Data Science: Clinical Bionformatics and Genetic Association Studies

Alberto Brusati

Introduction:

Clinical Bioinformatics

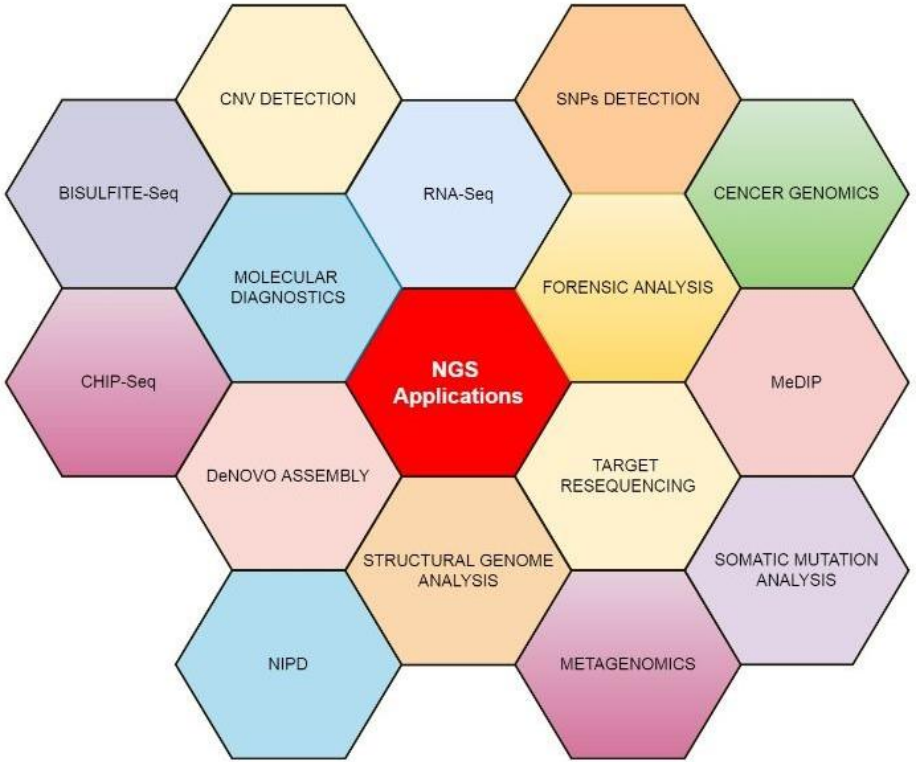
"Clinical bioinformatics is an area of healthcare science responsible for developing and improving methods for acquiring, storing, organising and analysing biological data that supports the delivery of patient care."

"Staff working in clinical bioinformatics use areas of computer science including software tools that generate useful biological knowledge by manipulating 'big data' "

Informatics in health care science is broken down into different areas:

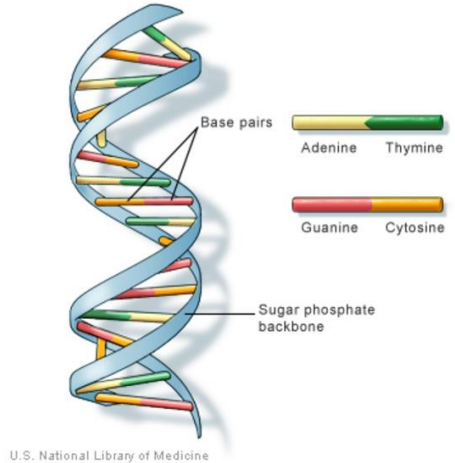
- clinical bioinformatics (genomics)
- clinical bioinformatics (health informatics)

NGS Applications



Biological background

genotype



phenotype

A mendelian trait is one that is controlled by a single locus, while complex phenotypes may be influenced by multiple variants in the genome. **AIM: analyze genetic variability!**

Type of genetic variability

Single base-pair substitution

These are also known as single nucleotide polymorphisms (SNPs) and can be any nucleic acid substitution:

1. Transition
    - interchange of the purine (Adenine/Guanine)
    - or pyrimidine (Cytosine/Thymine) nucleic acids
2. Transversion
    - interchange of a purine and pyrimidine nucleic acid

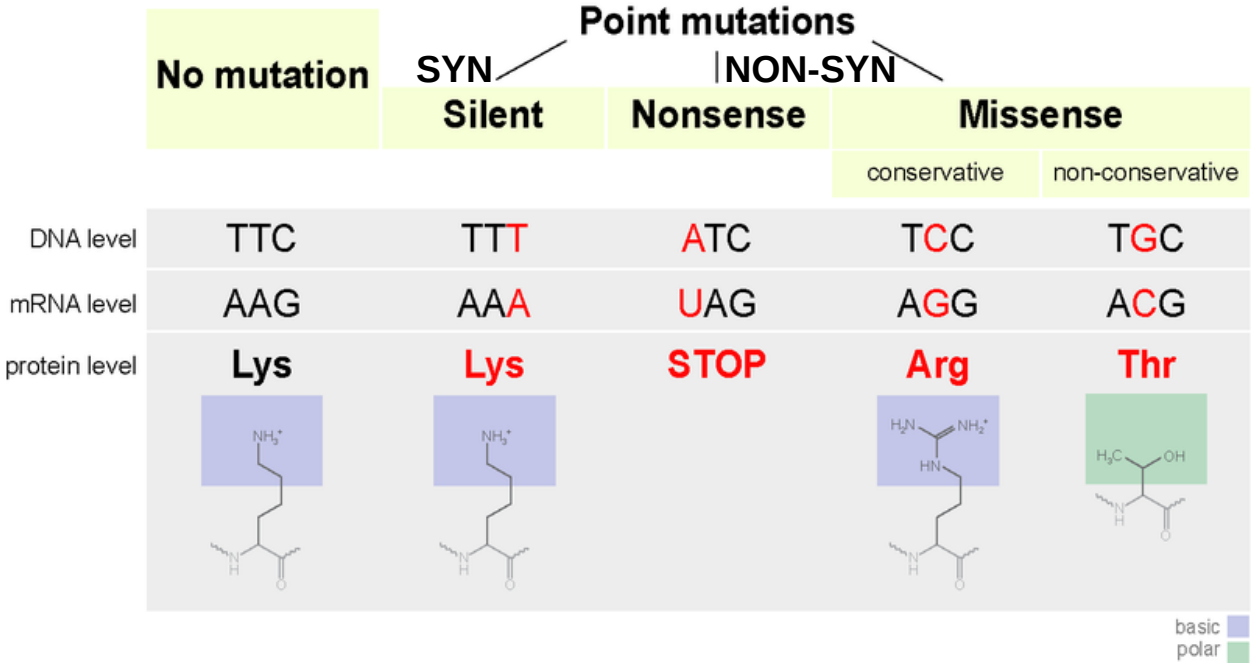
Insertion or deletion, also known as ‘indel’

Insertion or deletion of a single stretch of DNA sequence that can range from two to hundreds of base-pairs in length.

Structural variation

Typically used to describe genetic variation that occurs over a larger DNA sequence (example **CNV!!!!!!!!!!!!**).

Effects of genetic variability



Indels with a length divisible by three (i.e. whole codon indels) in coding regions will cause insertions or deletions of whole amino acids into the protein, and are known as **in-frame deletions or insertions.**

If the length is not divisible by three, this will cause a f**rameshift** where all codons downstream of the indel are shifted.

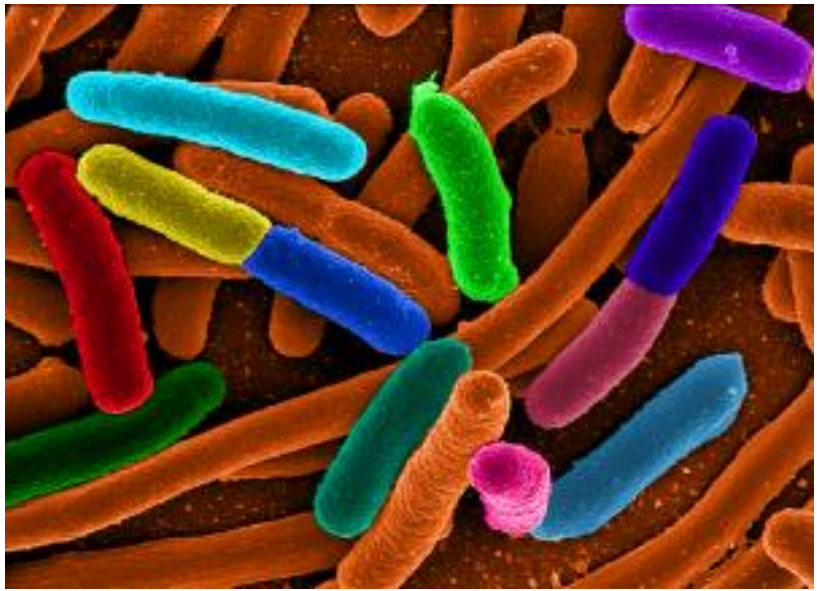
Effects of genetic variability: protein level

The effects of variants on protein structure can vary dramatically depending on the type of protein and the extent of variation.

core level: change of conformation and unfolding

surface level: protein-protein, or protein-nucleic acid interactions, any variation of amino acids on the binding surface could lead to a loss of function.

Variations in prokaryotes: a key of evolution

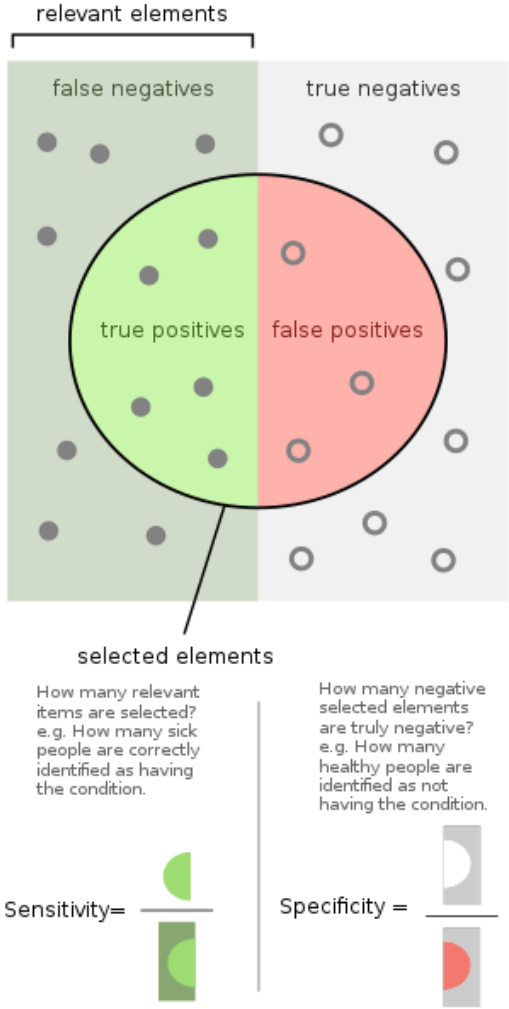


1. Prokaryotes tend to have haploid genomes, meaning that they only have one copy of each gene per individual.
2. Prokaryotes reproduce clonally (i.e. by making an exact copy of the cell), so recombination does not occur as a matter of course at every generation.
3. Prokaryotes typically reproduce much more rapidly than eukaryotes, and these short generation times can lead to more rapid adaptation.

Sensitivity and specificity

Sensitivity (true positive rate) is the probability of a positive test result, conditioned on the individual truly being positive.
HIGH SENSITIVITY = high probability that a sick subject will test positive; = low probability that a sick subject will test negative.
Higher sensitivity means low risk of false negative!

Specificity (true negative rate) is the probability of a negative test result, conditioned on the individual truly being negative.
HIGH SPECIFICITY = high probability that a healthy subject will test negative; = low probability that a healthy subject will test positive
Higher specificity means low risk of false positive!



Files And Formats

Introduction:

FastA format

It’s a standard format to indicate the DNA or protein sequence born in 1985. The term stands for "FAST-All" it’s Generally indicated with .fasta .fa. The file contains sequences with a header starting with '**>**' Each row consists of 60-70 or 80 bases.

```
>chromosome:GRCh38:15:48407706:48646449:-1
TTGGACACCGACTTAGAGAGAGCACCCTGGATTCAGGCCTGGGTGGCTTGAATTGAGCCTCAGAAGAGCCGCGTCTGG
AGTGGGCTCTCGACACCCAGGGCAAGTGGGGGCGGCAGAGCCCTCTCCTCGGTCGGCACAGCAGCCTCTGCCGCGGTC
CCGGCCTGCGACGCGCCCAGTCTTAGCCTCCCGGCCTCCGGCGTCTGCTGAGTGTCCGGCGGGAGAGGCGCAGGGAGC
GCGCTACCGGGAGGCGCGGGCAGCGGGGACTGGTTTTCTCTCGGGCCAGGGCCTCCGGGGCAACCGTCTCCAGCGCG
CATTCTTGGTGCAGGT
```

```
>MCHU - Calmodulin - Human, rabbit, bovine, rat, and chicken
ADQLTEEQIAEFKEAFSLFDKDGDGTITTKELGTVMRSLGQNPTEAELQDMINEVDADGNGTID
FPEFLTMMARKMKDTDSEEEIREAFRVFDKDGNGYISAAELRHVMTNLGEKLTDEEVDEMIREA
DIDGDGQVNYEEFVQMMTAK*
```

FastQ format

FASTQ format is a text-based format for storing both a biological sequence (usually nucleotide sequence) and its corresponding quality scores. Both the sequence letter and quality score are each encoded with a single ASCII character for brevity.

The quality score is indicated as **PHRED Score** $Q = -10 \log_{10} P$

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10000	99.99%
50	1 in 100000	99.999%

Fast
Q

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

FASTQ

```
[sbsuser@compute-00-00 RAW]$ cat 341_CGATGT_L001_R1_001.fastq |head
```

Read 1

```
@B0P8DQ1:71:HJM7YBCXX:1:1101:1074:2057 1:N:0:CGATGT
NCTGAGATGAGTCAGAGGAGAGCTAGAGTTGGGAACNGATCACCAGTGGCC
+
#.<<.A.A<GA.<.GGGGAGGG.GGGAGAGAAAGGA#<<GG.G<GGAG.GA
```

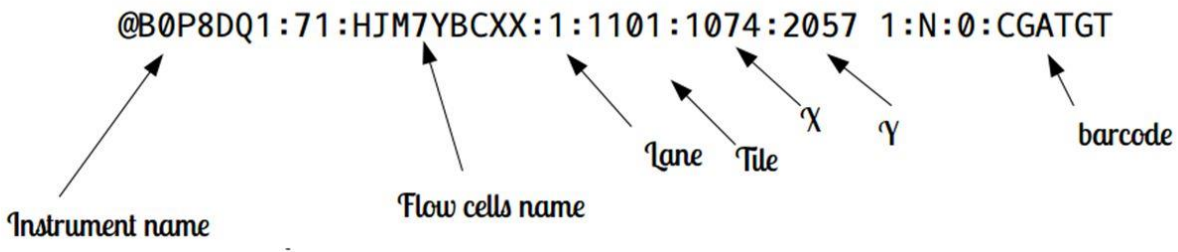
Read2

```
@B0P8DQ1:71:HJM7YBCXX:1:1101:1219:2071 1:N:0:CGATGT
NGTTGCTAAAGTAGGTAGAATGCAAACCTGAAGCTATTAGGAACTATATCT
+
#<<GGGGIIGGGIIIGGGIIGGGIIIIIGIIGGGIIIGGGGIIIIGGIIII
```

Read n

```
@B0P8DQ1:71:HJM7YBCXX:1:1101:1108:2122 1:N:0:CGATGT
CTATAGTCCTTAGCAAGACTTCTGAGTAAAATTAACTTTAATTCTTTAAAA
```

FASTQ



Element	Requirements	Description
@	@	Each sequence identifier line starts with @
<instrument>	Characters allowed: a-z, A-Z, 0-9 and underscore	Instrument ID
<run number>	Numerical	Run number on instrument
<flowcell ID>	Characters allowed: a-z, A-Z, 0-9	
<lane>	Numerical	Lane number
<tile>	Numerical	Tile number
<x_pos>	Numerical	X coordinate of cluster
<y_pos>	Numerical	Y coordinate of cluster
<read>	Numerical	Read number. 1 can be single read or read 2 of paired-end
<is filtered>	Y or N	Y if the read is filtered, N otherwise
<control number>	Numerical	0 when none of the control bits are on, otherwise it is an even number. See below.
<index sequence>	ACTG	Index sequence

.fastq

.bam /
.sam /
.cram

.vcf

BROAD INSTITUTE

VCF format

VCF is a text file format. The name stands for Variant Call Format and it contains meta-information lines, a header line, and then data lines each containing information about a position in the genome. The format also has the ability to contain genotype information on samples for each position.

```
##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
```

	Name	Brief description (see the specification for details).
1	CHROM	The name of the sequence (typically a chromosome) on which the variation is being called. This sequence is usually known as 'the reference sequence', i.e. the sequence against which the given sample varies.
2	POS	The 1-based position of the variation on the given sequence.
3	ID	The identifier of the variation, e.g. a dbSNP rs identifier, or if unknown a ".". Multiple identifiers should be separated by semi-colons without white-space.
4	REF	The reference base (or bases in the case of an indel) at the given position on the given reference sequence.
5	ALT	The list of alternative alleles at this position.
6	QUAL	A quality score associated with the inference of the given alleles.
7	FILTER	A flag indicating which of a given set of filters the variation has failed or PASS if all the filters were passed successfully.
8	INFO	An extensible list of key-value pairs (fields) describing the variation. See below for some common fields. Multiple fields are separated by semicolons with optional values in the format: `<key>=<data>[,data]` .
9	FORMAT	An (optional) extensible list of fields for describing the samples. See below for some common fields.
+	SAMPLEs	For each (optional) sample described in the file, values are given for the fields listed in FORMAT

VCF FILE

##source=GATK 1.6									
#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	MISEQ_12_02_2014_A04
chr2	1,9E+08	.	T	C	32.86	LowVariar	BaseQRan	GT:AD:DP	0/1:49,7:69:62.86:63,0,1164:40:33:0.125
chr2	1,9E+08	.	G	C	24.93	PASS	BaseQRan	GT:AD:DP	0/1:20,7:30:54.91:55,0,394:40:25:0.259
chr2	1,9E+08	rs3106796	A	G	169.21	PASS	BaseQRan	GT:AD:DP	0/1:13,9:22:99:199,0,467:40:99:0.409
chr2	1,9E+08	.	T	A	109.78	PASS	BaseQRan	GT:AD:DP	0/1:35,9:57:99:140,0,1045:40:99:0.205
chr2	1,9E+08	.	G	A	27.47	PASS	BaseQRan	GT:AD:DP	0/1:3,2:5:57.46:57,0,96:40:27:0.400
chr2	1,9E+08	.	T	C	11.01	LowGQX;L	DP=5;Dels	GT:AD:DP	1/1:0,2:2:3.01:41,3,0:40:3:1.000
chr2	1,9E+08		T	A	23.86	LowVariar	BaseQRan	GT:AD:DP	0/1:15,3:19:53.84:54,0,412:40:24:0.167

WHAT IT MAY BE IMPORTANT TO KNOW ABOUT A GENETIC VARIANT

- ITS GENOMIC POSITION
- THE GENE INVOLVED
- IF IT IS EXONIC OR NOT
- HOW IT IS FREQUENT IN THE POPULATION
- WHAT KIND OF AMINO ACID ALTERATION IT PRODUCES
- IF IT IS CONSERVED
- KNOW HOW MUCH IT CAN ALTER THE PROTEIN
- IF IT IS A KNOWN PATHOGENIC MUTATION

There are tools that allow you to intersect all this information and to merge the VCF files reporting the variants we have found and the data available about them.
THIS OPERATION IS CALLED

ANNOTATION

GWAS DIS	GWAS OR	GWAS BETA	GWAS PUBMED	GWAS SNP	GWAS P	SIFT score	SIFT pred	Polyphen2 HDIV score	Polyphen2 HDIV pred	Polyphen2 HVAR score
.	.	.	.	.	.	0.01	D	0.069	B	0.19
.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.

Polyphen2 HVAR pred	LRT score	LRT pred	MutationTaster score	MutationTaster pred	MutationAssessor score
B	0.000	D	1.000	D	1.66
.	.	.	.	.	.
.	.	.	.	.	.

MutationAssessor pred	FATHMM score	FATHMM pred	RadialSVM score	RadialSVM pred	LR score	LR pred	VEST3 score	CADD raw
L	-4.16	D	0.720	D	0.852	D	0.951	3.855
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.

GERP++ RS	phyloP46way placental	phyloP100way vertebrate	SiPhy 29way logOdds	Otherinfo
5.44	2.183	7.390	14.614	het
.	.	.	.	hom
.	.	.	.	het

gnomAD exome ALL	gnomAD exome AFR	gnomAD exome AMR	gnomAD exome ASJ	gnomAD exome EAS	gnomAD exome FIN	gnomAD exome NFE	gnomAD exome OTH	gnomAD exome SAS	gnomAD genome ALL	gnomAD genome AFR	gnomAD genome AMR	gnomAD genome ASJ	gnomAD genome EAS	gnomAD genome FIN	gnomAD genome NFE	gnomAD genome OTH
.	.	.	.	.	.	.	.	.	3.431e-05	0	0	0	0	0	7.182e-05	0

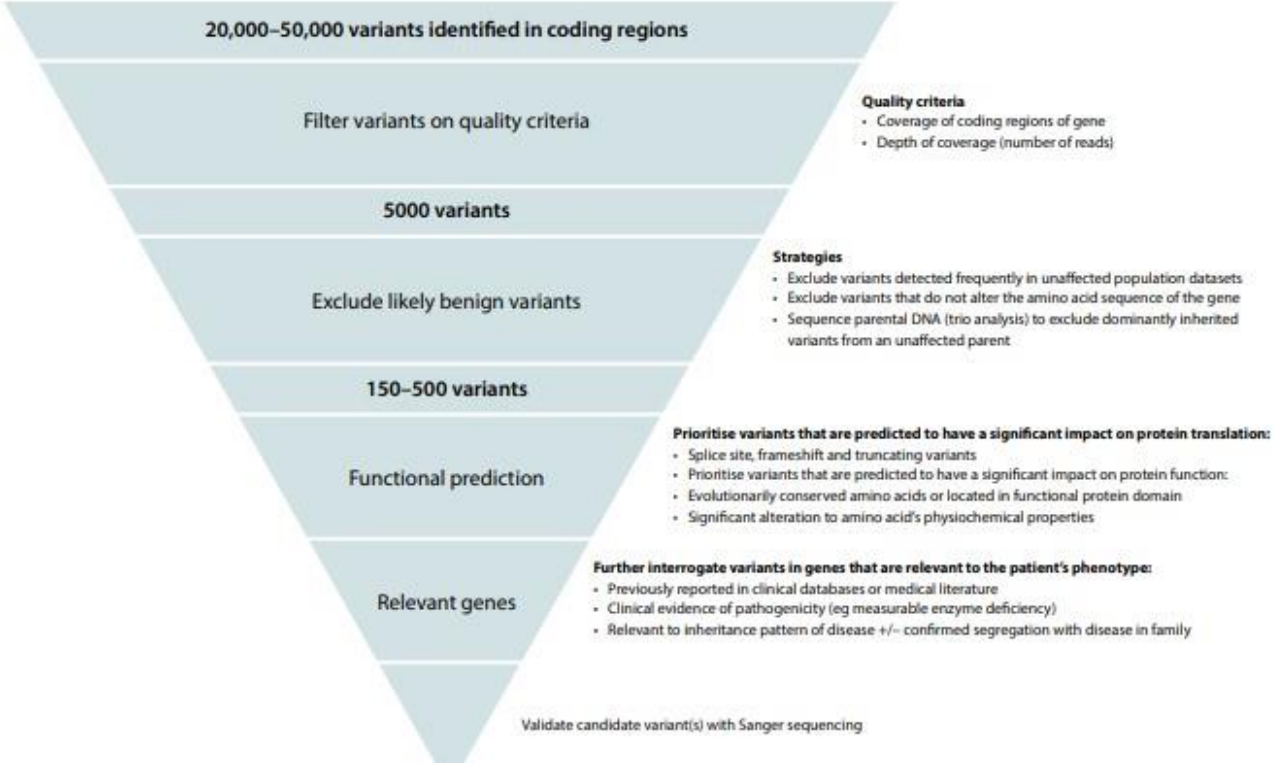
Filtering Genetic Variants

- Consider the prevalence of the disease (different from incidence).
- Consider the model of inheritance (Dominant or recessive)
- Make a calculation on the potential Heterozygotes/Homozygotes present in the population.
- Identify a threshold for the MAF
- Consider the penetrance of the disease. And adjust the threshold for the MAF

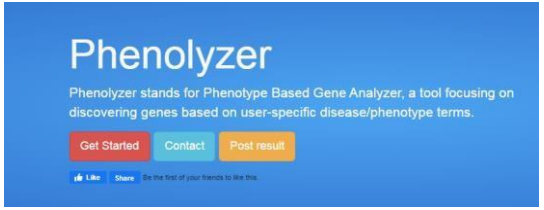
Filtering Genetic Variants

- Select only exonic and Splicing variants
- Based on ExonicFunc.refGene field we exclude all variants that are synonimous
- Exlude all variants reported in 1000G database and dbSNP
- Exclude all variants reported as benign in ClinVar_SIG
- Exlude all variants reported in gnomAD with a MAF
- Prediction tools (pay attention)

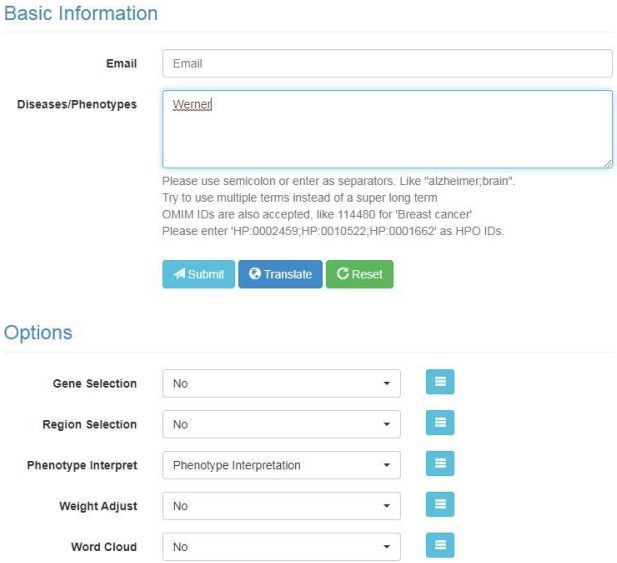
Filtering Genetic Variants

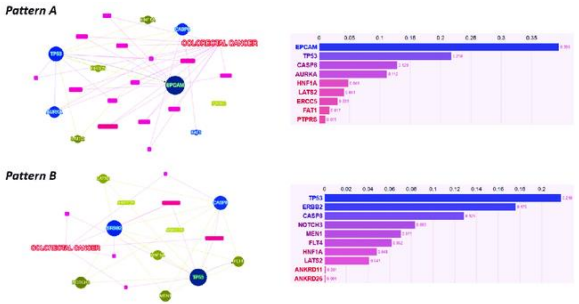


Filtering Genetic Variants



Prioritization of Genes





Filtering Genetic Variants

	Benign		Pathogenic			
	Strong	Supporting	Supporting	Moderate	Strong	Very strong
Population data	MAF is too high for disorder BA1/BS1 **OR** observation in controls inconsistent with disease penetrance BS2			Absent in population databases PM2	Prevalence in affecteds statistically increased over controls PS4	
Computational and predictive data		Multiple lines of computational evidence suggest no impact on gene /gene product BP4 Missense in gene where only truncating cause disease BP1 Silent variant with non predicted splice impact BP7 In-frame indels in repeat w/out known function BP3	Multiple lines of computational evidence support a deleterious effect on the gene /gene product PP3	Novel missense change at an amino acid residue where a different pathogenic missense change has been seen before PM5 Protein length changing variant PM4	Same amino acid change as an established pathogenic variant PS1	Predicted null variant in a gene where LOF is a known mechanism of disease PVS1
Functional data	Well-established functional studies show no deleterious effect BS3		Missense in gene with low rate of benign missense variants and path. missenses common PP2	Mutational hot spot or well-studied functional domain without benign variation PM1	Well-established functional studies show a deleterious effect PS3	
Segregation data	Nonsegregation with disease BS4		Cosegregation with disease in multiple affected family members PP1	Increased segregation data →		
De novo data				De novo (without paternity & maternity confirmed) PM6	De novo (paternity and maternity confirmed) PS2	
Allelic data		Observed in *trans* with a dominant variant BP2 Observed in *cis* with a pathogenic variant BP2		For recessive disorders, detected in trans with a pathogenic variant PM3		
Other database		Reputable source w/out shared data = benign BP6	Reputable source = pathogenic PP5			
Other data		Found in case with an alternate cause BP5	Patient's phenotype or FH highly specific for gene PP4			

Filtering Genetic Variants

Benign:

The variant is not considered to be the cause of the tested disease

Main evaluation criteria

- It is evident that the variant does not segregate with the disease in families with 2 or more affected individuals

Or at least 4 points among these options

- Control population minor allele frequency (eg, gnomAD) is considerable (MAF > 0.001) (prevalence of the disease must be taken into account)
- Homozygous variant in a gene with no association to the disease.
- Intact protein function observed in appropriate functional assay(s), eg,. a splice-region variant without abnormal splicing
- Co-occurrence with a pathogenic mutation in the same gene (phase unknown) or in another gene that clearly explains the proband's phenotype
- No disease association in small case-control study
- Majority of the in silico tools predict the substitution to be benign.

Filtering Genetic Variants

Likely Benign:

The variant is not considered to be the cause of the tested disease

Main evaluation criteria

- Control population minor allele frequency (eg,gnomAD) is considerable (MAF > 0.001) (disease prevalence must be taken into account)

or at least 2 points

- MAF < 0.001 in control populations but the variant is detected in healthy controls with no disease association in a case-control study/studies
- Homozygous variant in a gene with no association to the disease.
- Intact protein function observed in appropriate functional assay(s), eg,. a splice-region variant without abnormal splicing
- Co-occurrence with a pathogenic mutation in the same gene (phase unknown) or in another gene that clearly explains the proband's phenotype
- Majority of the in silico tools predict the substitution to be benign.

Filtering Genetic Variants

Variant of uncertain significance (VUS):

The variant has characteristics of being an independent disease-causing mutation, but insufficient or conflicting evidence exists

Main evaluation criteria

- The variant is typically very rare, predicted to be deleterious, and the gene has an association with the patient's phenotype.

Filtering Genetic Variants

Likely pathogenic variant

The identified variant is considered the probable cause of the patient's disease.

Main evaluation criteria

- A clear genotype-phenotype correlation exists. In these cases, it is essential to have thorough background information from the referring clinician about the patient's phenotype, which helps to determine the probable pathogenicity.

Filtering Genetic Variants

Likely pathogenic variant

The identified variant is considered the probable cause of the patient's disease.

Additional criteria are: at least 2 points

- Alterations resulting in premature truncation (eg, frameshift, nonsense, or consensus splice site (+/-1, 2) in a gene where loss of gene function has been established as a mechanism of pathogenicity for the patient's disease.
- Variant is novel or very rare in control populations (cannot be applied for ethnic backgrounds absent from control populations).
- Clear genotype-phenotype correlation exists (eg, MfS and FBN1)
- Missense variant predicted deleterious by a majority of in silico tools applied
- A variant predicted to have an effect to the splicing by majority of in silico tools applied
- Variant has been identified in ≥2 individuals (one of which can be the current patient) with the same disease manifestation
- Deficient protein function in appropriate functional assay(s)
- Well-characterized mutation at the same codon or same splice consensus site

Filtering Genetic Variants

Pathogenic variant

The variant is considered the cause of the patient's disease.

Main evaluation criteria

- The variant is well established as disease causing in the databases and literature, and a wide consensus on the variant pathogenicity exists. In these cases, significant family segregation has been verified and several publications support pathogenicity

or at least 5 points

- Variant is novel or very rare in control populations
- Loss of gene function has been established as a mechanism of pathogenicity
- A missense variant predicted deleterious by a majority of in silico tools applied
- De novo alteration in the setting of a novel disease in the family
- Variants considered deleterious in consensus splice site
- Deficient protein function in appropriate functional assay(s)
- Well-characterized other mutation at the same codon

Filtering Genetic Variants

Evidence Based Categorization of Somatic mutation

Unlike the interpretation of germline mutations which focuses on their pathogenicity in a specific clinical setting, in the case of somatic mutations, the interpretation should be focused on their impact in clinical practice. A somatic variant can in fact be considered a biomarker

- If it predicts sensitivity or resistance to therapy
- If it impairs the function of a gene target of a therapy
- If it serves as an inclusion criterion for a trial
- If it affects the prognosis
- If it helps diagnose cancer
- If it helps detect the presence of the tumor or residual disease

Filtering Genetic Variants

Rare Variants and Problems with medical reporting

- **NO RELEVANT GENE VARIANT FOR THE PHENOTYPE**
- **ONE OR MORE KNOWN AND PATHOGENIC VARIANTS IN GENES COMPATIBLE WITH CLINICAL INDICATIONS**

- **ONE OR MORE VARIANTS ONLY POTENTIALLY PATHOGENIC IN GENES COMPATIBLE WITH CLINICAL INDICATIONS**

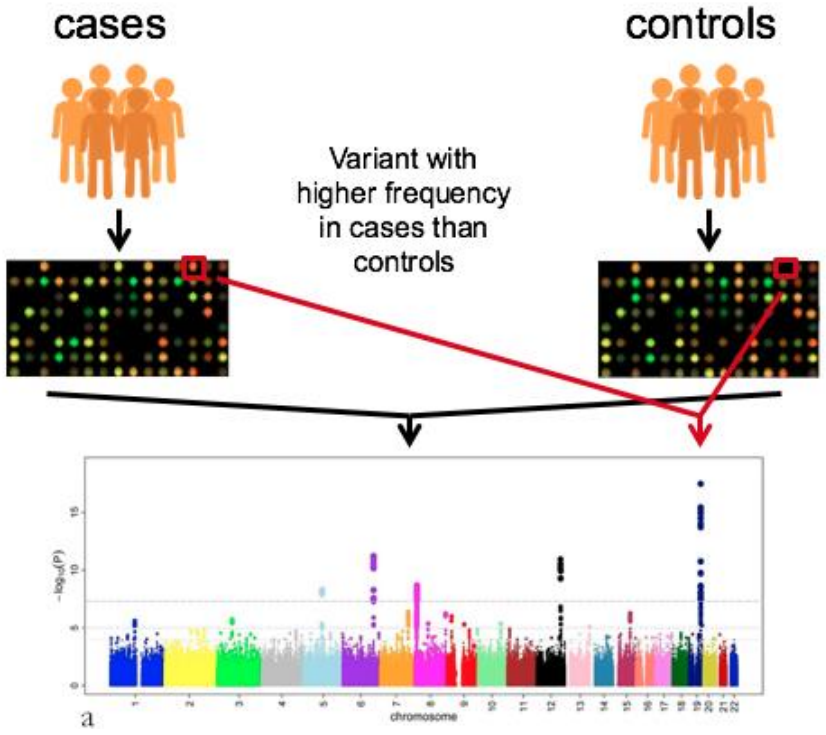
- **ONE OR MORE KNOWN AND PATHOGENIC VARIATIONS IN GENES NOT COMPATIBLE WITH CLINICAL INDICATIONS**

- **ONE OR MORE VARIATIONS ONLY POTENTIALLY PATHOGENIC IN GENES NOT COMPATIBLE WITH CLINICAL INDICATIONS**

Genome-Wide Association Studies

Many methods for associating variants with a phenotype, trait or disease rely on the fact that a variant leading to a phenotype is found at a higher frequency in cases (individuals with the phenotype) than controls (individuals without the phenotype).

Genome wide association studies (GWAS) involve genotyping individuals at common variants across the genome using genome wide SNP arrays. Variants associated with trait, or within the same haplotype as a variant associated with a trait, will be found at a higher frequency in cases than controls.

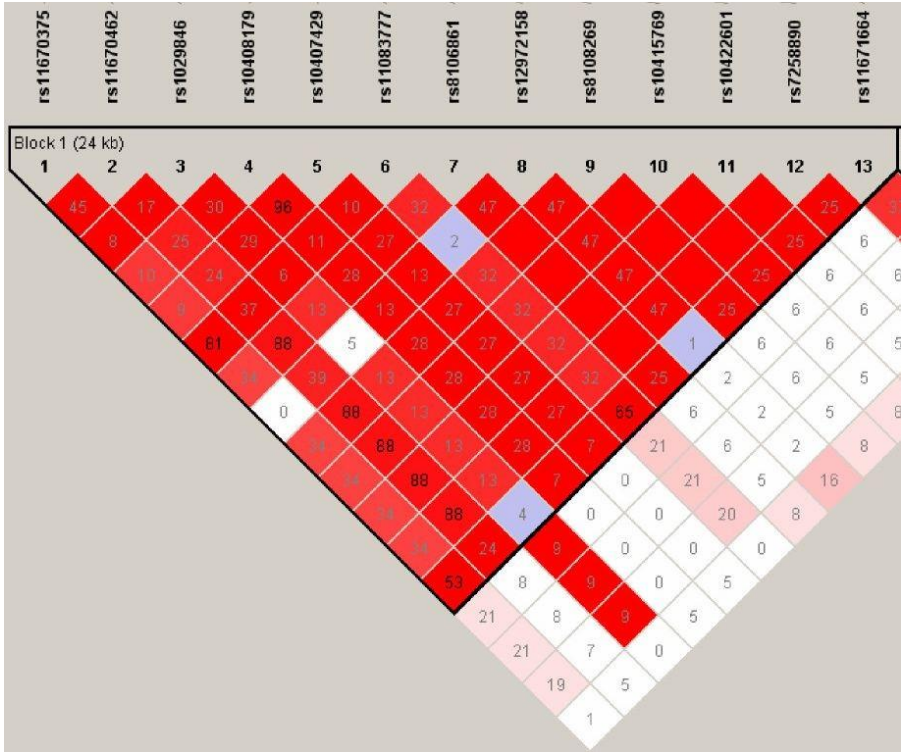


Which SNPs may i genotype?

It’s not necessary to genotype all SNPs in a genome. Using linkage disequilibrium we could collect all the information.

With specific tag-SNPs it’ possible to collect the information of 2M of SNPs.

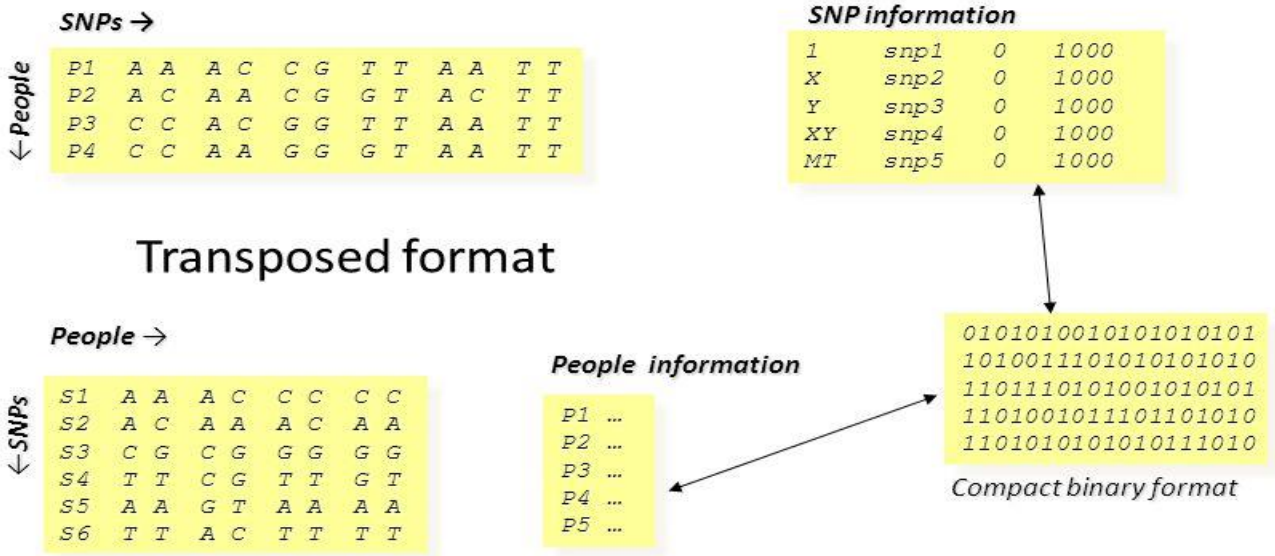
A correct density could cover the entire genome. (first array 100k, average array 450k/600k probes)



plink... Data Format

Whole genome association analysis toolset

PED and MAP format

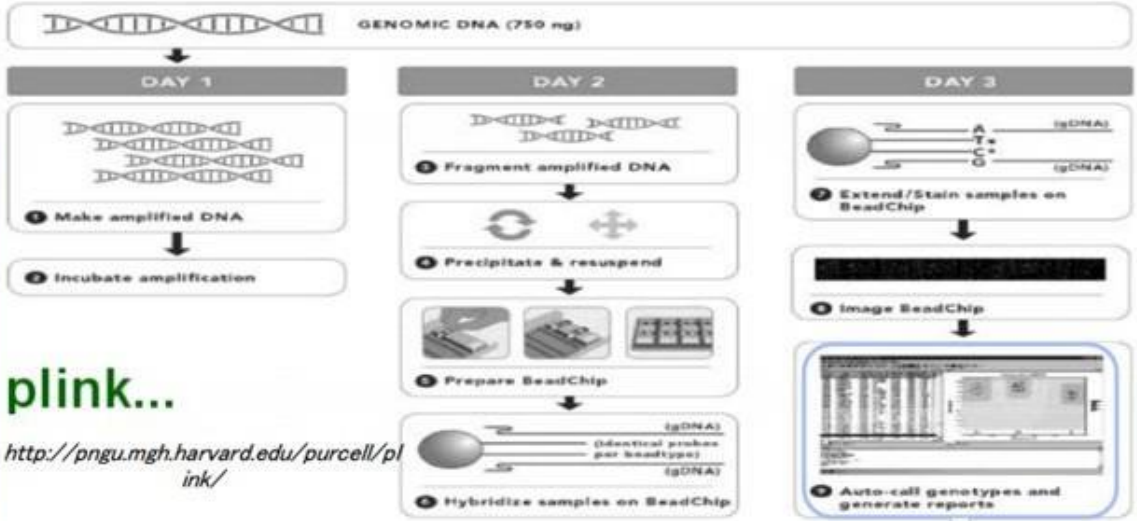


PED file:

- *Col 1:* Family ID
- *Col 2:* Individual ID
- *Col 3:* Paternal ID
- *Col 4:* Maternal ID
- *Col 5:* Gender
- *Col 6:* Phenotype
- *Col 7…n:* genotypes

MAP file:

- *Col 1:* Chromosome
- *Col 2:* rs# or snp identifier
- *Col 3:* Genetic distance (morgans)
- *Col 4:* Base-pair position (bp units)



plink...

http://pngu.mgh.harvard.edu/purcell/plink/

Elaborazione

ipert10001c	ipert10001c	0	0	2	1	C	C	G	G	A	A
ipert10002c	ipert10002c	0	0	2	1	T	C	G	G	A	A
ipert10003c	ipert10003c	0	0	1	1	T	T	G	G	A	A
ipert10004c	ipert10004c	0	0	1	1	T	C	G	G	A	C
ipert10006c	ipert10006c	0	0	2	1	T	C	G	G	C	A
ipert10007c	ipert10007c	0	0	2	1	T	C	G	G	C	A
ipert10008c	ipert10008c	0	0	1	1	C	C	G	G	A	A
ipert10009c	ipert10009c	0	0	2	1	C	0	G	G	A	C
ipert10010c	ipert10010c	0	0	2	1	0	0	G	G	A	C
ipert10012c	ipert10012c	0	0	2	1	C	C	G	G	A	C
ipert10013c	ipert10013c	0	0	2	1	C	C	G	G	A	C
ipert10014c	ipert10014c	0	0	2	1	T	C	G	G	A	A
ipert10016c	ipert10016c	0	0	1	1	T	C	G	G	A	A
ipert10017c	ipert10017c	0	0	1	1	T	C	G	G	A	C
ipert10018c	ipert10018c	0	0	1	1	T	T	G	G	A	A

How to analyze: quality control

#conversion to bed format (**NB not a genomic bed!!!!**)

plink --ped file.ped --map file.map --make-bed --out file

#sex mistake check:

plink --bfile file --check-sex --out file

#phenotype sex is compared to genotype sex, samples not corresponding will then removed.

How to analyze: quality control

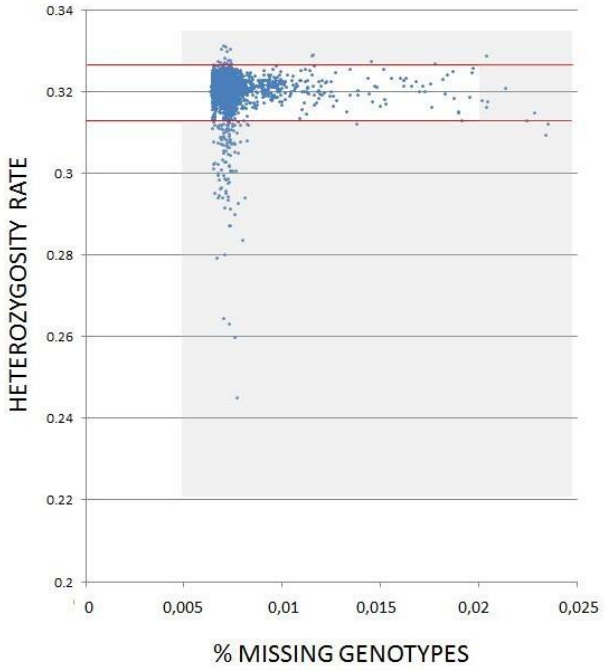
#remove SNPs missing or bad genotyped:

plink --bfile file -–missing --out file

#2 files produced: lmiss and imiss

#Het check (CI calcolation):

plink --bfile file --het --out file



How to analyze: quality control

#IBD check:

IBD probabilities are calculated:

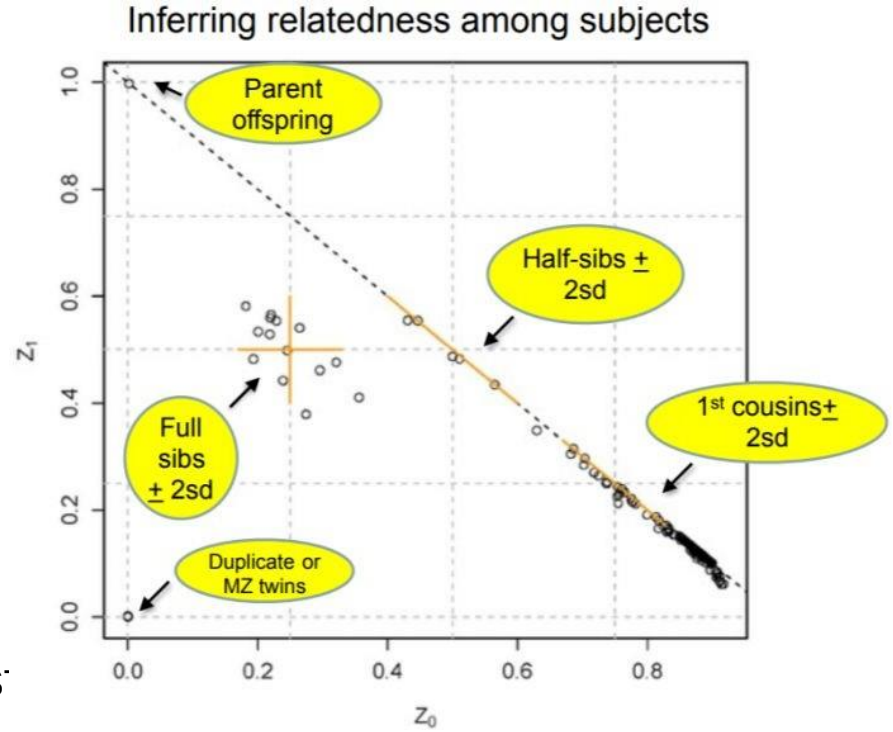
Z0 = P(share 0 IBD alleles)

Z1 = P(share 1 IBD allele)

Z2 = P(share 2 IBD alleles)

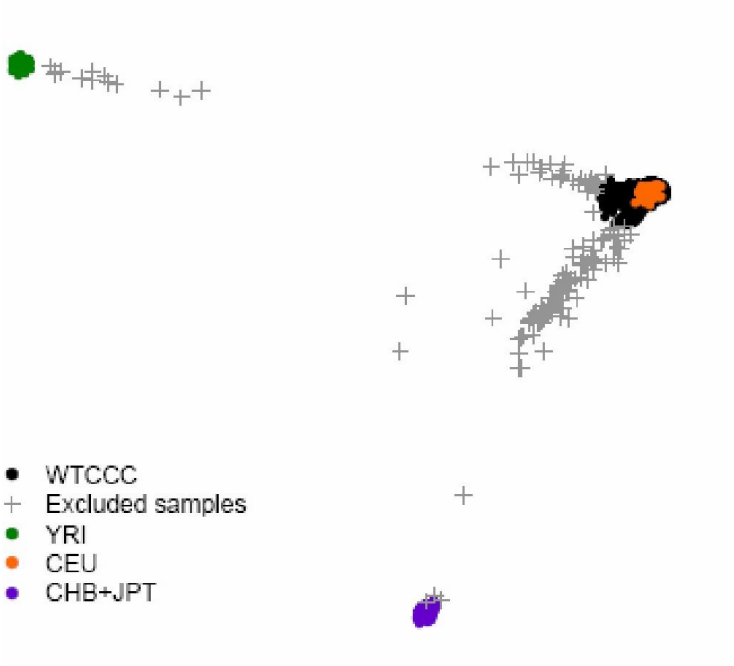
Plotting Z0 vs. Z1 we can identify related individuals

REMOVE INDIVIDUALS THAT FAILS ALL THESE TES



How to analyze: quality control

#BACKGROUND NOISE: POPULATION STRATIFICATION!



PCA of UK population:
remove all samples not from a specific population (EIGENSTRAT pipeline)

Ready to test the association:

```
#Plink2 has retired --assoc command, the new pipeline:

plink2 --pfile mydata \
      --glm genotypic interaction \
      --covar tmp.cov #age and sex, usually


There are additional parameters that take into accounts the additive effect or dominance deviation!
```

EXPECTED OUTPUT:

- CHR: the chromosome number of the SNP
- SNP: the identifier of the SNP
- BP: the physical position of the SNP on the genome map
- A1: the reference allele (in apostrophe) for the SNP
- A2: the alternate allele (in apostrophe) for the SNP
- NMISS: the number of subjects who do not have a missing value for the SNP
- BETA: the estimated coefficient for the SNP (on a logarithmic scale)
- SE: the standard error for the estimated coefficient
- L95: the lower limit of the 95% confidence interval for the estimated coefficient
- U95: the upper limit of the 95% confidence interval for the estimated coefficient
- STAT: the value of the Wald test for the SNP
- P: the p-value associated with the Wald test for the SNP. A significant p-value indicates that there is an effect of the SNP on the phenotype! But not the only one!!!!

TIPS: PLINK2.0 allows recoding VCF

Basic syntax:

plink --vcf my.vcf --make-bed --out out.vcf

….ERROR???? What can we do?

Large NGS file often contains extra contig, in this case use “--allow-extra-chr” parameter

https://www.cog-genomics.org/plink/2.0/

Sequencing and Introduction to Hail: Overview

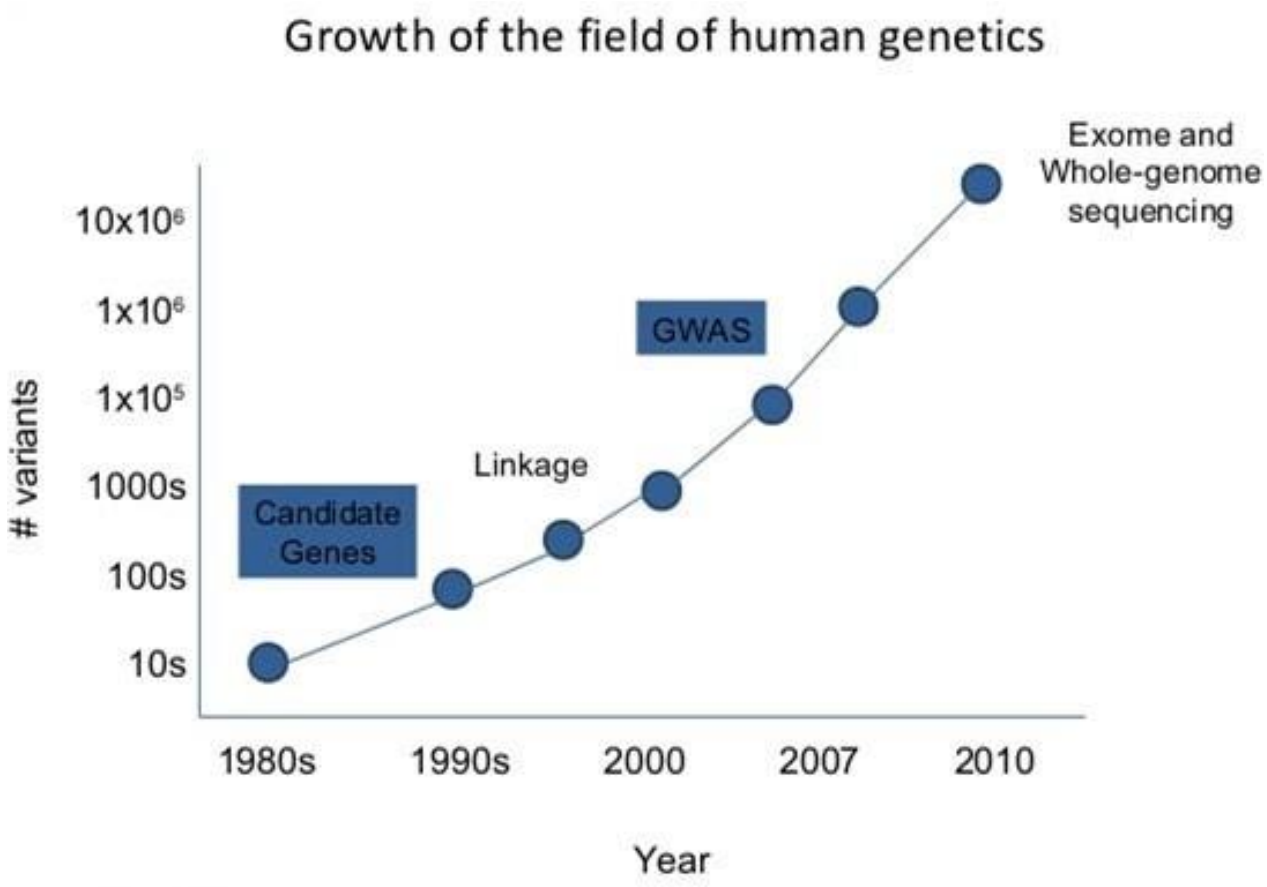
Credits to the valuable work of:

Kumar Veerapen, PhD

Learning Objectives

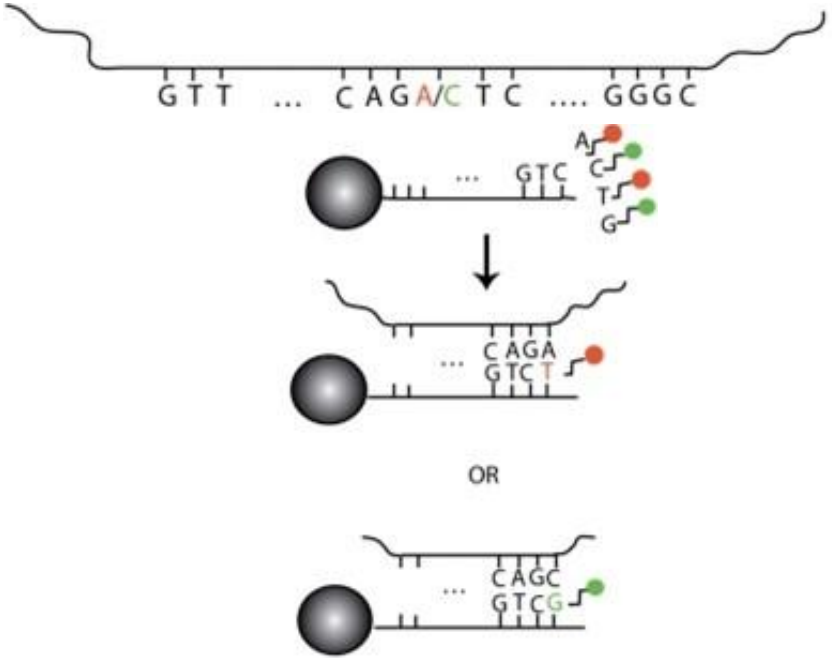
- To understand the overview of DNA sequencing methods
- To capture the need for Hail in the analysis of genomic datasets
- To be able to use basic Hail functions
- To apply basic GWAS analysis techniques using Hail on their own datasets
- To describe the use of PCA in Hail to decipher ancestries
- To obtain resources to further explore the extent of Hail capabilities
- To learn how to use Hail on public compute clouds

BROAD INSTITUTE



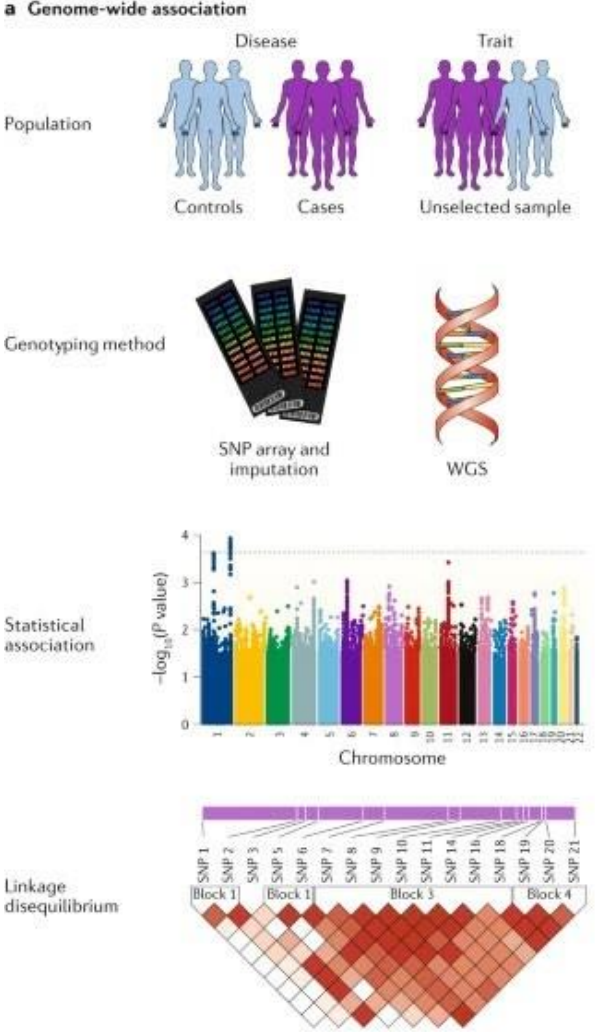
Adapted from: Lekki Wood. Baylor College of Medicine. Slideshare.net

Genotyping chip

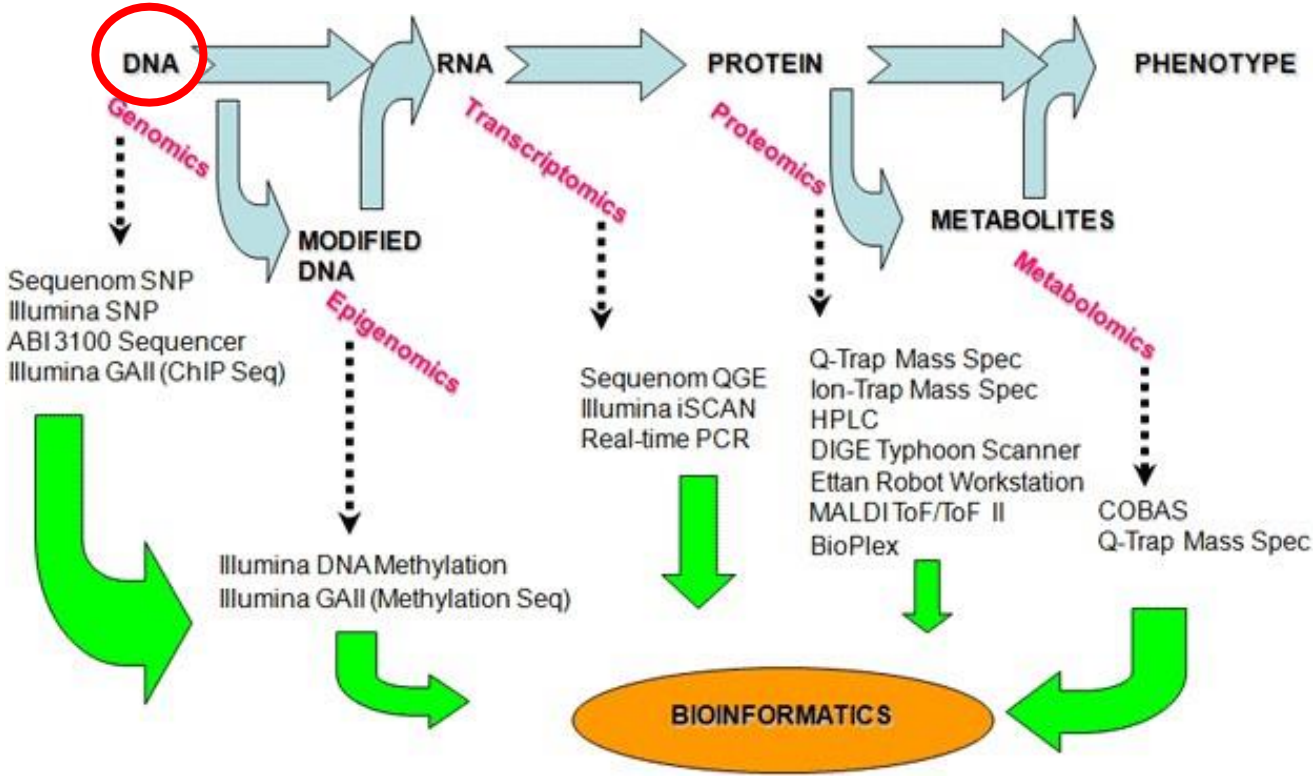


Single nucleotide polymorphism arrays:
a decade of biological, computational and technological advances

Hailiang Huang, MGH



Tam et al (2019). *Nat Gen Rev*

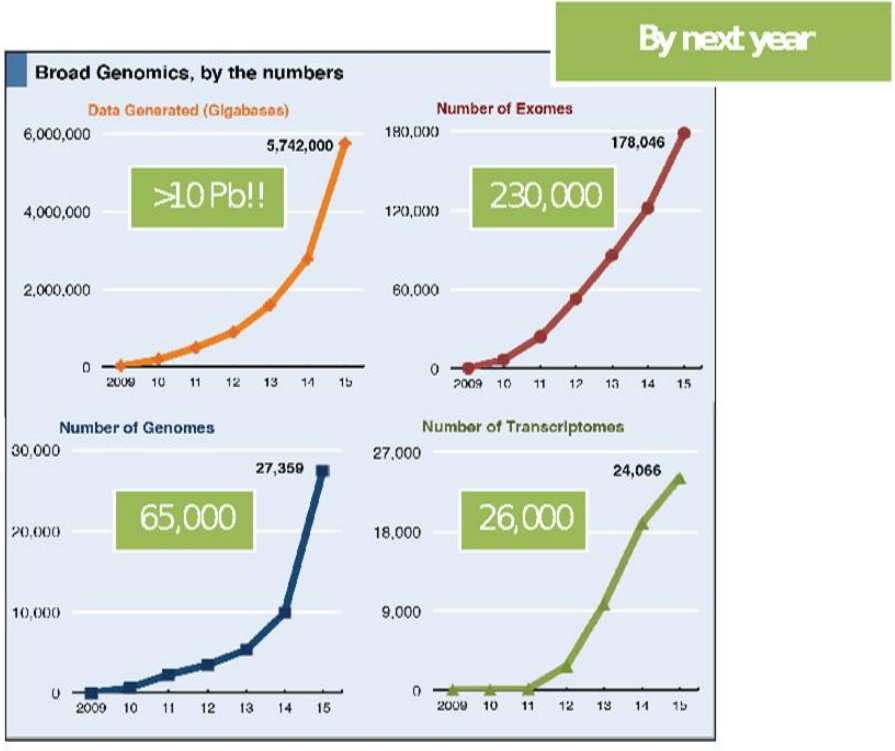

DNA
RNA
PROTEIN
PHENOTYPE
Genomics
Transcriptomics
Proteomics
MODIFIED DNA
METABOLITES
Epigenomics
Metabolomics
Sequenom SNP
Illumina SNP
ABI 3100 Sequencer
Illumina GAII (ChIP Seq)
Sequenom QGE
Illumina iSCAN
Real-time PCR
Q-Trap Mass Spec
Ion-Trap Mass Spec
HPLC
DIGE Typhoon Scanner
Ettan Robot Workstation
MALDI ToF/ToF II
BioPlex
COBAS
Q-Trap Mass Spec
Illumina DNA Methylation
Illumina GAII (Methylation Seq)
BIOINFORMATICS

Learning Objectives

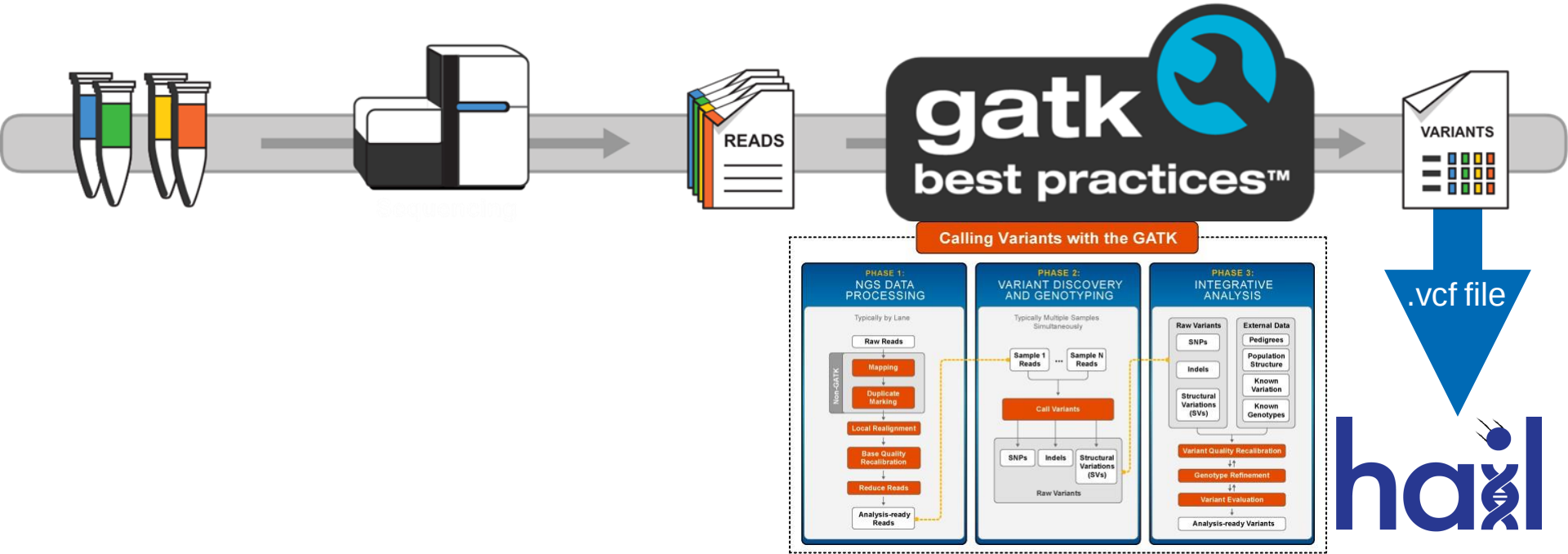
- To capture the need for Hail in the analysis of genomic datasets
    - Why Hail?
    - What can you use Hail for?

BROAD INSTITUTE

Accelerating Genomic Data e.g. Call Sets, variant files etc



What is Hail's role in callset generation?



What is Hail?

"On a scale from zero to dplyr, the Hail 0.2 interface scores an 8/10 for general-purpose data analysis." - Konrad K., lead analyst, gnomAD

Learn more at Hail.is

*We can't read your minds, so talk to us
discuss.hail.is

Open-Source Data Science Library

Slice, dice, query, and model any kind of data

Scalability

Easy to use with both small and **biobank-scale** genomic data

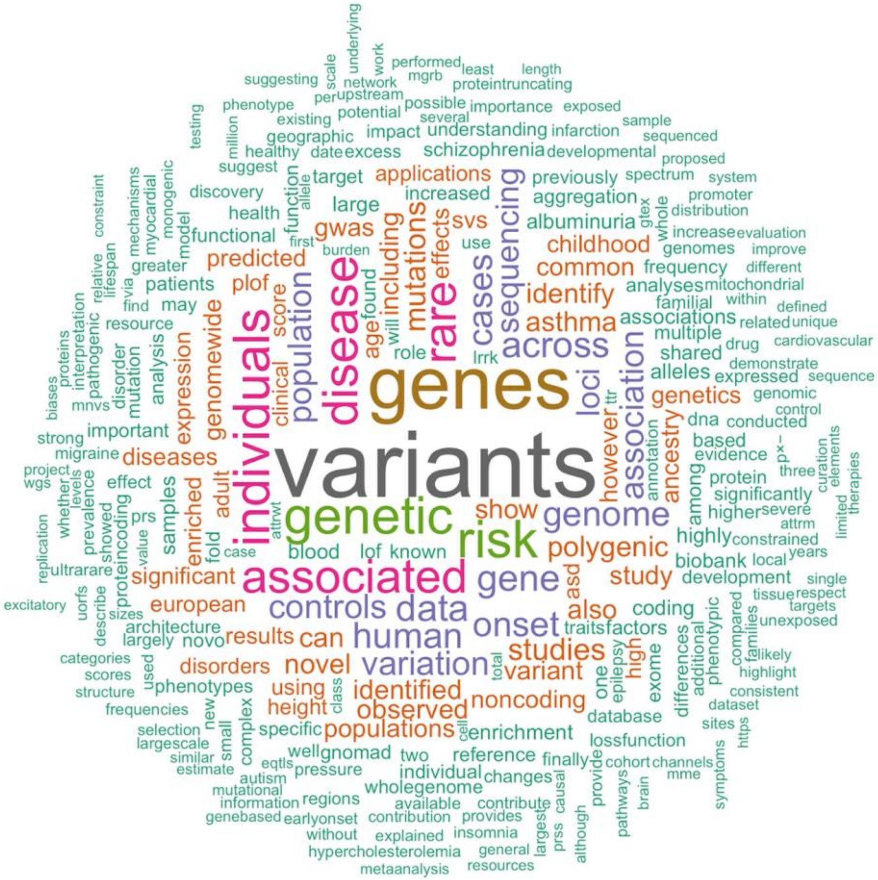
Unified Genomic Data Representation

The MatrixTable is a single interface for working with all kinds of genomic data

Community

Forum and chatroom for people interested in thinking + talking about genomic data analysis

How has Hail been used? (hail.is/references.html)

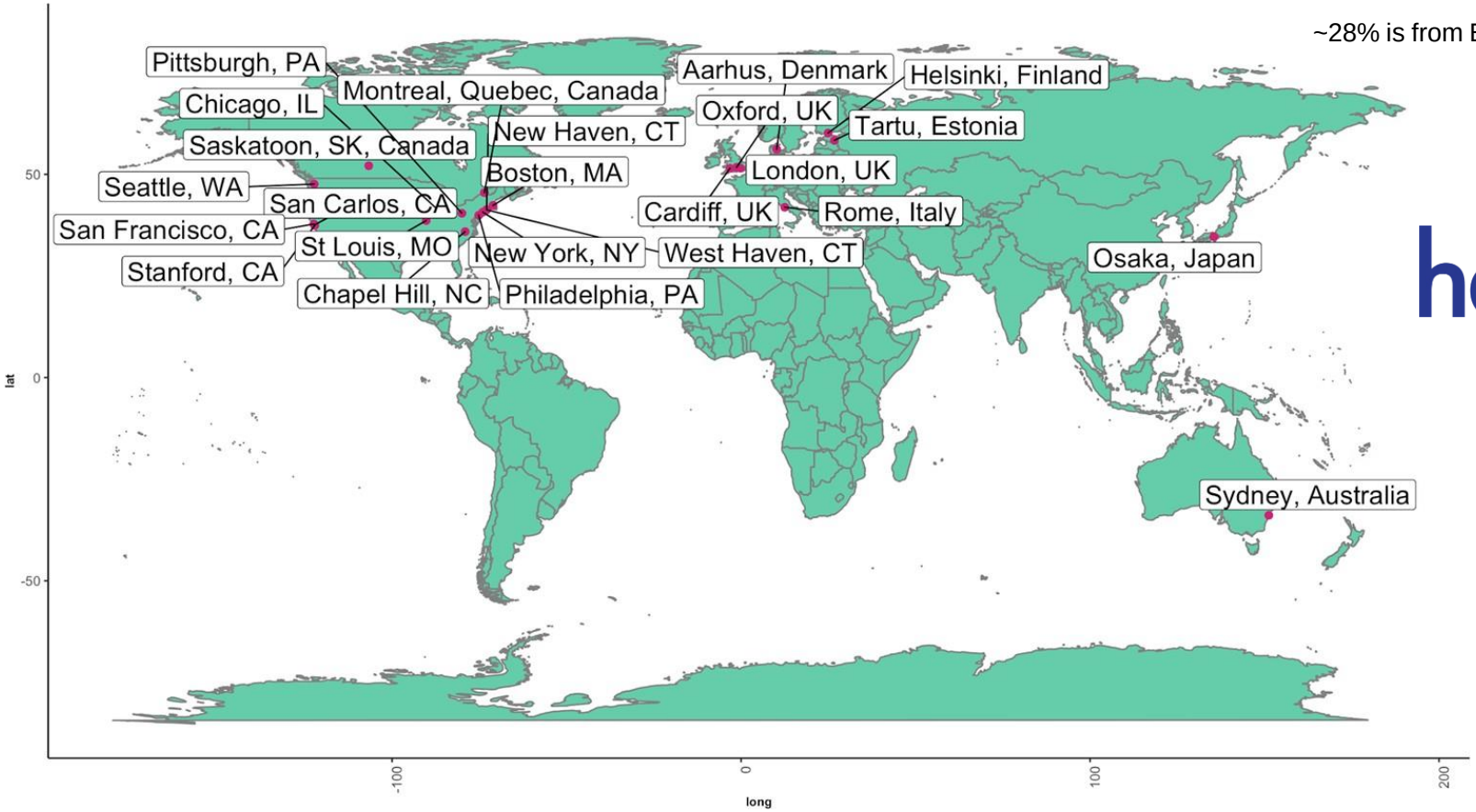


Notes:

- 51 abstracts (07/20/2020)
- Word appearing > 4x

Where has Hail been used?

~28% is from Boston, MA



Why would you use Hail?

Hail as a data science library

Data slinging

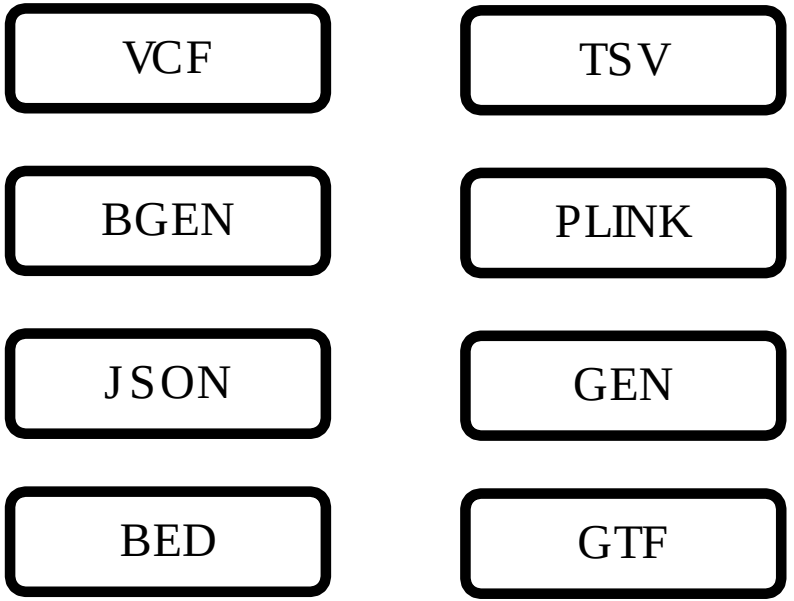
Analytical toolbox

Hail as a data science library

Data slinging

Analytical toolbox

- **Read and write common formats**
- Filter, group, aggregate
- Annotation
- Visualization



Hail as a data science library

Data slinging

Analytical toolbox

- Read and write common formats
- **Filter, group, aggregate**
- Annotation
- Visualization

- Compute mean depth per variant or per sample
  - Among heterozygotes
  - Grouped by ancestry labels & sex
- Count transitions & transversions called per sample

Hail as a data science library

Data slinging

Analytical toolbox

- Read and write common formats
- Filter, group, aggregate
- **Annotation**
- Visualization

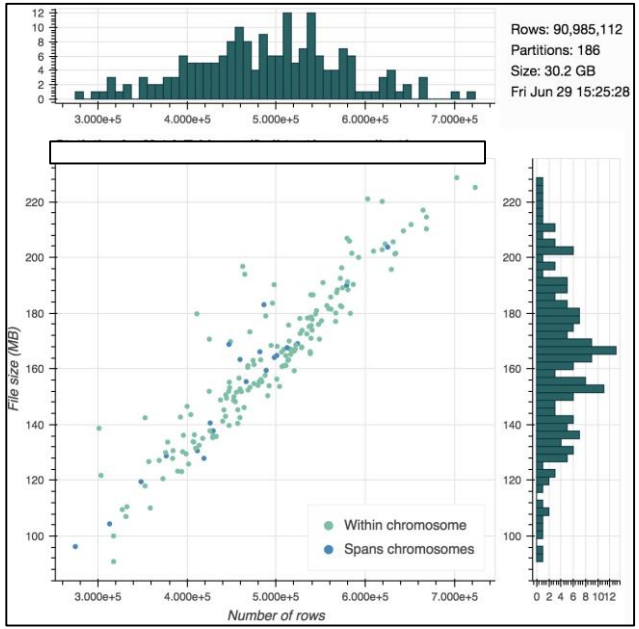
- Built-in wrapper for the Variant Effect Predictor (VEP). We did the setup so you don't have to!
- Join with annotations by variant, locus, interval, gene
- Annotation database

Hail as a data science library

Data slinging

Analytical toolbox

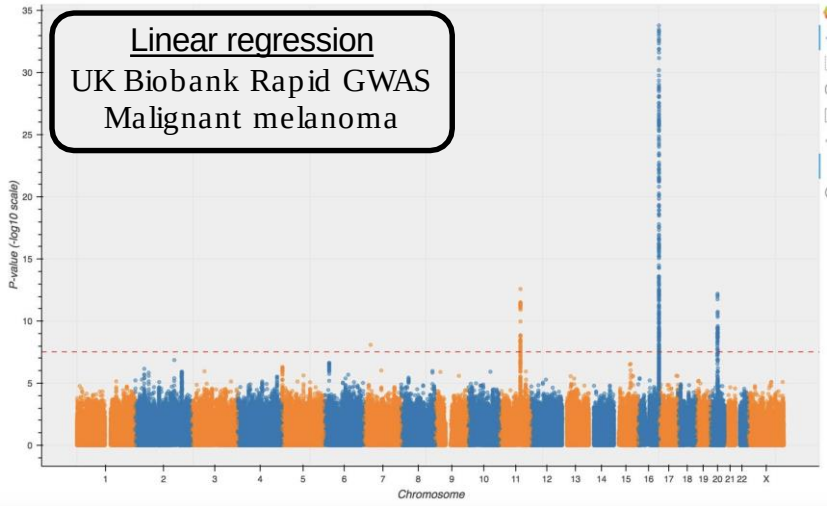
- Read and write common formats
- Filter, group, aggregate
- Annotation
- **Visualization**



Hail as a data science library

Data slinging

Analytical toolbox

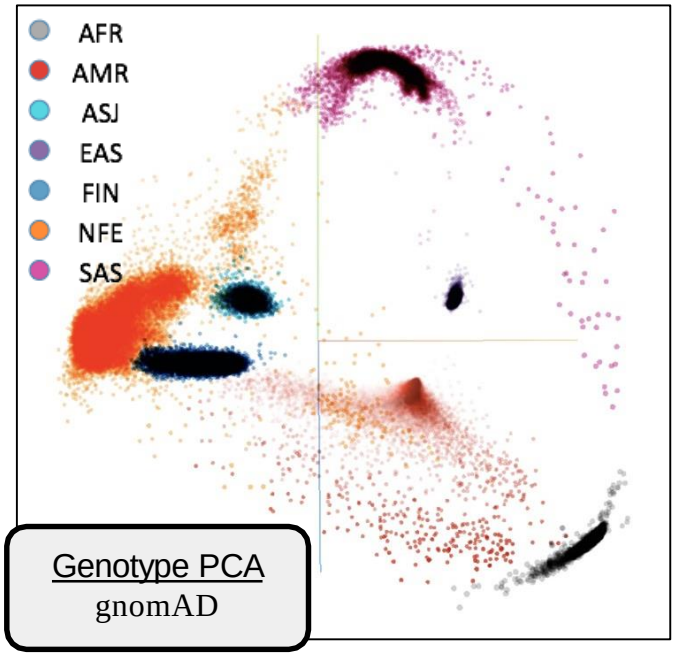


- **Statistical methods for genetics**
- Linear algebra

Hail as a data science library

Data slinging

Analytical toolbox



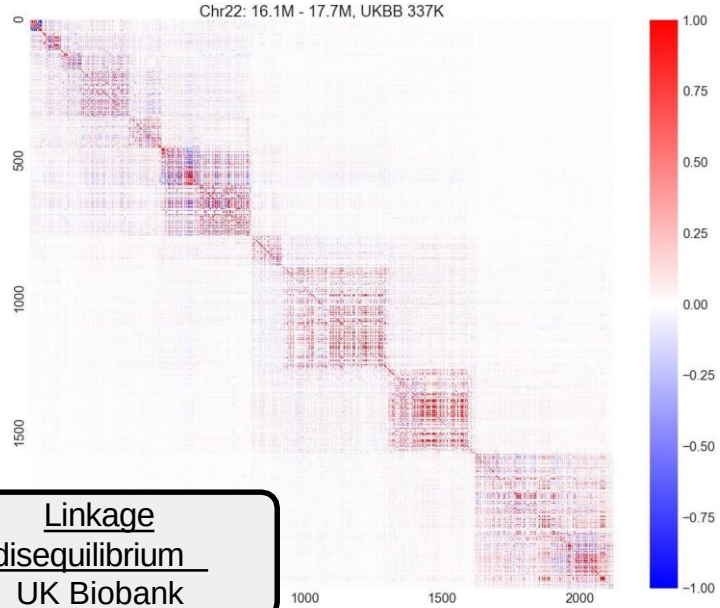
Genotype PCA
gnomAD

- **Statistical methods for genetics**
- Linear algebra

Hail as a data science library

Data slinging

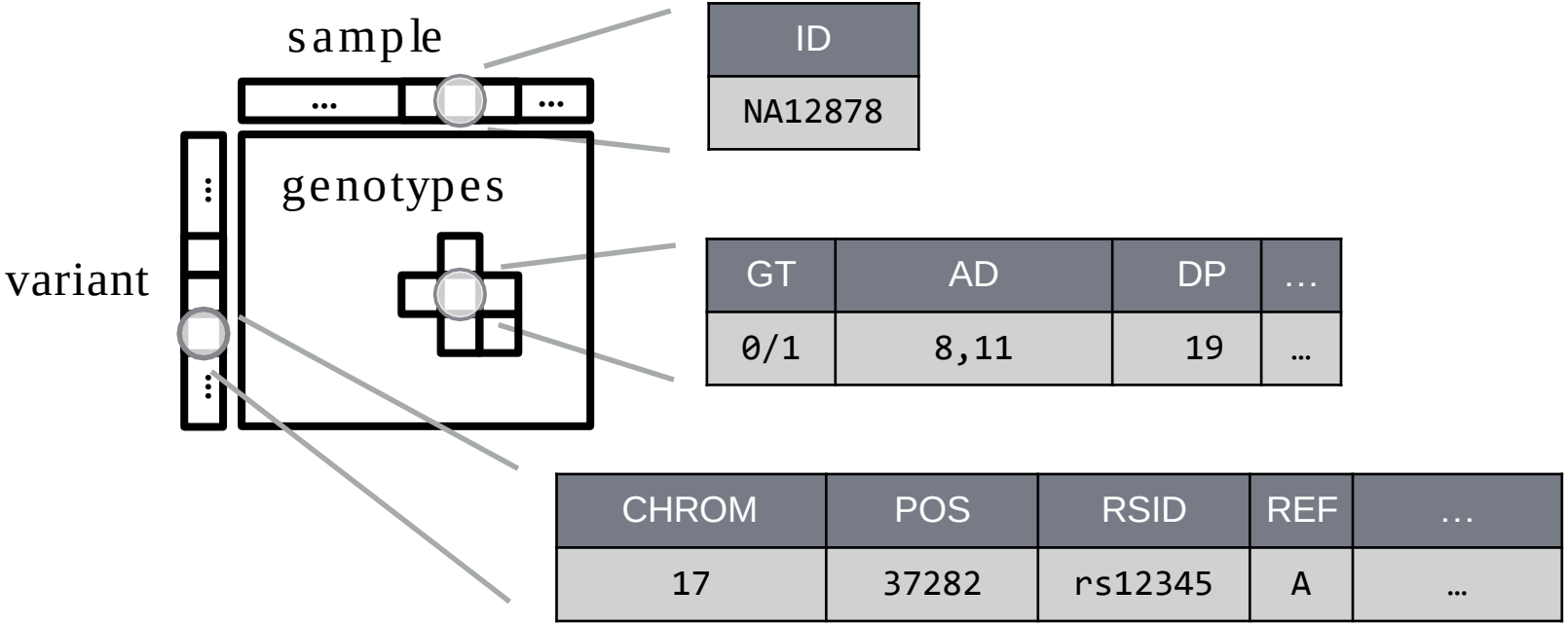
Analytical toolbox



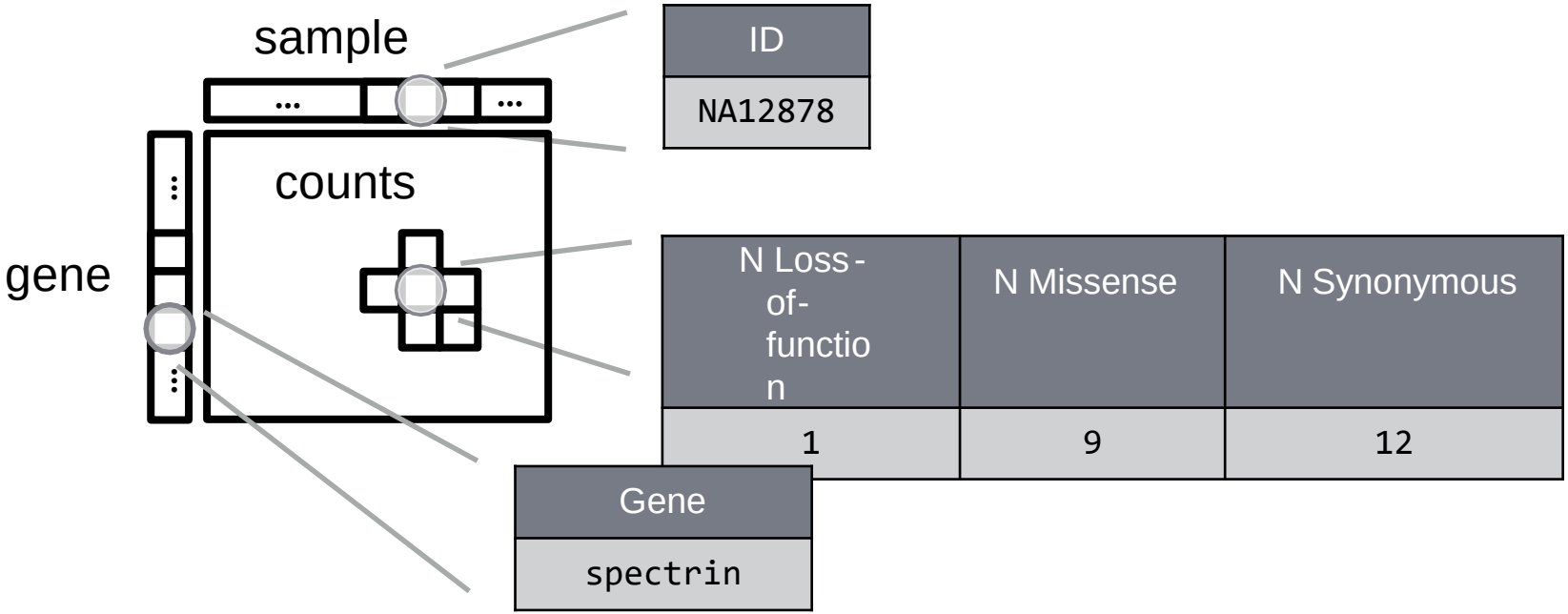
Linkage disequilibrium
UK Biobank

- Statistical methods for genetics
- **Linear algebra (early stages)**

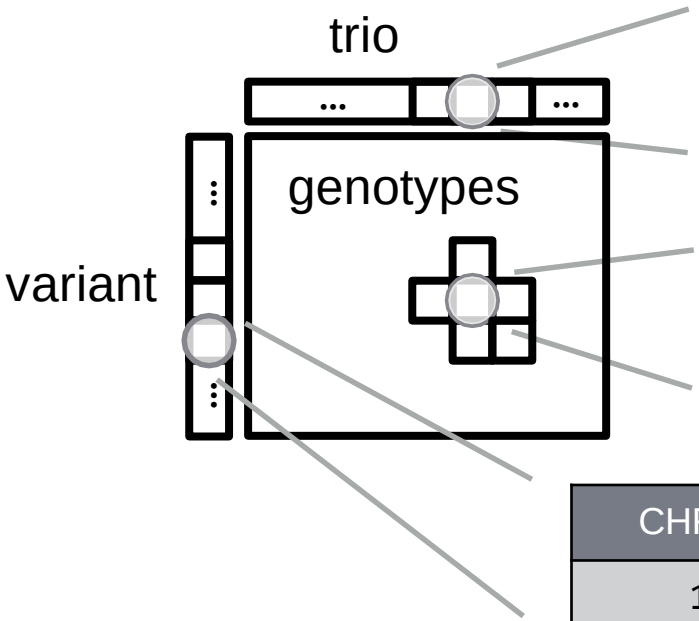
Variant Call Format (VCF)



Rare variant aggregation



Trio data

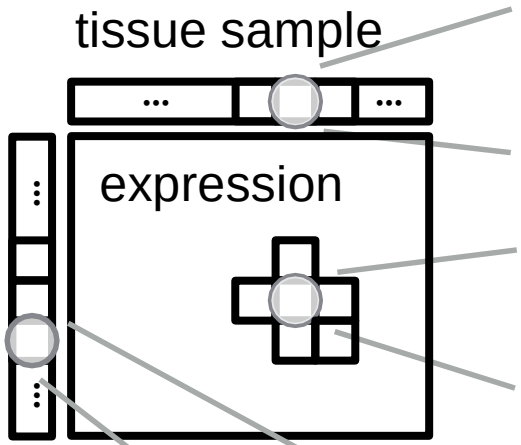


Proband	Mother	Father	…
NA12878	NA12891	NA12892	…

proband_GT	mother_GT	father_GT	…
0/1	0/0	0/0	…

CHROM	POS	RSID	REF	…
17	37282	rs12345	A	…

Transcript expression

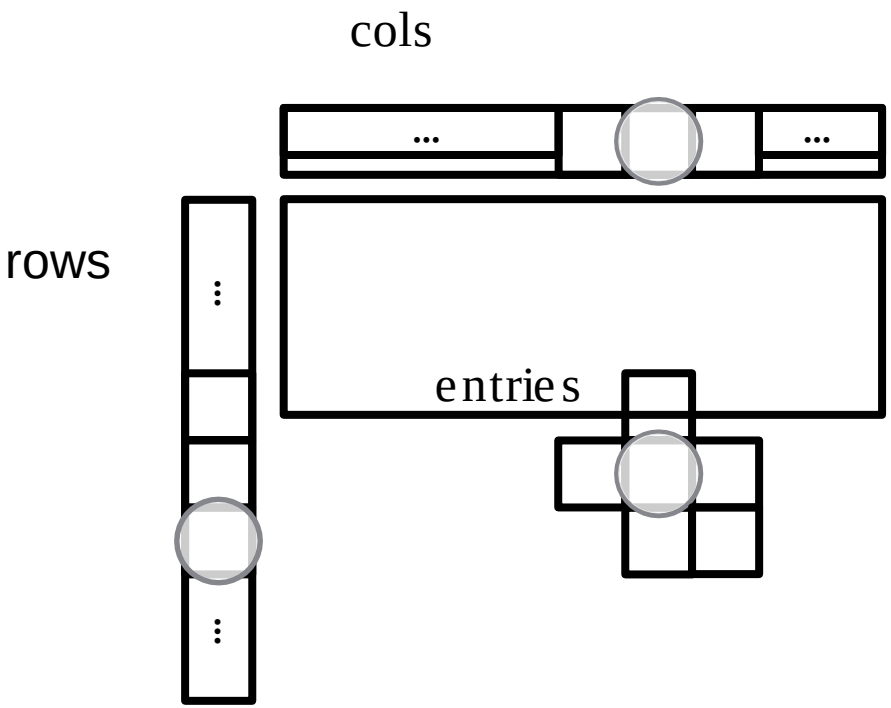


Sample	Tissue
NA12878	Liver

Raw_expression	Norm_expression
2	4.1232

Transcript	Gene
ENSG00000128487	spectrin

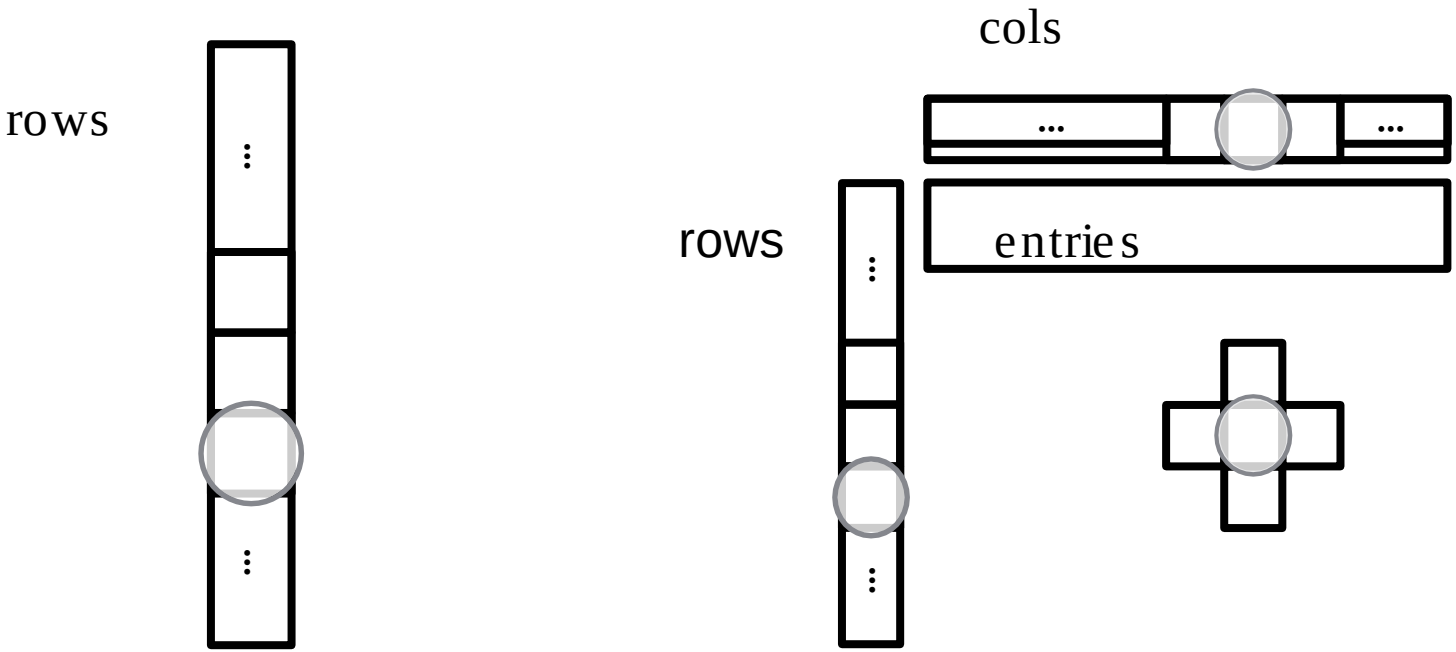
MatrixTable




```
----------------------------------------
Global fields:
    None
----------------------------------------
Column fields:
    's': str
----------------------------------------
Row fields:
    'locus': locus<GRCh37>
    'alleles': array<str>
    'rsid': str
    'qual': float64
    'filters': set<str>
    'info': struct {
        NEGATIVE_TRAIN_SITE: bool,
        AC: array<int32>,
        ...
        DS: bool
    }
----------------------------------------
Entry fields:
    'GT': call
    'AD': array<int32>
    'DP': int32
    'GQ': int32
    'PL': array<int32>
----------------------------------------
Column key:
    's': str
Row key:
    'locus': locus<GRCh37>
    'alleles': array<str>
----------------------------------------
```

Can be extended to rare variant aggregation, trio, transcript expression

Table



We have cheatsheets for this too!
https://hail.is/docs/0.2/cheatsheets.html

Large-scale datasets

- UK Biobank 500K => 5M?
  - … and many other biobanks
- gnomAD: 20K => 150K WGS
- TOPMed: >120K WGS
- All of Us: 1M
- Million Veterans Project: 1M
- ALS COMPUTE

From Bringing Data to Researchers

To Bringing Researchers to Data

Advantages:

- Many researchers can work on the same data at no additional cost
- No need to share resources with colleagues -- rent your own.
- High computer utilization means good cost efficiency
  - Pay for lots of CPUs when you need them, and pay nothing when you don’t
- Great security and fault-tolerance for data
- Democratization: don’t need access to institutional HPC cluster to participate in research (though do still need funding)

Disadvantages:

- Every operation has a cost, so an understanding of cost model is important
- Difficult to know how many resources to provision
  - Too small a cluster, you waste your time. Too large a cluster, you waste money.
- Cost overruns do happen
  - However, cloud providers will often refund accidental spend

Computational Landscape

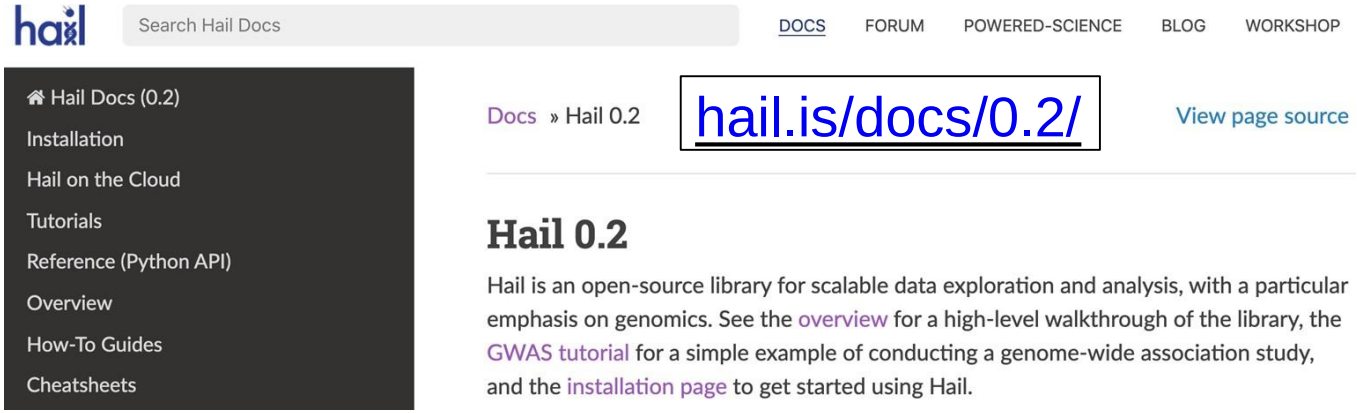
- Laptop/Desktop
  `pip install hail`
- Server, or a single node on High Performance Computing (HPC) cluster
  `pip install hail`
- High Performance Computing (HPC) cluster
  Institutional Spark cluster
  Hail *does not support* HPC schedulers like SLURM, UGER, and LSF
- Cloud
  Google Cloud Platform (GCP):

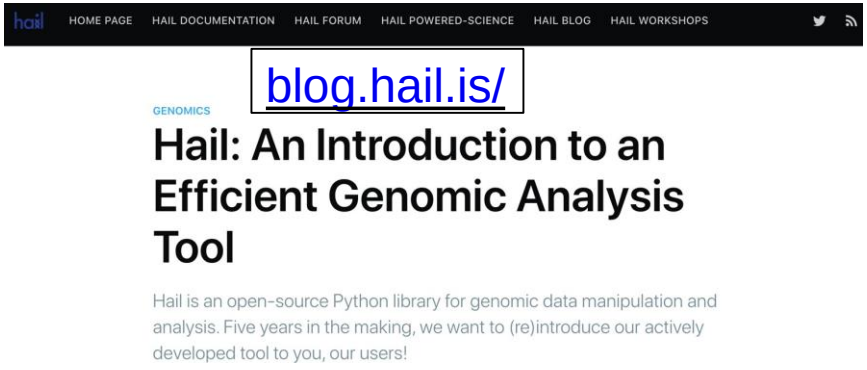
```
pip install hail

hailctl dataproc start CLUSTER
```

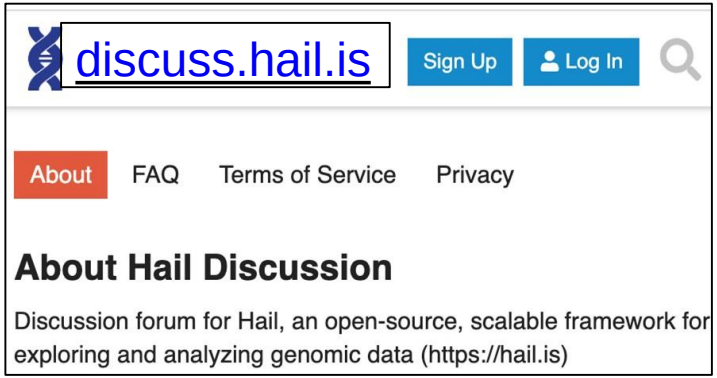
  Amazon Web Services (AWS): some support
  - https://github.com/hms-dbmi/hail-on-AWS-spot-instances
  - https://discuss.hail.is/t/spin-up-aws-emr-clusters-with-hail/818

Your next steps

```
pip install hail
```







Cloud Computing Platforms

amazon web services

best Hail support right now

Microsoft Azure

Cloud computing products from Google

Google Storage (sometimes referred to as "Google Buckets")

- Store data, Python notebooks, anything you want.
- ~$25 per TB per month.

Google Compute Engine (GCE)

- Rent a virtual machine, use it however you want.
- ~$0.05 per CPU per hour for standard VMs
- ~$0.01 per CPU per hour for preemptible VMs

Google Dataproc

- Rent a cluster running Apache Spark, which is Hail's distributed computing engine.
- GCE price, plus $0.01 per CPU per hour.

`hailctl`, the manager for Hail on the cloud

`hailctl` = "hail control"

- **`hailctl dataproc`** is the Hail cloud manager for Google.
  - **`hailctl emr`** (Amazon) and **`hailctl azure`** (Microsoft) planned.

Common cluster operations:

```
hailctl dataproc start MYCLUSTER --max-age 4h
hailctl dataproc connect MYCLUSTER notebook
hailctl dataproc submit MYCLUSTER script.py
hailctl dataproc stop MYCLUSTER
```

https://hail.is/docs/0.2/cloud/google_cloud.html#hailctl-dataproc

`gsutil`, file manager for Google Storage

`gsutil` = Google Storage Utilities

- Amazon and Microsoft clouds have their own analogs.

Create a new bucket (root directory)

```
gsutil mb gs://mybucket
```

List files in a bucket:

```
gsutil ls gs://mybucket
gsutil ls gs://mybucket/subfolder
```

Copy data to/from the cloud

```
gsutil cp gs://mybucket/file /Users/me/data/file
gsutil cp /Users/me/data/file gs://mybucket/file
```

https://cloud.google.com/storage/docs/gsutil

Cost management best practices

- **Develop small, run big**
  - Iterate on pipelines using a piece of the full dataset (make chr22 your bestie)
  - Run pipelines on large clusters when ready
- **Manage risk**
  - Set billing limits and alerts (you'll get an email if you start to overspend)
  - Always use **`--max-age`** or **`--max-idle`** flags on cluster creation
  - Use Buckets with retention policies (data deleted after X days) when possible
- **Plan Ahead**
  - Calculate costs ahead of time where possible
    - https://cloud.google.com/products/calculator/

https://hail.is/docs/0.2/cloud/general_advice.html