

1. An Introduction to Statistical Genetics

Iván Galván-Femenía¹, Jan Graffelman¹

¹Department of Statistics and Operations Research
Universitat Politècnica de Catalunya
Barcelona, Spain

November 3, 2016

ivan.galvan@upc.edu

Master in Innovation and Research in Informatics (MIRI)



UNIVERSITAT POLITÈCNICA
DE CATALUNYA
BARCELONATECH

Masters in **Computer Science** and **Engineering**

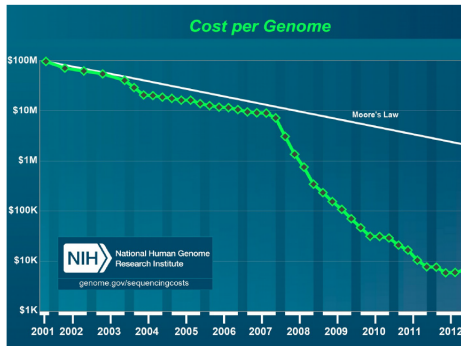
Planning for Statistical Genetics part of the BSG course

- 1 Introduction to statistical genetics
- 2 Hardy-Weinberg equilibrium
- 3 Linkage disequilibrium
- 4 Phase estimation
- 5 Population substructure
- 6 Genetic association analysis
- 7 Relatedness analysis (allele sharing)

*Computer exercises for each lesson (**R**, **Rstudio**)

Statistical genetics

- Statistical genetics is a branch of statistics that deals with the analysis of inherited traits and genetic data.
- Nowadays genetic data arises in different forms (sequences, markers,...).
- The size of the genetic databases has grown enormously over the years.

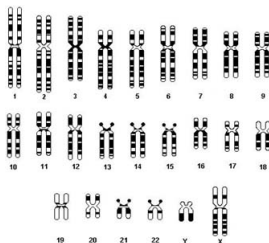


Type of studies in statistical genetics

- Population based studies (unrelated individuals)
- Family based studies (related individuals)

Some basic terminology I

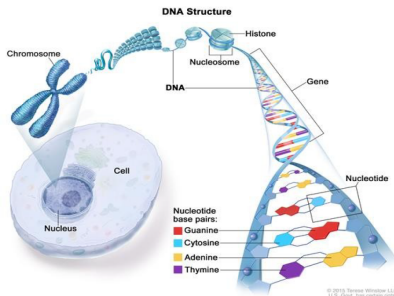
- A human being has 46 chromosomes in the nucleus of each cell, coming in 23 **homologous pairs** (22 pairs of **autosomes** and one pair of **sex** chromosomes (X/Y)).
- Of each pair one chromosome comes from the mother and one from the father.
- All genetic information of an individual constitutes his/her **genome**.
- > 3,000,000,000 base pairs (G, C, A, T).
- Humans share 99.9 % of DNA sequences.
- Human genome has more than 20,000 genes.



Chromosome	Length (mm)	Base pairs
1	85	249,250,621
2	83	243,199,373
3	67	198,022,430
4	65	191,154,276
5	62	180,915,260
6	58	171,115,067
7	54	159,138,663
8	50	146,364,022
9	48	141,213,431
10	46	135,534,747
11	46	135,006,516
12	45	133,851,895
13	39	115,169,878
14	36	107,349,540
15	35	102,531,392
16	31	90,354,753
17	28	81,195,210
18	27	78,077,248
19	20	59,128,983
20	21	63,025,520
21	16	48,129,895
22	17	51,304,566
X	53	155,270,560
Y	20	59,373,566

<https://en.wikipedia.org/wiki/Chromosome>

Some basic terminology II



- A **gene** corresponds to a piece of a chromosome, a stretch of DNA.
- The location of a gene in the genome is called its **locus**.
- Each individual inherits two copies of each gene, one from the father and one from the mother.
- The different forms 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 .
- The genetic makeup of an individual is called his/her **genotype**. For a gene with two alleles, this can be AA , Aa or aa .
- 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).

Traits and markers

- In many studies in statistical genetics, some trait (e.g. body mass index (BMI) or disease status) of an organism is considered to depend on one or more genetic variables.
- The position of genetic factors affecting a trait is often unknown.
- A marker is a genetic variable that shows variation over individuals, and has a known locus.
- The study of associations between markers and trait can be helpful in identifying the genetic factors that affect the trait (Lesson 6).

Markers

There are many different markers of which we consider

- SNPs (**Single Nucleotide Polymorphism**)
- Microsatellites or STRs (**Short Tandem Repeat**)
- RFLPs (Restriction Fragment Length Polymorphism)
- indels (insertion/deletion polymorphism)
- ...

Microsatellites or STRs (Short Tandem Repeat, 1/2)

- 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.
- Individuals vary in the number of repeats they have.
- Produces count data, with a limited number of outcomes.
- Microsatellites have many alleles.

Microsatellites or STRs (Short Tandem Repeat, 2/2)

- 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.

A glance at a STR database

	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	...
.
.
.

Single Nucleotide Polymorphism

- A Single Nucleotide Polymorphism (SNP) corresponds to the position of one base pair in the DNA chain.
- There are four nucleotides (A,T,G and C).
- In theory, a SNP is a categorical variable with $4 \times 4 = 16$ possible categories.
- In practice, the vast majority of SNPs is bi-allelic, so that only three genotypes occur.
- E.g. we may have a A/T polymorphism, with AA, AT and TT individuals.
- SNPs have become the most popular genetic markers.

A glance at a SNP database

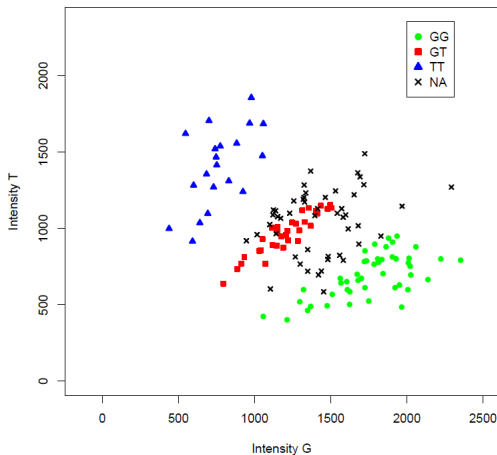
Id	SNP1	SNP2	SNP3	SNP4	SNP5	SNP6	SNP7	SNP8	SNP9	SNP10	...
NA18524	CC	CC	TT	TT	AT	AC	CC	AC	CT	GG	...
NA18526	CC	CC	CT	TT	AT	CC	CC	AC	CT	GG	...
NA18529	CC	CC	TT	TT	TT	AC	CG	AC	CT	GG	...
NA18532	CC	CC	TT	TT	TT	AC	CG	AC	CT	GG	...
NA18537	CC	CC	TT	TT	AT	CC	CC	AC	CT	GG	...
NA18540	CC	CC	CT	TT	AT	CC	CG	AC	CT	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	...
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	

SNP data

Note that:

- Missing data is a common problem (10% missing not unusual).
- SNP data is multivariate categorical data.
- SNPs occur about once in every 300 base pairs.
- Approximately 10 million SNPs in the human genome.
- Genotypes determined by a classification/clustering algorithm that use allele intensities.

Example of a Call plot



Elements of the descriptive analysis of one genetic marker

- Number and percentage of missing values (NA).
- Number of alleles.
- Genotype frequencies.
- Allele frequencies.
- Heterozygosity.
- Minor allele frequency (MAF).

Some notation (bi-allelic markers)

- 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.

Reading genetic data in R

```
load("c:/data/Chromosome1_CHBPopSubset.rda")
install.packages("genetics")
library(genetics)
Ysub[Ysub=="NN"] <- NA
Ysub[,1:5]
Ysub[,3]
Geno <- genotype(Ysub[,3],sep="")
Geno
summary(Geno)
```

Summarizing a genetic marker in R

```
> Geno
[1] "T/T" "T/C" "T/T" "T/T" "T/T" "T/C" "T/T" "T/C" "T/C" "T/C" "T/T" "T/T"
[13] "C/C" "T/T" "T/T" "T/C" "T/T" "T/T" "T/T" "T/T" "T/C" "T/T" "T/T" "T/T"
[25] "T/C" "T/T" "T/C" NA    "T/C" "T/C" "T/T" "T/T" "T/T" "T/T" "T/T" "T/T"
[37] "T/T" NA    "T/T" "T/T" "T/C" "T/T" "T/T" "T/T" "T/T" "C/C" "T/T" "T/C"
Alleles: T C
> summary(Geno)

Number of samples typed: 46 (95.8%)

Allele Frequency: (2 alleles)
      Count Proportion
T         75         0.82
C         17         0.18
NA          4          NA

Genotype Frequency:
      Count Proportion
T/T      31         0.67
T/C      13         0.28
C/C       2         0.04
NA         2          NA

Heterozygosity (Hu) = 0.3045867
Poly. Inf. Content  = 0.2558924
>
```

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.

Computer exercise: SNPs

- ➊ Load http://www-eio.upc.es/~jan/data/bsg/Chromosome1_CHBPopSubset.rda
- ➋ Install the genetics package.
- ➌ Determine # SNPs and # individuals in the database.
- ➍ Change all NN for NA.
- ➎ Describe the first 3 SNPs with the summary command.
- ➏ Compute the % of missings per individual and plot these.
- ➐ Compute the % of missings per SNP and plot these.
- ➑ Are there any individuals/SNPs with an exceptional amount of missing data?
- ➒ Compute the allele frequencies of SNP3 from the genotype frequencies.
- ➓ Compute the MAF for all SNPs in the database, and make a histogram. What do you observe?

Computer exercise: STRs

- 1 Load <http://www-eio.upc.es/~jan/data/bsg/JapaneseSTRs.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.