Oxford Nanopore Sequencing and *de novo* Assembly of a Eukaryotic Genome
Supplemental Notes and Figures

Table of Contents

Supplemental Note 1. Flowcell performance

We chose to sequence the yeast genome so that we could carefully measure the accuracy and other data characteristics of the device on a tractable and well-understood genome. Our initial flow cells had somewhat low reliability and throughput, but improved substantially over time (Supplemental Figure 1). This is due to a combination of improvements in chemistry, protocols, instrument software, and shipping conditions. Some runs have produced upwards of 490 Mb of sequencing data per flow cell over a 48 hour period. Six of our highest yielding flow cells produced over 90% of the data generated for the assembly of yeast W303, and the top two flow cells produced more than 60% of the data. The MinION software can predict the number of functional pores before a run is started. In this study the number of predicted functional pores ranged from fewer than 100 to over 400. We found that if fewer than 350 functional pores were predicted, the resulting data amounts were severely limited.

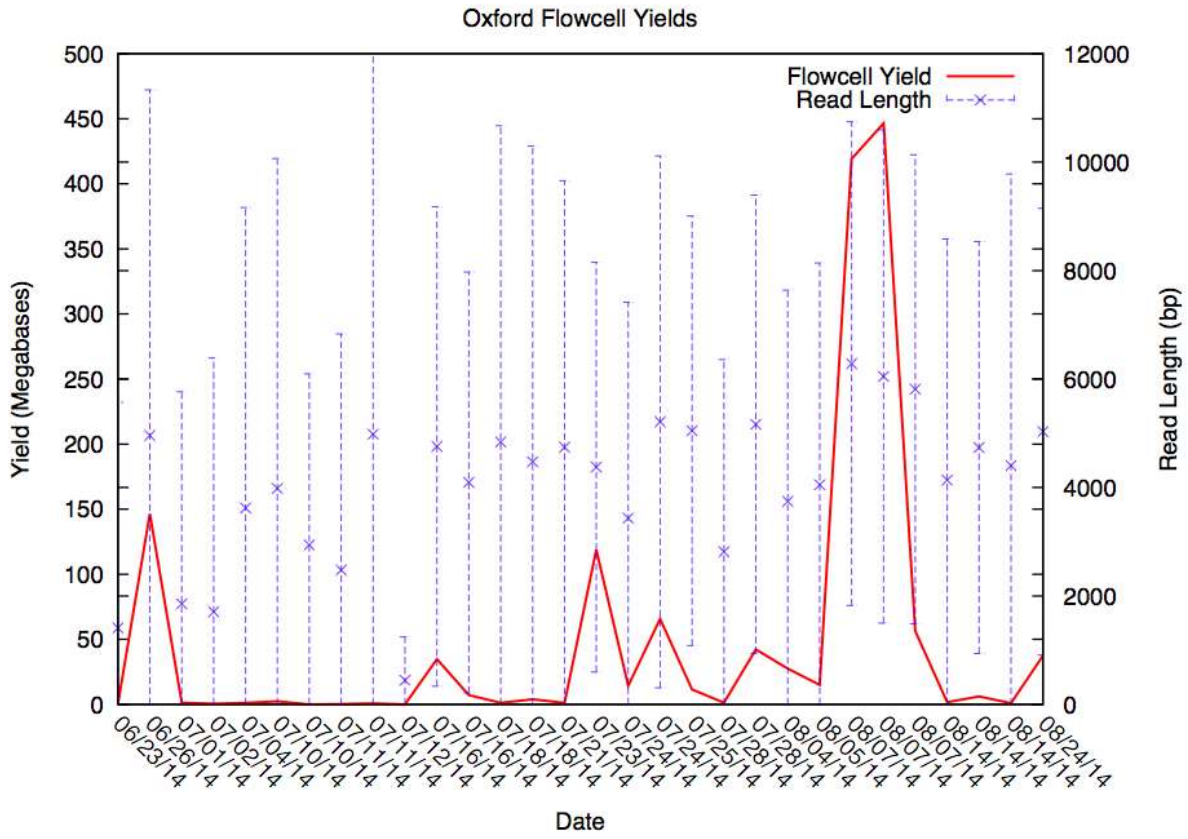


Figure 1. Flowcell yield and read lengths over time. Blue bars indicate the minimum, average and maximum length read from each flowcell. As the MAP program has progressed there has been a general trend to higher flowcell yield driven by improved library preparation and flowcell quality. Read lengths have remained constant over the course of program. Flowcells used from 6/23/14 to 7/16/14 are R6 chemistry, all remaining flowcells are R7.

Supplemental Note 2. Nanopore sequencing production over time

To evaluate the productivity of each flow cell we examined the cumulative base generation over time and found that sequencing throughput is nearly linear out to 48 hours. Although the instrument can be run indefinitely, the predefined protocol suggests ending the run at 48hrs after which sequence production begins to taper off.

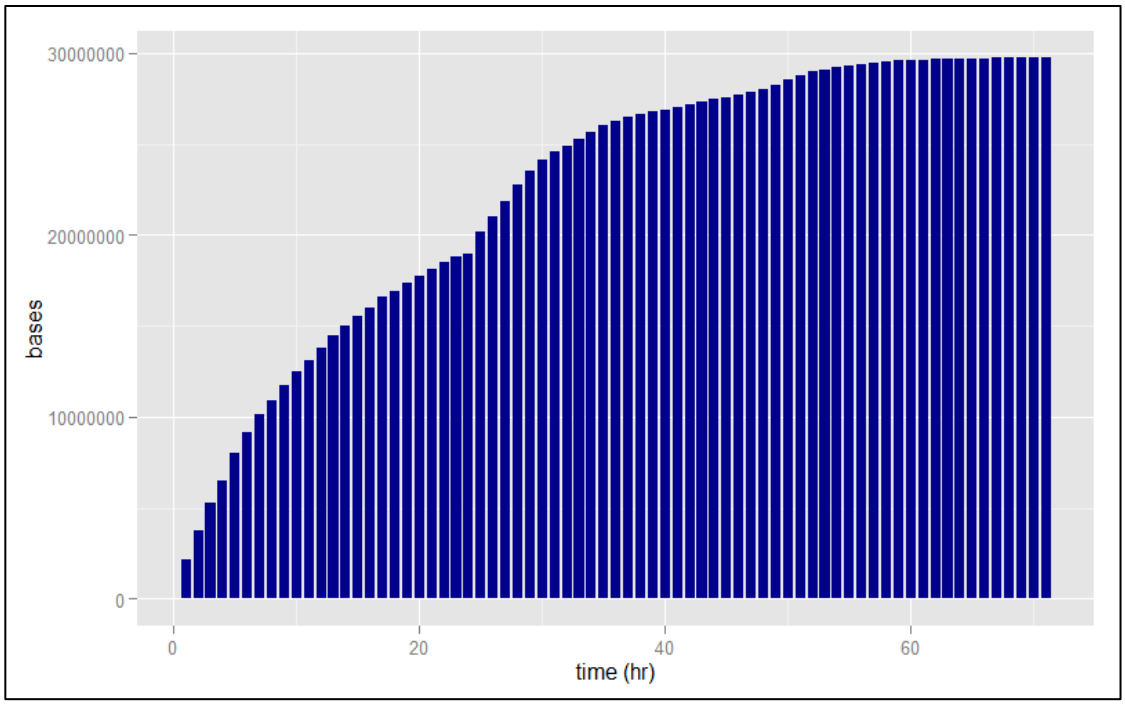


Figure S2. Yield over time bar chart. Displays the number of bases sequenced over each hour of a 72 hour run. The sharp increase in the flowcell yield observed at 24 hours is the point at which active pores are reselected and poor performing pores can be replaces with fresh pores.

Supplemental Note 3. Size selection and read lengths

The majority of the libraries used for this study were made with a shearing parameter of 10kb, without significant size selection. The stochastic loading of DNA fragments into individual pores suggests that the length distribution of the resulting reads should mirror the length distribution of the fragments in the sample. Figure S2A shows a Bioanalyzer trace for a library sheared at 10kb. The peak at 17kb is the upper limit marker for the trace. The majority of fragments are between 5 and more than 10kb. Figure S2B shows the fragment distribution of the same library post-sequencing. As is the case of the bioanalyzer trace, the majority of fragments are between 5 and more than 10kb indicating that many fragments are full length fragments. Different shearing parameters are also evaluated, 10kb 20kb and unsheared (Figure 2C). Like the previous figures, the mean fragment length was less than the shearing target and there are many fragments at 3.5kb, indicating an abundance of control DNA. As the input fragment size increased the sequencing fragment size increases as well with no apparent impact on performance.

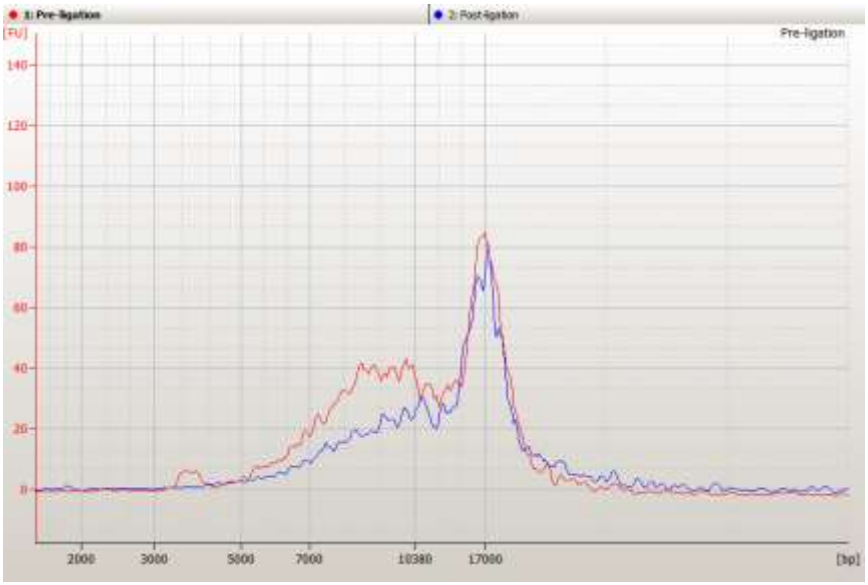


Figure S3A. Bioanalyzer traces prepared from an Oxford library sheared at 10kb before and after adapter ligation. The large peak at 17k is the upper marker.

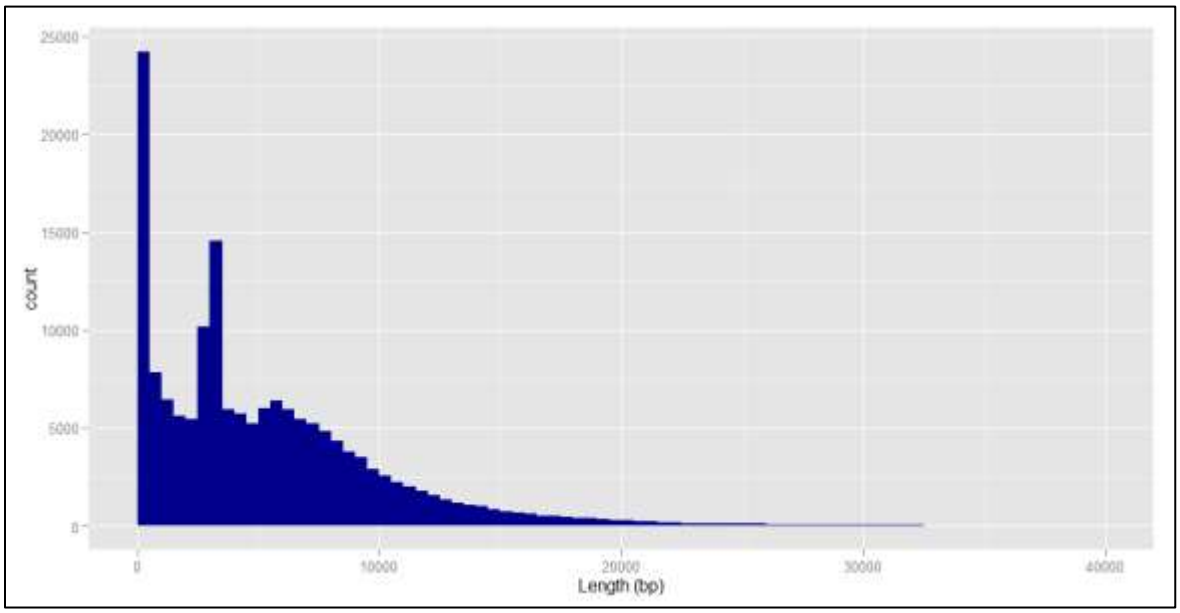


Figure S3B. Fragment lengths derived from sequencing. The large peak at 3500 is a control DNA spike in. For visualization purposes this figure has been truncated to 40000bp however that are a small number of reads that exceed that cutoff.

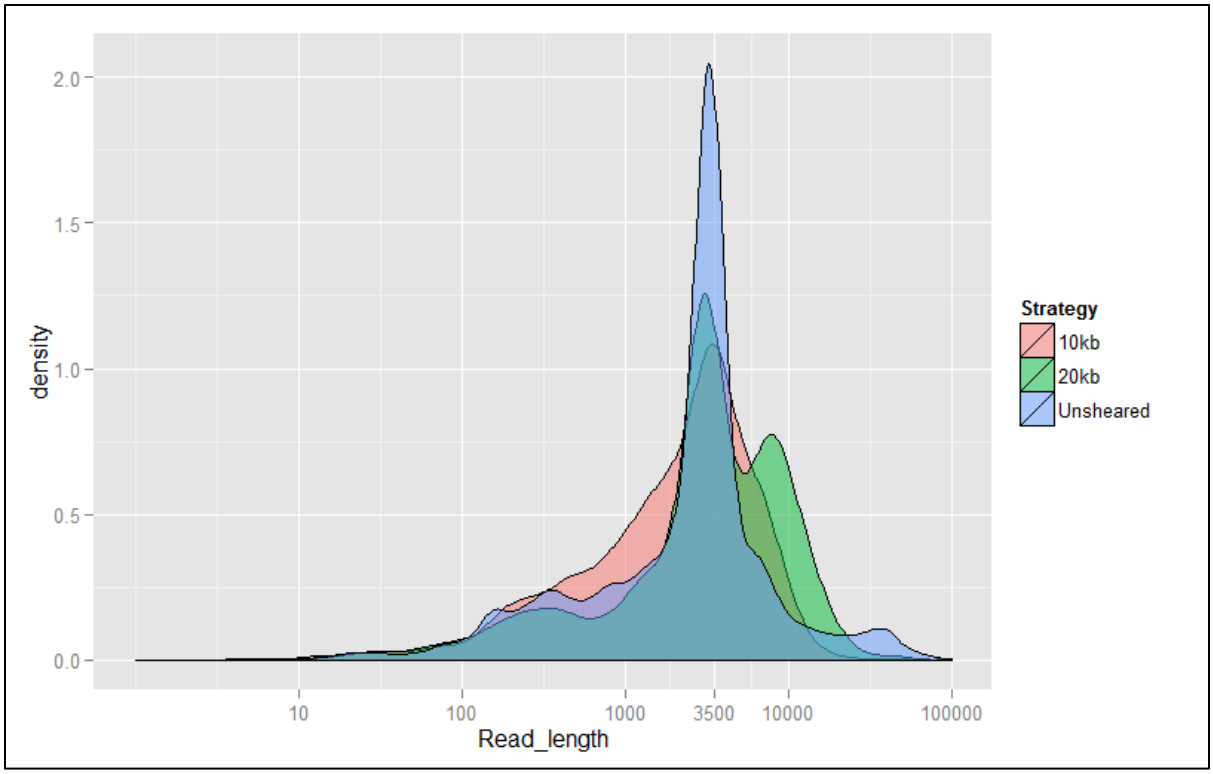


Figure S3C. Overlay of the fragment sizes at different shearing parameters. The large peak at 3500 is a control DNA spike in. A clear increase in in average fragment size is seen as shearing size is increased.

Supplemental Note 4. Base-caller 5-mer performance

To evaluate the general performance of the base caller, the expected abundance of each 5mer was calculated from the W303 reference genome and the totality of all reads generated by the flow cells. (Fifty-three out of 1024 5mers were found to deviate significantly. The majority of these 5mers had runs of at least 3 of the same nucleotide. Similar patterns of enrichment and depletion were seen for each of the different "types" of read: template (Figure S3A), complement(Figure S3B) and "2D" (the consensus of the template and complement reads) (Figure S3C). Figure S3D shows minor sequencing bias at high and low GC regions. Table S1 lists the 5mers that deviate from the expected abundance in the W303 genome. Fifty-three out of 1024 possible *5-mers* were found to deviate significantly from the expected abundance in the W303 genome. The majority of these *5-mers* had runs of at least 3 of the same nucleotide suggesting an enrichment for homopolymer errors. The most deficient *5-mer* was the nucleotide sequence TTTTT while the most enriched was ACCCG relative to its actual presence in the genome. There are many potential causes of this, including basecalling errors induced by homopolymers or sequencing bias for or against GC rich regions as has been previously reported for other sequencing technologies[11]

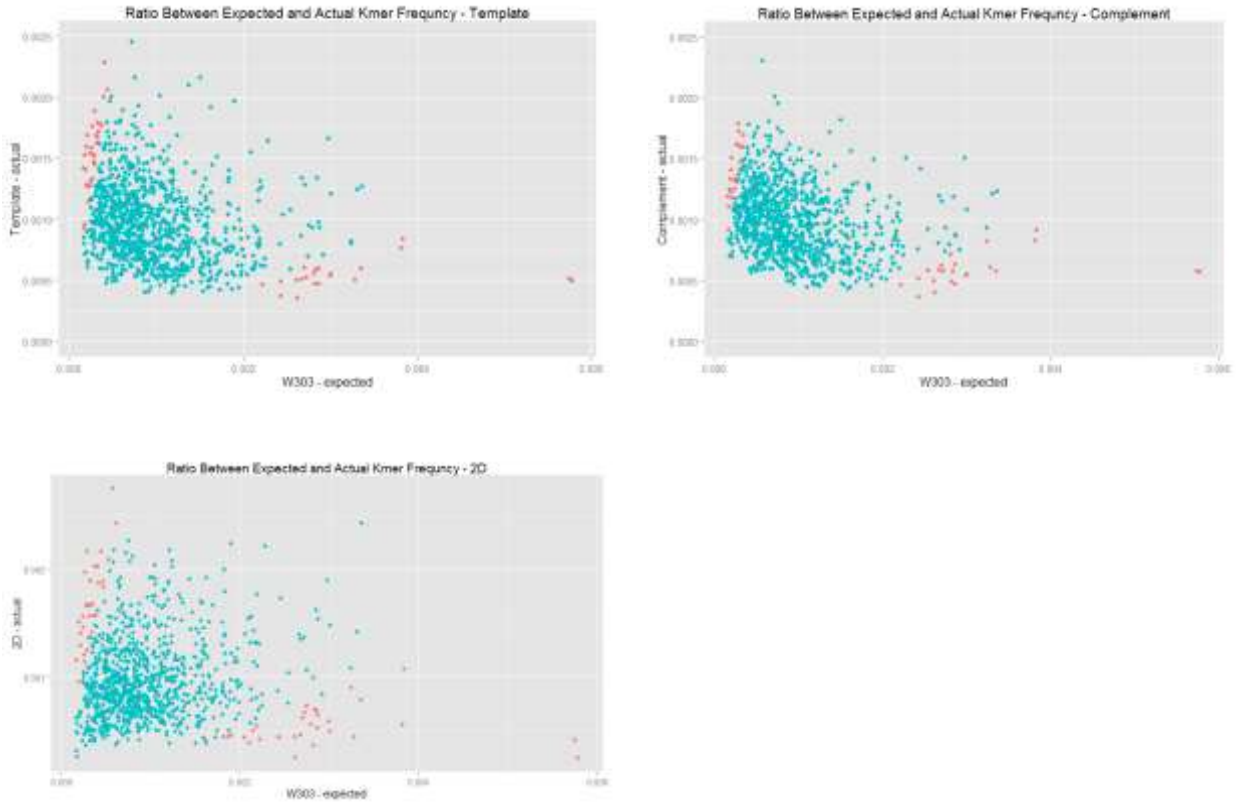


Figure S4. Plots of abundance of each Kmer derived from the reference assembly relative to the uncorrected ONT reads. Red points indicate Kmers whose abundance deviates significantly from the expected. Top-left: Template only. Top-right: Complement only. Bottom-left: 2D only.

	Template	Complement	2D	All		Template	Complement	2D	All
AAAAA	-3.053216224	-3.055223444	-3.541083529	-3.223118452	CGCCC	2.377197144	2.152552135	2.69156761	2.470425
AAAAT	-2.113162203	-2.103921716	-2.037259713	-2.040971679	CGCCG	2.057025334	2.072202308	2.579153211	2.261348
AAATT		-1.994225579			CGCGC		2.004493838		
AACCC			2.081268967		CGGAC			2.025032481	
AATAT	-2.08456369	-2.182007813	-2.103047878	-2.153604618	CGGGG	2.568402295	2.529033138		2.252666
AATTT			-2.002873207		CGTCC		2.084226986	2.275788612	2.215896
ACCCC	2.218668237		2.602201691	2.343629425	CGTGC	1.998172196			
ACCCG	2.497741879	2.390525097	3.095271843	2.82231925	CTCCC	2.34801972		2.385842355	2.419484
ACGCC			2.075635007		CTTTT			-2.143400025	
AGGGG	2.164427287				GACCC	2.022590088		2.55509798	2.254329
ATATA	-2.1372818	-2.231399946	-2.456480278	-2.336892428	GACCG			1.998093166	
ATATT	-2.289948465	-2.331255992	-2.752995248	-2.517458498	GCCCC	2.335097921		2.10023139	2.094073
ATCAT			-1.999774266		GCCCG	2.190178031	2.114655508	2.513116698	2.275965
ATTAT	-2.330518943	-2.420730233	-2.37641674	-2.466462359	GCGGG	1.992567672			
ATTCA			-2.033298413	-2.037492014	GGGGA	2.139721639			
ATTCT			-2.018764396		GGGGG	2.775112393	2.47951193		2.107828
ATTTT	-2.204298679	-2.202888279	-2.741977652	-2.363161782	GGGGT	2.107911159			
CAAAA	-2.128193022	-1.96199621			GGTCG		1.972109762		
CACCC			2.033975661	1.97937513	GTCCC			1.970182957	
CATTT			-2.013838676		GTGCG	2.146248053			
CCCCC	1.996529492				GTTTT			-2.080732806	
CCCCG	2.757440972	2.498174441	2.688839833	2.657496074	TATAT	-2.452863472	-2.422864105	-2.987537866	-2.70258
CCCCT	2.096567036		2.318021107	2.043279394	TATTC			-2.012805786	
CCCGA	2.502532174	2.4018232	3.043731955	2.778817404	TATTT	-2.221416723	-2.298897885	-2.542226338	-2.3971
CCCGC	2.741862706	2.595233813	2.71051705	2.763056074	TCAAA		-2.041187565		
CCCGG	2.713398257	2.547644305	2.935201588	2.797639338	TCATT			-2.306604091	-1.98677
CCCGT			2.091829824		TCCCG	2.207586302	2.019098438	2.451294148	2.323375
CCCTA			2.212135906		TCTTT	-2.093233206	-2.101719105	-2.203182023	-2.16229
CCCTC	2.046505455		2.399626203	2.101608973	TGTTT			-2.15652272	-2.00994
CCGAC	2.074318539	2.294426231	2.577433246	2.401777513	TTATT	-2.306560781	-2.390332264	-2.414863382	-2.42953
CCGAG	2.009455115	2.024826469			TTCAA	-2.152233614	-2.134392548	-2.090721106	-2.16476
CCGCC	2.093515908	2.106327759	2.276713825	2.254262651	TTCAT	-2.011714999	-2.087343394	-2.306216508	-2.21322
CCGCG	1.986576051	2.335008314			TTCTT	-2.284798939	-2.388558305	-2.201153767	-2.30487
CCGGC	2.068256101	2.150190025		2.063984908	TTGTT	-2.080864657	-2.100787632	-2.375856077	-2.25063
CCGGG	2.30821882	2.279267315	2.062563899	2.215593101	TTTAT	-2.013001147	-2.045124068	-2.16746485	-2.10943
CCTAC			1.983880719		TTTCA	-2.154038614	-2.100565268	-2.458578709	-2.28709
CCTCC	2.074418554		2.462908493	2.430225441	TTTCT	-2.247965116	-2.316517226	-2.359168742	-2.35036
CCTCG			1.998310453		TTTTA			-2.004523423	
CGAAC			2.049621182		TTTTC	-2.4134082	-2.309013352	-2.761322371	-2.52937
CGACC	2.115556349	2.466399958	2.57598353	2.491988615	TTTTG			-2.318213911	
CGACG		1.985581294			TTTTT	-3.085522636	-3.07631273	-3.957431003	-3.35579

Supplementary Table 1. List of significant Kmers from Complement, Template, 2D and combined reads, values represent standard deviations from expected. TTTTT was the most deficient Kmer in all cases. GGGGG was the most enriched in the Template strands, CCCGC was the most enriched in the Complement stands and ACCCG was the most enriched in the 2D and combined data.

Supplemental Note 5. Raw read accuracy and coverage

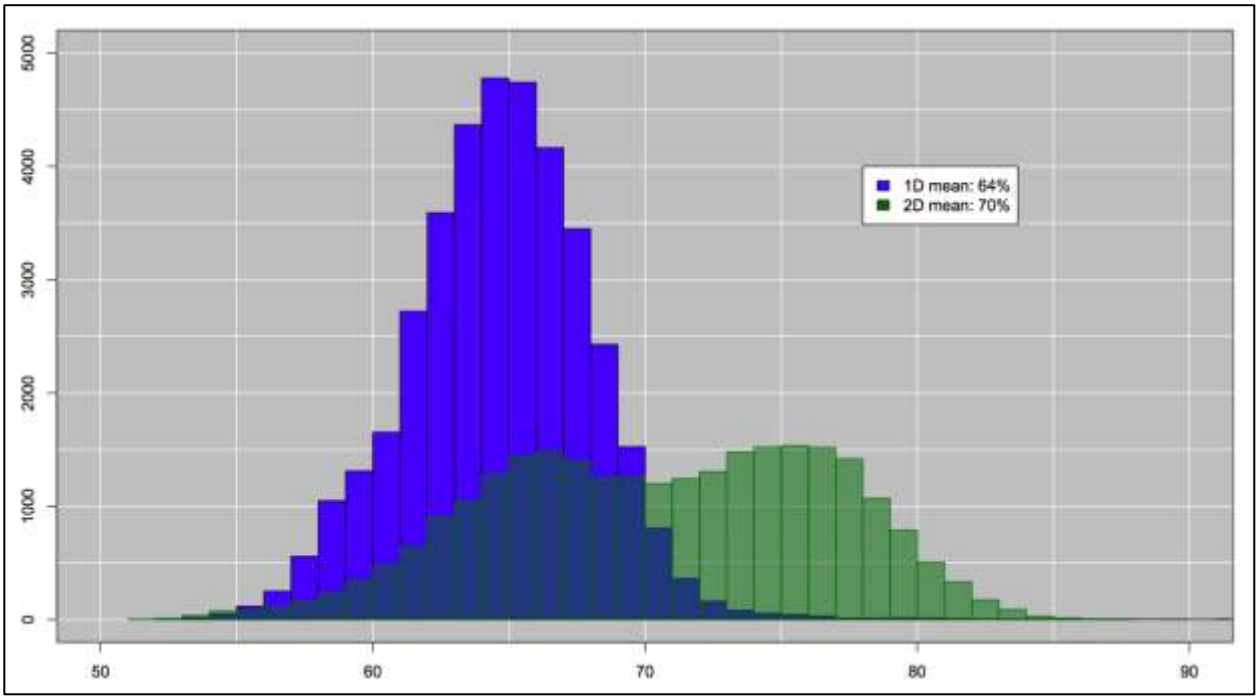


Figure S5A. Histogram of 1D and 2D read accuracy as determined by aligning the reads to the reference genome using BLASTN. The 1D reads averaged 64% accuracy, while the 2D reads average 70% accuracy. We speculate the bimodal distribution in the accuracy of the 2D is explained by a failure in 2D basecalling for some of 2D reads so that it reverts to 1D accuracy.

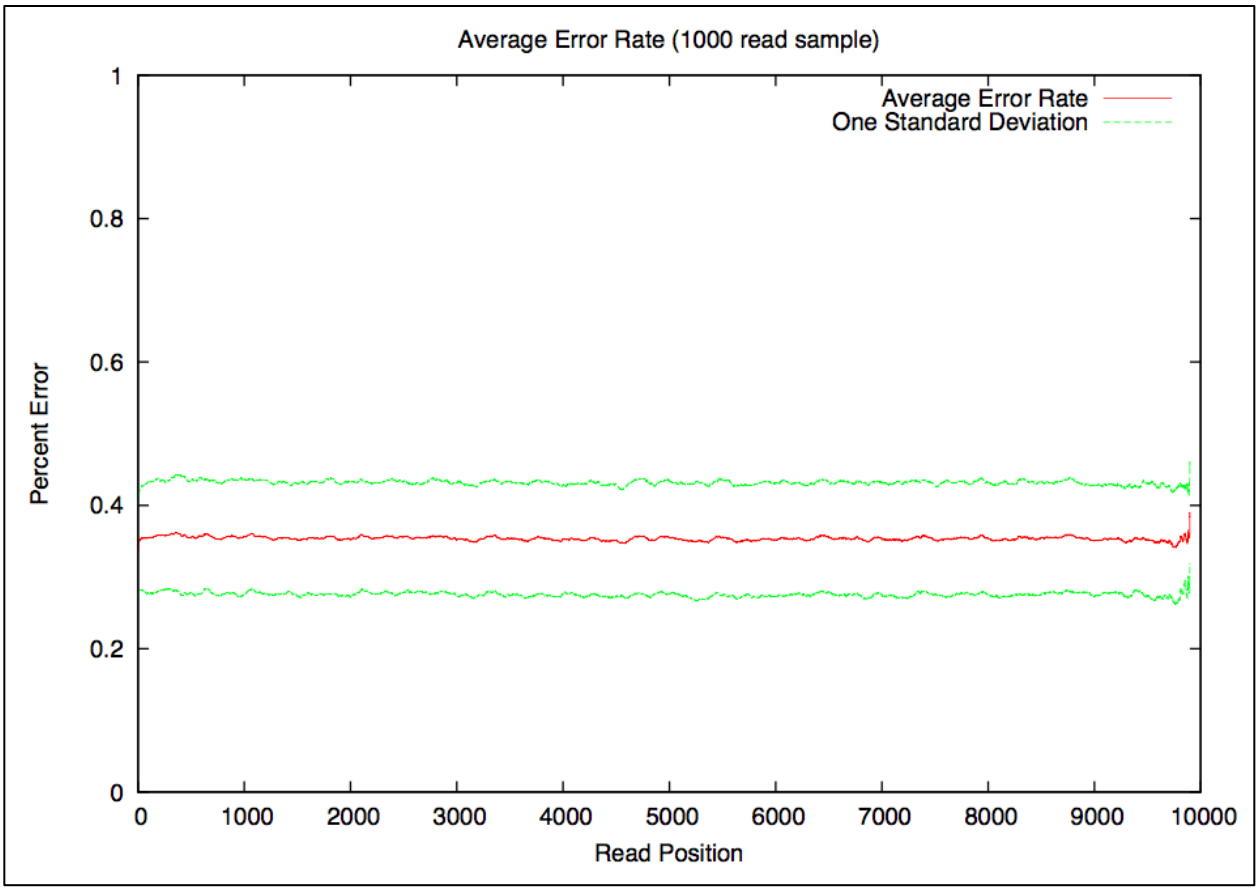


Figure S5B. Average error rate over the length of the read (red). Green lines indicate one standard deviation. 1000 reads with lengths between 9kb and 10kb were sampled and error rate was calculated for 100 basepair sliding windows using blastn alignments to the S288C reference genome

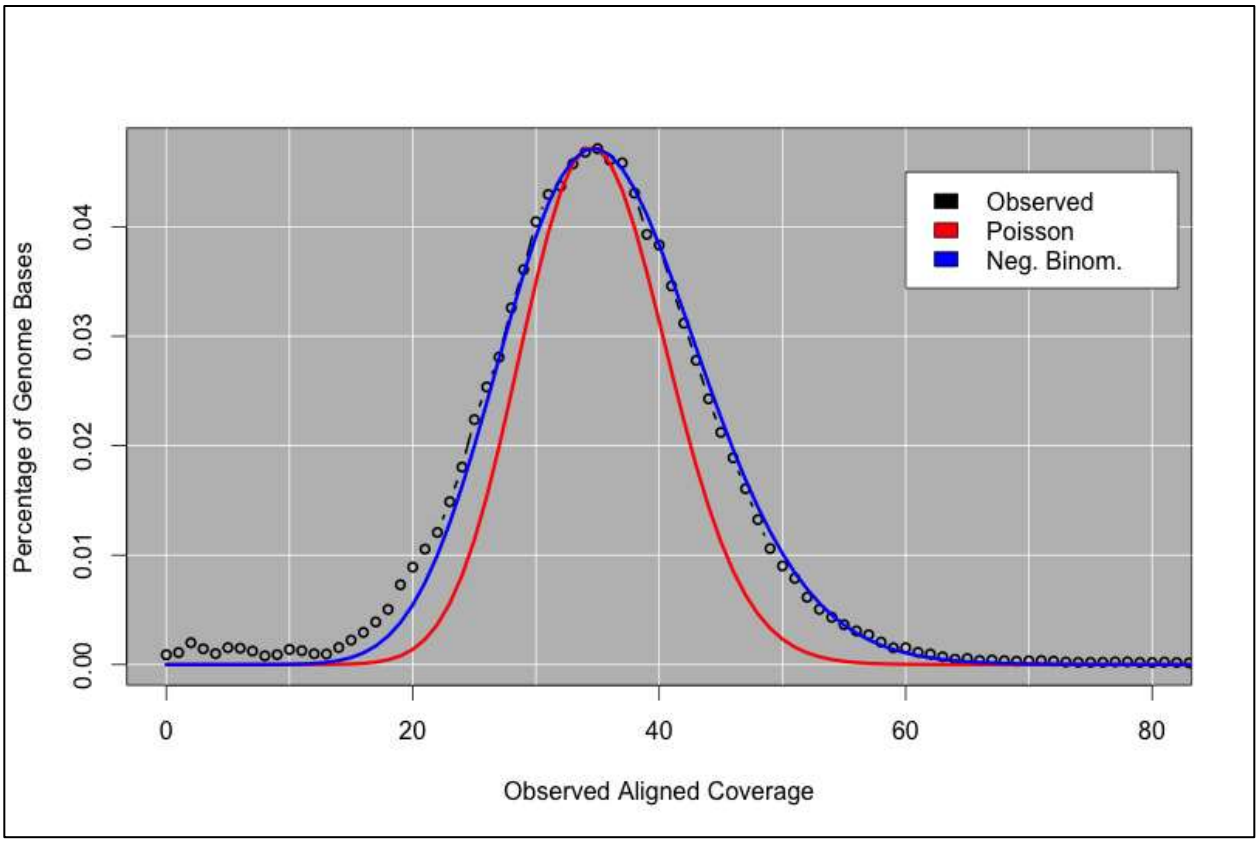


Figure S5C. Distribution of alignment coverage across genome. The observed coverage (black circles) approximates the expected Poisson distribution (red), although is better represented as a negative binomial distribution with a larger standard deviation.

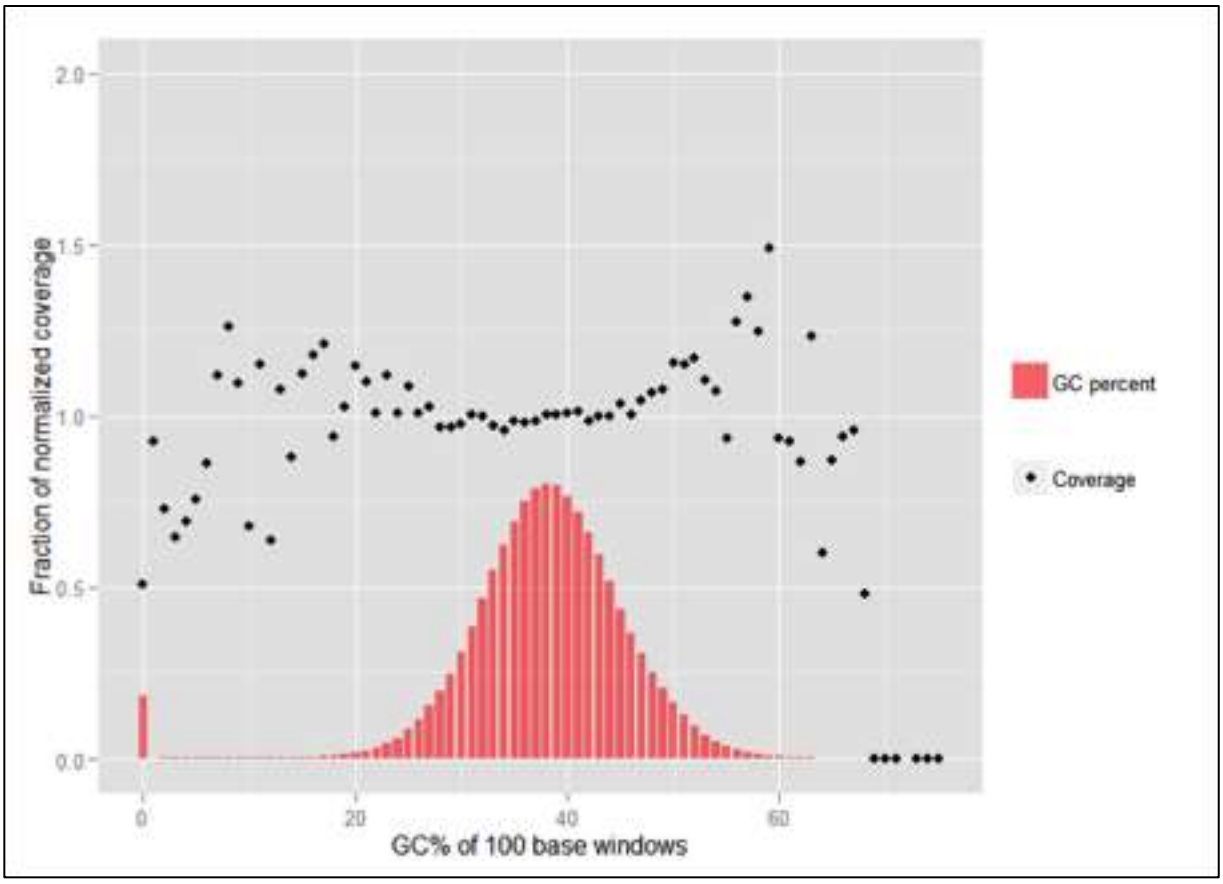


Figure S5D. Plot of coverage relative to the GC content of the W303 genome. Coverage (black dots) is mostly uniform at average GC content (30-50% GC) while high and low GC content shows a more variable coverage profile.

Supplemental Note 6. Nanocorr performance of yeast genome

The raw and error corrected reads, along with the final assemblies and parameters used for the error correction and assembly can be found on the nanocorr website:

http://schatzlab.cshl.edu/data/nanocorr/

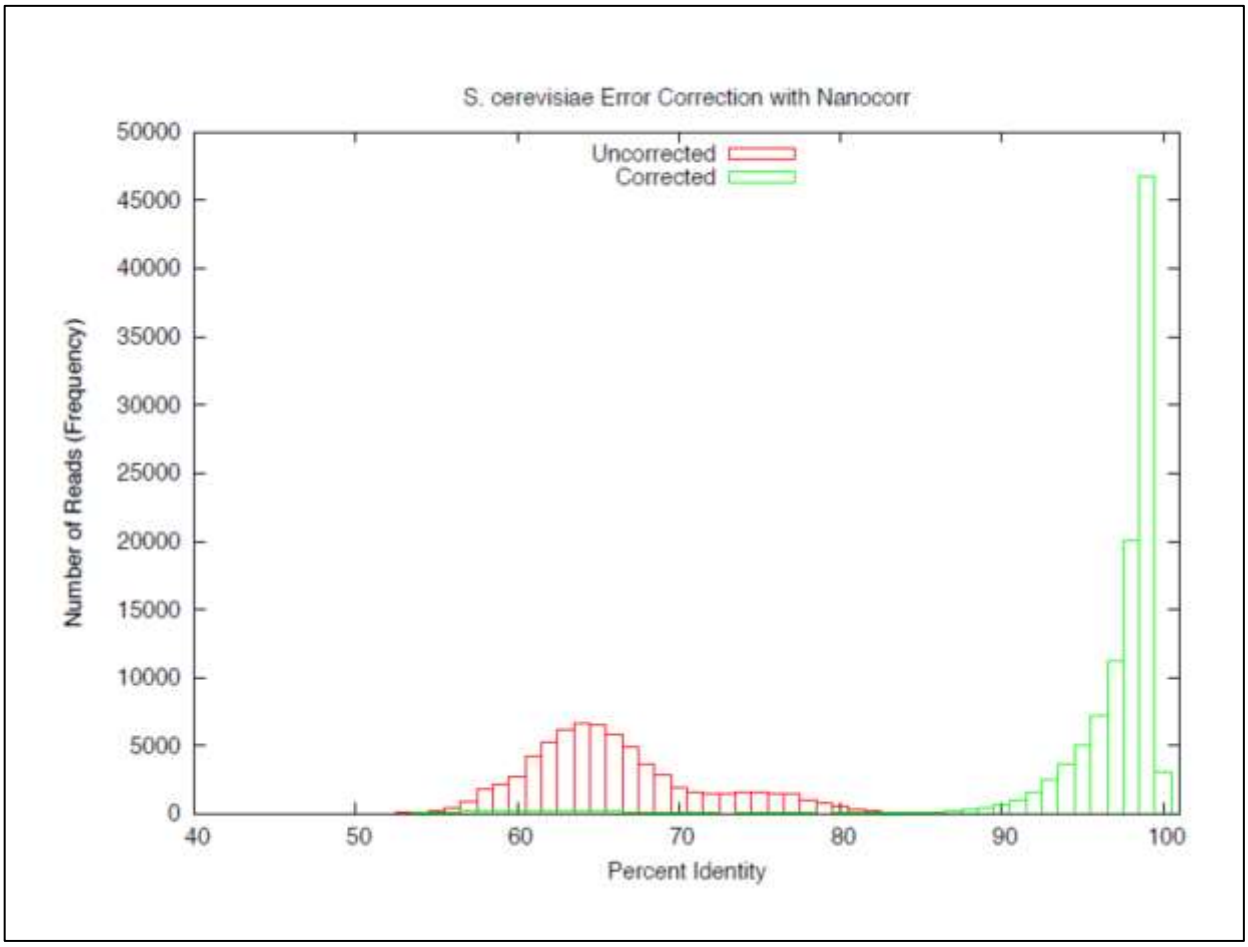


Figure S6A. Histogram of the percent identity of reads before and after error correction with Nanocorr in yeast. After correction, the long read accuracy improves to over 97% on average.

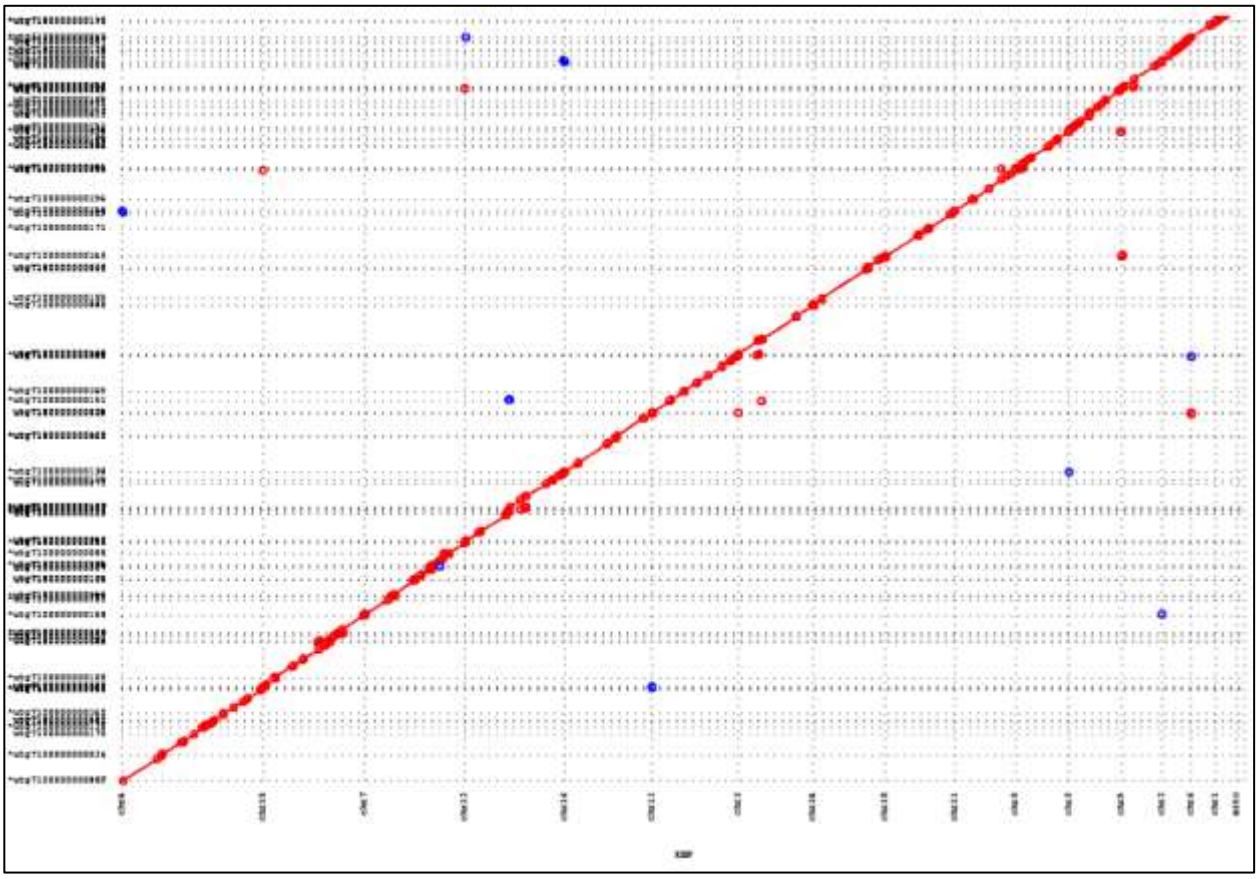


Figure S6B. Dot plot of Nanocorr-corrected Oxford Nanopore assembly (y-axis) of yeast versus the reference genome (x-axis).

Supplemental Note 7. Nanocorr performance of E. coli K12 genome

The raw and error corrected reads, along with the final assemblies and parameters used for the error correction and assembly can be found on the nanocorr website:

http://schatzlab.cshl.edu/data/nanocorr/

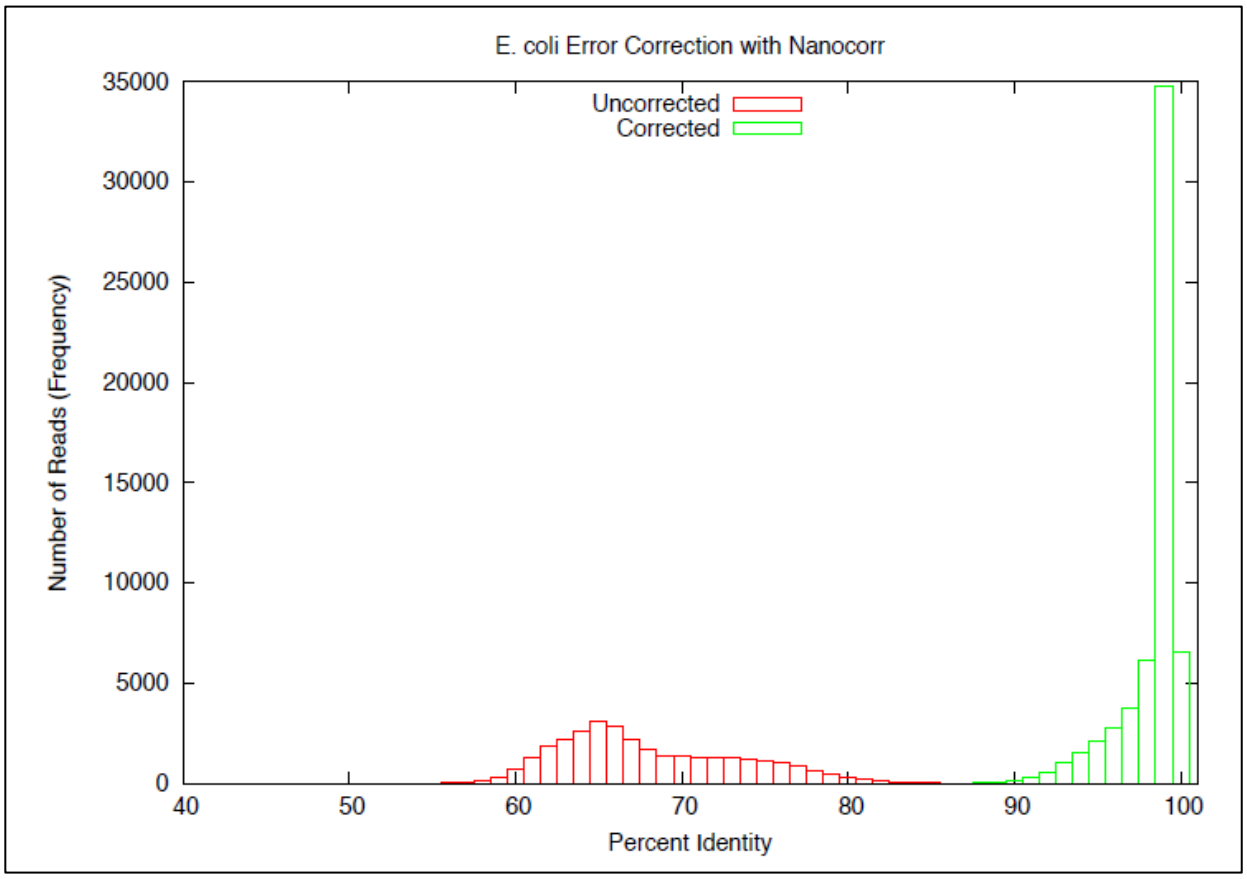


Figure S7A. Histogram of the percent identity of reads before and after error correction with Nanocorr in E. coli K12. After error correction, the long read accuracy improves to over 98% on average.

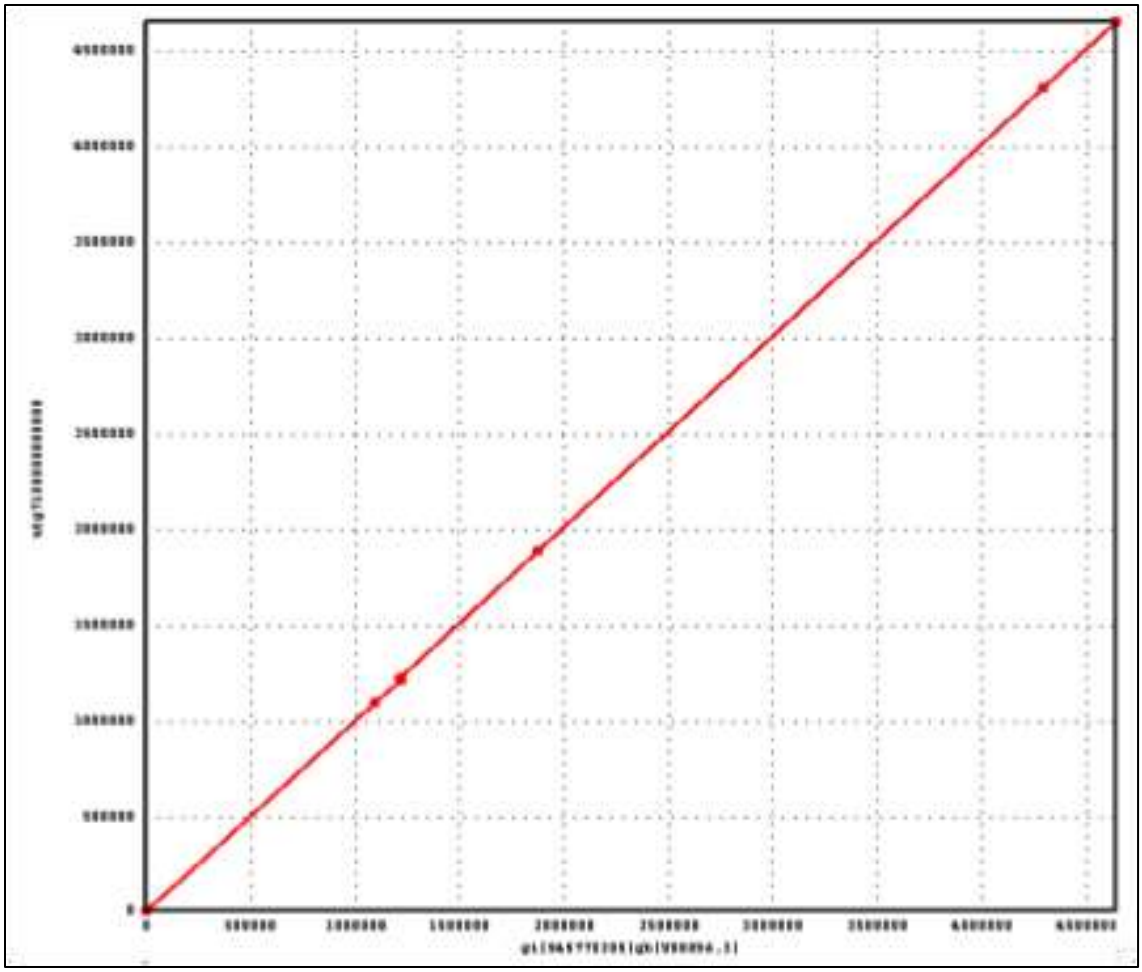
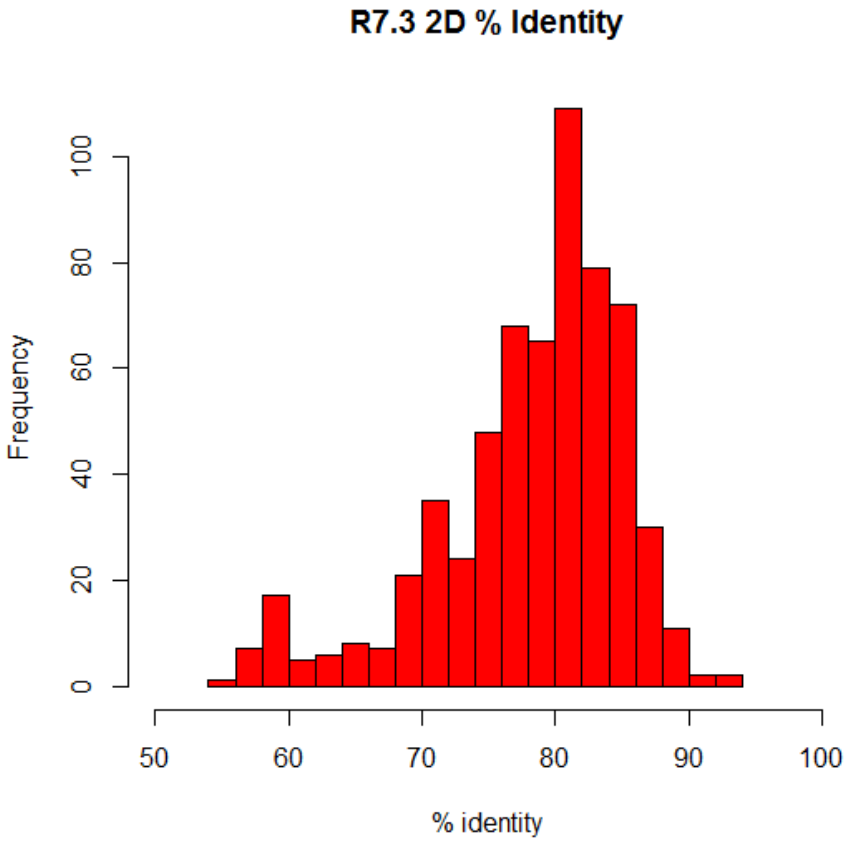


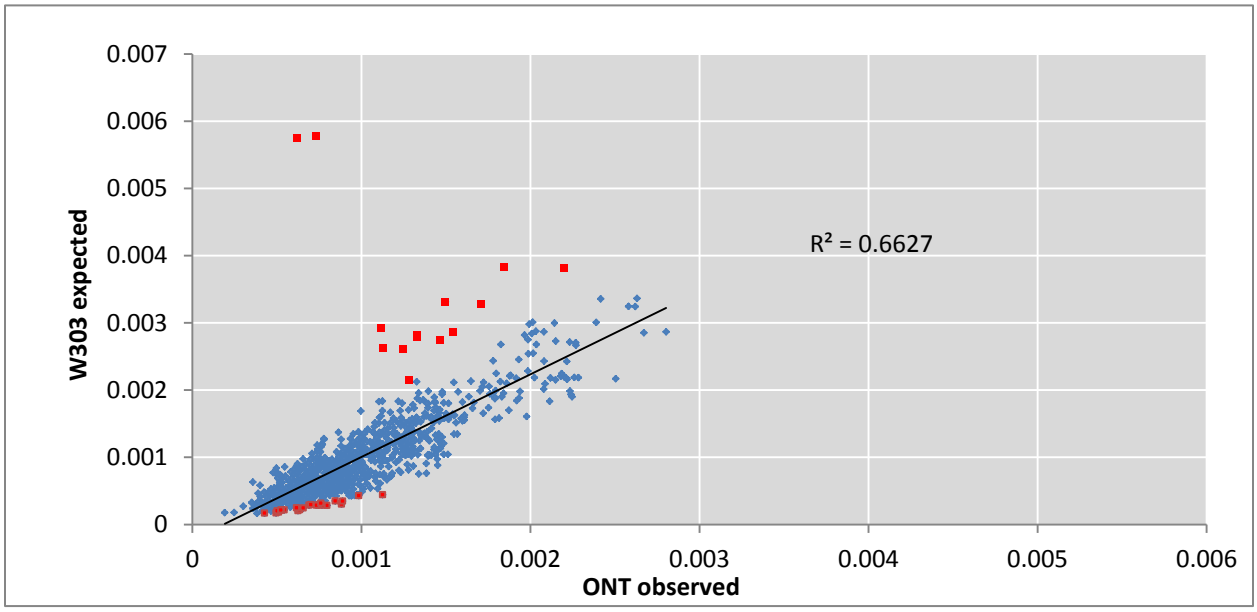
Figure S7B. Dot plot of Nanocorr-corrected Oxford Nanopore assembly (y-axis) of E. coli K12 versus the reference genome (x-axis). The nanocorr corrected assembly consisted of a single near perfect contig shown here as a single line along the diagonal, using dots to highlight the position of a few residual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

Supplemental Note 8 R7.3 performance

The chemistry of the Oxford Nanopore flowcells is dynamic and constantly evolving. We have had the opportunity to test out new chemistries over the course of our MAP participation. As the chemistry has evolved we have observed an improvement in both percent identify and homopolymer representation and we expect to see further improvements as time goes by.



Supplemental Figure S8A. Percent identity of 2D read from R7 and R7.3 flowcells. Average percent identity increased from 70% to 78%. Percent of 2D reads that align remain at 30%



Supplemental Figure S8B. Plots of abundance of each Kmer derived from the reference assembly relative to the uncorrected ONT reads. Red points indicate Kmers whose abundance deviates significantly from the expected. AAAAA and TTTTT remain the most significantly deviating 5-mers however a more linear trend is observed between the 5-mers seen in the reference vs the 5-mers seen in the nanopore data.

Supplemental Note 9. Flow cell details

Date notes the data a flow cell was run as noted in Figure 1 of the main document.

06/23/14 – Performed using ONT supplied ligase, overnight motor incubation , 1 ug DNA starting material

06/26/14 - Overnight motor incubation, 1 ug DNA starting material

07/01/14- Flowcell had been previously used for Lambda burn-in and washed per Oxford protocol, 1 ug DNA starting material, overnight motor incubation

07/02/14 Flowcell had been previously used on 06/23/14 for W303 DNA and washed per Oxford , 1 ug DNA starting material protocol, overnight motor incubation

07/04/14 Flowcell had been previously used on 06/26/14 for W303 DNA and washed per Oxford protocol, 1 ug DNA starting material, overnight motor incubation

07/10/14-1 Ampure bead concentration was 0.4X for all wash steps, DNA was sheared following Covaris instructions 1kb 1ug DN input, overnight motor incubation

07/10/14-2 DNA was sheared following Covaris instructions 10kb, 2ug DNA input, overnight motor incubation

07/11/14-1 DNA was sheared following Covaris instructions 10kb, 2ug DNA input, overnight motor incubation

07/11/14-2 1 Ampure bead concentration was 0.4X for all wash steps, DNA was sheared following Covaris instructions 10kb, 2ug DNA input, overnight motor incubation

07/12/14 Ampure bead concentration was 0.4X for all wash steps, DNA was sheared following Covaris instructions 10kb, 2ug DNA input, overnight motor incubation

07/16/14-1 DNA was sheared following Covaris instructions 10kb, 1ug DNA input , overnight motor incubation

07/16/14-2 DNA was sheared following Covaris instructions 10kb, 1ug DNA input, overnight motor incubation

07/18/14-1 DNA was sheared following Covaris instructions 10kb, 1ug DNA input, overnight motor incubation

07/18/14-2 DNA was sheared following Covaris instructions 10kb, 1ug DNA input, overnight motor incubation

07/21/14 Flowcell was washed following Oxford protocol prior to DNA loading, DNA was sheared following Covaris instructions 10kb, 1ug DNA input, overnight motor incubation

07/23/14 DNA was sheared following Covaris instructions 10kb, 1ug DNA input, 30min motor incubation

07/24/14-2 DNA was sheared following Covaris instructions 10kb, 1ug DNA input, 30min motor incubation

07/25/14 DNA was sized selected post shearing with blue pippin (Sage) at >10kb, DNA was sheared following Covaris instructions 10kb, 1ug DNA input, 30min motor incubation

07/18/14-1 Flowcell was washed following Oxford protocol prior to DNA loading, DNA was sheared following Covaris instructions 10kb, 1ug DNA input, 30min motor incubation

07/28/14-2 flow cell QCd at 0 available pores and appeared to have a crack, DNA was sheared following Covaris instructions 10kb, 1ug DNA input, 30min motor incubation

08/04/14 DNA was sheared following Covaris instructions 10kb, 1ug DNA input, 30min motor incubation

08/05/14 DNA was sheared following Covaris instructions 10kb, 1ug DNA input, 30min motor incubation

08/07/14-1 DNA was sheared following Covaris instructions 10kb, 2ug DNA input then split into two equal aliquots prior to ligation, adapter mix allows to incubate for 5 minutes with ligase prior to adding HP adapter, 30min motor incubation, 240ng DNA estimated added to flow cell

08/07/14-2 DNA was sheared following Covaris instructions 10kb, 2ug DNA input then split into two equal aliquots prior to ligation, adapter mix allows to incubate for 5 minutes with ligase prior to adding HP adapter, 30min motor incubation, 240ng DNA estimated added to flow cell

08/07/14-2 DNA was sheared following Covaris instructions 10kb, 2ug DNA input then split into two equal aliquots prior to ligation, adapter mix allows to incubate for 5 minutes with ligase prior to adding HP adapter, 30min motor incubation, 240ng DNA estimated added to flow cell

08/07/14-3 DNA was sheared following Covaris instructions 10kb, 2ug DNA input then split into two equal aliquots prior to ligation, adapter mix allows to incubate for 5 minutes with ligase prior to adding HP adapter, 30min motor incubation, 240ng DNA estimated added to flow cell

08/14/14-1 DNA was sheared following Covaris instructions 10kb, 2ug DNA input then split into two equal aliquots prior to ligation, adapter mix allows to incubate for 5 minutes with ligase prior to adding HP adapter, 30min motor incubation, 5ng DNA estimated added to flow cell

08/14/14-2 DNA was sheared following Covaris instructions 10kb, 2ug DNA input then split into two equal aliquots prior to ligation, adapter mix allows to incubate for 5 minutes with ligase prior to adding HP adapter, 30min motor incubation, 50ng DNA estimated added to flow cell

08/14/14-3 DNA was sheared following Covaris instructions 10kb, 2ug DNA input then split into two equal aliquots prior to ligation, adapter mix allows to incubate for 5 minutes with ligase prior to adding HP adapter, 30min motor incubation, 5ng DNA estimated added to flow cell

08/24/14 DNA was sheared following Covaris instructions 10kb, 2ug DNA input then split into two equal aliquots prior to ligation, adapter mix allows to incubate for 5 minutes with ligase prior to adding HP adapter, 30min motor incubation, 50ng DNA estimated added to flow cell

Supplemental Note 10. Materials and Methods

Yeast growth

An aliquot of yeast strain W303 was obtained from Dr. Gholson Lyon (CSHL). Four ml cultures in 15 mL falcon tubes of yeast were grown in YPD overnight in at 32°C to ~1x10^8 cells. The cells were purified using the Gentra Puregene Yeast/Bacteria kit (Qiagen, Valencia CA). DNA was stored at -20^{o}C for no more than 7 days prior to use.

Library preparation

Purified DNA was sheared to 10kb or 20kb fragments using a Covaris g-tube (Covaris, Woburn MA). Four ug of Purified DNA in 150 ul of DI water was loaded into a g-tube and spun at 6000 RPM Eppendorff 5424 for 120 sec (10kb) or 4200 RPM for 120 sec (20kb). All DNA was further purified by adding 0.4X AMPure beads. A twisted kimwipe was used to remove all visible traces of ethanol from the walls of the tube. The beads were allowed to air dry and DNA was eluted into 30ul of DI water.

The DNA concentration was measured with a Qubit fluorometer and an aliquot was diluted up to 80 ul. Five ul of CS DNA (Oxford Nanopore, Oxford UK) was added and the DNA was end-repaired using the NEBNext End Repair Module (NEB, Ipswich MA). The DNA was purified with AMPure beads and eluted in 25.2 ul of DI water. DNA A-tailing was performed with the NEBNext dA-Tailing module (NEB, Ipswich MA).

Blunt/TA ligase (NEB, Ipswich MA) was added to the A-tailed library along with 10 ul of the adapter mix (ONT, Oxford UK) and 10 ul of HP adapter (ONT, Oxford UK). The reaction was allowed to incubate at 25°C for 15 minutes. The DNA was purified with 0.4X of AMPure

beads. After removal of supernatant, the beads were washed 1X with 150 ul Wash Buffer (ONT, Oxford UK). After supernatant was removed the beads were briefly spun down and then re-pelleted and the remaining supernatant was removed. A twisted kimwipe was used to remove all traces of wash buffer from the wall of the tube. The DNA was resuspended in 25 ul of Elution Buffer (ONT, Oxford UK).

The DNA was quantified using a qubit to estimate the total ng of genomic+CS DNA in the final library. Ten ul of tether (ONT, Oxford UK) was added to the ligated library and allowed to incubate at room temperature for 10 minutes. Fifteen ul of HP motor was then added and allowed to incubate for 30 minutes or overnight.

Between 5 and 250 ng of the pre-sequencing library was diluted to 146 ul in EP Buffer (ONT, Oxford UK) and 4 ul of Fuel Mix (ONT, Oxford UK) was added and the the sequencing mix. The library was immediately loaded on to a flow cell.

Libraries were sequenced using the MinION device for between 48 and 72 hours. Whenever possible, DNA was handled with a wide bore pipette tip. Mixing of DNA with reagents was done by flicking or preferably pipetting with a wide bore tip. All material loaded onto a flow cell was loaded using a 1000 up pipette. Deviations from this protocol for each flow cell can be found in supplemental methods.

Flowcell disposition

Flowcell were received on ice and immediately stored at 4^{o}C. Ideally within 3 days each flow cell was QC'd with the minKnow software and the number of available pores was

recorded. The flowcells with 400 available pores or more were generally considered “good” and used first. Immediately prior to library loading the flowcell was removed from the 20°C refrigerator and flushed with 150ul of EP Buffer (ONT). The flowcell was allowed to incubate at room temperature for 10 minutes followed by a second EP flush and incubation.

For flowcells that were washed prior to the addition of additional library; the flow cells were washed with 150 ul of Solution A (ONT) followed by a 10 minute room temperature incubation. One-hundred and fifty ul for solution B (ONT) was then added and the flowcells were stored at 4°C until use. Prior to use the washed flowcells were flushed with EP Buffer (ONT) as previously described.

Read Alignment and Error Characteristics

Yield over time data extraction, individual flow cell statistics calculation, fasta/fastq generation were all performed using poretools[2]. Plots were generated using R (ggplot2) and gnuplot. Overall accuracy was calculated by aligning the raw Oxford Nanopore reads to the W303 pacbio assembly using Blast version 2.2.27+ with the following parameters:

-reward 5 -penalty -4 -gapopen 8 -gapextend 6 -task blastn -dust no -evalue 1e-10

High Scoring Segment Pairs were filtered using the LIS algorithm and a scoring function that penalizes overlaps while maximizing alignment lengths and accuracy. Overall accuracy was calculated by averaging the percent identity of all of the filtered HSP’s derived from all of the reads. Error rate over the read length was calculated by taking the HSP’s from a sampling of 1000 random reads in the dataset with read lengths between 9kb and 10kb. The identity was calculated for 100bp sliding windows over the length of the alignment and averaged over all of the alignments.

Read Correction and Assembly

Raw reads were extracted from the h5 files generated by the basecaller. As part of the Nanocorr algorithm, 30x coverage of 300bp paired end MiSeq data was then aligned to the nanopore reads using blastn with the following parameters:

-reward 5 -penalty -4 -gapopen 8 -gapextend 6 -task blastn -dust no -evalue 1e-10

The Nanocorr algorithm then filters the alignments by first removing those contained within a larger alignment and then an LIS Dynamic Programming algorithm was applied using a scoring scheme to minimize the overlaps in the alignments. The filtered set of alignments was then used to build a consensus using 'pbdagcon' (https://github.com/PacificBiosciences/pbdagcon.git). The software and documentation for the error correction software are available open source at https://github.com/jgurtowski/nanocorr. The error correct nanopore reads were then assembled using Celera Assembler version 8.2 (http://wgs-assembler.sourceforge.net/). Alignments and dotplots were generated using 'nucmer' and 'mummerplot' from the MUMmer version 3.23 package[3].

Feature Quantification

Each assembly was aligned to the S288C reference genome using *nucmer* from the MUMmer version 3.23 package. Alignments were filtered using the command *delta-filter -1*, also from the MUMmer 3.23 package to find the best non-redundant set of contigs. The non-redundant set of alignments was intersected with the feature coordinates from the S288C annotation obtained from the Saccharomyces Genome Database using BEDTools[4]

command *intersectBed* with the parameters : -u –wa –f 1.0. The features that were fully contained in an alignment were included in the tally seen in Figure 3.

References

1 Ross, M.G., et al. Characterizing and measuring bias in sequence data. *Genome Biol.* (2013), **14**, R51.

2 Loman, N. J., and Quinlan A.R. Poretools: a toolkit for analyzing nanopore sequence data. *Bioinformatics* (2014): btu55

3 S. Kurtz, A. Phillippy, A.L. Delcher, M. Smoot, M. Shumway, C. Antonescu, and S.L. Salzberg. Versatile and open software for comparing large genomes. *Genome Biology* (2004), 5:R12.

4 A.R. Quinlan and I,M. Hall. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* (2010) 26 (6): 841-842.