

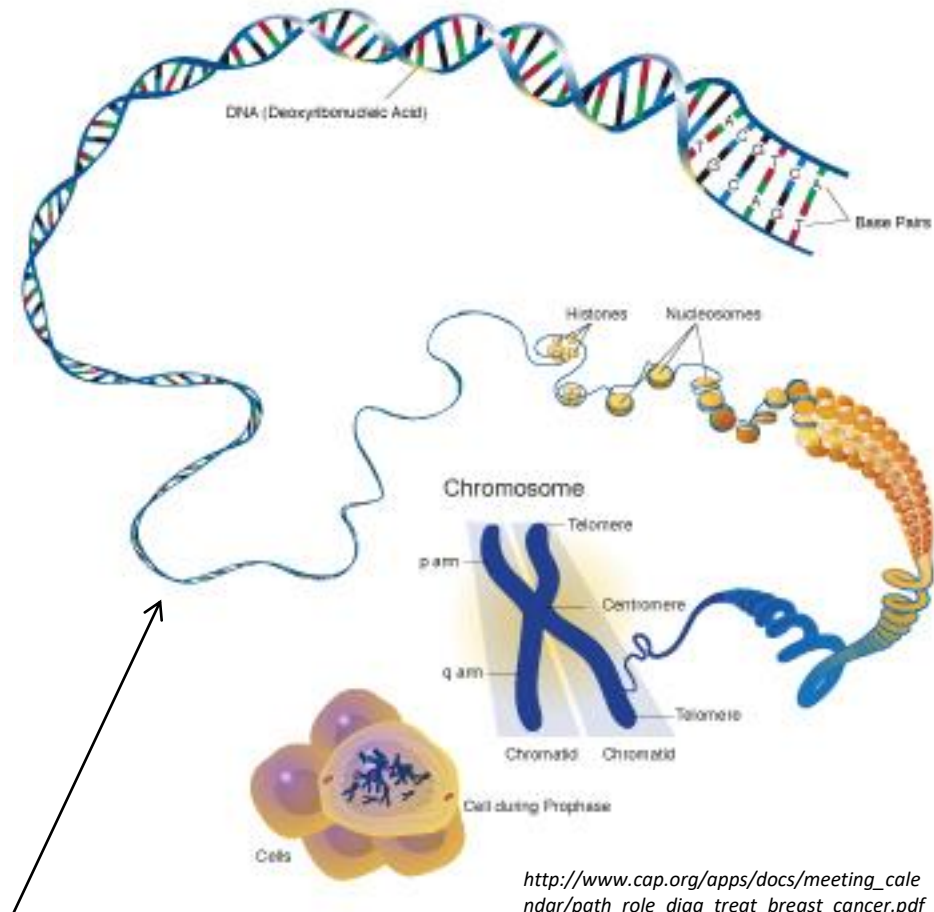
Reading the blueprint of life

Genomics & DNA sequencing

Prof. David Jones
d.t.jones@ucl.ac.uk

Genomics

- Genomic data
 - Size & structure
 - Organisation of genes
- Methods for gene prediction
 - By similarity
 - Ab initio
 - Pattern recognition



Where are the
genes???

http://www.cap.org/apps/docs/meeting_calendar/path_role_diag_treat_breast_cancer.pdf

Introduction

- The blueprint of life is contained in the DNA in the nuclei of eukaryotic cells and simply within prokaryotic cells – the *genome*
- Human genome project – just obtain the list of approximately 3×10^9 bases (As, Cs, Gs and Ts) in the 23 chromosomes.
- Extraction of useful information from this list and genome sequence of other organisms relies on computer-intensive data handling – i.e.
Bioinformatics!

C-value & Ploidy

- C-value - this is the size of the genome
- Ploidy is the number of HOMOLOGOUS sets of CHROMOSOMES in a biological cell
- Therefore a C-value is the total number of nucleotide bases found in the HAPLOID GENOTYPE (HAPLOTYPE)
- For DIPLOID organisms (e.g. humans) the C-value is half the total number of bases in the nucleus (i.e. every chromosome is paired with a homologue – one from the mother and one from the father)
- Human sex cells (sperm & egg cells) are HAPLOID i.e. only one copy of each homologous chromosome is present

Guess the C-value

HIV



19,750

Homo sapiens



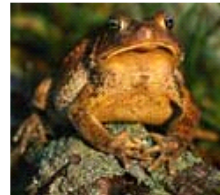
3,055,000,000

Rhinolophus ferrumequinum



1,929,400,000

Bufo bufo



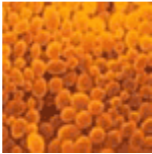
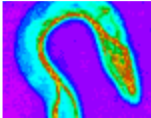
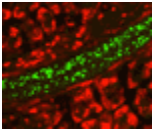



6,900,000,000

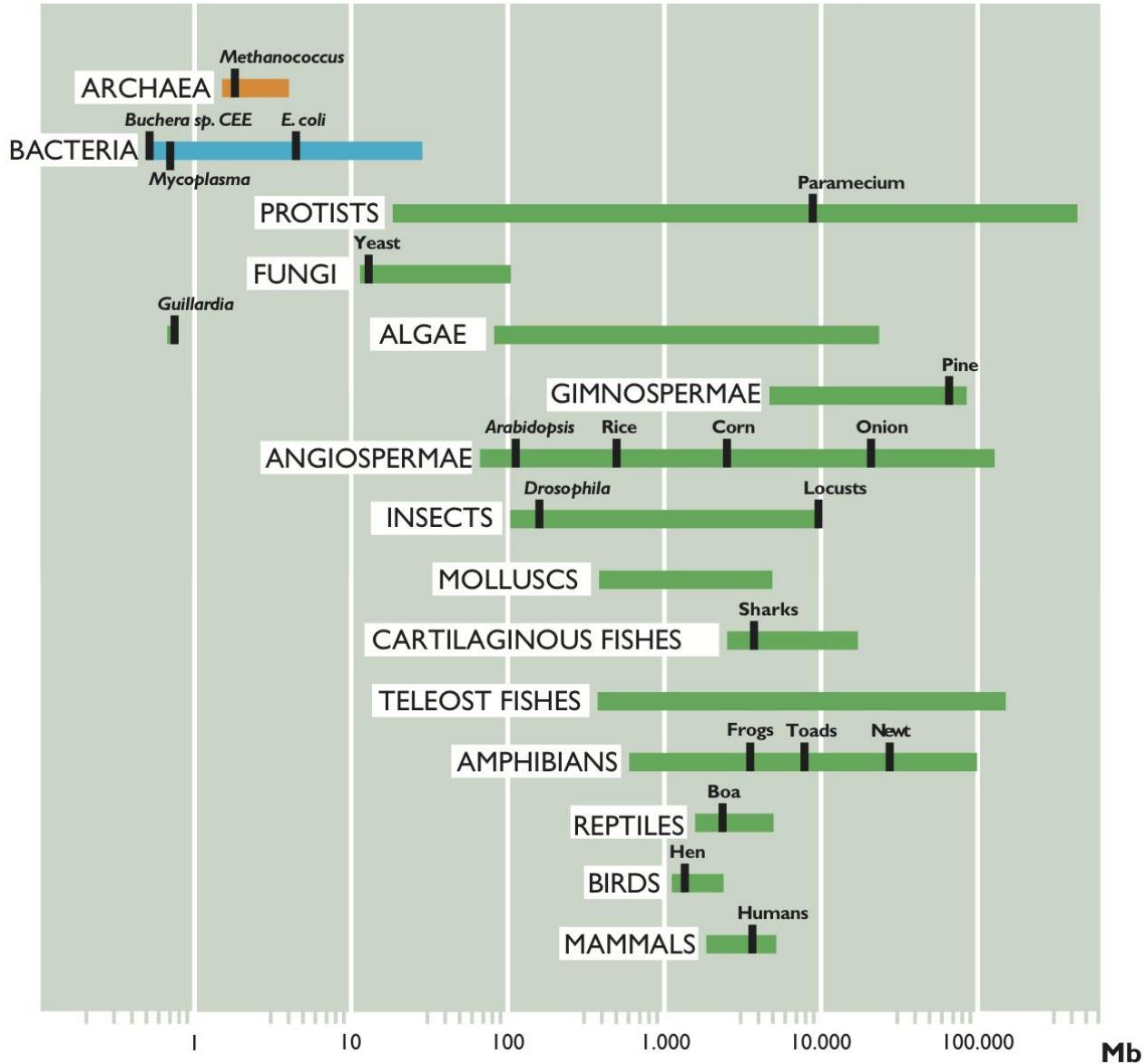
Amoeba dubia



670,000,000,000

Species	Size of genome	Number of genes
Human 	3.06 billion base pairs	20,441
Fruit fly (<i>Drosophila melanogaster</i>) 	120 million base pairs	13,601
Baker's yeast (<i>Saccharomyces cerevisiae</i>) 	12 million base pairs	6, 275
Worm (<i>Caenorhabditis elegans</i>) 	97 million base pairs	20,470
<i>E. coli</i> 	4.1 million base pairs	4,800
Arabidopsis (<i>Arabidopsis thaliana</i>) 	125 million base pairs	27,655

Genome Sizes



Species

Human



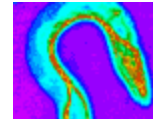
Fruit fly (*Drosophila melanogaster*)



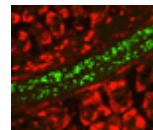
Baker's yeast (*Saccharomyces cerevisiae*)



Worm (*Caenorhabditis elegans*)



E. coli



Arabidopsis (*Arabidopsis thaliana*)



Type

Multicellular vertebrate

Multicellular
Invertebrate

Unicellular
Eukaryote

Multicellular
Invertebrate

Unicellular
Prokaryote

Multicellular
Plant

Human Gene Counts

(July 2022)

Gene Type	Count
Coding genes	19,813 (excl 651 readthrough*)
Non coding (RNA) genes	25,972
Pseudogenes	15,241

* Readthrough genes are “genes within genes” i.e. RNA splices that span two adjacent genes.

Pop Quiz

- C-value for a human is $\sim 3 \times 10^9$
- The average length of a protein is 200 amino acids
- There are $\sim 20,000$ human genes

So...

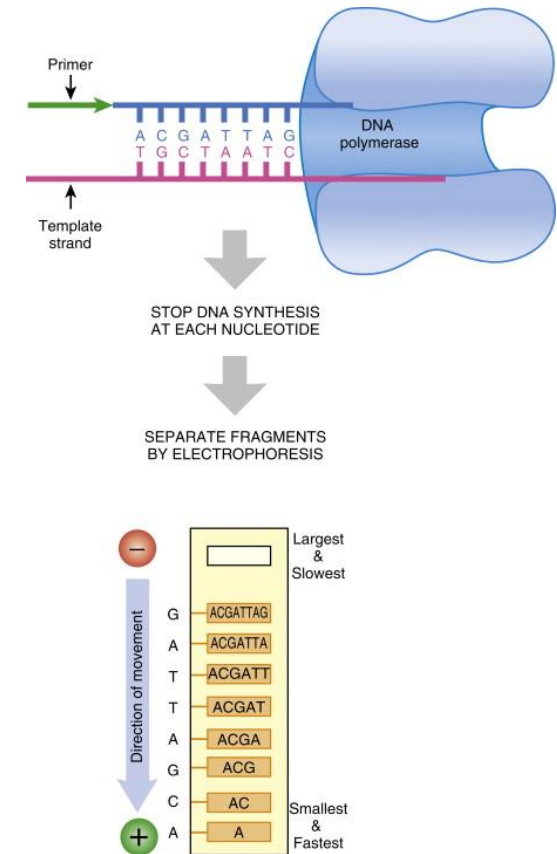
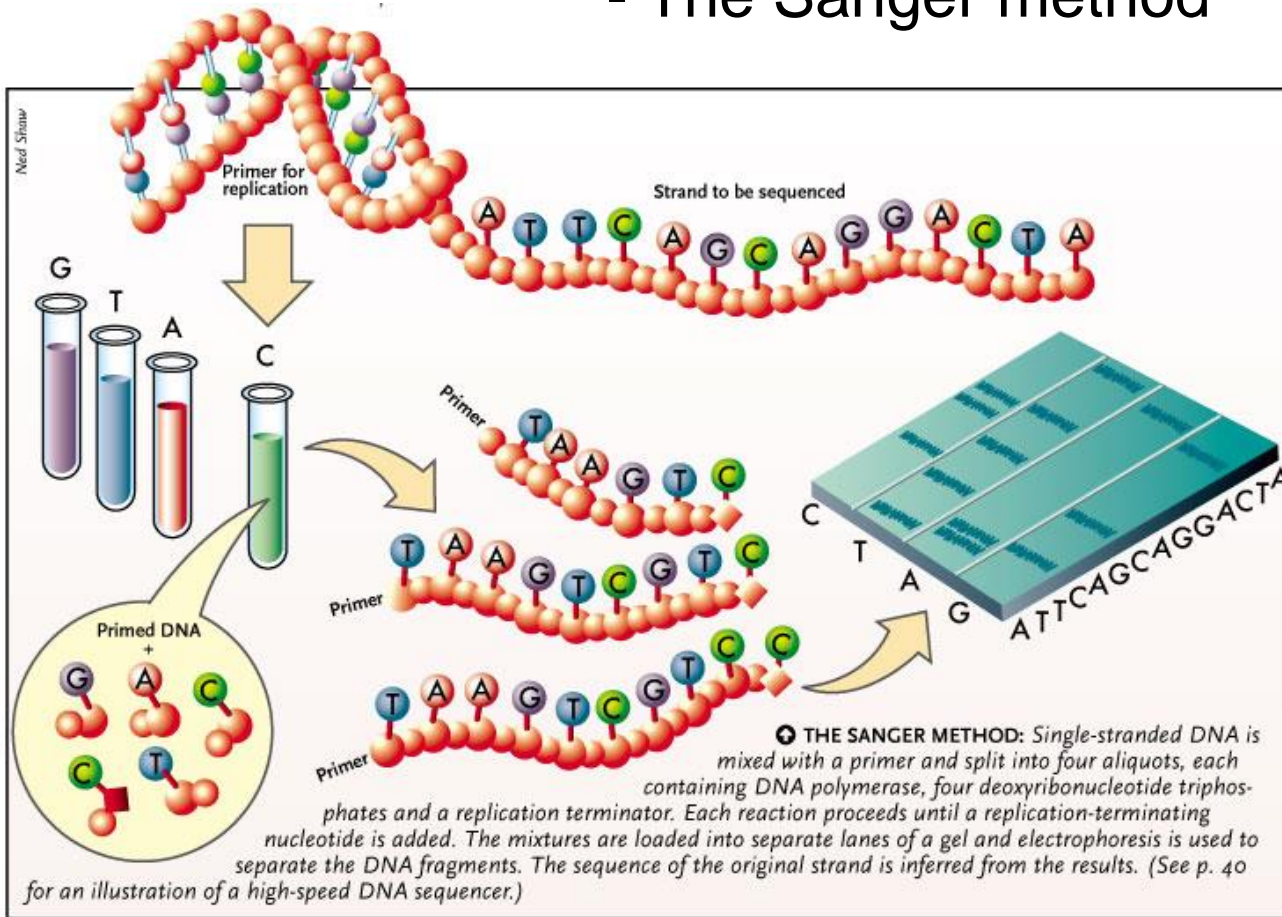
WHAT MINIMUM PERCENTAGE OF THE HUMAN GENOME
CODES FOR PROTEIN?

How to sequence DNA

- Denatured DNA is added to a reaction mix with:
 - a primer (to start complementary pairing),
 - DNA polymerase
 - Nucleotides, including special ones called **dideoxynucleotides**. These special nucleotides do not allow further nucleotides to be added to the chain. So in a mix with **dideoxy-A**, every time a dideoxy-A is added (small proportion of As), the reaction ends. This results in fragments of different length. The dideoxynucleotides are fluorescently tagged.
- Fragments can be separated out on a gel by **electrophoresis** and their length calculated. Working out DNA sequence ~ jigsaw puzzle.

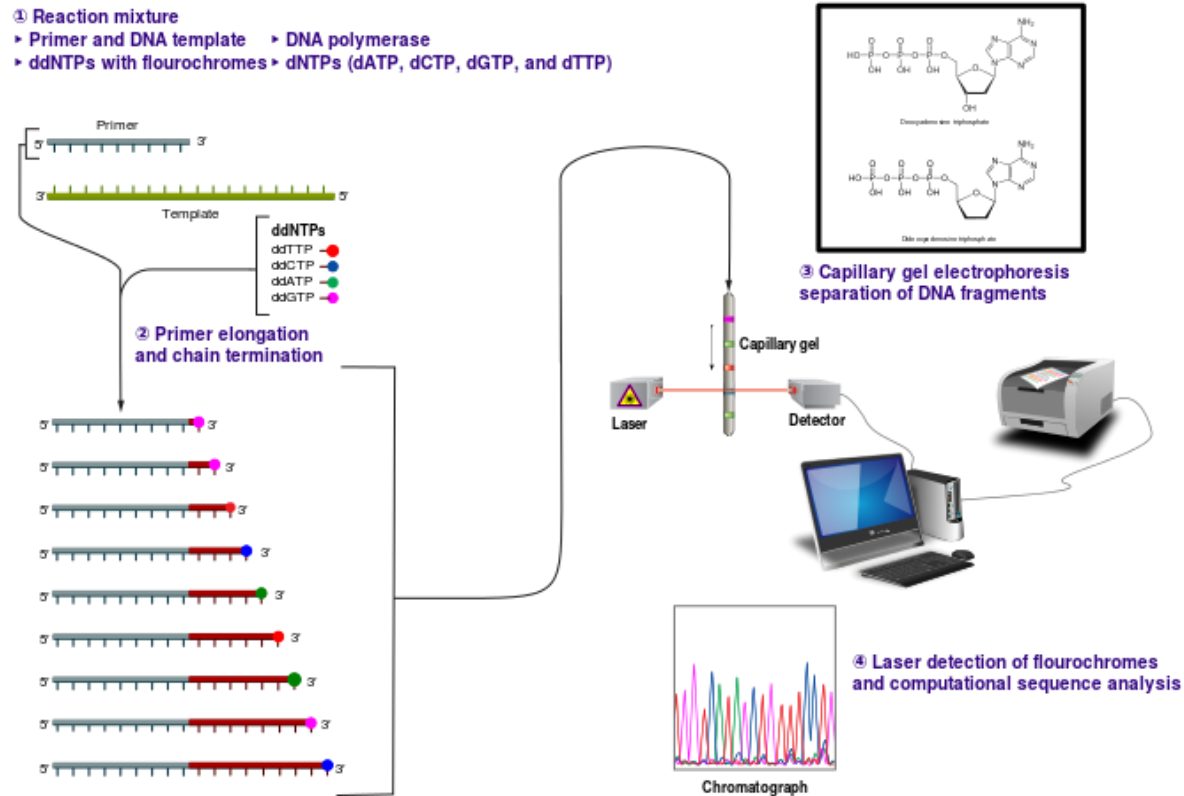
How to sequence DNA

- The Sanger method



High Throughput Sanger Sequencing

- Principles are similar to normal DNA sequencing
- Capillaries used instead of gel plates
- Instead of “spots”, continuous traces are read by laser
- Improved accuracy, speed and cost over the traditional method



Buzzwords to read up on...

- C-value
- Haploid/Diploid
- Chromosome
- Introns/exons
- Sanger sequencing
- High throughput sequencing
- DNA amplification
- Polymerase Chain Reaction (PCR)

How the human genome was sequenced

- High throughput sequencing technology works for around 500-1200 base-pair fragments of DNA
- Top of the line equipment can sequence 1200 bases with 99% accuracy
- Main task in genome sequencing is therefore to break the genome into pieces (called CONTIGS) which can be sequences ...
- ... and eventually (hopefully!) re-assemble the pieces into complete sequences afterwards

How to sequence a whole genome

- The DNA from the genome was chopped into bits- whole chromosomes are too large to deal with, so the DNA is broken into manageably-sized overlapping segments.
- The DNA was amplified by cloning into bacteria rather than PCR, which is now the main method.
- It is then **denatured** (i.e. melted), so that the two strands split apart.

How to select the fragments

- The Celera way
 - Smash up the genome randomly e.g. using sonication (ultrasound)
- The Human Genome Project Way
 - Build yourself a map
 - Genetic Map: based on known genetic features (e.g. inherited diseases)
 - Low accuracy + not anchored to actual sequence
 - Physical Map: based on sequence features
 - Accurate + anchored to specific regions of sequence
 - Cut genome according to mapped features

How to Assemble the fragments – The Hard Way



Acknowledgements : John Sgouros & Cancer Research UK (formerly ICRF)

Sequence Assembly

ERYCHEAPAN

OGYISVERYC

SEQUENCINGT

MODERNGENO

INGTECHNOLOGYI

APANDRELIABLE

GENOMESEQ

MODERNGENO

GENOMESEQ

SEQUENCINGT

INGTECHNOLOGYI

OGYISVERYC

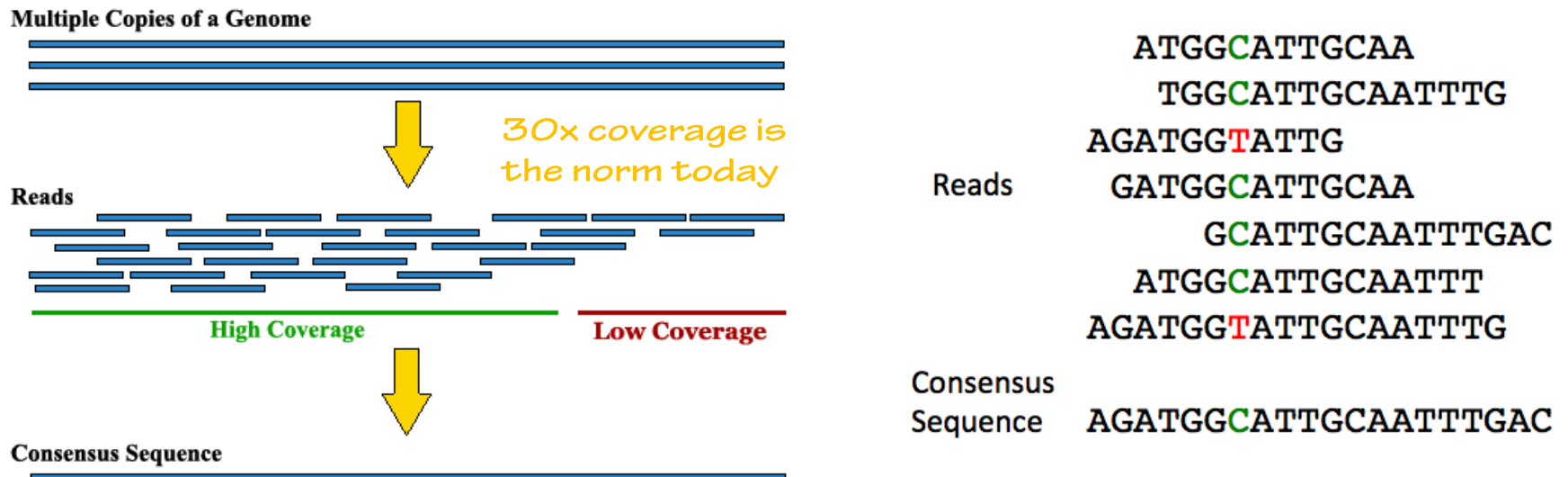
ERYCHEAPAN

APANDRELIABLE

MODERNGENOMESEQUENCINGTECHNOLOGYISVERYCHEAPANDRELIABLE

Acquisition of sequence data

- Genomes must be sequenced several times over on average, both to ensure complete coverage of the genome is achieved, and because sequencing data is somewhat error-prone.
- Increases in the efficiency of sequencing have led to a year on year increase in the rate of new sequence data acquisition:



How the human genome sequence was assembled

- The Celera way
 - Identify overlapping ends of sequenced fragments
 - Attempt to sort fragments (plus orient them) according to overlap scores
- The HGP way
 - First fit fragments to physical map (i.e. known start/end points)
 - Fill in gaps as with the Celera way
- The actual Celera way
 - Assemble as much as you can using the sort algorithm
 - Assemble the large fragments according to the map made freely available by your “competitors” i.e. the public HGP!
- PROBLEM: Human DNA contains highly repetitive regions of sequence which cannot be accurately sequenced or uniquely assembled

The Human Genome Project

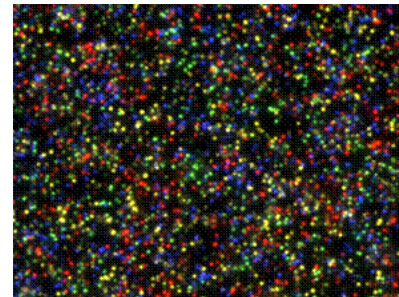
- 1990 - HGP started
- 1995 - Physical map published
- 2000 - Working draft sequence published
- 2003 - HGP “finished” (92% completed)
- .
- .
- .
- 2022 – Last 8% of the genome published

In 2022, the world record for complete sequencing of a single human genome stands at just over 5 *hours*!

Genome sequencing gets faster and cheaper

The Illumina NovaSeq X machine

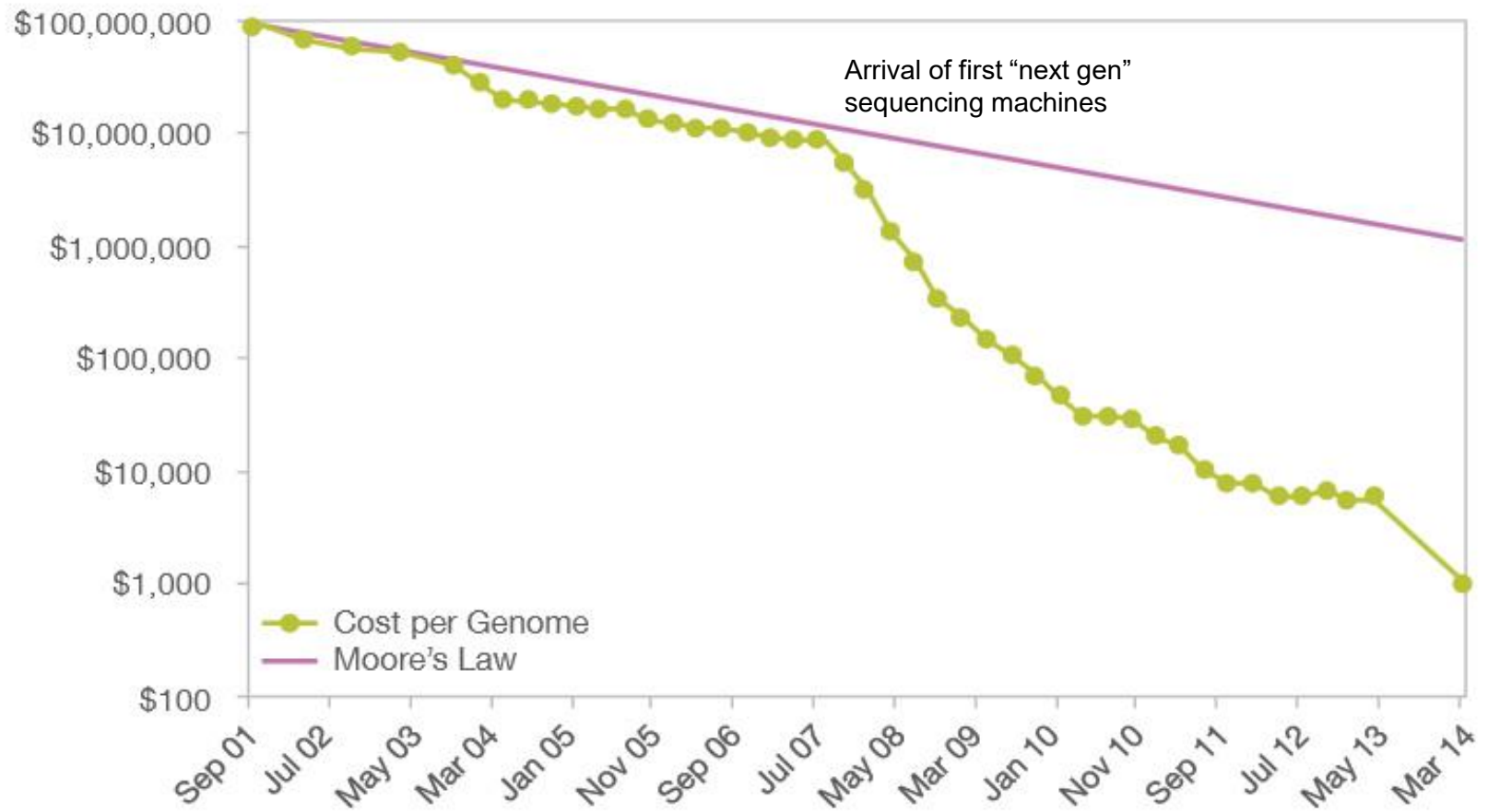
- Single reads per run 1.6 billion - 52 billion
- Read length 150 bp (x 2)
- Run time ~13 hr - 48 hr
- Machine cost ~ \$1 million
- Running cost per human genome ~ \$200



If you want a basic intro into how Illumina dye synthesis sequencing works, this short video is worth watching:

<https://www.youtube.com/watch?v=CZeN-IgjYCo>

Genome sequencing is now overtaking the growth in available computing power



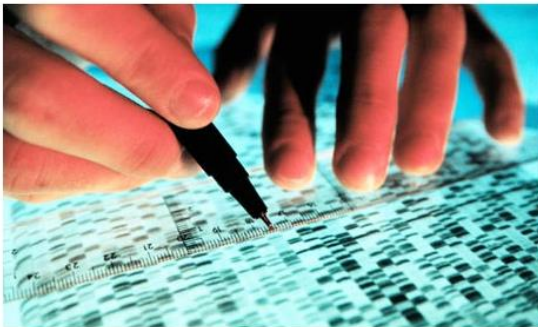
DNA of 100,000 people to be mapped for NHS

Government hopes public health programme will revolutionise treatment and prevention of cancer and other diseases

Peter Walker

theguardian.com, Monday 10 December 2012 10:14 GMT

[Jump to comments \(229\)](#)



David Cameron says that by 'unlocking the power of DNA data', the NHS will lead the global race for better healthcare. Photograph: Deco Images III/Alamy

Up to 100,000 people in England will have their entire genetic makeup mapped in the first stage of an ambitious public health programme the government hopes could revolutionise the treatment and prevention of cancer and other diseases.

Ministers have committed an initial £100m for the project, which aims to take advantage of the tumbling cost of mapping an individual's full DNA sequence to make genetic analysis a key component of some medical treatments. During an initial three to five-year period, up to 100,000 people with cancer or certain rare diseases will voluntarily have their DNA mapped.

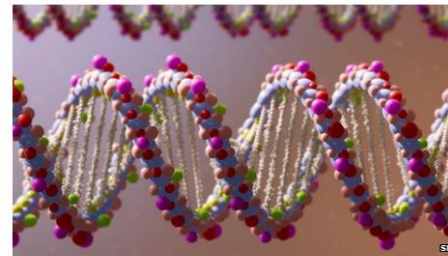
7 November 2013 Last updated at 00:03

[Share](#) [f](#) [t](#) [e](#) [d](#)

Massive DNA volunteer hunt begins

By James Gallagher

Health and science reporter, BBC News



Scientists are looking for 100,000 volunteers prepared to have their DNA sequenced and published online for anyone to look at.

The UK **Personal Genome Project** could provide a massive free tool for scientists to further understanding of disease and human genetics.

Participants will get an analysis of their DNA, but so will the rest of the world, and anonymity is not guaranteed.

They are warned there could be unknown consequences for them and relatives.

Unlocking the secrets of DNA could transform the understanding of disease.

Related Stories

[Most human gene variations mapped](#)

[Alzheimer's insight from DNA study](#)

[Acceleration in genome sequencing](#)

Project had sequenced 50,000 human genomes by Feb 2018

By Dec 2018, it reached its goal of sequencing 100,000 genomes!

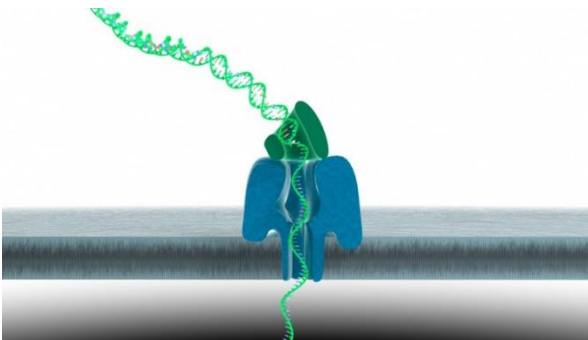
UK Biobank now (2024) has 500,000 genomes sequenced.

Nanopore sequencing – the “MinION” (DNA Sequencer on a USB Stick)



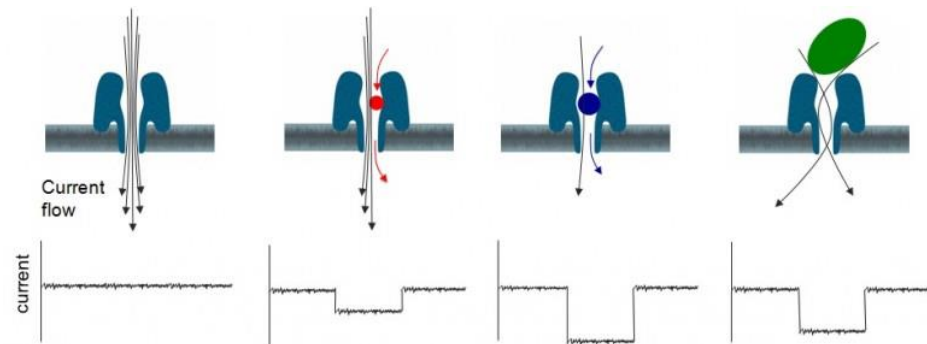
- Your own personal DNA sequencer on a USB stick for only £900!
- Can sequence 150 kilobases of DNA per hour
- No post-processing needed - sequence streams straight to your PC
- Downsides
 - Error rate high – was as high as 40% but now c. 15%
 - Device expires after 6 hours use!

<https://nanoporetech.com/technology/the-minion-device-a-miniaturised-sensing-system/the-minion-device-a-miniaturised-sensing-system>



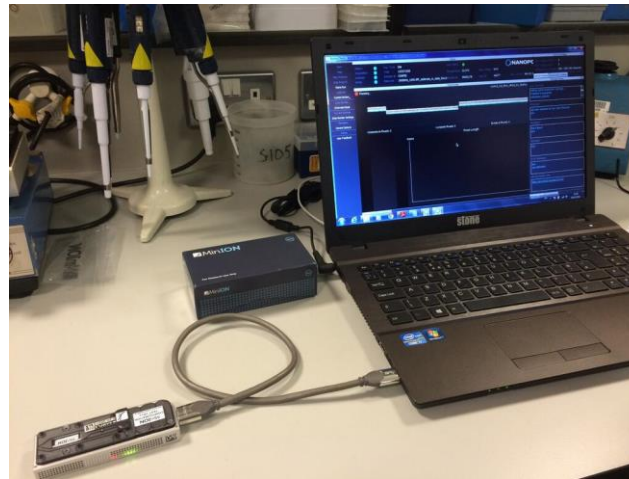
Each MinION has 512 of these nanopores

<https://www.freedomspheonix.com/News/105725-2012-02-21-minion-900-usb-powered-dna-sequencer-on-sale-this-year.htm>



<http://www.gizmag.com/minion-disposable-dna-sequencer/21513/>

Good for field work, though...



MinION Nanopore sequencer from Oxford Nanopore costs around £1000 per unit and plugs into a standard USB port!



Team from University of Birmingham used MinION sequencers to track the spread and mutation of the Ebola virus in the field, during the recent outbreak in West Africa

Conclusions

- Sequencing DNA involves:
 - Amplifying it by PCR or cloning
 - Chopping it up into manageable bits
 - Replicating it with fluorescently-tagged dideoxynucleotides
 - Running the different length fragments on a gel or through a capillary and reading the fluorescence signals
 - Assembling the pieces (sequences of manageable bits)
- Shotgun sequencing is faster than mapping-based assembly methods, but can have accuracy problems
- Next-generation sequencing methods are now widely used
 - Can sequence whole genomes in a week or two
 - Are replacing array technologies to monitor gene expression
 - Now frequently used on large patient cohorts (e.g. looking for mutations involved in disease)

Finding Genes

- The most important parts of the genome are the genes.
- Efforts have been made to identify genes out of sequence data.
- **Expressed sequence tags (ESTs)** are short pieces of sequence data that correspond to mRNAs found in cells of the organism.
- ESTs are produced by purifying mRNA from cells and then using an enzyme called **reverse transcriptase** to convert these to **copy DNA (cDNA)**. The DNA is then cloned in bacteria and sequenced.
- The sequence obtained is usually only short (c. 700 base pairs) and may not be very accurate, but ESTs still provide very useful information.

Gene prediction

- A weakness of ESTs is that it is very difficult to obtain them for genes which are expressed at a low level/ only under certain conditions, also slow, so
- People try to predict where in sequence the genes are.
- In prokaryotes, just look for long stretches of DNA without stop codon in any of the 6 reading frames.

Finding Genes in Prokaryotes

- Open reading frames

- There are 6 reading frames, 3 forwards:

5' 3'

atgcccgaagctgaatagcgtagaggggttttcatcatttgaggacgatgtataa

1 atg ccc aag ctg aat agc gta gag ggg ttt tca tca ttt gag gac gat gta taa
M P K L N S V E G F S S F E D D V *

2 tgc cca agc tga ata gcg tag agg ggt ttt cat cat ttg agg acg atg tat
C P S * I A * R G F H H L R T M Y

3 gcc caa gct gaa tag cgt aga ggg gtt ttc atc att tga gga cga tgt ata
A Q A E * R R G V F I I * G R C I

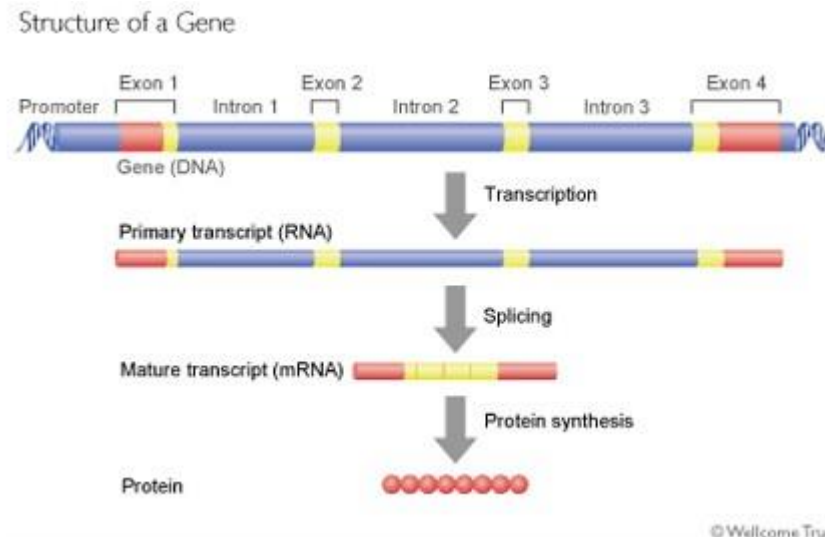
- And 3 backwards (on the other strand). A frame is said to be open if it contains long stretches without a stop codon.
- [Lower lines are single-letter amino acid codes, *=stop.]

Gene prediction in eukaryotes

- In bacteria, open reading frames (ORFs) are pretty much enough to indicate genes, but in eukaryotes finding genes is more complicated, because
 1. Eukaryotic DNA (e.g. human DNA) is roughly 97-98% noncoding - in such a large amount, ORFs may exist by chance.
 2. Eukaryotic DNA contains introns, so finding the start and the end of a gene is not enough- also have to find which bits (introns) to edit out of sequence. Also introns break up open reading frames.

Introns - reminder

- Mentioned in first lecture
- These are pieces of DNA within genes, which are transcribed but then spliced out of the RNA before it is translated.
- They make it much harder to find genes, since finding open reading frames is not enough, you also need to find where introns and exons start and end.



Prokaryotic Vs. Eukaryotic Gene Finding

Prokaryotes:

- Small genomes $\sim 10^6 - 10^7$ bp
 - High gene density (>90%)
 - No introns
- Gene identification relatively easy, with success rate $\sim 99\%$

Problems:

- Overlapping ORFs
- Short genes
- Finding TSS and promoters

Eukaryotes:

- Large genomes $10^7 - 10^{10}$ bp
 - Low gene density (<50%)
 - Intron/exon structure
- Gene identification a complex problem, gene level accuracy $\sim 50\%$

Problems:

- Many!

Gene Finding: Different Approaches

- **Similarity-based methods (extrinsic)** - use similarity to annotated sequences:
 - Known proteins
 - Known cDNAs
 - ESTs
- **Comparative genomics** - Aligning genomic sequences from different species
- ***Ab initio* gene-finding (intrinsic) – pattern recognition**
- **Hybrid approaches**

Similarity-based methods

- Based on sequence conservation due to functional constraints
- Use local alignment tools (Smith-Waterman algorithm, BLAST, FASTA) to search protein, cDNA, and EST databases
- Will not identify genes that code for proteins not already in databases (can identify ~50% new genes)
- Limits of the regions of similarity not well defined
- Very sensitive to so-called “frame shift errors”
- “Defunct” genes (pseudogenes) create false positives

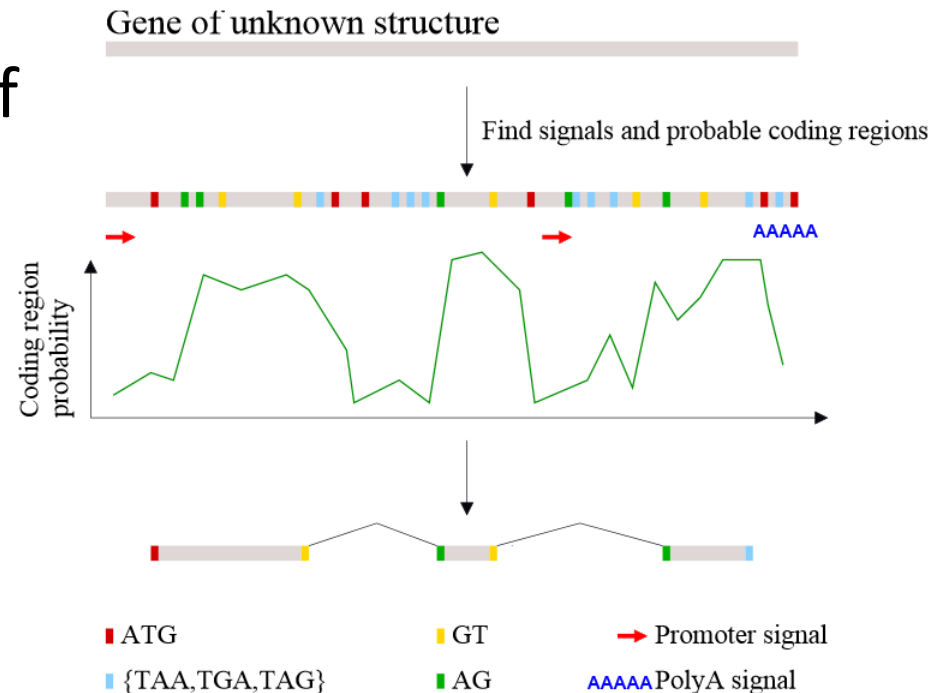
Comparative Genomics

- Based on the assumption that coding sequences are more conserved than non-coding
- Two approaches:
 - intra-genomic (gene families)
 - Does this region resemble a region elsewhere in the genome?
 - inter-genomic (cross-species)
 - Does this region resemble a region in the genome of another species?
- Alignment of homologous regions
- Difficult to define limits of similarity i.e. start/end points
- “Defunct” genes (pseudogenes) create false positives

Ab initio Gene Finding

- Using only sequence information
- Identify only coding exons of protein-coding genes (transcription start site, 5' and 3' UTRs are ignored)
- Combine statistical analysis of coding regions (biases in the coding sequence) with signal detection (i.e. parsing features in the sequence)

Ab initio method



Statistical analysis of genome sequences

Statistics can be global:

- **Base composition of genomes:**
 - *Bacteria (E. coli)*: 25% A, 25% C, 25% G, 25% T
 - *Malaria parasite (P. falciparum)*: 82%A+T
 - *Human*: 59%A+T

Or local:

- **Translation initiation:**
 - ATG is the near universal **motif** (codon) indicating the start of translation in DNA coding sequence (start of the first exon in a eukaryote).

Coding Statistics

- Biased usage of codons in the coding regions of genes is a universal feature of the genomes
 - uneven usage of amino acids in existing proteins
 - uneven usage of synonymous codons (correlates with the abundance of corresponding tRNAs)
 - Varies according to species
- We can use these features to distinguish between coding and non-coding regions of the genome
- Coding statistics - a function that for a given DNA sequence gives a likelihood that the sequence codes for a protein

Coding Statistics

- Many varieties described in the literature
 - codon usage
 - hexamer usage
 - GC content
 - compositional bias between codon positions
 - nucleotide periodicity
 - ...

Codon Usage

The Human Codon Usage Table															
Gly	GGG	17.08	0.23	Arg	AGG	12.09	0.22	Trp	TGG	14.74	1.00	Arg	CGG	10.40	0.19
Gly	GGA	19.31	0.26	Arg	AGA	11.73	0.21	End	TGA	2.64	0.61	Arg	CGA	5.63	0.10
Gly	GGT	13.66	0.18	Ser	AGT	10.18	0.14	Cys	TGT	9.99	0.42	Arg	CGT	5.16	0.09
Gly	GGC	24.94	0.33	Ser	AGC	18.54	0.25	Cys	TGC	13.86	0.58	Arg	CGC	10.82	0.19
Glu	GAG	38.82	0.59	Lys	AAG	33.79	0.60	End	TAG	0.73	0.17	Gln	CAG	32.95	0.73
Glu	GAA	27.51	0.41	Lys	AAA	22.32	0.40	End	TAA	0.95	0.22	Gln	CAA	11.94	0.27
Asp	GAT	21.45	0.44	Asn	AAT	16.43	0.44	Tyr	TAT	11.80	0.42	His	CAT	9.56	0.41
Asp	GAC	27.06	0.56	Asn	AAC	21.30	0.56	Tyr	TAC	16.48	0.58	His	CAC	14.00	0.59
Val	GTG	28.60	0.48	Met	ATG	21.86	1.00	Leu	TTG	11.43	0.12	Leu	CTG	39.93	0.43
Val	GTA	6.09	0.10	Ile	ATA	6.05	0.14	Leu	TTA	5.55	0.06	Leu	CTA	6.42	0.07
Val	GTT	10.30	0.17	Ile	ATT	15.03	0.35	Phe	TTT	15.36	0.43	Leu	CTT	11.24	0.12
Val	GTC	15.01	0.25	Ile	ATC	22.47	0.52	Phe	TTC	20.72	0.57	Leu	CTC	19.14	0.20
Ala	GCG	7.27	0.10	Thr	ACG	6.80	0.12	Ser	TCG	4.38	0.06	Pro	CCG	7.02	0.11
Ala	GCA	15.50	0.22	Thr	ACA	15.04	0.27	Ser	TCA	10.96	0.15	Pro	CCA	17.11	0.27
Ala	GCT	20.23	0.28	Thr	ACT	13.24	0.23	Ser	TCT	13.51	0.18	Pro	CCT	18.03	0.29
Ala	GCC	28.43	0.40	Thr	ACC	21.52	0.38	Ser	TCC	17.37	0.23	Pro	CCC	20.51	0.33

Observed frequency in exons/1000

Relative frequency
between synonymous codons

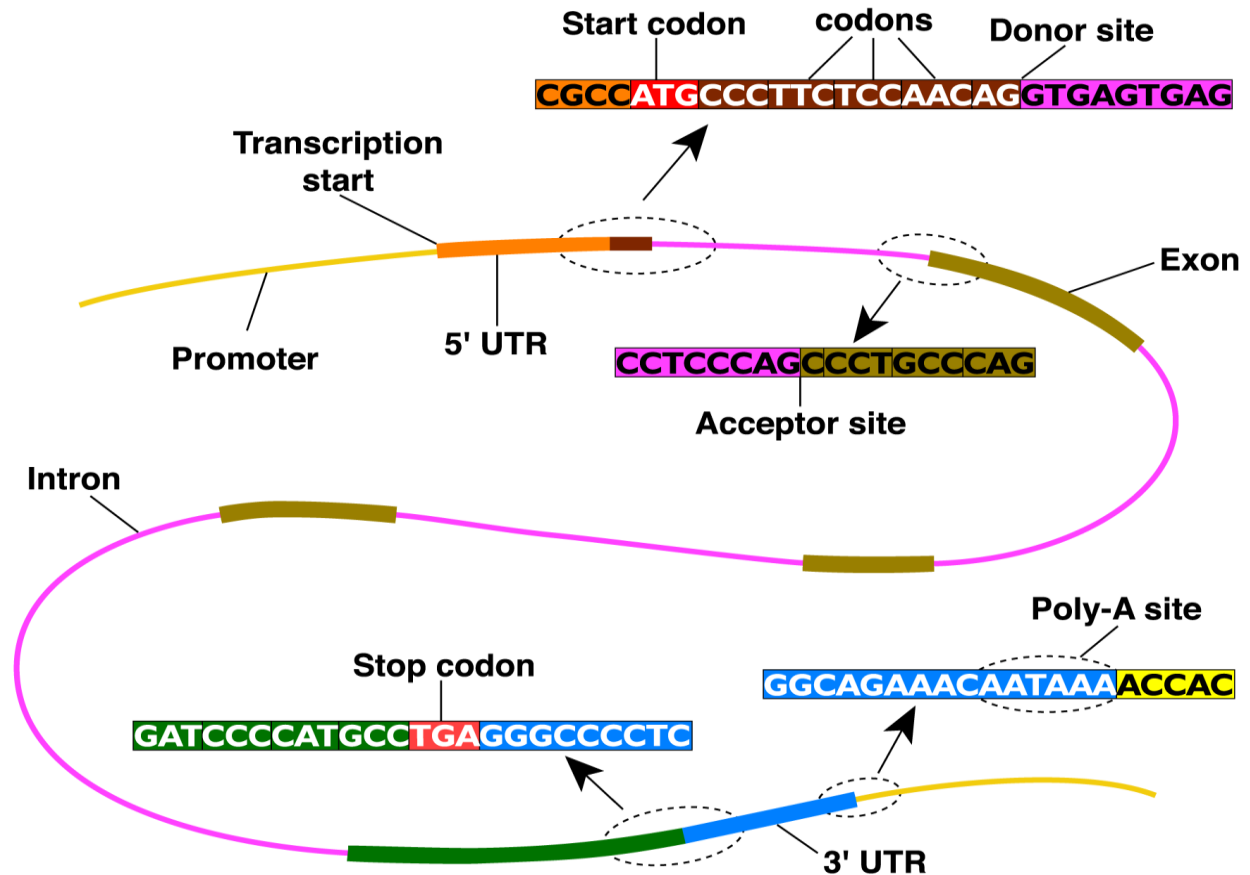
Codon Usage

- For each codon compute the log-likelihood ratio
- $LL(\text{Codon}) = \log (f(\text{Codon}) / f(\text{Null model}))$
- $F(\text{Null model})$ is simply the probability of selecting the codon by chance i.e. $1/64$
- A positive log-likelihood indicates a higher than background chance that the codon is part of a coding sequence

Signal Sensors

- Signal – a sequence feature in the DNA recognized by the transcription/translation processes in the cell:

Signals in a typical eukaryotic gene



Read up on all these regions in your textbooks!

Signal Sensors

- Various pattern recognition methods have been used for detecting signals:
 - Consensus sequences
 - Weight matrices
 - Decision trees
 - Hidden Markov Models (HMMs)
 - Neural networks
 - Support Vector Machines

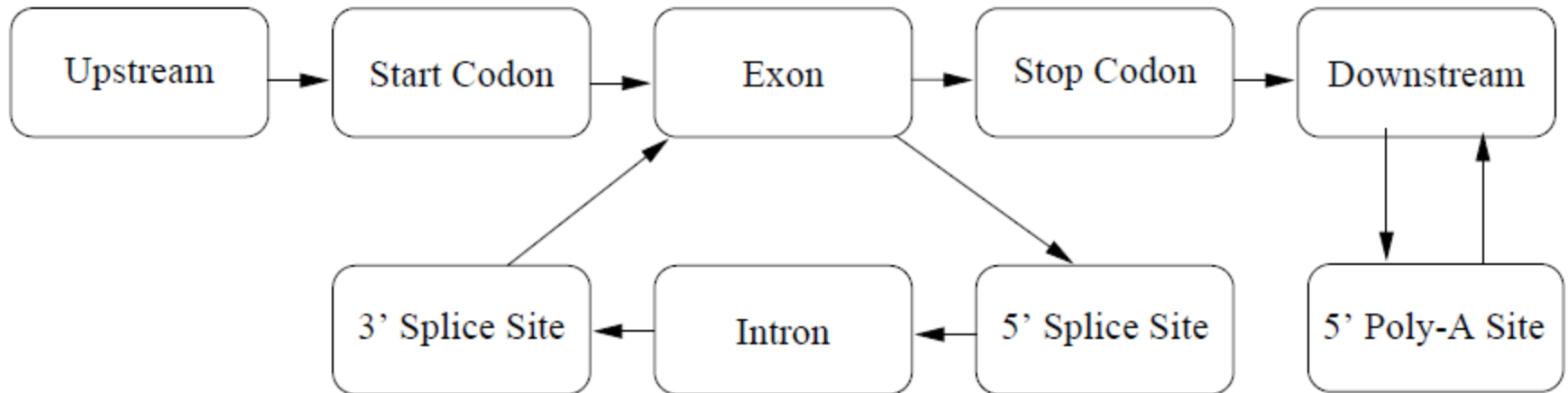
VEIL – a HMM based system

**Henderson, J., Salzberg, S., Fasman, K.H. 1997.
‘Finding genes in DNA with a Hidden Markov
Model’. *J Comput Biol.* 1997 Summer;4(2):127-41.**

Basic idea:

- Design a number of separate HMM models that capture properties of various regions (e.g. intron, exon)
- Train them separately.
- Various sub models can be replaced with new sub models which allows to experiment with alternate models.

Combined Model



Henderson, J., Salzberg, S., Fasman, K.H. 1997. 'Finding genes in DNA with a Hidden Markov Model'. J Comput Biol. 1997 Summer;4(2):127-41.

Exon Model

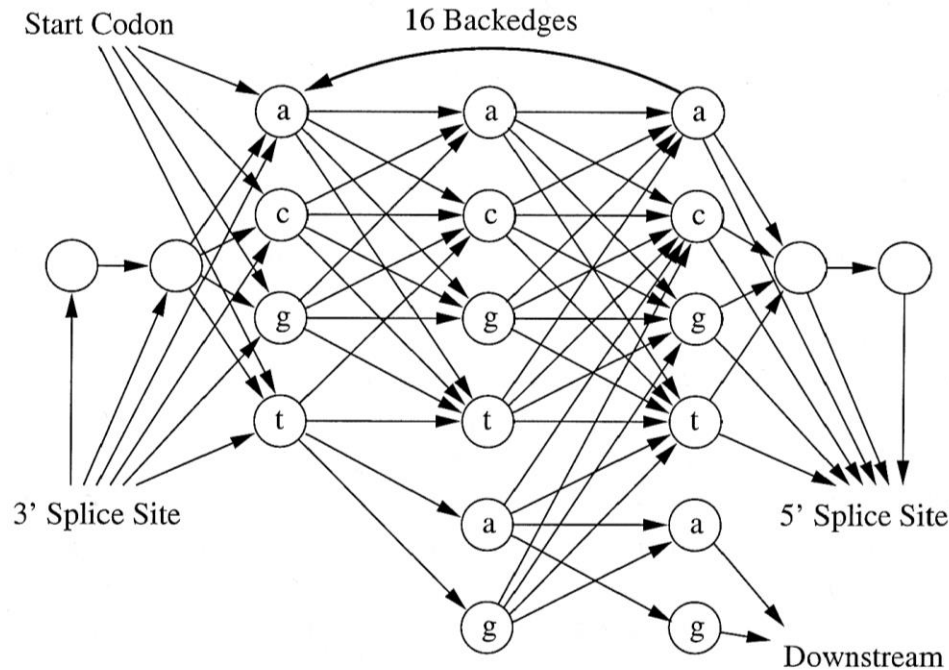


Figure 1: The exon and stop codon models in VEIL. This model can be entered in two ways: either just after outputting a start codon, or upon leaving the 3' splice site model, which follows the intron model. The three central columns of states correspond to the three codon positions. Each of these 12 states is labeled with the base that it can output. The system outputs bases three at a time, looping back after each codon. Note that the paths corresponding to a stop codon (TAA, TAG, and TGA) all force the system to exit from the model (four states at lower right of figure). Alternatively, the system can exit through the 5' splice site, in which case an intron must follow the exon. The two blank states on either end of the model can output any base; these “absorbing states” allow the model to align itself to the proper reading frame, as splice junctions need not respect codon boundaries.

Intron Model

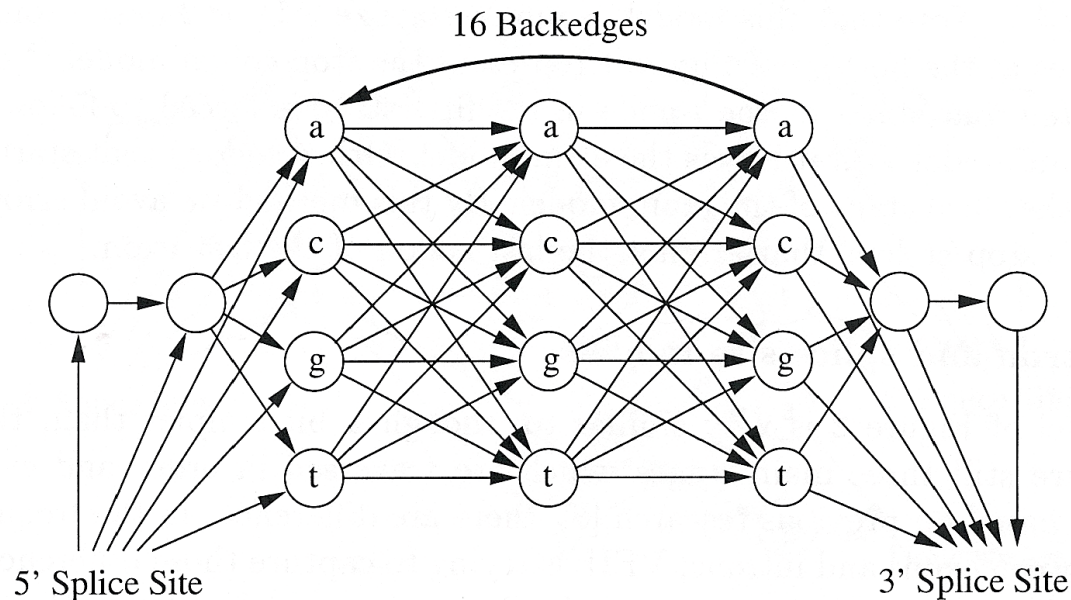
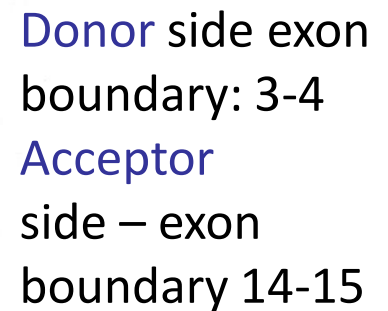


Figure 2: The intron model in VEIL. This model functions similarly to the exon model, with a few important differences. It reads bases three at a time in order to capture differences in the frequency of codon usage between coding and noncoding regions. Unlike the exon model, stop codons do not lead out of this model. The intron model must be entered and exited via splice junctions, which enforces the constraint that exons must appear on both sides of each intron.

There are statistics for census for donor and acceptor sides length (9 and 15 respectively). HMM corresponds to the consensus profile



Look up intron splicing in your textbooks. Most common intron donor/acceptor motif is AG ... GT.

Rest of the story...

- Training
 - Use known gene data set
 - Estimate transition probabilities in model by Expectation Maximization (EM)
- Parsing of new sequence
 - Viterbi algorithm
 - Align new sequence to model
 - Most probable path through states gives most likely parse of sequence

Ab initio Gene Finding is a Hard Problem™

- Genes are separated by large intergenic regions
- Genes are not continuous, but split in a number of (small) coding exons, separated by (larger) non-coding introns
 - in humans coding sequence comprise only a few percent of the genome and an average of 5% of each gene
- Sequence signals that are essential for elucidation of a gene structure are degenerate and highly unspecific
- Alternative splicing
- Repeat elements (>50% in humans) – some contain coding regions

Ab initio Gene Finding is being used less

- Now that we have so many sequenced and annotated genomes, there is now far less need for entirely *ab initio* gene finding methods
- Comparative methods now dominate when sequencing known organisms
- In metagenomics, however, *ab initio* techniques can still be essential tools because we initially don't even know what organisms the DNA sequences come from!
- Even here, though, methods which compare to protein sequence databanks can be very effective

Post-Genomic Analysis

- Assembly of genomes
 - Computationally expensive
 - Lack of data
- Gene finding
 - Computationally expensive
 - Weak algorithms
- Annotation (the rest of the course)
 - Difficulty in keeping-up with data flow
 - Difficulty in assessing reliability

Homework

- Critically read two papers (see Moodle):
 1. Pedersen, A.G., Nielsen, H. 1997. 'Neural network prediction of translation initiation sites in eukaryotes: perspectives for EST and genome analysis'. *Proc Int Conf Intell Syst Mol Biol.* 1997;5:226-33. PMID: 9322041.
 - Paper on prediction of translation initiation sites using neural nets
 2. Henderson, J., Salzberg, S., Fasman, K.H. 1997. 'Finding genes in DNA with a Hidden Markov Model'. *J Comput Biol.* 1997 Summer;4(2):127-41. PMID: 9228612.
 - Paper describing the VEIL gene-finding method
- Extra reading on current “state of the art”:
 - Goodswen SJ, Kennedy PJ, Ellis JT (2012) Evaluating High-Throughput *Ab Initio* Gene Finders to Discover Proteins Encoded in Eukaryotic Pathogen Genomes Missed by Laboratory Techniques. PLoS ONE 7(11): e50609.
<http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0050609>
 - Levy Karin, E., Mirdita, M. & Söding, J. MetaEuk (2020) Sensitive, high-throughput gene discovery, and annotation for large-scale eukaryotic metagenomics. Microbiome 8, 48.
<https://microbiomejournal.biomedcentral.com/articles/10.1186/s40168-020-00808-x>