Variant Analysis

Bioinformatics for Systems and Synthetic Biology

Emidio Capriotti
http://biofold.org/emidio

Biomolecules
Folding and
Disease

Department of Pharmacy and
Biotechnology (FaBiT)
University of Bologna

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA
A.D. 1088

Single Nucleotide Variants

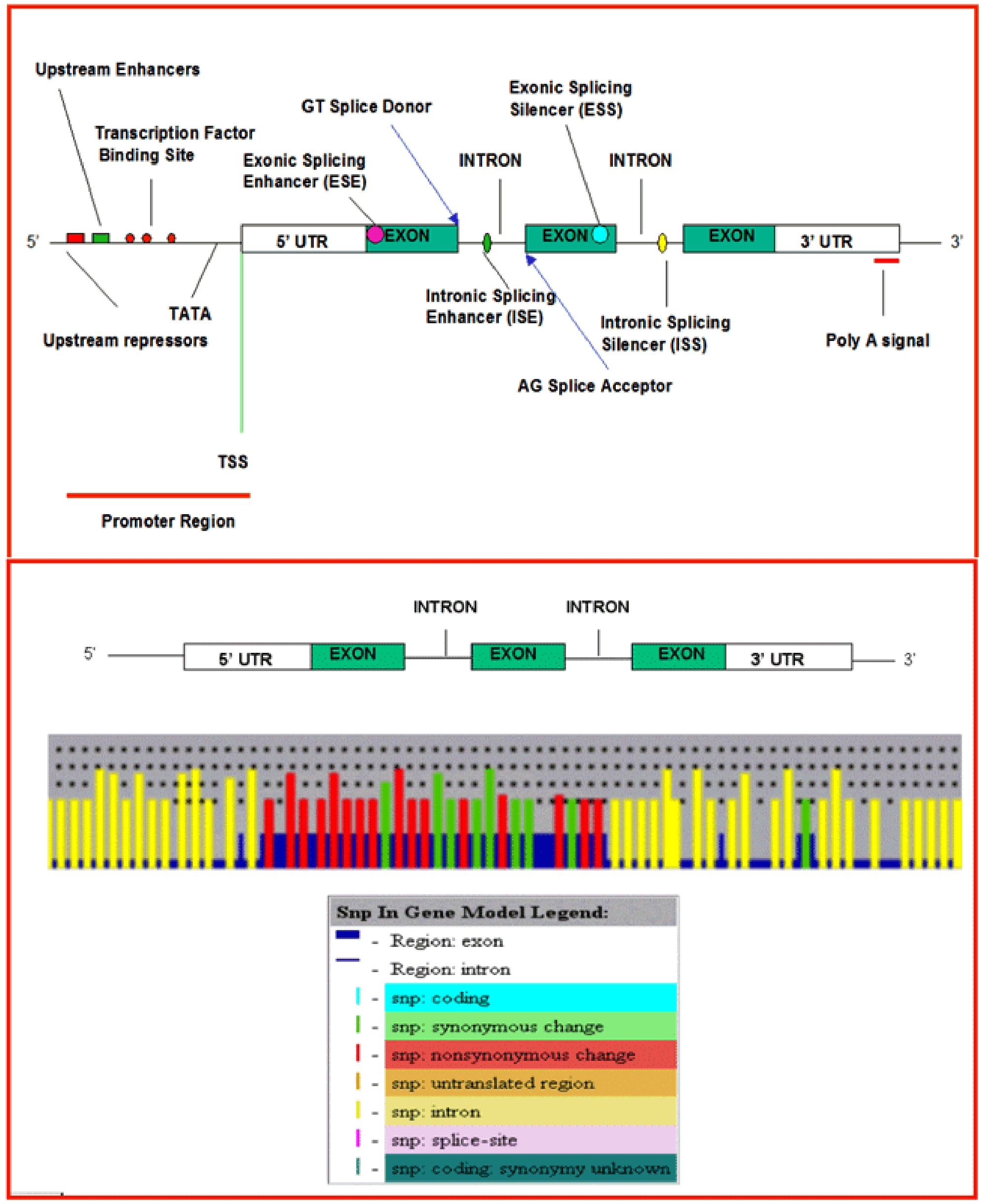
Single Nucleotide Variants (SNVs)
is a DNA sequence variation occurring when a single nucleotide A, T, C, or G in the genome differs between members of the species.
It is used to refer to Polymorphisms when the population frequency is ≥ 1%

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.



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

Effects of variants

It is important to understand the functional effect of Single Nucleotide Polymorphisms (SNPs) that are very common type of variations, but also the impact rare variants which have allele frequencies below than 1%

Impact of coding variants

- Properties of amino acid residue substitution
- The evolutionary history of an amino acid position
- Sequence–function relationships
- Structure–function relationships

Impact of non-coding variants

- Transcription
- Pre-mRNA splicing
- MicroRNA binding
- Altering post-translational modification sites

Cline and Karchin (2011) Bioinformatics, 27; 441-448.

1000 Genomes

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

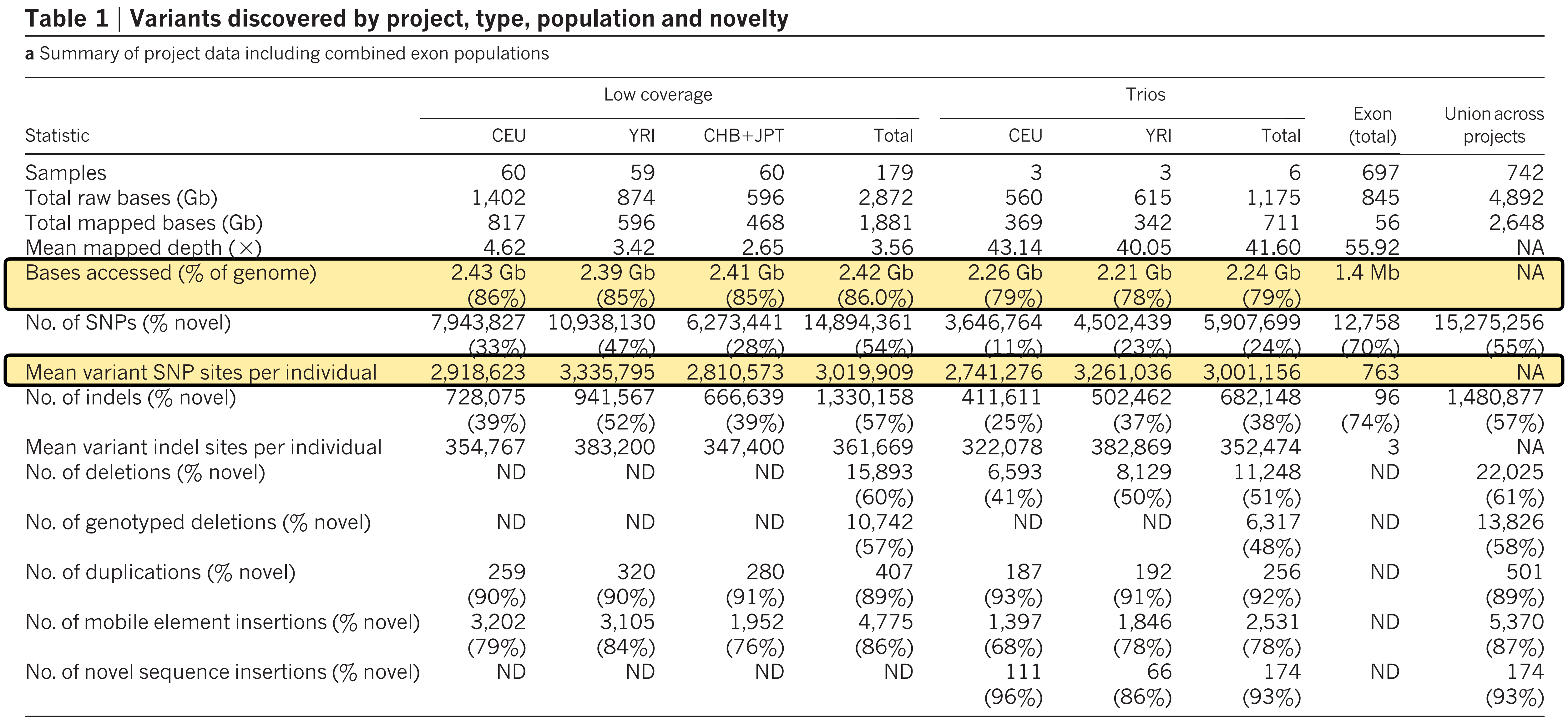


Table 1 | Variants discovered by project, type, population and novelty

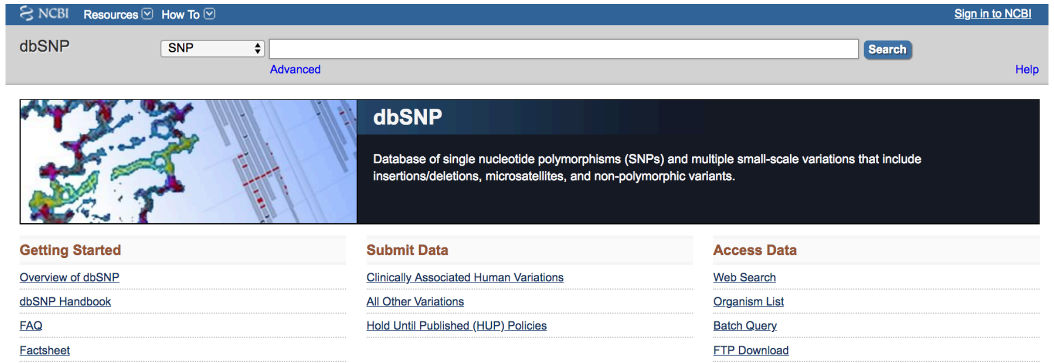
a Summary of project data including combined exon populations

	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.

SNVs and SAVs databases

dbSNP @ NCBI



dbSNP

Database of single nucleotide polymorphisms (SNPs) and multiple small-scale variations that include insertions/deletions, microsatellites, and non-polymorphic variants.

Getting Started
- Overview of dbSNP
- dbSNP Handbook
- FAQ
- Factsheet

Submit Data
- Clinically Associated Human Variations
- All Other Variations
- Hold Until Published (HUP) Policies

Access Data
- Web Search
- Organism List
- Batch Query
- FTP Download

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

Single Nucleotide Variants

Homo sapiens	**904,623,795**

SwissVar @ ExPASy

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

Single Amino acid Variants

Homo sapiens	83,996
Disease	*32,930*
Polymorphisms	*39,938*

Jan 2026

Variant Call Format

The final result of the variant calling procedure is a VCF file.

```
##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
. . . . . .
```

File Content

The file contains information about single nucleotide variants and indels of single or multiple samples.

For each variant the number of supporting reads for reference and alternative alleale

The original VCF does not contain any information functional effect of the variants.

Main data sources

Single genetic variants are collected in different databases:

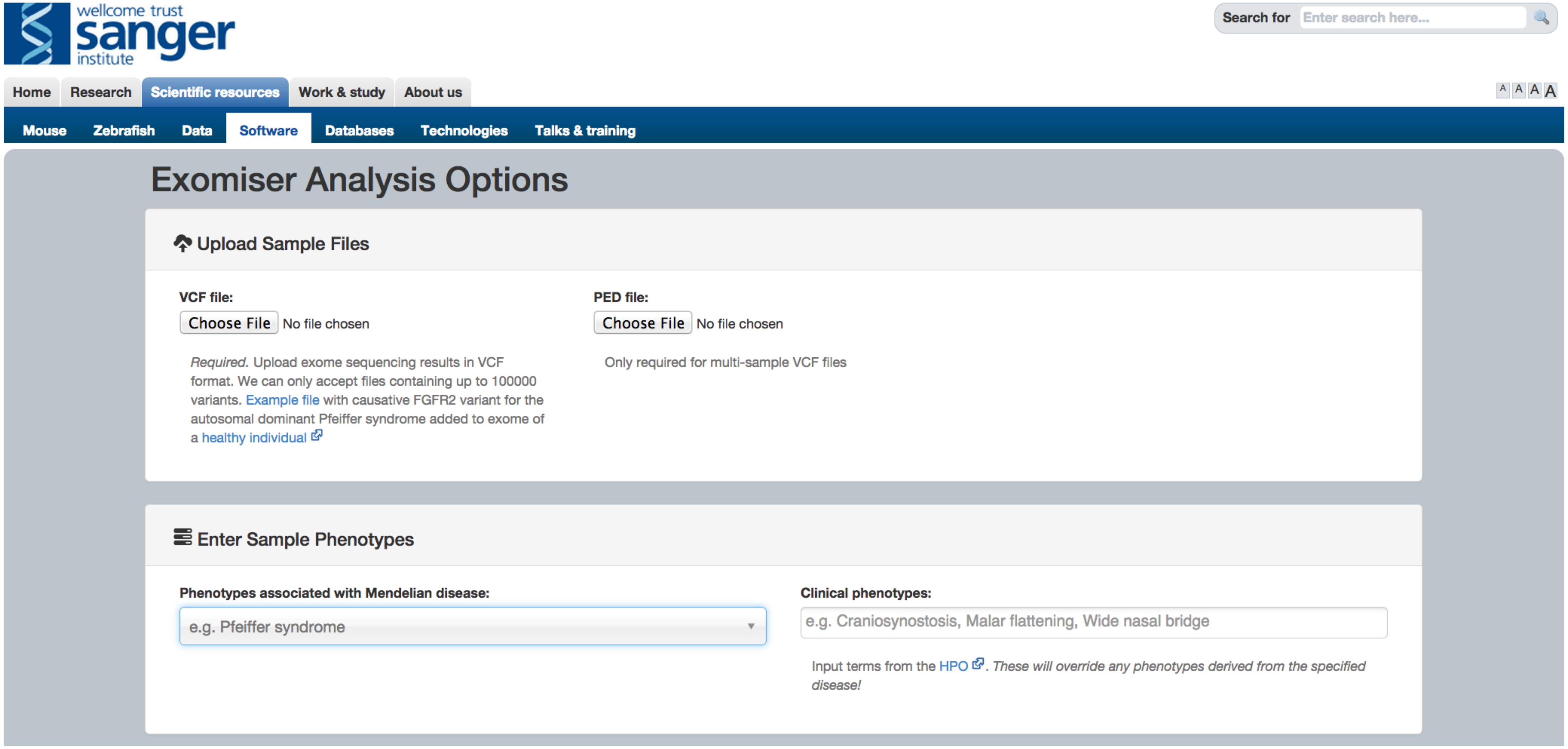
- dbSNP - variation from all species. http://www.ncbi.nlm.nih.gov/SNP/
- EVS - specific for human. http://evs.gs.washington.edu/EVS/
- ClinVar - Variants and human health. http://www.ncbi.nlm.nih.gov/clinvar/
- Cosmic - Somatic mutation in cancer. http://cancer.sanger.ac.uk/

This information is important for variant calling but useless for capturing the complexity of genotype/phenotype relationship. The VCF more informative because we can analyze co-occurring events. The major sources are:

- 1000 Genomes: WGS data of individuals http://www.1000genomes.org/
- TCGA: Cancer Genomes https://tcga-data.nci.nih.gov/

All in One

Exomizer is a variant analysis tools that tests presence of variants associated to specific phenotypes



Exomiser Analysis Options

Upload Sample Files

VCF file: Choose File No file chosen

Required. Upload exome sequencing results in VCF format. We can only accept files containing up to 100000 variants. Example file with causative FGFR2 variant for the autosomal dominant Pfeiffer syndrome added to exome of a healthy individual

PED file: Choose File No file chosen

Only required for multi-sample VCF files

Enter Sample Phenotypes

Phenotypes associated with Mendelian disease: e.g. Pfeiffer syndrome

Clinical phenotypes: e.g. Craniosynostosis, Malar flattening, Wide nasal bridge

Input terms from the HPO. *These will override any phenotypes derived from the specified disease!*

http://www.sanger.ac.uk/resources/software/exomiser/submit/

The complexity of cancer

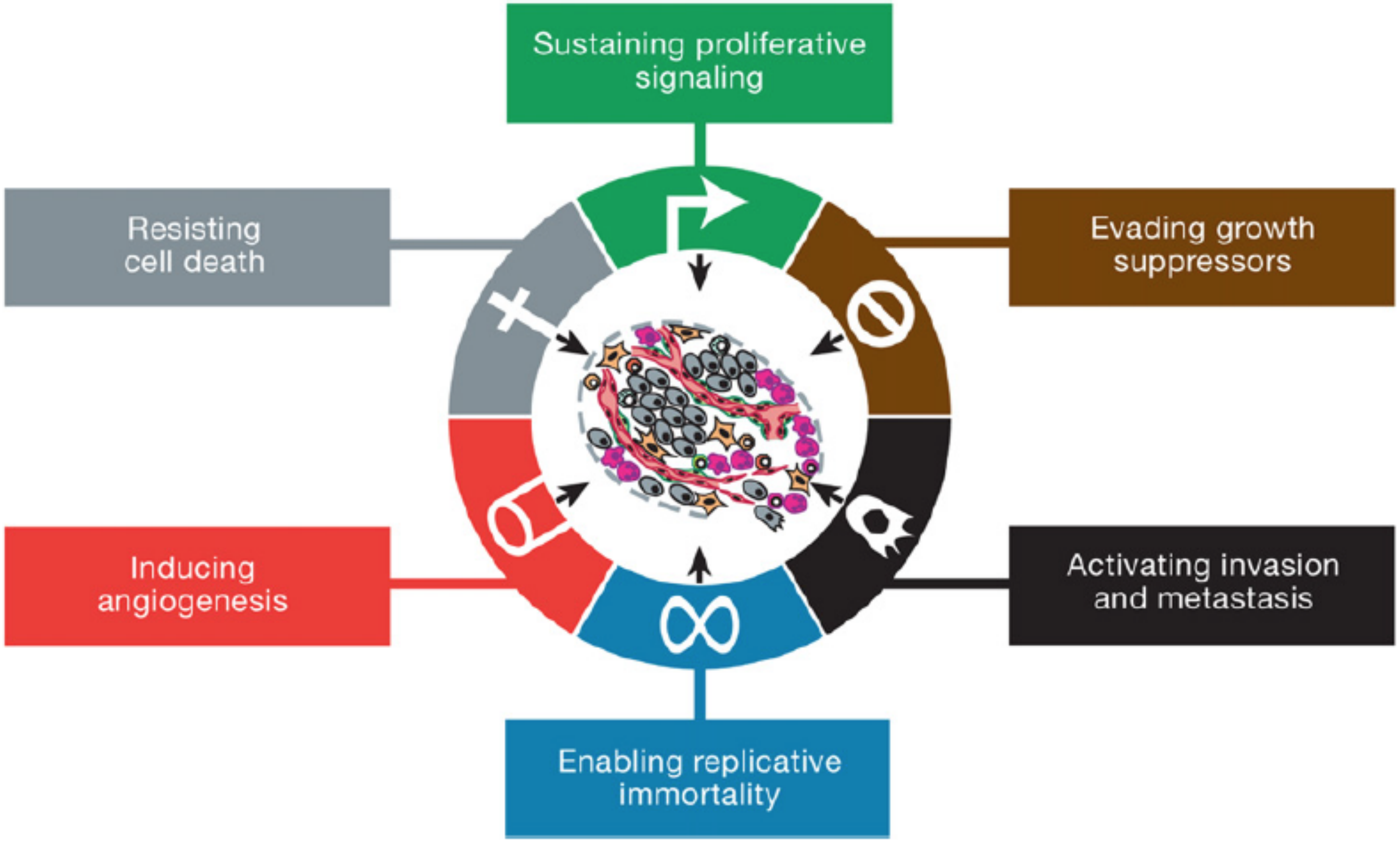
Cancer is **complex disorder** characterized by high level of mutation rate.

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.

Hallmarks of cancer

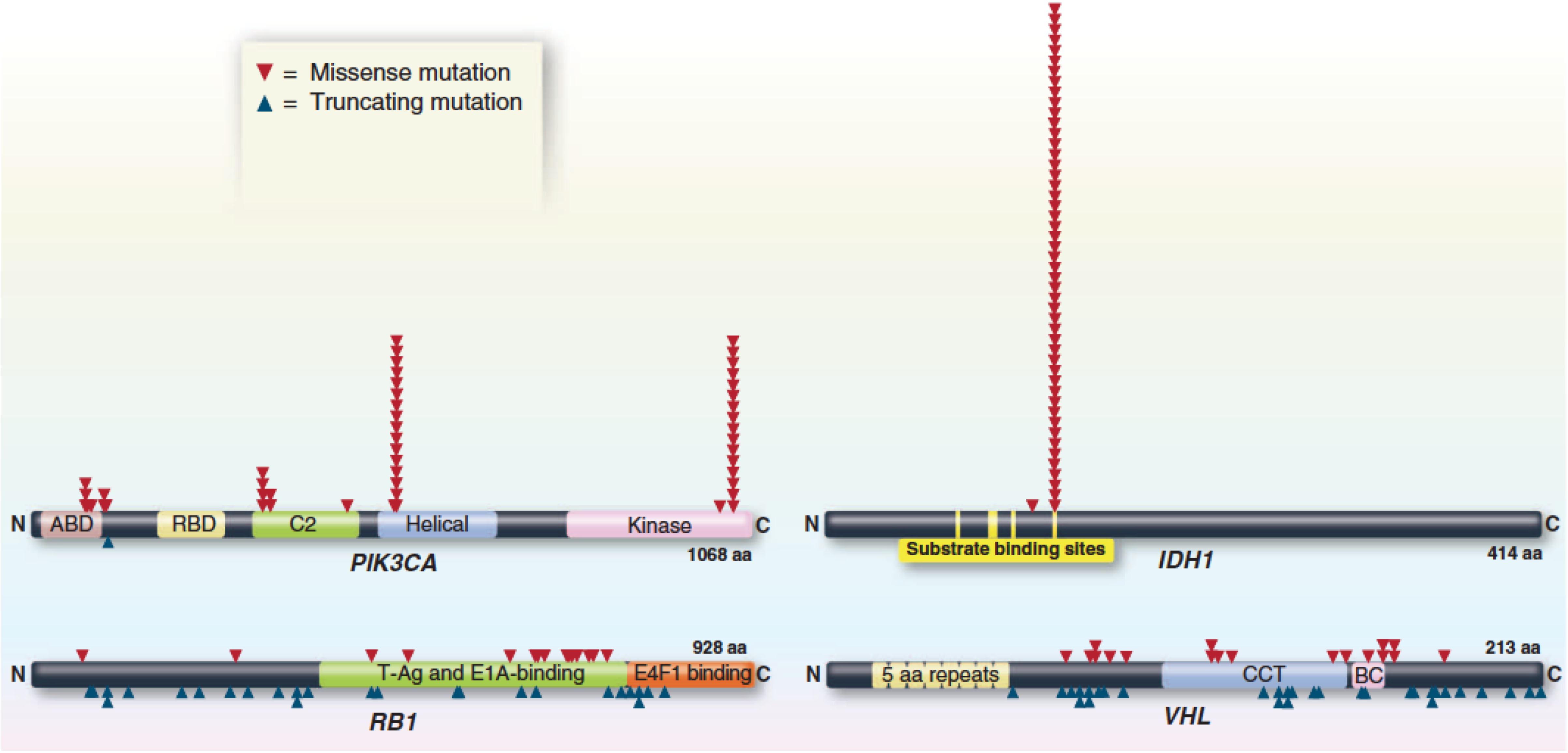
The six hallmarks of cancer - distinctive and complementary capabilities that enable tumor growth and metastatic dissemination.


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

Hanahan and Weinberg. Cell 2011, **144**:646

Oncogene vs Suppressor

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

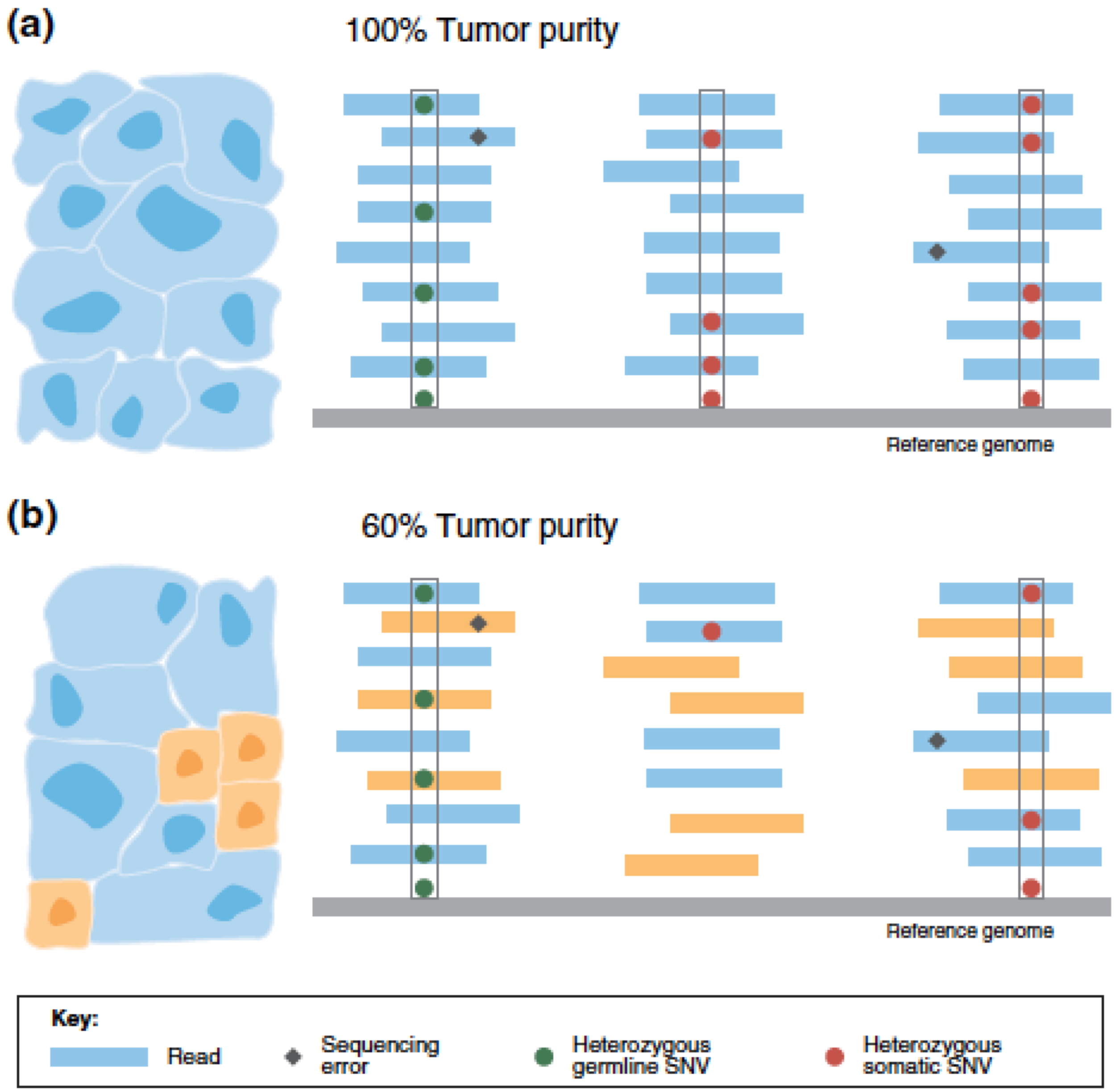


Vogelstein et al. Science 2013, **339**:1546

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.



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

How data looks like?

Variant Calling File (VCF) with germline and somatic variants

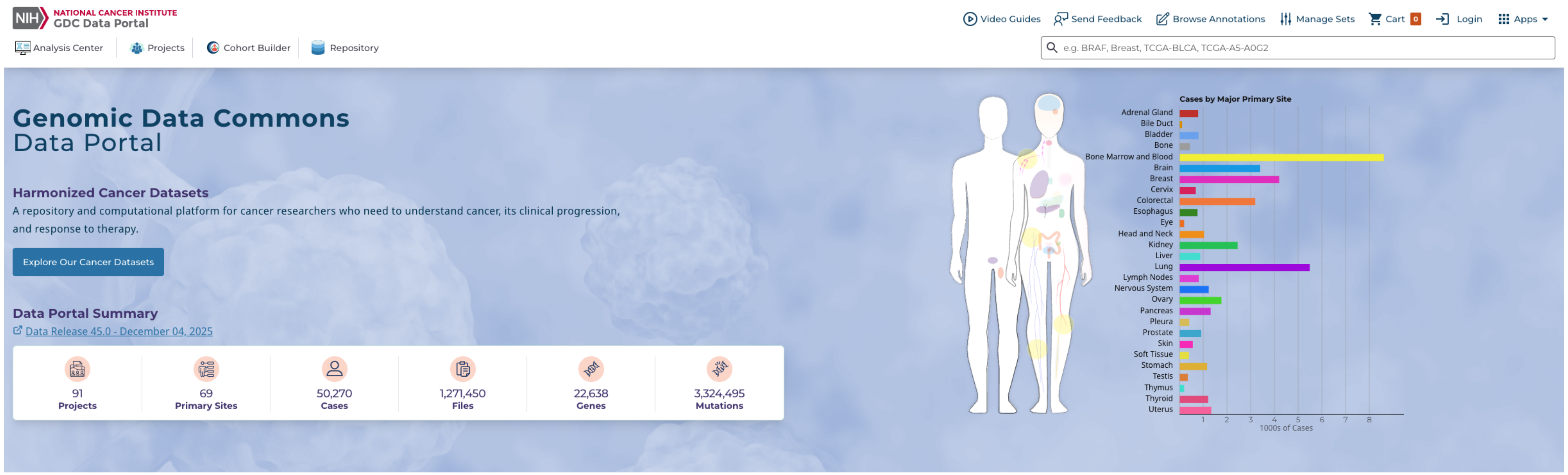
```
##fileformat=VCFv4.1
##tcgaversion=1.1
##reference=<ID=hg19,source=.>
##phasing=none
##geneAnno=none
##INFO=<ID=VT,Number=1,Type=String,Description="Variant type, can be SNP, INS or DEL">
##INFO=<ID=VLS,Number=1,Type=Integer,Description="Final validation status relative to non-adjacent Normal, ......”>
##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
. . . . . .
```

The TCGA data

The Cancer Genome Atlas Consortium

TCGA data (https://portal.gdc.cancer.gov/)

- 91 cancer projects (~50,270 cases)
- BAM files available



Genomic Data Commons
Data Portal

Harmonized Cancer Datasets

A repository and computational platform for cancer researchers who need to understand cancer, its clinical progression, and response to therapy.

Data Portal Summary

Data Release 45.0 - December 04, 2025

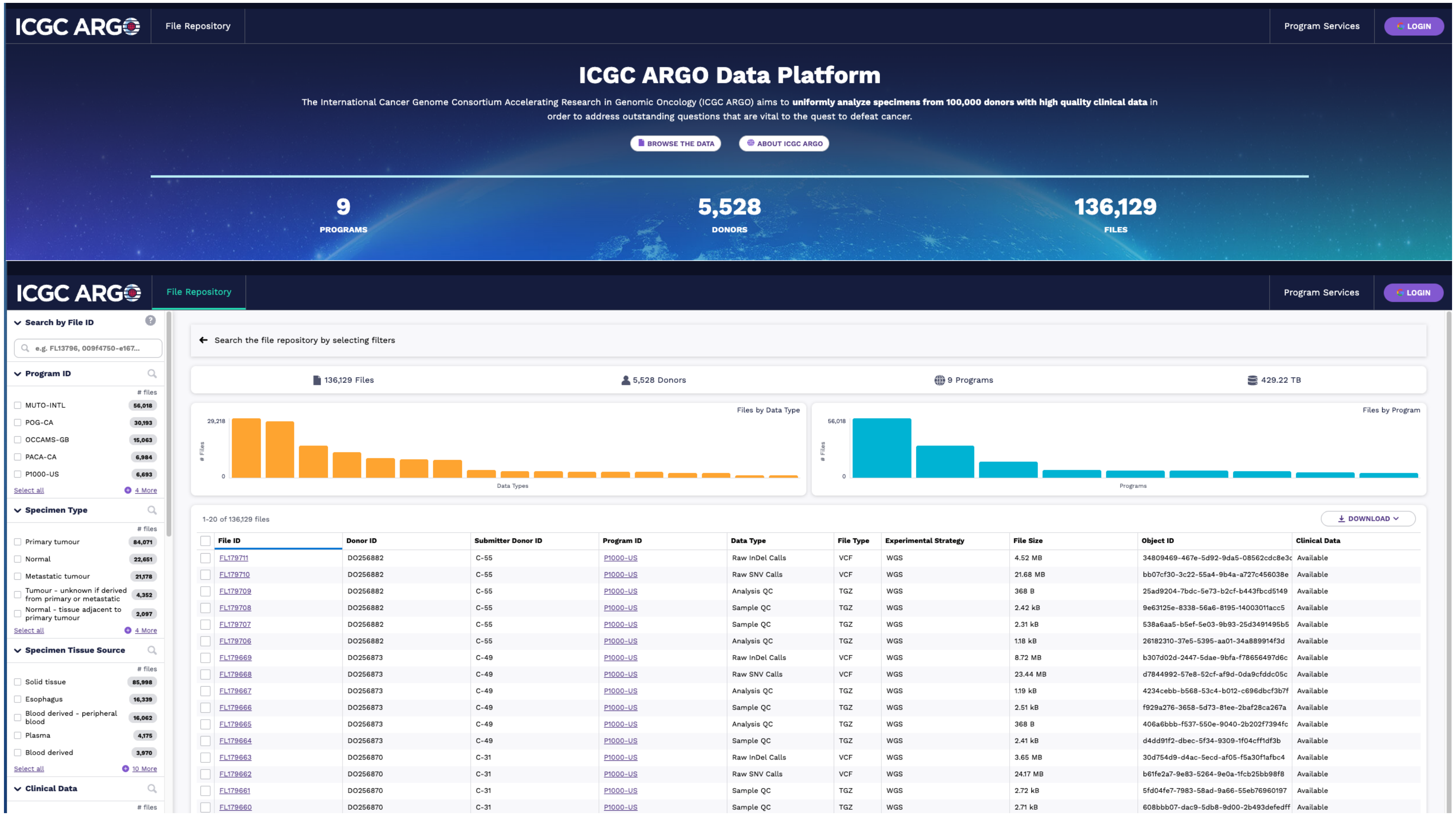
91	69	50,270	1,271,450	22,638	3,324,495
Projects	**Primary Sites**	**Cases**	**Files**	**Genes**	**Mutations**

The ICGC AR%GO

The International Cancer Genome Consortium

ICGC (https://platform.icgc-argo.org/)

- 5,528 Donors
- 429.22 TB data
- 77.4 million simple somatic mutations.



ICGC ARGO Data Platform

The International Cancer Genome Consortium Accelerating Research in Genomic Oncology (ICGC ARGO) aims to **uniformly analyze specimens from 100,000 donors with high quality clinical data** in order to address outstanding questions that are vital to the quest to defeat cancer.

9	5,528	136,129
PROGRAMS	DONORS	FILES

Search by File ID

e.g. FL13796, 009f4750-e167...

Program ID

	# files
□ MUTO-INTL	56,018
□ POG-CA	30,193
□ OCCAMS-GB	15,063
□ PACA-CA	6,984
□ P1000-US	6,693

Select all · 4 More

Specimen Type

	# files
□ Primary tumour	84,071
□ Normal	22,651
□ Metastatic tumour	21,178
□ Tumour - unknown if derived from primary or metastatic	4,352
□ Normal - tissue adjacent to primary tumour	2,097

Select all · 4 More

Specimen Tissue Source

	# files
□ Solid tissue	85,998
□ Esophagus	16,339
□ Blood derived - peripheral blood	16,062
□ Plasma	4,175
□ Blood derived	3,970

Select all · 10 More

Clinical Data

# files

Search the file repository by selecting filters

136,129 Files · 5,528 Donors · 9 Programs · 429.22 TB

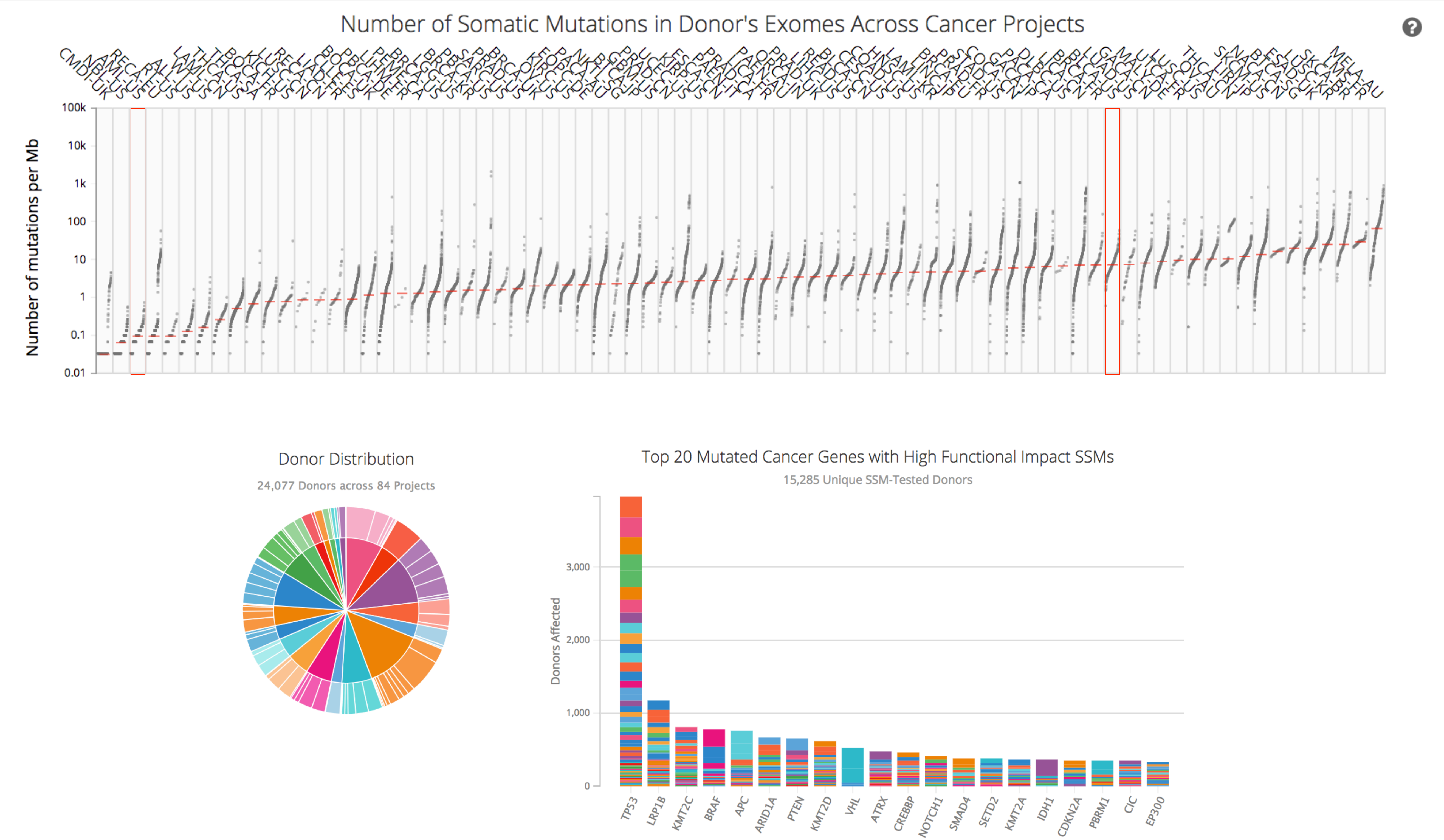
1-20 of 136,129 files

DOWNLOAD

□	File ID	Donor ID	Submitter Donor ID	Program ID	Data Type	File Type	Experimental Strategy	File Size	Object ID	Clinical Data
□	FL179711	DO256882	C-55	P1000-US	Raw InDel Calls	VCF	WGS	4.52 MB	34809469-467e-5d92-9da5-08562cdc8e3c	Available
□	FL179710	DO256882	C-55	P1000-US	Raw SNV Calls	VCF	WGS	21.68 MB	bb07cf30-3c22-55a4-9b4a-a727c456038e	Available
□	FL179709	DO256882	C-55	P1000-US	Analysis QC	TGZ	WGS	368 B	25ad9204-7bdc-5e73-b2cf-b443fbcd5149	Available
□	FL179708	DO256882	C-55	P1000-US	Sample QC	TGZ	WGS	2.42 kB	9e63125e-8338-56a6-8195-14003011acc5	Available
□	FL179707	DO256882	C-55	P1000-US	Sample QC	TGZ	WGS	2.31 kB	538a6aa5-b5ef-5e03-9b93-25d3491495b5	Available
□	FL179706	DO256882	C-55	P1000-US	Analysis QC	TGZ	WGS	1.18 kB	26182310-37e5-5395-aa01-34a889914f3d	Available
□	FL179669	DO256873	C-49	P1000-US	Raw InDel Calls	VCF	WGS	8.72 MB	b307d02d-2447-5dae-9bfa-f78656497d6c	Available
□	FL179668	DO256873	C-49	P1000-US	Raw SNV Calls	VCF	WGS	23.44 MB	d7844992-57e8-52cf-af9d-0da9cfddc05c	Available
□	FL179667	DO256873	C-49	P1000-US	Analysis QC	TGZ	WGS	1.19 kB	4234cebb-b568-53c4-b012-c696dbcf3b7f	Available
□	FL179666	DO256873	C-49	P1000-US	Sample QC	TGZ	WGS	2.51 kB	f929a276-3658-5d73-81ee-2baf28ca267a	Available
□	FL179665	DO256873	C-49	P1000-US	Analysis QC	TGZ	WGS	368 B	406a6bbb-f537-550e-9040-2b202f7394fc	Available
□	FL179664	DO256873	C-49	P1000-US	Sample QC	TGZ	WGS	2.41 kB	d4dd91f2-dbec-5f34-9309-1f04cff1df3b	Available
□	FL179663	DO256870	C-31	P1000-US	Raw InDel Calls	VCF	WGS	3.65 MB	30d754d9-d4ac-5ecd-af05-f5a30f1afbc4	Available
□	FL179662	DO256870	C-31	P1000-US	Raw SNV Calls	VCF	WGS	24.17 MB	b61fe2a7-9e83-5264-9e0a-1fcb25bb98f8	Available
□	FL179661	DO256870	C-31	P1000-US	Sample QC	TGZ	WGS	2.72 kB	5fd04fe7-7983-58ad-9a66-55eb76960197	Available
□	FL179660	DO256870	C-31	P1000-US	Sample QC	TGZ	WGS	2.71 kB	608bbb07-dac9-5db8-9d00-2b493defedff	Available

Somatic Mutations

Number of somatic mutations per sample vary significantly across cancer types



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

Donor Distribution

24,077 Donors across 84 Projects

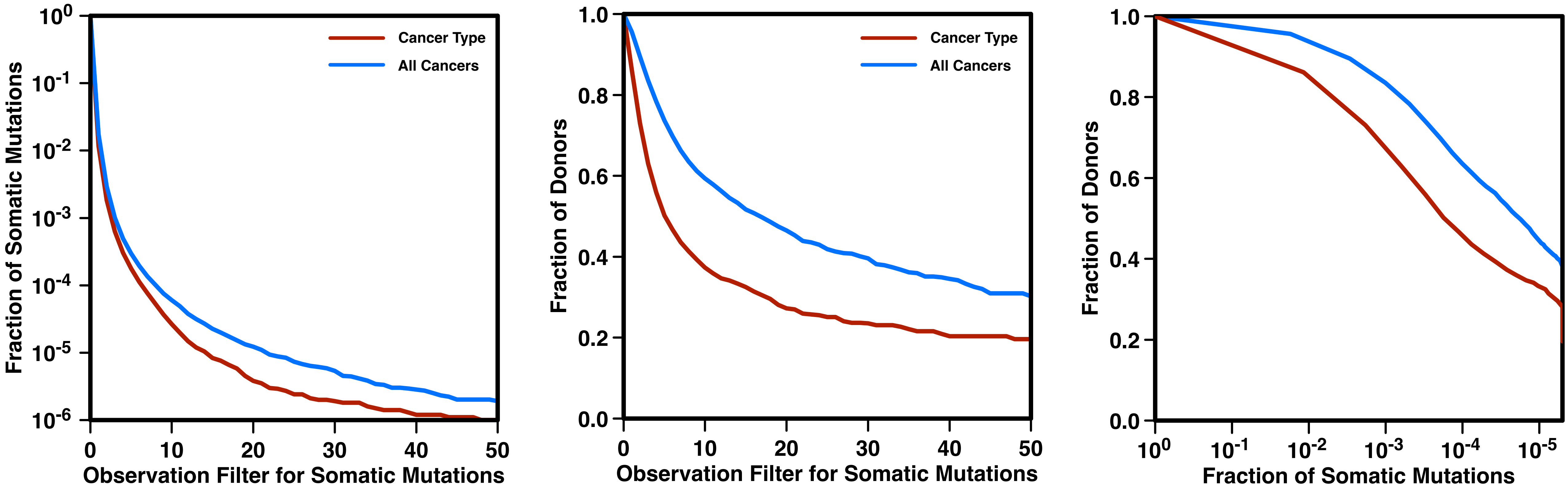
Top 20 Mutated Cancer Genes with High Functional Impact SSMs

15,285 Unique SSM-Tested Donors

Driver vs Passenger

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

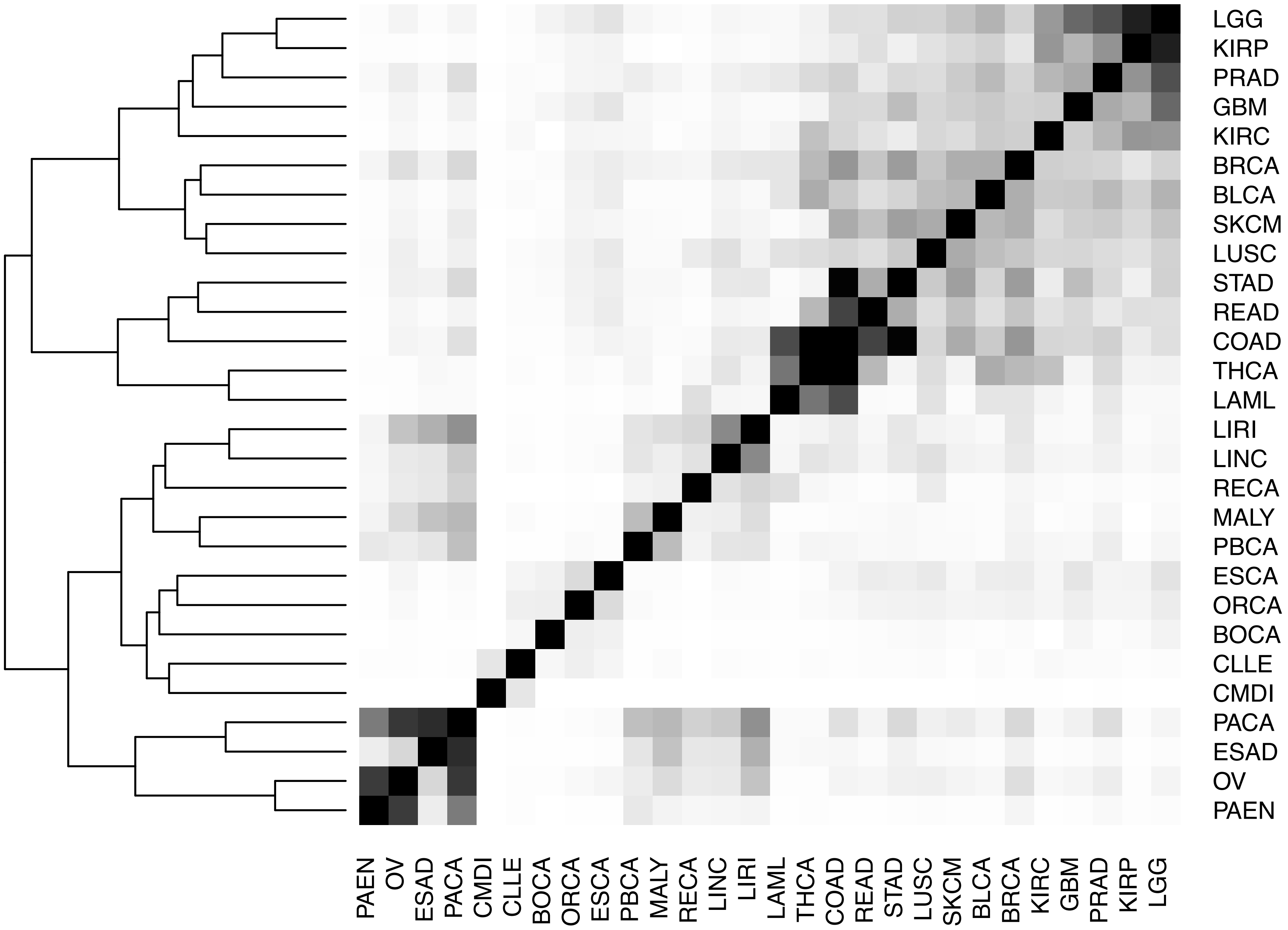
Selecting mutations that are repeated at least twice we filter out ~98% mutations
and are still able to recover ~96% of the patients



Tian R, Basu M, Capriotti E. BMC Genomics 2015

The Cancer Tree

The analysis of recurrent somatic mutations can be used to define similarities across cancer types.

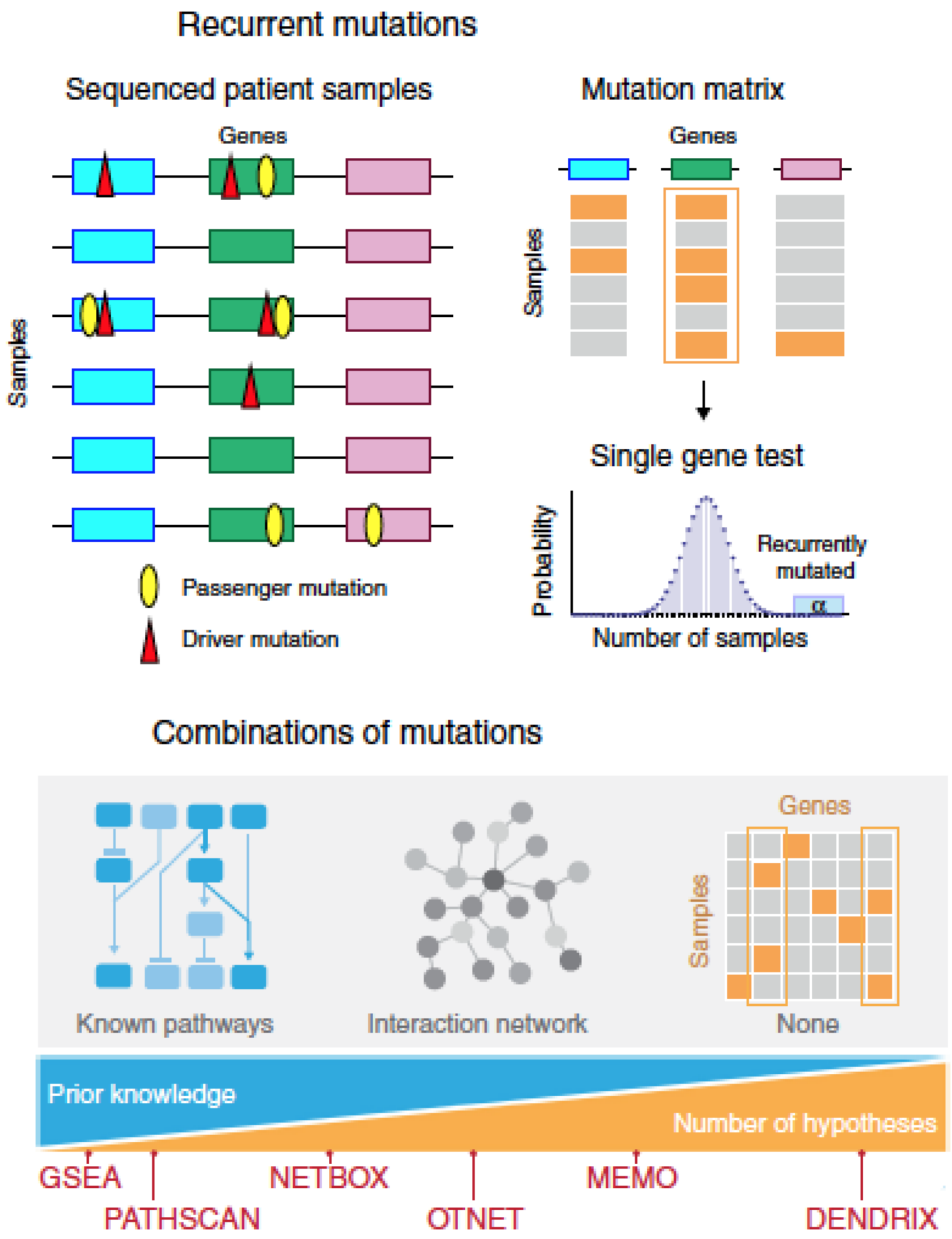


Recurrent variations

Recurrent mutations that are found in more samples than would be expected by chance 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 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

The main idea

Genes implicated in cancer should have **high mutation rate**

In comparison to normal, tumor cells should have higher occurrence of functional mutations in genes involved in the insurgence and progression of the disease.

Problem:

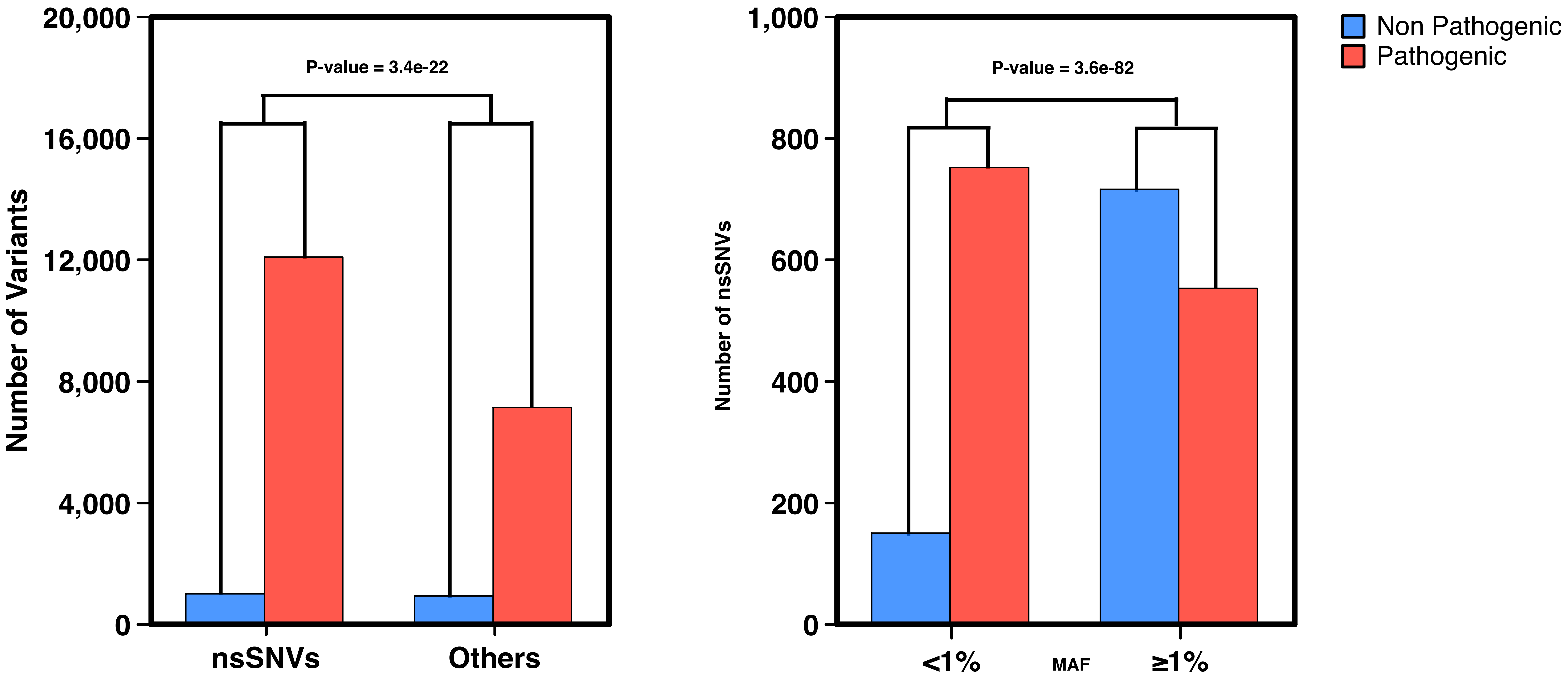
How can we select mutations with functional impact?

Average number of variants	~3,000,000
Average exome variants	~23,000
Average nonsynonymous single nucleotide variants	~10,000
Average rare (MAF≤0.5%) nonsynonymous single nucleotide variants	~300

The 1000 Genomes Project (2010). Nature. 467; 1062-1073.

Variants and MAF

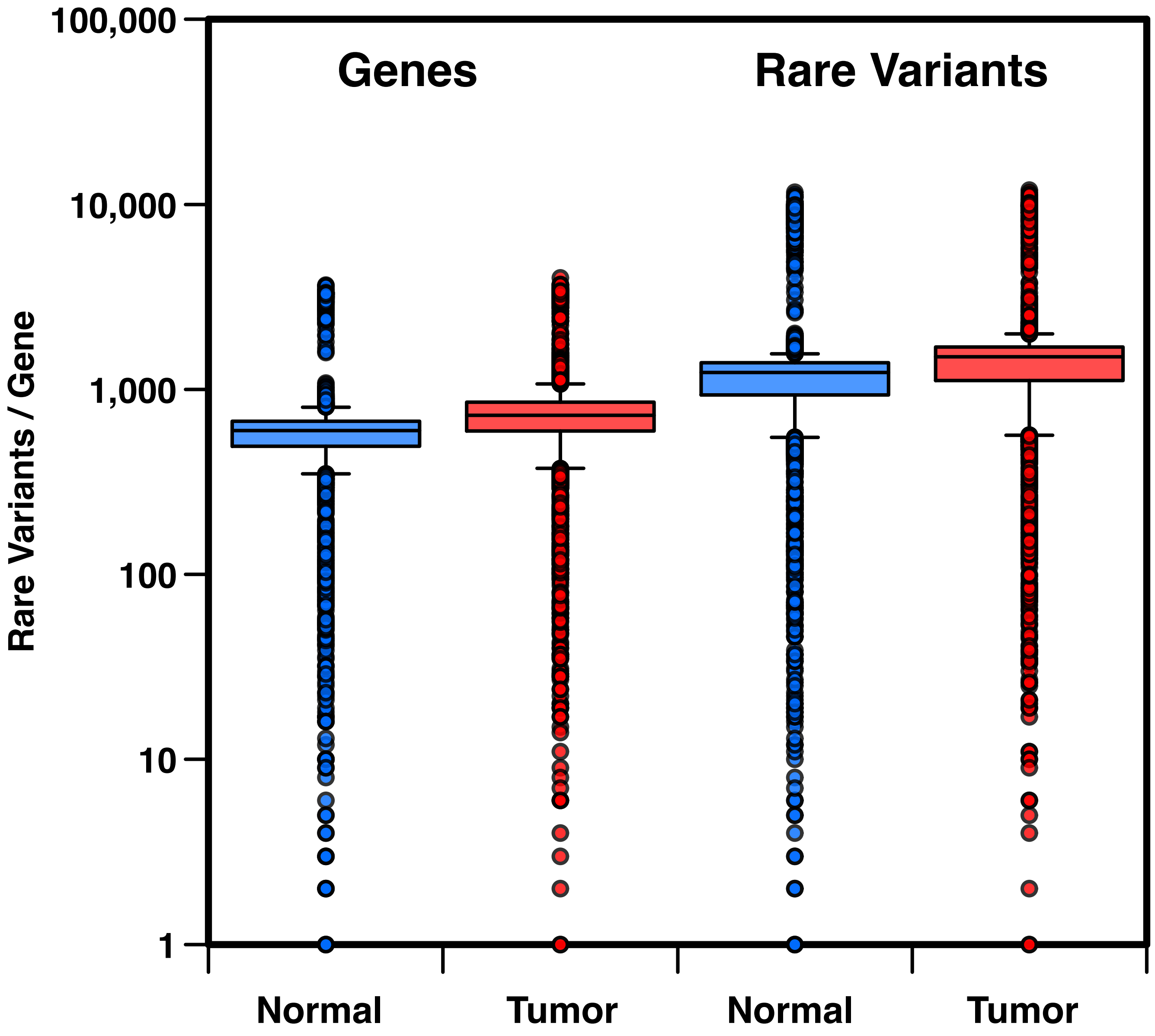
Rare variants are more likely to be associated to disease than high frequency variants



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

Rate Variants and Genes

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



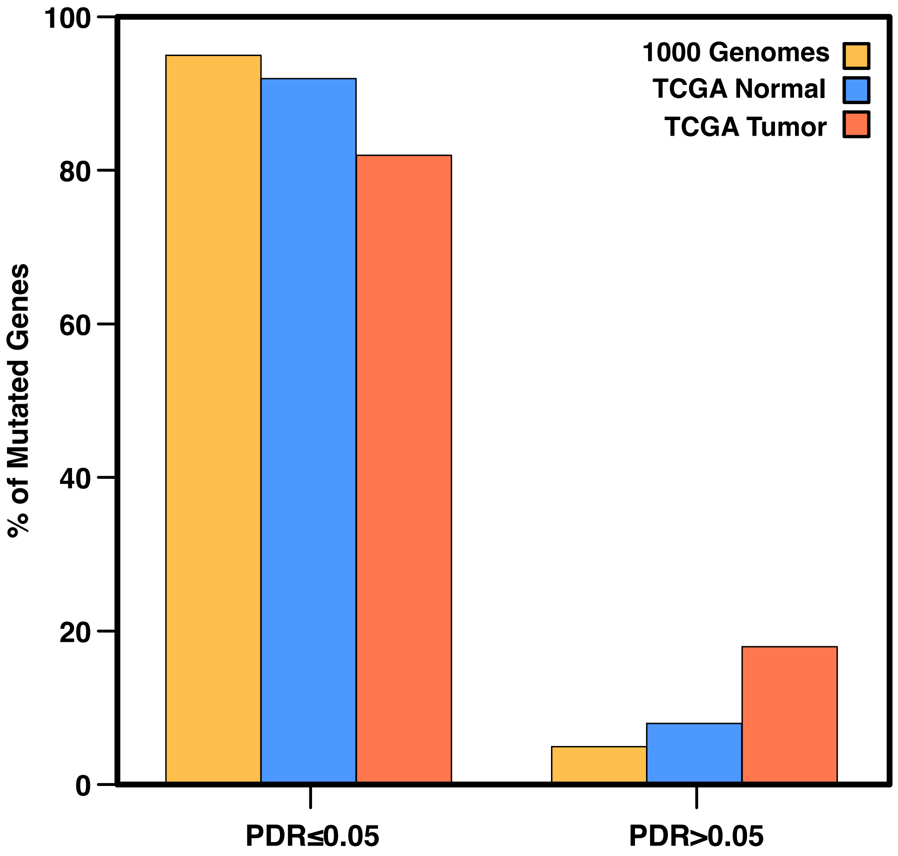
Mutation rates

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%


% of Mutated Genes
PDR≤0.05
PDR>0.05
1000 Genomes
TCGA Normal
TCGA Tumor