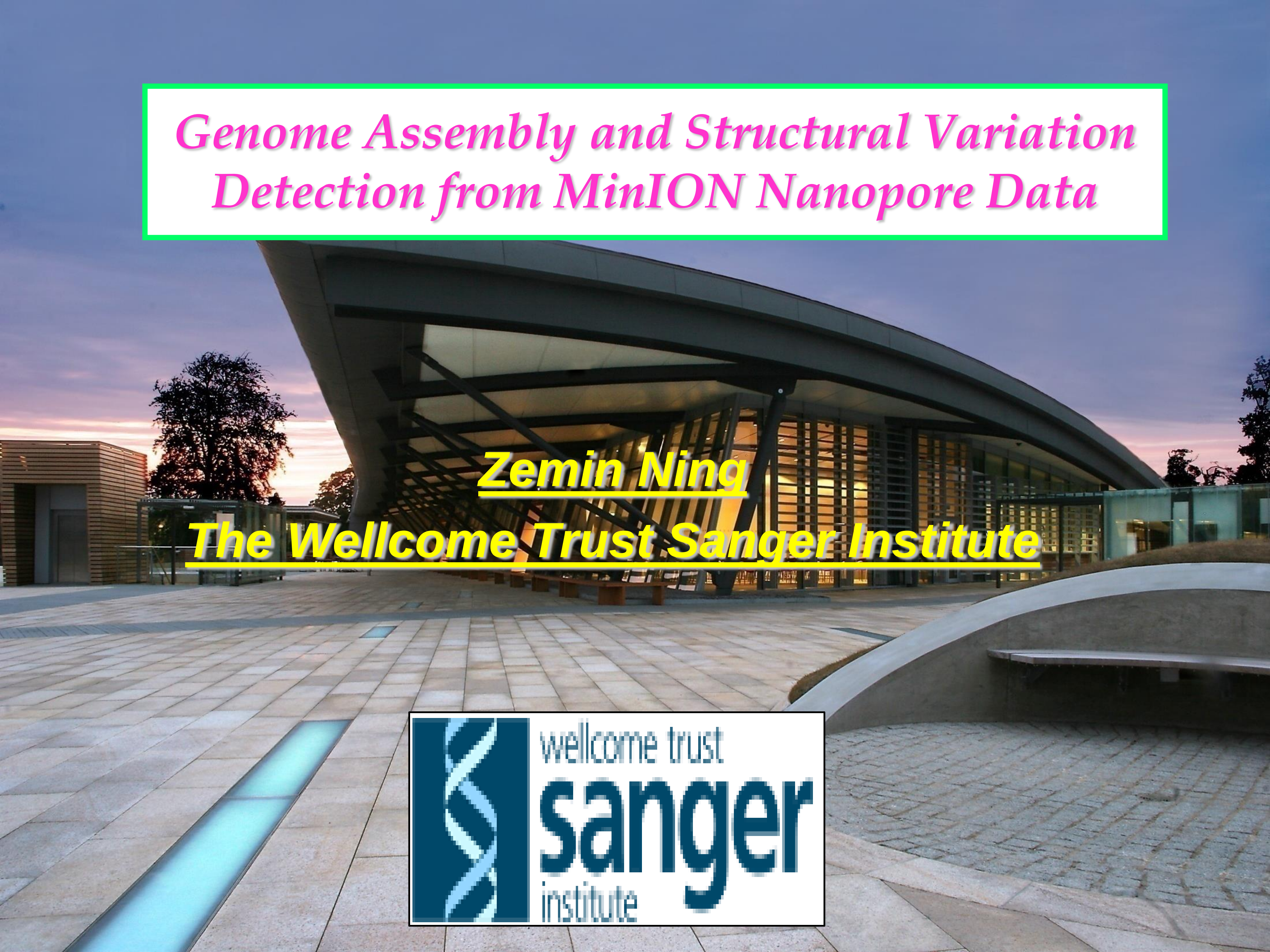
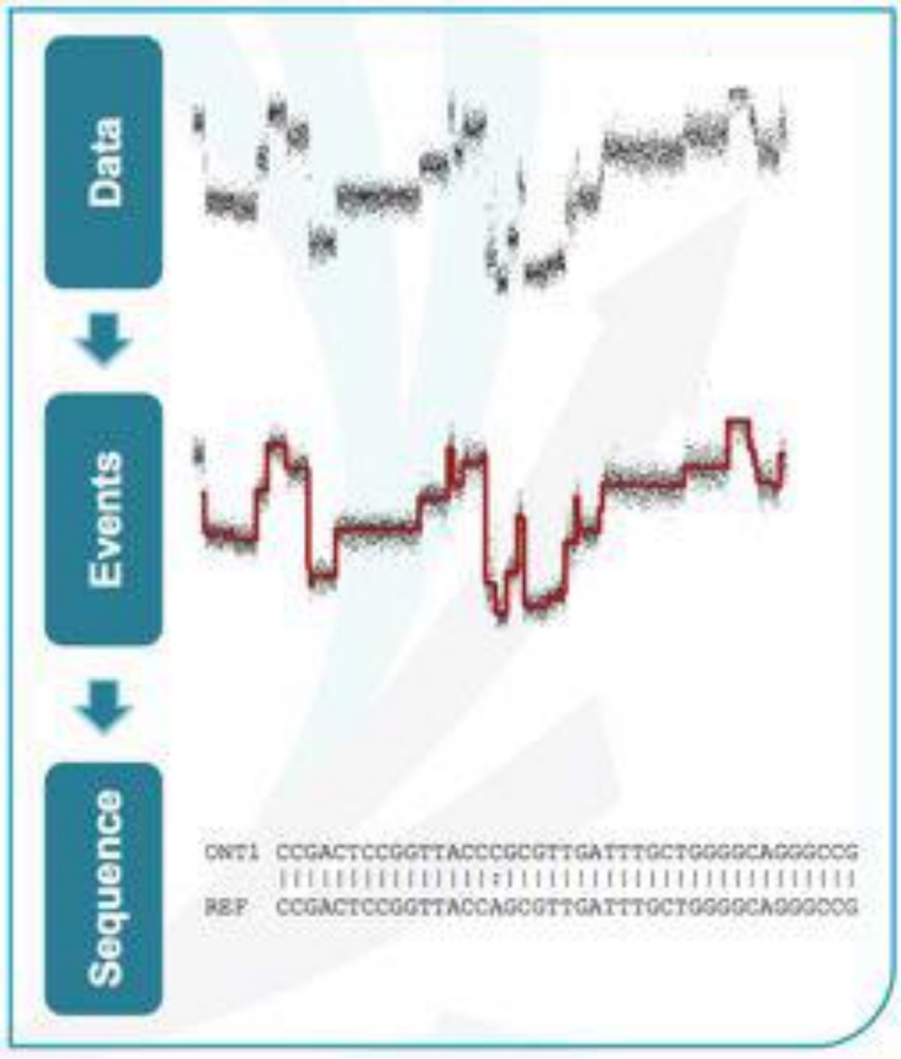
Genome Assembly and Structural Variation Detection from MinION Nanopore Data

Zemin Ning

The Wellcome Trust Sanger Institute

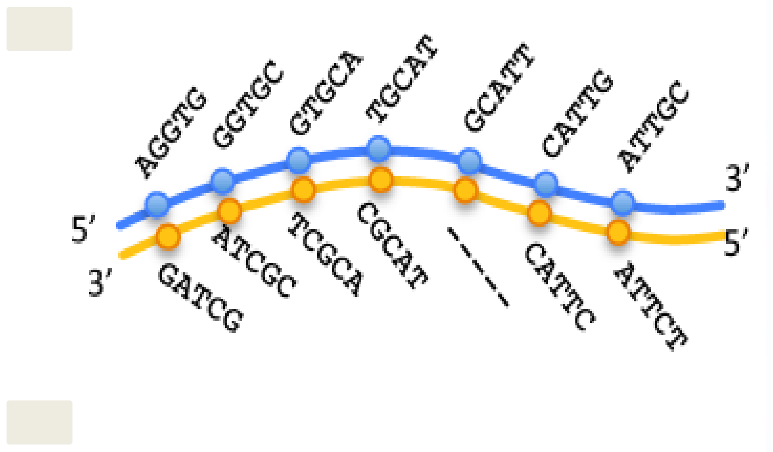


Base Calling of Nanopore



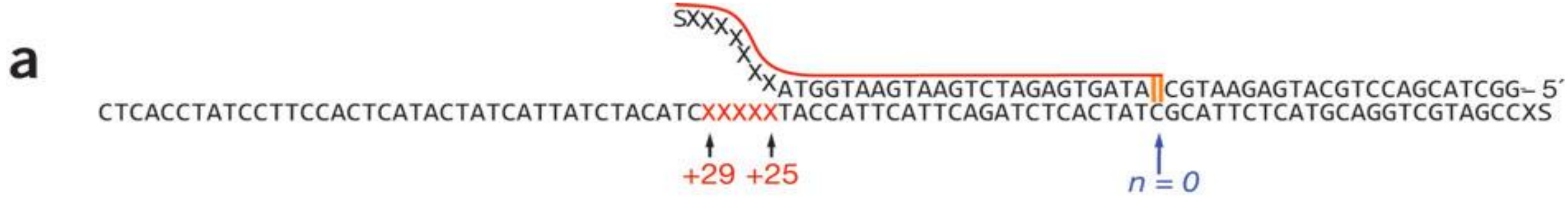
```
ONT1 CCGACTCCGGTTACCCGCGTTGATTTGCTGGGGCAGGGCCG
     |||||||||||||||:|||||||||||||||||||||||||
REF  CCGACTCCGGTTACCAGCGTTGATTTGCTGGGGCAGGGCCG
```

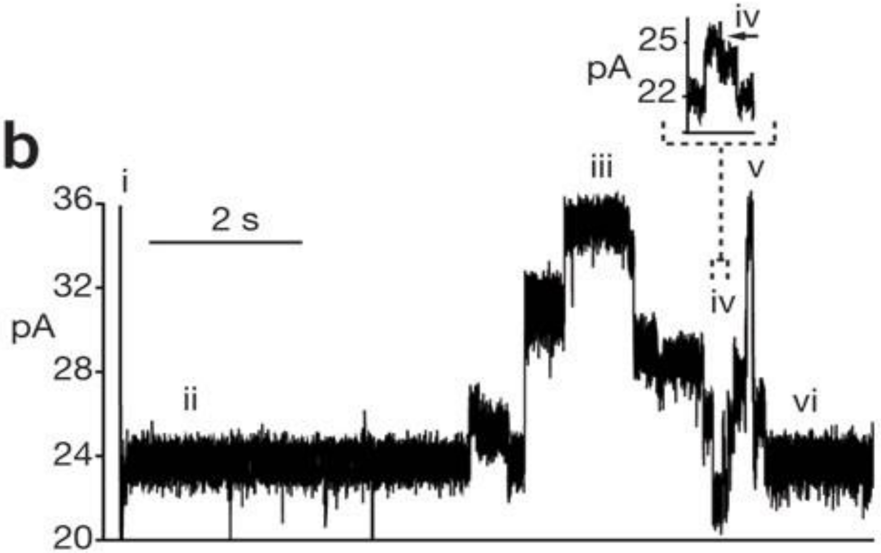
- Hidden Markov model
- Only four options per transition
- Pore type = distinct kmer length

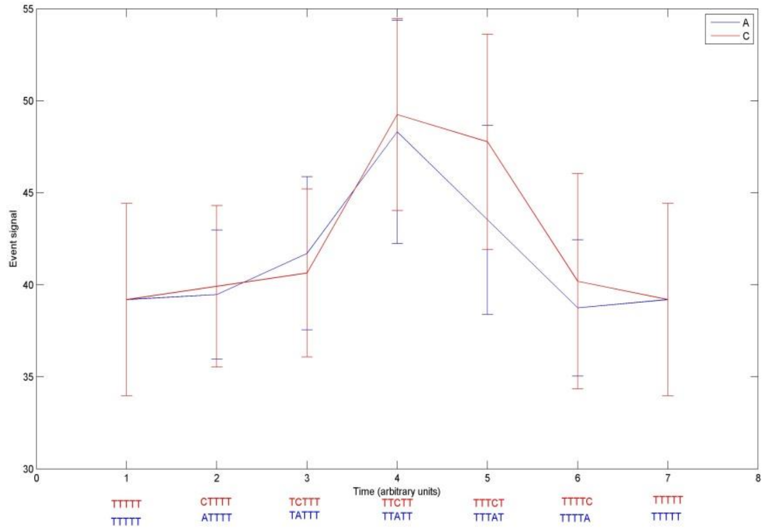


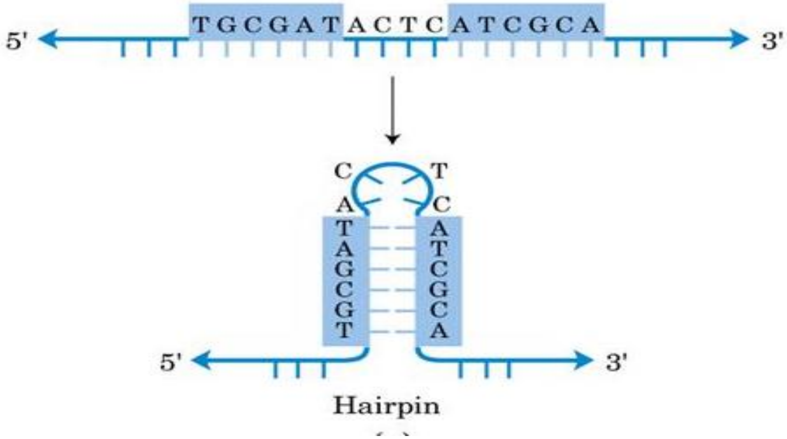
```
1D Temp 5′-GTGCATT-G-CTTGAGTTCTTT-CAGTT-A-3
           |||| || | |||||||||||| ||||| |
     2D 5′-CTGCATTCTCCGTGAATTCTTTCGAATTCT-3
           |||| || | |||||||||||| ||||| |
1D Comp 5′-TCGC-TT-TCCGCGAATTCCTT-GAATT-C-3
```

Basecalling currently is performed at Amazon


a
SXXXXXXXX
ATGGTAAGTAAGTCTAGAGTGATA CGTAAGAGTACGTCCAGCATCGG– 5′
CTCACCTATCCTTCCACTCATACTATCATTATCTACATCXXXXXTACCATTCATTCAGATCTCACTATCGCATTCTCATGCAGGTCGTAGCCXS
+29 +25
n = 0



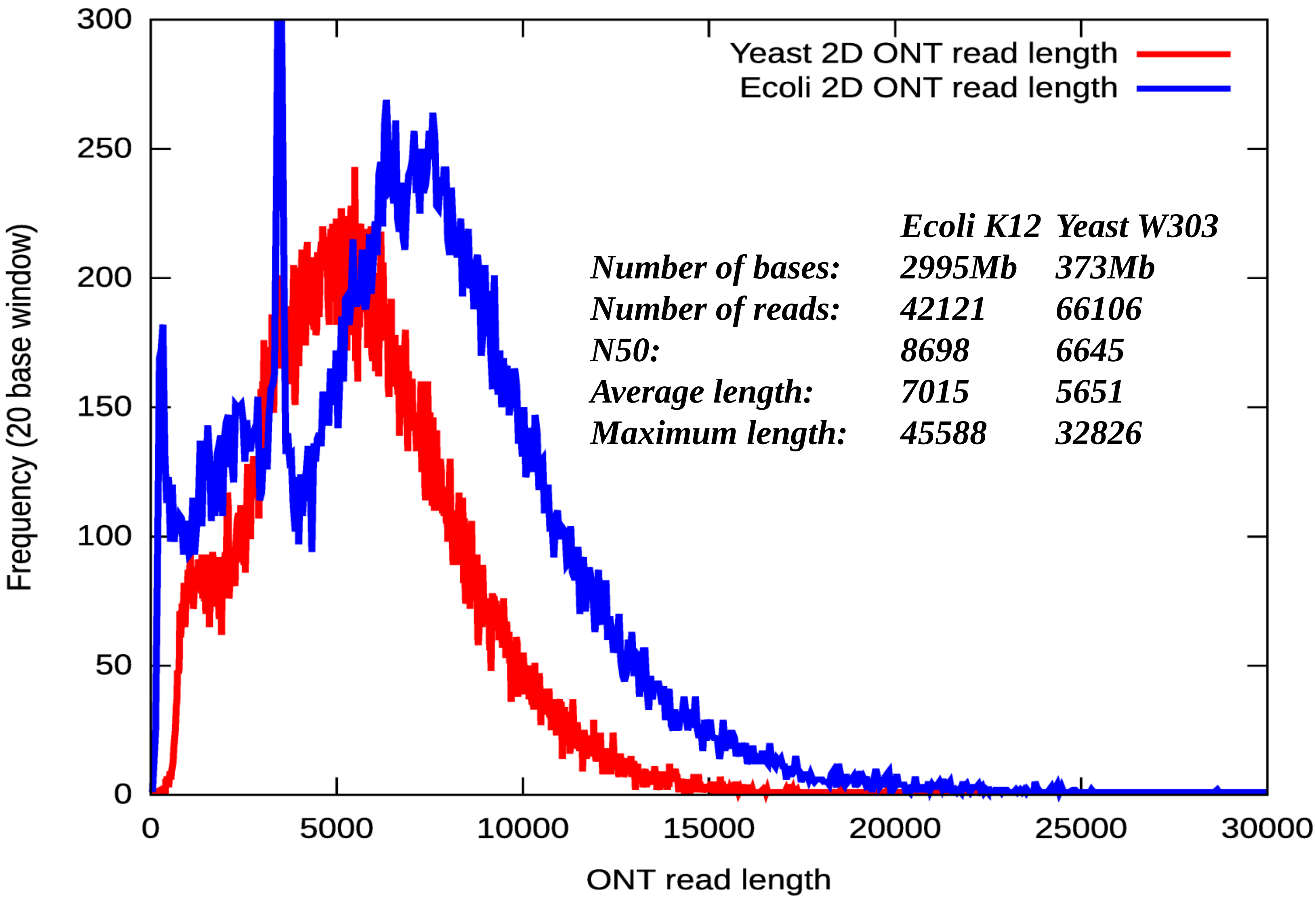




1D and 2D Base Calling

The 1D vs 2D barcoding refers to whether the complementary strand is used to improve basecalled data. Basically – it gives two shots when examining the same loci. The advantage being that the complementary strand will have a different kmer profile.

Read Length Distribution – Ecoli and Yeast



	Ecoli K12	*Yeast W303*
Number of bases:	*2995Mb*	*373Mb*
Number of reads:	*42121*	*66106*
N50:	*8698*	*6645*
Average length:	*7015*	*5651*
Maximum length:	*45588*	*32826*

Ecoli by UCSC: http://www.ebi.ac.uk/ena/data/view/ERS715551-ERS715552/

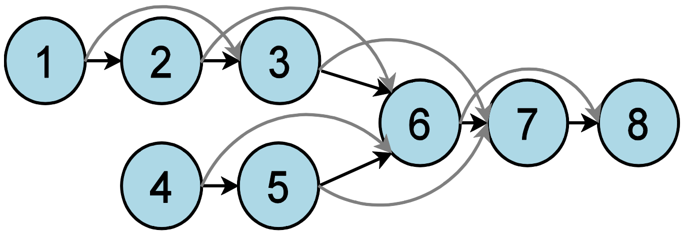
Yeast by CSH : http://labshare.cshl.edu/shares/schatzlab/www-data/nanocorr/

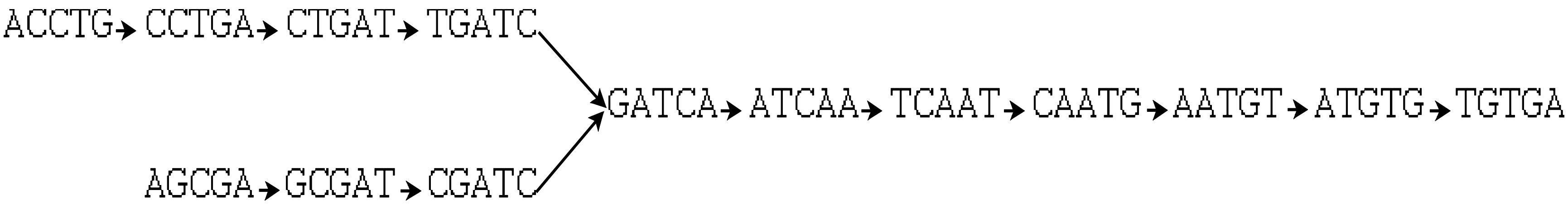
Assembly Method

Sequencing reads:

```
1 A C C T G A T C
2     C T G A T C A A
3       T G A T C A A T
4   A G C G A T C A
5       C G A T C A A T
6         G A T C A A T G
7             T C A A T G T G
8               C A A T G T G A
```

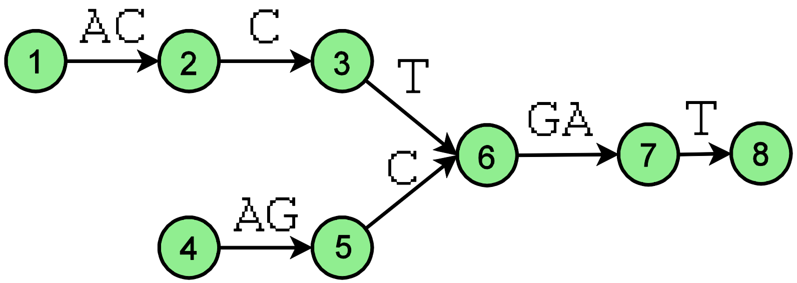
1. Overlap graph



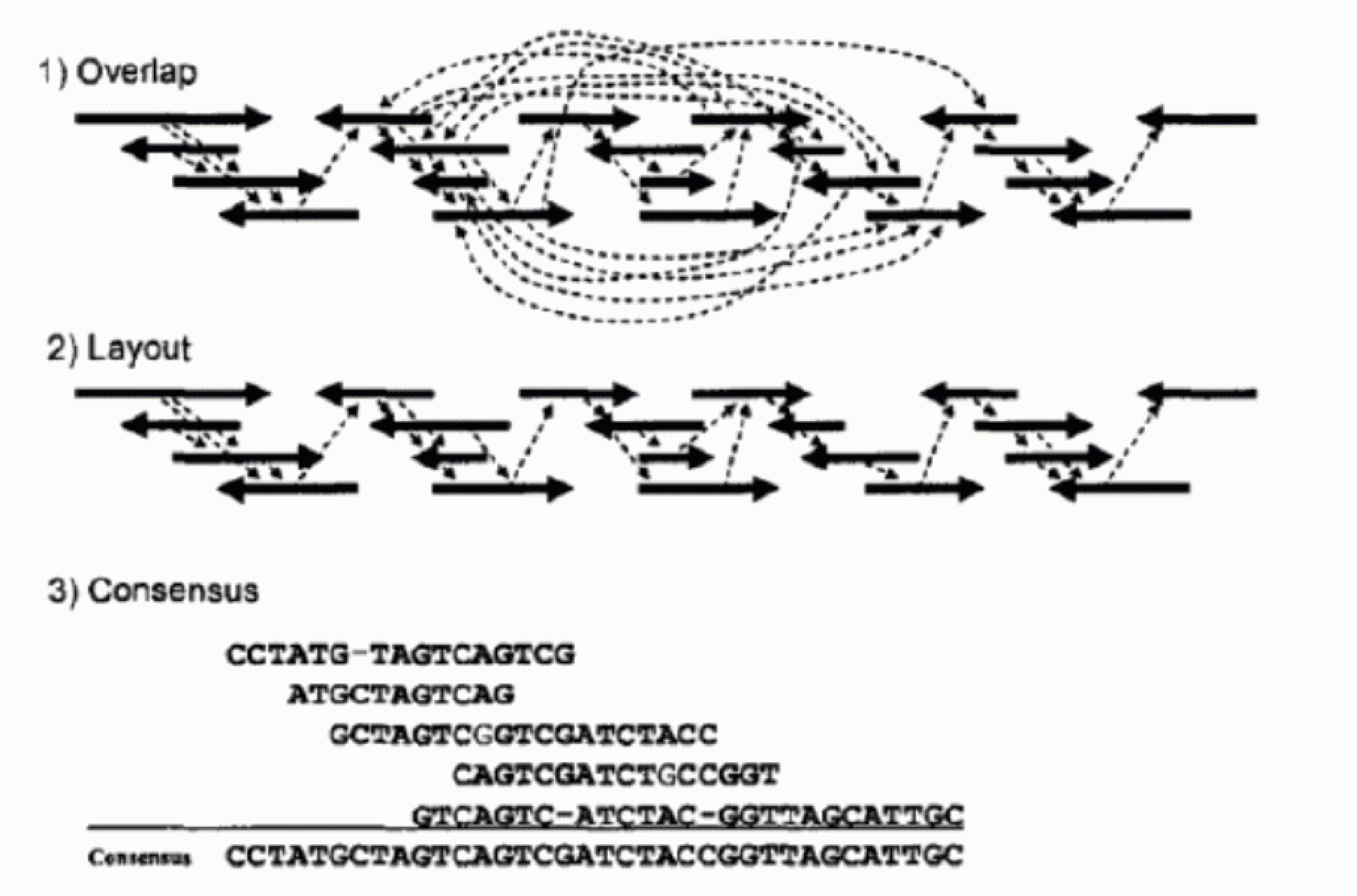


2. de Bruijn graph

3. String graph



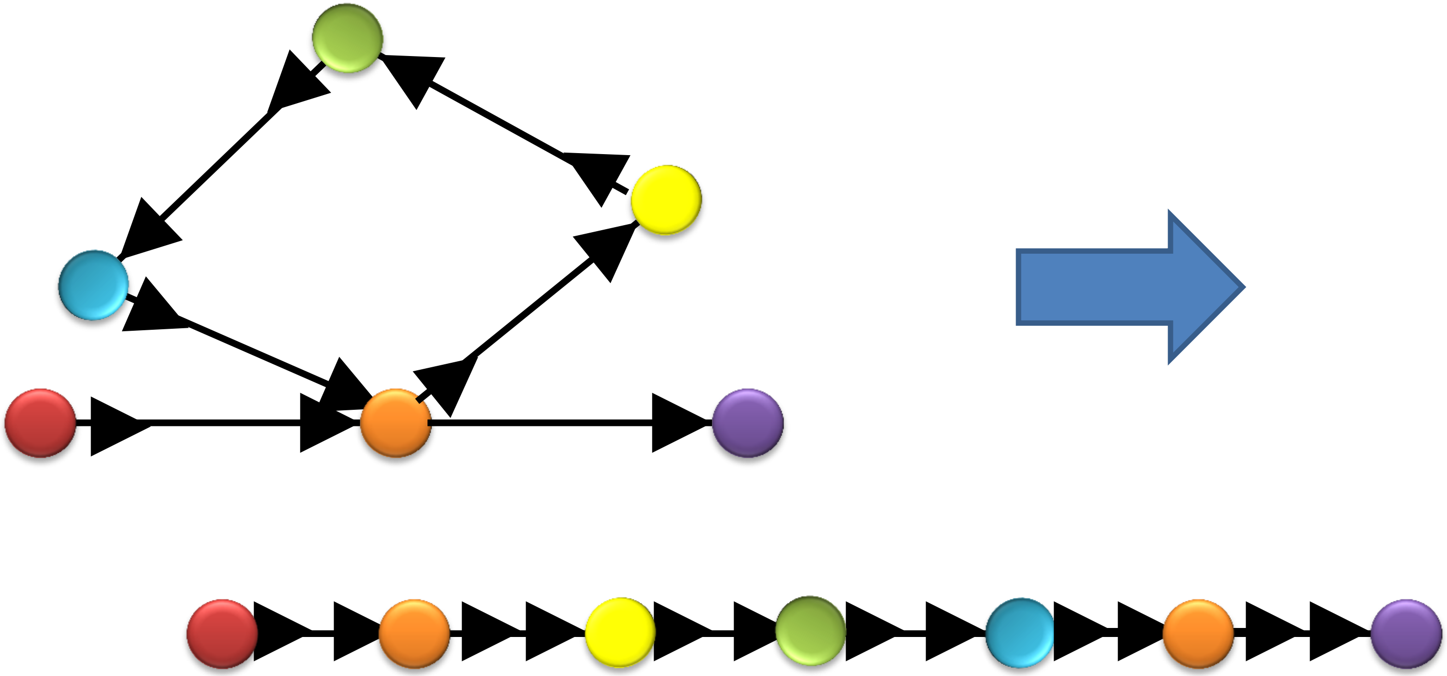
The Classic Overlap, Layout and Consensus Method



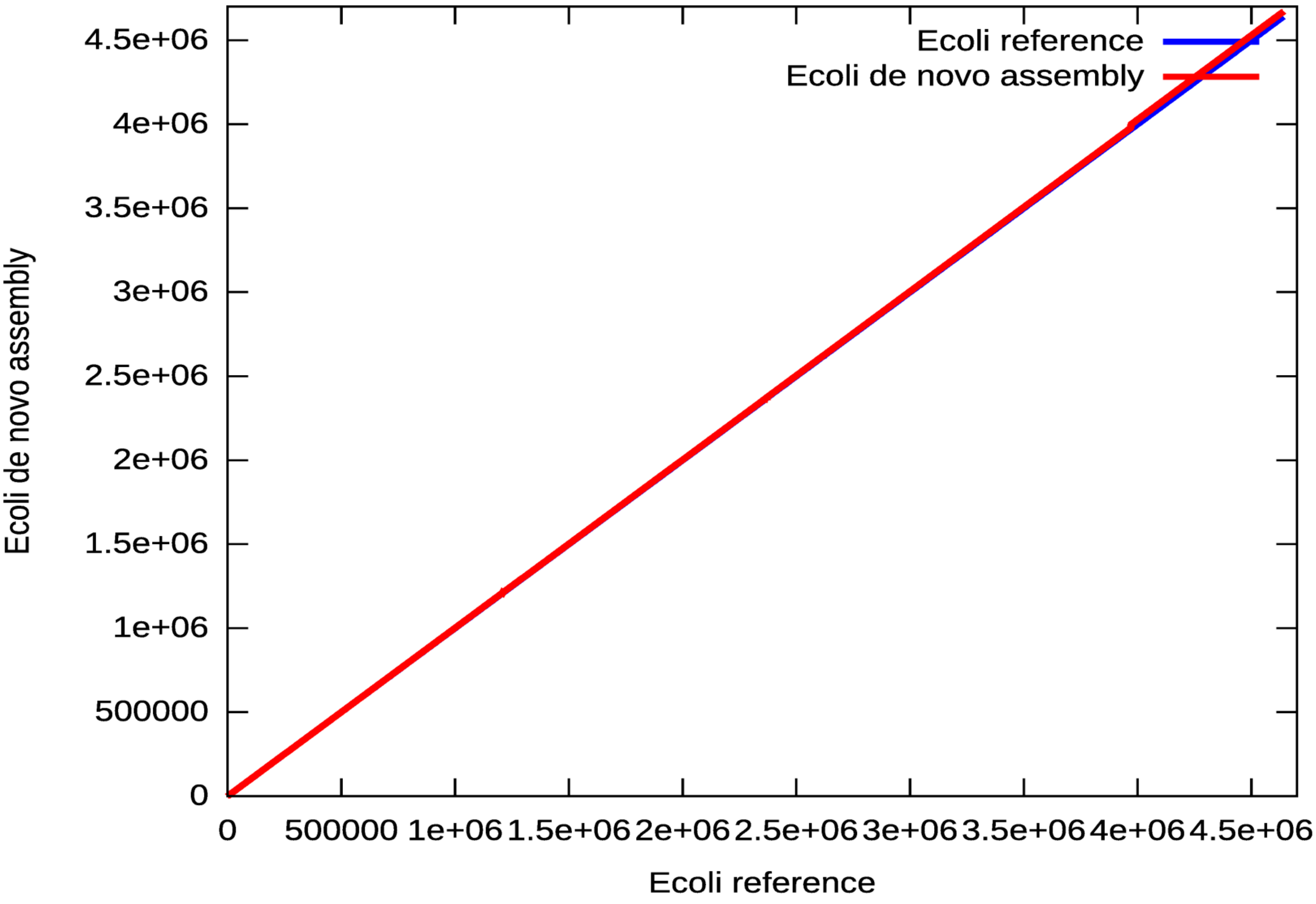
```
          CCTATG-TAGTCAGTCG
             ATGCTAGTCAG
               GCTAGTCGGTCGATCTACC
                     CAGTCGATCTGCCGGT
                   GTCAGTC-ATCTAC-GGTTAGCATTGC
Consensus CCTATGCTAGTCAGTCGATCTACCGGTTAGCATTGC
```

The Greedy Graph Based Method

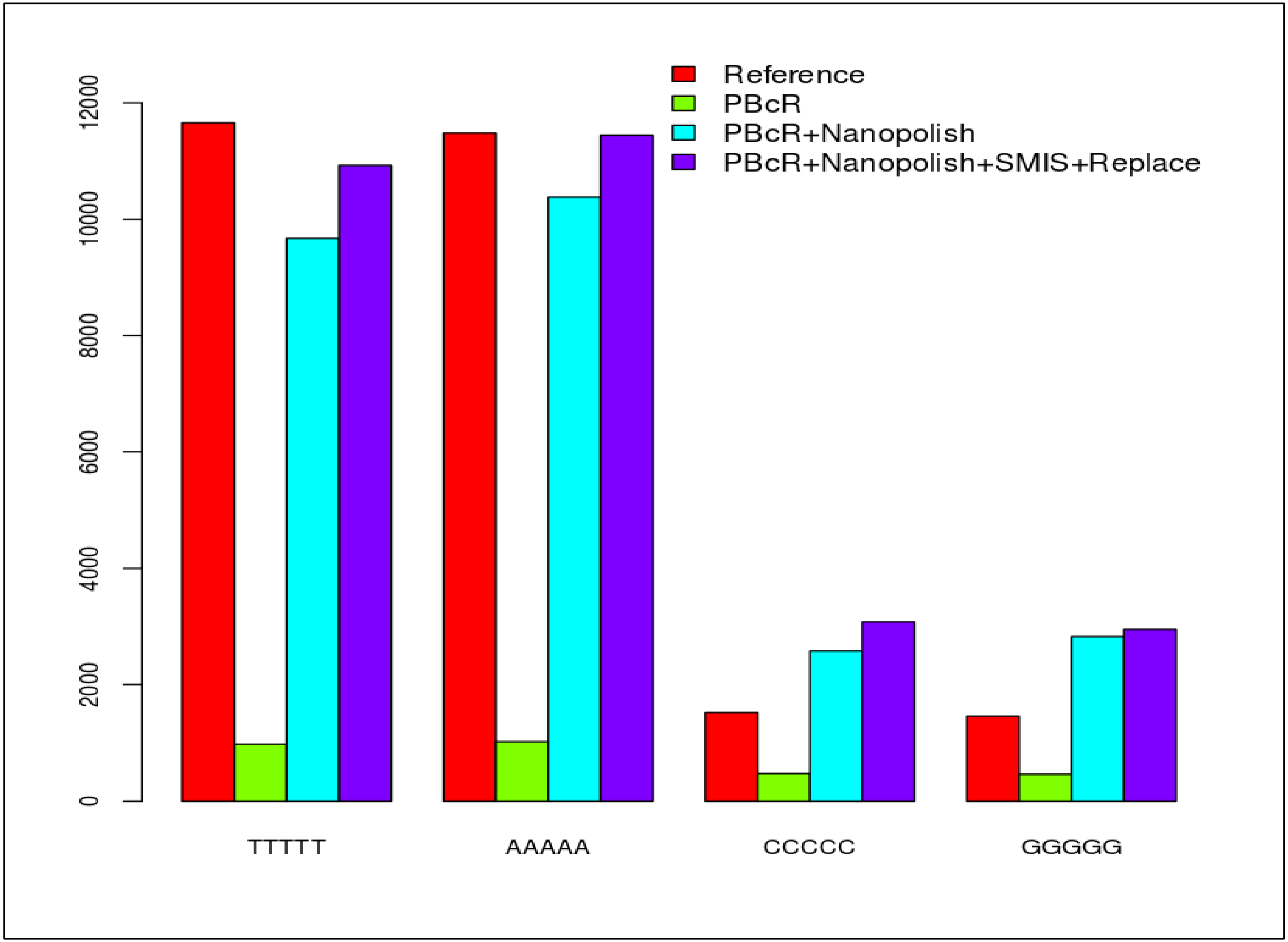
The greedy algorithms are implicit graph algorithms. They drastically simplify the graph by considering only the high-scoring edges. As an optimization, they may actually instantiate just one overlap for each read end they examine.



One Contig for The Ecoli Genome



Missing Homoplymers Recovered by Nanopolish from the Event Data



Assembly of Ecoli from Different Methods

	Total bases	Contigs	Mismatch_bp	Indel_bp	Identity
1. PBcR with nanopolish	4542223	2	2422	19928	99.52
2. SMIS_overlap with nanopolish	4671545	1	2958	22231	99.46
3. Jared Simpson's assembly^		1	1202	17241*	99.5
4. SPAdes with ONT and MiSeq	4651303	1	321	2058	99.95

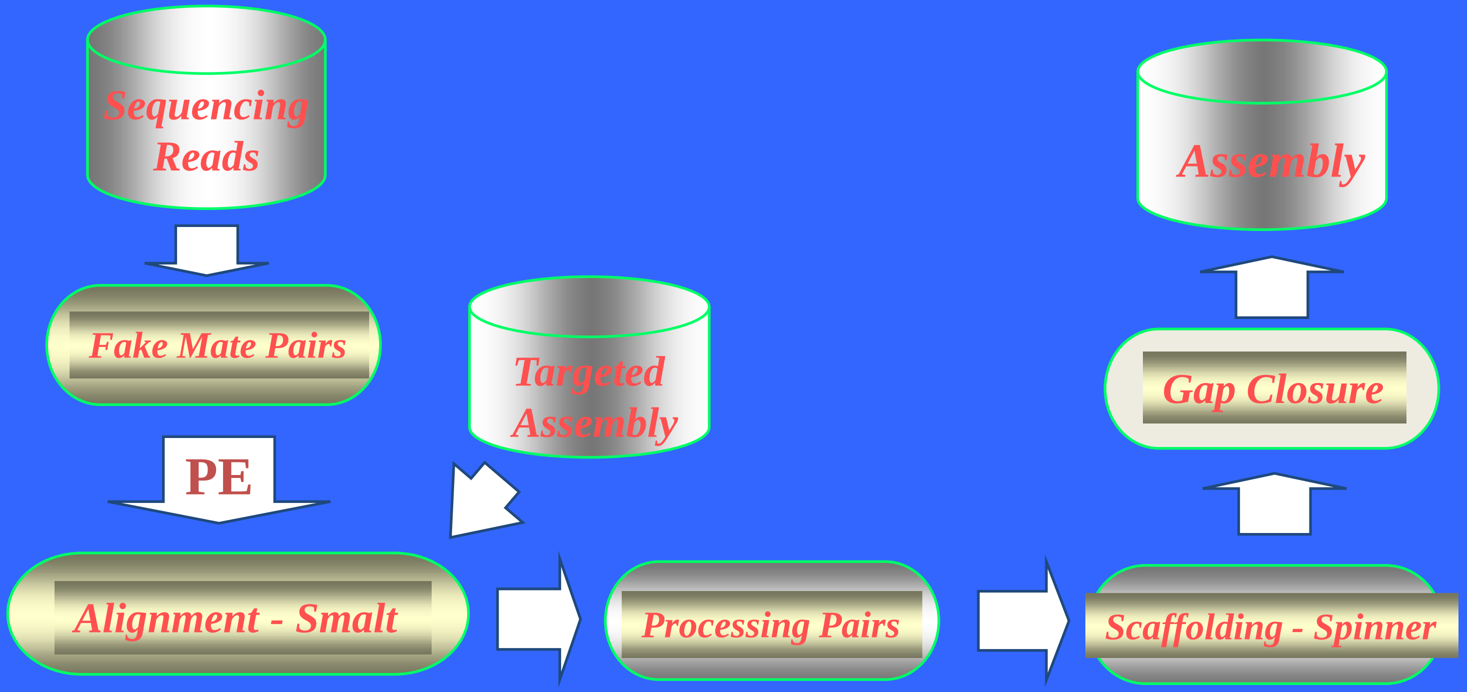
(i) Assemblies of 1,2,3 were obtained from ONT data only, while assembly 4 used both ONT and MiSeq reads;

(ii) Assemblies of 1 and 2 were obtained after using nanopolish;

(iii) * - in Assembly 3, the indel information is the number, rather the bases;

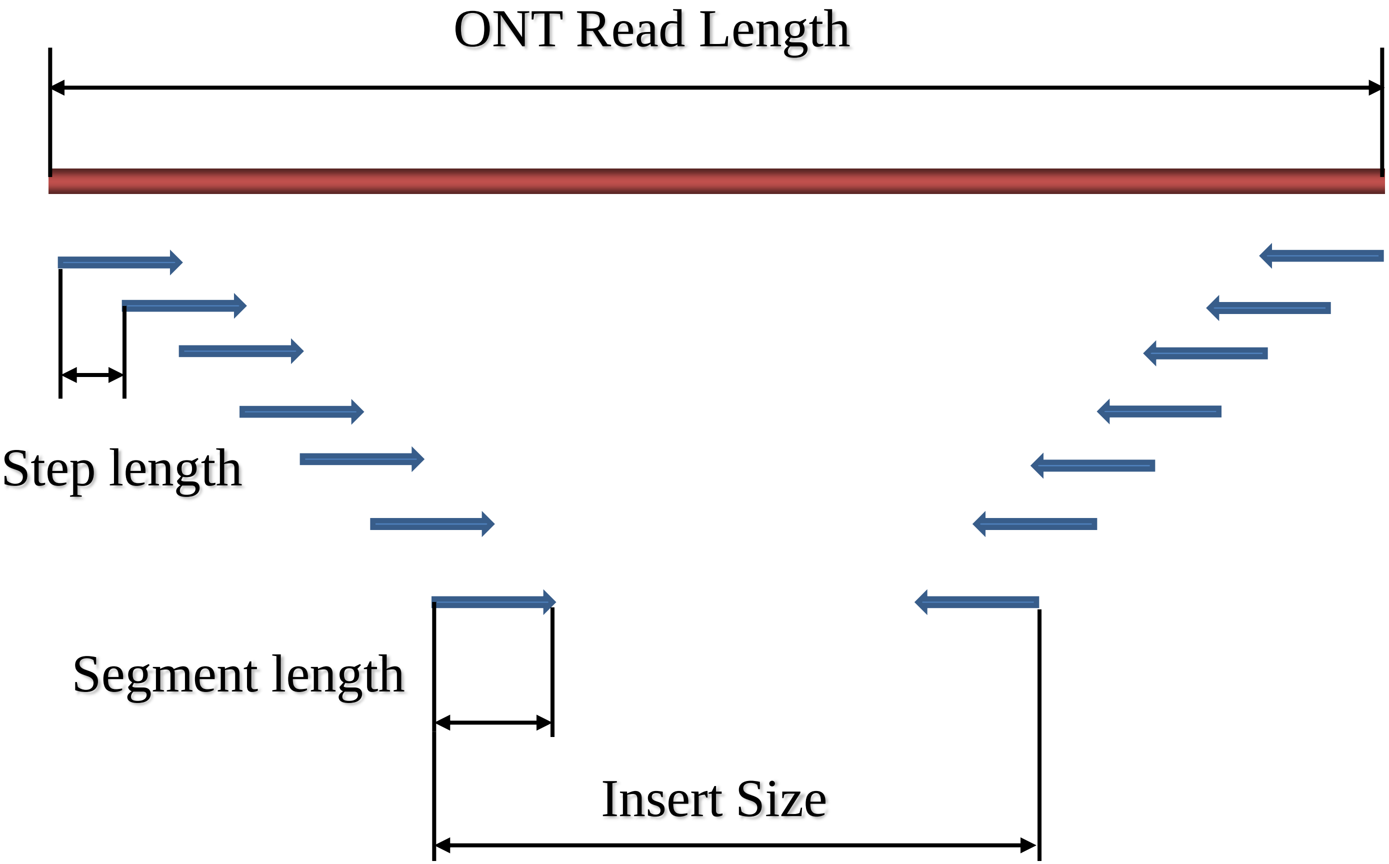
(iv) ^Loman NJ, Quick J, Simpson JT: A complete bacterial genome assembled *de novo* using only nanopore sequencing data. *Nat Methods.* 2015; 12(8): 733–735.

Single Molecular Integrated Scaffolding (SMIS)



SMIS: http://sourceforge.net/projects/phusion2/files/smis/

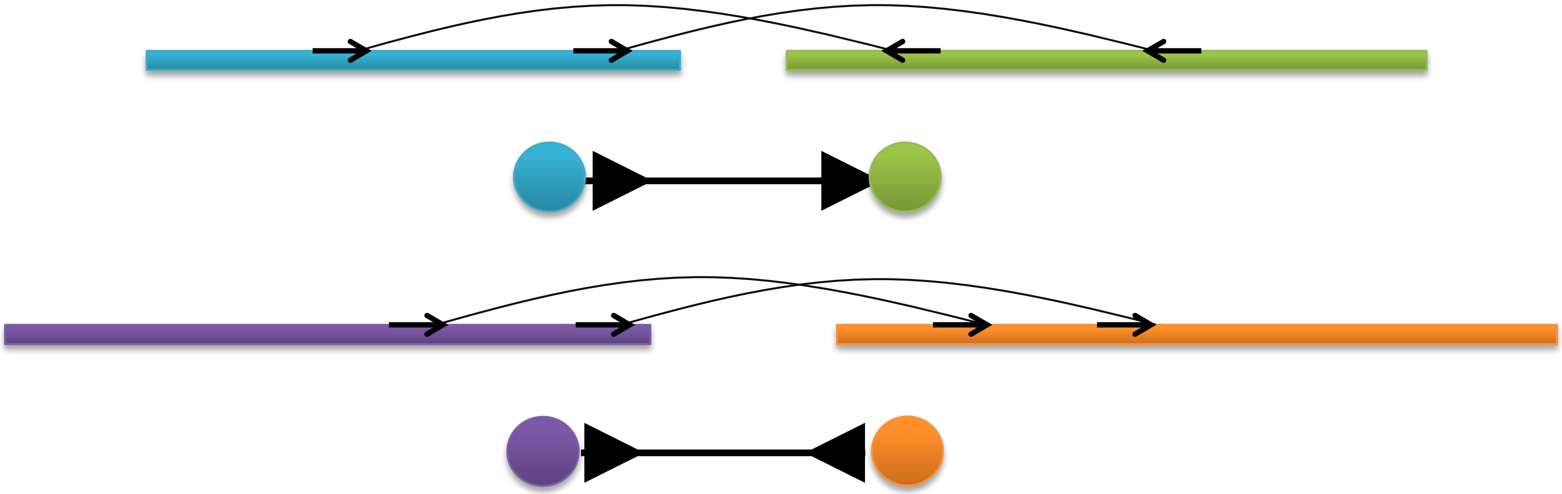
Fake Mate Pairs from ONT Reads



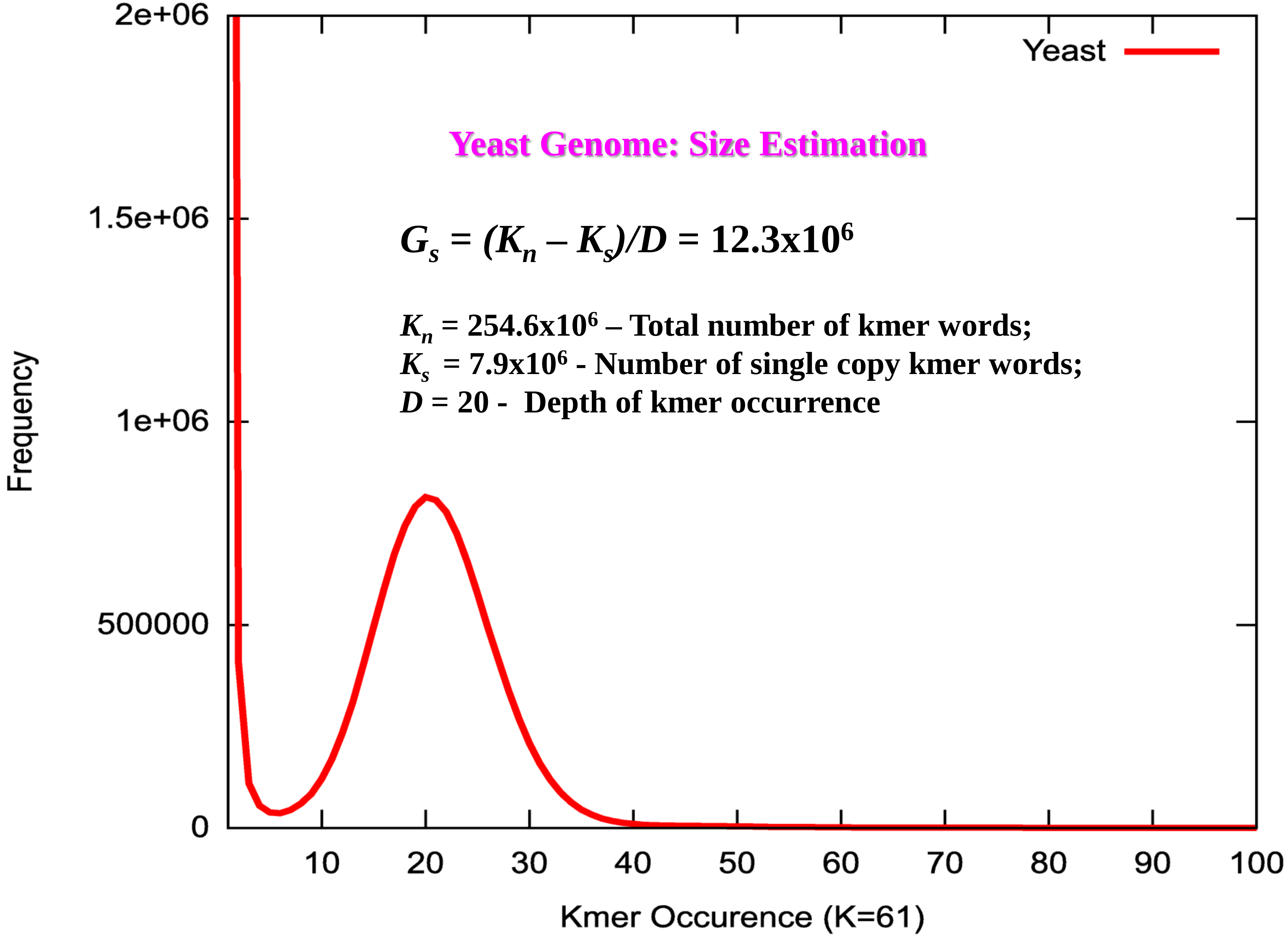
ONT Assisted Scaffolding

http://sourceforge.net/projects/phusion2/files/smis/

Mate pair data is used to scaffold contigs. Contigs, and pairs of contigs connected by pairs, define a bi-directional graph:



Using expected insert size, a estimate of the gap size can be given for each contig.



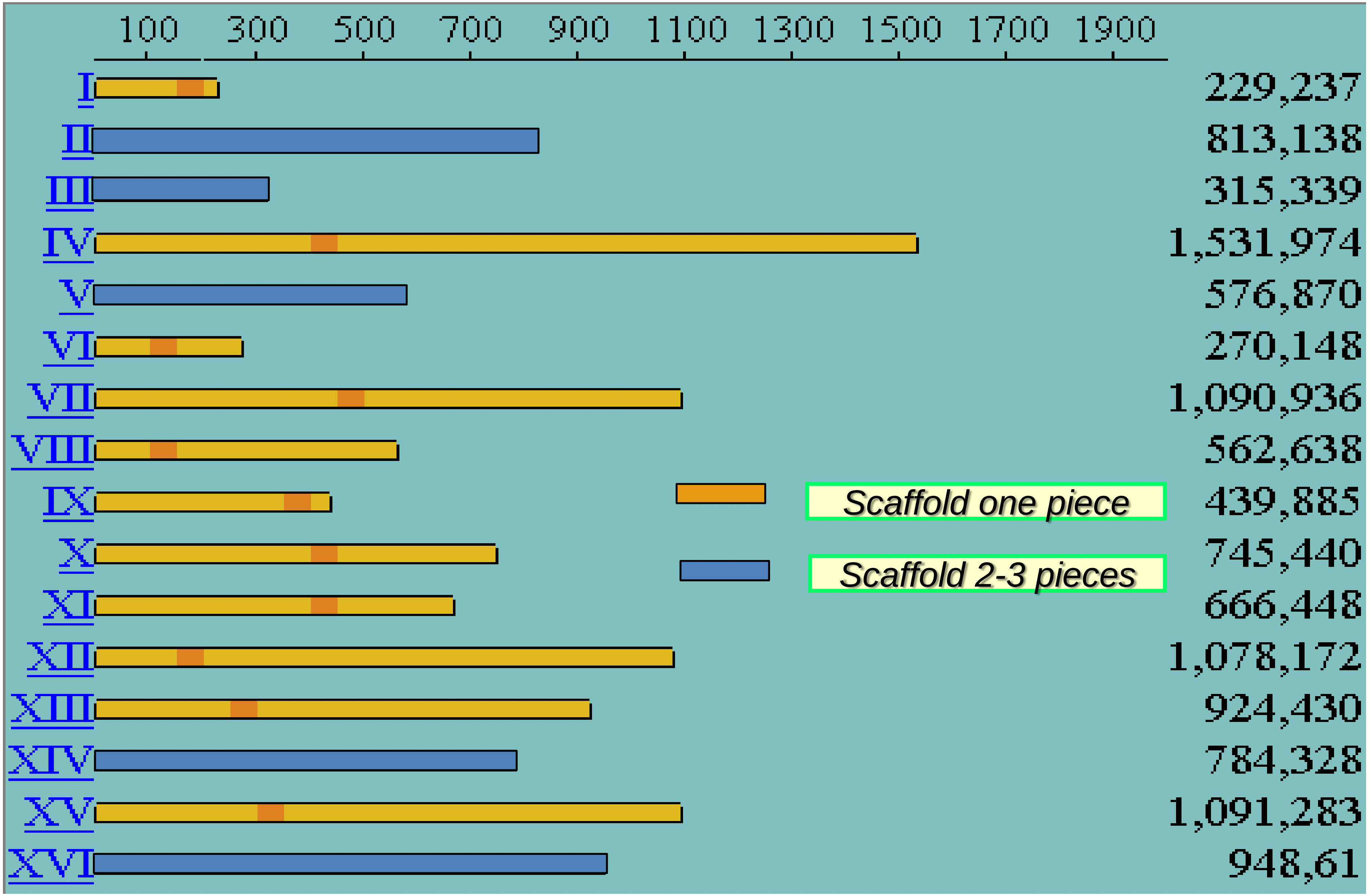
Yeast Genome: Size Estimation

$$G_s = (K_n – K_s)/D = 12.3\text{x}10^6$$

K_n = 254.6x10^6 – Total number of kmer words;
K_s = 7.9x10^6 - Number of single copy kmer words;
D = 20 - Depth of kmer occurrence

Saccharomyces cerevisiae complete genome

Scaffold N50 858Kb ; Contig N50 330Kb



Yeast W303 Assembly from PacBio Data using PBcB

- **Data:**

http://datasets.pacb.com.s3.amazonaws.com/2013/Yeast/

- **33 contigs and N50 = 777023**
- **12 out of 17 chromosomes are covered with a single contig**
- **99.95 % identity compared with assembly from Miseq**
- **No major homoplymer problems!**

Table 3 CSHL W303 Yeast Illumina Reads Used for Assembly[+]

Insert size	Library number	Total paired reads (m)	Read length (bp)	Sequence depth* (X)
550 bp	1	25.2	2x300	1200
550bp	1	6.0	2x300	300

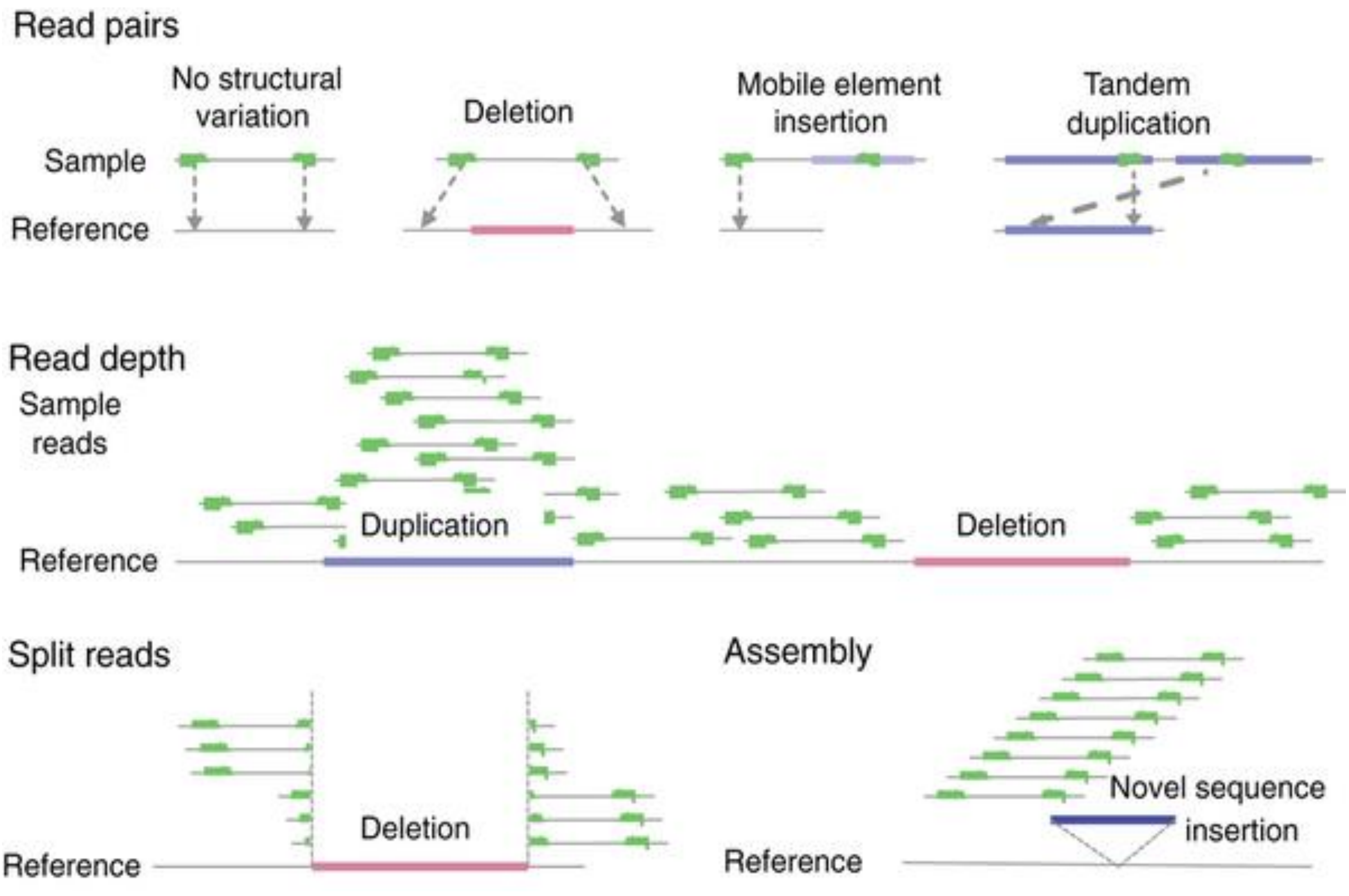
[+] The dataset was downloaded from http://labshare.cshl.edu/shares/schatzlab/www-data/nanocorr/

Table 4 W303 Yeast Assembly Stats

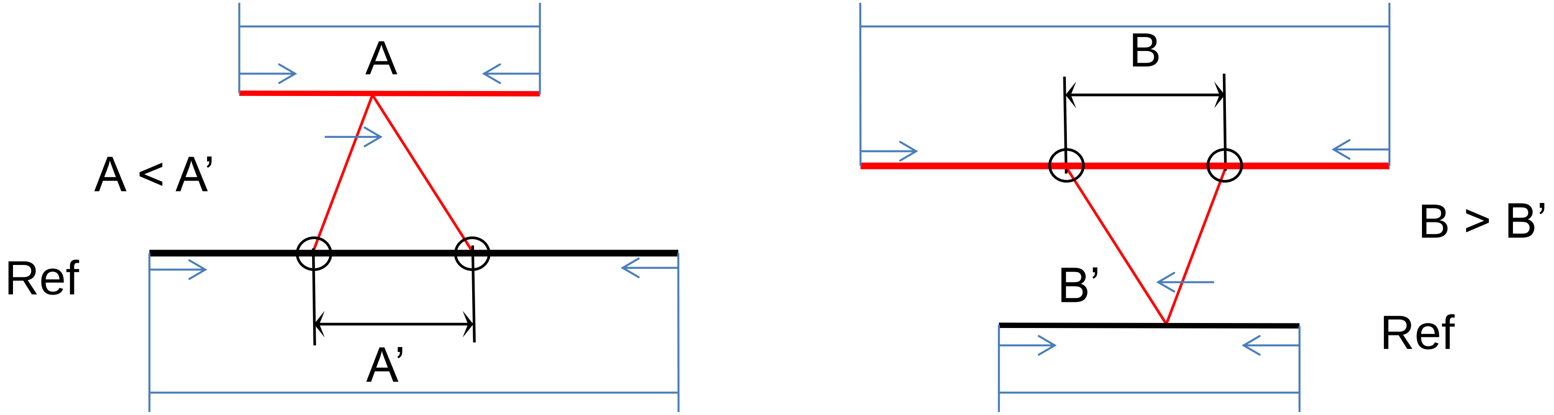
	Fermi	*SOAPdenovo**	*MaSuRCA*	*SMIS-Merge+*
Total bases of scaffolds (Mb)	11.8	11.7	11.9	11.8
Number of scaffolds	804	424	473	334
Scaffold N50 (bb)	124288	201711	247249	857808
Scaffold N90 (bp)	29458	58167	54929	251279
Maximum scaffold length (bp)	437507	4571744	701450	1442956
Total bases of contigs (Mb)	11.8	11.7	11.9	11.7
Number of contigs	804	432	495	385
Contig N50 (bp)	124288	186331	20203	329536
Contig N90 (bp)	29458	52862	5929	76150
Maximum contig length (bp)	437507	451744	75044	677392

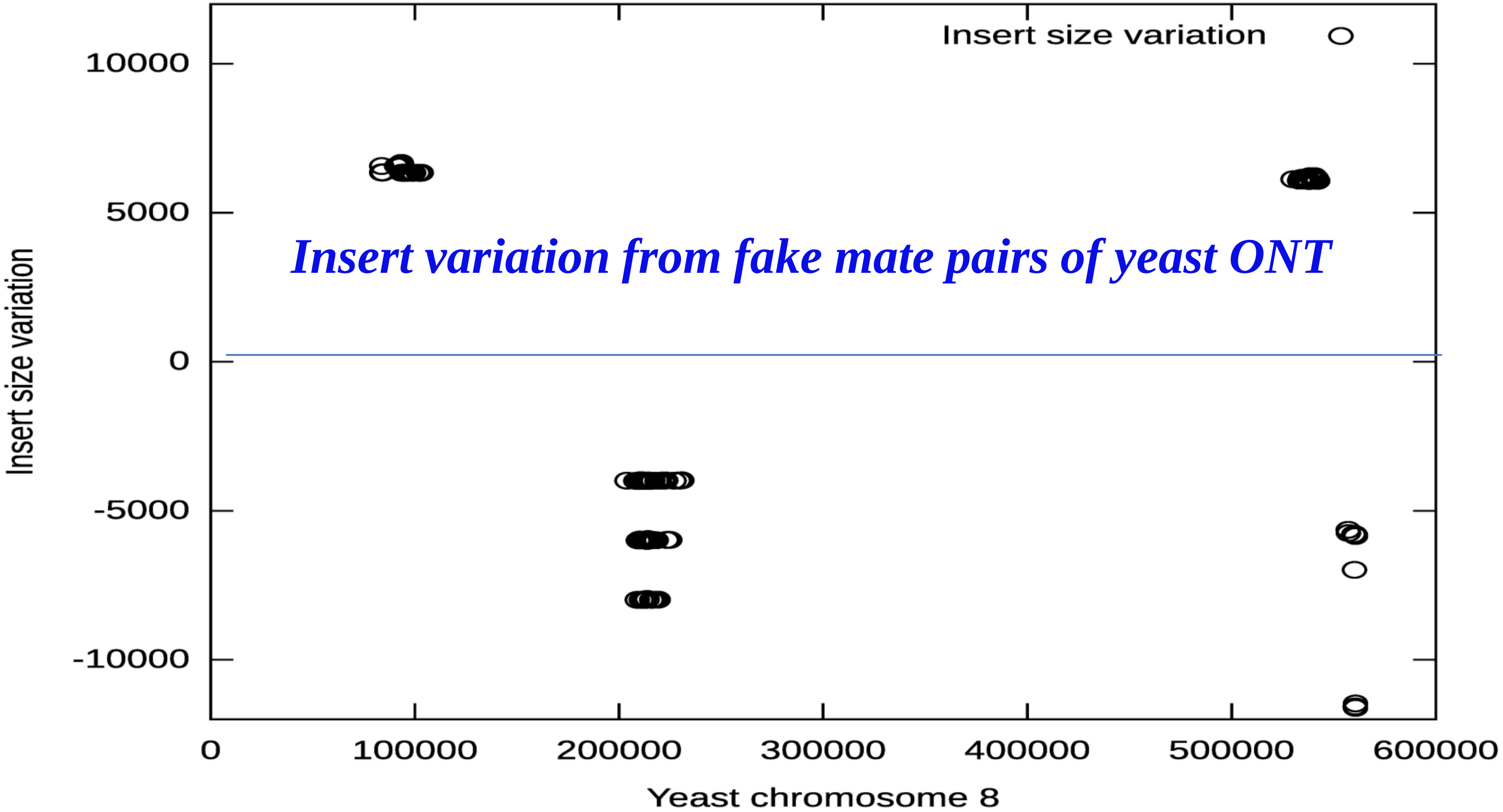
SOAPdenovo - reads were processed and base errors corrected using our own tools;*
SMIS-Merge+ - Scaffolding was performed using SMIS on the merged assembly and contigs were processed using our own tools.

Methods of Structural Variation Detection



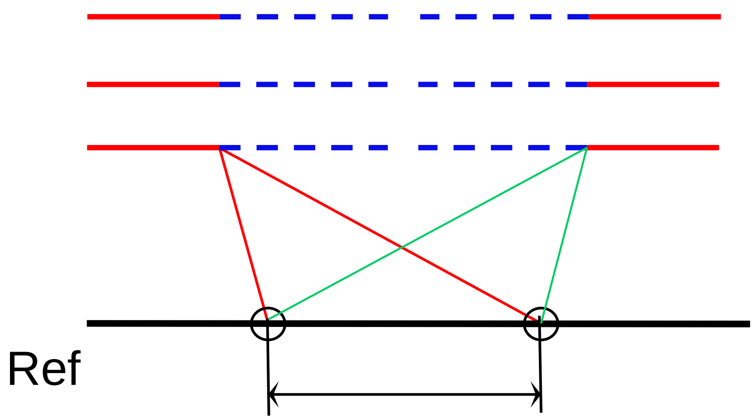
Read pairs – Examining Insert Size Variation

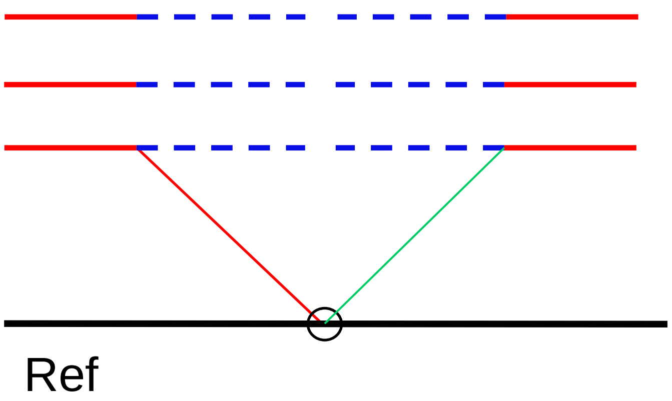




Insert variation from fake mate pairs of yeast ONT

Split Reads – Identifying Breakpoints





CIGAR:

M??H??	*H??M??*
M??S??	*S??M??*
M??H??	*H??M??*

CIGAR:

M??H??	*H??M??*
M??S??	*S??M??*
M??H??	*H??M??*

Parsing the alignment CIGAR strings and looking for common breakpoints with hard or soft clipping “H” or “S”

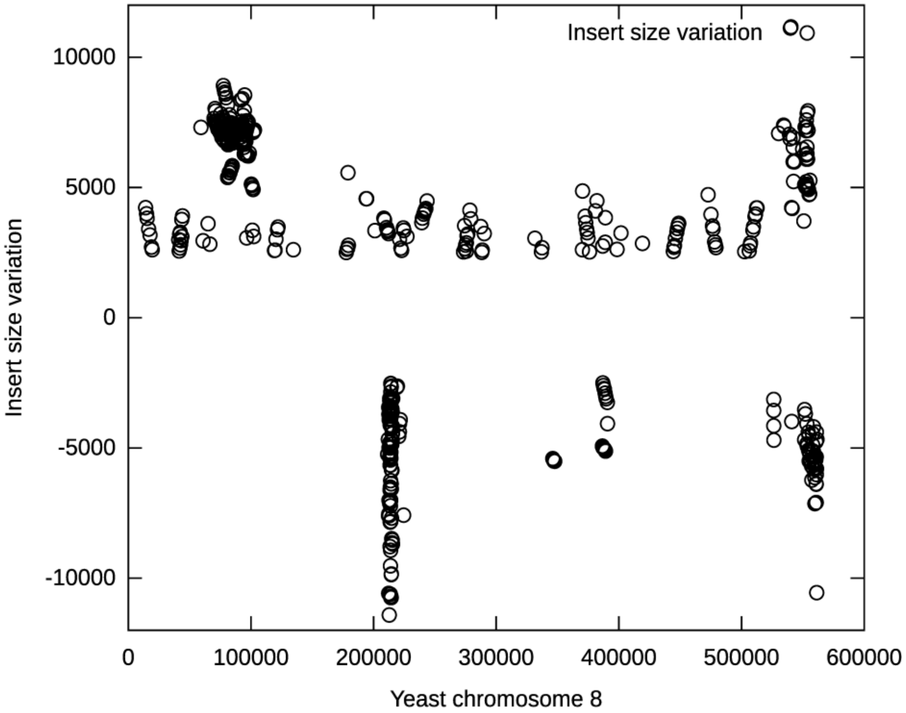
Normalised Insert Size Variation Factor

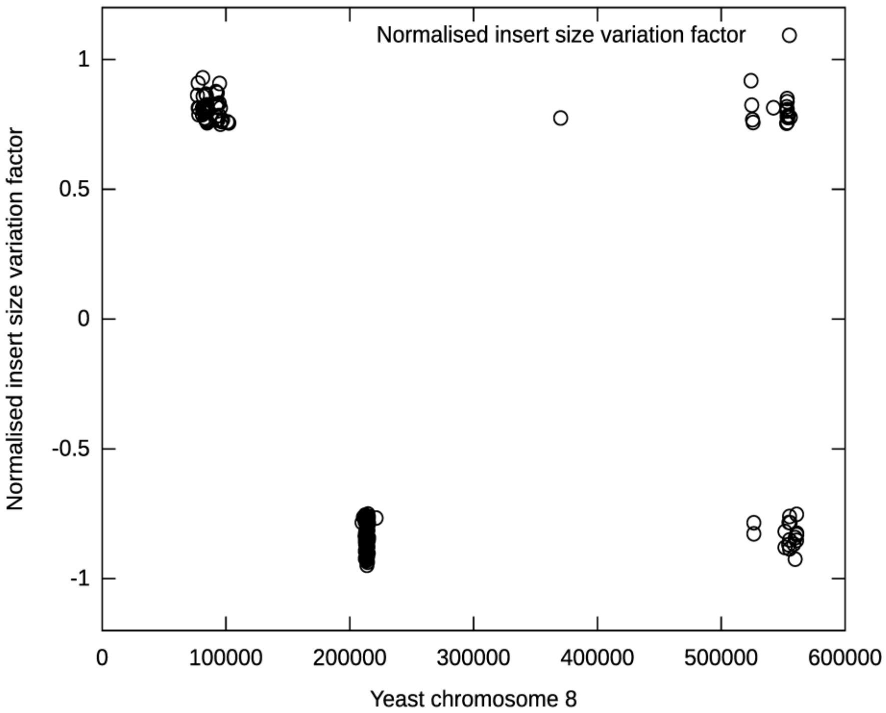
There are N mate pairs of sequences which can be mapped to a reference chromosome:

$$P_i = 1 - \left(\frac{C_i - C_{i\_1}}{D_i} \right)^{0.3} \qquad 0 \le i < N \text{ and } 0 \le \frac{C_i - Ci_{\_1}}{D_i} \le 1$$

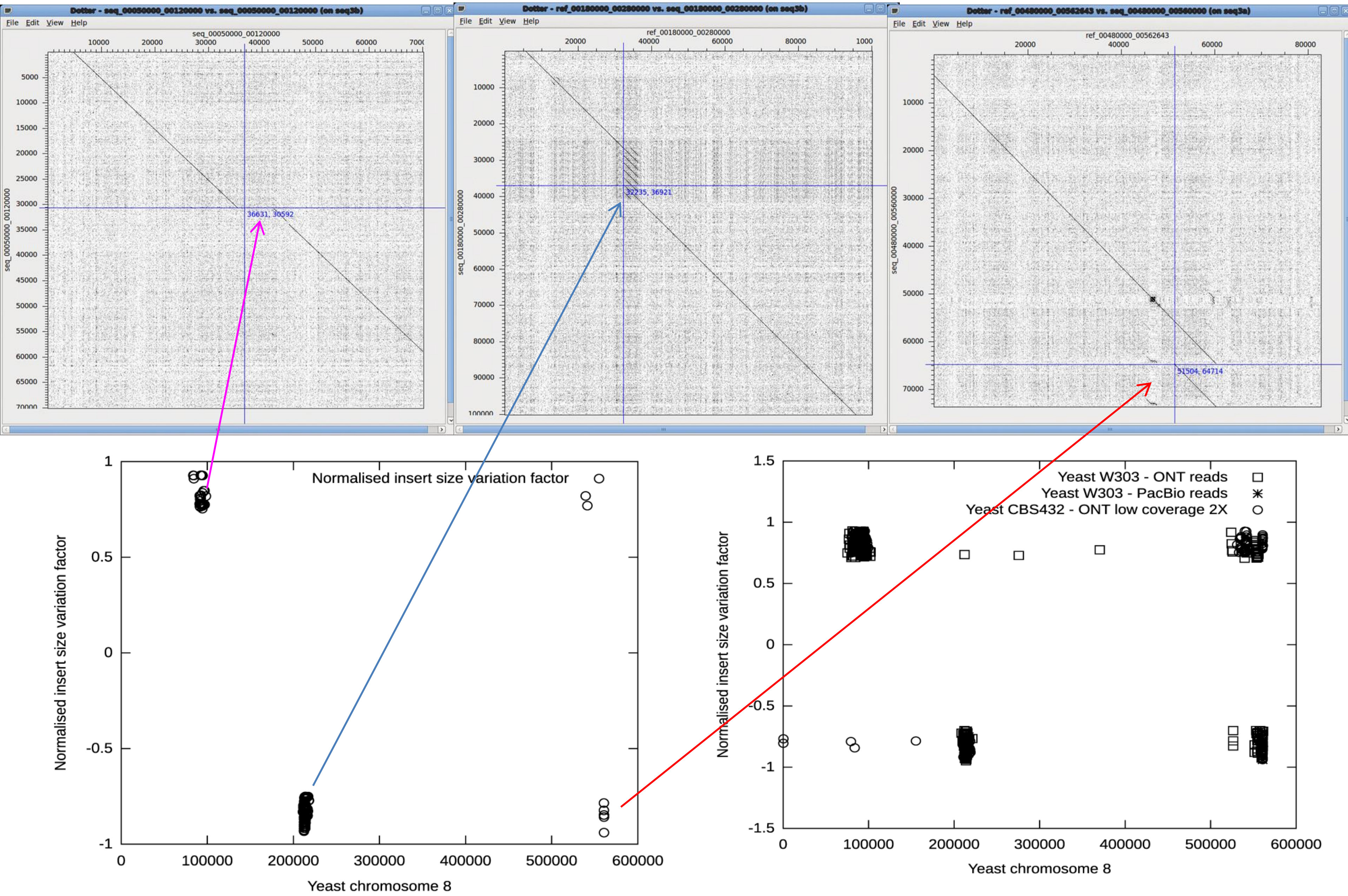
where C_i – *mapping coordinate of the ith pair on the chromosome;*

D_i – *insert size difference between the shredded distance and the value estimated from the alignment.*



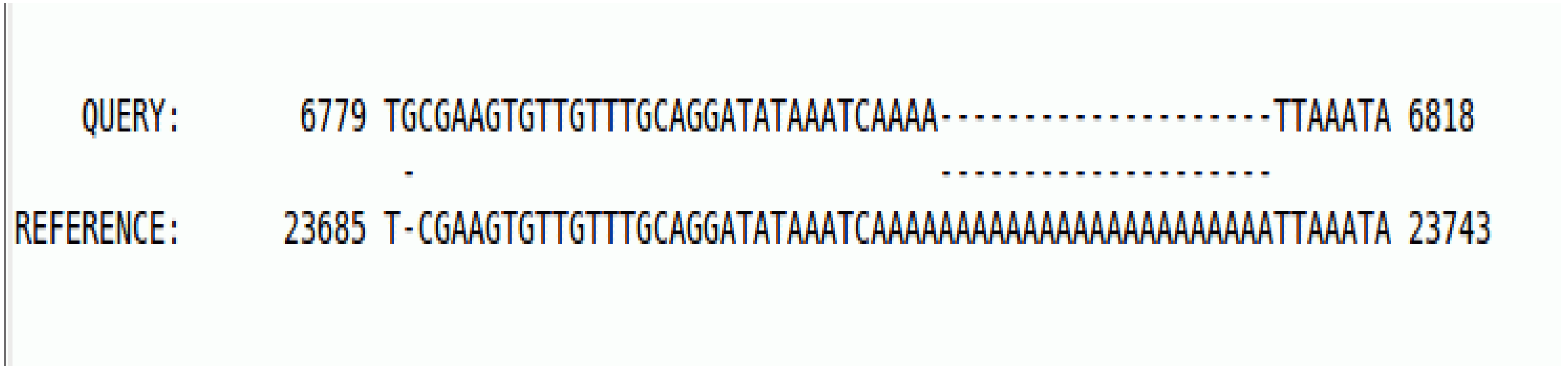


CNVs in Yeast Chr8 Comparison – SC288C vs W303



Summary:

- For de novo genome assemblies, nanopore data contributes to impressive contig/scaffold continuity;
- Missing homoplymers is the major issue on contig base quality;
- PacBio shows advantages in genome assembly, so far;
- Detection of structural variations is still a challenging task, while Oxford MinION data offers exciting chances.



```
    QUERY:       6779 TGCGAAGTGTTGTTTGCAGGATATAAATCAAAA--------------------TTAAATA 6818
                       -                               --------------------
REFERENCE:      23685 T-CGAAGTGTTGTTTGCAGGATATAAATCAAAAAAAAAAAAAAAAAAAAAAAATTAAATA 23743
```

Acknowledgements:

- Richard Durbin
- Louise Aigrain
- Francesca Giordano
- German Tischler
- Hannes Ponstingl
- James Bonfield
- Rob Davies
- Thomas Kean
- David Jackson
- Tony Cox

ONT Ecoli reads –	UCSC
Miseq Ecoli data –	CSHL
ONT Yeast data –	CSHL
Miseq Yeast reads -	CSHL
PacBio yeast data -	PacBio

MARC