

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

fasta.bioch.virginia.edu/biol4230

1

OMIM – Online Mendelian Inheritance in Man

The screenshot shows the OMIM Entry Statistics page. At the top, there's a navigation bar with links for About, Statistics, Downloads, Contact Us, MIMmatch, Donate, and Help. Below the navigation is a search bar labeled "Search OMIM..." and an "Options" button. The main content area is titled "OMIM Entry Statistics" and contains a table titled "Number of Entries in OMIM (Updated May 1st, 2017)". The table provides a breakdown of entries by MIM Number Prefix and category. The data is as follows:

MIM Number Prefix	Autosomal	X Linked	Y Linked	Mitochondrial	Totals
Gene description *	14,777	717	49	35	15,578
Gene and phenotype, combined +	77	0	0	2	79
Phenotype description, molecular basis known #	4,636	319	4	31	4,990
Phenotype description or locus, molecular basis unknown %	1,476	124	5	0	1,605
Other, mainly phenotypes with suspected mendelian basis	1,675	111	2	0	1,788
Totals	22,641	1,271	60	68	24,040

At the bottom of the page, there's a note about the intended use of OMIM and a copyright notice.

fasta.bioch.virginia.edu/biol4230

2

Variation/Clinical variant databases – NCBI Gene

The screenshot shows the NCBI Gene page for the gene *glutathione S-transferase mu 1* in *Homo sapiens*. The page includes a search bar, navigation links for 'Gene' and 'Advanced', and sections for 'Settings' and 'Send to'. The main content area displays detailed information about the gene, including its official symbol (*GSTM1*), full name (*glutathione S-transferase mu 1*), primary source (HGNC), and various aliases like *GSTM1*, *GSTM1-1*, *MU-1*, and *GSTM1a-1a*. It also provides details on its RefSeq status (REVIEWED), organism (*Homo sapiens*), lineage (Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo), and other known forms (H-B, GST1, GT14, GTM1). A summary section discusses the gene's function in detoxifying electrophilic compounds and its eight distinct classes. The right sidebar contains a 'Table of contents' with links to various genomic and protein-related databases.

NCBI Gene/refSNP/dbSNP

The screenshot shows the 'Variation' section of the NCBI Gene page. It includes a header with a 'Variation' link and a question mark icon. Below the header, there are several links: 'See variants in ClinVar', 'See studies and variants in dbVar', 'See Variation Viewer (GRCh37.p13)', 'See Variation Viewer (GRCh38)', and 'Genotypes'. Under 'Genotypes', there are links to 'See SNP Geneview Report' and 'See 1000 Genomes Browser (GRCh37.p13)'.

NCBI Gene/refSNP/dbSNP

The screenshot shows the NCBI dbSNP homepage. At the top, there is a search bar with 'SNP' selected and 'GSTM1' entered. Below the search bar is a message: 'Filters activated: Homo sapiens, SNP, missense. Clear all'. To the right of this message is a large graphic showing a map of SNP density across a chromosome, with various colored regions representing different SNP concentrations. To the right of the map, the text reads: 'dbSNP Database of single nucleotide polymorphisms (SNPs) and multiple small-scale variations that include insertions/deletions, microsatellites, and non-polymorphic variants.' Below the main search area, there are three columns of links:

- Getting Started:** Overview of dbSNP, dbSNP Handbook, FAQ, Factsheet.
- Submit Data:** Clinically Associated Human Variations, All Other Variations, Hold Until Published (HUP) Policies.
- Access Data:** Important RefSNP (RS) Attributes, Web Search, Batch Query, FTP Download.

Below these columns are three more sections:

- dbSNP News:** Announcements, Announcement Archive.
- NCBI Related Resources:** Variation Portal, Variation Tools.
- External Resources:** 1000 Genomes Project, HapMap, OMIM.

NCBI Gene/refSNP/dbSNP

The screenshot shows the NCBI dbSNP results page for the query 'GSTM1'. The results section displays three SNPs:

- rs412543 [Homo sapiens]**
CTTCTCTGATCCGAAAGCTACTCC [C/G] AGGGCTTAGTCCTCCCTCTAGCCCC
Chromosome: 1:109687322
Gene: GSTM1 (GeneView)
Functional Consequence: upstream variant 2KB
Validated: by 1000G, by cluster, by frequency by hapmap
HGVS: NC_000001.10:g.1102315892>C, NC_000001.11:g.109687322G>C, NC_009246.1:g.45370>C, NM_005961.3:c.-552G>C, NM_146421.2:c.-552G>C, XM_005270781.1:c.-552G>C, XM_005270782.1:c.-707C>C, XM_005270782.3:c.-707C>C, XM_005270783.1:c.-1469G>C, XM_005270783.3:c.-1469G>C
- rs737497 [Homo sapiens]**
ATCCCCTCCCATAGCCAAGACCGAG[A/G]GAGGAGACCCGGCACTACTGTGCC
Chromosome: 1:109688970
Gene: GSTM1 (GeneView)
Functional Consequence: intron variant
Validated: by 1000G, by 2hit allele, by cluster, by frequency
Global MAF: C=0.3131/1568
HGVS: NC_000001.10:g.1102315892>C, NC_000001.11:g.109688970>C, NC_009246.1:g.61757>C, NM_005961.3:c.178-78T>C, NM_146421.2:c.178-78T>C, XM_005270781.1:c.178-78T>C, XM_005270782.1:c.76-78T>C, XM_005270782.3:c.76-78T>C, XM_005270783.1:c.-54>C, XM_005270783.3:c.-54>C
- rs2239892 [Homo sapiens]**

On the right side of the results page, there are several filters and search fields:

- Filters: Manage Filters**
- Find related data**
- Database: Select**
- Search details**: GSTM1 [All Fields]
- Recent activity**:
 - GSTM1 (3132)
 - GSTM1 AND (snp[SnP_Class]) (2885)
 - GSTM1 AND (homo sapiens[Organism] AND snp[SnP_Class]) (462)
 - GSTM1 AND (homo sapiens[Organism] AND snp[SnP_Class] AND missense... (6 SNP)
 - rs10611170 (1)

NCBI Gene/refSNP/dbSNP

GeneView

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

[View more variation on this gene \(click to hide\)](#).

Clinical Source: in gene region cSNP has frequency double hit [Go](#)

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	SNP to mRNA	mRNA	Position	Allele change	Protein		
					Accession	Position	Residue change
missense	Fwd	NM_000561.3	637	C[G] → T[G]	NP_000562.2	187	R [Arg] → C [Cys]
missense	Fwd	XM_005270782.3	660	C[G] → T[G]	XP_005270839.1	153	R [Arg] → C [Cys]
missense	Fwd	XM_005270783.3	365	C[G] → T[G]	XP_005270840.1	83	R [Arg] → C [Cys]

NCBI Gene/refSNP/dbSNP

SNP linked to Gene (genelD: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_000562.2	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_000562.2			forward		plus strand	89, coding	

Region	Chr. position	mRNA pos	dbSNP rs# cluster id	Heterozygosity	Validation	MAF	Allele origin	3D	Clinically Associated	Clinical Significance	Function	dbSNP allele	Protein residue	Codon pos	Amino acid pos	PubMed
109687884	89	rs75699313B	0.000								missense	C	Thr [T]	2	4	
											contig reference	T	Ile [I]	2	4	
109687885	90	rs78100205A	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

Residue	Variation ID	Type	Evidence	Alleles	Ambig. code	Residues	Codons	SIFT	PolyPhen
3	COSM133249	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	☒	G/T	K	G, V	GGG, GTG	0	1
		Splice region variant							
14	COSM367666	Missense variant	☒	G/C	S	A, P	GCC, CCC	0.01	0.622
3	rs567320393	Missense variant	☒☒☒	C/A	M	H, Q	CAC, CAA	0.53	0.389
16	COSM367666	Missense variant	☒	G/C	S	A, P	GCC, CCC	0.66	0.167
4	rs536289169	Missense variant	☒☒☒	C/T	Y	A, V	GCC, GTC	0.03	0.07
5	COSM367666	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	COSM149163	Missense variant	☒	G/A	R	S, N	AGC, AAC	0.72	0.002
7	rs12068997	Synonymous variant	☒	C/T	Y	S	AGC, AGT	-	-
28	rs112778559	Synonymous variant	☒☒	T/C	Y	Y	TAT, TAC	-	-
30	COSM386213	Missense variant	☒	A/T	W	E, D	GAA, GAT	0.38	0.035
9									

ENSEMBL – filter missense

Residue	Variation ID	Type	Evidence	Alleles	Ambig. code	Residues	Codons	SIFT	PolyPhen
3	COSM133249	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	☒	G/T	K	G, V	GGG, GTG	0	1
		Splice region variant							
14	COSM367666	Missense variant	☒	G/C	S	A, P	GCC, CCC	0.01	0.622
3	rs567320393	Missense variant	☒☒☒	C/A	M	H, Q	CAC, CAA	0.53	0.389
16	COSM367666	Missense variant	☒	G/C	S	A, P	GCC, CCC	0.66	0.167
4	rs536289169	Missense variant	☒☒☒	C/T	Y	A, V	GCC, GTC	0.03	0.07
5	COSM367666	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	COSM149163	Missense variant	☒	G/A	R	S, N	AGC, AAC	0.72	0.002
7	rs3086213	Missense variant	☒	A/T	W	E, D	GAA, GAT	0.38	0.035
34	COSM133425	Missense variant	☒	G/A/T/C	-	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	rs57286828	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	COSM131614	Missense variant 4		T/C	Y	F, L	TTT, CTT	0.22	0.049
173	COSM374749	Missense variant 1		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	COSM398406	Missense variant 6		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	0.856
202	rs371247780	Missense variant		G/A	R	R, H	CGC, CAC	0.08	0.007
210	rs4498656	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					
		mm		μmol min⁻¹ mg⁻¹			
<i>Epoxide substrates</i>							
tSO (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
<i>Other substrates</i>							
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

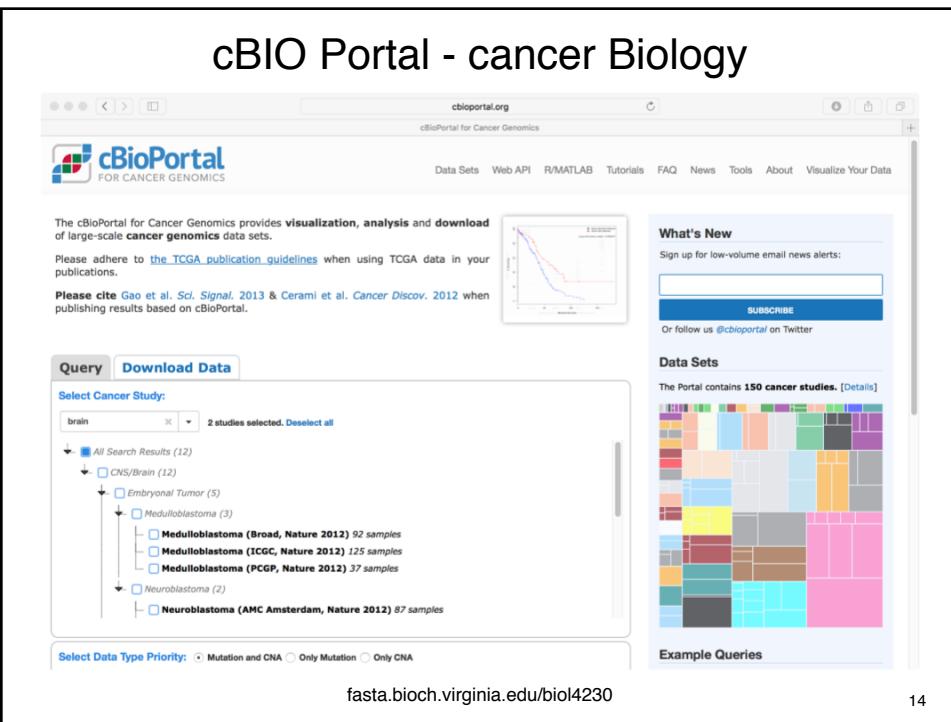
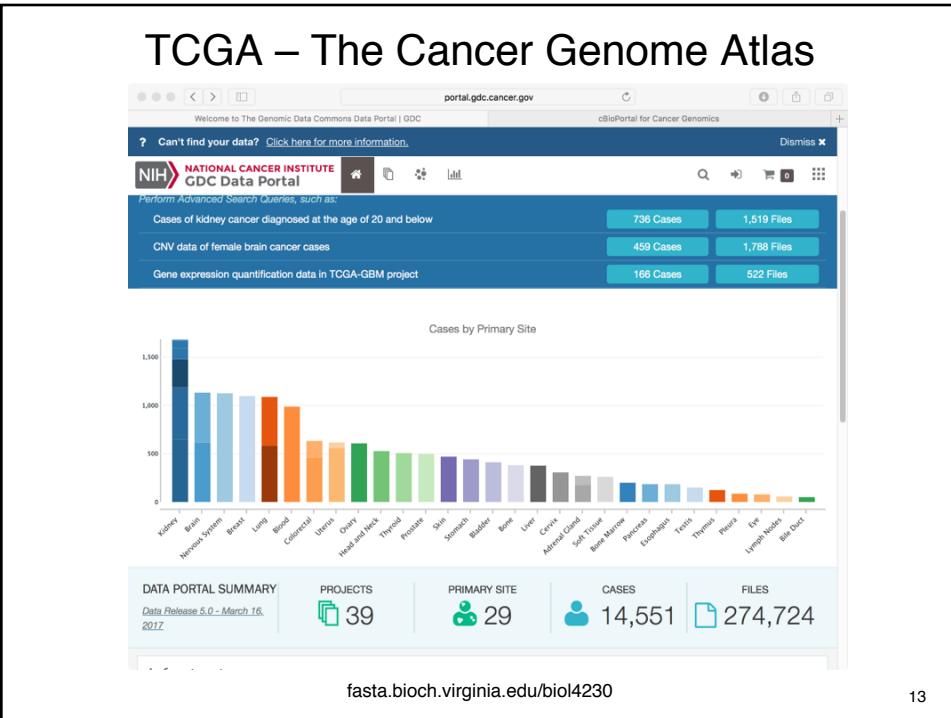
Ivarsson, Y. et al. (2003) *J Biol Chem* **278**, 8733

TCGA – The Cancer Genome Atlas

The screenshot shows the main homepage of the TCGA website. At the top, there's a navigation bar with links for Home, About Cancer Genomics, Cancers Selected for Study, Research Highlights, Publications, News and Events, and About TCGA. Below the navigation, there's a large banner for "Cancers Selected for Study" featuring a colorful 3D visualization of cancer cells. To the right of the banner, there's a "Launch Data Portal" button and a section for "Questions About Cancer" with links to visit cancer.gov, call 1-800-4-CANCER, and use LiveHelp Online Chat. On the left, there are links for "TCGA In Action" and "News and Announcements". The "News and Announcements" section includes a recent case study about a researcher named Shirley Pepke studying ovarian cancer. The footer contains links for "Images", "Videos and Animations", "Podcasts", and "Interactive". There's also a "Stay Connected" section with links for "Sign up for email updates" and "RSS newsfeeds".

fasta.bioch.virginia.edu/biol4230

12

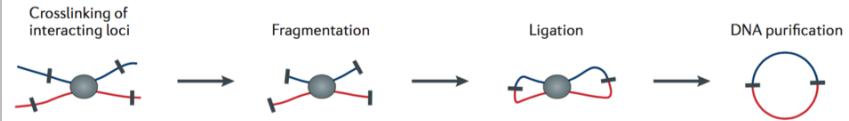


Higher order chromatin configuration: Chromatin Conformation Capture (3C)

Box 1 | 3C-based methods

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

a 3C: converting chromatin interactions into ligation products



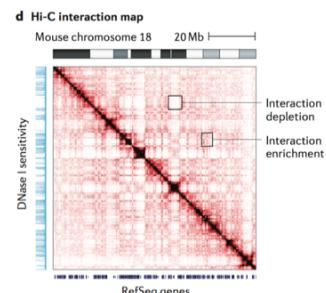
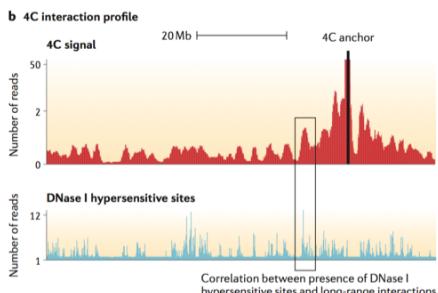
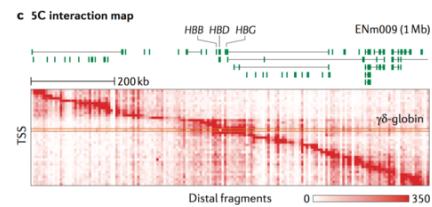
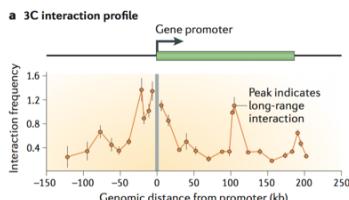
b Ligation product detection methods

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
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)

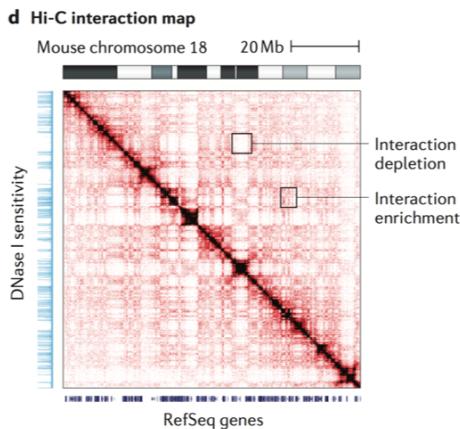


Dekker et al (2013)
Nat Rev Genet **14**, 390–403

fasta.bioch.virginia.edu/biol4230

16

Heat Maps and Clustering



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

fasta.bioch.virginia.edu/biol4230

17

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 and 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)

fasta.bioch.virginia.edu/Biol4230

18

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

fasta.bioch.virginia.edu/Biol4230

19

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

fasta.bioch.virginia.edu/Biol4230

20

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?