Genome sequencing and assembly

King Jordan

OUR SERVICES

The ABiL partnership combines the program management experience of IHRC with the bioinformatics expertise of Georgia Tech. ABiL offers a wide range of bioinformatics services to address data analysis and workforce development needs.

Omics Analysis

Genomics, proteomics, transcriptomics & epigenomics

Development

Development of bioinformatics analysis platforms

Molecular Typing

Genome-based typing, genotype-phenotype correlation

Consulting

Bioinformatics project conception, planning & execution

Big Data

High performance & cloud computing. Data mining & machine learning.

Training

Hands-on instructional modules for bioinformatics

http://abil.ihrc.com

Outline

- DNA and genome sequencing technology
- Genome sequence data and quality
- Genome assembly
  - Reference assembly
  - *De novo* assembly
- Assembly quality

Outline

- DNA and genome sequencing technology
- Genome sequence data and quality
- Genome assembly
  - Reference assembly
  - *De novo* assembly
- Assembly quality

Brief history of sequencing: terminology

- Sequencing read -> The trace of an individual DNA molecule/fragment as determined by the sequencing technology/platform

```
5'-ATCGATGGTATTATTGGCATAG-3'
```

- Sequencing Platform -> The machine that carries out the actual sequencing process *and* reads the bases, *i.e.* it generates reads

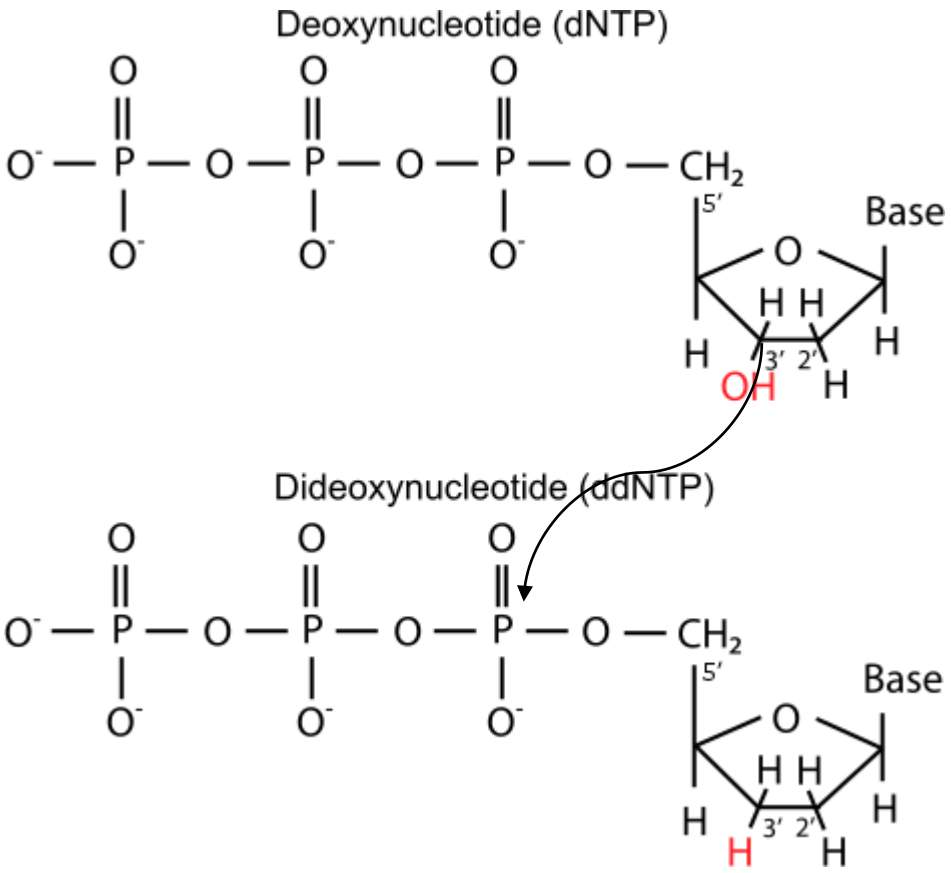
- Sequencing by synthesis -> Determining the bases which are present in a read by interrogating each base as it is added

Basics: Nucleic acid chain extension

- Goes in the 5’ to 3’ direction

  Depends on the presence of the 3’ hydroxyl to extend the chain

- Di-deoxy NTPs (ddNTPs) will terminate the chain -> shorter chain



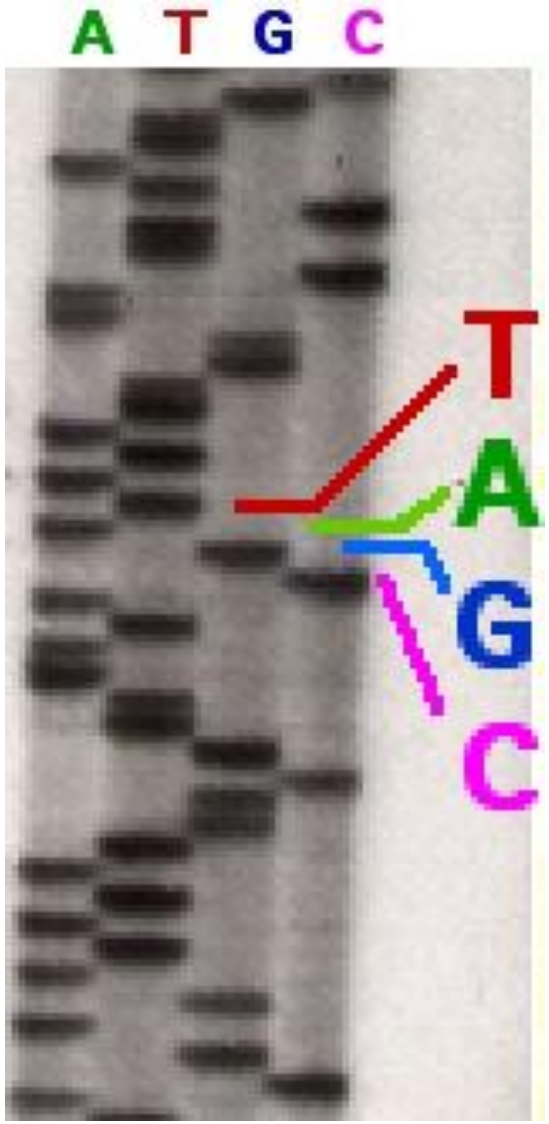
A brief history of sequencing: Sanger sequencing

- Chain termination sequencing – addition of radio-labeled di-deoxy (dd)NTPs to the sequencing reaction ddNTP stops extension

- Run the result on a gel and count the bands back to the top of the gel

A brief history of sequencing: Sanger sequencing

- You run the result on a gel, expose it to film and the radio-labeled ddNTPs mark the bases



A brief history of sequencing: Sanger sequencing

- Chain termination sequencing – addition of
  radio-labeled di-deoxy (dd)NTPs to the sequencing reaction
  ddNTP stops extension

- Run the result on a gel and count the bands back to the top of the gel

- Very inefficient
  Labor intensive
  A few kb per day

- Lots of early sequencing was done this way
  The first genes, the first viral genomes

This was the pre-genomic era

- The only genomes available were small, viral genomes
- The idea of sequencing any genome, let alone a large genome like ours was impossible
- One gene at a time. That was really it.
- Maybe the same gene across species for comparative purposes

Bioinformatics in the 80s

The term wasn't even around then

But I know a guy who claims to have invented it

Sequence analysis & signal processing

Gene prediction from primary sequence

Phylogenies galore – early evolution studies

Single/small gene sets and rRNA

1995 -> H. influenza – first free-living organism

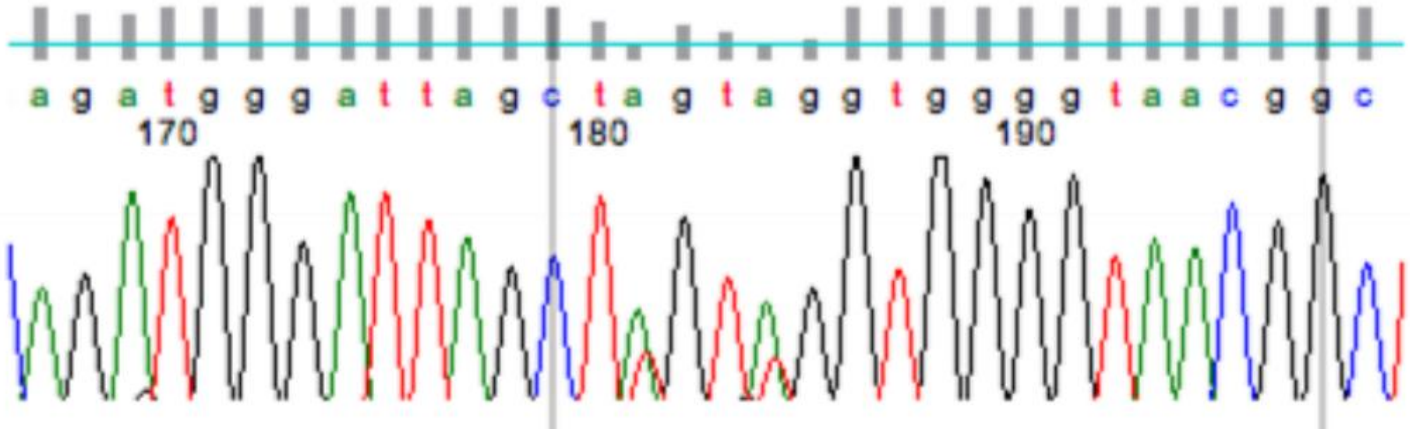
1996 -> S. cerevisiae – first eukaryote

1997 -> E. coli

Capillary sequencing completed the human genome

Still based on ddNTPs

ddNTPs are fluorescently labeled



All run in one reaction -> ¼ the space

Easier to automate – multiple 10s of kb per day

Capillary sequencing completed the human genome

Figure 3 The automated production line for sample preparation at the Whitehead Institute, Center for Genome Research. The system consists of custom-designed factory-style conveyor belt robots that perform all functions from purifying DNA from bacterial cultures through setting up and purifying sequencing reactions.

The early-mid 2000's – Genomic Era

Capillary sequencing on ABI machines allowed for unprecedented sequencing capacity

Still very labor intensive

2000 -> *A. thaliana*

2001 -> *H. sapiens* – automated sequencing in huge facilities, hundreds of capillary machines

2002 -> *M.musculus* – The mammal model

2005 -> *P. troglodytes* – Hugely important for comparative studies

2008 -> *B.floridae* – The Florida lancelet. Actually really important

Dozens of plants

Great time for evolutionary genomics

Something changed in the mid 2000s

One note: Sanger sequencing is still aroud

- What if you don't need to sequence an entire genome? What if you don't even need to sequence and entire gene?

  We have a collaborator who works in cystic fibrosis (there are also some interesting population genetic / evolutionary things there)

  She doesn't care about the whole genome, just 1 gene, and then one small part of that gene

  For her, one of these newer methods would be a waste.

  A few primer pairs are all she needs, three amplicons to cover the exon I think

  Then she does Sanger

Sequencing by synthesis

- This was the first SBS (sequencing-by-synthesis) machine to reach the market – 454

  Bases of a DNA molecule are read as a complimentary molecule is synthesized
  As opposed to the whole complimentary molecule being synthesized and then read out

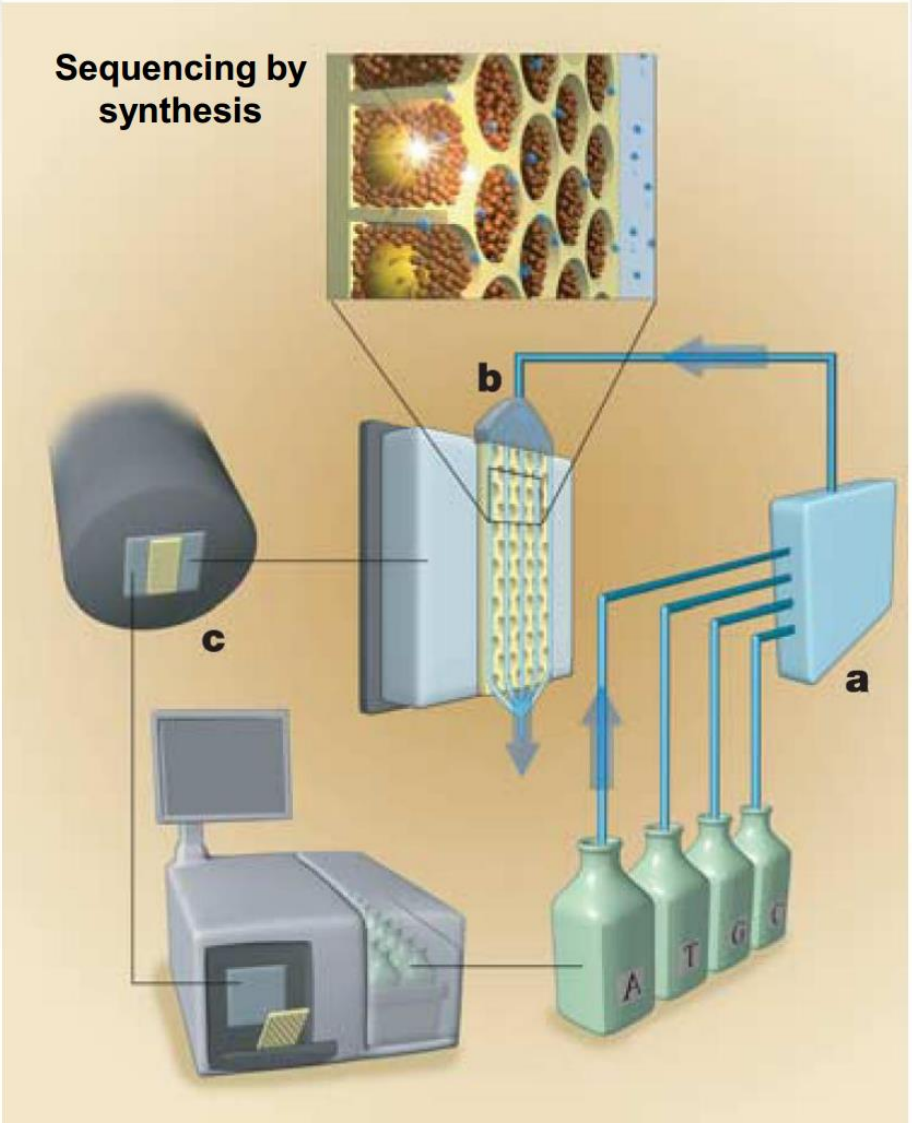
- Much smaller volumes of reagents

  Many, many reads at the same time

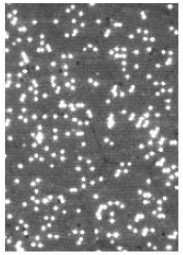
  454 was originally a few 100K vs 96 in capillary

- Illumina and Ion Torrent/Proton/PGM and PacBio are also SBS technologies – the bases are read as they're added onto the growing molecule

Pyrosequencing

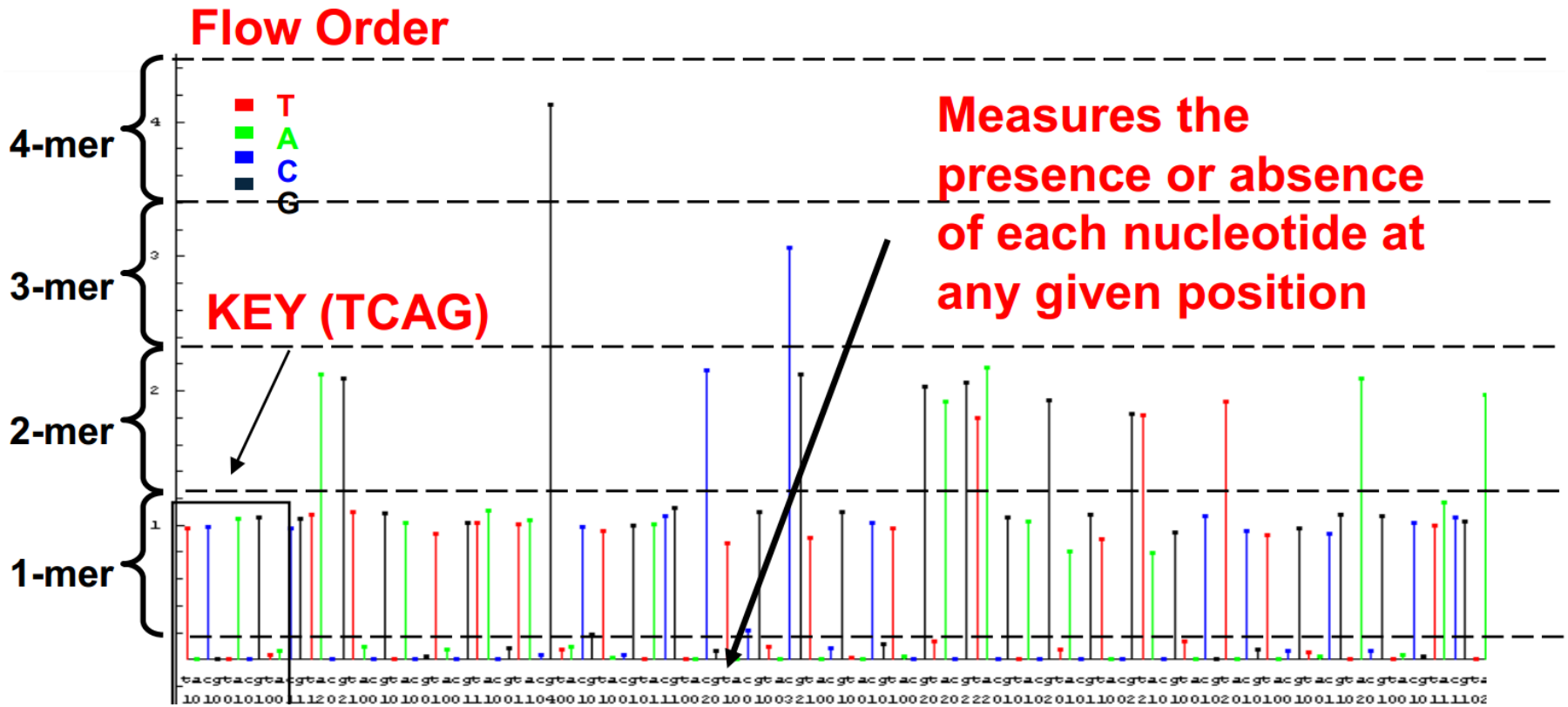


Photons generated are captured by CCD camera



Pyrosequencing

454: Flowgram (.sff file format)



454 was fantastic!

- It was a ***massive***, unbleiveable leap forward in sequencing technology
  - We could get hundreds of thousands of reads in one go
  - The small scale makes all the difference
  - 96 or so previously
- It revolutionized sequencing. Bacterial genomes could now be sequenced in one go – multiple genomes even
- And yet, it was only the beginning of modern sequencing
  - Actually now totally obsolete

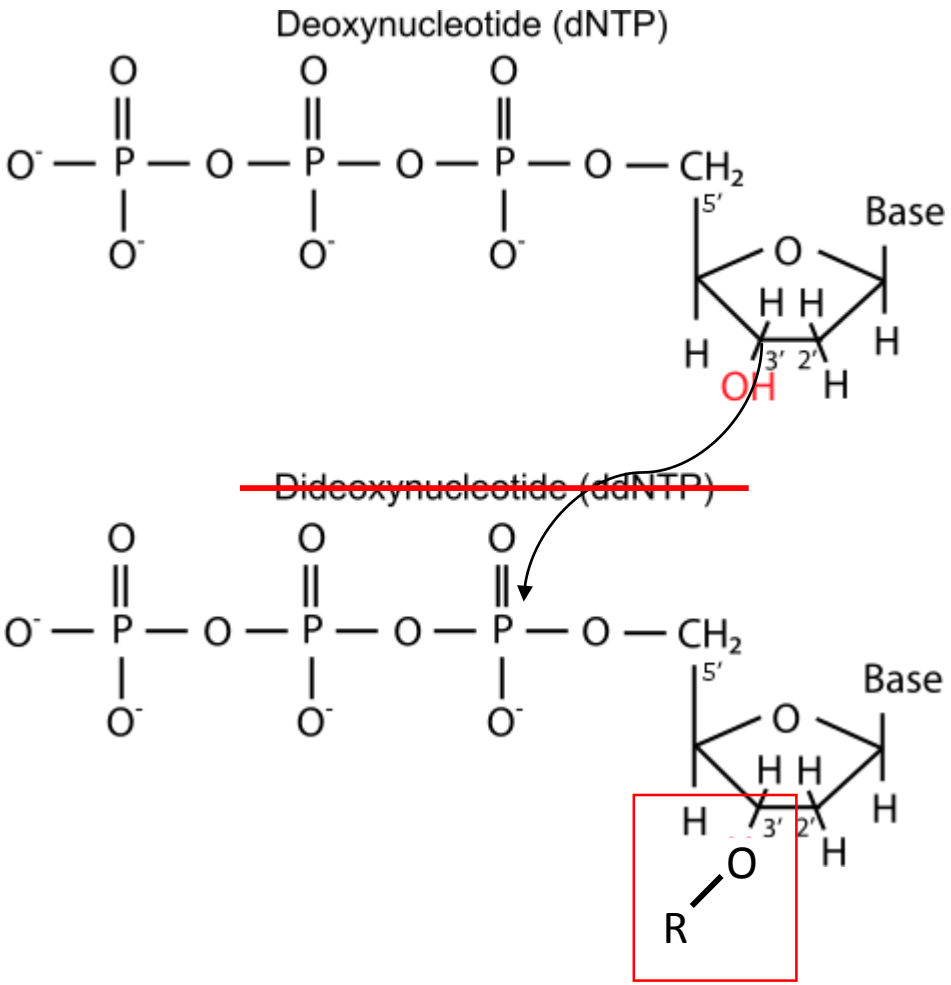
Next: Illumina sequencing technology

- Three things:
  1. Small scale, very tiny reactions, even smaller than 454, allowing for much greater density
  2. Reversible terminator chemistry - allows the controlled addition of one base at a time
  3. Optics. The unsung hero of the sequencing revolution. This allows for the insane density and number of reads that you can fit on an Illumina flow cell

- It wouldn't be an understatement to say that Illumina is the basis of modern biology

In Illumina, this chain termination is reversible

- Goes in the 5’ to 3’ direction
  Depends on the presence of the 3’ hydroxyl to extend the chain
- The nucleotide added has a block group on the 3’ carbon
- This blocking group prevents the nucleophilic attack -> only one base is added.
- The group is cleaved after each cycle, freeing up the 3’ hydroxyl
- Should mention that the group is also a strong fluorphore



Illumina was/is as far ahead of 454 as 454 was ahead of sanger/capillary sequencing

- *ca*. 2008, 454 could manage ~400K reads. Illumina could give you 50 million reads, about 100X as many as 454

- That number has hugely increased over time, as has the length of the reads Illumina provides

  More and longer reads -> quadratic growth in sequencing yield over the past few years

- Illumina sequencing probably accounts for 95% of sequencing data generated – that's being very generous to Ion and PacBio

Modern Illumina Machine: HiSeq 2500

- 10 genomes
- 150 Nextera Rapid Capture exomes
- 80 whole transcriptome RNA samples

(Assumes 100 gigabases (Gb) per genome at 2x125bp, 5 Gb per exome at 2x100bp, 50 million reads per transcriptome)

Increase daily throughput

The HiSeq v4 reagent kits generate up to 1 terabase (1Tb) of data per 6-day run (up to 500 Gb per flow cell), increasing daily throughput to 167 Gb per day. The new v4 reagents increase the number of clusters by 33% compared to the TruSeq SBS Kit v3, adding additional capacity for counting assays.

KIT NAME	OUTPUT MAX (PER 2 FLOW CELL)	NO. OF READS	MAX READ LENGTH	TIME
HiSeq SBS V4 Kits	Up to 1 Tb	Up to 4 billion	2 x 125 bp	6 days

The first machines produced < 30 million reads

The HiSeq is extremely expensive

MiSeq Specifications

Cluster Generation and Sequencing

MISEQ REAGENT KIT V2

READ LENGTH	TOTAL TIME*	OUTPUT
1 × 36 bp	~4 hrs	540-610 Mb
2 × 25 bp	~5.5 hrs	750-850 Mb
2 × 150 bp	~24 hrs	4.5-5.1 Gb
2 × 250 bp	~39 hrs	7.5-8.5 Gb

MISEQ REAGENT KIT V3

READ LENGTH	TOTAL TIME*	OUTPUT
2 × 75 bp	~21 hrs	3.3-3.8 Gb
2 × 300 bp	~56 hrs	13.2-15 Gb

Enough for 80 high quality bacterial genomes

MiSeq is ***very*** common for bacteria genomics

Small reads are an issue

- With Roche we lost a few hundred base pairs in our reads, with Illumina we lost a large fraction
- It is much better now (100-300 bp) as compared to 35 bp before but it is still too small for a lot of applications
- E.g. repeat resolution in genome assembly is a challenge with smaller reads and is typically left unresolved due to lack of data
- If only we can get more sequence using the same technology…

Extending read lengths

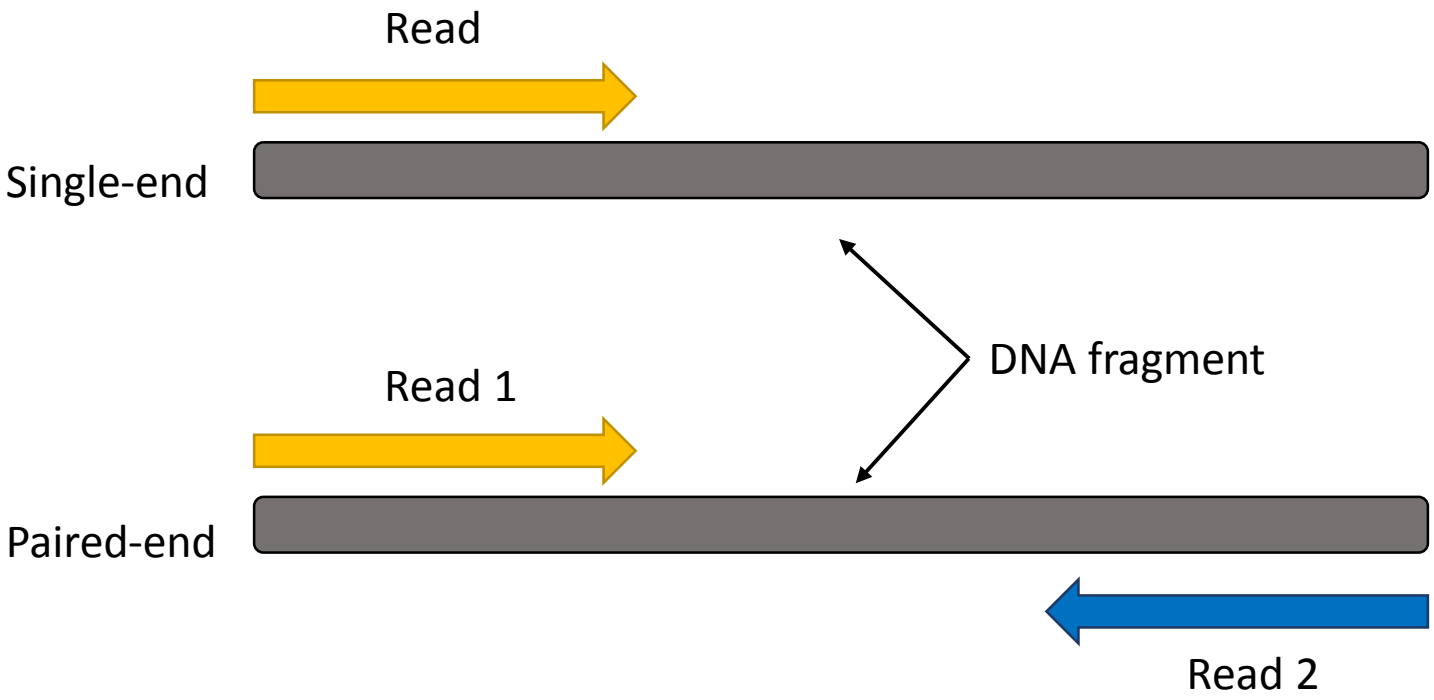
- If the machine is basically adding more bases and reading them, why not let it add a few more bases to it?

- A: Loss in fidelity
  The current read length is the length that can be read correctly with high confidence
  Anything longer than that is difficult to achieve for a variety of reasons including loss in transcriptase fidelity

- How about reading the DNA fragment twice, once from each end?

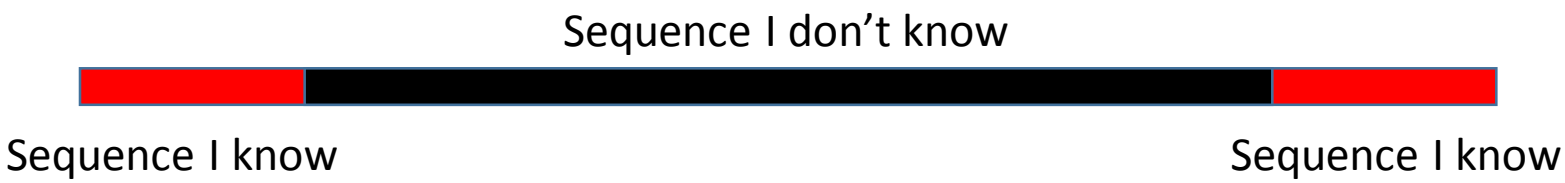
Paired-end sequencing

- In paired-end sequencing, a DNA fragment is read twice – once from each end (recall the Illumina video!)

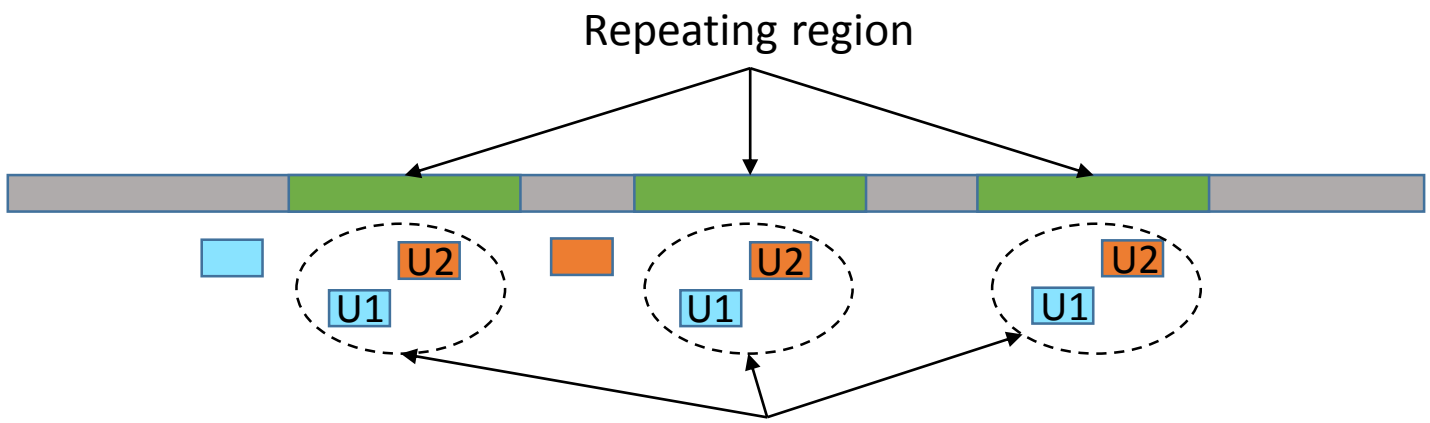


Paired-end sequencing

- But wait… this didn't really increase my read length!
- Yes, but it gave me one important piece of information – how far away a one read of the pair is from the other

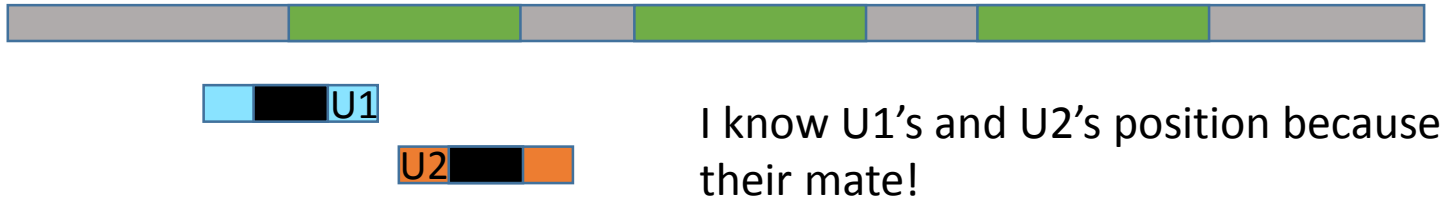


How does that help me?



Case 1: single-end

Where are these reads coming from???



I know U1's and U2's position because of their mate!

Case 2: paired-end

Paired-end sequencing

Paired-end sequencing provides many advantages over single-end sequencing:

- Sequencing from both ends of each fragment is a more efficient use of the fragment library
- *A priori* information of connected reads (paired reads) helps in improving alignment

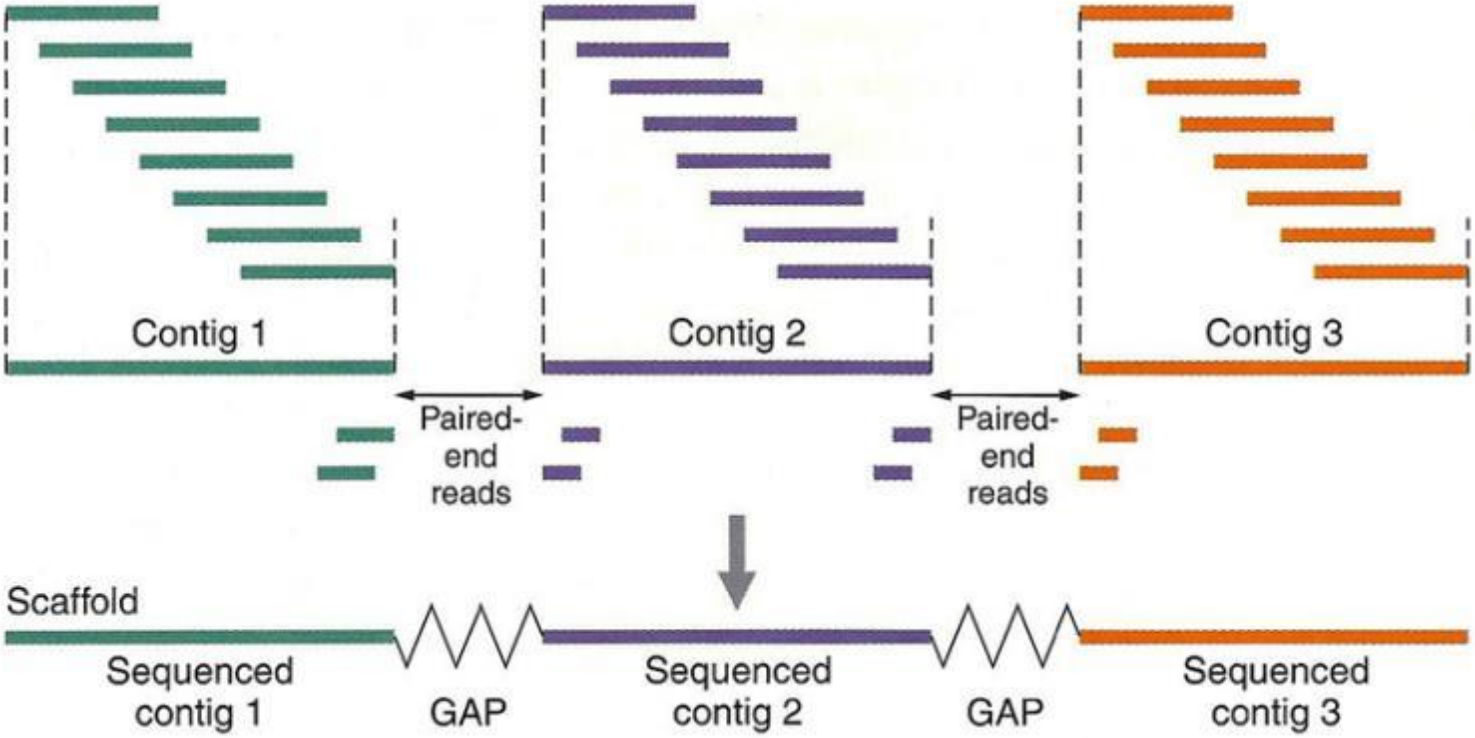
Paired-end sequencing

Paired-end sequencing provides many advantages over single-end sequencing:

- It also helps in resolving chromosomal rearrangements like insertions, deletions and inversions
- Scaffolding becomes possible due to the known connections between reads

Paired-end sequencing

Scaffolding



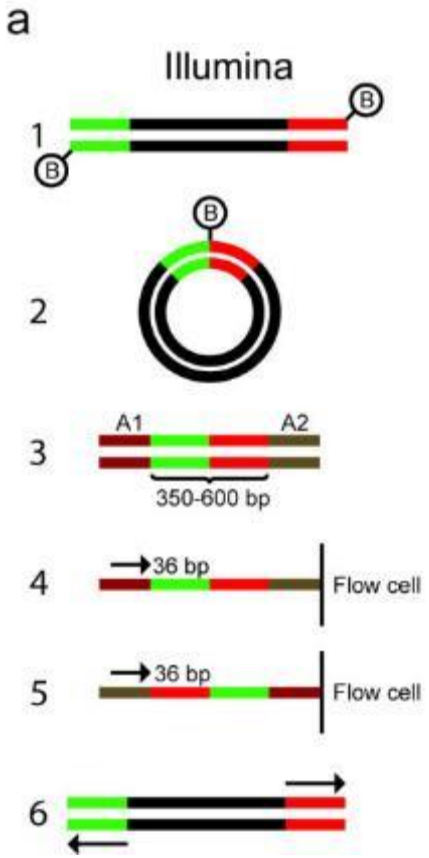
Paired end vs Mate pairs

Library preparations:

Paired-end libraries

- Break the DNA into fragments
- Attach adaptors to both ends
- Sequence from each end one by one

- *Mate-pair libraries*
  - Circularize the DNA using a biotinylated nucleotide
  - Shear and capture biotinylated fragments
  - Previously distant ends are now in proximity
  - Sequence normally



Berglund, E.C., et al. 2011. *Inv Gen*. 2:23

Paired end vs Mate pairs

- **Insert size**: Paired-end libraries have shorter insert size (<1kbp) whereas mate-pairs can have a considerably longer insert size (2-5kbp).

- **Read orientation**: Paired-end reads come in forward-reverse orientation whereas mate-pair reads come in reverse-forward orientation.

Illumina generates (relatively) short reads

- Longest I've seen is 2x375bp on the MiSeq

  That's pretty long considering Illumina machines produced 30bp reads in 2008

- This actually works very well for bacterial genomes – you can get a nearly complete assembly of the bacterial chromosome (and plasmids) from reads of this size.

  I'll show you a very nice assembly of *B. anthracis* using 2x150 bp reads, and that's a pretty big genome for a bacteria

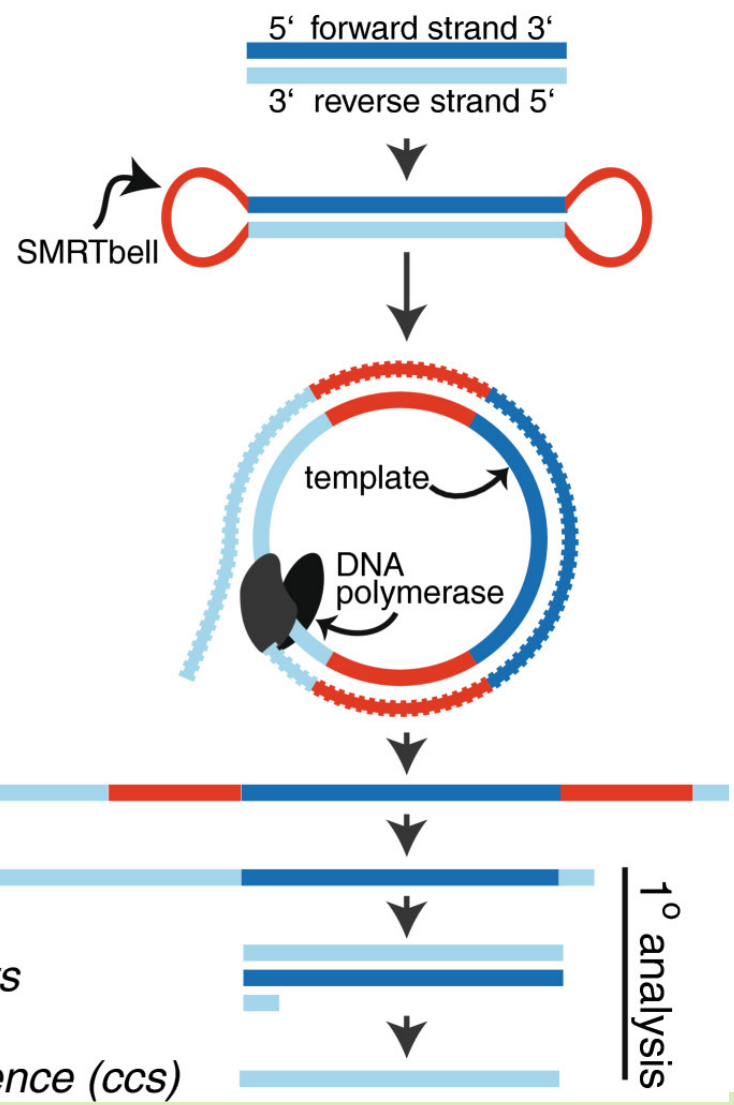
- But what if you want a *complete* genome, no gaps, nothing. One perfect circle?

Pacific Biosciences

- Pacific Biosciences (PacBio) belongs in the third generation of sequencing machines
- Single-molecule real-time (SMRT) sequencing
- Produces considerably long reads and is able to finish genomes by assembly

Pacific Biosciences

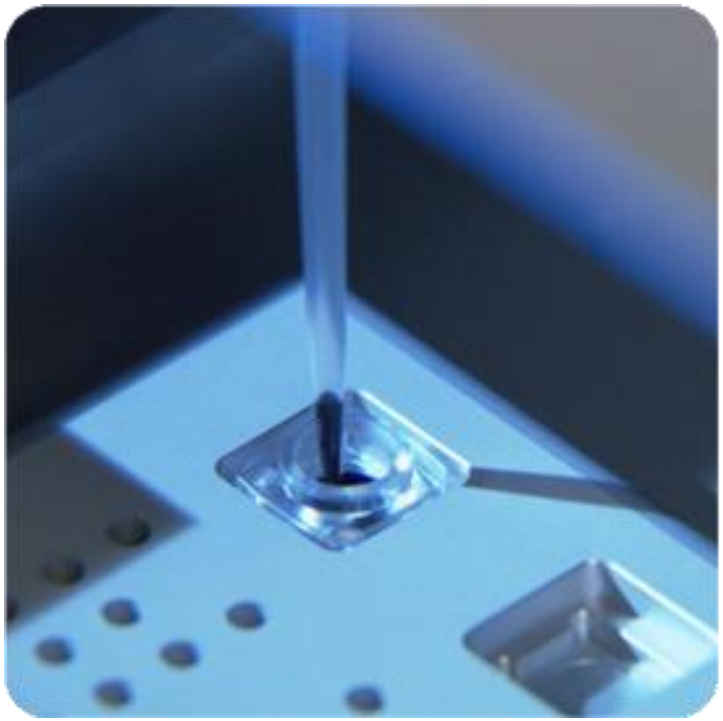
1. generate amplicon
2. ligate adaptors
3. sequence
4. data analysis



Fichot and Norman. 2013. *Microbiome*. 1:10

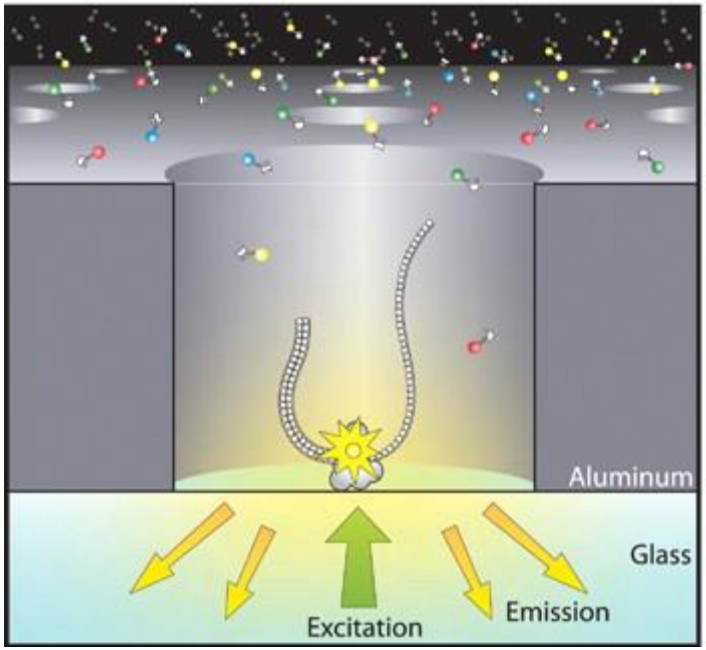
Pacific Biosciences

Pacific Biosciences



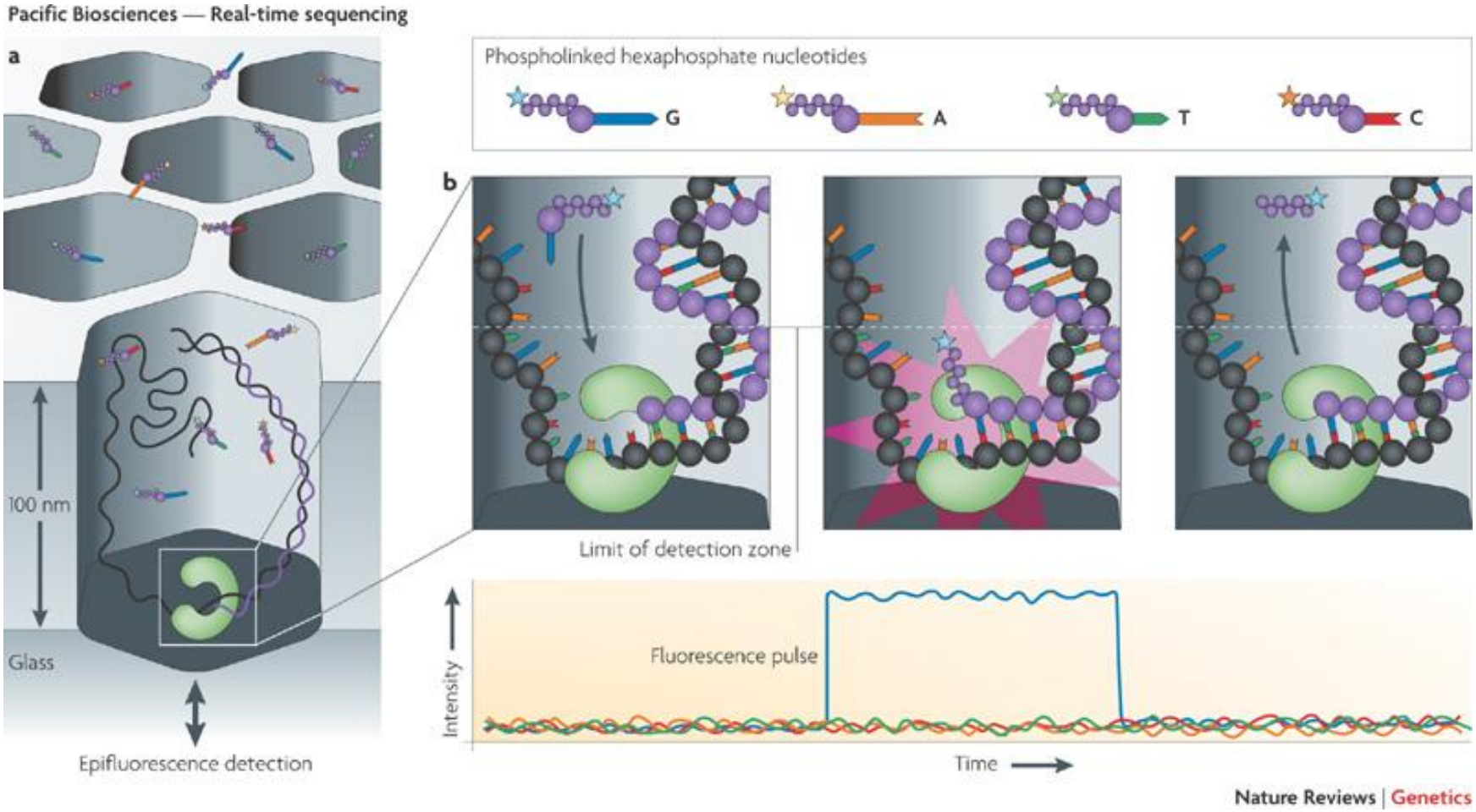
Pacific Biosciences

Zero-mode waveguide (ZMW),
a very fancy and very small well



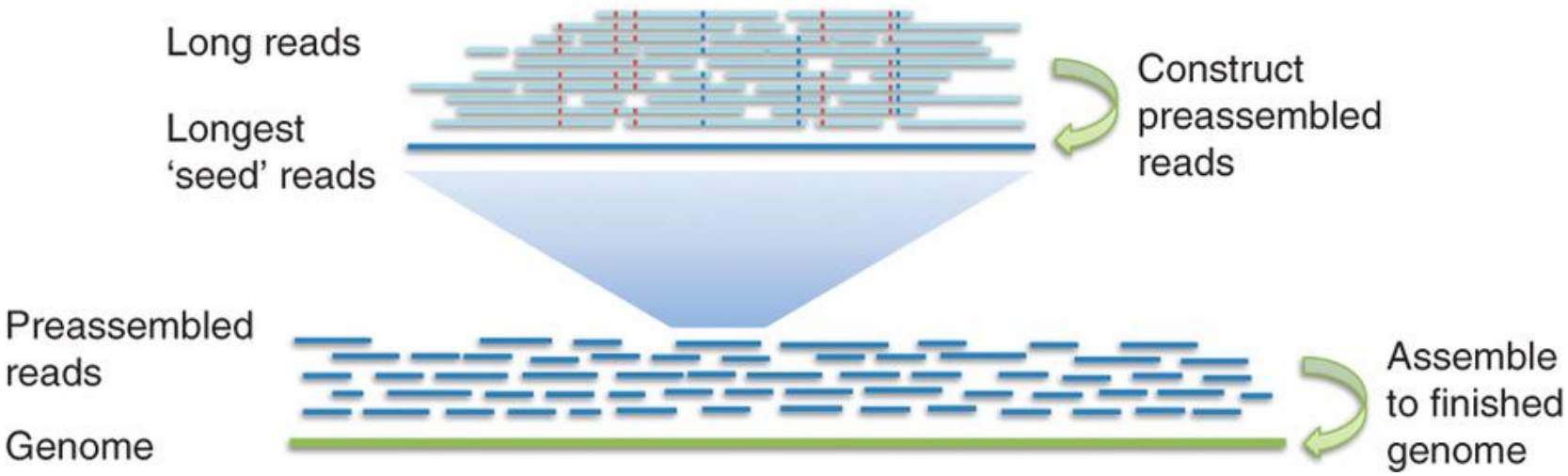
Eid, J. et al. 2009. *Science*. 323:133-8

Pacific Biosciences



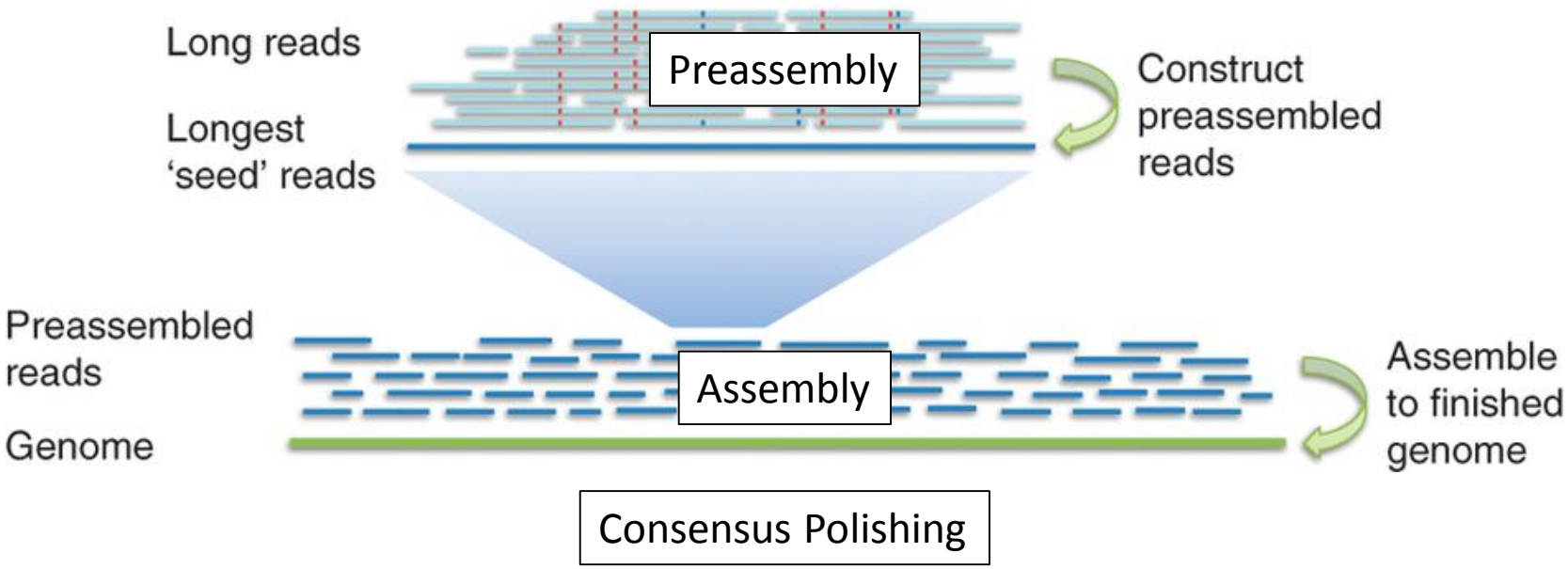
Metzker, M.L. 2010. *Nat Rev Gen*. 11:31-46

HGAP Assembly



Chin, C.S. et al. 2013. *Nat Met.* 10: 563–569

HGAP Assembly



Chin, C.S. et al. 2013. *Nat Met.* 10: 563–569

Side note: HGAP has had issues with plasmids

- The HGAP assembly we recently received was missing any plasmid sequences which were know to be in the isolates
- The plasmids were confirmed by Sanger, and resistances, so they had to be there
- A second, related assembly, the FALCON assembler, was used and found the plasmids in ~1/2 of the isolates
- This may have been an issue of library preparation or size selection; better size selection and HGAP may have been just fine.

PacBio Sequencing generates *very* long reads

- Upwards of 20,000bp v. 2x300bp reads from Illumina
  - But not nearly as many reads produced
  - Also more error-prone reads, *i.e.* the bases read are more likely to be incorrect
- The longer reads can yield substantially better assemblies
  - We've completed several *B.anthracis* genomes using a combination of PacBio and Illumina sequencing

Which one do you pick?

- We'll talk more about sequencing depth/coverage later

- Basically:

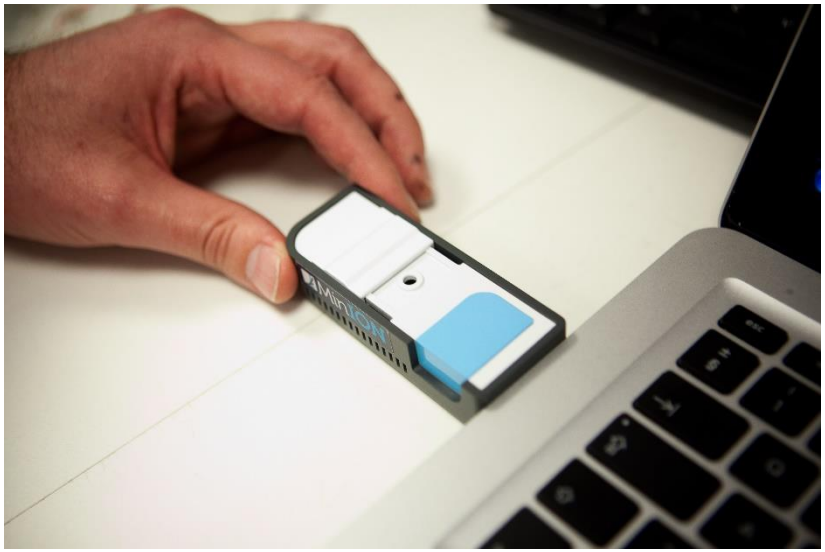
  Always do paired-end reads. There's never a good reason not to in modern times.

  Don't buy a new machine. Everybody around you has one and they're not using it right now.

  If you *have* to buy a machine, buy a MiSeq. None of you are doing human whole-genome sequencing

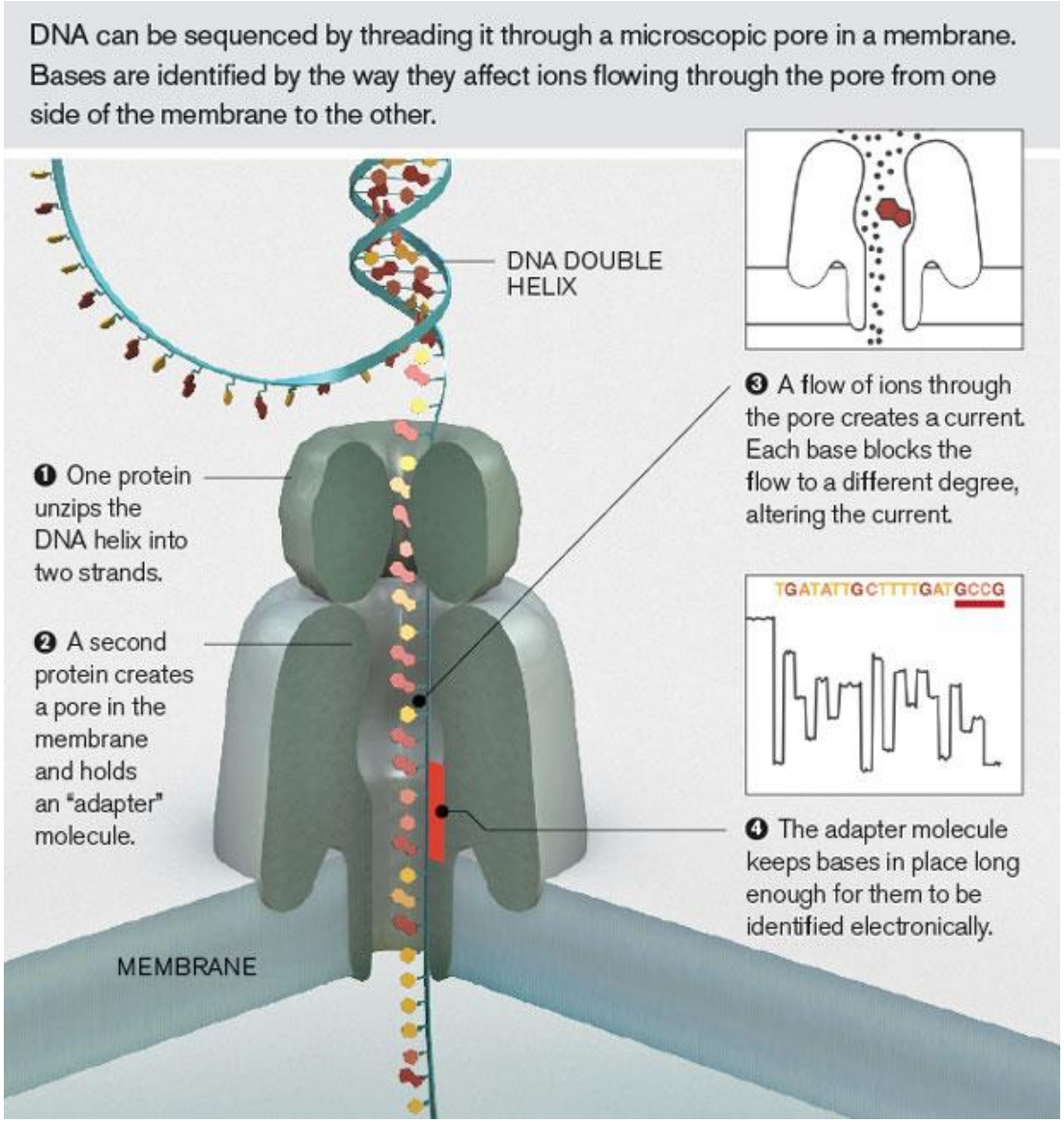
  Do PacBio sequencing if having *complete* genome is important

Oxford Nanopore



- Oxford Nanopore is another upcoming technology that sequences a DNA/RNA molecule using a nanopore mounted on a synthetic polymer

Oxford Nanopore



Oxford Nanopore

- The MinION is perhaps the most exciting of their products which lets you sequence and analyze the data right on your laptop

Outline

- DNA and genome sequencing technology
- Genome sequence data and quality
- Genome assembly
  - Reference assembly
  - *De novo* assembly
- Assembly quality

Sequencing platforms produce two things for each base they see

1- Just the actual base itself. What wavelength did the machine pick up when the reads were high with a light source?

2-A set of quality scores. One per base call generated

Always have the exact same number of quality scores as you have basecalls

These two things come as a FASTQ file

- FASTA-like with quality scores
- You've probably all gotten a big FASTQ file back from your sequencing people at some point
- If you look at the contents of the file, it probably looks mostly like gibberish

Format Description: FASTQ

```
@HWE1FGTJ-GH13-454470/1
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAA
+
!''*((((***+))%%%++)(%%%%).1***-+*'')
```

Read Identifier Line

Sequence

Description Line

Encoded Quality Scores

Decoding Quality

- Quality scores are encoded as their representative ASCII values with some offset
- E.g. Sanger quality is often represented as offset by 33
- If you have a quality of 30, the quality with offset will be 30 + 33 = 63
- Its corresponding ASCII character is the symbol '?'

ASCII Encoding

Decimal	Hex	Char	Decimal	Hex	Char	Decimal	Hex	Char	Decimal	Hex	Char
0	0	[NULL]	32	20	[SPACE]	64	40	@	96	60	`
1	1	[START OF HEADING]	33	21	!	65	41	A	97	61	a
2	2	[START OF TEXT]	34	22	"	66	42	B	98	62	b
3	3	[END OF TEXT]	35	23	#	67	43	C	99	63	c
4	4	[END OF TRANSMISSION]	36	24	$	68	44	D	100	64	d
5	5	[ENQUIRY]	37	25	%	69	45	E	101	65	e
6	6	[ACKNOWLEDGE]	38	26	&	70	46	F	102	66	f
7	7	[BELL]	39	27	'	71	47	G	103	67	g
8	8	[BACKSPACE]	40	28	(	72	48	H	104	68	h
9	9	[HORIZONTAL TAB]	41	29	)	73	49	I	105	69	i
10	A	[LINE FEED]	42	2A	*	74	4A	J	106	6A	j
11	B	[VERTICAL TAB]	43	2B	+	75	4B	K	107	6B	k
12	C	[FORM FEED]	44	2C	,	76	4C	L	108	6C	l
13	D	[CARRIAGE RETURN]	45	2D	-	77	4D	M	109	6D	m
14	E	[SHIFT OUT]	46	2E	.	78	4E	N	110	6E	n
15	F	[SHIFT IN]	47	2F	/	79	4F	O	111	6F	o
16	10	[DATA LINK ESCAPE]	48	30	0	80	50	P	112	70	p
17	11	[DEVICE CONTROL 1]	49	31	1	81	51	Q	113	71	q
18	12	[DEVICE CONTROL 2]	50	32	2	82	52	R	114	72	r
19	13	[DEVICE CONTROL 3]	51	33	3	83	53	S	115	73	s
20	14	[DEVICE CONTROL 4]	52	34	4	84	54	T	116	74	t
21	15	[NEGATIVE ACKNOWLEDGE]	53	35	5	85	55	U	117	75	u
22	16	[SYNCHRONOUS IDLE]	54	36	6	86	56	V	118	76	v
23	17	[ENG OF TRANS. BLOCK]	55	37	7	87	57	W	119	77	w
24	18	[CANCEL]	56	38	8	88	58	X	120	78	x
25	19	[END OF MEDIUM]	57	39	9	89	59	Y	121	79	y
26	1A	[SUBSTITUTE]	58	3A	:	90	5A	Z	122	7A	z
27	1B	[ESCAPE]	59	3B	;	91	5B	[	123	7B	{
28	1C	[FILE SEPARATOR]	60	3C	<	92	5C	\	124	7C	|
29	1D	[GROUP SEPARATOR]	61	3D	=	93	5D	]	125	7D	}
30	1E	[RECORD SEPARATOR]	62	3E	>	94	5E	^	126	7E	~
31	1F	[UNIT SEPARATOR]	63	3F	?	95	5F	_	127	7F	[DEL]

Offsets and Quality Ranges

Technology	Quality Type	Offset	Quality Range
Sanger	Phred	33	[0,40]
Solexa	Solexa	64	[-5,40]
Illumina 1.3+	Phred	64	[0,40]
Illumina 1.5+	Phred	64	[3,40]
Illumina 1.8+	Phred	33	[0,41]
IonTorrent	Phred	33	[0,40]

Quality control: What is it?

- What is QC in the case of NGS data?

- Ensuring the best possible quality of raw sequencing data for downstream analyses:

  Removing sequencing adapters/primers

  Trimming low quality bases

  Removing PCR duplicates

- ***You have to assume that the source of the FASTQ data was lying when they said they did these things***

- QC can't fix problems in the input DNA or gross mistakes in the sequencing process

Quality control: Why?

- Why do we care about the quality of sequencing libraries? What can go wrong if we don't QC well?
- It ***greatly*** affects the downstream analysis
- Lingering adapters -> poor genome assembly and mapping
- Low quality bases/PCR duplicates -> poor variant calling
- Simply, you have to make sure that your experiment worked

Sequencing adapters and barcodes

- Sequencing adapters are vital to the sequencing process, but they are bad for downstream analysis – they aren't part of the genome and won't align correctly

- Adapters and barcodes in sequencing data

```
@HWI-123-23:i3
GATGATATTTGACTATGAGT
+
FFIIIIFFI[[AIIFIFFFI
@HWI-123-67:i3
GATGATTTATCCGGTCGGTAGGA
+
FFIIIIFFI[[AIIFIFFFI(())
```

Removing PCR duplicates

- ***Ideally*** 1 DNA/RNA fragment -> 1 read. Rarely the case unless you start with a huge excess of DNA/RNA.
  - You typically have to amplify your input DNA using PCR
  - These duplicated sequences don't represent genuine DNA fragments
  - Really important for calling variants.
  - What if you have a bad segment that just happens to be amplified a whole bunch of times?

Removing PCR duplicates

- It's very easy to generate false positive variant calls from PCR artifacts
  We'll come back to this

- Single-cell sequencing really fell prey to this issue
  You're only starting with one copy! You have to amplify!
  Any PCR mistake early in the amplification will be carried through!

Removing PCR duplicates

- ***Of all the things in QC, duplicates are the most application-specific***

- In RNA-seq, you very rarely remove duplicates unless you know you have a very good reason too

  Doing odd experiments with very low input, for example

- In ChIP-seq, you almost always remove PCR duplicates

FastQC:

Babraham Bioinformatics

About | People | Services | Projects | Training | Publications

FastQC

Function	A quality control tool for high throughput sequence data.
Language	Java
Requirements	A suitable Java Runtime Environment The Picard BAM/SAM Libraries (included in download)
Code Maturity	Stable. Mature code, but feedback is appreciated.
Code Released	Yes, under GPL v3 or later.
Initial Contact	Simon Andrews

http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

FastQC is a common tool for NGS QC

- ***Every single*** FASTQ file I receive goes through FastQC
- Pretty much always my first step in any NGS analysis I do
- It will analyze, but won't fix problems in your data. There are a bunch of other tools for that

- FastQC will check for:

  Regions of poor quality in the flowcell

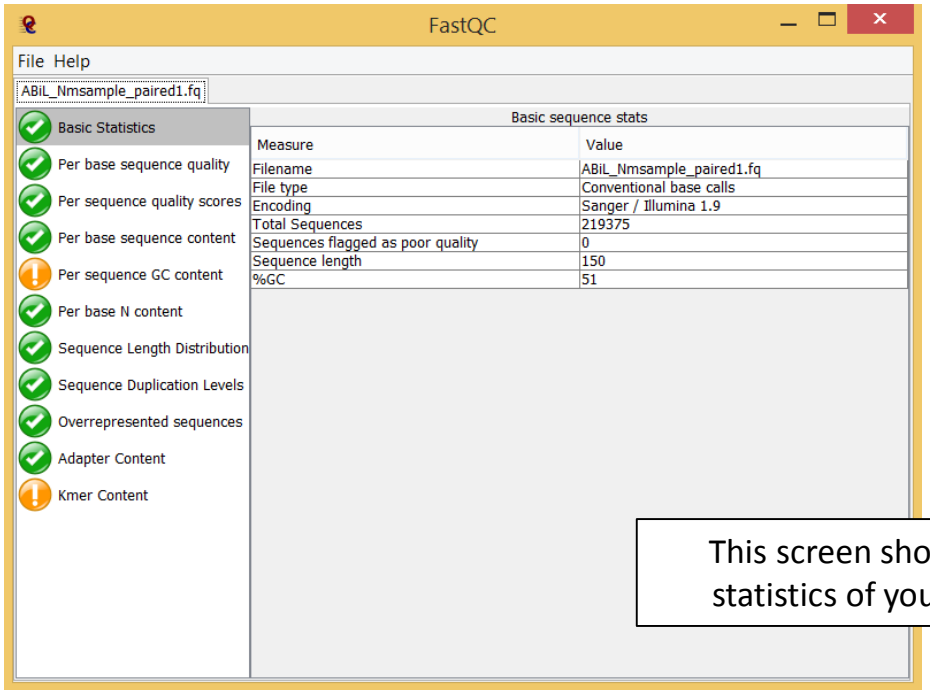
  Overrepresented sequences

  Adapters and primers

  Sequence duplication levels

  GC biases

Quality Assessment

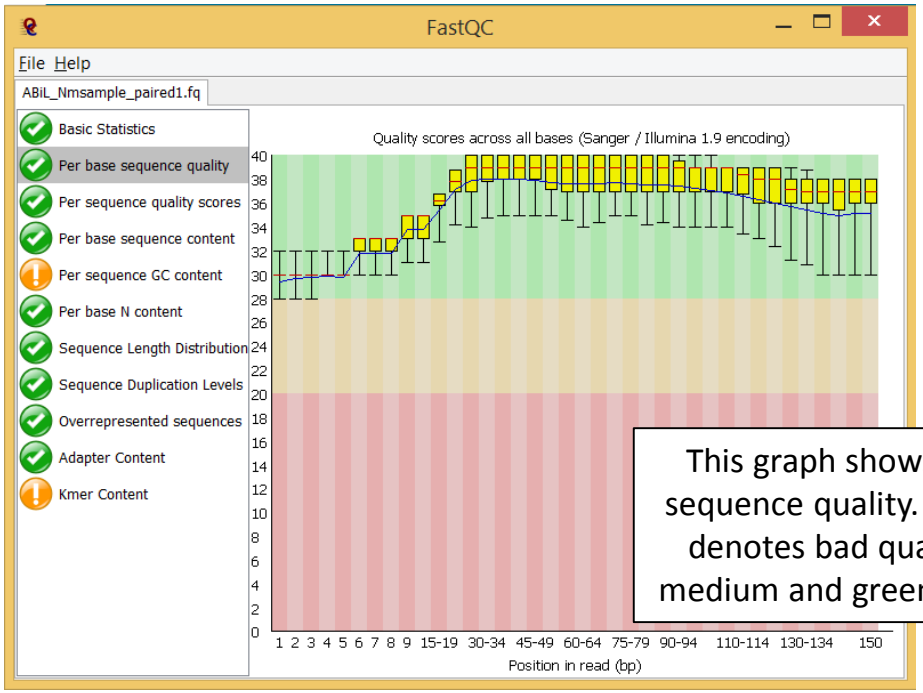


This screen shows the basic statistics of your input data

Genome coverage

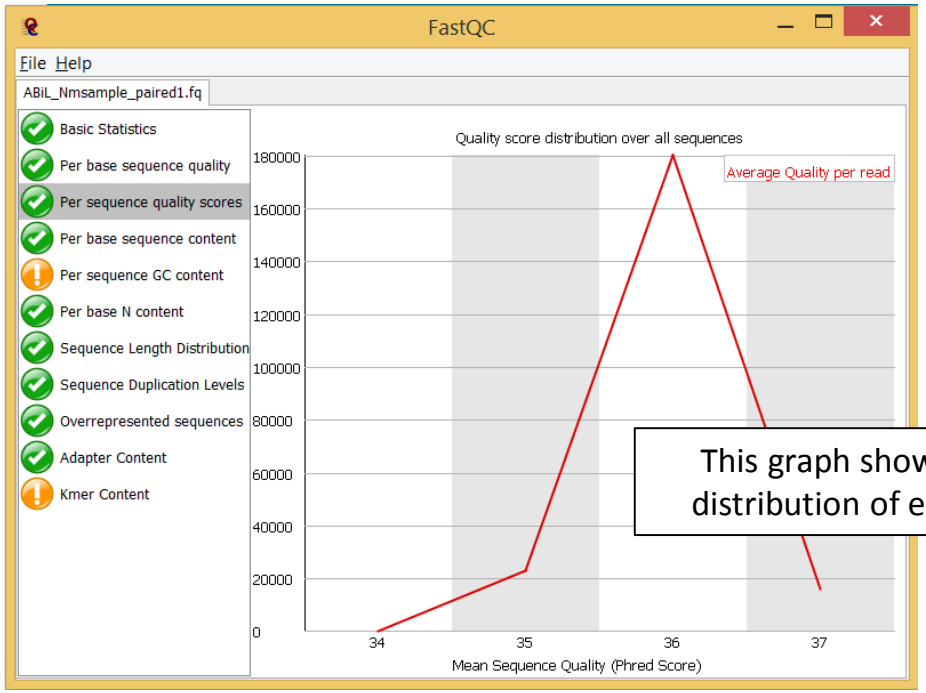
- Genome coverage refers to the average number of times each base is sequenced
- It is usually represented with an ‘x’ at the end
- If each base in the genome is sequenced once, the genomic coverage is 1x
- If each base is covered twice, the coverage is 2x and so on
- A coverage of 30x and above is considered good coverage for haploid genomes, *e.g.*, bacteria

Basecall quality



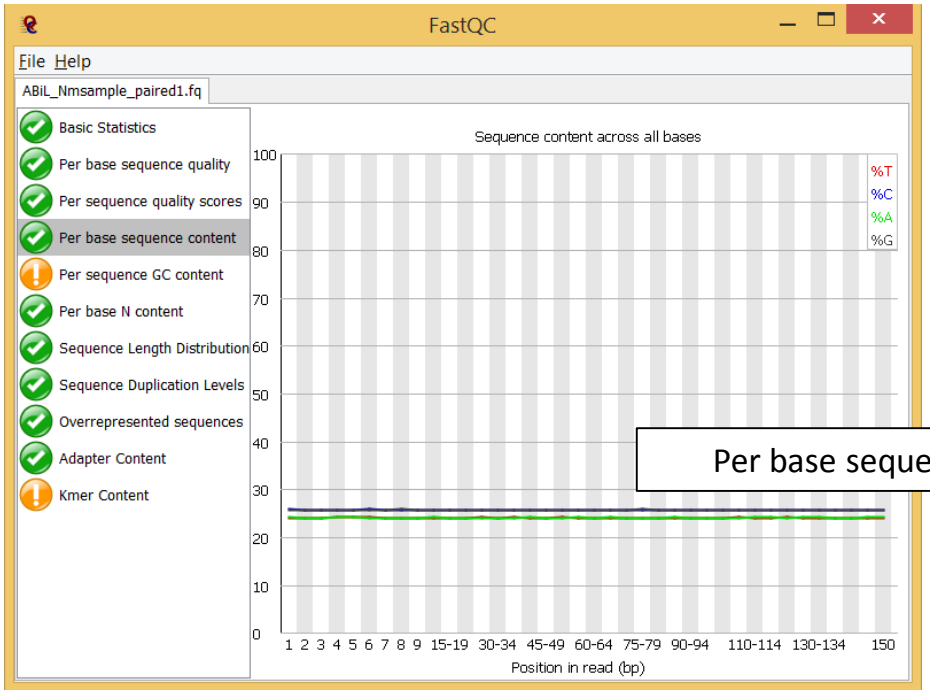
This graph shows the per base sequence quality. The red region denotes bad quality, orange is medium and green is good quality

Quality Assessment



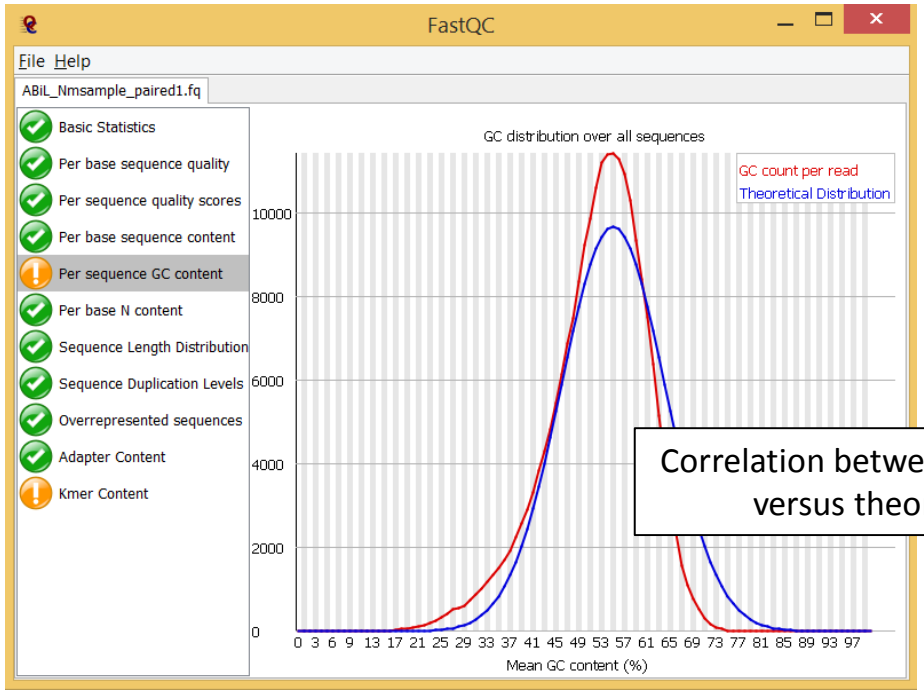
This graph shows the quality distribution of each sequence

Quality Assessment



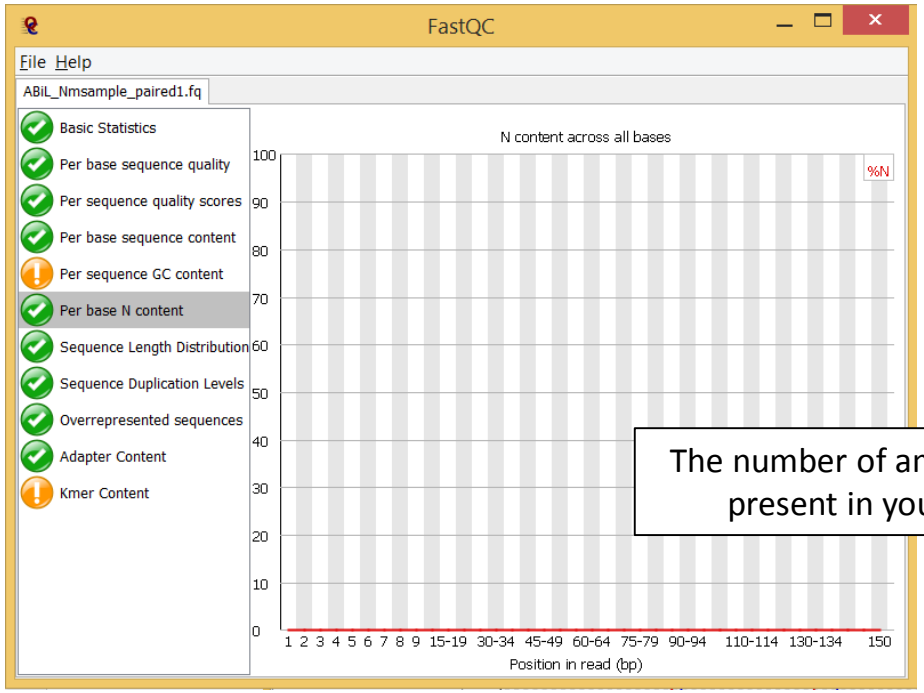
Per base sequence content

Quality Assessment



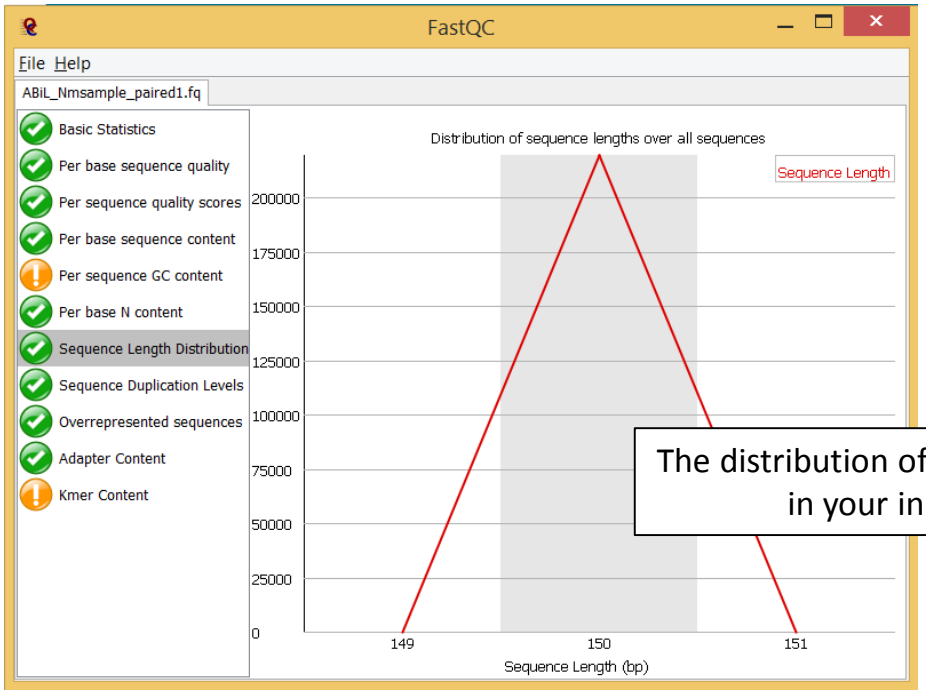
Correlation between observed GC versus theoretical GC

Quality Assessment



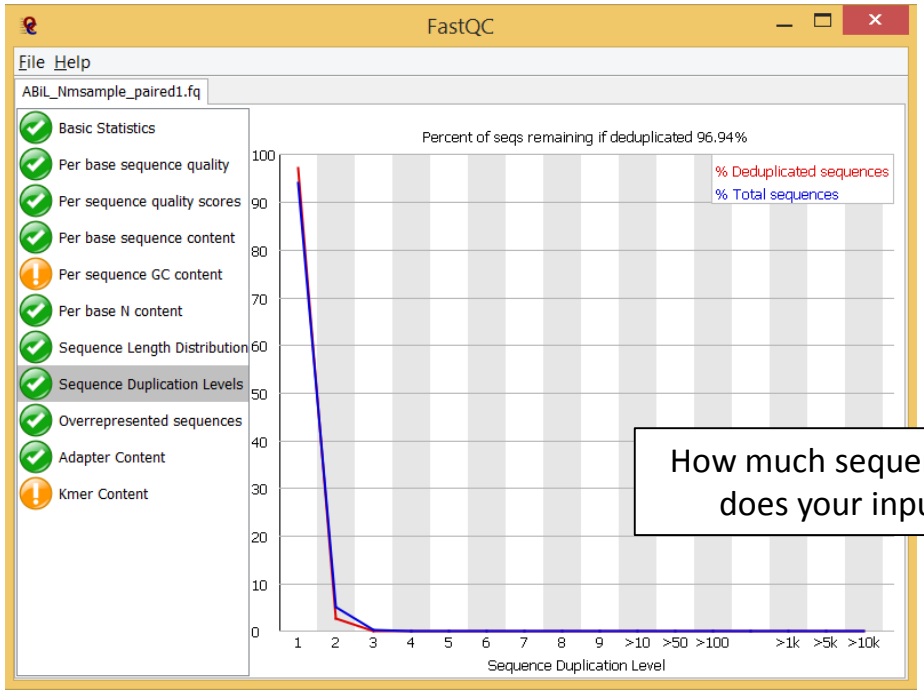
The number of ambiguous bases present in your input file

Quality Assessment



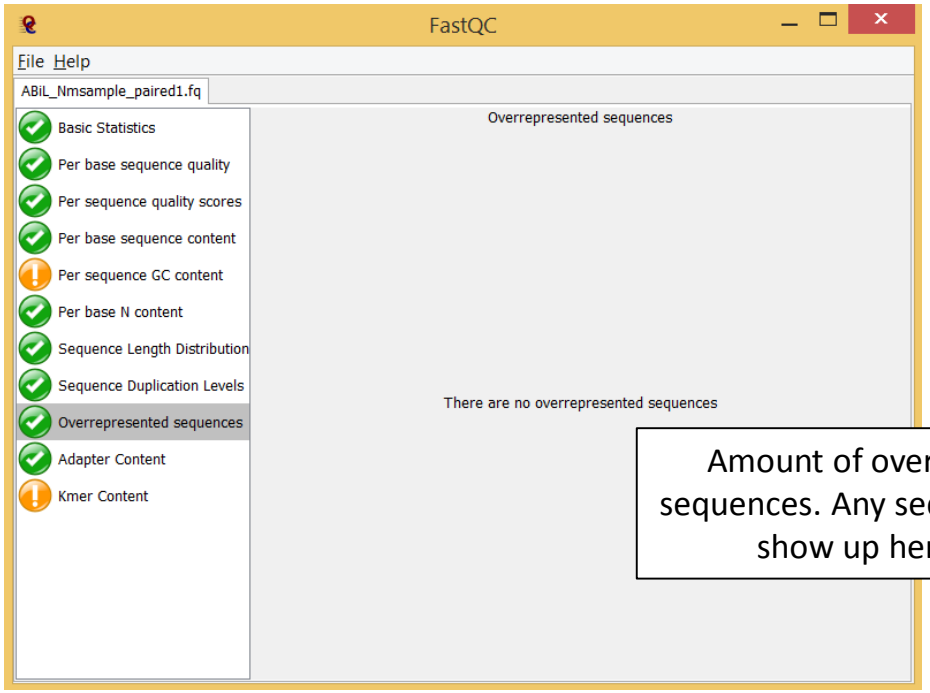
The distribution of the read length in your input file

Quality Assessment



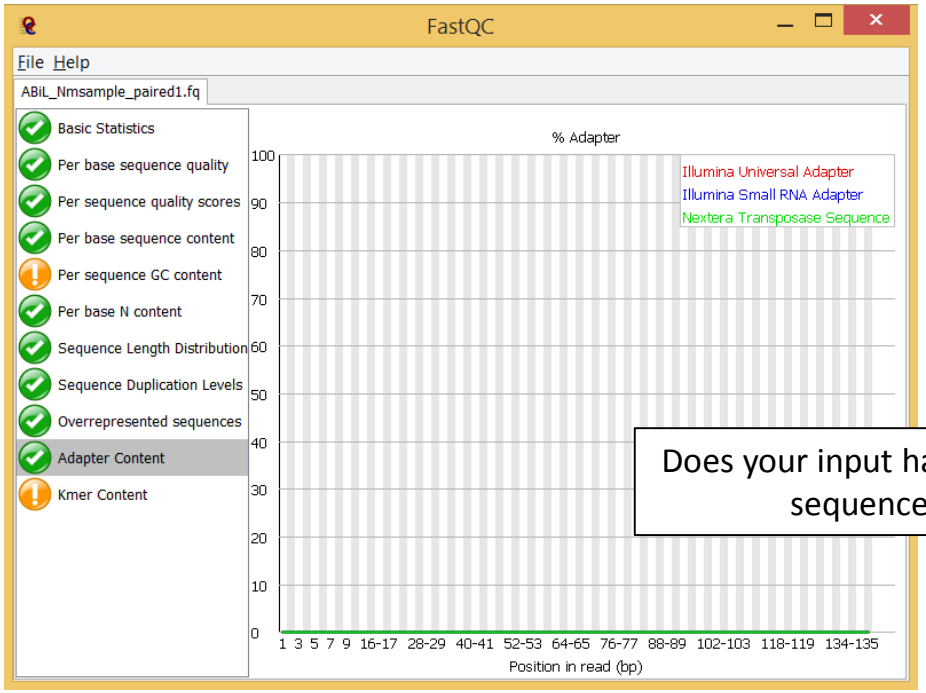
How much sequence duplication does your input file have?

Quality Assessment



Amount of overrepresented sequences. Any sequence bias will show up here as well

Quality Assessment

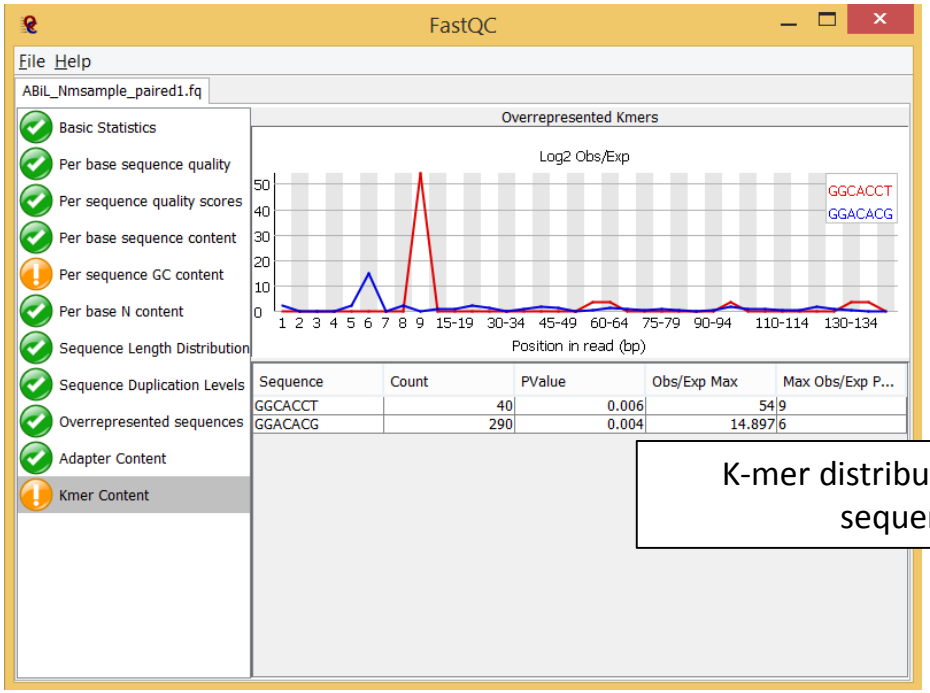


Does your input have any adapter sequences in it?

K-mers

- Sequences of length k (= 1, 2, 3, …)
- All possible DNA 1-mers or monomers are A, C, G and T
- All possible DNA 2-mers or dimers are AA, AC, AG, AT, CA, CC, CG, CT, GA, GC, GG, GT, TA, TC, TG and TT
- There are 64 possible 3-mers or tri-mers and 4^k possible k-mers

Quality Assessment

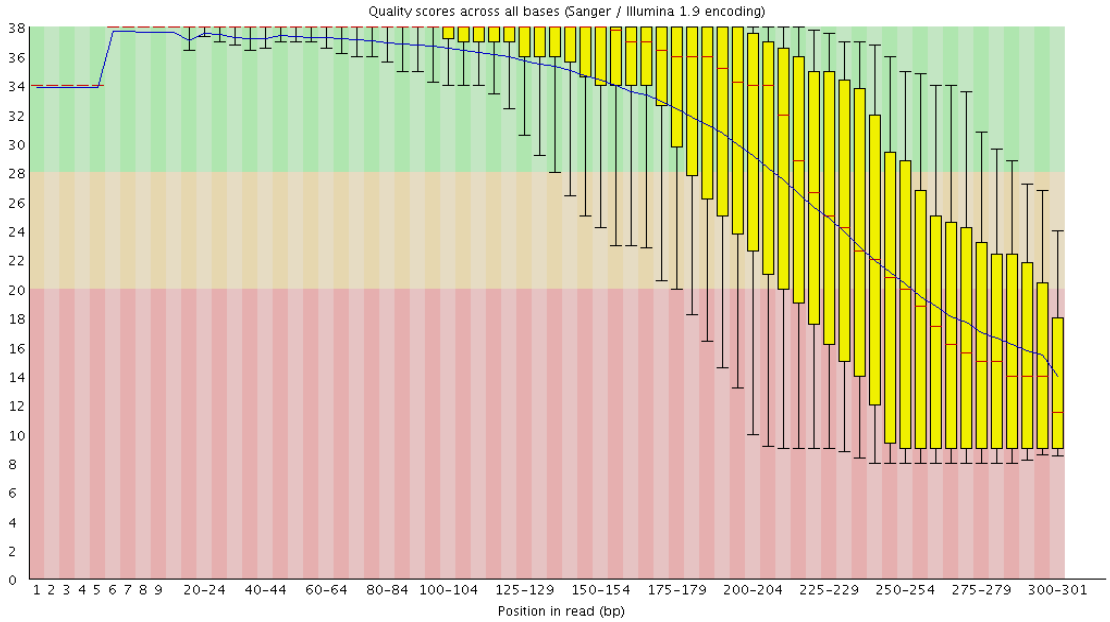


K-mer distribution in your sequence

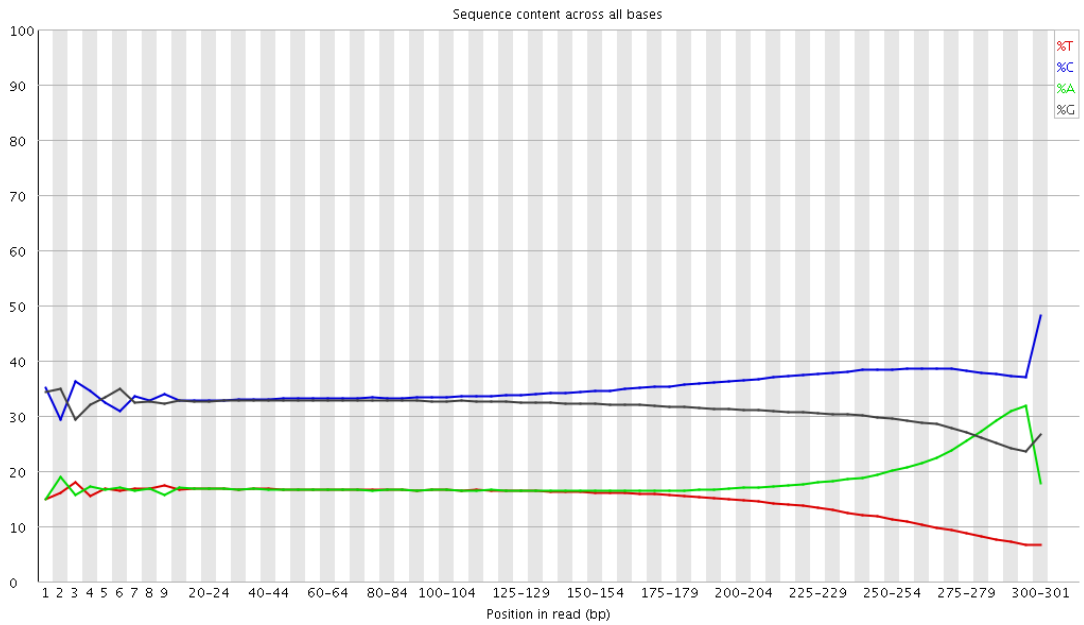
That was a *lovely* read set

- They're rarely ever that pretty in reality
- Here's one we got off of a MiSeq with the V3 chemistry a few months ago…
- Actually the third read set from V3 chemistry we've seen that didn't look so great

Bad 3’-end quality scores



Huge sequencing biases

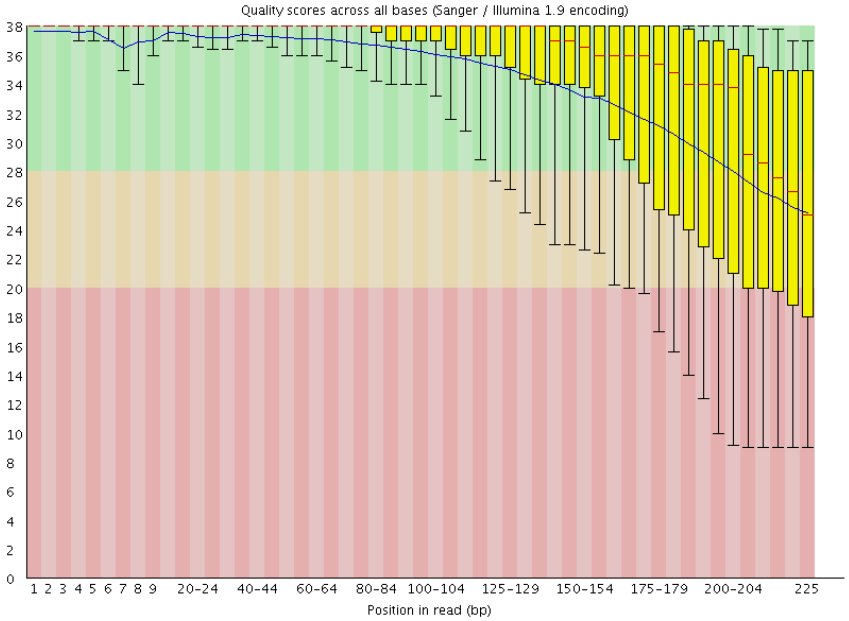


Bad fragment lengths

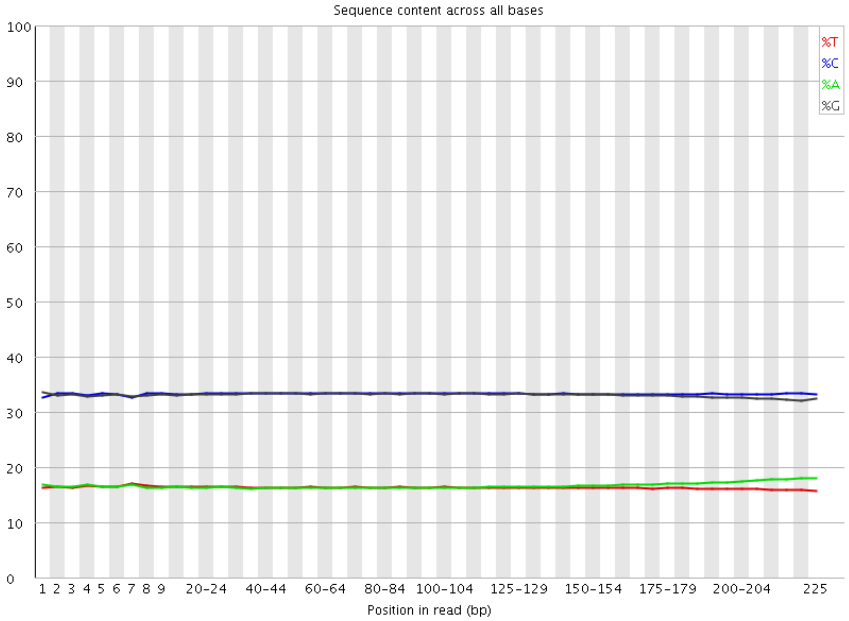
Long fragments generate true paired-end reads

Short fragments generate overlappping reads

If we chop the reads up, we do get some improvement



If we chop the reads up, we do get some improvement



Your bioinformatician should *always* do a QC run

- Skipping the QC run is ***inviting*** disaster
- No matter where the reads came from, you have to do QC
- You cannot assume that they did QC (or if they did, that they did a good job at it)
- There are many downstream QC checks that pertain to your individual datatype and analyses

Outline

- DNA and genome sequencing technology
- Genome sequence data and quality
- Genome assembly
  - Reference assembly
  - *De novo* assembly
- Assembly quality

The Sequencing Problem

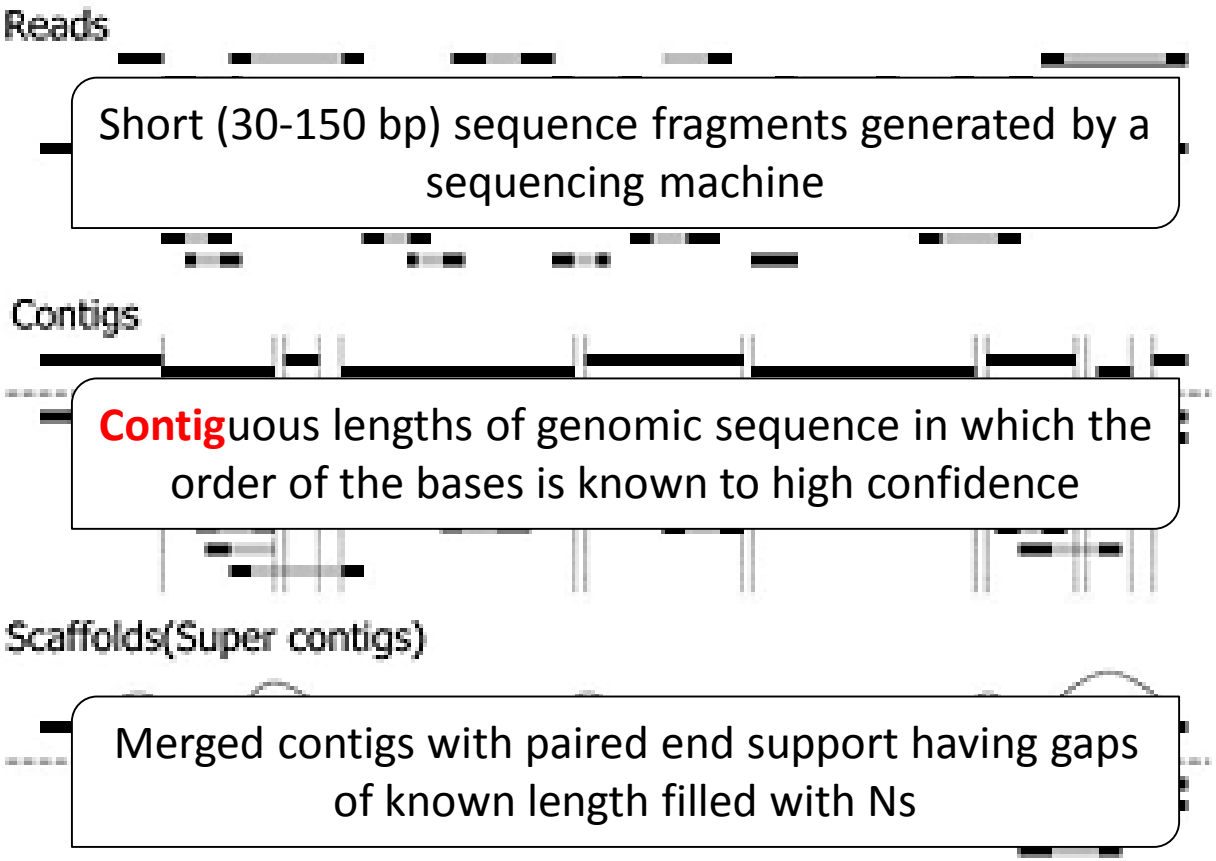
- Present day DNA sequencing technology can’t read the whole genome in a single pass
- To be able to read (sequence) the genome, it first needs to be fragmented into small pieces

The Sequencing Problem

- This leads to the need to put the pieces together (assemble) to recreate the whole genome

- To help ourselves, we first make multiple copies of the genome and then break it randomly – which leads to the **assembly problem**

Terminology



Adopted from: https://www.k.u-tokyo.ac.jp/pros-e/person/shinichi_morishita/genome-assembly.jpg

The Assembly Problem

Microbial Genome

Amplification

Fragmentation & Sequencing of Reads

Assembly

Expectation

The Assembly Problem

Microbial Genome

Amplification

Fragmentation & Sequencing of Reads

Assembly

Reality

The Assembly Problem

- This “recreation” or assembling of the genome is not straight-forward
- Consider this example:

It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness

Amplification and fragmentation

It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness

It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness

It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness

It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness

The Assembly Problem

it was the best of

was the best of times,

the best of times, it

best of times, it was

of times, it was the

of times, it was the

times, it was the worst

times, it was the age

Repeat issue

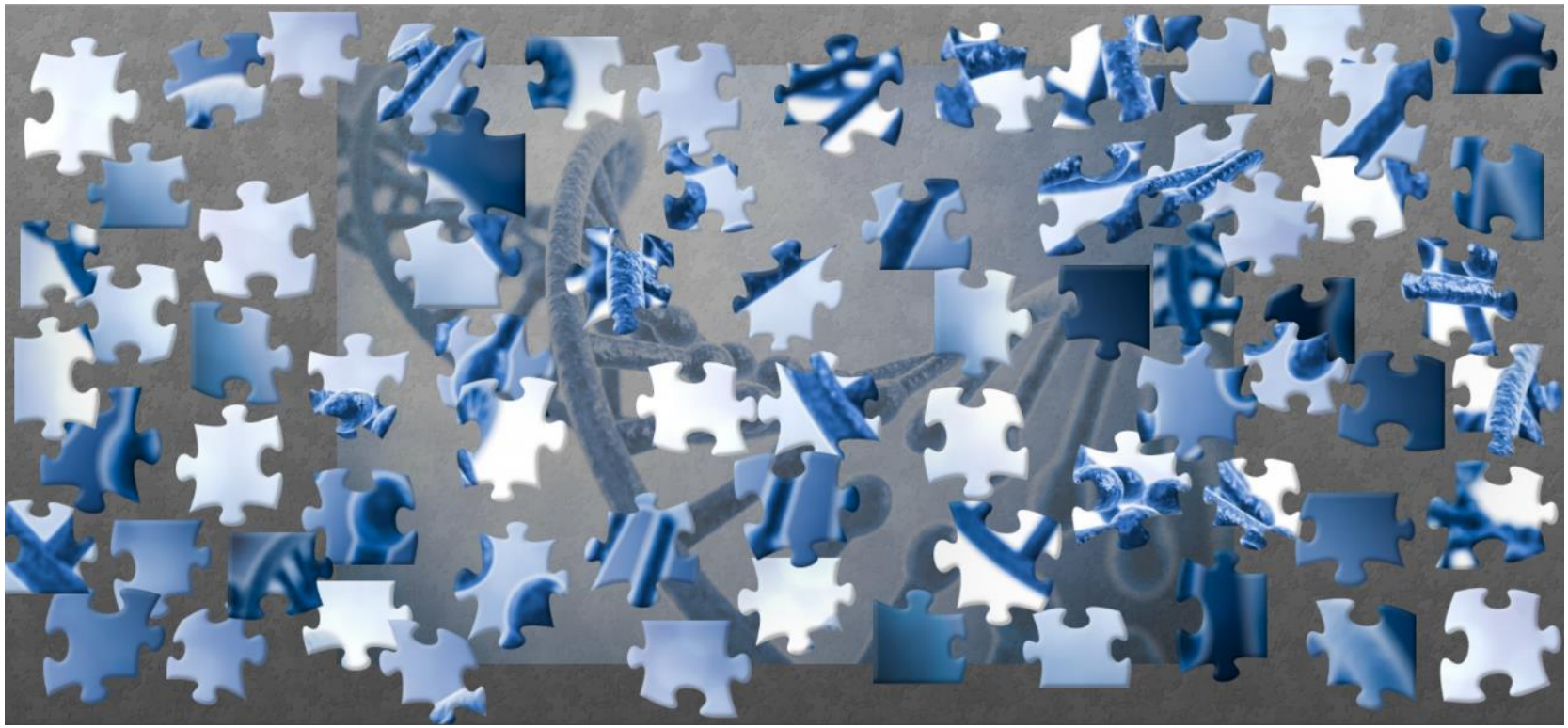
Multiple paths

Outline

- DNA and genome sequencing technology
- Genome sequence data
- Genome assembly
  - Reference assembly
  - *De novo* assembly
- Assembly quality

Reference Assembly (i.e., Genome Mapping)

If our sequencing reads are the jigsaw pieces…



Genome Mapping

… genome mapping is putting the pieces on top of the complete picture

Genome Mapping

Reference genome / transcriptome

```
...GTGGGCCGGCAATTCGATATCGCGCATATATTTCGGCGCATGCTTAGC...
```

Reads (unmapped)

```
1 GCATATATTT
2 GCATATATTT
3 TGGGCCGGCA
4 ATTCGATATC
5 ATATTTCGGC
6 CCGGCAATTC
7 TCGCGCATAT
8 CATGCTTAGC
9 GATATCGCGC
```

More specifically, genome mapping is the mapping of reads to a reference genome

Genome Mapping

Reference genome / transcriptome

```
...GTGGGCCGGCAATTCGATATCGCGCATATATTTCGGCGCATGCTTAGC...
    TGGGCCGGCA              GCATATATTT     CATGCTTAGC
        CCGGCAATTC              ATATTTCGGC
              ATTCGATATC  GCATATATTT
                      TCGCGCATAT
                  GATATCGCGC
```

Reads
(mapped)

More specifically, genome mapping is the alignment of reads to a reference genome

Genome Mapping

- Genome mapping is preferable to *de novo* assembly if a good reference genome exist
- This makes genome analysis faster than *de novo* as the annotation can be directly transferred from the reference genome
- Ideal for larger organisms such as fungi and humans
- Fails to detect large insertion/deletion/rearrangement events

Applications

Genome mapping has many applications, such as:

- **Genome Resequencing**
  Rapid assembly of genomes using existing reference genome

- **Variant Calling**
  Calling variants in highly similar strains

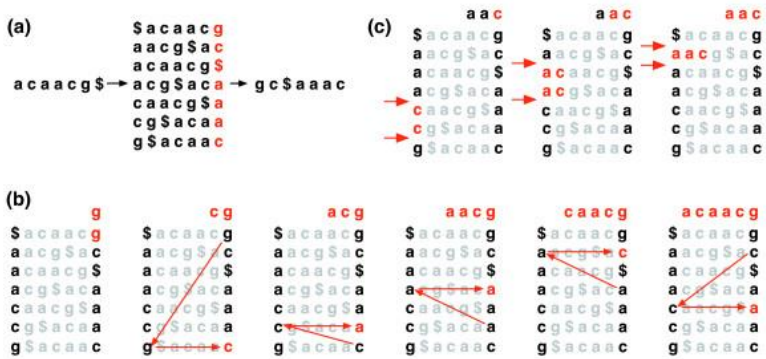
- **Population Genomics**
  Analyzing populations based on mapping and variants

- **Genome Alignment and Comparison**
  Assessing similarity at genomic level between different species/genus

Genome Mapping Paradigms

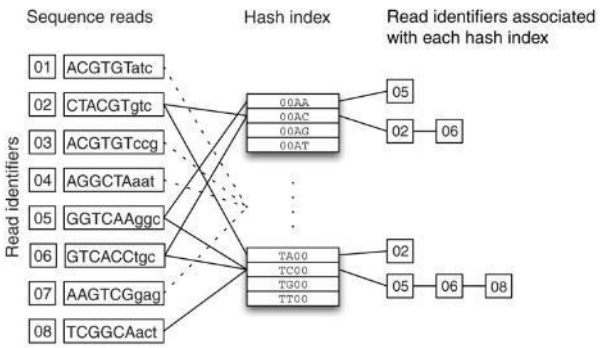
- Similar to *de novo* paradigms, there are two popular paradigms in genome mapping:

Burrows-Wheeler Transform



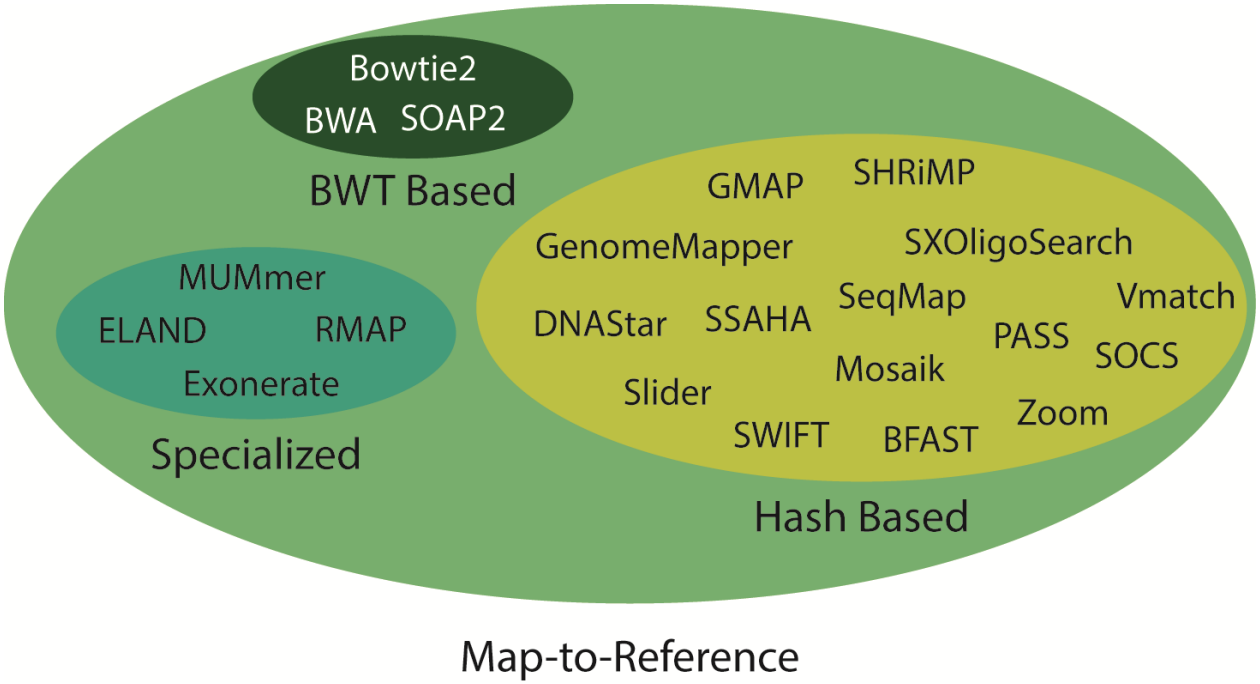
Langmead et al. 2009. *Genome Biol.* 10:R25

Hash Based

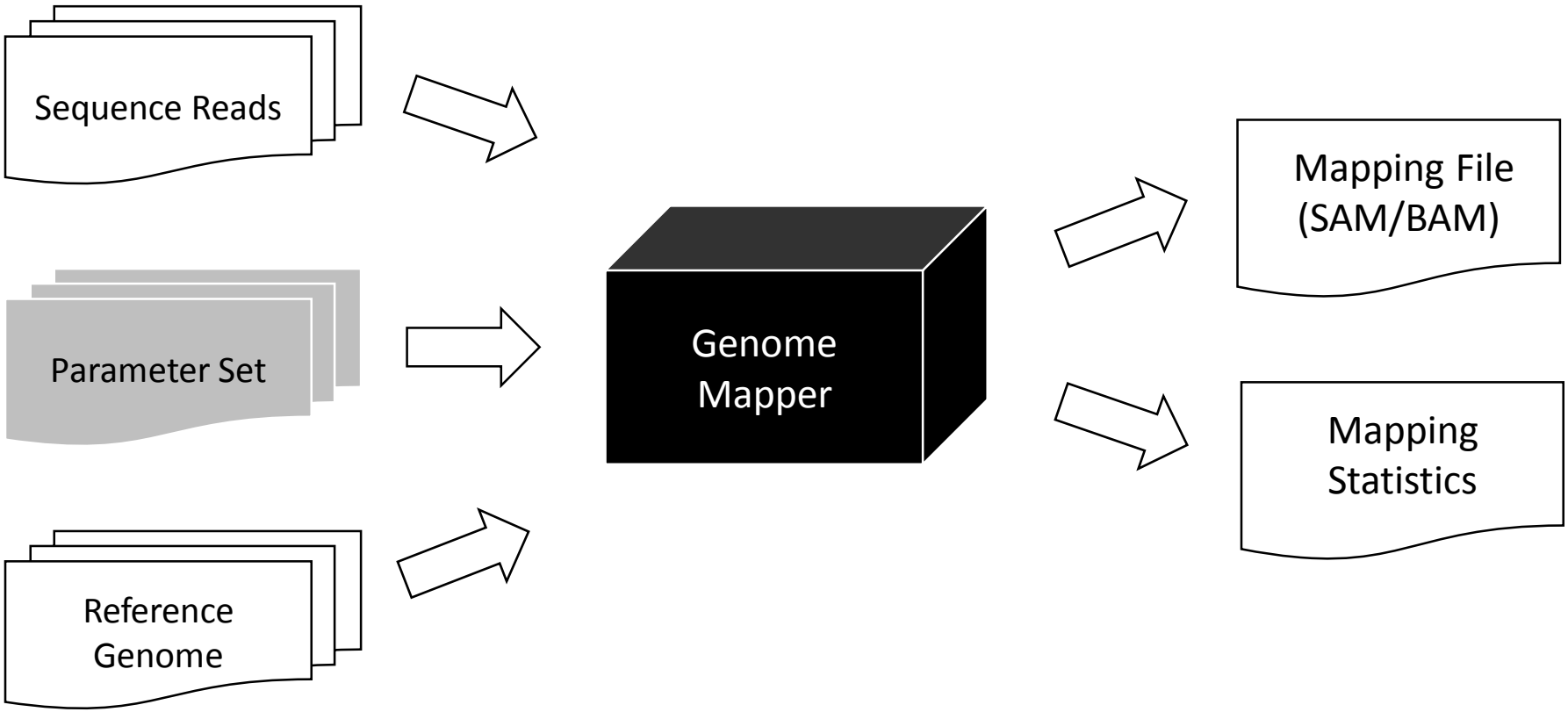


Flicek et al. 2009. *Nature Methods.* 6:S6-S12

Genome Mapping Programs



Genome Mapping Programs



SAM format

- Short for Sequence Alignment/Map format
- It is a tab-delimited **text** format
- Defines how each read is mapped/unmapped to which section of the genome
- Consists of optional header section followed by alignment section

SAM format

```
@SQ   SN:FAM18_C_genome   LN:2194961
@PG   ID:bwa    PN:bwa    VN:0.7.10-r789 CL:bwa mem -t 4 genome.fasta r1.fq r2.fq
ABiLNm-SM1-438990   99   FAM18_C_genome 840075    60   150M =    840445    520
GCATTGATTTTTCAATACGGAATCGATGATGTTAATGCAAGCACCGCGACTACCGTCGTCAGCAGCTGTGCCGCAATATCGAAACCGGGCAAGCAAATCCCTACTTTAGAAGATGCAAAAAAAGAATTGAAGATTCCGGATCAGGATAAG
??????BBDDBDDDBEG>GFGGIHIE>H/IIFHHHIHIHIIHHIHHHEHHHHFGIDHIHHHHHI?HHIIGIHHHIEC@IHHCIIIIFHHH@IGIHGHDFHGD@IHICH8HFHHHGHHHFEFFBGEHHG,EEFEFE*EFGEEG8EGF(GC
NM:i:0    MD:Z:150  AS:i:150  XS:i:0
ABiLNm-SM1-438990   147  FAM18_C_genome 840445    60   150M =    840075    -520
AATATACGGATAAATTCGACAGCTCCACAAATGCTGTTACGCCCGCCGGGGTGGAGGTTTTATTGAGTAGCGGTACTGATACCAAGATTGCCGCTTCTTCAGACAATCGTATTTATGCTTACCGTATCGATGCGACAATACGCGGGGGAA
-.AGG*GEFEEEEG<(*GFEEGEHEFEG,*EDEDHDEFHBHHEGHEGHEH+HDIHHGFFHH8E7IHF@H-HD?+5=+HHHHAIGFDFHHEHHAHHHIEEFIHB-+FHD7GI=HHAIH9A7AIIFIFHHFFCGGGDDB9BDD<BB=5??A?
NM:i:1    MD:Z:108A41    AS:i:145  XS:i:0
```

The FLAGs describe how the reads are being mapped

MAPQ describes how good the mapping is

CIGAR describes the summary of mapping

Col	Field	Type		f description
1	QNAME	String	[!-?A-~][1,255]	Query template NAME
2	FLAG	Int	$[0,2^{16}-1]$	
3	RNAME	String	*|[!-()+-<>-~][!-~]	
4	POS	Int	$[0,2^{31}-1]$	POSition
5	MAPQ	Int	$[0,2^{8}-1]$	
6	CIGAR	String	*|([0-9]+[MIDNSHPX	
7	RNEXT	String	*|=|[!-()+-<>-~][!-~]*	ad
8	PNEXT	Int	$[0,2^{31}-1]$	Position of the mate/next read
9	TLEN	Int	$[-2^{31}+1,2^{31}-1]$	observed Template LENgth
10	SEQ	String	*|[A-Za-z=.]+	segment SEQuence
11	QUAL	String	[!-~]+	ASCII of Phred-scaled base QUALity+33

BAM format

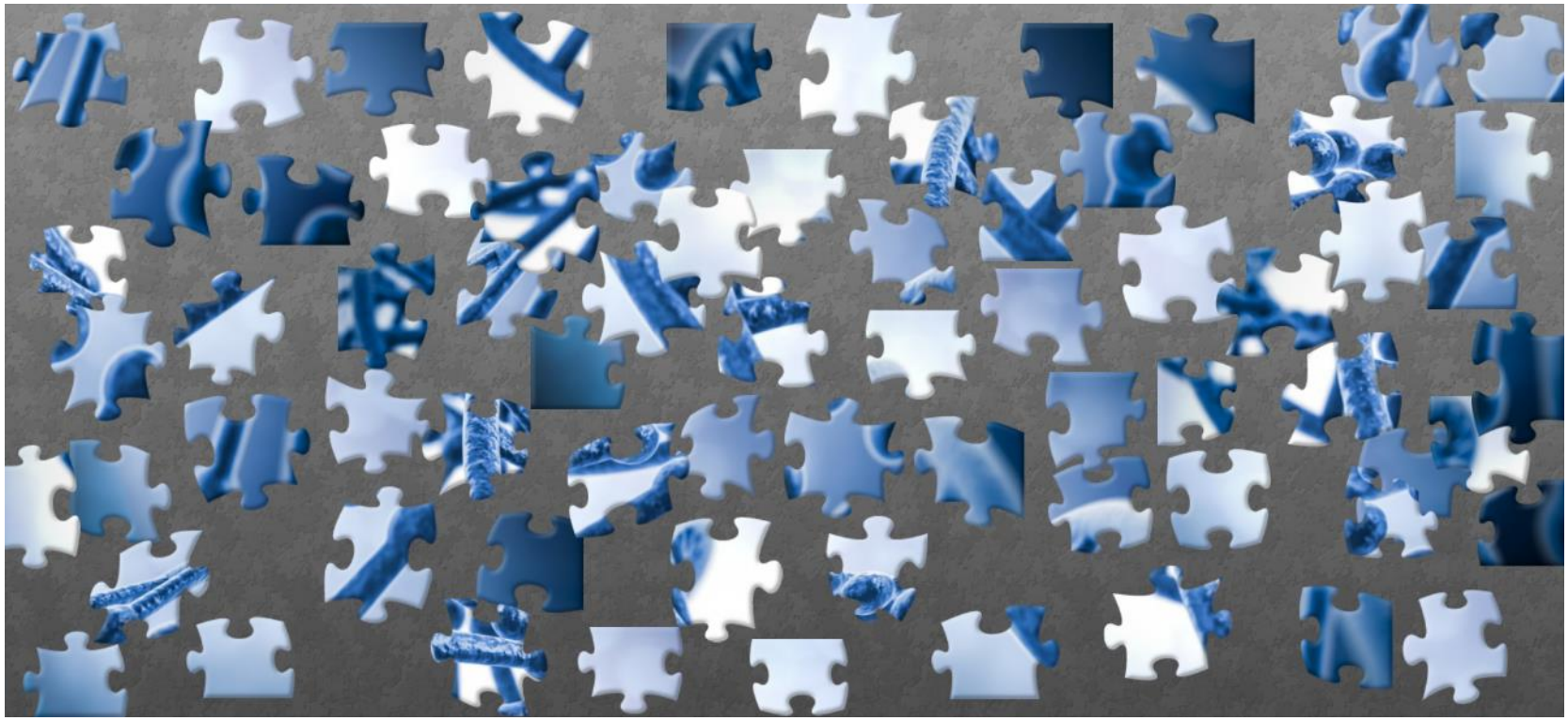
- Stands for Binary Alignment/Map format
- Binary version of the SAM file – cannot be read by normal text editors
- Special decompressors/decoders are required
- Compressed and faster to read/process

Outline

- DNA and genome sequencing technology
- Genome sequence data and quality
- Genome assembly
  - Reference assembly
  - *De novo* assembly
- Assembly quality

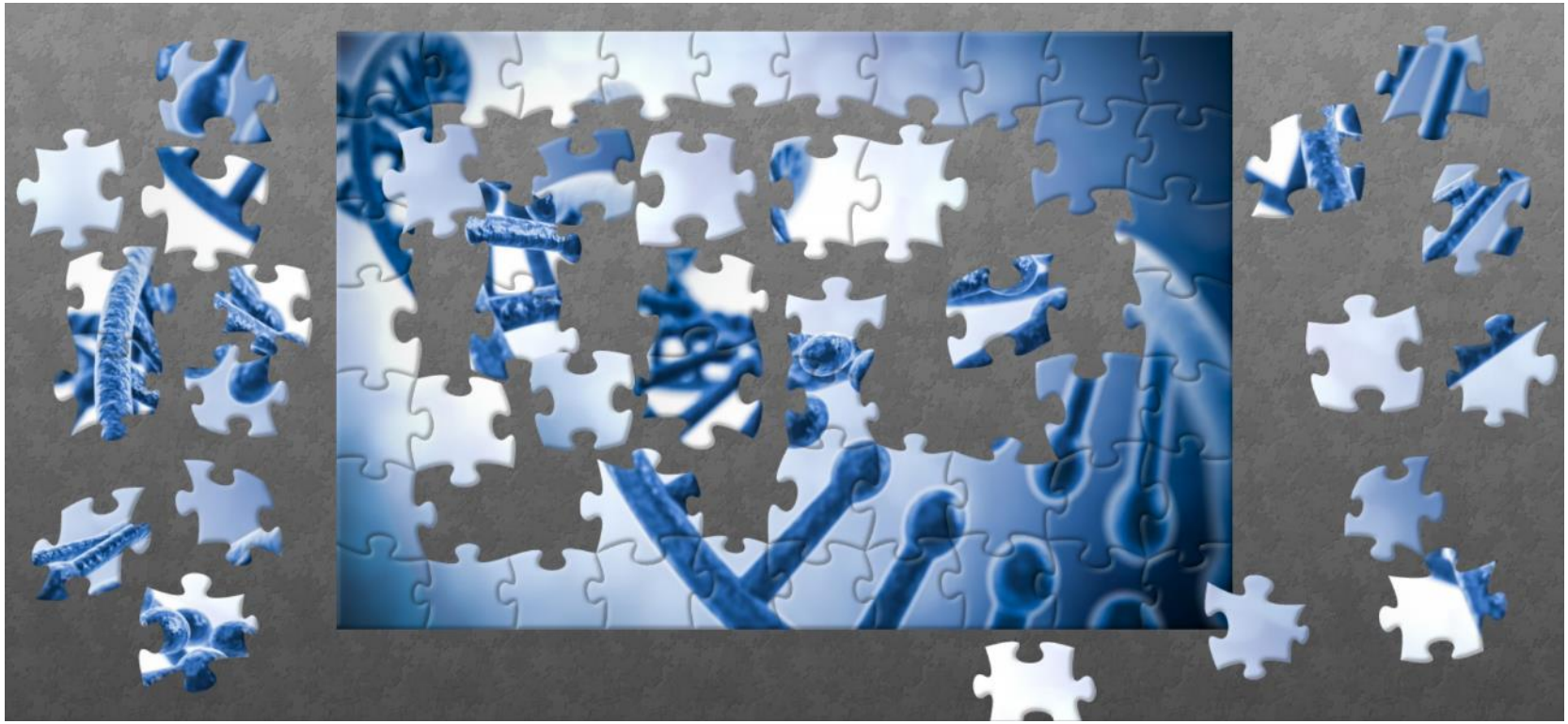
De novo Assembly

If the sequencing reads are considered jigsaw pieces…



De novo Assembly

… a *de novo* assembly is trying to construct the jigsaw puzzle without the picture



De novo Assembly

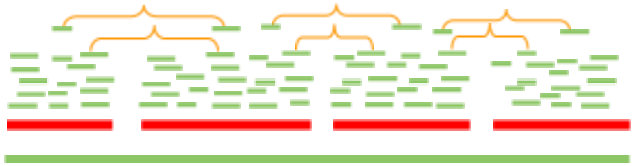
- i.e., trying to construct the genome based on the sequencing reads without the aid of a reference genome
- This process is naturally computationally intensive and requires substantial expertise to yield “good” assemblies
- In a way, this is more of an art than a computational process

Assembly steps

i. Find overlapping reads

ii. Merge good pairs of reads into longer contigs

iii. Link contigs to form supercontigs



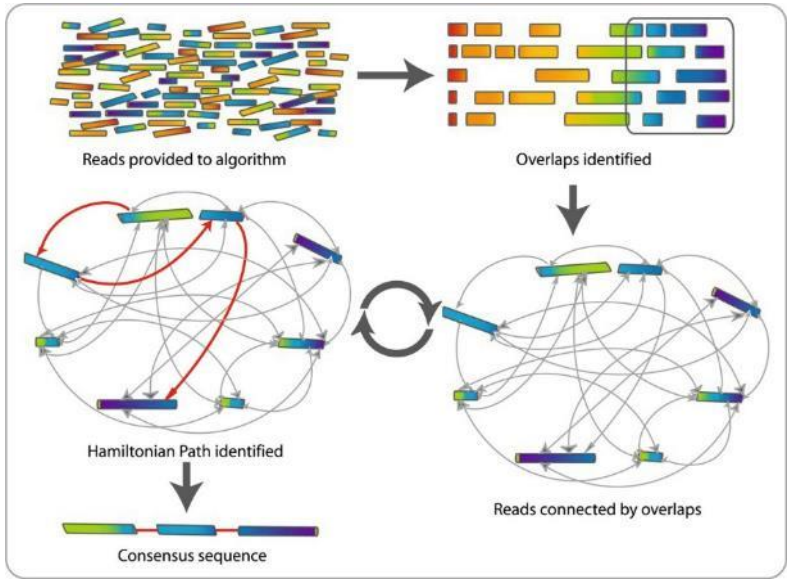
iv. Derive consensus sequence

..ACGATTACAATAGGTT..

Assembly paradigms

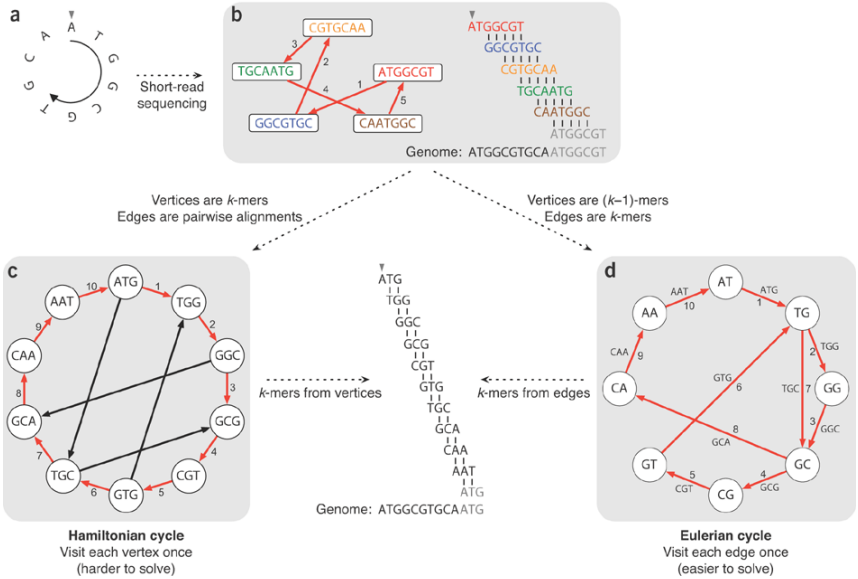
- A number of paradigms have been proposed for genome assembly, the notable of which are:

Overlap-Consensus-Layout



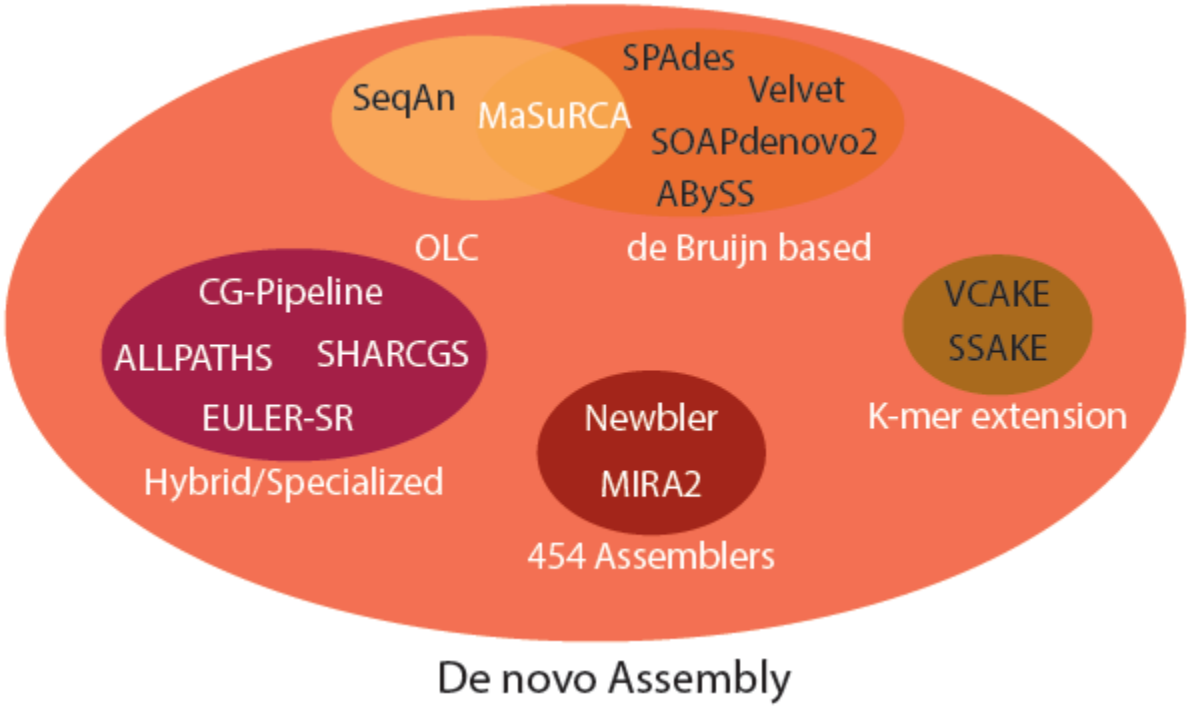
Commins et al. 2009. *Biol Proc Online*. 11:52-78

de Bruijn Graph



Compeau et al. 2011. *Nat Biotech*. 29:987-991

Genome Assembly Programs



Genome Assembly Programs

- The assembler's performance depends on the input data:
  - Organism
  - Sequencing platform
  - Sequencing chemistry
  - Sequencing quality
  - Assembly parameters

Genome Assembly Programs

- It is difficult to pick a best assembler for any assembly project
- A good approach for ensuring good assemblies is to be agnostic
- i.e., try as many as assemblers as possible and then choose the best resulting assembly

De novo Genome Assembly Programs

What is GAGE?

GAGE is an evaluation of the very latest large-scale genome assembly algorithms. We have organized this "bake-off" as an attempt to produce a realistic assessment of genome assembly software in a rapidly changing field of next-generation sequencing. The main results of GAGE have now been published in the journal Genome Research: GAGE: A critical evaluation of genome assemblies and assembly algorithms.

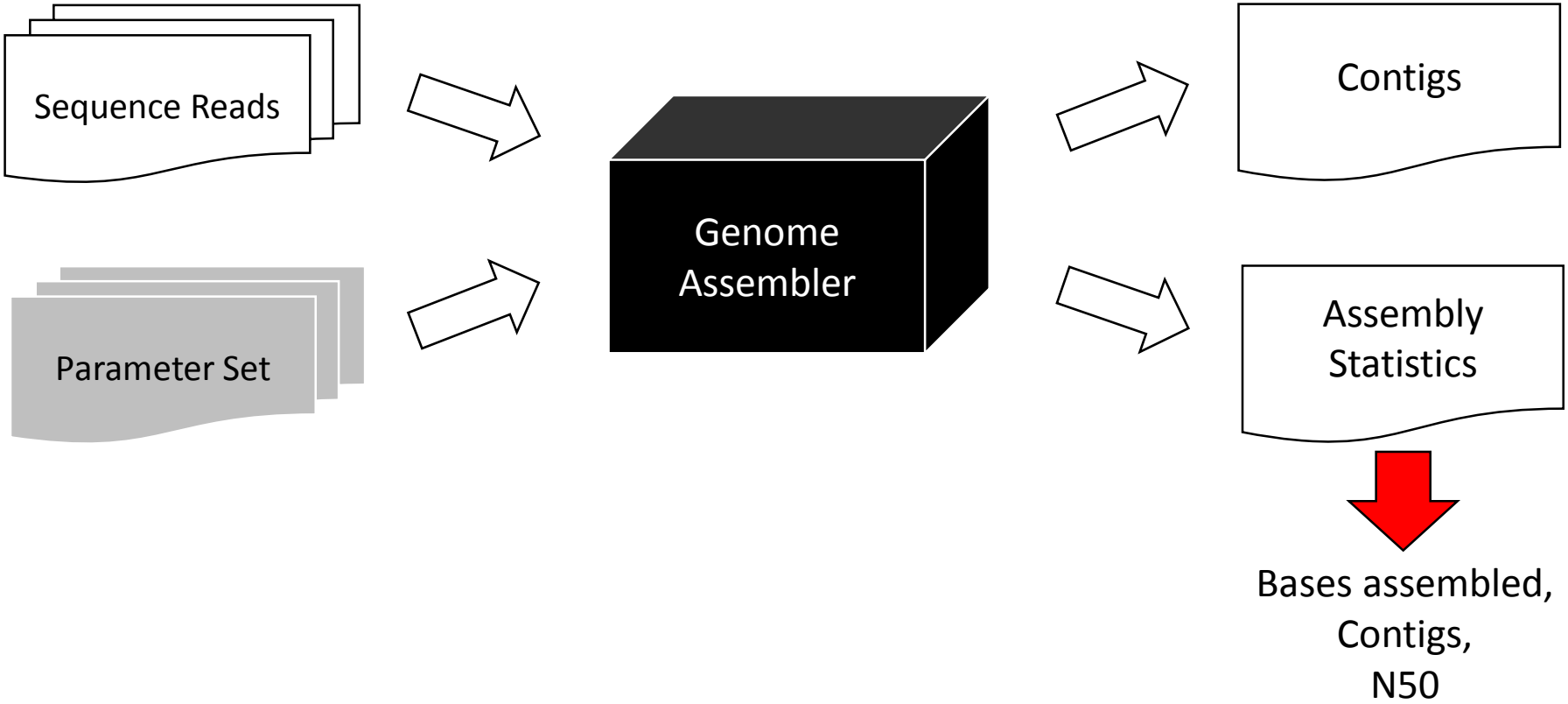
CSH PRESS GENOME RESEARCH

Assemblathon 1: A competitive assessment of de novo short read assembly methods

Dent A. Earl, Keith Bradnam, John St. John, et al.

Genome Res. published online September 16, 2011
Access the most recent version at doi:10.1101/gr.126599.111

Genome Assembly Programs



Outline

- DNA and genome sequencing technology
- Genome sequence data and quality
- Genome assembly
  - Reference assembly
  - *De novo* assembly
- Assembly quality

Assembly Statistics

What statistics can help us in assessing assemblies?

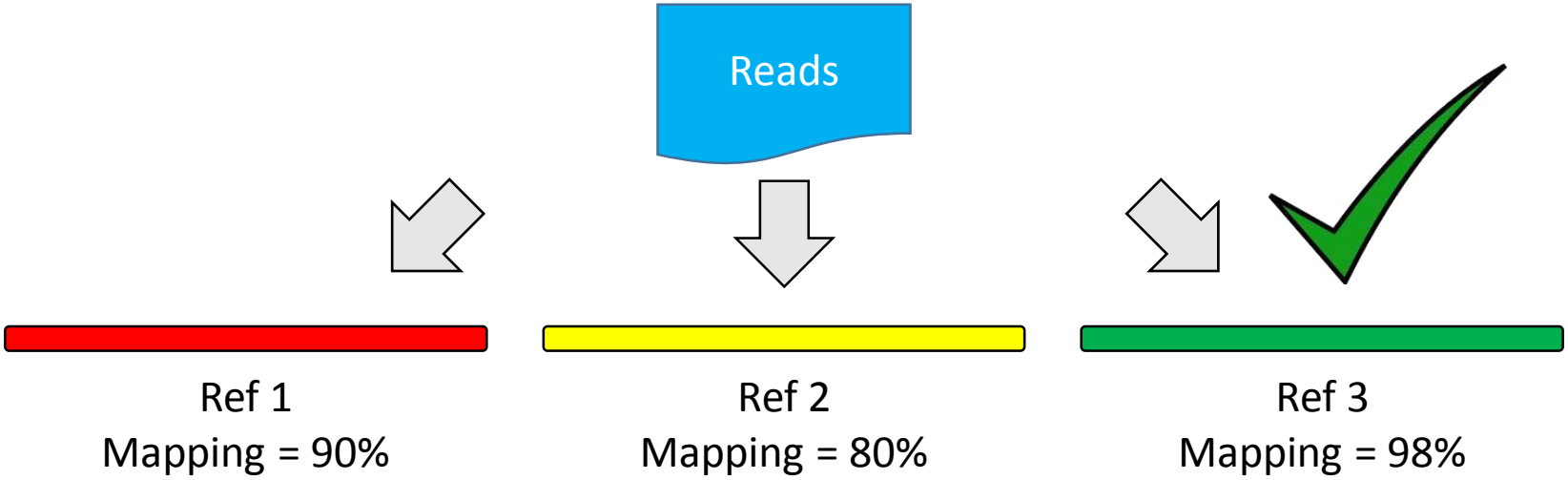
Assessing the “goodness” of mapping (reference assembly)

- The goal of genome mapping is to correctly and unambiguously map all the sequence reads to the corresponding position in the reference genome

- A mapping is deemed good if it can get as close to this goal as possible

- Goodness of mapping is measured by one parameter:
  - Percent Overall Alignment

Genome Mapping & Similarity



Assembly Statistics

What statistics can help us in assessing assemblies?

- ↑ Number of bases assembled
- ↓ Number of assembled contigs (length > 500bp)
- ↑ N50 value

- You can also combine these statistics to come to a unified number

Assembly Statistics

What statistics can help us in assessing assemblies?

- ↑ Number of bases assembled
- ↓ Number of assembled contigs (length > 500bp)
- ↑ N50 value

- You can also combine these statistics to generate a single score

$$Assembly\ score = \frac{\#\ bases\ assembled\ \times N50}{\#\ contigs}$$

Assembly Statistics

What statistics can help us in assessing assemblies?

- ↑ Number of bases assembled
- ↓ Number of assembled contigs (length > 500bp)
- ↑ N50 value

- You can also combine these statistics to generate a single score

$$Assembly\ score = \log_{10}\left(\frac{\#\ bases\ assembled\ \times N50}{\#\ contigs}\right)$$

N50 is a common measure of assembly quality

If these are the assembled contigs:

and I order them by length in an ascending manner

N50 is the length of contig which is in the center

Assembly Comparison

- So how can we smartly compare assemblies?
- As with any other domain of bioinformatics, tools have been developed to tackle this task
- QUAST is one such tool