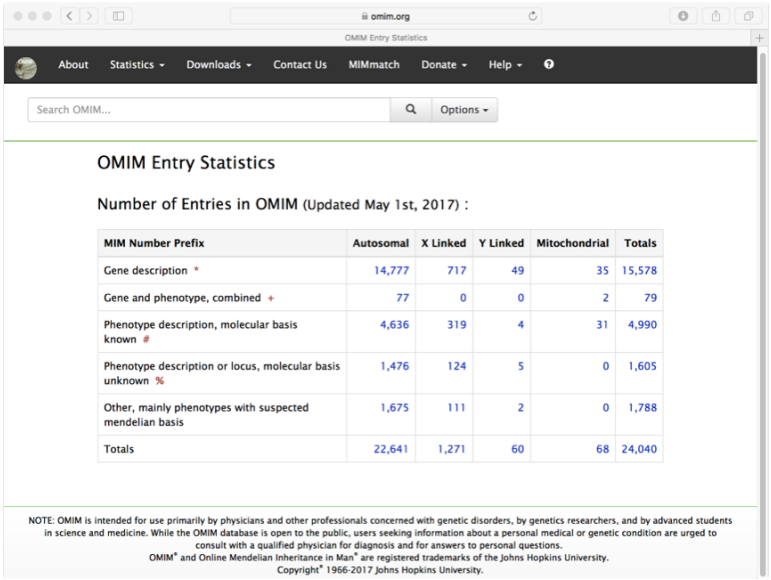
Bioinformatics and Functional Genomics wrap up

Biol4230 Tues, May 1, 2018
Bill Pearson wrp@virginia.edu 4-2818 Pinn 6-057

Things not covered:

- Variation and disease databases
  - OMIM, DBsnp, ClinVar, TCGA
  - Brookes, A. J. & Robinson, P. N. Human genotype-phenotype databases: aims, challenges and opportunities. *Nat Rev Genet* **16,** 702–715 (2015).
- mapping sequencing reads
  - SAM/BAM files to genomes
  - to exomes / transcripts
- peak finding (for ChIP seq, epigenetic marks)
- machine learning strategies
  - neural nets, SVMs, PCA
- Higher-order structure in chromatin
- Gene regulatory networks

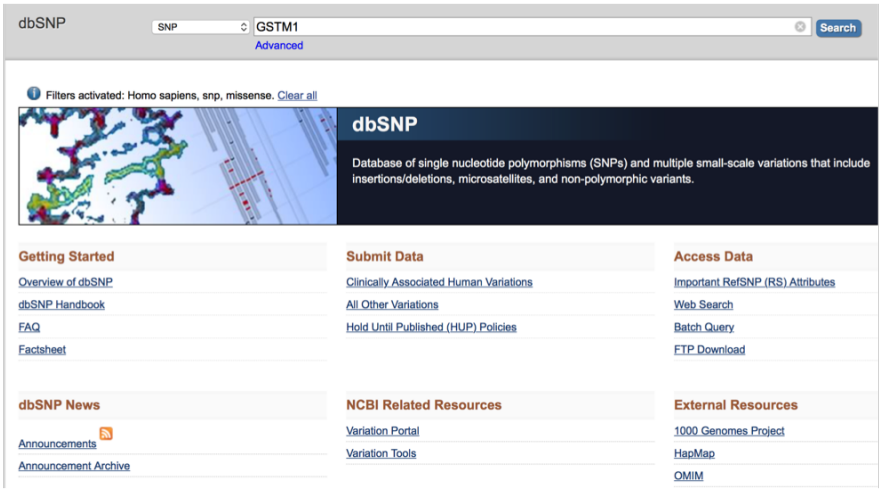
OMIM – Online Mendelian Inheritance in Man



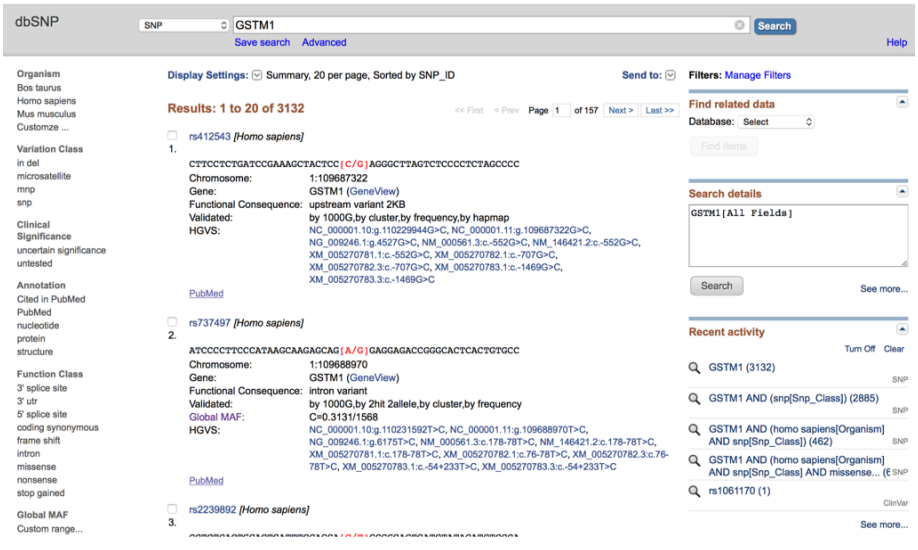
Variation/Clinical variant databases – NCBI Gene

NCBI Gene/refSNP/dbSNP

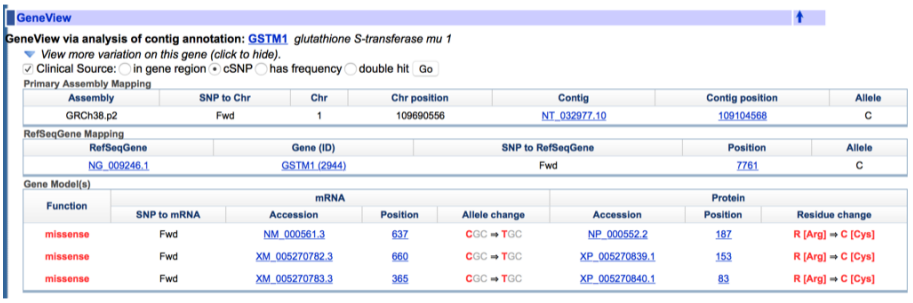
NCBI Gene/refSNP/dbSNP



NCBI Gene/refSNP/dbSNP



NCBI Gene/refSNP/dbSNP



GeneView

GeneView via analysis of contig annotation: GSTM1 *glutathione S-transferase mu 1*

View more variation on this gene (click to hide).

Clinical Source: in gene region cSNP has frequency double hit Go

Primary Assembly Mapping

Assembly	SNP to Chr	Chr	Chr position	Contig	Contig position	Allele
GRCh38.p2	Fwd	1	109690556	NT_032977.10	109104568	C

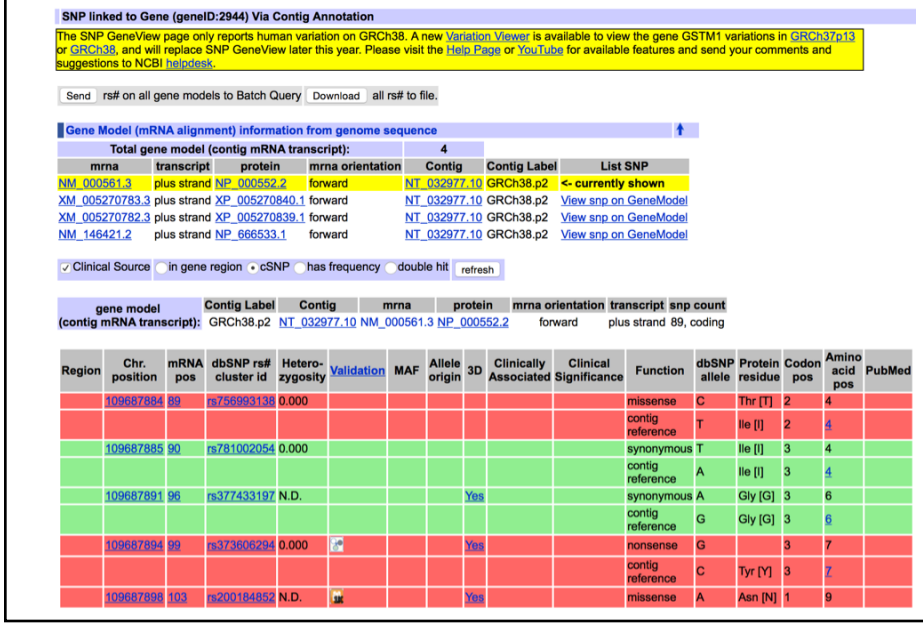
RefSeqGene Mapping

RefSeqGene	Gene (ID)	SNP to RefSeqGene	Position	Allele
NG_009246.1	GSTM1 (2944)	Fwd	7761	C

Gene Model(s)

Function	mRNA				Protein		
	SNP to mRNA	Accession	Position	Allele change	Accession	Position	Residue change
missense	Fwd	NM_000561.3	637	CGC ⇒ TGC	NP_000552.2	187	R [Arg] ⇒ C [Cys]
missense	Fwd	XM_005270782.3	660	CGC ⇒ TGC	XP_005270839.1	153	R [Arg] ⇒ C [Cys]
missense	Fwd	XM_005270783.3	365	CGC ⇒ TGC	XP_005270840.1	83	R [Arg] ⇒ C [Cys]

NCBI Gene/refSNP/dbSNP



SNP linked to Gene (geneID:2944) Via Contig Annotation

The SNP GeneView page only reports human variation on GRCh38. A new Variation Viewer is available to view the gene GSTM1 variations in GRCh37p13 or GRCh38, and will replace SNP GeneView later this year. Please visit the Help Page or YouTube for available features and send your comments and suggestions to NCBI helpdesk.

Send rs# on all gene models to Batch Query Download all rs# to file.

Gene Model (mRNA alignment) information from genome sequence

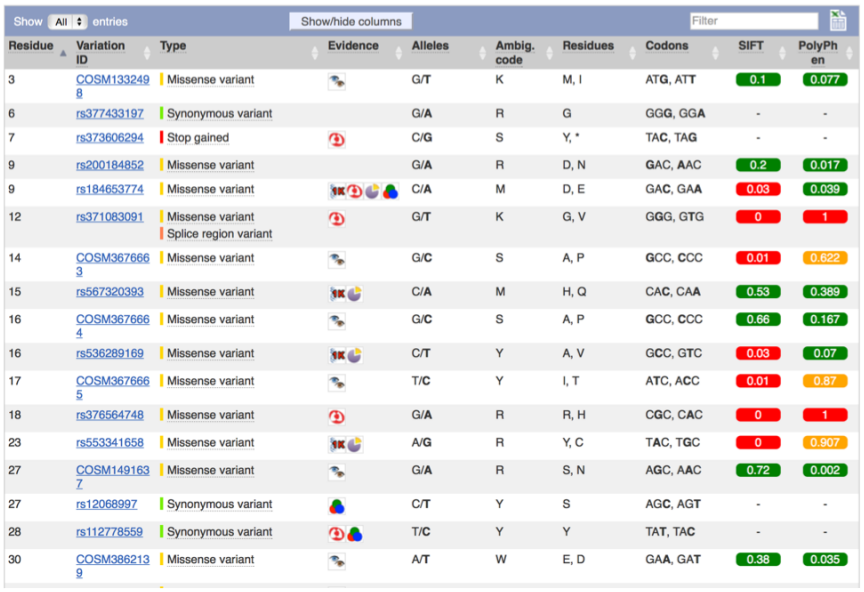
Total gene model (contig mRNA transcript):				4		
mrna	transcript	protein	mrna orientation	Contig	Contig Label	List SNP
NM_000561.3	plus strand	NP_000552.2	forward	NT_032977.10	GRCh38.p2	<- currently shown
XM_005270783.3	plus strand	XP_005270840.1	forward	NT_032977.10	GRCh38.p2	View snp on GeneModel
XM_005270782.3	plus strand	XP_005270839.1	forward	NT_032977.10	GRCh38.p2	View snp on GeneModel
NM_146421.2	plus strand	NP_666533.1	forward	NT_032977.10	GRCh38.p2	View snp on GeneModel

Clinical Source in gene region cSNP has frequency double hit refresh

gene model (contig mRNA transcript):	Contig Label	Contig	mrna	protein	mrna orientation	transcript	snp count
	GRCh38.p2	NT_032977.10	NM_000561.3	NP_000552.2	forward	plus strand	89, coding

Region	Chr. position	mRNA pos	dbSNP rs# cluster id	Hetero-zygosity	Validation	MAF	Allele origin	3D	Clinically Associated	Clinical Significance	Function	dbSNP allele	Protein residue	Codon pos	Amino acid pos	PubMed
	109687884	89	rs756993138	0.000							missense	C	Thr [T]	2	4	
											contig reference	T	Ile [I]	2	4	
	109687885	90	rs781002054	0.000							synonymous	T	Ile [I]	3	4	
											contig reference	A	Ile [I]	3	4	
	109687891	96	rs377433197	N.D.				Yes			synonymous	A	Gly [G]	3	6	
											contig reference	G	Gly [G]	3	6	
	109687894	99	rs373606294	0.000				Yes			nonsense	G		3	7	
											contig reference	C	Tyr [Y]	3	7	
	109687898	103	rs200184852	N.D.				Yes			missense	A	Asn [N]	1	9	

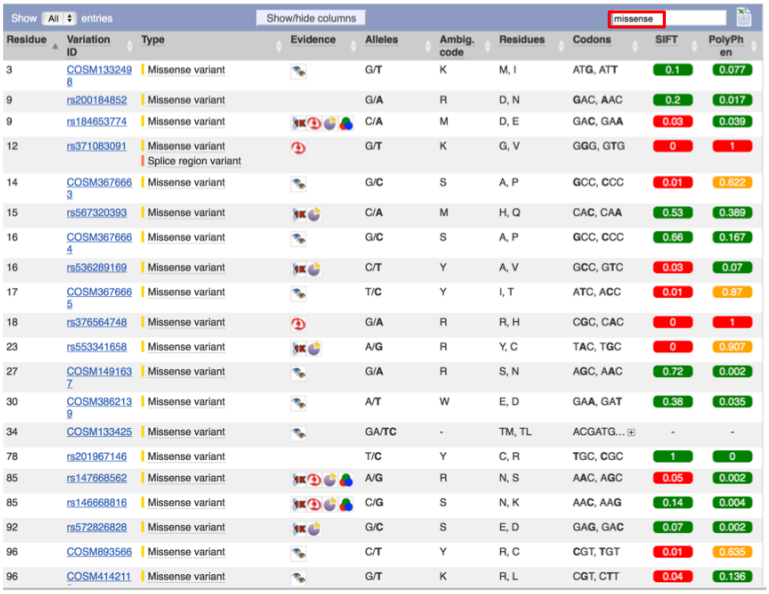
ENSEMBL



Show All entries | Show/hide columns | Filter

Residue	Variation ID	Type	Evidence	Alleles	Ambig. code	Residues	Codons	SIFT	PolyPhen
3	COSM1332498	Missense variant		G/T	K	M, I	ATG, ATT	0.1	0.077
6	rs377433197	Synonymous variant		G/A	R	G	GGG, GGA	-	-
7	rs373606294	Stop gained		C/G	S	Y, *	TAC, TAG	-	-
9	rs200184852	Missense variant		G/A	R	D, N	GAC, AAC	0.2	0.017
9	rs184653774	Missense variant		C/A	M	D, E	GAC, GAA	0.03	0.039
12	rs371083091	Missense variant Splice region variant		G/T	K	G, V	GGG, GTG	0	1
14	COSM3676663	Missense variant		G/C	S	A, P	GCC, CCC	0.01	0.622
15	rs567320393	Missense variant		C/A	M	H, Q	CAC, CAA	0.53	0.389
16	COSM3676664	Missense variant		G/C	S	A, P	GCC, CCC	0.66	0.167
16	rs536289169	Missense variant		C/T	Y	A, V	GCC, GTC	0.03	0.07
17	COSM3676665	Missense variant		T/C	Y	I, T	ATC, ACC	0.01	0.87
18	rs376564748	Missense variant		G/A	R	R, H	CGC, CAC	0	1
23	rs553341658	Missense variant		A/G	R	Y, C	TAC, TGC	0	0.907
27	COSM1491637	Missense variant		G/A	R	S, N	AGC, AAC	0.72	0.002
27	rs12068997	Synonymous variant		C/T	Y	S	AGC, AGT	-	-
28	rs112778559	Synonymous variant		T/C	Y	Y	TAT, TAC	-	-
30	COSM3862139	Missense variant		A/T	W	E, D	GAA, GAT	0.38	0.035

ENSEMBL – filter missense



Show All entries | Show/hide columns | missense

Residue	Variation ID	Type	Evidence	Alleles	Ambig. code	Residues	Codons	SIFT	PolyPhen
3	COSM1332498	Missense variant		G/T	K	M, I	ATG, ATT	0.1	0.077
9	rs200184852	Missense variant		G/A	R	D, N	GAC, AAC	0.2	0.017
9	rs184653774	Missense variant		C/A	M	D, E	GAC, GAA	0.03	0.039
12	rs371083091	Missense variant Splice region variant		G/T	K	G, V	GGG, GTG	0	1
14	COSM3676663	Missense variant		G/C	S	A, P	GCC, CCC	0.01	0.622
15	rs567320393	Missense variant		C/A	M	H, Q	CAC, CAA	0.53	0.389
16	COSM3676664	Missense variant		G/C	S	A, P	GCC, CCC	0.66	0.167
16	rs536289169	Missense variant		C/T	Y	A, V	GCC, GTC	0.03	0.07
17	COSM3676665	Missense variant		T/C	Y	I, T	ATC, ACC	0.01	0.87
18	rs376564748	Missense variant		G/A	R	R, H	CGC, CAC	0	1
23	rs553341658	Missense variant		A/G	R	Y, C	TAC, TGC	0	0.907
27	COSM1491637	Missense variant		G/A	R	S, N	AGC, AAC	0.72	0.002
30	COSM3862139	Missense variant		A/T	W	E, D	GAA, GAT	0.38	0.035
34	COSM133425	Missense variant		GA/TC	-	TM, TL	ACGATG...	-	-
78	rs201967146	Missense variant		T/C	Y	C, R	TGC, CGC	1	0
85	rs147668562	Missense variant		A/G	R	N, S	AAC, AGC	0.05	0.002
85	rs146668816	Missense variant		C/G	S	N, K	AAC, AAG	0.14	0.004
92	rs572826828	Missense variant		G/C	S	E, D	GAG, GAC	0.07	0.002
96	COSM893566	Missense variant		C/T	Y	R, C	CGT, TGT	0.01	0.635
96	COSM414211	Missense variant		G/T	K	R, L	CGT, CTT	0.04	0.136

ENSEMBL – protein variation (missense)

170	COSM1316144	Missense variant		T/C	Y	F, L	TTT, CTT	0.22	0.049
173	COSM3747491	Missense variant		G/C	S	K, N	AAG, AAC	0.07	0.023
173	rs74837985	Missense variant		G/C	S	K, N	AAG, AAC	0.07	0.023
179	rs72549312	Missense variant		C/T	Y	P, L	CCA, CTA	0.04	0.174
180	rs369344514	Missense variant		A/G	R	N, D	AAT, GAT	0	0.98
184	COSM3984066	Missense variant		T/G	K	F, V	TTC, GTC	0	0.925
187	rs72549313	Missense variant		C/T	Y	R, C	CGC, TGC	0.05	0.74
194	rs199721250	Missense variant		T/C	Y	I, T	ATC, ACC	0.01	0.656
202	rs371247780	Missense variant		G/A	R	R, H	CGC, CAC	0.08	0.007
210	rs449856	Missense variant		T/A	W	S, T	TCA, ACA	1	0.001
213	rs533860247	Missense variant		G/A	R	A, T	GCT, ACT	0	0.97

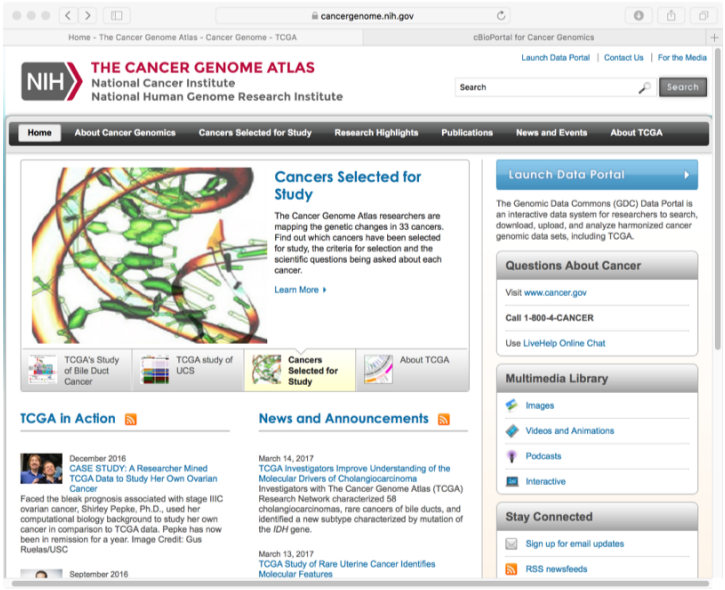
TABLE II

Specific activities of wild-type and mutant human Mu class GSTs with alternative electrophilic substrates

Electrophile	GSH	Specific activity					
		GST M2-2 wild type	GST M2-2 T210S	GST M2-2 T210S/F104T	GST M2-2 T210S/F104T/A130E	GST M1-1 wild type	GST M1-1 S210T
	mM	μmol min^{-1} mg^{-1}					
Epoxide substrates							
*t*SO (0.15 mM)	4.0	0.00020 ± 0.00003	0.17 ± 0.03	0.19 ± 2	0.28 ± 1	3.00 ± 0.02	0.026 ± 0.001
SO (1.6 mM)	5.0	0.037 ± 0.001	1.28 ± 0.06	1.24 ± 0.08	1.23 ± 0.04	2.7 ± 0.08	0.10 ± 0.01
NPG (1.0 mM)	2.0	0.12 ± 0.01	3.5 ± 0.1	2.4 ± 0.1	2.2 ± 0.1	4.5 ± 0.2	0.05 ± 0.006
Other substrates							
Aminochrome (0.3 mM)	1.0	120 ± 7	108 ± 6	82 ± 7	132 ± 8	0.73 ± 0.02	0.94 ± 0.05
CyanoDMNG (1.0 mM)	1.0	208 ± 4	116 ± 2	181 ± 4	135 ± 3	0.47 ± 0.01	0.36 ± 0.02
CDNB (1.0 mM)	1.0	426 ± 5	482 ± 14	547 ± 12	600 ± 16	136 ± 6	112 ± 3

Ivarsson, Y. et al. (2003) *J Biol Chem* **278,** 8733

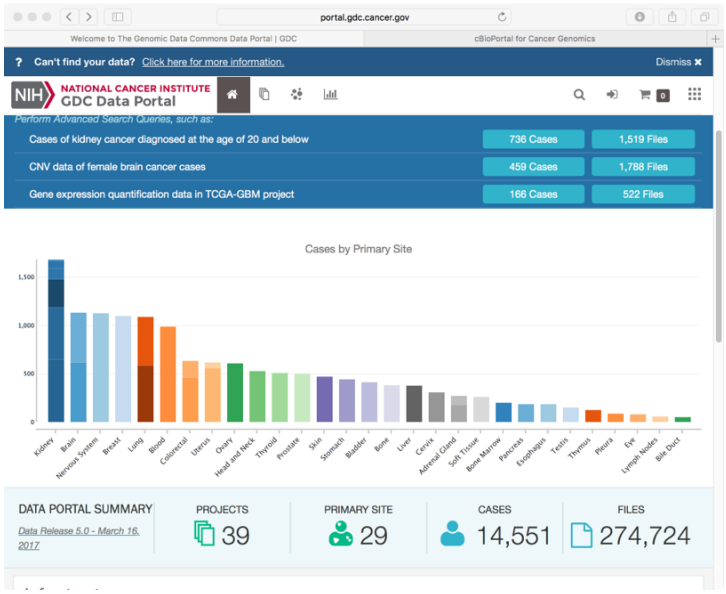
TCGA – The Cancer Genome Atlas



fasta.bioch.virginia.edu/biol4230

12

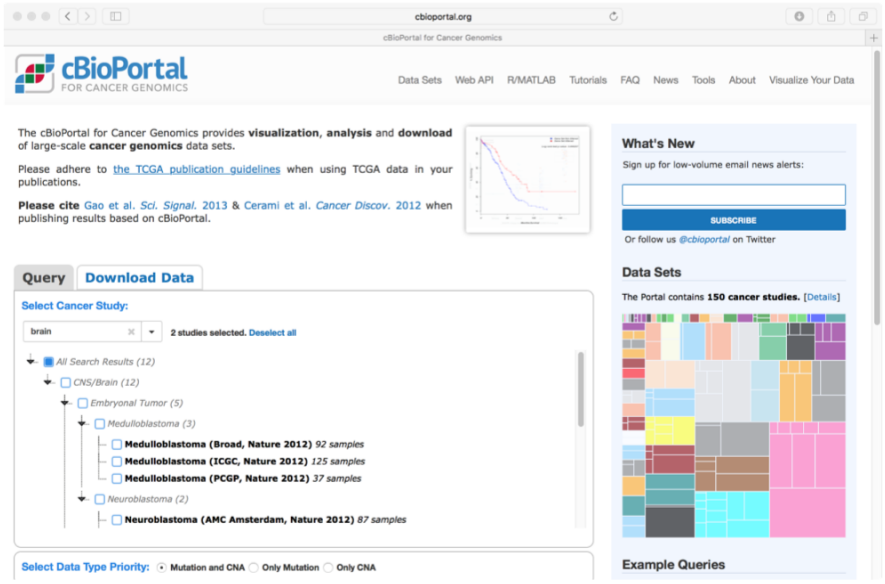
TCGA – The Cancer Genome Atlas



fasta.bioch.virginia.edu/biol4230

13

cBIO Portal - cancer Biology



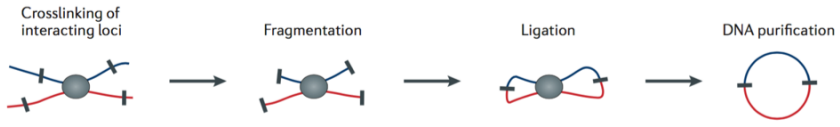
fasta.bioch.virginia.edu/biol4230

14

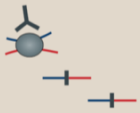
Higher order chromatin configuration: Chromatin Conformation Capture (3C)

Box 1 | **3C-based methods** Dekker et al (2013) *Nat Rev Genet* **14,** 390–403

a 3C: converting chromatin interactions into ligation products



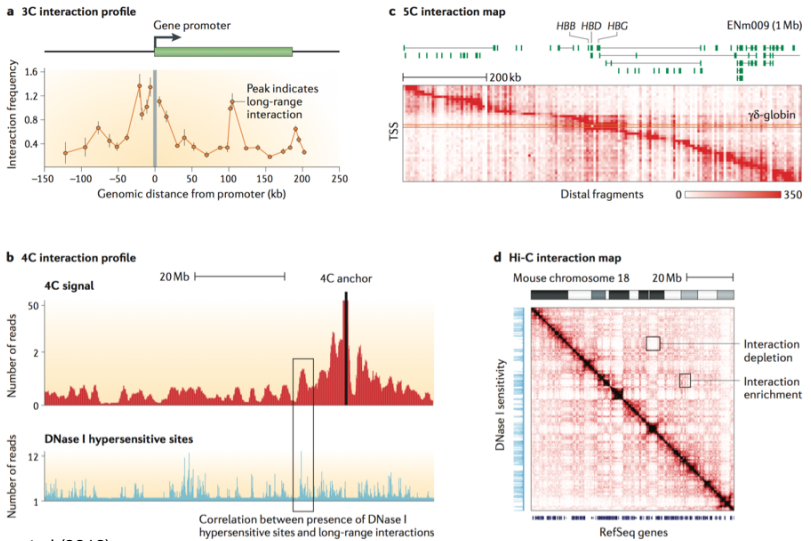
b Ligation product detection methods

3C	4C	5C	ChIA–PET	Hi-C
One-by-one All-by-all	One-by-all	Many-by-many	Many-by-many	All-by-all
			• DNA shearing • Immunoprecipitation 	• Biotin labelling of ends • DNA shearing 
PCR or sequencing	Inverse PCR sequencing	Multiplexed LMA sequencing	Sequencing	Sequencing

fasta.bioch.virginia.edu/biol4230

15

Higher order chromatin configuration: Chromatin Conformation Capture (3C)

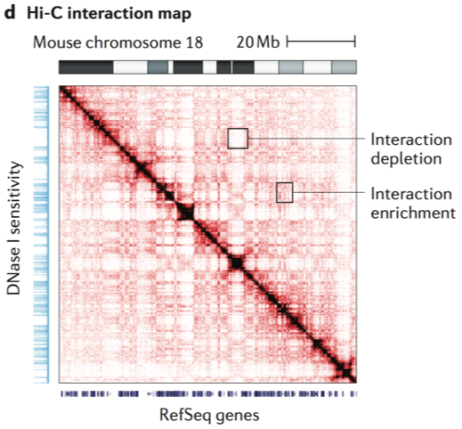


Dekker et al (2013)
Nat Rev Genet **14,** 390–403

fasta.bioch.virginia.edu/biol4230

16

Heat Maps and Clustering



Dekker et al (2013)
Nat Rev Genet **14,** 390–403

Biol4230 – what we did cover

- The human genome project
  - building complete "genomes" from pieces
- RNA expression analysis
  - RNA abundance vs protein abundance
  - the RNA abundance problem – many orders of magnitude between lowest an highest
  - looking for differential expression – how to normalize?
  - correcting for multiple tests (FDR)
  - looking for sets of co-regulated genes:
    - over-representation analysis (GO terms)

Biol4230 – what we did cover

- Identifying functional sites
  - not homologous
  - short, not well conserved
  - not significant (in the entire genome context)
  - represent with PWM (position weight matrix, PSSM)
  - estimation with missing data (alignment/PWM)
  - predicting binding from protein structure

Bioinformatics – the big picture

- Lots and lots and lots of data
  - is it "clean" enough?
    - do discrepancies in the data reflect biology, or technology
  - what inferences/conclusions are reliable?
    - E()< 10^{-6} implies homology
  - what assumptions have been made?
    - multiple sequence alignment requires homology
    - GO experimental terms are "better" than BLAST results
    - the database is complete
    - the protein predictions are accurate

Bioinformatics – the big picture

- Why could this result be wrong?
- Does it make sense
  - can 100% identical sequences have different functions?
- What is the control?
  - what kinds of errors does the control detect?
  - what kinds of errors does it miss?