Rapid Short-Read Sequencing and Aneuploidy Detection Using MinION Nanopore Technology

Shan Wei and Zev Williams[1]
Department of Obstetrics and Gynecology and Women's Health, Albert Einstein College of Medicine. Bronx, New York 10461

ABSTRACT MinION is a memory stick–sized nanopore-based sequencer designed primarily for single-molecule sequencing of long DNA fragments (>6 kb). We developed a library preparation and data-analysis method to enable rapid real-time sequencing of short DNA fragments (<1 kb) that resulted in the sequencing of 500 reads in 3 min and 40,000–80,000 reads in 2–4 hr at a rate of 30 nt/sec. We then demonstrated the clinical applicability of this approach by performing successful aneuploidy detection in prenatal and miscarriage samples with sequencing in <4 hr. This method broadens the application of nanopore-based single-molecule sequencing and makes it a promising and versatile tool for rapid clinical and research applications.

KEYWORDS bioinformatics; cytogenetics; nanopore; single-molecule sequencing; rapid aneuploidy detection

MinION is the first commercialized nanopore-based sequencer for sequencing DNA (Loman and Watson 2015). It records, in real time, changes in electric current as an applied electrical field drives single-stranded DNA (ssDNA) through ~500 nanopores assembled on the memory stick–sized device. The MinION DNA library preparation and data analysis pipeline is designed to sequence and analyze, in parallel, ultra-long DNA fragments as long as 100 kb (Urban *et al.* 2015). The ultra-long-read-length sequences generated by MinION have been used for *de novo* genome assembly and non-reference scaffold building (Quick *et al.* 2014; Ashton *et al.* 2015; Jain *et al.* 2015; Kilianski *et al.* 2015). We hypothesize that using nanopore sequencers to sequence short DNA fragments could enable real-time rapid and parallel sequencing of many-fold more reads in a given time compared with long-fragment sequencing and thereby enable a wide range of new research and clinical applications such as prenatal and preimplantation diagnosis of aneuploidy in an office setting as well as small DNA and amplicon sequencing in the field or clinic.

In the standard MinION sequencing protocol, DNA is fragmented to an average length of >6 kb. DNA ends are then repaired and dA-tailed, and the resulting long DNA fragments are ligated to a kit adapter mix. The adapter mix consists of two DNA adapters: a Y-shaped adapter and a hairpin-shaped adapter (Figure 1A). The Y-shaped adapter has a leader strand that guides DNA to the nanopore and a preattached E5 motor protein that separates the complementary DNA strands and aids in passage of DNA through the pore. The hairpin-shaped adapter enables a "U-turn" at the hairpin and continued sequencing of the complementary strand of a double-strand DNA (dsDNA). The structure of the Y-adapter/template/hairpin-adapter allows the sequencer to generate a template read, a complementary read, and a calibration of these two reads (*i.e.*, a two-dimensional read for dsDNA). The two-dimensional (2D) reads improve sequencing quality from a single dsDNA molecule (Cherf *et al.* 2012; Jain *et al.* 2015). A His-tagged E3 motor protein, attached to the hairpin-shaped adapter during the ligation process, slows sequencing speeds of the complementary strand and is used for purification of DNA fragments ligated to the hairpin-shaped adapter using His-tag bead purification. Although the parallel sequencing capacity of MinION (512 channels) is much lower than that of other platforms (*e.g.*, Illumina MiSeq, 25×10^6, and Life Technologies Ion Proton, 80×10^6), MinION sequences individual nucleotides at a much faster rate (1200–1800 nt/min) compared with Ion Proton and MiSeq

doi: 10.1534/genetics.115.182311
Manuscript received August 27, 2015; accepted for publication October 20, 2015; published Early Online October 23, 2015.
Available freely online through the author-supported open access option.
Supporting information is available online at www.genetics.org/lookup/suppl/doi:10.1534/genetics.115.182311/-/DC1
Sequence data from this article have been submitted to the NCBI Sequence Read Archive as series PRJNA285489.
IRB Protocol #: 2011-642
[1]Corresponding author: Department of Obstetrics and Gynecology and Women's Health, Albert Einstein College of Medicine, 1300 Morris Park Ave., Price Center Rm. 474, Bronx, NY 10461. E-mail: zev.williams@einstein.yu.edu

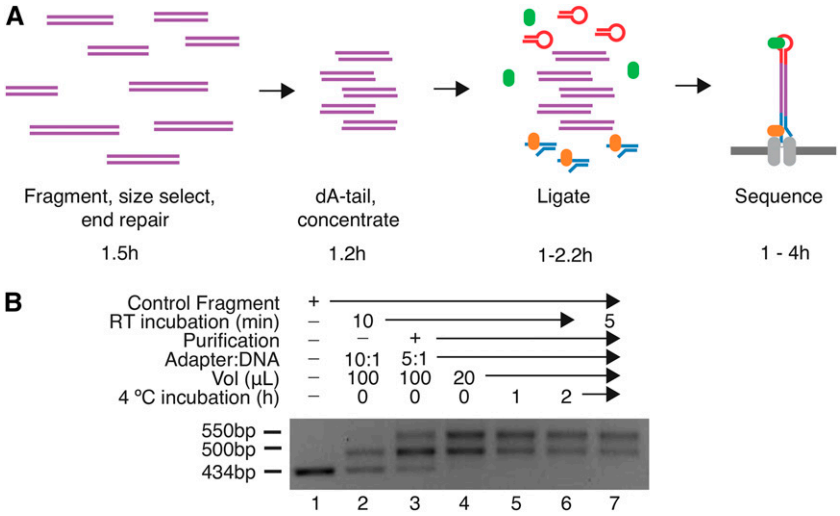


Figure 1 MinION short-fragment sequencing library preparation and optimization. (A) Schematic of the short-fragment sequencing library preparation. dsDNA (purple) is fragmented, size selected, end repaired, and concentrated. Increased concentrations of Y-shaped adapters (blue) with attached E5 proteins (orange) and hairpin adapters (red) are ligated onto the dsDNA, and E3 proteins (green) bind to the hairpin adapters. Electric current then drives a single strand of DNA through the nanopore (light gray). (B) Optimization of short-fragment library preparation. Lane 1, control DNA fragment; lane 2, ligation of control fragment and adapters using manufacturer's protocol; lanes 3–7, incremental improvements in ligation efficiency using purification of fragmented and dA-tailed template DNA (lane 3), reduced reaction volume (lane 4), incorporation of a 1- to 2-hr incubation at 4° (lanes 5 and 6), and reducing room-temperature incubation time to 5 min to reduce release of E5 proteins from adapters (lane 7).

(1 and 0.17 nt/min, respectively). Nanopore-based sequencing also has the distinct advantages that after completing sequencing of one DNA fragment, sequencing of another DNA fragment begins, and reads are generated in real time so that sequencing can be stopped when sufficient reads are obtained.

MinION nanopore genomic DNA library preparation and sequencing protocols were developed for long-fragment sequencing and consequently cannot be used for short-fragment library preparation directly. Moreover, because a short-fragment sequencing run generates significantly more individual reads within a given time than a long-fragment sequencing run, optimization of data analysis was required. However, the rapid quantification of short reads is of particular importance in some sequencing applications, such as ultra low-coverage sequencing (ULCS) for aneuploidy detection and identification of pathogens (Human Microbiome Project Consortium 2012; Palomaki *et al.* 2012; Dillies *et al.* 2013; Chen *et al.* 2014; Wells *et al.* 2014). As part of the MinION access program (MAP), we were given early access to beta test and develop the MinION. In this study, we explore the possibility of using a MinION nanopore sequencer as an ultra-portable device for rapid short-read sequencing. Here we report a library preparation and data-analysis method to enable rapid short-length sequencing on MinION nanopore sequencers and demonstrate its clinical applicability for aneuploidy detection in amniocentesis and miscarriage samples.

Materials and Methods

Development of ligation conditions

To assess the ligation efficiency, a short DNA control fragment was used for initial ligation reactions (Supporting Information, Table S1). The fragment was generated using PCR with M13 forward and reverse primers to amplify a 434-bp fragment from a pCR-Blunt vector (Invitrogen, cat. #K2700-20) using Q5 High-Fidelity DNA Polymerase (NEB, cat. #M0491S) (Table S1). A 50-µl PCR reaction was prepared following the manufacturer's protocol. The PCR reaction was subjected to a 30-sec initial denaturation at 98°, 25 cycles of 10-sec denaturation at 98°, a 30-sec annealing at 57°, and a 20-sec elongation at 72°. A final elongation step at 72° for 2 min was added to ensure complete amplification. The PCR product was purified using a QIAquick PCR Purification Kit following the manufacturer's protocol (Qiagen, cat. #28104). A 57-bp asymmetric adapter with a T overhang was used as a control adapter to assess ligation efficiency (Table S1). The control adapters were diluted to 0.4 µM in MinION adaptor buffer (50 mM NaCl and 10 mM Tris-HCl, pH 7.5) to simulate the 0.2-µM concentration of the Y-shaped and hairpin adapters in the adaptor mix (Oxford Nanopore, SQK-MAP004).

Ligation reactions were initially performed following the MinION Genomic Sequencing Kit protocol (Oxford Nanopore, SQK-MAP004). Control DNA fragments (0.2 pmol, 52 ng) were added to a 30-µl NEBNext dA-Tailing Module (NEB, cat. #E6053S) reaction [4 µl of control fragments, 21 µl of Qiagen Buffer EB, 3 µl of 10× NEBNext dA-tailing reaction buffer, and 2 µl of Klenow fragments (3′→5′ exo-)]. Reactions were performed at 37° for 30 min in a Bio-Rad C1000 Touch Thermal Cycler. All the dA-tailing reactions were added to a total volume of 100 µl [30 µl of dA-tailing reaction, 10 µl of control adapter, 10 µl of nuclease-free water, 50 µl of NEB Blunt/TA Ligase Master Mix (NEB, cat. #M0367S)] and incubated at room temperature (23–25°) for 10 min.

Because so few control fragments had adapters ligated on both ends (Figure 1B, lane 2), an alternative Klenow fragment (3′→5′ exo-) (NEB, cat. #M0212S) was used for dA-tailing, and the dA-tailing reactions were purified before being added to the ligation reactions. Control DNA fragments (250 ng) were subjected to a dA-tailing reaction [2.5 µl of NEBuffer II, 5 µl of 1 mM deoxyadenosine triphosphate (dATP), 1 µl of Klenow fragment (3′→5′ exo-), and nuclease-free water to a total volume of 25 µl]. After purification with 1.8-fold AMPure XP beads (Beckman Coulter, cat. #A63880) following the manufacturer's protocol for the SPRIselect

reagent (Beckman Coulter, cat. #B23317), the dA-tailed control fragment was eluted in 12 μl of 1/5 Qiagen Buffer EB (2 mM Tris-Cl, pH 8; Qiagen, cat. #19086) and diluted to 0.05 μM (13 ng/μl).

Overnight ligation reactions at 16° using T4 DNA ligase (NEB, cat. #M0202S) to ligate a 10:1 adapter:fragment mixture (4 pmol control adapter, 0.2 pmol control fragment in 2 μl 10× T4 DNA ligase buffer, 1 μl T4 DNA ligase, and nuclease-free water to 20-μl final volume) resulted in ∼75% of the control fragments having adapters on both ends, which would not be sufficient final products for downstream steps. Therefore, the reactions were run in duplicate and combined. Then 5:1 adapter:fragment ratios were used to preserve the adapters provided in the MinION kits.

The second ligation reactions were a replication of the manufacturer's ligation protocol using the purified dA-tailed DNA, as described previously (Figure 1B, lane 3), using 100 μl of ligation reaction with 0.4 pmol of DNA, 26 μl of Buffer EB, 10 μl of control adapter, 50 μl of Blunt/TA Ligase Master Mix (NEB, cat. #M0367S), and 10 μl of nuclease-free water (Ambion, cat. #AM9937). Reactions were incubated at room temperature for 10 min and purified using 1.8-fold AMPure XP beads, washed with the wash buffer in the SQK-MAP003 MinION Genomic DNA Sequencing Kit (750 mM NaCl, 10% PEG 8000, 50 mM Tris-HCl, pH 8.0), and eluted in 20 μl of Buffer EB.

The third ligation reactions were a reduced-volume system using purified dA-tailed DNA, as described previously (Figure 1B, lanes 4–7). A 20-μl ligation reaction containing 0.2 pmol of DNA (4 μl), 2 pmol of control DNA adaptor (5 μl), 10 μl of Blunt/TA Ligase Master Mix, and 1 μl of nuclease-free water was incubated for 10 min at room temperature, purified using onefold AMPure XP beads with the SQK-MAP003 wash buffer, and eluted in 20 μl of Buffer EB (Figure 1B, lane 4). Reactions were carried out at room temperature for 5–10 min, followed by 4° incubation for 1–2 hr (Figure 1B, lanes 5–7). Reactions were purified using onefold AMPure XP beads with SQK-MAP003 wash buffer and eluted in 20 μl of Buffer EB. Purified ligation products were run on 2% agarose gels. Portions of the ligation products were estimated using ImageJ (http://imagej.nih.gov/ij/) densitometry analysis with two technical replicates (Figure S5).

Nucleic acid manipulations

To facilitate maximum recovery of material, 1.5-ml low-retention microcentrifuge tubes (Phenix Research Products, cat #MH815S) and low-retention tips (Phenix Research Products, TFL series) were used unless stated otherwise (Figure S2). For all reactions performed in a thermal cycler, 0.2-ml PCR tubes were used (Axygen, cat. #PCR-02-C). An Agencourt SPRIStand Magnetic 6-tube Stand (Beckman Coulter, cat. #A29182) was used for pelleting of SPRIselect and AMPure XP bead–related purification; a DynaMag-2 magnet (Life Technologies, cat. #12321D) was used for His-tag bead isolation.

Genomic DNA samples

Genomic DNA (gDNA) samples from a karyotypically normal male and female, a male with trisomy 12, a male with trisomy 21, and a female with monosomy *X* were used for cytogenetic analysis using short-DNA-fragment ULCS with the MinION. Blood B-lymphocytes from karyotypically normal human male and female samples were obtained from the Coriell Institute Cell Repositories (GM12877 and GM12878) and cultured according to the protocol provided by the Coriell Institute. gDNA was extracted from cell cultures from the second passage using a QIAamp Blood DNA Mini Kit (Qiagen, cat. #51104) following the manufacturer's manual. gDNA from a male with trisomy 21 was provided by the Coriell Institute Cell Repositories (NG05397). DNA samples from a male with trisomy 12 and a female with monosomy *X* were obtained from the products of conception of miscarriage cases that had cytogenetic testing performed using G-band karyotyping. The studies were approved by the Institutional Review Board of Albert Einstein College of Medicine, and patients' consent was obtained. gDNA was extracted using an AllPrep DNA/RNA/Protein Mini Kit (Qiagen, cat. #80004) from the trophoblastic primary cell cultures of the chorionic villus. The quality of gDNA was examined on 0.8% agarose gel and quantified using a NanoDrop 1000 Spectrophotometer (Thermo Fisher Scientific). DNA was stored at −20° until needed.

Library preparation

For library preparation, 120 μl of 25 ng/μl gDNA in TE Buffer (pH 8.0) was fragmented using a Covaris S220 focused-ultrasonicator at the manufacturer's 500-bp setting in microTUBEs (Covaris, cat. #520045). For size selection, 100 μl of fragmented gDNA was used. Size selection was performed in a 1.5-ml DNA LoBind tube (Eppendorf, cat. #022431021) using SPRIselect reagent following the manufacturer's double-sized selection protocol using a right-side 0.55-fold, left-side 0.7-fold setting (Beckman Coulter, cat. #B23317). DNA was eluted in 40–50 μl of Buffer EB in a 1.5-ml DNA LoBind tube. Then 2 μl of DNA was used for a 2% gel electrophoresis to confirm fragment size. Purified DNA (3 μl) was saved for NanoDrop quantification. Size-selected DNA fragments were ∼350–600 bp in length (Figure S1C).

Buffer EB was added to size selected DNA to a final volume of 80 μl. End-repair reactions were performed using a NEBNext End Repair Module (NEB, cat. #E6050S) in a 1.5-ml DNA LoBind tube. Then 5 μl of DNA CS (Oxford Nanopore, SQK-MAP004), 10 μl of 10× NEBNext End Repair Reaction Buffer, and 5 μl of NEBNext End Repair Enzyme Mix were added to the size-selected DNA fragment and mixed by gently pipetting. The reactions were incubated at room temperature for 25 min and purified using 1.8-fold AMPure XP beads following the SPRIselect reagent protocol in a DNA LoBind tube. The end-repaired DNA was eluted in 22 μl of Buffer EB, and the DNA was quantified using a Qubit dsDNA HS Assay Kit (Life Technologies, cat. #Q32854).

End-repaired DNA was subjected to a dA-tailing reaction using a Klenow fragment (3′→5′ exo-) in a total volume of 25 μl in a sterile PCR tube. The reaction contained 2.5 μl of NEBuffer II, 1 μl of Klenow fragment (3′→5′ exo-), 16.5 μl

of end-repaired purified DNA, and 5 μl of dATP (1 mM). Reactions were incubated in a Bio-Rad C1000 Thermal Cycler at 37° for 45 min, purified using 1.8-fold AMPure XP beads, and then eluted in 12 μl of 1/5 Buffer EB. The purified product was quantified using NanoDrop and a Qubit dsDNA HS Assay Kit (Life Technologies, cat. #Q32854) and diluted to ~0.05 μM (~18 ng/μl) with 1/5 Buffer EB to be used as the dA-tailed DNA in subsequent reactions.

His-tag Dynabeads (10 μl) (Invitrogen, cat. #10103D) were washed in 1.5-ml low-retention tubes in a MinION Genomic DNA Sequencing Kit following the manufacture's protocol on a DynaMag-2 magnetic stand (Invitrogen, cat. #12321D). Washed beads were resuspended in 40 μl of undiluted wash buffer (SQK-MAP004) and kept on ice. Ligation reactions were performed in a 1.5-ml low-retention tube. Twenty-microliter reactions contain 4 μl of dA-tailed DNA (0.2 pmol), 5 μl of adaptor mix (1 pmol) (SQK-MAP004), 1 μl of HP adapter (1pmol) (SQK-MAP004), and 10 μl of Blunt/TA Ligase Master Mix (NEB, cat. #M0367S). The reactions were mixed by pipetting gently between each sequential addition and spun down briefly in a benchtop centrifuge. Ligation reactions were incubated at room temperature for 5 min follow by 4° for 2 hr. For each sample, 2 × 20 μl reactions were performed in separate tubes and combined for His-tag bead purification.

In 1.5-ml low-retention tubes, 40 μl of washed His-tag beads were added to the adapter-ligated DNA and carefully mixed by gentle pipetting. The mixture was incubated at room temperature for 5 min and placed on ice for 30 sec. His-tag bead purification was performed following the protocol of the MinION Genomic DNA Sequencing Kit (SQK-MAP004). Pelleted beads were resuspended 28 μl of the ELB elution buffer (SQK-MAP004) by gently pipetting 10 times. The suspension was incubated at room temperature for 5 min and placed on ice for 30 sec, and this was repeated once before placing the suspension back on the magnetic rack for pelleting. The eluate was transferred to a clean 1.5-ml low-retention tube, incubated on ice for 30 sec, and then placed on a magnetic rack for 2 min for pelleting any residual beads. The eluate then was carefully transferred to a 1.5-ml low-retention tube. This library was called the *presequencing mix*. 4 μl of the presequencing mix was used for quantification by a Qubit dsDNA HS Assay Kit.

MinION sequencing

150 μl of the priming mix (147 μl of EP buffer and 3 μl of fuel mix) was loaded on a MinION Flow Cell (R7.3) and incubated for 10 min. The priming process was repeated once. 150 μl of the MinION sequencing library (12 μl of the presequencing mix, 135 μl of EP buffer, and 3 μl of fuel mix) was gently mixed and loaded to the MinION Flow Cell. The MAP 48-hr gDNA sequencing protocol was used, and the sequencing reaction was stopped when sufficient data were collected.

Data analysis

Metrichor agent v2.26 was used to transfer local fast5 files, and 2D BaseCalling Rev1.14 was used to convert currency into base events (Oxford Nanopore Technologies). Poretools v0.5.0 (Loman and Quinlan 2014) was used to convert Fast5 to fastQ files. The first and last 50 bases were removed from each sequence using cutadapt v1.7.1 (Martin 2011), and sequences that were at least 50 bases long were kept after removal. Both one- (1D) and two-dimensional (2D) reads were aligned to the Ensembl GRCh37 human reference genome using BLAT (http://www.kentinformatics.com/) (Kent 2002) (Table S2). Fewer than 1% of 1D sequences passed the screening criteria (covers ≥40% of query and ≥80% alignment identity), and consequently, only 2D sequences were used for further analysis. 2D reads with a unique alignment (UA) match to a genomic location were retained for further analysis. Bowtie 2 also was tested for mapping 2D sequences to a human reference genome (Langmead and Salzberg 2012). Because Bowtie 2 was designed for high-throughput mapping of short sequences (50–200 bp), fewer than 5% of full-length 2D reads could be mapped. Bowtie 2 bwa-sw–like settings (Langmead and Salzberg 2012) developed for Roche 454 pyrosequencing data also were tested, and only 36% of the 2D reads were UA matches (data not shown). Therefore, we used Bowtie 2 to align the first 200 bp of the 2D reads and generated 45% UA matches in ~1 min (Table S3). 2D reads also were mapped to the reference genome with LAST (http://last.cbrc.jp/) using the settings that were reported to be most inclusive for alignment for MinION long reads (Frith *et al.* 2010; Quick *et al.* 2014, 2015). However, this produced fewer UA matches than the BLAT pipeline using the same screening criteria (Kent 2002) (Table S2). Hence, only the UA matches from the BLAT pipeline were used for the fast cytogenetic analysis with ULCS (Chen *et al.* 2014).

Digital karyotyping using ULCS

ULCS is a powerful tool for cytogenetic analysis (Chen *et al.* 2014). As a proof of concept, we performed the analysis on five samples, and a modified ULCS strategy was used for this study. A previous study indicated that the coefficient of variation (CV) in ULCS (<0.01-fold coverage) was lower than 15% on each autosome and that there was no significant difference in the autosomal CVs between the MiSeq and Ion Proton platforms (Chen *et al.* 2014). In a ULCS analysis, we assumed that the UA match on each chromosome (labeled as subscript i, i = 1, 2, …, 22, *X*, *Y*) fits the Poisson distribution (Fan and Quake 2010)

$$UA_i = n_i \varphi_i$$

where n_i is the number of reads needed to cover a chromosome i, and ϕ_i is the coverage of a chromosome i. The percentage of UA matches on each chromosome ($\%UA_i$) is determined by the length and copy number of each chromosome under the same coverage.

The lower limit of sequencing reads needed for ULCS was determined primarily by the UA matches assigned to chromosome *Y* because (1) it is one of the shortest chromosomes, and thus fewer DNA fragments would be sequenced from it, (2) less than 50% of chromosome *Y* has been sequenced and

annotated in the human reference genome, and hence more than half the chromosome *Y* reads would not be able to be mapped to the reference genome and then be counted, and (3) reads mapped to identical regions of chromosomes *X* and *Y* would not be considered as UA matches by the analysis pipeline. Moreover, cross-linking between chromosomes *X* and *Y* and the presence of repetitive elements will cause a small portion of misplacement of reads from chromosomes *X* and *Y* that will further reduced reads that could have been mapped to the *Y* chromosome.

To estimate the lower limit of UA_i needed for ULCS cytogenetic analysis, we used a normal approximation of the Poisson distribution in R (qpois function) to estimate the detection power of UA matches for aneuploidy(R Core Team 2012) (Figure 2F.). It was estimated that the when $UA_i = 41$, $p(x > 1.25\lambda) = 0.04$, $p(x > 1.5\lambda) = 0.0008$, and the detection power of aneuploidy is 90% (Figure 2B). When the $UA_i = 79$, the detection power of aneuploidy would be 95.6%. The corresponding total UA matches for $UA_Y \sim 79$ is ~15,000 in the normal male sample. Thus, 15,000 UA matches were randomly selected 30 times from the sequencing result of the normal male, and the average UA matches for each chromosome were used as a reference for normalization purpose (Ref_UA_i). To examine whether the 15,000 reference is representative of the human genome under a Poisson distribution, we compared the percentage of ungapped length (%UL) and the %UA of each chromosome. The ratio (Norm_Ref_%UA) on autosomes was 1.04 (SD = 0.0687, CV = 6.6%) (Table S6). The 15,000 reference represents the %UA of about a half the %UL of the sex chromosomes, which could be the result of depletion of non-unique alignments on homogeneous regions of the sex chromosomes. The mitochondrial chromosome (MT) is a multicopy small chromosome (Miller *et al.* 2003), and it was not included in the ULCS cytogenetics analysis. According to the Poisson distribution, the 99.9% confidence intervals of each chromosome of the normal male reference can be estimated as $Ref\_UA_i \pm 3.29\sqrt{Ref\_UA_i}$ under the same coverage.

To access the copy number of each chromosome of a query sample using 15,000 UA match reads (Table S7), we assumed that the number of UA reads on each chromosome (UA_i) fits the Poisson distribution, as described earlier. Using 15,000 UA reads, the normalized ratio between a query sample and the reference ($Norm\_\%UA_i$) was determined by the copy number of chromosomes

$$Norm\_\%UA_i = \frac{Query\_\%UA_i}{Ref\_\%UA_i} = \frac{Query\_n_i \times Query\_\varphi_i}{Ref\_n_i \times Ref\_\varphi_i}$$

To address the change in coverage, ϕ due to loss or gain of chromosomes, the corrected normalized $\%UA_i$ equals

$$Norm'\_\%UA_i = \frac{Norm\_\%UA_i}{\overline{Norm\_\%UA_i'}}$$

where $\overline{Norm\_\%UA_i'}$ is the average $Norm\_\%UA_i$ of normal autosomes, as determined by *Z*-score. For an unknown sample, the SD of $Norm\_\%UA_i$ of normal autosomes (SD_{normal}) was estimated based on known normal autosomes (within $Ref\_UA_i \pm 3.29\sqrt{Ref\_UA_i}$) in this study ($n = 105$, $SD_{normal} = 0.0489$). The *Z*-score was calculated for each chromosome as

$$Z - score_i = \frac{Norm\_\%UA_i - Mean\_\%UA_{autosome}}{SD_{normal}}$$

Chromosomes having a $|Z - score| > 3.29$ were considered to be abnormal chromosomes with $P < 0.001$. When the *Z*-score was greater than 3.29, we consider there to be a gain of a chromosome, and when the *Z*-score was less than −3.29, we consider there to be a loss of a chromosome. While the modified *Z*-score method would be less specific in detecting abnormalities on small autosomes than the *Z*-score method based on a census of each chromosome (Chen *et al.* 2014), it provided sufficient power for aneuploidy detection (>95%) (Figure 2C). The theoretical value of a normal autosome $Norm'\_\%UA_{normal} = 1$, that of a full trisomy of an autosome $Norm'\_\%UA_{trisomy} = 1.5$, that of a monosomy of an autosome $Norm'\_\%UA_{monosomy} = 0.5$, that of the *X* chromosome of a normal female $Norm'\_\%UA_{X\_female} > 1.5$, and that of the *Y* chromosome of a normal female or missing *Y* chromosome $Norm'\_\%UA_{Y\_female} < 0.5$.

We hypothesized that the corrected normalized $\%UA_i$ ($Norm'\_\%UA_i$) reflects the copy number of chromosomes. The $Norm'\_\%UA_i$ values were used to compute the adjusted *Z*-score (*Z*′-score). The $Norm'\_\%UA_i$ values of normal autosomes with a $|Z - score| > 3.29$ were summarized ($Mean\_Norm'\_\%UA = 0.9999$, $SD\_Norm'\_\%UA = 0.0481$). The *Z*′-score for each chromosome equals

$$Z' - score_i = \frac{Norm'\_\%UA_i - Mean\_Norm'\_\%UA}{SD\_Norm'\_\%UA}$$

In brief, 15,000 UA matches were randomly selected from the normal male sample—and this was repeated for a total of 30 times—and averaged for normalization purposes (Ref_UA). For each sample, the first 15,000 UA matches (Query_UA) were selected for gender determination and aneuploidy detection. The UA matches were summarized and counted for each chromosome (UA_i, i = 1, 2, . . ., *X*, *Y*), and corresponding percentages were calculated for each chromosome ($\%UA_i$) by $UA_i/15{,}000 \times 100$. The $\%UA_i$ value for each of the chromosomes of a query sample ($Query\_\%UA_i$) was normalized to the normal male reference ($Ref\_\%UA_i$) and corrected to detect the copy number of each chromosome ($Norm'\_\%UA_i$) (Table S7 and Figure 2A).

Results and Discussion

To maintain equivalent molar concentrations for short-DNA-fragment-length library preparations compared with long-fragment-length preparations, ~18-fold lower total nanograms of input DNA and improved ligation efficiency were required (Figure 1B, Table S5). We systematically modified the

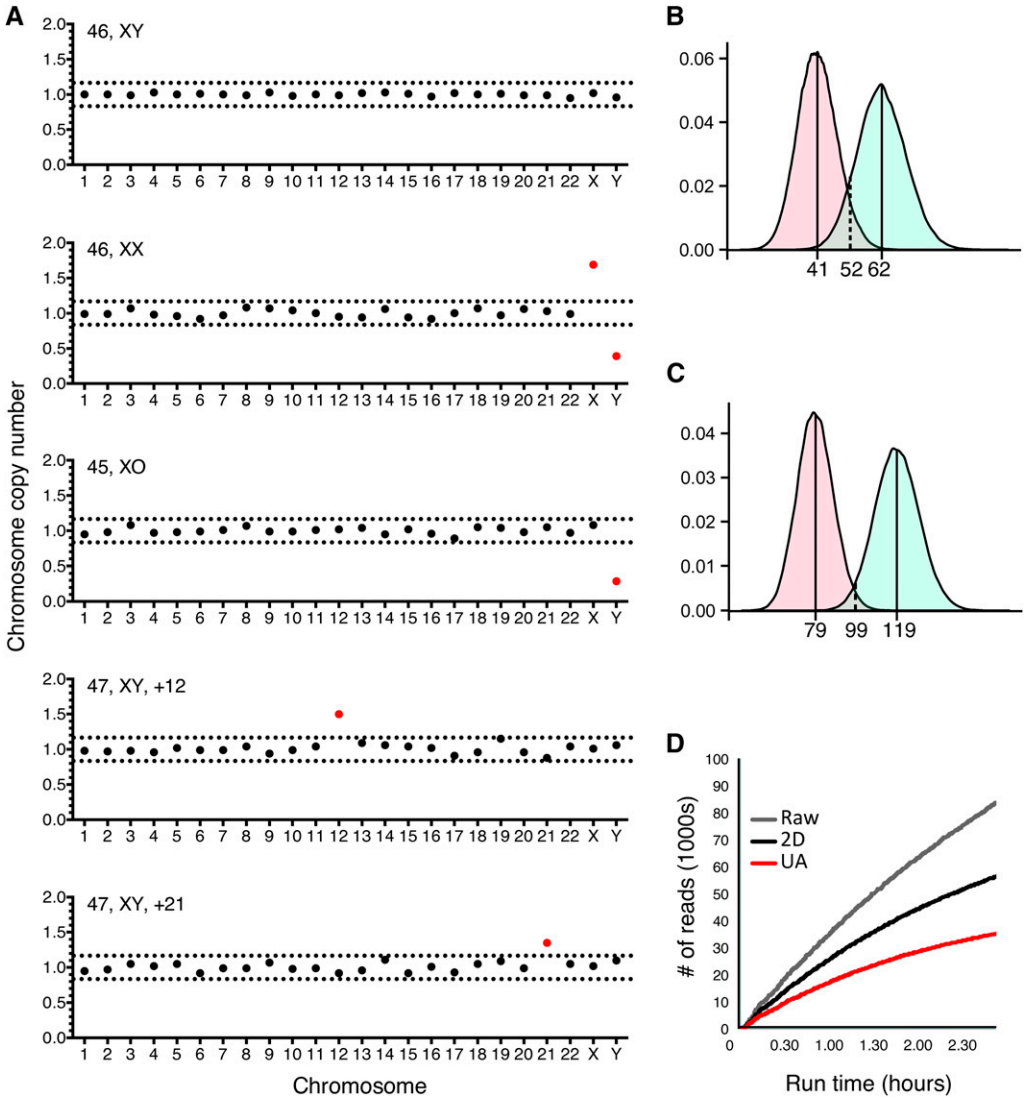


Figure 2 Cytogenetic analysis of gDNA using short-DNA-fragment MinION sequencing. (A) Short-DNA-fragment sequencing using MinION was able to correctly determine gender and detect aneuploidy in DNA samples from a normal male and female, a female with monosomy *X*, a male with trisomy 12, and a male with trisomy 21 ($P < 0.001$). The copy number of each chromosome was reflected by the corrected normalized percentage of UA matches ($Norm'\_\%UA_i$). Black dots represent chromosomes without significant copy number changes; red dots represent chromosomes with significant copy number changes compared with a normal male reference; dotted line represent 99.9% confidence intervals. (B) Theoretical lower UA required for aneuploidy detection under the Poisson distribution. When $\lambda = 41$, $p(x > 1.5\lambda) = 0.0008$ and $p_\beta(x' < 1.25\lambda) = 0.10$. (C) Theoretical lower detection power using the 15,000 reference under the Poisson distribution. The *Y* chromosome has the fewest UA matches assigned (79–80). When $\lambda = 79$, $p(x > 1.5\lambda) = 1.07 \times 10^{-5}$ and $p_\beta(x' < 1.25\lambda) = 0.034$. (D) Sequencing yield of a short-fragment library across time showing raw reads (gray line), 2D reads (black line), and reads uniquely aligned to the Hg19 reference genome (red).

protocol to improve ligation efficiency. To monitor ligation reactions, a 434-bp PCR product and a 57-bp control adapter duplex with a T overhang were used (Table S1). The intermediate ligation product with one adapter is ~490 bp, and the final product with adapters on both ends is ~550 bp. Use of the manufacturer's protocol resulted in <5% final products (Figure 1B, lane 2). By purifying dA-tailed DNA prior to ligation, the percentage of final products increased to 25% (Figure 1B, lane 3). Reducing reaction volumes from 100 to 20 μl further increased the percentage of final products to 48% (Figure 1B, lane 4). By combining a 10-min room temperature incubation with a 1- to 2-hr incubation at 4°, we were able to increase the percentage of final products to 61–63% (Figure 1B, lanes 5–7) without releasing the preattached E5 protein. Thus, by purifying and then concentrating dA-tailed DNA to reduce the reaction volume and by introducing a prolonged ligation at 4°, the ligation efficiency of the final products was increased from <5 to 63% (Figure 1B, lane 2 *vs.* lanes 6 and 7, Table S5) and provided sufficient material for downstream His-tag bead purification (Table S2).

To determine the optimal tool for data analysis of the increased number of reads obtained with sequencing of short DNA, we compared LAST—an alignment program recommended by MAP—with two similar programs, Bowtie 2 and Blat (Kent 2002; Frith *et al.* 2010; Langmead and Salzberg 2012) using a training library generated through a MinION short-DNA-sequencing run (Table S3). While Bowtie 2 and LAST completed alignments more quickly (1 and 14 min, respectively) than Blat (68 min), Blat generated more good alignments (65%) compared with Bowtie 2 and LAST (58 and 61%, respectively) for the same data sets, likely due to the tendency for MinION sequencing errors resulting in deletions (Kent 2002; Langmead and Salzberg 2012; Jain *et al.* 2015) (Table S2, Table S3, and Figure S3). Blat also generated more UA matches (62%) than Bowtie 2 and LAST (45 and 55%, respectively). Blat was used for alignment of the MinION short-DNA-sequencing results to provide the most

inclusive alignment results. Given sufficient computational resources on a high-performance server, increasing parallel threats can further reduce the run time.

To demonstrate the clinical utility of nanopore-based sequencing of short DNA fragments, we ran a pilot study testing the ability of this approach to diagnose aneuploidy. Fetal aneuploidy testing is routinely performed as a component of prenatal testing [*e.g.*, amniocentesis and chorionic villus sampling (CVS)], preimplantation genetic screening (PGS) of embryos in *in vitro* fertilization (IVF), and evaluation of miscarriage tissue (Brezina *et al.* 2012). A rapid diagnosis is clinically vital to enable timely management. In the case of prenatal samples obtained via an amniocentesis or CVS, rapid results will enable treatment before the pregnancy progresses to a more advanced gestational age, when treatment options are more limited, technically difficult, and dangerous to the mother. In the case of PGS, rapid testing will enable transfer of the embryo in a given IVF cycle without the need to freeze embryos. However, standard methods to diagnose aneuploidy, such as karyotyping and microarray analysis, take 7–21 days to complete. ULCS for the detection of aneuploidy is a new strategy for whole-genome aneuploidy detection that requires alignment of reads to a reference genome assembly but still requires 15–21 hr to complete and costly and technically advanced library preparation and sequencing platforms that cannot be readily used in a physician's office or in low-complexity settings (Chen *et al.* 2014; Wells *et al.* 2014). The ULCS approach for determining aneuploidy requires only that the reads are sufficiently long to enable unique alignment to the genome. Thus, a method to rapidly sequence large numbers of short DNA fragments in real time would enable rapid diagnosis of aneuploidy in settings outside an advanced laboratory facility.

Purified gDNA samples from a normal male and female, a male with trisomy 12, a male with trisomy 21, and a female with monosomy *X* were fragmented, size selected (350–600 bp), and processed as described earlier (Table S2). Sequencing short-DNA-fragment libraries prepared using our protocol with MinION generated ~500 unique reads after the first 3 min of sequencing and 43,000–87,000 raw reads and 27,000–58,000 2D reads (32–67%) after 4 hr of sequencing (Figure 2 and Table S4). This compares favorably with the traditional MinION sequencing protocol, which sequenced <12,000 reads after 36 hr (Quick *et al.* 2014; Kilianski *et al.* 2015). Noninvasive prenatal testing (NIPT) is another potential application of the nanopore sequencing technology that would benefit from our short-DNA-sequencing method. NIPT testing on the MinION nanopore sequencer using the standard protocol took 6–24 hr to generate up to 56,000 2D reads, and 15.6–23.9% of the 2D reads were UA reads (Cheng *et al.* 2015). In contrast, our assay generated up to 58,000 2D reads in <4 hr, and 39.1–69.7% of the 2D reads were UA reads. Our protocol and data analysis can be very helpful in improving NIPT analysis, including the run time and detection sensitivity (Table S4).

Using the short-fragment-length DNA sequencing library preparation and analysis pipeline, we obtained sufficient numbers of reads for successful determination of gender and aneuploidy ($P < 0.001$) in all samples within 2–4 hr (Figure 2A). Using a normal approximation of the Poisson distribution, the chance of a type II error for detecting aneuploidy (p_β-aneuploidy) was <0.05 (Figure 2C and Table S7). Because MinION is easily scalable, cytogenetic analysis can be done within 1–2 hr by running two MinION sequencers in parallel. MinION Flow Cells currently cost US$500–900, although efficiency improvements and multiplexing are likely to bring these costs down in the future. Furthermore, as sequencing chemistry and platforms are being actively improved, faster sequencing speeds (70 *vs.* 30 nt/sec) and platforms with higher throughput in the near future, such as the MinION MkII (4096 channels) and the PromethION (262,176 channels), will enable even faster short-read sequencing with higher coverage.

In summary, in addition to the intended role of MinION for sequencing long fragments of DNA, our results demonstrate that MinION also can be used for very rapid real-time acquisition of short DNA reads that can be used for time-sensitive aneuploidy detection in clinical care as well as sequencing of small DNA fragments and amplicons in the field or clinic. This ability can expand the utility of the MinION into new clinical and research applications.

Acknowledgments

We thank Thomas Tuschl, Jan Vijg, Steven Josefowicz, Barry Coller, and members of the Williams laboratory for their helpful input and advice with this project and manuscript. This research was supported by National Institutes of Health grants HD-068546 and U19-CA-179564.

Author contributions: Z.W. conceived of the project. S.W. performed the experiments and prepared figures. Z.W. and S.W. developed the method, analyzed the data, and wrote the manuscript.

Literature Cited

Ashton, P. M., S. Nair, T. Dallman, S. Rubino, W. Rabsch *et al.*, 2015 MinION nanopore sequencing identifies the position and structure of a bacterial antibiotic resistance island. Nat. Biotechnol. 33: 296–300.

Brezina, P. R., D. S. Brezina, and W. G. Kearns, 2012 Preimplantation genetic testing. BMJ 345: e5908.

Chen, S., S. Li, W. Xie, X. Li, C. Zhang *et al.*, 2014 Performance comparison between rapid sequencing platforms for ultra-low coverage sequencing strategy. PLoS One 9: e92192.

Cheng, S. H., P. Jiang, K. Sun, Y. K. Cheng, K. C. Chan *et al.*, 2015 Noninvasive prenatal testing by nanopore sequencing of maternal plasma DNA: feasibility assessment. Clin. Chem. 61: 1305–1306.

Cherf, G. M., K. R. Lieberman, H. Rashid, C. E. Lam, K. Karplus *et al.*, 2012 Automated forward and reverse ratcheting of DNA in a nanopore at 5-A precision. Nat. Biotechnol. 30: 344–348.

Dillies, M. A., A. Rau, J. Aubert, C. Hennequet-Antier, M. Jeanmougin *et al.*, 2013 A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. Brief. Bioinform. 14: 671–683.

Fan, H. C., and S. R. Quake, 2010 Sensitivity of noninvasive prenatal detection of fetal aneuploidy from maternal plasma using shotgun sequencing is limited only by counting statistics. PLoS One 5: e10439.

Frith, M. C., R. Wan, and P. Horton, 2010 Incorporating sequence quality data into alignment improves DNA read mapping. Nucleic Acids Res. 38: e100.

Human Microbiome Project Consortium, 2012 A framework for human microbiome research. Nature 486: 215–221.

Jain, M., I. T. Fiddes, K. H. Miga, H. E. Olsen, B. Paten *et al.*, 2015 Improved data analysis for the MinION nanopore sequencer. Nat. Methods 12: 351–356.

Kent, W. J., 2002 BLAT—the BLAST-like alignment tool. Genome Res. 12: 656–664.

Kilianski, A., J. L. Haas, E. J. Corriveau, A. T. Liem, K. L. Willis *et al.*, 2015 Bacterial and viral identification and differentiation by amplicon sequencing on the MinION nanopore sequencer. Gigascience 4: 12.

Langmead, B., and S. L. Salzberg, 2012 Fast gapped-read alignment with Bowtie 2. Nat. Methods 9: 357–359.

Loman, N. J., and A. R. Quinlan, 2014 Poretools: a toolkit for analyzing nanopore sequence data. Bioinformatics 30: 3399–3401.

Loman, N. J., and M. Watson, 2015 Successful test launch for nanopore sequencing. Nat. Methods 12: 303–304.

Martin, M., 2011 Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.Journal 17: 10–12.

Miller, F. J., F. L. Rosenfeldt, C. Zhang, A. W. Linnane, and P. Nagley, 2003 Precise determination of mitochondrial DNA copy number in human skeletal and cardiac muscle by a PCR-based assay: lack of change of copy number with age. Nucleic Acids Res. 31: e61.

Palomaki, G. E., C. Deciu, E. M. Kloza, G. M. Lambert-Messerlian, J. E. Haddow *et al.*, 2012 DNA sequencing of maternal plasma reliably identifies trisomy 18 and trisomy 13 as well as Down syndrome: an international collaborative study. Genet. Med. 14: 296–305.

Quick, J., A. R. Quinlan, and N. J. Loman, 2014 A reference bacterial genome dataset generated on the MinION portable single-molecule nanopore sequencer. Gigascience 3: 22 [erratum: Gigascience 4: 6 (2015)].

R Core Team, 2012 *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Urban, J. M., J. Bliss, C. E. Lawrence, and S. A. Gerbi, 2015 Sequencing ultra-long DNA molecules with the Oxford Nanopore MinION. *bioRxiv* .doi.org/10.1101/019281

Wells, D., K. Kaur, J. Grifo, M. Glassner, J. C. Taylor *et al.*, 2014 Clinical utilisation of a rapid low-pass whole genome sequencing technique for the diagnosis of aneuploidy in human embryos prior to implantation. J. Med. Genet. 51: 553–562.

Communicating editor: J. Shendure

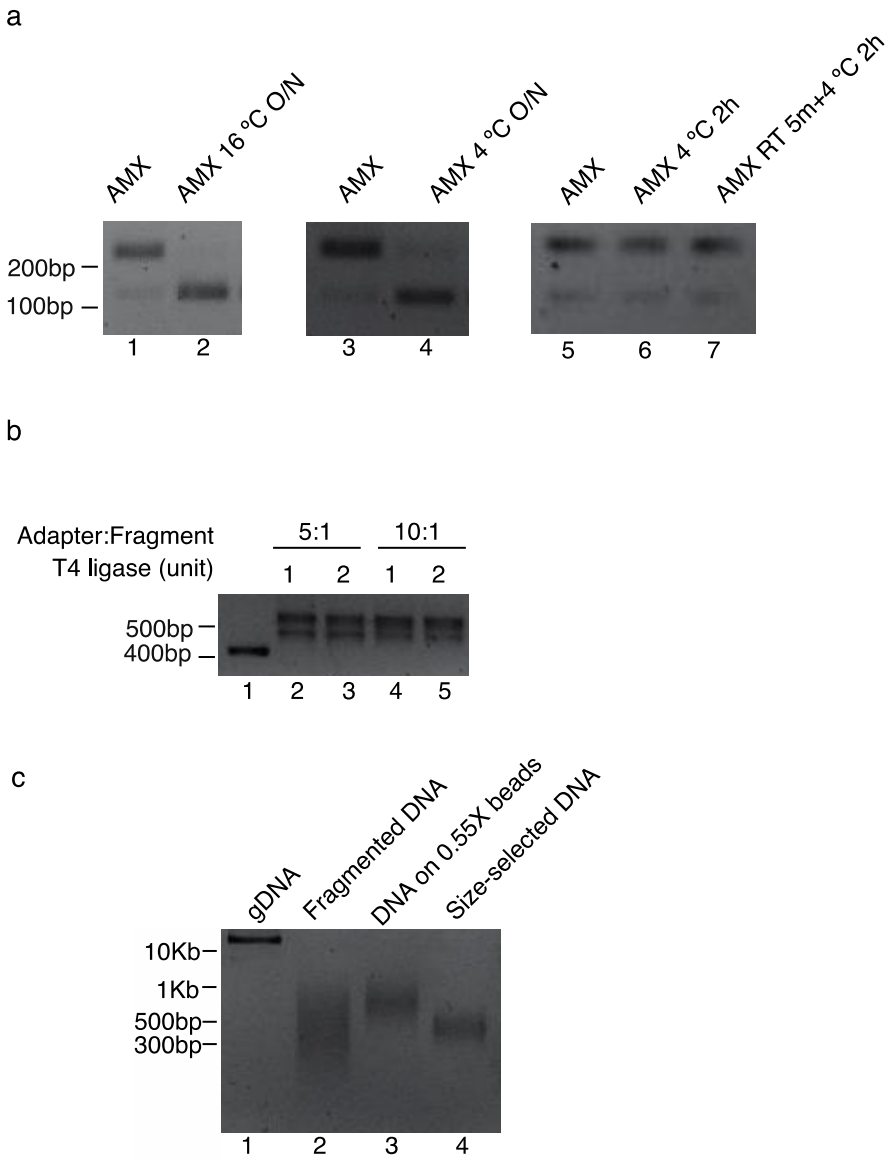
GENETICS

Supporting Information
www.genetics.org/lookup/suppl/doi:10.1534/genetics.115.182311/-/DC1

Rapid Short-Read Sequencing and Aneuploidy Detection Using MinION Nanopore Technology

Shan Wei and Zev Williams

DOI: 10.1534/genetics.115.182311

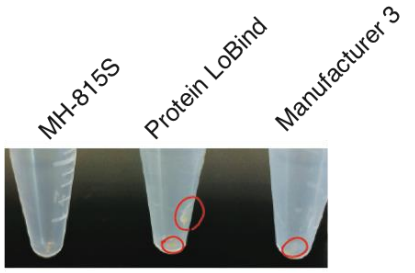


Supplementary Figure 1.

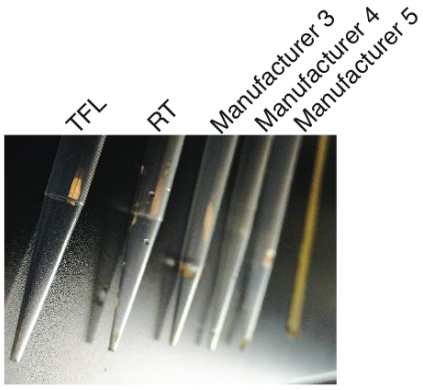
Figure S1 Development of MinION™ short-fragment library preparation method. (a) Evaluation of E5 detachment from Y-adapter in the adapter mixture. Gels showing levels of

detachment of E5 protein from the adapter mixture provided in kit (SQK-MAP004) under different conditions. Lane 1, 3, 5 show the untreated adapter mixture. Lane 2 and 4 show the results of adapter mixture incubated in T4 ligase buffer at 16 °C or 4 °C O/N respectively. Lane 6 and 7 show the results of adapter mix incubated in blunt/TA ligase master mix with extended 4 h incubation at 4 °C. Incubation of the adapter mixture at 16 °C or 4 °C O/N in 1X T4 ligation buffer results in detachment of the E5 protein (Lane 1 vs 2 and Lane 3 vs 4, respectively). Incubation of the adapter mixture in 1X blunt/TA ligase master mix at 4 °C for 2 h or RT 5min followed by 4 °C 2 h did not result in detachment of E5 protein (Lane 5 vs. Lane 6 and 7 respectively). (b) Ligation in 20 µL T4 ligation system. Gels showing ligation products using 20 µL T4 ligation system overnight incubation at 16 °C. 16 °C O/N incubation of adapter:fragment at 10:1 ratio results in ~75% control fragments ligated to adapters on both ends (Lane 4, 5 vs lane 1); 16 °C O/N incubation of adapter:fragment at 5:1 ratio results in ~65% control fragments ligated to adapters on both ends (Lane 2, 3 vs lane 1). Using the adapter:fragment at a 5:1 ratio enables 2 ligation reactions to be performed using the same kit reagents, resulting in ~130% (i.e., ~ 0.26 pmol) fragments with adapters on both ends for downstream experiments (Lane 1, control fragment; lane 2-3, ligate control adapter and fragment at a 1/2 ratio; lane 4-5, ligate control adapter and fragment at the ratio suggested in manufacturer's protocol.) (c) Fragmentation and size-selection of a MinION™ short-fragment library preparation. (Lane 1, gDNA (> 15kb); lane 2, fragmented gDNA; lane 3, large fragments (> 650bp) discarded during size-selection; lane 4, size-selected DNA fragments (350-600bp)).

a



b



Supplementary Figure 2.

Figure S2. Retention properties of plastic consumables. (a) Retention on tubes. The retention of His-Tag beads and wash buffer were tested on 4 low retention microcentrifuge tubes and 5 pipet tips including the kit recommended Protein LoBind tube and RT-1000 tip. After pipetting for 16 times and removal of the 250 μL washed His-Tag beads + 1/2 wash buffer, ~ 10 μL remnant. The VWR LR tube had seen less remnant, but still visible. The MH-815S was used for SRL library preparation unless stated. (b) Retention on tips. The tips were examined by pipetting the His-Tag beads purification reagents for 16 times. All the low retention tips seen different level of droplet retention, but overall the TFL tip are the cleanest with a tiny droplet at the point of the tip.

Supplementary Figure 3

Figure S3. Alignments of a 2D, template, read and complement reads. Alignments were performed using the USCS blat aligner (https://genome.ucsc.edu/). Color blocks represent alignments, while the gray scale of color blocks represent alignment identities. Red vertical lines within color blocks indicate mismatches. Two gray lines between color blocks indicate gaps. A 2D read has smaller gaps between alignment blocks and higher alignment identifies when comparing to the template and complementary reads.

Table S1. Sequence information

Control fragment sequence, 434bp
5'-
CAGGAAACAGCTATGACCATGATTACGCCAAGCTATTTAGGTGACGCGTTAGAATACTCAAGCTATGCATCAAGCTTGGTACCGAGCTCGGATCCACTAGTAACG
GCCGCCAGTGTGCTGGAATTCAGGCAAGCAGAAGACGGCATACGAGATCGTGATGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCTGCACAATGTGCACAT
GTACCCTAAAACTTAGAGTATAATAAAAATAAAAAATAAAAAAAGAAGTCCAAAAAAAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGT
CGCCGTATCATTCCTGAATTCTGCAGATATCCATCACACTGGCGGCCGCTCGAGCATGCATCTAGAGGGCCCAATTCGCCCTATAGTGAGTCGTATTACAATTCAC
TGGCCGTCGTTTTAC

M13F (-20) primer
5'- GTAAAACGACGGCCAG

M13R primer
5'- CAGGAAACAGCTATGAC

Control adaptor
Top: 5'- GGAAGCTTGACATTCTGGATCGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTT
Bottom 5'- **P**-AGATCGGAAGAGCACACGTCT-**AMINE 3'**

Table S2. MinION™ library preparation.

Sample	Karyotype	gDNA for fragments (ng)	Fragmented DNA for size selection (ng)	Purified size selected DNA (ng)	Purified End-Repaired DNA (ng)	DNA Input for dA-tail (ng)**	Purified dA-tailed DNA (ng)**	Purified dA-tailed DNA (ng)	dA-tailed DNA for ligation (ng)	DNA in pre-seq library (ng)**	% of dA-tailed DNA	Fold change of DNA in pre-seq library
1a*	46, XY	3000	1200**	533.8**	109.5**	91.25	N/A	N/A	91.25**	9.55	10.47	1.00
1b	46, XY	3000	2287	549.5	312.4	234.3	N/A	340.68	142.40	25.52	17.92	2.67
2	47, XY, +12	2251.2	2179	534.5	501.6	376.2	336	396.66	144.00	31.86	22.13	3.34
3	45, XO	3032.4	2494	465	272.8	204.6	178.56	216.48	144.32	31.08	21.54	3.25
4a	46, XX	2880	2923	451.5	462	346.5	331.2	390.12	144.00	37.52	26.06	3.93
4b	46, XX	3212.4	2716	452	411.4	308.55	266.4	298.20	144.00	38.08	26.44	3.99
5	47, XY, +21	2026.8	1740	205	367.4	275.55	242.4	258.48	144.00	36.68	25.47	3.84

* This sample was size-selected using 0.6-0.85X SPRIselect, 68ng of size-selected DNA was added to the end repair reaction.
** Measured by Qubit dsDNA HS assay kit.

Table S3. Software comparison.

Software	Parameter	Parallel	Run time (min)	2D	GA*	%GA	UA**	%UA
Bowtie2	--rdg 1,2 -D 25 -R 5 -N 1 -L 20 -i S,1,0.3 -k 5 --score-min L,0,-1.5	8	1	58,199	33,958	58.35	26,148	44.93
LAST	-s 2 -T 0 -Q 0 -r 1 -a 1 -b 1 -q 1	16	14	58,199	35,783	61.48	32,297	55.49
BLAT	-minIdentity=80 -minScore=40 -tileSize=10 -maxIntron=500	20	68	58,199	37,614	64.63	36,041	61.93

* GA, good alignment. In Bowtie2, GA is defined as alignment meets the minimum score as computed by Bowtie2. In LAST and and BLAT, GA is defined as alignments pass pslReps filter with the setting -minCover=0.40 -minAli=0.80 -nearTop=0.001 –singleHit.

** UA, unique alignment. In Bowtie2, UA is defined as alignment aligned exactly once as computed by Bowtie2. In LAST and BLAT, UA is defined as reads with only one GA passing the pslReps filter.

Table S4. MinION™ run summary.

Sample	Karyotype	Pore available	Pore used	% Pore used	Total reads	2D reads	% 2D reads	GA reads	% GA	UA reads	% UA reads	Run time	Time for 15K UA
1a*	46, XY	319	125	39.2	2,860	-	-	-	-	-	-	1:00	N/A
1b*	46, XY	342	325	95.0	120,683	38,640	32.02	18,357	47.5	17,939	46.4	7:42**	3:24
2	47, XY, +12	353	329	93.2	82,153	45,061	54.85	18,167	40.3	17,629	39.1	3:53	2:23
3	45, XO	288	202	70.1	55,146	36,687	66.53	20,583	56.1	19,984	54.5	4:21	2:18
4a	46, XX	457	389	85.1	61,282	36,375	59.36	24,520	67.4	23,809	65.5	3:54	1:37
4b	46, XX	225	216	96.0	43,574	27,883	63.99	20,240	72.6	19,424	69.7	6:31	2:44
5	47, XY, +21	472	466	98.7	87,006	58,189	66.88	37,614	64.6	36,041	61.9	2:57	0:53

* Samples were sequenced with R7.3 before platform upgrade.
** A 7h56min break was in between due to software upgrade.

Table S5. Gel densitometry analysis of figure 1b.

Lane	Band	Size (bp)	Mean (%)	SD
2	1	~550	2.7	1.3
	2	~490	47.6	2.6
	3	434	49.7	3.2
3	1	~550	25.1	4.5
	2	~490	60.0	6.1
	3	434	14.9	3.3
4	1	~550	48.2	1.5
	2	~490	50.4	0.9
	3	434	1.5	1.6
5	1	~550	60.8	4.5
	2	~490	39.1	4.5
	3	434	0.1	0.3
6	1	~550	63.0	4.2
	2	~490	36.9	4.2
	3	434	0.1	0.2
7	1	~550	63.5	4.7
	2	~490	36.4	4.6
	3	434	0.1	0.2

Table S6. Comparison of the15K normal male reference and the GRCh37 human reference genome.

Chromosome	Total length (bp)	Ungapped length (UL) (bp)	%UL	Ref_UA	Ref_%UA	Norm_Ref_%UA
1	249,250,621	225,280,621	7.87	1280	8.53	1.08
2	243,199,373	238,204,522	8.32	1346	8.97	1.08
3	198,022,430	194,797,140	6.81	962	6.41	0.94
4	191,154,276	187,661,676	6.56	960	6.40	0.98
5	180,915,260	177,695,260	6.21	909	6.06	0.98
6	171,115,067	167,395,067	5.85	895	5.97	1.02
7	159,138,663	155,353,663	5.43	825	5.50	1.01
8	146,364,022	142,888,922	4.99	715	4.76	0.95
9	141,213,431	120,143,431	4.20	643	4.28	1.02
10	135,534,747	131,314,747	4.59	759	5.06	1.10
11	135,006,516	131,129,516	4.58	720	4.80	1.05
12	133,851,895	130,481,395	4.56	698	4.66	1.02
13	115,169,878	95,589,878	3.34	475	3.17	0.95
14	107,349,540	88,289,540	3.09	452	3.01	0.98
15	102,531,392	81,694,769	2.86	444	2.96	1.04
16	90,354,753	78,884,753	2.76	477	3.18	1.15
17	81,195,210	77,795,210	2.72	476	3.17	1.17
18	78,077,248	74,657,248	2.61	425	2.83	1.08
19	59,128,983	55,808,983	1.95	290	1.93	0.99
20	63,025,520	59,505,520	2.08	347	2.32	1.11
21	48,129,895	35,106,642	1.23	213	1.42	1.16
22	51,304,566	34,894,566	1.22	190	1.27	1.04
X	155,270,560	151,100,560	5.28	379	2.53	0.48
Y	59,373,566	25,653,566	0.90	80	0.53	0.60
MT	16,569	16,569	0.00	40	0.27	460.51

Table S7. ULCS cytogenetics analysis.

Sample	Chromo-some	UA	%UA	Norm_ %UA	Z-score	p<0.001	Norm' _%UA	Z'-score	p<0.001
46, XY	1	1283	8.55	1.00	0.08	FALSE	1.00	0.08	FALSE
	2	1345	8.97	1.00	0.02	FALSE	1.00	0.02	FALSE
	3	951	6.34	0.99	-0.20	FALSE	0.99	-0.20	FALSE
	4	983	6.55	1.02	0.52	FALSE	1.03	0.54	FALSE
	5	905	6.03	1.00	-0.04	FALSE	1.00	-0.04	FALSE
	6	901	6.01	1.01	0.16	FALSE	1.01	0.17	FALSE
	7	823	5.49	1.00	-0.02	FALSE	1.00	-0.02	FALSE
	8	704	4.69	0.99	-0.26	FALSE	0.99	-0.27	FALSE
	9	661	4.41	1.03	0.62	FALSE	1.03	0.63	FALSE
	10	741	4.94	0.98	-0.44	FALSE	0.98	-0.45	FALSE
	11	721	4.81	1.00	0.06	FALSE	1.00	0.07	FALSE
	12	687	4.58	0.98	-0.29	FALSE	0.99	-0.30	FALSE
	13	484	3.23	1.02	0.43	FALSE	1.02	0.44	FALSE
	14	462	3.08	1.02	0.51	FALSE	1.03	0.52	FALSE
	15	447	2.98	1.01	0.16	FALSE	1.01	0.16	FALSE
	16	463	3.09	0.97	-0.56	FALSE	0.97	-0.57	FALSE
	17	483	3.22	1.02	0.35	FALSE	1.02	0.36	FALSE
	18	425	2.83	1.00	0.06	FALSE	1.00	0.06	FALSE
	19	292	1.95	1.01	0.16	FALSE	1.01	0.17	FALSE
	20	342	2.28	0.98	-0.27	FALSE	0.99	-0.28	FALSE
	21	211	1.41	0.99	-0.13	FALSE	0.99	-0.13	FALSE
	22	181	1.21	0.95	-0.92	FALSE	0.95	-0.94	FALSE
	X	387	2.58	1.02	0.45	FALSE	1.02	0.46	FALSE
	Y	77	0.51	0.96	-0.74	FALSE	0.96	-0.75	FALSE
46, XX	1	1248	8.32	0.97	-0.27	FALSE	0.99	-0.28	FALSE
	2	1315	8.77	0.98	-0.23	FALSE	0.99	-0.23	FALSE
	3	1015	6.77	1.06	1.37	FALSE	1.07	1.41	FALSE
	4	927	6.18	0.97	-0.46	FALSE	0.98	-0.47	FALSE
	5	863	5.75	0.95	-0.79	FALSE	0.96	-0.81	FALSE
	6	818	5.45	0.91	-1.53	FALSE	0.92	-1.57	FALSE
	7	790	5.27	0.96	-0.63	FALSE	0.97	-0.65	FALSE
	8	766	5.11	1.07	1.71	FALSE	1.08	1.77	FALSE
	9	677	4.51	1.05	1.33	FALSE	1.07	1.37	FALSE
	10	776	5.17	1.02	0.71	FALSE	1.04	0.73	FALSE
	11	713	4.75	0.99	0.04	FALSE	1.00	0.04	FALSE
	12	658	4.39	0.94	-0.94	FALSE	0.95	-0.96	FALSE
	13	443	2.95	0.93	-1.13	FALSE	0.94	-1.16	FALSE
	14	474	3.16	1.05	1.26	FALSE	1.06	1.30	FALSE
	15	411	2.74	0.92	-1.30	FALSE	0.94	-1.33	FALSE
	16	432	2.88	0.91	-1.69	FALSE	0.92	-1.74	FALSE
	17	471	3.14	0.99	0.04	FALSE	1.00	0.04	FALSE
	18	451	3.01	1.06	1.51	FALSE	1.07	1.56	FALSE
	19	279	1.86	0.96	-0.55	FALSE	0.97	-0.56	FALSE
	20	365	2.43	1.05	1.28	FALSE	1.06	1.32	FALSE
	21	216	1.44	1.02	0.56	FALSE	1.03	0.57	FALSE

	22	185	1.23	0.97	-0.29	FALSE	0.99	-0.30	FALSE
	X	635	4.23	1.67	14.02	TRUE	1.69	14.42	TRUE
	Y	31	0.21	0.39	-12.29	TRUE	0.39	-12.64	TRUE
	1	1221	8.14	0.95	-1.04	FALSE	0.95	-1.06	FALSE
	2	1332	8.88	0.99	-0.31	FALSE	0.98	-0.32	FALSE
	3	1043	6.95	1.08	1.63	FALSE	1.08	1.64	FALSE
	4	936	6.24	0.97	-0.61	FALSE	0.97	-0.62	FALSE
	5	897	5.98	0.99	-0.36	FALSE	0.98	-0.37	FALSE
	6	895	5.97	1.00	-0.11	FALSE	0.99	-0.11	FALSE
	7	836	5.57	1.01	0.17	FALSE	1.01	0.17	FALSE
	8	768	5.12	1.07	1.43	FALSE	1.07	1.45	FALSE
	9	637	4.25	0.99	-0.28	FALSE	0.99	-0.28	FALSE
45, XO	10	753	5.02	0.99	-0.25	FALSE	0.99	-0.26	FALSE
	11	731	4.87	1.02	0.21	FALSE	1.01	0.21	FALSE
	12	716	4.77	1.03	0.42	FALSE	1.02	0.42	FALSE
	13	495	3.30	1.04	0.77	FALSE	1.04	0.78	FALSE
	14	432	2.88	0.96	-0.98	FALSE	0.95	-1.00	FALSE
	15	454	3.03	1.02	0.35	FALSE	1.02	0.35	FALSE
	16	462	3.08	0.97	-0.74	FALSE	0.96	-0.75	FALSE
	17	426	2.84	0.90	-2.23	FALSE	0.89	-2.26	FALSE
	18	450	3.00	1.06	1.13	FALSE	1.05	1.13	FALSE
	19	303	2.02	1.04	0.80	FALSE	1.04	0.81	FALSE
	20	341	2.27	0.98	-0.47	FALSE	0.98	-0.47	FALSE
	21	225	1.50	1.06	1.08	FALSE	1.05	1.09	FALSE
	22	185	1.23	0.97	-0.63	FALSE	0.97	-0.64	FALSE
	X	412	2.75	1.09	1.66	FALSE	1.08	1.67	FALSE
	Y	23	0.15	0.29	-14.67	TRUE	0.29	-14.84	TRUE
	1	1230	8.20	0.96	-0.91	FALSE	0.98	-0.46	FALSE
	2	1287	8.58	0.96	-1.00	FALSE	0.97	-0.56	FALSE
	3	923	6.15	0.96	-0.93	FALSE	0.98	-0.49	FALSE
	4	910	6.07	0.95	-1.17	FALSE	0.96	-0.74	FALSE
	5	911	6.07	1.00	-0.05	FALSE	1.02	0.42	FALSE
	6	873	5.82	0.97	-0.62	FALSE	0.99	-0.16	FALSE
	7	803	5.35	0.97	-0.66	FALSE	0.99	-0.20	FALSE
	8	732	4.88	1.02	0.39	FALSE	1.04	0.88	FALSE
	9	595	3.97	0.93	-1.62	FALSE	0.94	-1.20	FALSE
47, XY, +12	10	735	4.90	0.97	-0.74	FALSE	0.99	-0.29	FALSE
	11	734	4.89	1.02	0.29	FALSE	1.04	0.78	FALSE
	12	1030	6.87	1.48	9.61	TRUE	1.50	10.41	TRUE
	13	510	3.40	1.07	1.41	FALSE	1.09	1.93	FALSE
	14	470	3.13	1.04	0.73	FALSE	1.06	1.23	FALSE
	15	454	3.03	1.02	0.34	FALSE	1.04	0.82	FALSE
	16	476	3.17	1.00	-0.15	FALSE	1.02	0.32	FALSE
	17	424	2.83	0.89	-2.33	FALSE	0.91	-1.93	FALSE
	18	400	2.67	0.94	-1.29	FALSE	0.96	-0.86	FALSE
	19	328	2.19	1.13	2.56	FALSE	1.15	3.12	FALSE
	20	327	2.18	0.94	-1.30	FALSE	0.96	-0.87	FALSE
	21	184	1.23	0.86	-2.87	FALSE	0.88	-2.49	FALSE

	22	194	1.29	1.02	0.33	FALSE	1.04	0.82	FALSE
	X	377	2.51	0.99	-0.24	FALSE	1.01	0.23	FALSE
	Y	83	0.55	1.04	0.65	FALSE	1.06	1.15	FALSE
	1	1217	8.11	0.95	-1.37	FALSE	0.95	-1.06	FALSE
	2	1302	8.68	0.97	-1.03	FALSE	0.97	-0.72	FALSE
	3	1009	6.73	1.05	0.63	FALSE	1.05	0.97	FALSE
	4	985	6.57	1.03	0.17	FALSE	1.02	0.50	FALSE
	5	954	6.36	1.05	0.66	FALSE	1.05	1.00	FALSE
	6	828	5.52	0.92	-1.90	FALSE	0.92	-1.60	FALSE
	7	820	5.47	0.99	-0.49	FALSE	0.99	-0.17	FALSE
	8	707	4.71	0.99	-0.58	FALSE	0.99	-0.26	FALSE
	9	690	4.60	1.07	1.14	FALSE	1.07	1.49	FALSE
47, XY, +21	10	744	4.96	0.98	-0.76	FALSE	0.98	-0.44	FALSE
	11	717	4.78	1.00	-0.45	FALSE	0.99	-0.13	FALSE
	12	645	4.30	0.92	-1.93	FALSE	0.92	-1.62	FALSE
	13	455	3.03	0.96	-1.22	FALSE	0.96	-0.90	FALSE
	14	500	3.33	1.11	1.83	FALSE	1.11	2.19	FALSE
	15	409	2.73	0.92	-1.99	FALSE	0.92	-1.69	FALSE
	16	483	3.22	1.01	-0.11	FALSE	1.01	0.22	FALSE
	17	441	2.94	0.93	-1.86	FALSE	0.93	-1.55	FALSE
	18	448	2.99	1.06	0.76	FALSE	1.05	1.10	FALSE
	19	318	2.12	1.10	1.59	FALSE	1.09	1.95	FALSE
	20	344	2.29	0.99	-0.56	FALSE	0.99	-0.23	FALSE
	21	288	1.92	1.35	6.87	TRUE	1.35	7.30	TRUE
	22	199	1.33	1.05	0.61	FALSE	1.05	0.95	FALSE
	X	389	2.59	1.03	0.15	FALSE	1.02	0.48	FALSE
	Y	88	0.59	1.10	1.67	FALSE	1.10	2.03	FALSE