Data Visualization & Annotation

8th International Biocuration Conference

24 April 2015 | Beijing, China

Rama Balakrishnan

Saccharomyces Genome Database

Gene Ontology Consortium

Stanford University, CA, USA

Monica Munoz-Torres

Berkeley Bioinformatics Open-Source Projects

Lawrence Berkeley National Lab, CA, USA

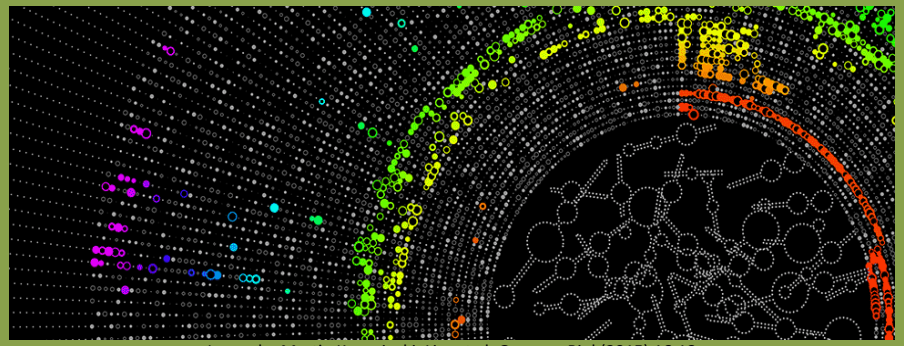
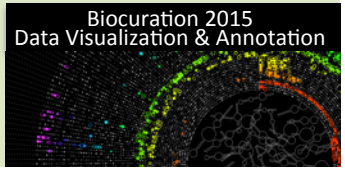


Image by Martin Krzywinski. Lim et al. Genome Biol (2015) 16:18

ISB

Outline

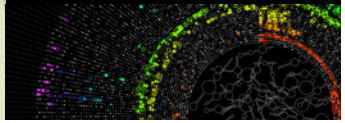
1. Introduction
   - Goals
   - Examples of genome visualization tools
2. Panelists
   - **Lorna Richardson**:
     eMouseAtlas and Image Informatics
   - **Justyna Szostak:**
     Curated Causal Biological Network Models
3. Discussion
   - Featuring you!



Goals of the workshop

1. To learn about tools available for human interpretation of genomic data, specifically in the context of annotation.

2. To open a space for discussion: genomic data are ever more abundant and heterogeneous, with widely varied sources, production techniques, and intrinsic experimental error.
   - How do we analyze these data?
   - What is the best way to interpret the stories the data are telling us?
   - How to put these together (overlay) visually?
   - Developers: what is the best way to disseminate and contribute code to make tool development easier?



Then and Now

Figures	**1**	**Figures**	**49**
Tables	**0**	**Tables**	**27**
References	**6**	**References**	**452**

No. 4356 April 25, 1953 NATURE 737

MOLECULAR STRUCTURE OF NUCLEIC ACIDS

A Structure for Deoxyribose Nucleic Acid

We wish to suggest a structure for the salt of deoxyribose nucleic acid (D.N.A.). This structure has novel features which are of considerable biological interest.

A structure for nucleic acid has already been proposed by Pauling and Corey[1]. They kindly made their manuscript available to us in advance of publication. Their model consists of three inter-twined chains, with the phosphates near the fibre axis, and the bases on the outside. In our opinion, this structure is unsatisfactory for two reasons:

(1) We believe that the material which gives the X-ray diagrams is the salt, not the free acid. Without the acidic hydrogen atoms it is not clear what forces would hold the structure together, especially as the negatively charged phosphates near the axis will repel each other. (2) Some of the van der Waals distances appear to be too small.

Another three-chain structure has also been suggested by Fraser (in the press). In his model the phosphates are on the outside and the bases on the inside, linked together by hydrogen bonds. This structure as described is rather ill-defined, and for this reason we shall not comment on it.

We wish to put forward a radically different structure for the salt of deoxyribose nucleic acid. This structure has two helical chains each coiled round the same axis (see diagram). We have made the usual chemical assumptions, namely, that each chain consists of phosphate diester groups joining β-D-deoxy-ribofuranose residues with 3',5' linkages. The two chains (but not their bases) are related by a dyad perpendicular to the fibre axis. Both chains follow right-handed helices, but owing to the dyad the sequences of the atoms in the two chains run in opposite directions.

Each chain loosely resembles Furberg's[2] model No. 1; that is, the bases are on the inside of the helix and the phosphates on the outside. The configuration of the sugar and the atoms near it is close to Furberg's 'standard configuration', the sugar being roughly perpendicular to the attached base. There is a residue on each chain every 3·4 A. in the z-direction. We have assumed an angle of 36° between adjacent residues in the same chain, so that the structure repeats after 10 residues on each chain, that is, after 34 A. The distance of a phosphorus atom from the fibre axis is 10 A. As the phosphates are on the outside, cations have easy access to them.

The structure is an open one, and its water content is rather high. At lower water contents we would expect the bases to tilt so that the structure could become more compact.

The novel feature of the structure is the manner in which the two chains are held together by the purine and pyrimidine bases. The planes of the bases are perpendicular to the fibre axis. They are joined together in pairs, a single base from one chain being hydrogen-bonded to a single base from the other chain, so that the two lie side by side with identical z-co-ordinates. One of the pair must be a purine and the other a pyrimidine for bonding to occur. The hydrogen bonds are made as follows: purine position 1 to pyrimidine position 1; purine position 6 to pyrimidine position 6.

If it is assumed that the bases only occur in the structure in the most plausible tautomeric forms (that is, with the keto rather than the enol configurations) it is found that only specific pairs of bases can bond together. These pairs are: adenine (purine) with thymine (pyrimidine), and guanine (purine) with cytosine (pyrimidine).

In other words, if an adenine forms one member of a pair, on either chain, then on these assumptions the other member must be thymine; similarly for guanine and cytosine. The sequence of bases on a single chain, does not appear to be restricted in any way. However, if only specific pairs of bases can be formed, it follows that if the sequence of bases on one chain, is given, then the sequence on the other chain is automatically determined.

It has been found experimentally[3,4] that the ratio of the amounts of adenine to thymine, and the ratio of guanine to cytosine, are always very close to unity for deoxyribose nucleic acid.

It is probably impossible to build this structure with a ribose sugar in place of the deoxyribose, as the extra oxygen atom would make too close a van der Waals contact.

The previously published X-ray data[5,6] on deoxyribose nucleic acid are insufficient for a rigorous test of our structure. So far as we can tell, it is roughly compatible with the experimental data, but it must be regarded as unproved until it has been checked against more exact results. Some of these are given in time following, communications. We were not aware of the details of the results presented there when we devised our structure, which rests mainly though not entirely on published experimental data and stereo-chemical arguments.

It has not escaped our notice that the specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic material.

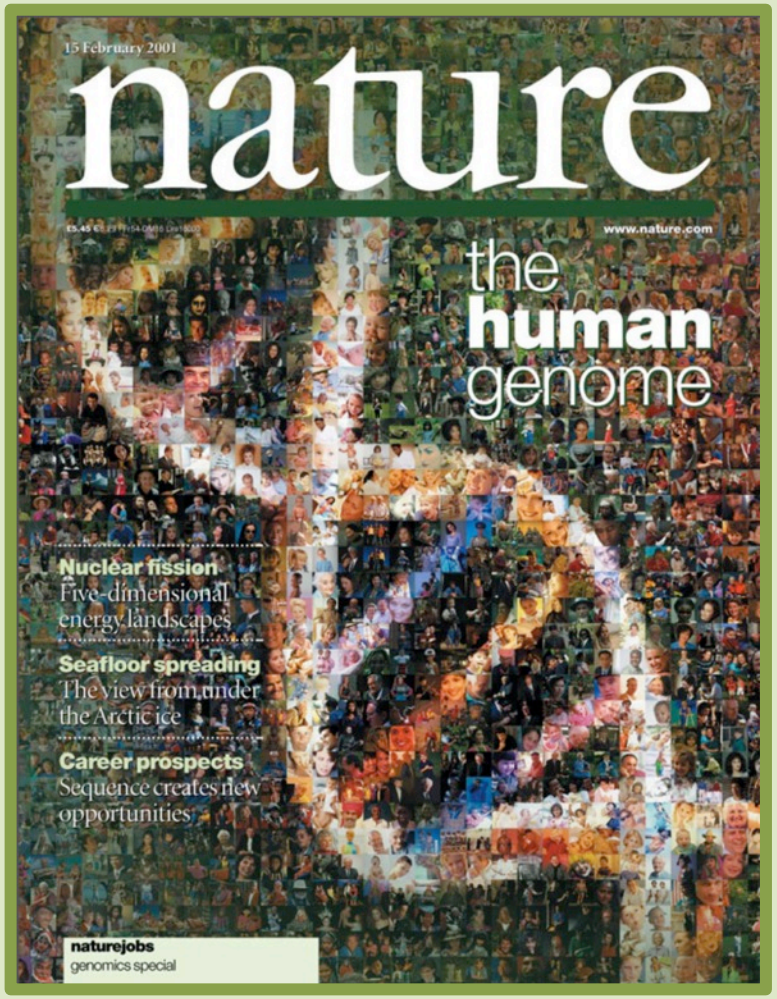
Full details of the structure, including the conditions assumed in building it, together with a set of co-ordinates for the atoms, will be published elsewhere.

We are much indebted to Dr. Jerry Donohue for constant advice and criticism, especially on interatomic distances. We have also been stimulated by a knowledge of the general nature of the unpublished experimental results and ideas of Dr. *M.* H. F. Wilkins, Dr. R. E. Franklin and their co-workers at King's College, London. One of us (J.D.W.) has been aided by a fellowship from the National Foundation for Infantile Paralysis.

J.D. WATSON
F.H. C. CRICK

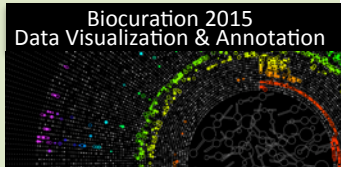
Medical Research Council Unit for the Study of the Molecular Structure of Biological Systems, Cavendish Laboratory, Cambridge. April 2.

[1] Pauling, L., and Corey, R. B. nature, 171, 346 (1953); Proc. U.S. Nat Acad. Sci., 39, 84 (1953).
[2] Furberg, S., Acta Chem. Scand., 6, 634 (1952).
[3] Chargaff, E., for references see Zamenhof, S., Brawerman, G., and Chargaff, E., Biochim. et Biophys. Acta, 9 402 (1952).
[4] Wyatt, G.R. J. Gen. Physiol., 36 201 (1952).
[5] Astbury, W.T., Symp. Soc. Exp. Biol. 1, Nucleic Acid, 66 (Camb. Univ. Press, 1947)
[6] Wilkins, M. H. F. and Randall, J. T. Biochim. et. Biophys. Acta, 10, 192 (1953).

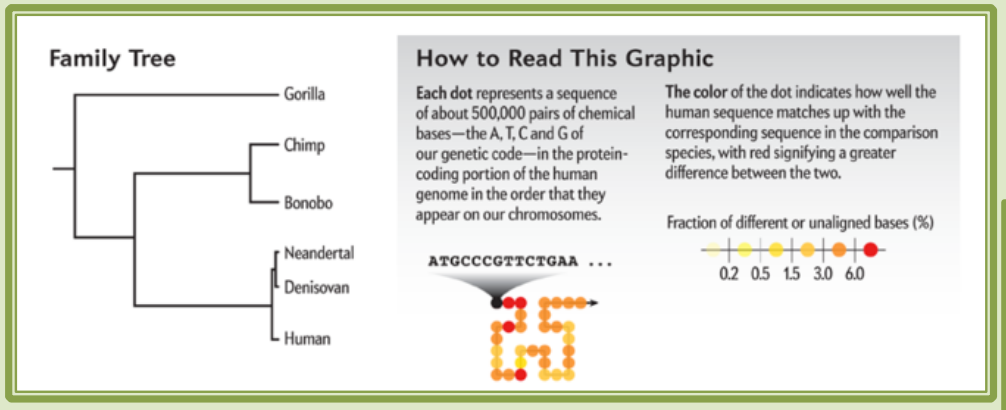


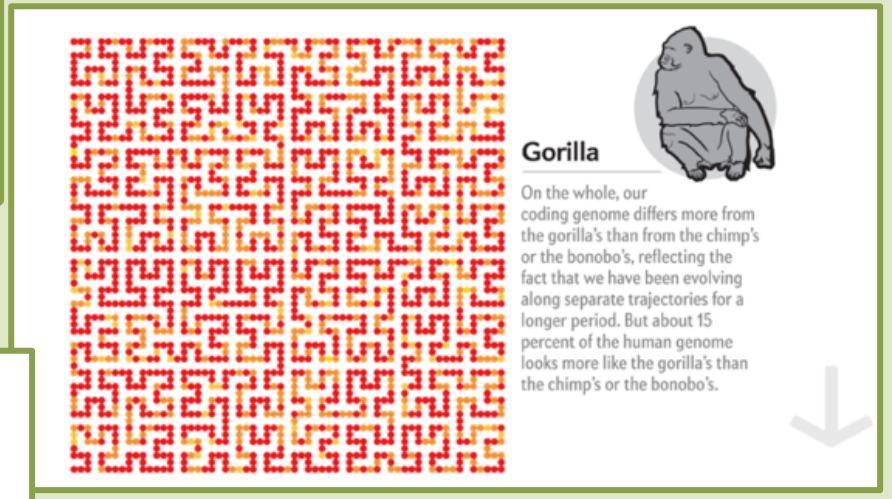
Genomic Data:
Heterogeneous & Abundant

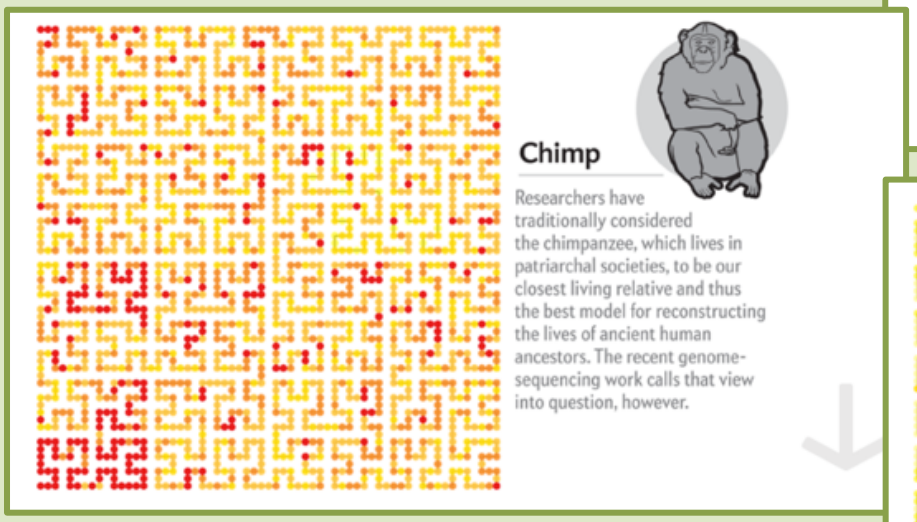
- Structural: gene models, transcriptomes, RNAseq, differential expression, etc.
- Functional: gene ontology, interactions, phenotypes, SNPs, complexes, protein abundance, diseases, images, etc.
- Some examples . . .

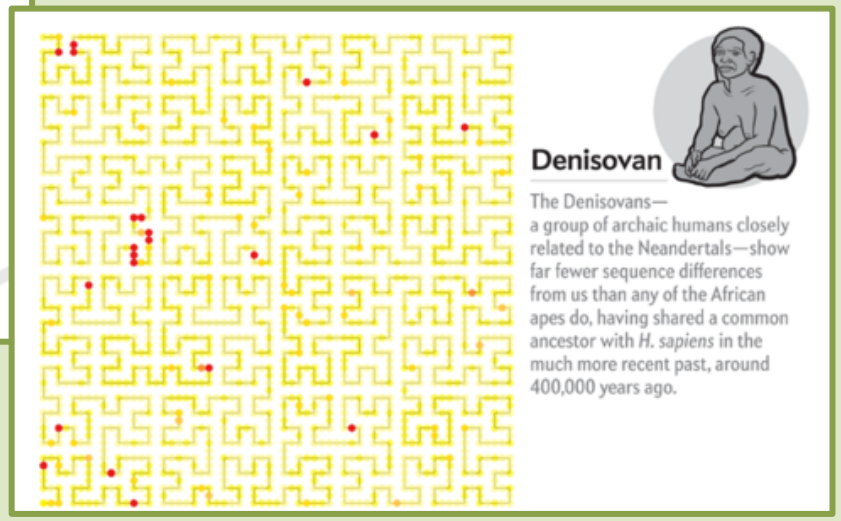


Genomic differences



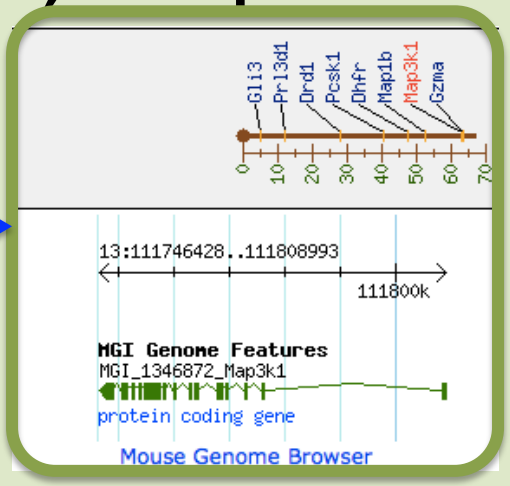




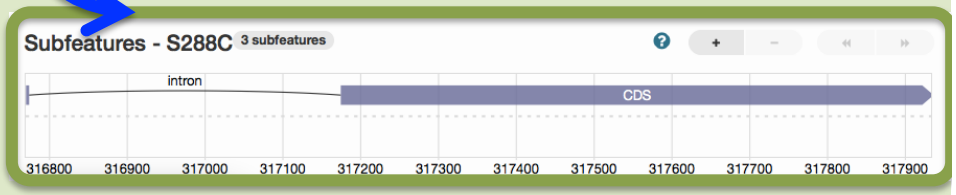


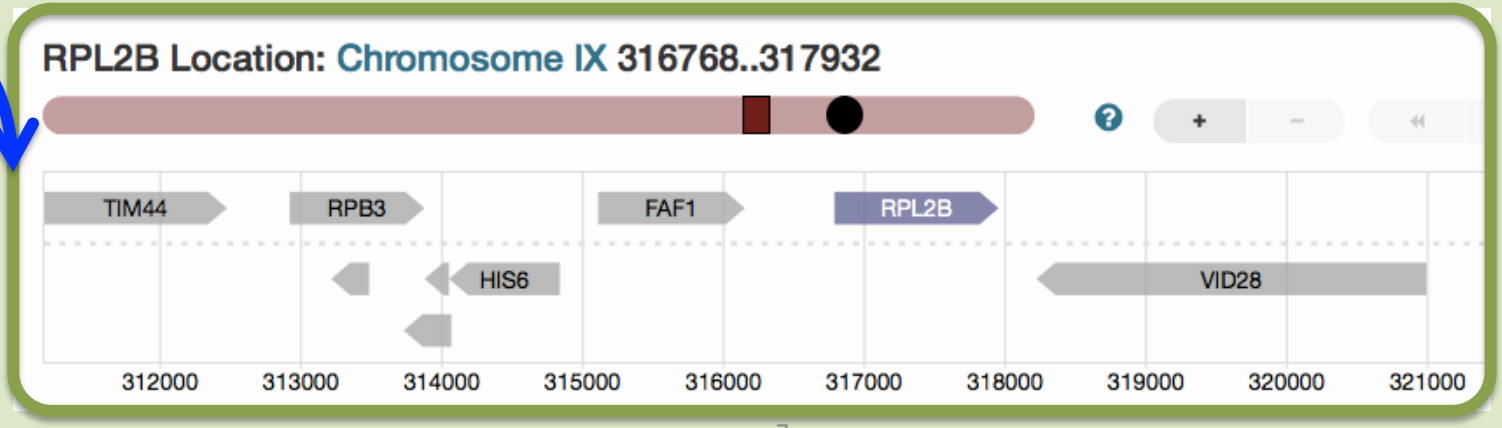
Gene structure, ideograms, maps

Genetic Map	Chromosome 13 63.36 cM Detailed Genetic Map ± 1 cM Mapping data(3)
Sequence Map	 Chr13:111746428-111808993 bp, - strand From VEGA annotation of GRCm38



Feature	Relative Coordinates	Coordinates
CDS	1..4	chrIX:316768..316771
intron	5..404	chrIX:316772..317171
CDS	405..1165	chrIX:317172..317932





Most of the curated data is text

Gene Ontology

Gene Ontology Details

Summary: Subunit of the PeBoW complex and the preribosomal complex; binds to the large ribosomal subunit rRNA and is involved in the processing of the rRNA and the biogenesis of the large ribosomal subunit

View computational annotations

Molecular Function

Manually Curated:
- large ribosomal subunit rRNA binding (IDA)

Biological Process

Manually Curated:
- ribosomal large subunit biogenesis (IPI)
- rRNA processing (IMP)

Cellular Component

Manually Curated:
- PeBoW complex (IDA, IPI)
- preribosome, large subunit precursor (IDA)

High-Throughput:
- nucleolus (IDA)
- nucleus (IDA)
- preribosome, large subunit precursor (IDA)

Phenotype

Phenotype Details

Classical Genetics

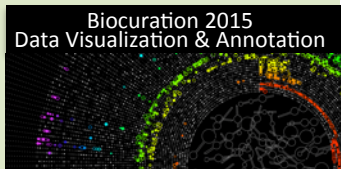
conditional:
- cell cycle passage through the metaphase-anaphase transition: delayed
- resistance to hydroxyurea: decreased
- spindle morphology: abnormal

Large-scale Survey

reduction of function:
- competitive fitness: decreased
- chromosome/plasmid maintenance: abnormal

conditional:
- colony sectoring: increased

null:
- haploinsufficient
- inviable



Visualizing interaction data

Interaction

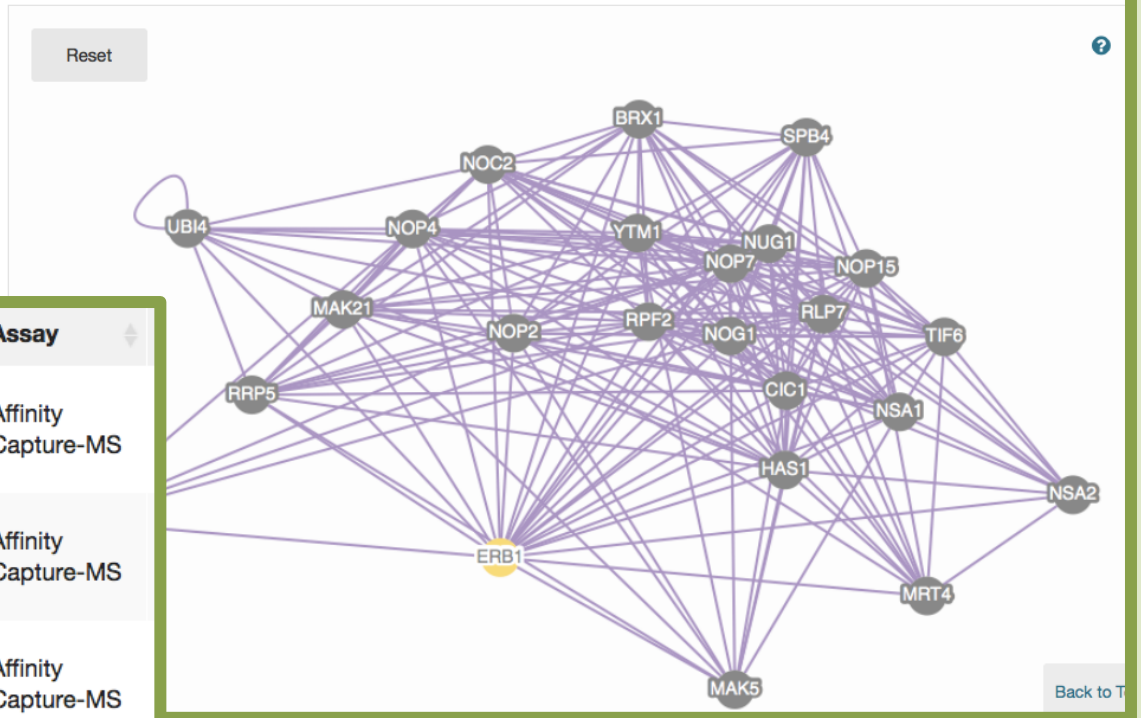
249 total interactions for 129 unique genes

Physical Interactions

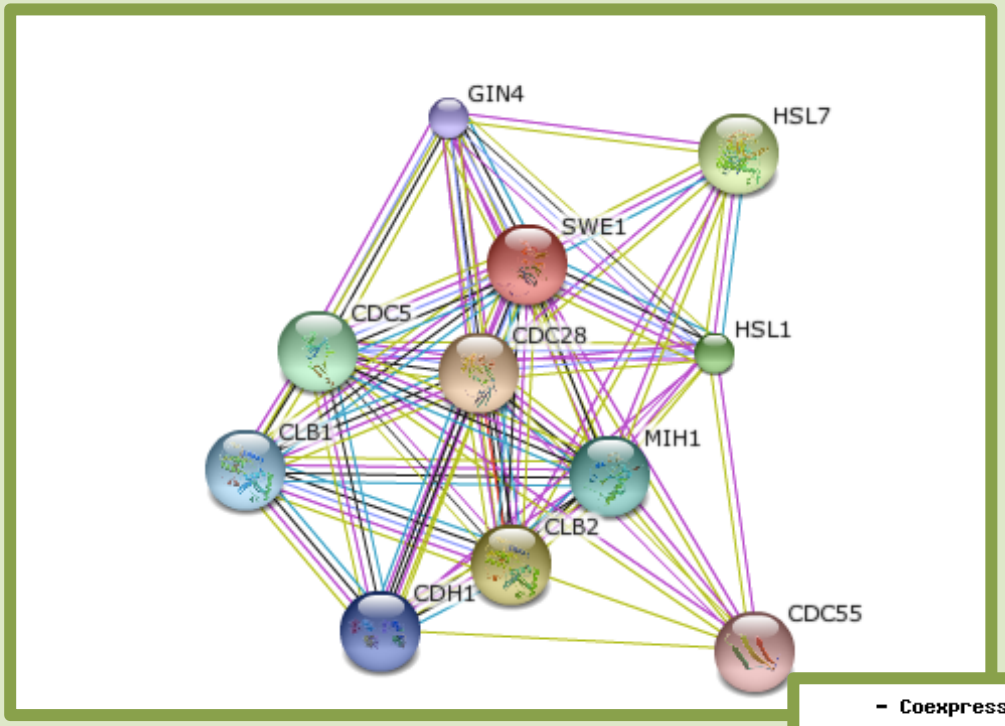
- Affinity Capture-MS: 220
- Affinity Capture-RNA: 6
- Affinity Capture-Western: 10
- Biochemical Activity: 4
- Co-purification: 2
- Reconstituted Co
- Two-hybrid: 2

	Interactor	Type	Assay
	ACO1	Physical	Affinity Capture-MS
	ARD1	Physical	Affinity Capture-MS
	ASK10	Physical	Affinity Capture-MS
ℹ	ATG1	Physical	Biochemical Activity
ℹ	BRE5	Physical	Affinity Capture-MS

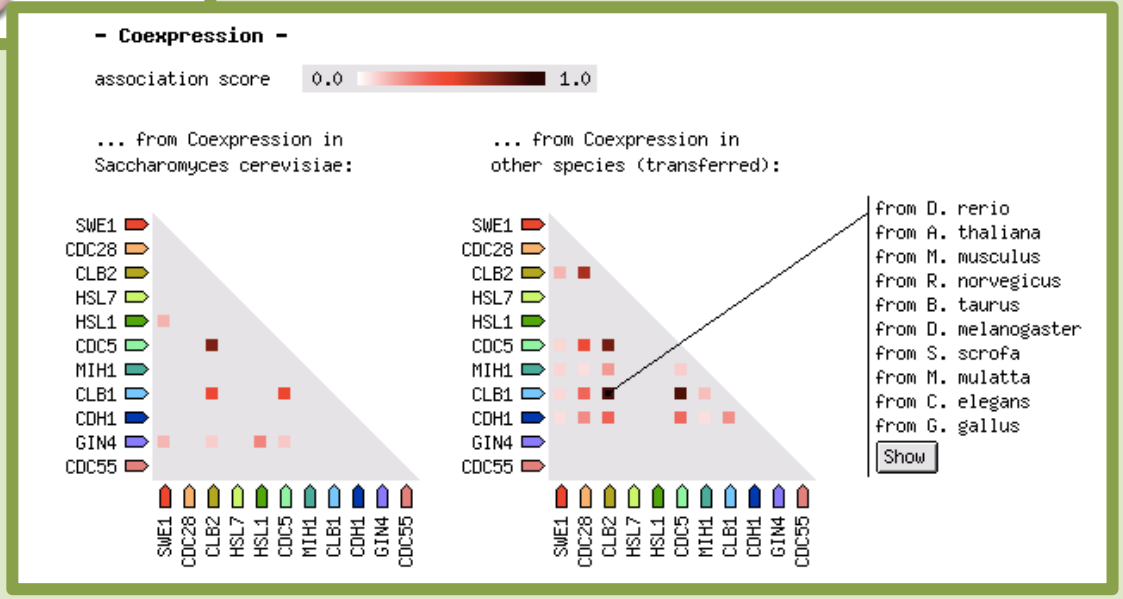
Interaction Network



Overlaying curated data



http://string-db.org



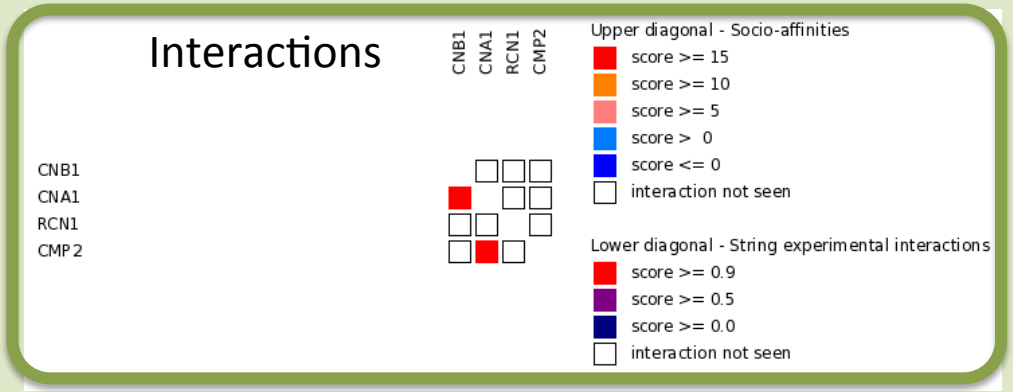
Complexes, interactions, and more

Complex SGD_GO:0005955
calcineurin complex subunits

Crosstalk (Help)

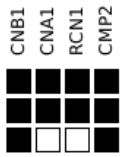
SGD GO:0008287
Krogan 57

Interactions



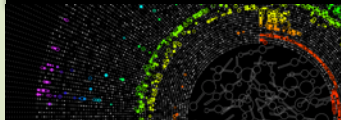
Subcellular Locations (Help)

calcineurin complex
protein serine/threonine phosphatase complex
cytoplasm

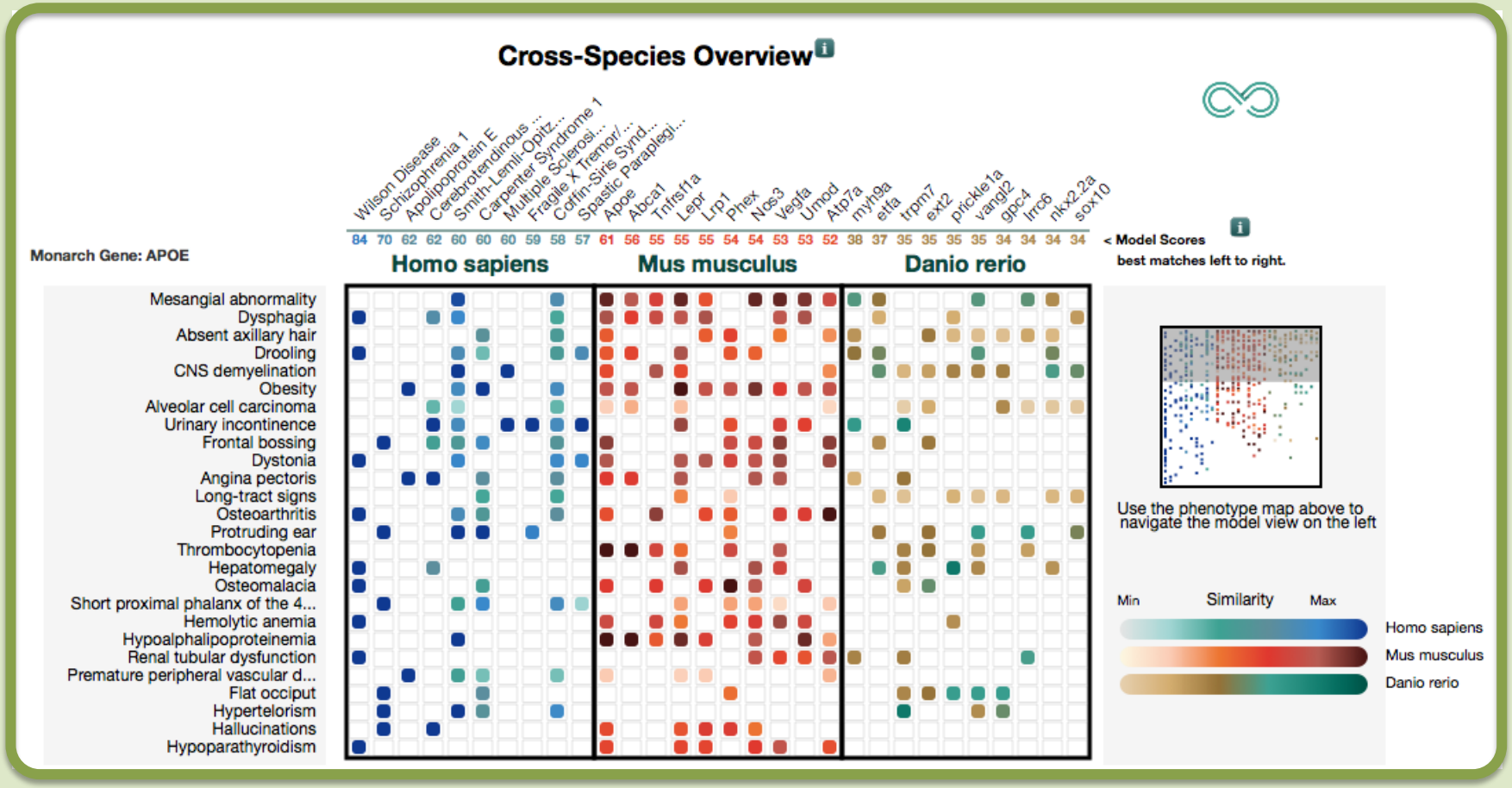


Copy Numbers (Help)

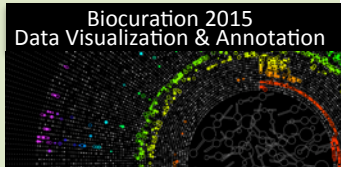
http://3drepertoire.russelllab.org/



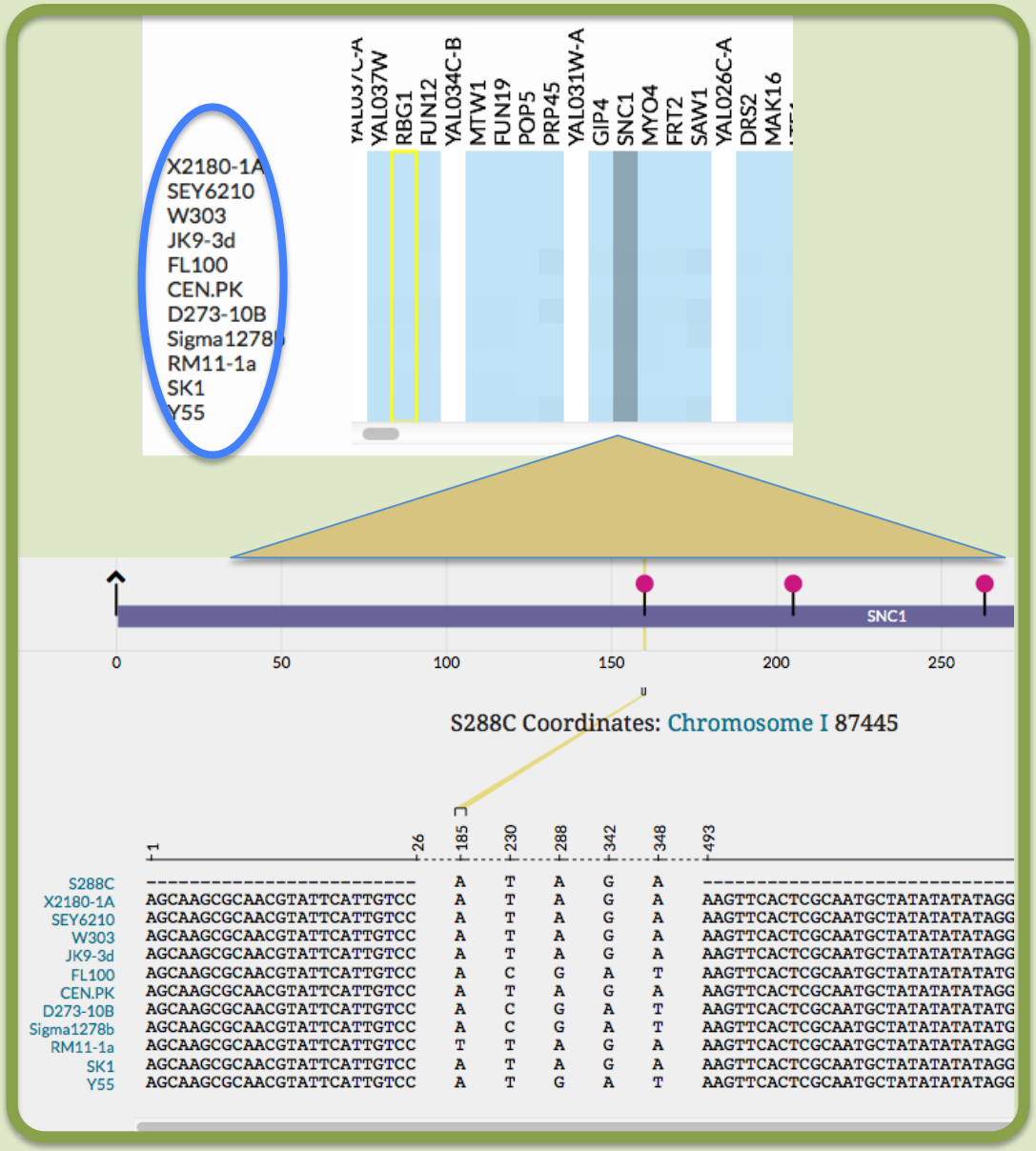
Phenotypes and diseases



http://monarchinitiative.org/



Sequence variations

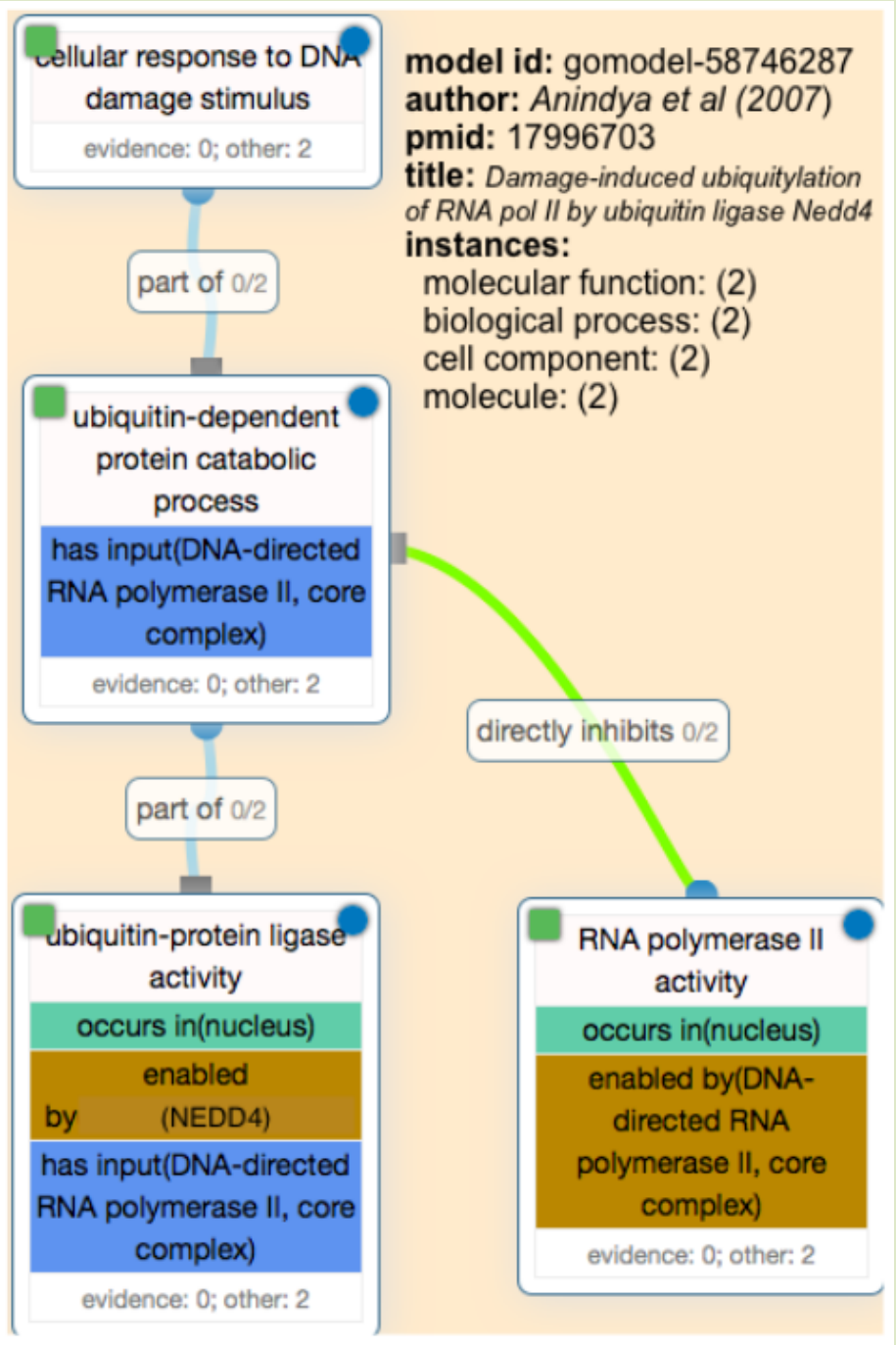


Sequence variation in various strains of *S. cerevisiae*

Molecular Model Editing Environment

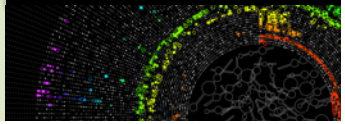
Noctua – prototype from GOC

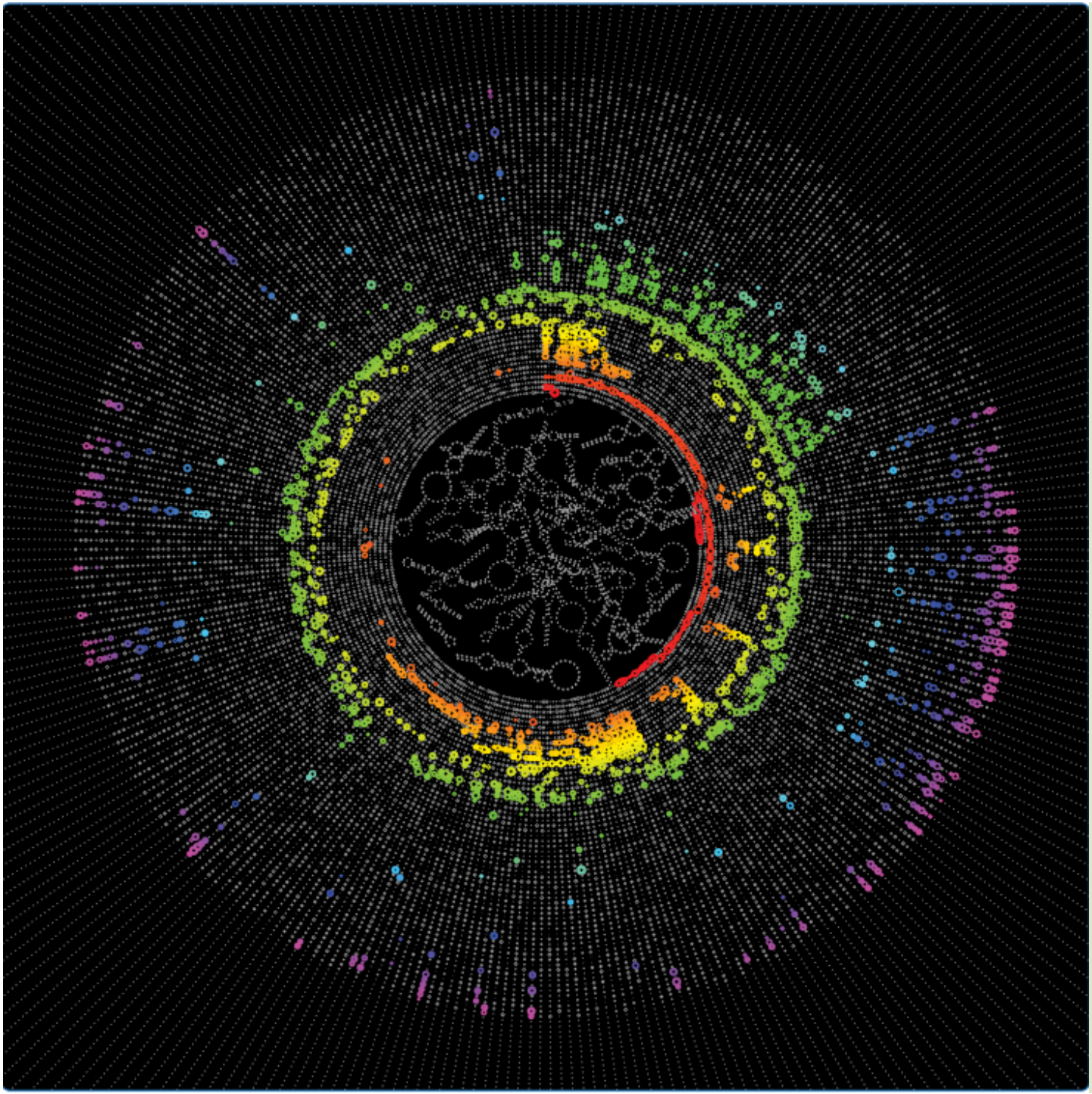
- Each node (box) is a function or process.
- Other nodes are folded in as OWL expressions.
- Users may add and drag elements
- Supports real time collaboration



Understanding the Data

Much of the interpretation requires human judgment. Visualization improves our understanding and increases our chances of extracting meaningful conclusions.

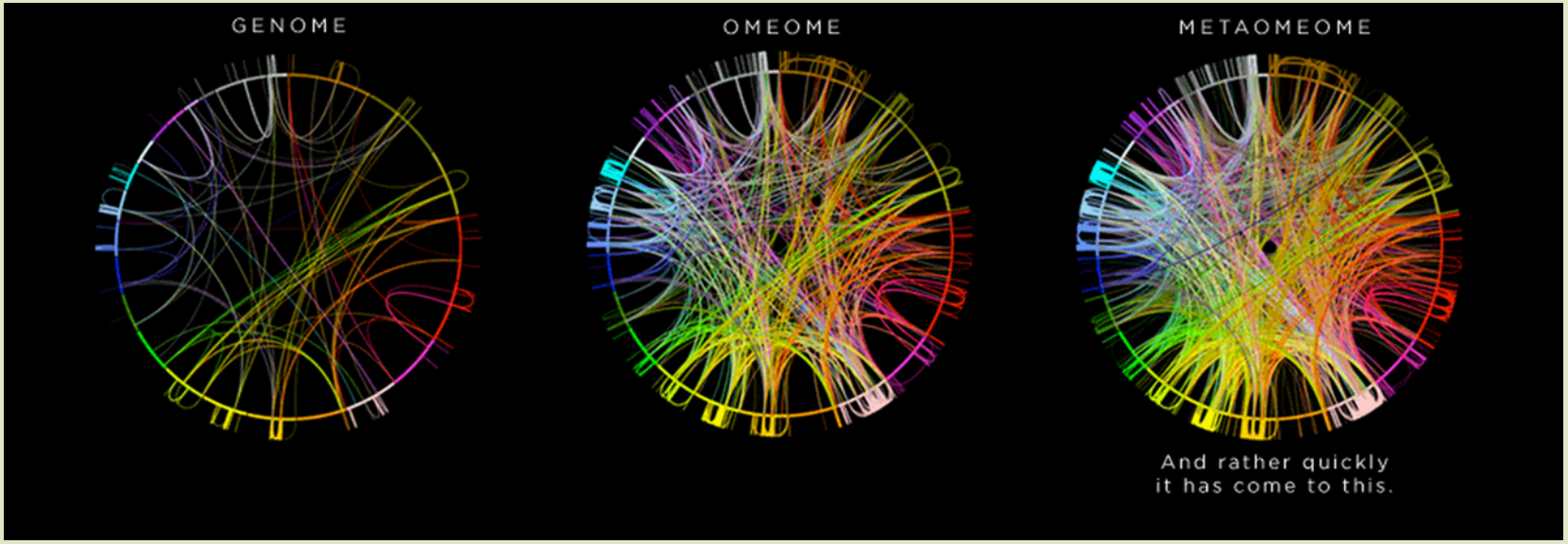




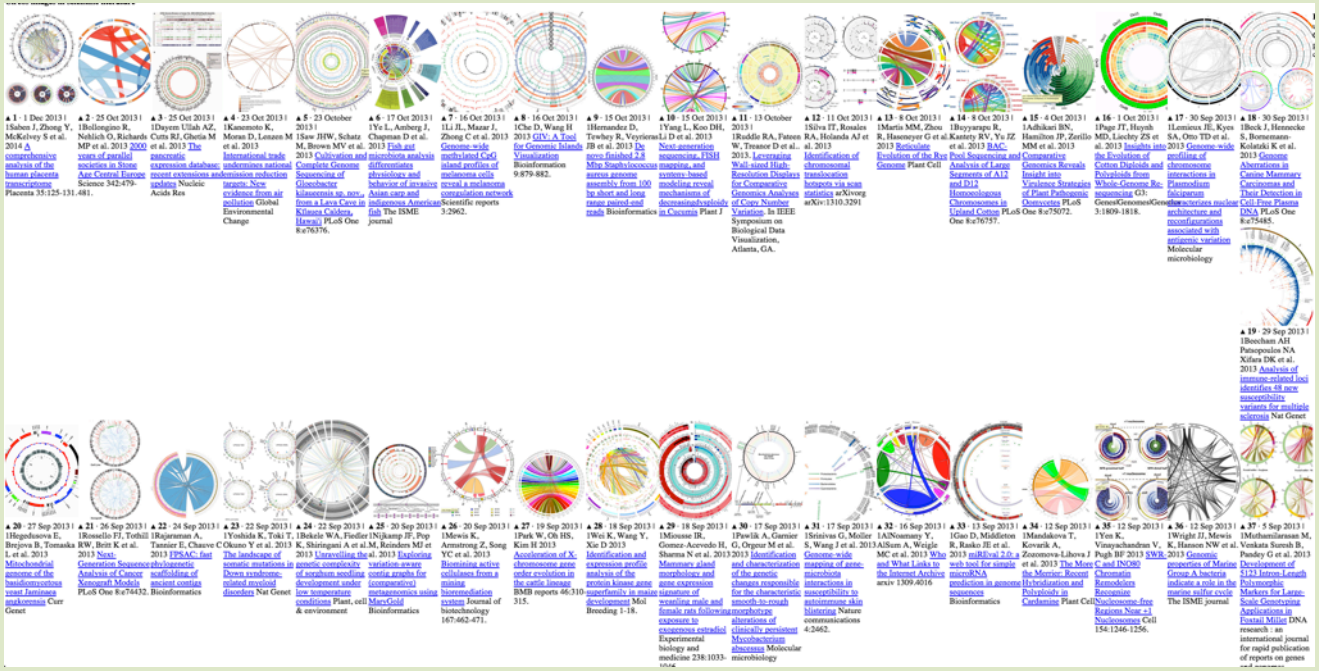
Cancer miRNome revealed survival differences in diffuse large B-cell lymphoma patients

Lim et al. *Genome Biol* **16**:18 (2015)

Circos

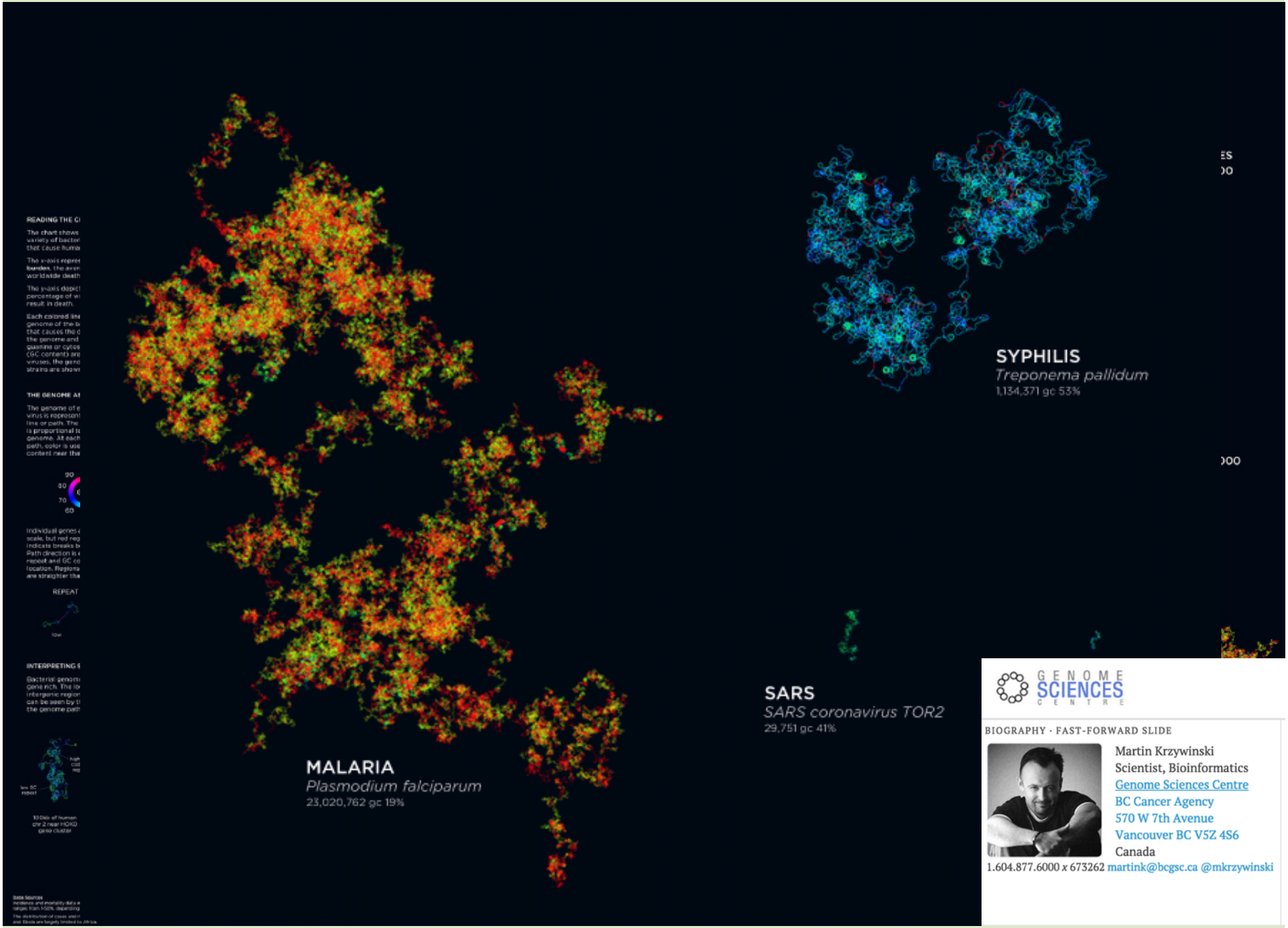


ENCODE



Circular Genome Data Visualization

- Human placenta transcriptome
- Pancreatic expression db
- Wall-sized High-res display for comparative analys. of CNV
- Chromosomal translocations
- Variant identification in multiple sclerosis
- Sorghum seedling development under Low Temp conditions
- Etc., etc., etc…



GENOME SCIENCES CENTRE

BIOGRAPHY - FAST-FORWARD SLIDE

Martin Krzywinski
Scientist, Bioinformatics
Genome Sciences Centre
BC Cancer Agency
570 W 7th Avenue
Vancouver BC V5Z 4S6
Canada
1.604.877.6000 *x* 673262 martink@bcgsc.ca @mkrzywinski

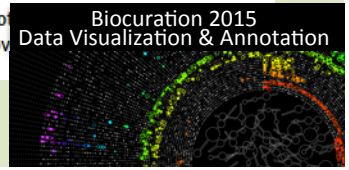
Visualizing sequencing data

Table 1 | Tools for visualizing sequencing data

Name	Cost	OS	Description	URL
Stand-alone tools				
ABySS-Explorer[25]	Free	Win, Mac, Linux	Interactive assembly structure visualization tool	http://tinyurl.com/abyss-explorer/
CLC Genomics Workbench	$	Win, Mac, Linux	Integrates NGS data visualization with analysis tools; user friendly	http://www.clcbio.com/
Consed[3]*	Free	Mac, Linux	Widely used; assembly finishing package; NGS compatible	http://www.phrap.org/
DNASTAR Lasergene[14]	$	Win, Mac	Analysis suite with an assembly finishing package; NGS compatible	http://www.dnastar.com/
EagleView[17]	Free	Win, Mac, Linux	Assembly viewer; compatible with single-end NGS	http://tinyurl.com/eagleview/
Gap[12,13]	Free	Linux	Widely used; assembly finishing package; Gap5 is NGS compatible	http://staden.sourceforge.net/
Hawkeye[6]	Free	Win, Mac, Linux (S)	Sanger sequencing assembly viewer	http://amos.sourceforge.net/hawkeye/
Integrative Genomics Viewer (IGV)*	Free	Win, Mac, Linux	Genome browser with alignment view support (**Table 2**); NGS compatible	http://www.broadinstitute.org/igv/
MapView[18]	Free	Win, Linux	Read alignment viewer; custom file format for fast NGS data loading	http://evolution.sysu.edu.cn/mapview/
MaqView	Free	Mac, Linux	Read alignment viewer; fast NGS data loading from Maq alignment files	http://maq.sourceforge.net/
Orchid	Free	Linux (S)	Assembly viewer customized to display paired-end relationships	http://tinyurl.com/orchid-view/
Sequencher	$	Win, Mac	Assembly finishing package	http://www.genecodes.com/
SAMtools tview[8]	Free	Win, Mac, Linux	Simple and fast text alignment viewer NGS compatible	http://samtools.sourceforge.net/
Web-based tools				
LookSeq[19]	Free		Uses AJAX; y axis for insert size; user configures data resources; NGS compatible	http://lookseq.sourceforge.net/
NCBI Assembly Archive Viewer[7]	Free		Graphical interface to contig and trace data in NCBI's Assembly Archive	http://tinyurl.com/assmbrowser/

Free means the tool is free for academic use; $ means there is a cost. OS, operating system: Win, Microsoft Windows; Mac, Macintosh OS X. Tools running on Linux usually also run on o
(S) indicates that compilation from source is required. "Assembly finishing package" enables interactive sequence editing and/or integration with tools for automated assembly improv
*Our recommendation

Nielsen et al. *Nature Met Suppl* 7: 3s. (2010)



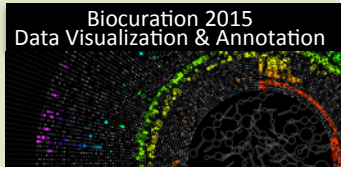
Genome Browsers

Table 2 | Genome browsers

Name	Description	URL
Stand-alone browsers		
Argo	Supports manual annotation of whole genomes	http://tinyurl.com/argo-combo
CGView[82]	Circular genome visualization	http://wishart.biology.ualberta.ca/cgview/
Gaggle[83]	Genome browser within an analysis framework; good microarray support	http://gaggle.systemsbiology.net/
Integrative Genomics Viewer (IGV)*	Flexible user interface; can integrate metadata as heat maps	http://www.broadinstitute.org/igv/
Integrated Genome Browser (IGB)[84]	GenoViz project genome browser; reusable visualization components	http://genoviz.sourceforge.net/
NCBI Genome Workbench	Displays sequence data in many views; integrated with BLAST	http://tinyurl.com/gbench/
Web-based browsers		
AnnoJ	Designed for NGS data; uses AJAX; assemble by html configuration	http://www.annoj.org/
Cancer Molecular Analysis Portal	Integrates clinical data; designed for TCGA project	https://cma.nci.nih.gov/cma-tcga/
Ensembl[31,32]*	Comprehensive genome browser and database; strong user support	http://www.ensembl.org/
GBrowse[28]*	GMOD[28]* component; back end of WormBase, FlyBase; v2.0 uses AJAX	http://gmod.org/wiki/Gbrowse
Genome Projector[42]	Offers circular and pathway views; user configures data resources	http://tinyurl.com/gprojector/
JBrowse[39]	Component of GMOD[28]*; AJAX interface; user configures data resources	http://jbrowse.org/
JGI	Supports live annotation; primary portal for JGI genome projects	http://genome.jgi-psf.org/
NCBI Map Viewer[33]	Vertically oriented viewer; integrated with NCBI resources and tools	http://tinyurl.com/mapview1/
UCSC Genome Browser[30]*	Comprehensive genome browser and database; strong user support	http://tinyurl.com/ucscbrowser/
UCSC Cancer Genomics Browser[43]	Integrates clinical data; offers a pathway view; portal for TCGA data	http://genome-cancer.ucsc.edu/
UTGB	Toolkit to construct personalized browser; uses AJAX; user configures data resources	http://utgenome.org/
X:map[41]	Customized to view Affymetrix exon arrays	http://xmap.picr.man.ac.uk/

All listed tools are free for academic use, and all are available for Microsoft Windows, Macintosh OS X and Linux. Tools running on Linux usually also run on other versions of Unix.

Nielsen et al. *Nature Met Suppl* 7: 3s. (2010)



Comparative Genomics Visualization

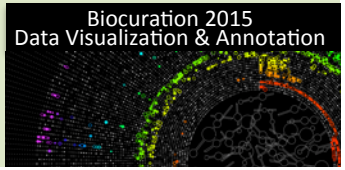
Table 3 | Tools for comparative genomics visualization

Name	Description	Data	URL
Web-based tools			
Cinteny[67]	Three-scale view of synteny calculated from user-specified markers	H	http://cinteny.cchmc.org/
CoGe SynMap[85]	Dot plots from DAGChainer[61] alignments; histograms of synonymous substitutions	H	http://tinyurl.com/synmap/
GenomeMatcher[63]	A rich, mostly dot plot–based viewer displaying alignments and annotation	F,E,G	http://tinyurl.com/genomematcher/
MEDEA*	A Flash-based suite of linked-track, dot-plot and global-synteny viewing tools	C	http://tinyurl.com/broadmedea/
MultiPipMaker[86]	Vertically arranged display of user-supplied multiple alignments	F	http://pipmaker.bx.psu.edu/pipmaker/
PhIGs[69]	Ideogram-style interactive display of orthologs across >75 genomes	H	
UCSC Genome Browser[72]*	Conservation tracks within popular UCSC genome browser	H,F,G	http://genome.ucsc.edu/cgi-bin/hgGateway/
VISTA[87]*	Conservation tracks connected to a variety of analysis tools	H	http://genome.lbl.gov/vista/index.shtml
VSV, VISTA-Dot*	Three-scale viewer for synteny and dynamic, interactive dot plots for whole-genome DNA alignments	H	http://genome.jgi-psf.org/synteny/
Stand-alone tools			
ACT[76]	Linked-track views; annotation track search; stacking of multiple genomes	E,GF,D	http://www.sanger.ac.uk/Software/ACT/
Circos[70]	Circle-graph presentation of synteny; animations for increased dimensionality	C	http://mkweb.bcgsc.ca/circos
CMap[88]	Stacked vertical depictions of arbitrary relations among DNA segments	D,S	http://gmod.org/wiki/CMap
Combo[77]	Dot-plot and linked-track views; integration of annotation in both views	G,F,C	http://tinyurl.com/argo-combo
GBrowse_syn	GMOD[28]* component; highly customizable linked-track view of synteny	D,S	http://gmod.org/wiki/GBrowse_syn
MizBee[71]	Synteny visualized using circular and linked-track views at multiple scales	C	http://mizbee.org/
Sybil[78]	Local and global views of synteny based on BlastP and protein clustering	D	http://sybil.sourceforge.net/
SynBrowse[75]	GMOD[28] component; local synteny based on gene order, orthology or structure	D	http://www.synbrowse.org/
SynView[79]	GMOD[28] component; synteny at different scales with multiple feature tracks	D	http://gmod.org/wiki/SynView

All tools listed are free and are either web-based or available for all three operating systems. The Data column describes the formats accepted for display within each tool: H, only alignment data hosted at the tool's website; F, FASTA format; E, EMBL/GenBank/DDJB format; G, gff format; C, a custom text-based format; D, designed for use with a user-hosted database; S, requires hosting from a user-supplied web server.

*Our recommendations

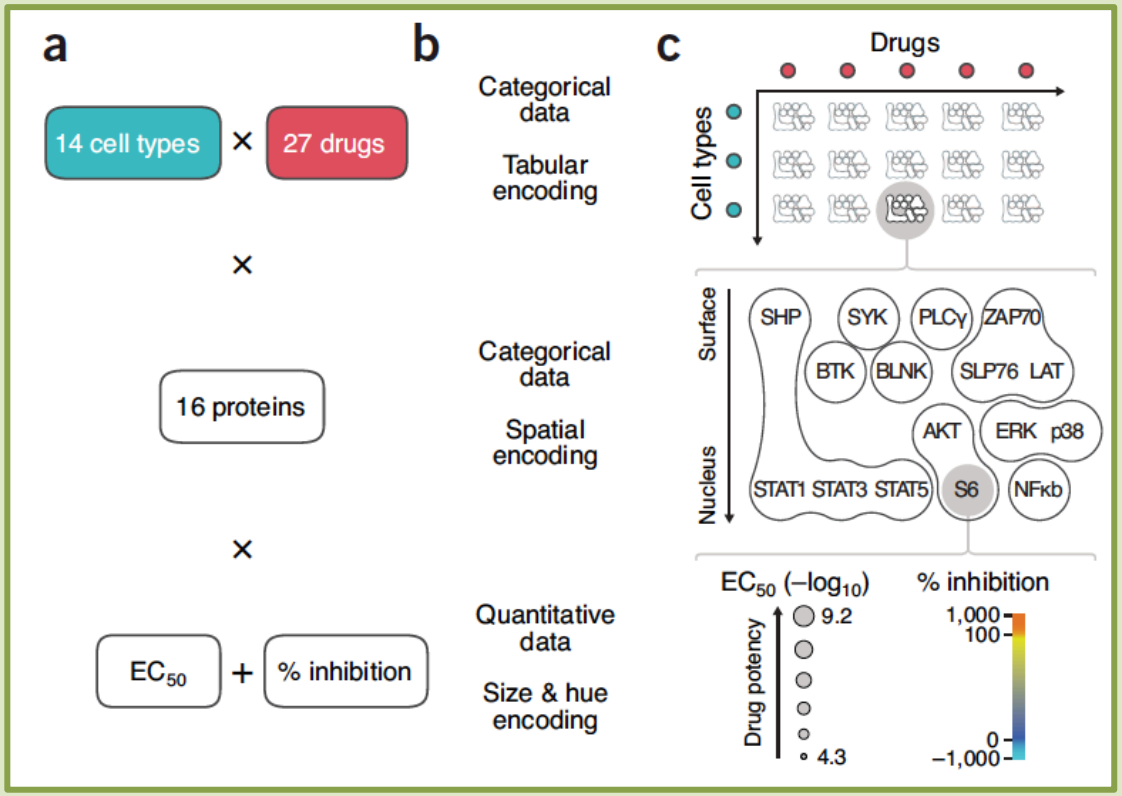
Nielsen et al. *Nature Met Suppl* 7: 3s. (2010)



Communicating Complex Data

Focus on meaning instead of structure—anchor the figure to relevant biology rather than to methodological details.

1) What are the interesting findings, and what representation would communicate them clearly?

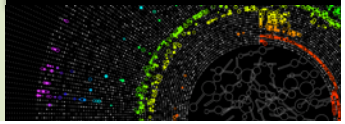


2) Forgo conventional approaches to displaying multidimensional data. Better to project the data onto familiar visual paradigms, such as a protein network or pathway, to saliently show biological effects in a functional context.

Krzywinski and Savig.
Nature Methods **10**:7, 595 (2013)

Storytelling

- Relate your data using the age-old custom of telling a story.
  - Stories have the capacity to delight and surprise and to spark creativity by making meaningful connections between data and the ideas, interests and lives of your readers.



Open-source: dissemination & contributions

- Genetic & genomic information is more valuable when shared
- Promote and encourage Open Science: transparency, reproducibility, data provenance. E.g. Open Bioinformatics Foundation http://open-bio.org
- Public repositories make software easily accessible and allow collaborative efforts, e.g. GitHub https://github.com/

Our Panelists

1. **Lorna Richardson**:
   eMouseAtlas and Image Informatics

2. **Justyna Szostak:**
   Curated Causal Biological Network Models