Oxford Nanopore sequencing, hybrid error correction, and de novo assembly of a eukaryotic genome

Sara Goodwin, James Gurtowski, Scott Ethe-Sayers, Panchajanya Deshpande, Michael C. Schatz, and W. Richard McCombie

Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 11724, USA

Genome Res. 2015 Oct 7.

Nanopore Technologies

- Oxford Nanopore Technologies Limited is a U.K.-based company which is developing and selling nanopore sequencing products for the direct, electronic analysis of single molecules

- The company was founded in 2005 as a spin-out from the University of Oxford by Hagan Bayley, Gordon Sanghera, and Spike Willcocks, with seed funding from the IP Group

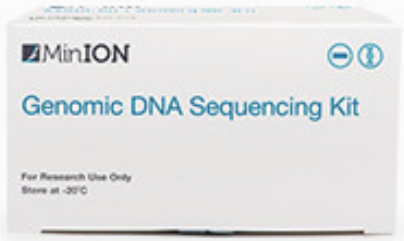
- £145 million in investment

- Products:
  - MinION: portable protein nanopore sequencing USB device is available through the MinION Access Programme (MAP)
  - PromethION: desktop high throughput device
  - GridION: a fully scalable nanopore analysis system

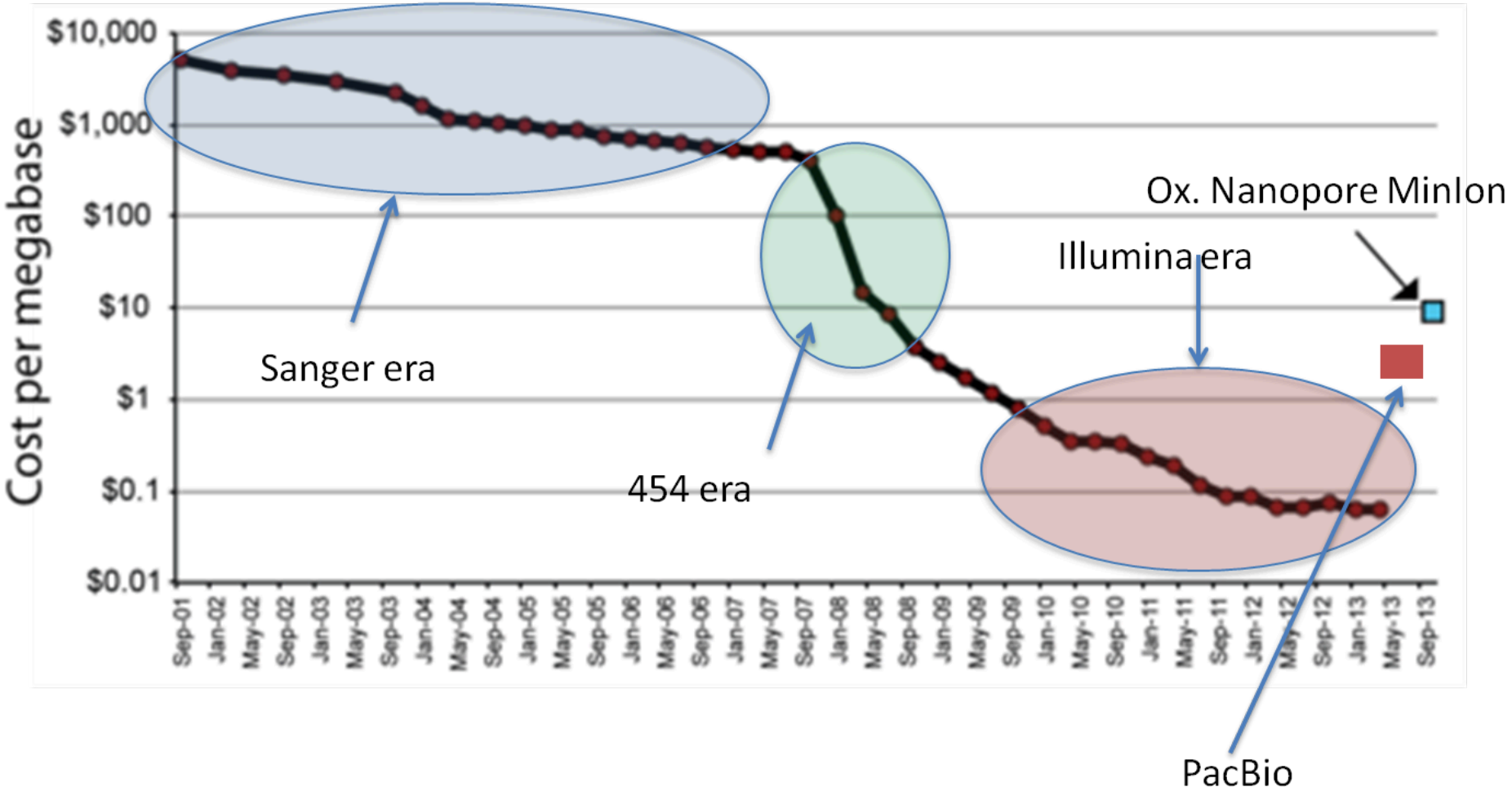
Nanopore MinION

- Upon registering to receive the MinION and paying your $1,000 fee, you will be provided a starter pack that includes a MinION MkI, sequencing kit and flow cells, relevant software and community based support. You will also receive a periodic supply of flow cells

Package Price	Number of flow cells	Price per flow cell
$900	1	$900.00
$9,480	12	$790.00
$16,200	24	$675.00
$24,000	48	$500.00



Cost per megabase



Why this paper?

- Developed a novel open-source hybrid error correction algorithm **Nanocorr** specifically for Oxford Nanopore reads
  - existing packages were incapable of assembling the long read lengths (5–50 kbp) at such high error rates (between ~5% and 40% error)

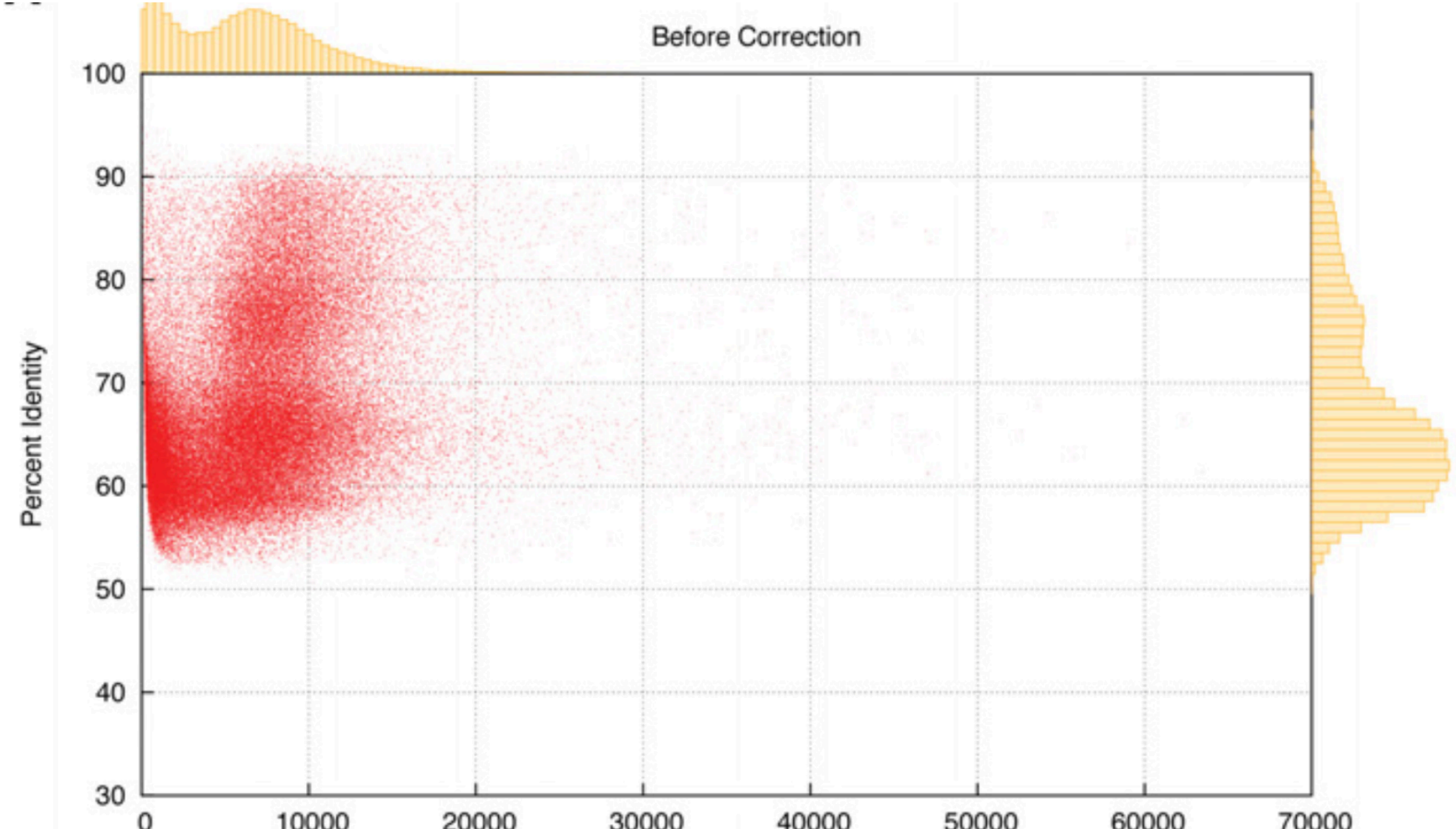
- Were able to perform a hybrid error correction of the nanopore reads using complementary **MiSeq** data and produce a de novo assembly that is highly contiguous and accurate
  - The contig N50 length is more than ten times greater than an Illumina-only assembly (678 kb versus 59.9 kbp) and has >99.88% consensus identity when compared to the reference

- The assembly with the long nanopore reads presents a **much more complete representation of the features of the genome** and correctly assembles gene cassettes, rRNAs, transposable elements, and other genomic features that were almost entirely absent in the Illumina-only assembly.

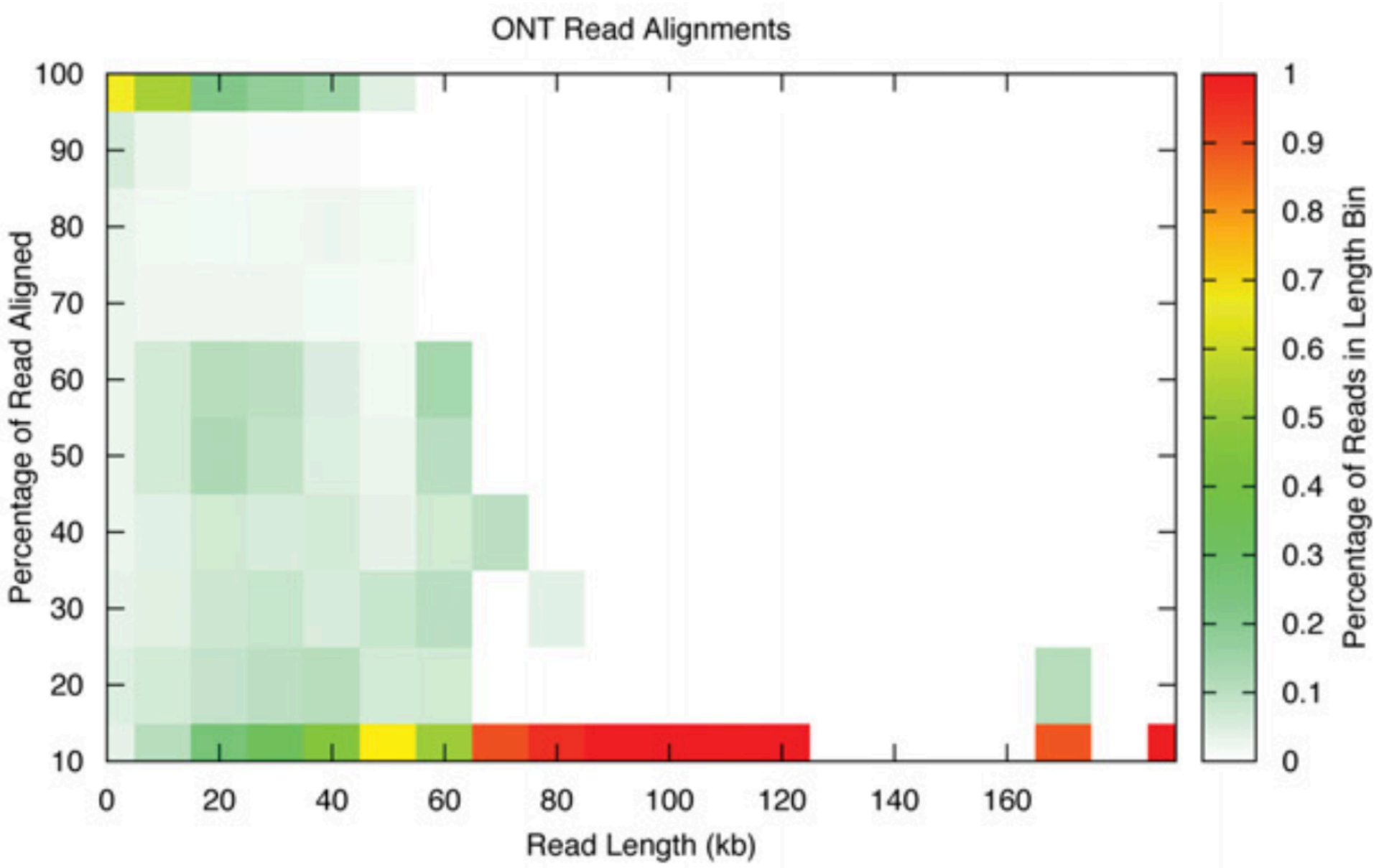
Nanopore sequencing of yeast

- *Saccharomyces cerevisiae* W303 strain
- Three device versions: R6.0 (11% of reads), R7.0 (49%), R7.3 (40%)
- 46 sequencing runs
- Generated >195x coverage of the genome
- In total 361,647 reads
- 44,028 “2D” reads (56%) and 105,771 “1D” reads (31%) aligned to the reference genome (BLAST)
- Avg read length 5548 bp, max 191,145 bp (“1D read”) and 57,453 bp (“2D read”)

Overall alignment identities



Overall read accuracy



Per-positsion error rates

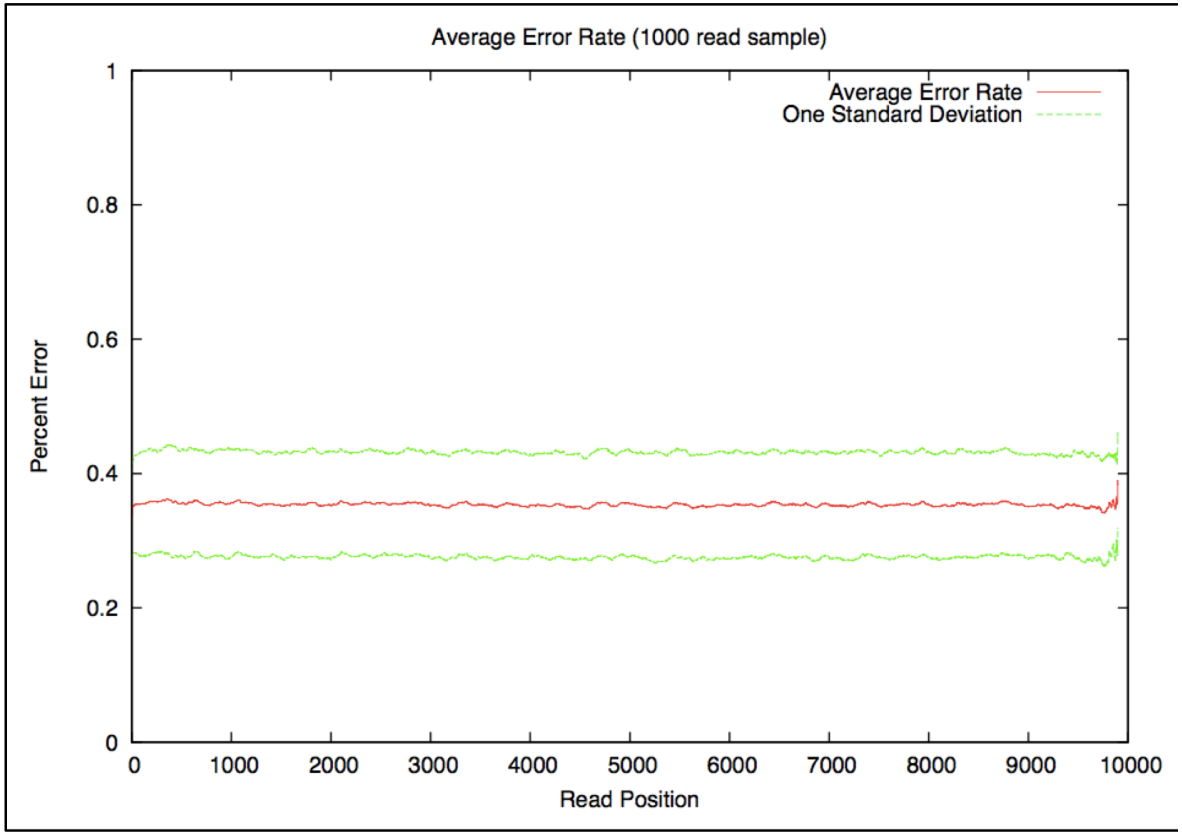


Figure S5B. Average error rate over the length of the read (red). Green lines indicate one standard deviation. 1000 reads with lengths between 9kb and 10kb were sampled and error rate was calculated for 100bp sliding windows using BLASTN alignments to the S288C reference genome

Distribution of coverage across GC%

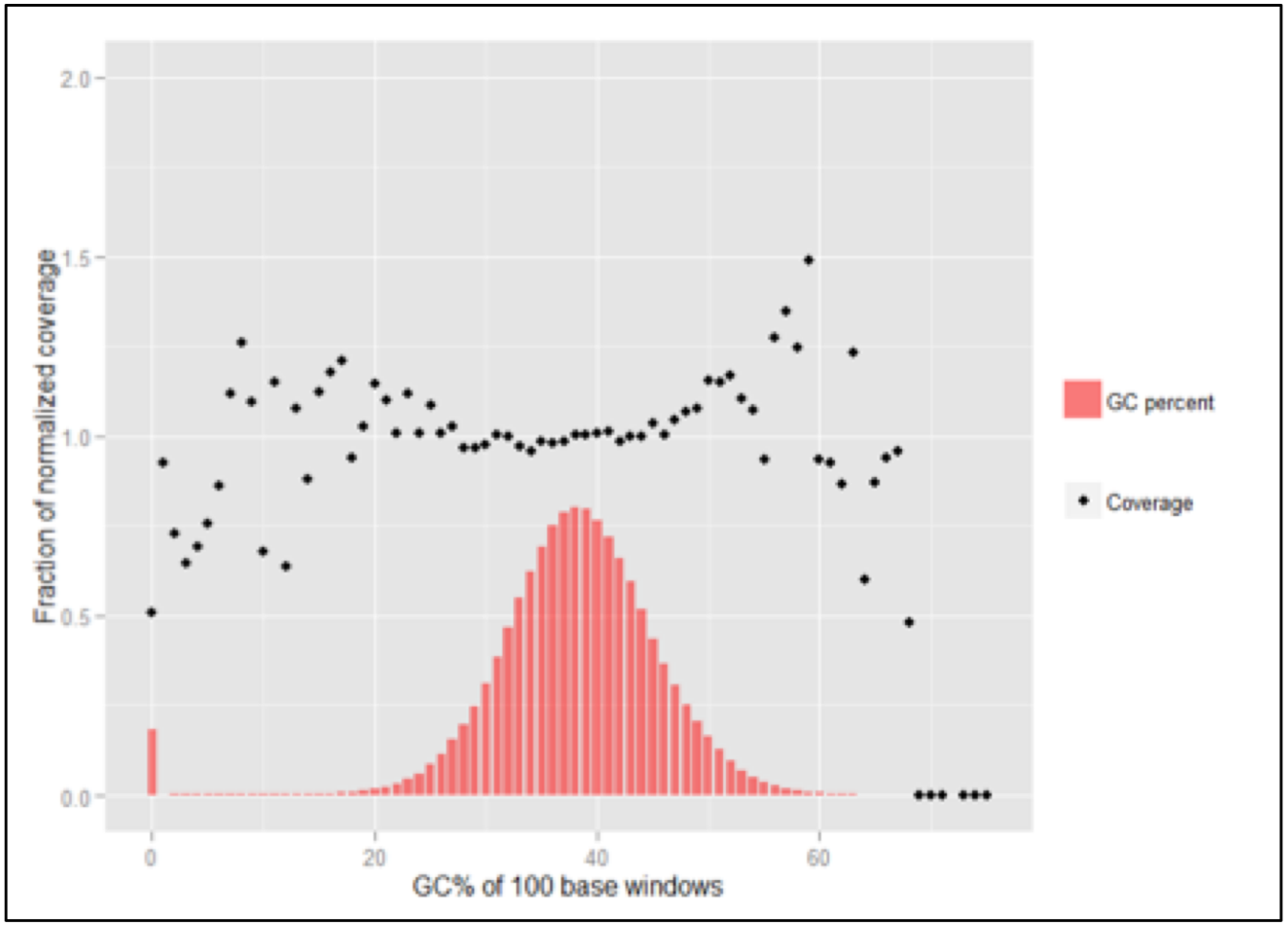
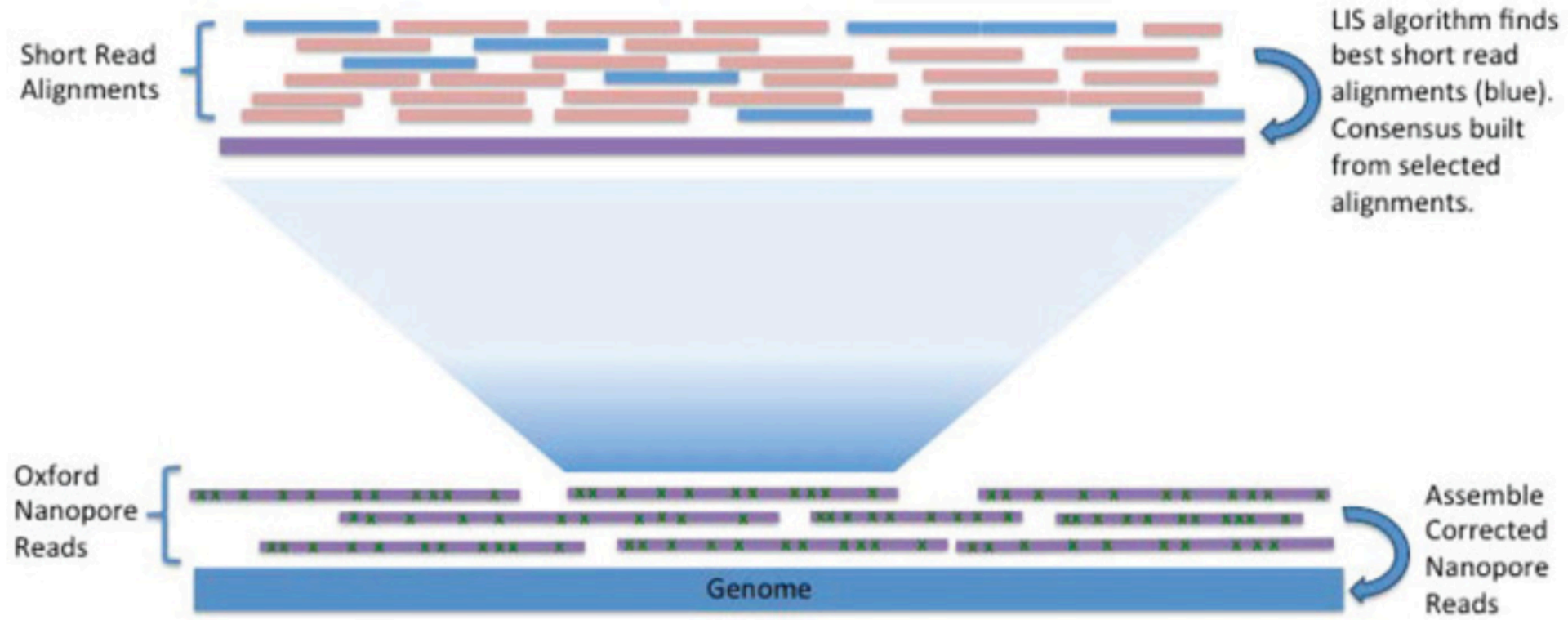


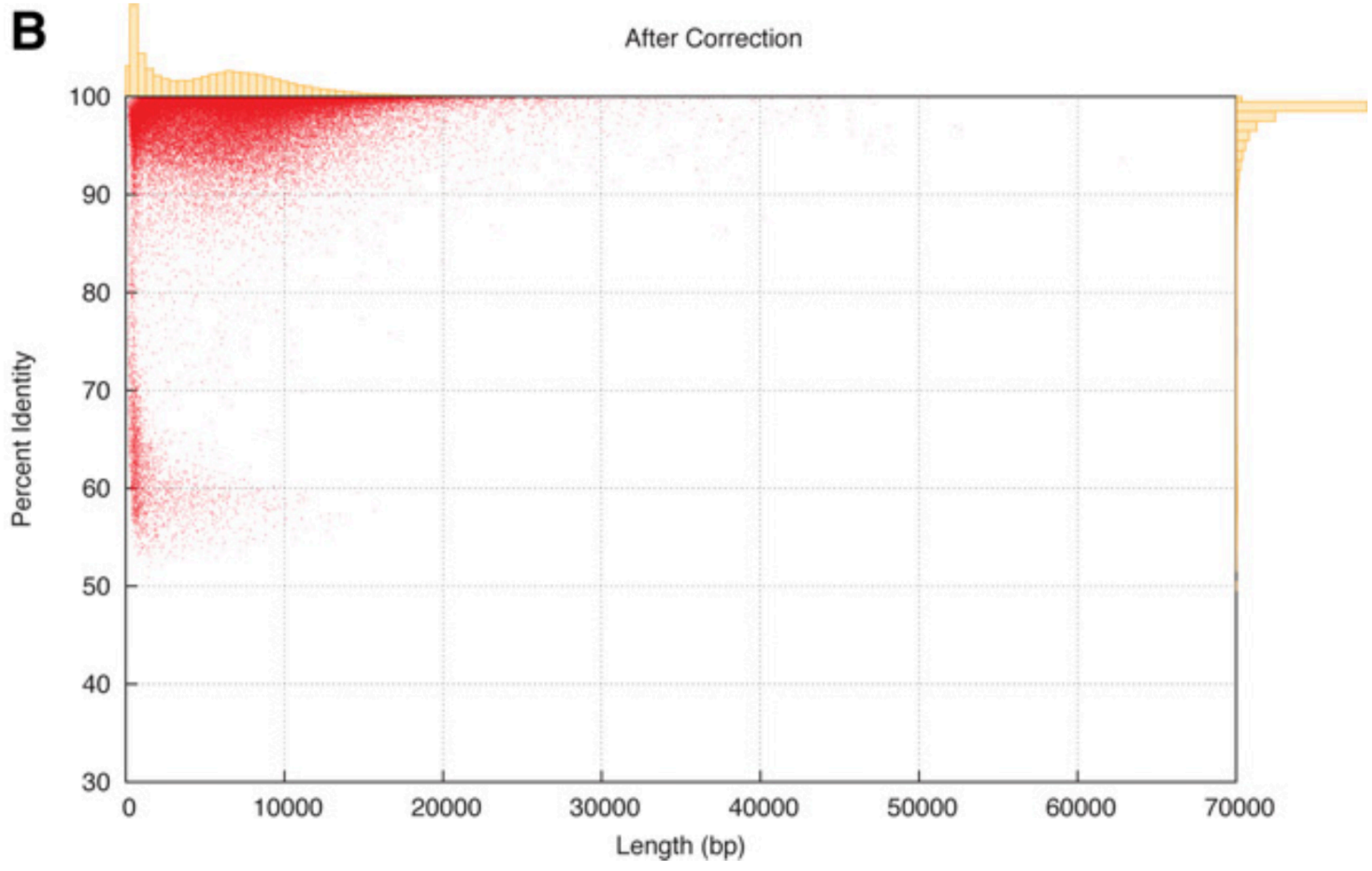
Figure S5D. Plot of coverage relative to the GC content of the W303 genome. Coverage (black dots) is mostly uniform at average GC content (30-50% GC) while high and low GC content shows a more variable coverage profile.

Hybrid error correction and de novo assembly

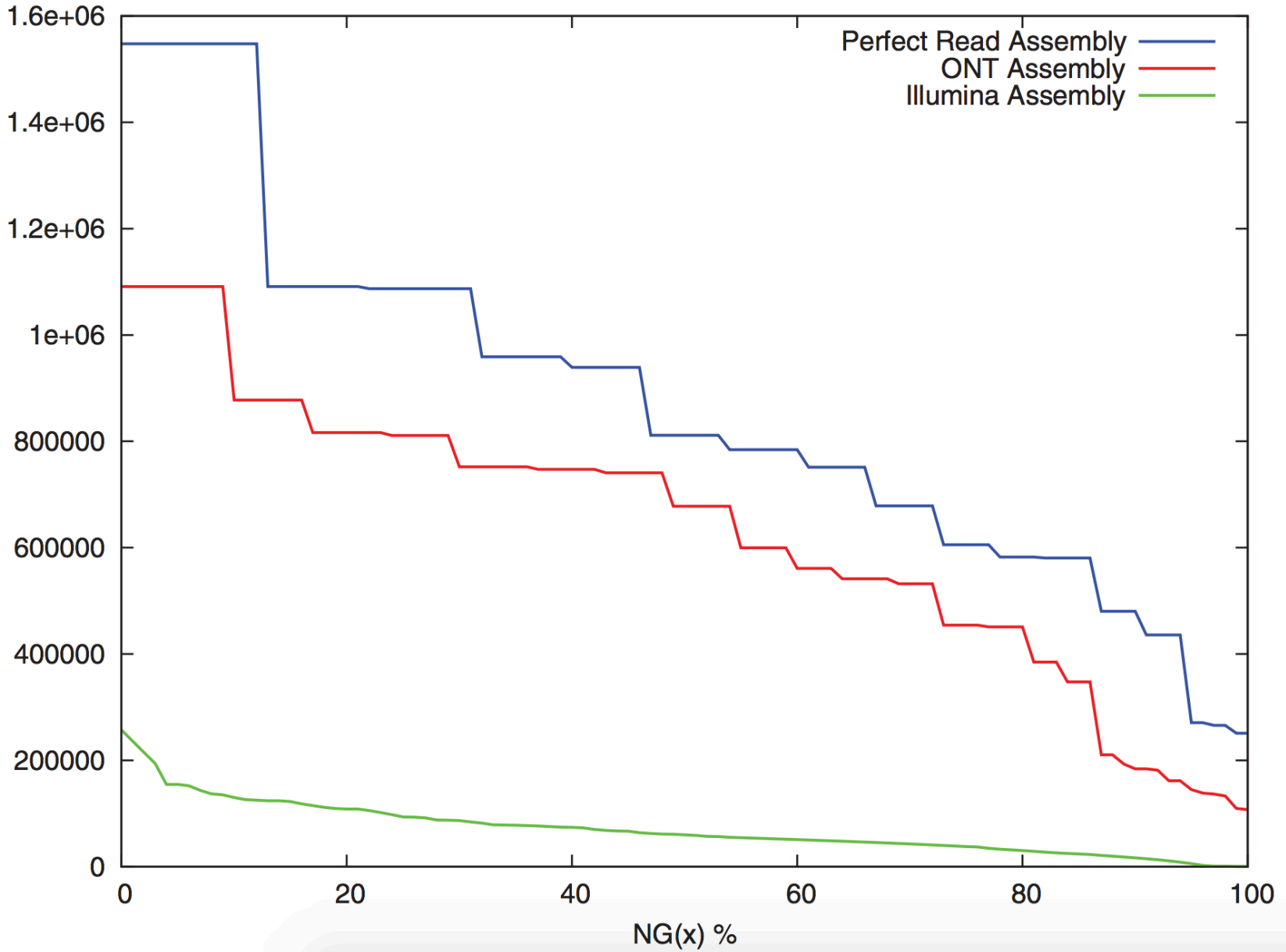
- Nanocorr algorithm:
  - Align 300bp PE MiSeq reats to long nanopore reads with BLASTN
  - Remove nanopore reads having no mapped MiSeq reads
  - LIS algorithm to find minimally overlapped reads
  - Build a consensus using *pbdagcon* tool (corrected nanopore reads)
  - Use Celera Assembler to assemble corrected reads into contigs
  - Use Pilon to revise nonredudant contigs by aligning MiSeq reads



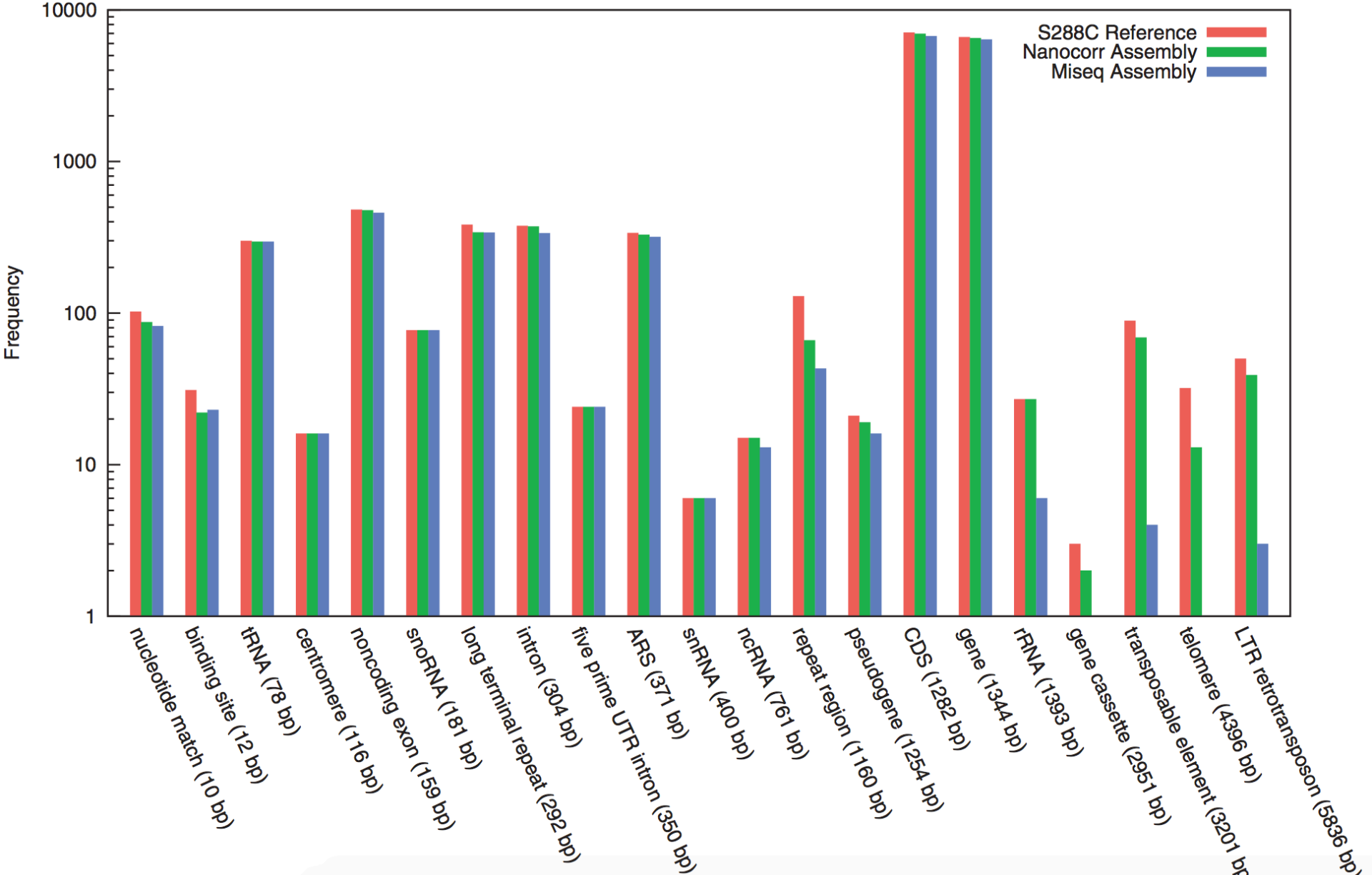
Overall alignment identities after correction



Assembly comparisons



Genomic features assembly



Dotplots of assemblies

20x coverage, reads >4kb length, N50 678kbp

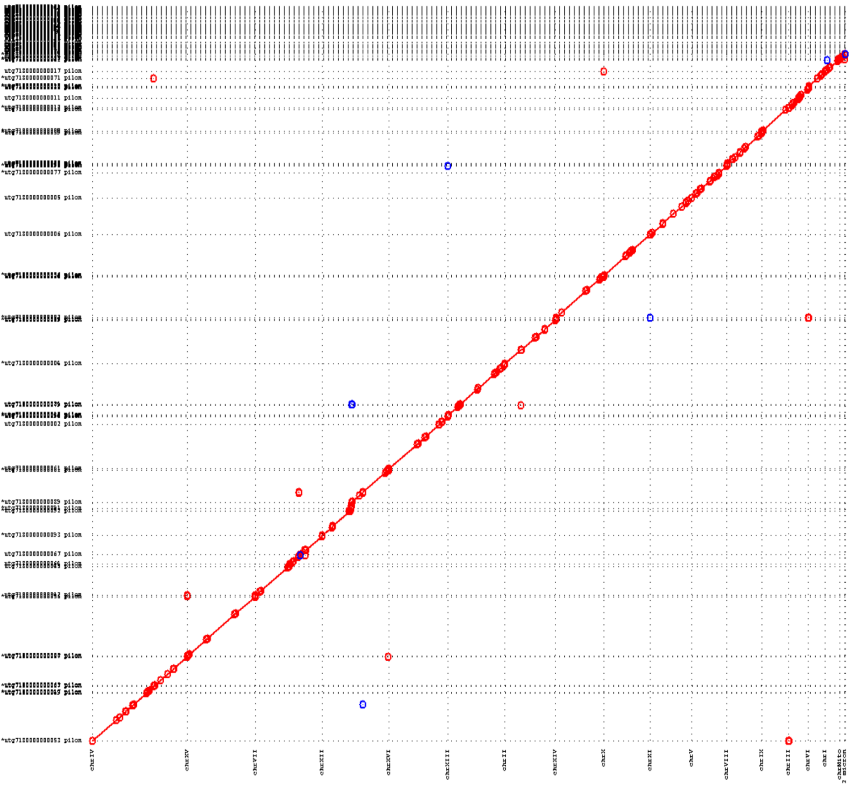


Figure S6B. Dot plot of Nanocorr-corrected Oxford Nanopore assembly (y-axis) of yeast versus the reference genome (x-axis).

28x coverage, reads >7kb length, N50 4.6Mbp

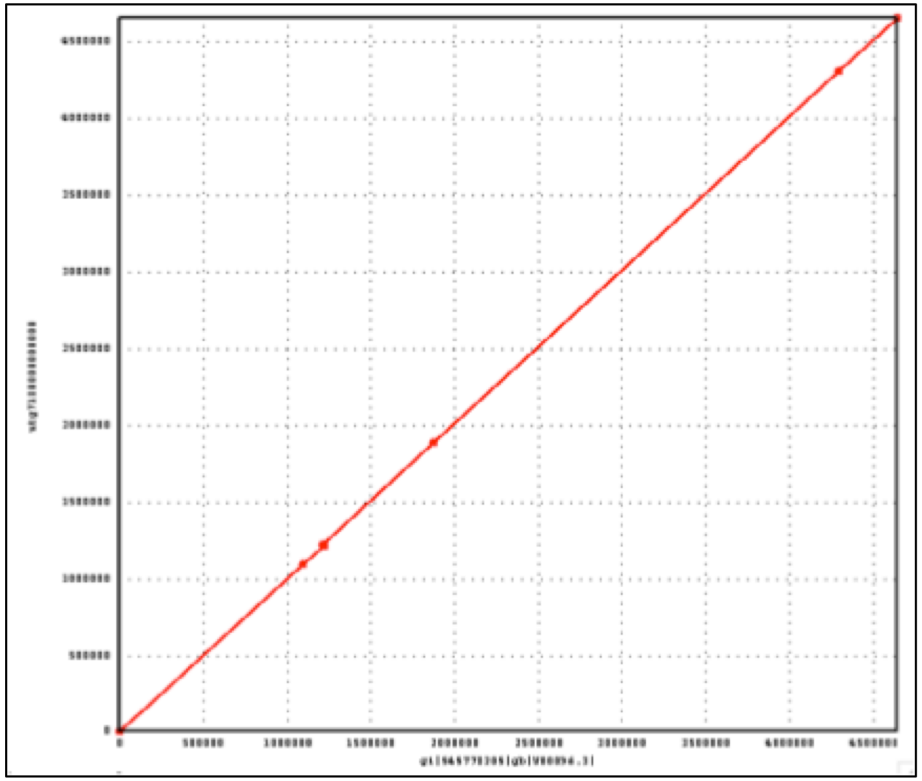


Figure S7B. Dot plot of Nanocorr-corrected Oxford Nanopore assembly (y-axis) of E. coli K12 versus the reference genome (x-axis). The nanocorr corrected assembly consisted of a single near perfect contig shown here as a single line along the diagonal, using dots to highlight the position of a few residual differences to the reference.

- Developed a novel open-source hybrid error correction algorithm **Nanocorr** specifically for Oxford Nanopore reads
  - existing packages were incapable of assembling the long read lengths (5–50 kbp) at such high error rates (between ~5% and 40% error)

- Were able to perform a hybrid error correction of the nanopore reads using complementary **MiSeq** data and produce a de novo assembly that is highly contiguous and accurate
  - The contig N50 length is more than ten times greater than an Illumina-only assembly (678 kb versus 59.9 kbp) and has >99.88% consensus identity when compared to the reference

- The assembly with the long nanopore reads presents a **much more complete representation of the features of the genome** and correctly assembles gene cassettes, rRNAs, transposable elements, and other genomic features that were almost entirely absent in the Illumina-only assembly.