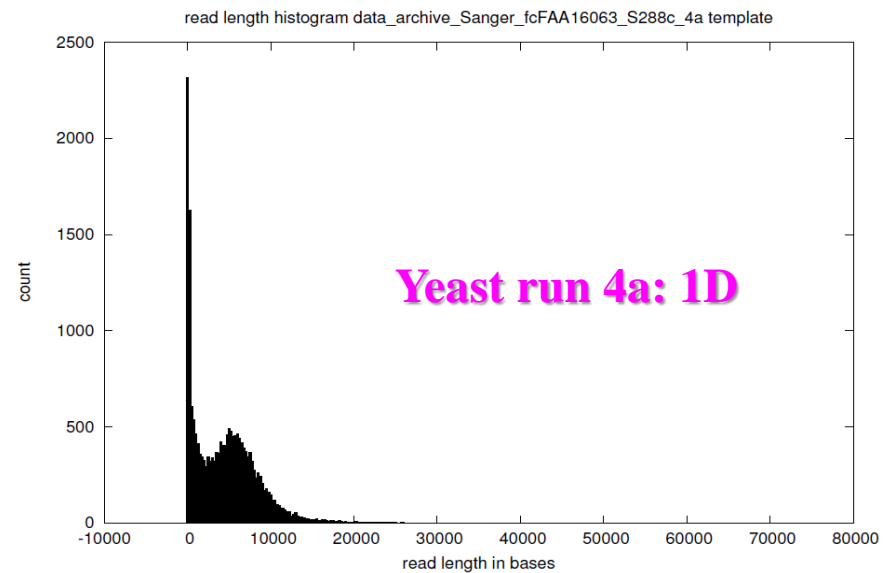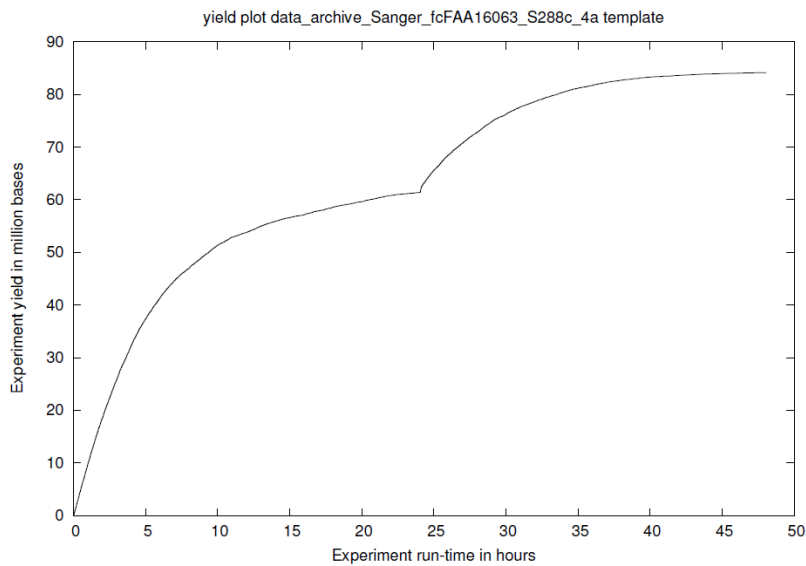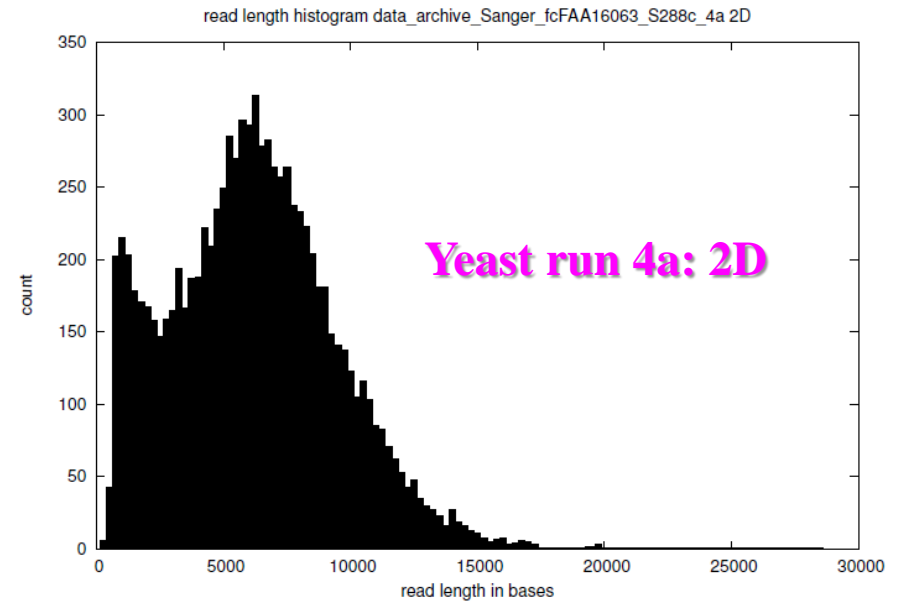# Oxford Nanopore
## The Data, the Alignment and the Assembly

**Zemin Ning**

**High Performance Assembly**

**Sequencing Informatics**

GGTTGTTTCTGTTGGTGCTG

**a**

SXXXXXXX
XXX
XX
XXX
ATGGTAAGTAAGTCTAGAGTGATA CGTAAGAGTACGTCCAGCATCGG~ 5´
CTCACCTATCCTTCCACTCATACTATCATTATCTACATCXXXXXTACCATTCATTCAGATCTCACTATCGCATTCTCATGCAGGTCGTAGCCXS

+29 +25          $n = 0$

**b**

iv
25
pA
22

iii

i
36
2 s
32
pA
28
ii
24
20

v
iv

vi

55
50
Event signal
45
40
35
30
0    1    2    3    4    5    6    7    8
Time (arbitrary units)

TTTTT   CTTTT   TCTTT   TTCTT   TTTCT   TTTTC   TTTTT
TTTTT   ATTTT   TATTT   TTATT   TTTAT   TTTTA   TTTTT

5' ← TGCGATACTCATCGCA → 3'

↓

C   T
A   C
T   A
A   T
G   C
C   G
G   C
T   A
5' ←     → 3'
Hairpin

## 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 at examining the same loci. The advantage being that the complementary strand will have a different kmer profile.

# *Throughput vs Time and Real Length Distribution*



yield plot data_archive_Sanger_fcFAA16063_S288c_4a 2D

read length histogram data_archive_Sanger_fcFAA16063_S288c_4a 2D

**Yeast run 4a: 2D**

yield plot data_archive_Sanger_fcFAA16063_S288c_4a template

read length histogram data_archive_Sanger_fcFAA16063_S288c_4a template

**Yeast run 4a: 1D**

# Read Length Distribution – 1D and 2D



Yeast S288C ONT reads sequenced – 1D & 2D

| | 1D | 2D |
|---|---|---|
| Number of bases: | 275Mb | 373Mb |
| Number of reads: | 47735 | 66106 |
| N50: | 8335 | 6645 |
| Average length: | 5765 | 5651 |
| Maximum length: | 143706 | 32826 |

Left terminal:

```
alignment:S:19 320   channel_202_read_95_0 chr14      126    4112 76256330 76260455    F    3279 75.
718

    QUERY:       126 TCTATCTCGT-CCACTTGGTCTATTTGCCTATCCT-GTGACAATTCCAAGAGCTCTTAAT 183
                     i  -   -                 -           vivvv
REFERENCE:  76256330 TCTATCTTGTTCCA-TTGGTCTATTTGCCTATCCTCGTGACAATTCCATACTATCTTAAT 76256388

    QUERY:       184 TGCTGTGGAATTCCACATTGA-TCTTA-TTTAATTTCATCTTT--ATCCTTC-TGCCTAT 238
                         -                    -    -            v    -  v
REFERENCE:  76256389 TGCTGTGGAATTCCAC-TTGAGTCTTACTTTAATTTCATCTTTGTATCCTTCGTGCCTAT 76256447

    QUERY:       239 CAGCATTGTGCTTGGCCACATAATATATGGCAAAGATATGTTAAGATGAAACCTAGCTCG 298
                      -  --  -         v iv  i-         vvvv-  v
REFERENCE:  76256448 CA-CA--GTGCTTGGC-ACATAATATATG-CAAATACTTGTTG-GATGAAAGGGC-CTCT 76256500

    QUERY:       299 -GTACTGGTGTTTCTG-ATCCTTCATCACAGGCAATGATCGGCTTAGG-TAGAA-GAAGG 354
                      -        v iv -  iivi      i-   v vviv-v     -   -
REFERENCE:  76256501 AGTACTGGTTTCTGTGGATTGCCATCAT-GG-AAGGCCCTG-TGAGGGTAGAAAGAAGG 76256557

    QUERY:       355 --ACCT-ATTCTTCTCCCAGGCCGCGTACCGGCTTGGAGGG--CCTT--ACGTGAGAGGC 407
                      --  i-    i          vi v vviv       --  i  -- v ivv  v
REFERENCE:        76                                                            6256617
```

Left overlay box:
**Mapping score:      19**
**Match identity:     75.2%**

```
    QUERY:                                                                 67
REFERENCE:  76256618                                                      76256672

    QUERY:       468 TAGTGAATATGGGAAGCTGTGA-AGAGTACGG-TGTTAGA--TGG-ACAGCATTATTGAA 522
                      i vv           -   - i      -  --      v iv i
REFERENCE:  76256673 TAGTGGAGCTGGGAAGCTGTGACAGAG-ATGGATGTTAGAGGTGGGACAGCAGTGGTGGA 76256731

    QUERY:       523 --ACGT-AGCTGGATCCC-TTCACAATGAGCCCGAT-GCACTGTTGCGTCCTCCTCCTCC 577
                      --  v  -            v---      -viv -v- -  -
REFERENCE:  76256732 GTAGGTTAGCTGGATCCCCTTCACAATGAT---GATAGCA-TCCAGC-TG-TC-T--TCC 76256782

    QUERY:       578 TTGGGGGAAGCCGAAGAGAAACAGGCATAACCTAA-GCTGTCTACCAAATCGA-CCTCAA 635
                      v- v      -  vvv    v       -    v -
REFERENCE:  76256783 TTGGGGGAAGG-GCAGAGAA-CATTAATAAACTAAAGCTGTC-ACCAAATCTAACCTCAA 76256839

    QUERY:       636 AGCCCTCCCATCAGCTCTTCCCCAACACCTAACCCA-ACTGCTCCTTTGCTCCACGAAGG 694
                      -          -   - -        i -iv -              v--
REFERENCE:  76256840 AGCC-TCCCATCAGC-CT-CCCCAACACCCA-CTGAGCTGCTCCTTTGCTCG--GAAGG 76256893

    QUERY:       695 GATAACCCTTGTCTCATCACGATAACGGTGGGTCATCTCCTCCC-ATCTCCCAATGGGAA 753
                      -  vi -  i  -   -- vv v vi                  -
REFERENCE:  76256894 -AGGACC-TCGTCTCA--AGCAGACTGGTGGGTCATCTCCTCCCATCTCCCAATGGGAA 76256949
```

Right terminal:

```
alignment:S:04 5493   channel_316_read_16_2 chr5      7199       1 136244417 136252515    C    4870 53.42
9

    QUERY:      7199 GCGAACATACGCG-T-GCCA-ACGC-TGACATCTTGATGAC-GTACAAACGGGATCTCAG 7145
                      -    v v v - -  - - - v-- -    v --- - vi  i   vvv
REFERENCE: 136244417 G-GAAGAGAGG-GATTGC-AGA-GAAT-AC-TCTGG--G-CAGTAGGAACAGGATAGAAG 136244467

    QUERY:      7144 CTTCA-AC-GATCC--TGCCGGGACAGACTGCCGGTAGAGTTAAA-AC---GCCGCAGGG 7093
                      -    v --     v--     ----   v  i iivv-v iv-  --- v v v  v
REFERENCE: 136244468 -TGCACACTG-TCAAATGCC----CAGTCTACTGTTT-ATTTGACTACTTTGACTCTGGC 136244520

    QUERY:      7092 GTGGATTTACGC--G---A--CACACGGTC--GTACGATAAACCCCGTTTT-TC-GTAAG 7044
                      v vv-  - v ----- v v iv -- i -    v v-- v-v-i -
REFERENCE: 136244521 TTCC-TTT-CTCCTGTTTATTCCCTCATTCTTGTTC-ATAAAGCA--TTTAATGAGCAA- 136244574

    QUERY:      7043 A-AAC----GCA-GCCAG-ATGGT-GG-T-GACACGCATTTACGAGG-C-C--GGGCTTC 6998
                      -  ---- - v v- v --- v i vv--- - ---- --
REFERENCE: 136244575 ACAACCTTTGCAAGGCACCATGCTAGGCTAGAAACACAGG-A-G-GGACACCAGGG-T-C 136244629

    QUERY:      6997 -GTTCA-GAAATTTGTACGATC-TTCGCGCCTTCTCG-C-G-CGT--TGGGTGCAT-TAG 6948
                      - v --  --- v v - i- v -- i--v--v - -v-
REFERENCE:      1362                                                         136244681
```

Right overlay box:
**Mapping score:       4**
**Match identity:      53.4%**

```
    QUERY:                                                                 6898
REFERENCE: 136244682 CAG-TAGCAATGCATCCCATGTGGAGTGG-ATACATTTAGTTTTACCTT-TA-GCAGATC 136244737

    QUERY:      6897 TAG-GCGAATCAGGATTACGATG-T--AACCGGATATT-TA--CGAATAAGGCTG--AA- 6848
                      -  i  iv vv  vv i - --  v v-- - -- vv- v  - v -- -
REFERENCE: 136244738 TAGAGCAAATTTGCTTTTGGGTGATTGAAACT--TATTCTATTCCT-TCAG-CAGTTAAT 136244793

    QUERY:      6847 CAGCTAACGTATT----T-A-ACTAACTT-GTTCCATATTGC-GTTTTAAACGATG--CT 6798
                      -- v     -   vv    - --- -v i -i  --
REFERENCE: 136244794 CA-CTAA-GTATTGGAGTGATA-TAAAGTAGT-CCA----GCTGATTCAA-CAATGTGCT 136244844

    QUERY:      6797 TGATGCGCTAGACGA-GGTCCGCGTAGCCTCTTTGTCA--TACGATCAATCTGCACGGAT 6741
                      -- - v- v-  - --- - vv-v  v - - i --  v i v
REFERENCE: 136244845 TG--GC-CT-GT-GCTGGT--G-G-A-CAG-TGTGTAAGCTA-GACCA--CTGCCCAGAA 136244890

    QUERY:      6740 ACGATTGTCGCTAATAATAACCGAATGTCACGCA-TCCACGCGT-TTTCGCATACGCGAT 6683
                      v ----   - i - - - - - i -vi - i-v-- i
REFERENCE: 136244891 AGG---G-CGCTA-TAG-A-CC-AAT-TCA-GTACTC-AGGTGTCTTTT-CTT--GCAAT 136244936

    QUERY:      6682 TGA--CTCGC--GAGGCGCCGCGCTTGGCACC--CGAGAG-AGTGCGTGATCGAG-AC-A 6632
                      v -- v -- vvvviivv -- -- vvv v-- - v- -
REFERENCE: 136244937 TTATTCTCCCCTGAGGATAAATTATT--CACCATCTCCATTA-T-C---ATC-ATTACTA 136244988
:_
```

Human ONT reads for alignment comparison – 1D & 2D

| | |
|---|---|
| *Number of bases:* | *61635202* |
| *Number of reads:* | *10542* |
| *N50:* | *7884* |
| *Average length:* | *5846* |
| *Maximum length:* | *31618* |

# Comparisons of Three Aligners

| | Match | Mismatch | Insertion | Deletion | Error_on_ref | Error_on_seq |
|---|---|---|---|---|---|---|
| **LAST** | 52579177 | 5366213 | 2513700 | 4911625 | 0.2035 | 0.2116 |
| **BWA-mem** | 53009646 | 5443432 | 2575758 | 4961183 | 0.2047 | 0.2127 |
| **SMALT*** | 53176721 | 2962671 | 3723882 | 6108984 | 0.2056 | 0.2137 |
| **SMALT+** | 54436869 | 3458839 | 3706162 | 6346271 | 0.2103 | 0.2193 |

| | Total_aligned_bp | Reads_unaligned | Ave_aligned_len | CPU time (s) |
|---|---|---|---|---|
| **LAST** | 60459090 | 0 | 5735 | 28,002 |
| **BWA-mem** | 60861909 | 23 | 5786 | 8,398 |
| **SMALT*** | 59863274 | 1 | 5679 | 32,400 |
| **SMALT+** | 61601870 | 0 | 5843 | 31,000 |

(1)  For LAST, take single best alignment only;
(2)  SMALT* -S 'match=1,subst=-2,gapopen=-2,gapext=-1'
(3)   SMALT+ -S 'match=2,subst=-1,gapopen=-1,gapext=-1'

Single Molecular Integrated Scaffolding (SMIS)

Sequencing Reads

Fake Mate Pairs

Targeted Assembly

PE

Alignment - Smalt

Processing Pairs

Scaffolding - Spinner

Gap Closure

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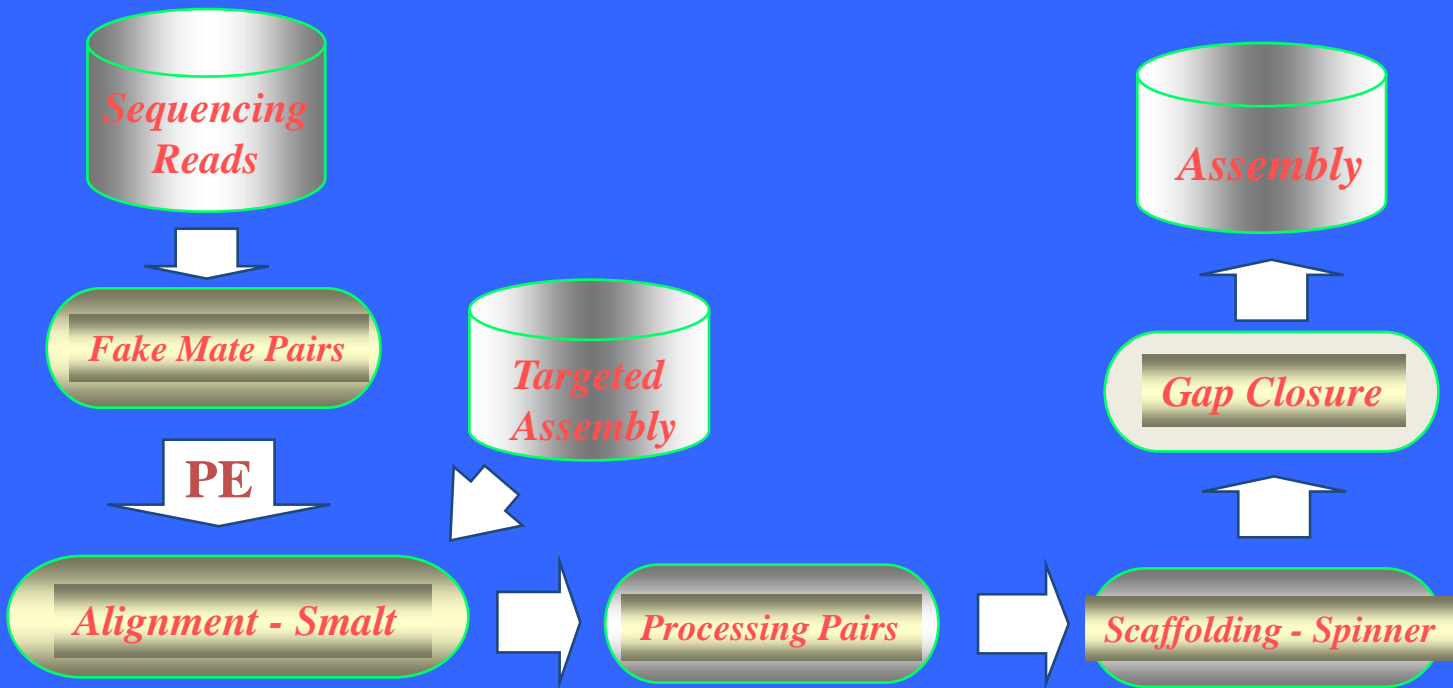
SMIS: http://sourceforge.net/projects/phusion2/files/smis/

Fake Mate Pairs from ONT Reads
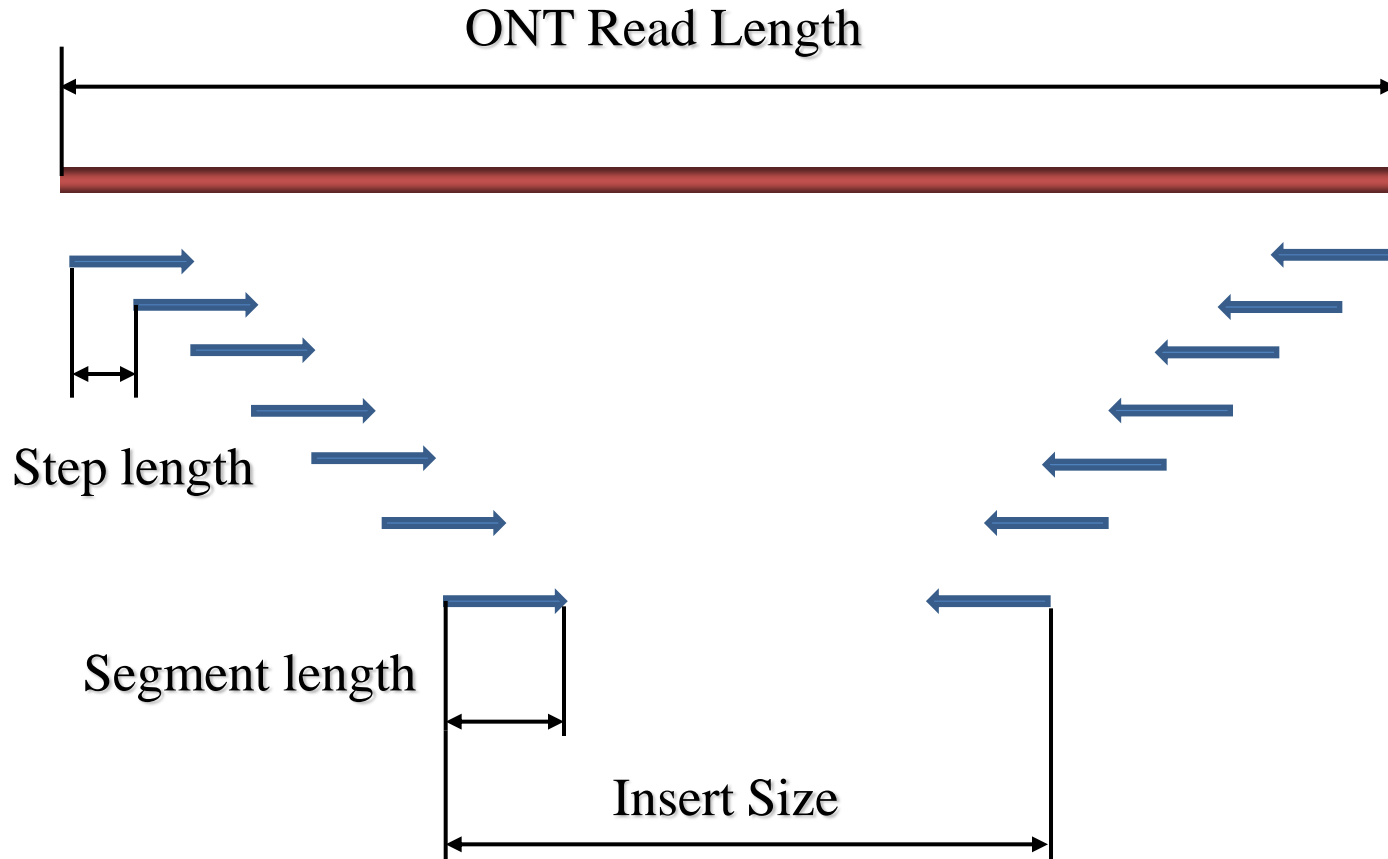
# Spinner – a scaffolding tool

*http://sourceforge.net/projects/phusion2/files/spinner/*

**Spinner uses mate pair data to scaffold contigs. Contigs, and pairs of contigs connected by pairs, define a bi-directional graph:**
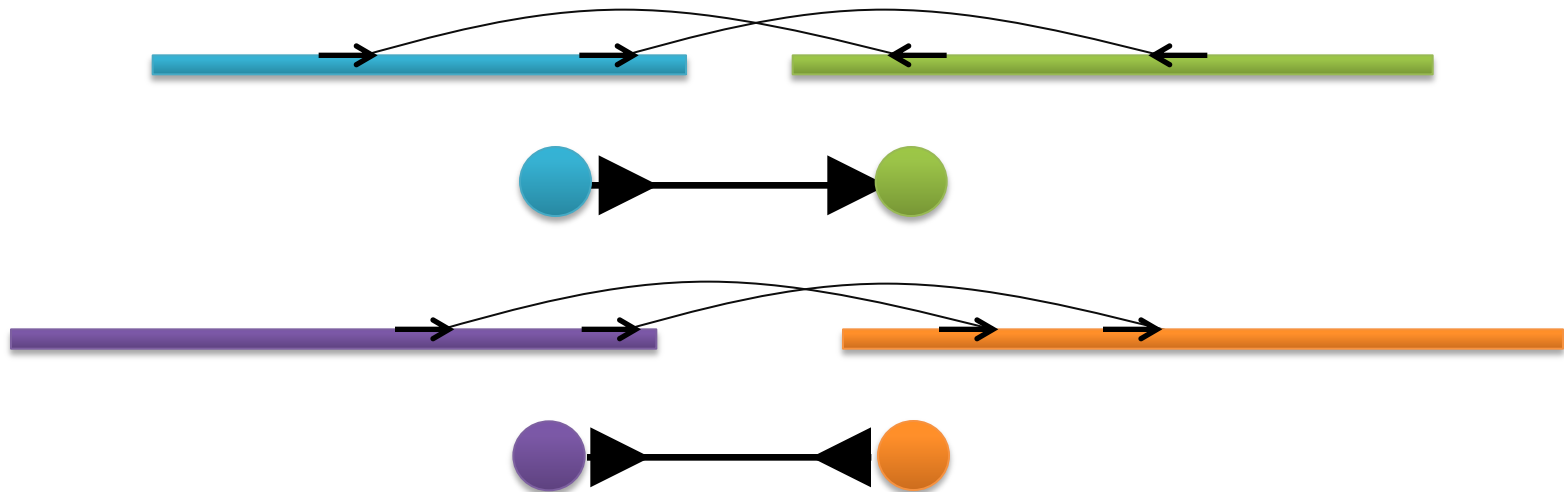


**Using expected insert size, a estimate of the gap size can be given for each contig.**

# Spinner – walks through a loop

**These techniques alone produces useful results. Further stages will be used to resolve repeats pairs that "jump over" repeats, and graph flow concepts.**
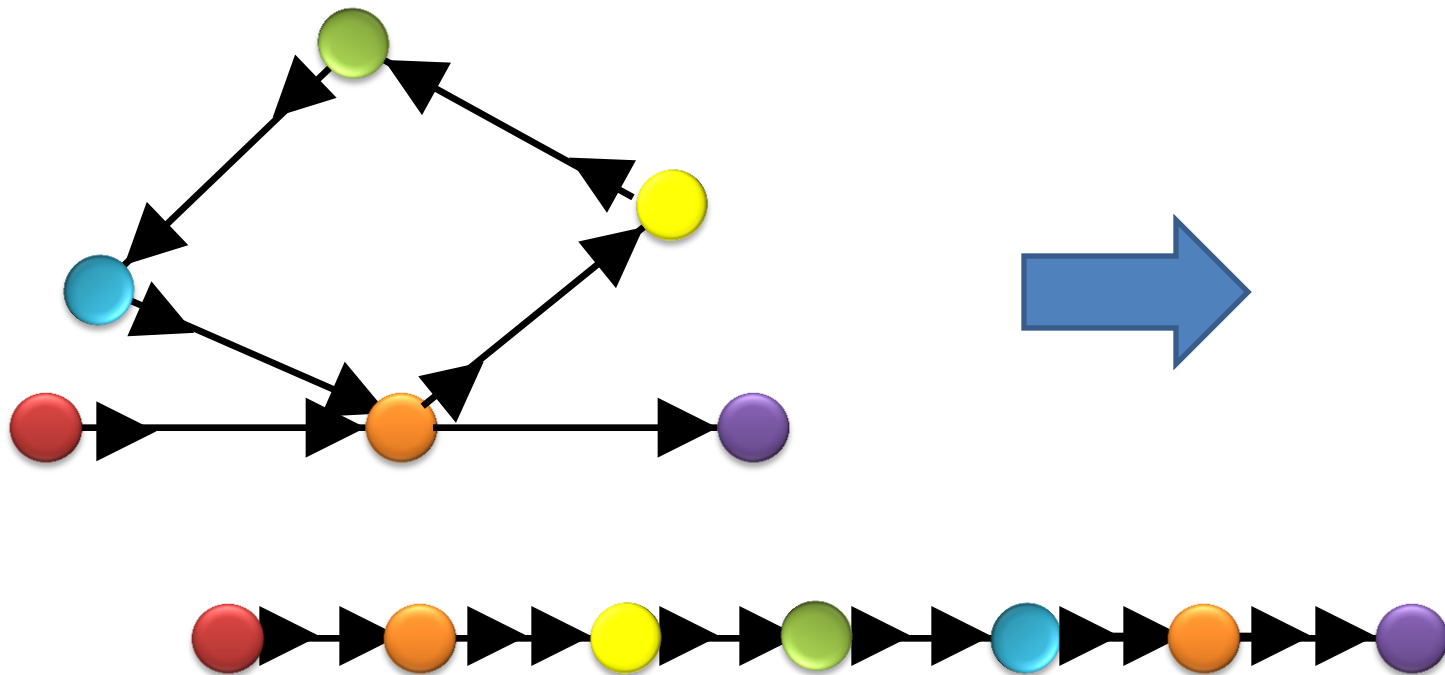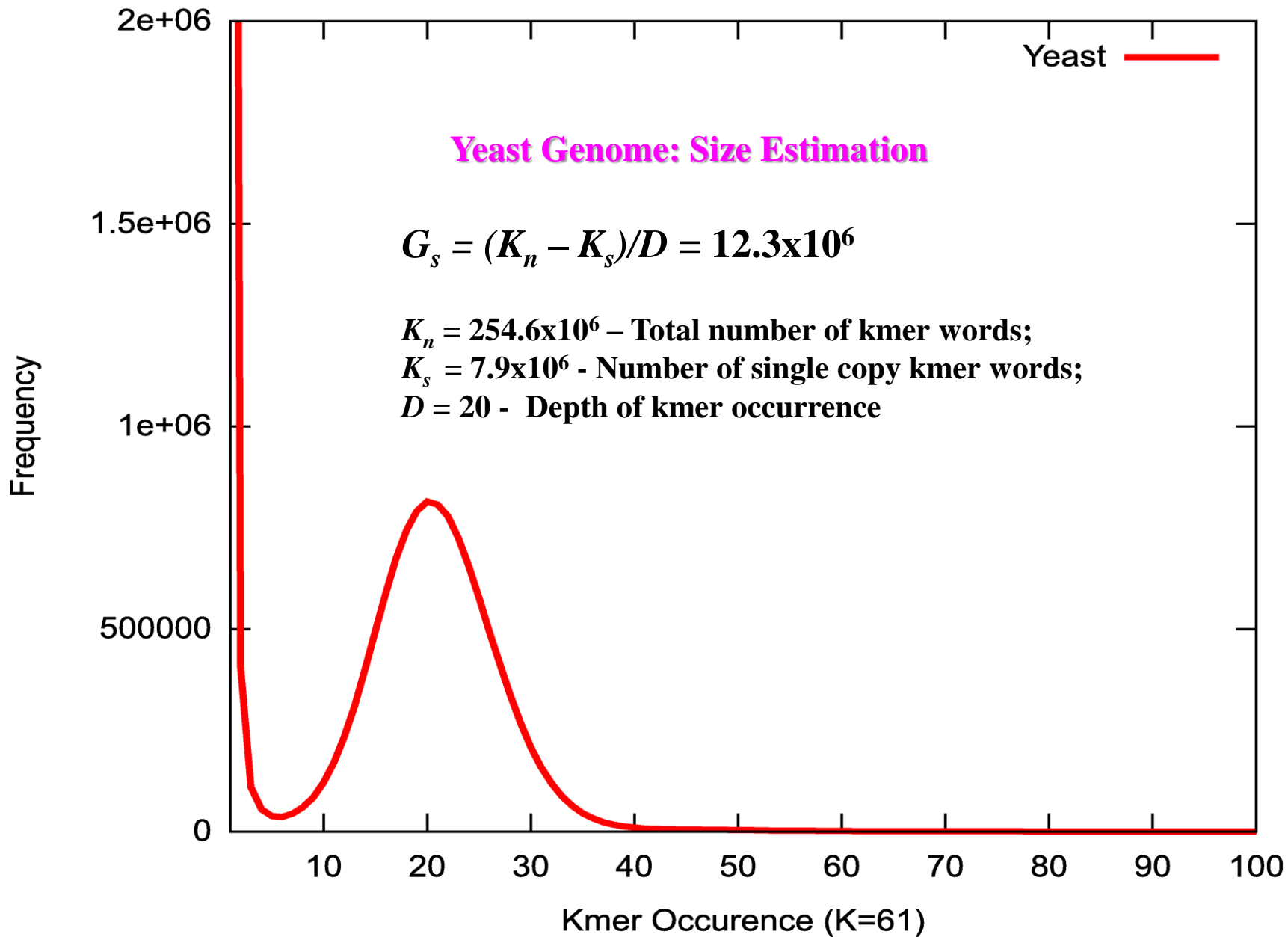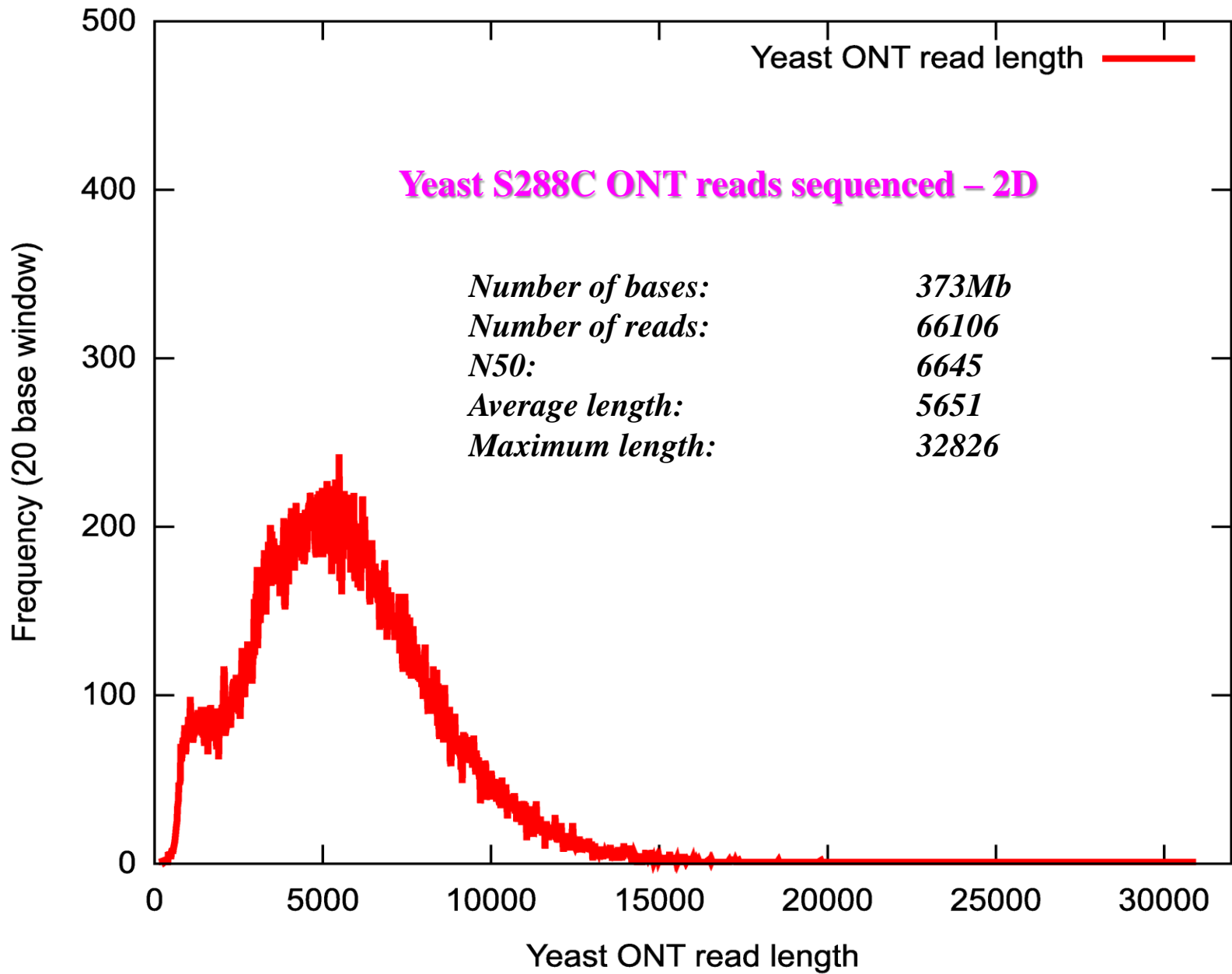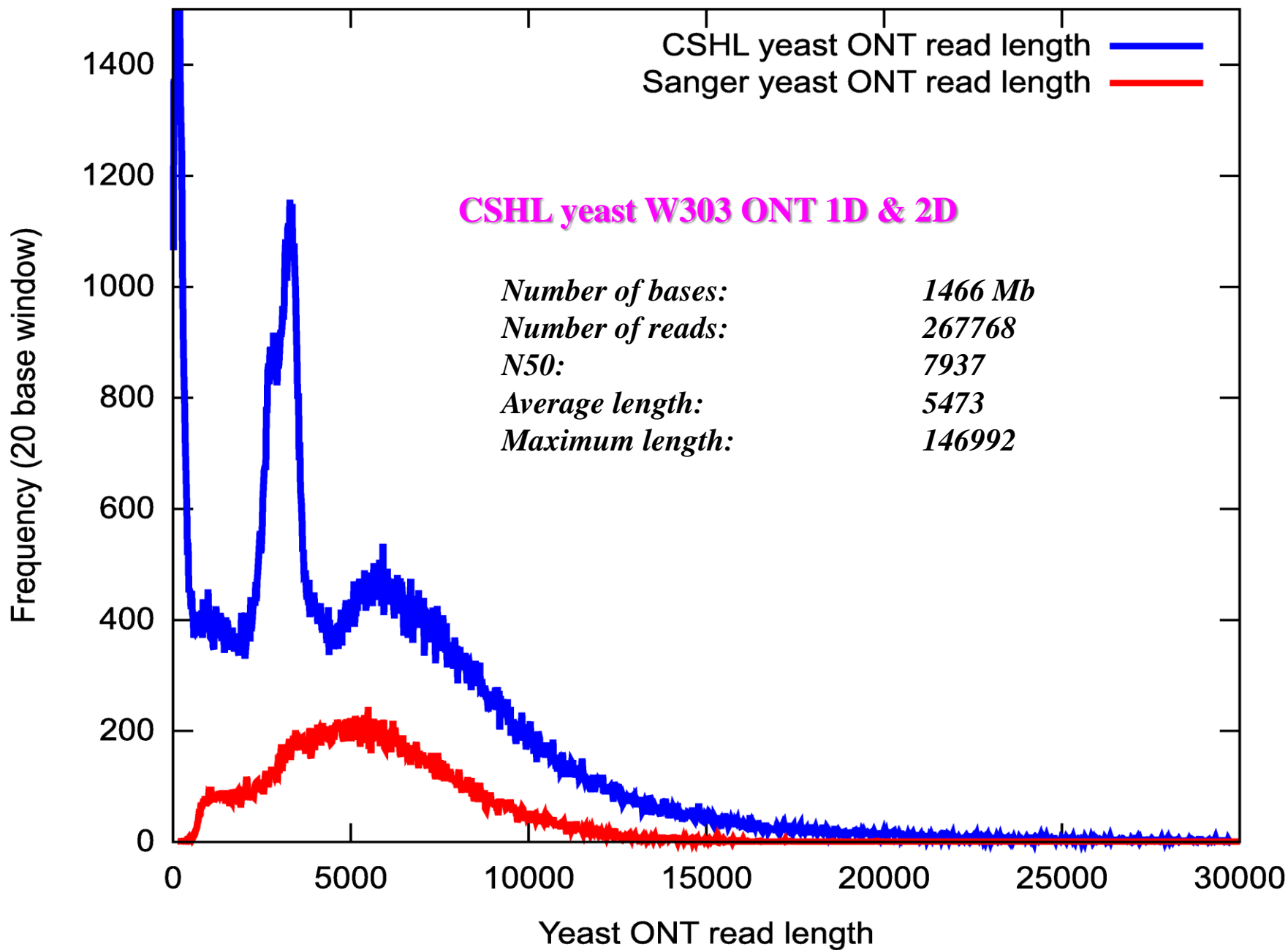
## Table 1 Illumina Reads Used for Assembly[+]

| Insert size | Library number | Total paired reads (m) | Read length (bp) | Sequence depth* (X) |
|---|---|---|---|---|
| 200bp | 1 | 3.6 | 2x76 | 46 |
| 200bp | 1 | 2 | 2x101 | 33.7 |
| Total | 2 | 5.6 | | 79.7 |

[+]The dataset was downloaded from this BioSample page: http://www.ncbi.nlm.nih.gov/biosample/839783

## Table 2 Yeast S288C Assembly Stats

| | *Fermi* | *SOAPdenovo** | *SMIS-Fermi* | *SMIS-SOAPdenovo+* |
|---|---|---|---|---|
| Total  bases of scaffolds (Mb) | 11.4 | 11.5 | 11.8 | 11.7 |
| Number of scaffolds | 1040 | 474 | 310 | 203 |
| Scaffold N50 (bb) | 20203 | 58059 | 342515 | 374531 |
| Scaffold N90 (bp) | 5929 | 16867 | 100124 | 103176 |
| Maximum scaffold length (bp) | 75044 | 234689 | 701224 | 703459 |
| Total  bases of contigs (Mb) | 11.4 | 11.5 | 11.4 | 11.5 |
| Number of contigs | 1040 | 476 | 1040 | 305 |
| Contig N50 (bp) | 20203 | 57497 | 20203 | 144573 |
| Contig N90 (bp) | 5929 | 16867 | 5929 | 31428 |
| Maximum contig length (bp) | 75044 | 234689 | 75044 | 455565 |

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

CSHL yeast W303 ONT 1D & 2D

| | |
|---|---|
| *Number of bases:* | *1466 Mb* |
| *Number of reads:* | *267768* |
| *N50:* | *7937* |
| *Average length:* | *5473* |
| *Maximum length:* | *146992* |
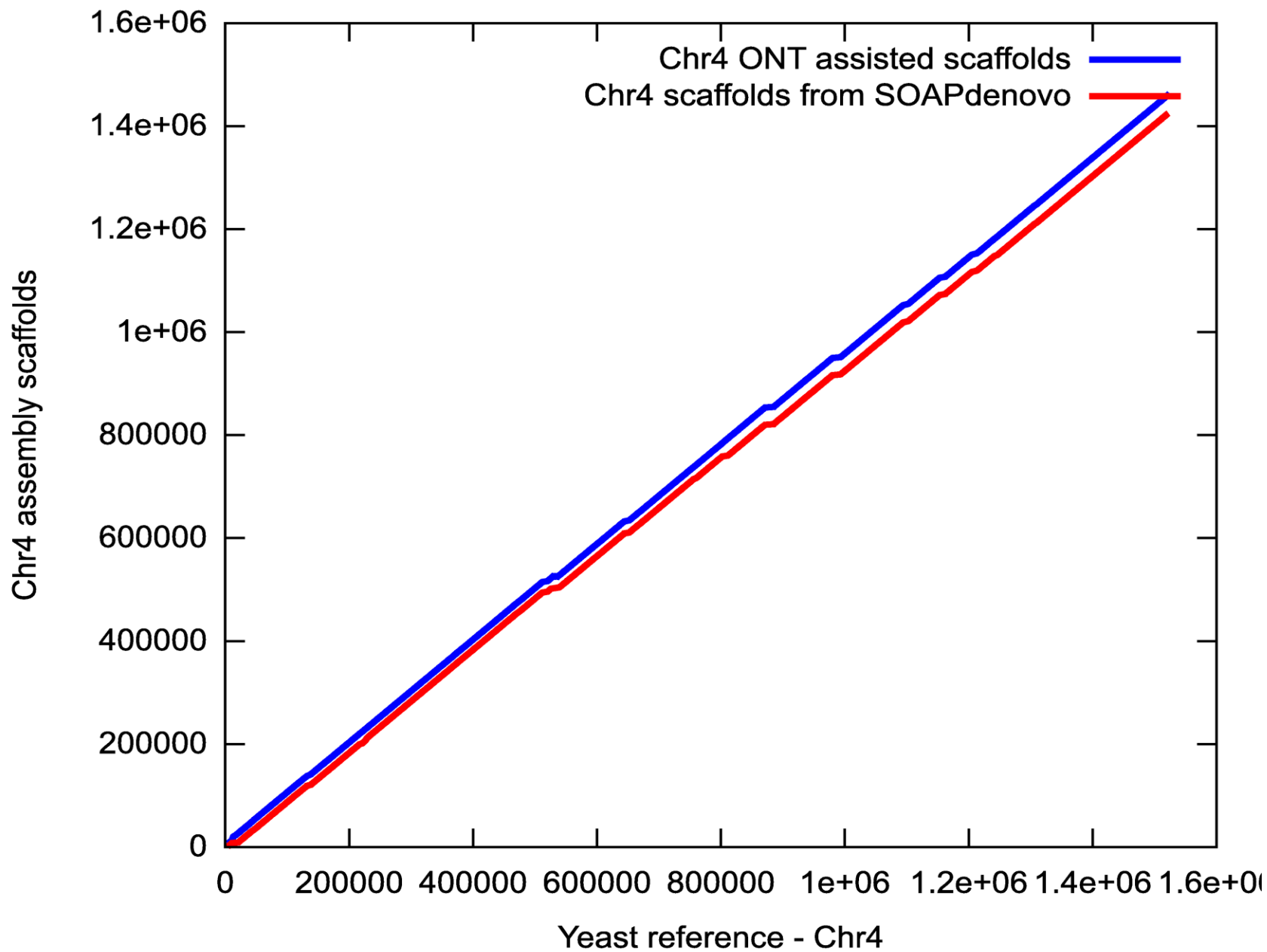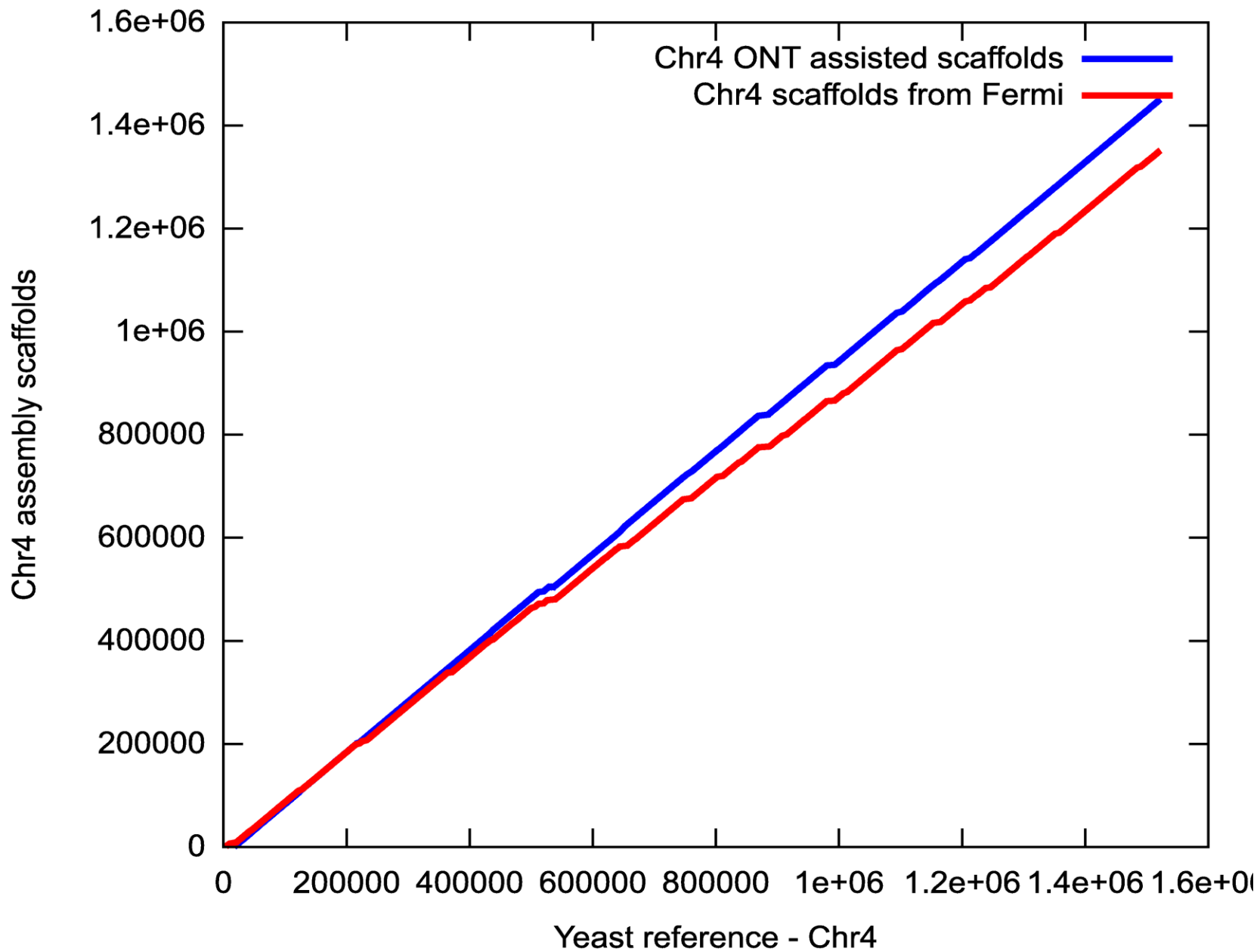
## Table 3 CSHL Illumina Reads Used for Assembly[+]

| 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.
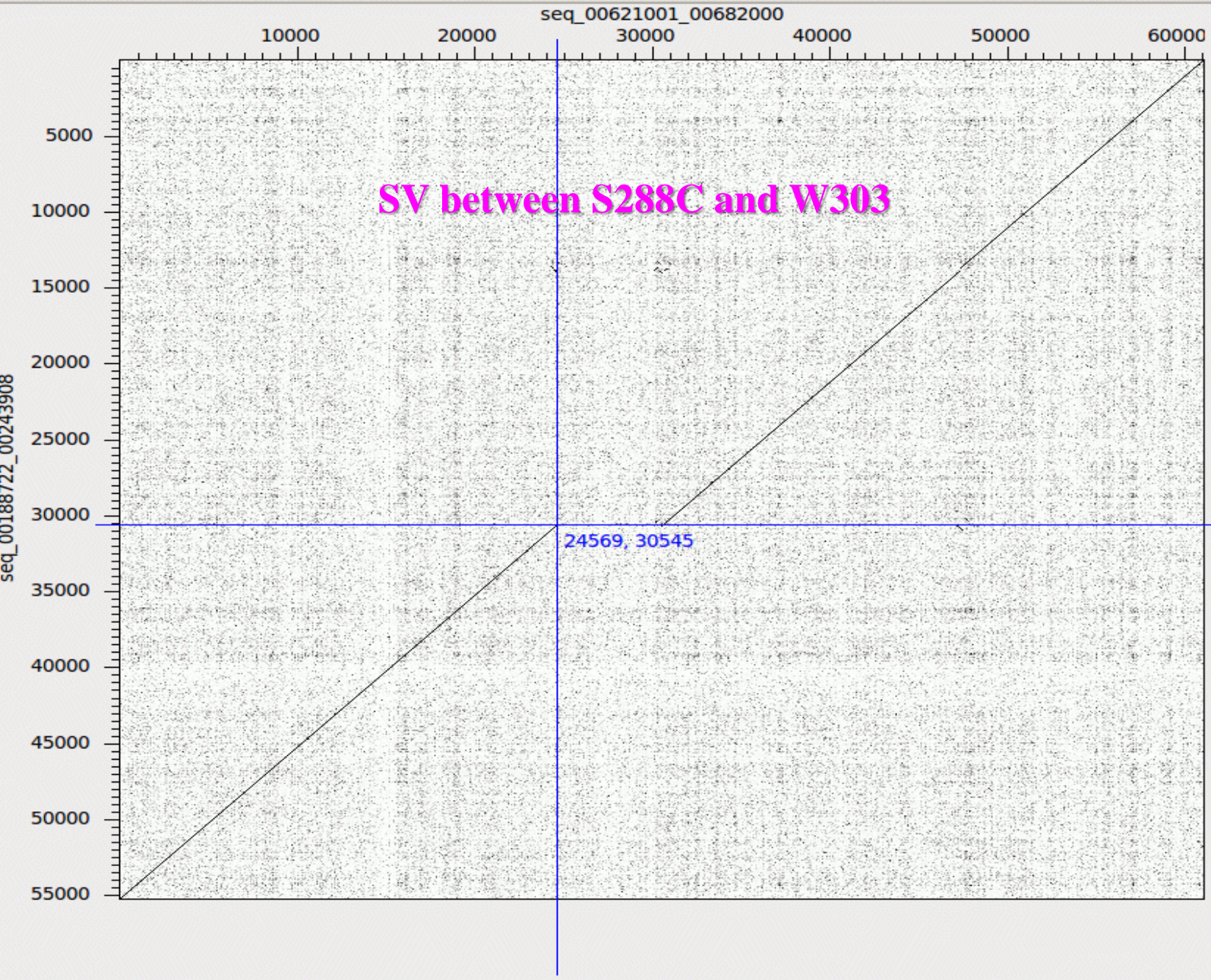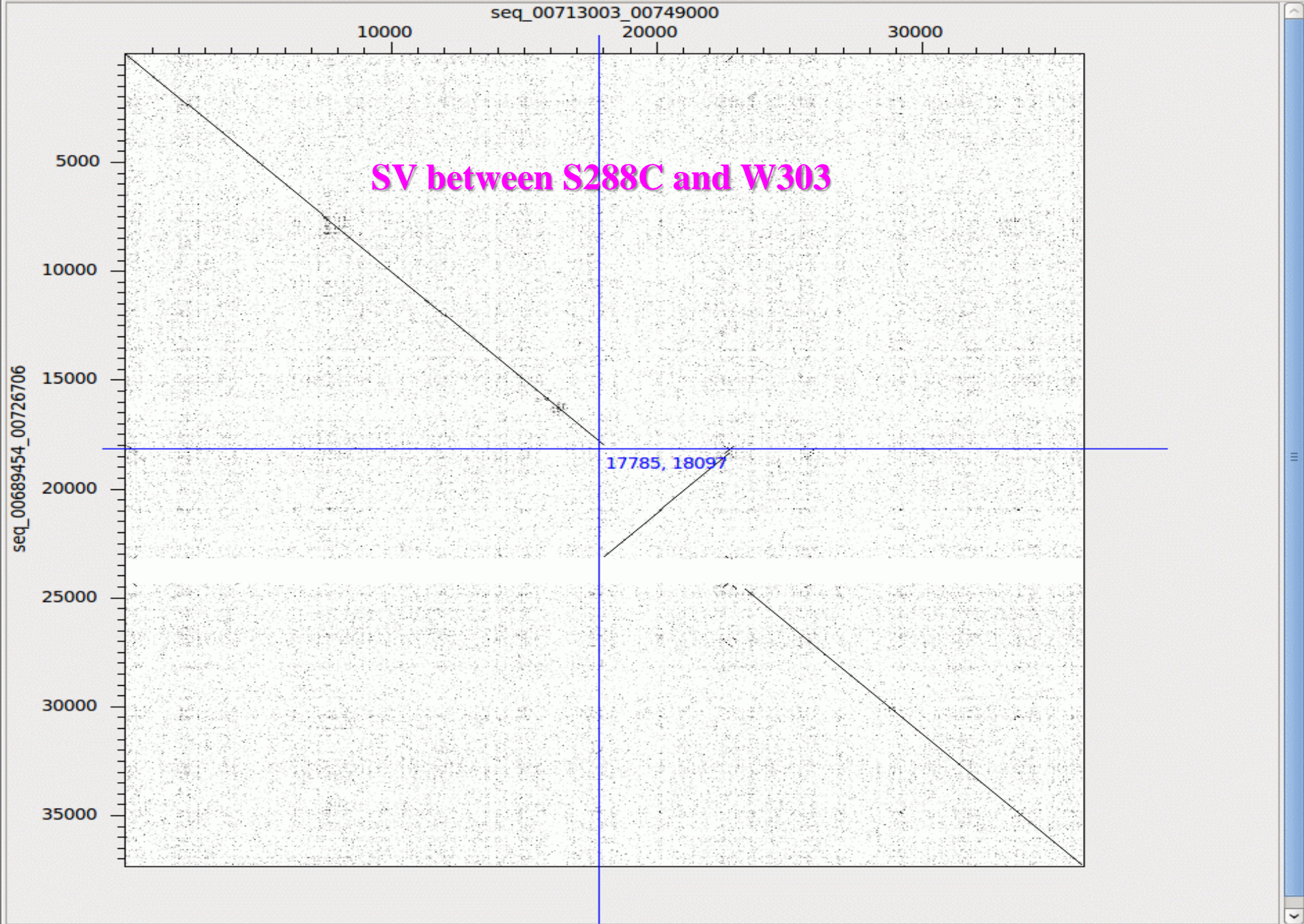
# *Acknowledgements:*

- ❑ **Richard Durbin**
- ❑ **Paul Coupland**
- ❑ **German Tischler**
- ❑ **Hannes Ponstingl**
- ❑ **James Bonfield**
- ❑ **Rob Davies**
- ❑ **Thomas Kean**
- ❑ **David Jackson**
- ❑ **Tony Cox**