Realising the Value of the Genome Sequence

Richard Durbin

Wellcome Trust Sanger Institute

rd@sanger.ac.uk

The Human Genome Project

- Sequencing 3 billion bases was an unprecedented technical and logistical challenge for biology.
- The initiative started in concept in 1985, in principle in 1990, and in earnest in 1995.
- The resulting genome sequence, and those of other organisms, have changed forever the way people do biology.

Less than 1/1,000,000 part of the genome sequence

The naked DNA sequence is a beginning not an end

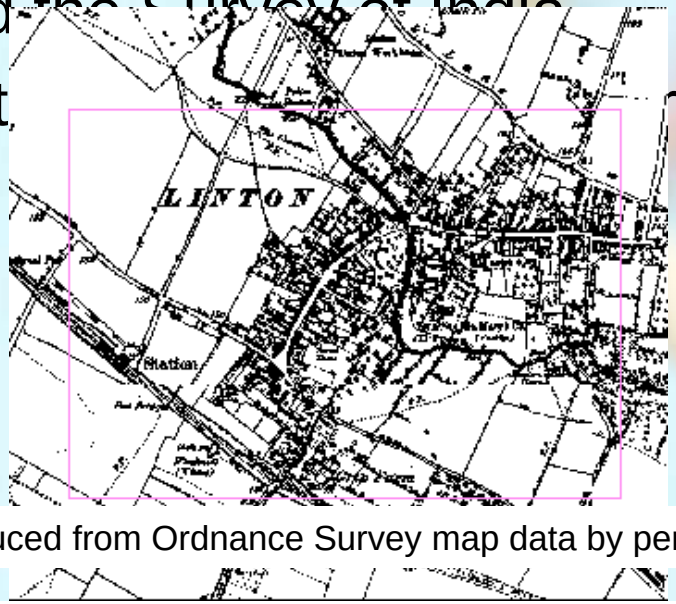
Interpreting the Human Genome Sequence

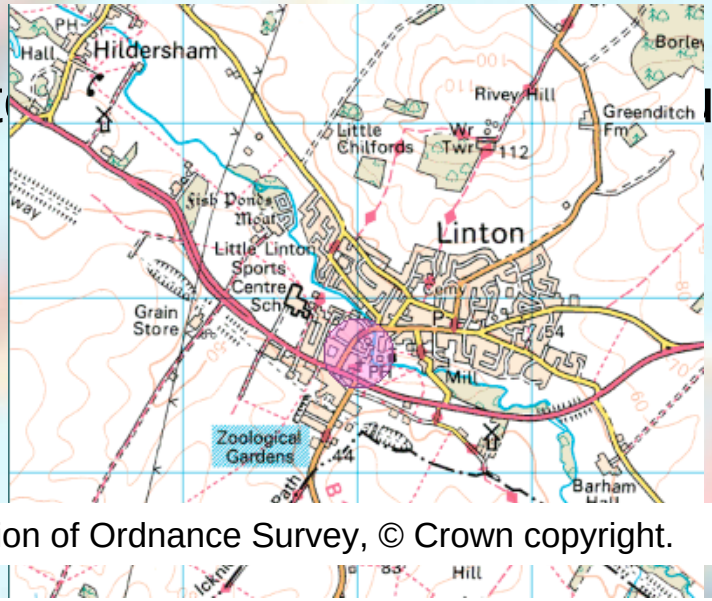
We need some sort of map

Mappa Mundi, c 1300 AD
Superimposed on the continents are drawings of the history of humankind and the marvels of the natural world. These 500 or so drawings include of around 420 cities and towns, 15 Biblical events, 33 plants, animals, birds and strange creatures, 32 images of the peoples of the world and 8 pictures from classical

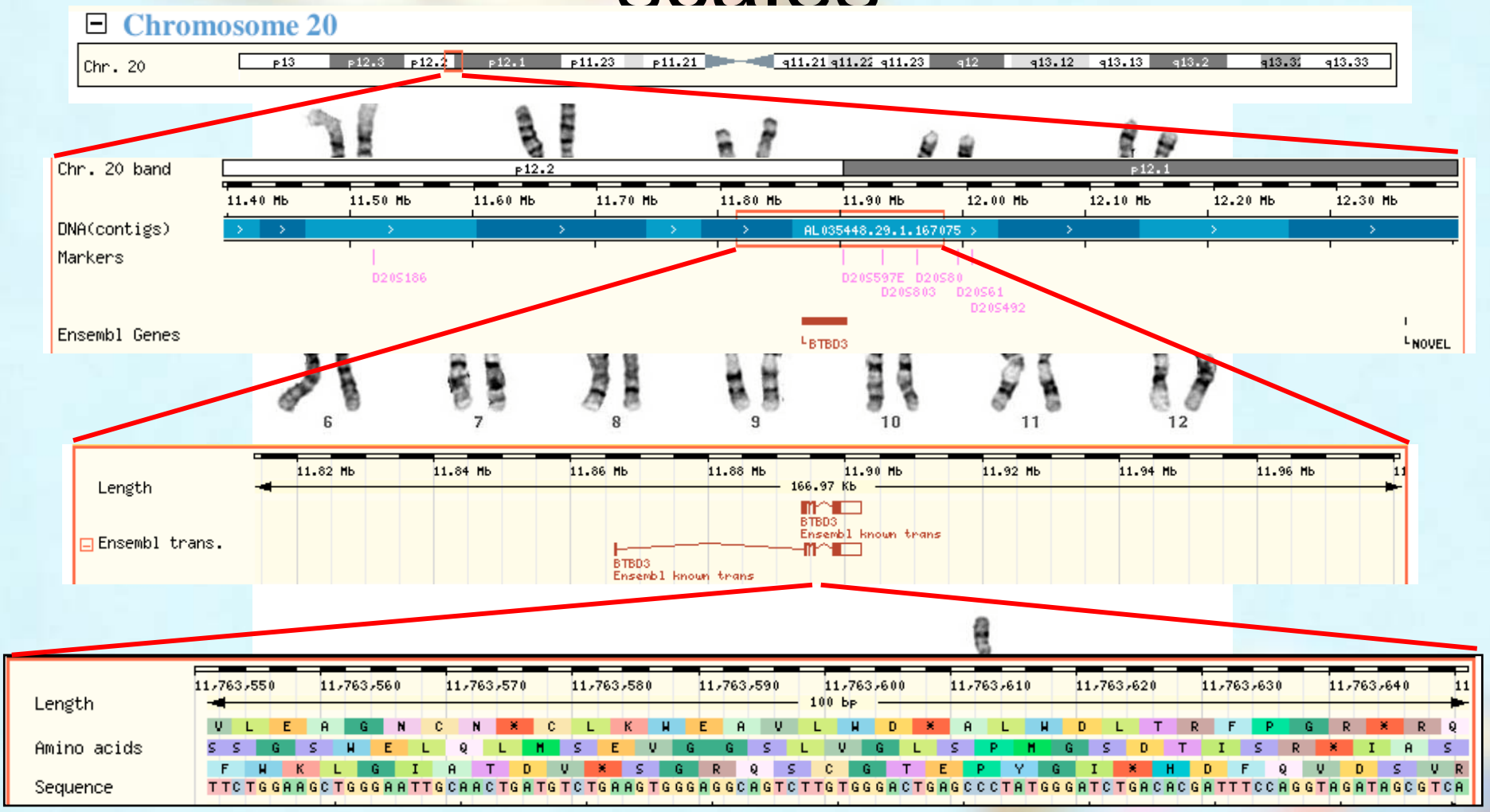
Progress in maps

- Approximate local maps in the Middle Ages and the Renaissance
- Accurate local maps and charts in the 18th century: Captain Cook
- Systematic maps in the 19th century: the Ordnance Survey and the Survey of India
- Dat rect ry





A genome map at different scales



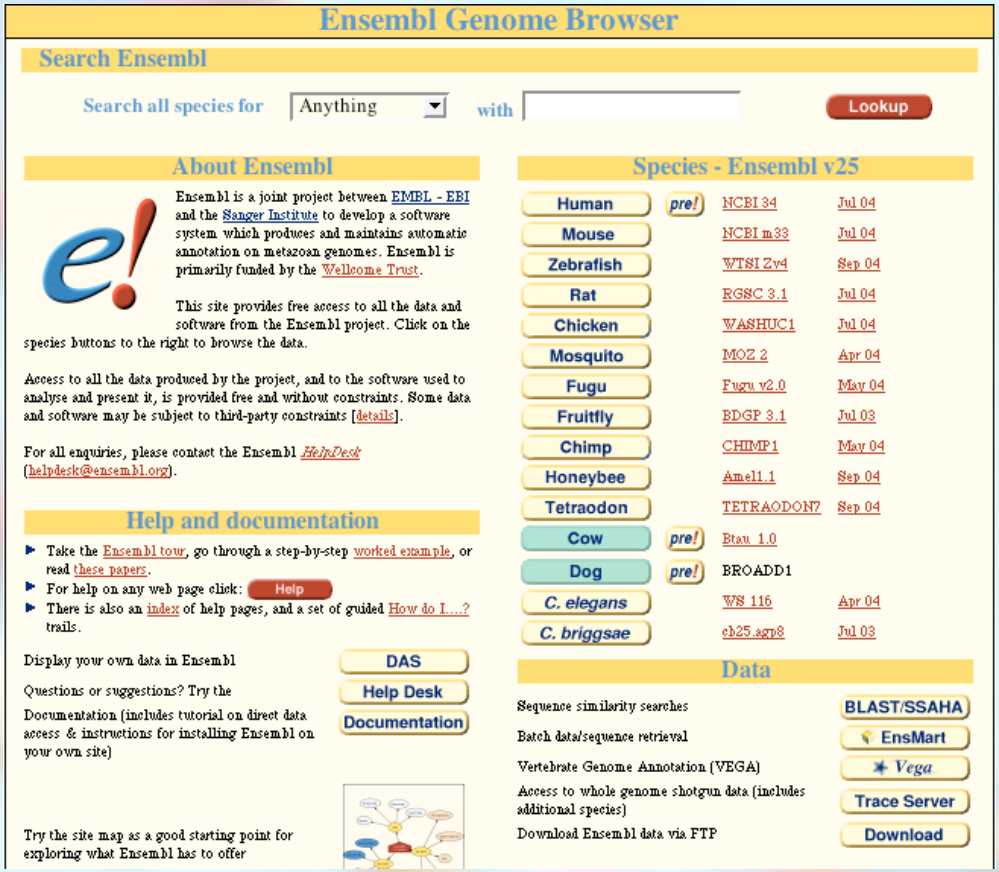
What is the genome sequence?

- It is pure information **Analyse on computers**
  - The heritable information to build an organism
  - High accuracy attainable -> a reference resource
- Finite and complete **Classify in databases**
  - About a CD-ROM's worth for human
  - All the genes are there
  - Basis for systematic experimental design/interpretation
- It is the product of evolution
  - Related genes have diverged by mutation: substitution, insertion and deletion

Transfer knowledge between related genes and species

Ensembl: Sanger/EBI Genome Resource

- Database containing genome sequence and many layers of annotation
- Comparative data relating 15 organisms' genomes
- Gene sets, which have provided the basis for human, mouse, rat, chicken publications
- Accessible in many ways: web, download

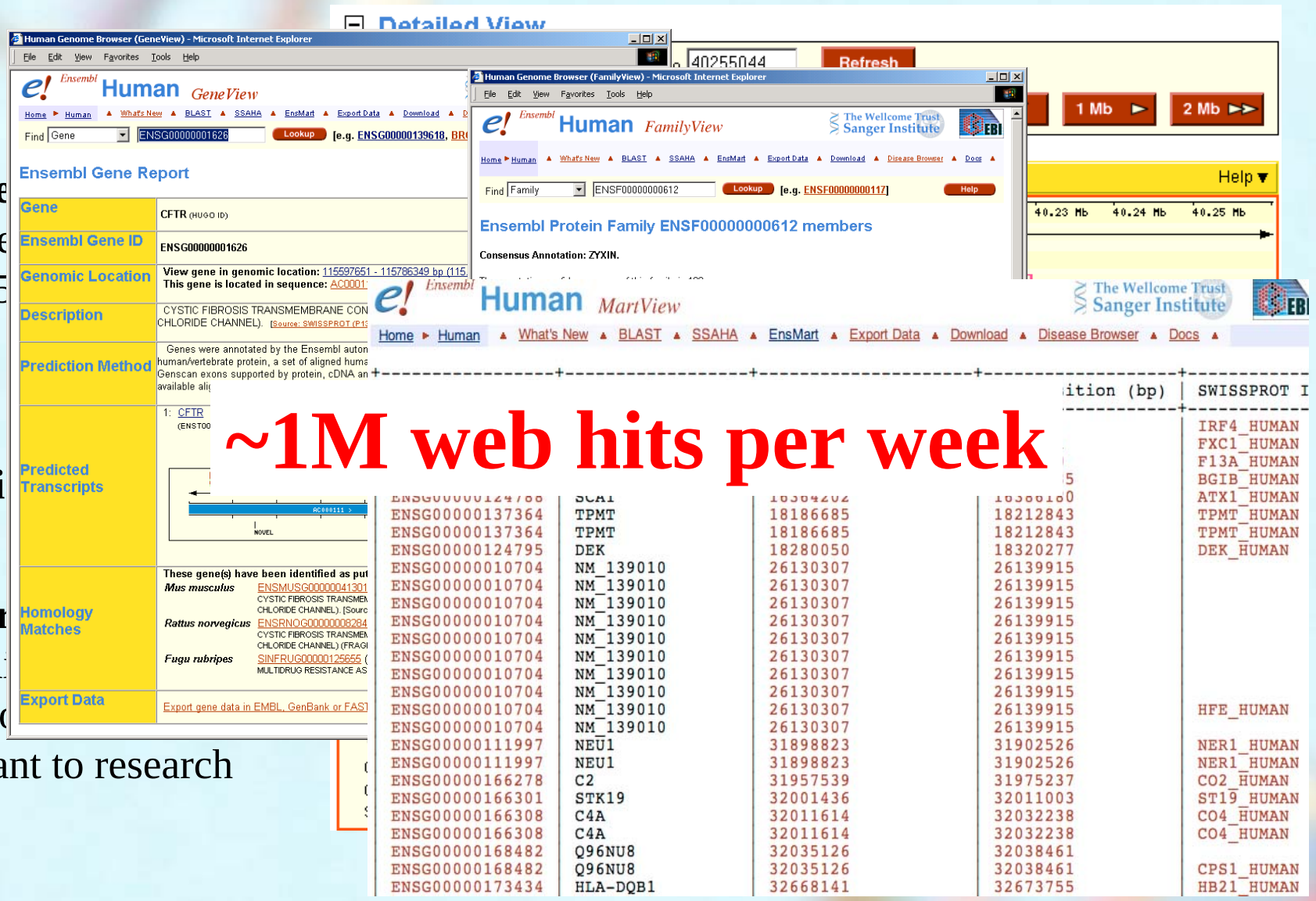


Some of the data in Ensembl

Gene
(Ense
23,75

Evi

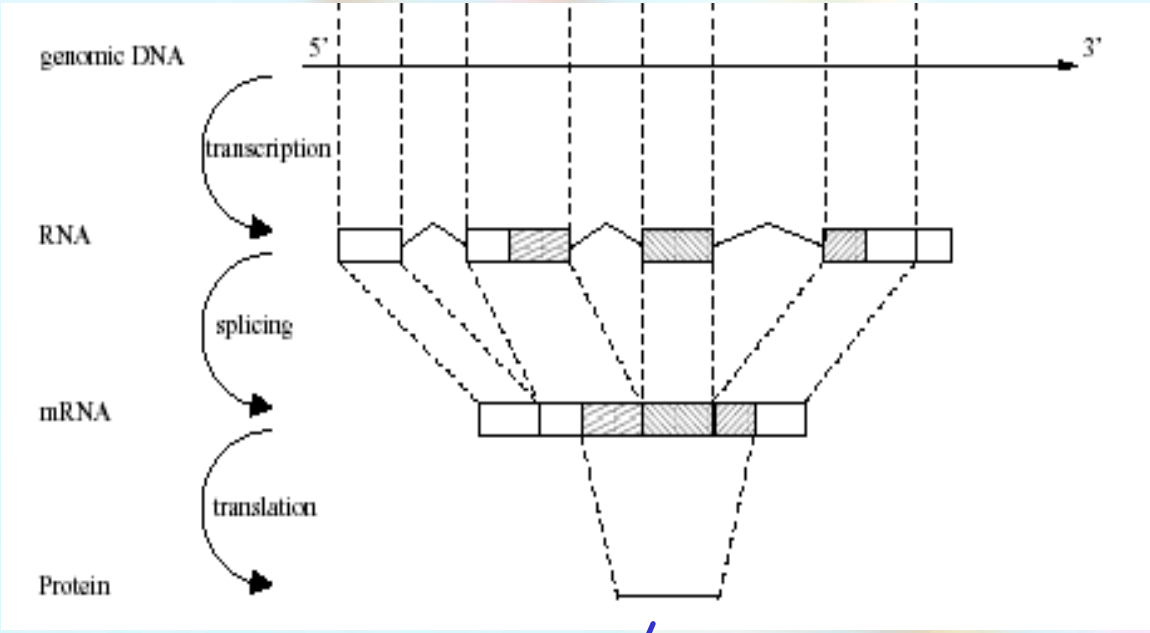
Other
~80 d
of info
relevant to research



Where are these genes?

```
AH6
     32000
     32500
     33000
     33500
31861 ggaaaaattaagttttagaagtgtttaaggtacttttctataattatttattataaaag
31921 atatagtcttcccttgtcatgacatgtggttaattctatgaaagtttgatagaattatga
31981 tattcacataaaacaaggttgtgatgtctggtgtttcacagtccgttgacttttgatgca
32041 attcttggtagacatcctccgactatgttttagatgtcattttcaagttttgcagtttct
32101 cgaaatattagaagccatgtctgcaccgaactgcgcacgaaaatatgatattgctcgtct
32161 ttccagcttgaattttcaaatttcccaatatgtttatcttagcttgataagcttaacttt
32221 tatattttcttattttgctgtgaaaattgttcatcaaaaatcgattttccaactttccac
32281 taaaatcttattatttcacaatttggtttctgcgaatcttcatcaacttttatacttatt
32341 ttccgcactccgaaggctcaacctggcatatttctatattgacgaaccatgtgttcctct
32401 aatatcggaagctgactgtctaccttacctcaaggttttagtaactggaataagtggaat
32461 gatctatggccaaactggtctacttttggaacgaggttgtgcaactttcatcaaagatta
32521 tgataagaagacaagcatgtttgttgggatagcaatatcaattgccattttgtttttgag
32581 cttgatcactggaaaaattataatttgggatgacccacttcaaggatatcttctttcgtg
32641 tgtttcctatccaagtcaaagtgttgaaagatcaagactatttgcaagtatttatacatt
32701 tatatcttcattcaatttggttttctcagttttgttaagaagatataacaaaaaactgga
32761 atattcgtaagctttctaagttgtttgaaaagtttactgttcgtcaatttctggaatttt
32821 agttaagaaagtcagatggcaaatcatgatgccttcataaaatgagtaaataaacctgat
32881 tagtttactattttgttcaaacttcaattttggaacgtattgcgcttaaaggtactgaaa
32941 actagtatgcacgaaaaacttcttactgtctactagatatctttaattgcctgaaacgcg
33001 gcaatatttagtgcaattcaacttccagacgtttgactctttgtaatttacttttgcgta
33061 atatctgatctctgaaatttctgaatagtatttttctgattagcttgtttttcttctcat
33121 tgtttccactacatttgcttccaaacttggaaaacaaatttttgaataaatctagaatat
33181 tctaatctggtttttgatgttttaaagttccattaatgtttttgagcgtaagaaatgtt
33241 tcaattttccagaacacccttttgttgttggcccgagatttcggaaaagagaagtgattga
33301 ctccacaagtacaatatgctttttgacatttgttcaattcatattcatgttcatttattc
33361 attcggaatattcacactgaaaactattcgaagcatgcttacttacagacagtactattt
33421 cattgttgtctggttctatgtaagcacttgaaatttattttaaaaagctgaaaattttat
33481 ttccagacaattccattcattgctgcgctgtttccaattctactagtttacaggattcgt
33541 tcttcacatgttagccgagtaacgattattaaaacatttacaaaaaccaagcaaacacaa
33601 gaagaacacattaagcaattgaaaaacgtttggaattgaacataatgattcctattttat
33661 gaaatctgaatttttgtaaatatgtgtatattttttggaataaataattgtcattagga
33721 aaaaaatcgagtgatcttcttttccgaattttcattttaatttcgagatagtaagaaaag
33781 ttgcaagtcatttgaattcaacgatttccttaatatttctgaatttattcttcaaatgt
33841 at
```

What does a gene look like? DNA makes RNA makes Protein



→ Function

Regions that code for protein

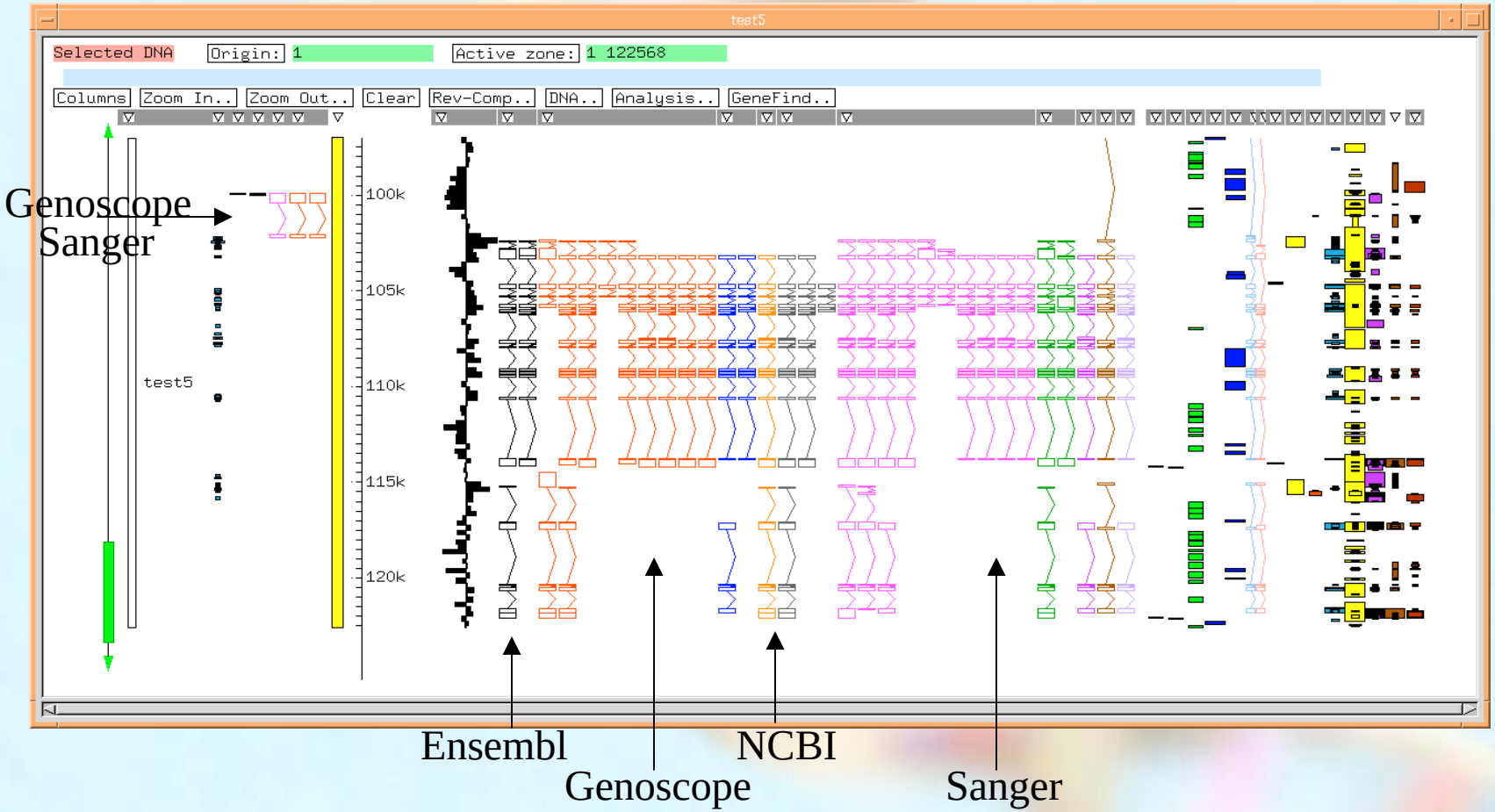
How do we identify the coding regions?

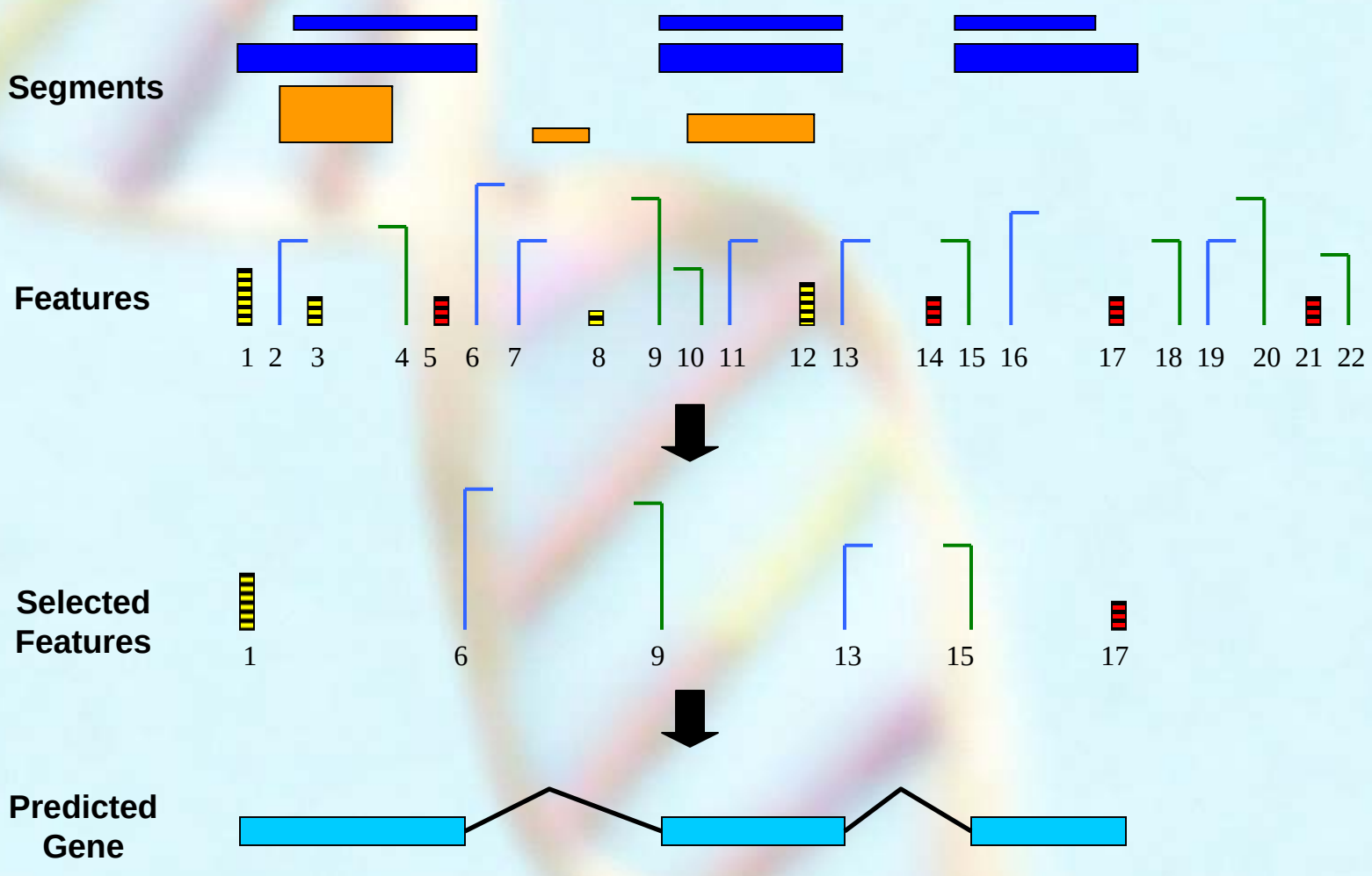
Experiments and computers

Methods for gene identification

- Experimental: sequence the RNA expressed from genes
  - Matching it to the genome is usually (but not always!) simple
  - Incomplete: not all mRNAs are cloned, and systematic projects largely ignore splice forms
  - Many transcripts are non-coding: RNA genes, reverse strand
- Computational approaches
  - Ab initio gene prediction: Hidden Markov Models
  - Using additional evidence from proteins or related sequence
- Integrating multiple types of evidence

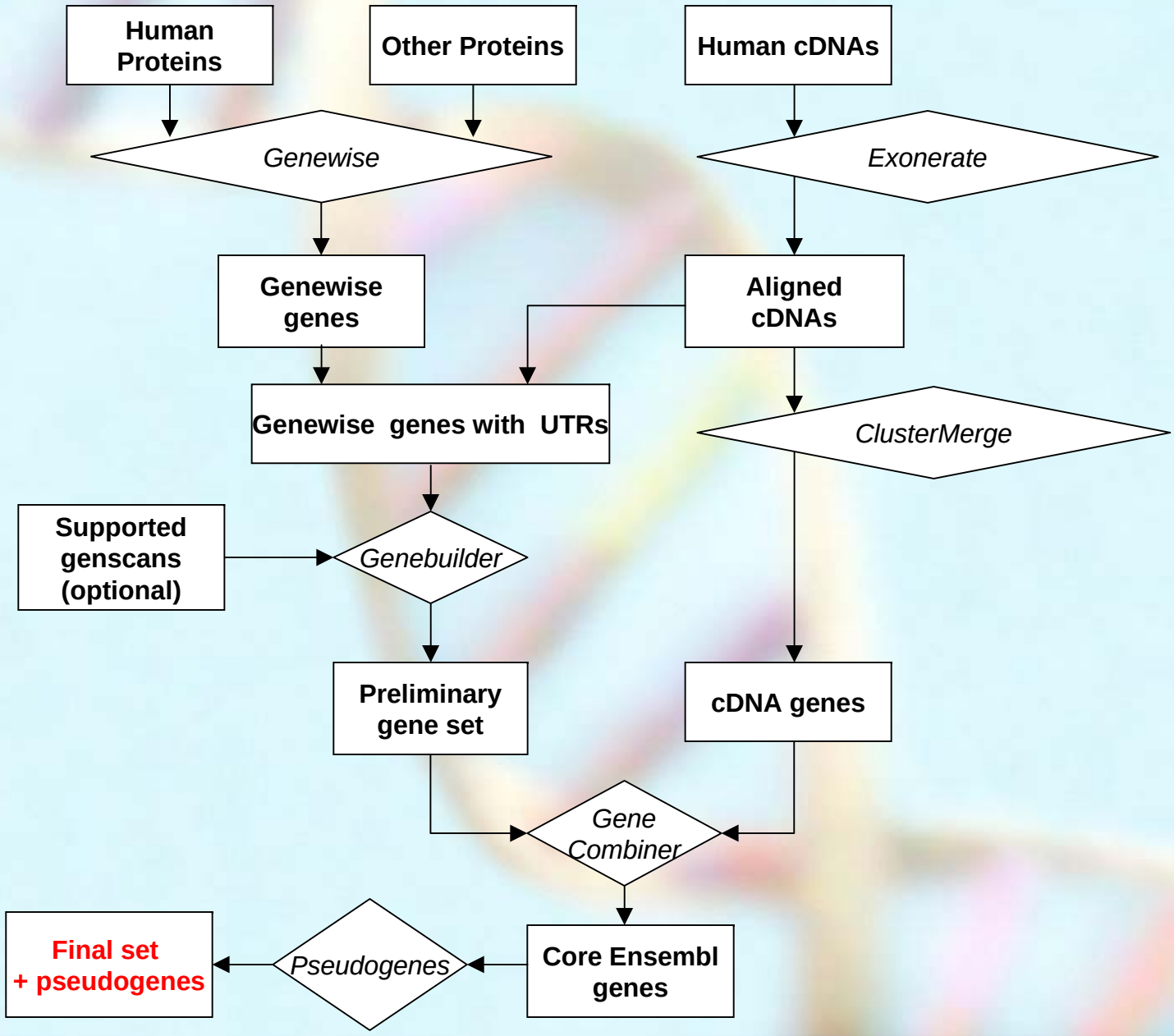
Variability in gene annotation



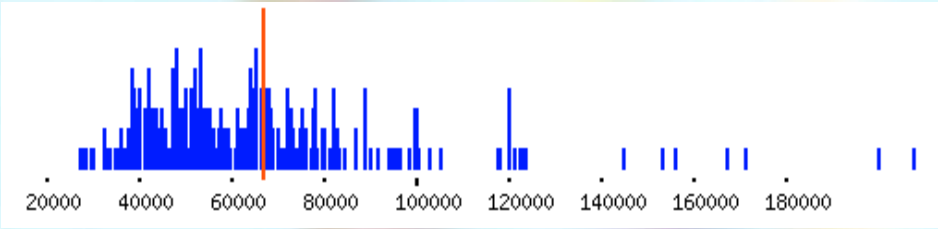


Programs select the gene structure that is most likely under a probabilistic model, given the evidence. The methods used are closely related to those used for computational speech recognition (based on hidden

Ensembl gene build overview



How many genes are there?

- Pre-genome estimates of gene number varied widely
  - During the 1980s and 1990s people thought 50-100,000
  - A couple of indirect estimates 1999 were lower: ~35,000
  - Expert predictions were spread, e.g. Cold Spring Harbor betting 

enesweep.html) *(1999)*

- Genome based
  - Feb 2001 genome sequence paper: 30-40k gene estimate
    - 14,882 known genes, 16,896 predicted (31,778 total)

How does having the sequence and the genes change things?

- Finding and characterising the human copy of a gene studied in mouse
  - Used to be a three year project
  - Now 5 seconds with a web click
- Three more advanced examples:
  - Changing the way we look for major genetic abnormalities
  - Looking for genes involved in cancer
  - Using model organisms to study gene function systematically

The human genome: how it looks down a microscope

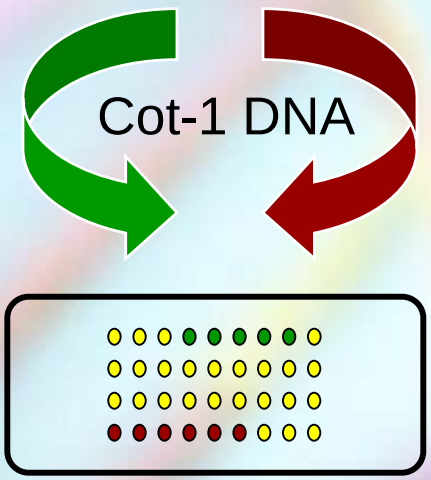
Chromosome defects in the clinic

- Some severe genetic defects are visible in the microscope
  - Deletions, duplications and rearrangements
- For children with mental retardation or significant physical abnormalities, about 5% show such defects
  - Diagnosis is technically demanding,
  - but very valuable to the families. They can understand the cause, and assess recurrence risk.

Looking for defects on DNA arrays

Patient

Genomic DNA



Reference

Genomic DNA

Hybridise to spots of DNA representing regions of genome sequence

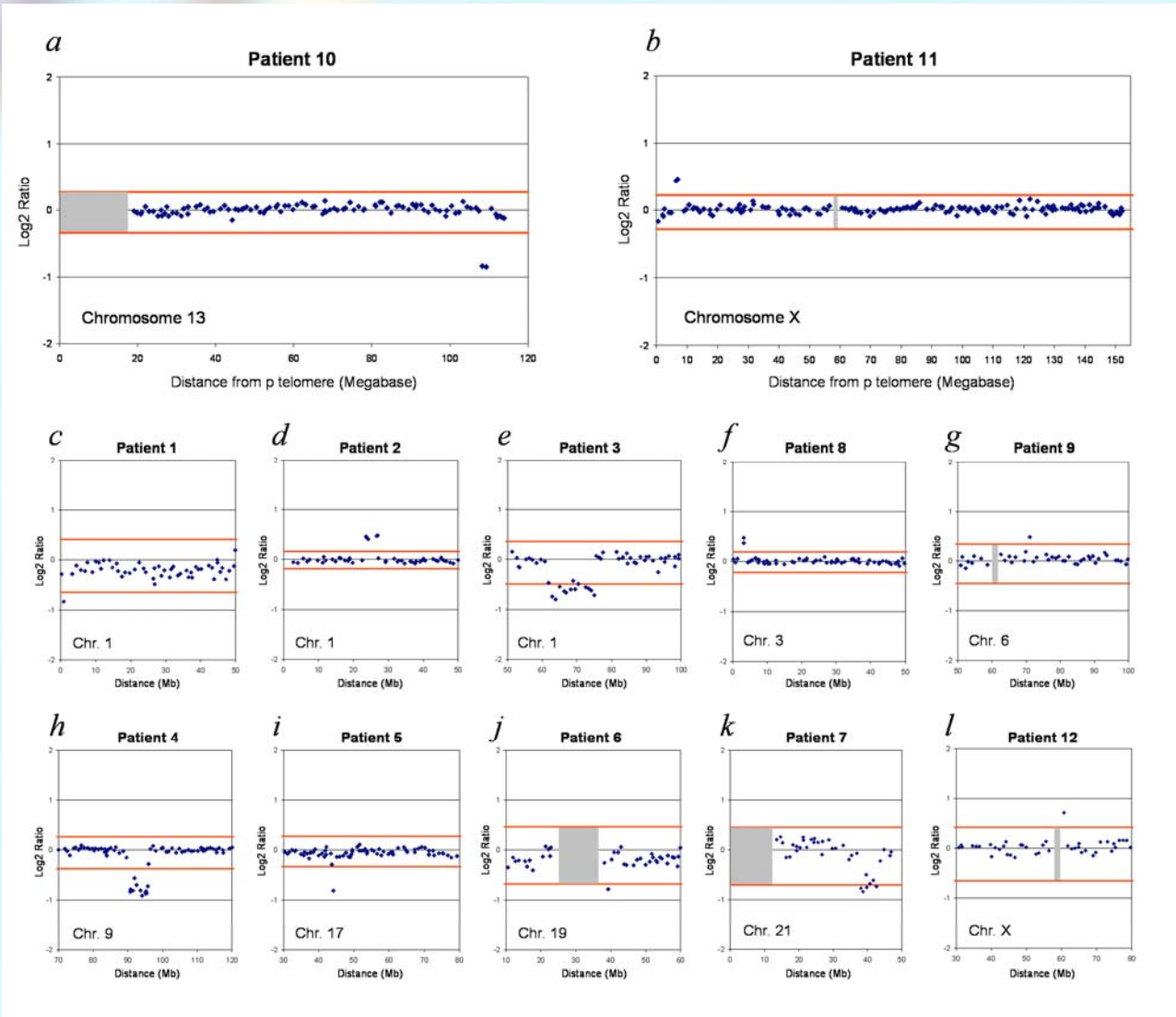
Genomic microarray (2nd generation)

- 3523 clones (one every megabase)
- plus telomeric clone set (S. Knight, Oxford)
- plus 167 clones containing known caner genes

- Five times the resolution of microscopy
- Instrument-based data collection
- Next generation slide will have 37,000 clones, giving another factor of 10 increase in resolution

Fiegler et al. (2003) Genes Chromosomes and Cancer 36:361-374

12 copy number changes in 50 patients (24%



Shaw-Smith et al, *J. Med. Genet.*

Addenbrooke's NHS
NHS Trust

Inserm

http://decipher.sanger.ac.uk

The Wellcome Trust Sanger Institute

Welcome to DECIPHER

DatabasE of Chromosomal Imbalance and Phenotype in Humans using Ensembl Resources

The DECIPHER database of submicroscopic chromosomal imbalance collects clinical information about chromosomal microdeletions/duplications and inversions and displays this information on the human genome map with the aims of:

- Increasing medical and scientific knowledge about chromosomal microdeletions/duplications
- Improving medical care and genetic advice for individuals/families with submicroscopic chromosomal imbalance
- Facilitating research into the study of genes which affect human development and health

Background

Chromosome analysis remains the single most useful tool in the diagnosis of children with developmental delay/learning disability and/or multiple congenital anomalies. The limit of resolution of a high quality Giemsa-banded karyotype is ~5Mb, and many such children have a normal result on routine karyotyping.

Challenges

Array-CGH offers the opportunity to detect submicroscopic chromosomal imbalance across the entire genome. With ~3,000 clones on a 1Mb array, and more than 30,000 clones on a whole genome tiling array, a large database is needed to capture

3rd Jun 2004

DECIPHER Database Released

The DECIPHER database of submicroscopic chromosomal imbalance collects clinical information about ...

more

10th Mar 2004

Majordomo mail list

There are two majordomo mailing lists dedicated to the DECIPHER Project,

Ensembl Human Genome Browser (CytoView) - Microsoft Internet Explorer

e Edit View Favorites Tools Help Links »

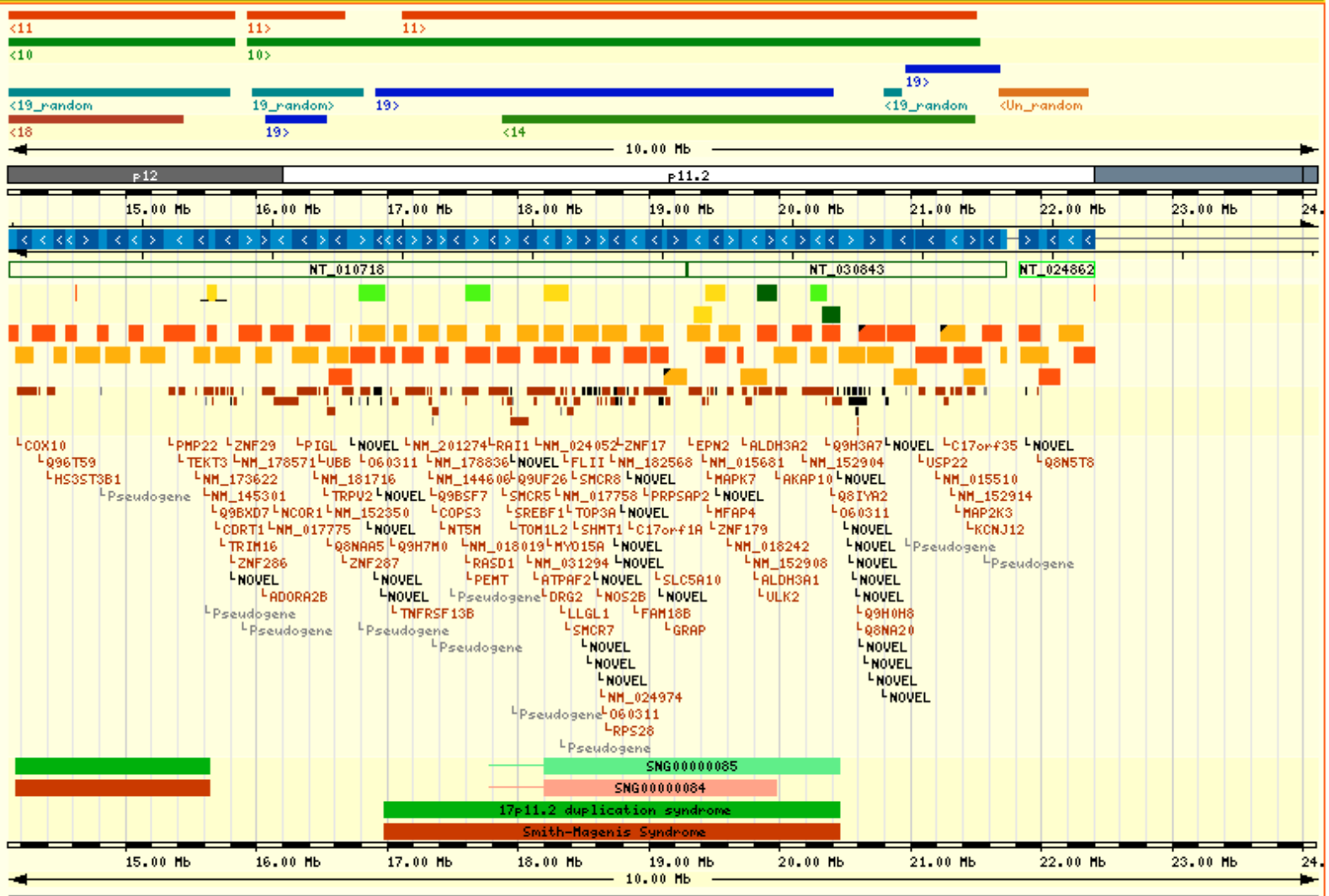
Detailed view

Jump to region: 17 bp 14104979.5 to 24104978.5 Refresh Band: Refresh

2 Mb | 1 Mb | Window | Zoom | Window | 1 Mb | 2 Mb

Features ▼ Compara ▼ DAS Sources ▼ Repeats ▼ Decorations ▼ Export ▼ Jump to ▼ Image size ▼ Help ▼

Mouse synteny
Rat synteny
Chimp synteny
Chicken synteny
Length
Chr. 17 band
DNA(contigs)
NT Contigs
1mb cloneset
Tilepath
Ensembl Genes
DECIPHER
Length
Gene legend

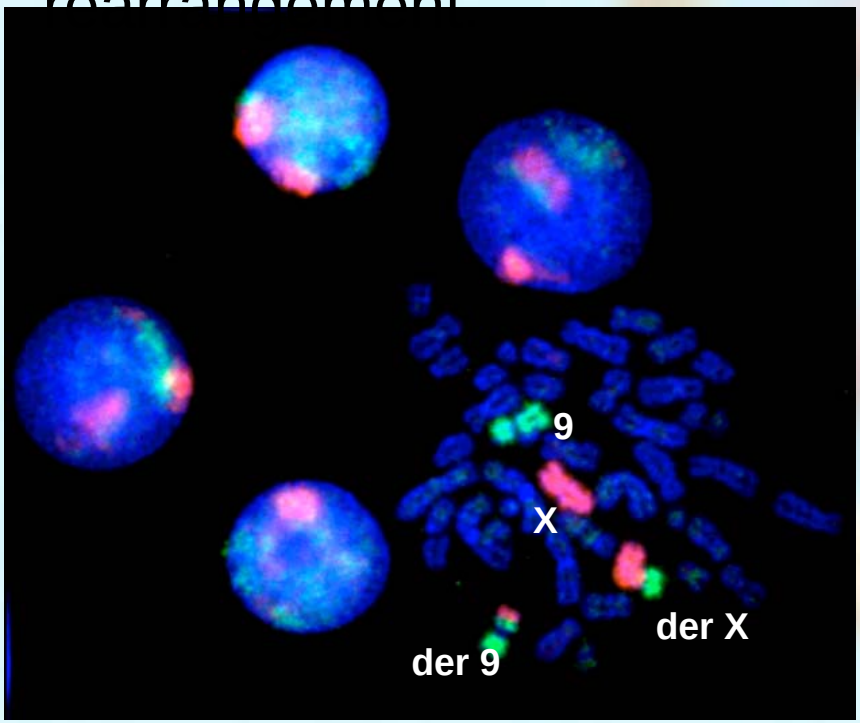


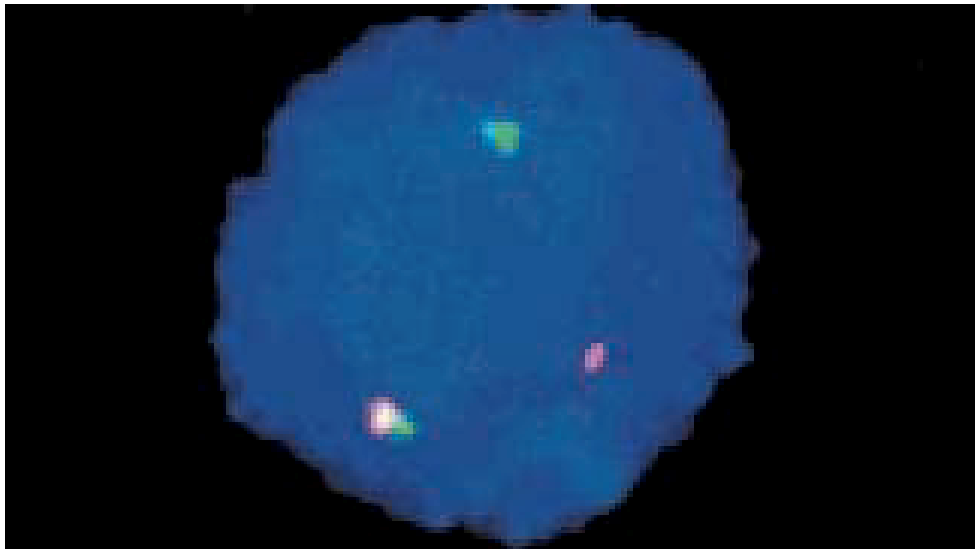
ENSEMBL PSEUDOGENES ENSEMBL PREDICTED GENES (KNOWN) ENSEMBL PREDICTED GENES (NOVEL)

There are currently 40 tracks switched off, use the menus above the image to turn these on.

Genome changes in cancer

Cancers are caused by some cells acquiring changes, or mutations, in their copy of the genome. In some cancers a gene is turned on incorrectly by a chromosome rearrangement





LSI BCR/ABL Dual Color, Single Fusion Translocation Probe hybridized to a nucleus containing the t(9;22). One orange, one green and one fusion (IOIGIF) signal pattern is observed.

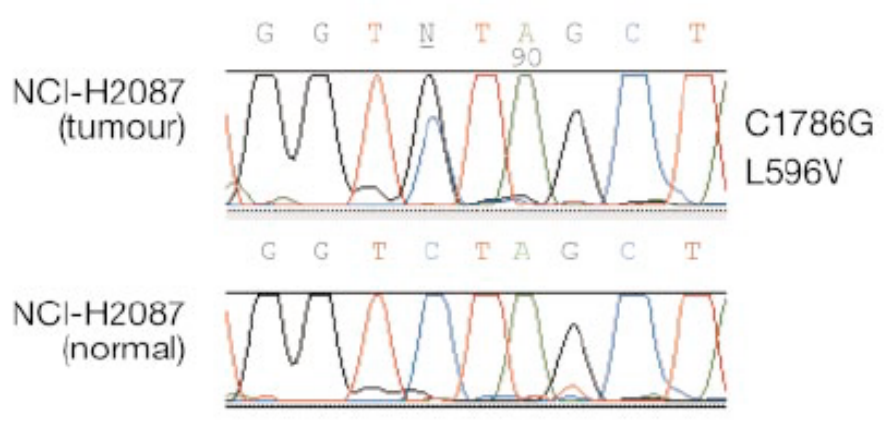
ABL can be inactivated by a specific drug, GLIVEC, curing the

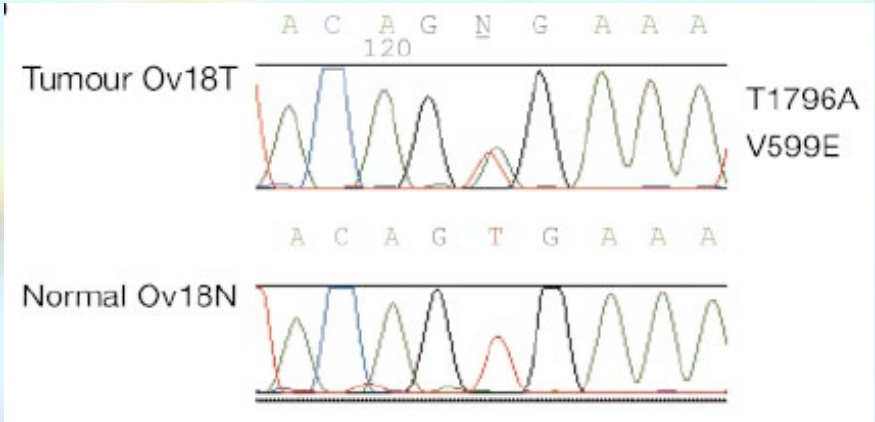
Cancer Genome Project

- Many mutations causing cancer are much smaller changes, often of single bases.
- Now we have the genome sequence, and a list of genes and their locations, we can look directly.
- Mike Stratton and colleagues at the Sanger Institute are searching for mutations in cancers, by resequencing gene-coding DNA from tumours matched to non-tumour controls.
- High throughput project: 25 000 genes x 50

BRAF is a cancer gene

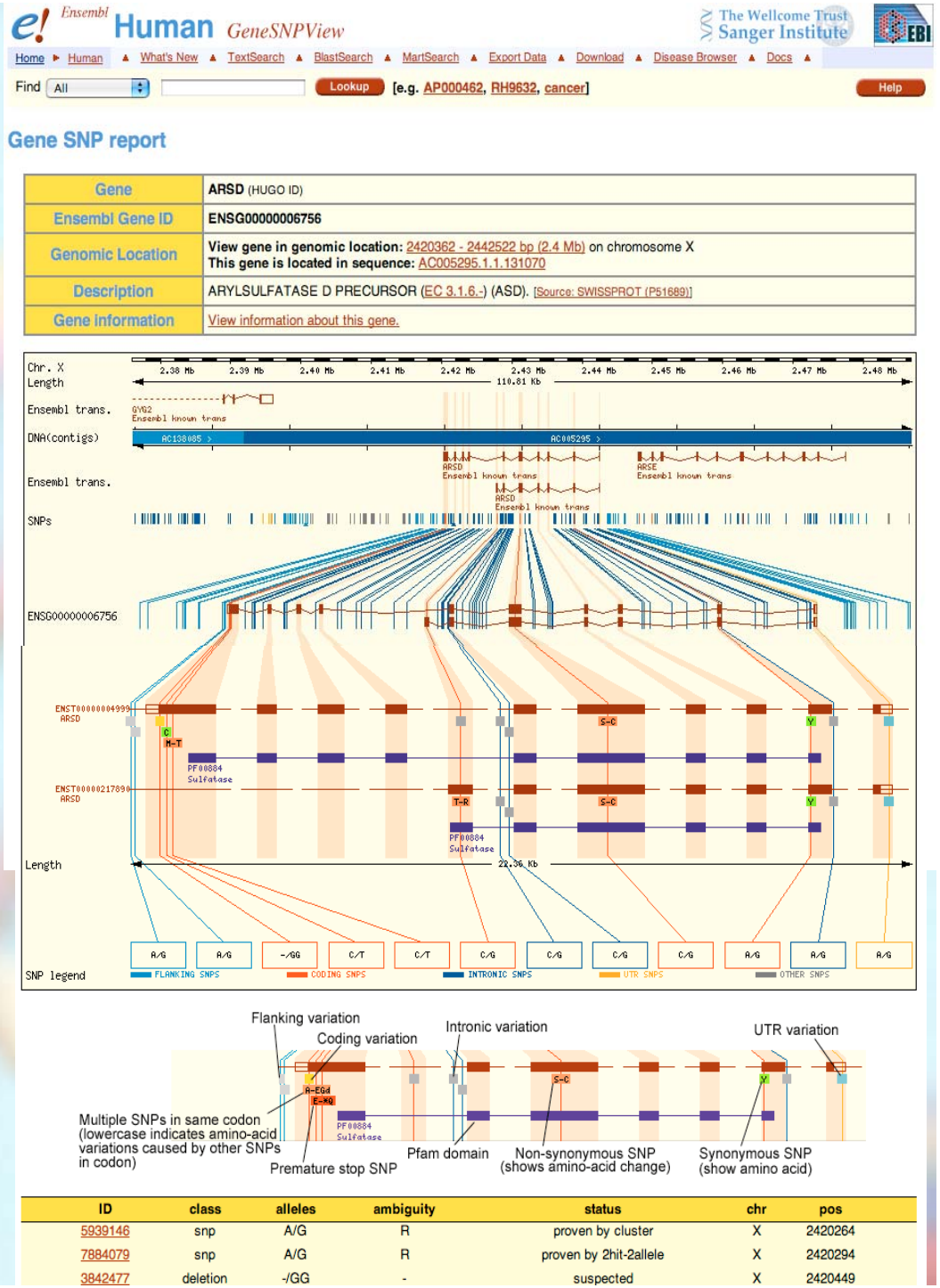
- Resequencing the BRAF gene DNA found mutations in cancers (Davies et al, 2002).





- Mutations are found in 66% of malignant melanomas and at lower levels in other cancers.
- Development of specific inhibitors is under

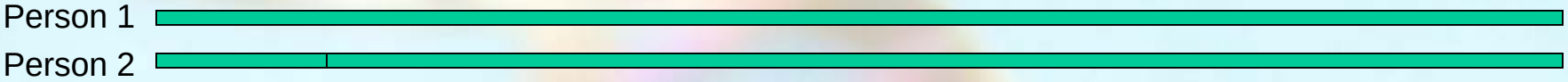
Ensembl was used to design the resequencing experiments, and can visualise the mutations in context.



ID	class	alleles	ambiguity	status	chr	pos
5939146	snp	A/G	R	proven by cluster	X	2420264
7884079	snp	A/G	R	proven by 2hit-2allele	X	2420294
3842477	deletion	-/GG	-	suspected	X	2420449

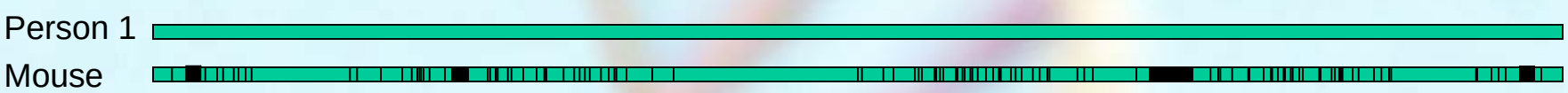
Conservation of genes across organisms

1 difference per 1000 base pairs between people

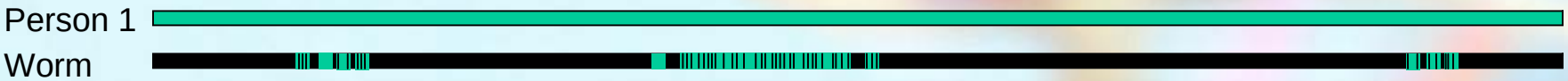


2 differences per 100 base pairs between people and chimp

40 differences per 100 base pairs between people and mice



Small regions of similarity between people and worms



The regions that code for proteins are most similar

60% of *C. elegans* genes have a recognizable human counterpart

```
C. elegans 7    LQCYHKGCGLLFDPKENDNEACTYHPGGPYFHDAYKIWTCCDKKSTDFGTWMNYKGCTRG 66
                L CY++GCG  FDP+ N ++ACTYHPG P FHDA K W+CC +++TDF  +++  GCT+G
Human      3    LLCYNRGCGQRFDPETNSDDACTYHPGVPVFHDALKGWSCCKRRTTDFSDFLSIVGCTKG 62

C. elegans 67   KHSNEKPVDIVKVAA-------VKEIRPEKEEDVIVWKGLNKSGKLDSKDATKRIEQNLN 119
                +H++EKP + VK          + E++P+ +E +I      ++ K  S D       NL
Human      63   RHNSEKPPEPVKPEVKTTEKKELCELKPKFQEHIIQAPKPVEAIKRPSPDEPM---TNLE 119

C. elegans 120  VEVTPGATAAIEK-KL----KEISEAAQSADIQIGAPCRNNGCSTEFDGSKN-KENCQHH 173
                ++++     A++K KL     E  +   + +I+IG  C+N GCS  + G ++ +E C +H
Human      120  LKISASLKQALDKLKLSSGNEENKKEEDNDEIKIGTSCKNGGCSKTYQGLESLEEVCVYH 179

C. elegans 174  PGAAIFHEGMKYWSCCNKKTSNFGAFLEQVGCTSGEHKFRNNEIVSKF---REDWFSSNG 230
                 G  IFHEGMKYWSCC +KTS+F  FL Q GCT G+H +   +   K    R D   + G
Human      180  SGVPIFHEGMKYWSCCRRKTSDFNTFLAQEGCTKGKHMWTKKDAGKKVVPCRHDLHQTGG 239

C. elegans 231  FVTINVYCRGALPETANIVSDGHTVRVSMKHGFGNASVDLDYDLWDEVIPEESRVVIGER 290
                  VI+VY + +LPE + + ++   + V +    G    D +  LW  +  + S V +
Human      240  EVTISVYAKNSLPELSRVEANSTLLNVIIVFE-GEKEFDQNVKLWGVIDVKRSYVTMTAT 298

C. elegans 291  KVEISLKQKHGTGWPRLKFDPELDAKNDEE 320
                K+EI++++     W  L    EL A   +E
Human      299  KIEITMRKAEPMQWASL----ELPAAKKQE 324
```

80% of human genes mutated in cancers have a *C. elegans* counterpart

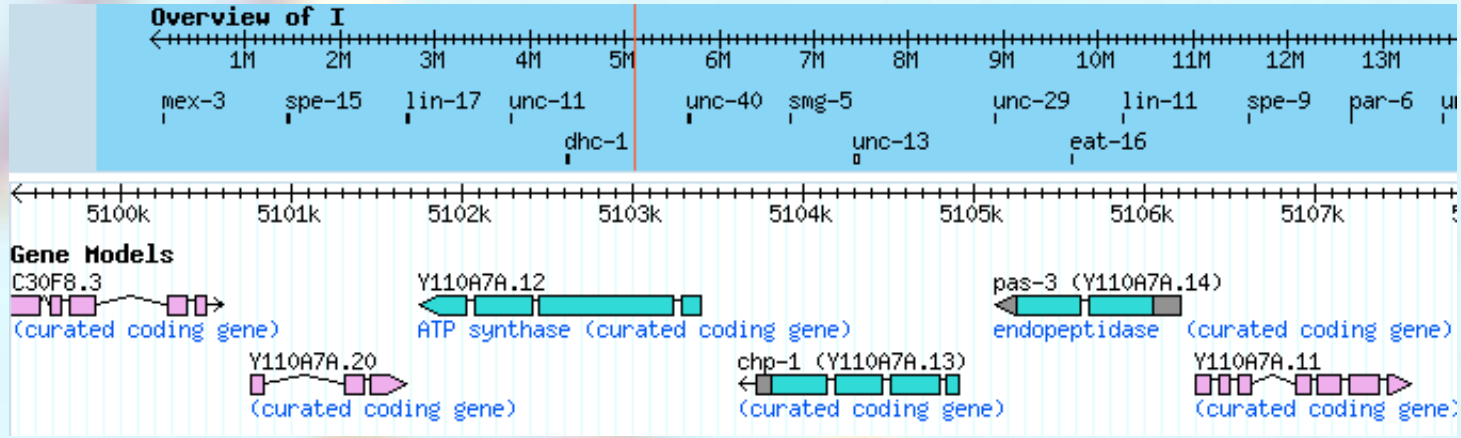
Why use worms?

We can investigate gene function in worms in ways that are impossible, or vastly more expensive, to do in higher animals.

e.g. we can look at what happens when each gene is removed.

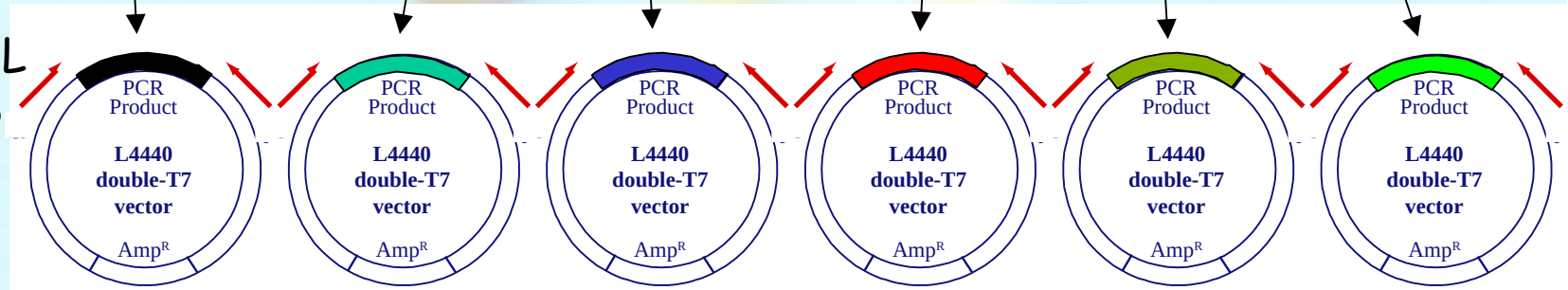
RNA interference disrupts a gene based on its sequence

MAKING THE WORM RNAi LIBRARY

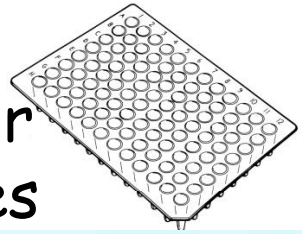


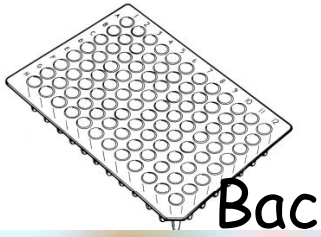
DNA fragments

BACTERIAL VECTORS (produces dsRNA)



Sequence for ~19,000 genes

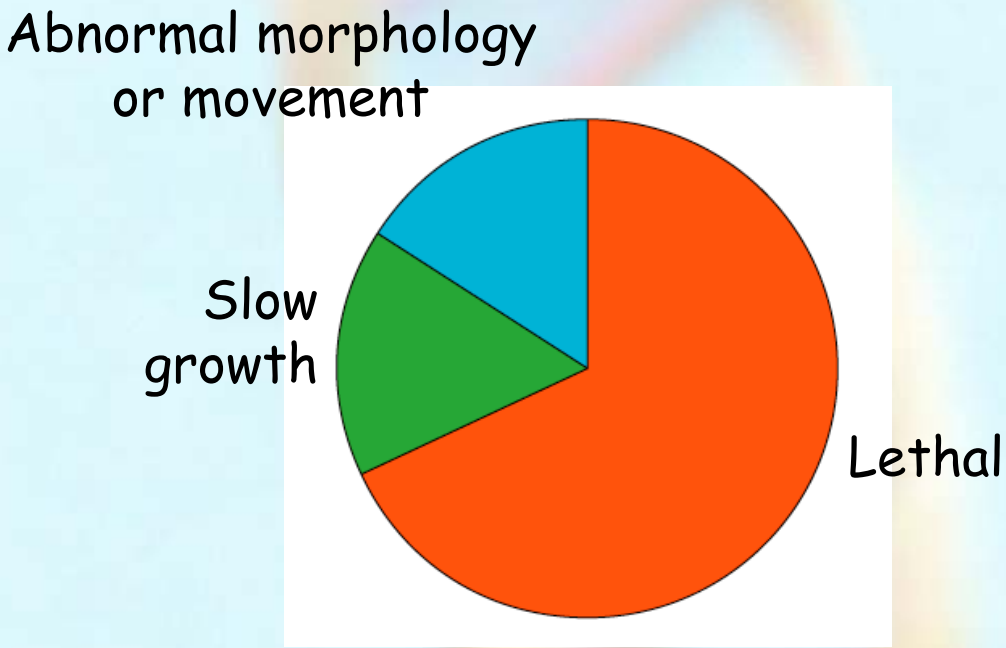




~17,000 Bacterial strains

Screen Summary

Screened 16,757 genes (Kamath et al, 2003)
1722 have an RNAi defect (only ~500 were known before)



1000 new gene functions were found by 3 people in two years

Genome-wide fat RNAi screen

Ashrafi et al (2003)

Detect fat by fluorescence after feeding worms Nile red (binds lipids)

Clones producing wt staining: **16,075**

Clones producing reduced staining (viable): **305**

Lipid/sterol metabolism, signal transduction molecules, transcription factors, channels, receptors

Clones producing increased staining (viable): **116**

Signal transduction molecules, transcription factors, channels, receptors

The same resource is being used by many research groups to study many different processes.

Viewing model organism phenotypes in Ensembl

Ensembl Gene Report

Gene	**MYH13** (HUGO ID)
Ensembl Gene ID	**ENSG00000006788**
Genomic Location	**View gene in genomic location:** 10404749 - 10475144 bp (10.4 Mb) on chromosome 17 **This gene is located in sequence:** AC005291.1.1.198582
Description	Myosin heavy chain, skeletal muscle, extraocular (MyHC-eo). [Source: SWISSPROT (Q9UKX3)]

...

GeneDAS Sources	☐ GAD (Genetic Association Database) ☐ HUGO_text (PubMed text-mining via HUGO symbol) ☑ New_source_1 ☐ Reactome (Knowledgebase of biological processes) ☐ UniProt (Protein knowledgebase) Manage Sources
	cele_phenotype family_07 Phenotypes of [C17E4.2] (C. elegans, associated through protein family): embryonic lethal (Emb) [SA:yk392e6] **cele_phenotype** family_08 Phenotypes of [K12F2.1] (C. elegans, associated through protein family): SLUggish (Slu) [JA:K12F2.1] // body morphology defect (Bmd) [Simmer:K12F2.1]

Open data resources

- All these studies (and many others) require access to the genome sequence in a computable form
  - Design of experiments
  - Interpretation of experiments
- To maximise the results of the studies we want to place them back on the genome
  - Make them accessible in the context of other results
  - The sequence as an index to biology
- We do this by having an open, extendable system
  - No constraints on what people do with the information

From the start, the Human Genome Project was open

- The Bermuda Statement, February 1996

  "All human genomic sequence information should be freely available and in the public domain in order to encourage research and development and to maximise its benefit to society."

- The (public) human genome project data were made available prior to publication:

  Assemblies of 1-2 kb are deposited in the public database (GenBank, EMBL) every 24 hours.

  No patents are filed.

HUMAN GENOMIC SEQUENCE GENERATED BY LARGE SCALE CENTRES:

RELEASE

- Automatic release of sequence assemblies >1 kb (preferably daily)
- Immediate submission of finished annotated sequence
- [redacted]
- Aim to have all sequence freely available and in the public domain for both research and development, in order to maximise its benefit to society.

POLICY

- The funding agencies are urged to foster these policies

Overhead from the Bermuda meeting at which the principles of free data release were agreed

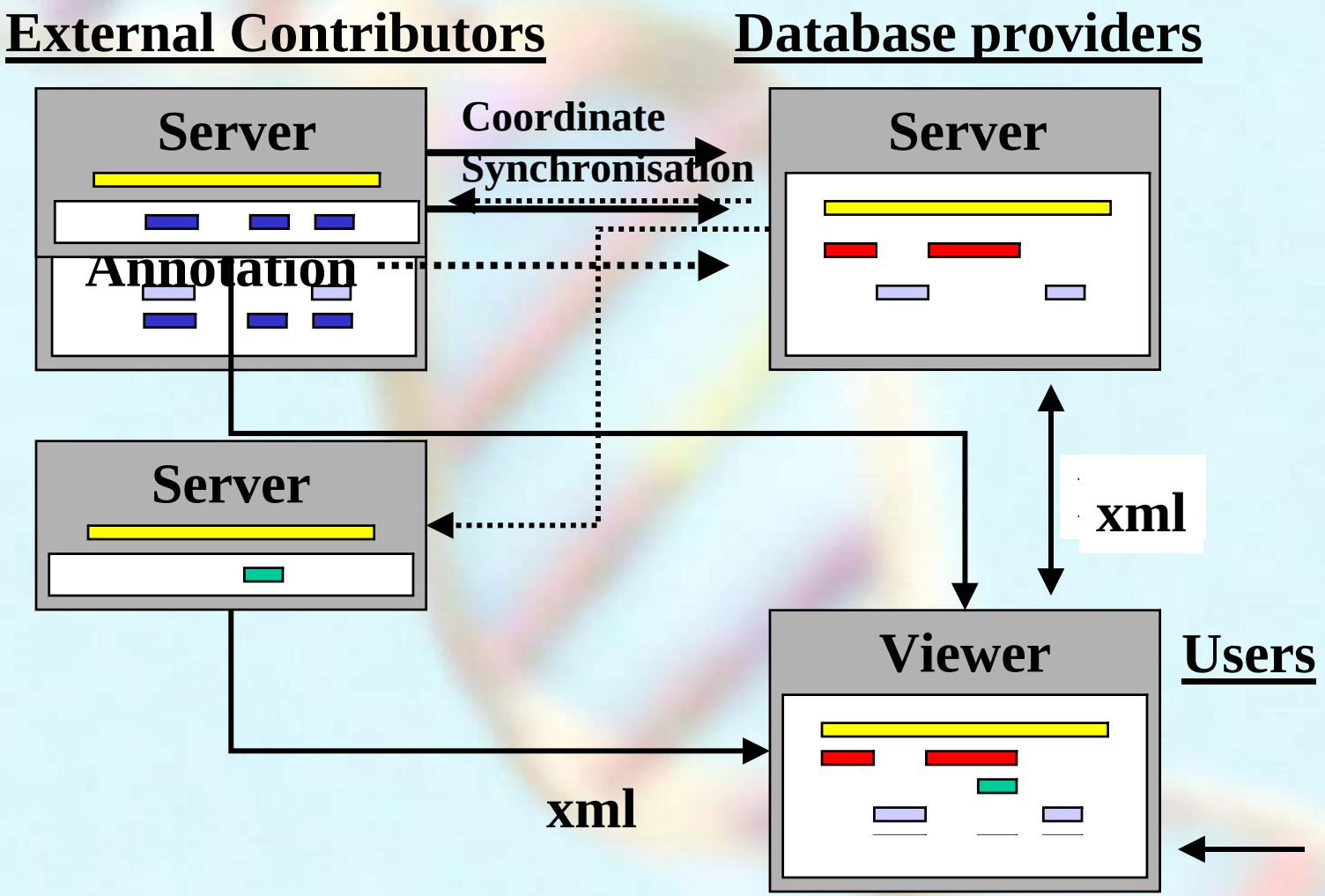
The value of working in the open

- Bermuda was in part to maximise the value of the product, but also to maximise effective collaboration
  - Transparency of progress and ability to verify by third parties
  - Decoupling of presentation and analysis from production
  - Rapid feedback from users
- It was also central to the competition with Celera
  - The Celera genome was only available under restrictive licence: free local use for academics and no redistribution
  - Working in the open ensured the ultimate quality of the public sequence and its becoming the reference sequence
  - (Bermuda was prior to the formation of Celera: Venter was there and signed up to it!)

Ensembl is an open project

- The data are freely available
  - The sequence is free.
  - The gene sets and other annotation we generate is free.
  - Many external free data sets are integrated and displayed.
  - The database can be downloaded for local use, or accessed directly. Many companies have private copies.
- The software is free
  - All Ensembl software and documentation is open access.
  - This includes the entire web site, low level API and data analysis pipeline.
- This and similar projects have fostered the development of open access protocols (e.g. BioPerl,

Distributed Annotation Server (Stein et al.)



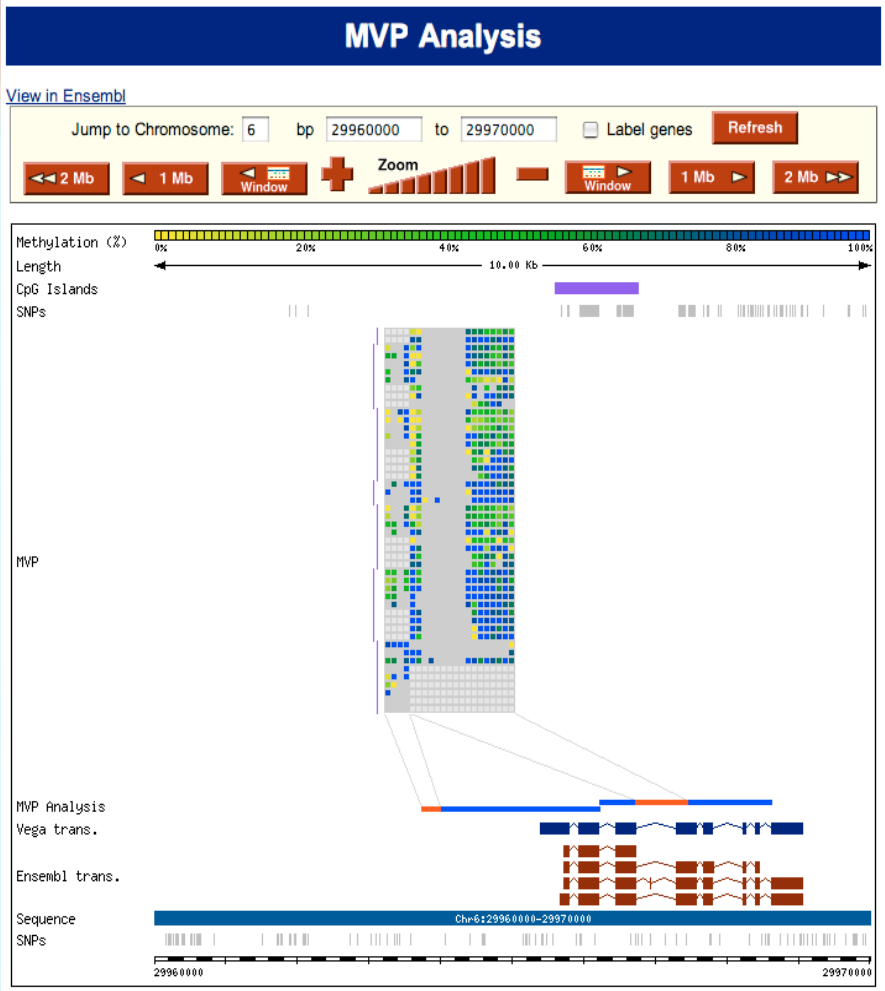
DAS for functional genomics

All data from DAS sources

HEP Human Epigenomic Project data

Vega genes

Ensembl



DAS is used extensively inside Sanger, on the Ensembl web site, and increasingly elsewhere (e.g. EU BioSapiens Network of Excellence)

DAS like model applied to other data types

- Features on a linear sequence
  - DNA, protein sequences, protein structures
  - MRC eScience protein family integration project (SCOP, CATH, Pfam, InterPro, MSD) developing DAS for protein structures.
- Annotation connected to stable identifiers
  - Genes, e.g. worm phenotypes, user note book

Other human genome data sources

- UCSC (genome.ucsc.edu) and NCBI (ncbi.nlm.nih.gov) are also open data resource providers
  - All three provide complex environments, linking in functional studies to support knowledge growth.
  - Others can use data in their own systems, or feed back into the central resources without fear of a transfer of “rights”.
- Competition in the open (together with collaborative sharing of progress) gives dynamic response to new research interests
- Many other resources provide information about genes or specific data types or organisms, all

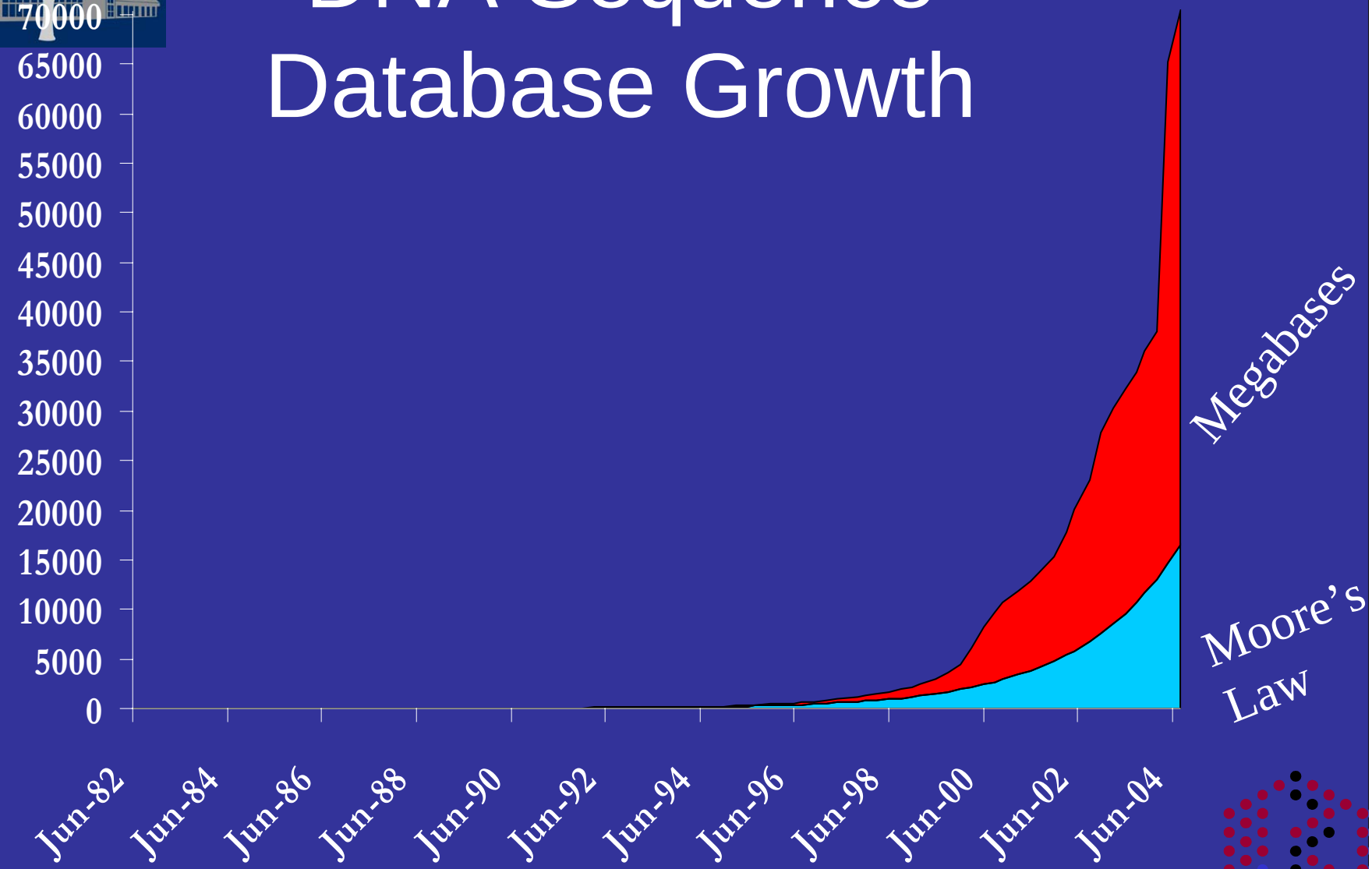
Open Access Principles

- **Publication is the central mechanism of academic research**
  - The goal of academic research is to advance knowledge.
  - It proceeds in a market fashion, with the product being publication of novel results.
  - Publication is necessary to achieve recognition and future support. Researchers want to maximise the impact of their publications. There are various proxies, but ultimately impact is how much their results influence others.
- Publication means making results available for others
  - The system relies on the ability of others to use published results in new research: data as well as ideas.
  - This is much more important than verifying results.

Where are we heading?

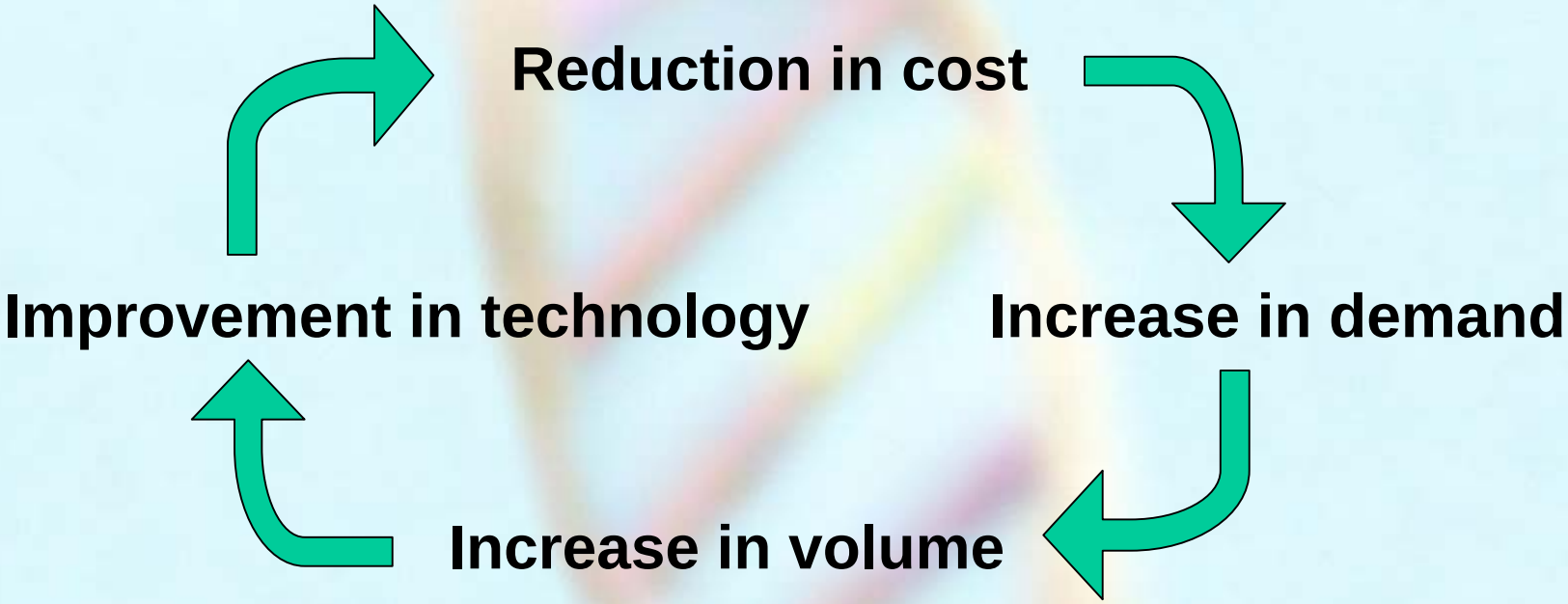
- The growth of genome sequence data
  - Sequence information is fundamental to biology: we have only just started!
- Many other data types from high throughput methods indexed on genome sequence and genes
- Genetics
  - The ultimate target is to sequence individual genomes
  - Moving from research to clinical use

DNA Sequence Database Growth



EMBL-EBI

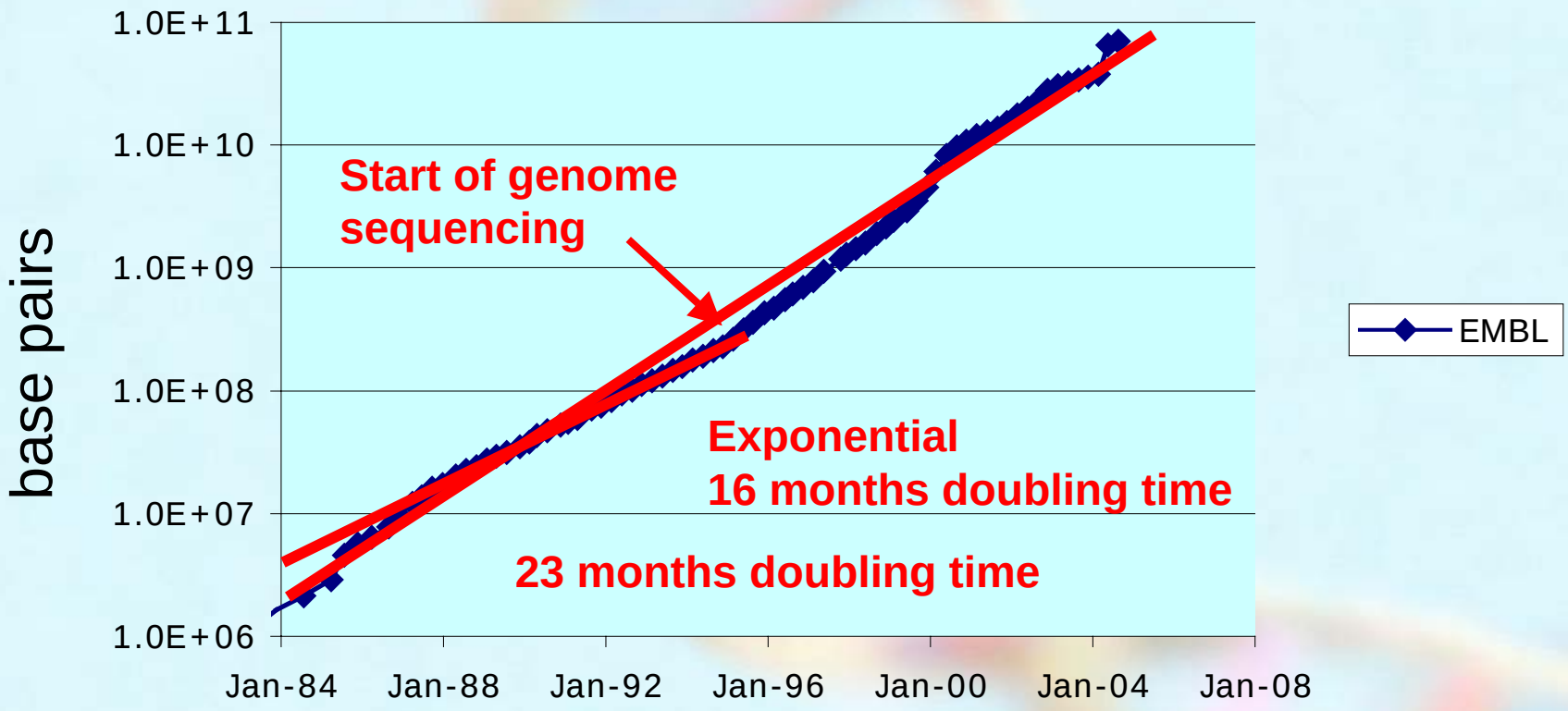
How is exponential growth sustained?



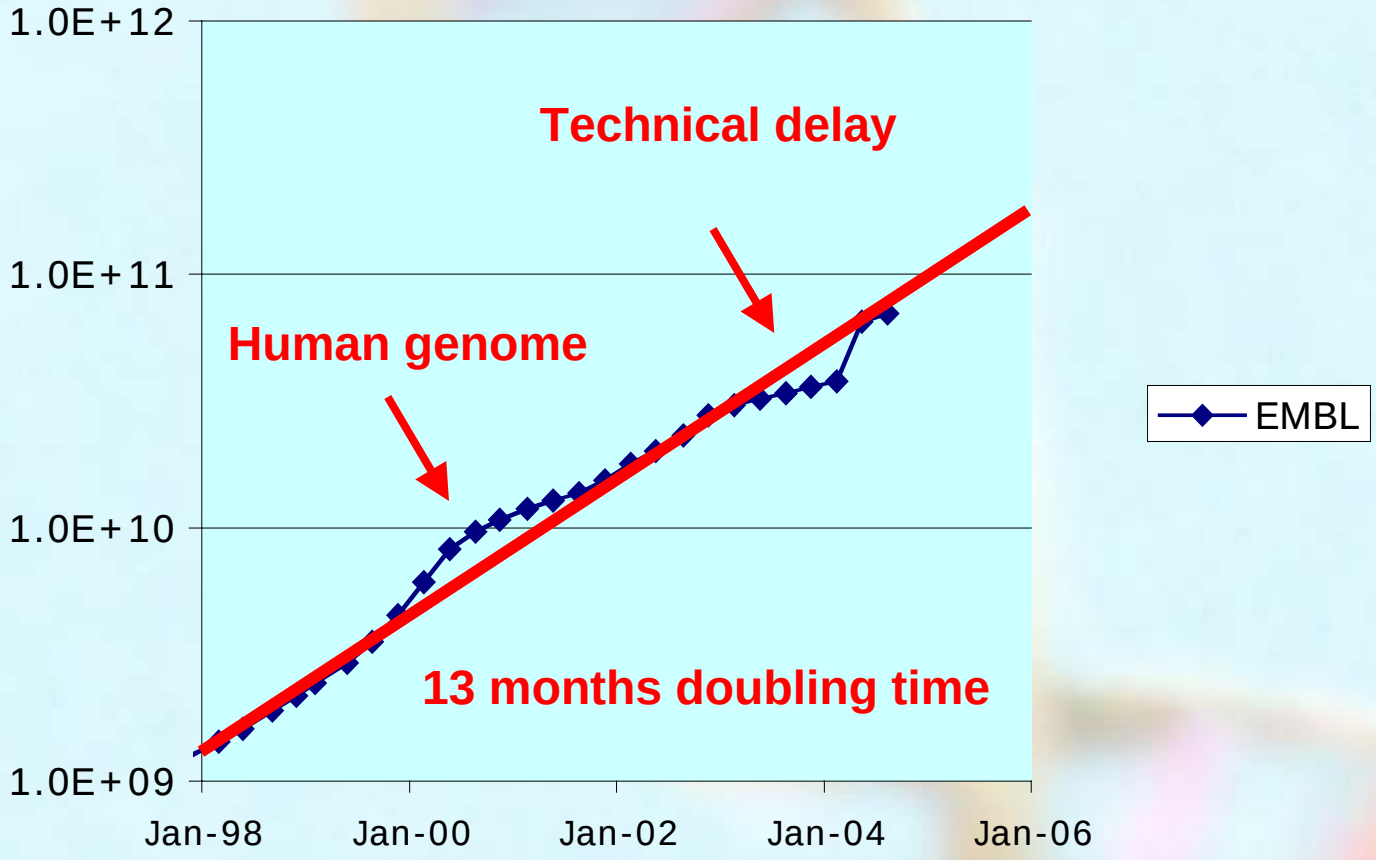
This clearly doesn't work for all technologies. Why computing and genome sequence data?

The key is that these are **information** activities; there is no inherent physical outcome, or constraint.

Growth in Public Sequence Data



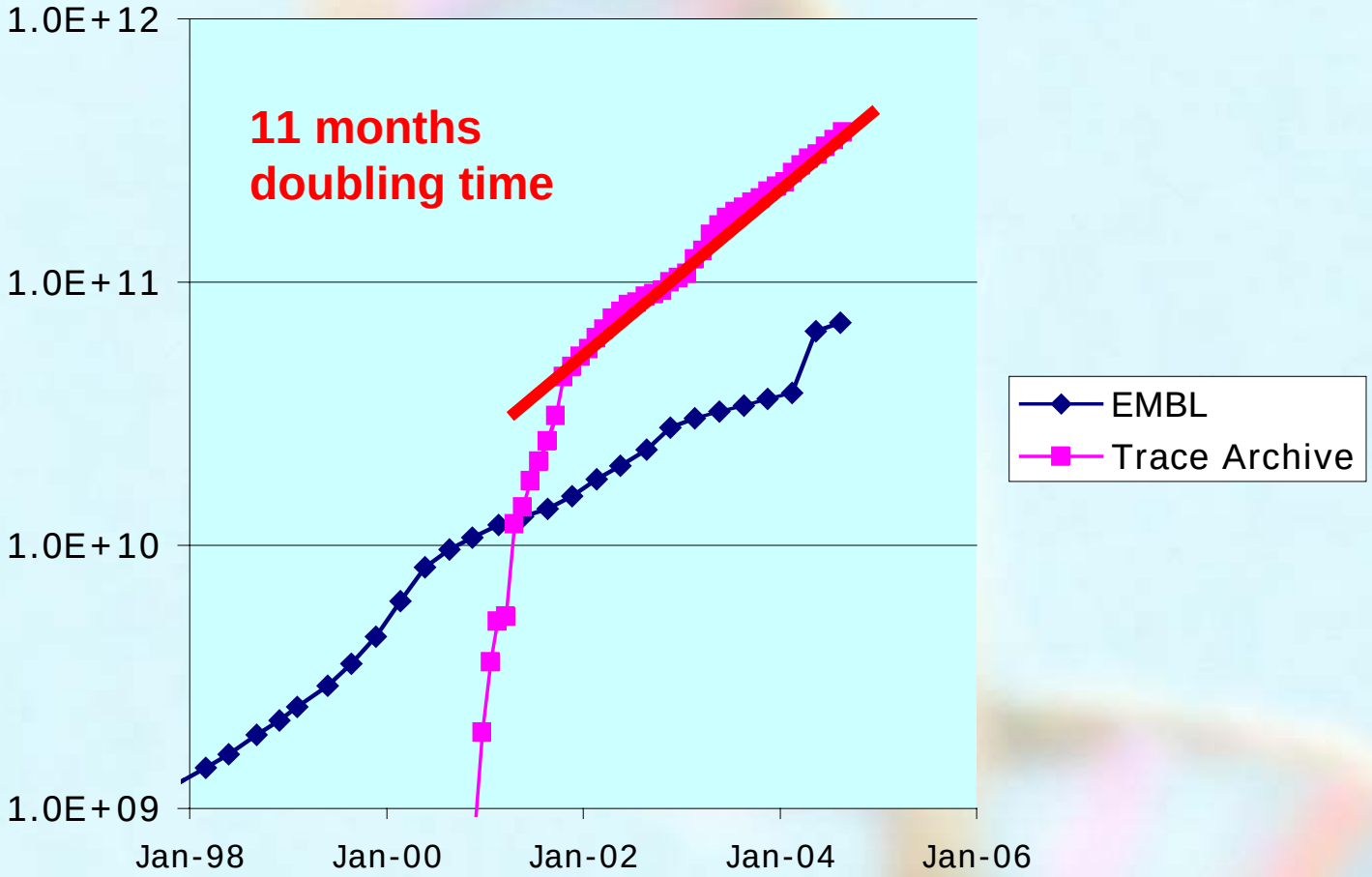
The last few years



The era of sequencing genomes

	Size (Mb)	Genes			Completion date
H. influenzae	2	1,700	1/1kb	Bacterium	1995
Yeast	13	6,000	1/2kb	Eukaryotic cell	1996
Nematode	100	20,000	1/5kb	Animal	1998
Human	3000	?30,000	1/100kb	Mammal	2000/3
Mouse, fish (3), Ciona (2), mosquito, rat, chimp, chicken, frog, pig, dog, cow, a marsupial, sea urchin, 200 bugs, more worms and flies					2002-5

A new repository: the Trace Archive



Lessons

- Sequencing is growing (at least) as fast as ever
  - Faster than Moore's law
- Lots of new genomes (EMBL/Genbank)
  - Genomes of organisms of interest in their own right: mouse, chicken, rice, pig
  - Related genomes, useful for comparative analysis
- Additional sequence growing faster (Traces)
  - Variation: human and others, e.g. cancer

Where next?

- No reason for growth in technology to slow
  - All sequencing so far has been done with Sanger chemistry. This can scale with more sensitive detection.
  - Alternative technologies are beginning to work
    - Hybridisation: Perlegen, Lynx
    - Enzymatic: 454, Solexa (UK, merging with Lynx)
- We have around 10 Human Genome Equivalents (HGEs) now (10x more raw data)
- Next major goal is resequencing individuals' genomes for research and health
  - To resequence a genome now would cost ~£10M

Resequencing human genomes

- What scale is required?
  - Population structure research requires hundreds
  - Disease genetics research requires thousands
  - Clinical use requires tens of thousands to millions
- On the (conservative) basis of 2-fold growth/year
  - In 5 years, population structure and resequence functional sequence (5%) for disease research
  - In 10 years, 10,000 HGEs: whole resequencing for research
  - In 20 years, 10 million HGEs: clinical use
  - In 30 years, 10 billion HGEs: sequence as part of the standard medical record

Coda: what about maps?

- Maps are fundamental to our knowledge and use of our environment

Coda: what about maps?

- We can now distribute and manipulate them with layered annotation using computers

Coda: what about maps?

- But such uses are blocked by restrictive licences and proprietary systems

Coda: what about maps?

- Why not open reference maps and Map-DAS as a public good?

Conclusions

- The result of the genome project is information
  - There have been many technical challenges in managing and using this information
  - But some of the biggest challenges, and benefits, were political, in establishing and developing openness.
- To maximise the value of the genome sequence
  - Give away the reference information, to everyone
  - Build knowledge from it, then annotate back
  - Continue to sequence!

Thanks to all my colleagues, and to the