Genome Informatics 2016 Module Overview

L1-5 Gos Micklem (CCBI, CSBC, Genetics) gm263@cam.ac.uk
Genomes; sequencing; sequence alignment; sequence assembly

L6-12 Alastair Crisp (Chem. Eng) eadc2@cam.ac.uk
Genome structure; genome annotation; sequence variation and consequences

L13-14 Myrto Kostadima (EBI) kostadim@ebi.ac.uk
Gene regulation

L15 Chris Wallace (CIMR): cew54@medschl.cam.ac.uk
GWAS; Hi-seq

L16: Review Session: Friday 25th November

1

All lectures in MR15
All labs in MR16 "CATAM Room" in the basement of Pavilion G.

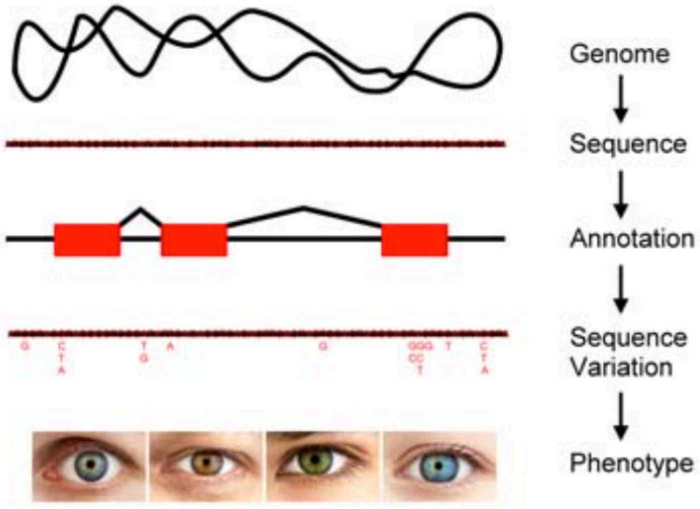
Lecture times unless noted otherwise below: Tuesday 1-2pm
Fridays 12-1pm

Practical sessions: Tuesdays 2-4pm Lecturer in attendance for first hour

Lecture	Lab	Assignments					
L1			Friday	7 October	12-1pm	GM	Sequencing
L2			Friday	7 October	2-3pm	GM	Sequencing
L3			Tuesday	11 October	1-2pm	GM	Alignment
	P1		Tuesday	11 October	2-4pm	GM	DP
L4			Friday	14 October	2-3pm	GM	Assembly
L5			Tuesday	18 October	1-2pm	GM	Genome 1
	P2	A1 SET	Tuesday	18 October	2-4pm	AC	Assembly
L6			Friday	21 October	12-1pm	AC	Genome 2
L7			Tuesday	25 October	1-2pm	AC	Annot
	P3		Tuesday	25 October	2-4pm	AC	ORF
L8		A1 DUE	Friday	28 October	12-1pm	AC	Annot
L9		A2 SET	Tuesday	1 November	1-2pm	AC	Annot
	P4		Tuesday	1 November	2-4pm	AC	InterMine etc
L10			Friday	4 November	12-1pm	AC	Comparative
L11			Tuesday	8 November	1-2pm	AC	Variation 1
	P5		Tuesday	8 November	2-4pm	AC	SNPs
L12			Friday	11 November	12-1pm	AC	Variation 2
			Tuesday	15 November	No lecture		
	P6	A2 DUE	Tuesday	15 November	2-4pm	AC	A2 presentations
L13			Thursday	17 November	10-11am	MK	Regulation 1
L14			Friday	18 November	11.30-12.30	MK	Regulation 2
L15			Tuesday	22 November	1-2pm	CW	GWAS + Hi-C
	P7		Tuesday	22 November	2-5pm	MK	Motif-finding
L16		A3 SET	Friday	25 November	12-1pm		
		A3 DUE	Friday	9 December			

2

What is Genome Informatics?



Genome

↓

Sequence

↓

Annotation

↓

Sequence Variation

↓

Phenotype

Why sequence genomes?

To aid molecular investigation of a species

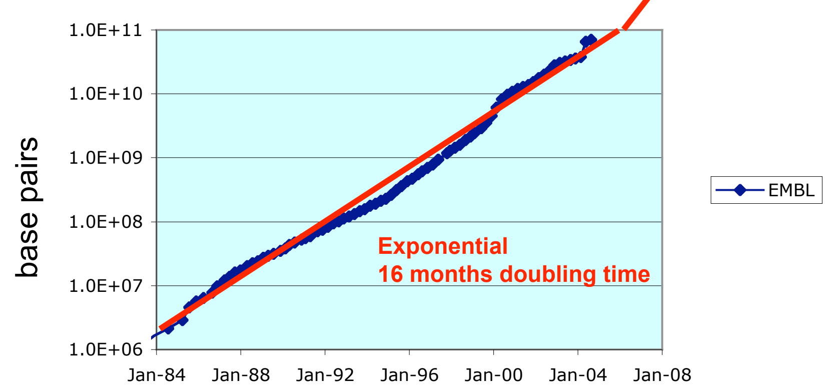
To discover the sequence variations in an individual

To help find the molecular lesions underlying disease

To aid in comparison of e.g. pathogenic vs non-pathogenic bacterial strains

To discover/ survey the organisms in a location ('metagenomics')

Growth of EMBL DNA sequence repository



DNA sequencing has recently become 1000 times faster and cheaper

Figure from Richard Durbin (WTSI)

5

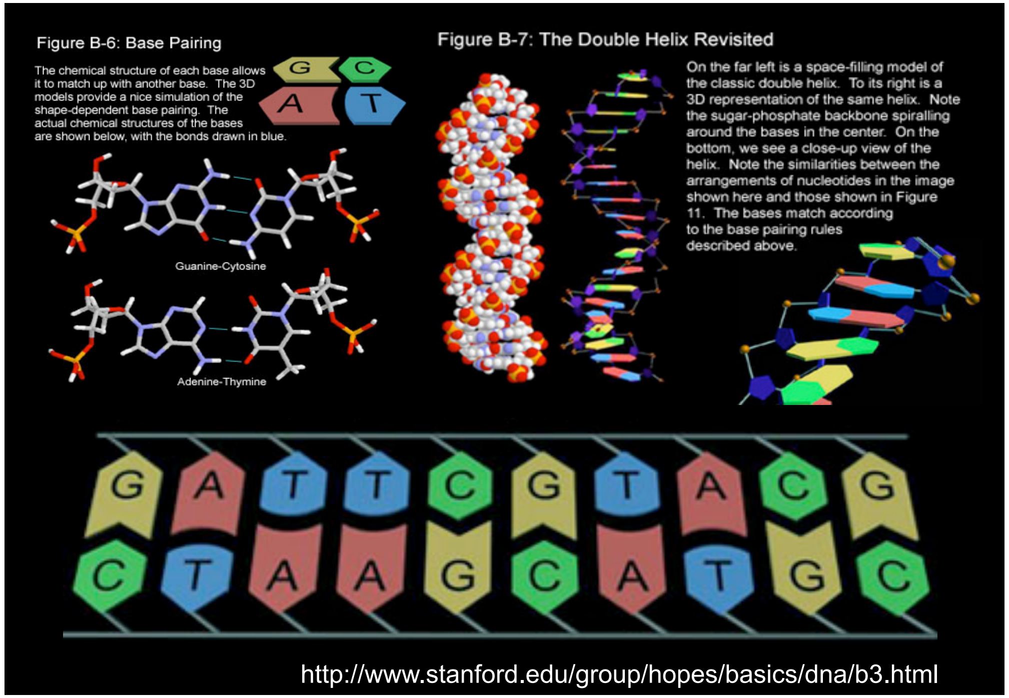


Figure B-6: Base Pairing

The chemical structure of each base allows it to match up with another base. The 3D models provide a nice simulation of the shape-dependent base pairing. The actual chemical structures of the bases are shown below, with the bonds drawn in blue.

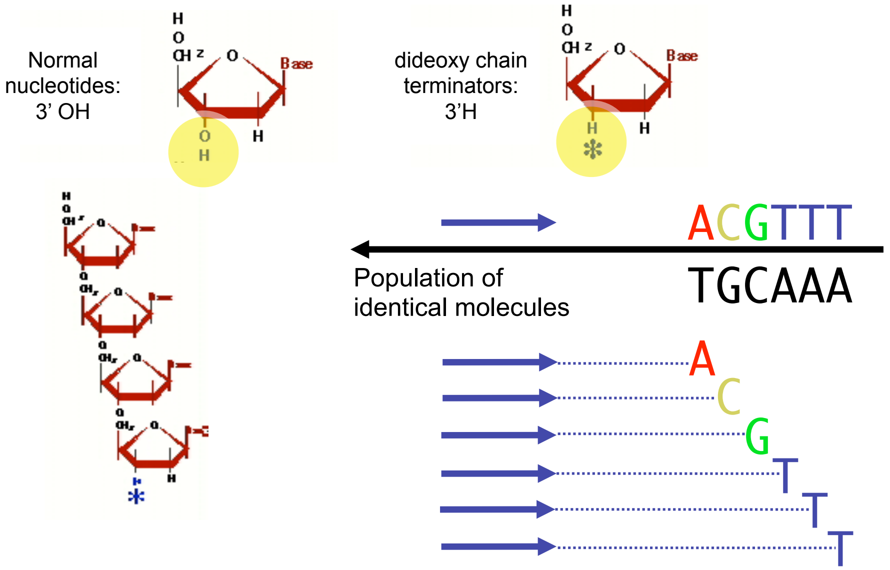
Figure B-7: The Double Helix Revisited

On the far left is a space-filling model of the classic double helix. To its right is a 3D representation of the same helix. Note the sugar-phosphate backbone spiralling around the bases in the center. On the bottom, we see a close-up view of the helix. Note the similarities between the arrangements of nucleotides in the image shown here and those shown in Figure 11. The bases match according to the base pairing rules described above.

http://www.stanford.edu/group/hopes/basics/dna/b3.html

6

Sanger dideoxy sequencing

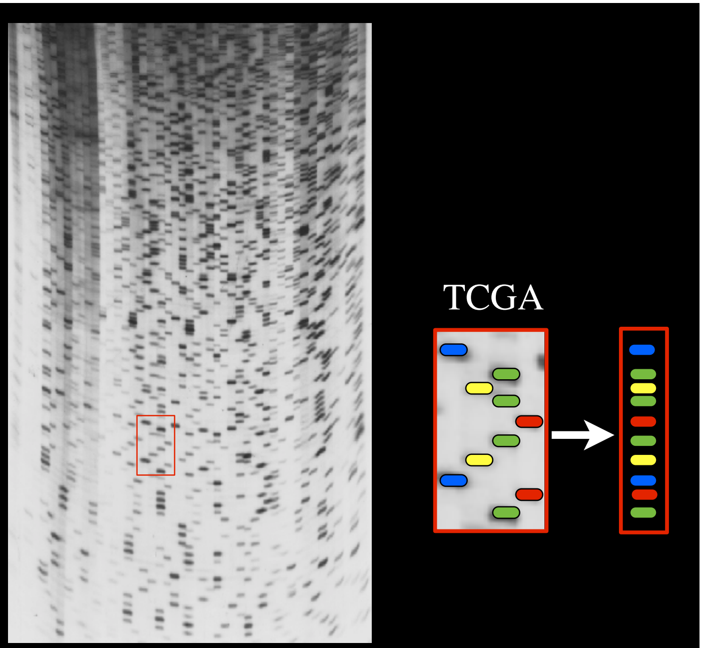


Normal nucleotides: 3' OH

dideoxy chain terminators: 3'H

Population of identical molecules

7



TCGA

http://homepages.inf.ed.ac.uk/rbf/HIPR2/images/dna1.gif

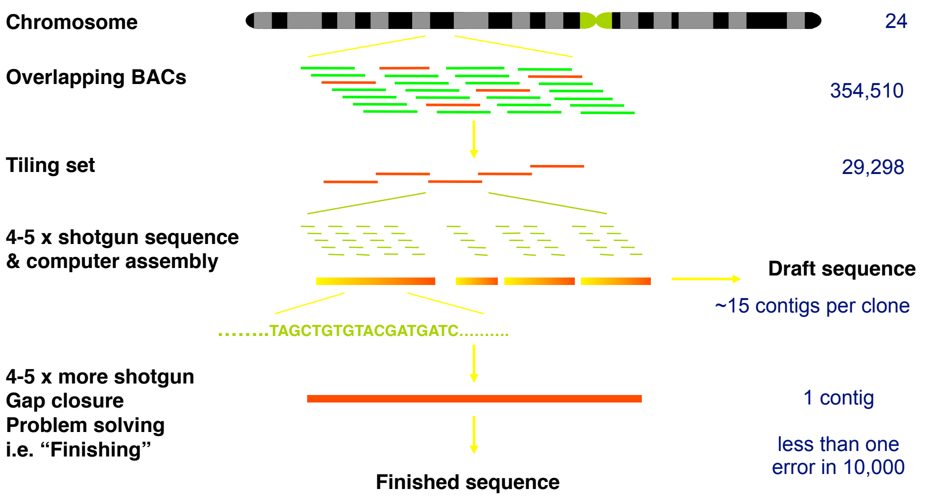
8

ABI 3700 - No more gel plates…

Genome Sequencing

Basic problem: how does one determine a genome sequence of say ~10^9 bases when can only read ~500 bases at a time?

Human Genome Project sequencing strategy



Chromosome 24

Overlapping BACs 354,510

Tiling set 29,298

4-5 x shotgun sequence & computer assembly

Draft sequence

~15 contigs per clone

........TAGCTGTGTACGATGATC..........

4-5 x more shotgun
Gap closure
Problem solving
i.e. "Finishing"

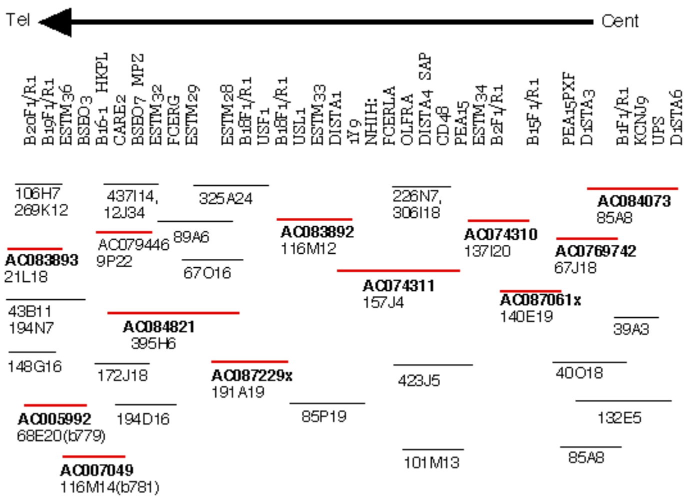
1 contig

less than one error in 10,000

Finished sequence

BAC contig that covers the QTL locus on Mouse chromosome 1.

1-17-01



Contig mapped by Weikuan Gu, at the JLP VA Medical Center, Loma LInda, California and being sequenced that the ACGT, University of Oklahoma, Bruce Roe's laboratory

Whole Genome Shotgun

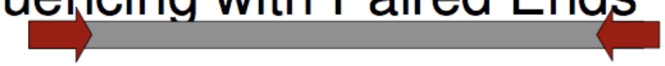
Issues:

- Cloning bias
- Assembly - potential for HUGE mistakes
  - repeats
  - computationally hard

But you don't have to wait for mapping...

13

Sequencing with Paired Ends



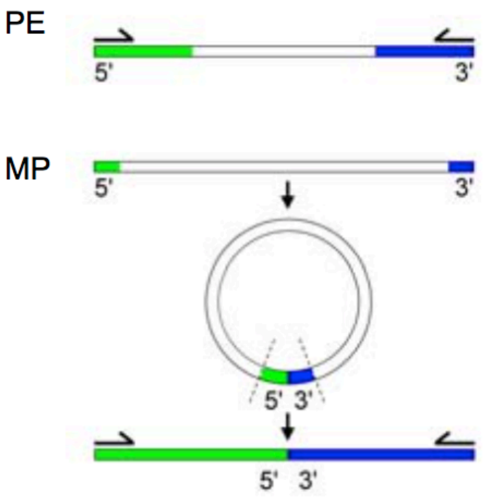
Reference	This is really the best way to do sequencing
Single-reads	This is
…	is really
…	really the
…	the best
…	sequencing
Paired-reads	This is (------26 characters-------) sequencing

Assembly becomes easier

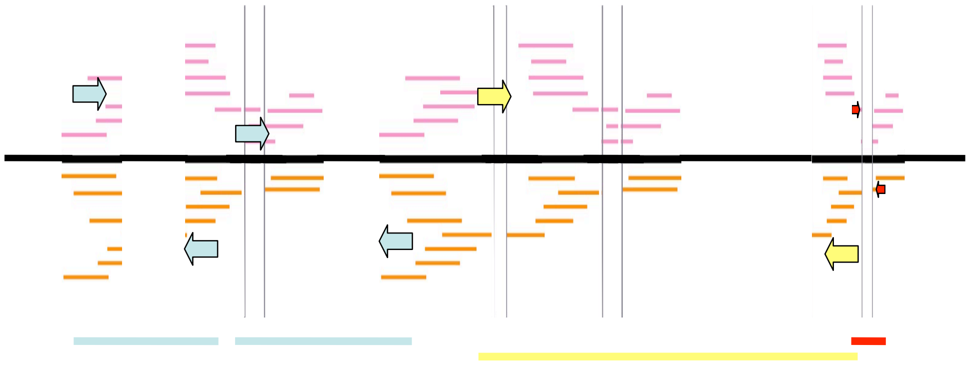
Illumina product literature

14

Paired Ends and Mate-pairs

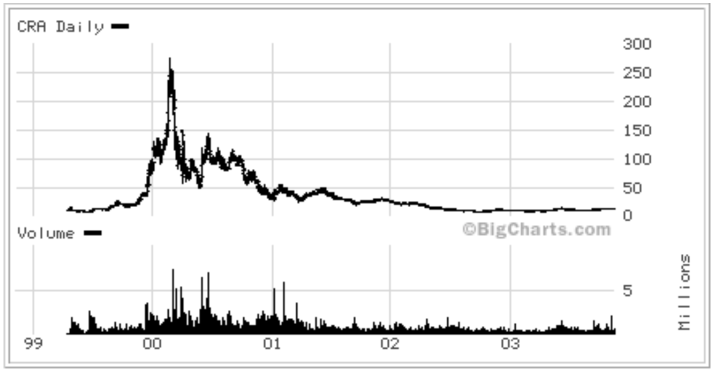


Whole Genome Shotgun



Relies on different libraries, carefully sized larger than repeats. Was controversial.

Celera share price



General Background Reading

Genomes 3 (College Libraries)
Terry Brown
ISBN 978-0815341383

Background on DNA structure:
Chapter 1 until page 12

Chapter 4 - Genome sequencing

Chapter 5 - Understanding a genome sequence: (parts that deal with computational rather than experimental approaches)

Chapter 6 - Understanding how a genome functions (not the sections on proteome, metabolome, wet lab experimental methods)

Chapter 7 - Eukaryotic Nuclear Genomes