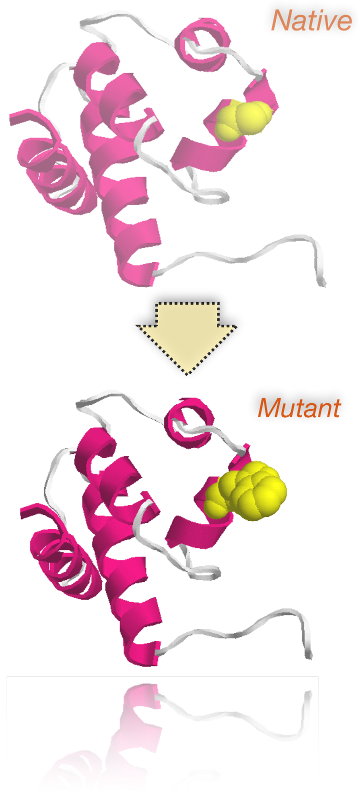
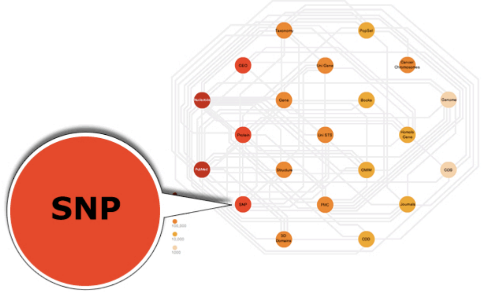
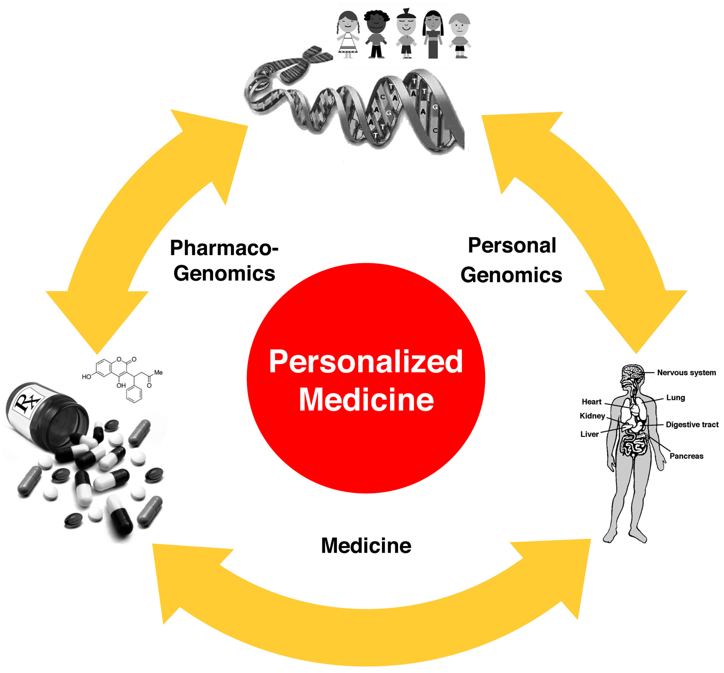
Detection, annotation and interpretation of short variants



Centro di Riferimento Oncologico
Aviano (Italy), September 6, 2017



OMIM
Online Mendelian Inheritance in Man
Johns Hopkins University



Emidio Capriotti
http://biofold.org/

Biomolecules Folding and Disease

Department of Biological, Geological,
and Environmental Sciences (BiGeA)
University of Bologna

Presentation outline

- Variation data resources:

  dbSNP, ClinVar, 1000Gemones

- Short variant detection:

  Matching reference genome. Variant calling procedures

- Variations in Cancer

  Cancer data resources, gene and variant classification

- Short variant annotation and interpretation:

  Annotation and prediction methods

Why genetic variants?

Genetic variation is fundamental to the evolution of all species and is what makes us individuals.

- To study the differences within and between populations, to understand the mechanism of adaptation, speciation, and the population structure.

- To characterize the relationship between genotype and phenotype.

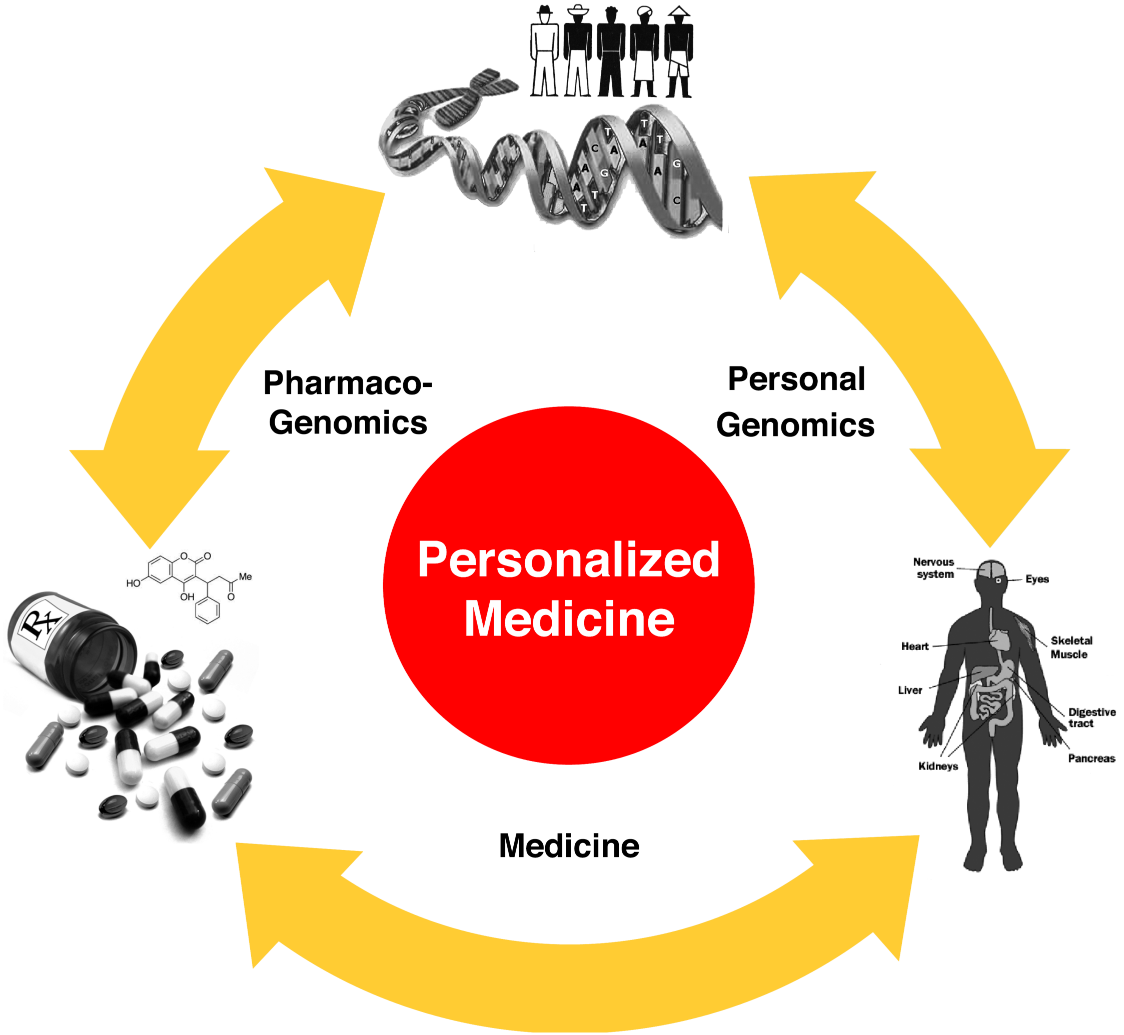
- Design new diagnostic protocols and therapeutic strategies.

Personalized medicine

Genotype test and exam sequencing, is cheap, and soon full genome sequencing cost will drop to $1000.

The future bioinformatics challenges for personalized medicine will be:

1. Processing Large-Scale Robust Genomic Data
2. Interpretation of the Functional Effect and the Impact of Genomic Variation
3. Integrating Systems and Data to Capture Complexity
4. Making it all clinically relevant



Fernald GH, et al (2011). Bioinformatics. 27: 1741-1748.

Single Nucleotide Variants

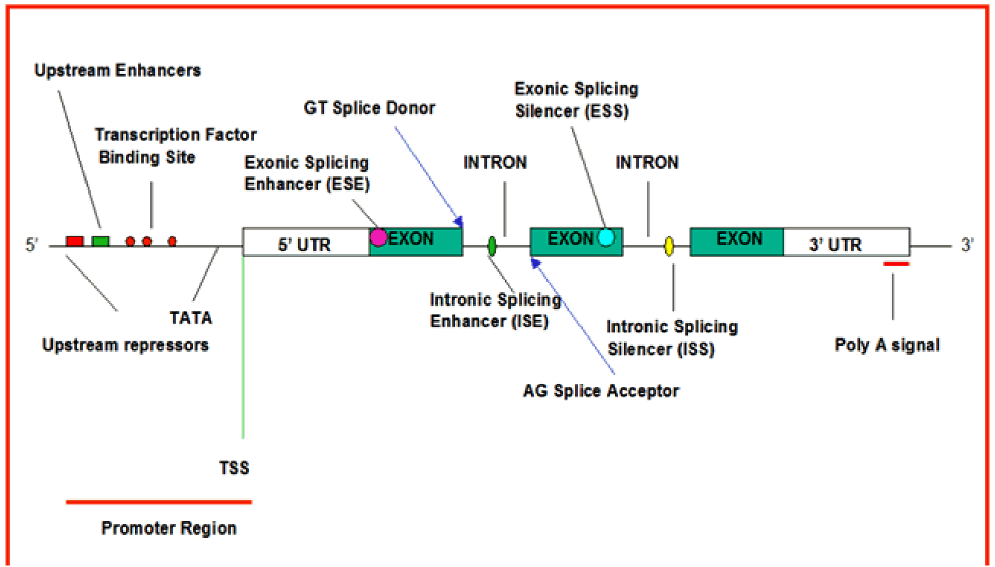
Single Nucleotide Variants (SNVs)
is a DNA sequence variation occurring when a single nucleotide A, T, C, or G in the genome differs between members of the species.
It is used to refer to Polymorphisms when the population frequency is ≥ 1%

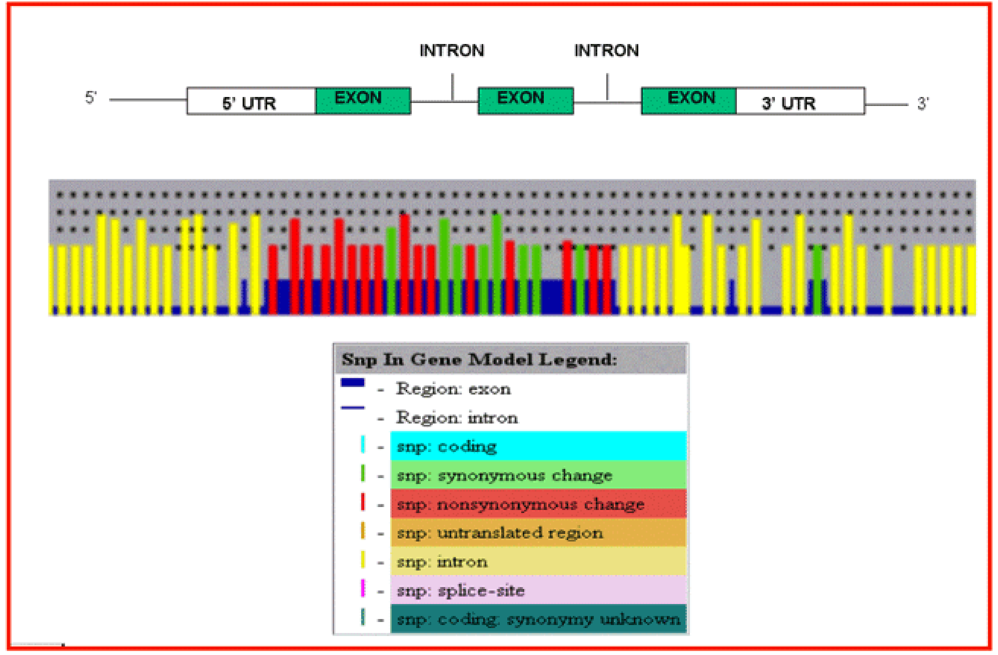
SNVs occur at any position and can be classified on the base of their locations.

Coding SNVs can be subdivided into two groups:

Synonymous: when single base substitutions do not cause a change in the resultant amino acid

Non-synonymous or Single Amino Acid Variants (SAVs): when single base substitutions cause a change in the resultant amino acid.

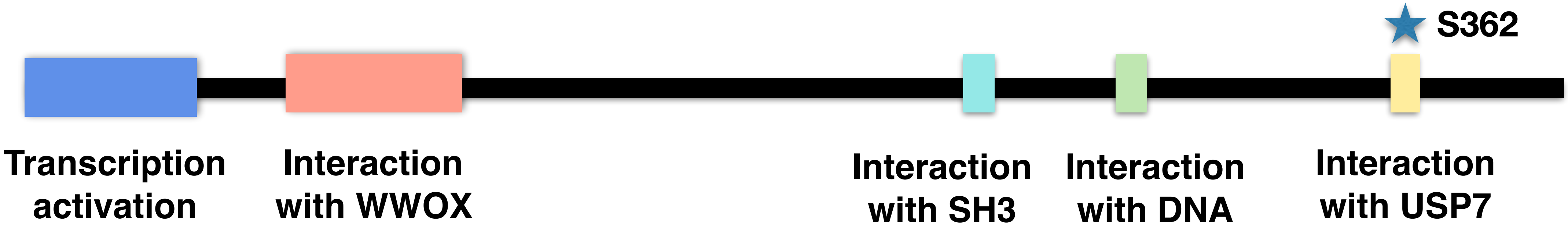

Upstream Enhancers
Transcription Factor Binding Site
GT Splice Donor
Exonic Splicing Silencer (ESS)
Exonic Splicing Enhancer (ESE)
INTRON
INTRON
5'
5' UTR
EXON
EXON
EXON
3' UTR
3'
TATA
Upstream repressors
Intronic Splicing Enhancer (ISE)
Intronic Splicing Silencer (ISS)
Poly A signal
AG Splice Acceptor
TSS
Promoter Region



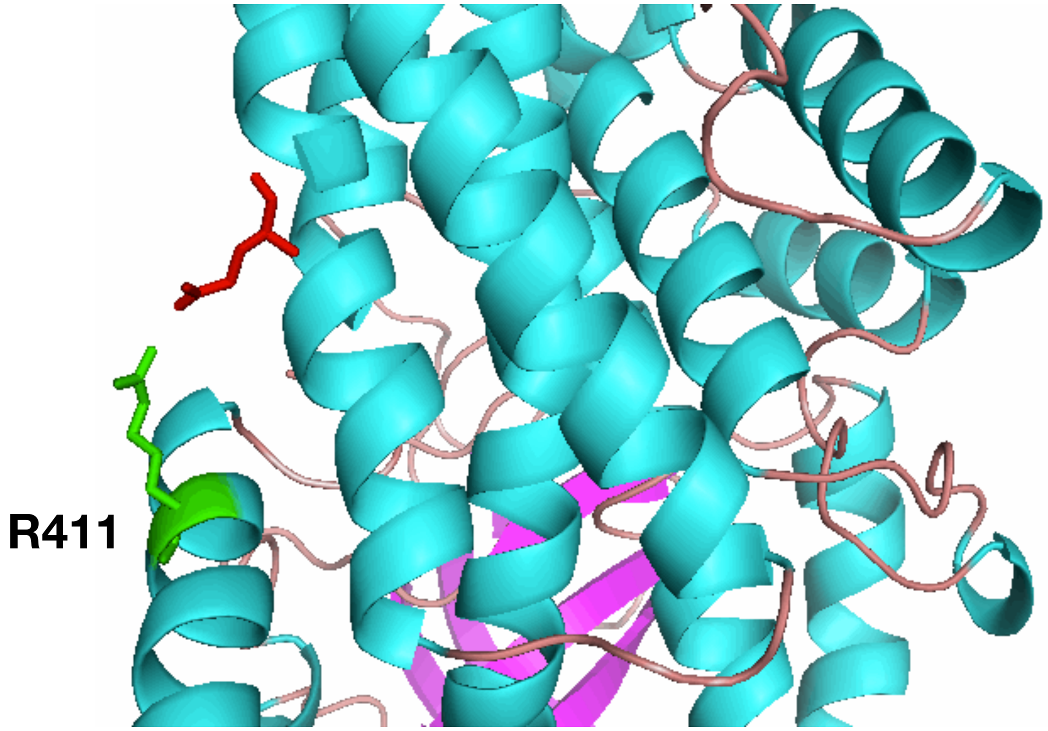
http://www.ncbi.nlm.nih.gov

Sequence, Structure & Function

Genomic variants in sequence motifs could affect protein function. Mutation S362A of P53 affect the interaction with hydrolase USP7 and the deubiquitination of the protein.



Nonsynonymous variants responsible for protein structural changes and cause loss of stability of the folded protein.
Mutation R411L removes the salt bridge stabilizing the structure of the IVD dehydrogenase.



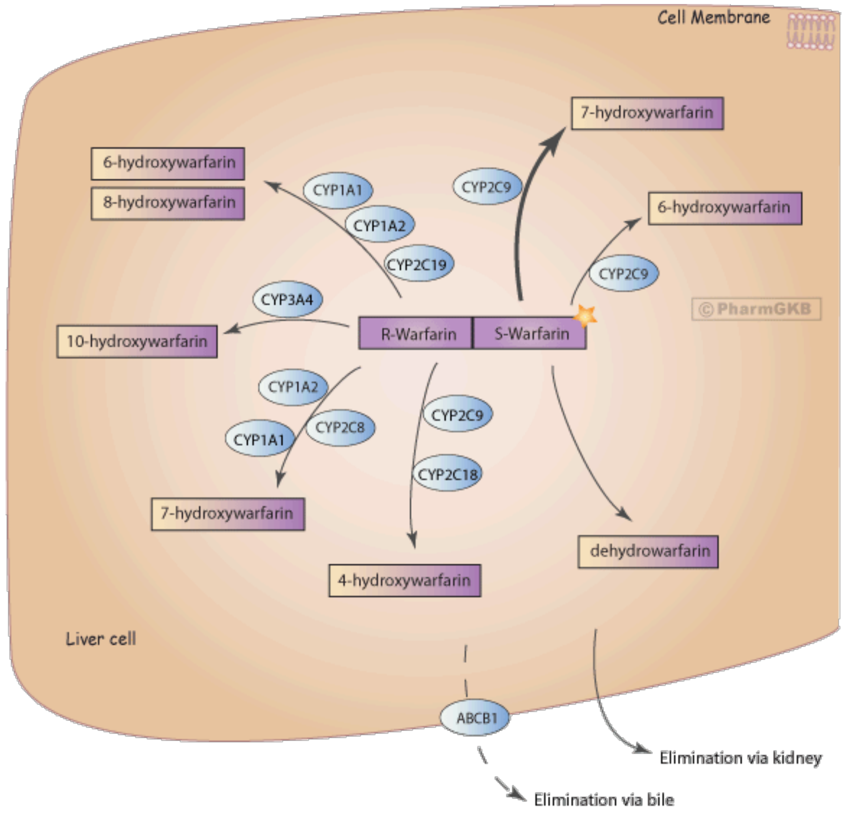
Variants and drug response

Pharmacogenomics aims at understanding how genetic variants influence drug efficacy and toxicity.

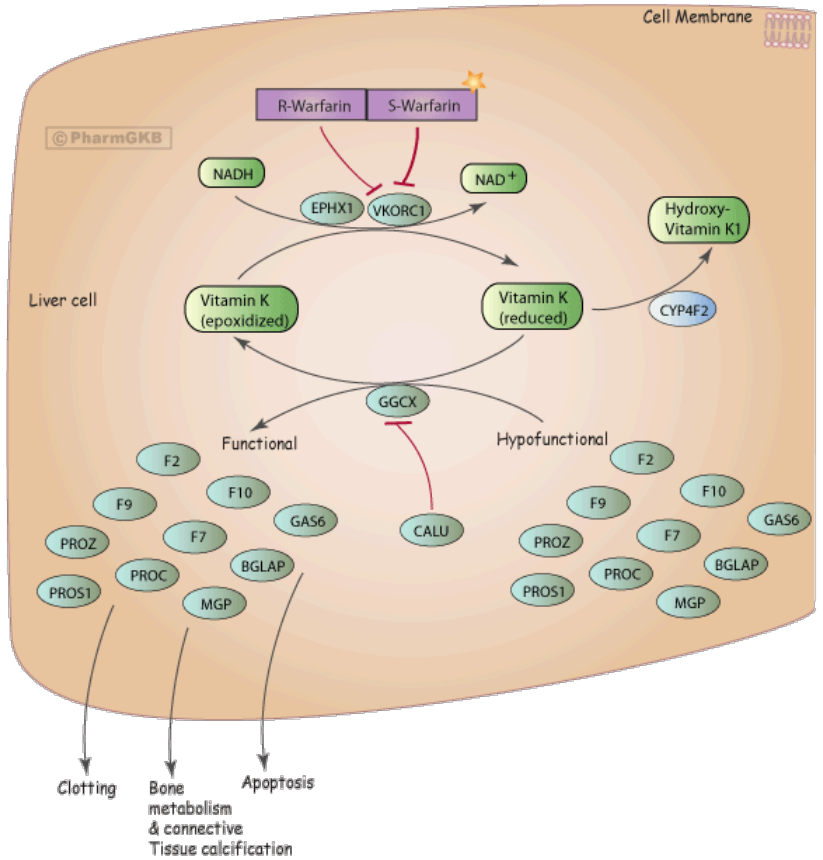
Pharmacokinetics variants: drug undergoes to bioinactivation via metabolic pathway. When the functionality of the pathway is compromised, a much higher concentrations of parent drug will accumulate.

Pharmacodynamics variants have an effect on the drug-receptor interactions and concentration. These variations have a directly impact on the dose-response relationships.

Warfarin and CYP2C9.



Warfarin and VKORC1

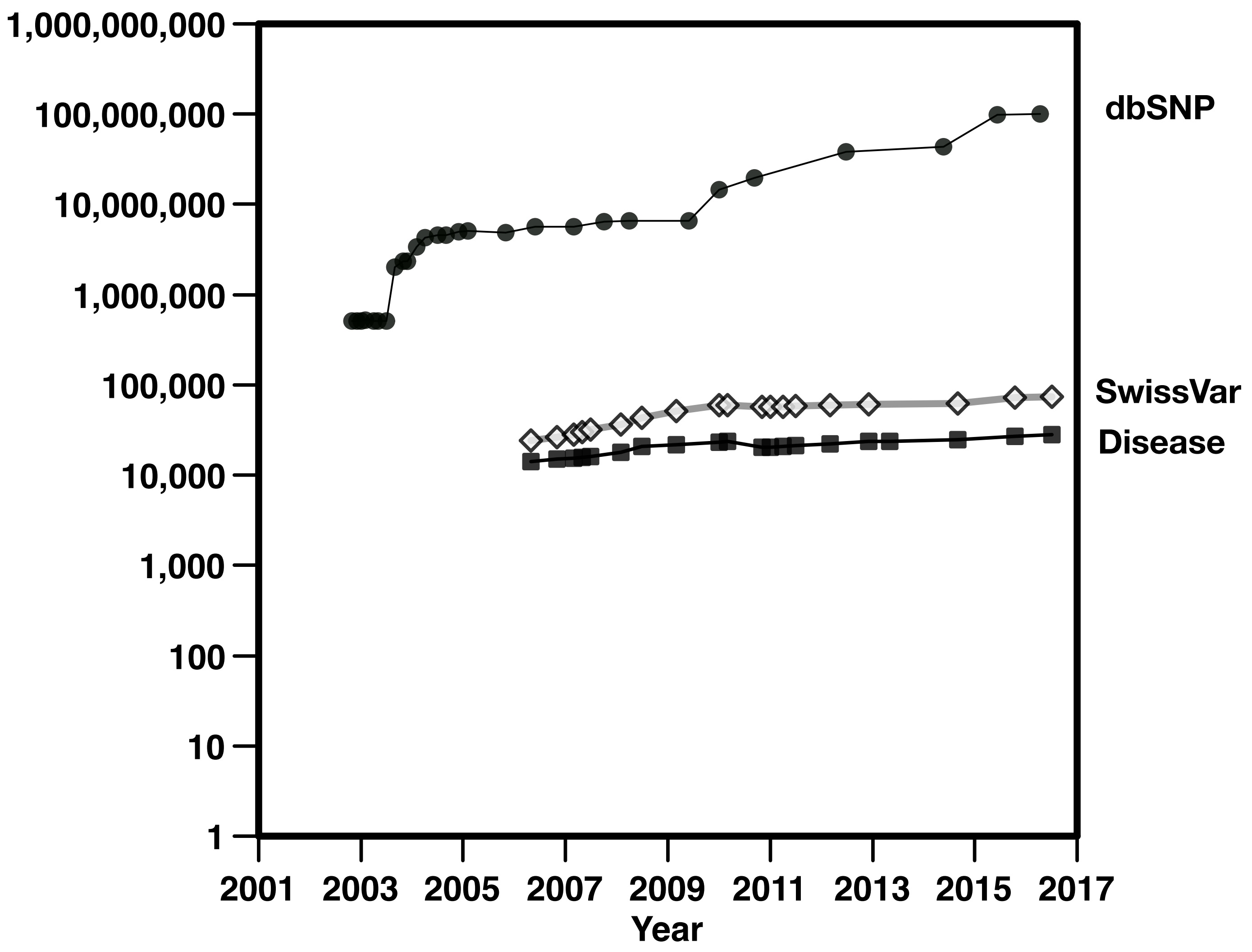


https://www.pharmgkb.org/

Variation data resources

Variant data growth

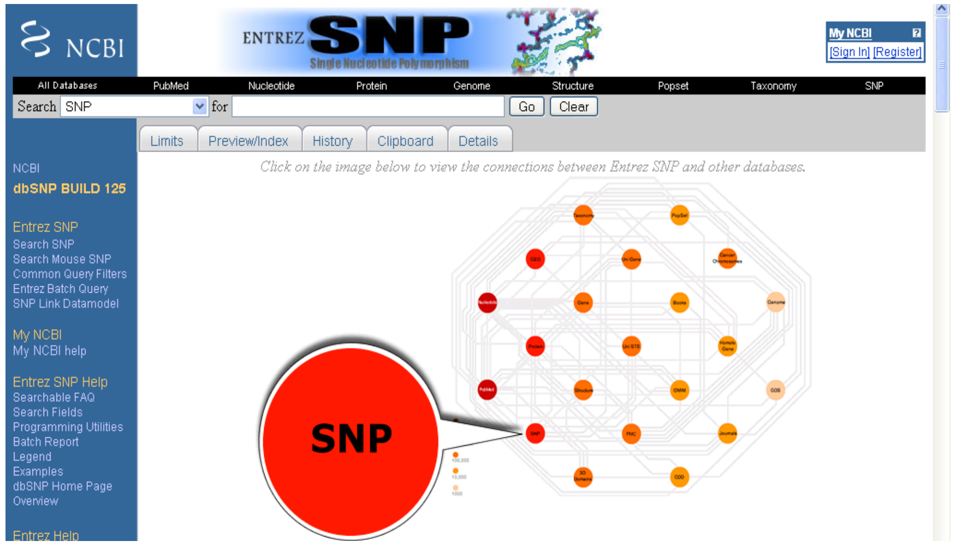
Single Nucleotide Variants (SNVs) are the most common type of genetic variations in human accounting for more than 90% of sequence differences (1000 Genome Project Consortium, 2012).



Capriotti et al. (2012). Briefings in Bioinformatics. 13; 495-512

SNVs and SAVs databases

dbSNP (2016/2017) @ NCBI



Click on the image below to view the connections between Entrez SNP and other databases.

http://www.ncbi.nlm.nih.gov/

Single Nucleotide Variants	
Homo sapiens	**135,967,291**
Bos taurus	39,722,628
Mus musculus	16,396,141

SwissVar (Jun 2017) @ ExPASy

http://www.expasy.ch/swissvar/

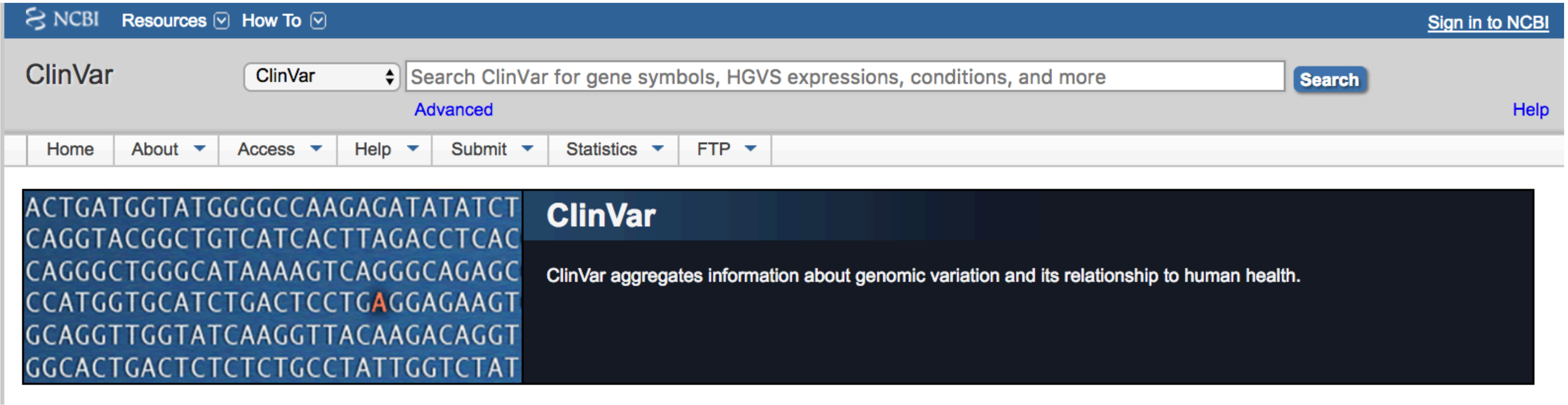
Single Amino acid Variants	
Homo sapiens	76,608
Disease	*29,529*
Polymorphisms	*39,779*

Jun 2017

Non-coding variants

Clinvar reports the clinical significance of ~280,000 short variants. Only 32,305 are annotated as Pathogenic and 17,180 as Benign.

Out of them ~89,000 variants are outside exotic regions, 3,164 are Pathogenic and 9,684 Benign.



NCBI
Resources
How To
Sign in to NCBI
ClinVar
ClinVar
Search ClinVar for gene symbols, HGVS expressions, conditions, and more
Search
Advanced
Help
Home
About
Access
Help
Submit
Statistics
FTP
ACTGATGGTATGGGGCCAAGAGATATATCT
CAGGTACGGCTGTCATCACTTAGACCTCAC
CAGGGCTGGGCATAAAAGTCAGGGCAGAGC
CCATGGTGCATCTGACTCCTGAGGAGAAGT
GCAGGTTGGTATCAAGGTTACAAGACAGGT
GGCACTGACTCTCTCTGCCTATTGGTCTAT
ClinVar

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

https://www.ncbi.nlm.nih.gov/clinvar/

Aug 2017

1000 Genomes

The 1000 Genomes Project aims to create the largest public catalogue of human variations and genotype data. Last versione released the genotype of ~2,500 individuals.

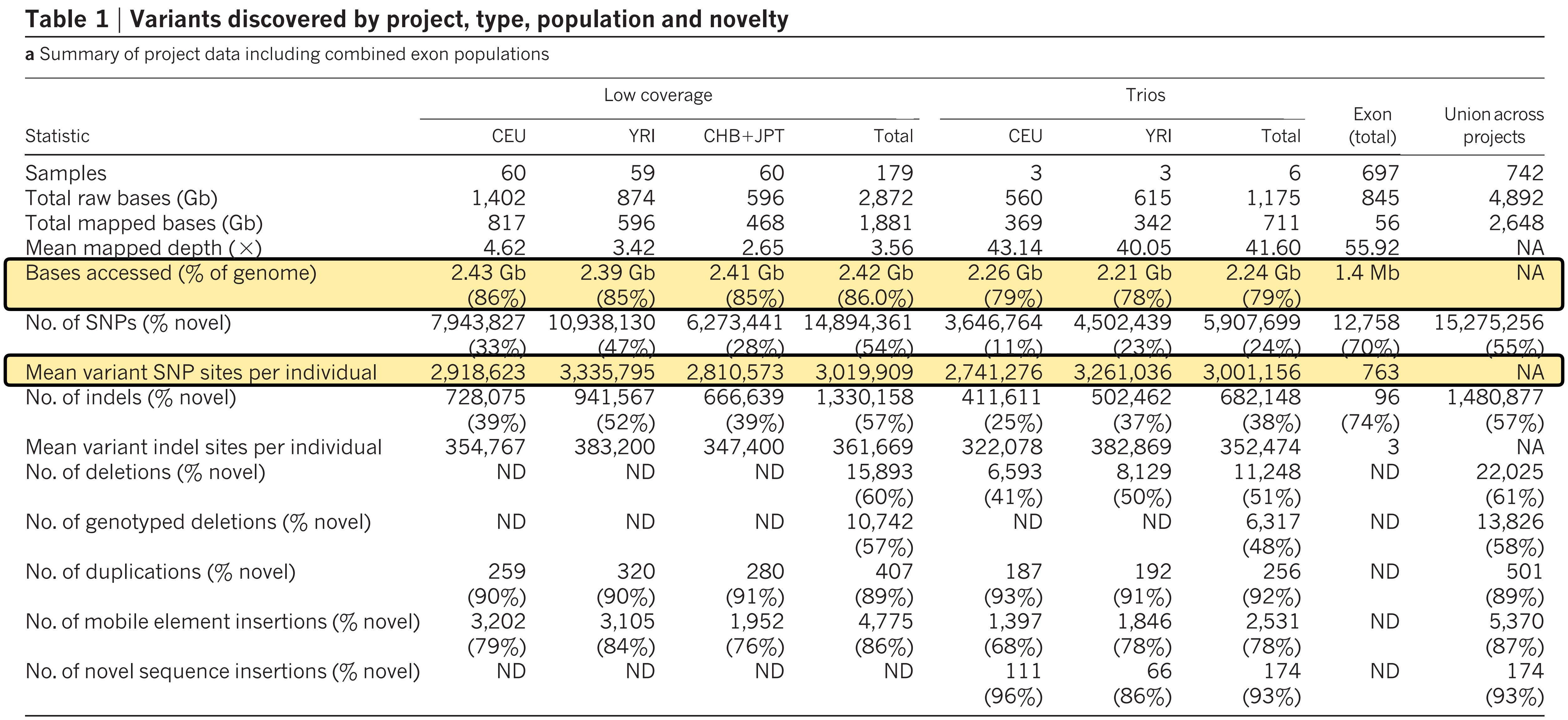


Table 1 | Variants discovered by project, type, population and novelty

a Summary of project data including combined exon populations

	Low coverage				Trios				
Statistic	CEU	YRI	CHB+JPT	Total	CEU	YRI	Total	Exon (total)	Union across projects
Samples	60	59	60	179	3	3	6	697	742
Total raw bases (Gb)	1,402	874	596	2,872	560	615	1,175	845	4,892
Total mapped bases (Gb)	817	596	468	1,881	369	342	711	56	2,648
Mean mapped depth (×)	4.62	3.42	2.65	3.56	43.14	40.05	41.60	55.92	NA
Bases accessed (% of genome)	2.43 Gb (86%)	2.39 Gb (85%)	2.41 Gb (85%)	2.42 Gb (86.0%)	2.26 Gb (79%)	2.21 Gb (78%)	2.24 Gb (79%)	1.4 Mb	NA
No. of SNPs (% novel)	7,943,827 (33%)	10,938,130 (47%)	6,273,441 (28%)	14,894,361 (54%)	3,646,764 (11%)	4,502,439 (23%)	5,907,699 (24%)	12,758 (70%)	15,275,256 (55%)
Mean variant SNP sites per individual	2,918,623	3,335,795	2,810,573	3,019,909	2,741,276	3,261,036	3,001,156	763	NA
No. of indels (% novel)	728,075 (39%)	941,567 (52%)	666,639 (39%)	1,330,158 (57%)	411,611 (25%)	502,462 (37%)	682,148 (38%)	96 (74%)	1,480,877 (57%)
Mean variant indel sites per individual	354,767	383,200	347,400	361,669	322,078	382,869	352,474	3	NA
No. of deletions (% novel)	ND	ND	ND	15,893 (60%)	6,593 (41%)	8,129 (50%)	11,248 (51%)	ND	22,025 (61%)
No. of genotyped deletions (% novel)	ND	ND	ND	10,742 (57%)	ND	ND	6,317 (48%)	ND	13,826 (58%)
No. of duplications (% novel)	259 (90%)	320 (90%)	280 (91%)	407 (89%)	187 (93%)	192 (91%)	256 (92%)	ND	501 (89%)
No. of mobile element insertions (% novel)	3,202 (79%)	3,105 (84%)	1,952 (76%)	4,775 (86%)	1,397 (68%)	1,846 (78%)	2,531 (78%)	ND	5,370 (87%)
No. of novel sequence insertions (% novel)	ND	ND	ND	ND	111 (96%)	66 (86%)	174 (93%)	ND	174 (93%)

1000 Genomes Project Consortium (2010). Nature. 467: 1061-1073.

Functional variants

An accurate estimation of the number of functional variants is given by the number of variants at conserved positions (GERP score >2).
The excess of deleterious rare variants is a significant fraction of the detected variants in the same class.

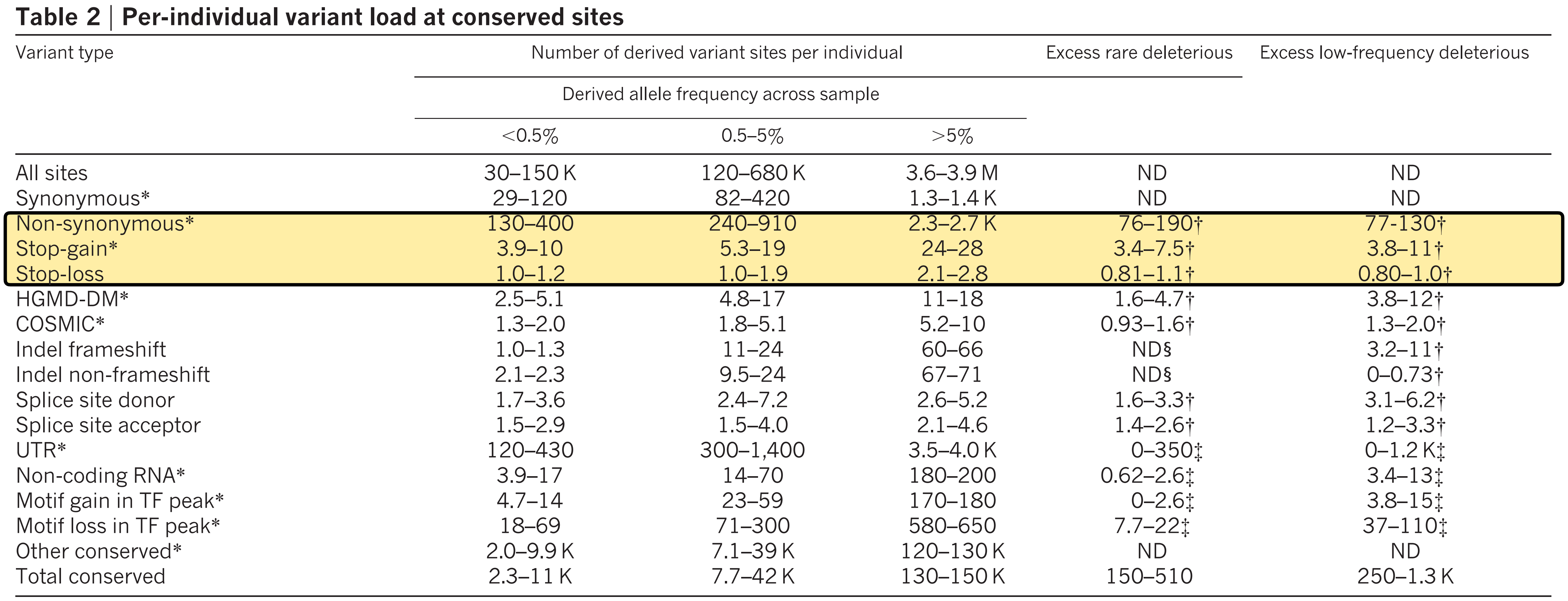


Table 2 | Per-individual variant load at conserved sites

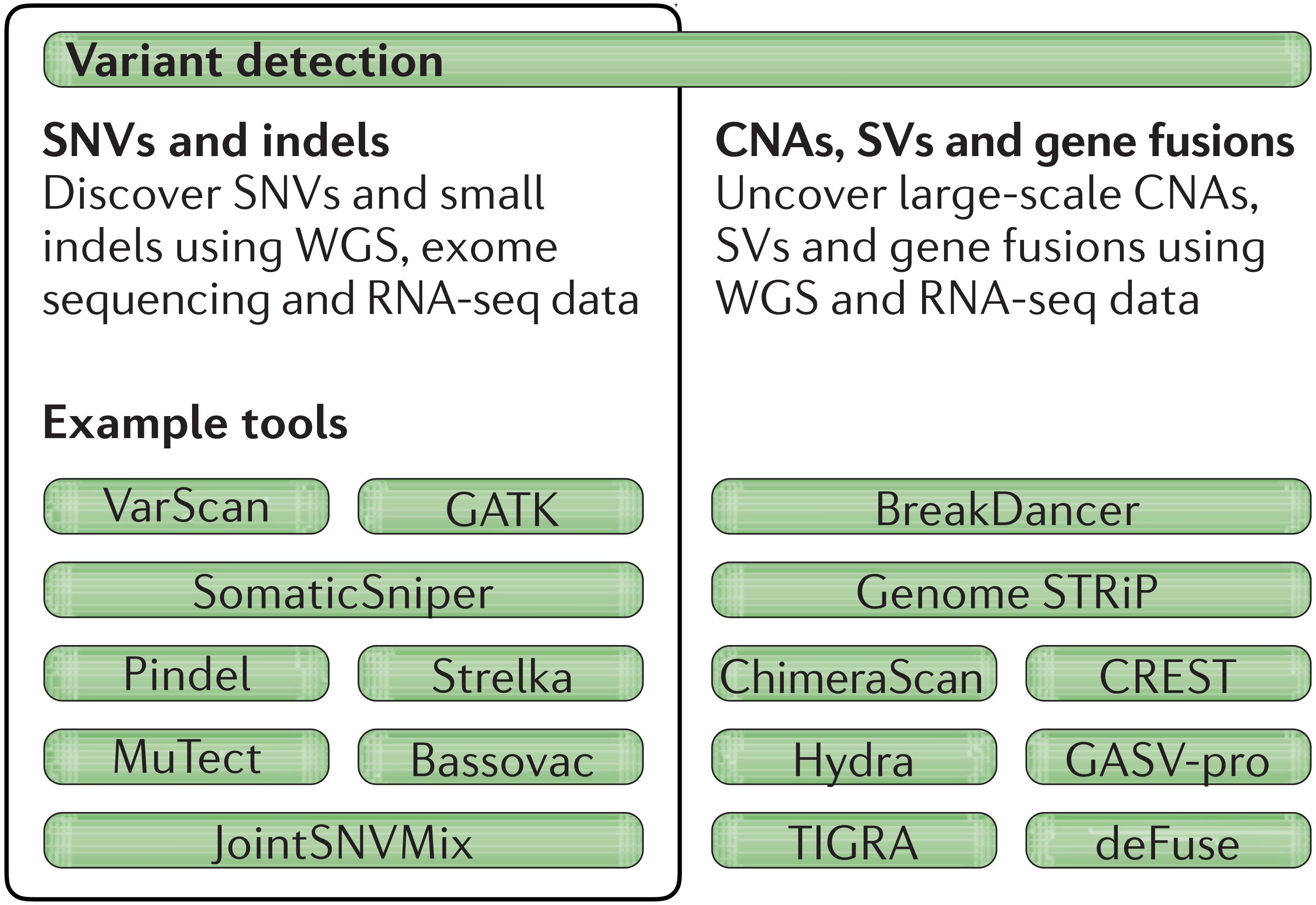
Variant type	Number of derived variant sites per individual			Excess rare deleterious	Excess low-frequency deleterious
	Derived allele frequency across sample				
	<0.5%	0.5–5%	>5%		
All sites	30–150 K	120–680 K	3.6–3.9 M	ND	ND
Synonymous*	29–120	82–420	1.3–1.4 K	ND	ND
Non-synonymous*	130–400	240–910	2.3–2.7 K	76–190†	77-130†
Stop-gain*	3.9–10	5.3–19	24–28	3.4–7.5†	3.8–11†
Stop-loss	1.0–1.2	1.0–1.9	2.1–2.8	0.81–1.1†	0.80–1.0†
HGMD-DM*	2.5–5.1	4.8–17	11–18	1.6–4.7†	3.8–12†
COSMIC*	1.3–2.0	1.8–5.1	5.2–10	0.93–1.6†	1.3–2.0†
Indel frameshift	1.0–1.3	11–24	60–66	ND§	3.2–11†
Indel non-frameshift	2.1–2.3	9.5–24	67–71	ND§	0–0.73†
Splice site donor	1.7–3.6	2.4–7.2	2.6–5.2	1.6–3.3†	3.1–6.2†
Splice site acceptor	1.5–2.9	1.5–4.0	2.1–4.6	1.4–2.6†	1.2–3.3†
UTR*	120–430	300–1,400	3.5–4.0 K	0–350‡	0–1.2 K‡
Non-coding RNA*	3.9–17	14–70	180–200	0.62–2.6‡	3.4–13‡
Motif gain in TF peak*	4.7–14	23–59	170–180	0–2.6‡	3.8–15‡
Motif loss in TF peak*	18–69	71–300	580–650	7.7–22‡	37–110‡
Other conserved*	2.0–9.9 K	7.1–39 K	120–130 K	ND	ND
Total conserved	2.3–11 K	7.7–42 K	130–150 K	150–510	250–1.3 K

1000 Genomes Project Consortium (2012). Nature. 491: 56-63.

Short variant detection

Variant detection

Variant detection can be performed using several tools. Some method are specific for particular types of variant.



Variant detection

SNVs and indels
Discover SNVs and small indels using WGS, exome sequencing and RNA-seq data

CNAs, SVs and gene fusions
Uncover large-scale CNAs, SVs and gene fusions using WGS and RNA-seq data

Example tools

VarScan
GATK
SomaticSniper
Pindel
Strelka
MuTect
Bassovac
JointSNVMix

BreakDancer
Genome STRiP
ChimeraScan
CREST
Hydra
GASV-pro
TIGRA
deFuse

How data looks like?

Variant Calling File (VCF) with germline and somatic variants

```
##fileformat=VCFv4.1
##tcgaversion=1.1
##reference=<ID=hg19,source=.>
##phasing=none
##geneAnno=none
##INFO=<ID=VT,Number=1,Type=String,Description="Variant type, can be SNP, INS or DEL">
##INFO=<ID=VLS,Number=1,Type=Integer,Description="Final validation status relative to non-adjacent Normal, ......”>
##FILTER=<ID=CA,Description="Fail Carnac (Tumor and normal coverage, tumor variant count, mapping quality, ......”>
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read depth at this position in the sample">
##FORMAT=<ID=AD,Number=.,Type=Integer,Description="Depth of reads supporting alleles 0/1/2/3...">
##FORMAT=<ID=BQ,Number=.,Type=Integer,Description="Average base quality for reads supporting alleles">
##FORMAT=<ID=SS,Number=1,Type=Integer,Description="Variant status relative to non-adjacent Normal,0=wildtype, ......">
##FORMAT=<ID=SSC,Number=1,Type=Integer,Description="Somatic score between 0 and 255">
##FORMAT=<ID=MQ60,Number=1,Type=Integer,Description="Number of reads (mapping quality=60) supporting variant">
#CHROM     POS        ID    REF   ALT   QUAL  FILTER     INFO            FORMAT                      NORMAL                  PRIMARY
1          10048      .     C     CCT   .     CA         VT=INS;VLS=5    GT:DP:AD:BQ:SS:SSC:MQ60     0/0:66:.,0:.:0:.:0      0/1:32:.,2:.:2:.:0
1          10078      .     CT    C     .     CA         VT=DEL;VLS=5    GT:DP:AD:BQ:SS:SSC:MQ60     0/0:25:.,0:.:0:.:0      0/1:13:.,2:.:2:.:0
1          10177      .     A     AC    .     CA         VT=INS;VLS=5    GT:DP:AD:BQ:SS:SSC:MQ60     0/0:57:.,0:.:0:.:0      0/1:22:.,2:.:2:.:0
. . . . . .
. . . . . .
1          900505     .     G     C     .     PASS       VT=SNP;VLS=5    GT:DP:AD:BQ:SS:SSC:MQ60     0/1:188:.,89:26:1:.:81  0/1:210:.,113:24:1:.:100
. . . . . .
. . . . . .
1          1991007    .     G     T     .     PASS       VT=SNP;VLS=5    GT:DP:AD:BQ:SS:SSC:MQ60     0/0:222:.,1:2:0:.:1     0/1:88:.,41:25:2:50:34
. . . . . .
```

Variant significance

The probability of observing a particular variant by chance can be calculated using different procedure.
VarScan2 uses Fisher’s exact test where the background distribution correspond all reads mapping the reference allele.

```
CHROM: chr17
POS: 560603
ID: .
REF: A
ALT: G
QUAL: .
FILTER: PASS
INFO: ADP=94;WT=0;HET=1;HOM=0;NC=0
FORMAT: ADP=94;WT=0;HET=1;HOM=0;NC=0
FORMAT: GT:GQ:SDP:DP:RD:AD:FREQ:PVAL:RBQ:ABQ:RDF:RDR:ADF:ADR
SAMPLE: 0/1:194:94:94:43:51:54.26%:3.3469E-20:40:40:14:29:13:38
```

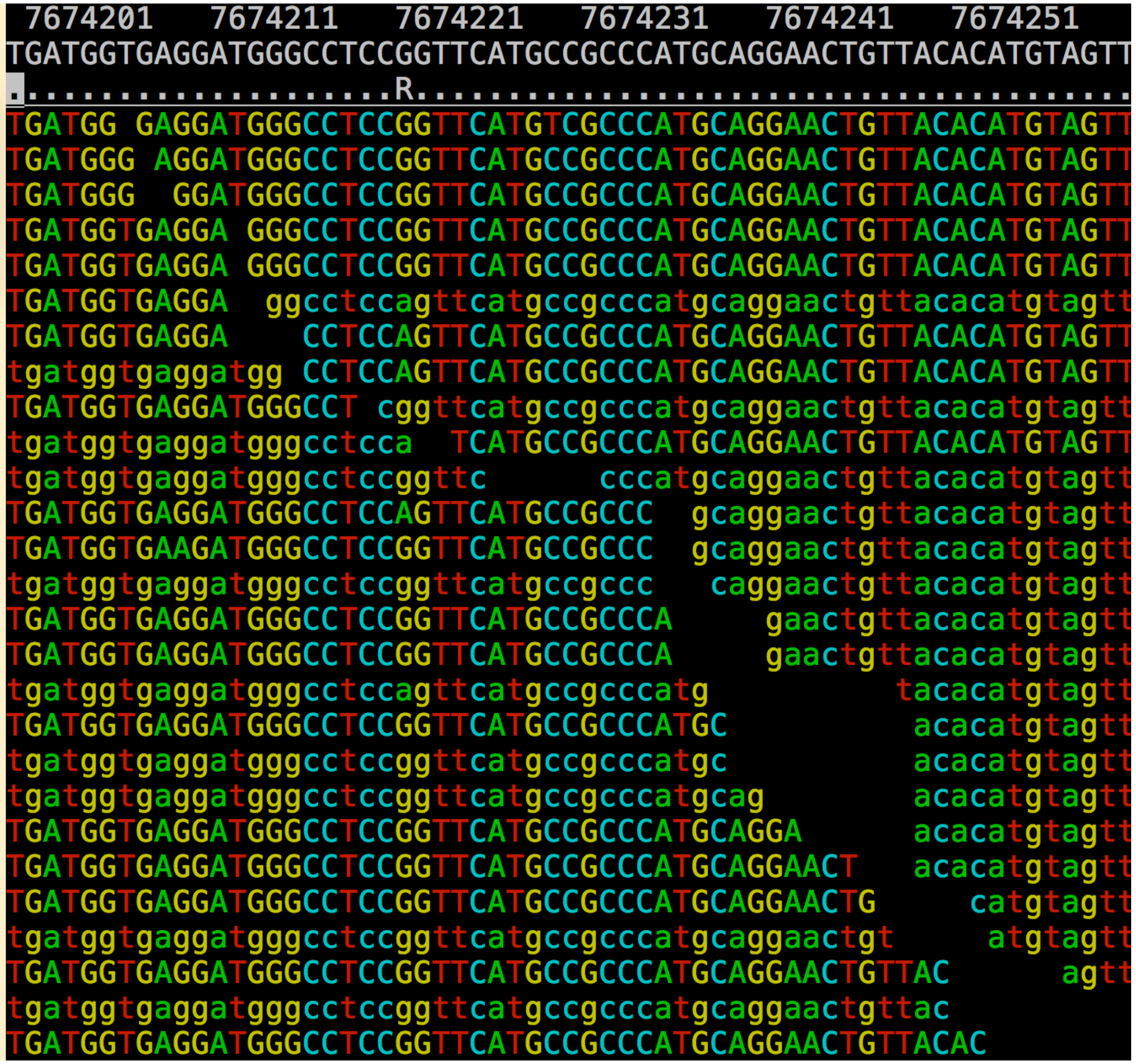
Contingency Table

	A	G
Data	43	51
Background	94	0

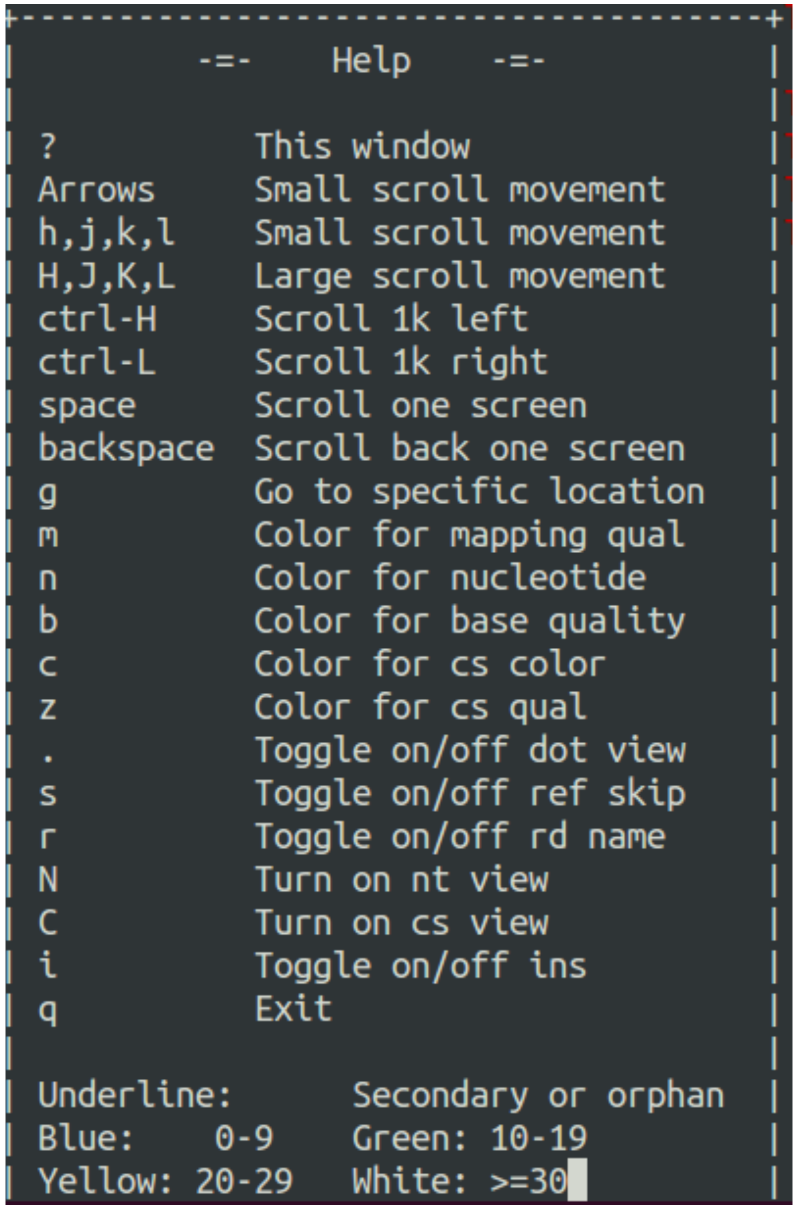
samtools view

Usage: samtools tview [options] <aln.bam> [ref.fasta]

```
samtools tview  -p chr17:7674200  bam/tumor_chr17.bam hg38/GRCh38.d1.vd1.fa
```

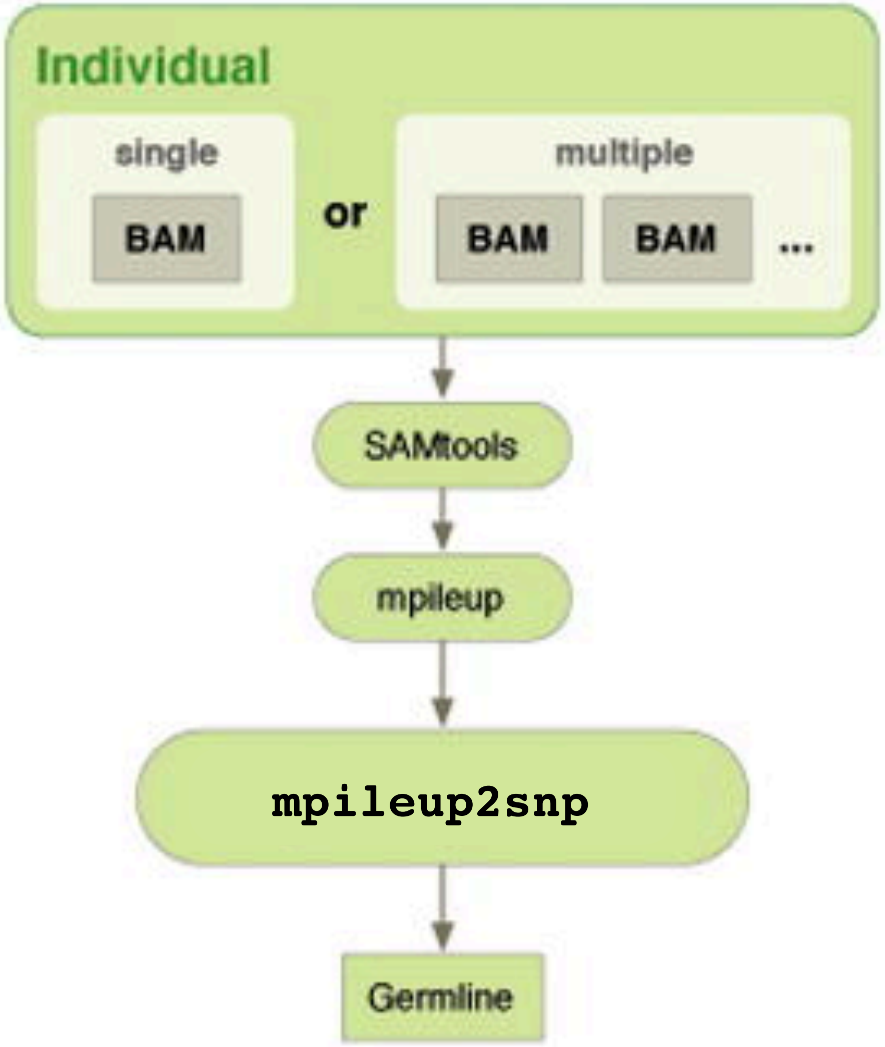


```
 7674201   7674211   7674221   7674231   7674241   7674251
TGATGGTGAGGATGGGCCTCCGGTTCATGCCGCCCATGCAGGAACTGTTACACATGTAGTT
.....................R.......................................
TGATGG GAGGATGGGCCTCCGGTTCATGTCGCCCATGCAGGAACTGTTACACATGTAGTT
TGATGGG AGGATGGGCCTCCGGTTCATGCCGCCCATGCAGGAACTGTTACACATGTAGTT
TGATGGG  GGATGGGCCTCCGGTTCATGCCGCCCATGCAGGAACTGTTACACATGTAGTT
TGATGGTGAGGA GGGCCTCCGGTTCATGCCGCCCATGCAGGAACTGTTACACATGTAGTT
TGATGGTGAGGA GGGCCTCCGGTTCATGCCGCCCATGCAGGAACTGTTACACATGTAGTT
TGATGGTGAGGA  ggcctccagttcatgccgcccatgcaggaactgttacacatgtagtt
TGATGGTGAGGA    CCTCCAGTTCATGCCGCCCATGCAGGAACTGTTACACATGTAGTT
tgatggtgaggatgg CCTCCAGTTCATGCCGCCCATGCAGGAACTGTTACACATGTAGTT
TGATGGTGAGGATGGGCCT cggttcatgccgcccatgcaggaactgttacacatgtagtt
tgatggtgaggatgggcctcca  TCATGCCGCCCATGCAGGAACTGTTACACATGTAGTT
tgatggtgaggatgggcctccggttc      cccatgcaggaactgttacacatgtagtt
TGATGGTGAGGATGGGCCTCCAGTTCATGCCGCCC  gcaggaactgttacacatgtagtt
TGATGGTGAAGATGGGCCTCCGGTTCATGCCGCCC  gcaggaactgttacacatgtagtt
tgatggtgaggatgggcctccggttcatgccgccc   caggaactgttacacatgtagtt
TGATGGTGAGGATGGGCCTCCGGTTCATGCCGCCCA     gaactgttacacatgtagtt
TGATGGTGAGGATGGGCCTCCGGTTCATGCCGCCCA     gaactgttacacatgtagtt
tgatggtgaggatgggcctccagttcatgccgcccatg          tacacatgtagtt
TGATGGTGAGGATGGGCCTCCGGTTCATGCCGCCCATGC          acacatgtagtt
tgatggtgaggatgggcctccggttcatgccgcccatgc          acacatgtagtt
tgatggtgaggatgggcctccggttcatgccgcccatgcag        acacatgtagtt
TGATGGTGAGGATGGGCCTCCGGTTCATGCCGCCCATGCAGGA      acacatgtagtt
TGATGGTGAGGATGGGCCTCCGGTTCATGCCGCCCATGCAGGAACT   acacatgtagtt
TGATGGTGAGGATGGGCCTCCGGTTCATGCCGCCCATGCAGGAACTG     catgtagtt
tgatggtgaggatgggcctccggttcatgccgcccatgcaggaactgt     atgtagtt
TGATGGTGAGGATGGGCCTCCGGTTCATGCCGCCCATGCAGGAACTGTTAC      agtt
tgatggtgaggatgggcctccggttcatgccgcccatgcaggaactgttac
TGATGGTGAGGATGGGCCTCCGGTTCATGCCGCCCATGCAGGAACTGTTACAC
```



```
+-------------------------------------------+
|          -=-    Help    -=-               |
|                                           |
| ?          This window                    |
| Arrows     Small scroll movement          |
| h,j,k,l    Small scroll movement          |
| H,J,K,L    Large scroll movement          |
| ctrl-H     Scroll 1k left                 |
| ctrl-L     Scroll 1k right                |
| space      Scroll one screen              |
| backspace  Scroll back one screen         |
| g          Go to specific location        |
| m          Color for mapping qual         |
| n          Color for nucleotide           |
| b          Color for base quality         |
| c          Color for cs color             |
| z          Color for cs qual              |
| .          Toggle on/off dot view         |
| s          Toggle on/off ref skip         |
| r          Toggle on/off rd name          |
| N          Turn on nt view                |
| C          Turn on cs view                |
| i          Toggle on/off ins              |
| q          Exit                           |
|                                           |
| Underline:      Secondary or orphan       |
| Blue:    0-9    Green: 10-19              |
| Yellow: 20-29   White: >=30               |
```

VarScan2 germline call



STEP1

```
samtools mpileup [options] in1.bam ……

samtools mpileup -B -q 1 -f
   hg38/GRCh38.d1.vd1.fa  bam/normal_chr17.bam
   >normal_chr17.mpileup
```

STEP 2

```
java -jar VarScan.v2.4.1.jar
 mpileup2snp mpileupfile [options]

java -jar VarScan.v2.4.1.jar  mpileup2snp
  normal_chr17.mpileup --min-coverage 10
  —min-var-freq 0.2 --p-value 0.05
  --output-vcf 1 > normal_chr17.snp.vcf
```

Koboldt et al. (2013). Curr Protoc Bioinformatics.

vcftools

Powerful tools for manipulating the variant call format (VCF) and binary variant call format (BCF)

Select a chromosome region

```
vcftools --vcf  1kgenomes/tp53_1kgenomes.vcf --chr 17
       --from-bp 7571752 --to-bp 7590868 --recode --stdout
```

Select variant with a minimum depth

```
vcftools --vcf 1kgenomes/tp53_1kgenomes.vcf --minDP 4 --recode
      —stdout
```

Select genotype of specific individuals

```
vcftools --vcf  1kgenomes/tp53_1kgenomes.vcf --indv HG00110
      --indv HG00113 --recode --stout
```

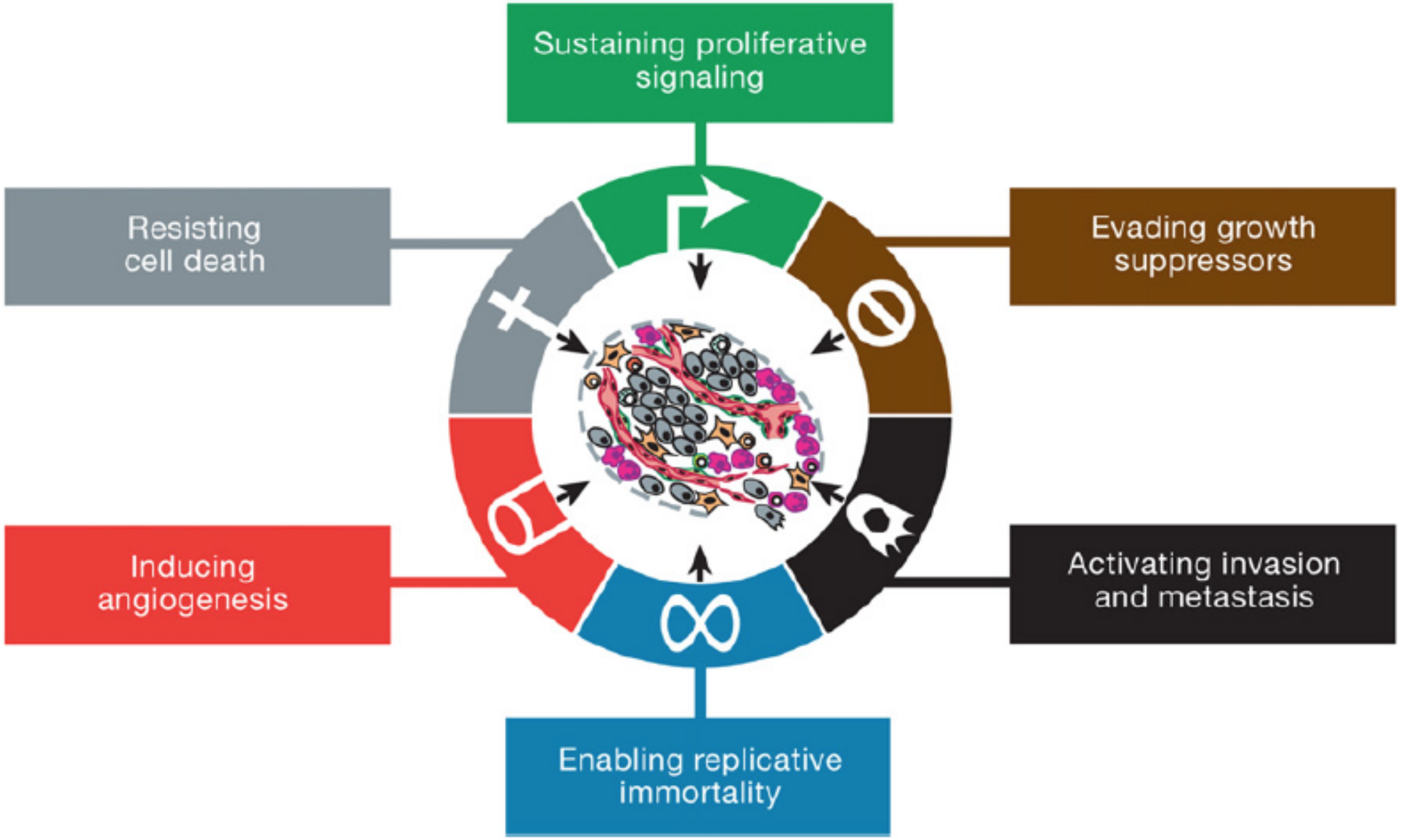
Compare vcf files

```
vcftools --vcf 1kgenomes/tp53_1kgenomes.vcf --diff
         1kgenomes/tp53_1kgenomes_ends.vcf  --diff-site  --stdout
```

Variations in Cancer

Hallmarks of cancer

The six hallmarks of cancer - distinctive and complementary capabilities that enable tumor growth and metastatic dissemination.


Sustaining proliferative signaling
Resisting cell death
Evading growth suppressors
Inducing angiogenesis
Activating invasion and metastasis
Enabling replicative immortality

Hanahan and Weinberg (2011) Cell, **144**:646

The complexity of cancer

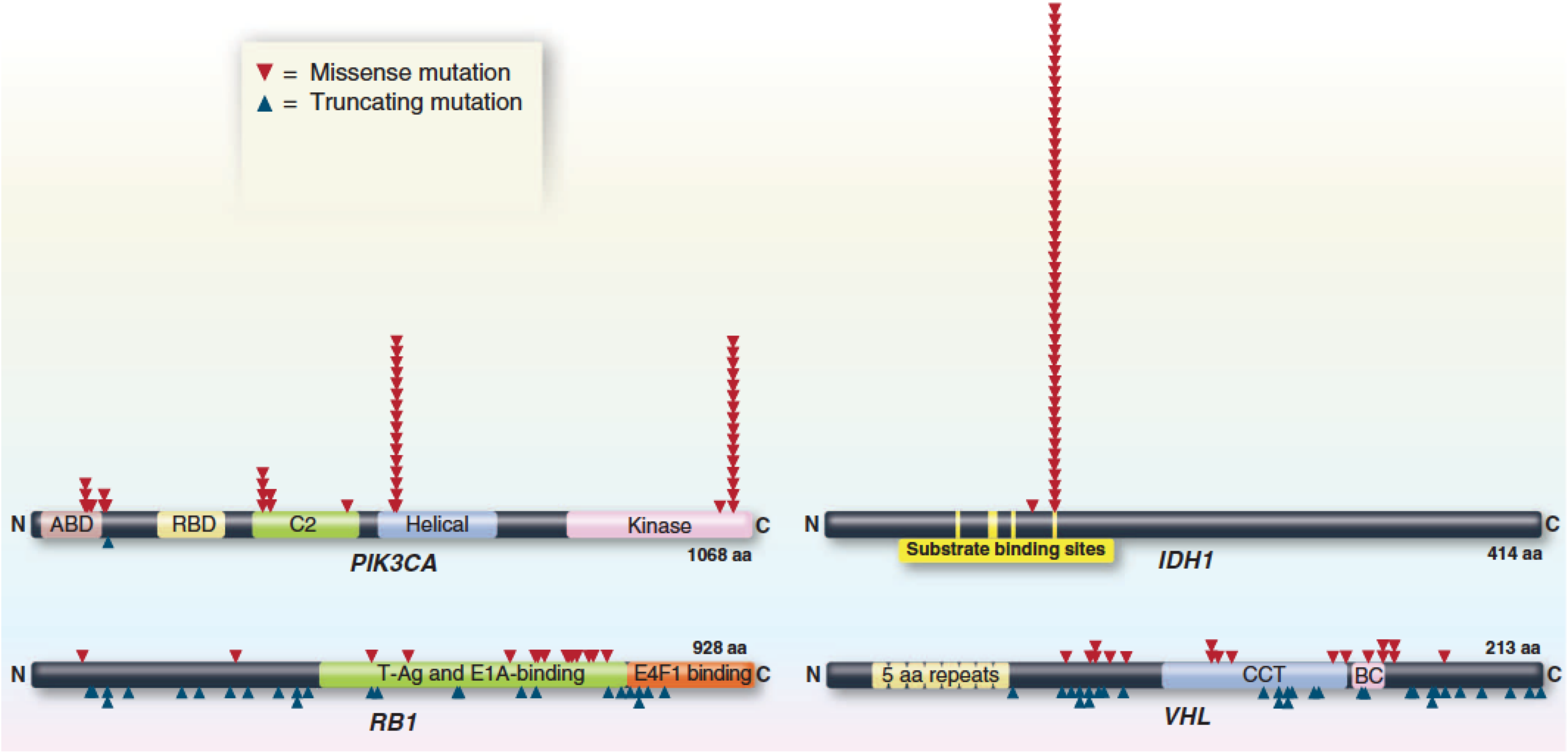
Cancer is **complex disorder** characterized by high level of mutation rate.

Mutations can be classified in germline and somatic whether they are inherited from parents or the result of error in DNA replication.

Another classification is between driver and passenger mutations whether they provide selective advantage with respect to normal cells increasing their proliferation rate or not.

Oncogene vs Suppressor

Oncogenes have highly recurrent mutations, tumor suppressors have sparse variants.



Vogelstein et al. (2013) Science , **339**:1546

Main challenges

Computational methods for cancer genome interpretation have been developed to address the following issues:

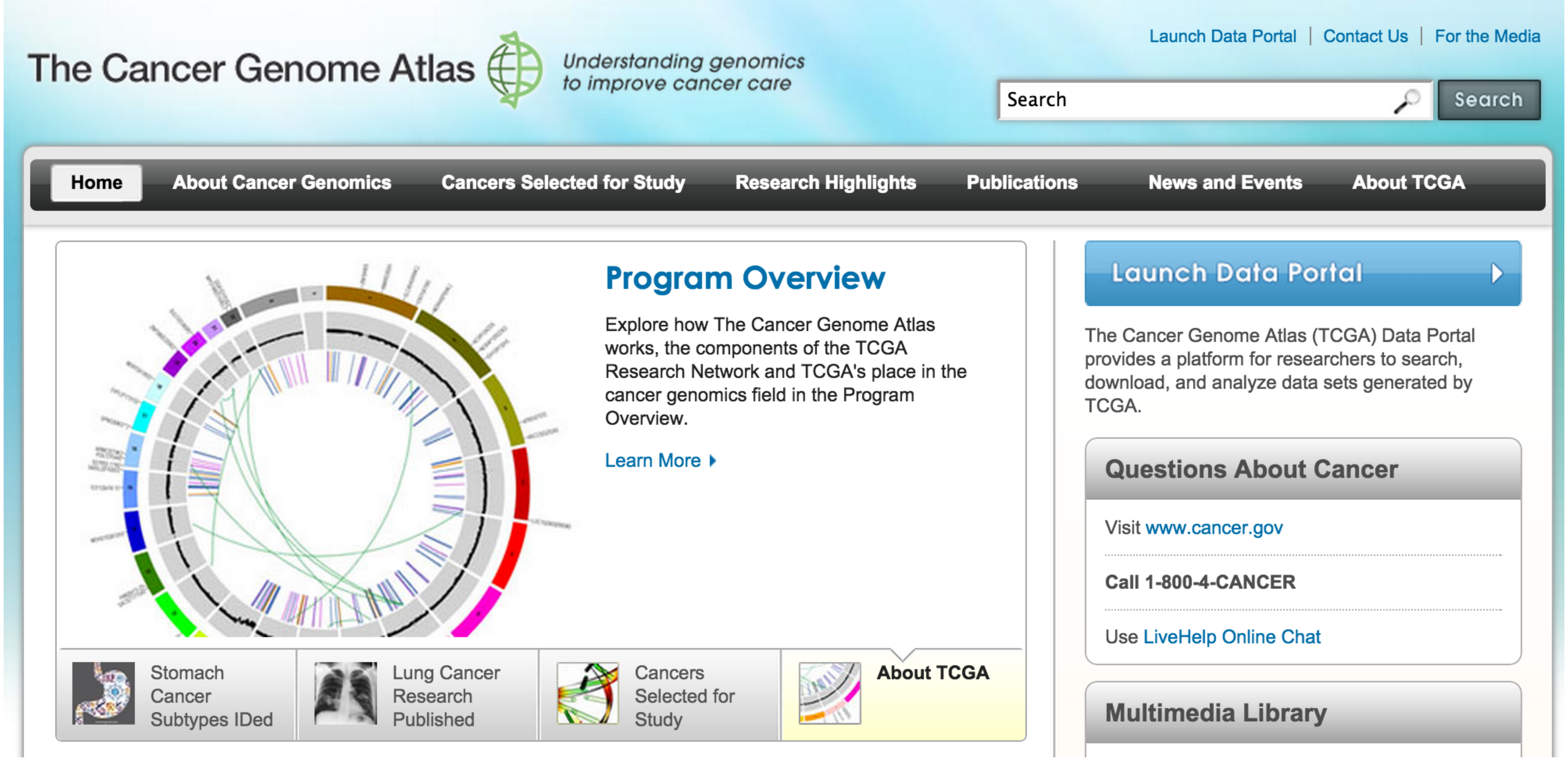
- Detection of recurrent somatic mutations and cancer driver genes;
- Prediction of driver variants and their functional impact;
- Estimate the impact of multiple variants at network and pathway level;
- Differentiate subclonal populations and their variation pattern.

The TCGA portal

The Cancer Genome Atalas Consortium

TCGA (http://cancergenome.nih.gov/)

- 36 cancer types
- BAM files available through the CGHub portal



Launch Data Portal | Contact Us | For the Media

The Cancer Genome Atlas

Understanding genomics to improve cancer care

Program Overview

Explore how The Cancer Genome Atlas works, the components of the TCGA Research Network and TCGA's place in the cancer genomics field in the Program Overview.

Learn More ▸

Launch Data Portal ▸

The Cancer Genome Atlas (TCGA) Data Portal provides a platform for researchers to search, download, and analyze data sets generated by TCGA.

Questions About Cancer

Visit www.cancer.gov

Call 1-800-4-CANCER

Use LiveHelp Online Chat

Multimedia Library

Stomach Cancer Subtypes IDed

Lung Cancer Research Published

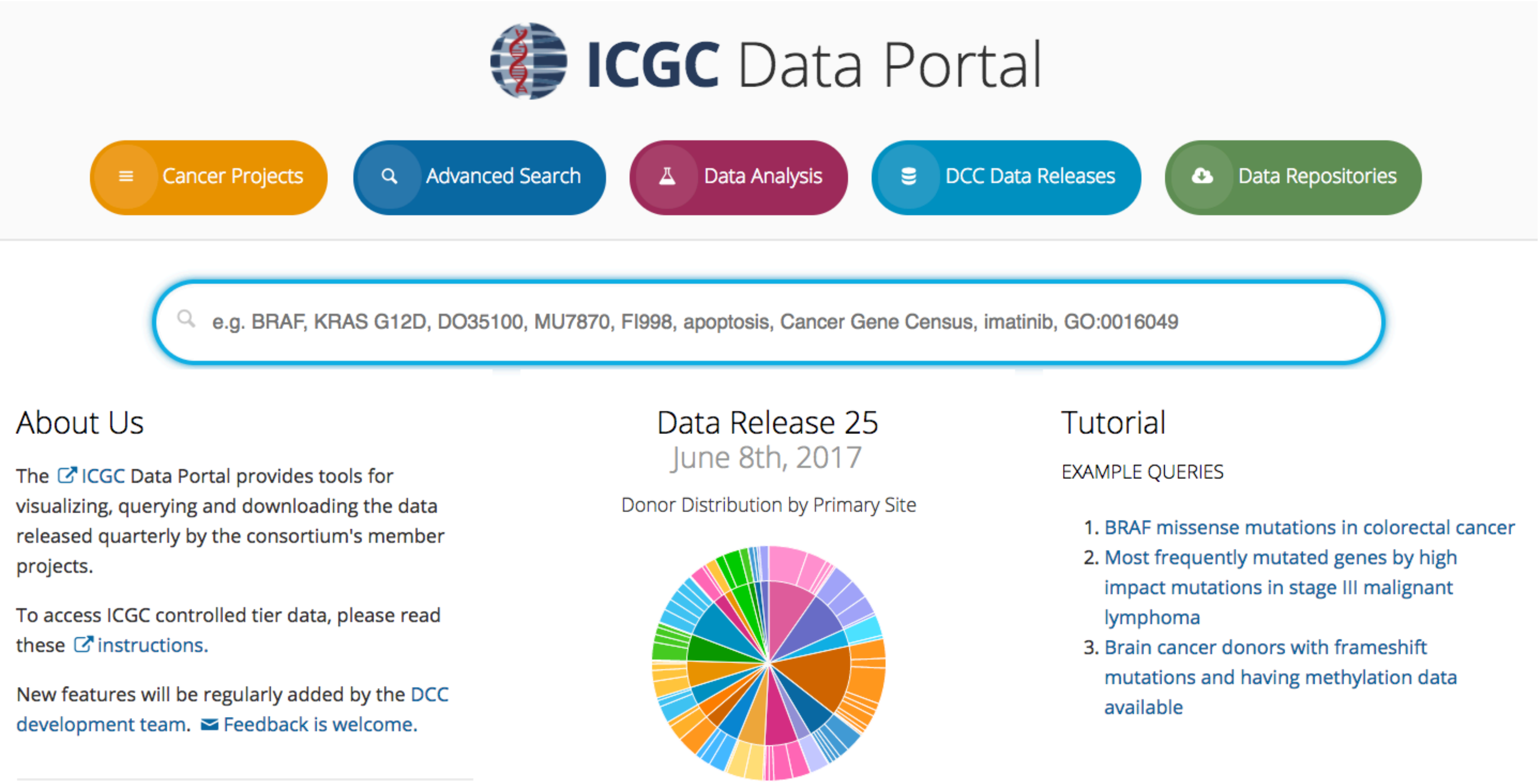
Cancers Selected for Study

About TCGA

The ICGC data portal

The International Cancer Genome Consortium

- 17,570 cancer patients
- 76 cancer projects in 21 primary sites
- more than 63 million simple somatic mutations.



ICGC Data Portal

Cancer Projects
Advanced Search
Data Analysis
DCC Data Releases
Data Repositories
e.g. BRAF, KRAS G12D, DO35100, MU7870, FI998, apoptosis, Cancer Gene Census, imatinib, GO:0016049

About Us

The ICGC Data Portal provides tools for visualizing, querying and downloading the data released quarterly by the consortium's member projects.

To access ICGC controlled tier data, please read these instructions.

New features will be regularly added by the DCC development team. Feedback is welcome.

Data Release 25

June 8th, 2017

Donor Distribution by Primary Site

Tutorial

EXAMPLE QUERIES

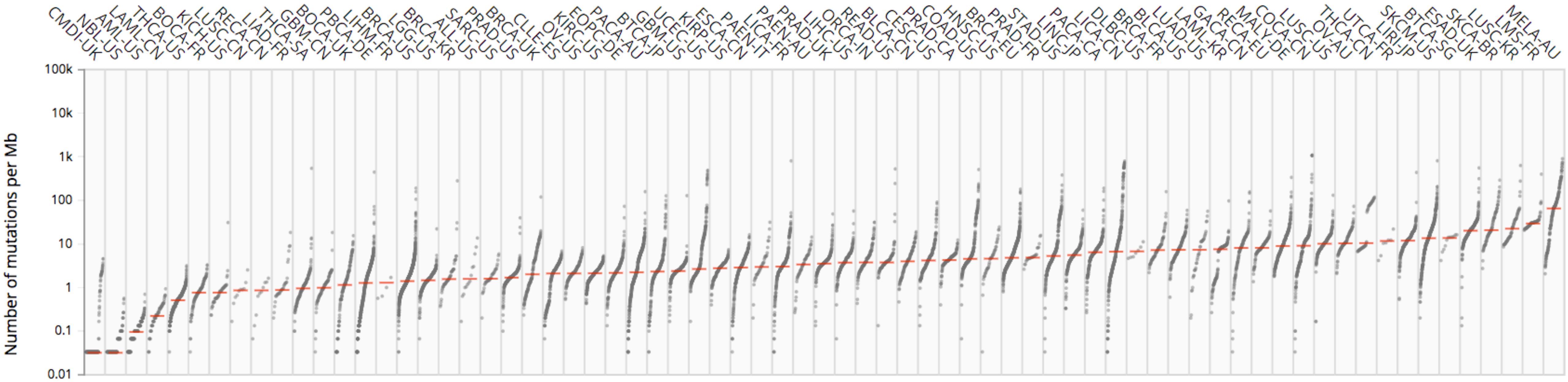
1. BRAF missense mutations in colorectal cancer
2. Most frequently mutated genes by high impact mutations in stage III malignant lymphoma
3. Brain cancer donors with frameshift mutations and having methylation data available

ICGC (https://dcc.icgc.org/)

Mutational landscape

The distribution of somatic variants varies significantly across cancer types

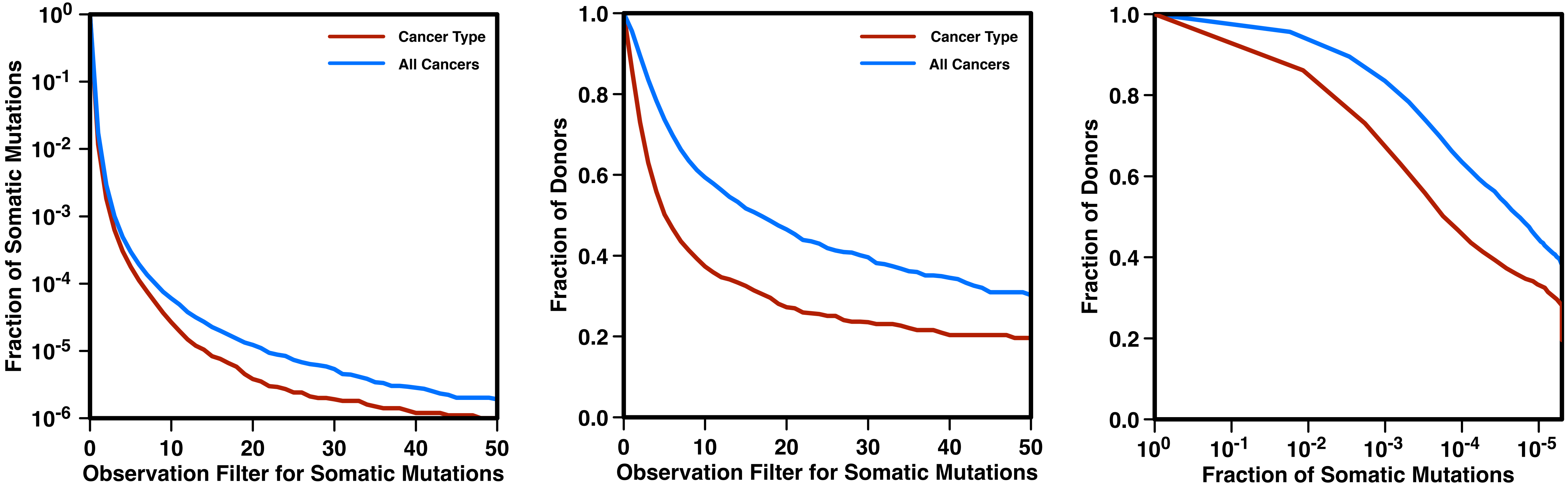
Number of Somatic Mutations in Donor's Exomes Across Cancer Projects



Driver vs Passenger

Number of recurrent mutations decrease exponentially.
On average a small fraction of variants a present in the majority of the samples.

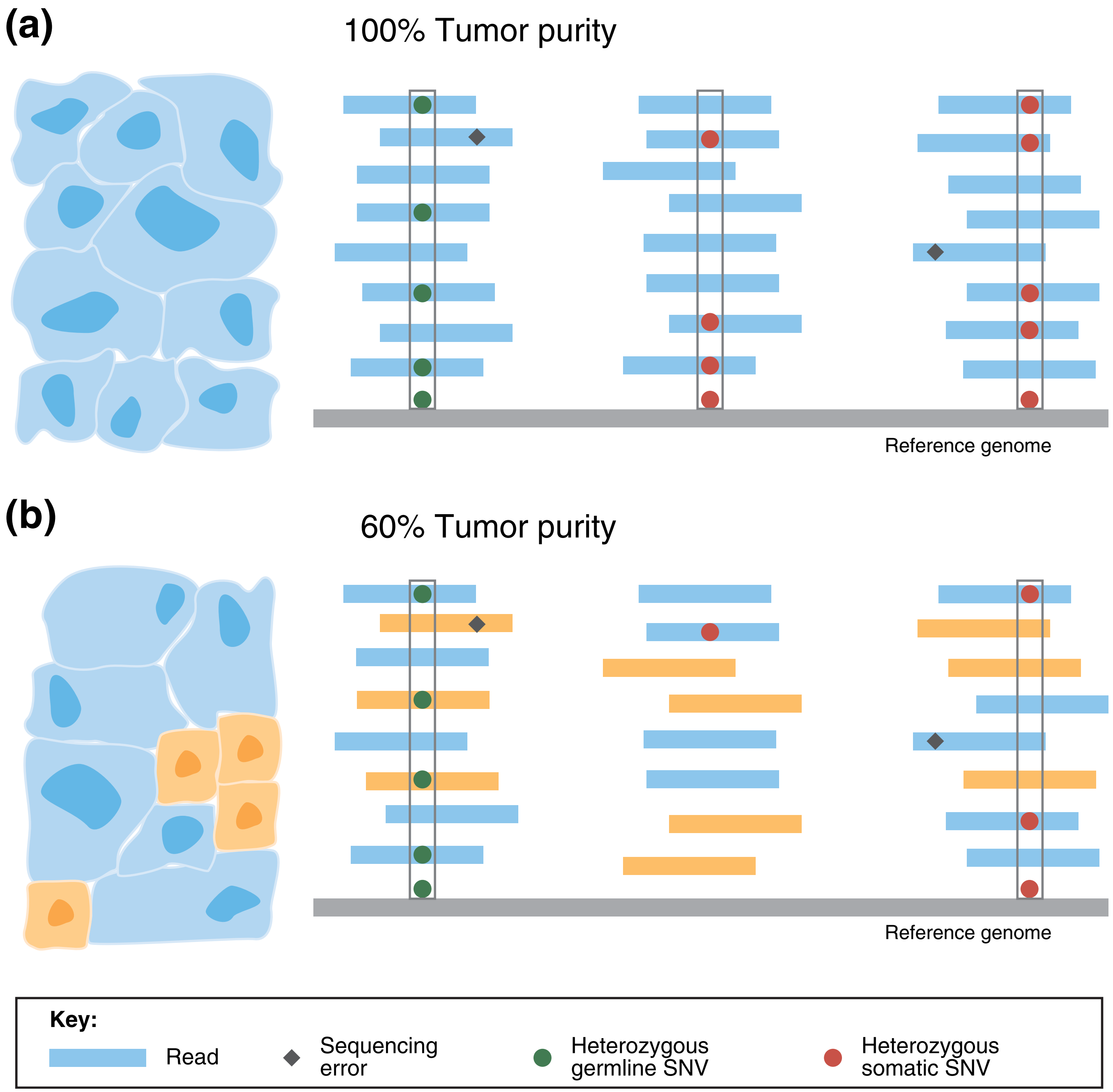
Selecting mutations that are repeated at least twice we filter out ~98% mutations
and are still able to recover ~96% of the patients



Tian R, Basu M, Capriotti E.(2015) BMC Genomics. 16 (Suppl. 8): S7.

Sample purity

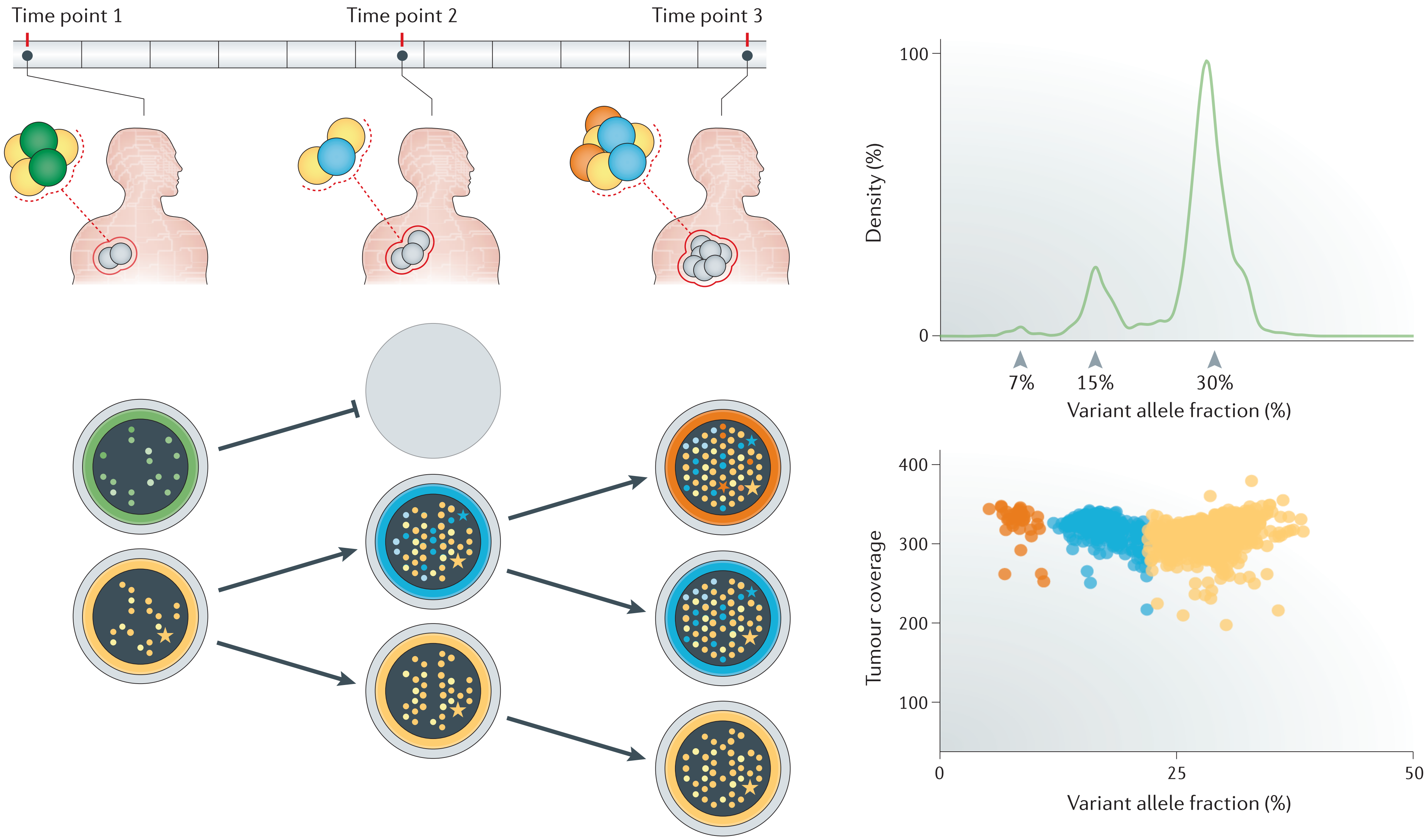
Impurity in the sample purity reduce the ability to detect variants



Raphael et al. (2014) Genome Medicine, **6**:5

Clonal evolution

On average tumor samples have ~150 more rare missense variants and mutated genes



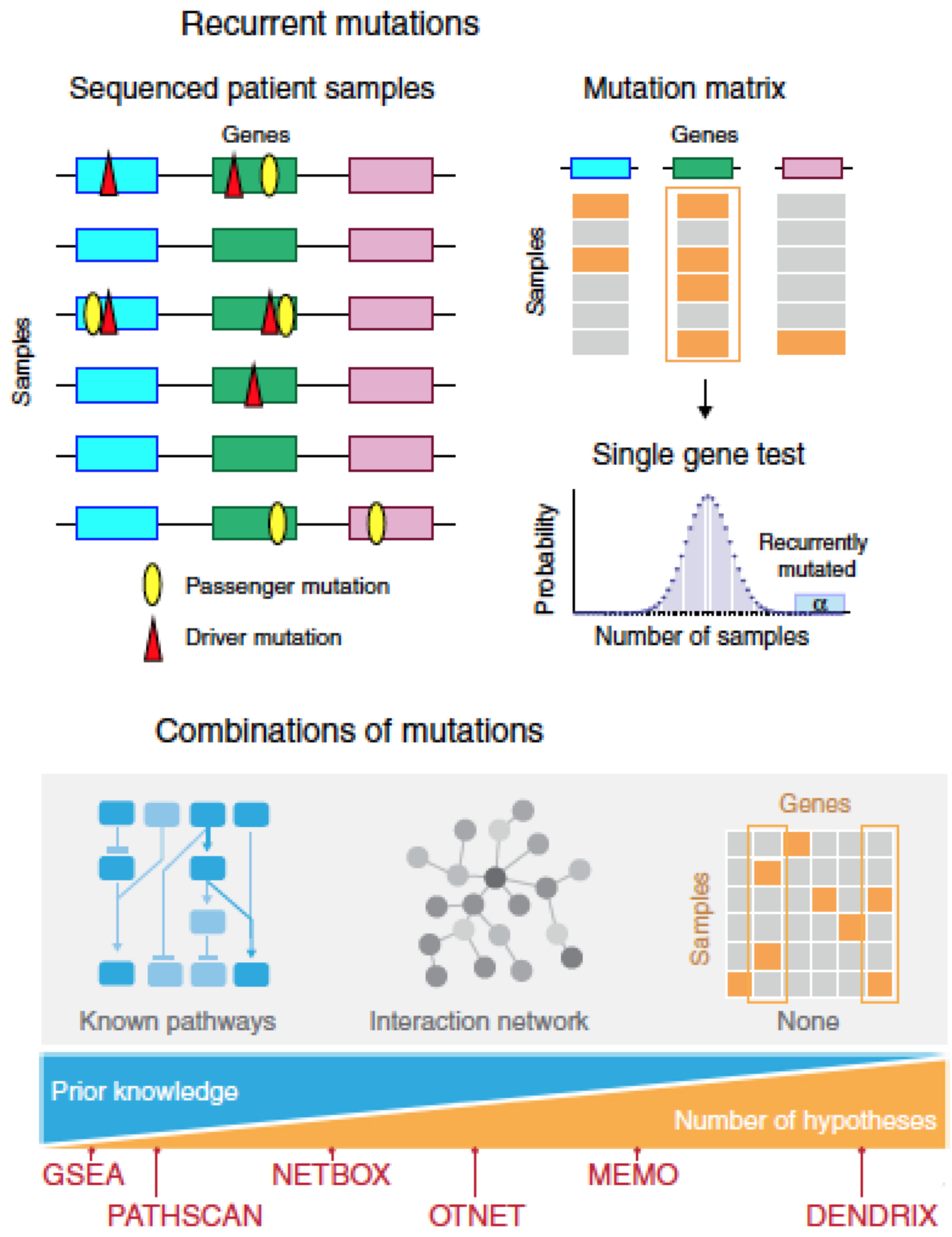
Ding et al (2014). Nat. Rev Genetics.

Recurrent variations

Recurrent mutations found in more samples than expected are good candidates for driver mutations.

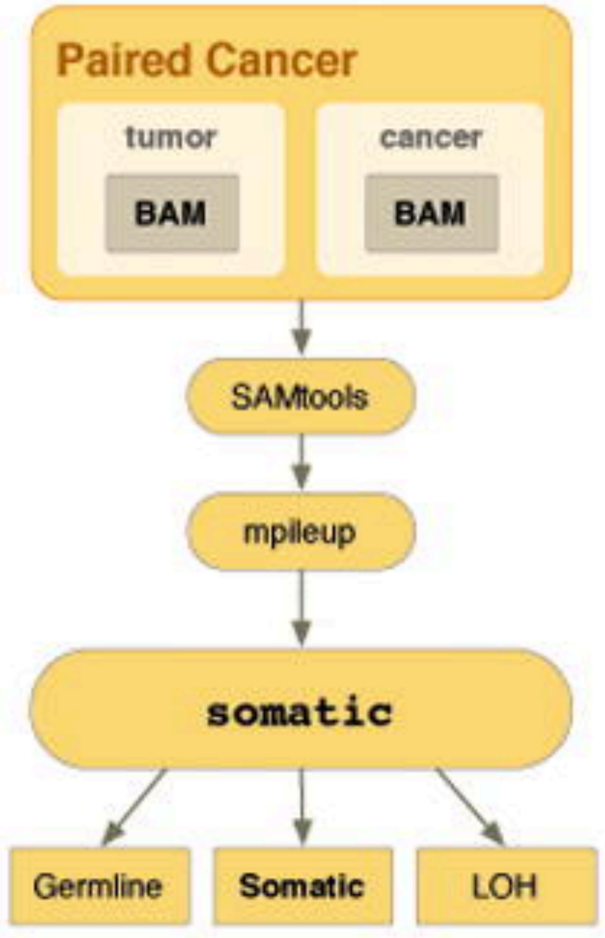
To identify such recurrent mutations, a statistical test is performed which usually collapses all the non-synonymous mutations in a gene.

Identification of recurrent mutations in predefined groups of genes such as pathways and protein-protein interaction networks and de novo identification of combinations, without relying on a priori definition.



Raphael et al. Genome Medicine 2014, **6**:5

VarScan2 somatic call (I)



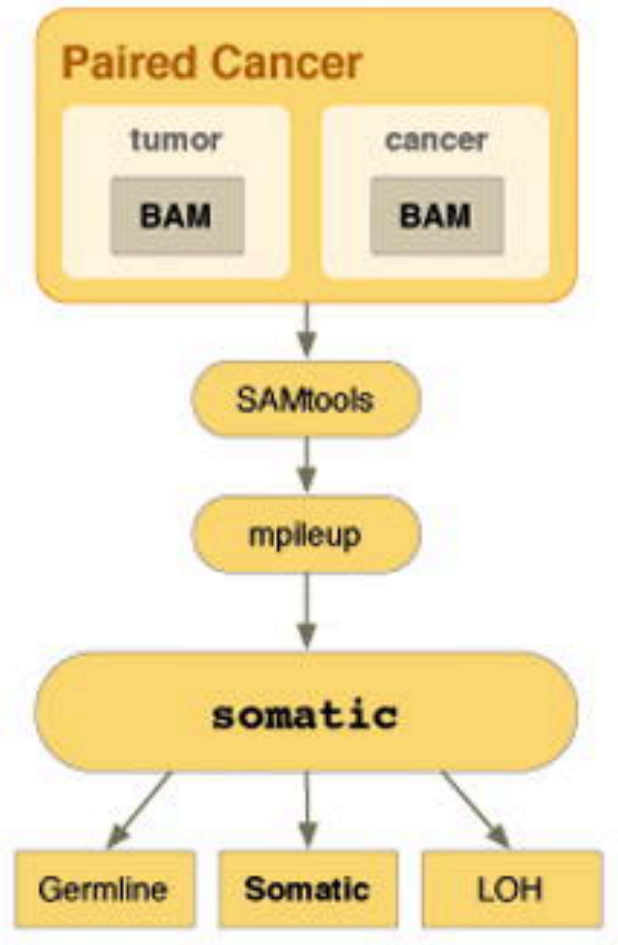
STEP 1

```
samtools mpileup [options] in1.bam in

samtools mpileup -B -q 1 -f
   hg38/GRCh38.d1.vd1.fa  bam/normal_chr17.bam
   bam/tumor_chr17.bam
   >normal_tumor_chr17.mpileup
```

STEP 2

```
java -jar VarScan.v2.4.1.jar
 somatic mpileupfile outfile.mpileup [options]

java -jar VarScan.v2.4.1.jar somatic
    normal_tumor.mpileup normal_tumor.vcf
    --output-vcf 1 --min-coverage 3
    --min-var-freq 0.08 --p-value 0.10
    --somatic-p-value 0.05
    --strand-filter 0 --mpileup 1
```

VarScan2 somatic call (II)



Paired Cancer

tumor

BAM

cancer

BAM

SAMtools

mpileup

somatic

Germline

Somatic

LOH

STEP 3

```
java -jar VarScan.v2.4.1.jar processSomatic
    variant_file

java -jar VarScan.v2.4.1.jar processSomatic
    normal_tumor.vcf.snp
```

STEP 4

```
java -jar VarScan.v2.4.1.jar somaticFilter
    somatic.snp.hc –indel-file
    somatic.indel.hc --output-file
    somatic.snp.hc.filter

java -jar VarScan.v2.4.1.jar somaticFilter
    normal_tumor.vcf.snp.Somatic.hc
    --indel-file
    normal_tumor.vcf.indel.Somatic.hc
    --output-file
    normal_tumor.vcf.Somatic.hc.filter
```

Somatic variant significance

The probability of observing a particular somatic variant by chance can be calculated using different procedure.
VarScan2 uses Fisher's exact test where the background distribution corresponding to threads in the normal sample.

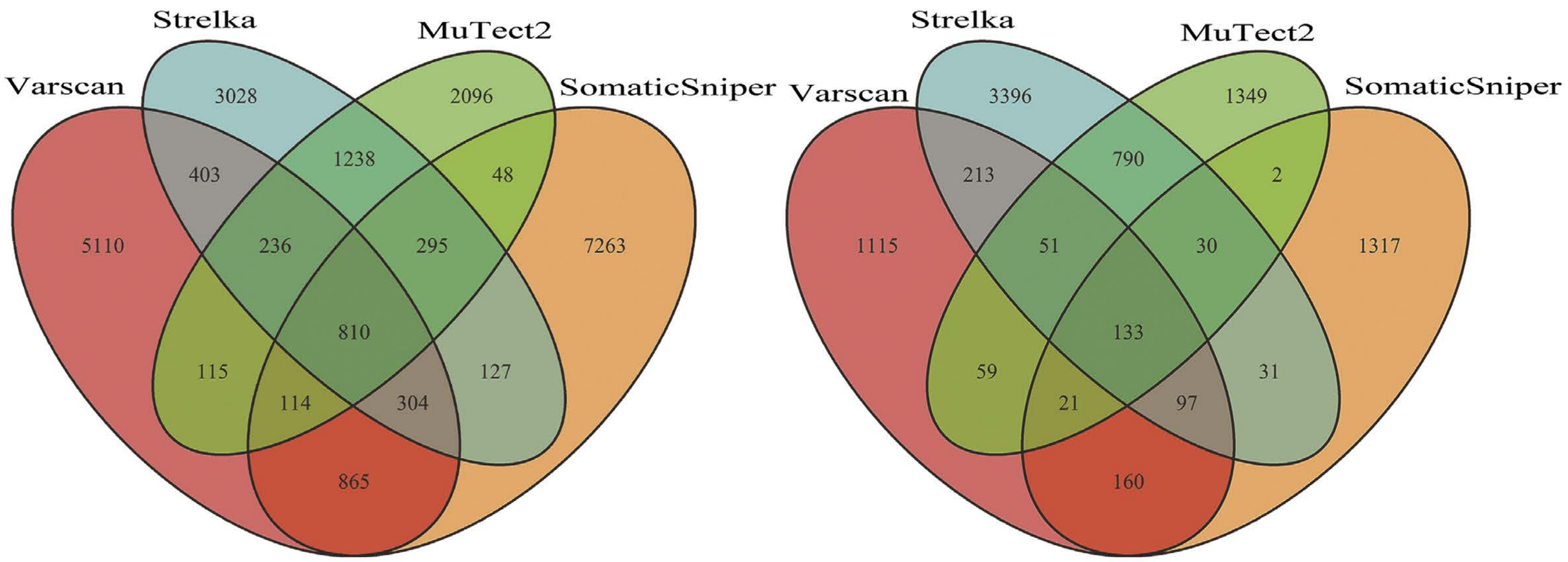
```
CHROM: chr17
POS: 7674221
ID: .
REF: G
ALT: A
QUAL: .
FILTER: PASS
INFO: DP=102;SOMATIC;SS=2;SSC=58;GPV=1E0;SPV=1.4006E-6
FORMAT: ADP=94;WT=0;HET=1;HOM=0;NC=0
FORMAT: GT:GQ:DP:RD:AD:FREQ:DP4
NORMAL: 0/0:.:47:47:0:0%:31,16,0,0
TUMOR: 0/1:.:55:36:19:34.55%:26,10,12,7
```

Contingency Table

	A	G
TUMOR	36	19
NORMAL	47	0

Variant callers survey

A survey of four somatic variant callers revealed that only a little fraction of detected variants are in common among methods



WES-Seq

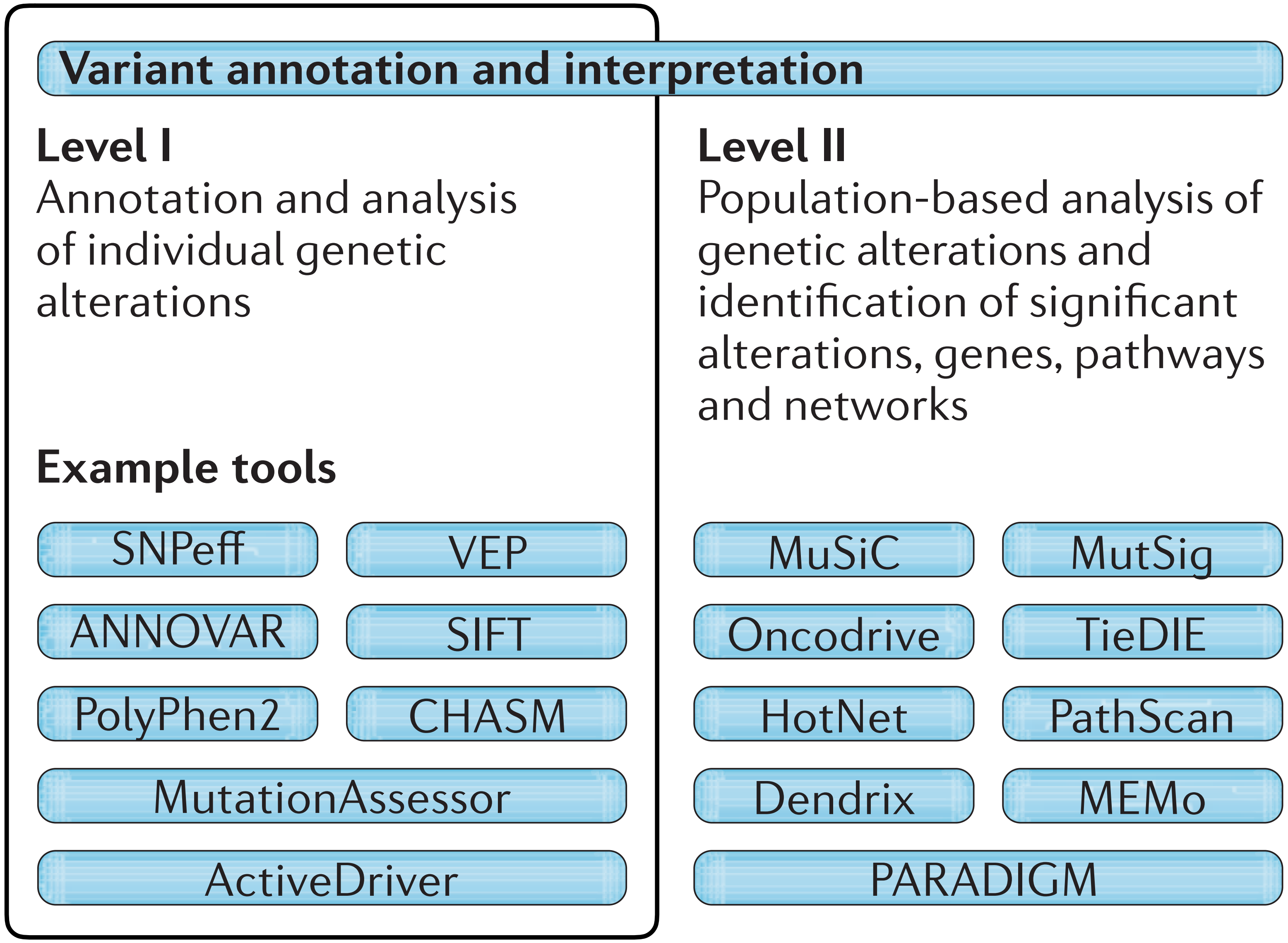
UDT-Seq

Cai et al (2016). Scientific Reports 6,: 36540 .

Short variant annotation and interpretation

Annotation and interpretation

Annotation define the effect of the variants and its location.
Variant interpretation consists in predicting its functional/phenotypic effect



Variant annotation and interpretation

Level I
Annotation and analysis of individual genetic alterations

Example tools

SNPeff, VEP, ANNOVAR, SIFT, PolyPhen2, CHASM, MutationAssessor, ActiveDriver

Level II
Population-based analysis of genetic alterations and identification of significant alterations, genes, pathways and networks

MuSiC, MutSig, Oncodrive, TieDIE, HotNet, PathScan, Dendrix, MEMo, PARADIGM

Aims of variant annotation

- Identify the gene(s) that overlaps with the variant
- Determine whether the variant is located in an exon
- Determine whether the variant is located in the coding sequence
- If the variant is a SNV, determine whether the encoded amino acid is changed, if so annotate as missense
- If the variant is located right before or after an exon/intron boundary, annotate as splicing
- If the variant removes/adds nucleotides from the CDS, annotate as deletion/insertion

VEP

Variant Effect Predictor (VEP) determines the effect of your variants (SNPs, insertions, deletions, CNVs or structural variants) on genes, transcripts, and protein sequence, as well as regulatory regions.

```
vep -i vcf_file -o annotated_vcf_–symbol --canonical --force
    --vcf --af  --offline --dir /nfs/vep/
```

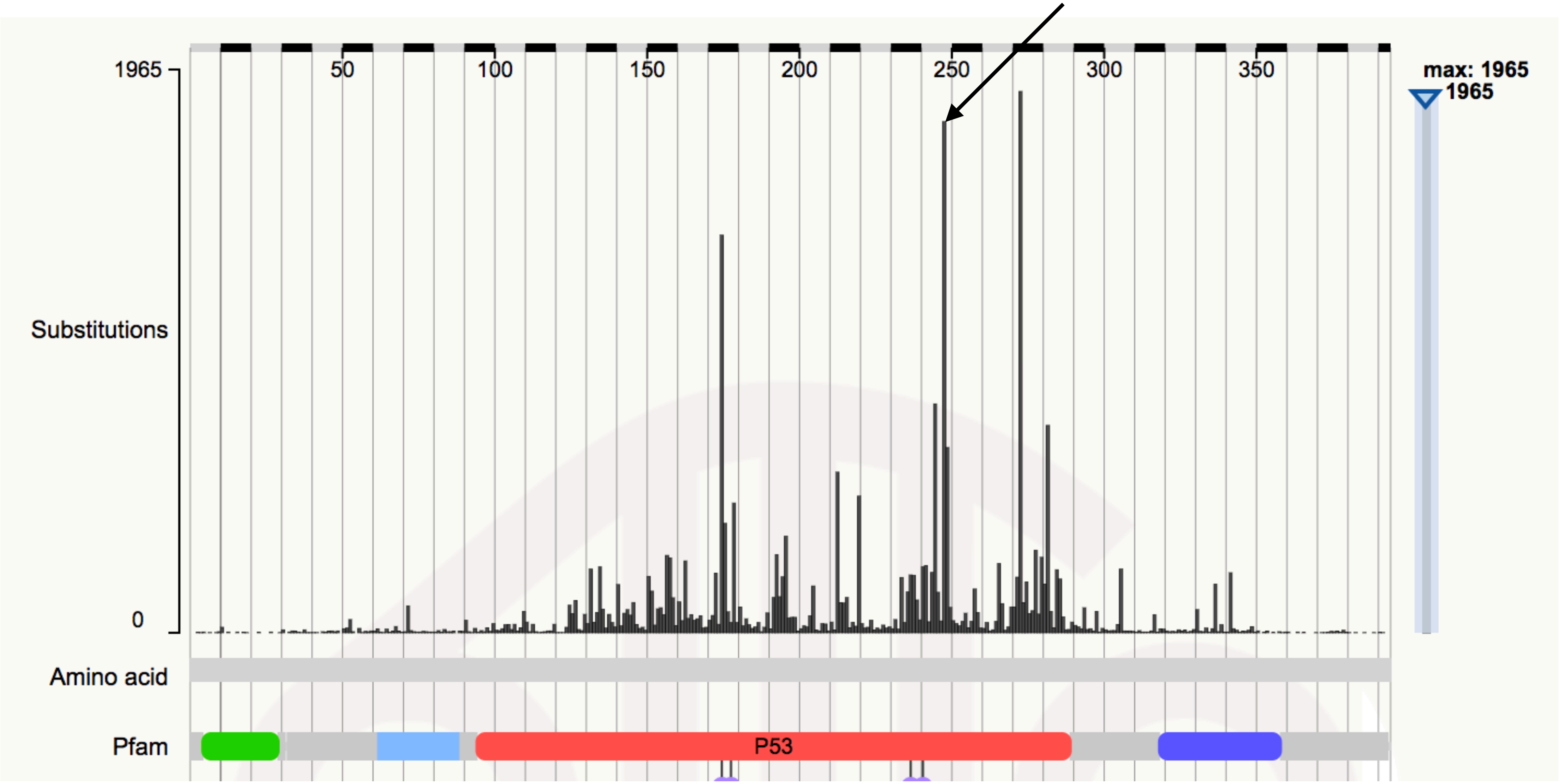
```
vep -i normal_tumor.vcf.snp.Somatic.hc.filter
    -o normal_tumor.vcf.snp.Somatic.hc.filter.vep  --symbol
    --canonical --force --vcf --af --offline --dir /nfs/vep
```

Looking at the VCF output, find out what is the effect of SNV in chromosome 17, position 7,674,221 from G to A.

COSMIC

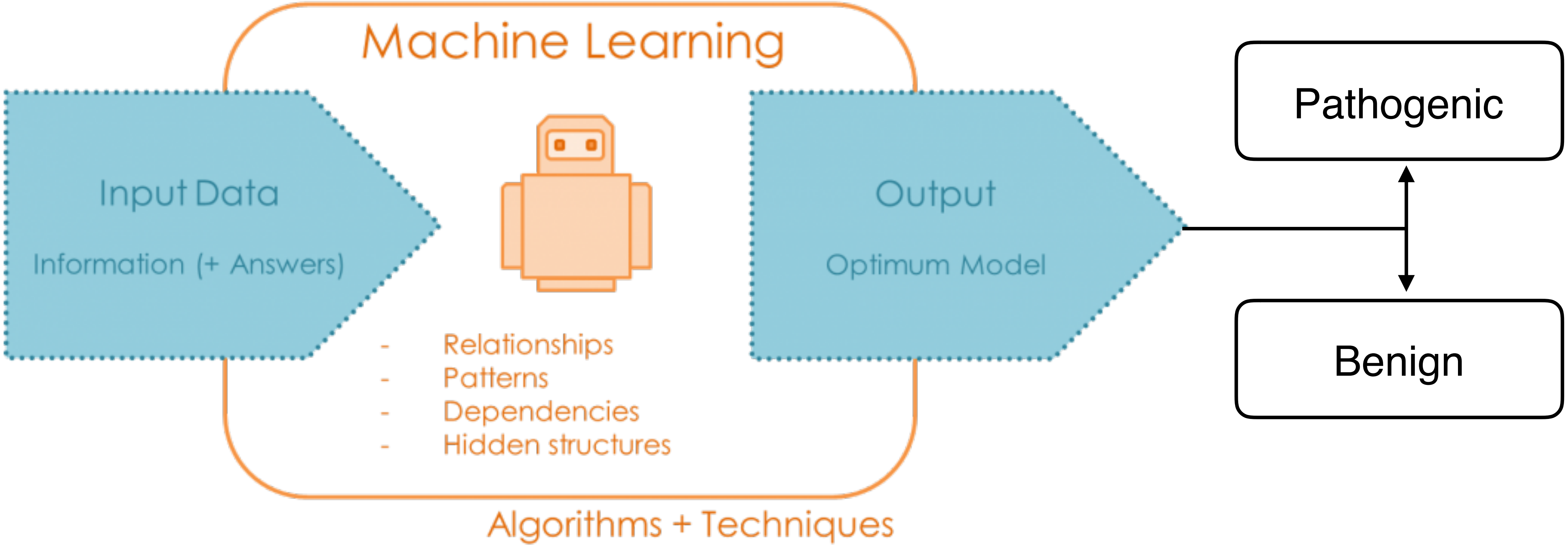
The Catalog of Somatic mutations in cancer (COSMIC) is the world's largest and most comprehensive resource for exploring the impact of somatic mutations in human cancer.

Mutations in position 248



Variant interpretation

Usually based learning algorithm which takes in input features associated to the variants and returns a probability for the variant to be Pathogenic or Benign



Machine Learning

Input Data

Information (+ Answers)

- Relationships
- Patterns
- Dependencies
- Hidden structures

Output

Optimum Model

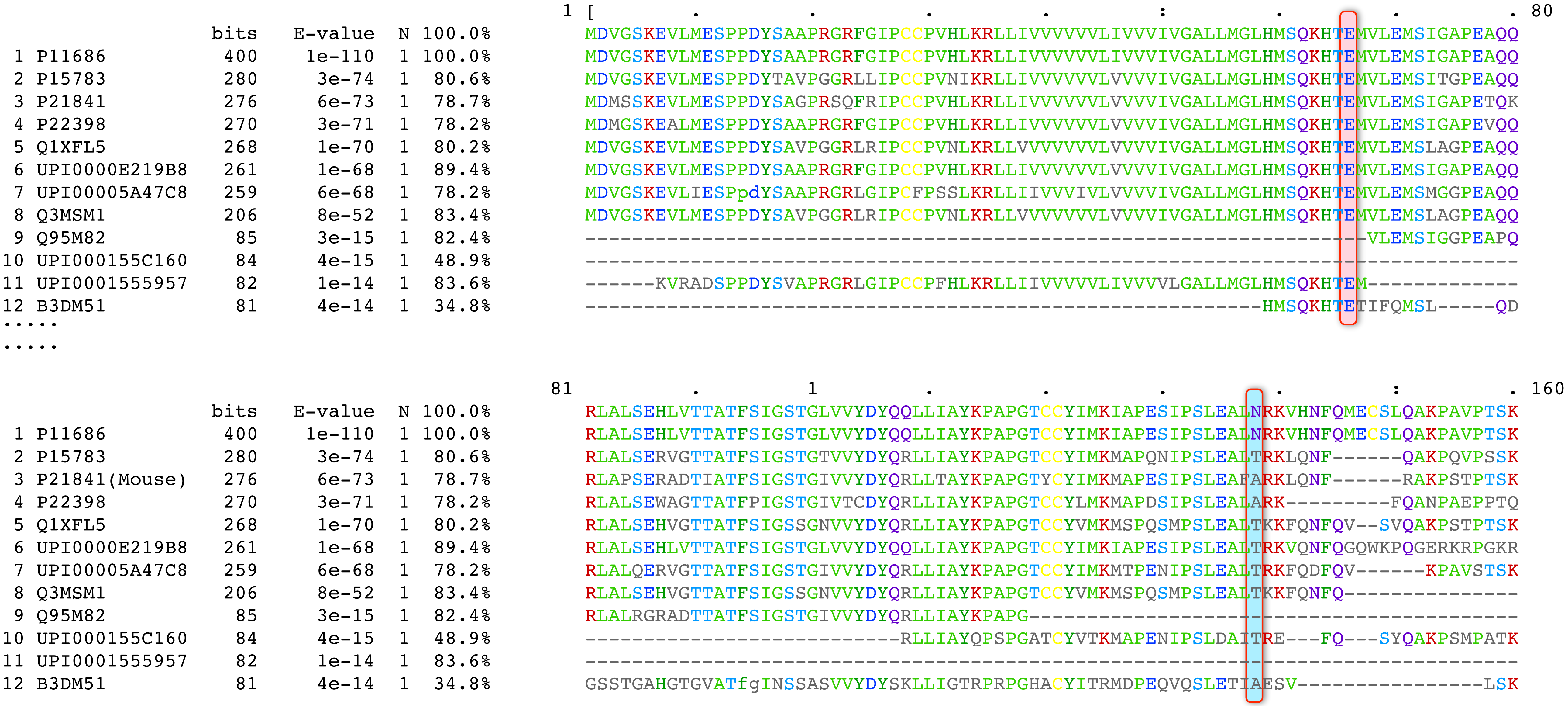
Pathogenic

Benign

Algorithms + Techniques

Conserved or not?

In positions 66 the Glutamic acid is highly conserved Asparagine in position 138 is mutated Threonine or Alanine

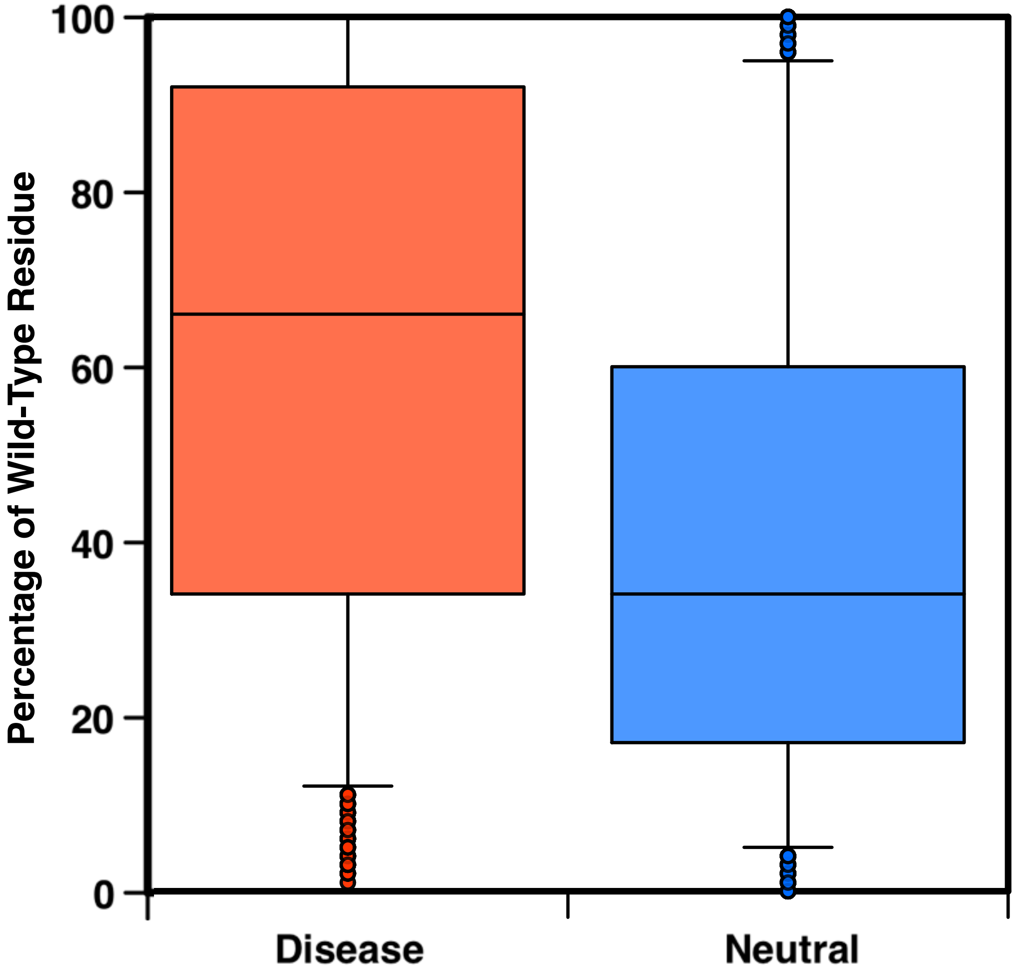


```
                                             1 [        .         .         .         .         :         .         .         . 80
                   bits    E-value  N 100.0%   MDVGSKEVLMESPPDYSAAPRGRFGIPCCPVHLKRLLIVVVVVVLIVVVIVGALLMGLHMSQKHTEMVLEMSIGAPEAQQ
 1 P11686           400     1e-110  1 100.0%   MDVGSKEVLMESPPDYSAAPRGRFGIPCCPVHLKRLLIVVVVVVLIVVVIVGALLMGLHMSQKHTEMVLEMSIGAPEAQQ
 2 P15783           280      3e-74  1  80.6%   MDVGSKEVLMESPPDYTAVPGGRLLIPCCPVNIKRLLIVVVVVVLVVVVIVGALLMGLHMSQKHTEMVLEMSITGPEAQQ
 3 P21841           276      6e-73  1  78.7%   MDMSSKEVLMESPPDYSAGPRSQFRIPCCPVHLKRLLIVVVVVVLVVVVIVGALLMGLHMSQKHTEMVLEMSIGAPETQK
 4 P22398           270      3e-71  1  78.2%   MDMGSKEALMESPPDYSAAPRGRFGIPCCPVHLKRLLIVVVVVVLVVVVIVGALLMGLHMSQKHTEMVLEMSIGAPEVQQ
 5 Q1XFL5           268      1e-70  1  80.2%   MDVGSKEVLMESPPDYSAVPGGRLRIPCCPVNLKRLLVVVVVVVLVVVVIVGALLMGLHMSQKHTEMVLEMSLAGPEAQQ
 6 UPI0000E219B8    261      1e-68  1  89.4%   MDVGSKEVLMESPPDYSAAPRGRFGIPCCPVHLKRLLIVVVVVVLVVVVIVGALLMGLHMSQKHTEMVLEMSIGAPEAQQ
 7 UPI00005A47C8    259      6e-68  1  78.2%   MDVGSKEVLIESPpdYSAAPRGRLGIPCFPSSLKRLLIIVVVIVLVVVVIVGALLMGLHMSQKHTEMVLEMSMGGPEAQQ
 8 Q3MSM1           206      8e-52  1  83.4%   MDVGSKEVLMESPPDYSAVPGGRLRIPCCPVNLKRLLVVVVVVVLVVVVIVGALLMGLHMSQKHTEMVLEMSLAGPEAQQ
 9 Q95M82            85      3e-15  1  82.4%   -------------------------------------------------------------------VLEMSIGGPEAPQ
10 UPI000155C160     84      4e-15  1  48.9%   --------------------------------------------------------------------------------
11 UPI0001555957     82      1e-14  1  83.6%   ------KVRADSPPDYSVAPRGRLGIPCCPFHLKRLLIIVVVVVLIVVVVLGALLMGLHMSQKHTEM-------------
12 B3DM51            81      4e-14  1  34.8%   ----------------------------------------------------------HMSQKHTETIFQMSL-----QD
.....
.....

                                            81          .         1         .         .         .         .         :         . 160
                   bits    E-value  N 100.0%   RLALSEHLVTTATFSIGSTGLVVYDYQQLLIAYKPAPGTCCYIMKIAPESIPSLEALNRKVHNFQMECSLQAKPAVPTSK
 1 P11686           400     1e-110  1 100.0%   RLALSEHLVTTATFSIGSTGLVVYDYQQLLIAYKPAPGTCCYIMKIAPESIPSLEALNRKVHNFQMECSLQAKPAVPTSK
 2 P15783           280      3e-74  1  80.6%   RLALSERVGTTATFSIGSTGTVVYDYQRLLIAYKPAPGTCCYIMKMAPQNIPSLEALTRKLQNF------QAKPQVPSSK
 3 P21841(Mouse)    276      6e-73  1  78.7%   RLAPSERADTIATFSIGSTGIVVYDYQRLLTAYKPAPGTYCYIMKMAPESIPSLEAFARKLQNF------RAKPSTPTSK
 4 P22398           270      3e-71  1  78.2%   RLALSEWAGTTATFPIGSTGIVTCDYQRLLIAYKPAPGTCCYLMKMAPDSIPSLEALARK---------FQANPAEPPTQ
 5 Q1XFL5           268      1e-70  1  80.2%   RLALSEHVGTTATFSIGSSGNVVYDYQRLLIAYKPAPGTCCYVMKMSPQSMPSLEALTKKFQNFQV--SVQAKPSTPTSK
 6 UPI0000E219B8    261      1e-68  1  89.4%   RLALSEHLVTTATFSIGSTGLVVYDYQQLLIAYKPAPGTCCYIMKIAPESIPSLEALTRKVQNFQGQWKPQGERKRPGKR
 7 UPI00005A47C8    259      6e-68  1  78.2%   RLALQERVGTTATFSIGSTGIVVYDYQRLLIAYKPAPGTCCYIMKMTPENIPSLEALTRKFQDFQV------KPAVSTSK
 8 Q3MSM1           206      8e-52  1  83.4%   RLALSEHVGTTATFSIGSSGNVVYDYQRLLIAYKPAPGTCCYVMKMSPQSMPSLEALTKKFQNFQ---------------
 9 Q95M82            85      3e-15  1  82.4%   RLALRGRADTTATFSIGSTGIVVYDYQRLLIAYKPAPG------------------------------------------
10 UPI000155C160     84      4e-15  1  48.9%   ---------------------------RLLIAYQPSPGATCYVTKMAPENIPSLDAITRE---FQ---SYQAKPSMPATK
11 UPI0001555957     82      1e-14  1  83.6%   --------------------------------------------------------------------------------
12 B3DM51            81      4e-14  1  34.8%   GSSTGAHGTGVATfgINSSASVVYDYSKLLIGTRPRPGHACYITRMDPEQVQSLETIAESV----------------LSK
```

Sequence profile

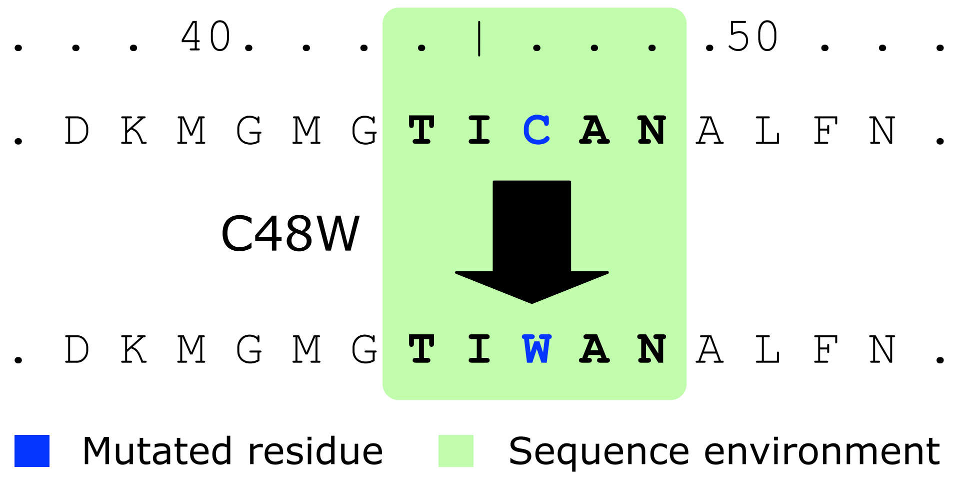
The protein sequence profile is calculated running BLAST on the UniRef90 dataset and selecting only the hits with e-value < 10^{-9}.

The frequency distributions of the wild-type residues for disease-related and neutral variants are significantly different (KS p-value=0).

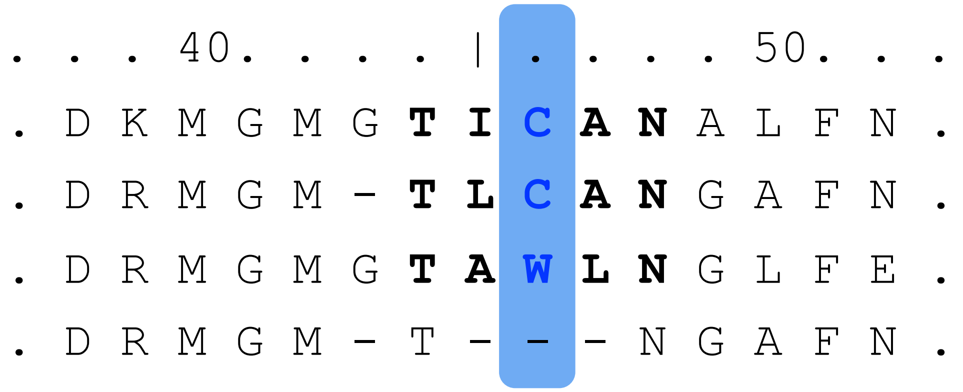


Capriotti et al (2012). Briefings in Bioinformatics. 13; 495-512.

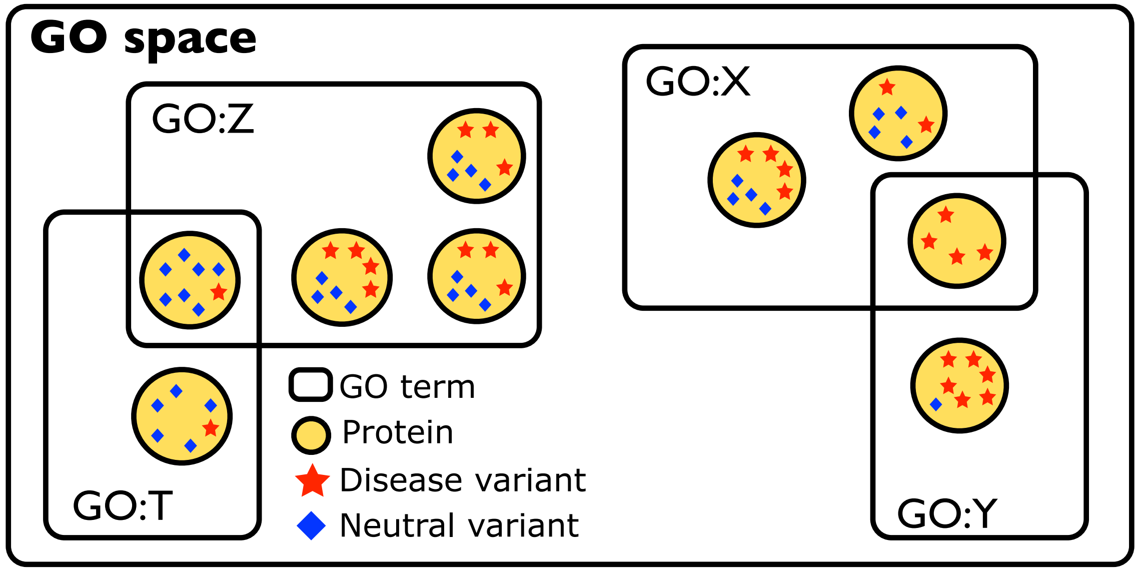
SNPs&GO input features



Sequence information is encoded in 2 vectors each one composed by 20 elements. The first vector encodes for the mutation and the second one for the sequence environment


. . . 40. . . . | . . . . 50. . .
. D K M G M G T I C A N A L F N .
. D R M G M – T L C A N G A F N .
. D R M G M G T A W L N G L F E .
. D R M G M – T – – – N G A F N .

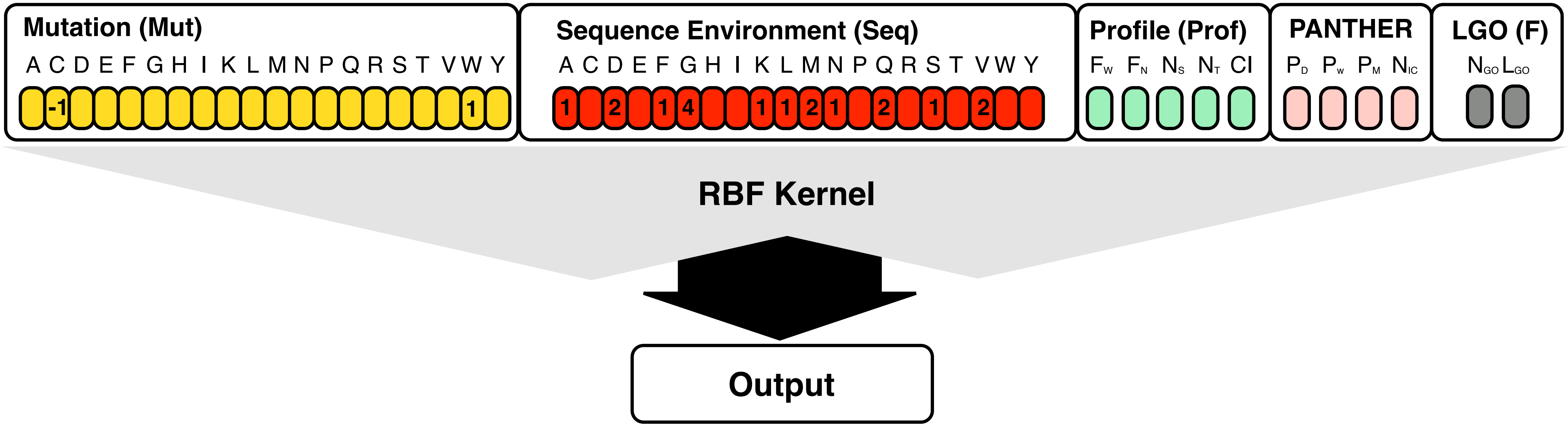
Protein sequence profile information derived from a multiple sequence alignment. It is encoded in a 5 elements vector corresponding to different features general and local features



The GO information are encoded in a 2 elements vector corresponding to the number unique of GO terms associated to the protein sequences and the sum of the logarithm of the total number of disease-related and neutral variants for each GO term.

SNPs&GO performance

SNPs&GO results in better performance with respect to previously developed methods.


Mutation (Mut)
A C D E F G H I K L M N P Q R S T V W Y
-1
1
Sequence Environment (Seq)
A C D E F G H I K L M N P Q R S T V W Y
1 2 1 4 1 1 2 1 2 1 2
Profile (Prof)
F_W F_N N_S N_T CI
PANTHER
P_D P_W P_M N_{IC}
LGO (F)
N_{GO} L_{GO}
RBF Kernel
Output

Method	Q2	P[D]	Q[D]	P[N]	Q[N]	C	PM
PolyPhen	0.71	0.76	0.75	0.63	0.64	0.39	58
SIFT	0.76	0.75	0.76	0.77	0.75	0.52	93
PANTHER	0.74	0.77	0.73	0.71	0.76	0.48	76
SNPs&GO	0.82	0.83	0.78	0.80	0.85	0.63	100

D = Disease related N = Neutral

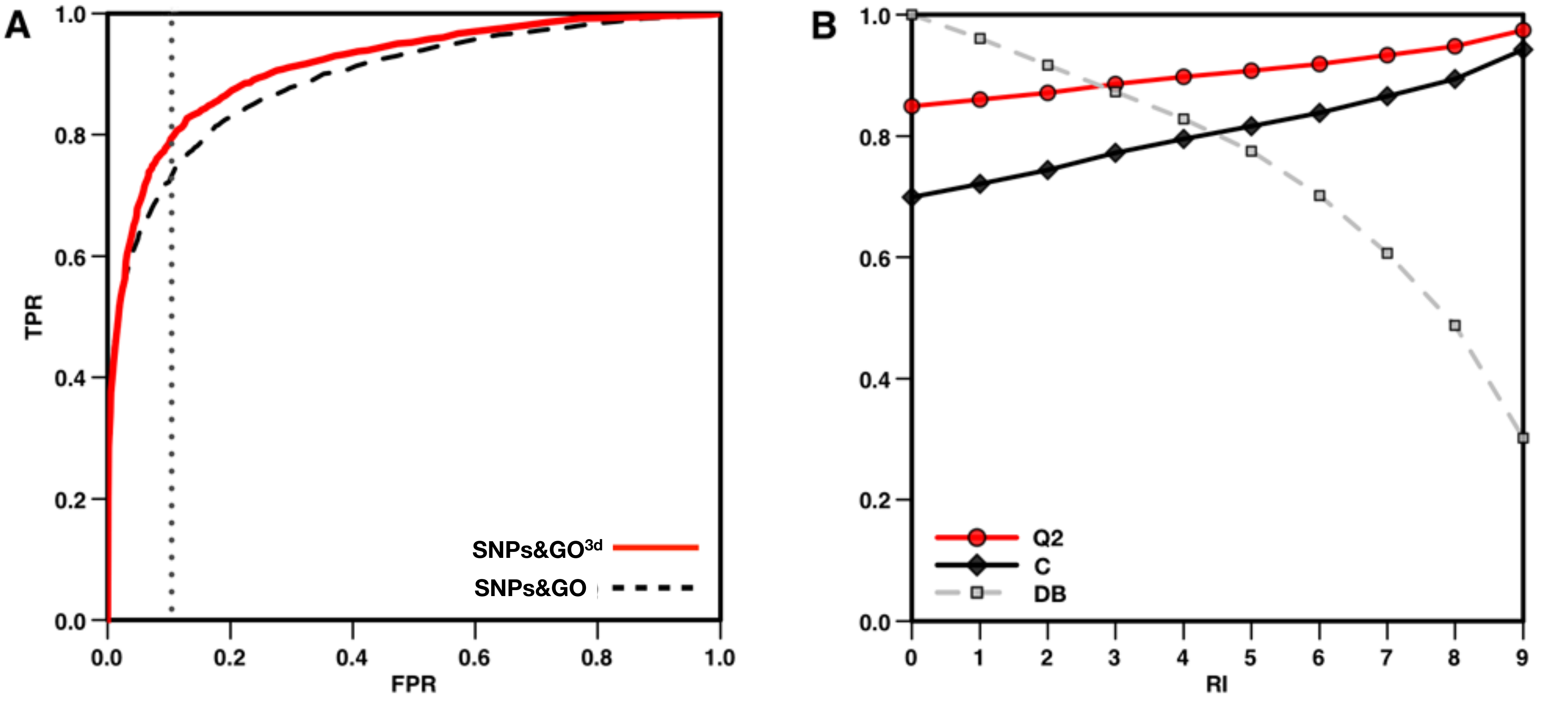
DB= 33672 nsSNVs

Calabrese et al. (2009) Human Mutation 30, 1237-1244.

Sequence vs Structure

The structure-based method results in better accuracy with respect to the sequence-based one. Structure based prediction are 3% more accurate and correlation coefficient increases of 0.06. If 10% of FP are accepted the TPR increases of 7%.

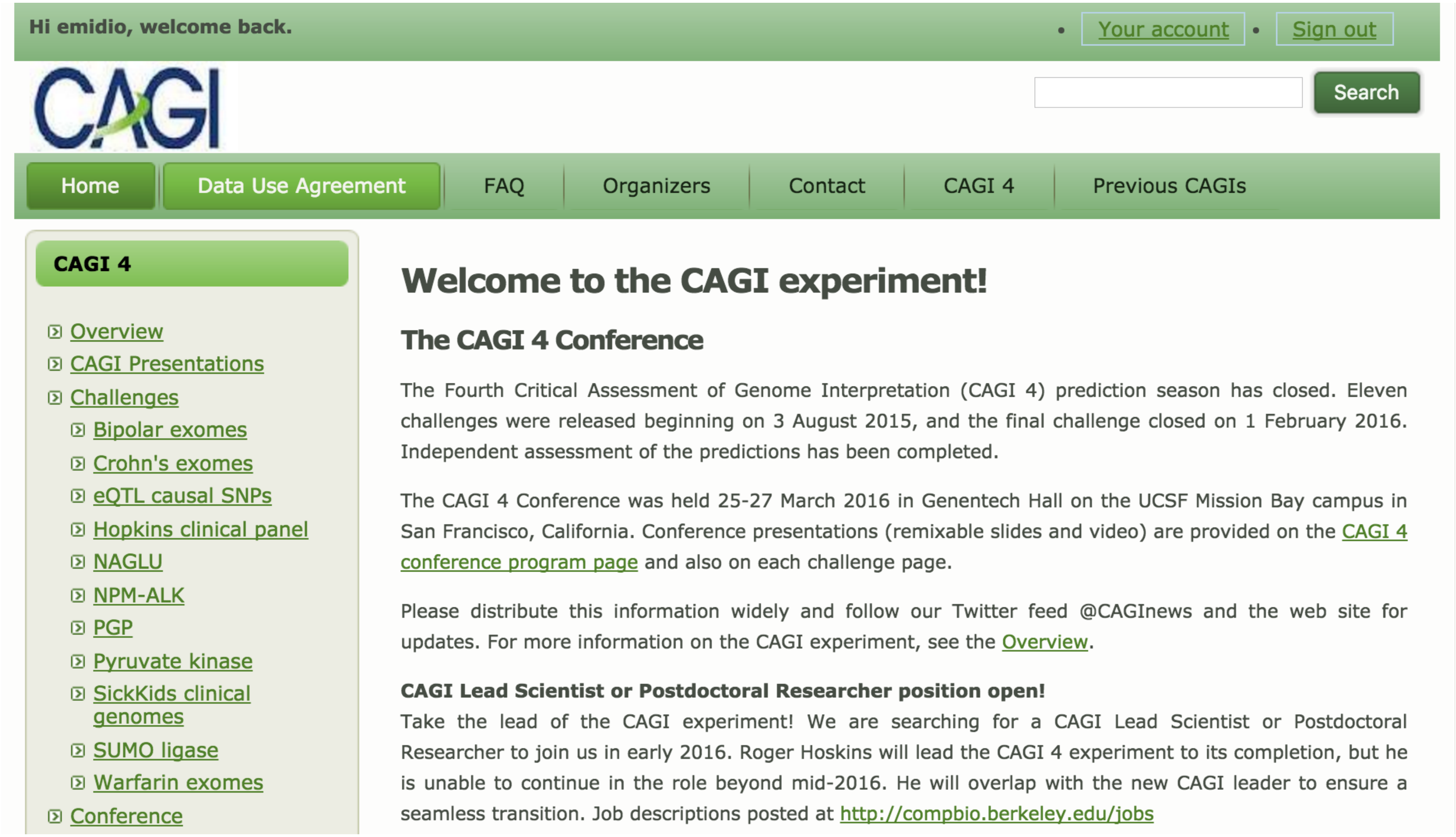
	Q2	P[D]	S[D]	P[N]	S[N]	C	AUC
SNPs&GO	0.82	0.81	0.83	0.82	0.81	0.64	0.89
SNPs&GO^{3d}	0.85	0.84	0.87	0.86	0.83	0.70	0.92



http://snps.biofold.org/snps-and-go

CAGI experiments

The Critical Assessment of Genome Interpretation is a community experiment to objectively assess computational methods for predicting the phenotypic impacts of genomic variation.



Welcome to the CAGI experiment!

The CAGI 4 Conference

The Fourth Critical Assessment of Genome Interpretation (CAGI 4) prediction season has closed. Eleven challenges were released beginning on 3 August 2015, and the final challenge closed on 1 February 2016. Independent assessment of the predictions has been completed.

The CAGI 4 Conference was held 25-27 March 2016 in Genentech Hall on the UCSF Mission Bay campus in San Francisco, California. Conference presentations (remixable slides and video) are provided on the CAGI 4 conference program page and also on each challenge page.

Please distribute this information widely and follow our Twitter feed @CAGInews and the web site for updates. For more information on the CAGI experiment, see the Overview.

CAGI Lead Scientist or Postdoctoral Researcher position open!
Take the lead of the CAGI experiment! We are searching for a CAGI Lead Scientist or Postdoctoral Researcher to join us in early 2016. Roger Hoskins will lead the CAGI 4 experiment to its completion, but he is unable to continue in the role beyond mid-2016. He will overlap with the new CAGI leader to ensure a seamless transition. Job descriptions posted at http://compbio.berkeley.edu/jobs

https://genomeinterpretation.org/

The P16 challenge

CDKN2A is the most common, high penetrance, susceptibility gene identified to date in familial malignant melanoma. p16^{INK4A} is one of the two oncosuppressor which promotes cell cycle arrest by inhibiting cyclin dependent kinase (CDK4/6).

Challenge: Evaluate how different variants of p16 protein impact its ability to block cell proliferation.

Provide a number between 50% that represent the normal proliferation rate of control cells and 100% the maximum proliferation rate in case cells.

SNPs&GO prediction

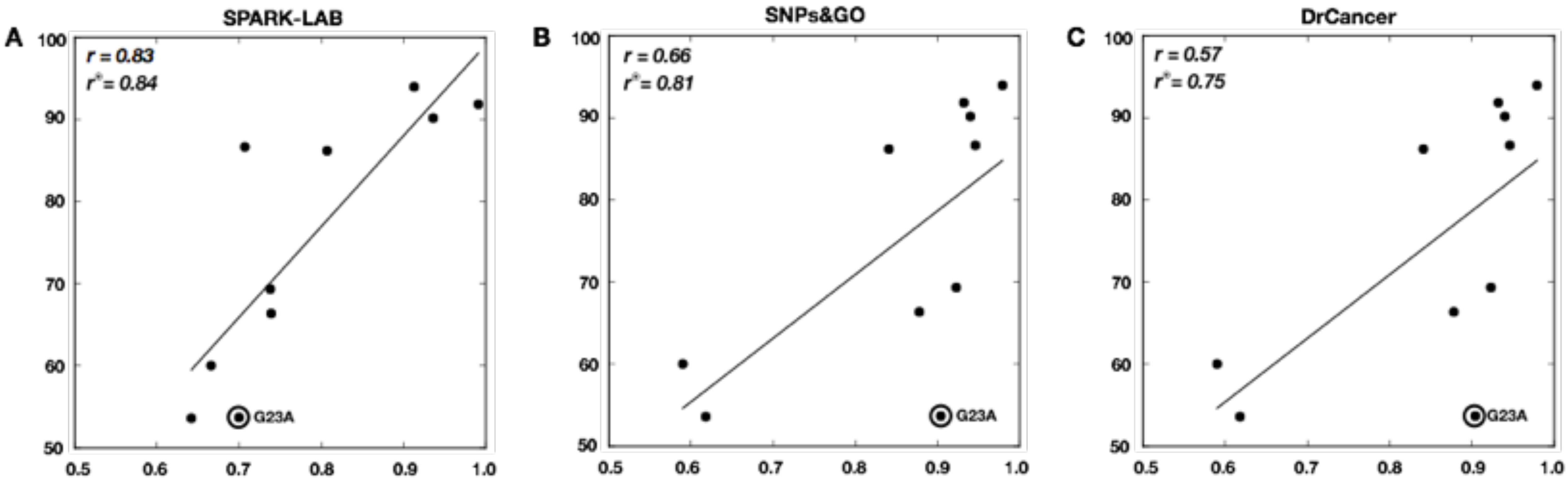
Proliferation rates predicted using the output of SNPs&GO without any optimization.

Variant	Prediction	Real	Δ	%WT	%MUT
G23R	0.932	0.918	0.014	84	0
G23S	0.923	0.693	0.230	84	1
G23V	0.940	0.901	0.039	84	0
G23A	0.904	0.537	0.367	84	2
G23C	0.946	0.866	0.080	84	0
G35E	0.590	0.600	0.010	12	14
G35W	0.841	0.862	0.021	12	0
G35R	0.618	0.537	0.081	12	4
L65P	0.878	0.664	0.214	15	1
L94P	0.979	0.939	0.040	56	0

P16 predictions

SNPs&GO resulted among the best methods for predicting the impact of P16INK4A variants on cell proliferation.

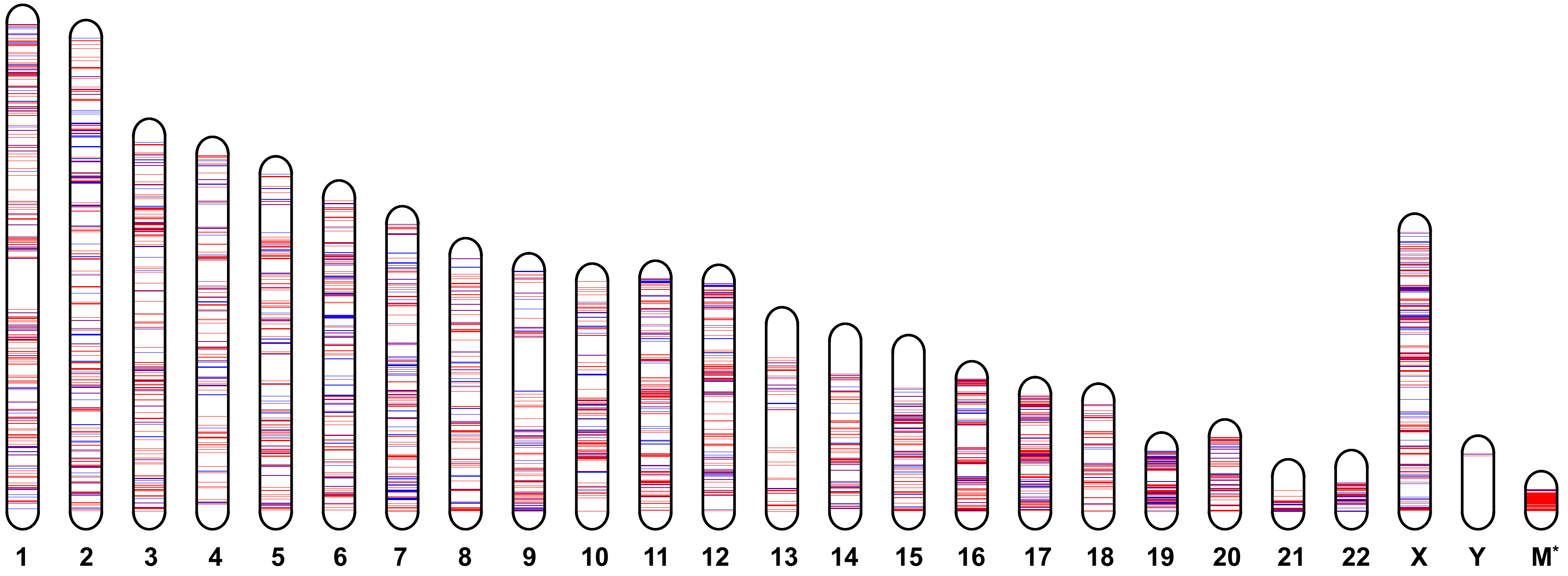
Method	Q2	AUC	MC	RMSE	$r_{Pearson}$	$r_{Spearman}$	$r_{KendallTau}$
SPARK-LAB	0.900	0.920	0.816	0.30	0.595	0.619	0.443
SNPs&GO	0.700	0.880	0.500	0.33	0.575	0.616	0.445
DrCancer	0.600	0.840	0.333	0.46	0.477	0.495	0.409



Capriotti et al. (2017) Human Mutations. PMID: 28102005.

Whole-genome predictions

Most of the genetic variants occur in non-coding region that represents >98% of the whole genome.

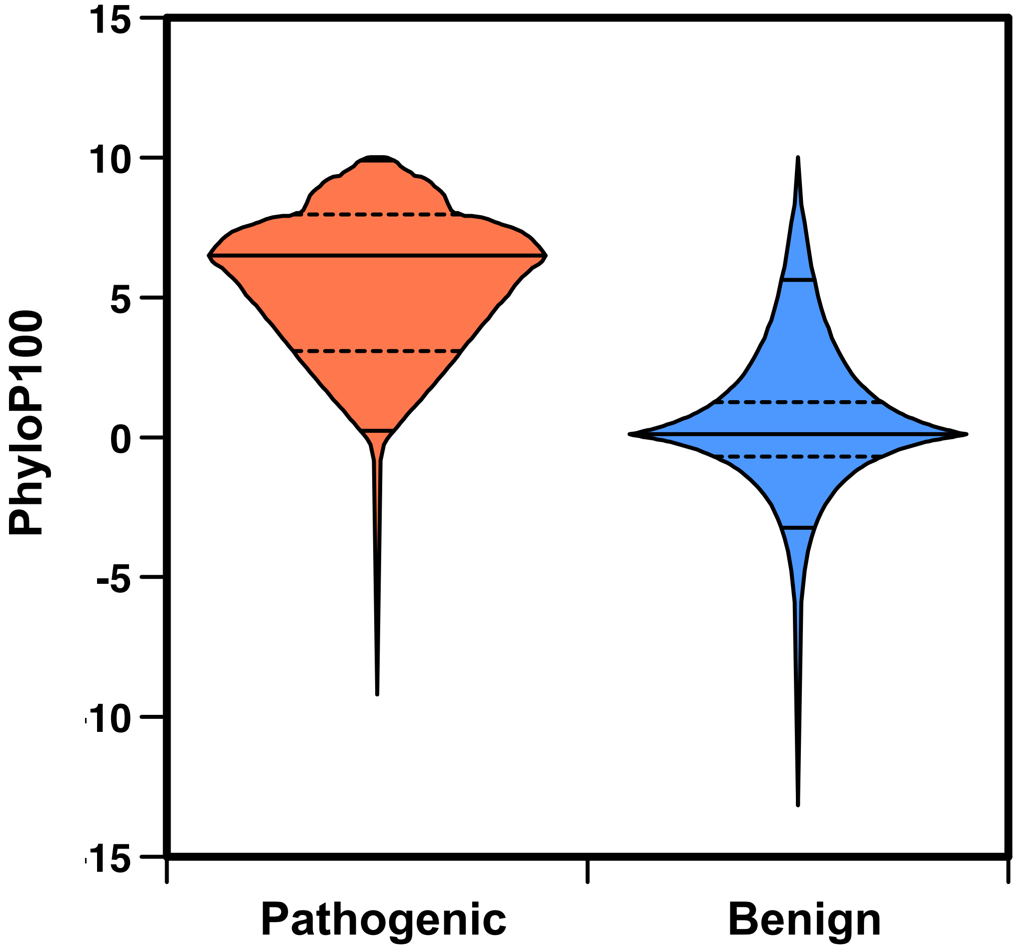


Predict the effect of SNVs in non-coding region is a challenging task because conservation is more difficult to estimate.

Sequence alignment is more complicated for sequences from non-coding regions.

PhyloP100 score

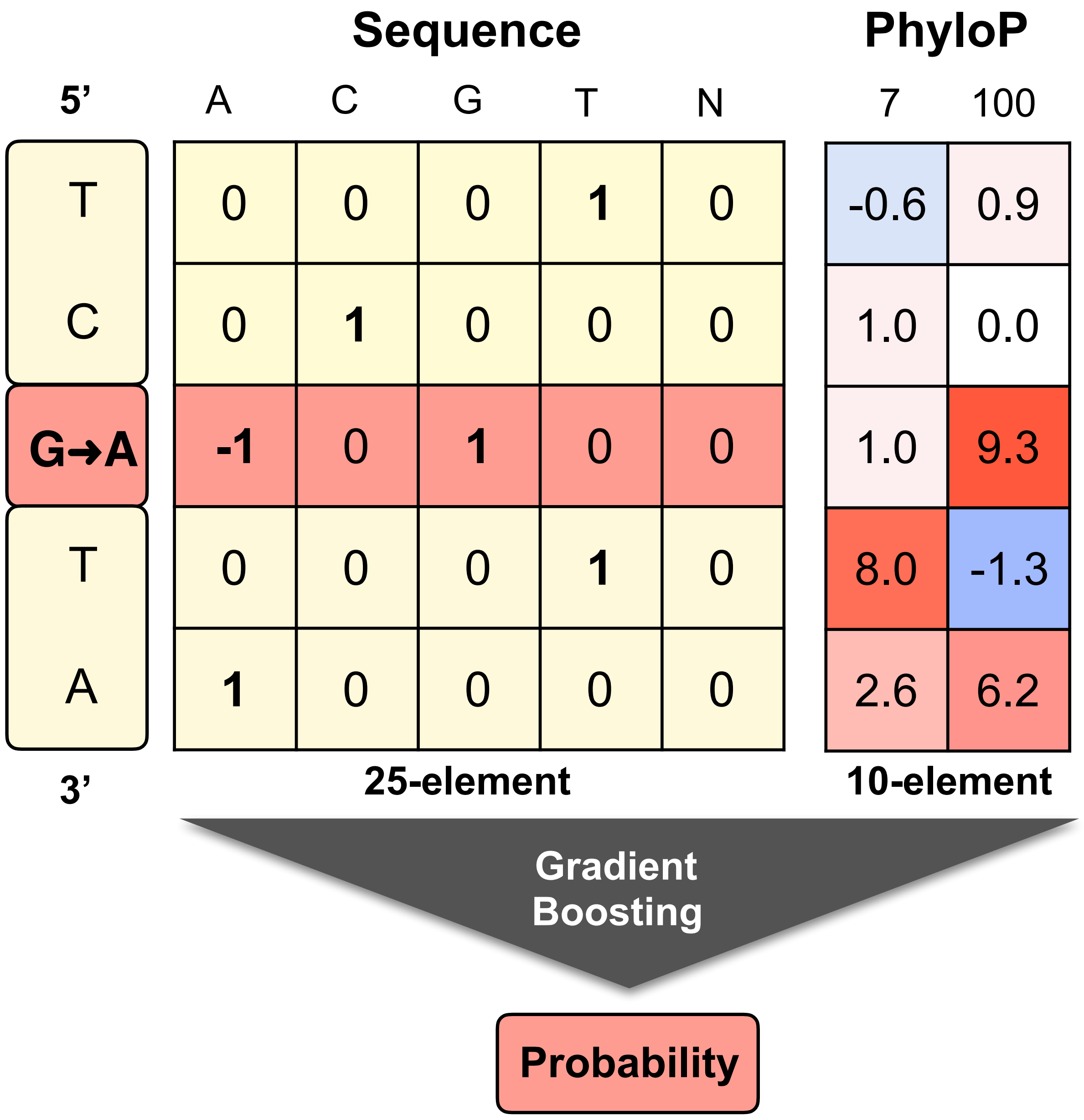
Conservation analysis based on the pre-calculated score available at the UCSC revealed a significant difference between the distribution of the PhyloP100 scores in Pathogenic and Benign SNVs.



PhyloP: Pollard et al,, Genome Research 2010

PhD-SNP^g

PhD-SNP^g is a simple method that takes in input 35 sequence-based features from a window of 5 nucleotides around the mutated position.



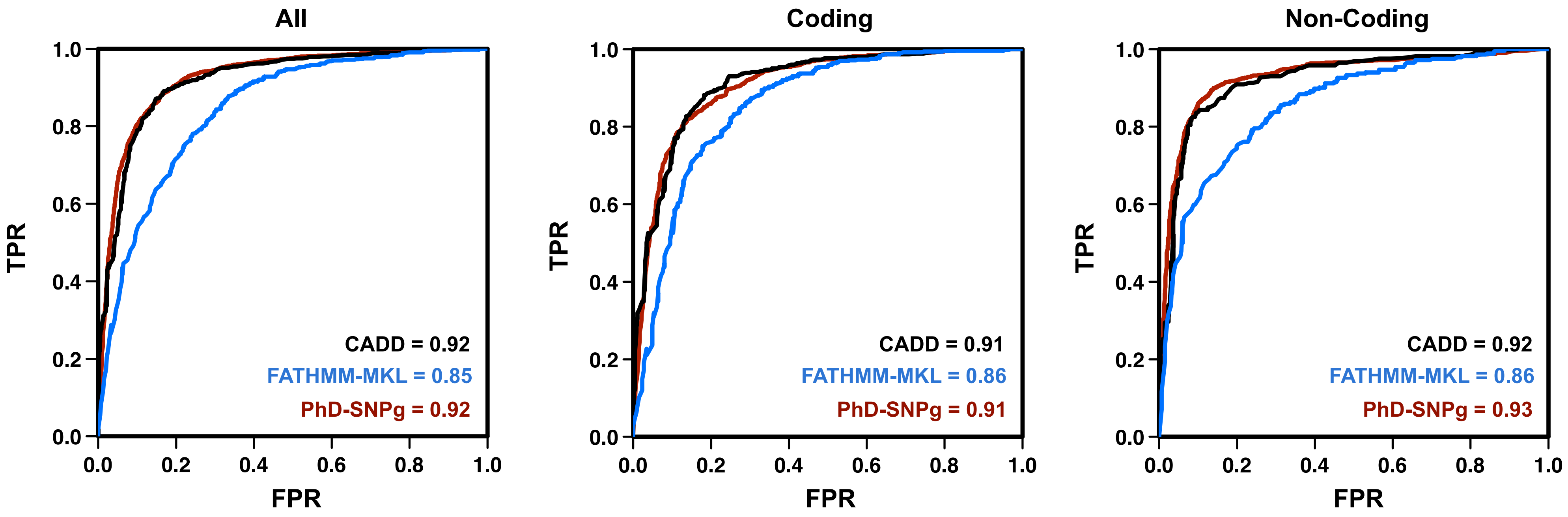
5’	Sequence A	C	G	T	N	PhyloP 7	100
T	0	0	0	**1**	0	-0.6	0.9
C	0	**1**	0	0	0	1.0	0.0
G→A	**-1**	0	**1**	0	0	1.0	9.3
T	0	0	0	**1**	0	8.0	-1.3
A	**1**	0	0	0	0	2.6	6.2
3’	25-element					10-element	

http://snps.biofold.org/phd-snpg/

Benchmarking

PhD-SNPg has been tested in cross-validation on a set of 35,802 SNVs and on a blind set of 1,408 variants recently annotated.

	Q2	TNR	NPV	TPR	PPV	MCC	F1	AUC
PhD-SNPg	0.861	0.774	0.884	0.925	0.847	0.715	0.884	0.924
Coding	0.849	0.671	0.845	0.938	0.850	0.651	0.892	0.908
Non-Coding	0.876	0.855	0.911	0.901	0.839	0.753	0.869	0.930



Capriotti and Fariselli. (2017) Nucleic Acids Res. PMID: 28482034.

Mutation rates

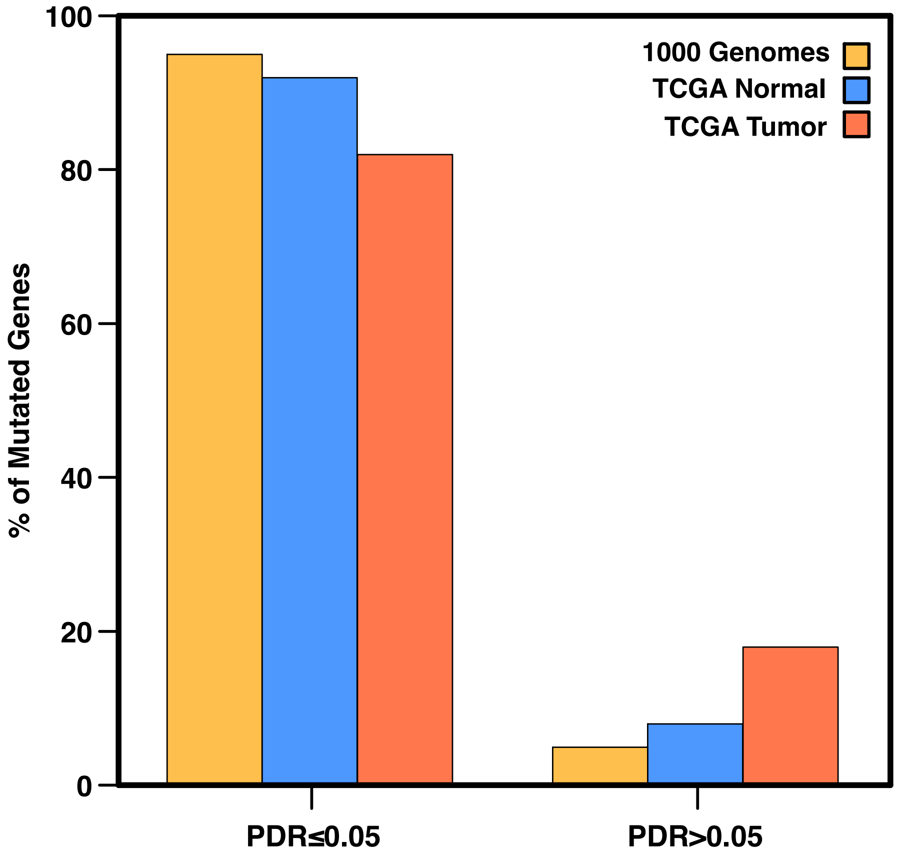
The analysis of 1000 Genomes, The Cancer Genome Atlas (TCGA) normal and tumor samples shows an increasing number of genes with rare nonsynonymous SNVs.

Cohort	%Genes PDR≤0.05	%Genes PDR>0.05
1000 Genomes	95%	5%
TCGA Normal	92%	8%
TCGA Tumor	82%	18%

Tumor = Colon Adenocarcinoma

PDR = Gene Putative Defective Rate

Fraction of samples in which a gene has ≥1 nonsynonymous variant with MAF≤0.5%

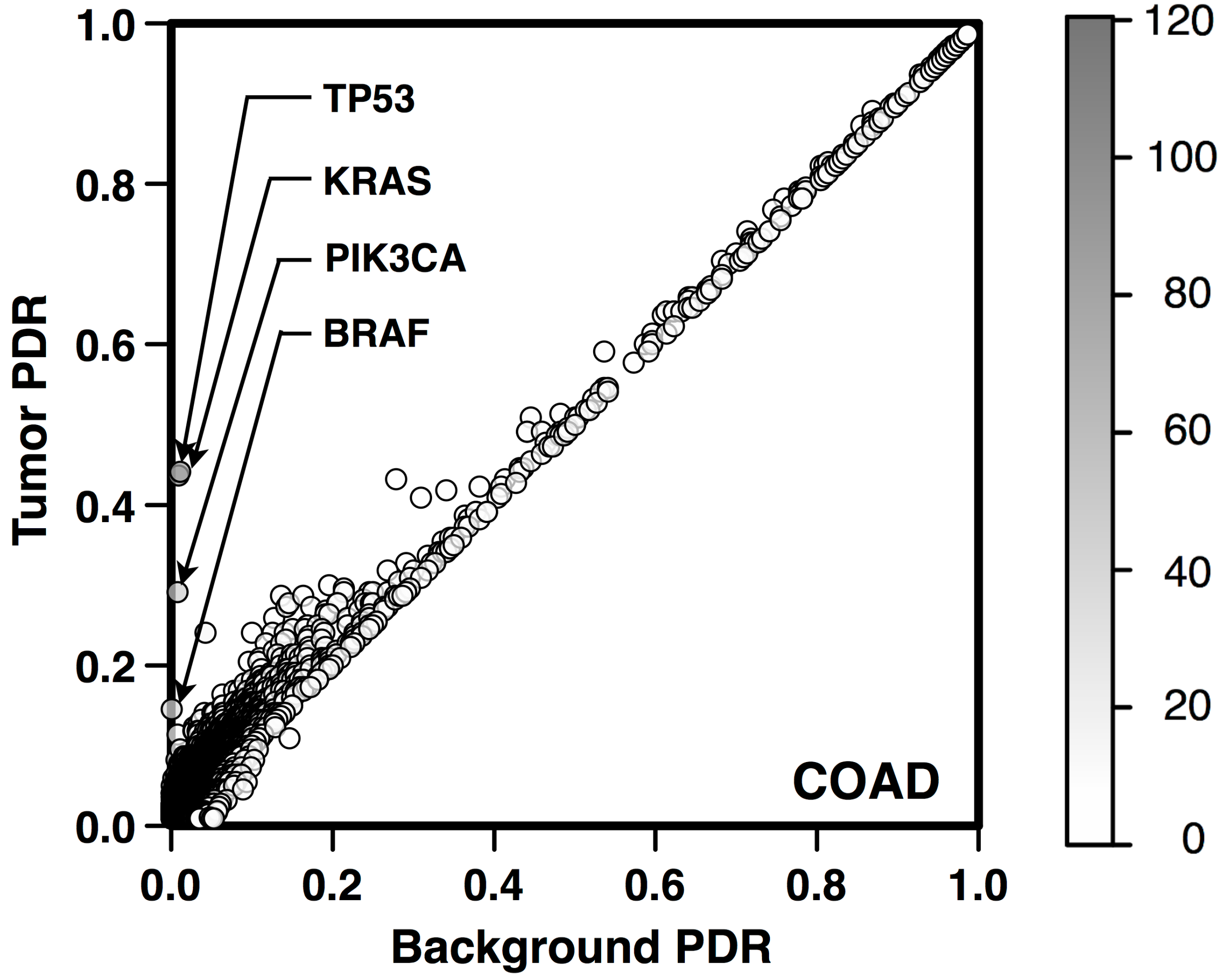


Gene prioritization

New method for cancer gene prioritization based on the comparison of the mutation rates in tumor samples vs normal and 1000 Genomes samples.

Gene	PDR[T]	PDR[B]	Score
KRAS	0.436	0.009	72.6
TP53	0.441	0.011	63.7
PIK3CA	0.291	0.007	39.4
BRAF	0.146	0.001	29.9

Colon Adenocarcinoma
PDR[T] = Putative Defective Rate Tumor
PDR[B] = Putative Defective Rate Background
Background = Max (Normal and 1000 Genomes)



Tian R, Basu M, Capriotti E (2014). Bioinformatics. 30: i572-i578

Acknowledgments

Structural Genomics @CNAG
Marc A. Marti-Renom
Davide Bau
David Dufour
Francois Serra

Computational Biology and Bioinformatics Research Group (UIB)
Jairo Rocha

Division of Informatics at UAB
Rui Tian
Shivani Viradia
Malay Basu
Diego E. Penha

Helix Group (Stanford University)
Russ B. Altman
Jennifer Lahti
Tianyun Liu
Grace Tang

Bologna Biocomputing Group
Rita Casadio
Pier Luigi Martell
Giuseppe Profiti
Castrense Savojardo
University of Padova
Piero Fariselli
University of Camerino
Mario Compiani

Mathematical Modeling of Biological Systems (University of Düsseldorf)
Markus Kollmann

Other Collaborations
Yana Bromberg, Rutger University, NJ
Francisco Melo, Universidad Catolica, Chile
Sean Mooney, Buck Institute, Novato
Cedric Notredame, CRG Barcelona
Gustavo Parisi, Univesidad de Quilmes
Frederic Rousseau, KU Leuven
Joost Schymkowitz, KU Leuven

FUNDING
Startup funding Dept. of Pathology UAB
NIH:3R00HL111322-04S1 Co-Investigator
EMBO Short Term Fellowship
Marie Curie International Outgoing Grant
Marie Curie Reintegration Grant
Marco Polo Research Project
BIOSAPIENS Network of Excellence
SPINNER Consortium

Biomolecules, Folding and Disease

http://biofold.org/

References

Hanahan D, Weinberg RA. (2011). Hallmarks of cancer: the next generation. Cell. 144: 646-74. PMID:21376230

Vogelstein B, et al. (2013). Cancer genome landscapes. Science. 339:1546-58. PMID: 23539594

Ding L, et al. (2014). Expanding the computational toolbox for mining cancer genomes. Nat Rev Genet. 15: 556-70. PMID: 25001846;

Raphael BJ, et al. (2014). Identifying driver mutations in sequenced cancer genomes: computational approaches to enable precision medicine. Genome Med. 6:5. PMID: 24479672

Lawrence MS, et al. (2014). Discovery and saturation analysis of cancer genes across 21 tumour types. Nature. 505:495-501. PMID: 24390350

Lawrence MS et al. (2013). Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature. 499: 214-8. PMID: 23770567

Khurana E, et al. (2013) Integrative annotation of variants from 1092 humans: application to cancer genomics. Science. 342:1235587.PMID: 24092746

Fernald GH, Capriotti E, Daneshjou R, Karczewski KJ, Altman RB. (2011). Bioinformatics challenges for personalized medicine. Bioinformatics. 27; 1741-1748. PMID: 21596790

Tian R, Basu MK, Capriotti E. (2015). Computational methods and resources for the interpretation of genomic variants in cancer. BMC Genomics. 16 (Suppl. 8): S7. PMID: 26111056