All your base(s) are belong to us

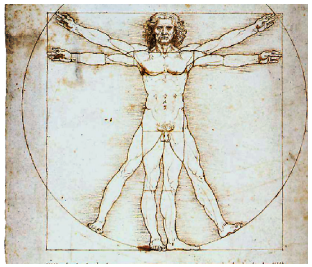
The dawn of the high-throughput DNA sequencing era

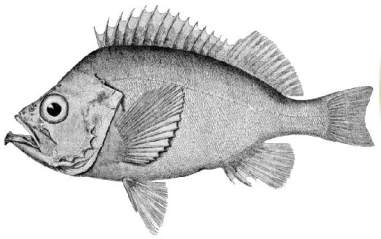
Magnus Manske

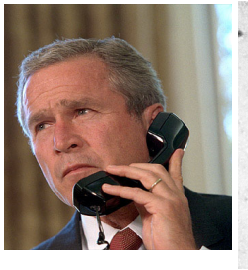
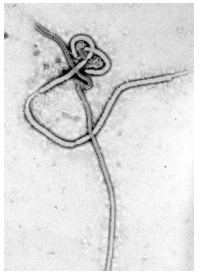
The place

Sanger Center, Cambridge, UK

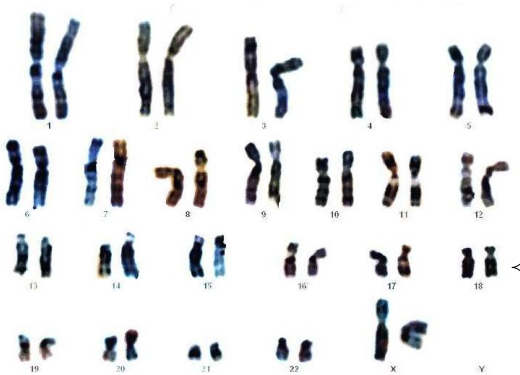
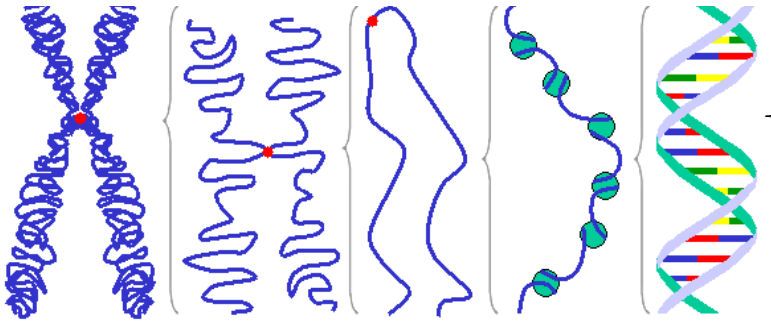
Basic biology





Level of complexity

Genome (all chromosomes of an organism)

Single chromosome

DNA

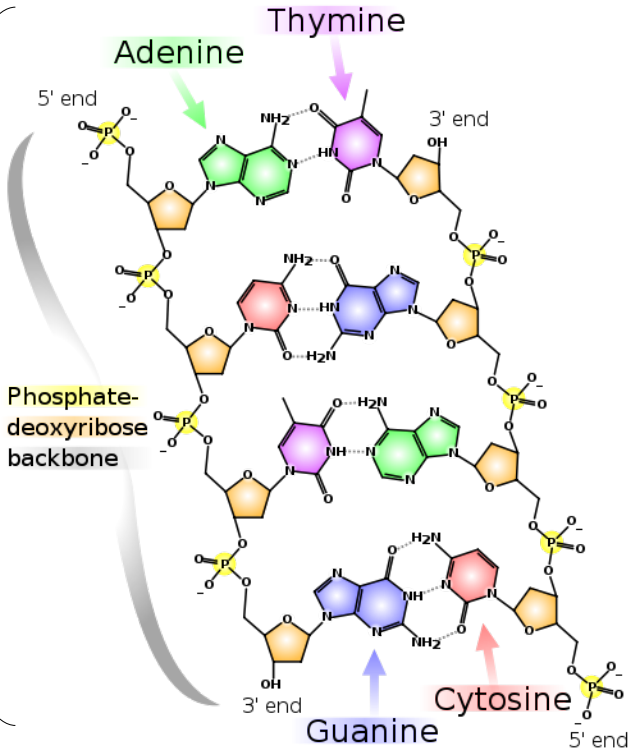


Image by Madeleine Price Ball [CC-BY-SA/GFDL]

Genomes and DNA

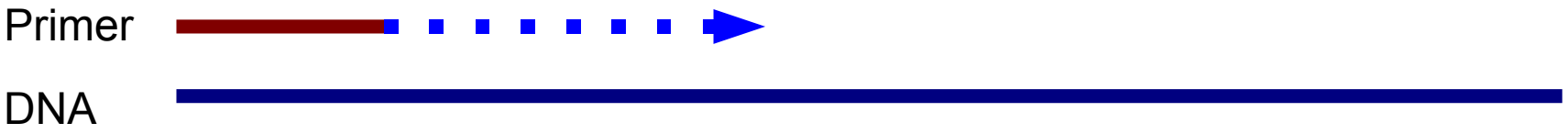
- DNA written as sequence of first letters of **A**denine, **C**ytosine, **G**uanine, **T**hymine
  e.g. ...AGGTTCGAGAGCCGCATGAGAC...
- Each “letter” called nucleotide or *base*
- Genome size does not necessarily correspond to organism size/complexity

Organism	Genome size	Order of magnitude
Bacteriophage MS2	3569 bases	10^3
Yeast	12 million	10^7
Human	3.2 billion	10^9
Amoeba dubia	670 billion	10^{11}

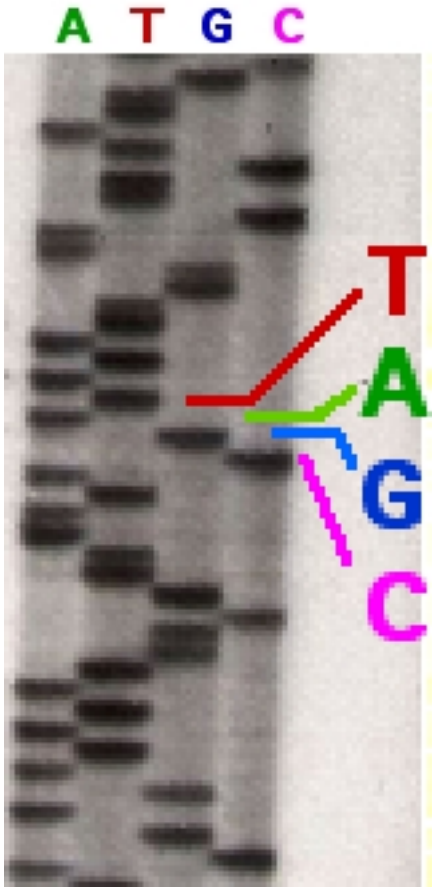
Sequencing – Episode 1

- First sequenced genome:
  Bacteriophage MS2 (1970s by Walter Fiers)
  3569 bases, using several man-decades
  RNA sequencing (RNA genome)
- Early DNA sequencing methods
  e.g. “wandering spot” (Gilbert and Maxam)
  yielding 24 base pairs per experiment
- Chain-termination sequencing method
  developed by Frederick Sanger in 1975
  first “rapid” DNA sequencing method

Chain termination sequencing

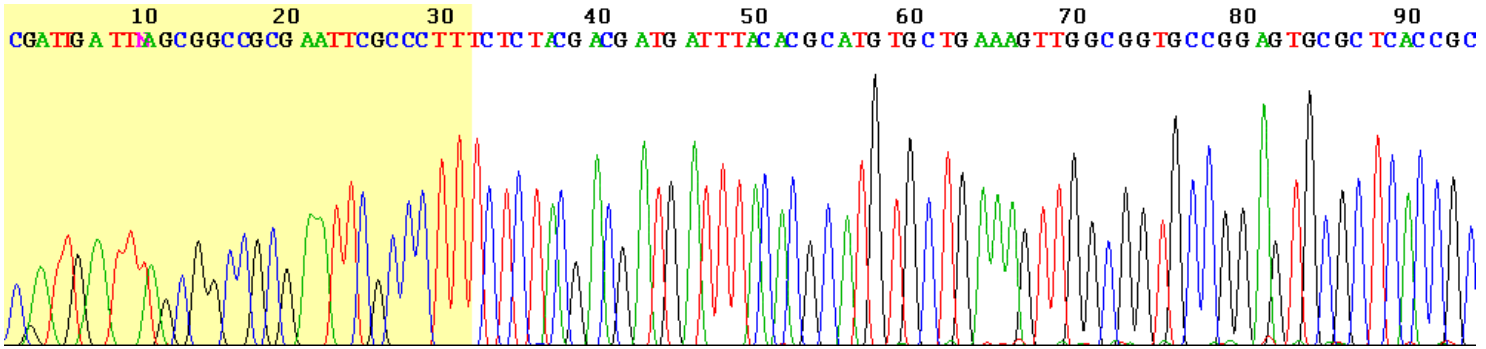


Radioactive gels



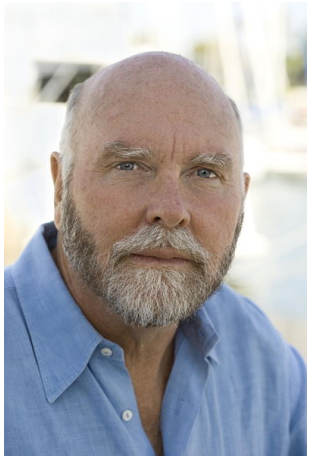
John Schmidt [GFDL]

Fluorescent dyes



- Sequences 300-1.000 bases en block
- Needs primer to define a starting point

Human Genome Project (1990-2003)



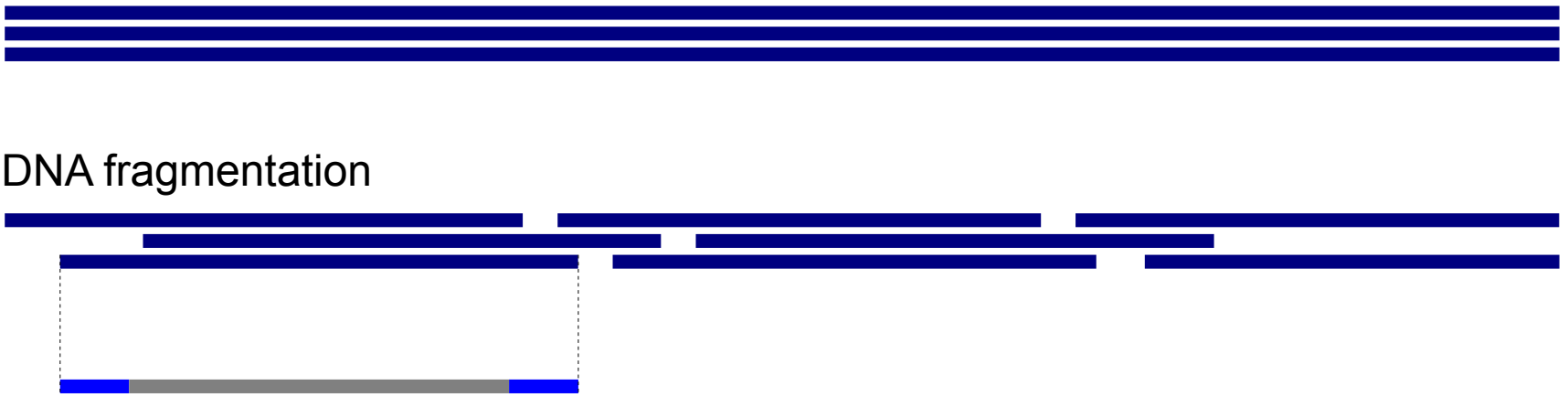
Craig Venter
PLoS [CC-BY-2.5]

Whole Genome Shotgun Sequencing

* Sequencing starts at random position
* Little need for scientific supervision
* Scale up by simply adding more machines and operators
* Use computers to reassemble the genome

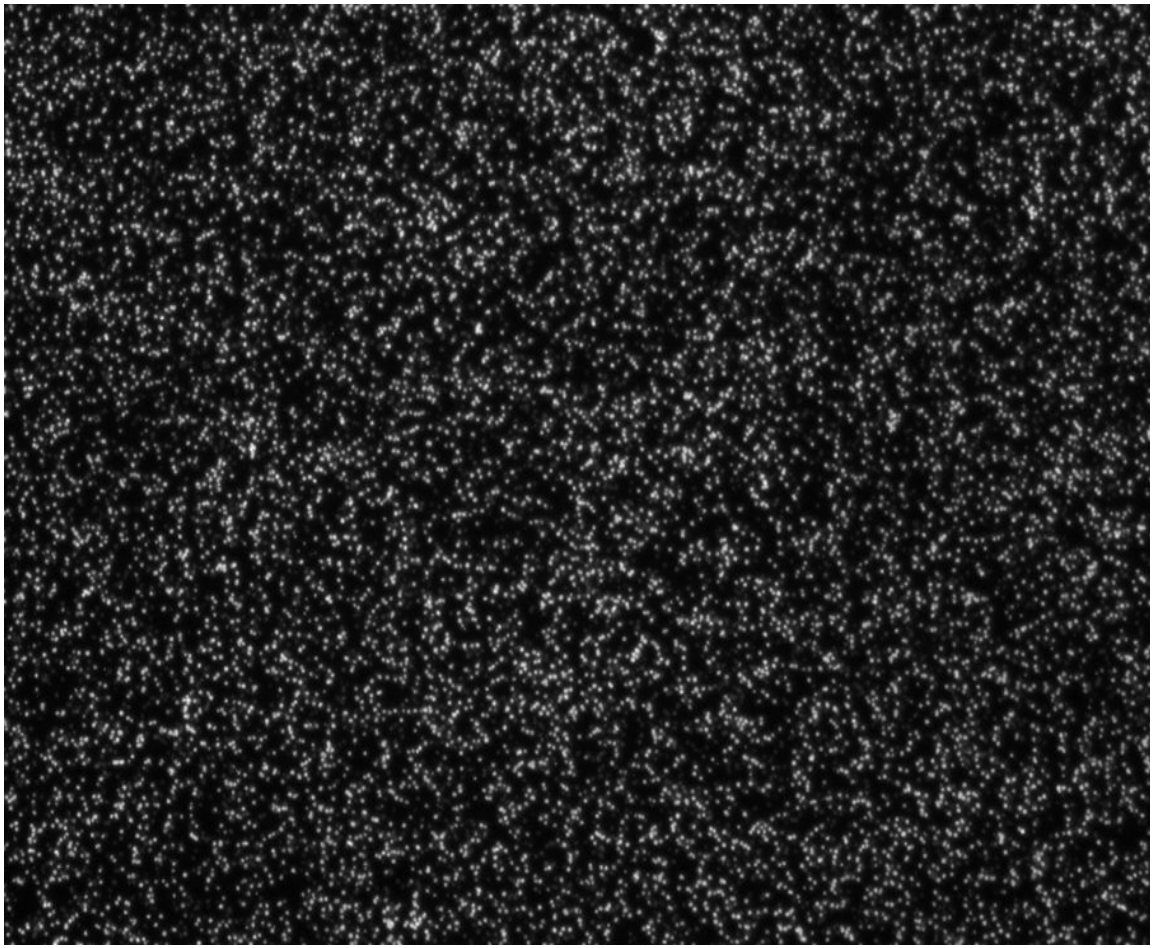
Solexa “new technology” sequencing

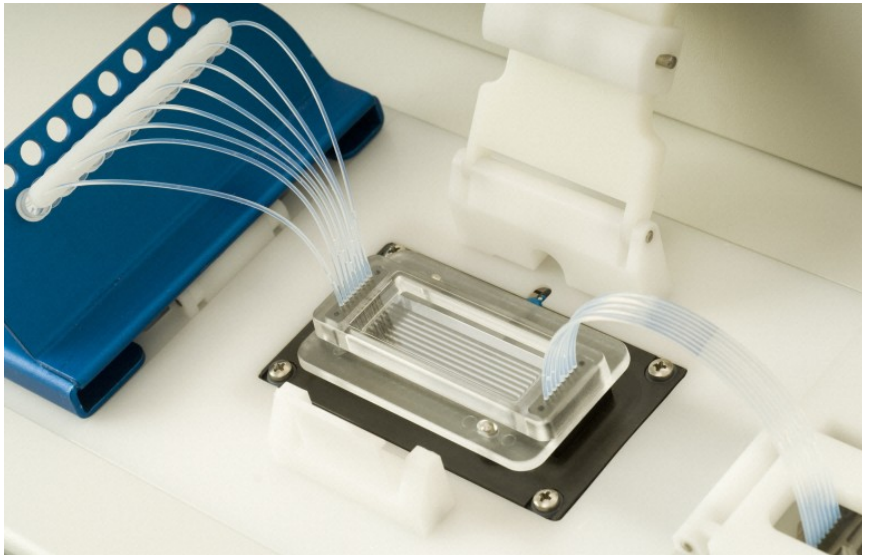
Chain termination method : Few, long reads
Solexa : Short reads, but buckets of them (shotgun)



DNA fragments are affixed on a glass “chip”
Single fragments are amplified in place to clusters
Sequencing through amplification using fluorescent dyes
Detection through digital photography

Some pictures





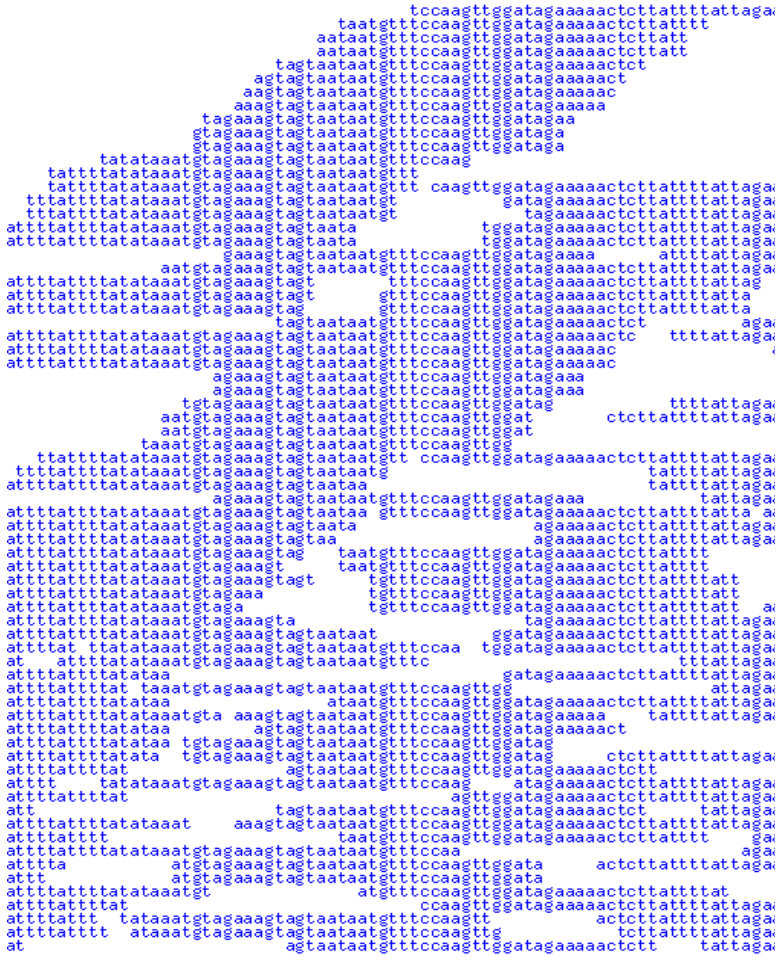
Pictures courtesy of Harold Swerdlow

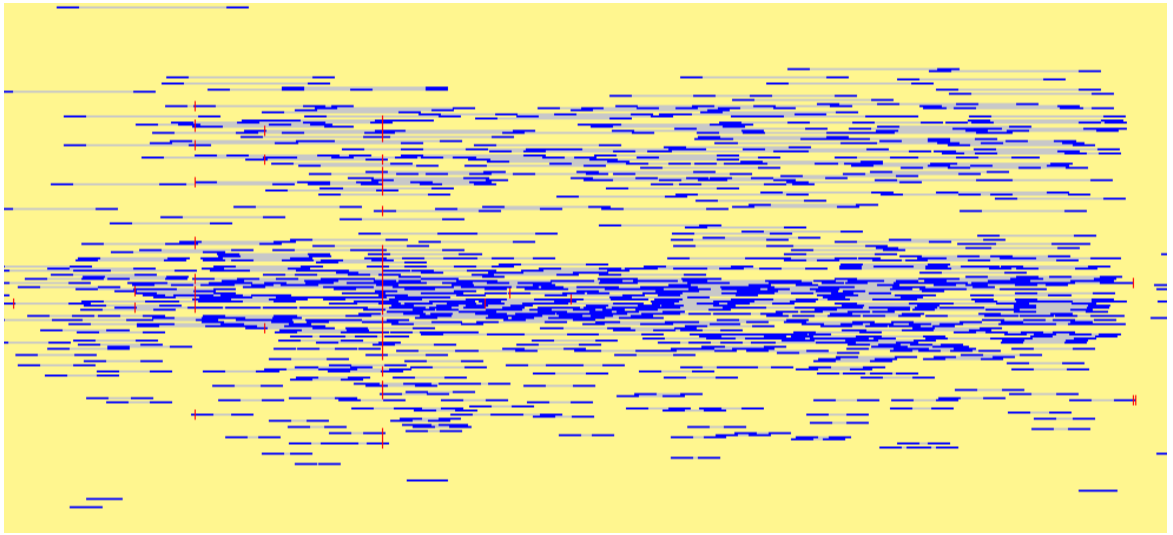
Data processing

Each run produces ~4TByte of image data
Processed into sequence data and quality estimate
Long lists of this:

```
@IL4_1106:1:1:539:515
GTACTATATATATAATATAAATATATATATAATAACA GCATTATAGAAAATATAAATATAAACATTTTTATGTT
+
>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>><<<<;; <<<<<<<<<<<<<<<<<<9<<<<<<<<<<<<<<<<
```

ca. 200
bases missing





New tech? So what?

- All DNA sequences so far stored in GeneBank
  200 gigabases (billion bases) at the end of 2007
  the sum of 35 years of world-wide sequencing
- In the first three months of 2008, the Sanger Center produced more than 200 gigabases of new sequencing data
- With 28 Solexa machines and 20 people
- Today: Capacity at Sanger is 7.5 terabases/year
  ~20 gigabases per run per machine
- Will probably double during next year

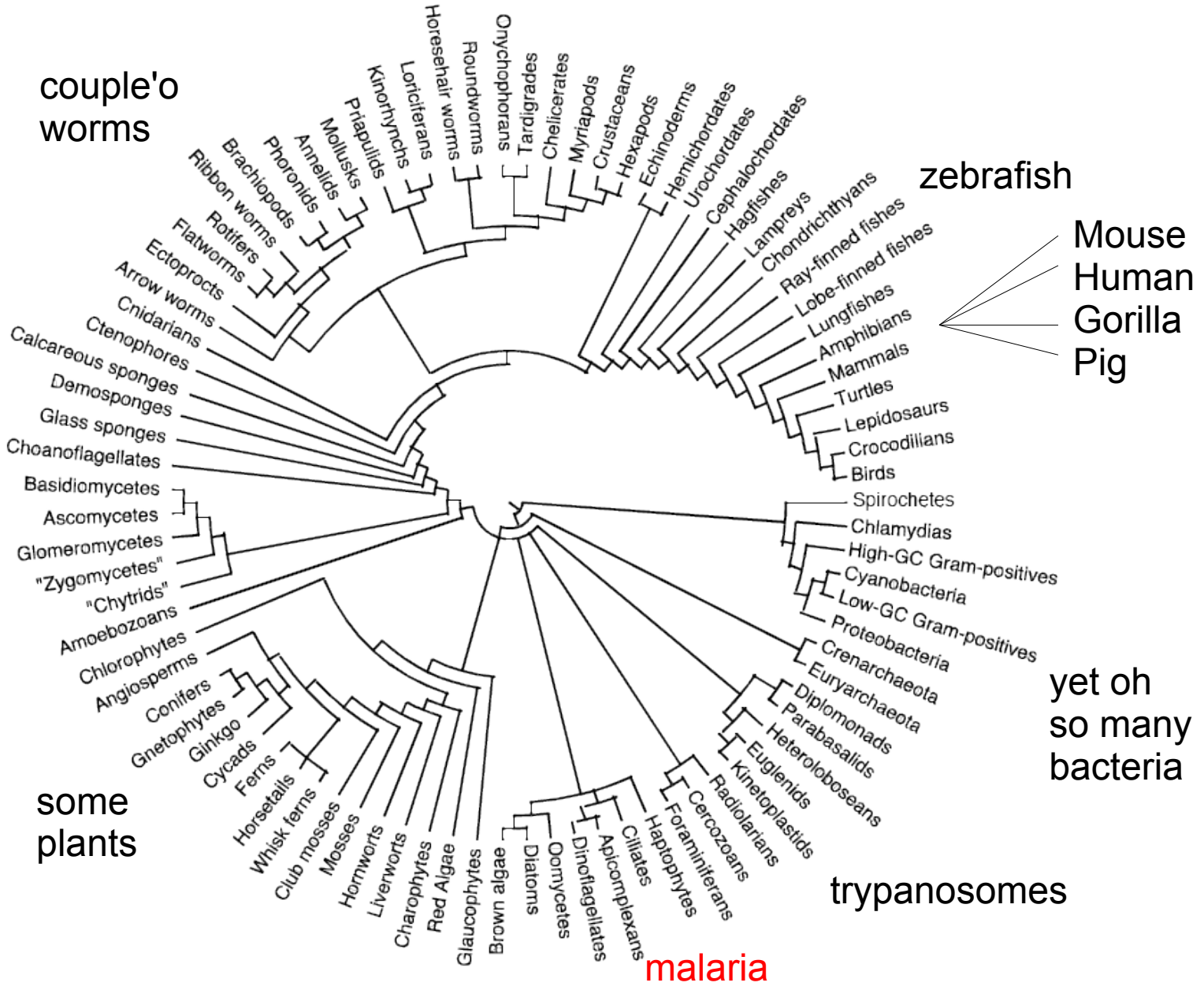
A paradigm shift

- Until this year : “I need my gene sequenced!”
- This year : “WTF should I do with all the sequencing data?”
- Reassembly is not fully automated
- New genomes need annotation
- Automatic annotation does not work well
- Manual annotators don't scale well

Fun with data storage and processing

- Output of all sequencing machines
  ~300 TByte/month
- Current storage capacity at Sanger
  ~3 PByte
- Backup is a problem
  it will take more time to restore backups from tape than to rerun the sequencing
- Currently ~4000 CPU cores in computing farm for image analysis / (re)assembly / etc.

What we use this for



couple'o worms

zebrafish

Mouse
Human
Gorilla
Pig

yet oh so many bacteria

some plants

trypanosomes

malaria

Fun stuff:

Neanderthal

Woolly mammoth

Platypus (alive!)

OK, how much for my genome?

- Human Genome Project:
  1 human genome = $500M
- Current chain termination sequencing:
  1 human genome = $10M
- Current Solexa sequencing:
  1 human genome (20x) : $100K
- Cheap enough e.g. for the 1000 genomes project

That still sounds expensive!

- But company XYZ says they'll do it for $400!
- They don't do sequencing – they do *genotyping*
- Checking for Single Nucleotide Polymorphisms
- A chip contains ~1 million known DNA variants
  e.g. for genetic diseases, phenotypic markers
- Cheaper, faster, and easier than sequencing
- Finds only known variants
- High error rate when used on individuals

GATTACA – sooner than you think?

Pacific Biotech

- Single enzyme sequencing
- 20 nm hole in a chip
- Multiplexed
- One human genome in a few minutes for ~ $100
- Announced for 2010

Helicos

- Single molecule sequencing

Nanopore

- Detects DNA as it passes through a nm-sized pore

Some UK news in context

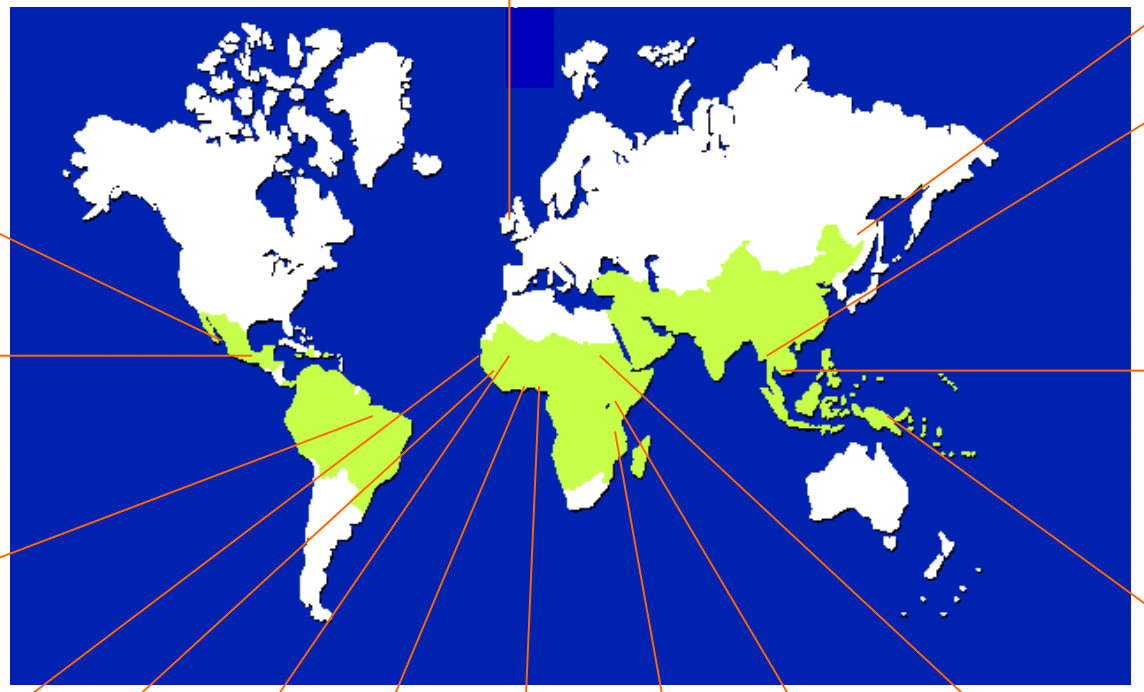
- “Forensic scientists could use DNA retrieved from a crime scene to predict the surname of the suspect, according to a new British study.” (BBC, 2006)
- “UK Police now hold DNA 'fingerprints' of 4.5m Britons [~7.5% of the population] … taken from individuals who were not charged with any offence, and have no criminal record” (Mail, 2007)
- “Police may be given power to take DNA samples in the street [from people for] littering, speeding or not wearing a seat belt” (The Guardian, 2007)
- “UK DNA Database Found To Violate Human Rights” (European Court of Human Rights, 2008)

Like any technology...

… it can be used for good and evil

- Malaria : 300-500 million infections per year
- 1-2 million deaths, mostly children under the age of 5
- Sequencing *Plasmodium falciparum* : 95 individual genomes now

Malaria sample distribution



3(UK)

1 (China)

Honduras (1)

3 (S.E. Asia)

El Salvador (1)

17 (Thailand)
20 (Cambodia)
1 (Vietnam)

Brazil (1)

Columbia (1)

10 Papua New Guinea

Gambia (3)

16 (Mali)

1 (Sierra Leon)

10 (Kenya)

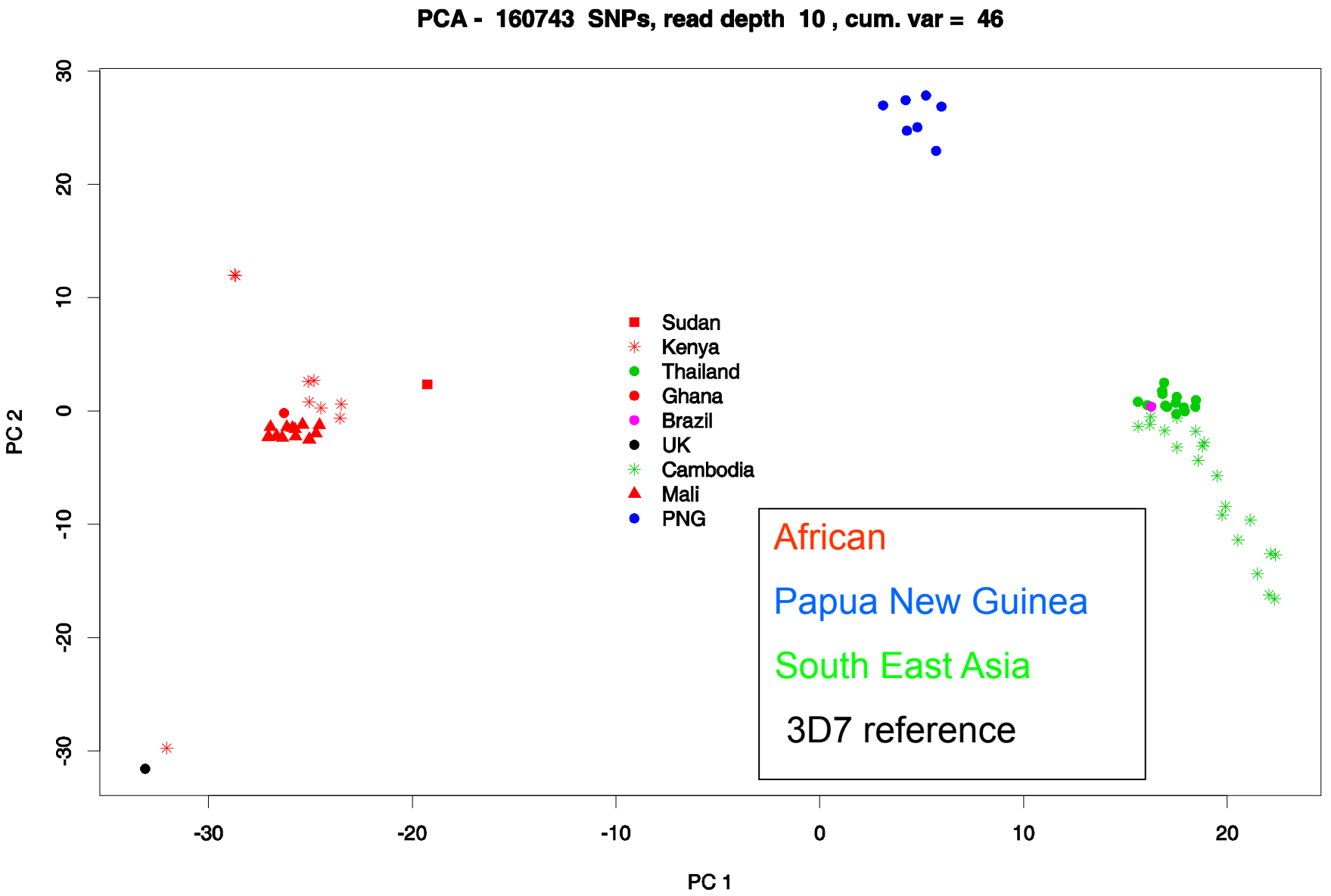
2 (Sudan)

Senegal (1)

2 (Ghana)

8 (Tanzania)

Variation clustering



Geographic visualizations

- Frequency
- PCA plots

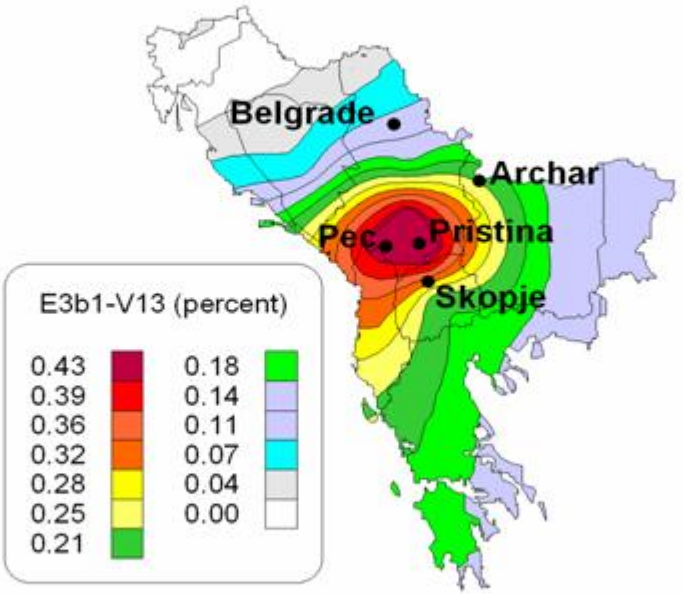
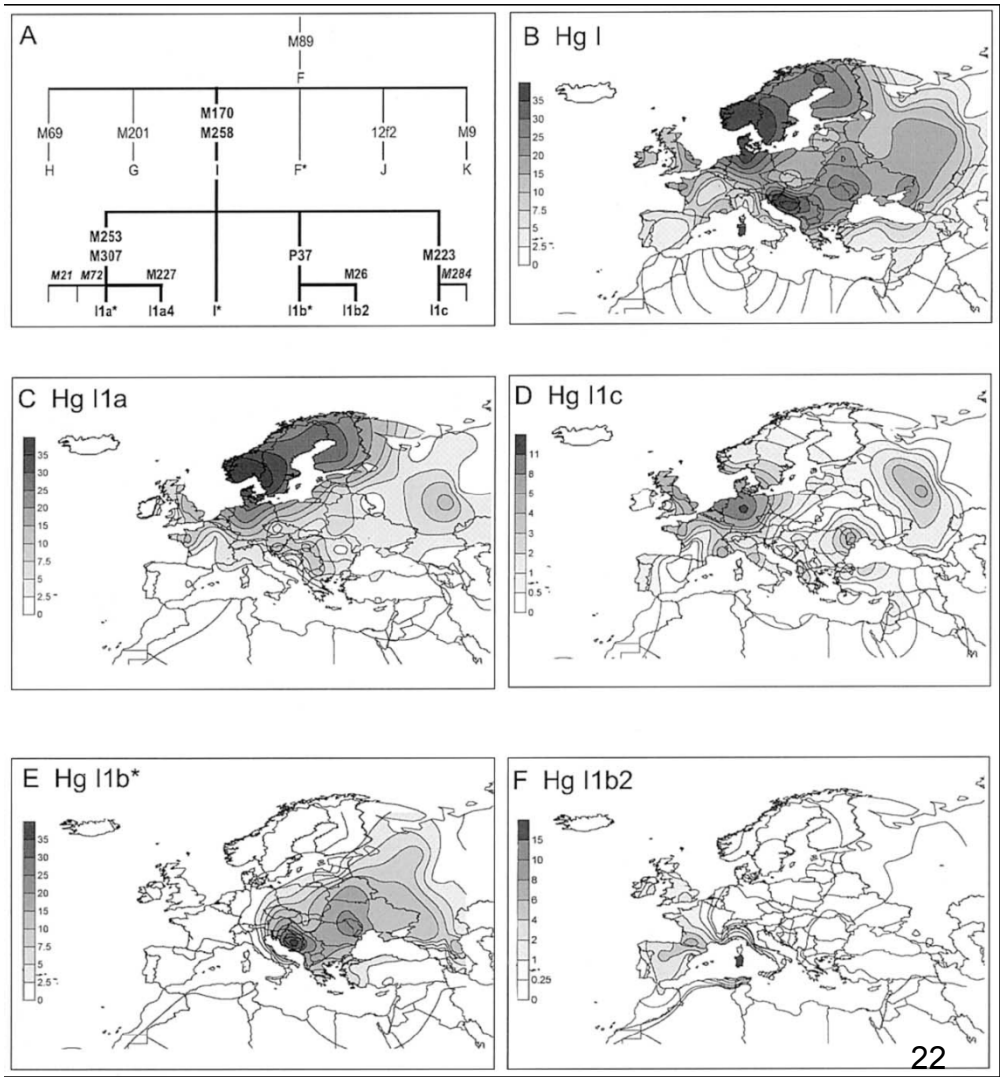
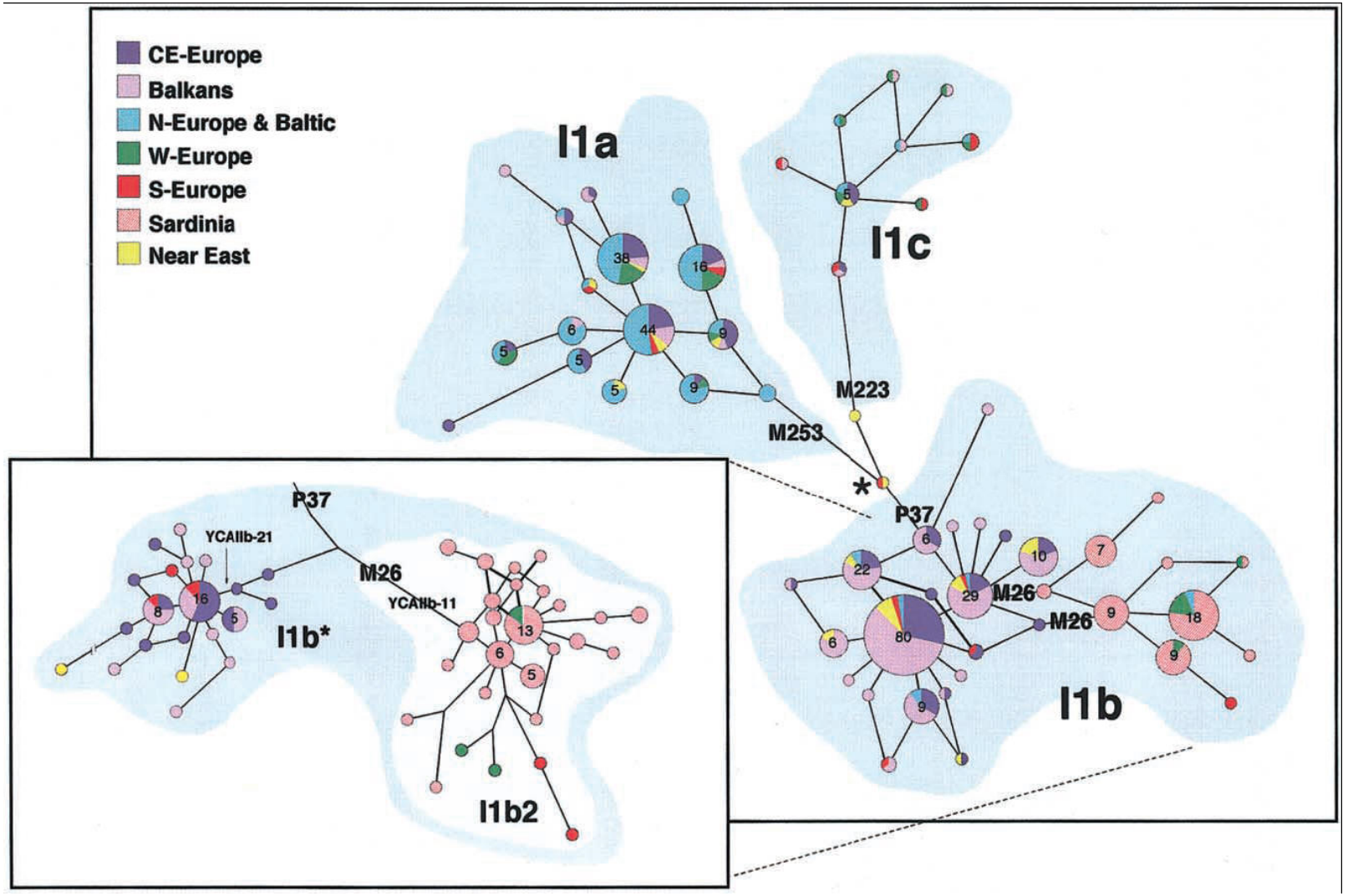


Figure 8. E-V13 Frequency Contour (Kriging Method), Southern Balkans.



Population “networks”



And much, much more...

- Multi-clonal infections
- Parasite travelling times and routes
- Insertions, deletions, inversions
- Copy number variation
- Detection of completely new variation (new genes etc.)

UAG (the end)