

# Variant annotation



- ❖ Introduction to variant annotation
- ❖ *Odds Ratio* and *Relative Risks*
- ❖ Finding the mutations causative of diseases
- ❖ Annotation tools (ANNOVAR, wANNOVAR, Variant Effect Predictor)
- ❖ Understanding the annotation and the ClinVar database
- ❖ Population structure
- ❖ Loss-of-function variants

**TRUMP** 



# Learn from the Master

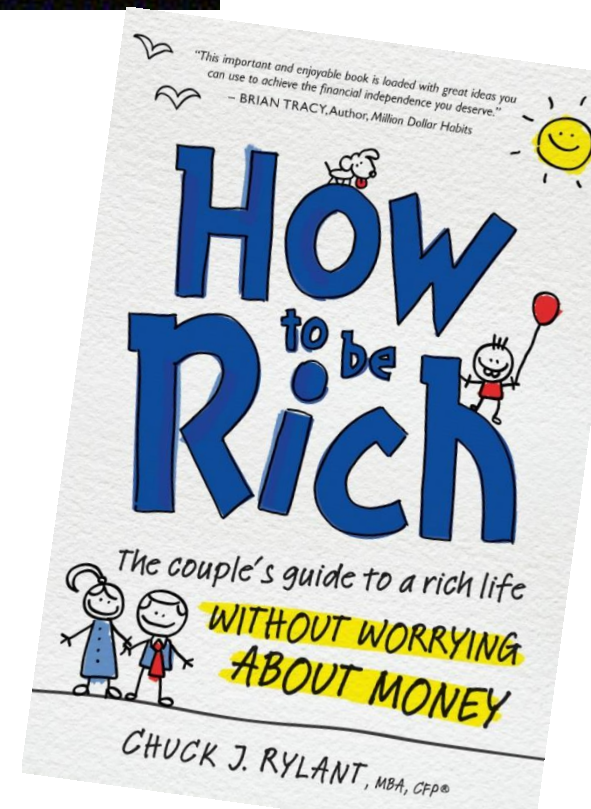
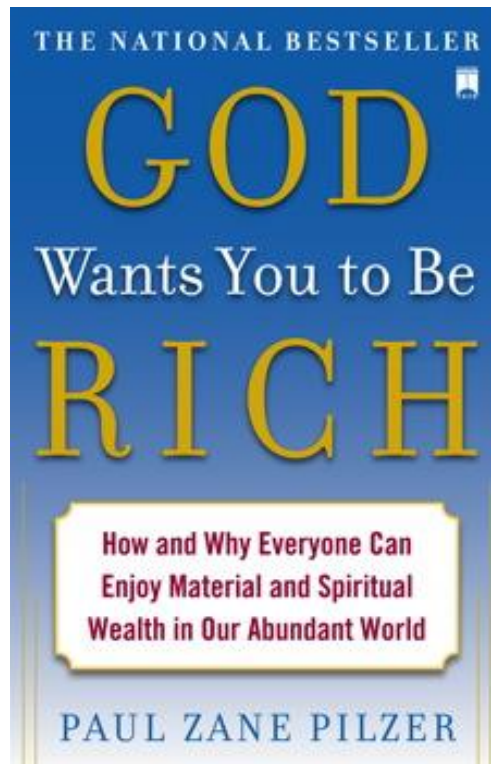
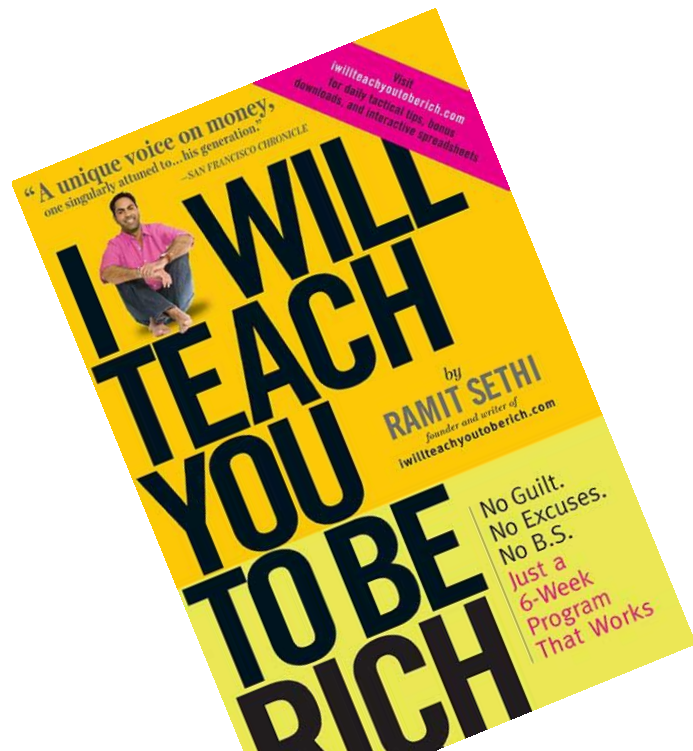
*"I can turn anyone into a successful real estate investor."*

  
 Donald J. Trump  
 Chairman, Trump University

**COMING TO YOUR AREA**  
**FREE**  
 INTRODUCTORY CLASS

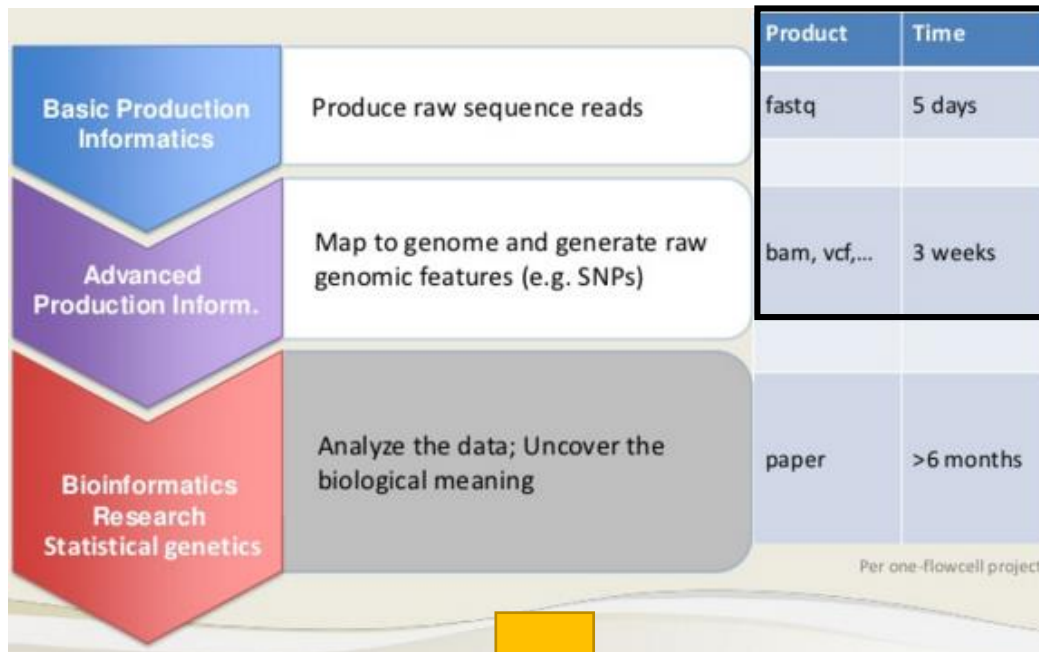
Phoenix  
 Jan 10 - 13

**Attend a FREE Real Estate Investing Workshop. Space is Limited.**



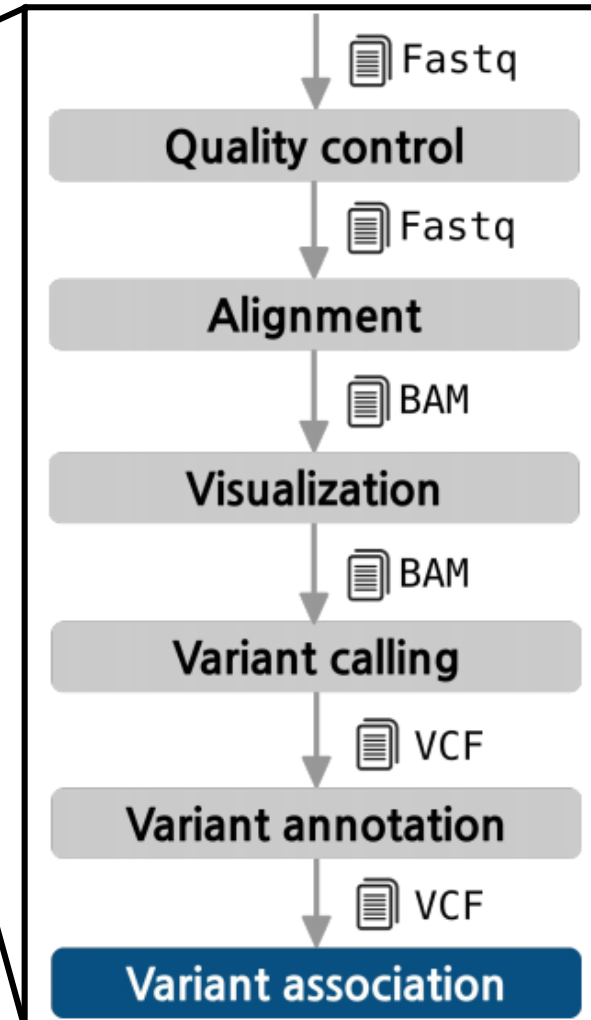


# The pipeline so far



532 variants are significantly associated with the disease with high odds ratio (OR).

**Now what ?**



# **Odds Ratios and Relative Risk**

# Odds and probabilities

Odds For Red



Odds can have any value from zero to infinity and they represent a ratio of desired outcomes versus the field.

Probability is defined as the fraction of desired outcomes in the context of every possible outcome with a value between 0 and 1.

Probability of Red





# Odds Ratios

An **odds ratio** (OR) is a measure of association between an exposure and an outcome. OR represents the **odds** that an outcome will occur given a particular exposure, compared to the **odds** of the outcome occurring in the absence of that exposure.

ORs are used to compare the relative odds of the occurrence of the outcome of interest (e.g., disease or disorder), given exposure to the variable of interest (e.g., health characteristic, aspect of medical history).

The odds ratio can also be used to determine whether a particular exposure is a risk factor for a particular outcome, and to compare the magnitude of various risk factors for that outcome.

- OR=1 Exposure does not affect odds of outcome
- OR>1 Exposure associated with higher odds of outcome
- OR<1 Exposure associated with lower odds of outcome

	Event (Fail)	No Event (Don't Fail)
Treatment (Tutoring)	a	b
Control (No Tutoring)	c	d

$$OR = \frac{\text{Odds of event in Treatment group}}{\text{Odds of event in Control group}} = \frac{a/b}{c/d} = ad/bc$$

# Let's calculate some ORs

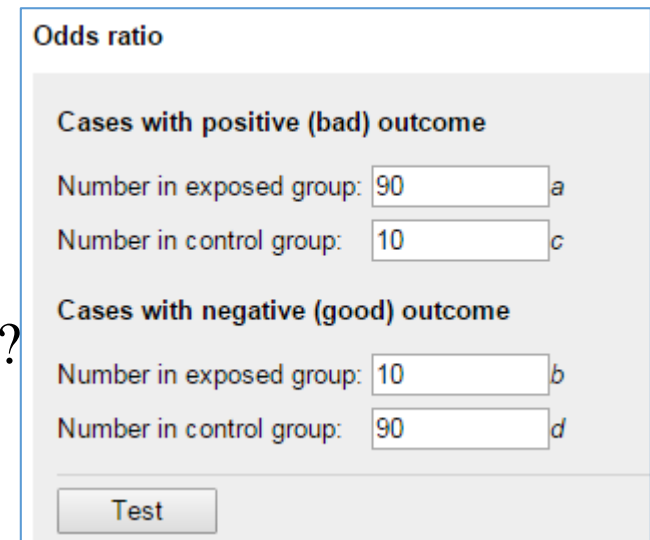
Online calculator: [https://www.medcalc.org/calc/odds\\_ratio.php](https://www.medcalc.org/calc/odds_ratio.php)

You hold a case-control study.

You recruited 100 patients and 100 matching controls.

If 90 of your cases and 10 of your controls have the alternative allele, what are the odds ratio?

Now assume that only 11 cases and 10 controls show the alternative allele, what are the odds ratio?



The screenshot shows a web-based calculator titled "Odds ratio". It is divided into two main sections. The first section, "Cases with positive (bad) outcome", contains two input fields: "Number in exposed group:" with the value "90" and a label "a", and "Number in control group:" with the value "10" and a label "c". The second section, "Cases with negative (good) outcome", also contains two input fields: "Number in exposed group:" with the value "10" and a label "b", and "Number in control group:" with the value "90" and a label "d". At the bottom of the form is a button labeled "Test".

Guidelines: oftentimes you will encounter statistically significant findings with OR that range between 1 and 1.5.  $OR < 2$  can be easily explained due to population structure. Be skeptical of OR smaller than 2.5-3.



# A disease allele?

To calculate overall risk for **breast cancer**, Johnson et al. (2007) counted the total alleles that had relative frequencies (in their control population) of less than 10%.

Counting potentially functional variants in <i>BRCA1</i> , <i>BRCA2</i> and <i>ATM</i> predicts breast cancer susceptibility						
Table 2. ORs for two primary breast cancers corresponding to increasing numbers of variant alleles for 25 SNPs in <i>BRCA1</i> , <i>BRCA2</i> , <i>ATM</i> , <i>CHEK2</i> and <i>TP53</i>						
No of variant alleles	All SNPs			Uncommon (MAF < 10%) SNPs		
	Controls <i>N</i> (%)	Cases <i>N</i> (%)	OR (95% CI)	Controls <i>N</i> (%)	Cases <i>N</i> (%)	OR (95% CI)
0	193 (7.91)	24 (5.15)	1.00 (Ref)	1373 (56.27)	226 (48.50)	1.00 (Ref)
1	369 (15.12)	67 (14.38)	1.46 (0.89–2.40)	794 (32.54)	162 (34.76)	1.24 (0.99–1.54)
2	493 (20.20)	85 (18.24)	1.39 (0.86–2.25)	229 (9.39)	57 (12.23)	1.51 (1.10–2.09)
3	516 (21.15)	112 (24.03)	1.75 (1.09–2.80)	35 (1.43)	19 (4.08)	2.90 (1.69–4.97)
4	392 (16.07)	76 (16.31)	1.56 (0.95–2.55)	7 (0.29)	1 (0.21)	
5	252 (10.33)	41 (8.80)	1.31 (0.76–2.24)	2 (0.08)	1 (0.21)	
6	140 (5.74)	32 (6.87)	1.84 (1.04–3.26)			
7	65 (2.66)	17 (3.65)	2.10 (1.06–4.16)			
8	16 (0.66)	8 (1.72)	4.02 (1.56–10.38)			
9+	4 (0.16)	4 (0.86)	8.04 (1.89–34.26)			
Total	2440 (100)	466 (100)	1.08 (1.02–1.14) per SNP <i>P</i> (trend) = 0.005	2440 (100)	466 (100)	1.30 (1.15–1.47) per SNP <i>P</i> (trend) = 0.00004

# Relative Risk

- RR is another way of measuring the effect of treatment.
- OR and RR are oftentimes used interchangeably, but they mean different things.
- OR is a ratio of two odds (obvious, right?)
- RR is a ratio of two probabilities.
- $RR=1$ , the treatment made no difference.  $RR>1$  the treatment has higher risk of failing compared to controls.  $RR<1$  the treatment has lower risk of failing compared to controls.

The probability of an event in the Treatment group is  $a/(a+b) = R1$   
The probability of an event in the Control group is  $c/(c+d) = R2$   
 $RR = R1/R2$

Ratio of event to non-events

a measure of events out of all possible events

	Event (Fail)	No Event (Don't Fail)
Treatment (Tutoring)	a	b
Control (No Tutoring)	c	d

$$OR = \frac{\text{Odds of event in Treatment group}}{\text{Odds of event in Control group}} = \frac{a/b}{c/d} = ad/bc$$

$$RR = \frac{\text{Risk of event in the Treatment group}}{\text{Risk of event in the Control group}} = \frac{a/(a+b)}{c/(c+d)}$$

# Let's calculate some RRs

Calculator: [https://www.medcalc.org/calc/relative\\_risk.php](https://www.medcalc.org/calc/relative_risk.php)

What is the **probability** of developing cancer among people with a mutation in BRAC2?

20% of cancer patients were carriers of the mutation.  
1% of cancer patients were not carriers of the mutation.

Here,  $a = 20$ ,  $b = 80$ ,  $c = 1$ , and  $d = 99$ .

What is the relative risk of cancer associated with this mutation?

BRCA2+

BRCA2-

<b>Exposed group</b>	
Number with positive (bad) outcome:	<input type="text" value="20"/> <i>a</i>
Number with negative (good) outcome:	<input type="text" value="80"/> <i>b</i>
<b>Control group</b>	
Number with positive (bad) outcome:	<input type="text" value="1"/> <i>c</i>
Number with negative (good) outcome:	<input type="text" value="99"/> <i>d</i>
<input type="button" value="Test"/>	

## Results

Relative risk	20.0000
95% CI	2.7361 to 146.1912
z statistic	2.952
Significance level	P = 0.0032
NNT (Harm)	5.263
95% CI	9.157 (Harm) to 3.693 (Harm)

# Exercise I

Answer the following questions:

1. Which of the following depends on the samples size: OR ,RR, or the *p-value* (explore with sample sizes of  $10^1$ ,  $10^2$ ,  $10^3$ )?
2. The table below shows how the risk ratio was calculated in the study examining the risk of wound infections when an incidental appendectomy was done during a staging laparotomy for Hodgkin disease. Calculate the RR and interpret the results.

Had Incidental Appendectomy?	Wound Infection	No Wound Infection	Total
Yes	7	124	131
No	1	78	79

3. A cohort study examined the association between smoking and lung cancer after following 400 smokers and 600 non-smokers for 15 years. At the conclusion of the study the investigators found a risk ratio = 17. Which of the following would be the best interpretation of this risk ratio?
  - ☐ a. There were 17 more cases of lung cancer in the smokers.
  - ☐ b. Smokers had 17% more lung cancers compared to non-smokers.
  - ☐ c. Smokers had 17 times more risk of lung cancer than non-smokers.
  - ☐ d. Smokers had 17 times the risk of lung cancer compared to non-smokers.
  - ☐ e. 17% of the lung cancers in smokers were due to smoking.

# Still confused between OR and RR?

Read the following example:

<http://www.theanalysisfactor.com/the-difference-between-relative-risk-and-odds-ratios/>

## Remember:

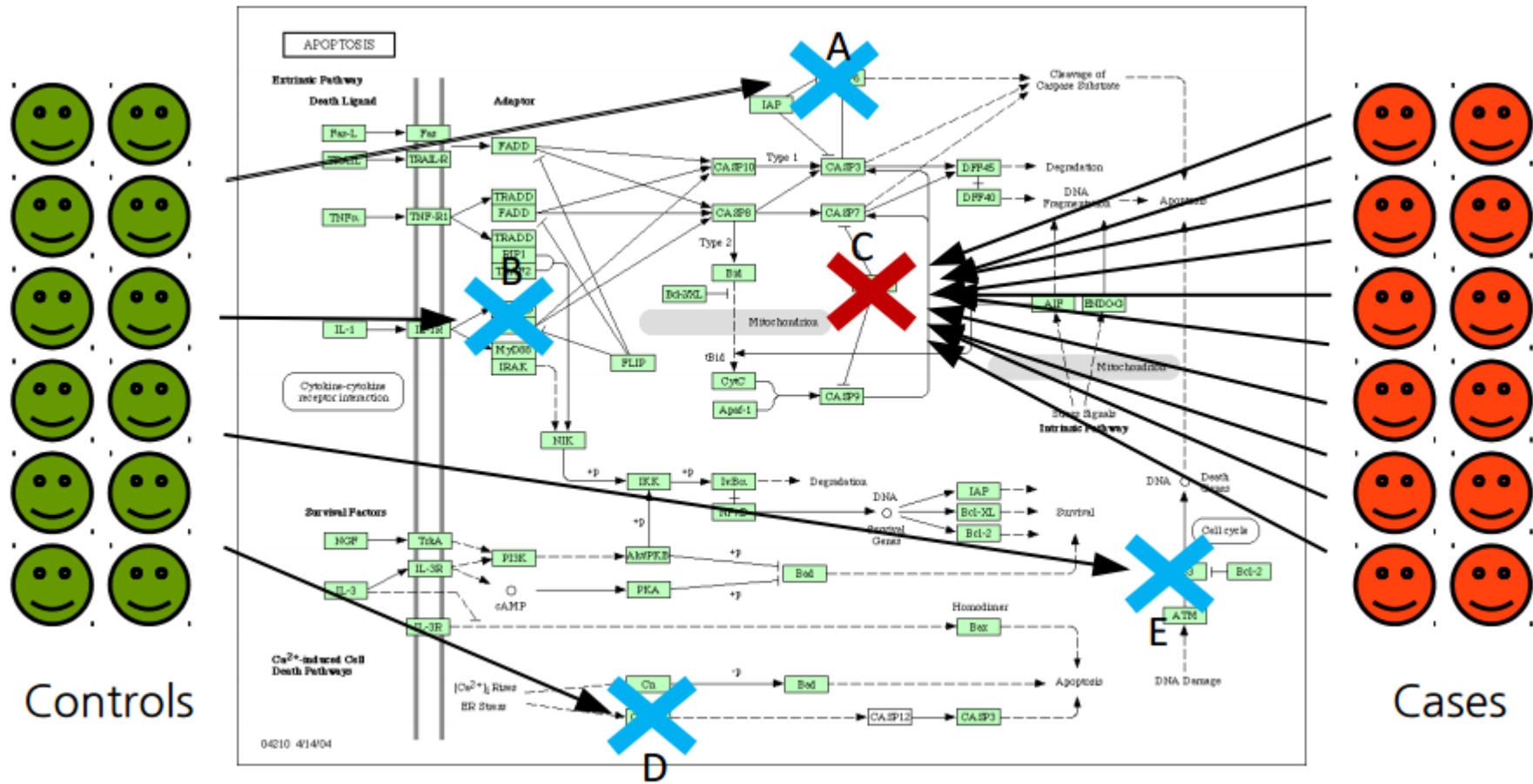
RR is the **probability** of an event occurring ( $a/a+b$ )/the probability of the event not occurring ( $c/c+d$ ).

OR is the **odds** of an even occurring ( $a/b$ )/the odds of the event not occurring ( $c/d$ ).

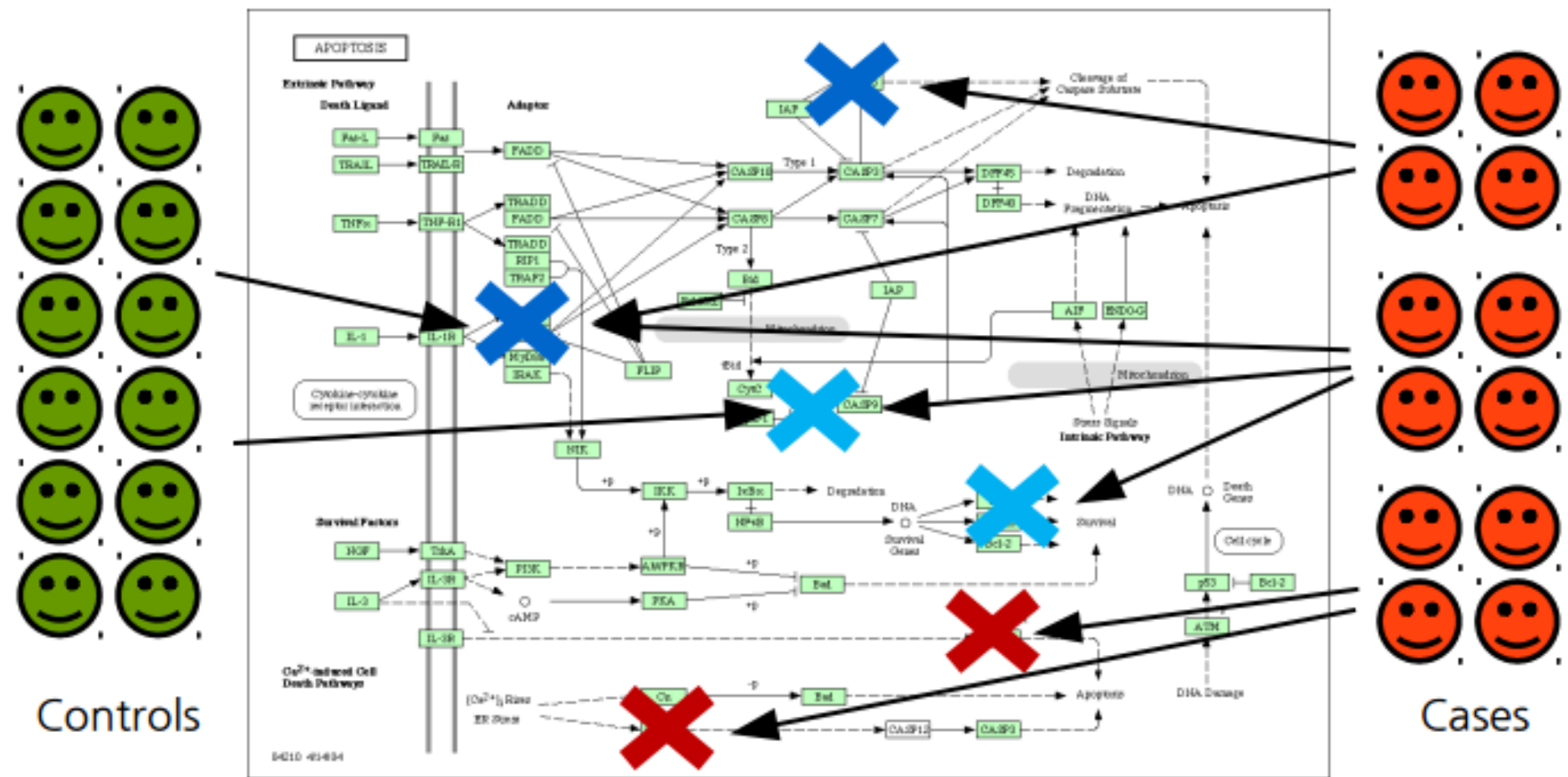
**What genomic mutations can cause a disease?**

# Finding the mutations causative of diseases

The simplest case: monogenic disease due to a single gene





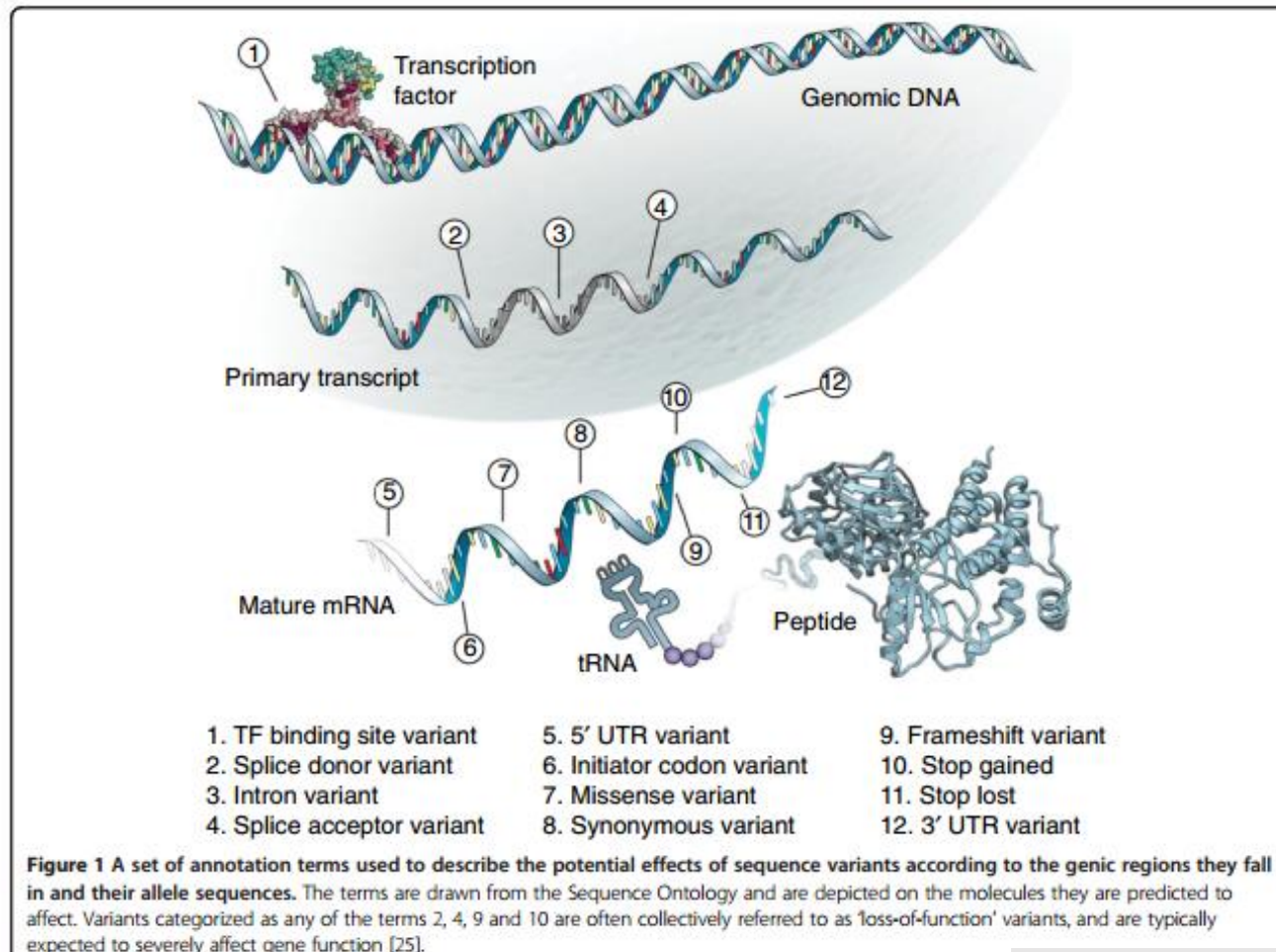


Clear individual gene associations are difficult to find in some diseases.

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
- Whole genome has up to 50 million variants on average
- After annotating there will be hundreds of deleterious variants



**Annotation tools  
provide useful data about the mutation**

# 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!**



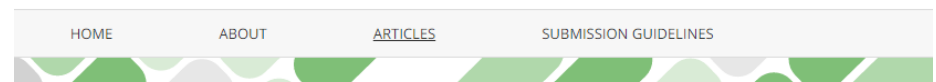
**How will you decide which of those is causal?**

# Annotating *de novo* mutations

VARiant PRIoritization SuM (*VARPRISM*) predicts the functional impact of *de novo* mutations and incorporates these quantitative predictions using a likelihood ratio test to evaluate evidence of *de novo* mutation load.

*VARPRISM* incorporates variant prioritization information to identify genes with a statistically significant excess of *de novo* mutations contributing to genetic diseases.

## Genome Medicine



[SOFTWARE](#) | [OPEN ACCESS](#)

### VARPRISM: incorporating variant prioritization in tests of *de novo* mutation association

[Hao Hu](#), [Hilary Coon](#), [Man Li](#), [Mark Yandell](#) and [Chad D. Huff](#) 

*Genome Medicine* 2016 8:91 | DOI: 10.1186/s13073-016-0341-9 | © The Author(s). 2016

Received: 1 March 2016 | Accepted: 2 August 2016 | Published: 25 August 2016



# 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 generate 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 uses > 400 categories to classify variants collected from different sources (see Variant annotation - dbNSFP3.1a.readme.txt in your Google Drive).
- ANNOVAR is implemented in GALAXY.

What are the problems with this type of information?

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

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

# WANNONAR

ANNONAR is a rapid, efficient tool to annotate functional consequences of genetic variation from high-throughput sequencing data. wANNONAR provides easy and intuitive web-based access to the most popular functionalities of the ANNONAR software

[Get Started](#) [About](#) [Contact](#)

<http://wannovar.wglab.org/>

```
##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
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	AACChange				
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	Otherinfo	Otherinfo
het	.	.	chr7	117149177	rs75961395	A	G	.	PASS	AL=A/G;ST=-	GT:GC	0/1:0.9120	



Search:  for

e.g. **BRCA2** or rat 5:62797383-63627669 or coronary heart disease

### Browse a Genome

The Ensembl project produces genome databases for vertebrates and other eukaryotic species, and makes this information freely available online.

### Popular genomes

Still using Human GRCh37?



Variant Effect Predictor



<http://www.ensembl.org/index.html>

**Variant Effect Predictor (VEP)**  
Works similarly to wANNOVAR,  
but faster

### Variant Effect Predictor ?

#### VEP for Human GRCh37

If you are looking for VEP for Human GRCh37, please go to [GRCh37 website](#).

Species:

Human (Homo sapiens)

Assembly: GRCh38.p5

Name for this data (optional):

Either paste data:

rs699  
rs144678492  
COSM354157

## What is Oncotator?

Oncotator is a web application for annotating human genomic point mutations and indels with data relevant to cancer researchers. Annotations are aggregated from the following resources:

### Genomic Annotations

- Gene, transcript, and functional consequence annotations using [GENCODE](#) for hg19.
- Reference sequence around a variant.
- GC content around a variant.
- Human DNA Repair Gene annotations from [Wood et al.](#)

### Protein Annotations

- Site-specific protein annotations from [UniProt](#).
- Functional impact predictions from [dbNSFP](#).

### Cancer Variant Annotations

- Observed cancer mutation frequency annotations from [COSMIC](#).
- Cancer gene and mutation annotations from the [Cancer GenCensus](#).
- Overlapping mutations from the [Cancer Cell Line Encyclopedia](#).
- Cancer gene annotations from the [Familial Cancer Database](#).
- Cancer variant annotations from [ClinVar](#).

### Non-Cancer Variant Annotations

- Common SNP annotations from [dbSNP](#).
- Variant annotations from [1000 Genomes](#).
- Variant annotations from [NHLBI GO Exome Sequencing Project \(ESP\)](#).

Please see the [help](#) page for detailed information regarding all annotations.

## Paste Mutations

```
7 55259515 55259515 T G
7 140453136 140453136 A T
1 120612003 120612004 GG -
8 145138175 145138176 - G
12 42538391 42538401 GGAGCGAGCAG -
```

Paste data in Oncotator [format](#) or use [example](#).

<http://portals.broadinstitute.org/oncotator/>

# Exploring genetic 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.

# Your toolkit



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

# 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

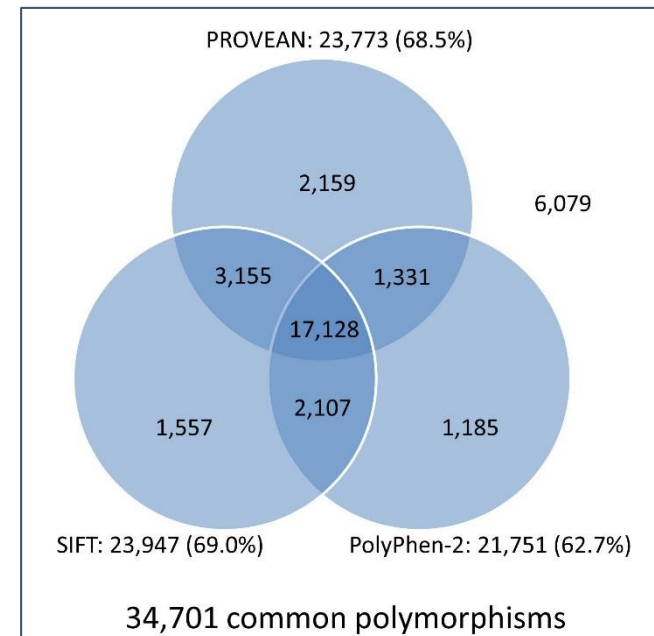
# Understanding the annotation

# 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





## 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. Moreover, relying on conservation alone can fail to capture human specific constraint.

Finally, different parts of proteins can have different functions, suggesting that it will ultimately be more informative to focus attention on functionally distinct portions of genes.

**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

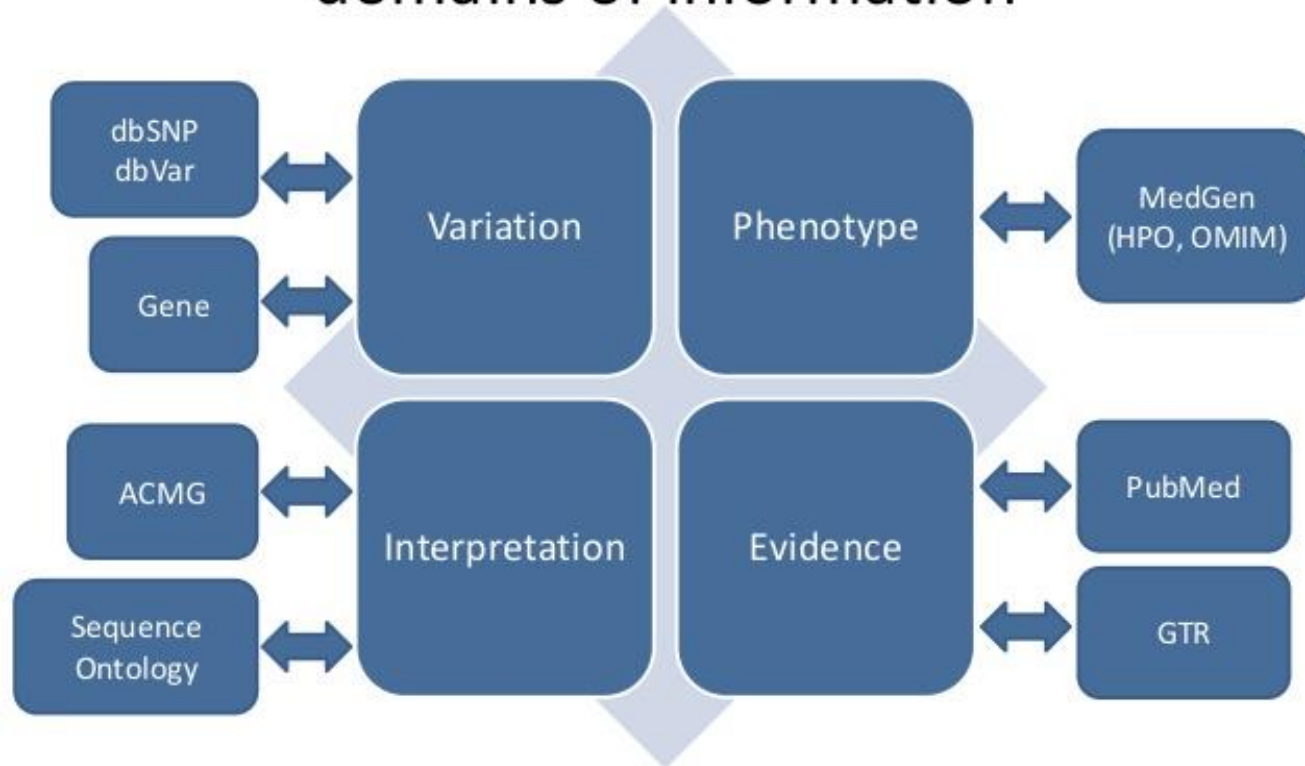
**Gene expression.** Variation in gene expressions in different tissues may imply on the functionality of the gene (e.g., GTEx).

ACTGATGGTATGGGGCCAAGAGATATATCT  
CAGGTACGGCTGTCATCACTTAGACCTCAC  
CAGGGCTGGGCATAAAAGTCAGGGCAGAGC  
CCATGGTGCATCTGACTCCTGAGGAGAAGT  
GCAGGTTGGTATCAAGGTTACAAGACAGGT  
GGCACTGACTCTCTCTGCCTATTGGTCTAT

## ClinVar

ClinVar aggregates information about genomic variation and its relationship to human health.

### ClinVar integrates four domains of information



Popular among clinicians...

# Exercise II

1. Use ClinVar to find how many pathogenic variants are in the *NFKB1* gene? How reliable is this information?
2. The CHD8 gene is known to be associated with autism. Annotate the CHD8 SNP Rs192989929 using any annotation tool. What is your diagnostic?
3. Use any annotation tool to annotate the following variants. What is your diagnostic?

rs121908745	rs78655421	rs77932196
	rs2572886	rs9264942
rs2745557	rs1143674	rs1445442
rs2075575	rs4800773	rs2075575
rs13186740	rs2042959	rs4867387
rs104886271	rs80359806	rs80359796
rs4481887	rs2153271	rs1042725
4. How will you decide which variants are causal?
5. Which genes harbor these variants? To which pathways they belong to?

# 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/>)

# Answers to Exercise I

1) Using the calculators we can calculate OR and RR. The  $p$ -value is the only variable that depends on the sample size.

Cases with positive (bad) outcome

Number in exposed group:  *a*

Number in control group:  *c*

Cases with negative (good) outcome

Number in exposed group:  *b*

Number in control group:  *d*

Test

Results

Odds ratio	1.8750
95 % CI:	0.2036 to 17.2702
z statistic	0.555
Significance level	P = 0.5790

Cases with positive (bad) outcome

Number in exposed group:  *a*

Number in control group:  *c*

Cases with negative (good) outcome

Number in exposed group:  *b*

Number in control group:  *d*

Test

Results

Odds ratio	2.5000
95 % CI:	0.2528 to 24.7202
z statistic	0.784
Significance level	P = 0.4332

Cases with positive (bad) outcome

Number in exposed group:  *a*

Number in control group:  *c*

Cases with negative (good) outcome

Number in exposed group:  *b*

Number in control group:  *d*

Test

Results

Odds ratio	2.5000
95 % CI:	1.9881 to 3.1438
z statistic	7.838
Significance level	P < 0.0001

2) In this study patients who underwent incidental appendectomy had 4.2 times the risk of post-operative wound infection compared to patients who did not undergo incidental appendectomy.

3) d.

35