktop” as follows: Fig. 2.1

On the computer “desktop” is the R icon:

In this book, the following special color scheme legend will be used for all statements during the computational activities in the R environment, to clarify the various inputs to and outputs from the computational process:

1. Texts in this book (Times New Roman font)
2. Line Input in R code (Verdana font)
3. Line output in R code (Verdana font)
4. *Line Comment Statements in R code (Italicized Times New Roman font)*

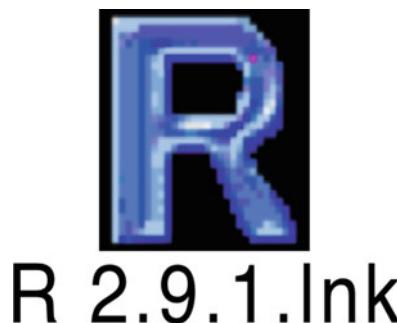


Fig. 2.1 The R icon on the computer desktop (The R 3.3.2 and the R 3.3.3 icon looks *exactly* the same as that for R 2.9.1)

Note The **#** sign is the Comment Character: all text in the line following this sign is treated as a comment by the R program, i.e., no computational action will be taken regarding such a statement. That is, the computational activities will proceed as though the comment statements are ignored. These comment statements help the programmer and user by providing some clarification of the purposes involved in the remainder of the R environment. The computations will proceed even if these comment statements are eliminated.

is known as the Number Sign, it is also known as the pound sign/key, the hash key, and, less commonly, as the octothorp, octothorpe, octathorp, octotherp, octathorpe, and octatherp!

To use R under Windows: Double-click on the R 3.3.2 icon

Upon selecting and clicking on R, the R-window opens, with the following declaration:

R version 3.3.2 (2016-10-31)
 Copyright (C) 2016 The R Foundation for Statistical Computing
 ISBN 3-900051-07-0
 R is free software and comes with ABSOLUTELY NO WARRANTY.
 You are welcome to redistribute it under certain conditions.
 Type 'license()' or 'licence()' for distribution details.
 R is a collaborative project with many contributors.
 Type 'contributors()' for more information and
 'citation()' on how to cite R or R packages in publications.
 Type 'demo()' for some demos, 'help()' for on-line help, or
 'help.start()' for an HTML browser interface to help.
 Type 'q()' to quit R.

```
> # This is the R computing environment.
> # Computations may begin now!
>
> # First, use R as a calculator, and try a simple arithmetic
> # operation, say: 1 + 1
> 1+1
> [1] 2  # This is the output!
>          # WOW! It's really working!
> # The [1] in front of the output result is part of R's way of printing
> # numbers and vectors. Although it is not so useful here, it does
> # become so when the output result is a longer vector
```

***** From this point on, this book is most beneficially read with the R environment at hand. It will be a most effective learning experience if one practises each R command as one goes along the textual materials!**

2.2.1 A First Session Using R

This section introduces some important and practical features of the R Environment (Fig. 2.2).

Login and start an R session in the Windows system of the computer (Fig. 2.3):

Statistical Data Analysis

Manuals

[An Introduction to R](#)

[The R Language Definition](#)

[Writing R Extensions](#)

[R Installation and Administration](#)

[R Data Import/Export](#)

[R Internals](#)

Reference

[Packages](#)

[Search Engine & Keywords](#)

Miscellaneous Material

[About R](#)

[Authors](#)

[Resources](#)

[License](#)

[Frequently Asked Questions](#)

[Thanks](#)

[NEWS](#)

[User Manuals](#)

[Technical papers](#)

Material specific to the Windows port

[CHANGES](#)

[Windows FAQ](#)

Fig. 2.2 Output of the R command

Packages in C:\Program Files\R\R-

2.14.1\library

<u>base</u>	<i>The R Base Package</i>
<u>boot</u>	<i>Bootstrap Functions (originally by Angelo Canty for S)</i>
<u>class</u>	<i>Functions for Classification</i>
<u>cluster</u>	<i>Cluster Analysis Extended Rousseeuw et al.</i>
<u>codetools</u>	<i>Code Analysis Tools for R</i>
<u>compiler</u>	<i>The R Compiler Package</i>
<u>datasets</u>	<i>The R Datasets Package</i>
<u>foreign</u>	<i>Read Data Stored by Minitab, S, SAS, SPSS, Stata, Systat, dBase, ...</i>
<u>graphics</u>	<i>The R Graphics Package</i>
<u>grDevices</u>	<i>The R Graphics Devices and Support for Colours and Fonts</i>
<u>grid</u>	<i>The Grid Graphics Package</i>
<u>KernSmooth</u>	<i>Functions for kernel smoothing for Wand & Jones (1995)</i>
<u>lattice</u>	<i>Lattice Graphics</i>
<u>MASS</u>	<i>Support Functions and Datasets for Venables and Ripley's MASS</i>
<u>Matrix</u>	<i>Sparse and Dense Matrix Classes and Methods</i>
<u>methods</u>	<i>Formal Methods and Classes</i>

Fig. 2.3 Package index

<u>mgcv</u>	<i>GAMs with GCV/AIC/REML smoothness estimation and GAMMs by PQL</i>
<u>nlme</u>	<i>Linear and Nonlinear Mixed Effects Models</i>
<u>nnet</u>	<i>Feed-forward Neural Networks and Multinomial Log-Linear Models</i>
<u>parallel</u>	<i>Support for Parallel computation in R</i>
<u>rpart</u>	<i>Recursive Partitioning</i>
<u>spatial</u>	<i>Functions for Kriging and Point Pattern Analysis</i>
<u>splines</u>	<i>Regression Spline Functions and Classes</i>
<u>stats</u>	<i>The R Stats Package</i>
<u>stats4</u>	<i>Statistical Functions using S4 Classes</i>
<u>survival</u>	<i>Survival analysis, including penalised likelihood.</i>
<u>tcltk</u>	<i>Tcl/Tk Interface</i>
<u>tools</u>	<i>Tools for Package Development</i>
<u>utils</u>	<i>The R Utils Package</i>

Fig. 2.3 (continued)

```

>
> # This is the R environment.
> help.start()  # Outputting the page shown in Fig. 2.1
>                 # Statistical Data Analysis Manuals[31]
starting httpd help server ... done
If nothing happens, you should open
'http://127.0.0.1:28103/doc/html/index.html' yourself
At this point, explore the HTML interface to on-line help
right from the desktop, using the mouse pointer to note
the various features of this facility available within
the R environment. Then, returning to the R environment:

```

```
> help.start()
```

Carefully read through each of the sections under "Manuals" – to obtain an introduction to the basic language of the R environment. Then look through the items under "Reference" to reach beyond the elementary level, including access to the available "R Packages" – all R functions and datasets are stored in packages. For example, if one selects the Packages Reference, the following R Package Index window will open up, showing **Figure 2.3**, listing a collection of R program packages under the R library: C:\Program Files\R\R-2.14.1\library

One may now access each of these R program packages, and use them for further applications as needed.

Returning to the R environment (Fig. 2.4):

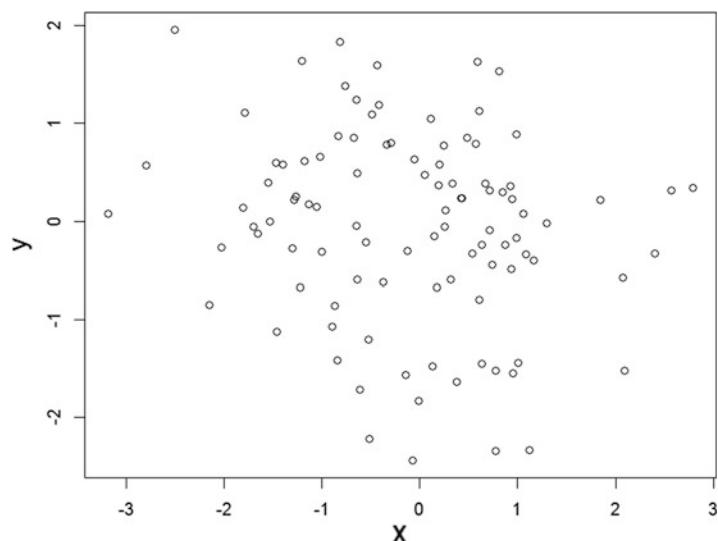
```
>
> x <- rnorm(100)
> # Generating a pseudo-random 100-vector x
> y <- rnorm(x)
> # Generating another pseudo-random 100-vector y
> plot (x, y)
> # Plotting x vs. y in the plane, resulting in a graphic
> # window: Fig. 2.4.
```

Remark For reference, Appendix 1 contains the CRAN documentation of the R function `plot()`, available for graphic outputting, which may be found by the R code segment:

```
> ?plot
```

CRAN has documentations for many R functions and packages.

Fig. 2.4 Graphical output for `plot(x, y)`



Again, returning to the R workspace, and enter (Figs. 2.5, 2.6, 2.7, 2.8, 2.9, 2.10 and 2.11):

```
>
>
> ls()  # (This is a lower-case "L" followed by "s", viz., the 'list'
>       # command.)
>       # (NOT 1 = "ONE" followed by "s")
>       # This command will list all the R objects now in the
>       # R workspace:
>       # Outputting:
[1] "E"  "n"  "s"  "x"  "y"  "z"
```

Again, returning to the R workspace, and enter:

```
>
> rm (x, y) # Removing all x and all y from the R workspace
> x          # Calling for x
Error: object 'x' not found
>          # Of course, the xs have just been removed!
> y          # Calling for y
Error: object 'y' not found # Because the ys have also been
                           # removed!
>
> x <- 1:10   # Let x = (1, 2, 3, 4, 5, 6, 7, 8, 9, 10)
> x          # Outputting x (just checking!)
[1] 1 2 3 4 5 6 7 8 9 10
> w <- 1 + sqrt(x)/2    # w is a weighting vector of
>                      # standard deviations
> dummy <- data.frame (x = x, y = x + rnorm(x)*w)
> # Making a data frame of 2 columns, x, and y, for inspection
> dummy    # Outputting the data frame dummy

      x      y
1    1 1.311612
2    2 4.392003
3    3 3.669256
4    4 3.345255
5    5 7.371759
6    6 -0.190287
7    7 10.835873
8    8 4.936543
9    9 7.901261
10   10 10.712029
>
> fm <- lm(y~x, data=dummy)
> # Doing a simple Linear Regression
> summary(fm) # Fitting a simple linear regression of y on x,
>               # then inspect the analysis, and outputting:
Call:
lm(formula = y ~ x, data = dummy)
```

Residuals:

Min	1Q	Median	3Q	Max
-6.0140	-0.8133	-0.0385	1.7291	4.2218

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	1.0814	2.0604	0.525	0.6139
x	0.7904	0.3321	2.380	0.0445 *

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.016 on 8 degrees of freedom

Multiple R-squared: 0.4146, Adjusted R-squared: 0.3414

F-statistic: 5.665 on 1 and 8 DF, p-value: 0.04453

```
> fm1 <- lm(y~x, data=dummy, weight=1/w^2)
> summary(fm1) # Knowing the standard deviation, then doing a
>                  # weighted regression and outputting:
Call:
lm(formula = y ~ x, data = dummy, weights = 1/w^2)
```

Residuals:

Min	1Q	Median	3Q	Max
-2.69867	-0.46190	-0.00072	0.90031	1.83202

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	1.2130	1.6294	0.744	0.4779
x	0.7668	0.3043	2.520	0.0358 *

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.356 on 8 degrees of freedom

Multiple R-squared: 0.4424, Adjusted R-squared: 0.3728

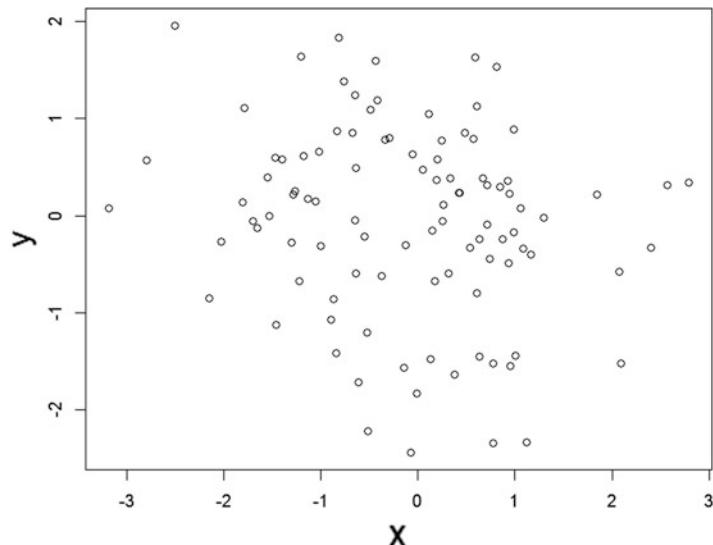
F-statistic: 6.348 on 1 and 8 DF, p-value: 0.03583

```
> attach(dummy) # Making the columns in the data
>                  # frame as variables
```

The following object(s) are masked by '.GlobalEnv': x

```
> lrf <- lowess(x, y) # a non-parametric local
>                  # regression function lrf
> plot (x, y) # Making a standard point plot, outputting: Fig. 2.5.
```

Fig. 2.5 Graphical output for plot (x, y)



Remark For reference, Appendix 1 contains the CRAN documentation of the R function `plot()`, available for graphic outputting, which may be found by the R code segment:

```
> ?plot

> # CRAN has documentations for many R functions and packages.
```

Again, returning to the R workspace, and enter:

```
>

> ls() # (This is a lower-case "L" followed by "s", viz., the 'list'
>       # command.)
>       # (NOT 1 = "ONE" followed by "s")
>       # This command will list all the R objects now in the
>       # R workspace:
>       # Outputting:

[1] "E" "n" "s" "x" "y" "z"
```

Again, returning to the R workspace, and enter:

```
>

> rm (x, y) # Removing all x and all y from the R workspace
> x          # Calling for x
Error: object 'x' not found
> # Of course, the xs have just been removed!
>

> y          # Calling for y
Error: object 'y' not found
> # Because the ys have been removed too!
>

> x <- 1:10 # Let x = (1, 2, 3, 4, 5, 6, 7, 8, 9, 10)
> x          # Outputting x (just checking!)
```

```

[1] 1 2 3 4 5 6 7 8 9 10
> w <- 1 + sqrt(x) / 2      # w is a weighting vector of
>                               # standard deviations
> dummy <- data.frame (x = x, y = x + rnorm(x)*w)
> # Making a data frame of 2 columns, x, and y, for inspection

> dummy      # Outputting the data frame dummy

      x      y
1 1 1.311612
2 2 4.392003
3 3 3.669256
4 4 3.345255
5 5 7.371759
6 6 -0.190287
7 7 10.835873
8 8 4.936543
9 9 7.901261
10 10 10.712029

> fm <- lm(y~x, data=dummy)
> # Doing a simple Linear Regression
> summary(fm)
> # Fitting a simple linear regression of y on x,
> # then inspect the analysis, and outputting:

Call:
lm(formula = y ~ x, data = dummy)

Residuals:
    Min      1Q      Median      3Q      Max
-6.0140 -0.8133 -0.0385  1.7291  4.2218

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 1.0814     2.0604    0.525  0.6139    
x           0.7904     0.3321    2.380  0.0445 *  
---
Signif. codes:  0 '****' 0.001 '***' 0.01 '**' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.016 on 8 degrees of freedom
Multiple R-squared: 0.4146,    Adjusted R-squared: 0.3414 
F-statistic: 5.665 on 1 and 8 DF,  p-value: 0.04453

> fm1 <- lm(y~x, data=dummy, weight=1/w^2)
> summary(fm1) # Knowing the standard deviation,
>                  # then doing a weighted
>                  # regression and outputting:
Call:
lm(formula = y ~ x, data = dummy, weights = 1/w^2)

```

Residuals:

Min	1Q	Median	3Q	Max
-2.69867	-0.46190	-0.00072	0.90031	1.83202

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	1.2130	1.6294	0.744	0.4779
x	0.7668	0.3043	2.520	0.0358 *

Signif. codes: 0 '****' 0.001 '***' 0.01 '**' 0.05 '*' 0.1 '.' 1

Residual standard error: 1.356 on 8 degrees of freedom

Multiple R-squared: 0.4424, Adjusted R-squared: 0.3728

F-statistic: 6.348 on 1 and 8 DF, p-value: 0.03583

```

> attach(dummy)  # Making the columns in the data frame as
>                  # variables
> lrf <- lowess(x, y)
> lrf
> plot (x, y) # Making a standard point plot,
>                  # outputting: Fig. 2.6

> abline(0, 1, lty=3) # adding in the true regression line:
>                  # (Intercept = 0, Slope = 1),
>                  # outputting: Fig. 2.7.

```

Fig. 2.6 Adding in the local regression line

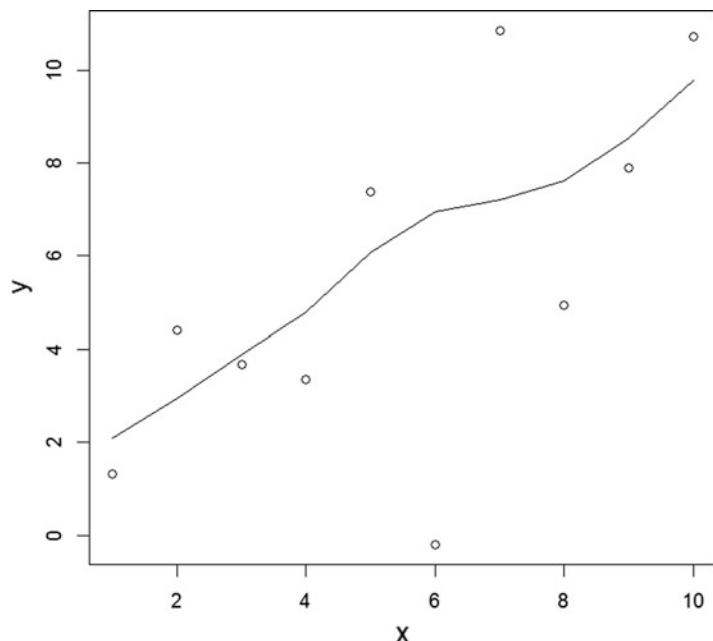
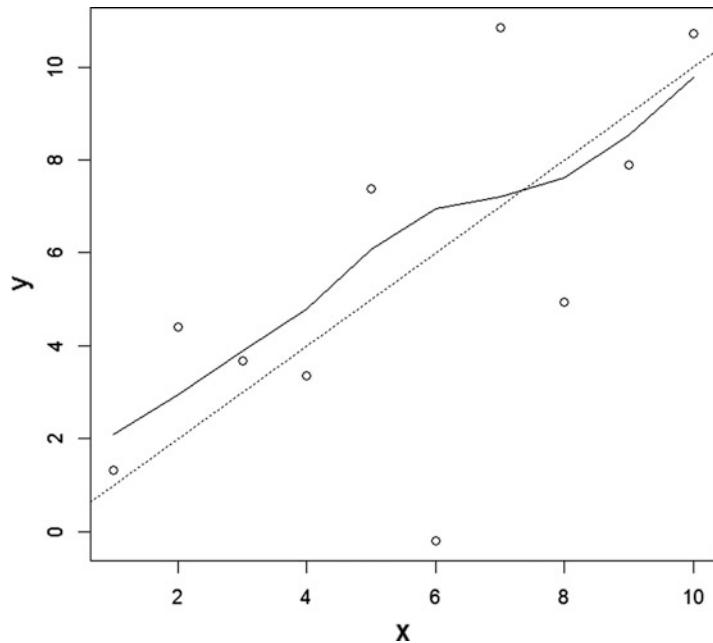


Fig. 2.7 Adding in the true regression line (Intercept = 0, Slope = 1)



```

> abline(coef(fm)) # adding in the unweighted regression line:
> # outputting Fig. 2.8

> abline(coef(fm1), col="red")
> # adding in the weighted regression line:
> # outputting Fig. 2.9.

> detach() # Removing data frame from the search
> # path
> plot(fitted(fm), resid(fm)),
> # Doing a standard diagnostic plot
+ xlab="Fitted values", # to check for
+ # heteroscedasticity**,
+ ylab="residuals", # viz., checking for differing
+ # variance.
+ main="Residuals vs Fitted")
# Outputting Fig. 2.10.
**Heteroskedasticity occurs when the variance of the error terms differ across
observations.

> qqnorm(resid(fm), main="Residuals Rankit Plot")
> # Doing a normal scores plot to check for
> # skewness, kurtosis, and outliers.
> # (Not very useful here.) Outputting Fig. 2.11.

```

Fig. 2.8 Adding in the unweighted regression line

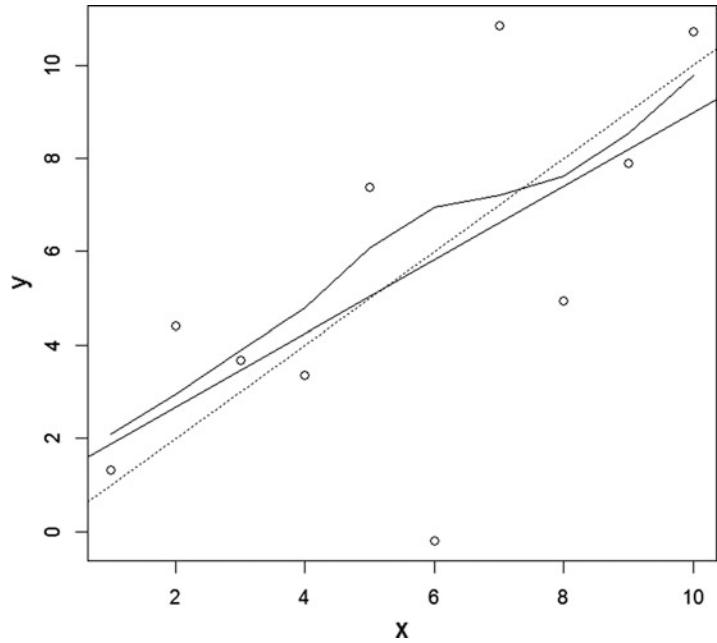


Fig. 2.9 Adding in the weighted regression line

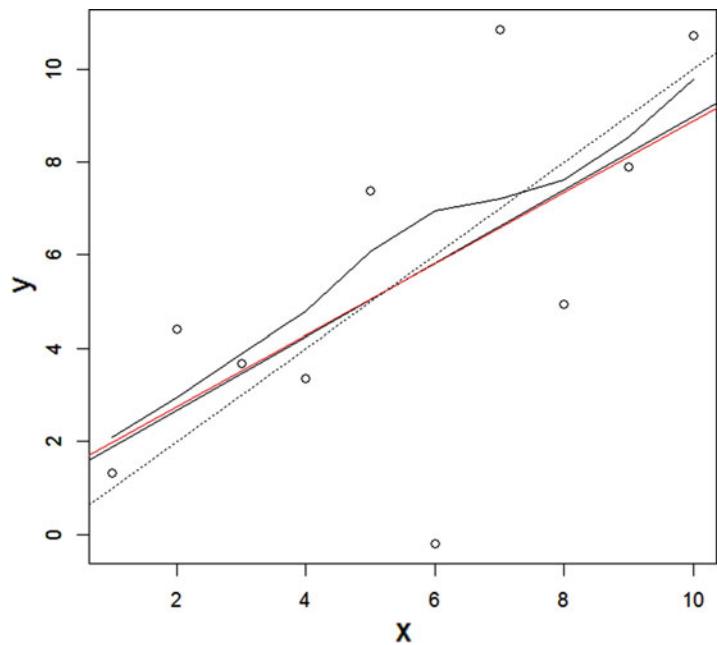


Fig. 2.10 A standard diagnostic plot to check for heteroscedasticity

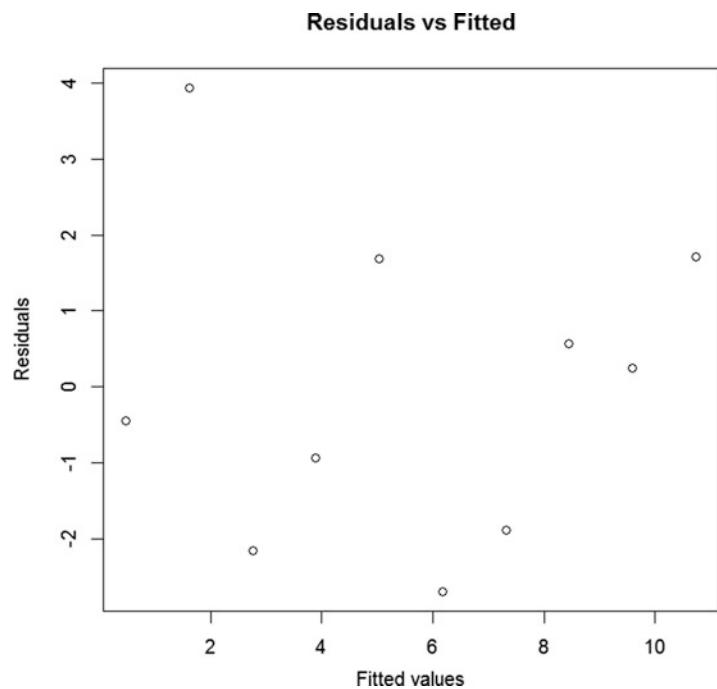
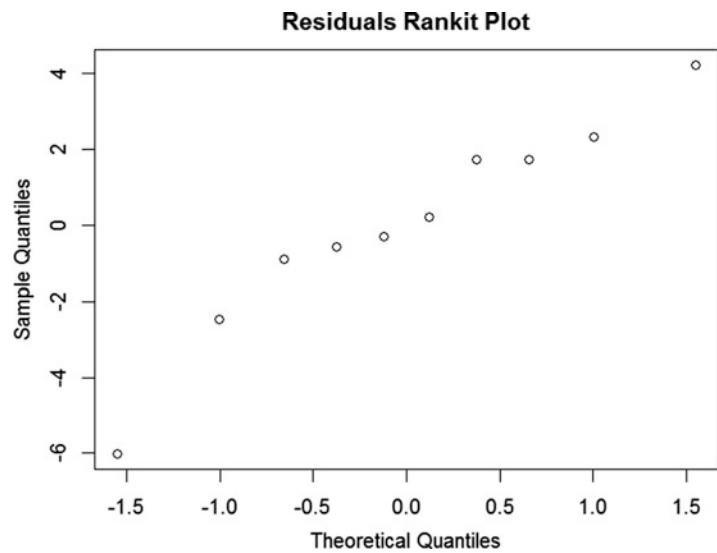


Fig. 2.11 A normal scores plot to check for skewness, kurtosis, and outliers



```

>
> rm(fm, fm1, lrf, x, dummy)
> # Removing these 5 objects
> fm
Error: object 'fm' not found      # Checked!
> fm1
Error: object 'fm1' not found    # Checked!
> lrf
Error: object 'lrf' not found    # Checked!
> x
Error: object 'x' not found      # Checked!
> dummy
Error: object 'dummy' not found # Checked!
# END OF THIS PRACTICE SESSION!

```

2.2.2 The R Environment

(THIS IS IMPORTANT!)

Getting through the First Session in the previous section, Sect. 2.2.1, shows:

Technically, R is an expression language with a simple syntax which is almost self-explanatory. It is case sensitive: so `x` and `X` are different symbols and refer to different variables. All alphanumeric symbols are allowed, plus `.` and `-`, with the restriction that a name must start with `.` or a letter, and if it starts with `.` the second character must not be a digit. The command prompt `>` indicates when R is ready for input.

This is where one types commands to be processed by R, which will happen when one hit the ENTER key.

Commands consist of either expressions or assignments.

When an expression is given as a command, it is immediately evaluated, printed, and the value is discarded. An assignment evaluates an expression and passes the value to a variable – but the value is not automatically printed. To printed the computed value, simple enter the variable again at the next command.

Commands are separated either by a new line, or separated by a semi-colon (`;`). Several elementary commands may be grouped together into one compound expression by braces: (`{` and `}`).

Comments, starting with a hashmark/number-sign (`#`), may be put almost anywhere: everything to the end of the line following this sign is a comment.

Comments may not be used in an argument list of a function definition or inside strings. If a command is not complete at the end of a line, R will give a different prompt, a “`+`” sign, by default:

On the second and subsequent lines, and continue to read input until the command is completed syntactically.

The result of a command is printed to the output device: if the result is an array, such as a vector or a matrix, then the elements are formatted with line break (wherever necessary) with the indices of the leading entries labeled in square brackets: `[index]`.

For example, an array of 15 elements may be outputted as:

```

> array(8, 15)
[1]  8  8  8  8  8  8  8  8  8
[11] 8  8  8  8  8

```

The labels ‘[1]’ and ‘[11]’ indicate the 1st and 11th elements in the output .

These labels are not part of the data itself!

Similarly, the labels for a matrix are placed at the start of each row and column in the output.

For example, for the 3×5 matrix M, it is outputted as:

```
>
> M <- matrix(1:15, nrow=3)
> M
      [,1] [,2] [,3] [,4] [,5]
[1,]    1    4    7   10   13
[2,]    2    5    8   11   14
[3,]    3    6    9   12   15
>
```

Note that the storage is a column-major, viz., the elements of the first column are printed out first, followed by those of the second column, etc. To cause a matrix to be filled in a row-wise manner, rather than the default column-wise fashion, the additional switch `byrow=T` will cause the matrix to be filled row-wise rather than by column-wise:

```
>
> M <- matrix(1:15, nrow=3, byrow=T)
> M
      [,1] [,2] [,3] [,4] [,5]
[1,]    1    2    3    4    5
[2,]    6    7    8    9   10
[3,]   11   12   13   14   15
>
```

The First Session also shows that there is a host of helpful resources imbedded in the R environment that one can readily access, using the on-line help provided by CRAN.

Review Questions for Sect. 2.2

1. Let us get started!

Please follow the step-by-step instructions given in the opening paragraphs of Sect. 2.2 to set up an R environment. The R window show look like this:

```
>
```

Great!

Now enter the following arithmetic operations: press “*Enter*” after each entry:

- (a) $2 + 3 <\text{Enter}>$
- (b) $13 - 7 <\text{Enter}>$
- (c) $17 * 23 <\text{Enter}>$
- (d) $100/25 <\text{Enter}>$
- (e) Did you obtain the following results:
5, 6, 391, 4?

2. Here is a few more: The *<Enter>* prompt will be omitted from now on!

- (a) 2^4
- (b) $\text{sqrt}(3)$
- (c) $1i$ [$1i$ is used for the complex unit i , where $i^2 = -1$.]
- (d) $(2 + 3i) + (4 + 5i)$
- (e) $(2 + 3i) * (4 + 5i)$

3. Here is a short session on using R to do complex arithmetic: just enter the following commands into the R environment, and report the results:

```
> th <- seq(-pi, pi, len = 20)
> th
(a) How many numbers are printed out?
> z <- exp(1i*th)
> z
(b) How many complex numbers are printed out?
> par(pty="s")
(c) Along the menu-bar at the top of the R environment:
*Select and left-click on “Window”; then
*Move downwards and select the 2nd option:
R Graphic Device 2 (ACTIVE)
*Go to the “R Graphic Device 2 (ACTIVE) Window”
(d) What is there?
> plot(z)
(e) Describe what is in the Graphic Device 2 Window.
```

2.3 R As a Calculator (Aragon 2011; Dalgaard 2002)

2.3.1 Mathematical Operations Using R

To learn to do statistical analysis and computations, one may start by considering the R programming language as a simple calculator!

Start from here:

just enter an arithmetic expression, press the **<Enter>** key, and the answer from the machine us found in the next line –

```
>
> 2 + 3
[1] 5
>
```

OK! What about other calculations? Such as:

$13 - 7$, 3×5 , $12/4$, 72 , $\sqrt{2}$, e^3 , $e^{i\pi}$,
 $\ln 5 = \log_e 5$, $(4 + \sqrt{3})(4 - \sqrt{3})$,
 $(4 + i\sqrt{3})(4 - i\sqrt{3})$, ... and so on.

Just try:

```
>
> 13 - 7
[1] 6
>
> 3*5
[1] 15
>
> 12/4
```

```
[1] 3
>
> 7^2
[1] 49
>
> sqrt(2)
[1] 1.414214
>
> exp(3)
[1] 20.08554
>
> exp(1i*pi)
# 1i is used for the complex number i = -1.
[1] -1-0i
```

[This is just the famous Euler's Identity equation: $e^{i\pi}+1 = 0$.]

```
> log(5)
[1] 1.609438
> (4 + sqrt(3))*(4 - sqrt(3))
[1] 13
```

[Checking: $(4+\sqrt{3})(4-\sqrt{3}) = 42 - (\sqrt{3})2 = 16 - 3 = 13$ (Checked!)]

```
> (4 + 1i*sqrt(3))*(4 - 1i*sqrt(3))
[1] 19+0i
```

[Checking: $(4+i\sqrt{3})(4-i\sqrt{3}) = 42 - (i\sqrt{3})2 = 16 - (-3) = 19$ (Checked!)]

Remark

The [1] in front of the computed result is R's way of outputting numbers. It becomes useful when the result is a long vector. The number N in the brackets [N] is the index of the first number on that line. For example, if one generated 23 random numbers from a normal distribution:

```
>
> x <- rnorm(23)
> x
[1] -0.5561324  0.2478934 -0.8243522  1.0697415  1.5681899
[6] -0.3396776 -0.7356282  0.7781117  1.2822569 -0.5413498
[11]  0.3348587 -0.6711245 -0.7789205 -1.1138432 -1.9582234
[16] -0.3193033 -0.1942829  0.4973501 -1.5363843 -0.3729301
[21]  0.5741554 -0.4651683 -0.2317168
>
```

Remark

After the random numbers have been generated, there is no output until one calls for x, viz., x has become a vector with 23 elements, call that a 23-vector!

The [11] on the third line of the output indicates that 0.3348587 (colored in red here for emphasis!) is the 11th element in the 23-vector x. The numbers of outputs per line depends on the length of each element as well as the width of the page.

2.3.2 Assignment of Values in R, and Computations Using Vectors and Matrices

R is designed to be a Dynamically-typed Language, viz., at any time one may change the data type of any variable. For example, one can first set `x` to be numeric as has been done so far, say: `x = 7`; next one may set `x` to be a vector, say: `x = c(1, 2, 3, 4)`; then again one may set `x` to a word object, such as “Hi!”. Just watch the following R environment:

```
>
> x <- 7
> x
[1] 7
> x <- c(1, 2, 3, 4)      # x is assigned to be a 4-vector.
> x
[1] 1 2 3 4
> x <- c("Hi!")  # x is assigned to be a character string.
> x
[1] "Hi!"
> x <- c("Greetings & Salutations!")
> x
[1] "Greetings & Salutations!"
> x <- c("The rain in Spain falls mainly on the
+      plain.")
> x
[1] "The rain in Spain falls mainly on the plain."
> x <- c("Biostatistics", "Human", "Genetic",
+      "Epidemiology")
> x
[1] "Biostatistics", "Human", "Genetic",
+      "Epidemiology"
>
```

2.3.3 Computations in Vectors and Simple Graphics

The use of arrays and matrices was introduced in Sect. 2.2.2. In finite mathematics, a matrix is a 2-dimensional array of elements, which are usually numbers. In R, the use of the matrix extends to elements of any type, such as a matrix of character strings. Arrays and matrices may be represented as vectors with dimensions.

In statistics in which most variables carry multiple values, therefore computations are usually performed between vectors of many elements.

These operations among multivariates result in large matrices. To demonstrate the results, often graphical representations are useful.

The following simple example illustrates these operations being readily accomplished in the R environment

```
>
> weight <- c(73, 59, 97)
> height <- c(1.79, 1.64, 1.73)
> bmi <- weight/height^2
> bmi  # Read the BMI Notes below
```

```
[1] 22.78331 21.93635 32.41004
> # To summarize the results proceed to compute
> # as follows:
> cbind(weight, height, bmi) # Outputting:
  weight     height     bmi
[1,]    73      1.79  22.78331
[2,]    59      1.64  21.93635
[3,]    97      1.73  32.41004
>
> rbind(weight, height, bmi) # Outputting:
     [,1]      [,2]      [,3]
weight 73.00000 59.00000 97.00000
height 1.79000 1.64000 1.73000
bmi    22.78331 21.93635 32.41004
>
```

Clearly, the functions cbind and rbind bind (viz., join, link, glue, concatenate) by column and by row, respectively, the vectors to form new vectors or matrices.

2.3.4 Use of Factors in R Programming

In the analysis of, for example, health science datasets, categorical variables are often needed. These categorical variables indicate subdivisions of the origin dataset into various classes, for example: age, gender, disease stages, degrees of diagnosis, etc. Input of the original dataset is generally delineated into several categories using a numeric code: 1 = age, 2 = gender, 3 = disease stage, etc. Such variables are specified as factors in R, resulting in a data structure that enables one to assign specific names to the various categories. In certain analyses, it is necessary for R to distinguish among categorical codes and variables whose values have direct numerical meanings.

A factor has 4 levels, consisting of 2 items:

- (1) a vector of integers between 1 and 4, and
- (2) a character vector of length four containing strings which describe the 4 levels.

Consider the following example:

**A certain type of cancer is being categorized into 4 levels: Levels 1, 2, 3, and 4, respectively.
 **The corresponding pain levels consistent with these diagnoses are: none, mild, moderate, and severe, respectively.
 **In the dataset, 5 case-subjects have been diagnosed in terms of their respective levels.

The following R code segment delineates the dataset:

```
> cancerpain <- c(1, 4, 3, 3, 2, 4)
> fcancerpain <- factor(cancerpain, level=1:4)
> levels(fcancerpain) <- c("none", "mild",
+                           "moderate", "severe")
```

The first statement creates a numerical vector `cancerpain` which encodes the pain levels of 6 case-subjects. This is being considered as a categorical variable for which, using the `factor` function, a factor `fcancerpain` is created. This may be called with one argument in addition to `cancerpain`, viz., `levels = 1 to 4`, which indicates that the input coding uses the values 1 – 4. In the final line, the pain level names are changed to the 4 specified character strings. The result is:

```
> fcancerpain
[1] none  severe  moderate  moderate  mild   severe
Levels: none  mild  moderate  severe
> as.numeric(fcancerpain)
[1] 1 4 3 3 2 4
> levels(fcancerpain)
[1] "none"      "mild"       "moderate"    "severe"
```

Remarks

The function `as.numeric()` outputs the numerical coding as numbers 1 to 4, and the function `levels()` outputs the names of the respective levels.

The original input coding in terms of the numbers 1 to 4 is no longer needed, There is an additional option using the function `ordered` which is similar to the function `factor` used here,

BMI (BMI Notes 2012)

The Body Mass Index (BMI), is a useful measure for human body fat based on an individual's weight and height – it does not actually measure the percentage of fat in the body. Invented in the early 19th century, BMI is defined as a person's body weight (in kilograms) divided by the square of the height (in meters). The formula universally used in health science produces a unit of measure of kg/m^2 :

$$\text{BMI} = \text{Body Mass (kg)} / \{\text{Height (m)}\}^2$$

A BMI chart may be used which, displaying BMI as a function of weight (horizontal axis) and height (vertical axis) with contour lines for different values of BMI or colors for different BMI categories: Fig. 2.12.

2.3.5 Simple Graphics

Generating graphical presentations is an important aspect of statistical data analysis. Within the R environment, one may construct plots that allows production of plots and control of the graphical features. Thus, with the previous example, the relationship between Body Weight and Height may be considered by first plotting one versus the other by using the following R code segments (Fig. 2.13):

```
>
> plot (weight, height)
> # Outputting: Fig. 2.13.
```

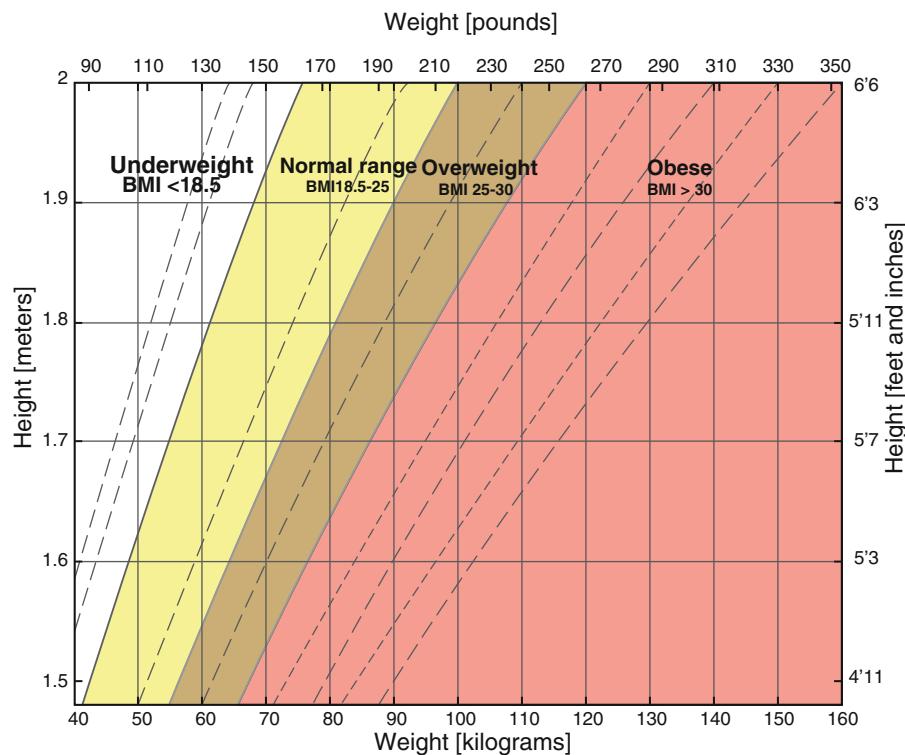


Fig. 2.12 A graph of BMI (Body Mass Index): the dashed lines represent subdivisions within a major class – the "Underweight" classification is further divided into "severe", "moderate", and "mild" subclasses. World Health Organization data (BMI Notes 2012)

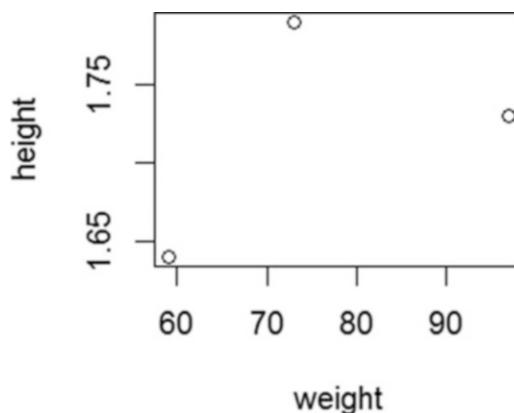


Fig. 2.13 An X-Y plot for > plot (weight, height)

Remarks

(1) Note the order of the parameters in the plot (x, y) command:

the first parameter is x (the independent variable – on the horizontal axis), and the second parameter is y (the dependent variable – on the vertical axis).

- (2) Within the R environment, there are many plotting parameters

that may be selected to modify the output. To get a full list of available options, return to the R environment and call for:

```
> ?plot # This is a call for "Help!" within the R environment.  
> # The output is the R documentation for:  
> plot {graphics} # Generic X-Y plotting
```

This is the official documentation of the R function `plot`, within the R package `graphics` – note the special notations used for `plot` and `{graphics}`. To fully make use of the provisions of the R environment, one should carefully investigate all such documentations. (R has many available packages, each containing a number of useful functions.) This document shows all the plotting options available with the R environment. A copy of this documentation is shown in Appendix 1 for reference.

For example, to change the plotting symbol, one may use the keyword `pch` (for “plotting character”) in the following R command: (Fig. 2.14)

```
> plot (weight, height, pch=8)  
> # Outputting: Fig. 2.14.
```

Note that the output is the same as that shown in Fig. 2.13, except that the points are marked with little “8-point stars”, corresponding to Plotting Character `pch = 8`.

In the documentation for `pch`, a total of 26 options are available, providing different plotting characteristics for points in R graphics. They are shown in Fig. 2.15.

The parameter `BMI` was chosen in order that this value should be independent of a person’s height, thus expressing as a single number or index indicative of whether a case-subject is overweight, and by what relative amount.

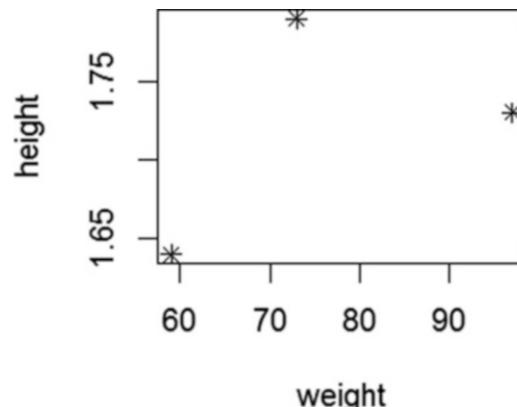


Fig. 2.14 An X-Y plot for `> plot (weight, height, pch=8)`



Fig. 2.15 Plotting symbols in R: `pch = n`, $n = 0, 1, 2, \dots, 25$

Of course, one may plot “height” as the abscissa (viz. the horizontal “x-axis”), and “weight” as the ordinate (viz., the vertical “y-axis”), as follows:

```
> plot(height, weight, pch=8)
# Outputting: Fig. 2.16.
```

Since a normal BMI is between 18.5 and 25, averaging $(18.5 + 25)/2 = 21.75$. For this BMI value, then the weight of a typical “normal” person would be $(21.75 \times \text{height}^2)$. Thus, one can superimpose a line-of-“expected”-weights at $\text{BMI} = 21.75$ on Fig. 2.16. This line may be accomplished in the R environment by the following code segments (Fig. 2.17):

```
> ht <- c(1.69, 1.64, 1.73)
> lines(ht, 21.75*ht^2) # Outputting: Fig. 2.17.
```

In the last plot, a new variable for heights (ht) was defined instead of the original (height) because:

1. The relation between height and weight is a quadratic one, and hence non-linear. Although it may not be obvious on the plot, it is preferable to use points that are spread evenly along the x-axis than to rely on the distribution of the original data.
2. As the values of height are not sorted, the line segments would not connect neighboring points but would run back and forth between distant points.

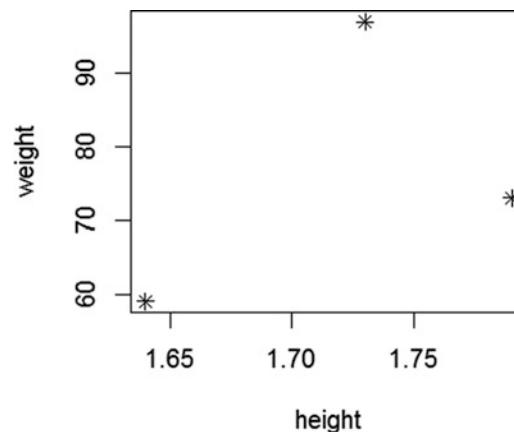


Fig. 2.16 An X-Y plot for > plot (height, weight, pch=8)

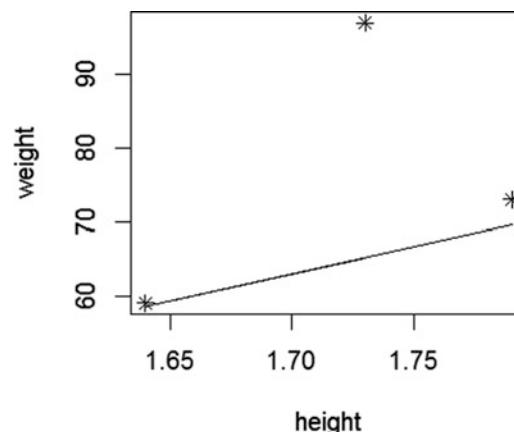


Fig. 2.17 Superimposed reference curve using line: (ht, 21.75*ht^2)

Remarks

1. In the final example above, R was actually doing the arithmetic of vectors.
2. Notice that the two vectors weight and height are both 3-vectors, making it reasonable to perform the next step.
3. The cbind statement, used immediately after the computations have been completed, forms a new matrix by binding together matrices horizontally, or column-wise. It results in a multivariate response variable. Similarly, the rbind statement does a similar operation vertically, or row-wise.
4. But, if for some reason (such as mistake in one of the entries) the two entries weight and height have different number of elements, then R will output an error message. For example:

```

>
> weight <- c(73, 59, 97) # a 3-vector
> height <- c(1.79, 1.64, 1.73, 1.48) # a 4-vector !
> bmi <- weight/height^2 # Outputting:
Warning message: # An Error message!
In weight/height^2 :
longer object length is not a multiple of shorter object length
>

```

2.3.6 x As Vectors and Matrices in Statistics

It has just been shown that a variable, such as x or M may be assigned as

- (1) a number, such as $x = 7$
- (2) a vector or an array, such as $x = c(1, 2, 3, 4)$
- (3) a matrix, such as

```

x =
      [,1] [,2] [,3] [,4] [,5]
[1,]    1    4    7   10   13
[2,]    2    5    8   11   14
[3,]    3    6    9   12   15

```

- (4) a character string, such as

```
x = "The rain in Spain falls mainly on the plain."
```

- (5) In fact, in R, a variable x may be assigned a complete dataset which may consist of a multiple-dimensional set of elements each of which may in turn be anyone of the above kinds of variables. For example, besides being a numerical vector, such as in (2) above, x may be:

a character vector, which is a vector of text strings whose elements are expressed in quotes, using double-, single-, or mixed-quotes:

```

> c("one", "two", "three", "four", "five")
> # Double-quotes
[1] "one"   "two"   "three" "four"  "five"

```

```

>
> c('one', 'two', 'three', 'four', 'five')
> # Single-quotes
[1] 'one'    'two'    'three'   'four'    'five'
>
> c("one", "two", "three", "four", "five")
> # Mixed-quotes
[1] "one"    'two'    "three"   'four'    "five"

```

However, if there is a mixed pair of quotes such as “xxxxx”, it will not be accepted! For example:

```
> c("one", "two", "three", "four", "five")
```

- (a) a logical vector, which takes the value TRUE or FALSE (or NA). For inputs, one may use the abbreviations T or F. These vectors are similarly specified using the c function:

```

> c(T, F, T, F, T)
[1] TRUE FALSE TRUE FALSE TRUE

```

In most cases, there is no need to specify logical vectors repeated. It is acceptable to use a single logical value to provide the needed options as vectors of more than one value will respond in terms of relational expressions. Observe:

```

> weight <- c(73, 59, 97)
> height <- c(1.79, 1.64, 1.73)
> bmi <- weight/height^2
> bmi # Outputting:
[1] 22.78331 21.93635 32.41004
> bmi > 25 # A single logical value will suffice!
[1] FALSE FALSE TRUE
>

```

2.3.7 Some Special Functions That Create Vectors

Three functions that create vectors are: c, seq, and rep

- (1) c, for “concatenate”, or, the joining of objects end-to-end (this was introduced earlier) – for example:

```

> x <- c(1, 2, 3, 4) # x is assigned to be a 4-vector.
> x
[1] 1 2 3 4

```

- (2) seq, for “sequence”, for defining an equidistant sequence of numbers – for example:

```

> seq(1, 20, 2)
# To output a sequence from 1 to 20, in steps of 2
[1] 1 3 5 7 9 11 13 15 17 19
> seq(1, 20)

```

```

# To output a sequence from 1 to 20, in steps of 1,
# (which may be omitted)
[1] 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17
[2] 18 19 20
> 1:20
# This is a simplified alternative to writing
# seq(1, 20)
[1] 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17
[2] 18 19 20
> seq(1, 20, 2.5)
# To output a sequence from 1 to 20, in steps of
# 2.5 .
[1] 1.0 3.5 6.0 8.5 11.0 13.5 16.0 18.5

```

(3) rep, for “replicate”, for generating repeated values.

This function takes two forms, depending on whether the second argument is a single number or a vector – for example:

```

> rep(1:2, c(3,5))
# Replicating the first element (1) 3 times, and
# then replicating the second element (2) 5 times
[1] 1 1 1 2 2 2 2 2 # This is the output.
> vector <- c(1, 2, 3, 4)

> vector      # Outputting the vector:
[1] 1 2 3 4
> rep(vector, 5)      # Replicating vector 5 times:
[1] 1 2 3 4 1 2 3 4 1 2 3 4 1 2 3 4 1 2 3 4

```

2.3.8 Arrays and Matrices

In finite mathematics, a matrix M is a 2-dimensional array of elements, generally numbers, such as

M	=	1	4	7	10	13
		2	5	8	11	14
		3	6	9	12	15

and the array is usually placed inside parenthesis (), or some brackets {}, [], etc. In R, the use of a matrix is extended to elements of many types: numbers as well as character strings.

For example, in R, the above matrix M is expressed as:

	[,1]	[,2]	[,3]	[,4]	[,5]
[1,]	1	4	7	10	13
[2,]	2	5	8	11	14
[3,]	3	6	9	12	15

2.3.9 Use of the Dimension Function `dim()` in R

In R, the above 3×5 matrix may be set up as vectors with dimension `dim(x)` using the following code segment:

```
> x <- 1:j
> x
[1] 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
> dim(x) <- c(3, 5)      # a dimension of 3 rows by 5 columns
```

Remark

Here a total of 15 elements, 1 through 15, are set to be the elements of the matrix `x`. Then the dimension of `x` is set as `c(3, 5)`, making `x` to become a 3×5 matrix. The assignment of the 15 elements follows a column-wise procedure, viz., the elements of the first column are allocated first, followed by those of the second column, then the third column, etc.

2.3.10 Use of the Matrix Function `matrix()` in R

Another way to generate a matrix is using the function `matrix()`.

The above 3×5 matrix may be created by the following 1-line code segment:

```
> matrix (1:15, nrow=3)
> matrix # Outputting:
```

```
[,1] [,2] [,3] [,4] [,5]
[1,] 1 4 7 10 13
[2,] 2 5 8 11 14
[3,] 3 6 9 12 15
```

However, if the 15 elements should be allocated by row, then the following code segment should be used:

```
> matrix (1:15, nrow=3, byrow=T)
> matrix # Outputting:
[,1] [,2] [,3] [,4] [,5]
[1,] 1 2 3 4 5
[2,] 6 7 8 9 10
[3,] 11 12 13 14 15
```

2.3.11 Some Useful Functions Operating on Matrices in R

`colnames`, `rownames`, and `t` (for transpose)

Using the previous example:

- (i) the 5 columns of the 3×5 matrix `x` is first assigned the names `C1, C2, C3, C4` and `C5` respectively, then
- (ii) the transpose is obtained, and finally
- (iii) one take the transpose of the transpose to obtain the original matrix `x`:

```

> matrix (1:15, nrow=3, byrow=T)
> matrix # Outputting:
     [,1]  [,2]  [,3]  [,4]  [,5]
[2,]    6    7    8    9   10
[3,]   11   12   13   1 4   15
> colnames(x) <- c("C1", "C2", "C3", "C4", "C5")
> x # Outputting:
     C1  C2  C3  C4  C5
[1,]  1   4   7  10  .gf
[2,]  2   5   8  11  14
[3,]  3   6   9  12  15
> t(x)
     [,1]  [,2]  [,3]
C1    1    2    3
C2    4    5    6
C3    7    8    9
C4   10   11   12
C5   13   14   15
> t(t(x)) # which is just x, as expected!

     C1  C2  C3  C4  C5
[1,]  1   4   7  10  13
[2,]  2   5   8  11  14
[3,]  3   6   9  12  15

```

Yet another way is to use the function LETTERS, which is a built-in variable containing the capital letters A through Z. Other useful vectors include letters, month.name, and month.abb for lower-case letters, month names, and abbreviated names of months, respectively. Take a look:

```

> X <-LETTERS
> X # Outputting:
[1] "A" "B" "C" "D" "E" "F" "G" "H" "I" "J" "K" "L" "M" "N" "O"
[16] "P" "Q" "R" "S" "T" "U" "V" "W" "X" "Y" "Z"
> M # Outputting:
[1] "January" "February" "March"      "April"      "May"
[6] "June"     "July"      "August"     "September" "October"
[11] "November" "December"
> m <- month.abb
> m # Outputting:
[1] "Jan" "Feb" "Mar" "Apr" "May" "Jun" "Jul" "Aug"
[9] "Sep" "Oct" "Nov" "Dec"

```

2.3.12 NA 'Not Available' for Missing Values in Datasets

NA is a logical constant of length 1 which contains a missing value indicator.

NA may be forced to any other vector type except raw. There are also constants

NA integer, NA real, NA complex, and NA character of the other atomic vector types which support missing values: all of these are **reserved** words in the R language.

The generic function is `.na` indicates which elements are missing.

The generic function `.na<-` sets elements to NA.

The reserved words in R's parser are: `if`, `else`, `repeat`, `while`, `function`, `for`, in `next`, `break`, NA complex, `NA character`, ... and `..1`, `..2`, etc., which are used to refer to arguments passed down from an enclosing function.

Reserved words outside `quotes` are always parsed to be references to the objects linked to in the foregoing list, and are not allowed as syntactic names. They are allowed as non-syntactic names.

2.3.13 Special Functions That Create Vectors

There are three useful R functions which are often used to create vectors:

- (1) `c` for “concatenate”, which was introduced in Sect. 2.3.2 for joining items together end-to-end, for example:

```
> c(2, 3, 5, 7, 11, 13, 17, 19, 23, 29)
> # The first 10 prime numbers
[1] 2 3 5 7 11 13 17 19 23 29
```

- (2) `seq` for “sequence”, is used for listing equidistant sequences of numbers, for example:

```
> seq(1, 20)      # Sequence from 1 to 20
[1] 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17
     18 19 20
> seq(1, 20, 1)  # Sequence from 1 to 20, in steps of 1
[1] 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17
     18 19 20
> 1:20            # Sequence from 1 to 20, in steps of 1
[1] 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17
     18 19 20
> seq(1, 20, 2)  # Sequence from 1 to 20, in steps of 2
[1] 1 3 5 7 9 11 13 15 17 19
> seq(1, 20, 3)  # Sequence from 1 to 20, in steps of 3
[1] 1 4 7 10 13 16 19
> seq(1, 20, 10) # Sequence from 1 to 20, in steps of 10
[1] 1 11
> seq(1, 20, 20) # Sequence from 1 to 20, in steps of 20
[1] 1
> seq(1, 20, 21) # Sequence from 1 to 20, in steps of 21
[1] 1
>
```

- (3) `rep` for “replicate”, is used to generate repeated values, and may be expressed in 2 ways, for example

```
> x <- c(3, 4, 5)
> rep(x, 4)      # Replicate the vector x 4-times.
[1] 3 4 5 3 4 5 3 4 5 3 4 5
> rep(x, 1:3)    # Replicate the elements of x:
```

```

# the first element once,
# the second element twice, and
# the third element three times.
[1] 3 4 4 5 5 5
> rep(1:3, c(3,4,5)) # For the sequence (1, 2, 3), replicate its
>                      # elements 3-, 4-, and 5-times, respectively
[1] 1 1 1 2 2 2 2 3 3 3 3 3

```

Review Questions for Sect. 2.3

1. A Tower of Powers – by computations using R

There is an interesting challenge in arithmetic which goes like this:

$$\sqrt{2}^{\sqrt{2}}$$

What is the value of $\sqrt{2}^{\sqrt{2}} \dots$?

viz. an infinity of ascending tower of powers of the square root of 2.

Solution: Let x be the value of this “Tower of Powers”, then it is easily seen that $\sqrt{2}^x = x$ itself ! Agree?

Watch the lowest $\sqrt{2}$.

And clearly it follows that $x = 2$, because $\sqrt{2}^2 = 2$.

This shows that the value of this “Infinite Tower of Powers of $\sqrt{2}$ ” is just 2.

Now use the R environment to verify this interesting result:

(a) Compute $\sqrt{2}$

```
> sqrt(2)
```

(b) Compute $\sqrt{2}^{\sqrt{2}}$

```
> sqrt(2)^sqrt(2) [a 2-Towers of 2-s]
```

(c) > sqrt(2)^sqrt(2)^sqrt(2) [a 3-Towers of $\sqrt{2}$ -s]

(d) > sqrt(2)^sqrt(2)^sqrt(2)^sqrt(2) [a 4-Towers of $\sqrt{2}$ -s]

(e) > sqrt(2)^sqrt(2)^sqrt(2)^sqrt(2)^sqrt(2) [a 5-Towers of $\sqrt{2}$ -s]

(f) Now try the following computations of 10-, 20-, 30-, and finally 40-Towers of Powers of $\sqrt{2}$, and finally reach the result of 2 (accurate to 6 places of decimal!).

```
> sqrt(2)^sqrt(2)^sqrt(2)^sqrt(2)^sqrt(2)^sqrt(2)^sqrt(2)^sqrt(2)^sqrt(2)^sqrt(2)
[1] 1.983668 [a 10-Towers of Powers of 2-s]
```

```
> sqrt(2)^sqrt(2)^sqrt(2)^sqrt(2)^sqrt(2)^sqrt(2)^sqrt(2)^sqrt(2)^sqrt(2)^sqrt(2)^sqrt(2)
+ sqrt(2)^sqrt(2)^sqrt(2)^sqrt(2)^sqrt(2)^sqrt(2)^sqrt(2)^sqrt(2)^sqrt(2)^sqrt(2)^sqrt(2)^sqrt(2)
+ sqrt(2)^sqrt(2)^sqrt(2)^sqrt(2)^sqrt(2)^sqrt(2)^sqrt(2)^sqrt(2)^sqrt(2)^sqrt(2)^sqrt(2)^sqrt(2)^sqrt(2)
+ sqrt(2)^sqrt(2)^sqrt(2)^sqrt(2)^sqrt(2)^sqrt(2)^sqrt(2)^sqrt(2)^sqrt(2)^sqrt(2)^sqrt(2)^sqrt(2)^sqrt(2)
[1] 1.999586 [a 20-Towers of Powers of 2-s]
```

```

> sqrt(2)^sqrt(2)^sqrt(2)^sqrt(2)^sqrt(2)^sqrt(2)^
+ sqrt(2)^sqrt(2)^sqrt(2)^sqrt(2)^sqrt(2)^sqrt(2)^
+ sqrt(2)^sqrt(2)^sqrt(2)^sqrt(2)^sqrt(2)^sqrt(2)^
+ sqrt(2)^sqrt(2)^sqrt(2)^sqrt(2)^sqrt(2)^sqrt(2)^
+ sqrt(2)^sqrt(2)^sqrt(2)^sqrt(2)^sqrt(2)^sqrt(2)^
[1] 1.999989      [a 30-Towers of Powers of 2-s]

> sqrt(2)^sqrt(2)^sqrt(2)^sqrt(2)^sqrt(2)^sqrt(2)^
+ sqrt(2)^sqrt(2)^sqrt(2)^sqrt(2)^sqrt(2)^sqrt(2)^
[1] 2   (Accurate to 6 decimal places!)
      [a 40-Towers of Powers of 2-s]

```

Thus, this R computation verifies the solution.

2.
 - (a) What are the equivalents in R for the basic mathematical operations of: +, -, x, / (division), $\sqrt{}$, squaring of a number
 - (b) Describe the use of factors in R programming. Give an example.
3. If $x = (0, 1, 2, 3, 4, 5)$, and $y = (0, 1, 4, 9, 16, 25)$, in R, plot: (a) y vs x , (b) x vs y , (c) \sqrt{y} vs x , (d) y vs \sqrt{x} , (e) \sqrt{y} vs \sqrt{x} , (e) \sqrt{x} vs \sqrt{y} .
4. Explain, given an example, how the following functions may be used to combine matrices to form new ones:
 - (a) cbind, (b) rbind.
5.
 - (a) Describe is the R function factor().
 - (b) Give an example of using factor() to create new arrays.
6. Using examples, illustrate two procedures for creating:
 - (a) a vector; (b) a matrix.
7. Describe, using examples, the following 3 functions for creating vectors:
 - (a) c, (b) seq, and (c) rep.
8.
 - (a) Use the function dim() to set up a matrix. Give an example.
 - (b) Use the function matrix() to set up a matrix. Give an example.
9. Describe, using an example, the use of the following functions operating on a matrix in R: t(), colnames(), and rownames().
10.
 - (a) What are reserved word in the R environment?
 - (b) In R, how is the logical constant NA used? Give an example.

Exercises for Sect. 2.3 (Virasakdi n.d.; Everitt and Hothorn 2006)

Enter the R environment, and do the following exercises using R programming:

1. Perform the following elementary arithmetic exercises:

(a) $7 + 31$; (b) $87 - 23$; (c) $3.1417 \times (7)2$; (d) $22/7$; (e) $e\sqrt{2}$

2. Body Mass Index (BMI) is calculated from your weight in kilograms and your height in meters:

$$\text{BMI} = \text{kg}/\text{m}^2$$

Using $1 \text{ kg} \approx 2.2 \text{ lb}$, and $1\text{m} \approx 3.3 \text{ f.} \approx 39.4 \text{ in}$

(a) Calculate your BMI.

(b) Is it in the “normal” range $18.5 \leq \text{BMI} \leq 25$?

3. In the MPH program, 5 graduate students taking the class in Introductory Epidemiology measured their weight (in kg) and height (in meters). The result is summarized in following matrix:

	John	Chang	Michael	Bryan	Jose
WEIGHT	69.1	62.5	74.3	70.9	96.6
HEIGHT	1.81	1.46	1.69	1.82	1.74

(a) Construct a matrix showing their BMI as the last row.

(b) Plot: (i) WEIGHT (on the y-axis) vs HEIGHT (on the x-axis)

i. (ii) HEIGHT vs WEIGHTi

ii. (iii) Assuming that the weight of a typical “normal”: person is $(21.75 \times \text{HEIGHT}^2)$, superimpose a line-of- “expected”-weight at $\text{BMI} = 21.75$ on the plot in (i).

4. (a) To convert between temperatures in degrees Fahrenheit (F) and Celsius (C), the following conversion formulas are used:

$$\begin{aligned} F &= (9/5)C + 32 \\ C &= (5/9) \times (F - 32) \end{aligned}$$

At standard temperature and pressure, the freezing and boiling points of water are 0 and 100 degrees Celsius, respectively. What are the freezing and boiling points of water in degrees Fahrenheit?

- (b) For $C = 0, 5, 10, 15, 20, 25, \dots, 80, 85, 90, 95, 100$, compute a conversion table that shows the corresponding F temperatures.

Note: To create the sequence of Celsius temperatures use the R function `seq(0, 100, 5)`.

5. Use the data in Table A below(Aragon 2011; Centers for Disease Control and Prevention 2005). Assume a person is initially HIV-negative. If the probability of getting infected per act is p , then the probability of not getting infected per act is

$$(1 - p).$$

Table A Estimated per-act risk (transmission probability) for acquisition of HIV, by exposure route to an infected source.

Exposure route	Risk per 10,000 exposures
Blood Transfusion (BT)	9000
Needle-sharing Injection-Drug Use (IDU)	67

Source: CDC (2005)

The probability of not getting infected after 2 consecutive acts is $(1 - p)^2$, and the probability of not getting infected after 3 consecutive acts is $(1 - p)^3$. Therefore, the probability of not getting infected after n consecutive acts is $(1 - p)^n$, and the probability of getting infected after n consecutive acts is $1 - (1 - p)^n$.

- For the non-blood transfusion transmission probability (per act risk) in Table A, calculate the risk of being infected after one year (365 days) if one carries out the Needle-sharing Injection-Drug Use (IDU) once daily for one year.
- Do these cumulative risks seem reasonable? Why? Why not?

Solution

```
> p <- 67/10000
> p
[1] 0.0067
> q <- (1 - p)
> q
[1] 0.9933
> q365 <- q^365
> q365
[1] 0.08597238
> p365 <- 1 - q365
> p365
[1] 0.9140276
# => Probability of being infected in a year = 91.40%.
#     A high risk, indeed!
```

2.4 Using R in Data Analysis in Human Genetic Epidemiology

In epidemiology, after preparing the collected datasets to undertake biostatistical analysis, the first step is to enter the datasets into the R environment. Once the datasets are placed

within the R environment, analysis will process the data to obtain results leading to creditable conclusions, and likely recommendations for definitive courses of actions to improve pertinent aspects of public and personal data. Several methods for dataset entry will be examined.

2.4.1 Entering Data at the R Command Prompt

Data Frames and Datasets (Virasakdi n.d.)

For many biostatistical investigators, the terms data frame and dataset may be used interchangeably.

Datasets In many applications, a complete dataset contain several data frames, including the real data that have been collected.

Data Frames Rules for data frames are similar to those for arrays and matrices, introduced earlier. However, data frames are more complicated than arrays. In an array, if just one cell is a character, then all the columns will be characters. On the other hand, a data frame can consist of:

- a column of “IDnumber”, which is numeric, and
- a column of “Name”, which is character.

In a data frame:

- each variable can have long variable descriptions,
- a factor can have “levels” or value levels.

These properties can be transferred from the original dataset in other software formats (such as SPS, Stata, etc.). They can also be created in R.

2.4.1.1 Creating a Data-Frame for R Computation using the EXCEL Spreadsheet (on a Windows Platform)

As an example using a typical set of real case-control epidemiologic research data, consider the dataset in Table B, from a clinical trial to evaluate the efficacy of maintenance chemotherapy for case-subjects with Acute Myelogenous Leukemia (AML), conducted at Stanford University, California, U.S.A., in 1977. After reaching a status of remission through treatment by chemotherapy, the patients who entered the study were assigned randomly to 2 groups:

- Maintained: this group received maintenance chemotherapy
- Non-maintained: this group did not, & is the Control Group.

The clinical trial was to ascertain whether maintenance chemotherapy prolonged the time until relapse (= “death”).

Procedure:

- To create an Acute Myelogenous Leukemia (AML) data file, called AML.csv, in Windows, &
- To input it into R as a data file AML

Table B Data for the AML maintenance clinical study (a + indicates a censored value)^a

Group	Duration for complete remission (weeks)
1=Maintained (11)	9,13,13+,18,23,28+,31,34,45+,48,161+ } 1=Uncensored
0=Non-maintained (12)	5, 5, 8, 8, 12, 16+,23,27,30,33,43,45} 0=Censored (+).

NB: The Non-maintained Group may be considered as MBD^a

The AML Clinical Study Data: Tableman & Kim (2004). – Table B-1: 23 data points, taken from – “Survival Analysis Using S: Analysis of Time-to-Event Data” by Mara Tableman and Jong Sung Kimz, published by Chapman & Hall/CRC, Boca Raton, 2004

^aThe cancer epigenome is characterised by specific DNA methylation and chromatin modification patterns. The proteins that mediate these changes are encoded by the epigenetics genes here defined as: DNA methyltransferases (DNMT), methyl-CpG-binding domain (MBD) proteins, histone acetyltransferases (HAT), histone deacetylases (HDAC), histone methyltransferases (HMT) and histone demethylases

(1) Creating a Data-frame for R Computation

1. Data Input, Using EXCEL:

(a) Open the EXCEL spreadsheet

(b) Type in data such that the variable names are in the row 1 of the EXCEL spreadsheet.

(c) Consider each row of data as an individual in the study.

(d) Start with column A.

2. Save as a .csv file:

(a) Click: “File” → “Save as” → and then, in the file name box (the upper box at the bottom) type: AML

(b) In the “Save in:” Box (at the top), choose: “Local Disc (C:)”

The file AML will be then be saved in the top level of the C:Drive, but another level may also be chosen.

In the “Save as Type” Box (the lower box at the bottom), scroll down, select, & click on: CSV (Comma delimited = Comma Separated Values)

To close out of EXCEL using the big “X” at the top- right-hand corner: Click X.

3. In Windows, check the C:Drive for the AML.csv file., viz., C:AML

4. Read AML into R:

(a) Open R

(b) Use the read.csv() function:

```
> aml <- read.csv("C:\crAML.csv", header = T,
+                   sep= ",")
```

(c) Actually, it can be also be done by:

```
> aml <- read.csv("C:\crAML.csv")
> # Read in the AML.csv file from the C:Drive of
> # the Computer, and call it aml
```

5. Output the AML.csv file for inspection

```
> aml # Outputting:
  weeks group status
    1     9     1     1
    2    13     1     1
    3    13     1     0
    4    18     1     1
    5    23     1     1
    6    28     1     0
    7    31     1     1
    8    34     1     1
    9    45     1     0
   10    48     1     1
   11   161     1     0
```

```

12      5      0      1
13      5      0      1
14      8      0      1
15      8      0      1
16     12      0      1
17     16      0      0
18     23      0      1
19     27      0      1
20     30      0      1
21     33      0      1
22     43      0      1
23     45      0      1
>

```

2.4.1.2 Obtaining a Data Frame from a Text File

Data from various sources are often entered using many different software programs.

They may be transferred from one format to another through the ASCII file format.

For example, in Windows, a text file is the most common ASCII file, usually having a ".txt" extension. There are other files in ASCII format, including the ".R" command file.

Data from most software programs can be outputted or saved as an ASCII file. From EXCEL, a very common spreadsheet program, the data can be saved as ".csv" (comma separated values) format. This is an easy way to interface between EXCEL spreadsheet files and R. Open the EXCEL file and 'save as' the csv format.

- Files with field separators:

As an example, suppose the file "csv1.xls" is originally an EXCEL spreadsheet. After 'save as' into csv format, the output file is called "csv1.csv", the contents of which is:

```

"name",    "gender",    "age"
"A",        "F",        20
"B",        "M",        30
"C",        "F",        40

```

The characters are enclosed in quotes and the delimiters (variable separators) are commas. Sometimes the file may not contain quotes, as in the file "csv2.csv".

```

name,    gender,    age
A,        F,        20
B,        M,        30
C,        F,        40

```

For both files, the R command to read in the dataset is the same.

```

> a <- read.csv("csv1.csv", as.is=TRUE)
> a

```

	name	gender	age
1	A	F	20
2	B	M	30
3	C	F	40

The argument 'as.is=TRUE' keeps all characters as they are, otherwise the characters would have been coerced into factors. The variable 'name' should not be factored but 'gender' should. The following command should therefore be entered:

```
> a$gender <- factor(a$gender)
```

Note that the object 'a' has class data frame and that the names of the variables within the data frame 'a' must be referenced using the dollar sign notation \$. Otherwise, R will state that the object 'gender' cannot be found.

For files with white space (spaces and tabs) as the separator, such as in the file "data1.txt", the command to use is read.table():

```
> a <- read.table("data1.txt", header=TRUE,
+                  as.is=TRUE)
```

- Files without field separators:

Consider the file "data2.txt" which in fixed field format without field separators.

	name	gender	age
1	A	F	20
2	B	M	30
3	C	F	40

To read in such a file, use the function read.fwf():

- Skip the first line, which is the header.
- The width of each variable and the column names must be specified.

```
> a <- read.fwf("data2.txt", skip=1, width=c(1,1,2),
+                  col.names = c("name", "gender", "age"),
+                  as.is=TRUE)
```

2.4.1.3 Data Entry and Analysis Using the Function data.entry()

The previous section deals with creating data frames by reading in data created from programs outside R, such as EXCEL. It is also possible to enter data directly into R by using the function data.entry(). However, if the data size is large (say more than 15 columns and/or more than 25 rows), the chance of human error is high with the spreadsheet or text mode data entry. A software program specially designed for data entry, such as Epidata, is more appropriate.

The web site of Epidata is: <http://www.epidata.dk>.

2.4.1.4 Data Entry Using Several Available R Functions

The dataset(Aragon 2011), in Table C, listing deaths among subjects who received a dose of tolbutamide or a placebo in the University Group Diabetes Program (1970), stratifying by age:

The R functions that can be used to import the data frame have been previously introduced in Sects. 2.3.3, 2.3.4, 2.3.5, 2.3.6, 2.3.7, 2.3.8, 2.3.9, 2.3.10, 2.3.11, 2.3.12, and 2.3.13.

Table C Deaths Among Subjects Who Received Tolbutamide or a Placebo in the University Group Diabetes Program (1970)^a

Age < 55		Age \geq 55			Combined	
Tolbutamide		Placebo			Tolbutamide	
Deaths	8	5	22	16	30	21
Survivors	98	115	76	69	174	184

^aAvailable at <http://www.medepi.net/data/ugdp.txt>

A convenient way to enter data at the command prompt is to use the R functions :

`c()`, `matrix()`, `array()`, `apply()`, `list()`, `data.frame()`,
and `odd.ratio()`,

as shown by the following examples and using the data in Table C.

```

> #Entering data for a vector
> vector1 <- c(8, 98, 5, 115) # Using data from
+                               Table C.
> vector1
[1] 8 98 5 115
>
> vector2 <- c(22, 76, 16, 69); vector2
> # Data from Table C.
[1] 22 76 16 69
>
> # Entering data for a matrix
> matrix1 <- matrix(vector1, 2, 2)
> matrix1
     [,1] [,2]
[1,]    8    5
[2,]   98   115
> matrix2 <- matrix(vector2, 2, 2); matrix2
     [,1] [,2]
[1,]   22   16
[2,]   76   69
>
> # Entering data for an array
> udata <- array(c(vector1, vector2), c(2, 2, 2))
> udata
, , 1
     [,1] [,2]
[1,]    8    5
[2,]   98   115
,
, , 2
     [,1] [,2]
[1,]   22   16
[2,]   76   69
>
> apply(udata, c(1, 2), sum); udata.tot
```

```

[,1]      [,2]
[1,]      30      21
[2,]     174     184
>
> # Entering a list
> x <- list(crude.data = udata.tot, stratified.data =
+           udata)
> x$crude.data
[,1]      [,2]
[1,]      30      21
[2,]     174     184
> x$stratified
, , 1
[,1]      [,2]
[1,]      8       5
[2,]     98     115
, , 2
[,1]      [,2]
[1,]     22      16
[2,]     76      69
>
> # Entering a simple data frame
> subjectname <- c("Peter", "Paul", "Mary")
> subjectnumber <- 1:length(subjectname)
> age <- c(26, 27, 28) # These are their true ages,
>                      # respectively, in 1964!
> gender <- c("Male", "Male", "Female")
> data1 <- data.frame(subjectnumber, subjectname,
+                      age, gender)
> data1
  subjectnumber subjectname  age gender
1            1       Peter  26   Male
2            2       Paul  27   Male
3            3       Mary  28 Female
>
> # Entering a simple function
> odds.ratio <- function(aa, bb, cc, dd){ aa*dd /
+                      (bb*cc) }
> odds.ratio(30, 174, 21, 184) # Data from Table C.
[1] 1.510673

```

2.4.1.5 Data Entry and Analysis Using the Function `scan()`(Teetor 2011)

The R function `scan()` is taken from the CRAN package base.

This function reads data into a vector or list from the console or file. This function takes the following usage form:

```
scan(file = "", what = double(), nmax = -1, n = -1,
      sep = "",
      quote = if(identical(sep, "\n")) "" else
      "'\"', dec = ".",
      skip = 0, nlines = 0, na.strings = "NA",
      flush = FALSE, fill = FALSE, strip.white =
      FALSE,
      quiet = FALSE, blank.lines.skip = TRUE,
      multi.line = TRUE,
      comment.char = "", allowEscapes =
      FALSE,
      fileEncoding = "", encoding = "unknown",
      text)
```

Argument

what The type of what gives the type of data to be read. The supported types are logical, integer, numeric, complex, character, raw and [list](#). If what is a list, it is assumed it is assumed that the lines of the data file are records each containing length (what) items (“fields”) and the list components components should have elements that are one of the first six types six types listed or NULL.

The what argument describes the tokens that scan() should expect in the input file.

For a detailed description of this function, execute:

```
> ?scan
```

The methodology of applying scan() is similar to c() , as described in the previous section: Sect. 2.4.1.4, except that it does not matter the numbers are being entered on different lines, it will still be a vector.

- Use scan() when accessing data from a file that has an irregular or a complex structure.
- Use scan() to read individual tokens and use the argument what to describe the stream of tokens in the file.
- scan() converts tokens into data, and then assemble the data into records.
- Use scan() along with readLines(), especially when one attempts to read an unorthodox file format. Together, these two functions will likely result in a successful processing the individual lines and tokens of the file!

The function readLines() reads lines from a file, and returns them to a list of character strings:

```
> lines <- readLines("input.text")
```

One may limit the number of lines to be read, per pass, by using the `n` parameter which gives the maximum number of lines to be read:

```
> lines <- readLines("input.txt", n=5)
> # read 5 lines and stop
```

The function `scan()` reads one token at a time, and handles it accordingly as instructed.

An example:

Assume that the file to be scanned and read contains triplets of data (like the Dates, and the corresponding daily Highs & Lows of financial markets):

```
15-Oct-1987 2439.78 2345.63
16-Oct-1987 2396.21 2207.73
19-Oct-1987 2164.16 1677.55
20-Oct-1987 2067.47 1616.23
21-Oct-1987 2087.07 1951.76
```

Use a list to operate `scan()` that it should expect a repeating, 3-token sequence:

```
> triplets <- scan("triples.txt", what=list(character(0),
+                           numeric(0), numeric(0)))
```

Give names to the list elements, and `scan()` will assign those names to the data:

```
> triplets <- scan("triples.txt",
+                     what=list(date=character(0),
+                               high=numeric(0), low=numeric(0)))
```

Reads 5 records.

```
> triples # Outputs:
$date
[1] "15-Oct-1987"  "15-Oct-1987"  "19-oct-1987"  "20-Oct-1987"  "21-oct-1987"
$high
[1] 2439.78 2396.21 2164.16 2067.47 2081.07
$low
[1] 2345.63 2207.73 1677.55 1616.21 1951.76
```

2.4.1.6 Data Entry and Analysis Using the Function `source()` (Venables and Smith 2004; Aragon 2011; Teeter 2011)

The R function `source()` is also taken from the CRAN package base. This function reads data into a vector or list from the console or file. It takes the following usage form:

`source()` causes R to accept its input from the named file or URL or connection.

Input is read and `parsed` from that file until the end of the file is reached, then the parsed expressions are evaluated sequentially in the chosen environment:

```
source(file, local = FALSE, echo = verbose, print.eval
      = echo,
      verbose =getOption("verbose"),
      prompt.echo =getOption("prompt"),
      max.deparse.length = 150, chdir = FALSE,
```

```
encoding =getOption("encoding"),
continue.echo =getOption("continue"),
skip.echo = 0, keep.source =
getOption("keep.source"))
```

For commands which are stored in an external file, such as ‘commands.R’ in the working directory ‘work’, they can be executed in an R environment with the command

```
> source("command.R")
```

The function source() instructs R to read the text and execute its contents. Thus, when one has a long, or frequently used, piece of R code, one may capture it inside a text file. This allows one to rerun the code without having to re-type it, and use the function source() to read and execute the code.

For example, suppose the file howdy.R contains the familiar greeting:

```
Print ("Hi, My Friend!")
```

Then by sourcing the file, one may execute the content of the file, as in the follow R code segment:

```
> source("howdy.R")
[1] "Hi, My Friend!"
```

Setting echo-TRUE will echo the same script lines before they are executed, with the R prompt shown before each line:

```
> source("howdy.R", echo=TRUE)
> Print("Hi, My Friend!")
[1] "Hi, My Friend!"
```

2.4.1.7 Data Entry and Analysis Using the Spreadsheet Interface in R (Aragon 2011)

This method consists of the following R functions in the package Utils.

Spreadsheet Interface for Entering Data

This is a spreadsheet-like editor for entering or editing data, with the following R functions:

```
data.entry(..., Modes = NULL, Names = NULL)
dataentry(data, modes)
de(..., Modes = list(), Names = NULL)
```

The arguments of these R functions are:

- ... A list of variables: currently these should be numerals or character vectors or a list containing such vectors.

Modes The modes to be used for the variables.

Names The names to be used for the variables.

data A list of numeric and/or character vectors.

modes A list of length up to that of data giving the modes of (some of) the variables. list() is allowed.

The function data.entry() edits an existing object, saving the changes to the original object name.

However, the function `edit()` edits of an existing object but not saving the changes to the original object name so that one must assign it to an object name (even if it is the original name). To enter a vector one needs to initialize a vector and then use the function `data.entry()`. For example:

Start by entering the R environment, and set

```
> x <- c(2, 4, 6, 8, 10)
> # X is initially defined as an array of 5 elements.
> x # Just checking - to make sure!
[1] 2 4 6 8 10 # x is indeed set to be an array of 5 elements
>
> data.entry(x) # Entering the Data Editor:

> # The Data Editor window pops up, and looking at
> # the first
> # column: it is now named "x", with the first 5
> # rows (all on first
> # column) filled, respectively, by the numbers 2,
> # 4, 6, 8, 10
> # One can now edit this dataset by, say, changing
> # all the
> # entries to 2, then closing the Data Editor window,
> # and
> # returning to the R console window:
> x
[1] 2 2 2 2 2 # x is indeed changed!
> # Thus one can change the entries for x via the
> # Data Editor,
> # and save the changes.
```

When using the functions `data.entry(x)` and `edit()` for data entry, there are a number of limitations (Aragon 2011):

- (i) Arrays and non-tabular lists cannot be entered using a spreadsheet editor.
- (ii) When using the function `edit()` to create a new data frame, one must assign an object name in order to save the data frame.
- (iii) This approach is not a preferred method of entering data because one often prefers to have the original data to be in a text editor or available to be read in from a data file.

2.4.1.8 Human Genetic Epidemiology Using R: The CRAN Package Genetics

To illustrate the ease of use of R in Human Genetic Epidemiology, consider examples in the CRAN package `Genetics`.

2.4.2 The Function `list()` and the Construction of `data.frame()` in R (Venables and Smith 2004; Dalgaard 2002; Teator 2011)

The Function `list()`

A list in R consists of an ordered collection of objects – its components, which may be of any modes or types. For examples, a list may consist of a matrix, a numeric vector, a complex vector, a logical value, a character array, a function, etc. Thus, some simple way to create a list would be:

Example 1 It is as easy as “1, 2, 3”!

```
> x <- 1
> y <- 2
> z <- 3
> list1 <- list(x, y, z) # Forming a simple list
> list1 # Outputting:
[[1]]
[1] 1

[[2]]
[1] 2

[[3]]
[1] 3
```

Moreover, the components are always numbered, and may be referred to as such.

Thus if `my.special.list` is the name of a list with 4 components, they may be referred to, individually, as:

```
my.special.list[[1]], my.special.list[[2]],
my.special.list[[3]], and my.special.list[[4]].
```

If one defines `my.special.list` as:

```
> my.special.list <- list(name="John", wife="Mary",
+   number.of.children=3, children.age=c(2, 4, 6))
```

then

```
> my.special.list[[1]] # Outputting:
[1] "John"
> my.special.list[[2]]
[1] "Mary"
> my.special.list[[3]]
[1] 3
> my.special.list[[4]]
[1] 2 4 6
```

The Number of Components in a List: the number of (top-level) components in a list may be found by the function `length()`. Thus:

```
> length(my.special.list)
[1] 4
```

viz., the list `my.special.list` has 4 components.

To combine a set of objects into a larger composite collection for more efficient processing, the list function may be used to construct a list from its components.

As an example, consider

```
> odds <- c(1, 3, 5, 7, 9, 11, 13, 15, 17, 19)
> evens <- c(2, 4, 6, 8, 10, 12, 14, 16, 18, 20)
> mylist <- list(before=odds, after=evens)
> mylist
$before
[1] 1 3 5 7 9 11 13 15 17 19
$after
[1] 2 4 6 8 10 12 14 16 18 20
> mylist$before
[1] 1 3 5 7 9 11 13 15 17 19
> mylist$after
[1] 2 4 6 8 10 12 14 16 18 20
```

Components of a List: Components of a list may be named. In such a case, the component may be referred to either

- (1) by giving the component name as a character string in place of the number in double square brackets, or
- (2) by giving an expression of the form

```
> name$component_name
for the same object.
```

Example 2: A family affair -

```
> my.special.list <- list(name="John", wife="Mary",
+   number.of.children=3, children.age=c(2, 4, 6))
> my.special.list # Outputting: $name
[1] "John"
$wife
[1] "Mary"
$number.of.children
[1] 3

$children.age
[1] 2 4 6
```

Thus, for this list:

```
my.special.list[[1]]
[1] "John"
> my.special.list$name
> # This is the same as my.special.list[[1]]
[1] "John"
> my.special.list[[2]]
[1] "Mary"
> my.special.list$wife
> # This is the same as my.special.list[[2]]
[1] "Mary"
```

```

> my.special.list[[2]]
[1] "Mary"
> my.special.list[[3]]
[1] 3
> my.special.list$number.of.children
> # This is the same as my.special.list[[3]]
[1] 3
> my.special.list[[4]]
[1] 2 4 6
> my.special.list$children.age
> # This is the same as my.special.list[[4]]
[1] 2 4 6

```

Extraction of a Variable from a List To extract the name of a component stored in another variable, one may use the names of the list components in double square brackets, viz., list1[["name"]]. The following R code segment may be used;

```

> x <- "name"; my.special.list[[John]]
[1] "John"

```

Constructing, Modifying, and Concatenating Lists:

New lists may be constructed from existing objects by the function list().

Thus, the form

```

> new.list <- list(name_1=object_1, ... name-
+                   n=object_n)

```

will set up a list, list1, of n components using object_1, ..., object_n for the components and giving them names as specified.

2.5 Univariate, Bivariate, and Multivariate Data Analysis

A univariate dataset has only one variable: {x}, e.g. {patient name}.

A bivariate dataset has two variables: {x1, x2}, or {x, y}, e.g. {patient name, gender}.

A multivariate dataset has more than two, or many, variables: {x1, x2, x3, ..., xn},

e.g. {case subject name, gender, age, weight, height, blood pressures, blood sugar level, heart rates, history of illnesses, etc.}

2.5.1 Univariate Data Analysis

As an example, enter the following code segments:

```
> x <- rexp(100); x
> # Outputting 100 exponentially-distributed
> # random numbers:
[1] 0.39136880 0.66948212 1.48543076 0.34692128 0.71533079 0.12897216
[7] 1.08455419 0.07858231 1.01995665 0.81232737 0.78253619 4.27512555
[13] 2.11839466 0.47024886 0.62351482 1.02834522 2.17253419 0.37622879
[19] 0.16456926 1.81590741 0.16007371 0.95078524 1.26048607 5.92621325
[25] 0.21727112 0.07086311 0.83858727 1.01375231 1.49042968 0.53331210
[31] 0.21069467 0.37559212 0.10733795 2.84094906 0.17899040 1.34612473
[37] 0.00290699 1.77078060 1.79505318 0.09763821 1.96568170 0.15911043
[43] 4.36726420 0.33652419 0.01196883 0.35657882 0.72797670 0.91958975
[49] 0.68777857 0.29100399 0.22553560 1.56909742 0.20617517 0.37169621
[55] 0.53173534 0.26034316 0.21965356 2.94355695 1.88392667 1.13933083
[61] 0.31663107 0.23899975 0.01544856 1.30674088 0.53674598 1.72018758
[67] 0.31035278 0.81074737 0.09104104 1.52426229 1.35520172 0.27969075
[73] 1.36320488 0.56317216 0.85022837 0.49031656 0.17158651 0.31015165
[79] 2.07315953 1.29566872 1.28955269 0.33487343 0.20902716 2.84732652
[85] 0.58873236 1.54868210 2.93994181 0.46520037 0.73687959 0.50062507
[91] 0.20275282 0.49697531 0.58578119 0.49747575 1.53430435 4.56340237
[97] 0.90547787 0.72972219 2.60686316 0.33908320
```

Note: The function `rexp()` is defined as follows:

`rexp(n, rate = 1)`

with arguments:

`x` vector

`n` number of observations. If `length(n) > 1`, the length is taken to be the number required.

The exponential distribution with rate λ has density:

$$f(x) = \lambda e^{-\lambda x}, \text{ for } x \geq 0.$$

If the rate λ is not specified, it assumes the default value of 1.

Remark

The function `rexp()` is one of the functions in R under Exponential in the CRAN package stats.

To undertake a biostatistical analysis of this set of univariate data, one may call up the function `univax()`, in the package `epibasix`, using the following code segments:

```
> library(epibasix)
> univar(x) # Outputting:
Univariate Summary
Sample Size: 100
Sample Mean: 1.005
Sample Median: 0.646
Sample Standard Deviation: 1.067
>
```

Thus, for this sample, size 100 elements, the mean, median and standard deviation have been computed.

For data analysis of univariate datasets, the R package *epibasix* may be used.

This CRAN(CRAN, n.d.) package covers many elementary financial functions for statistics and econometrics. It contains elementary tools for analysis of common financial problems, ranging from sample size estimation, through 2×2 contingency table analysis, and basic measures of agreement (kappa, sensitivity/specificity). Appropriate print and summary statements are also written to facilitate interpretation wherever possible. This work is appropriate for graduate financial engineering courses.

This package is a work in progress.

To start, enter the R environment and use the code segment:

```
> install.packages("epibasix")
```

Installing package(s) into 'C:/Users/bertchan/Documents/R/win-library/2.14'

(as 'lib' is unspecified)

--- Please select a CRAN mirror for use in this session ---

```
> # Select CA1
```

trying URL

'http://cran.cnr.Berkeley.edu/bin/windows/contrib/2.14/epibasix_1.1.zip'

Content type 'application/zip' length 57888 bytes (56 Kb)

opened URL

downloaded 56 Kb

package 'epibasix' successfully unpacked and MD5 sums checked

The downloaded packages are in

C:\Users\bertchan\AppData\Local\Temp\RtmpMFOrEn\downloaded_packages

With *epibasix* loaded into the R environment, to learn more about this package, follow these steps:

1. Go to the CRAN website: <http://cran.r-project.org>
2. Select (single-click) Packages, on the left-hand column
3. On the page: select E (for *epibasix*)

Available CRAN Packages By Name

[A](#) [B](#) [C](#) [D](#) [E](#) [F](#) [G](#) [H](#) [I](#) [J](#) [K](#) [L](#) [M](#) [N](#) [O](#) [P](#) [Q](#) [R](#) [S](#) [T](#) [U](#) [V](#) [W](#) [X](#) [Y](#) [Z](#)

4. Scroll down list of packages whose name starts with "E" or "e", and select: *epibasix*
5. When the *epibasix* page opens up, select: Reference manual: *epibasix.pdf*
6. The information is now on displayed, as follows:

Package 'epibasix'

January 2, 2012

Version 1.1

Date 2009-05-13

Author Michael A Rotondi <mrotondi@uwo.ca>

Maintainer Michael A Rotondi mrotondi@uwo.ca

Depends R (>= 2.01)

For another example, consider the same analysis on the first one hundred Natural Numbers, using the following R code segments:

```

> x <-1:100; x # Consider, and then output, the first 100
> # Natural Numbers
[1] 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18
[19] 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36
[37] 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54
[55] 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72
[73] 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88 89 90
[91] 91 92 93 94 95 96 97 98 99 100
> # ANOVA Tables: Summarized in the following tables,
> # ANOVA is used for
> # two different purposes:
> library(epibasix)
> univar(x) # Performing a univariate data analysis
> # on the vector x, and Outputting:

```

Univariate Summary

Sample Size: 100

Sample Mean: 50.5

Sample Median: 50.5

Sample Standard Deviation: 29.011

And that's it!

2.5.2 Bivariate and Multivariate Data Analysis (Daniel 2005)

When there are two variables, (X, Y), one need to consider two cases:

Case I: In the classical regression model, only Y called the dependent variable, is required to be random. X is defined as a fixed, non-random, variable and is the independent variable. Under this model, observations are obtained by pre-selecting values of X and determining the corresponding value of Y.

Case II: If both X and Y are random variables, it is called the Correlation Model – under which sample observations are obtained by selecting a random sample of the units of association (such as persons, characteristics (age, gender, locations, points of time, specific events/actions/..., or elements on which the two measurements are based) and by recording a measurement of X and of Y. In this case, values of X are not preselected but occurring at random, depending on the unit of association selected in the sample.

Regression Analysis:

Case I - Correlation analysis cannot be meaningfully under this model.

Case II – Regression analysis can be performed under the Correlation Model.

Correlation for 2 variables implies a co-relationship between the variables, and does not distinguish between them as to which is the dependent or the independent variable. Thus one may fit a straight line to the data either by minimizing $\sum(x_i - \bar{x})^2$ or by minimizing $\sum(y_i - \bar{y})^2$. The fitted regression line will, in general, be different in the two cases – and a logical question arises as to which line to fit.

Two situation do exit, and should be considered:

1. If the objective is to obtain a measure of strength of the relationship between the 2 variables, it does not matter which line is fitted – the measure calculated will be the same in either case.
2. If one need to use the equation describing the relationship between the 2 variables for the dependency of one upon the other, it does matter which line is to be fitted, The variable for which one wishes to estimate means or to make predictions should be treated as the depending variable. That is, this variable should be regressed with respect on the other variable.

Available R Packages for Bivariate Data Analysis:

Among the R packages for bivariate data analysis, a notable one available for sample size calculations for bivariate random intercept regression model there is the bivariate power.

An Example in Bivariate Data Analysis

As an example, this package may be used to calculate necessary sample size to achieve 80 percent power at 5 percent alpha level for null and alternative hypotheses that correlation between RI is 0 and 0.2, respectively, across 6 time points. Other covariance parameter are set as follows:

Correlation between residuals = 0;

Standard deviations: 1st RI = 1, 2nd RI = 2, 1st residual = 0.5, 2nd residual = 0.75

The following R code segment may be used:

```
> library(bivarRIpower)
> bivarcalcn(power=.80,powerfor='RI',timepts=6,
+ d1=1,d2=2, p=0,p1=.2,s1=.5,s2=.75,r=0,r1=.1)
# Outputting:
```

Variance parameters

Clusters	= 209.2
Repeated measurements	= 6
Standard deviations	
1st random intercept	= 1
2nd random intercept	= 2
1st residual term	= 0.5
2nd residual term	= 0.75
Correlations RI under H_o	= 0
RI under H_a	= 0.2
Residual under H_o	= 0
Residual under H_a	= 0.1
Con obs under H_o	= 0
Con obs under H_a	= 0.1831984
Lag obs under H_o	= 0
Lag obs under H_a	= 0.1674957

Correlation variances under H_o

Random intercept	= 0.005096138
Residual	= 0.0009558759
Concurrent observations	= 0.00358999
Lagged observations	= 0.003574277

Power (%) for correlations

Random intercept	= 80%
Residual	= 89.9%
Concurrent observations	= 86.4%
Lagged observations	= 80%

>

Bivariate Normal Distribution – Under the correlation model, the bivariates X and Y vary together in a joint distribution, which, if this joint distribution is a normal distribution, it is called a bivariate normal distribution, from which inferences may be made based on the results of sampling properly from the population. If the joint distribution is known to be non-normal, or if the form is unknown, inferential procedures are invalid. The following assumptions must hold for inferences about the population to be valid when sampling from a bivariate distribution:

- (i) For each value of X, there is a normally distributed sub-population of Y values.
- (ii) For each value of Y, there is a normally distributed sub-population of X values.
- (iii) The joint distribution of X and Y is a normal distribution called the Bivariate Normal Distribution.
- (iv) The sub-population of Y values all have the same variance.
- (v) The sub-population of X values all have the same variance.

Two random variables X and Y are said to be jointly normal if they can be expressed in the form

$$X = aU + bV \quad (2.1)$$

$$Y = cU + dV \quad (2.2)$$

where U and V are independent normal random variables.

If X and Y are jointly normal, then any linear combination

$$Z = s_1X + s_2Y \quad (2.3)$$

has a normal distribution. The reason is that if one has $X = aU + bV$ and $Y = cU + dV$ for some independent normal random variables U and V , then

$$\begin{aligned} Z &= s_1(aU + bV) + s_2(cU + dV) \\ &= (as_1 + cs_2)U + (bs_1 + ds_2)V \end{aligned} \quad (2.4)$$

Thus, Z is the sum of the independent normal random variables $(as_1 + cs_2)U$ and $(bs_1 + ds_2)V$, and is therefore normal.

A very important property of jointly normal random variables is that zero correlation implies independence.

Zero Correlation Implies Independence: If two random variables X and Y are jointly normal and are uncorrelated, then they are independent.

(This property can be verified using multivariate transforms,)

Multivariate Data Analysis(Daniel 2005)

Two similar, but distinct, approaches used for multivariate data analysis are:

1. The Multiple Linear Regression Analysis – assuming that a linear relationship exists between some variable Y , call the dependent variable, and n independent variables, $X_1, X_2, X_3, \dots, X_n$, which are called explanatory or predictor variables because of their use.

The assumptions underlying Multiple Regression Model Analysis are:

- (a) The X_i are non-random fixed variables, indicating that any

inferences drawn from sample data apply only to the set of X values observed, but not to larger collections of X . Under this regression model, correlation analysis is not meaningful.

- (b) For each set of X_i values, there is a sub-population of Y

values. Usually, one assumes that these Y values are normally distributed.

- (c) The variances of Y are all equal.

- (d) The Y values are independent of the different selected set of X values.

For multiple linear regression, the model equation is:

$$y_j = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + \beta_3 x_{3j} + \dots + \beta_n x_{nj} + e_j \quad (2.5)$$

where y_j is a typical value from one of the sub-populations of Y values, and the β_i values are the regression coefficients.

$x_{1j}, x_{2j}, x_{3j}, \dots, x_{nj}$ are, respectively, particular values of the independent variables $X_1, X_2, X_3, \dots, X_n$, and e_j is a random variable with mean 0 and variance σ^2 , the common variance of the sub-population of Y values. Generally, e_j is assumed normal and independently distributed.

When Eq. (2.1) consists of one dependent variable and two independent variables, the model becomes:

$$y_j = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + e_j \quad (2.6)$$

A plane in 3-dimensional space may be fitted to the data points. For models containing more than 2 variables, it is a hyperplane.

Let $f_j = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j}$

Now, $e_j = y_j - f_j$, forming a vector e .

If \underline{y} is the mean of the observed data,

viz., $\underline{y} = (1/n) \sum y_i$, for $i = 1, 2, 3, \dots, n$ then the variability of the dataset may be measured using three sums of squares (proportional to the variance of the data):

- (1) The Total Sum of Squares (proportional to the variance of the data):

$$SS_{\text{total}} = \sum (y_i - \underline{y})^2 \quad (2.7)$$

(2) The Regression Sum of Squares:

$$SS_{\text{reg}} = \sum (f_i - \bar{y})^2 \quad (2.8)$$

(3) The Sum of Squares of Residuals:

$$SS_{\text{res}} = \sum (y_i - f_i)^2 = \sum e_i^2 \quad (2.9)$$

The most general definition of the Coefficient of Multiple Determination is

$$R^2_{y,12\dots n} \equiv 1 - (SS_{\text{res}}/SS_{\text{total}}) \quad (2.10)$$

The parameter of interest in this model is the Coefficient of Multiple Determination, $R^2_{y,12,3,\dots,n}$, obtained by dividing the explained sum of squares by the total sum of squares:

$$R^2_{y,12\dots n} = \frac{\sum (y_i - \bar{y})^2}{\sum (y_i - \bar{y})^2} = \frac{SSR}{SSE} \quad (2.11)$$

where:

$\sum (y_i - \bar{y})^2$	= the Explained Variation
	= the original observed values from the calculated Y values
	= the sum of squared deviation of the calculated values from the mean of the observed Y values, or
	= the Sum of Squares due to Regression (SSR)
$\sum (y_i - \bar{y})^2$	= the Unexplained Variation
	= the sum of squared deviations of the original observations from the calculated values
	= the sum of squares about regression, or
	= the Error Sum of Squares (SSE)

The total variation is the sum of squared deviations of each observation of Y from the mean of the observations:

$$\sum (y_j - \bar{y})^2 = \sum (y_i - \bar{y})^2 + \sum (y_i - \bar{y})^2 \quad (2.12)$$

viz.,

$$SST = SSR + SSE \quad (2.13)$$

or, Total sum of squares = Explained (regression) sum.of squares

$$+ Unexplained (error) sum squares \quad (2.14)$$

The Multiple Correlation Model Analysis – the object of this approach is to gain insight into the strength of the relationship between variables.

The Multiple Regression Correlation Model Analysis Equation is:

$$y_j = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + \beta_3 x_{3j} + \dots + \beta_n x_{nj} + e_j \quad (2.15)$$

where y_j is a typical value from one of the sub-populations of Y values, the β_i are the regression coefficients, $x_{1j}, x_{2j}, x_{3j}, \dots, x_{nj}$ are, respectively, particular known values of the random variables $X_1, X_2, X_3, \dots, X_n$, and e_j is a random variable with mean 0 and variance σ^2 , the common variance of the sub-population of Y values. Generally, e_j is assumed normal and independently distributed.

This model is similar to Model Equation (2.5), with one important distinction:

in Eq. (2.5), the x_i are non-random variables, but

in Eq. (2.9), the x_i are random variables.

That is, in the Correlation Model, Eq. (2.9), there is a joint distribution of Y and the X_i that is called a Multivariate Distribution.

Under this model, the variables are no longer considered as being dependent or independent, because logically they are interchangeable, and either of the X_i may play the role of Y .

2. The Correlation Model Analysis

The Multiple Correlation Coefficient: To analyze the relationships among the variables, consider the multiple correlation coefficient, which is the square root of the coefficient of multiple determination and hence the sample value may be computed by taking the square root of Eq. (2.12), viz.,

$$\begin{aligned} R^2 y, 12 \dots n &= \sqrt{R^2 y, 12 \dots n} = \sqrt{\left\{ \sum (y_i - \bar{y})^2 / \sum (y_i - \bar{y})^2 \right\}} \\ &= \sqrt{SSR/SSE} \end{aligned} \quad (2.16)$$

3. Analysis of Variance (ANOVA) (http://en.wikipedia.org/wiki/Analysis_of_variance)

In **statistics**, ANalysis Of VAriance (ANOVA) is a collection of **statistical models** in which the observed **variance** in a particular variable is partitioned into components from different sources of variation. ANOVA provides a **statistical test** of whether or not the **means** of several groups are all equal, and therefore generalizes **t-test** to more than two groups. Doing multiple two-sample t-tests would result in an increased chance of committing a **Type I Error**. For this reason, ANOVAs are useful in comparing two, three, or multiple means.

ANOVA Tables: Summarized in the following tables, ANOVA is used for two different purposes:

- (1) To estimate and test hypotheses for simple linear regression about population variances, and
- (2) To estimate and test hypotheses about population means.

ANOVA Table for testing hypotheses about simple linear regression

Source	DF	Sum of squares	Mean squares	F-value	P-value
Model	1	$\sum (y_i - \bar{y})^2 = SSM_{Model}$	$SSM = MSM$	$MSG/MSE = F_{1,n-2}$	$Pr(F > F_{1,n-2})$
Residual	$n - 2$	$\sum e_i^2 = SSR_{Residual}$	$SSR/(n - 2) = MSE$		
Total	$n - 1$	$\sum (y_i - \bar{y})^2 = SST_{Total}$	$SST/(n - 1) = MST$		

Residuals are often called errors since they are the part of the variation that the line could NOT explain, so

$MSR = MSE = \text{sum of squared residuals}/df = \sigma^2$
 = estimate for variance of the population regression line
 $SSTot/(n-1) = MSTOT = s_y^2 = \text{the total variance of the } y_s$
 $F = t^2$ for Simple Linear Regression.

The larger the F (the smaller the p-value) the more of y's variation the line explained so the less likely H_0 is true. One rejects a hypothesis when the p-value $< \alpha$.

$R^2 = \text{proportion of the total variation of } y \text{ explained by the regression line}$
 $= SSM/SST$
 $= 1 - SSResidual/SST$

ANOVA table for testing hypotheses about population means

Source	df	Sum of squares	Mean squares	F value	P- value	p-value
Group (between)	$k-1$	$\sum n_i (\bar{x}_i - \bar{x})^2 = SSG$	$SSG/(k-1) = MSG$	$MSG/MSE = F_{k-1, N-k}$	$Pr = (F > F_{k-1, N-k})$	$Pr(F > F_{k-1, N-k})$
Error (within)	$N-k$	$\sum (n_i - 1)s_i^2 = SSE$	$SSE/(N-k) = MSE$			
Total	$N-1$	$\sum (x_{ij} - \bar{x})^2 = SSTot$	$SSTot/(N-1) = MST$			

$N = \text{total number of observations} = \sum n_i$, where $n_i = \text{number of observations for group } i$

The F test statistic has two different degrees of freedom: the numerator = $k-1$, and the denominator = $N - k \cdot F_{k-1, N-k}$

$$\begin{aligned}
 \text{NOTE : } SSE/(N-k) &= MSE = s^2 \\
 &= (\text{pooled sample variance}) \\
 &= \frac{(n_1 - 1)s_1^2 + \dots + (n_k - 1)s_k^2}{(n - 1) + \dots + (n_k - 1)} \\
 &= \hat{\sigma}^2 \\
 &= \text{estimate for assumed equal variance}
 \end{aligned}$$

(This is the 'average' variance for each group.)

$SSTot/(N-1) = MSTOT = s^2 = \text{the total variance of the data}$
 (assuming NO groups)

$F \approx \text{variance of the (between) samples means divided by the average variance of the data}$, the larger the F (the smaller the p-value) the more varied the means are so the less likely H_0 is true. It is rejected when the p-value $< \alpha$.

$R^2 = \text{proportion of the total variation explained by the difference in means} = \frac{SSG}{SSTot}$

Example 1

Package 'fAssets'

February 19, 2015

Title Rmetrics – Analyzing and Modelling Financial Assets

Date 2014-10-30

Version 3011.83

Author Rmetrics Core Team,

Diethelm Wuertz [aut],

Tobias Setz [cre],

Yohan Chalabi [ctb]

Maintainer Tobias Setz <tobias.setz@rmetrics.org>

Description Environment for teaching

``Financial Engineering and Computational Finance".

Depends R (>= 2.15.1), timeDate, timeSeries, fBasics

Imports fMultivar, robustbase, MASS, sn, ecodist, mvnormtest, energy

Suggests methods, mnormt, RUnit

Note SEVERAL PARTS ARE STILL PRELIMINARY AND MAY BE CHANGED IN THE FUTURE. THIS TYPICALLY INCLUDES FUNCTION AND ARGUMENT NAMES, AS WELL AS DEFAULTS FOR ARGUMENTS AND RETURN VALUES.

LazyData yes

License GPL (>= 2)

URL <https://www.rmetrics.org>

NeedsCompilation no

Repository CRAN

Date/Publication 2014-10-30 13:38:28

R topics documented:

fAssets-package	2
assets-arrange	6
assets-distance	7
assets-lpm	9
assets-meancov	10
1	
2 fAssets-package	
assets-modeling	12
assets-outliers	14
assets-selection	15
assets-testing	16

>

> `install.packages("fAssets")`

Installing package into 'C:/Users/Bert/Documents/R/win-library/3.2'

> `library(fAssets)`

<https://www.rmetrics.org> --- Mail to: info@rmetrics.org

Warning messages:

1. package 'fAssets' was built under R version 3.2.5
2. package 'timeDate' was built under R version 3.2.5
3. package 'timeSeries' was built under R version 3.2.5
4. package 'fBasics' was built under R version 3.2.5

```
> ls("package:fAssets")
[1] "abcArrange"                      "assetsArrange"
[3] "assetsBasicStatsPlot"             "assetsBoxPercentilePlot"
[5] "assetsBoxPlot"                    "assetsBoxStatsPlot"
[7] "assetsCoreEigenPlot"              "assetsCorgramPlot"
[9] "assetsCorImagePlot"               "assetsCorTestPlot"
[11] "assetsCumulatedPlot"             "assetsDendrogramPlot"
[13] "assetsDist"                      "assetsFit"
[15] "assetsHistPairsPlot"              "assetsHistPlot"
[17] "assetsLogDensityPlot"             "assetsLPM"
[19] "assetsMeanCov"                   "assetsMomentsPlot"
[21] "assetsNIGFitPlot"                 "assetsNIGShapeTrianglePlot"
[23] "assetsOutliers"                  "assetsPairsPlot"
[25] "assetsQQNormPlot"                 "assetsReturnPlot"
[27] "assetsRiskReturnPlot"              "assetsSelect"
[29] "assetsSeriesPlot"                 "assetsSim"
[31] "assetsSLPM"                      "assetsStarsPlot"
[33] "assetsTest"                      "assetsTreePlot"
[35] "binaryDist"                      "braycurtisDist"
[37] "canberraDist"                    "corDist"
[39] "covEllipsesPlot"                 "euclideanDist"
[41] "getCenterRob"                   "getCovRob"
[43] "hclustArrange"                  "jaccardDist"
[45] "kendallDist"                    "mahalanobisDist"
[47] "manhattanDist"                  "maximumDist"
[49] "minkowskiDist"                  "mutinfoDist"
[51] "mvenergyTest"                   "mvshapiroTest"
[53] "orderArrange"                   "pcaArrange"
[55] "sampleArrange"                  "sorensenDist"
[57] "spearmanDist"                   "statsArrange"
```

assets-selection Selecting Assets from Multivariate Asset Sets

Description

Select assets from Multivariate Asset Sets based on clustering.

Usage

```
assetsSelect(x, method = c("hclust", "kmeans"), control = NULL, ...)
```

Arguments

x any rectangular time series object which can be converted by the function `as.matrix()` into a matrix object, e.g. like an object of class `timeSeries`, `data.frame`, or `mts`.

method a character string, which clustering method should be used? Either `hclust` for hierarchical clustering of dissimilarities, or `kmeans` for k-means clustering.

control a character string with two entries controlling the parameters used in the underlying cluster algorithms. If set to `NULL`, then default settings are taken: For hierarchical clustering this is `method=c(measure="euclidean", method="complete")`,

and for `kmeans` clustering this is `method=c(centers=3, algorithm="Hartigan-Wong")`.

... optional arguments to be passed. Note, for the k-means algorithm the number of centers has to be specified!

Details

The function `assetsSelect` calls the functions `hclust` or `kmeans` from R's "stats" package. `hclust` performs a hierarchical cluster analysis on the set of dissimilarities `hclust(dist(t(x)))` and `kmeans` performs a k-means clustering on the data matrix itself.

Note, the hierarchical clustering method has in addition a `plot` method.

Value

if `use="hclust"` was selected then the function returns a S3 object of class "hclust", otherwise if `use="kmeans"` was selected then the function returns an object of class "kmeans".

For details we refer to the help pages of `hclust` and `kmeans`.

Author(s)

Diethelm Wuertz for the Rmetrics port.

References

Wuertz, D., Chalabi, Y., ChenW., Ellis A. (2009); Portfolio Optimization with R/Rmetrics, Rmetrics

eBook, Rmetrics Association and Finance Online, Zurich.

16 assets-testing

Examples

`## LPP -`

`# Load Swiss Pension Fund Data:`

`LPP <- LPP2005REC`

`colnames(LPP)`

`## assetsSelect -`

`# Hierarchical Clustering:`

`hclust <- assetsSelect(LPP, "hclust")`

`plot(hclust)`

`## assetsSelect -`

`# kmeans Clustering:`

`assetsSelect(LPP, "kmeans", control =`

`c(centers = 3, algorithm = "Hartigan-Wong"))`

`assets-`

```
> # LPP -
> # Load Swiss Pension Fund Data:
> LPP <- LPP2005REC
> colnames(LPP)
[1] "SBI"     "SPI"     "SII"     "LMI"     "MPI"
      "ALT"     "LPP25"   "LPP40"   "LPP60"
>
> ## assetsSelect -
> # Hierarchical Clustering:
> hclust <- assetsSelect(LPP, "hclust")
> plot(hclust)
> # Outputting: Fig. 2.18
```

```
>
> # assetsSelect -
> # kmeans Clustering:
> assetsSelect(LPP, "kmeans", control =
+ c(centers = 3, algorithm = "Hartigan-Wong"))
```

K-means clustering with 3 clusters of sizes 3, 1, 5

Cluster means:

	2005-11-01	2005-11-02	2005-11-03	2005-11-04	2005-11-07
1	-7.174567e-05	-0.0004228193	0.0070806633	0.0086669900	0.003481978
2	8.414595e-03	0.0025193420	0.0127072920	-0.0007027570	0.006205226
3	-9.685162e-04	-0.0021474822	-0.0004991982	-0.0005400452	0.000752058
	2005-11-08	2005-11-09	2005-11-10	2005-11-11	2005-11-14
1	0.0006761683	0.001840756	0.001091617	0.008291539	-0.0006469167
2	0.0003292600	-0.002378200	0.000922087	0.013334906	-0.0046930640
3	0.0014039486	-0.001694911	0.001119758	0.001929053	-0.0003708446
	2005-11-15	2005-11-16	2005-11-17	2005-11-18	2005-11-21
1	-0.0005408223	0.001686769	0.0048643273	0.005051224	0.001901881
2	0.0012668650	-0.007187498	0.0076581030	0.012527202	0.002659666
3	-0.0008278810	0.001879324	0.0003510514	-0.001005688	0.001511155
	2005-11-22	2005-11-23	2005-11-24	2005-11-25	2005-11-28
1	0.0037938340	0.0024438233	-0.000394091	0.0022790387	-0.006316902
2	0.0021424940	0.0035671280	-0.002559544	0.0033748180	-0.009816739
3	0.0003204556	0.0008044326	0.001046522	0.0007771762	-0.001717004
	2005-11-29	2005-11-30	2005-12-01	2005-12-02	2005-12-05
1	0.0031941750	-0.0037985373	0.011412029	0.0030362450	-0.003450339
2	0.0018864440	-0.0040370550	0.015977559	0.0070552900	-0.000268694
3	0.0001304264	-0.0001264086	0.002280999	0.0004551984	-0.001906796
	2005-12-06	2005-12-07	2005-12-08	2005-12-09	2005-12-12
1	-0.001325485	0.0007240550	-0.0066049723	0.0031775860	-0.001708432
2	0.002864672	-0.0026064460	-0.0034829450	0.0007472420	-0.000278182
3	0.002119992	-0.0007569246	0.0008520178	0.0006191402	0.000879002
	2005-12-13	2005-12-14	2005-12-15	2005-12-16	2005-12-19
1	0.0046313440	-0.004764346	0.0029376953	0.001213580	0.0004901883
2	0.0013586830	-0.007283576	-0.0073441080	0.004292528	0.0049806720
3	0.0008959768	-0.001095387	0.0007868516	0.001616794	0.0023614496
	2005-12-20	2005-12-21	2005-12-22	2005-12-23	2005-12-26
1	0.0071593677	0.007886516	0.0001159380	0.001659755	0.001861593
2	-0.0016334200	0.004319502	-0.0036299400	-0.002234004	0.000000000
3	0.0009404428	0.002617884	-0.0002855498	0.000507245	0.000246350
	2005-12-27	2005-12-28	2005-12-29	2005-12-30	2006-01-02
1	-0.001653545	-0.0005308263	0.004244265	-0.0020539910	0.0007404967
2	0.006571546	0.0016811710	0.007701353	-0.0039764730	0.0000000000
3	0.001191907	0.0015647450	0.001259883	-0.0007382012	0.0000960024
	2006-01-03	2006-01-04	2006-01-05	2006-01-06	2006-01-09
1	-0.0013538470	0.001985783	-0.001291249	0.0015069157	0.006899025
2	0.0065956320	0.012163609	-0.002285923	0.0022278130	-0.000174350
3	-0.0001111704	0.001657156	-0.000790521	0.0005815596	0.000574890

	2006-01-10	2006-01-11	2006-01-12	2006-01-13	2006-01-16	
1	-0.001533005	0.0040728800	0.004428087	-0.001791852	-0.0005044543	
2	-0.006484459	0.0109940950	0.005887780	0.000629145	0.0041673760	
3	-0.000713456	-0.0008200066	0.001761519	-0.002024939	-0.0000711238	
	2006-01-17	2006-01-18	2006-01-19	2006-01-20	2006-01-23	
1	-0.0064264427	-0.009959086	0.008732922	-0.005498331	-0.012300008	
2	-0.0060610800	-0.007104074	0.004901245	-0.008712876	0.000080500	
3	-0.0004887026	-0.000363750	0.001841324	-0.002770640	-0.000594795	
	2006-01-24	2006-01-25	2006-01-26	2006-01-27	2006-01-30	2006-01-31
1	0.0043293403	0.003928724	0.007596804	0.013018524	0.004510001	-0.000363935
2	-0.0009696750	-0.001245166	0.008348556	0.004847171	0.001426750	0.002736095
3	-0.0007514866	-0.001468543	0.001332284	0.002198388	0.000585831	0.000676559
	2006-02-01	2006-02-02	2006-02-03	2006-02-06	2006-02-07	
1	0.0025679990	-0.0033918090	0.001392298	0.003829826	-0.0016356553	
2	0.0033390250	-0.0052315370	0.006229563	0.000010100	-0.0003661590	
3	-0.0001937426	-0.0009799974	0.001561223	0.000789857	-0.0005261388	
	2006-02-08	2006-02-09	2006-02-10	2006-02-13	2006-02-14	
1	-0.0034102587	0.004724740	-0.0002670790	-0.0043036747	0.0071958650	
2	-0.0007797950	0.005323768	-0.0010290070	0.0030521410	-0.0005124970	
3	0.0002705448	0.001218430	-0.0001116748	-0.0002136692	0.0009441608	
	2006-02-15	2006-02-16	2006-02-17	2006-02-20	2006-02-21	
1	-0.0012863397	0.0045143087	-0.0008902317	-0.002237808	0.0050509303	
2	-0.0099343840	0.0129171990	0.0012844190	0.004250758	0.0063638350	
3	0.0002048986	0.0009410282	0.0012602358	0.001315374	0.0009478938	
	2006-02-22	2006-02-23	2006-02-24	2006-02-27	2006-02-28	
1	0.007771075	0.0008127150	0.0050935397	0.0050390380	-0.0108695373	
2	0.004433658	-0.0060669620	-0.0027457400	0.0053722110	-0.0119569840	
3	0.002287663	0.0003148286	-0.0003851528	0.0009570004	-0.0006375348	
	2006-03-01	2006-03-02	2006-03-03	2006-03-06	2006-03-07	
1	0.0038673107	-0.003960535	-0.004273008	0.0005488833	-0.001580974	
2	0.0105047210	-0.004177747	0.001755361	0.0004602100	-0.007028455	
3	0.0001534274	-0.003340033	-0.001016466	0.0001020288	-0.001403624	
	2006-03-08	2006-03-09	2006-03-10	2006-03-13	2006-03-14	
1	-0.004631919	0.0054281930	0.0077513163	0.0026288817	-0.0011077267	
2	-0.004230567	0.0035344700	0.0128746620	0.0082543960	-0.0005305960	
3	-0.000989756	0.0001691144	-0.0002940006	0.0006133858	0.0006779516	
	2006-03-15	2006-03-16	2006-03-17	2006-03-20	2006-03-21	
1	0.0039076300	-0.002874349	0.0035407443	-0.0002971253	0.0017513113	
2	-0.0017679700	0.000666818	0.0043129660	0.0021624820	-0.0002753190	
3	0.0001087012	0.000484471	-0.0001434976	0.0006474828	-0.0008360246	

	2006-03-22	2006-03-23	2006-03-24	2006-03-27	2006-03-28
1	0.0034417023	0.0042167183	-0.0005082437	-0.002477860	-0.005123098
2	0.0013273100	-0.0043185900	0.0027725900	-0.006945127	-0.000727865
3	0.0006235516	0.0002566278	0.0001903118	-0.000559792	-0.001748221
	2006-03-29	2006-03-30	2006-03-31	2006-04-03	2006-04-04
1	0.0083719110	-0.0023536467	0.0032189740	0.002929918	-0.0053531483
2	0.0009252440	0.0062064890	-0.0009292510	0.006944399	-0.0026184930
3	-0.0004704764	0.0003112066	-0.0003227242	-0.001220841	-0.0005425296
	2006-04-05	2006-04-06	2006-04-07	2006-04-10	2006-04-11
1	0.001672089	0.0035391920	0.0006688370	-0.0018577213	-0.0078612010
2	0.004752739	0.0015471340	0.0005119710	0.0000160000	-0.0095747430
3	0.000371034	0.0001308954	-0.0001606024	0.0004964256	0.0008645472
	2006-04-12	2006-04-13	2006-04-14	2006-04-17	2006-04-18
1	-0.001794527	-6.5767676e-05	0.0003451073	-0.01162293	0.0072958027
2	-0.002978612	3.927437e-03	0.0000000000	0.000000000	-0.0050312550
3	-0.001221078	-2.154967e-03	-0.0000179728	-0.00127895	0.0008801984
	2006-04-19	2006-04-20	2006-04-21	2006-04-24	2006-04-25
1	0.005262631	0.004346695	-0.0001275733	-0.0055535810	-0.0008195537
2	0.010893639	0.004201724	0.0048380810	-0.0016947400	-0.0041756790
3	0.001021676	0.001559365	-0.0000182248	-0.0001482958	-0.0017556356
	2006-04-26	2006-04-27	2006-04-28	2006-05-01	2006-05-02
1	0.001310372	-0.003174862	-0.0081970307	-0.0042204567	0.005113142
2	0.004699408	0.000883969	-0.0029437410	0.0000000000	0.003189327
3	-0.000656350	0.000053330	0.0002421188	-0.0007466388	0.001444051
	2006-05-03	2006-05-04	2006-05-05	2006-05-08	2006-05-09
1	-0.001263695	-0.0006502133	0.006613750	0.0063954017	-0.003357079
2	-0.011943802	0.0030146350	0.011235363	0.0061756300	0.002517020
3	-0.001037496	-0.0000032026	0.002150292	0.0007181488	-0.001401839
	2006-05-10	2006-05-11	2006-05-12	2006-05-15	2006-05-16
1	-0.0025826107	-0.008643966	-0.016678837	-0.006493649	-0.002878691
2	-0.0013872090	-0.001096609	-0.017989732	-0.013399941	0.003285150
3	-0.0005785598	-0.003397807	-0.004047129	-0.000274325	0.001044719
	2006-05-17	2006-05-18	2006-05-19	2006-05-22	2006-05-23
1	-0.010999777	-0.0085399823	0.005933202	-0.021077161	0.007398834
2	-0.028406916	-0.0097114170	0.001848007	-0.025997761	0.018970677
3	-0.003777801	0.0001129756	0.002052023	-0.001940486	0.002266990
	2006-05-24	2006-05-25	2006-05-26	2006-05-29	2006-05-30
1	-0.0004267907	0.0061746340	0.016697015	-0.0011805273	-0.017512962
2	-0.0111155890	0.0000000000	0.025842125	0.0017720070	-0.019842413
3	-0.0011767074	0.0006506848	0.002053482	0.0005608306	-0.003672446

	2006-05-31	2006-06-01	2006-06-02	2006-06-05	2006-06-06	
1	0.004247769	0.0048529913	0.002548541	-0.006215087	-0.007154035	
2	0.009323326	0.0015364830	0.006993059	0.000000000	-0.022326545	
3	0.001607127	-0.0008596914	0.002280503	-0.001082275	-0.001973220	
	2006-06-07	2006-06-08	2006-06-09	2006-06-12	2006-06-13	
1	-0.0015405703	-0.008894756	0.004967781	-0.004911457	-0.017072989	
2	0.0056383270	-0.027379310	0.012429170	-0.013895865	-0.023992295	
3	-0.0003462346	-0.001317298	0.002478040	-0.001395419	-0.003113317	
	2006-06-14	2006-06-15	2006-06-16	2006-06-19	2006-06-20	
1	-0.0008148080	0.0180758310	0.0023954587	0.001269003	-0.002988990	
2	0.0022722680	0.0215693860	-0.0029262460	0.008192221	0.003859327	
3	-0.0006585664	0.0004955194	-0.0000012182	-0.001387469	-0.001853536	
	2006-06-21	2006-06-22	2006-06-23	2006-06-26	2006-06-27	
1	0.0019448093	0.005720239	0.0003236553	0.0022263750	-0.004784092	
2	0.0033826850	0.007397751	0.0005459380	-0.0024934820	-0.006852493	
3	-0.0002515638	-0.000396063	-0.0009850436	-0.0002276698	-0.001709434	
	2006-06-28	2006-06-29	2006-06-30	2006-07-03	2006-07-04	2006-07-05
1	0.002418113	0.013946565	0.0007082933	0.0051894390	0.0036357180	-0.003018747
2	0.002764602	0.014041765	0.0144738140	0.0087848690	0.0019620600	-0.008807793
3	0.000155729	0.003295715	0.0007134052	0.0007825544	-0.0005997304	-0.002343919
	2006-07-06	2006-07-07	2006-07-10	2006-07-11	2006-07-12	
1	0.0006826483	-0.0053623837	0.006527871	-0.0026637643	-0.000460215	
2	0.0055633270	-0.0057979700	0.005982938	-0.0038346560	0.004470950	
3	0.0008409888	0.0001182382	0.001124756	0.0001879154	-0.000366662	
	2006-07-13	2006-07-14	2006-07-17	2006-07-18	2006-07-19	
1	-0.013003389	-0.0051940770	-0.0004672143	0.0002666863	0.010286439	
2	-0.015294320	-0.0103727210	-0.0022662700	-0.0056064500	0.018419580	
3	-0.002107831	0.0007863122	0.0008104944	-0.0013472198	0.001975698	
	2006-07-20	2006-07-21	2006-07-24	2006-07-25	2006-07-26	2006-07-27
1	0.001826095	-0.0080939067	0.013847674	0.0046651393	0.0005912200	-0.000995908
2	0.008023110	-0.0053387720	0.019239808	0.0006067870	0.0043343300	0.005560211
3	0.001140619	-0.0006294194	0.003391497	0.0009658202	0.0007748634	0.001358731
	2006-07-28	2006-07-31	2006-08-01	2006-08-02	2006-08-03	
1	0.0054530087	0.0009621367	-0.0016914860	0.001702776	3.388133e-05	
2	0.0102982490	0.0002672360	0.000000000	-0.003903872	-1.101953e-02	
3	0.0008377254	0.0012288024	-0.0002328868	-0.000336470	-1.946893e-03	
	2006-08-04	2006-08-07	2006-08-08	2006-08-09	2006-08-10	
1	0.0004074303	-0.005170347	0.001632508	0.0004397503	0.005330442	
2	0.0095373900	-0.012098329	0.000553560	0.0107776200	-0.005122863	
3	0.0025471316	-0.001348114	0.001391077	0.0003048744	0.000744234	

	2006-08-11	2006-08-14	2006-08-15	2006-08-16	2006-08-17	
1	0.0003234277	0.0042420490	0.006281944	0.002505853	0.0022312940	
2	0.0016952790	0.0080151810	0.014854679	0.004466055	0.0029039360	
3	-0.0003461608	-0.0005048628	0.002816991	0.001751071	0.0003821246	
	2006-08-18	2006-08-21	2006-08-22	2006-08-23	2006-08-24	
1	0.001322035	-0.0061511060	0.007819312	-0.0024164317	-0.0006088927	
2	-0.003261951	-0.0030139970	0.003136533	-0.0002148440	0.0020208950	
3	0.001275974	-0.0002004374	0.002229195	-0.0002547412	-0.0001443246	
	2006-08-25	2006-08-28	2006-08-29	2006-08-30	2006-08-31	
1	0.001879964	0.0001150750	0.0039964717	-0.0013533323	0.003702581	
2	0.000665318	0.0018015970	0.0060754350	0.0027585440	-0.001466397	
3	0.000606633	0.0008119902	-0.0000166692	0.0008165558	0.002074951	
	2006-09-01	2006-09-04	2006-09-05	2006-09-06	2006-09-07	
1	0.003825121	0.0022519797	0.002990688	-0.005576266	-0.0046239360	
2	0.003126659	0.0045617030	-0.000424630	-0.005904139	-0.0061051260	
3	0.001543151	-0.0000862626	0.000150264	-0.002215612	-0.0008681144	
	2006-09-08	2006-09-11	2006-09-12	2006-09-13	2006-09-14	2006-09-15
1	0.005100734	-0.006268633	0.0079822300	0.004665103	-0.0020439300	0.007261748
2	0.005038040	-0.008630266	0.0115443790	0.003047893	-0.0032925020	0.005426908
3	0.002150122	-0.002463850	0.0007386444	0.001504165	-0.0003697208	0.001212241
	2006-09-18	2006-09-19	2006-09-20	2006-09-21	2006-09-22	
1	-0.002313630	-0.0022077103	0.002879550	0.0017740240	-0.0120944313	
2	0.003786049	-0.0025686540	0.012243951	0.0039715930	-0.0091896050	
3	-0.001438723	0.0004212428	0.001333139	0.0009339414	0.0000184756	
	2006-09-25	2006-09-26	2006-09-27	2006-09-28	2006-09-29	2006-10-02
1	0.0015608587	0.007422705	0.0054981810	0.003993309	0.002248728	-0.004799167
2	-0.0027169680	0.012816618	0.0036014040	0.000706224	0.001409431	-0.005008288
3	0.0006968214	0.002279966	-0.0000253728	0.000284538	0.000532484	-0.001994171
	2006-10-03	2006-10-04	2006-10-05	2006-10-06	2006-10-09	
1	0.0011308367	0.007532294	0.007702274	0.0013383343	-1.842333e-06	
2	0.0006241600	0.006657881	0.007166048	0.0010145910	3.046625e-03	
3	-0.0000141456	0.001957197	0.001609076	-0.0009962644	2.434124e-04	
	2006-10-10	2006-10-11	2006-10-12	2006-10-13	2006-10-16	
1	0.0058490050	-0.001032232	0.005763076	0.0057859347	0.0016738733	
2	0.0089863210	0.001066884	0.004239901	-0.0020605700	-0.0018373360	
3	-0.0004574944	0.000480683	0.001357373	0.0004827066	-0.0004780952	
	2006-10-17	2006-10-18	2006-10-19	2006-10-20	2006-10-23	
1	-0.0072083463	0.0065625757	-0.004172731	0.0013755277	0.0067030360	
2	-0.0105018630	0.0091540220	0.001017744	0.0026080710	0.0055130300	
3	-0.0002134596	0.0004904236	-0.001158779	-0.0000334576	-0.0001258452	

	2006-10-24	2006-10-25	2006-10-26	2006-10-27	2006-10-30
1	0.0006909480	0.0007045297	0.0004360173	-0.006801505	-0.0013724143
2	-0.0035125790	0.0028470680	-0.0006172110	0.002649247	-0.0047902870
3	0.0001992902	0.0006721280	0.0005680150	0.000502184	0.0003469408
	2006-10-31	2006-11-01	2006-11-02	2006-11-03	2006-11-06
1	-0.001518782	1.051133e-05	-0.0013112907	0.0043942913	0.006986784
2	-0.008238610	4.875862e-03	0.0034278750	0.0059986060	0.010778959
3	0.001557833	1.065105e-03	-0.0000054624	-0.0000061736	0.002111045
	2006-11-07	2006-11-08	2006-11-09	2006-11-10	2006-11-13
1	-0.000492798	0.0010355810	-0.0042464430	-0.0022569473	0.0017332780
2	0.003869904	-0.0059257000	-0.0005821010	-0.0026741380	0.0015220670
3	0.002714624	0.0004917422	0.0007154908	-0.0007612248	-0.0003075366
	2006-11-14	2006-11-15	2006-11-16	2006-11-17	2006-11-20
1	0.005042735	0.00039623527	0.000571091	-0.0041334430	-0.0008046817
2	-0.001419646	0.0067483290	0.0000088600	-0.0046228550	0.0012852990
3	0.001633393	-0.0003903638	-0.000064044	-0.0004950332	0.0008012360
	2006-11-21	2006-11-22	2006-11-23	2006-11-24	2006-11-27
1	0.0035024650	-0.001423817	-0.0022464847	-0.0074566157	-0.008884057
2	0.0015966630	0.000878824	-0.0028429630	-0.0120525280	-0.014190477
3	0.0000896774	-0.000362714	-0.0009733886	-0.0007358458	-0.001985401
	2006-11-28	2006-11-29	2006-11-30	2006-12-01	2006-12-04
1	-0.0005720070	0.010011509	-0.0028207700	-0.0045097767	0.008035140
2	-0.0064998550	0.012901202	-0.0086468080	-0.0064254010	0.006552949
3	0.0004750374	0.001182122	0.0003041304	-0.0001392252	0.003841311
	2006-12-05	2006-12-06	2006-12-07	2006-12-08	2006-12-11
1	0.0006129080	0.0012128337	0.0014398583	-0.001522731	0.006351745
2	-0.0028113180	0.0071741430	0.0074404230	-0.002915097	0.006216417
3	-0.0002819442	-0.0002501772	-0.0002762668	0.000266215	0.000853780
	2006-12-12	2006-12-13	2006-12-14	2006-12-15	2006-12-18
1	0.0003822723	0.0032613870	0.0085999303	0.004852111	0.0007975687
2	0.0077836640	0.0016147920	0.0100852550	0.001294973	0.0045667980
3	0.0014112930	-0.0003111648	0.0006167068	0.002972998	-0.0018724352
	2006-12-19	2006-12-20	2006-12-21	2006-12-22	2006-12-25
1	-0.007259400	0.0039956217	-0.001713398	-0.001404090	-0.0004215033
2	-0.006190852	0.0015531190	0.000940550	-0.005018284	0.0000000000
3	-0.001259345	0.0002665752	0.002535426	-0.002011179	-0.0000406544
	2006-12-26	2006-12-27	2006-12-28	2006-12-29	2007-01-01
1	0.0026915907	0.009066652	0.0004164007	-0.001200784	-9.820667e-06
2	0.00000000000	0.010333820	-0.0020348850	-0.001101976	0.000000e+00
3	0.0004873476	0.002159424	-0.0010962306	-0.000159042	-1.140000e-05

	2007-01-02	2007-01-03	2007-01-04	2007-01-05	2007-01-08	
1	0.0016820867	0.004437189	-0.0002060123	-0.002683635	-0.001690284	
2	0.0000000000	0.014834908	0.0001279650	-0.002580941	-0.004541686	
3	0.0004246192	0.001879946	0.0006782908	-0.001927514	-0.001011925	
	2007-01-09	2007-01-10	2007-01-11	2007-01-12	2007-01-15	
1	0.0044103520	-0.0006727953	0.007703170	0.003633612	0.005514158	
2	0.0032808260	-0.0014696800	0.013481449	0.005427095	0.008135203	
3	0.0001659808	-0.0012173928	0.001547265	-0.000644134	0.001385737	
	2007-01-16	2007-01-17	2007-01-18	2007-01-19	2007-01-22	
1	-0.0002334037	-0.0009348027	0.0018915933	0.005485118	-0.001422125	
2	-0.0023385150	0.0036009040	-0.0000306000	0.004899725	-0.004463529	
3	0.0019940186	-0.00003052906	-0.0002513524	0.001870075	0.000533480	
	2007-01-23	2007-01-24	2007-01-25	2007-01-26	2007-01-29	
1	-0.0024795987	0.010078597	-0.005617319	0.0003207103	0.0026204570	
2	-0.0005070610	0.006027774	-0.001732155	-0.0095941300	0.0074250550	
3	-0.0000391516	0.001921089	-0.001420198	0.0000189644	0.0006710216	
	2007-01-30	2007-01-31	2007-02-01	2007-02-02	2007-02-05	2007-02-06
1	0.002703326	-0.0017867277	0.005665622	0.005261889	1.452767e-05	0.0016964537
2	0.004207812	-0.0000649000	0.009002711	0.004474549	2.320000e-05	0.0009123640
3	0.002512929	0.0001274554	0.001459012	0.002004385	1.534503e-03	-0.0001289984
	2007-02-07	2007-02-08	2007-02-09	2007-02-12	2007-02-13	
1	0.001088440	0.0005737837	-0.0000244410	-0.003474908	0.0039971257	
2	0.002937055	-0.0060861880	0.0065013820	-0.004102506	-0.0023852220	
3	-0.001036504	0.0008835816	-0.0007572888	-0.001322240	0.0002060698	
	2007-02-14	2007-02-15	2007-02-16	2007-02-19	2007-02-20	
1	0.003429143	0.0003378953	-0.0010496707	0.001237031	0.002481922	
2	0.006839015	0.0014878740	0.0030582220	0.000894606	-0.003270610	
3	0.001787848	0.0004864332	-0.0005123012	-0.002126213	0.001312554	
	2007-02-21	2007-02-22	2007-02-23	2007-02-26	2007-02-27	
1	-0.0003544737	0.004133925	-0.003687659	-0.0006432653	-0.028181932	
2	-0.0099202370	0.003907766	-0.000693145	-0.0035030230	-0.035746244	
3	0.0008831304	-0.0000198173	0.000224204	0.0003787182	-0.003557775	
	2007-02-28	2007-03-01	2007-03-02	2007-03-05	2007-03-06	2007-03-07
1	-0.006545974	-0.004681529	-0.007204097	-0.016965306	0.0131994283	0.001848668
2	-0.011946524	-0.001959064	0.002364749	-0.014462125	0.0116318040	0.011790712
3	-0.001276448	-0.000431924	-0.001183209	-0.000934475	0.0006908982	0.000013807
	2007-03-08	2007-03-09	2007-03-12	2007-03-13	2007-03-14	2007-03-15
1	0.01238125	0.0054817963	-0.0019347513	-0.011489525	-0.016127064	0.010698674
2	0.01192565	0.0003297920	-0.0031733170	-0.007777588	-0.028205491	0.014756610
3	0.00211282	0.0003630802	0.0003176836	-0.001476343	-0.003026142	0.003330527

```

2007-03-16 2007-03-19 2007-03-20 2007-03-21 2007-03-22 2007-03-23
1 -0.0040807663 0.011023104 0.006051645 0.008317470 0.006371441 0.003735611
2 0.0011294700 0.014488916 0.004157206 0.007430739 0.013969587 0.001698465
3 -0.0004977044 0.001539199 0.001811865 0.002705776 0.002193543 -0.000066116

2007-03-26 2007-03-27 2007-03-28 2007-03-29 2007-03-30
1 -0.0038874940 -0.0024984800 -0.007625166 0.008954610 0.0028986417
2 -0.0075840220 -0.0047538740 -0.009870398 0.011153959 0.0004970940
3 -0.0006071596 -0.0008749648 -0.001164874 0.001407261 -0.0002237774

2007-04-02 2007-04-03 2007-04-04 2007-04-05 2007-04-06
1 -0.0013403807 0.008656804 0.0031424653 -0.0014326240 -0.0004997497
2 -0.0014515930 0.010335207 0.0013071590 0.0050004920 0.0000000000
3 0.0000685862 0.002129822 0.0004029036 -0.0001666756 -0.0000500682

2007-04-09 2007-04-10 2007-04-11
1 0.006564997 0.0004269093 -0.001272110
2 0.000000000 0.0063294250 -0.001044170
3 0.000570819 -0.0000510372 -0.000411593

```

Clustering vector:

SBI	SPI	SII	LMI	MPI	ALT	LPP25	LPP40	LPP60
3	2	3	3	1	1	3	3	1

Within cluster sum of squares by cluster:

```

[1] 0.003806242 0.000000000 0.005432037
(between_SS / total_SS = 75.3 %)

```

Available components:

```

[1] "cluster"        "centers"        "totss"         "withinss"       "tot.withinss"
[6] "betweenss"      "size"          "iter"          "ifault"

```

>

+++++

Example 1 in Multivariate Data Analysis –

(see the illustrative example in 4.2.1)

ANOVA Analysis

ANOVA (Analysis of Variance) may be achieved using the R function `anova()`, which is in the CRAN package `stats`. The standard usage form of this function is

```
> anova(object, ...)
```

with arguments:

`object` ≡ an object for model fitting (e.g., `lm` or `glm`).
`...` ≡ additional objects of the same type.

Example 2 in Multivariate Data Analysis –

Appendix 1

Documentation for the plot function

plot {graphics}

R Documentation

Generic X-Y Plotting

Description

Generic function for plotting of R objects. For more details about the graphical parameter arguments, see [par](#).

For simple scatter plots, [plot.default](#) will be used. However, there are plot methods for many R objects, including [functions](#), [data.frames](#), [density](#) objects, etc. Use [methods\(plot\)](#) and the documentation for these.

Usage

`plot(x, y, ...)`

Arguments

- X the coordinates of points in the plot. Alternatively, a single plotting structure, function or any R object with a plot method can be provided.
- Y the y coordinates of points in the plot, optional if x is an appropriate structure.

Details

The two step types differ in their x-y preference: Going from (x1,y1) to (x2,y2) with $x_1 < x_2$, type = “s” moves first horizontal, then vertical, whereas type = “S” moves the other way around.

See Also

[plot.default](#), [plot.formula](#) and other methods; [points](#), [lines](#), [par](#).

For X-Y-Z plotting see [contour](#), [persp](#) and [image](#).

Examples

```
require(stats)
plot(cars)
lines(lowess(cars))
plot(sin, -pi, 2*pi) # see ?plot.function

## Discrete Distribution Plot:
plot(table(rpois(100,5)), type = "h", col = "red",
lwd=10, main="rpois(100,lambda=5)")

## Simple quantiles/ECDF, see ecdf\(\) {library(stats)} for a better one:
plot(x <- sort(rnorm(47)), type = "s", main = "plot(x, type = \"s\")")
points(x, cex = .5, col = "dark red")
```

Special References

ANOVA http://en.wikipedia.org/wiki/Analysis_of_variance

Aragon TJ (2011) Applied epidemiology using R (epir). UC Berkeley School of Public Health, and San Francisco Department of Public Health, Berkeley

BMI Notes (2012) Body mass index. http://en.wikipedia.org/wiki/Body_mass_index

Centers for Disease Control and Prevention (2005) Antiretroviral postexposure Prophylaxis after sexual, injection-drug use, or other nonoccupational exposure to HIV in the United States: recommendations from the U.S. Department of Health and Human Services. MMWR Recomm Rep 54(RR-2):1–20 Available from: <http://www.cdc.gov/mmwr/preview/mmwrhtml/rr5402a1.htm>

CRAN, The comprehensive R archive network: <http://cran.r-project.org/>

Dalgaard P (2002) Introductory statistics with R, Springer statistics and computing series, Springer, New York

Daniel WW (2005) Biostatistics – a foundation for analysis in the health sciences. Wiley, New York

Everitt BS, Hothorn T (2006) A handbook of statistical analysis using R. Chapman & Hall/CRC, Boca Raton

Statistician Job Search <http://jobs.amstat.org/jobs/4627784/biostatistician-1>

Statistics Canada <http://www.statcan.gc.ca/edu/power-pouvoir/ch3/5214785- eng.htm#a1>

Teeter P (2011) R Cookbook. O'Reilly Media, Sebastopol

Venables WN, Smith DM, and the R Development Core Team (2004) An introduction to R. Network Theory, Ltd., Bristol

Virasakdi C (n.d.) Analysis of epidemiological data using R and Epicalc. Epidemiology unit, Prince of Songkla University, Thailand: cvirasak@medicine.psu.ac.th



Abstract

This chapter considers the fundamental concepts in the theory of probability and applied statistics in epidemiology, including the biostatistical concepts and measures in genetic association and **familial aggregation** studies, including:

- Additional Approaches in **Familial Aggregation** Studies
- Twin Studies
- Adoption Studies
- Inbreeding Studies
- Randomization Test
- **Segregation** studies, **Linkage** studies, **Association** studies
- Genome-wide Association Studies (GWAS)
- Big Data and Human Genomics

Keywords

Theory of probability and applied statistics in epidemiology · Biostatistical concepts and measures in genetic association · Familial aggregation studies · Twin studies · Adoption studies · Inbreeding studies · Randomization test · Segregation studies · Linkage studies · Association studies · Genome-wide Association Studies (GWAS) · Big data and human genomics

3.1 Some Fundamental Concepts in the Theory of Probability and Applied Statistics in Epidemiology

A comprehensive presentation and discussion of the up-to-date fundamental theories in applied probability theory and of the theory of biostatistics, with illustrations of computations using the open-sourced computer program R, pertinent to the applications and the mathematical development of epidemiology and health sciences, have been recently presented in Chan (2015). In summary, the topics covered in that book included the following:

- Theories of Probability:
What is Probability?
Basic Properties of Probability
Probability Computations Using R
Applications of Probability Theory to Health Sciences
Typical Summary Statistics in Biostatistics: Confidence Intervals, Significance Tests, and Goodness of Fit
- Typical Statistical Inference in Biostatistics: Bayesian Biostatistics:
What is Bayesian Biostatistics?
Bayes's Theorem in Probability Theory
Bayesian Methodology and Survival Analysis (Time-to-Event) Models for Biostatistics in Epidemiology and Preventive Medicine
The Inverse Bayes Formulas
Modeling in Biostatistics

It is recommended that the reader should also access the on-line supplementary materials of that book, which may be found in the website:

www.springerpub.com/chan-biostatistics

and which, in particular, contains an additional chapter entitled “**Research-Level Applications of R**”.

3.2 Biostatistical Concepts and Measures in Genetic Association

The investigation of genetics in disease progresses through the following study designs, each answering a different question:

1. **Familial Aggregation studies:** Is there a genetic component to the disease, and what are the relative contributions of genes and environment?
2. **Segregation studies:** What is the **pattern of inheritance** of the disease (recessive or dominant)?
3. **Linkage studies:** On which part of which **chromosome** is the disease gene located.?
4. **Association studies:** Which **allele** of which gene is associated with the disease?

This traditional approach has proved highly successful in identifying monogenic disorders and locating the genes responsible.

Nowadays, the scope of genetic epidemiology has expanded to include common diseases for which many genes each make a smaller contribution (**polygenic, multifactorial or multi-genetic disorders**). This has developed rapidly in the first decade of the twenty-first century (2001–2010) following completion of the Human **Human Genome Project**, as advances in **genotyping** technology and associated reductions in cost has made it feasible to conduct large-scale genome-wide **association studies** that genotype many thousands of **single nucleotide polymorphisms** in thousands of individuals. **These have led to the discovery of many genetic polymorphisms that influence the risk of developing many common diseases.**

3.2.1 **Familial Aggregation Studies**

Here, the fundamental question to address is:

Is there a genetic component to the disease, and what are the relative contributions of genes and environment?

To begin with, one looks for evidence of clustering among close family members, as well as distant family members, viz., the evidence of clustering in families. In other words, one pursues a possible genetic etiology by demonstrating that the disease tends to run in families **more often than** would be by pure chance, and to note other relationships, and/or environmental factors. The investigations may be based on case control comparisons of families and their histories, or on twins (similar nature) or adoption (same nurture) studies.

The following progressive steps may be taken:

1. For any given disease: look for a potential genetic trait: to see if it tends to aggregate in families. This approach depends on the nature of the *phenotype*, viz., the observable outcome – in contradistinction with the *genotype*, viz., the unobserved genes.
2. If such a potential family is identified on account of an individual, called the *Proband*. Depending on the sampling scheme as well as the endpoint, the probands are NOT necessarily diseased case subjects even if a binary disease trait is being investigated. Thus, both unaffected and affected persons could be probands in a design of any case-control study.
3. In a typical investigation of a continuous trait, typically one might enroll a random series of probands from the general population, with their families, to examine the patterns of correlation in the phenotype among different types of relatives, such as parent-offsprings, siblings, other blood relatives, and so on.t
4. Using path analysis and/or variance components techniques, the resultant correlation matrices are examined to determine the proportion variance owing to genetic and shared environmental influences.
5. Moreover, it is possible to identify families through probands with increased values of the phenotype or some related correlations – such as
 - Heart diseases in a study of hypertension or levels of lipids
 - Melanoma in a study of nevi
 - Breast cancers in a study of benign breast diseases for breast cancers
 - Polyps in a study of colorectal cancers
 - etc.

For each case-subject, a structured familial history is first noted, noting both the familial cases as well as the family members at risk, including their times at risk and their ages. If at all possible, seek independent confirmations of the reported familial cases.

To analyze these data, two independent should be taken:

1. Using standard case-control methods, compare case-subjects and controls as independent individuals in terms of the history of the family as an exposure variable
2. Consider the family members od cases and of controls as two distinct cohorts, and apply standard person-year methods to compare their incidence of disease.

3.2.1.1 Additional Approaches in **Familial Aggregation Studies**

The following additional approaches may be used in familial aggregation studies:

3.2.1.2 Twin Studies

The aim of this approach is to separate the environmental effects (Nurture) from genetic effects (Nature). The ideal **Twin Methodology** consists of identifying pairs of biological twins through affected members, and contrasting the *Concordance Rates* of Monozygotic and Identical (MZ) twins and Dizygotic or Fraternal (DZ) twins. If it can be assumed that MZ and DZ twins share environmental factors to a comparably similar degree (Nurture) but differ in their genetic similarity (Nature), then this comparison may allow the estimation of **heritability** – that is, the proportion of the variance on the liability to diseases owing to common genetic factors, and the proportion owing a shared environment.

Variations of this twin design may include comparisons of biological twins reared apart and twin family studies – to shed some light on the relative contributions of Nature vs. Nurture! Recent data seem to show that, in the general population, Nurture plays a somewhat stronger role Nature, given that all other factors being the same or comparable!

3.2.1.3 Adoption Studies

Similar to epidemiologic investigations of biological twins reared separately, Adoption Studies is another approach to contrast the effects of nurture (environmental alone) and nature (genetic alone) by comparing case subjects who have different ancestry but share the same environment. For example, a review of the literature of breast cancer cases has yielded no adoption cases. However, in a study of the relatives of some 34 adoptees with schizophrenia and 34 control adoptees, it was found that:

- a 5% risk (14 cases/275 cases) appeared among their biological relatives
- a 0.4% risk (1 case/253 cases)
- No cases among the adoptive relatives of either group

3.2.1.4 Inbreeding Studies

Since the children of closely-related, or consanguineous, couples are likely to carry 2 copies of the same allele, and thus become at increased risk of diseases of recessive genes – those diseases for which 2 mutant alleles are necessary. Thus, inbreeding studies may be carried out as cohort studies or case-control studies – in either case, treating the exposure as proportional to the degree of inbreeding (such as Wright's Coefficient of Inbreeding W_{CI} (the probability of carrying 2 copies of the same allele from a common ancestor.

3.2.1.5 Randomization Tests

For large population-based series of families or case-control data, randomization tests may be used. The basic approach is to compare an index of familiality for the observed families with a distribution of hypothetical families of the same structure which have been obtained by repeat randomization of the assignment of individuals to families. For example, using this approach, it had been found that families of young onset of breast cancer case subjects had greater heterogeneity in risk than expected, viz., an excess of both high and low risk, whereas the distribution in control families was found to be consistently skewed toward low risk.

3.2.2 Segregation Studies

Segregation Studies address the question: “What is the **pattern of inheritance** of the disease: recessive or dominant?”

In epidemiology investigations of a genetic disease, **Segregation Studies** seek responses to the questions:

“What are the **patterns of inheritance** of the disease: recessive or dominant?”

The approach in these studies consists of the process of fitting statistical models to data on the phenotypes of the involved family members through affected probands of a dichotomous phenotype. This technique may be applied to variable age-at-onset disease traits or to continuous, dichotomous disease traits. The statistical approach includes the Method of Maximum Likelihood which seeks parameters that maximize the probability, or likelihood, of the fitted model by mathematically evaluating the values of these parameters that maximize the likelihood (in the probability sense) of the observed dataset.

Essentially, this maximum likelihood process considers the following elements of the model:

- (i) The **Penetrance Function**,
- (ii) The **Population Genotype Distribution** which, for a major Mendelian gene, may be a function of the allele frequency,
- (iii) The **Transmission Probabilities** within the families involved, and
- (iv) The **Method of Ascertainment**: since families ascertained through affected probands are likely to over-represent the gene frequency relative to the general population, an ascertainment correction is thereby necessary. Similarly, families with a large number of affected cases are more likely to be ascertained than those with few cases, which may result in the overestimation of penetrances.

As the underlying genotypes are generally unknown, the likelihood of the observed phenotypes may be calculated by summing over **all possible** combinations of genotypes that are compatible with the Mendelian laws and with the observed data.

It should not escape ones attention that the set of possible genotypes may very large: reaching 3^N for a single major gene with 2 alleles, where N is the number of family members – even if not every combination is, in fact, possible. Except for very small families, complete enumeration is not possible! On the other hand, there exist a remarkably efficient algorithm, called “PEELING”, for locating the set of legal genotypes which renders the computation possible: starting at the bottom of a pedigree and computing the probability of the genotypes of the parents of the case subject (given their phenotypes and the offsprings), and working up from there, at each step using the genotype probabilities that have been calculated at the lower levels of the pedigree.

An Example: Williams, W. R., and Anderson, D. E. (1984).- *General Epidemiology*, 1:7–20

“Genetic Epidemiology of Breast Cancer: Segregation Analysis of 200 Danish Pedigrees”

The familial occurrence of breast cancer is well known. It had been reported that, as far back as AD100, ancient Roman doctors were cognizant of pedigrees of cancer-prone families for over a century! Moreover, for the past 50 years, it has been noted that there have been increased risks originating from a family history of the disease. Notwithstanding this long historical record as well as a large number of studies on this subject, the etiologic significance of familial breast cancer is still being investigated as well as debated! It is also being contended that familial clustering reflects a chance phenomenon owing to the fact that breast cancer is common in the general public. Moreover, others contend that the phenomenon of such familial clustering reflects the consequence of close relatives

sharing common dietary practices and common customs, or exposure to common environmental hazards, while others contend that it is the result of an inherited susceptibility to the disease!

Early family investigations reported evidence of twofold to fourfold higher risks for the disease among groups of relatives of patients than among the control groups.

An investigation of the genetic epidemiology of breast cancer involving complex segregation analysis of 200 breast cancer pedigrees of Danish extraction is presented. The observed distribution of breast is compatible with transmission of an autosomal-dominant gene with no evidence for residual family resemblance. The gene frequency of the abnormal allele is 0.00756, and the displacement between homozygous genotype means is 1.695. The gene frequency accounts for a significant proportion of breast cancer in young women, whereas by an advanced age a majority (87%) of affected women are “phenocopies”. Genetic modeling of other breast cancer families and results of linkage studies are also reviewed.

3.2.3 **Linkage Studies**

Linkage Studies Address the Question: “On Which Part of Which [Chromosome](#) is the Disease Gene Located?”

Linkage analysis* maps genetic loci by the use of observations of related individuals. A usual introduction is to seek methods commonly used to map loci that pre-dispose to disease. Linkage analysis methods can be applied to both major gene disorders (parametric linkage) and complex diseases (model-free or non-parametric linkage). Evidence for linkage is most commonly expressed as a logarithm of the odds score. This study provides a framework for interpretation of these scores and discuss the role of simulation in assessment of statistical significance and estimation of power. Genetic and phenotypic heterogeneity can also affect the success of a study, and several methods exist to address such problems.

*[Teare, D. M.](#), and [Barrett, J. H.](#) (2005) – Lancet, 366(9490):1036-44, “Genetic Linkage Studies”

In linkage studies for genetic epidemiology, a common approach is to identify the approximate the approximate chromosomal location of a major gene – as a result of the recombination phenomenon based on the following guideline principles:

1. Genes on different chromosomes segregate independently: thus there can be no linkage between them, and
2. The probability θ of recombination between two loci on the same chromosome *increases* with the distance between them, reaching the limiting value of $\frac{1}{2}$, which is the same probability for two separate chromosomes.

Recombination rates may be expressed in units of 1%, called a centiMorgan (cM). Whenever a genetic marker is found to have a low recombination rate with a disease gene, it may be concluded that the disease gene is close to that marker. Thus, it is critical to determine the genotypes of various markers, whose locations are known, for various members of a multiple case family.

If both the disease gene alleles as well as the marker gene are diallelic, then

- denoting disease gene alleles as d and D , and
- denoting the marker alleles as m or M

then a doubly heterozygous case subject's genotype may be represented as

$$dm|DM \text{ or } dM|Dm,$$

distinguishing their phase.

In principle, one would look for parent-offspring pairs, or meiosis in which the parent is doubly heterozygous, and then count the number of occasions in which the transmitted haplotype is a recombinant or not.

The Direct Counting Method: On the basis of this count, the proportion of recombinants out of all meiosis is a direct estimate of the recombination fraction.

However, unless the diseased gene is fully penetrant with no phenocopies, one may not unambiguously ascertain the genotypes of the individuals. And even if that is definitively ascertained, one may not be able to determine the phase. Thus, in practice, linkage analysis uses one of two basic approaches:

Method I: The lod Score, or lods Method for Quantitative Traits

The **lod** score, or **lods** method is Likelihood Ratio (LR) Test calculated using the base 10 (instead of the Natural Logarithm base e , and without multiplying by 2), viz.,

$$\text{Lods} = \log_{10} [L(\underline{\theta})/L(\theta = 0.5)] = G^2/[2 \ln (10) = 4.6]$$

Here, the 5% significant level for a 1 df test corresponds to an LR test of 3.84, or a lods of 0.83. The lods value of 3, which is commonly used as a benchmark for an established linkage, corresponds to an LR chi-square of 13.82, or a p -value of 0.0002.

Method II: The Affected sib Pair Method for Quantitative Traits [Thomas 188]

For this approach, the working hypothesis is as follows:

Sib pairs that share, by chance, alleles that are **identical-by-descent** (IBD), viz., they are derived from the same parent or same grandparent, etc., tend to have greater similarity in their phenotypes than pairs that do not. Pairs that share two alleles IBD tend to have greater similarity. This hypothesis may be tested by regressing the squared differences in phenotypes on the number

In this approach, as in the affected sib Pair Method for Quantitative Traits [Thomas 188], it is not always possible to infer IBD sharing with any certainty. The Haseman-Elstrom Method (Haseman and Elston 1972) solves this problem by regressing the squared phenotype differences, defined as

$$D_i = (Y_{i1} - Y_{i2})^2$$

on the **expected** proportion of alleles shared IBD

$$\underline{\pi} = \pi_1/2 + \pi_2$$

where π_n is the probability of sharing n alleles IBD based on the available marker information. To allow for dominance, this regression model may be extended to:

$$E(D_i) = \beta \underline{\pi}_i + \gamma(\pi_{i1} - \underline{\pi})$$

It was further shown that $E(\beta) = 2\sigma_A^2 (1 - 2\theta)^2$

$$\text{and} \quad E(\gamma) = \sigma_D^2(1 - 2\theta)^4$$

where σ_A^2 and σ_D^2 are the additive and dominance components of the variance.

Several modifications of this approach had been proposed to refine the model, and to account for secondary and tertiary effects.

3.2.4 Association Studies

These studies are directed to the linked region which may include a number of genes with known functions that could be relevant to the etiology of the disease. By comparing the genotypes at these candidate loci between cases and controls, it provide critical association with the disease under study. Nevertheless, it should not escape ones attention that such associations may well be non-causal, thus reflecting linkage disequilibrium with the correct causal gene – as in fine mapping.

In general, using linkage analysis, a particular region may be identified which will contain a number of genes that have already been noted. Of these genes, one or more may have some specific function which is probably related to the disease under investigation – and may well be the target gene! One may note that it is a *Candidate Gene*!

Next, one proceeds to the studies of candidate gene associations to provide a direct test of the hypothesis that the candidate gene is the disease-causing gene that is being sought, viz., the ***Candidate Gene***! Now, further studies of the candidate gene ***associations*** would provide a direct test of the hypothesis that it is indeed the gene being sought! From this point on, one may follow a standard case-control design with population-based cases and unrelated controls. For example, one could then express the results in the usual standard 2x2 table, and check the association using a chi-square test.

Remarks:

1. A suitable choice of controls is needed to maintain the validity of this ed
2. Since ethnic origin is a concern (since different racial groups have diverse distributions among many genes, spurious results may occur if controls and cases are not properly matched!)
3. There are significant differences among various ethnic subgroups!
4. One way to develop a control series that is not affected by ethnic stratification is the use of family members as controls. It is common to use parents or siblings for this purpose. A *case-sibling design* may be considered as a standard case-control study. In case-parent trio designs, the parents are not used as controls, but rather the set of alleles carried by the parents themselves are not used as controls, but rather the set of alleles carried by the parents of the case that were not transmitted to the case. Thus, each case has two transmitted alleles and the control has two non-transmitted alleles. The data are arranged for matched case-control analysis, cross-tabulating the alleles the alleles which each parent transmitted to the case against those that are not transmitted.
5. This is the Transmission-Disequilibrium Test (TDT).
6. Note that an association does ***NOT necessarily imply that the gene has a causal effect on the disease owing to the phenomenon of Linkage Equilibrium!*** Whenever a marker gene and a disease gene are closely linked, their alleles may not be independently distributed in the whole population, i.e., carriers of the D allele may also be more likely to carry a particular marker allele. This may occur, for example, as a result of a near term mutation that does not yet have time to reach equilibrium among the descendants. Hence, association does ***NOT*** necessarily imply causation,

but may indicate simply a spurious result owing to an inappropriate choice of control or a linkage to a nearby gene. The following example illustrate this important point:

Example of an Association Study in Genetic Epidemiology: A Typical Research Study in the Genetics of Breast Cancer

Reference: Cui, J. S. (2003).- “Regressive logistic and proportional hazards disease models for within-family analyses of measured genotypes, with application to a CYP17 polymorphism and breast cancer”, *Genetic Epidemiology* **24**:161-172

In this study, a family-based case control design for studying a polymorphism in the CYP17 gene which regulates an early step in the pathway for estrogen synthesis pathway. Also, it was found that:

- (a) there exists a recessive effect of the *T* allele, with a relative risk of 1.5;
- (b) there are heterozygous mutations in the *ATM* gene - on the basis of population-based case control studies – which is associated with cancer generally, and breast cancer in particular, for which a meta-analysis of several associated studies showed an average relative risk of 3.9!
- (c) In its homozygous form, this particular gene produces the neurologic disease ataxia-telangiectasia which has, as one of its features, an extreme sensitivity to ionizing radiation, on the basis of population-based case control studies, heterozygous mutations in the *ATM* gene may be associated with cancer generally, and with breast cancer in particular, for which a meta-analysis of a number of studies have yielded a pooled relative risk of 3.9-to-1.0!
- (d) It is known that *ATM* encodes a protein that forms a complex with several others, including *BRCA1*, that is known to be responsible for the repair of double-strand breaks that are caused by radiation, thus the hypothesis that *ATM* and radiation mate interact in breast cancer is of interest.
- (e) Thus, this problem has become a subject of both family- and population-based research and studies.

Example of Computational Genetic Association Using R

GenCAT: Genetic Class Association Testing

Implementation of the genetic class level association testing (GenCAT) method from SNP level association data. Refer to: “Qian J, Nunez S, Reed E, Reilly MP, Foulkes AS (2016) <<https://doi.org/10.1371/journal.pone.0148218>> A Simple Test of Class-Level Genetic Association Can Reveal Novel Cardiometabolic Trait Loci. PLoS ONE 11(2): e0148218”.

Version:	1.0.3
Depends:	R (\geq 2.10), stats, dplyr , doParallel , ggplot2 , foreach , parallel methods
Suggests:	snpStats , knitr
Published:	2016-06-10
Author:	Eric Reed, Sara Nunez, Jing Qian, Andrea Foulkes
Maintainer:	Eric Reed <reeder at bu.edu>
License:	GPL-2
NeedsCompilation:	no
CRAN checks:	GenCAT results

Downloads:**Reference manual:**[GenCAT.pdf](#)**Vignettes:**[GenCAT Package](#)**Package source:**[GenCAT_1.0.3.tar.gz](#)**Windows binaries:**r-devel: [GenCAT_1.0.3.zip](#),r-release: [GenCAT_1.0.3.zip](#),r-oldrel: [GenCAT_1.0.3.zip](#)r-release: [GenCAT_1.0.3.tgz](#)**OS X El Capitan binaries:**r-oldrel: [GenCAT_1.0.3.tgz](#)**OS X Mavericks binaries:**[GenCAT archive](#)

Old sources:

Package ‘GenCAT’

June 10, 2016

Type Package

Title Genetic Class Association Testing

Version 1.0.3

Date 2016-06-10

Author Eric Reed, Sara Nunez, Jing Qian, Andrea Foulkes

Maintainer Eric Reed <reeder@bu.edu>

Description Implementation of the genetic class level association testing (Gen-CAT) method from SNP level association data. Refer to: ``Qian J, Nunez S, Reed E, Reilly MP, Foulkes AS (2016) <DOI:10.1371/journal.pone.0148218> A Simple

Test of Class-Level Genetic Association Can Reveal Novel Cardiometabolic

Trait Loci. PLoS ONE 11(2): e0148218”.

Suggests snpStats, knitr

Depends R (>= 2.10), stats, dplyr, doParallel, ggplot2, foreach, parallel, methods

License GPL-2

VignetteBuilder knitr

NeedsCompilation no

Repository CRAN

Date/Publication 2016-06-10 23:12:38

R topics documented:

GenCAT-package	2
CardioData	3
CardioMapped	4
coords	5
GenCAT	6
GenCAT_manhattan	7
geno	9
map2class	10

GenCAT_manhattan **Create Manhattan Plot of GenCAT Results**

Description

This function will create a Manhattan Plot from output of GenCAT function
GenCAT_manhattan

Usage

```
GenCAT_manhattan(GenCATout, sigThresh = NULL,
                  highlightPosi = FALSE,
                  labelPosi = FALSE,
                  sepChr = 8e+05,
                  plotTitle = "Manhattan Plot of GenCAT Results")
```

Arguments

GenCATout	An object of class, GenCATtest.
sigThresh	P-value threshold to highlight classes with strong signal from GenCAT test
highlightPosi	logical. If TRUE, classes with GenCAT p-value less than sigThresh will be shown in blue
labelPosi	logical. If TRUE, classes with GenCAT p-value less than sigThresh will be labelled.
sepChr	Specifies the space to put between chromosomes on the plot.
Plot Title	Character expression for plot title.

Details

GenCAT test is the class of the output of the GenCAT() function.

Author(s)

Eric Reed, Sara Nunez, Jing Qian, Andrea Foulkes

Examples

```
#####
#   Running GenCAT  #
#####

data("CardioMapped")
#Subset CardioMapped to decrease CPU time

CardioMappedSub<-CardioMapped[CardioMapped$chr < 15,]
set.seed(1)
CardioMappedSub<-CardioMappedSub[sample(1:nrow(CardioMappedSub), 100),]
library(snpStats)
data('geno')
```

```

genoData<-geno$genotypes
snpInfo<-geno$map
colnames(snpInfo)<-c('chr', 'SNP', 'gen.dist', 'position', 'A1', 'A2')
print(head(snpInfo))
GenCATtest <- GenCAT(CardioMappedSub, genoData=genoData, snpInfo = snpInfo)
#####
#Create Manhattan Plot
geno 9
#####
print(str(GenCATtest))
GenCAT_manhattan(GenCATtest, sigThresh = (0.05/nrow(GenCATtest$GenCAT)),
highlightPosi = TRUE, labelPosi = TRUE)

```

In the R domain:

```

> install.packages("GenCAT")
Installing package into 'C:/Users/Bert/Documents/R/win-library/3.3'
(as 'lib' is unspecified)
--- Please select a CRAN mirror for use in this session ---

```

A CRAN mirror is selected

```

trying URL 'https://mirrors.tuna.tsinghua.edu.cn/CRAN/bin/windows/contrib/3.3/
GenCAT_1.0.3.zip'
Content type 'application/zip' length 3778881 bytes (3.6 MB)
downloaded 3.6 MB

package 'GenCAT' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
      C:\Users\Bert\AppData\Local\Temp\RtmpKQVO0B\downloaded_packages
> library(GenCAT)
Loading required package: dplyr

Attaching package: 'dplyr'
The following objects are masked from 'package:stats':
  filter, lag
The following objects are masked from 'package:base':
  intersect, setdiff, setequal, union

Loading required package: doParallel
Loading required package: foreach
foreach: simple, scalable parallel programming from Revolution Analytics
Use Revolution R for scalability, fault tolerance and more.
http://www.revolutionanalytics.com

```

```
Loading required package: iterators
Loading required package: parallel
Loading required package: ggplot2
> ls("package:GenCAT")
[1] "GenCAT"                  "GenCAT_manhattan" "map2class"
>
> #####
> #Running GenCAT
> #####
> data("CardioMapped")
> #Subset CardioMapped to decrease CPU time
> CardioMappedSub<-CardioMapped[CardioMapped$chr < 15,]
> set.seed(1)
> CardioMappedSub<-
+  CardioMappedSub[sample(1:nrow(CardioMappedSub), 100),]
> library(snpStats)
Loading required package: survival
Loading required package: Matrix
> data('geno')
> genoData<-geno$genotypes
> snpInfo<-geno$map
> colnames(snpInfo)<-c('chr', 'SNP', 'gen.dist', 'position', 'A1',
+                      'A2')
> print(head(snpInfo))
      chr      SNP      gen.dist position A1 A2
rs624673  13 rs624673      NA 19743996  G  A
rs9511877  13 rs9511877      NA 19744070  A  G
rs638773   13 rs638773      NA 19744848  A  G
rs9511880  13 rs9511880      NA 19745096  G  T
rs482278   13 rs482278      NA 19745251  A  G
rs9507552  13 rs9507552      NA 19745903  G  A
> GenCATtest <- GenCAT(CardioMappedSub,
+                      genoData=genoData, snpInfo = snpInfo)
[1] "Running GenCAT on 41 classes on chromosome 13."
[1] "Running GenCAT on 43 classes on chromosome 14."
> #####
> #Create Manhattan Plot
>
> print(str(GenCATtest))
List of 5

  $ GenCAT    :'data.frame':    75 obs. of  6 variables:
  ..$ class : chr [1:75] "STK24" "LRCH1" "FLT1" "HTR2A" ...
  ..$ chr   : num [1:75] 13 13 13 13 13 13 13 13 13 13 ...
  ..$ n_SNPs: num [1:75] 1 1 1 1 1 1 1 1 1 1 ...
  ..$ n_Obs : num [1:75] 1 1 1 1 1 1 1 1 1 1 ...
  ..$ CsumT : num [1:75] 2.086 1.127 5.885 0.318 3.626 ...
  ..$ CsumP : num [1:75] 0.1486 0.2884 0.0153 0.5727 0.0569 ...
  $ Used     :'data.frame':    88 obs. of  9 variables:
  ..$ SNP     : chr [1:88] "rs4389009" "rs844520" "rs11149523" "rs9567737" ...
```

```

..$ effect_allele: chr [1:88] "G" "G" "G" "C" ...
..$ other_allele : chr [1:88] "A" "A" "A" "T" ...
..$ testStat      : num [1:88] -1.444 1.062 2.426 -0.564 -1.904 ...
..$ class         : chr [1:88] "STK24" "LRCH1" "FLT1" "HTR2A" ...
..$ chr           : num [1:88] 13 13 13 13 13 13 13 13 13 13 ...
..$ position      : num [1:88] 99193519 47264930 28995630 47421266 43639845 ...
..$ A1            : chr [1:88] "A" "G" "A" "C" ...
..$ A2            : chr [1:88] "G" "A" "G" "T" ...
$ notFound   :'data.frame':   12 obs. of  5 variables:
..$ SNP         : chr [1:12] "rs9572807" "rs7987481" "rs17253843" "rs7335275" ...
..$ effect_allele: chr [1:12] "C" "G" "G" "C" ...
..$ other_allele : chr [1:12] "T" "A" "A" "T" ...
..$ testStat      : num [1:12] -0.3015 -1.0024 0.0867 0.7788 -0.1335 ...
..$ class         : chr [1:12] "DACH1" "LRCH1" "GPC6" "ABCC4" ...
$ unMatched  :'data.frame':   0 obs. of  9 variables:
..$ SNP          : logi(0)
..$ effect_allele : logi(0)
..$ other_allele : logi(0)
..$ testStat      : num(0)
..$ class         : logi(0)
..$ chr           : num(0)
..$ position      : num(0)
..$ A1            : logi(0)
..$ A2            : logi(0)
$ TransStats   :'data.frame':   88 obs. of  2 variables:
..$ class         : chr [1:88] "STK24" "LRCH1" "FLT1" "HTR2A" ...
..$ transStat    : num [1:88] 1.444 1.062 -2.426 -0.564 1.904 ...
- attr(*, "class")= chr "GenCATtest"
NULL
> GenCAT_manhattan(GenCATtest, sigThresh =
+                     (0.05/nrow(GenCATtest$GenCAT)),
+                     highlightPosi = TRUE, labelPosi = TRUE)
> # Outputting: Figure 3.1
>

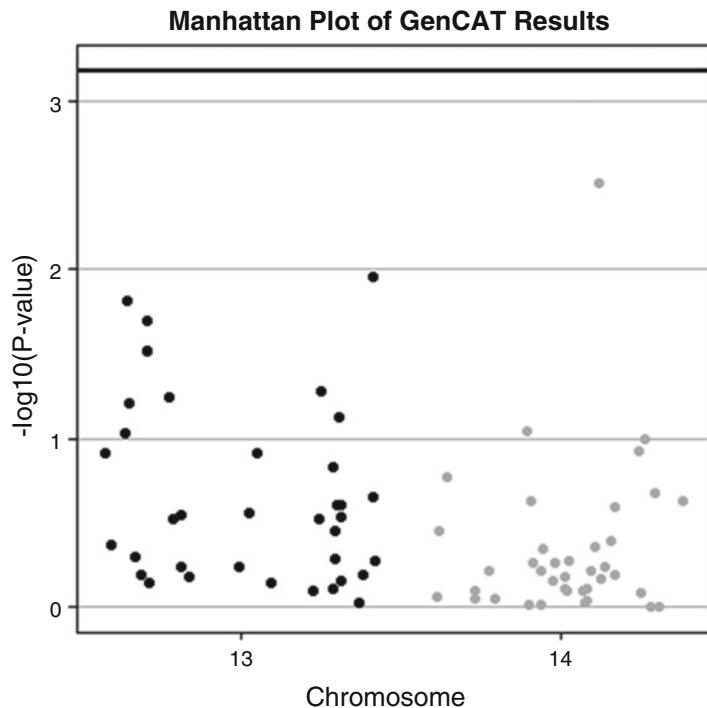
```

3.3 Genome-wide Association Studies (GWAS)

As with candidate gene association, studies of partial and whole genome-wide scans, called *Genome-Wide Association Studies*, GWAS, are performed to elucidate associations among SNPs and a genetic trait. Typical partial-genome scans generally involve 100 Kb and 500 Kb segments of the DNA, and whole-genome scans cover 500Kb to 1000Kb regions. “SNP” chips are available for greater efficiency.

Generally, the population-based genetic association studies have the common objective of relating genetic information to a phenotype or to some definitive clinical outcomes – and both may be considered as a **trait**:

Fig. 3.1 Manhattan of GenCAT results



*Thus, a **Binary Trait** refers to binary variables, the latter is defined as one that can have two values: such as *Diseased* or *Not Diseased*.

*A **Phenotype** is defined as a physical attribute, viz., the manifestation of a trait, in the sense of a measures of disease progression

3.3.1 A Worked Example of SNPs-based Whole Genome Association Study

Here below is a worked example taken from CRAN:

Package 'GenABEL', February 19, 2015

Type Package

Title genome-wide SNP association analysis

Version 1.8-0

Date 2013-12-09

Author GenABEL project developers

Contact GenABEL project developers <genabel.project at gmail.com>

Maintainer Yurii Aulchenko <yurii@bionet.nsc.ru>

Depends R (>= 2.15.0), methods, MASS, utils, GenABEL.data

Suggests qvalue, genetics, haplo.stats, DatABEL (>= 0.9-0), hglm, MetABEL, PredictABEL, VariABEL, bigRR

Description a package for genome-wide association analysis between quantitative or binary traits and single-nucleotide polymorphisms (SNPs).

License GPL (>= 2)

URL <http://www.genabel.org>, <http://forum.genabel.org>,
<http://genabel.r-forge.r-project.org/>

BugReports http://r-forge.r-project.org/tracker/?group_id=505

NeedsCompilation yes

Repository CRAN

Date/Publication 2013-12-27 14:47:02

add.plot function to plot additional GWAA results

Description

Add plot of results of GWA analysis

Usage

```
add.plot(x, ..., df = 1, col=c("lightgreen", "lightblue"), sort=TRUE, delta = 1)
```

Arguments

x object of type scan.gwaa-class, as returned by scan.glm, qtsscore, ccfast, emp.ccfast, emp.qtsscore, or scan.haplo; or of type scan.gwaa.2D-class, as returned by scan.haplo.2D or scan.glm.2D.

... additional arguments to be passed to plot

arrange_probabel_phe

df P-value at which df to add (1, 2 or "Pc1df")

col which colors to use to depict consecutive chromosomes

sort whether results should be plotted after sorting by chromosome and position

delta gap width between chromosomes

Value No value returned.

Author Yurii Aulchenko

See Also

plot,.snp.subset, scan.glm, qtsscore, ccfast, emp.qtsscore, emp.ccfast, scan.haplo, scan.haplo.2D, scan.glm.2D

Example

```
require(GenABEL.data)
data(srdata)
a <- ccfast("bt",srdata,snps=c(1:100))
plot(a)
a1 <- qtsscore(bt,srdata,snps=c(1:100))
add.plot(a1,col="red",type="l")
```

In the R domain:

```
>
> install.packages("GenABEL")
Installing package into 'C:/Users/Bert/Documents/R/win-library/3.3'
(as 'lib' is unspecified)
trying URL 'https://ftp.cc.uoc.gr/mirrors/CRAN/bin/windows/contrib/3.3/GenABEL_1.8-0.zip'
Content type 'application/zip' length 3643610 bytes (3.5 MB)
downloaded 3.5 MB

package 'GenABEL' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
  C:/Users/Bert/AppData/Local/Temp/RtmpQBWujH downloaded_packages
> library(GenABEL)
Loading required package: MASS
> ls("package:GenABEL")
 [1] "add.phdata"                      "add.plot"
 [3] "annotation"                      "arrange_probabel_phe"
 [5] "as.character.gwaa.data"          "as.character.snp.coding"
 [7] "as.character.snp.data"           "as.character.snp.strand"
 [9] "as.data.frame.gwaa.data"         "as.double.gwaa.data"
[11] "as.double.snp.data"              "as.genotype"
[13] "as.genotype.gwaa.data"           "as.genotype.snp.data"
[15] "as.hsgeno"                      "as.hsgeno.gwaa.data"
[17] "as.hsgeno.snp.data"              "autosomal"
[19] "blurGenotype"                   "catable"
[21] "ccfast"                         "check.marker"
[23] "check.trait"                    "checkPackageVersionOnCRAN"
[25] "chi2_CG"                        "chromosome"
[27] "cocohet"                        "coding"
[29] "coding<->"                   "convert.snp.affymetrix"
[31] "convert.snp.illumina"           "convert.snp.mach"
[33] "convert.snp.ped"                "convert.snp.text"
[35] "convert.snp.tped"               "crnames"
[37] "del.phdata"                    "descriptives.marker"
[39] "descriptives.scan"              "descriptives.trait"
[41] "dprfast"                        "effallele"
[43] "egscore"                        "egscore.old"
[45] "emp.ccfast"                   "emp.qtscore"
[47] "estlambda"                     "export.impute"
[49] "export.merlin"                 "export.plink"
[51] "extract.annotation.impute"     "extract.annotation.mach"
[53] "findRelatives"                 "formetascore"
[55] "GASurv"                        "generateOffspring"
[57] "getcall"                        "getfamily"
[59] "getLogLikelihoodGivenRelation" "grammar"
[61] "gtdata"                         "hom"
```

```

[63] "hom.old"                      "HWE.show"
[65] "ibs"                           "ibs.old"
[67] "idnames"                      "impute2databel"
[69] "impute2mach"                  "lambda"
[71] "load.gwaa.data"                "mach2databel"
[73] "makeTransitionMatrix"          "male"
[75] "map"                           "merge.gwaa.data"
[77] "merge.snp.data"                "mlreg"
[79] "mlreg.p"                      "mmscore"
[81] "nids"                          "npsubtreated"
[83] "nsnps"                         "patch_strand"
[85] "perid.summary"                 "PGC"
[87] "phdata"                        "phdata<="
[89] "plot.check.marker"             "plot.scan.gwaa"
[91] "plot.scan.gwaa.2D"             "polygenic"
[93] "polygenic_hglm"                "qtscore"
[95] "qvaluebh95"                   "r2fast"
[97] "r2fast.old"                   "recodeChromosome"
[99] "reconstructNPs"                "redundant"
[101] "refallele"                   "refresh.gwaa.data"
[103] "reg.gwaa"                     "results"
[105] "rhofast"                      "rntransform"
[107] "save.gwaa.data"                "scan.glm"
[109] "scan.glm.2D"                  "scan.haplo"
[111] "scan.haplo.2D"                "show"
[113] "show.ncbi"                     "snp.data"
[115] "snp.names"                     "snp.subset"
[117] "snpnames"                      "sortmap.internal"
[119] "sset"                           "strand"
[121] "strand<-"                     "summary.check.marker"
[123] "summary.snp.data"              "summary.snp.data_old"
[125] "VIFGC"                         "VIFGC_ovdom"
[127] "Xfix"                          "ztransform"
>
> require(GenABEL.data)
> data(srdta)
> a <- ccfast("bt", srdta, snps=c(1:100))
Warning in ccfast("bt", srdta, snps = c(1:100)) :
  11 people (out of 2500 ) excluded as not having cc status
> plot(a)
> # Outputting      Figure 3.2 GenABEL-"add.plot"-1
>
> a1 <- qtscore(bt, srdta, snps=c(1:100))
Warning messages:
1: In test.type(y, trait.type) : binomial trait is analysed as
   gaussian
2: In qtscore(bt, srdta, snps = c(1:100)) :
  11 observations deleted due to missingness

```

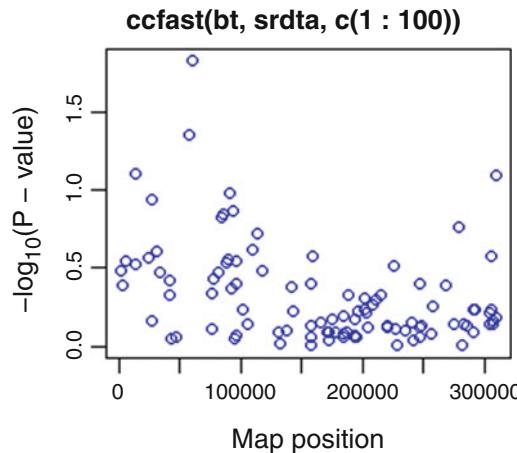


Fig. 3.2 GenABEL- "add.plot"-1

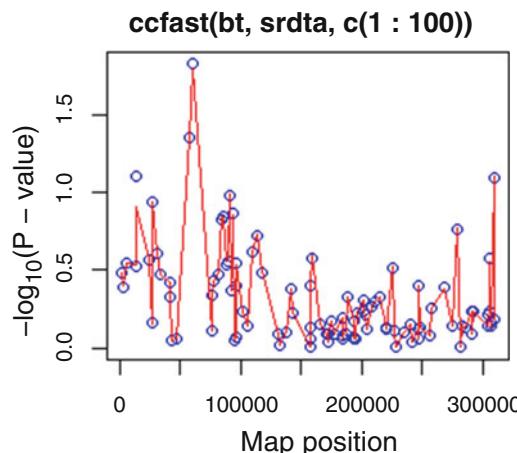


Fig. 3.3 GenABEL- "add.plot"-2

```

3: In qtscore(bt, srdta, snps = c(1:100)) : Lambda estimated < 1, set to 1
> add.plot(a1,col="red",type="l")
>
> # Outputting Figure 3.3 GenABEL- "add.plot"-2
>

>
> a1 <- qtscore(bt,srdta,snps=c(1:100))
Warning messages:
1: In test.type(y, trait.type) : binomial trait is analysed as gaussian
2: In qtscore(bt, srdta, snps = c(1:100)) :
   11 observations deleted due to missingness

```

```

3: In qtsscore(bt, srdta, snps = c(1:100)) :
  Lambda estimated < 1, set to 1
> add.plot(a1,col="red",type="l")
>
> # Outputting Figure GenABEL - "add.plot" - 2
>

```

3.4 Big Data and Human Genomics

3.4.1 What Is Big Data? [W]

Big Data is now an accepted term for describing datasets with such large complexity or volume that conventional data processing methods are *not* good enough! **Big Data** has been described disparately by different people. The most popular definition of **Big Data** is the 5Vs:

- *Volume
- *Velocity
- *Variety
- *Verification/Veracity, and
- *Value.

The definition of **Big Data** will likely be subjected to technological advances in the future. **Big Data** infrastructure is a framework, which covers important components databases, Massively Parallel Processing (MPP), and others, that is used for storing, processing, and analyzing **Big Data**. **Big Data** analytics covers collection, manipulation, and analyses of massive, diverse data sets that contain a variety of data types including genomic data and EHRs to reveal hidden patterns, cryptic correlations, and other intuitions on a **Big Data** infrastructure. Owing to its effectiveness, **Big Data** analytics is widely used in different research fields. One may describe how one type of **Big Data**, genomic data, is applied to improve clinical research and healthcare. We give an overview of the challenges in processing genomic data and EHRs, provide possible solutions to overcome these challenges using approaches that ensure the safety of genomic data, and present a **Big Data** solution for identifying clinically actionable variants in sequence data. We also discuss the requirement for the efficient integration of genomic information into EHRs.

3.4.2 What Is Genetic Big Data? And Where Is It Taking Genetics?

Researchers finished the first draft of the human genome in the year 2000. By 2013, it was possible to produce 18,000 high quality human genomes per year at \$1,000 per genome. Also, in 2013, it was reported that worldwide about [225,000 genomes had sequenced worldwide to date](#), It is also estimated that about five million complete human genomes may be sequenced by the year 2020!

Moreover, in addition to gathering genomic data from a wide population of participants, “**health hubs**” will also be set up around the world to gather a vast collection of physiological information about each genome they sequenced.

This data is also known as phenotype – the biological expression of genes in the body – and plays out in physical traits like blue eyes or genetic disorders like Huntington’s disease. By matching

phenotype to genotype and comparing them over large populations, researchers hope to decipher which genes or groups of genes are responsible for which biological traits.

This is **Big Data!** Like all big data ventures, it's meaningless without effective methodologies and tools of analysis.

All these human genome data have been online for 15 years, and one knows about as much about it now as one did when it was first sequenced. More data is better, he said, but data is not the goal – the aim is to *take data and generate knowledge with it*.

It is realized that machine learning algorithms are showing themselves capable of handling big data to make connections no human could likely see. Selected machine algorithms trained on lung tumor MRIs were able to discover new diagnostic features. And software proved more accurate than humans at predicting 5 year survival probabilities based on breast cancer biopsies.

These and other artificial intelligence and computational approaches may likewise, in the future, take the enormous sets of data generated by large genomic studies and extract knowledge from them.

3.4.3 Analysis of Human Genomics

Initial sequencing and analysis of the human genome International Human Genome Sequencing Consortium has been presented in: **Nature 409: 860-921 (2001)**. This report confirmed that the human genome holds a large amount of information about human development, physiology, medicine and evolution. The results, of an international collaboration to produce and make freely available a draft sequence of the human genome, is being presented. Included in this presentation also is an initial analysis of the data, describing some of the insights that can be gleaned from the sequence.

Herebelow is a summary of this report:

The scientific progress made falls naturally into four main phases:

1. The first established the cellular basis of heredity: the chromosomes.
 2. The second defined the molecular basis of heredity: the DNA double helix.
 3. The third unlocked the informational basis of heredity, with the discovery of the biological mechanism by which cells read the information contained in genes and with the invention of the recombinant DNA technologies of cloning and sequencing by which scientists can do the same.
 4. The fourth has been marked by a relentless drive to decipher first genes and then entire genomes, spawning the field of genomics. The fruits of this work already include the genome sequences of 599 viruses and viroids, 205 naturally occurring plasmids, 185 organelles, 31 eubacteria, seven archaea, one fungus, two animals and one plant.
-
- There seem to be about 30,000–40,000 protein-coding genes in the human genome—only about twice as many as in worm or fly. However, the genes are more complex, with more alternative splicing generating a larger number of protein products.
 - Hundreds of human genes appear likely to have resulted from horizontal transfer from bacteria at some point in the vertebrate lineage. Dozens of genes appear to have been derived from transposable elements.
 - Although about half of the human genome derives from transposable elements, there has been a marked decline in the overall activity of such elements in the hominid lineage. DNA transposons appear to have become completely inactive and long-terminal repeat (LTR) retrotransposons may also have done so.

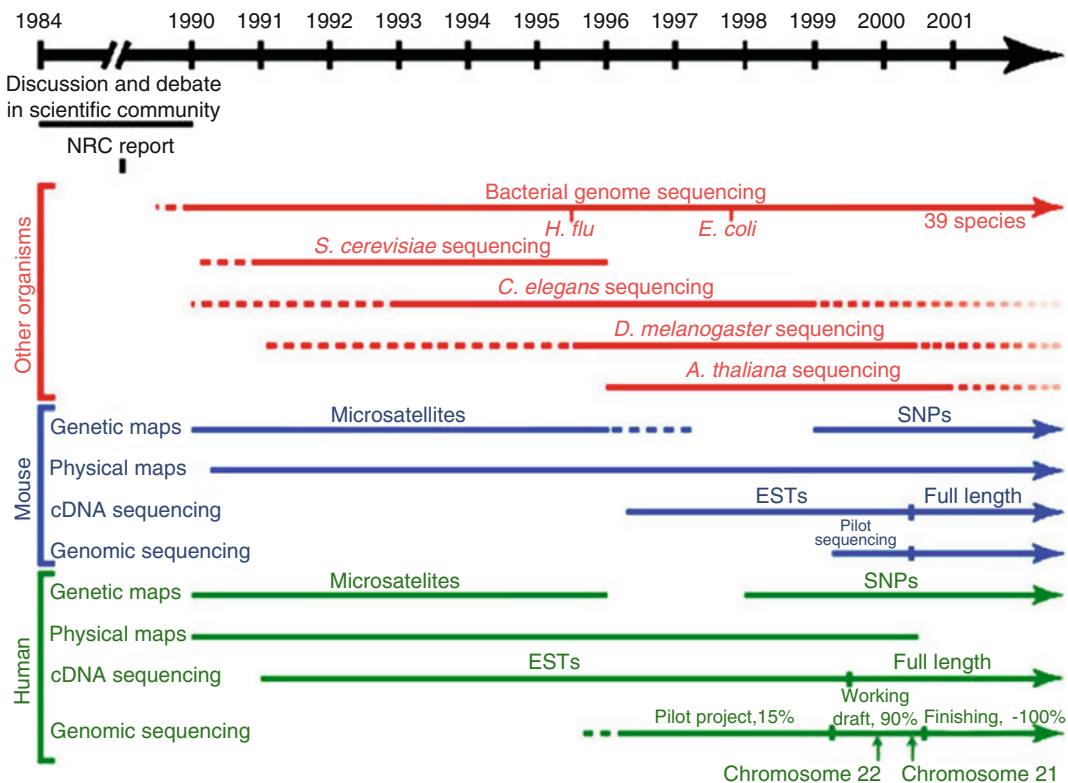


Fig. 3.4 Timeline of large-scale genomic analyses

- The pericentromeric and subtelomeric regions of chromosomes are filled with large recent segmental duplications of sequence from elsewhere in the genome. Segmental duplication is much more frequent in humans than in yeast, fly or worm.
- Analysis of the organization of Alu elements explains the longstanding mystery of their surprising genomic distribution, and suggests that there may be strong selection in favor of preferential retention of Alu elements in GC-rich regions and that these ‘selfish’ elements may benefit their human hosts.
- More than 1.4 million single nucleotide polymorphisms (SNPs) in the human genome have been identified. This collection should allow the initiation of genome-wide (Fig. 3.4)

References

- Chan BKC (2015) Biostatistics for epidemiology and public health using R. Springer, New York
- Haseman J, Elton R (1972) The investigation of linkage between a quantitative trait and a marker locus. *Behav Genet* 2:3–19; Quoted in Thomas, D. C. (2004)
- Teare DM, Barrett JH (2005) Genetic linkage studies. *Lancet* 366(9490):1036–1044
- [W] Wikipedia



Abstract

This chapter covers the study of human epidemiology, including family studies in genetic epidemiology, linkage analysis, genetic Mapping in human diseases, human genetic influences on diseases, genetic relationships in familial aggregation, and derivation of familial risk.

An Illustration is provided of a research project in genetic epidemiology research which included

- (1) Heritability Analysis
- (2) Molecular Variation Study Methods
- (3) Genomics for Human Genetic Epidemiology
 - Complex Traits and Mendelian Inheritance
 - Mendel's Laws
 - Hardy-Weinberg Principle
 - Gene Structure and Genetic Code
 - Genetic Linkage and Linkage Disequilibrium
 - Study Designs for of Rare Genetic Variations
 - Spectrum of Variation
 - Familial Factors in Human Genetic Epidemiology
 - *Human Genetic Association
 - Genetic Epidemiology Owing to Population Stratification
 - Environmental Effects on Genetic Epidemiology
 - Genetic Epidemiology and Public Health

Keywords

Heritability analysis · Molecular variation study methods · Genomics for human genetic epidemiology · Complex traits and mendelian inheritance · DNA sequencing · Study designs for genetic variants · Spectrum of variation · Familial factors in human genetic epidemiology · Linkage analysis · Family association studies · The Transmission Disequilibrium Test (TDT) · Human genetic association · Genetic epidemiology owing to population stratification · Environmental effects on genetic epidemiology · Genetic epidemiology and public health · Environmental factors on genetic epidemiology

4.1 The Study of Human Genetic Epidemiology

In 2010, the *Yale Journal of Biology and Medicine* (DeWan, A. T. (2010).- *Yale J Biol Med.*; 83(2): 87–90, published online 2010 June) introduced “Five Classic Articles in Genetic Epidemiology”. It declared that:

The application of epidemiological principles to human genetic studies was first discussed by Neel and Schull (1954). Soon thereafter, genetic epidemiology has been defined by practitioners with diverse backgrounds, and two important concepts underlying genetic epidemiology have emerged:

- (1) the study of the etiology of disease among groups of relatives to unravel the cause of family resemblance; and
- (2) the study of inherited causes of disease in populations (Morton and Chung 1975).

The essence of the work of a genetic epidemiologist is to assess the nature of genetic inheritance within families and populations, using classical epidemiology, statistical genetics, human genetics, and population genetics. It was found that more than highlighting any one gene discovery, one pushes forward ones ability to make these discoveries. These papers have greatly influenced how one think about and conduct the task of unraveling the genetic causes of disease. These include:

- (1) the description of the **Logarithm of Odds (LOD)** score for a sequential test of linkage in multiple families (Morton 1955);
 - (2) the first report of a generally available computer program to perform linkage analysis (Ott 1974);
 - (3) a theoretical paper that made the case for applying association mapping to identify the genetic basis of complex diseases (Risch and Merikangas 1996);
 - (4) the report that signified the initial completion of the public human genome project (Lander et al. 2001); and
 - (5) the first report of a successful genome-wide association study using SNP markers (Ozaki et al. 2002).
- (1) Morton, N. E. (1955). – “Sequential tests for the detection of linkage”, *Am J Hum Genet.* 7(3):277–318

The study of phenotypic variation owing to inherited genetic variation intuitively leads one directly to the family as the unit of study. While direct observation of recombination events between genotype and phenotype in a pedigree is the preferred method for assessing linkage to a particular genetic locus in a pedigree, incomplete data and unknown genetic model parameters make this method inefficient. Early statistical approaches to conduct linkage analysis tended to have limited applications in terms of pedigree size and were laborious to apply before the advent of computer programs to do so. Morton introduced the “**Logarithm (base 10) Of oDds**” (**LOD**) score method for computing linkage. LOD scores have the property of being additive and, thus, can be applied to multiple families subsequently summed to test for linkage over a number of independent families. Over the past 50 or so years, numerous extensions and modifications have been made to Morton’s original LOD score method, but this paper laid the foundation for how we continue to conduct linkage analysis today.

- (2) Ott, J. (1974).– “Estimation of the recombination fraction in human pedigrees: efficient computation of the likelihood for human linkage studies”, *Am J Hum Genet.* 26(5):588–597

The **L**ikelihoods in **P**EDigrees (**LIPED**) computer program, as proposed in paper by Ott, was the first widely available program for conducting genetic linkage studies. Up to this point, researchers

relied primarily on published LOD score tables to assess linkage. The publication of LIPED added a new tool to the genetic epidemiologist's toolbox: the computer. This program implemented the Elston-Stewart algorithm (Elston and Stewart 1971) for computing the likelihood in general pedigrees, allowing researchers to compute LOD scores in general pedigree with few exceptions. While this was not the first effort to exploit computers to conduct linkage analysis, this was the first program that was both portable and could compute likelihoods in more than two generation pedigrees. The use of computer programs, most freely available and written by scientists in the field, to analyze genetic data has become an invaluable tool to the genetic epidemiologist, with nearly 500 programs currently available (<http://linkage.rockefeller.edu/soft/list.html>).

- (3) Risch N, and Merikangas K. (1996). – “The future of genetics studies of complex human diseases”, *Science*. 273(5281):1516–1517

Geneticists had been successful at identifying genetic loci for single-gene Mendelian diseases through linkage analysis but less efficient in identifying loci for traits with more complex inheritance patterns such as schizophrenia and diabetes. In this paper, the authors demonstrate that for a locus with a **Genotypic Risk Ratio (GRR)** of 1.5 (the high end of the range of GRRs observed in current studies of complex traits) and a disease allele frequency greater than 10 percent, 37 to 72 times more families were needed to detect significant linkage compared to tests of association. This paper resulted in a shift in thinking *from linkage to association studies* for complex traits likely resulting from multiple loci each with modest effects. The major limitation to this approach was the lack of known polymorphisms needed to interrogate the entire genome for association, which was estimated to be 100,000 to 1 million. However, the requisite number of markers in the form of **Single-Nucleotide Polymorphisms (SNPs)** and the ability to both rapidly and cheaply genotype them would become available in less than 10 years from the publication of this paper, leading to myriad successful genome-wide association studies for complex diseases.

- (4) Lander, E. S., Linton LM, et al. (2001). – “Initial sequencing and analysis of the human genome”. *Nature*, 409(6822): 860–921

The sequencing of the entire human genome had long been discussed in the genetics literature. The complete sequence of the human genome revolutionized not only genetic epidemiology, but all of human genetics. A reference genome now existed to which all other human genome sequences could be compared. A much more accurate estimate of the total number of genes in the human genome was made of approximately 25,000 genes, much smaller than the 100,000 estimated in the years leading up to the complete sequence. For the genetic epidemiologist, this meant that more markers were available than ever before for the identification of genomic regions harboring disease susceptibility loci. More genomic annotation information was accessible to easily discover which genes were contained in linkage regions. This paper is chosen to announce the completion of the public project because this project made its data freely available as it was generated, but it also recognized that the private sequencing effort that was published at the same time (Venter et al. 2001). This sequencing effort took a completely different and faster whole-genome shotgun sequencing approach that in some ways pushed the public effort to finish faster. While the privately sequenced genome required an expensive subscription to gain full access to the fully annotated data, the sequencing approach they took built the foundation for whole-genome sequencing that is used today. However, it was the freely available and was constantly improving annotation of the public project that makes it a giant step forward for everyone in the genetics community.

- (5) Ozaki K, Ohnishi Y, et al. (2002). – “Functional SNPs in the lymphotoxin-alpha gene that are associated with susceptibility to myocardial infarction.” *Nat Genet.*, 32(4):650–654

This is the first report of a **Genome-Wide Association Study (GWAS)** that successfully identified a function variant associated with a disease, in this case a coding change in lymphotoxin-alpha associated with myocardial infarction. However, this study was more important than simply identifying a genetic variant for myocardial infarction; it was the first demonstration that the long discussed GWAS was viable. The study involved genotyping >92,000 SNPs, more markers than had ever been genotyped in an association study, using a multiplex PCR-based assay, the order of magnitude of markers that was mentioned by Risch and Marikangas when they discussed the current limitations to conducting true genome-wide association studies in the mid-1990s. While this was a significant advance, the genotyping of hundreds of thousands of SNPs in this manner is a daunting task. Over the next several years, microarray technology, generally used for gene expression assays, was adapted to conduct truly high-throughput whole genome SNP genotyping (Kennedy et al. 2003; Matsuzaki et al. 2004a, b). Then, in 2005, one noted the first publication of a GWAS, this time identifying a functional mutation in complement factor H associated with age-related macular degeneration (Klein et al. 2005), which used this new microarray technology. This has led to a flood of GWAS with loci being mapped for a diverse set of diseases, including asthma, Crohn’s disease, diabetes, and glaucoma. Current microarrays are now able to genotype up to 1 million SNPs. This has forced the genetic epidemiology and statistical genetics communities to develop new analytical techniques to deal with this vast amount of data and allow one to realize its full potential. As one continues to explore these enormous GWAS datasets, the field continues to quickly advance, and now one is witnessing the first set of whole-genome sequencing studies that have identified several medically relevant rare loci (Choi et al. 2009; Lupski et al. 2010; Roach et al. 2010), an exciting development that promises to further hone the ability to detect genetic loci contributing to heritable diseases within the population.

4.1.1 Family Studies in Genetic Epidemiology

In all of genetic epidemiology, a fundamental component is to discover disease susceptibility genes in family studies. Here, the LOD score linkage analyses is first undertaken: comparing the number of recombinant and non-recombinant offsprings in a search for evidence of co-segregation between a hypothetical disease gene and a recognized genetic marker,

Next, one may follow the Identical-By-Descent (IBD) path, using non-parametric linkage analysis, which depends on excess sharing of marker alleles approach – especially in the absence of multi-generational family data. In families in which parent-offsprings trios are available, genetic epidemiologic studies may be undertaken.

Where possible, family studies in genetic epidemiology may be undertaking linkage analysis to exome sequencing for finding disease susceptibility genes. This approach can provide major progress in the understanding of genetic susceptibility to some rare and familial form of complex diseases.

It is to be recognized that these approaches apply molecular discoveries to improve the diagnosis, prevention, and treatment of genetics-related diseases.

4.1.1.1 Linkage Analysis

For linkage analysis in family studies in genetic epidemiology, the “**Logarithm (base 10) Of Odds**” (**LOD**) score linkage analysis, is the pre-eminent approaches of mapping disease genes. It is based

upon the search for evidence of co-segregation of diseases with known genetic markers within families. The aim of linkage analysis is the identification of the approximate chromosomal locations of a disease susceptibility gene.

The biological basis, and hence the rationale of linkage analysis is as follows:

- (i) During meiosis and the formation of gametes, alleles on the same chromosome (viz., the linked genes) are inherited together except when crossing over, or recombination, occurs at the first division of meiosis.
- (ii) During prophase I of meiosis, homologous material and paternal chromosomes line up at the centromere. Next, recombination, or crossing over, takes place when sister chromatids on the paternal and maternal chromosomes exchange genetic materials. This is a normal process and it breaks up associations between alleles along the same chromosome during cell division.
- (iii) The closer together two genes are on the same chromosome, the less likely recombination between them occurs. Thus, the frequency of crossing over depends on the physical distance between two such genes, thus creating a way to identify deviations from independent segregation and to estimate the genetic distance between any two genes.
- (iv) Comparing:
 - A. this null hypothesis of ***no linkage*** (viz., independent assortment or 50% recombination) with
 - B. the alternate hypothesis of ***linkage*** (in which recombination is reduced),

provides a biostatistical test for linkage, and allows the genetic distance to be estimated.

- (v) This approach may be used to indicate evidence of co-segregation between a hypothetical gene controlling a disease phenotype (which provides evidence of linkage to map a previously unknown causal gene) and an observed marker.
- (vi) Since linkage is based upon the physical distance between two genes, it is a characteristic of loci along a chromosome, and not the alleles at these particular loci.

4.1.1.2 Linkage Analysis and Genetic Mapping in Human Diseases

Reference: Altshuler, D., Daly, M. J., and Eric S. Lander, E. S. (2008).-*Science*, Nov 7, 322 (5903):881–888. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2694957/>

Science. Author manuscript; available in PMC 2009 Jun 11. Published in final edited form as: Science. 2008 Nov 7; 322(5903): 881–888.

By the early 1900s, geneticists understood that Mendel's laws of inheritance underlie the transmission of genes in diploid organisms. It was noted that some traits are inherited according to Mendel's ratios, as a result of alterations in single genes, and methods were developed to map the genes responsible. It was also recognized that most naturally occurring trait variation, while showing strong correlation among relatives, involves the action of multiple genes as well as non-genetic factors. Although it was clear that these insights applied to humans as much as to animals and insects, it took most of the century to turn these concepts into practical tools for discovering genes contributing to human diseases. Starting in the 1980s, the use of naturally occurring DNA variation as markers to trace inheritance in families led to the discovery of thousands of genes for rare Mendelian diseases. Despite great hopes, the approach proved ***unsuccessful*** for common forms of human diseases—such as diabetes, heart disease, and cancer—that show very complex inheritance in the general population.

Recently, a new approach to genetic mapping has yielded the first general progress toward mapping loci that influence susceptibility to common human diseases. However, most of the genes and mutations underlying these findings remain to be defined and understood, and it remains unclear how much of the heritability of common disease they explain. In the following paragraphs, the intellectual foundations of genetic mapping will be considered, emerging lessons noted, questions discussed, and challenges that lie ahead carefully considered.

Genetic Mapping by Linkage and Association

Genetic mapping is the localization of genes underlying phenotypes on the basis of correlation with DNA variation, without prior hypotheses about biological function. The simplest form, called *linkage analysis*, had been conceived by Sturtevant for fruit flies in 1913 (Neel and Schull 1954). Linkage analysis involves crosses between parents that vary at a Mendelian trait and at many polymorphic variants (“markers”); because of meiotic recombination, any marker showing correlated segregation (“linkage”) with the trait must lie nearby in the genome. In the 1970s, the ability to clone and sequence DNA made it possible to tie genetic linkage maps in model organisms to the underlying DNA sequence, and thereby to molecularly clone the genes responsible for any Mendelian trait solely on the basis of their genomic position (Morton and Chung 1975; Morton 1955). Such studies typically involved three steps:

- (i) identifying the locus responsible through a genome-wide search;
- (ii) sequencing the region in cases and controls to define causal mutations; and
- (iii) studying the molecular and cellular functions of the genes discovered. Such “positional cloning” became a mainstay of experimental genetics, identifying pathways that are crucial in medical biology and physiology.

Linkage Analysis in Humans

For most of the 20th century, genome-wide linkage mapping was impractical in humans: Family sizes are small, and there were too few classical genetic markers to systematically trace inheritance. Progress in identifying the genes contributing to human traits was initially limited to studies of biological candidates such as blood-type antigens (4) and hemoglobin β protein in sickle-cell anemia (5). In 1980, Botstein and colleagues, building on their use of DNA polymorphisms to study linkage in yeast (6) and the finding of DNA polymorphism at the globin locus in humans (7, 8), proposed the use of naturally occurring DNA sequence polymorphisms as generic markers to create a human genetic map and systematically trace the transmission of chromosomal regions in families (9). The feasibility of genetic mapping in humans was soon demonstrated with the localization of Huntington disease in 1983 (10). A rudimentary genetic linkage map with ~400 DNA markers was generated by 1987 (11) and was extended to ~5000 markers by 1996 (12). Physical maps providing access to linked chromosomal regions were developed by 1995 (13). With these advances, by the 1980s, positional cloning became possible in humans, and the number of disorders tied to a specific gene grew from about 100 to about 2200 today (14).

Several lessons emerged from studies of Mendelian disease genes:

- (i) The “candidate gene” approach was woefully inadequate; most disease genes were completely unsuspected on the basis of previous knowledge,
- (ii) Disease-causing mutations often cause major changes in encoded proteins,
- (iii) Loci typically harbor many disease-causing alleles, mostly rare in the population, and
- (iv) Mendelian diseases often revealed great complexity, such as locus heterogeneity, incomplete penetrance, and variable expressivity.

Geneticists were eager to apply genetic mapping to common diseases, which also show familial clustering. Mendelian subtypes of common diseases: such as breast cancer (15), hypertension (16), and diabetes (17)] were elucidated, but mutations in these genes explained few cases in the population. In common forms of common disease, risk to relatives is lower than in Mendelian cases, and linkage studies with excellent power to detect a single causal gene yielded equivocal results.

These features were consistent with, but did not prove, a polygenic model. The idea that commonly varying traits might be polygenic in nature was offered by East in 1910 (18). By 1920, linkage mapping was used to identify multiple unlinked factors influencing truncate wings in *Drosophila* (19), and Fisher had developed a mathematical framework for relating Mendelian factors and quantitative traits (20). In the late 1980s, linkage mapping of complex traits was made feasible for experimental organisms through the use of genetic mapping in large crosses (21). ***Still, there was not much significant successes in humans.***

Genetic Association in Populations

A possible path forward emerged from population genetics and genomics. Instead of mapping disease genes by tracing transmission in families, one might localize them through association studies—that is, comparisons of frequencies of genetic variants among affected and unaffected individuals.

Genetic association studies were not a new idea. In the 1950s, such studies revealed correlations between blood-group antigens and peptic ulcer disease (4); in the 1960s and 1970s, common variation at the human leukocyte antigen (HLA) locus was associated with autoimmune and infectious diseases (22); and in the 1980s, apolipoprotein E was implicated in the etiology of Alzheimer's disease (23). Still, only about a dozen extensively reproduced associations of common variants (outside the HLA locus) were identified in the 20th century (24).

A central problem was that association studies of candidate genes were a shot in the dark: They were limited to specific variants in biological candidate genes, each with a tiny a priori probability of being disease-causing. Moreover, association studies were susceptible to false positives due to population structure, because there was no way to assess differences in the genetic background of cases and controls. Although many claims of associations were published, the statistical support tended to be weak and few were subsequently replicated (25).

In the mid-1990s, a systematic **genome-wide** approach to association studies was proposed (26–28): to develop a catalog of common human genetic variants and test the variants for association to disease risk. The focus on common variants as a mapping tool was used, based upon population genetics. The human population has recently grown rapidly from a small size. As predicted by classical theory (29), humans have limited genetic variation: The heterozygosity rate for single-nucleotide polymorphisms (SNPs) is ~1 in 1000 bases (30–32). Moreover, approximately 90% of heterozygous sites in each individual are common variants, typically shared among continental populations (33).

If most genetic variation in an individual is common, then why are mutations responsible for Mendelian diseases typically rare? Perhaps one reason is natural selection: Mutations that cause strongly deleterious phenotypes, as most Mendelian diseases appear to be, are lost to purifying selection. ***But if deleterious mutations are typically rare, then how could common variants play a role in disease?***

Common diseases often have late onset, with modest or no obvious impact on reproductive fitness. Mildly deleterious alleles can rise to moderate frequency, particularly in populations that have undergone recent expansion (34). Some alleles that were advantageous or neutral during human development might now confer susceptibility to disease because of changes in living conditions accompanying civilization. Lastly, disease-causing alleles could be maintained at high frequency if they were under balancing selection, with disease burden offset by a beneficial phenotype (as in sickle-cell disease and malaria resistance).

These lines of reasoning led to the so-called “common disease–common variant” (CD-CV) hypothesis: the proposal that common polymorphisms (classically defined as having a minor allele frequency of >1%) might contribute to susceptibility to common diseases (26–28). If that were so, genome-wide association studies (GWAS) of common variants might be used to map loci contributing to common diseases. The concept was not that all causal mutations at these genes should be common (to the contrary, a full spectrum of alleles is expected), only that some common variants exist and could be used to pinpoint loci for detailed study.

It took a decade to develop the tools and methods required to test the CD-CV hypothesis:

- (i) catalogs of millions of common variants in the human population,
- (ii) techniques to genotype these variants in studies with thousands of patients, and (iii)
- (iii) an analytical framework to distinguish true associations from noise and artifacts.

Cataloging SNPs and Linkage Disequilibrium

Pilot projects in the late 1990s showed that it was possible to identify thousands of SNPs and to perform highly multiplexed genotyping by means of DNA microarrays (35). A public-private partnership, the SNP Consortium, built an initial map of 1.4 million SNPs (32); this has grown to more than 10 million SNPs (36) and is estimated to contain 80% of all SNPs with frequencies of >10% (37).

As the SNP catalog grew, a critical question loomed:

Would GWASs require directly testing each of the ~10 million common variants for association to disease?

That is, *if only 5% of variants were tested, would 95% of associations be missed? Or could a subset serve as reliable proxies for their neighbors?* Experience from Mendelian diseases suggested that substantial efficiencies might be possible. Each disease-causing mutation arises on a particular copy of the human genome and bears a specific set of common alleles in cis at nearby loci, termed a haplotype. Because the recombination rate is low [~1 crossover per 100 megabases (Mb) per generation], disease alleles in the population typically show association with nearby marker alleles for many generations, a phenomenon termed Linkage Disequilibrium (LD): Fig. 4.1

Science. Author manuscript; available in PMC 2009 Jun 11.

Published in final edited form as:

Science. 2008 Nov 7; 322(5903): 881–888.

Sample sizes required for genetic association studies. The graphs, Figs. 4.1, 4.2, 4.3 and 4.4, show the total number N of samples (consisting of $N/2$ cases and $N/2$ controls) required to map a genetic variant as a function of the increased risk owing to the disease-causing allele (x -axis) and the frequency of the disease-causing allele (various curves). The required sample size is shown in the table on the right for various different kinds of association studies.

- The first three columns refer to GWASs using common variants across the entire genome; the columns correspond to different levels of statistical power to achieve a significant result at $P < 10^{-8}$.
- The fourth column refers to a search for rare variants where the frequency listed is the collective frequency of rare variants in controls, and the odds ratio is the excess in cases as compared to controls.

Sample sizes assume correction for a genome-wide search of about 20,000 protein-coding genes in the genome (aiming to achieve $P < 10^{-5}$ with one test performed per gene). The fifth column refers to a

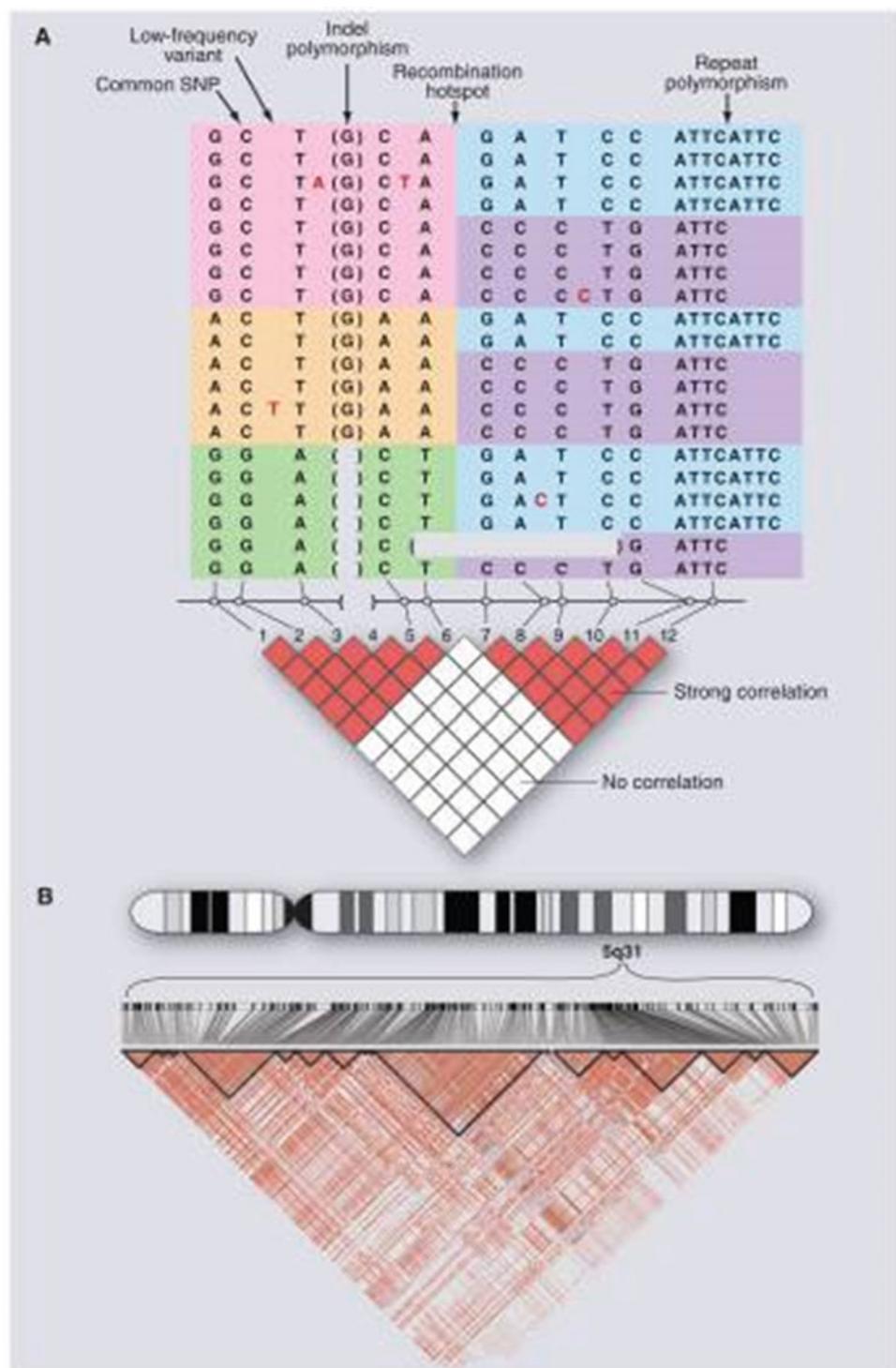


Fig. 4.1 DNA sequence variation in the human genome – 1

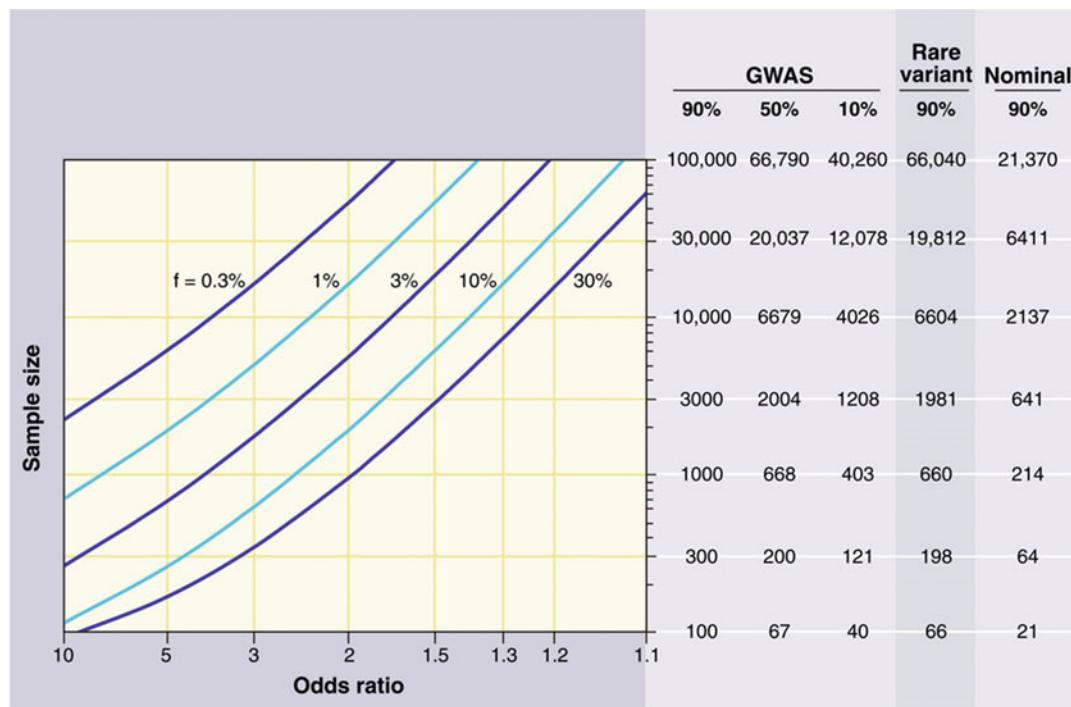


Fig. 4.2 DNA sequence variation in the human genome – 2

test of a single hypothesis (e.g., testing association with a single SNP). For example, in a GWAS, 1000 samples provide 90% statistical power to detect a 30% allele with a factor of 2 effect. In a genome-wide search via exon sequencing, 660 samples provide 90% power to detect a gene in which rare variants have aggregate population frequency 1% and convey a factor of ~ 8 increase in risk. The sample size to test almost all common SNPs in the human genome is only 5 times the sample size to test a single SNP.

Early studies had demonstrated LD of nearby polymorphisms at the globin locus (38), which proved useful in tracking sickle-cell mutation. In the mid-1980s, it was suggested that a genome-wide search might be performed in genetically isolated populations by scanning the genome for a haplotype shared among unrelated patients carrying the same founder mutation (39). Such “LD mapping” in essence treated the entire population as a very large and very old extended family. This method was soon found useful in fine-mapping the founder $\Delta 508$ mutation in the transmembrane conductance regulator CFTR as a cause of cystic fibrosis (40) and in screening the entire genome in isolated populations such as Finland (41).

The key question was whether the same approach could be used more generally to study common alleles in large human populations, where recombination had more time to whittle down haplotypes. A simulation study suggested that LD typically may be too short to be useful with a SNP every 5 kb (500,000 SNPs across the genome) providing very weak LD: average correlation $r^2 = 0.1$ (42). Studies of individual loci showed great heterogeneity in local LD (43).

As denser genetic maps became available, a clearer picture emerged. Nearby variants were observed to form a block-like structure consisting of regions characterized by little evidence for historical recombination and limited haplo-type diversity (44, 45). Within such regions, which soon proved general (46), genotypes of common SNPs could be inferred from a knowledge of only a few empirically determined tag SNPs (45–47). These patterns were shaped by hot and cold spots of recombination in the human genome (48–50), and also as historical population bottlenecks (51).

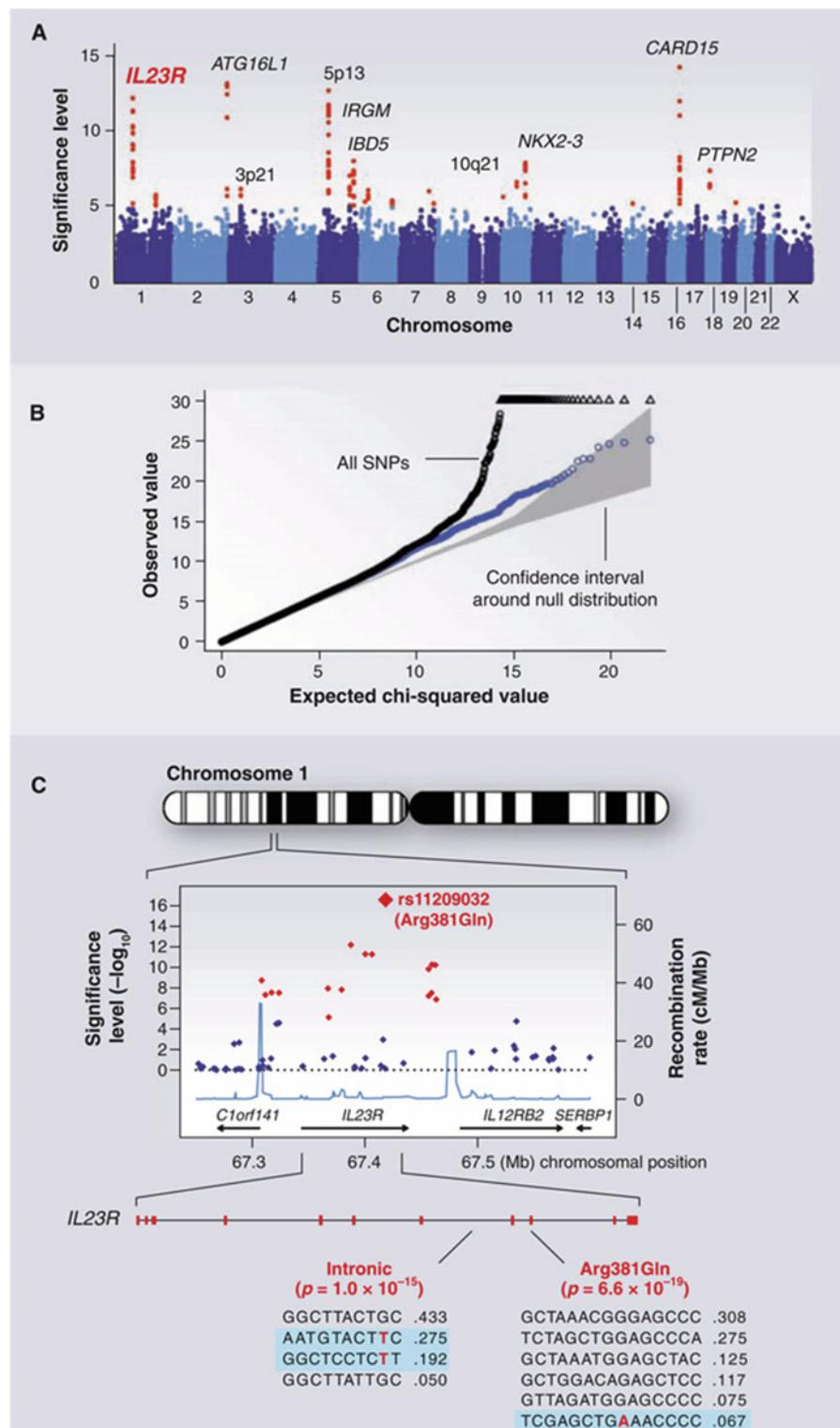


Fig. 4.3 DNA sequence variation in the human genome – 3

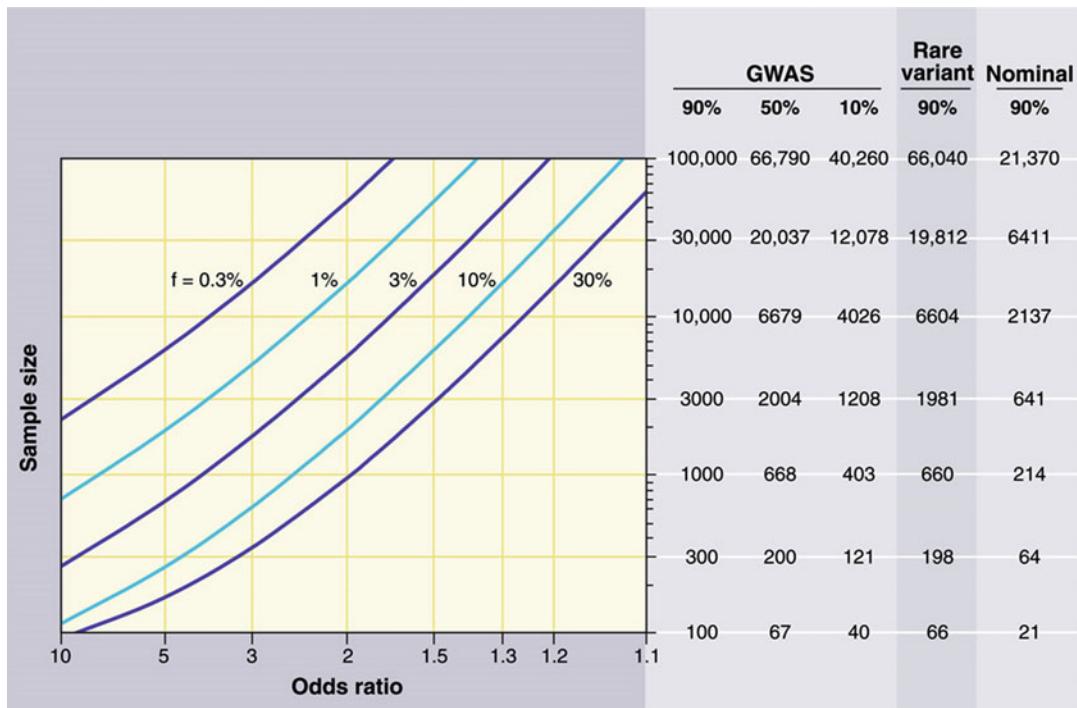


Fig. 4.4 Sample sizes required for genetic association studies

The International HapMap Project was established in 2002, with the goal of characterizing SNP frequencies and local LD patterns across the human genome in 270 samples from Asia, Europe, and West Africa. The project genotyped approximately one million SNPs by 2005 (37) and more than 3 million by 2007 (52). The sequence data collected by the project confirmed that the vast majority of common SNPs are strongly correlated to one or more nearby proxies: 500,000 SNPs provide excellent power to test >90% of common SNP variation in out-of-Africa populations, with about twice that number required in African populations (37).

Improved Parallel Genotyping

SNP genotyping was initially performed one SNP at a time, at a cost of approximately \$1 per measurement. Multiplex genotyping of hundreds of SNPs on DNA microarrays became available in 1998 (35), and the capacity per array grew from 10,000 to 100,000 SNPs in 2002 to 500,000 to 1 million SNPs in 2007! Along with this progress, the associated cost fell to \$0.001 per genotype, or less than \$1000 per sample for a whole-genome analysis! By 2006, the available advancing technologies could simultaneously genotype hundreds of thousands of SNPs at >99% completeness with >99% accuracy.

Copy-Number Variation

SNPs are only one type of genetic variation (Fig. 4.1). Using microarray technology, two groups in 2004 observed that individual copies of the human genome contain large regions (tens to hundreds of kilobases in size) that are deleted, duplicated, or inverted relative to the reference sequence (53, 54). Structural variants had been previously associated with developmental disorders and were often assumed to be pathogenic; the presence of so many segregating copy-number variations (CNVs) in

the general population was surprising. The generality of CNVs was soon established (55–59). Many CNVs display tight LD with nearby SNPs (56, 57) and thus can be proxied by nearby SNPs in GWASs. Others occur in regions that are difficult to follow with SNPs, are highly mutable, or are rare (58, 59). Hybrid genotyping platforms have recently been developed to genotype SNPs and CNVs simultaneously (60).

Biostatistical Analysis

Recognizing causal loci amid a genome's worth of random fluctuation required advances in statistical design, analysis, and interpretation. The risk of false negatives was illustrated by a study of type 2 diabetes (T2D) and the Pro¹² → Ala polymorphism in peroxisome proliferator-activated receptor γ . Whereas an initial positive report (61) had not been confirmed in four modest-sized replication studies, larger studies produced strong and consistent evidence of increased risk by a factor of 1.2 (62, 63). The negative studies were actually consistent with the level of increased risk, but simply lacked adequate power to detect it.

Conversely, stringent thresholds for statistical significance are needed to avoid false positives due to multiple hypothesis testing. Simulations indicated that a dense genome-wide scan of common variants involves the equivalent of \sim 1 million independent hypotheses (64). A significance level of $P = 5 \times 10^{-8}$ thus represents a finding expected by chance once in 20 GWASs. Large sample sizes would be needed to reach such a stringent threshold: Fig. 4.4.

Sample sizes required for genetic association studies. The graphs show the total number N of samples (consisting of $N/2$ cases and $N/2$ controls) required to map a genetic variant as a function of the increased risk due to the disease-causing allele (x ...)

Systematic biases could also cause false positives. Differences in ancestry between cases and controls would yield spurious associations (65), suggesting the need for family-based controls (66). It was later recognized that genome-wide studies provide their own internal control: Mismatched ancestry is readily detectable because it produces frequency differences at thousands of SNPs, which could not all reflect causal associations. Methods were developed to detect and adjust for such biases (67–69) as well as unexpected relatedness between subjects. Technical artifacts, which are particularly problematic if cases and controls are not genotyped in parallel (70), were overcome by improved genotyping methods, quality control, and stringent filtering. To maximize efficiency and power, several groups developed methods of selecting tag SNPs (47, 71–73) from empirical LD data and using them to impute genotypes at other SNPs not genotyped in clinical samples (74) on the basis of LD relationships in the HapMap.

Genome-Wide Associations Development

By early 2006, the tools were in place and studies were under way in many laboratories to resolve the hotly debated issue (75, 76) of whether genetic mapping of common SNPs would shed light on common disease. Since then, scores of publications have reported the localization of common SNPs associated with a wide range of common diseases and clinical conditions (age-related macular degeneration, type 1 and type 2 diabetes, obesity, inflammatory bowel disease, prostate cancer, breast cancer, colorectal cancer, rheumatoid arthritis, systemic lupus erythematosus, celiac disease, multiple sclerosis, atrial fibrillation, coronary disease, glaucoma, gallstones, asthma, and restless leg syndrome) as well as various individual traits (height, hair color, eye color, freckles, and HIV viral set point). Figure 4.3 illustrates data from a paradigmatic genome-wide association study of Crohn's disease performed by the Wellcome Trust Case Control Consortium.

GWAS for Crohn's disease. The panels show data from the study of Crohn's disease by the Wellcome Trust Case Control Consortium. (A) Significance level (P value on \log_{10} scale) for each of the 500,000 SNPs tested across the genome. SNP locations reflect ...

Various lessons have already emerged about genetic mapping by GWAS:

- (1) GWASs work. Before 2006, only about two dozen reproducible associations outside the HLA locus had been discovered (25). However, by early 2008, more than 150 relationships were identified between common SNPs and disease traits (table S1). In most diseases studied, GWASs have revealed multiple independent loci, although some traits have not yet yielded associations that meet stringent thresholds (e.g., hypertension). It is not clear whether this reflects inadequate sample size, phenotypic definition, or a different genetic architecture.
- (2) Effect sizes for common variants are typically modest. In a few cases, common variants with effects of a factor of ≥ 2 per allele have been found: APOE4 in Alzheimer's disease (23), CFH in age-related macular degeneration (77–79), and LOXL1 in exfoliative glaucoma (80). However, in most cases, the estimated effects are much smaller – mostly increases in risk by a factor of 1.1 to 1.5 per associated allele.
- (3) The power to detect associations had been low. Given the effect sizes now known to exist, and the need to exceed stringent biostatistical thresholds, the first wave of GWASs provided low power to discover disease-causing loci (81, 82). For example, achieving 90% power to detect an allele with 20% frequency and a factor of 1.2 effect at a biostatistical significance of 10^{-8} requires 8600 samples (Fig. 4.2). Thus, although it is unlikely that common alleles of large effect have been missed, GWASs of hundreds to several thousand cases have necessarily identified only a fraction of the loci that can be found with larger sample sizes. This prediction has been empirically confirmed in T2D (83), serum lipids (84, 85), Crohn's disease (86), and height (87–90). Across these four traits and diseases, individual GWASs together documented 29 associations. Increasing the power by pooling the samples to perform meta-analysis and replication genotyping has increased this yield to more than 100 replicated loci for these four conditions.
- (4) Association signals have identified small regions for study but have not yet identified causal genes and mutations. *Genetic mapping is a double-edged sword*: local correlation of genetic variants facilitates the initial identification of a region but makes it difficult to distinguish causal mutations. Fortunately, whereas family-based linkage methods typically yield regions of 2 to 10 Mb in span, GWASs typically yield more manageable regions of 10 to 100 kb.

These regions have yet to be scrutinized by fine-mapping and resequencing to identify the specific gene and variants responsible. Even when a locus is identified by SNP association, the causal mutation itself need not be a SNP. For example, the *IRGM* gene was associated with Crohn's disease on the basis of GWAS. Subsequent study suggests that the causal mutation is a deletion upstream of the promoter affecting tissue-specific expression (91).

- (5) A single locus can contain multiple independent common risk variants. Intensive study has already identified:
 - 7 independent alleles at 8q24 for prostate cancer (92),
 - 3 at complement factor H (CFH) for age-related macular degeneration (93, 94),
 - 3 at *IRF5* for systemic lupus erythematosus (95), and
 - 2 at *IL23R* for Crohn's disease (96).

Multiple distinct alleles with different frequencies and risk ratios seems to be the rule.

- (6) A single locus can have both common variants of weak effect and rare variants of large effect. In recent GWASs, studies of common SNPs enabled the identification of 19 loci as influencing low-

or high-density lipoprotein (LDL, HDL) or triglycerides (84, 85). 9 of these 19 were loci for the LDL receptor (LDLR) and familial hypercholesterolemia (FH). Similarly, the genes encoding Kir6.2, WFS1, and TCF2 are all known to cause Mendelian syndromes including T2D, as well as common SNPs with modest effects.

- (7) Because allele frequencies vary across human populations, the relative roles of common susceptibility genes can vary among ethnic groups. One typical example is the association of prostate cancer at 8q24: SNPs in the region play a role in all ethnic groups, but *the contribution is greater in African Americans*. This is not because the risk alleles yet found confer greater susceptibility in African Americans, but because they occur at higher frequencies (92), contributing to the higher incidence among African American men than among men of European ancestry.

The Functions and Phenotypic Associations of Genes Related to Common Diseases

Useful conclusions have also emerged regarding the functions and phenotypic associations of genes related to common diseases:

- (1) A subset of associations involve genes previously related to the disease. Of 19 loci meeting genome-wide significance in a recent GWAS of LDL, HDL, or triglyceride levels, 12 contained genes with known functions in lipid biology (84, 85). The gene for 3-hydroxy-3-methyl glutaryl-coenzyme A reductase (HMGCR), encoding the rate-limiting enzyme in cholesterol biosynthesis and the target of statin medications, was found by GWAS to carry common genetic variation influencing LDL levels (84, 85). Similarly, SNPs in the β -cell zinc transporter encoded by SLC30A8 were associated with risk of T2D (97).
- (2) Most associations do not involve previous candidate genes. In some cases, GWAS results immediately suggest new biological hypotheses—for example, the role of complement factor H in age-related macular degeneration (77–79), FGFR2 in breast cancer (98), and CDKN2A and CDKN2B in T2D (99–101). In many other cases, such as LOC387715/HTRAI with age-related macular degeneration (102), nearby genes have no known function.
- (3) Many associations implicate non-protein-coding regions. Although some associated non-coding SNPs may ultimately be found to be attributable to LD with nearby coding mutations, many are sufficiently far from nearby exons to make this outcome unlikely. Examples include the region at 8q24 associated with prostate, breast, and colon cancer, 300 kb from the nearest gene (103, 104), and the region at 9q21 associated with myocardial infarction and T2D, 150 kb from the nearest genes encoding CDKN2A and CDKN2B (99–101, 105–107).

A role for noncoding sequence in disease risk is not surprising: comparative genome analysis has shown that 5% of the human genome is conserved and thus functional; less than one-third of this 5% consists of genes that encode proteins (108). Non-coding mutations with roles in disease susceptibility will likely open new doors to understanding genome biology and gene regulation. Regulatory variation also suggests different therapeutic strategies: Modulating levels of gene expression may prove more tractable than replacing a fully defective protein or turning off a gain-of-function allele.

- (4) Some regions contain expected associations across diseases and traits. Crohn's disease, psoriasis, and ankylosing spondylitis have long been recognized to share clinical features; the association of the same common polymorphisms in IL23R in all three diseases points to a shared molecular cause (96, 109, 110). SNPs in STAT4 (signal transducer and activator of transcription 4) are associated with rheumatoid arthritis and systemic lupus, two diseases that share clinical features. Multiple variants associated with T2D are associated with insulin secretion defects in nondiabetic individuals (101, 111–116), highlighting the role of β -cell failure in the pathogenesis of T2D.
- (5) Some regions reveal surprising associations. For example, unexpected connections have emerged among T2D, inflammatory diseases (two loci), and cancer (four loci). A single intron of CDKAL1

was found to contain a SNP associated with T2D and insulin secretion defects (99–101, 116), and another with Crohn’s disease and psoriasis (117). A coding variant in glucokinase regulatory protein is associated with triglyceride levels and fasting glucose (101) but also with C reactive protein levels (118, 119) and Crohn’s disease (86). A SNP in TCF2 is associated with protection from T2D, as expected on the basis of Mendelian mutations at the same gene (120). Unexpectedly, the same association signal turned up in a GWAS for prostate cancer (121). Similarly, JAZF1 was identified as containing SNPs associated with T2D (83) and prostate cancer (122), and TCF7L2 with T2D (123) and colon cancer (124, 125).

From Common SNPs to the Full Allelic Spectrum

The current HapMap provides reliable representations for the vast majority of SNPs at frequencies above 5%, but its coverage declines rapidly for lower-frequency alleles (37). Such lower-frequency alleles may be important: Alleles with strong deleterious effects are constrained by natural selection from becoming too common. One may divide these alleles into two conceptually distinct classes:

- (1) Common variants with frequencies below 5%. By “common,” one refers to variants that occur at sufficient frequency to be cataloged in studies of the general population and measured (directly, or indirectly through LD) in association studies. In practice, this class may include allele frequencies in the range of 0.5% and above. A GWAS of 2000 cases and 2000 controls provides good power for a 1% allele causing a factor of 4 increase in risk (even at $P < 10^{-8}$) (Fig. 4.2).

The value of lower-frequency common variants is illustrated by PCSK9 (proprotein convertase subtilisin/kexin type 9). The gene encoding PCSK9 contains very rare mutations causing autosomal dominant hypercholesterolemia (discovered by linkage analysis), as well as high-frequency common variants with modest effects. The former are rare, and the latter too weak, to enable effective clinical study of PCSK9 with respect to coronary artery disease risk. Hobbs and Cohen sequenced the gene (126, 127) and identified low-frequency common variants (0.5 to 1%), which allowed epidemiological research documenting a protective effect on myocardial infarction (128).

- (2) Rare variants. Most Mendelian diseases involve rare mutations that are essentially never observed in the general population. Rare mutations likely also play an important role in common diseases. Because they are numerous and individually rare, it is not possible to create a complete catalog in the general population. Instead, they must be identified by sequencing in cases and controls in each study. Moreover, because each variant is too rare to prove statistical evidence of association, the mutations must be aggregated as a class to compare the overall frequency of cases versus controls.

A few examples are known through candidate gene studies. Rare nonsynonymous mutations in MC4R are found in patients with extreme early-onset obesity (129). Rare nonsynonymous mutations in ABCA1 are more common in patients with extremely low HDL than in those with high HDL (130). An excess of rare mutations in renal salt-handling genes has been associated with lower blood pressure and protection against hypertension (131).

The sample size required to perform a genome-wide search based on coding mutations depends on the background frequency (μ) of mutations that confer disease risk and the level (ω) of increased risk for each such mutation. ABCA1 is a favorable case because μ and ω are high (the gene has an unusually large coding region of ~7 kb, and mutations confer a factor of ~6 increase in risk). Achieving genome-wide significance will likely require resequencing studies of thousands of cases and controls, similar to GWASs (Fig. 4.2).

GWASs of rare variants are under way for large structural variants through the use of micro-array analysis. A recent GWAS of autism showed that a highly penetrant, recurrent microdeletion and micro-

duplication of a 593-kb region in 16p11.2 explains 1% of cases (132). Moreover, some recent studies report that patients with autism and schizophrenia may have an excess of rare deletions across the genome relative to unaffected controls (133,134). Although these studies did not identify specific loci (none of the novel loci were observed more than once), they suggest that the universe of rare structural changes contributing to each disease may be as large and diverse as that of common SNPs.

The Genetic Architecture of Common Disease

Variants so far identified by GWASs together explain only a small fraction of the overall inherited risk of each disease (for example, ~10% of the variance for Crohn's and ~5% for T2D).

Where is the remaining genetic variance to be found?

There are several answers:

- (1) At disease loci already identified by GWAS, the locus-attributable risk will most likely be higher than currently estimated. This is owing to the marker SNPs used in GWASs which will typically be imperfect proxies for the actual causal mutation that led to the association signal. The causal gene will often contain additional mutations not tagged by the initial marker SNPs, both common and rare. Determining the contribution of each gene will require intensive studies of variants at each locus.
- (2) Many more disease loci remain to be identified by GWAS. As noted previously, GWASs to date have had low statistical power and thus necessarily missed many loci with common variants of similar and smaller effects. The first studies did not have proxies for common structural variants and have failed to capture lower-frequency common variants (0.5 to 5%). Moreover, the vast majority of studies have been performed only in case subject samples of European ancestry. Larger, more comprehensive, and more diverse GWASs may reveal many more loci.
- (3) Some disease loci will contain only rare variants. Such loci cannot be identified by the study of common variants alone. They will require systematic resequencing of all genes in large samples (Fig. 4.2).
- (4) Current estimates of the variance explained are based on simplifying assumptions. Owing to the genotype-phenotype correlation has yet to be well established, the estimates assume that the variants interact in a simple additive manner. Nevertheless, gene-gene and gene-environment interactions play important roles in disease risk. Although searches have not yet found much evidence for epistasis [e.g., (93, 94, 135)], this may simply reflect limited power to assess the many possible modes of interaction, including pairwise interactions and threshold effects. Once patterns of association and interaction are understood, effects of specific gene and environmental exposures on each phenotype may be larger.

For these reasons, it is premature to make inferences about the overall genetic architecture of common disease. Only by systematically exploring each of these directions will a general picture emerge—with the likely outcome being that different diseases will each be characterized by a different balance of allele frequencies, interactions, and types. Although the proportion of genetic variance explained is certain to grow in the future, it is unlikely to approach 100% because of practical limitations, such as the difficulty of detecting common variants with extremely small effects, genes harboring rare variants at very low frequency, and complex interactions among genes and with the environment.

Disease Risk Versus Disease Mechanism

The primary value of genetic mapping is not risk prediction, but providing novel insights about mechanisms of disease. A knowledge of disease pathways (not limited to the causal genes and mutations) may suggest strategies for prevention, diagnosis, and therapy. From this perspective, the frequency of a genetic variant is not related to the magnitude of its effect, nor to the potential clinical value that may be obtained.

The classic example is Brown and Goldstein's studies of FH, which affects ~0.2% of the population and accounts for a tiny fraction of the heritability of LDL and myocardial infarction. Studies of FH led to the discovery of the LDL receptor and supported the development of HMGCR inhibitors (statins) for lowering LDL, the use of which is not limited to FH carriers.

Recently, GWASs have shown that common genetic variation in LDLR and HMGCR influences LDL levels (84, 85). Although SNPs in HMGCR have only a small effect (~5%) on LDL levels, drugs targeting the encoded protein decrease LDL levels by a much greater extent (~30%). This is due to the effect of an inherited variant is limited by natural selection and pleiotropy, whereas the effect of a drug treatment is not!

Looking at the Future

Given the long-standing success of genetic mapping in providing new insights into biology and disease etiology, and the recent proof that systematic association studies can identify novel loci, ones aim should be the identification of all pathways at which genetic variation contributes to common diseases. The key steps in achieving this goal are as follows:

(1) Expanding clinical studies

Current studies are underpowered for the types of SNP alleles that one now knows exist, and available evidence indicates that increasing sample size will yield substantial returns. A study of 1000 cases and 1000 controls provides only 1% power to detect a 20% variant that increases risk by a factor of 1.3, but a study of 5000 cases and 5000 controls provides 98% power (Fig. 4.2). Moreover, early data on rare single-nucleotide (130, 131) and structural variants (133, 134, 136) indicate that similarly large samples will be needed to achieve the levels of statistical significance required to detect rare events in a genome-wide search.

Almost all GWASs available to date have been performed in populations of European ancestry! Even if a variant has the same effect in all ancestry groups, it may be more readily detected in one population simply because it happens to have higher frequency. Genetic effects will likely vary across groups because of modification by environment and behavior, which may vary more across groups than does genotype.

Many important diseases remain to be studied by GWAS. Disease-related intermediate traits can also offer substantial insight, particularly in conjunction with clinical endpoints. For example, newly described variants on Chromosome-1 (near SORT1) are associated both with levels of LDL cholesterol (84, 85) and with risk of myocardial infarction (106); this provides not only increased statistical confidence, but also a biomarker for gene function and patho-physiological insight. Genetic variants that influence gene expression [e.g. (137)] hold promise for elucidating regulatory pathways. For example, mapping of modifiers of Mendelian mutation genes that influence the age of onset in carriers of BRCA1 and BRCA2 mutations – may suggest ways to reverse high risk owing to mutations.

Correlations between genetic variants and phenotypes are limited by the accuracy with which each is measured. The ability to measure genotype now far exceeds our ability to measure phenotype. Continuous ambulatory monitoring, imaging methods, and comprehensive approaches to biological samples all have promise in improving the accuracy of phenotype measurement.

Environmental exposures do play a larger role in human phenotypic variation than does genetic variation, but environmental exposures are fundamentally more difficult to measure. DNA is stable throughout life, with a single physical chemistry that enables generic approaches to measurement. Environmental exposures are heterogeneous and may be changing. Improved methods for measuring environmental exposures, perhaps based on epigenetic marks they leave, are needed.

(2) Expanding the range of genetic variation

The simplest beneficial approach is to re-sequence loci that have been definitively implicated in disease by Mendelian genetics or by GWAS. Since the prior probability of a true association is higher, such regions will be the best setting to develop methods for understanding the statistical significance and biological importance of rare mutations. Initially, resequencing of coding exons will be easiest to interpret. Rare coding mutations with large effect will be especially valuable, because physiological studies of mutation carriers can help illuminate the biological basis of the disease, and because coding mutations of large effect are more straightforwardly transferred to cellular and animal models for further mechanistic studies.

Extending GWASs to include structural variants and lower-frequency common variants will require comprehensive catalogs of genomic variation, as well as characterization of LD relationships. With new massively parallel sequencing technologies, an accurate map of all 1% alleles (both single-nucleotide and structural) may be achievable. To this end, a “1000 Genomes Project” was recently launched (138).

Some loci may harbor neither common variants nor rare structural variants, and thus will be missed by array and LD-based approaches. Discovering such genes will require sequencing in thousands of cases and controls. Initial studies will likely focus on exons, where functional mutations are enriched to the greatest extent. Highly parallel methods to capture hundreds of thousands of exons, and other targets of interest, are under development (139).

Multiple instances of new coding mutations at a locus (by comparing affected individuals with parents) could provide particularly powerful association information, because the human mutation rate is so low (in the range of 10^{-8}). However, identifying *de novo* mutations without being overwhelmed by false positives will require extraordinary sequencing accuracy (far better than finished genome sequence). As such studies will be expensive at first, priority should go to disorders with high heritability, where there is an unmet medical need, and for which other approaches have met with limited success. Psychiatric disorders might be considered as one such target.

Eventually, it will become practical to re-sequence entire genomes from thousands of cases and controls. The challenge of interpretations persist, because it is unclear either how to aggregate elements to achieve a large enough target size, or to develop ways to recognize function-altering changes.

Routine genome sequencing of deeply phenotyped cohorts will fundamentally change the nature of genetic mapping: from the current serial process (in which initial localization by linkage or GWAS is followed by scrutiny of DNA variation and phenotypes) to a joint estimation procedure combining variation information of all types, frequencies, and phenotypes to discover and characterize genotype-phenotype correlations. New biostatistical methods will be required to combine evidence from rare and common alleles at a locus and across multiple loci, phenotypes, and non-genetic exposures. A particular challenge will be to identify mutations in regions without known function or evolutionary conservation.

There may be inherent limits to ones ability to relate phenotypic variation and genotypic variation. To the extent that disease is influenced by tiny effects at hundreds of loci or highly heterogeneous rare mutations, it may well be impractical to assemble sufficiently large samples to give a complete accounting.

(3) Implications for Biology, Medicine, and Society

Genetic mapping is only a first step toward biological understanding and clinical application. Useful tools will include maps of evolutionary conservation (108) and chromatin state (140), as well as databases of cell-state signatures, such as genome-wide expression patterns, that may integrate aspects of cell biology under resting and provoked conditions (141). Creation of disease models, both in human cell culture and non-human animals, will be key. Physiological studies in patients classified by genotype may inform disease processes and lead to useful non-genetic bio-markers. Given the limits of human clinical research, rare alleles of strong effect may be more useful than common alleles of weak effect.

The high failure rate of clinical trials testifies to the limited predictive value of current approaches. **By focusing attention on genes and processes, human genetics has the potential to yield productive targets and predictive animal models. In clinical trials, the ability to stratify patients by genotype or biological pathway may reveal differences in therapeutic response.** Genetics may also increase the efficiency of outcome trials by focusing on patients at higher-than-average risk.

The extent to which genetic information will figure in “personalized medicine” will depend on:

- (i) whether predictive accuracy beyond conventional measures can be attained, and
- (ii) whether there are interventions whose effectiveness is improved by knowledge of a genetic test.

Knowledge of a common variant that increases T2D risk by 20% may eventually lead to new understanding and therapeutic strategies, but whether an increase in absolute risk (from 8% to 10%) is useful for patients remains to be seen. ***Although the knowledge of individual risk might promote greater adherence to a healthy lifestyle, human behavior is complex and risk estimates are challenging to interpret.*** Even where genotype can predict response to a drug with a narrow therapeutic window, it cannot be assumed that genetic testing will necessarily lead to improved clinical outcomes.

Our understanding of complex disease will be in constant flux. The pace of discovery, while scientifically exhilarating, poses daunting challenges. Direct-to-consumer marketing of genetic information is already under way. It will be a challenge for the public to understand the difference between relative and absolute risk, and to figure in their thinking the larger component of genetic and environmental factors not yet captured by today’s technologies. Rigorous assessment of health benefit and cost are needed, including costs of testing and treatment that may flow from an altered sense of risk. As genetic information is shown to be useful, equitable access will be critical.

Finally, one should ensure that the promise of research on genetic factors in complex disease does ***not*** encourage a mistaken sense of genetic determinism. This is important for behavioral traits, which are especially prone to misinterpretation and misguided policy. One must constantly remind the public, and oneself, that although genes play a role (and can lead one to new biological insight), ones traits are powerfully shaped by the environment, and the solutions to important problems will often lie outside ones genes.

4.2 Human Genetic Influences on Diseases

There are, at least, 2 approaches to describe and characterize whether or not there are genetic influences on a given disease, or at least some risk factor of interest:

- (1) using familial aggregation studies, and
- (2) estimating heritability by twin studies.

4.2.1 Genetic Relationships in a Familial Aggregation^[*]

[*] Thomas, D. C. (2004).- “Statistical Methods in Genetic Epidemiology”. Oxford University Press, New York, NY”

In a familial aggregation, genetic relationships are biostatistically governed by the probability that, at any autosomal locus, any 2 members of a pedigree may share 1 or 2 alleles from the same source.

Now consider 2 case subjects, q and r , in family p . Each case subject has 2 alleles: and they may share *none*, *one*, or *both* of them, and this sharing can be assessed either by descent or by state:

- 2 alleles are **Identical-By-Descent, IBD**, if they are derived from a common ancestor
- 2 alleles are **Identical-By-State, IBS**, if they are identical in terms of their DNA function and composition, but do not necessarily derived from a common ancestor.

Genetic relationships depend on their **IBD** status. Unilineal relatives are related through one side of the family: for example:

Unilineal relatives are related via one side of the family, for example: parent-child, half siblings, first cousins, etc., and they can share only zero or one allele: Fig. 4.5 (2): **IBD**. On the other hand, bilineal relatives are related through both paternal and maternal lines, for example: full sibling, double first cousins, etc., and then can share two, one or no alleles: Fig. 4.5 (1): **IBD**.

An Example

First consider a pair of siblings, and their IBD status: let the parents' alleles be labeled as: a, b , and c, d . Some of these may be IBS. However, not knowing the genotypes of their ancestors, it is not possible to determine whether they are IBD, and, for this exercise, one may distinguish them. The possible configurations of the IBD status for any 2 offsprings are tabulated as follows, in Table 4.1:

In Table 4.1, each of the 16 possible genotype pairs is **equally** likely, it is evident that by simply tabulating the number of possibilities by the number of alleles shared IBD(Z) that the probability P of sharing 0, 1, or 2 alleles IBD is:

$$P_0 = \frac{1}{4}, \quad P_1 = \frac{1}{2}, \quad P_2 = \frac{1}{4}$$

respectively. And the expected number of alleles shared IBD is:

$$P_1 + 2P_2 = \frac{1}{2} + 2\left(\frac{1}{4}\right) = 0.5 + 2 \times 0.25 = 0.5 + 0.5 = 1$$

Fig. 4.5 Examples of Unilinear and Bilinear relationships

(1) Bilinear Relationship:

parent-children => full siblings: **Child A and Child B**



(2) Unilinear Relationships:

(a) half-brother/sister=> half siblings: **Child a and Child b**



(b) First cousins => **Child X and Child Y**

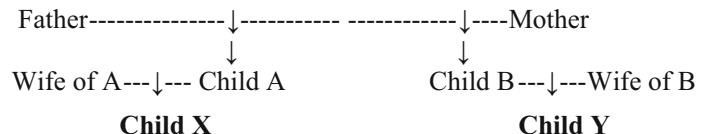


Table 4.1 All possible genotypes of a sibling pair by *Identical By Descent (IBD)* status assuming that their parents have genotypes *ab* and *cd*

Number of alleles shared IBD	Offspring genotypes	Probability
0	(<i>ac</i> , <i>bd</i>) (<i>ad</i> , <i>bc</i>) (<i>bc</i> , <i>ad</i>) (<i>bd</i> , <i>ac</i>)	$\frac{1}{4}$
1	(<i>ac</i> , <i>ad</i>) (<i>bc</i> , <i>bd</i>) (<i>ac</i> , <i>bc</i>) (<i>ad</i> , <i>bd</i>)	
2	(<i>ad</i> , <i>ac</i>) (<i>bd</i> , <i>bc</i>) (<i>bc</i> , <i>ac</i>) (<i>bd</i> , <i>ad</i>)	$\frac{1}{2}$
	(<i>ac</i> , <i>ac</i>) (<i>ad</i> , <i>ad</i>) (<i>bc</i> , <i>bc</i>) (<i>bd</i> , <i>bd</i>)	$\frac{1}{4}$

or equivalently, the Expected Proportion P of alleles shared IBD is $\frac{1}{2}$ (since each individual has 2 alleles).

This proportion, called the **Coefficient of Relationship**, equals 2^{-R} , where R is the

Degree of Relationship, viz.,

$$R = \left\{ \begin{array}{l} \downarrow \\ 0, \text{ for Identical (MZ) Twins} \\ 1, \text{ for First - Degree Relatives (siblings or parent - offspring pairs)} \\ 2, \text{ for Second - Degree Relatives (grandparents, half - siblings, Uncles and aunts), and} \\ 3, \text{ for Third - Degree Relatives (first cousins, etc.)} \\ 4, \text{ etc.} \end{array} \right.$$

4.2.2 Familial Risk of Diseases

Definition: **Familial Risk (FR)** is the *probability* that a case subject will be affected by a disease.

Definition: **Familial Relative Risk (FRR)** is the ratio of the risk of the case subject affected by the disease *for being a member of the family*, to the risk of one being affected by the disease *in the general population*.

Clearly, FR depends on:

- the gender, age, year-of-birth of the case subject,
- other significant risk factors such as: race, host factors, environmental factors,
- the number and types of relatives similarly affected,
- the number and ages of relatives at risk.

Remarks

1. Within the definition of FRR, it is to be understood that the 2 risks being compared are indeed comparable with respect to the factors of interests.
2. The magnitude of the relative risk may be modified by some of the foregoing factors.
3. Perhaps, a more suitable comparison than lifetime risks may be age-specific incidences, or mortality rates.
4. It is critical to distinguish the concept of *familial risk* from *genetic risk*. Genetic risk is the probability that a case subject will be affected by a disease, given his/her genotype. This may be defined as lifetime risks or age-specific incidence rates, as relative risks when comparing genotypes *AA* and *aa* for a single major locus.

4.2.2.1 Derivation of Familial Risk

For simplicity, consider the computations for a single genetic locus (similar calculations may be undertaken for multiple loci), with Fig. 4.6, in which *k* is affected, and one wishes to calculate the risk to Case Subject *j*. This is a simple DAG (Directed Acyclic Graph) used to conceptualize further computational procedures. In this DAG, one wishes to calculate the risk to Subject *p*, given that *q* is affected:

Genetic models are usually expressed in terms of three components:

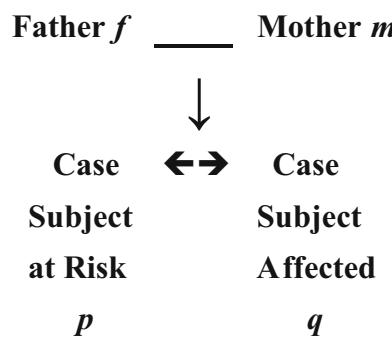
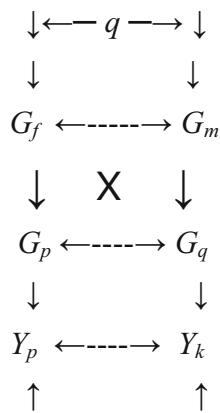


Fig. 4.6 A simple nuclear family with 2 offsprings affected: The Family Risk of the other offspring is to be calculated

Fig. 4.7 Directed Acyclic Graph (DAG) for the pedigree



1. the population genotype probabilities for unrelated individuals,
2. the transition probabilities for the relationship between parental and offspring genotypes, and
3. the penetrances.

These three components may be represented in the top, middle, and bottom portions of Fig. 4.7 (Morton 1955):

In Fig. 4.7:

- (a) Y_p and Y_k are Observed Quantities
- (b) the other 6 represent unobserved latent variables or model parameters
- (c) the line between G_f and G_m indicates the assumption of no assortive mating
- (d) the line between G_f and G_g indicates the assumption of independent segregation of offspring genotypes given the parental genotypes, and
- (e) the line between Y_p and Y_k indicates the assumption of conditional independence of phenotypes given the genotypes

Application of the Hardy-Weinberg Equilibrium

Let q be the frequency of the A allele, then $p = (1 - q)$ is the frequency of the a allele, and Q_g the genotype probabilities under Hardy-Weinberg Equilibrium

Then

$$Q_0 = \Pr(G = aa) = p^2 \quad (4.1)$$

$$Q_1 = \Pr(G = aA) = 2pq \quad (4.2)$$

$$Q_2 = \Pr(G = AA) = q^2 \quad (4.3)$$

Now, let $Y_i = 1$ indicate that individual i is affected, and

$Y_i = 0$ indicate that individual i is unaffected, and

let $f_g = \Pr(Y = 1 \mid G = g)$ denote the generic risk (viz., *penetrance*) for genotypes: $g = aa, aA, AA$.

Table 4.2 Transition Probabilities for Offspring's Genotype, Conditional on Their Parental Genotypes

Father's genotype	Mother's genotype	Offspring's genotype		
		dd	dD	DD
dd	dd	1	0	0
dd	dD	½	½	0
dd	DD	0	1	0
dD	dd	½	½	0
dD	dD	¼	½	¼
dD	DD	0	½	½
DD	dd	0	1	0
DD	dD	0	½	½
DD	DD	0	0	1

Finally, let T_{gjgngf} denote the transition probabilities

$$\Pr(G_j = g_j | G_m = g_m, G_f = g_f) \quad (4.4)$$

given by the following table: Table 4.2:

Consider the risk to an individual j , given the disease status Y_k of single sibling k :

$$\Pr(Y_j = 1 | Y_k = 1) = \Pr(Y_j = 1, Y_k = 1) / \Pr(Y_k = 1) \quad (4.5)$$

Now consider the numerator of the RHS of Eq. (4.5):

$$\Pr(Y_j = 1, Y_k = 1) = \sum_{gj} \sum_{gk} \Pr(Y_j = 1, Y_k = 1, G_j = g_j, G_k = g_k) \quad (4.6)$$

$$= \sum_{gj} \sum_{gk} \Pr(Y_j = 1, Y_k = 1 | g_j, g_k) \Pr(g_j, g_k) \quad (4.7)$$

$$= \sum_{gj} \sum_{gk} \Pr(Y_j = 1 | g_j, g_k) \Pr(Y_k = 1 | g_j, g_k) \Pr(g_j, g_k) \quad (4.8)$$

$$= \sum_{gj} \sum_{gk} \Pr(Y_j = 1 | g_j) \Pr(Y_k = 1 | g_k) \Pr(g_j, g_k) \quad (4.9)$$

$$= \sum_{gj} \sum_{gk} f_{gj, gk} \Pr(g_j, g_k) \quad (4.10)$$

Remarks

1. In the above collection of equations, beginning in Eq. (4.6) and for notational simplicity, “ $G_j = g_j$ ” has been written simply as g_j .
2. In the above derivation, the first step involves summing all the joint probabilities of the two genotypes and of the two phenotypes over all possible genotype combinations – to give Eq. (4.7).
3. Next, the joint probability function is decomposed into the conditional probability of phenotypes giving genotypes times the genotype probability function: Eq. (4.8).
4. These first 2 steps require only the classical laws of total and conditional probabilities, making no genetic assumptions.
5. The third step depends on the assumption that the 2 siblings' phenotypes are conditionally independent. This strong assumption is dependent on assuming that there are no other shared risk factors, such as genes or environmental factors, that influence both of their phenotypes. All these result in Eq. (4.9).

6. The final step, Eq. (4.10) is obtained by assuming that each sibling's phenotype depends only on his/her own genotype but not that of his/her sibling.

Next, by considering the sum over all possible parental genotypes, one may evaluate the last term in Eq. (4.10):

$$\Pr(g_j, g_k) = \Pr(G_j = g_j, G_k = g_k) \quad (4.11)$$

$$= \sum_{gm} \sum_{gj} \Pr(g_j, g_k | g_m, g_f) \Pr(g_m, g_j) \quad (4.12)$$

$$= \sum_{gm} \sum_{gj} \Pr(g_j | g_m, g_f) \Pr(g_k | g_m, g_f) \Pr(g_m, g_f) \quad (4.13)$$

$$= \sum_{gm} \sum_{gj} \Pr(g_j | g_m, g_f) \Pr(g_j | g_m, g_f) \Pr(g_m) \Pr(g_j) \quad (4.14)$$

$$= \sum_{gm} \sum_{gj} T_{gjgmgf} T_{gkgmgf} \Pr(g_m) Q_{gm} Q_{gf} \quad (4.15)$$

Remarks

1. Equation (4.11) is simply an expansion of the definition of the probability function.
2. Equation (4.12) is also based on the laws of total and conditional probability, as used previously.
3. Equation (4.13) is based upon the assumption that the genotypes of the offsprings are conditionally independent given their parents' genotypes, which is simply an expression of the Mendel's Law of Independent Segregation.
4. Equation (4.14) is based upon the assumption of random mating – i.e., that parental genotypes are independent. This is indeed a strong assumption, which is unlikely to be strictly in practice because allele frequencies vary by ethnic group, and couples tend to marry within their ethnic groups. Nevertheless, this assumption is very commonly accepted in genetic analysis!

Lastly, consider the denominator of the expression given by Eq. (4.5), viz., the population risk:

$$\Pr(Y_k = 1) = \sum_g \Pr(Y_k = 1 | G_k = g_k) \Pr(G_k = g) \quad (4.16)$$

$$= \sum_g f_g Q_g \quad (4.17)$$

$$= f_0 p^2 + 2f_1 p q + f_2 q^2 \quad (4.18)$$

Combining all the foregoing relevant equations, one may compute the **Sibling Recurrent Risk** as:

$$\Pr(Y_j = 1 | Y_k = 1) = \frac{\sum_{gm} \sum_{gj} \sum_{gk} f_{gj} f_{gk} T_{gjgmgf} T_{gkgmgf} Q_{gm} Q_{gf}}{\sum_{gj} f_{gj} Q_{gj}} \quad (4.19)$$

Relative Risks (RR)^[W]

In biostatistics and epidemiology, **Relative Risk** or **Risk Ratio (RR)** is the ratio of the probability of an event occurring (for example, developing a disease, being injured, etc.) in an exposed group to the probability of the event occurring in a comparison, non-exposed group. Relative risk includes two important features:

- (i) a comparison of risk between two “exposures” puts risks in context, and
- (ii) “exposure” is ensured by having proper denominators for each group representing the exposure.

Consider an example where the **probability** of developing lung cancer among smokers was 20% and among non-smokers 1%. This situation is expressed in the 2×2 table below:

Here, $a = 25$, $b = 75$, $c = 2$, and $d = 98$. Then the relative risk of cancer associated with smoking would be

$$\begin{aligned}
 \text{RR} &= \{a/(a+b)\}/\{c/(c+d)\} \\
 &= \{25/(25+75)\}/\{2/(2+98)\} \\
 &= (25/100)/(2/100) \\
 &= 25/2 \\
 &= 12.5
 \end{aligned}$$

That is, smokers would be twelve and a half times as likely as non-smokers to develop lung cancer!

The alternative term **Risk Ratio** is also used because it is the ratio of the risk in the exposed to the risk in the unexposed.

Remarks

1. Relative risk is used frequently in the biostatistical analysis of binary outcomes where the outcome of interest has relatively low probability. It is often used in clinical trial data, where it is used to compare the risk of developing a disease, in people not receiving the new medical treatment (or receiving a placebo) versus people who are receiving an established treatment. Alternatively, it is used to compare the risk of developing a side effect in people receiving a drug as compared to the people who are not receiving the treatment (or receiving a placebo). It is particularly attractive because it can be calculated by hand in the simple case, but is also amenable to regression modelling, typically in a Poisson regression application.
2. In a simple comparison between an experimental group and a control group:
 - (a) A relative risk (RR) of 1 means there is no difference in risk between the two groups.
 - (b) $\text{RR} < 1$ means the event is less likely to occur in the experimental group than in the control group.
 - (c) $\text{RR} > 1$ means the event is more likely to occur in the experimental group than in the control group.
3. When the event is not necessarily an adverse one, the term *Relative Probability* may be used instead.
4. The **Odds Ratio** (OR) may be viewed in similar terms as follows

Relative risk is different from the odds ratio, although it *asymptotically approaches it for small probabilities*. In the above example of association of smoking to lung cancer considered, if a is substantially smaller than b , then $a/(a+b) \approx a/b$. And, similarly, if c is much smaller than d , then $c/(c+d) \approx c/d$. Thus

$$\text{RR} = \{a/(a+b)\}/\{c/(c+d)\} \approx (a/b)/(c/d) = ad/bc = \text{OR, the Odds Ratio}$$

The odds ratio has much wider use in biostatistics, since logistic regression, often associated with clinical trials, works with the logarithm of the odds ratio, *not* the relative risk. Since the log of the odds

ratio is estimated as a linear function of the explanatory variables, the estimated odds ratio for 70-year-olds and 60-year-olds associated with type of treatment would be the same in a logistic regression models where the outcome is associated with drug and age, although the relative risk might be significantly different. In cases like this, ***statistical models of the odds ratio often reflect the underlying mechanisms more effectively.***

Since relative risk is a more intuitive measure of effectiveness, the distinction is important especially in cases of medium to high probabilities. If action A carries a risk of 99.9% and action B a risk of 99.0% then the relative risk is ***just over 1***, while the odds associated with action A are ***more than 10 times higher*** than the odds with B.

In epidemiological research, the odds ratio is commonly used for case-control studies, as odds, but not probabilities, are usually estimated (Ozaki et al. 2002). Relative risk is used in randomized controlled trials and cohort studies (Elston and Stewart 1971).

In statistical modelling, approaches like Poisson Regression (for counts of events per unit exposure) have relative risk interpretations: the estimated effect of an explanatory variable is multiplicative on the rate, and thus leads to a risk ratio or relative risk. Logistic regression (for binary outcomes, or counts of successes out of a number of trials) should be interpreted in odds-ratio terms: the effect of an explanatory variable is multiplicative on the odds and thus leads to an odds ratio.

Familial Relative Risks (FRR)

Genetic Relative Risks (GRR)

An Illustrative Example on Population Risk

Calculation of the Population Risk for a Partially Penetrant Dominant Gene and Penetrance

DATA: Penetrance

$f_1 = f_2 = 0.9$

Sporadic Case Rate of $f_0 = 0.1$

Allele Frequency of $q = 0.0$

Remarks

1. Table 4.4 shows the computation of the population risk for a *single* gene model with a specified selection of parameters.
2. The computations of sibling recurrence risk are complex, but may done by hand: using a spreadsheet, or by machine computation – described and illustrated in Chapter 5: using R.
3. Alternate relative measures include:

Familial Risk Ratio:

$$R_r = \left\{ \Pr(Y_j = 1 | Y_k = 1) / \Pr(Y_j = 1 | Y_k = 0) \right\} \quad (4.20)$$

Familial Odds Ratio: This quantity may be estimated by the case-control design, as illustrated in Table 4.3

$$R_o = \frac{\left\{ \Pr(Y_j = 1 | Y_k = 1) / \Pr(Y_j = 0 | Y_k = 1) \right\}}{\left\{ \Pr(Y_j = 1 | Y_k = 0) / \Pr(Y_j = 0 | Y_k = 0) \right\}} \quad (4.21)$$

Table 4.3 Table for determining Relative Risk (RR)

Risk	Disease status	
	Present	Absent
Smoker	a	b
Non-Smoker	c	d

Table 4.4 Computation for the Population Risk for a Partially Penetrant

Genotype	$Q_g = \Pr(g)$	$fg = \Pr(Y = 1 g)$	Product
aa	$(0.99)^2 = 0.9801$	0.1	$(0.9801) \times (0.1) = 0.09801$
aA	$2x(0.01)x(0.99) = 0.0198$	0.9	$(0.0198) \times (0.9) = 0.01782$
AA	$(0.01)^2 = 0.0001$	0.9	$(0.0001) \times (0.9) = 0.00009$
Total	1.0000		0.11592

Dominant Gene with: Penetrance $f_1 = f_2 = 0.9$
a Sporadic Case Rate of $f_0 = 0.1$, and
an Allele Frequency of $q = 0.01$

Risk Relative to the General Population:

$$R_g = \Pr(Y_j = 1 | Y_k = 1) / P = (Y_j = 1) \quad (4.22)$$

This parameter is usually denoted by λ_s , when applies to siblings, λ_m , when applies to MZ twins, and λ_o , when applies to offspring.

- Each of these 3 quantities (**Familial Risk Ratio**, **Familial Odds Ratio**, and **Risk Relative to the General Population**) may be calculated from the expressions for their component parts. However, these formulas do not provide much insight into the quantitative relationship between the types of risk.
- Next, consider the numerical quantity **Familial Relative Risks (FRR)**, in Table 4.5 below. This table shows the results of numerical computation of the FRR under dominant and recessive models for various combinations of allele frequency and **Genetic Relative Risk (GRR)**.
- From, it is seen that **the FRRs may be quite modest even when the Generic relative risk is very large!** This phenomenon may be a reflection of the following situations:
The sibling's disease may not necessarily be hereditary.
Even if it were, the individual at risk may not have inherited the mutant allele.
Even if the case subject did inherit the mutant allele, the individual may nevertheless not succumb to the disease.
On the other hand, if the sibling is unaffected, he may still carry the susceptibility genotype, while the individual at risk might have inherited it, and he might get the disease!

Table 4.5 Sample calculation of Familial Relative Risks (FRR) for various genetic models

		Dominant		Recessive	
	q =	0.01	0.05	0.10	0.30
GRR	Pr (g) =	0.0199	0.0975	0.01	0.09
2		1.01	1.04	1.00	1.03
5		1.13	1.36	1.04	1.26
10		1.57	2.00	1.20	1.74
50		6.97	4.32	4.16	3.45
100		11.8	4.75	8.24	3.99
1000		23.3	5.48	25.2	4.61
∞		25.9	5.57	30.3	4.69

4.2.2.2 An Illustration of a Research Project in Genetic Epidemiology Research

This research project has been reported in: [BMC Med Inform Decis Mak. 2015; 15: 95](#). Published online 2015 Nov 18. doi: [10.1186/s12911-015-0211-1](https://doi.org/10.1186/s12911-015-0211-1) Copyright/License [Request permission to reuse](#)

Copyright © Schultz et al. 2015

Open Access

This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated (Fig. 4.8).

4.2.3 Heritability Analysis^[G]

The term “heritability”, as used in genetic epidemiology, has a special biostatistical meaning based upon an additive polygenic model, and on partitioning the variance of a disease or risk factor into genetic (G) and environmental (E) components, viz., the value of a phenotype P (either dichotomous or continuous) may be modeled as:

$$P = P(\mu + G + E) \quad (4.23)$$

(1) The Simple Linear Model of Heritability

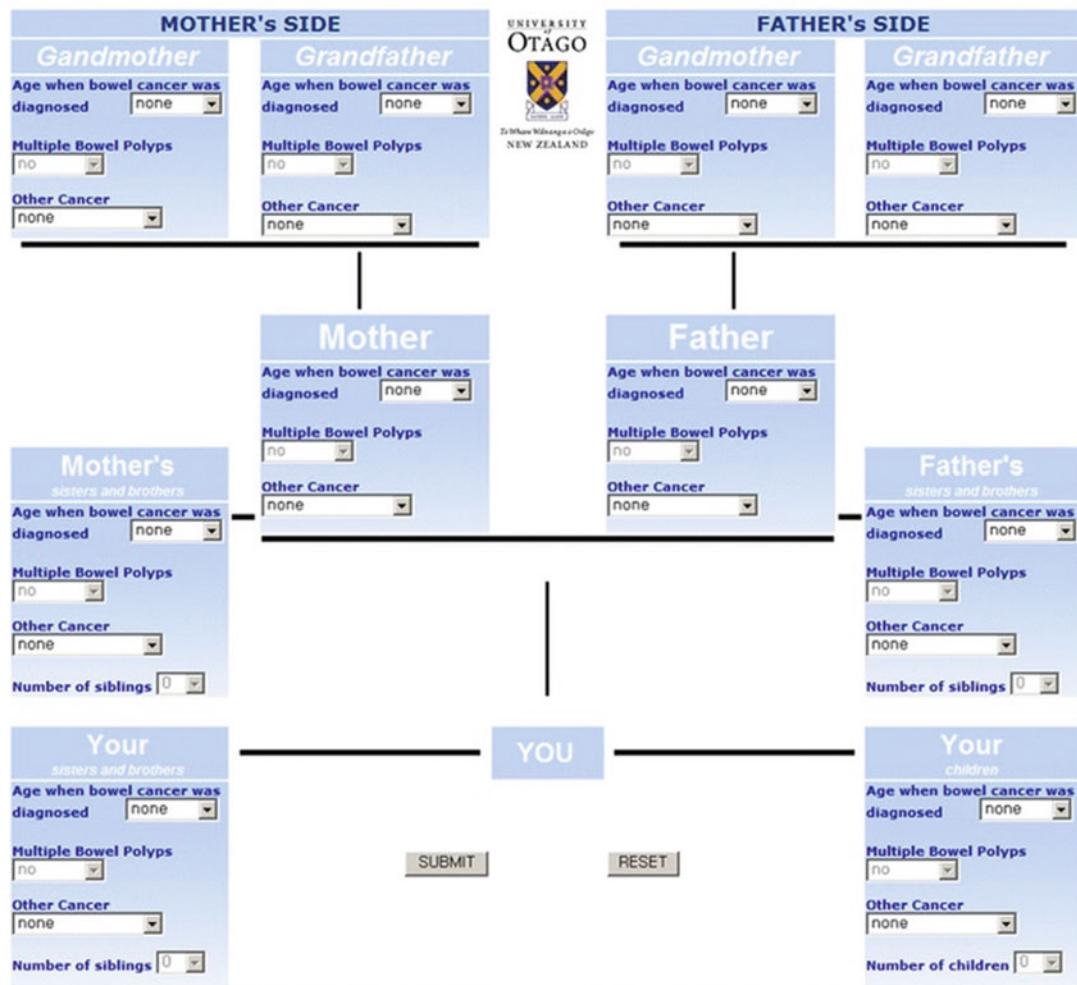
$$P = \mu + G + E \quad (4.24)$$

where μ is the mean of P , G are the genetic effects, and E are the environmental effects. According to Eq. 4.23, the variance of P , σ_P^2 , may be expressed as

$$\sigma_P^2 = \sigma_G^2 + \sigma_E^2 \quad (4.25)$$

Human Computer Interface Design Research Project

Diagram-Based assessment based on your family history



The diagram illustrates a Directed Acyclic Graph (DAG) for genetic model computations. It consists of several nodes arranged in a tree-like structure. At the top level, there are two sections: 'MOTHER's SIDE' and 'FATHER's SIDE'. Each section contains nodes for 'Grandmother' and 'Grandfather', each with fields for 'Age when bowel cancer was diagnosed' (dropdown menu with 'none' selected), 'Multiple Bowel Polyps' (dropdown menu with 'no' selected), and 'Other Cancer' (dropdown menu with 'none' selected). The 'MOTHER's SIDE' section also includes a logo for 'UNIVERSITY Otago' and text 'Te Ihu Wānanga Otago NEW ZEALAND'. Below these are 'Mother' and 'Father' nodes, each with similar fields. The 'Mother' node also includes a section for 'Mother's sisters and brothers' with fields for 'Age when bowel cancer was diagnosed' (dropdown menu with 'none' selected), 'Multiple Bowel Polyps' (dropdown menu with 'no' selected), 'Other Cancer' (dropdown menu with 'none' selected), and 'Number of siblings' (dropdown menu with '0' selected). The 'Father' node includes a section for 'Father's sisters and brothers' with similar fields. In the center, there is a large blue box labeled 'YOU' with a horizontal line extending to the left and right. Below 'YOU' are 'SUBMIT' and 'RESET' buttons. At the bottom, there are two more sections: 'Your sisters and brothers' and 'Your children', each with fields for 'Age when bowel cancer was diagnosed' (dropdown menu with 'none' selected), 'Multiple Bowel Polyps' (dropdown menu with 'no' selected), 'Other Cancer' (dropdown menu with 'none' selected), and 'Number of siblings' (dropdown menu with '0' selected) or 'Number of children' (dropdown menu with '0' selected).

Fig. 4.8 A DAG (Directed Acyclic Graph) for genetic model computations. (*BMC Med Inform Decision Maker.* 2015; 15: 95. Published online 2015 Nov 18. doi: 10.1186/s12911-015-0211-1. Copyright © Schultz et al. 2015)

from which one may define the **Heritability** of P , denoted as H^2_P , as

$$H^2_P = \sigma_G^2 / \sigma_P^2 \quad (4.26)$$

viz., the heritability H^2 is the proportion of the variance in P that is **attributed** to genetic effects. Indeed, H^2 varies from

0 (for small or no genetic influence)
to 1 (for strong genetic influences).

The genetic component of the variance may be further partitioned into additive effects, A, representing the effect of many genes, each with a small, additive effect, and dominance effect (D) between alleles at each locus, viz.,

$$\sigma_G^2 = \sigma_A^2 + \sigma_D^2 \quad (4.27)$$

with the “*narrow sense*” heritability is defined by

$$h^2 = \sigma_A^2 / \sigma_P^2 \quad (4.28)$$

and the “*broad sense*” heritability may be written as:

$$H^2 = (\sigma_A^2 + \sigma_D^2) / \sigma_P^2 \quad (4.29)$$

(2) The Polygenic Model of Heritability

Besides the simple linear model, another approach to model genetic heritability is based upon a polygenic model, using the liability scale which assumes that there are 2 unlinked genes contributing to the phenotype P, each with 2 alleles:

- Gene 1 with allele A and a, and
- Gene 2 with alleles B and b

In this model: each capital letter allele, A and B, increases an individual’s risk of disease by 1 unit on the liability scale. If the trait is dichotomous, then a threshold Z is assumed, such that case subjects with a liability score above this value succumb to the disease. Under this model, each possible genotype, the corresponding liability score, and the frequency of each genotype are shown in Table 4.6:

If one plots the frequency of genotypes by liability score, it is easy to see that the resulting distribution approximates a normal Gaussian distribution. And if one then assumes that there are 3 loci contributing to the heritability of the trait of interest, with the addition of allele C and c for the third gene, one may again tabulate all the possible genotypes and the corresponding liability scores, as shown in Table 4.7:

Again, if one plots the frequency of genotypes by liability score, it is easy to see, again, that the resulting distribution approximate a Gaussian distribution for the frequency of genotypes by liability score.

Table 4.6 Genotypes, liability scores, and genotype frequencies under a polygenic model with 2 loci

Genotypes	Liability score	Frequency
aabb	0	1
Aabb, aAbB, aaBb, aabB	+1	4
AAAb, AAbB, AsBb, aAbB, aAbB, aaBB	+2	6
AABb, AAbB, AaBB, aABB,	+3	4
AABB	+4	1

Table 4.7 Genotypes, liability scores, and genotype frequencies under a polygenic model with 3 loci

Genotypes	Liability score	Frequency
aabbcc	0	1
Aabbcc, aAbbcc, aaBbcc, aabBacc, aabbCc, aabbcC	+1	6
AAabbcc, AaBbcc, AabBccB,	+2	15
AABbcc, AAbBcc, AaBBcc,	+3	20
AABbcc, AABbCc, AAbbcC,	+4	15
AABBcc, AABbCc, AAbBCC,	+5	6
AABBCC	+6	1

If this exercise is continued with any number of genes contributing to the trait under the polygenic model, one would always end up with a similar distribution. Hence, with this model, one may estimate without further knowledge of the number of genes – as long as an additive model is used,

(3) Modeling Genetic Heritability Using Twins

To analyze and study genetic heritability of a biological disease or trait, a common approach is the use of twin studies. The approach is to compare mono-zygous (MZ, viz., identical) twins, who resulted from the division of a single zygote into two embryos and share 100% of their alleles, with di-zygous (DZ, fraternal) twins who result from the fertilization of two different ova, and on the average share 50% of their alleles, the proportion of phenotypic variance attributable to genetic influences can be estimated.

(All MZ co-twins are the same gender, whereas DZ co-twins can be of the same or different genders.)

Now assume only additive genetic effects (σ_A^2) and shared equal environments for MZ and DZ co-twins in a sample, then the covariances between the co-twins are:

$$\text{Cov}(MZ) = 1.0(\sigma_A^2) + 1.0(\sigma_B^2) \quad (4.30)$$

since MZ co-twins share **all** of their alleles, and

$$\text{Cov}(DZ) = 0.5(\sigma_A^2) + 1.0(\sigma_B^2) \quad (4.31)$$

since DZ co-twins share only **half** of their alleles on the average.

Now, σ_A^2 may then be estimated as

$$\begin{aligned} 2[\text{Cov}(MZ) - \text{Cov}(DZ)] &= 2[\{1.0(\sigma_A^2) + 1.0(\sigma_B^2)\} - \{0.5(\sigma_A^2) + 1.0(\sigma_B^2)\}] \\ &= 2[\sigma_A^2 + \sigma_B^2 - 0.5(\sigma_A^2) - \sigma_B^2] \\ &= \sigma_A^2 \end{aligned} \quad (4.32)$$

$$\text{and the heritability is } h^2 = \sigma_A^2 / \sigma_P^2 \quad (4.33)$$

Generally, heritability studies which are based on twins use the “ACE” model (which partitions the environmental variance based on twin studies) further **partitions** the environmental variance into shared and un-shared components, viz.,

$$\sigma_P^2 = \sigma_A^2 + \sigma_C^2 + \sigma_E^2 \quad (4.34)$$

Table 4.8 Proportions of Variance Explained by Genetic Factors, and Shared and Non-Shared Environment**

Cancer Site	Heritability	Shared Environment	Non-Shared Environment*
Stomach	0.28	0.10	0.62*
Colorectum	0.35*	0.05	0.60*
Pancreas	0.36	0.00	0.64*
Lung	0.26	0.12	0.62*
Breast	0.27*	0.06	0.67*
Prostate	0.42*	0.00	0.58*

* $p < 0.05$

where:

$$\sigma_A^2 = \text{variance owing to additive genes}$$

$$\sigma_C^2 = \text{variance owing to common or shared environmental effects genes and}$$

$$\sigma_E^2 = \text{variance owing to a unique or non-shared environment for co-twins}$$

$$\text{Then} \quad h^2(\text{ACE}) = 2(r_{\text{MZ}} - r_{\text{DZ}}) \quad (4.35)$$

where r is the intraclass correlation of P between co-twins.

Remarks on the Assumptions and Biases of Heritability Analysis

- Assumptions:** Clearly, many simplifying assumptions are used in undertaking heritability analysis. If these are not satisfied, biases will be created in the resulting estimates, resulting in under- or over-estimating the magnitude of genetic influences on a risk factor or a disease. Moreover, there are some misconceptions about heritability. When interpreting the results of heritability studies using twin, all these concerns should be recognized.
- Zygosity Determination:** A basic assumption is that zygosity can be accurately ascertained in a twin study. However, with new genomic technologies, zygosity determination is becoming more precise. Nevertheless, even these new techniques require the availability of DNA samples – which may not be feasible in very large-scale investigations. For example, in the following investigation based on the Scandinavian twin registrars described below in Table 4.8, all of the heritability estimates for cancer were based on questionnaire responses from co-twins:
- Generalizability:** Having obtained the heritability estimates, based on the foregoing twins study, an important issue would be whether these data are generalizable to other population groups and/or to other populations at large, especially to other ethnic populations or groups living in non-Nordic environments. Moreover, twin registries are usually based on recruitment of twins from specific populations. Generally, bias may be present owing to the magnitude of the heritability based upon the denominator, σ_P^2 , which is usually population specific. Hence, if σ_P^2 differs among group, then there may be decreases or increases in heritability estimates, even with the value of σ_C^2 unchanged.
- Same or Equally Shared Environments:** As described by Eqs. (4.30) through (4.35), another fundamental assumption in heritability analysis is the equally-shared environments for MZ and DZ co-twins. This may be written as:

$$\sigma_{E,\text{DZ}}^2 = \sigma_{E,\text{MZ}}^2 \quad (4.36)$$

If the assumption holds true, then the estimate of the additive variance, σ_A^2 , equals

the true value σ_A^2 :

$$s_A^2 = 2[(\sigma_A^2 + \sigma_{E,MZ}^2) - \{0.5(\sigma_A^2) + \sigma_{E,DZ}^2\}] \quad (4.37A)$$

$$= 2[\sigma_A^2 + \sigma_{E,MZ}^2 - 0.5(\sigma_A^2) - \sigma_{E,DZ}^2] \quad (4.37B)$$

$$= 2[0.5(\sigma_A^2)], \text{ using (4.36)} \quad (4.37C)$$

$$= \sigma_A^2 \quad (4.37D)$$

On the other hand, if the environment of DZ co-twins could be less similar than that for MZ, then

$$\sigma_{E,DZ}^2 > \sigma_{E,MZ}^2 \quad (4.38)$$

and

$$(\sigma_{E,DZ}^2 - \sigma_{E,MZ}^2) = W > 0 \quad (4.39)$$

Then

$$s_A^2 = \sigma_A^2 + W > \sigma_A^2$$

And heritability may be overestimated:

$$h^2 = s_A^2 / \sigma_P^2 \quad (4.40A)$$

$$= (\sigma_A^2 + W) / \sigma_P^2 \quad (4.40B)$$

$$> (\sigma_A^2 / \sigma_P^2) \quad (4.40C)$$

Remarks

1. It had been shown, in a study based on the Kaiser-Permanente Women Twins, that environmental and behavioral risk factors associated with coronary heart disease risk factors were more similar among MZ co-twins compared with DZ co-twins.
2. For example, smoking and alcohol consumption were more similar among MZ co-twins than DZ co-twins. But after adjusting for these factors, the heritability estimates for blood pressure and body mass were not significantly altered. Thus, at least in this study, this potential source of bias did not alter the outcome of these heritability estimates.
3. Environment-Gene Interactions

In heritability analysis, another common assumption is that there are no gene-environmental interactions involved. Recently, it was suggested that heritability may be significantly inflated by genetic interactions, whereby the concept of “phantom heritability” was introduced as follows:

Let $\pi_{\text{explained}}$ be the proportion of h^2 (narrow sense) explained by a set of known genetic variant, and write

$$\pi_{\text{explained}} = h^2_{\text{known}} / h^2_{\text{all}} \quad (4.41)$$

where

h^2_{known} = proportion of σ_P^2 explained by known additive variants, and

h^2_{all} = proportion of σ^2_p explained by *all* variants, whether known or not

It is generally assumed that

$$h^2_{\text{all}} = h^2_{\text{pop}} \quad (4.42)$$

where h^2_{pop} = “apparent” h^2 based upon phenotypic correlations in the population

$$= h^2_{\text{pop}}(\text{ACE}) \quad (4.43\text{A})$$

$$= 2(r_{\text{MZ}} - r_{\text{DC}}), \text{ from twin studies} \quad (4.43\text{B})$$

However, Eq. (4.42) holds only under a strictly additive model such as:

$$\sigma_p^2 = \sigma_A^2 + \sigma_C^2 + \sigma_E^2 \quad (4.34)$$

and if there are interactions, one needs to add terms to the model, such as:

$$\sigma_p^2 = \sigma_A^2 + \sigma_C^2 + \sigma_E^2 + \sigma_{G*G}^2 + \sigma_{G*E}^2 \quad (4.44)$$

where

σ_{G*G}^2 represents epistasis, and

σ_{G*E}^2 represents gene-environment interactions.

In this case:

$$h^2_{\text{pop}}(\text{ACE}) = h^2_{\text{all}} + W \quad (4.45)$$

where $W > 0$, and thus

$$h^2_{\text{all}} < h^2_{\text{pop}} \quad (4.46)$$

showing that the heritability based upon the ACE model will over-estimate the heritability attributable to all genetic variants in the presence of genetic interactions.

Austin, et al. (Morton 1955) called the term “*Phantom Heritability*” which was defined as:

$$\pi_{\text{phantom}} = 1 - (h^2_{\text{all}}/h^2_{\text{pop}}) > 0 \quad (4.47)$$

showing that when all genetic variants contributing to a phenotype are found, there will be a “gap” between the heritability based on these variants and the population-based (ACE) estimate of heritability.

4. Some Concluding Remarks:

An important question regarding a quantitative trait is whether the observed *variation* in the *character* is influenced by genes. This is not the same as asking whether genes play any role in the character’s development. Gene-mediated developmental processes lie at the base of every character, but **variations from individual to individual NOT necessarily the result of genetic variation**. Thus, the possibility of speaking any language at all depends critically on the structures of the

central nervous system as well as of the vocal cords, tongue, mouth, and ears, which depend in turn on the nature of the human genome. **There is no *environment* in which monkeys will speak!** Although the particular language that is spoken by humans varies from nation to nation, that variation is totally non-genetic.

The question of whether a trait is heritable is a question about the role that differences in genes play in the phenotypic differences between individuals or groups.

On Familiarity and Heritability

In principle, it is easy to determine whether any *genetic variation* influences the phenotypic variation among organisms for a particular trait. If genes are involved, then, on the average, biological relatives should resemble one another more than unrelated individuals do. This resemblance would be seen as a positive correlation between parents and offspring or between siblings (offspring of the same parents). Parents who are larger than the average would have offspring who are larger than the average; the more seeds that a plant produces, the more seeds that its siblings would produce. Such correlations between relatives, however, are evidence for genetic variation only if the relatives do *not* share common environments *more than nonrelatives do*. It is absolutely fundamental to distinguish heritability from *familiarity*.

Remarks

1. Traits are **familial** if and only if members of the same family share them, for whatever reason.
2. Traits are **heritable** if and only if the similarity arises from shared genotypes.

There are two general methods for establishing the heritability of a trait as distinct from its familial occurrence:

1. The first method depends on *phenotypic similarity* between relatives. For most of the history of genetics, this method has been the only one available; so nearly all the evidence about heritability for most traits in experimental organisms and in humans has been established by using this approach.
2. The second method, using *marker-gene segregation*, depends on showing that genotypes carrying different alleles of marker genes also differ in their average phenotype for the quantitative character. If the marker genes (which have nothing to do with the character under study) are seen to vary in relation to the character, presumably they are linked to genes that *do* influence the character and its variation. Thus, heritability is demonstrated even if the actual genes causing the variation are not known. This method requires that the genome of the organism being studied have large numbers of detectable genetically variable marker loci spread throughout the genome. Such marker loci can be observed from electrophoretic studies of protein variation or, in vertebrates, from immunological studies of blood group genes. (For example, within flocks, chickens of different blood groups show some significant difference in the weights of their eggs!)

4.2.4 Molecular Variation Study Methods

Since the introduction of molecular methods for the study of DNA sequence variation, a large numbers of variable nucleotide positions have been discovered in a great variety of organisms. This molecular

variation includes both single nucleotide replacements and insertions as well as deletions of longer nucleotide sequences. These variations may be detected by the gain or loss of sites of cleavage of restriction enzymes or by length variation of DNA sequences between two fixed restriction sites, both of which are a form of Restriction Fragment Length Polymorphisms (RFLPs). For example, different tomatoe strains carrying different RFLP variants differ in fruit characteristics.

Since so much of what is known or claimed about heritability still depends on phenotypic similarity between relatives, especially in human genetics, one may begin the examination of the problem of heritability by analyzing phenotypic similarity.

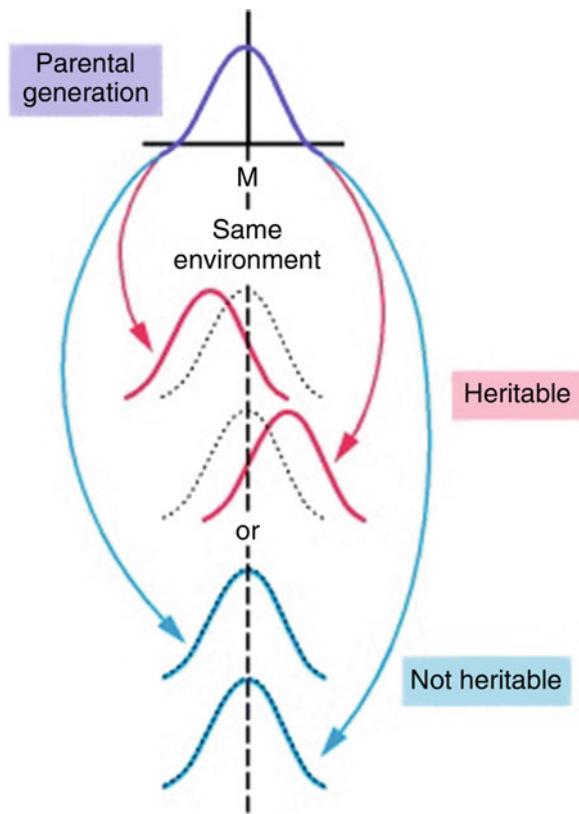
Phenotypic Similarity Among Relatives

In experimental organisms, one may separate environmental from genetic similarities. For example, the offspring of a cow producing milk at a high rate and the offspring of a cow producing milk at a low rate may be raised together in the same environment to see whether, despite the environmental similarity, each resembles its own parent. However, in humans, this is difficult to do. Owing to the nature of human societies, members of the same family not only share genes, but also have similar environments. Thus, the observation of simple familiarity of a trait is genetically uninterpretable. For example, people who speak Chinese have Chinese-speaking parents and people who speak English have English-speaking parents. Yet the massive experience of immigration to North America has demonstrated that these linguistic differences, although familial, are non-genetic. The highest correlations between parents and offspring for any social traits in the United States are those for political party and religious sect, but they are not heritable ! The distinction between familiarity and heredity is not so obvious. The Public Health Commission, which originally studied the vitamin-deficiency disease *pellegra* in the southern United States in 1910, came to the conclusion that it was genetic because it ran in families!

To determine whether a trait is *heritable* in human populations, one must use *Adoption Studies* to avoid the usual environmental similarity between biological relatives. *The ideal experimental subjects are identical twins reared apart, because they are genetically identical but environmentally different.* Such adoption studies must be so organized that there is *no correlation between the social environment of the adopting family and that of the biological family*. These requirements are difficult to meet. Thus, in practice, one knows little about whether human quantitative traits that are familial are also heritable! Skin color is heritable, as is adult height, but even for these traits one must be very careful. One knows that skin color is affected by genes from studies of cross-racial adoptions and observations that the offspring of black African slaves were black even when they were born and reared in Canada. But are the differences in height between Japanese and Europeans affected by genes? The children of Japanese immigrants who are born and reared in North America are taller than their parents but shorter than the North American average, so we might conclude that there is some influence of genetic difference. However, most second-generation Japanese-Americans are even taller than their first-generation American-born parents. It appears that some environmental-cultural influence or perhaps maternal effect is still felt in the first generation of births in North America. One cannot yet say whether genetic differences in height distinguish North Americans of, say, Japanese and Swedish ancestry.

Personality traits, temperament, and cognitive performance (such as IQ Scores, etc.), as well as a whole variety of behaviors such as alcoholism and of mental disorders such as schizophrenia, have been the subject of heritability studies in human populations. Many show familiarity. There is a positive correlation between the IQ scores of parents and the scores of their children (the correlation is about 0.5 in white American families), but the correlation does not distinguish familiarity from heritability. To make that distinction requires that the environmental correlation between parents and

Fig. 4.9 Standard method for testing heritability in experiment organisms



children be broken, so **adoption studies** are common. It is difficult to randomize the environments, even in cases of adoption, evidence of heritability for human personality and behavior traits remains equivocal despite the very large number of studies that exist. Prejudices about the causes of human differences are widespread and deep, and, as a result, the standard of evidence adhered to in studies of human heritability of IQ, for example, have been much more lax than in studies of milk yield in cows.

Figure 4.9 summarizes the usual method for testing heritability in experimental organisms. Individuals from both extremes of the distribution are mated with their own kind, and the offspring are raised in a common controlled environment. If there is an average difference between the two offspring groups, the trait is heritable.

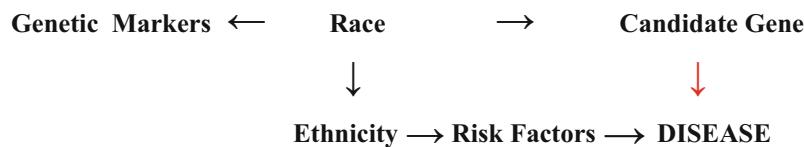
In experimental organisms, environmental similarity can often be readily distinguished from genetic heritability.

4.3 Genomics for Human Genetic Epidemiology

Currently, the available genomic technology applicable to genetic epidemiology studies include the following:

- Genome-wide SNP (*Single Nucleotide Polymorphisms*, pronounced “snips”) genotyping platforms
- Sanger and next generation (**NextGen**, massively parallel) DNA sequencing Methods
- Exome sequencing.

Fig. 4.10 A conceptual model for population stratification or confounding by ethnicity



Remarks

- (1) A number of workers in the field of genetic epidemiology (Thomas 259ff []) have indicated some concern that case-control or cohort designs using unrelated case-subjects may be susceptible to a form of confounding by ethnicity known (in the genetic epidemiologic literature) as *population stratification*, because there are often considerable gradients in gene frequency within the broad racial groups that are generally used for stratification. Such variation could be correlated with other unknown risk factors, as illustrated in Fig. 4.10:
- (2) In Fig. 4.10, one would like to test the association between “Candidate Gene” and “Disease”. The allele frequency for the candidate gene is determined by “Race”, which is also a determinant of other unmeasured “Risk Factors” and hence a confounder!
- (3) Here, “Genetic Marker” may be used to infer *race*, but self-reported “Ethnicity” may also be used for the purpose, and it can also be a surrogate for the other risk factors.

4.3.1 Complex Traits and Mendelian Inheritance

From their parents, humans (being diploids) normally inherit one of each of 22 autosomal chromosomes and a gender chromosome (X or Y). This means that biostatistical analysis for genetic epidemiology is basically different from that of traditional epidemiology! Thus, for genomic studies in human epidemiology, each case subject carries 2 alleles at any autosomal location in his/her genome. In this way, rather than the unit of analysis being an individual:

- (1) each of the two alleles must be considered, and
- (2) chromosomes, rather than individuals, are often the unit of analysis.

Individuals may be:

- (A) Homozygous, carrying 2 identical copies of a possible disease susceptibility allele of interest;
- (B) heterozygous, carrying only one of the alleles of interest; or
- (C) homozygous with 2 identical alleles that are not thought to be associated with any disease.

Hence, sample size for a study may be the number of chromosomes carrying the disease susceptibility allele. Here, one may use the ABO blood groups to demonstrate this approach as well as other fundamental concepts of genetics

Remarks on the ABO Blood Group

ABO phenotypes or serotypes consist of blood groups A, B and O. These genetically determined blood groups were discovered in 1900, and the molecular basis of the ABO system gene on Chromosome 9q34 was also found. The ABO locus encodes a glycosyl-transferase with different substrate

specificities, and has 3 main variants: the A and B variant code for A and B antigens, respectively, on the surface of red blood cells, whereas O represents the lack of an antigen owing to a single DNA base deletion.

Thus, a case subject may have 1 of 6 possible genotypes at this locus:

- (i) 3 homozygous: AA, BB, and OO, or
- (ii) 3 heterozygous AB, AO, and BO.

4.3.1.1 Mendel's Law

Described by Mendel in 1865, his Laws of Inheritance remain fundamental to the relevant Understanding of genetic susceptibility to diseases. The following 2 laws are most relevant to human genetic epidemiology:

1. ***Law of Equal Segregation:*** *The 2 alleles at a genetic locus segregate from each other and are passed on to an offspring with equal probability.*

Thus, for the ABO blood group, AB genotype has an equal probability of passing on the A allele and the B allele to his/her offspring.

2. ***Law of Independent Assortment:*** *Genes on different chromosomes are inherited independently.*

Thus, for the same chromosome in the foregoing example, genes located on any chromosome other than the location of the ABO locus would be inherited independently of the ABO alleles. Such genes on different chromosomes are not genetically “linked” to the ABO locus.

Thus, in genetic epidemiology, the term linkage means that 2 genes are close enough together on the same chromosome that they do not segregate independently.

4.3.1.2 Modes of Inheritance

The observed genotypes that may be described physiologically, biochemically, and physically, are inherited owing to a single underlying gene, depend on the mode of inheritance of that particular gene. These are Mendelian traits which may have the following modes of inheritance:

- (1) **Autosomal dominant:** viz., a gene that is expressed phenotypically in both Heterozygotes and homozygotes. Only one copy of a mutant allele is needed for expression of that trait.
- (2) **Recessive:** Both alleles are expressed phenotypically only in homozygotes; two copies of the mutant allele are needed for expression of the trait.
- (3) **Co-dominant or Additive Inheritance:** Both alleles are expressed in heterozygotes.
- (4) **X-linked:** Genes located on the X chromosome.
- (5) **Y-linked:** Genes located on the Y chromosome

These several modes of inheritance form the basis of segregation analysis: a biostatistical methodology that may be used to determine whether a specific mode identifying any major effects on any disease.

4.3.1.3 OMIM (Online Mendelian Inheritance in Man)

The OMIM is a useful online resource database for genetic epidemiology, being a comprehensive collection of human genes and genetic phenotypes: <https://www.ncbi.nlm.nih.gov/omim>

It is a comprehensive, authoritative compendium of human genes and genetic phenotypes which is available freely, and daily updated. Authored and edited at the McKusick-Nathans Institute of Genetic Medicine at Johns Hopkins University School of Medicine, directed by Dr. Ada Hamosh. Its official home is: omim.org

This resource forms a concise reviews of all known Mendelian diseases “with a particular focus on molecular relationships between genetic variation and phenotypic expressions. It is well referenced, containing links to other genetic resources. Historically, it is the continuation of the catalog of Mendelian traits and disorders of the work of Dr. Victor McKusick, whose catalog of Mendelian traits and disorders that were published in the 12-volume “Mendelian Inheritance in Man” (1966–8). As of this writing, some OMIM statistics are as follows:

OMIM Gene Map Statistics

OMIM Morbid Map Scorecard (Updated February 15th, 2018) :

Total number of phenotypes* for which the molecular basis is known	6153
Total number of genes with phenotype-causing mutation	3874

Phenotypes include:

- (1) single-gene Mendelian disorders and traits;
- (2) susceptibilities to cancer and complex disease (e.g., BRCA1 and familial breast-ovarian cancer susceptibility, and CFH and macular degeneration,
- (3) variations that lead to abnormal but benign laboratory test values (“nondiseases”) and blood groups (e.g., lactate dehydrogenase B deficiency, and ABO blood group system, ; and
- (4) select somatic cell genetic disease (e.g., GNAS and McCune-Albright syndrome, and IDH1 and glioblastoma multiforme)

Distribution of Phenotypes across Genes (Updated February 15th, 2018) :

Number of genes with 1 phenotype	2652
Number of genes with 2 phenotypes	729
Number of genes with 3 phenotypes	259
Number of genes with 4+ phenotypes	234

Dissected OMIM Morbid Map Scorecard (Updated February 15th, 2018):

Class of Phenotype	Phenotype	Gene *
Single gene disorders and traits	5114	3495
Susceptibility to complex disease or infection	696	498
“Nondiseases”	145	115
Somatic cell genetic disease	213	121

*Some genes may be counted more than once because mutations in a gene may cause more than one phenotype and the phenotypes may be of different classes (e.g., activating somatic BRAF mutation underlying cancer, and germline BRAF mutation in Noonan syndrome.

4.3.1.4 Complex Traits in Humans^[W]

Most human genetic traits may be classified as either monogenic or complex. Monogenic traits are strongly influenced by variation within a single gene and are recognized by their classic patterns of inheritance within families. While monogenic traits formed the basis for “classic” genetics, it has become clear that conditions whose inheritance strictly conforms to Mendelian principles are relatively rare.

Generally, complex traits are believed to result from variation within multiple genes and their interaction with behavioral and environmental factors. Complex traits do not follow readily predictable patterns of inheritance. This distinction between monogenic and complex traits, while useful, may be

overly simplistic. Traits that appear to be monogenic can be influenced by variation in multiple genes (“modifier genes”). Complex traits may be predominantly influenced by variation in a single gene.

A complex trait has been defined as:

- (1) a measured phenotype, such as disease status or a quantitative character, which is influenced by many environmental and generic factor, and potentially by interaction between them, or
- (2) any phenotype that does not exhibit classic Mendelian recessive or dominant inheritance attributable to a single gene when the single correspondence between genotype and phenotype breaks down.

Moreover, complex traits are also characterized by:

- (a) incomplete penetrance and phenocopies,
- (b) genetic heterogeneity
- (c) polygenic inheritance, and
- (d) gene-environment interactions and epistasis.

4.3.1.5 Hardy-Weinberg Equilibrium (HWE) Principle

In genetic epidemiology, an early key concept of population genetics is the Hardy-Weinberg Principle, independently described by Geoffrey. H. Hardy (an Oxford/Cambridge University mathematician and William Weinberg (a German geneticist).

On January 13, 1908, Wilhelm Weinberg read to an evening meeting in Stuttgart, Germany, a paper in which he “derived the general equilibrium principle for a single locus with two alleles”. Mendel (1866) had already initiated population genetics by considering the consequences of continued self-starting with the cross $Aa \times Aa$, obtaining $1 AA : 2 Aa : 1 aa$ in the first generation and $(2^n - 1)AA : 2Aa : (2^n - 1)aa$ in the n th generation, assuming for simplicity that each plant produced four seeds. With A dominant to a , this gives phenotypic proportions $2^n + 1$ “A”: $2^n - 1$ “a” as noted by Weinberg [although Mendel’s n th generation was his $(n - 1)$ th]. He did not explicitly refer to Mendel, but he was surely familiar with Mendel’s paper. He went on, “This situation appears much different when Mendelian inheritance is viewed under the influence of panmixia” and, starting with arbitrary proportions m and n (not the same n as before; $m + n = 1$) of each of the two homozygotes AA and aa , he obtained “by application of the symbolism of the binomial theorem” the daughter generation $m^2 AA + 2mnAa + n^2 aa$. Another generation of random mating led by direct calculation to the same proportions among the offspring and “We thus obtain the same distribution of pure types and hybrids for each generation under “panmixia”. Weinberg then used his result to work out the numbers of the two phenotypes to be expected among the relatives of an individual of known phenotype. Thus, he has established the “Hardy-Weinberg law” in the most obvious and direct manner.

Meanwhile in England, between the rediscovery of Mendel’s paper in 1900 and the publication by the mathematician Professor G. H. Hardy of the same result as Weinberg’s in July 1908 in the American weekly *Science* (HARDY 1908), one witnessed the story of how Britain’s foremost mathematician became involved in a simple problem of Mendelian genetics has been told many times.

The Hardy-Weinberg Equilibrium (HWE) Principle explains:

- (a) why a dominant trait does NOT replace a recessive trait in the population,
- (b) why, in tic populations, allele frequencies do not change from one generation to the next, and
- (c) for genetic epidemiology, the relationship between allele frequencies on chromosomes and genotypes in individuals are well-defined

To further describe the HWE, assume that one has a single genetic locus with allele X and x , for which the frequency of allele X in the population is

$$f(X) = p, \quad 0 < p < 1 \quad (4.48)$$

and

$$f(x) = q \quad (4.49)$$

As X and x are the only possible alleles, by definition

$$p + q = 1 \quad (4.50)$$

Since the alleles are on chromosomes, the denominators for these frequencies are the number of chromosomes carrying that allele, not the number of individuals. If it is assumed that in the population under investigation there is:

- (a) no genetic mutation,
- (b) random mating,
- (c) a large population size without random drift,
- (d) negligible migration out of and into the population, and
- (e) no mutual selection operating on the locus,

then it may be shown that, in one generation, the genotype frequency is given by:

$$f(XX) = ep^2 \quad (4.51)$$

$$f(Xx) = 2pq \quad (4.52)$$

$$f(xx) = q^2 \quad (4.53)$$

Also, over time, these frequencies will remain the *same*, in equilibrium, from one generation to the next so long as the foregoing assumptions hold. Moreover, the presence of the HWE Principle is equivalent to the alleles at that locus being independent from one another. It permits the computation of the genotype frequencies from known allele frequencies, as shown in the following worked example:

An Example of the Testing of the Hardy-Weinberg Equilibrium (HWE) Principle

This example illustrates the testing of whether a genetic code is in HWE. Consider the human chemokine receptor gene CCR5 for which it had been shown that homozygosity for a 32 bp deletion in this gene, $CCR5\Delta$, confers resistance to human immunodeficiency virus infection. In a research study of some 338 case-subjects, the observed genotypes are shown in the following table: Table 4.9

Of the total of 338 case-subjects in this study, 265 were homozygous for the CCR5 wild-type genotypes, 66 were heterozygous, and 7 were homozygous for the deletion.

To ascertain whether this locus is in HWE, one should first estimate the allele frequencies from the observed genotype counts:

$$\begin{aligned} p &= f(CCR5) \\ &= [2(CCR5/CCR5) + (CCR5/CCR5\Delta)]/2N \\ &= [2(265) + 66]/2(338) \\ &= [530 + 66]/676 \\ &= 596/676 \\ &= 0.881, 656, 8 \approx 0.882 \end{aligned} \quad (4.54)$$

Table 4.9 Testing of the HWE principle of gene alleles in 338 case-subjects

Genotype	Observed number of case-subjects	Expected frequency under HWE principle	Expected number of case-subjects under HWE
CCR5/CCR5	265	0.778	263.0
CCR5/CCR5 Δ	66	0.208	70.3
CCR5 Δ /CCR5 Δ	7	0.014	4.7
Total (N):	338		

$$\begin{aligned}
 q &= f(CCR5\Delta) \\
 &= [2(CCR5/CCR5\Delta) + (CCR5/CCR5\Delta)]/2N \\
 &= [2(7) + 66]/2(338) \\
 &= [14 + 66]/676 \\
 &= 80/676 \\
 &= 0.118,343,2 \approx 0.118
 \end{aligned} \tag{4.55}$$

and so

$$\begin{aligned}
 p + q &= 0.881,656,8 + 0.118,343,2 \\
 &= 1.000,000,0
 \end{aligned}$$

viz.,

$$p + q = 1 \tag{4.56}$$

thus, satisfying the HWE Principle!

One may then move on to determine the expected genotype frequencies and the number of case-subjects under HWE, viz., the *third column* in Table 4.9:

$$\begin{aligned}
 f(CCR5/CCR5) &= p^2 \\
 &= (0.881,656,8)^2 7 \\
 &= 0.777,318,7 \approx 0.777
 \end{aligned} \tag{4.57}$$

$$\begin{aligned}
 f(CCR5/CCR5\Delta) &= 2pq \\
 &= 2(0.881,656,8)(0.118,343,2) \\
 &= 0.208,676,2 \approx 0.209
 \end{aligned} \tag{4.58}$$

$$\begin{aligned}
 f(CCR5\Delta/CCR5\Delta) &= q^2 \\
 &= (0.118,343,2)^2 \\
 &= 0.014,005,1 \approx 0.014
 \end{aligned} \tag{4.59}$$

To obtain the expected number of case-subjects with each genotype under the HWE Principle shown in the 4th column of Table 4.9, simply multiply the calculated from

$$\begin{aligned}
 0.777,318,7 \times 338 &= 262.7337206 \approx 262.7 \\
 0.208,676,2 \times 338 &= 70.5325556 \approx 70.5 \\
 0.014,005,1 \times 338 &= 4.7337238 \approx 4.7
 \end{aligned}$$

Using a χ^2 - test to ascertain if the observed and expected numbers of case-subjects are statistically significantly different:

$$\chi^2 - (1 \text{ df}) = \Sigma \left[(\text{Observed} - \text{Expected})^2 / \text{Expected} \right] = 1.42. \text{ and } p - \text{value} = 0.25$$

Since the *p*-value of 0.25 is **not** significant, it may be concluded that the observed distribution of genotypes in the sample is consistent with the HWE Principle.

Deviations from the Hardy-Weinberg Equilibrium (HWE) Principle

It is possible for genetic loci to deviate from HWE in a study. Examples of such instants include:

- Migration: movement of case-subjects among sub-populations;
- Random genetic drift: chance fluctuations in allele frequencies in small populations as a result of random sampling among gametes;
- Natural selections: inherited differences in the ability to survive and reproduce; over time, superior survival and reproductive genotypes increase in the population;
- Non-random mating in the population, either assortative mating or inbreeding;
- Mutation: change in genetic material, viz., variations in the DNA sequence.

4.3.1.6 Gene Structure and Genetic Code

By the processes of translation and transcription, DNA sequences of the 4 bases (A, Adenine; G, Guanine; C, Cytosine; and T, Thymine) form the code for amino acids based upon 3-base triplets or “codons” that comprise the human genetic code.

Now, there are 64 such possible triplets ($4^3 = 64$), but only 20 amino acids. Thus, several different codons can specify the same amino acid. For example, arginine and leucine are each specified by 6 different Codons, whereas methionine and tryptophan are specified by a single codon. Also, 3 codons are “stop”, or nonsense, codons which terminate translation of mRNA into amino acids. These phenomena, called “degeneracy” has important implications for human genetic epidemiology because it implies that ***there can be point mutations in the DNA sequence that change a codon, but do not change the protein encoded by a potential disease susceptibility gene!*** Thus, for example, if a genetic in a candidate gene is found to be associated with a disease, but that variant does NOT change the protein, it can be uncertainty in determining whether that variant is causally related to disease susceptibility!

4.3.1.7 Genetic Linkage and Disequilibrium

Since the beginning of the study of human genetic epidemiology, family studies have been a fundamental research approach for mapping disease susceptibility genes.

In the human genetic epidemiology context, the term “linkage” has biologic and specific connotations. During meiosis and the formation of gametes, alleles on the same chromosome are inherited together except when recombination (crossing over) occurs during the first division of meiosis. Initially, homologous paternal and maternal chromosomes line up at the centromere during Prophase I of meiosis. Then recombination, or crossing over, takes place when sister chromatids on the paternal and maternal chromosomes line up at the centromere during Prophase I of meiosis. Recombination, or crossing over, takes place when sister chromatids on the paternal and maternal chromosomes exchange genetic material. This is a normal process which breaks up associations between alleles along the same chromosome during cell division.

On the same chromosome, the closer together 2 genes are the less likely recombination between them becomes. On the other hand, genes on different chromosomes always segregate independently,

according to Mendel's Law of Independent Assortment, and thus are not linked genetically. This biological process is the basis of LOD score linkage analysis, as described in Sect. 4.4.

4.3.1.8 DNA Sequencing

DNA sequencing, including exome sequencing, in contradistinction to genome-side SNP chips that are designed primarily to detect common variation, can detect rare genetic variants. This capability is important for complex diseases, and Mendelian disease may be because functionally deleterious variants are expected to be present in most genes, although they may be rare. This is a very active area of research.

Three Generations of DNA Sequencing

1. “First Generation” – Sanger Sequencing Method

This is the automated Sanger sequencing (1977), designed to read the sequence of base pairs in a region of DNA, and used to obtain the first entire sequence of a human genome for the Human Genome Project in 2001. Even with the rapidly evolving platforms for DNA sequencing, Sanger sequencing remains the “gold standard” for clinical sequencing, such as in cytogenetics. It is still the most accurate method for sequencing, having a well-defined chemistry with an average error rate of <1%. This process of DNA sequencing has 4 steps:

- (1) Generation of a nested set of DNA fragments, using chemical or enzymatic methods.
- (2) Separation of the fragments by electrophoresis.
- (3) Detection of bases using labeling chemistries using 4-color fluorescent dyes.
- (4) Analysis or base calling.

On the other hand, this approach requires a relatively large amount of DNA, and the throughput rate is relatively low compared with other available methods.

2. The NextGen (“Next Generation”) or “Massively Parallel” DNA Sequencing Method

This approach permits hundreds of millions of reads at a much lower cost. It is capable of sequencing huge numbers of different DNA sequences in a single reaction, hence the term “parallel, using mostly dye-labeled modified nucleotides. NextGen sequencing has short read lengths (30 – 400 bp) and lower sequencing Quality. This are the “Second-Generation” Methods – this method uses amplification-based approaches, viz., following sample collection, the steps used involve experiments for either human whole-genome re-sequencing or exome sequencing

3. The “Third Generation” of DNA Sequencing Technologies – “Single Molecule Systems”:

Unlike the “Second Generation” approaches, these **Third-Generation** approaches do not require amplification steps, thus avoiding potential errors, and may be used to sequence very small quantities of DNA, and will generally provide longer read lengths. These approaches, including exome sequencing, that can target specific areas of special interest in the genome, are the preferred techniques for genetic epidemiology research and study.

4.3.1.9 Study Designs for Genetic Variants

In selecting the study design to be used for a genetic epidemiology study, the frequency of potential disease-causing genetic variants is an important element to be considered. Depending upon the frequency of the less common allele at a locus, viz., “the minor allele frequency”, many different analysis approaches are applicable. These approaches are summarized in Table 4.10:

Table 4.10 Analysis approaches for genetic epidemiology study design

Minor allele frequency	Class of variant	Analysis approaches
5–50%	<i>Very Common</i>	Genome-wide Association Studies (GWAS)
1–5%	<i>Less Common</i>	Association analysis using variants from 1000 Genomes
Less than 1%	<i>Rare</i>	Exome sequencing, sequencing affected relatives in families, or extreme trait sequencing
Present only in affected probands and relatives	<i>Private</i>	Co-segregation (linkage) in affected families

Study Designs for Rare Genetic Variants

For identifying rare genetic variants in disease susceptibility, 2 distinct study designs available are:

(1) Sequencing relatives in families with multiple affected members

This approach involves sequencing the distantly related and affected family members – since the more distantly related these families are, the fewer genetic variants they will share. By assuming that the disease of interest is caused by the same variant in both relatives, they should share that causal variant. For example, first cousins share approximately 1/8 of their genomes **IBD** (Identity By Descent). However, when using this strategy, distant relatives will still share too much of their genomes to allow simple identification of the causal variants. Hence, one must screen or filter out the associated variants, based on function, allele frequency, and the type of gene affected.

(2) Extreme trait sequencing

The second strategy, well-suited for quantitative traits, is to sequence a small number of carefully chosen study participants from one or both ends of the phenotype distribution. The frequency of alleles contributing to disease susceptibility is expected to be enriched in these extremes of the distribution, and so may be identified based upon a relatively modest number of cases and controls.

The success potential for this strategy was demonstrated in a study of **HDL** (High-Density Lipoprotein) cholesterol. In this study, 3 candidate genes were sequenced in only 38 case-subjects taken from the upper and lower 5% of the HDL cholesterol distribution. The results showed that rare alleles had significant effects on low HDL cholesterol, a major factor for coronary heart disease. In one of the first findings from the NHLBI ESP project, extreme phenotypes were used to ascertain that variants in the DCTN4 gene, which encodes a dynactin protein, is associated with chronic *Pseudomonas aeruginosa* infection among people with cystic fibrosis. Such infections are related to reduced lung function and shorter mean survival! In this study, 43 case-subjects with an early age of onset for the infection were compared with 48 of the oldest individuals who did not have a chronic infection, and the presence of one or more missense mutations in the DCTN4 gene was associated with early age of onset of the infection, with Hazard Ratio = 1.9, and P = 0.004. It appeared that the success of this exome sequencing study of a complex trait with a small sample size is attributable to the well-matched extremes, the large effective size of the gene, and the relatively high combined MAF sequencing of rare coding variation of 2440 individuals from the ESP project demonstrate that large sample size may be needed for sufficient power in association studies of rare variants with complex diseases.

4.3.1.10 Genetic Linkage and Linkage Disequilibrium

(A) Genetic Linkage and Recombination

Genetic linkage analysis is the determination of the approximate chromosomal location of a gene by searching for evidence of congregation with other genes whose locations are known, viz., the Marker Genes. There are tendencies for co-segregation – in which 2 or more genes tend to be inherited together, and hence for individuals with similar phenotypes to share alleles at the marker locus.

(B) Recombination and Map Functions

Meiosis, a kind of cell division, results in reduction of chromosome number from diploid to haploid. Meiosis occurs only in germ cells and produces gametes, sperm, and ova that are genetically different from the parent cells. Like mitosis, meiosis is preceded by a round of DNA synthesis, but in meiosis this is followed by two rounds of cell division, called Meiosis I and II. Meiosis I differs from mitosis in that it is not the sister chromatids that separate but rather the maternal and paternal homologs. Maternal and paternal homologs, each consisting of 2 sister homologs, pair up, forming a *bivalent*, a structure consisting of 4 chromosomes of the same type. When the nucleus divides, both chromatids of the maternal chromosome go one way, while the 2 chromatids of the paired paternal chromosome go the other way. Mendel's Law of Independent Assortment states that segregation of paternal and maternal homologs to the 2 daughter cells is independent from chromosome to chromosome. Thus, there are at least 2^{23} (approximately 8.4 million) possible types of sperm or ova that can be produced by one person.

Through the 2 processes of *meiosis*, genes are transmitted from parents to offsprings:

Meiosis I results in 2 daughter cells, each 23 duplicated chromosomes:

- (i) in males, these cells are the same size, and both progress to the second round of cell division.
- (ii) in females, uneven division of cytoplasm during meiosis I results in one large cell and a small polar body that does not further divide.

Meiosis II is identical to cell division in mitosis (except for uneven division of the cytoplasm in females). The 2 chromatids of each chromosome separate, and cell division results in 2 haploid cells.

Meiosis can take a long time: in women, Meiosis I begins during embryogenesis and is completed many years later at the time of ovulation. Each of the 22 autosomes in a gamete would be an exact copy of one of the two parental homologs. A process called *meiotic recombination* mixes the genetic material of the homologs during meiosis, so that each chromosome present in the gamete has contributions from both parents.

(C) Linkage Disequilibrium (LD)

The concept of LD, developed by population geneticists in the 1960s, is a fundamental concept in genetic epidemiology. LD may be defined as the non-random association or correlation between alleles at 2 loci. It is equivalent to alleles occurring on the same chromosome more or less often than expected from their individual allele frequencies.

Referring, again, to **Fig. 4.1 (DNA Sequence Variation in the Human Genome – 1)**:

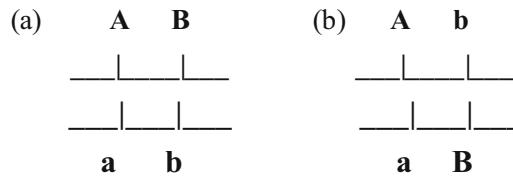


Fig. 4.11 Possible Haplotypes of Linked Biallelic Genes A and B

- (a) The first 6 nucleotides in this figure are in strong LD with each other, viz., among the pink chromosomes, every time G occurs in Column 1:

- a C occurs in Column 2, and
 - a T occurs in Column 3, etc.
- (b) Among the yellow chromosomes, every time A occurs in column 1:
- a C occurs in Column 2, and
 - a T occurs in Column 3, etc.

but different nucleotides are found in Columns 5 and 6.

- (c) These patterns of association, or haplotypes, are common in the human genome.

The International HapMap project was established to create a genome-wide database of these patterns.

The strong correlation among the first six nucleotides on the plate (Fig. 4.1) are indicated by red boxes below these columns, as are the blue and purple haplotypes on the right side of the figure. On either side of the recombination hotspot, there is little or no correlation, indicated by the white boxes.

(D) LD Calculations

To show LD calculations, consider linked genes A and B with 2 alleles each: Fig. 4.11:

In this example, the genotyped individual is a double heterozygote with genotypes Aa and Bb at these 2 loci, respectively. If the capital letter alleles are on the same chromosome: Fig. 4.11a, then the haplotypes are A-B and a-b. Taken together, these 2 haplotypes are referred to as a **diplotype**.

Note that it is also possible that the capital letter alleles are on different chromosomes: Fig. 4.11b. In such case, the haplotypes are then A-b and a-B.

To determine if LD is present, one compares the frequency of haplotype A-B, viz., $f(A-B)$, with the allele frequencies of A and B, viz., $f(A)$ and $f(B)$, in the population under discussion:

If

$$f(A-B) = f(A) \times f(B) \quad (4.60)$$

then A and B are considered to be in equilibrium. This is equivalent to alleles A and B being independent of each other.

However, if

$$f(A-B) \neq f(A) \times f(B) \quad (4.61)$$

then A and B are **not** independent, and disequilibrium is present.

Moreover, if

$$f(A - B) > f(A) \times f(B) \quad (4.62)$$

then alleles A and B are correlated, and are in LD with each other.

To calculate the magnitude of LD, one should first determine the pairwise disequilibrium coefficient:

$$D = f(A - B) - f(A) \times f(B) \quad (4.63)$$

However, this may not be a satisfactory approach because of the possibility of the values of D may be dependent on the allele frequencies and D can be positive or negative. To cater for such problems, one should next determine the maximum possible value of D, given the allele frequencies, that can be expresses as:

$$D_{\max} = \text{Minimum}[f(A)xf(b), f(a) \times f(B)] \quad (4.64)$$

One then standardize D using D_{\max} , and take the absolute value:

$$D' = |D/D_{\max}| \quad (4.65)$$

This estimate of LD may be shown to be useful for detecting recombination between 2 genetic loci. However, for determining the correlation, another measure of LD, r^2 , may be used:

$$r^2 = D^2 / [f(A) \times f(a) \times f(B) \times f(b)] \quad (4.66)$$

Both D' and r^2 range from 0 (indicating little LD) to 1 (indicating strong LD between alleles A and B).

For genetic epidemiology, in many applications of LD, one is interested in whether known markers are in LD with a possible disease susceptibility gene. This is illustrated in Fig. 4.12a, b:

In such a setting, one is searching for a functional SNP, say T, possible in the coding region of the disease susceptibility gene. In most genetic linkage and association studies, however, one is *unlikely* to have genotype data on that functional SNP. On the other hand, one may have genotypes of surrounding SNPs that are denoted A, B, and C: as illustrated in Fig. 4.12. In a hypothetical example, SNP A is located **within** the disease susceptibility gene, but is not the functional variant, while SNPs B and C are located outside of the gene, but are nearby on the same chromosome.

To illustrate this situation, assume that each of the markers is biallelic, with the allele frequencies shown in Table 4.11:

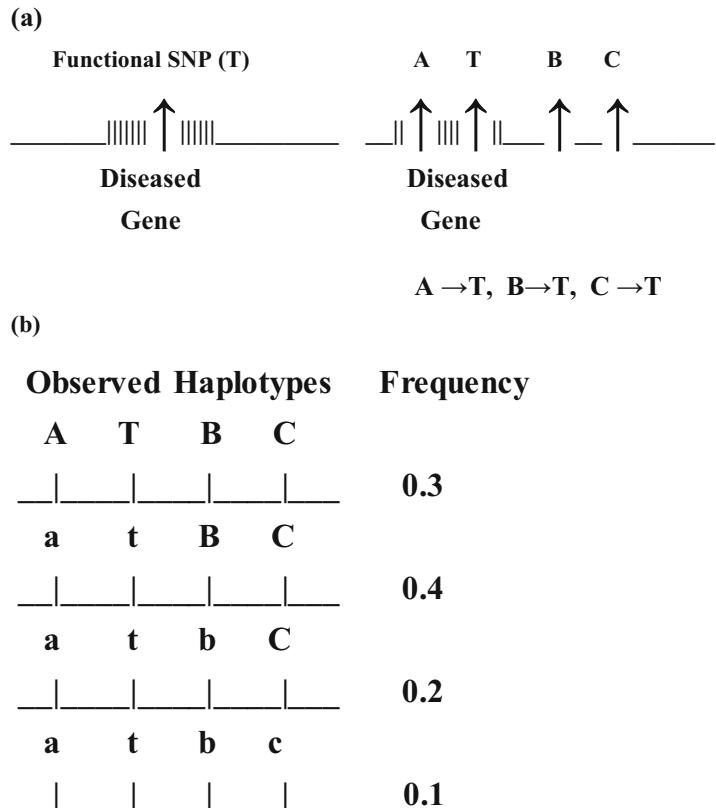
Thus, to calculate the LD between the possible disease susceptibility Loci T and Marker B:

$$\begin{aligned} D &= f(T - B) - f(T) \times f(B) \\ &= 0.3 - 0.3 \times 0.7 \\ &= 0.3 - 0.21 \end{aligned} \quad (4.67)$$

$$\begin{aligned} D_{\max} &= \text{minimum}[f(T)xf(b), f(t)xf(B)] \\ &= \text{minimum}[0.3 \times 0.3, 0.7 \times 0.7] \\ &= \text{minimum}[0.09, 0.49] \end{aligned} \quad (4.68)$$

Fig. 4.12

(a) Unknown functional variant T in diseased gene, and known SNPs A, B, and C (b) Observed haplotypes and frequencies in a hypothetical study sample for 4 bi-allelic loci: A, T, B, C

**Table 4.11** Linkage Disequilibrium Between 4 Biallelic Foci: See Fig. 4.12b

Allele Frequencies	D with T	D' with T	r^2 with T
$f(A) = 0.3, f(a) = 0.7$	0.21	1.00	1.00
$f(T) = 0.3, f(t) = 0.7$			
$f(B) = 0.7, f(b) = 0.3$	0.09	1.00	0.18
$f(C) = 0.9, f(c) = 0.1$	0.03	1.00	0.05

$$\begin{aligned}
 D' &= [D/D_{\max}] \\
 &= 0.09/0.09 \\
 &= 1
 \end{aligned} \tag{4.69}$$

$$\begin{aligned}
 r^2 &= D^2 / [f(T) \times f(t) \times f(B) \times f(b)] \\
 &= (0.09)^2 / [0.3 \times 0.7 \times 0.7 \times 0.3] \\
 &= 0.0081 / 0.0441 \\
 &= 0.18367
 \end{aligned} \tag{4.70}$$

Hence, in the illustrative example, locus T and locus B are:

- (i) in “complete” LD, based on D' , see the calculated result in Eq. (4.69), but
- (ii) **not** based on r^2 , according to the calculated result in Eq. (4.70)!

This apparent dichotomy is due to the fact that the allele B occurs on 2 different haplotypes: one with allele T, and the other with allele t: see Fig. 4.12b!

- (iii) A similar pattern is observed for the C locus: C-B and C-b !
- (iv) In contradistinction, the A allele occurs only on the haplotype with allele T.
- (v) Thus, allele A is in “perfect” LD with T, with both C and c, for which $r^2 = 1.0$.

In the human genome, there are many regions that have high LD, known as “haplotype blocks”, most genetic epidemiology research and studies determine LD between all combinations of genetic markers on a particular area of interest on a chromosome. Moreover, the size of such regions in the genome varies greatly, with significant differences between various ethnic groups. For instance, the mean size of haplotype blocks is 22 kb in Asian or Europeans populations, but is 11 kb for those of recent African ancestries.

4.3.1.11 Spectrum of Rare Genetic Variations

Most monogenic diseases are caused by mutations that reduce the function or stability of a single protein by altering its three-dimensional structure. These mutations include point mutations (e.g., changes in single nucleotides that alter the amino acid sequence), insertions, or deletions in the DNA sequence that encodes the protein; or changes in the non-coding DNA that interfere with gene splicing. To improve ones understanding of biological mechanisms of disease risks, it would be very helpful to first understand the role of gene-environment and gene-gene interactions (epistasis) in disease susceptibility which may lead to increasing the biostatistical power, accuracy, and precision for detecting genetic and environmental effects.

Two distinct study designs should be considered as strategies for identifying rare genetic variants underlying disease susceptibility:

- (1) Sequencing relatives in families with multiple affected individuals, and
- (2) Extreme trait sequencing.

4.3.1.12 Sequencing Relatives in Families with Multiply Affected Cases

This strategy involves sequencing the distantly related and affected family members: the more distantly related these family members are, the fewer genes variants they will share. However, by assuming that the disease of interest is caused by the same variant in both relatives, they should share that causal variant.

For example, first cousins share approximately 1/8 of their genomes IBD (Identity-By-Descent). However, when applying this approach, distant relatives may still share too much of their genomes to permit simple identification of the causal variants. Hence, the associated variants should first be filtered out, basing on functions, allele frequencies, and/or the types of genes affected. Moreover, for reducing the variants of interest, one may use evidence of modest linkage based on LOD (Limit-Of-Detection) score analysis.

4.3.1.13 Extreme Trait Sequencing

The second strategy, well-suited for quantitative traits, is to sequence a small number of selected case-subjects from one or both ends of the phenotype distribution. The frequency of alleles contributing to disease susceptibility is expected to be enhanced in these extremes of the distribution, and thus may be identified based upon a relatively small number of cases and controls.

The potential for this strategy was successively shown in a study of HDL (High-Density Lipoprotein) cholesterol. In that study, three candidate genes were sequenced in only 38 case-subjects from the lower and upper 5% of the HDL cholesterol distribution. The results showed that rare alleles had significant effects on low HDL cholesterol: a major risk factor for coronary heart disease. In one of the first findings from the NHLBI ESP project, extreme phenotypes were used to ascertain that variants in the *DCTN4* gene, which encodes a dynactin protein, is associated with chronic *Pseudomonas aeruginosa* infection among case-subjects with cystic fibrosis. Such infections are related to reduced lung function and shorter mean survival. In that study:

- 43 case-subjects with an early age of onset for the infection were compared with
- 48 cases-subjects of the oldest individuals who did not have a chronic infection,

The presence of one or more missense mutations in the *DCTN4* gene was associated with early age at onset of the infection:

$$\text{HR (Hazard Ratio)} = 1.9 \ (p = 0.004) \quad (4.71)$$

It appeared that the success of this exome sequencing study of a complex trait with a small sample size is due to the well-matched extremes, the large effect size of the gene, and the relatively high combined Minor Allele Frequency (MAF) of 0.065 for the rare variant examined. Initial analyses based upon deep sequencing of rare coding variation of 2440 case-subjects from ESP project show that large sample sizes of rare variants may be required for sufficient power in association studies of rare variant with many complex diseases.

4.4 Factors in Human Genetic Epidemiology

Familial factors in human genetic epidemiology consist of the following important issues:

- (1) Family studies which remain a fundamental aspect of human genetic epidemiology for discovering genes which have disease susceptibility
- (2) LOD score linkage analysis which compares the number of recombinant and non-recombinant offsprings in a search for evidence of co-segregation between a hypothetical disease gene and a recognized genetic marker.

Remarks

LOD – a measure of the strength in Genetic Linkage

One of the fundamental approaches for tracking disease genes is the use of the **LOD** linkage score: LOD stands for **Logarithm of the Odds (to the base 10)**. Thus, a LOD score of three or more is generally taken to indicate that two gene loci are close to each other on the chromosome. A LOD score of three means the odds are a thousand to one, in favor of genetic linkage, since:

$$\log_{10}(Y/X) = 3 \iff (Y/X) = 10^3 = 1,000 \quad (4.72)$$

- (3) Non-parametric linkage analysis uses excess sharing of marker alleles Identical By Descent (IBD), particularly siblings, for a linkage analysis approach in the absence of multi-generational data.
- (4) Genetic association studies may also be used for families, often using parents-offspring trios.
- (5) Family studies in genetic epidemiology have progressed from linkage analysis to exome sequencing for finding disease susceptibility genes, providing important progress in the understanding of genetic susceptibility to rare and familial forms of diseases, as well as common, complex diseases.
- (6) These advances are supported with examples from on-going research on familial cancers of the breast, pancreas, and prostate, etc.
- (7) The goal of these research approaches is to use molecular discoveries to improve diagnosis, treatment, and prevention of diseases.

4.4.1 Linkage Analysis

The objective of linkage analysis is to identify the approximate chromosomal location of a disease susceptibility gene. Using LOD score linkage analysis, the fundamental approach for mapping disease genes is based on searching for evidence of co-segregation of disease with known generic markers within families. Thus the objective of linkage analysis is to identify the approximate location of disease susceptibility genes. Thus, the biological and physiological bases, as well as the concomitant interpretation of linkage analysis is clear.

During meiosis and the formation of gametes, alleles on the same chromosome (the linked gene) are inherited together except when crossing over (viz., recombination) occurs during the first division of meiosis:

- (1) Homologous material and paternal chromosomes exchange genetic material. centromere during Prophase I of meiosis.
- (2) Then recombination, or crossing-over, occurs when sister chromatids on the paternal and maternal chromosomes exchange genetic material.
Although this is a normal process, it breaks up associations between alleles along the same chromosome during cell division.
- (3) The closer together the 2 genes are on the same chromosome, the less likely recombination between them becomes. Note that, according to Mendel's Law, genes on different chromosomes segregate independently, and therefore are not genetically linked.
- (4) Therefore, the frequency of crossing over depends on the physical distance between 2 syntenic genes, creating a means to identify deviation from independent segregation, and to estimate the genetic distance between two genes.ioslt hypothesis of no linkage with the alternate hypothesis of linkage gives a biostatistical test for linkage, and allows the genetic distance to be estimated.

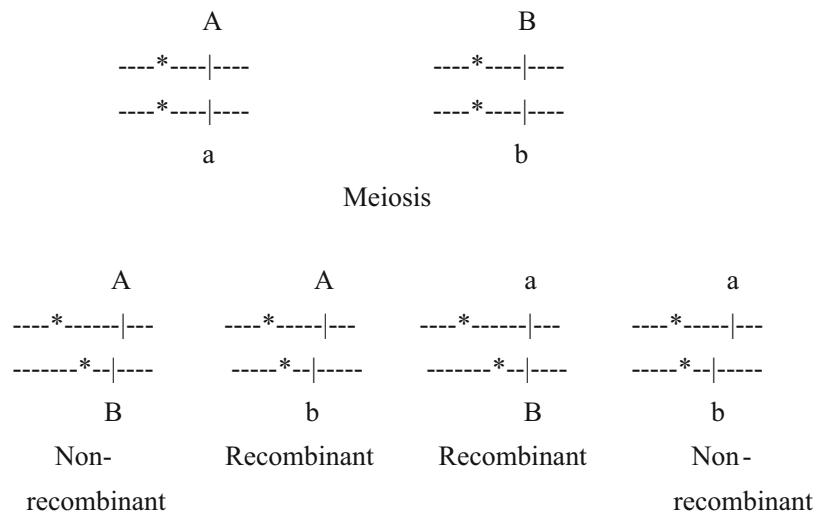
The foregoing approach may be used to show evidence of co-segregation between an observed marker and a hypothetical gene controlling a disease phenotype which provides evidence of linkage to map a previously unknown causal gene. Since linkage is based upon the physical distance between 2 genes, it is a property of loci along a chromosome, not the 2 alleles at these loci.

An Illustration of Linkage Analysis

For 2 homologous genes of chromosomes, Figure 4.13 illustrates 2 pairs of such genes that a, with locations of unlinked genes A and B, and gametes resulting from meiosis:

- Genes A and B are **not** linked, but are on different chromosomes.

Fig. 4.13 Two homologous pairs of chromosomes, locations of unlinked genes A and B, and gametes resulting from meiosis



- Both apparent non-recombinant (A-B and a-b), and recombinant (A-b and a-B) gametes are all equally possible under Mendel's Law of Independent Assortment
- $\theta = P(\text{observing a recombinant gamete}) = \frac{1}{2}$

Hence, θ ranges from 0 (for complete linkage), to $\frac{1}{2}$ (for no linkage), and values between 0 and $\frac{1}{2}$ indicating possible linkage of genes A and B.

LOD scores are biostatistics for calculating the linkage between a known genetic marker and a potential disease gene, and are based on unscaled likelihood function $L(\theta)$ that are proportional to the probabilities:

$$L(\theta) = \text{Likelihood}(\theta | \text{Family Data}) \propto P(\text{Family Data} | \theta) \quad (4.73)$$

Equation (4.73) may be used to compute the likelihood of observing the co-Segregation of a causal disease allele and the observed markers.

4.4.2 Family Association Studies

Linkage studies focus on testing hypothesis about the recombination rates, θ , and are based upon families, from large multiplex families covering 3 more generations to pairs of affected siblings drawn from a nuclear family. Association studies, on the other hand, test for differences in allele, genotype, or haplotype frequencies between groups of unrelated people or test for deviation from Mendelian expectations transmission of marker alleles within families determined through an affected case. In either case, association tests do not focus on recommendation rates within families. Rather, they test a composite null hypothesis of no linkage: $\theta = 0.5$, or no **Linkage Disequilibrium** (LD), which reflects correlation owing to tight linkage between an observed marker allele and an unobserved high-risk allele.

LD, also known as Gametic Phase Disequilibrium, represents a deviation from two-locus Hardy-Weinberg Equilibrium (HWE) in which allele frequencies at different loci are products of the individual allele frequencies. The presence of LD implies allele frequencies at different loci are correlated with one another. For unlinked loci, mixing among distinct populations can create detectable gametic disequilibrium or deviation from 2-locus HWE which will transform back to true equilibrium within a few generations if there is random mating. When 2 loci are linked, the reduced recombination, $\theta \ll 0.5$, between them permits the LD to persist for many generations. In this way, LD becomes a result of tight linkage and can be used to map genes through association studies via either unrelated individuals or a minimal family structure such as children and their parents.

Since association studies depend on LD, which spans much smaller distances, typically of the order of hundreds of kb of physical distance, tests for LD have the advantage of resulting in a much smaller chromosomal region of signal. On the other hand, LD is also a complex reflection of the genetic history of the population, and LD blocks vary across populations. Hence, stable populations that are developmentally older will have had more time to approach linkage equilibrium, resulting in smaller blocks of LD, rendering it more difficult to cover the whole genome. However, younger populations will have larger LD blocks spanning greater physical differences, requiring fewer markers to cover the whole genome.

The presence of locus heterogeneity (where more than one locus exists) creates an Intrinsic weakness in linkage analysis. However, this may be addressed by adding another parameter in the LOD score method, although this may create some reduction in the biostatistical power. Similarly, locus heterogeneity will reduce power to map genes using non-parametric linkage methods. However, association studies are susceptible to the detrimental effects of allele heterogeneity (multiple mutations in the same gene – leading to diseases) since these multiple mutations are not likely to be in LD with the same markers around the gene.

4.4.2.1 The Transmission Disequilibrium Test (TDT)

As a test of observed versus expected patterns of transmission of marker alleles in families tested by a case, family-based tests of association is often used. The TDT has been introduced, using the smallest possible family structure, such as an affected child and both parents, known as a “case-parents trio” or “triad”: Fig. 4.14

In this test configuration, one is primarily concerned with a marker M with 2 alleles: M₁ and M₂. This study design is using the TDT for both association and linkage between the allele of interest (M₁, in this case) and the disease. This analytical approach requires at least one parent: in the trio to be heterozygous for the marker: denoted as M₁M₂. Now, under Mendel’s Law of Equal Segregation, Sect. 4.3.1.1, the M₁ and M₂ alleles are equally likely to be transmitted from the heterozygote parent to an affected child if the marker is totally independent of any gene controlling risk to disease. If the marker M is linked to, and in LD with, an unobserved causal gene, then one would expect the M₂ allele to be

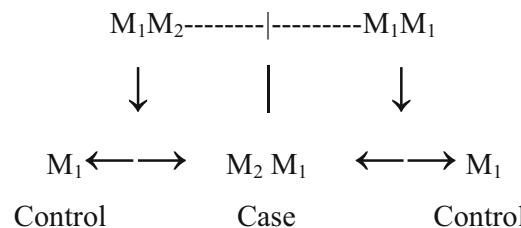


Fig. 4.14 Transmission disequilibrium test for a family trio: with an affected offspring (Case) and a heterozygote parent

transmitted more often from the heterozygote parent: Fig. 4.14, and the underlying hypotheses may be expressed as:

$$H_0 : P(\text{Parent transmits } M_2 | \text{parent is heterozygous } M_1M_2) = 0.5 \quad (4.74)$$

$$H_1 : P(\text{Parent transmits } M_2 | \text{parent is heterozygous } M_1M_2) > 0.5 \quad (4.75)$$

The TDT determines if the segregation of the M_1 and M_2 alleles is not equal by comparing the frequency of the transmitted versus the non-transmitted alleles. Each case-parent trio contributes 2 matched case-control pairs to the table, one from each parent, representing a matched pair of transmitted versus non-transmitted alleles: Fig. 4.14. The test will compare whether the probability of transmission of the M_2 allele from a heterozygous parent to an affected child differs from 0.5. For the TDT test, the data may be shown as in Table 4.12, where:

B is the count of transmitted M_2 alleles,
C is the count of non-transmitted alleles, and
N is the total count of transmitted alleles.

Under Mendelian inheritance, these frequencies are expected to be equal, viz., $B = C$ in Table 4.12. The Null Hypothesis for the allelic TDT test may be written as:

$$H_0 : D(1 - 2\theta) = 0 \quad (4.76)$$

where $D = LD$, and $\theta = \text{recombination fraction}$, which represents no LD, viz., $D = 0$, and $\theta = 0.5$, corresponding to no linkage. Thus

$$D(1 - 2\theta) = 0 \quad (4.77)$$

represents a composite null hypothesis of no linkage or no association between the marker and the hypothesized disease gene. The alternate hypothesis may then be expressed as:

$$H_0 : D(1 - 2\theta) \neq 0 \quad (4.78)$$

viz., there is no association: $D \neq 0$, and no linkage: $\theta \neq 0.5$, between the marker and the disease gene.

Using the layout in Table 4.12, the biostatistical test for H_0 is then a McNemar's Test with one degree of freedom:

$$\chi^2 = [B - C]^2 / [B + C] \quad (4.79)$$

If this χ^2 test is biostatistically significant, leading to rejection of the composite null hypothesis, it may then be concluded that the M_2 allele is transmitted to the affected offspring more or less often than expected under Mendelian inheritance. Thus this marker is linked and associated with some unobserved gene controlling risk to the disease under investigation.

Table 4.12 Transmission disequilibrium test for heterozygous (M_1/M_2) parents transmitting N Alleles

		Number of transmitted alleles		
		M_2	M_1	Total
Observed	B	C		N
Expected	N/2	N/2		

Table 4.13 Transmission disequilibrium test for Class 1 Alleles at a 5' VNT marker in the insulin gene and risk of IDDM

	Number of transmitted alleles		
	Class 1	Other	Total
Observed	78	46	124
Expected	62	62	

$$\chi^2 = (78 - 46)^2 / (78 + 46) = 8.258,065, P = 0.004$$

Worked Example I: Application of Transmission Disequilibrium Test–Insulin Gene and Insulin-Dependent Diabetes Mellitus (IDDM)

Austin, using the format of 4.12, and published data, showed an allelic TDT test for the Class-1 alleles, for the smallest fragment sizes, at a VNTR marker located within the 5'-region of the insulin gene and risk of Type-1 Diabetes (IDDM). For a sample of some 94 families with 2 or more IDDM children, there are 57 heterozygote parents who transmitted 124 alleles to the affected offsprings. Of these, 78 alleles, or 63%, were Class 1, and 46 alleles, or 37%, were other alleles, compared with 62 alleles, or 50%, each expected under Mendelian inheritance: Table 4.13:

Remarks

Since this test statistic is significant, these data provide evidence for association and linkage between the marker and a gene controlling risk to IDDM (which may be the Insulin Gene itself!)

Worked Example II: Application of Transmission Disequilibrium Test–Prostate Cancer Genetics

Approach I: Gathering of Epidemiologic Data from the Public

Prostate cancer (PCa) is the leading cause of cancer in the USA, as well as the second leading cause of cancer-related deaths. Besides race and age factors, family history is a strong risk factor. Past epidemiologic studies have demonstrated that men with one or more first-degree affected relatives have a 2- to 3-fold increase in risk compared with men without any reported family history. Family history profile is often used to stratify 3 subsets of PCa cases:

Subset 1 – Sporadic: cases with no known family history of PCa:

75–80% of cases.

Subset 2 – Familial: cases with 1 first-degree or 2 second-degree relatives with the disease:

15–20% of cases.

Subset 3 – Hereditary (HPCa): cases with 2 or more first-degree relatives with the disease, or cases from families in which there are 3 generations – maternal or paternal lineage – of men with PCa, or cases with at least 1 first-degree relative with the disease when both were diagnosed with early-onset PCa:

5% of cases

In summary, observational studies, segregation studies, twin studies, and family-based linkage studies have provided strong evidence regarding an inherited component to the etiology of PCa, with a 42% of disease incidence estimated to be due to genetic predisposition.

Contributions to PCa susceptibility are being determined in recent studies to determine some of the genetic variants and mutations that contribute to PCa susceptibility. As from 1995, the Prostate Cancer Genetic Research Study (PROGRESS) was started – with the objective recruiting families which have a defined pattern of HPCa for linkage studies directed towards mapping genes. Across North America,

families were tested through public media as well as through contacts with members of relevant professional organizations.

Approach II: Linkage Analysis

Two genome-wide linkage studies have been completed on Progress families:

- (1) The first was based on genotyping 441 microsatellite markers, and
- (2) The second was based on genotyping of 6000 SNP markers.

The second effort identified several possible evidence for linkage at:

- (a) 7q21 (heterogeneity LOD, HLOD = 1.87),
- (b) 8q22 (non-parametric) and 15q13-q14 in families of European ancestry,
- (c) 2q24 in 12 African-American HPCa families

with aggressive PCa phenotype showed additional linkage regions. When early of diagnosis was considered (below 65 years), the linkage signal on 15q13-q 14 became more suggestive for a locus harboring an HPCa gene mutation.

This approach has been found to be useful for stratifying families for linkage analysis into subgroups with HPCa and primary brain cancer, kidney cancer, or colon cancer. These results seem to indicate how strategies which use clinical data and family characteristics may create more homogeneous subsets of HPCa families, and may be useful for identifying genetic linkage. Moreover, one may pool data across multiple studies to enhance biostatistical power for finding linkage in defined subsets of families.

Worked Example III: International Consortium for Prostate Cancer Genetics (ICPCaG)

Beginning in 1998, several groups with collected data of HPCa families from Australia, Finland, North America, Norway, Sweden, and the United Kingdom jointly form the ICPCaG, and the first combined analyses consisted of 772 HPCa families and evaluated a linkage region on 1q24-25, for which the data failed to provide a definitive confirmation! However, the next ICPCaG effort, which included 1233 HPCa families, revealed significant linkage at 22q12 in HPCa families, with at least 5 affected members. Further analysis showed of this result narrowed the region to a 15kb interval spanning only 1 gene: Apolipoprotein L3, or APOL3!

Summary Remark:

Advances in sequencing technology are yielding important insights into the genetics of HPCa, as well as some HPCa-associated mutations. These advances should expand the concept and knowledge of genetic susceptibility to PCa, and to improved way to diagnose, treat, as well as to prevent this common, but complex, cancer.

4.5 Human Genetic Association

The achievements of Human Genome Project (HGP), and subsequent advancements in genotyping, leads to an influx of new developments in genetics. Moreover, technology has provided scientists with a comprehensive data on human genomes as human genome is now capable to incite in depth and precise data information that allows access to detailed DNA sequences in order to analyze clinical questions. The methods used have been optimized to examine the application of Genome Wide Association Studies (GWAS) with population based forensic investigations. GWAS-associated approaches have also been incorporated in routine clinical practice.

4.6 Genetic Epidemiology Owing to Population Stratification

This issue is now in a critical stage of development: while some early work seemed to indicate that population stratification has great impact on genetic epidemiologic results, recent work seems to indicate otherwise!

1. In the early 1990s [Austin 3ff], genetic epidemiologists undertook genetic association studies of unrelated case subjects in search of susceptibility genes for many complex diseases, using either *genome-wide* or *candidate-gene* approaches. In the candidate-gene approach, genes whose functions are known to be involved in disease pre-disposition, as well as specific polymorphisms within those genes, were studied. However, Genome-Wide Association Studies (GWAS) do not rely on known etiologic mechanisms and examine genetic markers spanning the entire genome to identify new genes or confirm previously known susceptibility genes. They use high-throughput techniques for genotyping millions of SNPs on each sample. To maintain high biostatistical powers to detect the small effect sizes found in most GWAS, large samples are needed, and are usually based on large-scale collaborations spanning multiple studies.
2. In recent researches, the following results seem to be more acceptable:

It has been argued that biases from population stratification (the mixture of individuals from heterogeneous backgrounds) may undermine the credibility and reliability of studies in genetic epidemiology. Work had been undertaken to investigate the amount of biases that may be due to population stratification in U.S. studies of cancer among non-Hispanic Caucasians of European origin. [GJ] Using an expression of the Confounding Risk Ratio (the ratio of the effect of the genetic factor on risk of diseases, with and without adjustment for ethnicity) to measure the potential relative bias from population stratification, it was found that:

- Using empirical data on the frequency of the N-acetyltransferase (NAT2) slow acetylation genotype and incidence rates of female breast cancer and male bladder cancer in non-Hispanic U.S. Caucasians with ancestry from 8 European countries to assess the bias in a hypothetical population-based U.S. study that not take ethnicity into account.
- Theoretical computations were undertaken to obtain the bias over a large range of allele frequencies and disease rates.
- The results showed that, ignoring ethnicity, there is a bias of 1% or less in the empirical studies of NAT2
- Evaluation of a wide range of allele frequencies and representative ranges of cancer rates that exist across European populations show that the risk ratio is biased by less than 10% in U.S. studies except under extreme conditions.
- The bias decreased as the number of ethnic strata increases.
- It may be concluded that there will be *only a small bias* from population stratification in a well-designed case-control study of genetic factors that ignores ethnicity among non-Hispanic Caucasians of European origin.
- For other ethnic groups, further work seems to be needed to estimate the effect of population stratification within other populations.

In all likelihood, it appears that any bias owing to population stratification may be small in well-designed and analyzed case-control studies of Caucasian populations.

Methods for controlling for population stratification include:

- (1) matching,
- (2) restricting or adjusting for ethnic background, and
- (3) methods specific to genetic epidemiology such as
 - family-based methods including Transmission Disequilibrium Tests (TDT), or
 - biostatistical methods using unlinked genetic markers to detect, quantify, and correct for stratification.

4.7 Environmental Effects on Genetic Epidemiology^[Google]

Concerns for the environmental effects on genetic epidemiology are brought about by the goals to gain improved knowledge for the prevention of diseases, particularly immune disorders. With several child immune disorders such as Type-1 Diabetes and Multiple Food Allergies on the increase, genetic factors, though important, seem *not* sufficient to explain the increase in these diseases. Thus workers in this special aspect of genetic epidemiology aim to improve the understanding of all the factors driving this increased risk of child immune disorders and other diseases. The goal of this effort is to generate further knowledge that will lead to direct policy and prevention activity to decrease the risk of children developing *Type-1 Diabetes mellitus, multiple sclerosis, or child allergies*.

One of the largest studies in this group is the **Barwon Infant Study (BIS)**, in Australia. It is a birth cohort study with relatively intensive data and biospecimen collection at regular intervals from pregnancy until the child reaches school age. The Barwon Infant Study (BIS) is a major birth cohort study being conducted by the Child Health Research Unit (CHRU) at Barwon Health in collaboration with the Murdoch Children's Research Institute (MCRI) and Deakin University, in the State of Victoria, Australia. The objective of the BIS is to generate new knowledge on the best way to provide infants and children with a healthy start to life. More than 1000 pregnant women from the Barwon region of Victoria, Australia, were recruited between 2010 – 2013, and their children are now part of the invaluable BIS cohort.

Two typical projects in this area of investigating the environmental effects on genetic epidemiology are:

Project 1: The relationship between the maternal and infant microbiome and early life markers of cardiovascular risk

Cardiovascular disease (CVD) is a major source of morbidity and mortality and novel paradigms are required to inform prevention strategies. Although the majority of CVD research and prevention activity targets adults, CVD has its origins in early life. There are considerable interests in the potential relationship between early life gut microbiome and atherosclerosis, the inflammatory process that underlies CVD. The gut contains 10 times as many bacteria as there are cells in the human body. The composition and activity of the gut microbiome is influenced by modifiable factors such as mode of birth, microbial exposure, antibiotics and diet. In turn, the gut microbiome has a profound impact on immune development and function. It has been shown that maternal vaginal colonization with Group B Streptococcus in the third trimester of pregnancy is strongly associated with the infants aortic intima media thickness in the first months of life. This association is seen only among infants delivered vaginally, suggesting infant inoculation with the maternal microbiome is likely to be of relevance.

The Approach: This research program involves an investigation of the relationship between the various potential environmental determinants of the infant gut microbiome and markers at CVD risk

measured at 4 years of age. The project will prosecute a world class investigation of the relationship between the early life human microbiome and the development of markers of CVD risk in preschool aged children.

Project 2: A population based investigation of the relationship between maternal stress during pregnancy and offspring behavior and mental health

This work uses the Barwon Infant Study (BIS) platform to investigate the relationship between maternal antenatal stress and offspring neurodevelopment. The BIS protocol includes validated measurements of maternal stress and mental health during pregnancy and the first years of the child's life. The work would be involved in administering the 2 year BIS review, which would include developing expertise in a number of validated measures of childhood socio-emotional development (such as the Bailey's Developmental Inventory and Achenbach Child Behavior Checklist). There are a number of ways in which the BIS biosamples could be used to enhance this project. For example, there is a planned investigation of the relationship between maternal antenatal stress and the offspring's epigenetic profile. There is also substantial opportunity to develop the synergies between the proposed study and other aspects of BIS.

The Approach: Maternal stress and mental health are measured repeatedly during pregnancy and the first years of the child's life using the Perceived Stress Scale and the Edinburgh Postnatal Depression Scale. The child's socio-emotional development is primarily measured at 2 years of age using the Bailey's Development Inventory and the Achenbach Child Behavior Checklist. A wide range of covariates have been captured. Synergistic work regarding the antenatal factors and the infant's epigenetic profile are under way. A variety of analysis techniques will be used.

4.7.1 Environmental Factors on Genetic Epidemiology^[Google]

It is strongly believed that cancer is generally a polygenic multifactorial disease, which makes ***environment an important modifier in the risk of cancer.*** It is estimated that only 1 percent of cancers are caused by "cancer syndromes" and up to 5 percent result from highly penetrant single-gene mutations. Thus, the majority are polygenic. Studies with various animal and in vitro models, initiation and promotion models, adenoma carcinoma models, and immortalized human cells provide evidence that polygenic mechanisms are important in cancer, at least in experimental systems.

Most of the known cancer syndromes are monogenic and conform to a two-stage model of development; that is, they require inactivation of two copies of a tumor suppressor gene in order to initiate. These syndromes tend to be dominant Mendelian conditions, which can be assessed in family studies covering two or more generations. However, such studies provide no data on recessive Mendelian conditions and have a limited resolving power in polygenic conditions. Consequently, apart from highly penetrant single-gene mutations, the risks posed by low-penetrance single-gene mutations, polygenes, and recessive genes are not well understood.

Hemminki described a study of data obtained from 44,000 same-sex twin pairs to assess cancer risks for co-twins of twins with cancer. There were almost 10,000 pairs in which one of the members had cancer. The analysis of environmental and inherited contributions was based on correlations between monozygotic twins who share the genome completely, that is, 100 percent concordance in their genomes. A similar concordance was carried out with dizygotic twins, the difference being the assumption that only 50 percent of the genes are common. The assumption is that the environment is affecting monozygotic and dizygotic twins similarly. Some of these different effects will then be 100 percent, or 1. The nonshared random environmental effect was the largest factor for all cancers,

accounting for 58 to 82 percent of the mounting evidence supports the concept that cancer is generally a polygenic multifactorial disease, which makes environment an important modifier in the risk of cancer, stated Kari Hemminki, Karolinska Institute. It is estimated that only 1 percent of cancers are caused by “cancer syndromes” and up to 5 percent result from highly penetrant single-gene mutations; thus, the majority are polygenic. Studies with various animal and in vitro models, initiation and promotion models, adenoma carcinoma models, and immortalized human cells provide evidence that polygenic mechanisms are important in cancer, at least in experimental systems.

Almost all of the known cancer syndromes are monogenic and conform to a two-stage model of development; that is, they require inactivation of two copies of a tumor suppressor gene in order to initiate. These syndromes tend to be dominant Mendelian conditions, which can be assessed in family studies covering two or more generations. However, such studies provide no data on recessive Mendelian conditions and have a limited resolving power in polygenic conditions. Consequently, apart from highly penetrant single-gene mutations, the risks posed by low-penetrance single-gene mutations, polygenes, and recessive genes are poorly understood.

Heritable and Environmental Effects from Twin Studies

Twin studies as tools for understanding genes, the environment, and cancer

Genetic: if monozygotic twins are more similar for a given trait than dizygotic twins

Shared Environment (e.g., diet and childhood experiences): if there is twin similarity not accounted for by genetic effects

Nonshared Environment: anything that is not hereditary and not shared between relatives, that is, sporadic causes of cancer total variation.

The Largest Sample

A Swedish family cancer database, containing 10 million people, is the largest population-based data set ever used for studies on familial cancer, said Hemminki. The data are used to develop estimates for the environmental and inherited components in cancer, using the genetic relationships among family members to calculate the effects of genotype, shared environment, and non-shared environment. The database has been used in modeling cancer causation and has revealed that *environmental causes explained most of the total variation for all neoplasms except thyroid cancer*, for which heritable causes were largest. There also appears to be a subgroup of cancer patients who develop a second cancer to which there is a strong genetic predisposition, that often cannot be predicted by a family history. This phenomenon is typical of polygenic diseases.

Hemminki reported that the twin and family data quantified non-shared environmental effects as ranging from 40 to 90 percent for different cancers. It is of interest to note that this effect was large for some cancers of identified environmental causes, such as lung and cervical cancers. In contrast, shared environment—common family experiences and habits—accounted for 0 to 30 percent of cancer etiology.

For all cancer, the genetic effect was estimated to be 26 percent; however, there is evidence supporting heritability for all cancers.

The data presented by Hemminki on twins, families, and second cancers provide additional support to the multistage theory of carcinogenesis. If most cancers are indeed polygenic, this should be adequately considered in study designs for gene mapping approaches. Linkage analysis in families of multiple affected individuals is not sufficient to identify cancer-related genes, said Hemminki.

Instead, what are needed are large case-control studies with stringent clinical criteria so that the different types of cancer can be distinguished and there is a large enough sample size to enable even the rare homozygotes to be scored, emphasized Hemminki. In addition, it will be important to study people with multiple cancers or second cancers, because they can provide a good indication of whether polygenic effects are operating.

4.8 Genetic Epidemiology and Public Health^[Google]

Genetic Epidemiology and Public Health: The Evolution from Theory to Technology

Genetic epidemiology represents a hybrid of epidemiologic designs and statistical models that explicitly consider both genetic and environmental risk factors for disease. It is a relatively new field in public health. Meanwhile, the field has gone through a major evolution, changing from a field driven by theory, without the technology for genetic measurement or computational capacity to apply much of the designs and methods developed, to a field driven by rapidly expanding technology in genomic measurement and computational analyses while epidemiologic theory struggles to keep up. In this discussion, several progressing eras of genetic epidemiology are presented, spanning this evolution:

- from theory to technology,
- what have been learned,
- what have been added to the broader field of public health, and
- what remains to be done.

Genetic epidemiology represents a hybrid of epidemiologic designs and statistical models that explicitly consider both genetic and environmental risk factors for complex diseases, that is, those diseases that have some genetic component to their etiology but are not exclusively Mendelian. In the past 4 decades, genetic epidemiology has emerged as an important area of public health research, one that could eventually have multiple influences on public health practice. Although most of genetic epidemiology focuses on research, the area of “public health genetics” has incorporated these tools into the practice of public health. Understanding how genes function and interact and characterizing the role of genetic predisposition in human disease may help accomplish many of the goals of public health, including:

- (1) improved prediction of individuals at risk,
- (2) design and implementation of targeted biologic interventions, and
- (3) deeper insights into the biology of a disease, all of which should combine to improve prevention and intervention strategies.

Findings from genetic epidemiologic research may also guide policy and recommendations for public health services and may help tailor pharmacological therapies to prevent adverse events and maximize efficacy based on an individual’s genetic makeup, an area now termed pharmacogenetics.

Much of the current development in genetic epidemiology has been focused on the medical concept of individualized medicine. Nevertheless, the population-based approach of genetic epidemiology is the true history of how public health genetics has grown and flourished. The most striking changes in genetic epidemiology have been in response to technological improvements that have moved from theoretical applications of population genetics in public health to studies of genomic sequence information at the individual level. Thus, although the early years of this field witnessed great

theoretical advances, the technology needed to support testing of those theories was lacking. Much of the design, biostatistical, and computational advances in the past several decades have been driven by ever-expanding technology for measuring the human genome. With this in mind, one may describe 3 eras in genetic epidemiology:

- (1) the era before polymorphic DNA markers were discovered,
- (2) the “pre-genome” era, and
- (3) the “post-genome” era.

Here, first consider these different eras, what defined them, and how they contributed to genetic epidemiology, and relate them to the current era in which genomic sequencing has become more affordable and can now be incorporated directly into research and practice.

(1) The Era Before Polymorphic DNA Markers were Discovered (1956–1979)

Recalling that the work of Watson and Crick was only reported in 1953, therefore, the field of genetics, especially its applications in medicine and public health, is still relatively young. During the latter part of the 20th century, genetics developed into a formal science with a clear definition of Mendelian inheritance and its consequences at the population level. Mendelian diseases are generally individually rare, but collectively they create a major public health burden; thus, efforts to identify genetic backgrounds and develop effective screening and treatment strategies are completely justified. However, identification of causal DNA variants, a biochemical understanding of these disorders, and development of DNA-based screening tests were rarely possible in this era. Genetic markers were limited to blood groups (ABO) or serum protein markers. Their locations in the human genome were not precisely known, and available markers did not come close to covering the entire genome. However, population-based surveillance systems to monitor and control some Mendelian genetic diseases, such as phenylketonuria, were implemented in the 1960s, showing early on the potential impact of the study of genetics on public health.

The statistical principles of linkage analysis were first developed in the middle of the 20th century using maximum likelihood estimation to identify genes or chromosomal regions that harbored disease mutations based on multiplex families (i.e., those with ≥ 2 affected members). The first linkage analysis tools were parametric models for categorical traits (a.k.a. logarithm of odds score approaches), which were proposed in 1953–1955. These biostatistical methods used multiplex families to identify genetic markers that segregated through families in a pattern similar to that seen with a putative disease gene. It was assumed that the gene causing the disease must be genetically linked (i.e., in close proximity) to the observed genetic markers. This design and statistical method required researchers to make multiple assumptions about the properties of inheritance of the disease within a family, such as whether there was Mendelian dominant or recessive inheritance and expression of a disease gene, the risk of disease conferred by the risk allele (denoted *penetrance*), the risk of disease in the absence of this risk allele (denoted *phenocopy rate*), and the prevalences (or allele frequencies) at the genetic marker in the population and at the causal gene. Models that required fewer assumed parameters can be traced to Penrose in 1935, although these affected relative pair designs and statistical approaches were not commonly pursued until much later.

The population concept of Linkage Disequilibrium (LD) was also defined in this period. It described how genetic variants on the same chromosome are genetically and spatially correlated in a population sample. This spatial correlation, which reflects the population history and proximity of 2 genes in the genome, ultimately allowed tests for genetic association in unrelated individuals using simple tests of association between a genetic marker and disease without knowing the actual causal allele, as discussed below.

This era could have been termed “the era of genetic theory”, because major advances in genetics and biostatistical methods occurred during this time. However, there were not enough useful markers mapped and cataloged to support linkage or association approaches that could be applied across the genome to identify disease genes using either family or population-based study designs. The scientific principles of both linkage analyses and population genetics (including LD) were all developed and recognized in this era, although implementation based on genome-wide marker sets and user-friendly computer programs were not available until the last quarter of the 20th century. Thus, the impact of genetic epidemiology on public health was delayed until technology caught up.

(2) The Pre-Genome Period (1980–2001)

Beginning around 1980, the discovery of polymorphic DNA markers (either restriction fragment length polymorphisms or simple tandem repeats, also called microsatellite markers) led to the development of genome-wide DNA marker panels. These panels were first used for genome-wide linkage analysis to map Mendelian genes. The use of these highly polymorphic, multi-allelic markers (i.e., markers with several alleles each) maximized the information content of families and led the development of multiplex genotyping technologies (i.e., measuring genotypes at many loci in a single assay) that increased efficiency and marked the beginning of high-throughput genotyping efforts. The linkage analysis approach relied on valid genetic and physical maps of marker locations and their distances along the chromosomes that were created using resources such as the Marshfield Medical Research Foundation genetic maps. This expansion also prompted a surge in new statistical techniques to handle both analysis of individual marker linkage (single-point linkage analysis) and analyses that included multiple markers simultaneously (multipoint linkage analysis), which was itself an important step in high-dimensional genetics data management and analysis. In this era, one moved from simple linkage theory to conducting genome-wide linkage screens with approximately 500 polymorphic simple tandem repeat markers scattered throughout the genome using multi-generation multiplex families. For diseases with Mendelian inheritance, there was a very high probability of successfully mapping chromosomal regions that harbored causal genes. For example, the identification of mutations in the amyloid precursor protein and presenilins for early-onset familial Alzheimer’s disease led to major breakthroughs in the biological understanding of Alzheimer’s disease.

Genome-wide linkage studies using multiplex families were also attempted for complex diseases, but the success rate was much lower largely because of the inherent causal heterogeneity of complex diseases. Different families reflected the effects of different disease genes, and different combinations of genes and/or environmental contributed to risk. In many countries, recruiting large, multiplex families was difficult because of secular trends in family size and the late ages of onset for many common diseases. To address this, methods of nonparametric linkage analysis based on affected sibling pairs (or other types of affected relatives) were developed and used alongside likelihood-based logarithm of odds score methods. These approaches successfully identified genes, such as the breast cancer susceptibility genes 1 and 2 (*BRCA1* and *BRCA2*), that cause inherited forms of breast cancer (a complex disease) despite considerable locus heterogeneity among families. Only about 5% of breast cancer cases can be attributed to mutations in recognized cancer-susceptibility genes, but understanding the etiology of breast and ovarian cancers in this high-risk subset does allow effective intervention on some individuals, driving a substantial screening effort in modern times. Thus, linkage analysis has led to effective screening methods for genes that control the risk of common diseases, as well as to subsequent intervention strategies.

Owing to the applications of genetic epidemiology in public health in this era were based on families with rare mutations in a single gene that exerted a large effect on risk, roughly following Mendelian dominant or recessive patterns, this period could also have been termed the Mendelian era

of genetic epidemiology. The field of public health genetics incorporated genetic knowledge and intervention strategies based on genetic counseling for several Mendelian diseases, such as population-level implementation of prenatal and newborn testing for metabolic diseases, such as phenylketonuria. More recent public health genetics efforts to control Mendelian diseases such as β -thalassemia in high-risk Mediterranean countries through population level screening, public health education, and genetic counselling has resulted in remarkable declines in the occurrence of these diseases.

Toward the end of the 20th century, the field began to recognize the utility of traditional epidemiologic designs for genetic epidemiology, representing a shift away from the collection of large, multiplex families to smaller nuclear families and ultimately a focus on case-control and cohort studies of unrelated individuals. This was in part due to recognizing how exploiting LD could also allow researchers to effectively map or localize genes at shorter genetic distances and perhaps with greater efficiency. Technologically, there was also a shift to the use of Single Nucleotide Polymorphism (SNP) markers instead of simple tandem repeat markers, which was driven by technologic improvements in microarray genotyping arrays (also known as SNP chips) and by the realization that although each individual biallelic SNP was intrinsically less informative, collectively they could “tag” most of the haplotype blocks present in a population and thereby cover the entire genome.

In summary, the impact of this era could be characterized as genetic discovery that informed screening and prevention. Linkage analysis results were less fruitful for non-Mendelian diseases for which multiple genetic, environmental, or combined etiologies exist. This set the stage for a move to large-scale genetic association analyses that would dominate the next era, particularly as genotyping and sequencing technology began to exceed an exponential pace. For example, the costs of sequencing the genome have decreased far faster than predicted by Moore’s law for advances in technology.

(3) The Post-Genome Era (2002–2010)

The sequencing of the human genome in 2001, and the development of large SNP marker panels soon after, along with annotation of the between-person variability (heterozygosity) and between-marker spatial correlations (due to LD), led the way for whole-genome association studies beginning early in the 21st century. This era has included wide-scale use of genome-wide association studies (GWAS) based on SNP markers, which became easier to genotype on a large scale via emerging SNP array technology. GWAS typically involve at least 100,000 separate SNPs and currently may include 5 million SNPs. By exploiting the concept of LD, these SNPs can effectively “tag” most unmeasured genetic variation in a population, so that the specific causal genetic variant need not be on the panel itself for investigators to detect associations that would at least implicate key genetic regions containing causal genes. This is the same concept as linkage but for much shorter genetic distances and without requiring family-based designs. Thus, the search for a particular causal variant could be much narrower, and samples of unrelated individuals could be used, as in traditional epidemiologic and registry-based designs.

GWAS have been enormously successful in identifying genetic risk factors for complex diseases. Since they are based on preselected DNA markers, they may not identify actual causal variants within a gene, but they do highlight which genes are consistently associated with risk for a given disease. Such discoveries, even without specific causal variants, broadened the biological understanding of how genes may control risk or influence quantitative disease phenotypes. However, building public health policies upon even the best of these genetic risk factors is not simple. For the majority of complex diseases, there have been a large number of variants identified as being genome-wide significant through GWAS, with many disorders now showing 50 or more associated genetic loci. However, it has been difficult to identify a single specific genetic variant that can be targeted for the development of therapeutic interventions. Further, there have not been many single variants with risk effects strong

enough to provide realistic predictive ability for screening individuals or for diagnostic use. Polygenic risk scores based on weighted sums of multiple genetic risk factors, often thousands, show some promise in the identification of high-risk groups, but the clinical utility of polygenic risk scores for most disorders remains unknown. This period could rightly be called “the era of complex diseases in genetic epidemiology”, because it reflects a shift in focus from gene discovery for Mendelian diseases to discovery for complex diseases in which multiple genetic and environmental risk factors influence risk across the lifespan. Studies in which unrelated subjects from traditional epidemiologic designs (either case-control or cohorts) were used became the most common genetic epidemiology study designs. This also led to recognition of the need for extremely large sample sizes to provide enough statistical power to identify associations with only modest effects on risk (odds ratios in the range of < 1.5). This point has driven the successful collaborative consortium framework for genetic epidemiology, in which multiple studies are combined to carry out either pooled analysis or meta-analysis of a common phenotype across several studies, enabling sample sizes to accumulate to the 5- and 6-digit ranges. This joint analysis also requires careful epidemiologic design in the pooling of data and ensuring the same measurement of phenotypes across studies, a process called phenotype harmonization.

Another major advance in this time period was the increased ability to measure multiple types of DNA variation, including not only SNPs and rare single base mutations (termed single nucleotide variants) but also copy number variants, which include deletions and amplifications both rare and common. Measurement of these variations in turn allowed researchers to test for possible impacts on risk of disease. Thus, in psychiatric epidemiology, rare deletions and mutations in the same genes have increasingly been found to be associated with particular disorders.

The greatest frustration with GWAS has been the inability to account for a substantial proportion of the estimated heritability for many complex diseases. Heritability is a broad concept, best defined as the proportion of variance in a quantitative trait (or variance in the liability to a qualitative trait) attributable to unobserved genetic factors divided by the total phenotypic variance in that trait. This concept also dates back to practices in the early part of the 20th century. Heritability is typically estimated from familial correlations (for quantitative phenotypes) or from measures of familial aggregation in risk (for qualitative phenotypes). Although many of the complex diseases studied with GWAS have estimated heritability at or above 50%, meaning that over half of the variance should be due to some genetic factors, even the sum of highly significant markers consistently associated in multiple GWAS over large samples of individuals fail to add up to more than a small fraction of the total heritability. This has been dubbed *the missing heritability dilemma*, and it suggests several possibilities:

- (1) There are many more unknown genetic risk factors to be found;
- (2) heritability has been overestimated, possibly because of shared environmental risk factors that also influence the phenotype;
- (3) SNPs in GWAS are inadequate proxies for the causal variant(s) and thus provide greatly diluted estimates of effect sizes; or
- (4) these apparent genetic risk factors identified by GWAS are dominated by false positives.

Much of the work in the post-GWAS era, discussed below, is now focused on addressing this missing heritability challenge in one way or another.

Despite the lingering limitations of discovery, much has been accomplished through GWAS. Since the first GWAS-based discovery of the association of Complement Factor H gene (CFH) with age-related macular degeneration, more than 2000 robust associations of genes with common diseases have been identified. These genetic risk factors have been used to improve risk prediction, diagnostic

classification, and drug development (both for potential efficacy and to minimize adverse events). For example, genetic information in the solute carrier organic anion transporter family, member 1B1 gene (*SLCO1B1*) can predict who is at highest risk of myopathy as a consequence of statin therapy. This can improve adherence and minimize adverse events that have a direct effect on public health. This has fueled excitement about precision medicine based on genetic information that can influence and educate clinical medicine. Recently, it was advocated for a similar emphasis on precision prevention that highlights the role and importance of genetics in public health.

WHERE ARE WE NOW?

Around 2010 or so, we entered a separate post-GWAS era, where the current challenges are how to use the enormous wealth of genetic data already generated from families and population-based studies to improve public health and how to merge these data with next-generation sequencing data, as well as other biomarker and health data. A common emerging principle is the utility of combining genetic risk factors to characterize the genetic predisposition of a person, without direct identification of specific causal variants. Polygenic risk scores are now being used to capture heritable versus nonheritable risk, to provide estimated risk probabilities for individuals, and to improve discrimination among heterogeneous subsets of a disease. These risk scores may also be a way forward in driving discovery of environmental risk factors by using a single scalar metric of genetic liability per individual while also considering environmental exposures either through effect modification or cumulative risk or as part of a separate sufficient cause model. In addition, biological insights are emerging based on the particular sets of genes that have only moderate associations with disease individually but are enriched among genetic risk factors found by GWAS or on genes that show co-expression in particular tissues related to a disease. Although the field has identified only a few particular causal genetic variants for further translational work in public health, it has been extremely useful in identifying new biological pathways that should be considered in the prevention or treatment of the disease, and this is now a prevailing principle.

Sequencing technology is now driving the current era of genetic epidemiology. Sequencing has the advantage of identifying all variants in a region of the genome: rare and low-frequency single nucleotide variants, as well as common tagging SNPs. This new phase has been driven by 2 forces—the advancements in massive parallel or next-generation sequencing technology that is becoming affordable even for samples sizes of the scale seen in epidemiology (47, 48) and the impending completion of GWAS analyses and meta-analyses for many (if not most) complex diseases in large studies, which has identified numbers of genes as significant for disease etiology and so must now be considered areas for new biological research. Although GWAS are largely based on tagging SNPs showing a strong association with a phenotype despite their not being directly causal (all based on LD between the observed SNP and some unobserved causal element), sequencing holds the promise of identifying all sequence variants (both common and rare) in a chromosomal segment (or genome) without the need to rely on proxy information. Thus, it is up to the researcher to identify which genetic variants could be directly causal. This can be done with relative ease when the variant is in a coding region, is nonsynonymous, and obviously leads to a severe disruption in the gene product. However, predicting function of variants outside of coding regions remains a major challenge.

Potentially causal variants identified through sequencing are often quite rare, occurring only one or few times in a sample of cases and often never observed in control samples. At a population level, these infrequent observations make statistical testing and causal inference difficult. However, a gene that likely affects disease biology may be disrupted in different ways in different individuals, leading to similar diseases in each. Although any particular causal variant is rare in the population, the occurrence of any rare causal variant in that gene across a sample of cases versus controls would implicate that gene. Several statistical methods are available that pool or collapse rare variants of the same gene into a

group and collectively tested for association with disease. Whole exome or genome sequencing studies hold out the hope of identifying causal genes, regardless of whether the risk alleles have low or high frequencies in the population. However, the genome is quite large and still poorly understood, so analysis of genomic sequencing data remains a major challenge.

On the translation front, genetic discoveries continue to inform personalized medicine. One major emphasis for epidemiologists will be the design of and analytic considerations for prediction models versus etiologic research. A consortium-derived odds ratio for any single SNP (or some combination metric, like the polygenic risk score) based on several population-based and/or clinically derived samples cannot easily be translated into a meaningful prediction algorithm with real clinical or population-screening utility. This requires study designs targeted specifically toward the population undergoing the screening or clinical test, and statistical analyses focused more on accurate metrics will be needed when testing the original causal inference that results from a discovery phase study.

We have not reached the ultimate goal of fully understanding complex diseases (and we might be far from it) if the causal pathway involves mechanistic gene-gene and gene-environment interactions. Epidemiologic studies remain underpowered to detect higher-order models of risk effects despite sample sizes in the tens of thousands, especially if these interactive risks are of moderate effect sizes. The emergence of mega-consortium samples of more than 100,000 genotyped individuals for common complex diseases may improve our ability to identify gene-gene interactions, but harmonization of environmental measurements and their possible heterogeneity across populations remain important barriers in executing such large-scale efforts for gene-environment studies. Indeed, whole new paradigms may be needed to accomplish this ultimate goal of genetic epidemiology.

Finally, other genomic technologies have also pushed the frontier of genetic epidemiology. A person's inherited genome is fixed before birth, but the genome necessarily interacts with dynamic gene expression architecture, often in response to endogenous and exogenous environmental triggers, which themselves will change the protein and metabolome content of a particular tissue at a particular moment in time. Thus, both clinical epidemiology and analytic epidemiology are now working to integrate proteomics, metabolomics, transcriptomics, and epigenomics into genetic epidemiologic studies. This new effort brings the fundamentals of epidemiologic methods back into genetic epidemiology, because the field can no longer rely on a static measurement of the genome made at any time in the disease process or at any age without concern about reverse causation, timing of risk susceptibility, or confounding by factors other than simple ancestry.

The future will bring unimaginable insights and new public health implementations based on genomic discoveries. In only 60 years, one had gone from the original identification of the double helix in the middle of the 20th century to the ability to sequence the entire genome of individuals in the first quarter of the 21st century. In just the past 15 years, one has gone from having an incomplete draft sequence on a few individuals to having full genome sequence catalogued for thousands of individuals, soon to be hundreds of thousands. One has moved from a primarily theoretical field to Mendelian discoveries that have directed public health screening and yielded therapeutic insights, and now to massive data generation that can inform prediction, screening, therapeutic discovery, and individualized medicine when attention is paid to epidemiologic principles. It is an exciting time for the field of genetic epidemiology in this regard; as genomic technologies continue to produce the potential for unprecedented molecular data, there will be increasing need for careful design, analytic, and inferential methods.

Special References

- Choi M, Scholl UI et al (2009) Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proc Natl Acad Sci USA* 106(45):19096–19101. [PubMed]
- Elston RC, Stewart J (1971) A general model for the genetic analysis of pedigree data. *Hum Hered* 21(6):523–542. [PubMed]
- Kennedy GC, Matsuzaki H et al (2003) Large-scale genotyping of complex DNA. *Nat Biotechnol* 21(10):1233–1237. [PubMed]
- Klein RJ, Zeiss C et al (2005) Complement factor H polymorphism in age-related macular degeneration. *Science* 308 (5720):385–389. [PubMed]
- Lander ES, Linton LM et al (2001) Initial sequencing and analysis of the human genome. *Nature* 409(6822):860–921. [PubMed]
- Lupski JR, Reid JG et al (2010) Whole-genome sequencing in a patient with charcot-marie-tooth neuropathy. *N Engl J Med* 362:1181–1191. [PubMed]
- Matsuzaki H, Dong S et al (2004a) Genotyping over 100,000 SNPs on a pair of oligonucleotide arrays. *Nat Methods* 1 (2):109–111. [PubMed]
- Matsuzaki H, Loi H et al (2004b) Parallel genotyping of over 10,000 SNPs using a one-primer assay on a high-density oligonucleotide array. *Genome Res* 14(3):414ff. [PubMed]
- Morton NE (1955) Sequential tests for the detection of linkage. *Am J Hum Genet* 7(3):277–318
- Morton NE, Chung CS (1975) Genetic epidemiology. Academic New York, New York
- Neel JV, Schull WJ (1954) Human heredity. University of Chicago Press, Chicago
- Ott J (1974) Estimation of the recombination fraction in human pedigrees: efficient computation of the likelihood for human linkage studies. *Am J Hum Genet* 26(5):588–597
- Ozaki K, Ohnishi Y et al (2002) Functional SNPs in the lymphotoxin-alpha gene that are associated with susceptibility to myocardial infarction. *Nat Genet* 32(4):650–654. [PubMed]
- Risch N, Merikangas K (1996) The future of genetics studies of complex human diseases. *Science* 273 (5281):1516–1517. [PubMed]
- Roach JC, Glusman G et al (2010) Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science* 2010. [PubMed]
- Venter JC, Adams MD et al (2001) The sequence of the human genome. *Science* 291(5507):1304–1351. [PubMed]



Human Genetic Epidemiology Using R

5

Abstract

This chapter illustrates the biostatistical human genetics with an introduction to the T-Test, and the Manhattan Plots in statistics using Human Genetic Data concepts leading to procedures for Statistical Tests and Utilities for Genetic Association. Several R packages are chosen: iGasso, GenABEL, and HardyWeinberg. Other topics discussed include:

- Regression Decision Trees and Classifications
- Multi-dimensional Analysis in Genetic Epidemiology
- Regression Decision Trees and Classifications

Keywords

Regression decision trees · Multi-dimensional analysis · Regression decision trees · Multi-dimensional analysis

R

This chapter introduces the extensive use of the computer program R as applied to calculations in human genetic epidemiology[5]. R (R Development Core Team, 2018) is an open-sourced, viz., free, computer program primarily developed for statistics and graphics, originally by Robert Gentleman and Ross Ihaka of the Department of Statistics in the University of Auckland, New Zealand in 1997! Since then, there exists an R development core group of about 20 people who have write-access to the R source code. The original R environment, having evolved from the S/S-PLUS languages, was *not* primarily designed for statistics and biostatistics. Nevertheless, ever since its development in the 1990s, it has become the tool of choice of many practitioners in modern statistical methodologies, especially among those working in the field of biostatistics as applied to preventive.

A number of these R packages have been developed for a wide range of human genetics, including classical areas such as phylogenetics, quantitative trait linkage (QTL) mapping, haplotype mapping, linkage disequilibrium, next generation sequencing, genome-wide association studies (GWAS), etc.

Other resources include the following:**1. *Bioconductor***

Another resource for the analysis of genomic and genetic data, using R, is the collection ***Bioconductor***, which includes many hundreds of R packages for high-throughput genomic data.

2. *List of Genetic Analysis Software*

This list was created by **Dr. Wentian Li**, when he was at Columbia University (1995-1996). It was later moved to Rockefeller University (1996-2002) and is now being maintained as a [github repository](http://lab.rockefeller.edu/ott/links) by Dr. Gao Wang: <http://lab.rockefeller.edu/ott/links>

3. *European Genome-phenome Archive, EGA:*

<https://www.ebi.ac.uk/ega/home>

The EGA **archives** a large number of datasets, the access to which is controlled by a Data Access Committee (DAC).

4. *The European Bioinformatics Institute <EMBL-EBI*

<https://www.ebi.ac.uk/>

The European Bioinformatics Institute (EMBL-EBI) shares data from life science experiments, performs basic research in computational biology and offers an extensive user training program, supporting researchers in academia and industry. We are part of EMBL, **Europe's** flagship laboratory for the life sciences.

5. *EGA European Genome-Phenome Archive*

<https://ega.crg.eu/>

The **European Genome-phenome Archive** (EGA) is a service for permanent archiving and sharing of all types of personally identifiable genetic and phenotypic data resulting from biomedical research projects. The EGA contains exclusive data collected from individuals whose consent agreements authorize data release

6. *European Bioinformatics Institute (EMBL-EBI)*

<https://www.facebook.com/EMBLEBI/>

The UK Biobank delivers its genetic data via the European Genome Phenome Archive and the Center for Genomic Regulation: <https://www.ebi.ac.uk/>

7. *The European Nucleotide Archive (ENA)*

<http://www.ebi.ac.uk/ena>

Europe's primary nucleotide sequence resource, captures and presents globally comprehensive nucleic acid sequence and associated information; covering the spectrum from raw data to assembled and functionally annotated genomes. The ENA is a repository for the world public domain nucleotide sequence data output. ENA content covers a spectrum of data types including raw reads, assembly data and functional annotation. ENA has faced a dramatic growth in **genome** assembly. Other subsidiaries include:

The European Nucleotide Archive - PubMed Central Canada

The European Genome-phenomeArchive (EGA)

8. *NHGRI-Related News Archive - National Human Genome*

<https://www.genome.gov/19016944/NHGRIRelated-News-Archive>

9. *The European Bioinformatics Institute (EMBL-EBI):*

The New York University School of Medicine and

The Ontario Institute for Cancer Research (OICR)

The University of California Santa Cruz: Investigators at the UC Santa

The Cruz Genomics Institute have optimized performance of a mobile-phone-sized MinIONTM

10. Wellcome Library | Collecting Genomics

<https://wellcomelibrary.org/what-we-do/.../collecting-genomics/>

The Wellcome Library is safeguarding the history of modern genomics by collecting the records of scientists and organizations involved in pioneering genomics research, in preserving their own archives. It is working with the European Molecular Biology Laboratory.



5.1 Biostatistical Human Genetics

Biostatistics is a field of science that uses mathematics and computing science to investigate a number of life science related problems, including those pertaining to genetics, medicine, biology, agriculture, etc.

Source: The American Society of **Human Genetics**;

URL: http://www.ashg.org/education/gena/NatureNurture_L2_corrected.pdf ...

Statistical Significance (*T*-Test, or *t*-Test)

Definition

A *T*-Test statistical significance indicates whether or not the difference, between the averages of two groups, is most likely reflecting a “real” difference in the population from which the groups were sampled. A *T*-Test is an analysis of two population means by the use of statistical examination. A *T*-Test with two samples is commonly used with small sample sizes, testing the difference between the samples when the *variances* of those two *distributions* are unknown.

A *T*-Test looks at the *t*-statistic, the *t*-distribution, and the *degrees of freedom* to determine the probability of difference between populations. The test statistic in the test is known as the *T*-statistic.

To conduct a test with three or more variables, an [Analysis of Variance \(ANOVA\)](#) should be used.

The 'T-Test'

This is a form of hypothesis testing, the *T*-Test is one of many tests used for this purpose. Statisticians must use tests other than the *T*-Test to examine more variables, as well as for test with larger sample sizes. For a large sample size, statisticians use a *z*-test. Other testing options include the chi-square test and the *f*-test.

Statistical Analysis of the *T*-Test

The formula used to calculate the test is a ratio: The top portion of the ratio is the easiest portion to calculate and understand, as it is simply the difference between the means or averages of the two samples. The lower half of the ratio is a measurement of the dispersion, or variability, of the scores. The bottom part of this ratio is known as the **standard error** of the difference. To compute this part of the ratio, the variance for each sample is determined and is then divided by the number of individuals the compose the sample, or group. These two values are then added together, and a square root is taken of the result.

An Example

For example, consider that an analyst wanting to study the amount that Pennsylvanians and Californians spend, per month, on clothing. It would not be practical to record the spending habits of every individual (or family) in both states, thus a *sample* of spending habits is taken from a selected group of individuals from each state. The group may be of any small to moderate size — for this example, assume that the sample group is 200 individuals.

The average amount for Pennsylvanians comes out to \$500; the average amount for Californians is \$1,000. The *T-Test* questions whether the difference between the groups is representative of a true difference between people in Pennsylvania and people in California in general or if it is likely a meaningless statistical difference. In this example, if, theoretically, all Pennsylvanians spent \$500 per month on clothing and all Californians spent \$1,000 per month on clothing, it is highly unlikely that 200 randomly selected individuals all spent that exact amount, respective to state. Thus, if an analyst or statistician yielded the results listed in the example above, it is safe to conclude that the difference between sample groups is indicative of a significant difference between the populations, as a whole, of each state.

Sampling Distribution

In statistics, a **sampling distribution** or **finite-sample distribution** is the **probability distribution** of a given statistic based on a **random sample**. Sampling distributions are important in statistics because they provide a major simplification to obtain a measure of **statistical inference**. Specifically, they allow analytical considerations to be based on the sampling distribution of a **statistic**, rather than on the **joint probability distribution** of all the individual sample values.

The **sampling distribution** of a statistic is the **distribution** of that statistic, considered as a **random variable**, when derived from a **random sample** of a large size. It may be considered as the distribution of the statistic for *all possible samples from the same population* of a given sample size. The sampling distribution depends on the underlying **distribution** of the population, the statistic being considered, the sampling procedure employed, and the sample size used. There is often considerable interest in whether the sampling distribution may be approximated by an **asymptotic distribution**, which corresponds to the limiting case either as the number of random samples of finite size, taken from an infinite population and used to produce the distribution, tends to infinity, or when just one equally-infinite-size "sample" is taken of that same population.

For example, consider a **normal** population with mean μ and variance σ^2 . Assume one repeatedly take samples of a *given* size from this population and then calculate the **arithmetic mean** \bar{x} for each sample — this statistic is called the **sample mean**. The distribution of these means, or averages, is called the "sampling distribution of the sample mean". This distribution is normal (n is the sample size) since the underlying population is normal, although sampling distributions may also often be close to normal even when the population distribution is not (cf. Central Limit Theorem). An alternative to the sample mean is the sample **median**. When calculated from the same population, it has a different sampling distribution to that of the mean and is generally not normal (but it may be close for *large* sample sizes).

The mean of a sample from a population having a normal distribution is an example of a simple statistic taken from one of the simplest **statistical populations**. For other statistics and other populations the formulas are more complicated, and often they do not exist in **closed-form**. In such cases the sampling distributions may be approximated through **Monte-Carlo simulations**, **bootstrap** methods, or **asymptotic distribution** theory.

Probability Distribution

In probability theory and statistics, a **probability distribution** is a mathematical function that may be thought of as providing the probabilities of occurrence of different possible outcomes in an experiment. The probability distribution of a statistic is obtained through a large number of samples drawn from a given population.

Sampling Distribution

A lot of data drawn and used by academicians, statisticians, researchers, marketers, analysts, etc. are actually samples, *not* population. A **sample** is a subset of a **population**. For example, a medical researcher that wanted to compare the average weight of all babies born in North America from 1995 to 2005 to those born in South America within the same time period cannot, within a reasonable and limited amount of time and resources, draw the data for the entire population of over a million childbirths that occurred over the ten-year time frame. He will instead only use the weight of, say 100 babies, in each continent to make a conclusion. The weight of 200 babies used is the sample and the average weight calculated is the sample mean.

Example 5.1

Figure 5.1 graphically illustrate the “Dollars Spent on Movies Per Month” for consumers living two states of the United States of America: namely in the states of New York and Kansas:

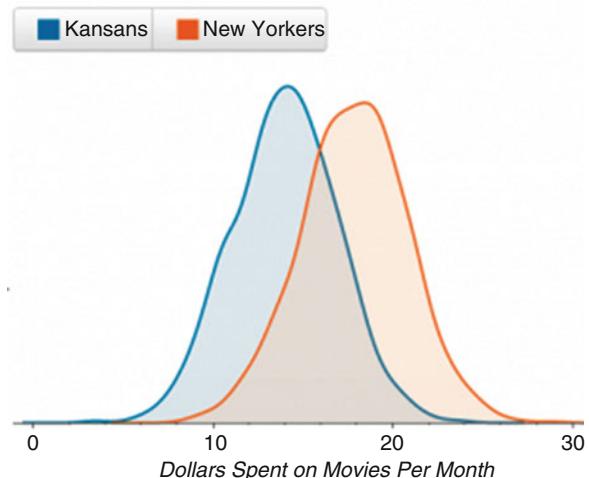
Assuming there is a large enough sample size, the difference between these two groups is probably statistically significant. Here, one may probably assume statistical significance for the unranked “Independent Samples *T*-Test”, the most common form of *T*-Test.

For this example:

Assume that one is interested in whether the average New Yorker spends more than the average Kansan per month on movies.

One may ask a sample of 4 people from each state about their movie spending. One might observe a difference in those averages (like \$15 for the average Kansan and \$19 for the average New Yorker). That difference is not *statistically significant*; it could easily just be random luck of which 4 people one randomly sampled that makes one group appear to spend more money than the other. If instead one

Fig. 5.1 Distributions for moneys spent monthly by consumers of 2 States of the U.S.A. – in New York and in Kansas



asks 1,000 New Yorkers and 1,000 Kansans and still see a big difference, that difference is *less* likely to be caused by the sample being unrepresentative.

If one asked 1,000,000 New Yorkers and 1,000,000 Kansans, the result would likely be statistically significant even if the difference between the group was only a penny! The [T-Test's effect size](#) complements its statistical significance, describing the magnitude of the difference, whether or not the difference is statistically significant.

Definition

A *statistically significant* t-test result is one in which a difference between two groups is unlikely to have occurred because the sample happened to be atypical. Statistical significance is determined by the size of the difference between the group averages, the sample size, and the standard deviations of the groups. For practical purposes statistical significance suggests that the two larger populations from which one samples are “actually” different.

5.1.1 Some Preliminary Remarks on the T-Test in Statistics

A Definition of the Two-sample T-Test^[1]

The T-test is a statistical test of the Null Hypothesis H_0 that the means μ_1 and μ_2 of two populations, 1 and 2, are the same. It is usually written as:

$$H_0 : \mu_1 = \mu_2 \quad (5.1)$$

Thus, for example, one may define μ_1 as the population mean for case subjects with the aa genotype, and μ_2 as the population mean for case subjects with the AA or Aa genotypes. Then the two-sample T-test statistic, for equal variances, is

$$t = \left(\underline{y}_1 - \underline{y}_2 \right) / \sqrt{[s_p^2 \{ (1/n_1) + (1/n_2) \}]} \quad (5.2)$$

where \underline{y}_1 and \underline{y}_2 are the sample means of the quantitative trait for genotype

groups 1 and 2,

s_p^2 is the pooled estimate of the variance, and

n_1 and n_2 are the sample sizes, respectively.

and, under the Null Hypothesis, this statistic has a T-distribution with $(n_1 + n_2 - 2)$ degrees of freedom.

Remarks:

1. A non-parametric analog to the foregoing Two-sample T-Test is the **Wilcoxon Rank-Sum Test** (also known as the **Mann-Whitney U-Test**) and is more suitable than the T-test if the trait is not normally distributed and if the sample is small.
2. The Wilcoxon rank-sum test is a rank-based test, and may be used for testing the null hypothesis that the medians of a quantitative trait in each of two populations are equal.

3. Both the *T*-test and the Wilcoxon Rank-Sum Test had been implemented in R using the R functions: `t.test()` and `wilcox.test()` respectively, as shown in the following worked example:

Worked Examples of the *T*-Test and the Wilcoxon Rank-Sum Test for Calculating the Suitable Sample Sizes for Statistical Testing, using R:

Example 5.2

From the CRAN website:

Package ‘samplesize’

December 24, 2016

Type	Package
Title	Sample Size Calculation for Various t-Tests and Wilcoxon-Test
Version	0.2-4
Date	2016-12-22
Author	Ralph Scherer
Maintainer	Ralph Scherer < shearer.ra76@gmail.com >
Description	Computes sample size for Student's t-test and for the Wilcoxon-Mann-Whitney test for categorical data. The t-test function allows paired and unpaired (balanced / unbalanced) designs as well as homogeneous and heterogeneous variances. The Wilcoxon function allows for ties.
License	GPL (>= 2)
URL	https://github.com/shearer/samplesize
BugReports	https://github.com/shearer/samplesize/issues
NeedsCompilation	no
Repository	CRAN
Date/Publication	2016-12-24 11:24:04

Example 4A

`n.ttest` `n.ttest` computes sample size for paired and unpaired t-tests.

Description

`n.ttest` computes sample size for paired and unpaired t-tests. Design may be balanced or unbalanced. Homogeneous and heterogeneous variances are allowed.

Usage

```
n.ttest (power = 0.8, alpha = 0.05, mean.diff = 0.8, sd1 = 0.83,
         sd2 = sd1, k = 1, design = "unpaired",
         fraction = "balanced", variance = "equal")
```

Arguments

Power	Power (1 - Type-II-error)
alpha	Two-sided Type-I-error
mean.diff	Expected mean difference
sd1	Standard deviation in group 1
sd2	Standard deviation in group 2

k	Sample fraction k
design	Type of design. May be paired or unpaired
fraction	Type of fraction. May be balanced or unbalanced
variance	Type of variance. May be homo- or heterogeneous

Value

Total sample size	Sample size for both groups together
Sample size group 1	Sample size in group 1
Sample size group 2	Sample size in group 2

Author Ralph Scherer

References Bock J., Bestimmung des Stichprobenumfangs fuer biologische Experimente und kontrollierte klinische Studien. Oldenbourg 1998

Examples

```
n.ttest(power = 0.8, alpha = 0.05, mean.diff = 0.80, sd1 = 0.83,
       k = 1, design = "unpaired", fraction = "balanced",
       variance = "equal")

n.ttest(power = 0.8, alpha = 0.05, mean.diff = 0.80, sd1 = 0.83,
       sd2 = 2.65, k = 0.7, design = "unpaired",
       fraction = "unbalanced", variance = "unequal")
```

Example 4B

n.wilcox.ord Sample size for Wilcoxon-Mann-Whitney for ordinal data

Description

Function computes sample size for the two-sided Wilcoxon test when applied to two independent samples with ordered categorical responses.

Usage

```
n.wilcox.ord(power = 0.8, alpha = 0.05, t, p, q)
```

Arguments

power	required Power
alpha	required two-sided Type-I-error level
t	sample size fraction n/N, where n is sample size of group B and N is the total sample size
p	vector of expected proportions of the categories in group A, should sum to 1,
q	vector of expected proportions of the categories in group B, should be of equal length as p and should sum to 1

Details

This function approximates the total sample size, N , needed for the two-sided Wilcoxon test when comparing two independent samples, A and B, when data are ordered categorical according to Equation 12 in Zhao et al. (2008). Assuming that the response consists of D ordered categories $C1; \dots; CD$. The expected proportions of these categories in two treatments A and B must be specified as numeric vectors $p1; \dots; pD$ and $q1; \dots; qD$, respectively. The argument t allows to compute power for an unbalanced design, where $t = n_B/N$ is the proportion of sample size in treatment B.

Value

total sample size Total sample size
 m Sample size group 1
 n Sample size group 2

Author Ralph Scherer

References

Zhao YD, Rahardja D, Qu Yongming (2008).- “Sample size calculation for the Wilcoxon-Mann-Whitney test adjusting for ties”. **Statistics in Medicine** 27:462-468

Examples

```
## example out of:  

## Zhao YD, Rahardja D, Qu Yongming.  

## Sample size calculation for the Wilcoxon-Mann-Whitney test ## adjusting for ties.  

## Statistics in Medicine (2008) 27:462-468
```

```
n.wilcox.ord(power = 0.8, alpha = 0.05, t = 0.53,  

             p = c(0.66, 0.15, 0.19), q = c(0.61, 0.23, 0.16))
```

In the R domain:

```
>  

> install.packages("samplesize")  

Installing package into 'C:/Users/Bert/Documents/R/win-library/3.3'  

(as 'lib' is unspecified)  

trying URL 'https://mirrors.tuna.tsinghua.edu.cn/CRAN/bin/windows/contrib/3.3/  

samplesize_0.2-4.zip'  

Content type 'application/zip' length 18449 bytes (18 KB)  

downloaded 18 KB  

package 'samplesize' successfully unpacked and MD5 sums checked
```

The downloaded binary packages are in

```
C:\Users\Bert\AppData\Local\Temp\RtmpmMmtf0\downloaded_packages  

> library(samplesize)  

> ls("package:samplesize")  

[1] "n.ttest"      "n.wilcox.ord"  

>
```

```
> # Example A:
>
> n.ttest  # Outputting:
function (power = 0.8, alpha = 0.05, mean.diff = 0.8, sd1 = 0.83,
         sd2 = sd1, k = 1, design = "unpaired", fraction = "balanced",
         variance = "equal")
{
  if (variance == "equal" & sd1 != sd2) {
    warning("Variance is set to equal, but sd's are different. This makes no
sense!")
  }
  if (fraction == "unbalanced" & k == 1) {
    warning("Groups are chosen unbalanced, but fraction argument k is 1")
  }
  if (design == "paired" & fraction == "unbalanced") {
    warning("Argument -unbalanced- is not used. Paired design is balanced")
  }
  if (design == "paired" & k != 1) {
    warning("Argument -k- is set to 1. Paired design is balanced")
  }
  if (design == "paired" & variance == "unequal") {
    warning("Paired design assumes and uses equal variances")
  }
  if (design == "paired") {
    fraction = "balanced"
  }
  if (design == "paired") {
    variance = "equal"
  }
  if (design == "unpaired" & variance == "unequal") {
    warning("Arguments -fraction- and -k- are not used, when variances are
unequal")
  }
  if (power > 1 | power < 0) {
    stop("Power must be between 0 and 1.0")
  }
  if (power < 0.5) {
    warning("Are you sure that Power should be lower than 50 % ?")
  }
  if (alpha > 1 | alpha < 0) {
    stop("Type-I-error must be between 0 and 1.0")
  }
  if (alpha > 0.1) {
    warning("Are you sure that the two-sided Type-I-Error should be larger
than 10 % ?")
  }
  if (k < 0) {
    stop("Fraction k must be greater than zero")
  }
  conf.level <- 1 - alpha/2
```

```
n.start <- 4
switch(variance, unequal = {
  k <- sd2/sd1
  n1.pri <- n.start/(1 + k)
  n2.pri <- (k * n.start)/(1 + k)
  n1 <- max(n1.pri, 2)
  n2 <- max(n2.pri, 2)
  gamma <- sd1/(sd1 + sd2)
  c <- mean.diff/(sd1 + sd2)
  df_approx <- 1/((gamma)^2/(n1 - 1) + (1 - gamma)^2/(n2 - 1))
  tkrit.alpha <- qt(conf.level, df = df_approx)
  tkrit.beta <- qt(power, df = df_approx)
  n.temp <- ((tkrit.alpha + tkrit.beta)^2)/(c^2)
  while (n.start <= n.temp) {
    n.start <- n1 + n2 + 1
    n1 <- n.start/(1 + k)
    n2 <- (k * n.start)/(1 + k)
    df_approx <- 1/((gamma)^2/(n1 - 1) + (1 - gamma)^2/(n2 - 1))
    tkrit.alpha <- qt(conf.level, df = df_approx)
    tkrit.beta <- qt(power, df = df_approx)
    n.temp <- ((tkrit.alpha + tkrit.beta)^2)/(c^2)
  }
  output <- list('Total sample size' = ceiling(n1) + ceiling(k *
    n1), 'Sample size group 1' = ceiling(n1), 'sample size group 2' =
ceiling(n2))
  return(output)
}, equal = {
{
  switch(design, paired = {
    n.start <- 2
    c <- mean.diff/sd1
    tkrit.alpha <- qt(conf.level, df = n.start -
      1)
    tkrit.beta <- qt(power, df = n.start - 1)
    n.temp <- ((tkrit.alpha + tkrit.beta)^2)/(c^2)
    while (n.start <= n.temp) {
      n.start <- n.start + 1
      tkrit.alpha <- qt(conf.level, df = n.start -
        1)
      tkrit.beta <- qt(power, df = n.start - 1)
      n.temp <- ((tkrit.alpha + tkrit.beta)^2)/(c^2)
    }
    output <- list('Total sample size' = n.start)
    return(output)
}, unpaired = {
```

```

switch(fraction, balanced = {
  c <- mean.diff/(2 * sd1)
  tkrit.alpha <- qt(conf.level, df = n.start -
    1)
  tkrit.beta <- qt(power, df = n.start - 1)
  n.temp <- ((tkrit.alpha + tkrit.beta)^2)/(c^2)
  while (n.start <= n.temp) {
    n.start <- n.start + 1
    tkrit.alpha <- qt(conf.level, df = n.start -
      1)
    tkrit.beta <- qt(power, df = n.start - 1)
    n.temp <- ((tkrit.alpha + tkrit.beta)^2)/(c^2)
  }
  n1 <- ceiling(n.start/2)
  n2 <- ceiling(n.start/2)
  output <- list('Total sample size' = 2 * n1,
    'Sample size group 1' = n1, 'Sample size group
    2' = n2)
  return(output)
}, unbalanced = {
  df <- n.start - 2
  c <- (mean.diff/sd1) * (sqrt(k)/(1 + k))
  tkrit.alpha <- qt(conf.level, df = df)
  tkrit.beta <- qt(power, df = df)
  n.temp <- ((tkrit.alpha + tkrit.beta)^2)/(c^2)
  while (n.start <= n.temp) {
    n.start <- n.start + 1
    tkrit.alpha <- qt(conf.level, df = n.start -
      2)
    tkrit.beta <- qt(power, df = n.start - 2)
    n.temp <- ((tkrit.alpha + tkrit.beta)^2)/(c^2)
  }
  n1 <- n.start/(1 + k)
  n2 <- k * n1
  output <- list('Total sample size' = ceiling(n1) +
    ceiling(n2), 'Sample size group 1' = ceiling(n1),
    'Sample size group 2' = ceiling(n2), Fraction = k)
  return(output)
})
})
}
return(output)
})
}
<environment: namespace:samplesize>
>
>

```

```
> # Example B:
>
> n.wilcox.ord # Outputting:
function (power = 0.8, alpha = 0.05, t, p, q)
{
  t = t
  q = q
  p = p
  if (power <= 0 | power >= 1) {
    stop("Power must be a numeric value between 0 and 1")
  }
  if (alpha <= 0 | alpha >= 1) {
    stop("alpha must be a numeric value between 0 and 1")
  }
  if (t <= 0 | t >= 1) {
    stop("t must be a numeric value between 0 and 1")
  }
  np <- length(p)
  nq <- length(q)
  if (np != nq) {
    stop("p and q must be vectors of equal length")
  }
  if (np < 2) {
    stop("p and q must be vectors of with at least two elements")
  }
  sp <- sum(p)
  sq <- sum(q)
  if (abs(sp - 1) > .Machine$double.eps * 10) {
    p <- p/sp
    warning("The elements in p did not sum up to 1 and have been rescaled")
  }
  if (abs(sq - 1) > .Machine$double.eps * 10) {
    q <- q/sq
    warning("The elements in q did not sum up to 1 and have been rescaled")
  }
  alpha_half = alpha/2
  Z1 <- qnorm(alpha_half)
  Z2 <- qnorm(1 - power)
  pq1 <- function(p, q) {
    D <- length(p)
    PQ1 <- 0
    for (i in 2:D) {
      PQ1 <- PQ1 + p[i] * sum(q[1:(i - 1)])
    }
    return(PQ1)
  }
  p.t <- (1 - t) * p
  q.t <- t * q
  pq.t <- p.t + q.t
  pq.t.3 <- pq.t^3
```

```

t.sum <- sum(pq.t.3)
pq <- cbind(p, q)
pq.sum <- sum(apply(pq, 1, prod))
N <- (((Z1 + Z2)^2) * (1 - t.sum))/(12 * t * (1 - t) * (pq1(p = p,
q = q) + 0.5 * pq.sum - 0.5)^2)
samplesize <- ceiling(N)
m <- round(ceiling(N) * (1 - t), 0)
n <- round(ceiling(N) * t, 0)
return(list('total sample size' = samplesize, m = m, n = n))
}
<environment: namespace:samplesize>
>
>
> n.ttest(power = 0.8, alpha = 0.05, mean.diff = 0.80,
+           sd1 = 0.83, k = 1,
+           design = "unpaired", fraction = "balanced", variance =
+           "equal")
>
> # Outputting:
$'Total sample size'
[1] 36

$'Sample size group 1'
[1] 18

$'Sample size group 2'
[1] 18

>
>
> n.ttest(power = 0.8, alpha = 0.05, mean.diff = 0.80,
+           sd1 = 0.83, sd2 = 2.65, k = 0.7,
+           design = "unpaired", fraction = "unbalanced",
+           variance = "unequal")
>
> # Outputting:
$'Total sample size'
[1] 153

$'Sample size group 1'
[1] 37

$'sample size group 2'
[1] 116

Warning message:
In n.ttest(power = 0.8, alpha = 0.05, mean.diff = 0.8, sd1 = 0.83, :
  Arguments -fraction- and -k- are not used, when variances are unequal
>

```

5.2 Human Genetic Data Concepts

Human genetic association occurs when one or more [genotypes](#) within a population co-occur with a [phenotypic](#) trait [more often than would be expected by chance occurrence](#). Studies of genetic association consist of testing whether single-locus alleles or genotype frequencies, or more generally, multi-locus [haplotype](#) frequencies, differ between two groups of individuals - usually diseased subjects and healthy controls. Genetic association studies are based upon the principle that genotypes can be compared "directly", i.e. with the sequences of the actual [genomes](#) or exomes via whole genome sequencing or whole exome sequencing. Before the 2010s, earlier DNA sequencing methods were used. Genetic association may occur between phenotypes, such as visible characteristics such as flower color or height, between a phenotype and a genetic polymorphism, such as a [single nucleotide polymorphism](#) (SNP), or between two genetic polymorphisms. Association between genetic polymorphisms occurs when there is non-random association of their alleles as a result of their proximity on the same chromosome; this is known as [genetic linkage](#).

[Linkage disequilibrium](#) (LD) is the study of population genetics for the non-random association of alleles at two or more loci, not necessarily on the same chromosome. It is not the same as linkage, which is the phenomenon whereby two or more loci on a chromosome have reduced recombination between them because of their physical proximity to each other. LD describes a situation in which some combinations of alleles or genetic markers occur more or less frequently in a population than would be expected from a random formation of haplotypes from alleles based on their frequencies.

Genetic association studies are undertaken to determine whether a genetic variant is associated with a disease or trait: if association is present, a particular allele, genotype or haplotype of a polymorphism or polymorphisms will be seen more often than expected by chance in an individual carrying the trait. For example, a person carrying one or two copies of a high-risk variant is at increased risk of developing the associated disease or having the associated trait.

5.2.1 The Study of Human Genetic Variation

Classical types of human genetic association studies include:

1. Case-Control Designs

Case control studies are a classical epidemiological tool. Case-control studies use case-subjects who already have a disease, trait or other condition and determine if there are characteristics of these patients that differ from those who do not have the disease or trait.

In genetic case-control studies, the frequency of alleles or genotypes is compared between the cases and controls. The cases will have been diagnosed with the disease under study, or have the trait under test; the controls, who are either known to be unaffected, or who have been randomly selected from the population.

A difference in the frequency of an allele or genotype of the polymorphism under test between the two groups indicates that the genetic marker may increase risk of the disease or likelihood of the trait, or be in [linkage disequilibrium](#) with a polymorphism which does. Haplotypes can also show association with a disease or trait.

It should not escape ones attention that the case-control design is that genotype and haplotype frequencies vary between ethnic or geographic populations. If the case and control populations are not

well matched for ethnicity or geographic origin then false positive association may occur because of the confounding effects of population stratification.

2. Family Based Designs

Family based association designs aim to avoid the potential confounding effects of population stratification by using the parents or using unaffected siblings as controls for the case, which is their affected offspring/ siblings. The most commonly used test is the Transmission Disequilibrium Test, or TDT. Two similar tests are used: the **Transmission Disequilibrium Test (TDT)** and the **Haplod-Relative-Risk (HRR)**. Both measure association of genetic markers in nuclear families by transmission from parent to offspring. If an allele increases the risk of having a disease then that allele is expected to be transmitted from parent to offspring more often in populations with the disease.

3. Quantitative Trait Association

A quantitative trait (cf. **quantitative trait locus**) is a measurable trait that shows continuous variation, such as height or weight. Quantitative traits often have a 'normal' distribution in the population. In addition to the case control design, quantitative trait association can also be performed using an unrelated population sample or family trios in which the quantitative trait is measured in the offspring.

Statistical Programs for Association Analysis

There are many computer packages for analyzing genetic association, such as:

- **Package** nplplot Plotting statistics along a chromosome
- **Package** SNPassoc Whole genome association studies.
- **Package** ParseCNV, and Golden Helix's **SNP & Variation Suite**.

However simple genotypic or allelic association with a dichotomous trait can be measured using the chi-squared test for significance.

Example 5

nplplot Plotting statistics along a chromosome

Package nplplot

February 20, 2015

Version 4.5

Date 2014-04-30

Title Plotting linkage and association results

Author Robert V Baron <rvb5@pitt.edu>, Nandita Mukhopadhyay, Xinyu Tang, Daniel E. Weeks <weeks@pitt.edu>

Maintainer Daniel E. Weeks <weeks@pitt.edu>

Description

This package provides routines for plotting linkage and association results along a chromosome, with marker names displayed along the top border.

There are also routines for generating BED and BedGraph custom tracks for viewing in the UCSC genome browser.

The data reformatting program Mega2 uses this package to plot output from a variety of programs.

License	GPL (>= 3)
URL	http://watson.hgen.pitt.edu/register
Repository	CRAN
Date/Publication	2014-05-01 09:32:57
Needs Compilation	no
nplplot	Plotting Statistics Along a Chromosome

Description

Plots linkage or association statistics along a chromosome, contained within a data frame or a file.

Marker names are displayed along the top border.

Usage

```
nplplot(plotdata=NULL, filename=NULL, yline=2.0, ymin=0,
        ymax=3.0,
        header=TRUE, yfix=FALSE, title=NULL, draw.lgnd=TRUE,
        xlabel="", ylabel="", lgndx=NULL, lgndy=NULL,
        lgndtxt=NULL, cex.legend = 0.7, cex.axis=0.7, tcl=1,
        bw=TRUE, my.colors=NULL, ltypes=NULL, ptypes=NULL,
        na.rm=TRUE, plot.width=0.0, ...)
```

Arguments

plotdata	A data frame containing marker names in the first column, marker map positions in the second column, and statistical scores in column 3 onwards.
filename	A table format file containing the plot data as described above.
header	TRUE or FALSE depending on whether the plotdata or file has a header line.
yline	Y-value for displaying a horizontal cut-off line. If 'yfix' is set to TRUE and Y-line falls outside of [ymin, ymax], then the cut-off line is omitted.
ymin, ymax	Y-axis minimum and maximum values. If non-NULL values are provided, and yfix is set to TRUE, then the plot area will be cropped to these values. If yfix is set to FALSE, then ymin and ymax values are ignored.
yfix	TRUE or FALSE to denote whether plot area should be cropped to the ymin, ymax values. This has no effect if ymin, ymax values are NULL.
title	Used as the subtitle of the plot.
xlabl	X-axis label. May interfere with the display of the subtitle provided as the title argument.
ylabl	Y-axis label.
draw.lgnd	TRUE or FALSE denoting whether a plot legend should be displayed.

lgndx	X coordinate for the legend box, passed to the legend command. Ignored if draw.legend is set to FALSE. If set to NULL with draw.legend set to TRUE, the X-coordinate is automatically calculated.
lgndy	Y coordinate for the legend box, passed to the legend command. Ignored if draw.legend is set to FALSE. If set to NULL with draw.legend set to TRUE, the Y-coordinate is automatically calculated.
lgndtxt	Vector of strings to use in the legend.
cex.legend	Character scaling for legend, passed as the cex argument to the legend command.
cex.axis	Character scaling for the axis, passed to the axis command for drawing the top border.
tcl	Length of ticks for the top border, passed to the axis command.
bw	TRUE or FALSE depending on whether plots should be drawn in color. If set to FALSE, then the colors defined by my.colors are used.
my.colors	Vector of color specifications as described in the par command. Ignored if bw above is set to FALSE. If bw is set to TRUE and my.colors is set to NULL, the rainbow palette will be used.
ltypes	Vector of line types for the plots. Each non-zero line type is passed on to a lines command. Use 0 or 'none' if a line is to be skipped. If NULL, no lines will be drawn. For line types see the par command. If set to "default", line-types 1 through the number of plots is used.
ptypes	Vector of characters giving the point types, to be passed onto the points command. Use 'none' if no points are to be drawn for a score column. If NULL, then no points will be displayed. If both the line-type and point-type specification for a results column is set to 'none', then that column will not be plotted.
na.rm	TRUE or FALSE depending on whether points with Y-coordinates set to NAs should be skipped. Setting na.rm to TRUE eliminates discontinuities in the plots.
plot.width	A number giving the width of the plot in inches. This is used to decide whether some marker names should be dynamically hidden, if they are too close to each other along the top border. If set to 0, the default page-size is used to set the width.
...	Further graphical parameters to be passed onto the 'plot', 'lines', and 'points' commands.

Details

The nplplot() function draws multiple curves within a single plot by automatically calling 'plot', 'lines', and 'points' multiple times, thus making it easy for the user to plot many columns of result using a single plot command. It is intended for the display of linkage and association analysis results such as LOD scores and P-values. It allows the marker names to be displayed along the top border of the plot, as well as a significance threshold line. The input plot data has to be in a specific tabular format with each column separated by white-space:

Here is an example:

Marker	Position	score1	score2	score3
d1s228	0.00	0.546	0.345	0.142
d1s429	1.00	0.346	0.335	0.252
d1s347	2.00	0.446	0.245	0.342

This example file contains a header, therefore the header argument should be set to TRUE.

Lines 2-4 contain scores at various marker positions. Missing scores can be denoted with either "." or "NA". The position column cannot have missing data. There can be any number of score columns within a file and will be plotted as separate curves within the same plot. Each file is plotted as a separate plot.

Value

TRUE or FALSE depending on whether the plot data was successfully plotted.

See Also nplplot.multi, nplplot.old

Examples

```
# plot with legend
par(omi=c(0.05, 0.05, 0.5, 0.05))
data(lods1, package="nplplot")
nplplot(plotdata=lods1, draw.lgnd=TRUE)

# plot without legend
data(lods2, package="nplplot")
nplplot(plotdata=lods2, draw.lgnd=FALSE)

# plotting from a data file
datadir <- paste(system.file("data", package="nplplot"), .Platform$file.sep,
sep="")
nplplot(filename=paste(datadir, "lods2.txt.gz", sep=""))
nplplot.
```

In the R domain:

```
>
> install.packages("nplplot")
Installing package into 'C:/Users/Bert/Documents/R/win-library/3.3'
(as 'lib' is unspecified)
--- Please select a CRAN mirror for use in this session ---
# A CRAN mirror is selected
trying URL 'https://mirrors.tuna.tsinghua.edu.cn/CRAN/bin/windows/contrib/3.3/
nplplot_4.5.zip'
Content type 'application/zip' length 61911 bytes (60 KB)
downloaded 60 KB

package 'nplplot' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
  C:/Users/Bert/AppData/Local/Temp/RtmpuqPF7s/downloaded_packages
>
> library(nplplot)
> ls("package:nplplot")
[1] "bedplot"        "genomeplot"      "nplplot"         "nplplot.multi"
```

```
[5] "nplplot.old"    "prepareplot"
> nplplot # Outputting:
function (plotdata = NULL, filename = NULL, yline = 2, ymin = 0,
         ymax = 3, header = TRUE, yfix = FALSE, title = NULL, draw.lgnd = TRUE,
         xlabl = "", ylabl = "", lgndx = NULL, lgndy = NULL, lgndtxt = NULL,
         cex.legend = 0.7, cex.axis = 0.7, tcl = 1, bw = TRUE, my.colors = NULL,
         ltypes = NULL, ptypes = NULL, na.rm = TRUE, plot.width = 0,
         ...)
{
  leg = NULL
  ltyps = NULL
  pchs = NULL
  if (!is.null(title)) {
    maint <- title
  }
  if (is.null(plotdata) && is.null(filename)) {
    print("Both plotdata and file are NULL, no data to plot!")
    return(F)
  }
  if (!is.null(plotdata) && !is.null(filename)) {
    print("Both plotdata and file are specified, using plotdata.")
  }
  if (!is.null(plotdata)) {
    lods <- plotdata
    print("lods read in as table")
    if (ncol(lods) <= 2) {
      print(paste("plotdata does not have any data."))
      return(F)
    }
  }
  else {
    if (!is.null(filename)) {
      lods <- read.table(filename, header = header, sep = "",
                          na.strings = c("NA", "."))
      if (ncol(lods) <= 2) {
        print(paste("File ", filename, "does not have any data."))
        return(F)
      }
    }
    rows <- dim(lods)[1]
    cols <- dim(lods)[2]
    dist <- lods[, 2]
    ydatamax <- max(lods[, 3:cols], na.rm = TRUE)
    ydatamin <- min(lods[, 3:cols], na.rm = TRUE)
    xdatamin <- min(dist, na.rm = TRUE)
    xdatamax <- max(dist, na.rm = TRUE)
    if (is.null(ymax)) {
      ymax1 <- ydatamax + 0.2
    }
  }
}
```

```
else {
  if (yfix == TRUE) {
    ymax1 <- ymax
  }
  else {
    ymax1 <- max(ydatamax, yline, ymax)
  }
}
if (is.null(ymin)) {
  ymin1 <- ydatamin - 0.1
}
else {
  if (yfix == TRUE) {
    ymin1 <- ymin
  }
  else {
    ymin1 <- min(ymin, ydatamin)
  }
}
if ((yfix == TRUE) && ((yline > ymax1) || (yline < ymin1))) {
  if (yline > ymax1) {
    print(paste("Y-line exceeds Y-max fixed at ",
               as.character(ymax1)))
  }
  else if (yline < ymin1) {
    print(paste("Y-line falls below Y-min fixed at ",
               as.character(ymin1)))
  }
  print("Y-line will not be drawn.")
  yline <- NA
}
if (is.null(ltypes) && is.null(ptypes)) {
  ltypes = 1:(cols - 2)
}
plot(c(xdatamin, xdatamax), c(ymin1, ymax1), type = "n",
      lty = 1, xlab = xlabel, ylab = ylabel, xlim = c(xdatamin,
      xdatamax), ylim = c(ymin1, ymax1), tcl = tcl, ...)
if (!is.na(yline)) {
  abline(h = yline, col = "grey40", ...)
}
if (!is.null(title)) {
  title(sub = maint, line = 2)
}
if (bw == FALSE) {
  if (is.null(my.colors)) {
    my.colors <- rainbow(cols - 2)
  }
}
for (k in 3:cols) {
  scores <- lods[, k]
```

```

ifelse(!is.null(ltypes), lt <- ltypes[k - 2], lt <- "none")
ifelse(!is.null(ptypes), ch <- ptypes[k - 2], ch <- "none")
this.x <- dist
this.y <- scores
if (bw == FALSE) {
  color <- my.colors[k - 2]
}
else {
  color <- "black"
}
if (na.rm) {
  this.x <- this.x[!is.na(this.y)]
  this.y <- this.y[!is.na(this.y)]
}
if (lt == 0) {
  lt = "none"
}
if (ch == 0) {
  ch = "none"
}
if (lt == "none" && ch != "none") {
  points(this.x, this.y, pch = ch, col = color, ...)
  pchs[k - 2] <- ch
  ltys[k - 2] <- "blank"
}
if (lt != "none" && ch == "none") {
  lines(this.x, this.y, type = "l", lty = lt, xaxt = "n",
        yaxt = "n", col = color, ...)
  pchs[k - 2] = " "
  ltys[k - 2] = lt
}
if (lt != "none" && ch != "none") {
  lines(this.x, this.y, type = "b", lty = lt, pch = ch,
        xaxt = "n", yaxt = "n", col = color, ...)
  pchs[k - 2] <- ch
  ltys[k - 2] <- lt
}
if (ch == "none" && lt == "none") {
  print("Skipping this line")
}
}
markers <- as.character(lods[, 1])
lbl <- markers[markers != "-"]
lloc <- dist[markers != "-"]
if (plot.width == 0) {
  plot.width <- (par()$pin)[1]
}
last.pos <- lloc[1]/(lloc[length(lloc)] - lloc[1]) * plot.width
for (j in 2:length(lloc)) {
  diff <- lloc[j]/(lloc[length(lloc)] - lloc[1]) * plot.width -

```

```
        last.pos
    if (diff < 0.15) {
        lbl[j] <- "      "
    }
    else {
        last.pos <- last.pos + diff
    }
}
axis(3, tck = 0.05, at = lloc, labels = lbl, cex.axis = cex.axis,
     las = 2)
if (draw.lgnd) {
    if (is.null(lgndtxt) || length(lgndtxt) < (cols - 2)) {
        leg <- names(lods)[3:cols]
    }
    else {
        leg <- lgndtxt
    }
    if (!is.null(lgndx)) {
        lgx <- lgndx
    }
    else {
        lgx <- xdatamin + 0.05 * (xdatamax - xdatamin)
    }
    if (!is.null(lgndy)) {
        lgy <- lgndy
    }
    else {
        lgy <- ymin1 + 0.9 * (ymax1 - ymin1)
    }
    if (!is.null(my.colors)) {
        leg.color <- my.colors
    }
    else {
        leg.color <- "black"
    }
    legend(lgx, lgy, leg, col = leg.color, lty = ltys, pch = pchs,
           cex = cex.legend)
}
return(TRUE)
}
<environment: namespace:nplplot>
> nplplot.multi
function (filenames, plotdata = NULL, col = 2, row = 2, mode = "l",
          output = "screen", headerfiles = NULL, lgnd = "page", customtracks = FALSE,
          mega2mapfile = NULL, pagewidth = NULL, pageheight = NULL,
          topmargin = 0.25, ...)
{
    if (!is.null(plotdata))
        numfiles = length(plotdata)
    else numfiles <- length(filenames)
```

```
if (is.null(pagewidth)) {
  pagewidth <- 8
}
if (is.null(pageheight)) {
  pageheight <- 10.5
}
if (mode == "p") {
  full.width <- pagewidth
  full.height <- pageheight
  landscape <- FALSE
}
else {
  full.width <- pageheight
  full.height <- pagewidth
  landscape <- TRUE
}
pl.wd <- full.width/col
if (output == "screen") {
  do.call(names(dev.cur()), list())
}
else {
  output1 <- output
  fileparts <- unlist(strsplit(output1, "\cr."))
  if (fileparts[length(fileparts)] == "ps") {
    print("Creating postscript file")
    postscript(width = full.width, height = full.height,
               file = output, paper = "letter", horizontal = landscape)
  }
  else {
    print("Creating pdf file")
    pdf(width = full.width, height = full.height, file = output,
         paper = "special")
  }
}
par(omi = c(0.05, 0.05, topmargin, 0.05))
par(mfrow = c(row, col), ask = FALSE)
retval <- TRUE
if (length(lgnd) > 1) {
  for (lg in lgnd) {
    if (!is.numeric(lg)) {
      print("Invalid list value for lgnd, use plot numbers.")
      print("No legends will be drawn.")
      lgnd <- "page"
      break
    }
  }
}
else {
  if (!is.logical(lgnd) && (lgnd != "page") && !is.numeric(lgnd)) {
    print("Invalid value for lgnd, use PAGE, TRUE/FALSE or plot numbers.")
  }
}
```

```
        print("Setting to PAGE")
        lgnd <- "page"
    }
}

if ((customtracks == TRUE) && !is.null(mega2mapfile)) {
    chrlist <- NULL
}

for (i in 1:numfiles) {
    title <- ""
    yline <- 2
    ymin <- 0
    ymax <- 4
    yfix <- FALSE
    ltypes <- NULL
    ptypes <- NULL
    my.colors <- NULL
    bw <- TRUE
    lgndx <- NULL
    lgndy <- NULL
    lgndtxt <- NULL
    cex.legend <- 0.7
    ylabel <- "Scores"
    tcl <- 0.3
    cex.axis <- 0.9
    draw.lgnd <- TRUE
    if (!is.null(headerfiles)) {
        if (i <= length(headerfiles)) {
            hdrfile <- headerfiles[i]
        }
        else {
            hdrfile <- headerfiles[length(headerfiles)]
        }
        lines <- readLines(con = hdrfile, n = -1)
        for (l in 1:length(lines)) {
            eval(parse(text = lines[l]))
        }
    }
    if (length(lgnd) == 1) {
        if (i == lgnd) {
            draw.lgnd <- TRUE
        }
        else if (lgnd == "page") {
            draw.lgnd <- ((i == 1) || (i%%(col * row)) ==
                1)
        }
        else if (is.logical(lgnd)) {
            draw.lgnd <- lgnd
        }
        else {
            draw.lgnd <- FALSE
        }
    }
}
```

```
        }
    }
  else {
    for (lg in lgnd) {
      if (is.numeric(lg) && i == lg) {
        draw.lgnd <- TRUE
        break
      }
      else {
        draw.lgnd <- FALSE
      }
    }
  }
  if (!is.null(plotdata)) {
    title = paste("Chromosome", names(plotdata[i]))
    retval1 <- nplplot(filename = NULL, plotdata = plotdata[[i]],
      header = TRUE, yline = yline, ymin = ymin, ymax = ymax,
      yfix = yfix, title = title, draw.lgnd = draw.lgnd,
      xlabl = "", ylabl = ylabel, lgndx = lgndx, lgndy = lgndy,
      lgndtxt = lgndtxt, cex.legend = cex.legend, bw = bw,
      my.colors = my.colors, ltypes = ltypes, ptypes = ptypes,
      tcl = tcl, cex.axis = cex.axis, plot.width = pl.wd,
      ...)
  }
  else {
    retval1 <- nplplot(filename = filenames[i], plotdata = NULL,
      header = TRUE, yline = yline, ymin = ymin, ymax = ymax,
      yfix = yfix, title = title, draw.lgnd = draw.lgnd,
      xlabl = "", ylabl = ylabel, lgndx = lgndx, lgndy = lgndy,
      lgndtxt = lgndtxt, cex.legend = cex.legend, bw = bw,
      my.colors = my.colors, ltypes = ltypes, ptypes = ptypes,
      tcl = tcl, cex.axis = cex.axis, plot.width = pl.wd,
      ...)
  }
  retval <- retval && retval1
  if ((customtracks == TRUE) && !is.null(mega2mapfile)) {
    if (is.null(plotdata)) {
      a <- unlist(strsplit(filenames[i], ".", fixed = TRUE))
      chrlist[i] <- switch(a[length(a)], X = "X", Y = "Y",
        XY = "XY", a[length(a)])
      if (i == 1) {
        prefix <- paste(a[1:length(a) - 1], sep = ".")
      }
    }
    else {
      chrlist[i] = names(plotdata[i])
      if (i == 1) {
        a <- unlist(strsplit(filenames[i], ".", fixed = TRUE))
        prefix <- paste(a[1:length(a) - 1], sep = ".")
      }
    }
  }
}
```

```
        }
    }
    if (output == "screen") {
        if ((i%(col * row)) == 0) {
            par(ask = TRUE)
        }
    }
}
if ((customtracks == TRUE) && !is.null(mega2mapfile)) {
    prepareplot(plotdata, chrlist, mega2mapfile, output = "both")
    for (i in (1:numfiles)) {
        if (!is.null(plotdata))
            if (nchar(chrlist[i]) == 1)
                chrlist[i] = paste("0", chrlist[i], sep = "")
        bedplot(paste("bed.data", chrlist[i], sep = "."))
    }
    retval1 <- genomeplot("GG.data.all")
    if (retval1 == FALSE) {
        print("Unable to create BED plot files and Genome Graph plot files.")
    }
}
if (names(dev.cur()) == "postscript" || names(dev.cur()) ==
    "pdf") {
    dev.off()
}
if (retval) {
    return(TRUE)
}
else {
    return(FALSE)
}
}
<environment: namespace:nplplot>
>
>
> # plot with legend
> par(omi=c(0.05, 0.05, 0.5, 0.05))
> data(lods1, package="nplplot")
> nplplot(plotdata=lods1, draw.lgnd=TRUE) # Outputting:
[1] "lods read in as table":
[1] TRUE
> # Outputting: Figure 5.2
># Outputting: Figure 5.3
>
> # plot without legend
> data(lods2, package="nplplot")
> nplplot(plotdata=lods2, draw.lgnd=FALSE) # Outputting:
[1] "lods read in as table"
[1] TRUE
> # Outputting: Figure 5.4
```

```

>
> # plotting from a data file
> datadir <- paste(system.file("data", package="nplplot"),
+                     .Platform$file.sep, sep="")
> nplplot(filename=paste(datadir, "lodscore2.txt.gz", sep=""))
> # Outputting:
[1] TRUE
> # Outputting: Figure 5.5
>

```

Fig. 5.2 Preparation of space for following graphical outputs



Fig. 5.3 nplplot
(plotdata=lods1, draw.
lgnd=TRUE) **Output**

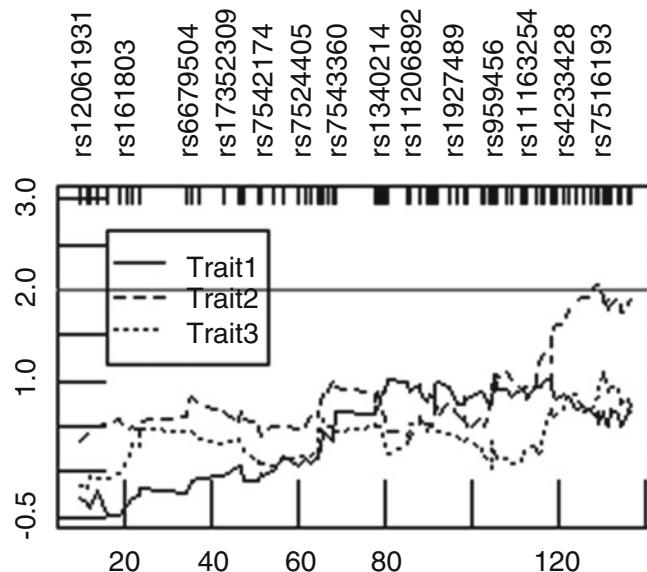


Fig. 5.4 >
nplplot
(plotdata=lods2, draw.
lgnd=FALSE) **Output**

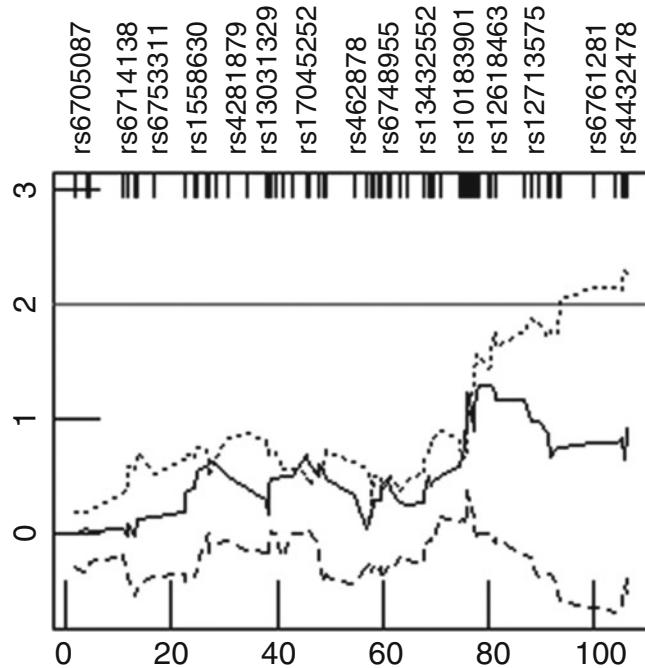
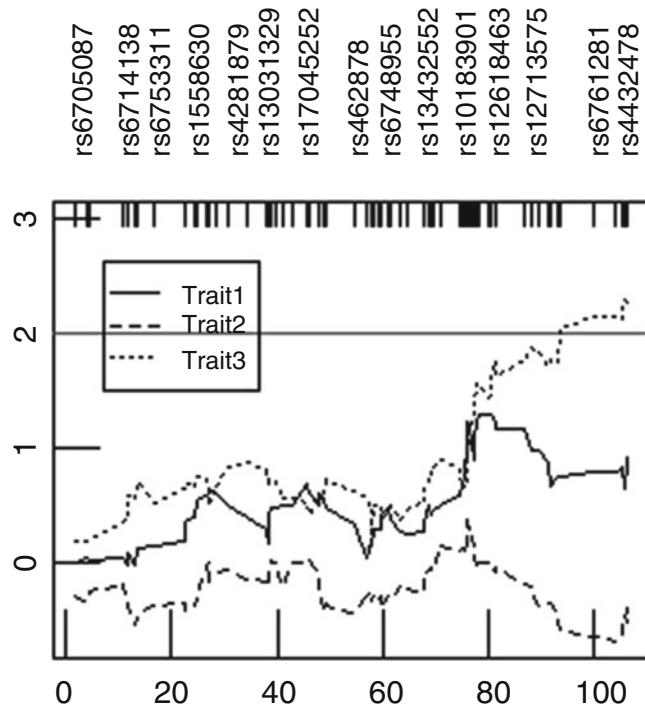


Fig. 5.5 >
nplplot
(filename=paste(datadir,
"lods2.txt.gz", sep=""))
Output



Example 5 Package ‘GenABEL’

February 19, 2015

Type	Package
Title	Genome-Wide SNP Association Analysis
Version	1.8-0
Date	2013-12-09
Author	GenABEL project developers
Contact	GenABEL project developers <genabel.project at gmail.com>
Maintainer	Yurii Aulchenko <yurii@bionet.nsc.ru>
Depends	R (>= 2.15.0), methods, MASS, utils, GenABEL.data
Suggests	qvalue, genetics, haplo.stats, DatABEL (>= 0.9-0), hglm, MetABEL, PredictABEL, VariABEL, bigRR
Description	a package for genome-wide association analysis between quantitative or binary traits and Single-Nucleotide Polymorphisms (SNPs).
License	GPL (>= 2)
URL	http://www.genabel.org , http://forum.genabel.org , http://genabel.r-forge.r-project.org/
BugReports	http://r-forge.r-project.org/tracker/?group_id=505
NeedsCompilation:	yes
Repository	CRAN
Date/Publication:	2013-12-27 14:47:02

R topics documented:

add.plot Function to Plot Additional GWAA Results

Description Add plot of results of GWA analysis**Usage**

```
add.plot(x, ..., df = 1, col=c("lightgreen", "lightblue"),
          sort=TRUE, delta = 1)
```

Arguments

x	object of type scan.gwaa-class, as returned by scan.glm, qtscore, ccfast, emp.ccfast, emp.qtscore, or scan.haplo; or of type scan.gwaa.2D-class, as returned by scan.haplo.2D or scan.glm.2D.
...	additional arguments to be passed to plot
df	P-value at which df to add (1, 2 or "Pc1df")
col	which colors to use to depict consecutive chromosomes
sort	whether results should be plotted after sorting by chromosome and position
delta	gap width between chromosomes
Value	No value returned.
Author	Yurii Aulchenko

See Also

plot,.snp.subset, scan.glm, qtsscore, ccfast, emp.qtsscore, emp.ccfast, scan.haplo, scan.haplo.2D, scan.glm.2D

Examples

```
require(GenABEL.data)
data(srdata)
a <- ccfast("bt",srdata,snps=c(1:100))
plot(a)
a1 <- qtsscore(bt,srdata,snps=c(1:100))
add.plot(a1,col="red",type="l")
```

In the R domain

```
>
> install.packages("GenABEL")
Installing package into 'C:/Users/Bert/Documents/R/win-library/3.3'
(as 'lib' is unspecified)
trying URL 'https://ftp.gwdg.de/pub/misc/cran/bin/windows/contrib/3.3/GenABEL_1.8-0.zip'
Content type 'application/zip' length 3644015 bytes (3.5 MB)
downloaded 3.5 MB

package 'GenABEL' successfully unpacked and MD5 sums checked

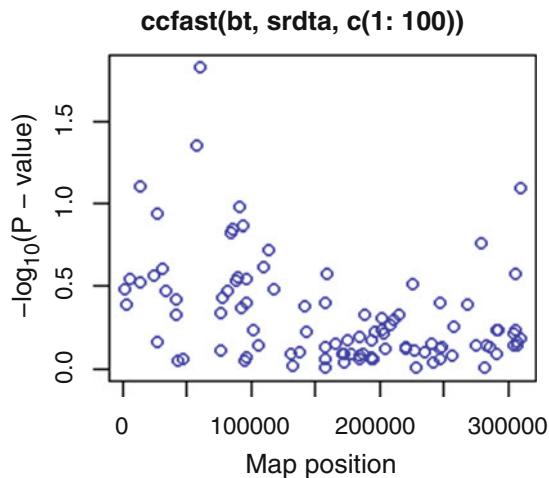
The downloaded binary packages are in
  C:\Users\Bert\AppData\Local\Temp\RtmpqObidp\downloaded_packages

> library(GenABEL)
Loading required package: MASS
Loading required package: GenABEL.data

> ls("package:GenABEL")
[1] "add.phdata"                  "add.plot"
[3] "annotation"                 "arrange_probabel_phe"
[5] "as.character.gwaa.data"    "as.character.snp.coding"
[7] "as.character.snp.data"      "as.character.snp.strand"
[9] "as.data.frame.gwaa.data"   "as.double.gwaa.data"
[11] "as.double.snp.data"        "as.genotype"
[13] "as.genotype.gwaa.data"     "as.genotype.snp.data"
[15] "as.hsgeno"                  "as.hsgeno.gwaa.data"
[17] "as.hsgeno.snp.data"        "autosomal"
[19] "blurGenotype"                "catable"
[21] "ccfast"                      "check.marker"
[23] "check.trait"                 "checkPackageVersionOnCRAN"
[25] "chi2_CG"                      "chromosome"
[27] "cocohet"                      "coding"
```

```
[29] "coding<-"                      "convert.snp.affymetrix"
[31] "convert.snp.illumina"          "convert.snp.mach"
[33] "convert.snp.ped"                "convert.snp.text"
[35] "convert.snp.tped"              "crnames"
[37] "del.phdata"                   "descriptives.marker"
[39] "descriptives.scan"             "descriptives.trait"
[41] "dprfast"                      "effallele"
[43] "egscore"                      "egscore.old"
[45] "emp.ccfast"                  "emp.qtscore"
[47] "estlambda"                   "export.impute"
[49] "export.merlin"                "export.plink"
[51] "extract.annotation.impute"    "extract.annotation.mach"
[53] "findRelatives"                "formetascore"
[55] "GASurv"                      "generateOffspring"
[57] "getcall"                      "getfamily"
[59] "getLogLikelihoodGivenRelation" "grammar"
[61] "gtdata"                       "hom"
[63] "hom.old"                      "HWE.show"
[65] "ibs"                           "ibs.old"
[67] "idnames"                      "impute2databel"
[69] "impute2mach"                  "lambda"
[71] "load.gwaa.data"               "mach2databel"
[73] "makeTransitionMatrix"          "male"
[75] "map"                           "merge.gwaa.data"
[77] "merge.snp.data"               "mlreg"
[79] "mlreg.p"                      "mmscore"
[81] "nids"                          "npsubtreated"
[83] "nsnps"                         "patch_strand"
[85] "perid.summary"                "PGC"
[87] "phdata"                        "phdata<="
[89] "plot.check.marker"             "plot.scan.gwaa"
[91] "plot.scan.gwaa.2D"             "polygenic"
[93] "polygenic_hglm"                "qtscore"
[95] "qvaluebh95"                  "r2fast"
[97] "r2fast.old"                   "recodeChromosome"
[99] "reconstructNPs"                "redundant"
[101] "refallele"                   "refresh.gwaa.data"
[103] "reg.gwaa"                     "results"
[105] "rhofast"                      "rntransform"
[107] "save.gwaa.data"               "scan.glm"
[109] "scan.glm.2D"                 "scan.haplo"
[111] "scan.haplo.2D"                "show"
[113] "show.ncbi"                    "snp.data"
[115] "snp.names"                   "snp.subset"
[117] "snpnames"                     "sortmap.internal"
[119] "sset"                          "strand"
[121] "strand<-"                   "summary.check.marker"
[123] "summary.snp.data"              "summary.snp.data.old"
[125] "VIFGC"                        "VIFGC_ovdom"
[127] "Xfix"                         "ztransform"
```

Fig. 5.6 Plotting additional GWAA results



```

> require(GenABEL.data)
> data(srdta)
> a <- ccfast("bt", srdta, snps=c(1:100))
Warning in ccfast("bt", srdta, snp

> s = c(1:100) :
 11 people (out of 2500 ) excluded as not having cc status

> plot(a)
> # Outputting: Figure 5.6
>

```

Example 6

Ccfast Fast Case-Control Analysis

Description

Fast case-control analysis by computing chi-square test from 2x2 (allelic) or 2x3 (genotypic) tables

Usage

```
ccfast(y, data, snpsubset, idsubset, times=1,
       quiet=FALSE, bcast=10, clambda=TRUE, propPs=1.0)
```

Arguments

- y character name of the vector of case-control status. Cases are denoted as 1 and controls as 0.
 data An object of gwaa.data-class

snpsubset	Index, character or logical vector with subset of SNPs to run analysis on. If missing, all SNPs from data are used for analysis.
idssubset	Index, character or logical vector with subset of SNPs to run analysis on. If missing, all people from data are used for analysis.
times	If more than one, the number of replicas to be used in derivation of empirical genome-wide significance. See emp.qtscore, which calls qtscore with times>1 for details
quiet	do not print warning messages
bcast	If the argument times > 1, progress is reported once in bcast replicas
clambda	If inflation facot Lambda is estimated as lower then one, this parameter controls if the original P1df (clambda=TRUE) to be reported in Pc1df, or the original 1df statistics is to be multiplied onto this “deflation” factor (clambda=FALSE). If a numeric value is provided, it is used as a correction factor.
propPs	proportion of non-corrected P-values used to estimate the inflation factor Lambda, passed directly to the estlambda
Value	Object of class scan.gwaa-class
Author	Yuriii Aulchenko
See Also	emp.ccfast, plot.scan.gwaa, scan.gwaa-class

Examples

```
require(GenABEL.data)
data(srdta)
a <- ccfast("bt", data=srdta, snps=c(1:10), ids=c(1:100))
a
a <- ccfast("bt", data=srdta)
plot(a)
```

In the R domain:

```
>
> require(GenABEL.data)
> data(srdta)
> a <- ccfast("bt", data=srdta, snps=c(1:10), ids=c(1:100))
Warning in ccfast("bt", data = srdta, snps = c(1:10), ids = c(1:100)) :
  11 people (out of 100 ) excluded as not having cc status

> a # Outputting:
***** 'scan.gwaa' object *****
*** Produced with:
ccfast(y = "bt", data = srdta, snpsubset = c(1:10), idsubset = c(1:100))
*** Test used: chi2 test for association, 2x2 and 2x3 tables
*** no. IDs used: 89 ( p1 p2 p3 , ... )
*** Lambda: NA
*** Results table contains 10 rows and 10 columns
```

Chromo-	Position	Strand	A1	A2	N	effB	se_effB	ch11.1df
some								

rs10	1	2500	+	T	G	82	1.4880952	1.9362912	0.59063625
rs18	1	3500	+	G	A	87	0.7348928	0.7499131	0.96034232
rs29	1	5750	-	G	T	78	1.6741071	1.6806166	0.99226847
rs65	1	13500	+	A	T	84	0.9365079	4.9406487	0.03592981
rs73	1	14250	+	A	G	85	1.2000000	3.015963	0.15831068
rs114	1	24500	+	A	T	85	1.3888889	2.1130089	0.43204834
rs128	1	27000	-	G	T	82	0.8043478	1.7401278	0.21366112
rs130	1	27250	+	A	G	86	1.1922504	2.3226615	0.26348936
rs143	1	31000	+	T	G	83	0.7893098	1.0381398	0.57807379
rs150	1	33250	+	C	A	87	0.6110599	0.3964254	2.37599072

	P1df	Pc1df	effAB	effBB	chi2.2df	
P2df						
rs10	0.4421732	0.4421732	2.258065e+00	0.0000000	2.9090909	0.23350648
rs18	0.3271006	0.3271006	1.500000e+00	0.5384615	2.7888742	0.24797258
rs29	0.3191886	0.3191886	2.592593e+00	0.0000000	3.5030089	0.17351270
rs65	0.8496605	0.8496605	2.533333e+00	1.6666667	1.4540965	0.48333356
rs73	0.6907166	0.6907166	1.000000e+04	1.0277778	2.0022110	0.36747297
rs114	0.5109856	0.5109856	1.989247e+00	0.0000000	2.4265521	0.29722197
rs128	0.6439129	0.6439129	1.636364e+00	0.6545455	0.6368282	0.72730156
rs130	0.6077330	0.6077330	2.461538e+00	2.0000000	1.3103448	0.51935251
rs143	0.4470683	0.4470683	2.272727e+00	0.5909091	6.1891256	0.04529481
rs150	0.1232134	0.1232134	8.518519e-01	0.4242424	2.7764064	0.24952324

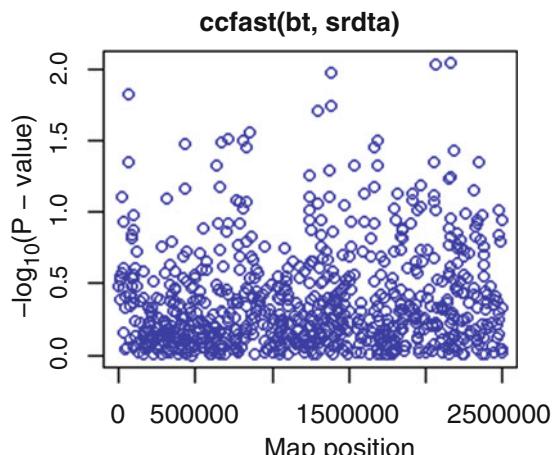

```
____ Use 'results(object)' to get complete results table ____
> a <- ccfast("bt", data=srdta)
Warning in ccfast("bt", data = srdta) :
  11 people (out of 2500 ) excluded as not having cc status

> plot(a)

> # Outputting: Figure 5.7

>
```

Fig. 5.7 Output of Ccfast fast case-control analysis



5.2.2 Manhattan Plots [Wikipedia]

A **Manhattan Plot**, as used in genetic epidemiology, is a [scatter plot](#), used to show data with a large number of data-points - many of non-zero amplitude, and with a distribution of higher-magnitude values, for instance in [Genome-Wide Association Studies \(GWAS\)](#). In GWAS Manhattan plots, genomic coordinates are displayed along the X-axis, with the negative [logarithm](#) of the association *P*-value for each [Single Nucleotide Polymorphism \(SNP\)](#) displayed on the Y-axis. Thus, each dot on the Manhattan plot signifies a SNP. As the strongest associations have the smallest *P*-values (e.g., 10^{-20}), their *negative logarithm will be the greatest* (e.g., $-(-20) = 20$).

Its name comes from the similarity of such a plot to the [Manhattan](#) Island [skyline](#) in New York City, viz., resembling the profile of skyscrapers towering above the lower level "buildings" which vary around a lower height.

Figure 5.8 shows a typical Manhattan Plot for GWAS:

Worked Example Using a Manhattan Plot Display

R Package gap

January 24, 2018

Version	1.1-21
Date	2018-1-23
Title	Genetic Analysis Package
Author	Jing Hua Zhao and colleagues with inputs from Kurt Hornik and Brian Ripley
Maintainer	Jing Hua Zhao <jinghua.zhao@mrc-epid.cam.ac.uk>
Depends	R (>= 2.10)
Suggests	BradleyTerry2, MASS, Matrix, MCMCglmm, R2jags, bdsmatrix, coda, coxme, foreign, grid, haplo.stats, kinship2, lattice,magic, mets, nlme, pedigree, pedigreeemm, regress, rms, survival
LazyData	Yes
LazyLoad	Yes
Description	It is designed as an integrated package for genetic data analysis of both population and family data. Currently, it contains functions for sample size calculations of both population-based and family-based designs, probability of familial disease aggregation, kinship calculation, statistics in linkage analysis,

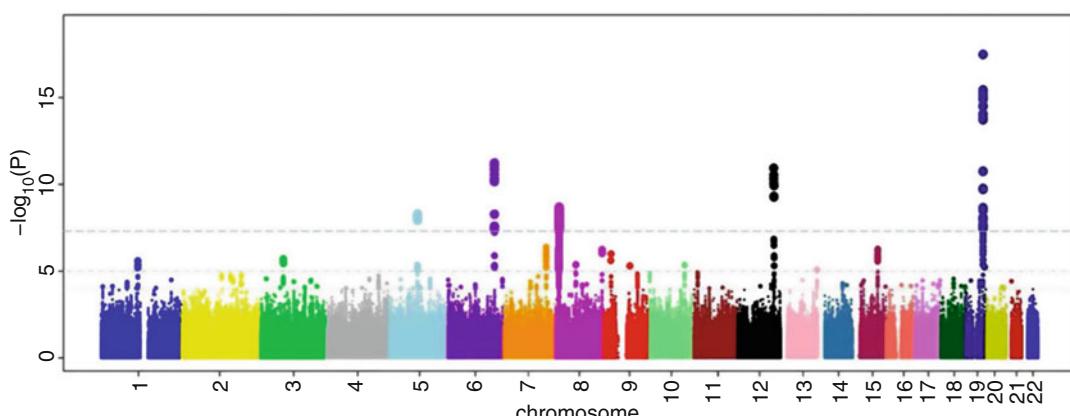


Fig. 5.8 A manhattan plot indicating some strongly associated risk loci

and association analysis involving genetic markers including haplotype analysis with or without environmental covariates.

License	GPL (>= 2)
URL	https://jinghuazhao.github.io
NeedsCompilation	yes
Repository	CRAN
Date/Publication	2018-01-24 11:40:17 UTC

The following examples are selected:

mhtplot **Manhattan Plots**

Description

To generate Manhattan plot, e.g., of genomewide significance (p values) and a random variable that is uniformly distributed. By default, a log10-transformation is applied. Note that with real chromosomal positions, it is also appropriate to plot some but not all chromosomes.

It is possible to specify options such as xlab and ylim when the plot is requested for data in other context.

Usage

```
mhtplot(data, control=mht.control(), hcontrol=hmht.control(), ...)
```

Arguments

data a data frame with three columns representing chromosome, position and p values
control A control function named `mht.control()` with the following arguments:

1. **type**. a flag with value "p" or "l" indicating if points or lines are to be drawn.
2. **usepos**. a flag to use real chromosomal positions as composed to ordinal positions with default value FALSE
3. **logscale**. a flag to indicate if the p value is to be log-transformed with default value TRUE
4. **base**. the base of the logarithm with default value 10
5. **cutoffs**. the cut-offs where horizontal line(s) are drawn with default value NULL
6. **colors**. the color for different chromosome(s), and random if unspecified with default values NULL
7. **labels**. labels for the ticks on x-axis with default value NULL
8. **srt**. degree to which labels are rotated with default value of 45
9. **gap**. gap between chromosomes with default value NULL
10. **cex**. cex for the data points
11. **yline**. Margin line position
12. **xline**. Margin line position

hcontrol A control function named `hmht.control()` with the following arguments:

1. `data`. chunk of data to be highlighted with default value `NULL`
2. `colors`. colors for annotated genes
3. `yoffset`. offset above the data point showing most significant p value with default value `0.5`
4. `cex`. shrinkage factor for data points with default value `1.5`
5. `boxed`. if the label for the highlighted region with default value `FALSE`

... other options in compatible with the R plot function

Value The plot is shown on or saved to the appropriate device.

Author Jing Hua Zhao

See Also `qqunif`

Examples

```
## Not run:
# foo example
test <- matrix(c(1,1,4,1,1,6,1,10,3,2,1,5,2,2,6,2,4,8),byrow=TRUE,6)
mhtplot(test)
mhtplot(test,mht.control(logscale=FALSE))

# fake example with Affy500k data
affy <-c(40220, 41400, 33801, 32334, 32056, 31470, 25835, 27457,
        22864, 28501, 26273, 24954, 19188, 15721, 14356, 15309,
        11281, 14881, 6399, 12400, 7125, 6207)
CM <- cumsum(affy)
n.markers <- sum(affy)
n.chr <- length(affy)
test <- data.frame(chr=rep(1:n.chr,affy),pos=1:n.markers,p=runif(n.markers))

# to reduce size of the plot
# bitmap("mhtplot.bmp",res=72*5)
oldpar <- par()
par(cex=0.6)
colors <- rep(c("blue","green"),11)
# other colors, e.g.
# colors <- c("red","blue","green","cyan","yellow","gray","magenta","red","-blue","green",
# "cyan","yellow","gray","magenta","red","blue","green","cyan","yellow","-gray",
# "magenta","red")
mhtplot(test,control=mht.control(colors=colors),pch=19,srt=0)
title("A simulated example according to EPIC-Norfolk QCed SNPs")
axis(2)
axis(1,pos=0,labels=FALSE,tick=FALSE)
abline(0,0)
# dev.off()
par(oldpar)
```

```
mhtplot(test,control=mht.control(usepos=TRUE,colors=colors,gap=10000),pch=19,
bg=colors)
title("Real positions with a gap of 10000 bp between chromosomes")
box()

png("manhattan.png",height=3600,width=6000,res=600)
opar <- par()
par(cex=0.4)
ops <- mht.control(colors=rep(c("lightgray","lightblue"),11),srt=0,yline=2.5,
xline=2)
mhtplot(mhtdata[,c("chr","pos","p")],ops,xlab="",ylab="",srt=0)
axis(2,at=1:16)
title("An adaptable plot as .png")
par(opar)
dev.off()

data <- with(mhtdata,cbind(chr,pos,p))
glist <- c("IRS1","SPRY2","FTO","GRIK3","SNED1","HTR1A","MARCH3","-WISP3","PPP1R3B",
"RP111","FDFT1","SLC39A14","GFRA1","MC4R")
hdata <- subset(mhtdata,gene%in%glist)[c("chr","pos","p","gene")]
color <- rep(c("lightgray","gray"),11)
glen <- length(glist)
hcolor <- rep("red",glen)
par(las=2, xpd=TRUE, cex.axis=1.8, cex=0.4)
ops <- mht.control(colors=color,yline=1.5,xline=3,labels=paste("chr",1:22,
sep=""),
srt=270)
hops <- hmht.control(data=hdata,colors=hcolor)
mhtplot(data,ops,hops,pch=19)
axis(2, pos=2, at=1:16)
title("Manhattan plot with genes highlighted",cex.main=1.8)

mhtplot(data,mht.control(cutoffs=c(4,6,8,16)),pch=19)
title("Another plain Manhattan plot")

## End(Not run)
```

In the R domain:

```
> install.packages("gap")
Installing package into 'C:/Users/Bert/Documents/R/win-library/3.3'
(as 'lib' is unspecified)
--- Please select a CRAN mirror for use in this session ---
```

A CRAN mirror is selected.

```
trying URL 'https://mirrors.tuna.tsinghua.edu.cn/CRAN/bin/windows/contrib/3.3/
gap_1.1-21.zip'
Content type 'application/zip' length 3319010 bytes (3.2 MB)
downloaded 3.2 MB

package 'gap' successfully unpacked and MD5 sums checked
```

The downloaded binary packages are in

```
C:\Users\Bert\AppData\Local\Temp\RtmpQV1ppn\downloaded_packages
> library(gap)
gap version 1.1-21
> ls("package:gap")
 [1] "a2g"                 "ab"                  "AE3 "
 [4] "aldh2"               "allele.recode"      "apoeapoc "
 [7] "asplot"               "b2r"                  "BFDP "
 [10] "bt"                  "ccsize"              "cf "
 [13] "chow.test"            "comp.score"         "cov.invlogit "
 [16] "Cox.est"              "Cox.T"                "crohn "
 [19] "DevH0dominant"        "DevH0dominant.est"  "DevH0recessive "
 [22] "DevH0recessive.est"   "DevHaGdominant"     "DevHaGdominant.est"
 [25] "DevHaGrecessive"      "DevHaGrecessive.est" "ESplot "
 [28] "fa"                  "fbsize"              "FPRP "
 [31] "fsnps"                "g2a"                  "g2a.c "
 [34] "gc.control"            "gc.em"                "gcode "
 [37] "gcontrol"              "gcontrol2"            "gcp "
 [40] "genecounting"          "geno.recode"          "getblstar "
 [43] "getPTE"                "gif"                  "grec2g "
 [46] "h2"                   "h2G"                  "h2GE "
 [49] "h2l"                  "hap"                  "hap.control "
 [52] "hap.em"                "hap.score"            "HapDesign "
 [55] "HapFreqSE"              "hla"                  "hmht.control "
 [58] "htr"                  "hwe"                  "hwe.cc "
 [61] "hwe.hardy"              "invlogit"             "k "
 [64] "KCC"                  "kin.morgan"          "klem "
 [67] "151"                  "LD22"                 "LDkl "
 [70] "logit"                 "lukas"                "m2plem "
 [73] "makeped"               "mao"                  "masize "
 [76] "MCMCgrm"               "metap"                "metareg "
 [79] "meyer"                 "mfblong"              "mht.control "
 [82] "mhtplot"                "mhtplot2"              "mia "
 [85] "micombine"              "mtdt"                 "mtdt2 "
 [88] "muvar"                 "mvmeta"              "nep499 "
```

```
[91] "PARn"                  "pbsize"                 "pbsize2"
[94] "PD"                     "pedtodot"                "pfc"
[97] "pfc.sim"                "pgc"                     "plem2m"
[100] "plot.hap.score"        "print.hap.score"        "qqfun"
[103] "qqunif"                 "read.ms.output"        "ReadGRM"
[106] "ReadGRMBin"             "ReadGRMPCA"              "ReadGRMPLINK"
[109] "revhap"                 "revhap.i"                "s2k"
[112] "se.exp"                 "se.invlogit"            "snp.ES"
[115] "snp.HWE"                 "snp.PAR"                 "solve_skol"
[118] "toETDT"                  "tscc"                     "ungcode"
[121] "VR"                      "whscore"                 "WriteGRM"
[124] "WriteGRMBin"             "WriteGRMSAS"             "x2"
[127] "z"

>
> # foo example
> test <- matrix(c(1,1,4,1,1,6,1,10,3,2,1,5,2,2,6,2,4,8),byrow=TRUE,6)
> mhtplot(test)
Plotting points 1 - 3
Plotting points 4 - 6
> mhtplot(test,mht.control(logscale=FALSE)) # Outputting: Figure 5.9
```

Fig. 5.9 Outline of:
mhtplot(test,mht.control
(logscale=FALSE))



```
>
>
> # foo example
> test <- matrix(c(1,1,4,1,1,6,1,10,3,2,1,5,2,2,6,2,4,8),byrow=TRUE,6)
> mhtplot(test)
Plotting points 1 - 3
Plotting points 4 - 6
> mhtplot(test,mht.control(logscale=FALSE))
Plotting points 1 - 3
Plotting points 4 - 6
>
>
> # fake example with Affy500k data
> affy <-c(40220, 41400, 33801, 32334, 32056, 31470, 25835, 27457,
+           22864, 28501, 26273, 24954, 19188, 15721, 14356, 15309,
+           11281, 14881, 6399, 12400, 7125, 6207)
> CM <- cumsum(affy)
> n.markers <- sum(affy)
> n.chr <- length(affy)
> test <-
+   data.frame(chr=rep(1:n.chr,affy),pos=1:n.markers,p=runif(n.markers))
> # to reduce size of the plot
> # bitmap("mhtplot.bmp",res=72*5)
>
> oldpar <- par()
> par(cex=0.6)
> colors <- rep(c("blue","green"),11)
> # other colors, e.g.
> # colors <-
+ c("red","blue","green","cyan","yellow","gray","magenta",
+   "red","blue","green",0"cyan","yellow","gray","magenta",
+   "magenta","red")
> mhtplot(test,control=mht.control(colors=colors),pch=19,srt=0)
Plotting points 1 - 40220
Plotting points 40221 - 81620
Plotting points 81621 - 115421
Plotting points 115422 - 147755
Plotting points 147756 - 179811
Plotting points 179812 - 211281
Plotting points 211282 - 237116
Plotting points 237117 - 264573
Plotting points 264574 - 287437
Plotting points 287438 - 315938
Plotting points 315939 - 342211
Plotting points 342212 - 367165
Plotting points 367166 - 386353
Plotting points 386354 - 402074
Plotting points 402075 - 416430
Plotting points 416431 - 431739
Plotting points 431740 - 443020
```

```
Plotting points 443021 - 457901
Plotting points 457902 - 464300
Plotting points 464301 - 476700
Plotting points 476701 - 483825
Plotting points 483826 - 490032
>
> # Outputting: Figure 5.10
> #
> title("A simulated example according to EPIC-Norfolk QCed SNPs")
> # Outputting: Figure 5.11
>
> axis(2)
>
> # Outputting: Figure 5.12
> #
> axis(1, pos=0, labels=FALSE, tick=FALSE)
>
> # Outputting: Figure 5.13
> #
> abline(0, 0)
> # Outputting: Figure 5.14
```

Fig. 5.10 Manhattan plot of genomewide significance, using: mhtplot (test,mht.control (logscale=FALSE))

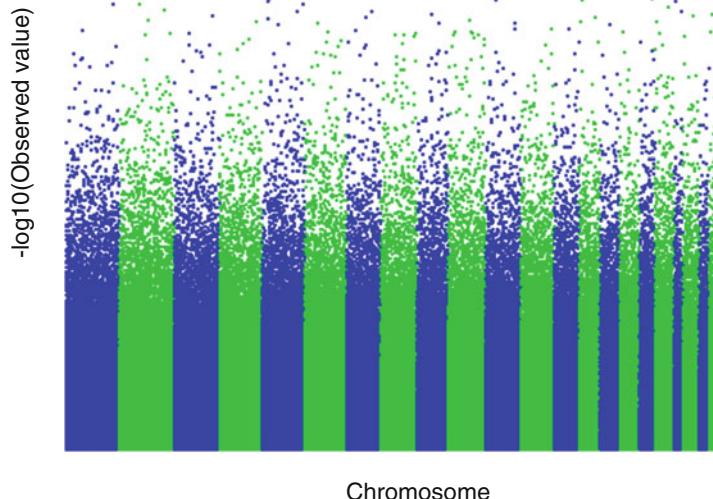


Fig. 5.11 Manhattan plot of genomewide significance, using: mhtplot (test,mht.control (logscale=FALSE))

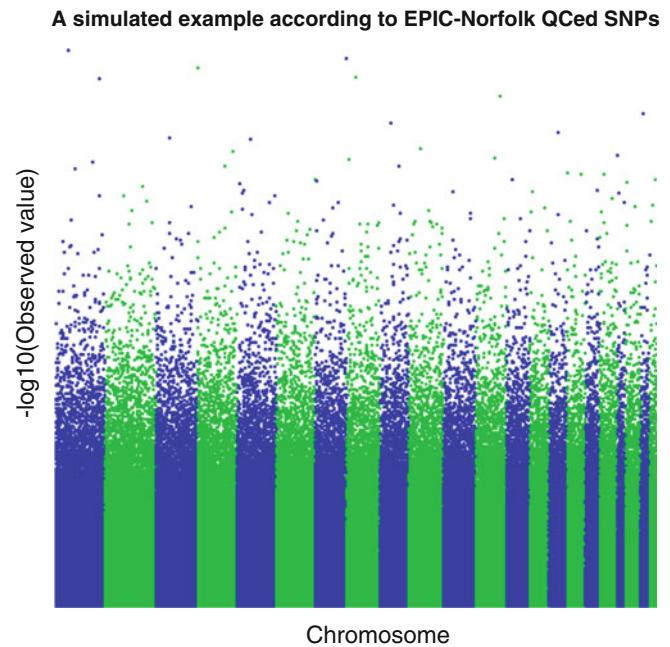


Fig. 5.12 Manhattan plot of genomewide significance, using: mhtplot (test,mht.control(logscale))

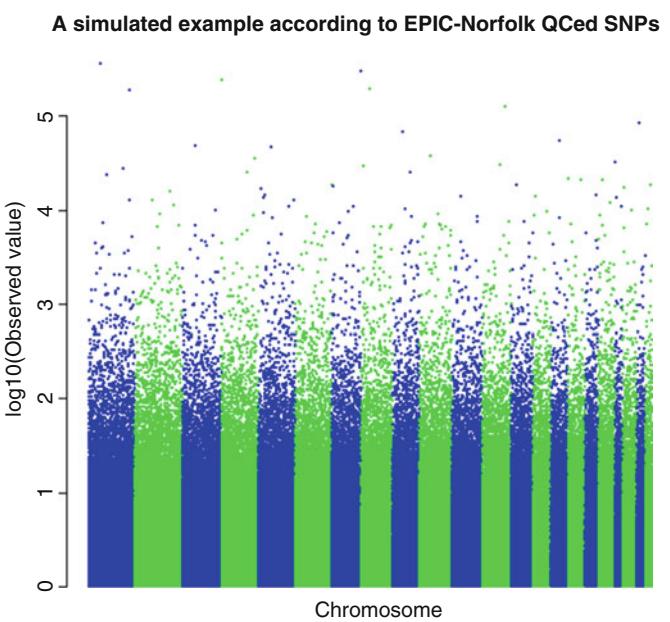


Fig. 5.13 Manhattan plot of genomewide significance, using: mhplot (test,mht.control(logscale))

A simulated example according to EPIC-Norfolk QCed SNPs

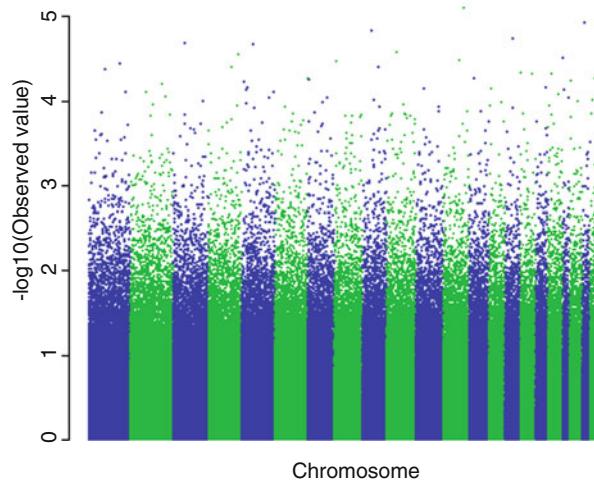
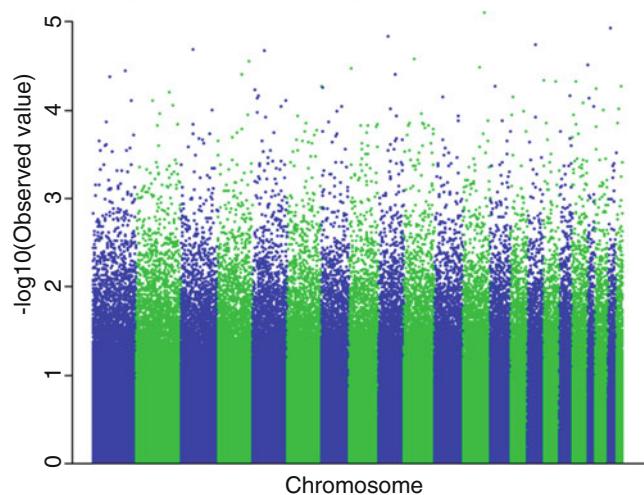


Fig. 5.14 Manhattan plot of genomewide significance, using: mhplot (test,mht.control(logscale))

A simulated example according to EPIC-Norfolk QCed SNPs



```
> #
> # dev.off()
> par(oldpar)
Warning messages:
1: In par(oldpar) : graphical parameter "cin" cannot be set
2: In par(oldpar) : graphical parameter "cra" cannot be set
3: In par(oldpar) : graphical parameter "csi" cannot be set
4: In par(oldpar) : graphical parameter "cxy" cannot be set
5: In par(oldpar) : graphical parameter "din" cannot be set
6: In par(oldpar) : graphical parameter "page" cannot be set
>
> mhtplot(test,control=mht.control(usepos=TRUE,
+           colors=colors,gap=10000), pch=19,bg=colors)
Plotting points 1 - 40220
Plotting points 40221 - 81620
Plotting points 81621 - 115421
Plotting points 115422 - 147755
Plotting points 147756 - 179811
Plotting points 179812 - 211281
Plotting points 211282 - 237116
Plotting points 237117 - 264573
Plotting points 264574 - 287437
Plotting points 287438 - 315938
Plotting points 315939 - 342211
Plotting points 342212 - 367165
Plotting points 367166 - 386353
Plotting points 386354 - 402074
Plotting points 402075 - 416430
Plotting points 416431 - 431739
Plotting points 431740 - 443020
Plotting points 443021 - 457901
Plotting points 457902 - 464300
Plotting points 464301 - 476700
Plotting points 476701 - 483825
Plotting points 483826 - 490032
> title("Real positions with a gap of 10000 bp between chromosomes")
> box()
> # Outputting: Figure 5.15
> # Outputting: Figure 5.16
>
```

Fig. 5.15 Manhattan plot of genomewide significance, using: mhplot (test,mht.control(logscale))

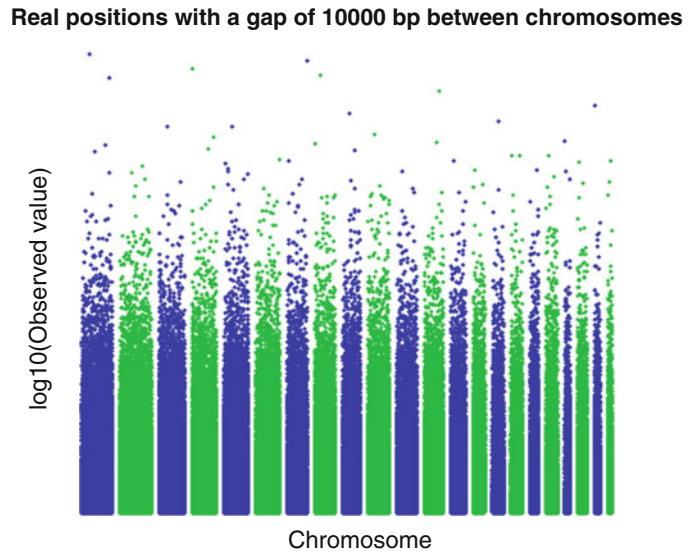
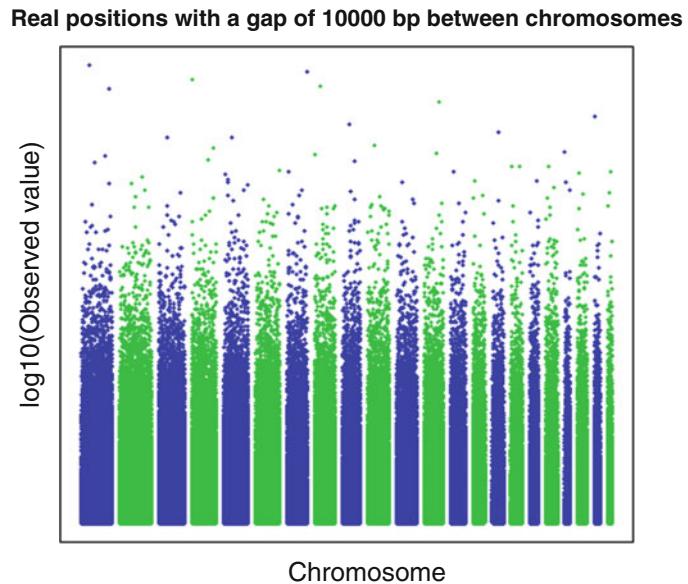


Fig. 5.16 Manhattan plot of genomewide significance, using: mhplot (test,mht.control(logscale))



5.3 Procedures for Multiple Comparison

In **biostatistics**, the multiple comparisons, multiplicity, or multiple testing problem occurs when one considers a set of **statistical inferences** simultaneously or infers a subset of parameters selected based on the observed values. It is also known as the “*look-elsewhere effect*”. The more inferences are made, the more likely erroneous inferences are to occur. Several statistical techniques have been developed to prevent this from happening, allowing significance levels for single and multiple comparisons to be

directly compared. These techniques generally require a stricter significance threshold for individual comparisons, so as to compensate for the number of inferences being made.

Analysis of multiple comparisons began in the 1950s with the work of Tukey and [Scheffé](#). Other methods, such as the [Closed Testing Procedure](#) (1976) and the Holm-Bonferroni method (1979), came later. In 1995, work on the False Discovery Rate began. In 1996, the first conference on multiple comparisons took place. This was followed by conferences around the world, usually taking place about every two years.

Multiple comparisons arise when a statistical analysis involves multiple simultaneous statistical tests, each of which has a potential to produce a "discovery."

A stated confidence level generally applies only to each test considered individually, but often it is desirable to have a confidence level for the whole family of simultaneous tests. Failure to compensate for multiple comparisons can have important real-world consequences, as illustrated by the following examples:

Example 1 Suppose the treatment is a new way of teaching writing to students, and the control is the standard way of teaching writing. Students in the two groups can be compared in terms of grammar, spelling, organization, content, and so on. As more attributes are compared, it becomes increasingly likely that the treatment and control groups will appear to differ on at least one attribute due to random sampling error alone.

Example 2 Suppose one considers the efficacy of a [drug](#) in terms of the reduction of any one of a number of disease symptoms. As more symptoms are considered, it becomes increasingly likely that the drug will appear to be an improvement over existing drugs in terms of at least one symptom.

In both examples, as the number of comparisons increases, it becomes more likely that the groups being compared will appear to differ in terms of at least one attribute.

One's confidence that a result will generalize to independent data should generally be weaker if it is observed as part of an analysis that involves multiple comparisons, rather than an analysis that involves only a single comparison.

For example, if one test is performed at the 5% level and the corresponding null hypothesis is true, there is only a 5% chance of incorrectly rejecting the null hypothesis. However, if 100 tests are conducted and all corresponding null hypotheses are true, the [expected number](#) of incorrect rejections (also known as [false positives](#) or [Type I errors](#)) is 5. If the tests are statistically independent from each other, the probability of at least one incorrect rejection is 99.4%.

Remarks:

The multiple comparisons problem also applies to [confidence intervals](#). A single confidence interval with a 95% [coverage probability](#) level will contain the population parameter in 95% of experiments. However, if one considers 100 confidence intervals simultaneously, each with 95% coverage probability, the expected number of non-covering intervals is 5. If the intervals are statistically independent from each other, the probability that at least one interval does not contain the population parameter is 99.4%. Techniques have been developed to prevent the inflation of false positive rates and non-coverage rates that occur with multiple statistical tests.

5.3.1 Worked Examples of Statistical Tests and Utilities for Genetic Association

5.3.1.1 Package iGasso

Package iGasso

June 4, 2016

Type	Package
Title	Statistical Tests and Utilities for Genetic Association
Version	1.4
Date	2016-06-03
Author	Dr. Kai Wang
Maintainer	Kai Wang kai-wang@uiowa.edu
Depends	lattice, CompQuadForm
Description	A collection of statistical tests for genetic association studies.
License	GPL (>= 2)
LazyLoad	yes
NeedsCompilation	no
Repository	CRAN
Date/Publication	2016-06-04 07:52:27
genome.plot	Genome-wide Plot of a Variable

Description

genome.plot plots the value of a variable across the genome.

Usage

```
genome.plot(mydata, style=1, type="h", sig.line=c(4, -4),
            sig.color=c("red", "red"), ...)
```

Arguments

mydata	a data frame containing three variables: y (numeric, the value of the variable to be plotted), chr (character, chromosome label), and pos (numeric, position, for instance, in base pair or centi-Morgan). Examples of y include -log10 of p-values and test statistic values.
style	either 1 (default) or 2.
type	a generic graphic parameter. Recommended values are "h" (default) and "b".
sig.line	vertical locations of significance lines.
sig.	colors of significance lines.
color	
...	other parameters to be passed to function xyplot in the lattice package.

Details

This function makes use of the function xyplot from package lattice.

Author Kai Wang kai-wang@uiowa.edu

Examples

```

y = rnorm(100)
chr = c(rep(1, 20), rep(3, 20), rep(10, 20), rep(19, 30), rep("X", 10))
pos = c(1:20, 1:20, 1:20, 1:30, 1:10)
mydata = data.frame(y=y, chr=chr, pos=pos)
mydata2 = data.frame(y=y^2, chr=chr, pos=pos)
genome.plot(mydata, sig.line=c(1, -1), ylab="T Statistic")
genome.plot(mydata, sig.line=c(1, -1), ylab="T Statistic", type="b")
genome.plot(mydata2, sig.line=c(2), ylab="y squared")
genome.plot(mydata, style=2, sig.line=c(1, -1), ylab="T Statistic")
genome.plot(mydata, style=2, sig.line=c(1, -1), ylab="T Statistic", type="b")

```

In the R domain:

```

> install.packages("iGasso")
Installing package into 'C:/Users/Bert/Documents/R/win-library/3.3'
(as 'lib' is unspecified)
--- Please select a CRAN mirror for use in this session ---

```

A CRAN mirror is selected.

```

trying URL 'https://mirrors.tuna.tsinghua.edu.cn/CRAN/bin/windows/contrib/3.3/
iGasso_1.4.zip'
Content type 'application/zip' length 33731 bytes (32 KB)
downloaded 32 KB

```

```
package 'iGasso' successfully unpacked and MD5 sums checked
```

```

The downloaded binary packages are in
  C:\Users\Bert\AppData\Local\Temp\Rtmp27KpvS\downloaded_packages
> library(iGasso)
Loading required package: lattice
Loading required package: CompQuadForm
Warning messages:
1: package 'iGasso' was built under R version 3.3.3
2: package 'CompQuadForm' was built under R version 3.3.3
> ls("package:iGasso")
[1] "genome.plot"    "KAT.coin"      "MFree.test"    "SKATplus"     "VSTF.test"
> genome.plot      # Listing the program: genome.plot
function (mydata, style = 1, type = "h", sig.line = c(4, -4),
  sig.color = c("red", "red"), ...)
{
  mydata = na.omit(mydata)
  t.chr = as.character(mydata$chr)
  t.chr = ifelse(t.chr %in% c("1", "2", "3", "4", "5", "6",
    "7", "8", "9"), paste("0", t.chr, sep = ""), t.chr)
  t.chr = factor(t.chr)

```

```
chr.name = levels(t.chr)
chr.name = ifelse(substr(chr.name, 1, 1) == "0",
                  substring(chr.name, 2), chr.name)

levels(t.chr) = chr.name
n.chr = length(chr.name)
mins = as.vector(tapply(mydata$pos, t.chr, FUN = "min"))
maxs = as.vector(tapply(mydata$pos, t.chr, FUN = "max"))
if (style == 1) {
  xyplot(y ~ pos | t.chr, data = mydata,
         xlab = "Chromosome",
         type = type, ..., layout = c(n.chr, 1),
         scales = list(x = list(relation = "free",
                               draw = FALSE)), par.settings = list(layout.widths =
         list(panel = maxs - mins), axis.line = list(lwd = 0.1),
         strip.border = list(lwd = 0.1)),
         strip = function(..., bg, par.strip.text)
                  strip.default(...,
                               bg = "pink", par.strip.text = list(cex = 0.75)),
         abline = list(h = sig.line, col = sig.color))
}
else if (style == 2) {
  xyplot(y ~ pos | t.chr, data = mydata, xlab
         = "Chromosome",
         type = type, ..., layout = c(n.chr, 1), strip = FALSE,
         scales = list(x = list(relation = "free", tck = c(0, 0),
                               at = as.vector((maxs + mins)/2, mode = "list")),
                     labels = as.vector(chr.name, mode = "list"))),
         par.settings = list(layout.widths = list(panel
         = maxs - mins), axis.line = list(lwd = 0.1)),
         abline = list(h = sig.line,
                       col = sig.color))
}
else stop("The value of shape should be either 1 or 2")
}

<environment: namespace:iGasso>
> y = rnorm(100)
> chr = c(rep(1, 20), rep(3, 20), rep(10, 20), rep(19, 30),
+         rep("X", 10))
> pos = c(1:20, 1:20, 1:20, 1:30, 1:10)
> mydata = data.frame(y=y, chr=chr, pos=pos)
> mydata2 = data.frame(y=y^2, chr=chr, pos=pos)
>
> # Running the program genome.plot
>
> genome.plot(mydata, sig.line=c(1, -1), ylab="T Statistic")
>
> # Outputting: Figure 5.17
>
> genome.plot(mydata, sig.line=c(1, -1),
+             ylab="T Statistic", type="b")
```

Fig. 5.17 > genome.plot
(mydata, sig.line=c(1, -1),
> ylab="T Statistic")

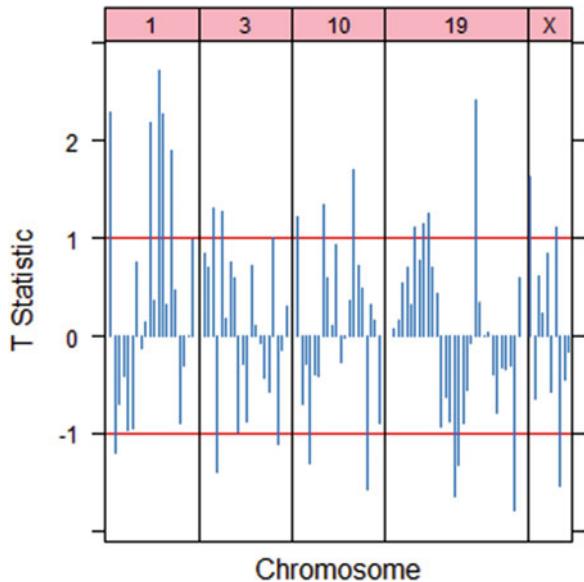
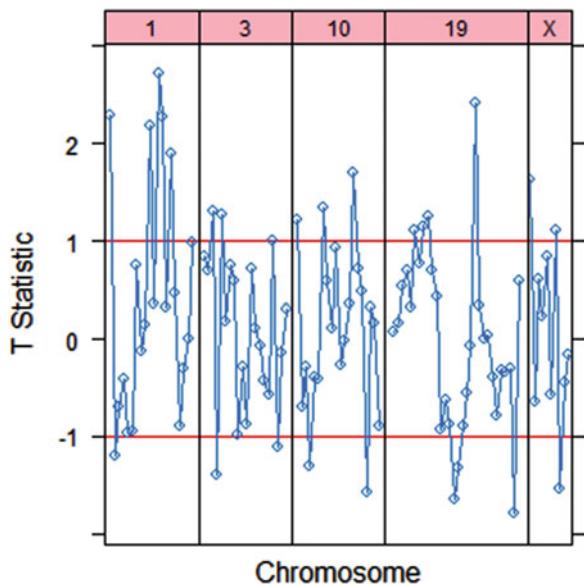


Fig. 5.18 > genome.plot
(mydata, sig.line=c(1, -1),
> ylab="T Statistic",
type="b")



```
>
> # Outputting: Figure 5.18
>
> genome.plot(mydata2, sig.line=c(2), ylab="y squared")
>
> # Outputting: Figure 5.19
>
> genome.plot(mydata, style=2, sig.line=c(1, -1),
+               ylab="T Statistic")
```

```
>  
> # Outputting: Figure 5.20  
>  
> genome.plot(mydata, style=2, sig.line=c(1, -1),  
+ ylab="T Statistic", type="b")  
>  
> # Outputting: Figure 5.21  
>
```

Fig. 5.19 > genome.plot
(mydata2, sig.line=c(2), +
ylab="y squared")

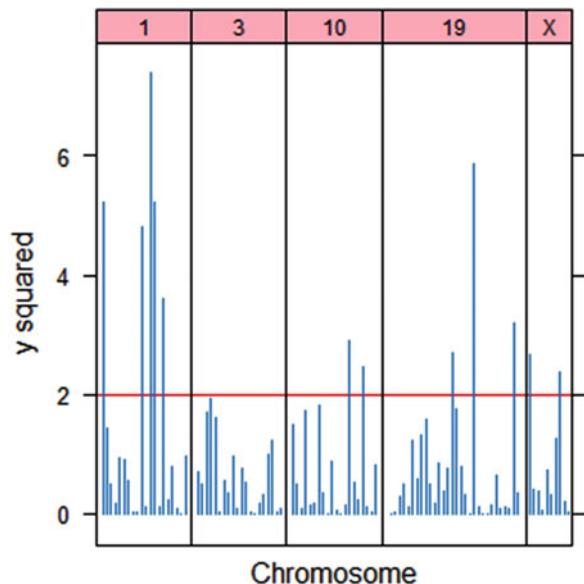


Fig. 5.20 > genome.plot
(mydata, style=2, sig.
line=c(1, -1), + ylab="T
Statistic")

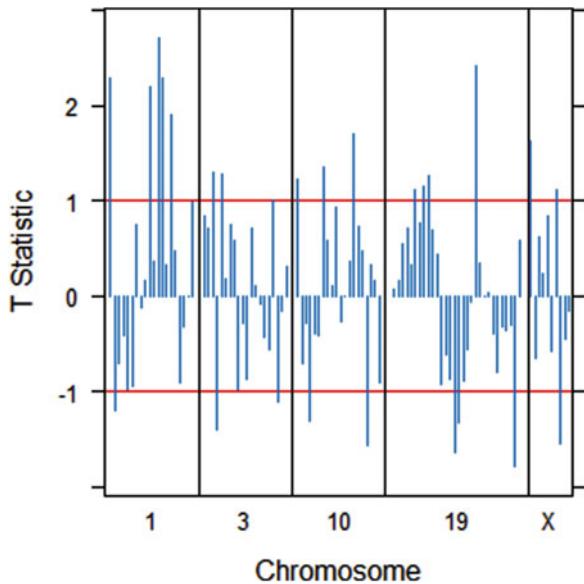
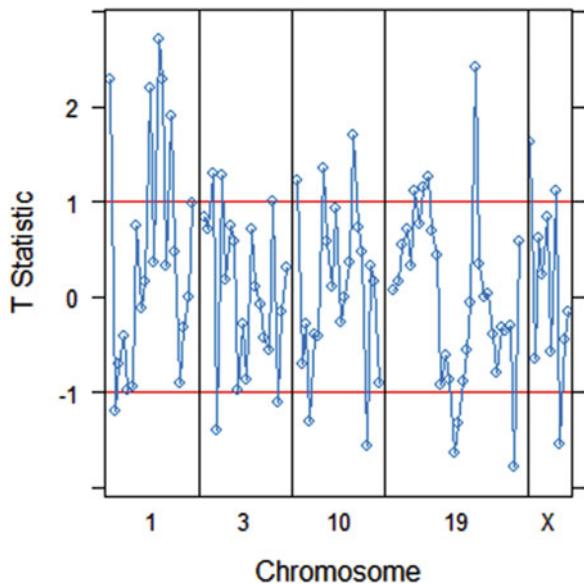


Fig. 5.21 > genome.plot
 (mydata, style=2, sig.line=c(1, -1), + ylab="T
 Statistic", type="b")



5.3.1.2 Package GenABEL

Package GenABEL
 February 19, 2015

Type	Package
Title	Genome-Wide SNP Association Analysis
Version	1.8-0
Date	2013-12-09
Author	GenABEL project developers
Contact	GenABEL project developers <genabel.project at gmail.com>
Maintainer	Yuriii Aulchenko <yuriii@bionet.nsc.ru>
Depends	R (>= 2.15.0), methods, MASS, utils, GenABEL.data
Suggests	qvalue, genetics, haplo.stats, DatABEL (>= 0.9-0), hglm, MetABEL, PredictABEL, VariABEL, bigRR
Description	a package for genome-wide association analysis between quantitative or binary traits and Single-Nucleotide Polymorphisms (SNPs).
License	GPL (>= 2)
URL	http://www.genabel.org , http://forum.genabel.org , http://genabel.r-forge.r-project.org/
BugReports	http://r-forge.r-project.org/tracker/?group_id=505
NeedsCompilation	yes
Repository	CRAN
Date/Publication	2013-12-27 14:47:02

In the R domain:

```
> install.packages("GenABEL")  
  
Installing package into 'C:/Users/Bert/Documents/R/win-library/3.3'  
(as 'lib' is unspecified)  
trying URL 'https://mirrors.tuna.tsinghua.edu.cn/CRAN/bin/windows/contrib/3.3/  
GenABEL_1.8-0.zip'  
Content type 'application/zip' length 3643450 bytes (3.5 MB)  
downloaded 3.5 MB  
  
package 'GenABEL' successfully unpacked and MD5 sums checked  
  
The downloaded binary packages are in  
  C:/Users/Bert/AppData/Local/Temp/Rtmp4udLAh/downloaded_packages  
  
> library(GenABEL)  
  
Loading required package: MASS  
Loading required package: GenABEL.data  
  
> ls("package:GenABEL")  
[1] "add.phdata"                      "add.plot"  
[3] "annotation"                      "arrange_probabel_phe"  
[5] "as.character.gwaa.data"          "as.character.snp.coding"  
[7] "as.character.snp.data"           "as.character.snp.strand"  
[9] "as.data.frame.gwaa.data"         "as.double.gwaa.data"  
[11] "as.double.snp.data"             "as.genotype"  
[13] "as.genotype.gwaa.data"          "as.genotype.snp.data"  
[15] "as.hsgeno"                      "as.hsgeno.gwaa.data"  
[17] "as.hsgeno.snp.data"             "autosomal"  
[19] "blurGenotype"                   "catable"  
[21] "ccfast"                         "check.marker"  
[23] "check.trait"                    "checkPackageVersionOnCRAN"  
[25] "chi2(CG)"  
[27] "cocohet"  
[29] "coding<-"  
[31] "convert.snp.illumina"          "convert.snp.mach"  
[33] "convert.snp.ped"                "convert.snp.text"  
[35] "convert.snp.tped"               "crnames"  
[37] "del.phdata"                     "descriptives.marker"  
[39] "descriptives.scan"              "descriptives.trait"  
[41] "dprfast"                        "effallele"  
[43] "egscore"                         "egscore.old"  
[45] "emp.ccfast"                     "emp.qtscore"  
[47] "estlambda"                      "export.impute"  
[49] "export.merlin"                   "export.plink"  
[51] "extract.annotation.impute"      "extract.annotation.mach"  
[53] "findRelatives"                  "formetascore"
```

```

[55] "GASurv"                               "generateOffspring"
[57] "getcall"                               "getfamily"
[59] "getLogLikelihoodGivenRelation" "grammar"
[61] "gtdata"                                "hom"
[63] "hom.old"                               "HWE.show"
[65] "ibs"                                    "ibs.old"
[67] "idnames"                               "impute2databel"
[69] "impute2mach"                            "lambda"
[71] "load.gwaa.data"                         "mach2databel"
[73] "makeTransitionMatrix" "male"
[75] "map"                                    "merge.gwaa.data"
[77] "merge.snp.data"                         "mlreg"
[79] "mlreg.p"                                "mmscore"
[81] "nids"                                    "npsubtreated"
[83] "nsnps"                                   "patch_strand"
[85] "perid.summary"                           "PGC"
[87] "phdata"                                 "phdata<-"
[89] "plot.check.marker"                      "plot.scan.gwaa"
[91] "plot.scan.gwaa.2D"                      "polygenic"
[93] "polygenic_hglm"                          "qtscore"
[95] "qvaluebh95"                            "r2fast"
[97] "r2fast.old"                            "recodeChromosome"
[99] "reconstructNPs"                          "redundant"
[101] "refallele"                             "refresh.gwaa.data"
[103] "reg.gwaa"                               "results"
[105] "rhofast"                                "rntransform"
[107] "save.gwaa.data"                          "scan.glm"
[109] "scan.glm.2D"                            "scan.haplo"
[111] "scan.haplo.2D"                          "show"
[113] "show.ncbi"                               "snp.data"
[115] "snp.names"                               "snp.subset"
[117] "snpnames"                               "sortmap.internal"
[119] "sset"                                    "strand"
[121] "strand<-"                             "summary.check.marker"
[123] "summary.snp.data"                      "summary.snp.data_old"
[125] "VIFGC"                                  "VIFGC_ovdom"
[127] "Xfix"                                   "ztransform"

```

add.plot Function to Plot Additional GWAA Results

Description Add plot of results of GWA analysis

Usage

```
add.plot(x, ..., df = 1, col=c("lightgreen", "lightblue"), sort=TRUE, delta = 1)
```

Arguments

- x** object of type scan.gwaa-class, as returned by scan.glm, qtscore, ccfast, emp.ccfast, emp.qtscore, or scan.haplo; or of type scan.gwaa.2D-class, as returned by scan.haplo.2D or scan.glm.2D.
... additional arguments to be passed to plot
df P-value at which df to add (1, 2, or "Pc1df")
col which colors to use to depict consecutive chromosomes sort whether results should be plotted after sorting by chromosome and position
delta gap width between chromosomes
Value No value returned.
Author Yurii Aulchenko

See Also

plot,.snp.subset, scan.glm, qtscore, ccfast, emp.qtscore, emp.ccfast, scan.haplo, scan.haplo.2D, scan.glm.2D

Examples

```

require(GenABEL.data)
data(srdta)
a <- ccfast("bt",srdta,snps=c(1:100))
plot(a)
a1 <- qtscore(bt,srdta,snps=c(1:100))
add.plot(a1,col="red",type="l")

```

In the R domain:

```

> require(GenABEL.data)
> data(srdta)
> a <- ccfast("bt",srdta,snps=c(1:100))
Warning in ccfast("bt", srdta, snps = c(1:100)) :
  11 people (out of 2500 ) excluded as not having cc status

> plot(a)
> # Outputting: Figure 5.22
>
> a1 <- qtscore(bt,srdta,snps=c(1:100))
Warning messages:
1: In test.type(y, trait.type) : binomial trait is analysed as gaussian
2: In qtscore(bt, srdta, snps = c(1:100)) :
  11 observations deleted due to missingness
3: In qtscore(bt, srdta, snps = c(1:100)) : Lambda estimated < 1, set to 1
>
> add.plot(a1,col="red",type="l")
> # Outputting: Figure 5.23
>

> add.plot(a1,col="red",type="l")
> # Outputting: Figure 5.23
>

```

Fig. 5.22 `a <- ccfast("bt", srdta,snps=c(1:100)) > plot(a)`

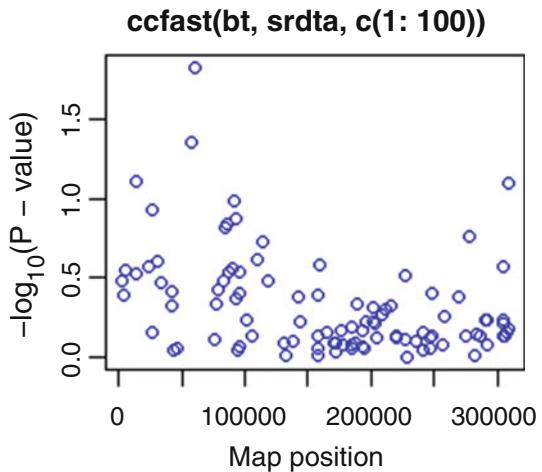
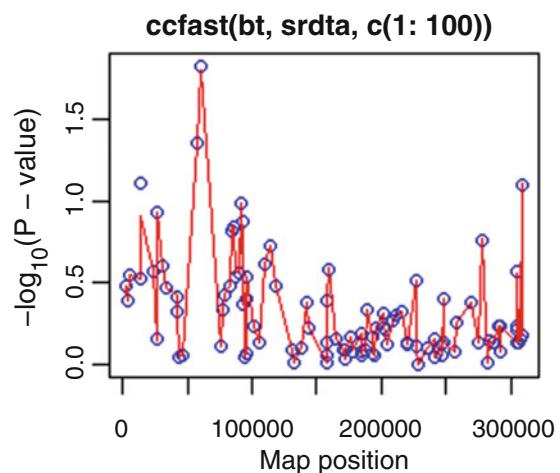


Fig. 5.23 `> add.plot(a1, col="red",type="l")`



5.3.1.3 Package HardyWeinberg

December 16, 2017

Type	Package
Title	Statistical Tests and Graphics for Hardy-Weinberg Equilibrium
Version	1.5.9
Date	2017-12-15
Authors	Jan Graffelman [aut, cre], Christopher Chang [ctb], Xavi Puig [ctb], Jan Wigginton [ctb]
Maintainer	Jan Graffelman jan.graffelman@upc.edu
Depends	R (>= 1.8.0), mice, Rsolnp

Description	Contains tools for exploring Hardy-Weinberg equilibrium (Hardy 1908; Weinberg 1908) for bi and multi-allelic genetic marker data. All classical tests (chi-square, exact, likelihood-ratio and permutation tests) with bi-allelic variants are included in the package, as well as functions for power computation and for the simulation of marker data under equilibrium and disequilibrium. Routines for dealing with markers on the Xchromosome are included, including Bayesian procedures. Some exact and permutation procedures also work with multi-allelic variants. Special test procedures that jointly address Hardy-Weinberg equilibrium and equality of allele frequencies in both sexes are supplied, for the bi- and multi-allelic case. Functions for testing equilibrium in the presence of missing data by using multiple imputation are also provided. Implements several graphics for exploring the equilibrium status of a large set of bi-allelic markers: ternary plots with acceptance regions, log-ratio plots and Q-Q plots.
License	GPL (>= 2)
URL	https://www.r-project.org , http://www-eio.upc.edu/~jan
LinkingTo	Rcpp
Imports	Rcpp
NeedsCompilation	yes
Repository	CRAN
Date/Publication	2017-12-16 22:37:44 UTC

Example 1:

HardyWeinberg-package

Graphical Tests for Hardy-Weinberg Equilibrium

Description

The package HardyWeinberg offers tools for exploring diallelic genetic marker data. It offers all classical tests (chi-square, exact, likelihood-ratio and permutation tests) for Hardy-Weinberg equilibrium, functions for power computation and for the simulation of marker data under equilibrium and disequilibrium. Functions for testing equilibrium in the presence of missing data by using multiple imputation are provided. The package also supplies various graphical tools such as ternary plots with acceptance regions, log-ratio plots and Q-Q plots for exploring the equilibrium status of a large set of diallelic markers.

Details

Package: HardyWeinberg

Type: Package

Version: 1.5.9

Date: 2017-17-08

License: GPL Version 2 or later.

The most important function of the package is HWTernaryPlot that can be used to create ternary plots with acceptance regions for HWE. Other routines implement statistical tests for HWE such as HWChisq and HWLratio

Author Jan Graffelman

Maintainer: Jan Graffelman jan.graffelman@upc.edu

References

- Weir, B.S. (1996) Genetic Data Analysis II. Sinauer Associates, Massachusetts.
- Graffelman, J. and Morales, J. (2008) Graphical tests for Hardy-Weinberg equilibrium based on the ternary plot. *Human Heredity* 65(2):77-84. <https://doi.org/10.1159/000108939>.
- Graffelman, J. (2015). Exploring Diallelic Genetic Markers: The HardyWeinberg Package. *Journal of Statistical Software* 64(3): 1-23. <http://www.jstatsoft.org/v64/i03/>.

```
library(HardyWeinberg)
# draw random SNPs from a population that is in HWE
set.seed(123)
m <- 100 # number of markers
n <- 100 # sample size
X <- HWData(n,m)
out <- HWTernaryPlot(X,100,region=1,vertex.cex=2,signifcolour=TRUE)
```

In the R domain:

```
> install.packages("HardyWeinberg")
Installing package into 'C:/Users/Bert/Documents/R/win-library/3.3'
(as 'lib' is unspecified)
--- Please select a CRAN mirror for use in this session ---
A CRAN mirror is selected.
trying URL 'https://mirrors.tuna.tsinghua.edu.cn/CRAN/bin/windows/contrib/3.3/
HardyWeinberg_1.5.9.zip'
Content type 'application/zip' length 1338952 bytes (1.3 MB)
downloaded 1.3 MB

package 'HardyWeinberg' successfully unpacked and MD5 sums checked
```

The downloaded binary packages are in

```
C:\Users\Bert\AppData\Local\Temp\RtmpuCrHvZ\downloaded_packages
```

```
> library(HardyWeinberg)
Loading required package: mice
Loading required package: Rsolnp
> ls("package:HardyWeinberg")
[1] "af"                      "AFtest"                  "agcounts"
[4] "alleles"                 "dgraffelmanweir.bi" "dlevene"
[7] "dlevene.bi"              "EAFFExact"              "fisherz"
[10] "GenerateSamples" "genlabels"                "HWABO"
[13] "HWAIC"                  "HWAlltests"              "HWAlr"
[16] "HWAlrPlot"               "HWChisq"                 "HWChisqMat"
[19] "HWChisqStats"            "HWClo"                   "HWClr"
[22] "HWClrPlot"               "HWCondProbAB"            "HWD"
[25] "HWData"                  "HWExact"                 "HWExactMat"
[28] "HWExactPrevious" "HWExactStats"              "HWF"
[31] "HWGenotypePlot"           "HWI1r"                   "HWI1rPlot"
[34] "HWLRAllTests"             "HWLratio"                "HWLRtest"
```

```

[37] "HWMissing"           "HWPerm"                  "HWPerm.mult"
[40] "HWPosterior"          "HWPower"                 "HWQqplot"
[43] "HWTernaryPlot"         "HWTriExact"               "ifisherz"
[46] "mac"                   "maf"                     "MakeCounts"
[49] "MakeFactor"            "n.alleles"                "recode"
[52] "strsort"                "ThetatoF"                 "toTriangular"
[55] "UniqueGenotypeCounts"  "vaf"

>
> # draw random SNPs from a population that is in HWE
> set.seed(123)
> m <- 100 # number of markers
> n <- 100 # sample size
> X <- HWData(n,m)
> out <- HWTernaryPlot(X,100,region=1,vertex.cex=2,signifcolour=TRUE)
> # Outputting: Figure 5.24

```

Example 2: HWClrPlot

Plot genetic markers in centered log-ratio coordinates

Description

HWClrPlot creates a scatter plot of the centred log-ratio coordinates of bi-allelic genetic markers. Hardy-Weinberg equilibrium is indicated by a straight line in the plot.

Usage

```
HWClrPlot(X, zeroadj = 0.5)
```

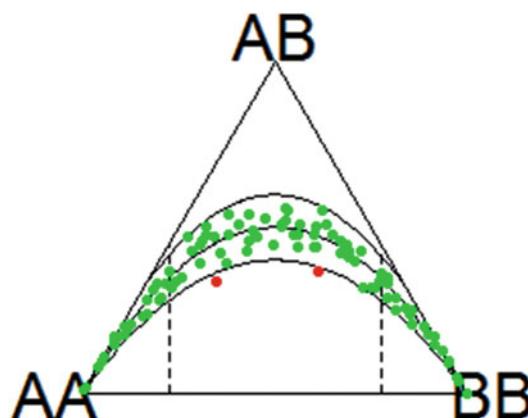


Fig. 5.24 HWTernaryPlot

Arguments

X A matrix of genotype counts (columns AA, AB, BB)
 zeroadj Zero-adjustment parameter. Zero counts in the count matrix are substituted by zeroadj which is 0.5 by default.

Value

NULL

Author Jan Graffelman (jan.graffelman@upc.edu)

References

Graffelman, J., and Egozcue, J. J. (2011).- Hardy-Weinberg equilibrium: a non-parametric compositional approach. In: Vera Pawlowsky-Glahn and Antonella Buccianti (eds.) Compositional Data Analysis: Theory and Applications, John Wiley & Sons, Ltd, pp. 207-215

See Also

HWAlrPlot, HWIrrPlot

Example

```
X <- HWClo(HWData(100,100))
```

```
HWClrPlot(X)
```

In the R domain:

```
>
> X <- HWClo(HWData(100,100))
> HWClrPlot(X)
NULL
> # Outputting: Figure 5.25
```

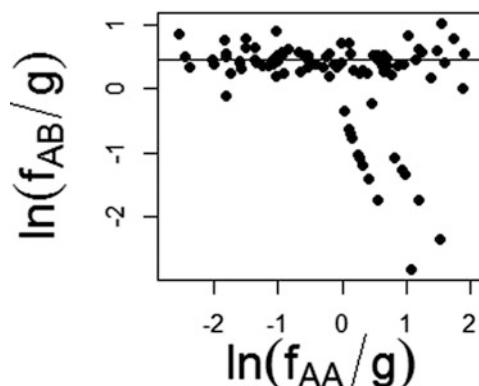
Clr-transformation

Fig. 5.25 Plot of genetic markers in centered log-ratio coordinates

5.3.1.4 Package HardyWeinberg

HWGenotypePlot Scatter Plot of the Genotype Frequencies

Description

HWGenotypePlot makes a scatterplots of the AB or BB frequency versus the AA frequency and represents a blue curve indicating the Hardy-Weinberg equilibrium condition.

Usage

```
HWGenotypePlot(X, plottype = 1, xlab = expression(f[AA]),
               ylab = ifelse(plottype == 1, expression(f[AB]),
                             expression(f[BB])), asp = 1, pch = 19, xlim = c(0, 1),
               ylim = c(0, 1), cex = 1, cex.axis = 2, cex.lab = 2, ...)
```

Arguments

X	A matrix of genotype counts or frequencies with three columns (AA, AB, BB)
plottype	plottype=1 produces a plot of AB versus AA, plottype=2 produced a plot of BB versus AA.
xlab	A label for the x axis
ylab	A label for the y axis
asp	Aspect ratio (1 by default)
pch	Plotting character (19 by default)
xlim	Limits for the x axis (0-1 by default)
ylim	Limits for the y axis (0-1 by default)
cex	Character expansion factor (1 by default)
cex.axis	Character expansion factor for the axes (2 by default)
cex.lab	Character expansion factor for labels of axis (2 by default)
...	Additional arguments for the plot function
Value	NULL
Author	Jan Graffelman <jan.graffelman@upc.edu>
See	HWTernaryPlot
Also	

Examples

```
n <- 100 # sample size
m <- 100 # number of markers
Xc <- HWClo(HWData(n,m))
HWGenotypePlot(Xc,plottype=1,main="Heterozygote-homozygote
scatterplot")
```

In the R domain:

```

>
> n <- 100 # sample size
> m <- 100 # number of markers
> Xc <- HWClo(HWData(n,m))
>
> HWGenotypePlot(Xc,plottype=1,main="Heterozygote-homozygote
+                               scatterplot")
>
NULL
>
> # Outputting: Figure 5.26

```

Challeneges for Unobservable Phase^[F]

[F] Foulkes, A. S., Yucel, R., and Reilly, M. P. <https://www.ncbi.nlm.nih.gov/por> and Foulkes, A. S. (2009).- “Applied Statistical Genetics with R, For Population-based Association Studies”, Springer “Use R! Series, Springer, New York, NY

In population-based genetic investigations of unrelated individuals, a fundamental analytic challenge is the *unobservable nature of allelic phase*. The biostatistical challenges and analytic techniques for characterizing haplotype associations rest in the context of unknown phase. The *haplotypic phase* refers to the specific alignment of alleles on a single homologous chromosome and is generally *not* observable in the context of population-based investigations of unrelated individuals. As the SNPs under study are often markers for the true disease-causing variant, haplotypes may include more

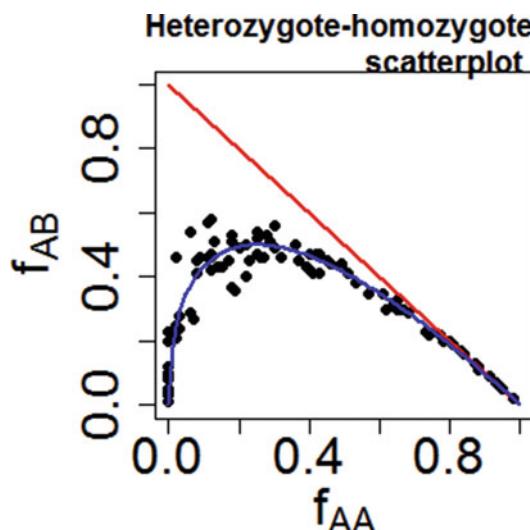


Fig. 5.26 Scatterplot of genotype frequencies: HWGenotypePlot

variability in the disease trait than genotype alone. Some biostatistical approaches to inferring haplotypic phase have been suggested. In general, the goals of these methods are generally twofold:

1. On the one hand, interest lies in estimating populationlevel haplotype frequencies; that is, the prevalence of specific haplotypes in the general population.
2. Investigators are also interested in making inference about the association between haplotypes and a trait.

Both goals will be addressed: the First deals with methods for estimation of haplotype frequencies that do not involve knowledge about a trait or disease. phenotype; and the second focuses on methods that involve both estimation of

haplotype frequencies and testing for association between these haplotypes and a measured trait.

5.3.2 Worked Examples of Statistical Tests and Utilities for Genetic Association

In the analysis of genetic associations with disease in the general population, several models should be considered. To distinguish among three types of models:

1. the genetic model for locus (haplotype or single SNP) effects,
2. the model for multi-locus association, and
3. the population genetic model (often called the coalescence process).

*The **genetic model** refers to how two genotype or haplotype copies (one on each chromosome) act in concert. Commonly described genetic models include additive, recessive, or dominant. In the simple case of a single SNP, the possible genotypes are *AA*, *Aa*, and *aa*. If the effects on phenotype of carrying *A* or *a* are denoted $e(A)$ and $e(a)$, respectively, then the additive model assumes that the effect of *AA* is

$$e(AA) = 2^* e(A) \quad (5.3)$$

while the effect of the *Aa* genotype is

$$e(Aa) = e(A) + e(a) \quad (5.4)$$

A **dominant** model assumes that the effect of *Aa* is the same as that of *AA*, while a **recessive** model assumes that an effect is observed only in the presence of two *a* alleles.

Moreover, the **model for multi-locus association** relates to how multiple SNPs or genes interact with one another in explaining the variability in phenotype, regardless of the placement of alleles on chromosomes. Models for association are typically additive or multiplicative, although alternative formulations have also been described. For example, suppose one observes *AA*, *Aa*, or *aa* at one position and *BB*, *Bb*, or *bb* at another position. An additive model for association assumes that the phenotypic effect of presenting the *AA* and *BB* genotypes is the **sum** of $e(AA)$ and $e(BB)$, while a multiplicative model for association would assume that this is given by the **product** of $e(AA)$ and $e(BB)$.

Finally, the **population genetic model** refers to the process by which SNPs are inherited in combination over generations. Measures of LD between two SNPs are highly dependent on the coalescence model assumption. While there remains some uncertainty as to which population genetic

models are most appropriate, a neutral coalescent infinite-many-site model with recombination has been used commonly because of its ability to capture efficiently the complicated genealogic dependence structure. Applicable software may be found at: <http://home.uchicago.edu/rhudson1/source/mksamples.html>.

This model assumes constant population size and does not account for variation in regional rates of recombination and mutation and the occurrence of gene conversions and multiple mutations as has been observed in actual human genetic data. The population genetic model is useful for estimating posterior haplotype probabilities, which requires one to impute unobserved haplotype information. While one does not focus additional attention on the selection of these models, the importance of considering the genetic assumptions regarding the population will identifying a method for estimating posterior haplotype probabilities.

In the original formulation of mixed modeling for genotype–phenotype association data [22], a subject who is heterozygous at two positions, so that the observed genotype is Aa for SNP 1 and Bb for SNP 2, would be assigned to the genotype group defined by (Aa, Bb) . However, the true haplotype pair (diplotype) for this individual could be (1) (AB, ab) : A and B are on the same chromosome, and a and b are on the same chromosome, or (2) (Ab, aB) : A and b are on the same chromosome and a and B are on the same chromosome. While this additional layer of information is usually unobserved, the probabilities that each haplotype pair is the true haplotype pair can be estimated, as described in [3].

Haplotype pairs (in this simple example, a set of two of AB , Ab , aB , or ab) can be regarded as clusters, rendering the mixed-effects model a natural framework for analysis. While a small number of clusters can render instability in model fitting, in general, the number of potential haplotypes under consideration within a gene (or combinations of haplotypes across genes as discussed below) is large. In some instances this cluster membership is fully observable. For example, if a subject is homozygous for all SNPs within a gene, then the haplotypes for that individual are known deterministically. However, if an individual is heterozygous for two or more SNPs within a gene, then the true haplotypes are not observable. In the following sections we present a multiple imputation approach to incorporating this uncertainty in phase. Notably, this approach is applicable to the analysis of haplotype–phenotype associations in settings in which we observe multiple SNPs within a single gene, as well as multiple SNPs across many genes.

5.3.2.1 Some Worked Examples from CRAN

Example 5.3

CRAN Package haploR

November 1, 2017

Type	Package
Title	Query 'HaploReg' and 'RegulomeDB'
Version	1.6.2
Date	2017-10-31
Maintainer	Ilya Y. Zhbannikov <ilya.zhbannikov@duke.edu>
Description	A set of utilities for querying 'HaploReg' http://archive.broadinstitute.org/mammals/haploreg/haploreg.php and 'RegulomeDB' < http://www.regulomedb.org > web-based tools. The package connects to 'HaploReg' or 'RegulomeDB', searches and downloads results, without opening web pages, directly from R environment. Results are stored in a data frame that can be directly used in various kinds of downstream analyses.
RoxygenNote	6.0.1
Suggests	knitr, rmarkdown, openxlsx

VignetteBuilder	knitr
Imports	httr, XML, tibble, RUnit, plyr
Depends	R (>= 3.3.2)
Encoding	UTF-8
License	GPL-3
NeedsCompilation	no
Authors	Ilya Y. Zhbannikov [aut, cre], Anatoliy I Yashin [ctb] (Anatoliy supervised the technical/software development and is a Principal Investigator of this project)
Repository	CRAN
Date/Publication	2017-10-31 23:30:02 UTC

R topics documented:

- (1) getExtendedView
- (2) getStudyList
- (3) queryHaploreg
- (4) queryRegulome

- (1) getExtendedView

This function queries HaploReg web-based tool in order to Extended view for SNP of interest

Description

This function queries HaploReg web-based tool in order to Extended view for SNP of interest

Usage

```
getExtendedView(snp, url = "http://archive.broadinstitute.org/
mammals/haploreg/detail_v4.1.php?query=&id=")
```

Arguments

- snp A SNP of interest.
- url A url to HaploReg. Default: <"http://archive.broadinstitute.org/mammals/haploreg/detail_v4.1.php?query=&Value

Value

A list of tables t1, t2, ..., etc depending on information contained in HaploReg database.

Examples

```
tables <- getExtendedView(snp="rs10048158")
tables
```

In the R domain :

```
> install.packages("haploR")
Installing package into 'C:/Users/Bert/Documents/R/win-library/3.3'
(as 'lib' is unspecified)
--- Please select a CRAN mirror for use in this session ---
```

A CRAN mirror is selected.

```
trying URL 'https://mirrors.tuna.tsinghua.edu.cn/CRAN/bin/windows/contrib/3.3/
haploR_1.6.2.zip'
Content type 'application/zip' length 320129 bytes (312 KB)
downloaded 312 KB
```

```
package 'haploR' successfully unpacked and MD5 sums checked
```

The downloaded binary packages are in

```
C:\Users\Bert\AppData\Local\Temp\RtmpYD6v2m\downloaded_packages
```

```
> library(haploR)
```

```
> ls("package:haploR") # Outputting:
```

```
[1] "getExtendedView"    "getStudyList"    "queryHaploreg"    "queryRegulome"
```

```
> getExtendedView # Outputting:
```

```
function (snp, url =
```

```
"http://archive.broadinstitute.org/mammals/haploreg/detail_v4.1.php?query=&id=")
{
  ext.url <- paste(url, snp, sep = " ")
  page <- htmlParse(ext.url)
  tables <- readHTMLTable(page)
  tables[[1]] <- tables[[1]][-1, ]
  colnames(tables[[1]]) <- c("chr", "pos hg19", "chr2", "pos (hg38)",
    "Reference", "Alternate", "AFR", "AMR", "ASN", "EUR",
    "by GERP", "by SiPhy", "dbSNP functional annotation")
  if (length(tables) >= 2) {
    colnames(tables[[2]]) <- as.character(unlist(c(tables[[2]][1,
      ])))
    tables[[2]] <- tables[[2]][-1, ]
  }
  table.names <- paste("t", 1:length(tables), sep = " ")
  names(tables) <- table.names
  return(tables)
}
```

<environment: namespace:haploR>

```

>
> getExtendedView
function (snp, url = "http://archive.broadinstitute.org/mammals/haploreg/detail_v4.1.php?query=&id=")
{
  ext.url <- paste(url, snp, sep = "")
  page <- htmlParse(ext.url)
  tables <- readHTMLTable(page)
  tables[[1]] <- tables[[1]][-1, ]
  colnames(tables[[1]]) <- c("chr", "pos hg19", "chr2", "pos (hg38)",
    "Reference", "Alternate", "AFR", "AMR", "ASN", "EUR",
    "by GERP", "by SiPhy", "dbSNP functional annotation")
  if (length(tables) >= 2) {
    colnames(tables[[2]]) <- as.character(unlist(c(tables[[2]][1,
      ])))
    tables[[2]] <- tables[[2]][-1, ]
  }
  table.names <- paste("t", 1:length(tables), sep = "")
  names(tables) <- table.names
  return(tables)
}
<environment: namespace:haploR>
>
> tables <- getExtendedView(snp="rs10048158")
> tables # Outputting:
$t1
  chr pos hg19  chr2 pos (hg38) Reference Alternate AFR AMR ASN EUR
2 chr17 64236318 chr17 66240200          T          C 0.93 0.66 0.92 0.56
  by GERP by SiPhy dbSNP functional annotation
2     No      No           none

$t2
  Source Distance  Direction           ID/Link Common name
2 GENCODE      NA Within gene ENSG00000091583.6          APOH
3 RefSeq       5'      10761           NM_000042          APOH
                                         Description
2 apolipoprotein H (beta-2-glycoprotein I) [Source:HGNC Symbol;Acc:616]
3 apolipoprotein H (beta-2-glycoprotein I) [Source:HGNC Symbol;Acc:616]

$t3
  Epigenome ID (EID)      Group      Mnemonic
1            E017    IMR90    LNG.IMR90
2            E002      ESC      ESC.WA7
3            E008      ESC      ESC.H9
4            E001      ESC      ESC.I3
5            E015      ESC      ESC.HUES6
6            E014      ESC      ESC.HUES48
7            E016      ESC      ESC.HUES64
8            E003      ESC      ESC.H1
9            E024      ESC    ESC.4STAR

```

10	E020	iPSC	IPSC.20B
11	E019	iPSC	IPSC.18
12	E018	iPSC	IPSC.15b
13	E021	iPSC	IPSC.DF.6.9
14	E022	iPSC	IPSC.DF.19.11
15	E007	ES-deriv	ESDR.H1.NEUR.PROG
16	E009	ES-deriv	ESDR.H9.NEUR.PROG
17	E010	ES-deriv	ESDR.H9.NEUR
18	E013	ES-deriv	ESDR.CD56.MESO
19	E012	ES-deriv	ESDR.CD56.ECTO
20	E011	ES-deriv	ESDR.CD184.ENDO
21	E004	ES-deriv	ESDR.H1.BMP4.MESO
22	E005	ES-deriv	ESDR.H1.BMP4.TROP
23	E006	ES-deriv	ESDR.H1.MSC
24	E062	Blood & T-cell	BLD.PER.MONUC.PC
25	E034	Blood & T-cell	BLD.CD3.PPC
26	E045	Blood & T-cell	BLD.CD4.CD25I.CD127.TMEMPC
27	E033	Blood & T-cell	BLD.CD3.CPC
28	E044	Blood & T-cell	BLD.CD4.CD25.CD127M.TREGPC
29	E043	Blood & T-cell	BLD.CD4.CD25M.TPC
30	E039	Blood & T-cell	BLD.CD4.CD25M.CD45RA.NPC
31	E041	Blood & T-cell	BLD.CD4.CD25M.IL17M.PL.TPC
32	E042	Blood & T-cell	BLD.CD4.CD25M.IL17P.PL.TPC
33	E040	Blood & T-cell	BLD.CD4.CD25M.CD45RO.MPC
34	E037	Blood & T-cell	BLD.CD4.MPC
35	E048	Blood & T-cell	BLD.CD8.MPC
36	E038	Blood & T-cell	BLD.CD4.NPC
37	E047	Blood & T-cell	BLD.CD8.NPC
38	E029	HSC & B-cell	BLD.CD14.PC
39	E031	HSC & B-cell	BLD.CD19.CPC
40	E035	HSC & B-cell	BLD.CD34.PC
41	E051	HSC & B-cell	BLD.MOB.CD34.PC.M
42	E050	HSC & B-cell	BLD.MOB.CD34.PC.F
43	E036	HSC & B-cell	BLD.CD34.CC
44	E032	HSC & B-cell	BLD.CD19.PPC
45	E046	HSC & B-cell	BLD.CD56.PC
46	E030	HSC & B-cell	BLD.CD15.PC
47	E026	Mesench	STRM.MRW.MSC
48	E049	Mesench	STRM.CHON.MRW.DR.MSC
49	E025	Mesench	FAT.ADIP.DR.MSC
50	E023	Mesench	FAT.MSC.DR.ADIP
51	E052	Myosat	MUS.SAT
52	E055	Epithelial	SKIN.PEN.FRSK.FIB.01
53	E056	Epithelial	SKIN.PEN.FRSK.FIB.02
54	E059	Epithelial	SKIN.PEN.FRSK.MEL.01
55	E061	Epithelial	SKIN.PEN.FRSK.MEL.03
56	E057	Epithelial	SKIN.PEN.FRSK.KER.02
57	E058	Epithelial	SKIN.PEN.FRSK.KER.03
58	E028	Epithelial	BRST.HMEC.35
59	E027	Epithelial	BRST.MYO

60	E054	Neurospoh	BRN.GANGEM.DR.NRSPHR
61	E053	Neurospoh	BRN.CRTX.DR.NRSPHR
62	E112	Thymus	THYM
63	E093	Thymus	THYM.FET
64	E071	Brain	BRN.HIPP.MID
65	E074	Brain	BRN.SUB.NIG
66	E068	Brain	BRN.ANT.CAUD
67	E069	Brain	BRN.CING.GYR
68	E072	Brain	BRN.INF.TMP
69	E067	Brain	BRN.ANG.GYR
70	E073	Brain	BRN.DL.PRFRNLT.CRTX
71	E070	Brain	BRN.GRM.MTRX
72	E082	Brain	BRN.FET.F
73	E081	Brain	BRN.FET.M
74	E063	Adipose	FAT.ADIP.NUC
75	E100	Muscle	MUS.PSOAS
76	E108	Muscle	MUS.SKLT.F
77	E107	Muscle	MUS.SKLT.M
78	E089	Muscle	MUS.TRNK.FET
79	E090	Muscle	MUS.LEG.FET
80	E083	Heart	HRT.FET
81	E104	Heart	HRT.ATR.R
82	E095	Heart	HRT.VENT.L
83	E105	Heart	HRT.VNT.R
84	E065	Heart	VAS.AOR
85	E078	Sm. Muscle	GI.DUO.SM.MUS
86	E076	Sm. Muscle	GI.CLN.SM.MUS
87	E103	Sm. Muscle	GI.RECT.SM.MUS
88	E111	Sm. Muscle	GI.STMC.MUS
89	E092	Digestive	GI.STMC.FET
90	E085	Digestive	GI.S.INT.FET
91	E084	Digestive	GI.L.INT.FET
92	E109	Digestive	GI.S.INT
93	E106	Digestive	GI.CLN.SIG
94	E075	Digestive	GI.CLN.MUC
95	E101	Digestive	GI.RECT.MUC.29
96	E102	Digestive	GI.RECT.MUC.31
97	E110	Digestive	GI.STMC.MUC
98	E077	Digestive	GI.DUO.MUC
99	E079	Digestive	GI.ESO
100	E094	Digestive	GI.STMC.GAST
101	E099	Other	PLCNT.AMN
102	E086	Other	KID.FET
103	E088	Other	LNG.FET
104	E097	Other	OVRY
105	E087	Other	PANC.ISLT
106	E080	Other	ADRL.GLND.FET
107	E091	Other	PLCNT.FET
108	E066	Other	LIV.ADLT
109	E098	Other	PANC

110	E096	Other	LNG	
111	E113	Other	SPLN	
112	E114	ENCODE2012	LNG.A549.ETOHO02.CNCR	
113	E115	ENCODE2012	BLD.DND41.CNCR	
114	E116	ENCODE2012	BLD.GM12878	
115	E117	ENCODE2012	CRVX.HELAS3.CNCR	
116	E118	ENCODE2012	LIV.HEPG2.CNCR	
117	E119	ENCODE2012	BRST.HMEC	
118	E120	ENCODE2012	MUS.HSMM	
119	E121	ENCODE2012	MUS.HSMMT	
120	E122	ENCODE2012	VAS.HUVEC	
121	E123	ENCODE2012	BLD.K562.CNCR	
122	E124	ENCODE2012	BLD.CD14.MONO	
123	E125	ENCODE2012	BRN.NHA	
124	E126	ENCODE2012	SKIN.NHDFAD	
125	E127	ENCODE2012	SKIN.NHEK	
126	E128	ENCODE2012	LNG.NHLF	
127	E129	ENCODE2012	BONE.OSTEO	
			Description	
1		IMR90 fetal lung fibroblasts	Cell Line	
2			ES-WA7 Cells	
3			H9 Cells	
4			ES-I3 Cells	
5			HUES6 Cells	
6			HUES48 Cells	
7			HUES64 Cells	
8			H1 Cells	
9			ES-UCSF4 Cells	
10			iPS-20b Cells	
11			iPS-18 Cells	
12			iPS-15b Cells	
13			iPS DF 6.9 Cells	
14			iPS DF 19.11 Cells	
15		H1 Derived Neuronal Progenitor	Cultured Cells	
16		H9 Derived Neuronal Progenitor	Cultured Cells	
17			H9 Derived Neuron	Cultured Cells
18			hESC Derived CD56+ Mesoderm	Cultured Cells
19			hESC Derived CD56+ Ectoderm	Cultured Cells
20			hESC Derived CD184+ Endoderm	Cultured Cells
21			H1 B.P. Derived Mesendoderm	Cultured Cells
22			H1 B.P. Derived Trophoblast	Cultured Cells
23			H1 Derived Mesenchymal Stem Cells	
24		Primary mononuclear cells	from peripheral blood	
25			Primary T cells	from peripheral blood
26	Primary T cells	effector/memory enriched	from peripheral blood	
27			Primary T cells	from cord blood
28		Primary T regulatory cells	from peripheral blood	
29			Primary T helper cells	from peripheral blood
30		Primary T helper naive cells	from peripheral blood	
31			Primary T helper cells	PMA-I stimulated

32 Primary T helper 17 cells PMA-I stimulated
33 Primary T helper memory cells from peripheral blood 1
34 Primary T helper memory cells from peripheral blood 2
35 Primary T CD8+ memory cells from peripheral blood
36 Primary T helper naive cells from peripheral blood
37 Primary T CD8+ naive cells from peripheral blood
38 Primary monocytes from peripheral blood
39 Primary B cells from cord blood
40 Primary hematopoietic stem cells
41 Primary hematopoietic stem cells G-CSF-mobilized Male
42 Primary hematopoietic stem cells G-CSF-mobilized Female
43 Primary hematopoietic stem cells short term culture
44 Primary B cells from peripheral blood
45 Primary Natural Killer cells from peripheral blood
46 Primary neutrophils from peripheral blood
47 Bone Marrow Derived Cultured Mesenchymal Stem Cells
48 Mesenchymal Stem Cell Derived Chondrocyte Cultured Cells
49 Adipose Derived Mesenchymal Stem Cell Cultured Cells
50 Mesenchymal Stem Cell Derived Adipocyte Cultured Cells
51 Muscle Satellite Cultured Cells
52 Foreskin Fibroblast Primary Cells skin01
53 Foreskin Fibroblast Primary Cells skin02
54 Foreskin Melanocyte Primary Cells skin01
55 Foreskin Melanocyte Primary Cells skin03
56 Foreskin Keratinocyte Primary Cells skin02
57 Foreskin Keratinocyte Primary Cells skin03
58 Breast variant Human Mammary Epithelial Cells (vHMEC)
59 Breast Myoepithelial Primary Cells
60 Ganglion Eminence derived primary cultured neurospheres
61 Cortex derived primary cultured neurospheres
62 Thymus
63 Fetal Thymus
64 Brain Hippocampus Middle
65 Brain Substantia Nigra
66 Brain Anterior Caudate
67 Brain Cingulate Gyrus
68 Brain Inferior Temporal Lobe
69 Brain Angular Gyrus
70 Brain_Dorsolateral_Prefrontal_Cortex
71 Brain Germinal Matrix
72 Fetal Brain Female
73 Fetal Brain Male
74 Adipose Nuclei
75 Psoas Muscle
76 Skeletal Muscle Female
77 Skeletal Muscle Male
78 Fetal Muscle Trunk
79 Fetal Muscle Leg
80 Fetal Heart
81 Right Atrium

82 Left Ventricle
83 Right Ventricle
84 Aorta
85 Duodenum Smooth Muscle
86 Colon Smooth Muscle
87 Rectal Smooth Muscle
88 Stomach Smooth Muscle
89 Fetal Stomach
90 Fetal Intestine Small
91 Fetal Intestine Large
92 Small Intestine
93 Sigmoid Colon
94 Colonic Mucosa
95 Rectal Mucosa Donor 29
96 Rectal Mucosa Donor 31
97 Stomach Mucosa
98 Duodenum Mucosa
99 Esophagus
100 Gastric
101 Placenta Amnion
102 Fetal Kidney
103 Fetal Lung
104 Ovary
105 Pancreatic Islets
106 Fetal Adrenal Gland
107 Placenta
108 Liver
109 Pancreas
110 Lung
111 Spleen
112 A549 EtOH 0.02pct Lung Carcinoma Cell Line
113 Dnd41 TCell Leukemia Cell Line
114 GM12878 Lymphoblastoid Cells
115 HeLa-S3 Cervical Carcinoma Cell Line
116 HepG2 Hepatocellular Carcinoma Cell Line
117 HMEC Mammary Epithelial Primary Cells
118 HSMM Skeletal Muscle Myoblasts Cells
119 HSMM cell derived Skeletal Muscle Myotubes Cells
120 HUVEC Umbilical Vein Endothelial Primary Cells
121 K562 Leukemia Cells
122 Monocytes-CD14+ R001746 Primary Cells
123 NH-A Astrocytes Primary Cells
124 NHDF-Ad Adult Dermal Fibroblast Primary Cells
125 NHEK-Epidermal Keratinocyte Primary Cells
126 NHLF Lung Fibroblast Primary Cells
127 Osteoblast Primary Cells

Chromatin states (Core 15-state model)

4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53

54
55
56
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100
101
102
103

```
104
105
106
107
108
109
110
111
112
113
114
115
116
117
118
119
120
121
122
123
124
125
126
127
    Chromatin states(25-state modelusing 12 imputed marks)      H3K4me1
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
```

26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47 H3K4me1_Enh
48
49 H3K4me1_Enh
50
51 H3K4me1_Enh
52 H3K4me1_Enh
53 H3K4me1_Enh
54
55
56 H3K4me1_Enh
57
58 H3K4me1_Enh
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75

76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91 H3K4me1_Enh
92
93
94
95
96
97
98
99
100
101
102
103
104
105
106
107
108
109
110
111
112 H3K4me1_Enh
113
114
115 H3K4me1_Enh
116
117
118
119
120 H3K4me1_Enh
121
122
123 H3K4me1_Enh
124 H3K4me1_Enh
125 H3K4me1_Enh

126
127 H3K4me3 H3K27ac H3K9ac DNase
1 H3K27ac_Enh
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47

```
48
49
50
51          H3K9ac_Pro
52          DNase
53          H3K27ac_Enh      DNase
54
55
56
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
```

```
98
99
100
101
102
103
104
105
106
107
108
109
110
111
112
113
114
115      H3K27ac_Enh
116
117
118
119
120 H3K4me3_Pro H3K27ac_Enh
121
122
123
124      DNase
125
126
127

$t4
                                         Trait p-value      PMID
1 &beta2-Glycoprotein I (&beta2-GPI) plasma levels    1E-6 23279374

$t5
      Study ID
1 GTEx2015_v6
2 GTEx2015_v6
3 Koopman2014
                                         Paper Title
1 The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene
regulation in humans
2 The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene
regulation in humans
3
      PMID          Tissue Correlated gene      p-value
1 25954001 Heart_Atrial_Appendage      PRKCA  7.3104710229491e-07
2 25954001 Heart_Left_Ventricle      PRKCA  3.02712315271318e-13
3                      Heart          PRKCA   3.03E-08
```

```

$t6
Position Weight Matrix ID(Library from Kheradpour and Kellis, 2013) Strand
1                               Mrg1::Hoxa9_1      -
2                               Pou2f2_known3      +
Ref  Alt
1 11.4 -0.6
2 10.4 10.5
Match on:Ref: AAAGGGAATGAGGTTAACAGATGAGAGCTTGATAATAAGATTCTGGAAATGCATTCGAlt:
AAAGGGAATGAGGTTAACAGATGAGAGCCTGATAATAAGATTCTGGAAATGCATTCG
1
TGACAGKTTWAYGR
2
NNNRTAATNRBDD

>

```

(2) getStudyList

This function queries HaploReg web-based tool in order to see a list of GWAS.

Description

This function queries HaploReg web-based tool in order to see a list of GWAS.

Usage

```
getStudyList(url = "http://archive.broadinstitute.org/mammals/haploreg/haploreg.php")
```

Arguments

url A url to HaploReg. Default: <http://archive.broadinstitute.org/mammals/haploreg/haploreg.php>

Value

A list of studies. Each study is itself a list of two: name, id.

Examples

```
studies <- getStudyList()
studies
```

In the R domain :

```
> studies <- getStudyList()

> studies #Outputting:
```

```
'Neuroblastoma (Maris JM, 2008, 1 SNP)'

-----
-----
-----
```

```
[1] "Tanning (Nan H, 2009, 10 SNPs)"

$`Tanning (Nan H, 2009, 10 SNPs)`$id
[1] "1595"

$`YKL-40 levels (Ober C, 2008, 1 SNP)`$name
$`YKL-40 levels (Ober C, 2008, 1 SNP)`$name
[1] "YKL-40 levels (Ober C, 2008, 1 SNP)"

$`YKL-40 levels (Ober C, 2008, 1 SNP)`$id
[1] "2624"
```

>

(3) queryHaploreg

This function queries HaploReg web-based tool and returns results.

Description

This function queries HaploReg web-based tool and returns results.

Usage

```
queryHaploreg(query = NULL, file = NULL, study = NULL, ldThresh = 0.8,
               ldPop = "EUR", epi = "vanilla", cons = "siphy",
               genotypes = "gencode",
               url = "http://archive.broadinstitute.org/
                      mammals/haploreg/haploreg.php",
               timeout = 100, encoding = "UTF-8", querySNP = FALSE,
               fields = NULL, verbose = FALSE)
```

Arguments

query Query (a vector of rsIDs).
 file A text file (one refSNP ID per line).
 study A particular study. See function `getHaploRegStudyList(...)`. Default: NULL.
 ldThresh LD threshold, r2 (select NA to only show query variants). Default: 0.8.
 ldPop 1000G Phase 1 population for LD calculation. Can be: "AFR", "AMR", "ASN". Default: "EUR".

epi	Source for epigenomes. Possible values: vanilla for ChromHMM (Core 15-state model); imputed for ChromHMM (25-state model using 12 imputed marks); methyl for H3K4me1/H3K4me3 peaks; acetyl for H3K27ac/H3K9ac peaks. Default: vanilla.
cons	Mammalian conservation algorithm. Possible values: gerp for GERP, siphy for SiPhy-omega, both for both. Default: siphy.
genotypes	Show position relative to. Possible values: gencode for Gencode genes; refseq for RefSeq genes; both for both. Default: gencode.
url	HaploReg url address. Default: <http://archive.broadinstitute.org/mammals/haploreg/haploreg.php>
timeout	A timeout parameter for curl. Default: 100
encoding	sets the encoding for correct retrieval web-page content. Default:UTF-8
querySNP	A flag indicating to return query SNPs only. Default: FALSE
fields	A set of fields to extract. Refer to the package vignette for available fields. Default: All.
verbose	Verbosing output. Default: FALSE.

Value

A data frame (table) with results similar to HaploReg uses.

Examples

```
library(haploR)
data <- queryHaploreg(c("rs10048158", "rs4791078"))
head(data)
```

In the R domain:

```
>
> library(haploR)
> data <- queryHaploreg(c("rs10048158", "rs4791078"))
> head(data) # Outputting:
# A tibble: 6 x 35
   chr pos_hg38     r2  'D' `is_query_snp`      rsID    ref    alt    AFR    AMR
   ASN     EUR GERP_cons SiPhy_cons
   <fctr> <fctr> <fctr> <fctr>      <fctr>    <fctr> <fctr> <fctr> <fctr> <fctr>
   <fctr> <fctr> <fctr> <fctr>
  1    17 66240200     1     1      1 rs10048158     T     C  0.93  0.66
  0.92  0.56      0     0      0
  2    17 66240200     0.82   0.93      0 rs10048158     T     C  0.93  0.66
  0.92  0.56      0     0      0
  3    17 66231972     0.82   0.99      0 rs11079645     G     T  0.88  0.63
  0.9   0.51      0     0      0
  4    17 66248387     0.99      1      0 rs12603947     T     C  0.93  0.66
  0.92  0.56      0     0      0
  5    17 66248387     0.81   0.92      0 rs12603947     T     C  0.93  0.66
  0.92  0.56      0     0      0
  6    17 66214285     0.98   0.99      0 rs1971682      G     C  0.86  0.67
  0.91  0.57      0     0      0
```

```
# ... with 21 more variables: Chromatin_States <fctr>, Chromatin_States_Imputed
# <fctr>, Chromatin_Marks <fctr>, DNase <fctr>, Proteins <fctr>, eQTL <fctr>,
# gwas <fctr>, grasp <fctr>, Motifs <fctr>, GENCODE_id <fctr>, GENCODE_name
# <fctr>, GENCODE_direction <fctr>, GENCODE_distance <fctr>, RefSeq_id <fctr>,
# RefSeq_name <fctr>, RefSeq_direction <fctr>, RefSeq_distance <fctr>,
dbSNP_functional_annotation <fctr>, query_snp_rsid <fctr>, Promoter_histone-
marks <fctr>,
# Enhancer_histone_marks <fctr>
>
```

(4) queryRegulome

This function queries RegulomeDB www.regulomedb.org web-based tool and returns results in a data frame.

Description

This function queries RegulomeDB www.regulomedb.org web-based tool and returns results in a data frame.

Usage

```
queryRegulome(query = NULL, format = "full",
              url = "http://www.regulomedb.org/results", timeout = 100,
              check_bad_snps = TRUE, verbose = FALSE)
```

Arguments

query Query (a vector of rsIDs).

format An output format. Can be one of the following:

full - plain text,

bed - BED (Browser Extensible Data) format,

see e.g. <<https://genome.ucsc.edu/FAQ/FAQformat.html#format5.1>>,

gff - GFF (General Feature Format),

see e.g. <https://genome.ucsc.edu/FAQ/FAQformat.html#format3> Only full is currently supported.

url Regulome url address. Default: <<http://www.regulomedb.org/results>>

timeout A timeout parameter for curl. Default: 10

check_bad_snps Checks if all SNPs are annotated. Default: TRUE

verbose Verbosing output. Default: FALSE.

Value

A list of two:

1. a data frame (table) and
2. a list of bad SNP IDs. Bad SNP ID are those IDs that were not found in 1000 Genomes Phase 1 data

Examples

```
data <- queryRegulome(c("rs4791078", "rs10048158"))
head(data[["res.table"]])
head(data[["bad.snp.id"]])
```

In the R domain :

```
>
> data <- queryRegulome(c("rs4791078", "rs10048158"))
> head(data[["res.table"]]) # Outputting:
# A tibble: 2 x 5
  '#chromosome' coordinate      rsid
  <fctr>      <fctr>      <fctr>
1     chr17    64236317 rs10048158
2     chr17    64210013 rs4791078
# ... with 2 more variables: hits <fctr>, score <fctr>
> head(data[["bad.snp.id"]]) # Outputting:
# A tibble: 0 x 1
# ... with 1 variables: rsID <fctr>
>
```

5.3.2.2 A Worked Example from the Work of Foulkes – “Applied Statistical Genetics with R, For Population-Based Association Studies” ^[F]

^[F] **Chapter 5 in:** Foulkes, A. S. (2009).- “Applied Statistical Genetics with R, For Population-based Association Studies”, Springer “Use R! Series”, Springer, New York, NY

Example 5.4 (EM approach to haplotype frequency estimation).

In this example, we estimate the population-level frequencies of haplotypes within the *actn3* gene for African Americans and Caucasians separately based on the FAMuSS data. We begin by calling the *haplo.stats* package and creating a genotype matrix: for example, using joint probability approaches in a standard Kaplan-Meier methodology: see, for example, Chan (2016). The genotype matrix has a pair of adjacent columns for each SNP such that each column corresponds to one of the two observed alleles at the corresponding site. The order of the columns is assumed to correspond to the order of the sites on the chromosome. Recall that we start with four SNPs within the *actn3* gene and so the following code is required:

```
> install.packages("haplo.stats")
```

```

> library(haplo.stats) > attach(fms)

> Geno <- cbind(substr(actn3_r577x,1,1), substr(actn3_r577x,2,2), + substr
  (actn3_rs540874,1,1), substr(actn3_rs540874,2,2), + substr
  (actn3_rs1815739,1,1), substr(actn3_rs1815739,2,2), + substr
  (actn3_1671064,1,1), substr(actn3_1671064,2,2))

> SNPnames <- c("actn3_r577x", "actn3_rs540874", "actn3_rs1815739",
  + "actn3_1671064")

```

We then subset African Americans and Caucasians and apply the haplo.em() function to each group. This function applies a modified version of the EM approach described above, in which sets of loci are progressively included and in turn haplotype pairs with small estimated probabilities are excluded. The haplo.em.control() function is used within the haplo.em() function call to specify the minimum posterior probability of a haplotype pair. Pairs that 5.1 Haplotype estimation 133 have an estimated frequency lower than this threshold will be removed from the list of possible pairs.

```

> Geno.C <- Geno[Race=="Caucasian" & !is.na(Race),]

> HaploEM <- haplo.em(Geno.C, locus.label=SNPnames,
  +
  control=haplo.em.control(min.posterior=1e-4))

> HaploEM

# Note that the results may differ slightly each run since different
# starting values are used

=====
Haplotypes
=====

actn3_r577x    actn3_rs540874    actn3_rs1815739    actn3_1671064    hap.freq

1      C          A          C          G  0.00261
2      C          A          T          A  0.00934
3      C          A          T          G  0.01354
4      C          G          C          A  0.47294
5      C          G          C          G  0.01059
6      T          A          C          A  0.00065
7      T          A          T          G  0.39891
8      T          G          C          A  0.08557
9      T          G          T          A  0.00065
10     T          G          T          G  0.00520
=====
```

```

          Details
=====
lnlike = -1285.406
lr stat for no LD = 2780.769 ,      df = 5 ,      p-val = 0

> Geno.AA <- Geno [Race=="African Am" & !is.na(Race) , ]

> HaploEM2 <- haplo.em(Geno.AA, locus.label=SNPnames,
+
+           control=haplo.em.control(min.posterior=1e-4))

> HaploEM2
=====

          Haplotypes
=====

actn3_r577x    actn3_rs540874    actn3_rs1815739    actn3_1671064    hap.freq
1              C                  A                  C                  A  0.01140
2              C                  A                  C                  G  0.08130
3              C                  A                  T                  G  0.03764
4              C                  G                  C                  A  0.57762
5              C                  G                  C                  G  0.01156
6              T                  A                  C                  A  0.00032
7              T                  A                  T                  G  0.17166
8              T                  G                  C                  A  0.10833
9              T                  G                  C                  G  0.00016
=====

          Details
=====

lnlike = -84.97891
lr stat for no LD = 119.7087 ,      df = 4 ,      p-val = 0

```

The column entitled `hap.freq` is the estimated population-level haplotype frequency. Note that the row numbers in the two outputs above do not necessarily correspond to the same haplotypes. Based on this output, we can see that the most prevalent haplotype is the same in African Americans and Caucasians and is given by $h_4 = \text{CGCA}$. The estimated prevalence of this haplotype is higher for African Americans ($\theta_4 = 0.58$) than for Caucasians ($\theta_4 = 0.47$). On the other hand, the estimated prevalence of the $h_7 = \text{TATG}$ haplotype is markedly lower in African Americans ($\theta_7 = 0.17$) than in Caucasians ($\theta_7 = 0.40$).

Based on the estimated population-level haplotype probabilities, we can calculate the probabilities of each possible haplotype pair for an observation in our sample. Consider again the simple example in which an individual presents with the genotype Aa and Bb across two SNPs. This individual's haplotype pair is ambiguous, though we know it is either

$$H_1 = (AB, ab) \text{ or } H_2 = (Ab, aB) \quad (5.5)$$

Now suppose the haplotype frequencies are $\theta_1, \theta_2, \theta_3$, and θ_4 for haplotypes AB, Ab, aB and ab , respectively. Assuming independence, we know that the posterior probability of H_1 is $p_1 = 2\theta_1\theta_4$, while the probability of H_2 is $p_2 = 2\theta_2\theta_3$ given the observed genotype.

If additional haplotype pairs are present in our population, we must additionally divide each of these probabilities by the sum $p_1 + p_2$ since we are conditioning on one of the two haplotype pairs for this individual. A demonstration of how we can calculate these posterior probabilities is given in the following example.

Example 5.5 Calculating posterior haplotype probabilities

In this example, one illustrates how to determine the posterior probability of each haplotype pair that is consistent with the observed genotype for an individual. Let us return to **Example 5.1** and recall that HapoEM is the result of applying the haplo.em() function to the SNPs within the *actn3* gene on the Caucasian subgroup within the FAMuSS study. The associated object HapoEM\$nreps is a vector of length equal to the number of individuals in our sample with elements equal to the number of haplotype pairs that are consistent with the observed genotype. For example, consider the first five elements of this vector, given by

```
> HapoEM$nreps[1:5]

indx.subj
1 2 3 4 5
1 2 2 2 1
```

This tells us that there is one haplotype pair consistent with the observed genotype for the first and fifth individuals and two pairs that are consistent with each of the observed genotypes for the second, third and fourth individuals. The corresponding potential haplotypes for these five individuals are given by the associated vectors HapoEM\$hap1code and HapoEM\$hap2code as shown below, where the coding corresponds to the numbering system we saw for Caucasians in **Example 5.1**. The indx.subj vector tells us the corresponding record number and contains the sequence of numbers from 1 to the number of observations in our sample, with each element of the sequence repeated according to the value in HapoEM\$nreps.

```
> HapoEM$indx.subj[1:8]

[1] 1 2 2 3 3 4 4 5

> HapoEM$hap1code[1:8]

[1] 4 8 7 3 7 8 4 4
```

```
> HaploEM$hap2code[1:8]
```

```
[1] 4 3 4 8 4 3 7 4
```

Based on this output, one sees that the first and fifth individuals have the haplotype pair (4, 4), while the second, third, and fourth individuals are all ambiguous between (3, 8) and (4, 7). The posterior probabilities associated with these pairs are given by

```
> HaploEM$post[1:8] # Outputting:
```

```
[1] 1.000000000 0.006102808 0.993897192 0.006102808
```

```
[5] 0.993897192 0.006102808 0.993897192 1.000000000
```

Notably, the sum of these probabilities within any single individual is equal to 1. We can also calculate these probabilities directly based on the estimated haplotype frequencies. To see this, first note that the ten haplotype probabilities given in the first table of output in **Example 5.1** are contained in the vector

```
> HapProb <- HaploEM$hap.prob
```

```
> HapProb
```

```
[1] 0.0026138447 0.0093400121 0.0135382727 0.4729357032 0.0105890282
```

```
[6] 0.0006518550 0.3989126969 0.0855667219 0.0006548104 0.0051970549
```

Now consider one of our individuals who is ambiguous between the pairs (3, 8) and (4, 7). Assuming independence, which was already assumed in the estimation procedure, estimated probabilities of each of these pairs are given respectively by

```
> p1 <- 2*prod(HapProb[c(3,8)])
```

```
> p2 <- 2*prod(HapProb[c(4,7)])
```

```
> p1 / (p1 + p2)
```

```
[1] 0.006102807
```

```
> p2 / (p1 + p2)
```

```
[1] 0.9938972
```

As expected, these values are equivalent to the probabilities given in the second and third elements of HaploEM\$post.

Investigators often fill in unknown haplotypes by assigning each individual the haplotype pair with the highest corresponding posterior probability and then treating these as known in subsequent

analysis. We caution the reader against proceeding in this manner since valuable information on the uncertainty in the assignment is lost. Instead, methods may be applied if the ultimate goal is to characterize haplotype–trait association. Finally, one notes that testing hypotheses involving haplotype frequencies within the EM context requires consideration of the uncertainty in the estimation procedure. For example, suppose one is interested in testing the null hypothesis that the $h_4 = \text{CGCA}$ haplotype frequencies are equal for Caucasians and African Americans. This requires knowledge about the unknown variance/covariance matrix of the estimates. Derivation of this matrix may be obtained by inverting the observed information matrix and using Louis' method for the EM framework. Alternatively, the observed information matrix may be approximated with the empirical observed information.

5.4 Regression Decision Trees and Classifications^[Google]

Decision Trees are a type of algorithm for predictive modeling machine learning. The classical decision tree algorithms have been used for years, and modern variations like **Random Forest** are among the most powerful techniques available. The **decision tree algorithm**, also known by its more modern name **CART** (Classification And Regression Trees).

Aspects of the CART algorithm for machine learning include”

- (a) The representation used by learned CART models that is actually stored on disk.
- (b) A CART model may be learned from training data.
- (c) A learned CART model may be used to make predictions on unseen data.
- (d) Additional resources are available for learning more about CART and related algorithms (Fig. 5.27).



Fig. 5.27 A possible image for classification and regression trees for decision machine learning. (Photo by [Wonderlane](#))

Regression Decision Trees

Classification And Regression Trees (**CART**) is a term introduced by Breiman to refer to *Decision Tree* algorithms that can be used for classification or regression predictive modeling problems. Classically, this algorithm is referred to as “decision trees”, but on some platforms, like R, they are referred to by the more modern term CART.

The CART algorithm provides a foundation for important algorithms like Bagged Decision Trees, Random Forest, and Boosted Decision Trees (Figs. 5.28 and 5.29).

“I’ve created a handy mind map of 60+ algorithms organized by type.”

“Download it, print it and use it.”

CART Model Representation

The representation for the CART model is a binary tree.

This is a binary tree from algorithms and data structures. Each root node represents a single input variable (x) and a split point on that variable (assuming the variable is numeric).

The leaf nodes of the tree contain an output variable (y) which is used to make a prediction.

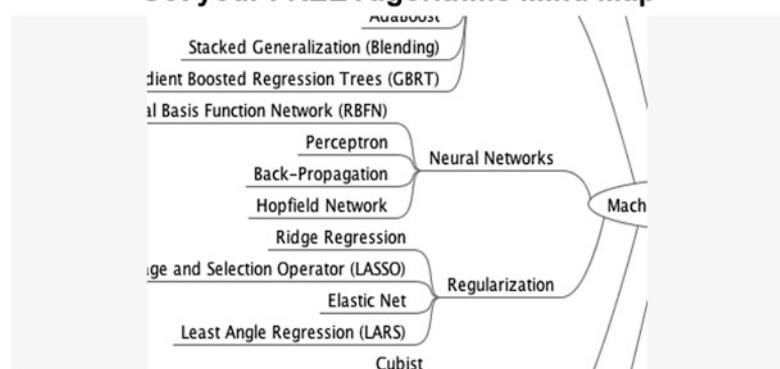
Given a dataset with two inputs (x) of height in centimeters and weight in kilograms the output of sex as male or female, below is a crude example of a binary decision tree (completely fictitious for demonstration purposes only) (Fig. 5.30).

The tree can be stored to file as a graph or a set of rules. For example, below is the above decision tree as a set of rules:

1	If Height > 180 cm Then Male
2	If Height <= 180 cm AND Weight > 80 kg Then Male
3	If Height <= 180 cm AND Weight <= 80 kg Then Female
4	Make Predictions With CART Models

Fig. 5.28 Sample of the handy machine learning algorithms mind map

Get your FREE Algorithms Mind Map



Download For FreeFrom: Jason @ ML Mastery

jason@machinelearningmastery.com via dripemail2.com xclusive access to the machine

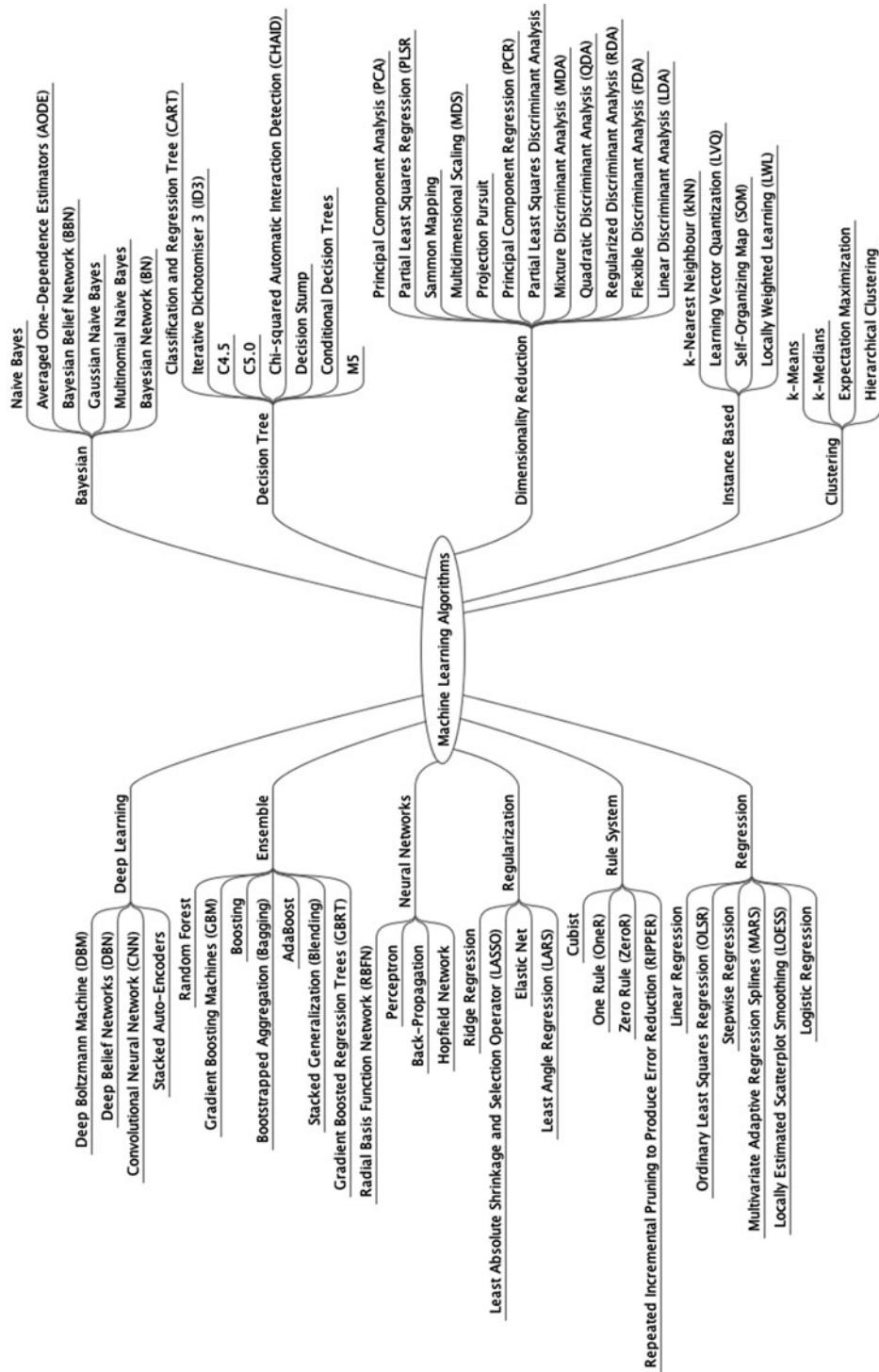
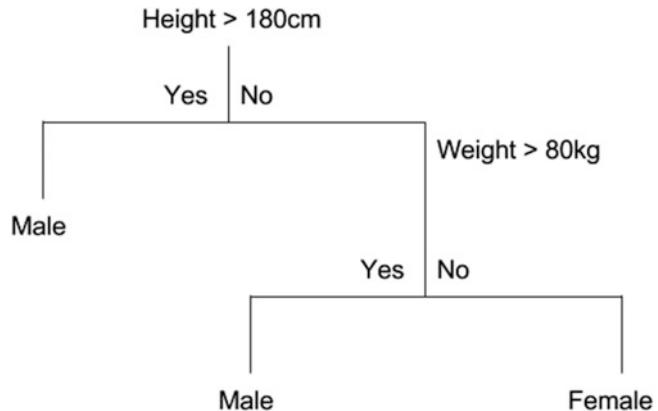


Fig. 5.29 Machine learning algorithms mind map

Fig. 5.30 Example decision tree



With the binary tree representation of the CART model described above, making predictions is relatively straightforward.

Given a new input, the tree is traversed by evaluating the specific input started at the root node of the tree.

A learned binary tree is actually a partitioning of the input space. One may think of each input variable as a dimension on a p -dimensional space. The decision tree split this up into rectangles (when $p = 2$ input variables) or some kind of hyper-rectangles with more inputs.

New data is filtered through the tree and lands in one of the rectangles and the output value for that rectangle is the prediction made by the model. This gives one some feeling for the type of decisions that a CART model is capable of making, e.g. boxy decision boundaries.

For example, given the input of [height = 160 cm, weight = 65 kg], we would traverse the above tree as follows:

1	Height > 180 cm: No
2	Weight > 80 kg: No
3	Therefore: Female

5.5 Multi-dimensional Analysis in Genetic Epidemiology

[B] Bull, S. B., Andrulus. I. L., and Paterson, A. D. (2018).- “Statistical challenges in high-dimensional molecular and genetic epidemiology”, *The Canadian Journal of Statistics*, Volume 46, Special Issue on “Big Data and the Statistical Sciences” Pages 24–40.

To investigate the many factors that can influence complex trait expression or disease courses, genetic association studies may offer a useful approach. As ***measurement techniques continually evolve, classical epidemiologic studies based on existing cohorts may raise methodology challenges***:

1. Molecular genetic prognostic factors in the history of node-negative breast cancer are studied using a combination of hypothesis generating and testing approaches.
2. Genome-wide association methods are applied to identify genes for multiple traits in an extended follow-up data from case-subjects of a therapeutic treatment program in Type-1 Diabetes.

5.5.1 Biomedical Background Challenges to Genetic Epidemiology

Each human being has 23 pairs of chromosomes inherited from their parents - one copy of each chromosome from the father and one copy from the mother. *Normally* each cell in our body includes the same DNA, which consists of **more than 3 billion pairs of nucleotides**. The sequence of nucleotides (A, T, C, G) along a chromosome is the DNA genetic code. A difference between nucleotides at a specific position in the sequence is called a **SNP** (Single Nucleotide Polymorphism) or a single nucleotide variant (SNV). An SNP is a single nucleotide change that is observed in at least 1% of a population, whereas SNV denotes variation without any restriction on variant frequency. Other more complex types of variation involving multiple nucleotides include insertions/deletions (Indels), and copy number variants (CNVs). There are roughly 8 million common SNPs, that is, those with variants that occur with frequency greater than 5% in the human population (1000 Genomes Project Consortium 2015).

The DNA sequence is highly structured, including genes, as well as nearby regulatory regions and intergenic regions of mostly unknown function between genes. It is estimated that there are 20,000 – 25,000 genes in the human genome (The ENCODE Project Consortium 2012); **Figure 5.31** illustrates that a gene, located on a chromosome segment, is structured as a promoter region, and alternating regions of exons (coding regions) and introns (non-coding regions). Transcription begins in the promoter and splicing, which removes introns, takes DNA code to mRNA, leading to RNA gene

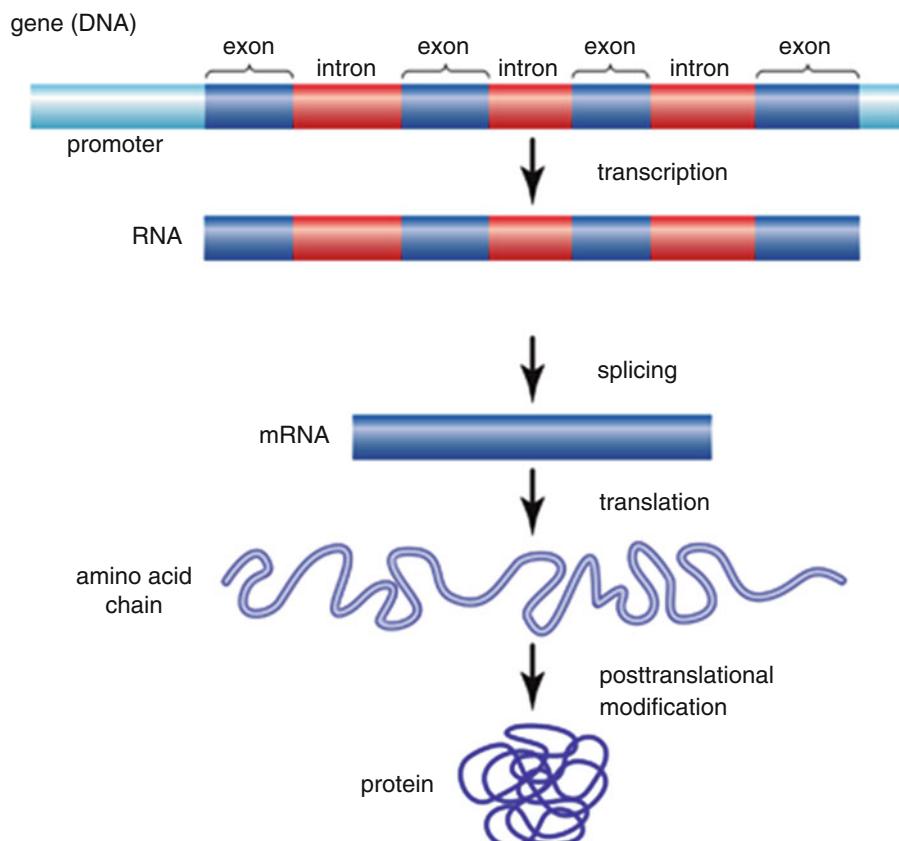


Fig. 5.31 Constructive illustration of a human gene

expression. For example, in translation and modification, amino acids coded by sets of three RNA nucleotides are assembled into a polypeptide chain, and one or more chains are linked to form a protein (Figure 5.31 (from the Encyclopedia Britannica 2015).

Transcription begins in the promoter and splicing, which removes introns, takes DNA code to mRNA, leading to RNA gene expression. For example, in translation and modification, amino acids coded by sets of three RNA nucleotides are assembled into a polypeptide chain, and one or more chains are linked to form a protein. Occurring at the cell level the process of: DNA → RNA. Protein is known as the “central dogma of biology.” Regulatory regions, outside genes, produce factors that can act on the promoter to modulate transcription. In addition, there are epigenetic modifications such as DNA methylation that can affect this process.

Structure and function of a gene (DNA → RNA → protein). Various technologies can measure DNA, RNA, and protein levels, usually in aggregation of many cells from relevant biological samples such as tumour tissue or blood. Molecular cancer epidemiology is concerned with populations of cells, particularly acquired alterations of cells in tumours, and how an individual's tumor characteristics influence disease prognosis, namely disease recurrence and mortality. Genetic epidemiology is concerned with inherited DNA variation, measured in blood or other convenient cells, and the association of DNA variants with individual-level characteristics such as risk factors for complex disease or susceptibility to disease itself. Figure 5.32 gives a brief list of molecular technologies that will be presented in the discussions that follow.

Molecular Prognostic Analysis – Node-Negative Breast Cancer

This research program began as a study to test the hypothesis that a specific tumor alteration had prognostic value for the risk of disease recurrence after standard treatment. At the time the study was designed in the late 1980s it was an open question whether emerging chemotherapy should be offered to women with axillary lymph node-negative (ANN) tumors as this group had overall good prognosis. An expressed concern was that many women would have to be treated unnecessarily with potentially toxic chemotherapy to benefit a few. Alterations in HER2/neu/erbB2 tumour DNA were postulated as a prognostic factor to identify those who might benefit from such therapy. Prospective multi-centre recruitment of newly diagnosed women was initiated, including collection of frozen tumour tissue whenever possible, and all eligible patients were monitored for post-treatment distant disease recurrence. Primary analysis, conducted when 34 metastatic disease events had been confirmed, found HER2/neu/erbB2 DNA amplification to be independently prognostic with a hazard ratio estimate of 2.4, after taking account of traditional prognostic factors. This study was one of the first to evaluate

Fig. 5.32 Some technologies in genetic epidemiology

Gene expression arrays:	mid- to high-throughput RNA RNA arrays
Tissue microarrays:	low-throughput protein expression TMA arrays (immunohistochemical)
Genotyping arrays:	high-throughput DNA GWAS SNP arrays Exome arrays
Targeted Sequencing:	low-throughput DNA Sanger sequencing
Next Generation Sequencing (NGS):	high-throughput Whole exome DNA seq Whole genome DNA seq

DNA amplification as a prognostic factor in node-negative disease, and the cohort has been followed clinically for many years, becoming an invaluable ongoing research resource. Numerous studies of HER2/neu/erbB2 in research labs and multicentre clinical trials eventually led to a clinical success story in which a targeted therapy (Herceptin) was found to be effective in women with HER2 positive tumors.

The primary study used a customized labor-intensive technology designed to measure amplification of a specific gene target in DNA from cells in each patient's tumor. DNA amplification is a particular form of acquired DNA alteration in which a chromosome segment is duplicated and the number of copies is greater than normal. The research program evolved into further studies as new scientific questions and new technologies emerged, and as the original cohort expanded and matured over time, notably with respect to the number of disease recurrence events.

GWAS studies aim to discover genomic variants associated with the trait of interest followed by replication in an independent dataset. GWAS arrays are designed to capture potential “causal” association indirectly by use of so-called tag SNPs, so direct measurement of the “causal” SNP is not necessary. A “causal” variant would be one that can ultimately be connected to a gene perturbation that leads to altered disease risk or trait value. To identify a chromosomal region of interest it is sufficient to detect association with a tag SNP that is correlated (in linkage disequilibrium, LD, with the unknown causal genetic variant. To improve resolution, individual-level imputation methods based on reference data such as available in the 1000 Genomes Project (2015) are widely used to estimate genotype values for 8–9 million unmeasured SNPs from 0.3 to 1 million measured SNPs. The SNPs included on commercial genotyping arrays such as the GWAS 1M array and the exome array (enriched for exonic variation) are not randomly chosen, but are highly selected to capture genome-wide variation.

Estimation under Selection

By multiple testing of single SNPs across the genome, typical GWAS association analysis amounts to high-dimensional variable selection. SNPs can be prioritized by *P*-value, ranking from the most significant, and strict criteria applied to control the global genome-wide type 1 error. A conventional criterion requires a *P*-value less than 5×10^{-8} for genome-wide significance. Optimistic bias in effect estimates for significant and/or higher ranked SNPs is a major consequence of such strict selection thresholds. This bias is worse when power is low, that is, for small effects, low frequency variants, and small samples, and affects both true positive and false positive associations. If the naïve estimate is taken at face value and used in sample size determination for replication, the study will be underpowered. This form of selection bias is known as the winner's curse or Beavis effect. Ideally, to obtain an unbiased estimate, the effect size of a discovered SNP association would be estimated in an independent sample, but this is not always practically feasible. To obtain effect estimates with reduced bias Sun and co-workers develop a non-parametric bootstrap resampling solution based on genome-wide analysis.

Regional Hypothesis Testing

GWAS analysis aims to comprehensively survey the genome for variants associated with a quantitative trait or disease status, and thereby identify a region to study more carefully, that is, by fine-mapping analysis. There are several motivations for a search targeted at a gene or a chromosomal region rather than an approach based on single SNP analysis: the gene is a natural biological unit for protein production; testing regions defined by sets of SNPs rather than single SNPs reduces the multiple testing

burden somewhat; a global test may be more robust to population differences and may be more sensitive to complex genetic architectures, for example, those involving multiple causal variants.

For a quantitative trait (Y) such as blood lipid levels, multi-SNP linear regression of a set of common SNPs ($X_j, j = 1, \dots, K$) can be specified by a regression model with K explanatory variables:

where X_j is coded as the number of copies (0, 1, or 2) of the non-reference variant in a SNP genotype. A global Wald test has an asymptotic chi-squared distribution with K degrees of freedom (df), one df for each SNP:

which is a quadratic test statistic based on the usual least squares coefficient estimates and associated variance covariance matrix estimate $\Sigma = \sigma^2(X^\top X)^{-1}/n$ obtained in a sample of n observations. The global null hypothesis of no association, $\beta_1 = \dots = \beta_K = 0$, is usually tested against the broad alternative hypothesis that at least one $\beta_j \neq 0$.

Among the set of SNPs included in the region-based regression a small number may be truly “causal” in the sense of having variants that affect RNA expression and/or protein level. Other SNPs (e.g., tag SNPs), carried on the same ancestral chromosome and correlated with causal variant(s), can serve as good surrogates to indirectly detect association. Yet other SNPs, uncorrelated with any causal variants, may be carried on a chromosome that has no association with the trait of interest, and will be consistent with a SNP-specific null hypothesis. Multiple causal variants may be correlated or uncorrelated, acting jointly or independently of one another.

To improve power to detect gene-level association in a manner that adapts to the local genetic correlation structure Yoo et al. (2013, 2017) propose a test statistic with reduced df oriented toward a restricted alternative (Li and Lagakos 2006). The idea underlying the test is that clustering the constituent SNPs according to the correlation structure within the region will combine information from a causal variant and/or its correlated neighbours, and such “causal” clusters will be separable from null clusters.

The number and composition of clusters are chosen by a network graph algorithm that identifies cliques in the network of SNPs such that all pairwise SNP correlations within a clique exceed a prespecified threshold value, with SNPs recoded to have positive pairwise correlation (Bron and Kerbosch 1973; Yoo et al. 2015). The contrast matrix C thus combines variant effects within the same cluster in a weighted linear combination, and then combines quadratic cluster-specific sums of squares and cross-products: the G_M test statistic has df equal to the number of clusters (Yoo et al. 2017). The restricted alternative hypothesis is that at least one of the cluster-specific linear combinations is associated with the trait. This test is directional in the sense of Li and Lagakos (2006) who compare the non-centrality parameters of an unrestricted global test and a linear combination directional test that is a function of the unrestricted effect estimates, and give a geometric interpretation. Under the global null hypothesis, G_M has an asymptotic chi-squared distribution with reduced $df_{L < K}$. Because the clusters and the coding are determined without using the trait data MLC does not incur a model selection penalty.

In the DCCT/EDIC candidate gene study of individuals with type 1 diabetes, application of MLC statistics to a set of 10 common SNPs that cluster into five subsets in the CETP gene confirms a known association with HDL-cholesterol detected previously in the general population (Teslovich et al. 2010; Yoo et al. 2017). In simulation studies of type 1 error and power in each of 1,000 genes using common SNP genotypes for 1,000 individuals derived from an Asian population of common variation (The International HapMap 3 Consortium, 2010), Yoo et al. (2017) found MLC to compare favourably to existing methods, especially as the number of “causal” variants increases. While MLC statistics are

valid, and more powerful than the generalized Wald statistic for a large majority of the 1,000 genes, on average MLC power is similar to that of alternative gene-based marginal methods, including variance-component statistics. In the absence of knowledge about the underlying genetic architecture, that is, the number and effect-size distribution of causal variants, there can be no best method. Nevertheless the observation that power across genes is less variable for MLC compared to other methods implies that MLC is reasonably robust and may perform better overall in genome-wide analysis.

Design and Analysis—Two Phase Sampling Studies

Once a SNP or a region with evidence of genetic association is detected, statistical fine-mapping studies aim to acquire information on all possible causal variants in a region and evaluate relative evidence for causality (Faye et al. 2013; Spain and Barrett 2015). Two-phase designs present opportunities for gains in cost efficiencies in both molecular and genetic epidemiology study design, although most work to date has focused on the GWAS setting (Thomas et al. 2009, 2013; Lin et al. 2013; Schaid et al. 2013). A primary motivation for two-phase designs stems from the prohibitive cost of any emerging technology. For example, by selecting a subset of informative individuals for expensive sequencing, cost efficiencies can be gained compared to sequencing everyone. In settings such as tumour studies, preservation of precious tissue samples is an additional motivation for the use of sampling.

In targeted sequencing of chromosomal regions detected by GWAS analysis Phase 1 is the GWAS: the sample size is large (e.g., $N = 5,000$), millions of SNPs are tested, and a GWAS SNP Z that meets a genome-wide significance criterion is detected. Phase 2 involves dense sequencing targeted to a region that includes the GWAS SNP Z : the GWAS sample is stratified on the Z_i genotype, individuals are sampled within strata at different rates (e.g., for a total sample of $n = 2,000$) to reduce sequencing costs, and each of several hundred sequence variants (e.g., $G_j, j = 1, \dots, m$) may be genotyped and tested. Combined analysis of data from both phases can achieve high relative efficiency when Z and G_j are well correlated. Valid and efficient methods for analysis of data thus obtained include estimating equations, inverse probability weighting, and semi-parametric maximum likelihood (Lawless et al. 1999; Zhao et al. 2009; Chen et al. 2012; Zeng and Lin 2014). A Bayesian approach to fine mapping enables comparison among variants using Bayes factors to select a credible set of variants from among those sequenced (WTCCC 2012; Chen et al. 2014; Spain and Barrett 2015). Compared to simple random sampling, which ignores genetic information from GWAS, tag-SNP-based stratified sample allocation reduces the number of variants in the credible interval and is more likely to promote the causal sequence variant into confirmation studies (Chen et al. 2014). For studies of quantitative traits the use of trait-dependent sampling, alone or in combination with genotype-dependent sampling, can also improve cost-efficiency but inference is complicated by ascertainment on the outcome (Yilmaz and Bull 2001; Lin et al. 2013; Derkach et al. 2015; Espin-Garcia et al. 2016).

Summary and Prospects

This review highlights a few selected methodological issues in molecular and genetic epidemiology, and is by no means comprehensive. Illustrations have been drawn from two longitudinal cohort studies that have evolved over time, incorporating emerging genomic technologies and integration with well-characterized individual-level data, and presenting statistical problems in study design and data analysis. Although not discussed here family studies involving high-dimensional multi-omics data remain important scientifically and present additional interesting methodological problems. For example, publications from recent Genetic Analysis Workshops report evaluations of methods for study design, model specification, treatment of missing data, and statistical computation (e.g., Bickeböller et al. 2014; Li et al. 2014; Cantor and Cordell 2016; Wijsman 2016).

While opportunities to pursue new scientific questions are often predicated on new technologies, working with developing molecular technologies involves messy data and mid-study technology improvements. Application of quality control procedures is essential as well as intellectual investment in understanding measurement issues and scientific questions. We can expect “next generation” sequencing technologies to continue to present new opportunities for statistical innovation (Mechanic et al. 2012; Lange et al. 2014; Goodwin et al. 2016; Pulit et al. 2017). A major challenge is the development of study designs encompassing statistical modelling and analysis appropriately informed by biological knowledge such as prediction of variant effects on protein coding and reference data that can be used to infer unmeasured features (Gamazon et al. 2015; Spain and Barrett 2015; McCarthy et al. 2016; Spencer et al. 2016). In discovery settings we need statistical inference that is robust to model misspecification and accounts for data-based model selection. With the proliferation of global ‘omics data we can anticipate continuing development of methods to integrate genomics, transcriptomics, proteomics, epigenomics, metabolomics, exposomics, and microbiome data (Khoury 2014).

Referenced Acknowledgements

Shelley Bull acknowledges research support from the Canadian Institutes of Health Research & the Natural Sciences and Engineering Research Council (Canada). Andrew Paterson and Shelley Bull acknowledge support from the Juvenile Diabetes Research Foundation (International). Irene Andrulis and Shelley Bull acknowledge support from the Canadian Institutes of Health Research.

Ancillary

After the initial hypothesis testing studies of tumor DNA for HER2 gene amplification and p53 gene mutations in the primary patient cohort, the emergence of high-throughput microarrays for measurement of genome-wide RNA gene expression (GE) and DNA copy number (CN) stimulated the conduct of discovery studies consisting of two-group comparisons in selected smaller subgroups (e.g., He et al. 2011). Across the genome duplications and deletions of chromosome segments produce, respectively, gains or losses in DNA copy number. The later hypothesis generating studies of tumour DNA used array comparative genomic hybridization (aCGH), a technology that quantifies copy number gains or losses in a tumour sample (relative to a normal control) at each of a large number of locations distributed across the genome.

2.1 Genomic Data Integration

One of the microarray substudy analyses illustrates some of the challenges that arise from high-dimensional hypothesis testing and complex spatial data structure (Asimit et al. 2011). In this case there were a small number of tumour samples (n) and two types of molecular genetic measurements (GE and CN) at a large number of genomic locations (p) with $n < p$. The specific locations are determined by microarray construction in which so-called gene probes designed to measure gene-specific DNA or RNA levels in a sample are placed on the array in an ordered fashion according to the array design (see Theisen 2008 for an instructive description of the array hybridization process that produces quantitative values). The GE and CN values obtained represent the aggregate of heterogeneous cells within a tumor.

Following Richardson et al. (2016) data integration can be defined as statistical analysis that aims to answer biological questions by joint modelling of different types of genomic data in the same set of samples. Different approaches to integration lead to different formulations for inference and hypothesis testing. Duplications and deletions of chromosome DNA segments are regional in nature, and CN has typically been analyzed within an individual, using statistical methods to call CN alterations as individual copy number states from the aCGH quantitative measures, and inferring change-points in the underlying copy number along the chromosome (Xing et al. 2007). In contrast, GE analysis is

usually conducted among individuals, considering one gene probe at a time, and ignoring genomic location, spatial correlation, and distance between probes. The best approach to joint analysis in this context, incorporating both within- and among-individual comparisons, is not obvious, given the high-dimension data structure, and the potential impact of multiple testing and over-fitting.

For the microarray substudy RNA and DNA extracted from the same patient's tumour were applied separately to a pair of arrays, which yielded paired measures of GE and CN, respectively, for each of the gene probes in 68 tumour samples. Let i index tumours and j index gene probe; the latter with genome-wide dimensionality of approximately 19,000. The approach taken for integrated analysis of these data is based on the idea that if a gene is biologically important, an association between CN and GE should be detectable among the tumour samples, and chromosome regions with strong positive association at multiple gene probes are more likely to harbour alterations that drive tumour development and/or disease progression (Asimit et al. 2011). The relationship between two types of measurements of the same gene, a so-called *cis*-effect, can be thought of as “vertical” association. Probe-specific association is examined first, using linear regression to model association between GE and CN—that is, at gene probe j , RNA expression level (Y) depends on DNA copy number level (X):

For instance, the scatter plot in Panel A of Box 2 suggests evidence of positive association at one probe (location 7.358507 Mb, chromosome 17); the regression is fit by robust regression to reduce effects of isolated outliers that can occur in microarray data. Covariates for other tumour characteristics such as lymphovascular invasion (LVI) can also be included in the regression:

Box 2: Identification and evaluation of CN-GE association genomic regions. Figures from Asimit et al. (2011), Copyright © 2011; reproduced by permission of John Wiley & Sons, Ltd.

Then, because multiple associations within a small region are considered to be more convincing, regions of CN-GE association across the “horizontal” chromosome direction are detected using scan statistics that account for genomic distances between probes and the segmental nature of chromosome CN values, as described by Asimit et al. (2011). Region detection proceeds by considering all possible windows of $(r + 1)$ probes shifted along the chromosome for $r = 2, 3, \dots, 10$ inter-probe distances. The scan statistic is $S_{j, r}(k)$ where $k \leq r$ is the number of inter-probe distances between successive “significant” probes in the region from probe j to prob $\epsilon j + r$. $S_{j, r}(k)$ is the sum of inter-probe distances, and probe-specific significance is determined by the criterion $P_j < \Theta$ for the test of hypothesis $H_0 : \beta_j = 0$ versus $H_A : \beta_j > 0$. Under the assumption that the k inter-probe distances are exponentially distributed with rate λ , $S_{j, r}(k)$ follows a gamma (k, λ) distribution. A regional test of $H_0 : K = K_0$ versus $H_A : K > K_0$ evaluates whether the number of significant probes is greater than expected over the observed distance $t = S_{j, r}(k)$ where $K_0 = \lambda_0 r - 1$ and λ_0 is the rate parameter estimated under the global null hypothesis of no CN-GE associations. If the regional probability is less than a criterion the set of probes identifies a region of association not likely to occur by chance (Fig. 5.33).

To address potential effects of multiple testing and departures from parametric assumptions Asimit et al. (2011) also develop a nonparametric bootstrap that compares the frequencies with which each probe appears within a detected region in bootstrap samples of the original data against the frequencies obtained in data generated under the global null hypothesis of no CN-GE association. The vertical lines in Panel B of Box 2 are bootstrap frequencies of probes identified in regions on chromosome 17 with probe P -value threshold $\theta = 0.1$ and regional threshold $\alpha = 0.01$. For three of the four original regions detected (shown across the bottom of the figure) the bootstrap frequencies are much higher than the global null frequencies, denoted by the solid points, suggesting stronger evidence for true positive association.

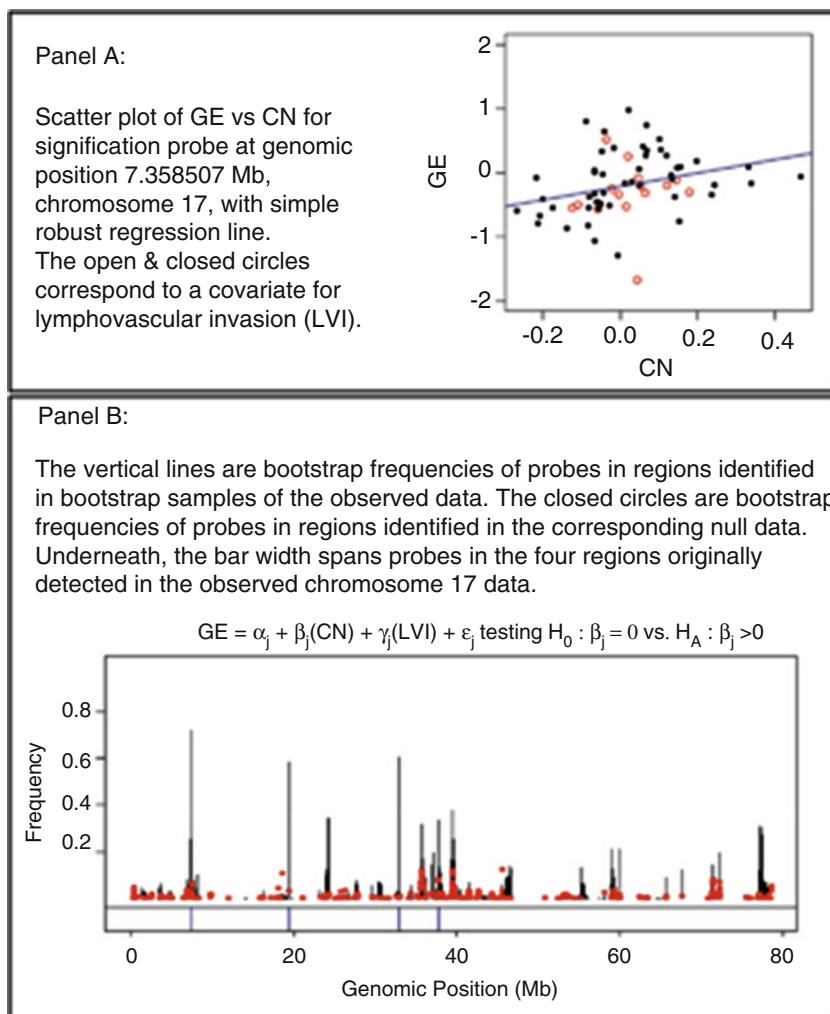


Fig. 5.33 Box 2 - microarray substudy analyses

Modelling under Heterogeneity

Recent studies in the ANN cohort have addressed questions in translational hypothesis testing which are relevant to patient prognosis and treatment. Tissue Microarray (TMA) technology measures protein expression using immunohistochemistry (IHC) which is a standard method applied to archived tumour samples in clinical pathology (Mulligan et al. 2008, 2016; Feeley et al. 2013). TMAs can handle hundreds of samples, but in doing so expend tumour tissue. The statistical issue here is how to account for patient heterogeneity. In the node-negative cohort Kaplan–Meier (K–M) survival curves for time to distant disease recurrence differ among four groups of women classified by tumour subtype (Figure 2); subtype classification is based on a combination of several TMA protein biomarker values. The K–M survival curves are non-proportional over the follow-up period and two of the curves cross, clearly violating the standard proportional hazards (PH) model assumption. Moreover, survival shows a pattern of long stable plateau with heavy censoring in the tail suggesting there is a fraction of individuals who do not continue to be at risk. A possible interpretation is that the study subjects consist of a mixture of long-term disease-free survivors (i.e., cured by standard therapy) and

susceptible patients who will experience recurrence at varying durations after diagnosis. Because there are unknown prognostic factors that would explain variation in susceptibility and time to recurrence each of the subtypes is a mixture of cured and susceptible patients. As those at risk are removed from the risk sets, leaving only cured individuals under observation, the survival curves flatten out.

Survival data for time to disease recurrence in ANN breast cancer.

A PH mixture cure model addresses this type of heterogeneity by specifying a survival probability $S(t|x)$ at time t given covariates x for a mixture of two groups of individuals (Farewell [1982](#); Yilmaz et al. [2013](#)). The study sample consists of a mixture of long-term survivors (cured) and susceptible women who will experience recurrence at some time after diagnosis. The probability of cure $p(x)$ is modelled by logistic regression as a function of covariates (such as tumour subtype) and time to recurrence in susceptible women is modelled by a Weibull survival model. The two association parameter vectors (denoted by α in the logistic model, β in the Weibull model) allow identification of different factors for early recurrence versus longterm survival. Forse et al. ([2013](#)) apply a similar PH mixture cure model in the node-negative cohort to evaluate the prognostic importance of podocalyxin protein expression (PODXL), a biomarker discovered in the microarray studies. Although ANN women with tumours expressing high PODXL have, on average, a less favourable risk profile according to traditional prognostic factors, paradoxically a higher proportion of women with high PODXL expressing tumours experience long-term disease-free survival (DFS). The mixture-cure model analysis helps to resolve counterintuitive results produced in standard PH model analysis by demonstrating that tumour overexpression of PODXL is associated with poor prognosis characteristics and earlier recurrence times in the “susceptible” group, yet is nevertheless associated with improved cure rates in ANN breast cancer.

Genome-Wide Association Analysis—Diabetes Complications

Studies of complications in individuals with type 1 diabetes originally recruited for a randomized clinical trial (RCT) illustrate one of the early examples in the field of genetic epidemiology that took an existing well-characterized longitudinal cohort and applied new technologies to answer questions about the role of genetic variation in susceptibility to complex disease (Box 3). The Diabetes Control and Complications Trial (DCCT 1993) was a pivotal RCT of intensive therapy designed to control blood glucose levels; elevated levels are understood to be harmful to cells, leading to kidney and eye complications. It was followed by Epidemiology of Diabetes Interventions and Complications (EDIC 1999) the post-trial ongoing follow-up study, and the DCCT/EDIC Genetics Study was later initiated to identify genetic susceptibilities to complications and related traits (Al-Kateb et al. [2007](#), [2008](#); Paterson et al. [2010](#)). Available genotyping technologies for genetic association evolved over time, beginning with a lower density custom array, and extending to the current use of high-density commercial arrays and multi-study multi-platform meta-analysis (Hosseini et al. [2015](#); Roshandel et al. [2016](#)) (Fig. 5.34).

Direct and indirect association in genome wide association study. Figure provided courtesy of Y. J. Yoo; used by permission.

The path diagram in Box 3 represents a conceptual model for genetic influences on glycemia and long-term complications (Paterson and Bull [2012](#)). Variation in certain genes may influence glycemia levels in type 1 diabetes (solid black line). Other genetic loci may independently influence risk for retinal and renal complications without acting through glycemia. Some of these loci will be specific to either retinal or renal complications (dashed line, dash-dotted line), while others will have effects on

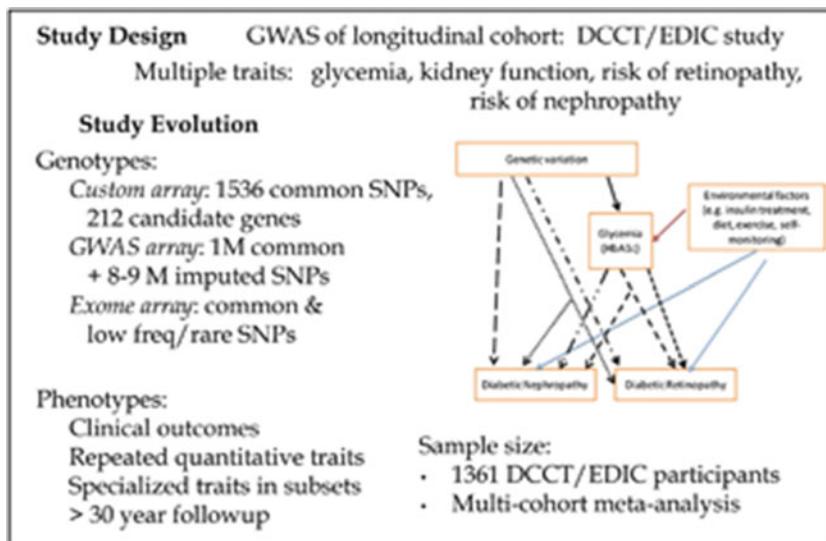


Fig. 5.34 Research program in genetic epidemiology of type 1 diabetes complications. Figure from Paterson & Bull (2012), Copyright © 2012; reproduced by permission of Springer Science+Business Media, LLC

both renal and retinal outcomes (split dotted line). Glycemia may in turn be associated with long-term diabetic complications, together with various environmental factors.

GWAS studies aim to discover genomic variants associated with the trait of interest followed by replication in an independent dataset. As illustrated in Figure 3 GWAS arrays are designed to capture potential “causal” association indirectly by use of so-called tag SNPs, so direct measurement of the “causal” SNP is not necessary. A “causal” variant would be one that can ultimately be connected to a gene perturbation that leads to altered disease risk or trait value (Spain and Barrett 2015). To identify a chromosomal region of interest it is sufficient (albeit with some loss of power) to detect association with a tag SNP that is correlated (in linkage disequilibrium, LD) with the unknown causal genetic variant. To improve resolution, individual-level imputation methods based on reference data such as available in the 1000 Genomes Project (2015) are widely used to estimate genotype values for 8–9 million unmeasured SNPs from 0.3 to 1 million measured SNPs. The SNPs included on commercial genotyping arrays such as the GWAS 1M array and the exome array (enriched for exonic variation) are not randomly chosen, but are highly selected to capture genome-wide variation.

3.1 Estimation under Selection

By multiple testing of single SNPs across the genome, typical GWAS association analysis amounts to high-dimensional variable selection. SNPs can be prioritized by *P*-value, ranking from the most significant, and strict criteria applied to control the global genome-wide type 1 error. A conventional criterion requires a *P*-value less than 5×10^{-8} for genome-wide significance (Dudbridge and Gusnanto 2008). Optimistic bias in effect estimates for significant and/or higher ranked SNPs is a major consequence of such strict selection thresholds. This bias is worse when power is low, that is, for small effects, low frequency variants, and small samples, and affects both true positive and false positive associations. If the naïve estimate is taken at face value and used in sample size determination for replication, the study will be underpowered. This form of selection bias is known as the winner’s curse or Beavis effect (Xu 2003). Ideally, to obtain an unbiased estimate, the effect size of a discovered SNP association would be estimated in an independent sample, but this is not always practically feasible. To obtain effect estimates with reduced bias Sun and co-workers develop a non-parametric bootstrap resampling solution based on genome-wide analysis (Sun and Bull 2005, Faye et al. 2011).

They estimate threshold and ranking bias by imitating GWAS discovery and replication for each bootstrap sample:

Here (k) denotes the k th ranked SNP obtained in genome-wide testing. The magnitude of the naïve estimate, $\hat{\beta}_{N(k)}$, is reduced by a shrinkage factor that is an average of the difference between discovery and estimation effect estimates taken over a large number, $B_{(k)}$, of bootstrap samples indexed by i . The “discovery” estimate is calculated from the observations in the bootstrap sample, and the “replication” estimate $\hat{\beta}_{Ei(k)}$ is calculated from the out-of-sample observations. The $*$ denotes an additional adjustment (not shown, see Faye et al. 2011) for correlation between bootstrap-sample and out-of-sample effect estimates. Because bias depends on SNP variant frequency, the shrinkage factors are weighted according to variance ratios between the k th ranked SNP in the i th bootstrap sample and the k th ranked SNP in the original data (Faye et al. 2011).

This approach is broadly adaptable in that any well-defined selection threshold criterion can be applied in each bootstrap sample, and stratification by P -value rank accounts for competition among SNPs. Application to a GWAS of glycemia in the DCCT/EDIC study using an implementation in “BR-squared” (“Bias Reduced estimates via Bootstrap Resampling”) software (Sun et al. 2011) reduced estimated effect sizes for the top SNPs by more than 50%. “BR-squared” also handles case-control designs and extensions to cohort studies with time-to-event outcomes (Poirier et al. 2015).

Regional Hypothesis Testing

GWAS analysis aims to comprehensively survey the genome for variants associated with a quantitative trait or disease status, and thereby identify a region to study more carefully, that is, by fine-mapping analysis. There are several motivations for a search targeted at a gene or a chromosomal region rather than an approach based on single SNP analysis: the gene is a natural biological unit for protein production; testing regions defined by sets of SNPs rather than single SNPs reduces the multiple testing burden somewhat; a global test may be more robust to population differences and may be more sensitive to complex genetic architectures, for example, those involving multiple causal variants (Asimit et al. 2009; Lehne Lewis and Schlitt 2011; Shi and Weinberg 2011; Stringer et al. 2011).

For a quantitative trait (Y) such as blood lipid levels, multi-SNP linear regression of a set of common SNPs ($X_j, j = 1, \dots, K$) can be specified by a regression model with K explanatory variables. A global Wald test has an asymptotic chi-squared distribution with K degrees of freedom (df), one df for each SNP which is a quadratic test statistic based on the usual least squares coefficient estimates and associated variance covariance matrix estimate $\Sigma = \sigma^2(X^T X)^{-1}/n$ obtained in a sample of n observations. The global null hypothesis of no association, $\beta_1 = \dots = \beta_K = 0$, is usually tested against the broad alternative hypothesis that at least one $\beta_j \neq 0$.

Among the set of SNPs included in the region-based regression a small number may be truly “causal” in the sense of having variants that affect RNA expression and/or protein level. Other SNPs (e.g., tag SNPs), carried on the same ancestral chromosome and correlated with causal variant(s), can serve as good surrogates to indirectly detect association. Yet other SNPs, uncorrelated with any causal variants, may be carried on a chromosome that has no association with the trait of interest, and will be consistent with a SNP-specific null hypothesis. Multiple causal variants may be correlated or uncorrelated, acting jointly or independently of one another.

To improve power to detect gene-level association in a manner that adapts to the local genetic correlation structure Yoo et al. (2013, 2017) propose a test statistic with reduced df oriented toward a restricted alternative (Li and Lagakos 2006). The idea underlying the test is that clustering the constituent SNPs according to the correlation structure within the region will combine information

from a causal variant and/or its correlated neighbours, and such “causal” clusters will be separable from null clusters.

The number and composition of clusters are chosen by a network graph algorithm that identifies cliques in the network of SNPs such that all pairwise SNP correlations within a clique exceed a prespecified threshold value, with SNPs recoded to have positive pairwise correlation (Bron and Kerbosch 1973; Yoo et al. 2015). The contrast matrix C thus combines variant effects within the same cluster in a weighted linear combination, and then combines quadratic cluster-specific sums of squares and cross-products: the G_M test statistic has df equal to the number of clusters (Yoo et al. 2017). The restricted alternative hypothesis is that at least one of the cluster-specific linear combinations is associated with the trait. This test is directional in the sense of Li and Lagakos (2006) who compare the non-centrality parameters of an unrestricted global test and a linear combination directional test that is a function of the unrestricted effect estimates, and give a geometric interpretation. Under the global null hypothesis, G_M has an asymptotic chi-squared distribution with reduced $df_{L < K}$. Because the clusters and the coding are determined without using the trait data MLC does not incur a model selection penalty.

In the DCCT/EDIC candidate gene study of individuals with type 1 diabetes, application of MLC statistics to a set of 10 common SNPs that cluster into five subsets in the CETP gene confirms a known association with HDL-cholesterol detected previously in the general population (Teslovich et al. 2010; Yoo et al. 2017). In simulation studies of type 1 error and power in each of 1,000 genes using common SNP genotypes for 1,000 individuals derived from an Asian population of common variation (The International HapMap 3 Consortium, 2010), Yoo et al. (2017) found MLC to compare favourably to existing methods, especially as the number of “causal” variants increases. While MLC statistics are valid, and more powerful than the generalized Wald statistic for a large majority of the 1,000 genes, on average MLC power is similar to that of alternative gene-based marginal methods, including variance-component statistics. In the absence of knowledge about the underlying genetic architecture, that is, the number and effect-size distribution of causal variants, there can be no best method. Nevertheless the observation that power across genes is less variable for MLC compared to other methods implies that MLC is reasonably robust and may perform better overall in genome-wide analysis.

Design and Analysis—Two Phase Sampling Studies

Once a SNP or a region with evidence of genetic association is detected, statistical fine-mapping studies aim to acquire information on all possible causal variants in a region and evaluate relative evidence for causality (Faye et al. 2013; Spain and Barrett 2015). Two-phase designs present opportunities for gains in cost efficiencies in both molecular and genetic epidemiology study design, although most work to date has focused on the GWAS setting (Thomas et al. 2009, 2013; Lin et al. 2013; Schaid et al. 2013). A primary motivation for two-phase designs stems from the prohibitive cost of any emerging technology. For example, by selecting a subset of informative individuals for expensive sequencing, cost efficiencies can be gained compared to sequencing everyone. In settings such as tumour studies, preservation of precious tissue samples is an additional motivation for the use of sampling.

In targeted sequencing of chromosomal regions detected by GWAS analysis Phase 1 is the GWAS: the sample size is large (e.g., $N = 5,000$), millions of SNPs are tested, and a GWAS SNP Z that meets a genome-wide significance criterion is detected. Phase 2 involves dense sequencing targeted to a region that includes the GWAS SNP Z : the GWAS sample is stratified on the Z_i genotype, individuals are sampled within strata at different rates (e.g., for a total sample of $n = 2,000$) to reduce sequencing costs, and each of several hundred sequence variants (e.g., $G_j, j = 1, \dots, m$) may be genotyped and tested. Combined analysis of data from both phases can achieve high relative efficiency when Z and G_j

are well correlated. Valid and efficient methods for analysis of data thus obtained include estimating equations, inverse probability weighting, and semi-parametric maximum likelihood (Lawless et al. 1999; Zhao et al. 2009; Chen et al. 2012; Zeng and Lin 2014). A Bayesian approach to fine mapping enables comparison among variants using Bayes factors to select a credible set of variants from among those sequenced (WTCCC 2012; Chen et al. 2014; Spain and Barrett 2015). Compared to simple random sampling, which ignores genetic information from GWAS, tag-SNP-based stratified sample allocation reduces the number of variants in the credible interval and is more likely to promote the causal sequence variant into confirmation studies (Chen et al. 2014). For studies of quantitative traits the use of trait-dependent sampling, alone or in combination with genotype-dependent sampling, can also improve cost-efficiency but inference is complicated by ascertainment on the outcome (Yilmaz and Bull 2001; Lin et al. 2013; Derkach et al. 2015; Espin-Garcia et al. 2016).

Prospects

This review highlights a few selected methodological issues in molecular and genetic epidemiology, and is by no means comprehensive. Illustrations have been drawn from two longitudinal cohort studies that have evolved over time, incorporating emerging genomic technologies and integration with well-characterized individual-level data, and presenting statistical problems in study design and data analysis. Although not discussed here family studies involving high-dimensional multi-omics data remain important scientifically and present additional interesting methodological problems. For example, publications from recent Genetic Analysis Workshops report evaluations of methods for study design, model specification, treatment of missing data, and statistical computation (e.g., Bickeböller et al. 2014; Li et al. 2014; Cantor and Cordell 2016; Wijsman 2016).

While opportunities to pursue new scientific questions are often predicated on new technologies, working with developing molecular technologies involves messy data and mid-study technology improvements. Application of quality control procedures is essential as well as intellectual investment in understanding measurement issues and scientific questions. We can expect “next generation” sequencing technologies to continue to present new opportunities for statistical innovation (Mechanic et al. 2012; Lange et al. 2014; Goodwin et al. 2016; Pulit et al. 2017). A major challenge is the development of study designs encompassing statistical modelling and analysis appropriately informed by biological knowledge such as prediction of variant effects on protein coding and reference data that can be used to infer unmeasured features (Gamazon et al. 2015; Spain and Barrett 2015; McCarthy et al. 2016; Spencer et al. 2016). In discovery settings we need statistical inference that is robust to model misspecification and accounts for data-based model selection. With the proliferation of global ‘omics data we can anticipate continuing development of methods to integrate genomics, transcriptomics, proteomics, epigenomics, metabolomics, exposomics, and microbiome data (Khoury 2014).

5.5.2 Worked Examples in Epidemiology

Worked Example 1

Package randomForestSRC

Random Forests for Survival, Regression, and Classification (RFSRC)

A unified treatment of Breiman's random forests for survival, regression and classification problems based on Ishwaran and Kogalur's Random Survival Forests (RSF) package. Now extended to include

multivariate and unsupervised forests. Also includes quantile regression forests for univariate and multivariate training/ testing settings. The package runs in both serial and parallel (OpenMP) modes.

Version:	2.5.1
Depends:	R (\geq 3.1.0)
Imports:	parallel
Suggests:	glmnet, survival, pec, proddlim, mlbench
Published:	2017-10-17
Author:	Hemant Ishwaran, Udaya B. Kogalur
Maintainer:	Udaya B. Kogalur <ubk at kogalur.com>
BugReports:	https://github.com/kogalur/randomForestSRC/issues/new
License:	GPL (\geq 3)
URL:	http://web.ccs.miami.edu/~hishwaran http://www.kogalur.com https://github.com/kogalur/randomForestSRC
NeedsCompilation:	yes
Citation:	randomForestSRC citation info
Materials:	NEWS
In views:	HighPerformanceComputing , MachineLearning , Survival
CRAN checks:	randomForestSRC results

Downloads:

Reference manual:	randomForestSRC.pdf
Package source:	randomForestSRC_2.5.1.tar.gz
Windows binaries:	r-devel: randomForestSRC_2.5.1.zip , r-release: randomForestSRC_2.5.1.zip , r-oldrel: randomForestSRC_2.5.1.zip
OS X El Capitan binaries:	r-release: randomForestSRC_2.5.1.tgz
OS X Mavericks binaries:	r-oldrel: randomForestSRC_2.5.1.tgz
Old sources:	randomForestSRC archive

Reverse dependencies:

Reverse depends:	ggRandomForests
Reverse imports:	boostmtree , fifer , sprinter , SurvRank
Reverse suggests:	CFC , edarf , IPMRF , mlr , ModelGood , pec , pmml , riskRegression

In the R domain:

plot.survival **Plot of Survival Estimates**

Description Plot various survival estimates.

Usage

```
## S3 method for class 'rfsrsrc'
plot.survival(x, plots.one.page = TRUE,
show.plots = TRUE, subset, collapse = FALSE,
haz.model = c("spline", "ggamma", "nonpar", "none"),
k = 25, span = "cv", cens.model = c("km", "rfsrsrc"), ...)
```

Arguments

x	An object of class (rfsrc, grow) or (rfsrc, predict).
plots.one.	Should plots be placed on one page?
page	
show.plots	Should plots be displayed?
subset	Vector indicating which individuals we want estimates for. All individuals are used if not specified.
collapse	Collapse the survival and cumulative hazard function across the individuals specified by 'subset'? Only applies when 'subset' is specified.
haz.model	Method for estimating the hazard. See details below. Applies only when 'subset' is specified.
k	The number of natural cubic spline knots used for estimating the hazard function. Applies only when 'subset' is specified.
span	The fraction of the observations in the span of Friedman's super-smoother used for estimating the hazard function. Applies only when 'subset' is specified.
cens.model	Method for estimating the censoring distribution used in the inverse probability of censoring weights (IPCW) for the Brier score:

km: Uses the Kaplan-Meier estimator.

rfsr: Uses random survival forests.

... Further arguments passed to or from other methods.

Details

If 'subset' is not specified, generates the following three plots (going from top to bottom, left to right):

1. Forest estimated survival function for each individual (thick red line is overall ensemble survival, thick green line is Nelson-Aalen estimator).
2. Brier score (0=perfect, 1=poor, and 0.25=guessing) stratified by ensemble mortality. Based on the IPCW method described in Gerd et al. (2006). Stratification is into 4 groups corresponding to the 0-25, 25-50, 50-75 and 75-100 percentile values of mortality. Red line is the overall (non-stratified) Brier score.
3. Plot of mortality of each individual versus observed time. Points in blue correspond to events, black points are censored observations.

When 'subset' is specified, then for each individual in 'subset', the following three plots are generated:

1. Forest estimated survival function.
2. Forest estimated cumulative hazard function (CHF) (displayed using black lines). Blue lines are the CHF from the estimated hazard function. See the next item.
3. A smoothed hazard function derived from the forest estimated CHF (or survival function). The default method, 'haz.model="spline"', models the log CHF using natural cubic splines as described in Royston and Parmar (2002). The lasso is used for model selection, implemented using the glmnet package (this package must be installed for this option to work). If 'haz.model="ggamma"', a three-parameter generalized gamma distribution (using the parameterization described in Cox et al 2007) is fit to the smoothed forest survival function, where smoothing is imposed using Friedman's

supersmoother (implemented by supsmu). If 'haz.model="nonpar"', Friedman's supersmoother is applied to the forest estimated hazard function (obtained by taking the crude derivative of the smoothed forest CHF). Finally, setting 'haz.model="none"' suppresses hazard estimation and no hazard estimate is provided.

At this time, please note that all hazard estimates are considered experimental And users should interpret the results with caution.

Note that when the object x is of class (rfsrc, predict) not all plots will be produced. In particular, Brier scores are not calculated.

Only applies to survival families. In particular, fails for competing risk analyses.

Use plot.competing.risk in such cases.

Whenever possible, out-of-bag (OOB) values are used.

Value

Invisibly, the conditional and unconditional Brier scores, and the integrated Brier score (if they are available).

Authors Hemant Ishwaran and Udaya B. Kogalur

References

Cox C., Chu, H., Schneider, M. F. and Munoz, A. (2007). Parametric survival analysis and taxonomy of hazard functions for the generalized gamma distribution. *Statistics in Medicine* 26:4252-4374.

Gerds T.A and Schumacher M. (2006). Consistent estimation of the expected Brier score in general survival models with right-censored event times, *Biometrical J.*, 6:1029-1040.

Graf E., Schmoor C., Sauerbrei W. and Schumacher M. (1999). Assessment and comparison of prognostic classification schemes for survival data, *Statist. in Medicine*, 18:2529-2545.

Ishwaran H. and Kogalur U.B. (2007). Random survival forests for R, *Rnews*, 7(2):25-31.

Royston P. and Parmar M.K.B. (2002). Flexible parametric proportional-hazards and proportionalodds models for censored survival data, with application to prognostic modelling and estimation of treatment effects, *Statist. in Medicine*, 21::2175-2197.

See Also plot.competing.risk, predict.rfsrc, rfsrc

Examples

```
## Not run:  
## veteran data  
data(veteran, package = "randomForestSRC")  
plot.survival(rfsrc(Surv(time, status)~ ., veteran), cens.model = "rfsrc")  
  
## pbc data  
data(pbc, package = "randomForestSRC")  
pbc.obj <- rfsrc(Surv(days, status) ~ ., pbc, nsplit = 10)  
  
# default spline approach  
plot.survival(pbc.obj, subset = 3)  
plot.survival(pbc.obj, subset = 3, k = 100)
```

```
# three-parameter generalized gamma is approximately the same
# but notice that its CHF estimate (blue line) is not as accurate
plot.survival(pbc.obj, subset = 3, haz.model = "ggamma")

# nonparametric method is too wiggly or undersmooths
plot.survival(pbc.obj, subset = 3, haz.model = "nonpar", span = 0.1)
plot.survival(pbc.obj, subset = 3, haz.model = "nonpar", span = 0.8)

## End(Not run)
```

In the R domain:

```
> install.packages("randomForestSRC")
Installing package into 'C:/Users/Bert/Documents/R/win-library/3.3'
(as 'lib' is unspecified)
--- Please select a CRAN mirror for use in this session ---
```

A **CRAN** mirror is selected.

```
trying URL 'https://mirrors.tuna.tsinghua.edu.cn/CRAN/bin/windows/contrib/3.3/
randomForestSRC_2.5.1.zip'
Content type 'application/zip' length 1310850 bytes (1.3 MB)
downloaded 1.3 MB
```

```
package 'randomForestSRC' successfully unpacked and MD5 sums checked
```

```
The downloaded binary packages are in
C:/Users/Bert/AppData/Local/Temp/Rtmpmeafn5/downloaded_packages
> library(randomForestSRC)
```

```
randomForestSRC 2.5.1
```

Type `rfsrc.news()` to see new features, changes, and bug fixes.

```
> ls("package:randomForestSRC") # Outputting:
[1] "cindex"                      "find.interaction"
[3] "find.interaction.rfsrc"       "get.mv.error"
[5] "get.mv.predicted"            "get.mv.vimp"
[7] "impute"                      "impute.rfsrc"
[9] "max.subtree"                 "max.subtree.rfsrc"
[11] "partial"                     "partial.rfsrc"
[13] "plot.competing.risk"        "plot.competing.risk.rfsrc"
[15] "plot.rfsrc"                  "plot.survival"
[17] "plot.survival.rfsrc"         "plot.variable"
[19] "plot.variable.rfsrc"         "predict.rfsrc"
[21] "print.rfsrc"                 "quantileReg"
[23] "quantileReg.rfsrc"           "rfsrc"
[25] "rfsrc.news"                  "rfsrcSyn"
```

```
[27] "rfsrcSyn.rfsrc"           "stat.split"
[29] "stat.split.rfsrc"         "var.select"
[31] "var.select.rfsrc"         "vimp"
[33] "vimp.rfsrc"
>
> plot.survival # Outputting:
function (x, plots.one.page = TRUE, show.plots = TRUE, subset,
         collapse = FALSE, haz.model = c("spline", "ggamma", "nonpar",
         "none"), k = 25, span = "cv", cens.model = c("km", "rfsrc"),
         ...)
{
  if (is.null(x)) {
    stop("object x is empty!")
  }
  if (sum(inherits(x, c("rfsrc", "grow")), TRUE) == c(1, 2)) !=
    2 & sum(inherits(x, c("rfsrc", "predict")), TRUE) == c(1,
    2)) != 2) {
    stop("This function only works for objects of class '(rfsrc, grow)' or
  '(rfsrc, predict)' .")
  }
  if (x$family != "surv") {
    stop("this function only supports right-censored survival settings")
  }
  if (sum(inherits(x, c("rfsrc", "predict")), TRUE) == c(1,
    2)) == 2) {
    pred.flag <- TRUE
  }
  else {
    pred.flag <- FALSE
  }
  if (is.null(x$predicted.oob)) {
    pred.flag <- TRUE
  }
  haz.model <- match.arg(haz.model, c("spline", "ggamma", "nonpar",
    "none"))
  if (!missing(subset) && haz.model == "spline") {
    if (requireNamespace("glmnet", quietly = TRUE)) {
    }
    else {
      warning("the 'glmnet' package is required for this option: reverting
to 'ggamma' method instead")
      haz.model <- "ggamma"
    }
  }
  cens.model <- match.arg(cens.model, c("km", "rfsrc"))
  if (!is.null(x$yvar) && !is.null(x$imputed.indv)) {
    x$yvar[x$imputed.indv, ] = x$imputed.data[, 1:2]
  }
  event.info <- get.event.info(x)
  if (missing(subset)) {
```

```

subset <- 1:x$n
subset.provided <- FALSE
}
else {
  if (is.logical(subset))
    subset <- which(subset)
  subset <- unique(subset[subset >= 1 & subset <= x$n])
  show.plots <- subset.provided <- TRUE
  if (length(subset) == 0) {
    stop("'subset' not set properly.")
  }
}
if (!pred.flag && !subset.provided && (x$n < 2 | x$ndead <
  1)) {
  stop("sample size or number of deaths is too small for meaningful
analysis")
}
if (is.null(x$predicted.oob)) {
  mort <- x$predicted[subset]
  surv.ensb <- t(x$survival[subset, , drop = FALSE])
  chf.ensb <- x$chf[subset, , drop = FALSE]
  y.lab <- "Mortality"
  title.1 <- "Survival"
  title.2 <- "Cumulative Hazard"
  title.3 <- "Hazard"
  title.4 <- "Mortality vs Time"
}
else {
  mort <- x$predicted.oob[subset]
  surv.ensb <- t(x$survival.oob[subset, , drop = FALSE])
  chf.ensb <- x$chf.oob[subset, , drop = FALSE]
  y.lab <- "OOB Mortality"
  title.1 <- "OOB Survival"
  title.2 <- "OOB Cumulative Hazard"
  title.3 <- "OOB Hazard"
  title.4 <- "OOB Mortality vs Time"
}
if (!subset.provided) {
  surv.mean.ensb <- rowMeans(surv.ensb, na.rm = TRUE)
}
if (subset.provided && collapse) {
  surv.ensb <- rowMeans(surv.ensb, na.rm = TRUE)
  chf.ensb <- rbind(colMeans(chf.ensb, na.rm = TRUE))
}
if (!pred.flag && !subset.provided) {
  km.obj <- matrix(unlist(mclapply(1:length(event.info$time.interest),
    function(j) {
      c(sum(event.info$time >= event.info$time.interest[j],
        na.rm = TRUE), sum(event.info$time[event.info$cens !=
        0] == event.info$time.interest[j], na.rm = TRUE)
    }
  )))
}

```

```

    })), ncol = 2, byrow = TRUE)
Y <- km.obj[, 1]
d <- km.obj[, 2]
r <- d/(Y + 1 * (Y == 0))
surv.aalen <- exp(-cumsum(r))
sIndex <- function(x, y) {
  sapply(1:length(y), function(j) {
    sum(x <= y[j])
  })
}
censTime <- sort(unique(event.info$time[event.info$cens ==
  0]))
censTime.pt <- c(sIndex(censTime, event.info$time.interest))
if (length(censTime) > 0) {
  if (cens.model == "km") {
    censModel.obj <- matrix(unlist(mclapply(1:length(censTime),
      function(j) {
        c(sum(event.info$time >= censTime[j], na.rm = TRUE),
        sum(event.info$time[event.info$cens ==
          0] == censTime[j], na.rm = TRUE))
      })), ncol = 2, byrow = TRUE)
    Y <- censModel.obj[, 1]
    d <- censModel.obj[, 2]
    r <- d/(Y + 1 * (Y == 0))
    cens.dist <- c(1, exp(-cumsum(r)))[1 + censTime.pt]
  }
  else {
    newd <- cbind(x$yvar, x$xvar)
    newd[, 2] <- 1 * (newd[, 2] == 0)
    cens.dist <- t(predict(x, newd, outcome = "test")$survival.oob)
  }
}
else {
  cens.dist <- rep(1, length(censTime.pt))
}
brier.obj <- matrix(unlist(mclapply(1:x$n, function(i) {
  tau <- event.info$time
  event <- event.info$cens
  t.unq <- event.info$time.interest
  cens.pt <- sIndex(t.unq, tau[i])
  if (cens.model == "km") {
    c1 <- 1 * (tau[i] <= t.unq & event[i] != 0)/c(1,
      cens.dist)[1 + cens.pt]
    c2 <- 1 * (tau[i] > t.unq)/cens.dist
  }
  else {
    c1 <- 1 * (tau[i] <= t.unq & event[i] != 0)/c(1,
      cens.dist[, i])[1 + cens.pt]
    c2 <- 1 * (tau[i] > t.unq)/cens.dist[, i]
  }
}))
```

```

(1 * (tau[i] > t.unq) - surv.ensb[, i])^2 * (c1 +
c2)
}}), ncol = length(event.info$time.interest), byrow = TRUE)
brier.score <- matrix(NA, length(event.info$time.interest),
4)
mort.perc <- c(min(mort, na.rm = TRUE) - 1e-05, quantile(mort,
(1:4)/4, na.rm = TRUE))
for (k in 1:4) {
  mort.pt <- (mort > mort.perc[k]) & (mort <= mort.perc[k +
1])
  brier.score[, k] <- apply(brier.obj[mort.pt, , drop = FALSE],
2, mean, na.rm = TRUE)
}
brier.score <- as.data.frame(cbind(brier.score, apply(brier.obj,
2, mean, na.rm = TRUE)))
colnames(brier.score) <- c("q25", "q50", "q75", "q100",
"all")
}
if (subset.provided) {
  sggamma <- function(q, mu = 0, sigma = 1, Q) {
    sigma <- exp(sigma)
    q[q < 0] <- 0
    if (Q != 0) {
      y <- log(q)
      w <- (y - mu)/sigma
      expnu <- exp(Q * w) * Q^-2
      ret <- if (Q > 0)
        pgamma(expnu, Q^-2)
      else 1 - pgamma(expnu, Q^-2)
    }
    else {
      ret <- plnorm(q, mu, sigma)
    }
    1 - ret
  }
  dggamma <- function(x, mu = 0, sigma = 1, Q) {
    sigma <- exp(sigma)
    ret <- numeric(length(x))
    ret[x <= 0] <- 0
    xx <- x[x > 0]
    if (Q != 0) {
      y <- log(xx)
      w <- (y - mu)/sigma
      logdens <- -log(sigma * xx) + log(abs(Q)) + (Q^-2) *
log(Q^-2) + Q^-2 * (Q * w - exp(Q * w)) - lgamma(Q^-2)
    }
    else logdens <- dlnorm(xx, mu, sigma, log = TRUE)
    ret[x > 0] <- exp(logdens)
    ret
  }
}

```

```

hggamma <- function(x, mu = 0, sigma = 1, Q) {
  dggamma(x = x, mu = mu, sigma = sigma, Q = Q)/sgamma(q = x,
    mu = mu, sigma = sigma, Q = Q)
}

haz.list <- mclapply(1:nrow(chf.ensb), function(i) {
  if (haz.model == "ggamma") {
    x <- event.info$time.interest
    y <- t(surv.ensb)[i, ]
    ll <- supsmu(x, y, span = span)
    fn <- function(z) {
      mean((y - sggamma(x, mu = z[1], sigma = z[2],
        Q = z[3]))^2, na.rm = TRUE)
    }
    init <- c(0, 1, 0)
    optim.obj <- optim(init, fn)
    if (optim.obj$convergence != 0)
      warning("fit.ggamma failed to converge")
    parm <- optim.obj$par
    list(x = x, y = hggamma(x, parm[1], parm[2],
      parm[3]))
  }
  else if (haz.model == "spline") {
    tm <- event.info$time.interest
    shift.time <- ifelse(min(tm, na.rm = TRUE) <
      0.001, 0.001, 0)
    log.tm <- log(tm + shift.time)
    shift.chf <- 1
    y <- log(chf.ensb[i, ] + shift.chf)
    k <- max(k, 2)
    knots <- unique(c(seq(min(log.tm), max(log.tm),
      length = k), 5 * max(log.tm)))
    m <- length(knots)
    kmin <- min(knots)
    kmax <- max(knots)
    if (m < 2) {
      stop("not enough knots (confirm that the number of unique
event times > 2")
    }
    x <- do.call(cbind, mclapply(1:(m + 1), function(j) {
      if (j == 1) {
        log.tm
      }
      else {
        lj <- (kmax - knots[j - 1])/(kmax - kmin)
        pmax(log.tm - knots[j - 1], 0)^3 - lj * pmax(log.tm -
          kmin, 0)^3 - (1 - lj) * pmax(log.tm - kmax,
          0)^3
      }
    })))
    cv.obj <- tryCatch({

```

```
glmnet::cv.glmnet(x, y, alpha = 1)
}, error = function(ex) {
  NULL
})
if (!is.null(cv.obj)) {
  coeff <- as.vector(predict(cv.obj, type = "coef",
    s = "lambda.1se"))
}
else {
  warning("glmnet did not converge: setting coefficients to zero")
  coeff <- rep(0, 1 + ncol(x))
}
sfn <- coeff[1] + x %*% coeff[-1]
x.deriv <- do.call(cbind, mclapply(1:m, function(j) {
  lj <- (kmax - knots[j])/(kmax - kmin)
  3 * (pmax(log.tm - knots[j], 0)^2 - lj * pmax(log.tm -
    kmin, 0)^2 - (1 - lj) * pmax(log.tm - kmax,
    0)^2)
)))
sfn.deriv <- coeff[2] + x.deriv %*% coeff[-c(1:2)]
haz <- sfn.deriv * exp(sfn)/(tm + shift.time)
haz[haz < 0] <- 0
haz <- supsmu(tm, haz)$y
haz[haz < 0] <- 0
list(x = tm, y = haz)
}
else if (haz.model == "nonpar") {
  x <- event.info$time.interest[-1]
  y <- pmax(diff(chf.ensb[i, ]), 0)
  haz <- supsmu(x, y, span = span)
  haz$y[haz$y < 0] <- 0
  haz
}
else if (haz.model == "none") {
  NULL
}
})
}
if (show.plots) {
  old.par <- par(no.readonly = TRUE)
  if (plots.one.page) {
    if (pred.flag && !subset.provided) {
      if (!is.null(x$yvar)) {
        par(mfrow = c(1, 2))
      }
      else {
        par(mfrow = c(1, 1))
      }
    }
    else {

```

```
if (haz.model != "none") {
  par(mfrow = c(2, 2))
}
else {
  par(mfrow = c(1, 2))
}
}
else {
  par(mfrow = c(1, 1))
}
par(cex = 1)
if (!subset.provided && x$n > 500) {
  r.pt <- sample(1:x$n, 500, replace = FALSE)
  matplot(event.info$time.interest, surv.ensb[, r.pt],
    xlab = "Time", ylab = title.1, type = "l", col = 1,
    lty = 3, ...)
}
else {
  matplot(event.info$time.interest, surv.ensb, xlab = "Time",
    ylab = title.1, type = "l", col = 1, lty = 3,
    ...)
}
if (!pred.flag && !subset.provided) {
  lines(event.info$time.interest, surv.aalen, lty = 1,
    col = 3, lwd = 3)
}
if (!subset.provided) {
  lines(event.info$time.interest, surv.mean.ensb, lty = 1,
    col = 2, lwd = 3)
}
rug(event.info$time.interest, ticksize = -0.03)
if (plots.one.page) {
  title(title.1, cex.main = 1.25)
}
if (subset.provided) {
  matplot(event.info$time.interest, t(chf.ensb), xlab = "Time",
    ylab = title.2, type = "l", col = 1, lty = 3,
    ...)
if (haz.model != "none") {
  matlines(haz.list[[1]]$x, do.call(cbind, mclapply(haz.list,
    function(ll) {
      cumsum(ll$y * c(0, diff(ll$x)))
    })), type = "l", col = 4, lty = 3, ...)
  rug(event.info$time.interest, ticksize = -0.03)
}
if (plots.one.page) {
  title(title.2, cex.main = 1.25)
}
}
```

```

if (subset.provided && haz.model != "none") {
  plot(range(haz.list[[1]]$x, na.rm = TRUE),
    range(unlist(mclapply(haz.list,
      function(l1) {
        l1$y
      })), na.rm = TRUE), type = "n", xlab = "Time",
    ylab = title.3, ...)
  void <- lapply(haz.list, function(l1) {
    lines(l1, type = "l", col = 1, lty = 3)
  })
  rug(event.info$time.interest, ticksize = -0.03)
  if (plots.one.page) {
    title(title.3, cex.main = 1.25)
  }
}
if (!pred.flag && !subset.provided) {
  matplot(event.info$time.interest, brier.score, xlab = "Time",
    ylab = "OOB Brier Score", type = "l", lwd = c(rep(1,
      4), 2), col = c(rep(1, 4), 2), lty = c(1:4,
      1), ...)
  point.x = round(length(event.info$time.interest) *
    c(3, 4)/4)
  text(event.info$time.interest[point.x], brier.score[point.x,
    1], "0-25", col = 4)
  text(event.info$time.interest[point.x], brier.score[point.x,
    2], "25-50", col = 4)
  text(event.info$time.interest[point.x], brier.score[point.x,
    3], "50-75", col = 4)
  text(event.info$time.interest[point.x], brier.score[point.x,
    4], "75-100", col = 4)
  rug(event.info$time.interest, ticksize = 0.03)
  if (plots.one.page)
    title("OOB Brier Score", cex.main = 1.25)
}
if (!subset.provided && !is.null(x$yvar)) {
  plot(event.info$time, mort, xlab = "Time", ylab = y.lab,
    type = "n", ...)
  if (plots.one.page) {
    title(title.4, cex.main = 1.25)
  }
  if (x$n > 500)
    cex <- 0.5
  else cex <- 0.75
  points(event.info$time[event.info$cens != 0], mort[event.info$cens != 0],
    pch = 16, col = 4, cex = cex)
  points(event.info$time[event.info$cens == 0], mort[event.info$cens == 0],
    pch = 16, cex = cex)
  if (sum(event.info$cens != 0) > 1)
    lines(supsmu(event.info$time[event.info$cens != 0][order(event.info$time[event.info$cens != 0])])

```

```

0]]], mort[event.info$cens != 0][order(event.info$time[event.
info$cens !=
0]]]), lty = 3, col = 4, cex = cex)
if (sum(event.info$cens == 0) > 1)
  lines(supsmu(event.info$time[event.info$cens ==
0][order(event.info$time[event.info$cens ==
0]]]), mort[event.info$cens == 0][order(event.info$time[event.
info$cens ==
0]]]), lty = 3, cex = cex)
  rug(event.info$time.interest, ticksize = -0.03)
}
par(old.par)
}
if (!pred.flag && !subset.provided) {
  Dint <- function(f, range, grid) {
    a <- range[1]
    b <- range[2]
    f <- f[grid >= a & grid <= b]
    grid <- grid[grid >= a & grid <= b]
    m <- length(grid)
    if ((b - a) <= 0 | m < 2) {
      0
    }
    else {
      (1/(2 * diff(range))) * sum((f[2:m] + f[1:(m -
1)]) * diff(grid))
    }
  }
  invisible(cbind(time = event.info$time.interest, brier.score,
  integrate = unlist(mclapply(1:length(event.info$time.interest),
  function(j) {
    Dint(f = brier.score[1:j, 4], range = quantile(event.info$time.
interest,
  probs = c(0.05, 0.95), na.rm = TRUE), grid = event.info$time.
interest[1:j])
  })))
}
<environment: namespace:randomForestSRC>
>
>
> ## Not run:
> ## veteran data
> data(veteran, package = "randomForestSRC")
> plot.survival(rfsrc(Surv(time, status)~ ., veteran), cens.model = "rfsrc")
> # Outputting: Figure 34A
>
> ## pbc data
> data(pbc, package = "randomForestSRC")
> pbc.obj <- rfsrc(Surv(days, status) ~ ., pbc, nsplit = 10)

```

```

>
> # default spline approach
> plot.survival(pbc.obj, subset = 3)
> # Outputting: Figure 34B
>
> plot.survival(pbc.obj, subset = 3, k = 100)
> # Outputting: Figure 34C
>
> # three-parameter generalized gamma is approximately the same
> # but notice that its CHF estimate (blue line) is not as accurate
> plot.survival(pbc.obj, subset = 3, haz.model = "ggamma")
> # Outputting: Figure 34D
>
> # nonparametric method is too wiggly or undersmooths
> plot.survival(pbc.obj, subset = 3, haz.model = "nonpar", span = 0.1)
> # Outputting: Figure 34E
>
> plot.survival(pbc.obj, subset = 3, haz.model = "nonpar", span = 0.8)
> # Outputting: Figure 34F
>
>
>

```

Fig. 34A Survival and mortality plots: `plot.survival(rfsrc(Surv(time, status)~., veteran), cens.model = "rfsrc")`

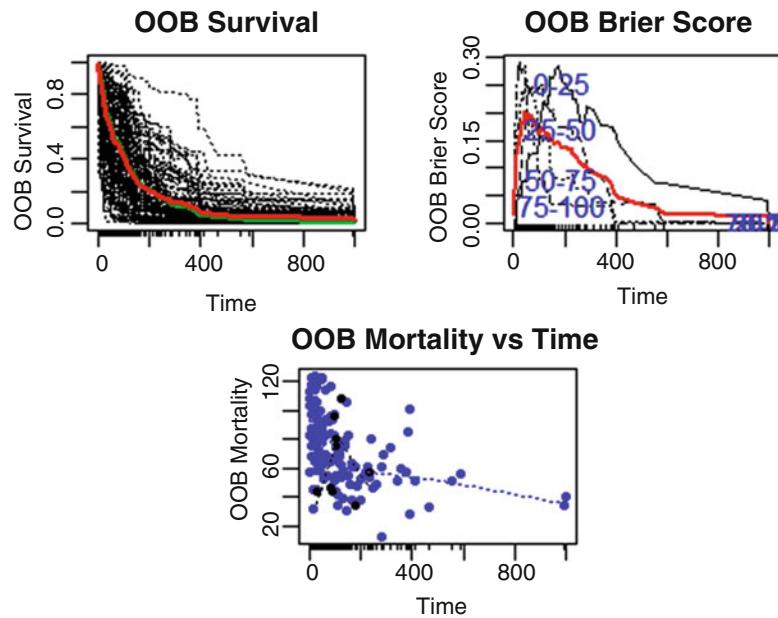


Fig. 34B Survival and hazard plots using splines:
plot.survival(pbc.obj,
subset = 3)

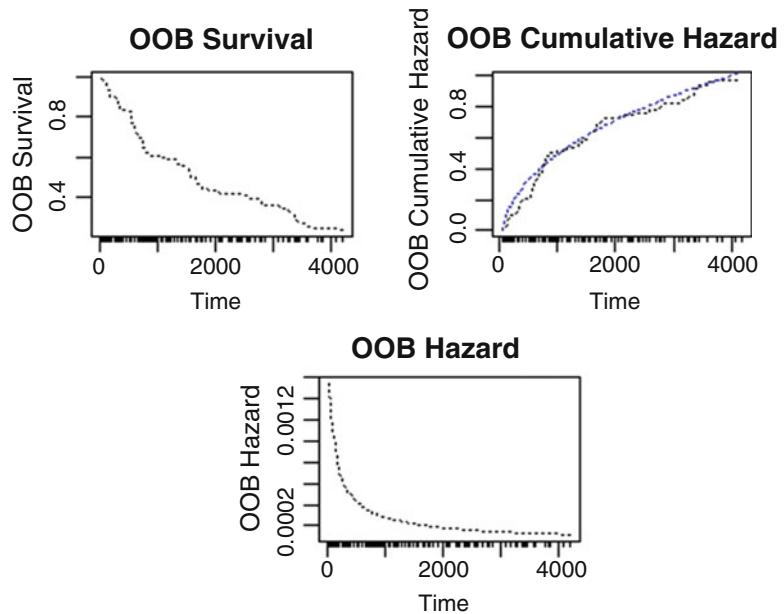


Fig. 34C Survival plots
plot.survival(pbc.obj,
subset = 3, k = 100)

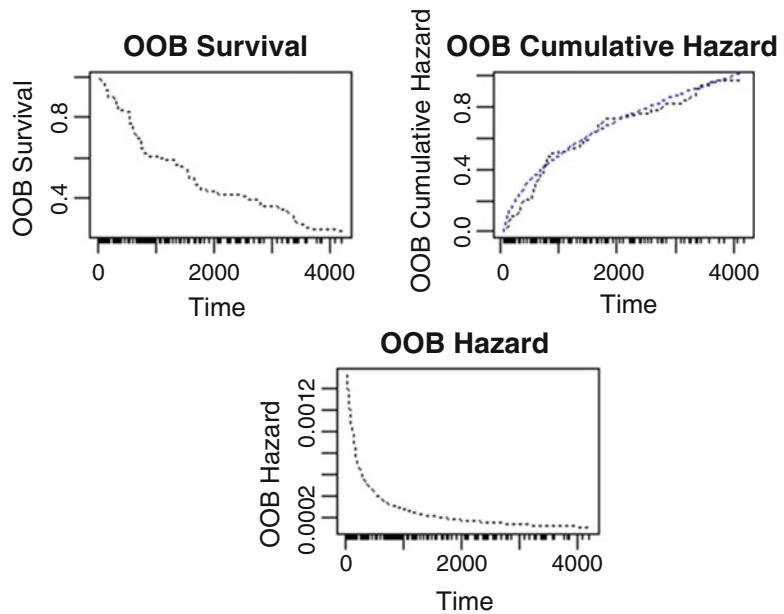


Fig. 34D 3-parameter survival plots plot.survival (pbc.obj, subset = 3, haz. model = "ggamma")

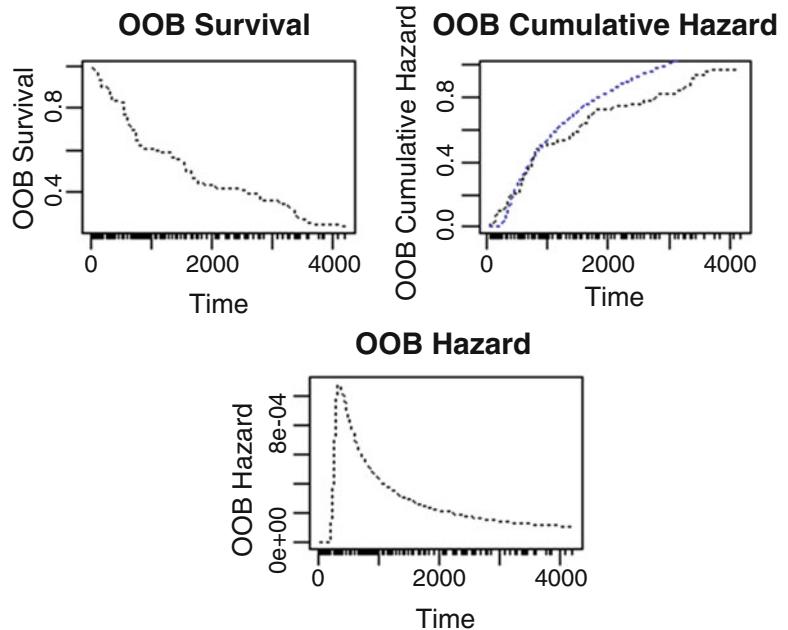


Fig. 34E Non-parametric survival plots plot.survival (pbc.obj, subset = 3, haz. model = "nonpar", span = 0.1)

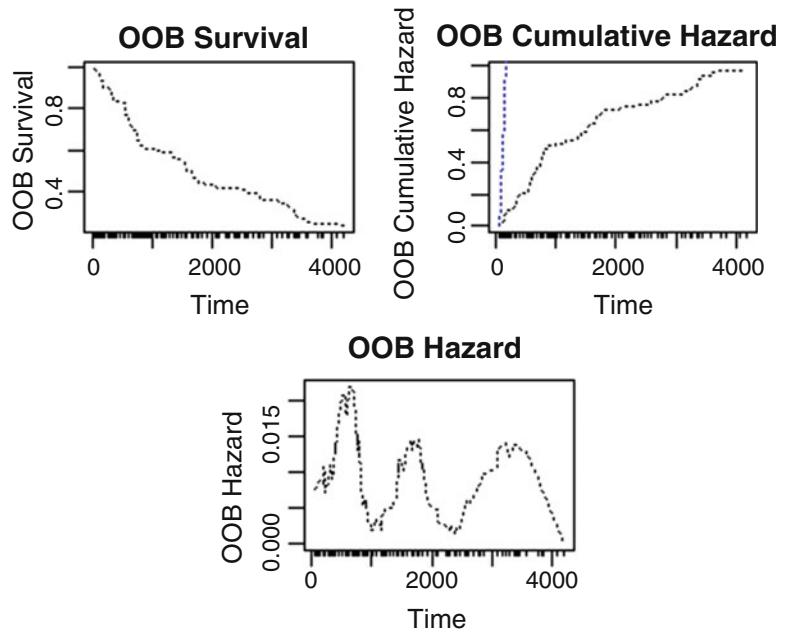
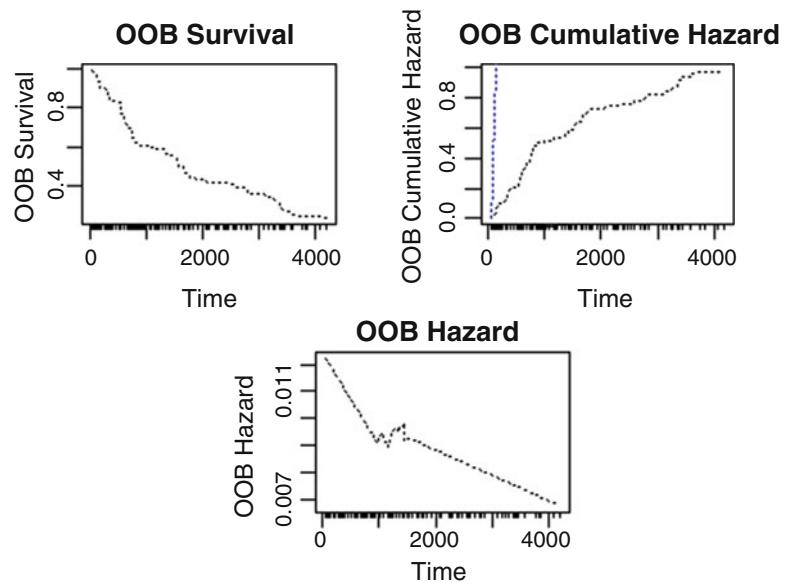


Fig. 34F Non-parametric survival plots, with smoothing plot.survival (pbc.obj, subset = 3, haz. model = "nonpar", span = 0.8)



References

- Austin MA et al (2013b) Genetic epidemiology, methods and applications. CABI, Boston
Chan BKC (2015) Biostatistics for epidemiology and public health using R. Springer Publishing, New York
Elston RC, Johnson WD (2008b) Basic biostatistics for geneticists. United Kingdom
Foulkes AS (2009b) "Applied statistical genetics with R" in the Springer "Use R!" series. Springer, New York
Thomas DC (2004b) Statistical methods in genetic epidemiology. Oxford University Press, New York

References¹

A

- Altshuler D, Daly MJ, Lander ES (2008) *Science* 322(5903):881–888
- Amberger JS, Bocchini CA, Schietecatte F, Scott AF, Hamosh (2017) OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders, by American Diabetes Association. <http://diabetes.org>
- Austin MA (2013) *Genetic epidemiology: methods and applications*, Modular Text Series. CABI Publishing, Wellingford <http://www.foxnews.com/health/2017/03/01/could-alzheimers-really-bankrupt-medicare-and-medicaid.html>
- Austin MA, Beaty TH, Dotson WD, Edwards K, Fullerton SM, Gwinn M, Khouri MJ, McKnight B, Ottman R, Psaty BM, Schwartz SM, Stanford JL, Thornton TA (2013) *Genetic epidemiology, methods and applications*. CABI, Boston, p 02111

B

- Ben-Shlomo Y, Brookes ST, Hickman M (2013) *Epidemiology, evidence-based medicine and public health*, 6/e, Wiley-Blackwell, Hoboken, NJ07030. <https://www.betterhealth.vic.gov.au/health/conditionsandtreatments/genes-and-genetics>
- Bondy ML (1990) *Genetic epidemiology of childhood brain tumors*. Texas Medical Center dissertations (via ProQuest). AA119109972. <http://digitalcommons.library.tmc.edu/dissertations/AA19109972>
- Bredesen DE (2017) *The end of Alzheimer's – the first program to prevent and reverse cognitive decline*, Avery – an Imprint of Penguin Random House LLC, New York, NY 10014
- Bull SB, Andrusis IL Paterson AD (2018) Statistical challenges in high-dimensional molecular and genetic epidemiology. *Can J Stat* 46(Special Issue on “Big Data and the Statistical Sciences”):24–40

C

- Carrington M, Hoelzel AR (eds) (2001) *Molecular epidemiology –practical approach*. Oxford University Press, New York
- Chan B (1978) *A new school mathematics for Hong Kong*. Ling Kee Publishing Co., Hong Kong 10 Volumes: 1A, 1B, 2A, 2B, 3A, 3B, 4A, 4B, 5A, 5 B 6 Workbooks: 1A, 1B, 2A, 2B, 3A, 3B
- Chan BKC (2016) *Biostatistics for epidemiology and public health using R*. Springer, New York (with additional materials on the Publisher's website)
- Chan BKC (2017) *Applied probabilistic calculus for financial engineering: an introduction using R*. Wiley, Hoboken

¹ References from <http://www.foxnews.com/health/2017/03/01/could-alzheimers-really-bankrupt-medicare-and-medicaid.html>

- Choi M, Scholl UI et al (2009) Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proc Natl Acad Sci U S A* 106(45):19096–19101 [PubMed]
- Cui JS (2003) Regressive logistic and proportional hazards disease models for within-family analyses of measured genotypes, with application to a CYP17 polymorphism and breast cancer. *Genet Epidemiol* 24:161–172

E

- Elston RC, Johnson WD (2008) Basic biostatistics for geneticists and epidemiologists, a practical approach. Wiley, The Atrium, Southern Gatet, Chichester, West Sussex PO19 SSQ, United Kingdom
- Elston RC, Stewart J A general model for the genetic analysis of pedigree data. *Hum Hered* 21(6):523–542 [PubMed]

F

- Foulkes AS (2009) Applied statistical genetics with R for population-based association studies. Springer Use R! Series. New York. <http://www.springer.com/series/6991>
- Foulkes AS, Yucel R Reilly MP <https://www.ncbi.nlm.nih.gov/por>

G

Genetic Epidemiology Review References:

- 1000 Genomes www.1000genomes.org
- Aird I, Bentall HH, Mehigan JA, Roberts JA (1954) *Br Med J* 2:315
- Albert TJ et al (2007) *Nat Methods* 4:903
- Altenburg E, Muller HJ (1920) *Genetics* 5:1
- Altshuler D, Daly M (2007) *Nat Genet* 39:813
- Altshuler D et al (2000) *Nat Genet* 26:76
- Amundadottir LT et al (2006) *Nat Genet* 38:652
- Antonarakis SE, Boehm CD, Giardina PJ, Kazazian HH (1982) *J Proc Natl Acad Sci USA* 79:137
- Ardlie KG, Kruglyak L, Seielstad M (2002) *Nat Rev Genet* 3:299
- Barrett JC et al (2008) *Nat Genet* 40:955
- Bell GI, Polonsky KS (2001) *Nature* 414:788
- Bender W et al (1983) *Science* 221:23
- Botstein D, White RL, Skolnick M, Davis RW (1980) *Am J Hum Genet* 32:314
- Burton PR et al (2007) *Nat Genet* 39:1329
- Cargill M et al (2007) *Am J Hum Genet* 80:273
- Chapman JM, Cooper JD, Todd JA, Clayton DG (2003) *Hum Hered* 56:18
- Clarke L, Carbon J (1980) *Proc Natl Acad Sci U S A* 77:2173
- Clayton DG et al (2005) *Nat Genet* 37:1243
- Cohen JC et al (2004) *Science* 305:869
- Cohen J et al (2005) *Nat Genet* 37:161
- Cohen JC, Boerwinkle E, Mosley TH Jr, Hobbs HH (2006) *N Engl J Med* 354:1264
- Collins FS, Guyer MS, Chakravarti A (1997) *Science* 278:1580
- Couzin J (2002) *Science* 296:1391
- Crawford DC et al (2004) *Nat Genet* 36:700
- Daly MJ, Rioux JD, Schaffner SF, Hudson TJ, Lander ES (2001) *Nat Genet* 29:229
- de Bakker PIW et al (2005) *Nat Genet* 37:1217
- Deeb SS et al (1998) *Nat Genet* 20:284
- Devlin B, Roeder K (1999) *Biometrics* 55:97

- Diabetes Genetics Initiative of Broad Institute of Harvard MIT, Lund University, Novartis Institutes for BioMedical Research (2007) *Science* 316:1331. (Published online 26 April 2007)
- Dib C et al (1996) *Nature* 380:152
- Donis-Keller H et al (1987) *Cell* 51:319
- Duerr RH et al (2006, 1461) *Science*:314
- East E (1910) *Am Nat* 44:65
- Easton DF et al (2007) *Nature* 447:1087
- Edwards AO et al (2005) *Science* 308:421
- Emilsson V et al (2008) *Nature* 452:423
- Entrez SNP www.ncbi.nlm.nih.gov/sites/entrez?db=snp
- Fisher RA (1918) *Trans R Soc Edinburgh* 152:399
- Florez JC, Hirschhorn JN, Altshuler D (2003) *Annu Rev Genomics Hum Genet* 24:257
- Florez JC et al (2006) *N Engl J Med* 355:241
- Folsom AR et al (2008) *Diabetes Care* 31:905
- Freedman ML et al (2006) *Proc Natl Acad Sci U S A* 103:14068
- Gabriel SB et al (2002) *Science* 296:2225
- Graham RR et al (2007) *Proc Natl Acad Sci U S A* 104:6758
- Grant SFA et al (2006) *Nat Genet* 38:320
- Grarup N et al (2007) *Diabetes* 56:3105
- Gudbjartsson DF et al (2008) *Nat Genet* 40:609
- Gudmundsson J et al (2007) *Nat Genet* 39:977
- Gusella JF et al (1983) *Nature* 306:234
- Haiman CA et al (2007) *Nat Genet* 39:638
- Haines JL et al (2005) *Science* 308:419
- Harris H (1996) *Proc R Soc London, Ser B* 164:298
- Hastbacka J et al (1992) *Nat Genet* 2:204
- Hazra A et al (2008) *Cancer Causes Control* 19:975
- Helgadottir A et al (2007) *Science* 316:1491
- Hinds DA et al (2005, 1072) *Science*:307
- Hinds DA, Kloek AP, Jen M, Chen X, Frazer KA (2006) *Nat Genet* 38:82
- Hirschhorn JN, Altshuler DJ (2002) *Clin Endocrinol: Metab* 87:4438. [PubMed]
- Hirschhorn JN, Lohmueller K, Byrne E, Hirschhorn K (2002) *Genet Med* 4:45
- Hudson TJ et al (1945) 1995. *Science* 270
- Iafrate AJ et al (2004) *Nat Genet* 36:949
- Ingram VM (1956) *Nature* 178:792
- International HapMap Consortium (2005) *Nature* 437:1299
- International HapMap Consortium (2007) *Nature* 449:851
- Jeffreys AJ (1979) *Cell* 18:1
- Ji W et al (2008) *Nat Genet* 40:592
- Johnson GC et al (2001) *Nat Genet* 29:233
- Kan YW, Dozy AM (1978) *Proc Natl Acad Sci USA* 75:5631
- Kathiresan S et al (2008) *Nat Genet* 40:189
- Kerem B et al (1989) *Science* 245:1073
- Kimura M, Ota T (1973) *Genetics* 75:199
- Klein J, Sato AN (2000) *New Engl J Med* 343:782
- Klein RJ et al (2005) *Science* 308:385
- Knowler WC, Williams RC, Pettitt DJ, Steinberg AG (1988) *Am J Hum Genet* 43:520
- Kotowski IK et al (2006) *Am J Hum Genet* 78:410
- Kruglyak L (1999) *Nat Genet* 22:139
- Lamb J et al (2006) *Science* 313:1929
- Lander ES (1996) *Science* 274:536
- Lander ES, Botstein D (1986) *Cold Spring Harb Symp Quant Biol* 51:49
- Lettre G et al (2008) *Nat Genet* 40:584
- Lewontin R (1972) In: Dobzhansky T, Hecht MK, Steere WC (eds) *Evolutionary Biology*, vol 6. Appleton-Century-Crofts, New York, pp 391–398
- Li WH, Sadler LA (1991) *Genetics* 129:513
- Li M et al (2006) *Nat Genet* 38:1049
- Lifton RP (2004) *Harvey Lect* 100:71
- Locke DP et al (2006) *Am J Hum Genet* 79:275

- Maller J et al (2006) *Nat Genet* 38:1055
- Marchini J, Howie B, Myers S, McVean G, Donnelly P (2007) *Nat Genet* 39:906
- McCarroll SA et al (2006) *Nat Genet* 38:86
- McCarroll SA et al (2008) *Nat Genet* 40:1166
- McCarroll SA et al (2008) *Nat Genet* 40:1107
- McPherson R et al (2007) *Science* 316:1488
- McVean GA et al (2004) *Science* 304:581
- Mikkelsen TS et al (2007) *Nature* 448:553
- Online "Mendelian Inheritance in Man". www.ncbi.nlm.nih.gov/sites/entrez?db=omim
- Pascoe L et al (2007) *Diabetes* 56:3101
- Paterson AH et al (1988) *Nature* 335:721
- Patil N et al (2001) *Science* 294:1719
- Pearce CL, Pike M, Lander ES, Hirschhorn NJ (2003) *Nat Genet* 33:177
- Pe'er I, Yelensky R, Altshuler D, Daly MJ (2008) *Genet Epidemiol* 32:381
- Petes TD, Botstein D (1977) *Proc Natl Acad Sci U S A* 74:5091
- Price AL et al (2006) *Nat Genet* 38:904
- Pritchard JK, Rosenberg NA (1999) *Am J Hum Genet* 65:220
- Redon R et al (2006) *Nature* 444:444
- Reich DE, Lander ES (2001) *Trends Genet* 17:502
- Reich DE et al (2002) *Nat Genet* 32:135
- Reiner AP et al (2008) *Am J Hum Genet* 82:1193
- Ridker PM et al (2008) *Am J Hum Genet* 82:1185
- Rioux JD et al (2007) *Nat Genet* 39:596
- Risch N, Merikangas K (1996) *Science* 273:1516
- Rivera A et al (2005) *Hum Mol Genet* 14:3227
- Sachidanandam R et al (2001) *Nature* 409:928
- Samani NJ et al (2007) *N Engl J Med* 357:443
- Sanna S et al (2008) *Nat Genet* 40:198
- Saxena R et al (2006) *Diabetes* 55:2890
- Scott LJ et al (2007, 1341) *Science*:316
- Sebat J et al (2004) *Science* 305:525
- Sebat J et al (2007) *Science* 316:445
- Sladek R et al (2007) *Nature* 445:881
- Spielman RS, McGinnis RE, Ewens WJ (1993) *Am J Hum Genet* 52:506
- Staiger H et al (2007) *PLoS One* 2:e832
- Steinthorsdottir V et al (2007) *Nat Genet* 39:770
- Strittmatter WJ, Roses AD (1996) *Annu Rev Neurosci* 19:53
- Sturtevant A (1913) *J Exp Zool* 14:43
- Thomas G et al (2008) *Nat Genet* 40:310
- Thorleifsson G et al (2007) *Science* 317:1397
- Tishkoff SA, Verrelli BC (2003) *Annu Rev Genomics Hum Genet* 4:293
- Tuzun E et al (2005) *Nat Genet* 37:727
- Walsh T et al (2008) *Science* 320:539
- Wang DG et al (1998) *Science* 280:1077
- Waterston RH et al (2002) *Nature* 420:520
- Weedon MN et al (2008) *Nat Genet* 40:575
- Weiss KM, Terwilliger JD (2000) *Nat Genet* 26:151
- Weiss LA et al (2008) *N Engl J Med* 358:667
- Welch PL, King MC (2001) *Hum Mol Genet* 10:705
- Wellcome Trust Case Control Consortium (2007) *Nature* 447:661
- Willer CJ et al (2008) *Nat Genet* 40:161
- Winckler W et al (2007) *Diabetes* 56:685
- Wolf N et al (2008) *J Med Genet* 45:114
- Zeggini E et al (2007) *Science* 316:1336
- Zeggini E et al (2008) *Nat Genet* 40:638

[G] Google

G & G

<https://www.betterhealth.vic.gov.au/health/conditionsandtreatments/genes-and-genetics0016>

H

Hahne F, Huber W, Gentleman R, Falcon S (2008) Bioconductor case studies, Springer Use R! Series, New York, NY 10013. <http://www.springer.com/series/6991>

J

JDRF (Juvenile Diabetes Research Foundation) (2017) “What is Type-1 Diabetes (T1D)?”, jdrf.org

K

Kennedy GC, Matsuzaki H et al (2003) Large-scale genotyping of complex DNA. *Nat Biotechnol* 21(10):1233–1237 PubMed

Klein RJ, Zeiss C et al (2005) Complement factor H polymorphism in age-related macular degeneration. *Science* 308 (5720):385–389. [PubMed]

L

Lander ES, Linton LM et al (2001) Initial sequencing and analysis of the human genome. *Nature* 409(6822):860–921 PubMed

Lupski JR, Reid JG et al (2010) Whole-genome sequencing in a patient with Charcot-Marie-Tooth Neuropathy. *N Engl J Med* 362:1181–1191 PubMed

M

Matsuzaki H, Dong S et al (2004) Genotyping over 100,000 SNPs on a pair of oligonucleotide arrays. *Nat Methods* 1 (2):109–111. [PubMed]

Matsuzaki H, Loi H et al (2004) Parallel genotyping of over 10,000 SNPs using a one-primer assay on a high-density oligonucleotide array. *Genome Res* 14(3):414ff. [PubMed]

Molecular Epidemiology Review References

What is molecular epidemiology? aacr.org. Retrieved 2008-02-19

Field N (2014) Strengthening the reporting of molecular epidemiology for infectious diseases (STROME-ID): an extension of the STROBE statement. *Lancet Infect Dis* 14(4):341–352. [https://doi.org/10.1016/S1473-3099\(13\)70324-4](https://doi.org/10.1016/S1473-3099(13)70324-4) PMID 24631223

- Goering R (6 August 2010). Pulsed field gel electrophoresis: A review of application and interpretation in the molecular epidemiology of infectious disease. *Infect Genet Evol* 10(7): 866–875. doi:<https://doi.org/10.1016/j.meegid.2010.07.023>. PMID 20692376.
- Kilbourne ED (Apr 1973). The molecular epidemiology of influenza. *J Infect Dis.* 127 (4): 478–487. doi:<https://doi.org/10.1093/infdis/127.4.478>. PMID 4121053.
- Kuller LH (2012) Invited commentary: the 21st century epidemiologist—a need for different training? *Am J Epidemiol* 176(8):668–671
- Morton NE (1955) Sequential tests for the detection of linkage. *Am J Hum Genet* 7(3):277–318
- Morton NE, Chung CS (1975) Genetic epidemiology. Academic New York, New York
- Ogino S, King EE, Beck AH, Sherman ME, Milner DA, Giovannucci E (2012) Interdisciplinary education to integrate pathology and epidemiology: Towards molecular and population-level health science. *Am J Epidemiol* 176:659–667
- Ogino S, Beck AH, King EE, Sherman ME, Milner DA, Giovannucci E (2012) Ogino et al. respond to “The 21st century epidemiologist”. *Am J Epidemiol* 176:672–674
- Ogino S, Lochhead P, Chan AT, Nishihara R, Cho E, Wolpin BM, Meyerhardt AJ, Meissner A, Schernhammer ES, Fuchs CS, Giovannucci E (2013) Molecular pathological epidemiology of epigenetics: emerging integrative science to analyze environment, host, and disease. *Mod Pathol* 26:465–484
- Porta M (2002) Incomplete overlapping of biological, clinical, and environmental information in molecular epidemiological studies: a variety of causes and a cascade of consequences. *J Epidemiol Community Health* 56(10):734–738. <https://doi.org/10.1136/jech.56.10.734> PMID 1732039
- Porta M (ed) Greenland S, Hernán M, dos Santos Silva I, Last JM (associate eds) (2014). *A dictionary of epidemiology*, 6th edn. Oxford University Press, New York. ISBN 9780199976737
- Schulte PA, Perera FP (1993) Molecular epidemiology: principles and practice. Academic, p 588. ISBN 0-12-632346-1
- Sherman ME, Howatt W, Blows FM, Pharoah P, Hewitt SM, Garcia-Closas M (2010) Molecular pathology in epidemiologic studies: a primer on key considerations. *Cancer Epidemiol Biomark Prev* 19(4):966–972
- Slattery M (2002) The science and art of molecular epidemiology. *J Epidemiol Community Health* 56(10):728–729. <https://doi.org/10.1136/jech.56.10.728> PMID 1732025
- Tevfik Dorak M (2008-03-03) *Introduction to genetic epidemiology*. *J Stat Softw* now at its 45th volume, is available at <http://www.jstatsoft.org/v45>

N

- Neel JV, Schull WJ (1954) Human heredity. University of Chicago Press, Chicago
- Newschaffer CJ et al (2007) The epidemiology of autism spectrum disorders. *Annu Rev Public Health* 28:235–258. <https://doi.org/10.1146/annurev.pubhealth.28.021406.144007> <http://edition.cnn.com/2017/04/05/health/autism-cord-blood-stem-cells-duke-study/index.html>
- No. 3, Big Science Media, LLC, 6900 Dallas Parkway, Suite 200, Plano, TX 75024
- Nucleic Acids Res 2015 Jan 28; 43(Database issue): D789–D798. Published online 2014 Nov 26. <https://doi.org/10.1093/nar/gku1205>
- Nussbaum RL, McInnes RR, Willard HF, Hamosh A (2016) Thompson & Thompson: genetics in medicine, 8/e, Elsevier, Philadelphia, PA 19103 “Genome – Your Health is Personal”, Fall 2015, ISSN 2374-5800, vol 2

O

- Ott J (1974) Estimation of the recombination fraction in human pedigrees: efficient 97 computation of the likelihood for human linkage studies. *Am J Hum Genet* 26(5):588
- Ozaki K, Ohnishi Y et al (2002) Functional SNPs in the lymphotoxin-alpha gene that are associated with susceptibility to myocardial infarction. *Nat Genet* 32(4):650–654 PubMed

P

- Parkinson’s Disease <https://www.healthline.com/health/what-causes-parkinsons-disease#loss-of-dopamine>

Q

- Ott J (1974) Estimation of the recombination fraction in human pedigrees: efficient computation of the likelihood for human linkage studies. *Am J Hum Genet* 26(5):588–597
- Qian J, Nunez S, Reed E, Reilly MP, Foulkes AS (2016) A simple test of class-level genetic association can reveal novel cardiometabolic trait loci. *PLoS ONE* 11(2):e0148218. <https://doi.org/10.1371/journal.pone.0148218>

R

- Risch N, Merikangas K (1996) The future of genetics studies of complex human diseases. *Science* 273(5281):1516–1517
PubMed
- Roach JC, Glusman G et al (2010) Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science* 2010 [PubMed]

S

- Sinnott EW, Dunn LC, Dobzhansky T (1950) Principles of genetics, 4/e, McGraw-Hill Publications in the Botanical Sciences Series. In: Sinnott EW (ed). McGraw-Hill Book Company, Inc., New York

T

- Teare DM, Barrett JH (2005) Genetic linkage studies. *Lancet* 366(9490):1036–1044
- Thomas DC (2004) Statistical methods in genetic epidemiology. Oxford University Press, New York, NY 10016. <http://www.oup.com>

U

- “What is molecular epidemiology?”. Molecular epidemiology homepage. University of Pittsburgh. 28 July 1998.
Retrieved 15 Jan 2010

V

- Venter JC, Adams MD et al (2001) The sequence of the human genome. *Science* 291(5507):1304–1351 PubMed

W

- Weitzman J, Weitman M (eds) (2017) “30-Second Genetics – The 50 Most Revolutionary Discoveries in Genetics, Each Explained in Half a Minute”, Metro Books – An Imprint of Sterling Publishing Co., Inc., New York, NY 10036
- White EG (1905) The ministry of healing. pp 261–266. <http://whiteestate.org/search/search.asp>

[W] Wikipedia

Williams WR, Anderson DE (1984) Genetic epidemiology of breast cancer: segregation analysis of 200 Danish pedigrees. *Genet Epidemiol* 1:7–20

X

Xinyu Tang, Daniel E. Weeks weeks@pitt.edu <http://www.genabel.org>, <http://forum.genabel.org> <http://genabel.r-forge.r-project.org/>

Y

Yale Journal of Biology and Medicine (DeWan AT (2010) Yale J Biol Med 83(2): 87–90, published online 2010 June) introduced “Five Classic Articles in Genetic Epidemiology”.

Z

Zhao YD, Rahardja D, Qu Y (2008) Sample size calculation for the Wilcoxon-Mann-Whitney test adjusting for ties. *Stat Med* 27:462–468

Zhao YD jinghua.zhao@mrc-epid.cam.ac.uk

General References for Human Genetic Epidemiology

<http://www.genabel.org>, <http://forum.genabel.org>

<http://genabel.r-forge.r-project.org/>

BugReports http://r-forge.r-project.org/tracker/?group_id=505

European Genome-phenome Archive, EGA

<https://www.ebi.ac.uk/ega/home>

The EGA archives a large number of datasets, the access to which is controlled by a Data Access Committee (DAC)

The European Bioinformatics Institute <MBL-EBI

<https://www.ebi.ac.uk/>

EGA European Genome-Phenome Archive

<https://ega.crg.eu/>

European Bioinformatics Institute (EMBL-EBI)

<https://www.facebook.com/EMBLEBI>

The European Nucleotide Archive (ENA)

<http://www.ebi.ac.uk/ena>

NHGRI-Related News Archive - National Human Genome

...

<https://www.genome.gov/19016944/NHGRIRelated-News-Archive>

The European Bioinformatics Institute (EMBL-EBI)

The New York University School of Medicine and
The Ontario Institute for Cancer Research (OICR)

The University of California Santa Cruz: Investigators at the UC Santa Cruz Genomics Institute have optimized performance of a mobile-phone-sized MinION™ ...

Wellcome Library | Collecting Genomics

<https://wellcomelibrary.org/what-we-do/.../collecting-genomics/>

The American Society of Human Genetics

URL: http://www.ashg.org/education/gena/NatureNurture_L2_corrected.pdf

URL <https://github.com/shearer/samplesize>

BugReports <https://github.com/shearer/samplesize/issues>

Index

A

- Adoption studies, 126, 182, 183
- Alleles, 14, 126, 148, 231
- Alzheimer disease (AzD), x, 26
- Applied human genetic epidemiology, 145–214
- Applied statistical human genetics, 46
- Applied statistics in epidemiology, 123, 124
- Association studies, 5, 45, 46, 124, 130, 136–142, 147, 148, 151, 152, 156, 157, 160, 162, 192, 195, 198–204, 212, 217, 231, 232, 265, 303, 321, 344
- Autisms, 19, 20, 34, 42, 160, 161
- Autism spectrum disorders (ASD), viii, 34
- Autosomes, 10, 13, 25, 193

B

- Bias, 178, 179, 205, 314, 321, 322
- Big data, 142–144
- Big data and human genomics, 142–144
- Bioinformatics, 7, 8, 10, 218
- Biostatistical concepts, 124–136
- Biostatistical human genetics, 43, 219–230
- Biostatistics, 12, 43, 46, 54, 72, 123, 124, 170, 171, 200, 219, 263
- BRCA1* gene, 20, 131, 162, 186, 211
- Breast cancer, 11, 20, 126–128, 131, 143, 151, 159, 205, 211, 311, 313, 320

C

- Cancer, x, viii, 4, 5, 8, 9, 11, 14, 15, 19, 20, 24, 26, 36–38, 43, 73, 88, 126–128, 131, 143, 149, 151, 157–160, 171, 178, 186, 203–205, 207–209, 211, 311, 313, 320, 344
- genetics and genomics, 36, 37
- heritability, 178, 208, 213
- Candidate gene studies, 160
- Case subjects, viii, 26, 28, 30, 38, 40, 73, 88, 125, 126, 165, 176, 184, 188–190, 192, 198, 205, 222, 231, 311
- Charcot-Marie-Tooth (CMT) disease, ix
- Chromosomes, 9, 128, 166, 231
- Cohort studies, 126, 172, 212, 316, 322, 324
- Colorectal cancer, 125, 157
- Complex traits in humans, 186, 187

D

- Data, 5, 48, 125, 146, 218
- Data analysis using R programming, 47–122
- Dataset, 48, 59, 73, 78, 82, 87, 88, 90, 91, 97, 100, 102, 106, 142, 148, 218, 309, 314, 321
- Derivation of familial risk, 167–169
- Diabetes
 - Type-1,2, viii–xi, 157–161, 164, 203, 206, 311, 315, 320, 321, 323
- Disease, 2, 73, 124, 146, 231
- Disequilibrium, 130, 152, 190, 193–197, 200–203, 206, 210, 231, 232, 275, 314, 321
- DNA, 4, 6, 8–11, 13, 20, 21, 23, 24, 30, 88, 136, 143, 144, 149, 150, 153–156, 163, 165, 178, 181–183, 185, 190, 191, 193, 206, 210–213
- DNA sequencing, 21, 30, 144, 183, 191, 231

E

- Environmental factors, 15, 26, 35, 38, 44, 125, 164, 167, 169, 186, 207, 321
- Epidemiology, 5, 54, 123, 146, 217
- Ethnicity, 184, 205
- Extreme trait sequencing, 192, 197, 198

F

- Familial aggregation studies, 45, 46, 124–126, 165
- Familial factors, 198
- Familial risk, 167–173
- Family, 1, 98, 125, 146, 220
 - association studies, 200–203
 - studies, 126, 148–164, 190, 198, 199, 316, 324

G

- Gaussian distribution, 176
- Gene structure, 190
- Gene structure and genetic code, 190
- Genetic, 4, 72, 124, 146, 218
 - big data, 142, 143
 - code, 14, 188, 190, 312
 - epidemiology, 1–46, 87–100, 124, 127, 128, 145–212, 311–324
 - epidemiology and public health, 209–215
 - linkage, 128, 146, 150, 190, 193, 198, 231

Genetic (cont.)

- mapping, 149–164
- medicine, 18, 23, 40
- variants, 9, 151, 158, 162, 163, 180, 191, 192, 197, 203, 210, 214

Genome, 5, 88, 124, 146, 217

Genome-wide association studies (GWAS), 5, 136–142, 152, 158–161, 163, 192, 204, 205, 212–214, 252, 299, 313, 314, 316, 321, 324

Genomics for human genetic epidemiology, 183–198

Genotype, 14, 124, 146, 222

H

Hardy-Weinberg Principle, 187

Heritability analysis, 174–181

Human genetic epidemiology (HGE), 1–46, 87–100, 145–215, 217–341

Human genetics, xi, 1, 87, 123, 146, 217

- association, 204, 231
- influences on diseases, 165–183
- variation, 231–251

Human genome, 5–12, 19, 23–43, 45, 124, 142–144, 146, 147, 152–156, 159, 181, 191, 193, 194, 197, 204, 210, 212, 218, 312

Human genomics, 23, 142–144

Human molecular epidemiology, 43, 44

I

Inbreeding studies, 126

J

Journal of Medicine (New England J. M.), 33

K

Kilobytes (Kb), 20, 102, 136, 154, 158–161, 197, 201, 204, 225, 235, 266, 284

L

Linkage analysis, 30, 128–130, 146–164, 191, 193, 198–201, 204, 208, 210–212, 252

Linkage disequilibrium, 130, 152, 154, 156, 157, 159, 163, 193–197, 200–202, 210–212, 217, 231, 281, 314, 321

Linkage studies, 5, 45, 46, 124, 128–130, 146, 200, 203, 204, 211

M

Manhattan plots, 133–136, 252–263

Mendelian inheritance, 31, 33, 184–198, 202, 203, 210, 211

Mendel's laws, 149, 170, 185, 191, 193, 199–201

Modes of inheritance, 185

Molecular variation study, 181–183

Multi-dimensional analysis in genetic epidemiology, 311–341

Multiple comparison, 263–308

Mutation, 8, 11–15, 18, 20, 21, 24–27, 30–34, 37–39, 42, 43, 130, 131, 148, 150–152, 154, 158–164, 186, 188, 190, 192, 197, 198, 201, 203, 204, 207, 208, 210, 211, 213, 282, 317

N

Natural selection, 151, 160, 162, 190

Nature, ix, 8, 10, 15, 16, 39–44, 125–126, 143, 146, 147, 151, 163, 164, 181, 182, 280, 317, 318

Nurture, ix, 15, 125, 126

O

Online Mendelian Inheritance in Man (OMIM), 31, 186, 187

P

Population, viii, 2, 105, 125, 146, 219

Population stratification, 184, 205, 206, 232

Preventive medicine, 1–5, 54, 124

Public health, xviii, 1–5, 34, 40, 42, 54, 182, 209–215

Q

Quality, 8, 10, 44, 49, 51–53, 142, 157, 191, 213, 275, 317, 324

R

Randomization test, 126

Rare genetic variations, 197

Regression decision trees, 308–311

R Software, ix, 53–70, 88

S

Segregation studies, 45, 46, 124, 127, 128, 203

Single nucleotide polymorphisms (SNP), 45, 124, 131, 136, 137, 144, 146–148, 152, 154, 156, 158, 160, 162, 183, 191, 193, 195, 204, 212, 214, 231, 232, 246, 252, 270, 281–283, 303, 312–316, 321–324

Spectrum of variation, 145

T

Theory of probability, 123, 124

Traditional Chinese Medicine (TCM), vii, viii

Traits, 13, 14, 24, 26, 44, 125, 127, 129, 131, 132, 136, 137, 142, 143, 147, 149–151, 157–159, 162, 164, 176, 177, 180–198, 208, 210, 213, 222, 231, 232, 237, 246, 270, 281, 308, 311, 314–316, 320–324

Transmission-disequilibrium test (TDT), 130, 201–204, 206, 232

T-test in statistics, 222–230

Tumors, viii, ix, 36–38, 143, 207, 208, 313, 314, 317

Twin studies, 126, 165, 177, 180, 203, 208

U

Univariate, bivariate and multivariate data analysis, 100–121

Unobservable phase, 280

V

Variations, 13, 14, 23, 25, 26, 29, 38, 42, 126, 156, 180, 182, 186, 190, 197, 213, 308

W

Wilcoxon rank-sum test, 222, 223

X

X-linked genes, 185

Y

Y-linked genes, 185

Z

Zygosity, 178, 188, 212