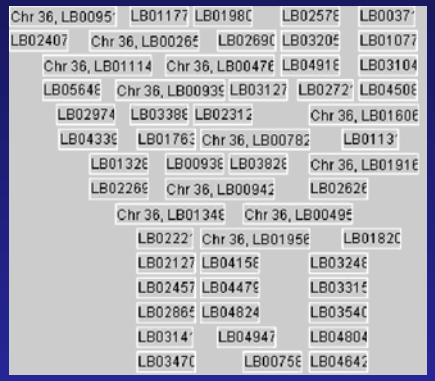
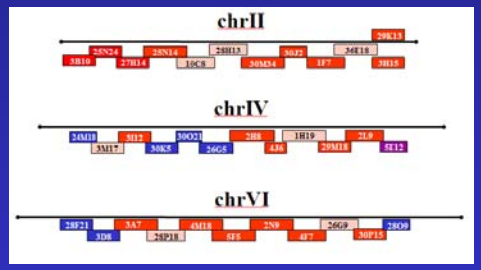
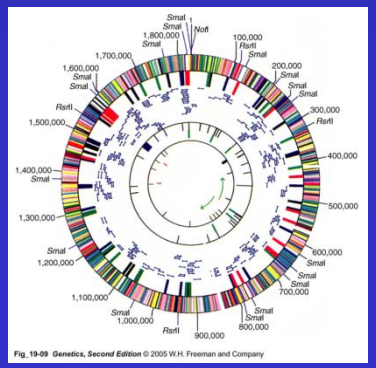
Anatomy of a genome project

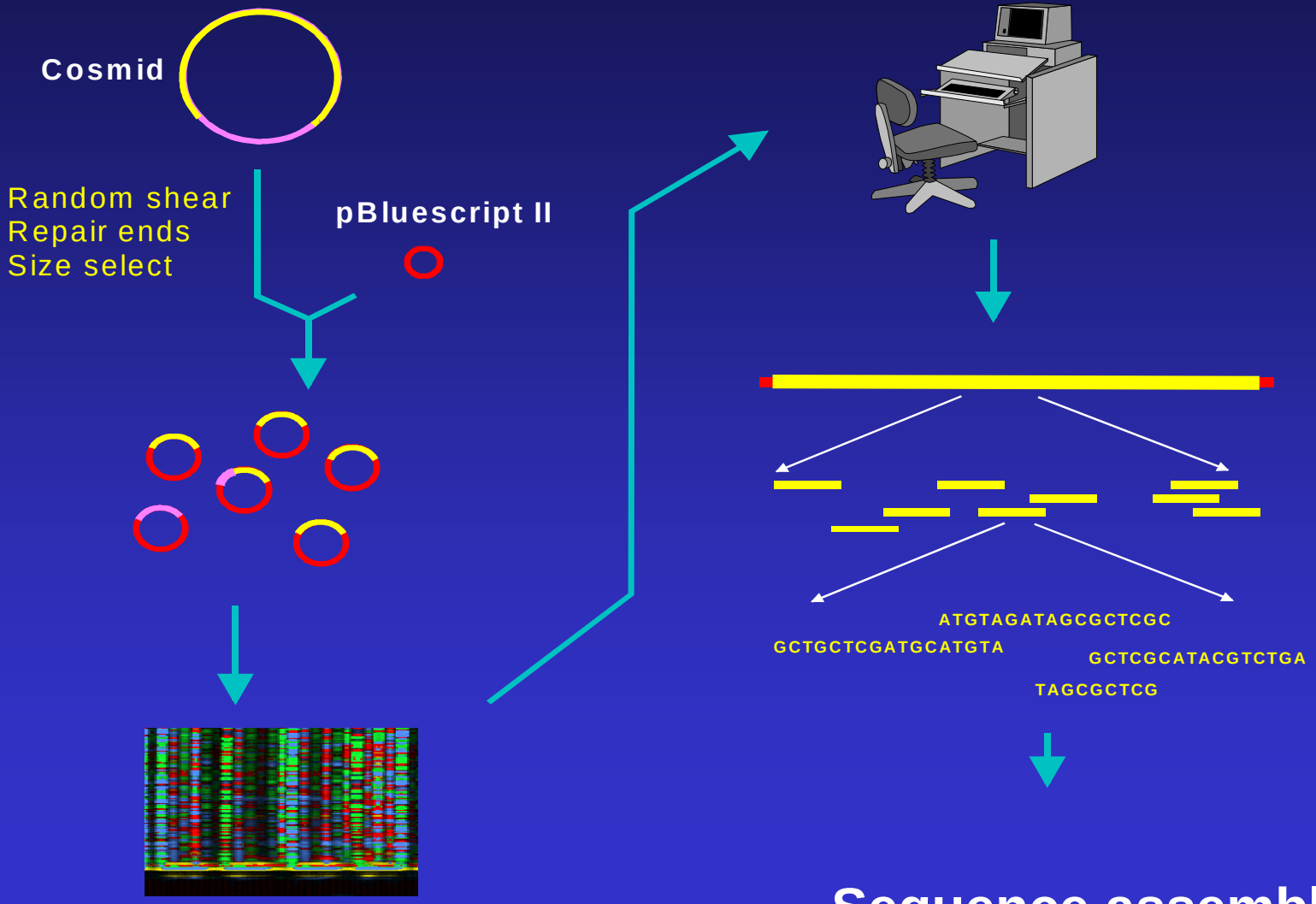






- **Map first, then sequence**
  - **Large insert clones (Cosmids, BACs, PACs, YACs)**
  - **Restriction mapping (fingerprinting)**
  - **STS mapping (PCR or hybridization)**
  - **Minimal tile path of clones > shotgun sequence**
  - **Slow, but accurate; scalable and divisible**
- **Map-as-you-go**
  - **Large insert clones**
  - **End-sequence**
  - **Shotgun sequence seed clones, then select clones with least overlap**
  - **More rapid, gene discovery early; scalable and divisible**
- **Whole genome/chromosome shotgun**
  - **Large and small insert clones**
  - **End-sequence then assemble**
  - **Computationally intensive**
  - **Not scalable or divisible**

Shotgun sequencing



Sequence assembly

Assembly

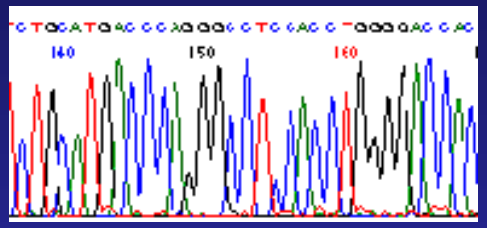
Finishing

Cosmid (40 kb)
500 - 1000 reactions
5 minutes

BAC (100 kb)
1000 - 2000 reactions
15 minutes

Chromosome (0.3-6 Mb)
10,000 - 200,000 reactions
0.5 – 10 hrs

Genome (25 – 60 Mb)
1 - 50 million reactions
Several days



Trace Data

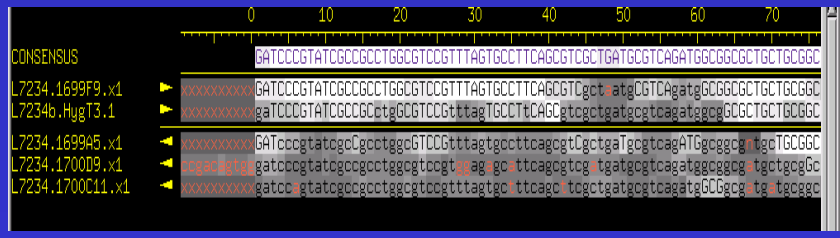
ACAAGTTCTAAC
TTCTAACAGGCGCATATCGG
CTCACAAGT

Individual Sequences

Contigs

Finishing Reactions

Consensus Sequence



Next Generation Sequencing

- **Massively parallel** (millions of reads *c.f.* 96)
- **No cloning**
- **Several different technologies**

	Platform		
	Roche(454)	**Illumina**	**SOLiD**
Sequencing chemistry	Pyrosequencing	Polymerase-based sequencing-by-synthesis	Ligation-based sequencing
Amplification approach	Emulsion PCR	Bridge amplification	Emulsion PCR
Paired ends/separation	Yes/3 kb	yes/200 bp	Yes/3 kb
Mb/run	100 Mb	1300 Mb	3000 Mb
Time/run (paired ends)	7 h	4 days	5 days
Read length	250 bp	32–40 bp	35 bp
Cost per run (total direct[a])	$8439	$8950	$17 447
Cost per Mb	$84.39	$5.97	$5.81

[a]Total direct costs include the reagents and consumables, the labor, instrument amortization cost and the disc storage space required for data storage/access.

- **Short reads**
- **Data handling problems**
- **“Draft” genomes, RNA-seq, ChIP-seq**

Gene prediction/annotation

- **What is a gene?**
- **Where are they?**
  - Gene prediction/structural annotation
- **What do they do?**
  - Functional annotation

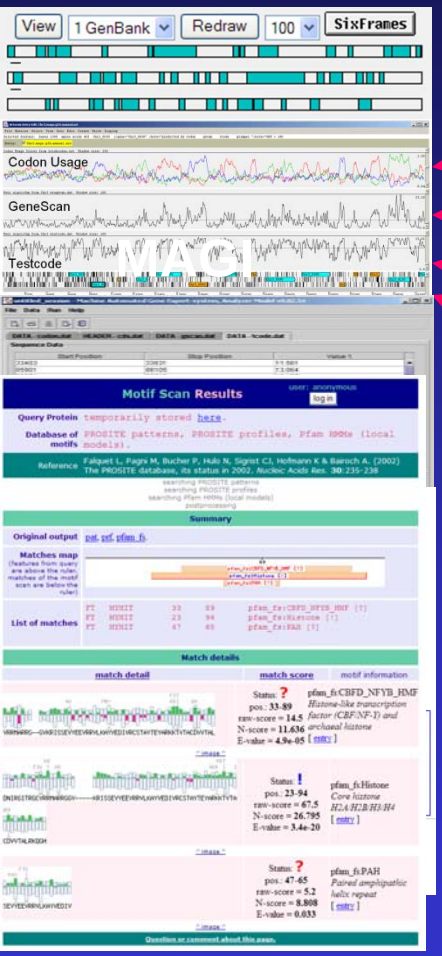
What is a gene ?

- **Genetic definition**
  - Region of genome that contains all the information required for synthesis of a product (protein or RNA)
  - Includes both coding and non-coding sequences.
- **Working definition**
  - Region of genome that encodes a potential protein
  - CDS (coding sequence)

Gene prediction methods

- **Open Reading Frame (ORF) analysis**
  - Simple, but limited
- **Extrinsic methods**
  - Sequence similarity to known genes
  - Misses novel genes
- **Intrinsic methods**
  - Rely on sequence content differences between coding and non-coding DNA
- **Consensus methods**
  - Combine multiple methods

Gene prediction/annotation



- **Find possible protein-coding genes (Open Reading Frames)**
- **Use statistical methods to determine likelihood**
  - Codon bias (**Codon Usage**)
  - Nucleotide bias (**GeneScan**)
  - Period3 constraint (**Testcode**)
  - Hidden Markov modeling (**Glimmer**)
- **Combine predictions (MAGI)**
- **Search sequence databases**
  - Sequence similarity – **Blast, COGs**
  - Patterns – **Prosite, Prints, Prodom**
  - Profiles – **Blocks, Pfam, Prosite**
  - Domains – **CDsearch**
- **Gene Ontology (GO) annotation**
- **Find RNA genes (Blast, tRNAScan)**

- **Automated *vs.* manual = speed *vs.* accuracy**