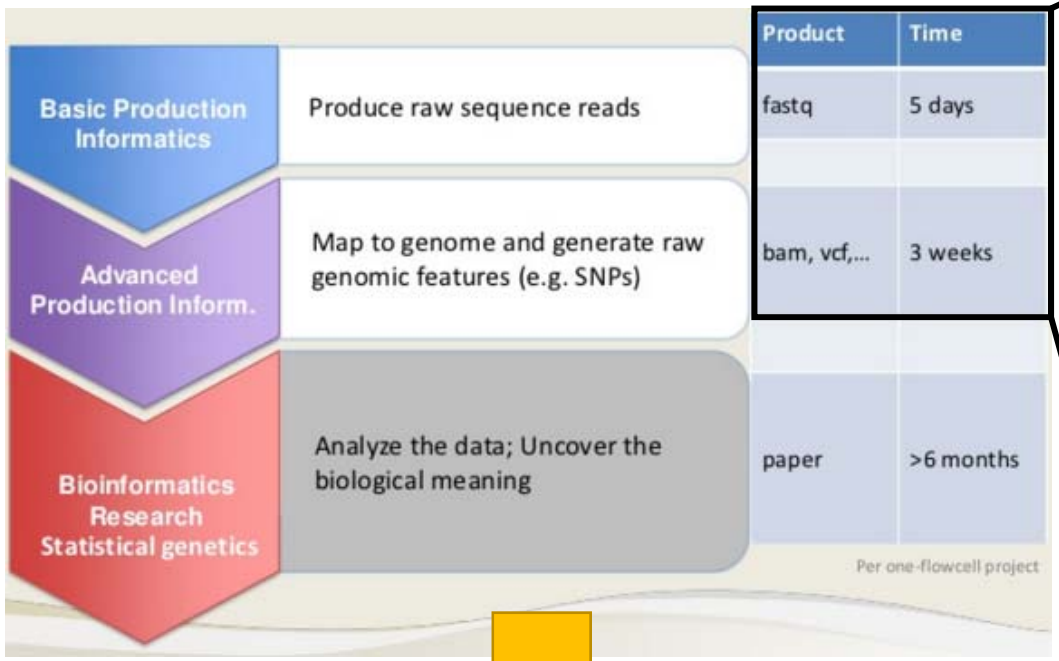
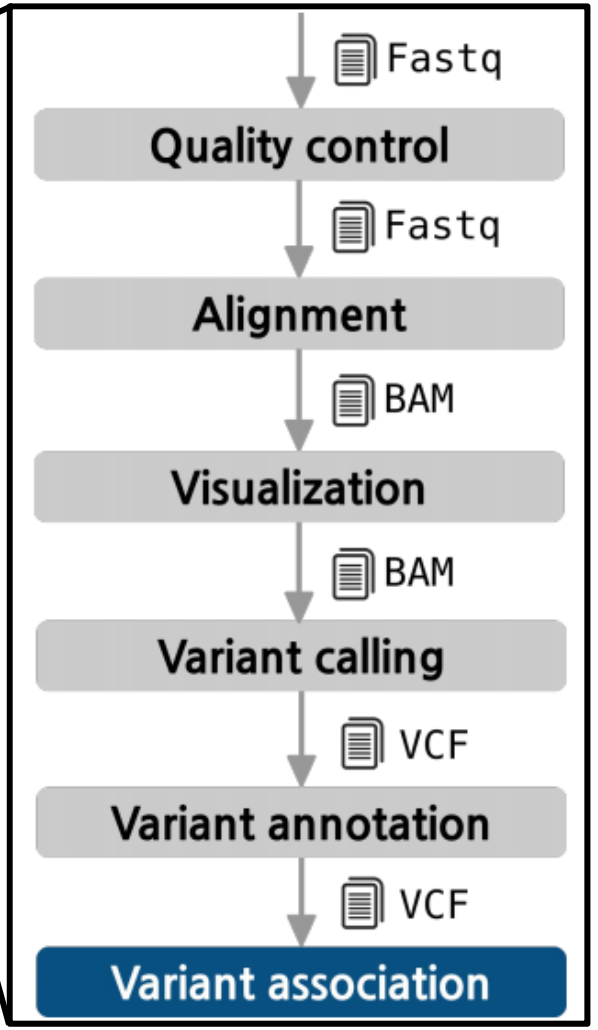
Variant prioritization

- The pipeline so far
- Finding the mutations causative of diseases
- The challenge
- So you think you found a deleterious mutation…
- Annotating the variants with ANNOVAR
- Understanding the annotation
- Your toolkit
- Class exercise

The pipeline so far



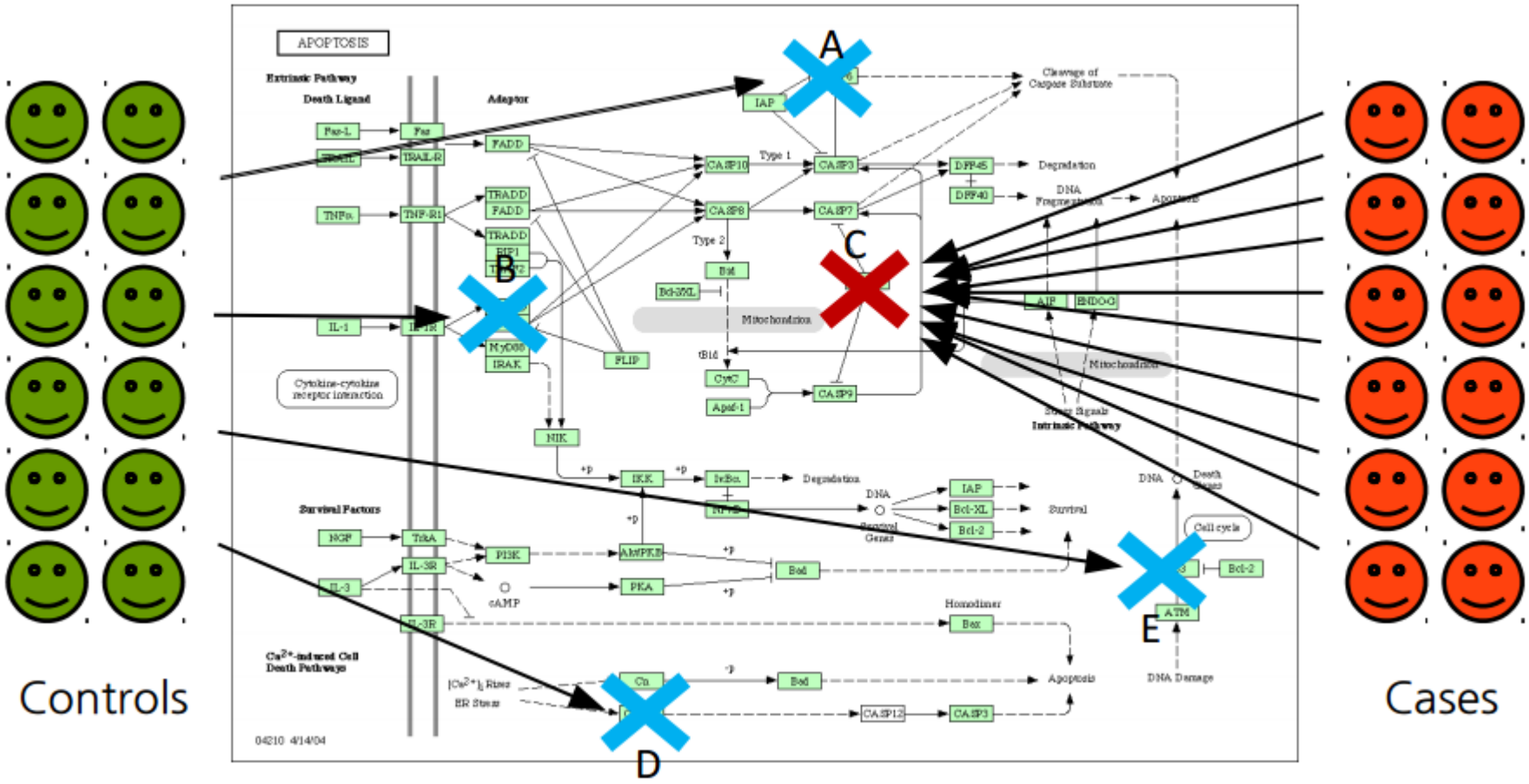


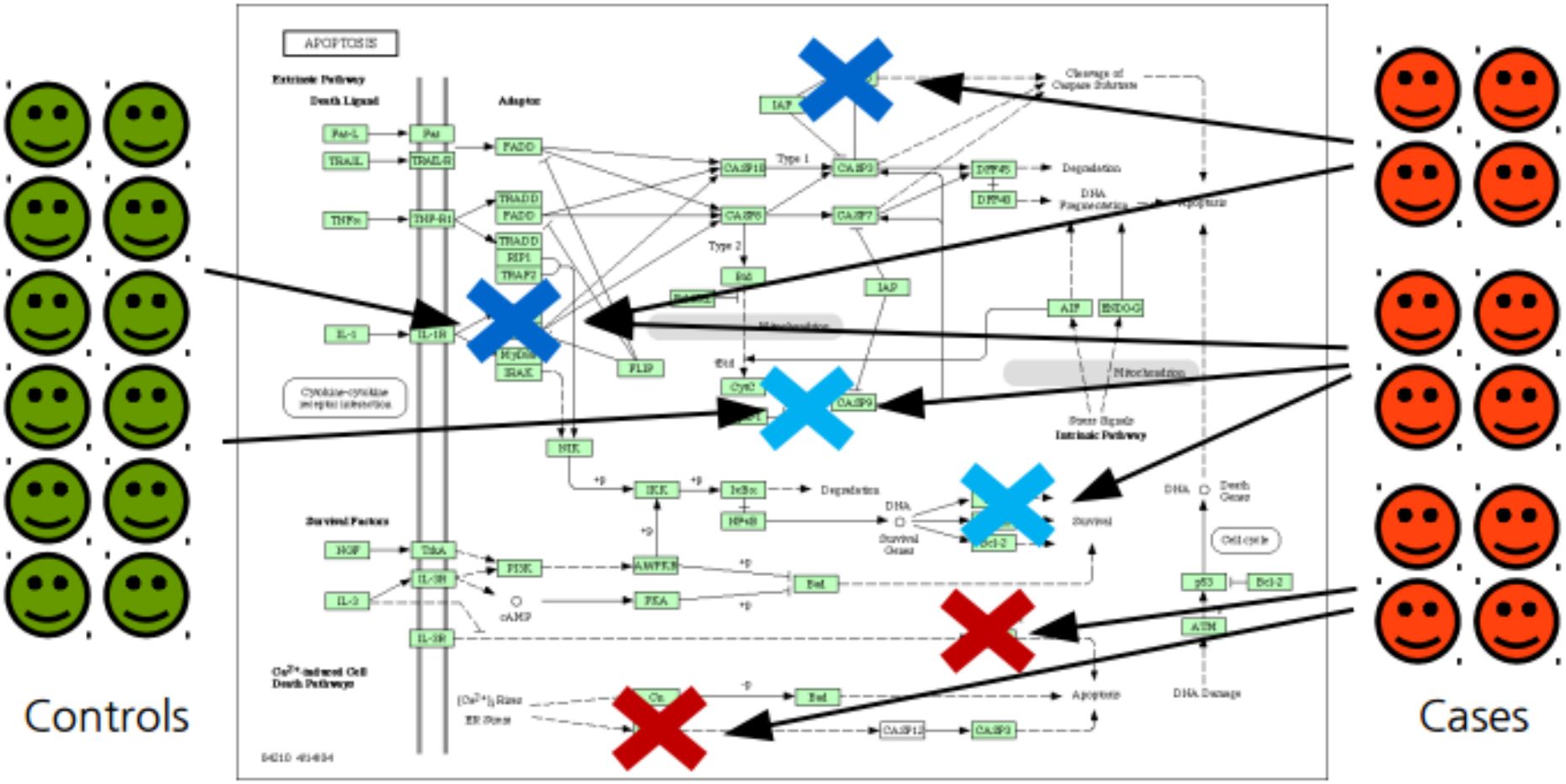
532 variants are significantly associated with the disease.

Now what ?

Finding the mutations causative of diseases

The simplest case: monogenic disease due to a <u>single</u> gene



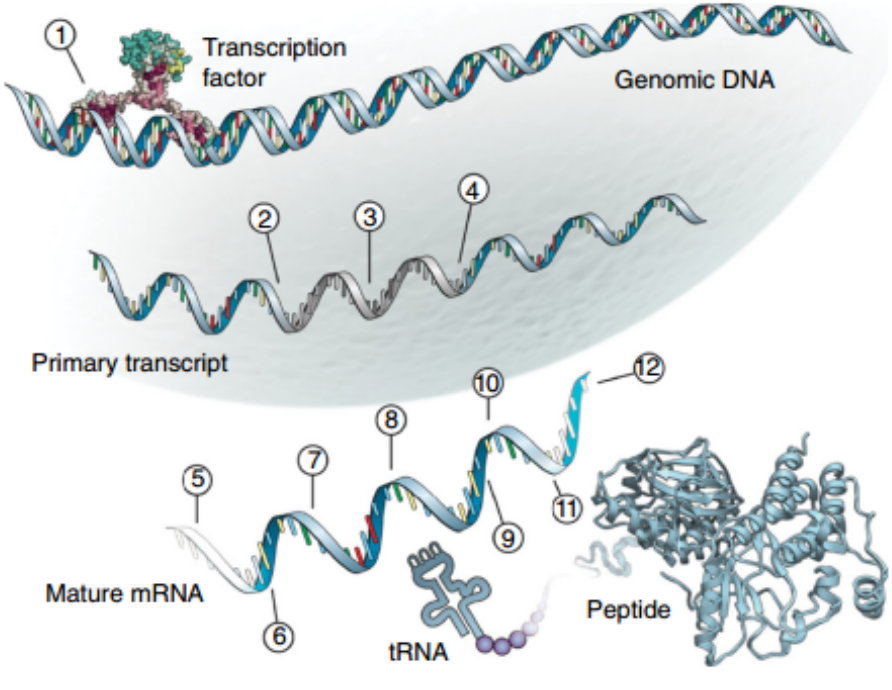


Clear individual gene associations are difficult to find in some diseases.

The same phenotype can be due to different mutations and different genes (or combinations). Many cases have to be used to obtain significant associations to many markers The only common element is the pathway (yet unknown) affected.

The challenge

- Disease related mutations can be anywhere within and around the gene.
- Each individual exome carries between 25,000 and 50,000 variants
- A whole genome can carry 3.5 million variants on average
- After annotating there will be hundreds of deleterious variants



1. TF binding site variant
2. Splice donor variant
3. Intron variant
4. Splice acceptor variant
5. 5′ UTR variant
6. Initiator codon variant
7. Missense variant
8. Synonymous variant
9. Frameshift variant
10. Stop gained
11. Stop lost
12. 3′ UTR variant

Figure 1 A set of annotation terms used to describe the potential effects of sequence variants according to the genic regions they fall in and their allele sequences. The terms are drawn from the Sequence Ontology and are depicted on the molecules they are predicted to affect. Variants categorized as any of the terms 2, 4, 9 and 10 are often collectively referred to as 'loss-of-function' variants, and are typically expected to severely affect gene function [25].

Ritchie and Flicek. Genome Medicine. 2014. 6:87

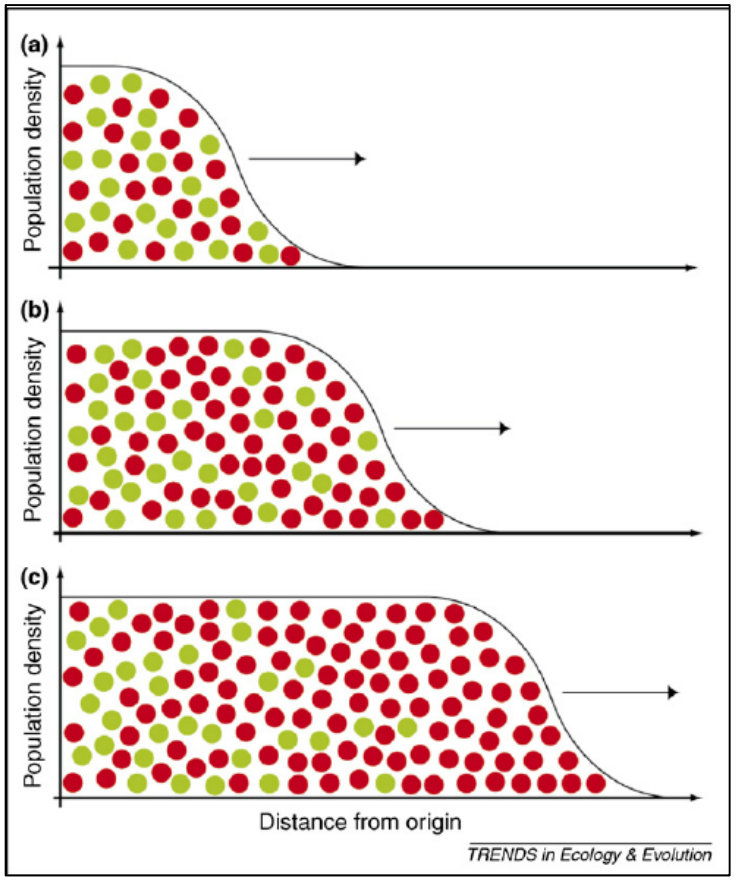
So what if you found an association?

Genetic drift leads to changes in allele frequencies causing false association

(a) Initial conditions show an equal proportion of two alleles (red and green).

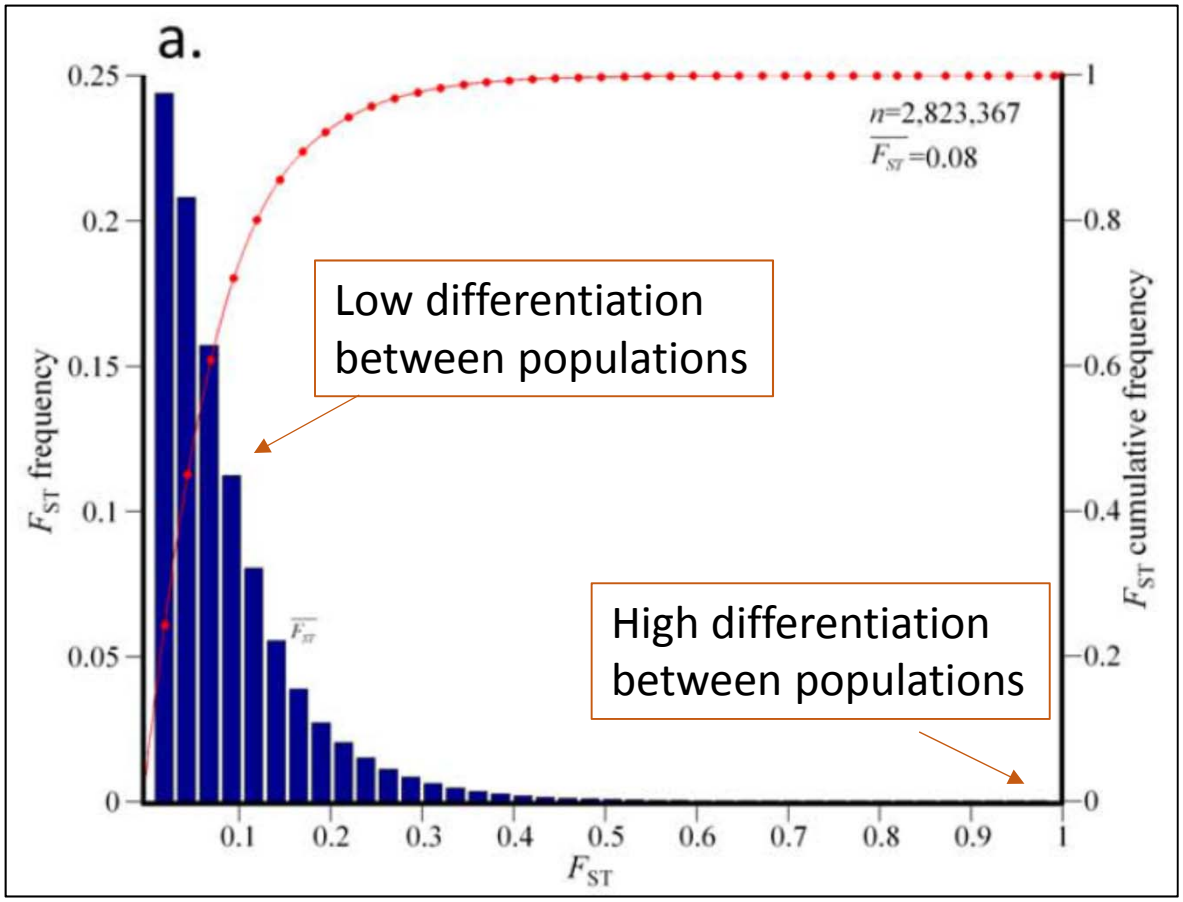
(b) The red allele increases in frequency owing to local drift.

(c) The red allele has become fixed by drift. Populations will only carry this allele.



Excoffier L, Ray N.. Surfing during population expansions promotes genetic revolutions and structuration. Trend Ecol Evol 23: 347-351

Allelic changes due to genetic drift can be misleading



Distribution of locus-specific F_{ST} in three continental populations (Europeans, Africans, and Asians).
F_{ST} values were obtained for 2.8M autosomal SNPs.

Adaptation Genetic drift

Pop1: ACG**C**TCA**G**CTAGCATAG
Pop2: ACG**A**TCA**C**CTAGCATAG

F_{ST}

Variants with opposite directions of effect in populations, will have high F_{ST} and higher levels of *P*-values which may be mistaken for true association

Rare variants are the worst!

Solutions: account for population structure

Elhaik et al. 2012. Plos one. 7.
Zawistowski et al. 2014. EJHG. 22. 1137–1144

Complexity of population structure may be difficult to account for

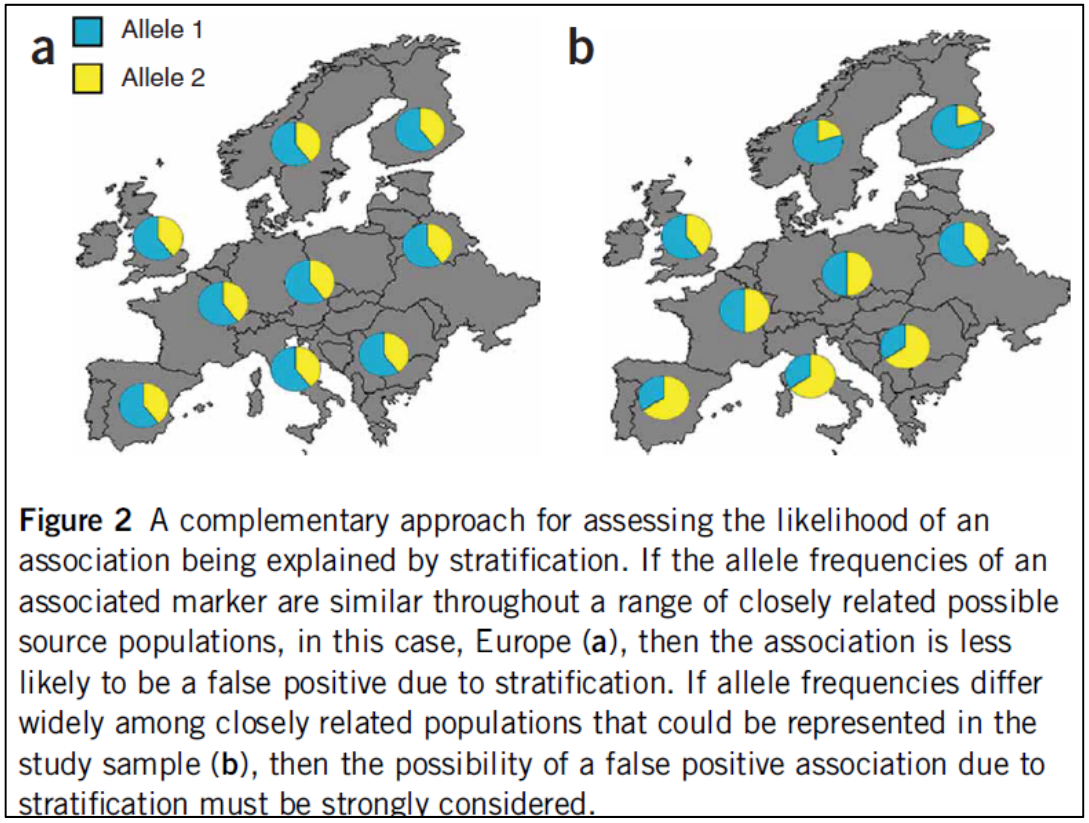


Figure 2 A complementary approach for assessing the likelihood of an association being explained by stratification. If the allele frequencies of an associated marker are similar throughout a range of closely related possible source populations, in this case, Europe (**a**), then the association is less likely to be a false positive due to stratification. If allele frequencies differ widely among closely related populations that could be represented in the study sample (**b**), then the possibility of a false positive association due to stratification must be strongly considered.

Questions to consider:

- How to define a population? (e.g., European-Americans, Scots, Ashkenazic Jews)
- How to model population structure? (e.g., PCA, Structure)
- How to account for population structure? (e.g., study design, statistical correction)

Solutions:

Large number of individuals from homogeneous population groups

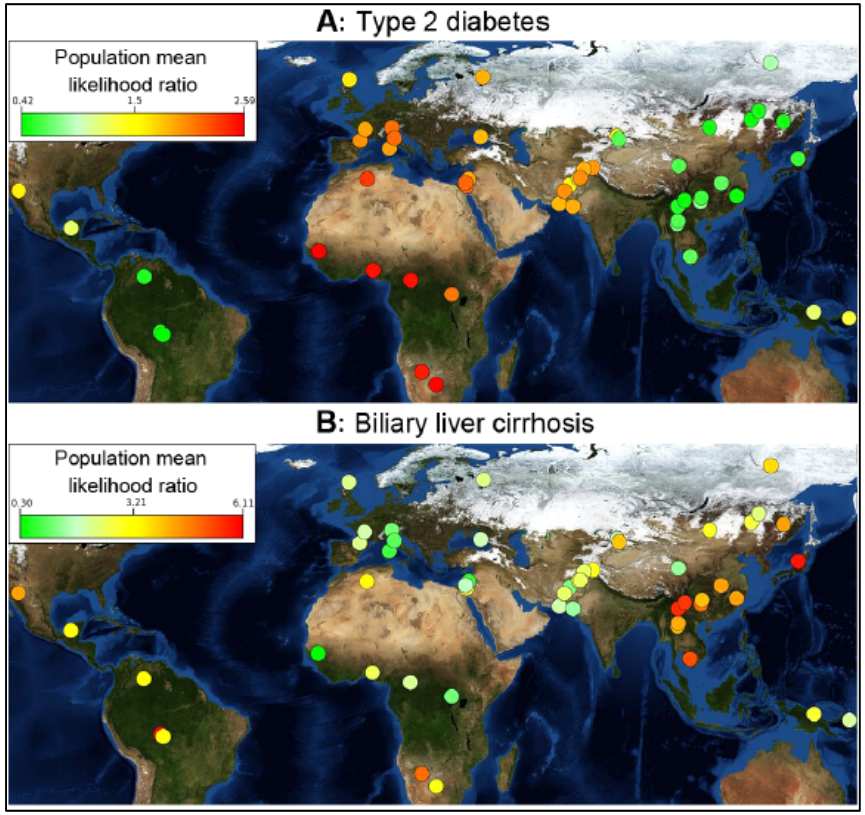
Admixture models capture population structure better

Beacon approaches to find the allele prevalence in different populations

Campbell et al. Demonstrating stratification in a European American population. Nat Genet. 2005;37(8):868-72.

By contrast, human populations do exhibit differences

Disease susceptibility



Corona et al.. *PLoS Genet.* 2013;9(5):e1003447

Drug response

Subject group	Reference group	Effect	Reference
Chinese	Caucasian, other S. Asian	Chinese patients are at substantially lower risk than European patients for cardiovascular complications after diabetes diagnosis, whereas South Asian patients were at comparable risk. Mortality after diabetes diagnosis is markedly lower for both minority populations.	Shah et al., 2013
Hispanic	Hispanic	The prevalence of hypertension and diabetes varies significantly among Hispanics by country of origin.	Pabon-Nau, et al., 2010
S. Asian, African	Caucasian	Type 2 diabetes 3- to 6-fold more likely in Africans and Asians, depending on ethnicity	Diabetes UK reports
Asian, African	Caucasian	Endometriosis is more common in Asian women and less common in African women as compared to Caucasian women	Gerlinge et al. 2012
Chinese	Caucasian	In glaucoma, multiple observable eye parameters differ between these racial groups; suggest potentially different mechanisms in occludable angle development in the two racial groups	Wang et al., 2013

B. R. Shah *et al.*, *Diabetes Care* **36**, 2670-2676 (2013).
L. P. Pabon-Nau *et al.*, *J. Gen. Intern. Med.* **25**, 847-852 (2010).
C. Gerlinger *et al.*, *BMC Womens Health* **12**, 9 (2012).
Y. E. Wang *et al.*, *Invest. Ophthalmol. Vis. Sci.* **54**, 7717-7723 (2013).

So what if you found a deleterious alleles?

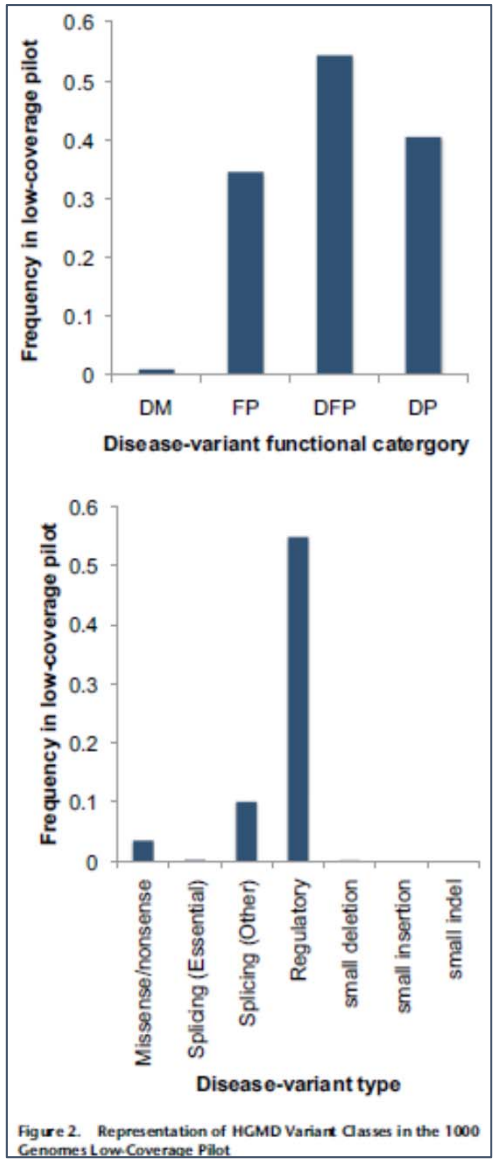


Figure 2. Representation of HGMD Variant Classes in the 1000 Genomes Low-Coverage Pilot

Xue et al. 2012. AJHG. 91, 1022–1032,

CAUTION!

On average, each healthy person carries:

- ~11,000 synonymous variants
- ~11,000 non-synonymous variants
- 250 to 300 loss-of-function variants in annotated genes
- 50 to 100 variants previously implicated in inherited disorders

1000 Genomes Project Consortium. *A map of human genome variation from population-scale sequencing.* **Nature**. 2010 Oct 28;467(7319):1061-73. PubMed PMID: 20981092

Solutions: experimental validation, larger cohorts, replication

DM, disease-causing mutation in HGMD.
FP, in vitro or in vivo functional polymorphism but with no disease association reported as yet.
DFP, disease-associated polymorphism with additional supporting functional evidence with disease and that has evidence of being of direct functional importance.
DP, disease-associated polymorphism with a disease or phenotype and that is assumed to be functional, although there might not yet be any direct evidence of a functional effect.

Real life example

You are studying a very rare disorder.
It is a devastating disorder. No known variants. No treatment. No cure.
You were lucky enough to find 7 families and obtained their complete-genome data.

One of the families is known to be inbred. You excluded it from the study.
Looking at the literature for Europeans, Africans, and Asians, you interpret your inbreeding statistic (pi hat) of the other families as an indication of no inbreeding.

Due to genetic privacy you have no demographic information about the patients.
You know that the inbred family was provided by a Middle Eastern country.
The 6 other families are from different American hospitals.

- You found a **deleterious variant** that segregated in all families.
- The gene's annotation is **highly suggestive** of causation.
- **A replication study** in unrelated patients from Arizona **confirmed the** segregation of the variant compared to European-American controls.
- **All the results are highly statistically significant.**

You are **very excited** and want to submit your findings to *Nature*.
Your scientific Spidey-sense urges you to rethink your study design.

The challenge

Most of the SNPs are probably not causative, but can we know which of them are?

Biological validation

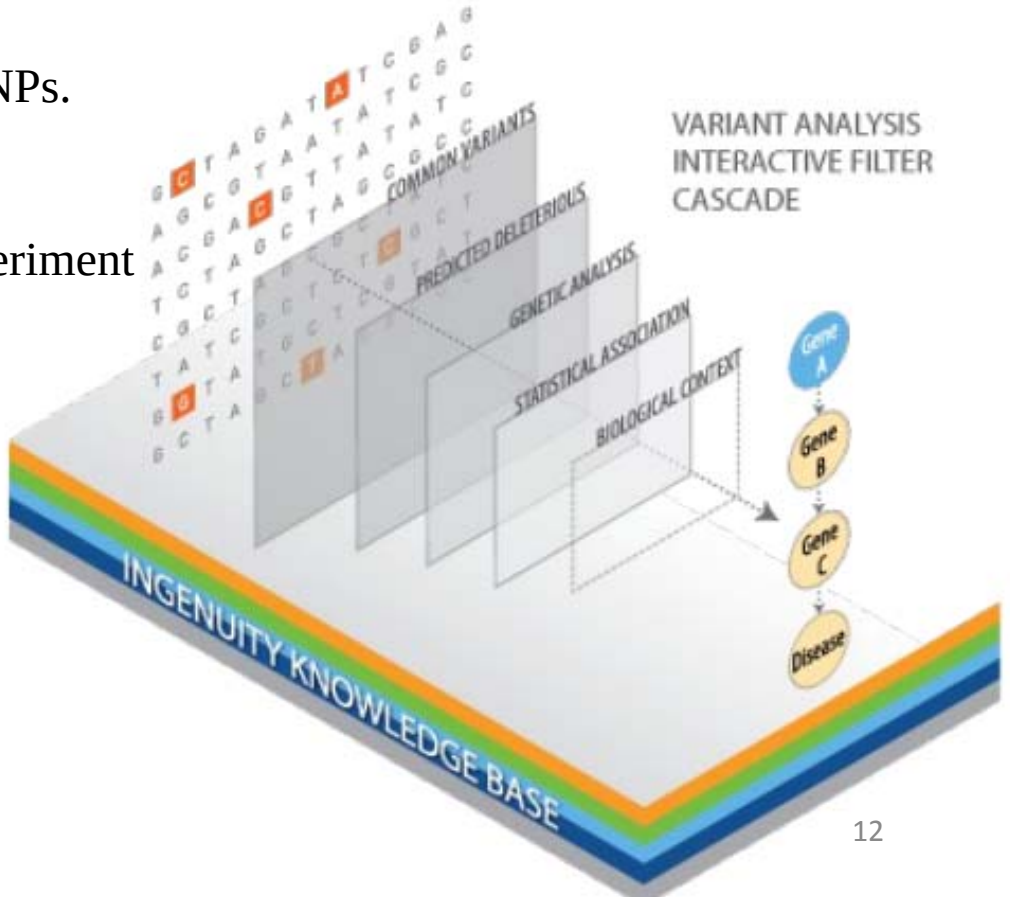
- Re-sequencing with *different* methods (e.g., Sanger) will allow you to confirm the existence of a variant.
- But you cannot do that for hundreds of SNPs.

Experimental valuation on animal models

- You can test whether gene knock-out experiment will produce a similar phenotype.
- But its expensive, slow, and has its own problems.

Bioinformatic validation

- A proxy to the reality.
- Cheap! but may be wrong!



How to decide?

ducibility Reproducibility Reproducibility Reproducibility Reproducibil
ducibility Reproducibility Reproducibility Reproducibility Reproducibil
ducibility Reproducibility Reproducibility Reproducibility Reproducibil
ducibility Reproducibility Reproducibility Reproducibility Reproducibil
ducibility Reproducibility Reproducibility Reproducibility Reproducibil
ducibility Reproducibility Reproducibility Reproducibility Reproducibil
ducibility Reproducibility Reproducibility Reproducibility Reproducibil
ducibility Reproducibility Reproducibility Reproducibility Reproducibil
ducibility Reproducibility Reproducibility Reproducibility Reproducibil
ducibility Reproducibility Reproducibility Reproducibility Reproducibil
ducibility Reproducibility Reproducibility Reproducibility Reproducibil
ducibility Reproducibility Reproducibility Reproducibility Reproducibil

Annotating the Variant

You did the perfect genetic study:

- You chose a highly heritable disorder.
- You collected many cases and controls.
- You studied the complete genome.
- You accounted for biases in the sequencing process and population structure.
- You did a replication study and removed variants that were not replicated.

You still have several hundreds candidate variants!

How will you decide which of those is causal?

Variant prioritization

We can prioritize variants in several wats giving more weight to SNPs based on certain characteristics. To have an idea where to look for you are required to have an *a priori* hypothesis or knowledge of the disease you study.

In psychiatric disorders, the priority is rare variants (at the time being). Therefore, minor allele frequency is considered as one prioritization characteristic. Variants that show expression in the brain are also favored.

In diabetes, we look at functionality.

What is the problem with this approach?

Using ANNOVAR for annotation

(http://annovar.openbioinformatics.org/en/latest/)

- ANNOVAR is not a prediction tool – it is a shell.
- ANNOVAR will not decide for you. It will generates a series of predictors made by other tools and combine it together in a user-friendly manner.
- ANNOVAR obtains some of its data from DBNSFP (https://sites.google.com/site/jpopgen/dbNSFP), which has > 400 categories to classify variants collected from different sources.
  https://drive.google.com/file/d/0B60wROKy6OqccXV4UXNyNkJwNUk/view?pref=2&pli=1
- ANNOVAR is implemented in GALAXY.

What are the problems with this type of information?

Table 4 Comparison of recent tools to systematically annotate variants[a]

Tool	Type	Input method	Protein annotation	Regulatory annotation	Other
SeattleSeq (http://snp.gs.washington.edu/SeattleSeqAnnotation/)	Server	Variants	Deleteriousness scores	Conservation scores	dbSNP clinical association data
ANNOVAR[57]	Software	Variants, regions	User defined: user downloads desired variation, conservation, coding and noncoding functional annotations		
ENSEMBL VEP[56]	Server	Variants, regions	Deleteriousness scores	Regulatory motif alteration scores	OMIM, GWAS data
VAAST[58]	Software	Variants	Deleteriousness scores	Conservation scores	Aggregation to discover rare variants in case-control studies
HaploReg[54]	Server	Variants, studies	dbSNP consequence data	Chromatin state, protein binding, DNase, conservation, regulatory motif alteration scores	GWAS data, eQTL, LD calculation, enrichment analysis per study
RegulomeDB[55]	Server	Variants, regions	Not applicable	Histone modification, protein binding, DNase, conservation, regulatory motif alteration scores	eQTL, reporter assays, combined score analysis per variant

[a]Many such tools have been released as databases or software in the past decade; a sampling of the most recent are listed here.

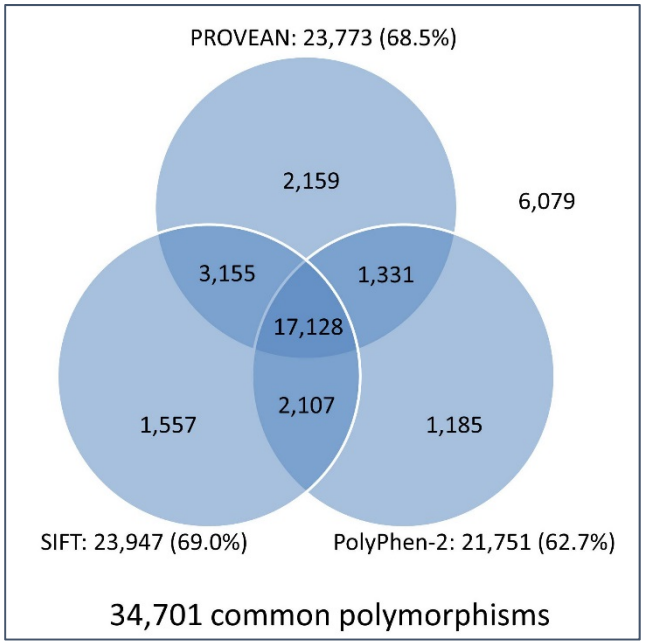
L. D. Ward, M. Kellis, *Nat. Biotechnol.* **30**, 1095-1106 (2012).

Understanding the annotation

Functional predictions

For variants found in protein-coding mutations, knowledge of protein structure and function, and the unambiguous nature of the genetic code, have allowed the development of a class of predictive algorithms that can score the severity of missense and nonsense variants (e.g., SIFT and PolyPhen).

These tools product a decision about variants in the form of:
Damaging/Possibly damaging/Benign



http://provean.jcvi.org/about.v1.0.php

Conservation scores

Even in the absence of conserved sequence, the conservation of biochemical activity can be indicative of conserved functional elements, even when the corresponding sequence features are not detectable by traditional alignment and constraint measures owing to turnover.

Because some fraction of protein binding and RNA transcription may be nonfunctional ‘noise,’ cross-species analysis of transcription factor binding or gene expression can help reveal the subset of elements that are most likely to be functional. However, lineage-specific elements may nevertheless be important and may not be captured through this method.

Common tools: Phylop, phastCons

Other categories

Noncoding variants. Several resources, including HaploReg, RegulomeDB, and ENSEMBL's Variant Effect Predictor annotate noncoding common variants from association studies using conservation, functional genomics and regulatory motif data.

Gene set enrichment analysis. Prior knowledge of gene interrelationships can be leveraged gene expression studies to discover differentially regulated pathways, even where single genes in those pathways change expression too little to rise to statistical significance. These methods for gene-set enrichment analysis (GSEA) are being applied to GWAS, where, similarly, genetic risk is expected to be concentrated along biological pathways and multiple testing diminishes the statistical significance of associations considered individually.

Regulatory element enrichment analysis. A recent study used chromatin state maps to discover an enrichment of cell type–specific enhancers among the top associations in several GWAS, demonstrating the ability of high-resolution functional genomics maps to serve as a type of pathway annotation

Gene expression. Variation in gene expressions in different tissues may imply on the functionality of the gene (e.g., GTEx).

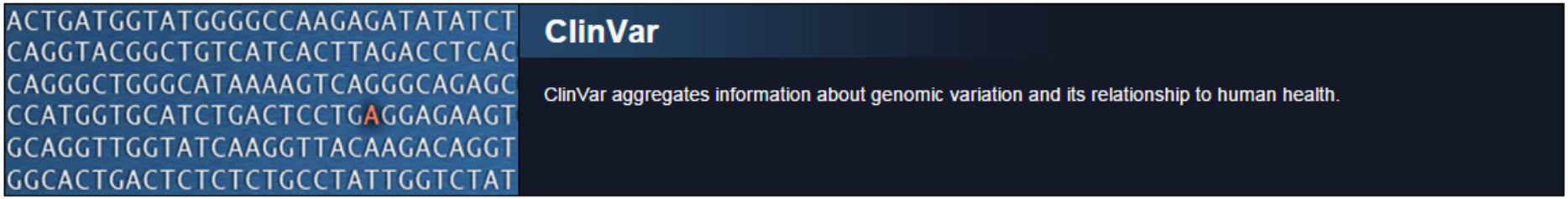
Allele frequency

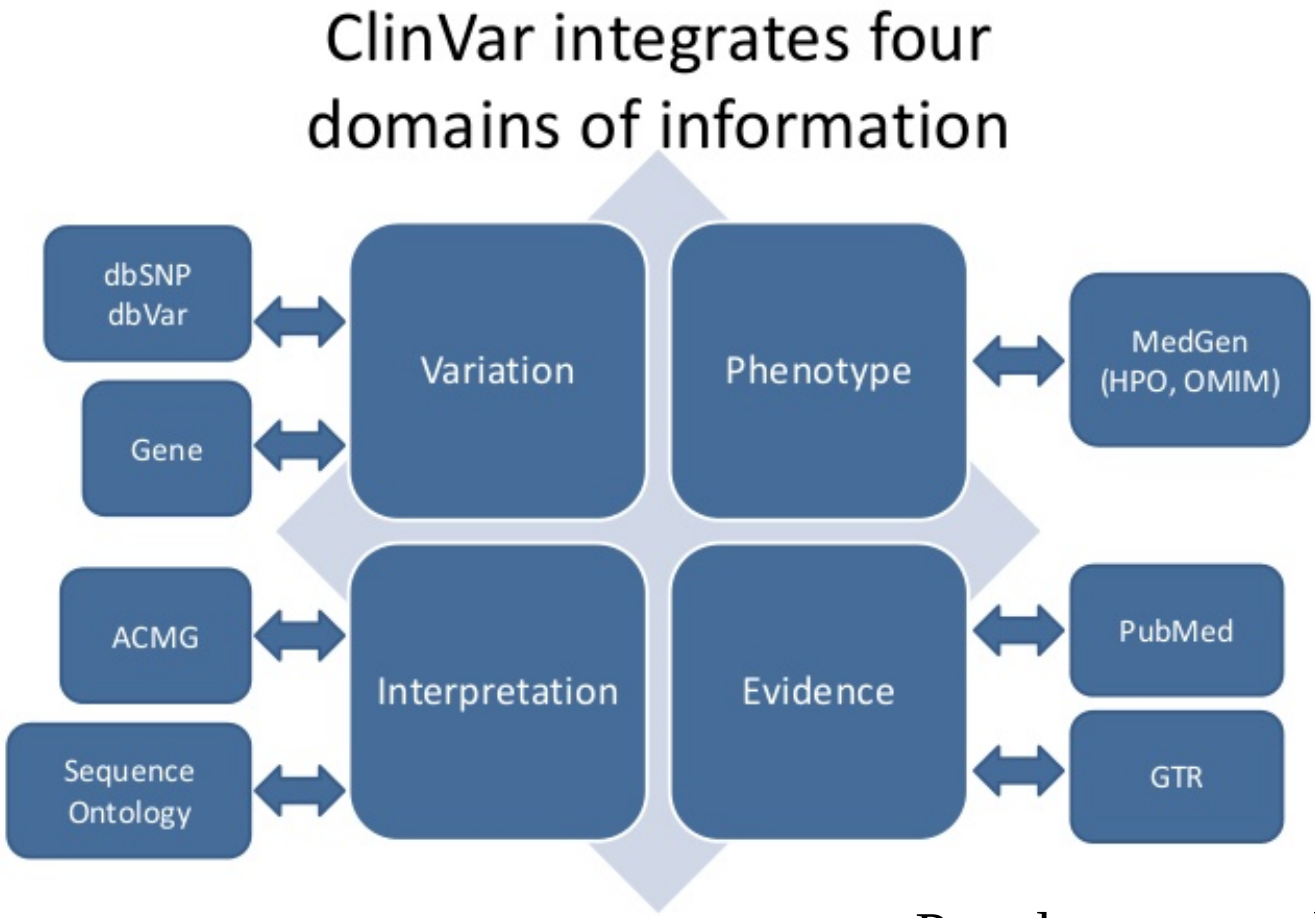
Knowledge of your population is critical.

You can identify the geographical origin of a population using GPS.
GPS will give you an idea of candidate populations.

You can use EVS, HGDP, 1000 Genomes, or Exac (broad institute) to get the allele frequency of your variants.

E. Elhaik *et al.*, *Nat Commun* **5**, (2014).





Popular among clinicians…

Your toolkit

Tool	Application	Comments	URL	Reference
Annotation based on overlap with and proximity to functional elements				
Ensembl Genome Browser	Manual variant annotation and genomic context	Web server, data also available via Perl and REST APIs	http://www.ensembl.org	[10]
UCSC Genome Browser	Manual variant annotation and genomic context	Web server, data also available for download using the UCSC table browser	http://www.genome.ucsc.edu	[11]
Bedtools	Automatic high performance feature overlap and proximity	Command line tool and Python interface	http://bedtools.readthedocs.org	[12]
Bedops	Automatic high performance feature overlap and proximity	Command line tool	http://bedops.readthedocs.org	[13]
HaploReg	Web server identifying non-coding annotations for variants and haplotypes	Web server with pre-computed results for several GWAS	http://www.broadinstitute.org/mammals/haploreg/	[14]
Biologically informed rule-based annotation				
Ensembl Variant Effect Predictor (VEP)	Wide support for variant annotation, emphasis on genic variants, but also incorporates regulatory elements and TF motifs from JASPAR	Downloadable software, web server, Perl and REST APIs, plugin system to add functionality	http://www.ensembl.org/vep	[17]
ANNOVAR	Annotation of genic variants, can also identify overlaps with other annotated elements	Downloadable software	http://www.openbioinformatics.org/annovar/	[18]
VAT	Annotation of genic variants	Downloadable software	http://vat.gersteinlab.org	[20]
SnpEff	Annotation of genic variants, companion tool SnpSift can filter results by annotations	Downloadable software	http://snpeff.sourceforge.net	[19]
RegulomeDB	Identifies overlaps with non-coding elements and applies heuristic rules to predict consequences	Web server	http://regulome.stanford.edu	[24]

More tools…

Tool	Application	Comments	URL	Reference
Annotation based on sequence motifs				
JASPAR	Open access database of TF binding PWMs	Queryable interface and database downloads	http://jaspar.genereg.net	[26]
MEME suite	Several tools for handling PWMs	Web services and downloadable tools	http://meme.nbcr.net	[27]
MOODS	Tool for aligning PWMs to sequences	Command line tool	http://www.cs.helsinki.fi/group/pssmfind/	[28]
Human Splicing Finder	Tool for computing the effects of mutations on splicing	Web server	http://www.umd.be/HSF/	[29]
Annotation based on constraint estimated from multiple sequence alignments				
GERP	Nucleotide resolution conservation scores	Downloadable software, pre-computed scores and elements for human and mouse genomes	http://mendel.stanford.edu/SidowLab/downloads/gerp/	[31]
PHAST package	Suite of tools for phylogenetic analyses, including phastCons and phyloP	Downloadable software and R package	http://compgen.bscb.cornell.edu/phast/	[32,33
SCONE	Position-specific conservation scores	Downloadable software	http://genetics.bwh.harvard.edu/scone/	[34]
SIFT	Predicts deleterious AASs) based on conservation and physico-chemical principles	Downloadable software and web server	http://sift.bii.a-star.edu.sg	[35]
FATHMM	Uses a hidden Markov model to identify AASs likely to be deleterious	Downloadable software and web server, VEP plugin	http://fathmm.biocompute.org.uk	[39]

Only few more tools…

Tool	Application	Comments	URL	Reference
Integrative approaches using supervised learning algorithms				
PolyPhen	Predicts deleterious AASs based on several sequence and structural features	Downloadable software and web server, pre-computed predictions for all possible substitutions	http://genetics.bwh.harvard.edu/pph2/	[41]
MutationTaster	Classifier which can predict deleterious variants in genic regions, including coding regions and splice sites	Web server	http://www.mutationtaster.org	[42]
MutationAssessor	Predicts deleterious AASs based on evolutionary conservation	Web server, pre-computed scores for all possible substitutions	http://www.mutationassessor.org	[43]
SNAP	Predicts deleterious AASs based on a range of protein level information	Downloadable software and web server	http://www.rostlab.org/services/SNAP/	[44]
PhD-SNP	Predicts deleterious AASs based on protein sequence information	Downloadable software and web server	http://snps.biofold.org/phd-snp/	[45]
Condel	Tool that integrates predictions from multiple AAS prediction tools	Downloadable software and web server, VEP plugin	http://bg.upf.edu/fannsdb/	[46]
CAROL	Tool that integrates scores from SIFT and PolyPhen using a weighted Z method	Downloadable R script, VEP plugin	http://www.sanger.ac.uk/resources/software/carol/	[47]
GWAVA	Classifier identifying likely functional regulatory variants	Downloadable software and database of pre-computed scores and annotations for known variants, VEP plugin	http://www.sanger.ac.uk/resources/software/gwava/	[48]
CADD	Integrated classifier that can score all classes of variants	Web server, pre-computed scores for all possible SNVs, VEP plugin	http://cadd.gs.washington.edu	[51]
Phenotype association techniques that can incorporate functional information				
fgwas	Command line tool for incorporating functional information into a GWAS	Downloadable software	http://www.github.com/joepickrell/fgwas	[52]
SKAT	A test for association between a set of variants and dichotomous or quantitative phenotypes	Downloadable software	http://www.hsph.harvard.edu/skat/	[53]
VT	Tests for pooled association of multiple rare variants and phenotypes	Downloadable software	http://genetics.bwh.harvard.edu/vt/dokuwiki/start	[54]
VAAST	Probabilistic tool to identify causal genes and variants in disease	Downloadable software, free for academic use, license required for commercial usage	http://www.yandell-lab.org/software/vaast.html	[55,56]

And many others!!!

I finished my analysis, now what?

Some web sites offer free, user-friendly analysis. A partial list of websites that accept transcript IDs and/or gene IDs:

- DAVID (http://david.abcc.ncifcrf.gov/tools.jsp)
- ConsensusPathDB (http://cpdb.molgen.mpg.de/)
- NetGestalt (http://www.netgestalt.org/)
- Molecular Signatures Database (http://www.broadinstitute.org/gsea/msigdb/index.jsp)
- PANTHER (http://www.pantherdb.org/)
- Cognoscente (http://vanburenlab.medicine.tamhsc.edu/cognoscente.shtml)
- Pathway Commons (http://www.pathwaycommons.org/)
- Reactome (http://www.reactome.org/)
- PathVisio (http://www.pathvisio.org/)
- Moksiskaan (http://csbi.ltdk.helsinki.fi/moksiskaan/)

There are free solutions that require some programming knowledge, including several packages in Bioconductor (http://bioconductor.org/).

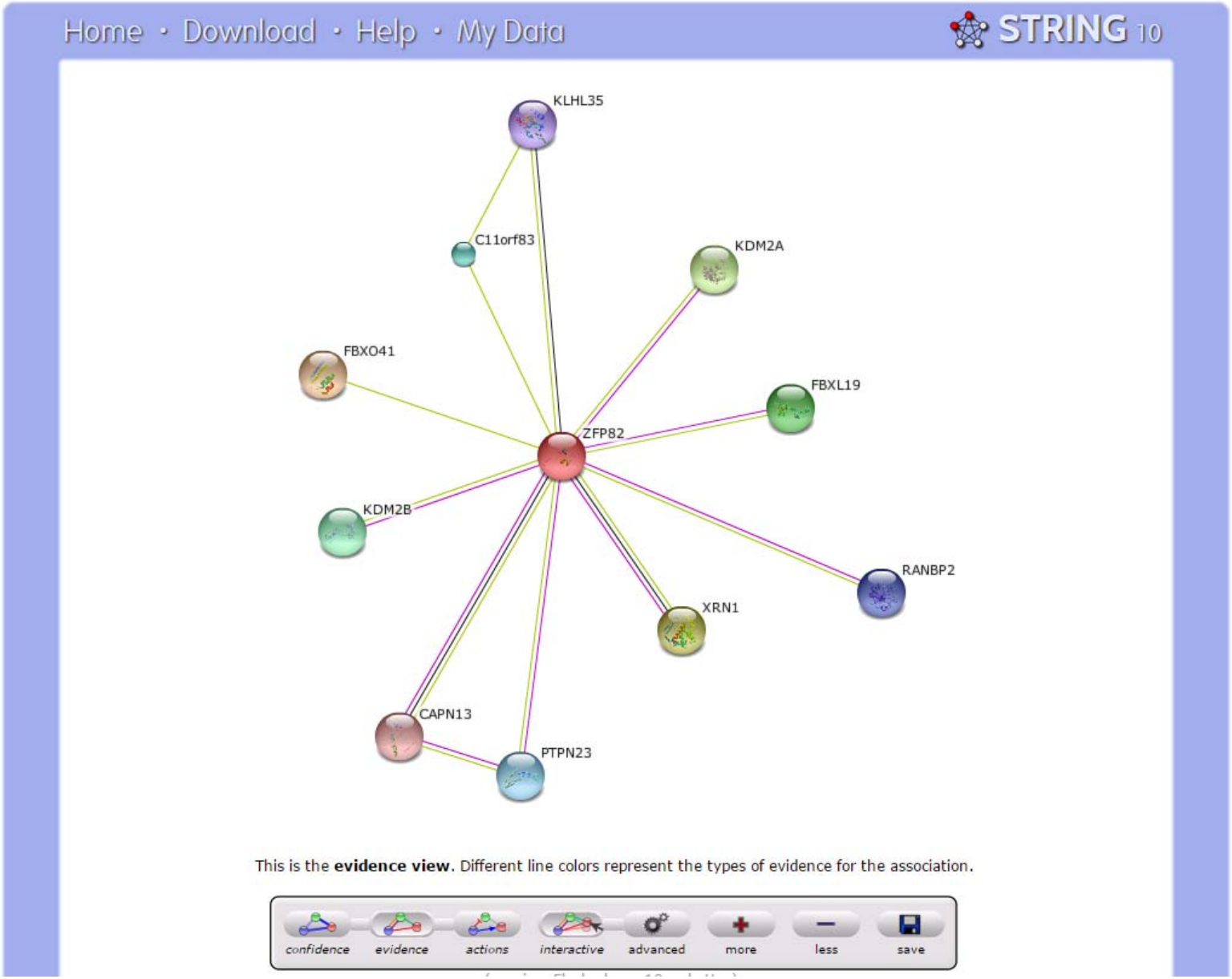
Commercial solutions include:

- Ingenuity (http://www.ingenuity.com/products/ipa)
- Advaita iPathwayGuide (https://apps.advaitabio.com/ipg/home)
- Metacore (http://lsresearch.thomsonreuters.com/)

STRING

(http://string-db.org/)

Known and Predicted Protein-Protein Interactions



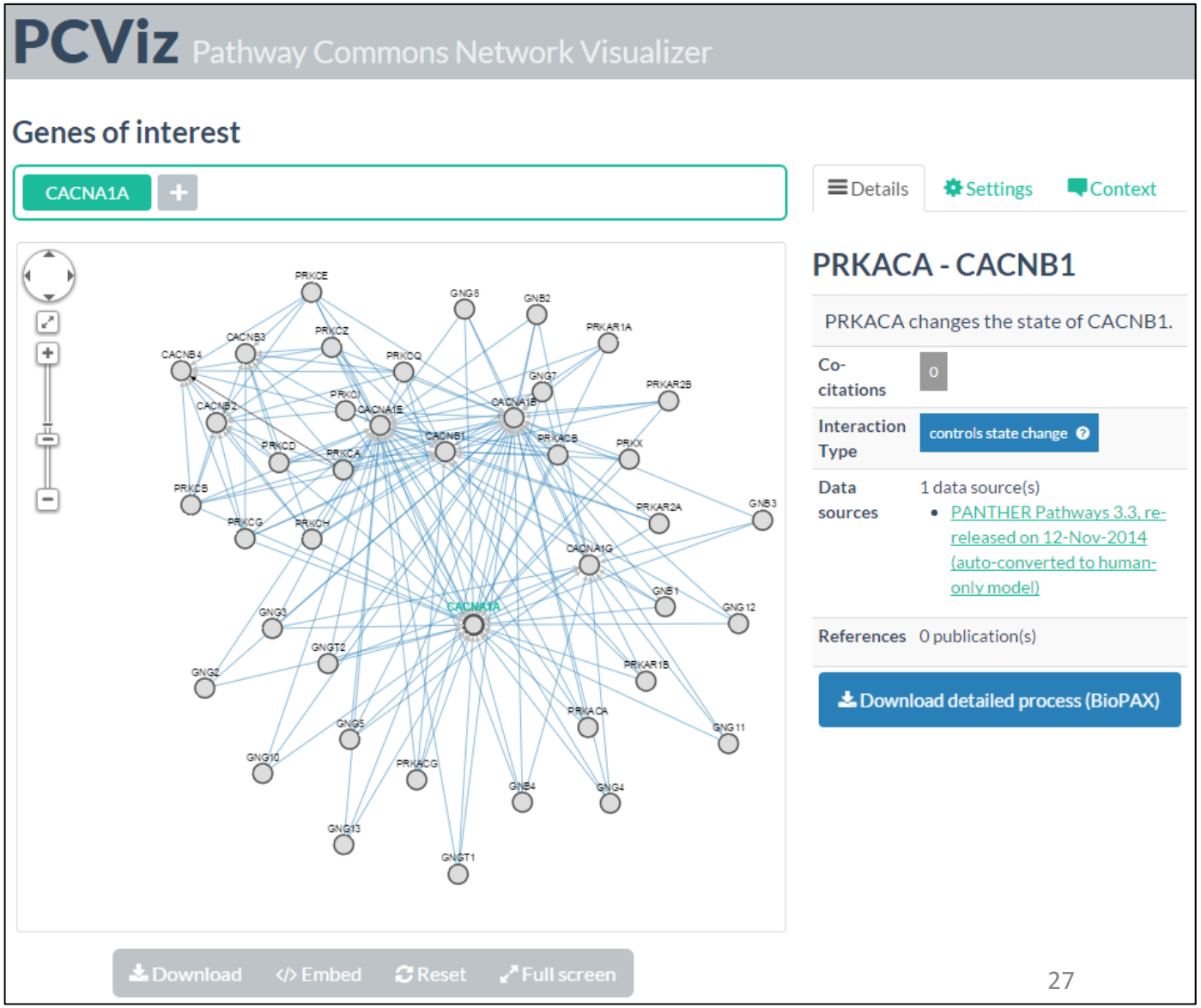
Pathway Commons

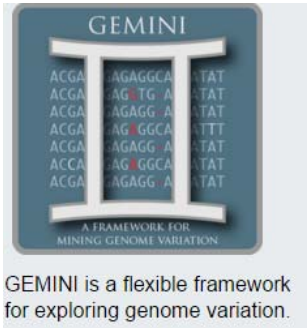
(http://www.pathwaycommons.org/)

A SNP may interfere with a gene whose function is unknown, but it is associated with a well studied gene.

Important questions:

1. What is the nature of that "association?"
2. Is there a biological confirmation for the association?
3. What type of "pathway" is this?
4. Who defined the "pathway" and was it validated?





GEMINI is a flexible framework for exploring genome variation.

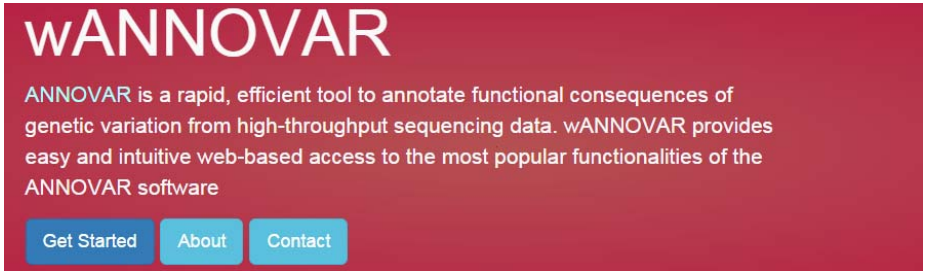
GEMINI

https://gemini.readthedocs.org/en/latest/

- GEMINI (GEnome MINIng) - a flexible framework for exploring genetic variation in the context of the wealth of genome annotations available for the human genome.

- GEMINI provides a simple, flexible, and powerful system for exploring genetic variation for disease and population genetics.

- Accepts VCF file. Each variant is automatically annotated by comparing it to several genome annotations from source such as ENCODE tracks, UCSC tracks, OMIM, dbSNP, KEGG, and HPRD. All of this information is stored in portable SQLite database that allows one to explore and interpret both coding and non-coding variation using “off-the-shelf” tools or an enhanced SQL engine.



http://wannovar.usc.edu/index.php

```
##fileformat=VCFv4.1
##contenttype=HumanOmni25M-8v1-1_B.bpm
##sourcereport=FinalReport_HumanOmni25M-8v1-1_PG0001217-BLD.txt
##genomemap=HumanOmni25M-8v1-1.NCBI37.map.txt
##contig=<ID=NA,length=0,Description="Contig undetermined-for genotyped alleles that do not map to the reference.">
##INFO=<ID=AL,Number=1,Type=String,Description="Array Alleles">
##INFO=<ID=ST,Number=1,Type=String,Description="ProbeStrand">
##FILTER=<ID=GTEX,Description="Genotype excluded from reference sequence mapping.">
##FILTER=<ID=NOCALL,Description="Genotype not called on array.">
##FORMAT=<ID=GC,Number=1,Type=Float,Description="GencallScore">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##workflow_type=Illumina_GenotypingToVCF
##workflow_version=v1.4
#CHROM  POS        ID          REF  ALT  QUAL  FILTER  INFO          FORMAT
                   WG0231131-DNAD01-PG0001217-BLD
chr7    117149177  rs75961395  A    G    .     PASS    AL=A/G;ST=-   GT:GC
                   0/1:0.9120
```

Results

Chr	Start	End	Ref	Alt	Func	Gene	GeneDetail	ExonicFunc	AAChange
chr7	117149177	117149177	A	G	exonic	CFTR		synonymous SNV	CFTR:NM_000492:exon3:c.G254G:p.G85G

Otherinfo	Otherinfo	Otherinfo	Otherinfo	Otherinfo	Otherinfo	Otherinfo	Otherinfo	Otherinfo	Otherinfo	Otherinfo	Otherinfo	Otherinf
het	.	.	chr7	117149177	rs75961395	A	G	.	PASS	AL=A/G;ST=-	GT:GC	0/1:0.91

Exercise

1. Use wANNOVAR to annotate the following variants.

rs121908745	rs78655421	rs77932196
Rs333	rs2572886	rs9264942
rs2745557	rs1143674	rs1445442
rs2075575	rs4800773	rs2075575
rs13186740	rs2042959	rs4867387
rs104886271	rs80359806	rs80359796
rs852287	rs2337193	rs5832208
rs4481887	rs2153271	rs1042725

2. Examine your annotation and try to interpret the values provided by each predictor.

3. How will you decide which variants are causal?

4. Which genes harbor these variants? To which pathways they belong to?