Farielberry Lab

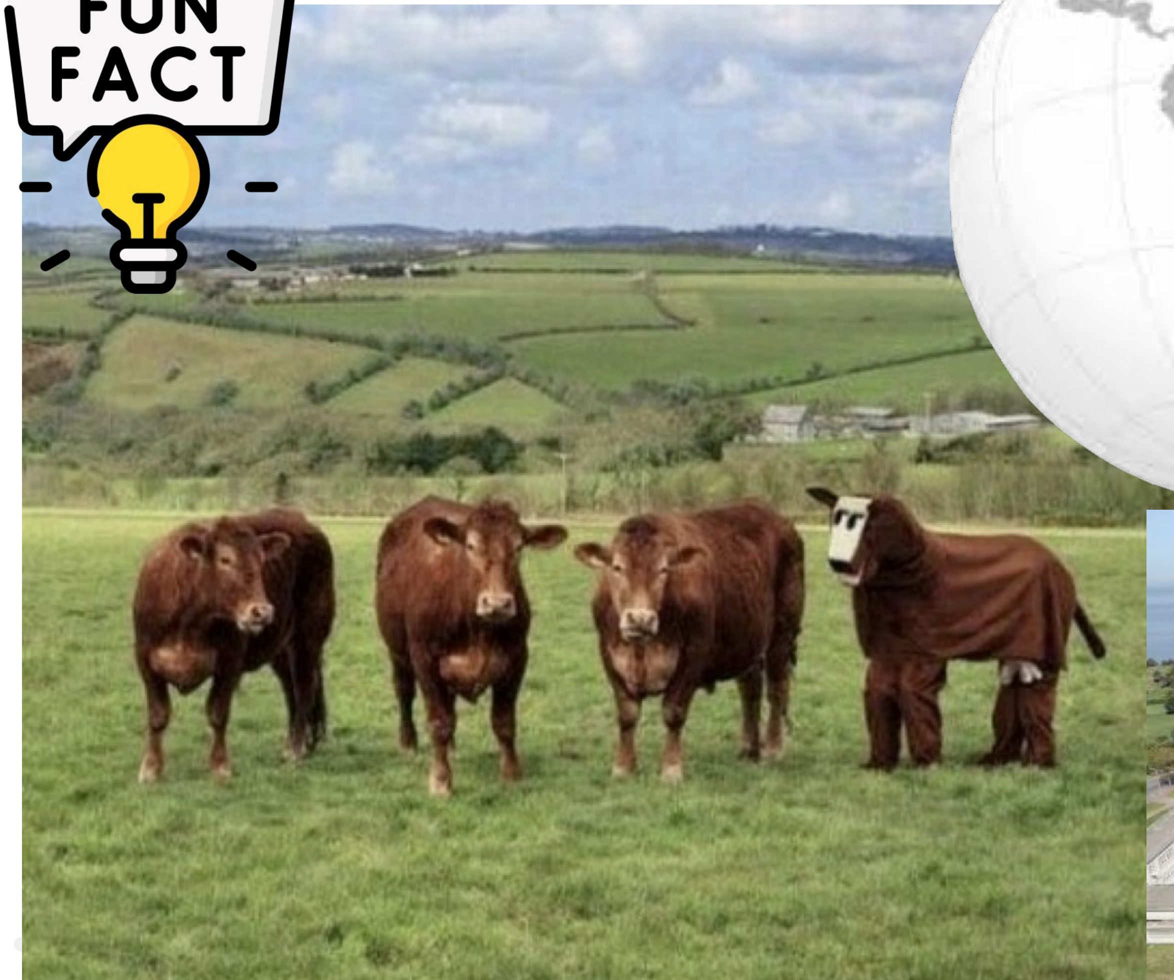
Deep Learning for Genomic Prediction

FACULTAD DE INGENIERÍA

UNIVERSIDAD DE LA REPÚBLICA URUGUAY

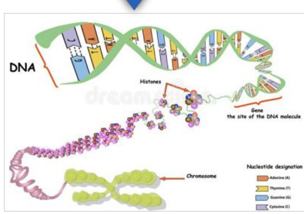
Uruguay

FUN
FACT

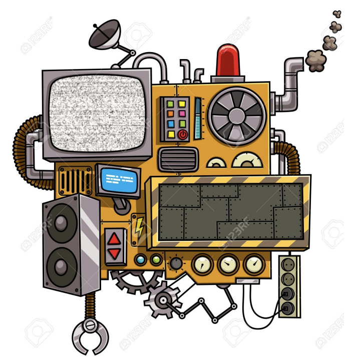


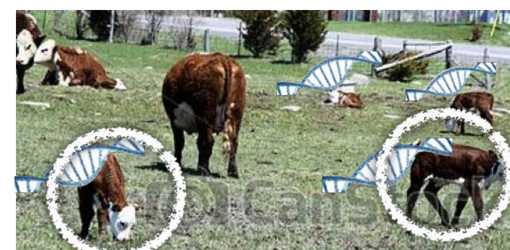
Genomic prediction

Phenotype: observable traits of an individual (traits, disease resistance, production).



Phenotype + Genotype





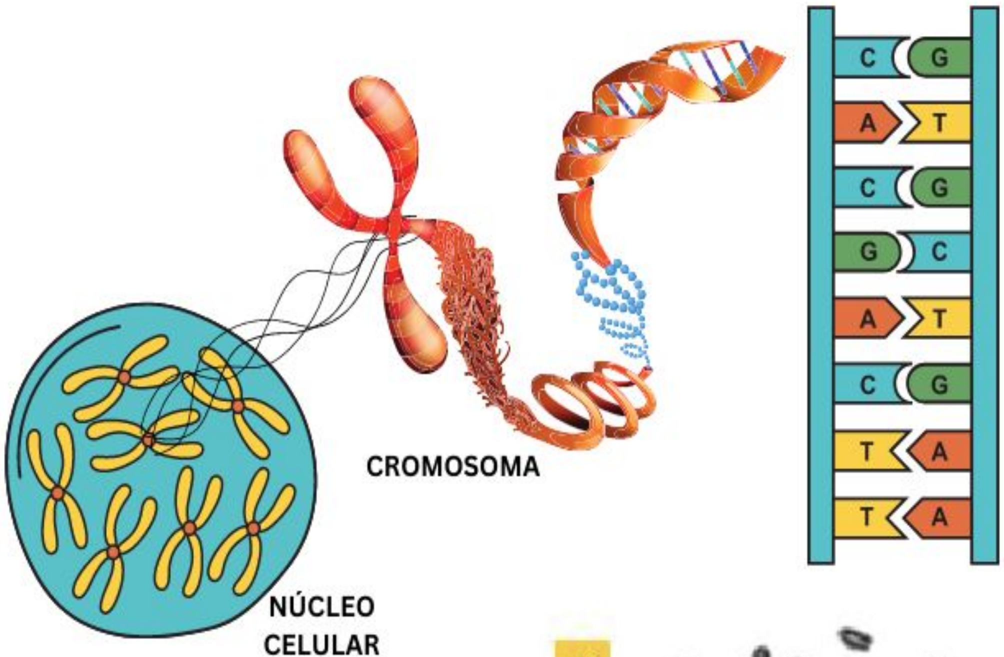
Farielberry Lab

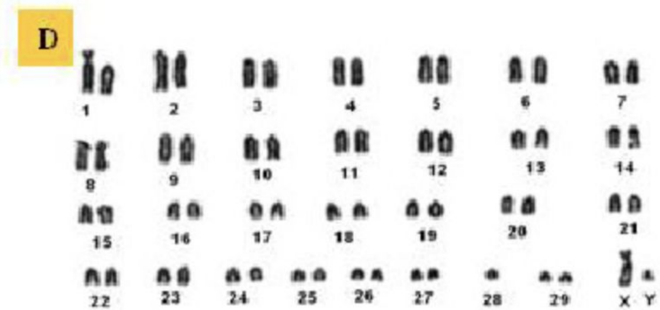
Genomics basics

FACULTAD DE INGENIERÍA

UNIVERSIDAD DE LA REPÚBLICA URUGUAY

DNA

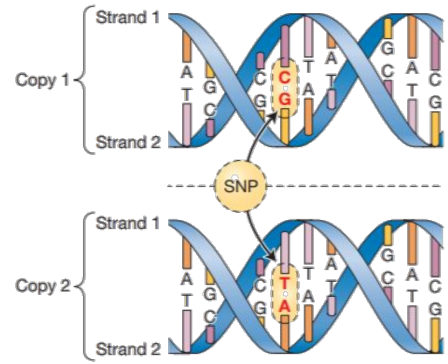


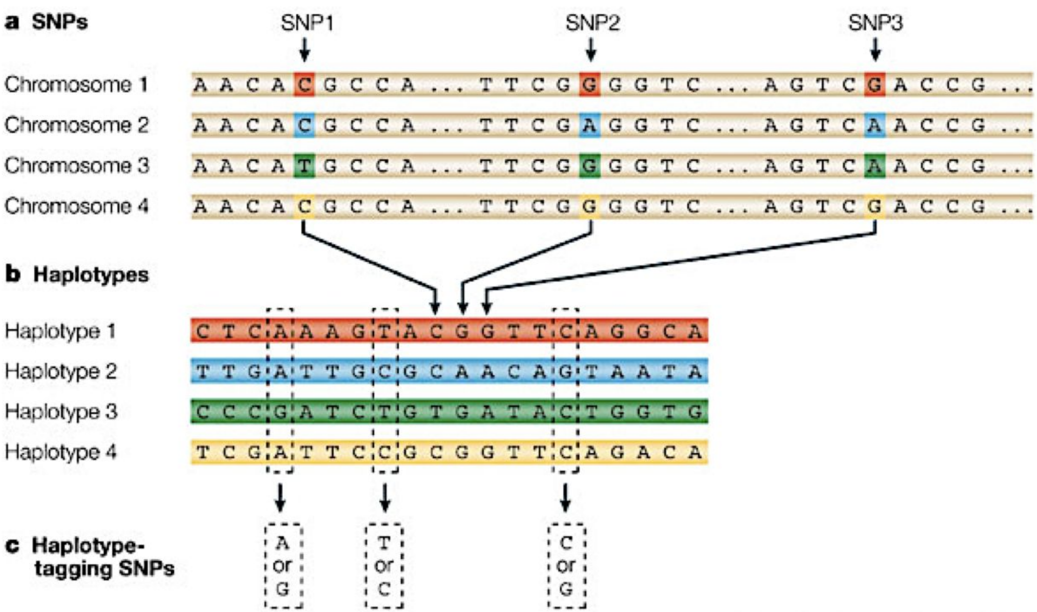


https://www.funpecrp.com.br/gmr/year2012/vol11-1/pdf/gmr1626.pdf

Single Nucleotide Polymorphism (SNP)

Strand 2:
Copy 1: TCC**C**TAGAC
Copy 2: TCC**T**TAGAC





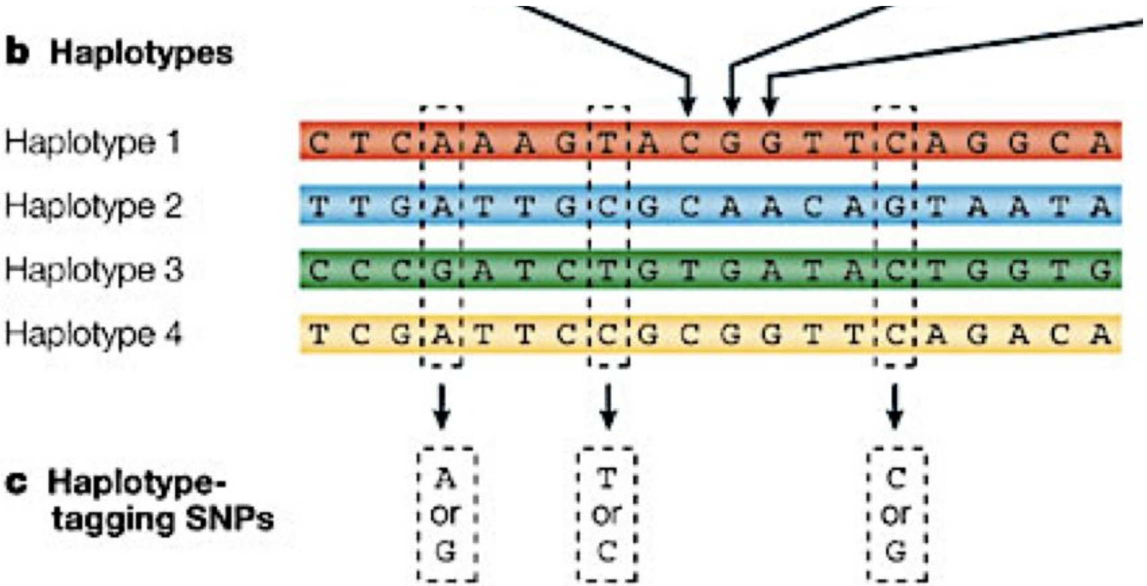
Nature Reviews | Immunology

Whole genome sequencing (WGS)

WGS after variant calling

SNP array

Haplotypes

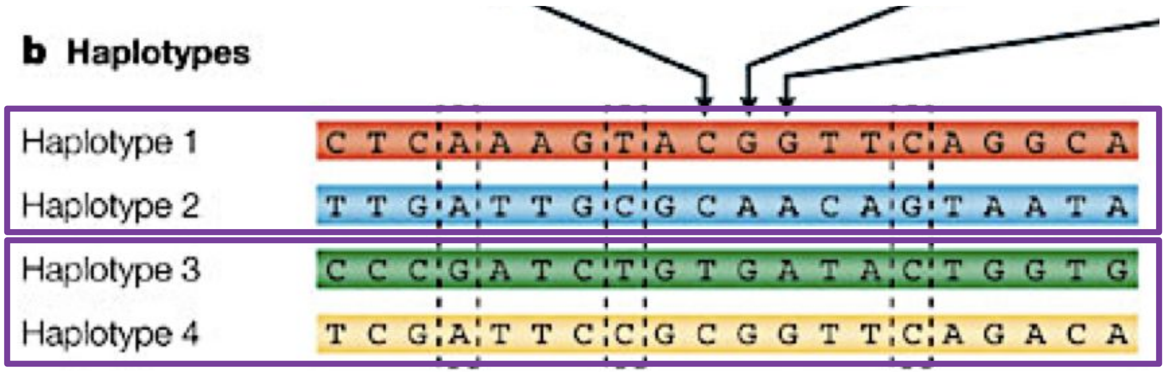


Bi-allelic SNPs:

- **1**: less frequent
- **0**: more frequent

Haplotype 1:	0	0	0
Haplotype 2:	0	1	1
Haplotype 3:	1	0	0
Haplotype 4:	0	1	0

Genotypes in diploid individuals



Genotype 1

Genotype 2

Additive codification

0+0= 0

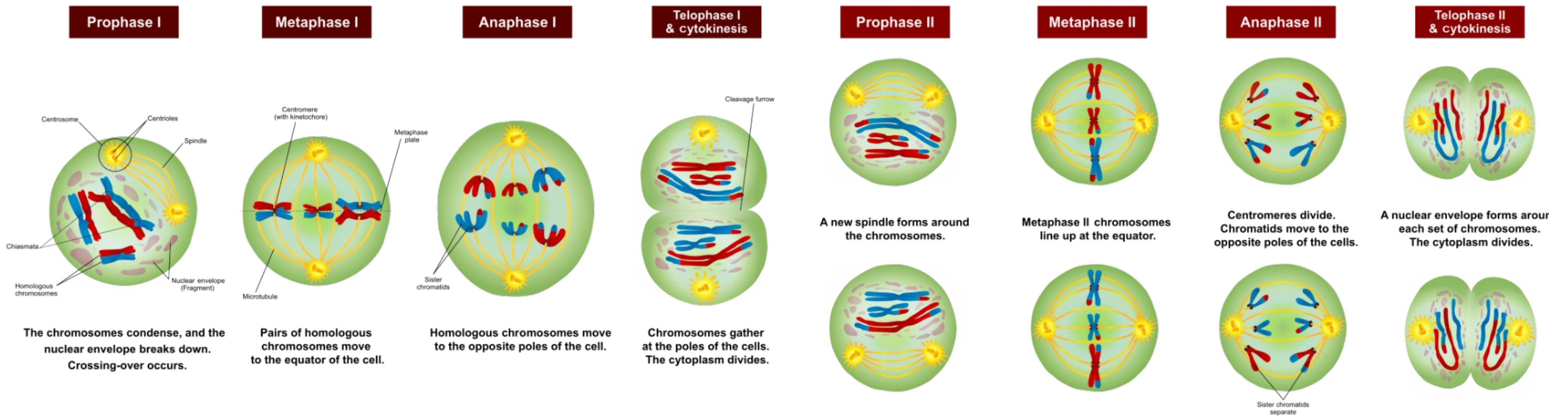
0+1= 1

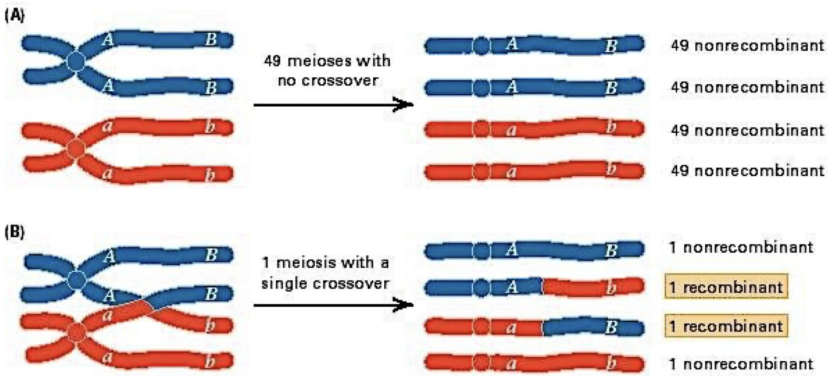
1+1= 2

Haplotype 1:	0	0	0	1
Haplotype 2:	0	1	1	1
Haplotype 3:	1	0	0	0
Haplotype 4:	0	1	0	0

Genotype 1:	0	1	1	2
Genotype 2:	1	1	0	0

Linkage disequilibrium (LD)





Humans: $r \approx 10^{-8}$ (1cM/Mb)

Recombination "probability": $$r = \frac{1+1}{4 \cdot 49 + 4 \cdot 1} = \frac{2}{200}$$

https://en.wikipedia.org/wiki/Meiosis

Linkage disequilibrium (LD)

SNP 2

		B	b	
SNP 1	A	p_{AB}	p_{Ab}	p_A
	a	p_{Ab}	p_{ab}	p_a
		p_B	p_b	1

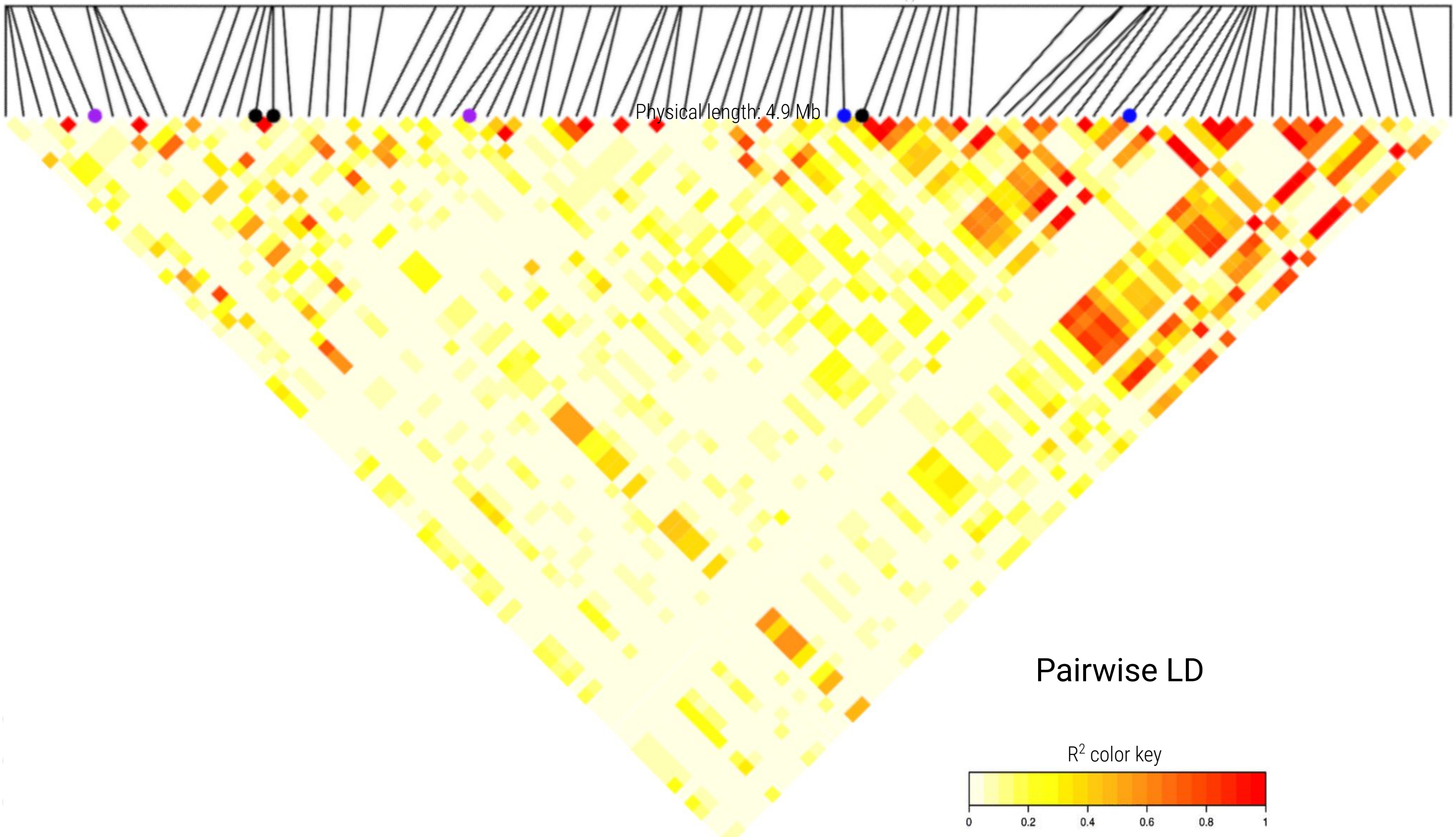
Under equilibrium (independence)

$$p_{AB} = p_A \cdot p_B$$

Linkage disequilibrium

$$D_{AB} = p_{AB} - p_A \cdot p_B$$

SNPs can be in LD despite being far away.



Chen et al. (2017). Genetics, 206(4), 1791-1806.

Data preparation

Variant call format (.vcf)

```
##fileformat=VCFv4.3
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS      ID         REF   ALT    QUAL  FILTER   INFO                              FORMAT        NA00001          NA00002          NA00003
20     14370    rs6054257  G     A      29    PASS     NS=3;DP=14;AF=0.5;DB;H2           GT:GQ:DP:HQ   0|0:48:1:51,51   1|0:48:8:51,51   1/1:43:5:.,.
20     17330    .          T     A      3     q10      NS=3;DP=11;AF=0.017               GT:GQ:DP:HQ   0|0:49:3:58,50   0|1:3:5:65,3     0/0:41:3
20     1110696  rs6040355  A     G,T    67    PASS     NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ   1|2:21:6:23,27   2|1:2:0:18,2     2/2:35:4
20     1230237  .          T     .      47    PASS     NS=3;DP=13;AA=T                   GT:GQ:DP:HQ   0|0:54:7:56,60   0|0:48:4:51,51   0/0:61:2
20     1234567  microsat1  GTC   G,GTCT 50    PASS     NS=3;DP=9;AA=G                    GT:GQ:DP      0/1:35:4         0/2:17:2         1/1:40:3
```

https://en.wikipedia.org/wiki/Variant_Call_Format

But we want 0s, 1s and 2s

$$X = \left.\underbrace{\begin{pmatrix} 1 & 0 & 2 & \dots & 0 & 0 & 2 \\ 0 & 1 & 2 & \dots & 0 & 2 & 2 \\ 0 & 1 & 2 & \dots & 1 & 0 & 1 \\ 0 & 1 & 1 & \dots & 2 & 1 & 1 \\ 1 & 2 & 2 & \dots & 0 & 0 & 1 \end{pmatrix}}_{\text{p SNPs}}\right\}\text{n individuals}$$

Plink

PLINK 1.9 | PLINK 2.0 home | plink2-users | Error messages | **File formats** | PLINK 2.0 index

Introduction, downloads
D: 15 Sep 2023
Recent version history
What's new?
Coming next

[Jump to search box]

General usage
Getting started
Flag usage summaries
Column set descriptors
Citation instructions

Standard data input
PLINK 1 binary (.bed)
PLINK 2 binary (.pgen)
Autoconversion behavior
VCF/BCF (.vcf[.gz], .bcf)
Oxford genotype (.bgen)
Oxford haplotype (.haps)
PLINK 1 text (.ped, .tped)
PLINK 1 dosage
Sample ID conversion
Dosage import settings
Generate random
Unusual chromosome IDs
Allele frequencies
Phenotypes
Covariates
'Cluster' import
Reference genome (.fa)

Input filtering
Sample ID file
Variant ID file
Interval-BED file
--extract-col-cond
QUAL, FILTER, INFO
Chromosomes
SNPs only
Simple variant window
Multiple variant ranges
Deduplicate variants
Sample/variant thinning
Pheno./covar. condition
Missingness
Category subset
--keep-col-match
Missing genotypes
Number of distinct alleles
Allele frequencies/counts
Hardy-Weinberg

File format reference

This page describes specialized PLINK 2.0 input and output file formats which are identifiable by file extension. (Most extensions not listed here have very simple one-entry-per-line or two-entry-per-line text formats.)

Unless otherwise specified, all multicolumn text files generated by PLINK 2.0 are tab-delimited, with one header line starting with '#'. In the column summaries, columns which are present unless removed by the column set descriptor are **boldface**, and columns which only appear under some data/flag/modifier combination(s) are *italicized*.

Jump to: .acount | .adjusted | .afreq | .bcf | .bed | .bgen | .bim | .bins | .clumps | .cov | .eigenvec{,.allele,.var} | .fam | .fst.summary | .fst.var | .gcount | .gen | .glm.firth | .glm.linear | .glm.logistic[.hybrid] | .grm | .grm.N.bin | .grm.bin | .haps | .hardy | .hardy.x | .het | .*.id | .kin0 | .king[.bin] | .legend | .map | .pdiff | .ped | **.pgen{,.pgi}** | **.psam** | **.pvar** | .raw | .rel[.bin] | .sample | .scount | .sdiff | .sdiff.summary | .smiss | .sscore | .ssf.tsv | .tfam | .tped | .traw | .vcf | .vmiss | .vscore | .vscore.bin

.acount, .afreq (allele count/frequency report)

Produced by --freq.

A text file with a header line, and then one line per variant with the following columns:

Header	Column set	Contents
CHROM	**chrom**	Chromosome code
POS	pos	Base-pair coordinate
ID	**(required)**	Variant ID
REF	**ref**	Reference allele
ALT1	alt1	Alternate allele 1
ALT	**alt**	All alternate alleles, comma-separated
PROVISIONAL_REF?	**maybeprovref**, provref	Reports whether REF allele is provisional
'REF_FREQ'/'REF_CT'	reffreq	Reference allele frequency/dosage
'ALT1_FREQ'/'ALT1_CT'	alt1freq	Alternate allele 1 frequency/dosage
'ALT_FREQS'/'ALT_CTS'	**altfreq**, alteq, alteqz	Comma-separated freqs/dosages for all alts; 'e requests '1=<ALT1 value>,2=<ALT2 value>,...' formatting with zero-values omitted, 'eqz' includes

Buscar proyectos

vcfpy 0.13.6

```
pip install vcfpy
```

Introduction to vcfR

Brian J. Knaus

2023-02-10

vcfR is a package intended to help visualize, manipulate and quality filter data in VCF files.

More documentation for vcfR can be found at the vcfR documentation website.

Plink is widely use and really easy (command lines)

Easy to change between different data formats.

Standard data input

- PLINK 1 binary (.bed)
- PLINK 2 binary (.pgen)
  - Autoconversion behavior
- VCF/BCF (.vcf[.gz], .bcf)
- Oxford genotype (.bgen)
- Oxford haplotype (.haps)
- PLINK 1 text (.ped, .tped)
- PLINK 1 dosage
- Sample ID conversion
- Dosage import settings
- Generate random
- Unusual chromosome IDs
- Allele frequencies
- Phenotypes
- Covariates
- 'Cluster' import
- Reference genome (.fa)

Linkage disequilibrium filter: keep "independent" SNPs

Linkage disequilibrium

All of the following calculations only consider founders. If your dataset has a shortage of them, PLINK 1.9 --make-founders may come in handy.

Since two-variant r^2 only makes sense for biallelic variants, these collapse multiallelic variants down to most common allele vs. the rest.

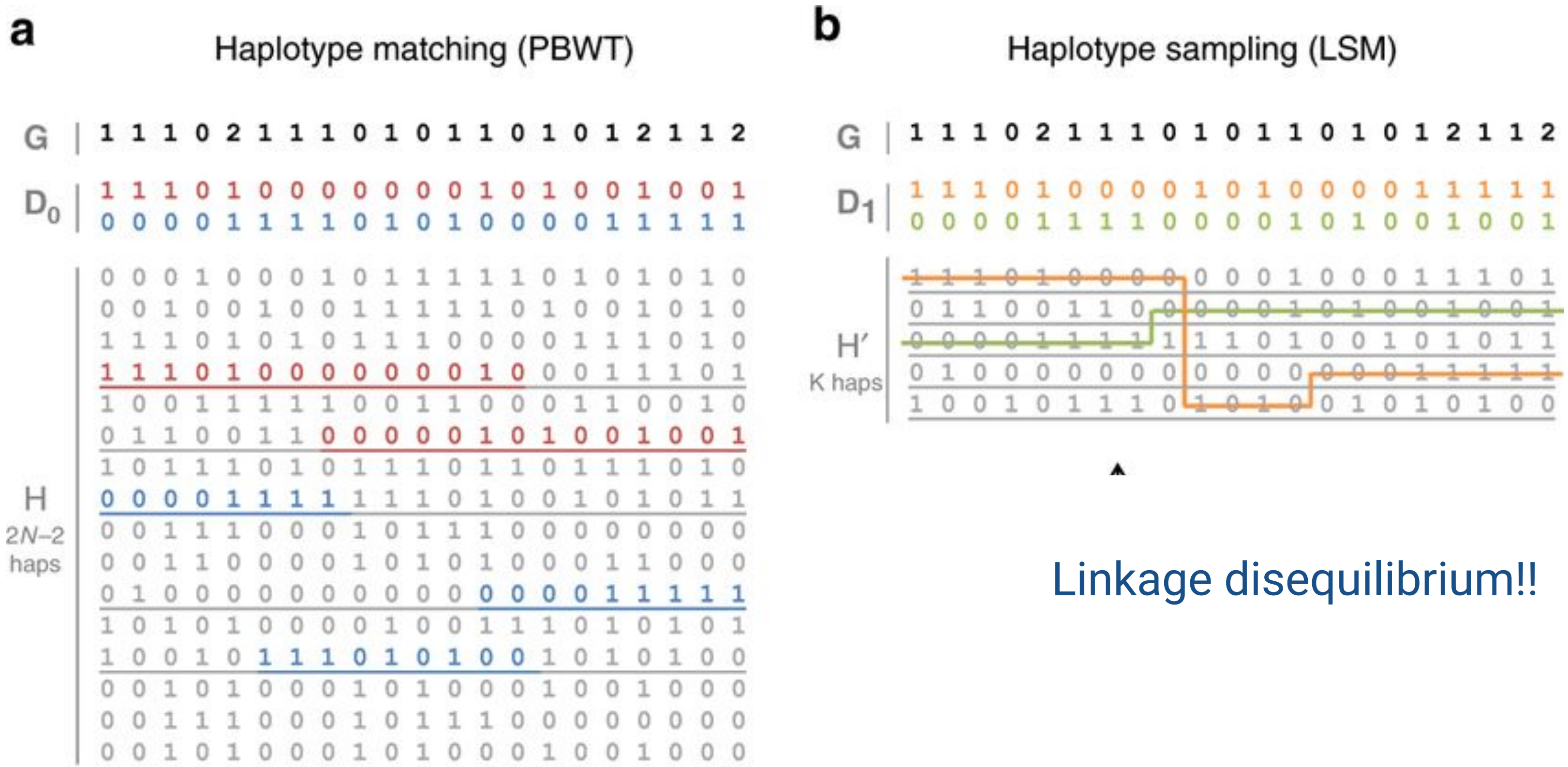
Variant pruning

```
--indep-pairwise <window size>['kb'] [step size (variant ct)]
                 <unphased-hardcall-r^2 threshold>
--indep-pairphase <window size>['kb'] [step size (variant ct)]
                  <phased-hardcall-r^2 threshold>
--indep <window size>['kb'] [step size (variant ct)] <VIF threshold>
--indep-order <mode>
```

Other filters:

- HW (Hardy-Weinberg Equilibrium)
- MAF (Minor Allele Frequencies)

Phasing: from genotypes to haplotypes



Linkage disequilibrium!!

https://www.researchgate.net/publication/337603515_Accurate_scalable_and_integrative_haplotype_estimation
https://github.com/odelaneau/shapeit5

Imputation

Typical imputation scenario

Linkage disequilibrium!!

HapMap or 1,000 Genomes — **Reference haplotypes**

0	0	1	1	1	0	0	1	1	0	0	0	1	1	1
0	0	0	0	0	1	1	1	0	1	1	1	0	0	1
1	1	1	1	1	0	0	0	1	0	0	0	0	0	0
1	0	1	1	0	0	0	1	1	1	1	1	0	0	1

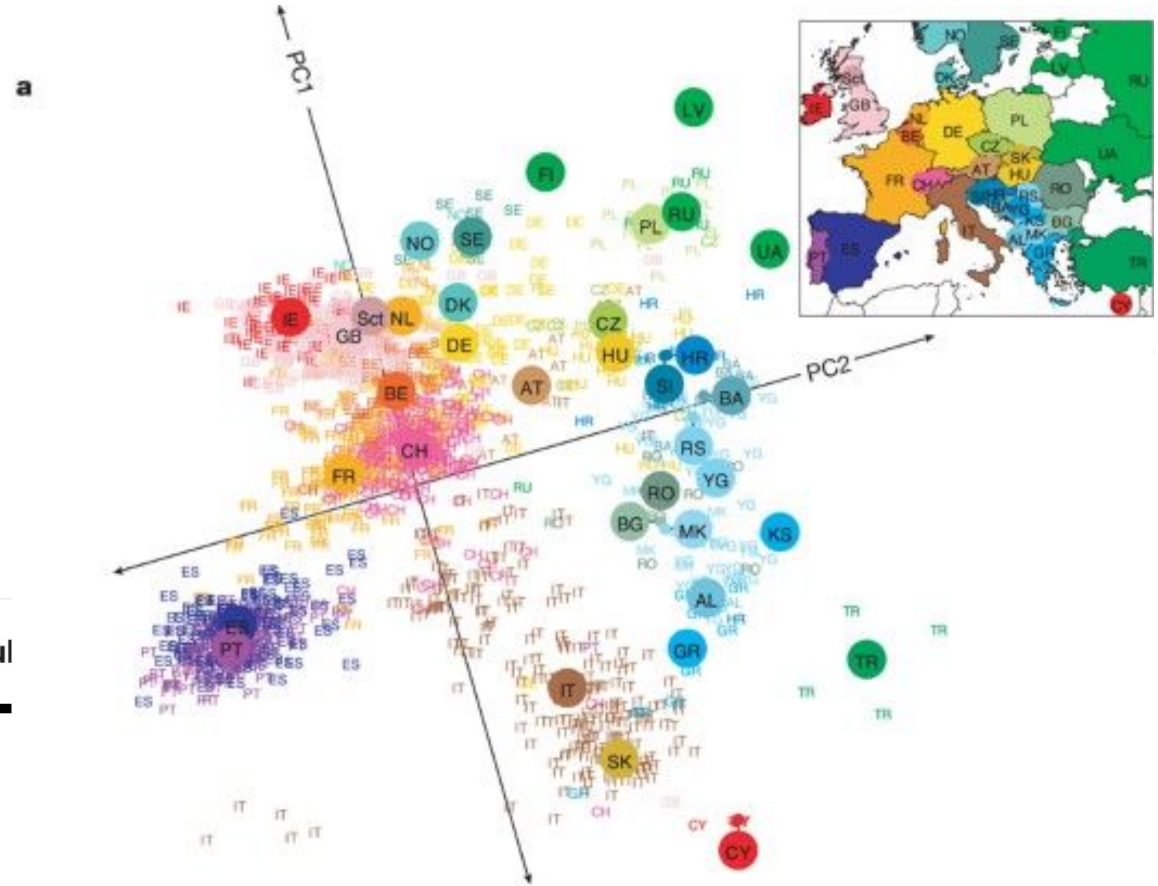
Cases and controls typed on SNP chip — **Study genotypes**

1	?	?	?	2	?	0	?	?	?	?	0	1	?	1
1	?	?	?	1	?	0	?	?	?	?	?	0	?	0
0	?	?	?	1	?	1	?	?	?	?	1	0	?	1
1	?	?	?	2	?	0	?	?	?	?	0	1	?	1
?	?	?	?	2	?	0	?	?	?	?	0	0	?	0
1	?	?	?	1	?	1	?	?	?	?	1	0	?	?
0	?	?	?	2	?	0	?	?	?	?	0	1	?	1
1	?	?	?	1	?	1	?	?	?	?	1	1	?	2

https://mathgen.stats.ox.ac.uk/impute/impute_v2.html

Data visualization

Is important to control for population structure or other sampling biases!



nature

Explore content ⌄ About the journal ⌄ Pul

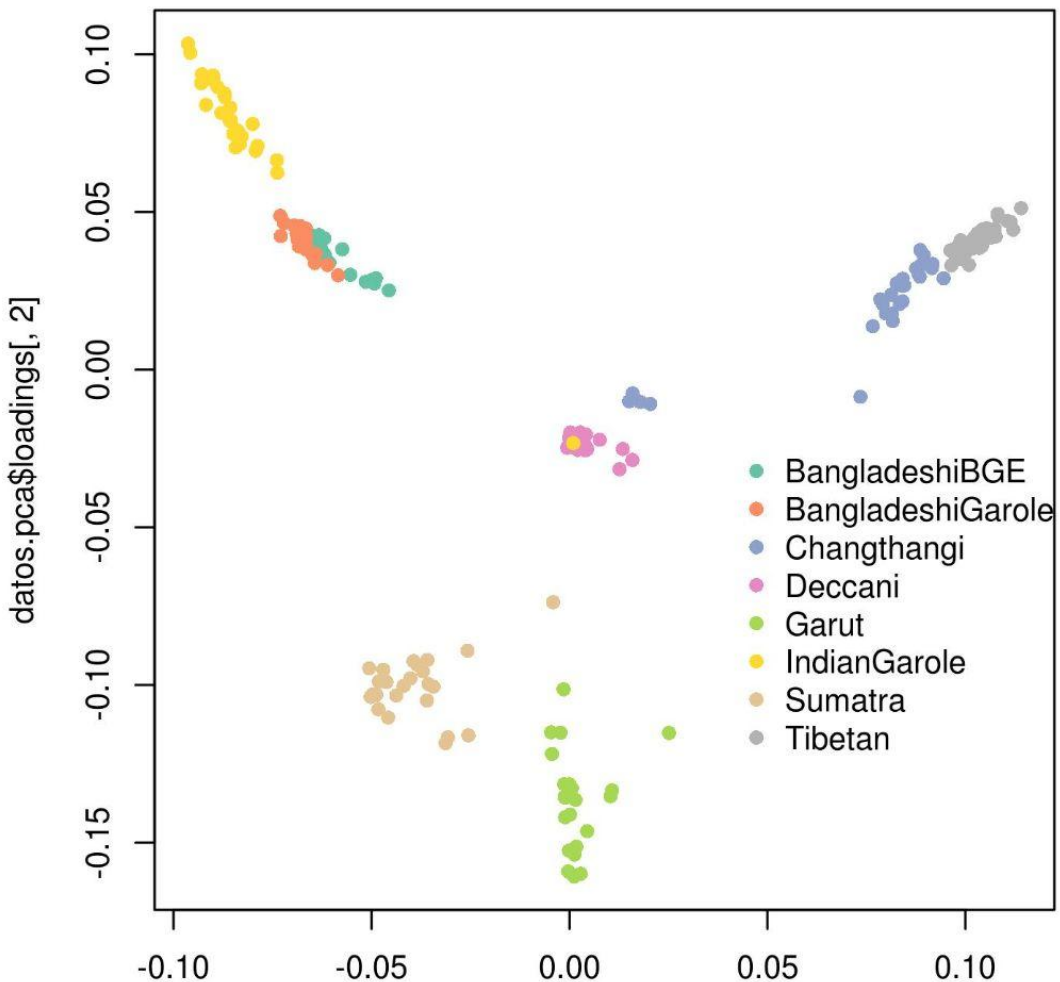
nature › letters › article

Published: 31 August 2008

Genes mirror geography within Europe

John Novembre, Toby Johnson, Katarzyna Bryc, Zoltán Kutalik, Adam R. Boyko, Adam Auton, Amit Indap, Karen S. King, Sven Bergmann, Matthew R. Nelson, Matthew Stephens & Carlos D. Bustamante

There could be mislabeled or incongruent data!



Sometimes not checking the data could lead to retracting articles

Retracted article

See the retraction notice

Science. 2010 Jul 1;2010. doi: 10.1126/science.1190532. Epub 2010 Jul 1.

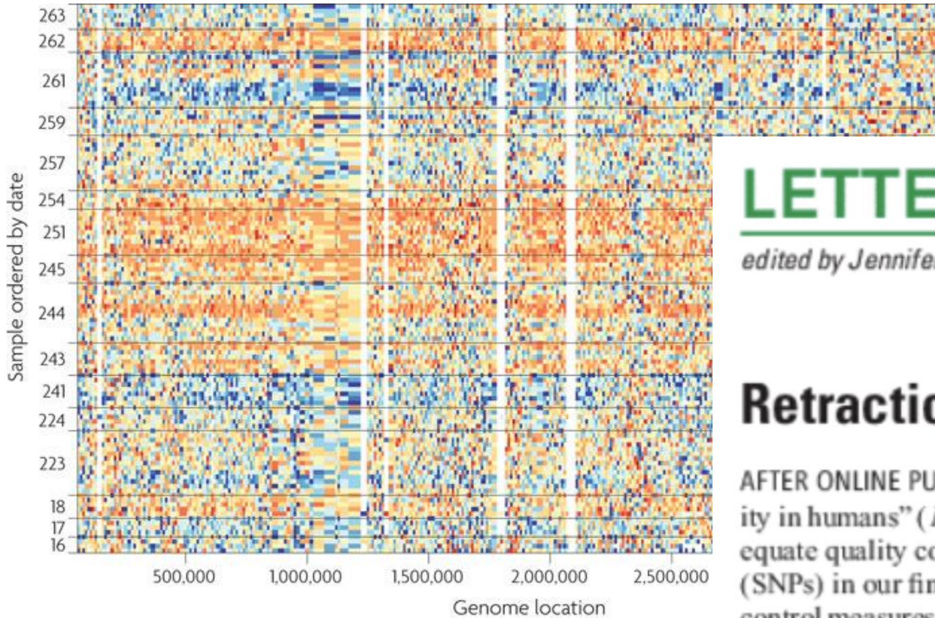
Genetic signatures of exceptional longevity in humans

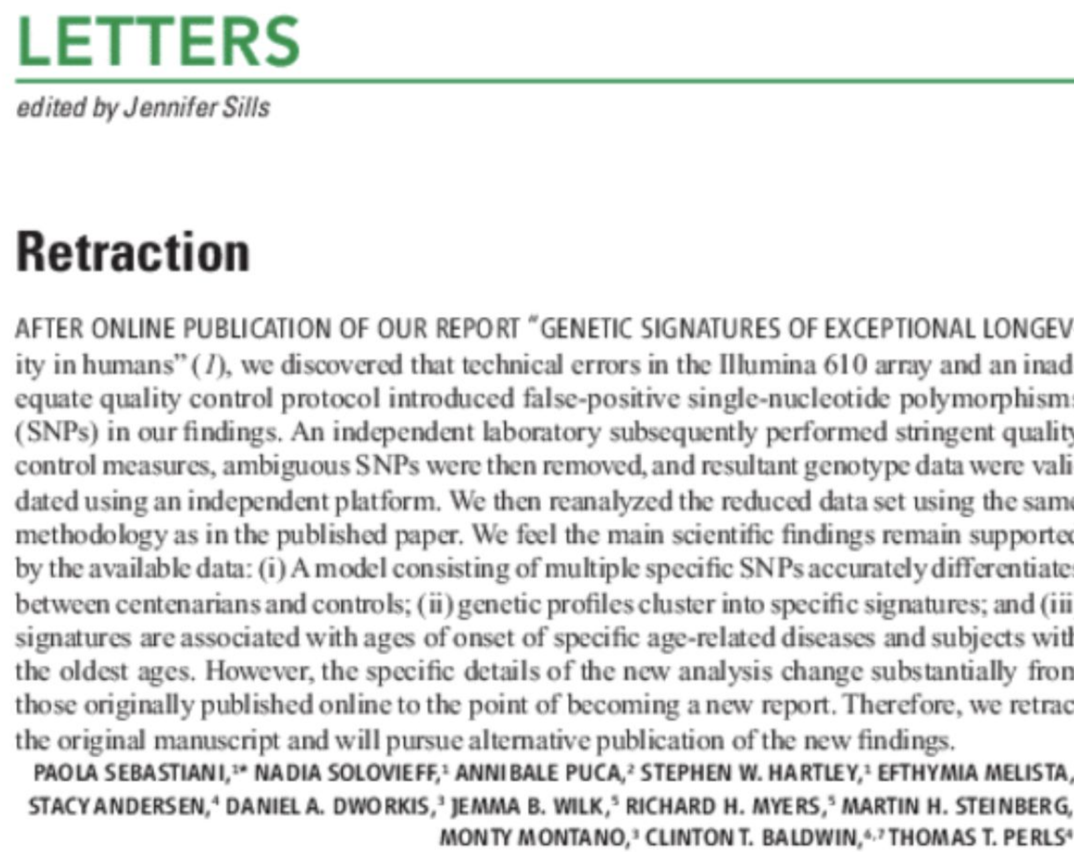
Paola Sebastiani [1], Nadia Solovieff, Annibale Puca, Stephen W Hartley, Efthymia Melista, Stacy Andersen, Daniel A Dworkis, Jemma B Wilk, Richard H Myers, Martin H Steinberg, Monty Montano, Clinton T Baldwin, Thomas T Perls

Affiliations + expand

PMID: 20595579 DOI: 10.1126/science.1190532 Paperpile

Confounded array type with the outcome





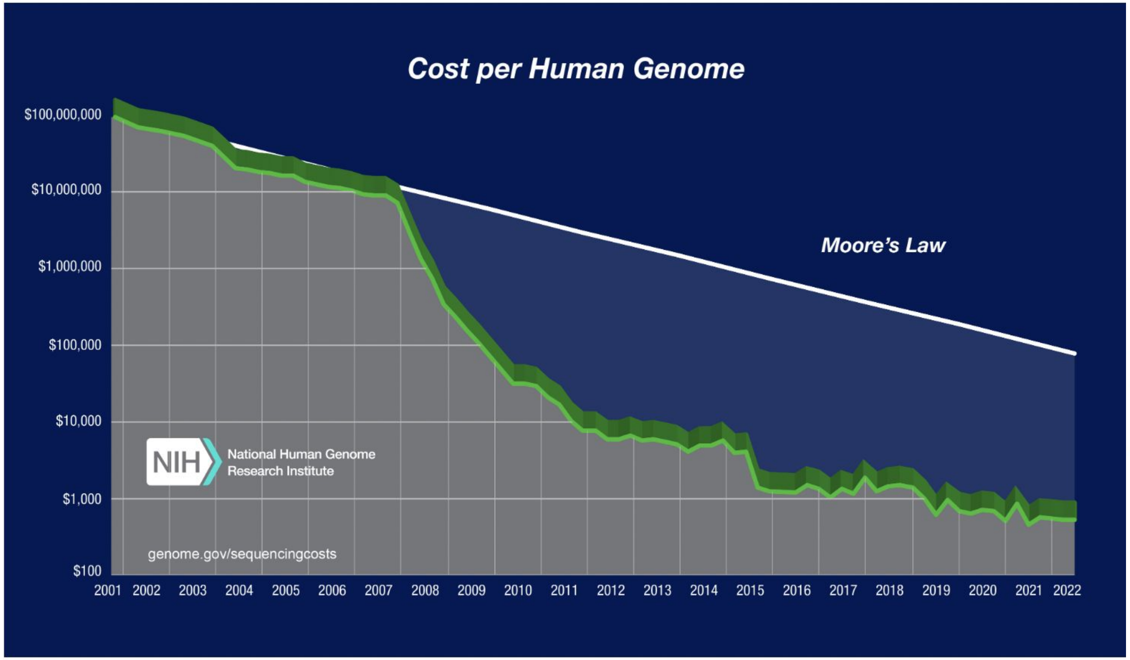
Farielberry Lab

Genomic prediction

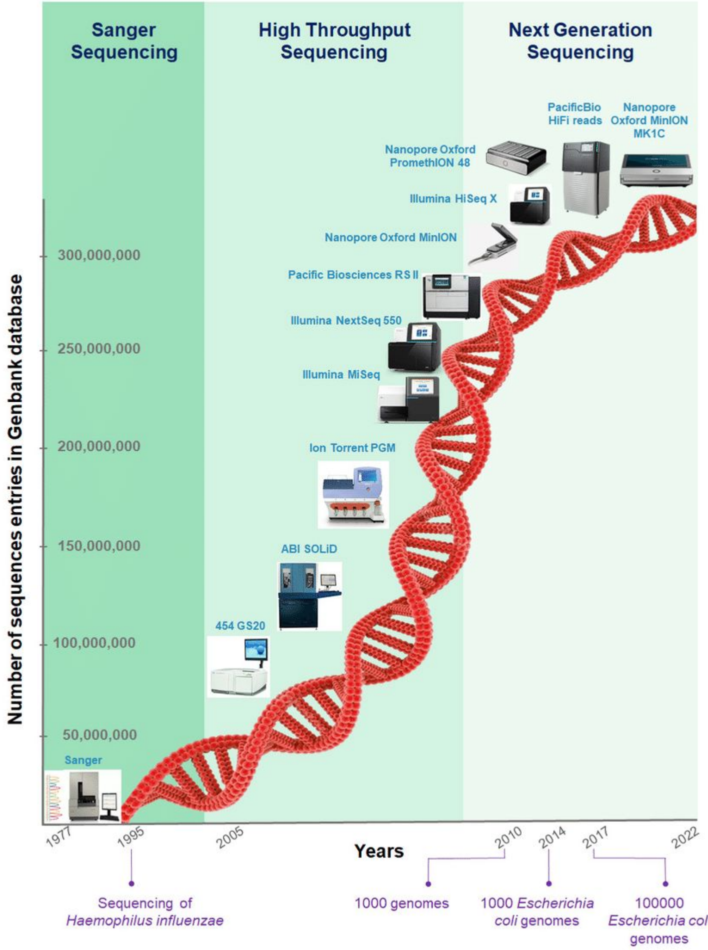
FACULTAD DE INGENIERÍA

UNIVERSIDAD DE LA REPÚBLICA URUGUAY

Genomic information keeps growing...



Cost per genome data - 2022

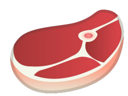
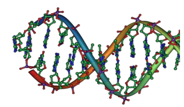


https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data

Nature vs. nurture

Genotype

Phenotype $$P = G + E$$ Environment

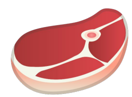
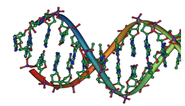
  

observable trait of an individual

Heritability $$H^2 = \frac{\mathrm{Var}(G)}{\mathrm{Var}(P)}$$

We want to find a function that links the genetic information with the phenotype

$$P = G + E$$

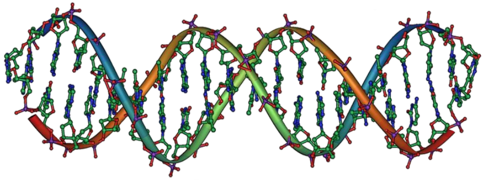
  

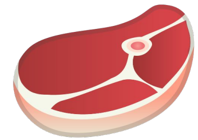
$$P = \Phi(G) + \epsilon$$

If there is good data about the environment, that links this information too

$$P = \Phi(G, E) + \epsilon$$

But we have some SNPs (for now)





$$\text{n individuals} \left\{ X = \left\{ \underbrace{\begin{pmatrix} 0 & 2 & \cdots & 2 \\ 1 & 2 & \cdots & 2 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 1 \\ 2 & 2 & \cdots & 1 \end{pmatrix}}_{\text{p SNPs}} \right. \right.$$

$$¿\ \Phi(\cdot)\ ?$$

$$\begin{pmatrix} 0.84 \\ 1.21 \\ \vdots \\ -0.34 \\ 0.1 \end{pmatrix}$$

$$Y = \Phi(X) + \epsilon$$

$$\underset{\Phi \in \mathcal{C}}{\operatorname{argmin}} \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}\left(\mathbf{y}_i, \Phi\left(\mathbf{x}_i\right)\right)$$

What wo we want to know about

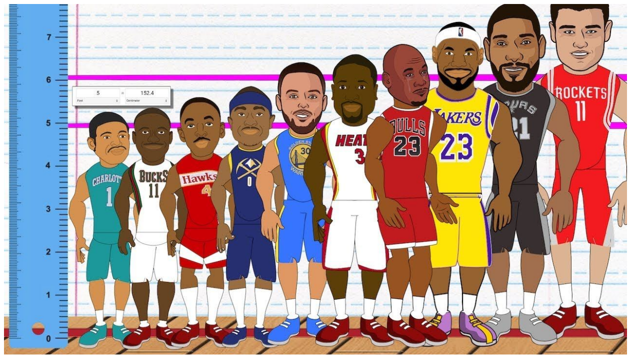
$$¿\ \Phi(\cdot)\ ?$$

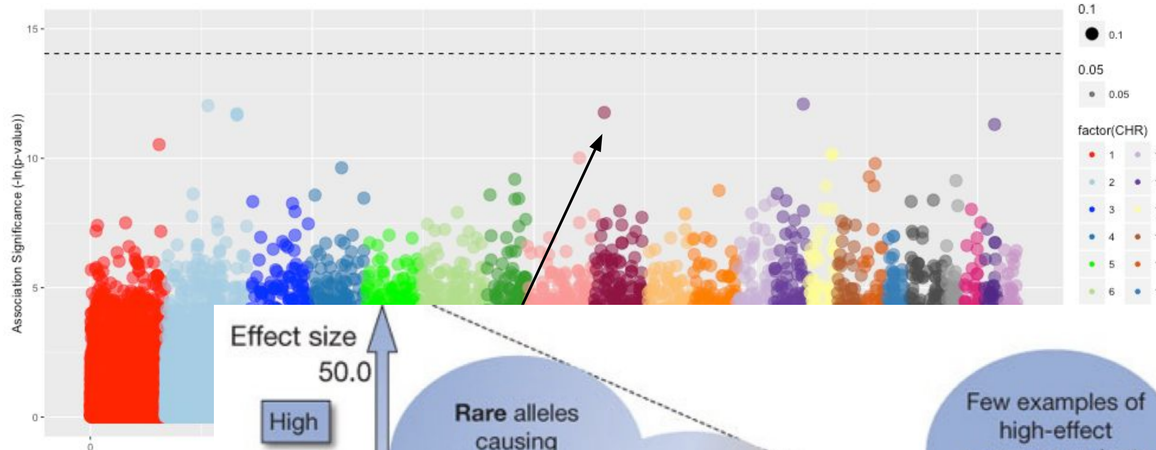
The predictions

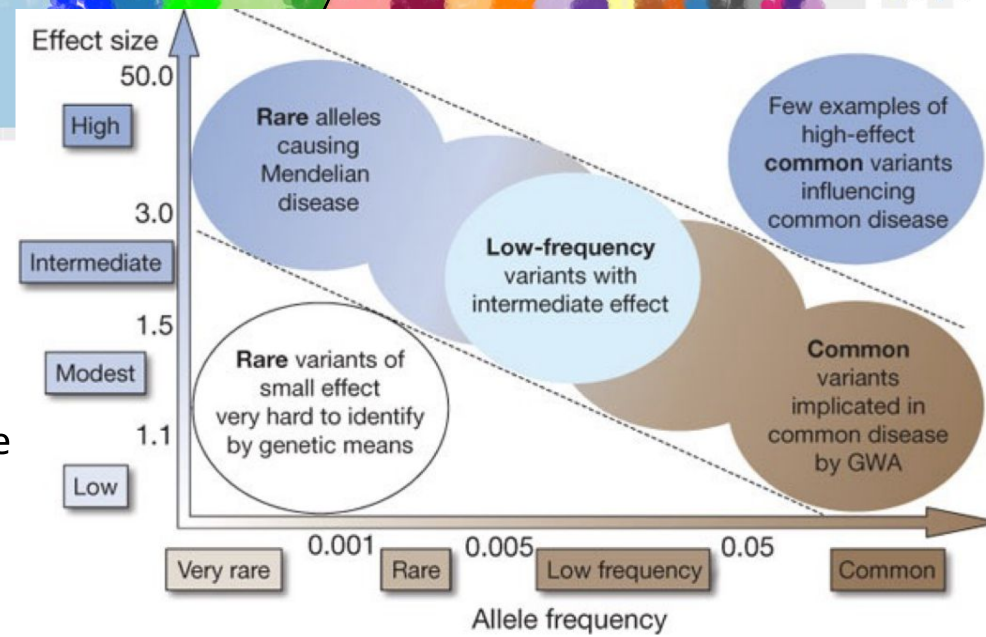
The function itself

Futures extraction

Genome Wide Association Study (GWAS)







$$y = \beta_i x_i + \epsilon,\ i \in \{1, \ldots, p\}$$

$$\begin{cases} H_0 : \beta_i = 0 \\ H_1 : \beta_i \neq 0 \end{cases}$$

=> p-value

Angelina Jolie

Based on genomic information: 87% of developing a breast cancer

Precision medicine may never be very precise - but it may be good for public health

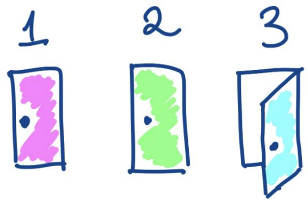
Simply Statistics | Home | Featured | About

Simply Statistics

Jan. 13, 2022
Roger Peng

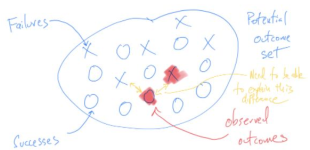
Narrative Failure in Data Analysis

A data analysis can fail if it doesn't present a coherent story and "close all the doors". Such a failure is not simply a problem with communication, but often indicates a problem with the details of the analysis itself.



Nov. 10, 2021
Roger Peng

Thinking About Failure in Data Analysis



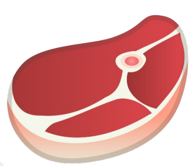
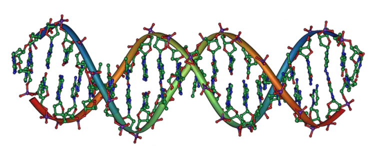
https://simplystatistics.org/

Multiple Marker Regression

$$y = \boldsymbol{X}\beta + \boldsymbol{e}$$

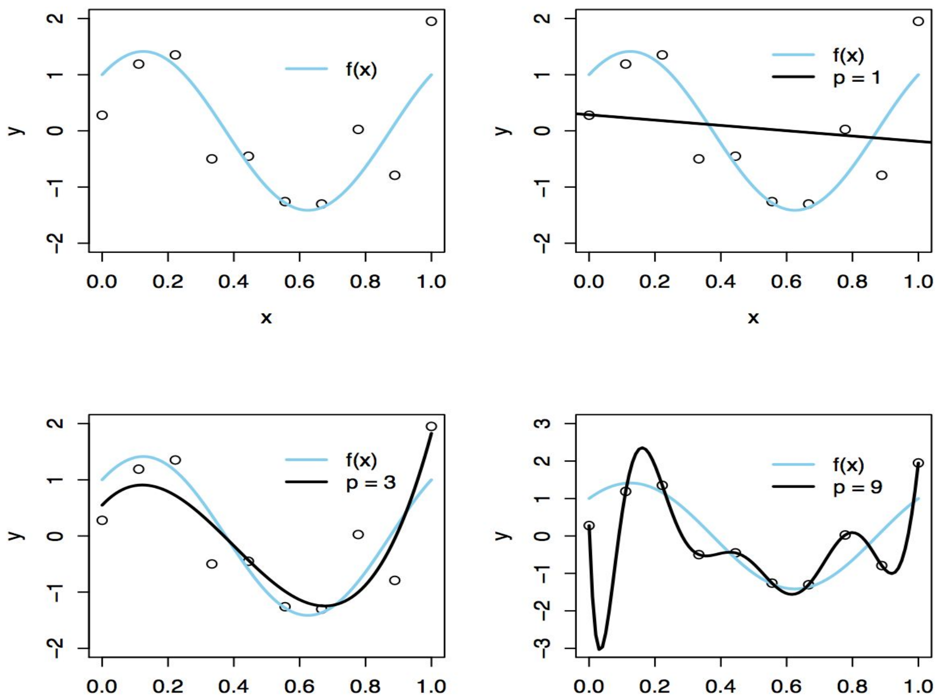
$$p \gg n$$

$$n\left\{\begin{bmatrix}\boxed{0.84}\\ 1.21\\ \vdots\\ -0.34\end{bmatrix}\right. = \overbrace{\begin{array}{c}SNP_0 \\ \left[\begin{array}{c|cccc} 2 & 1 & \dots & 0 & 2\\ 1 & 0 & \dots & 1 & 1\\ \vdots & \vdots & \ddots & \vdots & \vdots\\ 2 & 1 & \dots & 1 & 0\end{array}\right]\end{array}}^{p} \left.\begin{array}{c}\beta_0 \\ \begin{bmatrix}\boxed{0.1}\\ 0\\ \vdots\\ -0.3\end{bmatrix}\end{array}\right\}p + \begin{bmatrix}0.01\\ -0.2\\ \vdots\\ -0.04\end{bmatrix}$$

 $=$  β $+$ 

$$\hat{\beta} = argmin||y - X\beta||^2 \qquad \sim N(0, \boldsymbol{I}\sigma_{\mathrm{e}}^2)$$

Overfitting due to high number of variables



$\hat{\beta}_j$	$p=1$	$p=3$	$p=9$
$\hat{\beta}_0$	0.286	0.548	0.279
$\hat{\beta}_1$	−0.473	6.272	−237.909
$\hat{\beta}_2$	0	−30.338	5486.367
$\hat{\beta}_3$	0	25.346	−46686.042
$\hat{\beta}_4$	0	0	203251.273
$\hat{\beta}_5$	0	0	−509682.308
$\hat{\beta}_6$	0	0	765827.927
$\hat{\beta}_7$	0	0	−680299.555
$\hat{\beta}_8$	0	0	329140.427
$\hat{\beta}_9$	0	0	−66798.508
$\sum_{j=0}^{9} \hat{\beta}_j^2$	0.305	1602.479	1.465×10^{12}

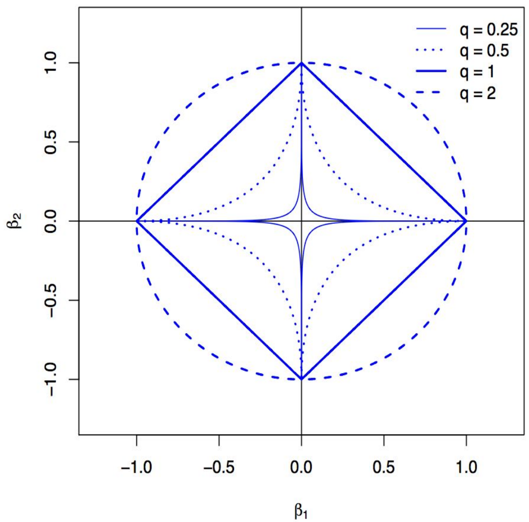
Need of penalization!!!!

Problem: Prediction of new samples will be bad!

Thanks to Sebastian Castro for the slides on multiple marker regressions.

Penalizations

$$\hat{\beta} = \arg\min \|y - X\beta\|^2 + \lambda\|\beta\|^q$$



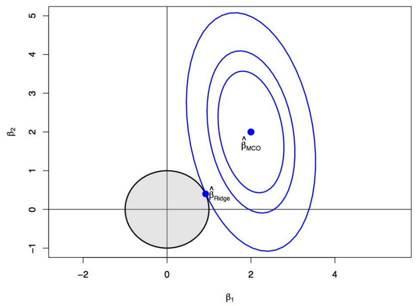
q=2: *Ridge Regression*

q=1: *Lasso*

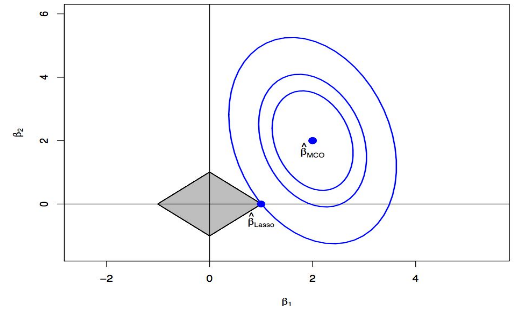
Ridge and Lasso combinations: *Elastic Net*

Shrinkage

Ridge



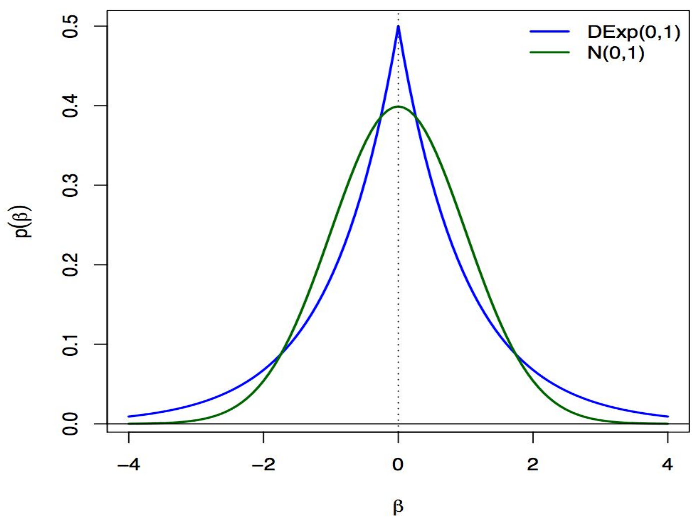
Lasso



L^q Penalizations shrink all the coefficients at the same time

Bayesian methods:

Choose the betas from a known distribution

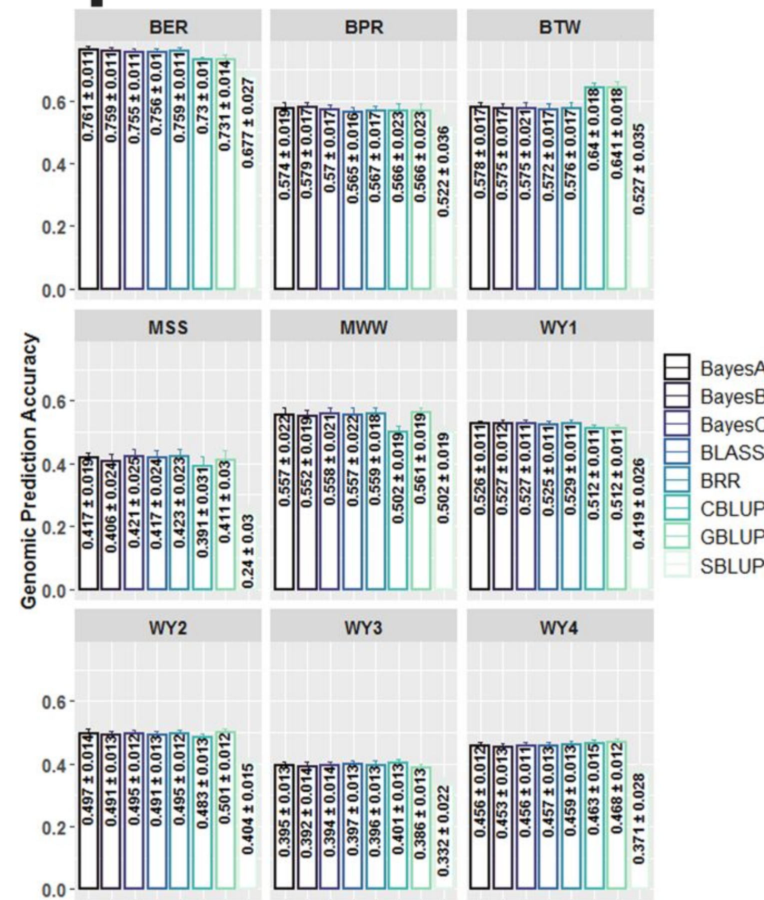


Article | Published: 04 May 2022

Performance of Bayesian and BLUP alphabets for genomic prediction: analysis, comparison and results

Prabina Kumar Meher, Sachin Rustgi & Anuj Kumar

Heredity **128**, 519–530 (2022) | Cite this article



https://www.nature.com/articles/s41437-022-00539-9

Predicting human height as the mean of the parents is more accurate than using the genomic information (in 2008)

The case of the missing heritability

- *"When scientists opened up the human genome, they expected to find the genetic components of common traits and diseases. But they were nowhere to be seen………"*

Nature news feature 6 Nov 2008

European Journal of Human Genetics

Search

Journal home > Archive > Articles > Full text

- Journal home
- Advance online publication
  - About AOP
- Current issue
- Archive
- Practical Genetics
- Gene Cards
- Focuses
- News

Article

European Journal of Human Genetics (2009) **17,** 1070–1075; doi:10.1038/ejhg.2009.5; published online 18 February 2009

Predicting human height by Victorian and genomic methods

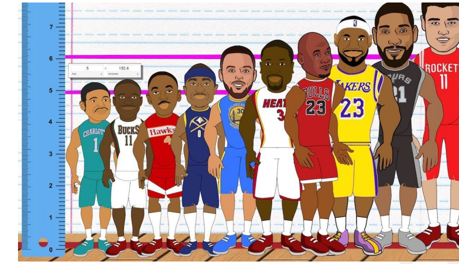
Yurii S Aulchenko[1,2,7], Maksim V Struchalin[1,3,7], Nadezhda M Belonogova[2,4], Tatiana I Axenovich[2], Michael N Weedon[5], Albert Hofman[1], Andre G Uitterlinden[6], Manfred Kayser[3], Ben A Oostra[1], Cornelia M van Duijn[1], A Cecile J W Janssens[1] and Pavel M Borodin[2,4]

Discussion - Top

In this work, we compared genomic and Victorian approaches to predict human height. In our data, the 54-loci genomic profile explained 4–6% and Victorian Galton's mid-parental values explained 40% of the height variance. Adding genomic information to the mid-parental values provided only a small (1.3%) increase in the proportion of variance explained.

by now, probably already include those with the largest effect sizes. Merely because the variants with the larger effect sizes are most easily captured, the detection of new height genes will require progressively bigger sample sizes (eg, to detect a locus explaining 0.1% of the variance at genome-wide significance $P<5 \times 10^{-8}$ with a power of 80%, one would need to study 40000 people, whereas to detect a locus explaining 0.01%, one would need 400000 people).[5]

From GWAS to genomic prediction (2011)



Discussion

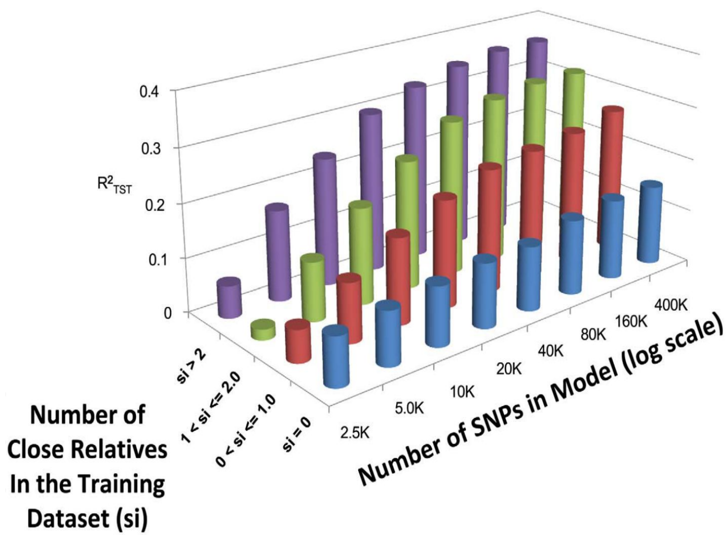
In this work, we compared genomic and Victorian approaches to predict human height. In our data, the 54-loci genomic profile explained 4–6% and Victorian Galton's mid-parental values explained 40% of the height variance. Adding genomic information to the mid-parental values provided only a small (1.3%) increase in the proportion of variance explained.

by now, probably already include those with the largest effect sizes. Merely because the variants with the larger effect sizes are most easily captured, the detection of new height genes will require progressively bigger sample sizes (eg, to detect a locus explaining 0.1% of the variance at genome-wide significance $P<5 \times 10^{-8}$ with a power of 80%, one would need to study 40000 people, whereas to detect a locus explaining 0.01%, one would need 400000 people).[5]

Beyond Missing Heritability: Prediction of Complex Traits

Robert Makowsky*, Nicholas M. Pajewski¤, Yann C. Klimentidis, Ana I. Vazquez, Christine W. Duarte, David B. Allison, Gustavo de los Campos

Department of Biostatistics, University of Alabama at Birmingham, Birmingham, Alabama, United States of America



With a bigger sample, the task becomes easier

HIGHLIGHTED ARTICLE
GENETICS | GENOMIC PREDICTION

Accurate Genomic Prediction of Human Height

2018

Louis Lello,* Steven G. Avery,* Laurent Tellier,*,†,‡ Ana I. Vazquez,§ Gustavo de los Campos,§, and Stephen D. H. Hsu*,†,1**

*Department of Physics and Astronomy, §Department of Epidemiology and Biostatistics, and **Department of Statistics and Probability, Michigan State University, East Lansing, Michigan 48824, †Cognitive Genomics Laboratory, Shenzhen Key Laboratory of Neurogenomics, China National GeneBank, BGI-Shenzhen, 518083, China, and ‡Department of Biology, Functional Genetics, University of Copenhagen, DK-2200, Denmark

ORCID ID: 0000-0001-5692-7129 (G.d.l.)

n = 488,371
p = 645,589

20.000 SNPs explain 50% of the variation

LASSO

Penalized linear regression

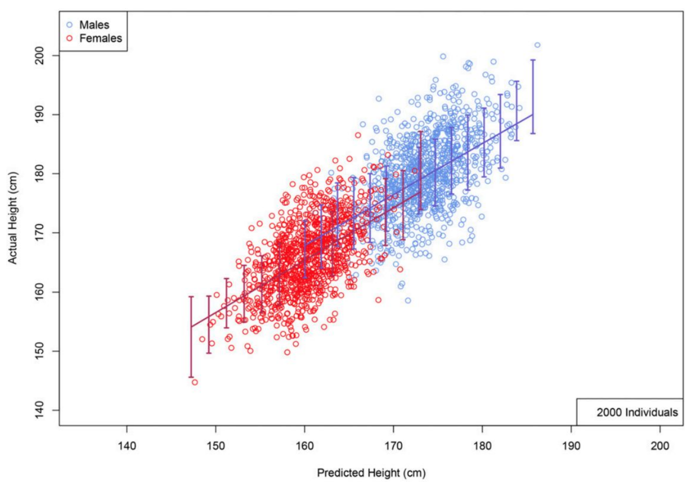
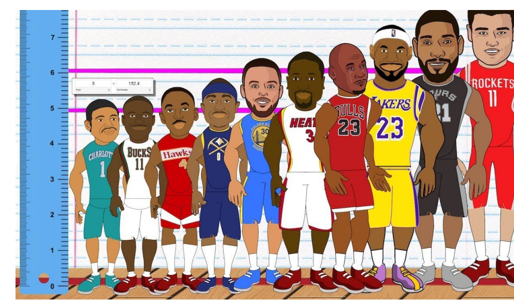


Figure A1 Actual height (centimeter) *vs.* predicted height (centimeter) using 2000 randomly selected individuals (roughly equal numbers of males and females; no corrections for age or sex) from the ARIC dataset. Error bars indicate ± 1 SD range computed using larger validation set.



The missing heritability was found… or not?

Article | Open Access | Published: 12 October 2022

A saturated map of common genetic variants associated with human height

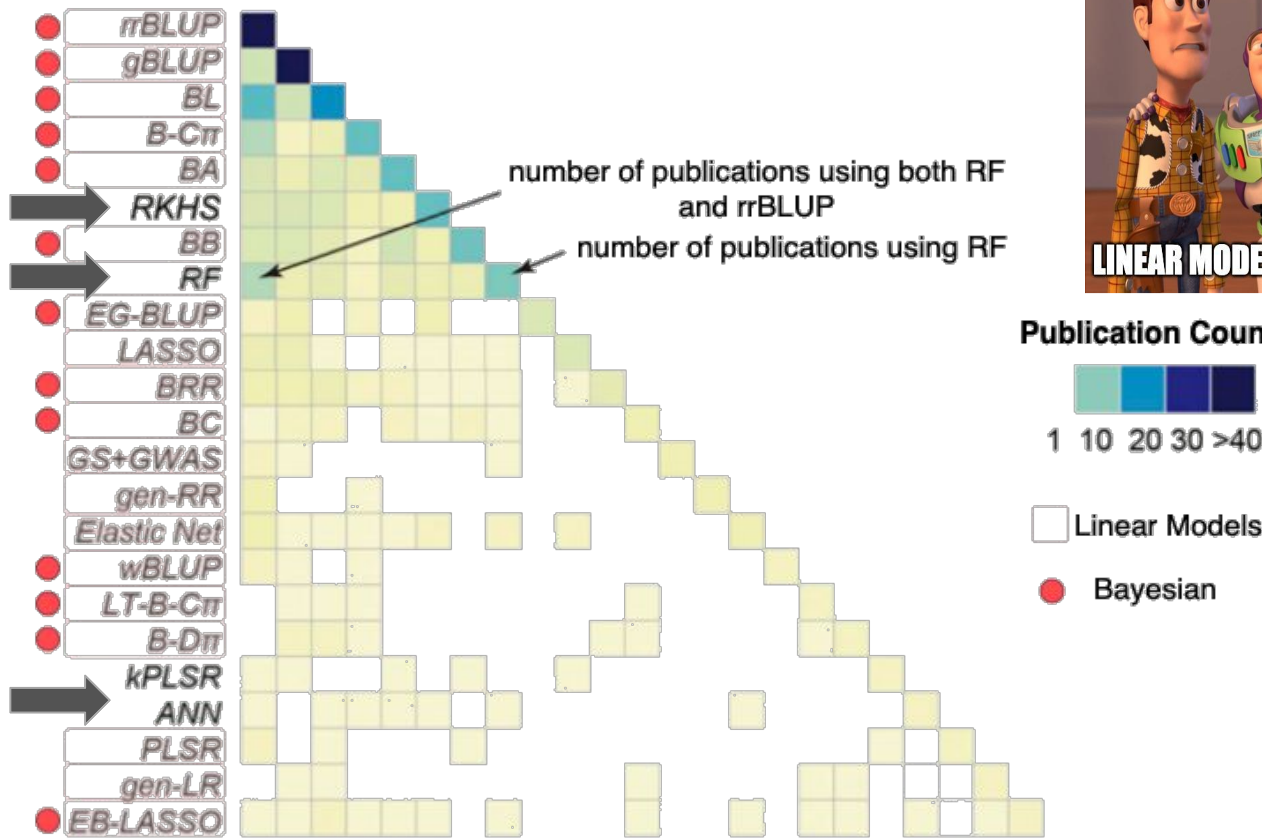
Loïc Yengo, Sailaja Vedantam, Eirini Marouli, Julia Sidorenko, Eric Bartell, Saori Sakaue, Marielisa Graff, Anders U. Eliasen, Yunxuan Jiang, Sridharan Raghavan, Jenkai Miao, Joshua D. Arias, Sarah E. Graham, Ronen E. Mukamel, Cassandra N. Spracklen, Xianyong Yin, Shyh-Huei Chen, Teresa Ferreira, Heather H.

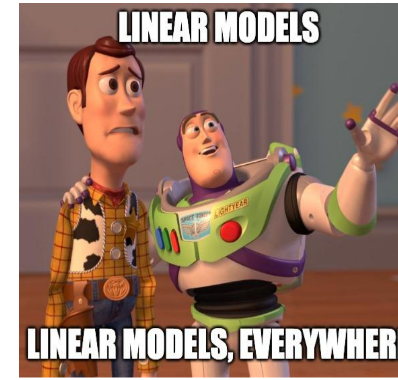
Abstract

Common single-nucleotide polym
50% of phenotypic variation in hun
associated regions requires huge s
association study of 5.4 million ind
independent SNPs that are significantly associated with height account for nearly all of the
common SNP-based heritability. These SNPs are clustered within 7,209 non-overlapping

density are enriched for biologically relevant genes. In out-of-sample estimation and prediction, the 12,111 SNPs (or all SNPs in the HapMap 3 panel[2]) account for 40% (45%) of phenotypic variance in populations of European ancestry but only around 10–20% (14–24%) in populations of other ancestries. Effect sizes, associated regions and gene prioritization

Review of the most used models for genomic prediction

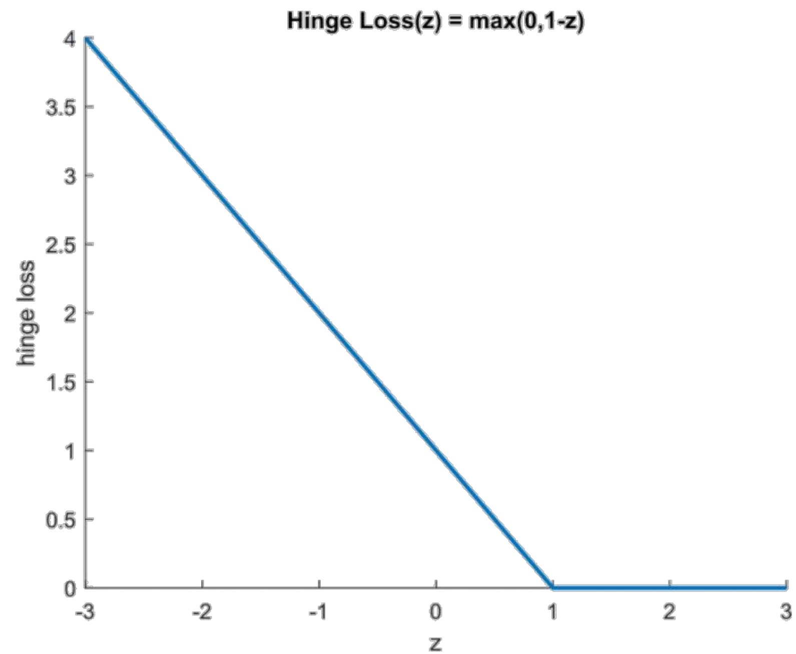




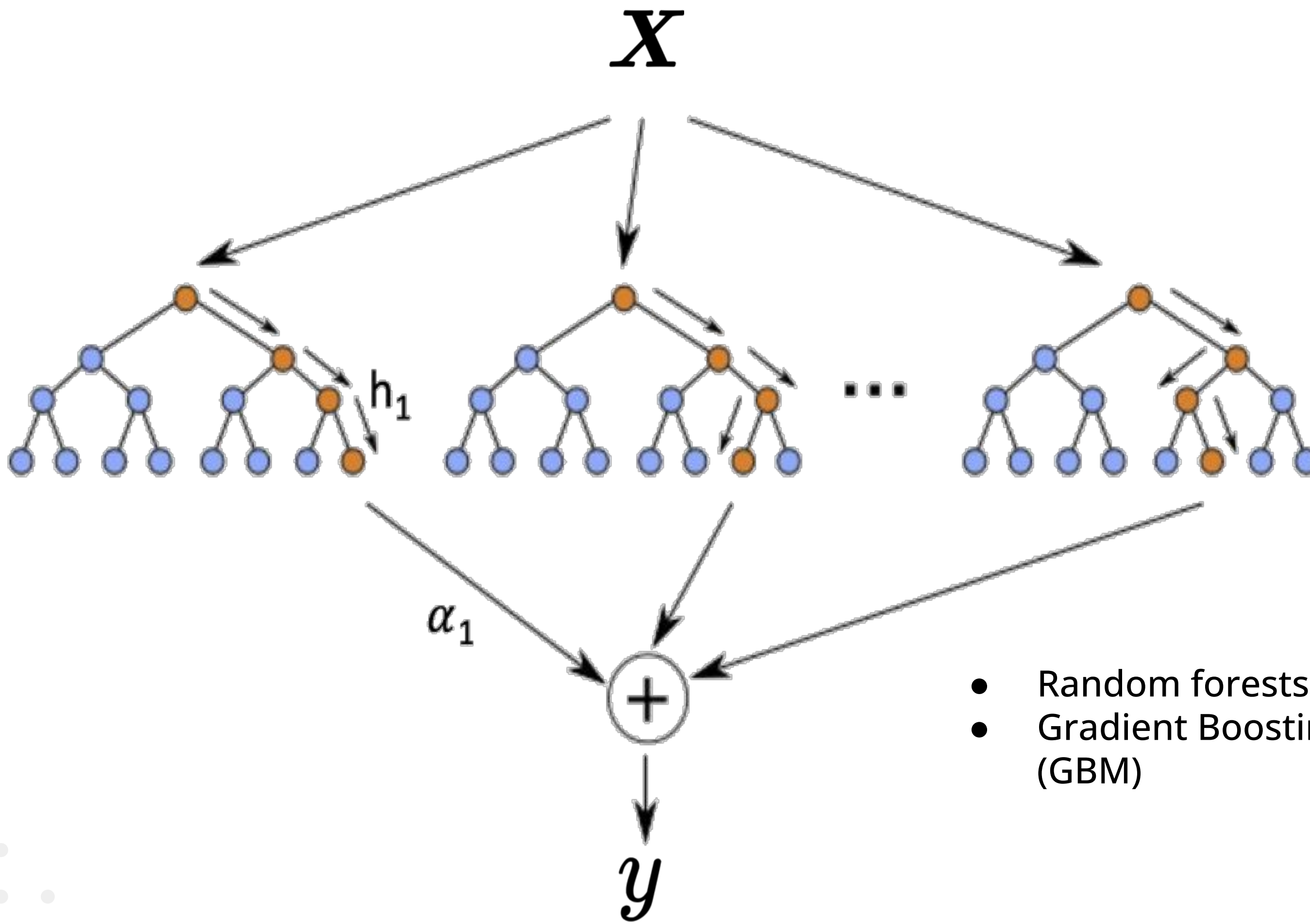
Azodi et al, *Benchmarking algorithms for genomic prediction of complex traits.* (2019)

Support Vector Regression (SVR) / RKHS

- Reproducing Kernel Hilbert Space (RKHS) is popular in genomic prediction.
- RKHS are SVR using a Hinge loss function.



Decision trees



- Random forests (RF)
- Gradient Boosting Methods (GBM)

For further reading...

AMERICAN Scientist

Genomic Prediction in the Big Data Era

BY GUSTAVO DE LOS CAMPOS, DANIEL GIANOLA

A simple model from the early 20th century remains our best tool for using DNA to predict disease risk and other complex traits.

BIOLOGY • MATHEMATICS • MEDICINE • TECHNOLOGY • GENETICS • STATISTICS

https://www.americanscientist.org/article/genomic-prediction-in-the-big-data-era