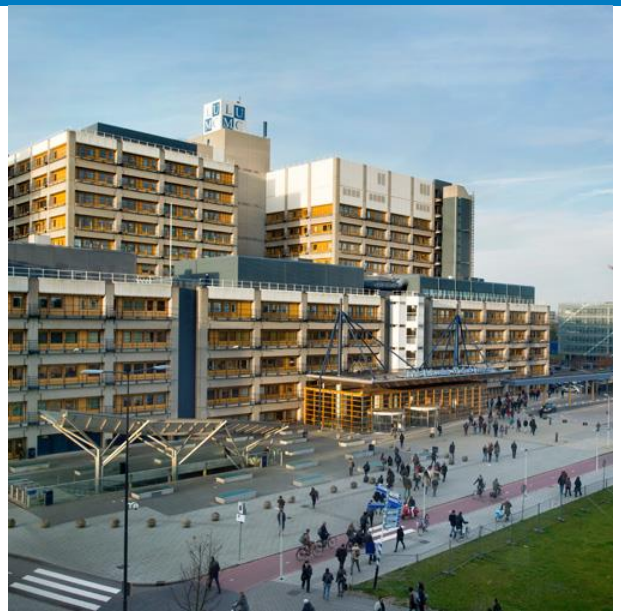
Leids Universitair
Medisch Centrum

Finding functional relevant genes



Yolande F. M. Ramos
y.f.m.ramos@lumc.nl
Molecular Epidemiology
LEIDEN, THE NETHERLANDS

AIMS of this lecture

- Understanding **genomic variation**, SNPs
- Functional **relevant variation**
- Use of online **databases**

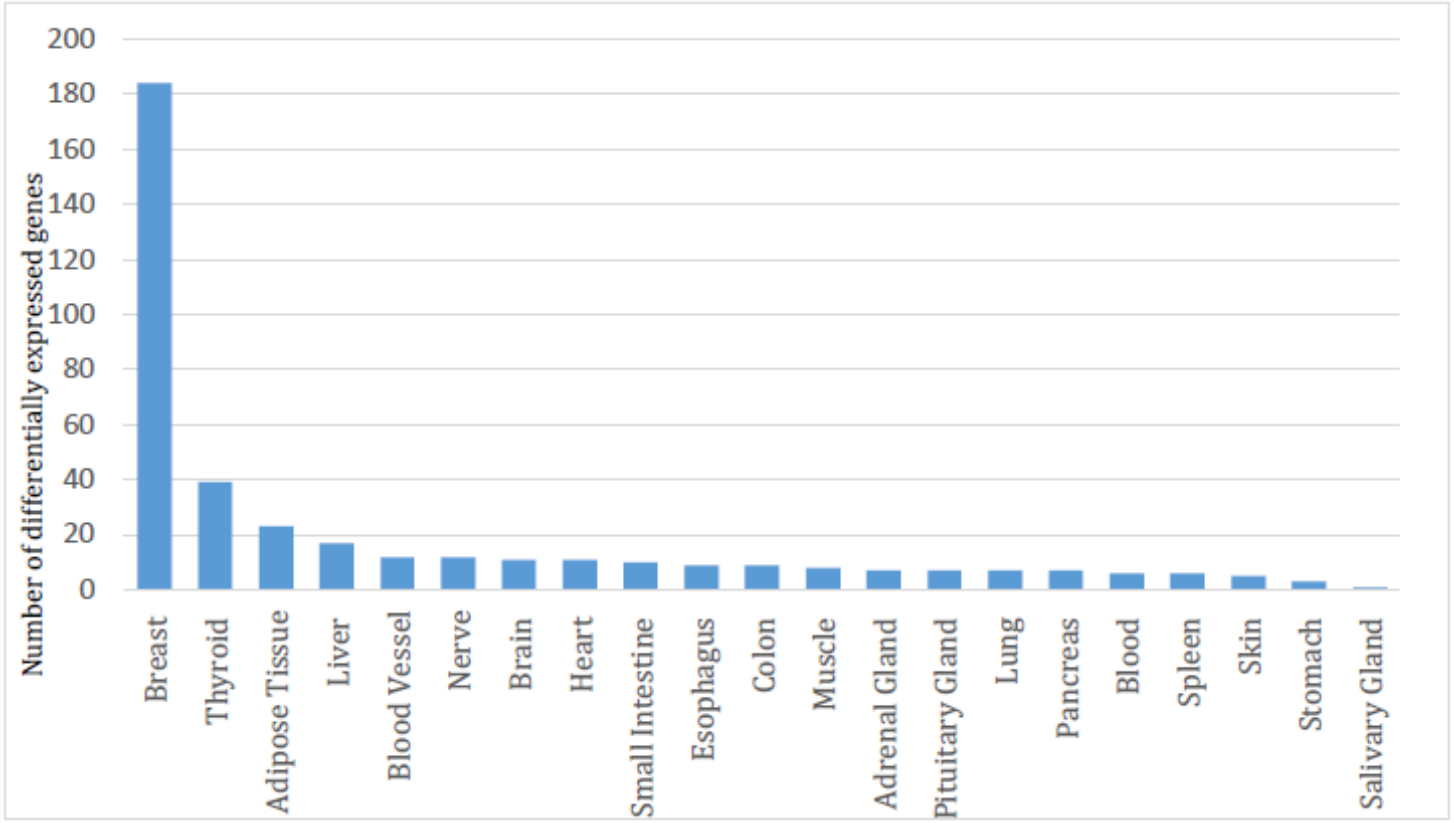
AIMS of this lecture

- Understanding **genomic variation**, SNPs
- Functional **relevant variation**
- Use of online **databases**

Human Genome

GENE TALLY

Scientists still don't agree on how many protein-making genes the human genome holds, but the range of their estimates has narrowed in recent years.



©nature

Willyard 2018; Pertea *et al* 2018, doi: http://dx.doi.org/10.1101/332825

Human Genome

- Consists of ~3.3 billion basepairs
- Total ~20,000 protein-coding genes
- Whole-genome sequence equals ~825 Mb
  - Data storage requires Terabytes!
- Sequence variation: repeats/deletions, SNPs…

Single Nucleotide Polymorphisms:
Specific nucleotides in the genome showing variation across the population in comparison to the reference sequence

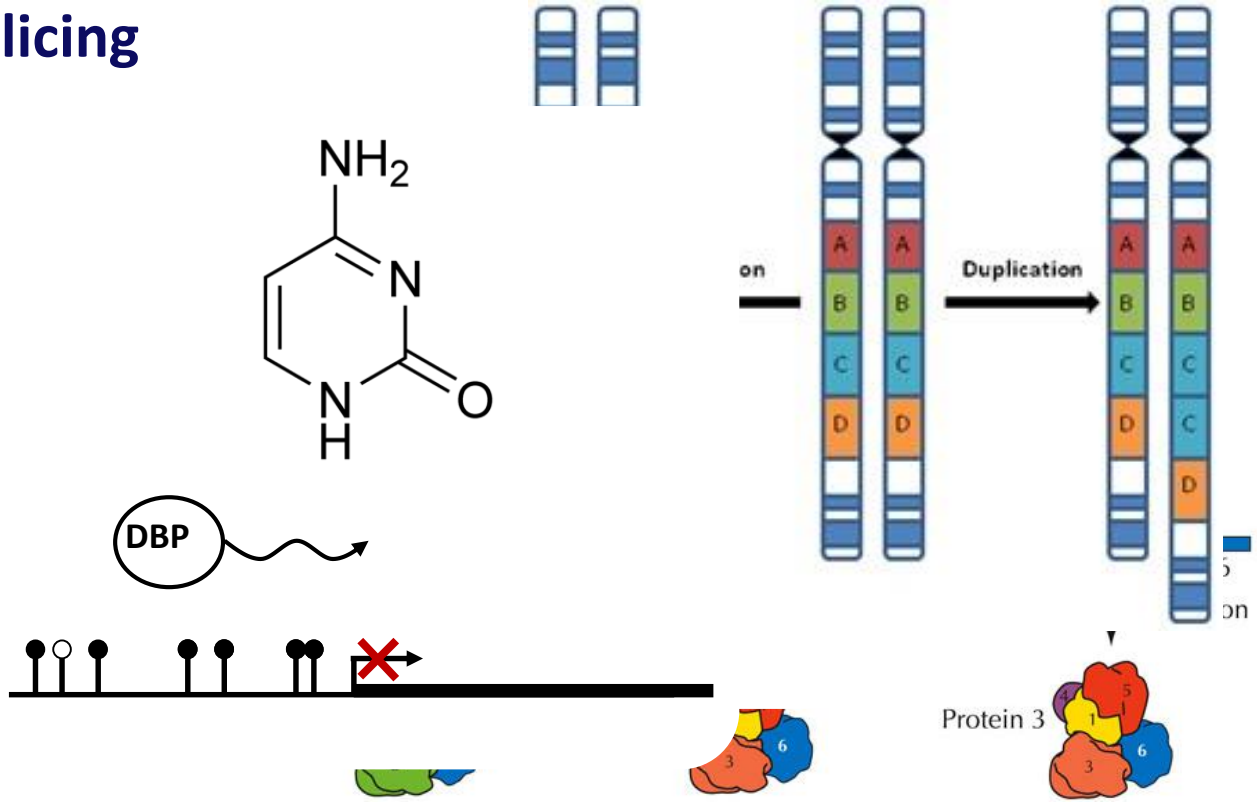
Single Nucleotide Polymorphism

AIMS of this lecture

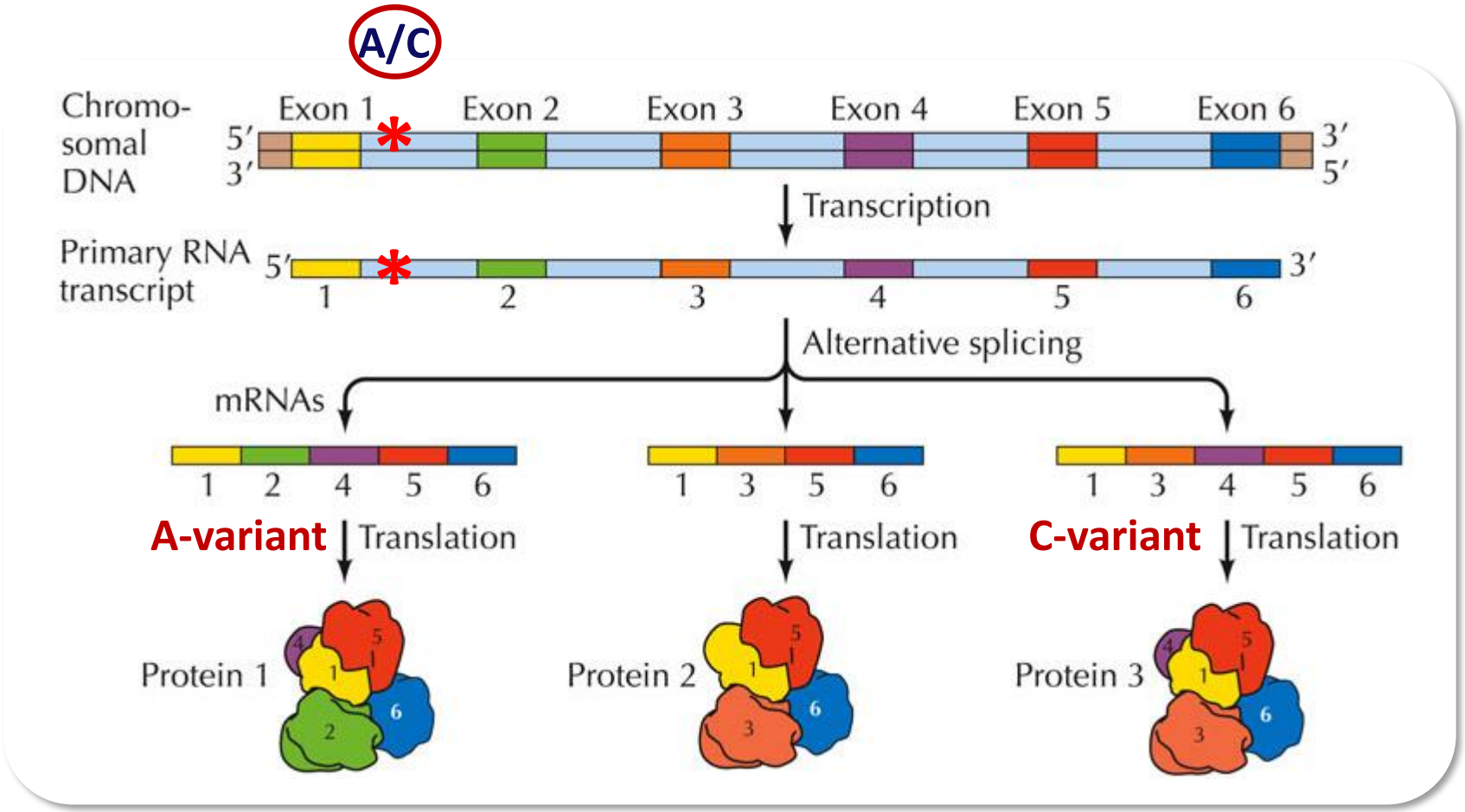
- Understanding **genomic variation**, SNPs
- Functional **relevant variation**
- Use of online **databases**

Functional variants

- **Methylation**
- **Copy number variation (CNV)**
- **Alternative splicing**
- **Single Nucle**
  - Intronic
  - Exonic
  - Promotor
  - Regulator



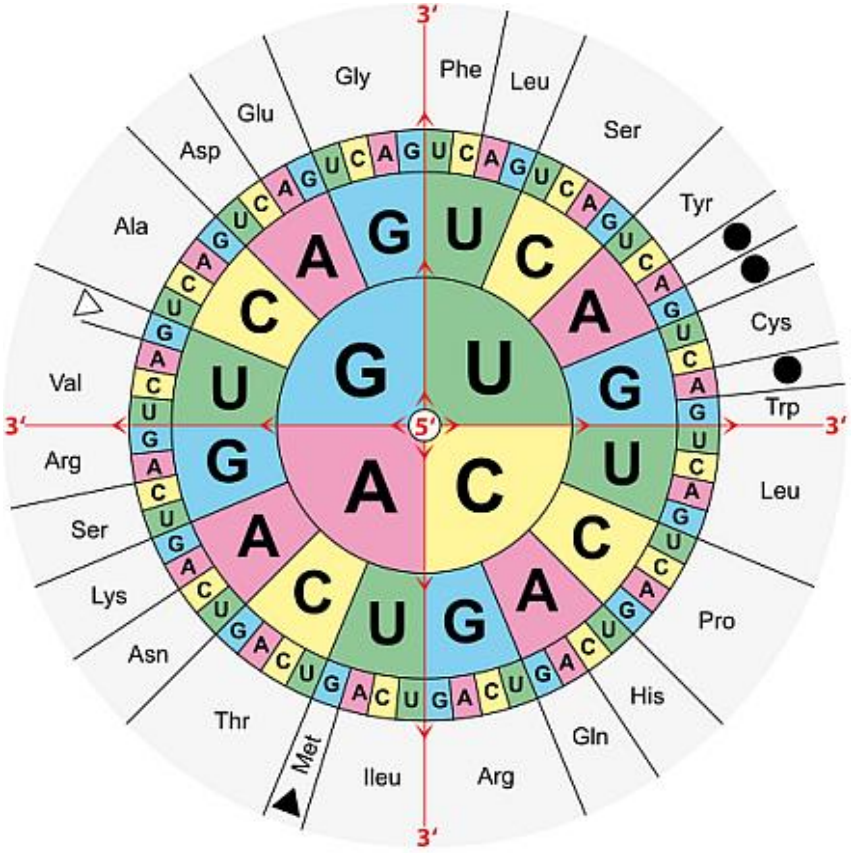
Intronic SNPs



Exonic SNPs

Coding SNPs:

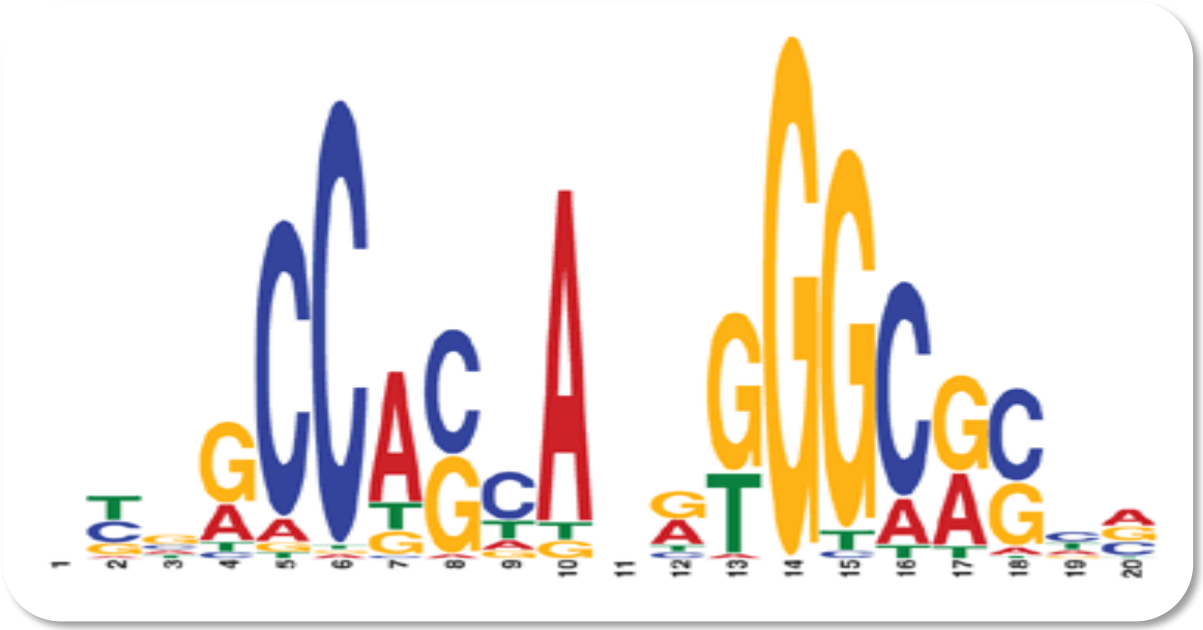
- Synonymous or **non**-synonymous



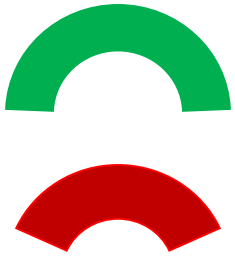
Codon: G A A Glu
Codon: G A **G** Glu
Codon: G **G** A Gly
Codon: **T** A A **STOP**

Promotor SNPs

Disruption of protein recognition sites

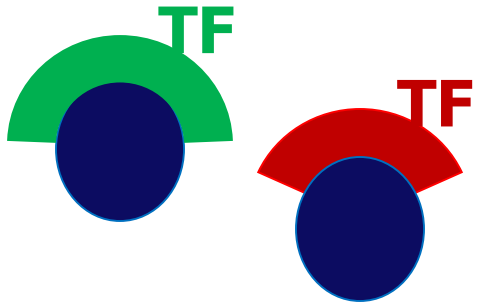


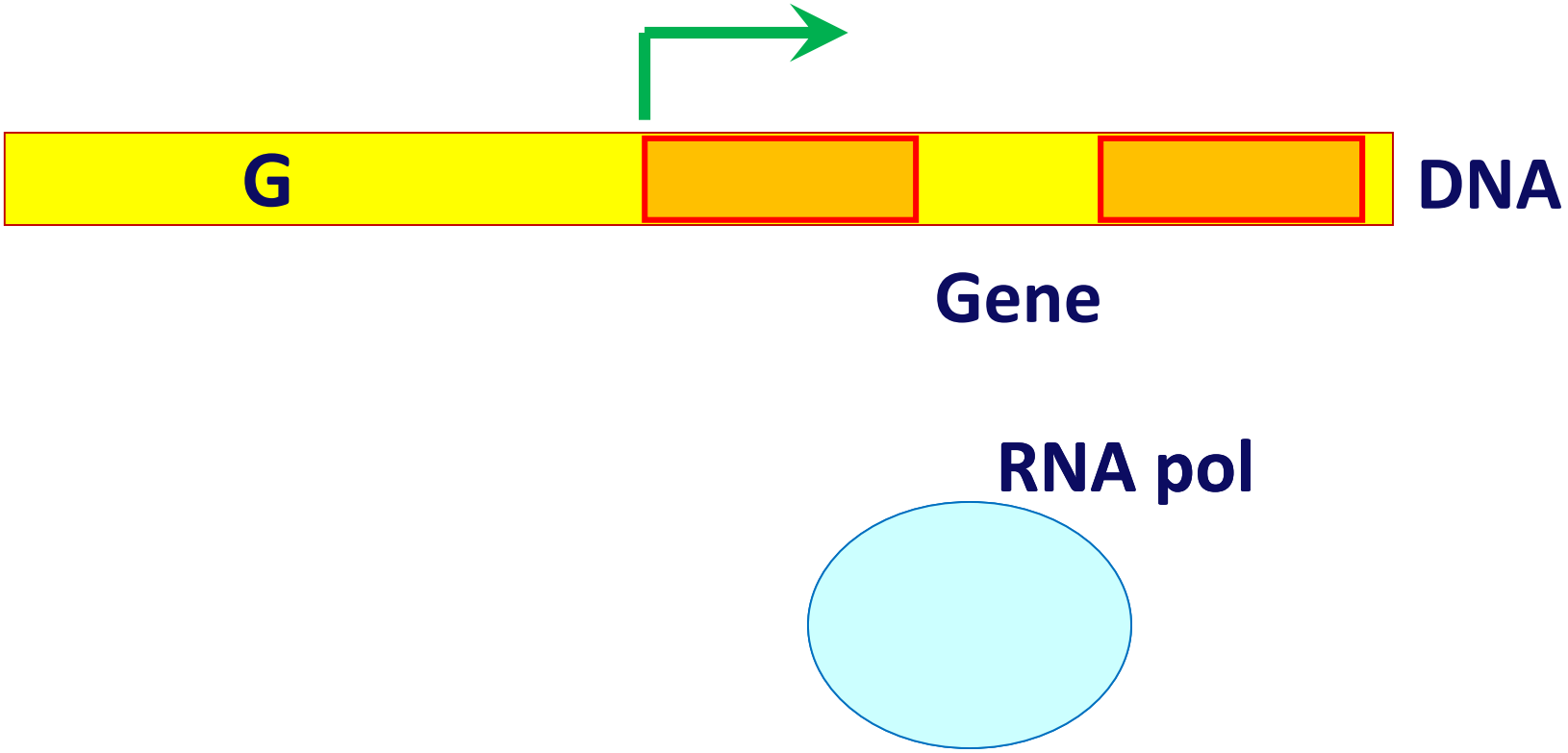
Promotor SNPs



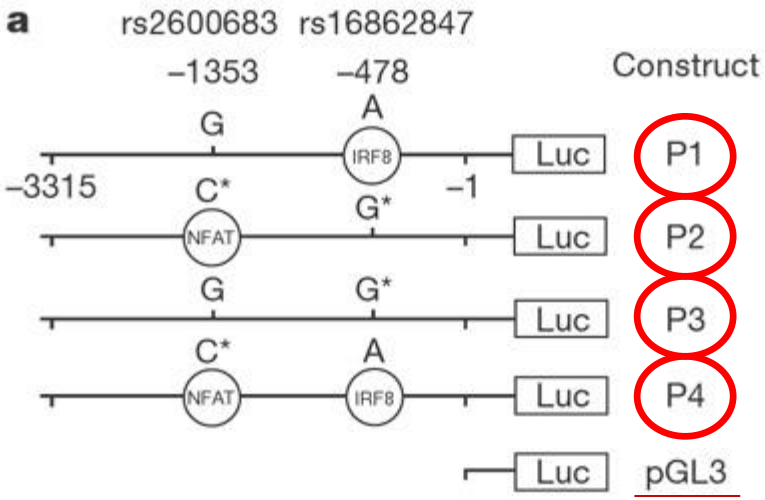
SNP G recognition

SNP A recognition

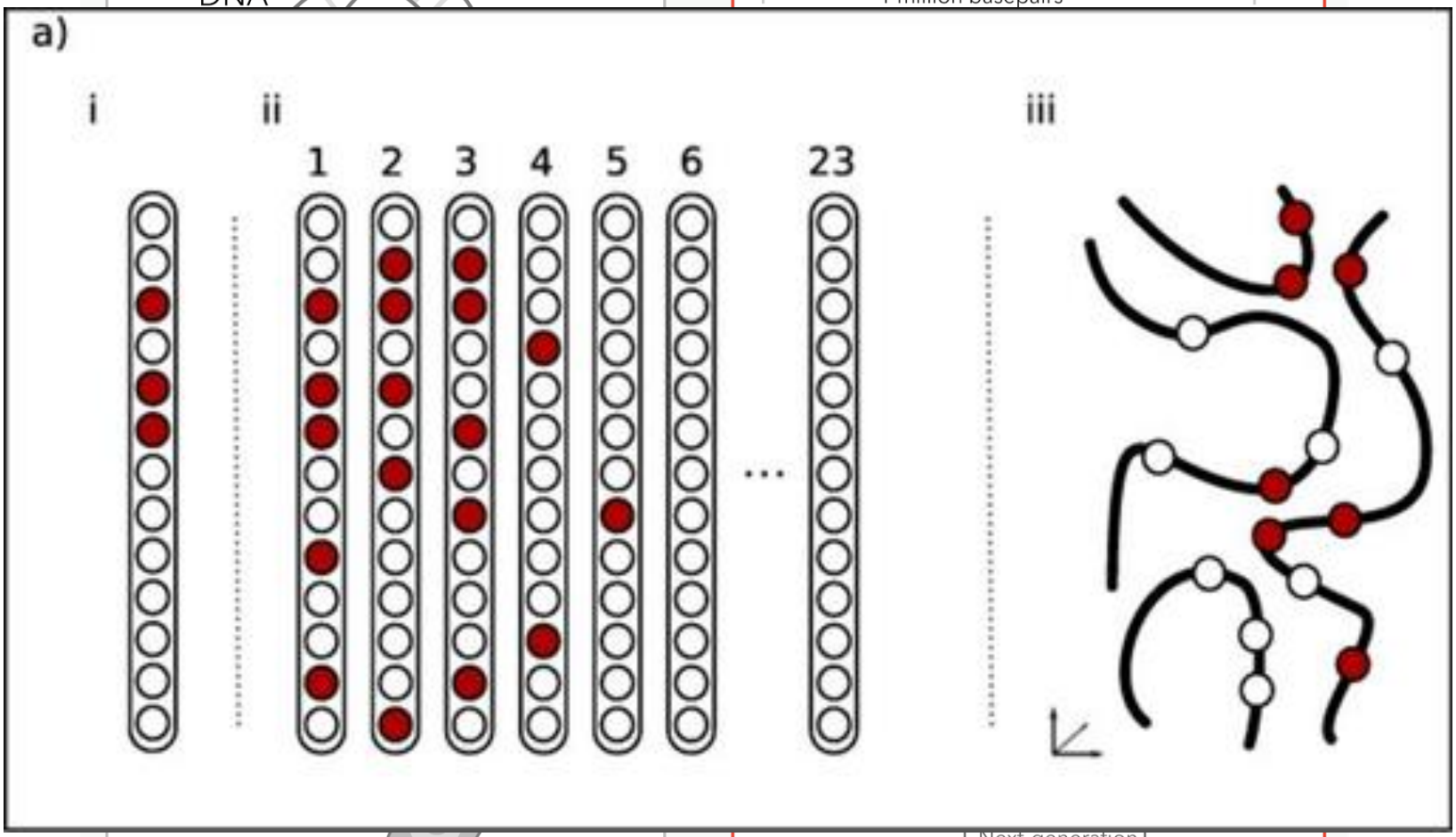




Promotor SNPs

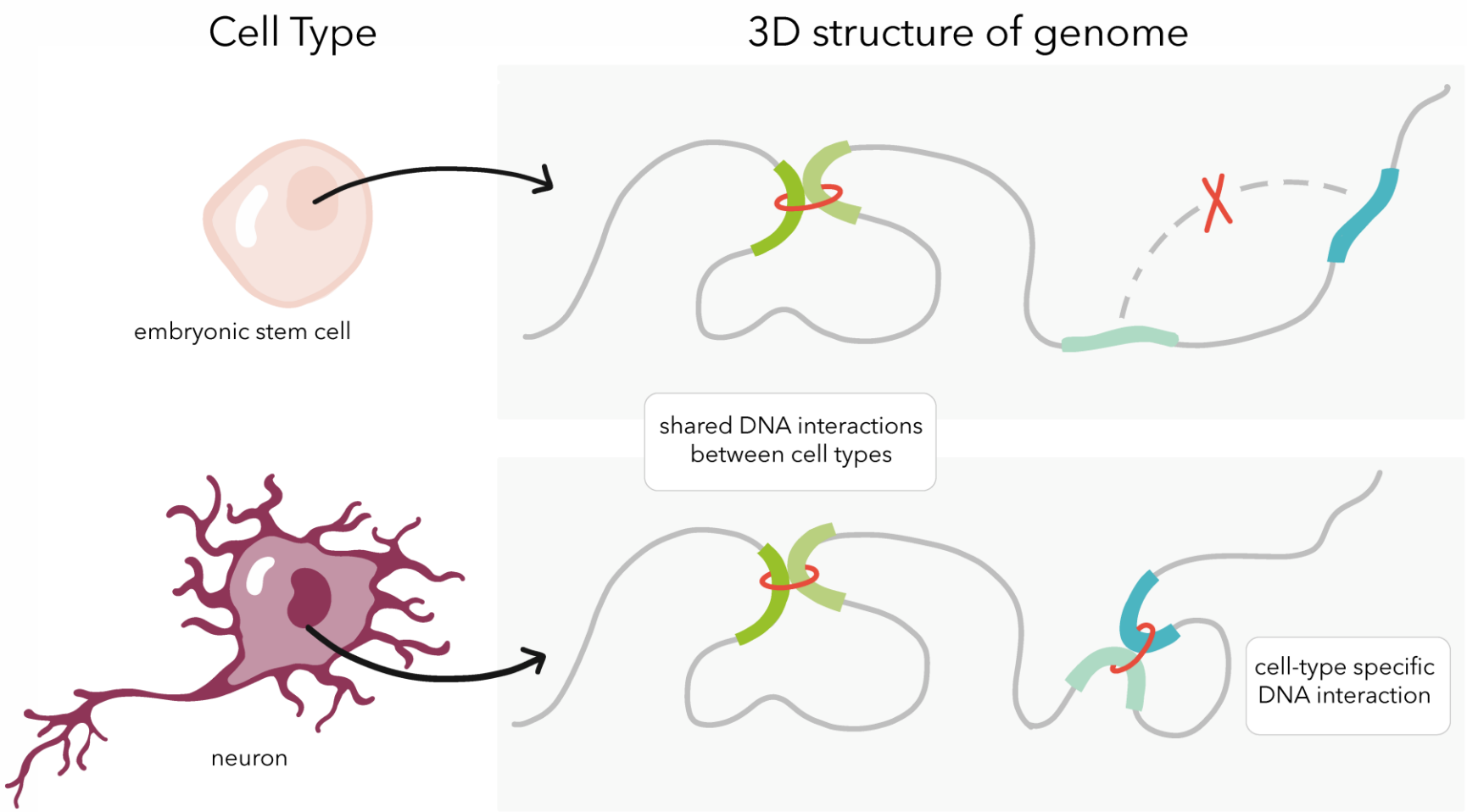


Distal regulatory elements



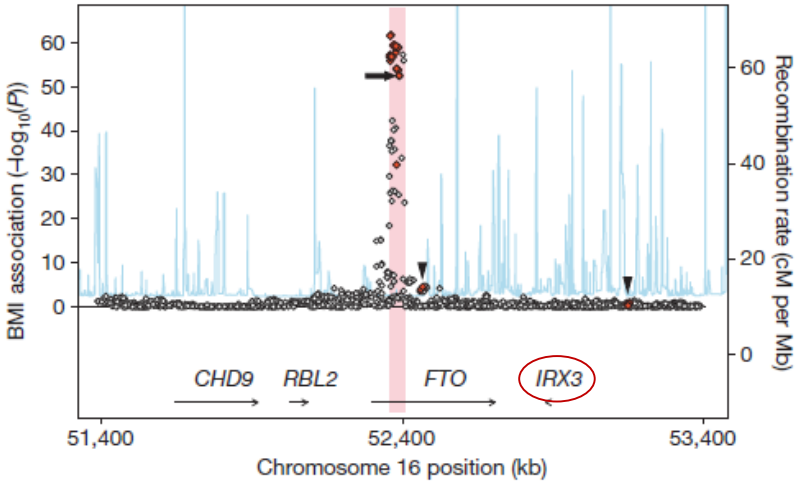
Thévenin *et al* 2014; doi: 10.1093/nar/gku667; Johnson 2018

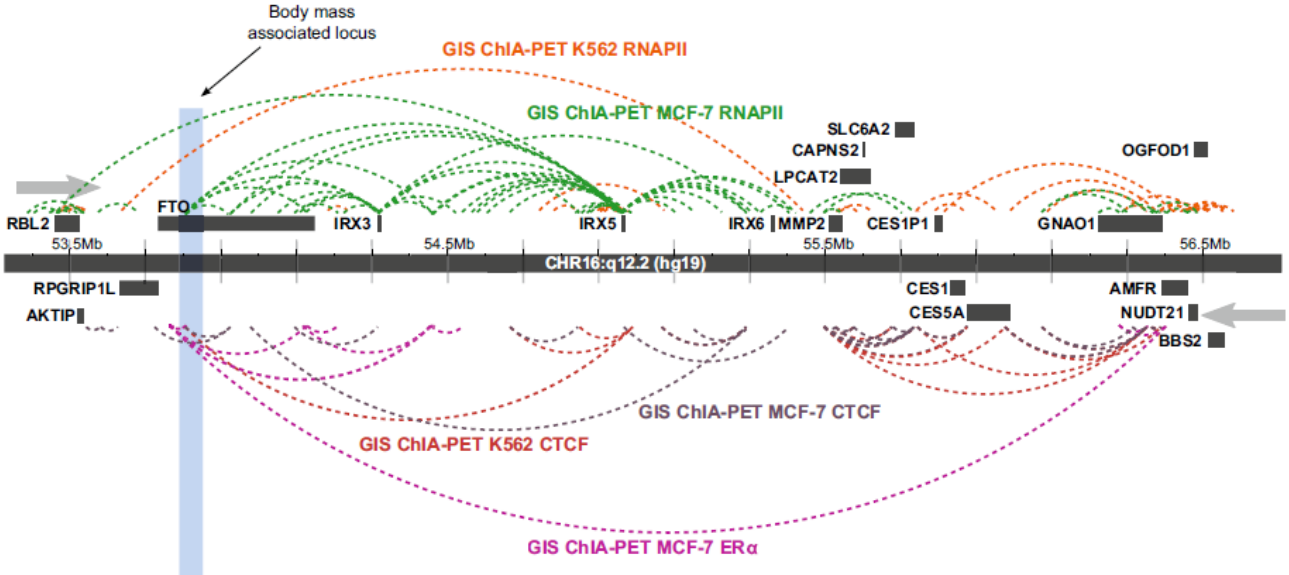
Distal regulatory elements

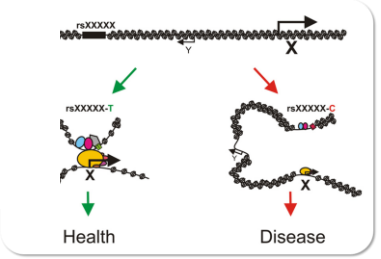


https://www.youtube.com/watch?v=1Fyq9ul9N9Q&feature=youtu.be

Distal regulatory elements







Smemo *et al* 2014, doi:10.1038/nature13138; Rask-Andersen *et al* 2015, doi:10.1007/s00439-015-1599-5

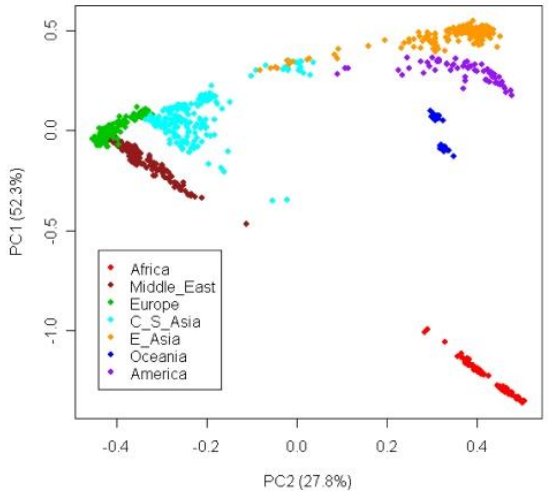
Summary – disease relevant SNPs

Functional SNPs

- **Alternative splicing**
- **Codon change**
- **Variation in regulatory elements**
  Promotor
  Distal regulatory element

Criteria relevant SNPs

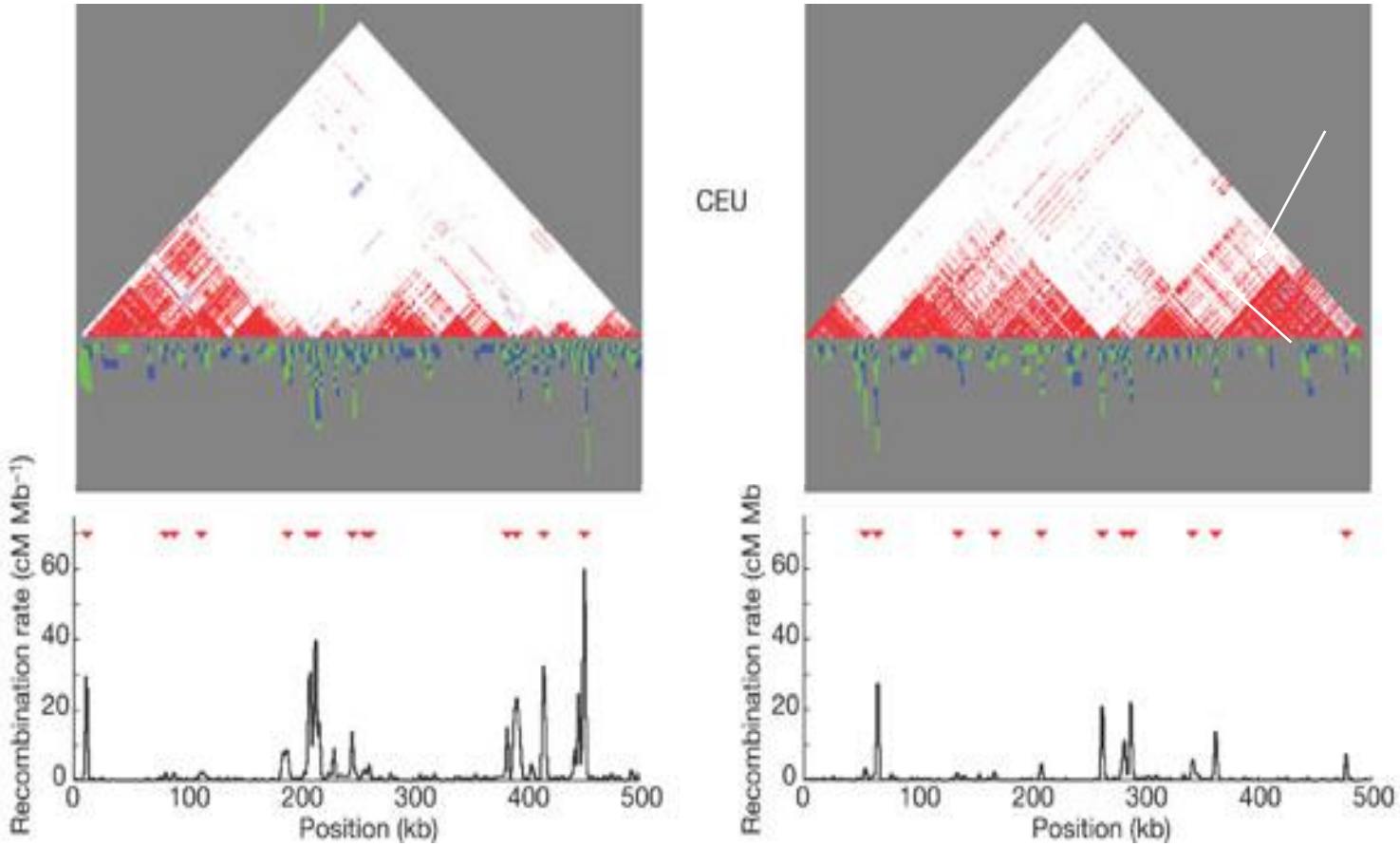
- Functional relevant variant
- Polymorphic in specific population
- Assay design possible (repeat, GC-rich)
- Tagging SNP (possibly: tagging functional SNP)



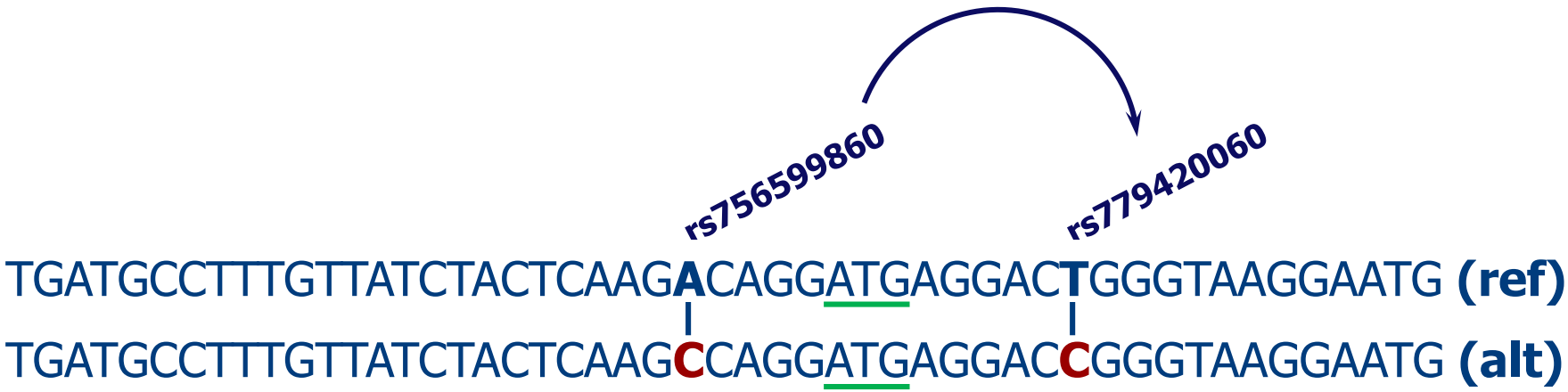
Li et al 2008, doi:10.1126/science.1153717

Tagging SNP

Linkage across ‘blocks’: genetic variation is limited (several SNPs carry same information)



Tagging SNP


rs756599860
rs779420060
TGATGCCTTTGTTATCTACTCAAGACAGGATGAGGACTGGGTAAGGAATG (ref)
TGATGCCTTTGTTATCTACTCAAGCCAGGATGAGGACCGGGTAAGGAATG (alt)

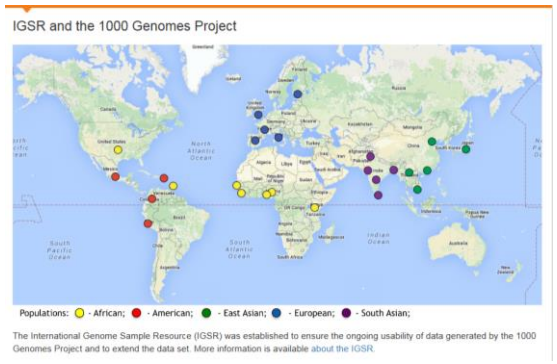
AIMS of this lecture

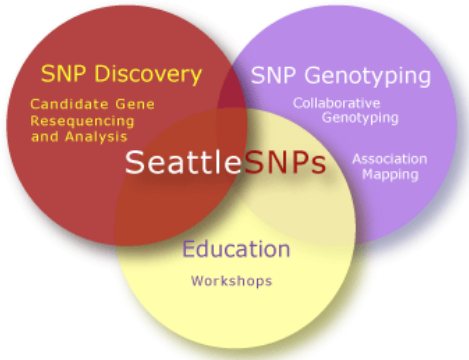
- Understanding **genomic variation**, SNPs
- Functional **relevant variation**
- Use of online **databases**

Human Genome

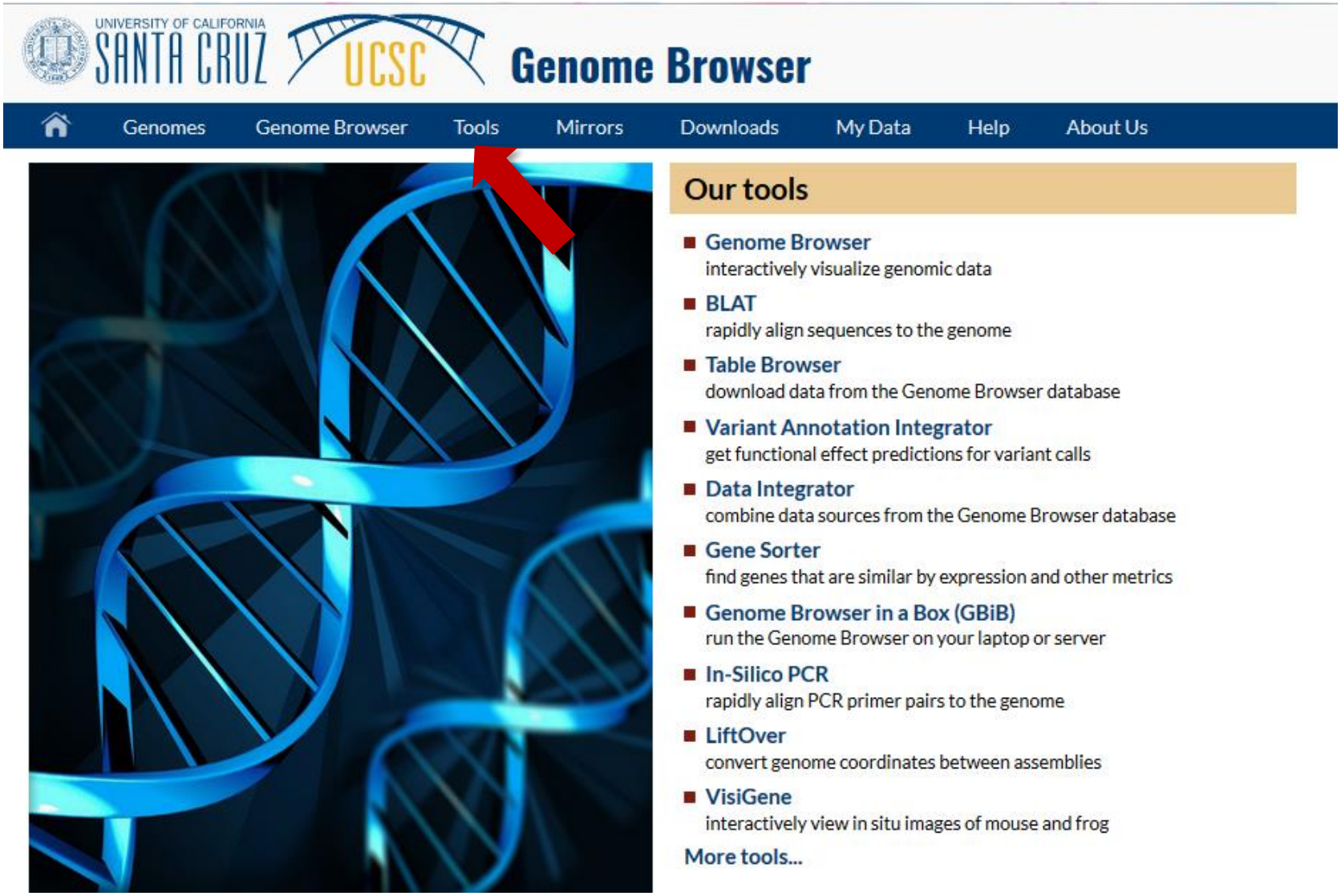
- Consists of ~3.3 billion basepairs
- Whole-genome sequence equals ~825 Mb
  - Data storage requires Terabytes!
- Online databases contain information on thousands of individuals and millions of polymorphisms

Entrez Gene

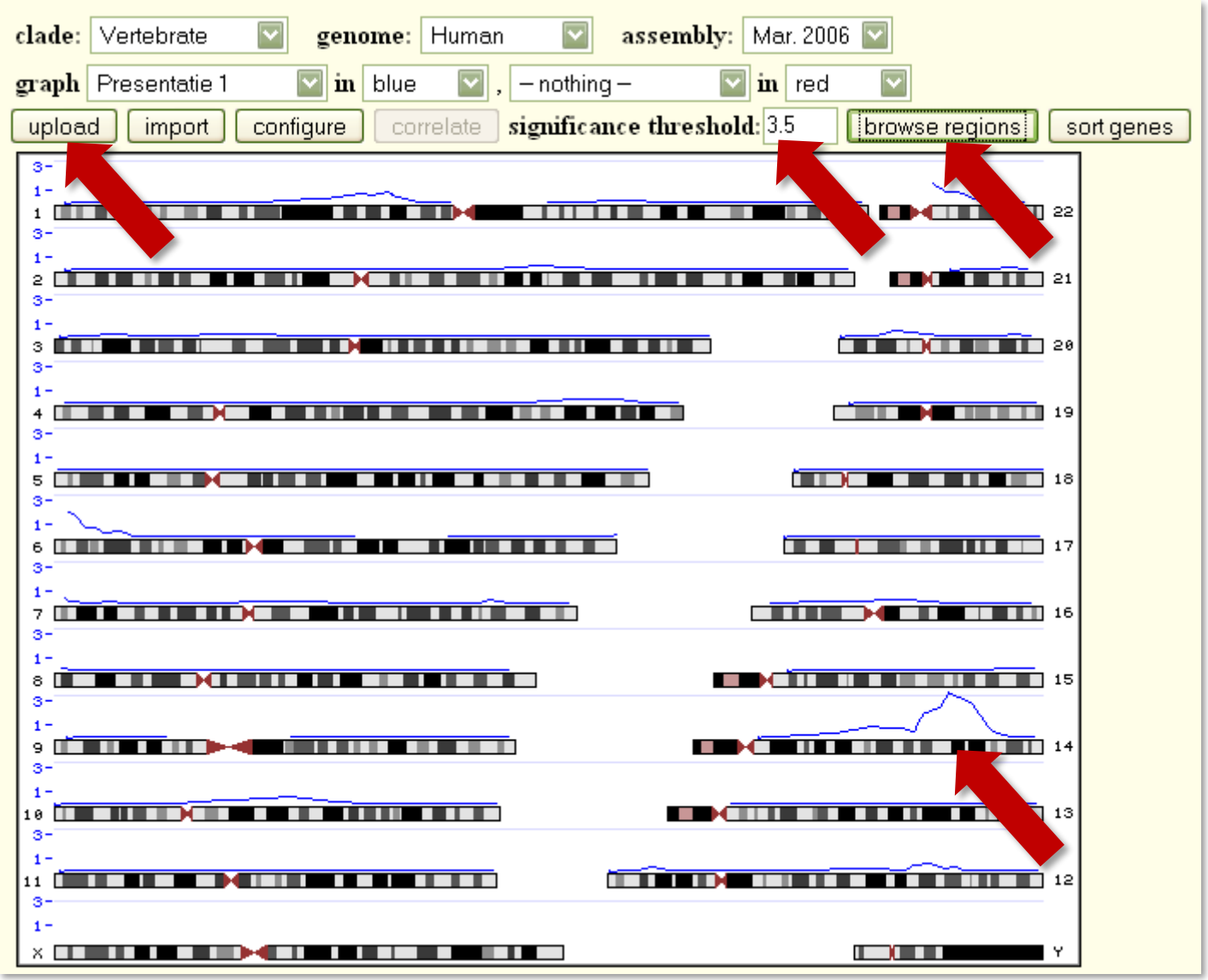




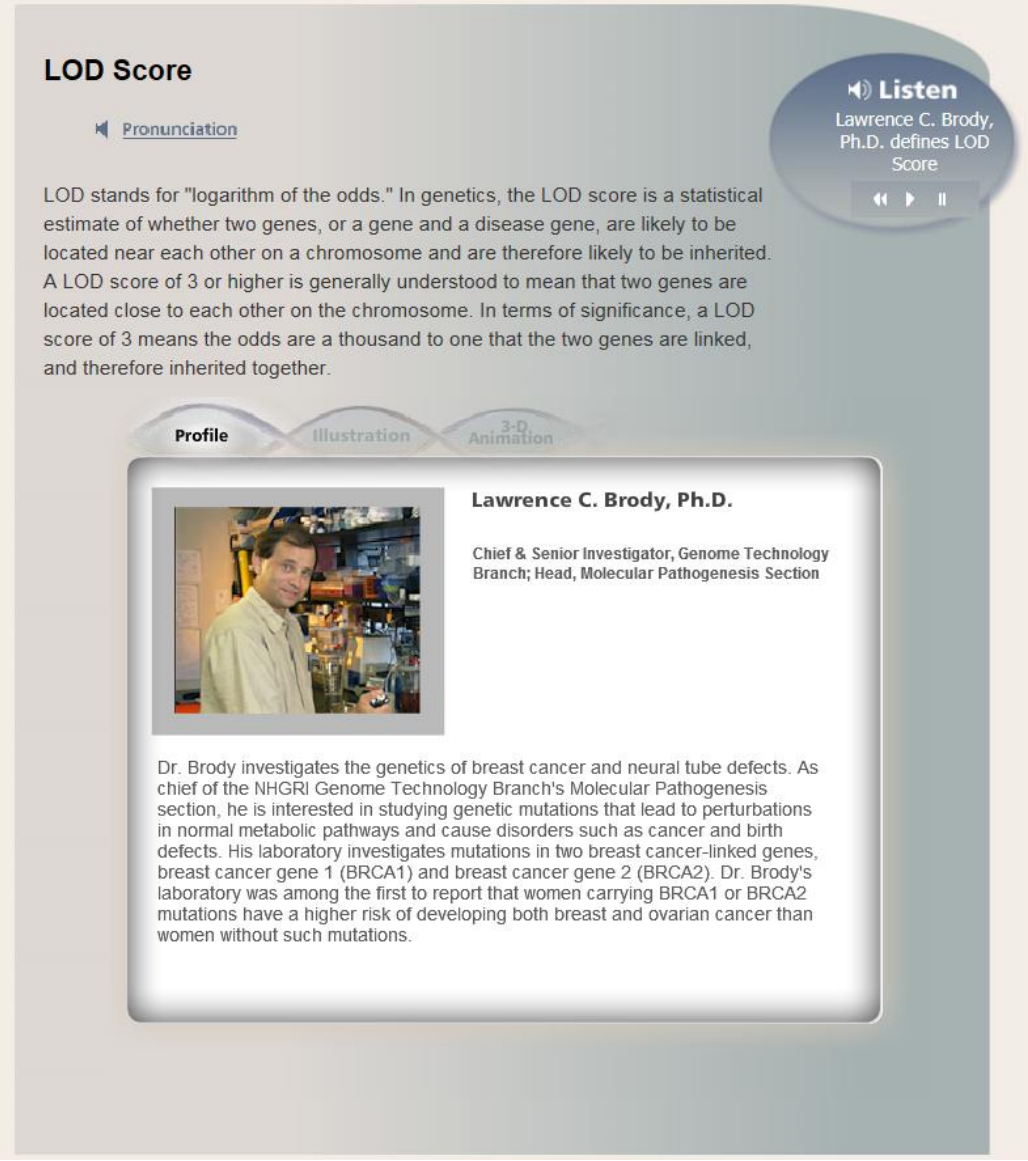
Example case – UCSC Genome Browser



Example case – Genome Graphs



Example case - Linkage Datafile

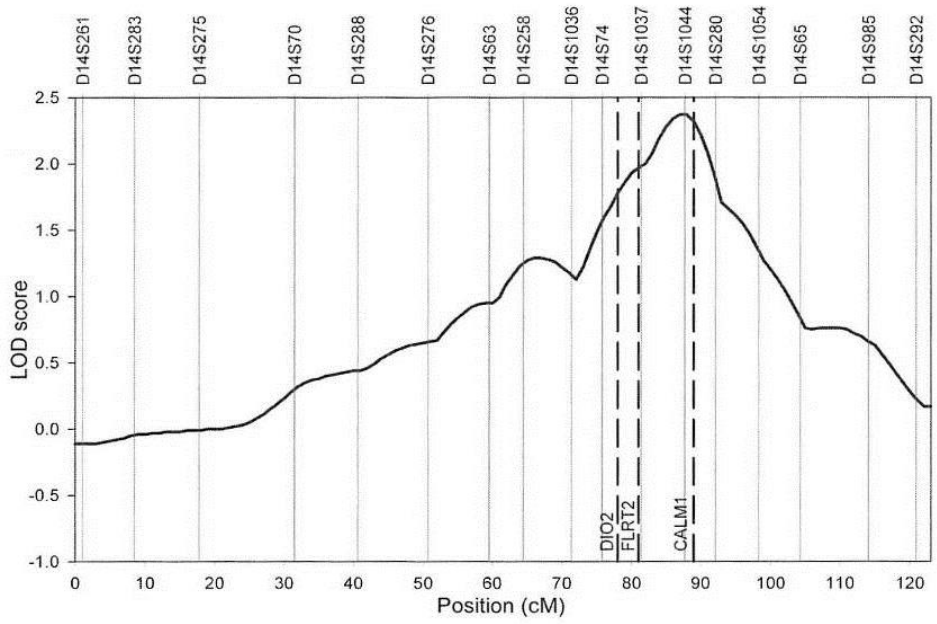


LOD score:

logarithm (log10) of the odds; for LOD=3 the odds is 10^3 (1000) to 1 that the genomic region is linked with trait tested.

Example case - Linkage Datafile

- Linkage at chromosome **14q32.11**
- 3 genes within 1-LOD drop interval
  - *DIO2*
  - *FLTR2*
  - *CALM1*



Marker	LOD
D1S214	0
D1S450	0
D1S2667	0
D1S2697	0
D1S199	0.05
D1S234	0.104
D1S255	0
D1S2797	0.006
D1S2890	0.049
D1S230	0.014
D1S2841	0.326
D1S207	0.298
D1S2868	0.73
D1S2793	0.599
D1S206	0.876
D1S495	0.733
D1S2626	0.667
D1S2778	0.218

Meulenbelt *et al* 2008, doi:10.1093/hmg/ddn082

Surplus of information

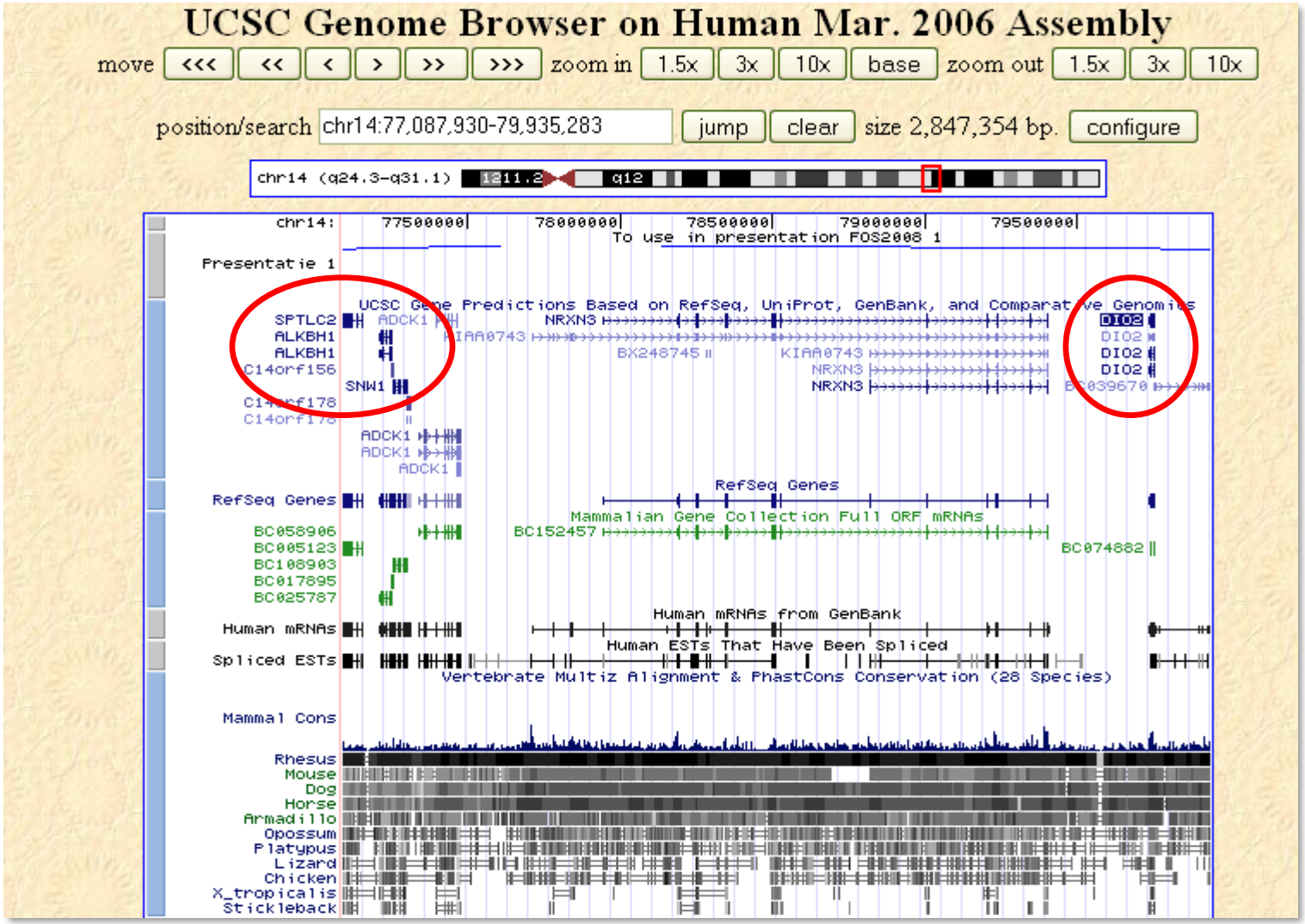
Filter information by:

- Population of interest (CEU, CHB+JPT, etc.)
- Chromosome / locus
- Gene
- Intronic / Exonic / Promotor / 5’ UTR/…
- Basepair

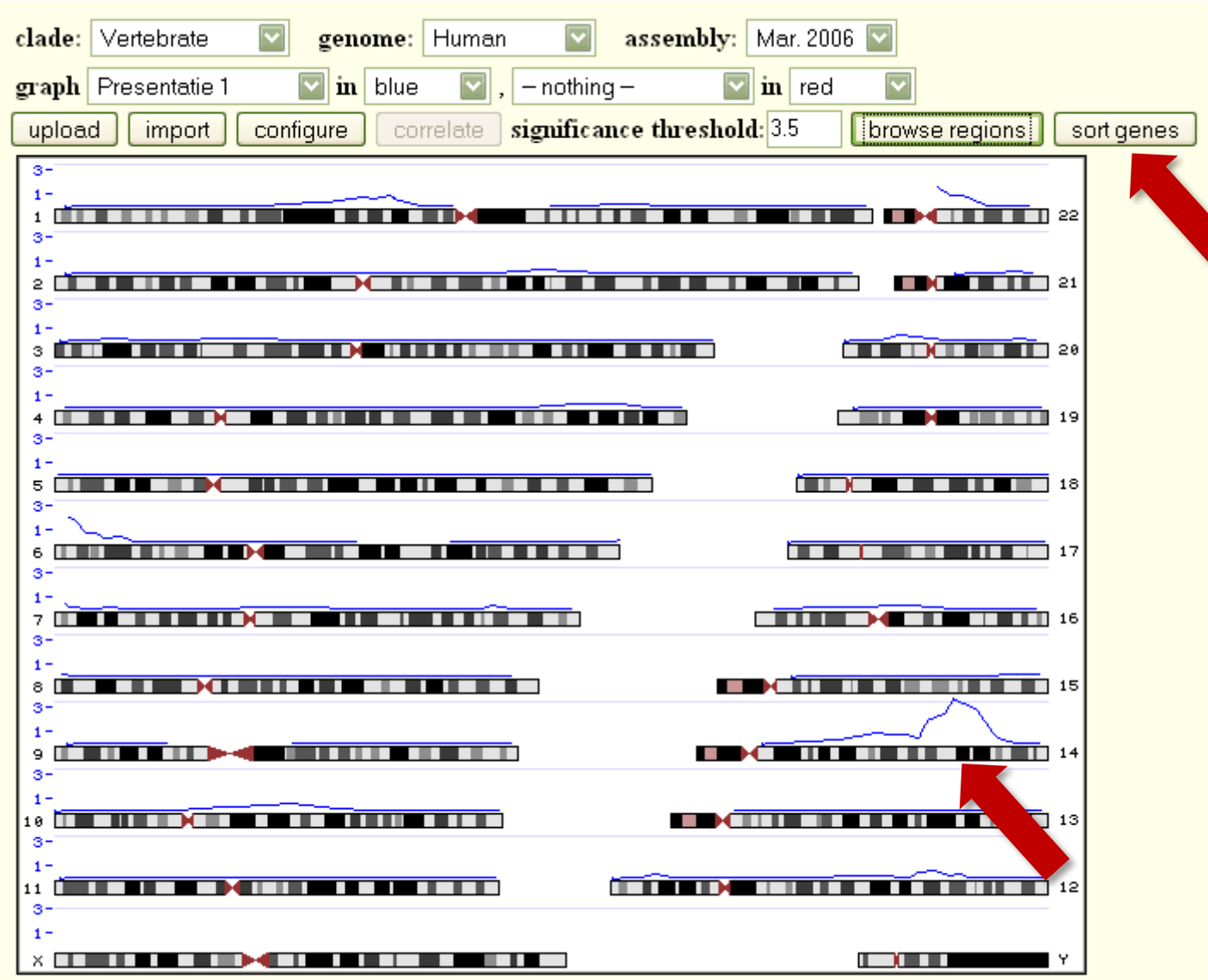
(Some) criteria for disease relevant SNPs:

- If exonic: non-synonymous; damaging?
- If not: in regulatory element?

Example case - Genome Browser



Example case - Genome Graphs



Example - Gene sorter

Home Genomes Blat Tables Gene Sorter PCR PDF/PS Session FAQ Help

Finding Candidate Genes for Gene Sorter

Thresholding *Presentatie 1* at 3.4. There are 1 regions covering 4246031 bases.
Installed a Gene Sorter filter that selects only genes in these regions.

go to gene sorter

Home

UCSC Human Gene Sorter

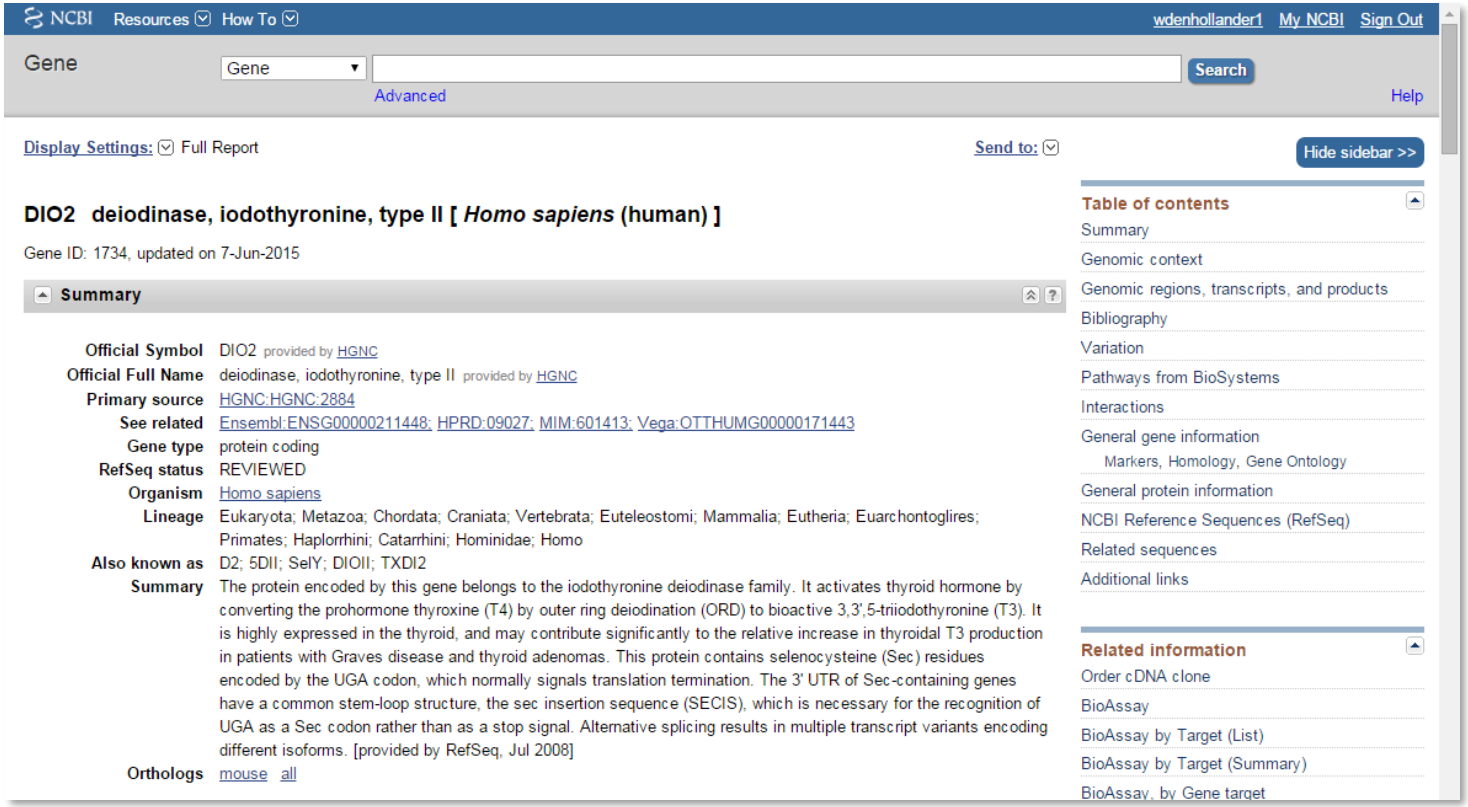
genome Human assembly Mar. 2006 search uc001aaa.1 Go!
sort by Expression (GNF Atlas2) configure filter (now on) display 50 output sequence text

#	Name	VisiGene	fetal brain	whole brain	amygdala	thymus	bone marrow	PB-CD4+ Tcells	skin	adipocyte	pancreatic islets	heart	lung	kidney	liver	ovary	testis	BLASTP E-Value	Genome Position	Description
1	TMEM63C	172819																n/a	chr14 76,756,757	transmembrane protein 63C
2	NGB	179188																n/a	chr14 76,804,497	neuroglobin
3	POMT2	179894																n/a	chr14 76,834,011	putative protein O-mannosyltransferase
4	GSTZ1	176852																n/a	chr14 76,862,400	glutathione transferase zeta 1 isoform 1
5	TMED8	n/a																n/a	chr14 76,895,098	transmembrane emp24 domain containing 8
6	C14orf174	177840	n/a					n/a					n/a					n/a	chr14 76,920,427	hypothetical protein LOC161394
7	C14orf148	172443	n/a					n/a					n/a					n/a	chr14 76,950,867	hypothetical protein LOC122945
8	C14orf133	168163																n/a	chr14 76,978,214	hypothetical protein LOC63894

Description and known literature

- Gene function — **NCBI / OMIM**
- Literature — **PubMed**
- Expression
  - eQTLs — **Genevar / GTEx**
  - Tissue of interest — **BioGPS / GTEx**
  - Alternative splicing — **Ensembl / GTEx**
- Known genetic variation — **UCSC / HaploReg / SNPper**

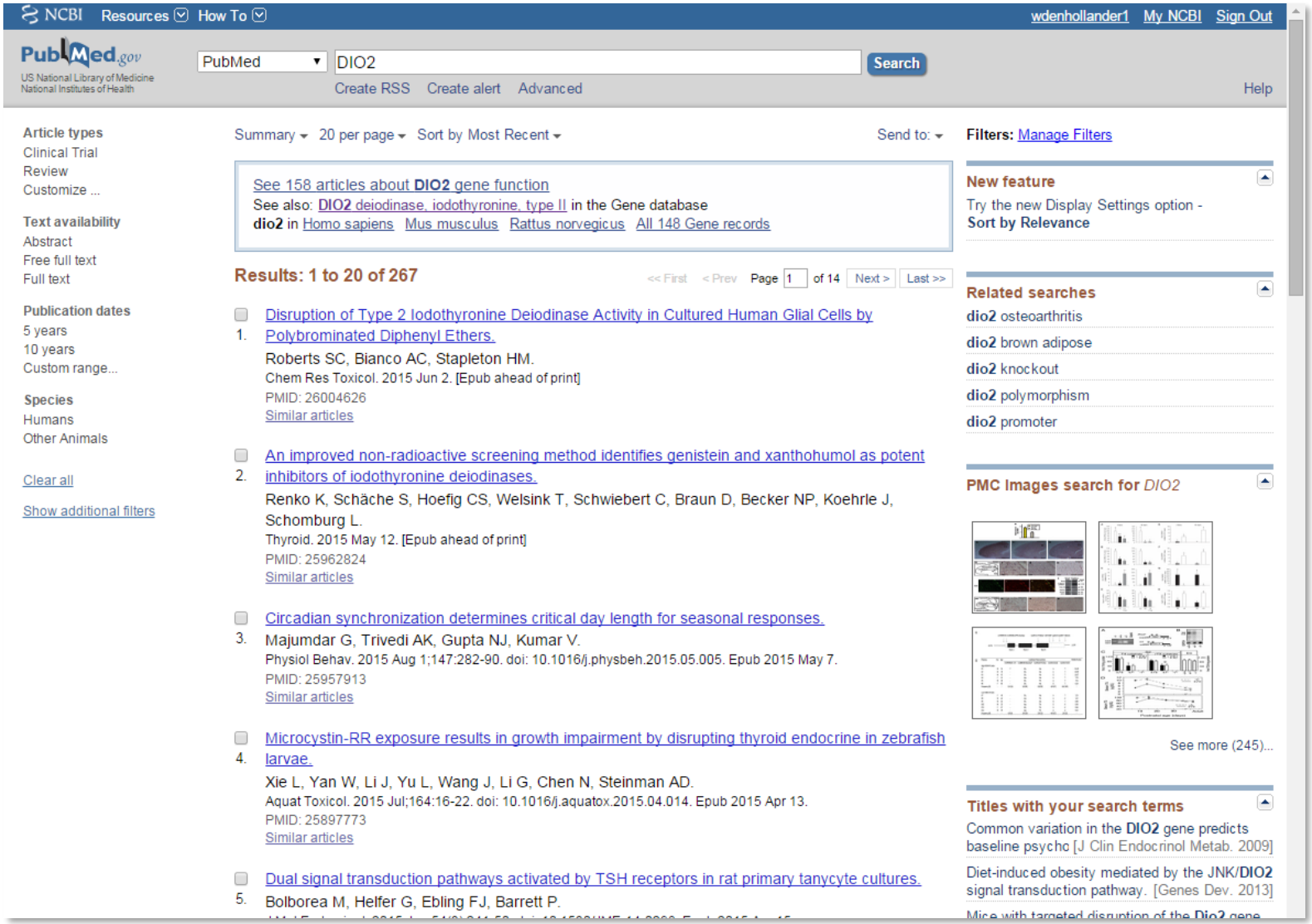
NCBI – Entrez Gene



GeneRIFs!

Gene Reference into Function

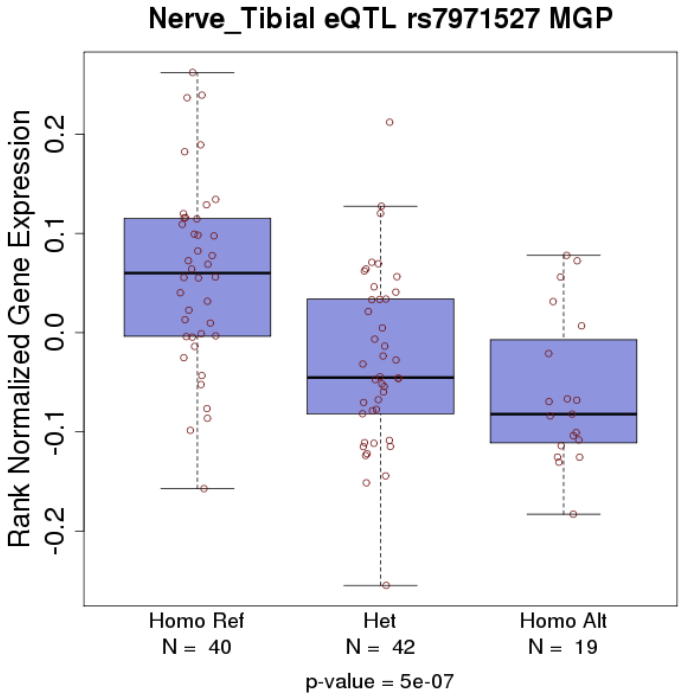
Literature - PubMed



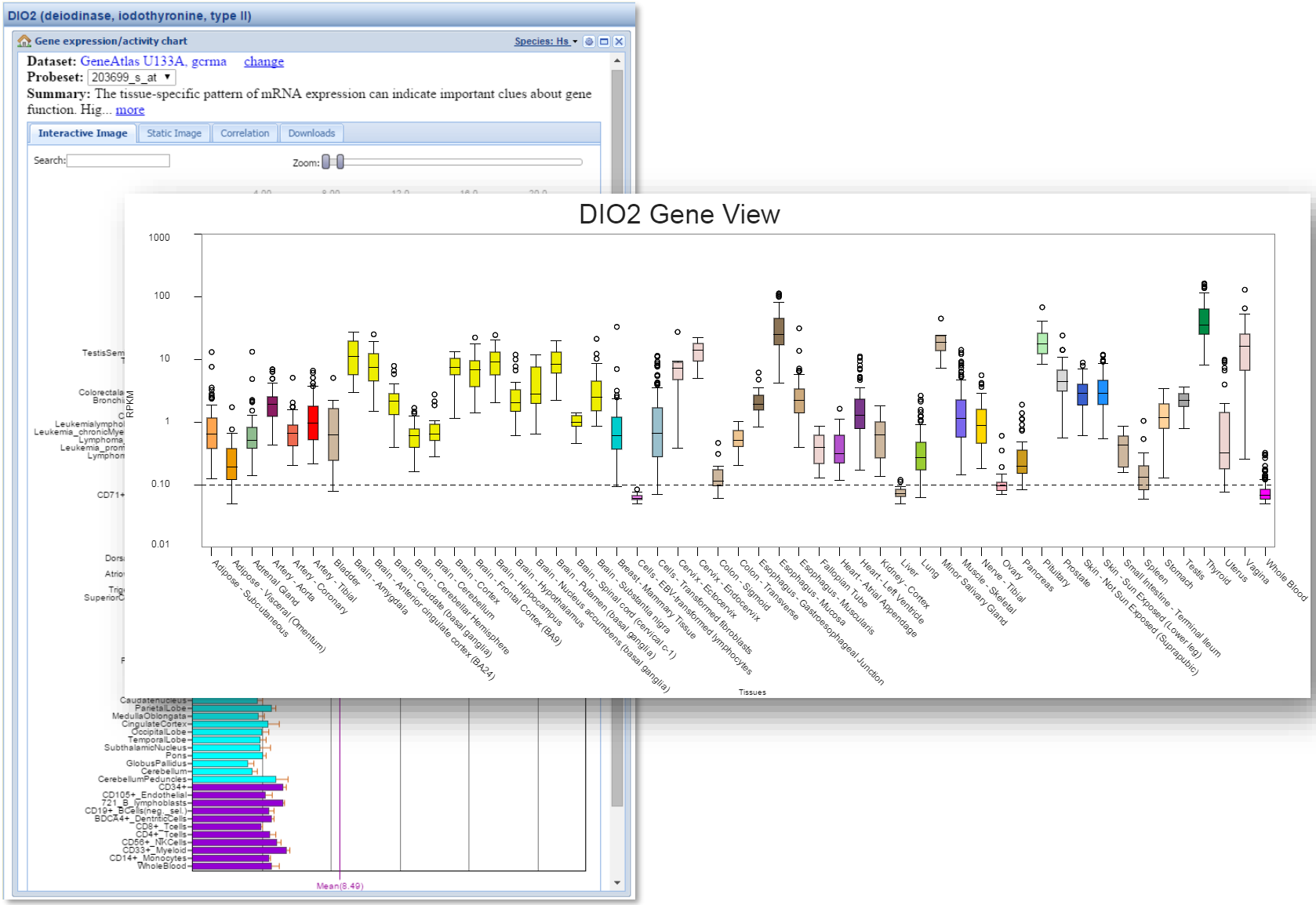
eQTLs – Genevar & GTEx

Expression Quantitative Trait Locus

- SNP affects gene expression
- Either *in cis* or *in trans*



Expression – Tissue of Interest (BioGPS & GTEx)



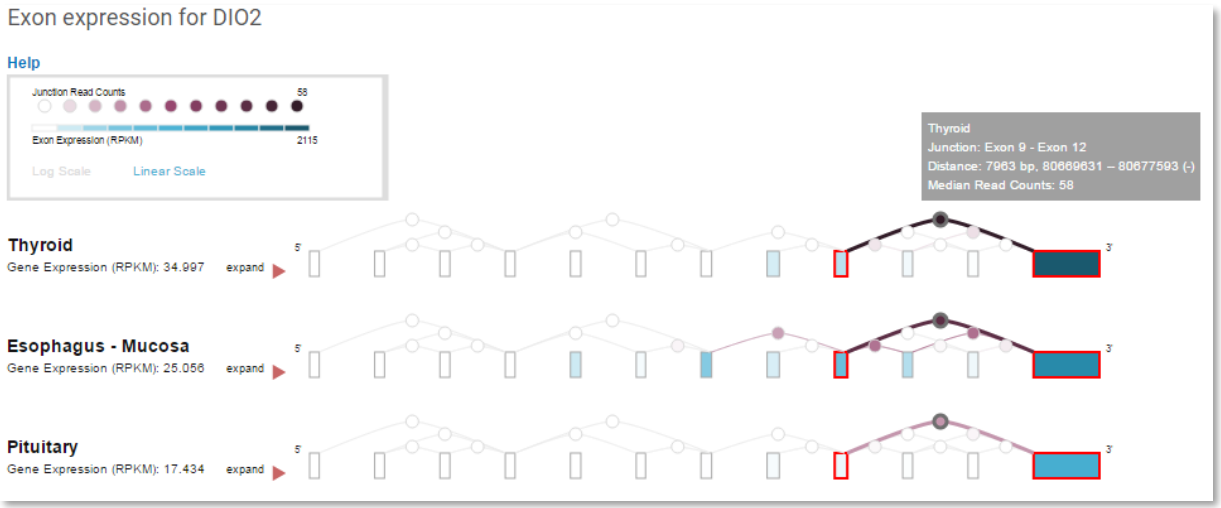
Expression – Alternative splicing

Ensembl gives an overview of reported splice variants

Show All entries | Show/hide columns (1 hidden) | Filter

Name	Transcript ID	bp	Protein	Biotype	CCDS	UniProt	RefSeq	Flags
DIO2-001	ENST00000557010	6367	273aa	Protein coding	CCDS45146	Q92813	NM_000793 NP_000784	TSL:2 GENCODE basic APPRIS P1
DIO2-002	ENST00000438257	6136	273aa	Protein coding	CCDS45146	Q92813	NM_013989 NP_054644	TSL:1 GENCODE basic APPRIS P1
DIO2-003	ENST00000555750	1049	309aa	Protein coding	CCDS55934	A0A0A0MTQ2	NM_001007023 NP_001007024	TSL:1 GENCODE basic
DIO2-201	ENST00000422005	6272	145aa	Protein coding	-	J3KQY5	NM_001242502 NM_001242503 NP_001229431 NP_001229432	TSL:5 GENCODE basic
DIO2-004	ENST00000556811	984	58aa	Protein coding	-	H0YJQ8	-	CDS 5' incomplete TSL:1
DIO2-006	ENST00000554188	569	55aa	Protein coding	-	G3V3A8	-	CDS 3' incomplete TSL:4
DIO2-010	ENST00000557125	517	56aa	Protein coding	-	G3V2A7	-	TSL:3 GENCODE basic
DIO2-007	ENST00000553594	491	6aa	Protein coding	-	-	-	CDS 3' incomplete TSL:4
DIO2-009	ENST00000553968	417	6aa	Protein coding	-	-	-	CDS 3' incomplete TSL:3
DIO2-005	ENST00000555844	776	37aa	Nonsense mediated decay	-	H0YJ42	-	CDS 5' incomplete TSL:1
DIO2-008	ENST00000556384	267	No protein	Processed transcript	-	-	-	TSL:3

GTEx shows tissue specific splice information



Known genetic variation - HaploREG

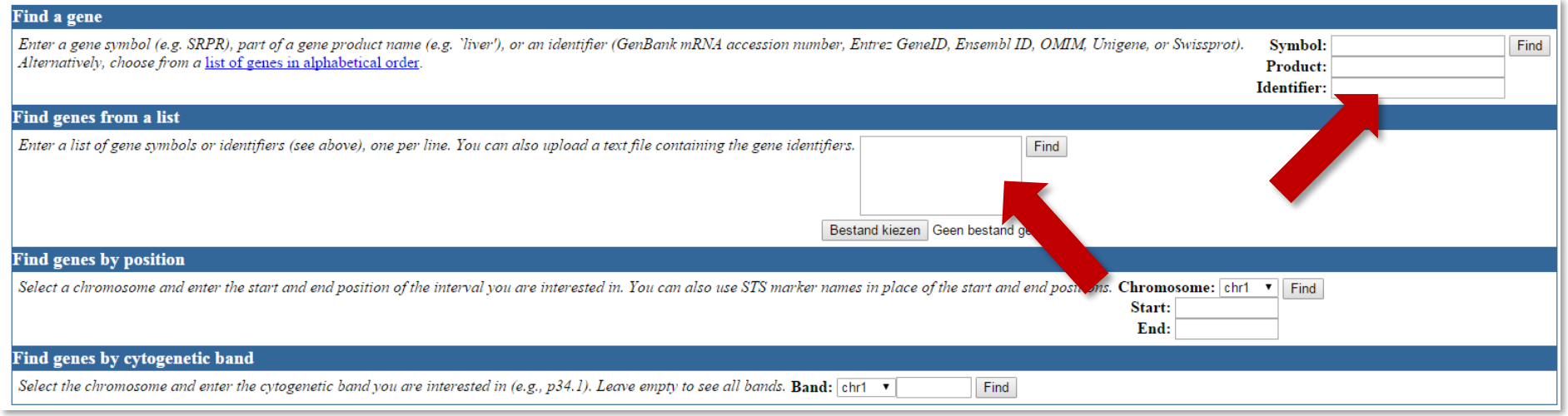
- SNPs in LD
- Alleles
- Population frequencies
- Protein binding
- DNAse sensitivity
- …

Query SNP: rs225014 and variants with r^2 >= 0.8

pos (hg19)	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	eQTL tissues	Motifs changed	Drivers disrupted	GENCODE genes	dbSNP func annot
chr14:80634365	chr14:80168022	0.84	0.96	rs4903903	T	C	0.22	0.41	0.35	0.32			FAT	MUS,PANC,MUS			STAT		30kb 3' of DIO2	
chr14:80635104	chr14:80168761	0.84	0.96	rs12588985	A	G	0.22	0.41	0.35	0.32							STAT		29kb 3' of DIO2	
chr14:80638440	chr14:80172097	0.84	0.96	rs1491504	G	T	0.22	0.41	0.35	0.32							10 altered motifs		25kb 3' of DIO2	
chr14:80641302	chr14:80174959	0.84	0.96	rs4899763	C	T	0.22	0.41	0.35	0.32							SRF		23kb 3' of DIO2	
chr14:80641672	chr14:80175329	0.84	0.96	rs74064450	T	C	0.22	0.41	0.35	0.32							Arid5b,CHOP::CEBPalpha		22kb 3' of DIO2	
chr14:80642320	chr14:80175977	0.84	0.96	rs2216086	C	T	0.35	0.43	0.35	0.32				PANC			GR		22kb 3' of DIO2	
chr14:80642939	chr14:80176596	0.84	0.96	rs4899764	G	T	0.22	0.41	0.35	0.32							5 altered motifs		21kb 3' of DIO2	
chr14:80643110	chr14:80176767	0.84	0.96	rs4899765	C	T	0.22	0.41	0.35	0.32				SKIN			11 altered motifs		21kb 3' of DIO2	
chr14:80648603	chr14:80182260	0.84	0.96	rs2005885	C	G	0.22	0.42	0.41	0.32							E2A,Pitx2,SETDB1		15kb 3' of DIO2	
chr14:80649239	chr14:80182896	0.83	0.96	rs759441	T	A	0.21	0.42	0.41	0.31							Hand1,Pax-6,Zbtb12		15kb 3' of DIO2	
chr14:80649565	chr14:80183222	0.84	0.96	rs4903904	G	A	0.22	0.42	0.41	0.32							Elf3,STAT		14kb 3' of DIO2	
chr14:80655946	chr14:80189603	0.82	0.95	rs74064456	A	T	0.23	0.42	0.44	0.32				ESDR					7.9kb 3' of DIO2	
chr14:80658261	chr14:80191918	0.86	0.96	rs56017760	T	C	0.23	0.43	0.44	0.32							COMP1,CTCF		5.6kb 3' of DIO2	
chr14:80660670	chr14:80194327	0.84	0.96	rs74064457	G	A	0.24	0.42	0.44	0.32							Sox		3.2kb 3' of DIO2	
chr14:80661203	chr14:80194860	0.86	0.95	rs35191251	A	C	0.23	0.43	0.44	0.33									2.7kb 3' of DIO2	
chr14:80662335	chr14:80195992	0.85	0.99	rs56025506	C	T	0.23	0.41	0.43	0.30							BAF155,Nkx3,Pou5f1		1.5kb 3' of DIO2	
chr14:80667579	chr14:80201236	0.87	0.99	rs225015	G	A	0.48	0.45	0.43	0.31							BCL,STAT		DIO2	3'-UTR
chr14:80669580	chr14:80203237	1	1	rs225014	T	C	0.50	0.46	0.44	0.34				4 organs	6 bound proteins		RXRA		DIO2	missense
chr14:80673242	chr14:80206899	0.94	1	rs12437279	C	T	0.23	0.42	0.43	0.32			4 organs	ESDR,BRN,BRN					DIO2	intronic

Known genetic variation - SNPper

- No direct information mRNA or protein consequences
- Focussed on regional variation
- Submit lists of SNPs



Subsequent steps

- Check for redundancy (LD):
  - Haploview with HapMap data
  - Design assay
    - If unsuccessful, use alternative SNPs (based on LD)
  - Measure SNPs

- Determine functionality: **wet lab!**

Questions

QUESTIONS?

Questions & Excercise

- Databases; hands on experience
- Select your own dataset (Data_x) and follow manual
- Share tips & tricks
- Sometimes the web based tools are a bit messy

- Ask questions!

https://www.labroots.com/trending/cell-and-molecular-biology/5499/impact-neanderthal-dna-human-gene-expression