The Human Genome Project

Biol4230 Tues, March 20, 2018
Bill Pearson wrp@virginia.edu 4-2818 Pinn 6-057

- A brief history of DNA and genomes
- The Human Genome Project
- The Draft Human Genome (2001)
  - history and strategy
  - quality metrics
  - human biology
  - viewing genomes
  - computing on genomes
- Next Generation Genomes

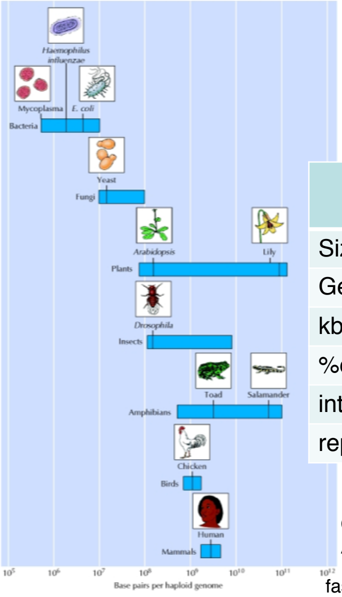
To learn more:

1. Pevsner, Chapter 19 pp. 791 – Human Genome
2. Pevsner, Chapter 18 pp. 729 – Eukaryotic Genomes
3. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (2001).
4. Venter, J. C. *et al.* The sequence of the human genome. *Science* **291,** 1304–1351 (2001).

The human genome sequence

- Assembled from pieces
  - PFP clone by clone, Celera Whole Genome Shotgun
  - Some regions hard to clone, some regions (repeats) hard to assemble
  - not complete, not perfect
- Determined from multiple individuals
  - an initial set of SNPs (single nucleotide polymorphisms) that can track variation
- Gene prediction (ab initio) is useless
  - virtually all gene predictions based on earlier evidence
  - no new gene types
  - many new genes (additional paralogs, duplications)

What is in a genome?

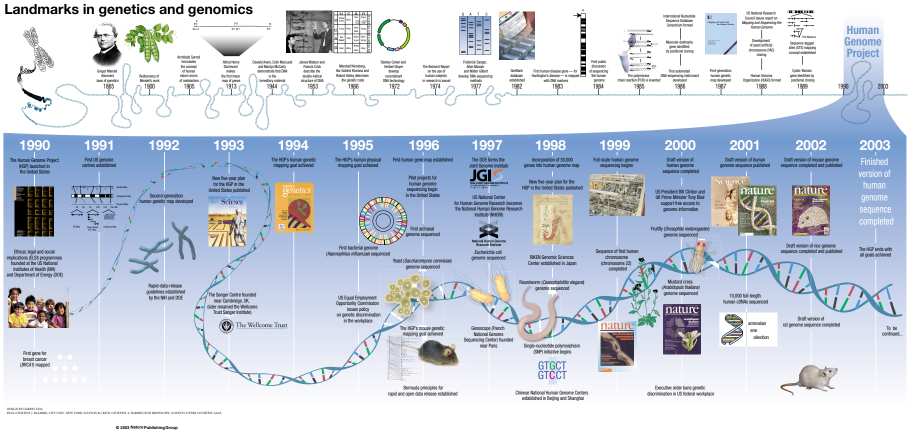


	E. col	Plas.	Yeast	Plant (ARATH)	Homo
Size(Mb)	4.64	22.8	12.5	115	3289
Genes	4288	5268	5770	25.5K	~25K
kb/Gene	0.95	4.34	2.09	4.53	27
%coding	87.8	52.6	70.5	28.8	1.3
introns	0	7406	272	107K	53K
repeat%	<1	<1	2.4	15	46

Pevsner, Table 16-1

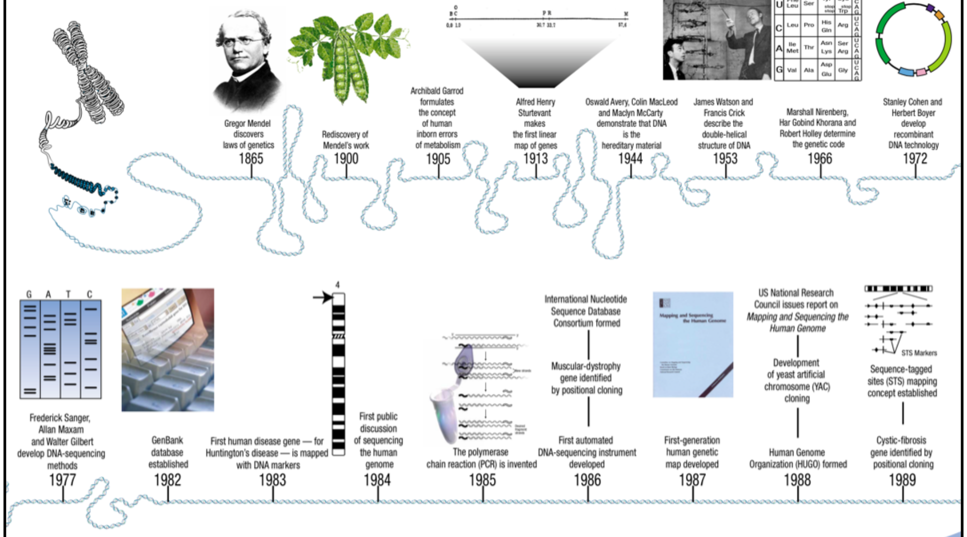
Cooper, GM (2000) The Cell: A Molecular Approach. 2nd edition. Fig 4.1

A history of genomes


Landmarks in genetics and genomics
Gregor Mendel discovers laws of genetics
1865
Rediscovery of Mendel's work
1900
Archibald Garrod formulates the concept of human inborn errors of metabolism
1905
Alfred Henry Sturtevant makes the first linear map of genes
1913
Oswald Avery, Colin MacLeod and Maclyn McCarty demonstrate that DNA is the hereditary material
1944
James Watson and Francis Crick describe the double-helical structure of DNA
1953
Marshall Nirenberg, Har Gobind Khorana and Robert Holley determine the genetic code
1966
Stanley Cohen and Herbert Boyer develop recombinant DNA technology
1972
The Belmont Report on the use of human subjects in research is issued
1974
Frederick Sanger, Allan Maxam and Walter Gilbert develop DNA-sequencing methods
1977
GenBank database established
1982
First human disease gene — for Huntington's disease — is mapped with DNA markers
1983
First public discussion of sequencing the human genome
1984
The polymerase chain reaction (PCR) is invented
1985
International Nucleotide Sequence Database Consortium formed
Muscular-dystrophy gene identified by positional cloning
First automated DNA-sequencing instrument developed
1986
First-generation human genetic map developed
1987
US National Research Council issues report on Mapping and Sequencing the Human Genome
Development of yeast artificial chromosome (YAC) cloning
Human Genome Organization (HUGO) formed
1988
STS Markers
Sequence-tagged sites (STS) mapping concept established
Cystic-fibrosis gene identified by positional cloning
1989
Human Genome Project
1990
2003
1990
The Human Genome Project (HGP) launched in the United States
Ethical, legal and social implications (ELSI) programmes founded at the US National Institutes of Health (NIH) and Department of Energy (DOE)
First gene for breast cancer (BRCA1) mapped
1991
First US genome centres established
Genetic Map
Cytogenetic Map
Physical Map
RH Map
Clone-Based STS Map
Sequence Map
1992
Second-generation human genetic map developed
Rapid-data-release guidelines established by the NIH and DOE
1993
New five-year plan for the HGP in the United States published
The Sanger Centre founded near Cambridge, UK, (later renamed the Wellcome Trust Sanger Institute)
The Wellcome Trust
1994
The HGP's human genetic mapping goal achieved
1995
The HGP's human physical mapping goal achieved
First bacterial genome (Haemophilus influenzae) sequenced
US Equal Employment Opportunity Commission issues policy on genetic discrimination in the workplace
1996
First human gene map established
Pilot projects for human genome sequencing begin in the United States
First archaeal genome sequenced
Yeast (Saccharomyces cerevisiae) genome sequenced
The HGP's mouse genetic mapping goal achieved
Bermuda principles for rapid and open data release established
1997
The DOE forms the Joint Genome Institute
JGI DOE JOINT GENOME INSTITUTE
US National Center for Human Genome Research becomes the National Human Genome Research Institute (NHGRI)
National Human Genome Research Institute
Escherichia coli genome sequenced
Genoscope (French National Genome Sequencing Center) founded near Paris
1998
Incorporation of 30,000 genes into human genome map
New five-year plan for the HGP in the United States published
RIKEN Genomic Sciences Center established in Japan
Roundworm (Caenorhabditis elegans) genome sequenced
Single-nucleotide polymorphism (SNP) initiative begins
GTGCT
GTCCT
Chinese National Human Genome Centers established in Beijing and Shanghai
1999
Full-scale human genome sequencing begins
Sequence of first human chromosome (chromosome 22) completed
2000
Draft version of human genome sequence completed
US President Bill Clinton and UK Prime Minister Tony Blair support free access to genome information
Fruitfly (Drosophila melanogaster) genome sequenced
Mustard cress (Arabidopsis thaliana) genome sequenced
Executive order bans genetic discrimination in US federal workplace
2001
Draft version of human genome sequence published
10,000 full-length human cDNAs sequenced
ammalian ene ollection
2002
Draft version of mouse genome sequence completed and published
Draft version of rice genome sequence completed and published
Draft version of rat genome sequence completed
2003
Finished version of human genome sequence completed
The HGP ends with all goals achieved
To be continued...
DESIGN BY DARRYL LEJA
PEAS COURTESY J. BLAMIRE, CITY UNIV. NEW YORK; WATSON & CRICK COURTESY A. BARRINGTON BROWN/SPL; SCIENCE COVERS COURTESY AAAS
© 2003 Nature Publishing Group

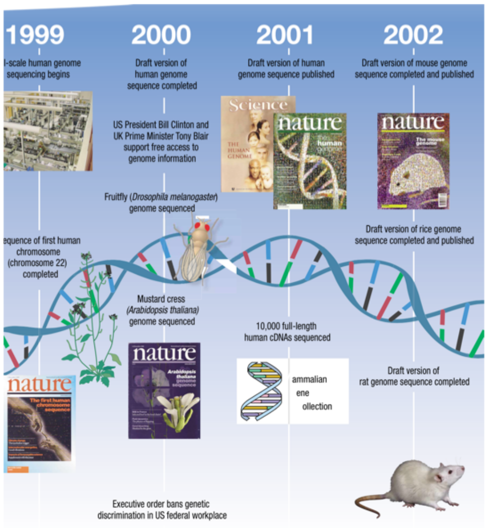
Collins, Nature (2003) 422:835

Landmarks in genetics and genomics


Gregor Mendel discovers laws of genetics
1865
Rediscovery of Mendel's work
1900
Archibald Garrod formulates the concept of human inborn errors of metabolism
1905
Alfred Henry Sturtevant makes the first linear map of genes
1913
Oswald Avery, Colin MacLeod and Maclyn McCarty demonstrate that DNA is the hereditary material
1944
James Watson and Francis Crick describe the double-helical structure of DNA
1953
Marshall Nirenberg, Har Gobind Khorana and Robert Holley determine the genetic code
1966
Stanley Cohen and Herbert Boyer develop recombinant DNA technology
1972
G A T C
Frederick Sanger, Allan Maxam and Walter Gilbert develop DNA-sequencing methods
1977
GenBank database established
1982
First human disease gene — for Huntington's disease — is mapped with DNA markers
1983
4
First public discussion of sequencing the human genome
1984
The polymerase chain reaction (PCR) is invented
1985
International Nucleotide Sequence Database Consortium formed
Muscular-dystrophy gene identified by positional cloning
First automated DNA-sequencing instrument developed
1986
Mapping and Sequencing the Human Genome
First-generation human genetic map developed
1987
US National Research Council issues report on Mapping and Sequencing the Human Genome
Development of yeast artificial chromosome (YAC) cloning
Human Genome Organization (HUGO) formed
1988
STS Markers
Sequence-tagged sites (STS) mapping concept established
Cystic-fibrosis gene identified by positional cloning
1989

Collins, Nature (2003) 422:835

The Human Genome Project

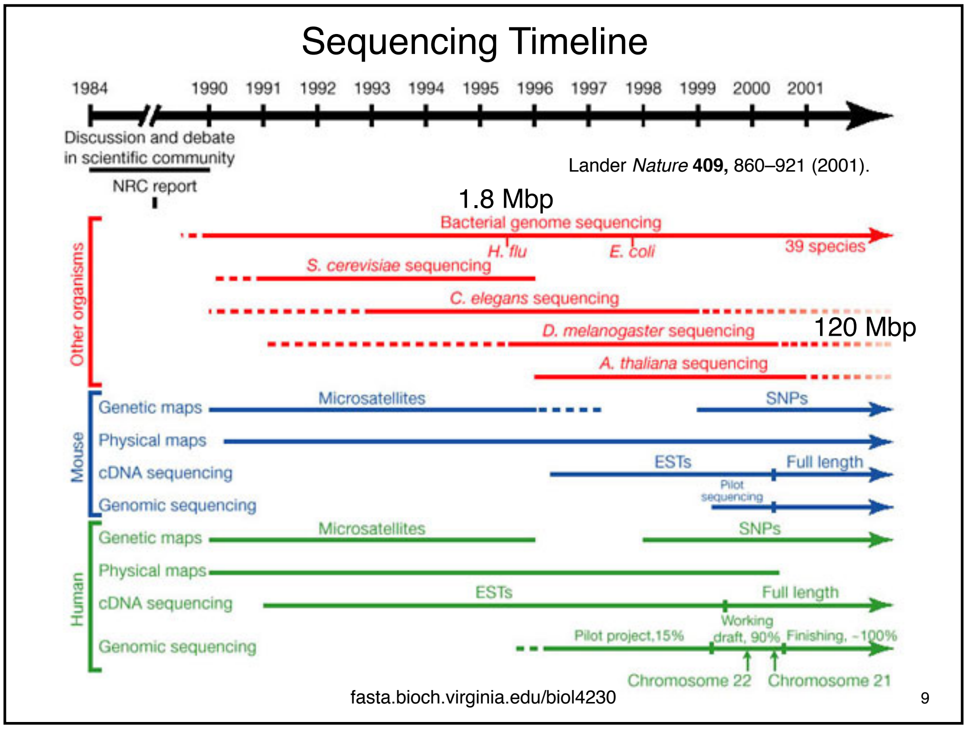

1999
-scale human genome sequencing begins
equence of first human chromosome (chromosome 22) completed
nature
The first human chromosome sequence
2000
Draft version of human genome sequence completed
US President Bill Clinton and UK Prime Minister Tony Blair support free access to genome information
Fruitfly (*Drosophila melanogaster*) genome sequenced
Mustard cress (*Arabidopsis thaliana*) genome sequenced
nature
Arabidopsis thaliana genome sequence
Executive order bans genetic discrimination in US federal workplace
2001
Draft version of human genome sequence published
Science
THE HUMAN GENOME
nature
the human genome
10,000 full-length human cDNAs sequenced
ammalian
ene
ollection
2002
Draft version of mouse genome sequence completed and published
nature
The mouse genome
Draft version of rice genome sequence completed and published
Draft version of rat genome sequence completed

Collins, Nature (2003) 422:835

Sequencing capacity (2004)
  40 million lanes/year ~ 8 billion bases – Whitehead Inst. (1 bacterial genome/day)
>40 billion bases/year, world-wide
  (1000 bacterial genomes/year; 1 mammalian genome/year)

Sequencing capacity (2011) – Illumina sequencing 200 billion bases/week/machine, ~30,000 human genomes/year
Sequencing capacity (2015) – at least 300,000 human genomes/year

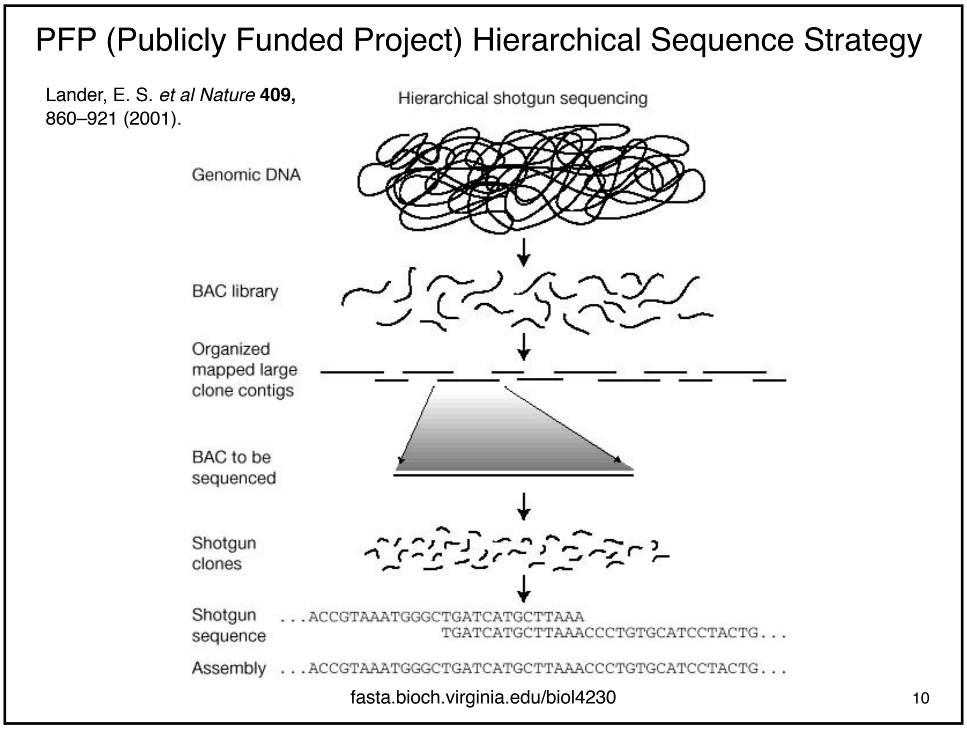
Sequencing Timeline



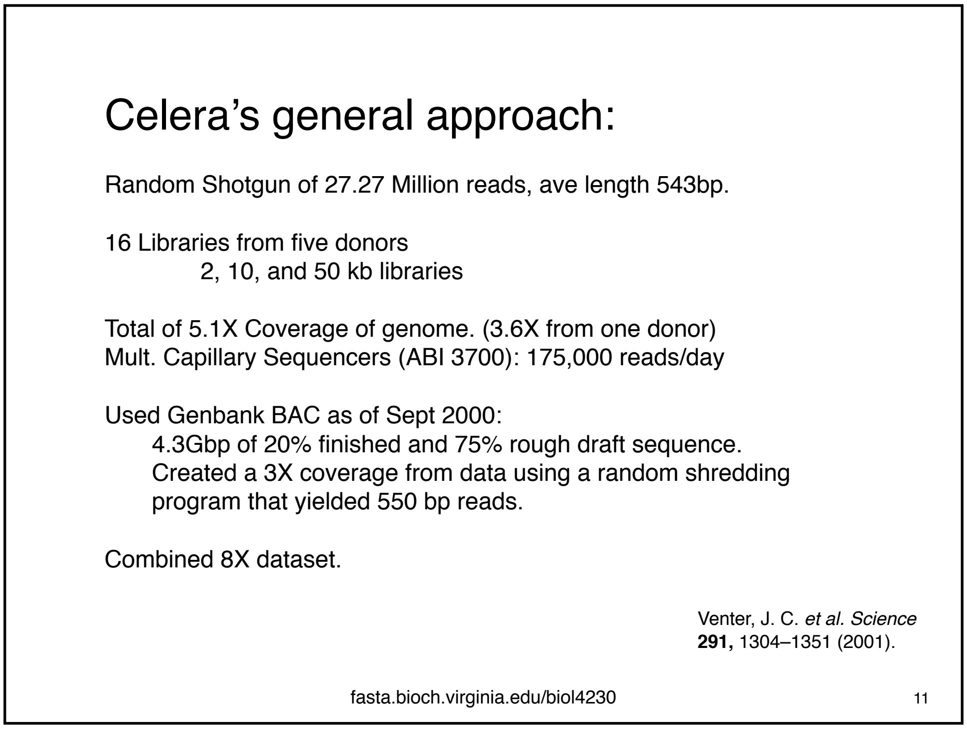
Lander *Nature* **409,** 860–921 (2001).

PFP (Publicly Funded Project) Hierarchical Sequence Strategy

Lander, E. S. *et al Nature* **409,** 860–921 (2001).



```
Shotgun    ...ACCGTAAATGGGCTGATCATGCTTAAA
sequence                        TGATCATGCTTAAACCCTGTGCATCCTACTG...
Assembly   ...ACCGTAAATGGGCTGATCATGCTTAAACCCTGTGCATCCTACTG...
```



Celera’s general approach:

Random Shotgun of 27.27 Million reads, ave length 543bp.

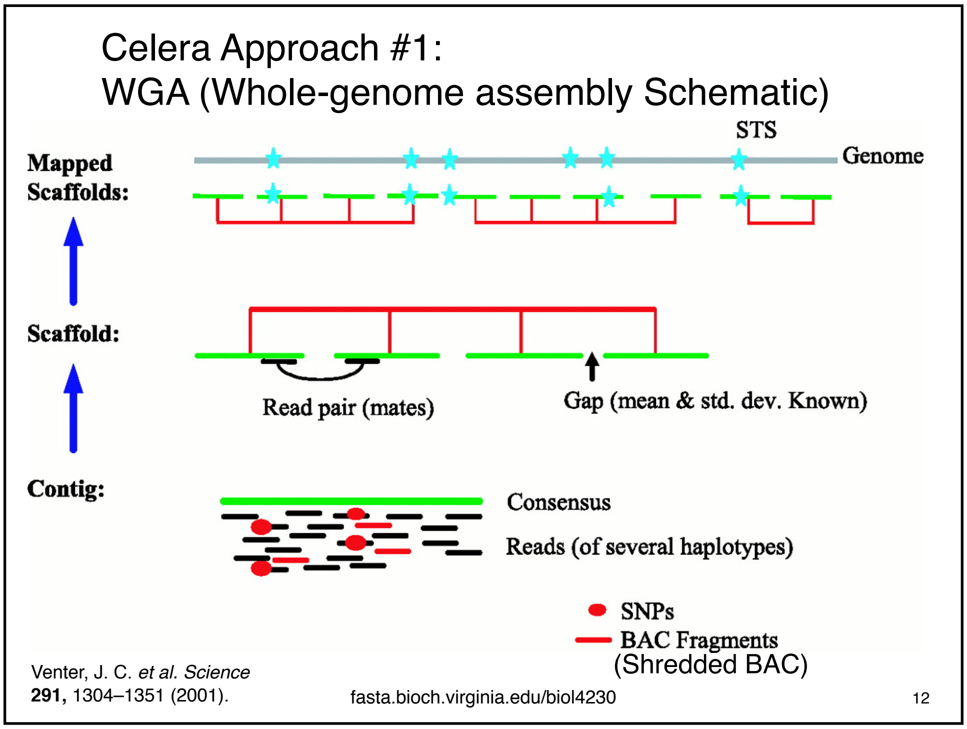
16 Libraries from five donors
2, 10, and 50 kb libraries

Total of 5.1X Coverage of genome. (3.6X from one donor)
Mult. Capillary Sequencers (ABI 3700): 175,000 reads/day

Used Genbank BAC as of Sept 2000:
4.3Gbp of 20% finished and 75% rough draft sequence.
Created a 3X coverage from data using a random shredding program that yielded 550 bp reads.

Combined 8X dataset.

Venter, J. C. *et al. Science*
291, 1304–1351 (2001).

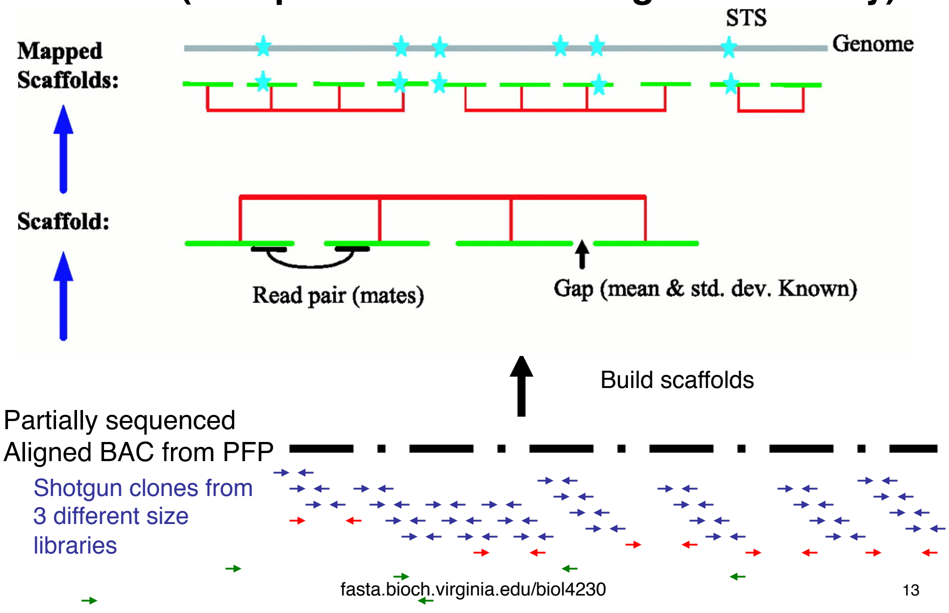


Celera Approach #1:
WGA (Whole-genome assembly Schematic)

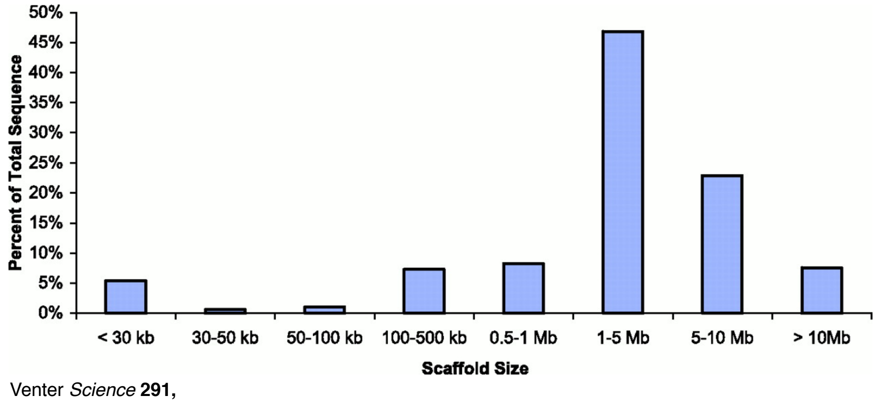
(Shredded BAC)

Venter, J. C. *et al. Science*
291, 1304–1351 (2001).

Celera Approach #2: CSA (Compartmentalized Shotgun Assembly)


STS
Genome
Mapped Scaffolds:
Scaffold:
Read pair (mates)
Gap (mean & std. dev. Known)
Build scaffolds
Partially sequenced Aligned BAC from PFP
Shotgun clones from 3 different size libraries

Distribution of Scaffold sizes (Not the same as sequence Contigs)



Venter *Science* **291,** 1304–1351 (2001).

CSA versus PFP for Chr 8 over 1Mbp

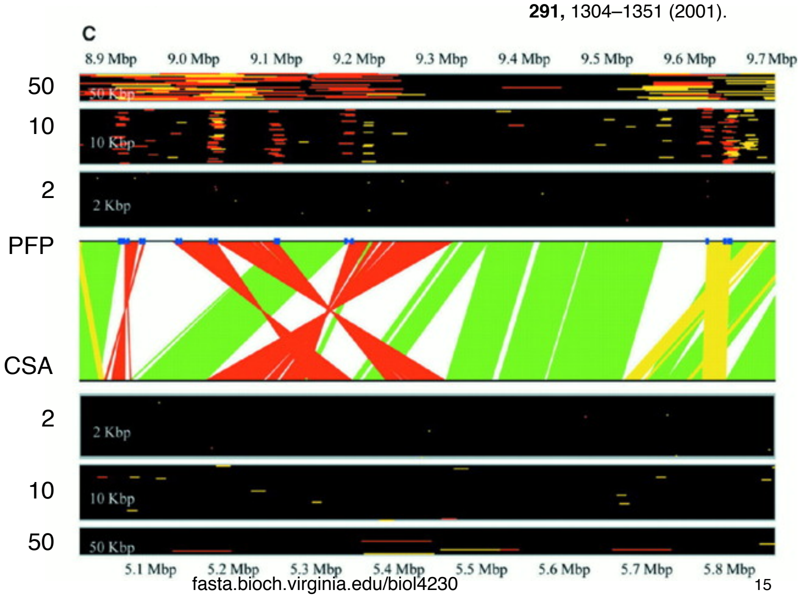
🟨 Blocks in same orient., but out of order.

Ind. Seq. have incorrect distance between ends

🟥 Blocks in opposite orient.

Ind. Seq. Are mis-oriented.

Venter, J. C. *et al. Science* **291,** 1304–1351 (2001).



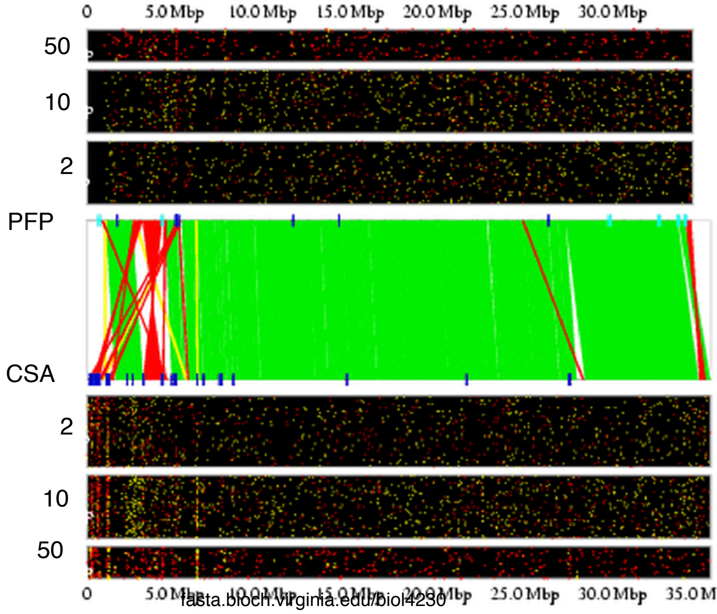
CSA versus PFP for Chr 22 (Finished)

🟨 Blocks in same orient., but out of order.

Ind. Seq. have incorrect distance between ends

🟥 Blocks in opposite orient.

Ind. Seq. Are mis-oriented.



CSA versus PFP for Chr 8

🟨

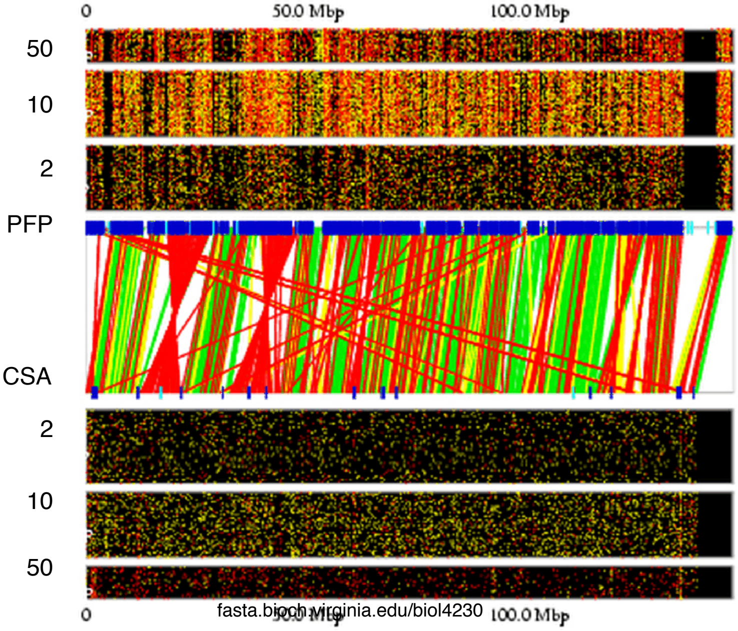
Blocks in same orient., but out of order.

Ind. Seq. have incorrect distance between ends

🟥

Blocks in opposite orient.

Ind. Seq. Are mis-oriented.



Determining Gene Number in Genome is Hard

Developed a homology/evidence based system called **Otto**.

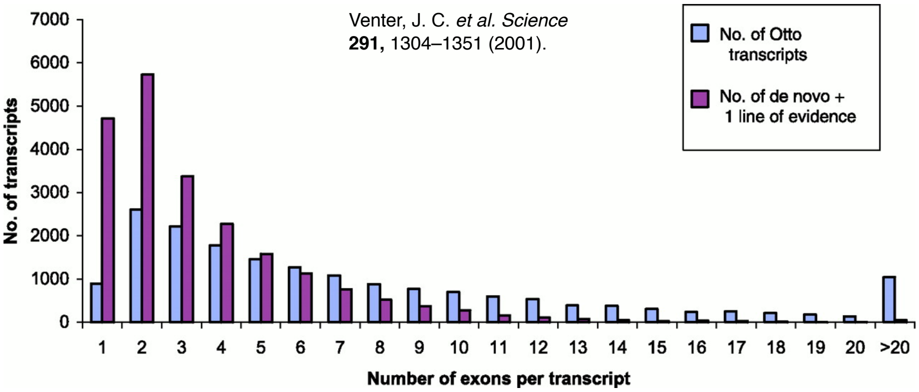
Otto searches scaffold sequences for homology against known protein, (RefSeq,) EST, and runs 3 de novo gene prediction programs to see if areas of homology are consistent with a gene transcript.

De novo sequences include all gene-prediction transcripts from GRAIL, Genscan, and FgenesH sorted on the basis of matches to EST, protein, or other mouse rat libraries.

	Predicted Genes [26,588 - ~39k]		
evidence	>=1	>=2	>=3
Otto	17, 968	17,501	15,877
De Novo	21,350	8,619	4,947

Venter *Science* **291,** 1304–1351 (2001).

Distribution of transcripts with varying exon number

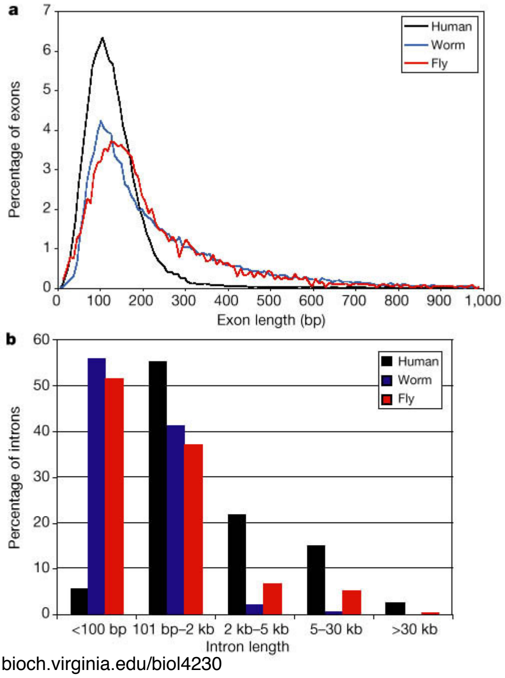


Venter, J. C. *et al. Science* **291,** 1304–1351 (2001).

Fig. 9. Comparison of the number of exons per transcript between the 17,968 Otto transcripts and 21,350 de novo transcript predictions with at least one line of evidence that do not overlap with an Otto prediction. Both sets have the highest number of transcripts in the two-exon category, but the de novo gene predictions are skewed much more toward smaller transcripts. In the Otto set, 19.7% of the transcripts have one or two exons, and 5.7% have more than 20. In the de novo set, 49.3% of the transcripts have one or two exons, and 0.2% have more than 20.

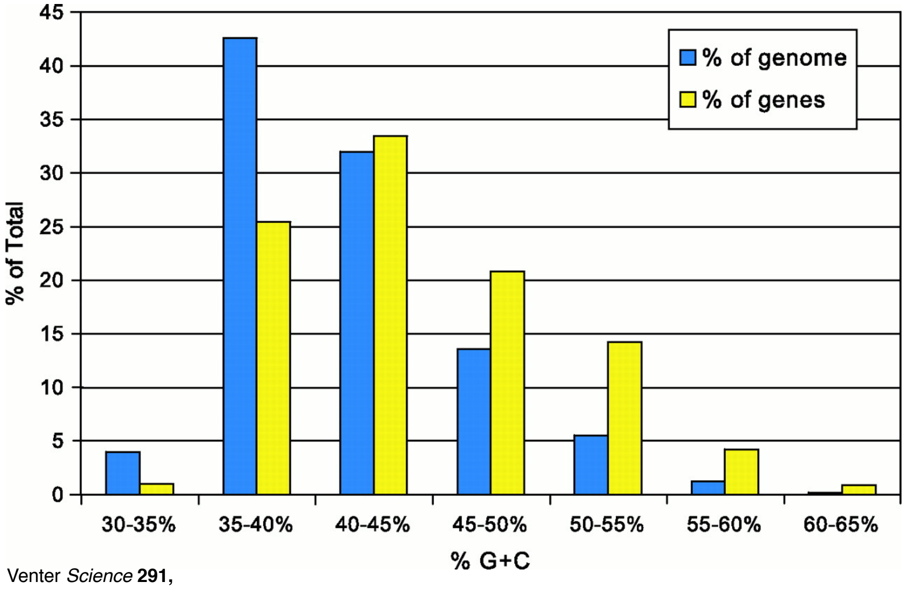
Exon Length

Intron Length



Lander *Nature* **409,** 860–921 (2001), Fig. 35

G+C content and gene density



Venter *Science* **291,**
1304–1351 (2001).

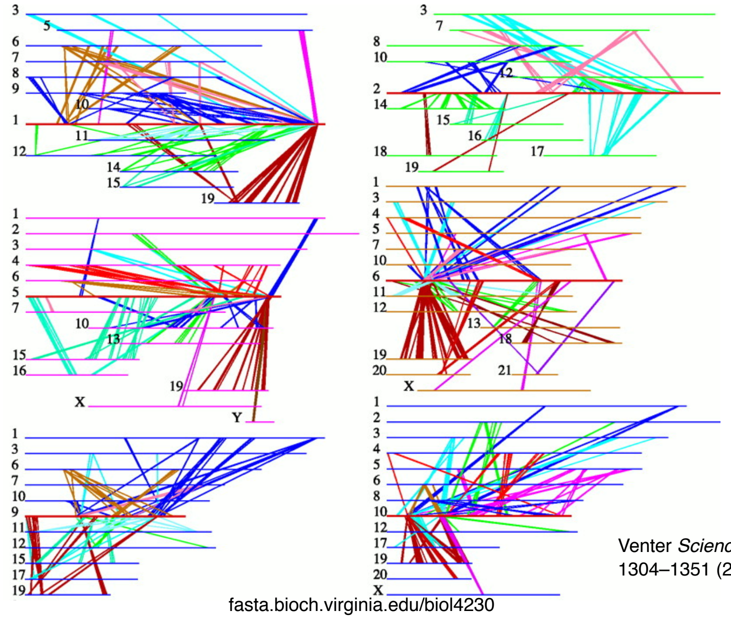
Segmental duplication versus Retrotransposition

Retrotransposition of mRNA's into the genome converts a gene with introns into an intronless gene flanked by direct repeats and containing a poly A at the 3' end.

Most of these genes will be reverse transcribed badly
==>become inactivated genes (pseudogenes.)
901 found, 97 appear to be functional

Segmental duplication is a duplication on the DNA level within or between chromosomes.

Representatives of 1077 Segmental Genomic duplications



Venter *Science* **291,** 1304–1351 (2001).

Detailed view of Duplication events on Chr18 and and 20

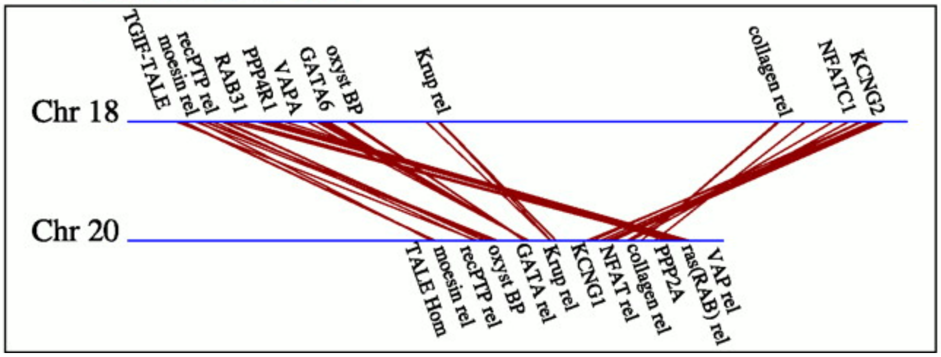


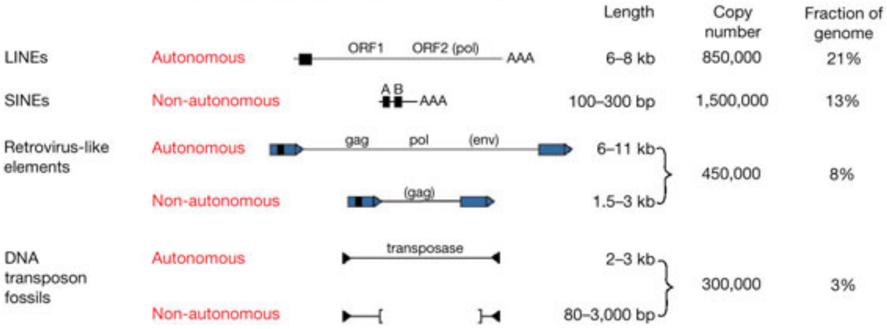
FIg. 13. Segmental duplications between chromosomes in the human genome. The 24 panels show the 1077 duplicated blocks of genes, containing 10,310 pairs of genes in total. Each line represents a pair of homologous genes belonging to a block; all blocks contain at least three genes on each of the chromosomes where they appear. Each panel shows all the duplications between a single chromosome and other chromosomes with shared blocks. The chro- mosome at the center of each panel is shown as a thick red line for emphasis. Other chromosomes are displayed from top to bot- tom within each panel ordered by chromosome number. The inset (bottom, center right) shows a close-up of one duplication between chromosomes 18 and 20, expanded to display the gene names of 12 of the 64 gene pairs shown.

Venter *Science* **291,** 1304–1351 (2001).

Comparison of Repeats from Celera and PFP

Alu	9.9%
Mammalian interspersed repeat (MIR)	2.3%
Medium reiteration (MER)	1.7%
Long Terminal Repeat (LTR)	5.3%
Long interspersed nucleotide element (LINE)	16.1%
Total	35%

Classes of interspersed repeat in the human genome



		Length	Copy number	Fraction of genome
LINEs	Autonomous	6–8 kb	850,000	21%
SINEs	Non-autonomous	100–300 bp	1,500,000	13%
Retrovirus-like elements	Autonomous	6–11 kb	450,000	8%
	Non-autonomous	1.5–3 kb		
DNA transposon fossils	Autonomous	2–3 kb	300,000	3%
	Non-autonomous	80–3,000 bp		

Lander *Nature* **409,** 860–921 (2001), Fig. 35

SNP (Single Nucleotide Polymorphism) frequency varies in the genome more than predicted.

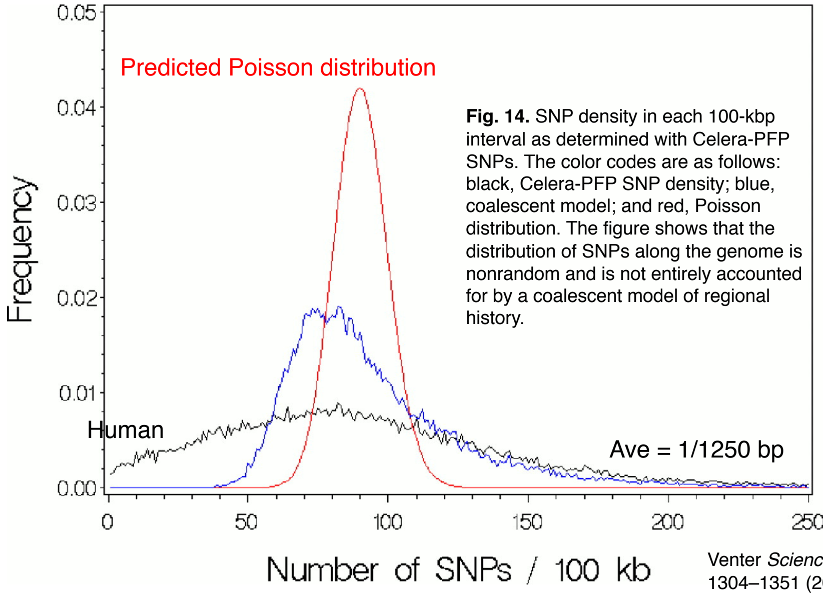
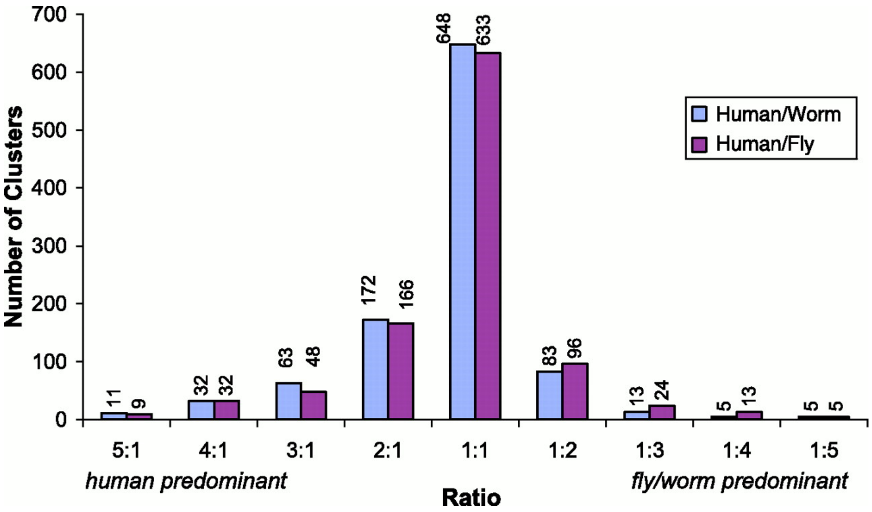


Fig. 14. SNP density in each 100-kbp interval as determined with Celera-PFP SNPs. The color codes are as follows: black, Celera-PFP SNP density; blue, coalescent model; and red, Poisson distribution. The figure shows that the distribution of SNPs along the genome is nonrandom and is not entirely accounted for by a coalescent model of regional history.

Venter *Science* **291,** 1304–1351 (2001).

Gene duplication in complete protein clusters (Lek)



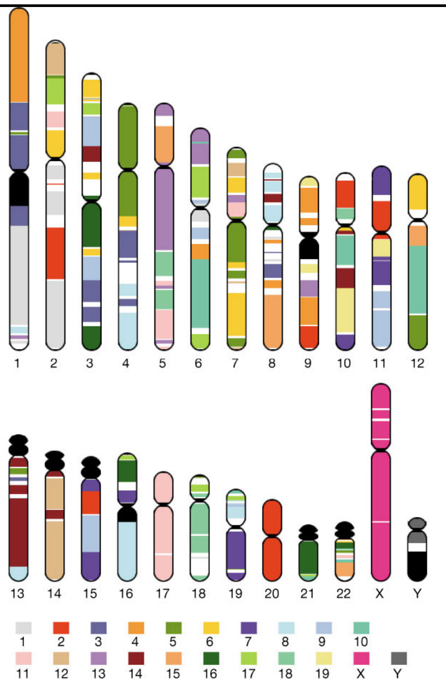
Venter *Science* **291,** 1304–1351 (2001).

fasta.bioch.virginia.edu/biol4230

Human Chromosomes colored by mouse chromosomes

Figure 46 Conserved segments in the human and mouse genome. Human chromosomes, with segments containing at least two genes whose order is conserved in the mouse genome as colour blocks. Each colour corresponds to a particular mouse chromosome. Centromeres, subcentromeric heterochromatin of chromosomes 1, 9 and 16, and the repetitive short arms of 13, 14, 15, 21 and 22 are in black.

Lander *Nature* **409,** 860–921 (2001), Fig. 35



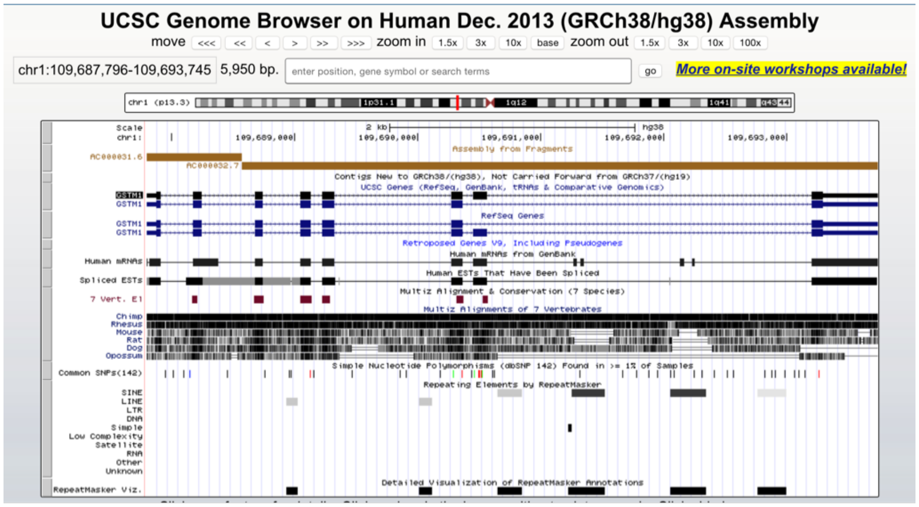
Human Chr.

Human Chromosomes

Mouse Chromosome

fasta.bioch.virginia.edu/biol4230

Retrieving the data: Genome Browsers (UCSC)



UCSC Genome Browser on Human Dec. 2013 (GRCh38/hg38) Assembly

move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x 100x

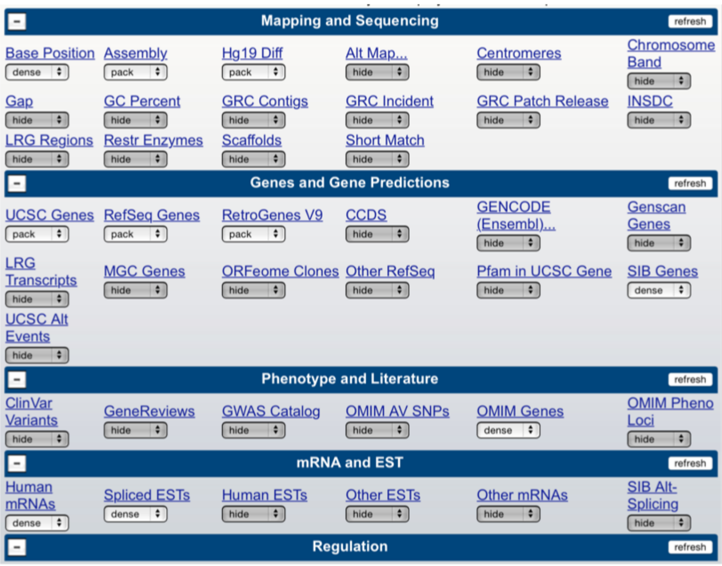
chr1:109,687,796-109,693,745 5,950 bp. enter position, gene symbol or search terms go **More on-site workshops available!**

chr1 (p13.3)
1p31.1
1q12
1q41
q43
44
Scale
chr1:
2 kb
hg38
109,689,000
109,690,000
109,691,000
109,692,000
109,693,000
Assembly from Fragments
AC000031.6
AC000032.7
Contigs New to GRCh38/(hg38), Not Carried Forward from GRCh37/(hg19)
UCSC Genes (RefSeq, GenBank, tRNAs & Comparative Genomics)
GSTM1
GSTM1
RefSeq Genes
GSTM1
GSTM1
Retroposed Genes V9, Including Pseudogenes
Human mRNAs from GenBank
Human mRNAs
Human ESTs That Have Been Spliced
Spliced ESTs
Multiz Alignment & Conservation (7 Species)
7 Vert. El
Multiz Alignments of 7 Vertebrates
Chimp
Rhesus
Mouse
Rat
Dog
Opossum
Simple Nucleotide Polymorphisms (dbSNP 142) Found in >= 1% of Samples
Common SNPs(142)
Repeating Elements by RepeatMasker
SINE
LINE
LTR
DNA
Simple
Low Complexity
Satellite
RNA
Other
Unknown
Detailed Visualization of RepeatMasker Annotations
RepeatMasker Viz.

fasta.bioch.virginia.edu/biol4230

29

Genome Browsers (UCSC)



Mapping and Sequencing refresh

Base Position dense	Assembly pack	Hg19 Diff pack	Alt Map... hide	Centromeres hide	Chromosome Band hide
Gap hide	GC Percent hide	GRC Contigs hide	GRC Incident hide	GRC Patch Release hide	INSDC hide
LRG Regions hide	Restr Enzymes hide	Scaffolds hide	Short Match hide		

Genes and Gene Predictions refresh

UCSC Genes pack	RefSeq Genes pack	RetroGenes V9 pack	CCDS hide	GENCODE (Ensembl)... hide	Genscan Genes hide
LRG Transcripts hide	MGC Genes hide	ORFeome Clones hide	Other RefSeq hide	Pfam in UCSC Gene hide	SIB Genes dense
UCSC Alt Events hide					

Phenotype and Literature refresh

ClinVar Variants hide	GeneReviews hide	GWAS Catalog hide	OMIM AV SNPs hide	OMIM Genes dense	OMIM Pheno Loci hide

mRNA and EST refresh

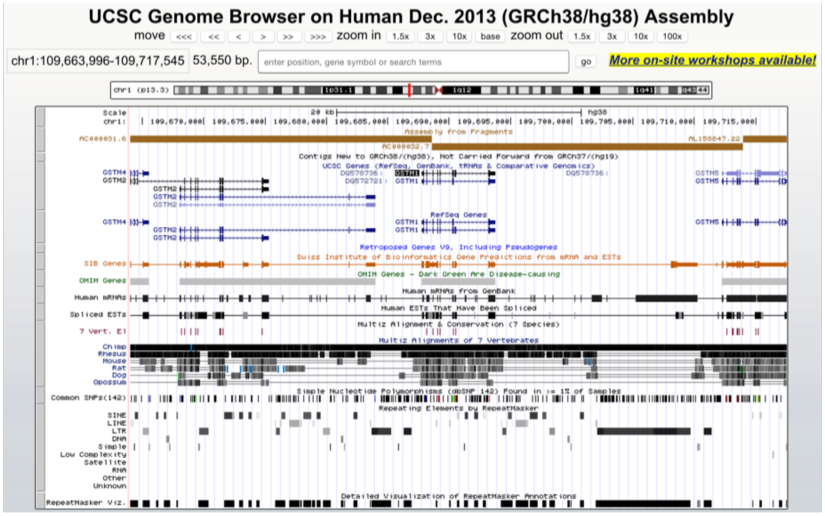
Human mRNAs dense	Spliced ESTs dense	Human ESTs hide	Other ESTs hide	Other mRNAs hide	SIB Alt-Splicing hide

Regulation refresh

fasta.bioch.virginia.edu/biol4230

30

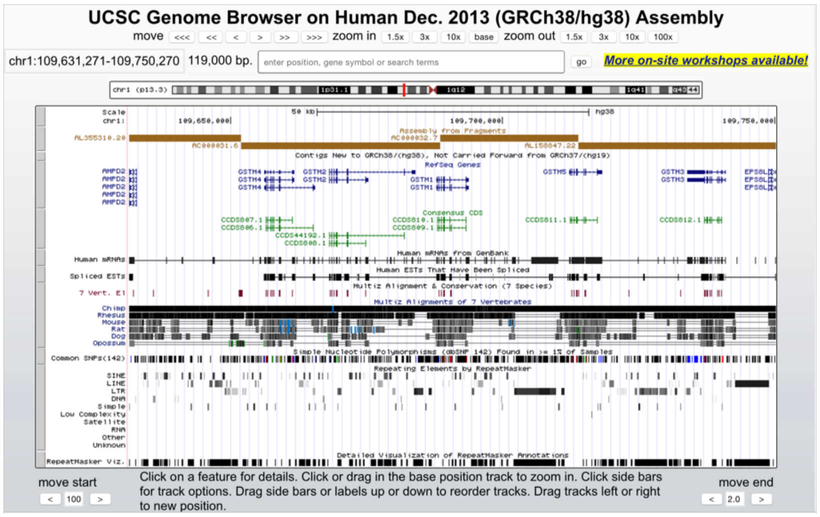
Genome Browsers (UCSC)



UCSC Genome Browser on Human Dec. 2013 (GRCh38/hg38) Assembly

move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x 100x
chr1:109,663,996-109,717,545 53,550 bp.
enter position, gene symbol or search terms
go
More on-site workshops available!
chr1 (p13.3)
Scale
chr1:
20 kb
hg38
109,670,000
109,675,000
109,680,000
109,685,000
109,690,000
109,695,000
109,700,000
109,705,000
109,710,000
109,715,000
Assembly from Fragments
AC000031.6
AC000032.7
AL158847.22
Contigs New to GRCh38/(hg38), Not Carried Forward from GRCh37/(hg19)
UCSC Genes (RefSeq, GenBank, tRNAs & Comparative Genomics)
GSTM4
GSTM2
GSTM1
GSTM5
DQ570736
DQ572721
RefSeq Genes
Retroposed Genes V9, Including Pseudogenes
Swiss Institute of Bioinformatics Gene Predictions from mRNA and ESTs
SIB Genes
OMIM Genes - Dark Green Are Disease-causing
OMIM Genes
Human mRNAs from GenBank
Human mRNAs
Human ESTs That Have Been Spliced
Spliced ESTs
Multiz Alignment & Conservation (7 Species)
7 Vert. El
Multiz Alignments of 7 Vertebrates
Chimp
Rhesus
Mouse
Rat
Dog
Opossum
Simple Nucleotide Polymorphisms (dbSNP 142) Found in >= 1% of Samples
Common SNPs(142)
Repeating Elements by RepeatMasker
SINE
LINE
LTR
DNA
Simple
Low Complexity
Satellite
RNA
Other
Unknown
Detailed Visualization of RepeatMasker Annotations
RepeatMasker Viz.

Genome Browsers (UCSC)



UCSC Genome Browser on Human Dec. 2013 (GRCh38/hg38) Assembly

move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x 100x
chr1:109,631,271-109,750,270 119,000 bp.
enter position, gene symbol or search terms
go
More on-site workshops available!
chr1 (p13.3)
Scale
chr1:
50 kb
hg38
109,650,000
109,700,000
109,750,000
Assembly from Fragments
AL355310.20
AC000032.7
AC000031.6
AL158847.22
Contigs New to GRCh38/(hg38), Not Carried Forward from GRCh37/(hg19)
RefSeq Genes
AMPD2
GSTM4
GSTM2
GSTM1
GSTM5
GSTM3
EPS8L3
Consensus CDS
CCDS807.1
CCDS806.1
CCDS44192.1
CCDS808.1
CCDS810.1
CCDS809.1
CCDS811.1
CCDS812.1
Human mRNAs from GenBank
Human mRNAs
Human ESTs That Have Been Spliced
Spliced ESTs
Multiz Alignment & Conservation (7 Species)
7 Vert. El
Multiz Alignments of 7 Vertebrates
Chimp
Rhesus
Mouse
Rat
Dog
Opossum
Simple Nucleotide Polymorphisms (dbSNP 142) Found in >= 1% of Samples
Common SNPs(142)
Repeating Elements by RepeatMasker
SINE
LINE
LTR
DNA
Simple
Low Complexity
Satellite
RNA
Other
Unknown
Detailed Visualization of RepeatMasker Annotations
RepeatMasker Viz.
move start
< 100 >
move end
< 2.0 >

Click on a feature for details. Click or drag in the base position track to zoom in. Click side bars for track options. Drag side bars or labels up or down to reorder tracks. Drag tracks left or right to new position.

Genome data (UCSC table browser)

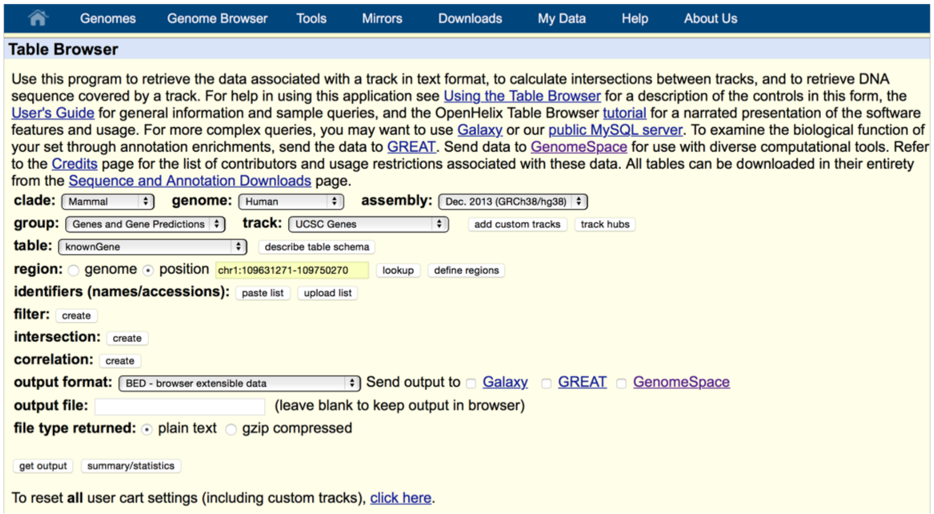


Table Browser

Use this program to retrieve the data associated with a track in text format, to calculate intersections between tracks, and to retrieve DNA sequence covered by a track. For help in using this application see Using the Table Browser for a description of the controls in this form, the User's Guide for general information and sample queries, and the OpenHelix Table Browser tutorial for a narrated presentation of the software features and usage. For more complex queries, you may want to use Galaxy or our public MySQL server. To examine the biological function of your set through annotation enrichments, send the data to GREAT. Send data to GenomeSpace for use with diverse computational tools. Refer to the Credits page for the list of contributors and usage restrictions associated with these data. All tables can be downloaded in their entirety from the Sequence and Annotation Downloads page.

clade: Mammal **genome:** Human **assembly:** Dec. 2013 (GRCh38/hg38)

group: Genes and Gene Predictions **track:** UCSC Genes add custom tracks track hubs

table: knownGene describe table schema

region: ○ genome ● position chr1:109631271-109750270 lookup define regions

identifiers (names/accessions): paste list upload list

filter: create

intersection: create

correlation: create

output format: BED - browser extensible data Send output to □ Galaxy □ GREAT □ GenomeSpace

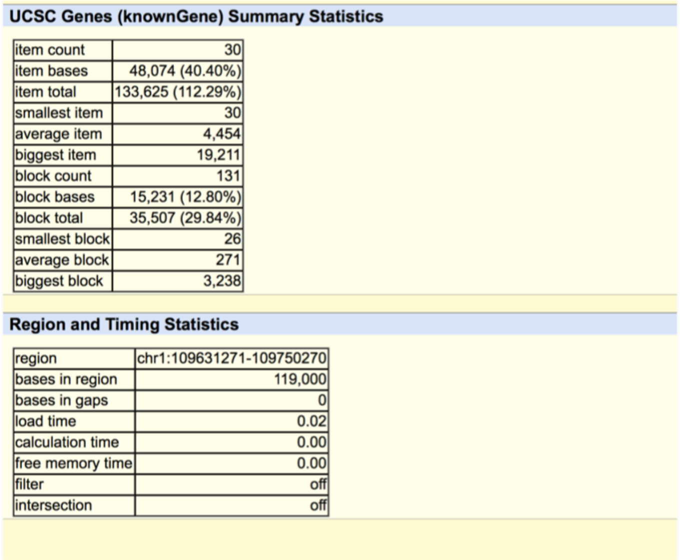
output file: (leave blank to keep output in browser)

file type returned: ● plain text ○ gzip compressed

get output summary/statistics

To reset **all** user cart settings (including custom tracks), click here.

Genome data (UCSC table browser)



UCSC Genes (knownGene) Summary Statistics

item count	30
item bases	48,074 (40.40%)
item total	133,625 (112.29%)
smallest item	30
average item	4,454
biggest item	19,211
block count	131
block bases	15,231 (12.80%)
block total	35,507 (29.84%)
smallest block	26
average block	271
biggest block	3,238

Region and Timing Statistics

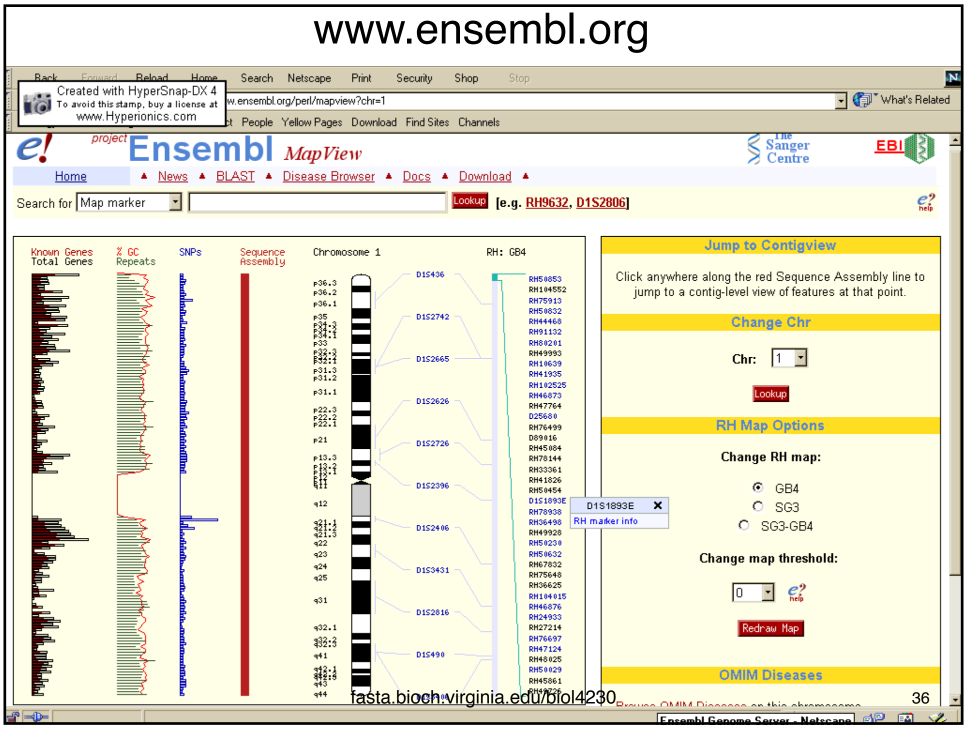
region	chr1:109631271-109750270
bases in region	119,000
bases in gaps	0
load time	0.02
calculation time	0.00
free memory time	0.00
filter	off
intersection	off

Genome data (UCSC table browser)

```
chr1   hg38_refGene       st_codon 109620269 109620271 0.000000 +  .         gene_id "NM_139156"; transcript_id "NM_139156";
chr1   hg38_refGene       CDS      109620269 109620278 0.000000 +  0         gene_id "NM_139156"; transcript_id "NM_139156";
chr1   hg38_refGene       exon     109619813 109620278 0.000000 +  .         gene_id "NM_139156"; transcript_id "NM_139156";
chr1   hg38_refGene       CDS      109625303 109625433 0.000000 +  2         gene_id "NM_139156"; transcript_id "NM_139156";
chr1   hg38_refGene       exon     109625303 109625433 0.000000 +  .         gene_id "NM_139156"; transcript_id "NM_139156";
chr1   hg38_refGene       CDS      109625662 109625792 0.000000 +  0         gene_id "NM_139156"; transcript_id "NM_139156";
chr1   hg38_refGene       exon     109625662 109625792 0.000000 +  .         gene_id "NM_139156"; transcript_id "NM_139156";
chr1   hg38_refGene       CDS      109626160 109626228 0.000000 +  1         gene_id "NM_139156"; transcript_id "NM_139156";
chr1   hg38_refGene       exon     109626160 109626228 0.000000 +  .         gene_id "NM_139156"; transcript_id "NM_139156";
chr1   hg38_refGene       CDS      109626319 109626427 0.000000 +  1         gene_id "NM_139156"; transcript_id "NM_139156";
chr1   hg38_refGene       exon     109626319 109626427 0.000000 +  .         gene_id "NM_139156"; transcript_id "NM_139156";
chr1   hg38_refGene       CDS      109626726 109626912 0.000000 +  0         gene_id "NM_139156"; transcript_id "NM_139156";
chr1   hg38_refGene       exon     109626726 109626912 0.000000 +  .         gene_id "NM_139156"; transcript_id "NM_139156";
chr1   hg38_refGene       CDS      109627175 109627316 0.000000 +  2         gene_id "NM_139156"; transcript_id "NM_139156";
chr1   hg38_refGene       exon     109627175 109627316 0.000000 +  .         gene_id "NM_139156"; transcript_id "NM_139156";
chr1   hg38_refGene       CDS      109627429 109627518 0.000000 +  1         gene_id "NM_139156"; transcript_id "NM_139156";
chr1   hg38_refGene       exon     109627429 109627518 0.000000 +  .         gene_id "NM_139156"; transcript_id "NM_139156";
chr1   hg38_refGene       CDS      109627774 109627903 0.000000 +  1         gene_id "NM_139156"; transcript_id "NM_139156";
chr1   hg38_refGene       exon     109627774 109627903 0.000000 +  .         gene_id "NM_139156"; transcript_id "NM_139156";
chr1   hg38_refGene       CDS      109628083 109628277 0.000000 +  0         gene_id "NM_139156"; transcript_id "NM_139156";
chr1   hg38_refGene       exon     109628083 109628277 0.000000 +  .         gene_id "NM_139156"; transcript_id "NM_139156";
chr1   hg38_refGene       CDS      109628364 109628495 0.000000 + 0    gene_id "NM_139156"; transcript_id "NM_139156";
```

GFF/GTF format

www.ensembl.org



Search for Map marker [Lookup] [e.g. RH9632, D1S2806]

Jump to Contigview

Click anywhere along the red Sequence Assembly line to jump to a contig-level view of features at that point.

Change Chr

Chr: 1 [Lookup]

RH Map Options

Change RH map:

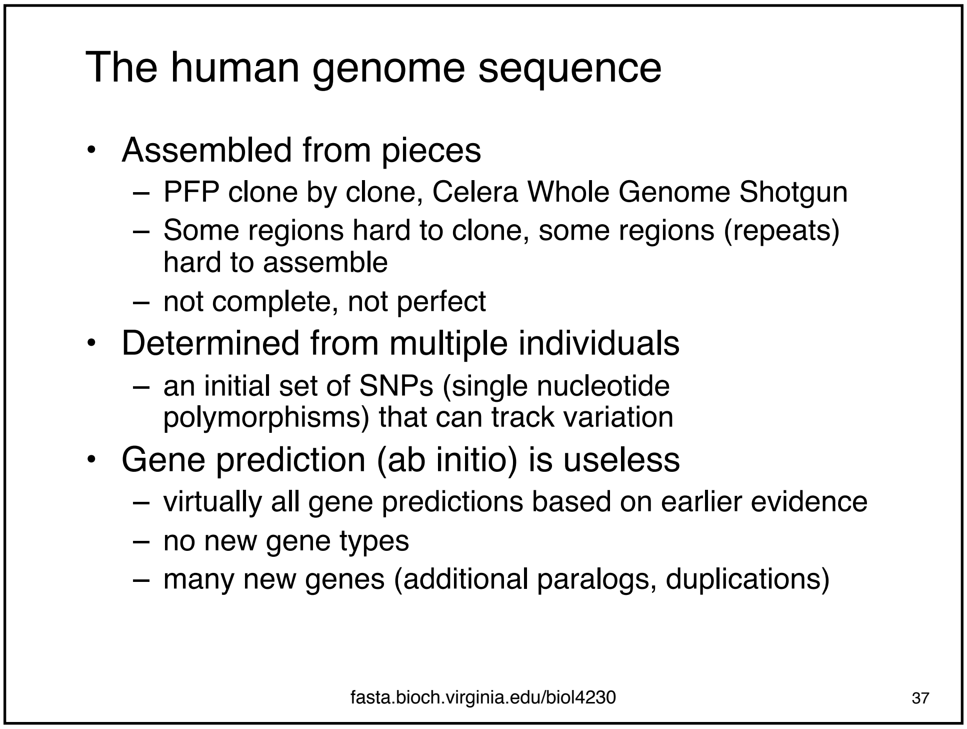
● GB4
○ SG3
○ SG3-GB4

Change map threshold:

0

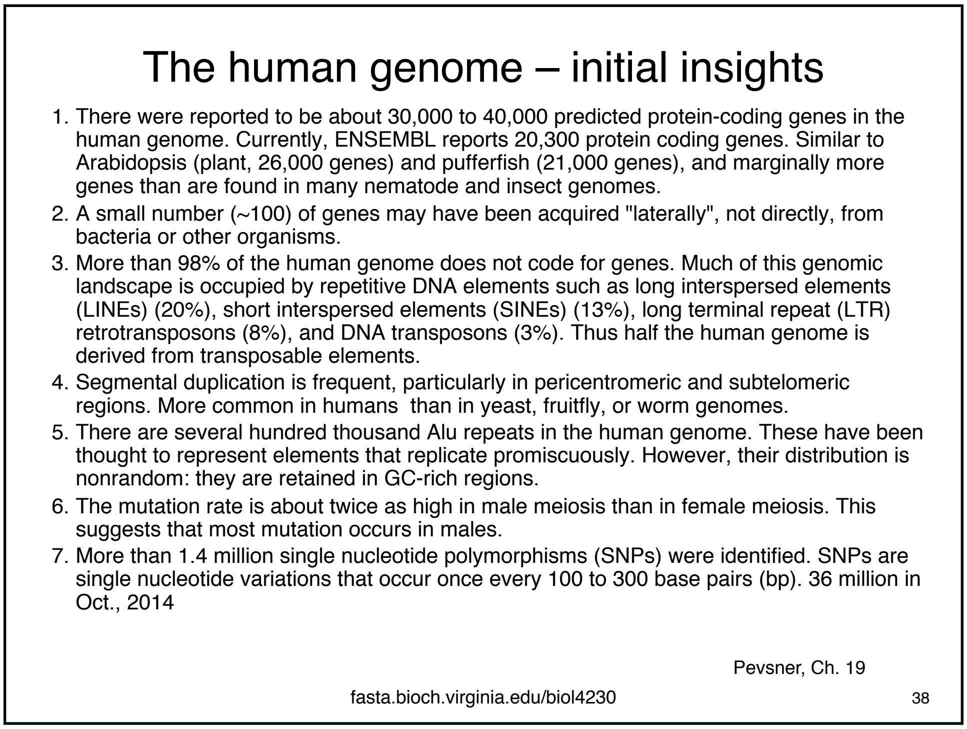
[Redraw Map]

OMIM Diseases



The human genome sequence

- Assembled from pieces
  - PFP clone by clone, Celera Whole Genome Shotgun
  - Some regions hard to clone, some regions (repeats) hard to assemble
  - not complete, not perfect
- Determined from multiple individuals
  - an initial set of SNPs (single nucleotide polymorphisms) that can track variation
- Gene prediction (ab initio) is useless
  - virtually all gene predictions based on earlier evidence
  - no new gene types
  - many new genes (additional paralogs, duplications)



The human genome – initial insights

1. There were reported to be about 30,000 to 40,000 predicted protein-coding genes in the human genome. Currently, ENSEMBL reports 20,300 protein coding genes. Similar to Arabidopsis (plant, 26,000 genes) and pufferfish (21,000 genes), and marginally more genes than are found in many nematode and insect genomes.
2. A small number (~100) of genes may have been acquired "laterally", not directly, from bacteria or other organisms.
3. More than 98% of the human genome does not code for genes. Much of this genomic landscape is occupied by repetitive DNA elements such as long interspersed elements (LINEs) (20%), short interspersed elements (SINEs) (13%), long terminal repeat (LTR) retrotransposons (8%), and DNA transposons (3%). Thus half the human genome is derived from transposable elements.
4. Segmental duplication is frequent, particularly in pericentromeric and subtelomeric regions. More common in humans than in yeast, fruitfly, or worm genomes.
5. There are several hundred thousand Alu repeats in the human genome. These have been thought to represent elements that replicate promiscuously. However, their distribution is nonrandom: they are retained in GC-rich regions.
6. The mutation rate is about twice as high in male meiosis than in female meiosis. This suggests that most mutation occurs in males.
7. More than 1.4 million single nucleotide polymorphisms (SNPs) were identified. SNPs are single nucleotide variations that occur once every 100 to 300 base pairs (bp). 36 million in Oct., 2014

Pevsner, Ch. 19

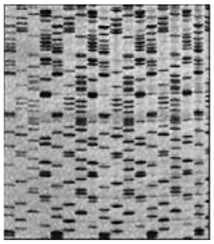
Sequencing capacity (2011) – Illumina sequencing 200 billion bases/week/machine, ~30,000 human genomes/year

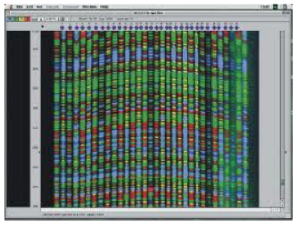
Sequencing capacity (2015) – at least 300,000 human genomes/year

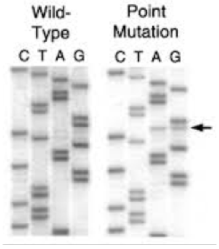
DNA sequencing technology pre-1998

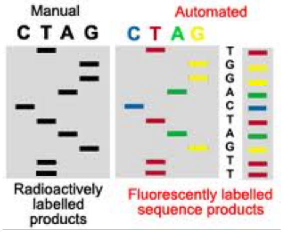
1977 – 1985
radioactive

1985 – 1995
dye-terminators

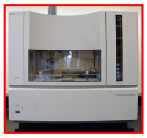
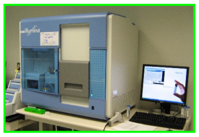
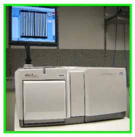







Manual
C T A G
Automated
C T A G
T G G A C T A G T T
Radioactively labelled products
Fluorescently labelled sequence products

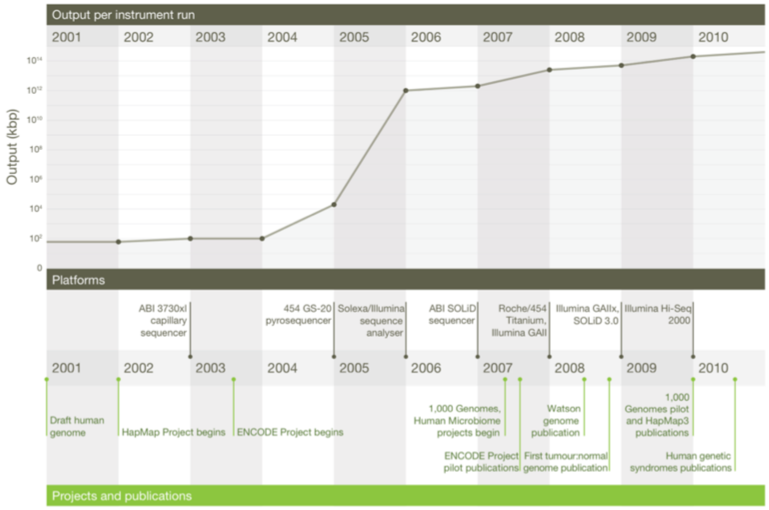
Adapted from Richard Wilson, School of Medicine, Washington University, “Sequencing the Cancer Genome” http://tinyurl.com/5f3alk

	3730	454 FLX	HiSeq
Genome size:		3 Gb == 3000	Mb
Req'd coverage:	6	12	24
bp/read	600	500	200
Reads/run	96	1,000,000	180,000,000
bp/run	57,600	500,000,000	4.E+10
#/runs req'd	312,500	72	2
Cost per run	$ 48	$ 7,500	$ 5,000
Total cost	**$15,000,000**	**$ 540,000**	**$ 10,000**

source: Francis Ouellette, OICR

Next (2^{nd},3^{rd},4^{th}) Generation Technologies



Mardis, *Nature* **470,** 198–203 (2011).

Next (2^{nd},3^{rd},4^{th}) Generation Technologies

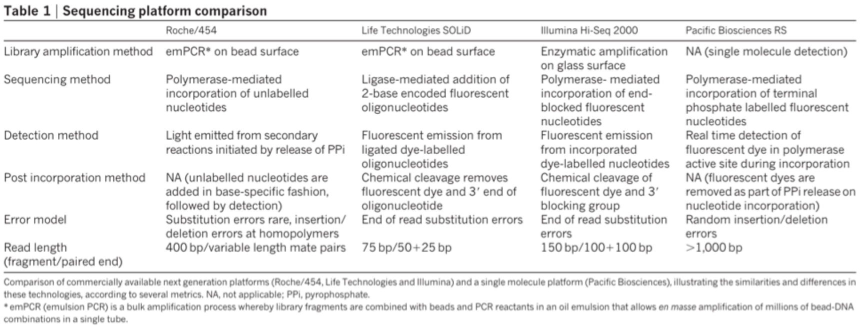


Table 1 | Sequencing platform comparison

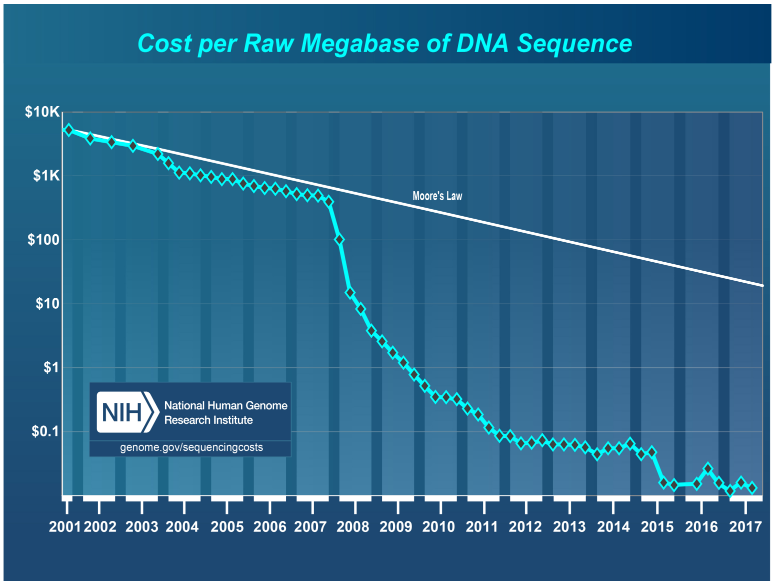
	Roche/454	Life Technologies SOLiD	Illumina Hi-Seq 2000	Pacific Biosciences RS
Library amplification method	emPCR* on bead surface	emPCR* on bead surface	Enzymatic amplification on glass surface	NA (single molecule detection)
Sequencing method	Polymerase-mediated incorporation of unlabelled nucleotides	Ligase-mediated addition of 2-base encoded fluorescent oligonucleotides	Polymerase- mediated incorporation of end-blocked fluorescent nucleotides	Polymerase-mediated incorporation of terminal phosphate labelled fluorescent nucleotides
Detection method	Light emitted from secondary reactions initiated by release of PPi	Fluorescent emission from ligated dye-labelled oligonucleotides	Fluorescent emission from incorporated dye-labelled nucleotides	Real time detection of fluorescent dye in polymerase active site during incorporation
Post incorporation method	NA (unlabelled nucleotides are added in base-specific fashion, followed by detection)	Chemical cleavage removes fluorescent dye and 3′ end of oligonucleotide	Chemical cleavage of fluorescent dye and 3′ blocking group	NA (fluorescent dyes are removed as part of PPi release on nucleotide incorporation)
Error model	Substitution errors rare, insertion/deletion errors at homopolymers	End of read substitution errors	End of read substitution errors	Random insertion/deletion errors
Read length (fragment/paired end)	400 bp/variable length mate pairs	75 bp/50+25 bp	150 bp/100+100 bp	>1,000 bp

Comparison of commercially available next generation platforms (Roche/454, Life Technologies and Illumina) and a single molecule platform (Pacific Biosciences), illustrating the similarities and differences in these technologies, according to several metrics. NA, not applicable; PPi, pyrophosphate.

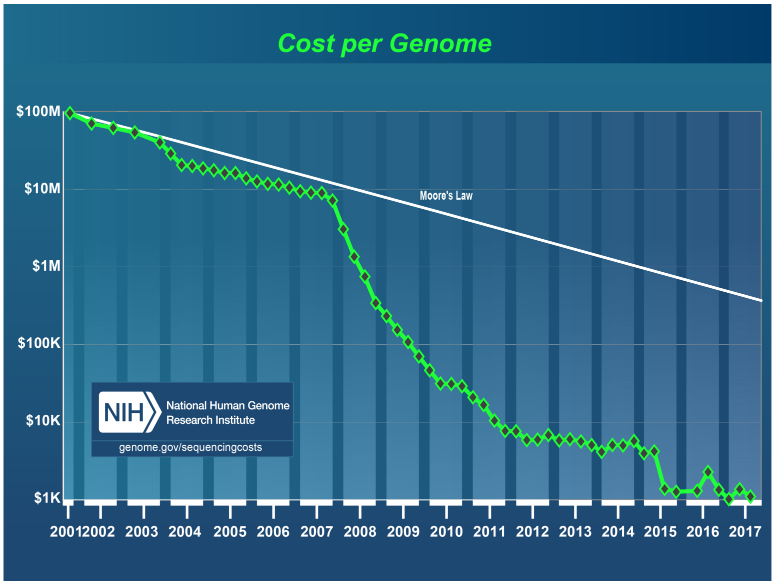
* emPCR (emulsion PCR) is a bulk amplification process whereby library fragments are combined with beads and PCR reactants in an oil emulsion that allows *en masse* amplification of millions of bead-DNA combinations in a single tube.

Mardis, *Nature* **470,** 198–203 (2011).

Sequencing Costs



Sequencing Costs



NGS applications

- genome (re)sequencing
  - *de novo* genomes: MiSeq Bact, small Euks, PacBio
  - SNP discovery and genotyping (barcoded pools), population genetics: Illumina
  - targeted, “deep” gene resequencing
  - metagenomics
- structural/copy-number variation
  - Tumor genome SV/CNV: Illumina/PET
- RNA-seq: transcriptomics
- ChIP/CLIP/etc-seq: DNA/RNA-binding, or DNA/RNA-modification
- Chromatin conformation capture (3C, Hi-C, etc.)

three phases of analysis

- primary
  - conversion of raw machine signal into sequence and qualities, QC, filtering, trimming, etc.
- secondary
  - read alignment to reference genome or transcriptome
  - or *de novo* assembly of reads into contigs
- tertiary
  - SNP discovery/genotyping
  - transcript clustering/quantification (RNA)
  - peak discovery/quantification (ChIP)

primary analyses

- Illumina GA Pipeline:
  - Firecrest: raw image -> clusters
  - Bustard: clusters -> sequence reads
  - Gerald/Eland -> raw alignment, sequence updates
- 454/IonTorrent: convert flowgram to FASTQ
- PacBio: decode video images as FASTQ, etc.
- core labs do these primary analyses for you
- “raw” image/video files are huge, and not stored
  - new primary analysis tools can’t be re-run on old data

FASTQ read format

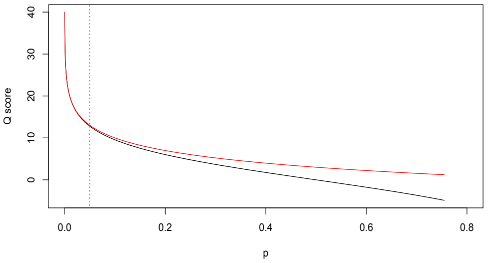
```
@HWUSI-EAS100R:6:73:941:1973#0/1
GATTTGGGGTTCAAAGCAGTATCGATCAAATA
+HWUSI-EAS100R:6:73:941:1973#0/1
!''*((((***+))%%%++)(%%%%).1***-
```

“Phred-scaled” base qualities

$$Q_{Phred} = -10\log_{10}\left(P_{err}\right)$$

$$Q_{Solexa} = -10\log_{10}\left(\frac{P_{err}}{1-P_{err}}\right)$$

$$Q_{Sanger} = 33 + \min(Q_{Phred}, 40)$$



```
 SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................................
 ..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX......................
 ...............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII......................
 .................................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ......................
 LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL....................................................
 !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
 |                         |    |        |                              |                     |
33                        59   64       73                            104                   126
S - Sanger        Phred+33,  raw reads typically (0, 40)
X - Solexa        Solexa+64, raw reads typically (-5, 40)
I - Illumina 1.3+ Phred+64,  raw reads typically (0, 40)
J - Illumina 1.5+ Phred+64,  raw reads typically (3, 40)
   with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)
   (Note: See discussion above).
L - Illumina 1.8+ Phred+33,  raw reads typically (0, 41)
```

conversion of sol/ill qualities

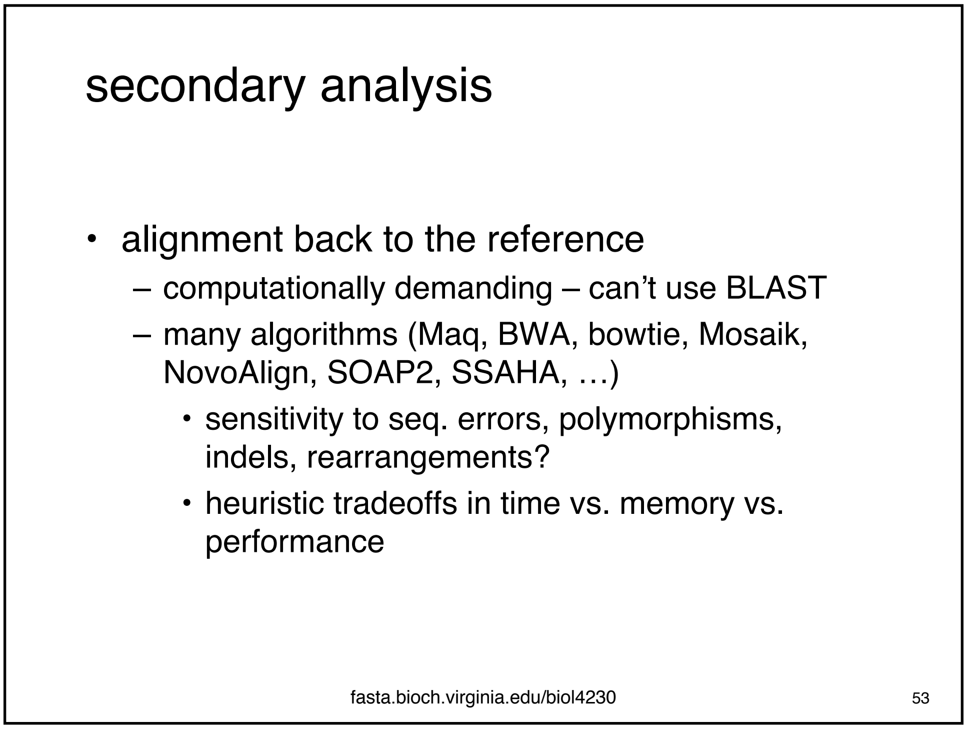
- “maq” package contains `sol2sanger` and `ill2sanger` utilities to convert to standard Phred-scaled quality encoding
- Or, usegalaxy.org “FASTQ Groomer”

read filtering/trimming/QC

- newer Illumina pipelines deliver unfiltered reads, with “chastity” filter tags:

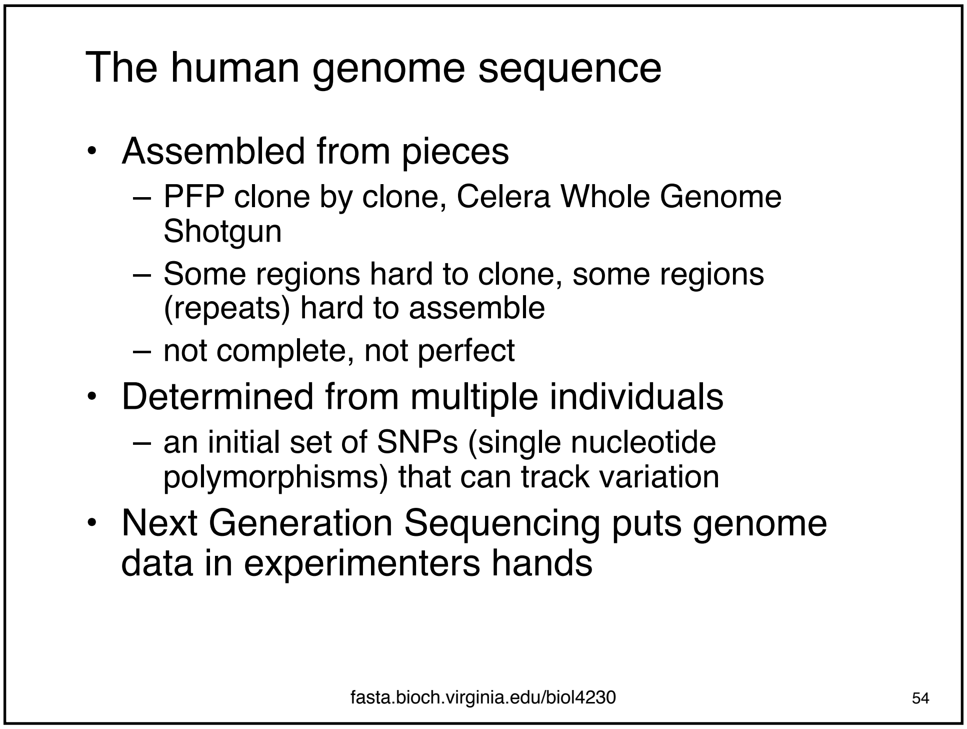
```
@EAS139:136:FC706VJ:2:5:1000:12850 1:Y:18:ATCACG
```

  - pipeline version dictates whether Y means “bad” (1.8+, recent) or “good” (pre-1.8)
- chimeric reads containing adapters, primers, etc. should be trimmed (sickle, scythe)
- barcoding, merging, data manipulations
- FASTQC
  - http://www.bioinformatics.babraham.ac.uk/projects/fastqc/



secondary analysis

- alignment back to the reference
  - computationally demanding – can't use BLAST
  - many algorithms (Maq, BWA, bowtie, Mosaik, NovoAlign, SOAP2, SSAHA, …)
    - sensitivity to seq. errors, polymorphisms, indels, rearrangements?
    - heuristic tradeoffs in time vs. memory vs. performance



The human genome sequence

- Assembled from pieces
  - PFP clone by clone, Celera Whole Genome Shotgun
  - Some regions hard to clone, some regions (repeats) hard to assemble
  - not complete, not perfect
- Determined from multiple individuals
  - an initial set of SNPs (single nucleotide polymorphisms) that can track variation
- Next Generation Sequencing puts genome data in experimenters hands