**SUPPLEMENTARY INFORMATION**

**for**

**Modeling BioNano optical data and simulation study of genome map assembly**

**Contents**

**S1 Supplementary Methods**

**S2 Supplementary BNX file**

**S3 Supplementary Figures**

**S4 Supplementary Tables**

**S1 Supplementary Methods**

**S1.1 Sample preparation and generation of BioNano molecule data**

To characterize the data properties and data variations of BioNano genome mapping technology, we obtained BioNano molecule data from eight organisms, five sets (*P. putida*, *S. coelicolor*, *S. pombe*, *D. melanogaster*, and *O. sativa ssp. Japonica*) generated by ourselves, and three sets (*E. coli* A7, A9 and PL) generated by BioNano Genomics (Supplemental Table 1). Samples of *P. putida*, *S. coelicolor*, *S. pombe*, *D. melanogaster*, and *O. sativa* ssp. japonica var. Nipponbare were provided by Drs. Sheng Yang, Zhongjun Qin, Yongzhen Xu, Erjun Ling, and Hongxuan Lin, respectively (all at Shanghai Institute of Plant Physiology and Ecology, CAS). These organisms had well-assembled reference genome sequences, which made them ideal to characterize BioNano molecule data by comparative analysis with reference genomes. In addition, the chosen organisms were homogeneous in genetic makeup, either having a single-copy genome or being inbred lines, which, by design, avoided possible complexity due to sequence heterogeneity.

Genomic DNA was extracted and labelled using the IrysPrep Reagent Kit (BioNano Genomics, CA, USA) according to the manufacturer’s standard protocol. Briefly, high-molecular-weight DNA was treated with nicking enzyme Nt.BspQI (New England BioLabs, MA) at 37°C for two hours in NEB Buffer 3. The nicked DNA was labelled with a fluorescent-dUTP nucleotide analogue using Taq polymerase (New England BioLabs) for one hour at 72°C before ligation with Taq ligase (New England BioLabs) in the presence of dNTPs. Star activities of nicking enzymes should be avoided by strictly following the manufacturer’s protocol. Next, the backbone of the labelled DNA was stained with YOYO-1 (Invitrogen). The labelled DNA was loaded onto the IrysChip (BioNano Genomics, CA, USA) by electrophoresis and was imaged automatically with the Irys system. The images were processed and analysed using the IrysView software package (BioNano Genomics, CA, USA).

**S1.2 Assigning SNR and intensity scores for simulated BioNano molecules**

The molecule imaging process can produce noise that impacts downstream applications. The distributions of signals and background noises were investigated within raw BioNano molecule data. To distinguish signals from background noises, BioNano devised the signal to noise ratio (SNR) score, which is the ratio of the labelling site signal intensity to the background signal intensity of surrounding areas. Usually, a greater SNR score indicates a higher confidence to call a signal. When the SNR scores of all molecules were log-transformed, the transformed scores were found to have a bimodal distribution (Supplementary Figure S8) in which signals and background noises were separated by a valley. To filter out background noise, the SNR threshold was determined using the classical histogram bimodal method (Qian and Huang, 1996). In our analysis, a threshold was determined for each scan (Supplementary Table S5) due to the variability of experiments and scan conditions. For the eight BioNano datasets, their labelling site densities after SNR filtering were evaluated and found to approximate the density values deduced from genome sequences (Supplementary Table S1).

For simulated molecules, we assigned all labelling sites, including false-positive (FP) sites, with SNR scores above a pre-set threshold level. For simplicity, we did not include background noise in our simulation. The distribution of log-transformed SNR scores of labelling sites from the experimental datasets approximated that of a Gaussian distribution (Supplementary Figure S9A). Additionally, the SNR scores from experimental data were comparable between true-positive and false-positive sites (Supplementary Figure S9B). Thus, SNR scores were generated based on log-normal distributions  and were randomly assigned to labelling sites on simulated BioNano molecules by BMSIM. In addition, for simplicity and to fulfil the requirements of the BioNano BNX format (S2 Supplementary BNX file), the intensity scores (which were positively correlated with the SNR scores) were generated by multiplying the SNR scores with a constant (*Cnst*) for each scan and were assigned to the labelling sites of simulated BioNano molecules.

**S1.3 Integrating steps for BioNano molecule simulator (BMSIM)**

BioNano Molecule SIMulator (BMSIM) explicitly incorporated BioNano data models (BioNano molecule length distribution, FN and FP signals, DNA molecules stretching variations, variation in optical resolution, and fragile sites) and the methods to generate chimeric molecules and assign SNR scores for simulated BioNano molecules.

**Step I**, generate BioNano molecules with random fragmentation and fragile site bias model. Using genomic sequences (.fasta file) as input, this step produced random BioNano molecules following the model of homogeneous Poisson process. In addition, fragile site bias model was used to estimate the breakage probability of candidate fragile sites marked for BioNano molecules. This only involved a small portion of fragments that contained fragile sites.

**Step II**, label nicking sites for BioNano molecules by *in silico* restriction digestion. Using nicking enzymes’ restriction sites pattern, this step labeled all the nicking sites for BioNano molecules. Our program supported all available nicking enzymes (Nt.BspQI, Nb.BbvCI, Nb.Bsml and Nb.BsrDI), as well as any artificial nicking sequences that users chose to define.

**Step III,** incorporate data models for FN sites, FP sites, stretching variations, optical resolution, and chimerism for BioNano molecules. As described earlier, for FN labels, we treated each nicking site as a Bernoulli event with probability of success *p*. For FP labels, FP sites were generated using the model of Poisson Process (Methods section 2.3). For stretching bias, the stretching variation factor ****, was generated following the Gaussian model:  (Methods section 2.4). For variation in optical resolution, the likelihood of resolving two neighbouring sites was estimated with cumulative Gaussian:

.

Finally, the chimeric molecules were generated as described in Methods section 2.7.

**Step IV,** assign SNR and intensity scores forlabelling sites. For the nicking sites of BioNano molecules produced from previous steps, SNR and intensity scores were generated based on log-normal distributions (Supplementary Methods Section 1.2).

**Step V,** iterate for targeted coverage depth.To reach the targeted coverage depth of simulated BioNano data, steps I to IV were repeated until a predetermined coverage depth was reached. BMSIM produced a BNX file as an output that contained molecule length, labelling sites, and SNR score for each site (S2 Supplementary BNX file).

Many parameters used in simulation of BioNano data were estimated first from our experimentally generated data sets. Thus, to approximate the real data, the default parameters used in simulation were set with mean molecule length =70 Kb (equivalent to molecule N50 of ~170 Kb), enzyme nicking efficiency p=0.85, false nicking rate λ=0.00725/Kb, and chimera fraction for bimera b=85%, that for trimera t=13%, and that for quarmera q=2%. For stretch variation factor****, mean µ was set at 0.98 and standard deviation σ was set at 0.08. For SNR scores , mean µ was set at 3 and standard deviation σ was set at 0.66. For optical resolution, the likelihood of resolving:

,

mean µ was set at 1.2 Kb and standard deviation σ was set at 0.9. For fragile likelihood, constant  was set at 0.7758 and  was set at -0.006984.

Note that BMSIM used the Mersenne Twister algorithm (Matsumoto and Nishimura, 1998) to generate random numbers. The statistic distributions were generated using a Perl port of the C version of randlib (Krogh, 1987). Model parameter estimation was based on non-linear regression using the curve fitting tool (cftool) from MATLAB (R2016a). Model discrimination was based on the goodness of fit, which was evaluated by the coefficient of determination (R2).

**S1.4 Criterion to acquire accurate alignments**

The alignment of BioNano data with　reference genomes was performed with RefAligner from IrysView package, which was described in previous studies (Cao, et al., 2014; Kawakatsu, et al., 2016). For alignments, the standardized parameters were used as recommended (Du, et al., 2017; Kawakatsu, et al., 2016; Xiao, et al., 2015) (<https://forums.bionanogenomics.c>om/forum/data-analysis) except where indicated otherwise: “-nosplit 2 -BestRef 1 -biaswt 0 -Mfast 0 -FP 1.5 -FN 0.15 -sf 0.2 -sd 0.0 -A 5 -outlier 1e-3 **-**outlierMax 40 -endoutlier 1e-4 -S -1000 -sr 0.03 -se 0.2 -MaxSF 0.25 -MaxSE 0.5 -resbias 4 64 -maxmem 64 -M 3 3 **-**minlen 150 **-**T 1e-11 -maxthreads 32 -hashgen 5 3 2.4 1.5 0.05 5.0 1 1 3 -hash -hashdelta 10 –hashoffset 1 -hashmaxmem 64 -insertThreads 4 -maptype 0 -PVres 2 -PVendoutlier -AlignRes 2.0 -rres 0.9 -resEstimate -ScanScaling 2 -RepeatMask 5 0.01 -RepeatRec 0.7 0.6 1.4 -maxEnd 50 –usecolor 1 -stdout –stderr”.

(Detailed de[scription of parameters can be found at BioN](file:///K:\backup\20161222\BMSIMV5\BMSIMsubmitV1\Bioinformatics\BioinformaticsSubmitV1.2\responseToReviewers\scription%20of%20the%20parameter%20see%20BioN)ano Genomics website: [https://forums.bionanogenomics.com/](https://forums.bionanogenomics.com/forum/data-analysis/)).

Some key parameters for accurate alignment we used are as follows:

1. Parameter **M** designates how many alignment iterations to perform. After each iteration of alignment, noise parameters are estimated, and those noise parameters are used for the next iteration. We set this parameter at value (M 3 3), which gave accurate error estimate and alignment (https://forums.bionanogenomics.com/foru

-m/data-analysis)

1. **outlierMax** limits the maximum size for outlier (https://forums.bionanogenomics

.com/), which controls how tolerant RefAligner is for the size of outliers. If significant structural differences are expected between the sample and the reference, this argument may be modified or removed. Since our sample and reference are same, we set this parameter to 40, as recommended by BioNano Genomics (https://forums.bionanogenomics.com/forum/data-analysis).

1. **T** is the alignment significance threshold ([https://forums.bionanogenomics.com/](https://forums.bionanogenomics.com/forum/data-analysis/)). It should be set based on genome complexity, which approximately scales with the genome size and average label density. As recommended by BioNano Genomics, we set value of T at 1e-5/(genome size in Mb) (Shelton, et al., 2015).
2. **minlen** is the threshold for molecule length filter used in alignment. We used a modified parameter value (100 Kb) to study the properties of BioNano molecules at different length ranges.

To ensure accurate alignment, we also evaluated our alignment results using following indicators:

1. molecule-to-reference map rate (%) of the alignment results. If the rate is higher than the desired map rate, i.e. > 60% (recommended by BioNano Genomics), it is considered to be of high quality.
2. uniformity of the alignments. Alignments are visualized in IrysView to evaluate coverage and uniformity of alignments.

**S2 Supplementary BNX file**

The BNX Format

# BNX File Version:

# Label Channels:

# Nickase Recognition Site 1:

# Min Molecule Length (Kb):

# Label SNR Filter Type:

# Min Label SNR:

# Software Version:

#rh SourceFolder InstrumentSerial Time NanoChannelPixelsPerScan StretchFactor BasesPerPixel NumberofScans ChipId Flowcell LabelSNRFilterType MinMoleculeLength MinLabelSNR RunId

# Run Data C:\Users\Administrator\Desktop\v1.8\test-data\2012-08\2012-08\A7 11409 6244 s4 2012-08-22 17\_26\Detect Molecules 10032011101078/22/2012 5:26:10 PM 5540834 0.85 492.268598615917 3 20245,11409,8/20/2012,0006244 1 Dynamic 0 3.5 1

# Quality Score QX01: SNR

# Quality Score QX02: Ave Intensity

#0h LabelChannel MoleculeId Length AvgIntensity SNR NumberofLabels OriginalMoleculeId ScanNumber ScanDirection ChipId Flowcell RunId GlobalScanNumber

#0f int int float float float int int int int string int int int

#1h LabelChannel LabelPositions[N]

#1f int float

#2h LabelChannel LabelPositions[N]

#2h int float

#Qh QualityScoreID QualityScores[N]

#Qf str float

0 1 49718.2 0.055443 12.785 5 1 1 -1 20245,11409,8/20/2012,0006244 1 1 1

1 12316.1 20661.5 23351.5 30780.5 44097.4 49718.2

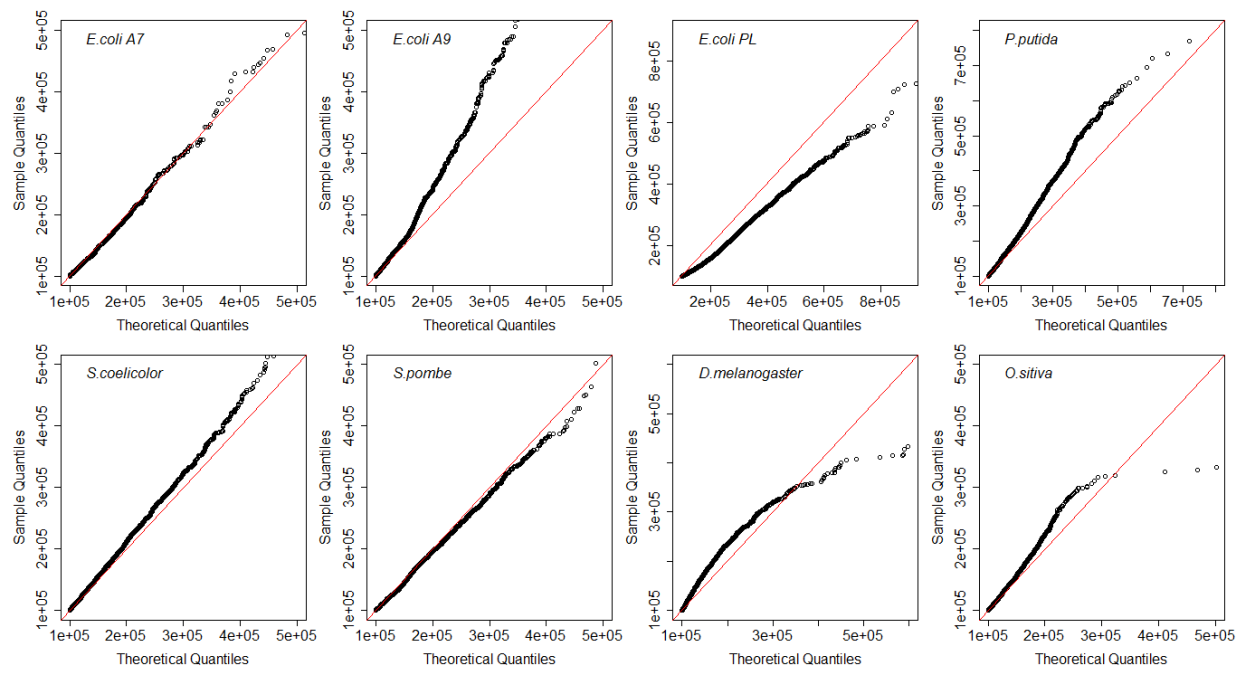
QX11 38.4613 19.6457 11.0611 12.5501 25.5341

QX12 0.0839 0.0444 0.0243 0.0195 0.0395

**S3 Supplementary Figures**

****

**Supplementary Figure S1 Length distribution of BioNano molecule data generated from eight organisms.** The solid (BioNano) and dashed (Exponential) lines represent the distributions of real molecules and that of theoretical exponential model, respectively.



**Supplementary Figure S2.** **Quantile plots (**Q-Q **plots) for the randomly generated, independent exponential length data (vertical axis) versus real BioNano molecule length population (horizontal axis).** Quantile plot compares two probability distributions by plotting their quantiles against each other. For two identical distributions, their quantile plot follows the 45° line y = x. The linearity of the points suggests that the BioNano length data are exponential distributed.

**Fig.3.tif**

**Supplementary Figure S3.** **Variations of stretch for BioNano molecule data generated from *O. sativa*.** We introduced ‘stretch variation factor’  as a measurement of stretch variation for each DNA molecule in BioNano system (Methods section 2.4). Stretch variation factor  between Irys chips (first row), between flowcells of same chips (second row), between runs for same flowcells (third row), and between scans within a run (forth row) were displayed.

supFig.3-20170209

**Supplementary Figure S4.** **Variations of stretch for BioNano molecule data generated from other seven datasets in Supplementary Table S1.** ‘Stretch variation factor’ **** is a measurement of stretch variation for each DNA molecule in BioNano system (Methods Section 2.4). Stretch variation factors between runs (right panel), and between scans (left panel) were displayed. C, chip; F, flowcell, R, run. C1F1R1 represent the first Run in Flowcell 1 of Chip 1.

supFig.6.tif

**Supplementary Figure S5. Distributions of stretch variation factor  for molecules of the same run in the eight datasets.** ‘Stretch variation factor’ **** is a measurement of stretch variation for each DNA molecule in BioNano system (Methods Section 2.4). The density plots show that the distribution of stretch variation factor (red line) approximates to Gaussian (blue dotted line).

supFigure6.tiff

**Supplementary Figure S6. Effects of different molecule lengths and alignment thresholds (T) in making the coverage distribution over-dispersed and/or skewed.**

**(A) Distribution of the simulated BioNano data sets with different N50 for *P. putida*.** We used BMSIM to simulate BioNano experiments with the *P. putida* genome, and generated BioNano data (~50×) of different N50 values, i.e., 100 Kb, 150 Kb, 200 Kb. We found that having larger N50 values for the genome appeared to make the distribution of coverage depth closer to that of a Poisson distribution (with less bias).

**(B) Altering the** **alignment threshold (T) did not improve the evenness of the coverage depth distribution.** We further check that if altering the alignment threshold (T) will improve the evenness of the coverage depth distribution of the simulated data of *P.putida* genome with N50=100Kb. As shown in the figure, with altering the alignment threshold (i.e., T=1e-5, T=1e-6, T=1e-7), the coverage depth distribution of *P. putida* (N50=100Kb) was still biased from the theoretical Poisson. Thus, altering the alignment parameter did not improve the evenness of the coverage depth distribution. T represents the p-value threshold in a pairwise comparison.



**Supplementary Figure S7. Comparison of real BioNano data coverage depth with simulative data coverage depth.** The second and third panel represent simulative data with and without fragile sites, respectively. The x-axes represent the coordinate of the genome of *P. putida*. The y-axes represent the relative coverage depth of the corresponding position along the chromosome. When the function for simulating fragile sites was disabled, the signature of fragile sites bias became un-recognizable.

logSNR

**Supplementary Figure S8. Density plot of the logarithmic transformation signal score SNR.** We showed 30 scans for each dataset in the figure.

C:\Users\apple\Desktop\BMSIM\materials\BMSIM\BMSIM_V4\Figures\sup\quality score\FP_TP_snr.pngFP_TP_snr**Supplementary Figure S9.** **Comparison of SNR distribution of FP and TP labels.** (A) Histogram of log transformed SNRs of FP and TP labels with an interpolated Gaussian curve (blue line for FP and red line for TP). (B) Boxplots of SNRs of FP and TP labels.

PputidaMapWithDiffLen.tiff

**Supplementary Figure S10. Effects of different filtered lenght cutoffs on the mapping distribution of *P. putida* BioNano molecules. W**e tested different length filter cutoffs, e.g. 100, 150, and 180 Kb, for *P. putida* data. The results showed higher filter cutoffs made the coverage distribution of mapped molecules closer to Poisson distribution. So it appears short BioNano molecules are more responsible for the biases in coverage, and filtering helps reduce the biases.



**Supplementary Figure S11. The relevance of the GC% of the genome and the nicking enzyme choice.** We downloaded 128 genome sequences from NCBI, covering broad phylogenetical systems, including microorganisms, plants (unicellular algae, moss, ferns, gymnosperm, monocotyledon, dicotyledon), animals (mammal, bird, reptile, amphibian, fish, echinoderm，mollush, arthropod, Nematomorpha, annulata, platyhelminth, coelenterate and unicellular animal) with variable genome size ranging from 2.3Mb~10.4Gb, and GC contents ranging from 16.9%~72.6% (Supplementary Table S7). Then, we use BMSIM *in silico* digest the genome sequence with four nick enzyme (Nt.BspQI, Nb.BbvCI, Nb.Bsml and Nb.BsrDI). As shown in the figure, the nicking enzymes Nb.Bsml and NbBsrDl for genomes with a low GC content (<25%), Nt.BspQI for those with a medium GC content (25-40%), and both Nt.BspQI and Nb.BbvCI for those with a higher GC content (>40%). The horizontal geen solid line mark the optimal labeling rates between 7 and 14/100kb; the vertical black dot line mark the locations of GC%.

**S4 Supplementary Tables**

**Supplementary Table S1** BioNano molecule data generated from eight organisms (in .xlsx format).

**Supplementary Table S2** Count of potential fragile sites of the six organisms in our study (in .xlsx format).

**Supplementary Table S3** Simulation of BioNano molecule data for four organisms and comparison with corresponding experimental datasets (in .xlsx format).

**Supplementary Table S4** Nick and label density of eight organisms simulatively nicked with different nicking enzymes (in .xlsx format)

**Supplementary Table S5** SNR threshold for each scan of the eight data sets (in .xlsx format).

**Supplementary Table S6** Summaries of the properties of BioNano data and the observation of these properties (in .xlsx format).

**Supplementary Table S7** Nick and label density of 128 organisms simulatively nicked with different nicking enzymes (in .xlsx format).

**Reference**

Cao, H.*, et al.* (2014) Rapid detection of structural variation in a human genome using nanochannel-based genome mapping technology, *Gigascience*, **3**, 34.

Du, H.*, et al.* (2017) Sequencing and de novo assembly of a near complete indica rice genome, *Nat Commun*, **8**, 15324.

Kawakatsu, T.*, et al.* (2016) Epigenomic Diversity in a Global Collection of Arabidopsis thaliana Accessions, *Cell*, **166**, 492-505.

Krogh, F.T. (1987) ACM algorithms policy, *ACM Transactions on Mathematical Software (TOMS)*, **13**, 183-186.

Matsumoto, M. and Nishimura, T. (1998) Mersenne twister: a 623-dimensionally equidistributed uniform pseudo-random number generator, *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, **8**, 3-30.

Qian, R.J. and Huang, T.S. (1996) Optimal edge detection in two-dimensional images, *IEEE Transactions on Image Processing*, **5**, 1215-1220.

Shelton, J.M.*, et al.* (2015) Tools and pipelines for BioNano data: molecule assembly pipeline and FASTA super scaffolding tool, *Bmc Genomics*, **16**, 734.

Xiao, S.*, et al.* (2015) Rapid construction of genome map for large yellow croaker (Larimichthys crocea) by the whole-genome mapping in BioNano Genomics Irys system, *Bmc Genomics*, **16**, 670.