

Managing and accessing biological information in research

(From DNA bases to image databases)

Stephen Taylor

Computational Biology Research Group



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?

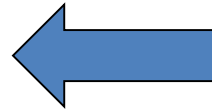
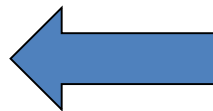
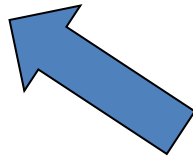
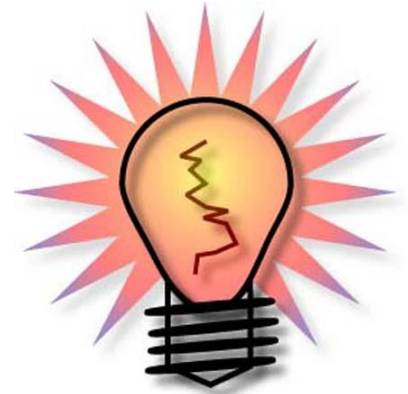
Experiment



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
 - Structure e.g. PDB
 - Pathways e.g. KEGG, Metacore
 - Interactions e.g. IntAct
 - Diseases e.g. OMIM
 - And many others ...
- Can be searched in a variety of ways
e.g. keyword, sequence, pattern,

Keyword

NCBI Resources How To bioinfblake My NCBI Sign Out

Search NCBI databases

Help

p53 Search

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
- Computational Biology Research Group provide a environment to do this

DNA Movie

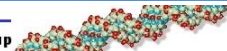
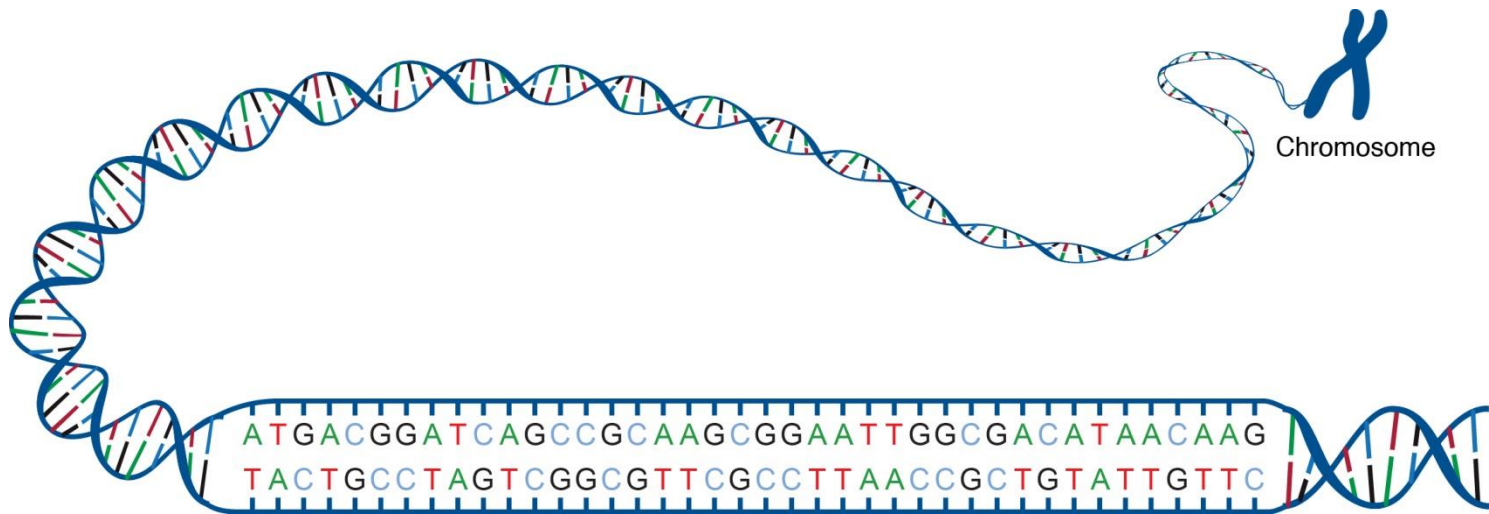
wehi.edu.au

Molecular visualizations of

DNA

1. *DNA Wrapping*

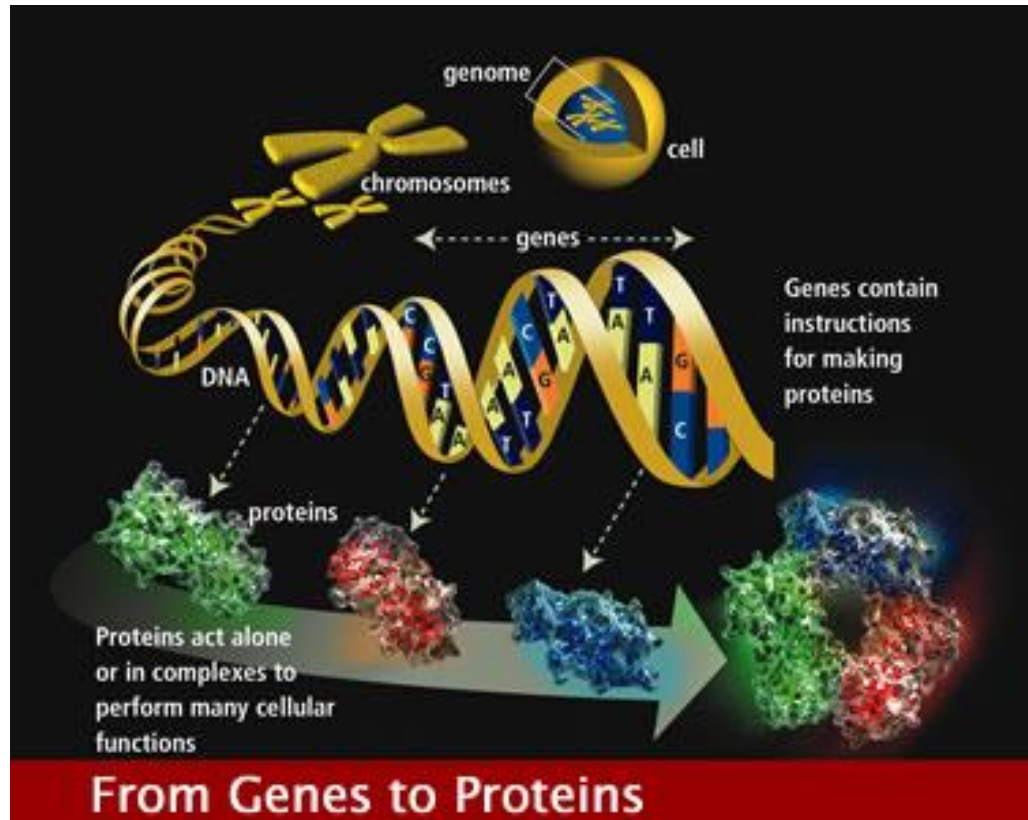
DNA



The Human Genome Project (1990-2003)

- 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

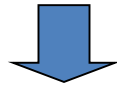
Central Dogma of Molecular Biology



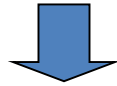
(http://www.ornl.gov/sci/techresources/Human_Genome/home.shtml)

Genome Bioinformatics

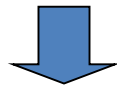
Assemble



Analyse



Annotate



Display

Assembly

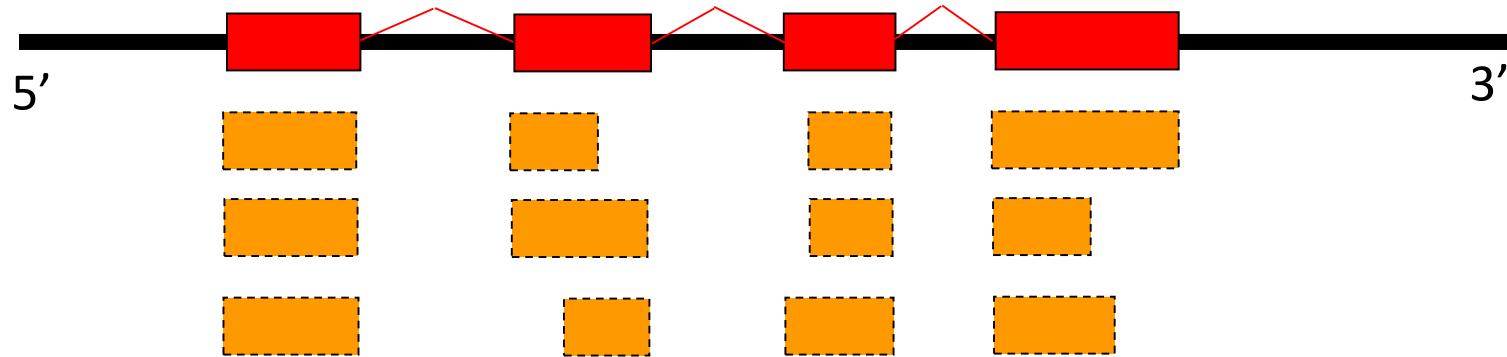
- Human genome is theoretically several long strings totalling 3 billion base pairs
 - Assembled via hundreds of thousands of overlapping units or contigs to make a single consensus sequence
 - Sequences collated using information stored on ABI sequencer
 - Sequence assembly bioinformatics tools used to
 - Automatically assemble fragments
 - Hand finish using computer tools
 - Required constant reassembly and rebuilds as new data comes in

Analyse

- Take the assembled string of nucleotides

AGTACGTAGTAGCTGCTGCTACGTGCGCTAGCTAGTACG
TCACGACGTAGATGCTAGCTGACTCGATGCAGACTGCTA
GCTGCCAGCGACTCAGCTACGACTAGCATCGGCGCTAG
CATCGGCAGC...

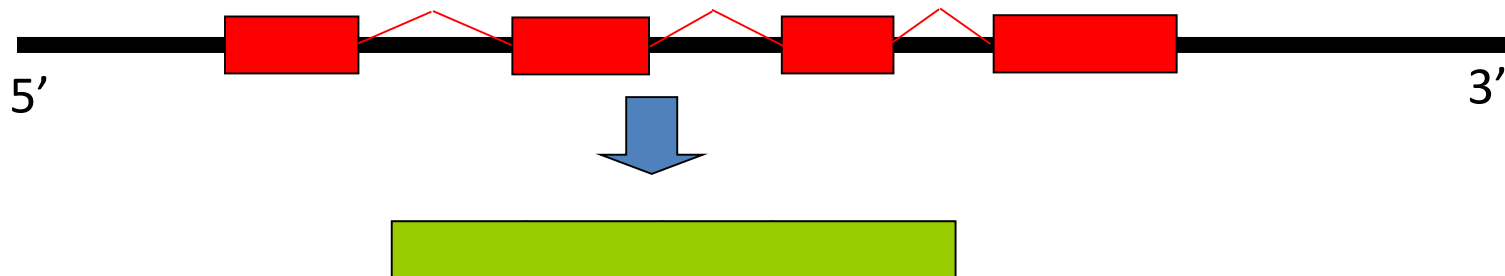
Model Genes



- Map existing RNA sequences to the genome
- Train algorithm to look for features e.g.
 - Splice sites
 - Start / Stop codons
 - Codon frequency
 - Promoters

Find Translated Protein(s)

- Translate DNA to theoretical protein



>Unknown Sequence

VLSPADKTNVKAAWGKVGAHAGEYGAELERMFLSFPTTKTYFPHFDLSHGSAQVKG
HGKKVADALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPA
EFTPAVHASLDKFLASVSTVLTSKYR

Find Function

- Major challenge in bioinformatics
 - Search the protein sequence vs database of proteins of known function*
 - Protein domains are evolutionarily conserved
 - Proteins that are similar in sequence across several species are likely to have a similar function
 - BLAST :
 - A query sequence
 - Sequence database (protein or nucleotide)
 - Inspection of significant hits
 - There are many other methods used to imply function!

Example

Query



Human, Unknown

Sequence database



Search Results



Chimp, Myoglobin



Pig, Myoglobin



Mouse, Myoglobin



Putative function = Human, Myoglobin

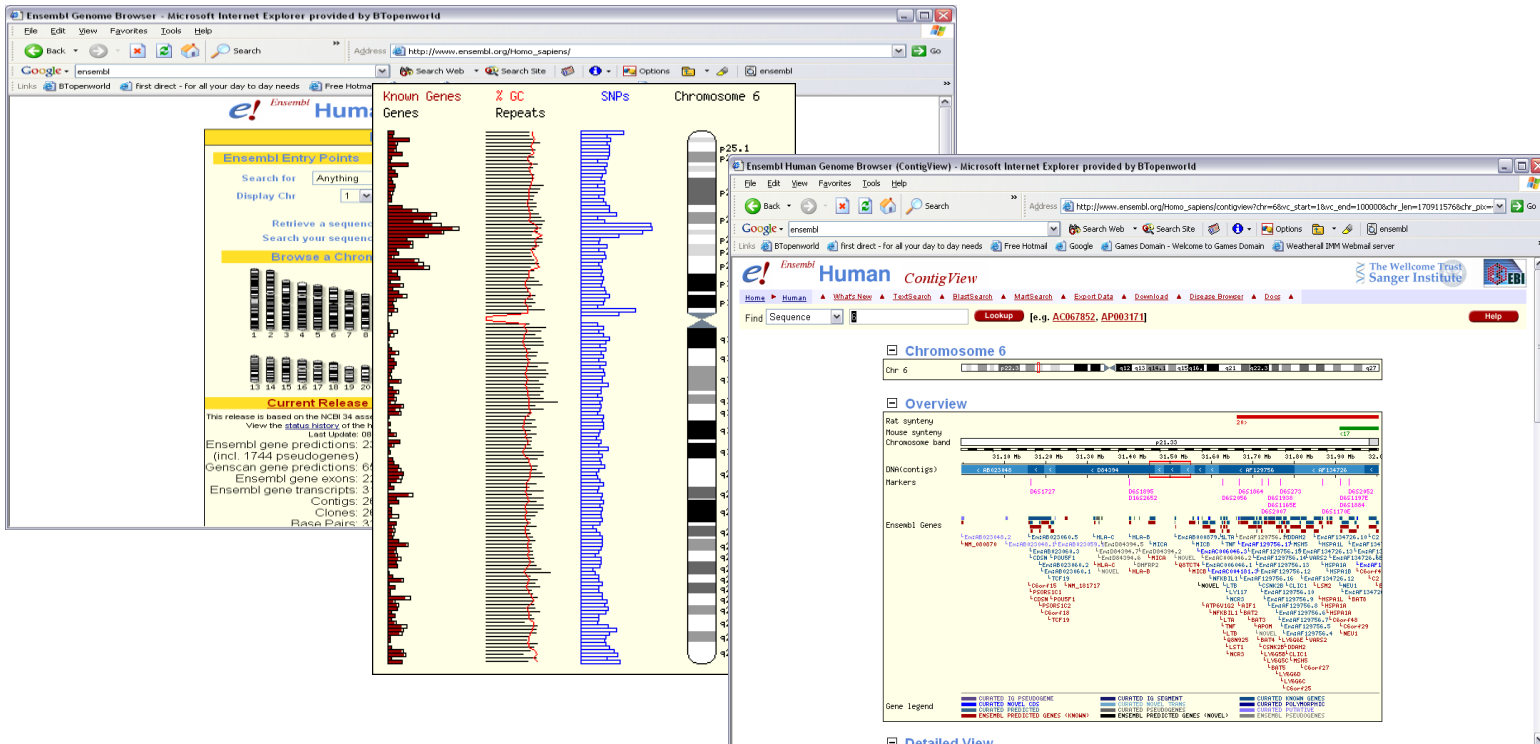
Annotate

- Results of raw gene analysis are FEATURES
- Integration of features, biological rules and knowledge make ANNOTATIONS
- Write these back to the database
- Automated what would take hundreds of scientists to do



Ensembl

- Ensembl Genome Browser (www.ensembl.org)



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

The screenshot displays the UCSC Genome Browser interface for Human Mar. 2006 Assembly. The main track shows the genomic region chr16 (p13.3) from 1,000,000 to 39,000,000 bp. The interface includes navigation tools (move, zoom in/out) and track selection options (default tracks, hide all, add custom tracks, configure, refresh). The tracks shown include:

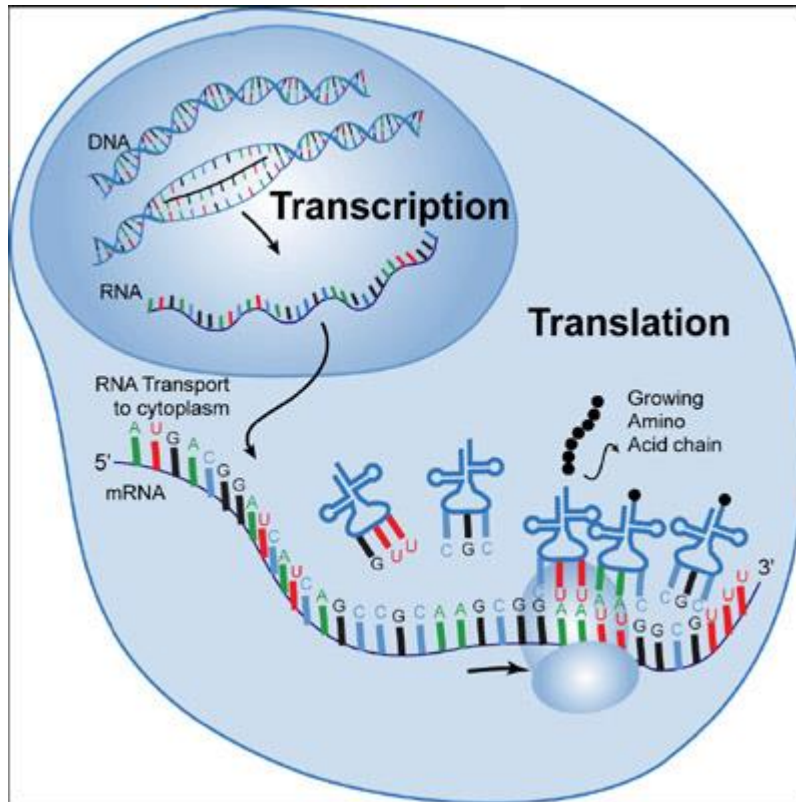
- STS Markers
- RefSeq Genes (e.g., POLR3K, NFG, HBZ, HEM, HBA1, LUC7L, HBE1, RGS11, RKN1, HNF1B, NHE4, DECR2)
- Human mRNAs
- Spliced ESTs
- CpG Islands
- Conservation (Vertebrate Multiz Alignment & Conservation)
- RepeatMasker

The interface also includes a 'Mapping and Sequencing Tracks' section with options for Base Position, Chromosome Band, STS Markers, FISH Clones, Recomb Rate, Map Contigs, Assembly, Gap, Coverage, and BAC End Pairs.

Post Genome (10 years on)

- What do all the genes do?
 - How do they interact?
 - How to cells specialise?
- Junk DNA – is not junk after all...
 - At least 80% genome seems to have function, usually regulation

Gene Expression

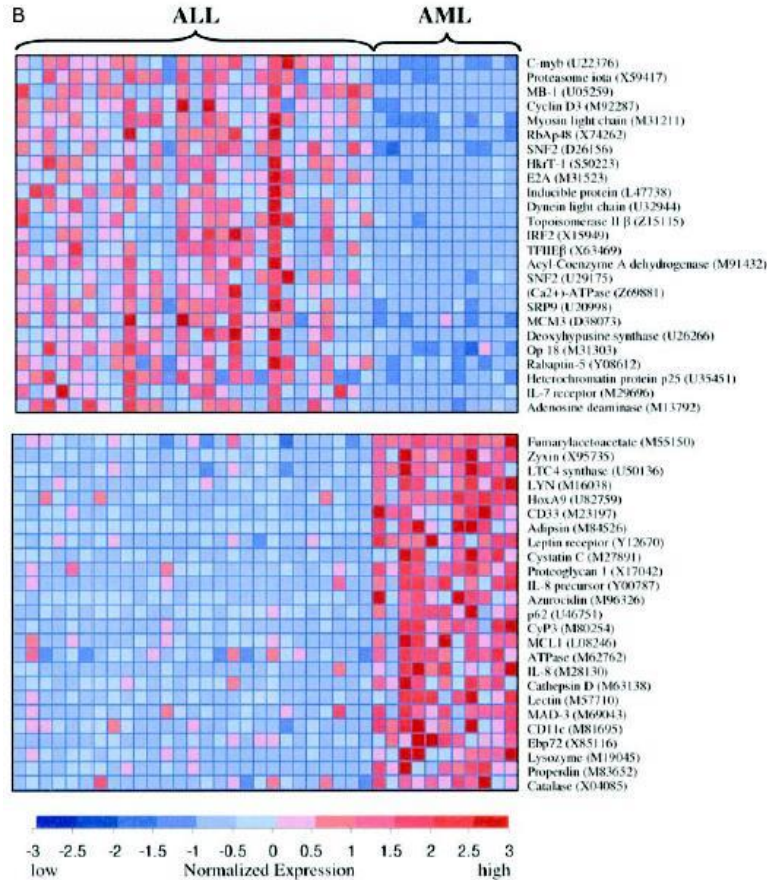


➔ **Proteins**

Microarrays



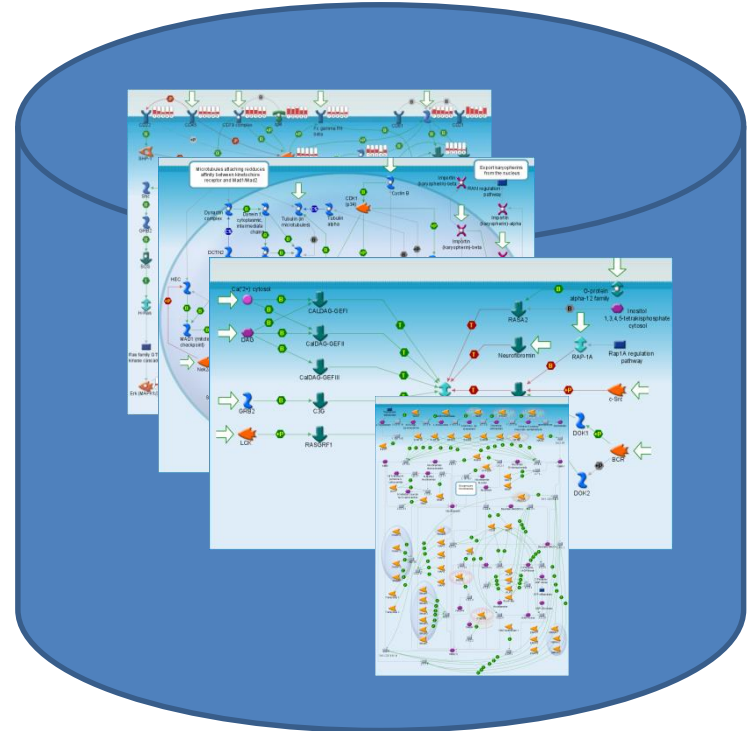
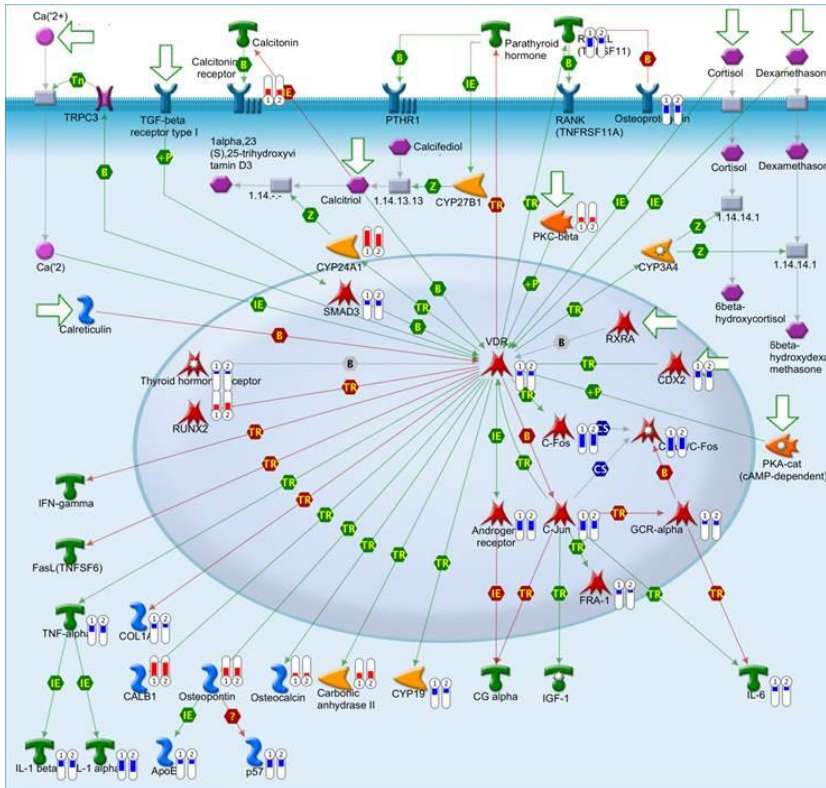
Expression Analysis



Treatment?

Golub et al. (1999)

Functional Annotation Databases



- Metacore
- Ingenuity
- DAVID

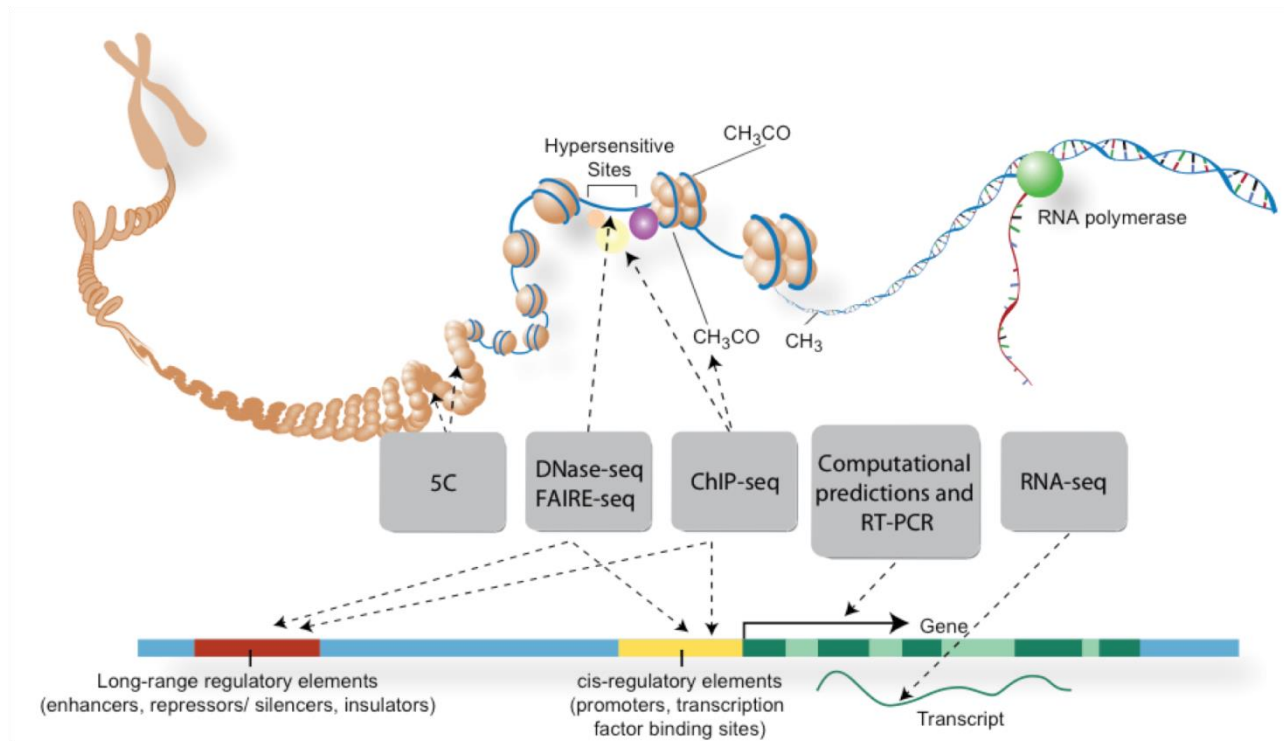
Gene Expression Omnibus (GEO)

The screenshot displays the NCBI Gene Expression Omnibus (GEO) interface. At the top, there are navigation tabs for Series, Samples, Platforms, and DataSets. A search bar indicates 1,163,446 samples. Below this is a table listing various samples with columns for Accession, Title, Sample type, Organism(s), Platform, Series, Supplementary, Contact, and Release date.

A pop-up window titled "Sample GSM952626" provides detailed information for a specific sample:

- 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 10µm 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 10µm 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).

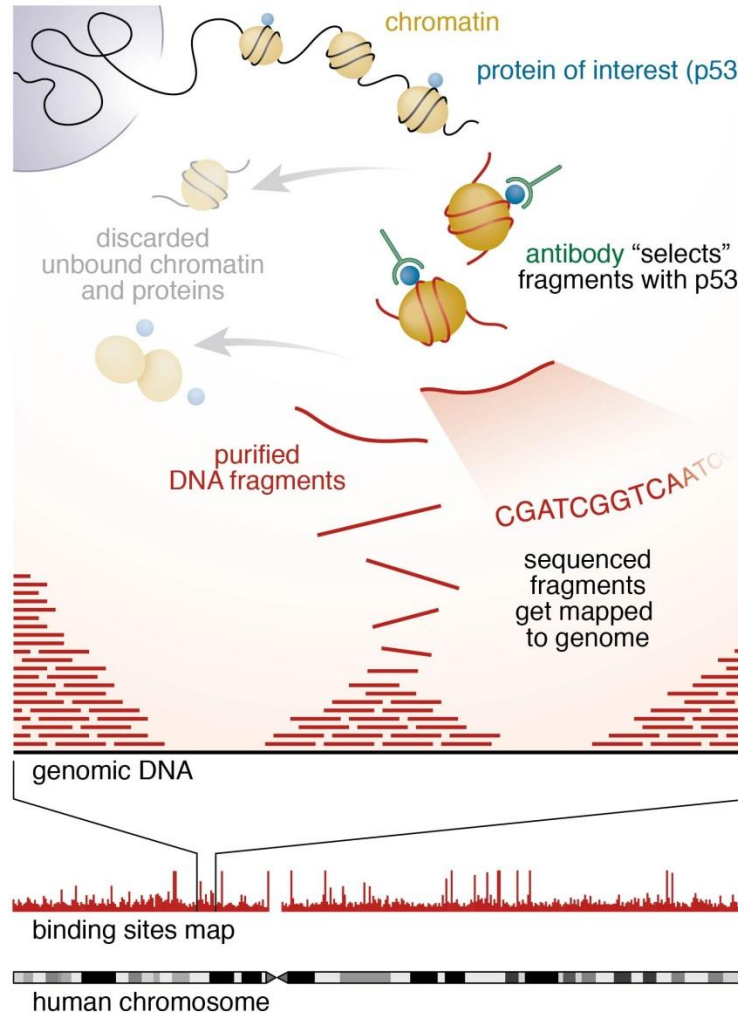
ENCODE (Encyclopedia of DNA Elements)



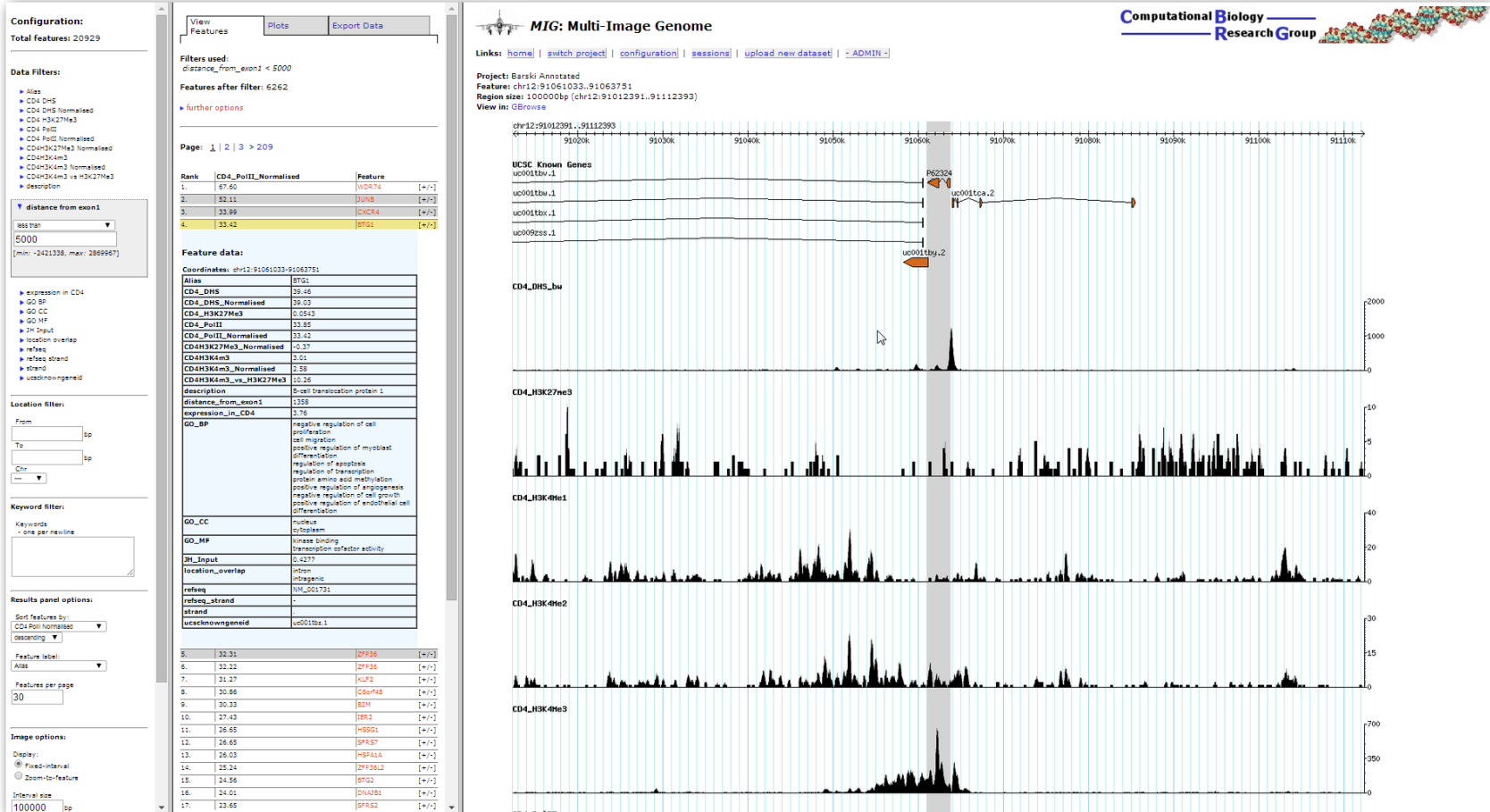
<http://genome.ucsc.edu/ENCODE/>

What Controls Expression?

ChIP-Seq



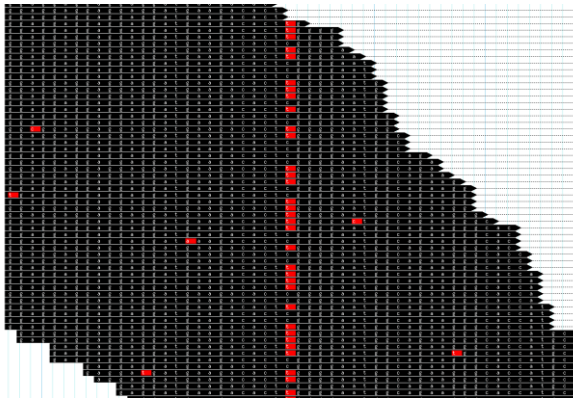
Multi-Image Genome Viewer



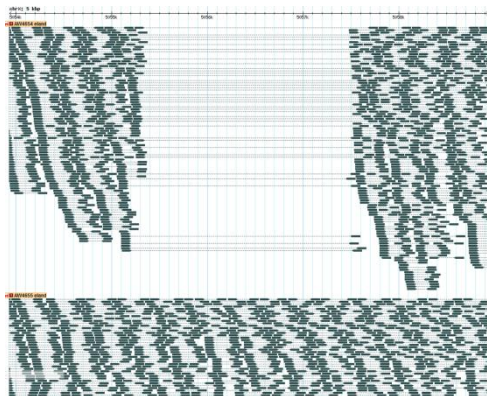
McGowan SJ, Hughes JR, Han ZP, Taylor S MIG: Multi-Image Genome viewer. *Bioinformatics* (2013) 29: 2477-8

DNA Mutations

Single base mutation



Insertion

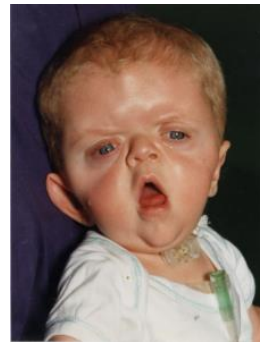
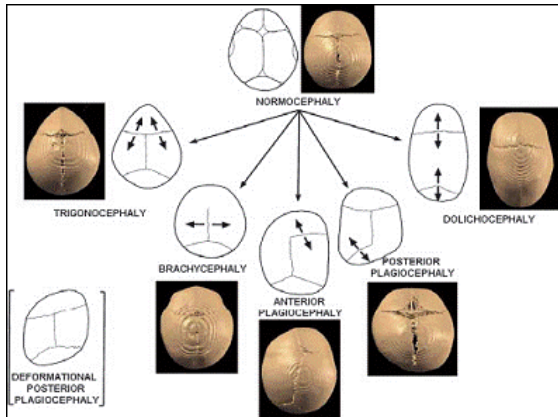


FAULTY GENE

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

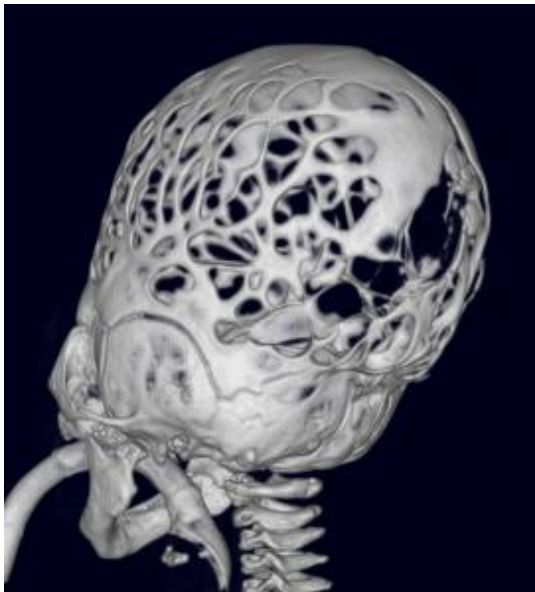
(<http://www.ncbi.nlm.nih.gov/SNP/>)

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 and holding a young child with blonde hair. The child is wearing a white shirt and blue pants. They are both looking towards the camera with slight smiles.

Mark Henderson Science Editor
August 3 2011 12:01AM

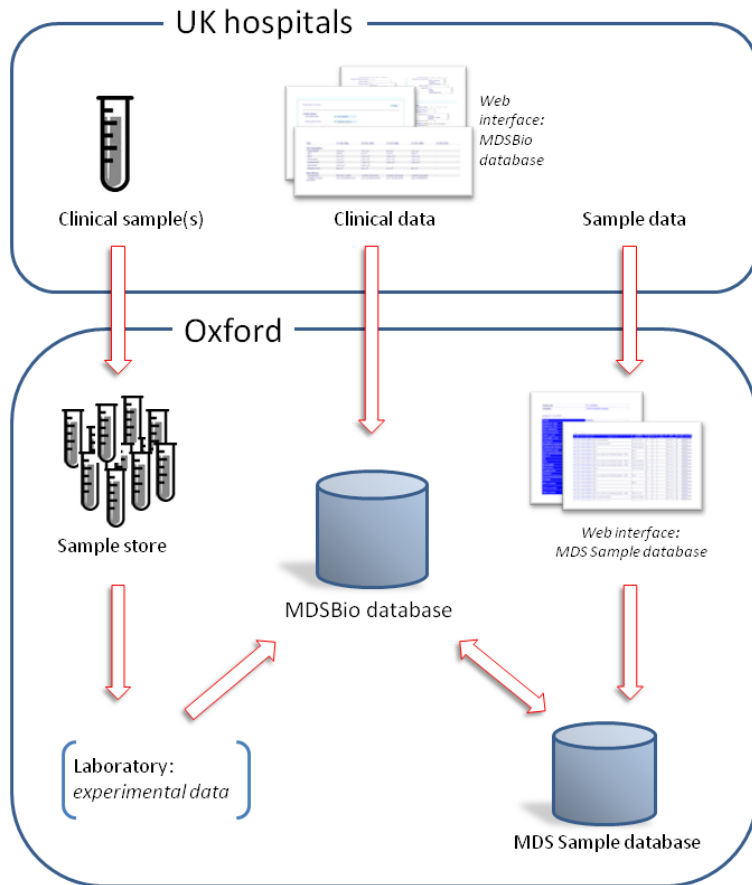
A four-year-old girl has become the first person in Britain to have her entire genetic code read to identify the cause of a disease, in a landmark development that illustrates how personal genetics is changing healthcare.

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

Post a comment

Recommend (4)

Clinical databases : MDSBio



Myelodysplastic Syndrome (MDS)

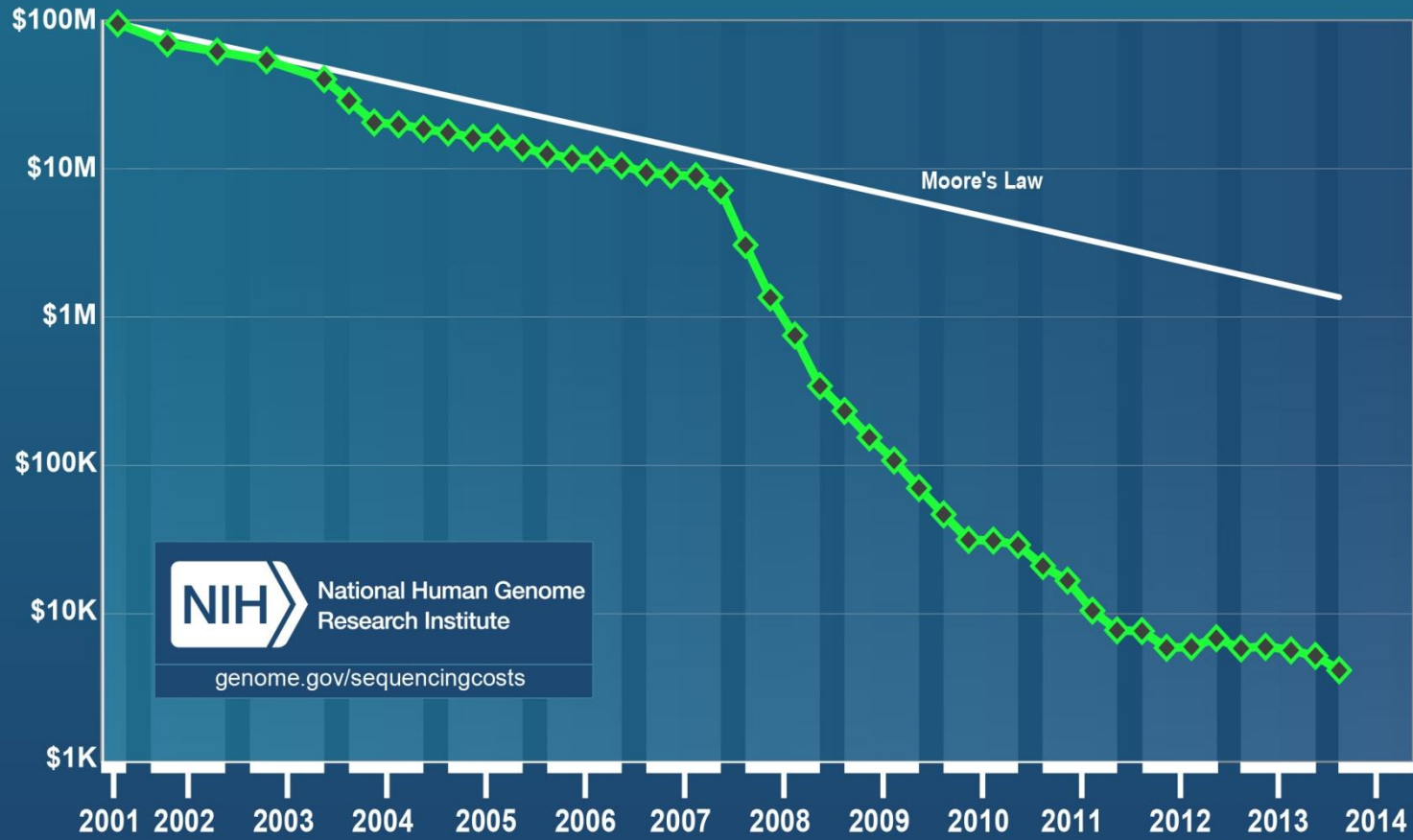
- Damaged bone marrow
- May lead to leukaemia
- Database provides diagnostic and follow-up bone marrow samples and complete follow-up clinical data from patients suspected to have a myeloid disorder to clinician scientists
- CBRG built original
- Infodev built production version

Paresh Vyas, WIMM



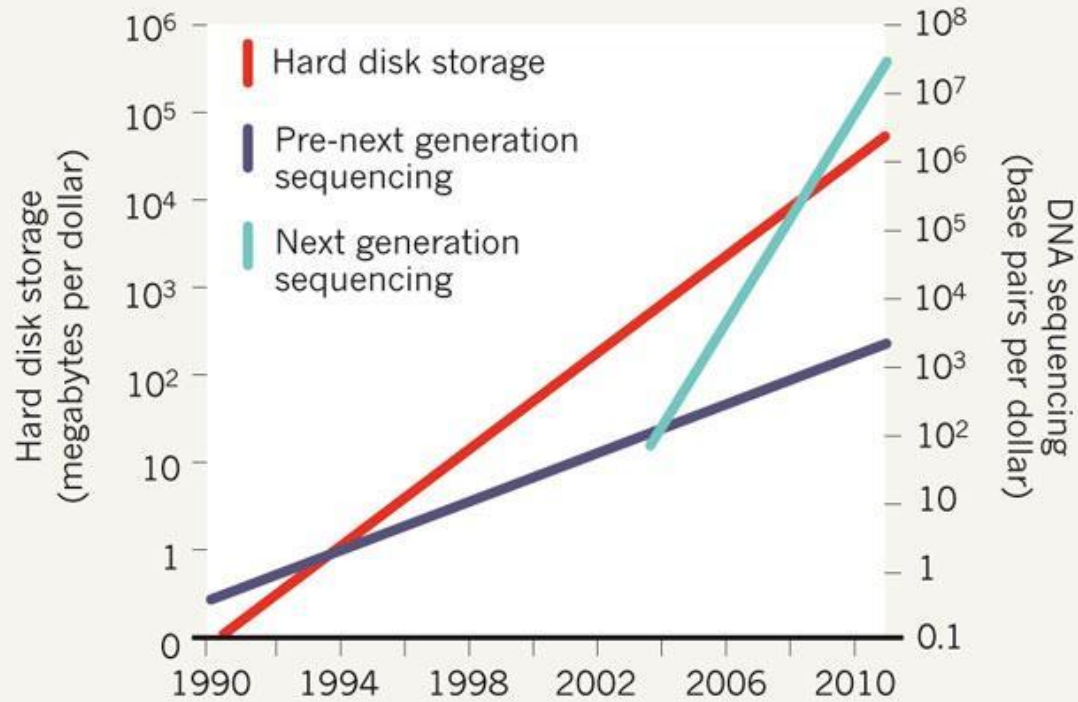
- 100,000 patients with rare inherited disease, common cancers and pathogens from the NHS in England
- Whole Genome Sequencing
- <http://www.genomicsengland.co.uk/>

Cost per Genome



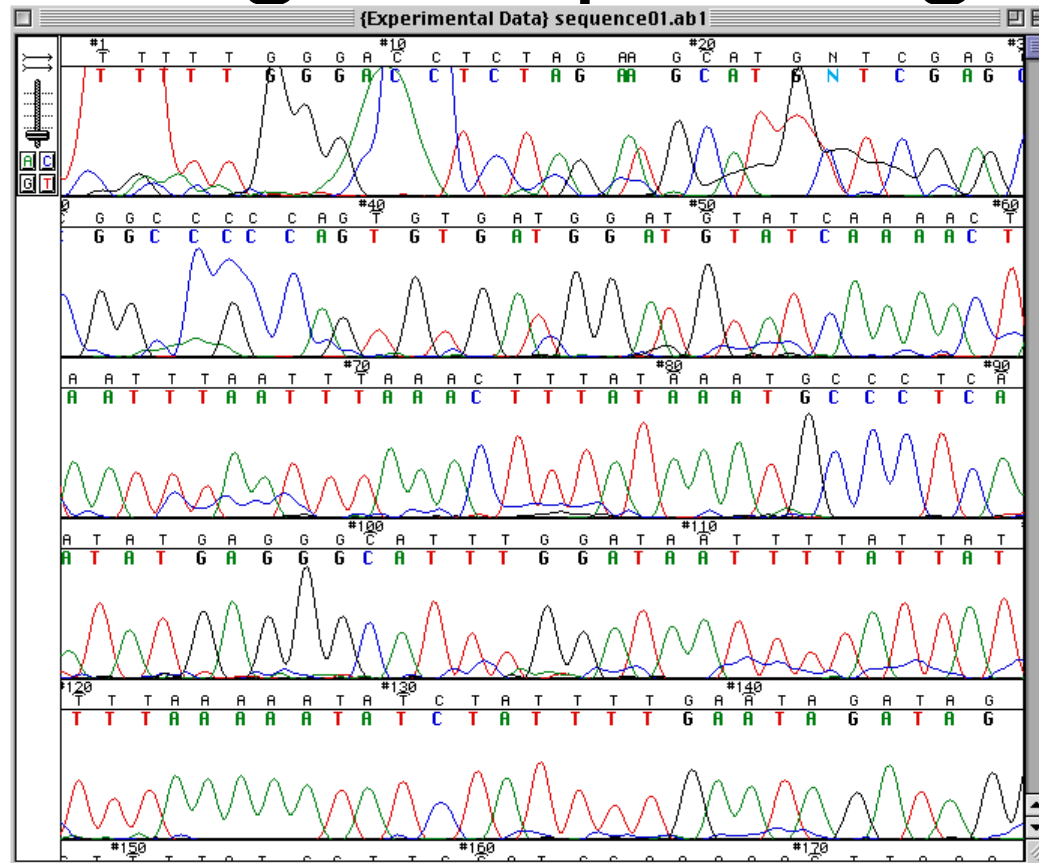
DNA AND CHIPS

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

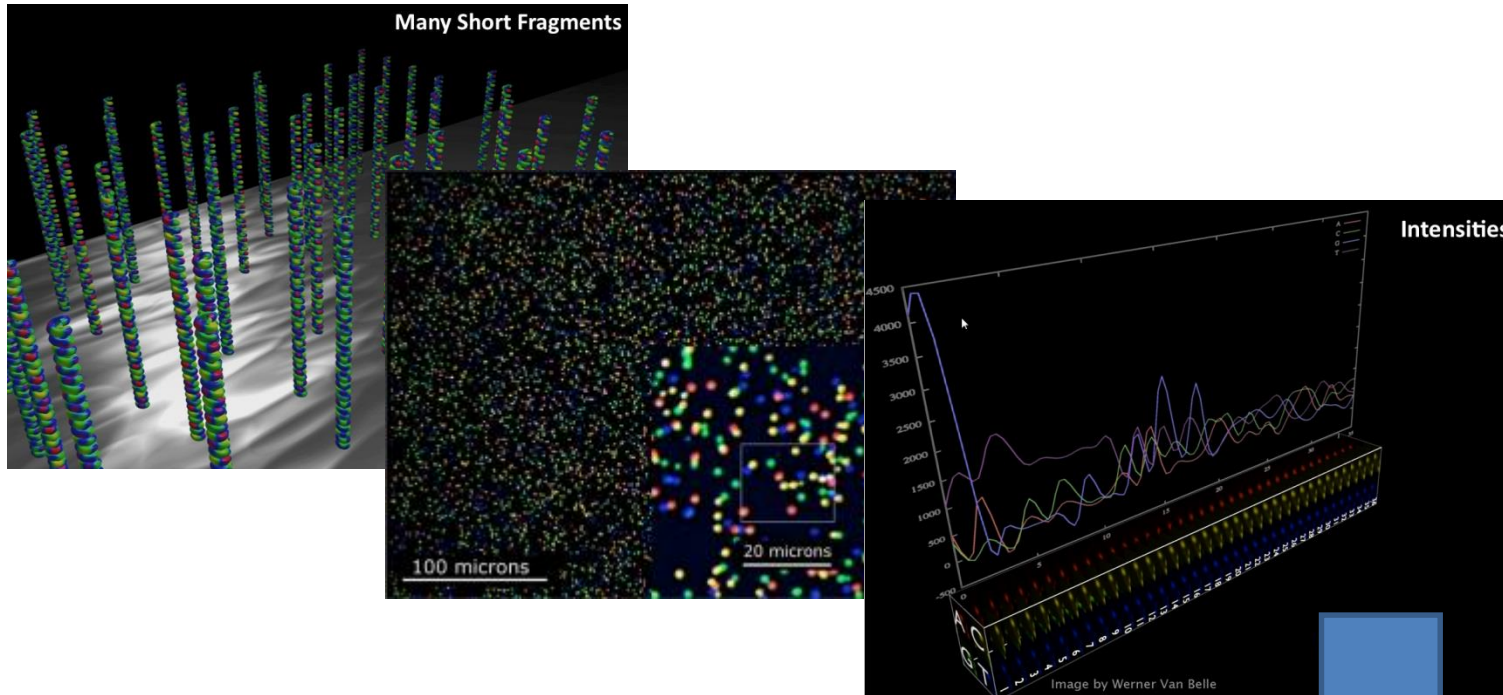


Source: L. D. Stein *Genome Biol.* 11, 207 (2010)

Sanger Sequencing



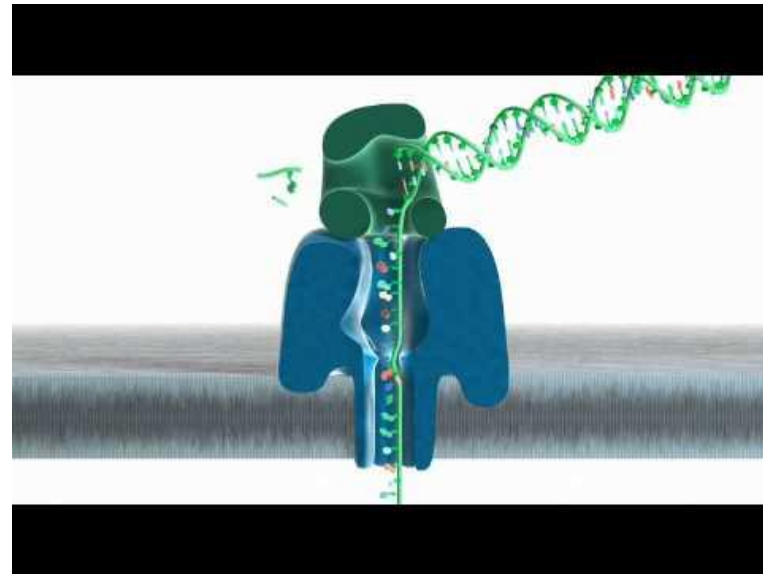
Next Generation Sequencing



```
@HWI-EAS209_0006_FC706VJ:5:58:5894:21141#ATCAG/1  
TTAATTGGTAAATAAATCTCCTAATAGCTTAGATNTTACCTNNNNNNNNNTAGTTTCTTGAGATTGTTGGGGAGACATTTTGTGATTGCCTTGAT  
+HWI-EAS209_0006_FC706VJ:5:58:5894:21141#ATCAG/1  
efcffffcfefffcfffffdff`feed]`_Ba^_[YBBBBBBBBBRTT\]] [dddd`ddd^dddadd^BBBBBBBBBBBBBBBBBBBBBBBB
```

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

Nanopore sequencing

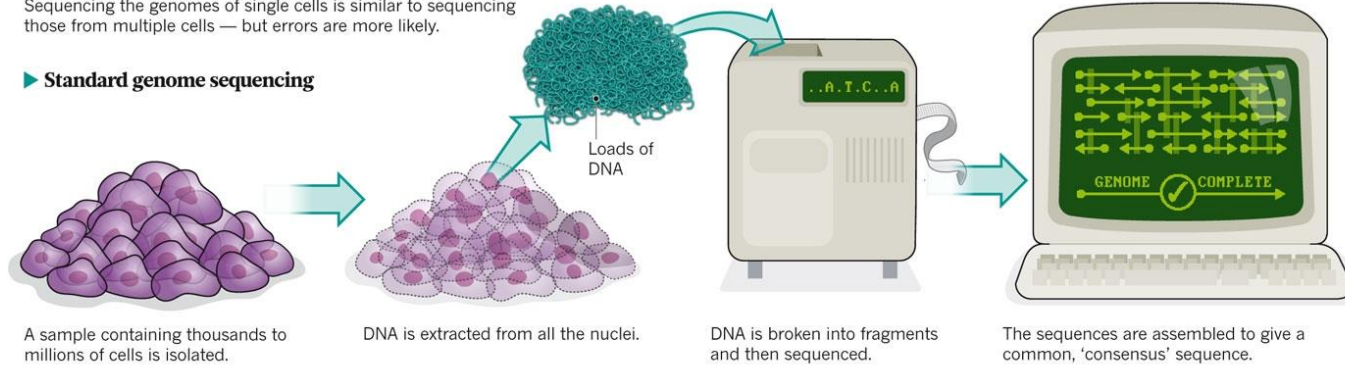


Single Cell Sequencing

ONE GENOME FROM MANY

Sequencing the genomes of single cells is similar to sequencing those from multiple cells — but errors are more likely.

► Standard genome sequencing



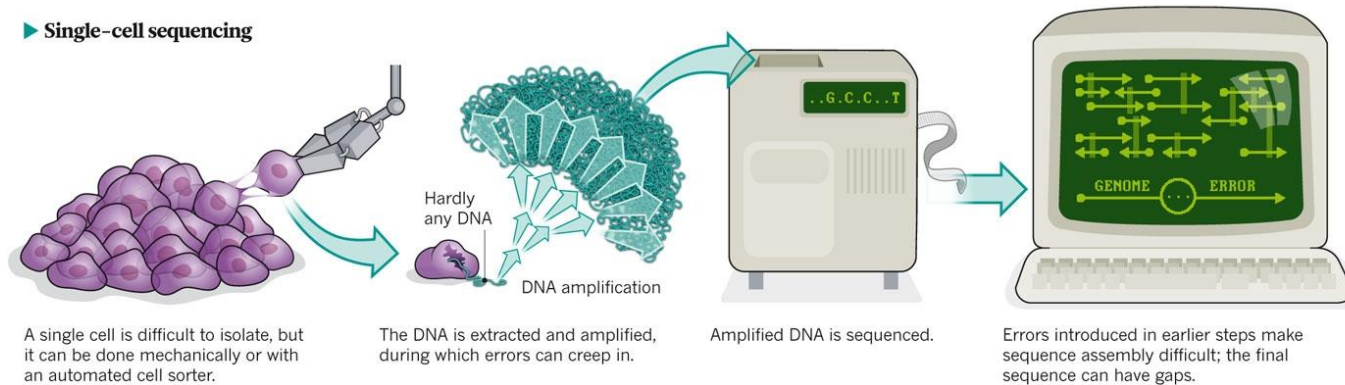
A sample containing thousands to millions of cells is isolated.

DNA is extracted from all the nuclei.

DNA is broken into fragments and then sequenced.

The sequences are assembled to give a common, 'consensus' sequence.

► Single-cell sequencing



A single cell is difficult to isolate, but it can be done mechanically or with an automated cell sorter.

The DNA is extracted and amplified, during which errors can creep in.

Amplified DNA is sequenced.

Errors introduced in earlier steps make sequence assembly difficult; the final sequence can have gaps.

<http://www.nature.com/news/single-cell-sequencing-jpg-7.7203?article=1.11710>

Data volumes

- Human genome
 - 3GB
- Human Brain
 - 10TB
- CBRG Servers
 - 150TB
- Large Hadron Collider
 - 200 PB
- Amount of data stored to date by man
 - 295 Exabytes
 - 30 x number of insects on earth
- All snowflakes that fall on earth per year
 - 1 Yottabyte

BIT	=	A BINARY DIGIT SET TO EITHER A 1 OR 0
BYTE	=	8 BITS
KB	KILOBYTE	= 1,000 BYTES
MB	MEGABYTE	= 1,000,000 BYTES
GB	GIGABYTE	= 1,000,000,000 BYTES
TB	TERABYTE	= 1,000,000,000,000 BYTES
PB	PETABYTE	= 1,000,000,000,000,000 BYTES
EB	EXABYTE	= 1,000,000,000,000,000,000 BYTES
ZB	ZETTABYTE	= 1,000,000,000,000,000,000,000 BYTES
YB	YOTTABYTE	= 1,000,000,000,000,000,000,000,000 BYTES

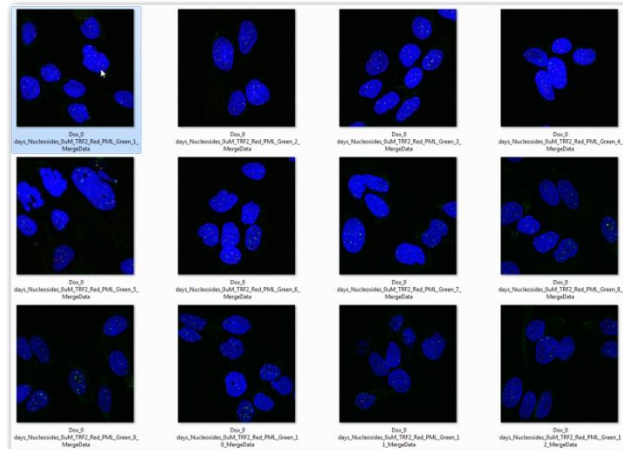
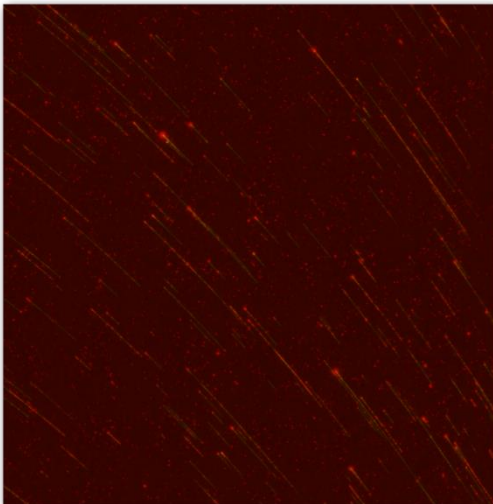
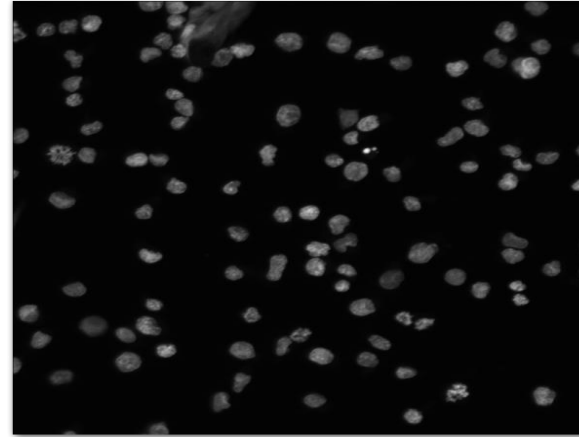
Source: http://c179631.r31.cf0.rackcdn.com/info_byte-final.png

High Throughput Imaging

- Golden age of microscopy
 - Breaking the light wavelength barrier
 - Fantastic optical CCD cameras
 - Automated image acquisition
 - Large disk storage systems
- Masses of images to analyse...
- More BIG DATA...

High Throughput Imaging

- Cells
- Nuclei
- DNA



Analysis Bottleneck

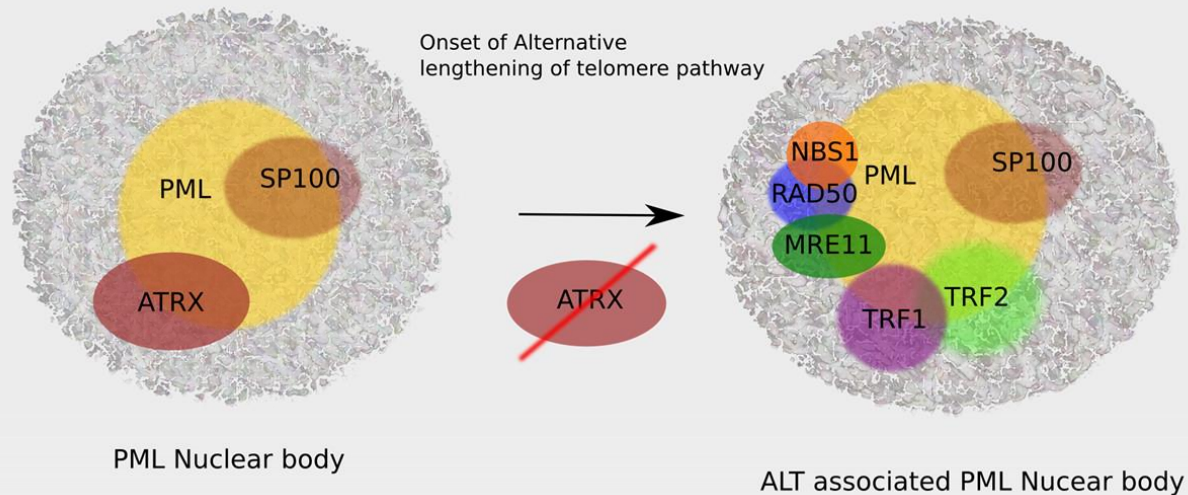
- Generate thousands of analysed images
 - how to check analysis is correct for vast numbers of images?
- Viewing images in context of metadata
- Sorting
- Filtering
- Exporting the subsets for further analysis

ATRX

- X-linked alpha thalassaemia mental retardation (ATR-X) syndrome
- Severe learning difficulties
- Characteristic facial appearance
- Recently associated with a subset of cancers and ALT pathway



ALT – Alternative Lengthening of Telomeres



Colocalisation provides diagnostic marker to assess ALT activity

Colocalisation

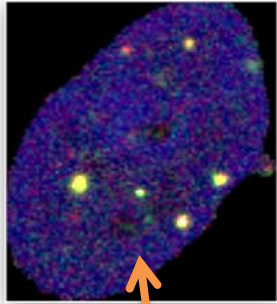
- Spatial overlap between two or more foci in the cell
 - Label proteins with fluorescent dyes
 - Measure overlapping pixels
 - Correlation analysis

Biological Question

- Example
 - To what extent do TRF2 and PML colocalise?

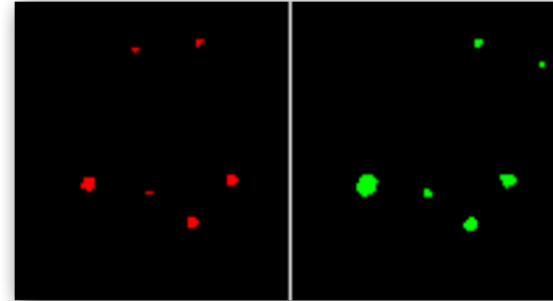
ROI Processing

Segmented Nuclei



ImageJ

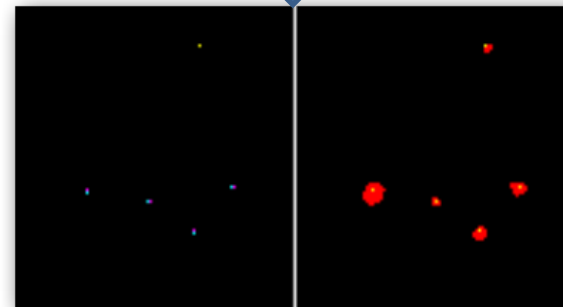
Split Channels / Threshold



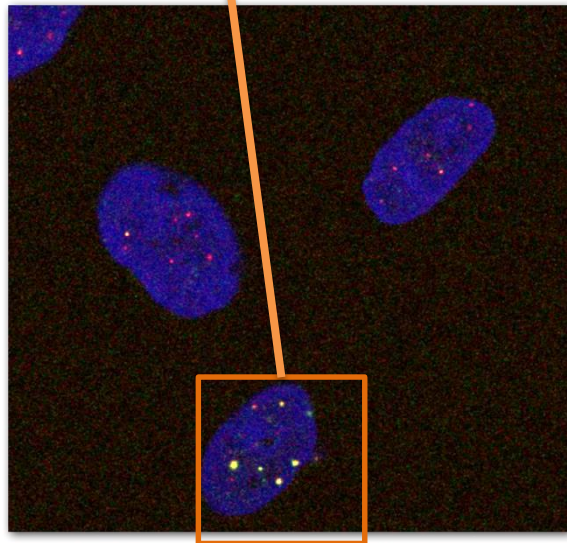
TRF2

PML

JACoP



Distance Based Colocalisation Centre of Mass Colocalisation



Widefield image

Thresholding

The image displays a workflow for thresholding a biological image. It starts with a color merge data image (top left) and a grayscale version (top middle). A 'Threshold' dialog box (top right) shows the process of selecting a threshold value. Below, three binary thresholded images are shown, labeled 'mean', 'otsu', and 'max entropy' (bottom row).

dox_13_days_1_MergeData.TIF_ROI_0.tif (300% 114x111 pixels; RGB; 49K

dox_13_days_1_MergeData.TIF_ROI_0-1.tif (grayscale 114x111 pixels; 8-bit; 12K

Threshold

Mean B&W

Dark background Stack histogram

Auto Apply Reset Set

dox_13_days_1_MergeData.TIF_ROI_0-1.tif (grayscale 114x111 pixels; 8-bit; 12K

dox_13_days_1_MergeData.TIF_ROI_0-1.tif (grayscale 114x111 pixels; 8-bit; 12K

dox_13_days_1_MergeData.TIF_ROI_0-1.tif (grayscale 114x111 pixels; 8-bit; 12K

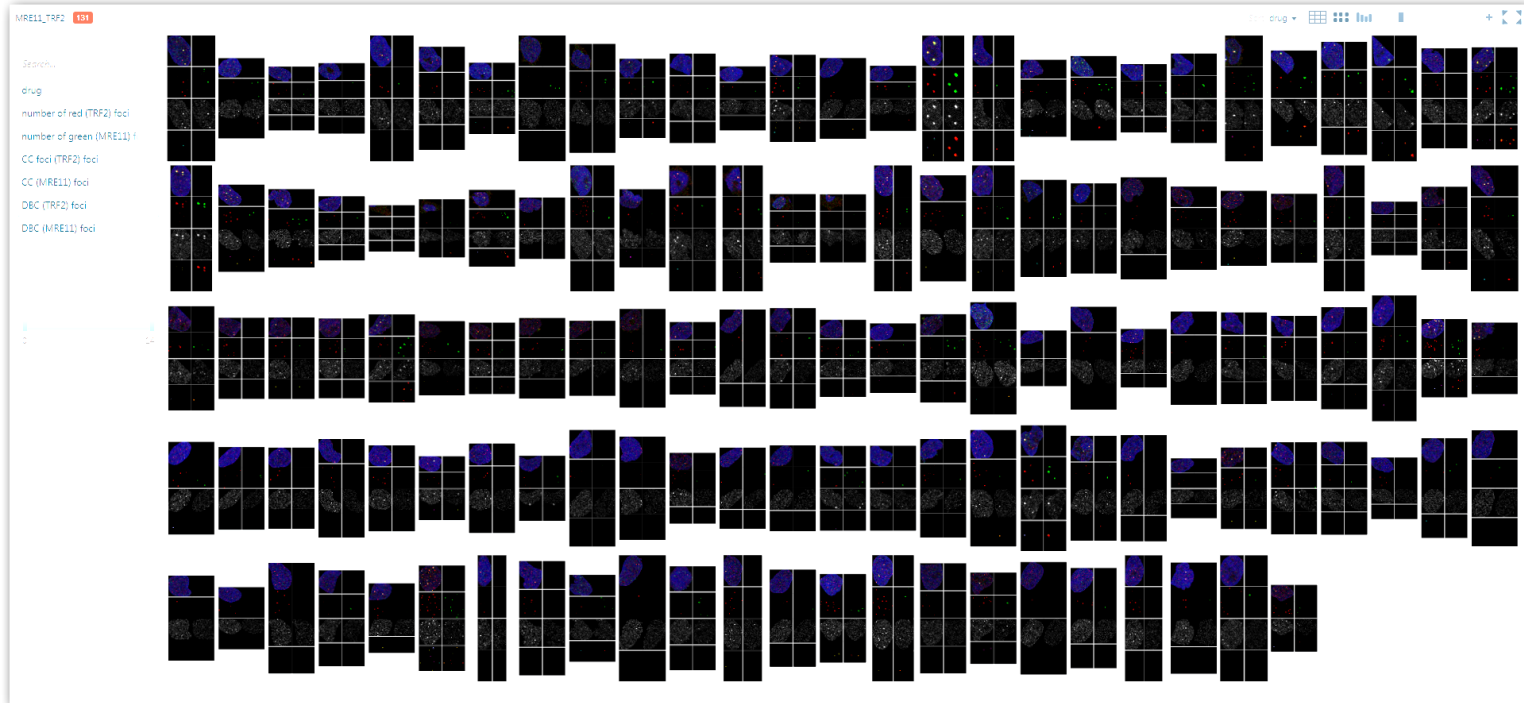
mean

otsu

max entropy

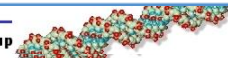
Which is the most accurate?

HTML5 PivotViewer

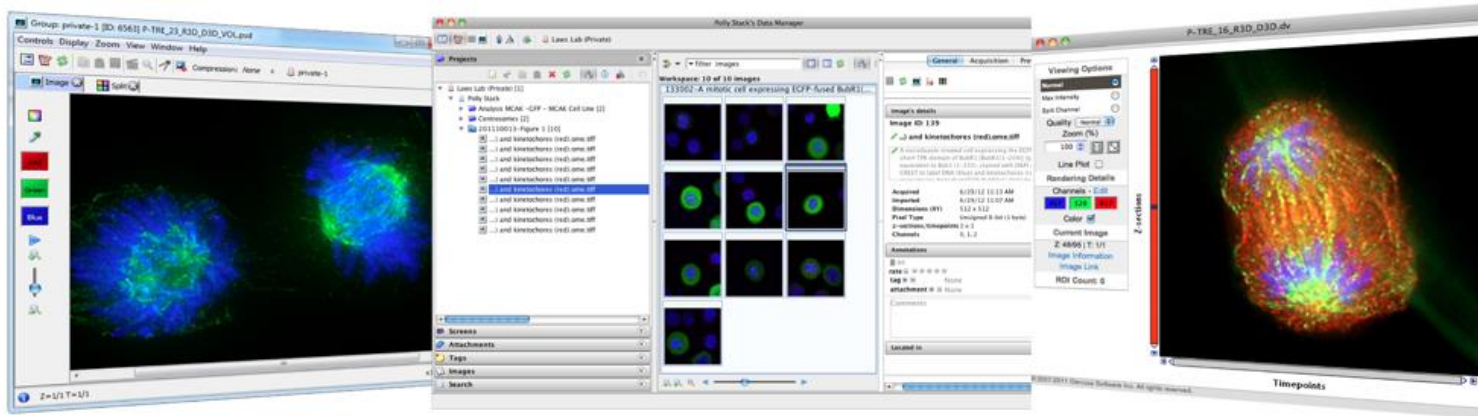


Taylor and Noble (2014), Bioinformatics

[DEMO](#)



OMERO Image Database



Final Thoughts

- Integration across all forms of data will lead to the most promising new leads in science
- Big data = big false leads
- Powerful tools required to cross query these resources
- Access to raw data is key

Acknowledgments

- CBRG
 - Simon McGowan
 - Zong-Pei Han
- Gibbons Group
 - David Clynes
- Patient Group
 - Jie Zuo
- Wilkie Group
- Coritsu
 - Roger Noble
 - Sam Conway