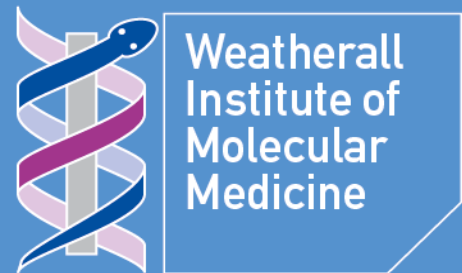


Methods and Techniques in Bioinformatics

(From DNA bases to image databases)

Stephen Taylor

MRC WIMM Centre of Computational Biology



MRC WIMM Centre for Computational Biology

Using computational biology to help understand complex biological systems and combat diseases, from blood disorders to cancer and diabetes.

IN THIS SECTION

About Us



Research Groups



People

Resources



Training



Account Request Form

CCB Account and Support



Contact us

Do you have a query or would like to find out who to contact within the Centre?

Email us at
ccb@imm.ox.ac.uk

Machine learning meets microscopy - Subjectivity dethroned

How can we remove subjectivity from science, whilst keeping the human in the loop?

[Read more](#)



LATEST PUBLICATIONS

A revised model for promoter competition based on multi-way chromatin interactions at the α -globin locus.

Oudelaar AM, et al, (2019), Nat Commun, 10

A Spontaneous Ring-Opening Reaction Leads to a Repair-Resistant Thymine Oxidation Product in Genomic DNA.

Sahakyan AB, et al, (2020), Chembiochem, 21, 320 - 323

Haplotype matching in large cohorts using the Li and Stephens model.

NEWS



Role-playing computer game helps players understand how vaccines work on a global scale

8 October 2020



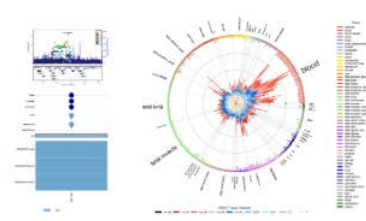
Marieke Oudelaar awarded prestigious Lise Meitner Excellence Program grant

8 October 2020

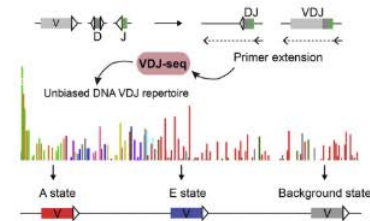
Research Groups



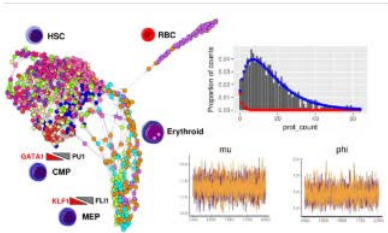
Hughes Group



Iotchkova Group: Statistical Genetics



Koohy Group: Machine Learning and Integrative Approaches in Immunology



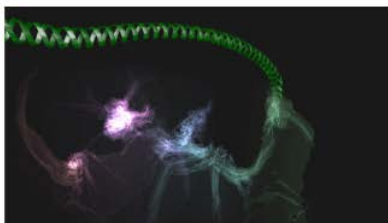
Morrissey Group: Quantitative biology of cell fate and tissue dynamics




Sahakyan Group: Integrative Computational Biology and Machine Learning



Sims Group: Computational Genomics



Taylor Group: Analysis, Visualisation and Informatics



Computational Biology and Bioinformatics is all about data...

- Definition
 - Bioinformatics is the computational analysis and storage of biological data
- Derivation
 - informatique – French for ‘data processing’
- Goal
 - To discover new biological insights using computers and biology

What is bioinformatics?

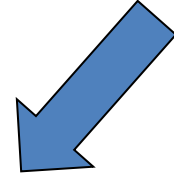
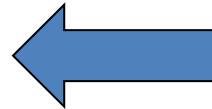
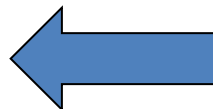
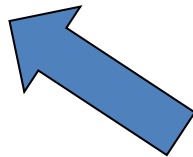
Experiment



Analysis

Sequence
Structure
Function
Evolution
Pathway
Interaction
Mutation
Expression

Hypothesis



Why use bioinformatics?



Find an answer quickly

Most *in silico* biology is faster than *in vitro*



Massive amounts of data
to analyse

Need to make use of all information

Not possible to do analysis by hand

Can't organise and store information only using
lab note books

Automation is key



However!

All results of computer analysis should to be
verified by biologists

Bioinformatics databases



Public databases are the most important entity in bioinformatics



Store knowledge about

Sequence e.g. EMBL/Genbank

HTS Experiments e.g. GEO

Structure e.g. PDB

Pathways e.g. KEGG, Metacore

Diseases e.g. OMIM




Genome Architecture: e.g. UCSC



Can be searched in a variety of ways


e.g. keyword, sequence, pattern

Keyword

 [Resources](#)  [How To](#) 

[bioinfobloke](#) [My NCBI](#) [Sign Out](#)

Search NCBI databases



About 1,673,010 search results for "p53"

Literature

Books	1,279	books and reports
MeSH	158	ontology used for PubMed indexing
NLM Catalog	108	books, journals and more in the NLM Collections
PubMed	71,937	scientific & medical abstracts/citations
PubMed Central	93,434	full-text journal articles

Health

ClinVar	225	human variations of clinical significance
dbGaP	22	genotype/phenotype interaction studies
GTR	110	genetic testing registry
MedGen	72	medical genetics literature and links
OMIM	583	online mendelian inheritance in man
PubMed Health	71	clinical effectiveness, disease and drug reports

Genomes

Assembly	1	genomic assembly information
BioProject	642	biological projects providing data to NCBI
BioSample	307	descriptions of biological source materials
Clone	0	genomic and cDNA clones
dbVar	1,464	genome structural variation studies
Epigenomics	0	epigenomic studies and display tools
Genome	5	genome sequencing projects by organism
GSS	36	genome survey sequences
Nucleotide	24,181	DNA and RNA sequences
Probe	3,507	sequence-based probes and primers
SNP	6,592	short genetic variations
SRA	440	high-throughput DNA and RNA sequence read archive
Taxonomy	0	taxonomic classification and nomenclature catalog

Genes

EST	796	expressed sequence tag sequences
Gene	7,879	collected information about gene loci
GEO DataSets	8,899	functional genomics studies
GEO Profiles	1,403,459	gene expression and molecular abundance profiles
HomoloGene	38	homologous gene sets for selected organisms
PopSet	94	sequence sets from phylogenetic and population studies
UniGene	414	clusters of expressed transcripts

Proteins

Conserved Domains	120	conserved protein domains
Protein	29,695	protein sequences
Protein Clusters	15	sequence similarity-based protein clusters
Structure	1,082	experimentally-determined biomolecular structures

Chemicals

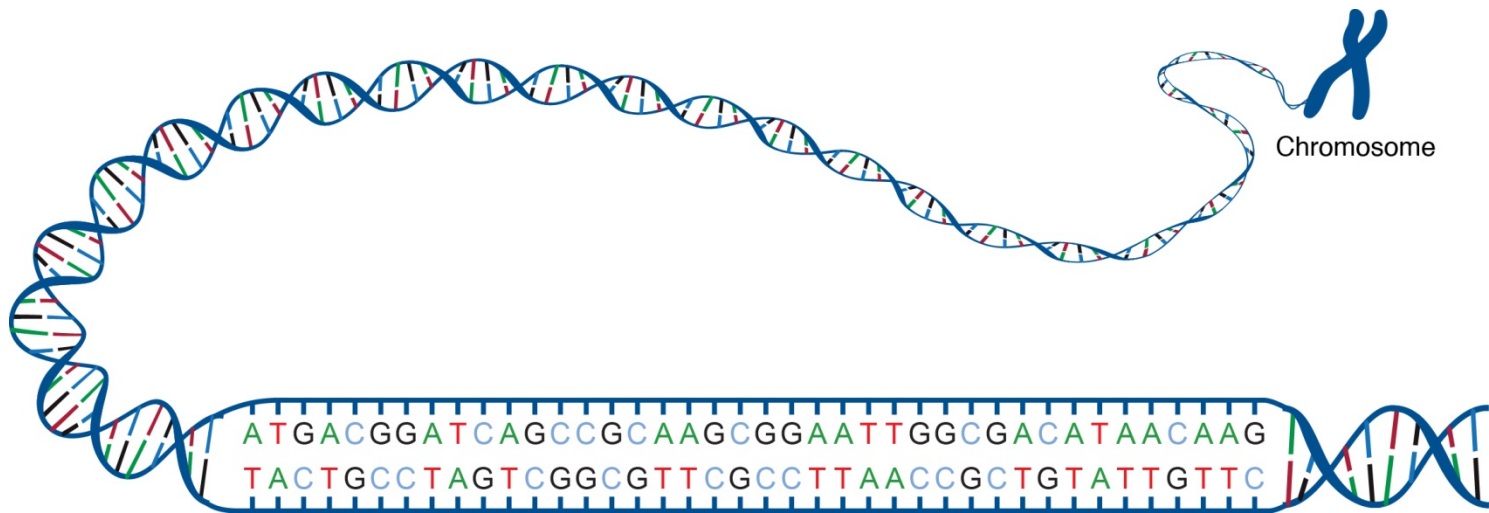
BioSystems	3,799	molecular pathways with links to genes, proteins and chemicals
PubChem BioAssay	10,848	bioactivity screening studies
PubChem Compound	8	chemical information with structures, information and links
PubChem Substance	650	deposited substance and chemical information



Bioinformatics Tools

- Hundreds of computer programs
- Many freely available
- Generally available on UNIX or LINUX
- Often interact with bioinformatics databases
- Many accessible via the WWW
- Some require very powerful computers to run on
- CCB provide a environment to do this

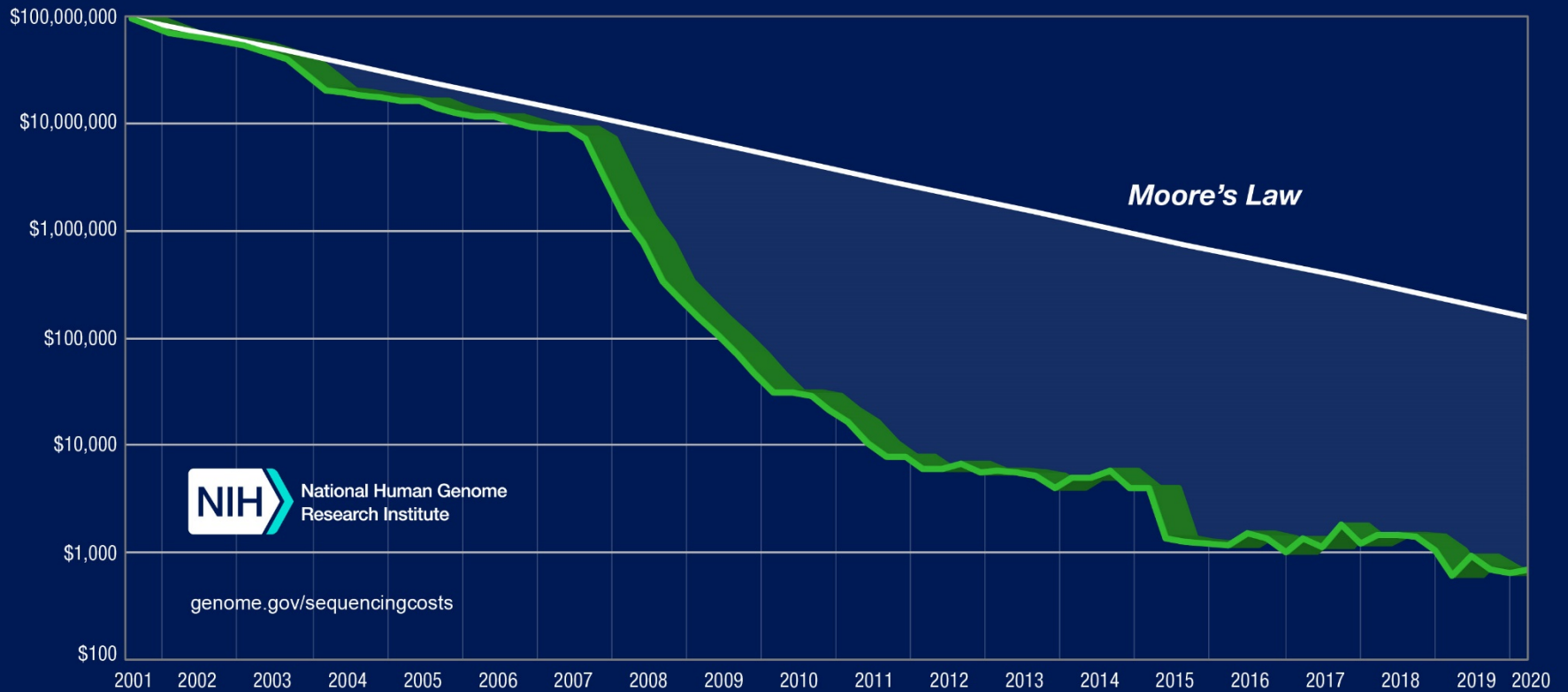
DNA



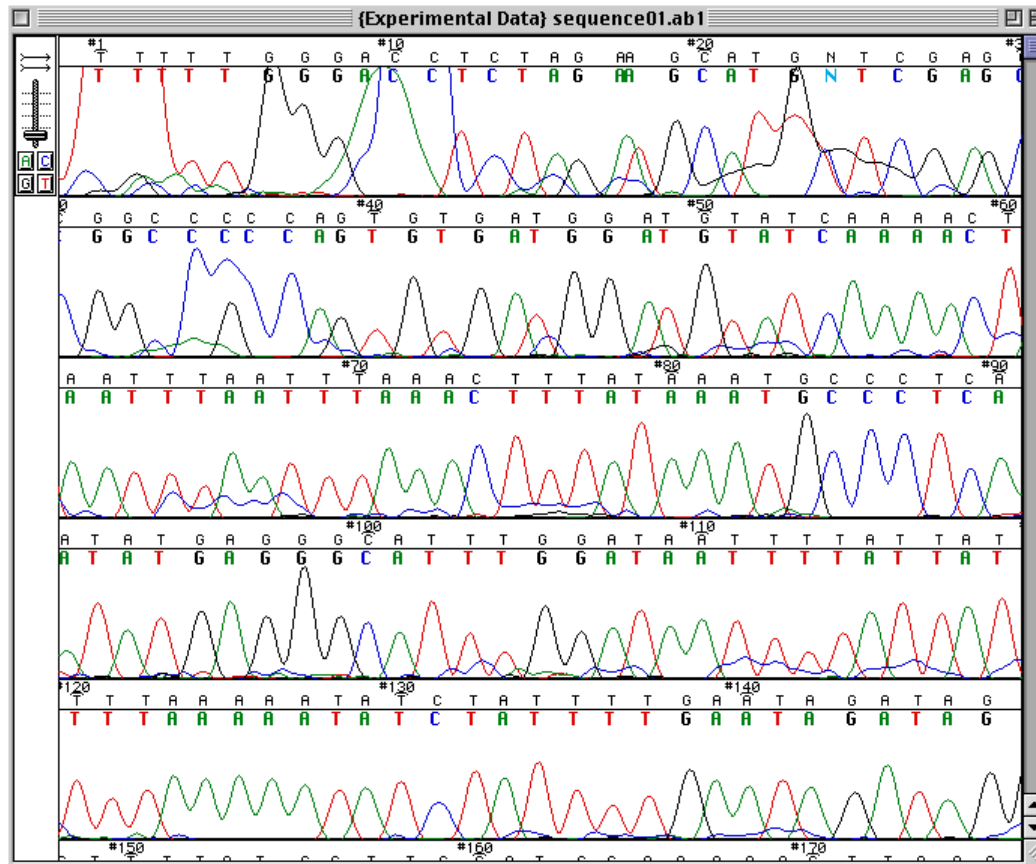
The Human Genome Project (1990- 2003)

- Cost \$3 billion
- Could not have been achieved without bioinformatics
- Goals
 - *identify* all the 20,500 genes in human DNA,
 - *determine* the sequences of the 3 billion chemical base pairs that make up human DNA
 - *store* this information in databases
 - *improve* tools for data analysis
 - *transfer* related technologies to the private sector, and
 - *address* the ethical, legal, and social issues (ELSI) that may arise from the project.
- Need to bring together and store vast amounts of information from
 - Lab equipment and experiments
 - Computer Analysis
 - Human Analysis
 - Make visible to the world's scientists

Cost per Human Genome

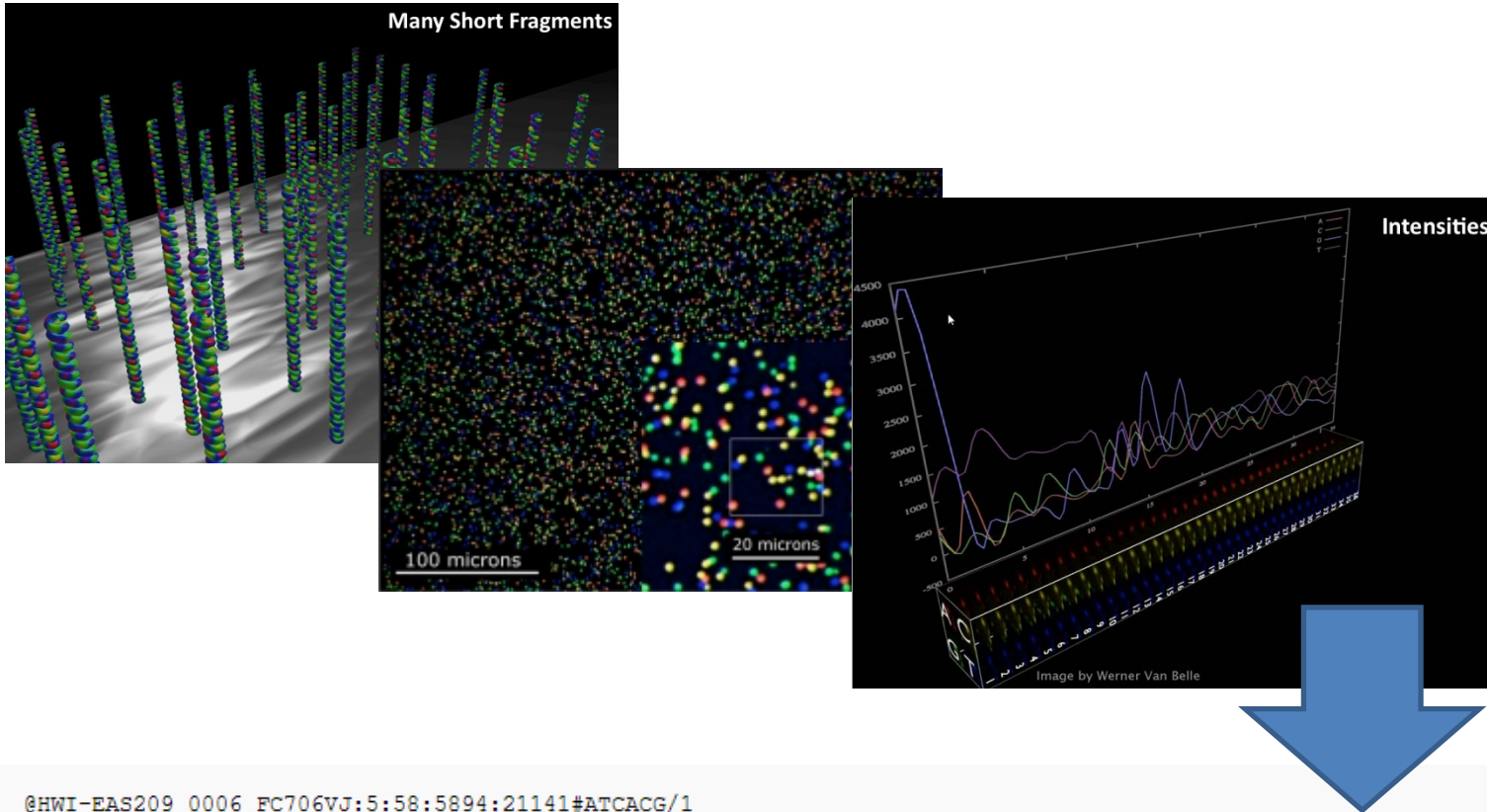


Sanger Sequencing



Read length typically 500-600bp (up to 800bp)

Next Generation Sequencing

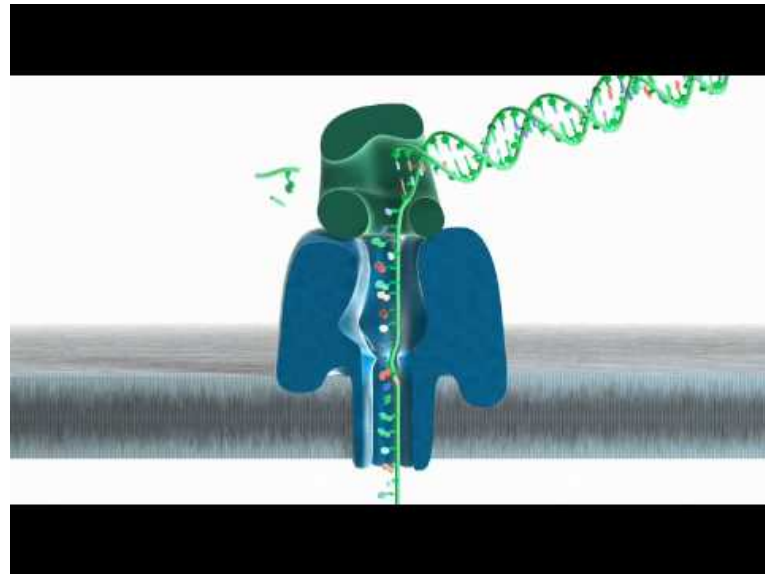


```
@HWI-EAS209_0006_FC706VJ:5:58:5894:21141#ATCACG/1
TTAATTGGTAAATAAATCTCCTAATAGCTTAGATTTTACCTTNNNNNNNNNTAGTTTCTTGAGATTGTGTGGGGGAGACATTTTTGTGATTGCCTTGAT
+HWI-EAS209_0006_FC706VJ:5:58:5894:21141#ATCACG/1
efcffffffcfeeffcffffffddfd`feed]`)_Ba_^_[YBBBBBBBBBRTT\]][]dddd`ddd^dddadd^BBBBBBBBBBBBBBBBBBBBBBBBBB
```

<http://werner.yellowcouch.org/Papers/pippres0802/index.html>

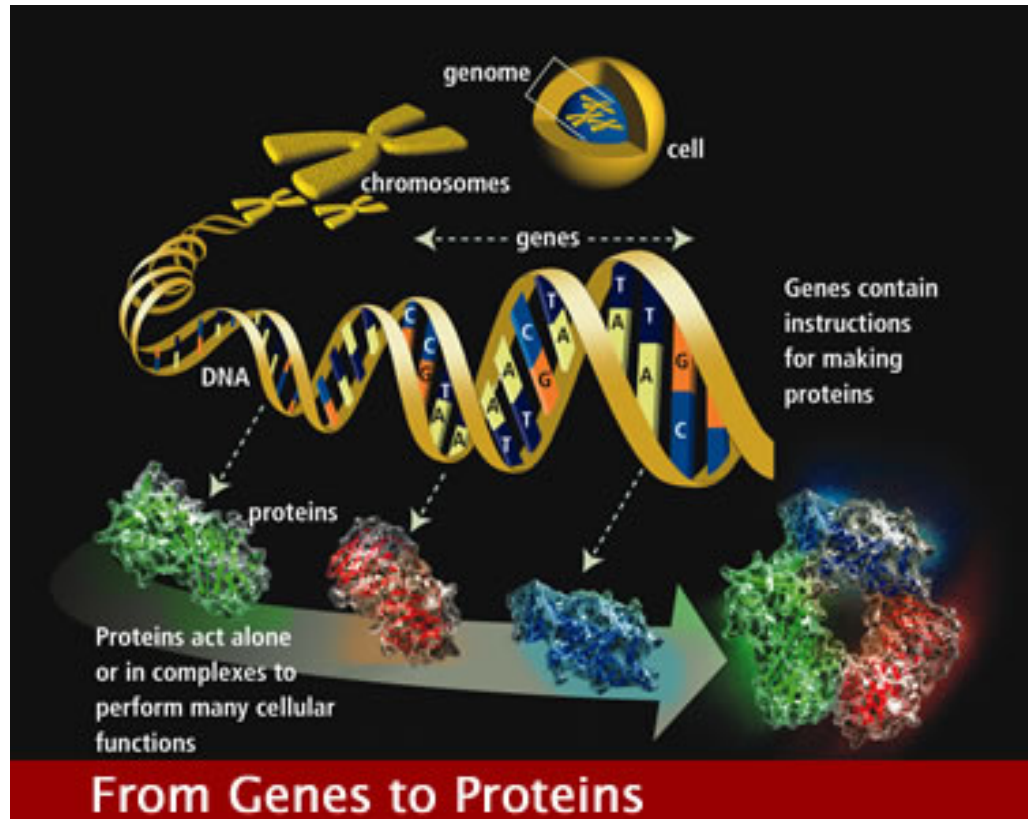
Paired end reads, 50-300bp

Nanopore sequencing



2Mb Read Lengths, genome assembly, direct RNA sequencing...

Organising Information



(http://www.ornl.gov/sci/techresources/Human_Genome/home.shtml)

The screenshot displays the UCSC Genome Browser interface for the Human Feb. 2009 (GRCh37/hg19) Assembly. The main view shows a genomic track for chromosome 16, specifically around position 67,227 bp. The track displays various genomic features, including genes (UCSC Genes, RefSeq Genes), transcripts, and other annotations. The interface includes navigation controls at the top (move, zoom in, zoom out) and a search bar. The bottom section contains controls for track search, default tracks, and track hubs, along with a 'Mapping and Sequencing' section.

Protein Domain and Structure Information

InterPro Domains: [Graphical view of domain structure](#)

[IPR002610](#) - Peptidase_S54_rhomboid

[IPR022764](#) - Peptidase_S54_rhomboid_dom

[IPR022241](#) - Rhomboid_SP

Pfam Domains:

[PF01694](#) - Rhomboid family

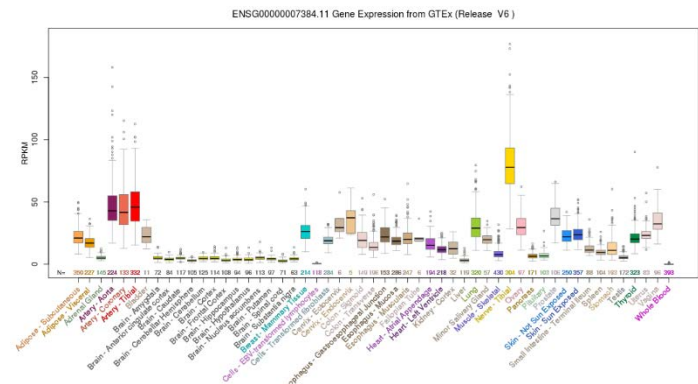
[PF12595](#) - Rhomboid serine protease

ModBase Predicted Comparative 3D Structure on [Q96CC6](#)

Front

Top

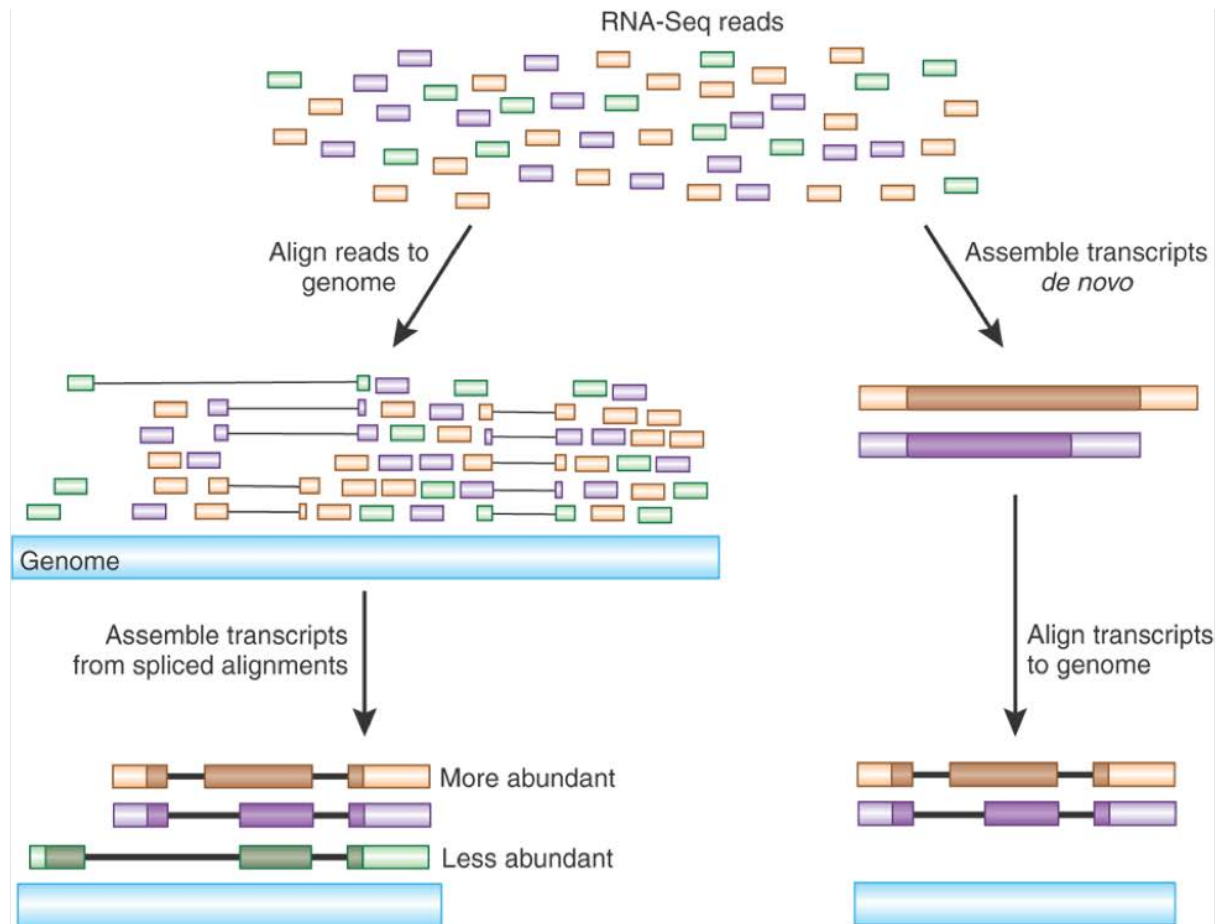
Side



Post Genome (19 years on)

- What do all the genes do?
 - How do they interact?
 - How to cells specialise?
- Junk DNA – is not junk after all...
 - 2% Genome contains genes
 - Between 80% (ENCODE) and 25% (Graur et al, 2017) genome seems to have function, usually regulation

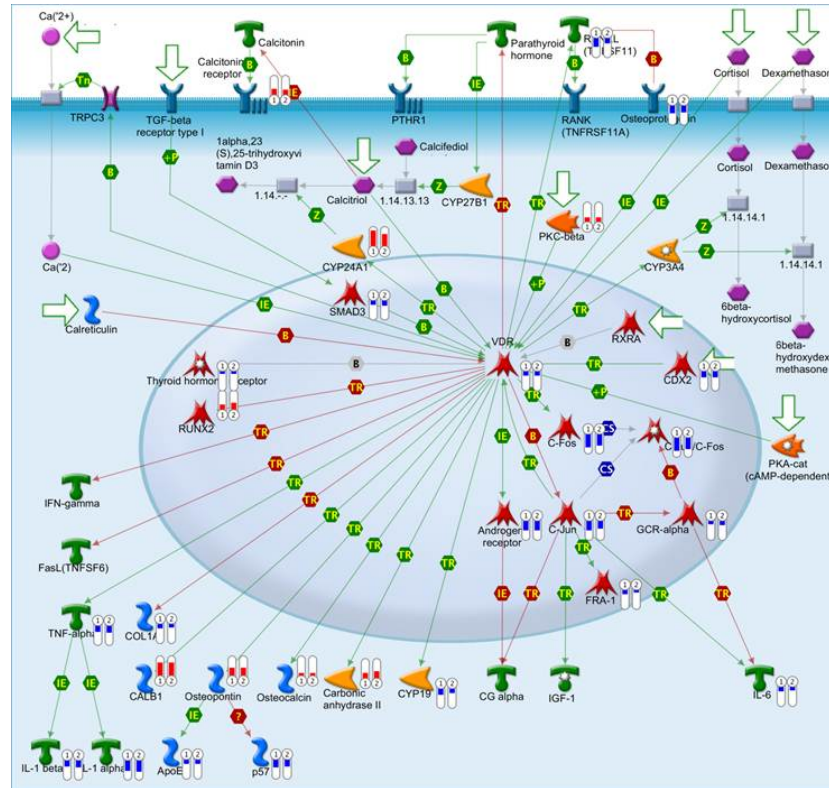
Expression Analysis (RNA-Seq)



Tools for Alignment

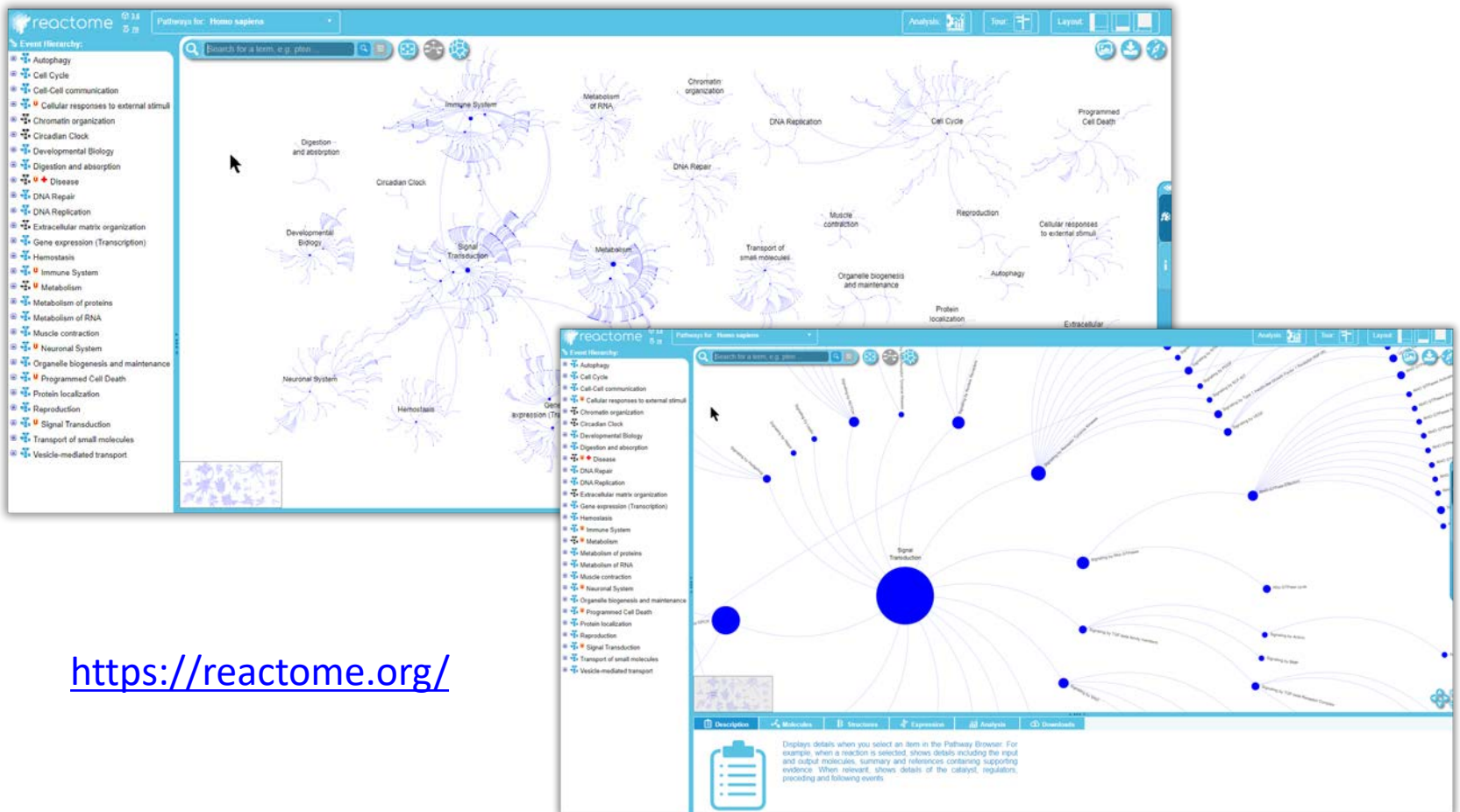
- **STAR (Spliced Transcripts Alignment to a Reference)**. Fast, but uses a lot of memory.
 - Alexander Dobin, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, Thomas R. Gingeras, STAR: ultrafast universal RNA-seq aligner, *Bioinformatics*, Volume 29, Issue 1, January 2013, Pages 15–21
- Normalisation and quantification of read counts use:
 - **edgeR**
 - edgeR: a Bioconductor package for differential expression analysis of digital gene expression data." *Bioinformatics*, **26**(1), 139-140)or DESeq2
 - **DESeq2**
 - Love MI, Huber W, Anders S (2014). "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2." *Genome Biology*, **15**, 550
- **Salmon** Very fast and does quantification. Uses *quasi-mapping* but no alignments to visualise.
 - Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., & Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods*.

Functional Annotation




- Metacore
- Ingenuity

Reactome



<https://reactome.org/>

Metascape



Step 1

Upload File Format

Single List

Multiple List

Or paste a gene list

Test Upload

single list

3 gene lists

Test Identifiers

Gene Symbol

RefSeq

Entrez Gene ID

Submit Cancel

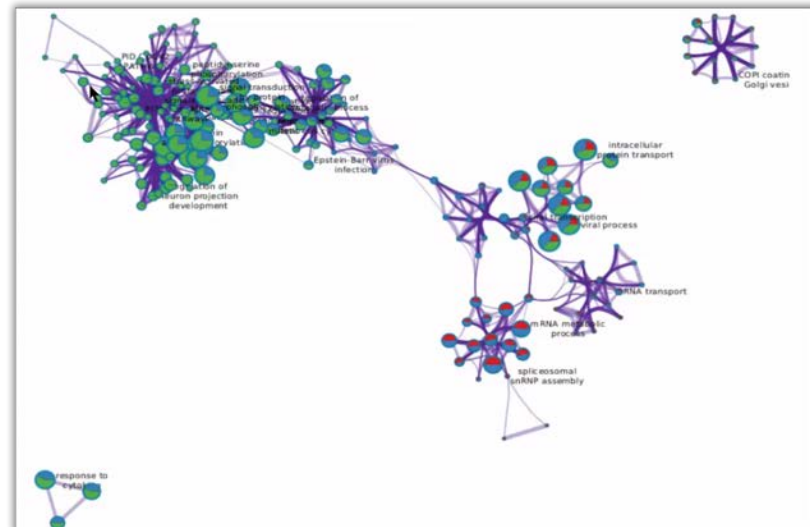
Step 2

Express Analysis Custom Analysis

working on Gene Enrichment

20

GroupID	Category	Term	Description	LogP	Inform	InList	Genes	Symbols
1_Summary	GO Biological Processes	GO:0016032	viral process	-18.851	49/771		156,527,2033, ADRBK1,ATP	
1_Member	GO Biological Processes	GO:0016032	viral process	-18.851	49/771		156,527,2033, ADRBK1,ATP	
1_Member	GO Biological Processes	GO:0044764	multi-organism cellular process	-18.549	49/784		156,527,2033, ADRBK1,ATP	
1_Member	GO Biological Processes	GO:0044419	interspecies interaction between organisms	-18.086	50/838		156,527,1536, ADRBK1,ATP	
1_Member	GO Biological Processes	GO:0044403	symbiosis, encompassing mutualism through parasitism	-18.086	50/838		156,527,1536, ADRBK1,ATP	
1_Member	GO Biological Processes	GO:0051351	positive regulation of ligase activity	-12.291	16/101		5347,5682,561 PLK1,PSMA1,	
1_Member	GO Biological Processes	GO:0051437	positive regulation of ubiquitin-protein ligase activity invo	-12.099	14/72		5347,5682,561 PLK1,PSMA1,	
1_Member	GO Biological Processes	GO:0000202	protein polyubiquitination	-11.468	17/212		331,4734,568: XIAP,NEDD4,	
1_Member	GO Biological Processes	GO:0051436	positive regulation of ubiquitin-protein transferase activity	-11.468	17/212		5347,5682,561 PLK1,PSMA1,	
1_Member	GO Biological Processes	GO:2000060	positive regulation of protein ubiquitination involved in ut-	-10.921	14/87		5347,5682,561 PLK1,PSMA1,	
1_Member	KEGG Pathway	hsa03050	Proteasome	-10.903	11/44		5682,5683,561 PSMA1,PSM	
1_Member	GO Biological Processes	GO:0051436	negative regulation of ubiquitin-protein ligase activity inv	-10.714	14/90		5347,5682,561 PLK1,PSMA1,	
1_Member	GO Biological Processes	GO:2000058	regulation of protein ubiquitination involved in ubiquitin-	-10.322	14/96		5347,5682,561 PLK1,PSMA1,	
1_Member	GO Biological Processes	GO:0051340	regulation of ligase activity	-10.312	16/135		5347,5682,561 PLK1,PSMA1,	
1_Member	GO Biological Processes	GO:0051444	negative regulation of ubiquitin-protein transferase activi	-10.260	14/97		5347,5682,561 PLK1,PSMA1,	
1_Member	GO Biological Processes	GO:0051352	negative regulation of ligase activity	-10.138	14/98		5347,5682,561 PLK1,PSMA1,	
1_Member	GO Biological Processes	GO:0051439	regulation of ubiquitin-protein ligase activity involved in n	-10.137	14/99		5347,5682,561 PLK1,PSMA1,	
1_Member	GO Biological Processes	GO:0032446	protein modification by small protein conjugation	-9.776	37/821		331,4734,492: XIAP,PLK1,P	
1_Member	GO Biological Processes	GO:0031398	positive regulation of protein ubiquitination	-9.579	17/174		331,5347,568: XIAP,PLK1,P	
1_Member	GO Biological Processes	GO:0031145	anaphase-promoting complex-dependent proteasomal ub	-9.565	14/109		5347,5682,561 PLK1,PSMA1,	
1_Member	GO Biological Processes	GO:0002479	antigen processing and presentation of exogenous peptide	-9.412	12/75		1536,5682,561 CYBB,PSMA1,	
1_Member	GO Biological Processes	GO:0051438	regulation of ubiquitin-protein transferase activity	-9.302	15/135		5347,5682,561 PLK1,PSMA1,	
1_Member	GO Biological Processes	GO:0042590	antigen processing and presentation of exogenous peptide	-9.140	12/79		1536,5682,561 CYBB,PSMA1,	
1_Member	GO Biological Processes	GO:1903322	positive regulation of protein modification by small protein	-9.091	17/187		331,5347,568: XIAP,PLK1,P	
1_Member	GO Biological Processes	GO:0044265	cellular macromolecule catabolic process	-9.055	38/913		3146,3837,411: HMGB1,KPNI	
1_Member	GO Biological Processes	GO:0006521	regulation of cellular amino acid metabolic process	-9.021	11/64		5682,5683,561 PSMA1,PSM	
1_Member	GO Biological Processes	GO:0042787	protein ubiquitination involved in ubiquitin-dependent pr	-8.962	16/166		4734,5347,561 NEDD4,PLK1,	
1_Member	GO Biological Processes	GO:0006977	DNA damage response, signal transduction by p53 class me	-8.872	11/66		5682,5683,561 PSMA1,PSM	



<http://metascape.org/>

Gene Expression Omnibus (GEO)

The screenshot displays the NCBI Gene Expression Omnibus (GEO) repository browser. The top navigation bar includes the NCBI logo and the GEO logo. Below the navigation bar, there are tabs for Series, Samples, Platforms, and DataSets. The main content area shows a list of samples with columns for Accession, Title, Sample type, Organism, Platform, Series, Supplementary, Contact, and Release date. A search bar is located at the top right of the sample list, showing 1,163,446 samples. A detailed view of sample GSM952626 is shown in a pop-up window, providing information about the sample's status, title, sample type, source name, organism, characteristics, growth protocol, extracted molecule, extraction protocol, label, and label protocol.

NCBI | **GEO** | **Gene Expression Omnibus**

Series | **Samples** | Platforms | DataSets | Summary | Advanced search

Search 1,163,446 samples Export

Page 1 of 36,173 Page size 20

Accession	Title	Sample type	Organism(s)	Ch	Platform	Series	Supplementary	Contact	Release date
GSM952626	SPC/cRaf mouse dysplasia 65.1 male 6 months	RNA	Mus musculus	1	GPL6096	GSE38948	CEL CHP	Kishor Babu Londhe	Jun 23, 2014
GSM952627	SPC/cRaf mouse dysplasia 67.3_71.5 male 5 months	RNA	Mus musculus	1	GPL6096	GSE38948	CEL CHP	Kishor Babu Londhe	Jun 23, 2014
GSM952628	SPC/cRaf mouse dysplasia 73.5 male 6 months	RNA	Mus musculus	1	GPL6096	GSE38948	CEL CHP	Kishor Babu Londhe	Jun 23, 2014
GSM952629	SPC/cRaf mouse dysplasia 73.7 male 6 months	RNA	Mus musculus	1	GPL6096	GSE38948	CEL CHP	Kishor Babu Londhe	Jun 23, 2014
GSM952630	non-transgenic mouse 65.0 male 7 months	RNA	Mus musculus	1	GPL6096	GSE38948	CEL CHP	Kishor Babu Londhe	Jun 23, 2014
GSM952631	non-transgenic mouse 67.5 female 7 months	RNA	Mus musculus	1	GPL6096	GSE38948	CEL CHP	Kishor Babu Londhe	Jun 23, 2014
GSM952632	non-transgenic mouse 92.7 female 11 months	RNA	Mus musculus	1	GPL6096	GSE38948	CEL CHP	Kishor Babu Londhe	Jun 23, 2014
GSM952633	non-transgenic mouse 92.7 female 11 months	RNA	Mus musculus	1	GPL6096	GSE38948	CEL CHP	Kishor Babu Londhe	Jun 23, 2014
GSM952634	non-transgenic mouse 92.7 female 11 months	RNA	Mus musculus	1	GPL6096	GSE38948	CEL CHP	Kishor Babu Londhe	Jun 23, 2014
GSM952635	SPC/cRaf mouse dysplasia 65.1 male 6 months	RNA	Mus musculus	1	GPL6096	GSE38948	CEL CHP	Kishor Babu Londhe	Jun 23, 2014
GSM952636	SPC/cRaf mouse dysplasia 67.3_71.5 male 5 months	RNA	Mus musculus	1	GPL6096	GSE38948	CEL CHP	Kishor Babu Londhe	Jun 23, 2014
GSM952637	SPC/cRaf mouse dysplasia 73.5 male 6 months	RNA	Mus musculus	1	GPL6096	GSE38948	CEL CHP	Kishor Babu Londhe	Jun 23, 2014
GSM952638	SPC/cRaf mouse dysplasia 73.7 male 6 months	RNA	Mus musculus	1	GPL6096	GSE38948	CEL CHP	Kishor Babu Londhe	Jun 23, 2014
GSM952639	non-transgenic mouse 65.0 male 7 months	RNA	Mus musculus	1	GPL6096	GSE38948	CEL CHP	Kishor Babu Londhe	Jun 23, 2014
GSM952640	non-transgenic mouse 67.5 female 7 months	RNA	Mus musculus	1	GPL6096	GSE38948	CEL CHP	Kishor Babu Londhe	Jun 23, 2014
GSM952641	non-transgenic mouse 92.7 female 11 months	RNA	Mus musculus	1	GPL6096	GSE38948	CEL CHP	Kishor Babu Londhe	Jun 23, 2014
GSM952642	non-transgenic mouse 92.7 female 11 months	RNA	Mus musculus	1	GPL6096	GSE38948	CEL CHP	Kishor Babu Londhe	Jun 23, 2014
GSM1314708	ECFC_L1_1	RNA	Human	1	GPL6096	GSE34416	SRA Experiment	Terri DiMaio	Jun 23, 2014
GSM1314709	ECFC_L1_2	RNA	Human	1	GPL6096	GSE34416	SRA Experiment	Terri DiMaio	Jun 23, 2014
GSM1314710	ECFC_L1_3	RNA	Human	1	GPL6096	GSE34416	SRA Experiment	Terri DiMaio	Jun 23, 2014

NCBI | **GEO** | **Gene Expression Omnibus**

HOME | SEARCH | SITE MAP

NCBI > GEO > Accession Display

GEO help: Mouse over screen elements for information.

Scope: Self Format: HTML Amount: Quick GEO accession: GSM952626

Sample GSM952626 Query DataSets for GSM952626

Status Public on Jun 23, 2014

Title SPC/cRaf mouse dysplasia 65.1 male 6 months

Sample type RNA

Source name dysplasia male

Organism Mus musculus

Characteristics age: 6 months
genotype: SPC/cRaf transgenic
tissue: lung dysplastic lesion
Sex: male

Growth protocol Four samples each for dysplastic and adenocarcinoma stages and 5 samples from healthy non-transgenic lungs were selected for laser micro-dissection. Lung tissue slices of 10um were prepared using a cryomicrotome (MICROM GmbH, Walldorf, Germany) and fixed over PEN membrane slide (Zeiss GmbH) and stained with Haematoxylin. The desired cells either dysplastic or transgenic (microscopically unaltered, normal) or adenocarcinoma or healthy non-transgenic alveolar cells were laser microdissected and collected in an adhesive cap using the LMPC (Laser Micro-dissection Pressure Catapulting) system.

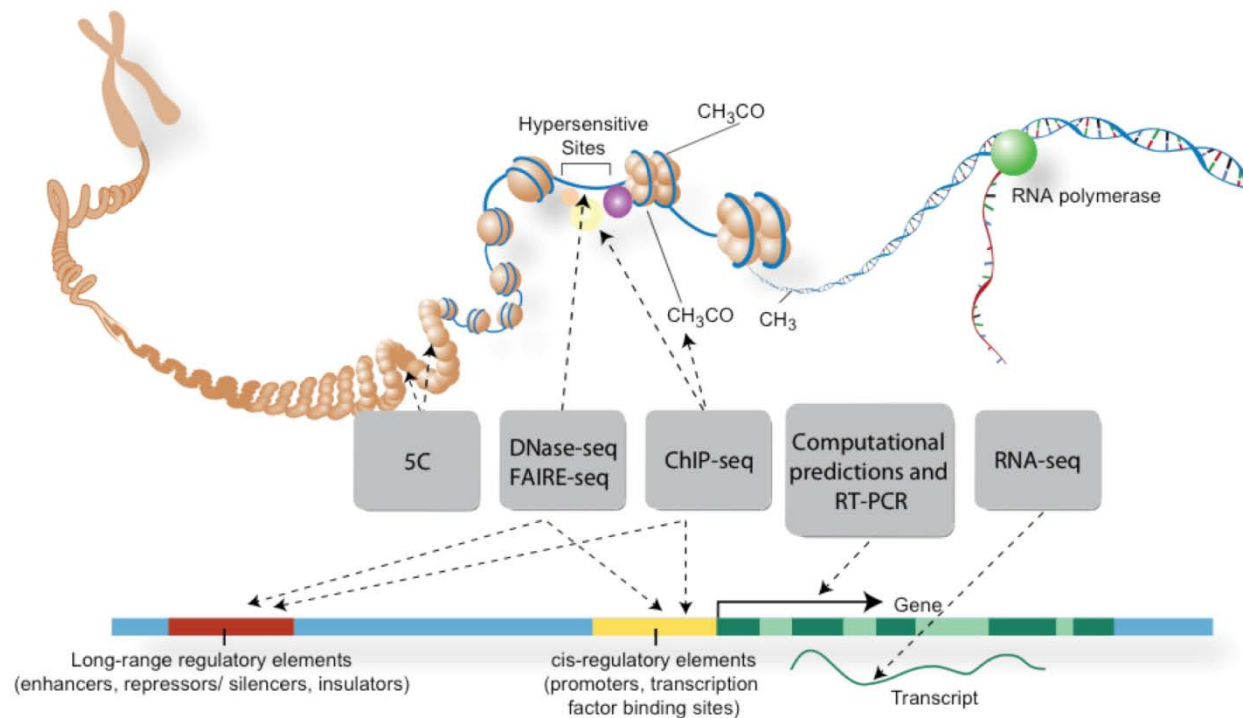
Extracted molecule total RNA

Extraction protocol Four samples each for dysplastic and adenocarcinoma stages and 5 samples from healthy non-transgenic lungs were selected for laser micro-dissection. Lung tissue slices of 10um were prepared using a cryomicrotome (MICROM GmbH, Walldorf, Germany) and fixed over PEN membrane slide (Zeiss GmbH) and stained with Haematoxylin. The desired cells either dysplastic or transgenic (microscopically unaltered, normal) or adenocarcinoma or healthy non-transgenic alveolar cells were laser microdissected and collected in an adhesive cap using the LMPC (Laser Micro-dissection Pressure Catapulting) system.

Label biotin

Label protocol rRNA reduction was done using Ribominus kit (Life technologies, Invitrogen, Carlsbad, California). Single-stranded cDNA was generated from the amplified cRNA with the WT cDNA Synthesis Kit (Affymetrix) and then fragmented and labeled with the WT Terminal Labeling Kit (Affymetrix).

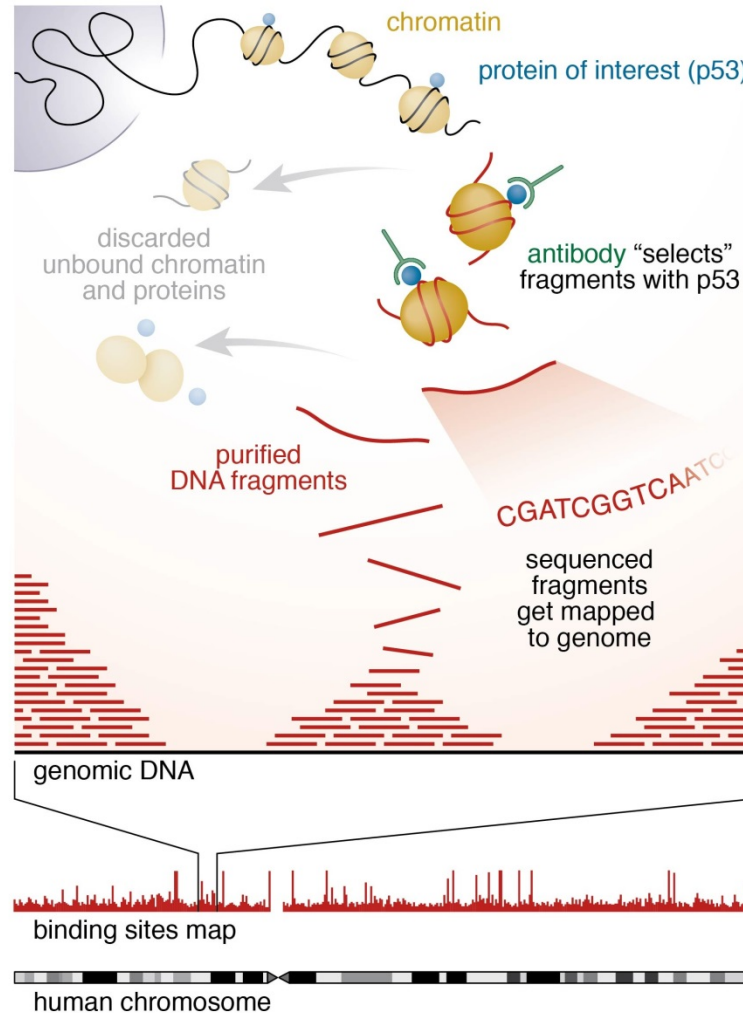
ENCODE (Encyclopedia of DNA Elements)



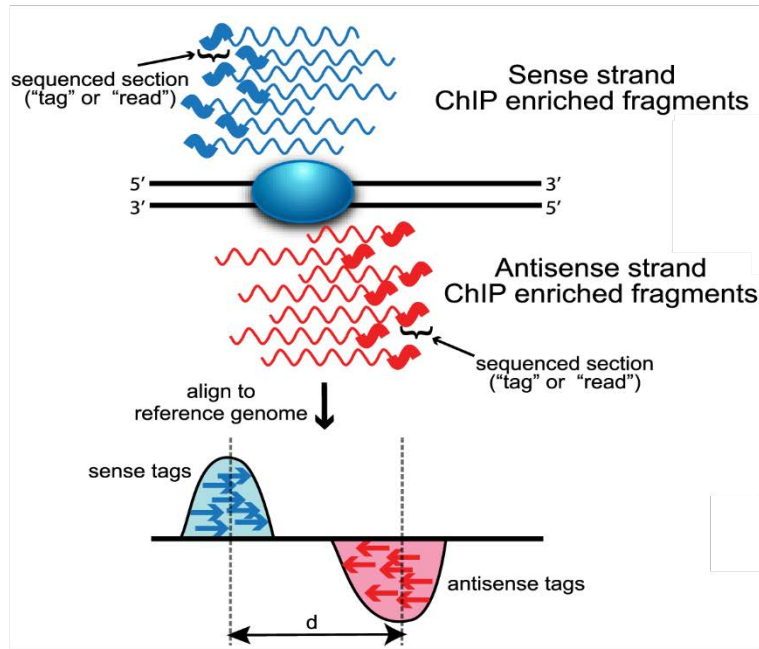
<http://genome.ucsc.edu/ENCODE/>

What Controls Expression?

ChIP-Seq



Tools for ChIP-Seq



1. Align using Bowtie
2. Peak call using Model-based Analysis of ChIP-seq (MACS)
3. Look for motif enrichment using HOMER
4. Functional annotation using GREAT

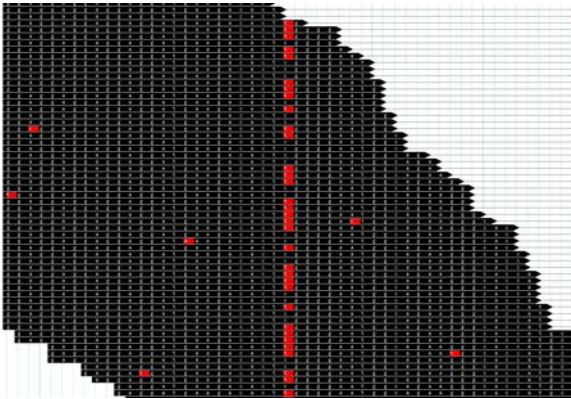
LANCETRON

<https://lanceotron.molbiol.ox.ac.uk/>

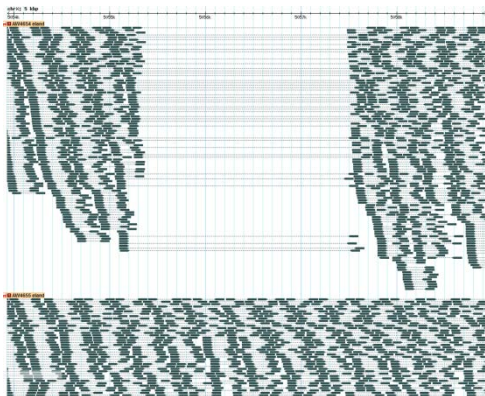
1. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 10:R25.
2. Zhang, Y., Liu, T., Meyer, C.A. *et al.* Model-based Analysis of ChIP-Seq (MACS). *Genome Biol* 9, R137 (2008) doi:10.1186/gb-2008-9-9-r137
3. Heinz S, Benner C, Spann N, Bertolino E *et al.* Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Mol Cell* 2010 May 28;38(4):576-589.
4. Cory Y McLean, Dave Bristor, Michael Hiller, Shoa L Clarke, Bruce T Schaar, Craig B Lowe, Aaron M Wenger, and Gill Bejerano. "GREAT improves functional interpretation of *cis*-regulatory regions". *Nat. Biotechnol.* 28(5):495-501, 2010

DNA Mutations

Single base mutation



Insertion



FAULTY GENE

The Single Nucleotide Polymorphism database (**dbSNP**) is a public-domain archive for a broad collection of simple genetic polymorphisms.

(<http://www.ncbi.nlm.nih.gov/SNP/>)

Tools for variant calling

SAMTOOLS

A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Li H Bioinformatics. 2011 Nov 1;27(21):2987-93.

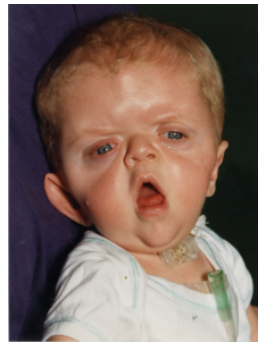
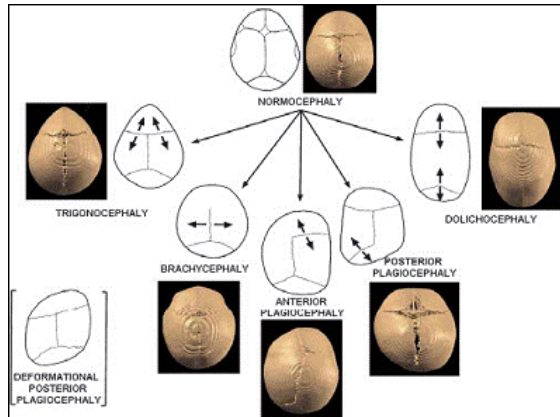


The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA, 2010 *GENOME RESEARCH* 20:1297-303



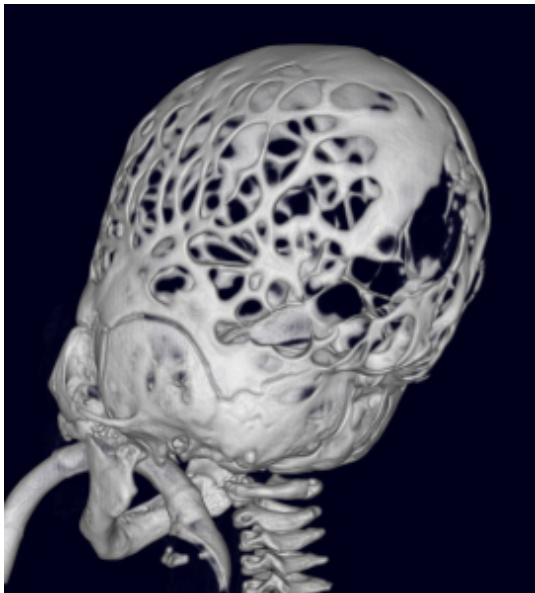
A unified haplotype-based method for accurate and comprehensive variant calling
Daniel P Cooke, David C Wedge, Gerton Lunter
bioRxiv 456103; doi: <https://doi.org/10.1101/456103>

Craniosynostosis



Andrew Wilkie, WIMM

Craniosynostosis



THE TIMES THE SUNDAY TIMES MY TIMES+ MY ACCOUNT Welcome Dr Simon McGowan

THE TIMES Genetics

News Opinion Business Money Sport Life Arts Puzzles Papers

Gene isolated as girl becomes first in Britain to have entire DNA code read

Article Graphic: the genome revolution

A photograph of a woman with dark hair, wearing a patterned top, sitting on a floral-patterned sofa. She is holding a young child with blonde hair, who is wearing a white shirt and blue pants. The child is holding a small white stuffed animal. The woman is looking at the child with a smile.

Mark Henderson Science Editor
August 3 2011 12:01AM

A four-year-old girl has become the first person in Britain to have her entire genetic code read to identify the cause of a disease, in a landmark development that illustrates how personal genetics is changing healthcare.

Katie Warner, who has a cranio-facial condition, with her mother Marie Mary Turner for The Times

Post a comment

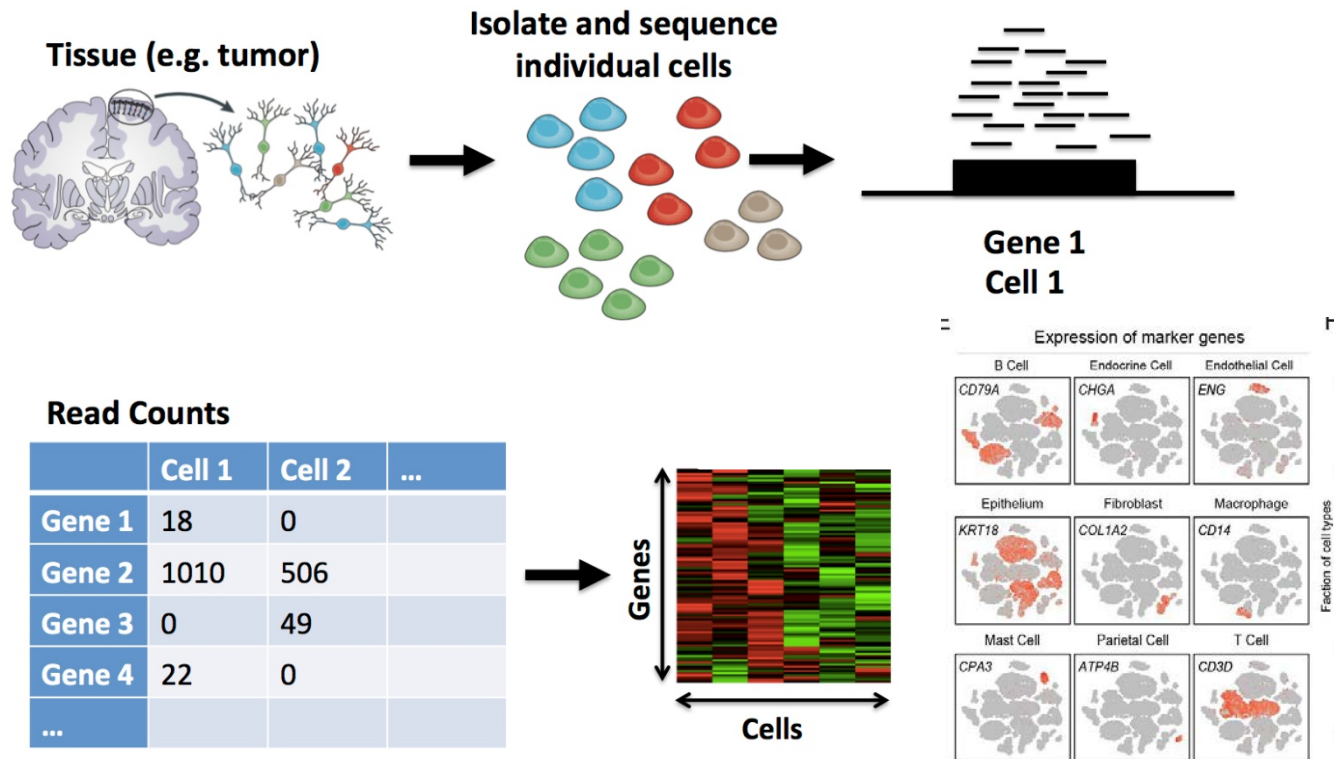
Recommend (4)

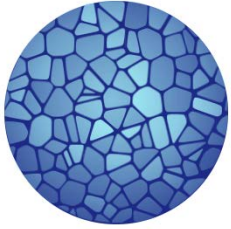


- Genomics England 100,000 Genomes Project (GEL)
- 100,000 patients with rare inherited disease, common cancers and pathogens from the NHS Whole Genome Sequencing
- <http://www.genomicsengland.co.uk/>
- Secure environment to do analysis

Single Cell Sequencing

Single-cell RNA-Seq (scRNA-Seq)





HUMAN
CELL
ATLAS

cellxgene

cellxgene tabula-muris

- ☒ cell_ontology_class >
- ☒ clusters_from_manuscript >
- ☒ clusters_leiden >
- ☒ clusters_louvain >
- ☒ free_annotation >
- ☒ mouse_id >
- ☒ mouse_sex >
- ☒ plate_barcode >
- ☒ subtissue >
- ☒ tissue >

Create new category

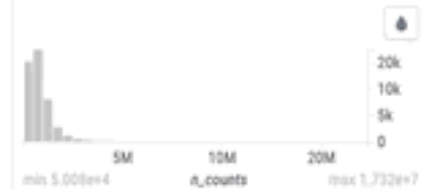
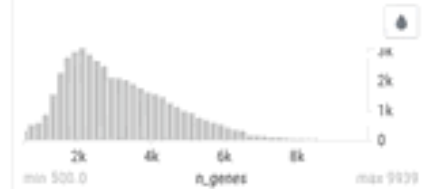


1: 0 cells 2: 0 cells **cell**



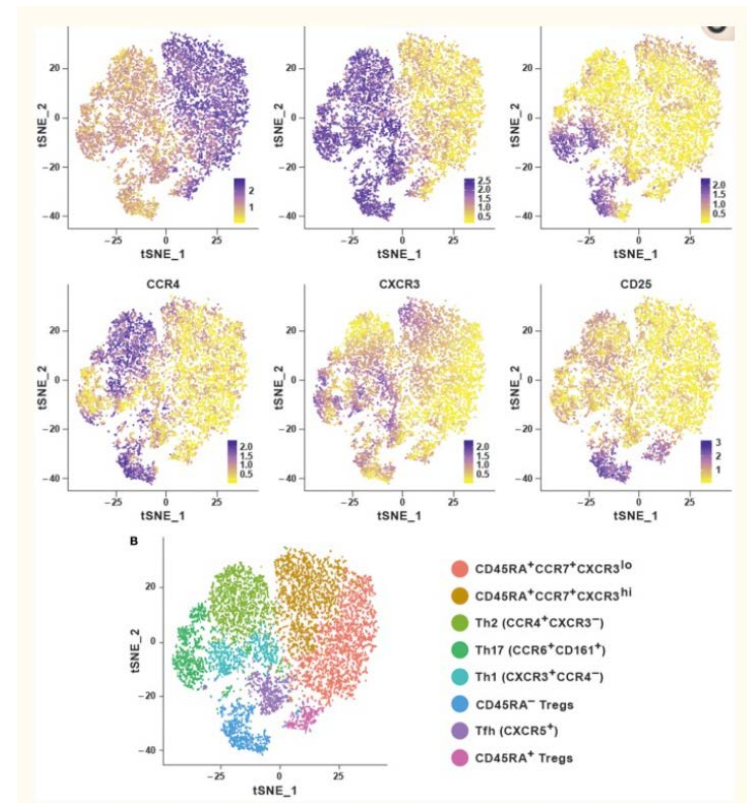
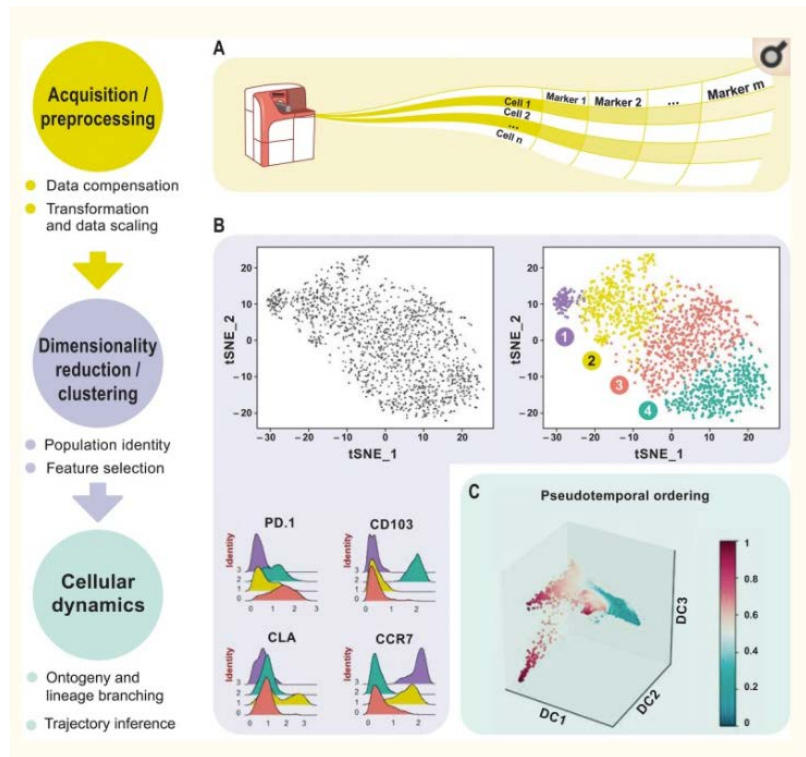
Autosuggest genes Bulk add genes

Search... **Add gene**



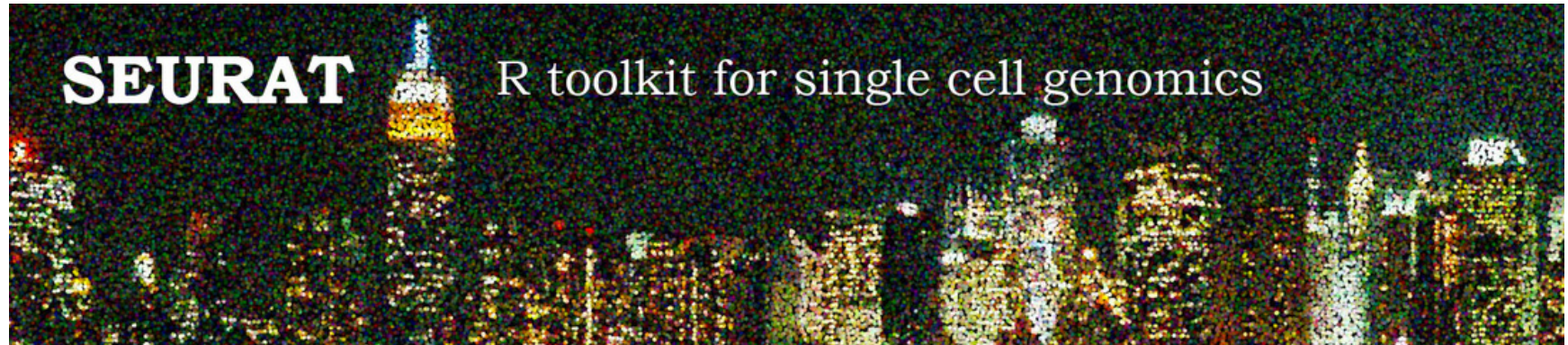
All saved

CyTOF – Single Cell Proteomics



Palit, Subarna, Christoph Heuser, Gustavo P. de Almeida, Fabian J. Theis, and Christina E. Zielinski. 2019. "Meeting the Challenges of High-Dimensional Single-Cell Data Analysis in Immunology." *Frontiers in Immunology* 10 (July): 1515.

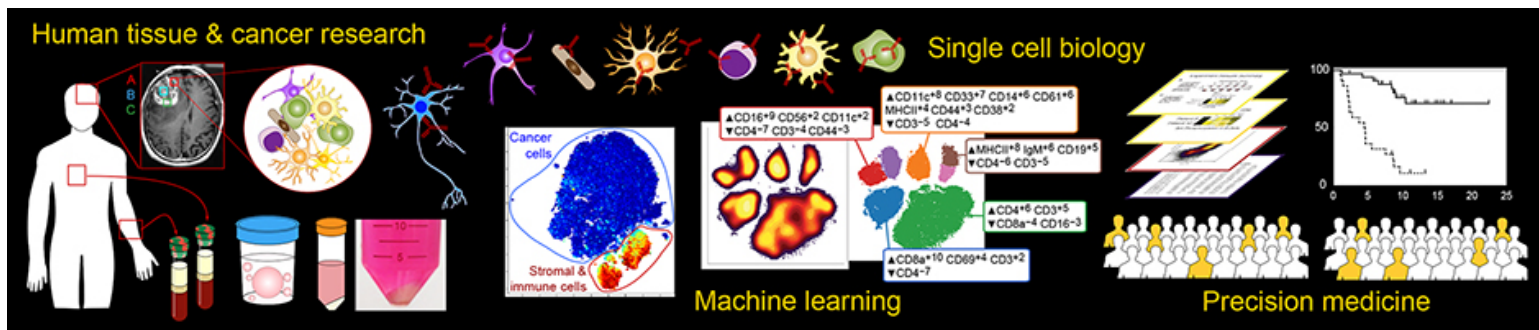
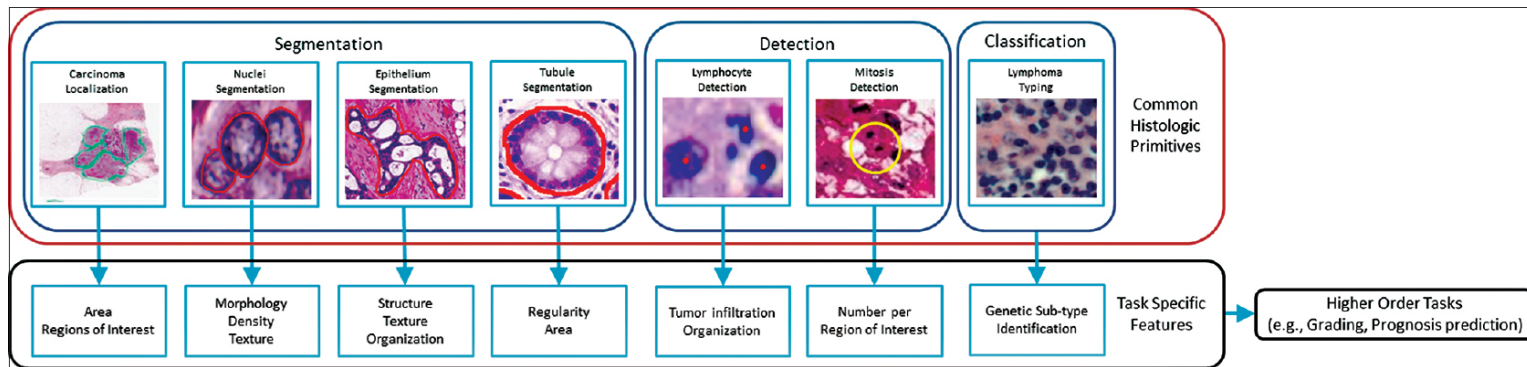
Single Cell Analysis



Overview of some of the established single-cell analysis methods.

Class	Methods	Description
Linear dimensionality reduction	PCA	Cannot account for the smooth nature of single-cell data
Non-linear dimensionality reduction	t-SNE	More intuitive representation of high-dimensional data on a lower manifold
	UMAP	Scales better and improves global structure of the data compared to t-SNE(see Box 1)
	HSNE	Scales better than conventional t-SNE(see Box 1)
	Diffusion maps	Explores continuity through progression of cell differentiation
Clustering methods; single-cell resolution is lost	SPADE	Hierarchical branched tree representation (see Box 2)
	FlowSOM	Self-organizing maps trained to detect cell populations (see Box 3)

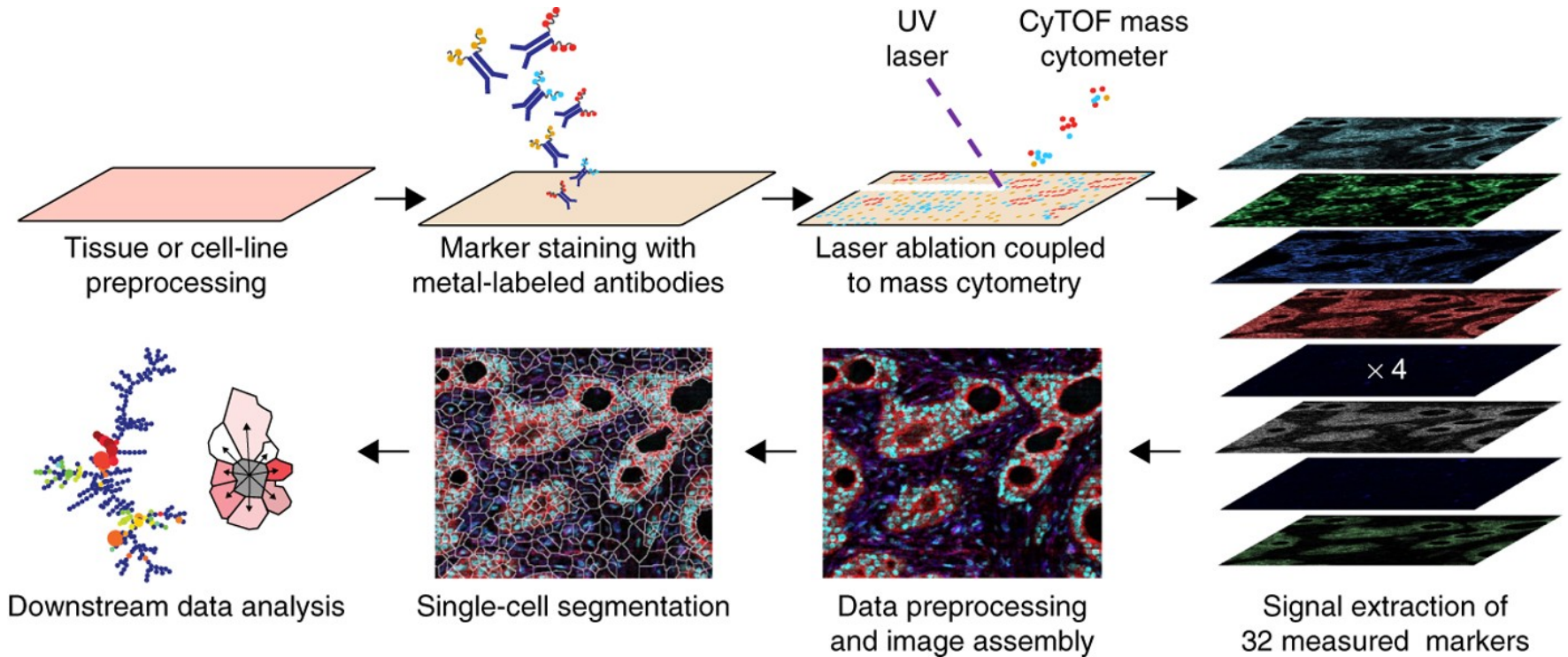
Machine Learning



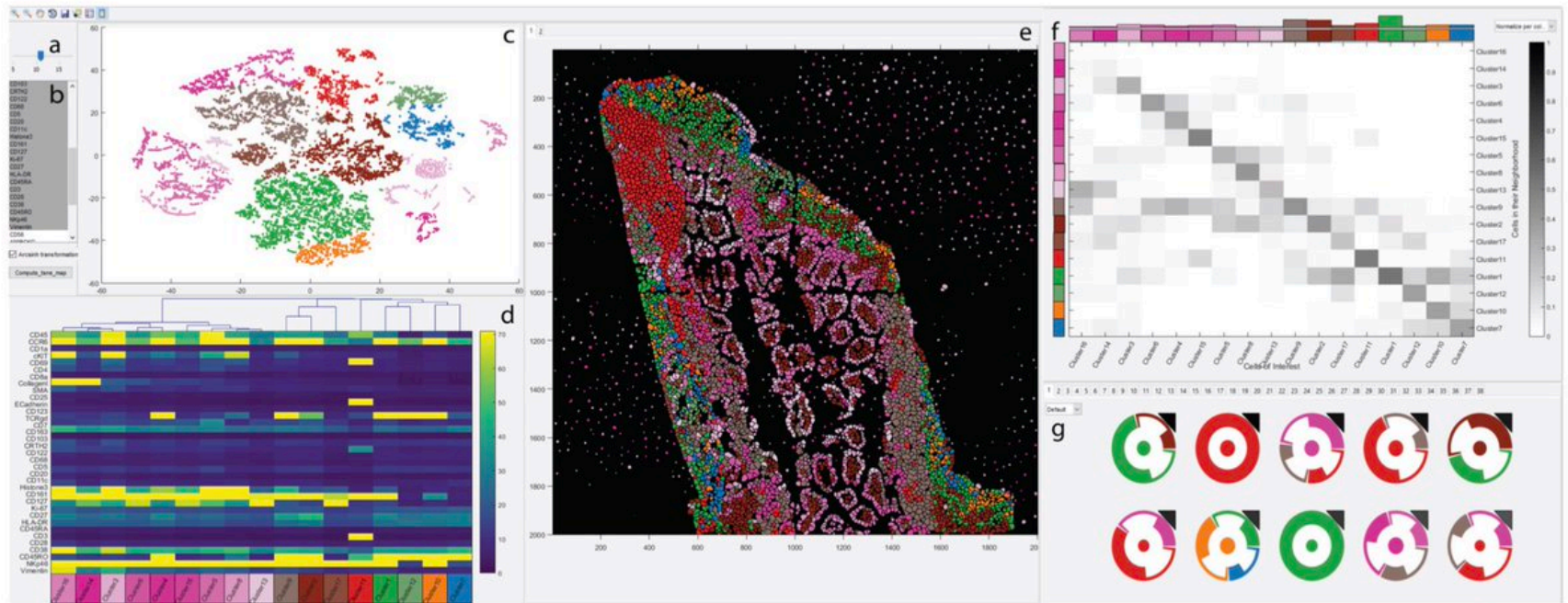
<https://my.vanderbilt.edu/irishlab/>

Also being applied increasing to all data types e.g. health records, DNA sequences

Spatial Proteomics



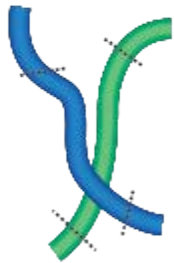
ImaCyte : Analyse Cell Microenvironment



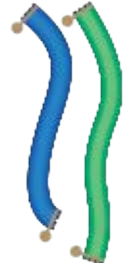
Somarakis, Antonios, Vincent Van Unen, Frits Koning, Boudewijn P. F. Lelieveldt, and Thomas Holtt. 2019. "ImaCytE: Visual Exploration of Cellular Microenvironments for Imaging Mass Cytometry Data." *IEEE Transactions on Visualization and Computer Graphics*, July. <https://doi.org/10.1109/TVCG.2019.2931299>.

Genome Modelling

1. Cut DNA strands with enzyme



2. Mark pieces for identification



3. Reseal DNA



4. Pull out sealed pieces



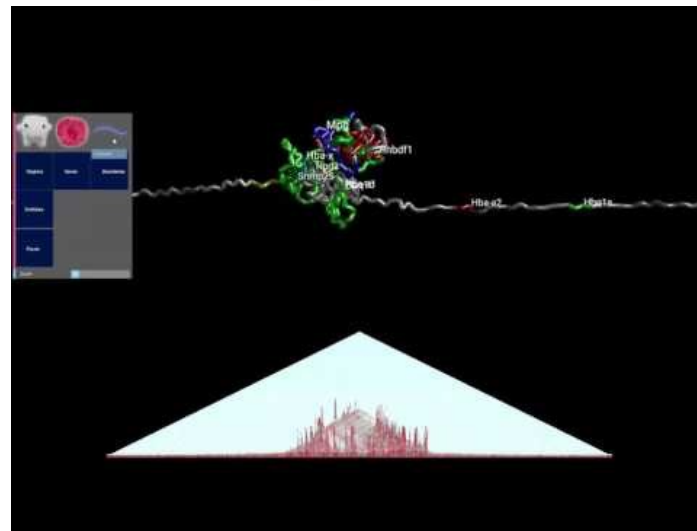
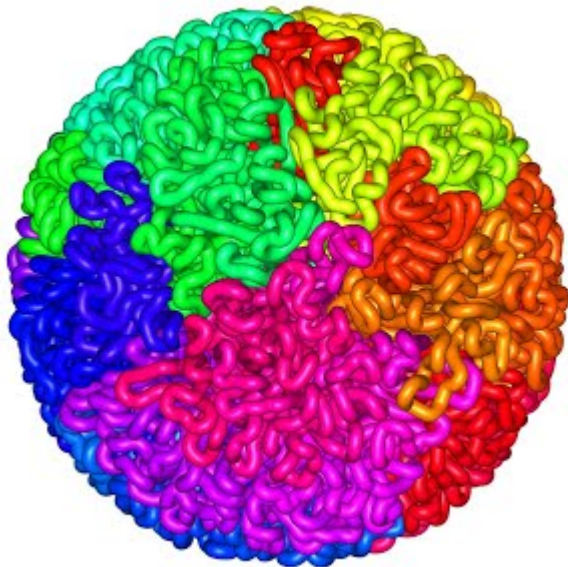
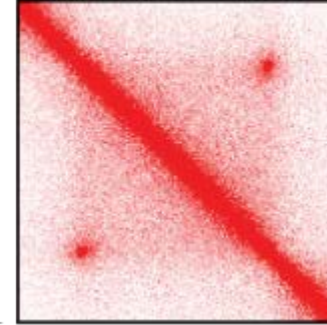
5. Read DNA



Chromosome 13

~1 million DNA letters

Chromosome 13

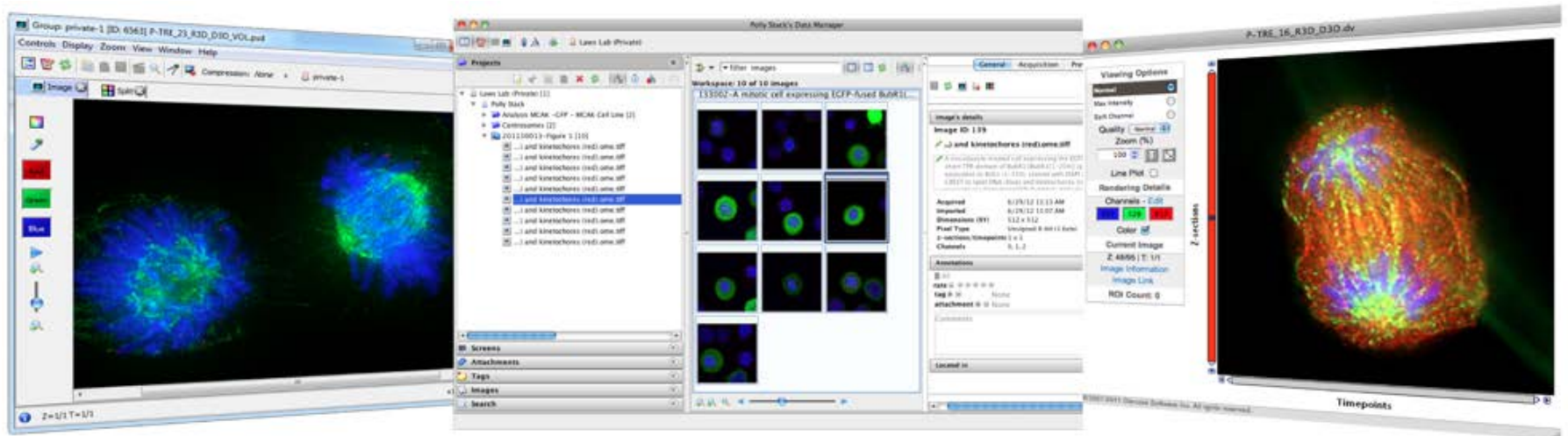


CSynth
Bio-Visualisation made interactive

Lieberman-Aiden, Erez, Nynke L. van Berkum, Louise Williams, Maxim Imakaev, Tobias Ragozy, Agnes Telling, Ido Amit, et al. 2009. "Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome." *Science* 326 (5950): 289–93.

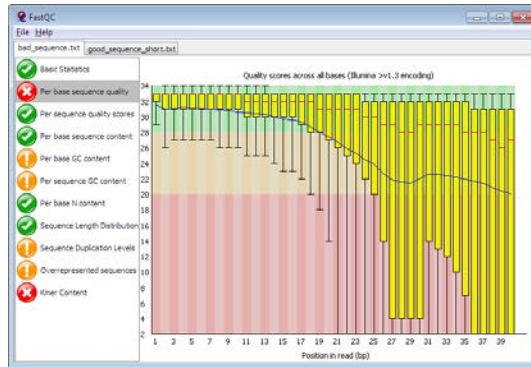
Todd, Stephen, Peter Todd, Simon J. McGowan, James R. Hughes, Yasutaka Kakui, Frederic Fol Leymarie, William Latham, and Stephen Taylor. 2020. "CSynth: An Interactive Modelling and Visualisation Tool for 3D Chromatin Structure." *Bioinformatics*, August. <https://doi.org/10.1093/bioinformatics/btaa757>.

OMERO Image Database

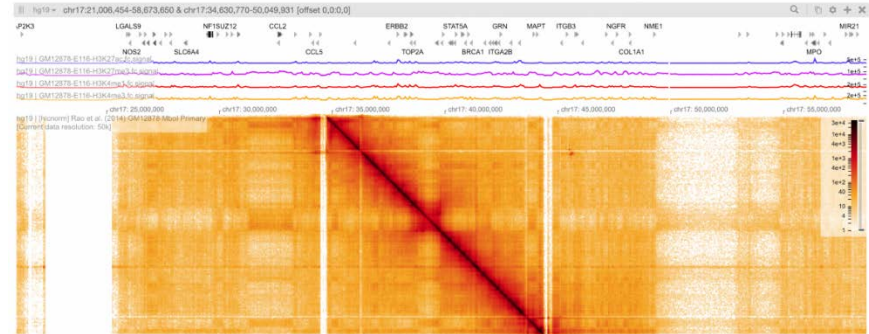


Visualisation

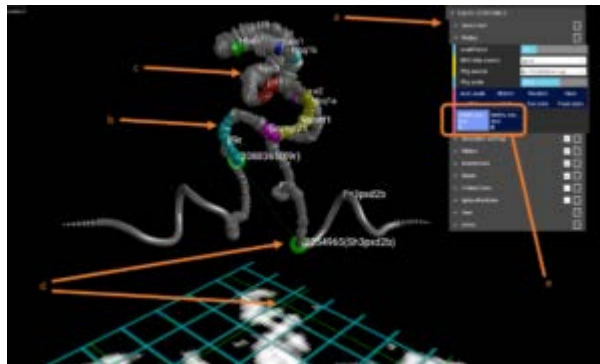
FASTQC



HiGlass



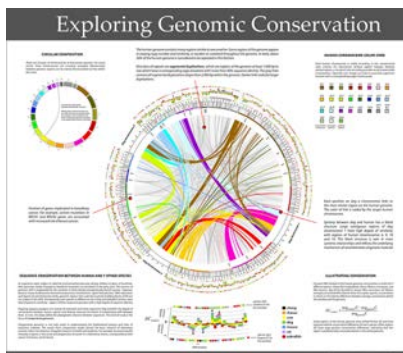
CSynth



Zegami



Circos

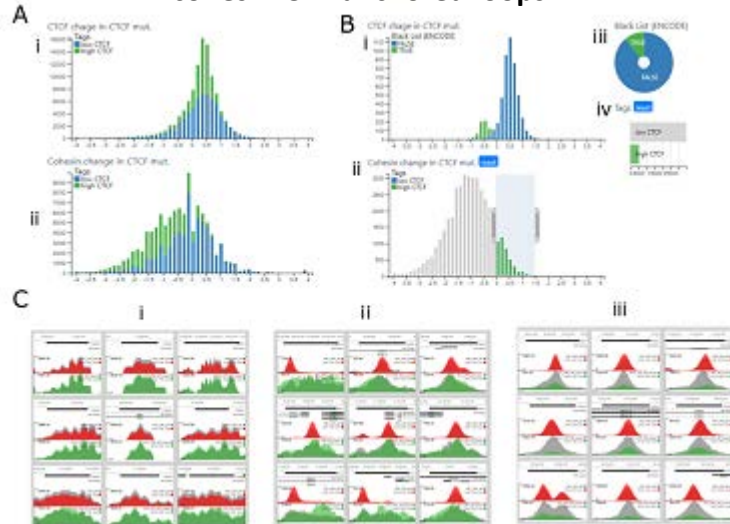


BabelVR

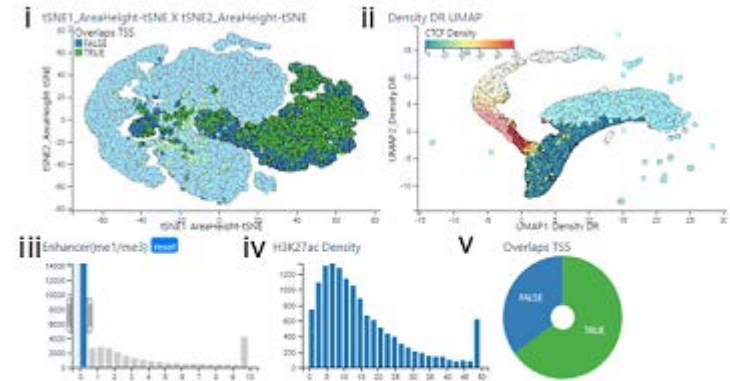


Multi Locus Viewer

The structural basis for
cohesin-CTF-anchored loops



Looking at ChIP-seq signals in enhancers and promoters



**Multi Locus View :An Extensible Web Based Tool for the
Analysis of Genomic Data**

Martin J Sergeant, Jim R Hughes, Lance Hentges, Damien J Downes, Stephen Taylor

doi: <https://doi.org/10.1101/2020.06.15.151837>

Under review “Nature Communications Biology”

Coding in R

- Very good for statistics
- Libraries
 - CRAN (12000 packages)
 - Bioconductor (1823 packages)
- Lots of methods for communicating results
- RStudio is nice graphical environment

Coding in Python

- Most popular language in bioinformatics (and probably data science)
- Used in industry and academic settings
- Very readable
- Great 'glue' for automation
- Lots of libraries for using matrices, machine learning, plotting etc
- <https://biopython.org/>

Python vs R

Parameter	R	Python
Objective	Data analysis and statistics	Deployment and production
Primary Users	Scholar and R&D	Programmers and developers
Flexibility	Easy to use available library	Easy to construct new models from scratch. I.e., matrix computation and optimization
Learning curve	Difficult at the beginning	Linear and smooth
Popularity of Programming Language. Percentage change	4.23% in 2018	21.69% in 2018
Average Salary	\$99,000	\$100,000
Integration	Run locally	Well-integrated with app
Task	Easy to get primary results	Good to deploy algorithm
Database size	Handle huge size	Handle huge size
IDE	Rstudio	Spyder, Ipython Notebook
Important Packages and library	tidyverse, ggplot2, caret, zoo	pandas, scipy, scikit-learn, TensorFlow, caret
Disadvantages	Slow High Learning curve Dependencies between library	Not as many libraries as R
Advantages	<ul style="list-style-type: none">• Graphs are made to talk. R makes it beautiful• Large catalog for data analysis• GitHub interface• RMarkdown• Shiny	<ul style="list-style-type: none">• Jupyter notebook: Notebooks help to share data with colleagues• Mathematical computation• Deployment• Code Readability• Speed• Function in Python

Learn both!

See review <https://www.guru99.com/r-vs-python.html>

CCB Training



INTRODUCTORY COURSES

Introductory short courses cover the Unix command line, programming in R and genomics workflows (ChIP-seq, RNAseq).

Find out more



OXFORD BIOMEDICAL DATA SCIENCE TRAINING PROGRAMME

This unique training programme consists of 10 week secondments, first building basic data science skills and then applying them to the analysis of your own biomedical data. **Find out more**

More information

- <https://www.imm.ox.ac.uk/research/units-and-centres/mrc-wimm-centre-for-computational-biology>
- Google “WIMM CCB”
- Tech Helpdesk : genmail@molbiol.ox.ac.uk
- General Questions : ccb@imm.ox.ac.uk