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?

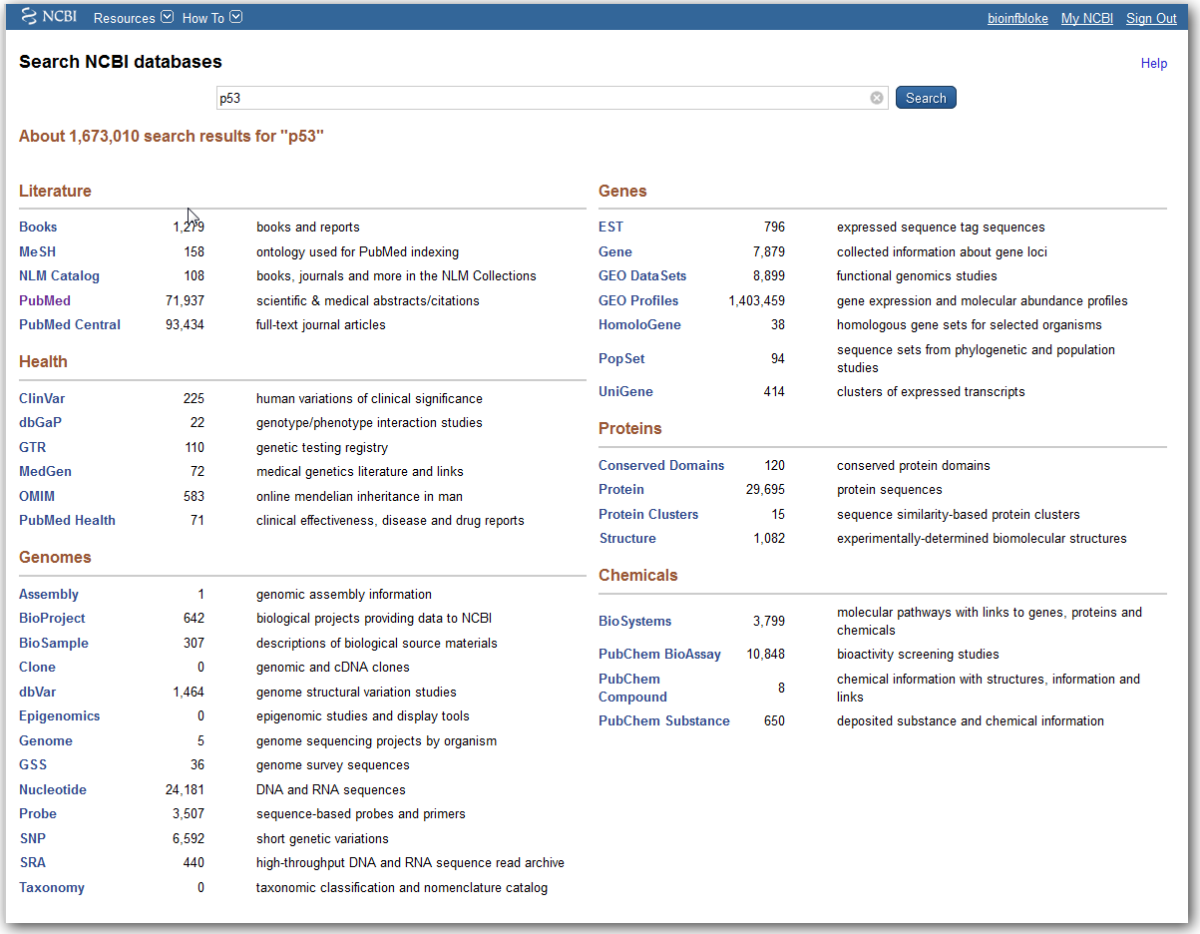
- Find an answer quickly
  - Most *in silico* biology is faster than *in vitro*
- Massive amounts of data to analyse
  - Need to make use of all information
  - Not possible to do analysis by hand
  - Can't organise and store information only using lab note books
  - Automation is key
- However!
  - 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,

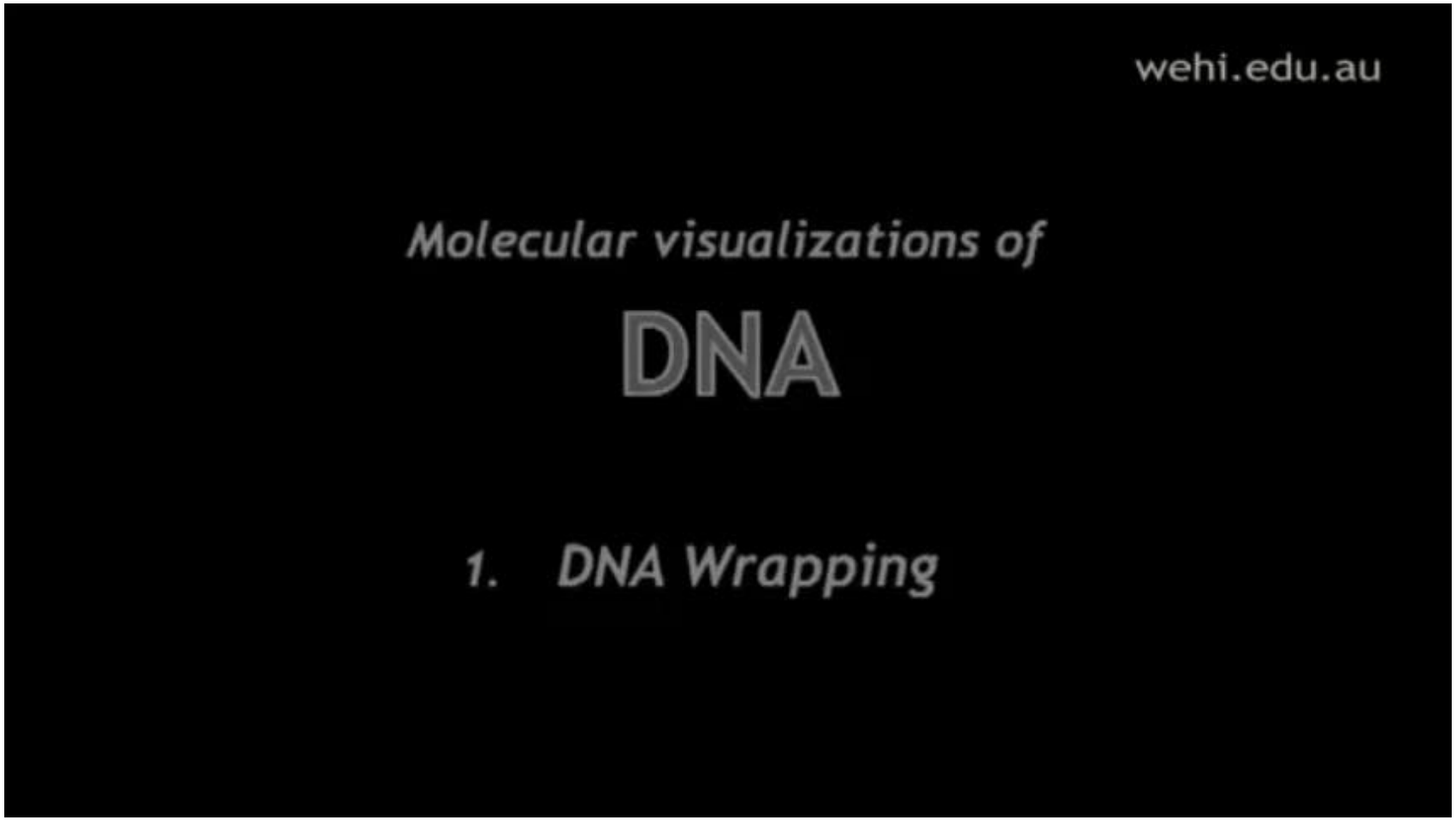
Keyword


NCBI Resources How To
bioinfbloke 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

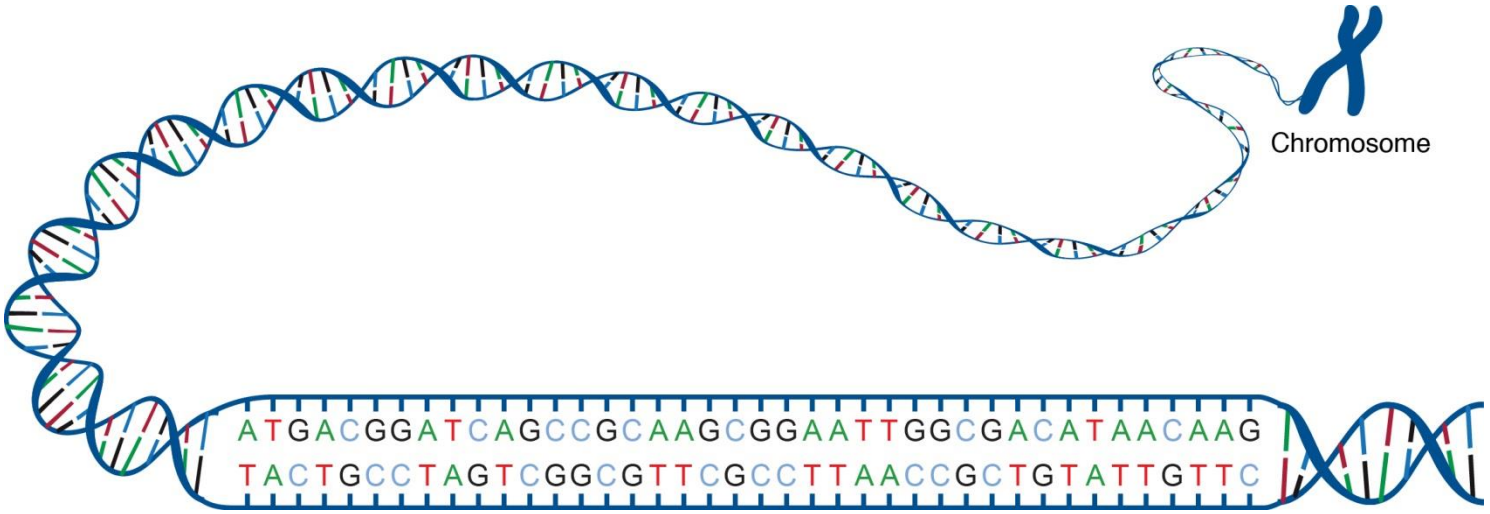
Bioinformatics Tools

- 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



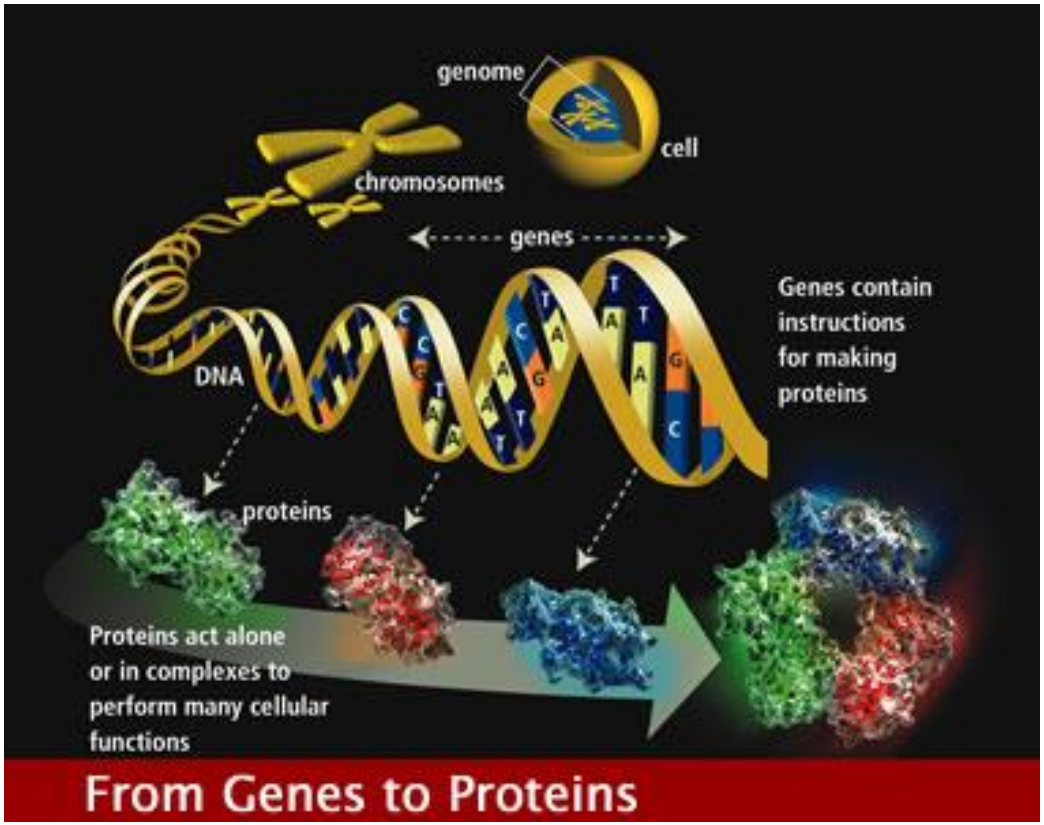
DNA


Chromosome
ATGACGGATCAGCCGCAAGCGGAATTGGCGACATAACAAG
TACTGCCTAGTCGGCGTTCGCCTTAACCGCTGTATTGTTC

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

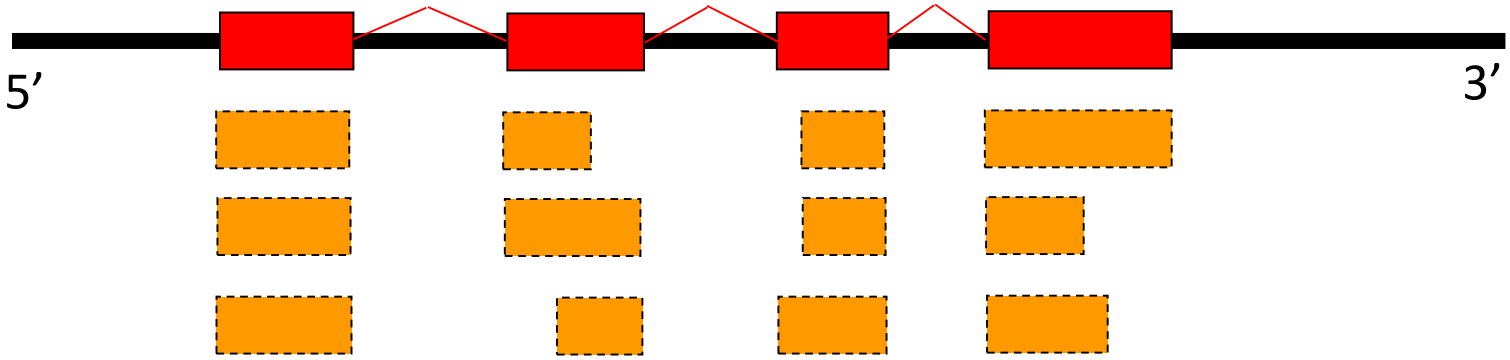
- Human genome is theoretically several long strings totalling 3 billion base pairs
  - Assembled via hundreds of thousands of overlapping units or contigs to make a single consensus sequence
  - Sequences collated using information stored on ABI sequencer
  - Sequence assembly bioinformatics tools used 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…

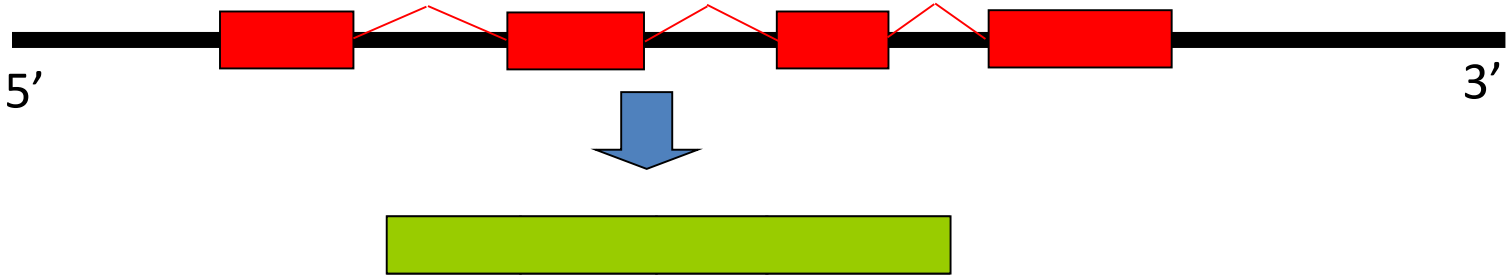
Model Genes



- 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
VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKG
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

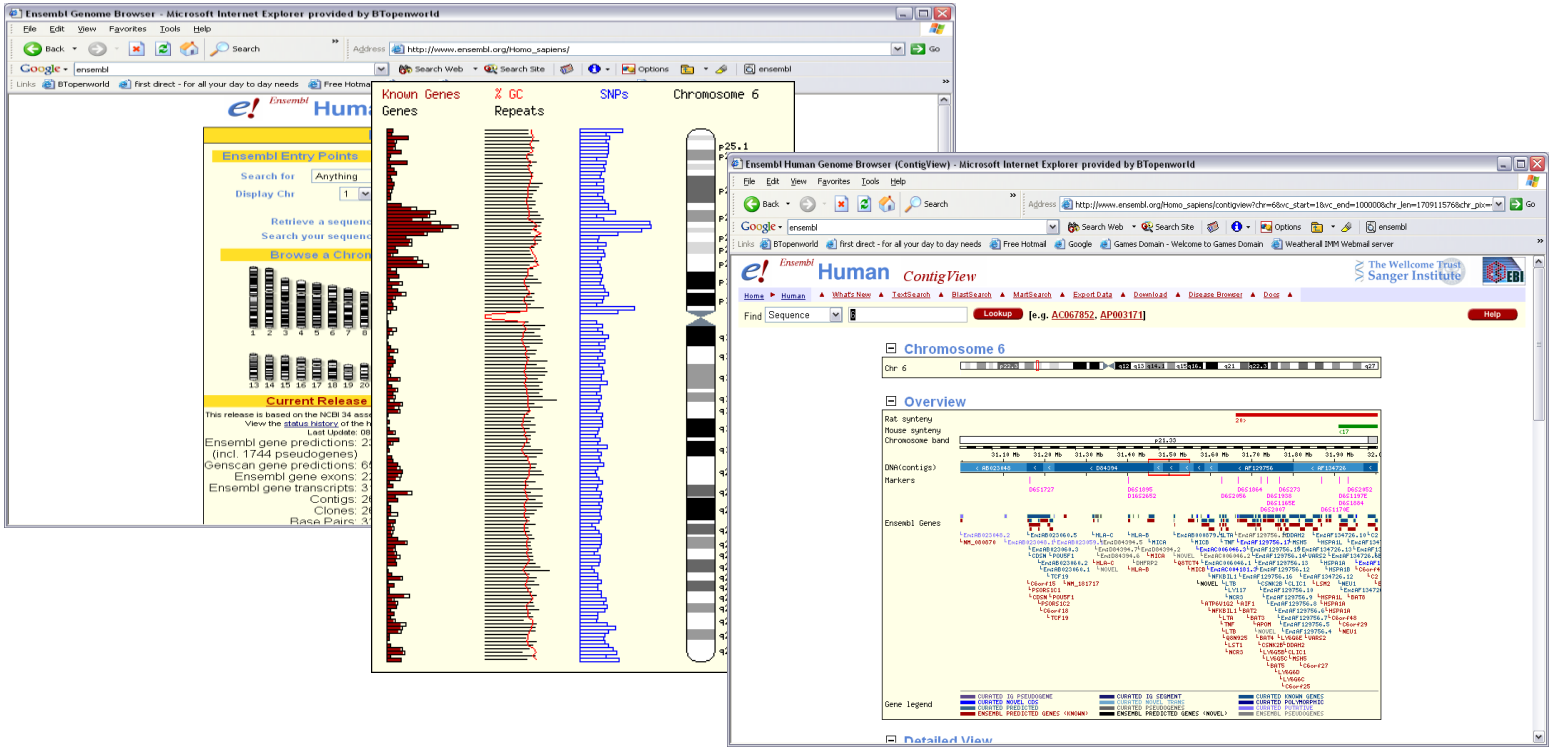
Putative function = Human, Myoglobin

Annotate

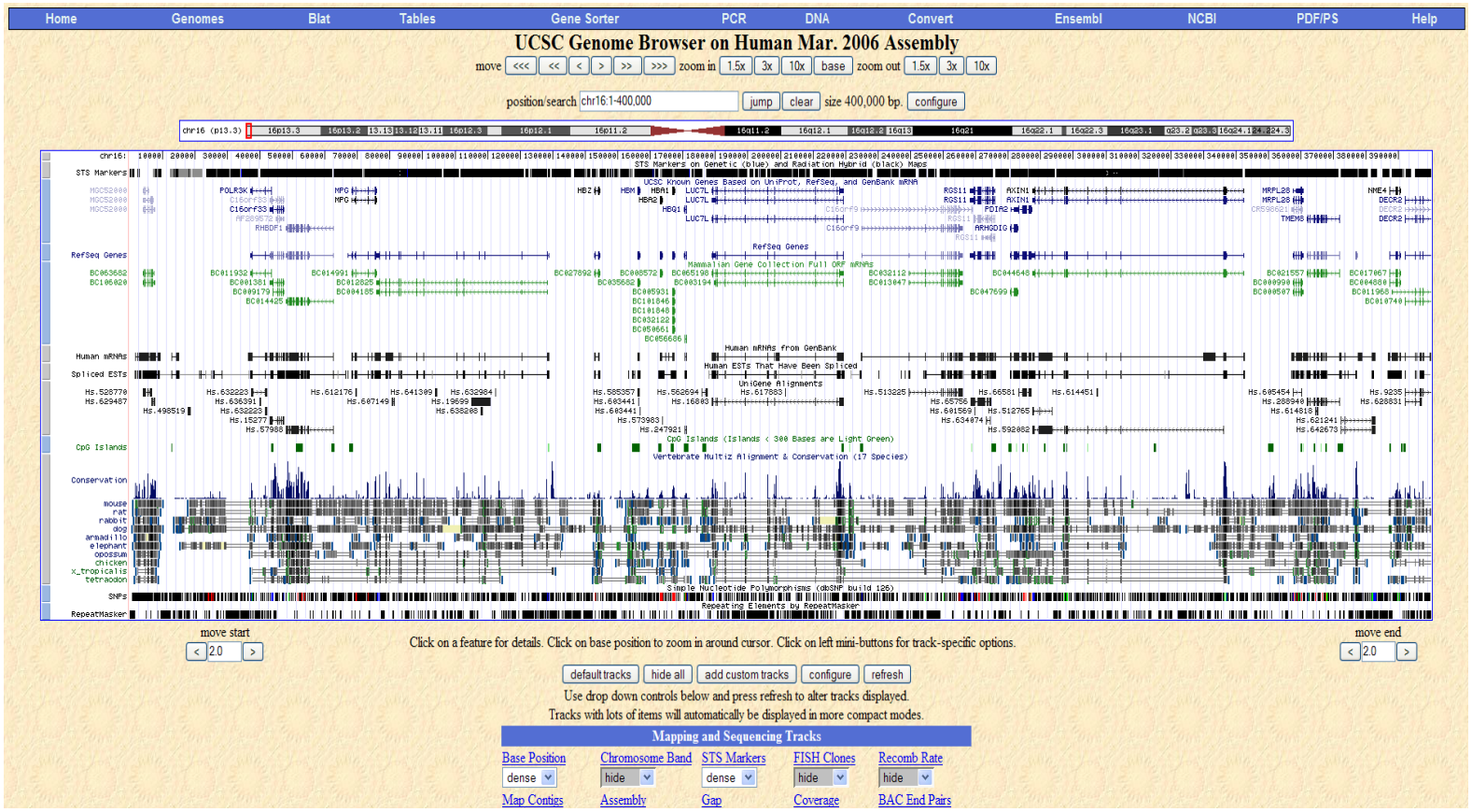
- Results of raw gene analysis are FEATURES
- Integration of features, biological rules and knowledge make ANNOTATIONS
- Write these back to the database
- Automated what would take hundreds of scientists to do

Ensembl

- Ensembl Genome Browser (www.ensembl.org)



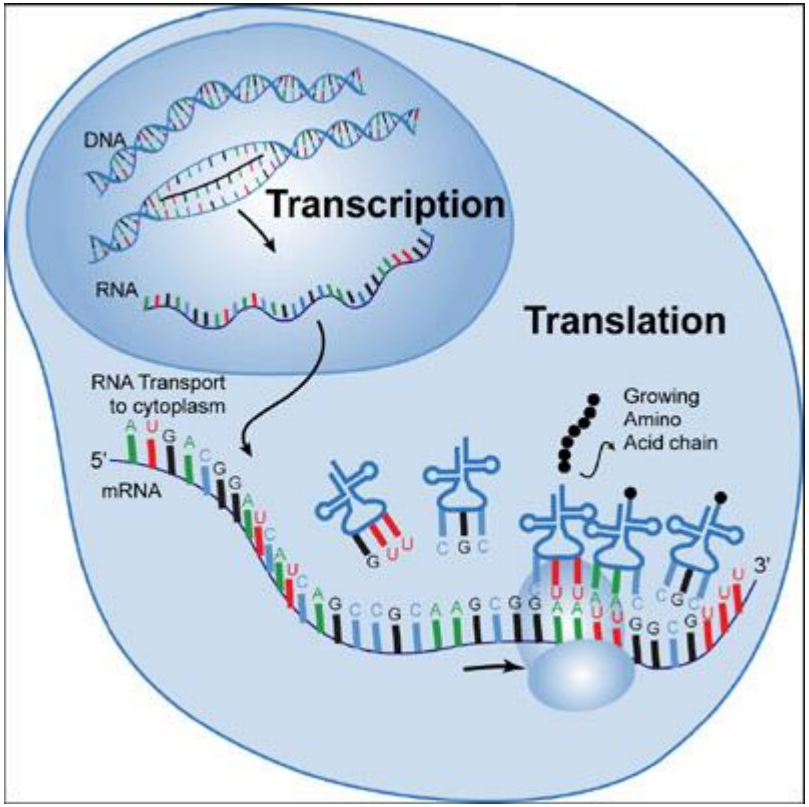
UCSC Genome Browser (http://genome.ucsc.edu/)



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

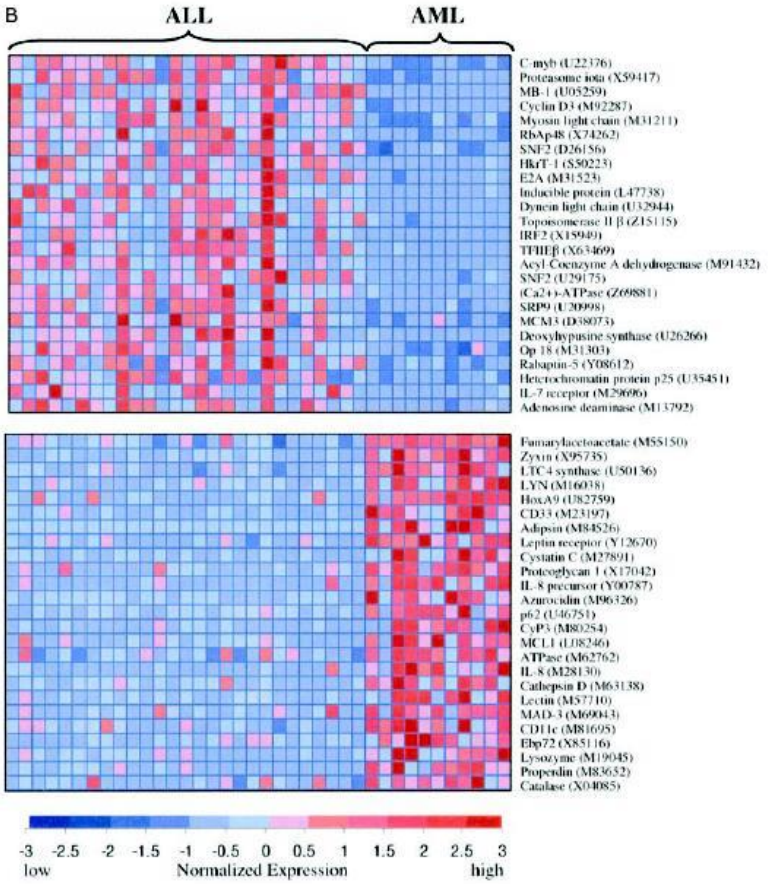
Gene Expression



Proteins

Microarrays

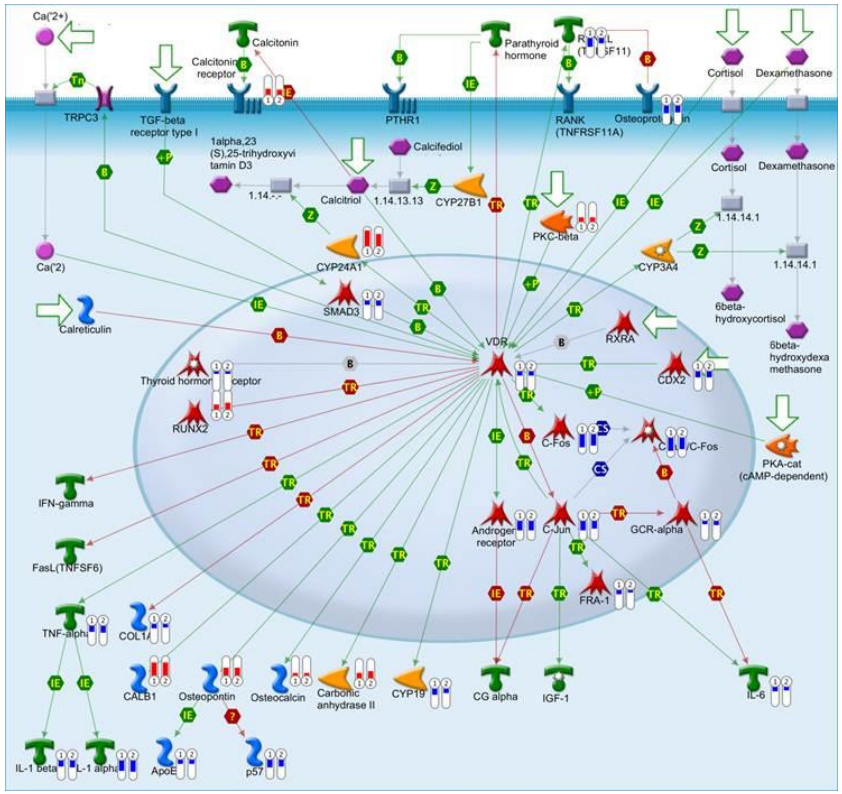
Expression Analysis

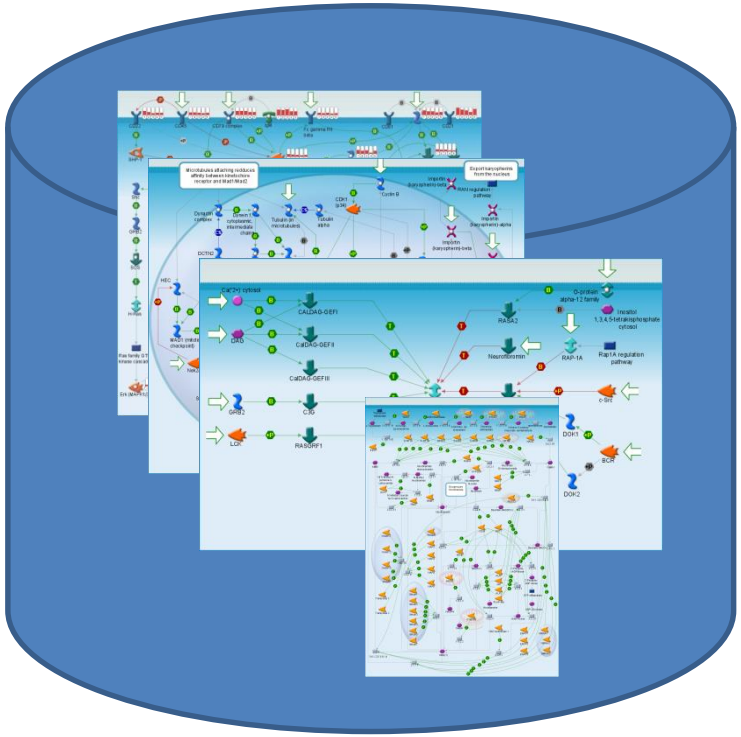


Treatment?

Golub et al. (1999)

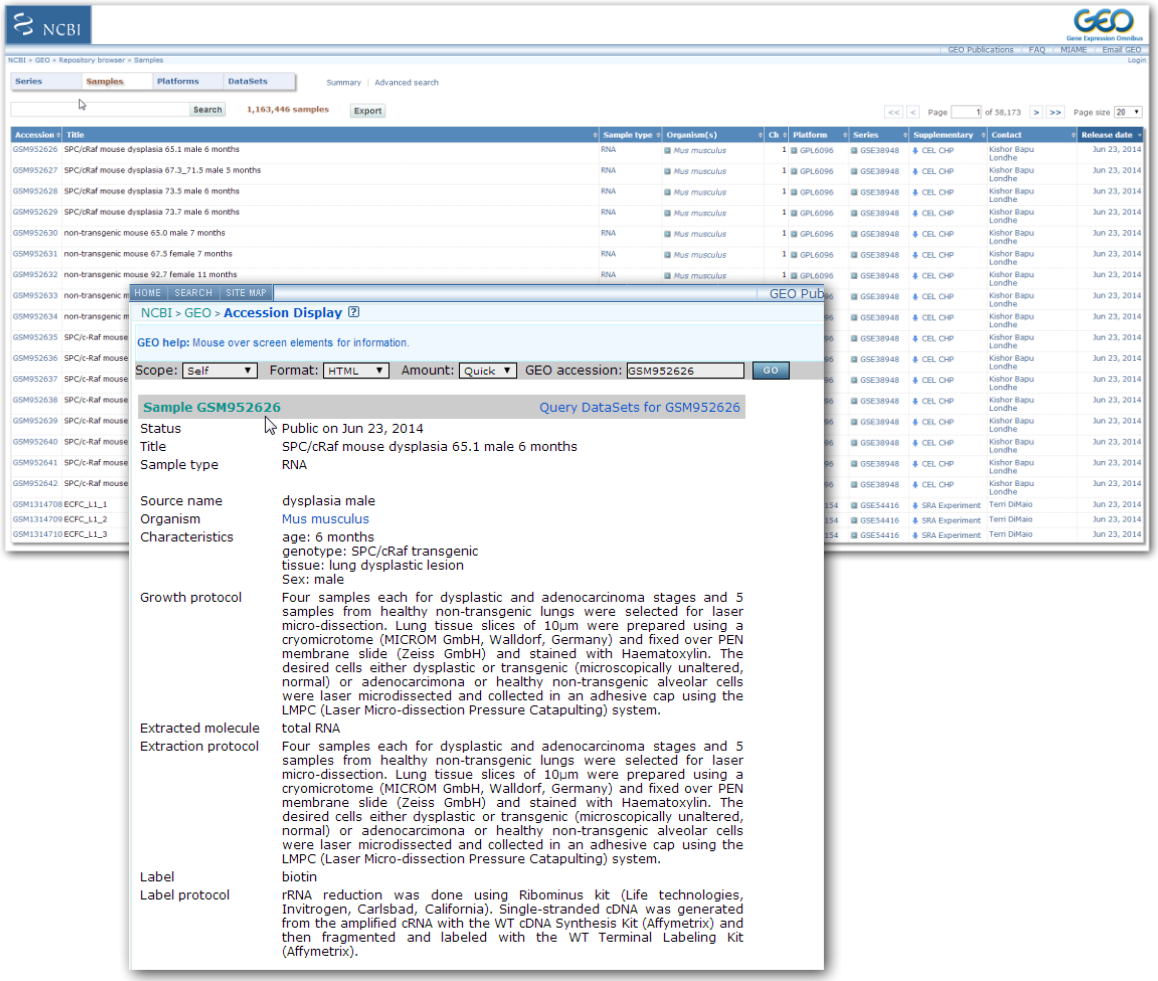
Functional Annotation Databases



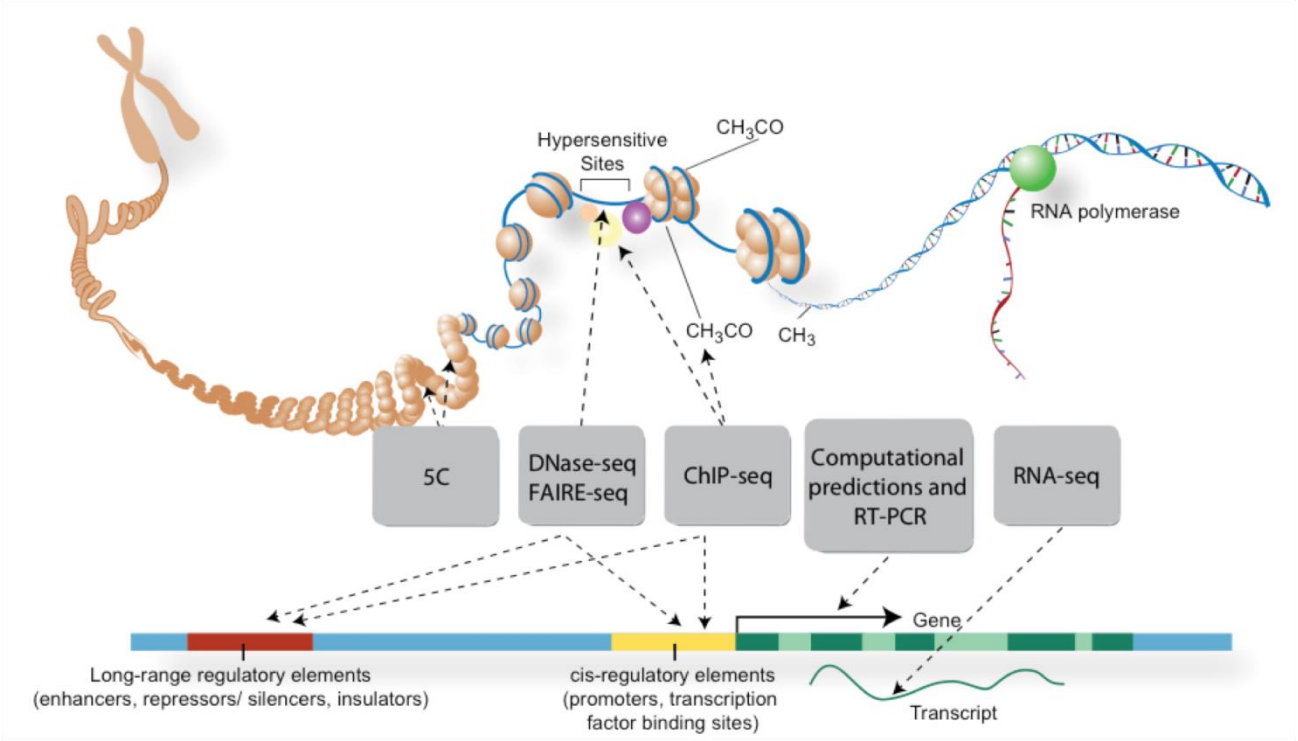


- Metacore
- Ingenuity
- DAVID

Gene Expression Omnibus (GEO)



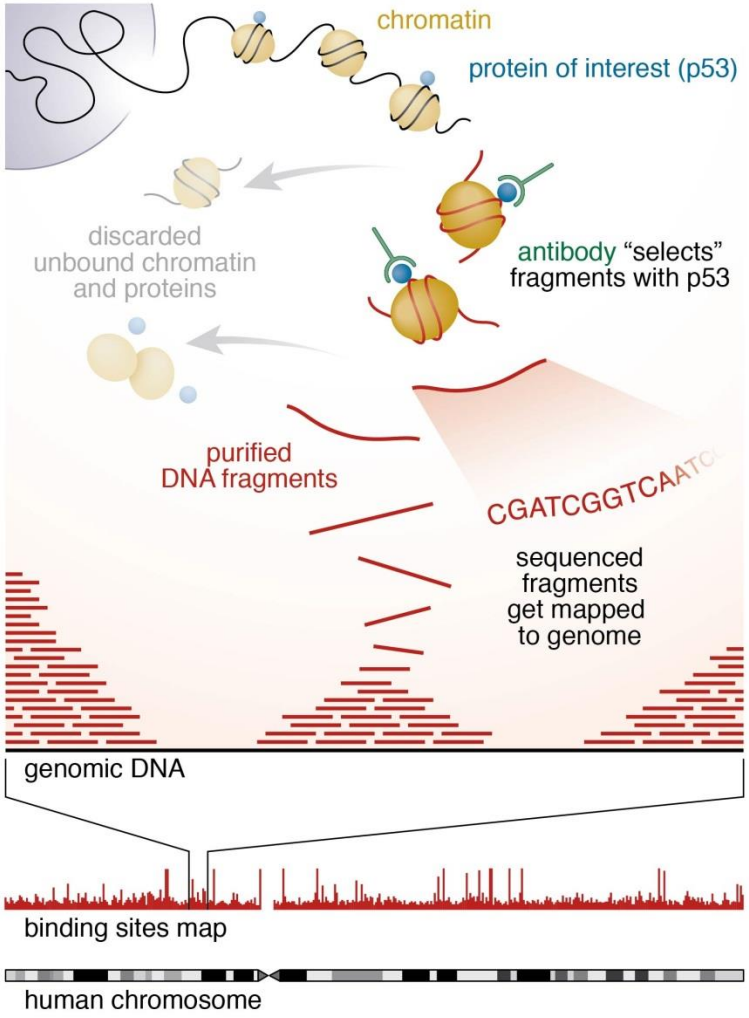
ENCODE (Encyclopedia of DNA Elements)



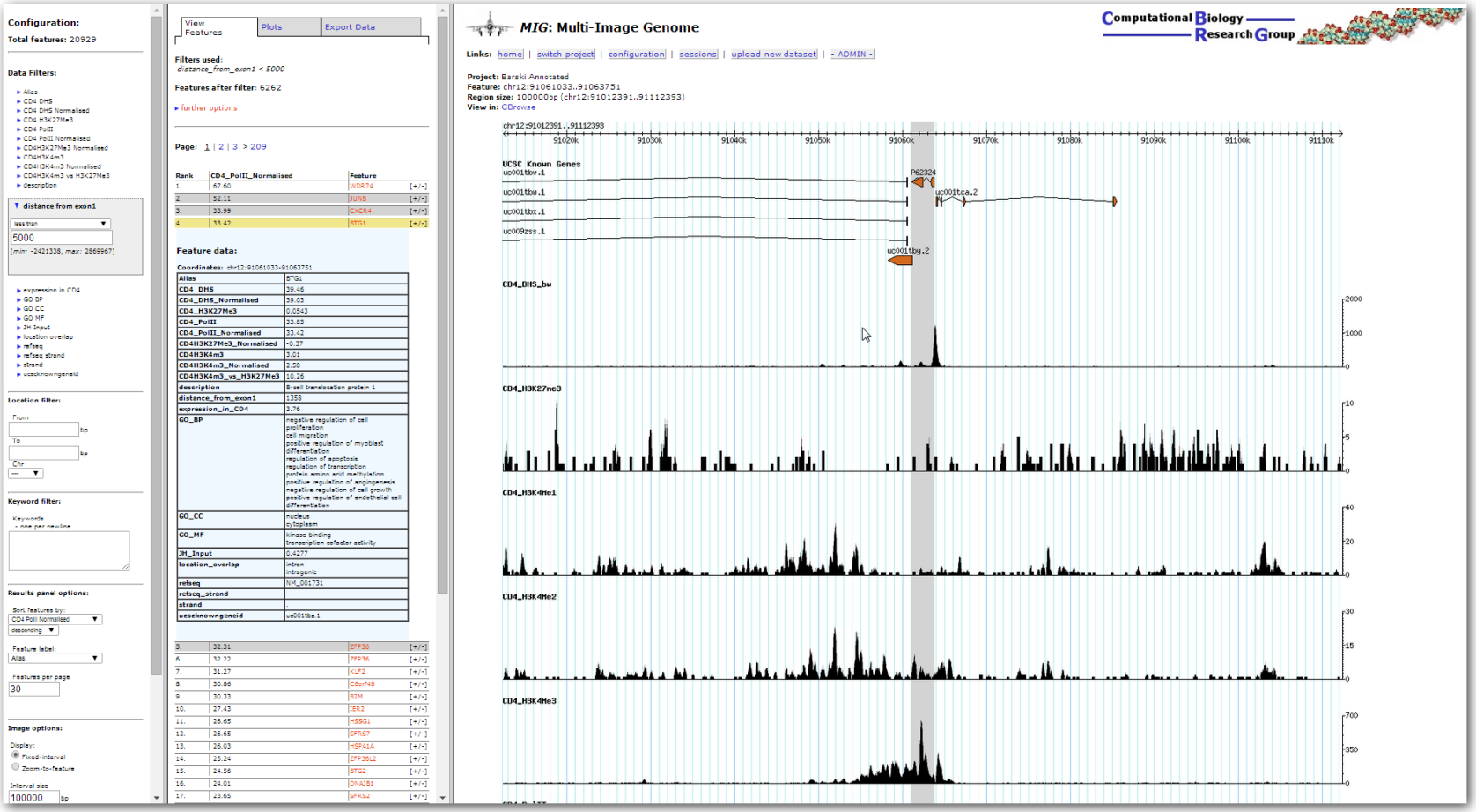
http://genome.ucsc.edu/ENCODE/

What Controls Expression?

ChIP-Seq



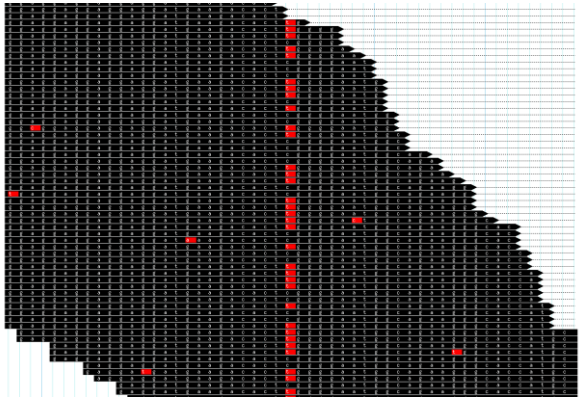
Multi-Image Genome Viewer



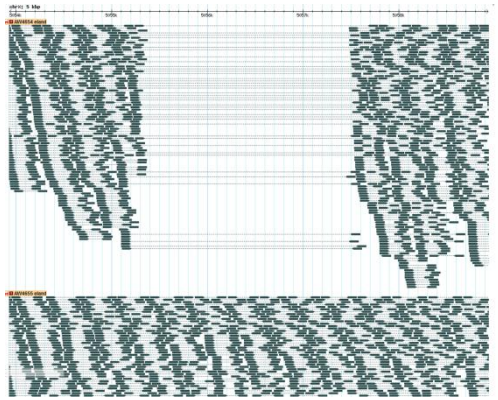
McGowan SJ, Hughes JR, Han ZP, Taylor S MIG: Multi-Image Genome viewer. *Bioinformatics* (2013) **29**: 2477-8

DNA Mutations

Single base mutation



Insertion

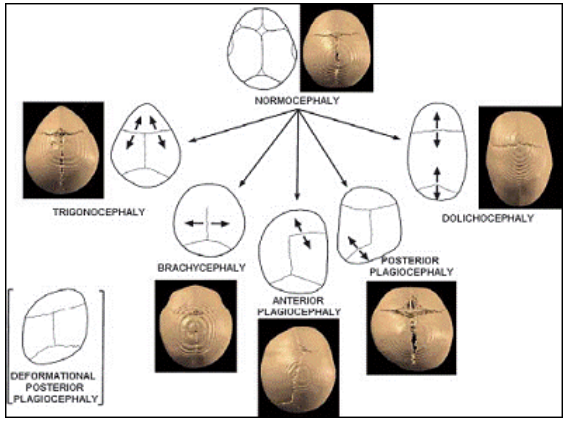


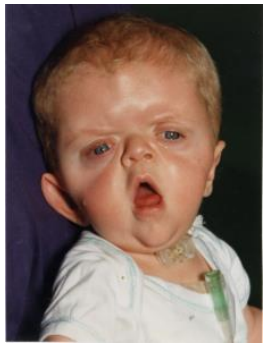
FAULTY GENE

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/)

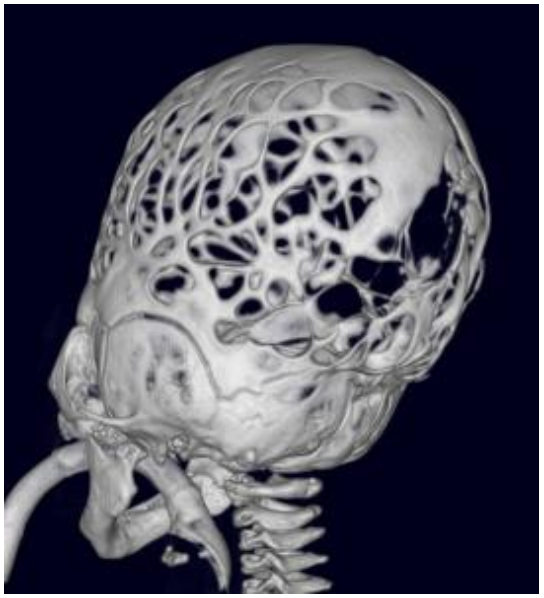
Craniosynostosis





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

Mark Henderson Science Editor
August 3 2011 12:01AM

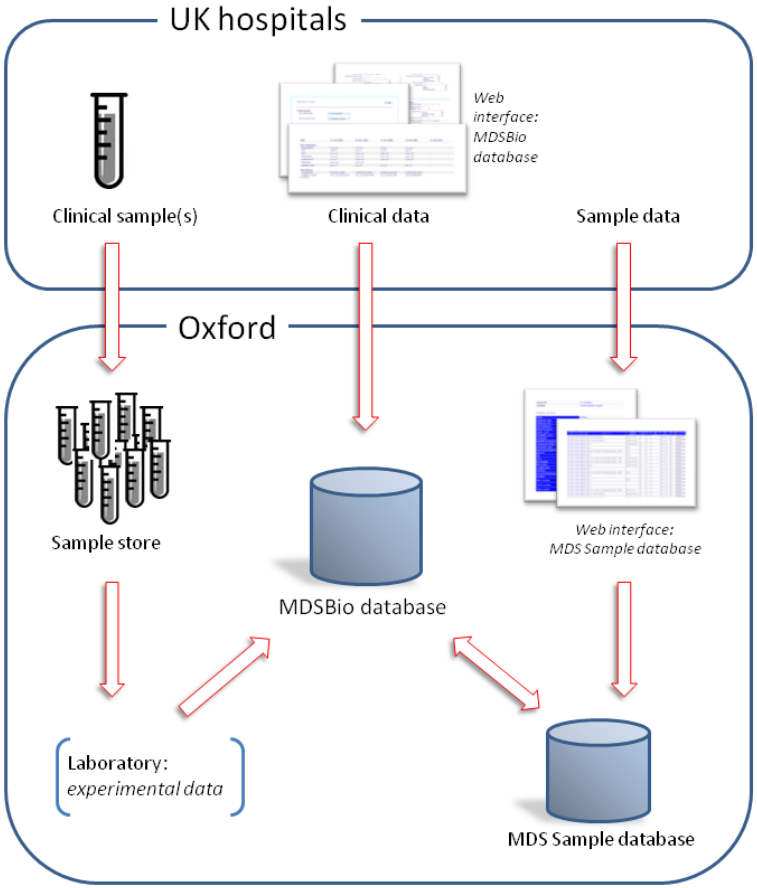
A four-year-old girl has become the first person in Britain to have her entire genetic code read to identify the cause of a disease, in a landmark development that illustrates how personal genetics is changing healthcare.

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

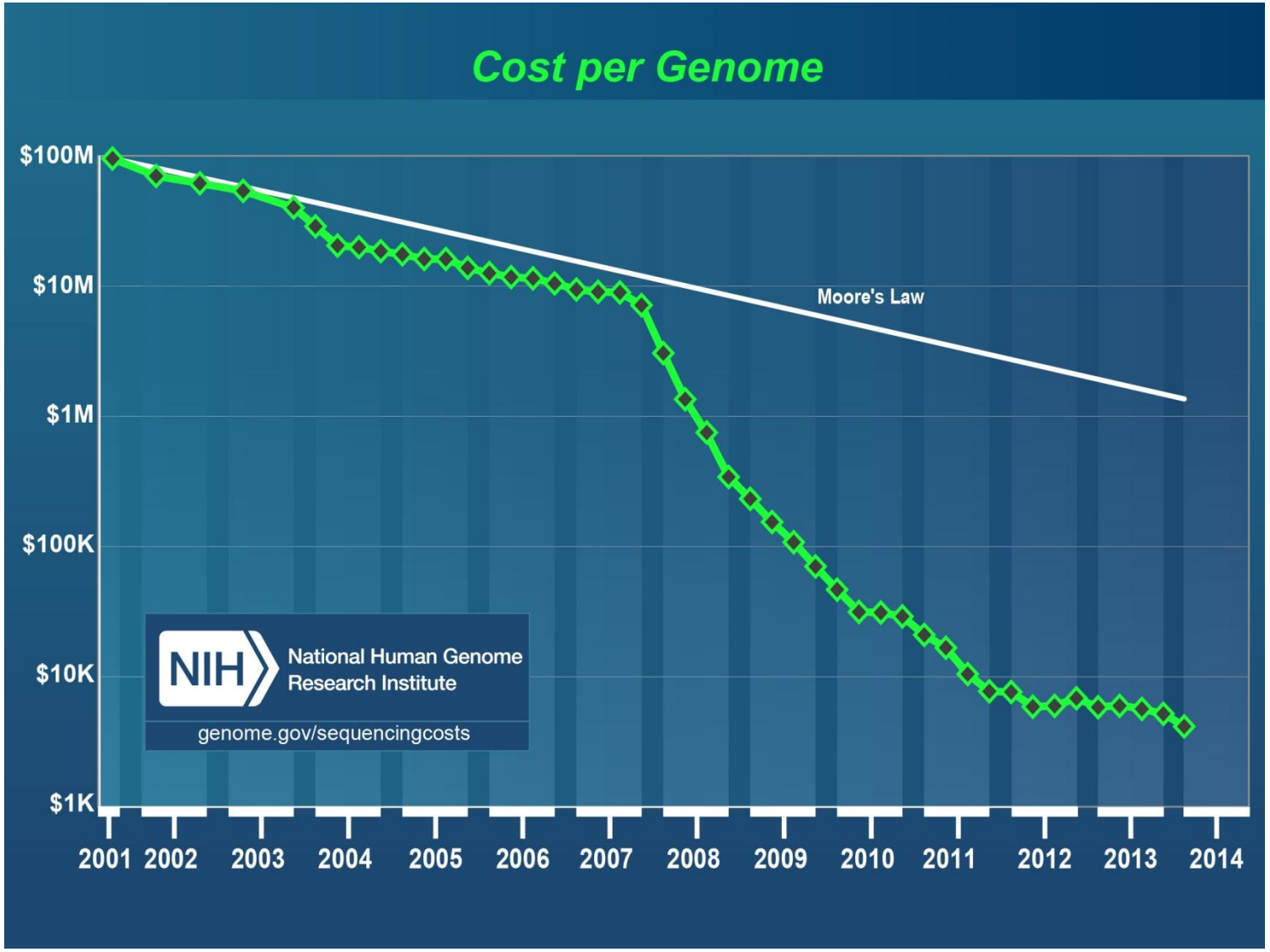


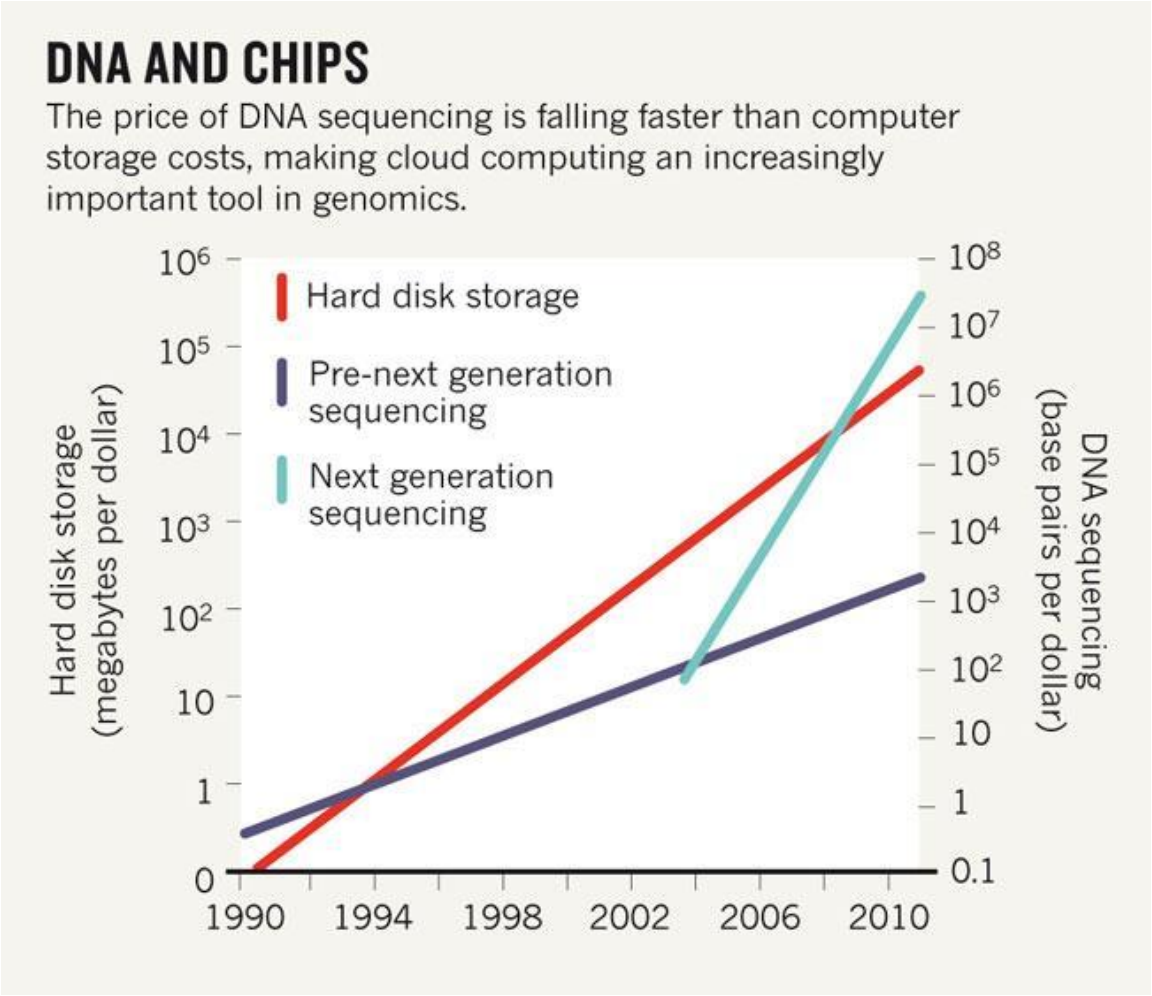
Myelodysplastic Syndrome (MDS)

- 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

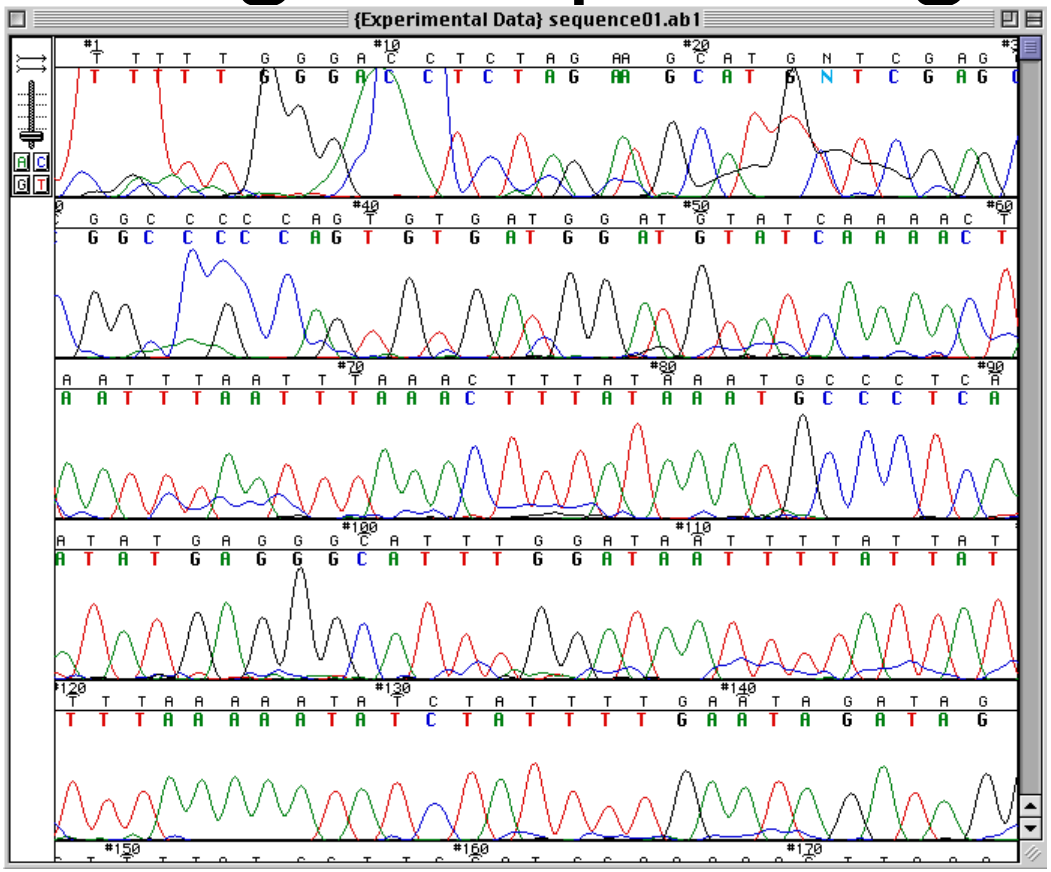
- 100,000 patients with rare inherited disease, common cancers and pathogens from the NHS in England
- Whole Genome Sequencing
- http://www.genomicsengland.co.uk/


Cost per Genome
$100M
$10M
$1M
$100K
$10K
$1K
Moore's Law
NIH
National Human Genome Research Institute
genome.gov/sequencingcosts
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014

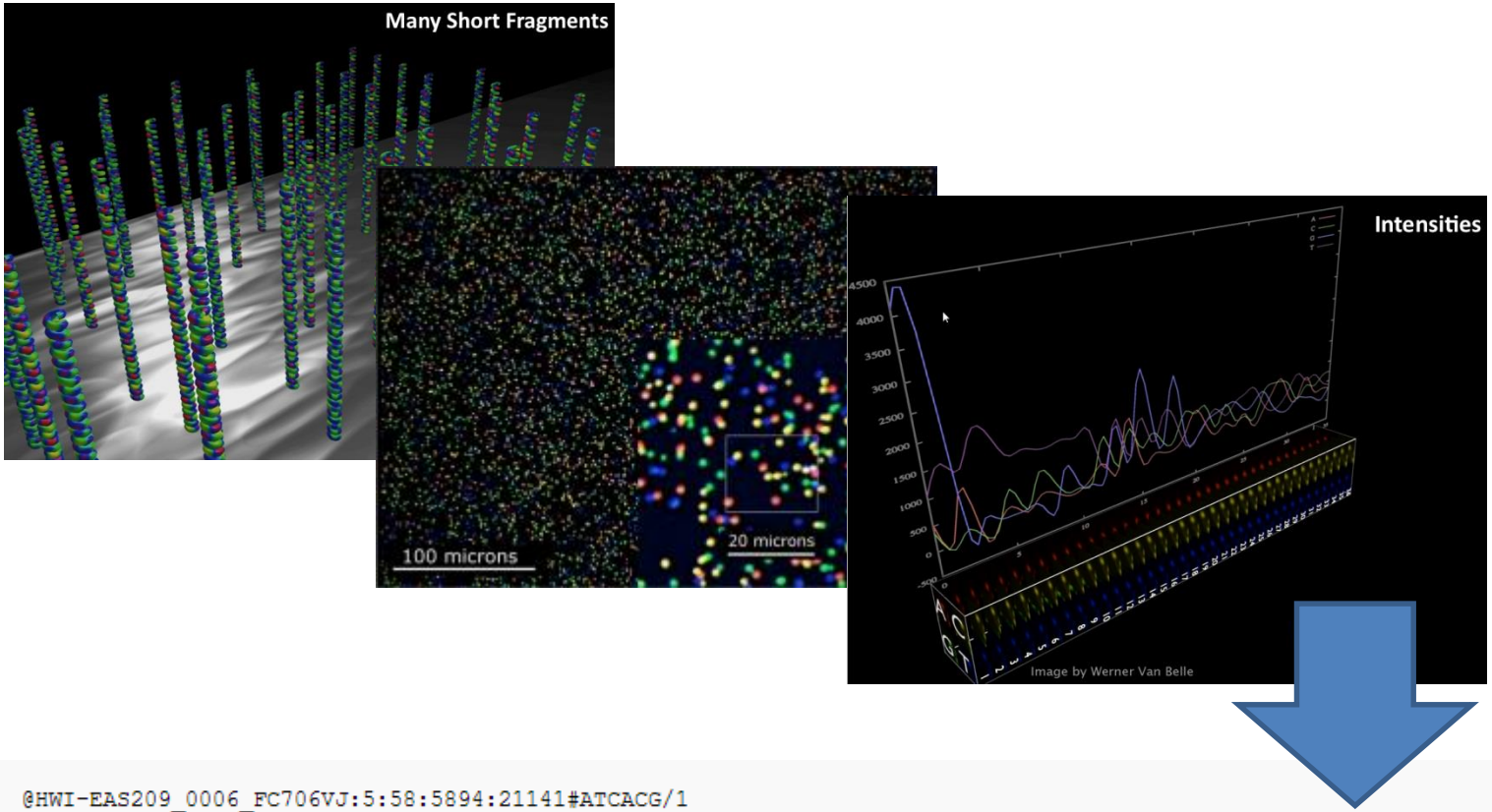


Source: L. D. Stein Genome Biol. 11, 207 (2010)

Sanger Sequencing



Next Generation Sequencing

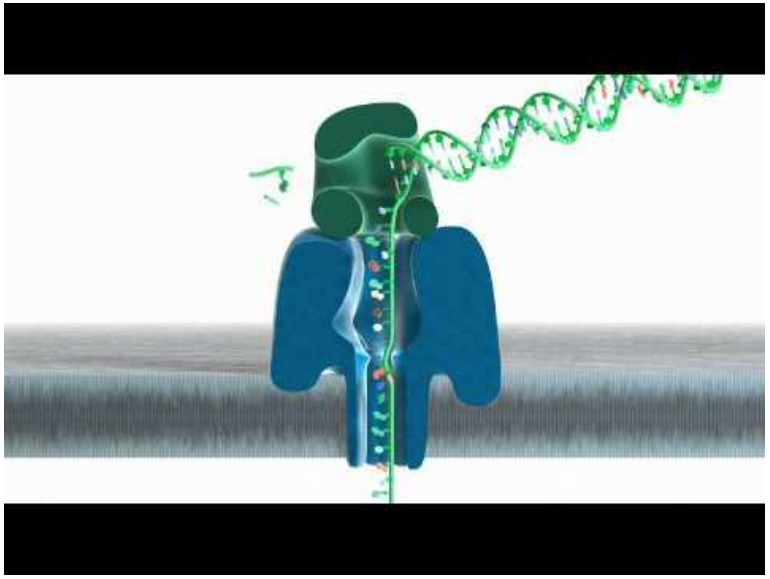



```
@HWI-EAS209_0006_FC706VJ:5:58:5894:21141#ATCACG/1
TTAATTGGTAAATAAATCTCCTAATAGCTTAGATNTTACCTTNNNNNNNNNNTAGTTTCTTGAGATTTGTTGGGGGAGACATTTTTGTGATTGCCTTGAT
+HWI-EAS209_0006_FC706VJ:5:58:5894:21141#ATCACG/1
efcfffffcfeefffcffffffddf`feed]`]_Ba_^__[YBBBBBBBBBRTT\]][]dddd`ddd^dddadd^BBBBBBBBBBBBBBBBBBBBBBB
```

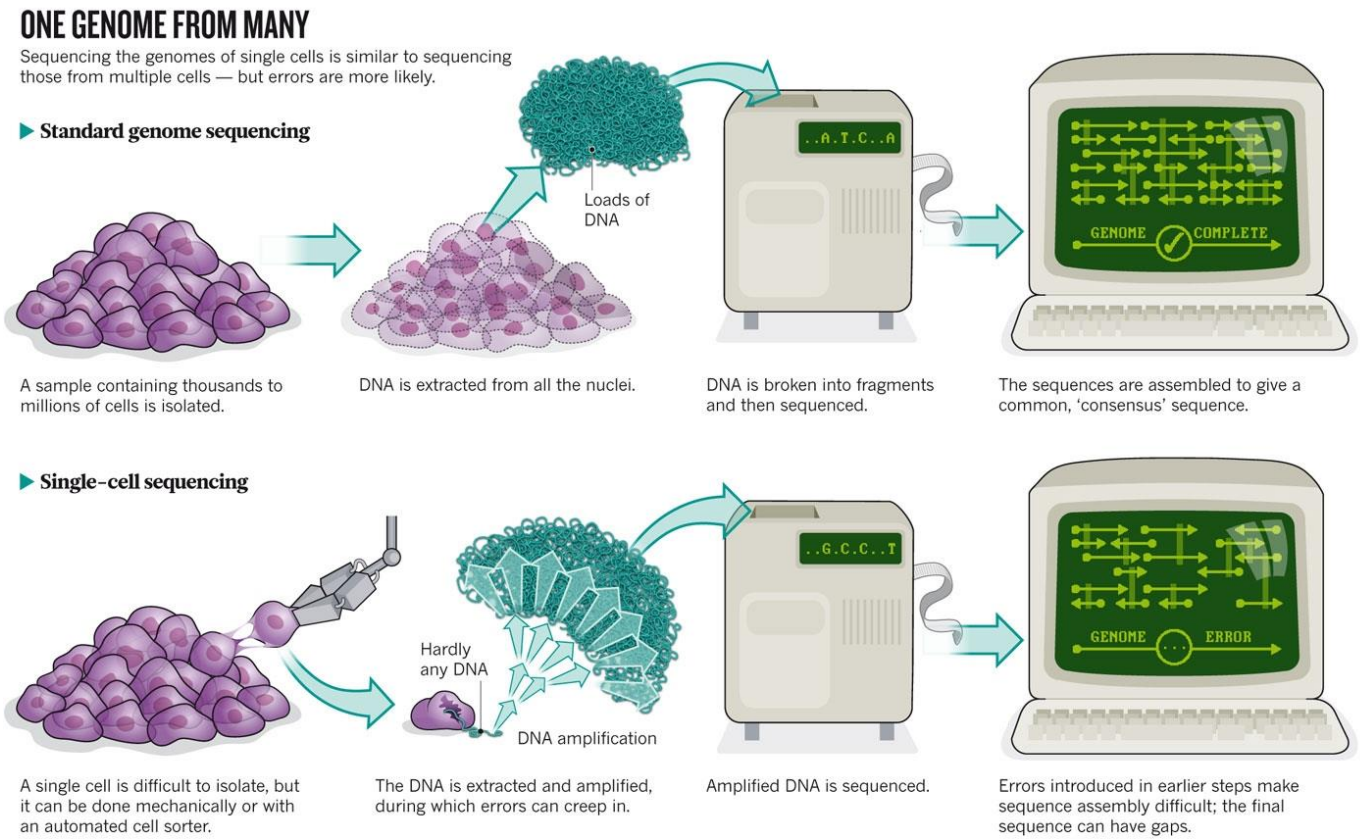
http://werner.yellowcouch.org/Papers/pippres0802/index.html

Nanopore sequencing

Oxford NANOPORE Technologies



Single Cell Sequencing



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

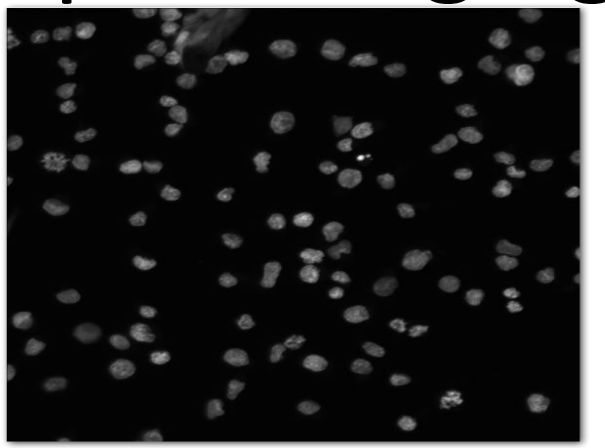
Source: http://c179631.r31.cf0.rackcdn.com/info_byte-final.png

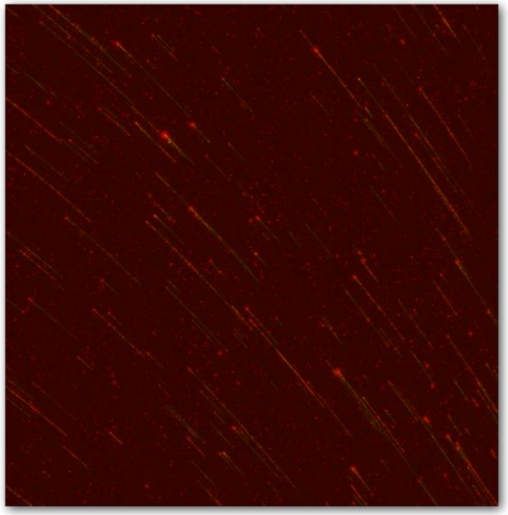
High Throughput Imaging

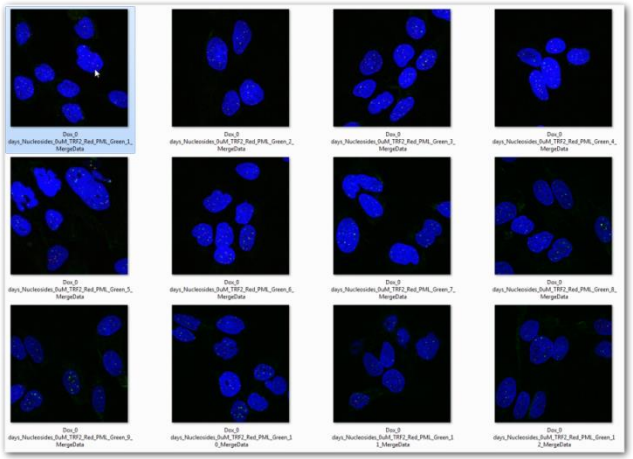
- Golden age of microscopy
  - Breaking the light wavelength barrier
  - Fantastic optical CCD cameras
  - Automated image acquistion
  - Large disk storage systems
- Masses of images to analyse...
- More BIG DATA...

High Throughput Imaging

- Cells
- Nuclei
- DNA







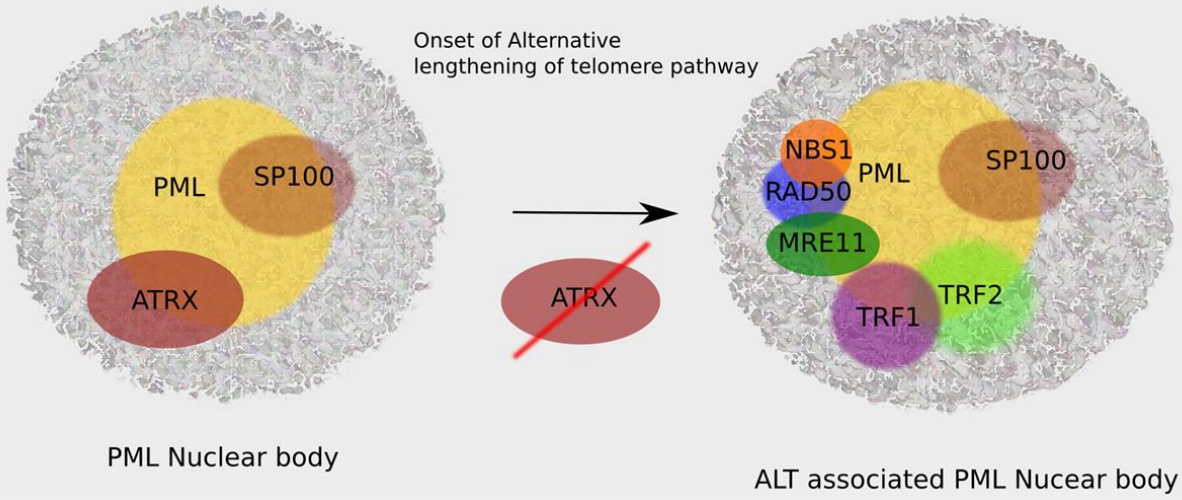
Analysis Bottleneck

- 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

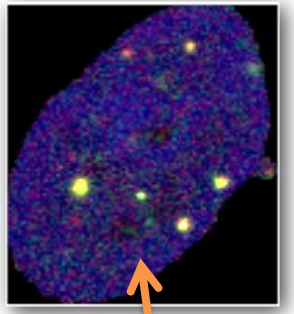
- Spatial overlap between two or more foci in the cell
  - Label proteins with fluorescent dyes
  - Measure overlapping pixels
  - Correlation analysis

Biological Question

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

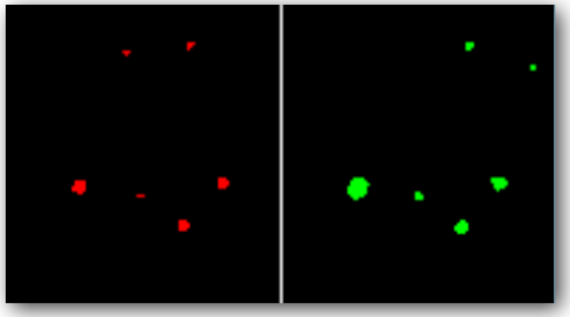
ROI Processing

Segmented Nuclei



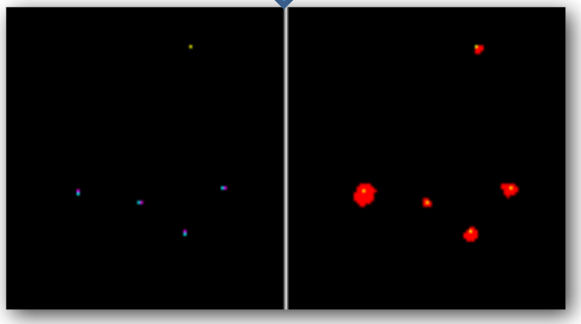
ImageJ

Split Channels / Threshold



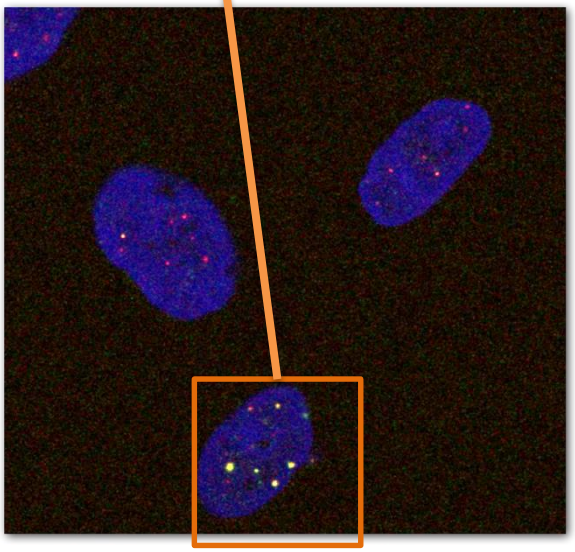
TRF2 PML

JACoP



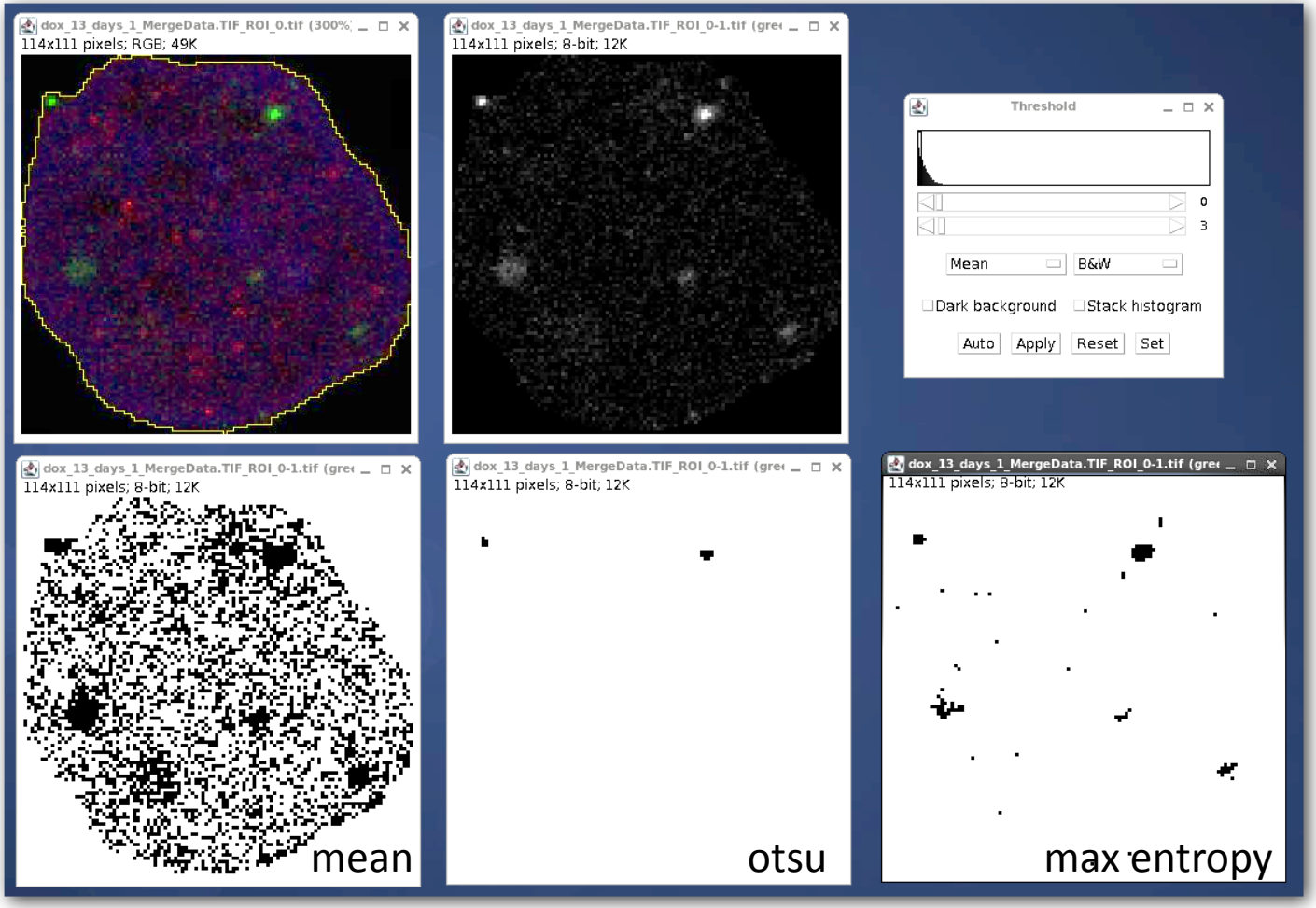
Distance Based Colocalisation

Centre of Mass Colocalisation



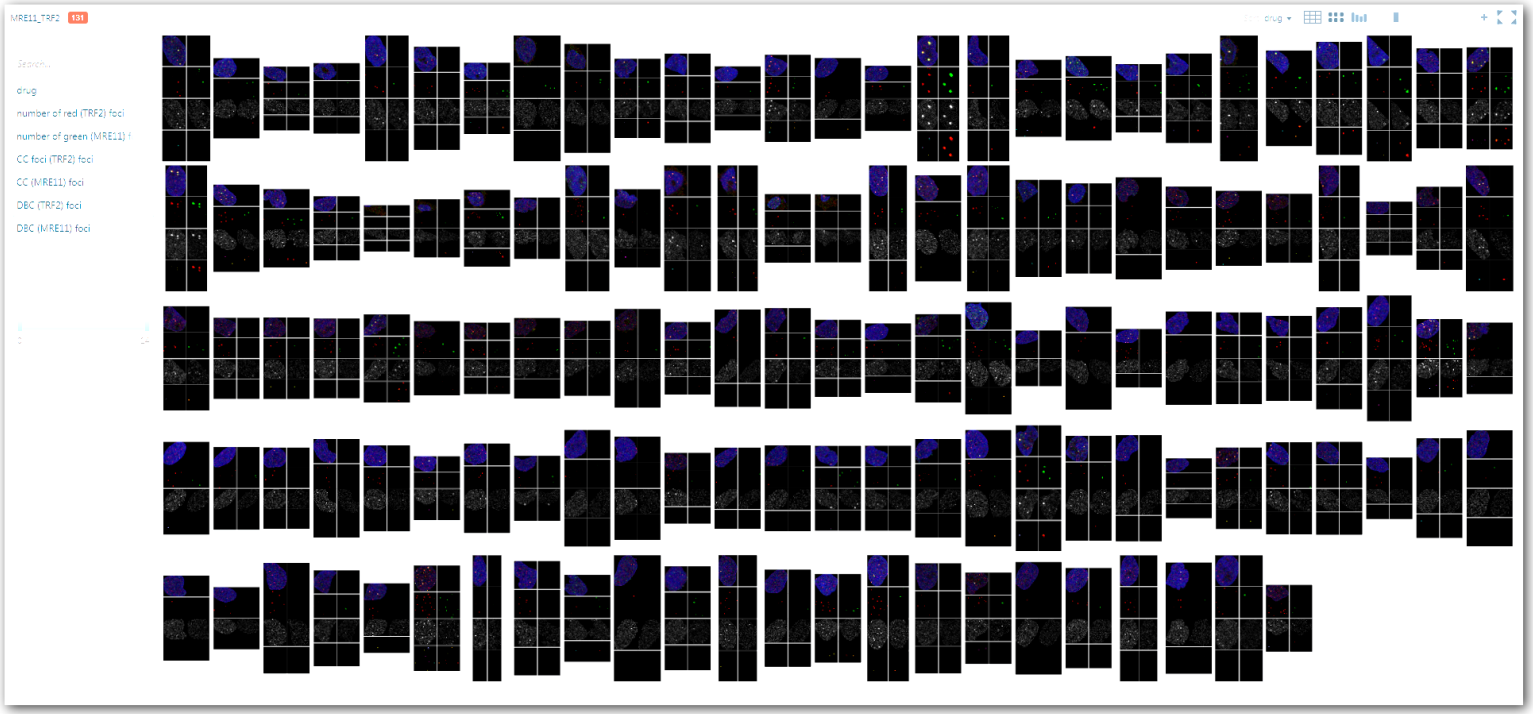
Widefield image

Thresholding



Which is the most accurate?

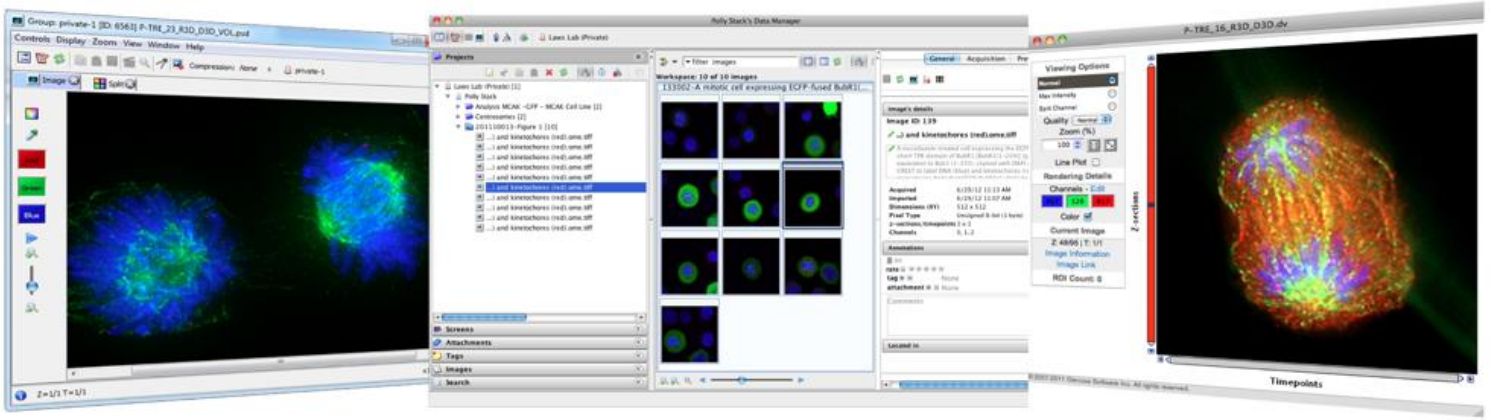
HTML5 PivotViewer



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