INF5380/INF9380 NORBIS course
High-performance computing in bioinformatics

Introduction to assembly and mapping

Torbjørn Rognes
torognes@ifi.uio.no

Department of Informatics, UiO
25 April 2016

UiO : Universitetet i Oslo

SEQUENCING

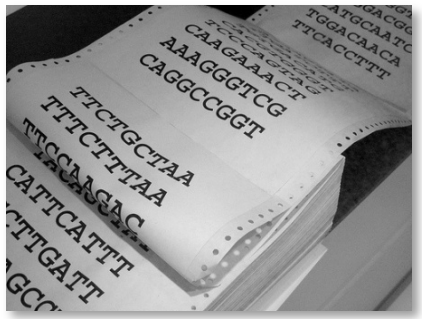
Genome sequencing

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

Obtain sample

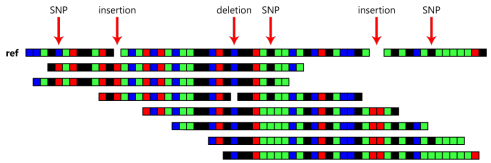
DNA extraction, preparation

Sequencing



Raw data

Sequence analysis



Interpretation

Sequencing systems

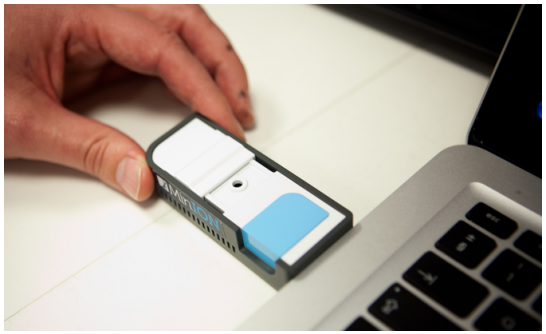
Roche (454)

ABI (SOLiD)

Ion Torrent

Illumina (Solexa)

Pacific Biosciences SMRT



Oxford Nanopore

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 properties

Box 1 | **Sequencing and mapping technologies**

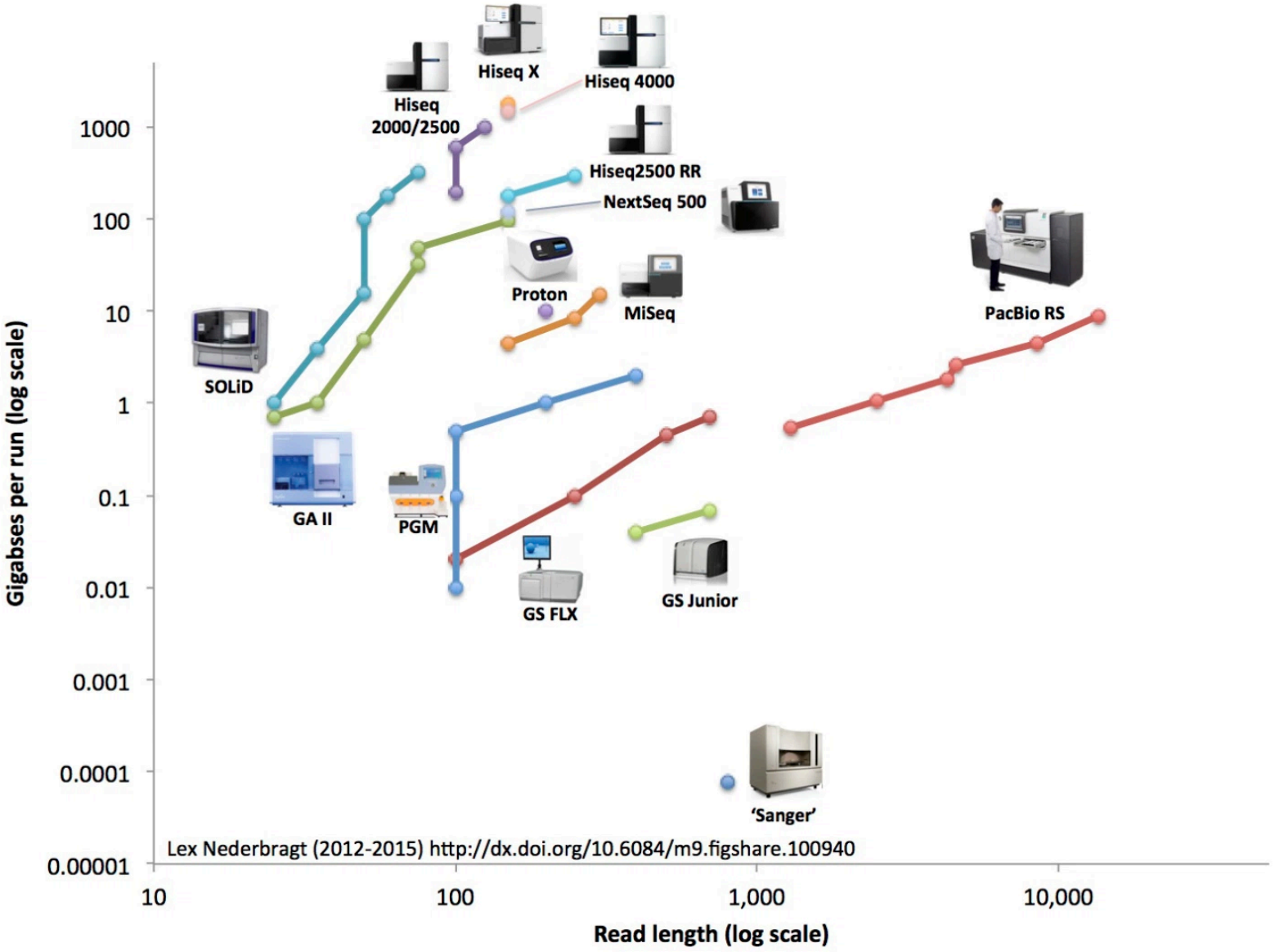
A full survey of sequencing technologies is beyond the scope of this article. The table below compares the currently available technologies in terms of several characteristics of importance for genome assembly: read length, error rate and the ability to generate paired-end reads natively. Note that potentially 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

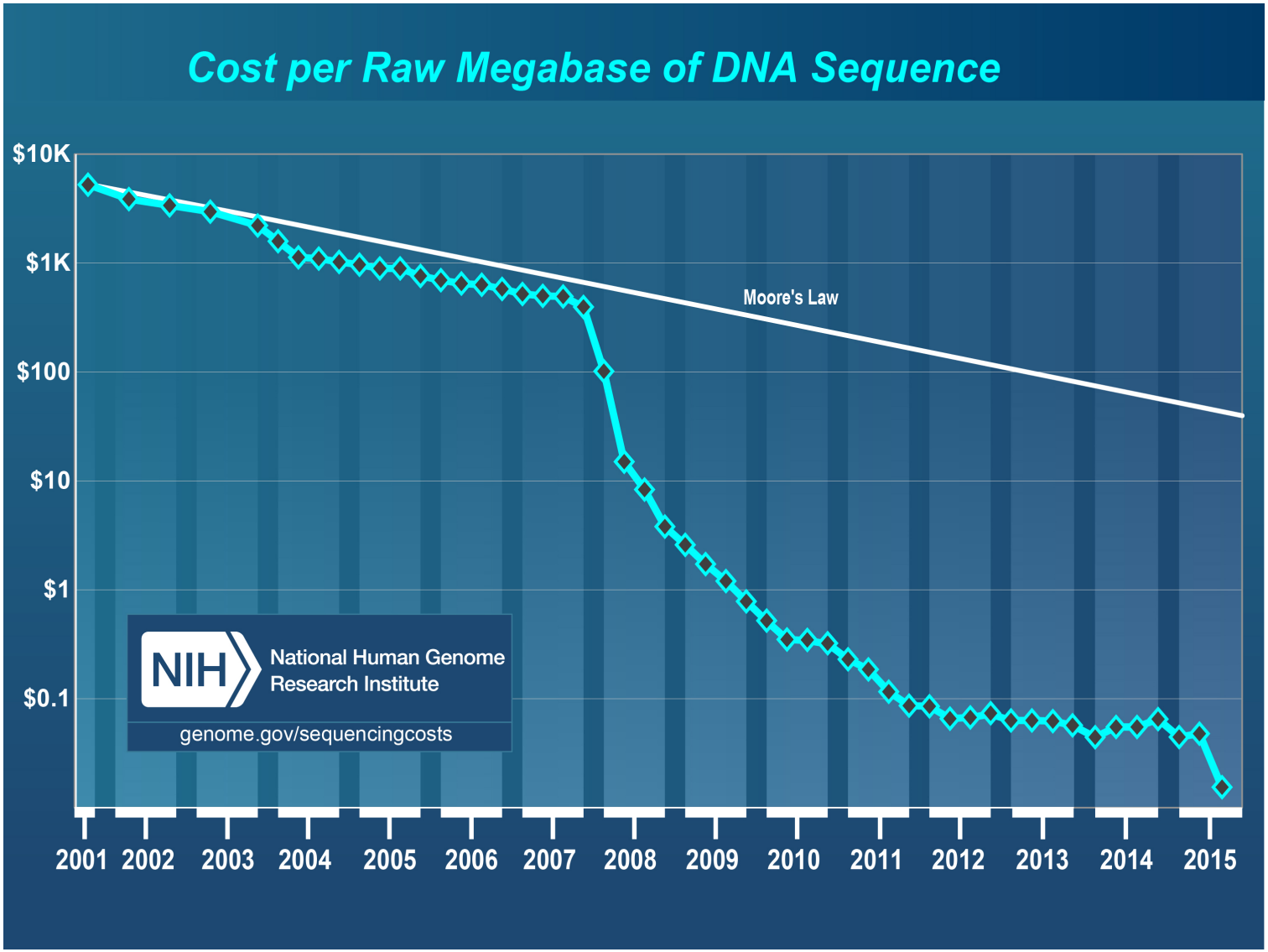
*454 and Ion Torrent technologies are prone to errors in homopolymer regions, which are segments of the genome in which the same nucleotide is repeated multiple times[94]. ‡Pacific Biosciences instruments produce reads with an exponential distribution of read lengths, only a few of which reach the multi-kb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

Developments in High Throughput Sequencing



Source: Lex Nederbragt (2012-2015) https://doi.org/10.6084/m9.figshare.100940

The cost of sequencing



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 followed by the same identifier as the first line. This identifier is optional.
- The fourth line contains characters that represent the quality scores for each nucleotide

FASTQ quality scores

- Quality value Q (or Phred quality score):
  $Q = -10 \log_{10} p$
  where p = probability of the base being incorrect

- The Q values are encoded as ascii characters by adding 33 to the value:
  $c = 33 + Q$

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

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

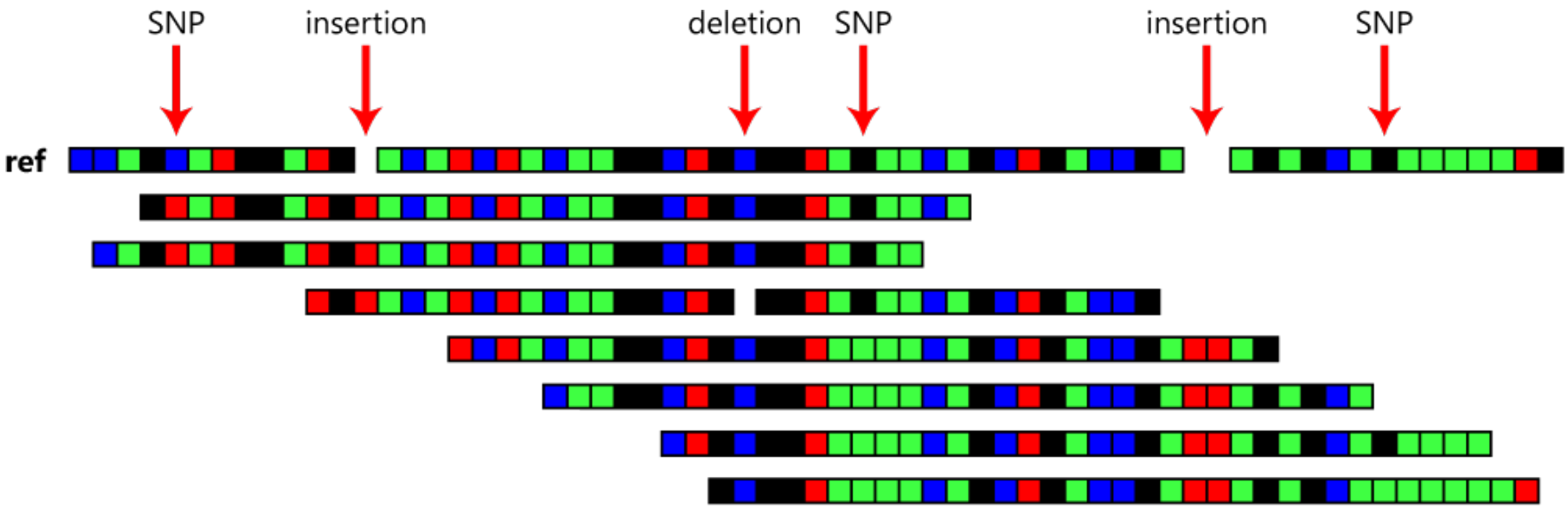
0	<NUL>	32	<SPC>	64	@	96	`
1	<SOH>	33	!	65	A	97	a
2	<STX>	34	"	66	B	98	b
3	<ETX>	35	#	67	C	99	c
4	<EOT>	36	$	68	D	100	d
5	<ENQ>	37	%	69	E	101	e
6	<ACK>	38	&	70	F	102	f
7	<BEL>	39	'	71	G	103	g
8	<BS>	40	(	72	H	104	h
9	<TAB>	41	)	73	I	105	i
10	<LF>	42	*	74	J	106	j
11	<VT>	43	+	75	K	107	k
12	<FF>	44	,	76	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>

Applications

- Resequencing
  - RNA-Seq
  - ChIP-Seq
  - Bisulphite sequencing
  - Variant detection
  - ...
- *De novo* Genome sequence assembly
- Metagenomics

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

- High coverage required

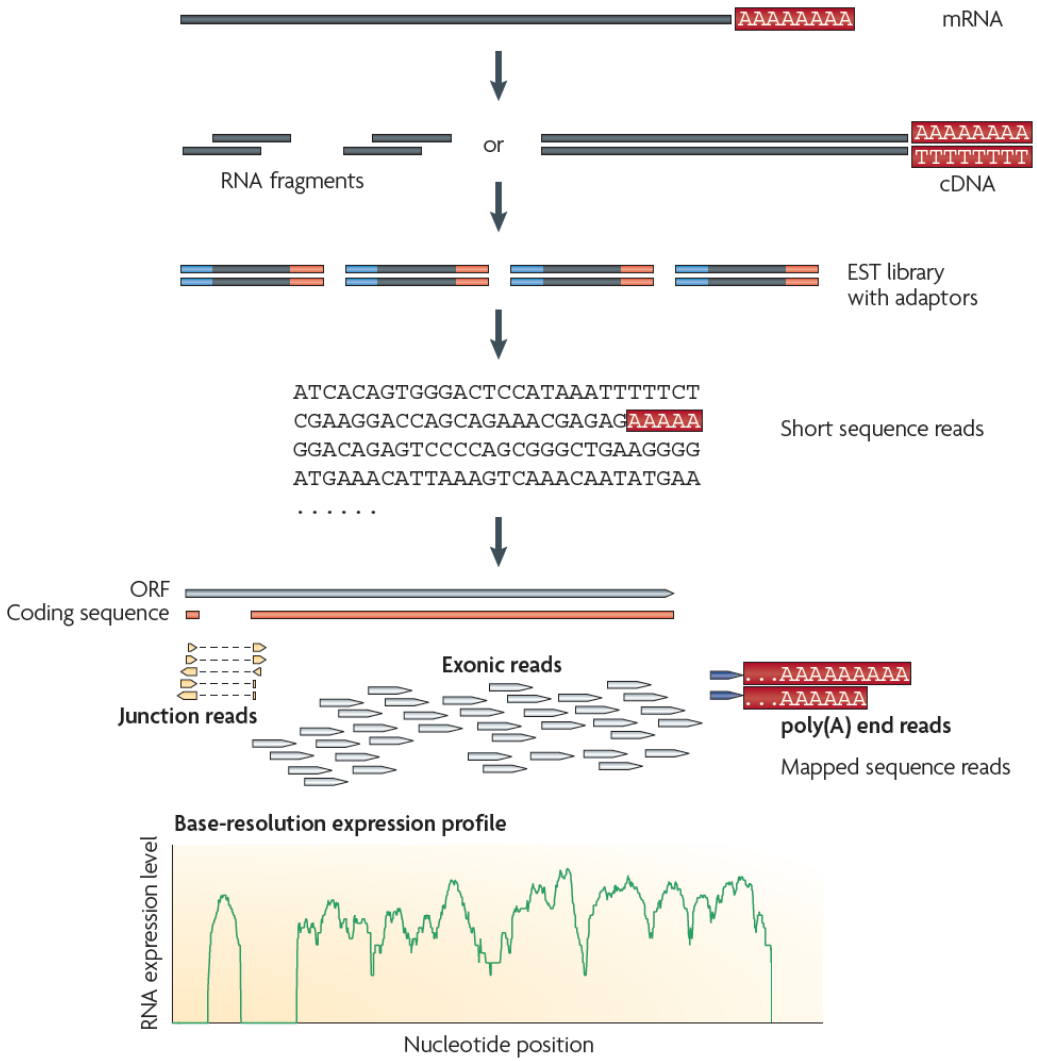
Gene expression (RNA-Seq)

- Gene expression analysis
- “transcriptomics”
- Replaces microarrays

- mRNAs
- Small RNAs (miRNA, piRNA…)

- Splice variants

- Counts the number of reads for each RNA



Software tools

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, Newbler, Phrap, TIGR, Arachne, Velvet, MaSuRCA, SPAdes, ALLPATHS-LG, ...

Mapping: Mapping sequence reads to a known genome sequence, used initially in resequencing procedures.
E.g.: BWA, Bowtie, SOAP2, Maq, BFAST, RMAP, ...

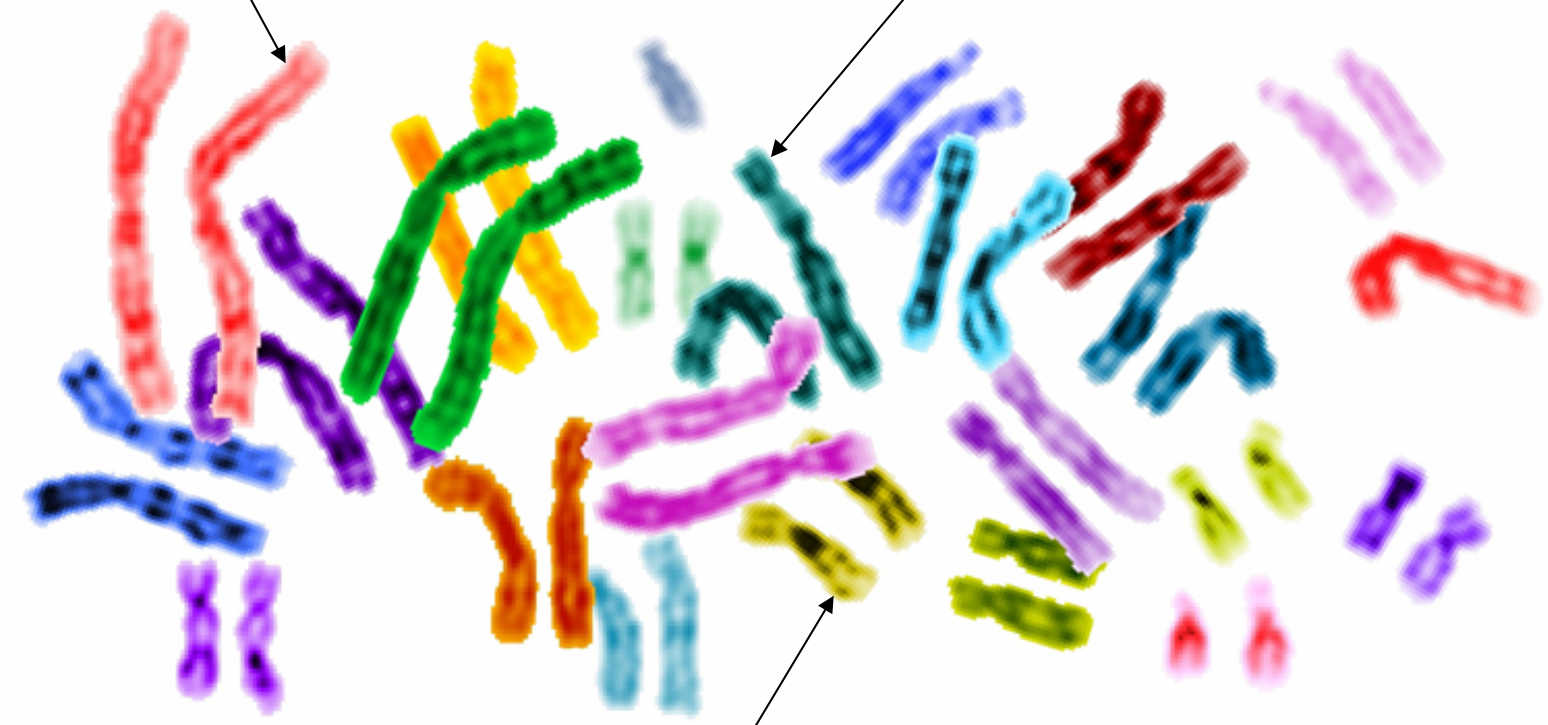
Variant detection: GATK, SAMtools, FreeBayes, Mutect, Strelka, ...

MAPPING

Mapping reads to a reference genome

TCAGTGCATTGCACCCTGGGTGTGATCGTGCATGTG

ACGTACGTACGTACGTACGTACGTACGTACGTACGT



CACGTACGTGGTTTTGCAACGTGCTGATGCTAGCAT

Mapping basics

Input:

- Millions to billions of (short) sequence reads, e.g. 100-200 bp
- A (large) reference sequence, e.g. 3Gbp human genome

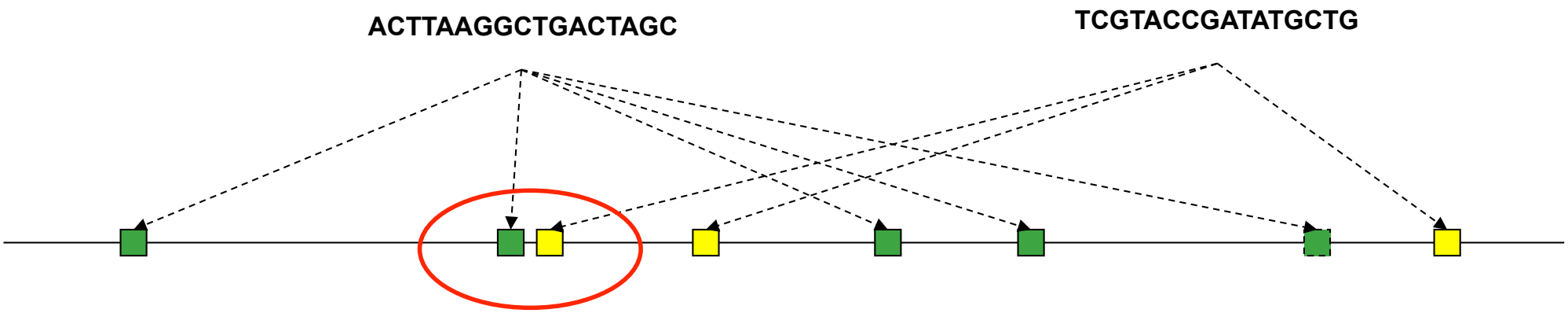
Goal:

- Identify positions in the genome that are most similar to the sequence reads
- Some differences (substitutions and gaps) must be allowed due to SNPs and small indels in the sequenced genome vs the reference genome, and to allow room for PCR and sequencing errors

Requirements:

- Sensitive: Don't overlook relevant matches
- Specific: Find the best match, not the second best
- Small: Limited amounts of computer memory
- Fast: We cannot wait for many days

Multiple mapping

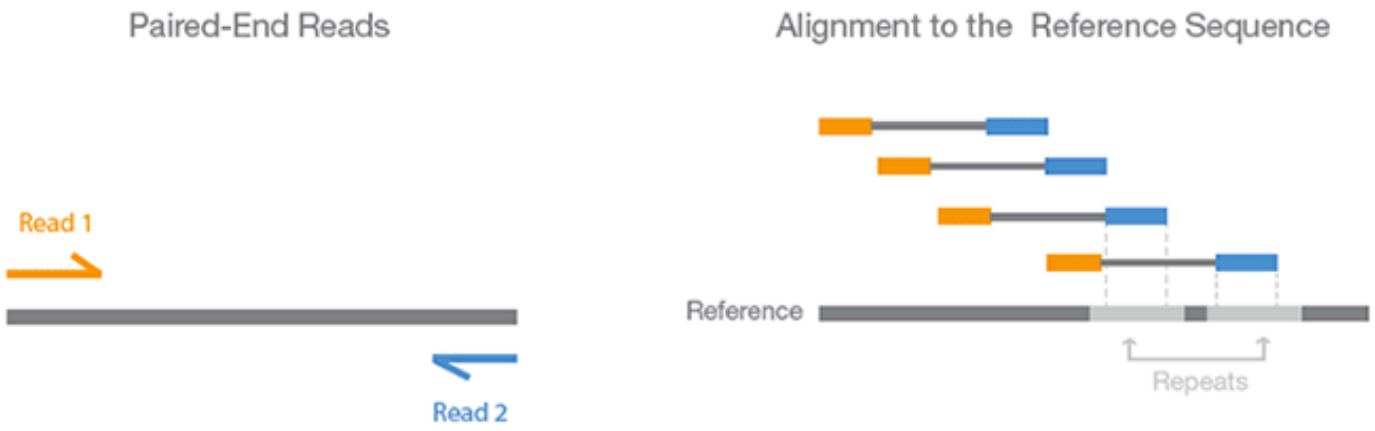


- Problem:
  - Short reads (e.g. 100bp) may not map uniquely due to long repeats (>100bp) in the genome
- Solutions:
  - Get longer reads
  - Get paired reads (pairs of reads with fixed distance)

Paired ends / 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

Figure 4. Paired-End Sequencing and Alignment



Paired-end sequencing enables both ends of the DNA fragment to be sequenced. Because the distance between each paired read is known, alignment algorithms can use this information to map the reads over repetitive regions more precisely. This results in much better alignment of the reads, especially across difficult-to-sequence, repetitive regions of the genome.

Source: Illumina

Mapping method

- Approach:
  - Initial rapid filtering using some form of indexing
  - Further examination of a small amount of interesting positions
- Index based on
  - Reference genome
  - Reads
- Heuristics:
  - Only a small part of the read is often used initially (the most reliable part of the sequence from Illumina sequencers)
  - Only 2 mismatches (and 1 gap) may be allowed in this region
- Quality of reads:
  - Some nucleotides in the reads are more certain than other
  - The probability of each called base being erroneous is indicated
  - Some programs take the quality values of the read into account
  - Allows mismatches at low quality positions

Mapping strategies

- Methods based on hashing or indexing of seed sequences
  - Continuous seeds
    - Single
      - RMAP, Novoalign, Mosaik
    - Multiple
      - FASTA, BLAST v 2, BLAT, SSAHA
  - Spaced seeds
    - Single
      - Eland, SOAP, Maq, SeqMap, BFAST
    - Multiple
      - ZOOM, SHRiMP, RazerS

- Methods based on Burrows-Wheeler Transform (BWT)
  - Bowtie, BWA, SOAP2, BWT

Indexing

- Indexing uses a table with all possible words of w symbols and their positions in the genome
- All positions in the genome where each word is found can be looked up directly
- Given alphabet size S, genome size N and word length w, it requires $(S^w + N) * \log_2 N$ bits
- Typical index width is 12-14 nucleotides
- Human genome, w=14:
  $(4^{14} + 3*10^9) * 4$ bytes
  = 1GB + 12GB = 13GB
- Example with string “ANANAS” and index width 2

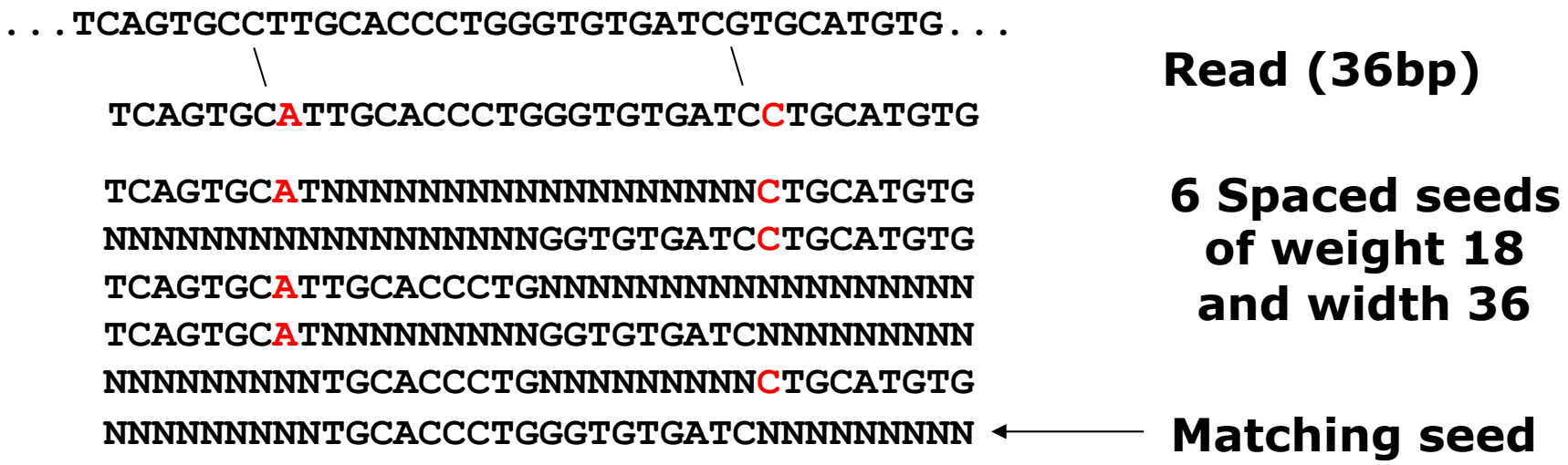
Index key	Positions
AA	
AN	1,3
AS	5
NA	2,4
NN	
NS	
SA	
SN	
SS	

Single seed mapping with *k* mismatches

- Divide the read into k+1 seeds
- Look up each seed in the index
- At least one seed must match
- Example below with k=2 and 36bp long reads
- Guaranteed to find all hits with up to k mismatches

Spaced seeds

- Spaced seeds are patterns of nucleotides that must match and some that do not need to match at specific positions
- Found to be more efficient / specific than continuous seeds
- The seeds have a width equal to the total number of positions and a weight equal to the number that must match
- Positions that must match are indicated with 1's and the other positions are indicated with 0's. Only 1-positions are indexed.
- E.g. 1001110011100l, weight 8 and width 14
- Often multiple spaced seeds (patterns) are used.

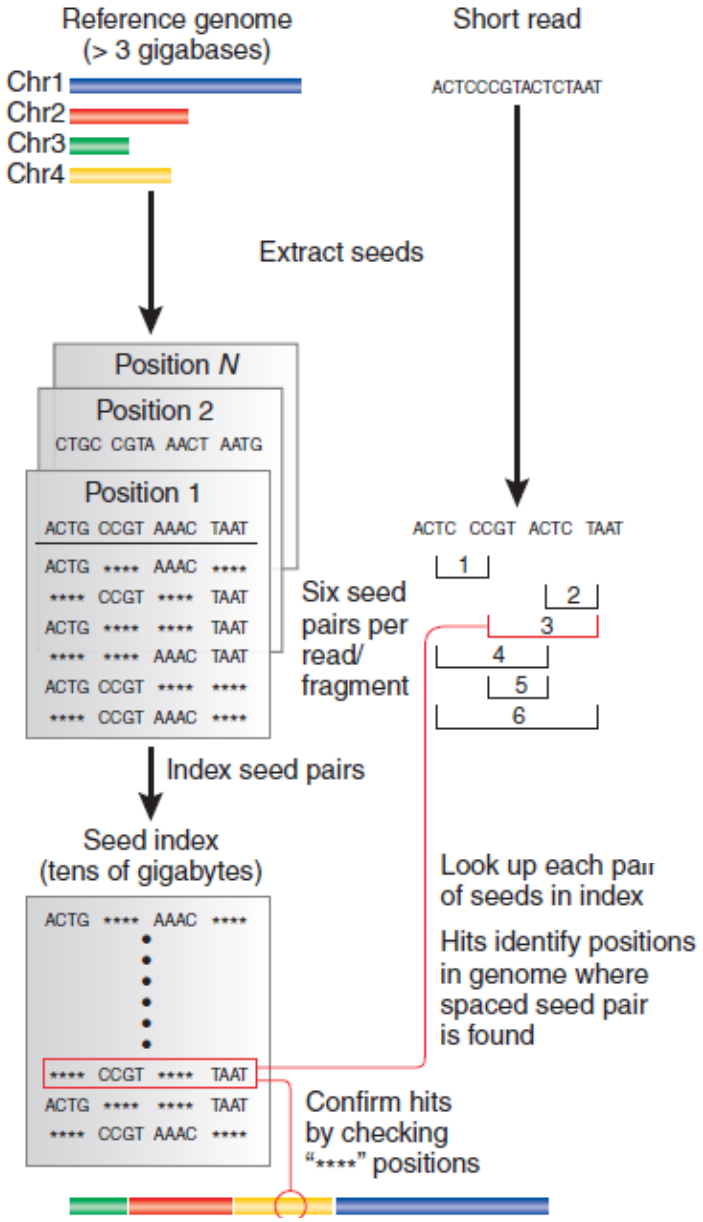


Spaced seed alignment: SOAP

- Indexes reference genome, several indices
- Uses 6 spaced seed templates
- Seeds contain four fragments of e.g. 8 nt, but only two fragments must match:

```
11111111 00000000 11111111 00000000
00000000 11111111 00000000 11111111
11111111 00000000 00000000 11111111
00000000 00000000 11111111 11111111
11111111 11111111 00000000 00000000
00000000 11111111 11111111 00000000
```

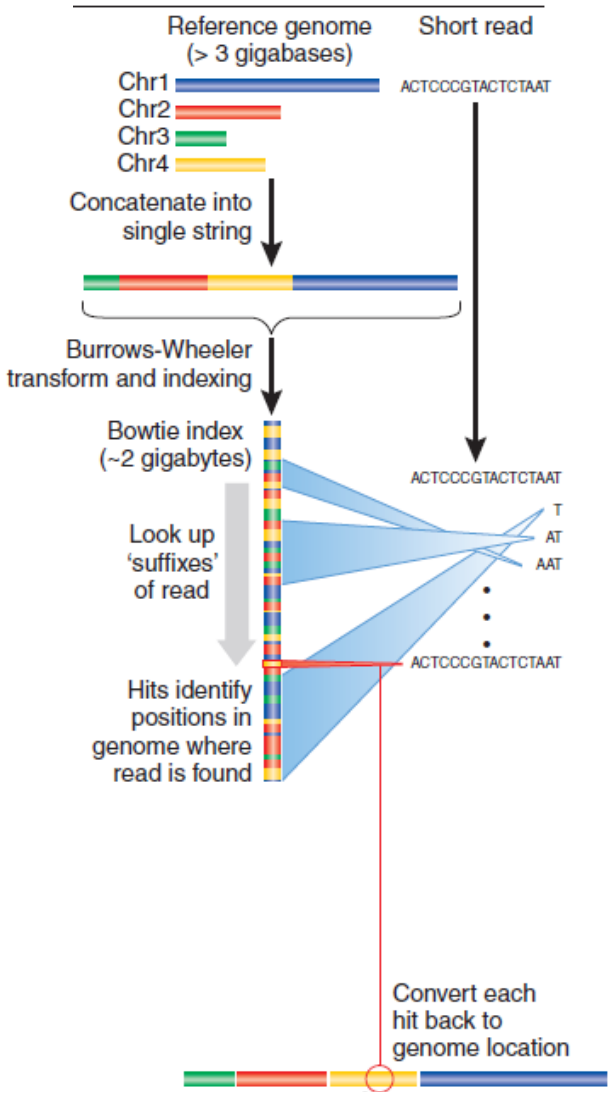
- Guaranteed to find all hits with max 2 mismatches



Burrows-Wheeler Transform (BWT)

- A kind of sequence index that allows a program to very rapidly find all perfect matches to given string
- A simple and elegant reversible transformation of a string
- Used in combination with a suffix array and a few other arrays
- Can be stored very compactly. The human genome can be stored in about 1.5GB.
- May be well compressed by storing only parts of the arrays, e.g. only 4 bit / nucleotide
- Takes about an hour to generate from the human genome (but can be used over and over)

BWT alignment: Bowtie / BWA



- The reference genome is indexed using BWT
- For each read
  - Initially, only a part of the read is used
  - Exact matches:
    - BWT *backward search*
  - Inexact matches:
    - Introduce substitutions or gaps at all possible positions in the read and then look for exact matches
    - To save time, the search tree may be pruned
- Multiple matches are identified
- Report matches to each read

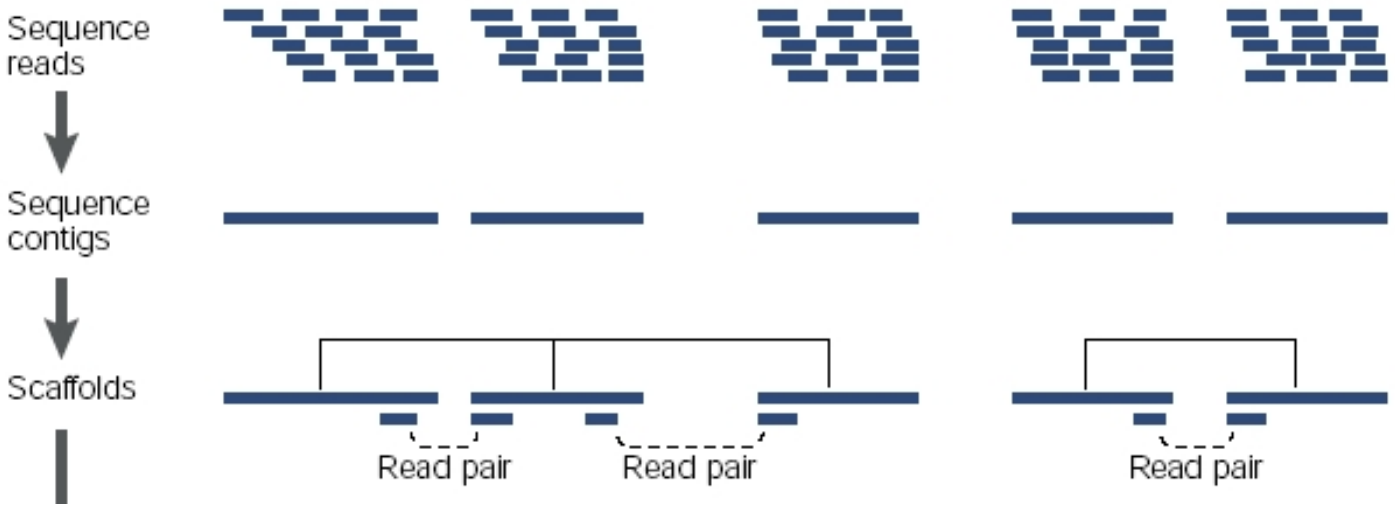
GENOME ASSEMBLY

Why assemble genomes?

- When a known very similar genome exist we are doing resequencing and can use mapping to place the reads
- Otherwise, "*de novo*" genome assembly is needed
- Almost any type of analysis requires an almost complete genome, for example
  - Gene finding
  - Transcription studies
  - Genome comparison between organisms

Complete genome *de novo* sequencing

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



Example: Reads of length 10

```
nøf,_tidde
snør,_det_
ddeli_bom.
,_den_snør
t_smør,_ti
Det_snør._
```

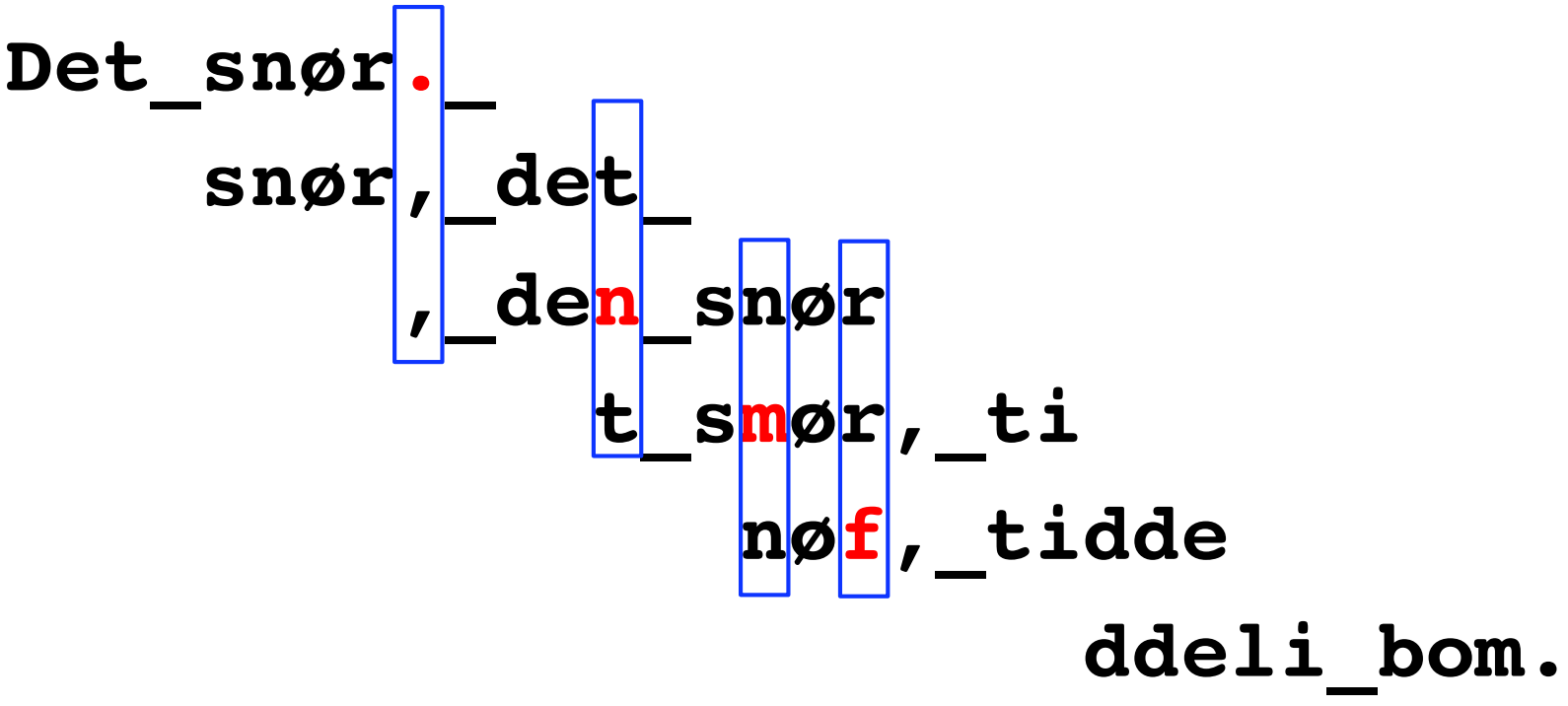
Example: Overlaps

```
                    nøf,_tidde
     snør,_det_
                                ddeli_bom.
          ,_den_snør
               t_smør,_ti
Det_snør._
```

Example: Layout

```
Det_snør._
    snør,_det_
        ,_den_snør
            t_smør,_ti
                nøf,_tidde
                    ddeli_bom.
```

Example: Consensus



Det_snør,_det_snør,_tiddeli_bom.

Repeat of length 9

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

Main assembly approaches

- Strategies
  - Greedy
  - Overlap – Layout – Consensus (OLC)
  - de Bruijn graph

- Other techniques
  - Comparative assembly: Uses information from related genomes to help during assembly (used in e.g. AMOS-cmp)

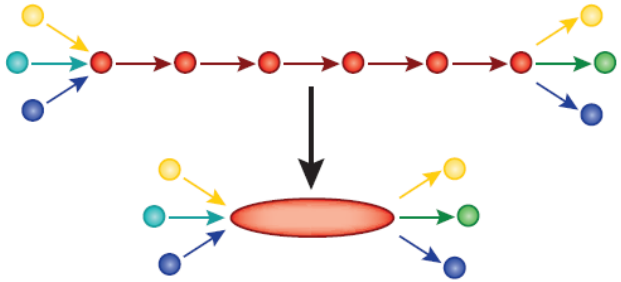
Overview of the assembly process

1. Fragment DNA and sequence

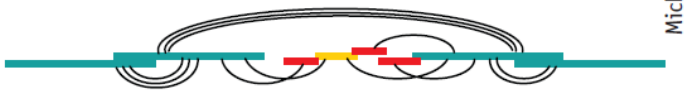
2. Find overlaps between reads

...AGCCTAGACCTACAGGATGCGCGACACGT
GGATGCGCGACACGTCGCATATCCGGT...

3. Assemble overlaps into contigs



4. Assemble contigs into scaffolds



Michael Schatz, Cold Spring Harbor

Genome assembly stitches together a genome from short sequenced pieces of DNA.

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

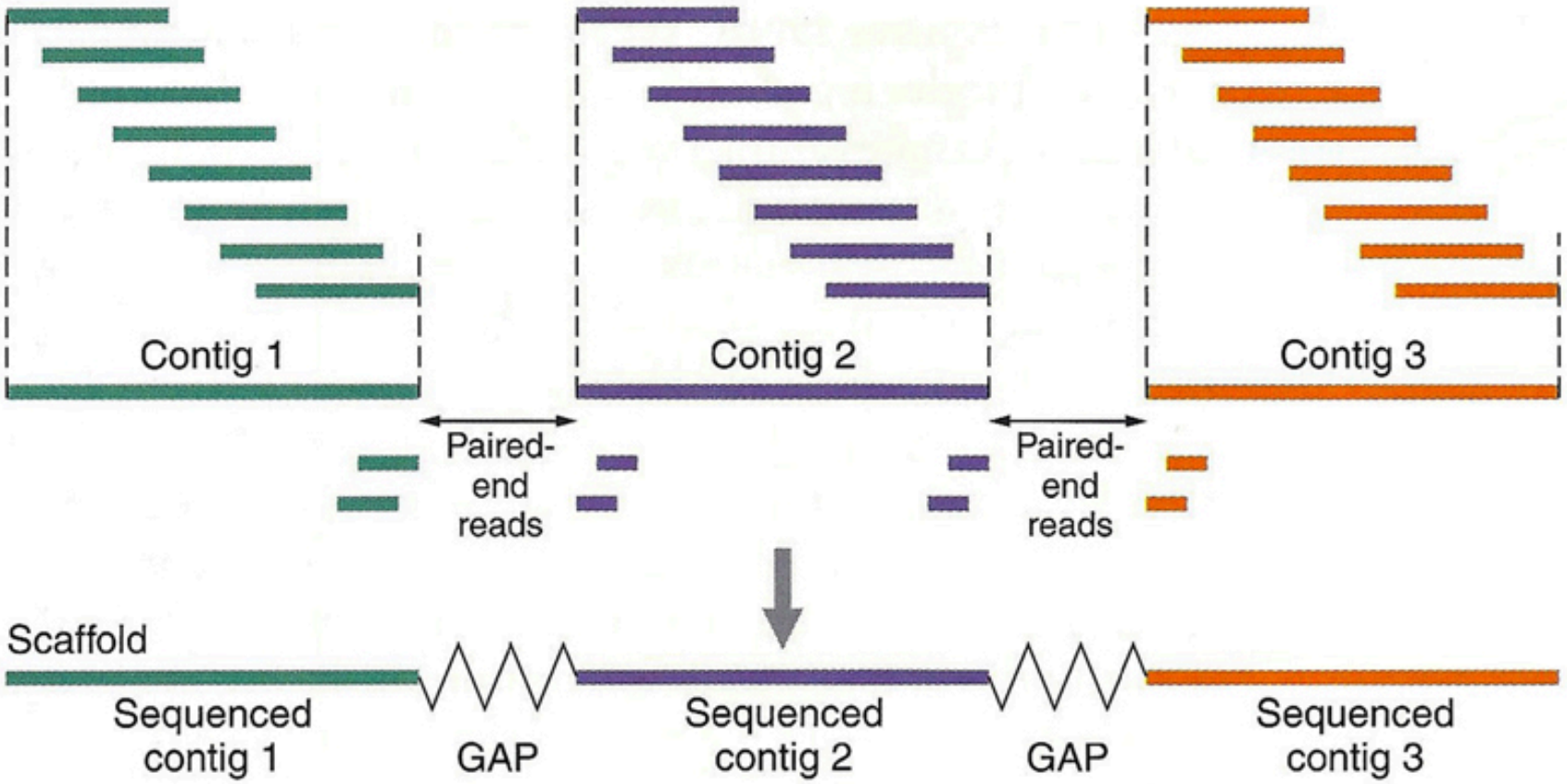
- Repeats
  - Our genomes are filled with many almost identical repeated sequences
  - Repeats longer than the read length makes it impossible to determine the exact location of the read
  - May cause compression or misassemblies
  - May be alleviated by longer reads or paired-end/mate pair reads

- Heterozygosity
  - Diploid organisms (e.g Humans) actually have two “genomes”, not one. Chromosome pairs 1-22 for all and XX for women. 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

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 at a fixed distance from each other, approximately.
- 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

Paired-end reads span gaps



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

Evaluation of assembly quality

- It is difficult to assess the quality of an assembly, especially when the correct genome is unknown.
- Common measures:
  - Total/average contig length
  - Total/average scaffold length
  - Number of contigs
  - Number of scaffolds
  - Breakpoints
  - Misassemblies
  - Incorrect bases, indels
  - Fraction of complete genome covered
  - Fraction of contigs/scaffolds covered
  - N_{50}

N_{50}

- The N_{50} length is the length of the shortest contig such that the sum of contigs of longer or equal length is at least 50% of the total length of all contigs.
- May also be used for scaffolds.
- May also be used with other percentages (e.g. N_{90}).
- May be misleading because it ignores assembly errors.