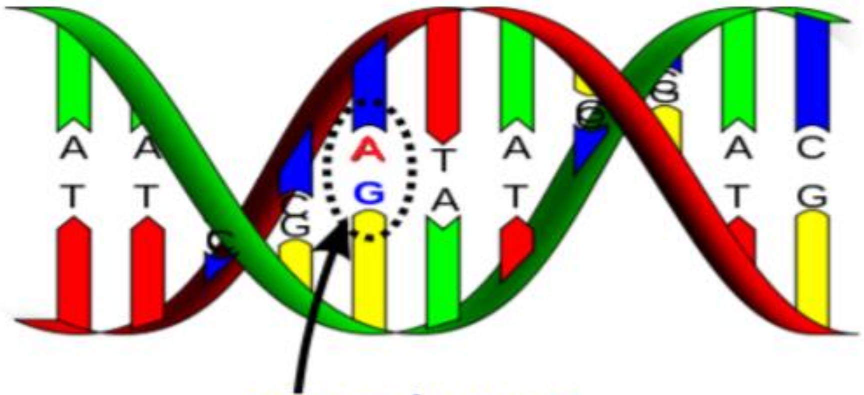
Archana Bhardwaj

GBIO0002

Important genetic terms

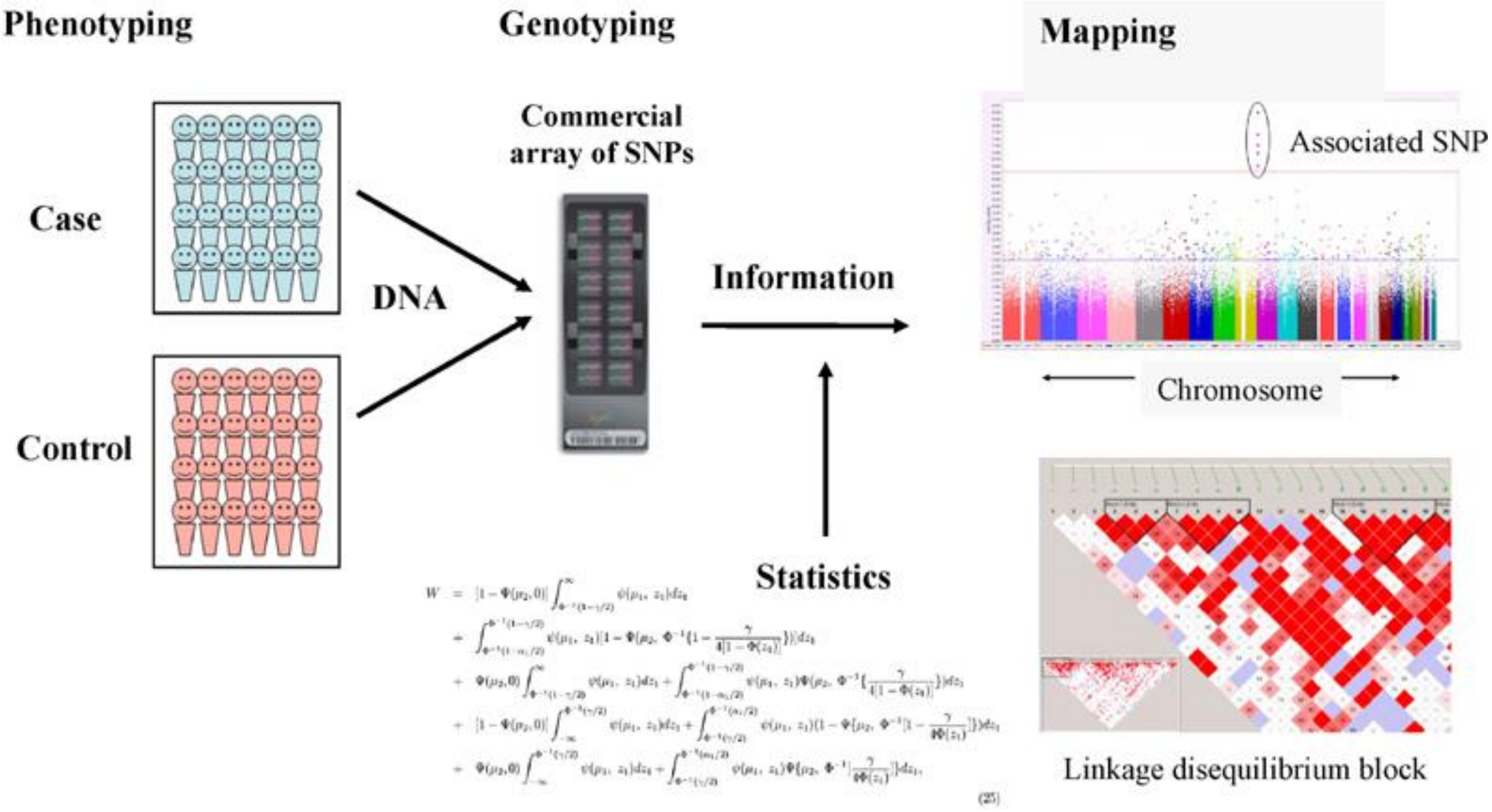
- Given position in the genome (i.e. locus) has several associated alleles (A and G) which produce genotypes r_A/r_G



SNP at locus X

- Haplotypes
  - Combination of alleles at different loci

GWAS



Phenotyping

Genotyping

Mapping

Case

Control

DNA

Commercial array of SNPs

Information

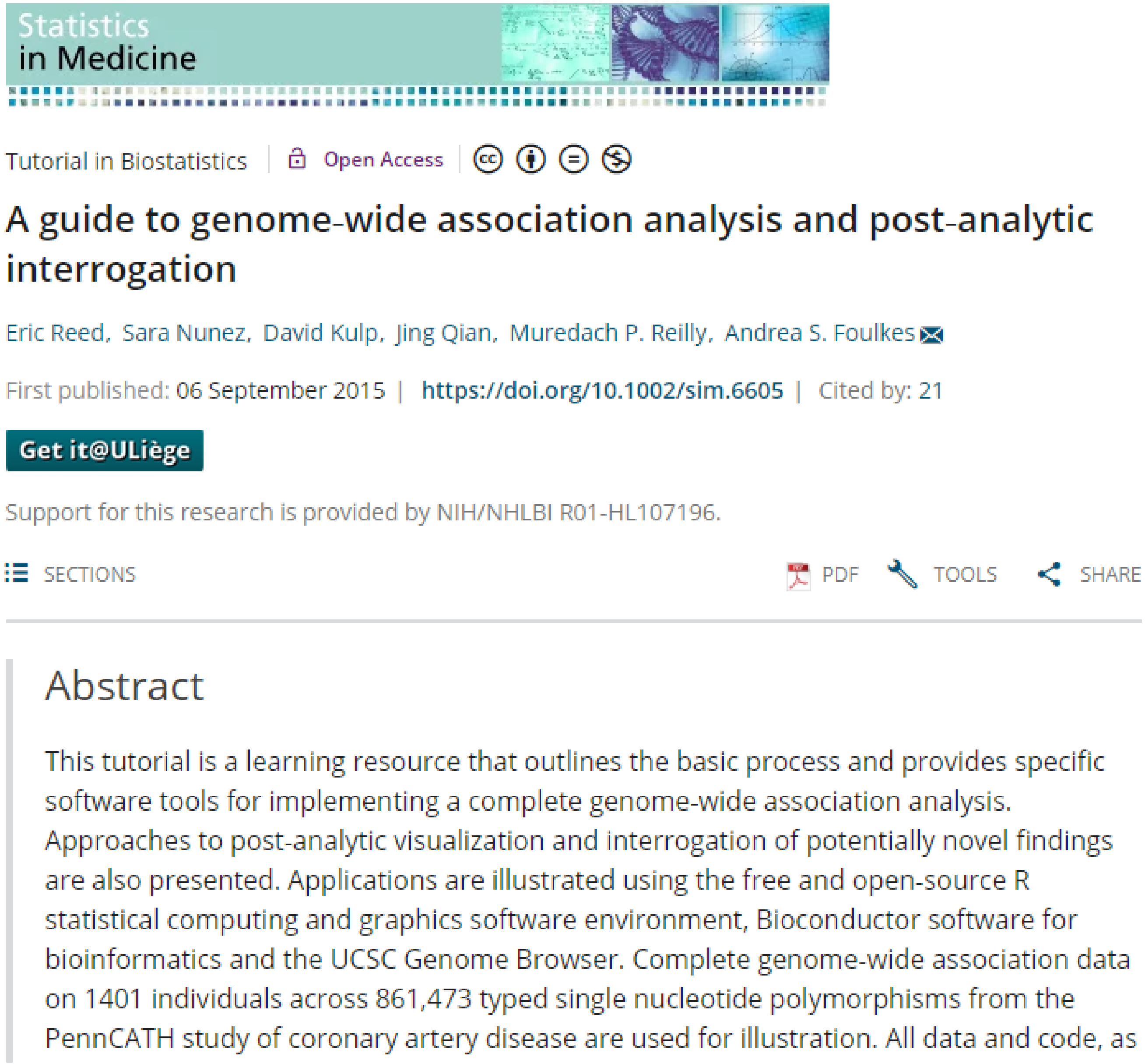
Statistics

Associated SNP

Chromosome

Linkage disequilibrium block

$$\begin{aligned} W &= [1-\Phi(\mu_2,0)]\int_{\Phi^{-1}(1-\gamma/2)}^{\infty}\phi(\mu_1, z_1)dz_1 \\ &+ \int_{\Phi^{-1}(1-\alpha_1/2)}^{\Phi^{-1}(1-\gamma/2)}\phi(\mu_1, z_1)[1-\Phi(\mu_2, \Phi^{-1}\{1-\frac{\gamma}{4[1-\Phi(z_1)]}\})]dz_1 \\ &+ \Phi(\mu_2,0)\int_{\Phi^{-1}(1-\gamma/2)}^{\infty}\phi(\mu_1, z_1)dz_1 + \int_{\Phi^{-1}(1-\alpha_1/2)}^{\Phi^{-1}(1-\gamma/2)}\phi(\mu_1, z_1)\Phi(\mu_2, \Phi^{-1}\{\frac{\gamma}{4[1-\Phi(z_1)]}\})dz_1 \\ &+ [1-\Phi(\mu_2,0)]\int_{-\infty}^{\Phi^{-1}(\gamma/2)}\phi(\mu_1, z_1)dz_1 + \int_{\Phi^{-1}(\gamma/2)}^{\Phi^{-1}(\alpha_1/2)}\phi(\mu_1, z_1)(1-\Phi\{\mu_2, \Phi^{-1}[1-\frac{\gamma}{4\Phi(z_1)}]\})dz_1 \\ &+ \Phi(\mu_2,0)\int_{-\infty}^{\Phi^{-1}(\gamma/2)}\phi(\mu_1, z_1)dz_1 + \int_{\Phi^{-1}(\gamma/2)}^{\Phi^{-1}(\alpha_1/2)}\phi(\mu_1, z_1)\Phi\{\mu_2, \Phi^{-1}[\frac{\gamma}{4\Phi(z_1)}]\}dz_1, \end{aligned} \tag{25}$$



Statistics in Medicine

Tutorial in Biostatistics | Open Access

A guide to genome-wide association analysis and post-analytic interrogation

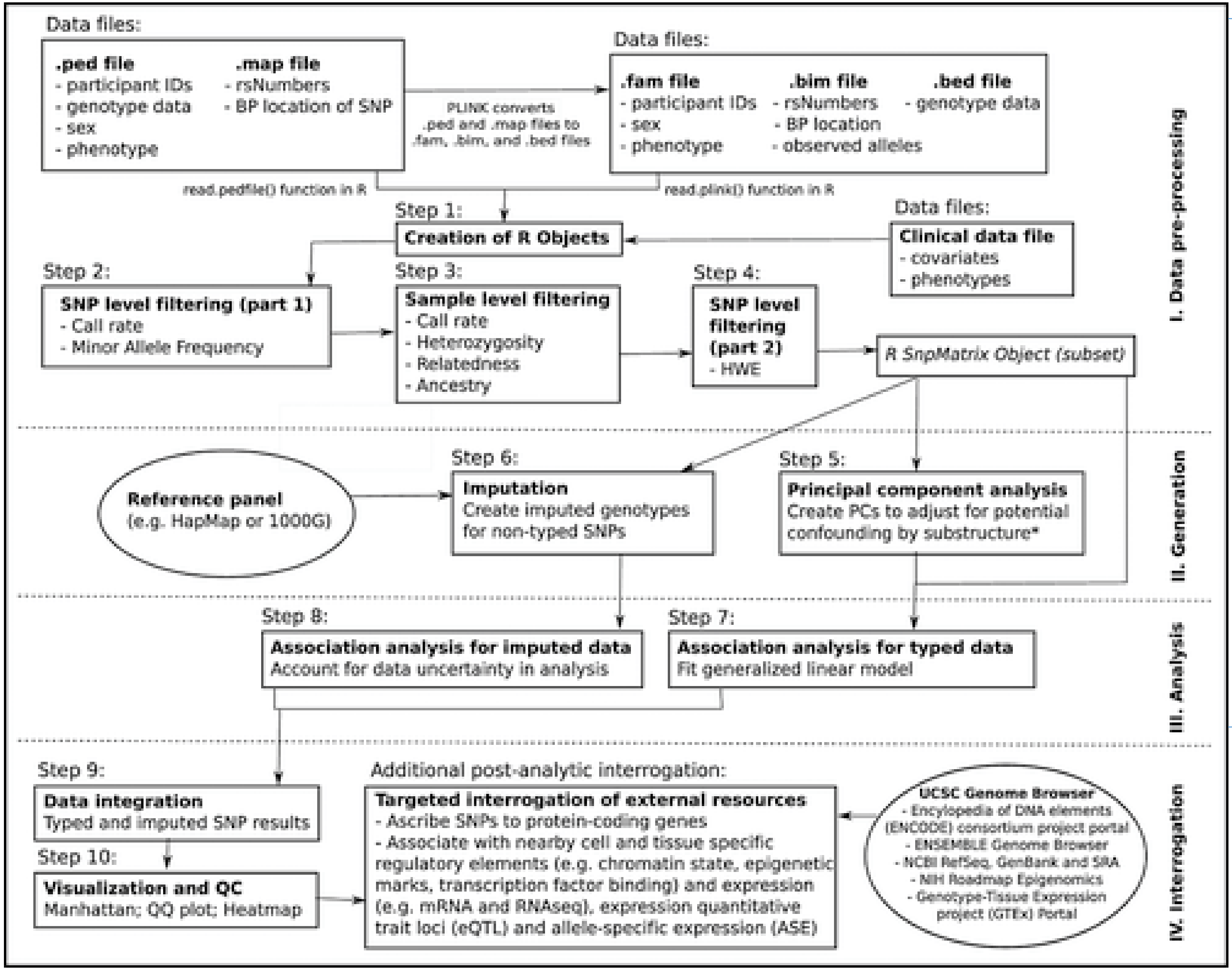
Eric Reed, Sara Nunez, David Kulp, Jing Qian, Muredach P. Reilly, Andrea S. Foulkes

First published: 06 September 2015 | https://doi.org/10.1002/sim.6605 | Cited by: 21

Support for this research is provided by NIH/NHLBI R01-HL107196.

Abstract

This tutorial is a learning resource that outlines the basic process and provides specific software tools for implementing a complete genome-wide association analysis. Approaches to post-analytic visualization and interrogation of potentially novel findings are also presented. Applications are illustrated using the free and open-source R statistical computing and graphics software environment, Bioconductor software for bioinformatics and the UCSC Genome Browser. Complete genome-wide association data on 1401 individuals across 861,473 typed single nucleotide polymorphisms from the PennCATH study of coronary artery disease are used for illustration. All data and code, as


Data files:
.ped file
- participant IDs
- genotype data
- sex
- phenotype
.map file
- rsNumbers
- BP location of SNP
PLINK converts .ped and .map files to .fam, .bim, and .bed files
Data files:
.fam file
- participant IDs
- sex
- phenotype
.bim file
- rsNumbers
- BP location
- observed alleles
.bed file
- genotype data
read.pedfile() function in R
read.plink() function in R
Step 1:
Creation of R Objects
Data files:
Clinical data file
- covariates
- phenotypes
Step 2:
SNP level filtering (part 1)
- Call rate
- Minor Allele Frequency
Step 3:
Sample level filtering
- Call rate
- Heterozygosity
- Relatedness
- Ancestry
Step 4:
SNP level filtering (part 2)
- HWE
R SnpMatrix Object (subset)
I. Data pre-processing
Reference panel
(e.g. HapMap or 1000G)
Step 6:
Imputation
Create imputed genotypes for non-typed SNPs
Step 5:
Principal component analysis
Create PCs to adjust for potential confounding by substructure*
II. Generation
Step 8:
Association analysis for imputed data
Account for data uncertainty in analysis
Step 7:
Association analysis for typed data
Fit generalized linear model
III. Analysis
Step 9:
Data integration
Typed and imputed SNP results
Step 10:
Visualization and QC
Manhattan; QQ plot; Heatmap
Additional post-analytic interrogation:
Targeted interrogation of external resources
- Ascribe SNPs to protein-coding genes
- Associate with nearby cell and tissue specific regulatory elements (e.g. chromatin state, epigenetic marks, transcription factor binding) and expression (e.g. mRNA and RNAseq), expression quantitative trait loci (eQTL) and allele-specific expression (ASE)
UCSC Genome Browser
- Encyclopedia of DNA elements (ENCODE) consortium project portal
- ENSEMBLE Genome Browser
- NCBI RefSeq, GenBank and SRA
- NIH Roadmap Epigenomics
- Genotype-Tissue Expression project (GTEx) Portal
IV. Interrogation

GWAS

Post GWAS

GWAS main philosophy

- GWAS = Genome Wide Association Studies
- IDEA = GWAS involve scan for large number of genetic markers across the whole genome of many individuals to find specific genetic variations associated with the disease and/or other phenotype
- Find the genetic variation(s) that contribute(s) and explain(s) complex diseases

GWAS visually

- GWAS tries to uncover links between genetic basis of the disease

- Which set of SNPs explain the phenotype?

Genotype	Phenotype
ATGCAGTT	control
TTGCAGTT	control
CTGCAGTT	control
ATGCGGTT	case
TTGCGGTT	case
CTGCCGTT	case

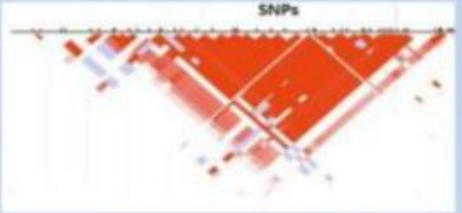
SNP

GWAS workflow

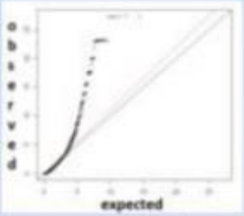
Large cohort (>1000) of cases and controls

Get genome information with SNP arrays

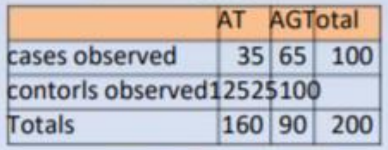
Find deviating from expected haplotypes visualize SNP-SNP interactions using HapMap



Detection of potential association signals and their fine mapping (e.g. detection of LD, stratification effect)

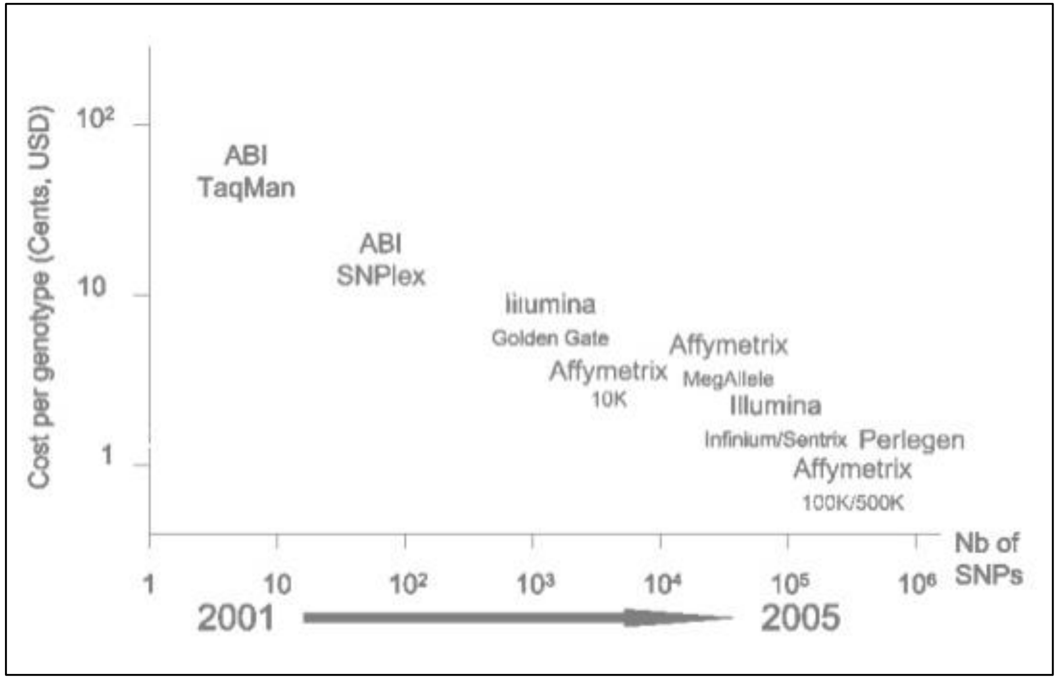


Replication of detected association in new cohot / subset for validation purposes



	AT	AG	Total
cases observed	35	65	100
contorls observed	125	25	100
Totals	160	90	200

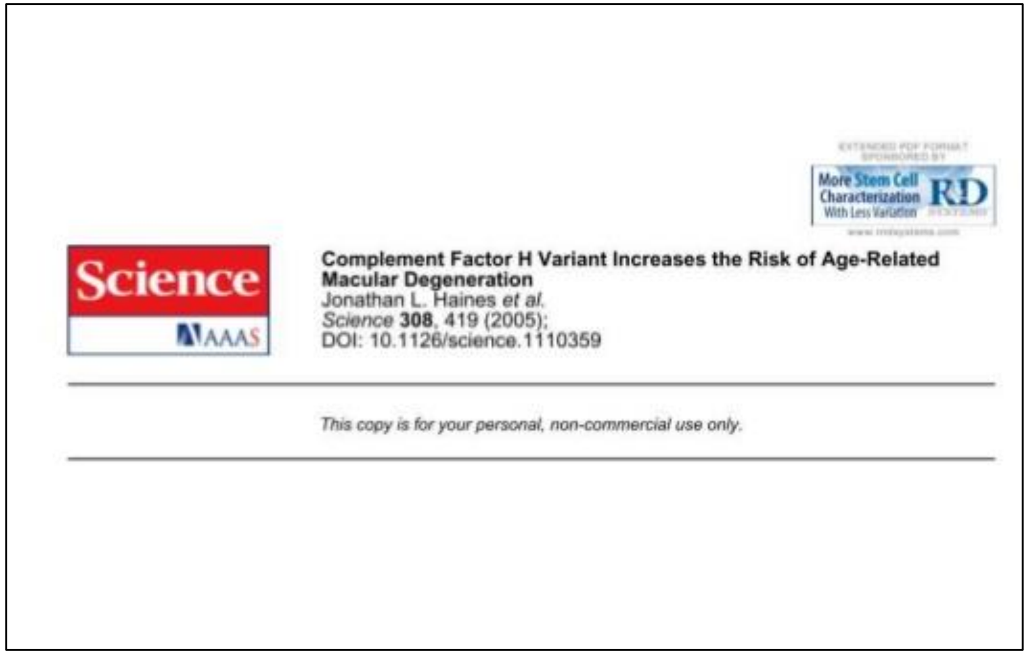
Biological / clinical validation



Running a GWAS: Getting your genotype data

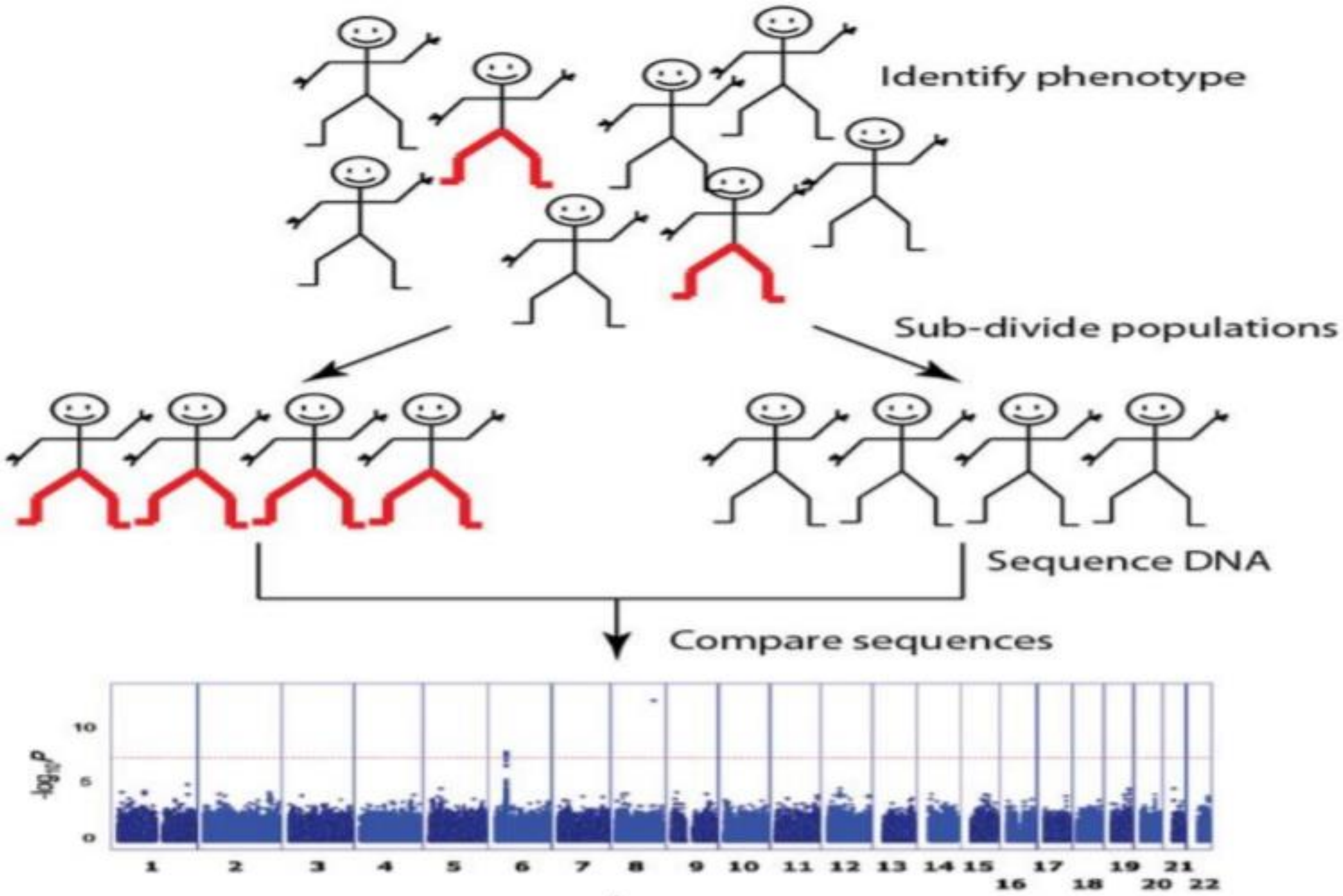
- Select your chip
- Complete your genotyping

The era of hypothesis generating research



Complement Factor H Variant Increases the Risk of Age-Related Macular Degeneration
Jonathan L. Haines *et al.*
Science **308**, 419 (2005);
DOI: 10.1126/science.1110359

This copy is for your personal, non-commercial use only.



Identify phenotype

Sub-divide populations

Sequence DNA

Compare sequences

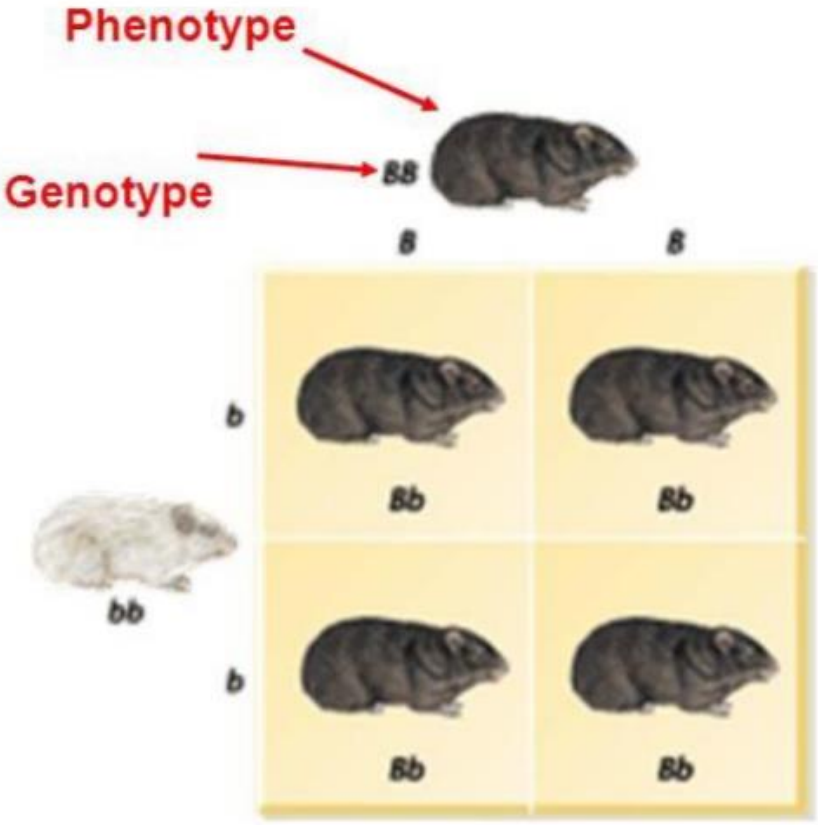
Identify SNPs

	Chromosomal Region 1	Chromosomal Region 2	Chromosomal Region 3	
Person 1	ACTTACGATCGA TGAATGCTAGCT	GTACTGTGGATA CATGACACCTAT	GCTATAGAGGG CGATATCTCCC	Person 1
Person 2	ACTTAAGATCGA TGAATTCTAGCT	GTACTATGGATA CATGATACCTAT	GCTATTGAGGG CGATAACTCCC	Person 2
Person 3	ACTTACGATCGA TGAATGCTAGCT	GTACTGTGGATA CATGACACCTAT	GCTATAGAGGG CGATATCTCCC	Person 3
	SNP1	SNP2	SNP3	

Verify

Relationship between Genotypes and Phenotypes

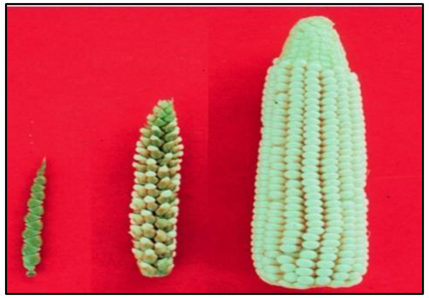
- **Genotype:** Indicates the alleles that the organism has inherited regarding a particular trait.
- **Phenotype:** The actual visible trait of the organism.



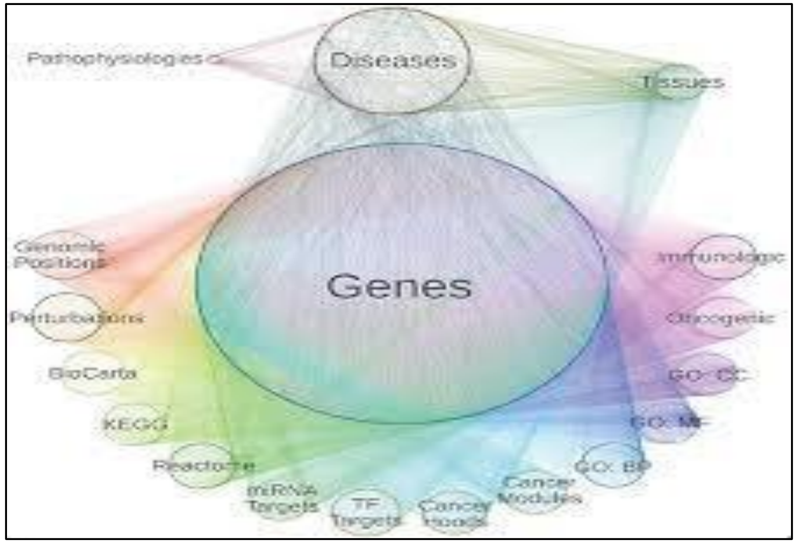
Uses of GWAS

➢Identify genes that are responsible for traits of interest:

- Humans
- Animals
- Plants

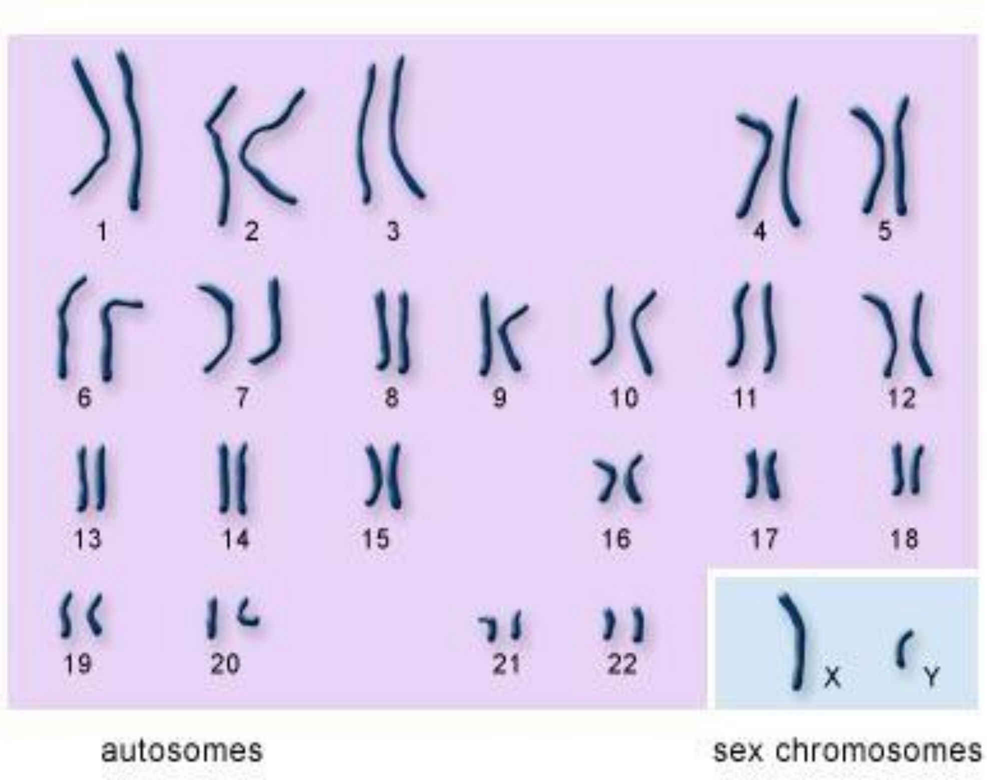


➢Understanding biological mechanisms related to the trait of interest



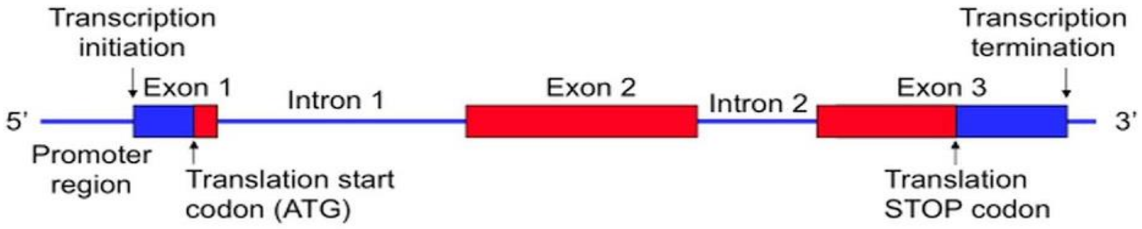
Human Genome Statistics

- Number of Chromosomes : 23 pairs
- Genome Size : 3,079,843,747 Base pairs
- No of Genes : 32,185

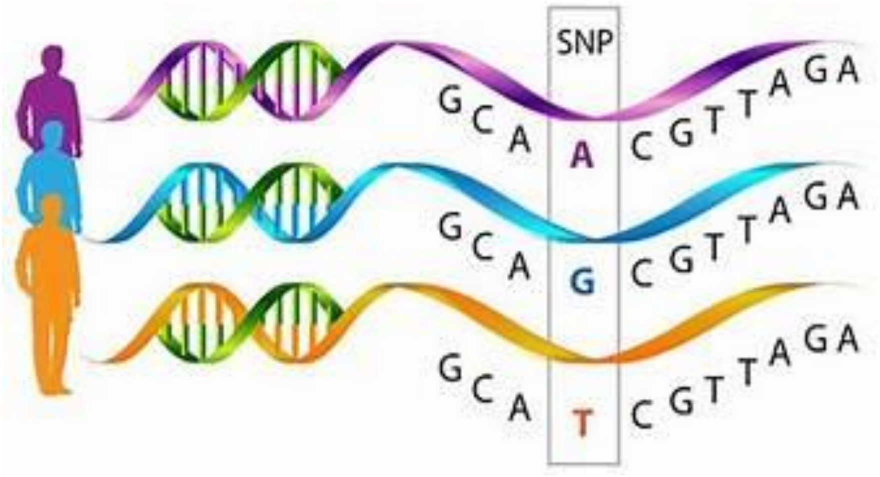


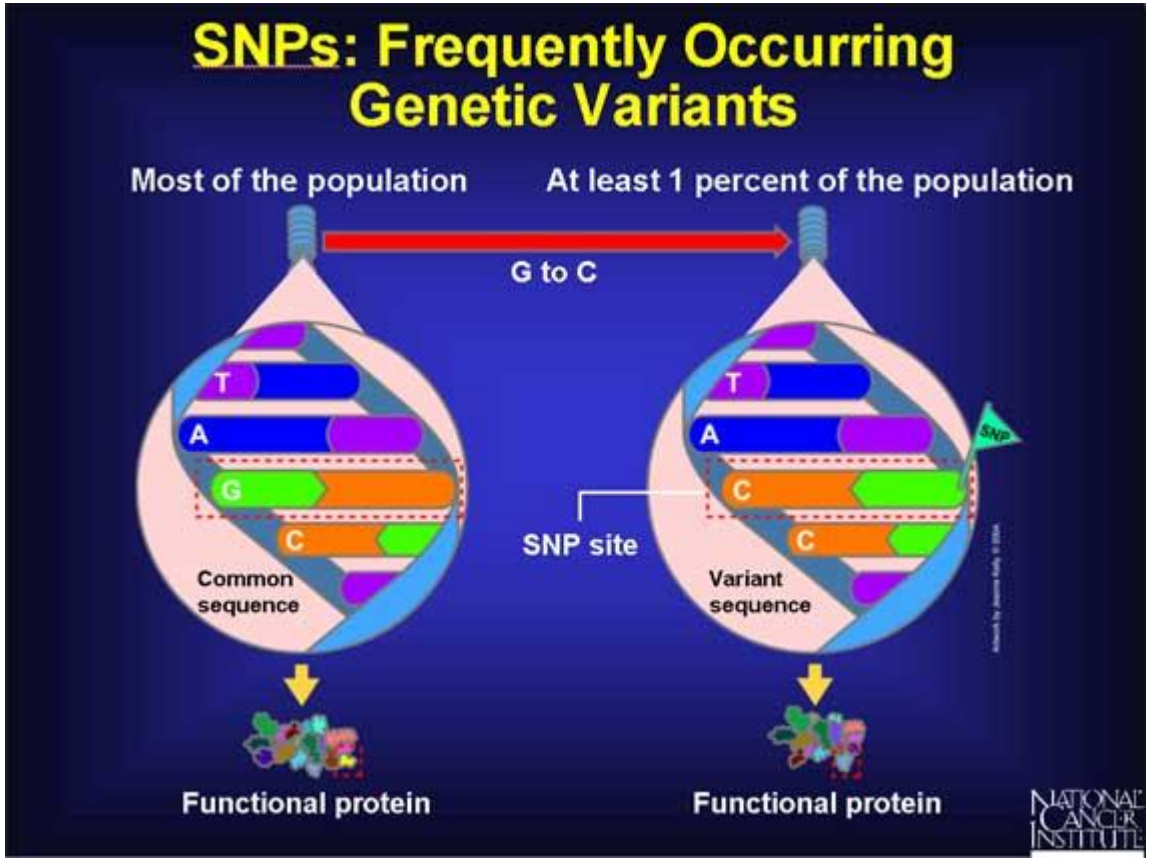
Gene: **This is a sequence of nucleotides in the DNA that codes for a molecule (e.g., a protein)**

Gene Structure



IMPORTANT FINDING





SNPs: Frequently Occurring Genetic Variants

Let us identify signal (in from of SNPs) from GWAS DATA

PLINK : Introduction

- PLINK is a free, open-source designed to perform a range of basic, large-scale analyses in a computationally efficient manner.
- PLINK is whole genome association analysis tool.
- PLINK has a well documented manual.
- Available for linux, MAC ansd MAC-DOS.
- Command line version is faster than graphical PLINK.

PLINK : Multi-feature tool

- Merge two or more files
- Extracts subsets (SNPs or individuals)
- Compress data in a binary file format
- PLINK has numerous useful features for managing and analyzing genetic data
- Read data in a variety of formats
- Recode and reorder files

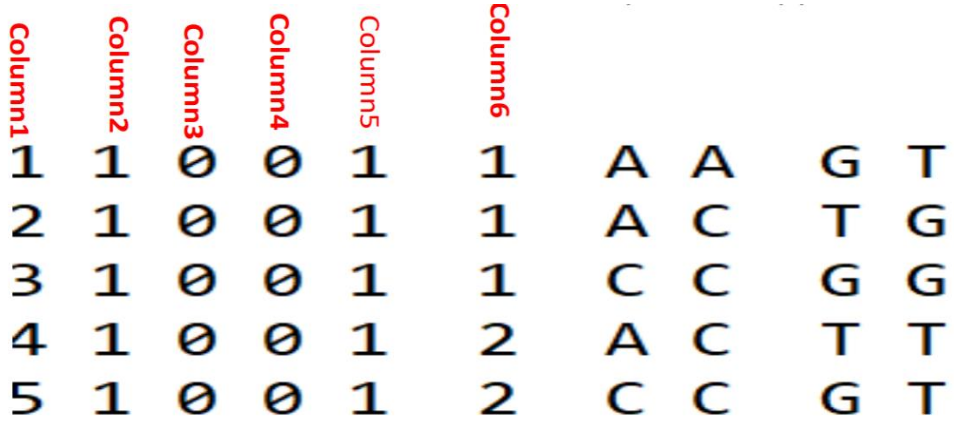
Input Files

- **Genotype data is a text file**
- **Pedigree file (.ped)**
- **Map file (.map)**
- **Genotype data is a compressed binary file**
- **Fam File (.fam)**
- **Bim file (.bim)**
- **Bed file (.bed)**

PED Input File

Pedigree File - the first six columns are mandatory:

- **Family ID**
- **Individual ID**
- **Paternal ID**
- **Maternal ID**
- **Sex (1=male; 2=female; other=unknown)**
- **Phenotype**


Column1
Column2
Column3
Column4
Column5
Column6

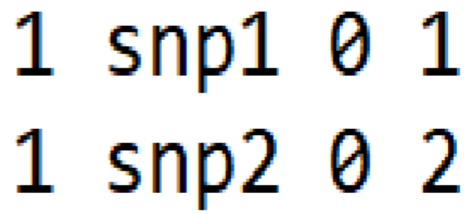
```
1 1 0 0 1  1  A A  G T
2 1 0 0 1  1  A C  T G
3 1 0 0 1  1  C C  G G
4 1 0 0 1  2  A C  T T
5 1 0 0 1  2  C C  G T
```

MAP Input File

MAP File has 4 columns:

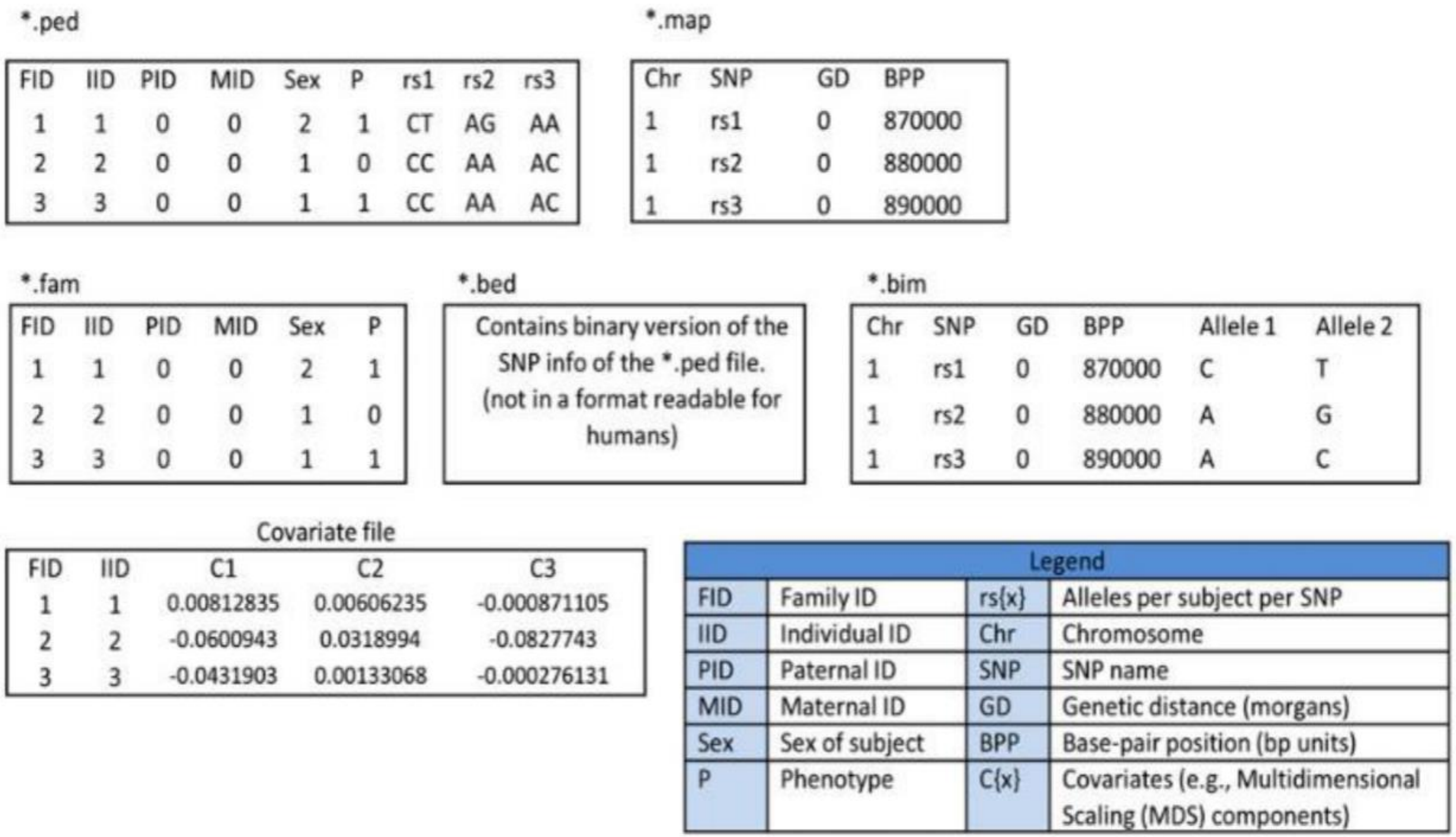
- **chromosome (1-22, X, Y or 0 if unplaced)**
- **rs# or snp identifier**
- **Genetic distance (morgans)**
- **Base-pair position (bp units)**

Column1 Column2 Column3 Column4



```
1 snp1 0 1
1 snp2 0 2
```

Others Input File



*.ped

FID	IID	PID	MID	Sex	P	rs1	rs2	rs3
1	1	0	0	2	1	CT	AG	AA
2	2	0	0	1	0	CC	AA	AC
3	3	0	0	1	1	CC	AA	AC

*.map

Chr	SNP	GD	BPP
1	rs1	0	870000
1	rs2	0	880000
1	rs3	0	890000

*.fam

FID	IID	PID	MID	Sex	P
1	1	0	0	2	1
2	2	0	0	1	0
3	3	0	0	1	1

*.bed

Contains binary version of the SNP info of the *.ped file. (not in a format readable for humans)

*.bim

Chr	SNP	GD	BPP	Allele 1	Allele 2
1	rs1	0	870000	C	T
1	rs2	0	880000	A	G
1	rs3	0	890000	A	C

Covariate file

FID	IID	C1	C2	C3
1	1	0.00812835	0.00606235	-0.000871105
2	2	-0.0600943	0.0318994	-0.0827743
3	3	-0.0431903	0.00133068	-0.000276131

Legend			
FID	Family ID	rs{x}	Alleles per subject per SNP
IID	Individual ID	Chr	Chromosome
PID	Paternal ID	SNP	SNP name
MID	Maternal ID	GD	Genetic distance (morgans)
Sex	Sex of subject	BPP	Base-pair position (bp units)
P	Phenotype	C{x}	Covariates (e.g., Multidimensional Scaling (MDS) components)

QC of genetic DATA

- **A vital step that should be part of any GWAS is the use of appropriate QC.**

- **Without extensive QC, GWAS will not generate reliable results because raw genotype data are inherently imperfect.**

- **Errors in the data can arise for numerous reasons, for example, due to poor quality of DNA samples, poor DNA hybridization to the array, poorly performing genotype probes, and sample mix-ups or contamination.**

QC of genetic DATA

The QC steps consist of filtering out of SNPs and individuals based on the following:

(1)individual and SNP missingness,

(2) inconsistencies in assigned and genetic sex of subjects (see sex discrepancy),

(3) minor allele frequency (MAF),

(4) deviations from Hardy–Weinberg equilibrium (HWE),

Important Commands

Step	Command	Function
1: Missingness of SNPs and individuals	--geno	Excludes SNPs that are missing in a large proportion of the subjects. In this step, SNPs with low genotype calls are removed.
	--mind	**Excludes individuals who have high rates of genotype missingness. In this step, individual with low genotype calls are removed.**
2: Sex discrepancy	**--check-sex**	**Checks for discrepancies between sex of the individuals recorded in the dataset and their sex based on X chromosome heterozygosity/homozygosity rates.**
3: Minor allele frequency (MAF)	**--maf**	**Includes only SNPs above the set MAF threshold.**
4: Hardy–Weinberg equilibrium (HWE)	**--hwe**	**Excludes markers which deviate from Hardy–Weinberg equilibrium.**

PLINK SESSION

- Data Preparation
- Quality Control
- Clustering
- GWAS

Example data

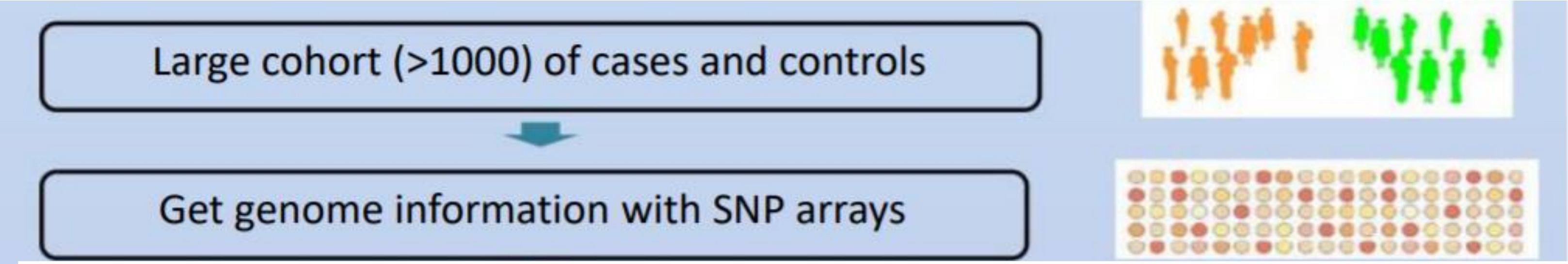
- Download the example data from the course website (PLINK FOLDER)

  – HapMap_3_r3_1.bed

  – HapMap_3_r3_1.bim

  – HapMap_3_r3_1.fam

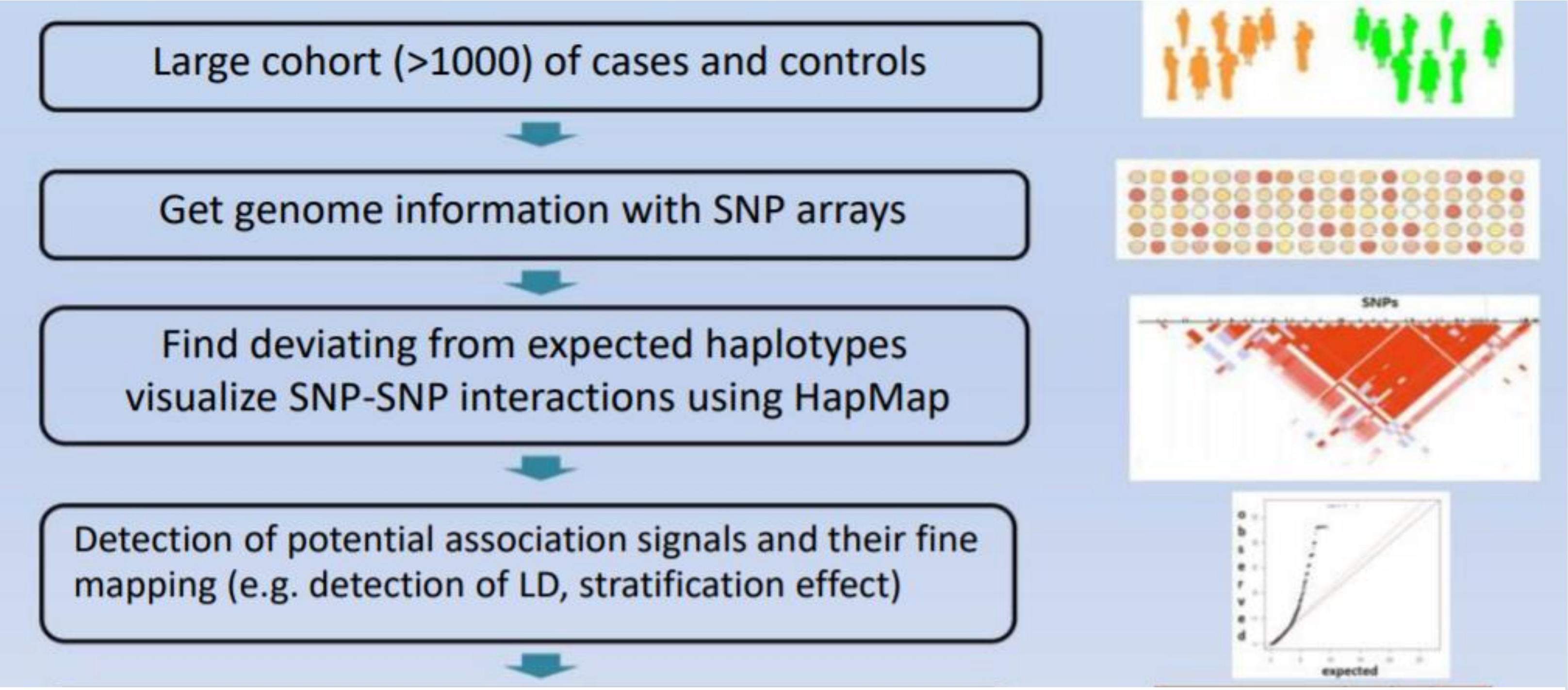
By looking into file extension, BED FORMAT



Large cohort (>1000) of cases and controls

Get genome information with SNP arrays

Here we have sample DATA (as our studied cohort).



Large cohort (>1000) of cases and controls

Get genome information with SNP arrays

Find deviating from expected haplotypes visualize SNP-SNP interactions using HapMap

Detection of potential association signals and their fine mapping (e.g. detection of LD, stratification effect)

Detection of LD, population stratification (comes under Filteration step)

Lets Perform Quality filteration

Quality control processes

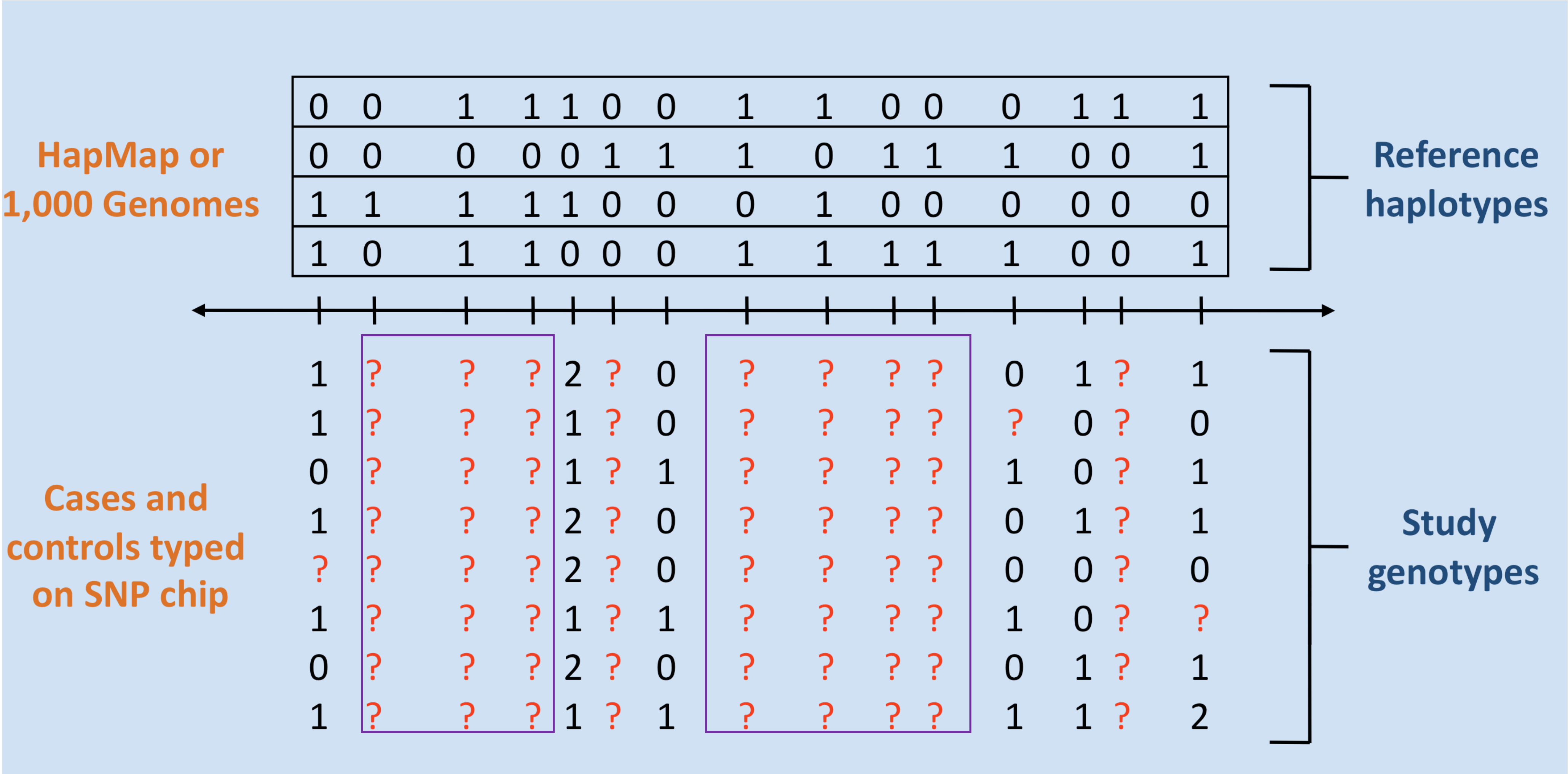
- Missing genotype

- Hardy-Weinberg Equilibrium

- Minor Allele frequency

- Linkage disequilibrium pruning

Missing genotype (1)



HapMap or 1,000 Genomes

Reference haplotypes

0	0	1	1	1	0	0	1	1	0	0	0	1	1	1
0	0	0	0	0	1	1	1	0	1	1	1	0	0	1
1	1	1	1	1	0	0	0	1	0	0	0	0	0	0
1	0	1	1	0	0	0	1	1	1	1	1	0	0	1

Cases and controls typed on SNP chip

Study genotypes

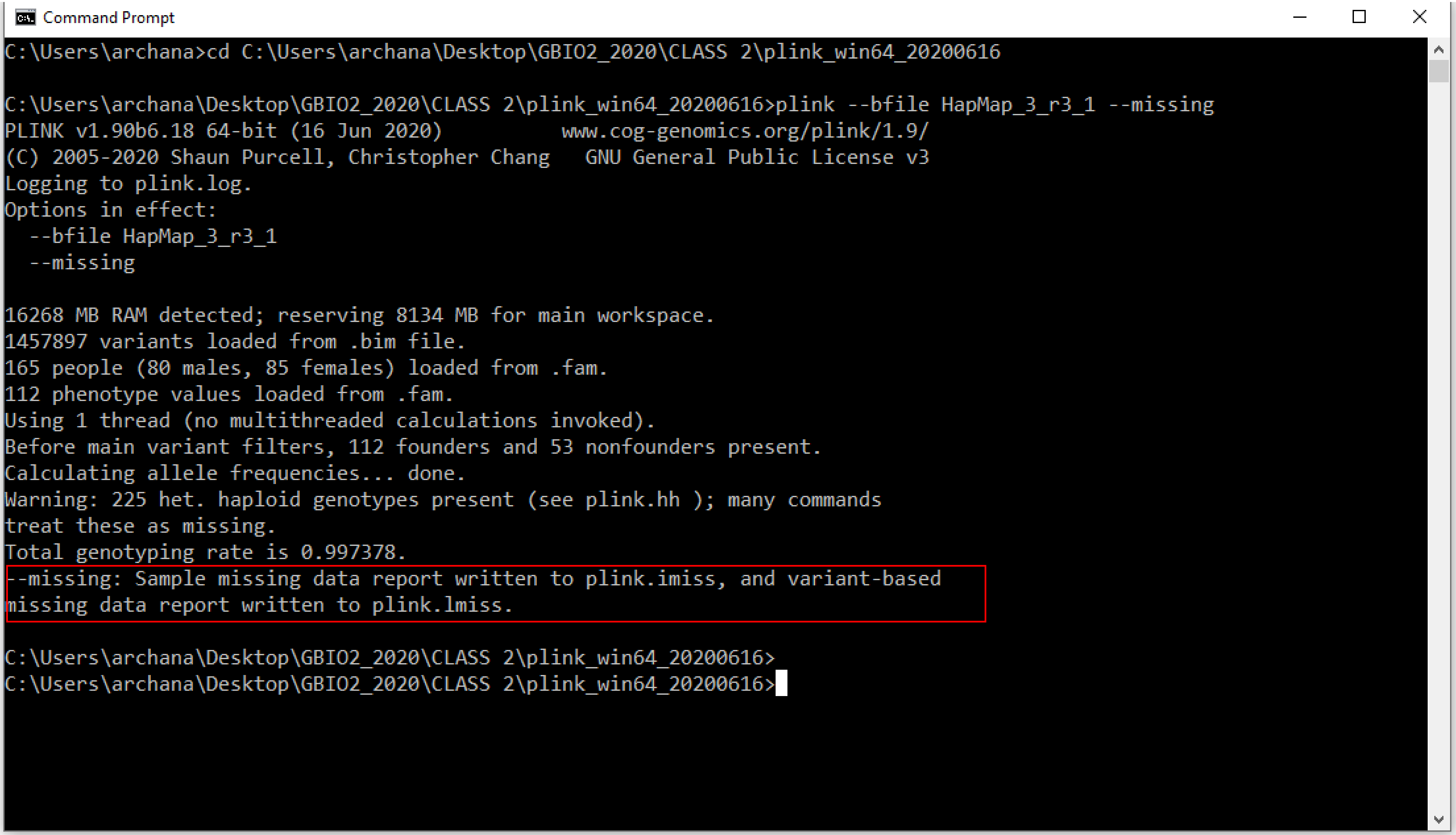
1	?	?	?	2	?	0	?	?	?	?	0	1	?	1
1	?	?	?	1	?	0	?	?	?	?	?	0	?	0
0	?	?	?	1	?	1	?	?	?	?	1	0	?	1
1	?	?	?	2	?	0	?	?	?	?	0	1	?	1
?	?	?	?	2	?	0	?	?	?	?	0	0	?	0
1	?	?	?	1	?	1	?	?	?	?	1	0	?	?
0	?	?	?	2	?	0	?	?	?	?	0	1	?	1
1	?	?	?	1	?	1	?	?	?	?	1	1	?	2

Missing genotype (2)

- Download Example files from website
- Copy all Files in PLINK Directory

```
plink --bfile HapMap_3_r3_1 --missing
```

- output:
    - plink.imiss and
    - plink.lmiss,
- These files show respectively the proportion of missing SNPs per individual and the proportion of missing individuals per SNP.



```
C:\Users\archana>cd C:\Users\archana\Desktop\GBIO2_2020\CLASS 2\plink_win64_20200616

C:\Users\archana\Desktop\GBIO2_2020\CLASS 2\plink_win64_20200616>plink --bfile HapMap_3_r3_1 --missing
PLINK v1.90b6.18 64-bit (16 Jun 2020)          www.cog-genomics.org/plink/1.9/
(C) 2005-2020 Shaun Purcell, Christopher Chang   GNU General Public License v3
Logging to plink.log.
Options in effect:
  --bfile HapMap_3_r3_1
  --missing

16268 MB RAM detected; reserving 8134 MB for main workspace.
1457897 variants loaded from .bim file.
165 people (80 males, 85 females) loaded from .fam.
112 phenotype values loaded from .fam.
Using 1 thread (no multithreaded calculations invoked).
Before main variant filters, 112 founders and 53 nonfounders present.
Calculating allele frequencies... done.
Warning: 225 het. haploid genotypes present (see plink.hh ); many commands
treat these as missing.
Total genotyping rate is 0.997378.
--missing: Sample missing data report written to plink.imiss, and variant-based
missing data report written to plink.lmiss.

C:\Users\archana\Desktop\GBIO2_2020\CLASS 2\plink_win64_20200616>
C:\Users\archana\Desktop\GBIO2_2020\CLASS 2\plink_win64_20200616>
```

Missing genotype (3)

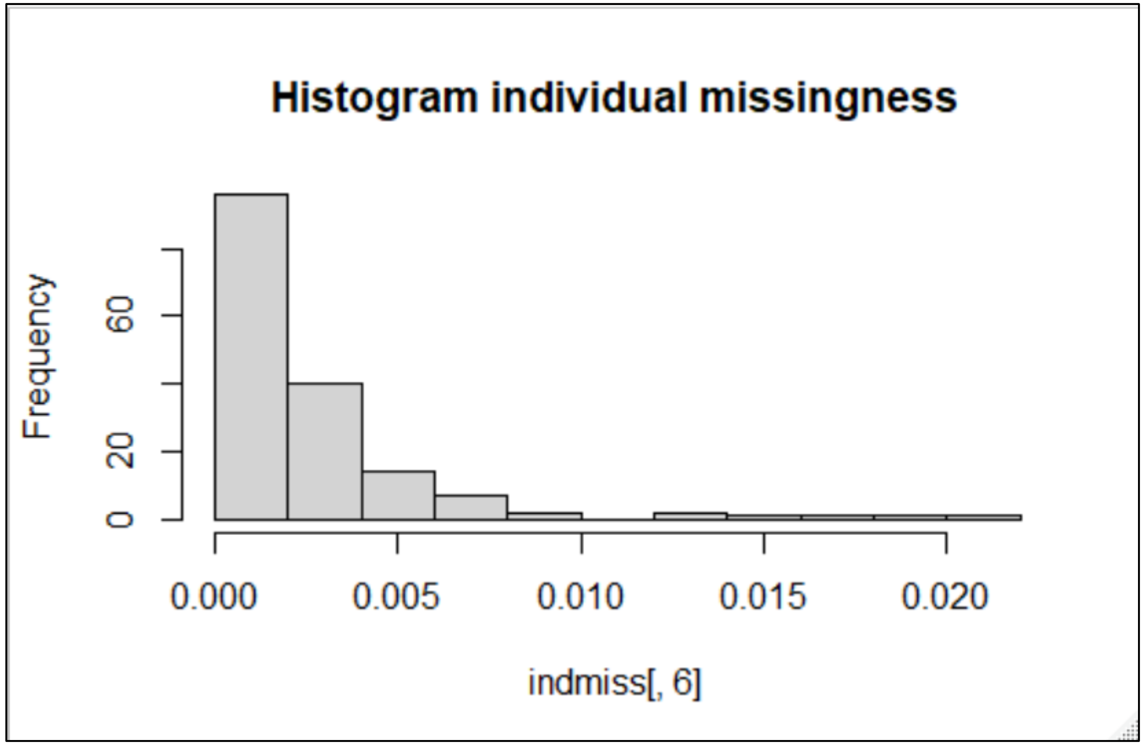
# Generate plots

indmiss<-read.table(file="plink.imiss", header=TRUE)
snpmiss<-read.table(file="plink.lmiss", header=TRUE)

hist(indmiss[,6],main="Histogram individual missingness")
#selects column 6, names header of file

hist(snpmiss[,5],main="Histogram SNP missingness")

#selects column 5, names header of file



Missing Rate Per Person (1)

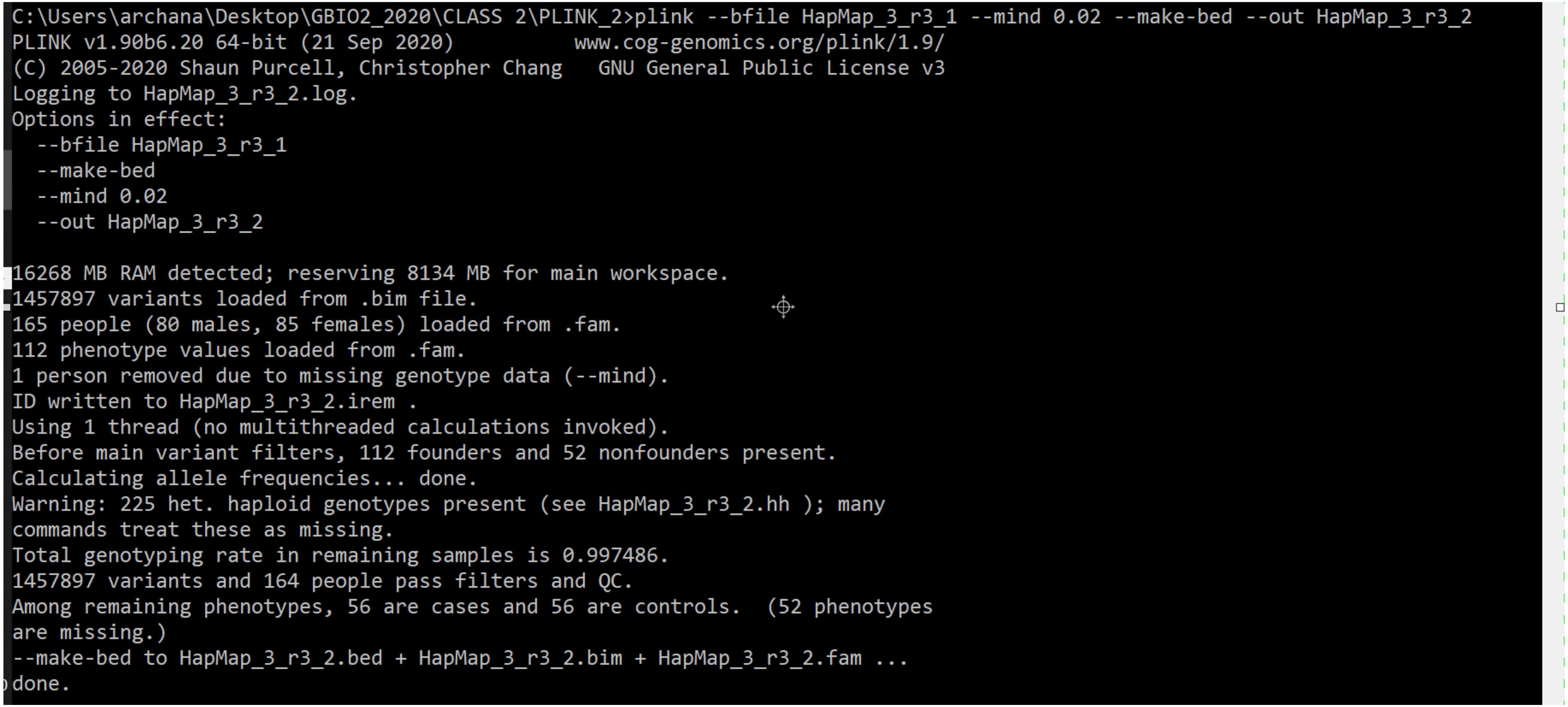
- **The initial step in all data analysis is to exclude individuals with too much missing Genotype data.**

- **A line in the terminal will appear, indicating how many individuals were removed due to low genotyping. If any individuals were removed, a file called plink.irem will be created, listing the Family and Individual IDs of these removed individuals.**

Missing Rate Per Person (2)

# Delete individuals with missingness >0.02.

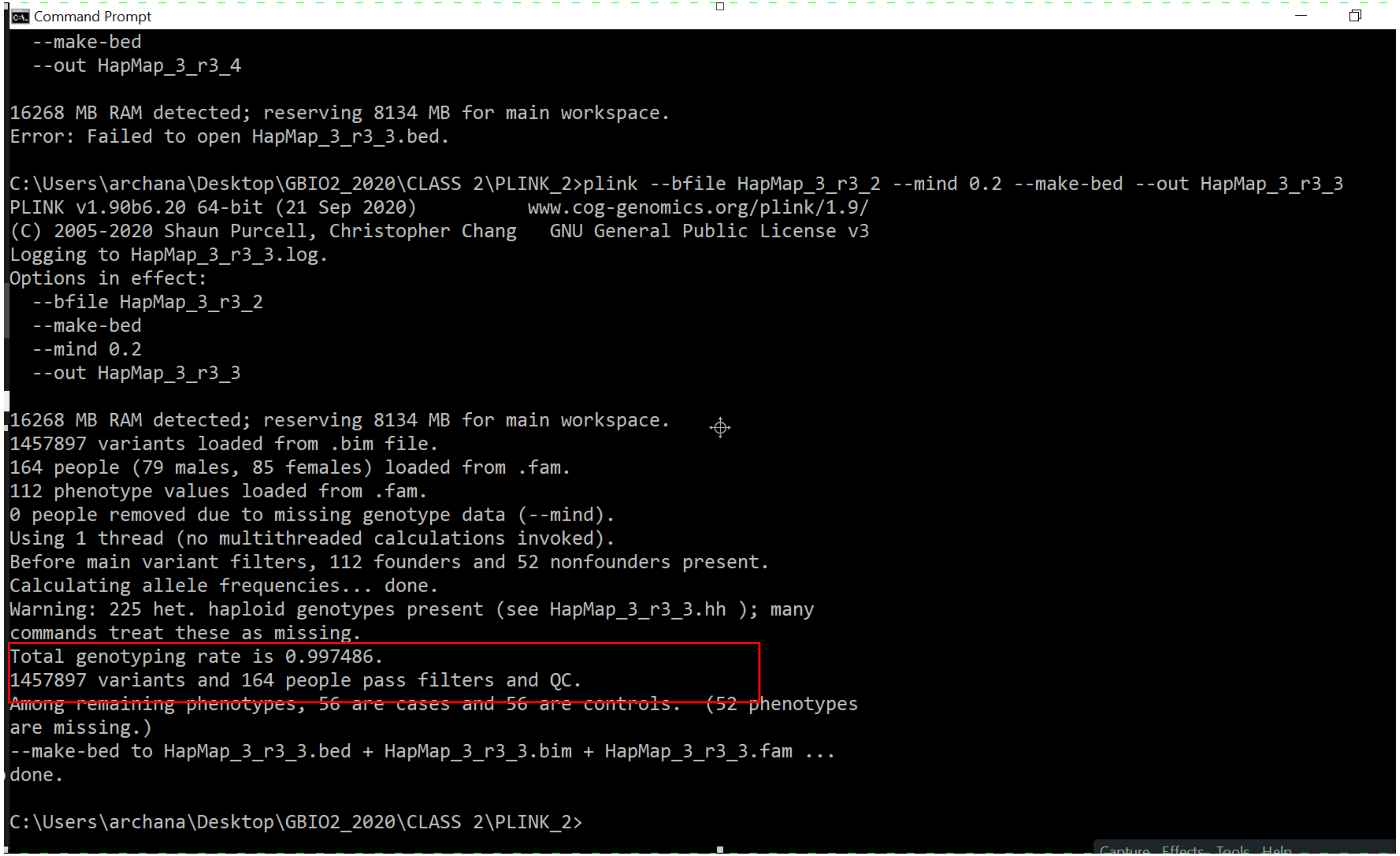
plink --bfile HapMap_3_r3_1 --mind 0.02 --make-bed --out HapMap_3_r3_2



```
C:\Users\archana\Desktop\GBIO2_2020\CLASS 2\PLINK_2>plink --bfile HapMap_3_r3_1 --mind 0.02 --make-bed --out HapMap_3_r3_2
PLINK v1.90b6.20 64-bit (21 Sep 2020)          www.cog-genomics.org/plink/1.9/
(C) 2005-2020 Shaun Purcell, Christopher Chang   GNU General Public License v3
Logging to HapMap_3_r3_2.log.
Options in effect:
  --bfile HapMap_3_r3_1
  --make-bed
  --mind 0.02
  --out HapMap_3_r3_2

16268 MB RAM detected; reserving 8134 MB for main workspace.
1457897 variants loaded from .bim file.
165 people (80 males, 85 females) loaded from .fam.
112 phenotype values loaded from .fam.
1 person removed due to missing genotype data (--mind).
ID written to HapMap_3_r3_2.irem .
Using 1 thread (no multithreaded calculations invoked).
Before main variant filters, 112 founders and 52 nonfounders present.
Calculating allele frequencies... done.
Warning: 225 het. haploid genotypes present (see HapMap_3_r3_2.hh ); many
commands treat these as missing.
Total genotyping rate in remaining samples is 0.997486.
1457897 variants and 164 people pass filters and QC.
Among remaining phenotypes, 56 are cases and 56 are controls.  (52 phenotypes
are missing.)
--make-bed to HapMap_3_r3_2.bed + HapMap_3_r3_2.bim + HapMap_3_r3_2.fam ...
done.
```

Missing Rate Per Person (3)

plink --bfile HapMap_3_r3_2 --mind 0.2 --make-bed --out HapMap_3_r3_3



```
  --make-bed
  --out HapMap_3_r3_4

16268 MB RAM detected; reserving 8134 MB for main workspace.
Error: Failed to open HapMap_3_r3_3.bed.

C:\Users\archana\Desktop\GBIO2_2020\CLASS 2\PLINK_2>plink --bfile HapMap_3_r3_2 --mind 0.2 --make-bed --out HapMap_3_r3_3
PLINK v1.90b6.20 64-bit (21 Sep 2020)          www.cog-genomics.org/plink/1.9/
(C) 2005-2020 Shaun Purcell, Christopher Chang   GNU General Public License v3
Logging to HapMap_3_r3_3.log.
Options in effect:
  --bfile HapMap_3_r3_2
  --make-bed
  --mind 0.2
  --out HapMap_3_r3_3

16268 MB RAM detected; reserving 8134 MB for main workspace.
1457897 variants loaded from .bim file.
164 people (79 males, 85 females) loaded from .fam.
112 phenotype values loaded from .fam.
0 people removed due to missing genotype data (--mind).
Using 1 thread (no multithreaded calculations invoked).
Before main variant filters, 112 founders and 52 nonfounders present.
Calculating allele frequencies... done.
Warning: 225 het. haploid genotypes present (see HapMap_3_r3_3.hh ); many
commands treat these as missing.
Total genotyping rate is 0.997486.
1457897 variants and 164 people pass filters and QC.
Among remaining phenotypes, 56 are cases and 56 are controls.  (52 phenotypes
are missing.)
--make-bed to HapMap_3_r3_3.bed + HapMap_3_r3_3.bim + HapMap_3_r3_3.fam ...
done.

C:\Users\archana\Desktop\GBIO2_2020\CLASS 2\PLINK_2>
```

Missing Rate Per SNP (1)

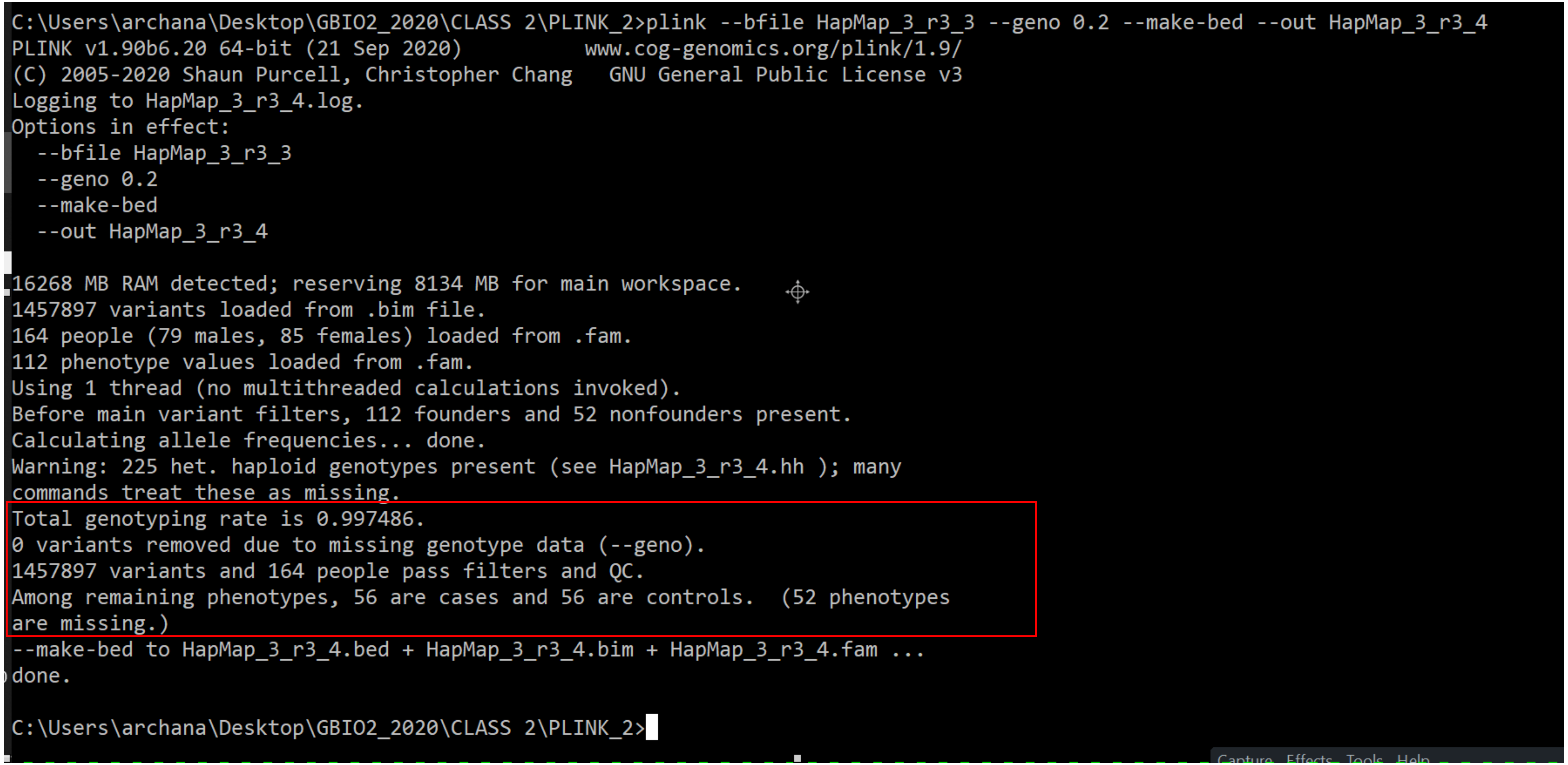
- Subsequent analyses can be set to automatically exclude SNPs on the basis of missing genotype rate, with the --geno option: the default is to include all SNPS (i.e. --geno 1).
- To include only SNPs with a 90% genotyping rate (10% missing) use

```
--bfile file --geno 0.1
```

- As with the --maf option, these counts are calculated after removing individuals with high missing genotype rates.

Missing Rate Per SNP(2)

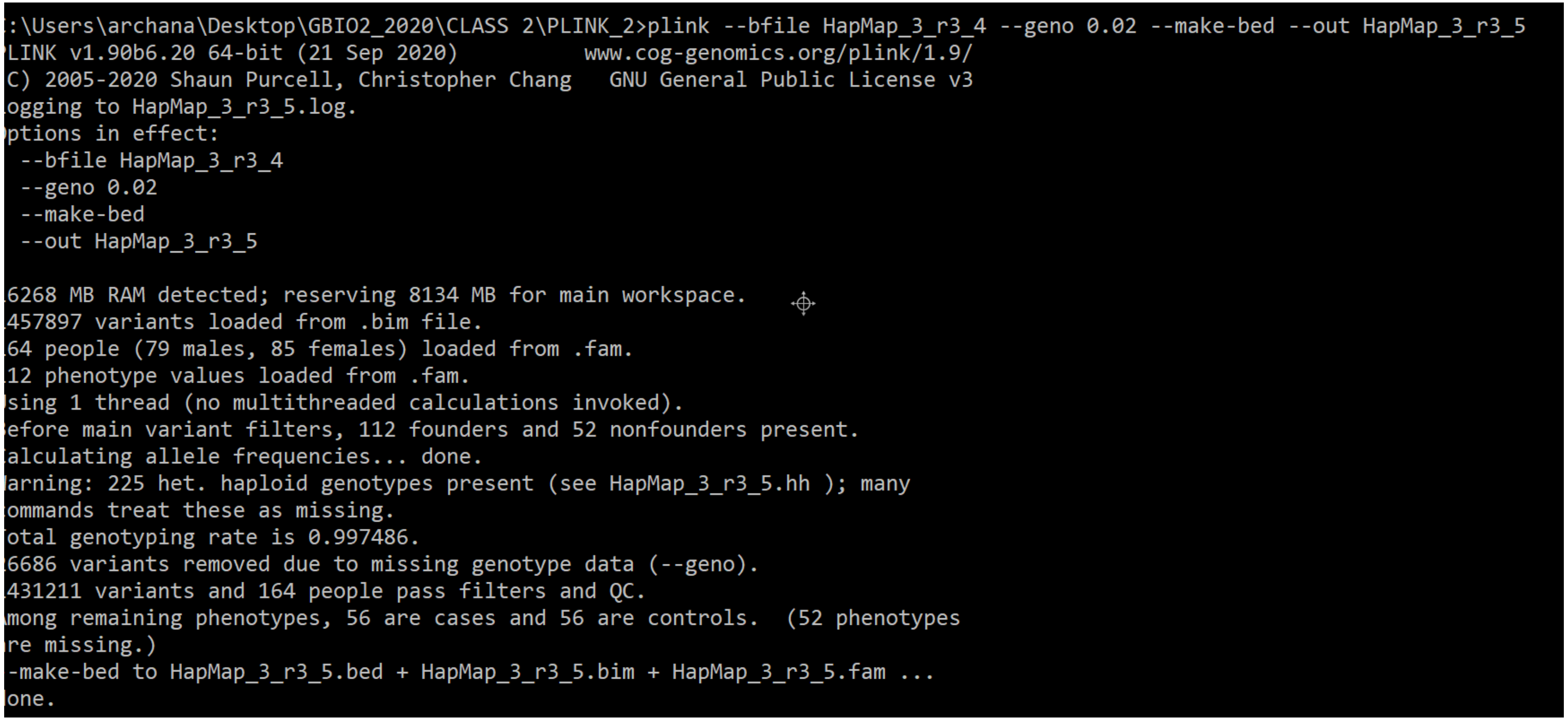
plink --bfile HapMap_3_r3_3 --geno 0.2 --make-bed --out HapMap_3_r3_4



```
C:\Users\archana\Desktop\GBIO2_2020\CLASS 2\PLINK_2>plink --bfile HapMap_3_r3_3 --geno 0.2 --make-bed --out HapMap_3_r3_4
PLINK v1.90b6.20 64-bit (21 Sep 2020)          www.cog-genomics.org/plink/1.9/
(C) 2005-2020 Shaun Purcell, Christopher Chang   GNU General Public License v3
Logging to HapMap_3_r3_4.log.
Options in effect:
  --bfile HapMap_3_r3_3
  --geno 0.2
  --make-bed
  --out HapMap_3_r3_4

16268 MB RAM detected; reserving 8134 MB for main workspace.
1457897 variants loaded from .bim file.
164 people (79 males, 85 females) loaded from .fam.
112 phenotype values loaded from .fam.
Using 1 thread (no multithreaded calculations invoked).
Before main variant filters, 112 founders and 52 nonfounders present.
Calculating allele frequencies... done.
Warning: 225 het. haploid genotypes present (see HapMap_3_r3_4.hh ); many
commands treat these as missing.
Total genotyping rate is 0.997486.
0 variants removed due to missing genotype data (--geno).
1457897 variants and 164 people pass filters and QC.
Among remaining phenotypes, 56 are cases and 56 are controls.  (52 phenotypes
are missing.)
--make-bed to HapMap_3_r3_4.bed + HapMap_3_r3_4.bim + HapMap_3_r3_4.fam ...
done.

C:\Users\archana\Desktop\GBIO2_2020\CLASS 2\PLINK_2>
```

Missing Rate Per SNP : Delete SNPs

# Delete SNPs with missingness >0.02.
plink --bfile HapMap_3_r3_4 --geno 0.02 --make-bed --out HapMap_3_r3_5



```
:\Users\archana\Desktop\GBIO2_2020\CLASS 2\PLINK_2>plink --bfile HapMap_3_r3_4 --geno 0.02 --make-bed --out HapMap_3_r3_5
LINK v1.90b6.20 64-bit (21 Sep 2020)          www.cog-genomics.org/plink/1.9/
C) 2005-2020 Shaun Purcell, Christopher Chang   GNU General Public License v3
ogging to HapMap_3_r3_5.log.
ptions in effect:
 --bfile HapMap_3_r3_4
 --geno 0.02
 --make-bed
 --out HapMap_3_r3_5

6268 MB RAM detected; reserving 8134 MB for main workspace.
457897 variants loaded from .bim file.
64 people (79 males, 85 females) loaded from .fam.
12 phenotype values loaded from .fam.
sing 1 thread (no multithreaded calculations invoked).
efore main variant filters, 112 founders and 52 nonfounders present.
alculating allele frequencies... done.
arning: 225 het. haploid genotypes present (see HapMap_3_r3_5.hh ); many
ommands treat these as missing.
otal genotyping rate is 0.997486.
6686 variants removed due to missing genotype data (--geno).
431211 variants and 164 people pass filters and QC.
mong remaining phenotypes, 56 are cases and 56 are controls.  (52 phenotypes
re missing.)
-make-bed to HapMap_3_r3_5.bed + HapMap_3_r3_5.bim + HapMap_3_r3_5.fam ...
one.
```

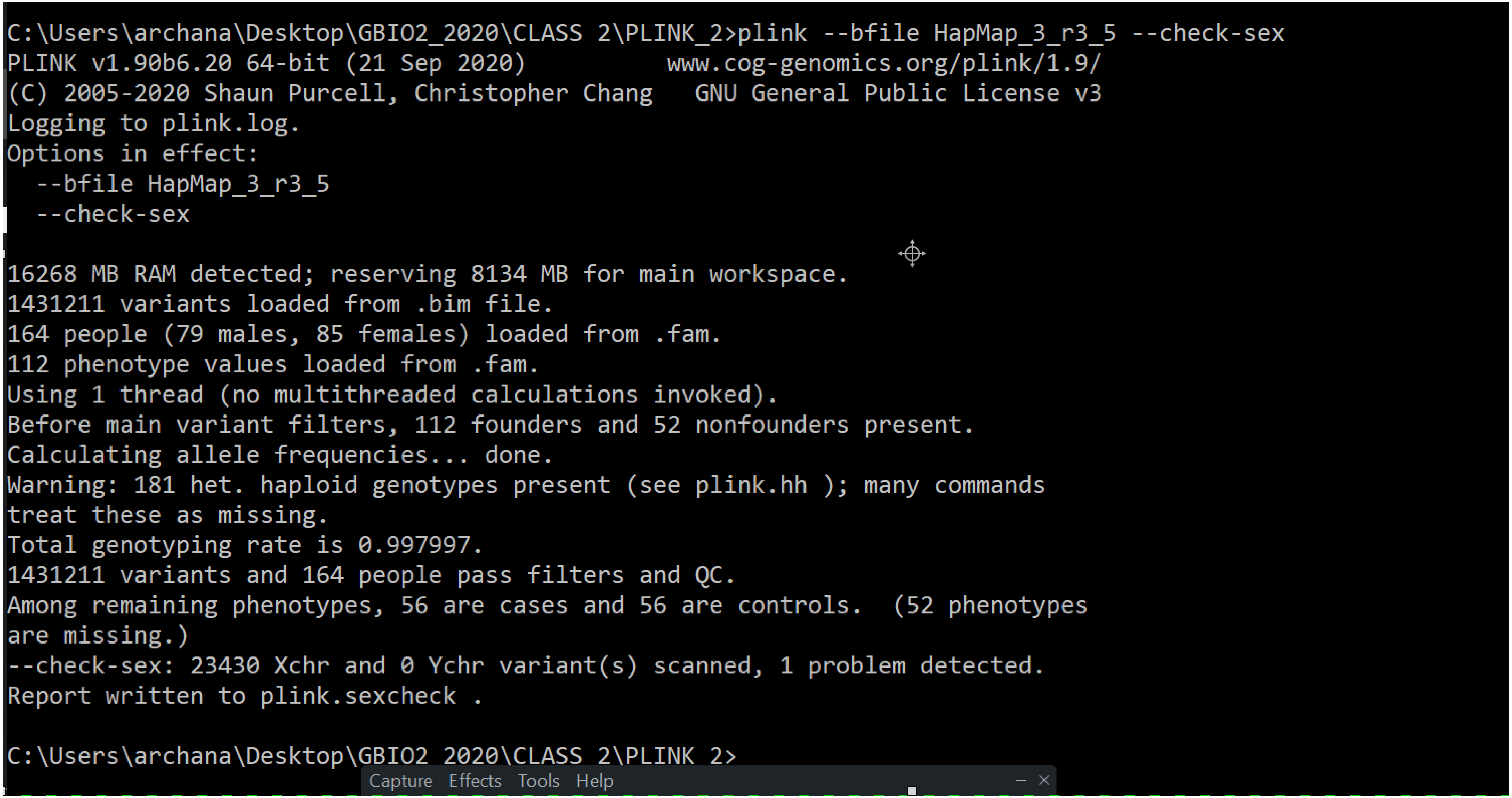
Check for sex discrepancy

- **Subjects who were a priori determined as females must have a F value of <0.2, and subjects who were a priori determined as males must have a F value >0.8.**

- **This F value is based on the X chromosome inbreeding (homozygosity) estimate.**

- **Subjects who do not fulfil these requirements are flagged "PROBLEM" by PLINK.**

```
plink --bfile HapMap_3_r3_5 --check-sex
```



```
C:\Users\archana\Desktop\GBIO2_2020\CLASS 2\PLINK_2>plink --bfile HapMap_3_r3_5 --check-sex
PLINK v1.90b6.20 64-bit (21 Sep 2020)          www.cog-genomics.org/plink/1.9/
(C) 2005-2020 Shaun Purcell, Christopher Chang   GNU General Public License v3
Logging to plink.log.
Options in effect:
  --bfile HapMap_3_r3_5
  --check-sex

16268 MB RAM detected; reserving 8134 MB for main workspace.
1431211 variants loaded from .bim file.
164 people (79 males, 85 females) loaded from .fam.
112 phenotype values loaded from .fam.
Using 1 thread (no multithreaded calculations invoked).
Before main variant filters, 112 founders and 52 nonfounders present.
Calculating allele frequencies... done.
Warning: 181 het. haploid genotypes present (see plink.hh ); many commands
treat these as missing.
Total genotyping rate is 0.997997.
1431211 variants and 164 people pass filters and QC.
Among remaining phenotypes, 56 are cases and 56 are controls.  (52 phenotypes
are missing.)
--check-sex: 23430 Xchr and 0 Ychr variant(s) scanned, 1 problem detected.
Report written to plink.sexcheck .

C:\Users\archana\Desktop\GBIO2 2020\CLASS 2\PLINK 2>
```

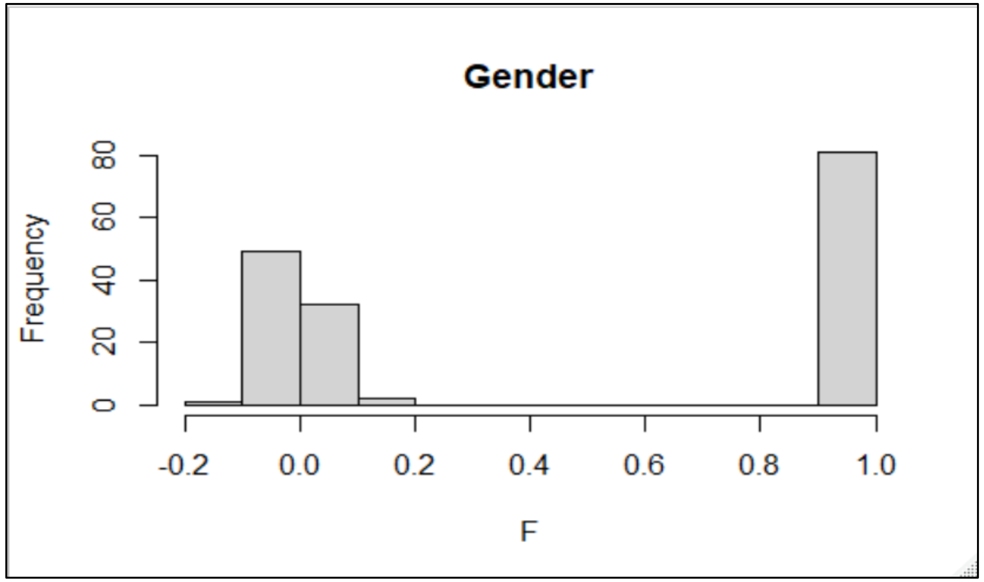
Generate plots to visualize

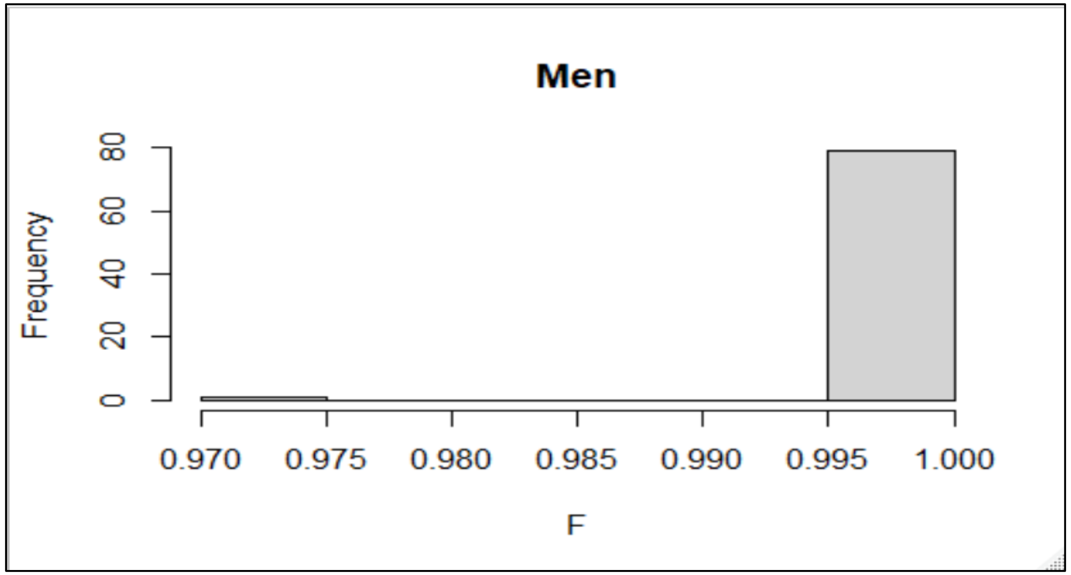
- # These checks indicate that there is one woman with a sex discrepancy, F value of 0.99.

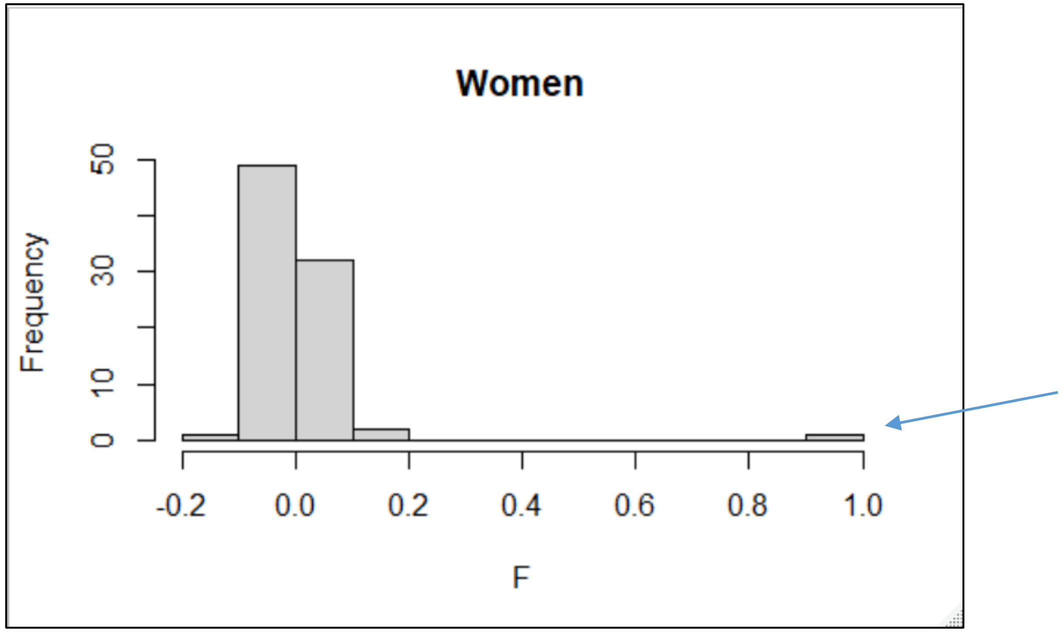
(When using other datasets often a few discrepancies will be found).

```
#READ plink.sexcheck

gender <- read.table(file.choose(), header=T)

hist(gender[,6],main="Gender", xlab="F")

male=subset(gender, gender$PEDSEX==1)
hist(male[,6],main="Men",xlab="F")

female=subset(gender, gender$PEDSEX==2)
hist(female[,6],main="Women",xlab="F")
```

Visualization







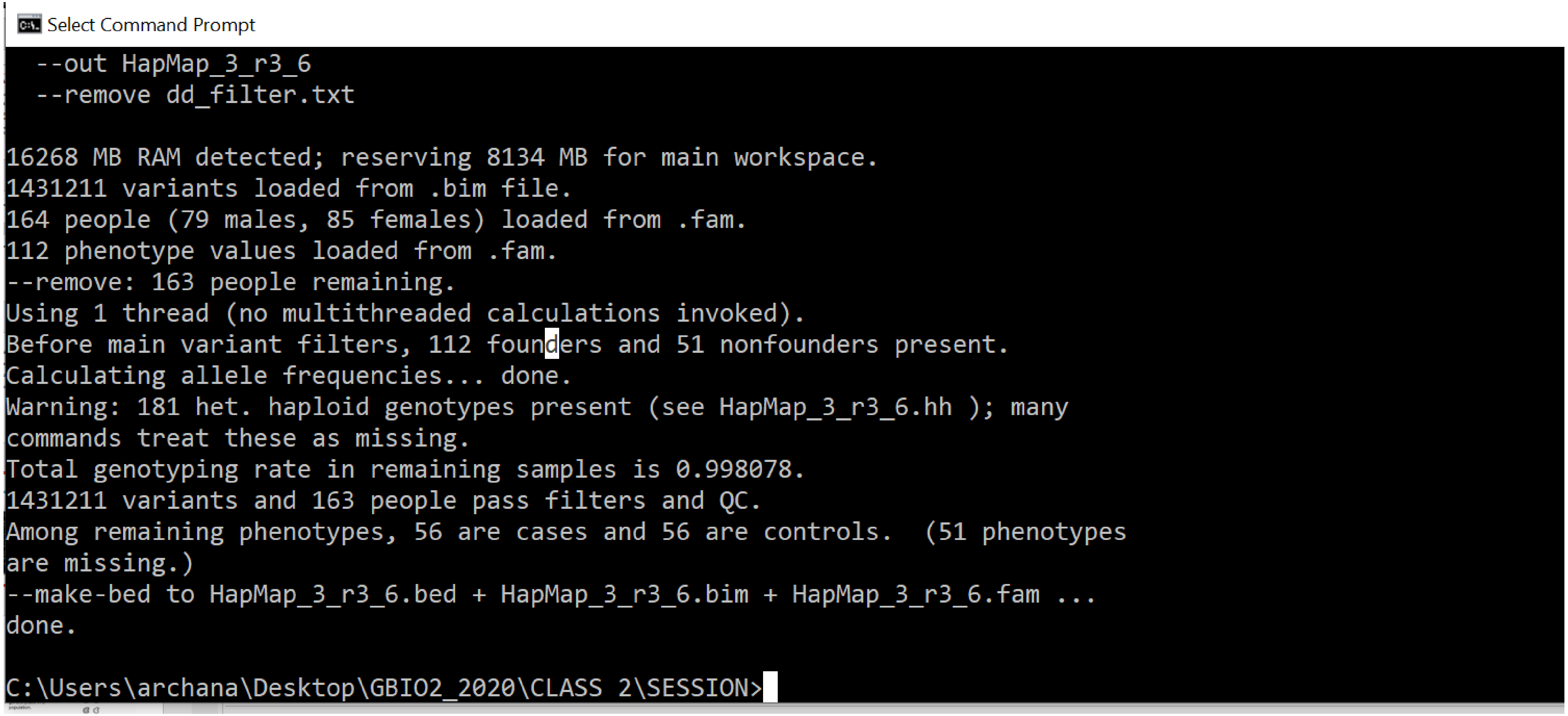
Delete individuals with sex discrepancy (1)

- **Read plink.sexcheck file**
- **Select specific row (164)**
- **Select first two column value**
- **Store information in dd_filter.txt**

Delete individuals with sex discrepancy (2)

- **This command removes the list of individuals with the status "PROBLEM".**

plink --bfile HapMap_3_r3_5 --remove dd_filter.txt --make-bed --out HapMap_3_r3_6

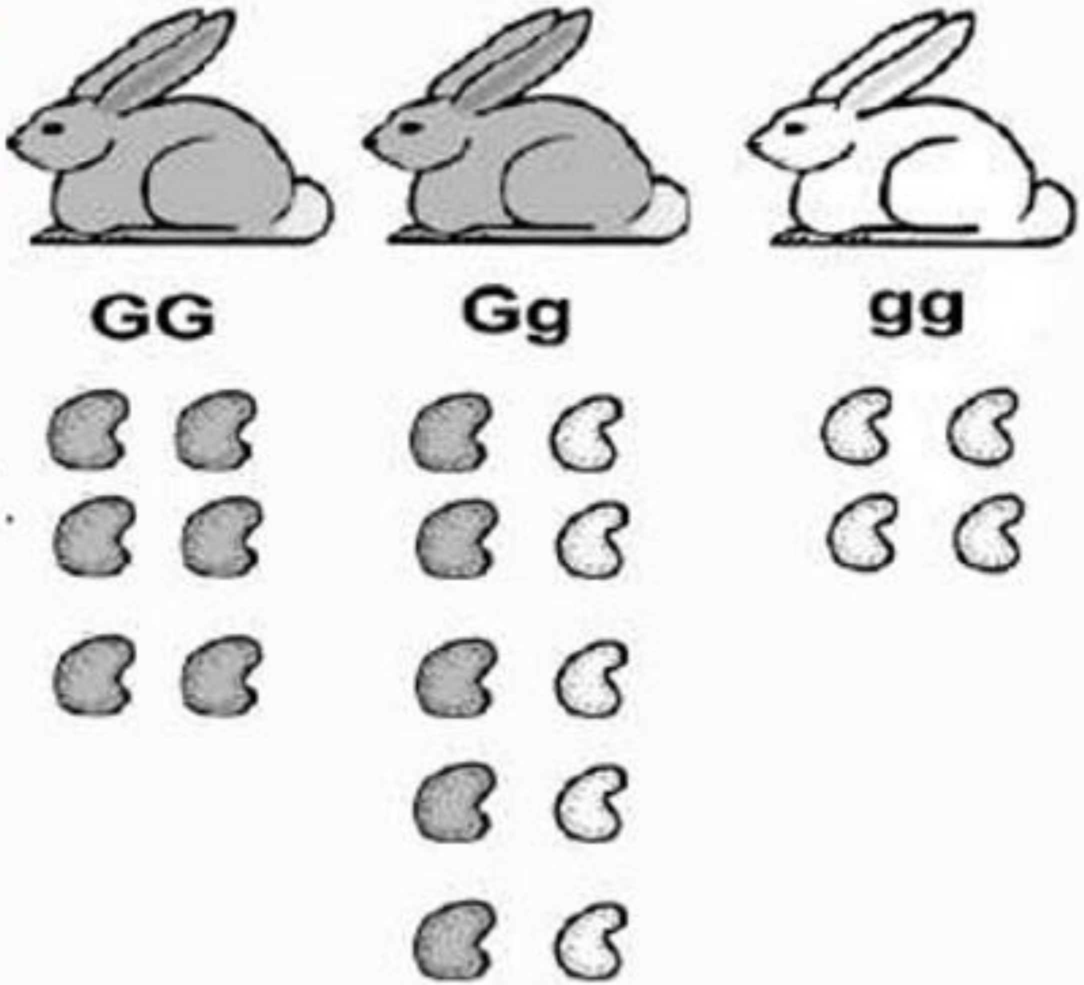


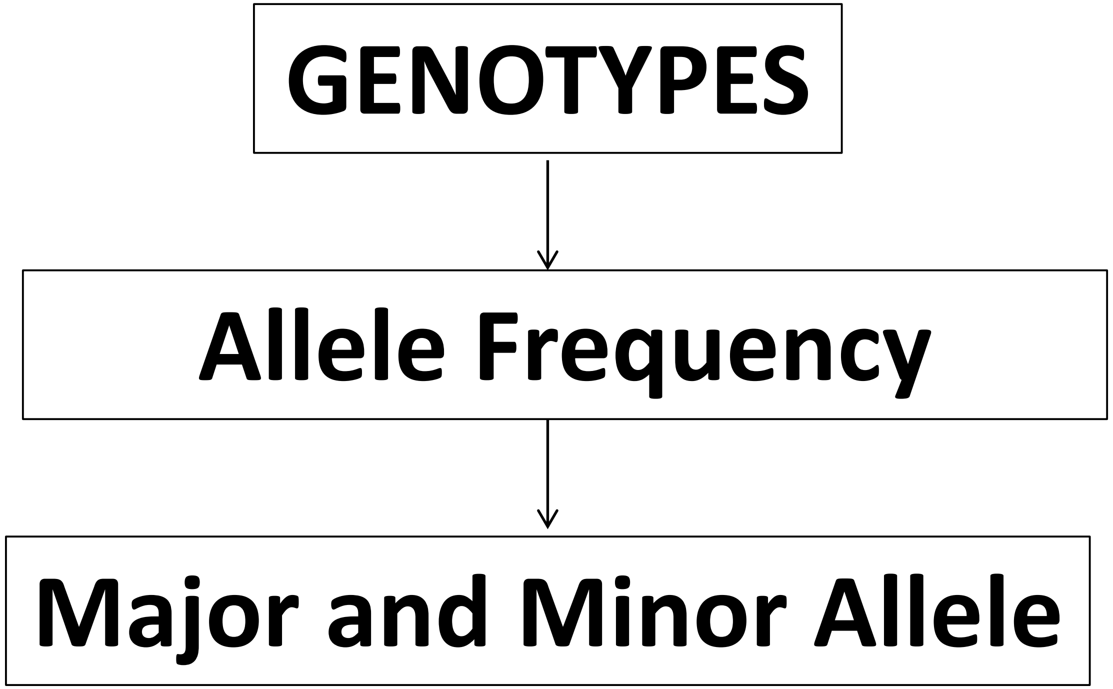
```
  --out HapMap_3_r3_6
  --remove dd_filter.txt

16268 MB RAM detected; reserving 8134 MB for main workspace.
1431211 variants loaded from .bim file.
164 people (79 males, 85 females) loaded from .fam.
112 phenotype values loaded from .fam.
--remove: 163 people remaining.
Using 1 thread (no multithreaded calculations invoked).
Before main variant filters, 112 founders and 51 nonfounders present.
Calculating allele frequencies... done.
Warning: 181 het. haploid genotypes present (see HapMap_3_r3_6.hh ); many
commands treat these as missing.
Total genotyping rate in remaining samples is 0.998078.
1431211 variants and 163 people pass filters and QC.
Among remaining phenotypes, 56 are cases and 56 are controls.  (51 phenotypes
are missing.)
--make-bed to HapMap_3_r3_6.bed + HapMap_3_r3_6.bim + HapMap_3_r3_6.fam ...
done.

C:\Users\archana\Desktop\GBIO2_2020\CLASS 2\SESSION>
```

Allele Frequency

how often an form of a gene shows up in a population over several generations

the number of copies of a particular allele divided by the number of copies of all alleles at the genetic place in a population.





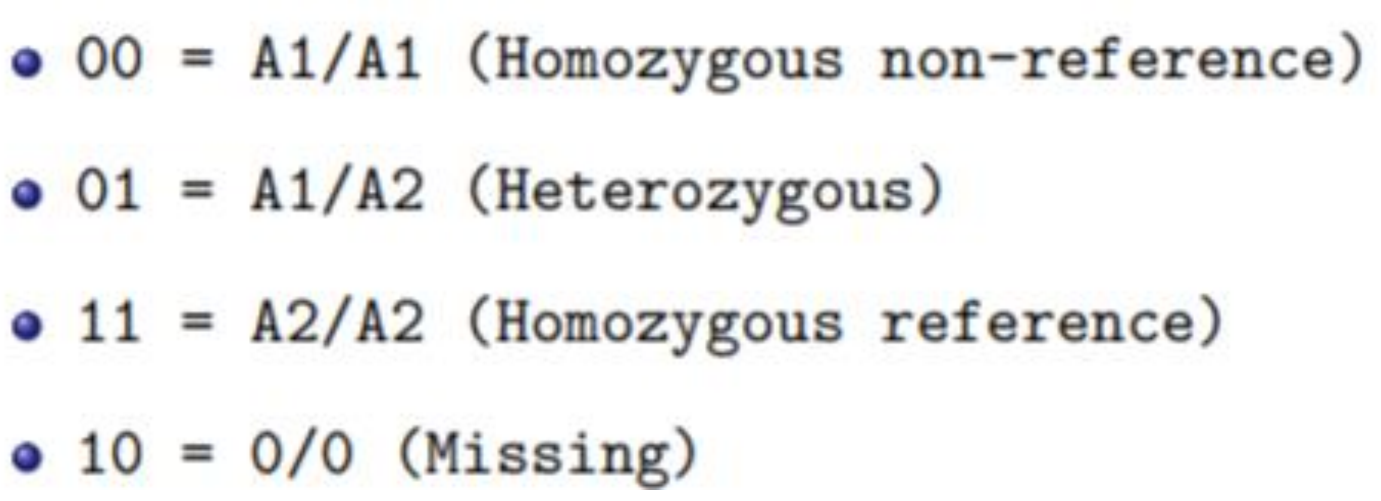
GENOTYPES

Allele Frequency

Major and Minor Allele

Genotypes

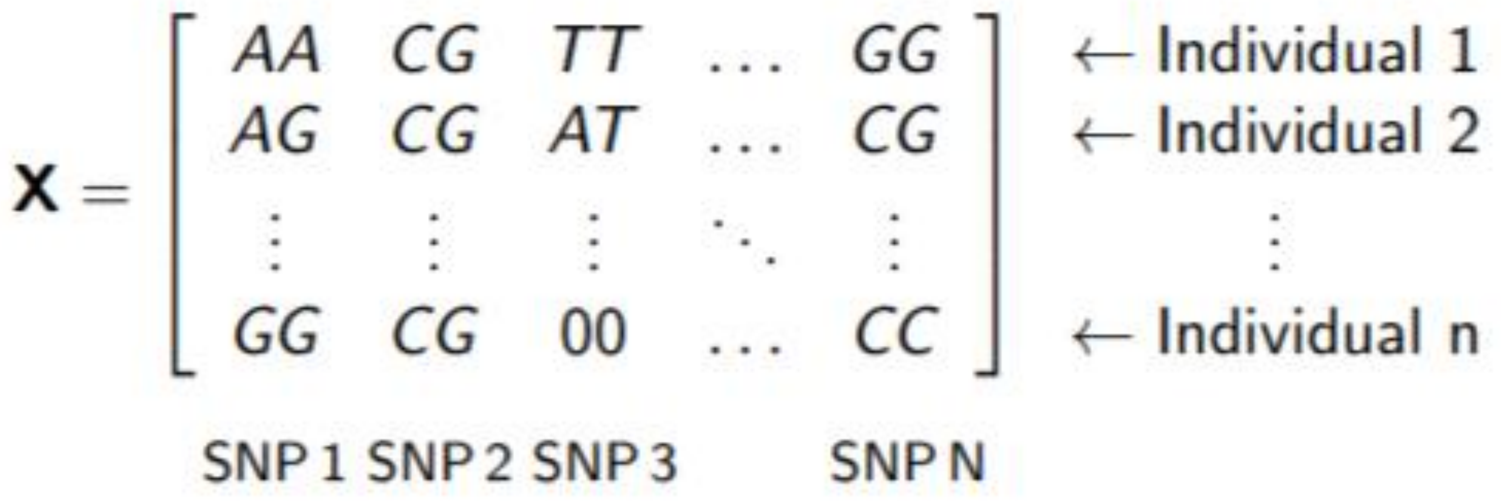
- **PLINK uses the following two-bit coding of genotypes**



- 00 = A1/A1 (Homozygous non-reference)
- 01 = A1/A2 (Heterozygous)
- 11 = A2/A2 (Homozygous reference)
- 10 = 0/0 (Missing)

Genotypes specific SNP matrix

- Suppose we have n individuals genotypes for N SNPs



$$\mathbf{X} = \begin{array}{cl} \begin{bmatrix} AA & CG & TT & \dots & GG \\ AG & CG & AT & \dots & CG \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ GG & CG & 00 & \dots & CC \end{bmatrix} & \begin{array}{l} \leftarrow \text{Individual 1} \\ \leftarrow \text{Individual 2} \\ \qquad\vdots \\ \leftarrow \text{Individual n} \end{array} \\ \begin{array}{ccccc} \text{SNP 1} & \text{SNP 2} & \text{SNP 3} & \quad & \text{SNP N} \end{array} & \end{array}$$

- The genotypes correspond to a matrix X of size n x p

Allele Frequency

- To generate a list of minor allele frequencies (MAF) for each SNP, based on all founders in the sample:

- This will create a file: **plink.frq** with five columns:

CHR	Chromosome
SNP	SNP identifier
A1	Allele 1 code (minor allele)
A2	Allele 2 code (major allele)
MAF	Minor allele frequency
NCHROBS	Non-missing allele count

Minor Allele Frequency (MAF)

- Once individuals with too much missing genotype data have been excluded, subsequent analyses can be set to automatically exclude SNPs on the basis of MAF (minor allele frequency).

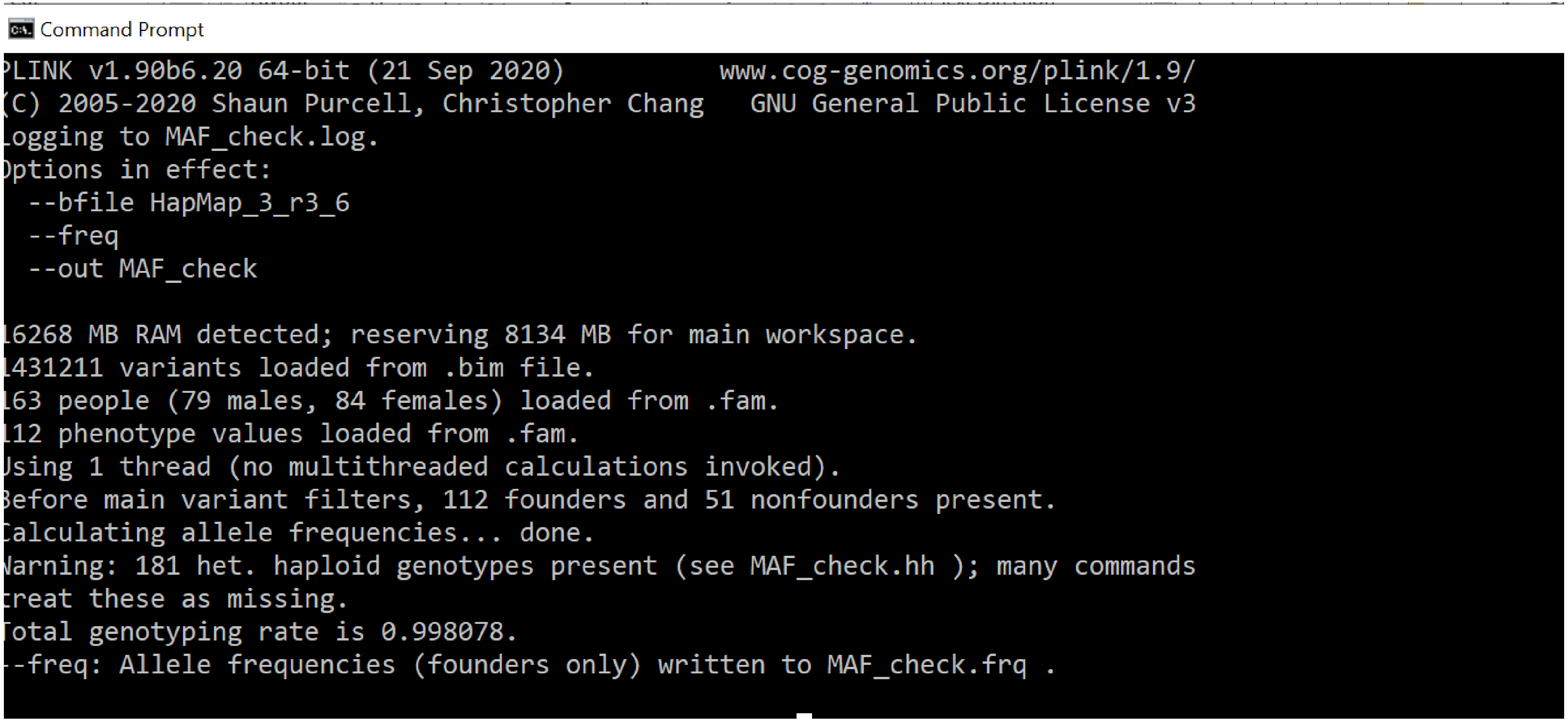
- Include SNPs with MAF >= 0.05.

- The default value is 0.01. This quantity is based only on founders

Minor Allele Frequency (MAF)

- **Minor allele frequency (MAF) is the frequency at which the second most common allele occurs in a given population**

plink --bfile HapMap_3_r3_6 --freq --out MAF_check



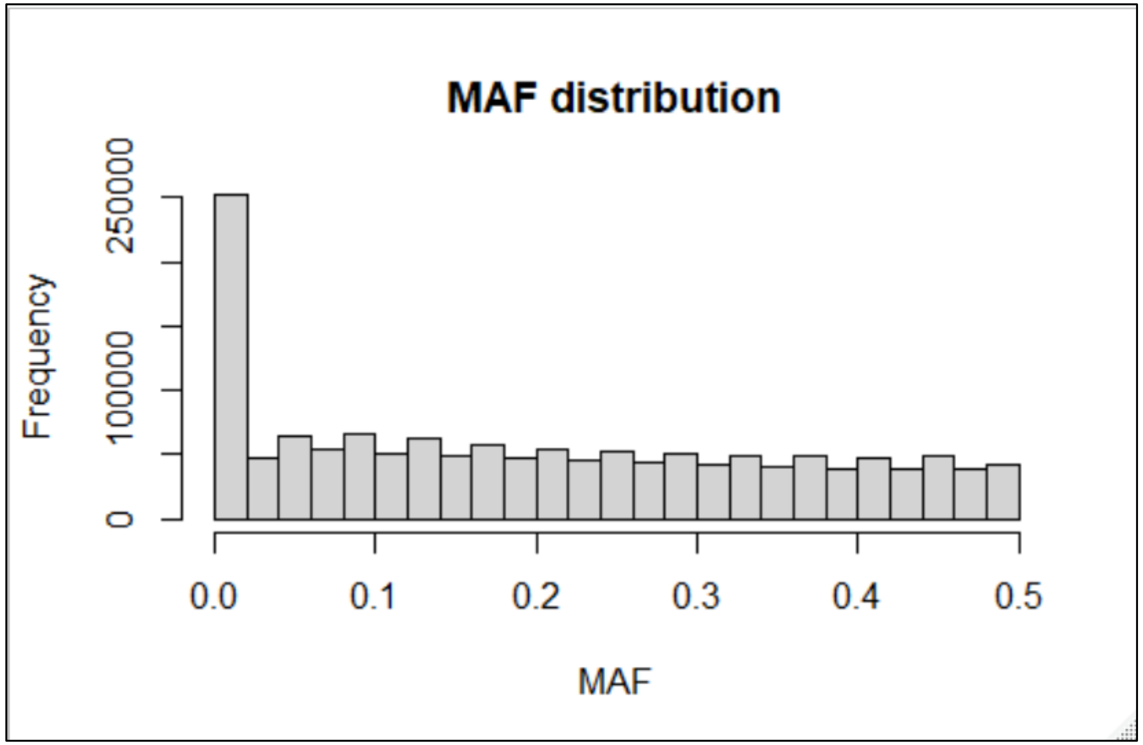
```
PLINK v1.90b6.20 64-bit (21 Sep 2020)          www.cog-genomics.org/plink/1.9/
(C) 2005-2020 Shaun Purcell, Christopher Chang   GNU General Public License v3
Logging to MAF_check.log.
Options in effect:
  --bfile HapMap_3_r3_6
  --freq
  --out MAF_check

16268 MB RAM detected; reserving 8134 MB for main workspace.
1431211 variants loaded from .bim file.
163 people (79 males, 84 females) loaded from .fam.
112 phenotype values loaded from .fam.
Using 1 thread (no multithreaded calculations invoked).
Before main variant filters, 112 founders and 51 nonfounders present.
Calculating allele frequencies... done.
Warning: 181 het. haploid genotypes present (see MAF_check.hh ); many commands
treat these as missing.
Total genotyping rate is 0.998078.
--freq: Allele frequencies (founders only) written to MAF_check.frq .
```

Exercise : Visualize the MAF

- **Read the MAF_check.frq**
- **Draw the histogram plot in R**

Visualize the MAF

```
maf_freq <- read.table("/path/MAF_check.frq", header =TRUE) #change “path” with working directory

hist(maf_freq[,5],main = "MAF distribution", xlab = "MAF")
```



Filtration based on MAF

# Remove SNPs with a low MAF frequency.

```
plink --bfile HapMap_3_r3_6 --maf 0.05 --make-bed --out HapMap_3_r3_7
```

# A conventional MAF threshold for a regular GWAS is between 0.01 or 0.05, depending on sample size.

Count SNPs under MAF < 0.01 ?

Hardy-Weinberg Equilibrium (1)

- To generate a list of genotype counts and Hardy-Weinberg test statistics for each SNP, use the option:

```
--hardy
```

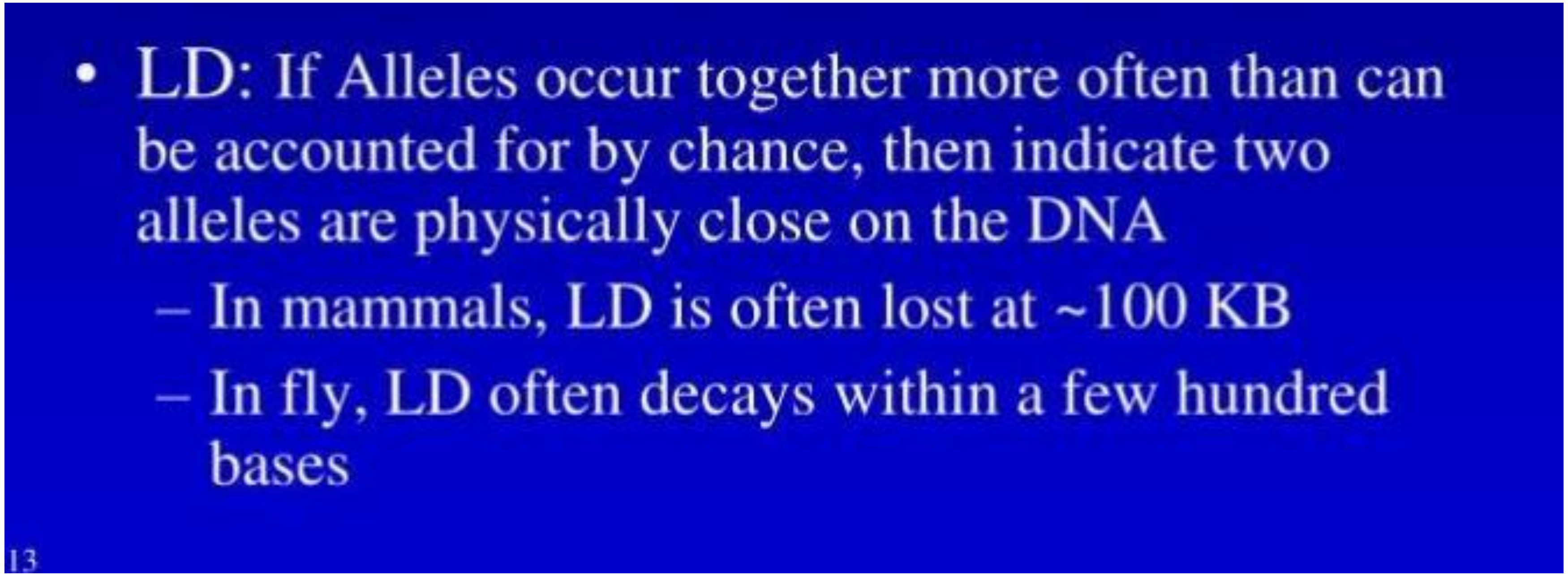
which creates a file: **plink.hwe.** The file has the following format

SNP	SNP identifier
TEST	Code indicating sample
A1	Minor allele code
A2	Major allele code
GENO	Genotype counts:11/12/22
O(HET)	observed hetrozygosity
E(HET)	Expected hetrozygosity
P	H-W p-value

Hardy–Weinberg equilibrium (2)

- **Selecting SNPs with HWE p-value below 0.00001**

```
plink --bfile HapMap_3_r3_7 --hwe 1e-6 --make-bed --out HapMap_hwe_filter_step1
```



- LD: If Alleles occur together more often than can be accounted for by chance, then indicate two alleles are physically close on the DNA
  - In mammals, LD is often lost at ~100 KB
  - In fly, LD often decays within a few hundred bases

13

- **Linkage disequilibrium (LD): This is a measure of non- random association between alleles at different loci at the same chromosome in a given population.**
- **SNPs are in LD when the frequency of association of their alleles is higher than expected under random assortment.**
- **LD concerns patterns of correlations between SNPs.**

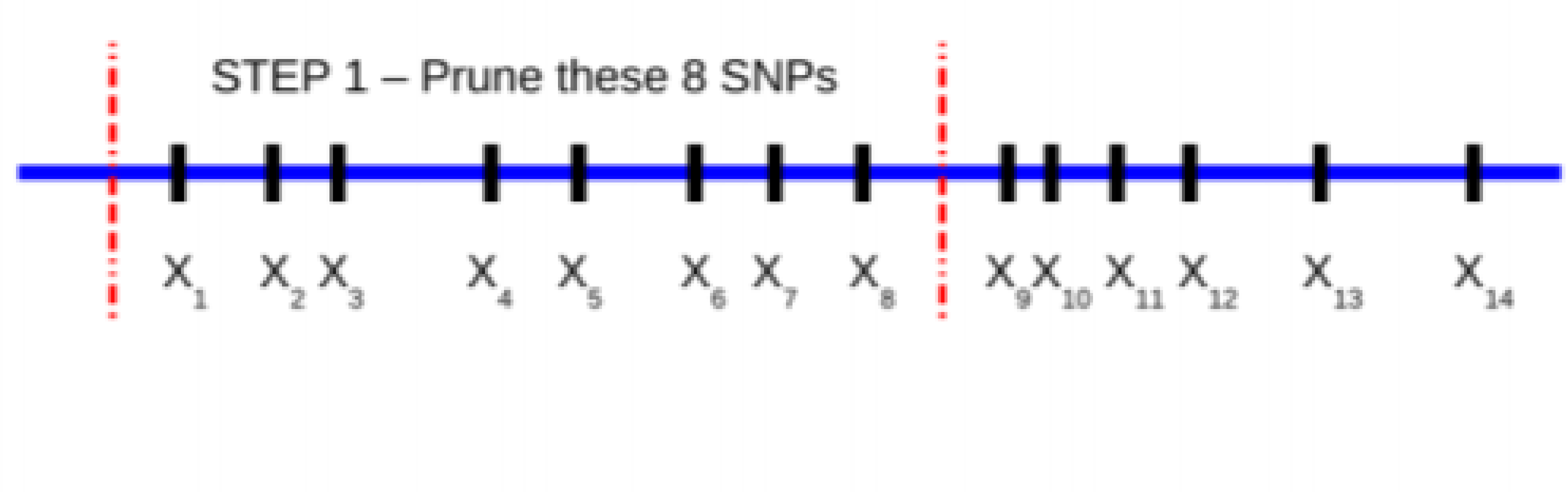
Linkage disequilibrium pruning (1)

▪**Sometimes it is useful to generate a pruned subset of SNPs that are in approximate linkage equilibrium with each other. This can be achieved via two commands:**

--indep which prunes based on the variance inflation factor (VIF), which recursively removes SNPs within a sliding window;

```
plink --bfile HapMap_3_r3_7 --indep 100 5 2 --make-bed --out HapMap_3_r3_8
```

Linkage disequilibrium pruning (2)


STEP 1 – Prune these 8 SNPs
X_1 X_2 X_3 X_4 X_5 X_6 X_7 X_8 X_9 X_{10} X_{11} X_{12} X_{13} X_{14}

Linkage disequilibrium pruning (3)

▪Each is a simlpe list of SNP IDs; both these files can subsequently be specified as the argument for a -- extract or --exclude command.

▪The parameters for --indep are: window size in SNPs (e.g. 50), the number of SNPs to shift the window at each step (e.g. 5), the VIF threshold. The VIF is 1/(1-R^2) where R^2 is the multiple correlation coefficient for a SNP being regressed on all other SNPs simultaneously.

▪That is, this considers the correlations between SNPs but also between linear combinations of SNPs.

How many snp in LD with window size “150”, “200” ?

clustering

```
plink.exe --bfile HapMap_3_r3_8   --cluster
```

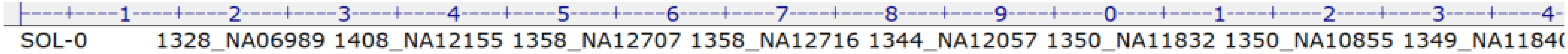
which generates four output files:

plink.cluster0
plink.cluster1
plink.cluster2
plink.cluster3

that contain similar information but in different formats. The
The *.cluster0 file contains some information on the clustering process. This file can be safely ignored by most users.
The *.cluster1 file contains information on the final solution, listed by cluster.
The *.cluster2 file contains the same information but listed one line per individual

The *.cluster3 file is in the same format as cluster2 (one line per individual) but contains all solutions (i.e. every step of the clustering from moving from N clusters each of 1 individual (leftmost column after family and individual ID) to 1 cluster (labelled 0) containing all N individuals (the final, rightmost column)

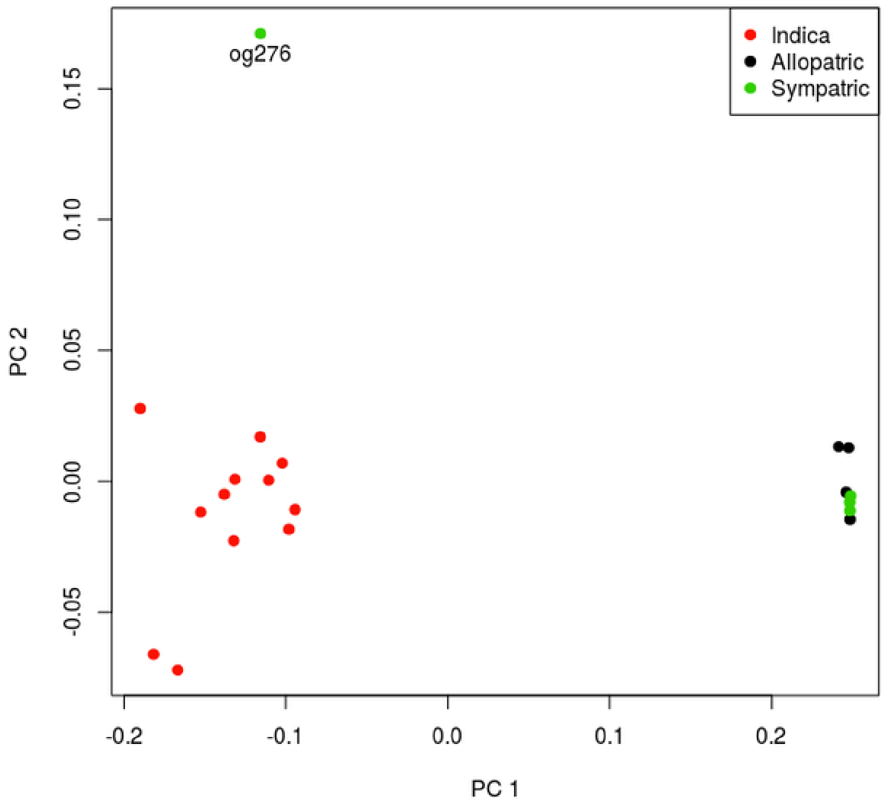
Plink.cluster1



```
SOL-0       1328_NA06989 1408_NA12155 1358_NA12707 1358_NA12716 1344_NA12057 1350_NA11832 1350_NA10855 1349_NA1184(
```

There is only one cluster.

What if we have more than one cluster?



Three cluster

We will perform this analysis in other R package

Association Analysis

- Case/control
- Multiple-testing correction

Basic case/control association test

To perform a standard case/control association analysis, use the option:

```
plink.exe --bfile HapMap_3_r3_8   --assoc --noweb
```

which generates a file

plink.assoc

which contains the fields:

CHR	Chromosome
SNP	SNP ID
BP	Physical position (base-pair)
A1	Minor allele name (based on whole sample)
F_A	Frequency of this allele in cases
F_U	Frequency of this allele in controls
A2	Major allele name
CHISQ	Basic allelic test chi-square (1df)
P	Asymptotic p-value for this test
OR	Estimated odds ratio (for A1, i.e. A2 is reference)

Adjustment for multiple testing

To generate a file of adjusted significance values that correct for all tests performed and other metrics, use the option:

```
plink.exe --bfile HapMap_3_r3_8 --assoc --adjust
```

which generates the file

plink.adjust

which contains the fields

CHR	Chromosome number
SNP	SNP identifer
UNADJ	Unadjusted p-value
GC	Genomic-control corrected p-values
BONF	Bonferroni single-step adjusted p-values
HOLM	Holm (1979) step-down adjusted p-values
SIDAK_SS	Sidak single-step adjusted p-values
SIDAK_SD	Sidak step-down adjusted p-values
FDR_BH	Benjamini & Hochberg (1995) step-up FDR control
FDR_BY	Benjamini & Yekutieli (2001) step-up FDR control

This file is sorted by significance value rather than genomic location, the most significant results being at the top.

Let us visualize GWAS result

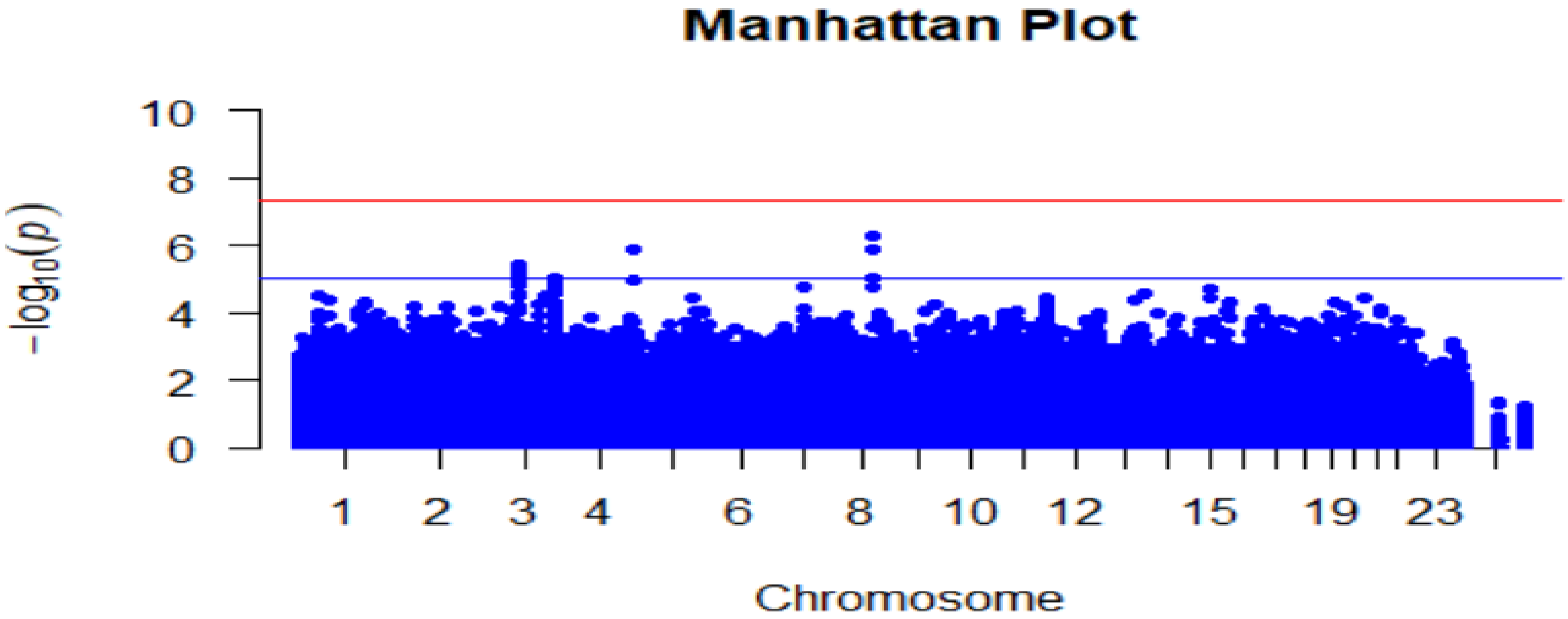
LETS INSTALL R Pakcage

1. Open R window
2. Install.packages("qqman")
3. Load in library

library("qqman")

```
gwas <- data.frame(read.table(file="plink.assoc",header=TRUE))
```

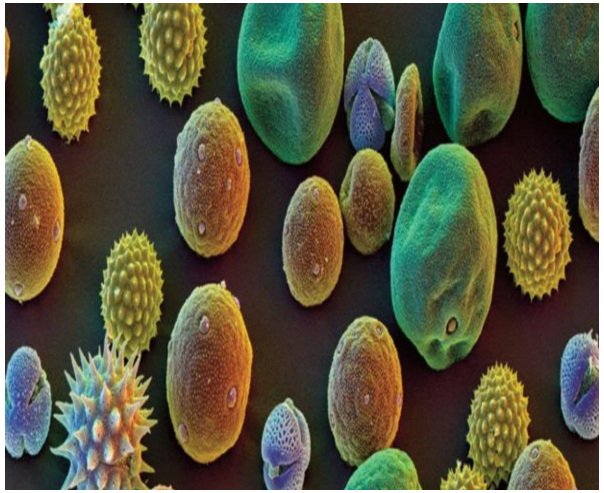
```
manhattan(gwas, main = "Manhattan Plot", ylim = c(0, 10),col="blue")
```



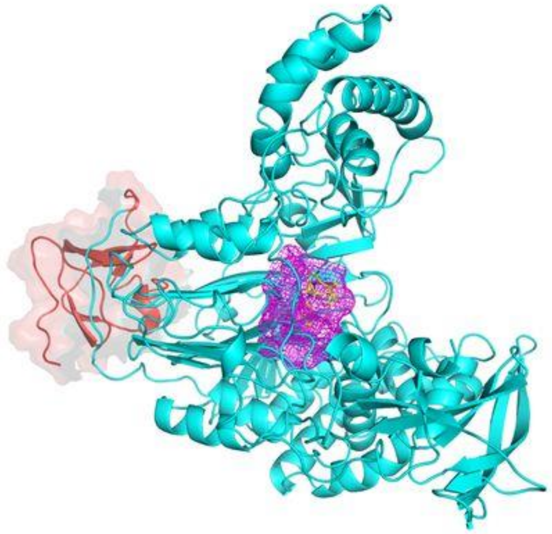
Unit of information in Bioinformatics

What ’’unit of information’’ do we deal within bioinformatics ?

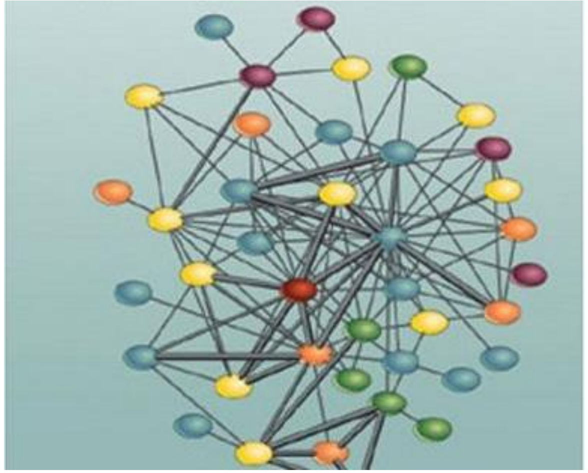
- DNA
- RNA
- Protein

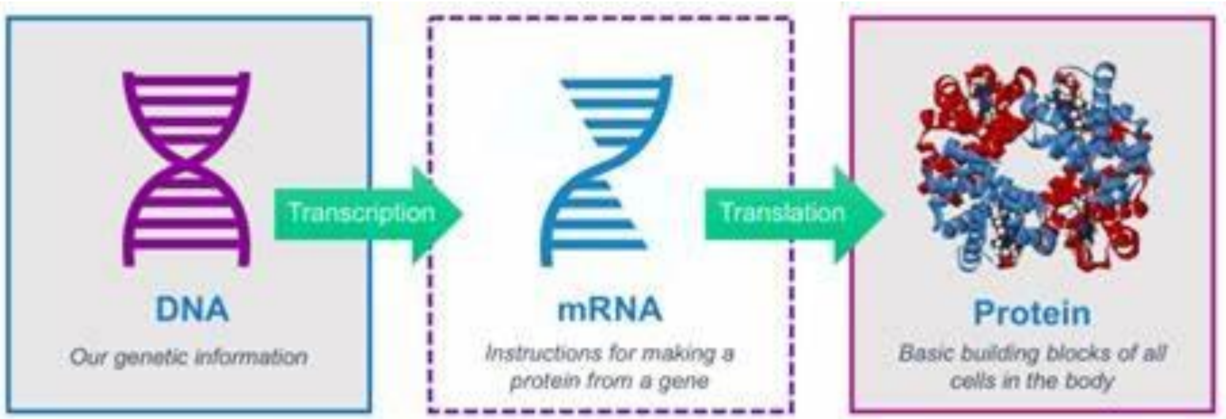


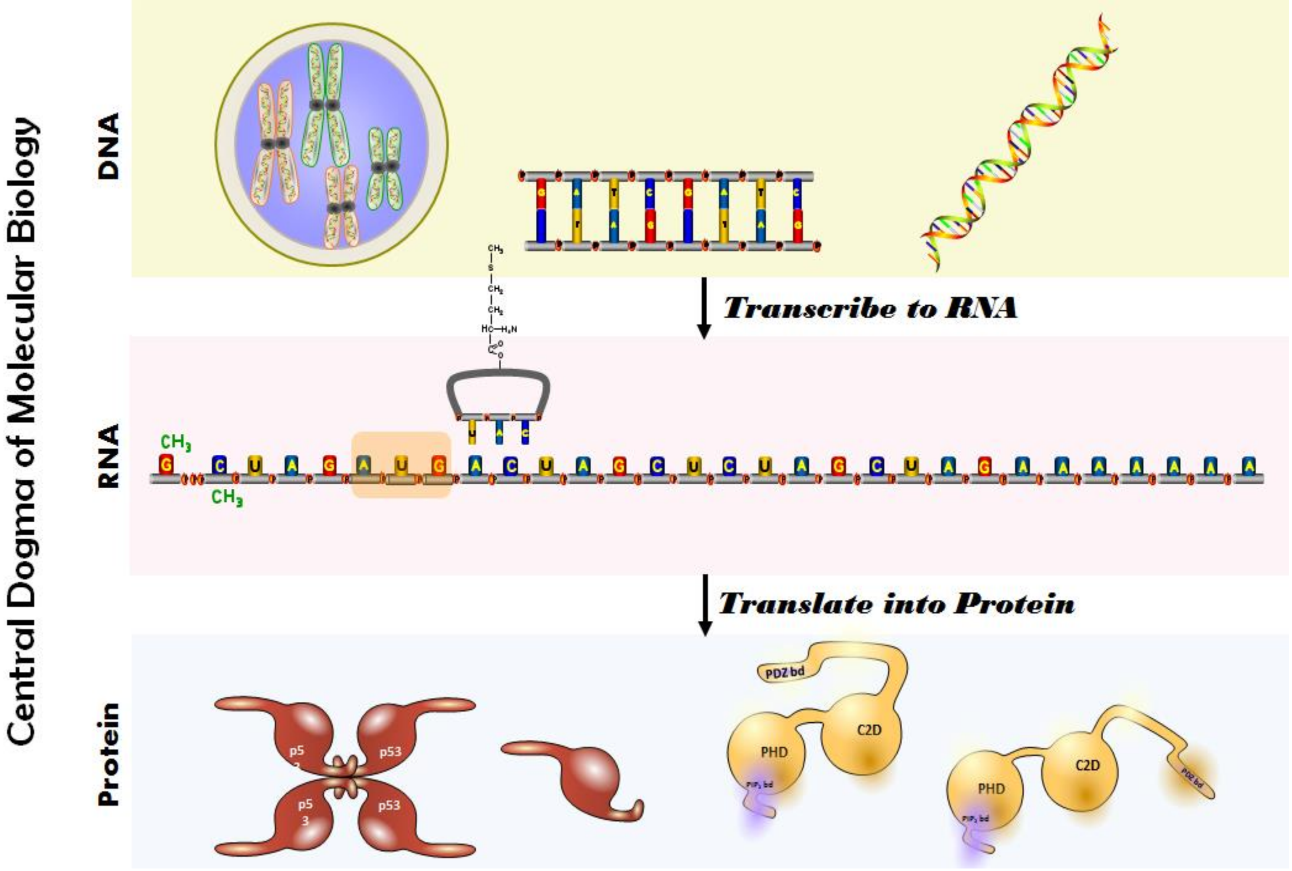
- Sequence
- Structure
- Evolution



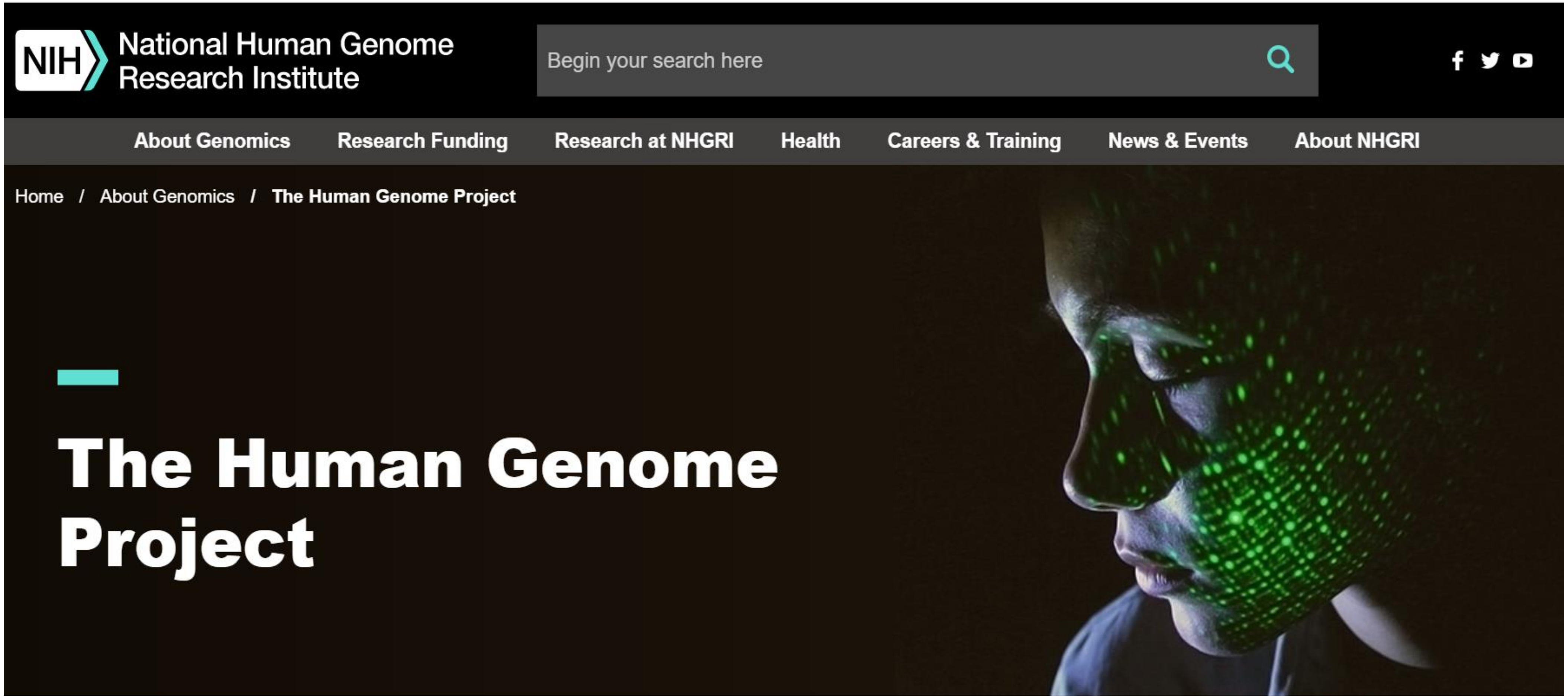
- Pathways
- Interactions
- Mutations




DNA
Our genetic information
Transcription
mRNA
Instructions for making a protein from a gene
Translation
Protein
Basic building blocks of all cells in the body


Central Dogma of Molecular Biology
DNA
Transcribe to RNA
RNA
CH3
CH3
Translate into Protein
Protein
p53
p53
p53
p53
PDZ bd
PHD
C2D
PHD
C2D

https://www.genome.gov/human-genome-project



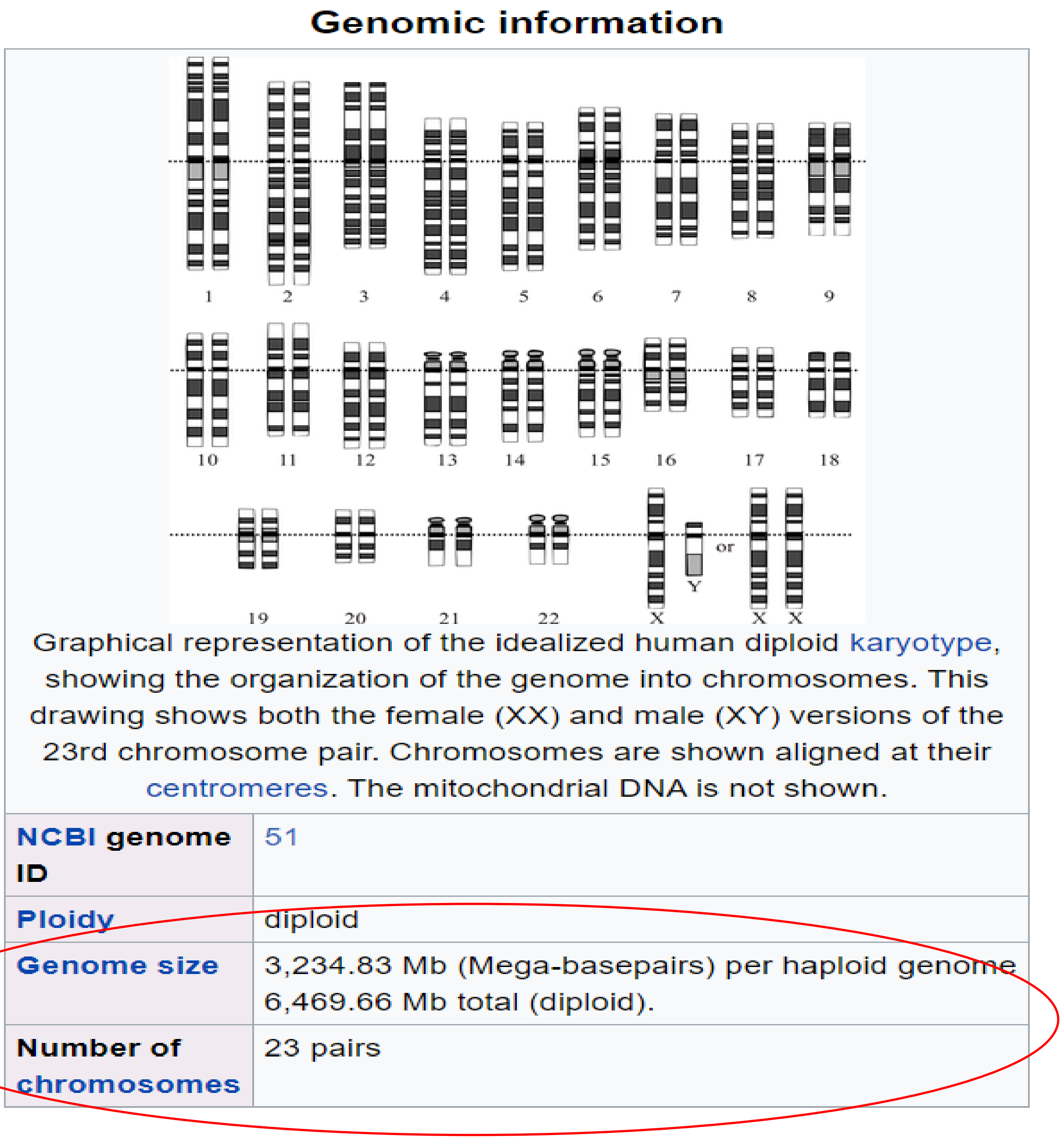
NIH National Human Genome Research Institute

Home / About Genomics / The Human Genome Project

The Human Genome Project

Human Genome- 1990-2003

The first printout of the human genome to be presented as a series of books, displayed at the Wellcome Collection, London



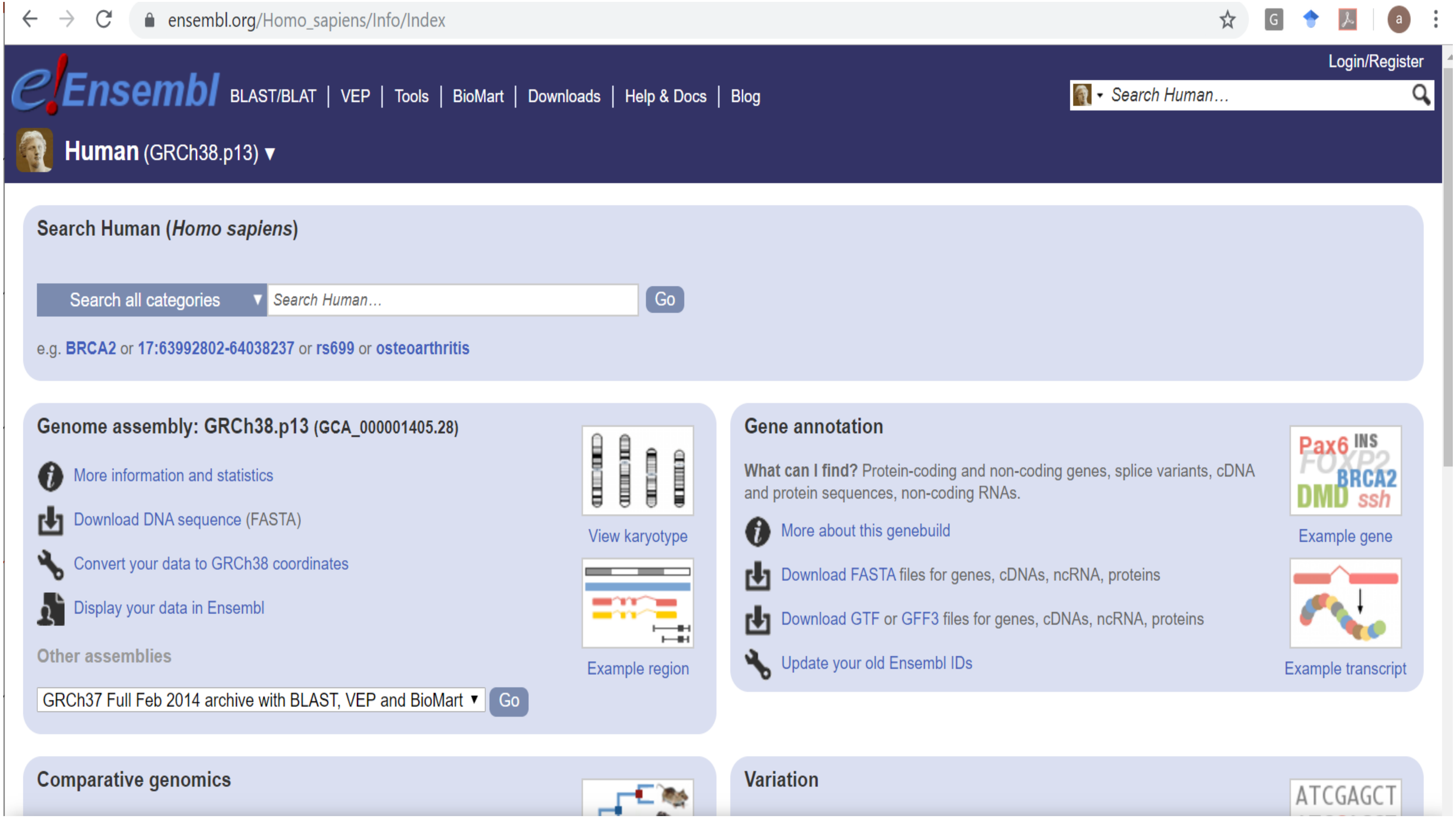
Genomic information

Graphical representation of the idealized human diploid karyotype, showing the organization of the genome into chromosomes. This drawing shows both the female (XX) and male (XY) versions of the 23rd chromosome pair. Chromosomes are shown aligned at their centromeres. The mitochondrial DNA is not shown.

NCBI genome ID	51
Ploidy	diploid
Genome size	3,234.83 Mb (Mega-basepairs) per haploid genome 6,469.66 Mb total (diploid).
Number of chromosomes	23 pairs

More information :

DNA sequence, RNA sequence, Protein sequence



Search Human (*Homo sapiens*)

e.g. BRCA2 or 17:63992802-64038237 or rs699 or osteoarthritis

Genome assembly: GRCh38.p13 (GCA_000001405.28)

- More information and statistics
- Download DNA sequence (FASTA)
- Convert your data to GRCh38 coordinates
- Display your data in Ensembl

Other assemblies

GRCh37 Full Feb 2014 archive with BLAST, VEP and BioMart ▾

Gene annotation

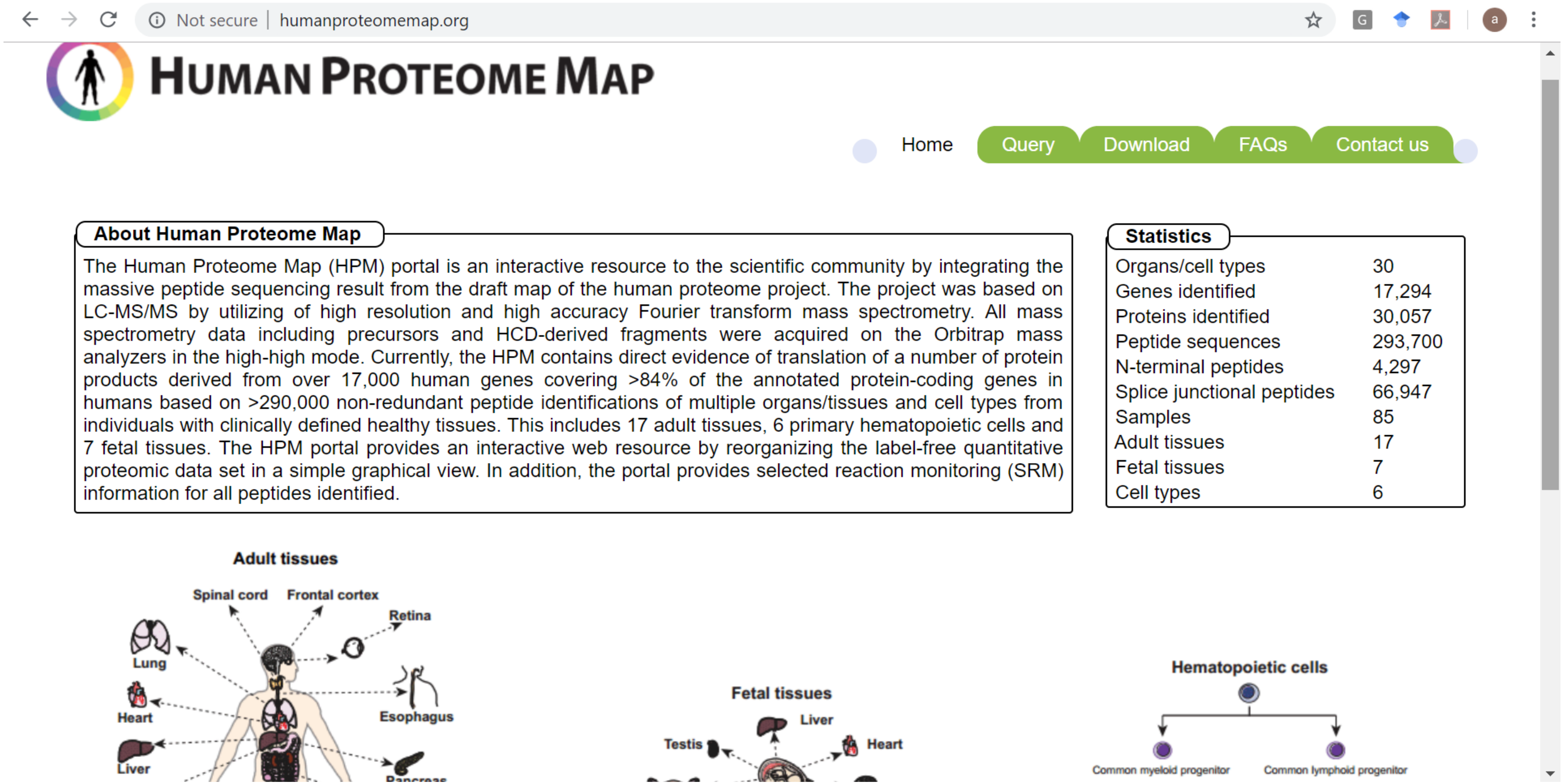
What can I find? Protein-coding and non-coding genes, splice variants, cDNA and protein sequences, non-coding RNAs.

- More about this genebuild
- Download FASTA files for genes, cDNAs, ncRNA, proteins
- Download GTF or GFF3 files for genes, cDNAs, ncRNA, proteins
- Update your old Ensembl IDs

Comparative genomics

Variation

http://humanproteomemap.org/ (Human Proteome Map (HPM))

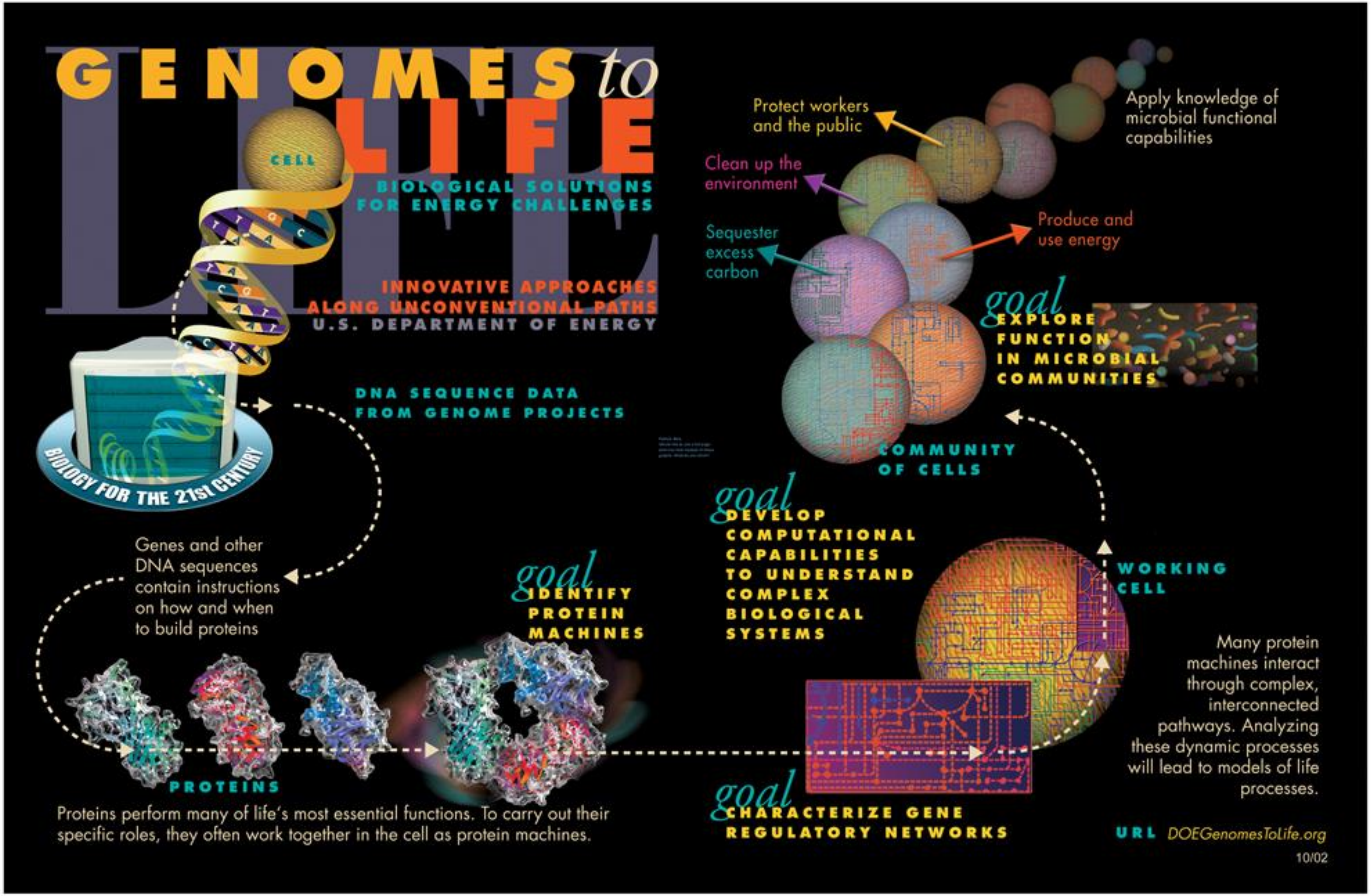


About Human Proteome Map

The Human Proteome Map (HPM) portal is an interactive resource to the scientific community by integrating the massive peptide sequencing result from the draft map of the human proteome project. The project was based on LC-MS/MS by utilizing of high resolution and high accuracy Fourier transform mass spectrometry. All mass spectrometry data including precursors and HCD-derived fragments were acquired on the Orbitrap mass analyzers in the high-high mode. Currently, the HPM contains direct evidence of translation of a number of protein products derived from over 17,000 human genes covering >84% of the annotated protein-coding genes in humans based on >290,000 non-redundant peptide identifications of multiple organs/tissues and cell types from individuals with clinically defined healthy tissues. This includes 17 adult tissues, 6 primary hematopoietic cells and 7 fetal tissues. The HPM portal provides an interactive web resource by reorganizing the label-free quantitative proteomic data set in a simple graphical view. In addition, the portal provides selected reaction monitoring (SRM) information for all peptides identified.

Statistics

Organs/cell types	30
Genes identified	17,294
Proteins identified	30,057
Peptide sequences	293,700
N-terminal peptides	4,297
Splice junctional peptides	66,947
Samples	85
Adult tissues	17
Fetal tissues	7
Cell types	6



GENOMES *to* LIFE

BIOLOGICAL SOLUTIONS FOR ENERGY CHALLENGES

INNOVATIVE APPROACHES ALONG UNCONVENTIONAL PATHS

U.S. DEPARTMENT OF ENERGY

DNA SEQUENCE DATA FROM GENOME PROJECTS

Genes and other DNA sequences contain instructions on how and when to build proteins

PROTEINS

Proteins perform many of life's most essential functions. To carry out their specific roles, they often work together in the cell as protein machines.

goal IDENTIFY PROTEIN MACHINES

goal CHARACTERIZE GENE REGULATORY NETWORKS

goal DEVELOP COMPUTATIONAL CAPABILITIES TO UNDERSTAND COMPLEX BIOLOGICAL SYSTEMS

WORKING CELL

Many protein machines interact through complex, interconnected pathways. Analyzing these dynamic processes will lead to models of life processes.

COMMUNITY OF CELLS

goal EXPLORE FUNCTION IN MICROBIAL COMMUNITIES

Apply knowledge of microbial functional capabilities

Protect workers and the public

Clean up the environment

Sequester excess carbon

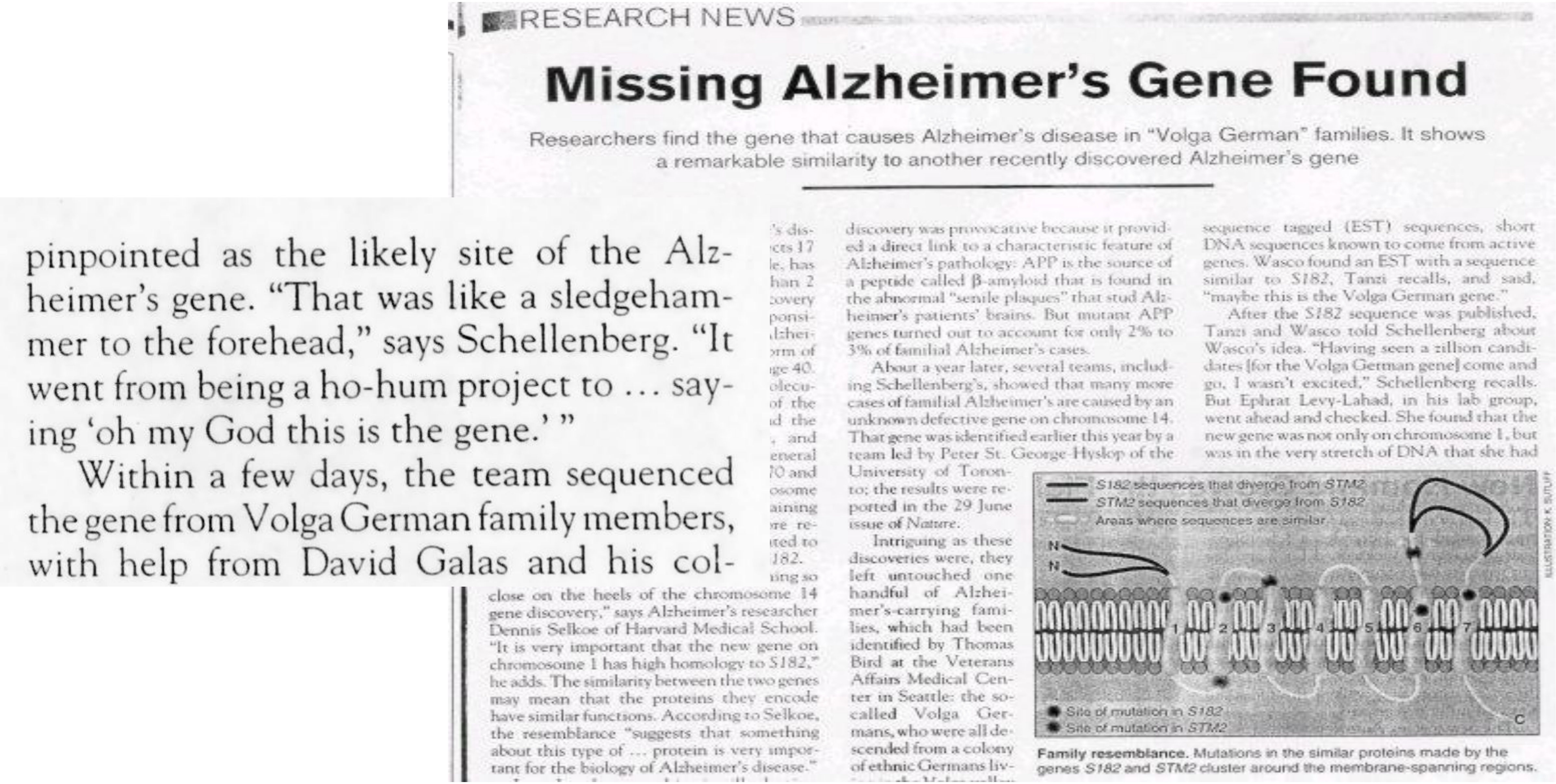
Produce and use energy

[illegible]

URL *DOEGenomesToLife.org*

10/02

Bioinformatics Significance



pinpointed as the likely site of the Alzheimer's gene. "That was like a sledgehammer to the forehead," says Schellenberg. "It went from being a ho-hum project to … saying 'oh my God this is the gene.' "

Within a few days, the team sequenced the gene from Volga German family members, with help from David Galas and his col-

RESEARCH NEWS

Missing Alzheimer's Gene Found

Researchers find the gene that causes Alzheimer's disease in "Volga German" families. It shows a remarkable similarity to another recently discovered Alzheimer's gene

's dis-
cts 17
le, has
han 2
covery
ponsi-
lzhei-
orm of
ge 40.
olecu-
of the
d the
, and
eneral
70 and
osome
aining
re re-
ted to
182.
ing so

close on the heels of the chromosome 14 gene discovery," says Alzheimer's researcher Dennis Selkoe of Harvard Medical School. "It is very important that the new gene on chromosome 1 has high homology to *S182*," he adds. The similarity between the two genes may mean that the proteins they encode have similar functions. According to Selkoe, the resemblance "suggests that something about this type of … protein is very important for the biology of Alzheimer's disease."

discovery was provocative because it provided a direct link to a characteristic feature of Alzheimer's pathology: APP is the source of a peptide called β-amyloid that is found in the abnormal "senile plaques" that stud Alzheimer's patients' brains. But mutant APP genes turned out to account for only 2% to 3% of familial Alzheimer's cases.

About a year later, several teams, including Schellenberg's, showed that many more cases of familial Alzheimer's are caused by an unknown defective gene on chromosome 14. That gene was identified earlier this year by a team led by Peter St. George-Hyslop of the University of Toronto; the results were reported in the 29 June issue of *Nature*.

Intriguing as these discoveries were, they left untouched one handful of Alzheimer's-carrying families, which had been identified by Thomas Bird at the Veterans Affairs Medical Center in Seattle: the so-called Volga Germans, who were all descended from a colony of ethnic Germans liv-

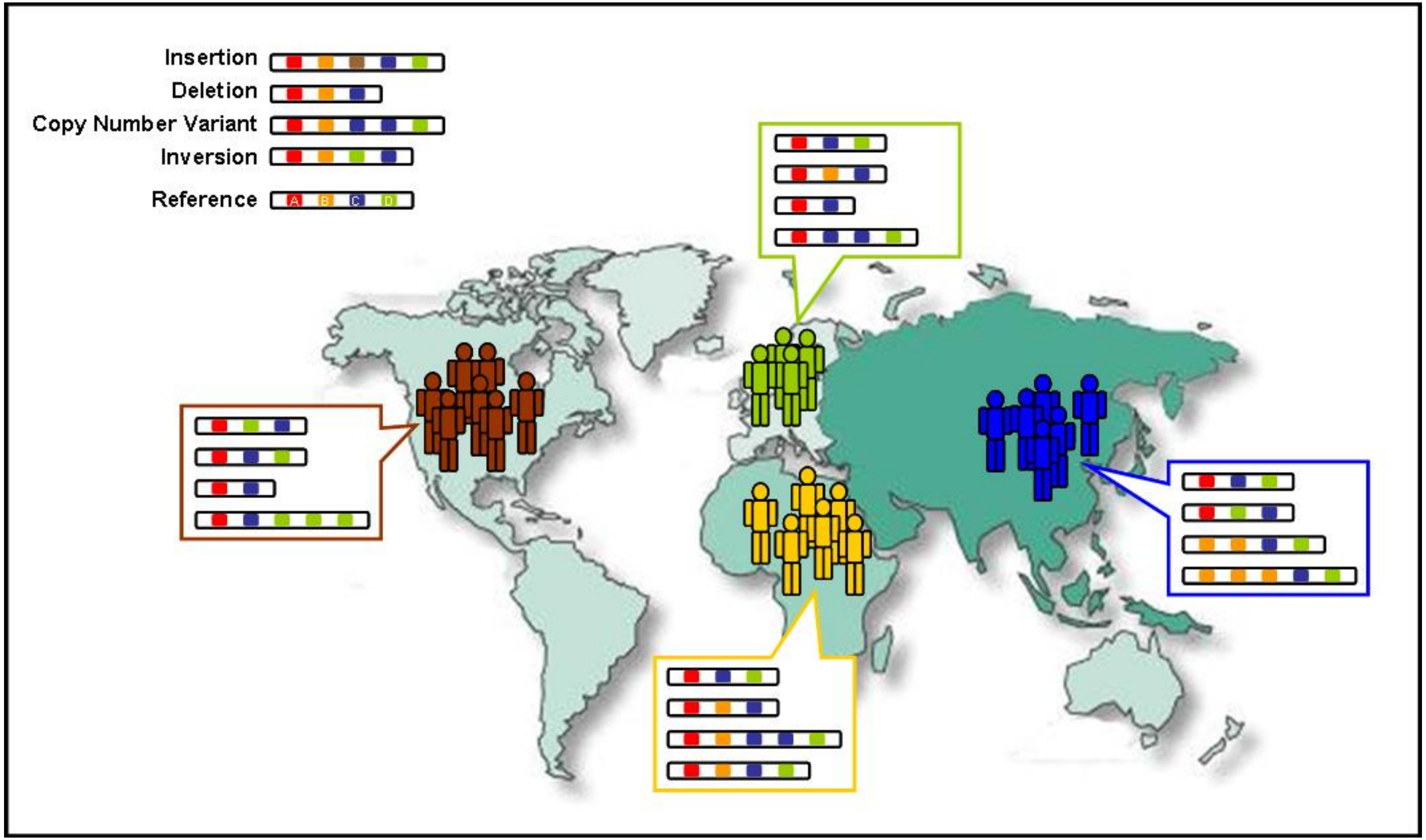
sequence tagged (EST) sequences, short DNA sequences known to come from active genes. Wasco found an EST with a sequence similar to *S182*, Tanzi recalls, and said, "maybe this is the Volga German gene."

After the *S182* sequence was published, Tanzi and Wasco told Schellenberg about Wasco's idea. "Having seen a zillion candidates [for the Volga German gene] come and go, I wasn't excited," Schellenberg recalls. But Ephrat Levy-Lahad, in his lab group, went ahead and checked. She found that the new gene was not only on chromosome 1, but was in the very stretch of DNA that she had

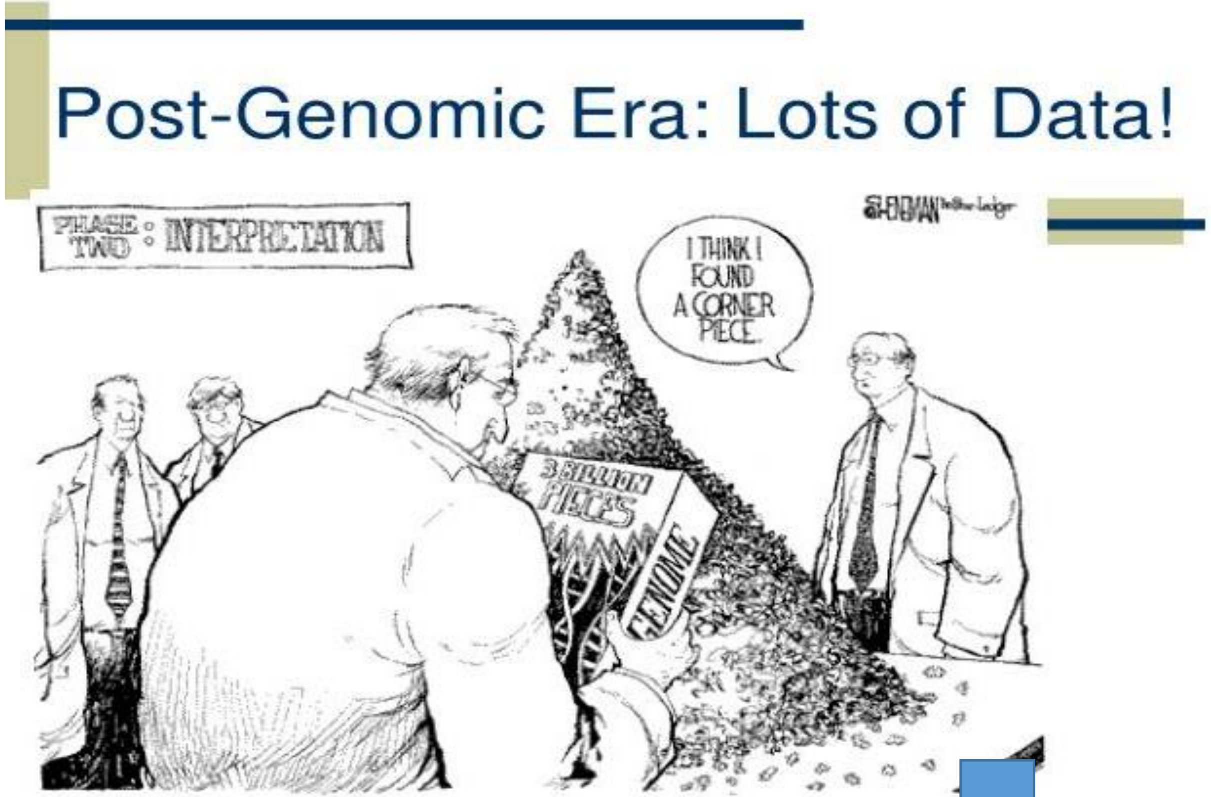
S182 sequences that diverge from *STM2*
STM2 sequences that diverge from *S182*
Areas where sequences are similar
N
N
1
2
3
4
5
6
7
C
● Site of mutation in *S182*
● Site of mutation in *STM2*
ILLUSTRATION: K. SUTLIFF

Family resemblance. Mutations in the similar proteins made by the genes *S182* and *STM2* cluster around the membrane-spanning regions.

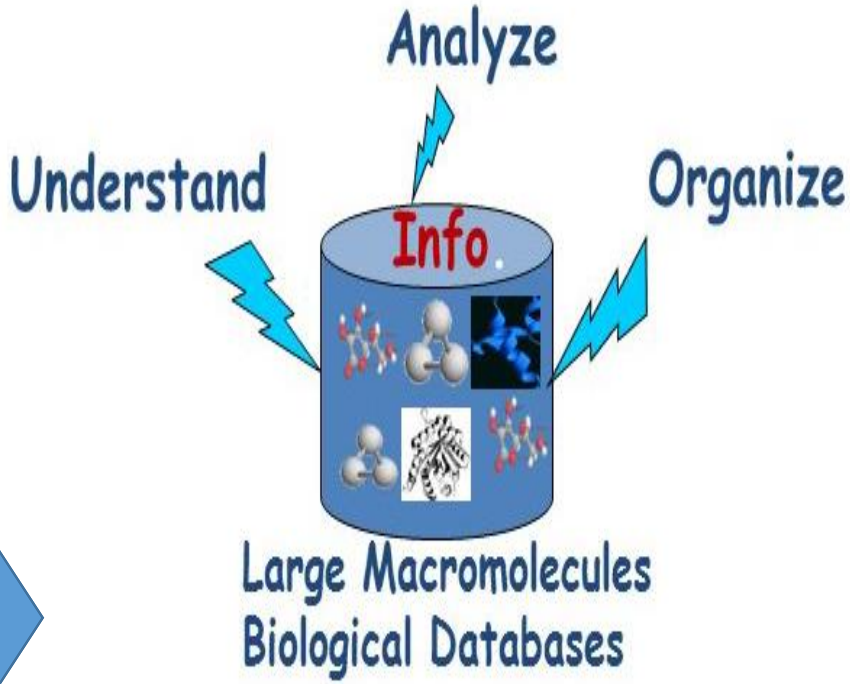
Changes in the number and order of genes (A-D) create genetic diversity within and between populations.



Why do we need DATABASES ?

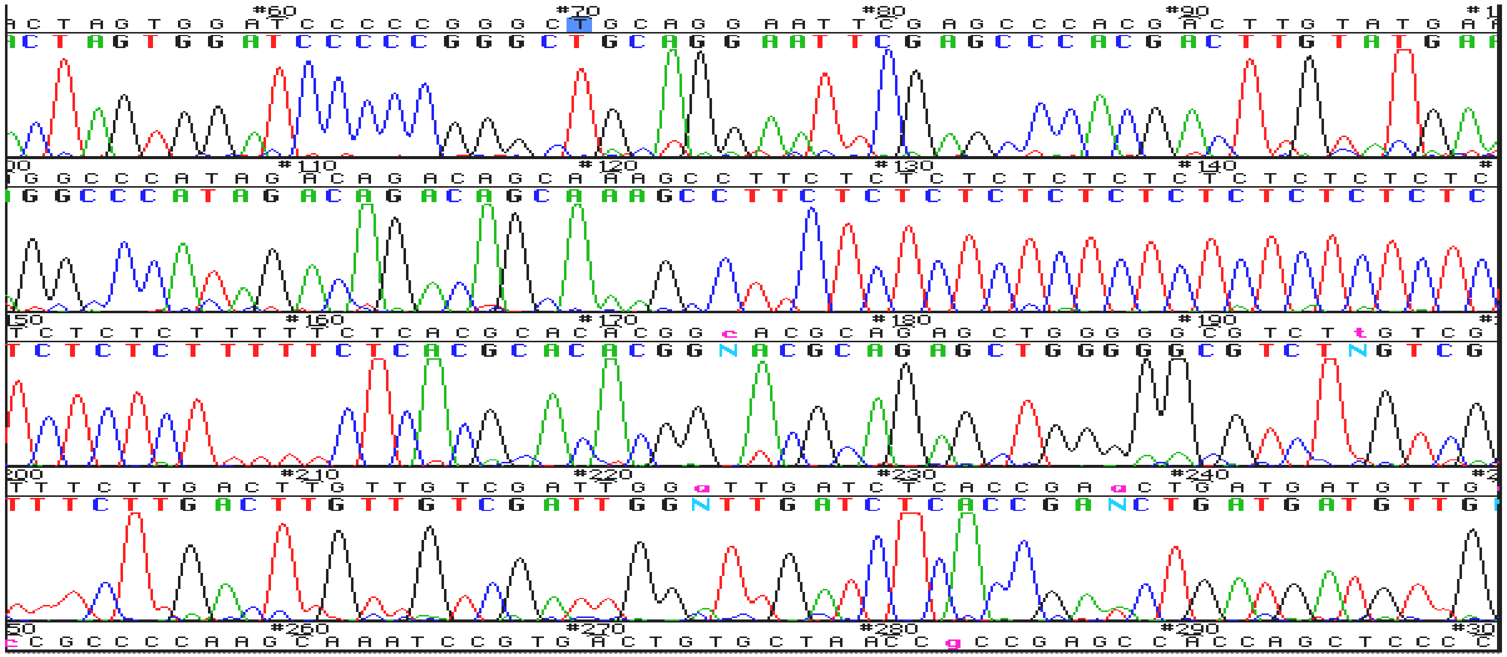


Post-Genomic Era: Lots of Data!



Large Macromolecules
Biological Databases

Genome sequencing generates lots of data

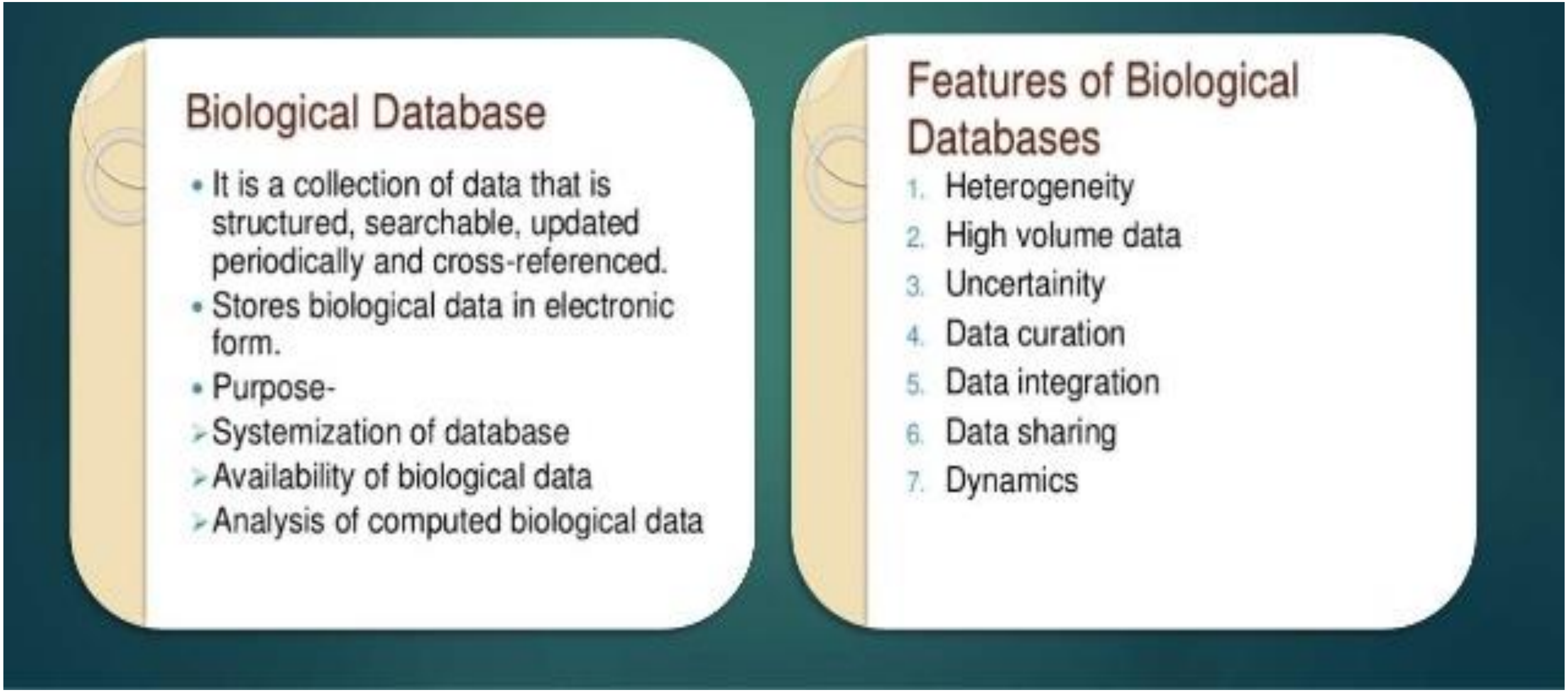

#60 #70 #80 #90
ACTAGTGGATCCCCCGGGCTGCAGGAATTCGAGCCCACGACTTGTATGAA
#110 #120 #130 #140
GGCCCATAGACAGACAGCAAAGCCTTCTCTCTCTCTCTCTCTCTCTC
150 #160 #170 #180 #190
TCTCTCTTTTTCTCACGCACACGGcACGCAGAGCTGGGGGCGTCTtGTCG
200 #210 #220 #230 #240
TTTCTTGACTTGTTGTCGATTGGaTTGATCTCACCGAaCTGATGATGTTG
50 #260 #270 #280 #290 #30
CGCCCCCAAGCAAATCCGTGACTGTGCTAACCgCCGAGCCACCAGCTCCCC

DATABASES

A database is a collection of data in an organized manner, which is accessible in various ways.

A database is an organized collection of data.

What are Biological Databases??



Biological Database

- It is a collection of data that is structured, searchable, updated periodically and cross-referenced.
- Stores biological data in electronic form.
- Purpose-
- ➢ Systemization of database
- ➢ Availability of biological data
- ➢ Analysis of computed biological data

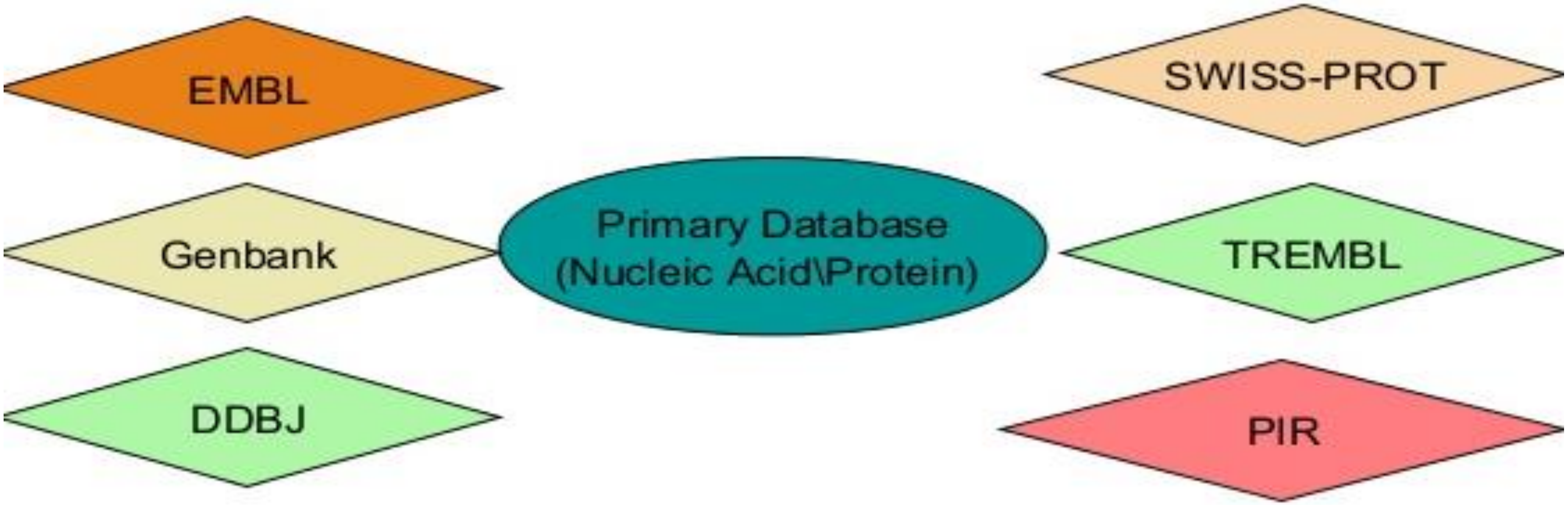
Features of Biological Databases

1. Heterogeneity
2. High volume data
3. Uncertainity
4. Data curation
5. Data integration
6. Data sharing
7. Dynamics

Types of Biological Databases??

There are many different types of database but for routine sequence analysis, the following are initially the most important.

- ➢ Primary databases
- ➢ Secondary databases
- ➢ Composite databases



Primary Database (Nucleic Acid\Protein)

EMBL

Genbank

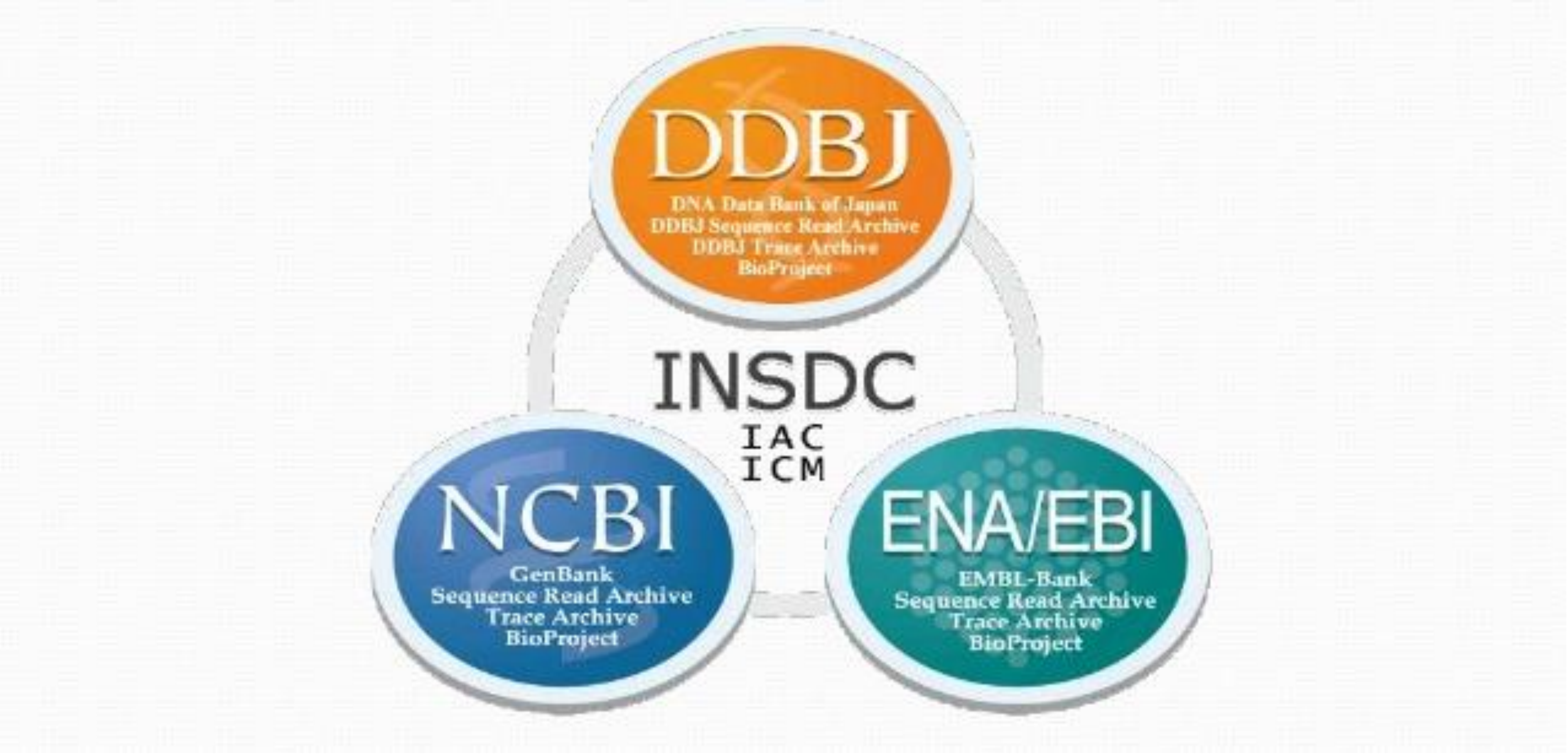
DDBJ

SWISS-PROT

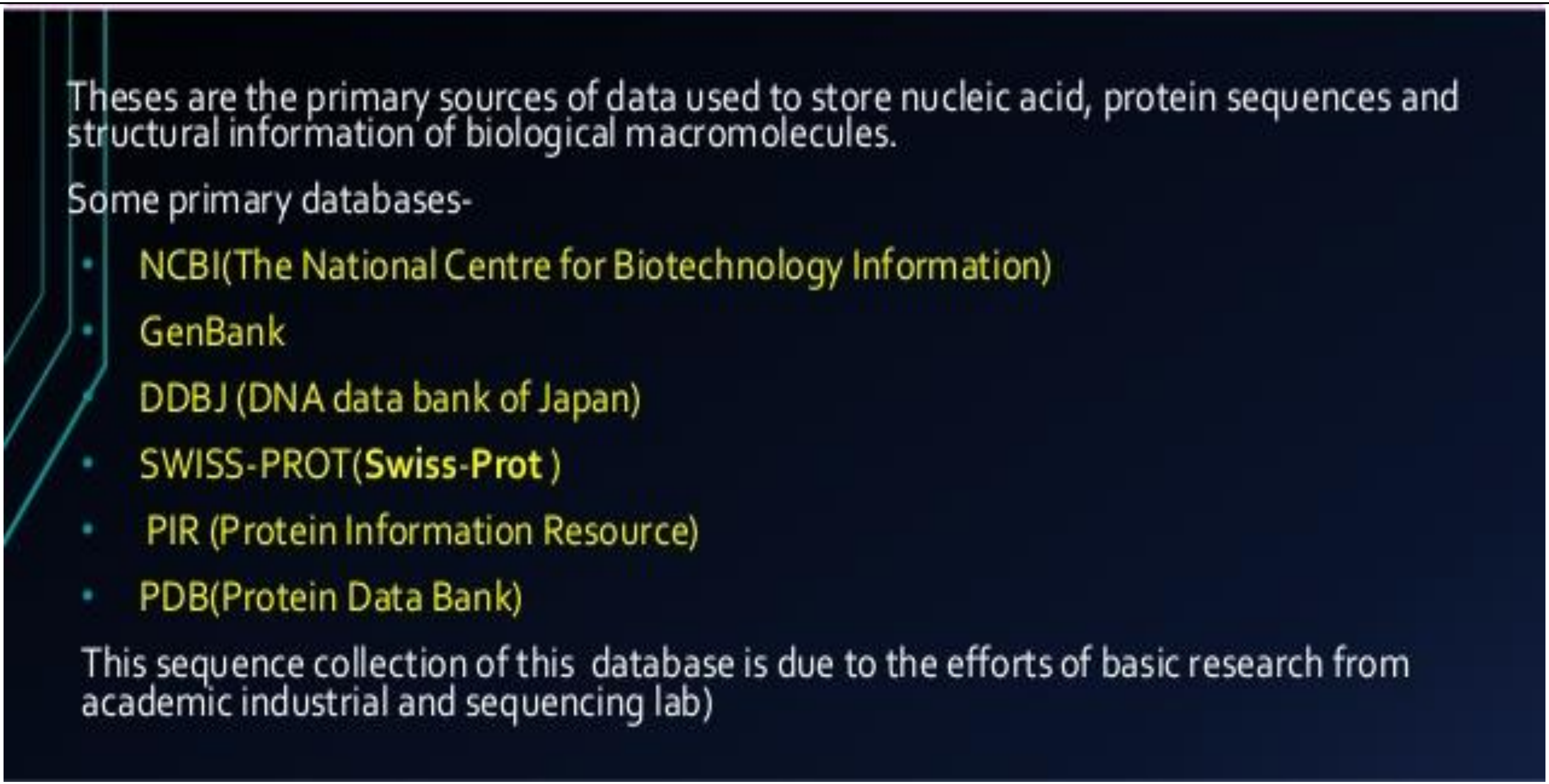
TREMBL

PIR

Interconnections between Databases



Primary Databases



Theses are the primary sources of data used to store nucleic acid, protein sequences and structural information of biological macromolecules.

Some primary databases-

- NCBI(The National Centre for Biotechnology Information)
- GenBank
- DDBJ (DNA data bank of Japan)
- SWISS-PROT(**Swiss-Prot**)
- PIR (Protein Information Resource)
- PDB(Protein Data Bank)

This sequence collection of this database is due to the efforts of basic research from academic industrial and sequencing lab)

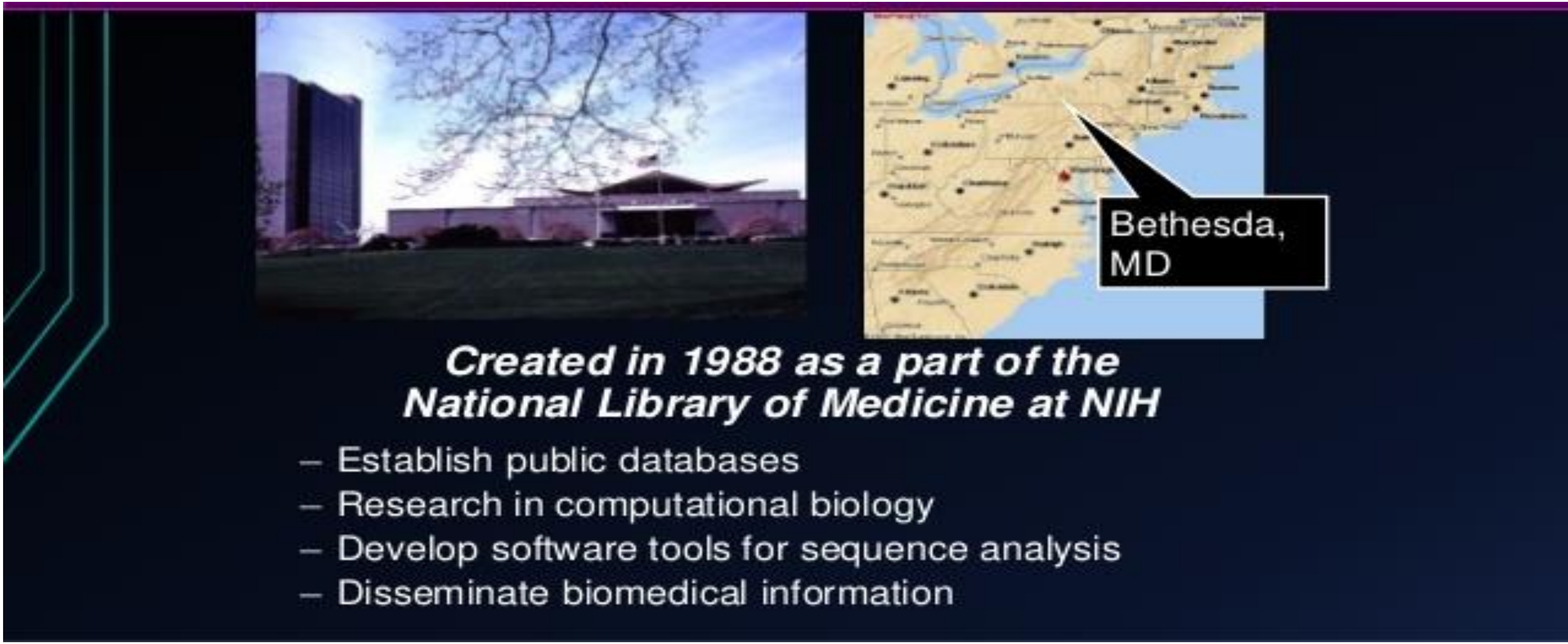
Classification : Primary Databases

- ✓ Sequence Information
  - ✓ DNA: EMBL, Genbank, DDBJ
  - ✓ Protein: SwissProt, TREMBL, PIR, OWL

- ✓ Genome Information
  - ✓ GDB, MGD, ACeDB

- ✓ Structure Information
  - ✓ PDB, NDB, CCDB/CSD

The National Center for Biotechnology Information

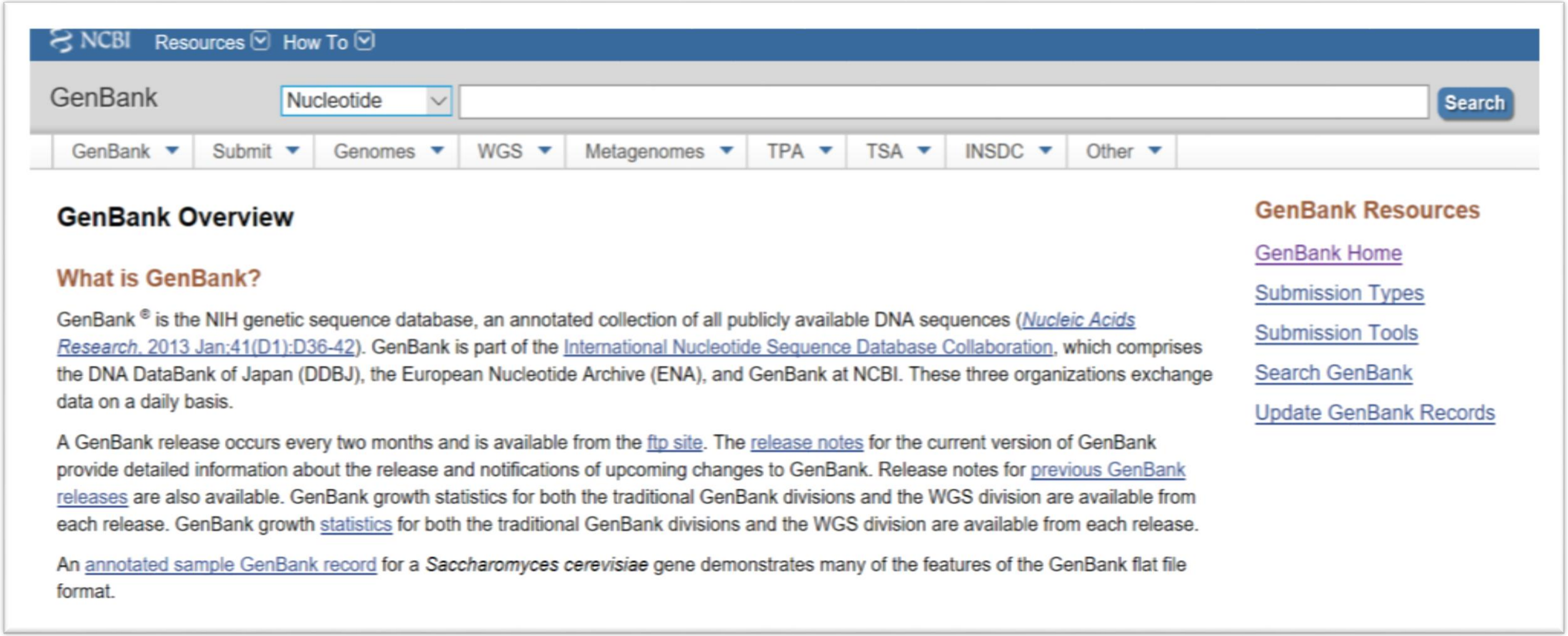


Created in 1988 as a part of the National Library of Medicine at NIH

- Establish public databases
- Research in computational biology
- Develop software tools for sequence analysis
- Disseminate biomedical information

Primary Databases - GenBank

- ✓ **Database from NCBI, includes sequences from publicly available resources**



GenBank Overview

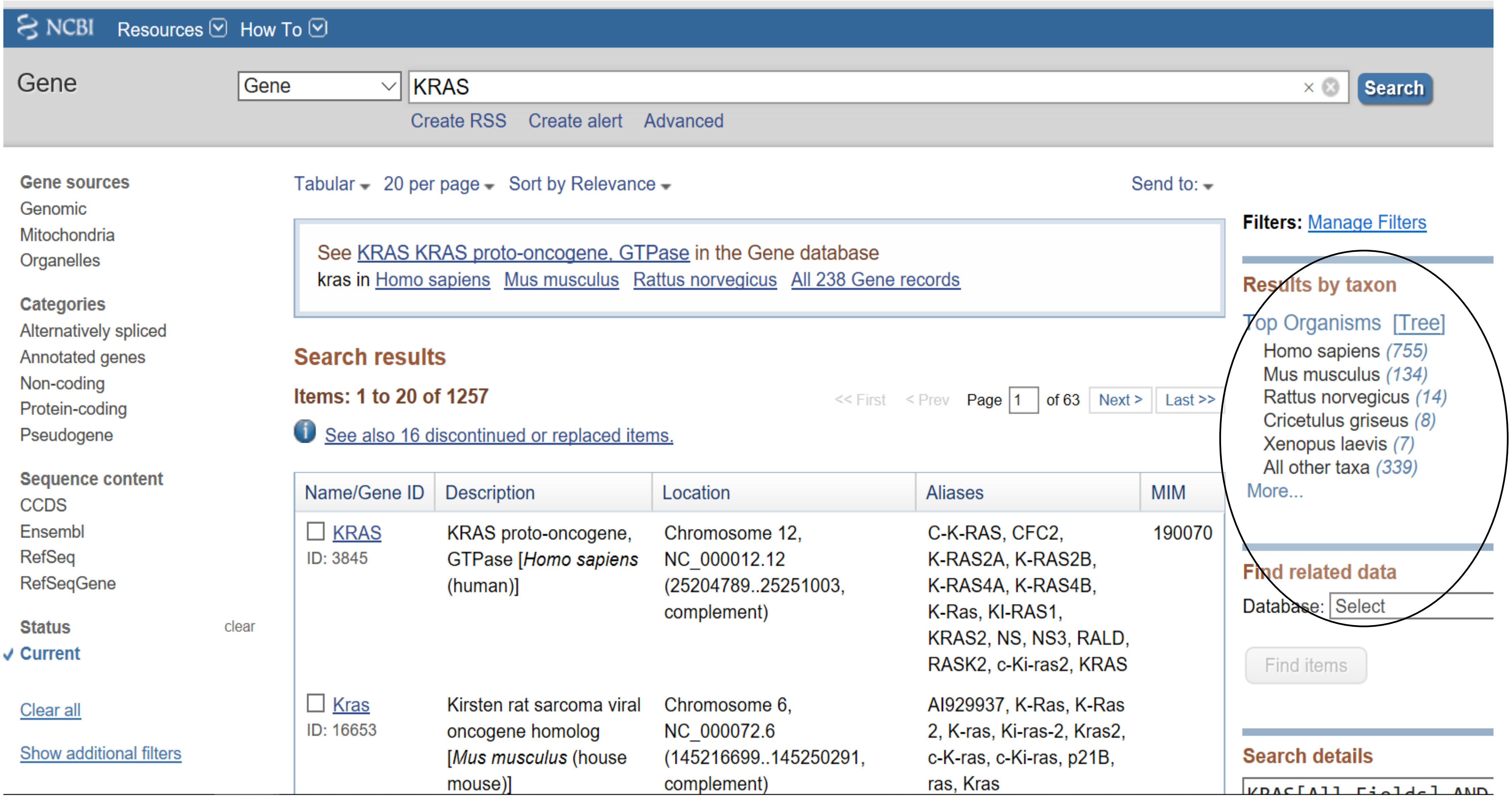
What is GenBank?

GenBank ® is the NIH genetic sequence database, an annotated collection of all publicly available DNA sequences (*Nucleic Acids Research*, 2013 Jan;41(D1):D36-42). GenBank is part of the International Nucleotide Sequence Database Collaboration, which comprises the DNA DataBank of Japan (DDBJ), the European Nucleotide Archive (ENA), and GenBank at NCBI. These three organizations exchange data on a daily basis.

A GenBank release occurs every two months and is available from the ftp site. The release notes for the current version of GenBank provide detailed information about the release and notifications of upcoming changes to GenBank. Release notes for previous GenBank releases are also available. GenBank growth statistics for both the traditional GenBank divisions and the WGS division are available from each release. GenBank growth statistics for both the traditional GenBank divisions and the WGS division are available from each release.

An annotated sample GenBank record for a *Saccharomyces cerevisiae* gene demonstrates many of the features of the GenBank flat file format.

✓ Open « Gene » and Search KRAS



Gene sources
- Genomic
- Mitochondria
- Organelles

Categories
- Alternatively spliced
- Annotated genes
- Non-coding
- Protein-coding
- Pseudogene

Sequence content
- CCDS
- Ensembl
- RefSeq
- RefSeqGene

Status clear
- ✓ Current

Clear all

Show additional filters

See KRAS KRAS proto-oncogene, GTPase in the Gene database
kras in Homo sapiens Mus musculus Rattus norvegicus All 238 Gene records

Search results

Items: 1 to 20 of 1257

<< First < Prev Page 1 of 63 Next > Last >>

See also 16 discontinued or replaced items.

Name/Gene ID	Description	Location	Aliases	MIM
□ KRAS ID: 3845	KRAS proto-oncogene, GTPase [*Homo sapiens* (human)]	Chromosome 12, NC_000012.12 (25204789..25251003, complement)	C-K-RAS, CFC2, K-RAS2A, K-RAS2B, K-RAS4A, K-RAS4B, K-Ras, KI-RAS1, KRAS2, NS, NS3, RALD, RASK2, c-Ki-ras2, KRAS	190070
□ Kras ID: 16653	Kirsten rat sarcoma viral oncogene homolog [*Mus musculus* (house mouse)]	Chromosome 6, NC_000072.6 (145216699..145250291, complement)	AI929937, K-Ras, K-Ras 2, K-ras, Ki-ras-2, Kras2, c-K-ras, c-Ki-ras, p21B, ras, Kras	

Filters: Manage Filters

Results by taxon

Top Organisms [Tree]

Homo sapiens	(755)
Mus musculus	(134)
Rattus norvegicus	(14)
Cricetulus griseus	(8)
Xenopus laevis	(7)
All other taxa	(339)

More...

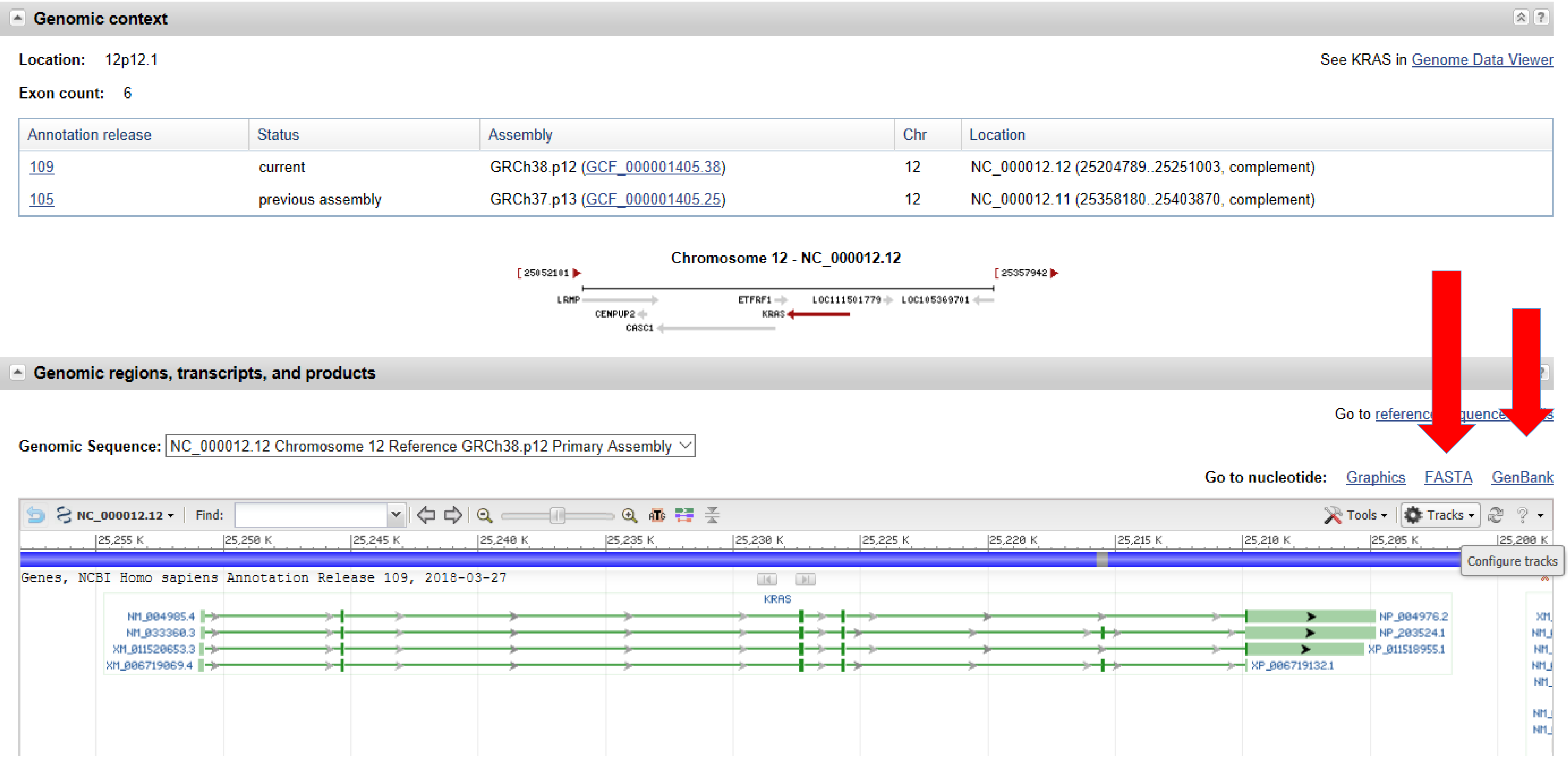
Find related data

Database: Select

Find items

Search details

KRAS[All Fields] AND



Genomic context

Location: 12p12.1

See KRAS in Genome Data Viewer

Exon count: 6

Annotation release	Status	Assembly	Chr	Location
109	current	GRCh38.p12 (GCF_000001405.38)	12	NC_000012.12 (25204789..25251003, complement)
105	previous assembly	GRCh37.p13 (GCF_000001405.25)	12	NC_000012.11 (25358180..25403870, complement)

Chromosome 12 - NC_000012.12

Genomic regions, transcripts, and products

Go to referenc quence ls

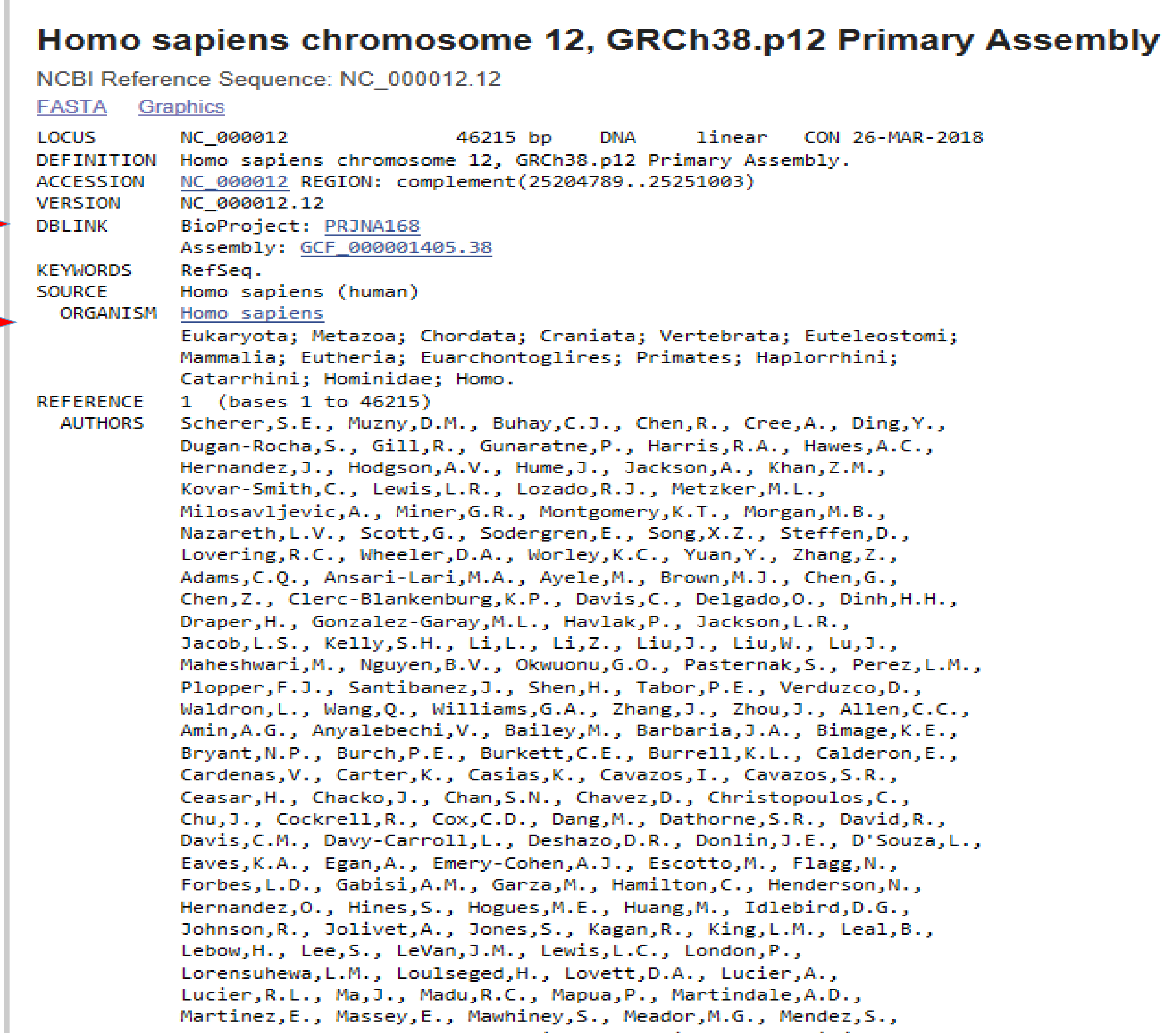
Genomic Sequence: NC_000012.12 Chromosome 12 Reference GRCh38.p12 Primary Assembly

Go to nucleotide: Graphics FASTA GenBank

Format

Accession – Key Identifier

Species

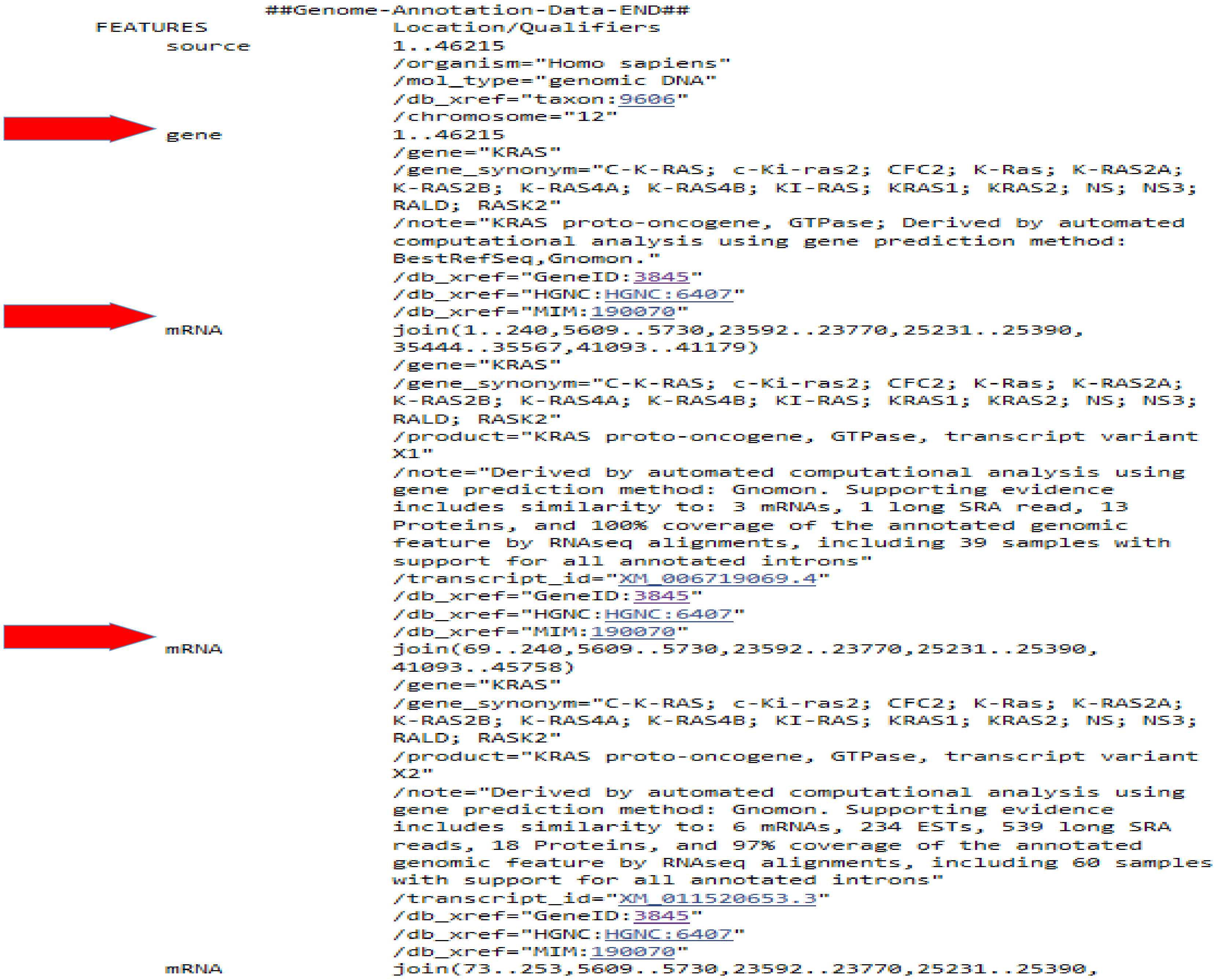


Homo sapiens chromosome 12, GRCh38.p12 Primary Assembly

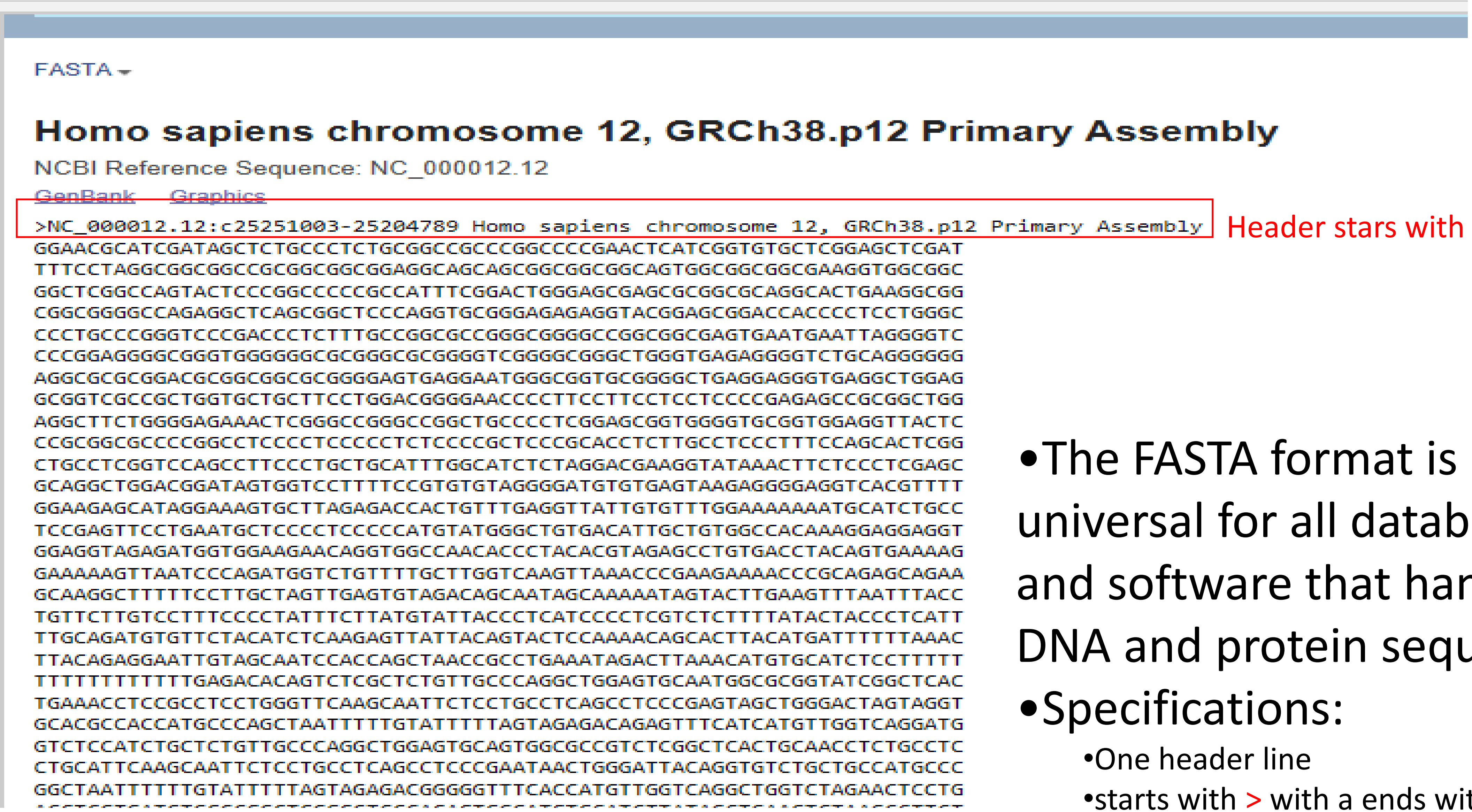
NCBI Reference Sequence: NC_000012.12

FASTA Graphics

```
LOCUS       NC_000012               46215 bp    DNA     linear   CON 26-MAR-2018
DEFINITION  Homo sapiens chromosome 12, GRCh38.p12 Primary Assembly.
ACCESSION   NC_000012 REGION: complement(25204789..25251003)
VERSION     NC_000012.12
DBLINK      BioProject: PRJNA168
            Assembly: GCF_000001405.38
KEYWORDS    RefSeq.
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
            Catarrhini; Hominidae; Homo.
REFERENCE   1  (bases 1 to 46215)
  AUTHORS   Scherer,S.E., Muzny,D.M., Buhay,C.J., Chen,R., Cree,A., Ding,Y.,
            Dugan-Rocha,S., Gill,R., Gunaratne,P., Harris,R.A., Hawes,A.C.,
            Hernandez,J., Hodgson,A.V., Hume,J., Jackson,A., Khan,Z.M.,
            Kovar-Smith,C., Lewis,L.R., Lozado,R.J., Metzker,M.L.,
            Milosavljevic,A., Miner,G.R., Montgomery,K.T., Morgan,M.B.,
            Nazareth,L.V., Scott,G., Sodergren,E., Song,X.Z., Steffen,D.,
            Lovering,R.C., Wheeler,D.A., Worley,K.C., Yuan,Y., Zhang,Z.,
            Adams,C.Q., Ansari-Lari,M.A., Ayele,M., Brown,M.J., Chen,G.,
            Chen,Z., Clerc-Blankenburg,K.P., Davis,C., Delgado,O., Dinh,H.H.,
            Draper,H., Gonzalez-Garay,M.L., Havlak,P., Jackson,L.R.,
            Jacob,L.S., Kelly,S.H., Li,L., Li,Z., Liu,J., Liu,W., Lu,J.,
            Maheshwari,M., Nguyen,B.V., Okwuonu,G.O., Pasternak,S., Perez,L.M.,
            Plopper,F.J., Santibanez,J., Shen,H., Tabor,P.E., Verduzco,D.,
            Waldron,L., Wang,Q., Williams,G.A., Zhang,J., Zhou,J., Allen,C.C.,
            Amin,A.G., Anyalebechi,V., Bailey,M., Barbaria,J.A., Bimage,K.E.,
            Bryant,N.P., Burch,P.E., Burkett,C.E., Burrell,K.L., Calderon,E.,
            Cardenas,V., Carter,K., Casias,K., Cavazos,I., Cavazos,S.R.,
            Ceasar,H., Chacko,J., Chan,S.N., Chavez,D., Christopoulos,C.,
            Chu,J., Cockrell,R., Cox,C.D., Dang,M., Dathorne,S.R., David,R.,
            Davis,C.M., Davy-Carroll,L., Deshazo,D.R., Donlin,J.E., D'Souza,L.,
            Eaves,K.A., Egan,A., Emery-Cohen,A.J., Escotto,M., Flagg,N.,
            Forbes,L.D., Gabisi,A.M., Garza,M., Hamilton,C., Henderson,N.,
            Hernandez,O., Hines,S., Hogues,M.E., Huang,M., Idlebird,D.G.,
            Johnson,R., Jolivet,A., Jones,S., Kagan,R., King,L.M., Leal,B.,
            Lebow,H., Lee,S., LeVan,J.M., Lewis,L.C., London,P.,
            Lorensuhewa,L.M., Loulseged,H., Lovett,D.A., Lucier,A.,
            Lucier,R.L., Ma,J., Madu,R.C., Mapua,P., Martindale,A.D.,
            Martinez,E., Massey,E., Mawhiney,S., Meador,M.G., Mendez,S.,
```



```
     ##Genome-Annotation-Data-END##
FEATURES             Location/Qualifiers
     source          1..46215
                     /organism="Homo sapiens"
                     /mol_type="genomic DNA"
                     /db_xref="taxon:9606"
                     /chromosome="12"
     gene            1..46215
                     /gene="KRAS"
                     /gene_synonym="C-K-RAS; c-Ki-ras2; CFC2; K-Ras; K-RAS2A;
                     K-RAS2B; K-RAS4A; K-RAS4B; KI-RAS; KRAS1; KRAS2; NS; NS3;
                     RALD; RASK2"
                     /note="KRAS proto-oncogene, GTPase; Derived by automated
                     computational analysis using gene prediction method:
                     BestRefSeq,Gnomon."
                     /db_xref="GeneID:3845"
                     /db_xref="HGNC:HGNC:6407"
                     /db_xref="MIM:190070"
     mRNA            join(1..240,5609..5730,23592..23770,25231..25390,
                     35444..35567,41093..41179)
                     /gene="KRAS"
                     /gene_synonym="C-K-RAS; c-Ki-ras2; CFC2; K-Ras; K-RAS2A;
                     K-RAS2B; K-RAS4A; K-RAS4B; KI-RAS; KRAS1; KRAS2; NS; NS3;
                     RALD; RASK2"
                     /product="KRAS proto-oncogene, GTPase, transcript variant
                     X1"
                     /note="Derived by automated computational analysis using
                     gene prediction method: Gnomon. Supporting evidence
                     includes similarity to: 3 mRNAs, 1 long SRA read, 13
                     Proteins, and 100% coverage of the annotated genomic
                     feature by RNAseq alignments, including 39 samples with
                     support for all annotated introns"
                     /transcript_id="XM_006719069.4"
                     /db_xref="GeneID:3845"
                     /db_xref="HGNC:HGNC:6407"
                     /db_xref="MIM:190070"
     mRNA            join(69..240,5609..5730,23592..23770,25231..25390,
                     41093..45758)
                     /gene="KRAS"
                     /gene_synonym="C-K-RAS; c-Ki-ras2; CFC2; K-Ras; K-RAS2A;
                     K-RAS2B; K-RAS4A; K-RAS4B; KI-RAS; KRAS1; KRAS2; NS; NS3;
                     RALD; RASK2"
                     /product="KRAS proto-oncogene, GTPase, transcript variant
                     X2"
                     /note="Derived by automated computational analysis using
                     gene prediction method: Gnomon. Supporting evidence
                     includes similarity to: 6 mRNAs, 234 ESTs, 539 long SRA
                     reads, 18 Proteins, and 97% coverage of the annotated
                     genomic feature by RNAseq alignments, including 60 samples
                     with support for all annotated introns"
                     /transcript_id="XM_011520653.3"
                     /db_xref="GeneID:3845"
                     /db_xref="HGNC:HGNC:6407"
                     /db_xref="MIM:190070"
     mRNA            join(73..253,5609..5730,23592..23770,25231..25390,
```



Homo sapiens chromosome 12, GRCh38.p12 Primary Assembly

NCBI Reference Sequence: NC_000012.12

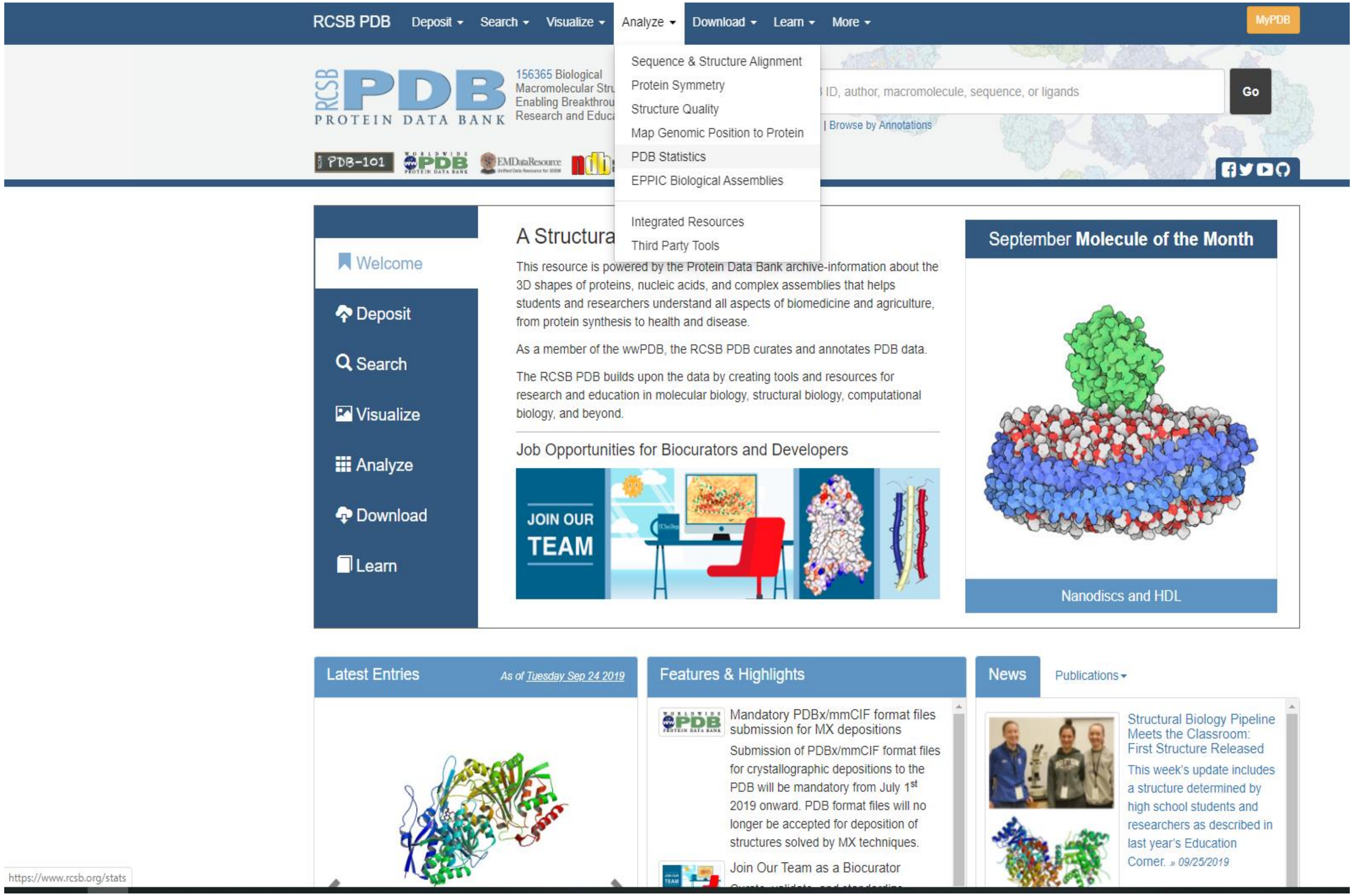
GenBank Graphics

```
>NC_000012.12:c25251003-25204789 Homo sapiens chromosome 12, GRCh38.p12 Primary Assembly
GGAACGCATCGATAGCTCTGCCCTCTGCGGCCGCCCGGCCCCGAACTCATCGGTGTGCTCGGAGCTCGAT
TTTCCTAGGCGGCGGCCGCGGCGGCGGAGGCAGCAGCGGCGGCGGCAGTGGCGGCGGCGAAGGTGGCGGC
GGCTCGGCCAGTACTCCCGGCCCCCGCCATTTCGGACTGGGAGCGAGCGCGGCGCAGGCACTGAAGGCGG
CGGCGGGGCCAGAGGCTCAGCGGCTCCCAGGTGCGGGAGAGAGGTACGGAGCGGACCACCCCTCCTGGGC
CCCTGCCCGGGTCCCGACCCTCTTTGCCGGCGCCGGGCGGGGCCGGCGGCGAGTGAATGAATTAGGGGTC
CCCGGAGGGGCGGGTGGGGGGCGCGGGCGCGGGGTCGGGGCGGGCTGGGTGAGAGGGGTCTGCAGGGGGG
AGGCGCGCGGACGCGGCGGCGCGGGGAGTGAGGAATGGGCGGTGCGGGGCTGAGGAGGGTGAGGCTGGAG
GCGGTCGCCGCTGGTGCTGCTTCCTGGACGGGGAACCCCTTCCTTCCTCCTCCCCGAGAGCCGCGGCTGG
AGGCTTCTGGGGAGAAACTCGGGCCGGGCCGGCTGCCCCTCGGAGCGGTGGGGTGCGGTGGAGGTTACTC
CCGCGGCGCCCCGGCCTCCCCTCCCCCTCTCCCCGCTCCCGCACCTCTTGCCTCCCTTTCCAGCACTCGG
CTGCCTCGGTCCAGCCTTCCCTGCTGCATTTGGCATCTCTAGGACGAAGGTATAAACTTCTCCCTCGAGC
GCAGGCTGGACGGATAGTGGTCCTTTTCCGTGTGTAGGGGATGTGTGAGTAAGAGGGGAGGTCACGTTTT
GGAAGAGCATAGGAAAGTGCTTAGAGACCACTGTTTGAGGTTATTGTGTTTGGAAAAAAATGCATCTGCC
TCCGAGTTCCTGAATGCTCCCCTCCCCCATGTATGGGCTGTGACATTGCTGTGGCCACAAAGGAGGAGGT
GGAGGTAGAGATGGTGGAAGAACAGGTGGCCAACACCCTACACGTAGAGCCTGTGACCTACAGTGAAAAG
GAAAAAGTTAATCCCAGATGGTCTGTTTTGCTTGGTCAAGTTAAACCCGAAGAAAACCCGCAGAGCAGAA
GCAAGGCTTTTTCCTTGCTAGTTGAGTGTAGACAGCAATAGCAAAAATAGTACTTGAAGTTTAATTTACC
TGTTCTTGTCCTTTCCCCTATTTCTTATGTATTACCCTCATCCCCTCGTCTCTTTTATACTACCCTCATT
TTGCAGATGTGTTCTACATCTCAAGAGTTATTACAGTACTCCAAAACAGCACTTACATGATTTTTTAAAC
TTACAGAGGAATTGTAGCAATCCACCAGCTAACCGCCTGAAATAGACTTAAACATGTGCATCTCCTTTTT
TTTTTTTTTTTTGAGACACAGTCTCGCTCTGTTGCCCAGGCTGGAGTGCAATGGCGCGGTATCGGCTCAC
TGAAACCTCCGCCTCCTGGGTTCAAGCAATTCTCCTGCCTCAGCCTCCCGAGTAGCTGGGACTAGTAGGT
GCACGCCACCATGCCCAGCTAATTTTTGTATTTTTAGTAGAGACAGAGTTTCATCATGTTGGTCAGGATG
GTCTCCATCTGCTCTGTTGCCCAGGCTGGAGTGCAGTGGCGCCGTCTCGGCTCACTGCAACCTCTGCCTC
CTGCATTCAAGCAATTCTCCTGCCTCAGCCTCCCGAATAACTGGGATTACAGGTGTCTGCTGCCATGCCC
GGCTAATTTTTTGTATTTTTAGTAGAGACGGGGGTTTCACCATGTTGGTCAGGCTGGTCTAGAACTCCTG
```

Header stars with ”>” sign

- The FASTA format is now universal for all databases and software that handles DNA and protein sequences
- Specifications:
  - One header line
  - starts with > with a ends with [return]

https://www.rcsb.org/



A Structura

This resource is powered by the Protein Data Bank archive-information about the 3D shapes of proteins, nucleic acids, and complex assemblies that helps students and researchers understand all aspects of biomedicine and agriculture, from protein synthesis to health and disease.

As a member of the wwPDB, the RCSB PDB curates and annotates PDB data.

The RCSB PDB builds upon the data by creating tools and resources for research and education in molecular biology, structural biology, computational biology, and beyond.

Job Opportunities for Biocurators and Developers

Mandatory PDBx/mmCIF format files submission for MX depositions

Submission of PDBx/mmCIF format files for crystallographic depositions to the PDB will be mandatory from July 1st 2019 onward. PDB format files will no longer be accepted for deposition of structures solved by MX techniques.

Join Our Team as a Biocurator

Structural Biology Pipeline Meets the Classroom: First Structure Released

This week's update includes a structure determined by high school students and researchers as described in last year's Education Corner. » 09/25/2019

Search ‘6Q6I’ : *Lysine decarboxylase A from Pseudomonas aeruginosa*

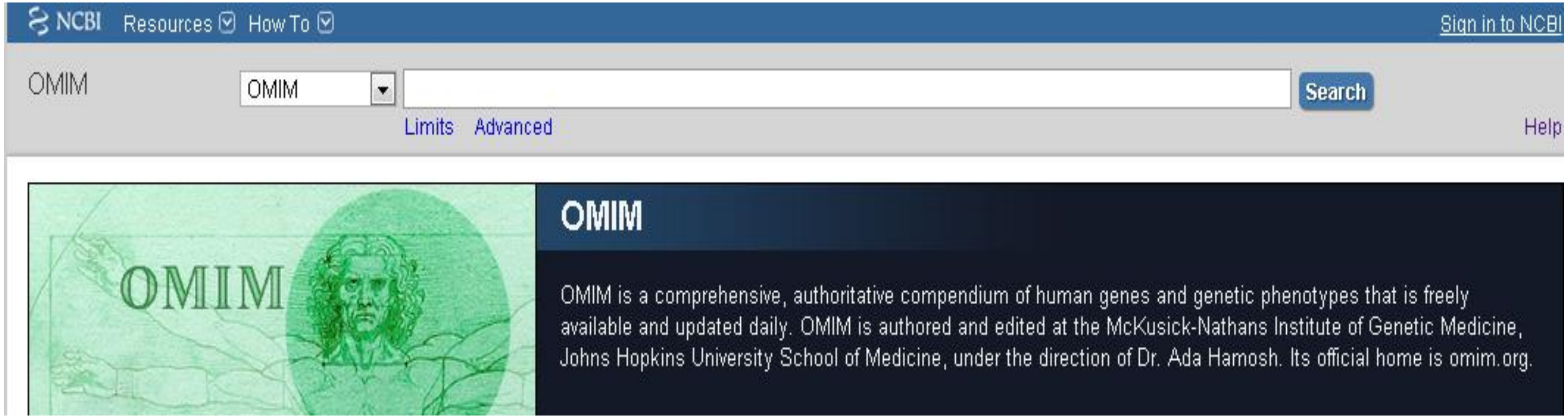
Classification: OXIDOREDUCTASE (type)

Organism(s): Pseudomonas aeruginosa

Expression System: Escherichia coli

OMIM database

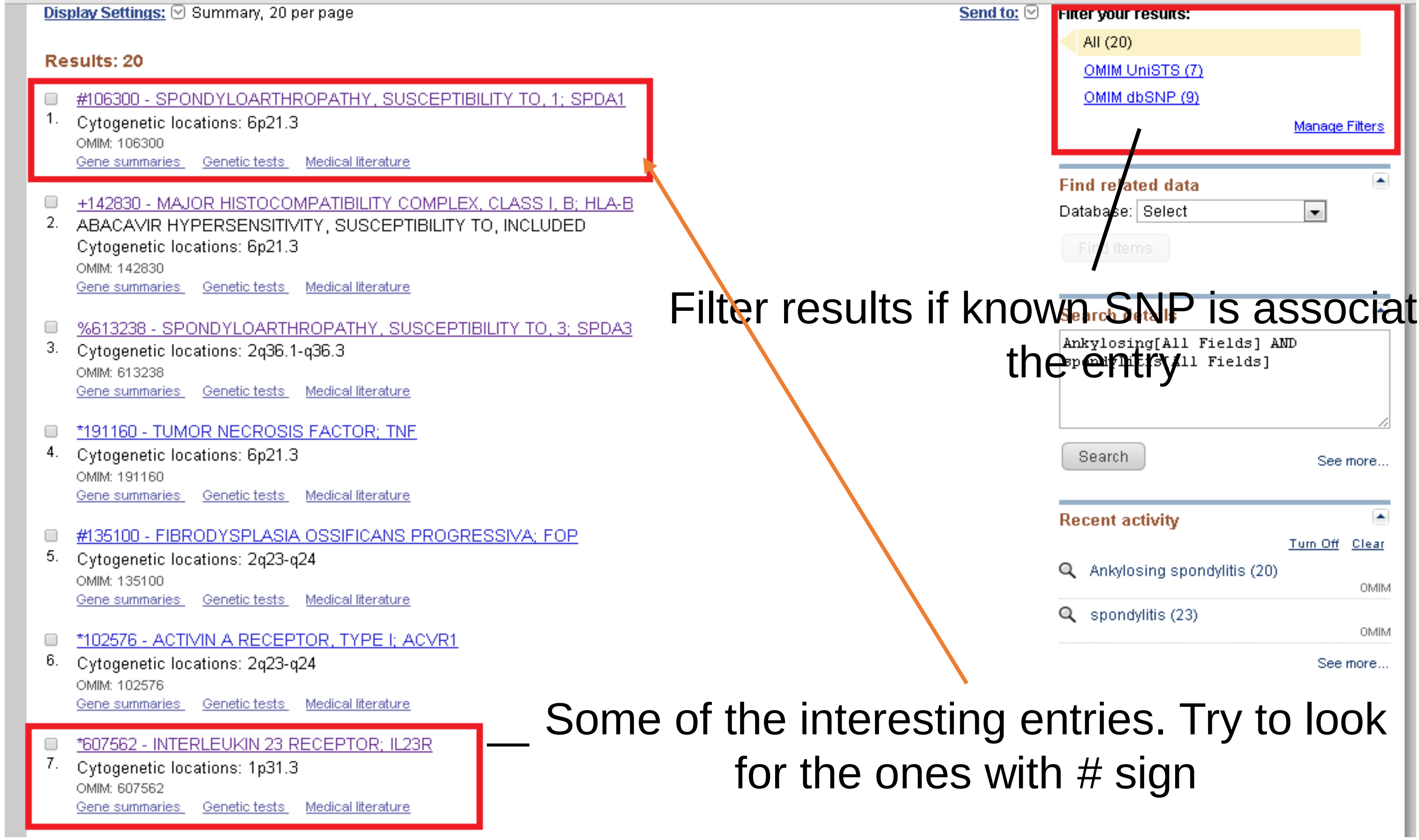
- Online Mendelian Inheritance in Man (OMIM)
- ”information on all known mendelian disorders linked to over 12,000 genes”
- “Started at 1960s by Dr. Victor A. McKusick as a catalog of mendelian traits and disorders”
- Linked disease data
- Links disease phenotypes and causative genes
- Used by physicians and geneticists



OMIM is a comprehensive, authoritative compendium of human genes and genetic phenotypes that is freely available and updated daily. OMIM is authored and edited at the McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, under the direction of Dr. Ada Hamosh. Its official home is omim.org.

OMIM-search results

- Look for the entires that link to the genes. Apply filters if needed



Display Settings: Summary, 20 per page

Send to:

Results: 20

1. □ #106300 - SPONDYLOARTHROPATHY, SUSCEPTIBILITY TO, 1; SPDA1
Cytogenetic locations: 6p21.3
OMIM: 106300
Gene summaries Genetic tests Medical literature

2. □ +142830 - MAJOR HISTOCOMPATIBILITY COMPLEX, CLASS I, B; HLA-B
ABACAVIR HYPERSENSITIVITY, SUSCEPTIBILITY TO, INCLUDED
Cytogenetic locations: 6p21.3
OMIM: 142830
Gene summaries Genetic tests Medical literature

3. □ %613238 - SPONDYLOARTHROPATHY, SUSCEPTIBILITY TO, 3; SPDA3
Cytogenetic locations: 2q36.1-q36.3
OMIM: 613238
Gene summaries Genetic tests Medical literature

4. □ *191160 - TUMOR NECROSIS FACTOR; TNF
Cytogenetic locations: 6p21.3
OMIM: 191160
Gene summaries Genetic tests Medical literature

5. □ #135100 - FIBRODYSPLASIA OSSIFICANS PROGRESSIVA; FOP
Cytogenetic locations: 2q23-q24
OMIM: 135100
Gene summaries Genetic tests Medical literature

6. □ *102576 - ACTIVIN A RECEPTOR, TYPE I; ACVR1
Cytogenetic locations: 2q23-q24
OMIM: 102576
Gene summaries Genetic tests Medical literature

7. □ *607562 - INTERLEUKIN 23 RECEPTOR; IL23R
Cytogenetic locations: 1p31.3
OMIM: 607562
Gene summaries Genetic tests Medical literature

Filter your results:
All (20)
OMIM UniSTS (7)
OMIM dbSNP (9)
Manage Filters

Find related data
Database: Select
Find items

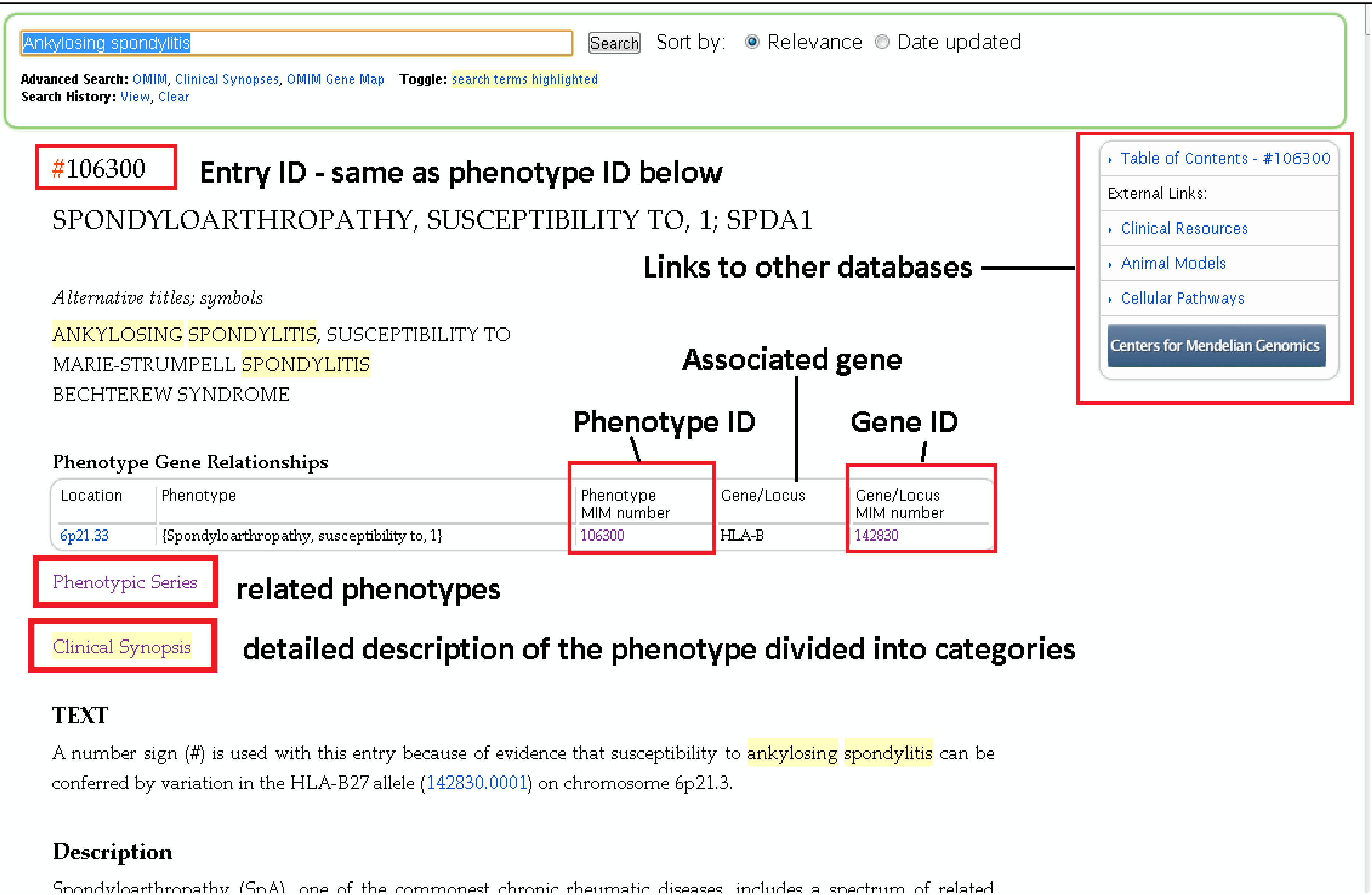
Search details
Ankylosing[All Fields] AND spondylitis[All Fields]
Search
See more...

Recent activity
Turn Off Clear
Ankylosing spondylitis (20) OMIM
spondylitis (23) OMIM
See more...

Filter results if known SNP is associated to the entry

Some of the interesting entries. Try to look for the ones with # sign

OMIM-entries


Ankylosing spondylitis
Search
Sort by: ● Relevance ○ Date updated
Advanced Search: OMIM, Clinical Synopses, OMIM Gene Map
Toggle: search terms highlighted
Search History: View, Clear
Table of Contents - #106300
External Links:
Clinical Resources
Animal Models
Cellular Pathways
Centers for Mendelian Genomics

#106300

Entry ID - same as phenotype ID below

SPONDYLOARTHROPATHY, SUSCEPTIBILITY TO, 1; SPDA1

Links to other databases

Alternative titles; symbols

ANKYLOSING SPONDYLITIS, SUSCEPTIBILITY TO
MARIE-STRUMPELL SPONDYLITIS
BECHTEREW SYNDROME

Associated gene

Phenotype ID

Gene ID

Phenotype Gene Relationships

Location	Phenotype	Phenotype MIM number	Gene/Locus	Gene/Locus MIM number
6p21.33	{Spondyloarthropathy, susceptibility to, 1}	106300	HLA-B	142830

Phenotypic Series

related phenotypes

Clinical Synopsis

detailed description of the phenotype divided into categories

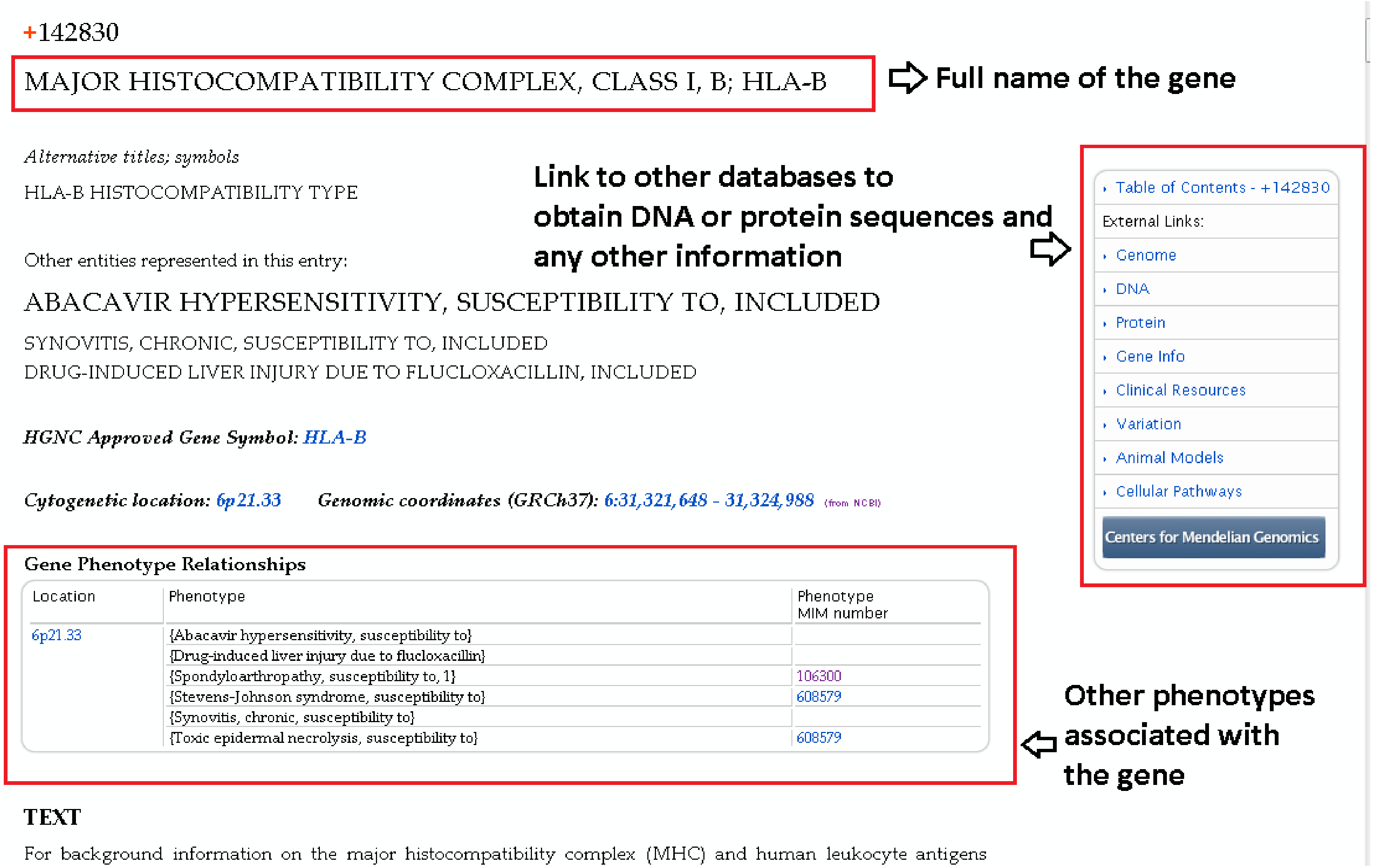
TEXT

A number sign (#) is used with this entry because of evidence that susceptibility to ankylosing spondylitis can be conferred by variation in the HLA-B27 allele (142830.0001) on chromosome 6p21.3.

Description

Spondyloarthropathy (SpA), one of the commonest chronic rheumatic diseases, includes a spectrum of related

OMIM Gene ID -entries



+142830

MAJOR HISTOCOMPATIBILITY COMPLEX, CLASS I, B; HLA-B

Full name of the gene

Alternative titles; symbols

HLA-B HISTOCOMPATIBILITY TYPE

Other entities represented in this entry:

ABACAVIR HYPERSENSITIVITY, SUSCEPTIBILITY TO, INCLUDED

SYNOVITIS, CHRONIC, SUSCEPTIBILITY TO, INCLUDED
DRUG-INDUCED LIVER INJURY DUE TO FLUCLOXACILLIN, INCLUDED

Link to other databases to obtain DNA or protein sequences and any other information

HGNC Approved Gene Symbol: ***HLA-B***

Cytogenetic location: ***6p21.33*** ***Genomic coordinates (GRCh37):*** ***6:31,321,648 - 31,324,988*** (from NCBI)

Gene Phenotype Relationships

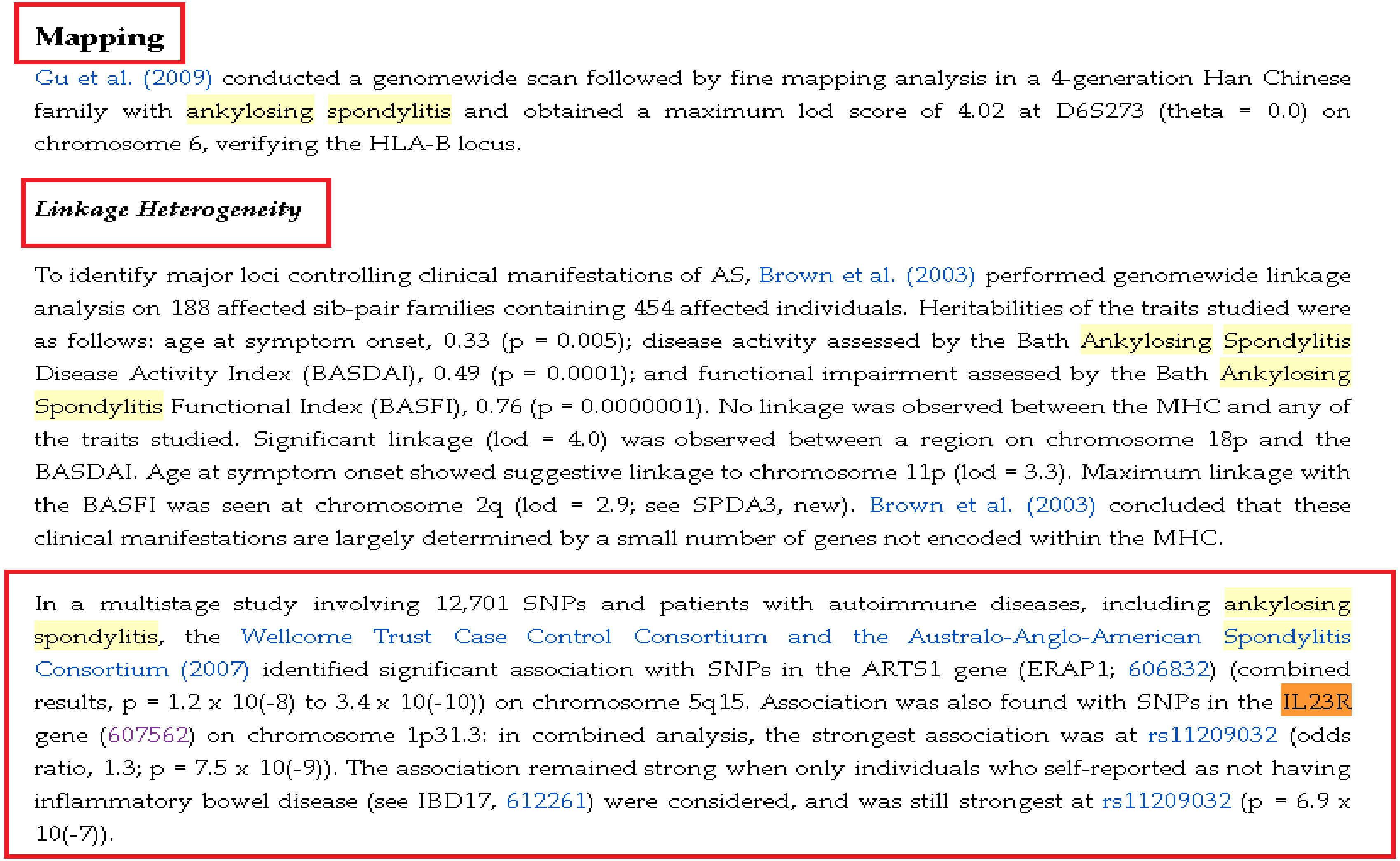
Location	Phenotype	Phenotype MIM number
6p21.33	{Abacavir hypersensitivity, susceptibility to}	
	{Drug-induced liver injury due to flucloxacillin}	
	{Spondyloarthropathy, susceptibility to, 1}	106300
	{Stevens-Johnson syndrome, susceptibility to}	608579
	{Synovitis, chronic, susceptibility to}	
	{Toxic epidermal necrolysis, susceptibility to}	608579

Other phenotypes associated with the gene

TEXT

For background information on the major histocompatibility complex (MHC) and human leukocyte antigens

OMIM-Finding disease linked genes



Mapping

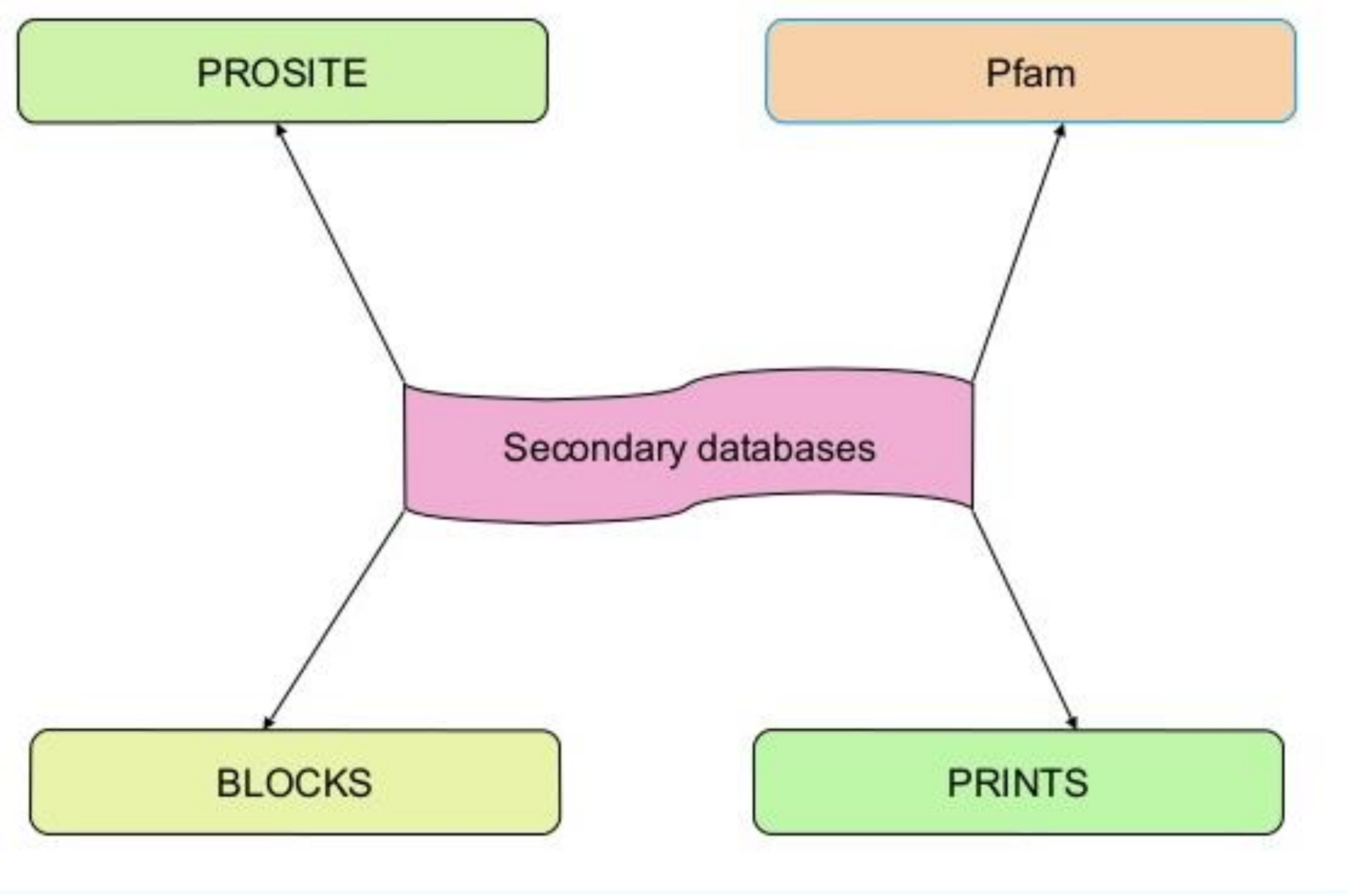
Gu et al. (2009) conducted a genomewide scan followed by fine mapping analysis in a 4-generation Han Chinese family with ankylosing spondylitis and obtained a maximum lod score of 4.02 at D6S273 (theta = 0.0) on chromosome 6, verifying the HLA-B locus.

Linkage Heterogeneity

To identify major loci controlling clinical manifestations of AS, Brown et al. (2003) performed genomewide linkage analysis on 188 affected sib-pair families containing 454 affected individuals. Heritabilities of the traits studied were as follows: age at symptom onset, 0.33 (p = 0.005); disease activity assessed by the Bath Ankylosing Spondylitis Disease Activity Index (BASDAI), 0.49 (p = 0.0001); and functional impairment assessed by the Bath Ankylosing Spondylitis Functional Index (BASFI), 0.76 (p = 0.0000001). No linkage was observed between the MHC and any of the traits studied. Significant linkage (lod = 4.0) was observed between a region on chromosome 18p and the BASDAI. Age at symptom onset showed suggestive linkage to chromosome 11p (lod = 3.3). Maximum linkage with the BASFI was seen at chromosome 2q (lod = 2.9; see SPDA3, new). Brown et al. (2003) concluded that these clinical manifestations are largely determined by a small number of genes not encoded within the MHC.

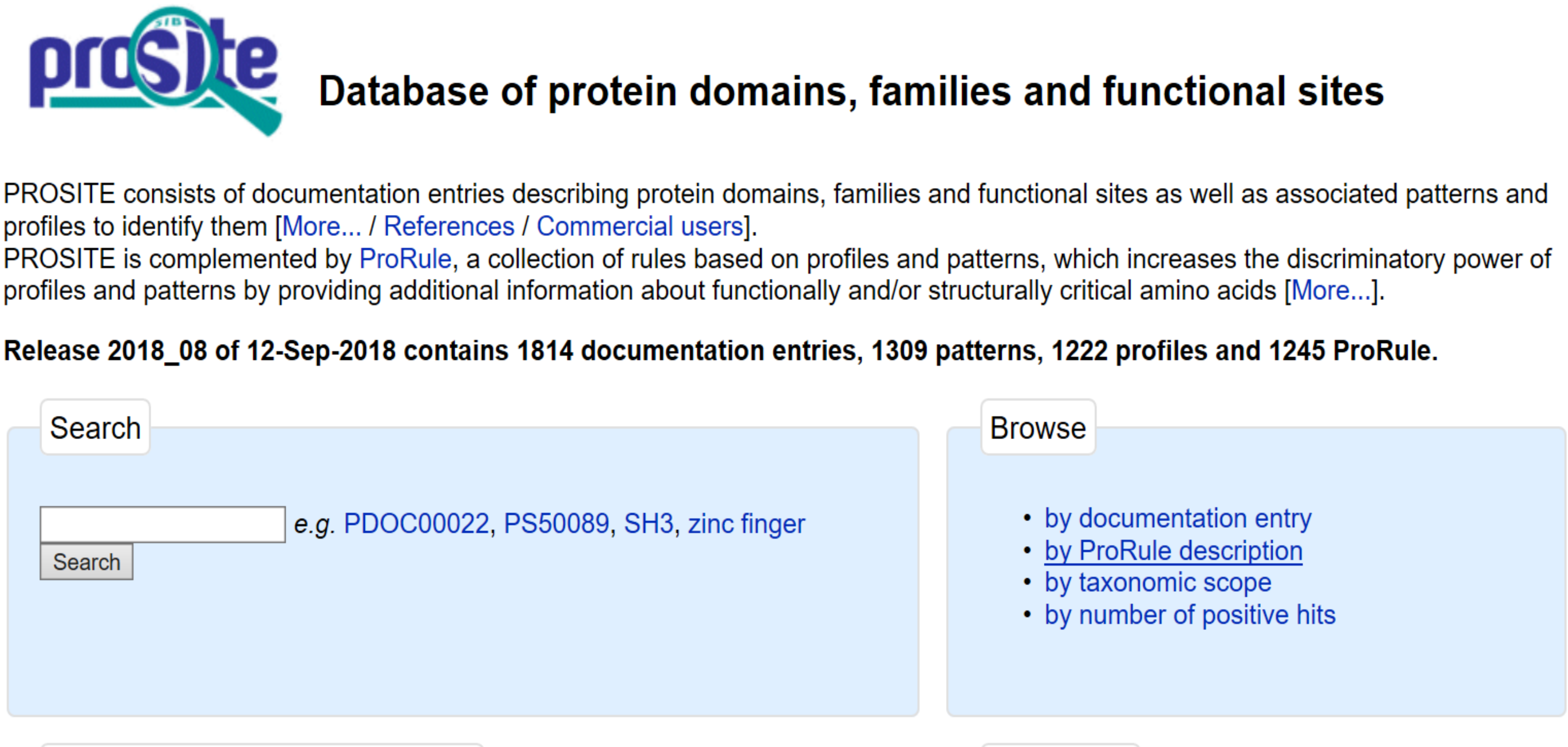
In a multistage study involving 12,701 SNPs and patients with autoimmune diseases, including ankylosing spondylitis, the Wellcome Trust Case Control Consortium and the Australo-Anglo-American Spondylitis Consortium (2007) identified significant association with SNPs in the ARTS1 gene (ERAP1; 606832) (combined results, p = 1.2 x 10(-8) to 3.4 x 10(-10)) on chromosome 5q15. Association was also found with SNPs in the IL23R gene (607562) on chromosome 1p31.3: in combined analysis, the strongest association was at rs11209032 (odds ratio, 1.3; p = 7.5 x 10(-9)). The association remained strong when only individuals who self-reported as not having inflammatory bowel disease (see IBD17, 612261) were considered, and was still strongest at rs11209032 (p = 6.9 x 10(-7)).

Secondary Databases



Secondary Database : PROSITE

- ✓ Open link https://prosite.expasy.org/



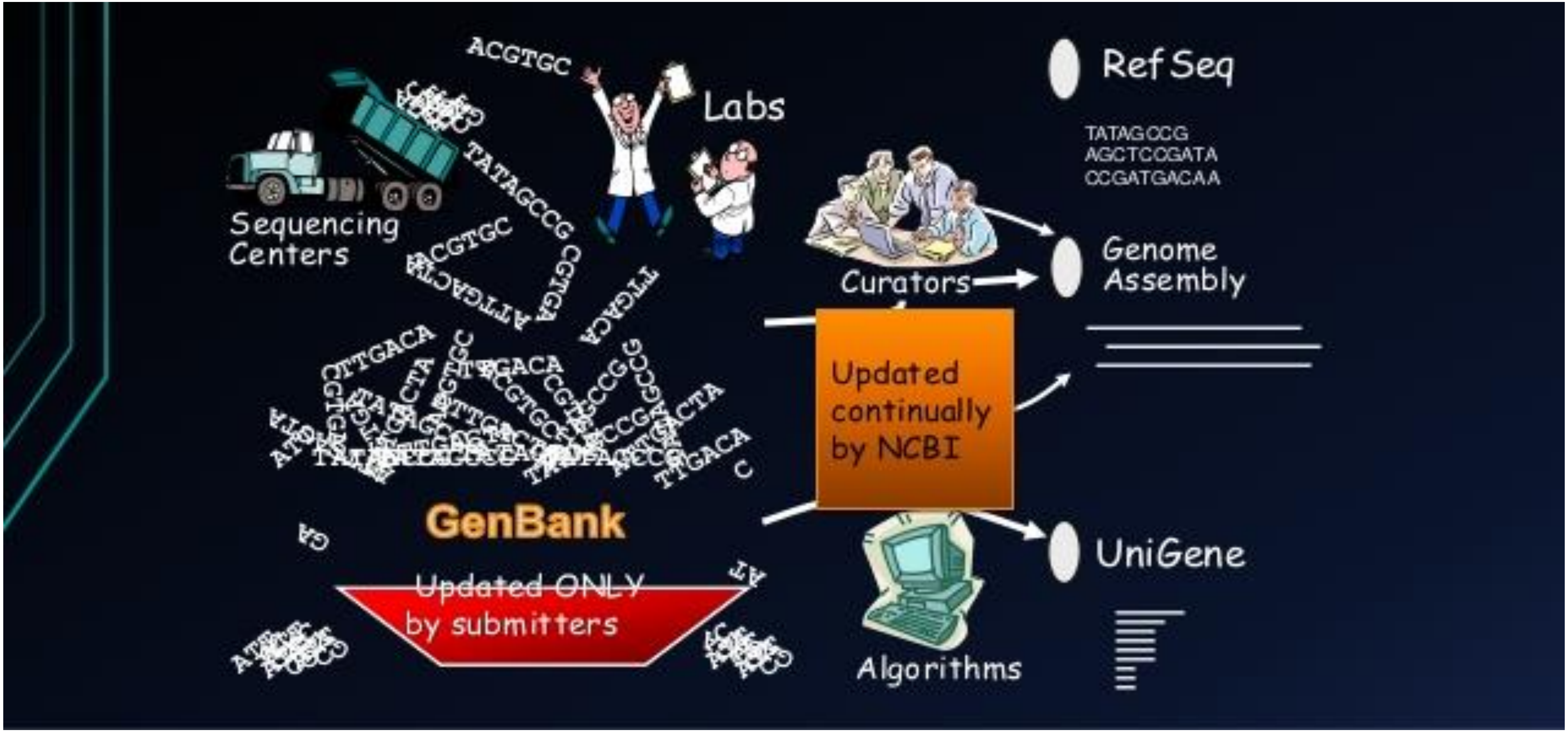
Database of protein domains, families and functional sites

PROSITE consists of documentation entries describing protein domains, families and functional sites as well as associated patterns and profiles to identify them [More... / References / Commercial users].
PROSITE is complemented by ProRule, a collection of rules based on profiles and patterns, which increases the discriminatory power of profiles and patterns by providing additional information about functionally and/or structurally critical amino acids [More...].

Release 2018_08 of 12-Sep-2018 contains 1814 documentation entries, 1309 patterns, 1222 profiles and 1245 ProRule.

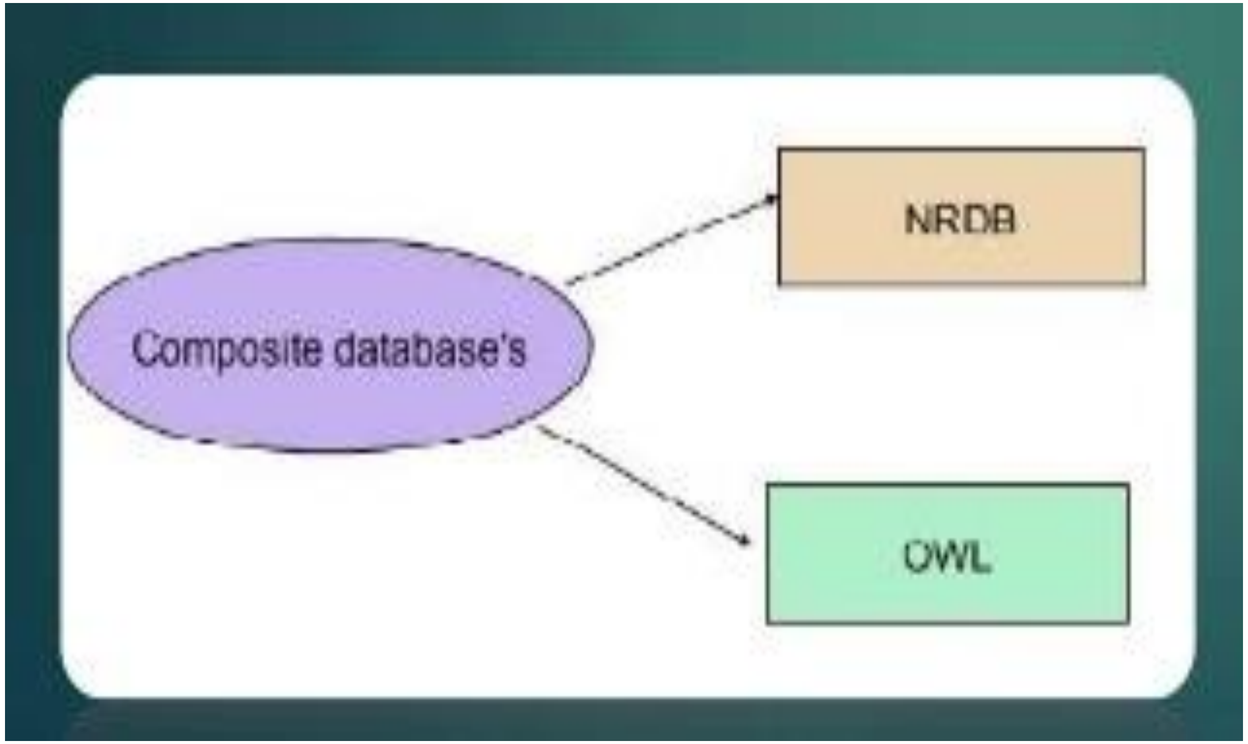
- ✓ Search homeobox

Primary vs Secondary Databases

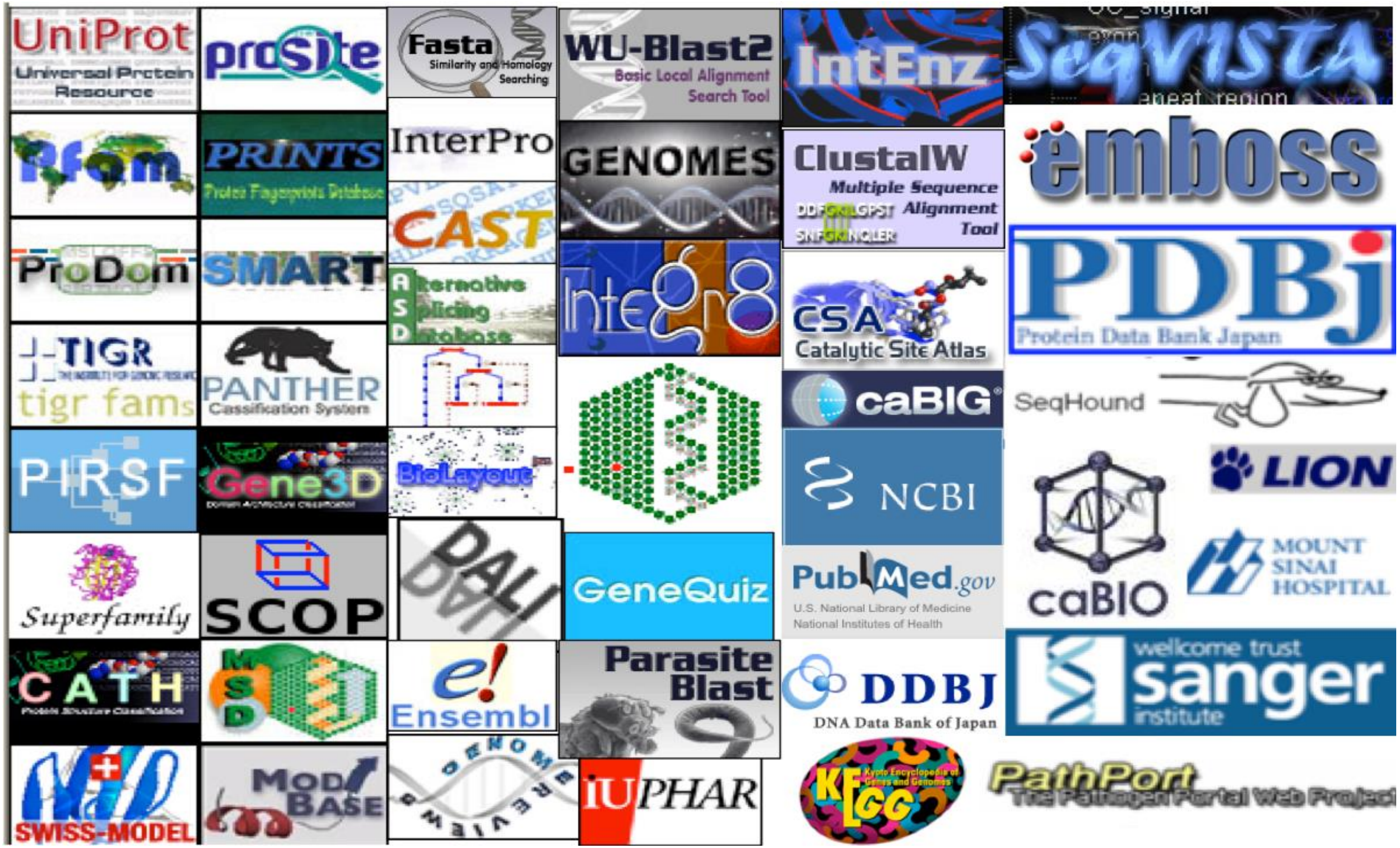


Composite Databases

- ✓ **Collection of various primary databases sequences**
- ✓ **Renders sequence searching highly efficient as it searches multiple resources**

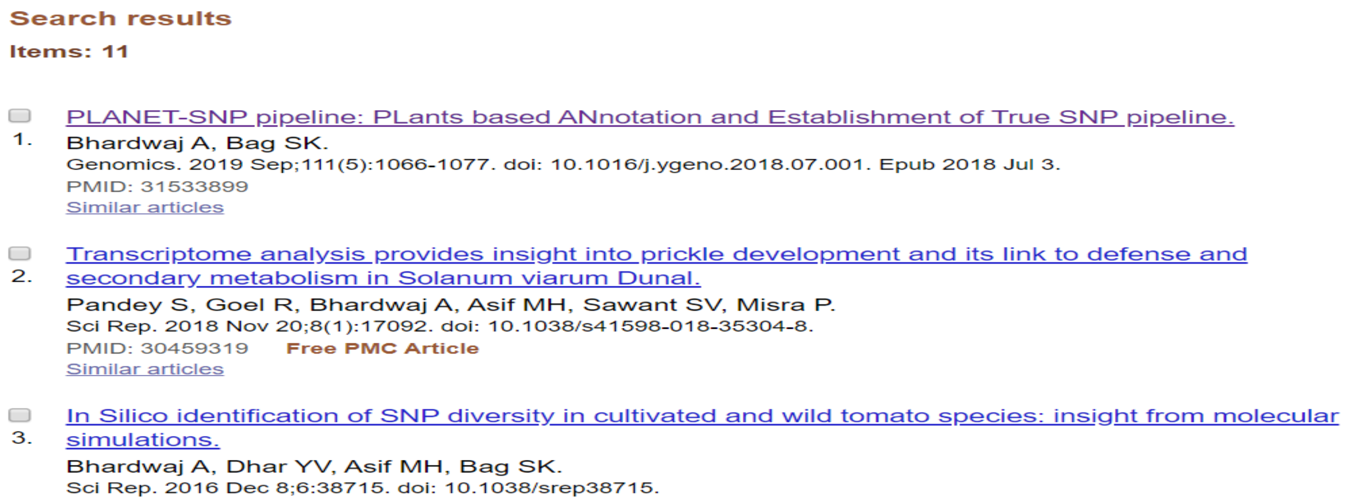


Other Databases



PubMed database

- PubMed is one of the best known database in the whole scientific community
- Most of biology related literature from all the related fields are being indexed by this database
- It has very powerful mechanism of constructing search queries
  - Many search fields
  - Logical operators (AND, OR)
- Provides electronic links to most journals
- Example of searching by author articles published within 2012-2013



Search results

Items: 11

□ 1. PLANET-SNP pipeline: PLants based ANnotation and Establishment of True SNP pipeline.
Bhardwaj A, Bag SK.
Genomics. 2019 Sep;111(5):1066-1077. doi: 10.1016/j.ygeno.2018.07.001. Epub 2018 Jul 3.
PMID: 31533899
Similar articles

□ 2. Transcriptome analysis provides insight into prickle development and its link to defense and secondary metabolism in Solanum viarum Dunal.
Pandey S, Goel R, Bhardwaj A, Asif MH, Sawant SV, Misra P.
Sci Rep. 2018 Nov 20;8(1):17092. doi: 10.1038/s41598-018-35304-8.
PMID: 30459319 **Free PMC Article**
Similar articles

□ 3. In Silico identification of SNP diversity in cultivated and wild tomato species: insight from molecular simulations.
Bhardwaj A, Dhar YV, Asif MH, Bag SK.
Sci Rep. 2016 Dec 8;6:38715. doi: 10.1038/srep38715.

Applications of Bioinformatics : Medical Implications

- ✓**Pharmacogenomics**
  - ✓Not all drugs work on all patients, some good drugs cause death in some patients
  - ✓So by doing a gene analysis before the treatment the offensive drugs can be avoided
  - ✓Also drugs which cause death to most can be used on a minority to whose genes that drug is well suited – volunteers wanted!
  - ✓Customized treatment
- ✓Gene Therapy
  - ✓Replace or supply the defective or missing gene
  - ✓E.g: Insulin and Factor VIII or Haemophilia

Applications of Bioinformatics : Diagnosis of Disease

✓Diagnosis of disease

  ❑Identification of genes which cause the disease will help detect disease at early stage e.g. Huntington disease -

✓Symptoms – uncontrollable dance like movements, mental disturbance, personality changes and intellectual impairment

✓Death in 10-15 years

✓The gene responsible for the disease has been identified

✓Contains excessively repeated sections of CAG

✓So once analyzed the couple can be counseled

Applications of Bioinformatics : Drug Design

- ✓Can go up to 15yrs and $700million
- ✓One of the goals of bioinformatics is to reduce the time and cost involved with it.
- ✓The process
    - ✓Discovery
        - ✓Computational methods can improves this
    - ✓Testing

All about Post GWAS

Post GWAS : Interpreting SNPs

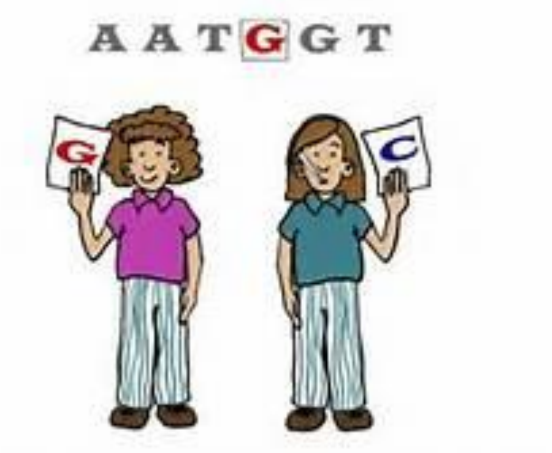
Look at the functionality of your SNP (SNPdoc)

Literature search – can you give biological plausibility?

Other tests: pathway analysis / Gene based tests

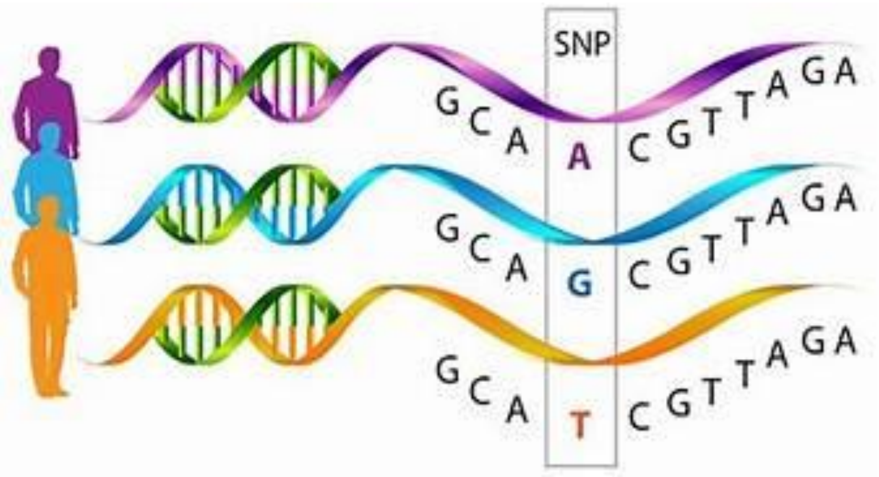
Manual Search = No

Multiple softwares are available

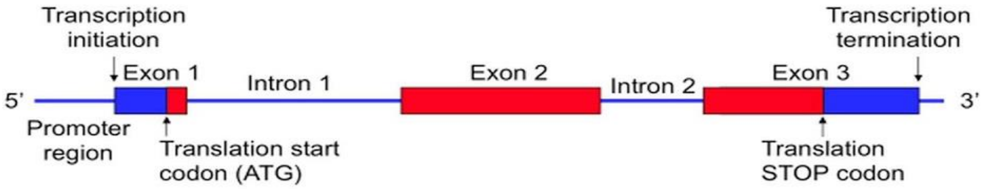


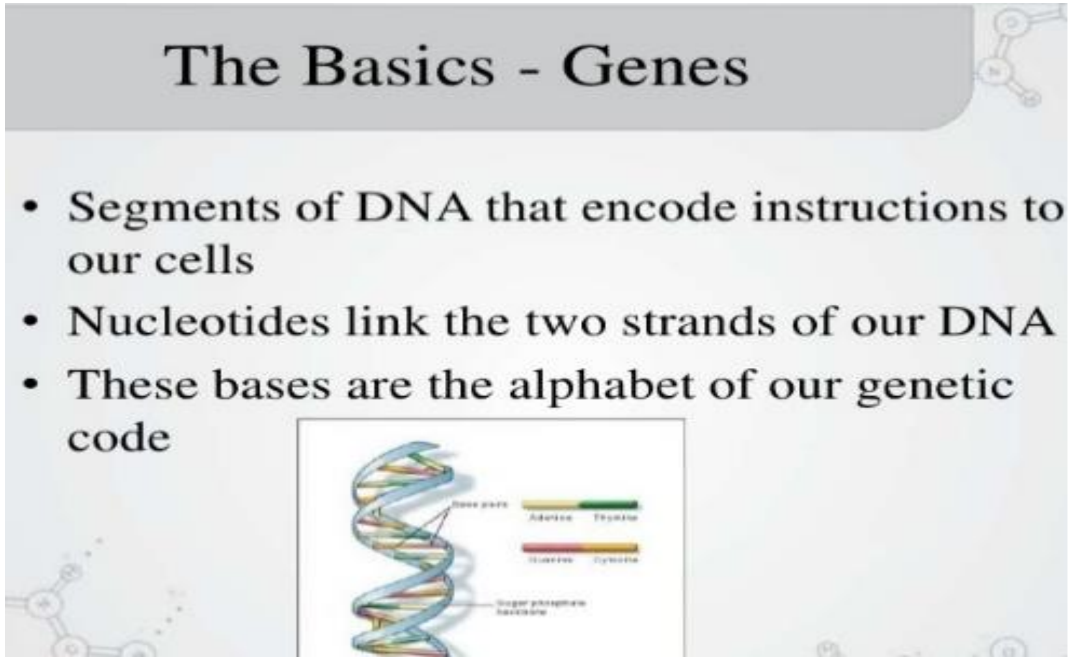
Genomic Positions of SNPs

IMPORTANT FINDING



Gene Structure

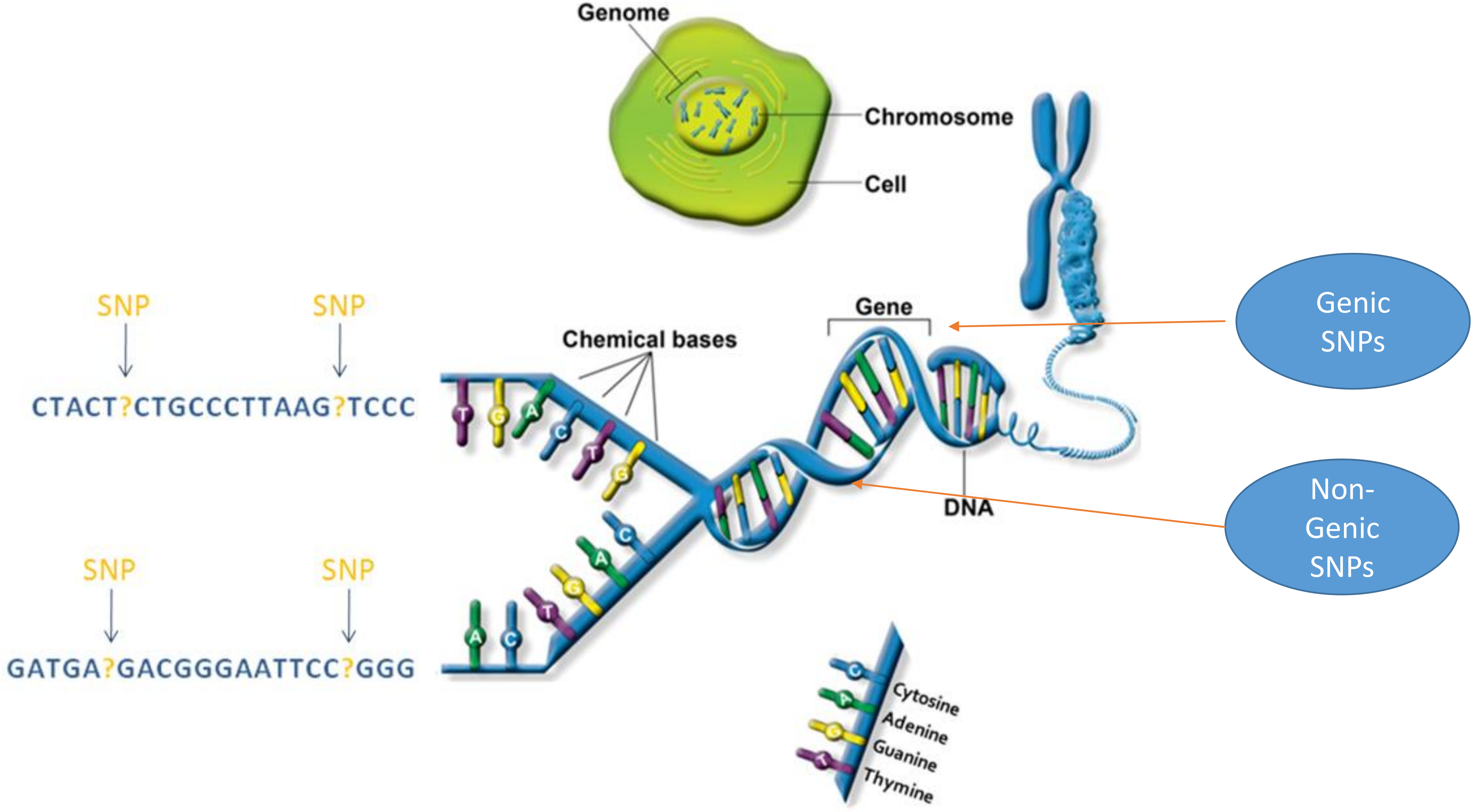




The Basics - Genes

- Segments of DNA that encode instructions to our cells
- Nucleotides link the two strands of our DNA
- These bases are the alphabet of our genetic code

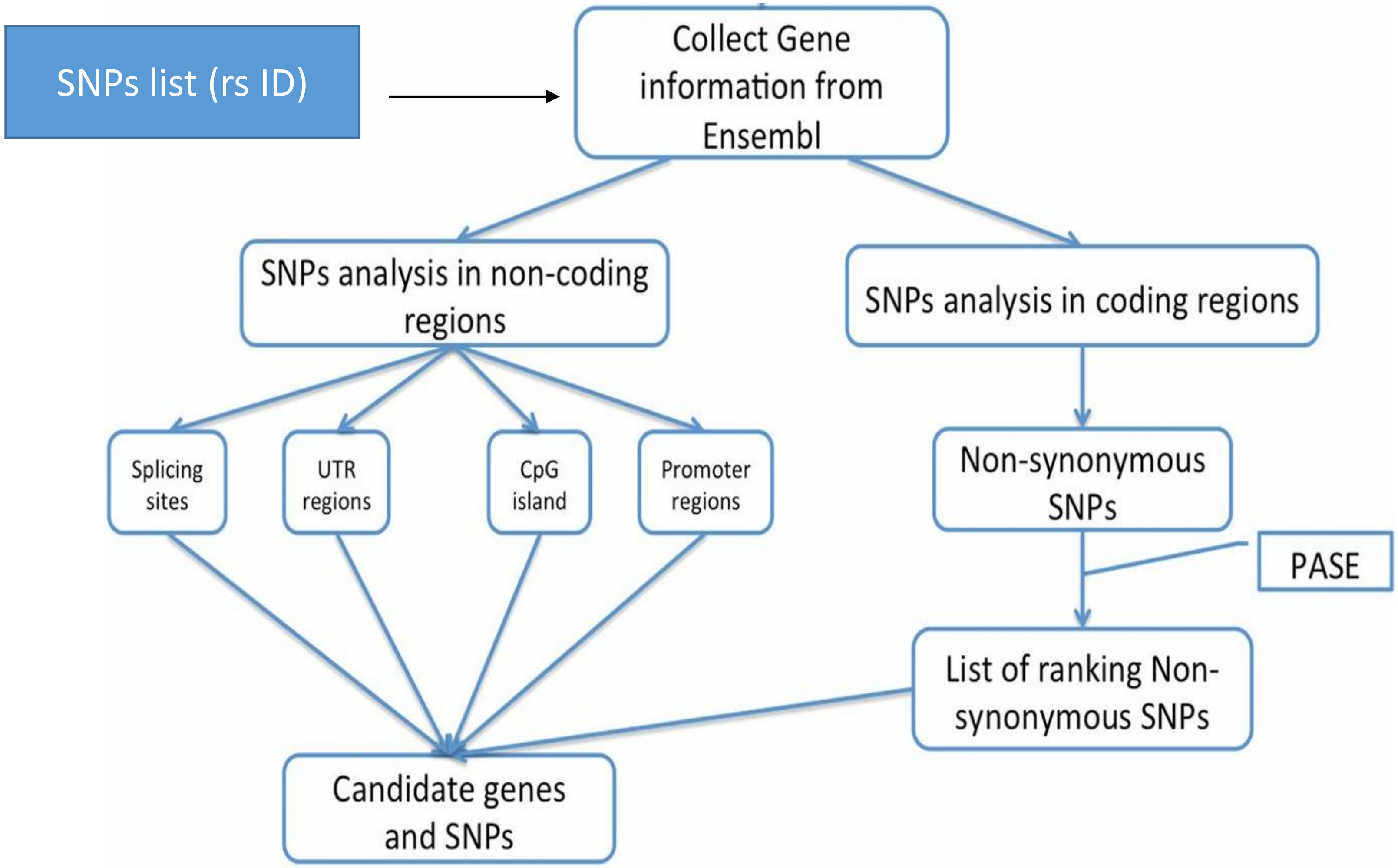
Genomic Positions of SNPs



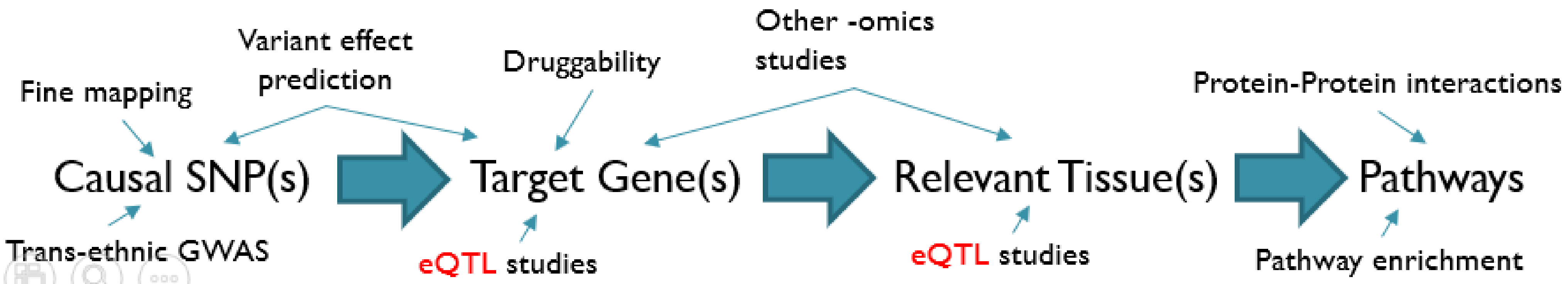
Genic SNPs

Non-Genic SNPs

Classification of SNPs (Based on Genomic Position)


SNPs list (rs ID)
Collect Gene information from Ensembl
SNPs analysis in non-coding regions
SNPs analysis in coding regions
Splicing sites
UTR regions
CpG island
Promoter regions
Non-synonymous SNPs
PASE
List of ranking Non-synonymous SNPs
Candidate genes and SNPs

Why : From SNPs to Genes



Causal SNP(s) → Target Gene(s) → Relevant Tissue(s) → Pathways

Fine mapping

Trans-ethnic GWAS

Variant effect prediction

Druggability

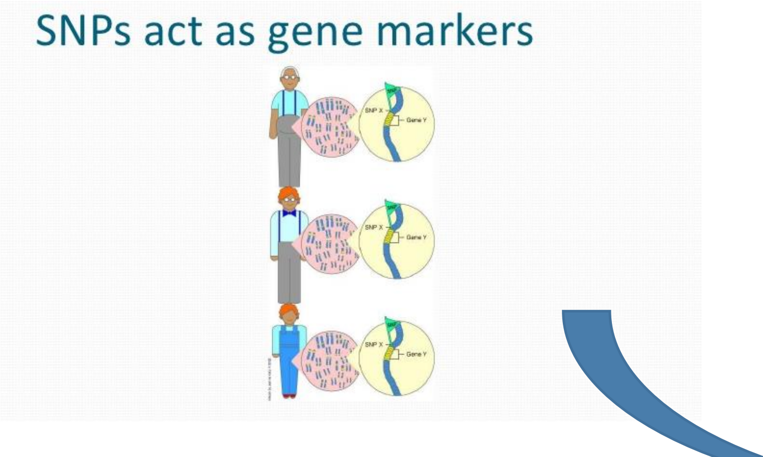
eQTL studies

Other -omics studies

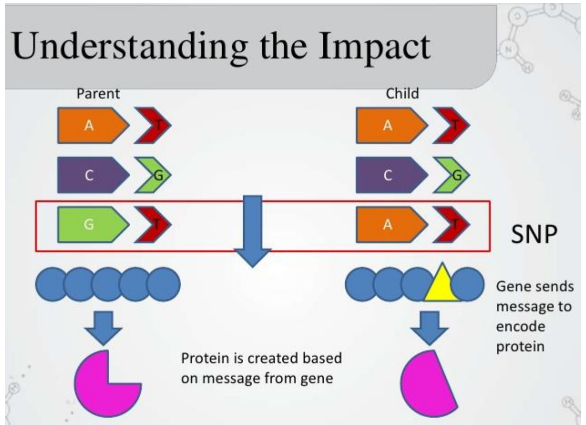
eQTL studies

Protein-Protein interactions

Pathway enrichment



SNPs act as gene markers



Understanding the Impact

Gene sends message to encode protein

Protein is created based on message from gene

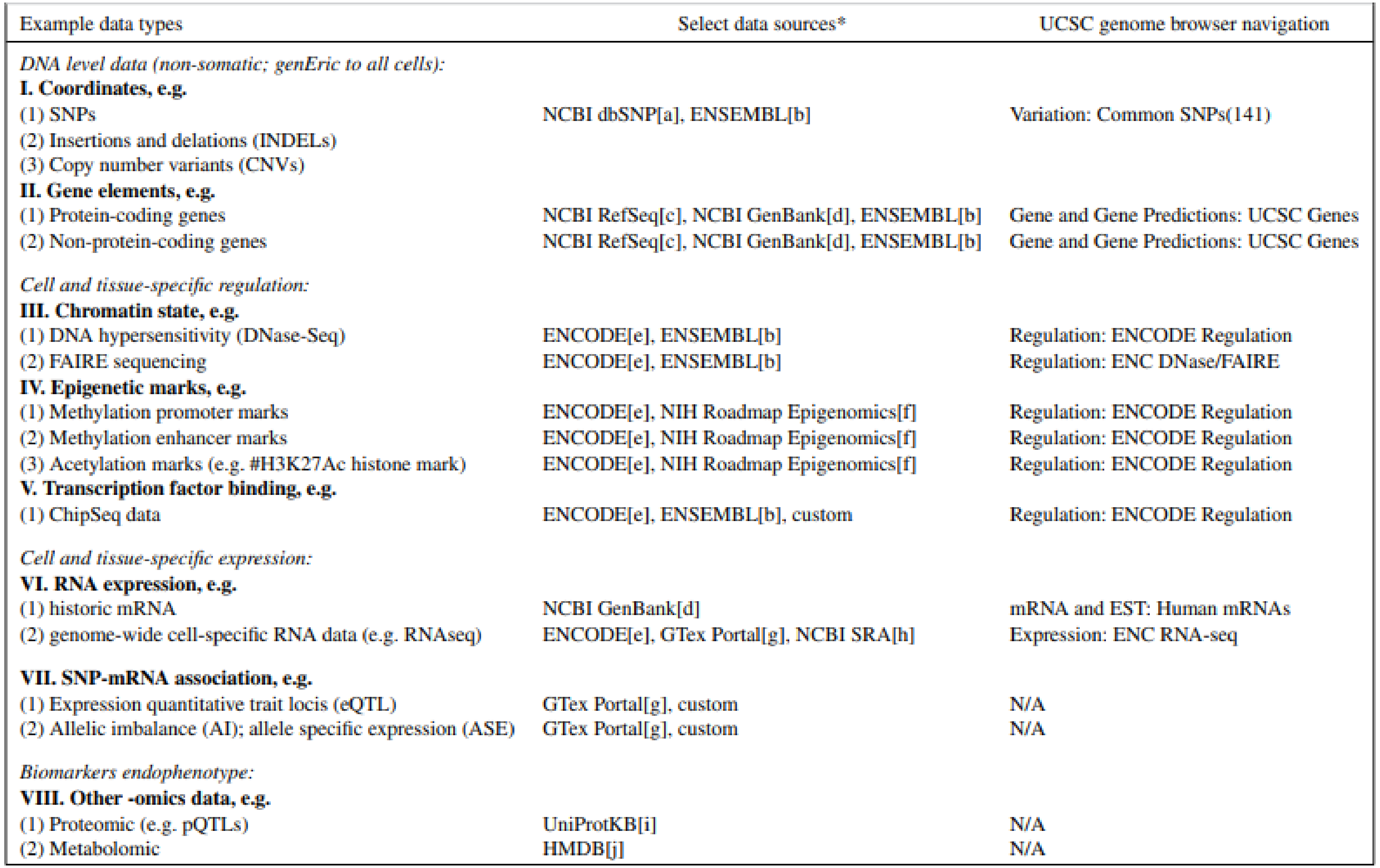
Examples: From SNPs to Genes

- **rs6311 and rs6313 are SNPs in the Serotonin 5-HT2A receptor gene on human chromosome 13.**

- **rs3091244 is an example of a triallelic SNP in the CRP gene on human chromosome 1.**

- **rs148649884 and rs138055828 in the FCN1 gene encoding M-ficolin crippled the ligand-binding capability of the recombinant M-ficolin.**

List of Data sources for Post GWAS



Example data types	Select data sources*	UCSC genome browser navigation
DNA level data (non-somatic; genEric to all cells):		
I. Coordinates, e.g.		
(1) SNPs	NCBI dbSNP[a], ENSEMBL[b]	Variation: Common SNPs(141)
(2) Insertions and delations (INDELs)		
(3) Copy number variants (CNVs)		
II. Gene elements, e.g.		
(1) Protein-coding genes	NCBI RefSeq[c], NCBI GenBank[d], ENSEMBL[b]	Gene and Gene Predictions: UCSC Genes
(2) Non-protein-coding genes	NCBI RefSeq[c], NCBI GenBank[d], ENSEMBL[b]	Gene and Gene Predictions: UCSC Genes
Cell and tissue-specific regulation:		
III. Chromatin state, e.g.		
(1) DNA hypersensitivity (DNase-Seq)	ENCODE[e], ENSEMBL[b]	Regulation: ENCODE Regulation
(2) FAIRE sequencing	ENCODE[e], ENSEMBL[b]	Regulation: ENC DNase/FAIRE
IV. Epigenetic marks, e.g.		
(1) Methylation promoter marks	ENCODE[e], NIH Roadmap Epigenomics[f]	Regulation: ENCODE Regulation
(2) Methylation enhancer marks	ENCODE[e], NIH Roadmap Epigenomics[f]	Regulation: ENCODE Regulation
(3) Acetylation marks (e.g. #H3K27Ac histone mark)	ENCODE[e], NIH Roadmap Epigenomics[f]	Regulation: ENCODE Regulation
V. Transcription factor binding, e.g.		
(1) ChipSeq data	ENCODE[e], ENSEMBL[b], custom	Regulation: ENCODE Regulation
Cell and tissue-specific expression:		
VI. RNA expression, e.g.		
(1) historic mRNA	NCBI GenBank[d]	mRNA and EST: Human mRNAs
(2) genome-wide cell-specific RNA data (e.g. RNAseq)	ENCODE[e], GTex Portal[g], NCBI SRA[h]	Expression: ENC RNA-seq
VII. SNP-mRNA association, e.g.		
(1) Expression quantitative trait locis (eQTL)	GTex Portal[g], custom	N/A
(2) Allelic imbalance (AI); allele specific expression (ASE)	GTex Portal[g], custom	N/A
Biomarkers endophenotype:		
VIII. Other -omics data, e.g.		
(1) Proteomic (e.g. pQTLs)	UniProtKB[i]	N/A
(2) Metabolomic	HMDB[j]	N/A

Post GWAS : Terminology

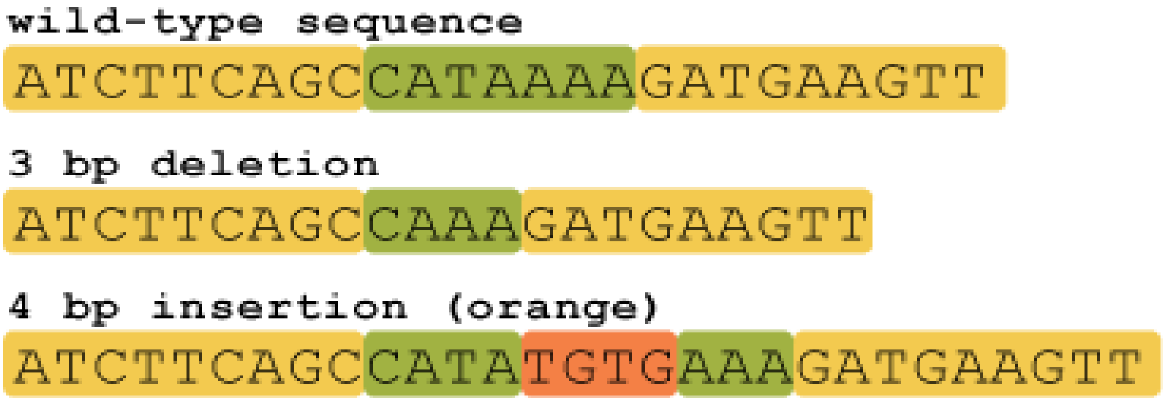
- Indels
- Epigenetic markers
- eQTL

SNPs could be linked to epigenetic markers and regulate the expression of other genes

What are indels ?

- Indels can be contrasted with a point mutation.

- An indel inserts and deletes nucleotides from a sequence, while a point mutation is a form of substitution that replaces one of the nucleotides without changing the overall number in the DNA.



wild-type sequence

ATCTTCAGCCATAAAAGATGAAGTT

3 bp deletion

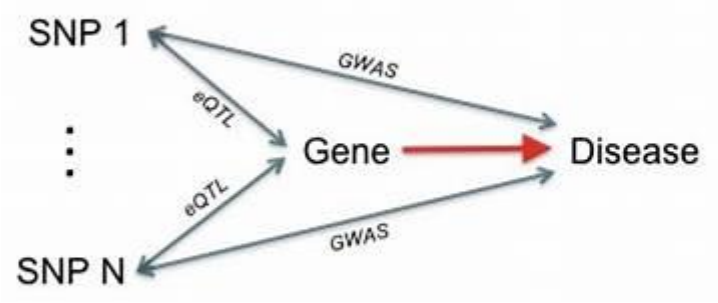
ATCTTCAGCCAAAGATGAAGTT

4 bp insertion (orange)

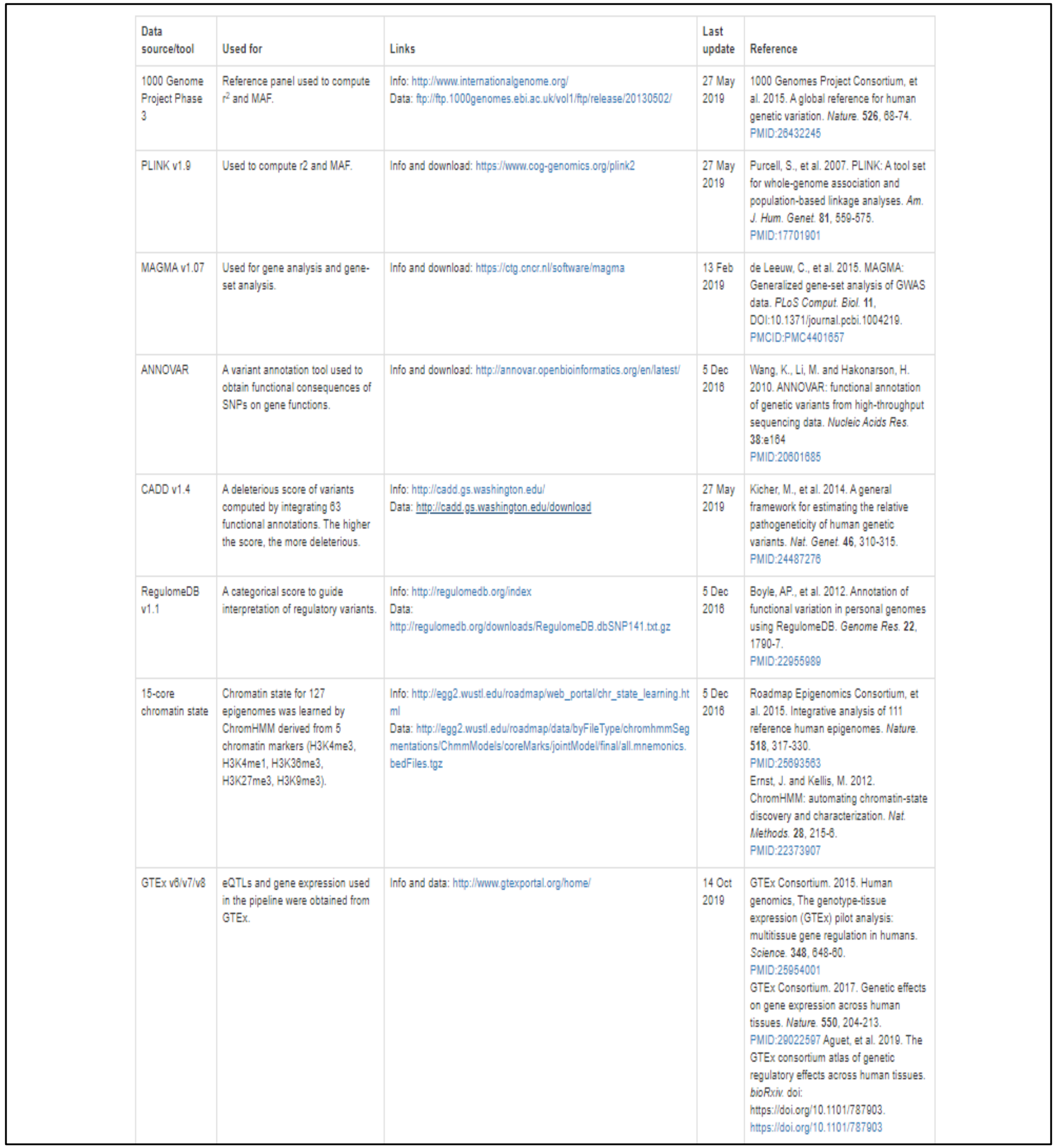
ATCTTCAGCCATATGTGAAAGATGAAGTT

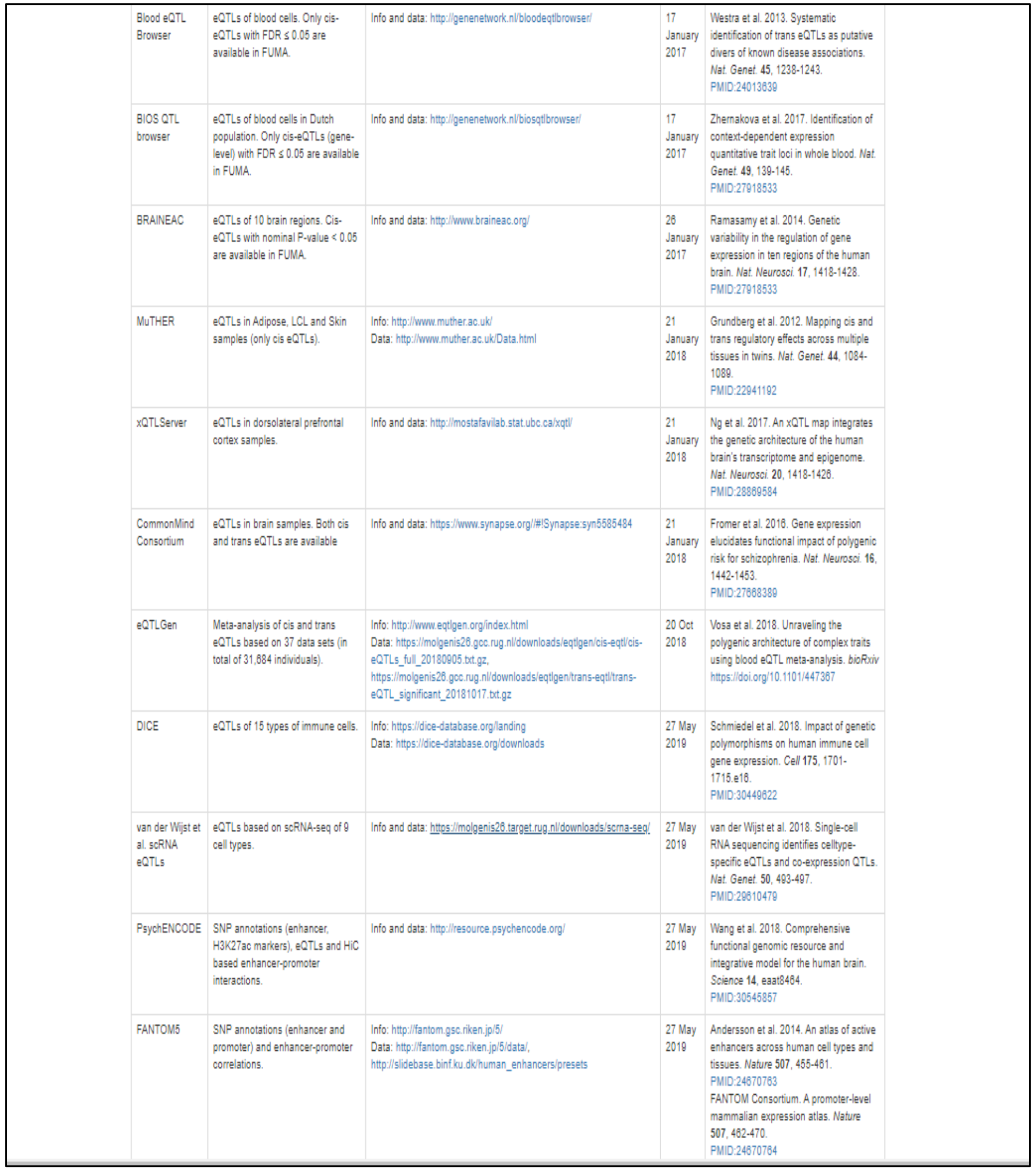
eQTL

- SNPs can be located in gene regions or intergenic ones.
- eQTL= expression Quantitative Trait Locus.
- This is a genomic locus that influences the expression level of mRNA (how much a gene is transcribed).
- This locus can be physically located close to the gene that gets regulated, or far away (even on another chromosome).



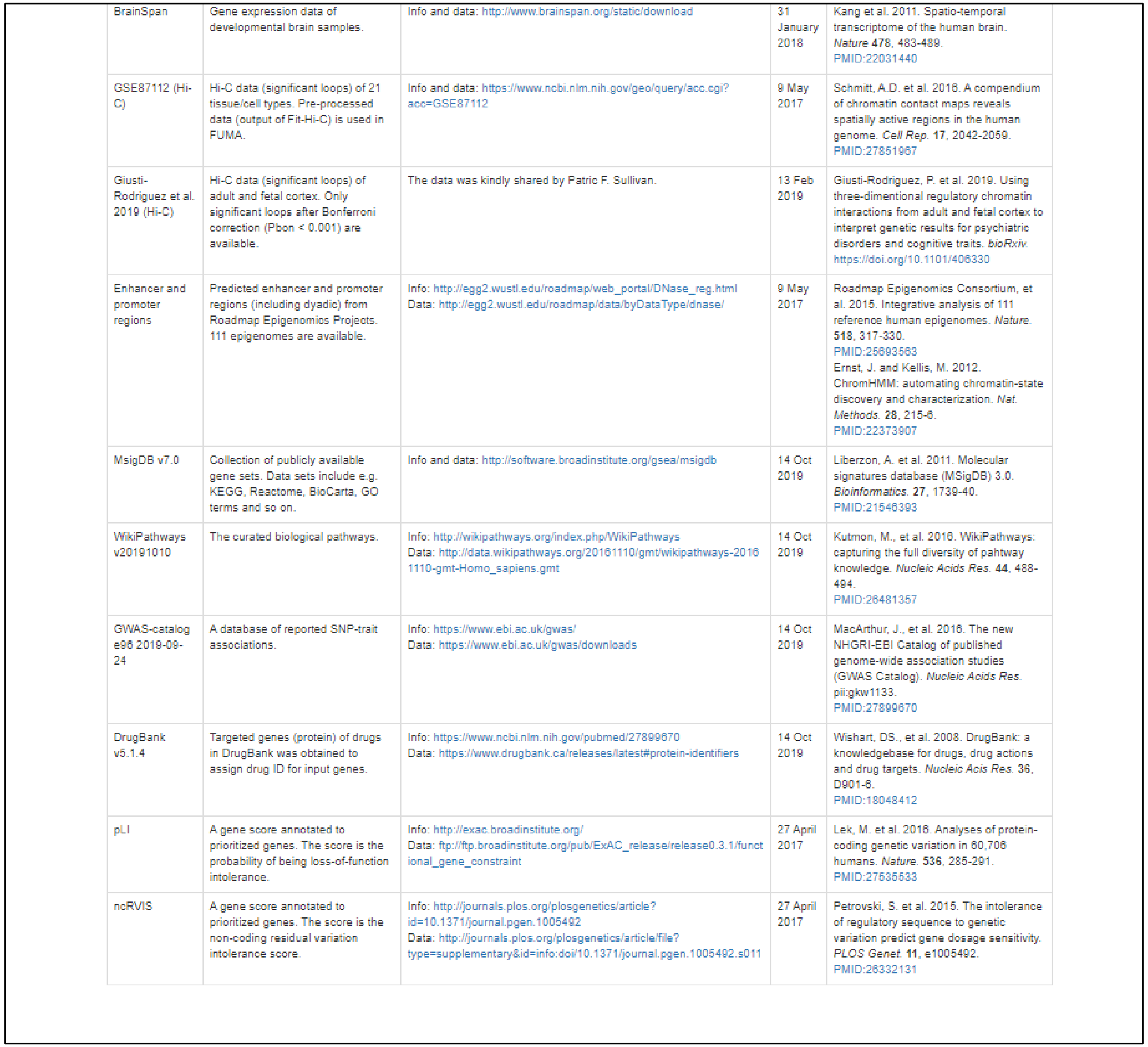
Databases and Softwares





Data source/tool	Used for	Links	Last update	Reference
1000 Genome Project Phase 3	Reference panel used to compute r^2 and MAF.	Info: http://www.internationalgenome.org/ Data: ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/	27 May 2019	1000 Genomes Project Consortium, et al. 2015. A global reference for human genetic variation. *Nature*. **526**, 68-74. PMID:26432245
PLINK v1.9	Used to compute r2 and MAF.	Info and download: https://www.cog-genomics.org/plink2	27 May 2019	Purcell, S., et al. 2007. PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559-575. PMID:17701901
MAGMA v1.07	Used for gene analysis and gene-set analysis.	Info and download: https://ctg.cncr.nl/software/magma	13 Feb 2019	de Leeuw, C., et al. 2015. MAGMA: Generalized gene-set analysis of GWAS data. *PLoS Comput. Biol.* **11**, DOI:10.1371/journal.pcbi.1004219. PMCID:PMC4401657
ANNOVAR	A variant annotation tool used to obtain functional consequences of SNPs on gene functions.	Info and download: http://annovar.openbioinformatics.org/en/latest/	5 Dec 2016	Wang, K., Li, M. and Hakonarson, H. 2010. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**:e164 PMID:20601685
CADD v1.4	A deleterious score of variants computed by integrating 63 functional annotations. The higher the score, the more deleterious.	Info: http://cadd.gs.washington.edu/ Data: http://cadd.gs.washington.edu/download	27 May 2019	Kicher, M., et al. 2014. A general framework for estimating the relative pathogeneticity of human genetic variants. *Nat. Genet.* **46**, 310-315. PMID:24487276
RegulomeDB v1.1	A categorical score to guide interpretation of regulatory variants.	Info: http://regulomedb.org/index Data: http://regulomedb.org/downloads/RegulomeDB.dbSNP141.txt.gz	5 Dec 2016	Boyle, AP., et al. 2012. Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res.* **22**, 1790-7. PMID:22955989
15-core chromatin state	Chromatin state for 127 epigenomes was learned by ChromHMM derived from 5 chromatin markers (H3K4me3, H3K4me1, H3K36me3, H3K27me3, H3K9me3).	Info: http://egg2.wustl.edu/roadmap/web_portal/chr_state_learning.html Data: http://egg2.wustl.edu/roadmap/data/byFileType/chromhmmSegmentations/ChmmModels/coreMarks/jointModel/final/all.mnemonics.bedFiles.tgz	5 Dec 2016	Roadmap Epigenomics Consortium, et al. 2015. Integrative analysis of 111 reference human epigenomes. *Nature*. **518**, 317-330. PMID:25693563 Ernst, J. and Kellis, M. 2012. ChromHMM: automating chromatin-state discovery and characterization. *Nat. Methods*. **28**, 215-6. PMID:22373907
GTEx v6/v7/v8	eQTLs and gene expression used in the pipeline were obtained from GTEx.	Info and data: http://www.gtexportal.org/home/	14 Oct 2019	GTEx Consortium. 2015. Human genomics, The genotype-tissue expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science*. **348**, 648-60. PMID:25954001 GTEx Consortium. 2017. Genetic effects on gene expression across human tissues. *Nature*. **550**, 204-213. PMID:29022597 Aguet, et al. 2019. The GTEx consortium atlas of genetic regulatory effects across human tissues. *bioRxiv*. doi: https://doi.org/10.1101/787903. https://doi.org/10.1101/787903
Blood eQTL Browser	eQTLs of blood cells. Only cis-eQTLs with FDR ≤ 0.05 are available in FUMA.	Info and data: http://genenetwork.nl/bloodeqtlbrowser/	17 January 2017	Westra et al. 2013. Systematic identification of trans eQTLs as putative divers of known disease associations. *Nat. Genet.* **45**, 1238-1243. PMID:24013639
BIOS QTL browser	eQTLs of blood cells in Dutch population. Only cis-eQTLs (gene-level) with FDR ≤ 0.05 are available in FUMA.	Info and data: http://genenetwork.nl/biosqtlbrowser/	17 January 2017	Zhernakova et al. 2017. Identification of context-dependent expression quantitative trait loci in whole blood. *Nat. Genet.* **49**, 139-145. PMID:27918533
BRAINEAC	eQTLs of 10 brain regions. Cis-eQTLs with nominal P-value < 0.05 are available in FUMA.	Info and data: http://www.braineac.org/	26 January 2017	Ramasamy et al. 2014. Genetic variability in the regulation of gene expression in ten regions of the human brain. *Nat. Neurosci.* **17**, 1418-1428. PMID:27918533
MuTHER	eQTLs in Adipose, LCL and Skin samples (only cis eQTLs).	Info: http://www.muther.ac.uk/ Data: http://www.muther.ac.uk/Data.html	21 January 2018	Grundberg et al. 2012. Mapping cis and trans regulatory effects across multiple tissues in twins. *Nat. Genet.* **44**, 1084-1089. PMID:22941192
xQTLServer	eQTLs in dorsolateral prefrontal cortex samples.	Info and data: http://mostafavilab.stat.ubc.ca/xqtl/	21 January 2018	Ng et al. 2017. An xQTL map integrates the genetic architecture of the human brain's transcriptome and epigenome. *Nat. Neurosci.* **20**, 1418-1426. PMID:28869584
CommonMind Consortium	eQTLs in brain samples. Both cis and trans eQTLs are available	Info and data: https://www.synapse.org//#!Synapse:syn5585484	21 January 2018	Fromer et al. 2016. Gene expression elucidates functional impact of polygenic risk for schizophrenia. *Nat. Neurosci.* **16**, 1442-1453. PMID:27668389
eQTLGen	Meta-analysis of cis and trans eQTLs based on 37 data sets (in total of 31,684 individuals).	Info: http://www.eqtlgen.org/index.html Data: https://molgenis26.gcc.rug.nl/downloads/eqtlgen/cis-eqtl/cis-eQTLs_full_20180905.txt.gz, https://molgenis26.gcc.rug.nl/downloads/eqtlgen/trans-eqtl/trans-eQTL_significant_20181017.txt.gz	20 Oct 2018	Vosa et al. 2018. Unraveling the polygenic architecture of complex traits using blood eQTL meta-analysis. *bioRxiv* https://doi.org/10.1101/447367
DICE	eQTLs of 15 types of immune cells.	Info: https://dice-database.org/landing Data: https://dice-database.org/downloads	27 May 2019	Schmiedel et al. 2018. Impact of genetic polymorphisms on human immune cell gene expression. *Cell* **175**, 1701-1715.e16. PMID:30449622
van der Wijst et al. scRNA eQTLs	eQTLs based on scRNA-seq of 9 cell types.	Info and data: https://molgenis26.target.rug.nl/downloads/scrna-seq/	27 May 2019	van der Wijst et al. 2018. Single-cell RNA sequencing identifies celltype-specific eQTLs and co-expression QTLs. *Nat. Genet.* **50**, 493-497. PMID:29610479
PsychENCODE	SNP annotations (enhancer, H3K27ac markers), eQTLs and HiC based enhancer-promoter interactions.	Info and data: http://resource.psychencode.org/	27 May 2019	Wang et al. 2018. Comprehensive functional genomic resource and integrative model for the human brain. *Science* **14**, eaat8464. PMID:30545857
FANTOM5	SNP annotations (enhancer and promoter) and enhancer-promoter correlations.	Info: http://fantom.gsc.riken.jp/5/ Data: http://fantom.gsc.riken.jp/5/data/, http://slidebase.binf.ku.dk/human_enhancers/presets	27 May 2019	Andersson et al. 2014. An atlas of active enhancers across human cell types and tissues. *Nature* **507**, 455-461. PMID:24670763 FANTOM Consortium. A promoter-level mammalian expression atlas. *Nature* **507**, 462-470. PMID:24670764

Databases and Softwares



BrainSpan	Gene expression data of developmental brain samples.	Info and data: http://www.brainspan.org/static/download	31 January 2018	Kang et al. 2011. Spatio-temporal transcriptome of the human brain. *Nature* **478**, 483-489. PMID:22031440
GSE87112 (Hi-C)	Hi-C data (significant loops) of 21 tissue/cell types. Pre-processed data (output of Fit-Hi-C) is used in FUMA.	Info and data: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE87112	9 May 2017	Schmitt, A.D. et al. 2016. A compendium of chromatin contact maps reveals spatially active regions in the human genome. *Cell Rep.* **17**, 2042-2059. PMID:27851967
Giusti-Rodriguez et al. 2019 (Hi-C)	Hi-C data (significant loops) of adult and fetal cortex. Only significant loops after Bonferroni correction (Pbon < 0.001) are available.	The data was kindly shared by Patric F. Sullivan.	13 Feb 2019	Giusti-Rodriguez, P. et al. 2019. Using three-dimentional regulatory chromatin interactions from adult and fetal cortex to interpret genetic results for psychiatric disorders and cognitive traits. *bioRxiv.* https://doi.org/10.1101/406330
Enhancer and promoter regions	Predicted enhancer and promoter regions (including dyadic) from Roadmap Epigenomics Projects. 111 epigenomes are available.	Info: http://egg2.wustl.edu/roadmap/web_portal/DNase_reg.html Data: http://egg2.wustl.edu/roadmap/data/byDataType/dnase/	9 May 2017	Roadmap Epigenomics Consortium, et al. 2015. Integrative analysis of 111 reference human epigenomes. *Nature.* **518**, 317-330. PMID:25693563 Ernst, J. and Kellis, M. 2012. ChromHMM: automating chromatin-state discovery and characterization. *Nat. Methods.* **28**, 215-6. PMID:22373907
MsigDB v7.0	Collection of publicly available gene sets. Data sets include e.g. KEGG, Reactome, BioCarta, GO terms and so on.	Info and data: http://software.broadinstitute.org/gsea/msigdb	14 Oct 2019	Liberzon, A. et al. 2011. Molecular signatures database (MSigDB) 3.0. *Bioinformatics.* **27**, 1739-40. PMID:21546393
WikiPathways v20191010	The curated biological pathways.	Info: http://wikipathways.org/index.php/WikiPathways Data: http://data.wikipathways.org/20161110/gmt/wikipathways-20161110-gmt-Homo_sapiens.gmt	14 Oct 2019	Kutmon, M., et al. 2016. WikiPathways: capturing the full diversity of pahtway knowledge. *Nucleic Acids Res.* **44**, 488-494. PMID:26481357
GWAS-catalog e96 2019-09-24	A database of reported SNP-trait associations.	Info: https://www.ebi.ac.uk/gwas/ Data: https://www.ebi.ac.uk/gwas/downloads	14 Oct 2019	MacArthur, J., et al. 2016. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res.* pii:gkw1133. PMID:27899670
DrugBank v5.1.4	Targeted genes (protein) of drugs in DrugBank was obtained to assign drug ID for input genes.	Info: https://www.ncbi.nlm.nih.gov/pubmed/27899670 Data: https://www.drugbank.ca/releases/latest#protein-identifiers	14 Oct 2019	Wishart, DS., et al. 2008. DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acis Res.* **36**, D901-6. PMID:18048412
pLI	A gene score annotated to prioritized genes. The score is the probability of being loss-of-function intolerance.	Info: http://exac.broadinstitute.org/ Data: ftp://ftp.broadinstitute.org/pub/ExAC_release/release0.3.1/functional_gene_constraint	27 April 2017	Lek, M. et al. 2016. Analyses of protein-coding genetic variation in 60,706 humans. *Nature.* **536**, 285-291. PMID:27535533
ncRVIS	A gene score annotated to prioritized genes. The score is the non-coding residual variation intolerance score.	Info: http://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1005492 Data: http://journals.plos.org/plosgenetics/article/file?type=supplementary&id=info:doi/10.1371/journal.pgen.1005492.s011	27 April 2017	Petrovski, S. et al. 2015. The intolerance of regulatory sequence to genetic variation predict gene dosage sensitivity. *PLOS Genet.* **11**, e1005492. PMID:26332131

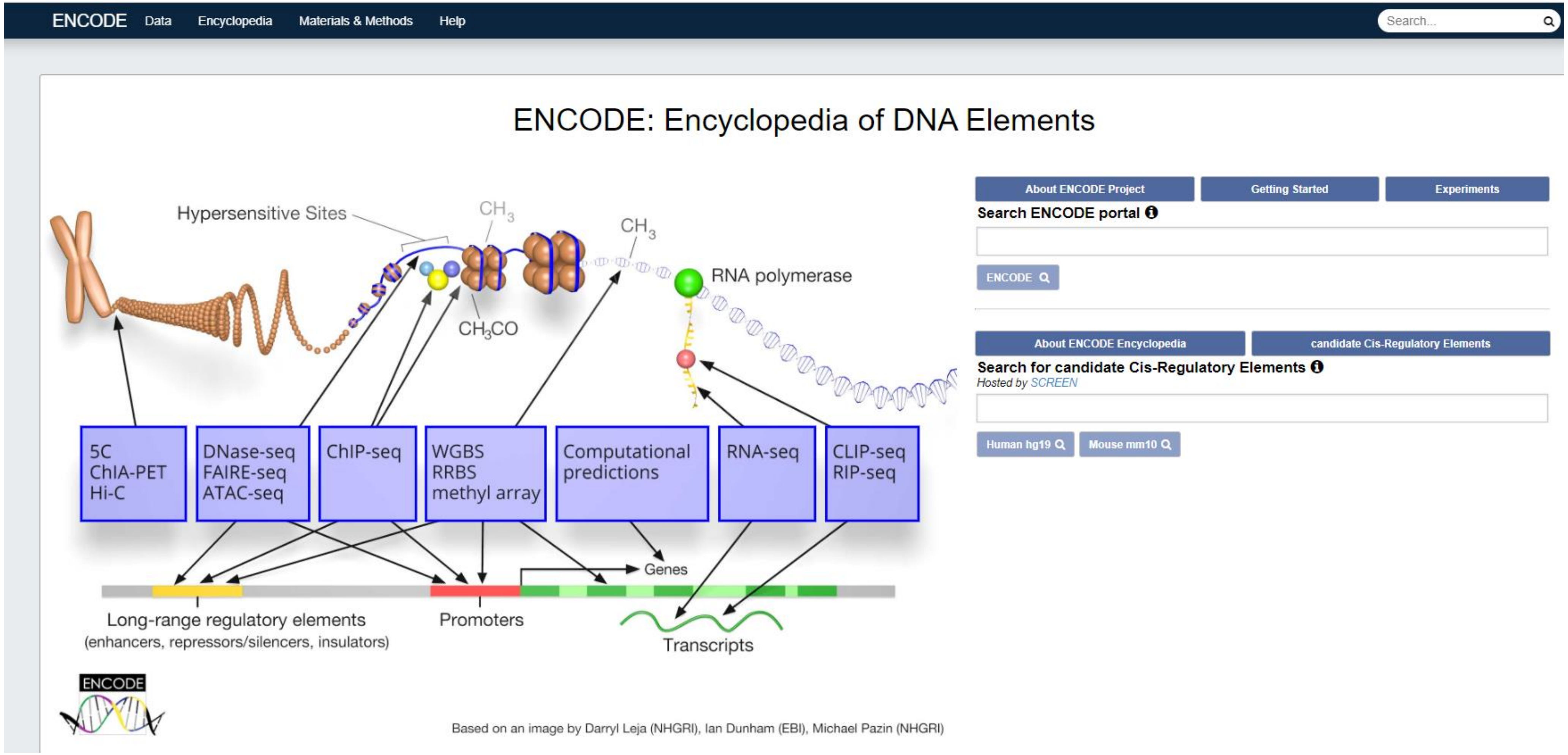
Data bases and web servers

Let us discuss :

- **ENCODE**
- **HelgoDB**
- **RegulomeDB**
- **UniprotKB**
- **ENSEMBL**
- **FUMA**

ENCODE: Encyclopedia of DNA Elements

https://www.encodeproject.org/



ENCODE: Encyclopedia of DNA Elements

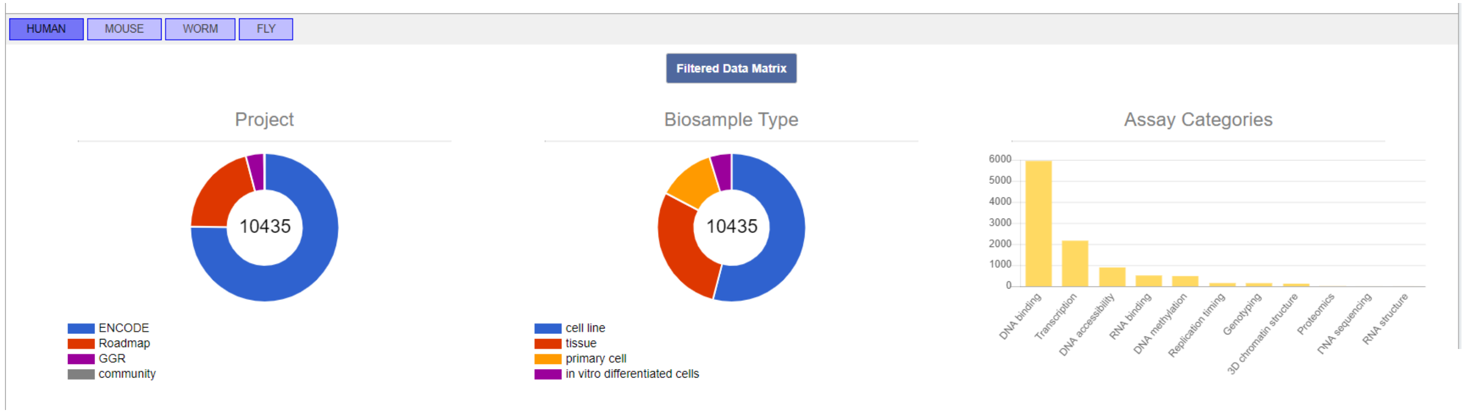
Based on an image by Darryl Leja (NHGRI), Ian Dunham (EBI), Michael Pazin (NHGRI)

Search ENCODE portal

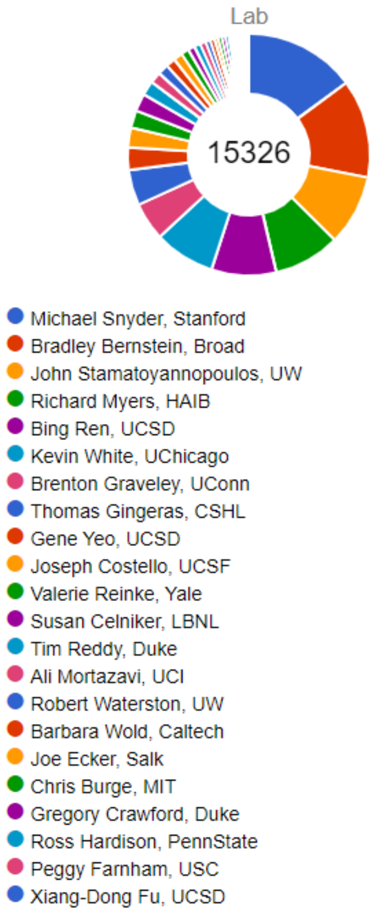
Search for candidate Cis-Regulatory Elements

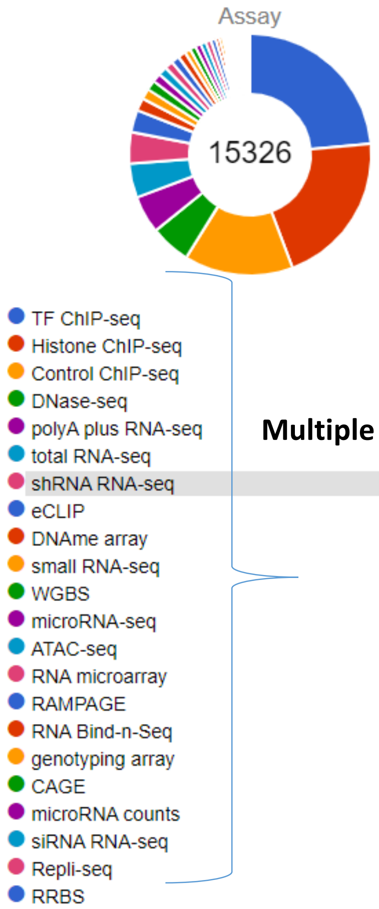
Hosted by *SCREEN*

Encode : Data structures



Most of data represents





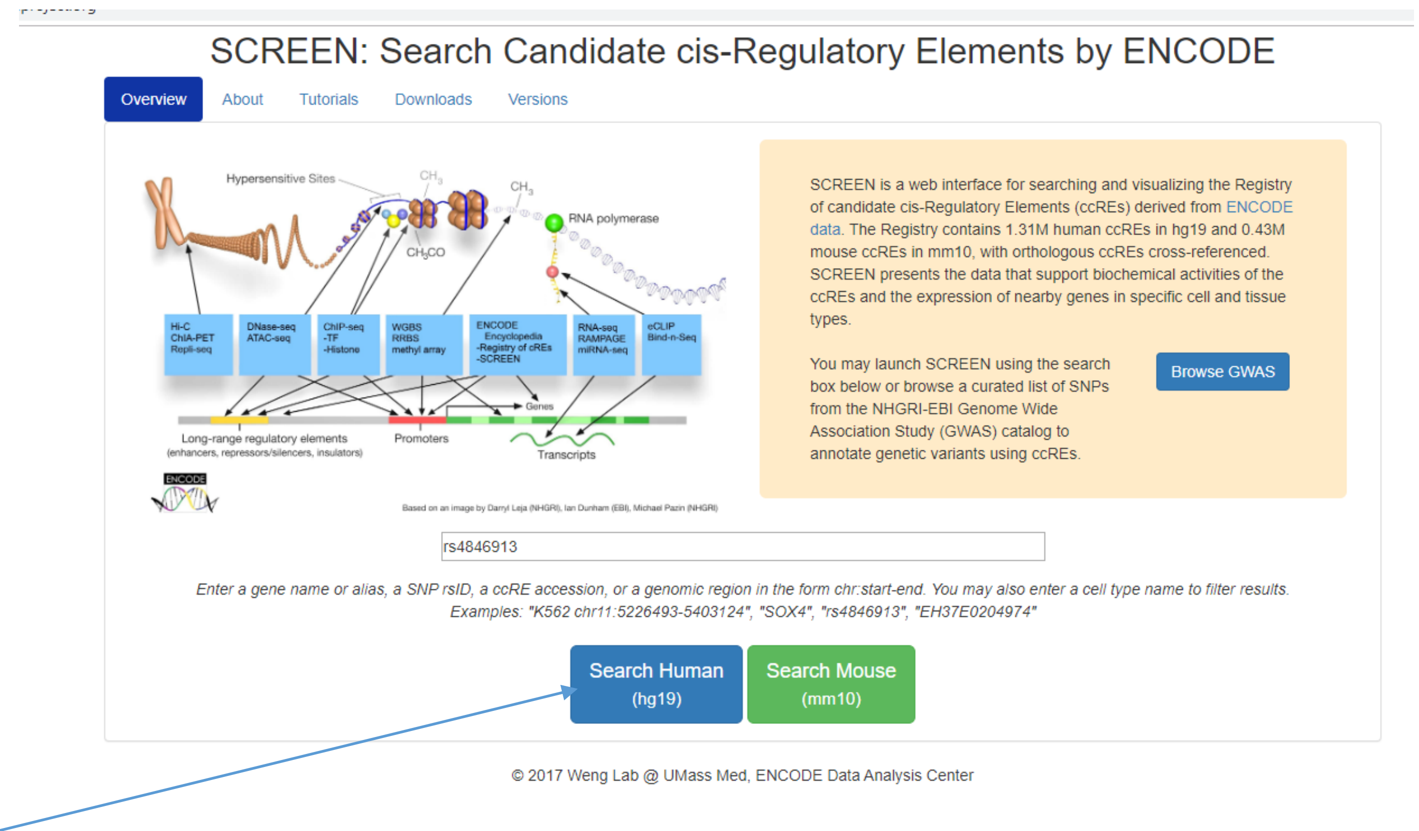
Multiple platform

Multiple resources

Let us use Encode

Go to link http://screen.encodeproject.org/

Enter snp id : rs4846913



SCREEN: Search Candidate cis-Regulatory Elements by ENCODE

SCREEN is a web interface for searching and visualizing the Registry of candidate cis-Regulatory Elements (ccREs) derived from ENCODE data. The Registry contains 1.31M human ccREs in hg19 and 0.43M mouse ccREs in mm10, with orthologous ccREs cross-referenced. SCREEN presents the data that support biochemical activities of the ccREs and the expression of nearby genes in specific cell and tissue types.

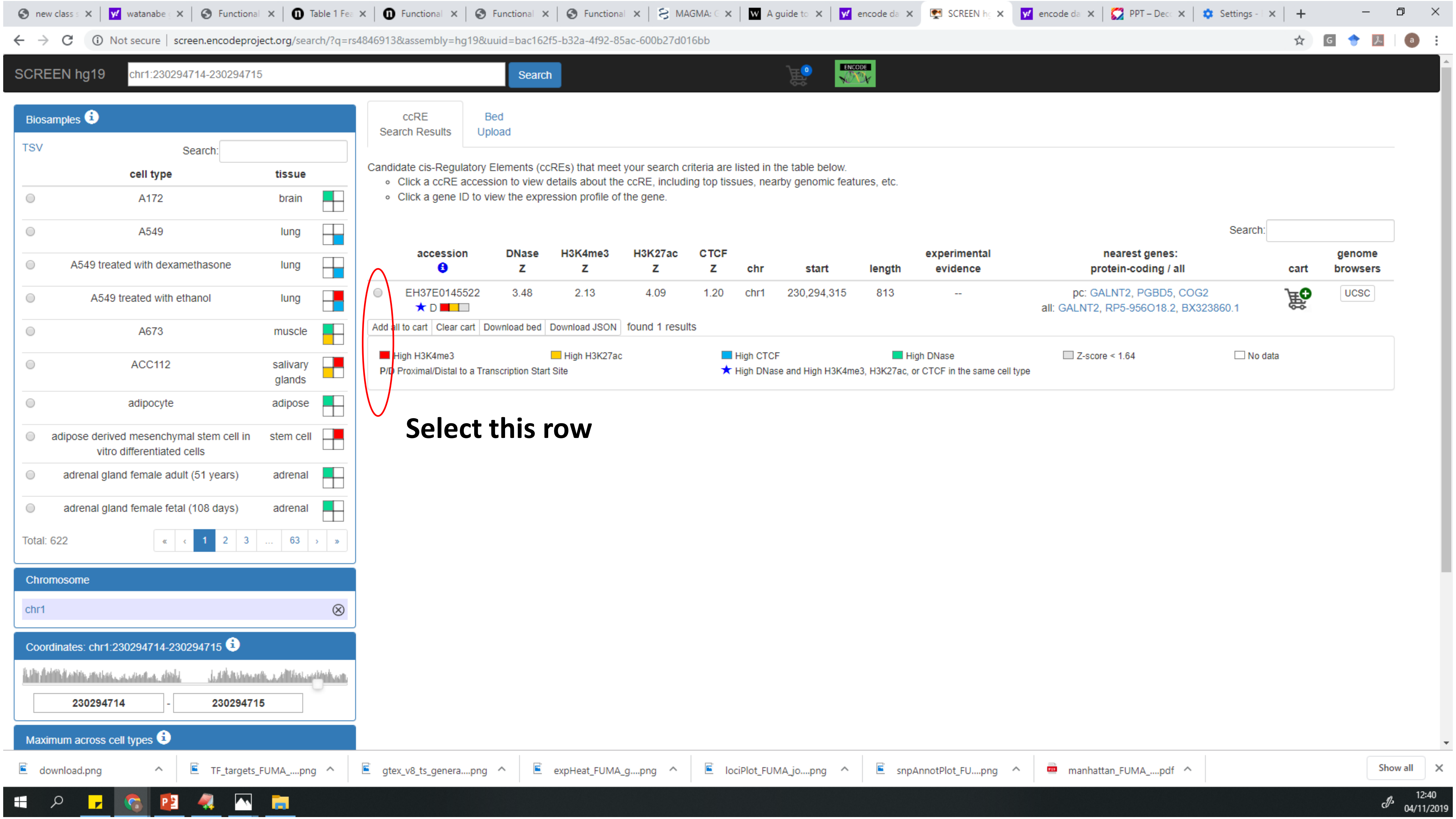
You may launch SCREEN using the search box below or browse a curated list of SNPs from the NHGRI-EBI Genome Wide Association Study (GWAS) catalog to annotate genetic variants using ccREs.

rs4846913

Enter a gene name or alias, a SNP rsID, a ccRE accession, or a genomic region in the form chr:start-end. You may also enter a cell type name to filter results. Examples: "K562 chr11:5226493-5403124", "SOX4", "rs4846913", "EH37E0204974"

© 2017 Weng Lab @ UMass Med, ENCODE Data Analysis Center

Click



Biosamples

TSV Search:

	cell type	tissue
○	A172	brain
○	A549	lung
○	A549 treated with dexamethasone	lung
○	A549 treated with ethanol	lung
○	A673	muscle
○	ACC112	salivary glands
○	adipocyte	adipose
○	adipose derived mesenchymal stem cell in vitro differentiated cells	stem cell
○	adrenal gland female adult (51 years)	adrenal
○	adrenal gland female fetal (108 days)	adrenal

Total: 622 « ‹ 1 2 3 ... 63 › »

Chromosome

chr1

Coordinates: chr1:230294714-230294715

230294714 - 230294715

Maximum across cell types

ccRE Search Results | Bed Upload

Candidate cis-Regulatory Elements (ccREs) that meet your search criteria are listed in the table below.

- Click a ccRE accession to view details about the ccRE, including top tissues, nearby genomic features, etc.
- Click a gene ID to view the expression profile of the gene.

Search:

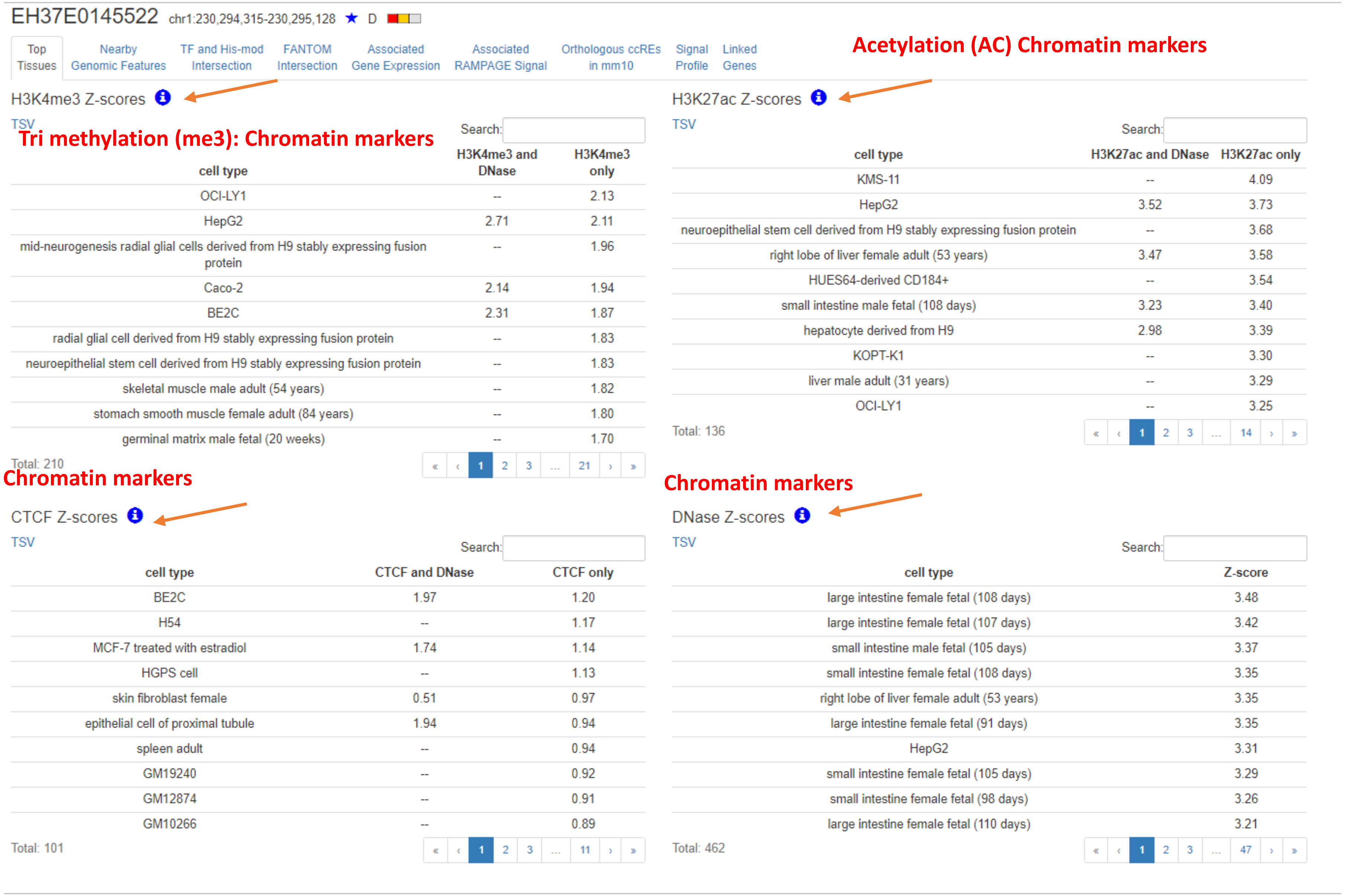
	accession	DNase Z	H3K4me3 Z	H3K27ac Z	CTCF Z	chr	start	length	experimental evidence	nearest genes: protein-coding / all	cart	genome browsers
○	EH37E0145522 ★ D	3.48	2.13	4.09	1.20	chr1	230,294,315	813	--	pc: GALNT2, PGBD5, COG2 all: GALNT2, RP5-956O18.2, BX323860.1		UCSC

Add all to cart | Clear cart | Download bed | Download JSON | found 1 results

🟥 High H3K4me3 🟨 High H3K27ac 🟦 High CTCF 🟩 High DNase ▨ Z-score < 1.64 □ No data
P/D Proximal/Distal to a Transcription Start Site ★ High DNase and High H3K4me3, H3K27ac, or CTCF in the same cell type

Select this row

Z score data from multiple chromatin markers in different cell types



EH37E0145522 chr1:230,294,315-230,295,128 ★ D

Top Tissues
Nearby Genomic Features
TF and His-mod Intersection
FANTOM Intersection
Associated Gene Expression
Associated RAMPAGE Signal
Orthologous ccREs in mm10
Signal Profile
Linked Genes

Acetylation (AC) Chromatin markers

Tri methylation (me3): Chromatin markers

H3K4me3 Z-scores

TSV

Search:

cell type	H3K4me3 and DNase	H3K4me3 only
OCI-LY1	--	2.13
HepG2	2.71	2.11
mid-neurogenesis radial glial cells derived from H9 stably expressing fusion protein	--	1.96
Caco-2	2.14	1.94
BE2C	2.31	1.87
radial glial cell derived from H9 stably expressing fusion protein	--	1.83
neuroepithelial stem cell derived from H9 stably expressing fusion protein	--	1.83
skeletal muscle male adult (54 years)	--	1.82
stomach smooth muscle female adult (84 years)	--	1.80
germinal matrix male fetal (20 weeks)	--	1.70

Total: 210

« ‹ 1 2 3 … 21 › »

H3K27ac Z-scores

TSV

Search:

cell type	H3K27ac and DNase	H3K27ac only
KMS-11	--	4.09
HepG2	3.52	3.73
neuroepithelial stem cell derived from H9 stably expressing fusion protein	--	3.68
right lobe of liver female adult (53 years)	3.47	3.58
HUES64-derived CD184+	--	3.54
small intestine male fetal (108 days)	3.23	3.40
hepatocyte derived from H9	2.98	3.39
KOPT-K1	--	3.30
liver male adult (31 years)	--	3.29
OCI-LY1	--	3.25

Total: 136

« ‹ 1 2 3 … 14 › »

Chromatin markers

CTCF Z-scores

TSV

Search:

cell type	CTCF and DNase	CTCF only
BE2C	1.97	1.20
H54	--	1.17
MCF-7 treated with estradiol	1.74	1.14
HGPS cell	--	1.13
skin fibroblast female	0.51	0.97
epithelial cell of proximal tubule	1.94	0.94
spleen adult	--	0.94
GM19240	--	0.92
GM12874	--	0.91
GM10266	--	0.89

Total: 101

« ‹ 1 2 3 … 11 › »

Chromatin markers

DNase Z-scores

TSV

Search:

cell type	Z-score
large intestine female fetal (108 days)	3.48
large intestine female fetal (107 days)	3.42
small intestine male fetal (105 days)	3.37
small intestine female fetal (108 days)	3.35
right lobe of liver female adult (53 years)	3.35
large intestine female fetal (91 days)	3.35
HepG2	3.31
small intestine female fetal (105 days)	3.29
small intestine female fetal (98 days)	3.26
large intestine female fetal (110 days)	3.21

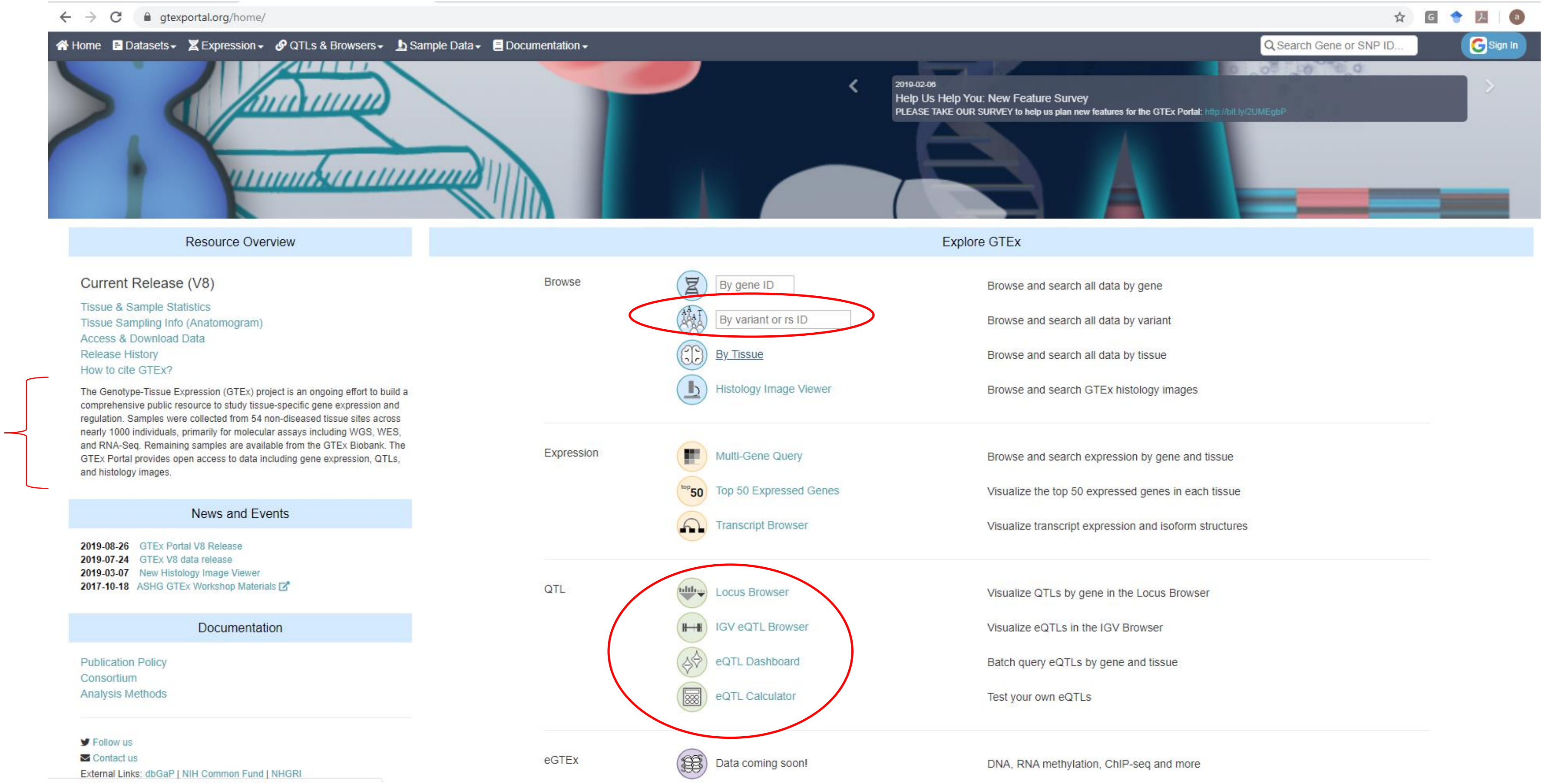
Total: 462

« ‹ 1 2 3 … 47 › »

GTEx : Genotype-Tissue Expression (GTEx)

Go to link https://gtexportal.org/home/

Enter snp id : rs712 [Homo sapiens]



2019-02-06
Help Us Help You: New Feature Survey
PLEASE TAKE OUR SURVEY to help us plan new features for the GTEx Portal: http://bit.ly/2UMEgbP

Resource Overview

Current Release (V8)

Tissue & Sample Statistics
Tissue Sampling Info (Anatomogram)
Access & Download Data
Release History
How to cite GTEx?

The Genotype-Tissue Expression (GTEx) project is an ongoing effort to build a comprehensive public resource to study tissue-specific gene expression and regulation. Samples were collected from 54 non-diseased tissue sites across nearly 1000 individuals, primarily for molecular assays including WGS, WES, and RNA-Seq. Remaining samples are available from the GTEx Biobank. The GTEx Portal provides open access to data including gene expression, QTLs, and histology images.

News and Events

2019-08-26	GTEx Portal V8 Release
2019-07-24	GTEx V8 data release
2019-03-07	New Histology Image Viewer
2017-10-18	ASHG GTEx Workshop Materials

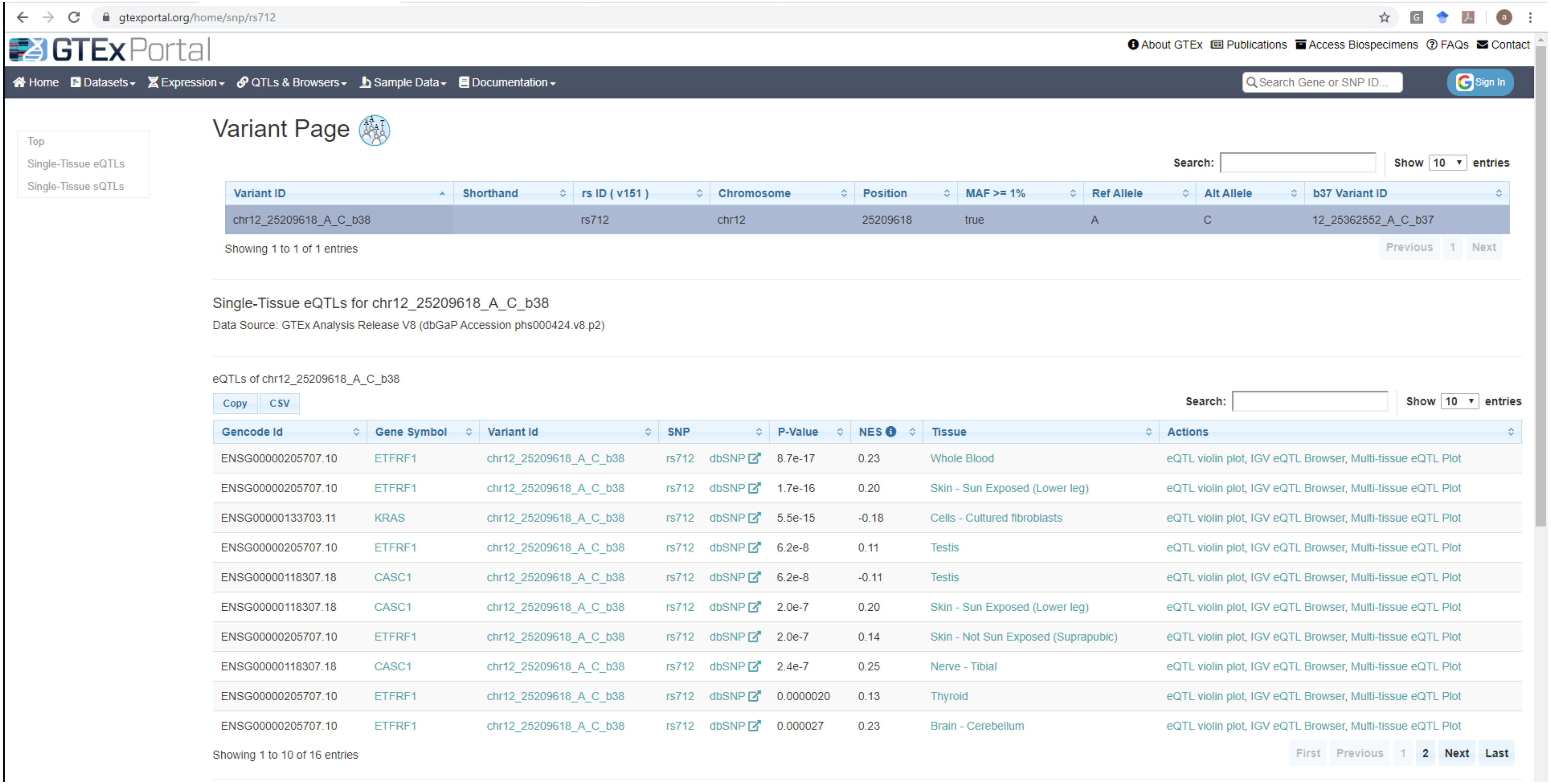
Documentation

Publication Policy
Consortium
Analysis Methods

Follow us
Contact us
External Links: dbGaP | NIH Common Fund | NHGRI

Explore GTEx

Browse	By gene ID	Browse and search all data by gene
	By variant or rs ID	Browse and search all data by variant
	By Tissue	Browse and search all data by tissue
	Histology Image Viewer	Browse and search GTEx histology images
Expression	Multi-Gene Query	Browse and search expression by gene and tissue
	Top 50 Expressed Genes	Visualize the top 50 expressed genes in each tissue
	Transcript Browser	Visualize transcript expression and isoform structures
QTL	Locus Browser	Visualize QTLs by gene in the Locus Browser
	IGV eQTL Browser	Visualize eQTLs in the IGV Browser
	eQTL Dashboard	Batch query eQTLs by gene and tissue
	eQTL Calculator	Test your own eQTLs
eGTEx	Data coming soon!	DNA, RNA methylation, ChIP-seq and more



Variant Page

Search: Show 10 entries

Variant ID	Shorthand	rs ID (v151)	Chromosome	Position	MAF >= 1%	Ref Allele	Alt Allele	b37 Variant ID
chr12_25209618_A_C_b38		rs712	chr12	25209618	true	A	C	12_25362552_A_C_b37

Showing 1 to 1 of 1 entries

Previous 1 Next

Single-Tissue eQTLs for chr12_25209618_A_C_b38

Data Source: GTEx Analysis Release V8 (dbGaP Accession phs000424.v8.p2)

eQTLs of chr12_25209618_A_C_b38

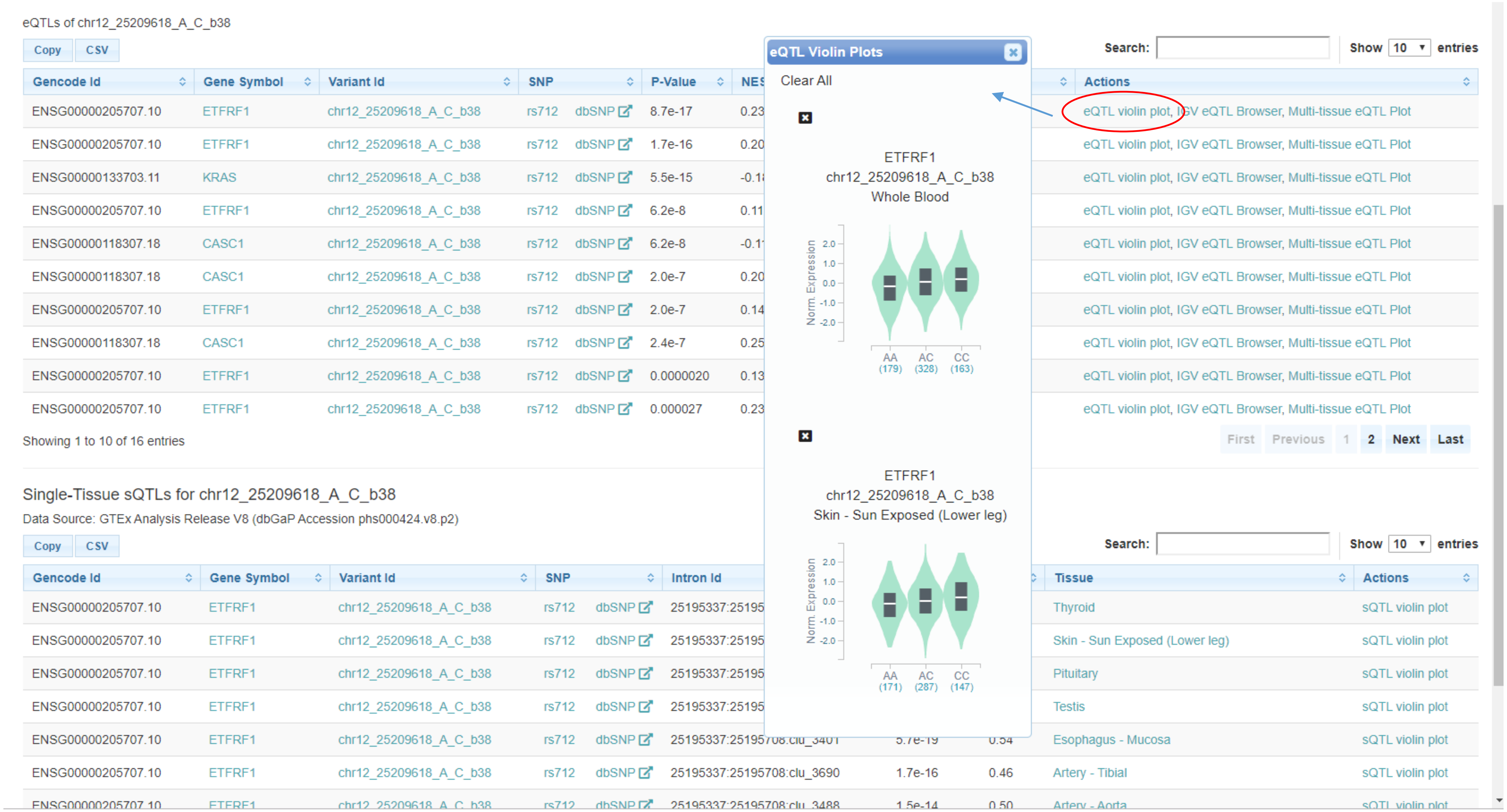
Copy CSV

Search: Show 10 entries

Gencode Id	Gene Symbol	Variant Id	SNP	P-Value	NES	Tissue	Actions
ENSG00000205707.10	ETFRF1	chr12_25209618_A_C_b38	rs712 dbSNP	8.7e-17	0.23	Whole Blood	eQTL violin plot, IGV eQTL Browser, Multi-tissue eQTL Plot
ENSG00000205707.10	ETFRF1	chr12_25209618_A_C_b38	rs712 dbSNP	1.7e-16	0.20	Skin - Sun Exposed (Lower leg)	eQTL violin plot, IGV eQTL Browser, Multi-tissue eQTL Plot
ENSG00000133703.11	KRAS	chr12_25209618_A_C_b38	rs712 dbSNP	5.5e-15	-0.18	Cells - Cultured fibroblasts	eQTL violin plot, IGV eQTL Browser, Multi-tissue eQTL Plot
ENSG00000205707.10	ETFRF1	chr12_25209618_A_C_b38	rs712 dbSNP	6.2e-8	0.11	Testis	eQTL violin plot, IGV eQTL Browser, Multi-tissue eQTL Plot
ENSG00000118307.18	CASC1	chr12_25209618_A_C_b38	rs712 dbSNP	6.2e-8	-0.11	Testis	eQTL violin plot, IGV eQTL Browser, Multi-tissue eQTL Plot
ENSG00000118307.18	CASC1	chr12_25209618_A_C_b38	rs712 dbSNP	2.0e-7	0.20	Skin - Sun Exposed (Lower leg)	eQTL violin plot, IGV eQTL Browser, Multi-tissue eQTL Plot
ENSG00000205707.10	ETFRF1	chr12_25209618_A_C_b38	rs712 dbSNP	2.0e-7	0.14	Skin - Not Sun Exposed (Suprapubic)	eQTL violin plot, IGV eQTL Browser, Multi-tissue eQTL Plot
ENSG00000118307.18	CASC1	chr12_25209618_A_C_b38	rs712 dbSNP	2.4e-7	0.25	Nerve - Tibial	eQTL violin plot, IGV eQTL Browser, Multi-tissue eQTL Plot
ENSG00000205707.10	ETFRF1	chr12_25209618_A_C_b38	rs712 dbSNP	0.0000020	0.13	Thyroid	eQTL violin plot, IGV eQTL Browser, Multi-tissue eQTL Plot
ENSG00000205707.10	ETFRF1	chr12_25209618_A_C_b38	rs712 dbSNP	0.000027	0.23	Brain - Cerebellum	eQTL violin plot, IGV eQTL Browser, Multi-tissue eQTL Plot

Showing 1 to 10 of 16 entries

First Previous 1 2 Next Last



eQTLs of chr12_25209618_A_C_b38

Copy CSV

Search: Show 10 ▾ entries

Gencode Id	Gene Symbol	Variant Id	SNP	P-Value	NES		Actions
ENSG00000205707.10	ETFRF1	chr12_25209618_A_C_b38	rs712 dbSNP	8.7e-17	0.23		eQTL violin plot, IGV eQTL Browser, Multi-tissue eQTL Plot
ENSG00000205707.10	ETFRF1	chr12_25209618_A_C_b38	rs712 dbSNP	1.7e-16	0.20		eQTL violin plot, IGV eQTL Browser, Multi-tissue eQTL Plot
ENSG00000133703.11	KRAS	chr12_25209618_A_C_b38	rs712 dbSNP	5.5e-15	-0.1		eQTL violin plot, IGV eQTL Browser, Multi-tissue eQTL Plot
ENSG00000205707.10	ETFRF1	chr12_25209618_A_C_b38	rs712 dbSNP	6.2e-8	0.11		eQTL violin plot, IGV eQTL Browser, Multi-tissue eQTL Plot
ENSG00000118307.18	CASC1	chr12_25209618_A_C_b38	rs712 dbSNP	6.2e-8	-0.1		eQTL violin plot, IGV eQTL Browser, Multi-tissue eQTL Plot
ENSG00000118307.18	CASC1	chr12_25209618_A_C_b38	rs712 dbSNP	2.0e-7	0.20		eQTL violin plot, IGV eQTL Browser, Multi-tissue eQTL Plot
ENSG00000205707.10	ETFRF1	chr12_25209618_A_C_b38	rs712 dbSNP	2.0e-7	0.14		eQTL violin plot, IGV eQTL Browser, Multi-tissue eQTL Plot
ENSG00000118307.18	CASC1	chr12_25209618_A_C_b38	rs712 dbSNP	2.4e-7	0.25		eQTL violin plot, IGV eQTL Browser, Multi-tissue eQTL Plot
ENSG00000205707.10	ETFRF1	chr12_25209618_A_C_b38	rs712 dbSNP	0.0000020	0.13		eQTL violin plot, IGV eQTL Browser, Multi-tissue eQTL Plot
ENSG00000205707.10	ETFRF1	chr12_25209618_A_C_b38	rs712 dbSNP	0.000027	0.23		eQTL violin plot, IGV eQTL Browser, Multi-tissue eQTL Plot

Showing 1 to 10 of 16 entries

First Previous 1 2 Next Last

Single-Tissue sQTLs for chr12_25209618_A_C_b38

Data Source: GTEx Analysis Release V8 (dbGaP Accession phs000424.v8.p2)

Copy CSV

Search: Show 10 ▾ entries

Gencode Id	Gene Symbol	Variant Id	SNP	Intron Id			Tissue	Actions
ENSG00000205707.10	ETFRF1	chr12_25209618_A_C_b38	rs712 dbSNP	25195337:25195			Thyroid	sQTL violin plot
ENSG00000205707.10	ETFRF1	chr12_25209618_A_C_b38	rs712 dbSNP	25195337:25195			Skin - Sun Exposed (Lower leg)	sQTL violin plot
ENSG00000205707.10	ETFRF1	chr12_25209618_A_C_b38	rs712 dbSNP	25195337:25195			Pituitary	sQTL violin plot
ENSG00000205707.10	ETFRF1	chr12_25209618_A_C_b38	rs712 dbSNP	25195337:25195			Testis	sQTL violin plot
ENSG00000205707.10	ETFRF1	chr12_25209618_A_C_b38	rs712 dbSNP	25195337:25195708:clu_3401	5.7e-19	0.54	Esophagus - Mucosa	sQTL violin plot
ENSG00000205707.10	ETFRF1	chr12_25209618_A_C_b38	rs712 dbSNP	25195337:25195708:clu_3690	1.7e-16	0.46	Artery - Tibial	sQTL violin plot
ENSG00000205707.10	ETFRF1	chr12_25209618_A_C_b38	rs712 dbSNP	25195337:25195708:clu_3488	1.5e-14	0.50	Artery - Aorta	sQTL violin plot

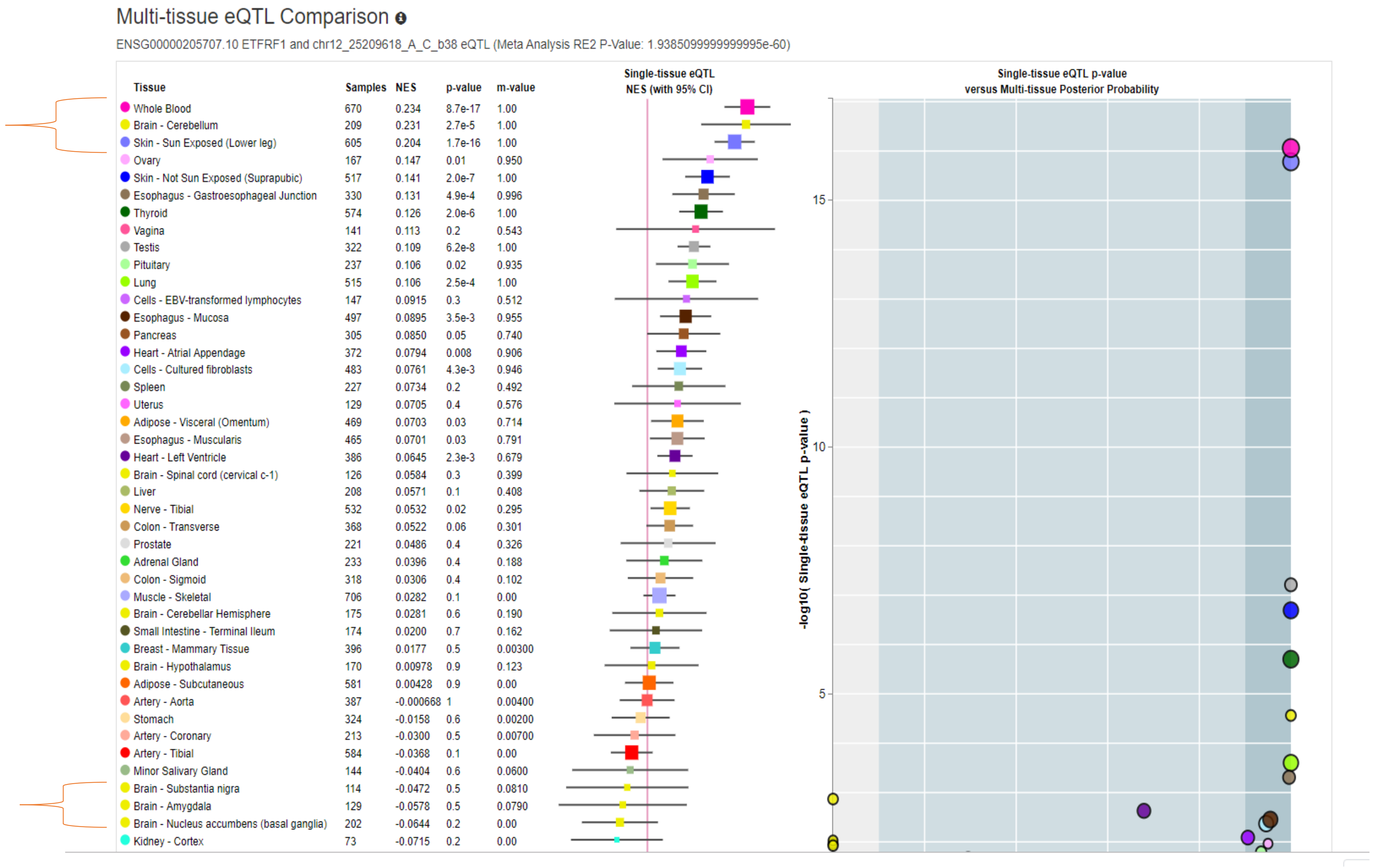
eQTL Violin Plots

Clear All

ETFRF1
chr12_25209618_A_C_b38
Whole Blood

ETFRF1
chr12_25209618_A_C_b38
Skin - Sun Exposed (Lower leg)

Indicates snps has high expression in human blood and skin tissues



Multi-tissue eQTL Comparison

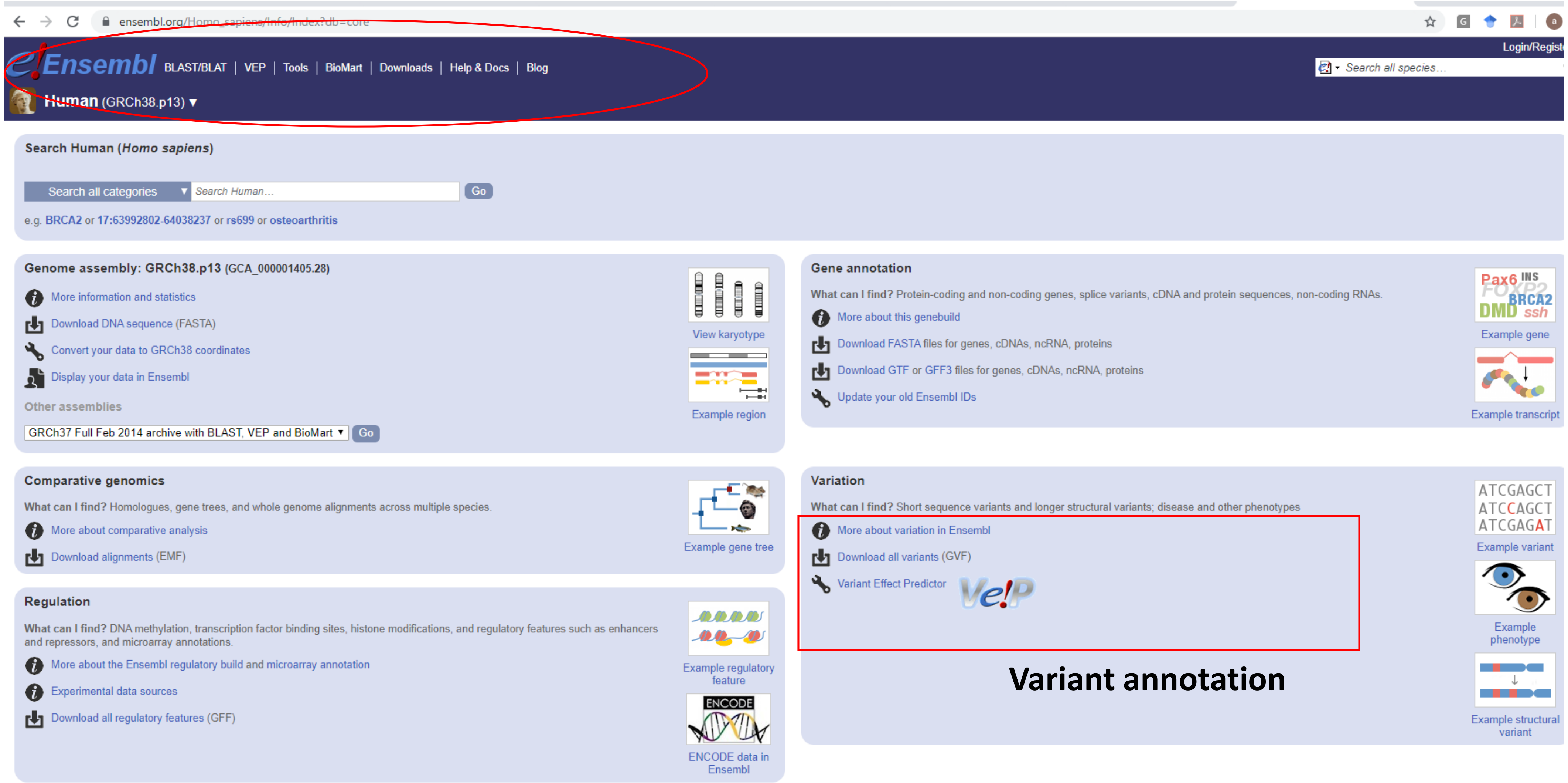
ENSG00000205707.10 ETFRF1 and chr12_25209618_A_C_b38 eQTL (Meta Analysis RE2 P-Value: 1.9385099999999995e-60)

	Tissue	Samples	NES	p-value	m-value
●	Whole Blood	670	0.234	8.7e-17	1.00
●	Brain - Cerebellum	209	0.231	2.7e-5	1.00
●	Skin - Sun Exposed (Lower leg)	605	0.204	1.7e-16	1.00
●	Ovary	167	0.147	0.01	0.950
●	Skin - Not Sun Exposed (Suprapubic)	517	0.141	2.0e-7	1.00
●	Esophagus - Gastroesophageal Junction	330	0.131	4.9e-4	0.996
●	Thyroid	574	0.126	2.0e-6	1.00
●	Vagina	141	0.113	0.2	0.543
●	Testis	322	0.109	6.2e-8	1.00
●	Pituitary	237	0.106	0.02	0.935
●	Lung	515	0.106	2.5e-4	1.00
●	Cells - EBV-transformed lymphocytes	147	0.0915	0.3	0.512
●	Esophagus - Mucosa	497	0.0895	3.5e-3	0.955
●	Pancreas	305	0.0850	0.05	0.740
●	Heart - Atrial Appendage	372	0.0794	0.008	0.906
●	Cells - Cultured fibroblasts	483	0.0761	4.3e-3	0.946
●	Spleen	227	0.0734	0.2	0.492
●	Uterus	129	0.0705	0.4	0.576
●	Adipose - Visceral (Omentum)	469	0.0703	0.03	0.714
●	Esophagus - Muscularis	465	0.0701	0.03	0.791
●	Heart - Left Ventricle	386	0.0645	2.3e-3	0.679
●	Brain - Spinal cord (cervical c-1)	126	0.0584	0.3	0.399
●	Liver	208	0.0571	0.1	0.408
●	Nerve - Tibial	532	0.0532	0.02	0.295
●	Colon - Transverse	368	0.0522	0.06	0.301
●	Prostate	221	0.0486	0.4	0.326
●	Adrenal Gland	233	0.0396	0.4	0.188
●	Colon - Sigmoid	318	0.0306	0.4	0.102
●	Muscle - Skeletal	706	0.0282	0.1	0.00
●	Brain - Cerebellar Hemisphere	175	0.0281	0.6	0.190
●	Small Intestine - Terminal Ileum	174	0.0200	0.7	0.162
●	Breast - Mammary Tissue	396	0.0177	0.5	0.00300
●	Brain - Hypothalamus	170	0.00978	0.9	0.123
●	Adipose - Subcutaneous	581	0.00428	0.9	0.00
●	Artery - Aorta	387	-0.000668	1	0.00400
●	Stomach	324	-0.0158	0.6	0.00200
●	Artery - Coronary	213	-0.0300	0.5	0.00700
●	Artery - Tibial	584	-0.0368	0.1	0.00
●	Minor Salivary Gland	144	-0.0404	0.6	0.0600
●	Brain - Substantia nigra	114	-0.0472	0.5	0.0810
●	Brain - Amygdala	129	-0.0578	0.5	0.0790
●	Brain - Nucleus accumbens (basal ganglia)	202	-0.0644	0.2	0.00
●	Kidney - Cortex	73	-0.0715	0.2	0.00

Single-tissue eQTL
NES (with 95% CI)
Single-tissue eQTL p-value
versus Multi-tissue Posterior Probability
-log10(Single-tissue eQTL p-value)
15
10
5

Ensembl Database

https://www.ensembl.org/index.html



ensembl.org/Homo_sapiens/Info/Index?db=core

Ensembl BLAST/BLAT | VEP | Tools | BioMart | Downloads | Help & Docs | Blog

Login/Regist

Search all species…

Human (GRCh38.p13) ▾

Search Human (*Homo sapiens*)

Search all categories ▾ Search Human… Go

e.g. BRCA2 or 17:63992802-64038237 or rs699 or osteoarthritis

Genome assembly: GRCh38.p13 (GCA_000001405.28)

- More information and statistics
- Download DNA sequence (FASTA)
- Convert your data to GRCh38 coordinates
- Display your data in Ensembl

View karyotype

Example region

Other assemblies

GRCh37 Full Feb 2014 archive with BLAST, VEP and BioMart ▾ Go

Gene annotation

What can I find? Protein-coding and non-coding genes, splice variants, cDNA and protein sequences, non-coding RNAs.

- More about this genebuild
- Download FASTA files for genes, cDNAs, ncRNA, proteins
- Download GTF or GFF3 files for genes, cDNAs, ncRNA, proteins
- Update your old Ensembl IDs

Example gene

Example transcript

Comparative genomics

What can I find? Homologues, gene trees, and whole genome alignments across multiple species.

- More about comparative analysis
- Download alignments (EMF)

Example gene tree

Variation

What can I find? Short sequence variants and longer structural variants; disease and other phenotypes

- More about variation in Ensembl
- Download all variants (GVF)
- Variant Effect Predictor

Example variant

Example phenotype

Example structural variant

Variant annotation

Regulation

What can I find? DNA methylation, transcription factor binding sites, histone modifications, and regulatory features such as enhancers and repressors, and microarray annotations.

- More about the Ensembl regulatory build and microarray annotation
- Experimental data sources
- Download all regulatory features (GFF)

Example regulatory feature

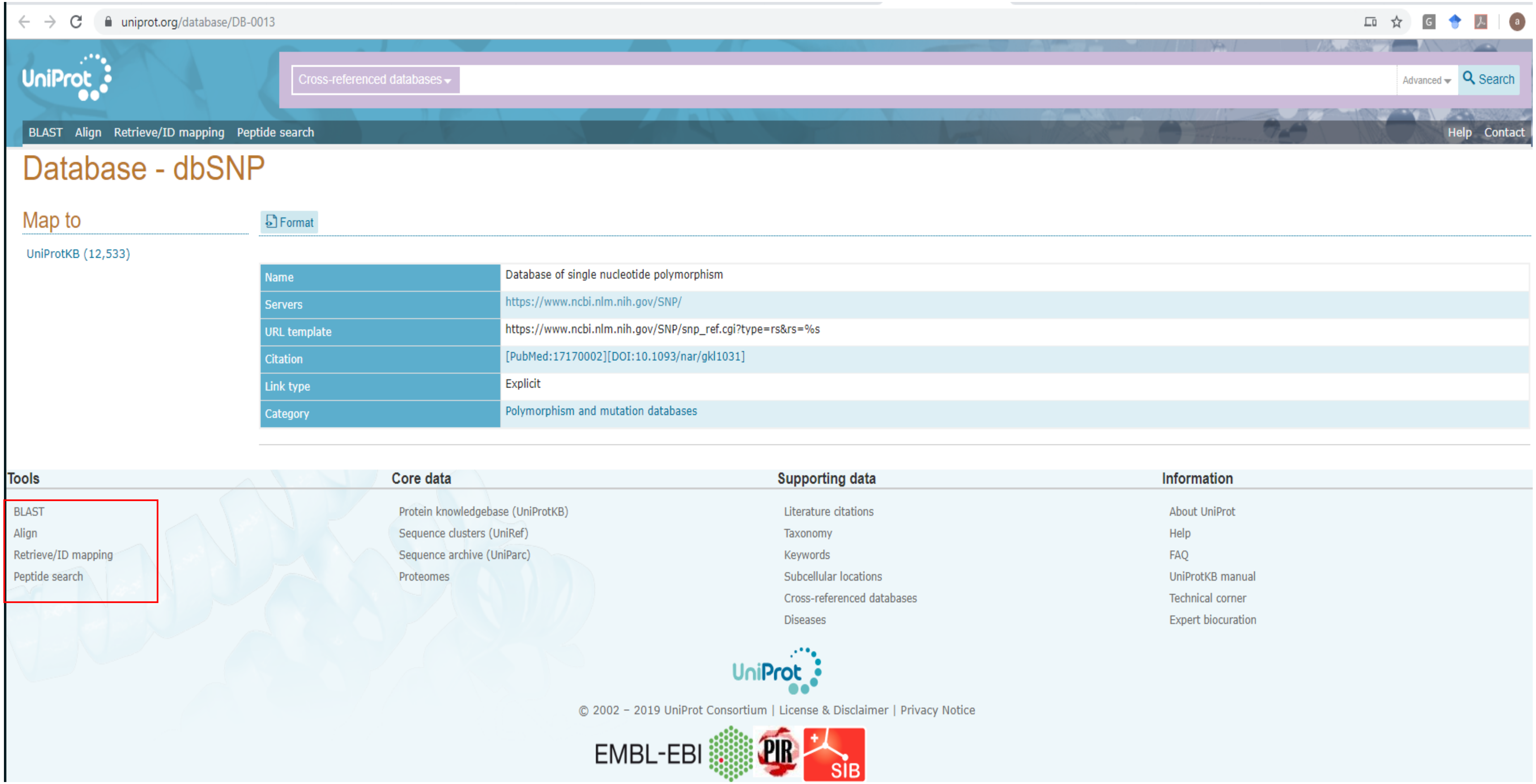
ENCODE data in Ensembl

UNIPROT KB

Available at https://www.uniprot.org/

- The UniProt Knowledgebase (UniProtKB) is the central hub for the collection of functional information on proteins, with accurate, consistent and rich annotation.

- In addition to capturing the core data mandatory for each UniProtKB entry (mainly, the amino acid sequence, protein name or description, taxonomic data and citation information), as much annotation information as possible is added.



Database - dbSNP

Map to

UniProtKB (12,533)

Format

Name	Database of single nucleotide polymorphism
Servers	https://www.ncbi.nlm.nih.gov/SNP/
URL template	https://www.ncbi.nlm.nih.gov/SNP/snp_ref.cgi?type=rs&rs=%s
Citation	[PubMed:17170002][DOI:10.1093/nar/gkl1031]
Link type	Explicit
Category	Polymorphism and mutation databases

Tools

- BLAST
- Align
- Retrieve/ID mapping
- Peptide search

Core data

- Protein knowledgebase (UniProtKB)
- Sequence clusters (UniRef)
- Sequence archive (UniParc)
- Proteomes

Supporting data

- Literature citations
- Taxonomy
- Keywords
- Subcellular locations
- Cross-referenced databases
- Diseases

Information

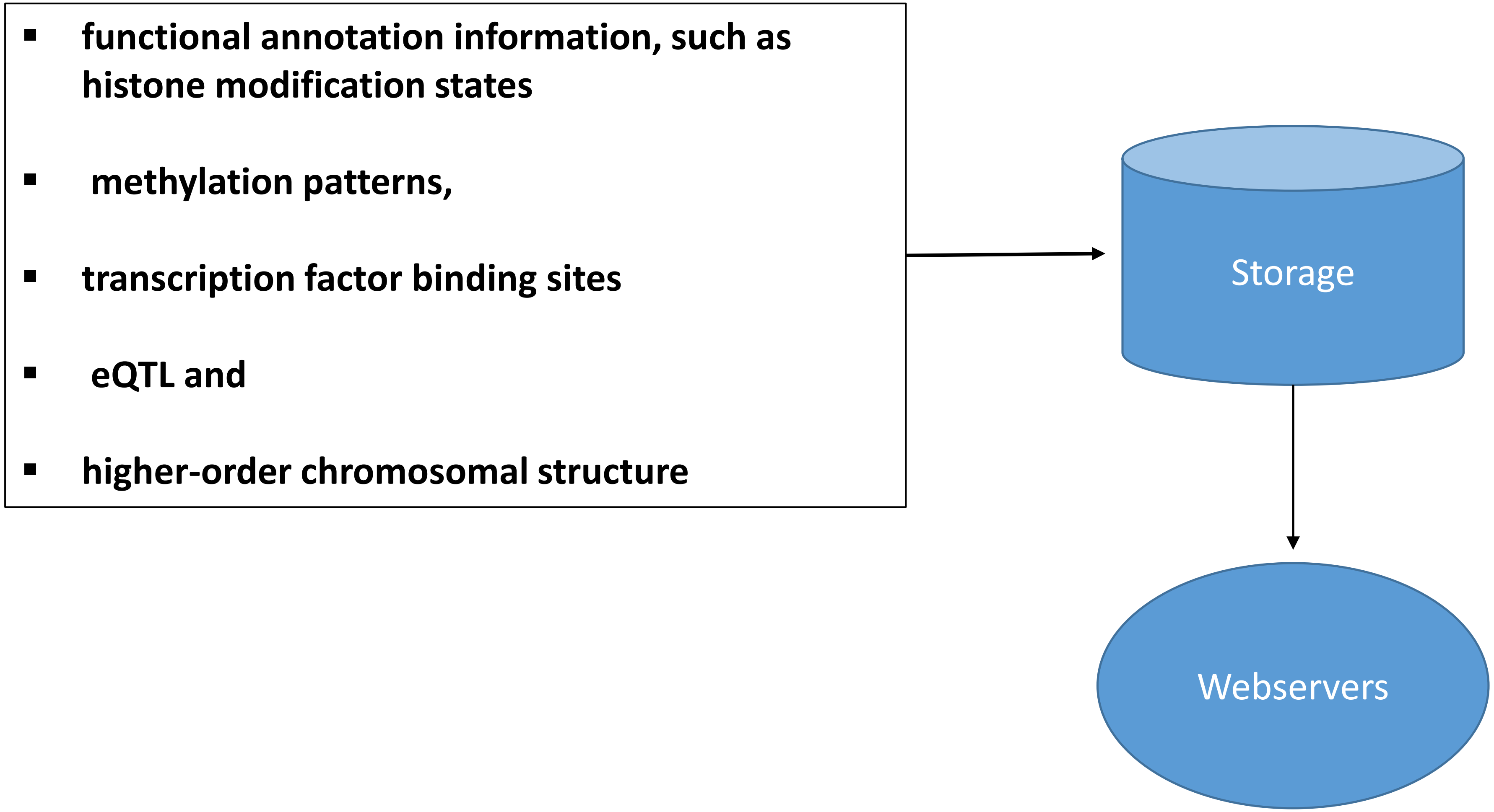
- About UniProt
- Help
- FAQ
- UniProtKB manual
- Technical corner
- Expert biocuration

Multiple web servers (for Post GWAS)

- **Identifying causal variants remains a key challenge in post-GWAS (genome-wide association study) era, as many GWAS single-nucleotide polymorphisms (SNPs) (including imputed ones) fall into non-coding regions.**

- **Its making it difficult to associate statistical significance with predicted functionality.**

- **Therefore, researches developed web-based multiple tools which overlays functional annotation information, such as histone modification states, methylation patterns, transcription factor binding sites, eQTL and higher-order chromosomal structure, to GWAS results.**



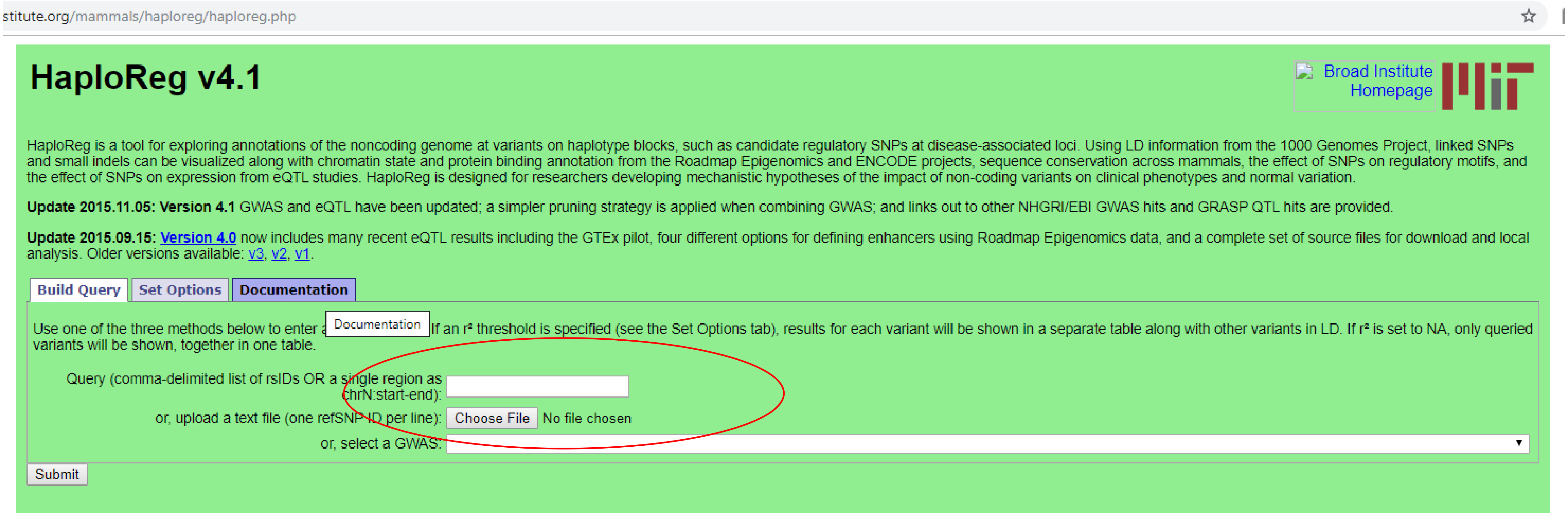
- functional annotation information, such as histone modification states
- methylation patterns,
- transcription factor binding sites
- eQTL and
- higher-order chromosomal structure

Storage

Webservers

HaploReg web server

https://pubs.broadinstitute.org/mammals/haploreg/haploreg.php



stitute.org/mammals/haploreg/haploreg.php

HaploReg v4.1

Broad Institute Homepage

HaploReg is a tool for exploring annotations of the noncoding genome at variants on haplotype blocks, such as candidate regulatory SNPs at disease-associated loci. Using LD information from the 1000 Genomes Project, linked SNPs and small indels can be visualized along with chromatin state and protein binding annotation from the Roadmap Epigenomics and ENCODE projects, sequence conservation across mammals, the effect of SNPs on regulatory motifs, and the effect of SNPs on expression from eQTL studies. HaploReg is designed for researchers developing mechanistic hypotheses of the impact of non-coding variants on clinical phenotypes and normal variation.

Update 2015.11.05: Version 4.1 GWAS and eQTL have been updated; a simpler pruning strategy is applied when combining GWAS; and links out to other NHGRI/EBI GWAS hits and GRASP QTL hits are provided.

Update 2015.09.15: Version 4.0 now includes many recent eQTL results including the GTEx pilot, four different options for defining enhancers using Roadmap Epigenomics data, and a complete set of source files for download and local analysis. Older versions available: v3, v2, v1.

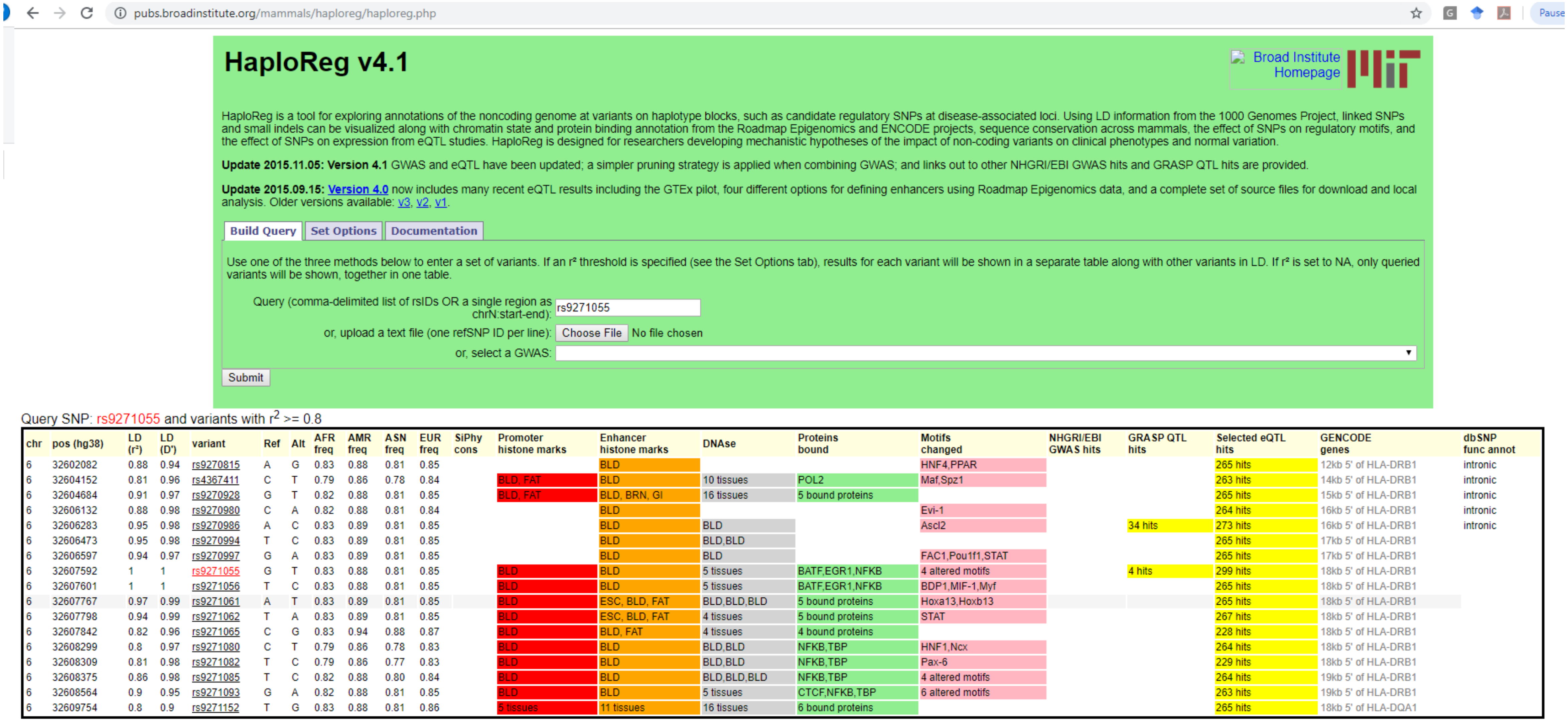
Use one of the three methods below to enter a If an r² threshold is specified (see the Set Options tab), results for each variant will be shown in a separate table along with other variants in LD. If r² is set to NA, only queried variants will be shown, together in one table.

Query (comma-delimited list of rsIDs OR a single region as chrN:start-end):

or, upload a text file (one refSNP ID per line): Choose File No file chosen

or, select a GWAS:

Submit



HaploReg v4.1

HaploReg is a tool for exploring annotations of the noncoding genome at variants on haplotype blocks, such as candidate regulatory SNPs at disease-associated loci. Using LD information from the 1000 Genomes Project, linked SNPs and small indels can be visualized along with chromatin state and protein binding annotation from the Roadmap Epigenomics and ENCODE projects, sequence conservation across mammals, the effect of SNPs on regulatory motifs, and the effect of SNPs on expression from eQTL studies. HaploReg is designed for researchers developing mechanistic hypotheses of the impact of non-coding variants on clinical phenotypes and normal variation.

Update 2015.11.05: Version 4.1 GWAS and eQTL have been updated; a simpler pruning strategy is applied when combining GWAS; and links out to other NHGRI/EBI GWAS hits and GRASP QTL hits are provided.

Update 2015.09.15: Version 4.0 now includes many recent eQTL results including the GTEx pilot, four different options for defining enhancers using Roadmap Epigenomics data, and a complete set of source files for download and local analysis. Older versions available: v3, v2, v1.

Use one of the three methods below to enter a set of variants. If an r² threshold is specified (see the Set Options tab), results for each variant will be shown in a separate table along with other variants in LD. If r² is set to NA, only queried variants will be shown, together in one table.

Query (comma-delimited list of rsIDs OR a single region as chrN:start-end): rs9271055

or, upload a text file (one refSNP ID per line): Choose File No file chosen

or, select a GWAS:

Query SNP: rs9271055 and variants with r^2 >= 0.8

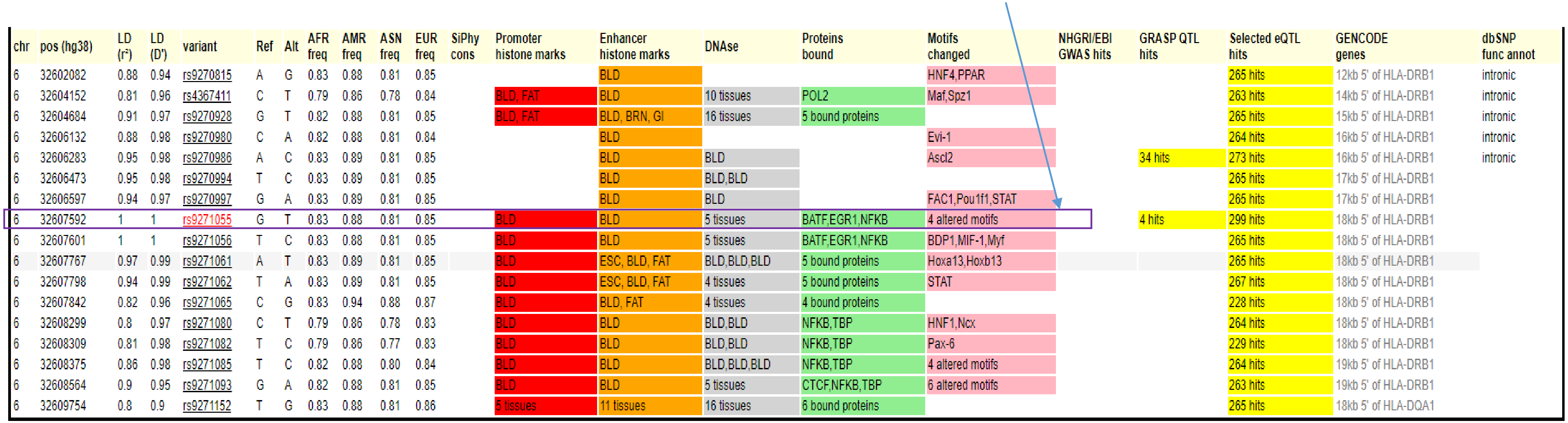
chr	pos (hg38)	LD (r²)	LD (D')	variant	Ref	Alt	AFR freq	AMR freq	ASN freq	EUR freq	SiPhy cons	Promoter histone marks	Enhancer histone marks	DNAse	Proteins bound	Motifs changed	NHGRI/EBI GWAS hits	GRASP QTL hits	Selected eQTL hits	GENCODE genes	dbSNP func annot
6	32602082	0.88	0.94	rs9270815	A	G	0.83	0.88	0.81	0.85			BLD			HNF4,PPAR			265 hits	12kb 5' of HLA-DRB1	intronic
6	32604152	0.81	0.96	rs4367411	C	T	0.79	0.86	0.78	0.84		BLD, FAT	BLD	10 tissues	POL2	Maf,Spz1			263 hits	14kb 5' of HLA-DRB1	intronic
6	32604684	0.91	0.97	rs9270928	G	T	0.82	0.88	0.81	0.85		BLD, FAT	BLD, BRN, GI	16 tissues	5 bound proteins				265 hits	15kb 5' of HLA-DRB1	intronic
6	32606132	0.88	0.98	rs9270980	C	A	0.82	0.88	0.81	0.84			BLD			Evi-1			264 hits	16kb 5' of HLA-DRB1	intronic
6	32606283	0.95	0.98	rs9270986	A	C	0.83	0.89	0.81	0.85			BLD	BLD		Ascl2		34 hits	273 hits	16kb 5' of HLA-DRB1	intronic
6	32606473	0.95	0.98	rs9270994	T	C	0.83	0.89	0.81	0.85			BLD	BLD,BLD					265 hits	17kb 5' of HLA-DRB1	
6	32606597	0.94	0.97	rs9270997	G	A	0.83	0.89	0.81	0.85			BLD	BLD		FAC1,Pou1f1,STAT			265 hits	17kb 5' of HLA-DRB1	
6	32607592	1	1	rs9271055	G	T	0.83	0.88	0.81	0.85		BLD	BLD	5 tissues	BATF,EGR1,NFKB	4 altered motifs		4 hits	299 hits	18kb 5' of HLA-DRB1	
6	32607601	1	1	rs9271056	T	C	0.83	0.88	0.81	0.85		BLD	BLD	5 tissues	BATF,EGR1,NFKB	BDP1,MIF-1,Myf			265 hits	18kb 5' of HLA-DRB1	
6	32607767	0.97	0.99	rs9271061	A	T	0.83	0.89	0.81	0.85		BLD	ESC, BLD, FAT	BLD,BLD,BLD	5 bound proteins	Hoxa13,Hoxb13			265 hits	18kb 5' of HLA-DRB1	
6	32607798	0.94	0.99	rs9271062	T	A	0.83	0.89	0.81	0.85		BLD	ESC, BLD, FAT	4 tissues	5 bound proteins	STAT			267 hits	18kb 5' of HLA-DRB1	
6	32607842	0.82	0.96	rs9271065	C	G	0.83	0.94	0.88	0.87		BLD	BLD, FAT	4 tissues	4 bound proteins				228 hits	18kb 5' of HLA-DRB1	
6	32608299	0.8	0.97	rs9271080	C	T	0.79	0.86	0.78	0.83		BLD	BLD	BLD,BLD	NFKB,TBP	HNF1,Ncx			264 hits	18kb 5' of HLA-DRB1	
6	32608309	0.81	0.98	rs9271082	T	C	0.79	0.86	0.77	0.83		BLD	BLD	BLD,BLD	NFKB,TBP	Pax-6			229 hits	18kb 5' of HLA-DRB1	
6	32608375	0.86	0.98	rs9271085	T	C	0.82	0.88	0.80	0.84		BLD	BLD	BLD,BLD,BLD	NFKB,TBP	4 altered motifs			264 hits	19kb 5' of HLA-DRB1	
6	32608564	0.9	0.95	rs9271093	G	A	0.82	0.88	0.81	0.85		BLD	BLD	5 tissues	CTCF,NFKB,TBP	6 altered motifs			263 hits	19kb 5' of HLA-DRB1	
6	32609754	0.8	0.9	rs9271152	T	G	0.83	0.88	0.81	0.86		5 tissues	11 tissues	16 tissues	6 bound proteins				265 hits	18kb 5' of HLA-DQA1	

Advantage

- It was developed to systematically mine chromatin state data, along with conservation data and regulatory motif alterations.

- It uses Gtex , Encode databases in backend.

- Most importanlly, it gives motif based regulatory impact of SNPs

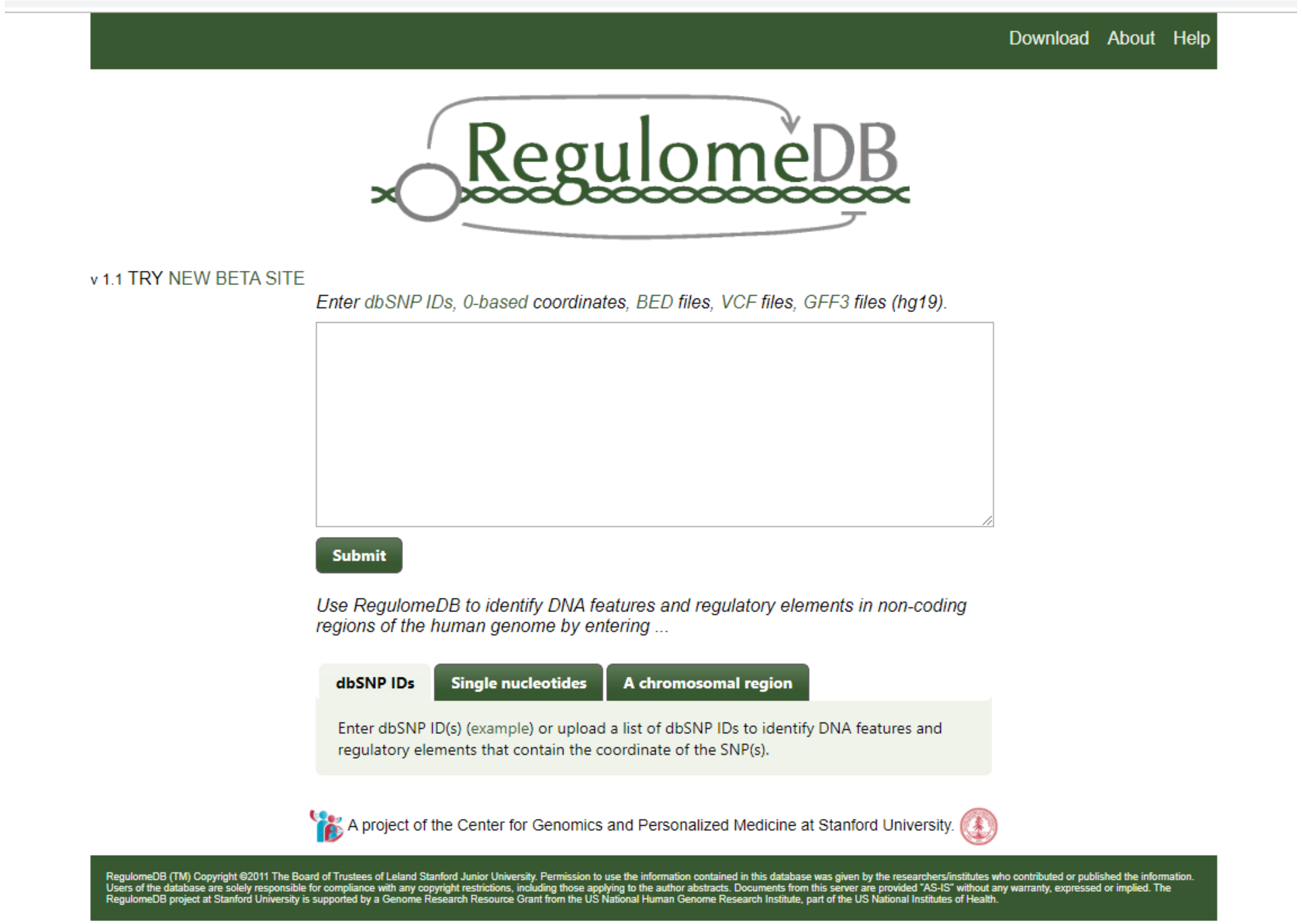


SNP causes 4 altered motifs due to change in nucleotide from G to T

chr	pos (hg38)	LD (r^2)	LD (D')	variant	Ref	Alt	AFR freq	AMR freq	ASN freq	EUR freq	SiPhy cons	Promoter histone marks	Enhancer histone marks	DNAse	Proteins bound	Motifs changed	NHGRI/EBI GWAS hits	GRASP QTL hits	Selected eQTL hits	GENCODE genes	dbSNP func annot
6	32602082	0.88	0.94	rs9270815	A	G	0.83	0.88	0.81	0.85			BLD			HNF4,PPAR			265 hits	12kb 5' of HLA-DRB1	intronic
6	32604152	0.81	0.96	rs4367411	C	T	0.79	0.86	0.78	0.84		BLD, FAT	BLD	10 tissues	POL2	Maf,Spz1			263 hits	14kb 5' of HLA-DRB1	intronic
6	32604684	0.91	0.97	rs9270928	G	T	0.82	0.88	0.81	0.85		BLD, FAT	BLD, BRN, GI	16 tissues	5 bound proteins				265 hits	15kb 5' of HLA-DRB1	intronic
6	32606132	0.88	0.98	rs9270980	C	A	0.82	0.88	0.81	0.84			BLD			Evi-1			264 hits	16kb 5' of HLA-DRB1	intronic
6	32606283	0.95	0.98	rs9270986	A	C	0.83	0.89	0.81	0.85			BLD	BLD		Ascl2		34 hits	273 hits	16kb 5' of HLA-DRB1	intronic
6	32606473	0.95	0.98	rs9270994	T	C	0.83	0.89	0.81	0.85			BLD	BLD,BLD					265 hits	17kb 5' of HLA-DRB1	
6	32606597	0.94	0.97	rs9270997	G	A	0.83	0.89	0.81	0.85			BLD	BLD		FAC1,Pou1f1,STAT			265 hits	17kb 5' of HLA-DRB1	
6	32607592	1	1	rs9271055	G	T	0.83	0.88	0.81	0.85		BLD	BLD	5 tissues	BATF,EGR1,NFKB	4 altered motifs		4 hits	299 hits	18kb 5' of HLA-DRB1	
6	32607601	1	1	rs9271056	T	C	0.83	0.88	0.81	0.85		BLD	BLD	5 tissues	BATF,EGR1,NFKB	BDP1,MIF-1,Myf			265 hits	18kb 5' of HLA-DRB1	
6	32607767	0.97	0.99	rs9271061	A	T	0.83	0.89	0.81	0.85		BLD	ESC, BLD, FAT	BLD,BLD,BLD	5 bound proteins	Hoxa13,Hoxb13			265 hits	18kb 5' of HLA-DRB1	
6	32607798	0.94	0.99	rs9271062	T	A	0.83	0.89	0.81	0.85		BLD	ESC, BLD, FAT	4 tissues	5 bound proteins	STAT			267 hits	18kb 5' of HLA-DRB1	
6	32607842	0.82	0.96	rs9271065	C	G	0.83	0.94	0.88	0.87		BLD	BLD, FAT	4 tissues	4 bound proteins				228 hits	18kb 5' of HLA-DRB1	
6	32608299	0.8	0.97	rs9271080	C	T	0.79	0.86	0.78	0.83		BLD	BLD	BLD,BLD	NFKB,TBP	HNF1,Ncx			264 hits	18kb 5' of HLA-DRB1	
6	32608309	0.81	0.98	rs9271082	T	C	0.79	0.86	0.77	0.83		BLD	BLD	BLD,BLD	NFKB,TBP	Pax-6			229 hits	18kb 5' of HLA-DRB1	
6	32608375	0.86	0.98	rs9271085	T	C	0.82	0.88	0.80	0.84		BLD	BLD	BLD,BLD,BLD	NFKB,TBP	4 altered motifs			264 hits	19kb 5' of HLA-DRB1	
6	32608564	0.9	0.95	rs9271093	G	A	0.82	0.88	0.81	0.85		BLD	BLD	5 tissues	CTCF,NFKB,TBP	6 altered motifs			263 hits	19kb 5' of HLA-DRB1	
6	32609754	0.8	0.9	rs9271152	T	G	0.83	0.88	0.81	0.86		5 tissues	11 tissues	16 tissues	6 bound proteins				265 hits	18kb 5' of HLA-DQA1	

RegulomeDB

Access to the database at http://RegulomeDB.org/



v 1.1 TRY NEW BETA SITE

Enter dbSNP IDs, 0-based coordinates, BED files, VCF files, GFF3 files (hg19).

Use RegulomeDB to identify DNA features and regulatory elements in non-coding regions of the human genome by entering ...

dbSNP IDs | **Single nucleotides** | **A chromosomal region**

Enter dbSNP ID(s) (example) or upload a list of dbSNP IDs to identify DNA features and regulatory elements that contain the coordinate of the SNP(s).

A project of the Center for Genomics and Personalized Medicine at Stanford University.

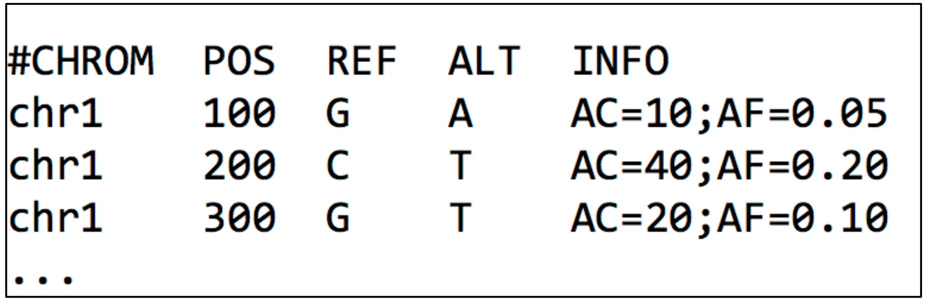
Input Files Format

- **The integrated database is fully searchable using common variant formats (VCF, BED, GFF3, rsIDs) and through file upload of the same formats.**

rsID FORMAT

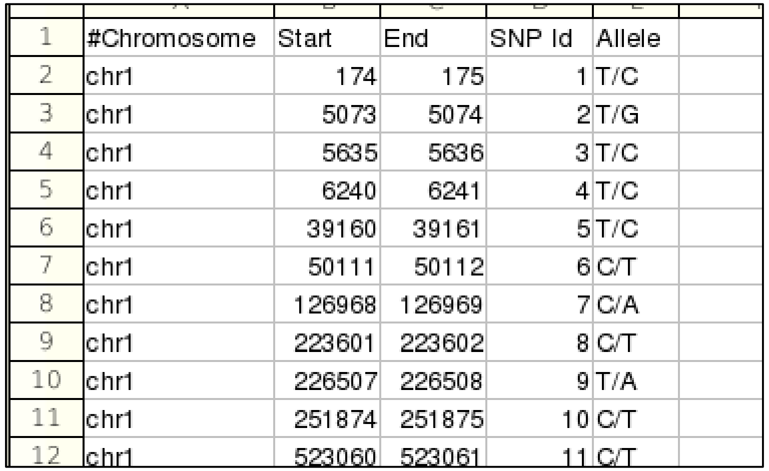
rs33914668
rs35004220
rs78077282
rs7881236

VCF FORMAT



```
#CHROM   POS   REF   ALT   INFO
chr1     100   G     A     AC=10;AF=0.05
chr1     200   C     T     AC=40;AF=0.20
chr1     300   G     T     AC=20;AF=0.10
...
```

BED FORMAT

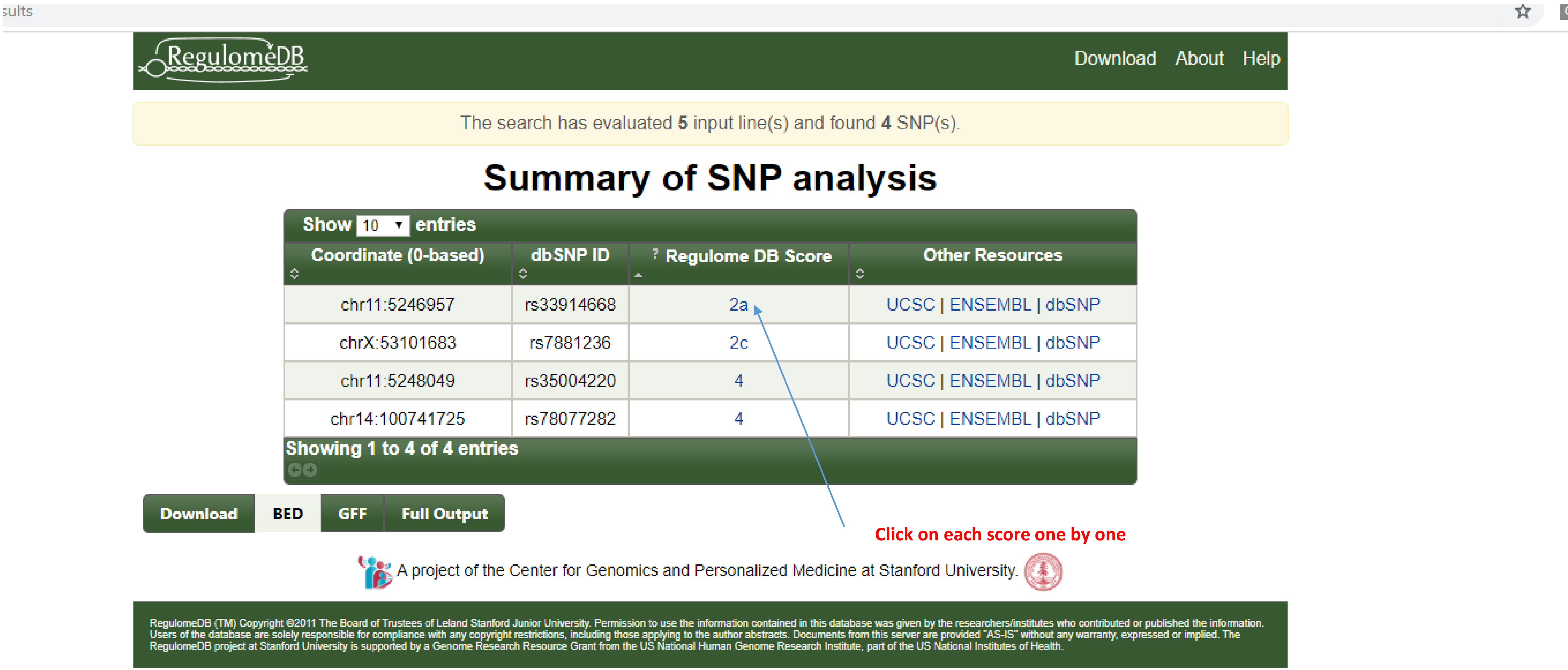


1	#Chromosome	Start	End	SNP Id	Allele	
2	chr1	174	175	1	T/C	
3	chr1	5073	5074	2	T/G	
4	chr1	5635	5636	3	T/C	
5	chr1	6240	6241	4	T/C	
6	chr1	39160	39161	5	T/C	
7	chr1	50111	50112	6	C/T	
8	chr1	126968	126969	7	C/A	
9	chr1	223601	223602	8	C/T	
10	chr1	226507	226508	9	T/A	
11	chr1	251874	251875	10	C/T	
12	chr1	523060	523061	11	C/T	

Output Files

- **The initial results table provides a list of the coordinates of the variants, a dbSNP rsID (if it exists), a score assigned by method, and links to external resources for each variant**

- **The list is sorted by our classification scheme, with the SNVs most likely to be functional listed first. This list of SNVs is also downloadable by the user for their own analysis.**



The search has evaluated **5** input line(s) and found **4** SNP(s).

Summary of SNP analysis

Show 10 entries

Coordinate (0-based)	dbSNP ID	Regulome DB Score	Other Resources
chr11:5246957	rs33914668	2a	UCSC | ENSEMBL | dbSNP
chrX:53101683	rs7881236	2c	UCSC | ENSEMBL | dbSNP
chr11:5248049	rs35004220	4	UCSC | ENSEMBL | dbSNP
chr14:100741725	rs78077282	4	UCSC | ENSEMBL | dbSNP

Showing 1 to 4 of 4 entries

Click on each score one by one

A project of the Center for Genomics and Personalized Medicine at Stanford University.

- **This display includes six major categories: Protein Binding, Motifs, Chromatin Structure, eQTLs, Histone Modifications, and Related Data (which includes gene information and other manual annotations).**

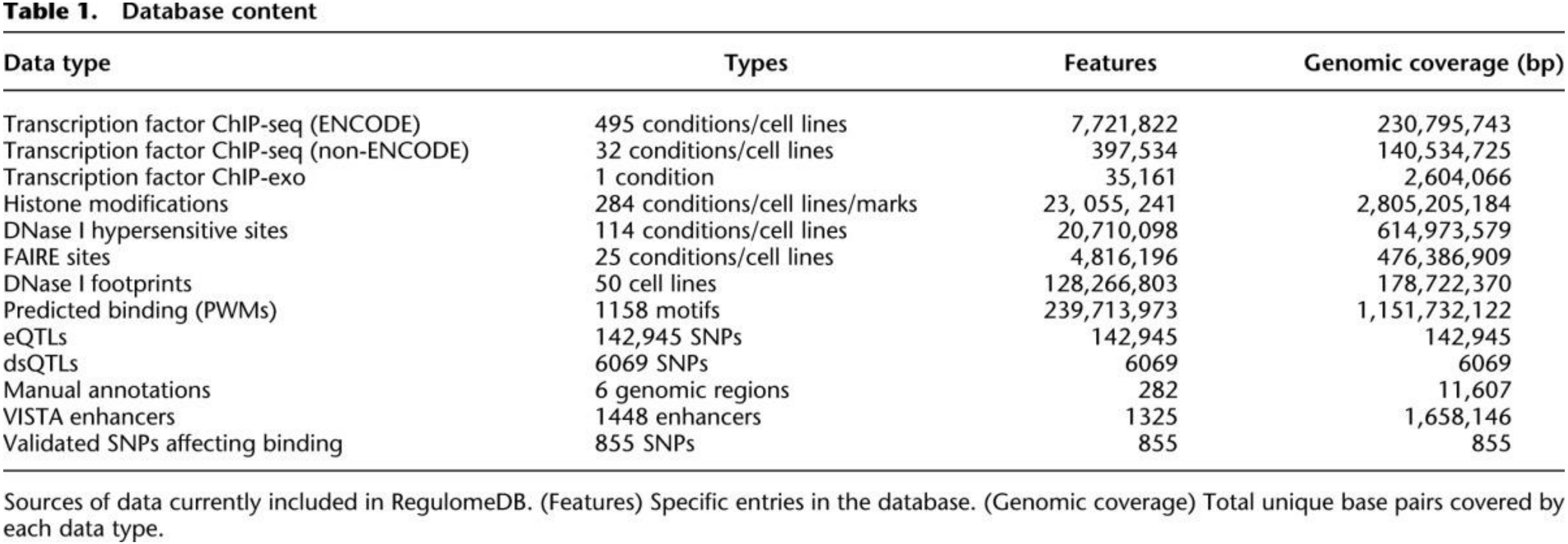
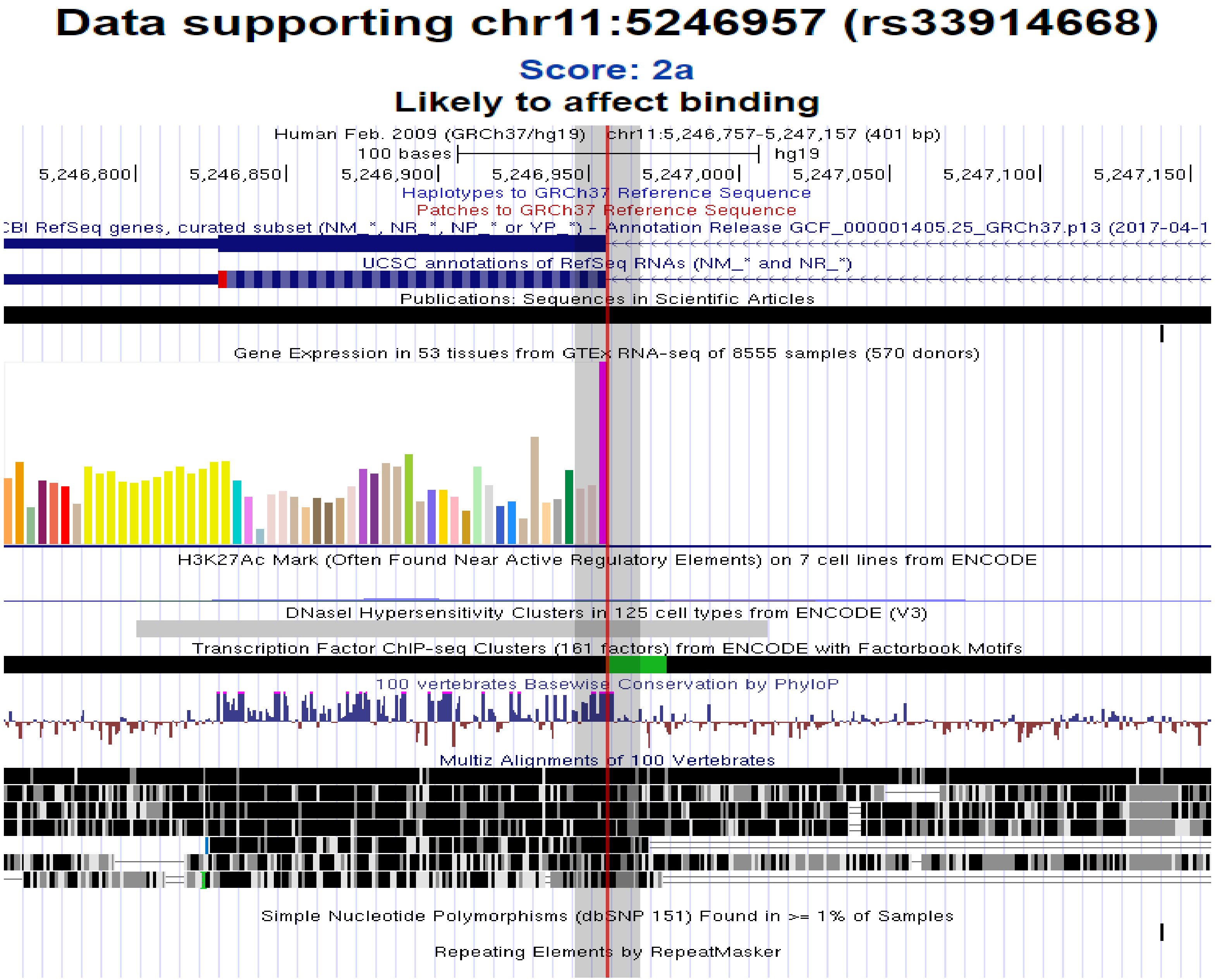


Table 1. Database content

Data type	Types	Features	Genomic coverage (bp)
Transcription factor ChIP-seq (ENCODE)	495 conditions/cell lines	7,721,822	230,795,743
Transcription factor ChIP-seq (non-ENCODE)	32 conditions/cell lines	397,534	140,534,725
Transcription factor ChIP-exo	1 condition	35,161	2,604,066
Histone modifications	284 conditions/cell lines/marks	23, 055, 241	2,805,205,184
DNase I hypersensitive sites	114 conditions/cell lines	20,710,098	614,973,579
FAIRE sites	25 conditions/cell lines	4,816,196	476,386,909
DNase I footprints	50 cell lines	128,266,803	178,722,370
Predicted binding (PWMs)	1158 motifs	239,713,973	1,151,732,122
eQTLs	142,945 SNPs	142,945	142,945
dsQTLs	6069 SNPs	6069	6069
Manual annotations	6 genomic regions	282	11,607
VISTA enhancers	1448 enhancers	1325	1,658,146
Validated SNPs affecting binding	855 SNPs	855	855

Sources of data currently included in RegulomeDB. (Features) Specific entries in the database. (Genomic coverage) Total unique base pairs covered by each data type.



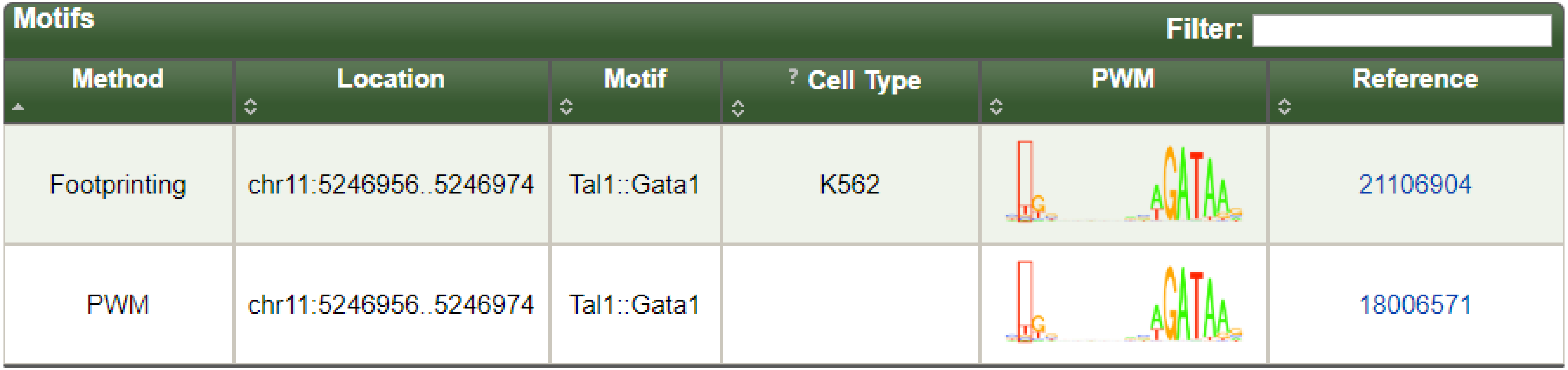
Data supporting chr11:5246957 (rs33914668)

Score: 2a

Likely to affect binding

Human Feb. 2009 (GRCh37/hg19) chr11:5,246,757-5,247,157 (401 bp)
100 bases
hg19
5,246,800
5,246,850
5,246,900
5,246,950
5,247,000
5,247,050
5,247,100
5,247,150
Haplotypes to GRCh37 Reference Sequence
Patches to GRCh37 Reference Sequence
CBI RefSeq genes, curated subset (NM_*, NR_*, NP_* or YP_*) - Annotation Release GCF_000001405.25_GRCh37.p13 (2017-04-1
UCSC annotations of RefSeq RNAs (NM_* and NR_*)
Publications: Sequences in Scientific Articles
Gene Expression in 53 tissues from GTEx RNA-seq of 8555 samples (570 donors)
H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE
DNaseI Hypersensitivity Clusters in 125 cell types from ENCODE (V3)
Transcription Factor ChIP-seq Clusters (161 factors) from ENCODE with Factorbook Motifs
100 vertebrates Basewise Conservation by PhyloP
Multiz Alignments of 100 Vertebrates
Simple Nucleotide Polymorphisms (dbSNP 151) Found in >= 1% of Samples
Repeating Elements by RepeatMasker

- Each of these categories provides detailed information about the transcription factor, cell line, and a literature source of the information to provide the user with direct access for addressing their hypothesis.

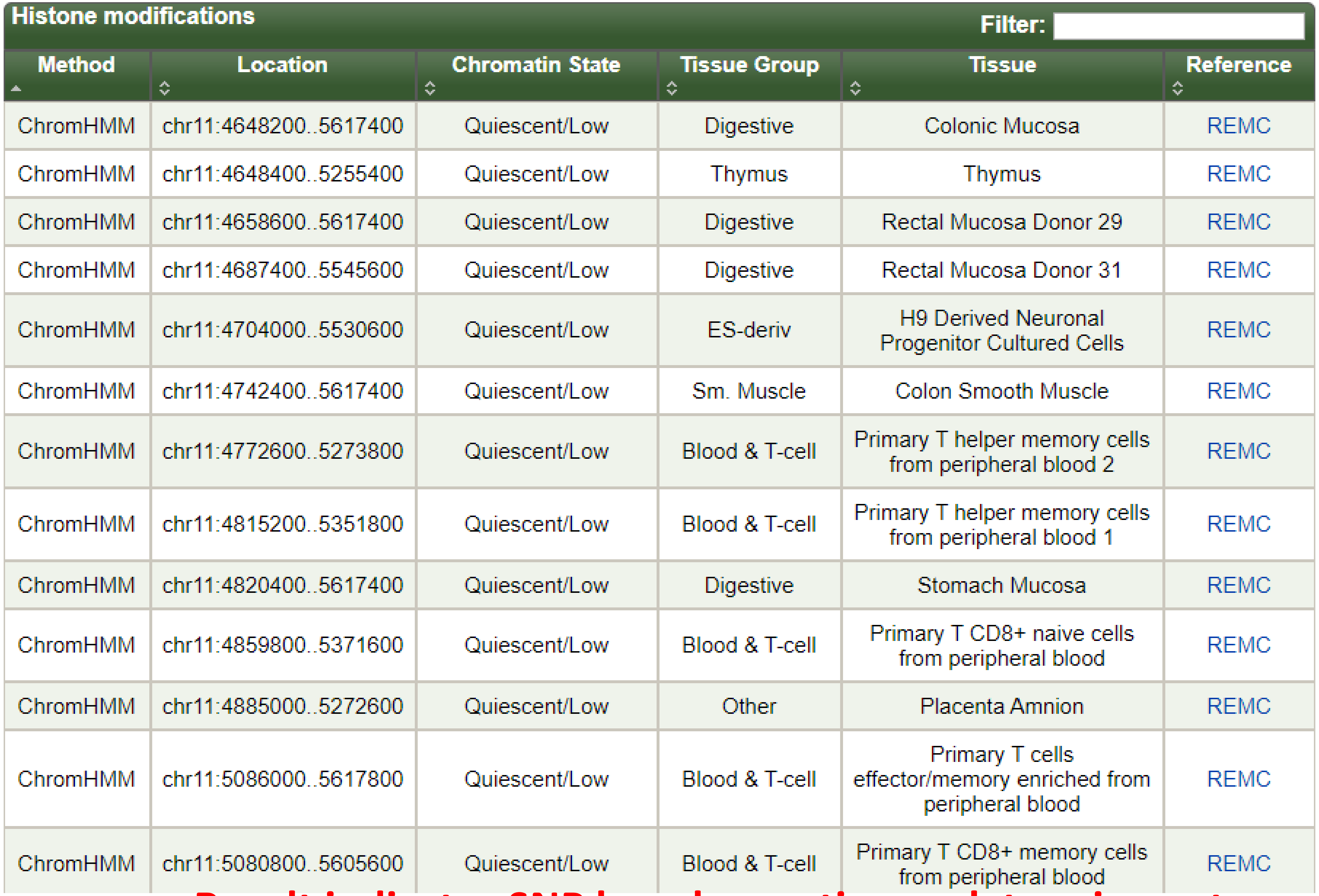


Motifs

Filter:

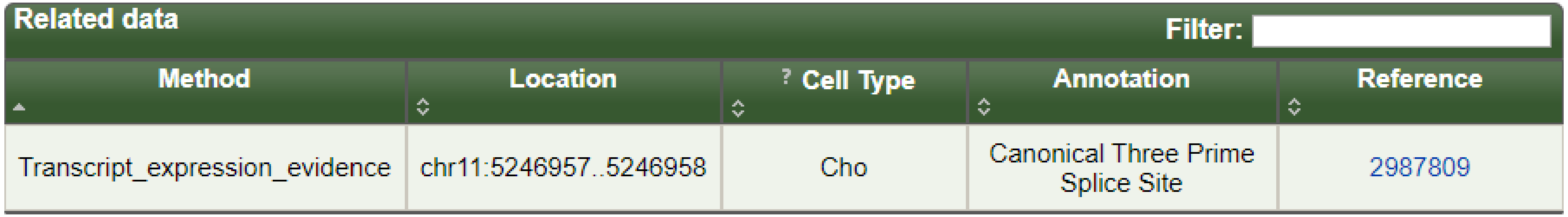
Method	Location	Motif	? Cell Type	PWM	Reference
Footprinting	chr11:5246956..5246974	Tal1::Gata1	K562		21106904
PWM	chr11:5246956..5246974	Tal1::Gata1			18006571

Result indicate SNP is present in Gata Motif which could have regulatory impact on the gene expresion



Method	Location	Chromatin State	Tissue Group	Tissue	Reference
ChromHMM	chr11:4648200..5617400	Quiescent/Low	Digestive	Colonic Mucosa	REMC
ChromHMM	chr11:4648400..5255400	Quiescent/Low	Thymus	Thymus	REMC
ChromHMM	chr11:4658600..5617400	Quiescent/Low	Digestive	Rectal Mucosa Donor 29	REMC
ChromHMM	chr11:4687400..5545600	Quiescent/Low	Digestive	Rectal Mucosa Donor 31	REMC
ChromHMM	chr11:4704000..5530600	Quiescent/Low	ES-deriv	H9 Derived Neuronal Progenitor Cultured Cells	REMC
ChromHMM	chr11:4742400..5617400	Quiescent/Low	Sm. Muscle	Colon Smooth Muscle	REMC
ChromHMM	chr11:4772600..5273800	Quiescent/Low	Blood & T-cell	Primary T helper memory cells from peripheral blood 2	REMC
ChromHMM	chr11:4815200..5351800	Quiescent/Low	Blood & T-cell	Primary T helper memory cells from peripheral blood 1	REMC
ChromHMM	chr11:4820400..5617400	Quiescent/Low	Digestive	Stomach Mucosa	REMC
ChromHMM	chr11:4859800..5371600	Quiescent/Low	Blood & T-cell	Primary T CD8+ naive cells from peripheral blood	REMC
ChromHMM	chr11:4885000..5272600	Quiescent/Low	Other	Placenta Amnion	REMC
ChromHMM	chr11:5086000..5617800	Quiescent/Low	Blood & T-cell	Primary T cells effector/memory enriched from peripheral blood	REMC
ChromHMM	chr11:5080800..5605600	Quiescent/Low	Blood & T-cell	Primary T CD8+ memory cells from peripheral blood	REMC

Result indicates SNP has chromatin regulatory impact



Related data

Filter:

Method	Location	? Cell Type	Annotation	Reference
Transcript_expression_evidence	chr11:5246957..5246958	Cho	Canonical Three Prime Splice Site	2987809

Result indicates SNP has expression in cho cell type and affect Splice site

Advantage of RegulomeDB

- **An integrated database to quickly generate prioritized hypotheses for the function of variants affecting both coding and noncoding regions in a genome by combining a large array of data sources into a single, integrated database.**

- **In particular, it include extensive information on annotated and computed regulatory elements in the human genome.**

- **Access to this novel approach via a simple and straightforward interface allows for easy query submission, and the scoring system provides for instant classification of significant variants.**

- **In addition, the SNV summary page will allow a user to quickly form a hypothesis as to the true functional consequence of a variant.**

- **While our examples deal with single nucleotide variants only, the database can also be used to annotate insertions and deletions.**

Comparision of HaploReg and RegulomeDB

- Ward and Kellis (2012) published the HaploReg database which aims to provide a similar annotation by providing an intersect of SNVs with chromatin state (Ernst and Kellis 2010).

- RegulomeDB database provides additional information well beyond this by prioritizing SNVs within general regulatory regions based on specific TF, chromatin, eQTL, and PWM information.

- Furthermore, RegulomeDB allow for a query of personal SNPs which account for a large proportion of variation in the population.

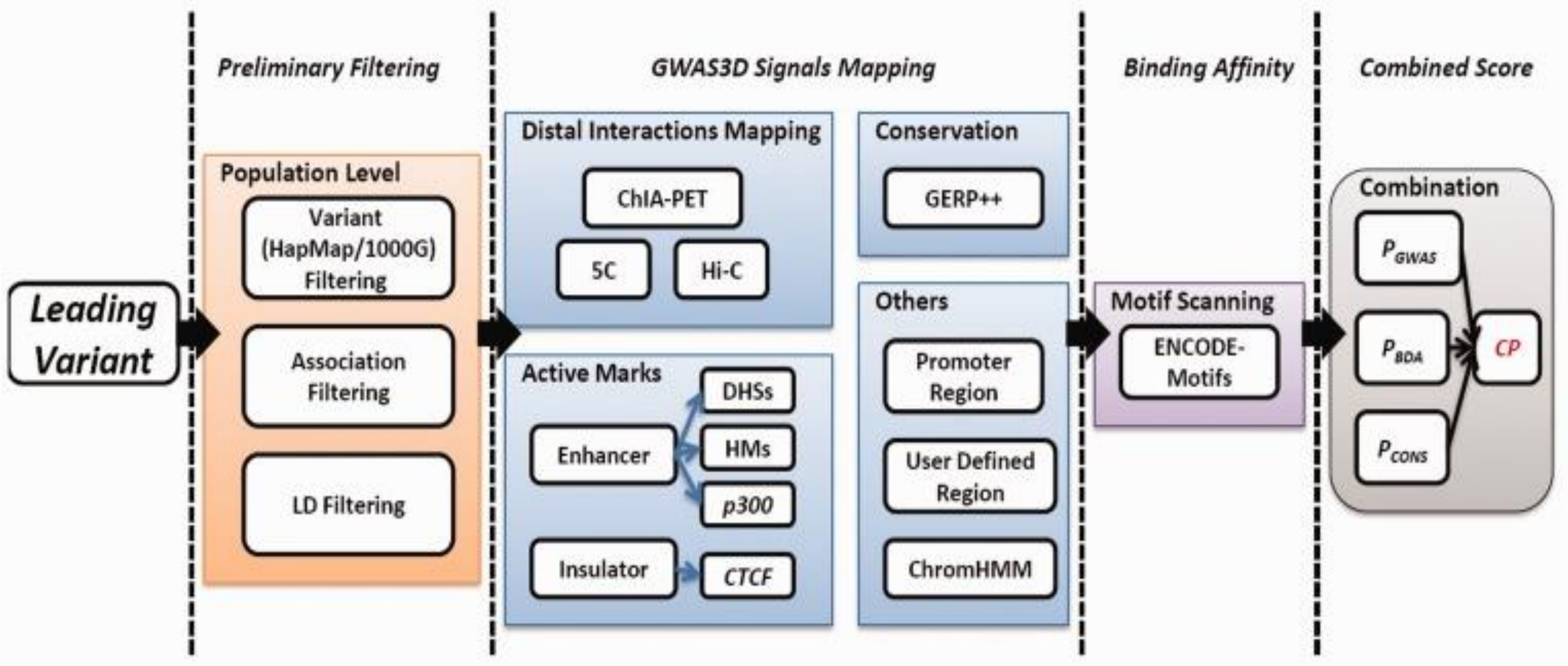
How many of these SNPs alter motifs sequence ?

rs4468290
rs11201609

GWAS3D/GWAS4D

- **GWAS3D: detecting human regulatory variants by integrative analysis of genome-wide associations, chromosome interactions and histone modifications**

http://mulinlab.tmu.edu.cn/gwas4d/gwas4d/gwas4d



From GWAS to Regulatory Function

- **Majority of GWAS risk loci localize to the noncoding genomic region with gene regulatory signal, suggesting that most trait/disease casual SNPs exert their phenotypic effects by altering gene expression. GWAS4D systematically analyzes GWAS summary data and identify context-specific regulatory variants by integrating latest multidimensional functional genomics resources and our recently published algorithms.**

Context-dependent Prediction

- **By incorporating roadmap 127 tissue/cell type-specific epigenomes data, GWAS4D uses joint likelihood framework to measure the regulatory probability of genetic variants in a context-dependent manner. It also estimates possible altered TFBSs using large-scale motif collections and annotates non-coding variant with comprehensive functional predictions.**

Link Variant to Target

Connecting non-coding variant to their gene targets under particular chromatin organization is crucial to understand variant regulatory mechanism. GWAS4D uniformly processes Hi-C data and reports significant interactions at 5kb resolution across tissues/cell types of multiple human organs and different development stages. It also equips a highly interactive visualization function for variant-target interaction.

Comparision with RegulomeDB and HaploReg

- **Compared with recent software and databases such as HaploReg and RegulomeDB, GWAS3D integrates more features and can be used in many scenarios.**

- **User can identify the most probable functional variant associated with interesting trait in one risk locus or prioritize the leading variants when given a full list of GWAS result or evaluate the deleteriousness of genetic variants affecting the gene regulation without any prior effect.**

- **GWAS3D also provides flexible configurations, such as human population, cell type specificity and TF family classification, for users to deal with different aspects of complex disease/trait. For example, user may select a matched cell type/tissue satisfying with a specific phenotype or manually define motifs of interested TFs used in following scanning when considering the tissue specificity of TFs.**
- **Recently, researchers found that the disease/trait-associated variants are highly related to active chromatin marks in relevant cell types. Therefore, these distinct features will greatly facilitate the discovery of regulatory variants under particular condition.**

Comparision with RegulomeDB and HaploReg

- **The computational process of our system is real-time, which is different from databases such as HaploReg and RegulomeDB, where the function annotations are pre-computed and stored in the database in advance.**

- **Therefore, it can dynamically deal with the genetic variants input by users with maximum flexibility.**

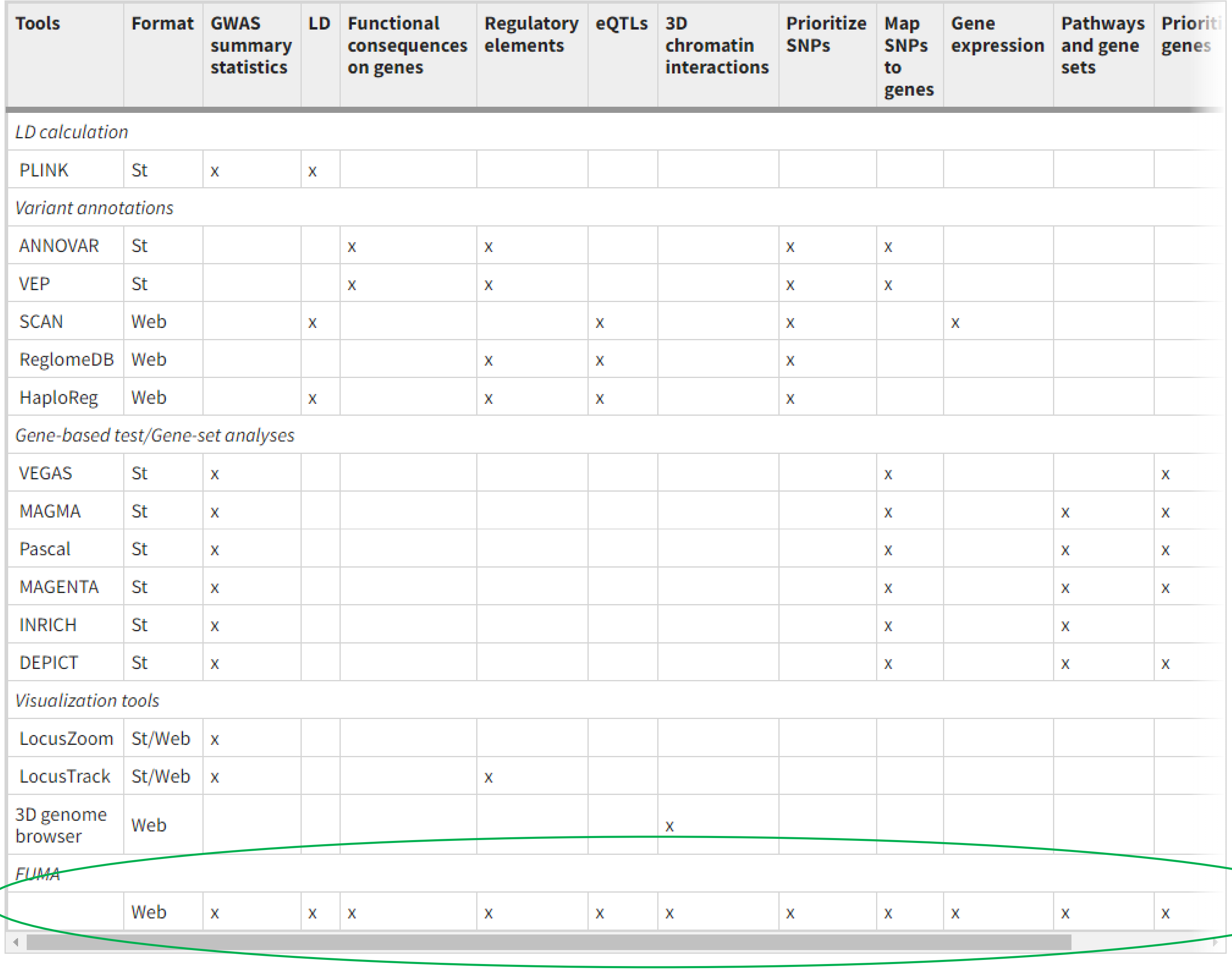
- **Despite large computational burden in the background when LD is considered, our system can finish the job of a meta GWAS data set (thousands of variants with moderate GWAS significance, $P < 1.0 \times 10^{-5}$) within a few hours even with LD from the 1000 Genomes Project. It will be much quicker when using HapMap LD.**

- **To exploit the regulatory properties of personal genomics data, GWAS3D accepts VCF-like format and can evaluate the deleteriousness of rare/novel variation altering gene regulation associated with personalized trait.**

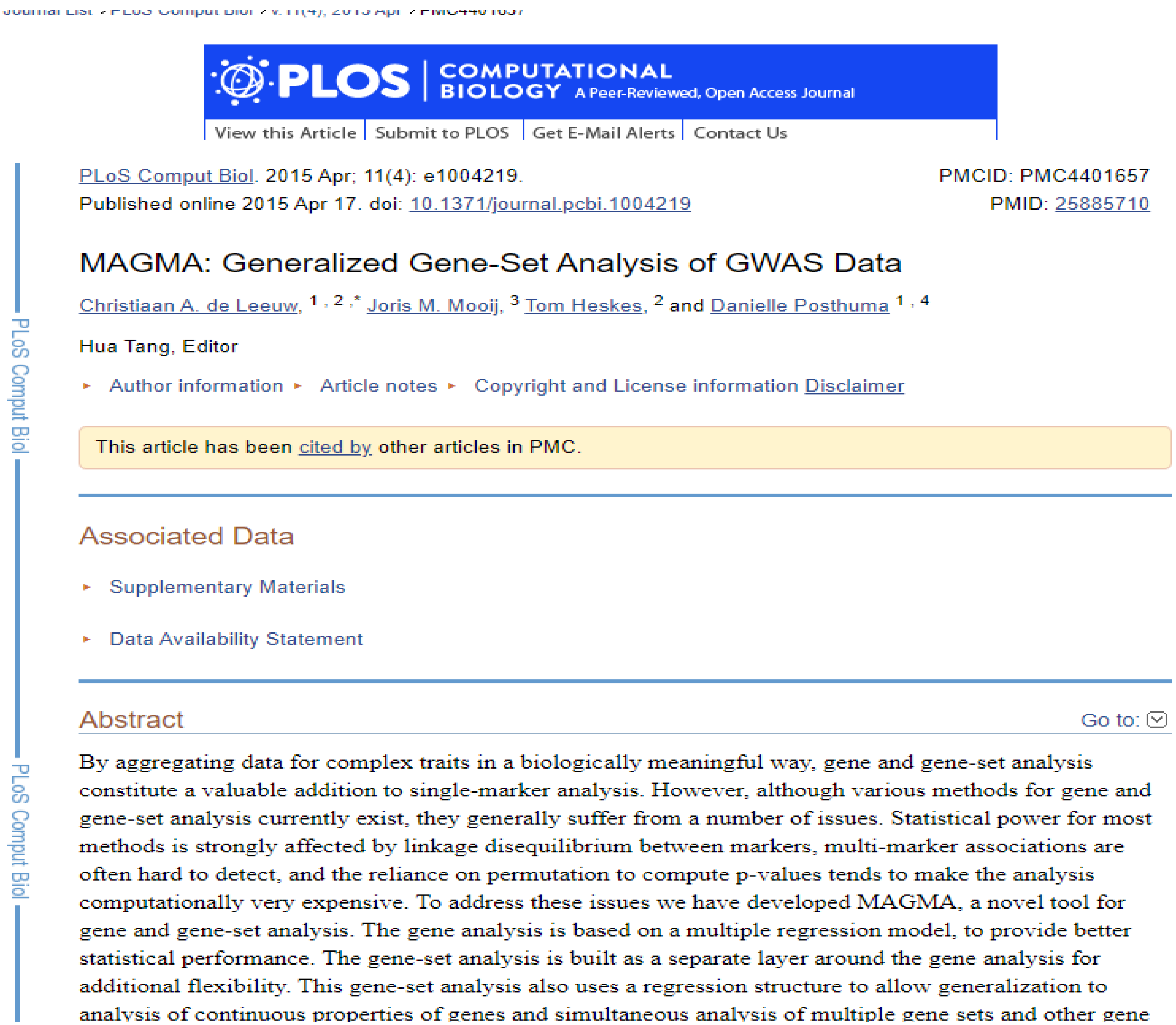
List of Tools

As discussed before

Analyses and visualization



Tools	Format	GWAS summary statistics	LD	Functional consequences on genes	Regulatory elements	eQTLs	3D chromatin interactions	Prioritize SNPs	Map SNPs to genes	Gene expression	Pathways and gene sets	Prioriti genes
LD calculation												
PLINK	St	x	x									
Variant annotations												
ANNOVAR	St			x	x			x	x			
VEP	St			x	x			x	x			
SCAN	Web		x			x		x		x		
ReglomeDB	Web				x	x		x				
HaploReg	Web		x		x	x		x				
Gene-based test/Gene-set analyses												
VEGAS	St	x							x			x
MAGMA	St	x							x		x	x
Pascal	St	x							x		x	x
MAGENTA	St	x							x		x	x
INRICH	St	x							x		x	
DEPICT	St	x							x		x	x
Visualization tools												
LocusZoom	St/Web	x										
LocusTrack	St/Web	x			x							
3D genome browser	Web						x					
FUMA												
	Web	x	x	x	x	x	x	x	x	x	x	x



Journal List > PLoS Comput Biol > v.11(4); 2015 Apr > PMC4401657
PLOS | COMPUTATIONAL BIOLOGY A Peer-Reviewed, Open Access Journal
View this Article
Submit to PLOS
Get E-Mail Alerts
Contact Us
PLoS Comput Biol
PLoS Comput Biol

PLoS Comput Biol. 2015 Apr; 11(4): e1004219.
Published online 2015 Apr 17. doi: 10.1371/journal.pcbi.1004219

PMCID: PMC4401657
PMID: 25885710

MAGMA: Generalized Gene-Set Analysis of GWAS Data

Christiaan A. de Leeuw,[1,2,*] Joris M. Mooij,[3] Tom Heskes,[2] and Danielle Posthuma[1,4]

Hua Tang, Editor

▸ Author information ▸ Article notes ▸ Copyright and License information Disclaimer

This article has been cited by other articles in PMC.

Associated Data

▸ Supplementary Materials

▸ Data Availability Statement

Abstract

Go to:

By aggregating data for complex traits in a biologically meaningful way, gene and gene-set analysis constitute a valuable addition to single-marker analysis. However, although various methods for gene and gene-set analysis currently exist, they generally suffer from a number of issues. Statistical power for most methods is strongly affected by linkage disequilibrium between markers, multi-marker associations are often hard to detect, and the reliance on permutation to compute p-values tends to make the analysis computationally very expensive. To address these issues we have developed MAGMA, a novel tool for gene and gene-set analysis. The gene analysis is based on a multiple regression model, to provide better statistical performance. The gene-set analysis is built as a separate layer around the gene analysis for additional flexibility. This gene-set analysis also uses a regression structure to allow generalization to analysis of continuous properties of genes and simultaneous analysis of multiple gene sets and other gene

Gene analysis

- The gene analysis in MAGMA is based on a multiple linear principal components regression model, using an F-test to compute the gene p-value.

- This model first projects the SNP matrix for a gene onto its principal components (PC), pruning away PCs with very small eigenvalues, and then uses those PCs as predictors for the phenotype in the linear regression model.

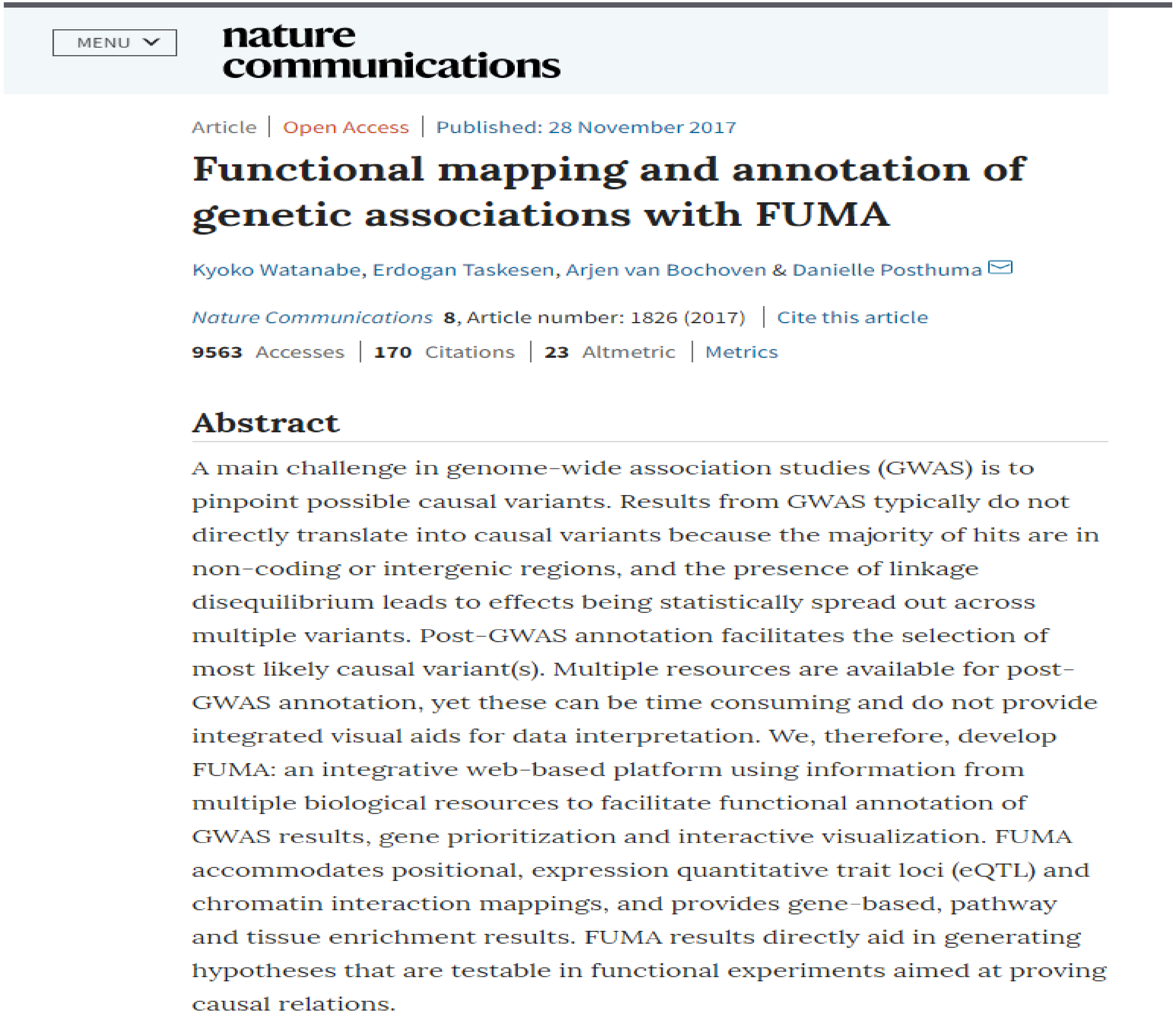
- This improves power by removing redundant parameters, and guarantees that the model is identifiable in the presence of highly collinear SNPs.

Gene-set analysis

- To perform the gene-set analysis, for each gene g the gene p-value p_g computed with the gene analysis is converted to a Z-value $z_g = \Phi^{-1}(1 - p_g)$, where Φ^{-1} is the probit function. This yields a roughly normally distributed variable Z with elements z_g that reflects the strength of the association each gene has with the phenotype, with higher values corresponding to stronger associations.

- Gene based and Gene set based analysis are included as feature of FUMA webserver

FUMA : interrogation of GWAS



nature communications

Article | Open Access | Published: 28 November 2017

Functional mapping and annotation of genetic associations with FUMA

Kyoko Watanabe, Erdogan Taskesen, Arjen van Bochoven & Danielle Posthuma ✉

Nature Communications **8**, Article number: 1826 (2017) | Cite this article

9563 Accesses | **170** Citations | **23** Altmetric | Metrics

Abstract

A main challenge in genome-wide association studies (GWAS) is to pinpoint possible causal variants. Results from GWAS typically do not directly translate into causal variants because the majority of hits are in non-coding or intergenic regions, and the presence of linkage disequilibrium leads to effects being statistically spread out across multiple variants. Post-GWAS annotation facilitates the selection of most likely causal variant(s). Multiple resources are available for post-GWAS annotation, yet these can be time consuming and do not provide integrated visual aids for data interpretation. We, therefore, develop FUMA: an integrative web-based platform using information from multiple biological resources to facilitate functional annotation of GWAS results, gene prioritization and interactive visualization. FUMA accommodates positional, expression quantitative trait loci (eQTL) and chromatin interaction mappings, and provides gene-based, pathway and tissue enrichment results. FUMA results directly aid in generating hypotheses that are testable in functional experiments aimed at proving causal relations.

http://fuma.ctglab.nl/

FUMA : Muti Steps

- The main purpose of FUMA is to use functional, biological information to prioritize genes based on GWAS outcomes.

- FUMA consists of two separate process; SNP2GENE and GENE2FUNC.

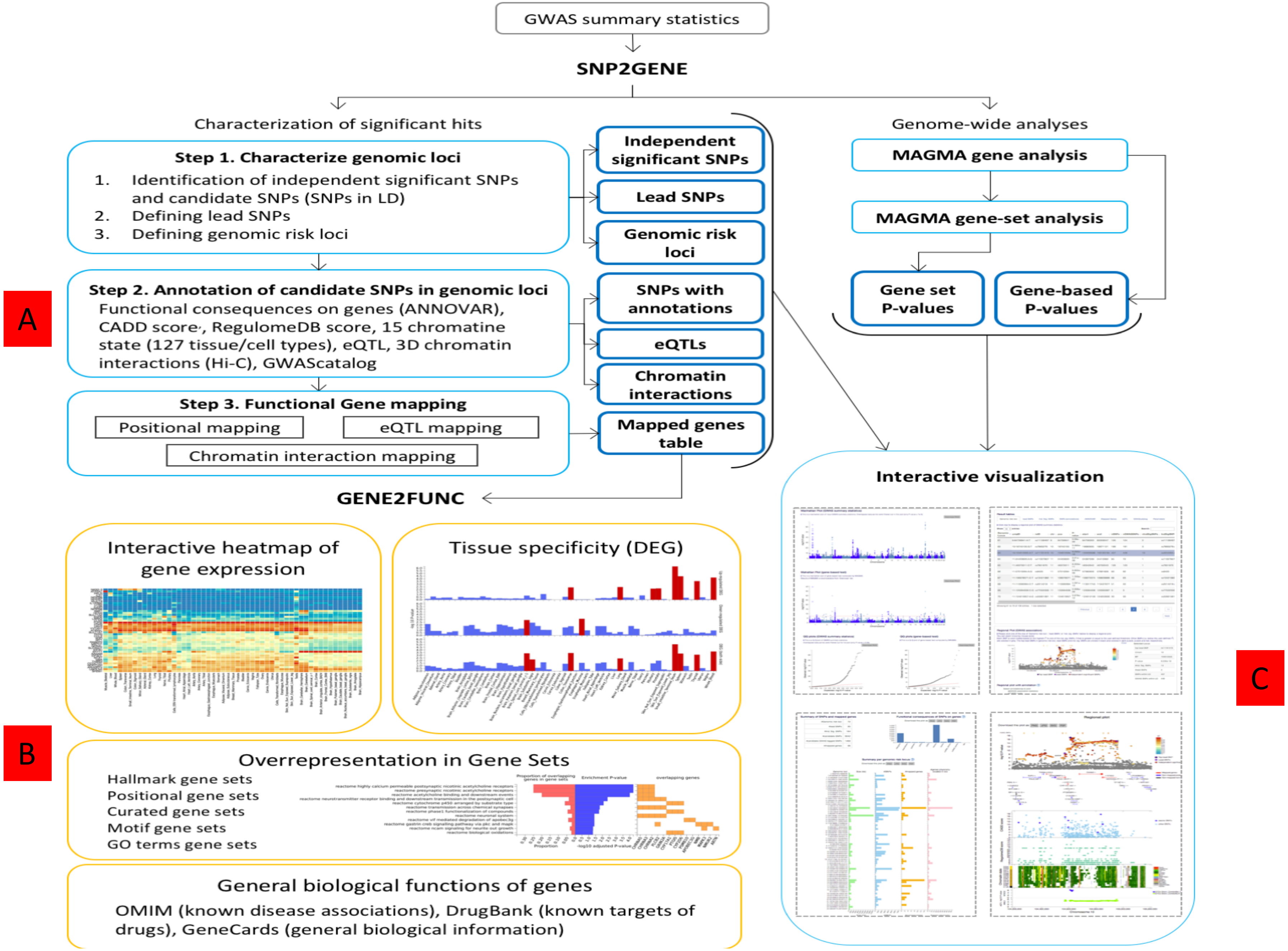
- To annotate and prioritize SNPs and genes from your GWAS summary statistics, go to SNP2GENE which compute LD structure, annotates functions to SNPs, and prioritize candidate genes.

- You can then use the prioritized genes as input to GENE2FUNC to check expression patterns and shared molecular functions between genes. GENE2FUNC can also be used for any list of pre-selected genes (i.e. created outside of SNP2GENE).

FUMA : Discuss

https://www.nature.com/articles/s41467-017-01261-5

Ready to use FUMA Webserver !!!



GWAS summary statistics

SNP2GENE

Characterization of significant hits

Step 1. Characterize genomic loci

1. Identification of independent significant SNPs and candidate SNPs (SNPs in LD)
2. Defining lead SNPs
3. Defining genomic risk loci

Step 2. Annotation of candidate SNPs in genomic loci

Functional consequences on genes (ANNOVAR), CADD score·, RegulomeDB score, 15 chromatine state (127 tissue/cell types), eQTL, 3D chromatin interactions (Hi-C), GWAScatalog

Step 3. Functional Gene mapping

Positional mapping

eQTL mapping

Chromatin interaction mapping

Independent significant SNPs

Lead SNPs

Genomic risk loci

SNPs with annotations

eQTLs

Chromatin interactions

Mapped genes table

Genome-wide analyses

MAGMA gene analysis

MAGMA gene-set analysis

Gene set P-values

Gene-based P-values

GENE2FUNC

Interactive heatmap of gene expression

Tissue specificity (DEG)

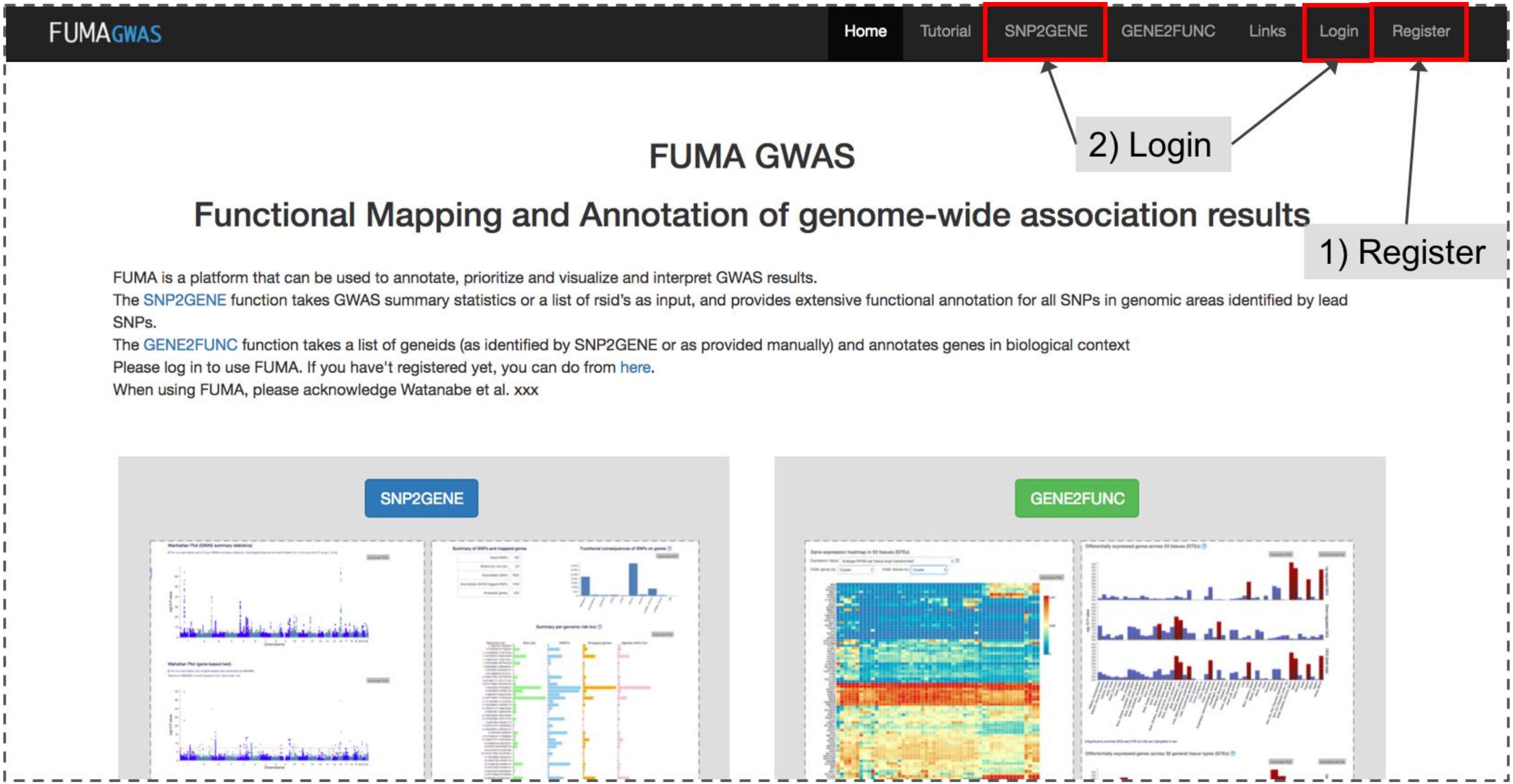
Overrepresentation in Gene Sets

Hallmark gene sets
Positional gene sets
Curated gene sets
Motif gene sets
GO terms gene sets

General biological functions of genes

OMIM (known disease associations), DrugBank (known targets of drugs), GeneCards (general biological information)

Interactive visualization



2) Login

1) Register

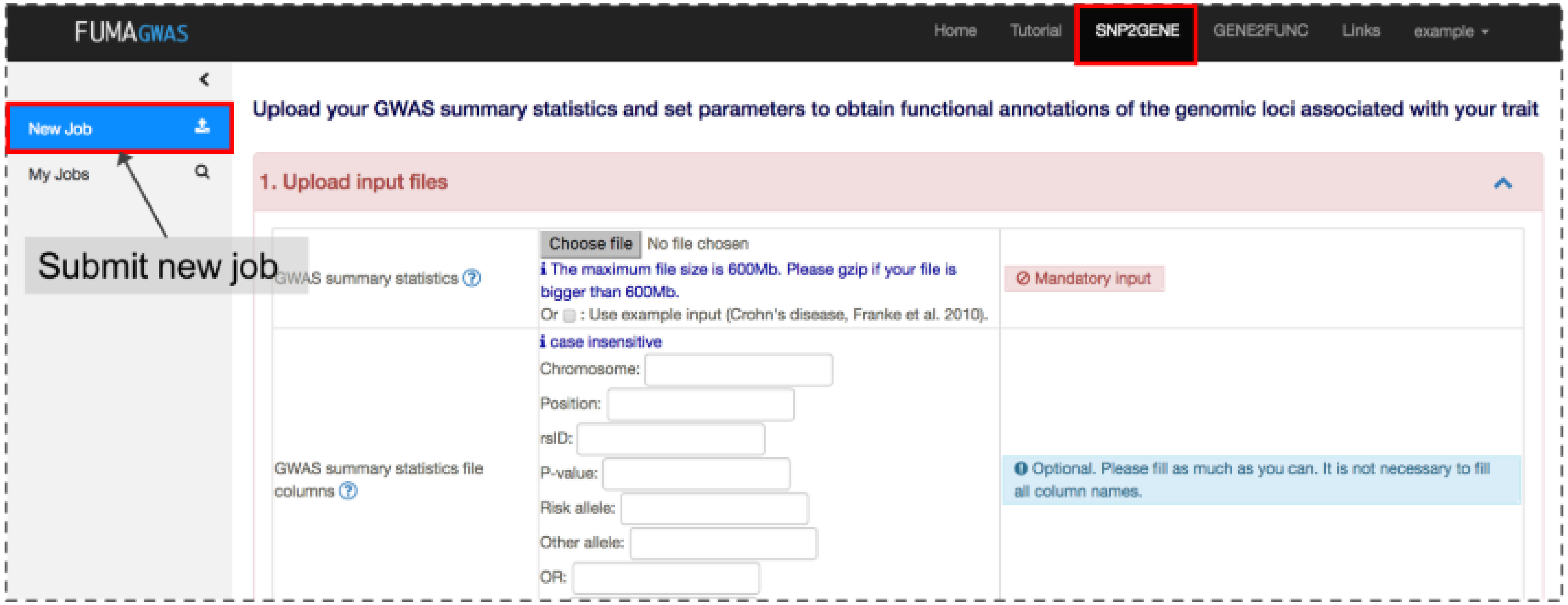
FUMA GWAS

Functional Mapping and Annotation of genome-wide association results

FUMA is a platform that can be used to annotate, prioritize and visualize and interpret GWAS results.
The SNP2GENE function takes GWAS summary statistics or a list of rsid's as input, and provides extensive functional annotation for all SNPs in genomic areas identified by lead SNPs.
The GENE2FUNC function takes a list of geneids (as identified by SNP2GENE or as provided manually) and annotates genes in biological context
Please log in to use FUMA. If you have't registered yet, you can do from here.
When using FUMA, please acknowledge Watanabe et al. xxx

2. Submit new job at SNP2GENE

A new job stats with a GWAS summary statistics file. A variety of file formats are supported. Please refer the section of Input files for details. If your input file is an output from PLINK, SNPTEST or METAL, you can directly submit the file without specifying column names.
The input GWAS summary statistics file could be a subset of SNPs (e.g. only SNPs which are interesting in your study), but in this case, MAGMA results are not relevant anymore.
Optionally, if you would like to pre-specify lead SNPs, you can upload a file with 3 columns; rsID, chromosome and position. FUMA will then use these SNPs to select LD-related SNPs for annotation and mapping, instead of using lead SNPs identified by FUMA (it requires to disable an option for "identify additional lead SNPs").
In addition, if you are interested in specific genomic regions, you can also provide them by uploading a file with 3 columns; chromosome, start and end position. FUMA will then use these genomic regions to select LD-related SNPs for annotation and mapping, instead of determining the regions itself.



Upload your GWAS summary statistics and set parameters to obtain functional annotations of the genomic loci associated with your trait

1. Upload input files

GWAS summary statistics ⓘ

No file chosen
The maximum file size is 600Mb. Please gzip if your file is bigger than 600Mb.
Or □ : Use example input (Crohn's disease, Franke et al. 2010).

Mandatory input

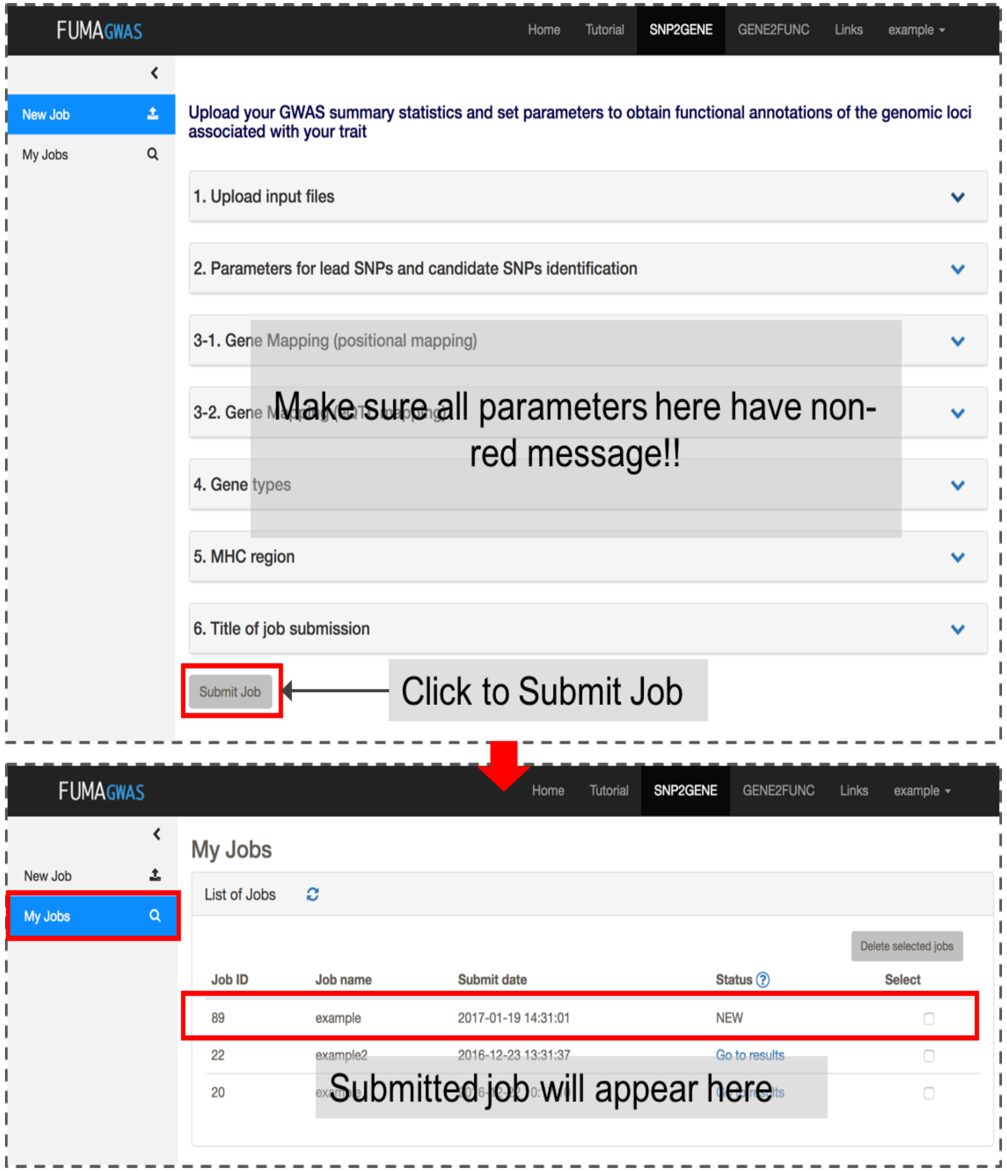
GWAS summary statistics file columns ⓘ

case insensitive
Chromosome:
Position:
rsID:
P-value:
Risk allele:
Other allele:
OR:

Optional. Please fill as much as you can. It is not necessary to fill all column names.

3. Set parameters

- On the same page as where you specify the input files, there are a variety of optional parameters that control the prioritization of genes.
- Please check your parameters carefully. The default settings are to perform identification of independent genome-wide significant SNPs at r^2 0.6 and lead SNPs at r^2 0.1, to maps SNPs to genes up to 10kb apart.
- To filter SNPs by specific functional annotations and to use eQTL mapping, please change parameters
- If all inputs are valid, 'Submit Job' button will be activated. Once you submit a job, this will be listed in My Jobs.



Upload your GWAS summary statistics and set parameters to obtain functional annotations of the genomic loci associated with your trait

1. Upload input files
2. Parameters for lead SNPs and candidate SNPs identification
3-1. Gene Mapping (positional mapping)
3-2. Gene Mapping (eQTL mapping)
4. Gene types
5. MHC region
6. Title of job submission

Make sure all parameters here have non-red message!!

Click to Submit Job

My Jobs

List of Jobs

Job ID	Job name	Submit date	Status	Select
89	example	2017-01-19 14:31:01	NEW	☐
22	example2	2016-12-23 13:31:37	Go to results	☐
20				☐

Submitted job will appear here

4. Check your results

After you submit files and parameter settings, a JOB has the status NEW which will be updated to QUEUES to RUNNING. Depending on the number of significant genomic regions, this may take between a couple of minutes and an hour. Once a JOB has finished running, you will receive an email. Unless an error occurred during the process, the email includes the link to the result page (this again requires login). You can also access to the results page from My Jobs page.

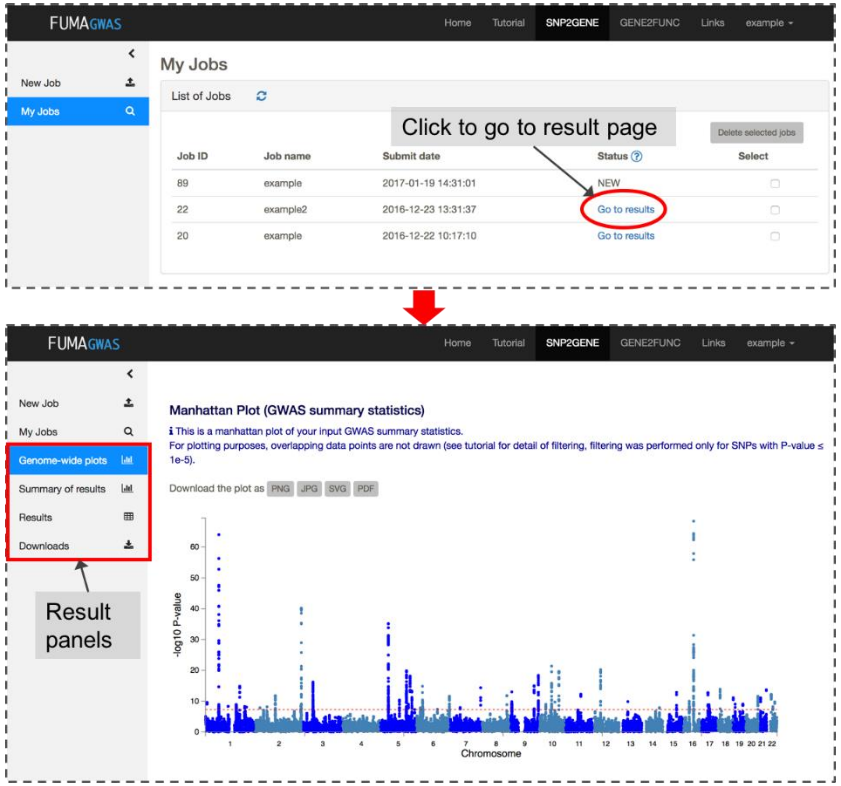
The result page displays 4 additional side bars.

Genome-wide plots: Manhattan plots and Q-Q plots for GWAS summary statistics and gene-based test by MAGMA, results of MAGMA gene-set analysis and tissue expression analysis.

Summary of results: Summary of results such as the number of lead and LD-related SNPs, and mapped genes for overall and per identified genomic risk locus.

Results: Tables of lead SNPs, genomic risk loci, candidate SNPs with annotations, eQTLs (only when eQTL mapping is performed), mapped genes and GWAS-catalog reported SNPs matched with candidate SNPs. You can also create interactive regional plots with functional annotations from this tab.

Downloads: Download all results as text files.



My Jobs

List of Jobs

Click to go to result page

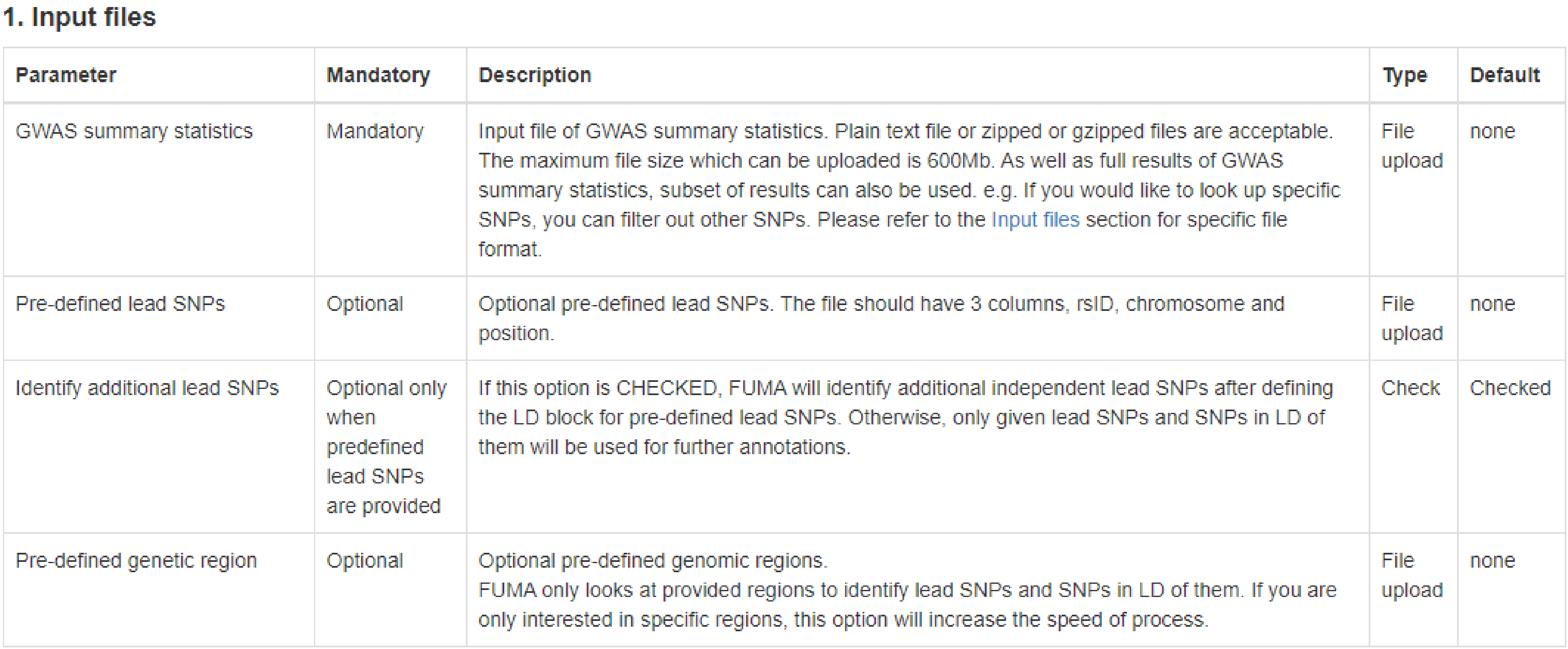
Job ID	Job name	Submit date	Status	Select
89	example	2017-01-19 14:31:01	NEW	□
22	example2	2016-12-23 13:31:37	Go to results	□
20	example	2016-12-22 10:17:10	Go to results	□

Manhattan Plot (GWAS summary statistics)

ℹ This is a manhattan plot of your input GWAS summary statistics.
For plotting purposes, overlapping data points are not drawn (see tutorial for detail of filtering, filtering was performed only for SNPs with P-value ≤ 1e-5).

Download the plot as

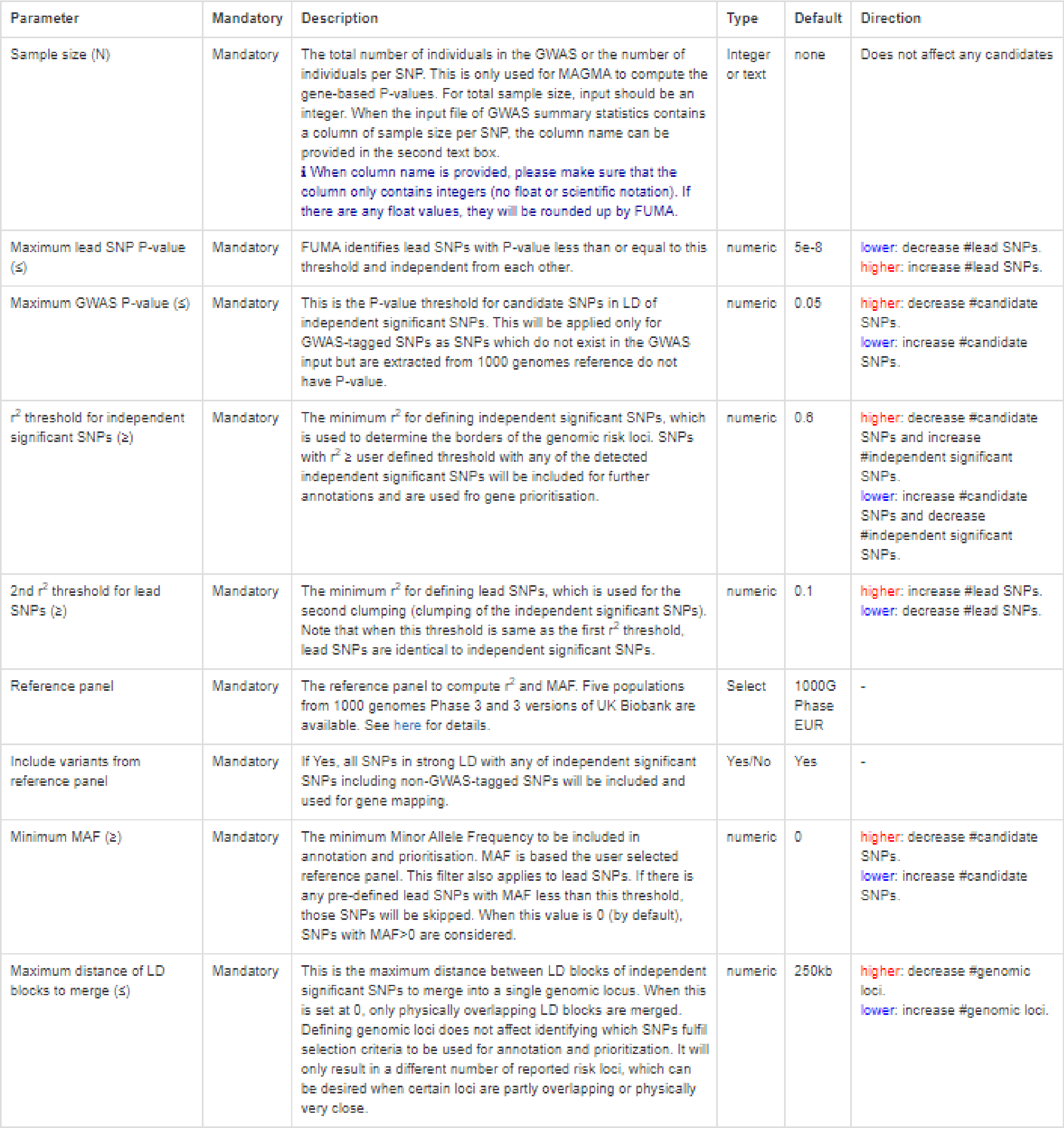
Result panels



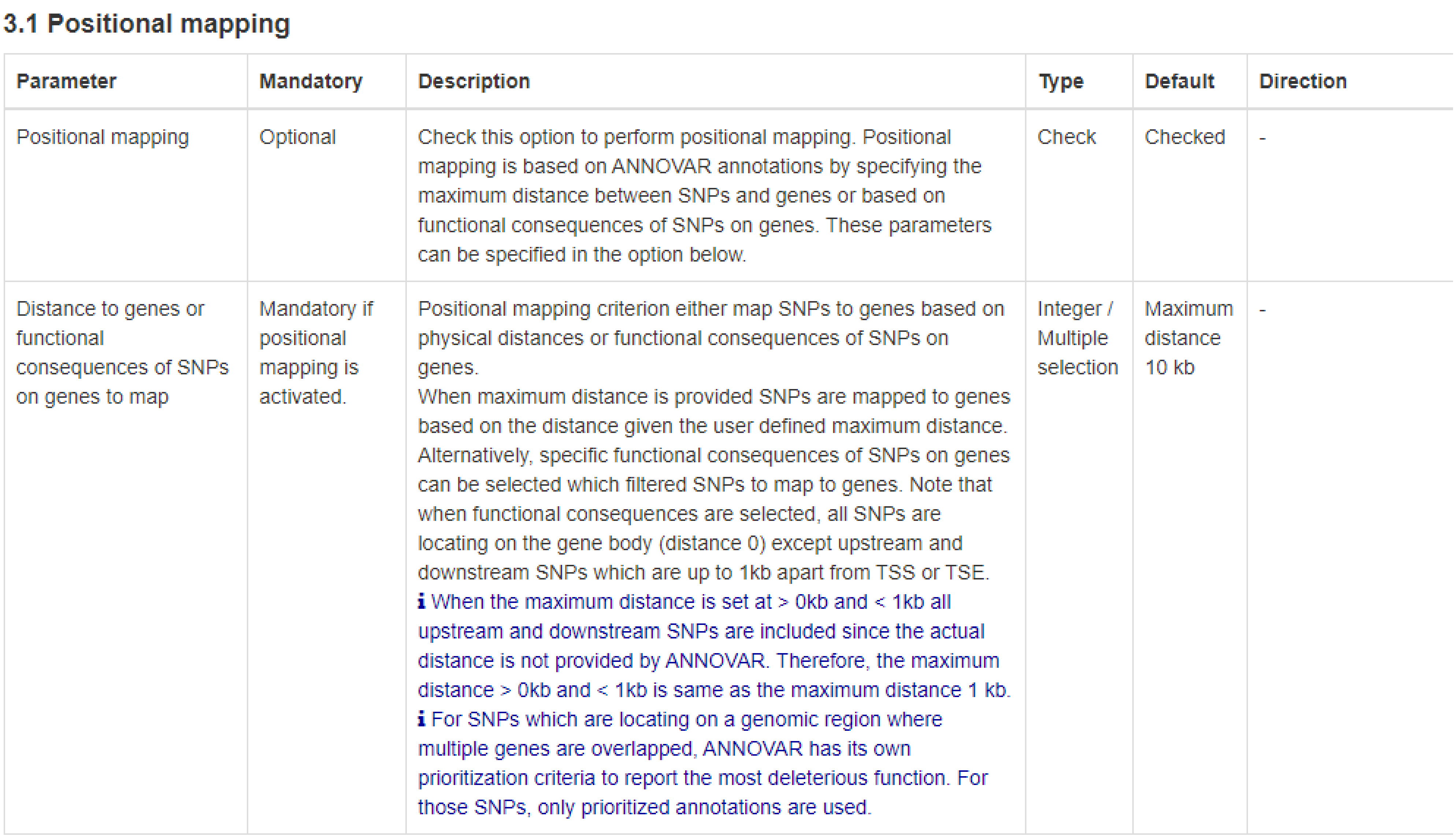
1. Input files

Parameter	Mandatory	Description	Type	Default
GWAS summary statistics	Mandatory	Input file of GWAS summary statistics. Plain text file or zipped or gzipped files are acceptable. The maximum file size which can be uploaded is 600Mb. As well as full results of GWAS summary statistics, subset of results can also be used. e.g. If you would like to look up specific SNPs, you can filter out other SNPs. Please refer to the Input files section for specific file format.	File upload	none
Pre-defined lead SNPs	Optional	Optional pre-defined lead SNPs. The file should have 3 columns, rsID, chromosome and position.	File upload	none
Identify additional lead SNPs	Optional only when predefined lead SNPs are provided	If this option is CHECKED, FUMA will identify additional independent lead SNPs after defining the LD block for pre-defined lead SNPs. Otherwise, only given lead SNPs and SNPs in LD of them will be used for further annotations.	Check	Checked
Pre-defined genetic region	Optional	Optional pre-defined genomic regions. FUMA only looks at provided regions to identify lead SNPs and SNPs in LD of them. If you are only interested in specific regions, this option will increase the speed of process.	File upload	none

FUMA : Parameter detail

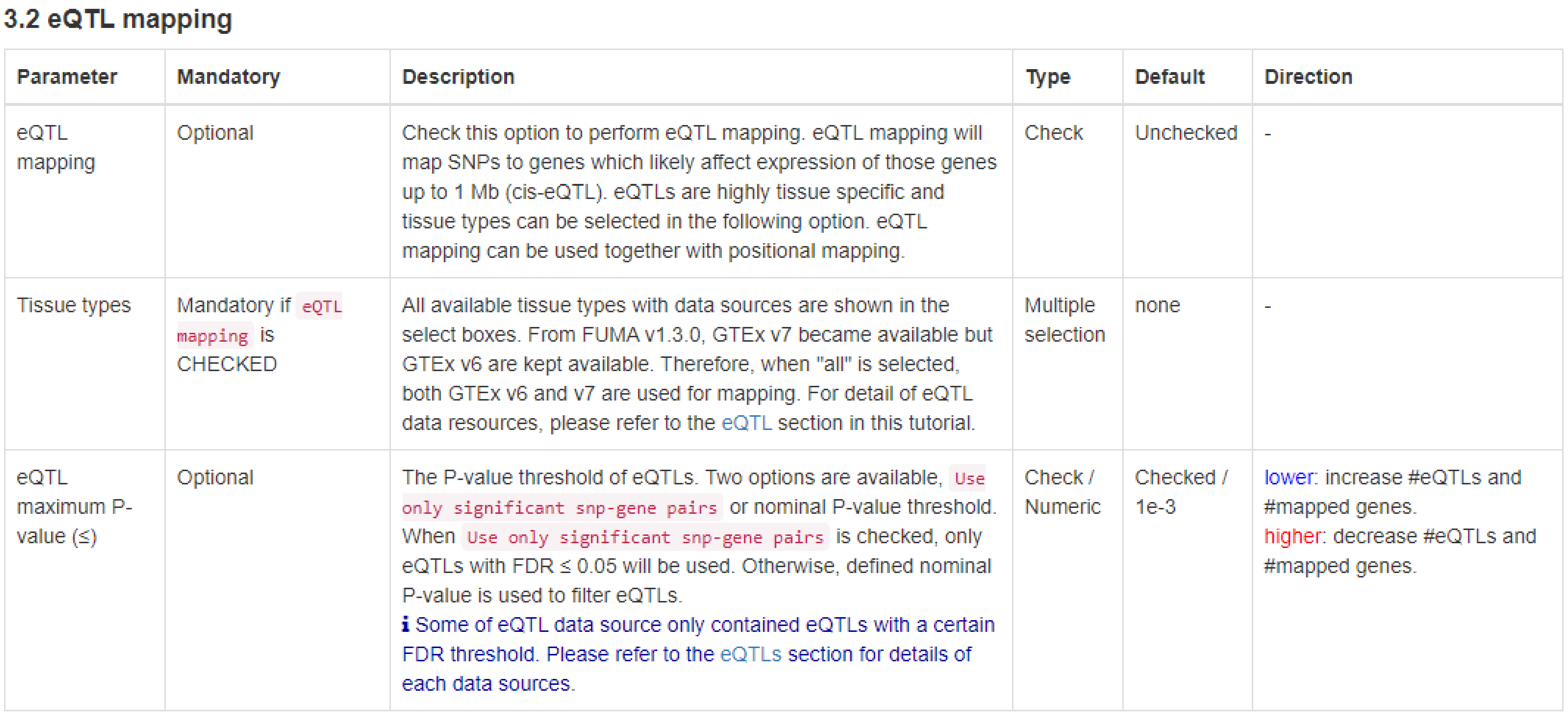


Parameter	Mandatory	Description	Type	Default	Direction
Sample size (N)	Mandatory	The total number of individuals in the GWAS or the number of individuals per SNP. This is only used for MAGMA to compute the gene-based P-values. For total sample size, input should be an integer. When the input file of GWAS summary statistics contains a column of sample size per SNP, the column name can be provided in the second text box. ℹ When column name is provided, please make sure that the column only contains integers (no float or scientific notation). If there are any float values, they will be rounded up by FUMA.	Integer or text	none	Does not affect any candidates
Maximum lead SNP P-value (≤)	Mandatory	FUMA identifies lead SNPs with P-value less than or equal to this threshold and independent from each other.	numeric	5e-8	lower: decrease #lead SNPs. higher: increase #lead SNPs.
Maximum GWAS P-value (≤)	Mandatory	This is the P-value threshold for candidate SNPs in LD of independent significant SNPs. This will be applied only for GWAS-tagged SNPs as SNPs which do not exist in the GWAS input but are extracted from 1000 genomes reference do not have P-value.	numeric	0.05	higher: decrease #candidate SNPs. lower: increase #candidate SNPs.
r^2 threshold for independent significant SNPs (≥)	Mandatory	The minimum r^2 for defining independent significant SNPs, which is used to determine the borders of the genomic risk loci. SNPs with r^2 ≥ user defined threshold with any of the detected independent significant SNPs will be included for further annotations and are used fro gene prioritisation.	numeric	0.6	higher: decrease #candidate SNPs and increase #independent significant SNPs. lower: increase #candidate SNPs and decrease #independent significant SNPs.
2nd r^2 threshold for lead SNPs (≥)	Mandatory	The minimum r^2 for defining lead SNPs, which is used for the second clumping (clumping of the independent significant SNPs). Note that when this threshold is same as the first r^2 threshold, lead SNPs are identical to independent significant SNPs.	numeric	0.1	higher: increase #lead SNPs. lower: decrease #lead SNPs.
Reference panel	Mandatory	The reference panel to compute r^2 and MAF. Five populations from 1000 genomes Phase 3 and 3 versions of UK Biobank are available. See here for details.	Select	1000G Phase3 EUR	-
Include variants from reference panel	Mandatory	If Yes, all SNPs in strong LD with any of independent significant SNPs including non-GWAS-tagged SNPs will be included and used for gene mapping.	Yes/No	Yes	-
Minimum MAF (≥)	Mandatory	The minimum Minor Allele Frequency to be included in annotation and prioritisation. MAF is based the user selected reference panel. This filter also applies to lead SNPs. If there is any pre-defined lead SNPs with MAF less than this threshold, those SNPs will be skipped. When this value is 0 (by default), SNPs with MAF>0 are considered.	numeric	0	higher: decrease #candidate SNPs. lower: increase #candidate SNPs.
Maximum distance of LD blocks to merge (≤)	Mandatory	This is the maximum distance between LD blocks of independent significant SNPs to merge into a single genomic locus. When this is set at 0, only physically overlapping LD blocks are merged. Defining genomic loci does not affect identifying which SNPs fulfil selection criteria to be used for annotation and prioritization. It will only result in a different number of reported risk loci, which can be desired when certain loci are partly overlapping or physically very close.	numeric	250kb	higher: decrease #genomic loci. lower: increase #genomic loci.



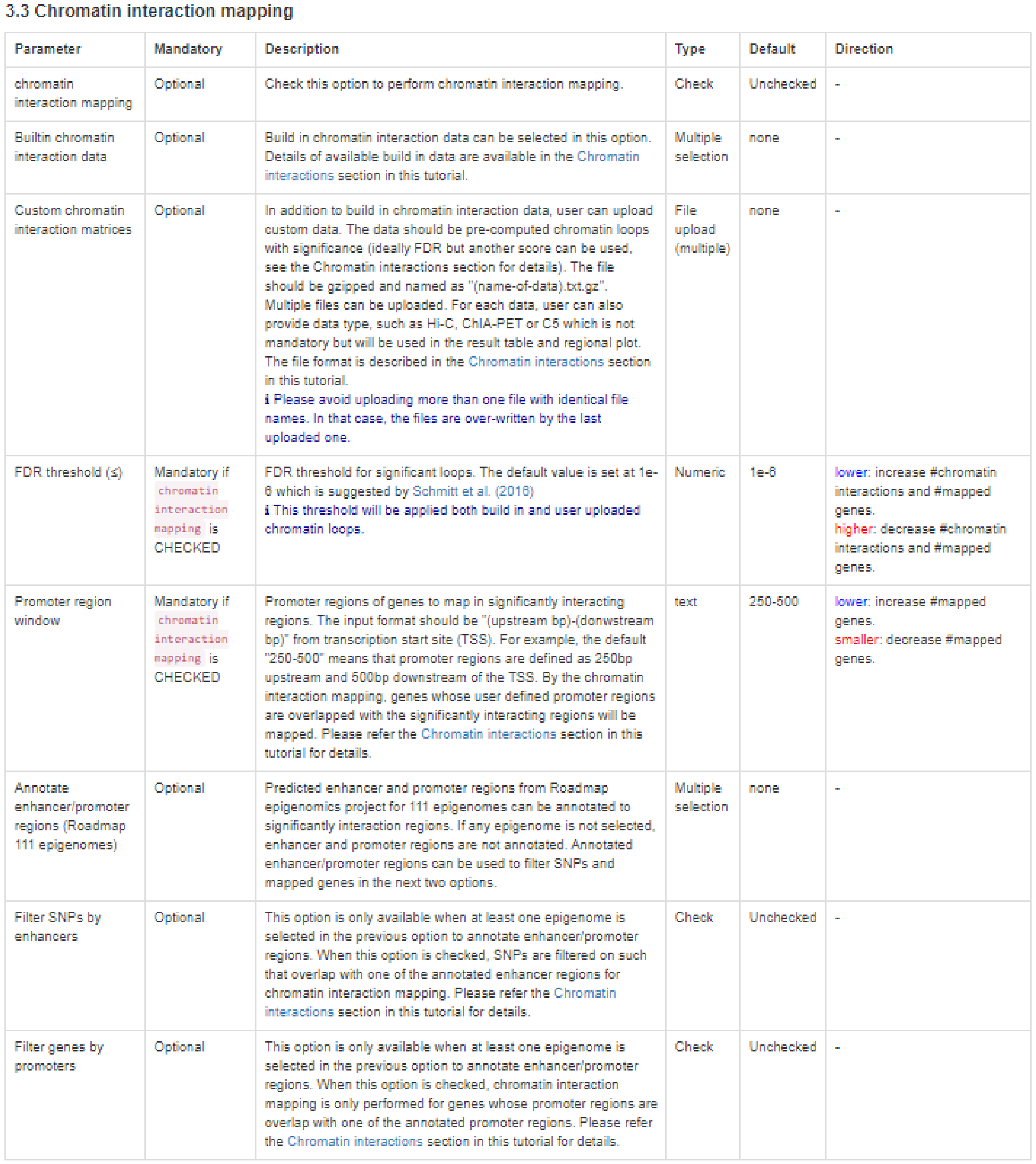
3.1 Positional mapping

Parameter	Mandatory	Description	Type	Default	Direction
Positional mapping	Optional	Check this option to perform positional mapping. Positional mapping is based on ANNOVAR annotations by specifying the maximum distance between SNPs and genes or based on functional consequences of SNPs on genes. These parameters can be specified in the option below.	Check	Checked	-
Distance to genes or functional consequences of SNPs on genes to map	Mandatory if positional mapping is activated.	Positional mapping criterion either map SNPs to genes based on physical distances or functional consequences of SNPs on genes. When maximum distance is provided SNPs are mapped to genes based on the distance given the user defined maximum distance. Alternatively, specific functional consequences of SNPs on genes can be selected which filtered SNPs to map to genes. Note that when functional consequences are selected, all SNPs are locating on the gene body (distance 0) except upstream and downstream SNPs which are up to 1kb apart from TSS or TSE. ℹ When the maximum distance is set at > 0kb and < 1kb all upstream and downstream SNPs are included since the actual distance is not provided by ANNOVAR. Therefore, the maximum distance > 0kb and < 1kb is same as the maximum distance 1 kb. ℹ For SNPs which are locating on a genomic region where multiple genes are overlapped, ANNOVAR has its own prioritization criteria to report the most deleterious function. For those SNPs, only prioritized annotations are used.	Integer / Multiple selection	Maximum distance 10 kb	-



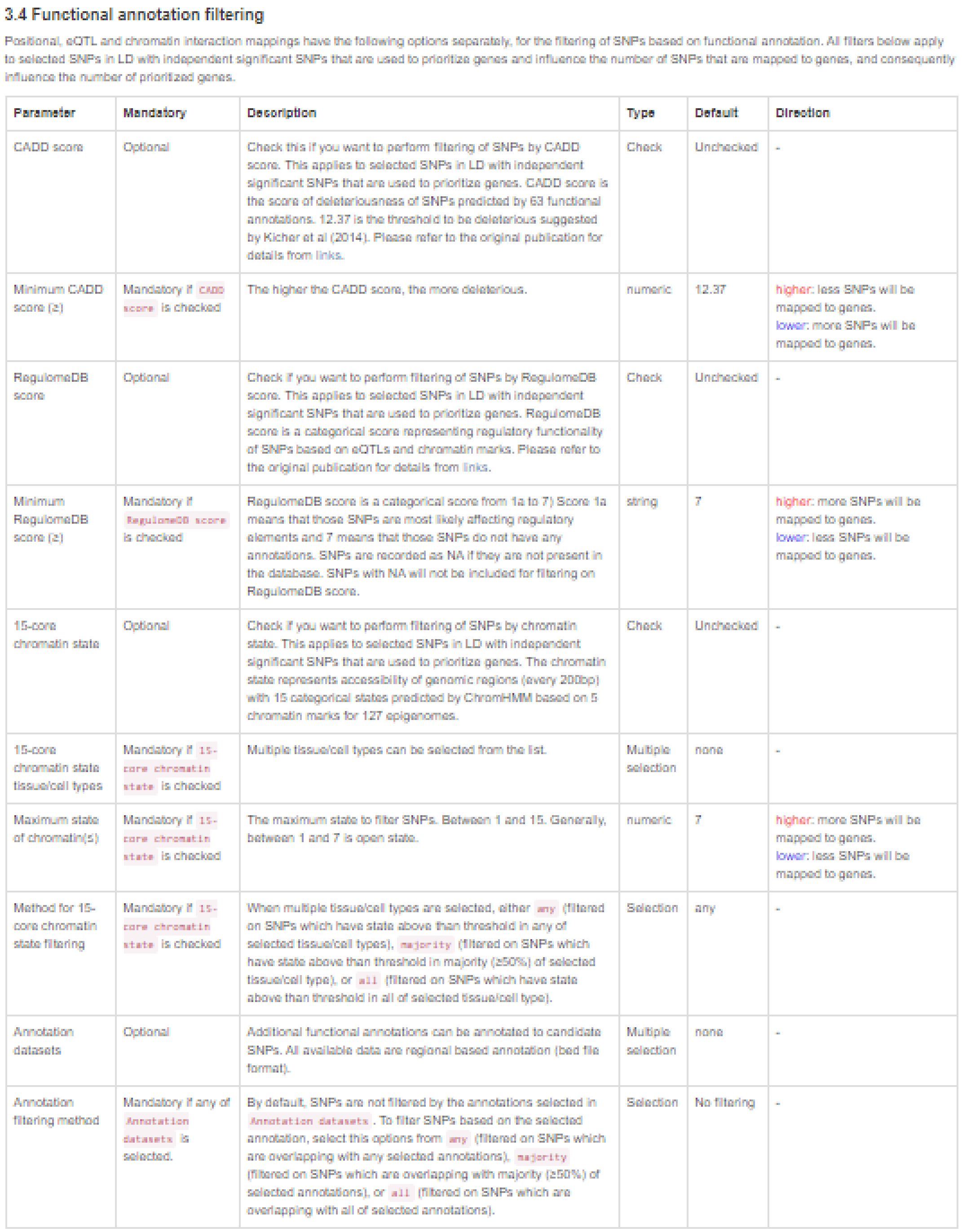
3.2 eQTL mapping

Parameter	Mandatory	Description	Type	Default	Direction
eQTL mapping	Optional	Check this option to perform eQTL mapping. eQTL mapping will map SNPs to genes which likely affect expression of those genes up to 1 Mb (cis-eQTL). eQTLs are highly tissue specific and tissue types can be selected in the following option. eQTL mapping can be used together with positional mapping.	Check	Unchecked	-
Tissue types	Mandatory if `eQTL mapping` is CHECKED	All available tissue types with data sources are shown in the select boxes. From FUMA v1.3.0, GTEx v7 became available but GTEx v6 are kept available. Therefore, when "all" is selected, both GTEx v6 and v7 are used for mapping. For detail of eQTL data resources, please refer to the eQTL section in this tutorial.	Multiple selection	none	-
eQTL maximum P-value (≤)	Optional	The P-value threshold of eQTLs. Two options are available, `Use only significant snp-gene pairs` or nominal P-value threshold. When `Use only significant snp-gene pairs` is checked, only eQTLs with FDR ≤ 0.05 will be used. Otherwise, defined nominal P-value is used to filter eQTLs. ℹ Some of eQTL data source only contained eQTLs with a certain FDR threshold. Please refer to the eQTLs section for details of each data sources.	Check / Numeric	Checked / 1e-3	lower: increase #eQTLs and #mapped genes. higher: decrease #eQTLs and #mapped genes.



3.3 Chromatin interaction mapping

Parameter	Mandatory	Description	Type	Default	Direction
chromatin interaction mapping	Optional	Check this option to perform chromatin interaction mapping.	Check	Unchecked	-
Builtin chromatin interaction data	Optional	Build in chromatin interaction data can be selected in this option. Details of available build in data are available in the Chromatin interactions section in this tutorial.	Multiple selection	none	-
Custom chromatin interaction matrices	Optional	In addition to build in chromatin interaction data, user can upload custom data. The data should be pre-computed chromatin loops with significance (ideally FDR but another score can be used, see the Chromatin interactions section for details). The file should be gzipped and named as "(name-of-data).txt.gz". Multiple files can be uploaded. For each data, user can also provide data type, such as Hi-C, ChIA-PET or C5 which is not mandatory but will be used in the result table and regional plot. The file format is described in the Chromatin interactions section in this tutorial. ℹ Please avoid uploading more than one file with identical file names. In that case, the files are over-written by the last uploaded one.	File upload (multiple)	none	-
FDR threshold (≤)	Mandatory if `chromatin interaction mapping` is CHECKED	FDR threshold for significant loops. The default value is set at 1e-6 which is suggested by Schmitt et al. (2016) ℹ This threshold will be applied both build in and user uploaded chromatin loops.	Numeric	1e-6	lower: increase #chromatin interactions and #mapped genes. higher: decrease #chromatin interactions and #mapped genes.
Promoter region window	Mandatory if `chromatin interaction mapping` is CHECKED	Promoter regions of genes to map in significantly interacting regions. The input format should be "(upstream bp)-(donwstream bp)" from transcription start site (TSS). For example, the default "250-500" means that promoter regions are defined as 250bp upstream and 500bp downstream of the TSS. By the chromatin interaction mapping, genes whose user defined promoter regions are overlapped with the significantly interacting regions will be mapped. Please refer the Chromatin interactions section in this tutorial for details.	text	250-500	lower: increase #mapped genes. smaller: decrease #mapped genes.
Annotate enhancer/promoter regions (Roadmap 111 epigenomes)	Optional	Predicted enhancer and promoter regions from Roadmap epigenomics project for 111 epigenomes can be annotated to significantly interaction regions. If any epigenome is not selected, enhancer and promoter regions are not annotated. Annotated enhancer/promoter regions can be used to filter SNPs and mapped genes in the next two options.	Multiple selection	none	-
Filter SNPs by enhancers	Optional	This option is only available when at least one epigenome is selected in the previous option to annotate enhancer/promoter regions. When this option is checked, SNPs are filtered on such that overlap with one of the annotated enhancer regions for chromatin interaction mapping. Please refer the Chromatin interactions section in this tutorial for details.	Check	Unchecked	-
Filter genes by promoters	Optional	This option is only available when at least one epigenome is selected in the previous option to annotate enhancer/promoter regions. When this option is checked, chromatin interaction mapping is only performed for genes whose promoter regions are overlap with one of the annotated promoter regions. Please refer the Chromatin interactions section in this tutorial for details.	Check	Unchecked	-



3.4 Functional annotation filtering

Positional, eQTL and chromatin interaction mappings have the following options separately, for the filtering of SNPs based on functional annotation. All filters below apply to selected SNPs in LD with independent significant SNPs that are used to prioritize genes and influence the number of SNPs that are mapped to genes, and consequently influence the number of prioritized genes.

Parameter	Mandatory	Description	Type	Default	Direction
CADD score	Optional	Check this if you want to perform filtering of SNPs by CADD score. This applies to selected SNPs in LD with independent significant SNPs that are used to prioritize genes. CADD score is the score of deleteriousness of SNPs predicted by 63 functional annotations. 12.37 is the threshold to be deleterious suggested by Kicher et al (2014). Please refer to the original publication for details from links.	Check	Unchecked	-
Minimum CADD score (≥)	Mandatory if `CADD score` is checked	The higher the CADD score, the more deleterious.	numeric	12.37	higher: less SNPs will be mapped to genes. lower: more SNPs will be mapped to genes.
RegulomeDB score	Optional	Check if you want to perform filtering of SNPs by RegulomeDB score. This applies to selected SNPs in LD with independent significant SNPs that are used to prioritize genes. RegulomeDB score is a categorical score representing regulatory functionality of SNPs based on eQTLs and chromatin marks. Please refer to the original publication for details from links.	Check	Unchecked	-
Minimum RegulomeDB score (≥)	Mandatory if `RegulomeDB score` is checked	RegulomeDB score is a categorical score from 1a to 7) Score 1a means that those SNPs are most likely affecting regulatory elements and 7 means that those SNPs do not have any annotations. SNPs are recorded as NA if they are not present in the database. SNPs with NA will not be included for filtering on RegulomeDB score.	string	7	higher: more SNPs will be mapped to genes. lower: less SNPs will be mapped to genes.
15-core chromatin state	Optional	Check if you want to perform filtering of SNPs by chromatin state. This applies to selected SNPs in LD with independent significant SNPs that are used to prioritize genes. The chromatin state represents accessibility of genomic regions (every 200bp) with 15 categorical states predicted by ChromHMM based on 5 chromatin marks for 127 epigenomes.	Check	Unchecked	-
15-core chromatin state tissue/cell types	Mandatory if `15-core chromatin state` is checked	Multiple tissue/cell types can be selected from the list.	Multiple selection	none	-
Maximum state of chromatin(s)	Mandatory if `15-core chromatin state` is checked	The maximum state to filter SNPs. Between 1 and 15. Generally, between 1 and 7 is open state.	numeric	7	higher: more SNPs will be mapped to genes. lower: less SNPs will be mapped to genes.
Method for 15-core chromatin state filtering	Mandatory if `15-core chromatin state` is checked	When multiple tissue/cell types are selected, either `any` (filtered on SNPs which have state above than threshold in any of selected tissue/cell types), `majority` (filtered on SNPs which have state above than threshold in majority (≥50%) of selected tissue/cell type), or `all` (filtered on SNPs which have state above than threshold in all of selected tissue/cell type).	Selection	any	-
Annotation datasets	Optional	Additional functional annotations can be annotated to candidate SNPs. All available data are regional based annotation (bed file format).	Multiple selection	none	-
Annotation filtering method	Mandatory if any of `Annotation datasets` is selected.	By default, SNPs are not filtered by the annotations selected in `Annotation datasets` . To filter SNPs based on the selected annotation, select this options from `any` (filtered on SNPs which are overlapping with any selected annotations), `majority` (filtered on SNPs which are overlapping with majority (≥50%) of selected annotations), or `all` (filtered on SNPs which are overlapping with all of selected annotations).	Selection	No filtering	-

4. Gene types

Biotype of genes to map can be selected. Please refer to Ensembl for details of biotypes.

Parameter	Mandatory	Description	Type	Default
Gene type	Mandatory	Gene type to map. This is based on gene_biotype obtained from BioMart of Ensembl build 85. Please see here for details	Multiple selection.	Protein coding genes.

5. MHC region

The MHC region is often excluded due to its complicated LD structure. Therefore, this option is checked by default. Please uncheck to include MHC region. Note that it doesn't change any results if there is no significant hit in the MHC region.

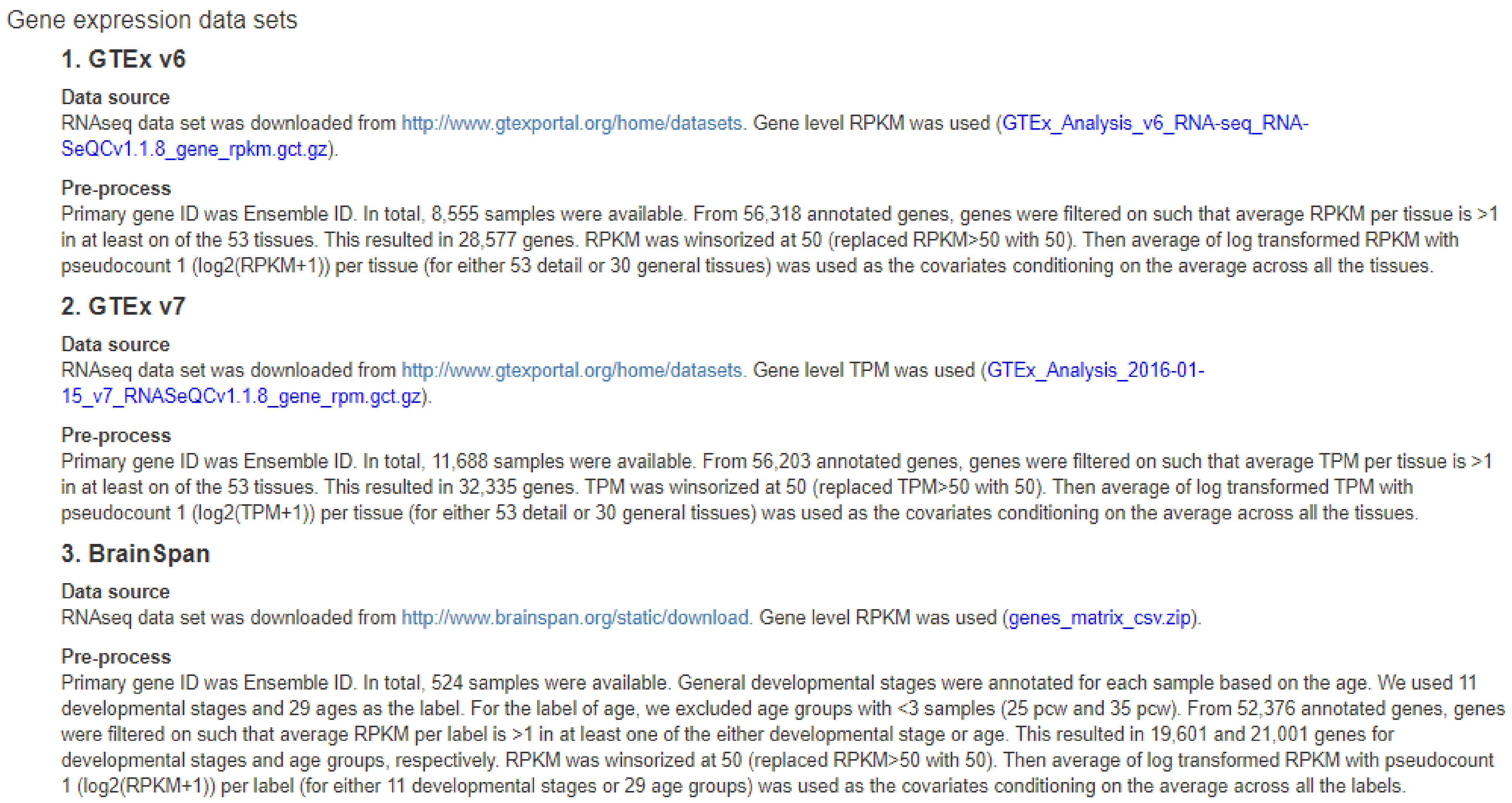
Parameter	Mandatory	Description	Type	Default
Exclude MHC region	Optional	Check if you want to exclude the MHC region. The default region is defined as between "MOG" and "COL11A2" genes.	Check	Checked
Options for excluding MHC region	Optional	MHC region can be excluded only from either annotations or MAGMA gene analysis, or from both by selecting this option.	Select	Only from annotations
Extended MHC region	Optional	User specified MHC region to exclude (for extended or shorter region). The input format should be like "25000000-34000000" on hg19.	Text	Null

6. MAGMA analysis

MAGMA gene and gene-set analyses are performed for the input summary statistics by default, but user can also select to omit MAGMA process that reduce the run time of SNP2GENE process. Gene expression data sets for MAGMA gene expression analysis can be also selected from here.

Parameter	Mandatory	Description	Type	Default
Perform MAGMA	Optional	UNCHECK to SKIP MAGMA analyses.	Check	Checked
MAGMA gene annotation window	Mandatory when `MAGMA` is active.	The window of the genes to assign SNPs (symmetric). e.g. when 5kb is selected, SNPs within 5kb window of a gene (both side) will be assigned to that gene. The option is available from 0, 5, 10, 15, 20kb window.	Select	0kb from both side of the genes
MAGMA gene expression analysis	Mandatory when `MAGMA` is active.	Gene expression data sets used for MAGMA gene-property analysis to test positive association between genetic associations and gene expression in a given label.	Select	GTEx v6

Gene expression database used by Fuma



Gene expression data sets

1. GTEx v6

Data source
RNAseq data set was downloaded from http://www.gtexportal.org/home/datasets. Gene level RPKM was used (GTEx_Analysis_v6_RNA-seq_RNA-SeQCv1.1.8_gene_rpkm.gct.gz).

Pre-process
Primary gene ID was Ensemble ID. In total, 8,555 samples were available. From 56,318 annotated genes, genes were filtered on such that average RPKM per tissue is >1 in at least on of the 53 tissues. This resulted in 28,577 genes. RPKM was winsorized at 50 (replaced RPKM>50 with 50). Then average of log transformed RPKM with pseudocount 1 (log2(RPKM+1)) per tissue (for either 53 detail or 30 general tissues) was used as the covariates conditioning on the average across all the tissues.

2. GTEx v7

Data source
RNAseq data set was downloaded from http://www.gtexportal.org/home/datasets. Gene level TPM was used (GTEx_Analysis_2016-01-15_v7_RNASeQCv1.1.8_gene_rpm.gct.gz).

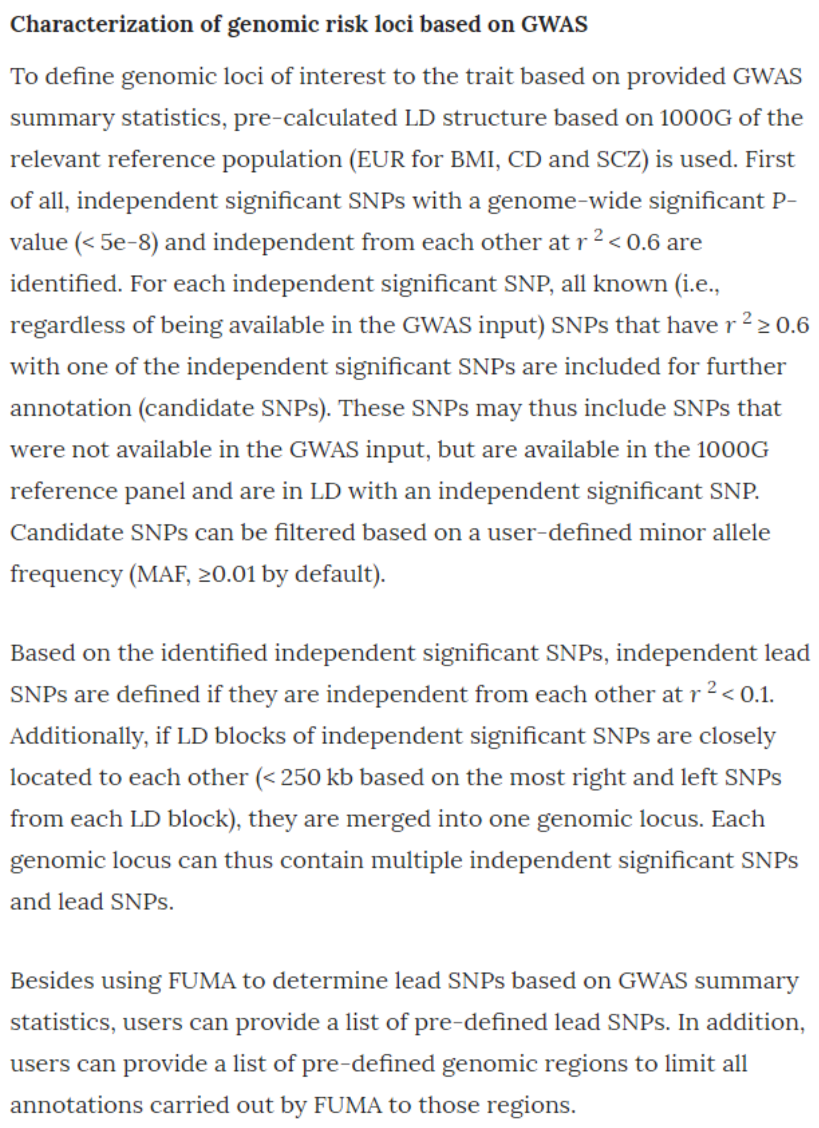
Pre-process
Primary gene ID was Ensemble ID. In total, 11,688 samples were available. From 56,203 annotated genes, genes were filtered on such that average TPM per tissue is >1 in at least on of the 53 tissues. This resulted in 32,335 genes. TPM was winsorized at 50 (replaced TPM>50 with 50). Then average of log transformed TPM with pseudocount 1 (log2(TPM+1)) per tissue (for either 53 detail or 30 general tissues) was used as the covariates conditioning on the average across all the tissues.

3. BrainSpan

Data source
RNAseq data set was downloaded from http://www.brainspan.org/static/download. Gene level RPKM was used (genes_matrix_csv.zip).

Pre-process
Primary gene ID was Ensemble ID. In total, 524 samples were available. General developmental stages were annotated for each sample based on the age. We used 11 developmental stages and 29 ages as the label. For the label of age, we excluded age groups with <3 samples (25 pcw and 35 pcw). From 52,376 annotated genes, genes were filtered on such that average RPKM per label is >1 in at least one of the either developmental stage or age. This resulted in 19,601 and 21,001 genes for developmental stages and age groups, respectively. RPKM was winsorized at 50 (replaced RPKM>50 with 50). Then average of log transformed RPKM with pseudocount 1 (log2(RPKM+1)) per label (for either 11 developmental stages or 29 age groups) was used as the covariates conditioning on the average across all the labels.

Fuma : Genomic risk loci Identification



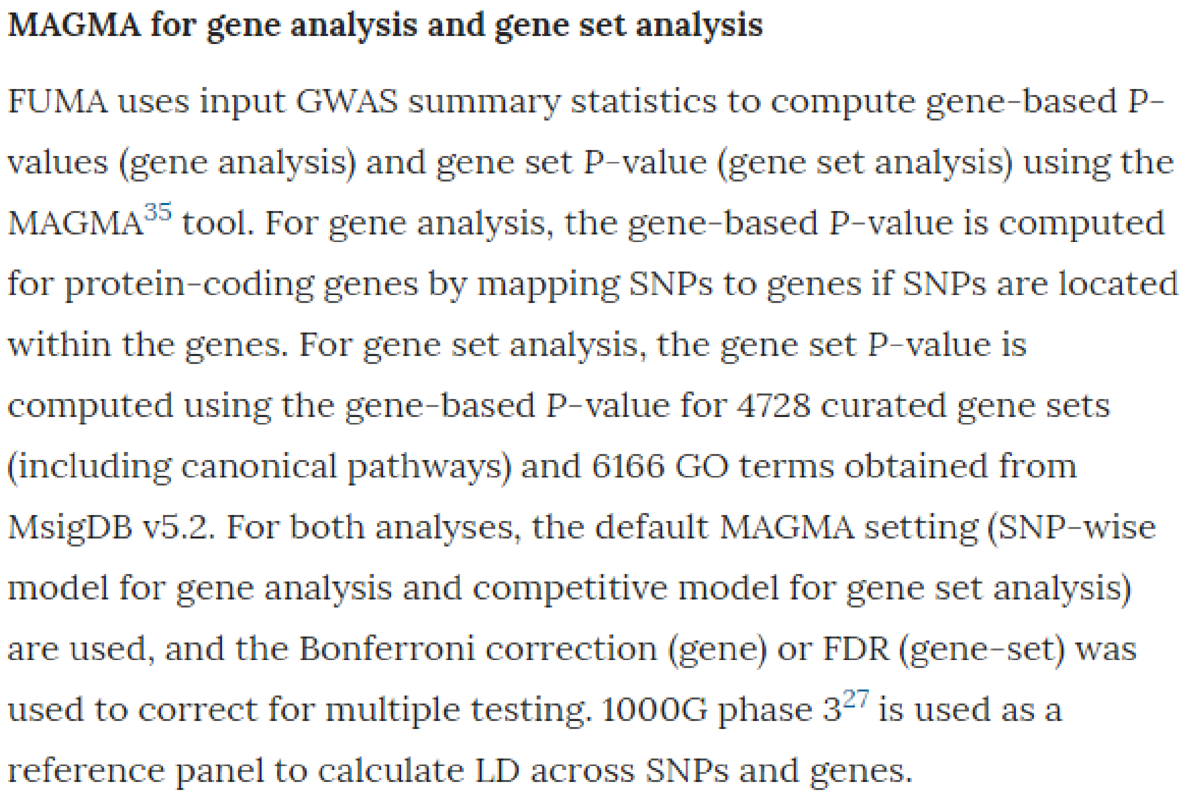
Characterization of genomic risk loci based on GWAS

To define genomic loci of interest to the trait based on provided GWAS summary statistics, pre-calculated LD structure based on 1000G of the relevant reference population (EUR for BMI, CD and SCZ) is used. First of all, independent significant SNPs with a genome-wide significant *P*-value (< 5e-8) and independent from each other at $r^2 < 0.6$ are identified. For each independent significant SNP, all known (i.e., regardless of being available in the GWAS input) SNPs that have $r^2 \geq 0.6$ with one of the independent significant SNPs are included for further annotation (candidate SNPs). These SNPs may thus include SNPs that were not available in the GWAS input, but are available in the 1000G reference panel and are in LD with an independent significant SNP. Candidate SNPs can be filtered based on a user-defined minor allele frequency (MAF, ≥0.01 by default).

Based on the identified independent significant SNPs, independent lead SNPs are defined if they are independent from each other at $r^2 < 0.1$. Additionally, if LD blocks of independent significant SNPs are closely located to each other (< 250 kb based on the most right and left SNPs from each LD block), they are merged into one genomic locus. Each genomic locus can thus contain multiple independent significant SNPs and lead SNPs.

Besides using FUMA to determine lead SNPs based on GWAS summary statistics, users can provide a list of pre-defined lead SNPs. In addition, users can provide a list of pre-defined genomic regions to limit all annotations carried out by FUMA to those regions.

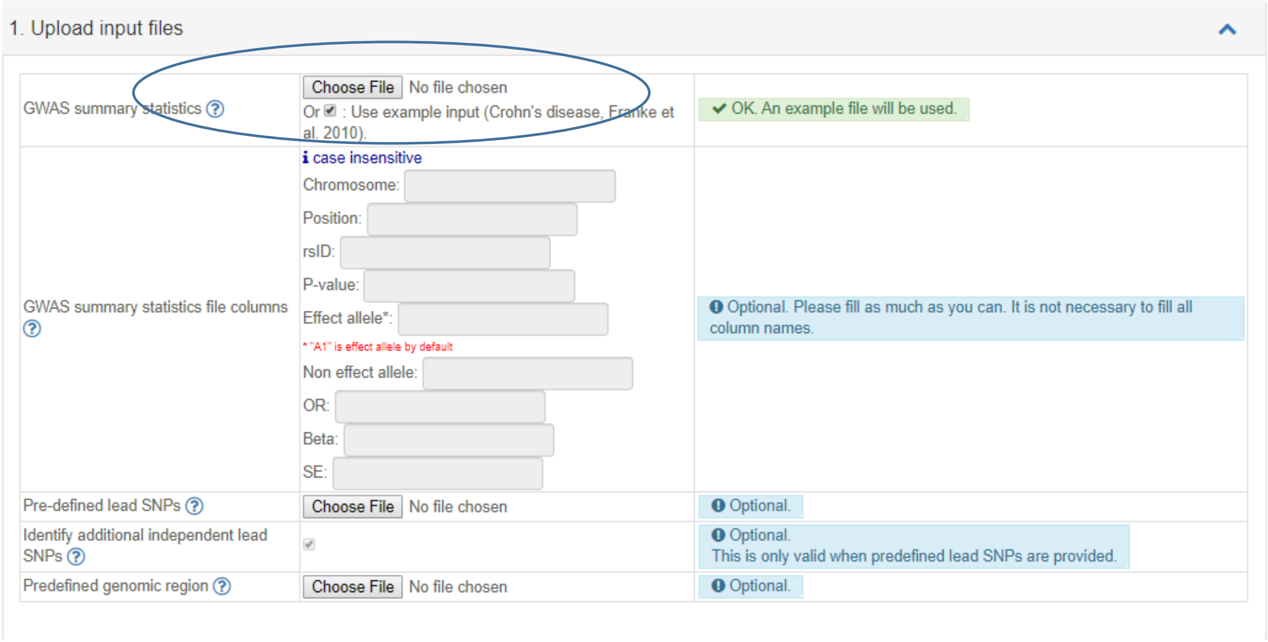
Fuma : Gene and Gene set analysis



MAGMA for gene analysis and gene set analysis

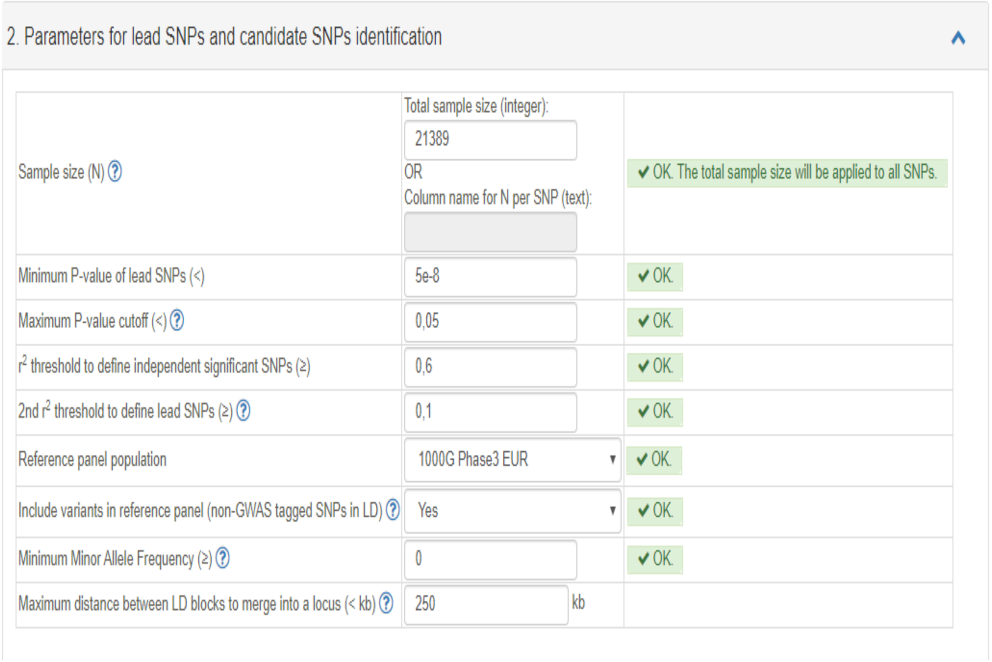
FUMA uses input GWAS summary statistics to compute gene-based P-values (gene analysis) and gene set P-value (gene set analysis) using the MAGMA[35] tool. For gene analysis, the gene-based P-value is computed for protein-coding genes by mapping SNPs to genes if SNPs are located within the genes. For gene set analysis, the gene set P-value is computed using the gene-based P-value for 4728 curated gene sets (including canonical pathways) and 6166 GO terms obtained from MsigDB v5.2. For both analyses, the default MAGMA setting (SNP-wise model for gene analysis and competitive model for gene set analysis) are used, and the Bonferroni correction (gene) or FDR (gene-set) was used to correct for multiple testing. 1000G phase 3[27] is used as a reference panel to calculate LD across SNPs and genes.

Lets run SNP2GENE



1. Upload input files

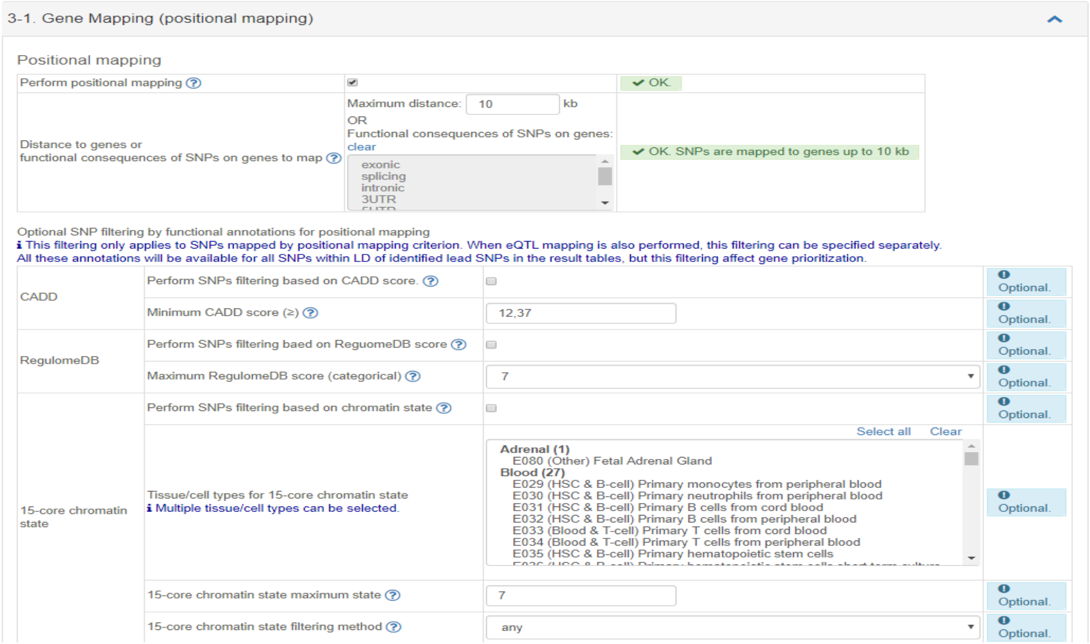
GWAS summary statistics ?	Choose File No file chosen Or ☑ : Use example input (Crohn's disease, Franke et al. 2010).	✓ OK. An example file will be used.
GWAS summary statistics file columns ?	ℹ case insensitive Chromosome: Position: rsID: P-value: Effect allele*: * "A1" is effect allele by default Non effect allele: OR: Beta: SE:	Optional. Please fill as much as you can. It is not necessary to fill all column names.
Pre-defined lead SNPs ?	Choose File No file chosen	Optional.
Identify additional independent lead SNPs ?	☑	Optional. This is only valid when predefined lead SNPs are provided.
Predefined genomic region ?	Choose File No file chosen	Optional.



2. Parameters for lead SNPs and candidate SNPs identification

Sample size (N) ?	Total sample size (integer):
	21389
	OR
	Column name for N per SNP (text):
		✓ OK. The total sample size will be applied to all SNPs.

Minimum P-value of lead SNPs (<)	5e-8	✓ OK.
Maximum P-value cutoff (<) ?	0,05	✓ OK.
r^2 threshold to define independent significant SNPs (≥)	0,6	✓ OK.
2nd r^2 threshold to define lead SNPs (≥) ?	0,1	✓ OK.
Reference panel population	1000G Phase3 EUR	✓ OK.
Include variants in reference panel (non-GWAS tagged SNPs in LD) ?	Yes	✓ OK.
Minimum Minor Allele Frequency (≥) ?	0	✓ OK.
Maximum distance between LD blocks to merge into a locus (< kb) ?	250 kb	



3-1. Gene Mapping (positional mapping)

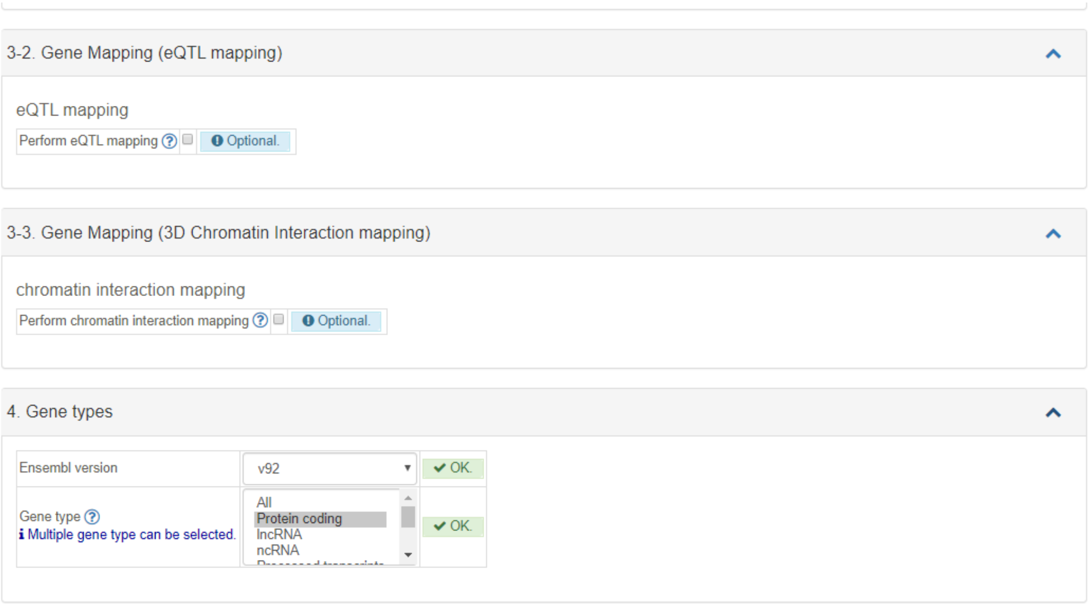
Positional mapping

Perform positional mapping ?	☑	✓ OK.
Distance to genes or functional consequences of SNPs on genes to map ?	Maximum distance: 10 kb OR Functional consequences of SNPs on genes: clear exonic splicing intronic 3UTR	✓ OK. SNPs are mapped to genes up to 10 kb

Optional SNP filtering by functional annotations for positional mapping

ℹ This filtering only applies to SNPs mapped by positional mapping criterion. When eQTL mapping is also performed, this filtering can be specified separately.
All these annotations will be available for all SNPs within LD of identified lead SNPs in the result tables, but this filtering affect gene prioritization.

CADD	Perform SNPs filtering based on CADD score. ?	□	Optional.
	Minimum CADD score (≥) ?	12,37	Optional.
RegulomeDB	Perform SNPs filtering baed on ReguomeDB score ?	□	Optional.
	Maximum RegulomeDB score (categorical) ?	7	Optional.
15-core chromatin state	Perform SNPs filtering based on chromatin state ?	□	Optional.
	Tissue/cell types for 15-core chromatin state ℹ Multiple tissue/cell types can be selected.	Select all Clear Adrenal (1) E080 (Other) Fetal Adrenal Gland Blood (27) E029 (HSC & B-cell) Primary monocytes from peripheral blood E030 (HSC & B-cell) Primary neutrophils from peripheral blood E031 (HSC & B-cell) Primary B cells from cord blood E032 (HSC & B-cell) Primary B cells from peripheral blood E033 (Blood & T-cell) Primary T cells from cord blood E034 (Blood & T-cell) Primary T cells from peripheral blood E035 (HSC & B-cell) Primary hematopoietic stem cells	Optional.
	15-core chromatin state maximum state ?	7	Optional.
	15-core chromatin state filtering method ?	any	Optional.



3-2. Gene Mapping (eQTL mapping)

eQTL mapping

Perform eQTL mapping ?	□	Optional.

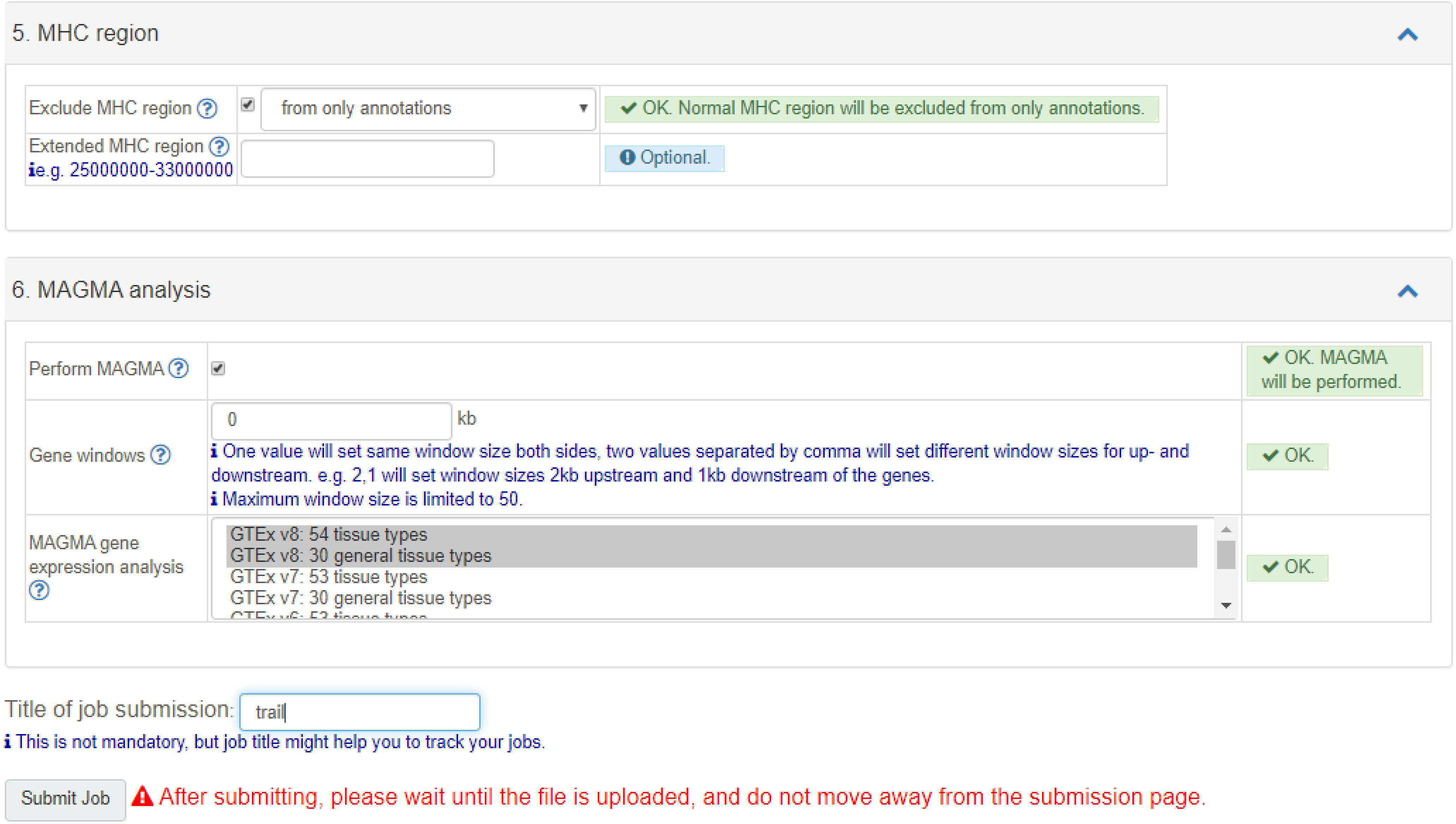
3-3. Gene Mapping (3D Chromatin Interaction mapping)

chromatin interaction mapping

Perform chromatin interaction mapping ?	□	Optional.

4. Gene types

Ensembl version	v92	✓ OK.
Gene type ? ℹ Multiple gene type can be selected.	All Protein coding lncRNA ncRNA	✓ OK.



5. MHC region

Exclude MHC region ?	☑ from only annotations ▾	✓ OK. Normal MHC region will be excluded from only annotations.
Extended MHC region ? ℹ e.g. 25000000-33000000		❶ Optional.

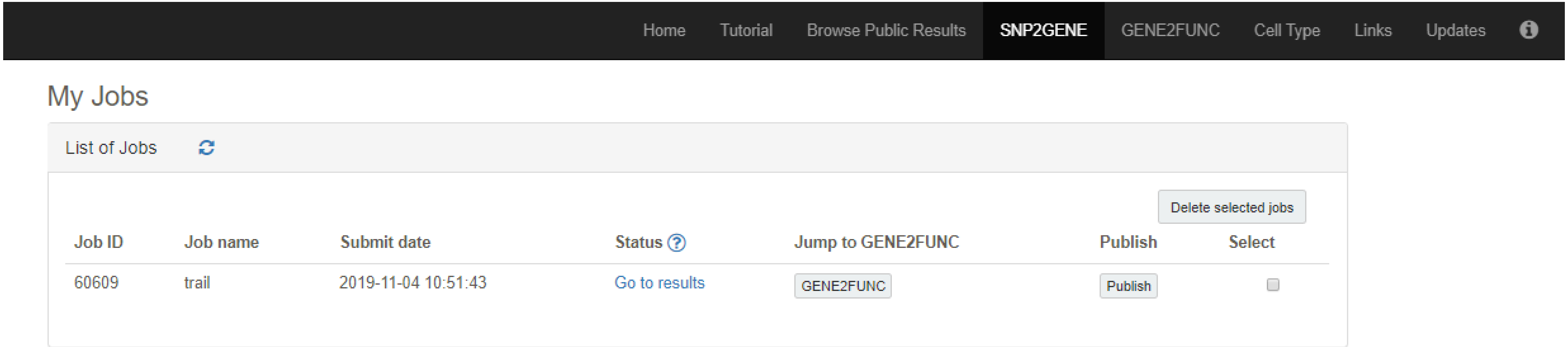
6. MAGMA analysis

Perform MAGMA ?	☑	✓ OK. MAGMA will be performed.
Gene windows ?	0 kb ℹ One value will set same window size both sides, two values separated by comma will set different window sizes for up- and downstream. e.g. 2,1 will set window sizes 2kb upstream and 1kb downstream of the genes. ℹ Maximum window size is limited to 50.	✓ OK.
MAGMA gene expression analysis ?	GTEx v8: 54 tissue types GTEx v8: 30 general tissue types GTEx v7: 53 tissue types GTEx v7: 30 general tissue types GTEx v6: 53 tissue types	✓ OK.

Title of job submission: trail

ℹ This is not mandatory, but job title might help you to track your jobs.

Submit Job ⚠ After submitting, please wait until the file is uploaded, and do not move away from the submission page.



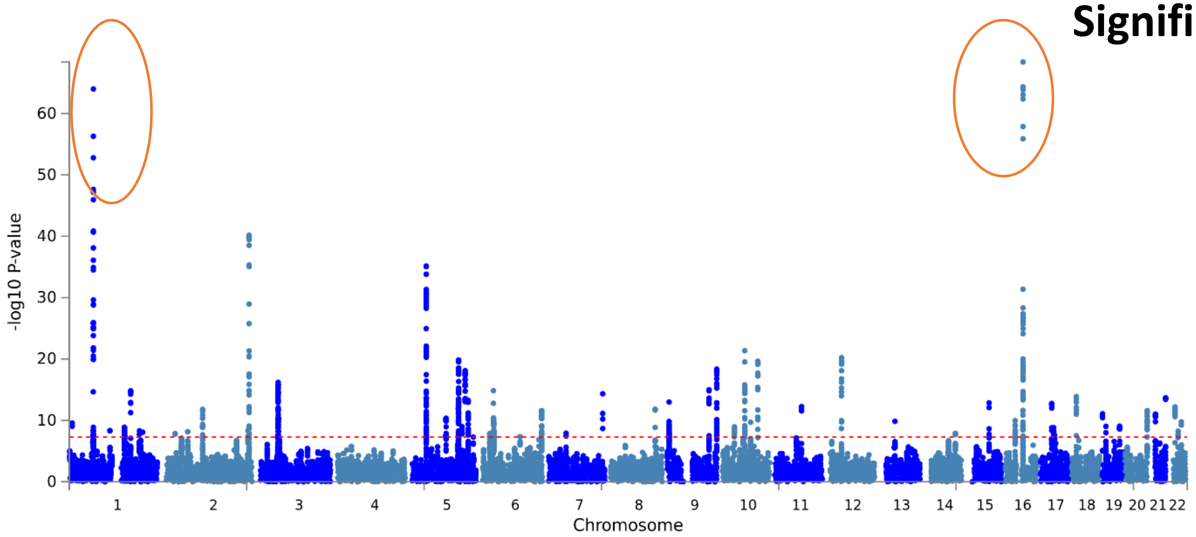
My Jobs

List of Jobs

Job ID	Job name	Submit date	Status ?	Jump to GENE2FUNC	Publish	Select
60609	trail	2019-11-04 10:51:43	Go to results	GENE2FUNC	Publish	□

Result

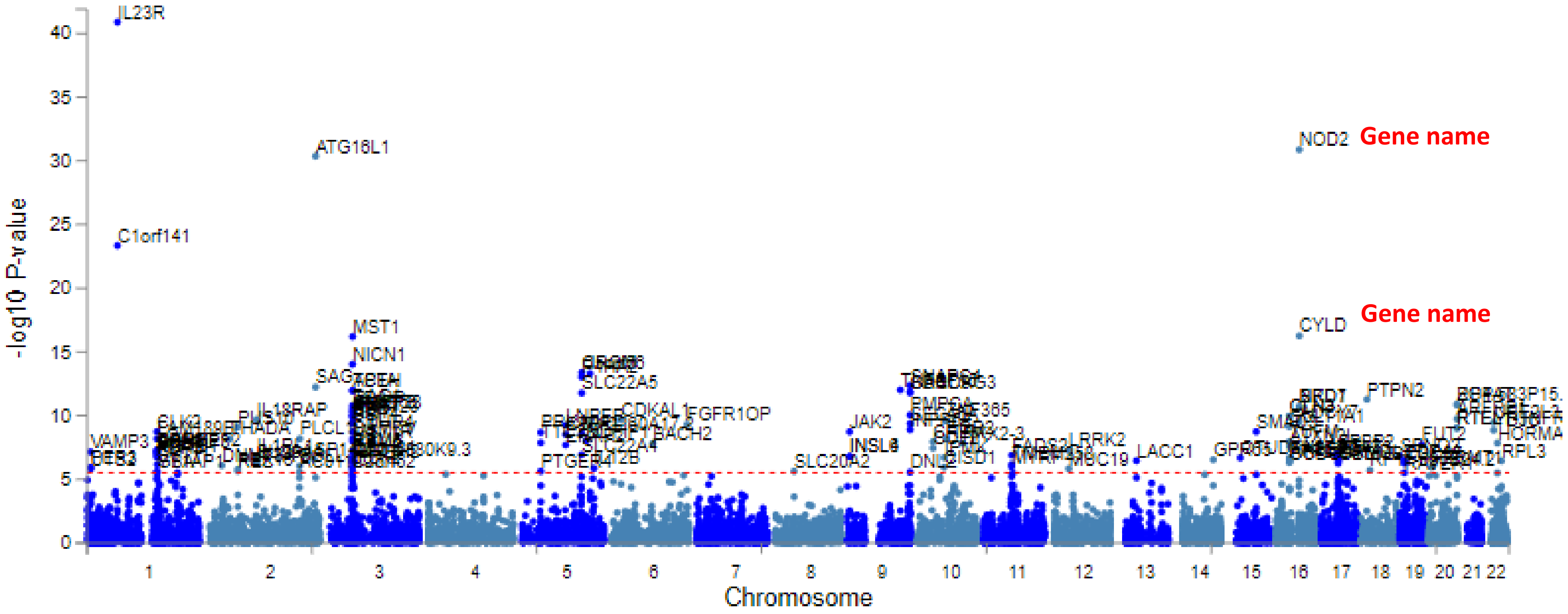
GWAS PLOTS



Significant signals in chromosome 1 and 16

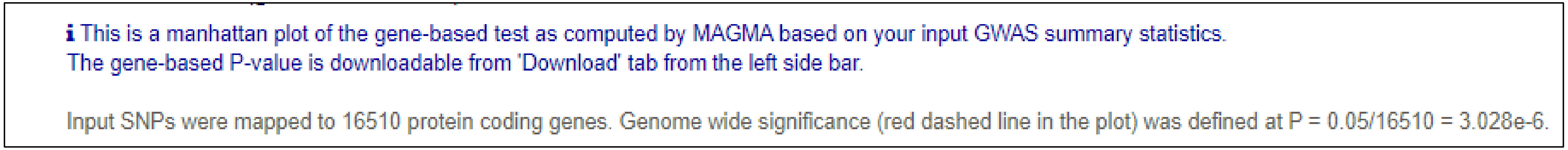
Manhattan Plot (GWAS summary statistics)

GWAS PLOTS (gene based test)



Gene name

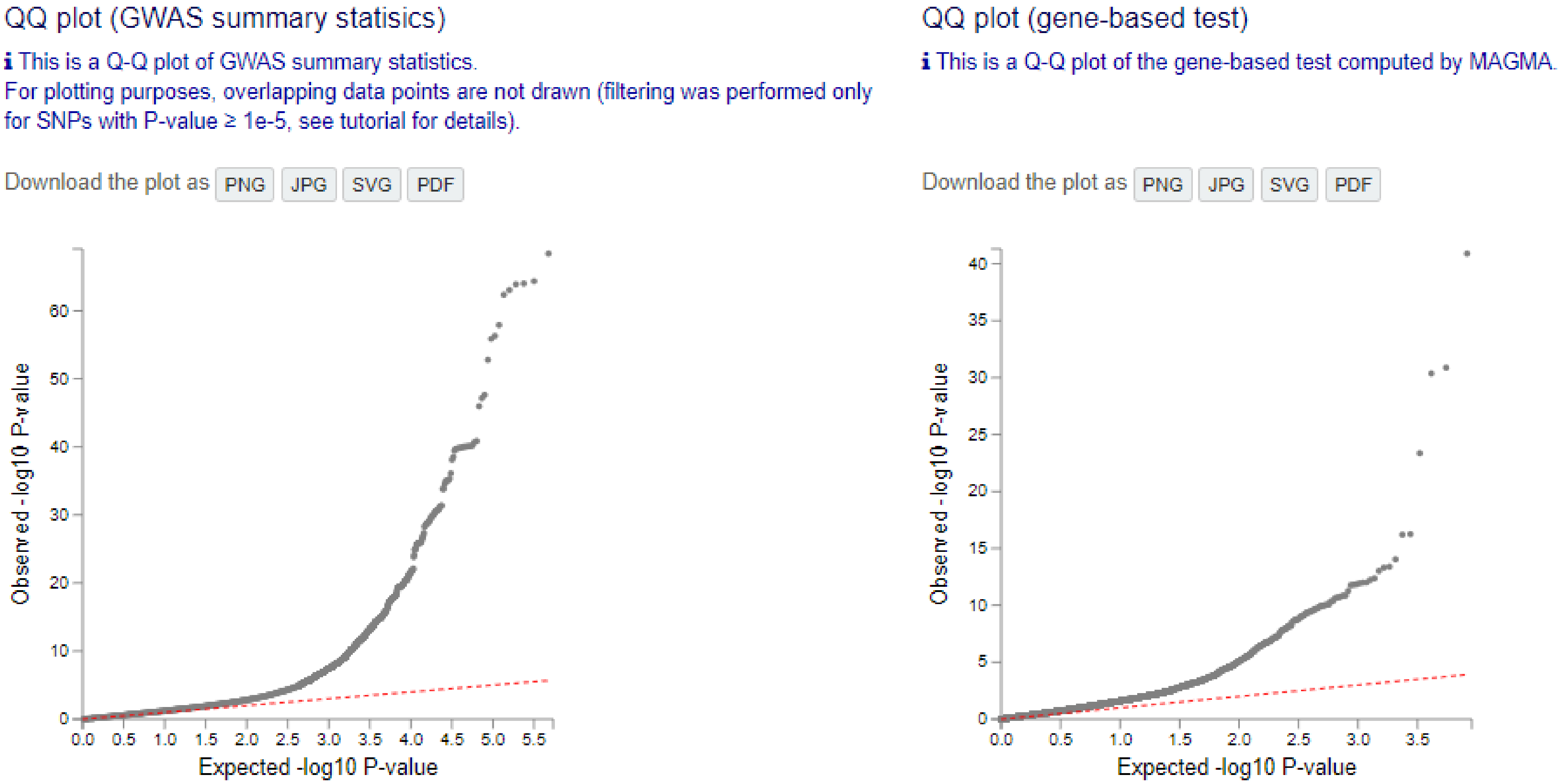
Gene name



ℹ This is a manhattan plot of the gene-based test as computed by MAGMA based on your input GWAS summary statistics.
The gene-based P-value is downloadable from 'Download' tab from the left side bar.

Input SNPs were mapped to 16510 protein coding genes. Genome wide significance (red dashed line in the plot) was defined at P = 0.05/16510 = 3.028e-6.

Q-Q PLOTS (GWAS/gene based test)



QQ plot (GWAS summary statisics)

ℹ This is a Q-Q plot of GWAS summary statistics.
For plotting purposes, overlapping data points are not drawn (filtering was performed only for SNPs with P-value ≥ 1e-5, see tutorial for details).

Download the plot as PNG JPG SVG PDF

QQ plot (gene-based test)

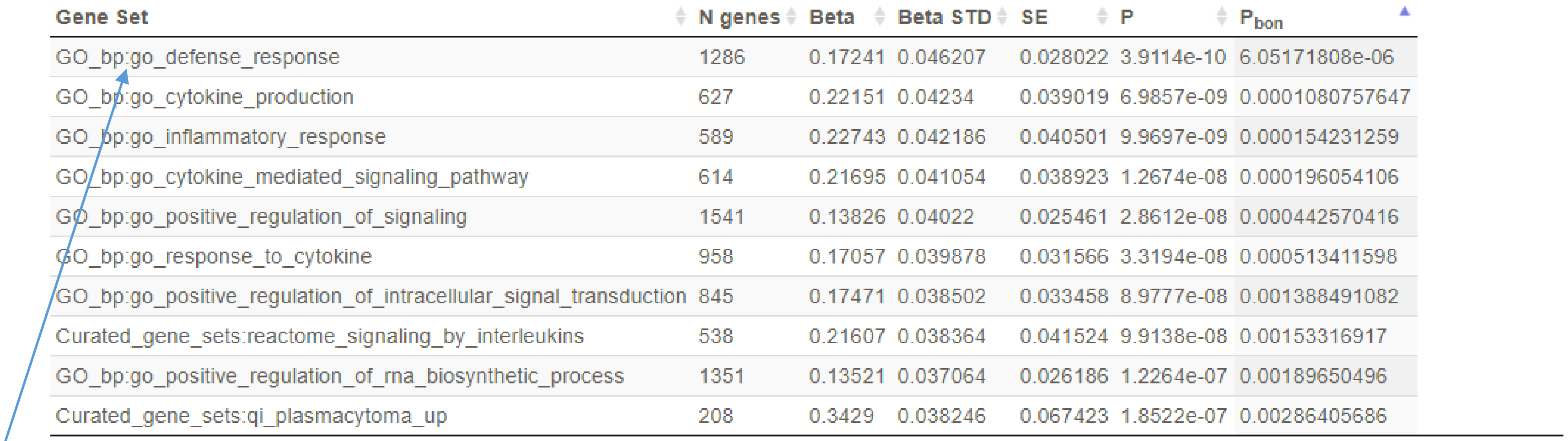
ℹ This is a Q-Q plot of the gene-based test computed by MAGMA.

Download the plot as PNG JPG SVG PDF

Slight variation in Plots (From SNPs to Gene based QQ plot)

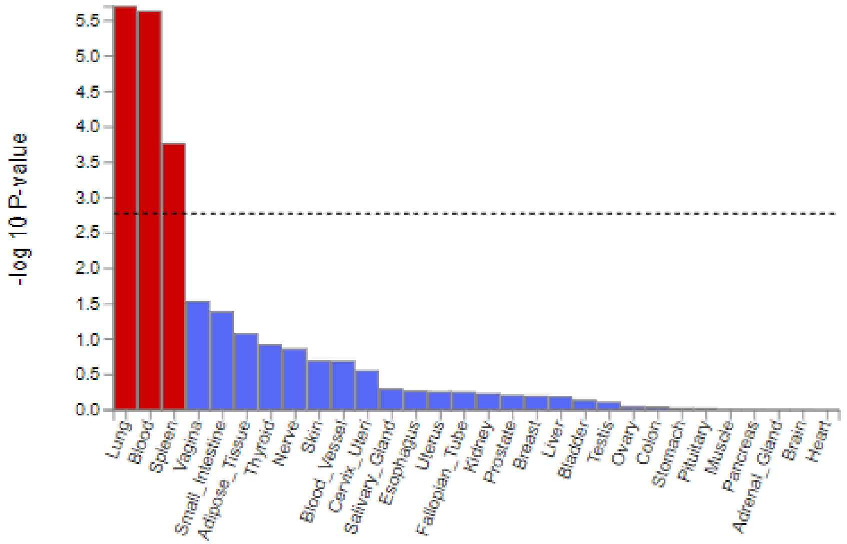
MAGMA gene set analysis

Over represented Gene ontology :



Gene Set	N genes	Beta	Beta STD	SE	P	P_{bon}
GO_bp:go_defense_response	1286	0.17241	0.046207	0.028022	3.9114e-10	6.05171808e-06
GO_bp:go_cytokine_production	627	0.22151	0.04234	0.039019	6.9857e-09	0.0001080757647
GO_bp:go_inflammatory_response	589	0.22743	0.042186	0.040501	9.9697e-09	0.000154231259
GO_bp:go_cytokine_mediated_signaling_pathway	614	0.21695	0.041054	0.038923	1.2674e-08	0.000196054106
GO_bp:go_positive_regulation_of_signaling	1541	0.13826	0.04022	0.025461	2.8612e-08	0.000442570416
GO_bp:go_response_to_cytokine	958	0.17057	0.039878	0.031566	3.3194e-08	0.000513411598
GO_bp:go_positive_regulation_of_intracellular_signal_transduction	845	0.17471	0.038502	0.033458	8.9777e-08	0.001388491082
Curated_gene_sets:reactome_signaling_by_interleukins	538	0.21607	0.038364	0.041524	9.9138e-08	0.00153316917
GO_bp:go_positive_regulation_of_rna_biosynthetic_process	1351	0.13521	0.037064	0.026186	1.2264e-07	0.00189650496
Curated_gene_sets:qi_plasmacytoma_up	208	0.3429	0.038246	0.067423	1.8522e-07	0.00286405686

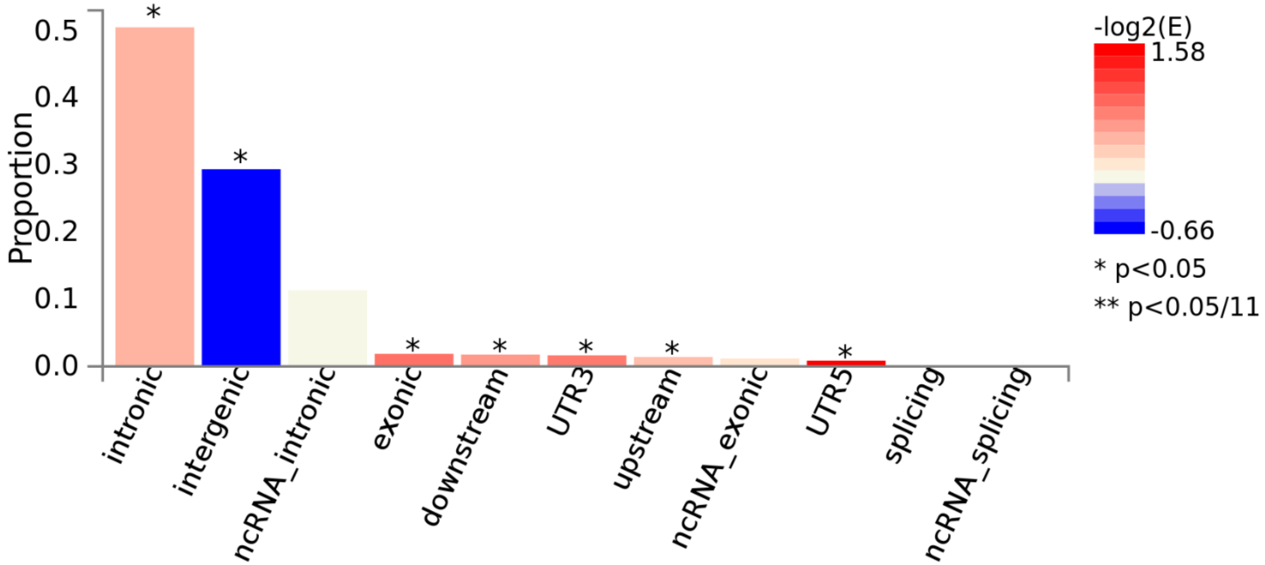
Defense response specific regulatory genes are highly significantly OR in this data.



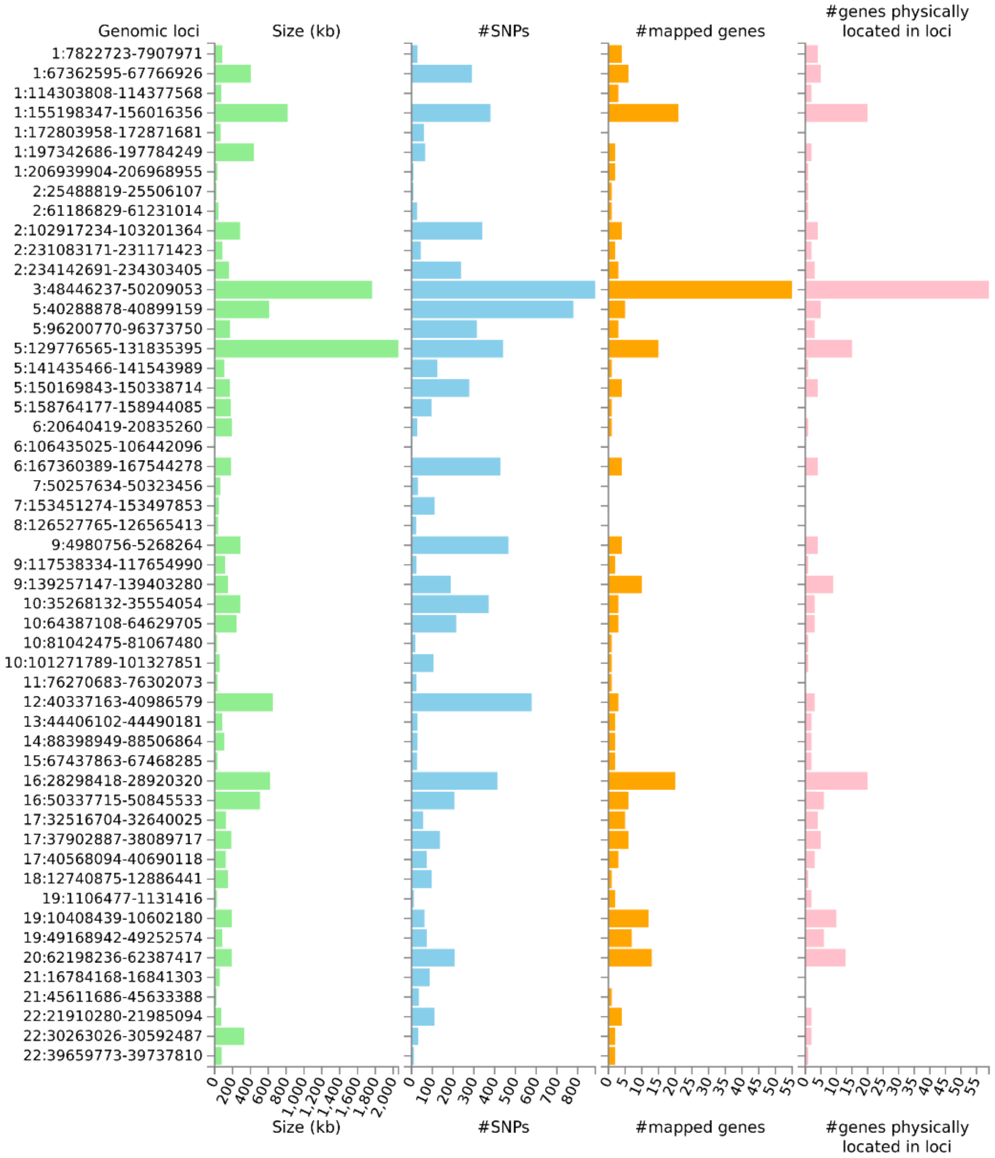
Signifiant expression observed in Lung, Blood and spleen tissue.

Summary of SNPs and mapped genes

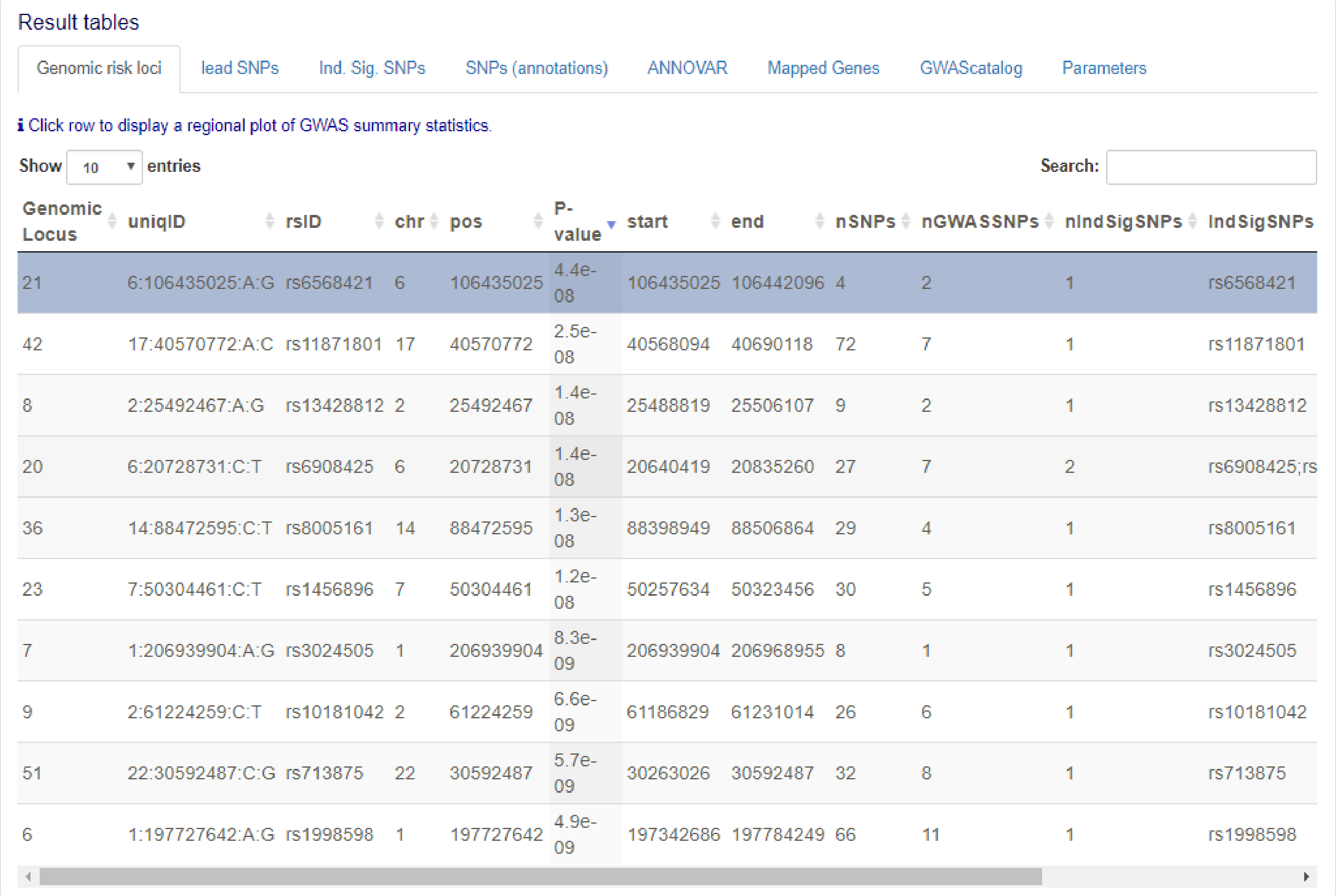
#Genomic risk loci	52
#lead SNPs	75
#Ind. Sig. SNPs	164
#candidate SNPs	8717
#candidate GWAS tagged SNPs	1247
#mapped genes	256



Distribution of SNPs



Fuma : Regional Plots



Result tables

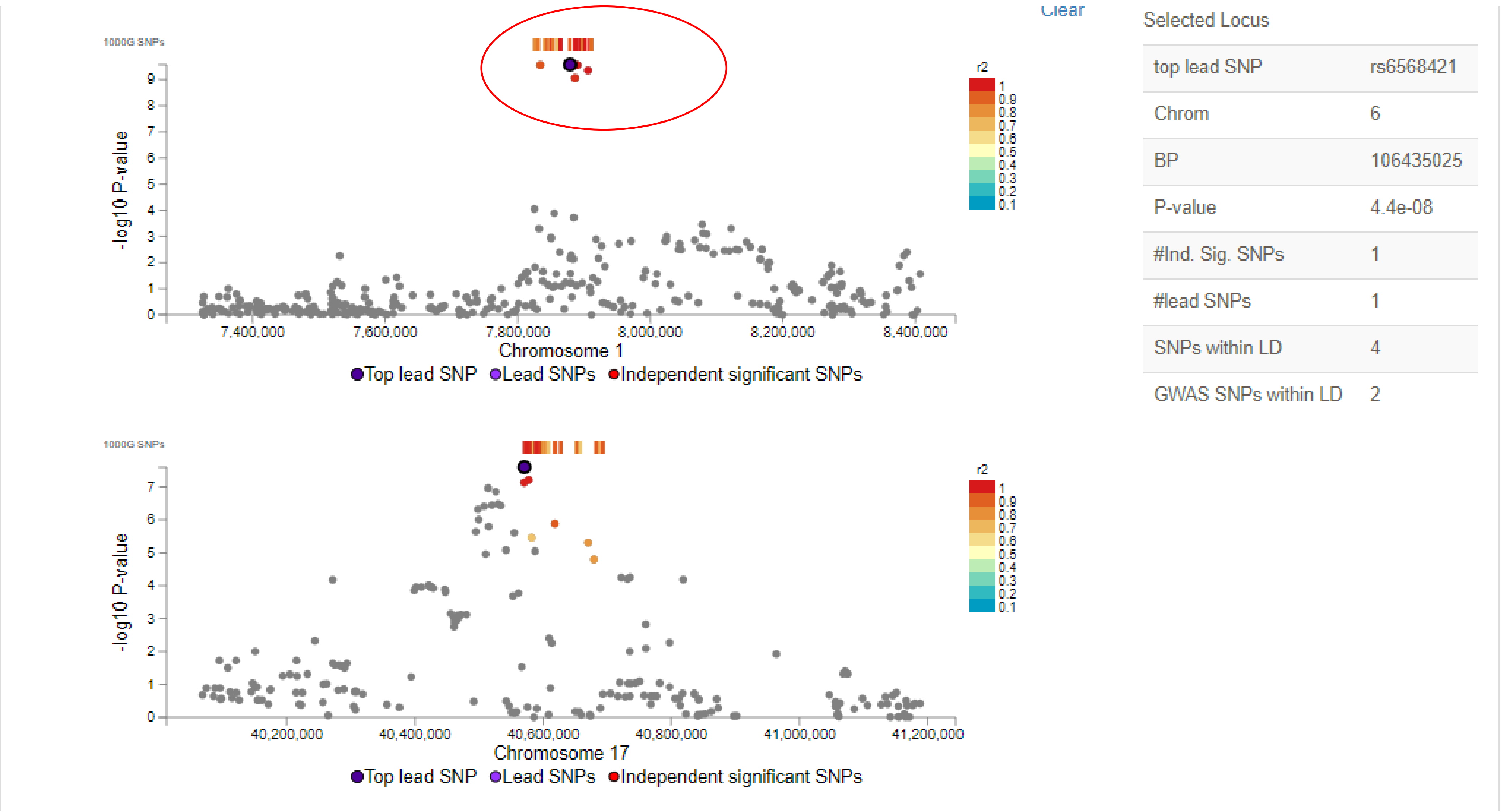
Genomic risk loci | lead SNPs | Ind. Sig. SNPs | SNPs (annotations) | ANNOVAR | Mapped Genes | GWAScatalog | Parameters

ℹ Click row to display a regional plot of GWAS summary statistics.

Show 10 entries

Search:

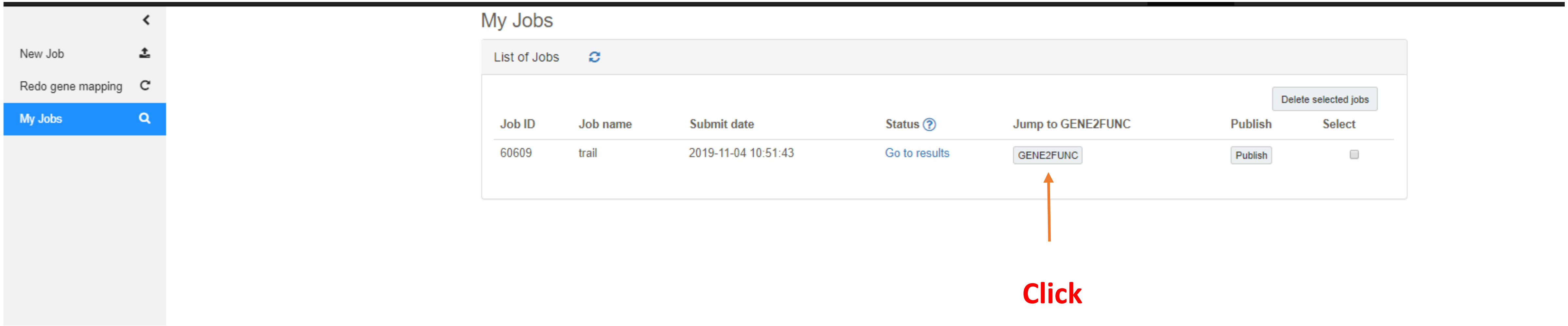
Genomic Locus	uniqID	rsID	chr	pos	P-value	start	end	nSNPs	nGWASSNPs	nIndSigSNPs	IndSigSNPs
21	6:106435025:A:G	rs6568421	6	106435025	4.4e-08	106435025	106442096	4	2	1	rs6568421
42	17:40570772:A:C	rs11871801	17	40570772	2.5e-08	40568094	40690118	72	7	1	rs11871801
8	2:25492467:A:G	rs13428812	2	25492467	1.4e-08	25488819	25506107	9	2	1	rs13428812
20	6:20728731:C:T	rs6908425	6	20728731	1.4e-08	20640419	20835260	27	7	2	rs6908425;rs
36	14:88472595:C:T	rs8005161	14	88472595	1.3e-08	88398949	88506864	29	4	1	rs8005161
23	7:50304461:C:T	rs1456896	7	50304461	1.2e-08	50257634	50323456	30	5	1	rs1456896
7	1:206939904:A:G	rs3024505	1	206939904	8.3e-09	206939904	206968955	8	1	1	rs3024505
9	2:61224259:C:T	rs10181042	2	61224259	6.6e-09	61186829	61231014	26	6	1	rs10181042
51	22:30592487:C:G	rs713875	22	30592487	5.7e-09	30263026	30592487	32	8	1	rs713875
6	1:197727642:A:G	rs1998598	1	197727642	4.9e-09	197342686	197784249	66	11	1	rs1998598



Selected Locus

top lead SNP	rs6568421
Chrom	6
BP	106435025
P-value	4.4e-08
#Ind. Sig. SNPs	1
#lead SNPs	1
SNPs within LD	4
GWAS SNPs within LD	2

Moving from SNP2Gene to Gene2FUNC



My Jobs

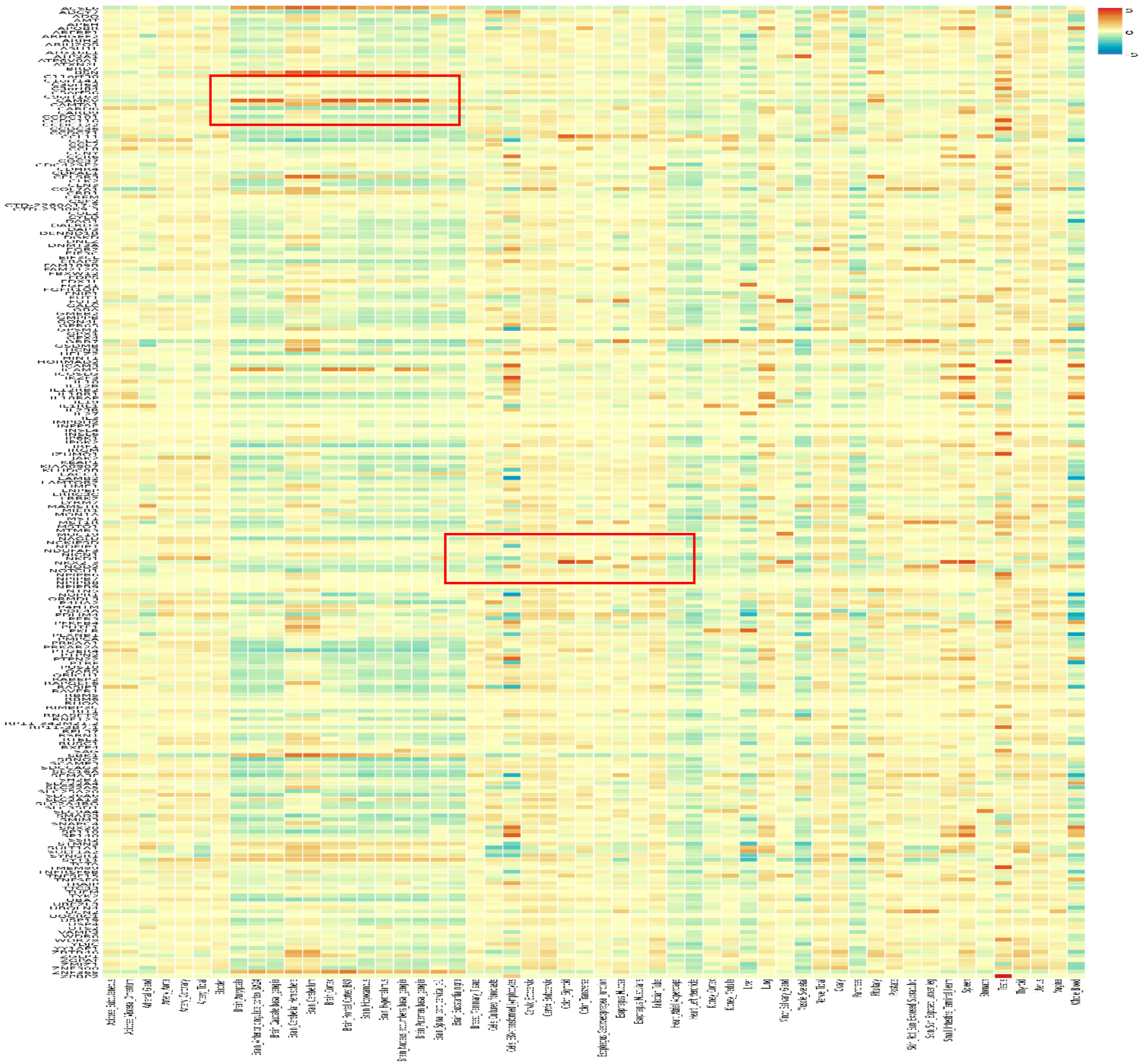
List of Jobs

Delete selected jobs

Job ID	Job name	Submit date	Status ?	Jump to GENE2FUNC	Publish	Select
60609	trail	2019-11-04 10:51:43	Go to results	GENE2FUNC	Publish	□

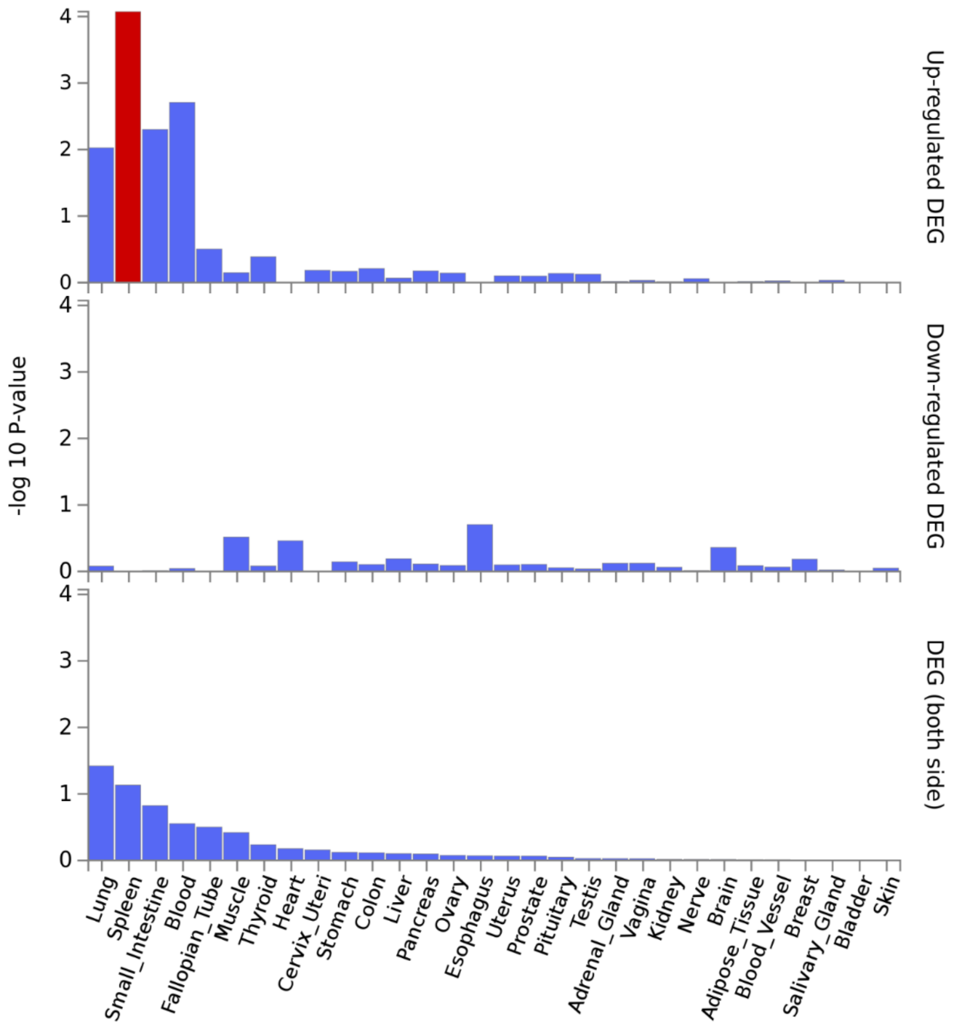
Click

Expression Heatmap plot



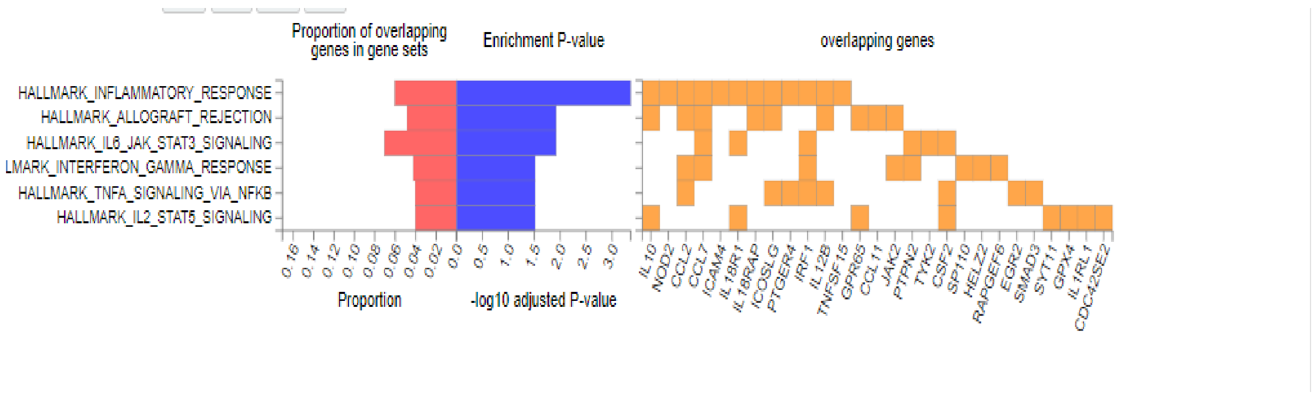
Dark red color : high expression

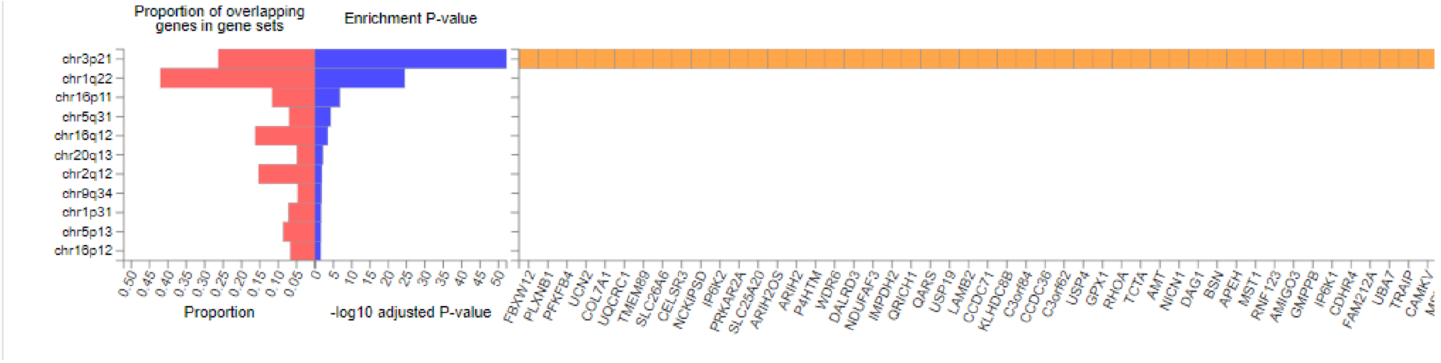
Tissue specific Expression

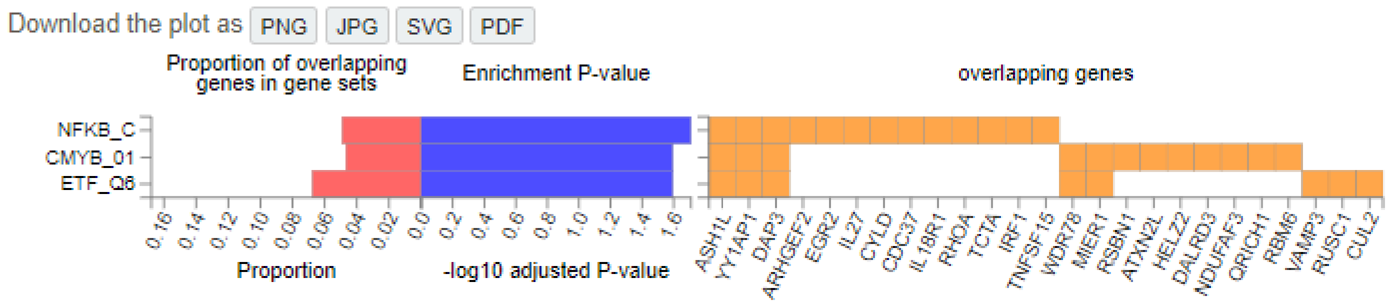


SNPs encoding genes have significant expression in spleen tissue

Functional Enrichment plots



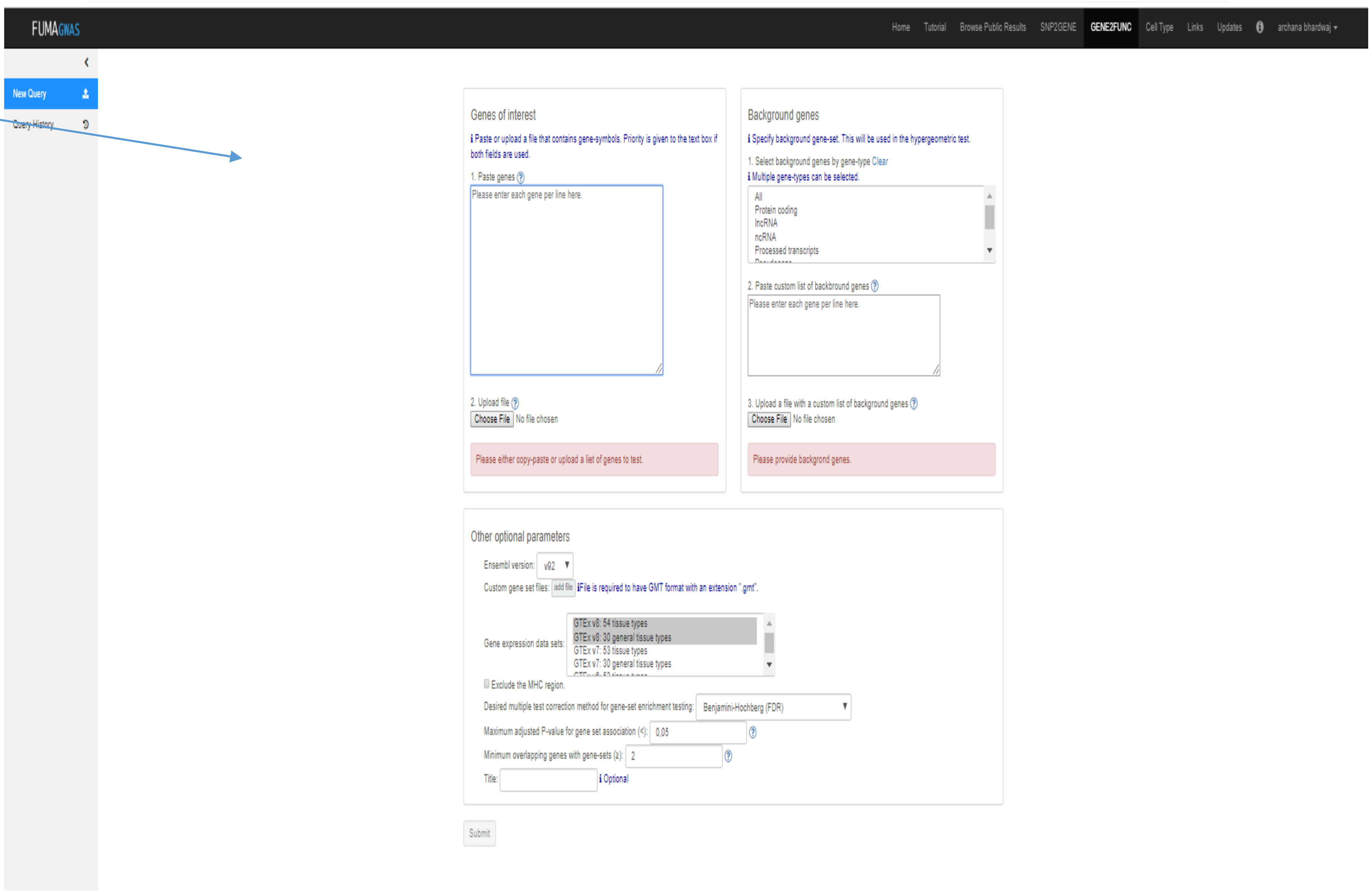




Let us run Gene2FUNC

Gene ID

ANKRD44
FOSL2
RAP1GAP
CARMIL1
CACNA1S
CYLD
ATG16L1
DOCK3
TTC33
INSL6
ADCY7
NKD1
KSR1
OSMR
BABAM2
IFNGR2
IL23R
NOD2
SPNS1
FOSL1
TEX41
AL138720.1
AC067751.1
ZNF512
LINC00824
AP005482.1
AC007493.1
LINC02178
LINC02178
AF246928.1
ATG16L1
AC008703.1



Genes of interest

ℹ Paste or upload a file that contains gene-symbols. Priority is given to the text box if both fields are used.

1. Paste genes ?

Please enter each gene per line here.

2. Upload file ?

Choose File No file chosen

Please either copy-paste or upload a liet of genes to test.

Background genes

ℹ Specify background gene-set. This will be used in the hypergeometric test.

1. Select background genes by gene-type Clear

ℹ Multiple gene-types can be selected.

All
Protein coding
lncRNA
ncRNA
Processed transcripts

2. Paste custom list of backbround genes ?

Please enter each gene per line here.

3. Upload a file with a custom list of background genes ?

Choose File No file chosen

Please provide backgrond genes.

Other optional parameters

Ensembl version: v92

Custom gene set files: add file ℹFile is required to have GMT format with an extension ".gmt".

Gene expression data sets:
GTEx v8: 54 tissue types
GTEx v8: 30 general tissue types
GTEx v7: 53 tissue types
GTEx v7: 30 general tissue types

□ Exclude the MHC region.

Desired multiple test correction method for gene-set enrichment testing: Benjamini-Hochberg (FDR)

Maximum adjusted P-value for gene set association (<): 0,05 ?

Minimum overlapping genes with gene-sets (≥): 2 ?

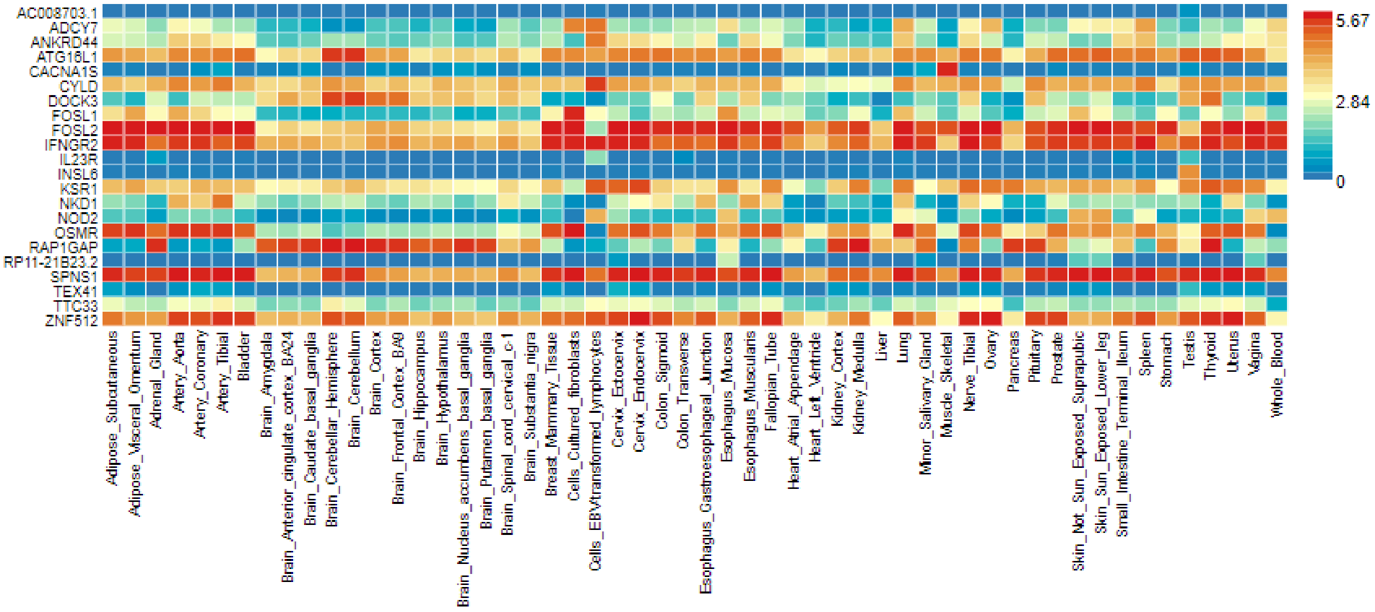
Title: ℹ Optional

Submit

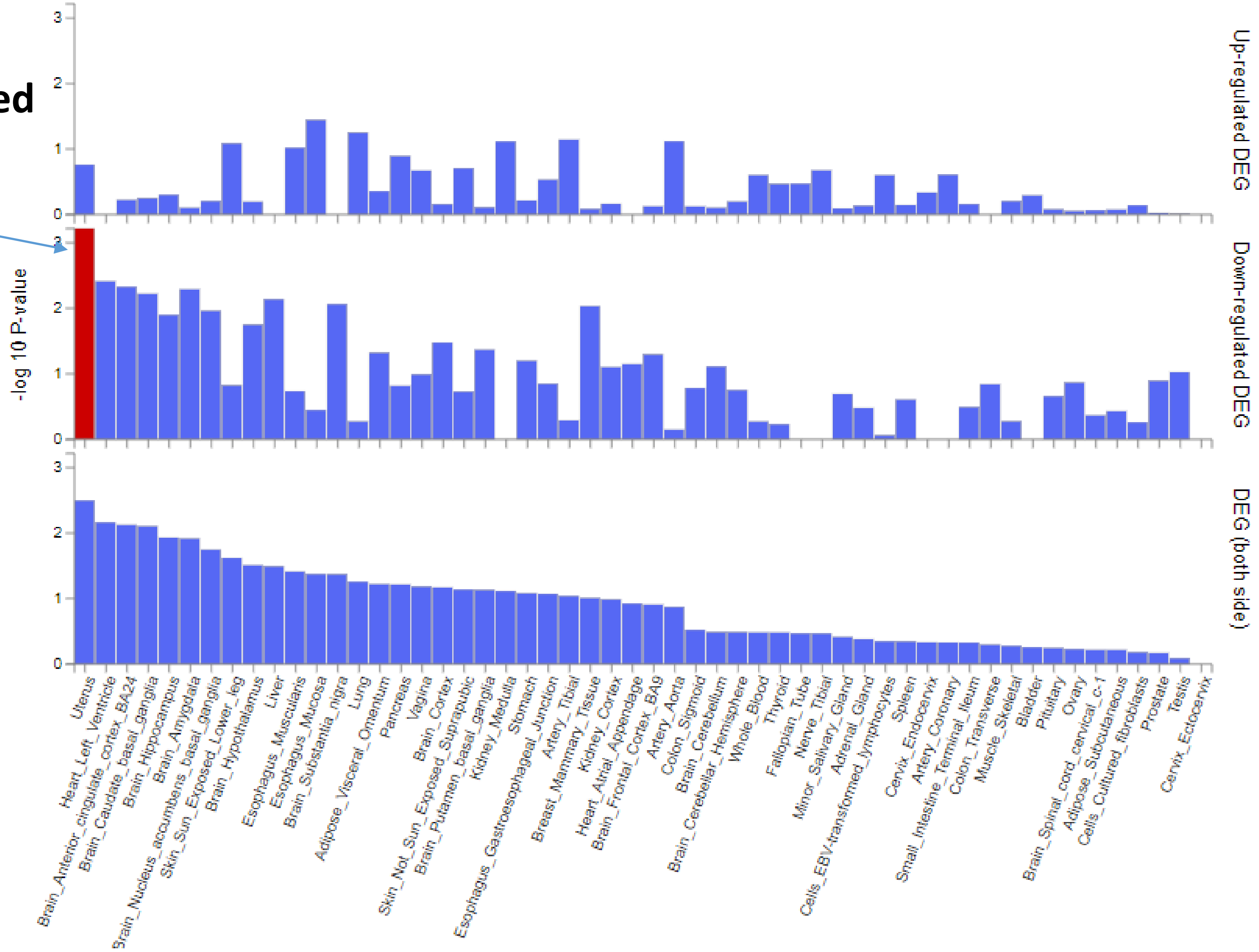
Summary of input genes

Number of input genes	32
Number of background genes	57241
Number of input genes with recognised Ensembl ID	26
Input genes without recognised Ensembl ID	CARMIL1, BABAM2, AL138720.1, LINC00824, AC007493.1, AF246928.1
Number of background genes with recognised Ensembl ID	57241
Background genes without recognised Ensembl ID	NA
Number of input genes with unique entrez ID	23
Number of background genes with unique entrez ID	35142

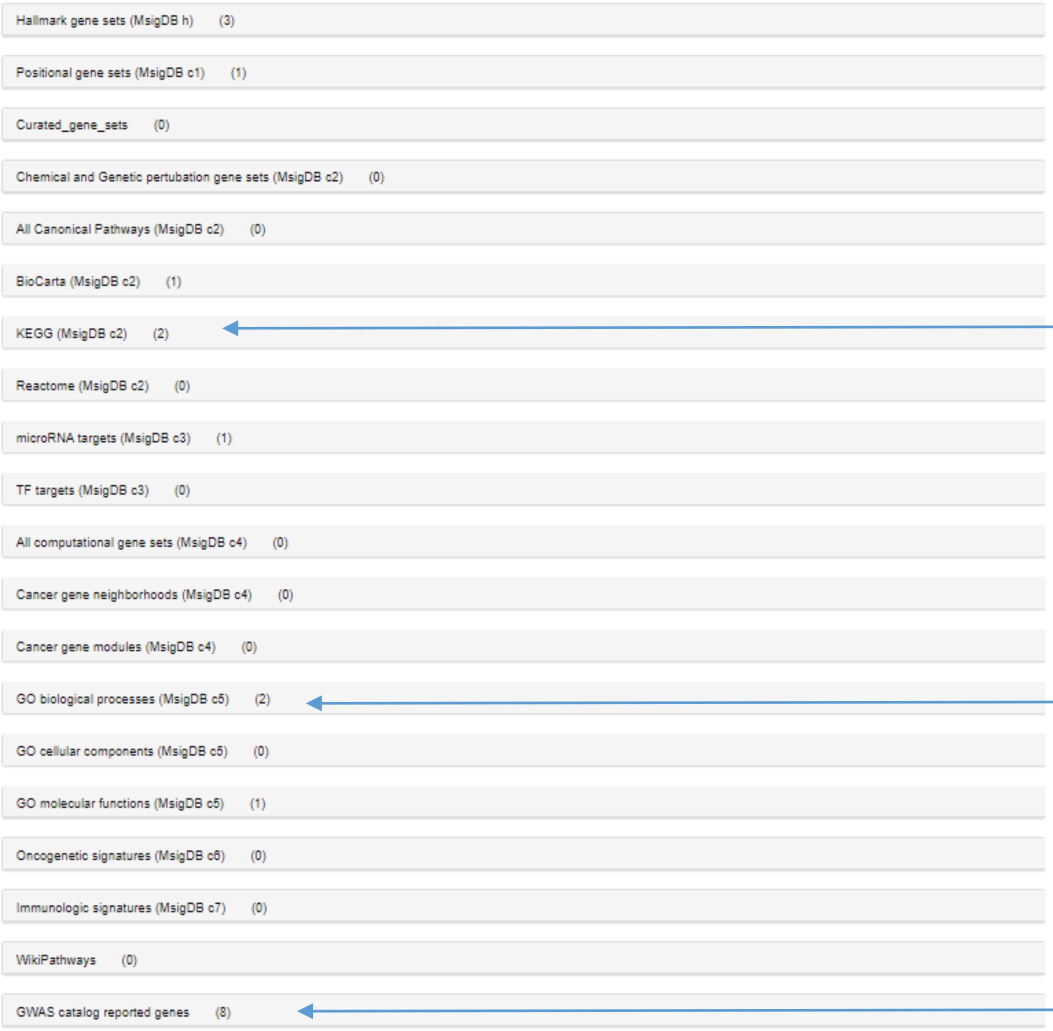
Download the plot as PNG JPG SVG PDF



Significantly differentially expressed



Enrichment : plots



Hallmark gene sets (MsigDB h) (3)

Positional gene sets (MsigDB c1) (1)

Curated_gene_sets (0)

Chemical and Genetic pertubation gene sets (MsigDB c2) (0)

All Canonical Pathways (MsigDB c2) (0)

BioCarta (MsigDB c2) (1)

KEGG (MsigDB c2) (2)

Reactome (MsigDB c2) (0)

microRNA targets (MsigDB c3) (1)

TF targets (MsigDB c3) (0)

All computational gene sets (MsigDB c4) (0)

Cancer gene neighborhoods (MsigDB c4) (0)

Cancer gene modules (MsigDB c4) (0)

GO biological processes (MsigDB c5) (2)

GO cellular components (MsigDB c5) (0)

GO molecular functions (MsigDB c5) (1)

Oncogenetic signatures (MsigDB c6) (0)

Immunologic signatures (MsigDB c7) (0)

WikiPathways (0)

GWAS catalog reported genes (8)

there are two signifiant pathways

there are two signifiant gene ontology

Informations found in GWAS catalog

Exercise

1. Classify SNPs list based on genomic location (genic and non genic)

2. Identify chromatin markers affected by given SNPs list .

3. Identify over represented KEGG pathways and GO enrichment based on SNPs encoding genes

4. Identify which tissue is differentially expressed due to given SNP list (via genes)?

rs4468290
rs11201609
rs4933212
rs701546
rs1241901
rs8087497
rs2409457
rs1666559
rs12943387
rs2036660