High-Throughput Sequencing and its applications

IN-BIOS5000/9000
Genome Sequencing Technologies, Assembly, Variant Calling and Statistical Genomics
22 October 2018

Torbjørn Rognes
Dept. of Informatics, UiO & Dept. of Microbiology, OUS
torognes@ifi.uio.no

UiO : University of Oslo

Oslo University Hospital

Overview

- Sequencing technologies & their general principles
- Important properties & sequencing development
- Sequence quality & file format
- Paired-end reads & mate pair sequencing
- Applications & basic bioinformatics tools
- Whole genome assembly
- Resequencing & read mapping
- Challenges

Genome sequencing

High-Throughput Sequencing (HTS), Deep sequencing, Next Generation Sequencing (NGS)

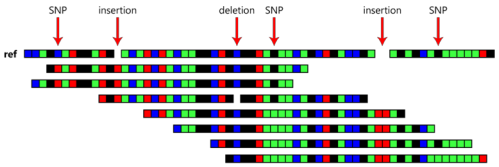
Obtain sample

DNA extraction, preparation

Sequencing

Raw data

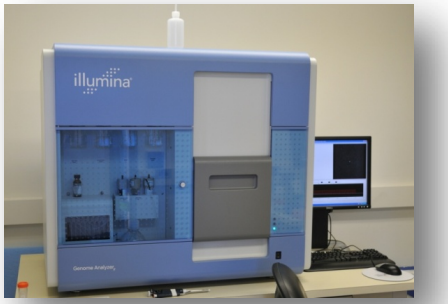
Sequence analysis



Interpretation

Illumina

- Sequencing by synthesis using fluorescence
- One fragment = one cluster = one read
- Read lengths up to 250bp, paired-end reads
- Dominant technology today
- Formerly known as Solexa
- HiSeq 2500 specifications:
  - Can sequence entire human genome in 27 hours at 30X coverage (2x100bp)
  - Up to 2x150 bp
  - Run time 7 hours to 11 days
  - Up to 6 billion 100bp reads in 11 days



GA IIx

HiSeq 2500

MiSeq

NovaSeq 6000

Sanger sequencing center

Other sequencing systems

Roche (454)

ABI (SOLiD)

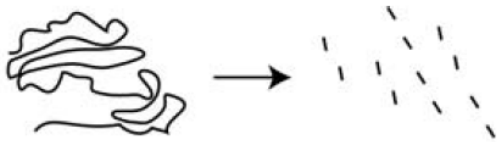
Ion Torrent

Pacific Biosciences SMRT and Sequel systems

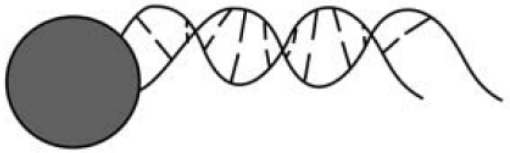
Oxford Nanopore

General HTS principles

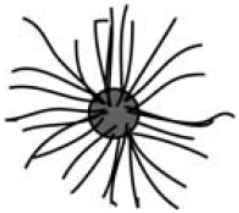
1) Randomly fragment many molecules of target DNA



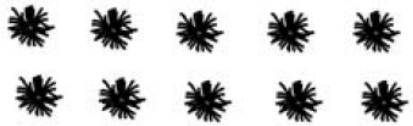
2) Immoblize individual DNA molecules on solid support



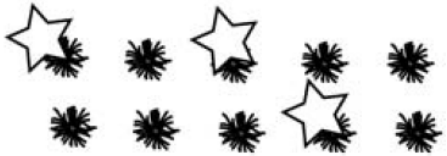
3) Amplify DNA in clonal 'polymerase colony'



4) Sequence DNA by adding liquid reagents to immoblized DNA colonies



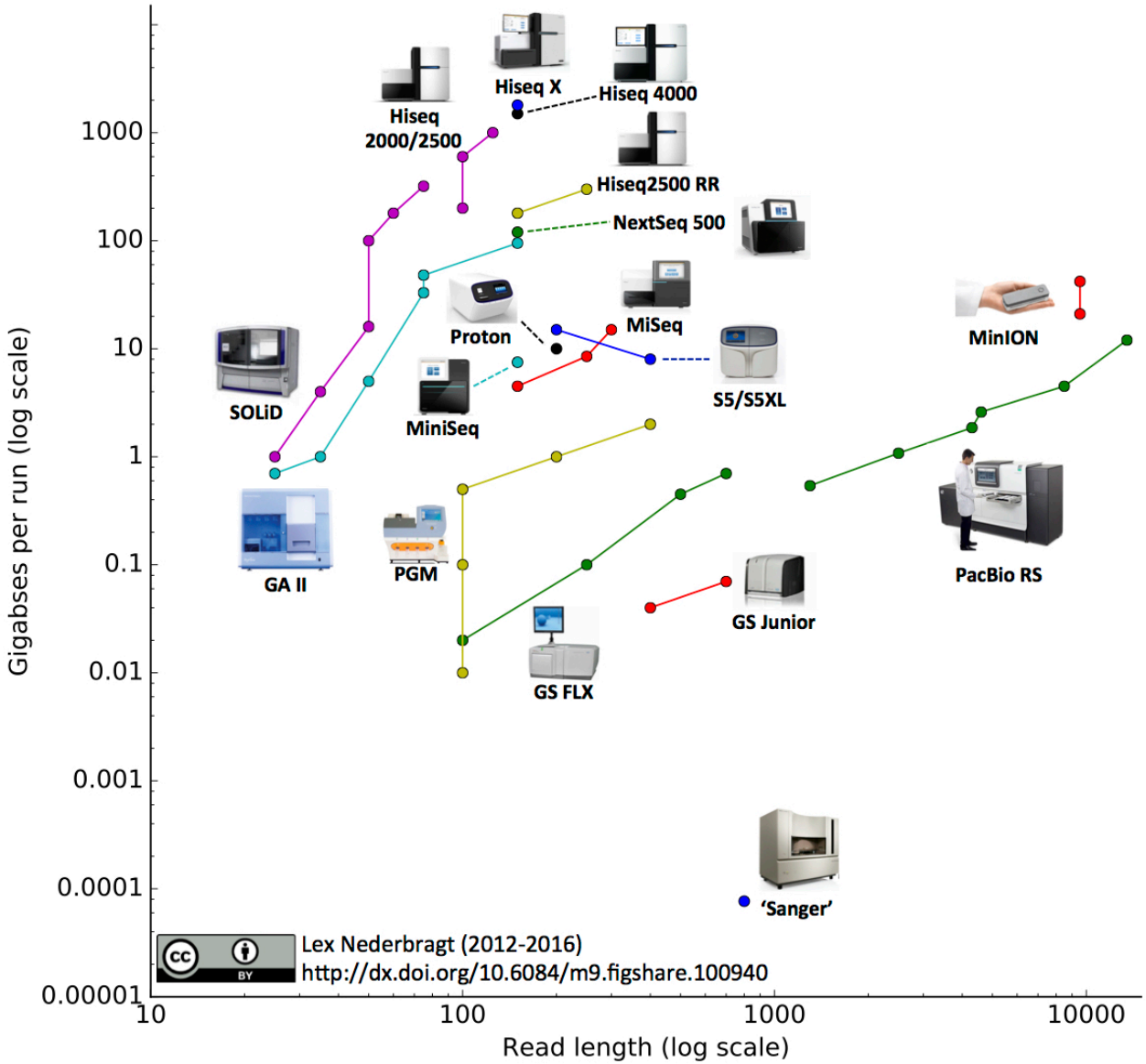
5) Interrogate sequence incorporation *in situ* after each cycle using fluorescence scanning or chemiluminescence



Important technology properties

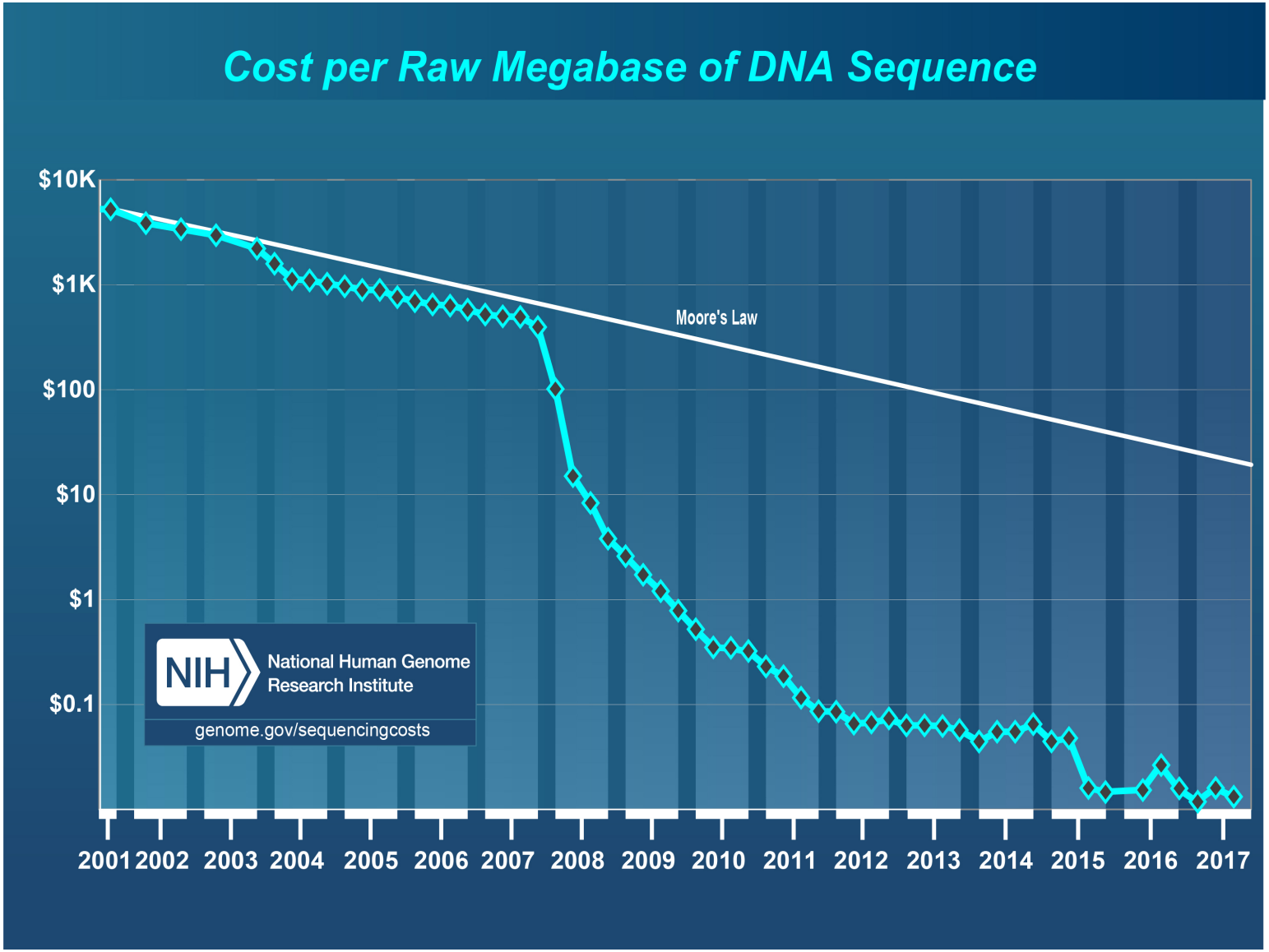
- Cost
  - Per base
  - Investment
- Read length
- Paired-end support
- Errors
  - Frequency
  - Profile (indels, substitutions)
  - Random or systematic?
- Speed or capacity (bases per day)
- PCR-based?
  - Single molecule
  - PCR amplification step
- Amount of lab work necessary

Sequencing technology development



Source: Lex Nederbragt (2012-2016) https://doi.org/10.6084/m9.figshare.100940

The cost of sequencing



Sequencing technology properties

Box 1 | **Sequencing and mapping technologies**

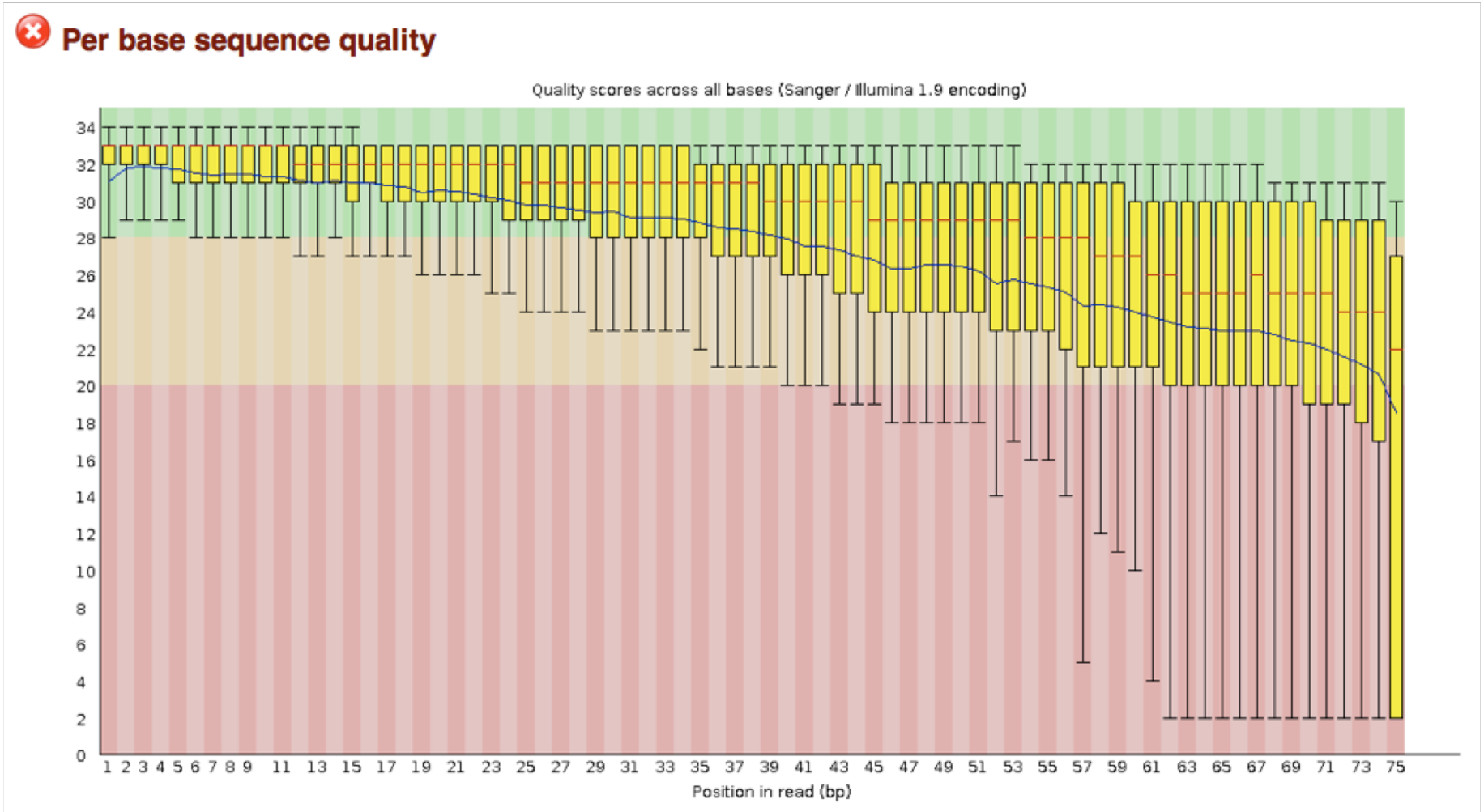
A full survey of sequencing technologies is beyond the scope of this article. The table below compares the currently available technologies in terms of several characteristics of importance for genome assembly: read length, error rate and the ability to generate paired-end reads natively. Note that potentially all sequencing technologies can be used to sequence mate-pair libraries obtained by the circularization of long DNA fragments[91]. Furthermore, long-range linking information can be obtained from genome-mapping technologies, such as optical mapping[92].

For most technologies, read lengths and error rates depend on the specific characteristics of the sequencing experiment. The values provided in the table are those that are encountered in typical recent projects.

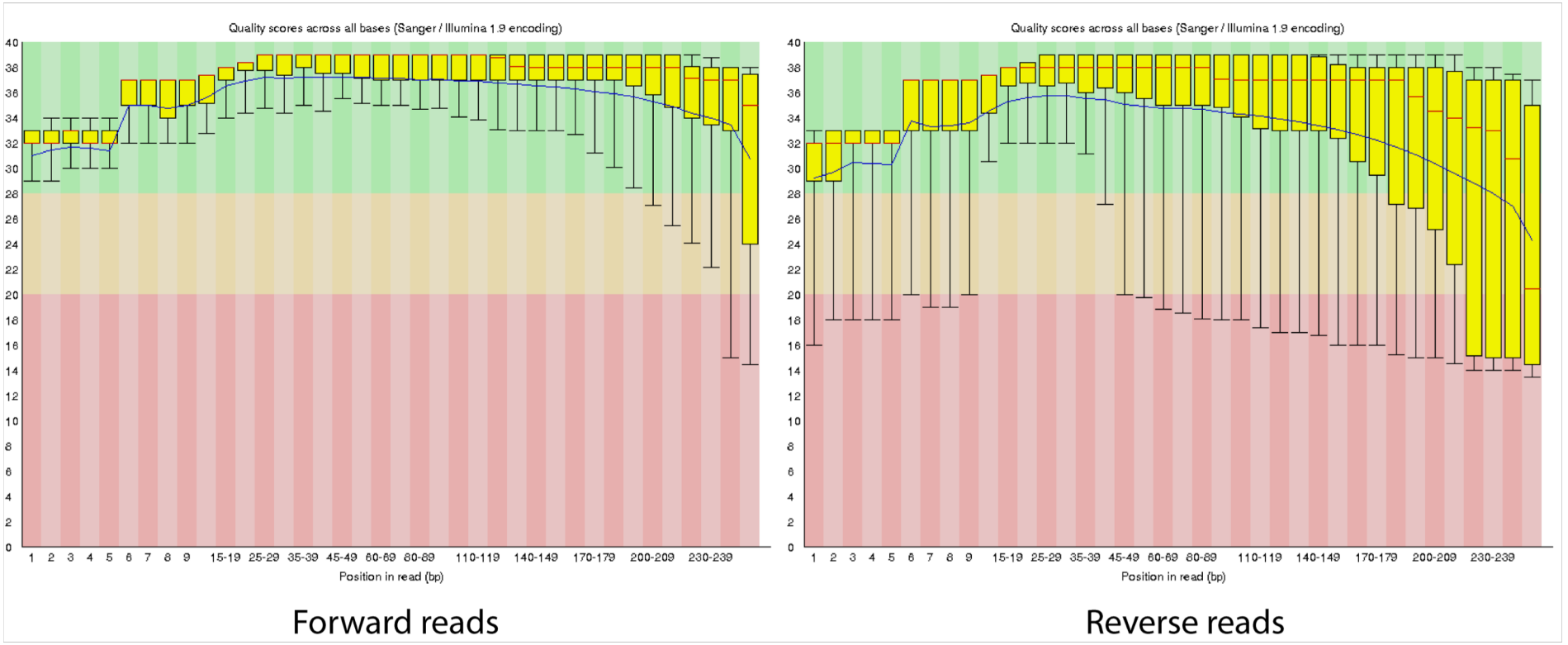
Technology	Read length (bp)	Error rate	Native paired-end read support	Refs
ABI/Solid	75	Low (~2%)	Yes	93
Illumina/Solexa	100–150	Low (<2%)	Yes	94
IonTorrent	~200	Medium (~4%)*	No	94
Roche/454	400–600	Medium (~4%)*	No	94
Sanger	Up to ~2,000 bp	Low (~2%)	Yes	
Pacific Biosciences	Up to ~15,000‡	High (~18%)	Yes (in strobe read mode)	39

*454 and Ion Torrent technologies are prone to errors in homopolymer regions, which are segments of the genome in which the same nucleotide is repeated multiple times[94]. ‡Pacific Biosciences instruments produce reads with an exponential distribution of read lengths, only a few of which reach the multi-kb range[10,11].

FASTQC - Quality plot of Illumina reads



Quality plots of Illumina MiSeq reads



The FASTQ format

- A sequence file format in plain text that includes quality scores for each nucleotide in the sequence
- Example:

```
@SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36
GGGTGATGGCCGCTGCCGATGGCGTCAAATCCCACC
+SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36
IIIIIIIIIIIIIIIIIIIIIIIIIIIII9IG9IC
```

- The first line starts with a '@' symbol followed by an identifier before the first space.
- The second line contains the actual sequence.
- The third line starts with a '+' symbol, optionally followed by the same identifier as the first line. Identifier rarely used.
- The fourth line contains characters that represent the quality scores for each nucleotide
- In principle, sequence and quality may span multiple lines, but rarely do.

FASTQ quality scores

- Quality value Q (or Phred quality score):
  $Q = -10 \log_{10} p$
  where p = error probability

- The Q values are encoded as ascii characters by adding 33 to the value after rounding to nearest integer:
  $c = 33 + \text{round}(Q)$

- Only Q values 0-93 used (often just 0-41), corresponding to characters 33-126, and to p values from 1 to $5 \cdot 10^{-10}$

- Example:
  p=0.0001 (high quality)
  $Q = -10 \log_{10} 0.0001 = -10 * -4 = 40$
  $c = 33 + Q = 33 + 40 = 73 =$ 'I'

- Older versions of the format (before 2011) differed slightly

From quality characters to p-values

$p = 10^{-(c-33)/10}$

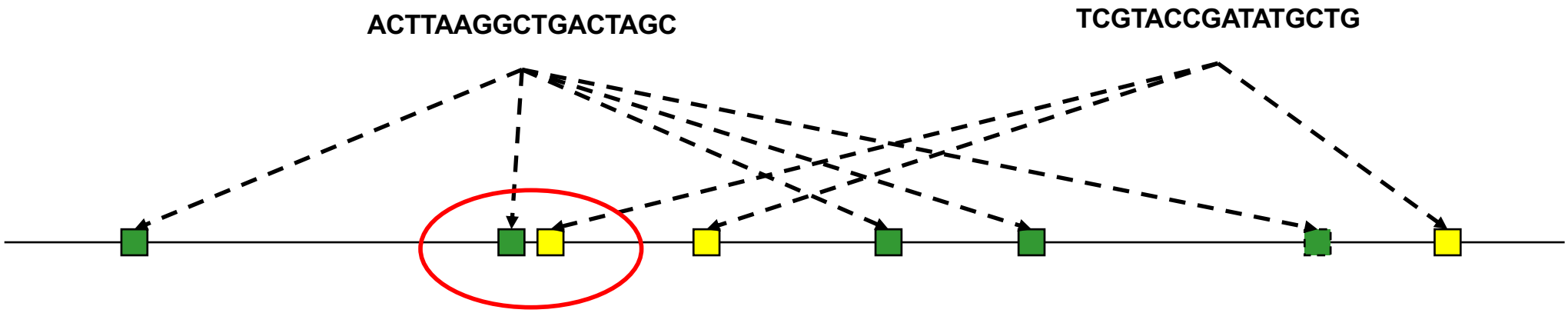
Example:

```
IIIIIIIIIIIIIIIIIIIIIIIIIIIIII9IG9IC
```

Char	ASCII	Q	p
I	73	40	0.0001
9	57	24	0.0040
G	71	38	0.00016
C	67	34	0.00040

0	<NUL>	32	<SPC>	64	@	96	`
1	<SOH>	33	!	65	A	97	a
2	<STX>	34	"	66	B	98	b
3	<ETX>	35	#	67	C	99	c
4	<EOT>	36	$	68	D	100	d
5	<ENQ>	37	%	69	E	101	e
6	<ACK>	38	&	70	F	102	f
7	<BEL>	39	'	71	G	103	g
8	<BS>	40	(	72	H	104	h
9	<TAB>	41	)	73	I	105	i
10	<LF>	42	*	74	J	106	j
11	<VT>	43	+	75	K	107	k
12	<FF>	44	,	76	L	108	l
13	<CR>	45	-	77	M	109	m
14	<SO>	46	.	78	N	110	n
15	<SI>	47	/	79	O	111	o
16	<DLE>	48	0	80	P	112	p
17	<DC1>	49	1	81	Q	113	q
18	<DC2>	50	2	82	R	114	r
19	<DC3>	51	3	83	S	115	s
20	<DC4>	52	4	84	T	116	t
21	<NAK>	53	5	85	U	117	u
22	<SYN	54	6	86	V	118	v
23	<ETB>	55	7	87	W	119	w
24	<CAN>	56	8	88	X	120	x
25	<EM>	57	9	89	Y	121	y
26	<SUB>	58	:	90	Z	122	z
27	<ESC>	59	;	91	[	123	{
28	<FS>	60	<	92	\	124	|
29	<GS>	61	=	93	]	125	}
30	<RS>	62	>	94	^	126	~
31	<US>	63	?	95	_	127	<DEL>

Multiple mapping



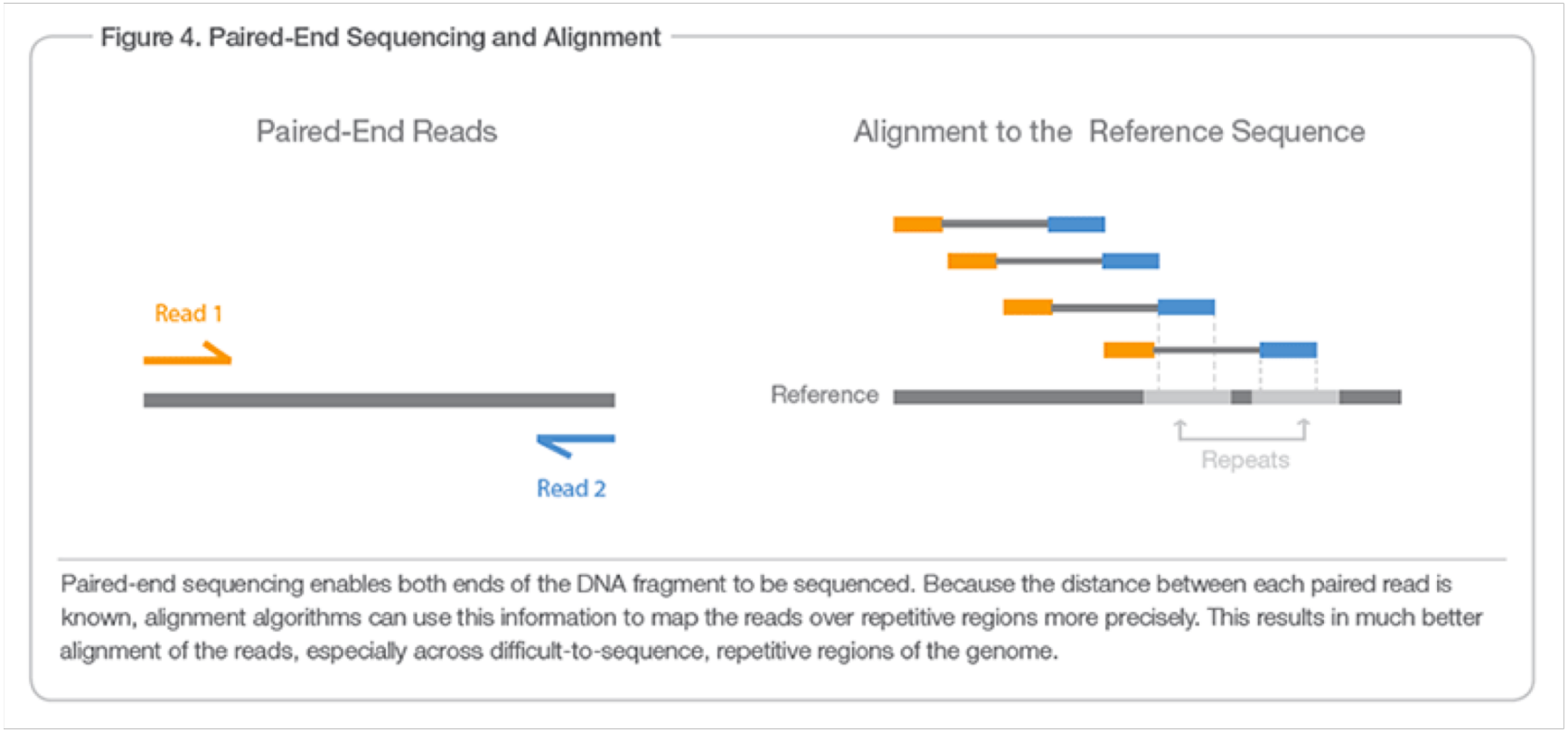
- Problem:
  - Short reads (e.g. 100bp) may not map uniquely due to long repeats (>100bp) in the genome
- Solutions:
  - Get longer reads
  - Get paired reads (pairs of reads with fixed distance)

Paired-end / mate pair sequencing

- Paired-end reads or mate pair reads are pairs of reads known to come from two close regions in the genome.
- They are located with an approximate fixed distance from each other.
- Typically paired ends are a ~100-500bp apart, while mate pairs are ~2-10kb apart
- Allows short reads to have a larger ”effective” size
- Performed by sequencing fragments from both ends
- Often used with Illumina reads
  - Typically 2 x 100 bp separated by 300bp
  - Allpaths-LG requires ~100 bp from fragments ~ 180bp
- May also overlap (e.g. 2x250bp from 400bp fragments)

Paired-end / mate pair sequencing

- Both ends of fragments of fixed length (a few kbp) may be sequenced
- Gives information on genomic distance between pairs of reads
- May be used to overcome some problems with short reads



Source: Illumina

Applications

Category	Examples of applications
Complete genome resequencing	Comprehensive polymorphism and mutation discovery in individual human genomes
Reduced representation sequencing	Large-scale polymorphism discovery
Targeted genomic resequencing	Targeted polymorphism and mutation discovery
Paired end sequencing	Discovery of inherited and acquired structural variation
Metagenomic sequencing	Discovery of infectious and commensal flora
Transcriptome sequencing	Quantification of gene expression and alternative splicing; transcript annotation; discovery of transcribed SNPs or somatic mutations
Small RNA sequencing	microRNA profiling
Sequencing of bisulfite-treated DNA	Determining patterns of cytosine methylation in genomic DNA
Chromatin immunoprecipitation–sequencing (ChIP-Seq)	Genome-wide mapping of protein-DNA interactions
Nuclease fragmentation and sequencing	Nucleosome positioning
Molecular barcoding	Multiplex sequencing of samples from multiple individuals

Basic bioinformatics tools

Mapping: Mapping sequence reads to a known genome sequence, used initially in resequencing procedures.

E.g.: BWA, Bowtie, SOAP2, Maq, BFAST, RMAP, ...

Assembly: Assembling together reads into a complete genome sequence, usually divided into a number of contigs and scaffolds. For sequencing entirely new genomes.

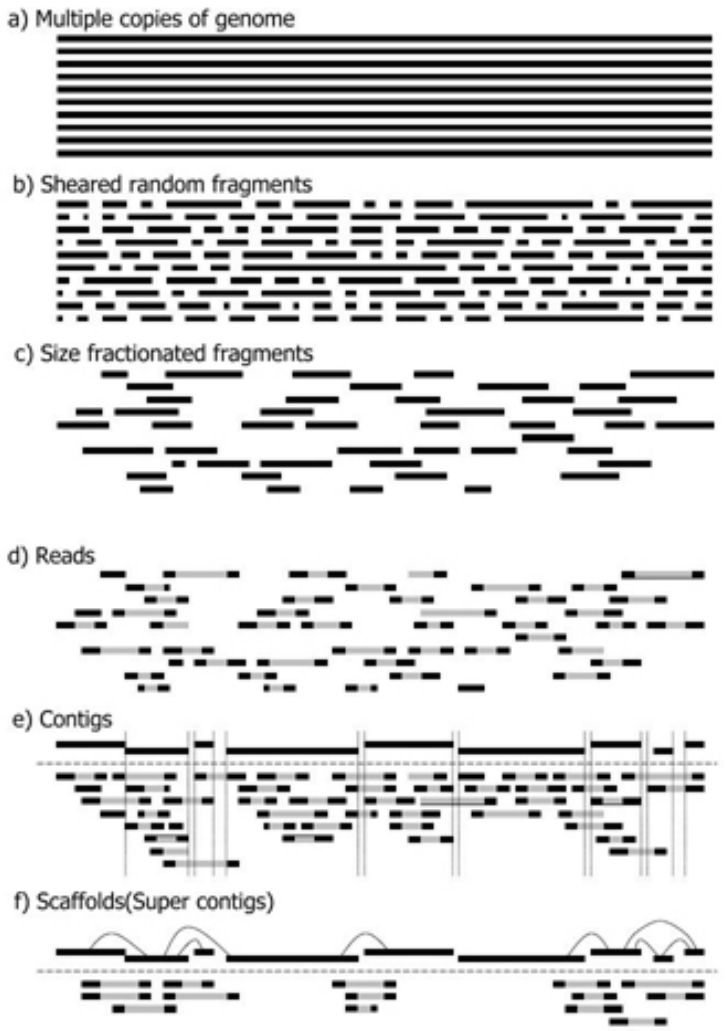
E.g.: Celera/CABOG, Newbler, Phrap, TIGR, Arachne, Velvet, MaSuRCA, SPAdes, ALLPATHS-LG, Abyss, ...

Other:

SNP discovery, Chip-Seq, RNA-Seq, Methyl-Seq, Metagenomics, other mutation/variant discovery

Whole genome *de novo* sequencing

- Whole genome sequencing results in millions of small pieces of the full genome
- The challenge is to puzzle these together in the right order
- Genome sizes ranging from 2Mbp (bacteria) to 3Gbp (human) to 150Gbp (plant)
- Read size from 30 bp to 10000 bp



Definitions

- **Reads** are raw sequences from the sequencers: short continuous sequences

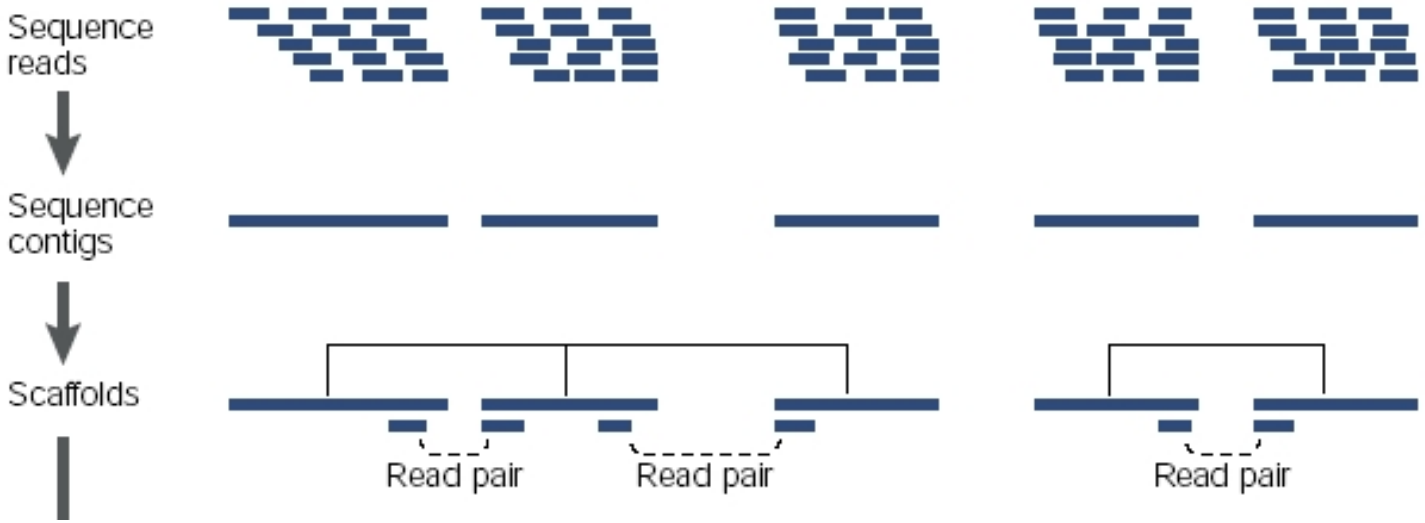
- **Contigs** are longer continuous sequences formed from reads that are partially overlapping. A consensus sequence is based on the reads.

- **Scaffolds** are even longer dis-continuous sequences formed from contigs using information about the distance between contigs and their orientation. Depends upon data from paired-end, mate-pairs or related info

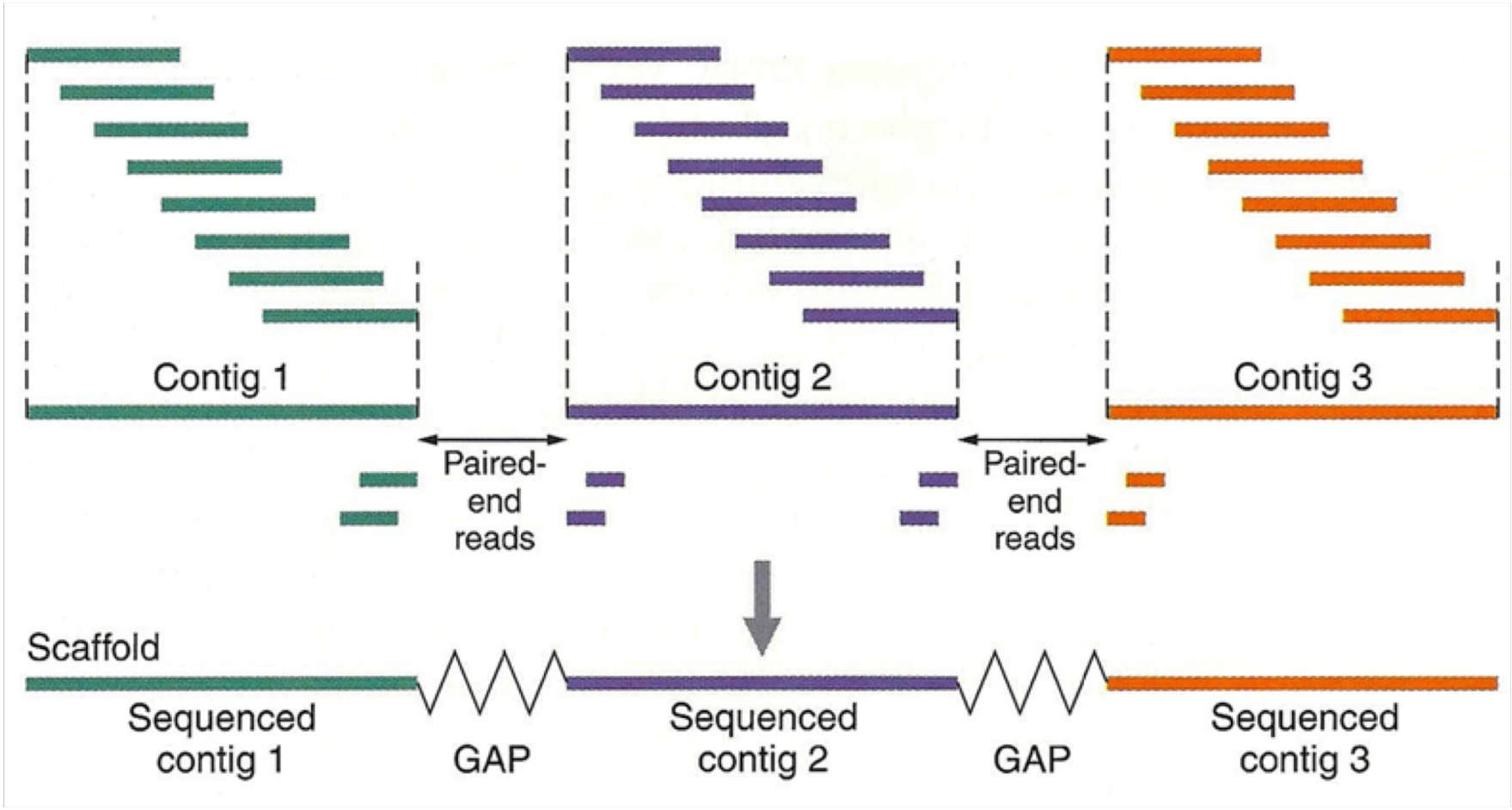
- **An assembly** is the collection of all scaffolds, ideally as few and long and correct as possible

Genome assembly

- Typically whole genome sequencing of novel bacterial species
- Short reads makes eukaryotes hard due to repeats, but not impossible, needs “paired ends”



Paired-end reads span gaps



Source: http://www.discoveryandinnovation.com/BIOL202/notes/lecture25.html

Problematic issues

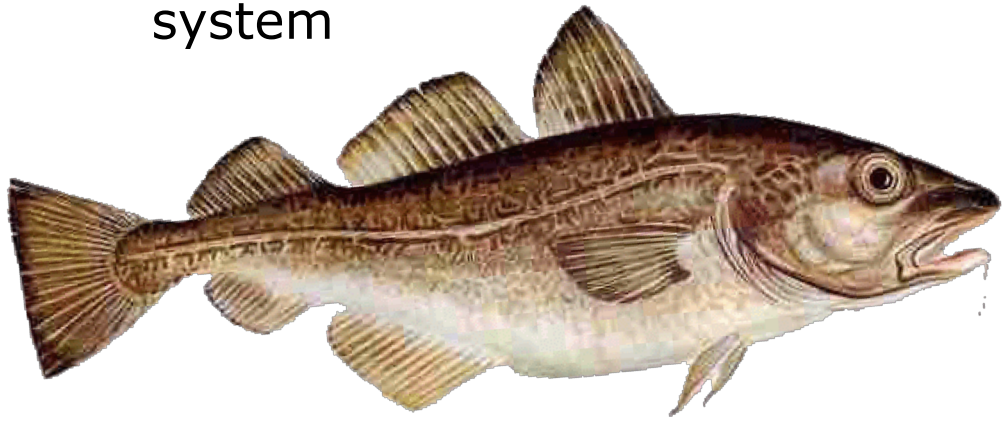
- Sequencing errors
  - Introduces false sequences into the assembly
  - May be alleviated by higher coverage / larger sequencing depth, or by error detection and correction

- Repeats
  - Genomes often contain many almost identical repeated sequences
  - Repeats longer than the read length makes it impossible to determine the exact location of the read
  - May cause compression or misassemblies
  - May be alleviated by longer reads or paired-end/mate pair reads

- Heterozygosity
  - Diploid organisms (e.g Humans) actually have two “genomes”, not one. Chromosome pairs 1-22 for all, plus XX or XY. One set of chromosomes from our mother and one from our father.
  - The two are mostly identical, but there are some differences
  - Causes “bubbles” in the assembly

The cod genome

- Atlantic cod
- Estimated genome size: 830Mbp
- Sequenced at Dept of Biology, UiO with collaborators
- Started with Roche 454 sequencing
- 40X coverage
- *de novo* assembly using Newbler and Celera Assembler
- 22154 genes identified
- Lacks part of common immune system



LETTER

doi:10.1038/nature10342

The genome sequence of Atlantic cod reveals a unique immune system

Bastiaan Star[1], Alexander J. Nederbragt[1], Sissel Jentoft[1], Unni Grimholt[1], Martin Malmstrom[1], Tone F. Gregers[2], Trine B. Rounge[1], Jonas Paulsen[1,3], Monica H. Solbakken[1], Animesh Sharma[4], Ola F. Wetten[5,6], Anders Lanzén[7,8], Roger Winer[9], James Knight[9], Jan-Hinnerk Vogel[10], Bronwen Aken[10], Øivind Andersen[11], Karin Lagesen[1], Ave Tooming-Klunderud[1], Rolf B. Edvardsen[12], Kirubakaran G. Tina[1,13], Mari Espelund[1], Chirag Nepal[4,8], Christopher Previti[8], Bård Ove Karlsen[14], Truls Moum[14], Morten Skage[1], Paul R. Berg[1], Tor Gjøen[15], Heiner Kuhl[16], Jim Thorsen[17], Ketil Malde[12], Richard Reinhardt[16], Lei Du[9], Steinar D. Johansen[14,18], Steve Searle[10], Sigbjørn Lien[13], Frank Nilsen[19], Inge Jonassen[4,8], Stig W. Omholt[1,13], Nils Chr. Stenseth[1] & Kjetill S. Jakobsen[1]

Atlantic cod (*Gadus morhua*) is a large, cold-adapted teleost that sustains long-standing commercial fisheries and incipient aquaculture[1,2]. Here we present the genome sequence of Atlantic cod, showing evidence for complex thermal adaptations in its haemoglobin gene cluster and an unusual immune architecture compared to other sequenced vertebrates. The genome assembly was obtained exclusively by 454 sequencing of shotgun and paired-end libraries, and automated annotation identified 22,154 genes. The major histocompatibility complex (MHC) II is a conserved feature of the adaptive immune system of jawed vertebrates[3,4], but we show that Atlantic cod has lost the genes for MHC II, CD4 and invariant chain (Ii) that are essential for the function of this pathway. Nevertheless, Atlantic cod is not exceptionally susceptible to disease under natural conditions[5]. We find a highly expanded number of MHC I genes and a unique composition of its Toll-like receptor (TLR) families. This indicates how the Atlantic cod immune system has evolved compensatory mechanisms in both adaptive and innate immunity in the absence of MHC II. These observations affect fundamental assumptions about the evolution of the adaptive immune system and its components in vertebrates.

We sequenced the genome of a heterozygous male Atlantic cod (NEAC_001, Supplementary Notes 1 and 2), applying a whole-genome shotgun approach to 40× coverage (estimated genome size of 830 megabases (Mb), Supplementary Note 4 and Supplementary Fig. 2) using 454 technology (Supplementary Note 3). Two programs (Newbler[6] and Celera[7], Supplementary Notes 5 and 6) produced assemblies with short contigs, yet with scaffolds of comparable size to those of Sanger-sequenced teleost genomes (Supplementary Note 10 and Supplementary Fig. 8). Although fragmentation due to short tandem repeats is difficult to address (Supplementary Note 7), we resolved numerous gaps attributable to heterozygosity (Supplementary Note 8). The assemblies differ in scaffold and contig length (Table 1), although their scaffolds align to a large extent (Supplementary Note 9 and Supplementary Fig. 7). We obtained about one million single nucleotide polymorphisms (SNPs) by mapping 454 and Illumina reads from the sequenced individual to the Newbler assembly (Supplementary Note 11). Both assemblies cover more than 98% of the reads from an extensive transcriptome data set, indicating that the proteome is well represented (Supplementary Note 13). The assemblies are consistent with four independently assembled bacterial artificial chromosome (BAC) insert clones (Supplementary Note 14 and Supplementary Fig. 9), and with the expected insert size of paired BAC-end reads (Supplementary Note 15 and Supplementary Fig. 10).

A standard annotation approach based on protein evidence was complemented by a whole-genome alignment of the Atlantic cod with the stickleback (*Gasterosteus aculeatus*), after repeat-masking 25.4% of the Newbler assembly (Supplementary Note 16 and Supplementary Table 6). In this way, 17,920 out of 20,787 protein-coding stickleback genes were mapped onto reorganized scaffolds (Supplementary Note 17). Additional protein-coding genes, pseudogenes and non-coding RNAs were annotated using the standard Ensembl pipeline. These approaches resulted in a final gene set of 22,154 genes (Supplementary Table 7). Comparative analysis of gene ontology classes indicates that the major functional pathways are represented in the annotated gene set (Supplementary Note 18 and Supplementary Fig. 11). We anchored 332 Mb of the Newbler assembly to 23 linkage groups of an existing Atlantic cod linkage map using 924 SNPs[8] (Supplementary Note 19 and Supplementary Table 8). These linkage groups have distinct orthology to chromosomes of other teleosts, on the basis of the number of co-occurring genes, showing that the whole-genome shotgun assembly reflects the expected chromosomal ancestry (Fig. 1, Supplementary Note 20 and Supplementary Table 9).

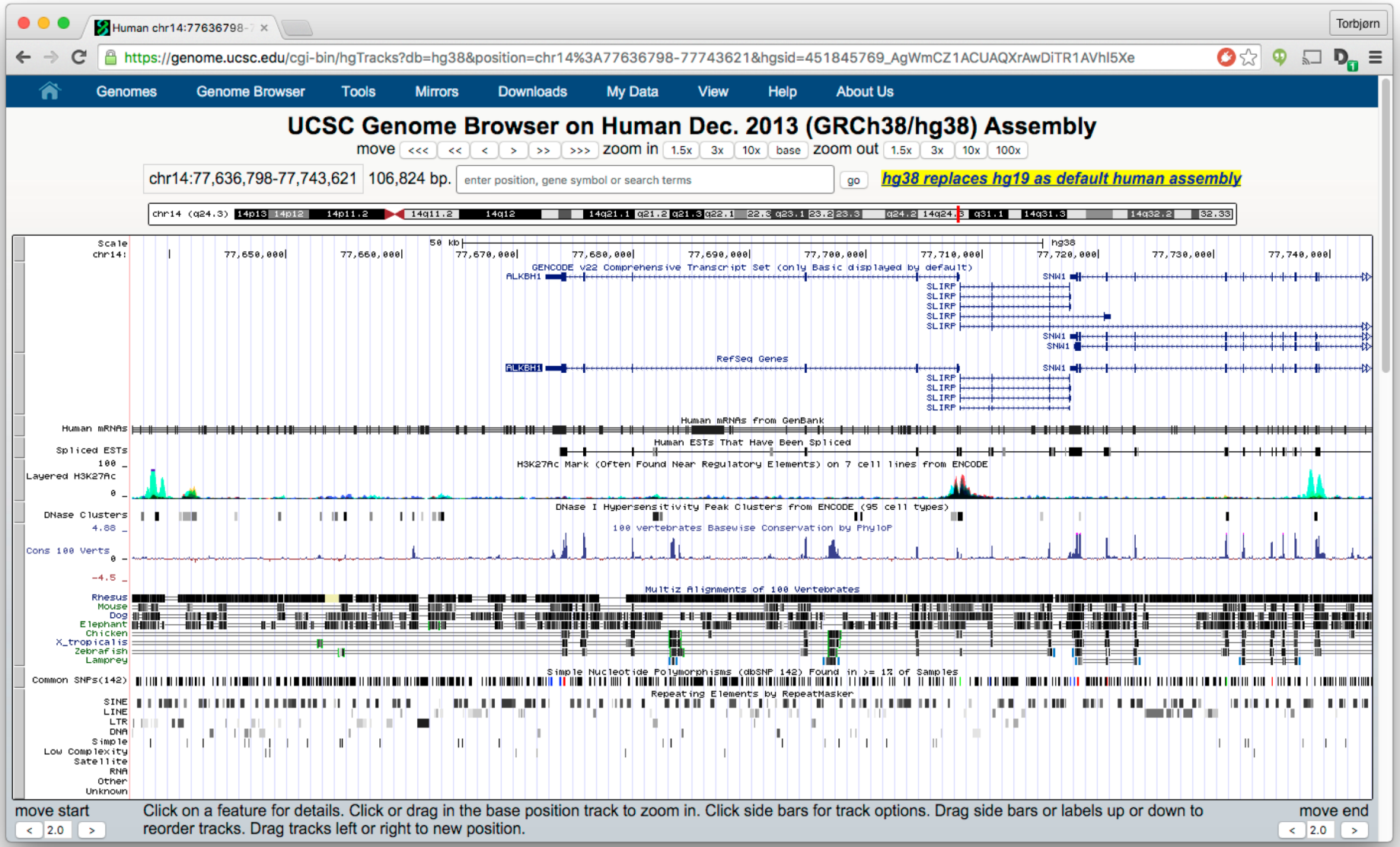
Table 1 | Assembly statistics

	Number	Bases (Mb)	N50L (bp)*	N50 (n)†	ML (bp)‡
Newbler					
Contigs§	284,239	536	2,778	50,237	76,504
Scaffolds	6,467	611	687,709	218	4,999,318
Entire assembly‖	157,887	753	459,495	344	4,999,318
Celera					
Contigs§	135,024	555	7,128	19,938	117,463
Scaffolds	3,832	608	488,312	373	2,810,583
Entire assembly‖	17,039	629	469,840	395	2,810,583

*Minimum sequence length in which half of the assembled bases occur.
†Number of sequences with lengths of N50L or longer.
‡Maximum length.
§Contigs longer than 500 bp.
‖Scaffolds and unplaced contigs.

[1]Centre for Ecological and Evolutionary Synthesis (CEES), Department of Biology, University of Oslo, PO Box 1066, Blindern, N-0316 Oslo, Norway. [2]Department of Molecular Biosciences, Centre for Immune Regulation, University of Oslo, Blindern, N-0316 Oslo, Norway. [3]Bioinformatics Core Facility, Institute for Medical Informatics, Oslo University Hospital, Montebello, N-0310 Oslo, Norway. [4]Department of Informatics, University of Bergen, N-5020 Bergen, Norway. [5]Department of Natural Sciences and Technology, Hedmark University College, P. O. Box 401 0, Bedriftsenteret, N-2306 Hamar, Norway. [6]Department of Animal and Aquacultural Sciences, University of Life Sciences, P.O. Box 5003, N-1432 Ås, Norway. [7]Department of Biology, Centre for Geobiology, University of Bergen, N-5020 Bergen, Norway. [8]Computational Biology Unit, Uni Computing, Uni Research AS, N-5020 Bergen, Norway. [9]454 Life Sciences, 15 Commercial Street, Branford, Connecticut 06405, USA. [10]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK. [11]Nofima Marine, P.O. Box 5010, N-1430 Ås, Norway. [12]Institute of Marine Research, P.O., Box 1870, Nordnes, N-5817 Bergen, Norway. [13]Department of Animal and Aquacultural Sciences, CIGENE, Centre for Integrative Genetics, Norwegian University of Life Sciences, PO Box 5003, 1432 Ås, Norway. [14]Faculty of Biosciences and Aquaculture, University of Nordland, N-8049 Bodø, Norway. [15]Department of Pharmaceutical Biosciences, School of Pharmacy, University of Oslo, P.O. Box 1068, Blindern, N-0316 Oslo, Norway. [16]Max Planck Institute for Molecular Genetics, Ihnestrasse 63-73, D-14195 Berlin-Dahlem, Germany. [17]Institute for Basic Sciences and Aquatic Medicine, School of Veterinary Sciences, N-0033 Oslo, Norway. [18]Department of Medical Biology, Faculty of Health Sciences, University of Tromsø, N-9037 Tromsø, Norway. [19]Department of Biology, PO Box 7803, University of Bergen, N-5020 Bergen, Norway.

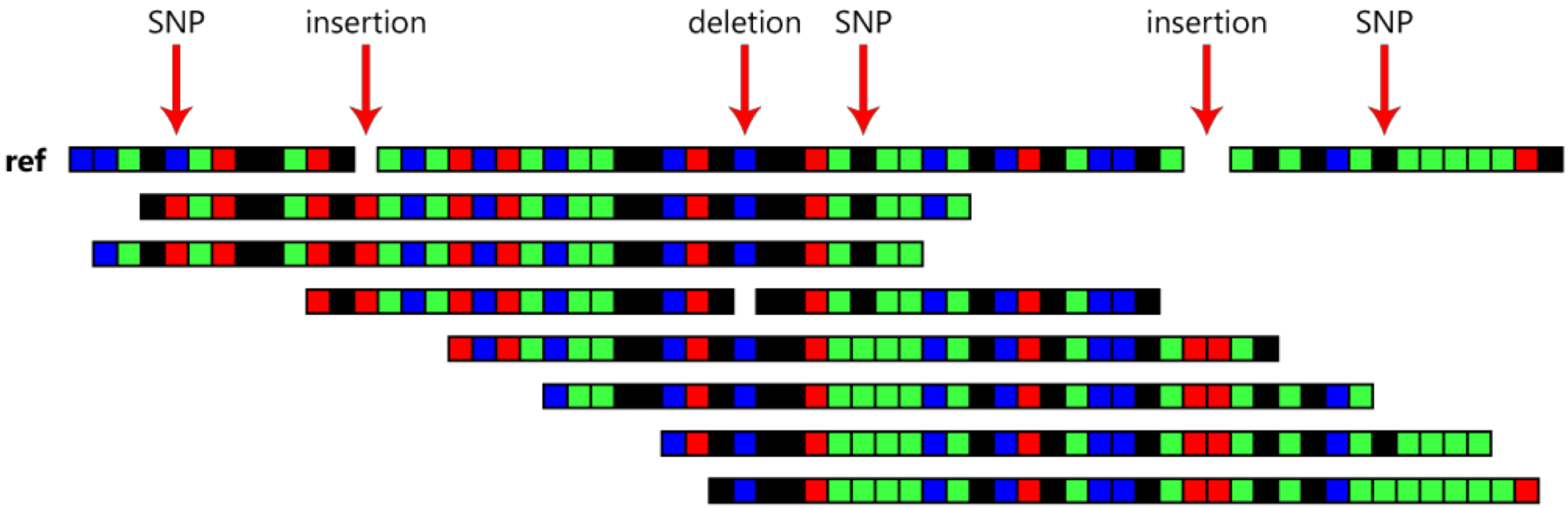
Genome browsers



Source: genome.ucsc.edu

Resequencing

- Sequencing DNA from a new individual when we already have a reference genome sequence
- Map reads to reference genome instead of assembly

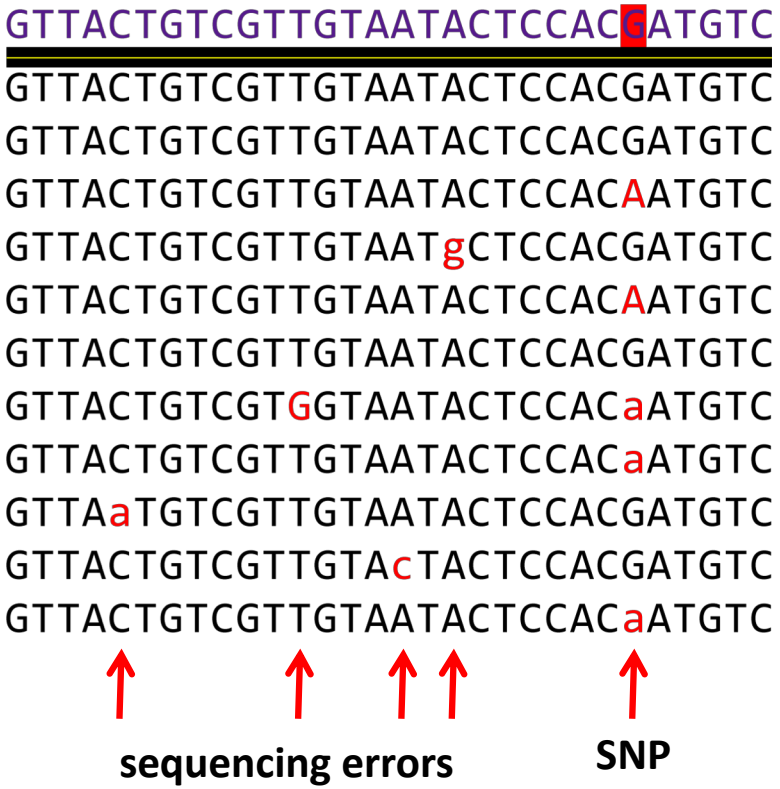


Variation detection by resequencing

- Natural variation discovery
- Mutation detection

- Single Nucleotide Polymorphisms (SNPs)
- Small insertions & deletions (Indels)
- Copy Number Variation (CNV)
- Large inversions, translocations etc

- Requires high coverage, that is, the average number of times each base is sequenced (typically 40X, but may require 100X)



Mapping and coverage

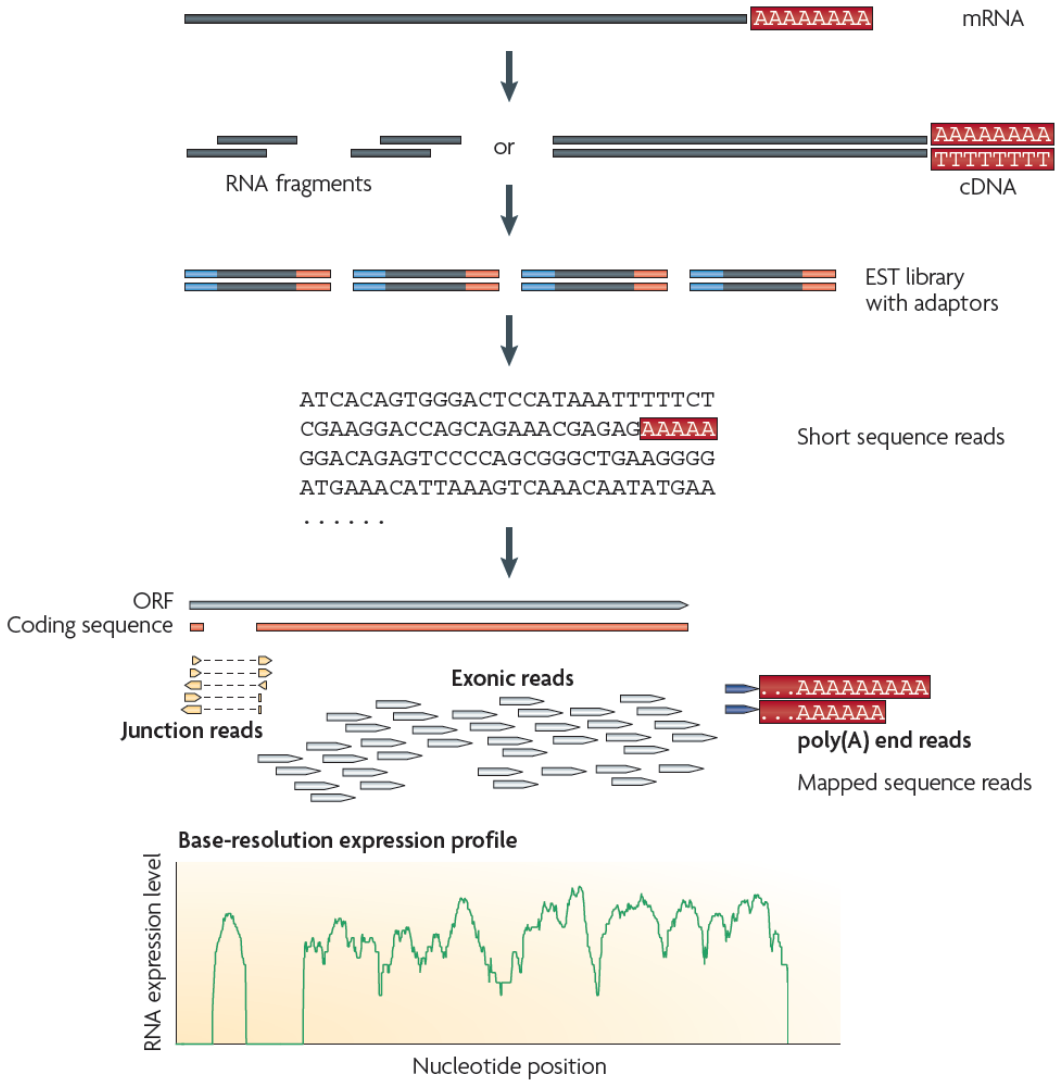
Gene expression (RNA-Seq)

- Gene expression analysis
- “transcriptomics”
- Replaces microarrays

- mRNAs
- Small RNAs (miRNA, piRNA…)

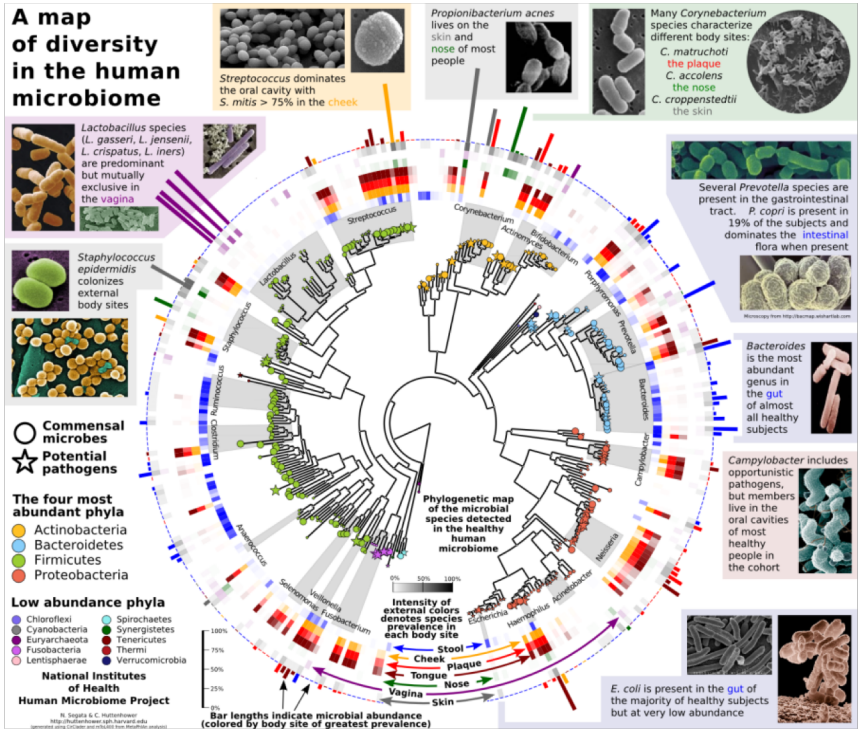
- Splice variants

- Counts the number of reads for each RNA



Metagenomics

- Samples contains collection of DNA/RNA from many microorganisms present in some niche - a microbial community
- Sources: Soil, ocean, mine, human body, the built environment, ...
- Ecological diversity studies
- Clinical studies (e.g. human gut)
- Big data: Many hundred million sequences



TARA OCEANS

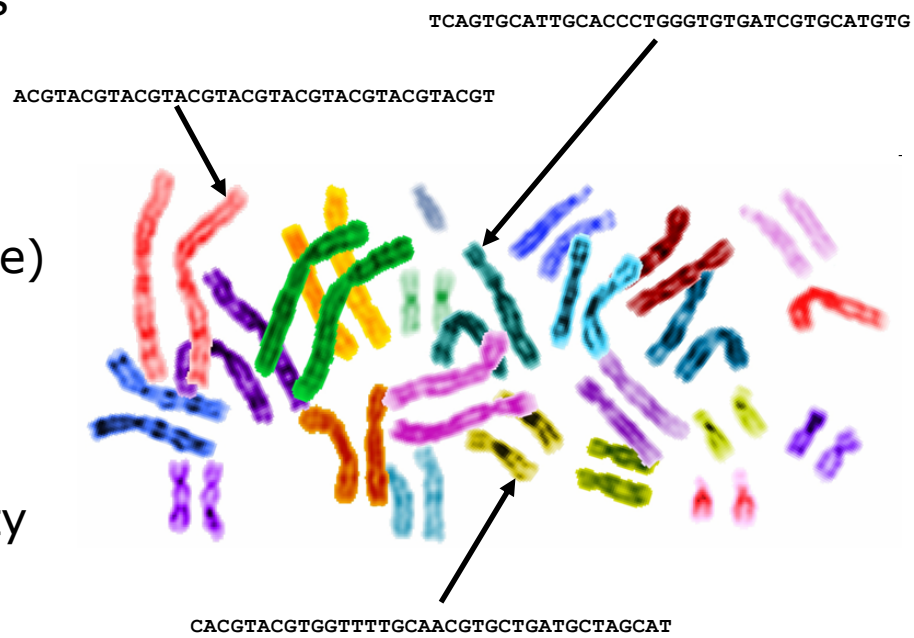
earth microbiome project

HMP

Human Microbiome Project

Mapping reads to a reference genome

Goal: Identify positions in the genome that are most similar to the sequence reads



Input data:

- 10-1000 million reads, each 30-300bp
- Sequencing errors (typ. ~1% error rate)

Reference genome:

- E.g. human genome, 3 Gbp
- Some genome variation, heterozygocity

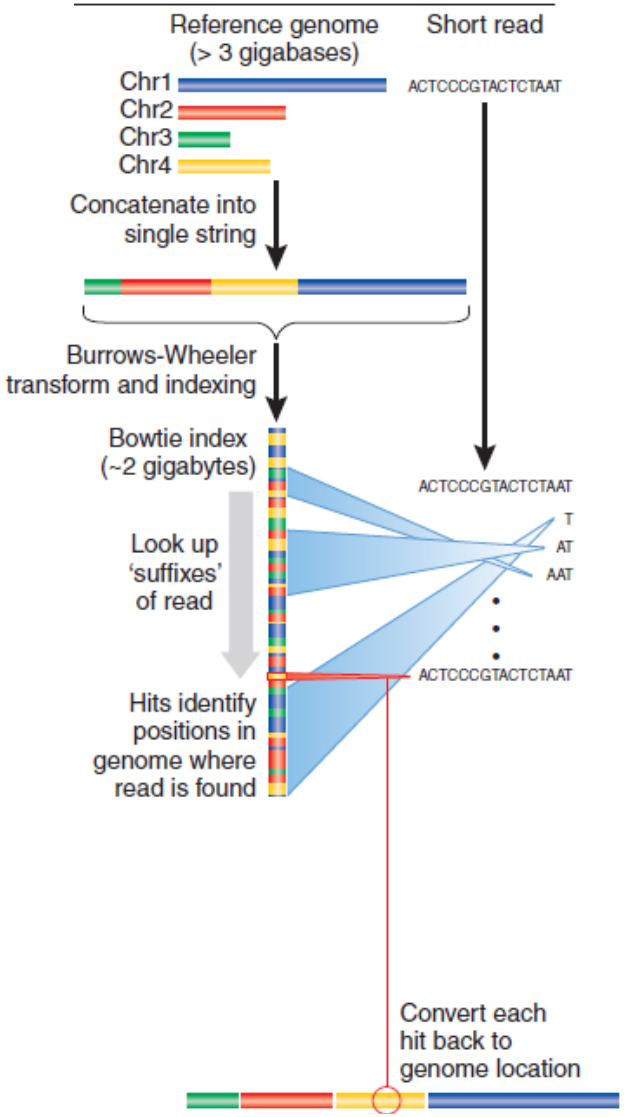
Output:

- 0, 1, or more potential genomic locations for each read
- Mapping quality assignment

Requirements:

- Sensitivity, specificity, speed, compactness

BWT alignment: Bowtie / BWA



- The reference genome is indexed using the Burrows-Wheeler Transform (BWT) and an FM-index is built.
- The BWT/FM-index allows exact matches to be found extremely fast and using little memory
- For each read, use the index to find one or more parts of the read that matches well
- Report best or all matches in the genome for each read
- Take paired-ends reads into account

Sequencing challenges

- Cost of actual sequencing is decreasing rapidly, but what about the cost of analysis?
- Lack of competent people for bioinformatics analysis
- Large storage needs due to the amounts of data generated. Terabytes of data.
- Compute intensive analysis (read mapping, assembly)
- Security and privacy issues related to sensitive human data


Thank you!