CMSC423: Bioinformatic Algorithms, Databases and Tools

Genome sequencing

In the news: Illumina

Illumina Announces Expedited Individual Genome Sequencing Service (IGS)

Breakthrough Turnaround Time for Human Genome Sequencing Available via Illumina's CLIA-Certified Laboratory

Press Release: Illumina, Inc. – Tue, Sep 11, 2012 6:30 AM EDT

Companies: Illumina Inc.

RELATED QUOTES

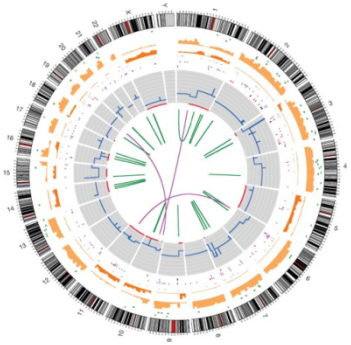
Symbol	Price	Change
ILMN	45.60	0.54

SAN DIEGO--(BUSINESS WIRE)--

Illumina, Inc. (ILMN) today announced the introduction of its rapid Individual Genome Sequencing (IGS) service with a turnaround time in as little as two weeks. Rapid turnaround whole genome sequencing services were announced by Illumina in June 2012, enabled by technology innovations to the HiSeq® platform. Now these advancements have been implemented in Illumina's CLIA-certified laboratory to enable the same fast turnaround for the IGS service.

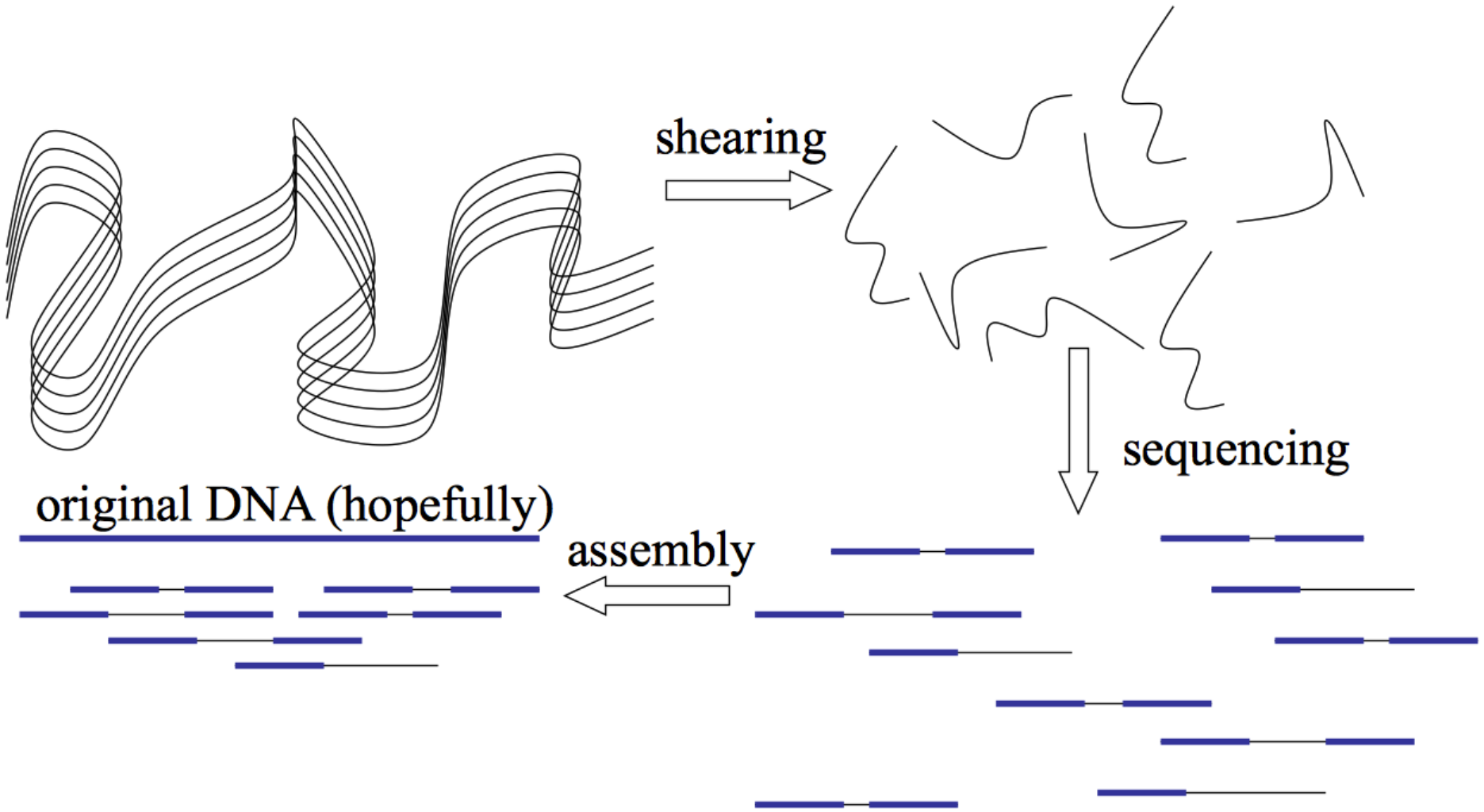
The IGS service is only available through a physician's order and is designed to assist clinicians with diagnosis and treatment decisions. As the only CLIA-certified, CAP-accredited whole genome sequencing service laboratory in the world, Illumina continues to increase access and lay the foundation for routine clinical use of whole-genome sequencing.

Biological Goals

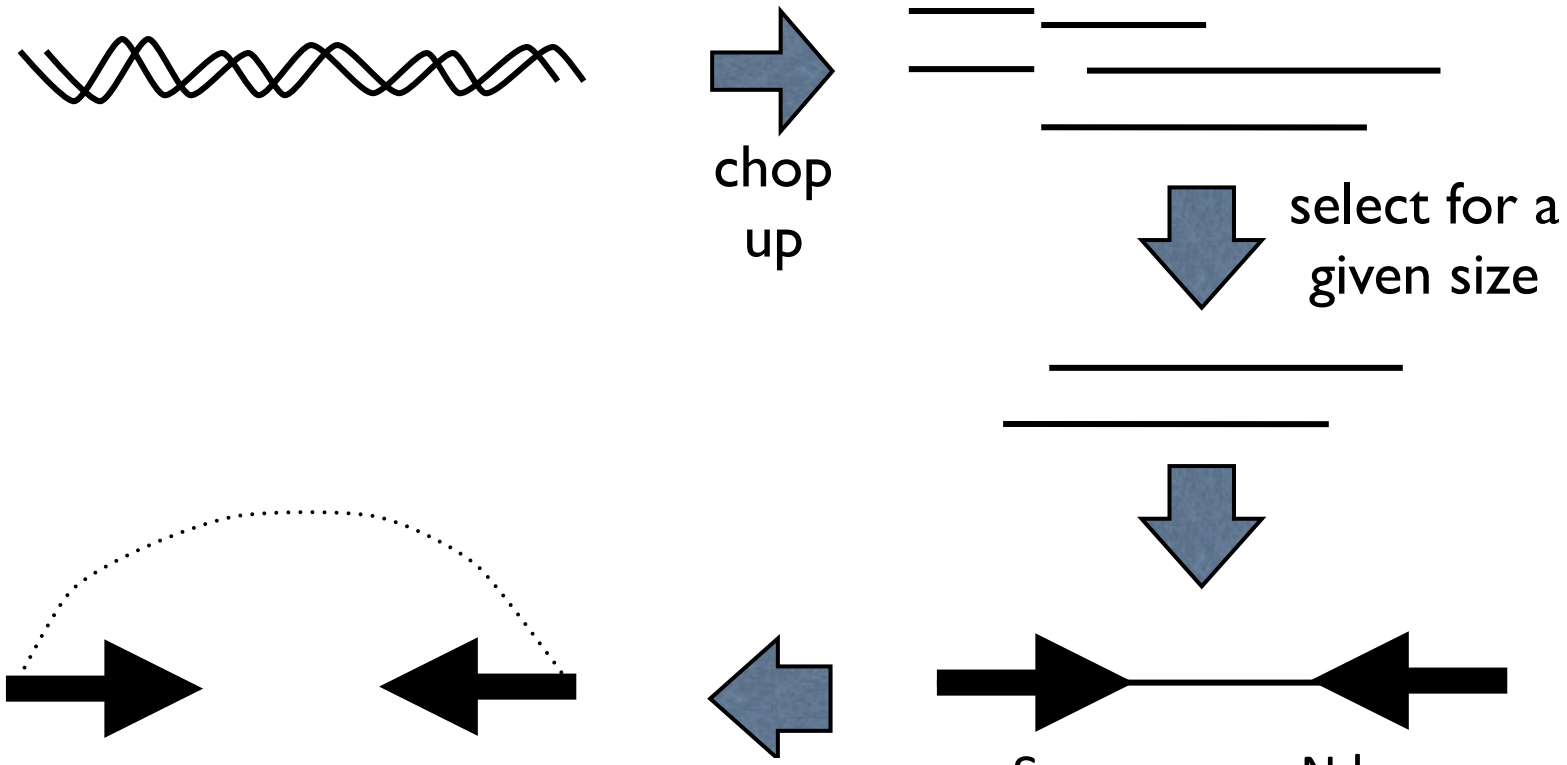


- Disease outbreaks
  - *V. cholerae* in Haiti
  - *B. anthracis* in Heroin users
- Learn what the cell is doing
  - The DNA transcribed into RNA to be translated to proteins
- Studying whole communities (metagenomics)
  - Human symbiotic bacteria
  - Ocean bacterial population
- Studying the dark matter
  - Studying individual cells (single-cell)

Sequencing



Paired Ends



Mate pair: 2 reads, of known orientation separated by an approximately known distance

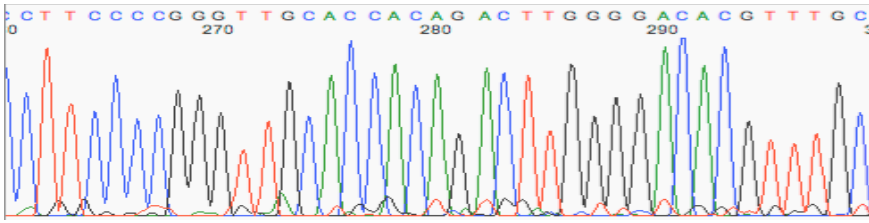
=> Long range information

Platform: 3730

Model	length	reads	bases
ABI 3730	800	96	80K

Applied Biosystems 3730xl
- DNA amplification in *E. coli*.
- Sanger dye-terminator PCR.
- Capillary gel electrophoresis.
- 90 min run makes 96 reads.
- 500 to 800 bp/read.
- Paired end range 2Kbp - 40Kbp.
- Mate rate almost 100%.
- Reads include vector at 5’ end.
- Quality drops gradually at 3’ end.
- Traces available for inspection.

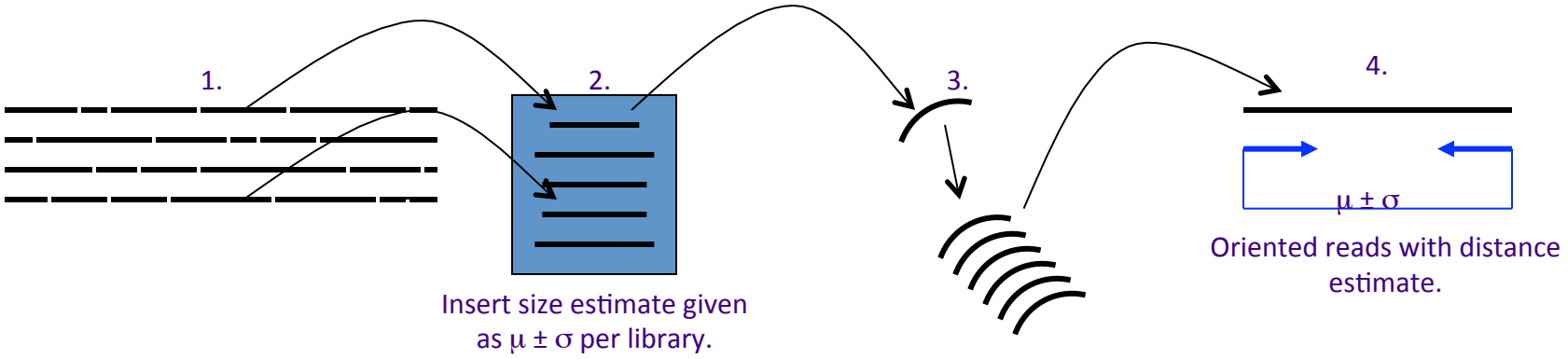
ABI 3730



Trace from 3730

Sanger-era Paired Ends

1. Randomly shear copies of the genome
2. Build libraries of similar-length fragments
3. Amplify individual fragments
4. Sequence two ends of each fragment

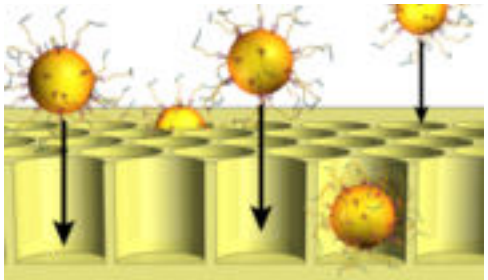


Platform: GS XL+

Model	length	reads	bases
ABI 3730	800	96	80K
454 XL+	700	1M	700M

Roche 454 Titanium XL+

- Successor to GS 20, FLX Standard, and FLX+
- Pyrosequencing
    - Emulsion PCR
    - Bead capture
    - Transfer to picoliter plate
    - Sequencing by synthesis of complementary strand
    - Image processing
- Paired-ends: Reads are cut from single circles (in lab), Size options are 3Kbp or 8Kbp or 20Kbp, Linker-positive reads cut into pairs by software, 50% of reads are linker-positive.
- Run time is 23hr.
- Clear ranges of 200bp-1000bp.
- Problems: Inaccurate counts of repeated bases, duplicate reads, lower output on paired-end.



Beads on a picotitre plate

454 Output

Example of using unix utilities from 454. Shown is sffinfo. See also sfffile.

```
$ sffinfo
Usage:  sffinfo [options...] [- | sfffile] [accno...]
Options:
       -a or -accno    Output just the accessions
       -s or -seq      Output just the sequences
       -q or -qual     Output just the quality scores
       -f or -flow     Output just the flowgrams
       -t or -tab      Output the seq/qual/flow as tab-delimited lines
       -n or -notrim   Output the untrimmed sequence or quality scores
```

Dump 1 read from SFF file for half-plate #2 →

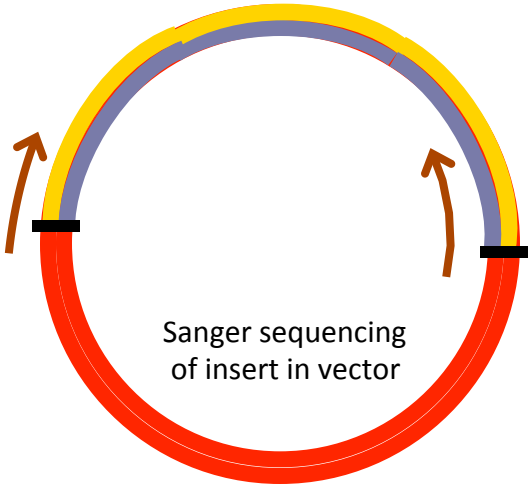
```
$ sffinfo  FFPMSHM02.sff  FFPMSHM02B382V
# of Flows:    400
# of Bases:    271
Clip Qual Left:   5
Clip Qual Right:  90
```

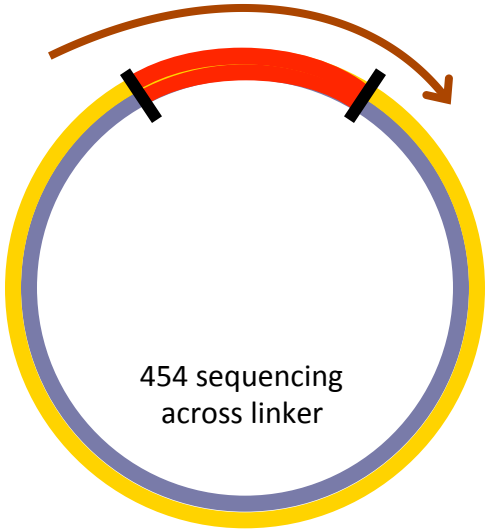
Each flow yields 0 or more bases →

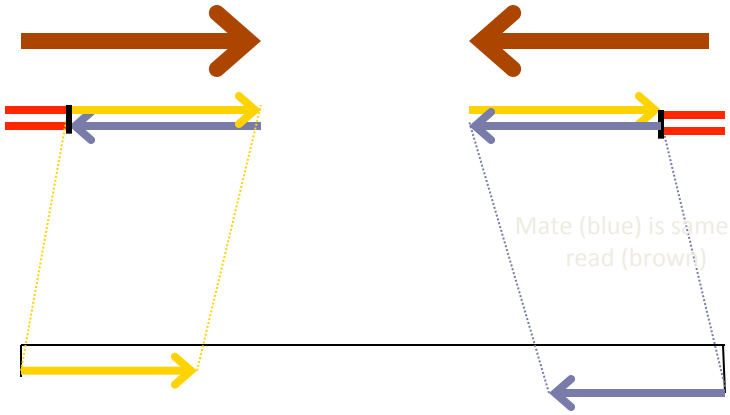
```
Flow Chars: TACGTACGTACGTACGTACGTACGTACGTACGTACGTACGTACGTACG…

Flowgram: 0.93  0.05  1.18  0.08  0.04  0.91  0.05  9.10  0.14  1.97…

Bases:  tcagGGGGGGGGAACCACTGTAACAAAGGAAAGTAGGTTGGTCTCGCGGTGAGTG…

Quality Scores: 32  32  32  16  16  16  16  16 16  16  16  16 23…
```

The first 4 bases (lower case) are the key sequence for calibration.
On this slide, the 4 colors show correspondence between flows and bases.

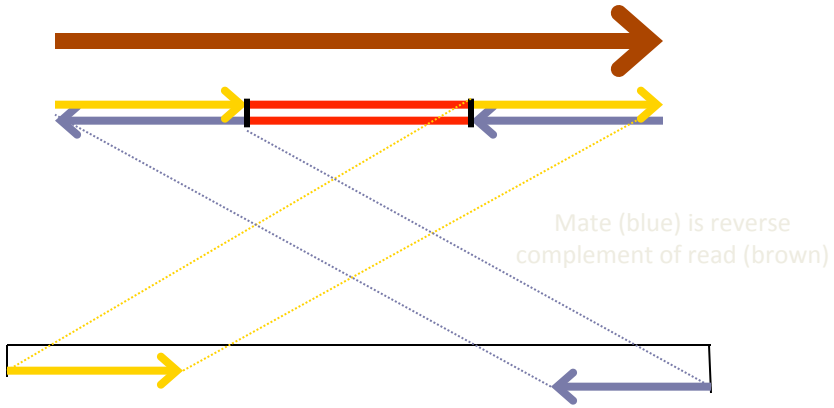
Sanger and 454 Paired Ends







Sanger mate pair from both
forward sequences, vector trimmed



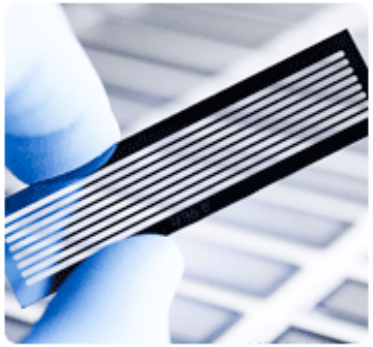
Sanger-style mate pair from
opposite strands, linker trimmed

Platform: Illumina

Model	length	reads	bases
ABI 3730	800	96	80K
454 XL+	700	1M	700M
Illumina HiSeq 2500	100	6B	600G

Illumina HiSeq 2500

- Bridge hybridization on flat surface
  - Cluster formation
  - Sequencing by synthesis
  - Image Processing.
- One 2x run in 11 days.
- Lengths have grown from 35 to 2x150.
- Reads have identical length initially, trimmed by software, Substitution more than indel, error accumulates at read end.
- Paired ends, sequencing both ends of a fragment, range of 35bp-700bp.
- Mate pairs, sequencing fragment cut from circle, range of 2Kbp-8Kbp.



Flow cell with 8 lanes

Illumina Output

s_3_1_0083_qseq.txt

```
HWUSI-EAS541 21 3 83 1787 876  0 1 TAAACAACCCCGGGCTCACCGAGCTGAGTCTGTAGC ^IV`J^JaT\F\D^DFN__ORH\GYIM\SSTWXF\T 0
HWUSI-EAS541 21 3 83 1787 673  0 1 CAGCAGAAGCAAGGCGGCTGCAGAGAGCAGAGAGAC X]\LaaJ`a\MZWWGP_]IZZWZa\IPFIMFPa_TX 0
HWUSI-EAS541 21 3 83 1787 1169 0 1 AACCTCCCTTGCCAGCTTCCCCCGATAGTTAATGGG ab\FHEW]`RT`VWIKPR^]QRX[GFZIQFR]_BBB 0
HWUSI-EAS541 21 3 83 1787 792  0 1 AGGCTGTTAACTGAGTATGGCCGCTCCTATTTGTTT WSabaaVIHYH[]MKUZBBBBBBBBBBBBBBBBBBB 0
HWUSI-EAS541 21 3 83 1787 1066 0 1 CGTTTAGCTCTATCTTACACCACTTCCTCTCCTTTC O__HYJXI_^GZ`bbbb^RN]IGRVI`K\YGZUH\[ 0
HWUSI-EAS541 21 3 83 1787 1723 0 1 AGCGAAGGGGCAGTGAAGGGACGAAAAGCCGAGCAG ZGHZ_^^OO_^BBBBBBBBBBBBBBBBBBBBBBB 0
HWUSI-EAS541 21 3 83 1787 1448 0 1 TCCTCTCCGAACGGTTGACAGCGCTCAGCATTGTAC a]b_b_BBBBBBBBBBBBBBBBBBBBBBBBBBBB 0
HWUSI-EAS541 21 3 83 1787 885  0 1 TGCCCATTGGAGGAAAAAAAATTACATAGATCTCCA ab_aGYDMZX^]`Z^]HWYD[_W\\HQWUMHWWaX_ 0
HWUSI-EAS541 21 3 83 1787 831  0 1 GCCCCGGC.TACGAATGCGGTCAGCGGAAAAAGCAG BBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB 0
HWUSI-EAS541 21 3 83 1787 1764 0 1 CGGCAATT.TTCTTATCAACACCACCCGAACGACTG aBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB 0
HWUSI-EAS541 21 3 83 1788 137  0 1 AAATCTG..AATGATTCTCCCTTCTGACGTACCACA BBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB 0
```

Machine ID | Lane | 1 or 2 of pair | Base calls | Quality values

s_3_1_0083_qseq.fastq

```
@HWUSI-EAS541:3:83:1787:876:#0/1 run=090401
TAAACAACCCCGGGCTCACCGAGCTGAGTCTGTAGC
+HWUSI-EAS541:3:83:1787:876#0/1
^IV`J^JaT\F\D^DFN__ORH\GYIM\SSTWXF\T
@HWUSI-EAS541:3:83:1787:673:0#0/1 run=090401
CAGCAGAAGCAAGGCGGCTGCAGAGAGCAGAGAGAC
+HWUSI-EAS541:3:83:1787:673:0#0/1
X]\LaaJ`a\MZWWGP_]IZZWZa\IPFIMFPa_TX
```

Alternate format, same data

<u>phred Quality Values, now used by Illumina</u>

QV=20 means $1/10^2$ chance of error.

QV=30 means $1/10^3$ chance of error.

QV=40 means $1/10^4$ chance of error.

<u>Solexa Quality Values, formerly used by Illumina</u>

QV=2 is encoded as ascii(2+64)='B'.

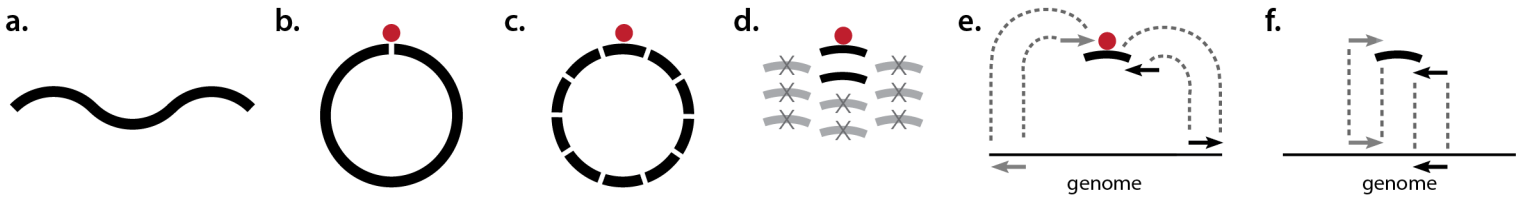
QV=20 is encoded as ascii(20+64)='T'.

QV=30 is encoded as ascii(30+64)='^'.

QV=40 is encoded as ascii(40+64)='h'.

*64 is an arbitrary constant. Sanger files used 33.

Illumina MP Protocol



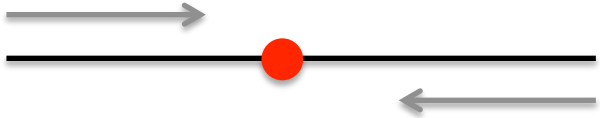
Illumina mate pair library construction protocol:

a. Shear genomic DNA. Size-select long fragments.
b. Circularize. Incorporate biotin marker (red) at the junction.
c. Shear.
d. Size-select short fragments. Enrich for biotin-positive fragments.
e. End reads from biotin-positive fragment has mate pair characteristics.
f. End reads from biotin-negative fragment has paired end characteristics.

Considerations:

- Biotin enrichment is incomplete.
- Presence/absence of biotin not captured by end read data.
- Some mate pair reads will span the junction. These are chimera.

Illumina MP Data



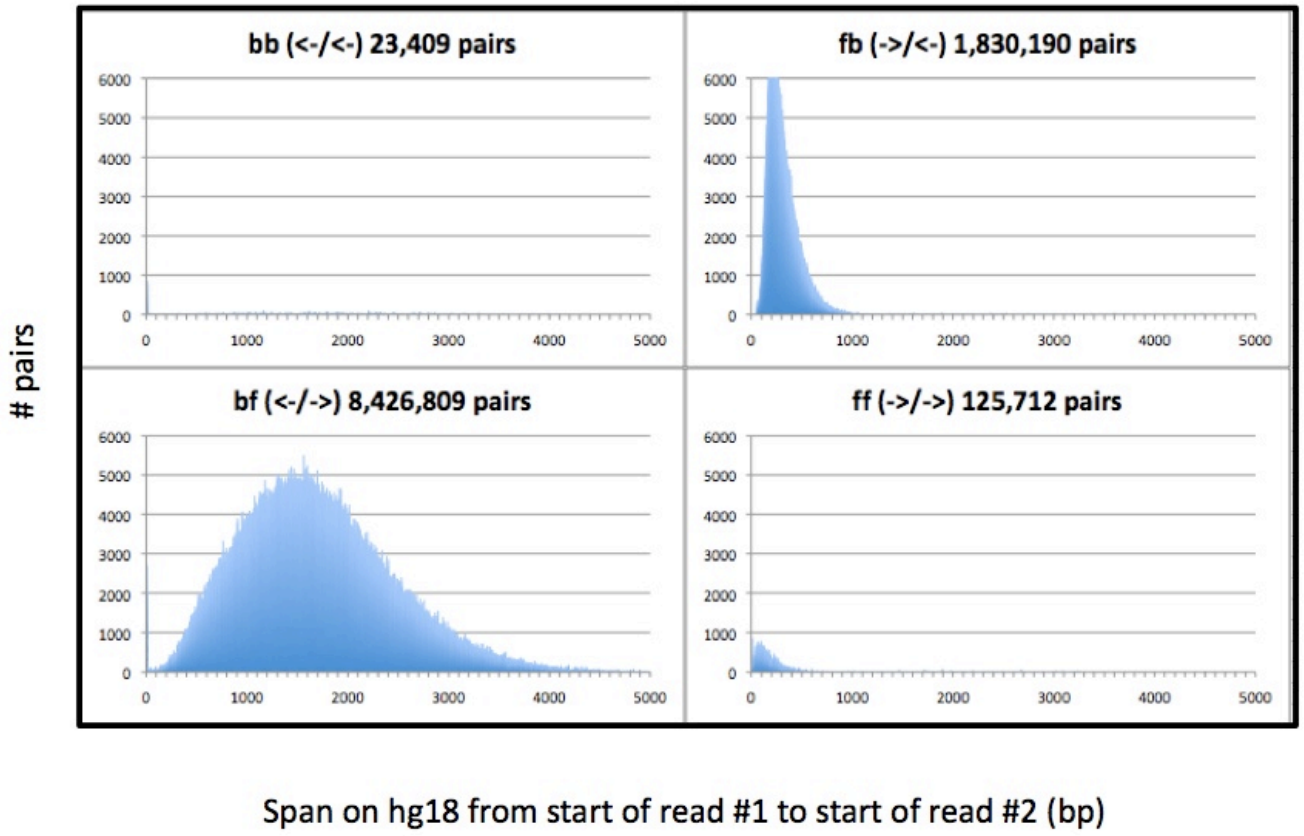
MP: Outie orientation, 3000 bp separation.

PE: Innie orientation, 100 bp separation.

Chimer: One read spans the junction.

Data from Illumina mate pair sequencing presents 3 pair types.

Illumina MP Example



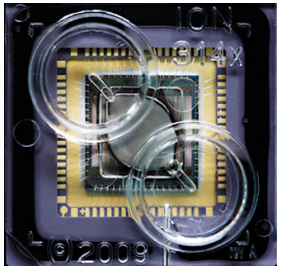
Example: Illumina 2x100 MP sequence from human genomic DNA mapped to NCBI reference with CLC software. Sequence generated at JCVI, January 2011. Analysis by Ewen Kirkness, JCVI.

Platform: ion torrent

Model	length	reads	bases
ABI 3730	800	96	80K
454 XL+	700	1M	700M
Illumina HiSeq 2500	100	6B	600G
Ion PGM	35-200	100K-5M	3M-1G

Life Ion PGM Sequences

- Pyrosequencing without images
  - Incorporation of a base releases Hydrogen ion
  - Sequencing by synthesis
  - pH sensor to detect ion release
  - Voltage processing
- Scalable sequencing by run-time and chip type
- 0.5-4.5hr per run.
- Problems: Inaccurate counts of repeated bases

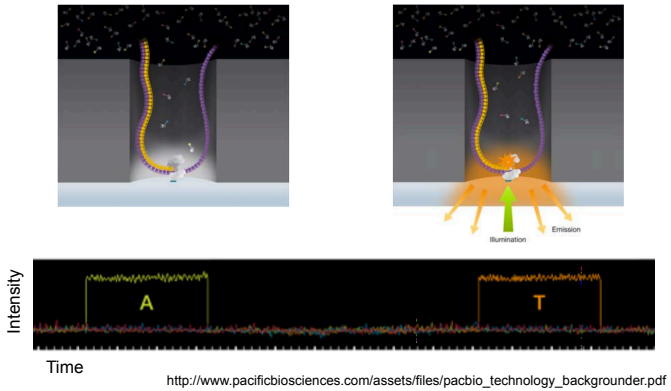


Platform: PacBio RS

Model	length	reads	bases
ABI 3730	800	96	80K
454 XL+	700	1M	700M
Illumina HiSeq 2500	100	6B	600G
Ion PGM	35-200	100K-5M	3M-1G
PacBio RS	1K-10K	75K	150M

Pacific Biosciences PacBio RS

- Single Molecule sequencer (one DNA strand)
  - Ligate adapters
  - Load molecules onto zero mode waveguides
  - Real-time polymerase sequencing
  - Video analysis
- 90 minutes per run.
- Reads are much longer than other technologies
- Can detect base modifications
- Reads have very high error rate (10-15%)
- Can read shorter reads multiple times to improve accuracy

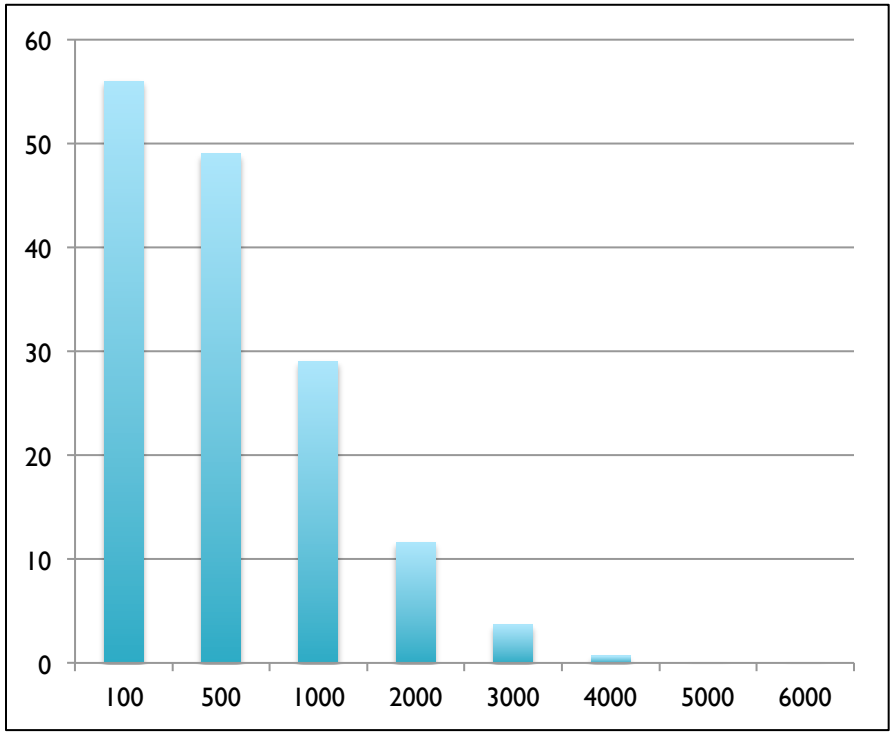


http://www.pacificbiosciences.com/assets/files/pacbio_technology_backgrounder.pdf

PacBio Sequencing Runs

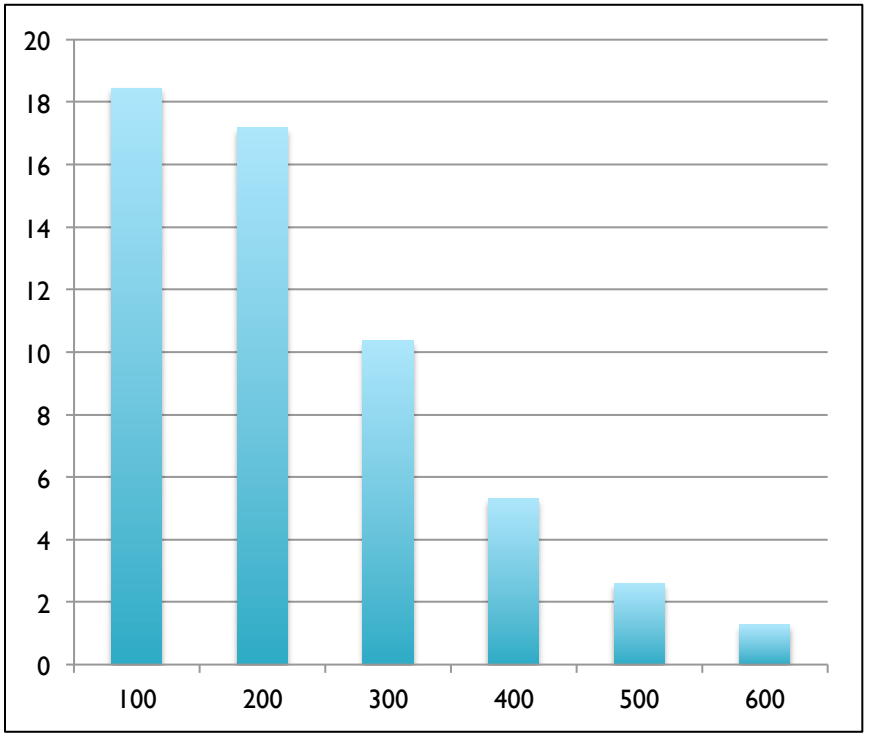
Yeast – Single-pass reads

969,445 reads after filtering
Mean: 710 +/- 663
Median: 558 Max: 8,495



Yeast – CCS reads

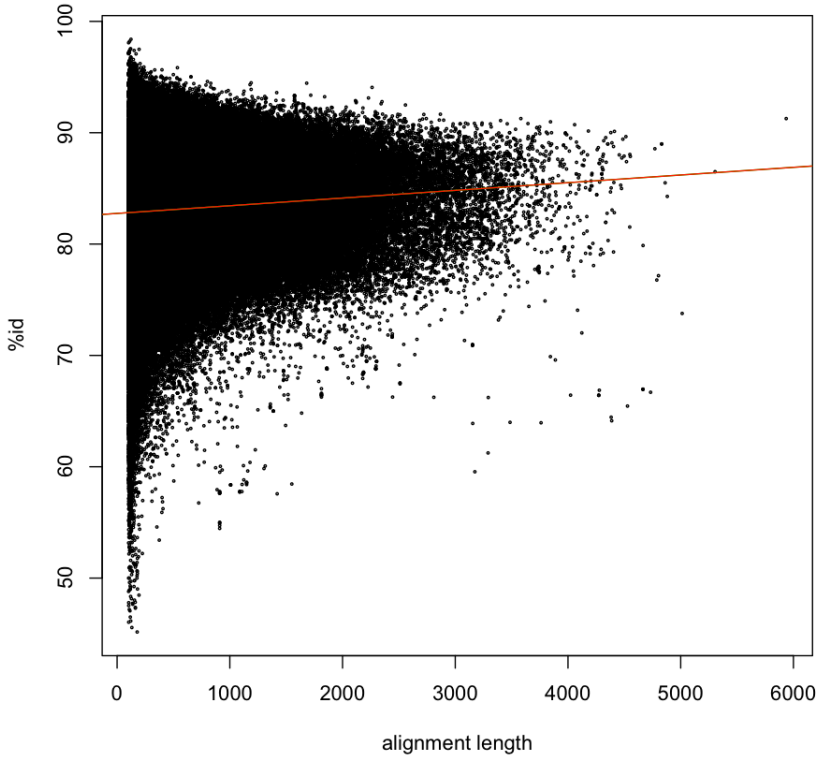
731,638 reads after filtering
Mean: 306 +/- 115
Median: 279 Max: 1,425



Read Accuracy

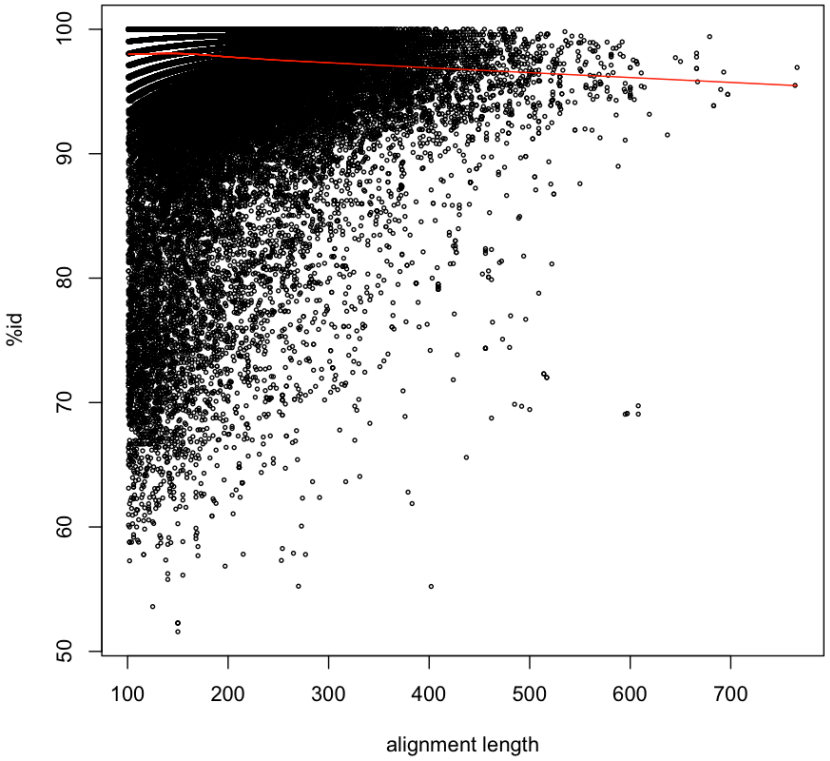
Yeast – Long reads

94% aligned reads
48% reads aligned >100bp
7% reads aligned >1kbp



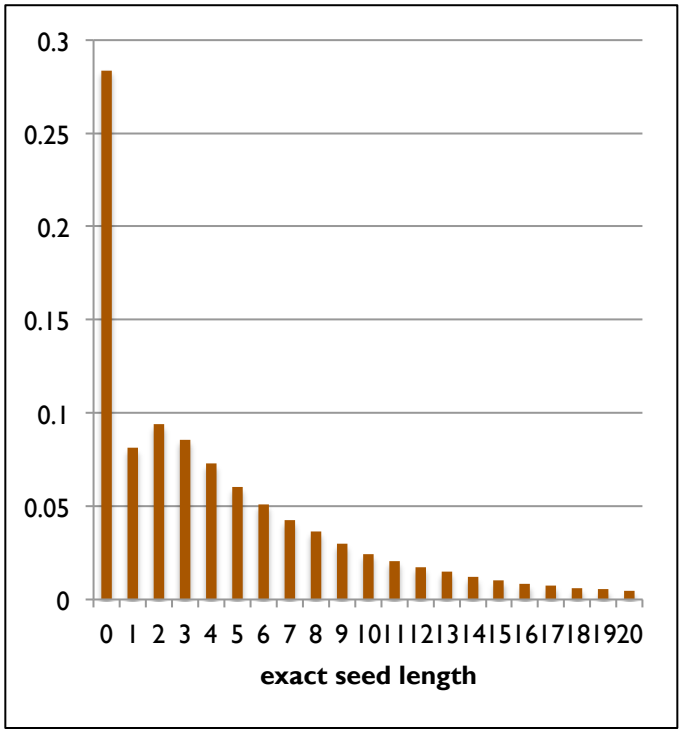
Yeast – CCS reads

99.93% aligned reads
98.2% reads aligned >100bp
38.8% reads aligned >300bp



Alignment Quality

Match	83.7%
Mismatch	1.4%
Insertions	11.5%
Deletions	3.4%




```
     4  TTGTAAGCAGTTGAAAACTATGTGTGGATTTAGAATAAAGAACATGAAAG
                ||||||||||||||||||||||||| ||||||| |||||||||||| |||
539752  TTGTAAGCAGTTGAAAACTATGTGT-GATTTAG-ATAAAGAACATGGAAG

    54  ATTATAAA-CAGTTGATCCATT-AGAAGA-AAACGCAAAAGGCGGCTAGG
                | |||||| |||||||||||||| |||| | |||||| |||||| ||||||
539800  A-TATAAATCAGTTGATCCATTAAGAA-AGAAACGC-AAAGGC-GCTAGG

   101  CAACCTTGAATGTAATCGCACTTGAAGAACAAGATTTTATTCCGCGCCCG
                | |||||| |||| ||  |||||||||||||||||||||||||||||||||
539846  C-ACCTTG-ATGT-AT--CACTTGAAGAACAAGATTTTATTCCGCGCCCG

   151  TAACGAATCAAGATTCTGAAAACACAT-ATAACAACCTCCAAAA-CACAA
                | |||||||| |||||||||||||| || ||    |||||||||| ||||||
539891  T-ACGAATC-AGATTCTGAAAACA-ATGAT----ACCTCCAAAAGCACAA

   199  -AGGAGGGGAAAGGGGGGAATATCT-ATAAAAGATTACAAATTAGA-TGA
                 ||||||   ||      |||||||| || |||||||||||||| || |||
539934  GAGGAGG---AA-----GAATATCTGAT-AAAGATTACAAATT-GAGTGA

   246  ACT-AATTCACAATA-AATAACACTTTTA-ACAGAATTGAT-GGAA-GTT
                |||  |||||||||  | |||||||||||||  |||  |||||||  ||||  |||
539974  ACTAAATTCACAA-ATAATAACACTTTTAGACAAAATTGATGGGAAGGTT

   291  TCGGAGAGATCCAAAACAATGGGC-ATCGCCTTTGA-GTTAC-AATCAAA
                || |||||||||  |||||||  |||  ||||  ||||||  |||||  |||||||
540023  TC-GAGAGATCC-AAACAAT-GGCGATCG-CTTTGACGTTACAAATCAAA

   338  ATCCAGTGGAAAATATAATTTATGCAATCCAGGAACTTATTCACAATTAG
                |||||||  |||||||||  ||||||  |||||  ||||||||||||||||||
540069  ATCCAGT-GAAAATATA--TTATGC-ATCCA-GAACTTATTCACAATTAG
```

Sample of 1M reads aligned with BLASR requiring >100bp alignment

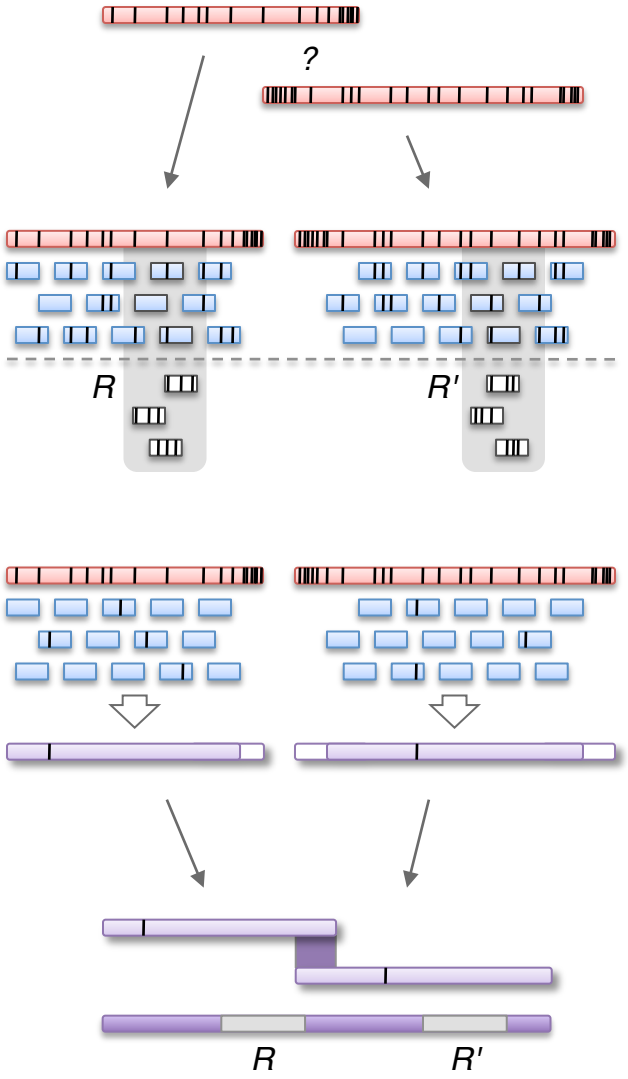
Hybrid correction and assembly

- Pre-process reads

- Map short to long reads
- Separate repeats

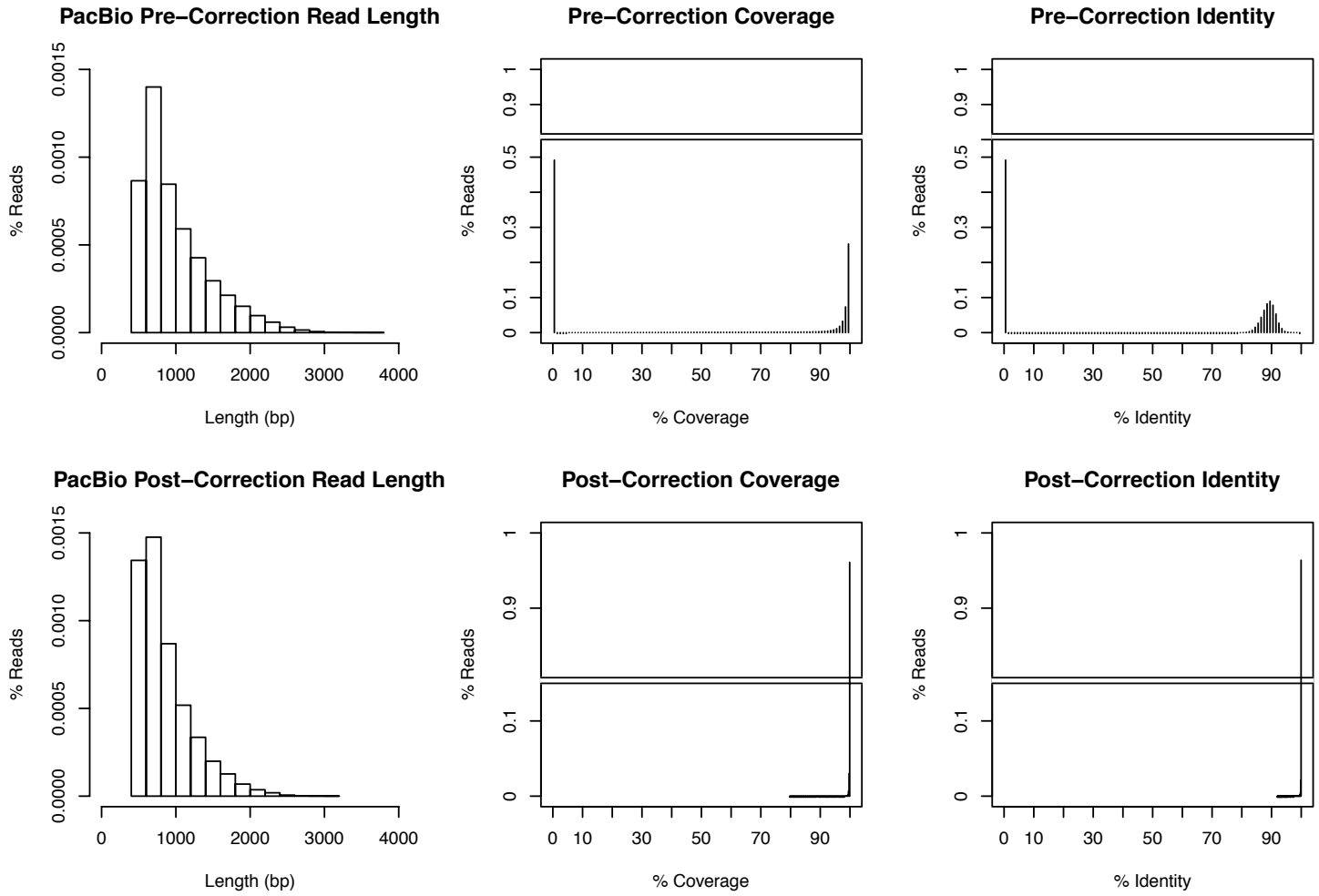
- Compute layout
- Trim at coverage gaps
- Compute consensus

- Assemble corrected reads with OLC strategy



(Koren *et. al.* 2012)

Return to Accurate Long Sequence



Correction results of 20x PacBio coverage of *E. coli* K12 corrected using 50x Illumina

Future

- Current technologies always improving
  - These slides are probably already out of date

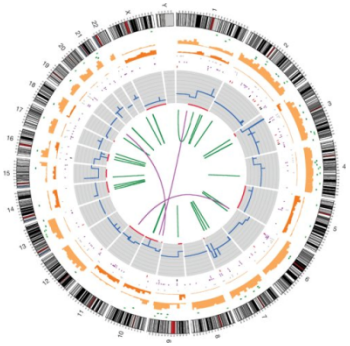
- Oxford Nanopore
  - Announced February 2012
  - Not publically available
  - Promise of single-cell real-time sequencing in a small package
    - Sequencing as a cluster for lots of data or a USB drive for portability
  - A DNA molecule is passed through a nanopore
  - Current across aperture identifies the base

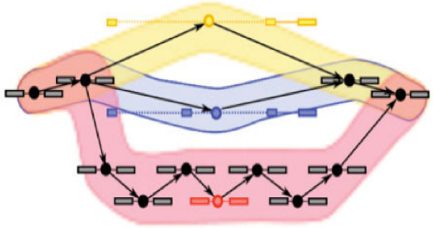
Sequencing Applications

- Novel genomes

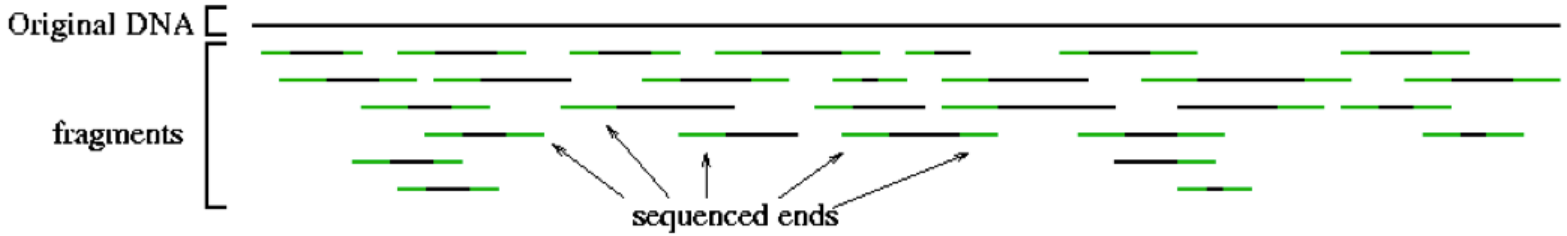
- Metagenomes

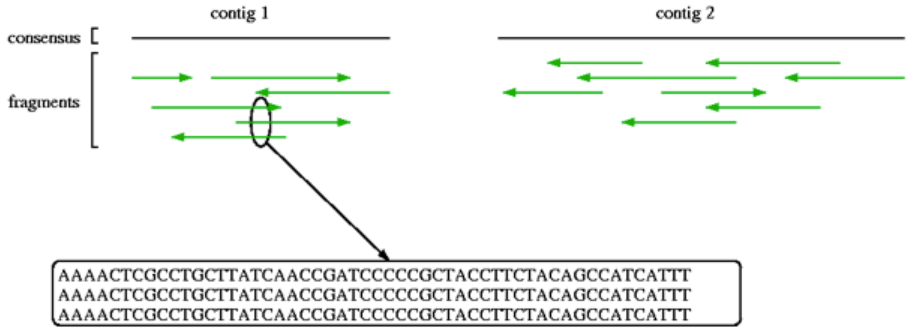
- Sequencing assays
  - Variations
  - Transcript analysis
  - …





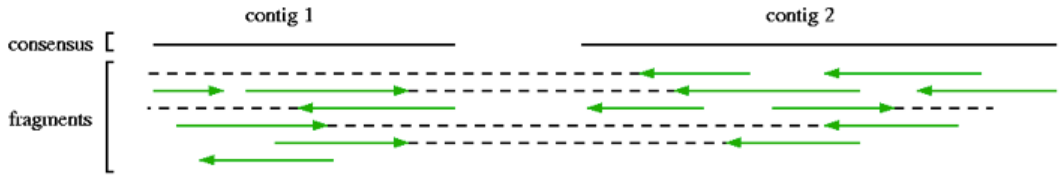
Genome Assembly





Assembly

Scaffolding

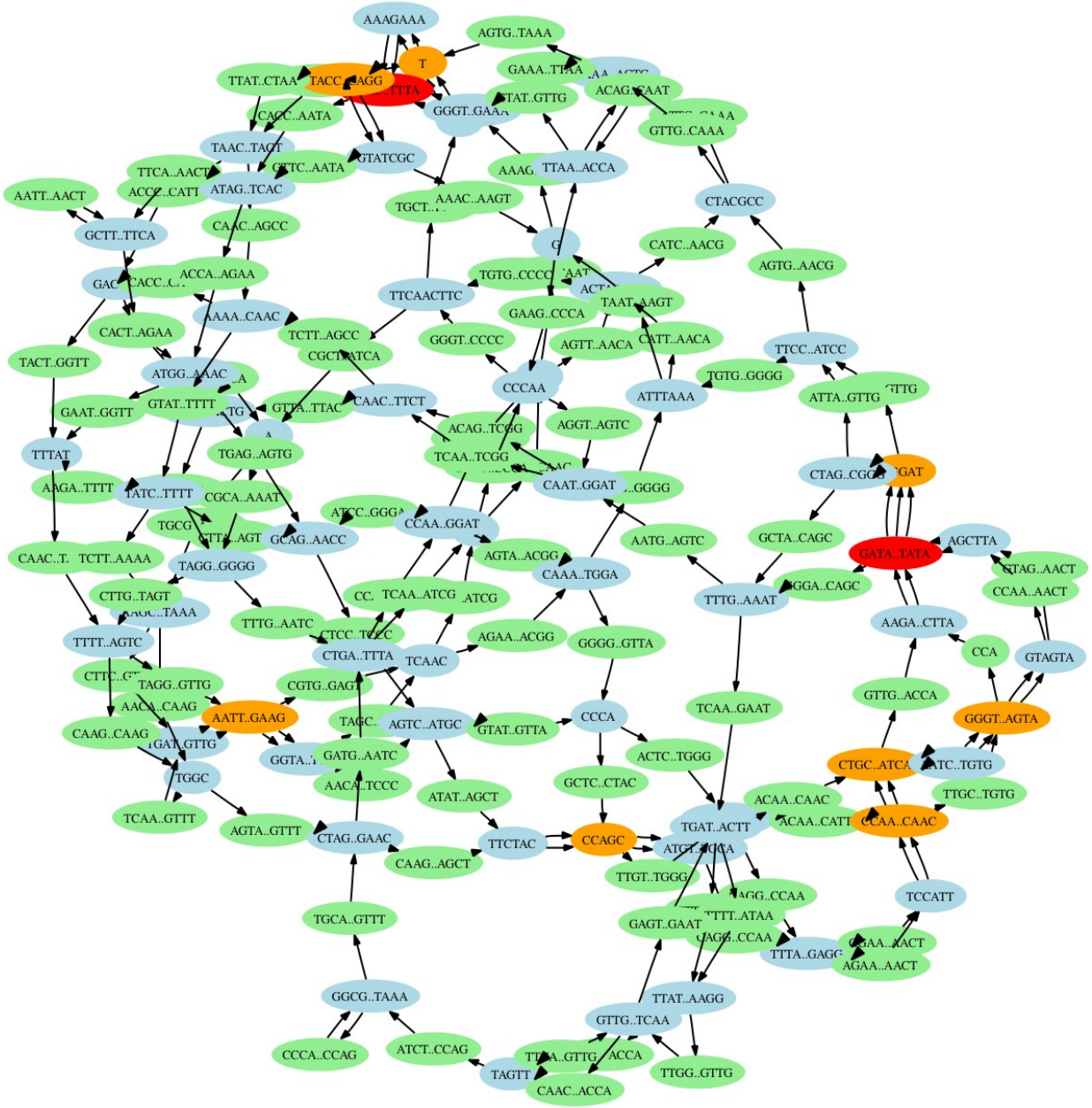


What is Assembly

- Break target into pieces we can read
- Convert sequence to a graph

  - Requires identifying segments with shared origin
  - Sequences occurring multiple times (repeats) make this ambiguous
  - Repeat sequences must be spanned with sufficient unique sequence to be unambiguous

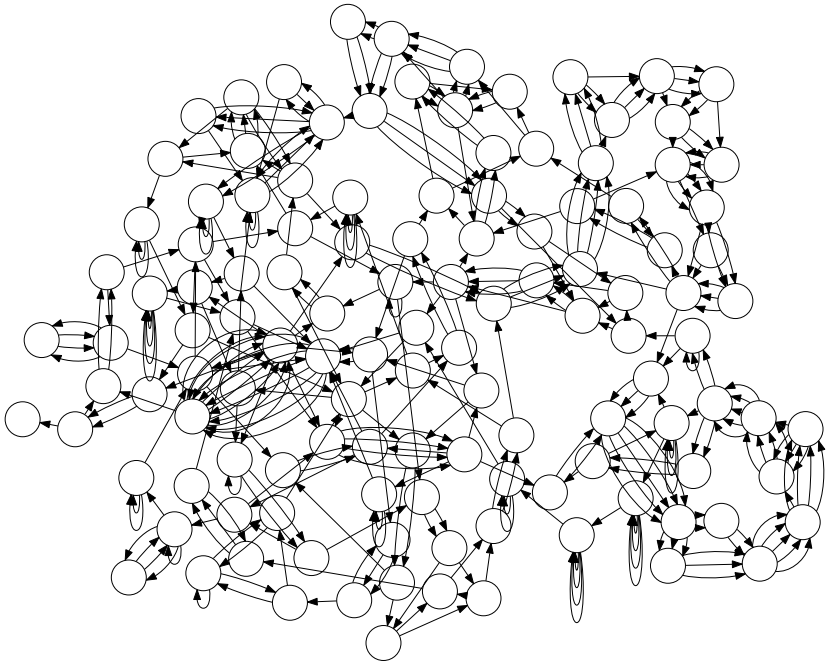
- Basically, a giant jigsaw puzzle



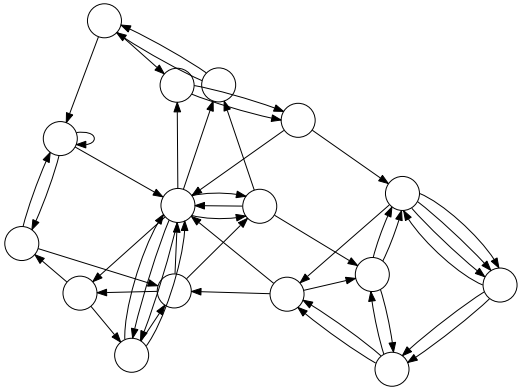
Mycoplasma genitalium, 600Kbp

Long Sequences Simply the Graph

k = 50



k = 1,000

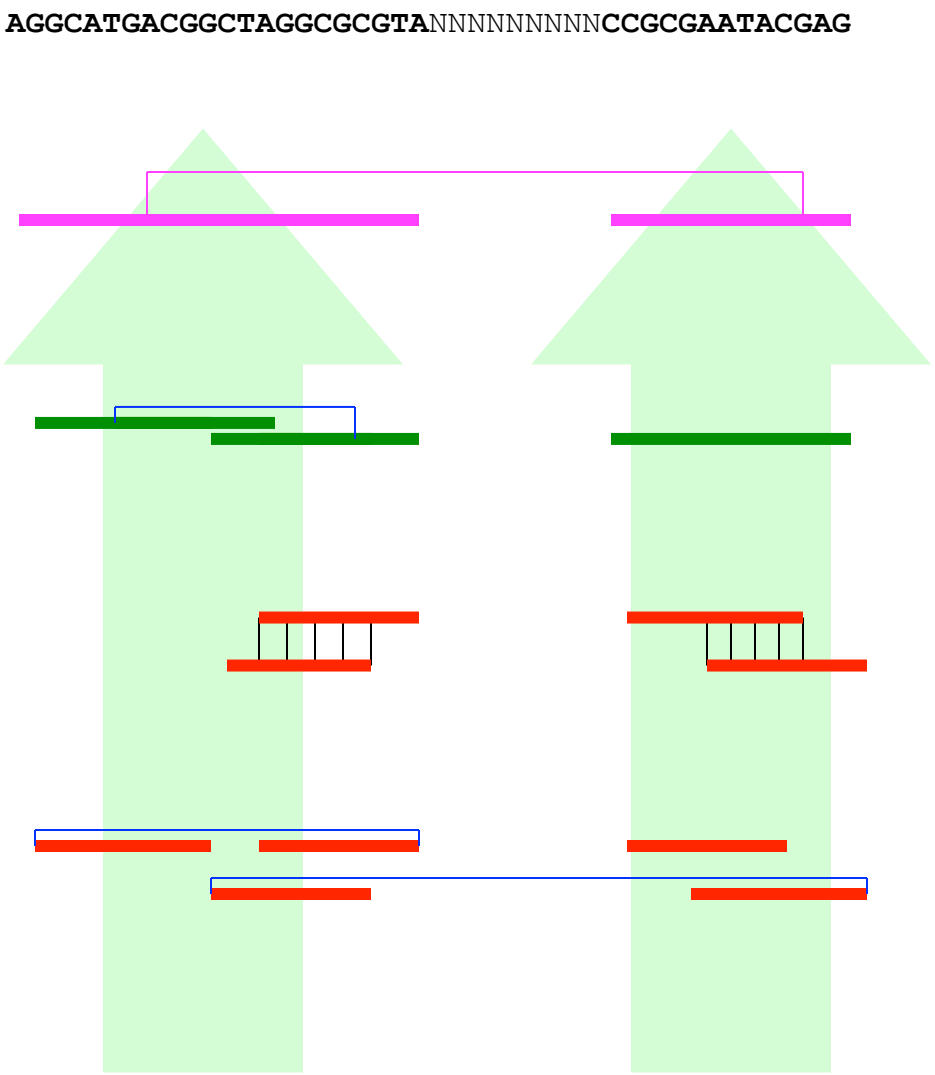


k = 5,000

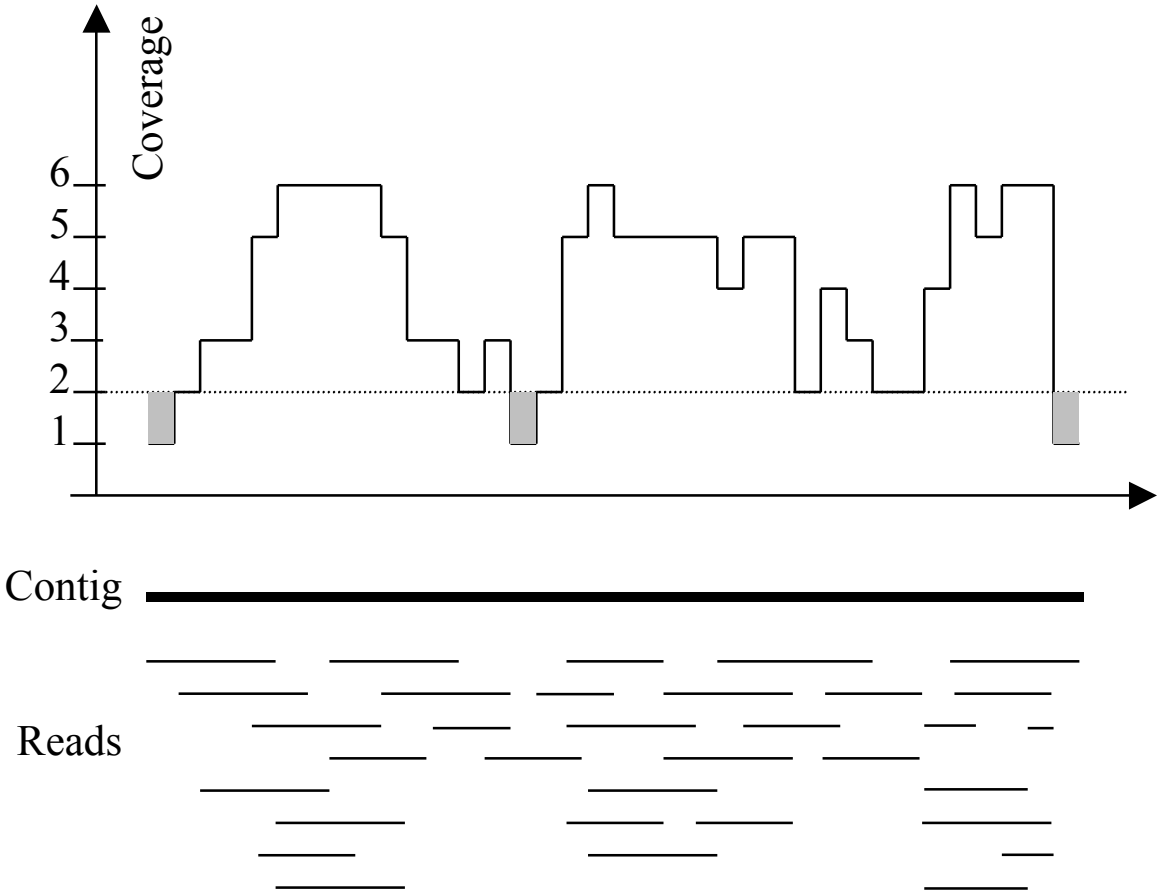
(Kingsford *et. al.* 2010)

Assembly Glossary

- Consensus
  - Multi-alignment of read sequences to generate a single representative sequence
- Scaffold = Contigs + Gaps
  - group of contigs that can be ordered and oriented with respect to each other (usually with the help of mate-pair data)
- Contigs
  - contiguous segment of DNA reconstructed (unambiguously) from a set of reads
- Overlaps
  - Shared sequences between the suffix of one read and the prefix of another
- Reads
  - small (50-2000bp) segment of DNA "read" by a sequencing instrument
- Mate-pair, paired ends
  - pair of reads whose distance from each other within the genome is approximately known



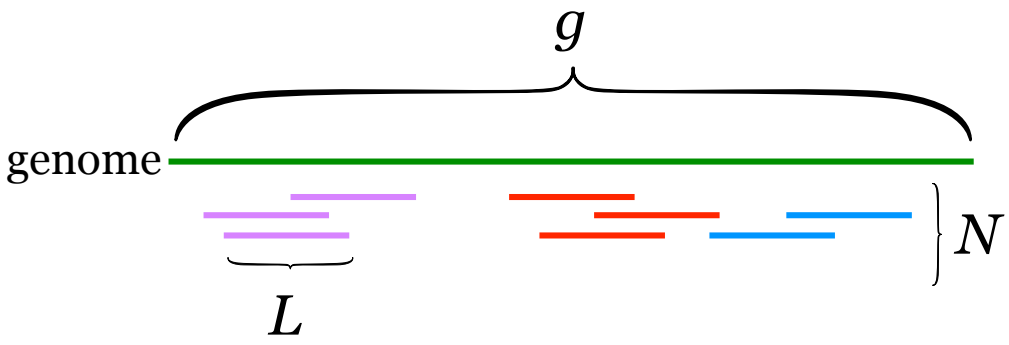
Impact of randomness – non-uniform coverage

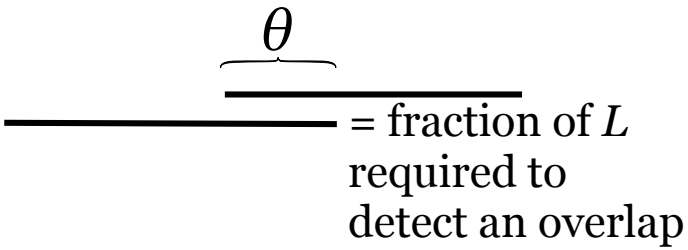


Imagine raindrops on a sidewalk

Lander-Waterman Statistics

How many reads to we need to be sure we cover the whole genome?





An ***island*** is a contiguous group of reads that are connected by overlaps of length ≥ θL.
(Various colors above)

Want: Expression for expected # of islands given N, g, L, θ.

Expected number of Islands

$\lambda := N/g$ = probability a read starts at a given position (assuming random sampling)

Pr(*k* reads start in an interval of length *x*)

x trials, want *k* "successes," small probability λ of success

Expected # of successes = λx

Poisson approximation to binomial distribution:

$$\Pr(k \text{ reads in length } x) = e^{-\lambda x} \frac{(\lambda x)^k}{k!}$$

Expected # of islands = $N \times$ Pr(read is at rightmost end of island)

$= N \times \Pr(0 \text{ reads start in } (1-\theta)L)$

$= Ne^{-\lambda(1-\theta)L} \frac{\lambda^0}{0!}$ (from above)

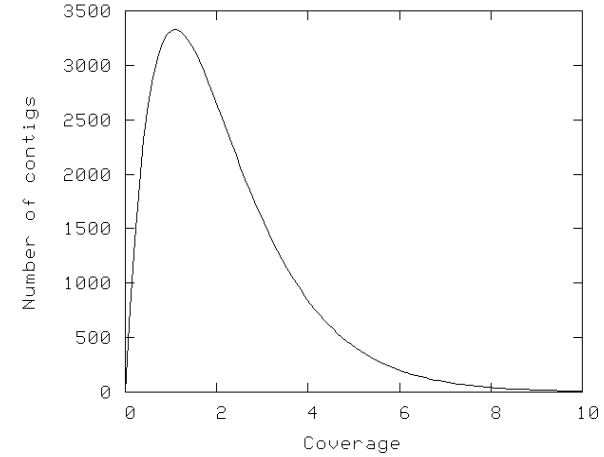
$= Ne^{-\lambda(1-\theta)L}$

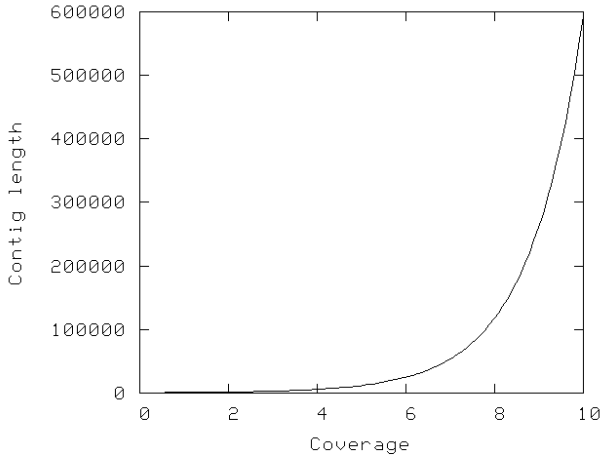
$= Ne^{-(1-\theta)LN/g}$ ← LN/g is called the **coverage** *c*.

Lander-Waterman Summary

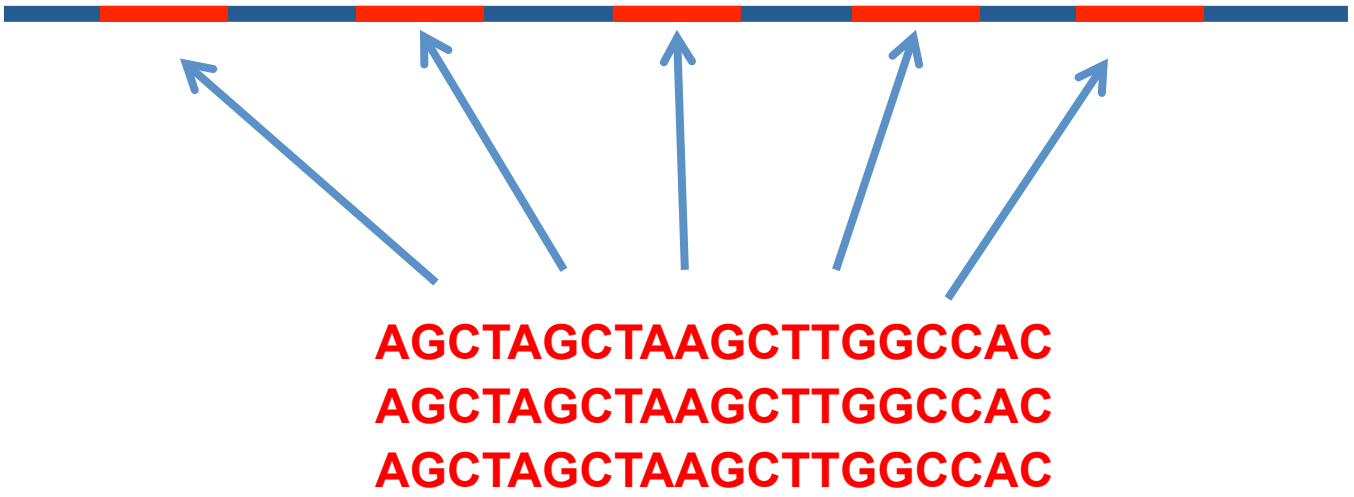
- L = read length
- T = minimum overlap
- G = genome size
- N = number of reads
- c = coverage (NL / G)
- $\sigma = 1 - T/L$

- E(#islands) = $Ne^{-c\sigma}$
- E(island size) = $L(e^{c\sigma} - 1) / c + 1 - \sigma$
- contig = island with 2 or more reads

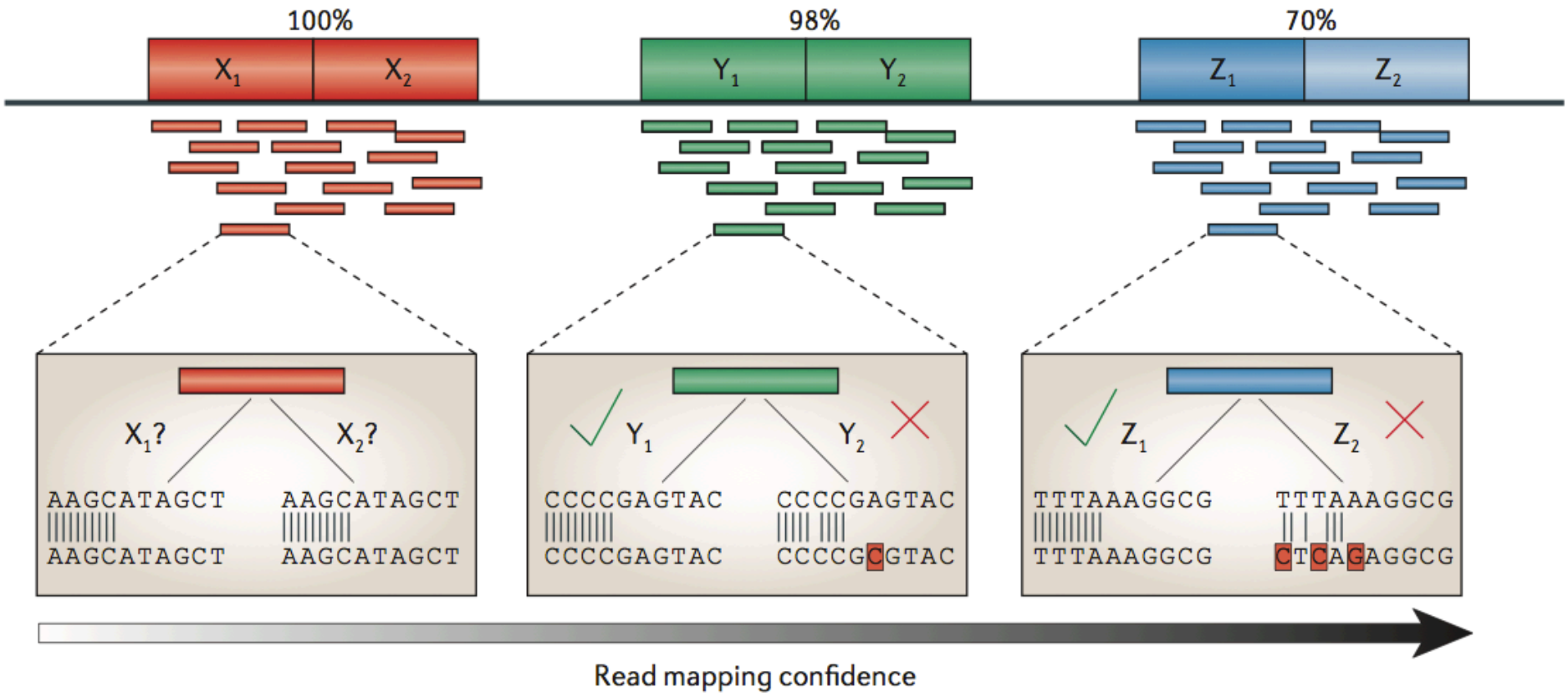




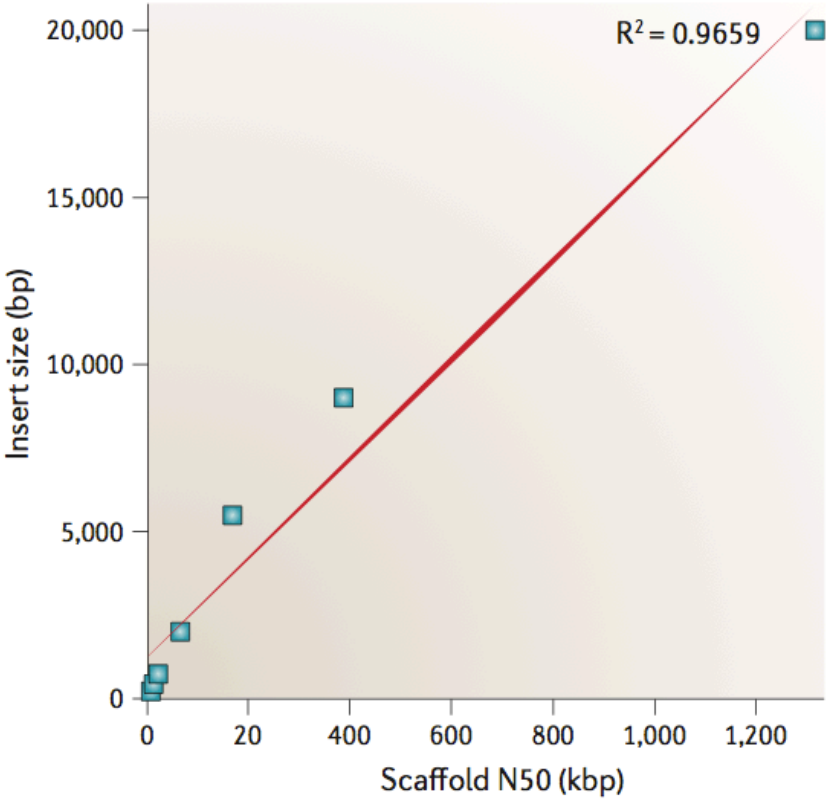
Sequence Alignment



Mapping confidence vs Repeat copy similarity



When all else fails -> $$$



- Repeats cover over 62% of the Potato genome
- First assembly produced tiny contigs (N50=697bp) and small scaffolds (N50=8kb)
- After additional Illumina mate-pair libraries and Sanger sequencing, the final N50 scaffold size was 1.3 Mbp, a 100x improvement!

(Treangen *et. al.* 2012)