Bioinformatics and Statistical Genetics

Elective specialization for MDS/MIRI/MAI students

Marta Castellano

Department of Statistics and Operations Research
Universitat Politècnica de Catalunya Barcelona, Spain

marta.castellano@upc.edu

UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

Syllabus

Bioinformatics and **Statistical Genetics**

1. Introduction to statistical genetics — 7 November 2023
2. Hardy-Weinberg equilibrium — 14 November 2023
3. Linkage disequilibrium and haplotype estimation — 21 November 2023
4. Population substructure — 28 November 2023
5. Genetic association analysis — 5 December 2023
6. Relatedness analysis (allele sharing) — 12 December 2023

General review 19 December 2023

EXAM 15 January 2024

Tuesdays 5-8pm

Statistical Genetics

Definition

Statistical genetics is a scientific field that deals with the analysis of inherited features (i.e. traits) and genetic data. The scientific study of inherited variation.

- Commonly used in the context of human genetics
- Nowadays genetic data arises in different forms (sequences, markers, ...)

Statistical Genetics

Why?

Statistical Genetics

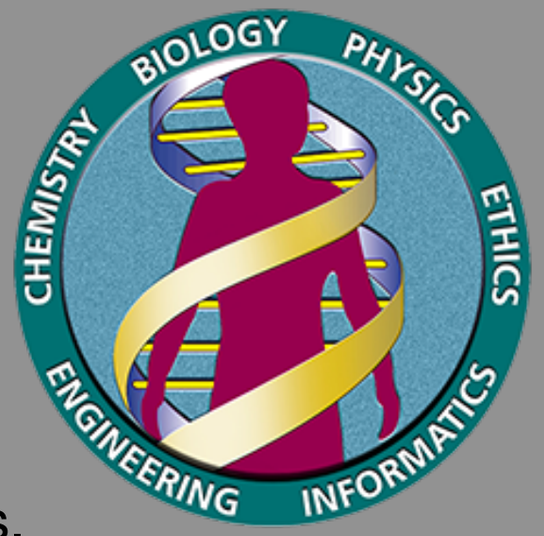
Why?

- Better understanding of the biological world
- Human wellbeing
    - Discovery and description of the genetic contribution to many human diseases
    - Forensics
    - …
- Understanding and describing evolution, in hand with classical physical anthropology approaches

National Institutes of Health, 2007. Understanding human genetic variation. NIH Curriculum Supplement Series.

Statistical Genetics

Why now?



- The size of the genetic databases has grown enormously over the years.

- The **human genome project** was launched in 1990…with a cost of $3 billion.
  - In 2000, the project produces a draft human genome sequence that accounts for the 90% of total human genome
  - In 2003, the project announces the finished version of human genome sequence, with the 92% of sampling (exceeding 99.99% accuracy)
  - Complete genome by the Telomere-to-Telomere initiative: March 2022

- Continuations:
  - 1000 Genomes Project (2008-2015) to establish the most detailed catalogue of human genetic variation.
  - International HapMap Project (2002-2009) to develop a haplotype map
  - China Kadoorie Biobank (2004-2020) largest prospective cohort study
  - …

Cohort study: a type of observational study focused on following a group of people (called a cohort) over a period of time, collecting data on their exposure to a factor of interest

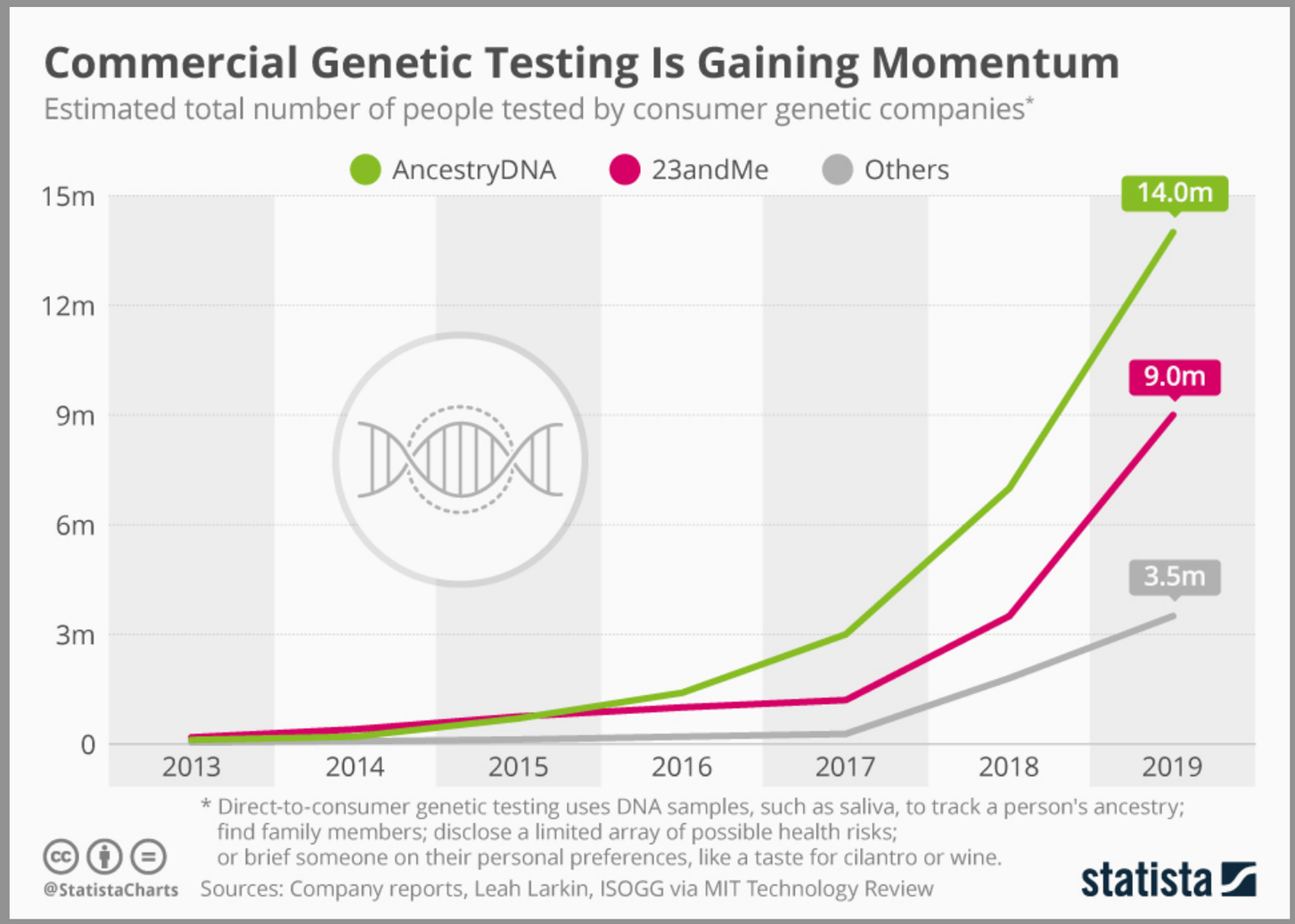
Statistical Genetics

Why now?

- Next Generation Sequencing platforms reduced the cost sequencing by 500 000-fold.
  - Current cost is about 200$ per whole genome sequencing…and keeps decreasing
- Current direct-to-consumer genetic testing are changing the scope of human genetic studies:

23andMe is the only company with FDA clearance to test for an increased risk of occurrence for several diseases: https://www.fda.gov/medical-devices/in-vitro-diagnostics/direct-consumer-tests#list

Department of Justice in USA is authorised to obtain data from DNA-testing websites since 2020



Statistical Genetics

How?

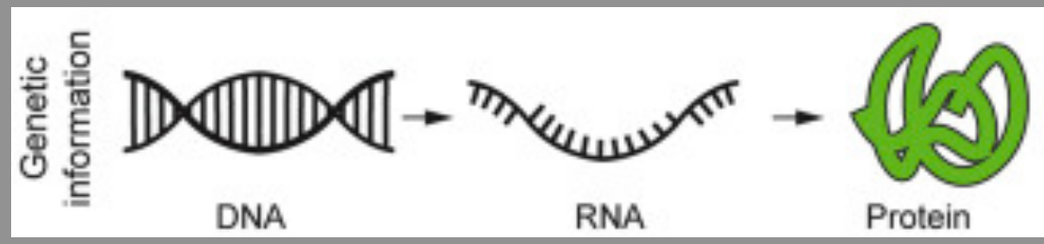
Content TODAY

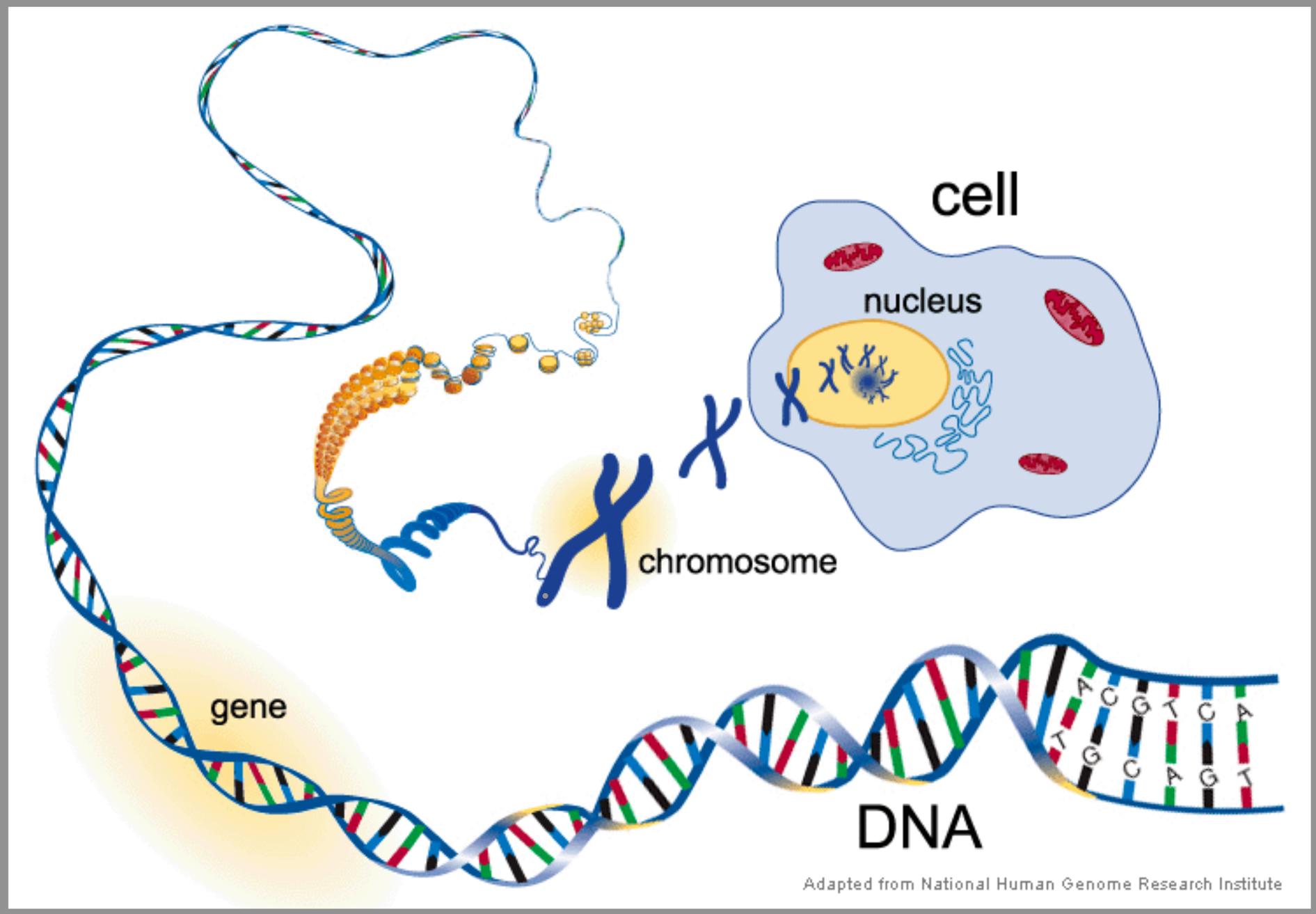
Introduction to statistical genetics

1. Basic terminology
2. Traits and genetic markers
   - RFLPs (Restriction Fragment Length Polymorphism)
   - STRs (Short Tandem Repeat)
   - SNP (Single Nucleotide Polymorphism)
3. Descriptive analysis of a genetic marker
4. Computer exercise

Statistical Genetics

Basic terminology



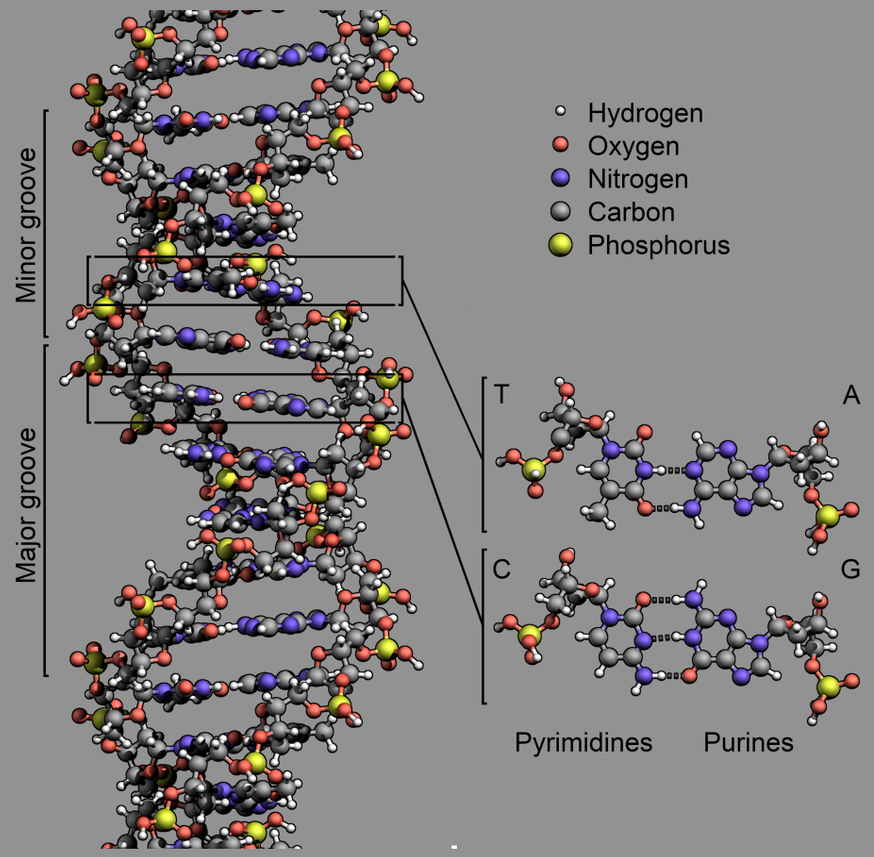


The classical image of a chromosome only appears during cell division during a process called chromatin condensation. Meanwhile, the genetic information is… loosely contained in the nucleus

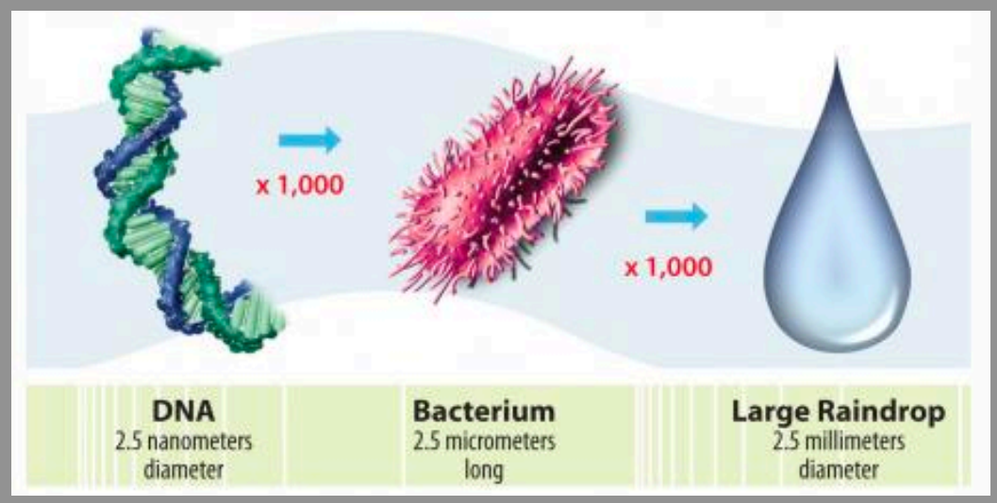
Statistical Genetics

Basic terminology

- **DNA (deoxyribonucleic acid):** is a molecule composed of two polynucleotide chains that coil around each other to form a double helix.
  - Each nucleotide is composed of one of four bases: C (cytosine), G (guanine), A (adenine) or T (tymine)
  - The bases are connected to each other via chemical bonds: A-T and C-G



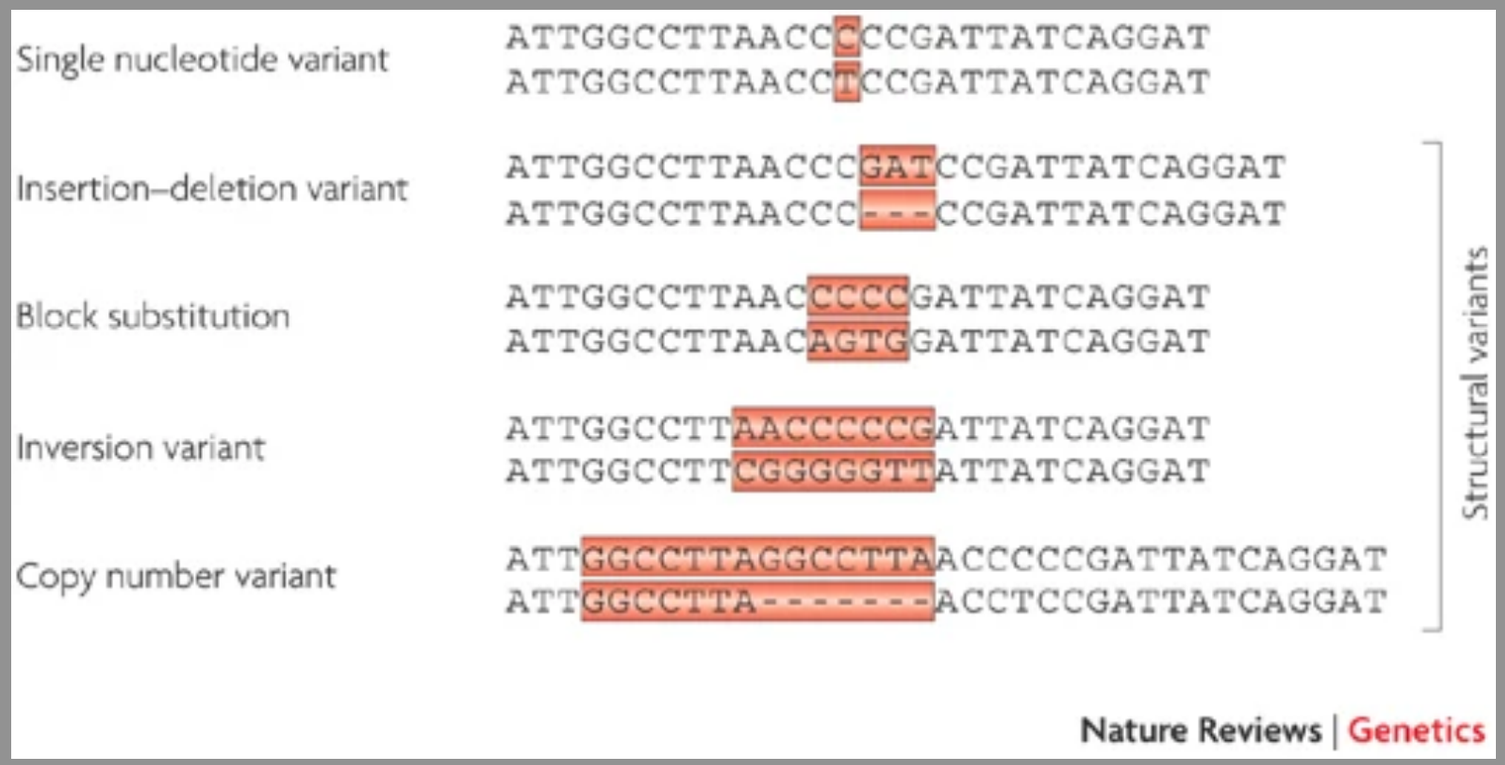
- The human genome comprises about 3×10^9 base pairs of DNA



Statistical Genetics

Basic terminology

- **Genetic variation** refers to differences among the genomes of members of the same species.
- **Genome**: is all the hereditary information—all the genes—of an organism.



Frazer, K.A., Murray, S.S., Schork, N.J. and Topol, E.J., 2009. Human genetic variation and its contribution to complex traits. Nature Reviews Genetics, 10(4), pp.241-251.

Statistical Genetics

Basic terminology

Sources of genetic variation

- Mutation
- Sexual reproduction (i.e. recombination)
- Genetic drift (the change in frequency of an existing gene variant in the population due to random chance)
- Other factors that alter genetic flow on a population
  - Non-overlapping generations.
  - Non-random mating (w.r.t the trait under study).
  - Population size
  - Migration (i.e. gene flow)
  - Environmental variance

Statistical Genetics

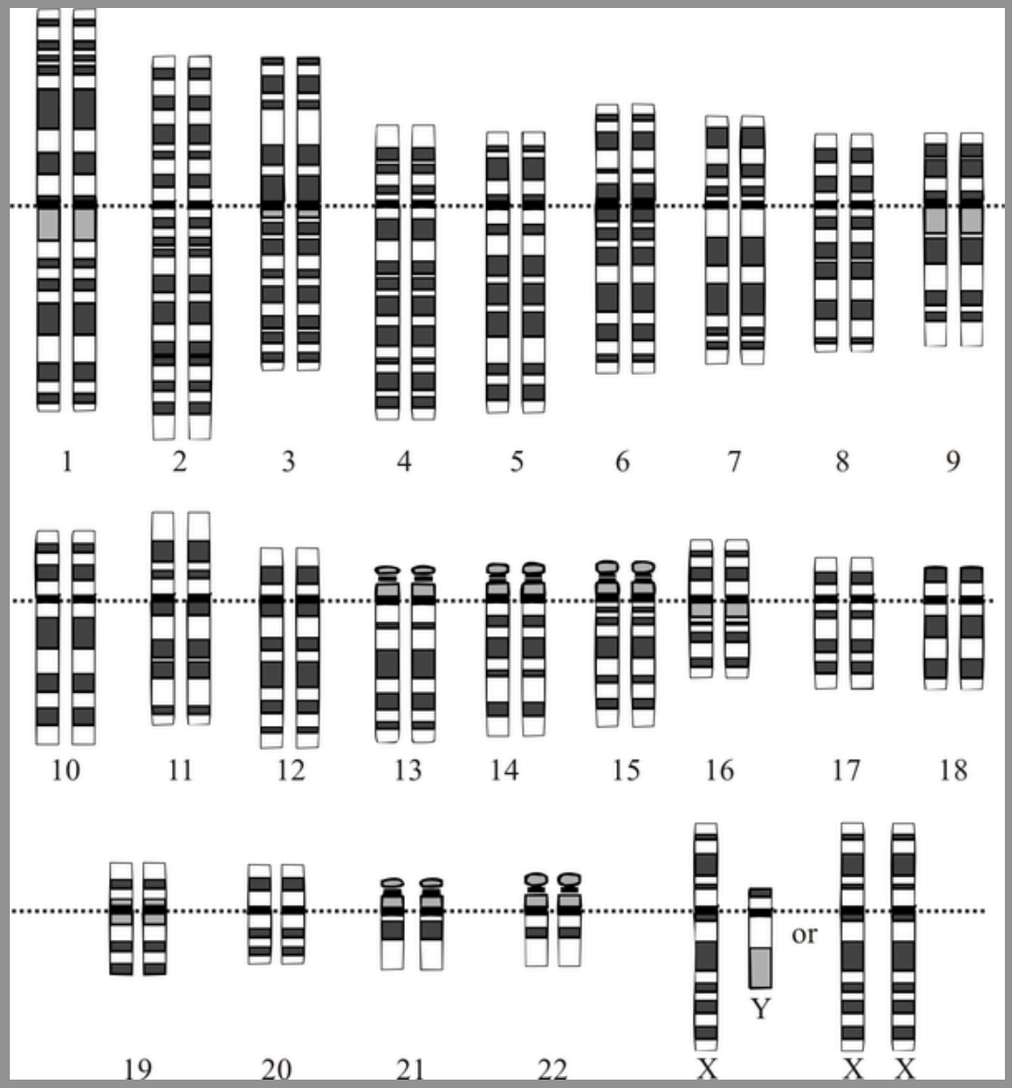
Other basic terminology

- A human being has 46 chromosomes in the nucleus of each cell, coming in 23 homologous pairs (22 pairs of autosomes and 1 pair of sex chromosomes (X/Y)).

- Reproductive cells have one copy of the genome and are **haploid**. Body cells of any individual are **diploid**.

- A **haplotype** (haploid genotype) is a group of alleles in an organism that are inherited together from a single parent.

Every cell contains all genetic information

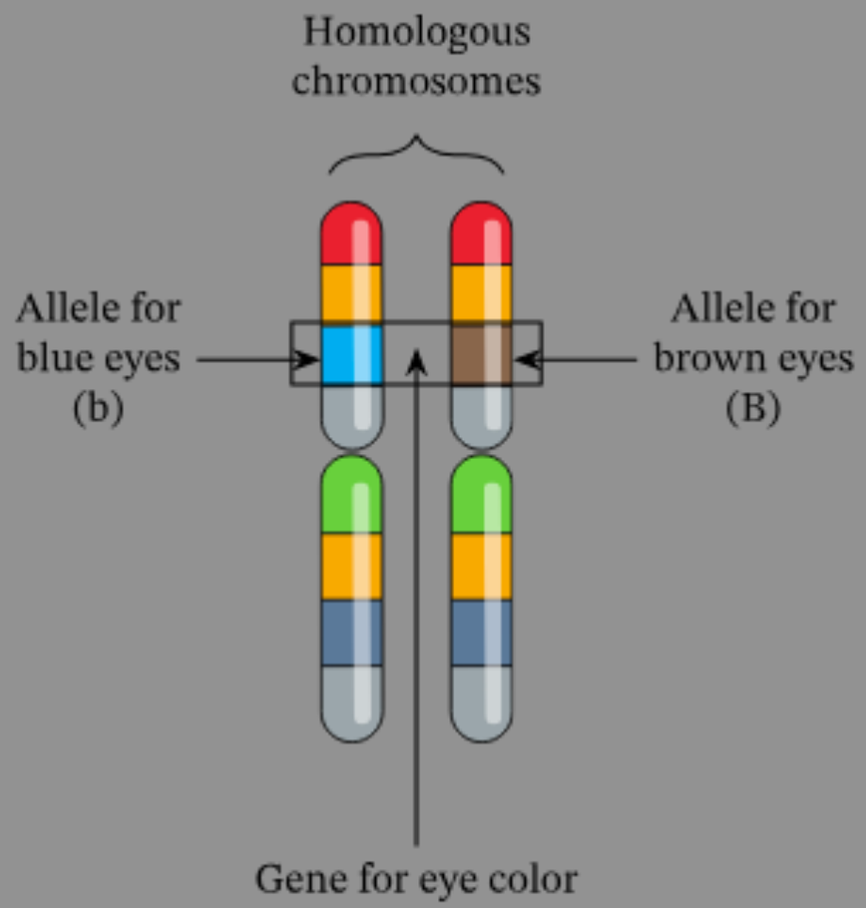


- An **haplotype** is a combination of alleles at different chromosomal regions that are closely linked and that tend to be inherited together.

Statistical Genetics

Other basic terminology

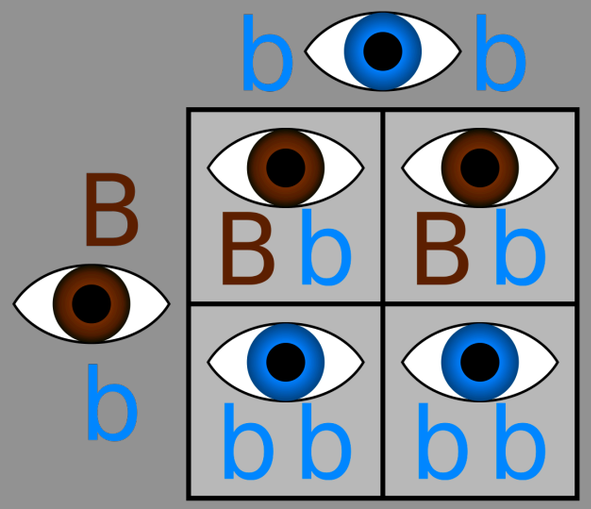
- A **gene** corresponds to a piece of a chromosome, a stretch of the genome that encodes information that determines some characteristic of the offspring.
- The location of a gene in the genome is called its **locus**.
- The multiple versions of a gene are called **alleles**, if there are two often indicated by A and *a*, A_1 and A_2, or A and B.
- Each individual inherits two copies of each gene, one from the father and one from the mother.
- The genetic makeup of an individual is called his/her **genotype**. For a gene with two alleles, this can be AA, Aa or aa

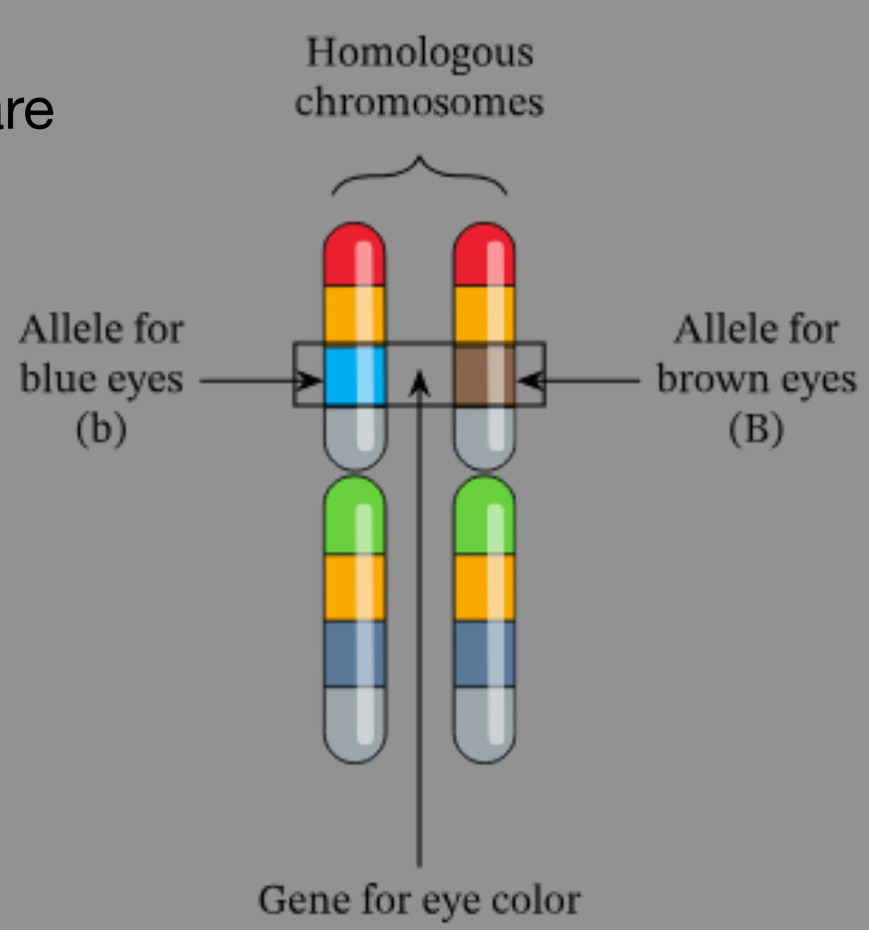


Statistical Genetics

Other basic terminology

- Alleles can be **dominant**, **recessive** or **codominant**.
- An individual that inherits the same allele from father and mother is **homozygous** (AA or aa).
- An individual with a different allele on each chromosome of a pair is **heterozygous** (Aa).
- Alleles can be visually displayed using a Punnett square





Content

Introduction to statistical genetics

1. Basic terminology
2. Traits and genetic markers
    - RFLPs (Restriction Fragment Length Polymorphism)
    - STRs (Short Tandem Repeat)
    - SNP (Single Nucleotide Polymorphism)
3. Descriptive analysis of a genetic marker
4. Computer exercise

Statistical Genetics

Traits and markers - moving to populations

- A **trait** (**phenotype**) is a specific characteristic of an individual. Traits can be determined by genes, environmental factors or by a combination of both.
    - Traits can be qualitative (such as eye color) or quantitative (such as height or blood pressure).
- In many studies in statistical genetics, some **trait** (e.g. yield or disease status) of an organism is considered to depend on one or more genetic variables.
- The position of genetic factors determining a trait is often unknown.

Statistical Genetics

Traits and markers - moving to populations

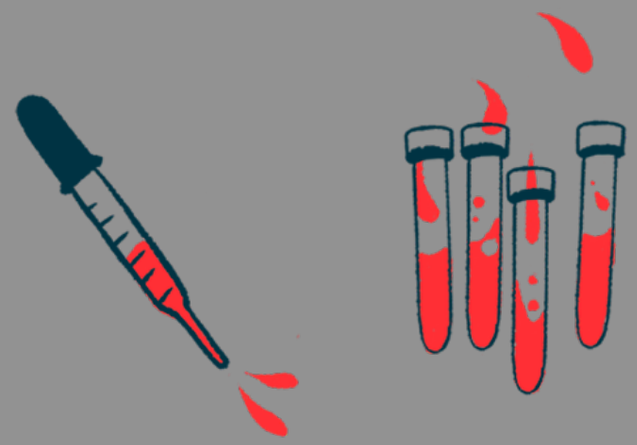
What is a marker?

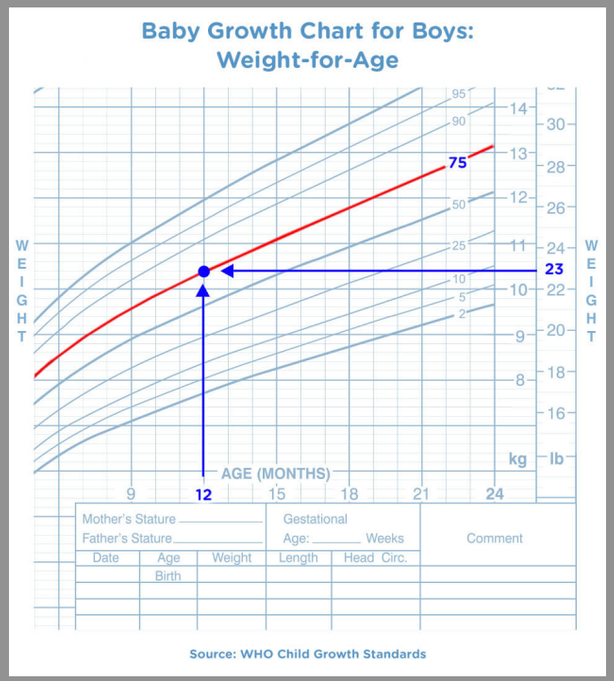
A characteristic that is **objectively** measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention.

Biomarkers Definitions Working Group, Atkinson Jr, A.J., Colburn, W.A., DeGruttola, V.G., DeMets, D.L., Downing, G.J., Hoth, D.F., Oates, J.A., Peck, C.C., Schooley, R.T. and Spilker, B.A., 2001. Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. Clinical pharmacology & therapeutics, 69(3), pp.89-95.

A broad subcategory of medical signs – that is, **objective** indications of medical state observed from outside the patient – which can be measured accurately and reproducibly.

Strimbu, K. and Tavel, J.A., 2010. What are biomarkers?. Current Opinion in HIV and AIDS, 5(6), p.463.

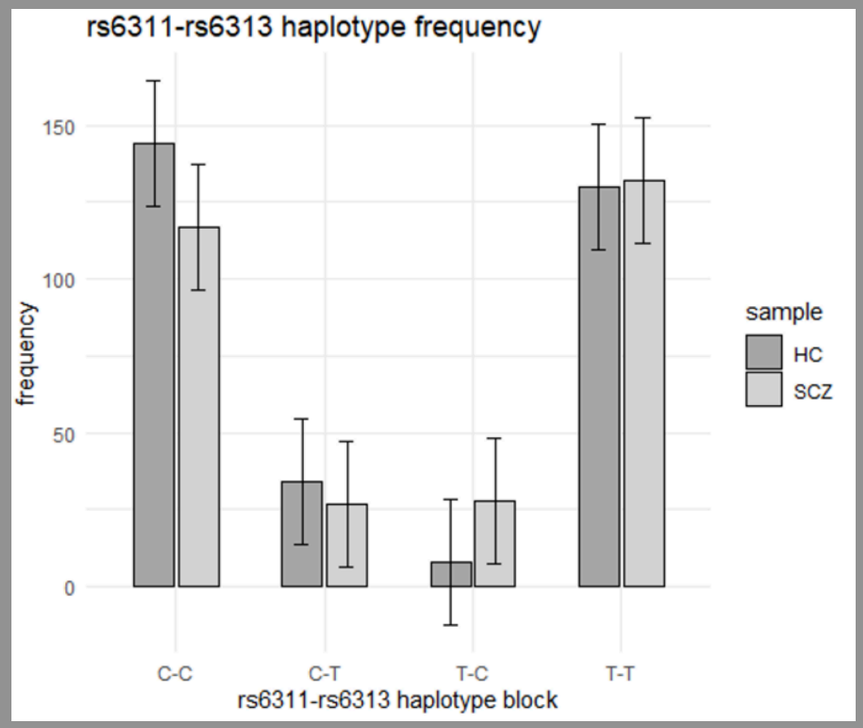




Statistical Genetics

Traits and markers - moving to populations

- A **genetic marker** is a genetic variable that has a known variation over individuals, has a known locus.
    - Example: rs6311 and rs6313 are genetic markers in the Serotonin 5-HT2A receptor gene on human chromosome 13.
- The study of associations between markers and phenotypes can be helpful in identifying the genetic factors that affect the trait/phenotype.



Massoud, S., Salmanian, M., Tabibian, M., Ghamari, R., Tavabe-Ghavami, T.S. and Alizadeh, F., 2022. Association evaluation of 5-HTR2A manifested salient contribution of rs6311 and rs6313 in increasing the occurrence of schizophrenia.

- Genetic markers can vary both within and between different populations
- Genetic markers can be observed occur throughout the entire genome, and most do not relate to an altered protein structure or gene expression.
- However, markers may still be associated with diseases due to their physical proximity, or “linkage,” to a disease-causing genetic variant

Statistical Genetics

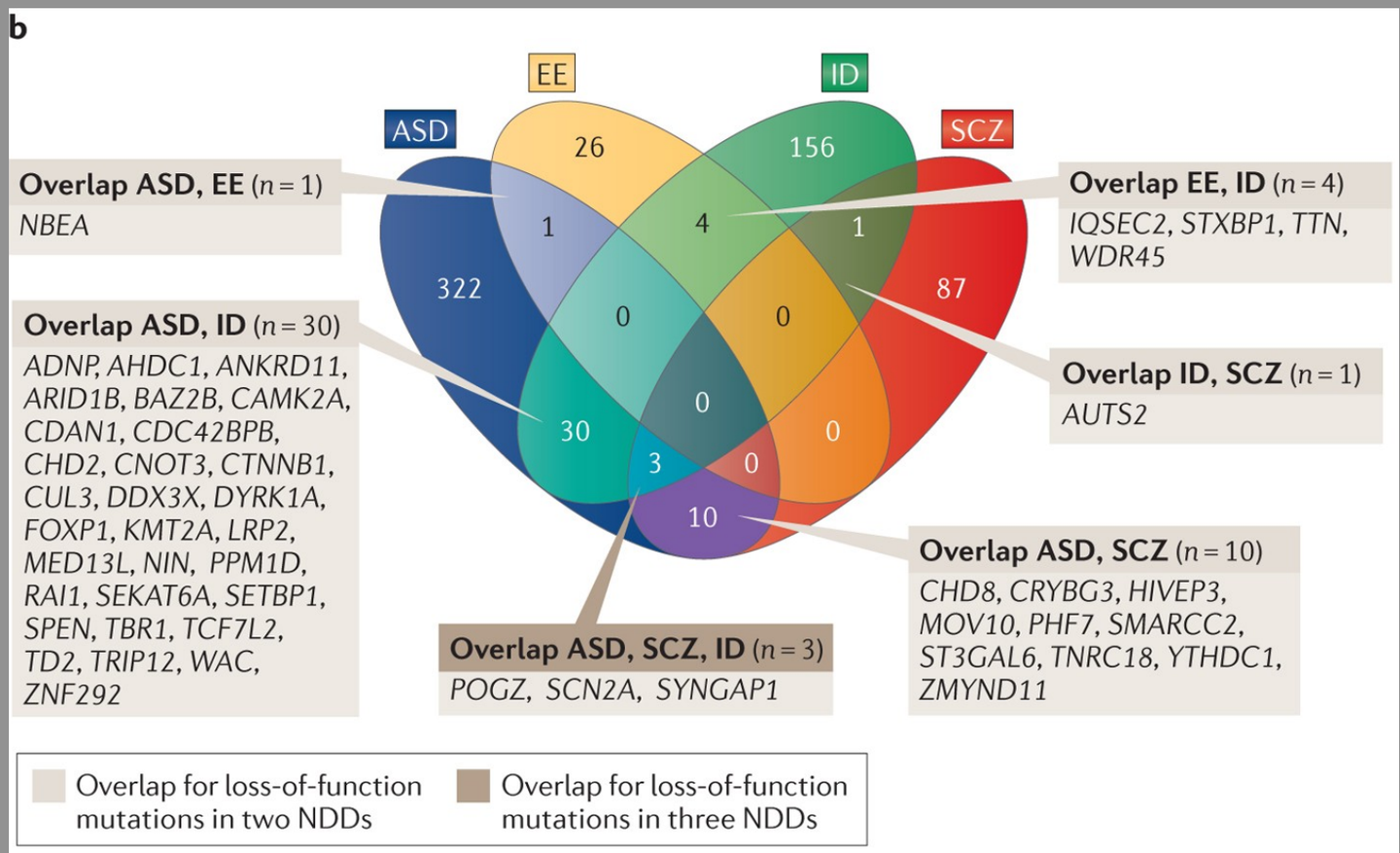
Traits and markers - moving to populations

Genetic markers in complex diseases

- Complex diseases are caused by a combination of genetic, environmental, and lifestyle factors, most of which have not yet been identified.

- In almost any complex trait that has been studied, **many loci contribute to standing genetic variation**…so that mutations in many genes contribute to genetic variation in the population.

- On average, the proportion of variance explained at the individual variants is small.



- Vissers, L.E., Gilissen, C. and Veltman, J.A., 2016. Genetic studies in intellectual disability and related disorders. *Nature Reviews Genetics*, *17*(1), pp.9-18.
- Génin, E., 2020. Missing heritability of complex diseases: case solved?. Human Genetics, 139(1), pp.103-113.

Statistical Genetics

Traits and markers - moving to populations

Genetic markers in complex diseases

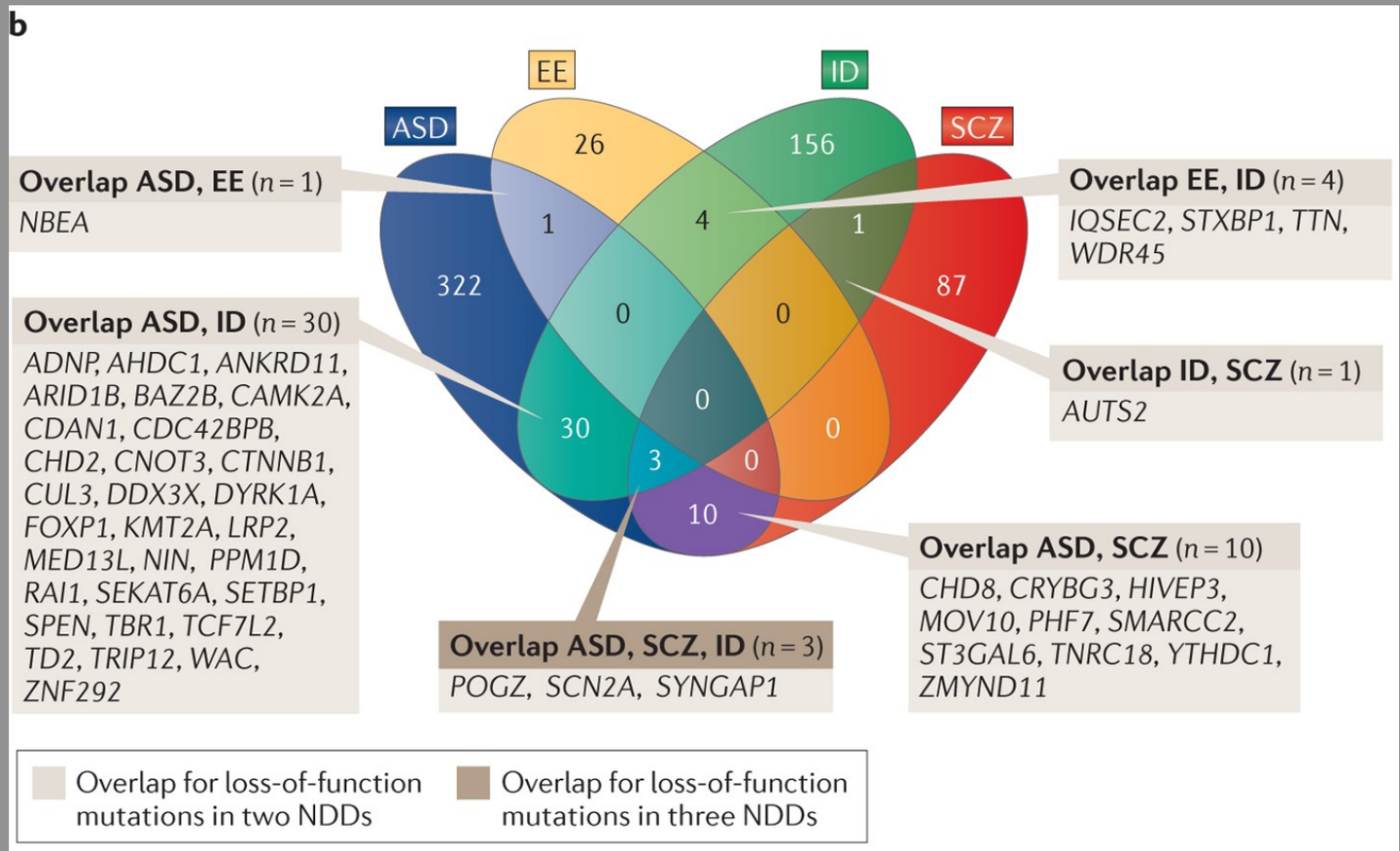
- Complex diseases are caused by a combination of genetic, environmental, and lifestyle factors, most of which have not yet been identified.

Example:

- The genes for which overlap in de novo mutations between neurodevelopmental disorders has been identified are listed.

- This does not imply that all of these mutations are the cause of these neurodevelopmental disorders.

ASD = Autism Spectrum Disorder
EE = epileptic encephalopathy
ID = intellectual disability
SCZ = Schizofrenia



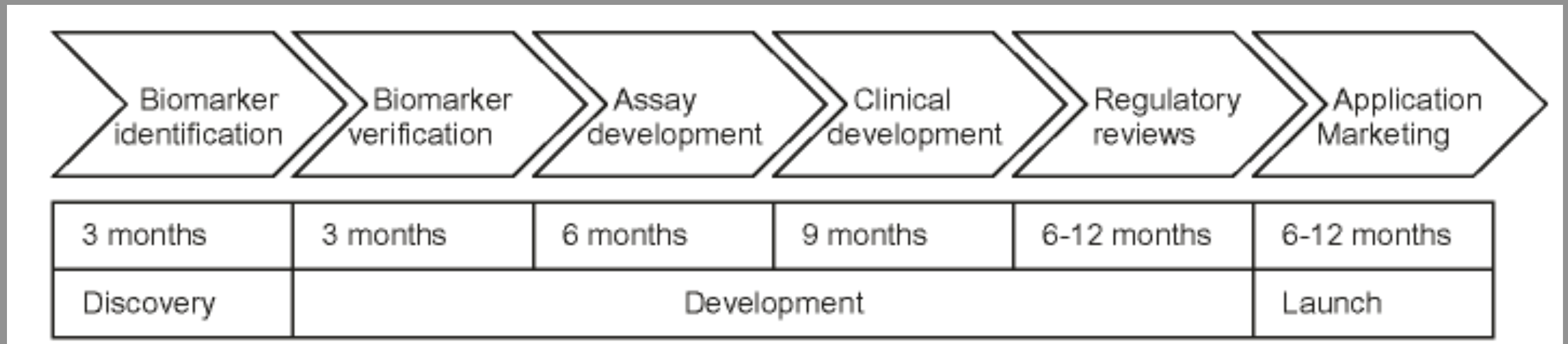
- Vissers, L.E., Gilissen, C. and Veltman, J.A., 2016. Genetic studies in intellectual disability and related disorders. *Nature Reviews Genetics*, *17*(1), pp.9-18.
- Génin, E., 2020. Missing heritability of complex diseases: case solved?. Human Genetics, 139(1), pp.103-113.

Nature Reviews | Genetics

Statistical Genetics

Traits and markers - moving to populations

From bench to market: the Clinical view



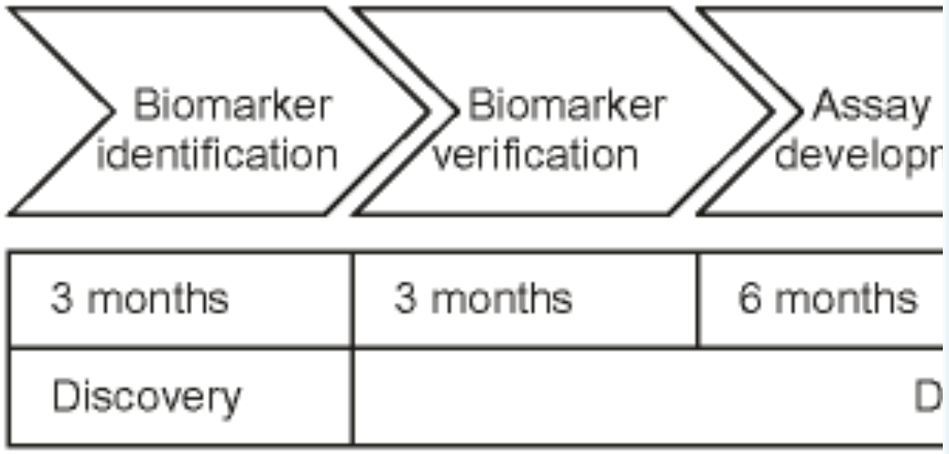
Stages and timelines of biomarker discovery, development and marketing

K.K. Jain, 2010, **The Handbook of Biomarkers**, DOI 10.1007/978-1-4939-7431-3_19

Statistical Genetics

Traits and markers - moving to populations

From bench to market: the Clinical view



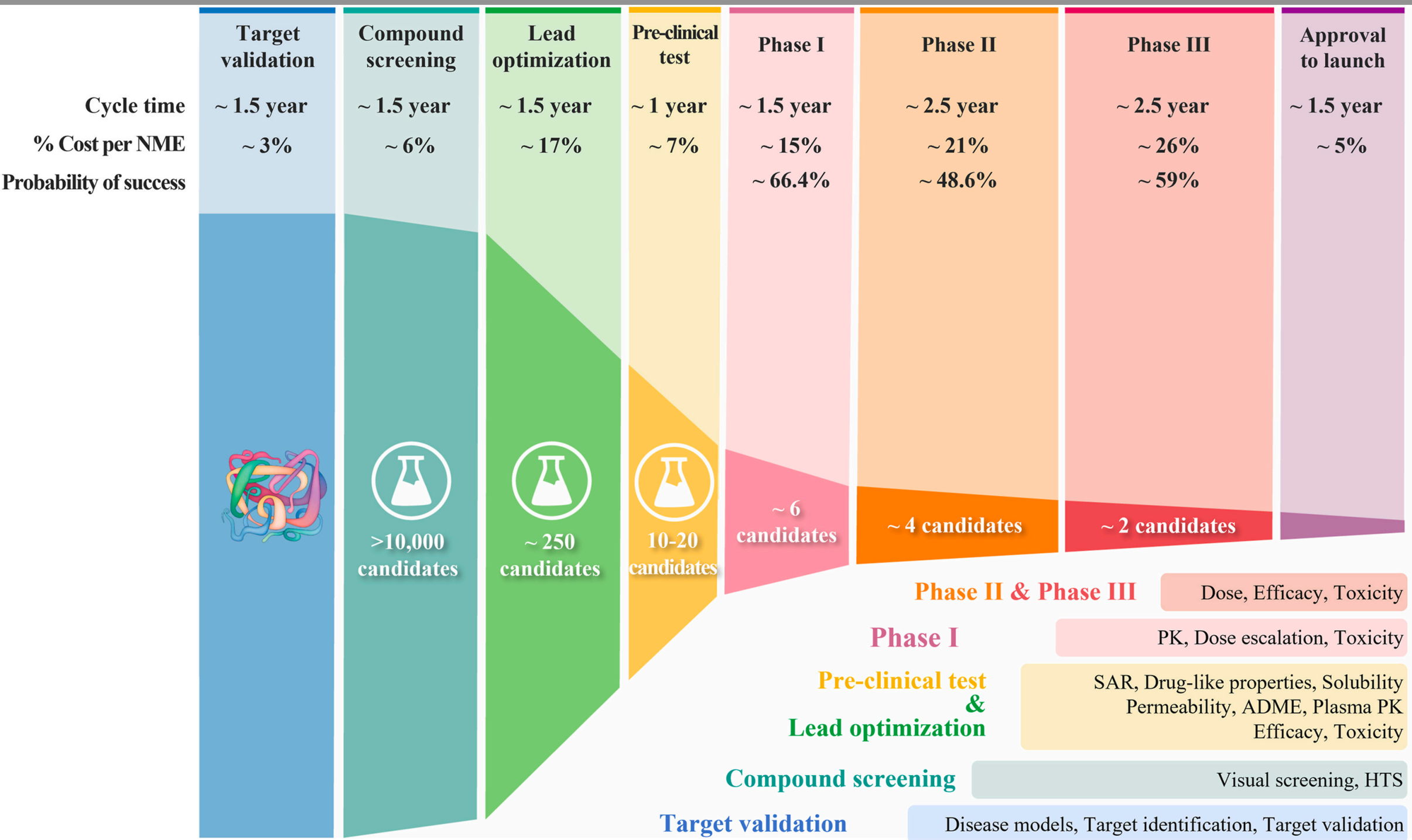
Stages and timelines of biomarker discovery, dev

What may be a more realistic timeline...

K.K. Jain, 2010, **The Handbook of Biomarkers**, DOI 10.1007/978-1-4939-7431-3_19

Statistical Genetics

Traits and markers - moving to populations



Statistical Genetics

Traits and markers - moving to populations

From bench to market: the Product Manager view

Brainstorming

Find a product that meets an unmet demand:

What is the problem that you want to solve?

Development

Research, design test possible solutions to your problem

How can you solve it?

Scale up

Ensure your solution meets market demands:

- Testing
- Scale Up Production
- Manufacturing processes
- Quality and Regulatory constraints

Launch

Reach your target client and beyond

- New product launch
- Marketing to maximize impact in the market

Statistical Genetics

Traits and markers - moving to populations

From bench to market: genetic markers at the FDA

At least 483 therapeutic products are listed in the FDA with genetic information found in the drug labelling.

- 42% of them in oncology
- 25% of them with gene CYP2D6

The CYP2D6 enzyme catalyses the metabolism of a large number of clinically important drugs including antidepressants, neuroleptics, some antiarrhythmics, lipophilic β-adrenoceptor blockers and opioids.
Variations of the CYP2D6 allele would alter enzymatic function to the point that a person can be classified into:

- poor metabolizer – little or no CYP2D6 function
- intermediate metabolizers – metabolize drugs at a rate somewhere between the poor and extensive metabolizers
- extensive metabolizer – normal CYP2D6 function
- ultrarapid metabolizer – multiple copies of the CYP2D6 gene are expressed, so greater-than-normal CYP2D6 function occurs

Examples of drugs with pharmacogenomic indications:

- **Codeine:** pain, coughing and diarrhea.
- **Setmelanotide:** chronic weight management with specific genetic mutations.
- **Citalopram**: antidepressant, whose dose needs to be adjusted for patients with specific mutations.
- **Herceptin, Gleevec and Erbitux:** treatment of cancer populations with specific genetic mutations.

Statistical Genetics

Traits and markers - moving to populations

DOSAGE AND ADMINISTRATION
Special Populations
20 mg/day is the maximum recommended dose for patients who are greater than 60 years of age, patients with hepatic impairment, and for CYP2C19 poor metabolizers or those patients taking cimetidine or another CYP2C19 inhibitor. (see WARNINGS)

WARNINGS
QT-Prolongation and Torsade de Pointes
The citalopram dose should be limited in certain populations. The maximum dose should be limited to 20 mg/day in patients who are CYP2C19 poor metabolizers or those patients who may be taking concomitant cimetidine or another CYP2C19 inhibitor, since higher citalopram exposures would be expected.

CLINICAL PHARMACOLOGY
Pharmacokinetics
Population Subgroups
(…) CYP2C19 poor metabolizers – In CYP2C19 poor metabolizers, citalopram steady state Cmax and AUC was increased by 68% and 107%, respectively. Celexa 20 mg/day is the maximum recommended dose in CYP2C19 poor metabolizers due to the risk of QT prolongation (see WARNINGS and DOSAGE AND ADMINISTRATION).
CYP2D6 poor metabolizers - Citalopram steady state levels were not significantly different in poor metabolizers and extensive metabolizers of CYP2D6.

- **Citalopram**: antidepressant, whose dose needs to be adjusted for patients with specific mutations.
- **Herceptin, Gleevec and Erbitux:** treatment of cancer populations with specific genetic mutations.

Statistical Genetics

Traits and markers - moving to populations

SIDE NOTES

- The terms **marker**, **variant** and **polymorphism** are often used interchangeably…although they have slightly different meanings.
    - A **genetic marker** is a genetic variable that has a known variation over individuals, and has a known physical location on a chromosome (locus).
    - A mutation is a change in a sequence of the genome of a particular organism, but **polymorphism** is a mutation that occurs in more than 1% of a particular population.
    - A **variant** is a sequence of the genome (or a whole genome, i.e. virus) that may contain one or more mutations.

- If in the population only one allele occurs at a site or locus, we shall say that it is **monomorphic**, or **monoallelic** in that population.
- DNA markers that are referred to as monomorphic markers. A monomorphic genetic marker consists only of homozygotes for one particular allele and cannot be used to differentiate genotypes.

Content

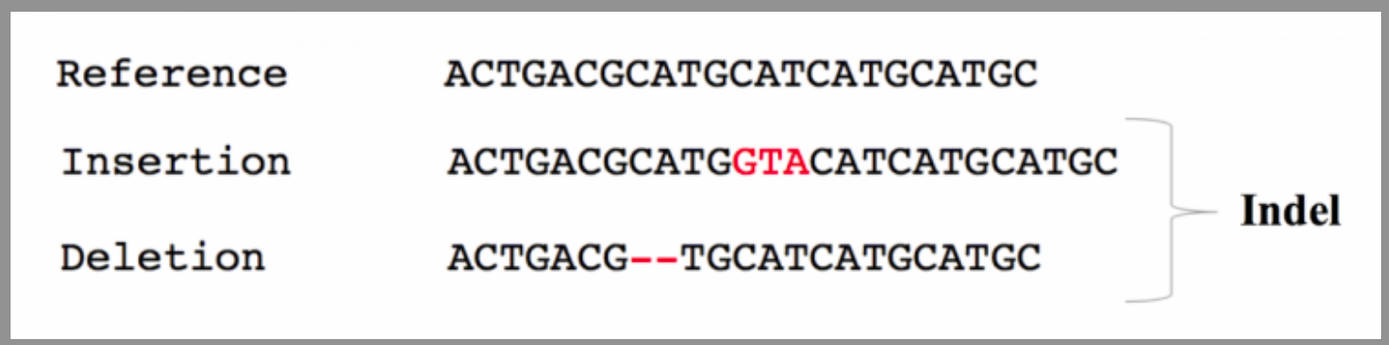
Introduction to statistical genetics

1. Basic terminology
2. Traits and genetic markers
    - RFLPs (Restriction Fragment Length Polymorphism)
    - STRs (Short Tandem Repeat)
    - SNP (Single Nucleotide Polymorphism)
3. Descriptive analysis of a genetic marker
4. Computer exercise

Statistical Genetics

Genetic markers

- There are many different markers of which we consider
  - RFLPs (Restriction Fragment Length Polymorphism)
  - Microsatellites or STRs (Short Tandem Repeat)
  - SNPs (Single Nucleotide Polymorphism)
  - Indels (insertion/deletion polymorphism)
  - …

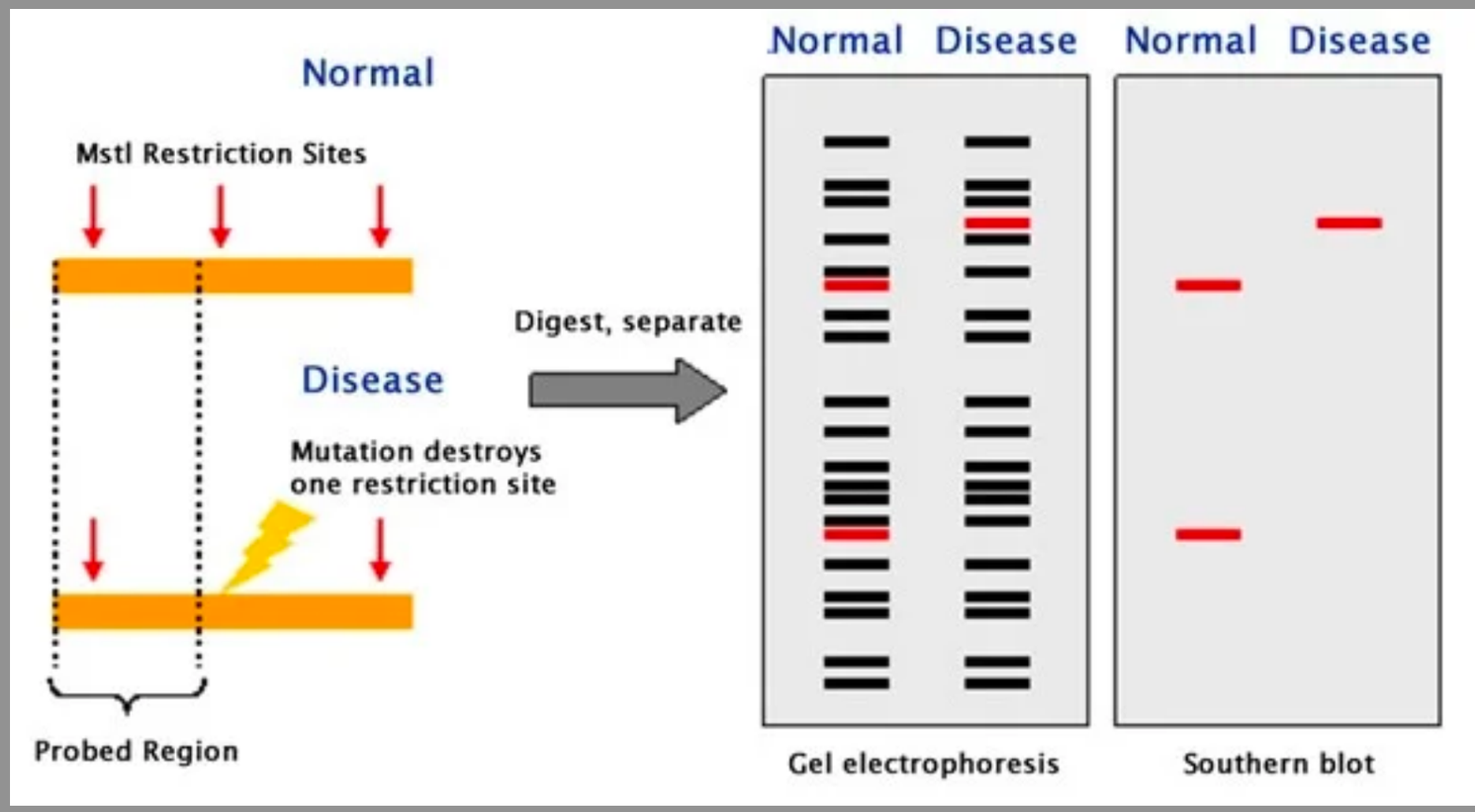


Statistical Genetics

Markers: RFLPs (Restriction Fragment Length Polymorphism)

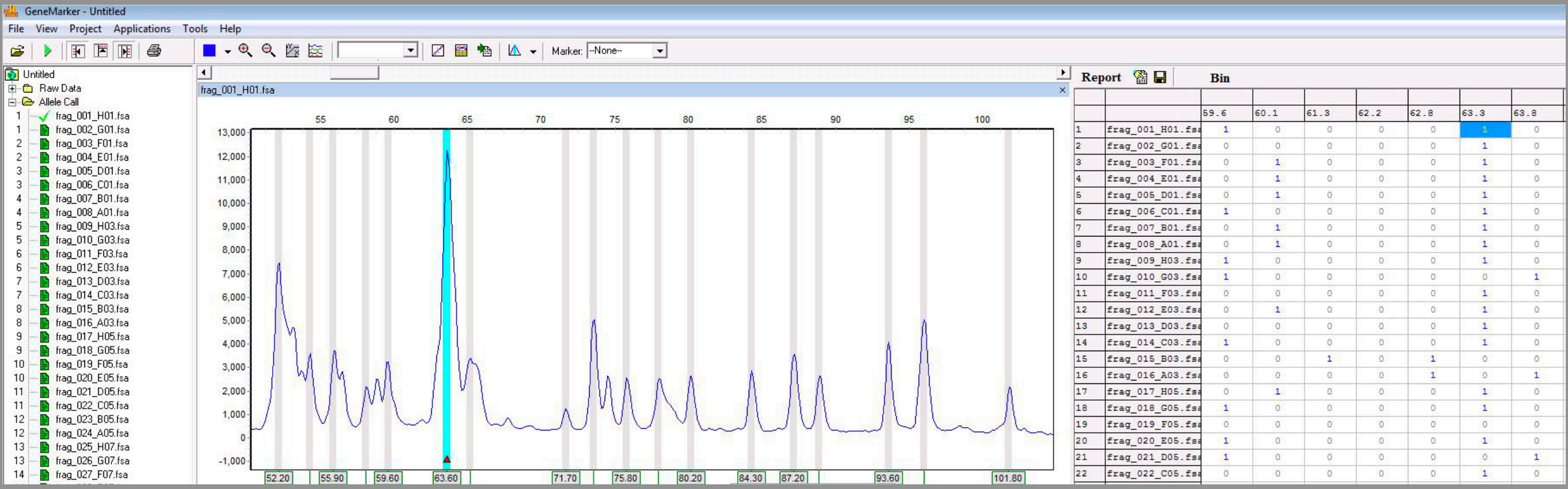
- A large number of restriction enzymes has been discovered that cut DNA at a specific motif.
  - Example: enzyme BamHI cuts DNA at the recognition sequence GGATCC/CCTAGG
- By digesting DNA with a restriction enzyme DNA fragments of variable length arise.
- First type genetic markers used for forensics, genetic disorders, or paternity testing. Now largely obsolete due to the emergence of next generation sequencing (NGS) or massive parallel sequencing (MPS).

- These can be separated on a gel in the laboratory, and presence/ absence of restriction sites can be inferred.
- Produces binary data.



Statistical Genetics

Markers: RFLPs (Restriction Fragment Length Polymorphism)

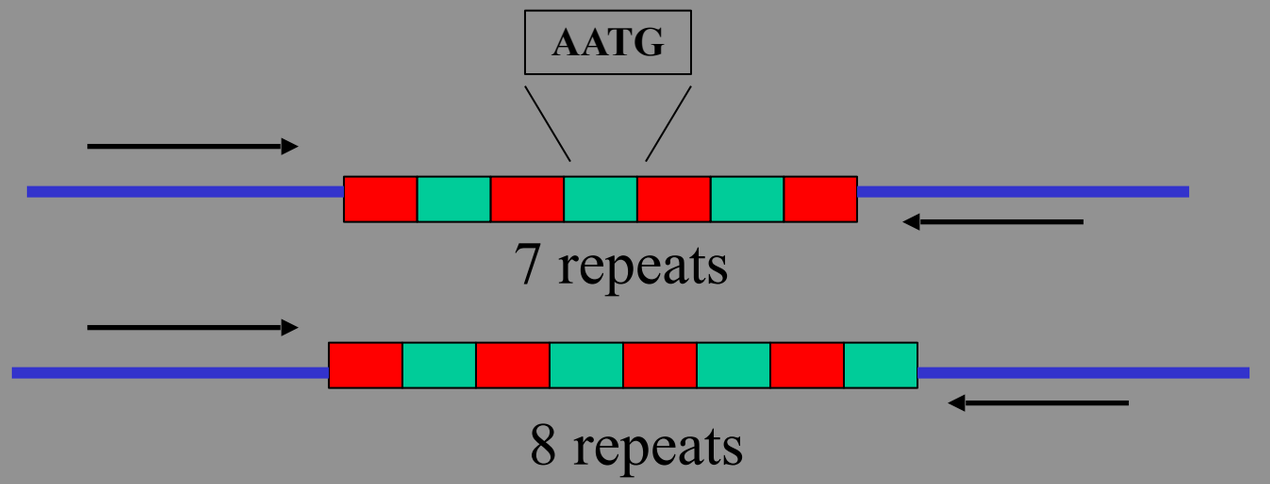


https://softgenetics.com/products/genemarker/t-rflp/#popup-image-modal

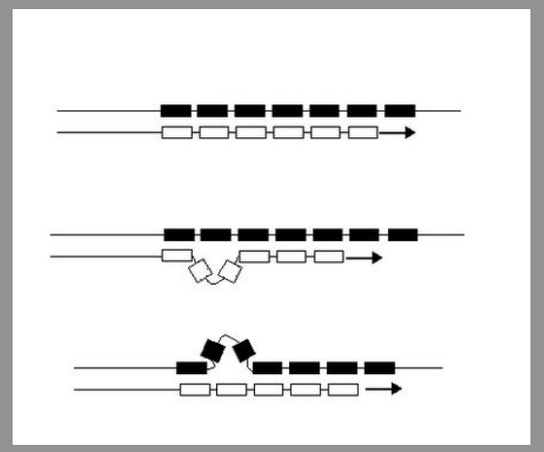
- Produces binary data.

Statistical Genetics

Markers: Microsatellites or STRs (Short Tandem Repeat)

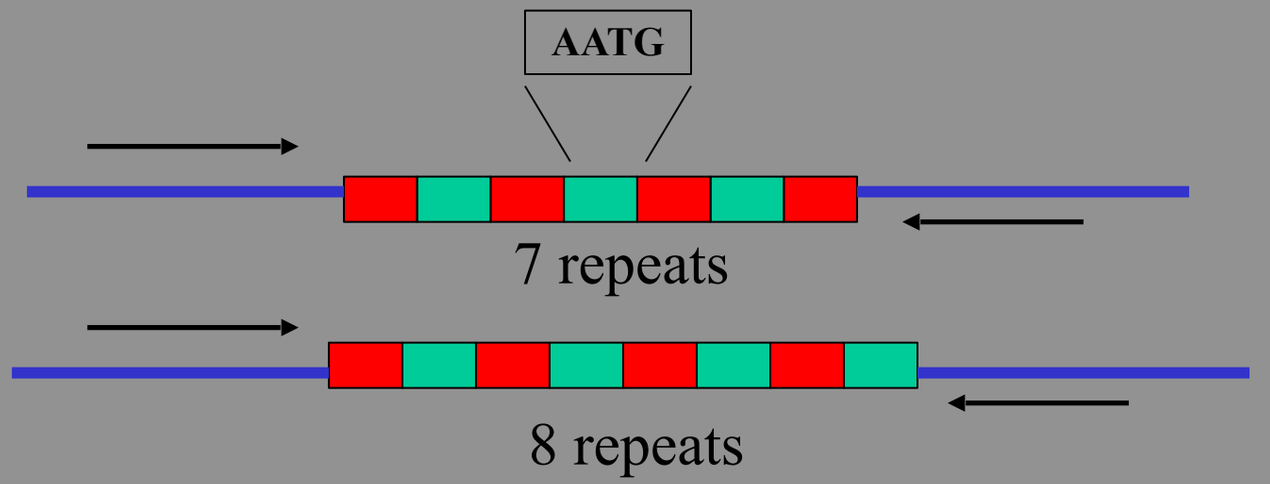


- Microsatellites consist of short sequences (e.g. ATT) that repeat a certain number of times (e.g. ATTATTATTATT).
  - A small (2-6) number of base pairs is repeated (5-50 times).
  - They have higher mutation rate than other areas of DNA
  - They are mostly located in non-coding regions
  - The number of repeats an individual has on each chromosome can vary



Statistical Genetics

Markers: Microsatellites or STRs (Short Tandem Repeat)



- At the population level…
  - Individuals vary in the number of repeats they have.
  - Microsatellites have many alleles.
- Microsatellites are widely used in ecology and in forensics. They are quick/cheap to obtain (PCR-based), requiring tiny amounts of tissue.
- Produces count data, with a limited number of outcomes.

Statistical Genetics

Markers: Microsatellites or STRs (Short Tandem Repeat)

- STRs can be coded in different ways:
  - reporting the number of repeats an individual has on each chromosome.
  - reporting the total size of the repeating sequences as the number of base pairs on each chromosome.
- Example:
  - a tri-nucleotide STR: ATT.
  - an individual has the DNA sequences (ATTATTATT,ATTATTCAA)
  - can be coded as (3/2) (repeats)
  - (9/6) (total size)
- In the statistical analysis mostly treated as categorical.

Statistical Ge

Markers: Microsatellite

- STR database

D8S1179

General Information

Other Names	Chromosomal Location	GenBank Accession
D6S502, D8	8q	GO8710; has 12 repeats

Repeat: [TCTA] = GenBank top strand

PCR Primer Information

Reported Primers	Ref.	PCR Primer Sequences
Set 1	369	5' - TTTTTGTATTTCATGTGTACATTCG - 3' 5' - CGTAGCTATAATTAGTTCATTTTCA - 3'
Set 2	PE ABI	AmpFlSTR® Profiler Plus™
Set 3	Promega	GenePrint® PowerPlex™ 2.1, GenePrint® PowerPlex™ 16

PCR Product Sizes of Observed Alleles

Allele (Repeat #)	Set 1	Set 2	Set 3	Repeat Structure	Ref.
7	157 bp	123 bp	203 bp	$[TCTA]_7$	716
8	161 bp	127 bp	207 bp	$[TCTA]_8$	369
9	165 bp	131 bp	211 bp	$[TCTA]_9$	369
10	169 bp	135 bp	215 bp	$[TCTA]_{10}$	369
11	173 bp	139 bp	219 bp	$[TCTA]_{11}$	369
12	177 bp	143 bp	223 bp	$[TCTA]_{12}$	369
13	181 bp	147 bp	227 bp	$[TCTA]_1[TCTG]_1[TCTA]_{11}$	369
14	185 bp	151 bp	231 bp	$[TCTA]_1[TCTG]_1[TCTA]_{12}$	369
15	189 bp	155 bp	235 bp	$[TCTA]_1[TCTG]_1[TCTA]_{13}$	369
16	193 bp	159 bp	239 bp	$[TCTA]_2[TCTG]_1[TCTA]_{13}$	369
17	197 bp	163 bp	243 bp	$[TCTA]_2[TCTG]_2[TCTA]_{13}$	369
18	201 bp	167 bp	247 bp	$[TCTA]_2[TCTG]_1[TCTA]_{15}$	369
19	205 bp	171 bp	251 bp	$[TCTA]_2[TCTG]_2[TCTA]_{15}$	716

Additional Information

Allelic Ladders: Commercially available from Applied Biosystems, Promega
Common Multiplexes: Profiler Plus, PowerPlex 2.1, PowerPlex 16

Statistical Genetics

Markers: Microsatellites or STRs (Short Tandem Repeat)

- STR database

Number of base pairs

	Id	STR1	STR2	STR3	STR4	STR5	STR6	STR7	STR8	STR9	···
1	794	129	264	142	156	157	171	205	183	196	···
2	794	155	292	146	156	166	179	205	187	196	···
3	795	145	288	138	168	157	171	205	195	196	···
4	795	150	292	142	172	166	175	210	203	196	···
5	796	155	292	138	156	157	167	205	183	184	···
6	796	155	300	142	156	169	171	205	199	196	···
7	797	150	264	142	156	157	171	205	187	196	···
8	797	155	292	146	176	163	175	205	187	196	···
9	798	150	292	138	156	157	171	205	183	187	···
10	798	155	300	146	160	166	171	205	207	190	···
11	799	155	296	146	152	157	167	205	179	196	···
12	799	155	296	146	176	157	171	210	183	196	···
13	800	145	264	138	156	157	163	205	187	190	···
14	800	160	296	146	156	157	171	210	199	196	···
15	801	155	264	142	156	157	175	205	183	196	···
16	801	155	292	146	184	166	179	209	199	199	···
17	802	145	292	138	176	157	159	193	183	187	···
18	802	155	296	142	180	166	171	201	187	187	···
19	803	155	280	142	172	166	163	205	183	196	···
20	803	155	300	142	176	169	175	213	187	196	···
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	

Statistical Genetics

Markers: Microsatellites or STRs (Short Tandem Repeat)

- With the advent of next generation sequencing (NGS) or massive parallel sequencing (MPS) the full sequence of STR alleles can be determined.
- This has led to the discovery of additional genetic variation (more alleles).
- Thus, when analyzing STR data, one can compares STR length (length-based or LB) but can also introduce information about their sequence (sequence-based or SB).
  - For sequence-based STRs the alleles are coded with an identifier for the sequence.

Statistical Genetics

Markers: Single Nucleotide Polymorphism (SNP)

- A Single Nucleotide Polymorphism (SNP) is a variation at a single position in a DNA sequence among individuals.

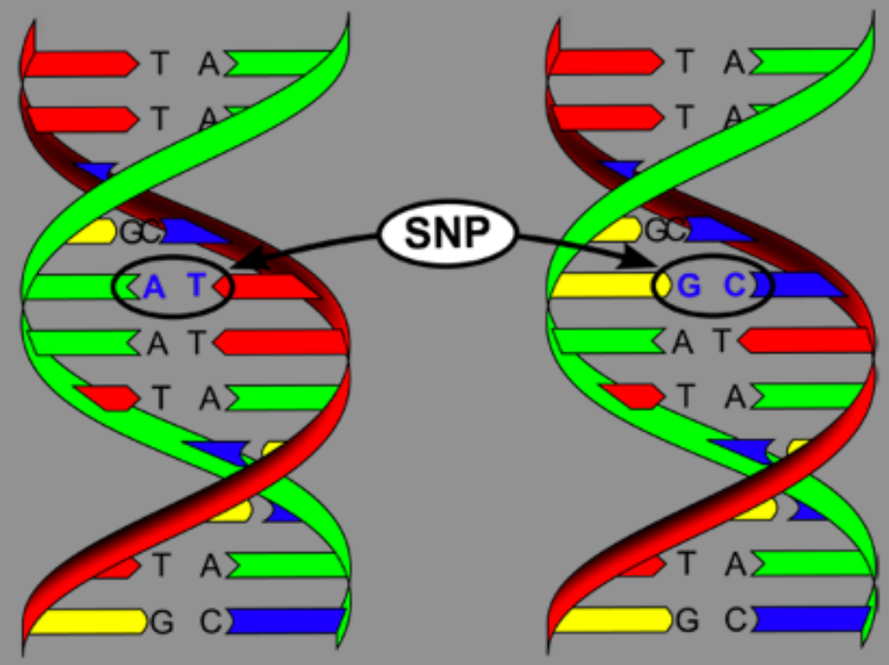
- There are four nucleotides (A,T,G and C).
  In theory, a SNP is a categorical variable with 10 possible categories (genotypes).

- In practice, the vast majority of SNPs is bi-allelic, so that only three genotypes occur.

  - Example: we may have a A/T polymorphism, with AA, AT and TT individuals.

- Bi-allelic SNPs often coded as count data (0=AA, 1=AB, 2=BB) of the minor allele.

- SNPs have become the most popular genetic markers.



Statistical Genetics

Markers: Single Nucleotide Polymorphism (SNP)

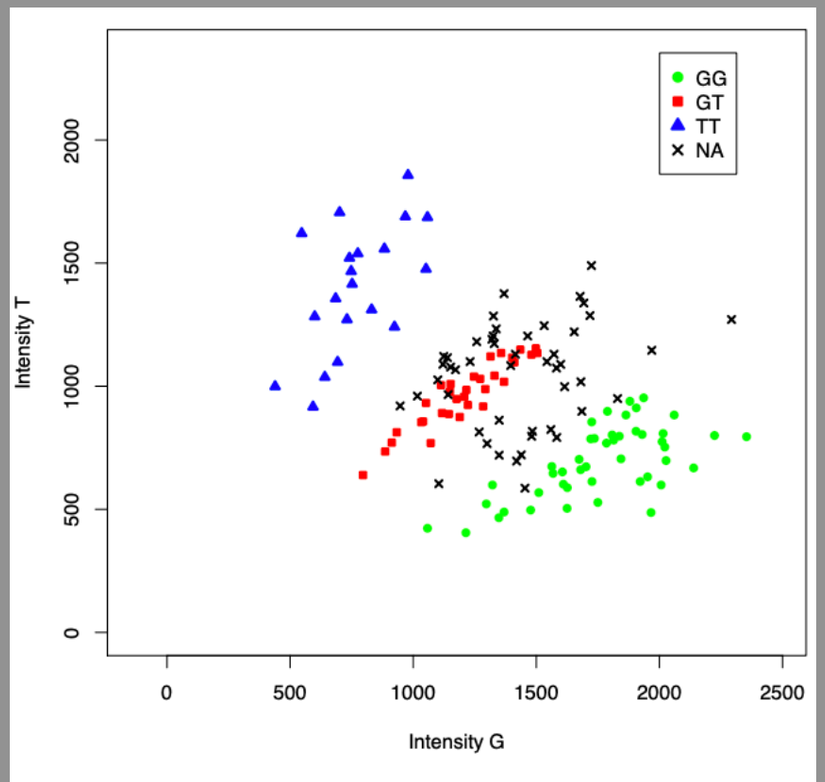
- SNP database

Id	SNP1	SNP2	SNP3	SNP4	SNP5	SNP6	SNP7	SNP8	SNP9	SNP10	···
NA18524	CC	CC	TT	TT	AT	AC	CC	AC	CT	GG	···
NA18526	CC	CC	CT	TT	AT	CC	CC	AC	CT	GG	···
NA18529	CC	CC	TT	TT	TT	AC	CG	AC	CT	GG	···
NA18532	CC	CC	TT	TT	TT	AC	CG	AC	CT	GG	···
NA18537	CC	CC	TT	TT	AT	CC	CC	AC	CT	GG	···
NA18540	CC	CC	CT	TT	AT	CC	CG	AC	CT	GG	···
NA18542	CC	CC	TT	TT	TT	CC	CG	AC	CT	GG	···
NA18545	CC	CC	CT	TT	AT	CC	CC	AC	CT	GG	···
NA18547	CC	CC	CT	TT	AT	CC	CC	AC	CT	GG	···
NA18550	CC	CC	CT	TT	AT	CC	CC	AC	CT	GG	···
NA18552	CC	CC	TT	TT	TT	CC	CG	AC	CT	GG	···
NA18555	CC	CC	TT	TT	TT	CC	CG	AC	CT	GG	···
NA18558	CC	NA	CC	TT	TT	CC	CG	CC	CT	GG	···
NA18561	CC	CC	TT	TT	TT	AC	CC	AC	CT	GG	···
NA18562	CC	CC	TT	TT	AT	AC	CG	AC	CT	GG	···
NA18563	CC	CC	CT	TT	AA	CC	CC	AA	CT	GG	···
NA18564	CC	CC	TT	TT	TT	AC	CC	AC	CT	GG	···
NA18566	CC	CC	TT	TT	TT	AC	CC	AC	CT	GG	···
NA18570	CC	CC	TT	TT	AT	AC	CC	AC	CT	GG	···
NA18571	CC	CC	TT	TT	AT	AC	CC	AC	CT	GG	···
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	

Statistical Genetics

Markers: Single Nucleotide Polymorphism (SNP)

- Missing data is a common problem (10% missing not unusual).
- SNP data is multivariate categorical data, but can also be considered count data.
- SNPs occur about once in every 300 base pairs. Over 84.7 million SNPs in the human genome (1000G; 2015).
- SNPs genotypes determined by a classification/clustering algorithm that use allele intensities.Given a test to genotype SNPs, we analyze SNP intensity, which lead to genotype clusters:



Content

Introduction to statistical genetics

1. Basic terminology
2. Traits and genetic markers
   - RFLPs (Restriction Fragment Length Polymorphism)
   - STRs (Short Tandem Repeat)
   - SNP (Single Nucleotide Polymorphism)
3. Descriptive analysis of a genetic marker
4. Computer exercise

Statistical Genetics

Descriptive analysis of a genetic marker

- Number and percentage of missing values (NA)
- Number of alleles (i.e. A and a)
- Genotype frequencies (i.e. AA, Aa or aa)
- Allele frequencies
- Heterozygosity (i.e. frequency of Aa)
- Minor and major allele
- Minor allele frequency (maf)

Statistical Genetics

Descriptive analysis of a genetic marker

- On bi-allelic makers
  - Allele frequencies: $p_A + p_B = 1$
  - MAF: $\min(p_A, p_B)$
  - Genotype frequencies: $f_{AA} + f_{AB} + f_{BB} = 1$
  - Observed heterozygosity: $H_o = f_{AB}$
  - Expected heterozygosity: $H_e = 1 - \sum_{i=1}^{K} p_i^2$
- Notes:
  - Note that for a genetic marker with K alleles there will be $\frac{1}{2}K(K+1)$ genotypes.
  - Mind the difference between population parameters and sample estimates.

Statistical Genetics

Descriptive analysis of a genetic marker

- On bi-allelic makers
  - Allele frequencies: $p_A + p_B = 1$
  - MAF: $\min(p_A, p_B)$
  - Genotype frequencies: $f_{AA} + f_{AB} + f_{BB} = 1$
  - Observed heterozygosity: $H_o = f_{AB}$
  - Expected heterozygosity: $H_e = 1 - \sum_{i=1}^{K} p_i^2$

Discrepancies give us information about the amount of genetic variation in populations

- High heterozygosity means lots of genetic variability.
- Low heterozygosity means little genetic variability.

For example…
- Ho < He inbreeding
- Ho > He mixing of previously isolated pop.

- Notes:
  - Note that for a genetic marker with K alleles there will be $\frac{1}{2}K(K+1)$ genotypes.
  - Mind the difference between population parameters and sample estimates.

Statistical Genetics

Descriptive analysis of a genetic marker

- Example of a bi-allelic maker (an A/T polymorphism with $n_{AA} = 46$, $n_{AT} = 39$ and $n_{TT} = 15$ counts:
  - $\hat{p}_A = \frac{n_{AA} + \frac{1}{2} n_{AT}}{n} = \frac{46 + \frac{1}{2} 39}{100} = 0.655$
  - $\hat{p}_T = \frac{n_{TT} + \frac{1}{2} n_{AT}}{n} = \frac{15 + \frac{1}{2} 39}{100} = 0.345$
  - $f_{AA} = \frac{n_{AA}}{n} = \frac{46}{100}$ and $f_{AT} = \frac{n_{AT}}{n} = \frac{39}{100}$ and $f_{TT} = \frac{n_{TT}}{n} = \frac{15}{100}$
- Allele frequencies: $p_A + p_T = 1$
- MAF: $\min(p_A, p_T) = 0.345$
- Genotype frequencies: $f_{AA} + f_{AT} + f_{TT} = 1$
- Observed heterozygosity: $H_o = f_{AT}$
- Expected heterozygosity: $H_e = 1 - \sum_{i=1}^{K} p_i^2 = 1 - (p_A^2 + p_T^2) = 0.445$

Statistical Genetics

Descriptive analysis of a genetic marker

- Public resources:

Catalogue of common human genetic variation, using openly consented samples from people who declared themselves to be healthy.

The 1000 genomes project: http://www.internationalgenome.org

Content

Introduction to statistical genetics

1. Basic terminology
2. Traits and genetic markers
   - RFLPs (Restriction Fragment Length Polymorphism)
   - STRs (Short Tandem Repeat)
   - SNP (Single Nucleotide Polymorphism)
3. Descriptive analysis of a genetic marker
4. Computer exercise

Statistical Genetics

Computer exercise STRs

1. Load http://www-eio.upc.es/~jan/Data/bsg/JapanaseSTRs.rda
2. Determine # STRs and # individuals in the database.
3. Change all -9 for NA.
4. Determine the number of alleles of the first STR.
5. Determine the allele counts for the first STR.
6. Determine the genotype counts for the first STR.
7. How many different genotypes are observed?
8. How many different genotypes are theoretically possible?
9. Determine the number of alleles for each STR, and make a barplot of the number of alleles.

Statistical Genetics

Computer exercise SNPs

1. Load http://www-eio.upc.es/~jan/Data/bsg/Chromosome1_CHBPopSubset.rda
2. Install the genetics package
3. Determine # SNPs and # individuals in the database
4. Change all NN for NA
5. Describe the first 3 SNPs with the summary command. Note that before you need to transform the SNP into the genotype class (i.e. SNP1.g <- genotype(SNP1,sep=""))
6. Compute the % of missings per individual and plot these
7. Compute the % of missings per SNP and plot these
8. Are there any individuals/SNPs with an exceptional amount of missing data?
9. Compute the allele frequencies of SNP3 from the genotype frequencies
10. Compute the MAF for all SNPs in the database, and make a histogram. What do you observe?

Statistical Genetics

References

- Foulkes, A.S. (2009) Applied statistical genetics with R. Springer.

- Laird, N.M. & Lange, C. (2011) The fundamentals of modern statistical genetics. Springer.

- Weir, B.S. (1996) Genetic Data Analysis II, Sinauer Associates, Massachusetts.