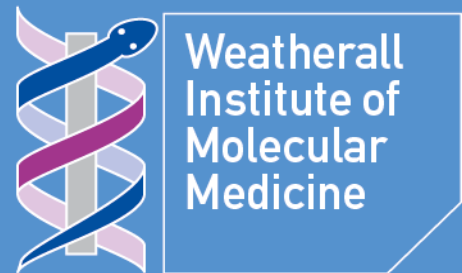


Methods and Techniques in Bioinformatics

(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](#) 

[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

From the bench to VR

Computational Biologist Stephen Taylor and his team were awarded an Innovation grant to develop a software package that allows researchers to use virtual reality for scientific research and public engagement.

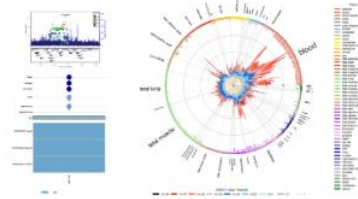
[Read more](#)



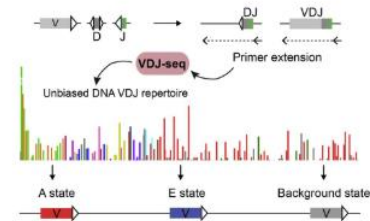
Research Groups



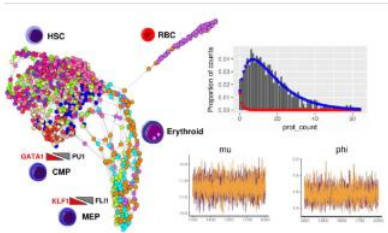
Hughes Group



Iotchkova Group: Statistical Genetics



Koohy Group: Machine Learning and Integrative Approaches in Immunology



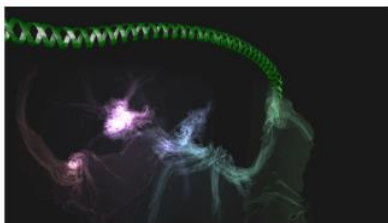
Morrissey Group: Quantitative biology of cell fate and tissue dynamics



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?

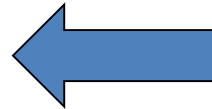
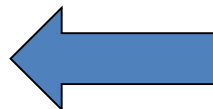
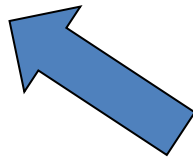
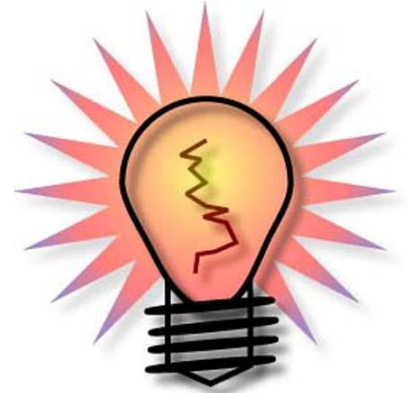
Experiment



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
 - HT Sequencing Experiments e.g. GEO
 - Structure e.g. PDB
 - Pathways e.g. KEGG, Metacore
 - Diseases e.g. OMIM
- 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

Help



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

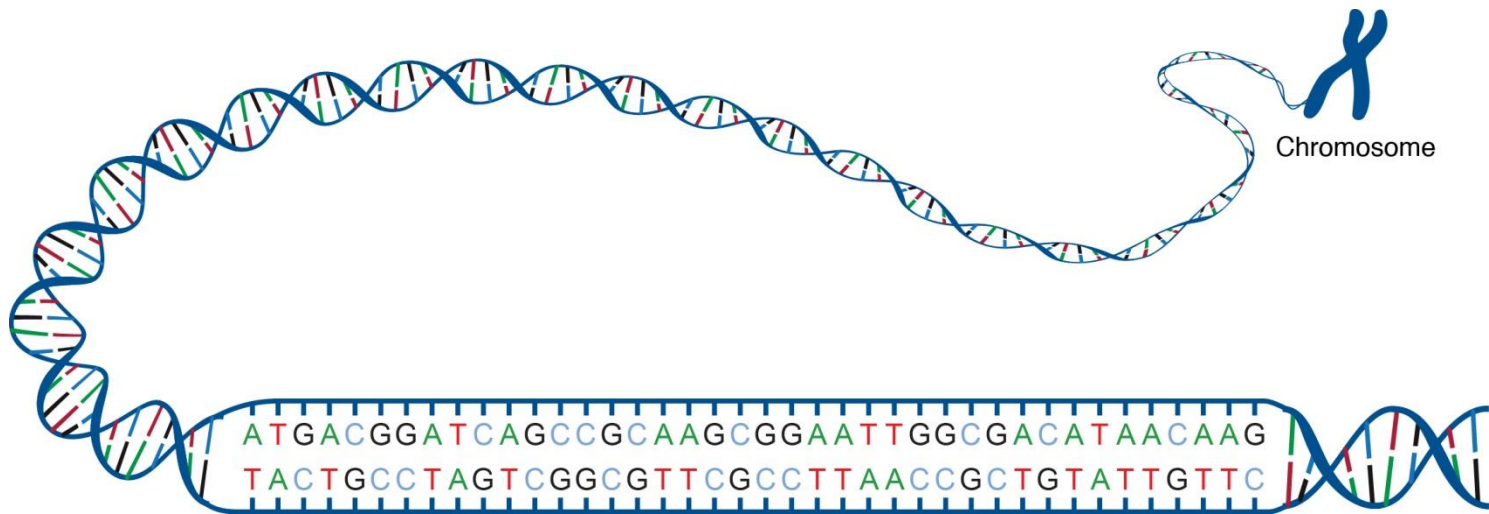
Chemicals

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

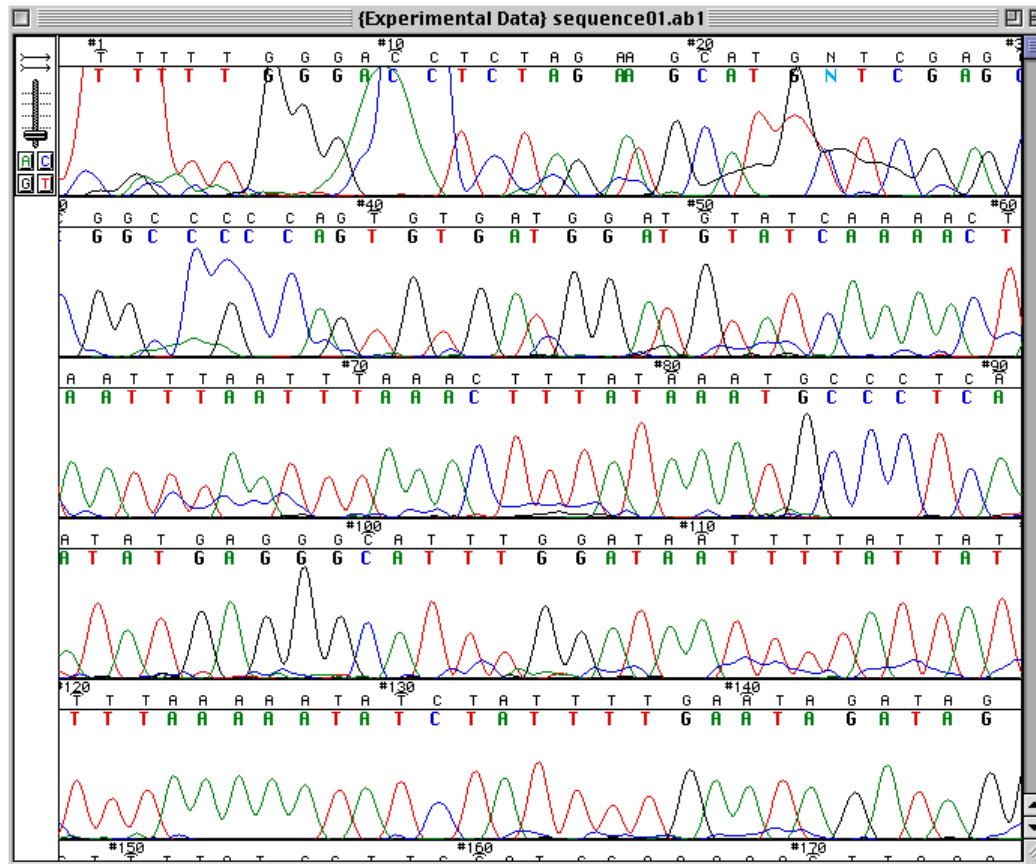
DNA



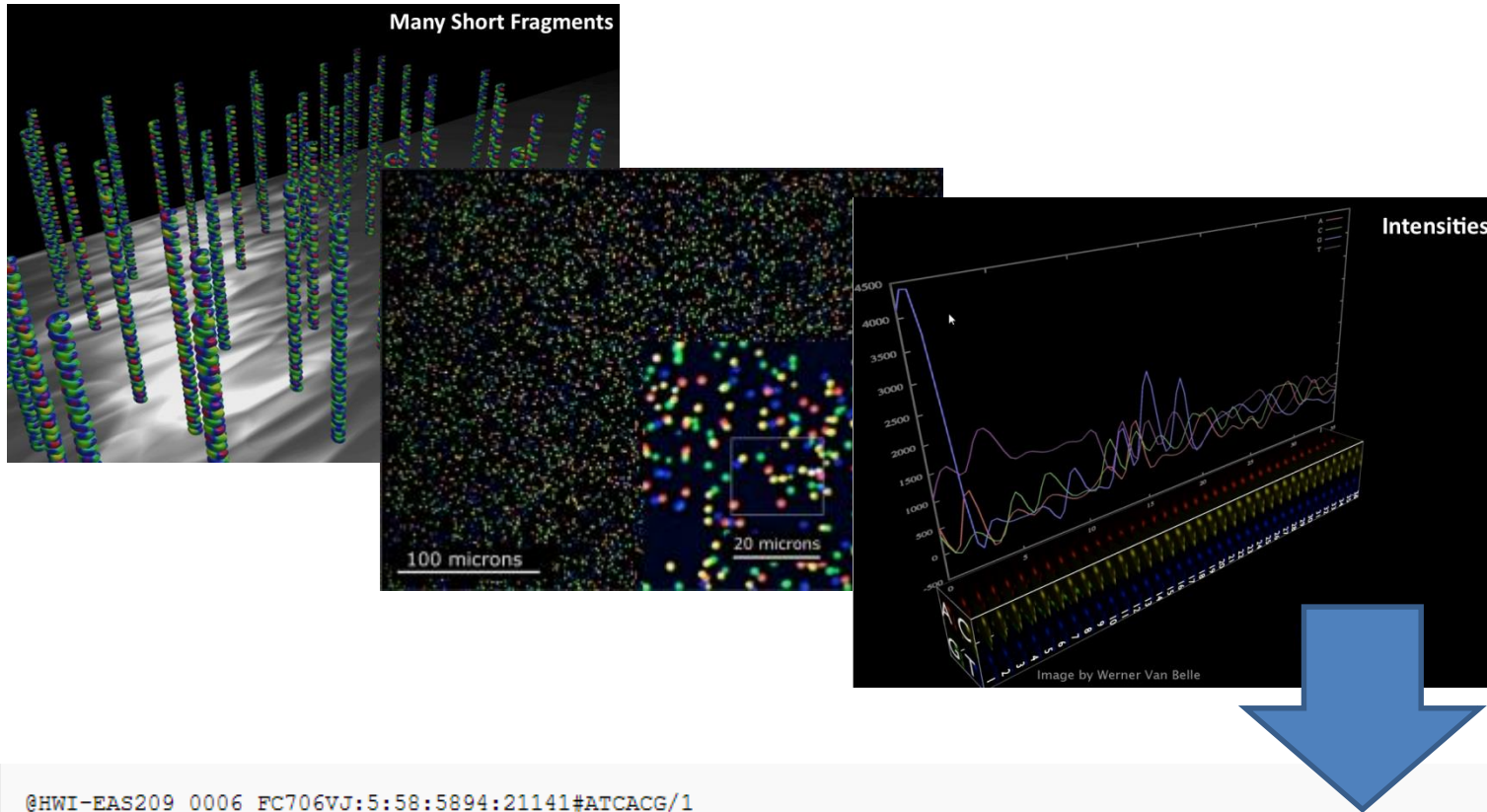
The Human Genome Project (1990-2003)

- 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
- Now (2019) can sequence a genome for less than \$1000

Sanger Sequencing



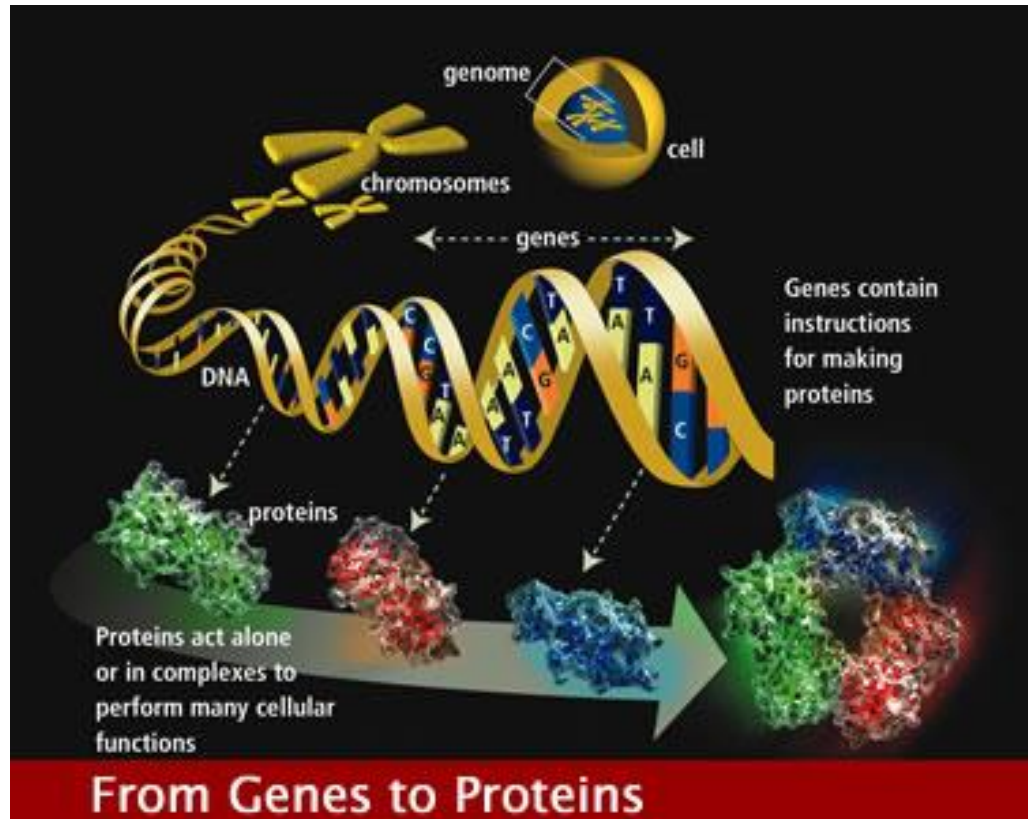
Next Generation Sequencing



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

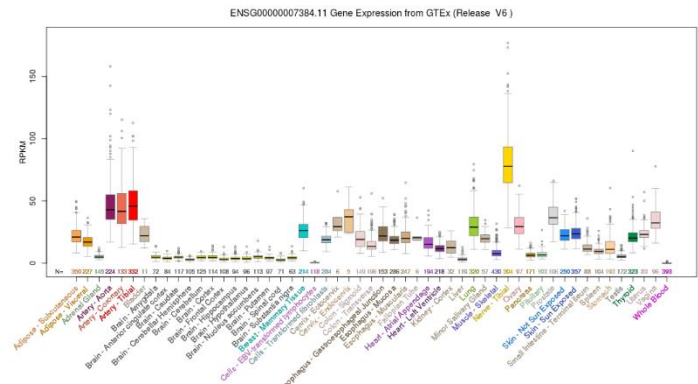
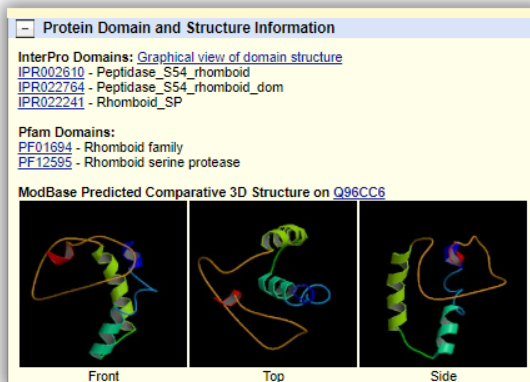
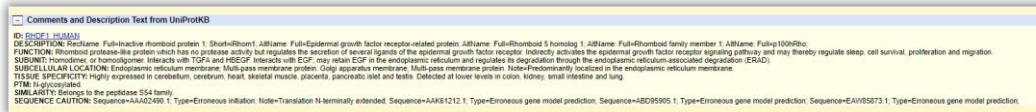
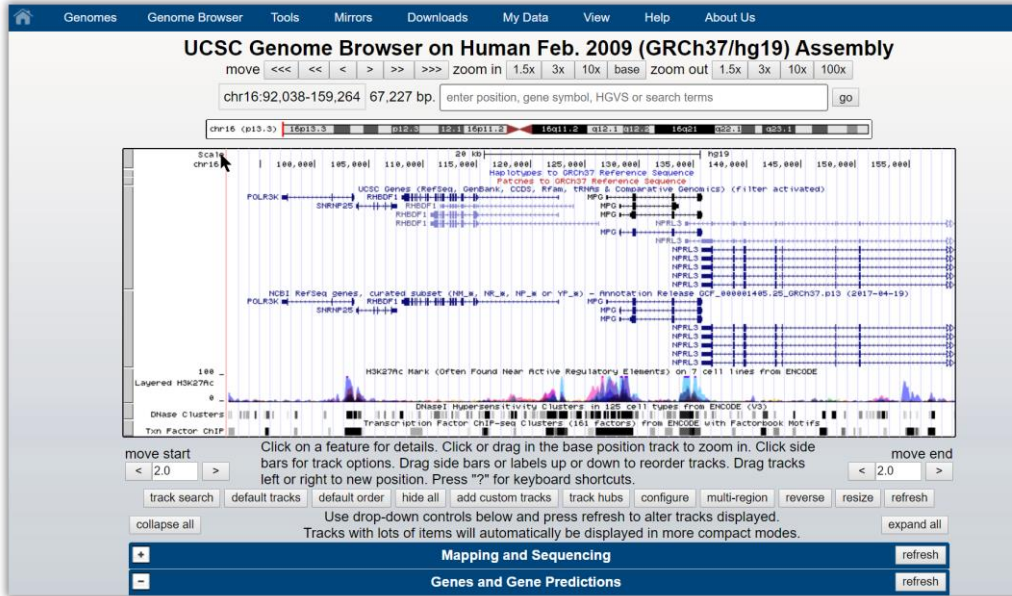
<http://werner.yellowcouch.org/Papers/pippres0802/index.html>

Central Dogma of Molecular Biology



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

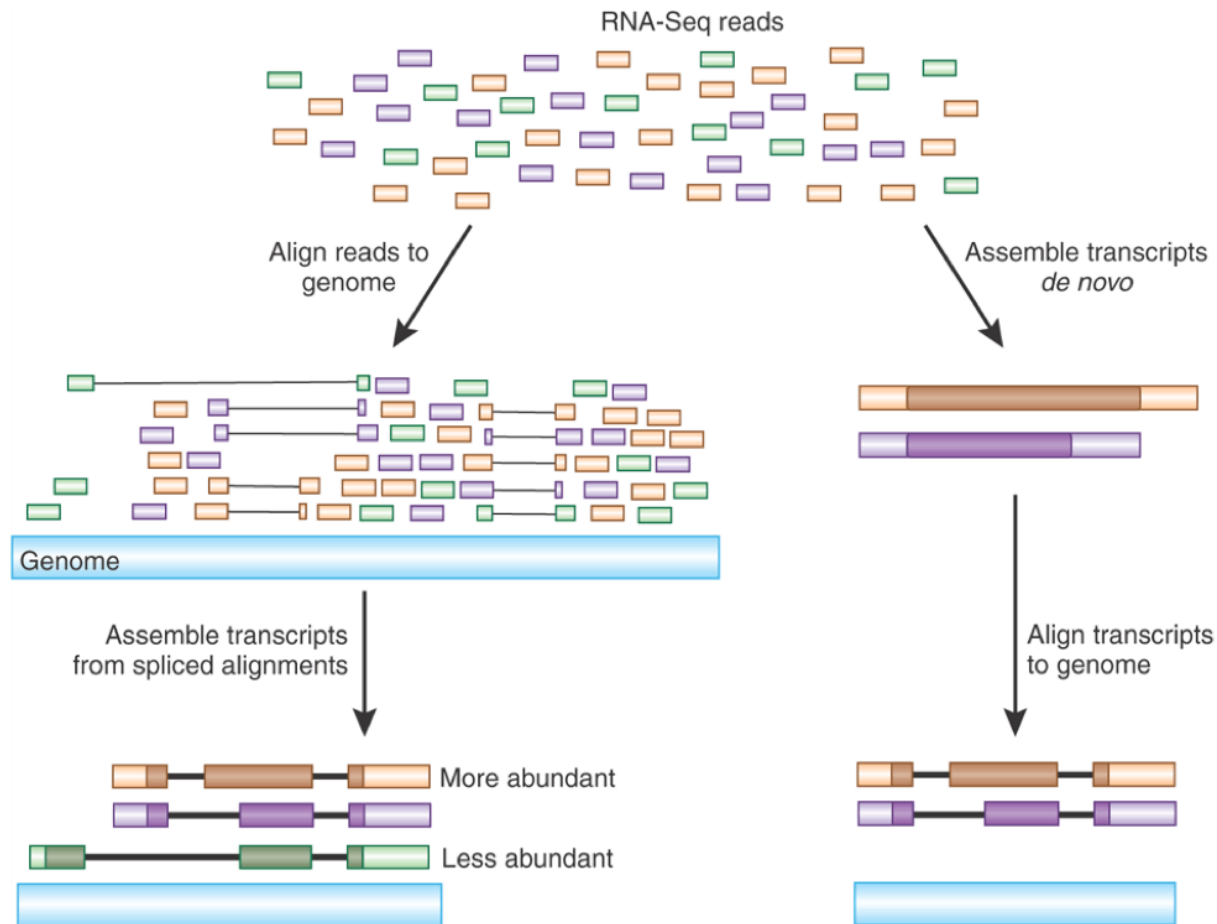
UCSC Genome Browser (<http://genome.ucsc.edu/>)



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

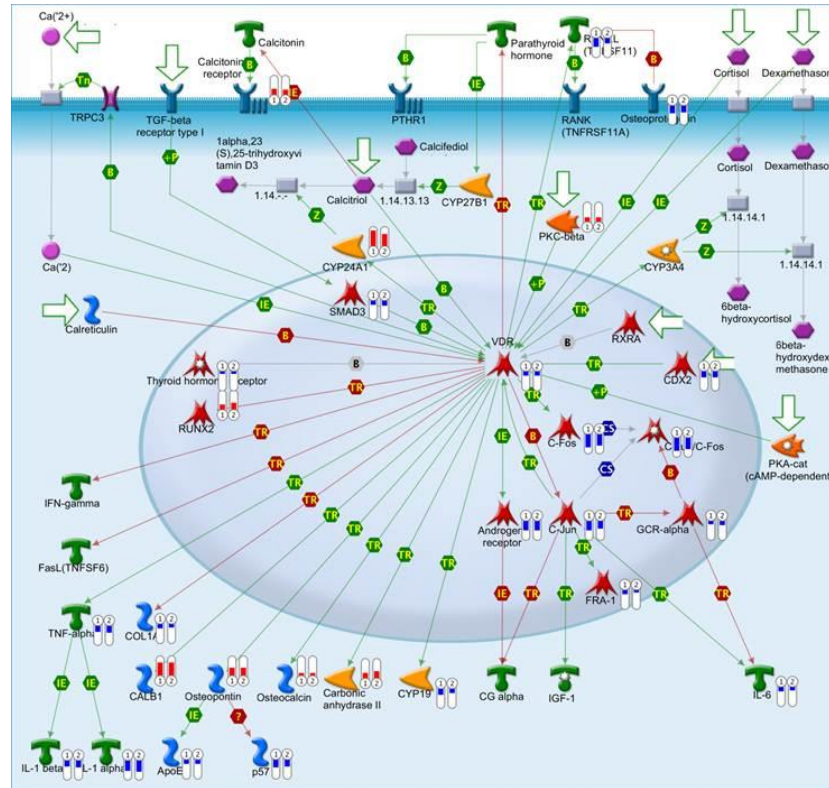
Expression Analysis (RNA-Seq)



Tools for RNA-Seq

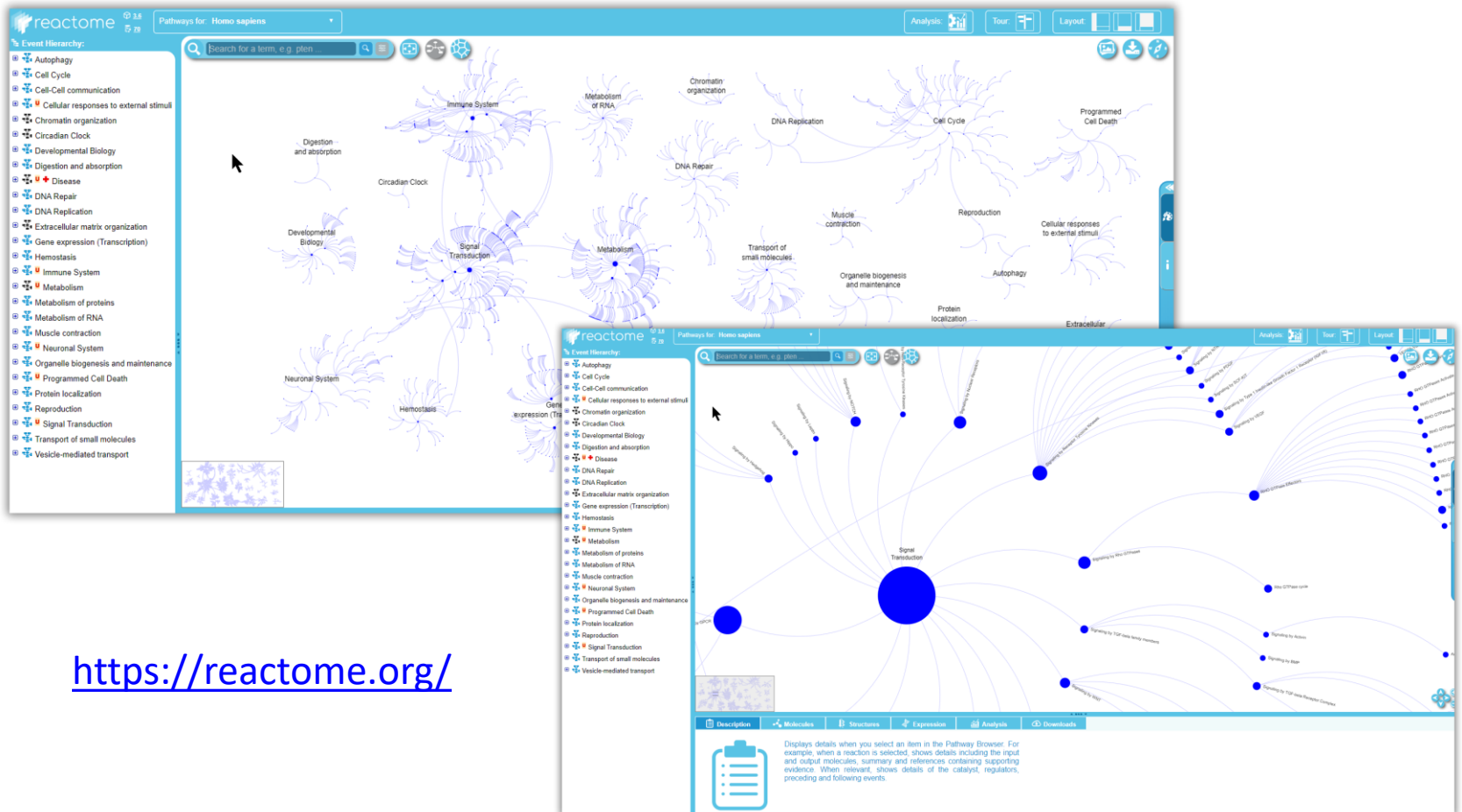
- **STAR (Spliced Transcripts Alignment to a Reference).** Fast, but uses a lot of memory.
 - Alexander Dobin, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, Thomas R. Gingeras, STAR: ultrafast universal RNA-seq aligner, *Bioinformatics*, Volume 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



Metascape

A Gene Annotation & Analysis Resource

Step 1

Or paste a gene list

UBL5,NDUF8,CHMP2B,PRPF8,NOS
TRIN,MFAP1,CWC22,PLCH2,PRPF31
ATP6AP1,DS3,CLN5,CHDC2,PIP,Z
NF473,DHX8,RAB5A,NUP98,NUP1,1,H
RK,SLC41A3,SNRPD3,SNRPD2,PCG

Submit

Cancel

Upload File Format

Single List

.xls

.xlsx

.csv

.txt

.tsv

Multiple List

.xls

.xlsx

.csv

.txt

.tsv

Test Upload

single list

3 gene lists

Test Identifiers

Gene Symbol

RefSeq

Entrez Gene ID

Step 2

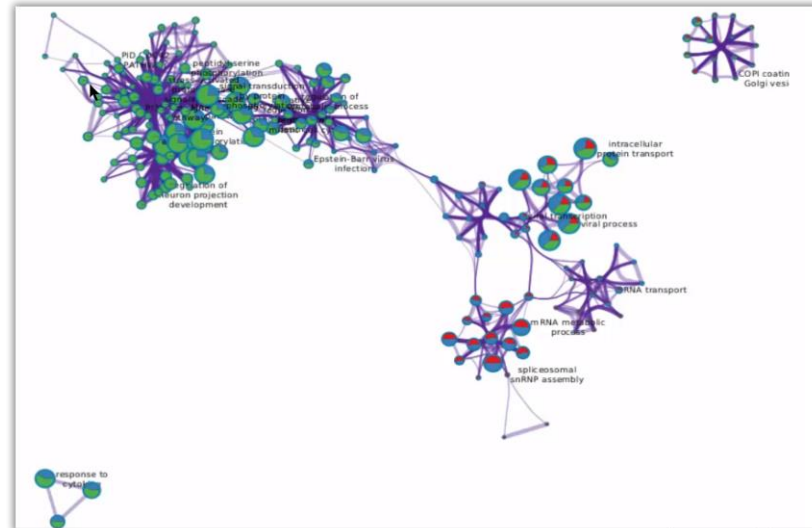
Express Analysis

Custom Analysis

working on Gene Enrichment

20

	A	B	C	D	E	F	G	H
	GroupID	Category	Term	Description	LogP	Interf_InList	Genes	Symbols
1	1_Summary	GO Biological Processes	GO:0016032	viral process	-18.851	49/771	156,527,2033	ADRBK1,ATP
2	1_Member	GO Biological Processes	GO:0016032	viral process	-18.851	49/771	156,527,2033	ADRBK1,ATP
3	1_Member	GO Biological Processes	GO:0046764	multi-organism cellular process	-18.549	49/784	156,527,2033	ADRBK1,ATP
4	1_Member	GO Biological Processes	GO:0044419	interspecies interaction between organisms	-18.086	50/838	156,527,1536	ADRBK1,ATP
5	1_Member	GO Biological Processes	GO:0044003	symbiosis, encompassing mutualism through parasitism	-18.086	50/838	156,527,1536	ADRBK1,ATP
6	1_Member	GO Biological Processes	GO:0051351	positive regulation of ligase activity	-12.291	16/101	5347,5682,56	PLK1,PSMA1
7	1_Member	GO Biological Processes	GO:0051347	positive regulation of ubiquitin-protein ligase activity involved in	-12.099	14/72	5347,5682,56	PLK1,PSMA1
8	1_Member	GO Biological Processes	GO:0002026	positive regulation of ubiquitin-protein transferase activity	-10.913	14/722	331,4734,568; XIAP, NEDD4,	
9	1_Member	GO Biological Processes	GO:0051344	positive regulation of ubiquitin-protein transferase activity	-11.468	9/56	5347,5682,56	PLK1,PSMA1
10	1_Member	GO Biological Processes	GO:0000060	positive regulation of protein ubiquitination involved in	-10.351	14/487	5347,5682,56	PLK1,PSMA1
11	1_Member	KEGG Pathway	hsa03005	Proteasome	-10.303	11/44	5682,5683,56	PSMA1,PSM
12	1_Member	GO Biological Processes	GO:0051436	negative regulation of ubiquitin-protein ligase activity involved in	-10.714	14/90	5347,5682,56	PLK1,PSMA1
13	1_Member	GO Biological Processes	GO:0000058	regulation of protein ubiquitination involved in ubiquitin-	-10.322	14/96	5347,5682,56	PLK1,PSMA1
14	1_Member	GO Biological Processes	GO:0051340	regulation of ligase activity	-10.312	16/135	5347,5682,56	PLK1,PSMA1
15	1_Member	GO Biological Processes	GO:0051444	negative regulation of ubiquitin-protein transferase activity	-10.260	9/97	5347,5682,56	PLK1,PSMA1
16	1_Member	GO Biological Processes	GO:0051352	negative regulation of ligase activity	-10.198	14/94	5347,5682,56	PLK1,PSMA1
17	1_Member	GO Biological Processes	GO:0051439	regulation of ubiquitin-protein ligase activity involved in	-10.137	14/99	5347,5682,56	PLK1,PSMA1
18	1_Member	GO Biological Processes	GO:0032446	protein modification by small protein conjugation	-9.776	37/821	331,4734,492	XIAP, NEDD4,
19	1_Member	GO Biological Processes	GO:0031198	positive regulation of protein ubiquitination	-9.579	17/179	331,5347,568; XIAP, PLK1,	
20	1_Member	GO Biological Processes	GO:0031145	anaphase-promoting complex-dependent proteasomal ubi-	-9.365	14/104	5347,5682,56	PLK1,PSMA1
21	1_Member	GO Biological Processes	GO:0002479	antigen processing and presentation of exogenous peptide	-9.412	12/75	1536,5682,56	CYBB,PSMA1
22	1_Member	GO Biological Processes	GO:0051438	regulation of ubiquitin-protein transferase activity	-9.302	15/135	5347,5682,56	PLK1,PSMA1
23	1_Member	GO Biological Processes	GO:0042590	antigen processing and presentation of exogenous peptide	-9.140	12/79	1536,5682,56	CYBB,PSMA1
24	1_Member	GO Biological Processes	GO:1903232	positive regulation of protein modification by small protein	-8.091	17/187	331,5347,568; XIAP, PLK1,	
25	1_Member	GO Biological Processes	GO:0042485	cellular macromolecule catabolic process	-8.055	38/613	3146,3837,41	HMGBl,KPNI
26	1_Member	GO Biological Processes	GO:000521	regulation of cellular amino acid metabolic process	-9.021	11/94	5682,5683,56	PSMA1,PSM
27	1_Member	GO Biological Processes	GO:0042787	protein ubiquitination involved in ubiquitin-dependent	-8.362	16/166	4734,5347,56	NEDD4, PLK1,
28	1_Member	GO Biological Processes	GO:0006977	DNA damage response, signal transduction by p53 class me-	-8.872	11/66	5682,5683,56	PSMA1,PSM



<http://metascape.org/>

Gene Expression Omnibus (GEO)

The screenshot displays the NCBI GEO repository browser interface. The top navigation bar includes the NCBI logo and links for GEO Publications, FAQ, NIAID, and Email GEO. Below the navigation bar, there are tabs for Series, Samples, Platforms, and DataSets. The 'Samples' tab is selected, showing a list of 1,163,446 samples. A search bar and an 'Export' button are visible. The main table lists samples with columns for Accession, Title, Sample type, Organism(s), Ch, Platform, Series, Supplementary, Contact, and Release date. A sample with Accession GSM952626 is highlighted. Below the table, a detailed view for sample GSM952626 is shown, including its status, title, sample type, source name, organism, characteristics, growth protocol, extracted molecule, extraction protocol, label, and label protocol.

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 Bapu Londhe	Jun 23, 2014
GSM952627	SPC/cRaf mouse dysplasia 67.3_71.5 male 3 months	RNA	Mus musculus	1	GPL6096	GSE38948	CEL CHP	Kishor Bapu Londhe	Jun 23, 2014
GSM952628	SPC/cRaf mouse dysplasia 73.5 male 6 months	RNA	Mus musculus	1	GPL6096	GSE38948	CEL CHP	Kishor Bapu Londhe	Jun 23, 2014
GSM952629	SPC/cRaf mouse dysplasia 73.7 male 6 months	RNA	Mus musculus	1	GPL6096	GSE38948	CEL CHP	Kishor Bapu Londhe	Jun 23, 2014
GSM952630	non-transgenic mouse 65.0 male 7 months	RNA	Mus musculus	1	GPL6096	GSE38948	CEL CHP	Kishor Bapu Londhe	Jun 23, 2014
GSM952631	non-transgenic mouse 67.5 female 7 months	RNA	Mus musculus	1	GPL6096	GSE38948	CEL CHP	Kishor Bapu Londhe	Jun 23, 2014
GSM952632	non-transgenic mouse 92.7 female 11 months	RNA	Mus musculus	1	GPL6096	GSE38948	CEL CHP	Kishor Bapu Londhe	Jun 23, 2014
GSM952633	non-transgenic mouse	RNA	Mus musculus	1	GPL6096	GSE38948	CEL CHP	Kishor Bapu Londhe	Jun 23, 2014
GSM952634	non-transgenic mouse	RNA	Mus musculus	1	GPL6096	GSE38948	CEL CHP	Kishor Bapu Londhe	Jun 23, 2014
GSM952635	SPC/cRaf mouse	RNA	Mus musculus	1	GPL6096	GSE38948	CEL CHP	Kishor Bapu Londhe	Jun 23, 2014
GSM952636	SPC/cRaf mouse	RNA	Mus musculus	1	GPL6096	GSE38948	CEL CHP	Kishor Bapu Londhe	Jun 23, 2014
GSM952637	SPC/cRaf mouse	RNA	Mus musculus	1	GPL6096	GSE38948	CEL CHP	Kishor Bapu Londhe	Jun 23, 2014
GSM952638	SPC/cRaf mouse	RNA	Mus musculus	1	GPL6096	GSE38948	CEL CHP	Kishor Bapu Londhe	Jun 23, 2014
GSM952639	SPC/cRaf mouse	RNA	Mus musculus	1	GPL6096	GSE38948	CEL CHP	Kishor Bapu Londhe	Jun 23, 2014
GSM952640	SPC/cRaf mouse	RNA	Mus musculus	1	GPL6096	GSE38948	CEL CHP	Kishor Bapu Londhe	Jun 23, 2014
GSM952641	SPC/cRaf mouse	RNA	Mus musculus	1	GPL6096	GSE38948	CEL CHP	Kishor Bapu Londhe	Jun 23, 2014
GSM952642	SPC/cRaf mouse	RNA	Mus musculus	1	GPL6096	GSE38948	CEL CHP	Kishor Bapu Londhe	Jun 23, 2014
GSM1314708	ECFC_L1_1	RNA	Human	1	GPL6096	GSE54416	SRA Experiment	Terri DiMaio	Jun 23, 2014
GSM1314709	ECFC_L1_2	RNA	Human	1	GPL6096	GSE54416	SRA Experiment	Terri DiMaio	Jun 23, 2014
GSM1314710	ECFC_L1_3	RNA	Human	1	GPL6096	GSE54416	SRA Experiment	Terri DiMaio	Jun 23, 2014

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.

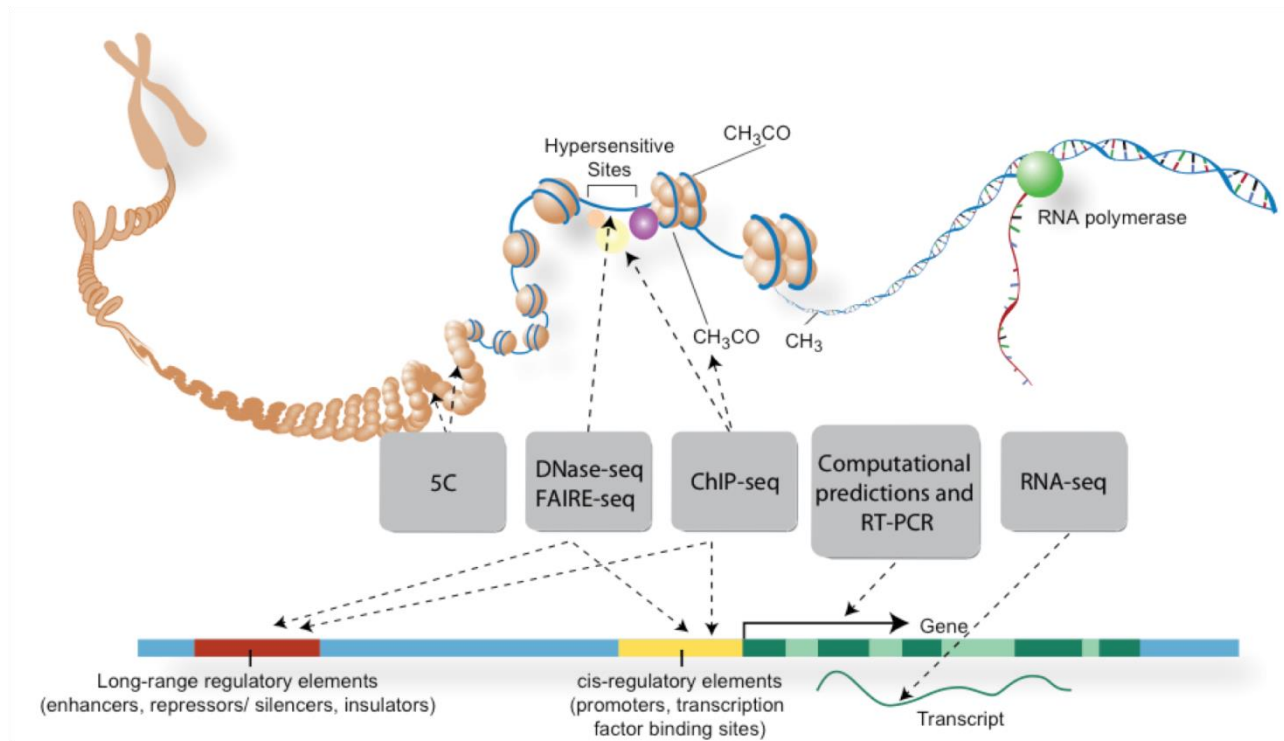
Extracted molecule: total RNA

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).

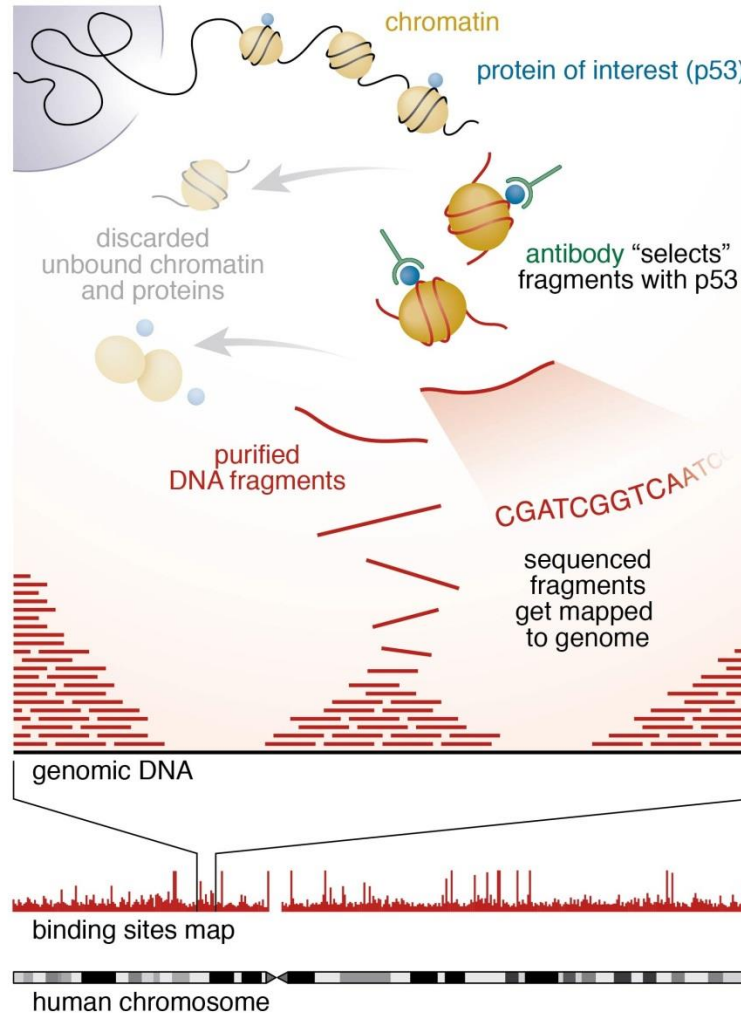
ENCODE (Encyclopedia of DNA Elements)



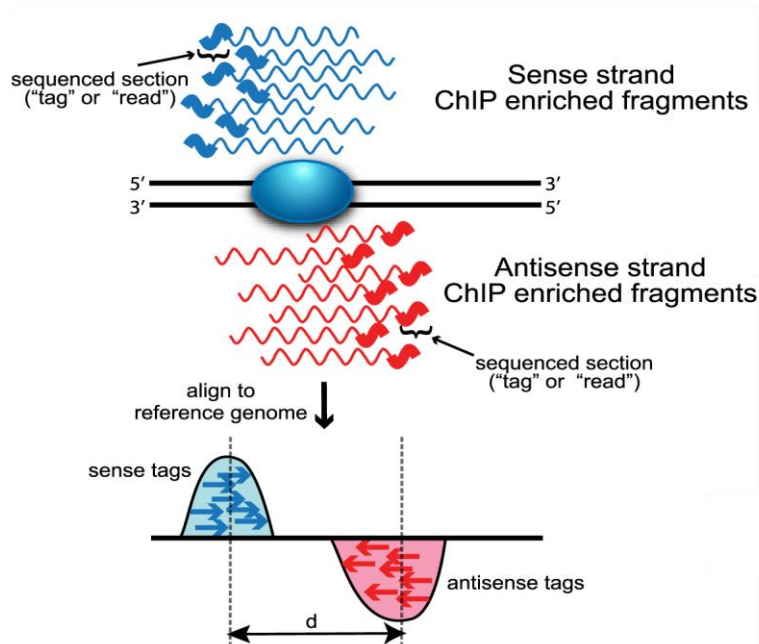
<http://genome.ucsc.edu/ENCODE/>

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

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

Configuration:

Total features: 20929

Data Filters:

- Alias
- CD4 DHS
- CD4 DHS Normalised
- CD4_H3K27Me3
- CD4_H3K4Me3
- CD4_H3K4Me3 Normalised
- CD4_H3K4Me3 vs H3K27Me3
- description

distance from exon1

less than
5000
(min: -2421338, max: 2869967)

Location filter:

From: bp
To: bp
Chr:

Keyword filter:

Keywords: one per row line

Results panel options:

Sort features by: CD4 PolII Normalised
descending

Feature label: Alias

Features per page: 30

Image options:

Display: ☒ fixed interval ☐ zoom-to-feature

Interval size: 100000 bp

MIG: Multi-Image Genome

Links: [home](#) | [switch project](#) | [configuration](#) | [sessions](#) | [upload new dataset](#) | [- ADMIN -](#)

Project: Barski Annotated
Feature: chr12:91061033..91063751
Region size: 10000bp (chr12:91012391..91112393)
View in: [GBrowse](#)

Filters used:
distance_from_exon1 < 5000

Features after filter: 6262

Further options

Page: 1 | 2 | 3 > 209

Rank	CD4_PolII_Normalised	Feature	
1	47.60	UCR74	[+/-]
2	52.11	UUN8	[+/-]
3	33.99	CXCR4	[+/-]
4	33.42	BTG1	[+/-]

Feature data:

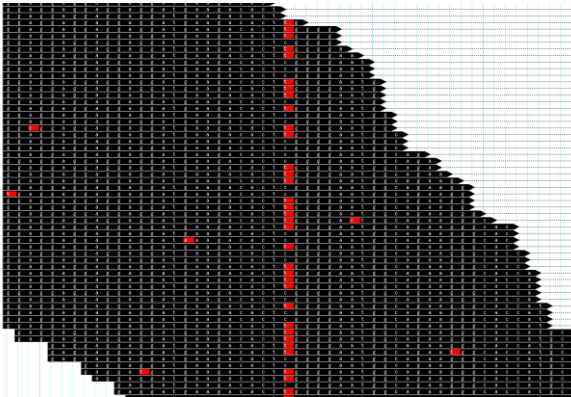
Coordinates: chr12:91061033..91063751

Alias	BTG1
CD4_DHS	39.46
CD4_DHS_Normalised	26.03
CD4_H3K27Me3	0.0543
CD4_PolII	33.65
CD4_PolII_Normalised	33.42
CD4H3K27Me3_Normalised	-0.37
CD4H3K4Me3	2.01
CD4H3K4Me3_Normalised	2.58
CD4H3K4Me3_vs_H3K27Me3	10.26
Description	G-protein transduction protein 1
distance_from_exon1	1358
expression_in_CD4	3.76
GO_BP	negative regulation of cell proliferation cell migration positive regulation of myoblast differentiation regulation of apoptosis regulation of transcription protein amino acid methylation positive regulation of angiogenesis negative regulation of cell growth positive regulation of endothelial cell differentiation
GO_CC	cytosol cytoplasm
GO_MF	kinase binding transcription coactivator activity
JM_Input	0.4277
location_overlap	intron
refseq	NM_001731
refseq_strand	-
strand	-
ucsknowngened	uc001tbl.1

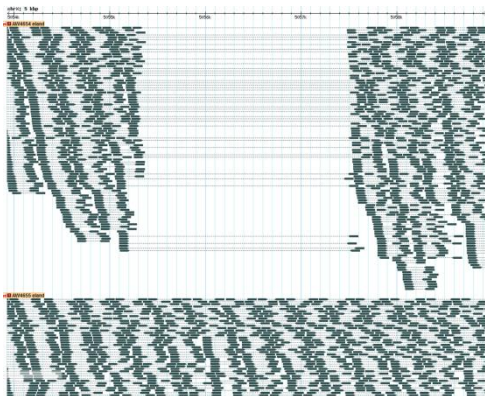
[McGowan SJ, Hughes JR, Han ZP, Taylor S](#) MIG: Multi-Image Genome viewer.
Bioinformatics (2013) **29**: 2477-8

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.

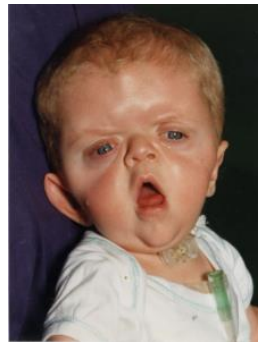
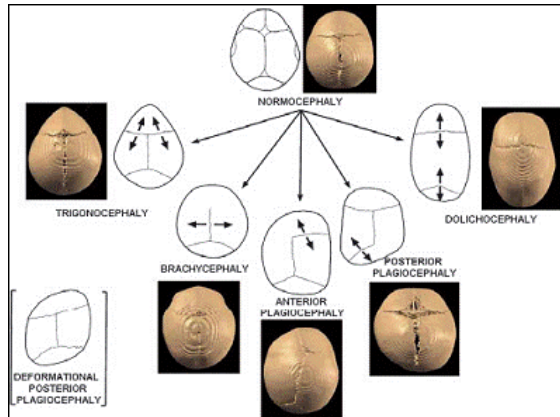


The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA, 2010 *GENOME RESEARCH* 20:1297-303



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, Katie Warner, sitting on a floral patterned sofa and holding a young child, Marie Turner. The woman is smiling and looking at the child, who is also smiling and holding a small white stuffed animal. The background is a plain wall.

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

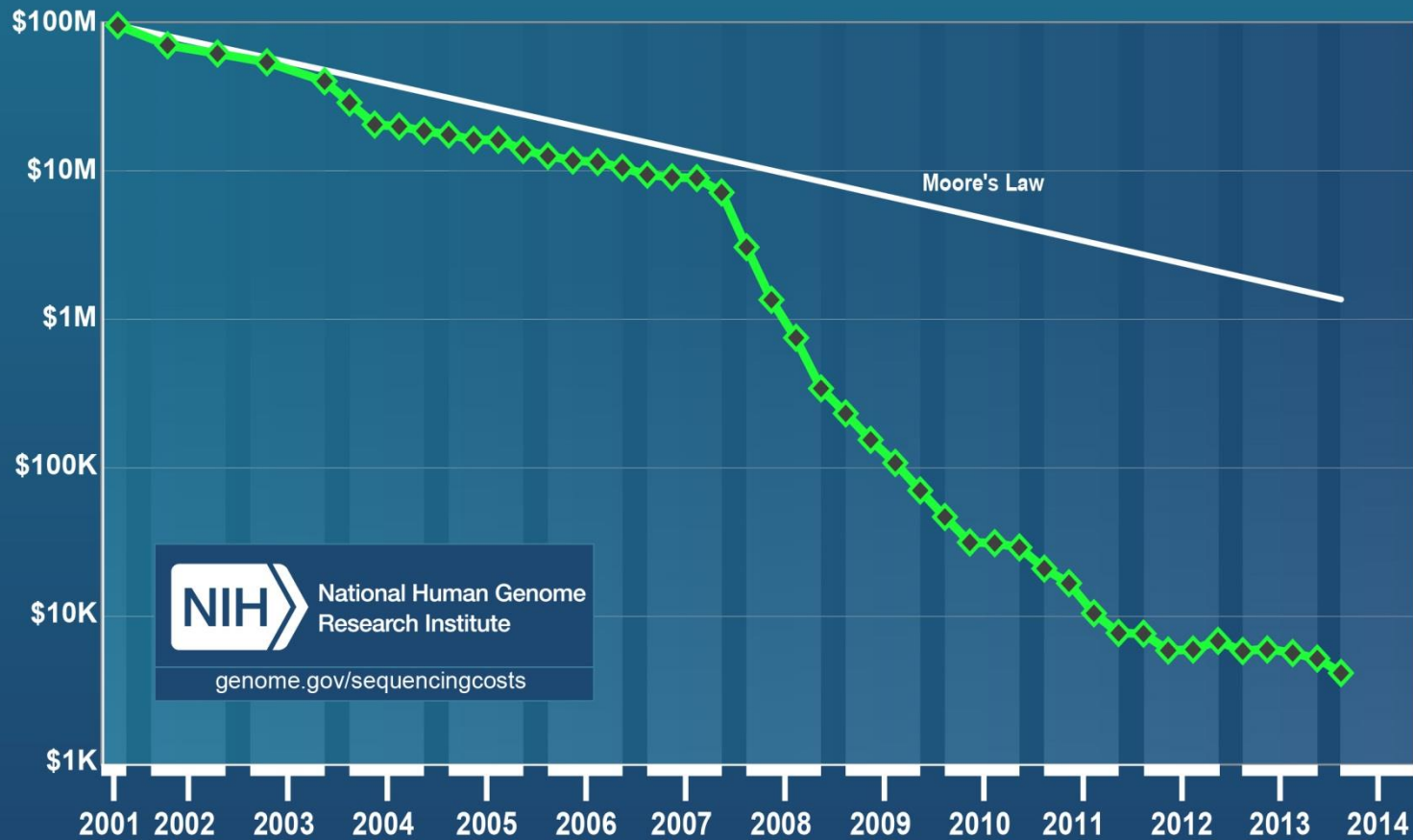
Post a comment

Recommend (4)



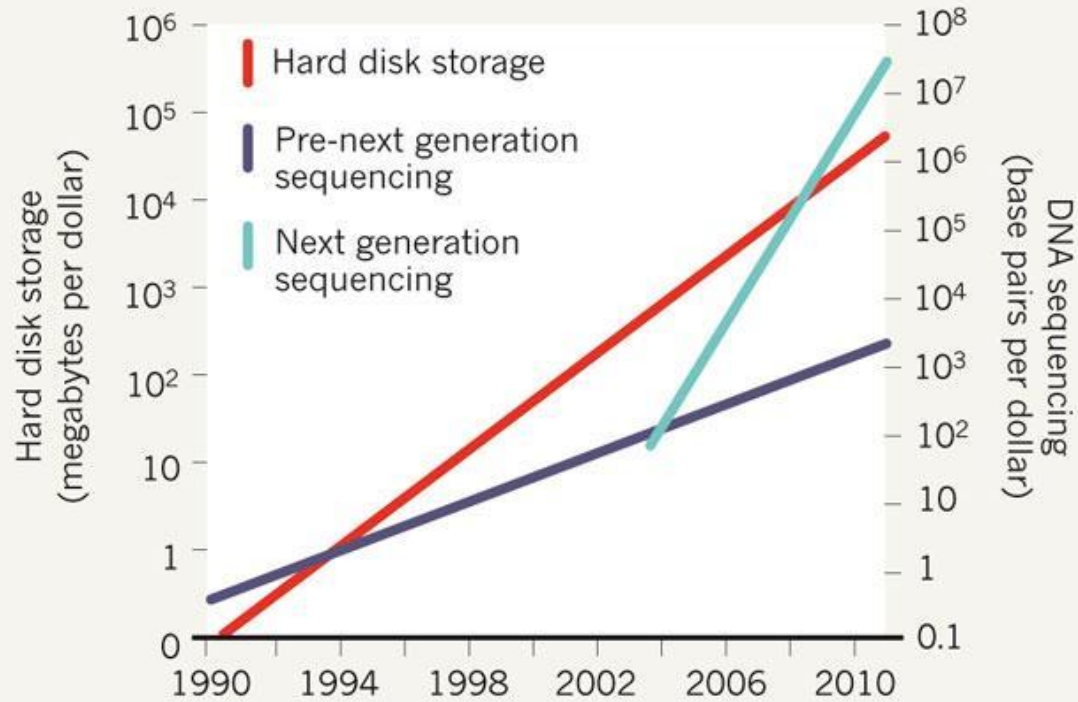
- 100,000 patients with rare inherited disease, common cancers and pathogens from the NHS in England
- Whole Genome Sequencing
- <http://www.genomicsengland.co.uk/>

Cost per Genome



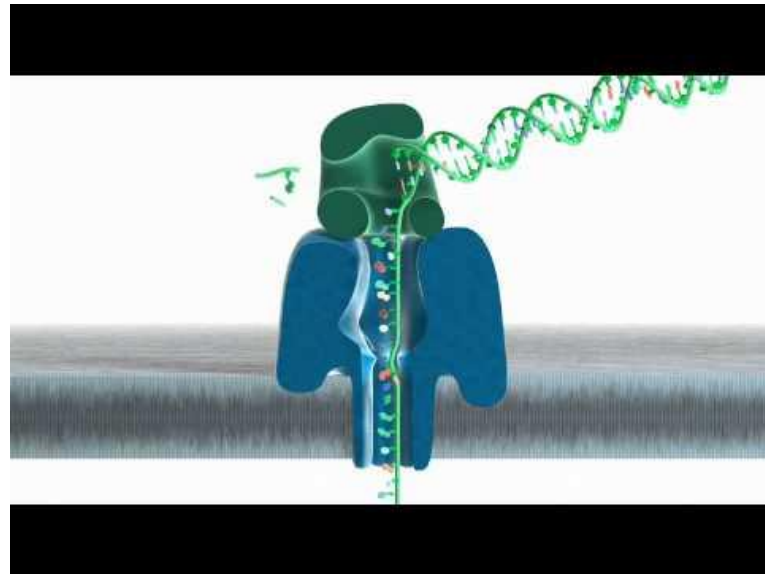
DNA AND CHIPS

The price of DNA sequencing is falling faster than computer storage costs, making cloud computing an increasingly important tool in genomics.



Source: L. D. Stein *Genome Biol.* 11, 207 (2010)

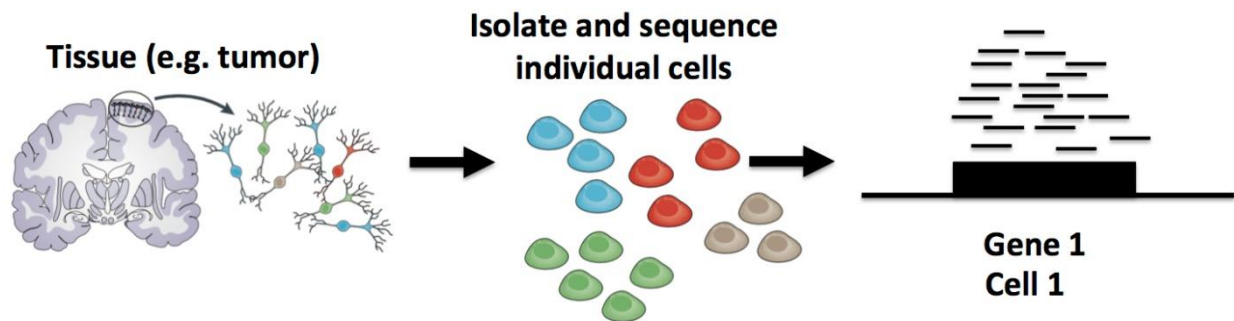
Nanopore sequencing



200KB Read Lengths, genome assembly, direct RNA sequencing...

Single Cell Sequencing

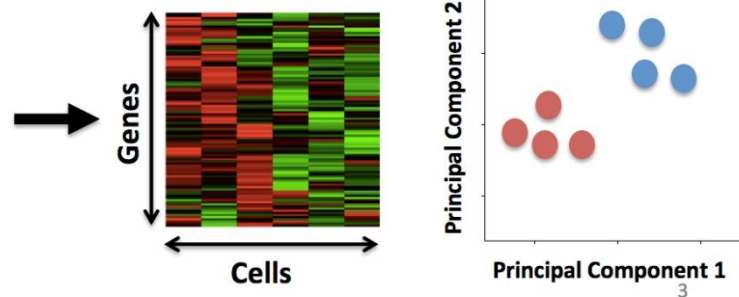
Single-cell RNA-Seq (scRNA-Seq)



Read Counts

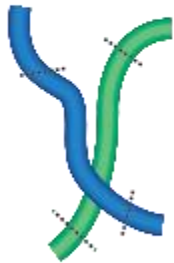
	Cell 1	Cell 2	...
Gene 1	18	0	
Gene 2	1010	506	
Gene 3	0	49	
Gene 4	22	0	
...			

Compare gene expression profiles of single cells



Genome Modelling

1. Cut DNA strands with enzyme



2. Mark pieces for identification



3. Reseal DNA



4. Pull out sealed pieces



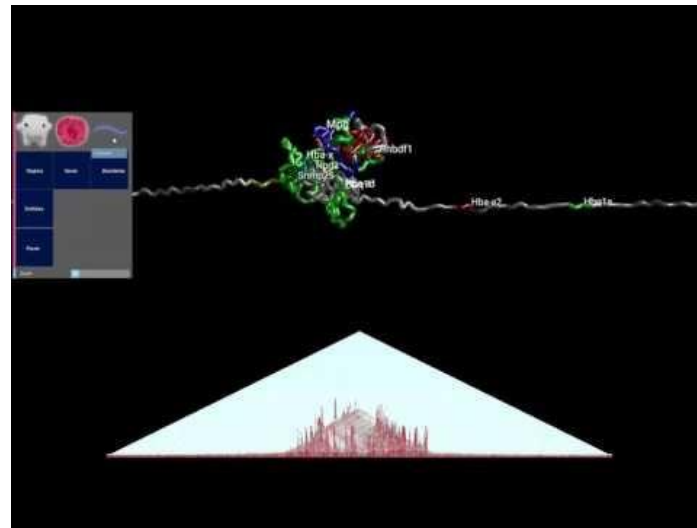
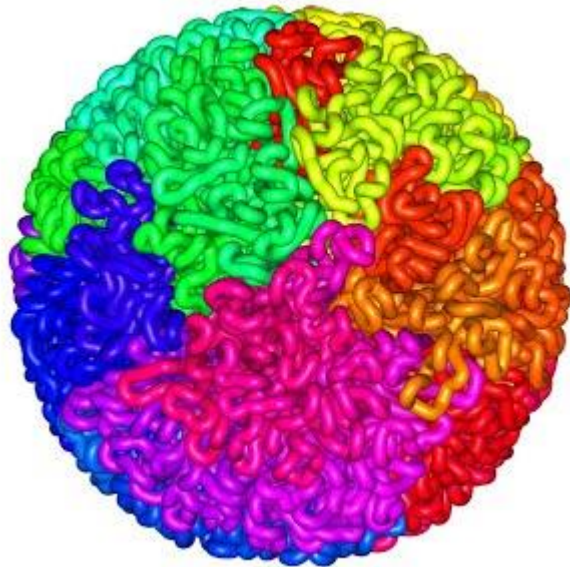
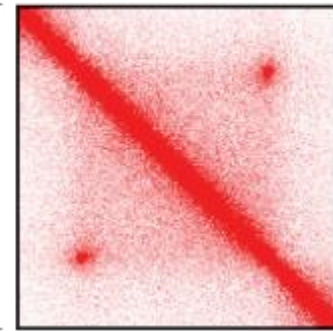
5. Read DNA



Chromosome 13

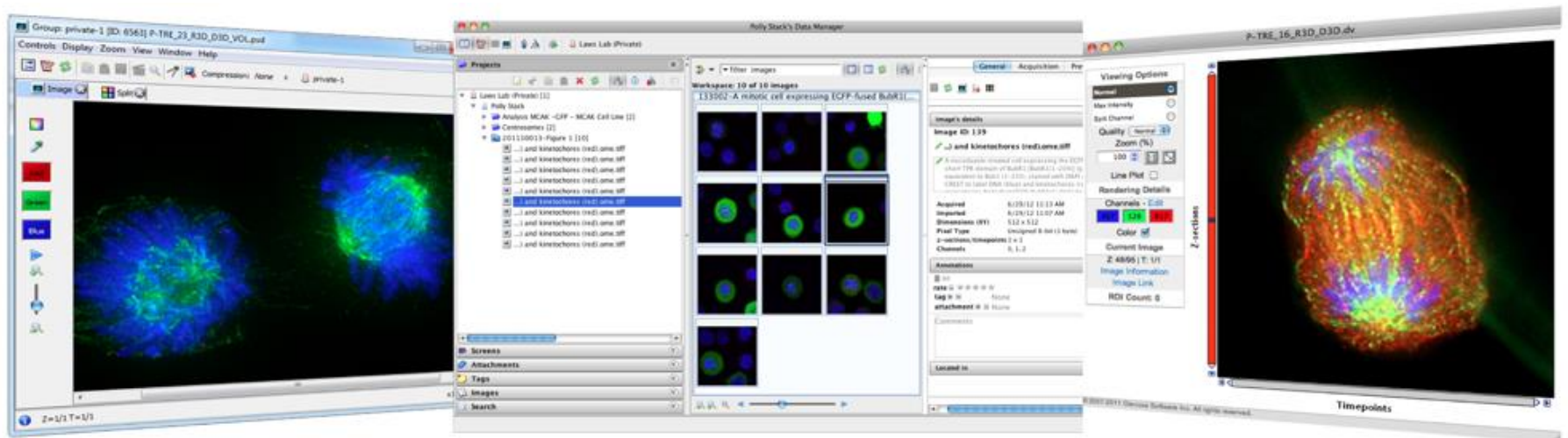
~1 million DNA letters

Chromosome 13

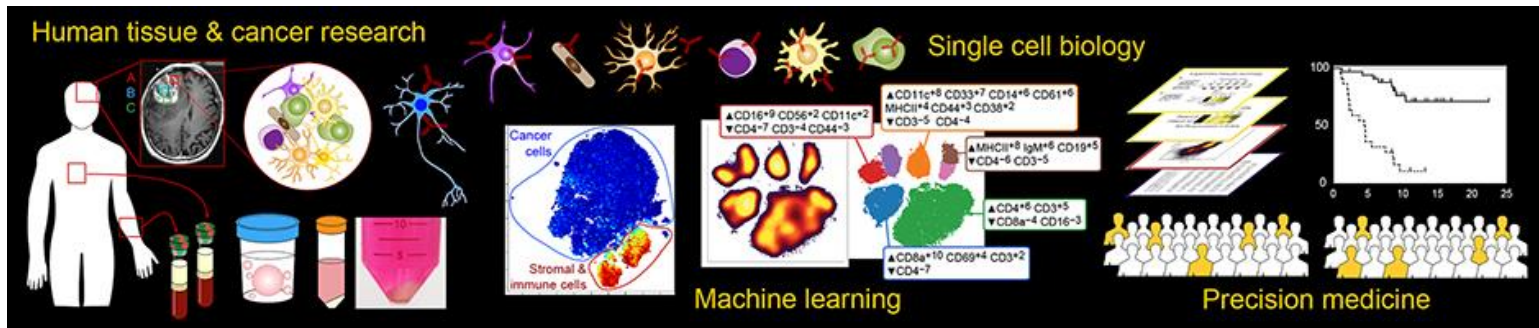
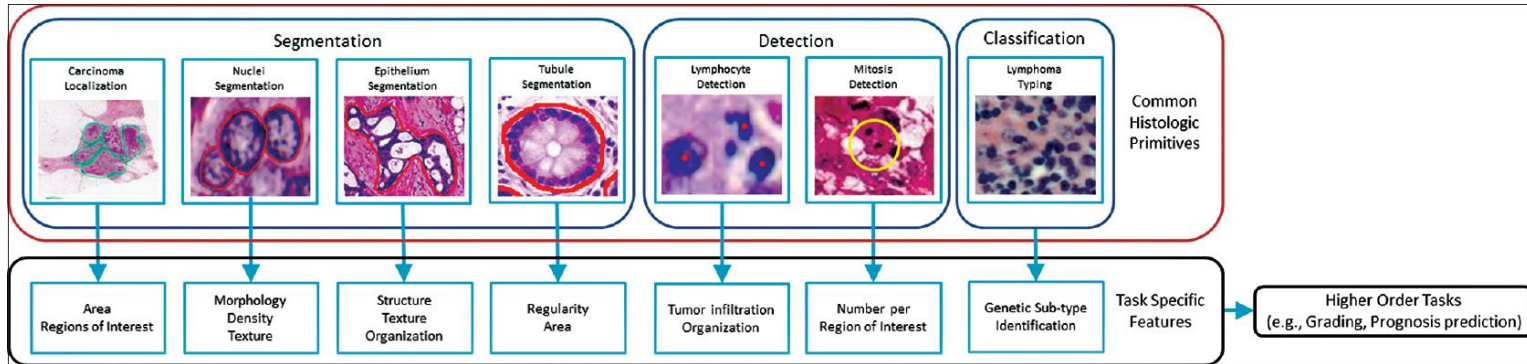


CSynth
Bio-Visualisation made interactive

OMERO Image Database



Machine Learning

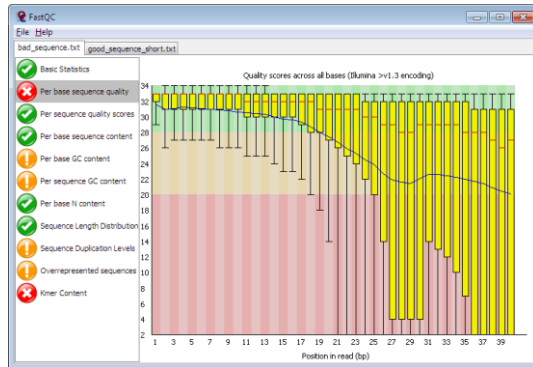


<https://my.vanderbilt.edu/irishlab/>

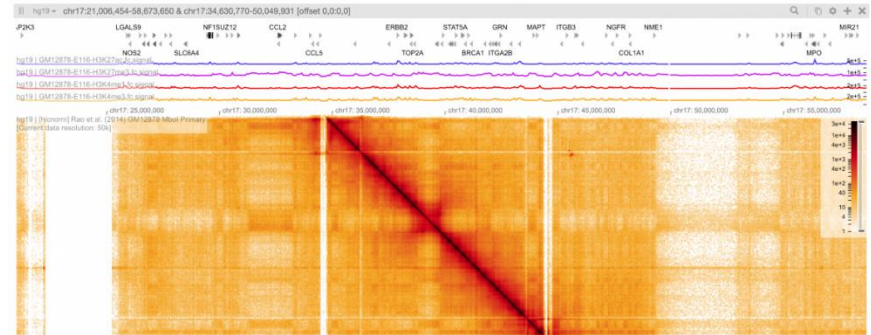
Also being applied increasingly to all data types e.g. health records, DNA sequences

Visualisation

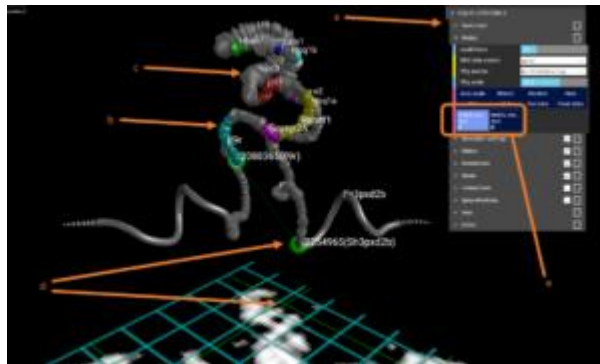
FASTQC



HiGlass



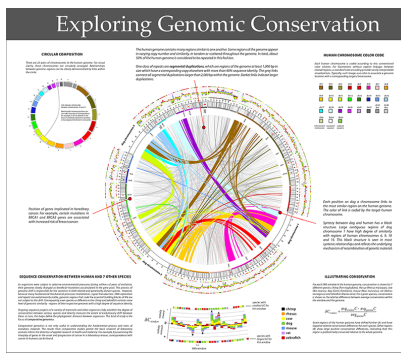
CSynth



Zegami



Circos



BabelVR



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