# SMURF-seq: efficient short-read sequencing on long-read sequencers

Rishvanth K. Prabakar

Quantitative and Computational Biology Section,
Department of Biological Sciences,
University of Southern California,
1050 Childs Way,
Los Angeles 90089, USA

April 21, 2020

# Abstract

Somatic copy number alterations (CNA) play a significant role cancer, and can be leveraged for diagnostic and personalized approaches to treatment. High-throughput short-read sequencing has been extremely efficient in copy number profiling; however, its applicability depends on the availability of instrument, and time to obtain profiles can vary from a few days to weeks. We present SMURF-seq, a protocol to efficiently sequence short DNA molecules on a long-read sequencer by randomly ligating them to form long molecules. Applying SMURF-seq using the highly portable and inexpensive Oxford Nanopore MinION yields up to 30 fragments per read, providing an average of 6.2 and up to 7.5 million mappable fragments per run, increasing information throughput for read-counting applications. We apply SMURF-seq on the MinION to generate copy number profiles, demonstrate that multiple samples can be multiplexed and sequenced in a single sequencing run, and show that concordant profiles are obtained with reads sequenced in the first 45 minutes of a run. A comparison with profiles from Illumina sequencing reveals that SMURF-seq attains similar accuracy. More broadly, with a fast and simple preparation method and a turnaround time measured in hours, the SMURF-seq approach could provide a highly efficient methodology for research and clinical laboratories where access to large-scale sequencing is limited.

Individuals are not stable things, they are fleeting. Chromosomes too are shuffled into oblivion, like hands of cards soon after they are dealt. But the cards themselves survive the shuffling. The cards are the genes. The genes are not destroyed by crossing-over, they merely change partners and march on. Of course they march on. That is their business. They are the replicators and we are their survival machines. When we have served our purpose we are cast aside. But genes are denizens of geological time: genes are forever.

*Richard Dawkins, The selfish gene*

# Acknowledgments

# Contents

# List of Figures

# Chapter 1

# Introduction

In the last decade, massively parallel high-throughput short-read sequencing has revolutionized the efficiency and breadth of applications for DNA sequencing (Kircher and Kelso, 2010). These high-throughput sequencing methods produce millions to billions of short reads in a single run, and have led to the development of many applications that depend on "read-counting" to measure the abundance of specific sequences in a sample. Examples include RNA-seq, ChIP-seq, and whole genome copy number profiling.

Recently, long-read technologies have been developed that are filling the gap left by short-read sequencers in applications such as genome assembly (Jain et al., 2018a; Loman et al., 2015), which benefit from connecting more distant sequences within a contiguous molecule. Among these the MinION instrument, from Oxford Nanopore Technologies, is highly portable and inexpensive and has shown its unique value for analysis outside of central sequencing facilities (Quick et al., 2016). Long-read sequencers such as the MinION typically produce vastly fewer reads from a sequencing run, and are therefore less efficient in applications that use sequenced reads purely as a means to count molecules. However, these technologies have the enormous advantage of operating in near real-time, with a turnaround time that can be measured in hours for some applications, rather than days or weeks.

Copy number variation (CNV) has been used successfully to understand a variety of diseases (Sebat et al., 2007) – notably cancers, which exhibit both extreme variation and recurrent trends that can be used for diagnostics and personalized approaches to treatment. For example, the amplification and loss of certain genes, such as *RB1* deletion and *MYCN* amplification in retinoblastoma, can be prognostic or even predictive for treatment (Berry et al., 2017). High-throughput short-read sequencing has been extremely effective in copy number profiling of cancers (Chiang et al., 2009), including profiling single tumor cells (Navin et al., 2011). However, for many potential users, the efficiency of high-throughput short-read sequencing in CNV analysis is determined by the availability of instruments and need for heavy multiplexing to hit reasonable cost per profile. A sequencing core is typically involved and an individual profile must wait for a "full" run before it can be processed. The MinION sequencer has an accessible buy-in and is easy to use. Unfortunately the MinION has optimal nucleotide throughput when producing reads that are orders of magnitude longer than needed for CNV profiling.

To make full use of the advantages offered by the MinION sequencer, we introduce sampling molecules using re-ligated fragments (SMURF)-seq, a protocol to efficiently sequence short DNA molecules on a long-read sequencer. The strategy of SMURF-seq is to concatenate short fragments into very long molecules ($\sim$8 kb) prior to sequencing.

# Chapter 2

# Background

## 2.1 Nanopore sequencing

**A brief history of nanopore sequencing** The concept of nanopore sequencing is based on the idea that as a single-stranded DNA (or RNA) translocates through a nanometer sized pore, a nanopore, in the presence of an electric field, the change in current level measured across the nanopore would be dependent on the nucleotide passing through the nanopore; Thus, measuring the current over time could be leveraged to determine the sequence of nucleotides (Fig. 2.1b). This idea of using transmemberane proteins an nanopores for sensing and sequencing nucleic acids was independently thought of by several researches including David Deamer, Hagan Bayley, and George Church (Bayley, 2015; Branton et al., 2009; Deamer et al., 2016).

Initial experiments showed that as a single-stranded DNA or RNA molecules could be driven through a *Staphylococcus aureus* $\alpha$-hemolysin in the presence of an electric field (Kasianowicz et al., 1996). The current through the pore remained constant in the absence of oligomers; and the presence of oligomers caused transient decreases in current, with the duration of the decrease proportional to the length of the oligomer. Further research demonstrated that decrease in amplitude of current could be used to differentiate between poly-purine and poly-pyrimidine sequences

of RNA (Akeson et al., 1999) and DNA (Meller et al., 2000). It was also observed that the DNA molecules translocate through the nanopore at few microseconds per base (Meller et al., 2000).

Although these experiments demonstrated the potential for nanopores to distinguish nucleic acid polymers, several challenges remained to be addressed to use this approach for reading individual bases on a DNA or RNA molecule. The most important of these were detecting the bases on a molecule at single nucleotide resolution and slowing the rate of translocation through the nanopore so that a readout can be obtained (Bayley, 2015; Branton et al., 2009). These challenges were resolved in the forthcoming years. A few notable milestones are described below.

In regard to identifying individual nucleobases, all four bases were identified in single-stranded DNA that have a terminal hairpin structure (Ashkenasy et al., 2005) and single-stranded DNA attached to streptavidin with a biotin linker (Purnell and Schmidt, 2009; Stoddart et al., 2009). These structures immobilized the DNA in the nanopore (either a wildtype $\alpha$-hemolysin or an engineered

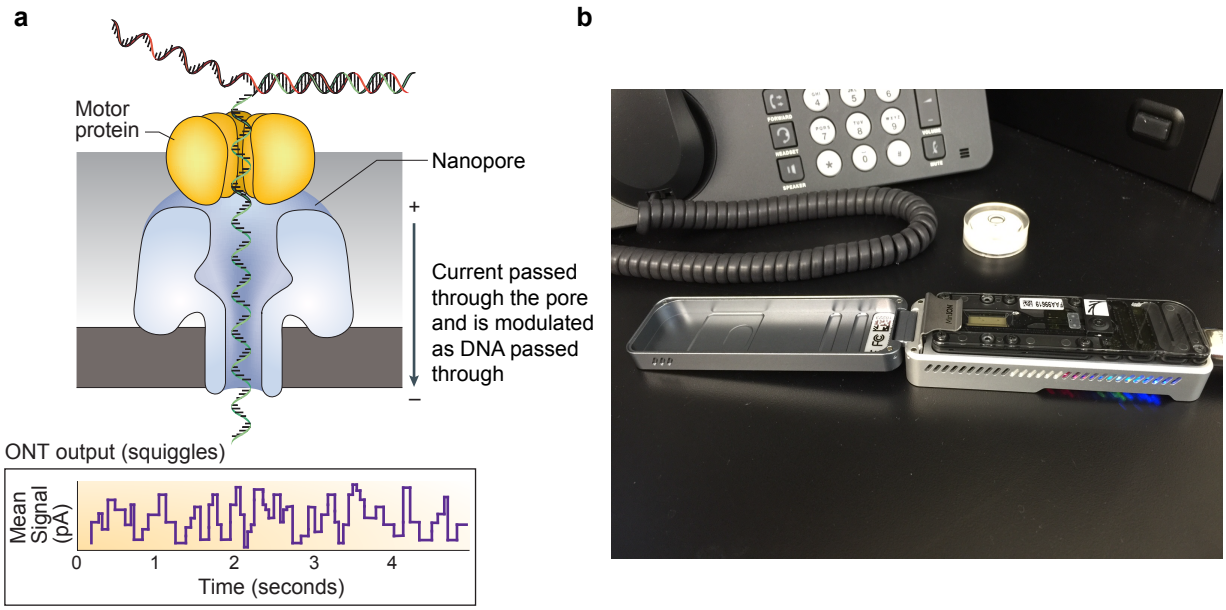


Figure 2.1: Nanopore sequencing. (a) Current across a nanopore is measured as a DNA molecule translocates through a nanopore. This figure is adapted from Figure 5Ab in Goodwin et al. (2016). (b) Oxford Nanopore Technologies MinION instrument.

form of it) allowing sufficient time for detection. Thus, with the appropriate sequence, each base could be resolved, and further, the location within a nanopore that is most sensitive to detect bases was also determined.

The frequency of translocation of DNA molecules through a $\alpha$-hemolysin pore was increased and the voltage threshold for translocation was decreased by engineering the pore to have positive charged groups in the lining of the lumen (Maglia et al., 2008). However, a disadvantage of the $\alpha$-hemolysin pores is that the region that is sensitive to the nucleotides in the pore is too wide, and so the current differences are small between nucleotides, making single nucleotide detection difficult. Pores derived from *Mycobacterium smegmatis* porin A has a narrower sensitive region, and could detect nucleotides at a higher resolution (Butler et al., 2008; Manrao et al., 2011).

In regard to controlling the rate of translocation through a nanopore, there were two approaches, exo-sequencing and strand-sequencing. In the exo-sequencing approach, individual bases from a DNA are cleaved into a nanopore with an exonuclease and unidentified one at a time (Astier et al., 2006; Clarke et al., 2009). Alternatively, in the strand sequencing approach a DNA molecule is threaded though a nanopore at a controlled rate and the bases are identified from the continuous change in current levels. Initial approaches in strand sequencing used a DNA polymerase (derived from *Escherichia coli* Kleenow fragment or bacteriophage T7) and recorded the current levels as the DNA translocates through a pore with the incorporation of each nucleotide (Benner et al., 2007; Chu et al., 2010; Cockroft et al., 2008; Gyarfas et al., 2009). Subsequently it was shown that $\phi$29 polymerase could be used to control ratcheting in both the forward and the reverse direction (Cherf et al., 2012; Lieberman et al., 2010; Manrao et al., 2012). $\phi$29 was used to sequence reads up to 4.5 kb from the $\phi$X174 genome using nanopores (Laszlo et al., 2014).

**Oxford Nanopore Technologies** Oxford Nanopore Technologies was founded in 2005 by Hagan Bayley and colleagues (Deamer et al., 2016). Oxford Nanopore Technologies announced the MinION instrument at the Advances in Genome Biology and Technology meeting in 2012 and

made it available to early access researches in 2014 (Bayley, 2015; Deamer et al., 2016).

The MinION sequencer (Fig. 2.1b) is a portable instrument requiring just a modest computer for control and data acquisition. A MinION flowcell of 2048 nanopores (at present, a derivative of *Escherichia coli* CsgG protein (Brown and Clarke, 2016)) embedded on a membrane, of which 512 pores can sequence molecules in parallel.

Whole genomes of several organisms including humans have been sequenced using the MinION instrument (Bowden et al., 2019; Jain et al., 2018a; Loman et al., 2015; Moss et al., 2020; Stancu et al., 2017). It has also been used in several other applications such as disease surveillance (Faria et al., 2016; Quick et al., 2016), metagenomics (Charalampous et al., 2019; Goordial et al., 2017; Leggett et al., 2020), direct RNA sequencing (Depledge et al., 2019; Garalde et al., 2018; Workman et al., 2019), and detecting methylated bases (Liu et al., 2019; Rand et al., 2017; Simpson et al., 2017) among others (Jain et al., 2016).

In addition to the MinION instrument, Oxford Nanopore Technologies currently offers various nanopore devices including the bench-top GridION and PromethION which allows parallel sequencing with up to 5 and 48 flowcells, respectively.

**Nanopore library preparation and sequencing** Sequencing nucleic acids requires prepossessing the sample DNA for compatibility with the underlying sequencing technology, a process traditionally referred to as library preparation. Sequencing on a nanopore machine usually requires fragmenting DNA molecules to the appropriate length and attaching sequencing adapters. Oxford nanopore technologies offers several commercially library preparation kits for both DNA and RNA samples. The most frequently used of these kits are the Ligation Sequencing Kit family and the Rapid Sequencing Kit family.

In theory, there is no limit on the length of a molecule that can be sequenced with a nanopore, and thus, the length is determined by the downstream application or the limitations of handling high molecular weight DNA. For the ligation sequencing kit (SQK-LSK108 1D DNA by ligation),

the recommended length is $\sim$8 kb when starting with 1ug of sample to ensure appropriate molar concentration in the subsequent steps. DNA molecules can be fragmented to the appropriate length using a variety of methods including the Covaris g-TUBE. These molecules are then optionally repaired to remove any nicks, and then the DNA ends are prepared to have a dA tail. Finally, sequencing adapters (that have a dT tail) are ligated to the end-prepared DNA. These adapters contain specific DNA sequenced with attached enzymes that regulate translocation of the DNA molecule into a nanopore. Library preparation with the ligation kit takes approximately 60 minutes.

The rapid library preparation kit (SQK-RAD003 Rapid sequencing) offers a faster method, by simultaneously fragmenting and tagging the ends of high molecular weight DNA (recommended $>$ 30 kb). Adapters are then attached to these tags. Library preparation with the rapid kit takes approximately 10 minutes.

Both of these kits offer barcoding capabilities for multiplexing several samples in a single sequencing run. For example, the Native Barcode Kit (EXP-NBD103) is used in addition to the ligation sequencing kit, and adds a barcode sequence to the end-prepared DNA molecules prior to ligating sequencing adapters. After library preparation with a unique barcode for each sample, they can pooled in appropriate molar concentrations before sequencing.

Oxford Nanopore Technologies offers severs other preparation kits, such as the $1D^2$ kit for higher accuracy reads and PCR based kits when starting with nanogram or picogram amounts of DNA.

After library construction the sample is ready to be sequenced. The flowcell is loaded on the sequencing machine, primed with the appropriate buffers, and the sample is loaded. After which the sequencing can be started and the reads are available as they are sequenced in real-time. The sequencing process is controlled by the MinKNOW tool, and can continue for up to 48 hours. The sequenced reads can either be base-called in real time using the MinKNOW or at a later time using the Guppy tool.

## 2.2 Copy number variation and profiling

**Copy number variation** Sources of genomic variation in humans include single-nucleotide polymorphisms (SNPs), short insertions and deletions, and repeats (). Another source of variation, copy number variation (CNV), is the change in number of copies of a region of the genome larger than 1 kb with respect to a reference genome (Feuk et al., 2006; Redon et al., 2006). The number of copies of a region of the genome could increase resulting in a copy number "gain" (also referred to as amplification or duplication), or decrease resulting in a copy number "loss" (also referred to as deletion). Changes in copy number can occur due to mechanisms such as homologous recombination and non-homologous DNA repair mechanisms (Hastings et al., 2009; Van Binsbergen, 2011).

Copy number variations contribute both to diversity in the human population and to disease (). Variations that contribute to diversity are present in the germline, whereas those that contribute to diseases could in the germline or somatic. (Somtic copy number variations are generally referred to as somatic copy number alterations, or CNA in short.)

**CNV in diversity** CNVs as a significant source of diversity in humans was established by Sebat et al. (2004) and Iafrate et al. (2004). Sebat et al. (2004) identified 221 copy number differences in 20 individuals with an average of 11 differences between individuals; these variations had an average length of 465 kb (median length of 222 kb). Iafrate et al. (2004) identified 225 regions in 55 individuals with an average of 12.4 differences between individuals; these variations ranged from 150 kb to 425 kb. These studies were extended to larger populations and to populations of different ancestry (Li et al., 2009; Redon et al., 2006). CNVs in the human genome and that those contribute to diversity are reviewed in (Feuk et al., 2006; Freeman et al., 2006; Zarrei et al., 2015).

**CNV in diseases**

**CNA in cancer**

**CNA profiling methods**

**Read counting based CNA profiling**

## 2.3 Prior protocols based on concatenating DNA molecules

**Serial analysis of gene expression (SAGE)** The concept of ligating short DNA molecules to improve the efficiency of sequencing was introduced in serial analysis of gene expression (SAGE) (), and subsequently its variants such as LongSAGE and SuperSAGE (). SAGE

**Variants of SAGE**

**Digital karyotyping**

**SMASH**

**Concat-seq**

# Chapter 3

# Sampling molecules using re-ligated fragments (SMURF)-seq

## 3.1 Naive approaches to read-counting on nanopore machines

Copy number profiling, and read-counting in general, can be done on nanopore sequencers with long reads ($\sim$8kb) following the standard sequencing procedure. Since nanopore machines are optimized for long read sequencing, this method has the advantage of using any standard library preparation protocol that are commercially available. Sequencing these long molecules using a nanopore keeps a pore occupied for a longer duration once a pore is loaded followed by an open pore waiting for a molecule to be reload in a pore. Further, technical nucleotides, such as sequencing adapters and barcodes, are sequenced one (or twice) every $\sim$8k bases, thus the fraction of time a nanopore spends sequencing technical nucleotides is low. However, read-counting applications do not benefit from longer reads beyond what is necessary for unique mapping to the reference genome. In these applications, for any fixed number of nucleotides sequenced, more information would be obtained if those nucleotides are organized as more DNA molecules, rather than longer contiguous fragments.

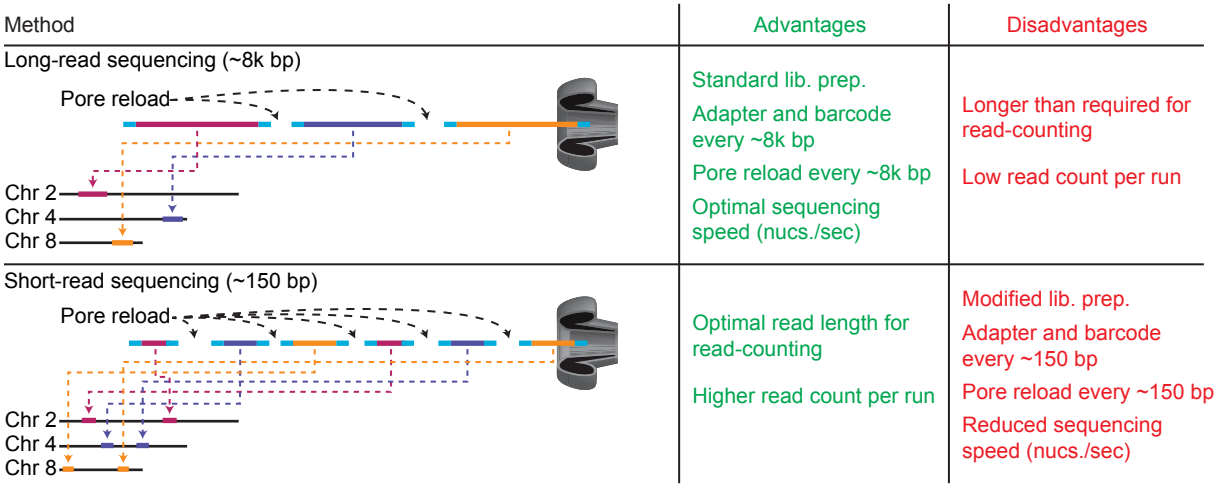


Figure 3.1: Naive approaches to read-counting on nanopore machines. Sequencing long-reads directly is optimized for nanopore machines but not for read-counting applications. Sequencing short-read is optimized for read-counting applications but not for nanopore sequencing.

An alternate approach to read-counting is to sequence short reads ($\sim$150 bp) directly on a nanopore sequencer. In general, for a given sample of DNA, a nanopore instrument will generate more reads if the corresponding molecules are shorter. Once a molecule is loaded into a pore, the time spent sequencing is less for shorter reads. In addition, for a fixed amount of DNA, shorter molecules result in higher molar concentration when loaded onto the machine, increasing the rate at which each pore captures molecules (Muthukumar, 2010; Wanunu et al., 2008). Therefore, sequencing short reads on a nanopore machine would generate more reads from a sequencing run than sequencing long reads. However, sequencing short reads requires ad-hoc modifications to the library preparation protocol as these are optimized for longer molecules. Sequencing these shorter molecules keeps a pore occupied for a shorter duration once a pore is loaded followed by waiting for a pore to be reloaded (but the reload time is usually shorter due to the higher molar concentration) Moreover, technical nucleotides are sequenced every $\sim$150bp, increasing the fraction of time a nanopore sequences the technical bases.

SMURF-seq approach combines the advantages of both of these methods while alleviating the

drawbacks by using a nanopore instrument as intended for long-read sequencing, while generating the desired short fragments. Using the SMURF-seq approach we generate higher read counts per run than sequencing long or short molecules directly.

## 3.2 SMURF-seq approach to read counting

The SMURF-seq protocol involves cleaving genomic DNA into short fragments, with length just sufficient for an acceptable rate of uniquely mapping fragments in the reference genome. These fragmented molecules are then randomly ligated back together to form artificial long DNA molecules, as required for long-read sequencing. The long re-ligated molecules are sequenced following the standard MinION library preparation protocol. After (or possibly concurrent with) sequencing, the SMURF-seq reads are mapped to the reference genome in a way that simultaneously splits them into their constituent fragments, each aligning to a distinct location in the genome (Fig. 3.2).

More specifically, genomic DNA was fragmented using restriction enzymes and ligated with T4 DNA ligase, with clean-up steps in between. SMURF-seq protocol is completely enzymatic and takes less than 90 minutes to complete (Fig. 3.3). The details of these steps are given below:

1. Restriction enzyme digestion: restriction enzymes recognize and cleave specific DNA sequences, typically producing sticky-ended DNA molecules. The choice of restriction enzyme used is primarily dependent on the size of the fragmented molecules produced. Based on the downstream application, they could also be influenced by other factors such as any bias they could introduce. An advantage of using restriction enzymes to fragment DNA molecules, over other fragmentation techniques, is that the fragmented molecules have a uniform ends (either overhangs with the same sequence or blunt-ends) and are thus compatible for ligation without an end-repair step in between.
2. Clean-up: the reaction containing the restriction enzymes and the fragmented DNA molecules

is cleaned to wash out the enzymes and retain the DNA molecules. The choice of clean-up kit used, also determines the length of the DNA molecules that are retained. We used a spin-column based clean-up that typically retains molecules that are over ∼70 bp. However, other clean-up kits, such as bead-based kits, could also be used at this step.

3. Re-ligation: fragmented DNA molecules with uniform ends are ligated at random with T4 DNA ligase enzymes. The most important factor in a ligation reaction is the concentration of compatible DNA ends (Dugaiczyk et al., 1975). At high concentrations, the chances are higher for ligation between two molecules than a molecule self-ligating. At low concentrations, the chances are higher for self-ligation. Thus, the main consideration during the ligation step is the duration of the ligation reaction, as the molar concentration of DNA molecules decrease with time. Too little time would lead to insufficient ligation, resulting in molecules of length that

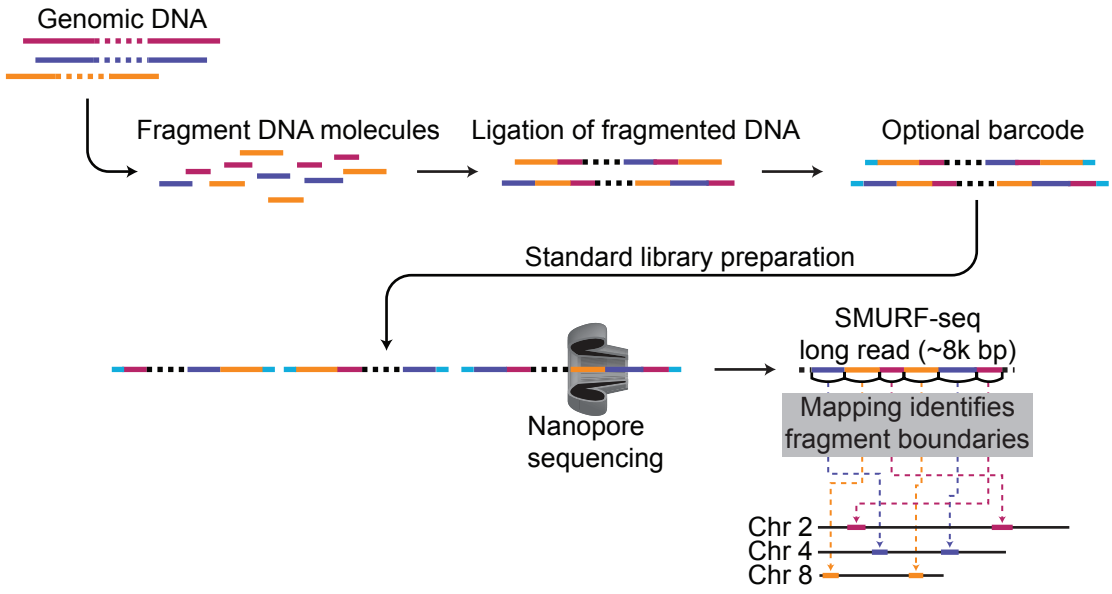


Figure 3.2: SMURF-seq approach to sequencing short fragments. SMURF-seq efficiently sequences short fragments of DNA for read-counting applications with a reference genome on long-read sequencers, and yields up to 30 countable fragments per sequenced read. SMURF-seq sequences short DNA molecules by generating long concatenated molecules from these. SMURF-seq reads are aligned by splitting them into multiple fragments, each aligning to a distinct region in the genome.

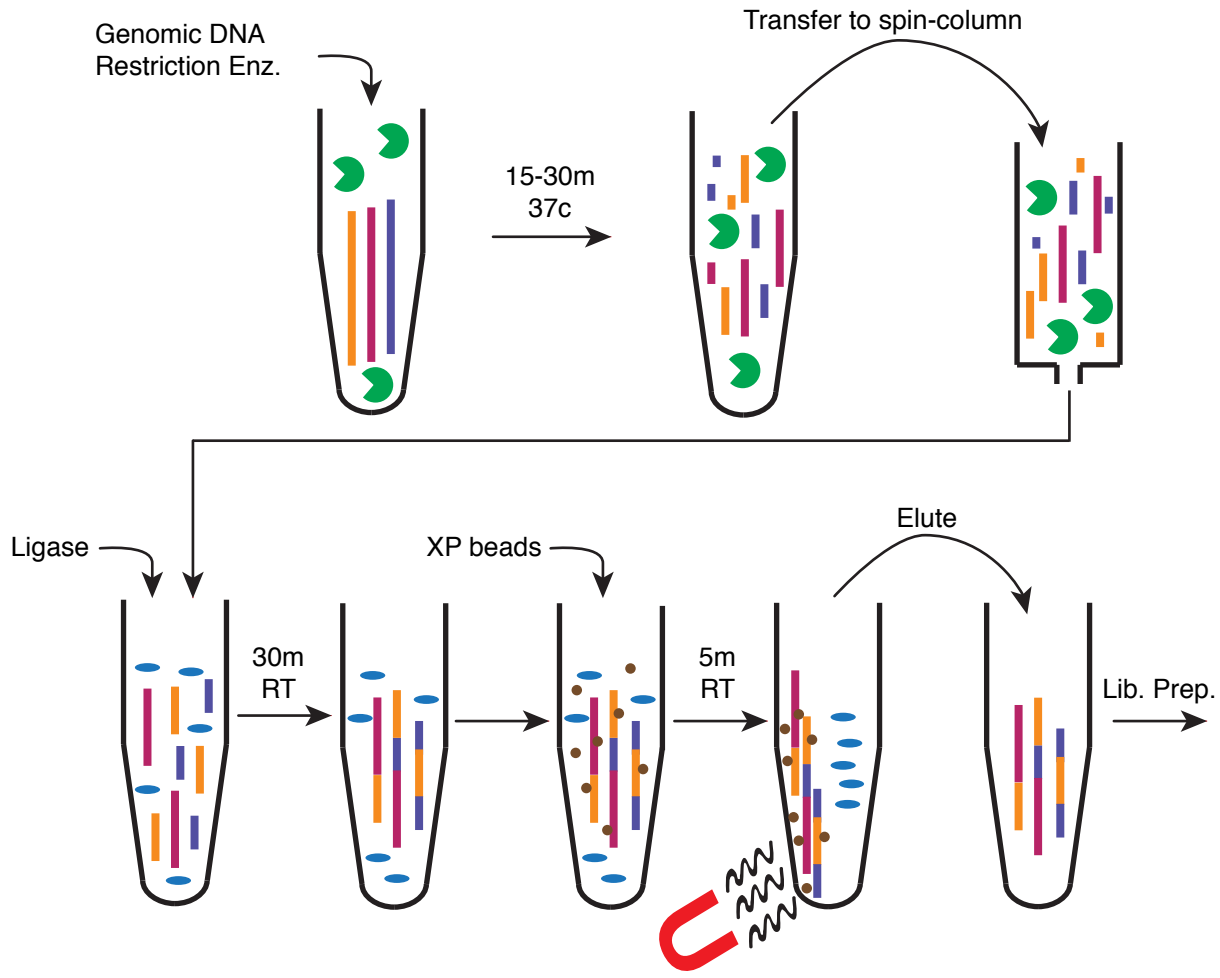


Figure 3.3: Schematic of SMURF-seq protocol. SMURF-seq consists of four steps: restriction enzyme digestion, spin-column clean-up, re-ligation of fragmented DNA, and Ampure XP beads clean-up.

do not achieve optimal SMURF-seq efficiency. On the other extreme, too much time would result in circular molecules that are incompatible with the most downstream library preparation process. A typical ligation reaction would contain both short and circularized molecules, and achieving a balance between these determines the efficiency of SMURF-seq. Other factors such as the temperature and buffer contents also affect the ligation process. In our experiments, the ligation reaction was performed at a DNA concentration of $25\,\mathrm{ng/\mu l}$ ($500\,\mathrm{ng}$ of DNA in $10\,\mu\mathrm{l}$ nuclease free water and $10\,\mu\mathrm{l}$ DNA ligase) for $30\,\mathrm{min}$.

4. Bead-based clean-up: the reaction containing the ligase enzymes and ligated DNA molecules is cleaned to retain only the ligated molecules. We used a bead-based clean-up at this step to avoid damage to long DNA molecules that are typical of spin-column based methods.

DNA molecules that are resultant of the SMURF-seq protocol are long DNA molecules that are several kilo-bases long, and therefore, any standard library preparation kits that are available for nanopore machines can be used with SMURF-seq molecules. These molecules can also be barcoded with one (or two) barcode sequences per molecule. Thus, the SMURF-seq approach overcomes the disadvantages of sequencing long or short DNA molecules directly on a nanopore machine for read-counting applications, and improves its efficiency for read-counting applications.

We also tested dsDNA Fragmentase enzymes (New England Biolabs) and acoustic shearing (Covaris) to fragment DNA. However, these methods require an additional end-repair step after fragmentation and the ligated molecules failed to reach the lengths we obtained by using restriction fragmentation.

In our applications, we used SaqAI restriction enzyme, which recognizes the sequence TTAA and produces molecules with mean lengths of 150.2 bp (Fig. 3.4a). The fragmented DNA molecules are then ligated randomly to form longer molecules using T4 DNA ligase enzyme (Fig. 3.4b). In our experiments, the resulting long DNA molecules were sequenced using the Oxford Nanopore Technologies 1D DNA by ligation kit (SQK-LSK108) or the rapid sequencing kit (SQK-RAD003) following the standard manufacturers protocol. We also multiplexed samples using the 1D native barcoding genomic DNA kit (EXP-NBD103) followed by library preparation using the 1D DNA by ligation kit. The detailed SMURF-seq protocol is given in Appendix A.

## 3.3 Mapping SMURF-seq reads

The reads sequenced using SMURF-seq can be mapped to a reference genome by first identifying short matches within the reads, corresponding to parts of the individual fragments, and then ex-

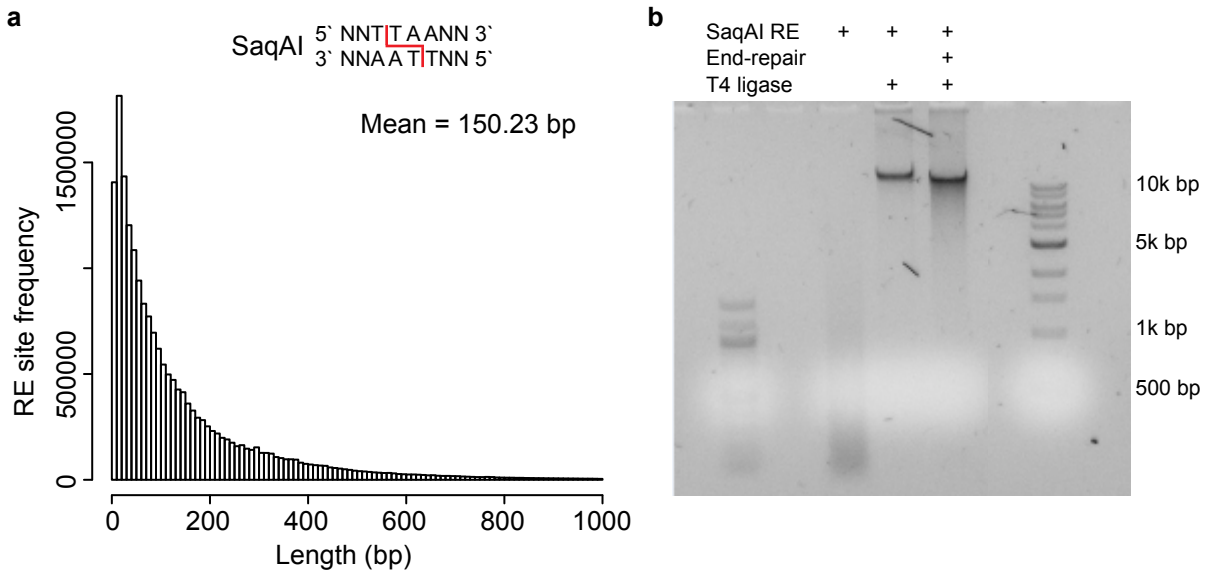


Figure 3.4: Restriction digestion and ligation of DNA molecules. (a) Distribution of length between restriction sites computed by measuring the distance between the recognition sites on the human reference genome. SaqAI recognizes the sequence TTAA and leaves a 2 bp overhang. (b) Negative gel image of fragmented and ligated normal diploid DNA using SaqAI restriction enzyme and T4 DNA ligase. Sticky-end and blunt-end ligation (by end-repair) of fragmented DNA are shown, and both yield ligated molecules of approximately the same length.

tending those to locate fragment boundaries. As currently implemented, SMURF-seq fragments are longer than $\sim$150bp, and mapping these reads is is handled nicely using the seed-and-extend paradigm implemented in many existing long-read mapping tools. Although none of these tools were designed to align SMURF-seq reads, several long read aligners include steps designed for split-read alignment, which can be leveraged from aligning SMURF-seq reads.

Aligning SMURF-seq reads with long-read mapping tools involve variations of the following steps:

- Identifying seeds: Mapping tools have a step of identifying seeds, which are short exactly matching parts of the read with parts of the reference genome. Choices in how seeds are defined and used are often made for mapping speed. The total size of SMURF-seq data sets is currently (relatively) small, so speed is not our primary concern. We favor the most sensitive seed strategy, but depending on implementation too many seed hits could lead to ambiguity later in the

mapping process.

- Chaining seeds: The identified seeds are further extended into proto alignments, and filtered to avoid aligning potentially false positive seed hits.
- Aligning within the chains: In this stage a Smith-Waterman alignment is performed, typically allowing users to specify a mismatch penalty along with penalties for both gap-open and gap-extend.
- Selecting best alignments: When high-scoring alignments overlap within a read, one of them (or both) could be trimmed or one is selected and the other discarded. The choices made here could lead to discarding entire fragments.

Mapping tools have several parameter options, in general, these are related to: (1) the seeding and chaining algorithm used by the individual tool. (2) The Smith-Waterman alignment scores, i.e. the match score, and the mismatch and indel penalty. The seeding and chaining parameters control the number of proto alignments that are further refined by aligning parts of the read to the reference genome using the specified alignment scores.

The Smith-Waterman alignment score used to align fragments to the reference genome is crucial for determining the optimal fragment length. On one extreme, a match score of 1 with a mismatch and indel penalty of 0 will result in one identified fragment covering the entire read and mapping perfectly, but will always map ambiguously. On the other extreme, a match score of 1 with a mismatch and indel penalty of $-\infty$ will result in any mismatch or indel on the read to be considered as a fragment boundary. Therefore to align SMURF-seq reads, we need to determine optimal alignment scores to use.

We evaluated mapping tools on simulated SMURF-seq data generated by concatenating random fragments from real Oxford Nanopore reads. This emulates idealized SMURF-seq reads. Within the simulated reads, the boundaries of each fragment are known *a priori*, as are their mapping locations when in the context of their original long reads. We used this information to evaluate mapping tools in terms of (1) how well they identify fragments purely for the purpose of counting

molecules, which is the primary information used in CNV analysis, and (2) how well they identify individual mapping bases within reads. After mapping these reads, we calculated precision and recall for identifying both the correct fragment locations, and the individual mapping bases within the fragments (i.e. the correct fragment boundaries). Using this simulation setup, we determined the optimal Smith-Waterman alignment score for use with SMURF-seq reads.

### 3.3.1 Simulating SMURF-seq reads to evaluate mapping programs

To test these mapping tools, we chose to create simulated reads with the technical characteristics we expect in idealized SMURF-seq data. We first selected a fragment length $\ell$ and a number $k$ of fragments per read. Then, for a given WGS nanopore data set, we took the set of mapped long reads as determined by BWA-MEM (with `-x ont2d` option). Each of the mapped reads was split into fragments of length $\ell$ (with a random offset of $0$ to $\ell - 1$ at the start of the long read). Each fragment was validated by requiring that it did not overlap a deadzone in the genome (as determined by the deadzone program available from https://github.com/smithlabcode/utils for 40 bp). The reason for excluding deadzones is that even when a short fragment has a "known" mapping location when it is part of a longer read, we cannot compare its reported mapping location as a short fragment with that known location, since we expect any good mapping algorithm to identify that the fragment maps ambiguously. Among these validated fragments, subsets of $k$ were sampled uniformly at random and concatenated (in random order and orientation) to form simulated SMURF-seq reads.

The first and last fragments in a read should be slightly easier to identify and map than the rest, since one of their boundaries is known. Using the above procedure, we select $k = 20$ so that the simulated reads have a sufficient number of fragments to eliminate the influence of the first and last fragments in each read on the results. There is no need to have large $k$ otherwise.

By lowering $\ell$ and making the fragments shorter, the task of mapping the fragments becomes more challenging. Real SMURF-seq reads have fragment lengths determined by restriction site density, size selection and other aspects of the experiments. But in testing mapping algorithms and

optimizing parameters, there is no disadvantage to making the task more challenging. We only need to be able to distinguish the relative performance of different mapping tools and parameter combinations. Real SMURF-seq reads have varying fragment lengths, but in evaluating mapping tools, there is no need to randomize fragment lengths. None of the algorithms we evaluated are capable of either deducing or leveraging the fact that all simulated fragments have the same length. We selected $\ell = 100$, which begins to challenge the various mapping strategies. These values of $\ell$ are slightly lower than the average in real SMURF-seq data.

### 3.3.2 Evaluating performance using simulated SMURF-seq reads

Within the simulated reads, the boundaries of each fragment are known *a priori*, as are their mapping locations. We used this information to evaluate mapping tools in terms of (1) how well they identify fragments purely for the purpose of counting molecules, which is the primary information used in CNV analysis, and (2) how well they identify individual mapping bases within reads. The latter criteria becomes important in challenging cases and will be increasingly important as fragment sizes are reduced.

Performance on identifying fragments: After mapping these simulated reads, each mapping result is called a predicted fragment. Each predicted fragment is considered a positive prediction, and we assume an arbitrary order over positive predictions. A positive prediction is a true positive if:

- The predicted fragment maps uniquely.
- The mapping locations of at least half the bases in the predicted fragment are equal to the original mapping locations for those bases, and those bases are all part of the same original fragment (we assume that it is unlikely for two fragments on a simulated read to have the same mapping location but opposite orientation, and thus do not check for the orientation of a fragment). In this case, we say the predicted fragment is associated with that original fragment.
- The predicted fragment is the first among predicted fragments associated the same original frag-

ment.

False positives are predicted fragments that are not true positives. Any original fragment with no associated predicted fragment is a false negative. These criteria penalize splitting one original fragment or merging two original fragments. By defining true positives, false positives and false negatives we are able to calculate precision, recall, and F-score for a particular mapping strategy.

Performance on identifying individual mapping bases: After mapping simulated reads, each mapping result is decomposed into individual nucleotides and associated with a location in the genome. Those locations are retained. We keep multiplicities, so when two mapped fragments overlap in the genome we count certain nucleotides twice. These are the predicted positive bases in the reference. The condition positive bases are those known *a priori* from the simulation. The original fragment mapping locations may overlap in the reference genome, leading to multiplicities in the condition positive bases, but with low probability. The true positives are the intersection of the condition positive and the predicted positive bases. When there are multiplicities of mapped fragments and simulated fragments overlapping the same bases in the reference genome, this is determined by taking the smaller of the two values. After removing the true positives bases, the remaining predicted positive bases are false positives, and the remaining condition positive bases are false negatives. These criteria penalize mapping approaches that do not cover the entire simulated SMURF-seq reads, and also penalize approaches that predict fragments that overlap within the read. The true positives, false positives, and false negatives here allow us to assign precision and recall in terms of individual bases and corresponding F-scores. Although the reference bases for both predicted positive and condition positive could involve multisets, since our simulations used relatively low coverage this almost never happened.

To generate simulated reads we used the standard long reads from four sequencing runs (Flowcell ID: FAB42704, FAB42810, FAB49914, and FAF01253) in the public dataset available at https://github.com/nanopore-wgs-consortium/NA12878/blob/master/Genome.md (Jain et al., 2018a,b). We downloaded the raw data from EBI (Run accession: ERR2184696, ERR2184704, ERR2184712,

and ERR2184722) and base-called these with Guppy (version: 2.3.5).

### 3.3.3 Initial selection of mapping tools

We tested the following mapping tools: BWA-MEM(Li, 2013), Minimap2(Li, 2018), LAST(Kiełbasa et al., 2011), GraphMap(Sović et al., 2016), BLASR(Chaisson and Tesler, 2012), rHAT(Liu et al., 2015), and LAMSA(Liu et al., 2017). These were selected either because they are known to perform well on certain mapping tasks or have unique properties that plausibly could help in mapping SMURF-seq reads. We tested each of these using default parameters on simulated reads and downsampled real SMURF-seq reads (data not shown). Among these BWA-MEM, Minimap2, and LAST had higher accuracy on simulated data, and the other tools identified at most 15 fragments per read on real data. Thus, we explored performance of BWA-MEM (0.7.17), LAST (963), and Minimap2 (2.15) in more detail, varying parameters to improve performance.

We remark that none of these tools were designed to map SMURF-seq reads; results we report here do not reflect the overall performance of the various mapping tools, only that the three aforementioned tools happened to perform relatively well on a task for which they were not directly designed for.

### 3.3.4 Determining the optimal Smith-Waterman score for SMURF-seq reads

In order to determine the optimal alignment score, we kept the seeding related parameters constant, and varied the alignment score combinations to perform a grid search. We varied the mismatch penalty from 1 to 6, gap open penalty from 0 to 4, and gap extend penalty form 1 to 4. The match score was fixed at 1. Thus for each tool we tested 120 ($6 \times 5 \times 4$) combinations of alignment scores.

The seeding and chaining related parameters for each tool was set at follows (along with the four alignment scores):

- BWA-MEM: `-x ont2d -k 12 -W 12 -T 30`

- Minimap2: `-w 1 -m 10 -s 30`
- LAST (NEAR): `lastal -Q0 -e 20` and `last-split -m 1 -s 30`

We set the seeding and chaining parameters in a liberal manner to allow for higher sensitivity than the default parameter of each tool, and the minimum alignment score to output was set at 30.

After aligning the simulated reads, we calculated the average precision and recall, each for the mapped fragment locations and nucleotides, for the four datasets. The F-score was computed for each, and the mean of the F-scores was used to determine the optimal alignment parameter for each tool. Based on these results BWA-MEM outperformed other tools for aligning SMURF-seq reads. BWA-MEM performed best with a mismatch, open, and extension penalty of 2, 1, 1 respectively.

To further refine the optimal alignment parameter for BWA-MEM, we aligned the simulated reads with parameter values around the value described above with a higher resolution. We varied the mismatch penalty from 1.5 to 2.5, and open and extend penalties from 0.5 to 1.5 in increments of 0.25. However, BWA-MEM does not accept floating point values for alignment score parameters. To overcome this, we scaled the alignment score proportionately to have integer values, i.e we varied the mismatch penalty from 6 to 10, open and extend penalties from 2 to 6, and fixed the match score at 4 (125 combinations). Based on these results, the highest accuracy was obtained with the mismatch, open, and extension penalty of 2.5, 1.5, 0.75 respectively (corresponding scaled values are 10, 6 and 3). We used these optimal alignment scores for mapping real SMURF-seq read, and all the CNV profiles presented are based on these.

## 3.4 Generating higher fragment counts with SMURF-seq

A typical Oxford MinION sequencing run generates approximately 500k reads (length $\sim$8 kb) (Jain et al., 2018a; Tyson et al., 2018) using the standard library preparation and sequencing protocols. Several studies have used these long reads for copy number profiling (Euskirchen et al., 2017; Magi et al., 2019). We tested the ability of the Oxford MinION instrument to sequence short

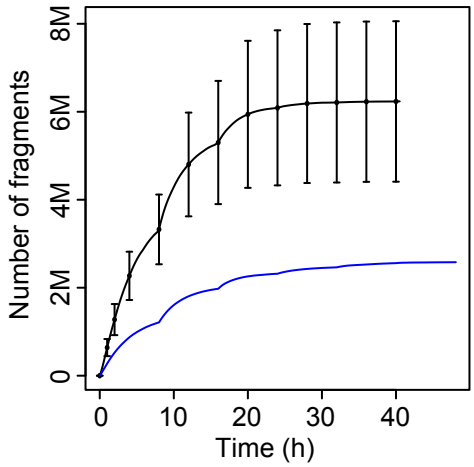


Figure 3.5: SMURF-seq generates fragments at a faster rate than sequencing short molecules directly. Number of fragments obtained from reads plotted as a function of sequencing time. For SMURF-seq, the average number of fragments from runs using the 1D sequencing by ligation kits are plotted (Error bars indicate one standard deviation). For short-read sequencing run, each read is considered as one fragment.

DNA molecules by sequencing restriction enzyme (SaqAI) digested normal diploid genome. The sequencing run produced 2.58 million reads with a mean read length of 630.93 bp. Using the same instrument, with SMURF-seq, we report here an average of 6.2 million mapped fragments per run, which is substantially more fragments than directly sequencing long or short reads directly. Further, the SMURF-seq approach generated fragments at a substantially faster rate than sequencing short molecules directly (Fig. 3.5).

The most important factor in the performance of SMURF-seq over sequencing short molecules directly is that sequencing concatenated fragments effectively eliminates the pore reload time for all but the first fragment in each read. However, there are a variety of additional factors that favor further optimization of the approach employed by SMURF-seq. First, reduction of resources spent on technical nucleotides: SMURF-seq uses a single barcode and sequencing adapter per read consisting of multiple fragments; sequencing short reads uses one barcode and adapter per fragment, adding approximately 50 bases to each fragment. This increases the time to sequence

each short read. In sequencing short reads, as the reads get shorter the time consumed by these technical bases increases. In SMURF-seq, sequencing either shorter fragments in fixed length reads, or longer reads containing fragments of fixed average length, both reduce the time consumed sequencing these technical bases. In the limit, assuming 100bp DNA fragments, sequencing those fragments as short-reads corresponds to 33% technical nucleotides; for SMURF-seq, the portion of technical nucleotides remains low. Second, more nucleotides sequenced at full speed: We observed that the speed of sequencing was lower when sequencing short molecules. For example, the average sequencing speed was 315.54 bases per second for sequencing the diploid genome without SMURF-seq, and 400.29 bases per second when sequencing using SMURF-seq on the MinION sequencer. Third, leveraging optimizations to long-read protocols: The rapidly evolving nanopore library construction kits are continually optimized for long-read sequencing, and would likely require significant ad-hoc modifications to optimize sequencing of short molecules of length optimal for read-counting applications.

## 3.5 Efficient CNV profiling using SMURF-seq

To demonstrate the utility of SMURF-seq, we generated CNV profiles of normal diploid and highly rearranged cancer genomes. The mapped fragments were grouped into variable length "bins" across the genome and bin counts were used to generate CNV profiles as described in (Baslan et al., 2012; Kendall and Krasnitz, 2014).

### 3.5.1 Accurate CNV profiles using SMURF-seq

We sequenced a normal diploid female genome with SMURF-seq, resulting in 270.8k reads (mean read length of 6.75 kb) in a single run. These reads were split into 7.28 million fragments (26.87 mean fragments per read; Fig. 3.6). A CNV profile for this normal diploid genome, with the expected (approximately flat) appearance can be seen in Fig. 3.7a. A replicate of this experiment

resulted in 497.9k reads (mean read length of 3.7 kb), which were split into 7.55 million fragments (15.16 mean fragments per read).

The Rapid sequencing kit form Oxford Nanopore Technologies offers an extremely fast (10 minute) and simple (2 step) protocol for library preparation. We verified that the SMURF-seq procedure behaves similarly using the Rapid Sequencing Kit. The 213.38k sequenced reads had a mean read length of 3.9 kb, and were split into 2.81 million fragments.

Next we applied SMURF-seq to the breast cancer line SK-BR-3, generating 147.0k reads with

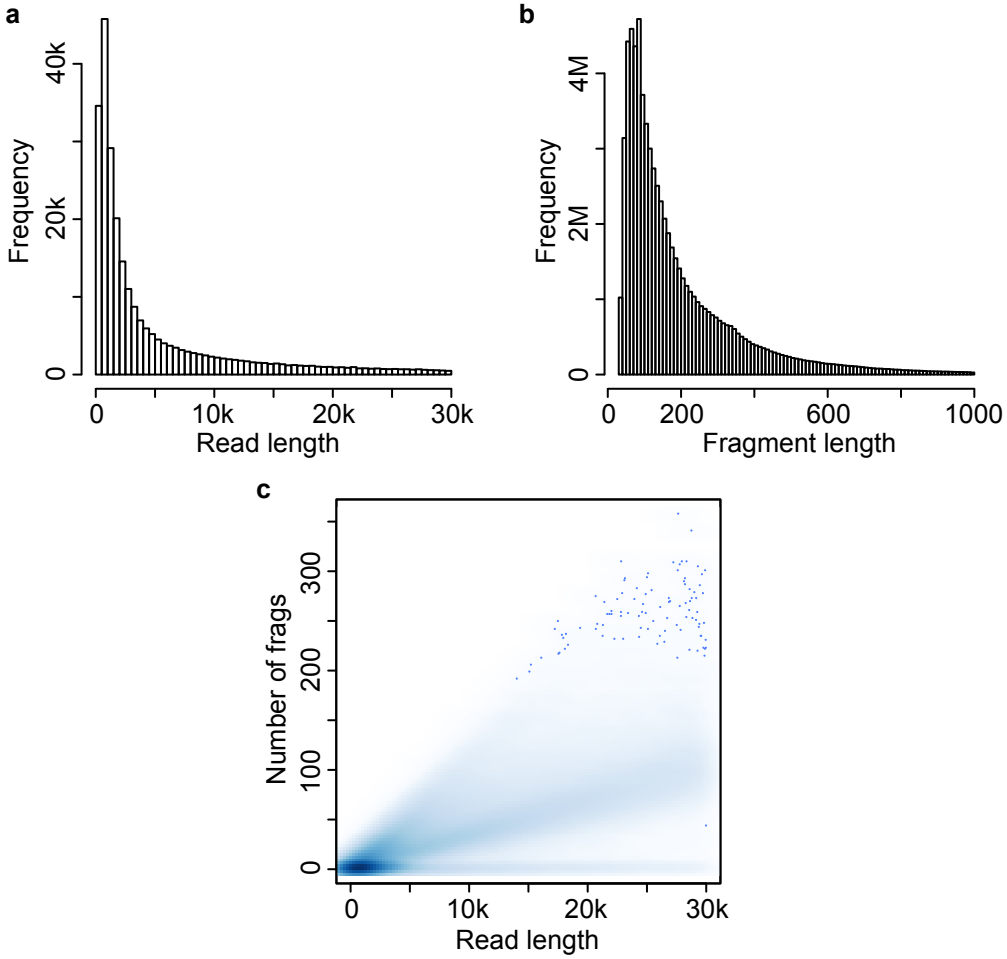


Figure 3.6: Read and fragment lengths from a SMURF-seq sequencing run. (a) Sequenced read length distribution. (b) Mapped fragment length distribution. (c) Scatter plot of read length and the number of fragments contained in the read.

mean length of 7.62 kb, which were split into 4.52 million fragments (30.78 mean fragments per read). We then obtained a CNV profile using 5,000 bins, corresponding to an average bin size of approximately 600 kb (Fig. 3.7b).

To provide a quantification of accuracy in terms of individual CNV events we conducted whole-genome sequencing (WGS) on the same SK-BR-3 using Illumina (5.56 million reads; 130 bp, single-end). We used this to define a ground truth by calling CNV events for each of the pre-

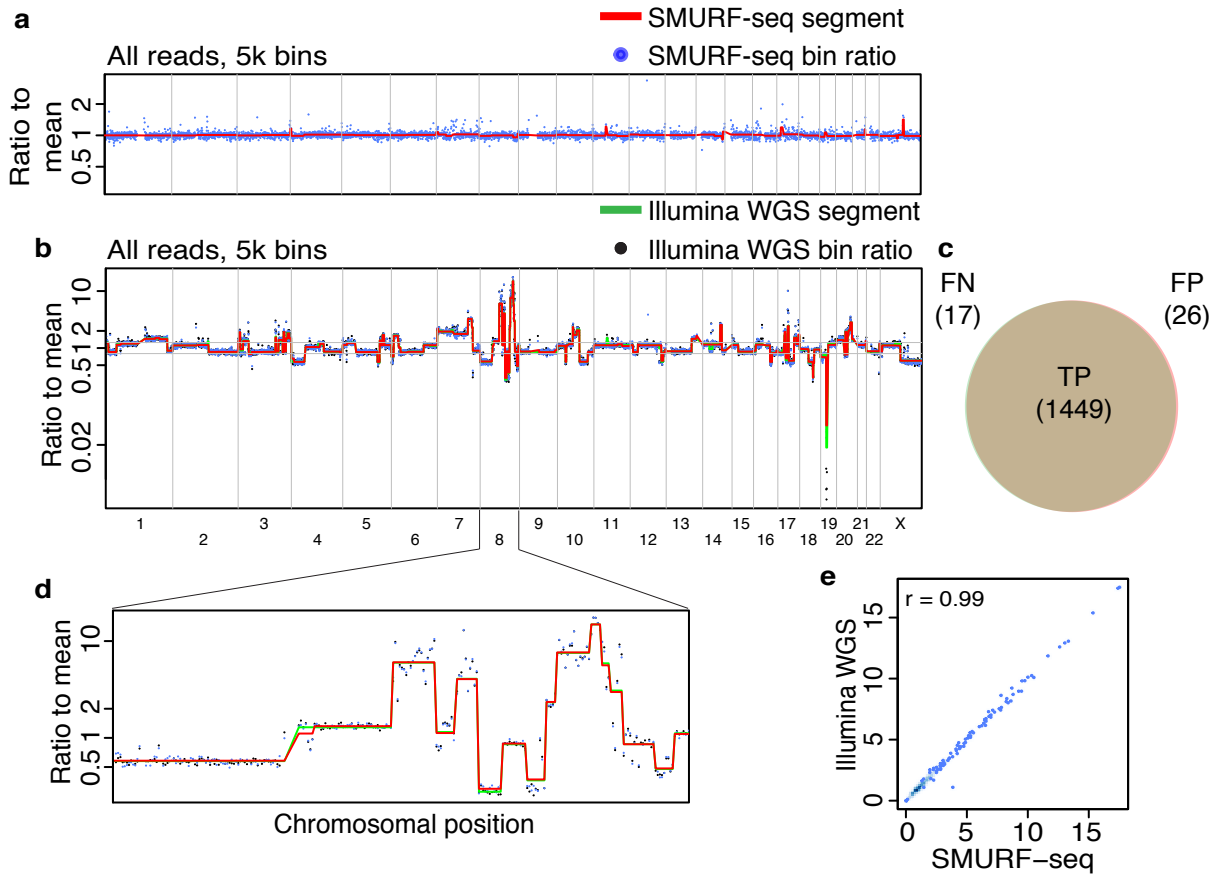


Figure 3.7: Accurate copy number profiles with SMURF-seq. (a) CNV profile of a normal diploid genome. Each blue point is a bin ratio to mean and the red line is the segmented bin ratio. (b) Superimposed CNV profiles of SK-BR-3 genome generated using SMURF-seq and Illumina WGS reads. (c) Venn diagram illustrating the accuracy of event calls using SMURF-seq compared with Illumina WGS. (d) Zoom-in of copy number changes on chromosome 8. (e) Scatter plot of bin ratio of SK-BR-3 genome using SMURF-seq and Illumina WGS reads. Pearson correlation of the data is shown.

defined bins (both amplifications and deletions) based on segmented signal with a cutoff of 1.25/0.8 (Fig. 3.7b) (Berry et al., 2017; Dago et al., 2014). This resulted in 1,466 events (886 amplifications, 580 deletions) from 4,953 bins. We then called events using the identical procedure with SMURF-seq data from the same SK-BR-3 sample. The precision and recall for SMURF-seq relative to the Illumina calls was 0.982 and 0.988, respectively (Fig. 3.7c). Fig. 3.7d shows a zoom-in of a region with extreme copy number alterations. The bin ratios for the Illumina WGS and the SMURF-seq profiles are highly correlated (Pearson $r$ = 0.99; Fig. 3.7e). A replicate of this experiment resulted 132.64k reads (mean read length of 7.3 kb), which were split into 4.02 million fragments (30.31 mean fragments per read).

We also generated higher-resolution CNV profiles at 20,000 and 50,000 bins, corresponding to an average of approximately 150 kb and 60 kb in length respectively (Fig. 3.8a, b). The profiles obtained at these resolutions have a high correlation with the profiles obtained using Illumina WGS (Pearson $r > 0.97$; Fig. 3.8c, d).

### 3.5.2 Concordant profiles from fewer countable fragments

Several cancer-related studies have employed CNV profiling based on low-coverage WGS (Kader et al., 2016; Macintyre et al., 2018). It has previously been demonstrated that 250k reads are sufficient for accurate genome-wide CNV profiling of single cells (Baslan et al., 2015). At the same time, the CNV profiles from a population of cells has been shown to have a high correlation with single-cell profiles (Baslan et al., 2015; Navin et al., 2011). We reasoned that using 250k fragments for CNV profiling using a population of cells would give useful profiles if they remained sufficiently accurate. By down-sampling our SMURF-seq data, we verified that 10k reads, approximately 250k fragments, result in highly-correlated CNV profiles (Pearson $r$ = 0.98; Fig. 3.9a, b).

Given the total capacity of the MinION instrument, this indicates that multiple samples can effectively be barcoded and multiplexed in a single sequencing run. To verify this we sequenced

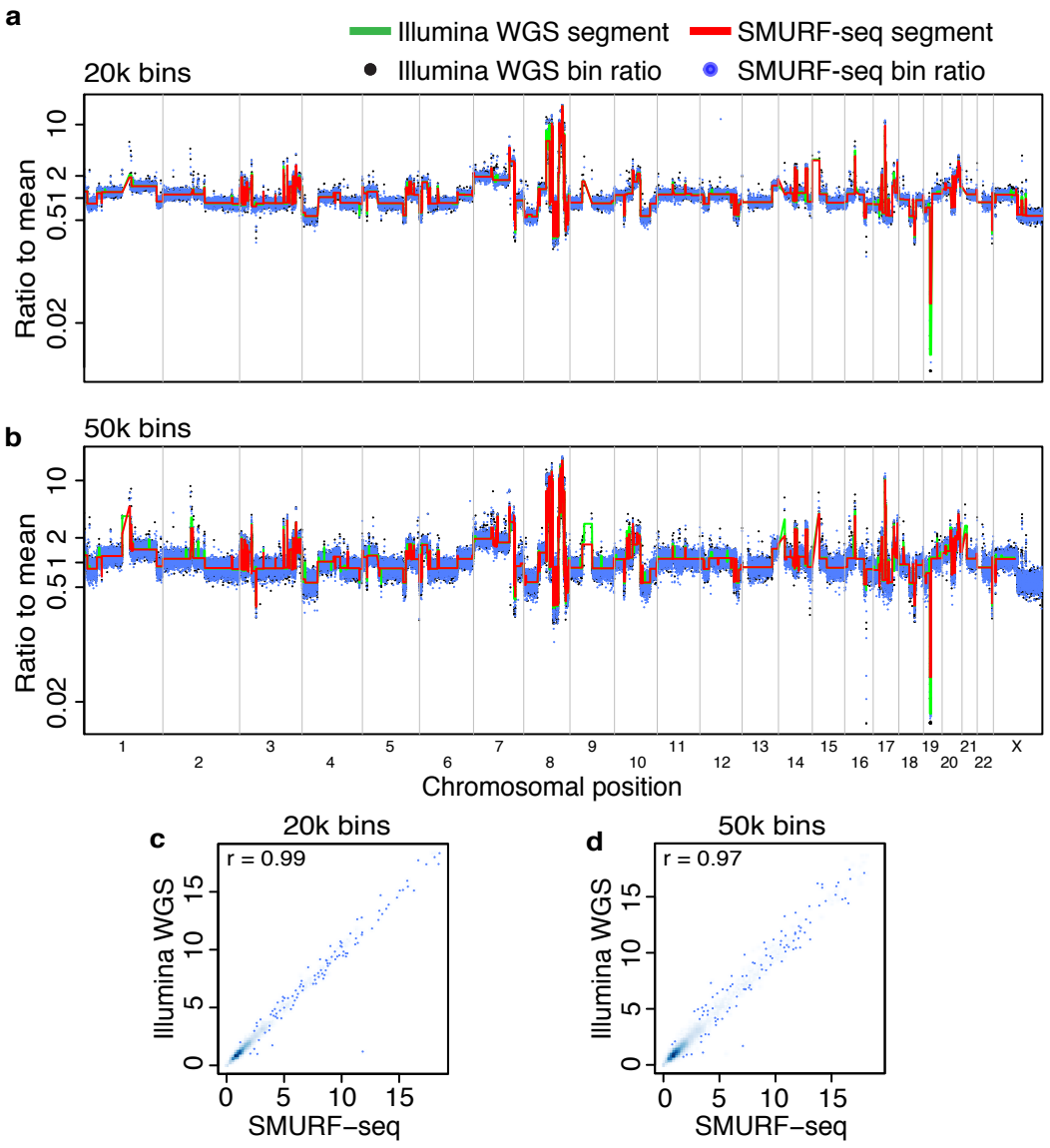


Figure 3.8: High resolution CNV profile generated using SMURF-seq is highly concordant with the profile generated with Illumina WGS. (a, b) Superimposed CNV profiles of SK-BR-3 genome generated using SMURF-seq and Illumina WGS at 20,000 and 50,000 bin resolutions. (c, d) Scatter plot of bin ratios of SK-BR-3 genome using SMURF-seq and Illumina WGS reads at 20,000 and 50,000 bin resolutions.

two DNA samples (normal diploid female and SK-BR-3) in a single run. These samples were processed with SMURF-seq protocol and then barcoded following the standard library construction. After demultiplexing and mapping the reads, the diploid genome had a CNV profile as expected

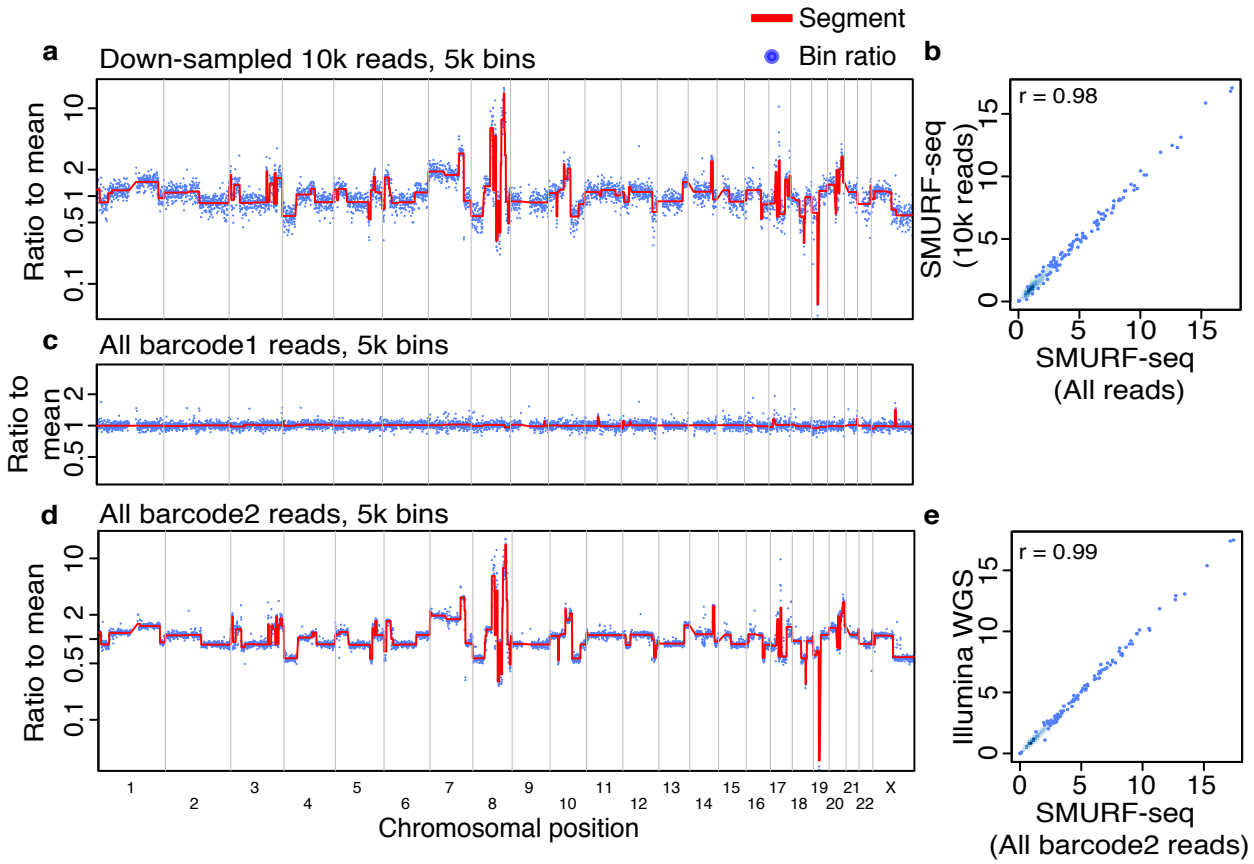


Figure 3.9: Multiple SMURF-seq CNV profiles by multiplexing in a single run. (a) CNV profile of SK-BR-3 genome with down-sampled 10k SMURF-seq reads. (b) Scatter plot of normalized bin counts of the original SMURF-seq data and data down-sampled to 10k SMURF-seq reads. Pearson correlation of the data is shown. (c) CNV profile of barcode01 (Normal diploid genome) reads. (d) CNV profile of barcode02 (SK-BR-3 cancer genome) reads. (e) Scatter plot of bin ratios of SK-BR-3 genome using multiplexed SMURF-seq and Illumina WGS reads.

(Fig. 3.9c) and the SK-BR-3 CNV profile was nearly identical to the profile obtained using Illumina WGS (Pearson $r$ = 0.99; Fig. 3.9d, e). All the sequencing runs and the availability of sequence data is summarized in Appendix B.

Further, we verified that the CNV profile with reads generated in the first 45, 90, and 180 minutes of starting a sequencing run had a high correlation to the profile with reads from the complete run (Pearson $r > 0.98$; Fig. 3.10).

In summary, our results demonstrate that SMURF-seq can generate more information for CNV

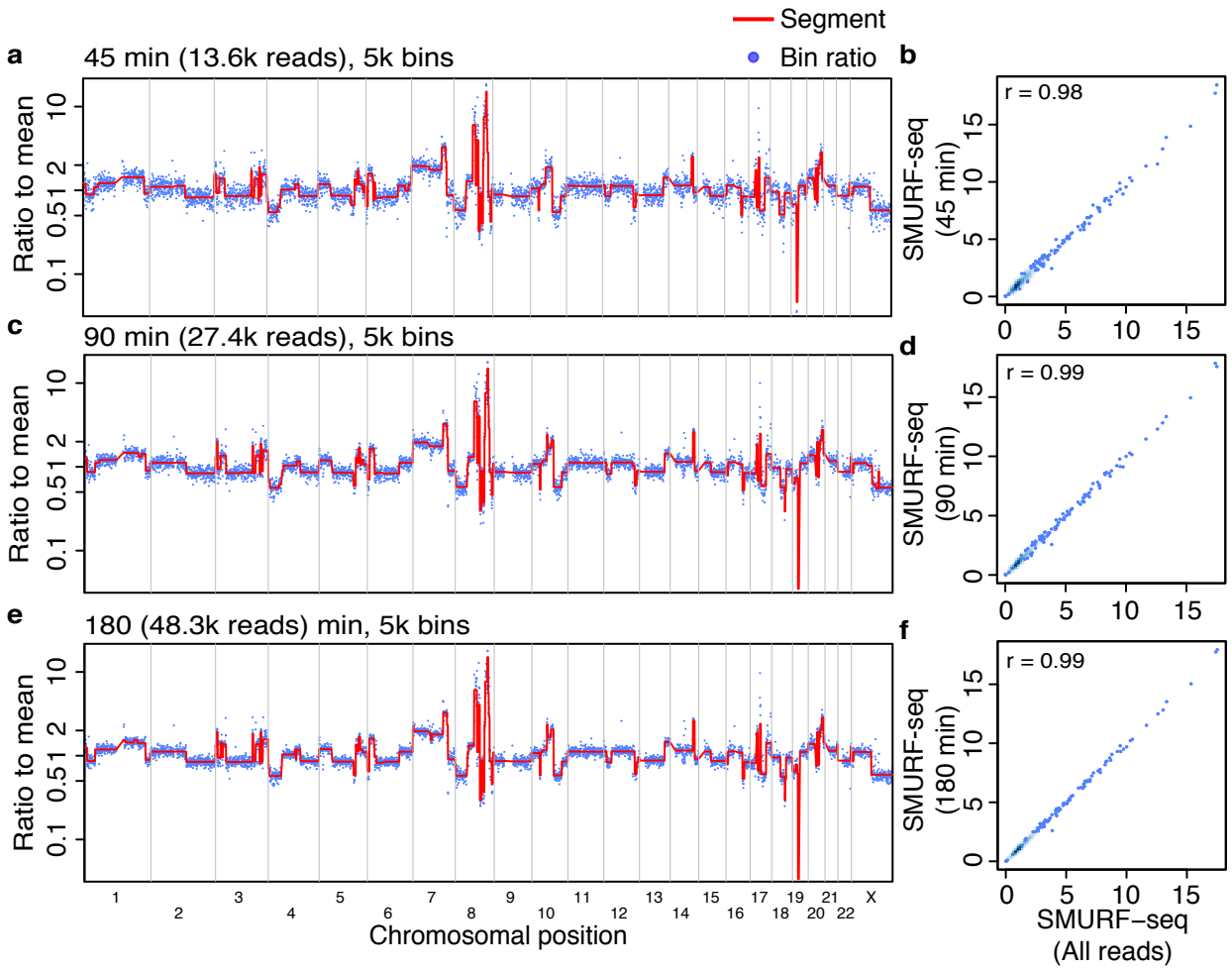


Figure 3.10: CNV profile with reads obtained in first few minutes of sequencing. (a, c, e) CNV profile with reads obtained in the first 45, 90, and 180 minutes of sequencing. (b, d, f) Scatter plot of bin ratios of the original SMURF-seq data and data obtained in first 45, 90, and 180 minutes of sequencing.

analysis in a single run of the Oxford MinION sequencer, compared with either producing long reads in the usual way or direct short-read sequencing on the same instrument. This increased information is in the form of increased numbers of distinct DNA fragments sequenced, and can be leveraged in multiple ways. Applying SMURF-seq on a single sample for a full run corresponds to higher counts for downstream analysis. In CNV analysis, increased counts either add confidence for a fixed resolution, or can allow higher resolution analysis (i.e. smaller bins) at the same level

of confidence. Alternatively, the increased information throughput can effectively reduce the time required to produce the same number of counts for CNV analysis by terminating the sequencing earlier. Finally, the increased information yield can be directed towards reducing the cost of generating CNV profiles by allowing a greater degree of multiplexing. For CNV analysis at resolutions permitted by 250k mapped fragments, our results show SMURF-seq allows roughly 20 and up to 30 samples in a single run, compared with 10 per run directly using short-read sequencing.

## 3.6 Future of SMURF-seq

# Chapter 4

# Identifying fragment boundaries on a SMURF-seq read

## 4.1 Motivation

New sequencing methods motivate development of new algorithms for mapping and analysis of sequences generated using these methods. A few significant developments include BLAST (Altschul et al., 1990) and FASTA (Pearson and Lipman, 1988) motivated by database searches with the advent of Sanger sequencing, BWA (Li and Durbin, 2009) and Bowtie (Langmead et al., 2009) inspired by high-throughput short-read sequencing, and BLASR (Chaisson and Tesler, 2012) by single-molecule long-read sequencing. SMURF-seq has enabled efficient short-read sequencing for read-counting applications on portable long-read machines. However, efficient methods tailored for mapping SMURF-seq reads are still lacking; Especially as SMURF-seq protocol evolves and the fragments become shorter, and thus, making the mapping process challenging in terms of identifying accurate fragment locations and boundaries.

As currently implemented, SMURF-seq protocol uses a single restriction enzyme (SaqAI) to fragment DNA molecules to $\sim$150 bp. However, depending on the downstream application, the

fragment lengths need to be just long enough to ensure unique mappability to a sufficient fraction of the genome. Fragments could be made shorter using methods discussed in section **??**. As an example, for copy-number profiling (at low resolutions, as used for tumor samples) the fragment lengths could be as short as 40 bp.

We used BWA-MEM (Li, 2013) to align SMURF-seq reads generated with the current protocol, which consists of fragments that are typically over 100 bp. Though not designed to align SMURF-seq reads, BWA-MEM is designed for split read alignment, and it works sufficiently well at these fragment lengths. SMURF-seq reads can also be aligned with other mapping tools capable of split-read alignment such as Minimap2 (Li, 2018) and LAST (Kiełbasa et al., 2011). However, all of these tools are either designed for aligning short reads with low sequencing error or long reads with high sequencing error.

Aligning SMURF-seq reads, especially as the fragments get shorter, would differ from these tools in the following aspects: (1) A seeding approach designed for short fragments sequenced with a high error-rate. (2) An approach to estimate the number of fragments on a SMURF-seq read and determine the optimal fragment boundaries.

The initial step of a typical alignment tool, called seeding, is to find candidate locations of a read on the reference genome, and limit the downstream steps to these locations. This is usually using a hash table based data structure to find exact matches (Altschul et al., 1990, 1997; Kent, 2002) or non-contiguous exact matches (Chen et al., 2009; Ma et al., 2002), or by using a suffix tree based data structure (Kurtz et al., 2004; Langmead et al., 2009; Li, 2013; Li and Durbin, 2009, 2010). The choice of data structure and parameters such as the length of the match, number of matches, etc. are determined by several factors such as the read length and the error profile of the underlying sequencing technology. For example, the optimal seeds parameters for expressed sequence tags and whole genome sequencing (prior to short-read high-throughput sequencing) was determined in BLAT (Kent, 2002), similarly for low-error rate short-reads in RazerS (Weese et al., 2009), and high-error rate long-reads in BLASR (Chaisson and Tesler, 2012). The fragments could be as short

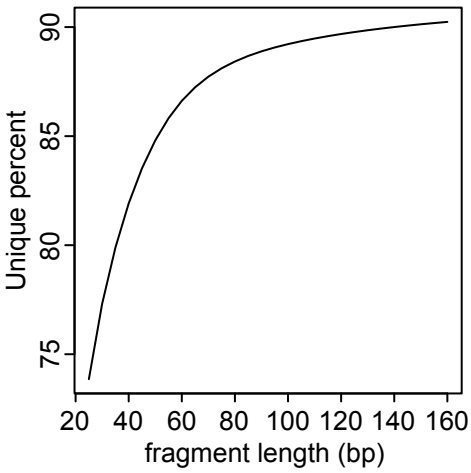


Figure 4.1: The fraction of genome that is uniquely mappable decreases with fragment length.

as 40 bp in a SMURF-seq read, and aligning these reads requires identifying candidate locations for these fragments in the presence of the characteristic error profile of nanopore machines.

The significance of having accurate fragment boundaries is understood by looking at the fraction of the genome that is uniquely mappable. For the human reference genome (hg19), when allowing no mismatches or indels, the fraction of the genome that is uniquely mappable reduces by 0.06% when going from 150 to 145 bp, whereas it reduces by 2.02% when going from 40bp to 35bp (Fig. 4.1). Thus, as the fragments get shorter, the probability of a fragment that originated from a unique location on the genome to misalign to an ambiguous location or vice-versa increases. Further, as sequencing errors are considered the difference in unique mappability due to having inaccurate fragment boundaries is likely to increase. Thus, as the fragments become shorter, having accurate fragment boundaries would improve the sensitivity of aligning SMURF-seq reads.

Although both seeding and determining the accurate fragmentation is crucial to align a SMURF-seq read, in this study we focus only on the second. To this end, we define the fragment identification problem for identifying the number of fragments and the fragment boundaries on a SMURF-seq read. We approach the fragment identification problem by defining a score function for aligning a SMURF-seq read, study the null score distribution of aligning reads and reference generated at

random, and estimate the number of fragments in a SMURF-seq read by comparing its alignment score with the null distribution for all possible fragmentations of a read. Then we show the accuracy of our method using from simulated genomes and SMURF-seq reads. Further, we empirically show that this method could also be used with a general score function.

## 4.2 Background

In the early days of DNA sequencing, as the number of nucleotides sequenced grew, comparison of DNA sequences became an indispensable tool to a biologist. DNA sequence comparison can be broadly classified into global alignment (Needleman and Wunsch, 1970) and local alignment (Smith et al., 1981). A global alignment seeks an optimal alignment between two sequences such that each base of one sequence is aligned to each base of the other sequences. On the other hand, a local alignment seeks an optimal alignment between any subsequences of the sequences being compared.

Comparison of two sequences, even unrelated or random sequences, always produces an optimal alignment. This motivated the development of approaches to differentiate a "meaningful" alignment from alignment of unrelated sequences. These methods determine the significance of an alignment by comparing the alignment score with a null distribution of alignment scores of unrelated sequences. Determining the appropriate null distribution was the subject of an enormous amount of research, some of which are summarized below.

In the context of local alignment, at the time of the initial studies on the score distribution of unrelated sequences, mathematical tools to understand the null distributions were still lacking, and these studies relied on empirical distributions generated from aligning unrelated sequences. In (Smith et al., 1985), it is shown that the similarity score is proportional to the logarithm of the length of the sequences being compared, and the standard deviation is independent of the sequence length. The significance of an alignment was determined from the number of standard

deviations over mean of the alignment score. These studies (Lipman et al., 1984) also highlight that the statistical properties (Smith et al., 1983), such as nucleotide frequencies or codon usage, of the sequences affect the distribution of the alignment scores. Generating a null distribution from an incorrect model could lead to an alignment of unrelated sequences being dubiously declared significant. Several methods are available to generate random sequences preserving these statistical properties (Altschul and Erickson, 1985; Fitch, 1983).

Erdos and Renyi (Erdös and Révész, 1975) presented results for the length of the longest headrun in a the first $n$ tosses of a biased coin. The length of the longest headrun in coin tosses is equivalent to the number of matches between two DNA sequences when shifts in the starting and ending positions of the sequences are not allowed, with the probability of head equal to the probability of match between letters of the DNA alphabet. In (Arratia and Waterman, 1985), this is generalized to matches between DNA sequences, while allowing shifts. These results indicate that allowing shifts doubles the length of the longest headrun. Results for the longest headrun allowing for up to $k$ mismatches and sequences generated from a Markov chain are also considered. In (Arratia et al., 1986; Gordon et al., 1986) the distribution of the longest matches is shown to a have an extreme value distribution with mean that is proportional to the logarithm of the sequences lengths and variance independent of sequence length. Here, when considering only matches, the asymptotic extreme value distribution is shown by considering a maximum of geometric distributions, and when mismatches are allowed, it is shown by considering a maximum of negative binomial distributions. An alternate approach is a Poisson approximation for the distribution of the longest match (Arratia et al., 1989).

An crucial aspect of in aligning nucleic acid and protein sequences is using the appropriate score function. For example, PAM (Dayhoff et al., 1978) and BLOSUM (Henikoff and Henikoff, 1992) matrices are commonly for protein sequences. The score function used alters the score of the aligned sequences and thus the alignment score distribution of unaligned sequences. However, the approach based on the length of the headruns does not consider the score function used for an

alignment. In (Karlin and Altschul, 1990; Karlin et al., 1990), it is shown that the maximal score of aligning unrelated sequences using any score function (that has at least one positive score and the expected score is negative) takes the form of an extreme value distribution, and and explicit formulas that its parameters are provided. Further, the number of high-scoreing alignments is closely approximated by the Poisson distribution. Thus, enabling the calculation of the probability that an alignment of random sequences having a score greated than any given value (usually the score of aligning two sequeneces of interest).

## 4.3 Fragment Identification problem

Let $\Sigma$ be an alphabet. A string $X$ is a sequence of letters $a_0 a_1 \dots a_{n-1}$, where $a_i \in \Sigma$; $|X|$ denotes the length of the string $X$; and $X[i \dots j] = a_i \dots a_{j-1}$ is a substring of $X$.

The reference string $T$ is generated from the DNA alphabet $\Sigma = \{A, T, G, C\}$, with $|T| = n$. A SMURF-seq read $S$ is generated by concatenating substrings (called fragments) of $T$, with no information available *a priori* about the number, length, orientation (forward or reverse-complement), and the position on $T$ of these fragments. Further, $S$ contains sequencing errors with a rate $\rho$. Let $|S| = m$ and $m \ll n$.

A fragment set $P$ is an set of start locations of fragments on $S$. $P \subset \{0 \dots m-1\}$ and $|P| = k$, with the rule that $0$ is in $P$ always. By convention we consider the set $P$ to be ordered such that if $i < j$ then $P_i < P_j$. For a fragment set $P$, $\sum_{i=1}^{k} P_{i+1} - P_i = m$ and we say that the $i^{\text{th}}$ of $S$ is the substring $S[P_i \dots P_{i+1}]$, with $P_{k+1} = m$.

For a given $T$ and $S$, the fragment identification problem is to determine the elements of the fragment set $P$ such that it corresponds to the start locations of fragments contained in $S$.

## 4.4 Approach to the fragment identification problem

We approach the fragment identification problem by defining a score function as follows: For a given fragment set $P$, we define the score of aligning $S$ to $T$ as:

$$score_T(S, P) = \sum_{i=1}^{k} \max\{score(T[u \dots v], S[P_i \dots P_{i+1}]) : 0 \leq u < v \leq n\}.$$

This allows us to consider the fragment identification problem as two inter-related problems: (1) Determining $k$, the size of the fragment set, and (2) given $k$, determining the elements of $P$ such that $score_T(S, P)$ is maximized.

By the score function defined above, to determine the elements of the fragment set $P$, requires the knowledge of the number of fragments $k$ and this is not known *a priori*. Further, the $k$ that maximizes the score function would almost never correspond to the optimal fragment set. As an example, taking $k = m - 1$ which corresponds to taking each base as a fragment would maximize the score, however, this is a non-sensical alignment.

We propose to estimate the number of fragments $k$ by aligning a read to the reference genome with different $k$. For each of these fragmentations, we determine the p-value by comparing the alignment score with the null distribution generated from aligning reads generated at random to a reference genome generated at random. Finally, we choose the fragmentation with lowest p-value as the optimal fragmentation.

The fragment identification problem differs from the alignment problems described in section 4.2 in a crucial manner. For the fragment identification problem we have the reference genome, and it is assumed that the reads always arise from this genome; the score distribution of sequences generated at random is used to determine the optimal number of fragments on a SMURF-seq read. Whereas in the context of local alignment the score distribution of aligning random reads are used to determine a "meaningful" alignment by comparing the alignment score of sequences with the random null distribution.

## 4.5 Aligning SMURF-seq reads and identifying fragment boundaries

Having a score function for SMURF-seq reads enables a statistical approach to estimate the number of fragments on a read. The proposed method depends on an algorithm that can identify the optimal fragment boundaries, given the number of fragments, such that the sum of the alignment score of each fragment is maximized.

### 4.5.1 Fragment boundary identification under exact matching

We first examine the fragment identification problem assuming the score function requires exact matching

$$score(a, b) = \begin{cases} 1 & \text{if } a = b \\ -\infty & \text{otherwise.} \end{cases}$$

The fragment identification problem then becomes an exact matching problem where the goal is to minimize the number of fragments such that $score_T(P, S)$ is maximized.

A simple linear time solution to this problem can be obtained as follows. First, we assume some data structure for $T$ has been constructed in linear time and allows for longest prefix matches to be computed in time proportional to the length of the query string. The data structure could be a suffix tree (McCreight, 1976), or a more space efficient and a modern structure like an FM-index (Ferragina and Manzini, 2000). The principle of the algorithm can be seen by starting at the beginning of $S$, and identifying the longest prefix match of $S$ in $T$. Then retain $j$ as the first position of where this longest prefix matches in $T$, and denote the first mismatching position on $S$ as $i$. Repeat the procedure solving the subproblem of fragment identification for $S[i \dots m]$. Repeating these steps, the algorithm iteratively solves the longest prefix match problems, retaining as $P_{i+1}$ the position of mismatch that terminates matching during iteration $i$. The following pseudocode

describes the procedure.

**Algorithm 1** ExactFragmentMatching($T, S$):

```
1: i ← 0
2: while i < m do
3:       P ← P ∪ {i}
4:       i ← longest-match-length(S, i, T)
5: return P
```

**Proof:** Consider an optimal solution to this problem, where the identified fragment set $P_{\text{opt}}$ has minimal size. To prove the optimality of our algorithm we need to show that it finds the same number of fragments as the optimal solution, i.e. $|P| = |P_{\text{opt}}|$.

The first iteration of the greedy algorithm will find the longest prefix match. If the optimal solution has its first fragmentation ending before $P_1$, i.e. $P_{\text{opt1}} < P_1$. Then the longest match starting at $P_{\text{opt1}}$ will end at or before $P_2$, the end of the second fragment found by the greedy algorithm. If it ends at $P_2$ then the greedy algorithm has the same number of fragments as the optimal solution so far. And it cannot end before $P_2$, because then the optimal solution will have more fragments than found by the greedy algorithm. Moreover, we cannot have $P_{\text{opt1}} > P_1$ as this would imply a longer prefix than found by the longest prefix match exists. With this reasoning we can say that this greedy approach will find just as little fragments as the optimal solution.

### 4.5.2 Fragment boundary identification allowing mismatches and indels

Let $M$ denote a table with $m+1$ rows, $n+1$ columns and $k+1$ dimensions, where $k$ is the maximum number of fragments ($1 \leq k \leq m$). $\max_{0 \leq j \leq n} M(i, j, l)$ represents the best fragmentation of $S[1 \ldots i]$ with $l$ fragments. The entries of $M$ are computed as follows:

**Algorithm 2** FragBoundaryIdentification$(T, S, k)$

$M(i, j, 0) \leftarrow -\infty$ for all $0 \leq i \leq m, 0 \leq j \leq n$
$M(0, j, 1) \leftarrow 0$ for all $0 \leq j \leq n$
$M(i, 0, 1) \leftarrow M(i-1, 0, 1) + score(S[i], \_)$ for all $0 \leq i \leq m$
$M(l-1, j, l) \leftarrow -\infty$ for all $2 \leq l \leq k, 0 \leq j \leq n$
$M(i, 0, l) \leftarrow -\infty$ for all $2 \leq l \leq k, l \leq i \leq m$
**for** $l \leftarrow 1$ **to** $k$ **do**
    **for** $i \leftarrow l$ **to** $m$ **do**
        **for** $j \leftarrow 1$ **to** $n$ **do**

$$M(i, j, l) \leftarrow \max \begin{cases} M(i-1, j-1, l) + score(S[i], T[j]) \\ M(i-1, j, l) + score(S[i], \_) \\ M(i, j-1, l) + score(\_, T[j]) \\ \max_{0 \leq h \leq n} M(i-1, h, l-1) + score(S[i], T[j]). \end{cases}$$

**Time and space complexity:** Each entry of $M$ is computed in constant time by storing the value of $\max_{0 \leq j \leq n} M(i-1, j, l-1)$ for every row of $M$ in a separate array. The algorithm runs in $O(knm)$ time and uses $O(knm + km)$ space, where the additional $O(km)$ is used to store the values of $\max_{0 \leq j \leq n} M(i-1, j, l-1)$.

The optimal alignment and fragment boundaries are determined from the usual traceback procedure starting from $\max_{1 \leq j \leq n} M(m, j, k)$ and ending in $M_{1 \leq j \leq n}(1, j, 1)$, with the exception of storing if a new fragment maximized the score at a cell.

Intuitively, this algorithm is similar to the local alignment algorithm but instead of picking an empty alignment when the score of an extension is negative, this algorithm starts a new fragment when the score of extending a fragment is less than score of starting a new fragment. In terms of an alignment graph, each node has a zero-weight incoming edge from the node corresponding to $\max_{0 \leq j \leq n} M(i-1, j, l-1)$, in addition to the weighted match/mismatch and indel edges (Fig. 4.2).

Although this algorithm provides the exact solution to determine optimal fragment boundaries for each $k$, it has several limitations for use with real SMURF-seq reads in practice. As discussed in section **??**, when mapping SMURF-seq reads the approximate mapping locations of fragments would be known after seeding. Further, if a mapping algorithms performs a preliminary align-

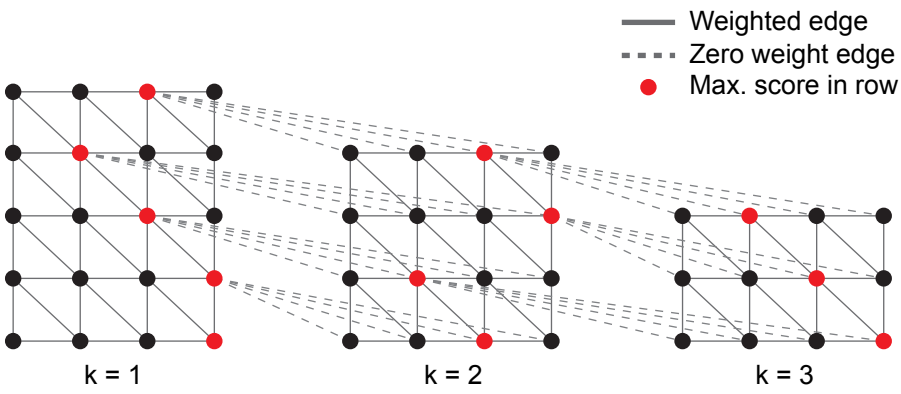


Figure 4.2: Alignment graph for fragment boundary identification algorithm with an arbitrary score function. The direction of arrows are omitted for clarity. The horizontal edges are directed from left to right and all other edges are directed from top to bottom.

ment between parts of the SMURF-seq read and the reference genome, the approximate fragment boundaries will also be know.

### 4.5.3 Identifying fragment boundaries in practice

The initial step in any traditional mapping algorithm is to determine the approximate mapping locations of the read on the reference genome using a seeding approach. Some algorithms further generate proto-alignments using the initial seeds. For example, BWA-MEM joins seeds that are close to on the read and reference coordinates and have the same orientation into a "chain" (). Typically, the final step is to perform a Smith-Waterman alignment between the read and a small region on the reference genome.

Any algorithm to aligning SMURF-seq reads can be expected to have similar steps. After seeding a SMURF-seq read, seeds from different fragments on a read would cluster at different locations on the genome, likely several different location for each fragment due to repeats or seeding parameters (such as seed length). These seed hits could be further processed to generate preliminary alignments between parts of the read and the reference genome. Thus, after these steps it is reasonable to assume that approximate number of fragments and the boundaries of these fragments

are known.

The final step of the algorithm would be finalize the number of fragments and fragment boundaries on a read. For a given number of fragments, the fragment boundaries can be easily determined such that the alignment score is maximized (this problem is equivalent to finding a longest path in a single-source directed acyclic graph). For finding the number of fragments on a SMURF-seq read, for each fragmentation, we use the alignment score distribution of random sequences with the same fragment set to determine the p-value for a fragmentation using the procedure described in the next section.

## 4.6 Alignment score of a SMURF-seq read

The alignment score of a SMURF-seq read given by equation **??** is the sum of the alignment score of each of individual fragment for a particular fragmentation with $k$ fragments, and algorithms given in section **??** determine the best fragmentation so that the alignment score is maximized.

The alignment score of a SMURF-seq read is a non-decreasing function with the number of fragments, as the score function does not penalize for increasing the number of fragments on a read. Thus, a SMURF-seq read aligned with $k$ fragments can always be aligned with $k + 1$ with a score at least as high as with $k$ fragments.

As an example of how the alignment score of a SMURF-seq read grows as a function of $k$, consider a read consisting of $f$ fragments (typically $> 20$ for a SMURF-seq read), however, $f$ is not known *a priori*. If the read is aligned to the reference genome as one fragment, it is likely going to be aligned to a random location on the reference genome; Alternatively, one of the fragments in the read could align close to its true location, and the flanking fragments would align to flanking regions on the genome (essentially aligning fragments to a random location on the genome for these flanking fragments). Similarly, if the read is aligned as two fragments, it is likely going to align to two random location on the genome, or at most two fragments could align to regions close

to their true location. With the alignment score with two fragments at least as high as with one fragment. As the read is aligned with increasing $k$ up to $f - 1$, the number of the number of bases on a read that are aligned to its true location on the genome will increase, and number of based aligned to random location would decrease. Although there will always bases that are not mapped to its true location. At $k = f$ all the fragments on the read would be mapped to their true locations with no bases aligned to random locations. As $k$ increased beyond $f$, each fragment will be split into shorter fragments, the fragmentation location likely being at locations of sequencing errors (with a small increase in the alignment score). And if there are no more sequencing errors, the split can be anywhere on the read, with the alignment score remaining the same. At $k = m$, i.e. each base on the read aligns as one fragment with the highest possible alignment score.

Consider a simulated SMURF-seq read from a genome of length 50 kb, the read has 20 fragments each of length 40 bp, and the read does not have any sequencing errors. Fig. 4.3a shows the alignment score as a function of the number of fragments $k$, with scoring a match as 1, a mismatch as 0, and not allowing indels. At $k = 1$, the score is at the lowest, and increases with $k$. At $k = 20$ every fragment is mapped to it true location, and since the read does not have any sequencing errors, the score is equal to the read length. For $k > 20$, the score remains at the maximum.

When a SMURF-seq read has sequencing errors, the alignment score increases with the number of fragments, but would not reach the peak at $k = 20$ (Fig 4.3b; read similar to 4.3a, but with 10% mismatch errors). The alignment score continues to increase beyond $k = 20$ but at a lower rate, and would eventually equal the read length as the number of fragments increase. Similarly, Fig. 4.3c ducts the alignment score for a read with 20% mismatch errors. However, the due to higher errors, the score is lower at $k = 20$, and the curve "flattens".

For a real SMURF-seq read, the fragment lengths cannot expected to be constant. Fig. 4.3d shows the score of of read with 20 fragments, with each fragment of lengths sampled from a $20 + \text{Geom}(20)$ (mean fragment length of 40 bp) distribution. The alignment score has increase, but the change in slope at $k = 20$ is not prominent.

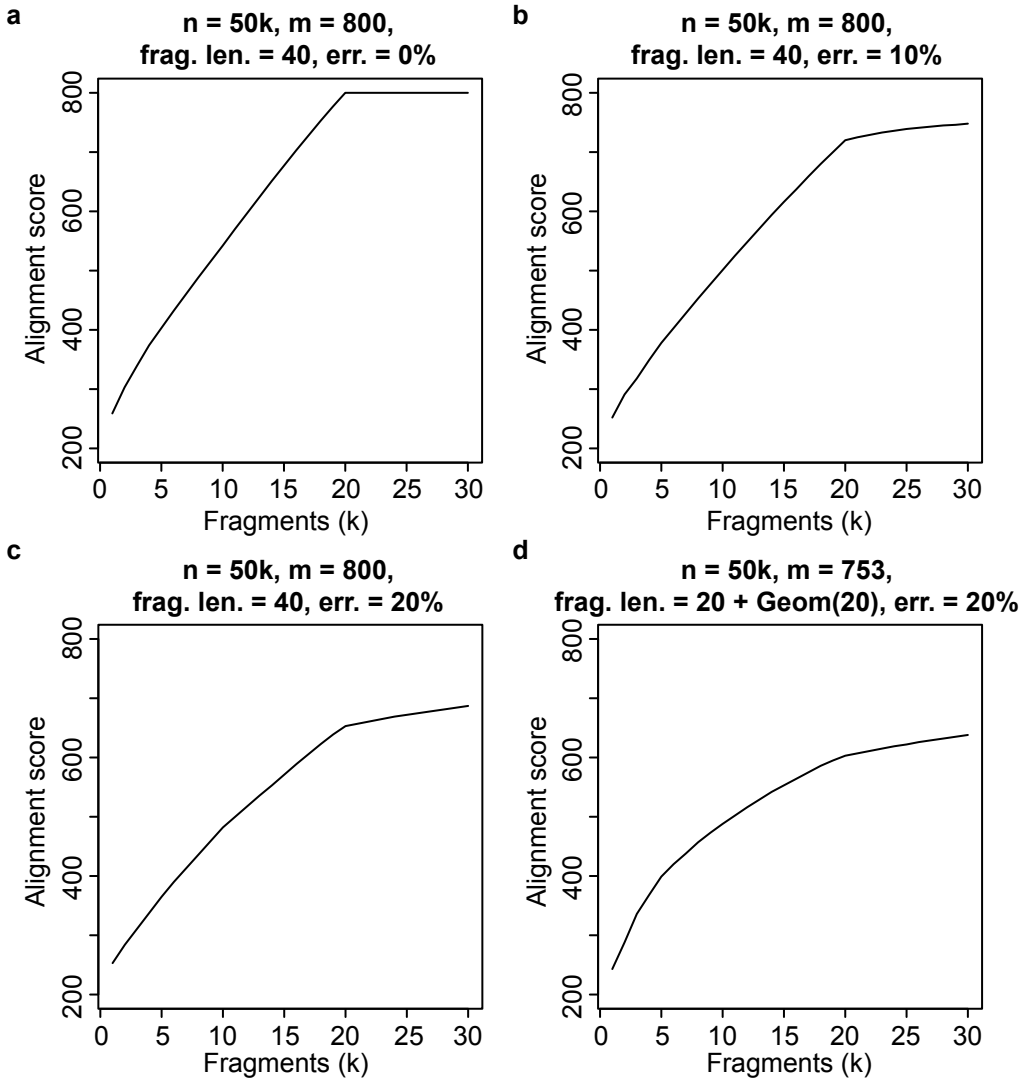


Figure 4.3: Alignment score of SMURF-seq read as a function of number of fragments. (a) Alignment score of a SMURF-seq read with 20 fragment (40 bp each) that does not have any errors. (b) Alignment score with 10% errors. (c) Alignment score with 20% errors. (d) Alignment score of a read with 20 fragments generated from $20 + \text{Geom}(20)$ distribution and with 20% errors.

## 4.7 Score distribution under a random model

Calculation of p-value for aligning a SMURF-seq read with a given fragmentation requires the null distribution of aligning reads generated at random with the same with the same fragmentation. The problem of finding the null distribution is defined as: consider strings $T$ and $S$ are generated by

drawing letters independently from the same distribution from an alphabet $a \in \Sigma$ with probability $p_a$ such that $\sum_{a \in \Sigma} p_a = 1$. For a given fragment set $P$ containing $k$ elements, we need to determine the distribution of $score_T(S, P)$. We use the following score function to obtain the distribution of $score_T(S, k)$:

$$score(a, b) = \begin{cases} 1 & \text{if } a = b \\ 0 & \text{if } a \neq b \\ -\infty & \text{otherwise.} \end{cases}$$

To determine the distribution of $score_T(S, P)$, we first consider the score distribution when $k = 1$, i.e. the entire read aligns as one fragment. Then, we consider the score distribution when $k > 1$ as the sum of $k = 1$ distributions. We also empirically show that the form of the null distribution when using a generalized scoring function is similar to the distribution obtained with score function defined above.

### 4.7.1 Score distribution of one fragment

The score distribution of $score_T(S, 1)$ has similarities to the score distribution of local alignment () and profile alignment (), but also differs from these. The distribution of $score_T(S, 1)$ differs from the local alignment as we require an end-to-end alignment of $S$ to a substring of $T$, and also differs from the profile score distribution since the letters of $S$ are generated at random. Based on these results, the distribution of $score_T(S, 1)$ is likely to follow an extreme value distribution.

Let $X_j$ denote the score of aligning $S$ with $T[j \dots j + m - 1]$, then

$$X_j = \sum_{i=0}^{m-1} score(S[i], T[j + i]), j = 0, \dots, n - m + 1.$$

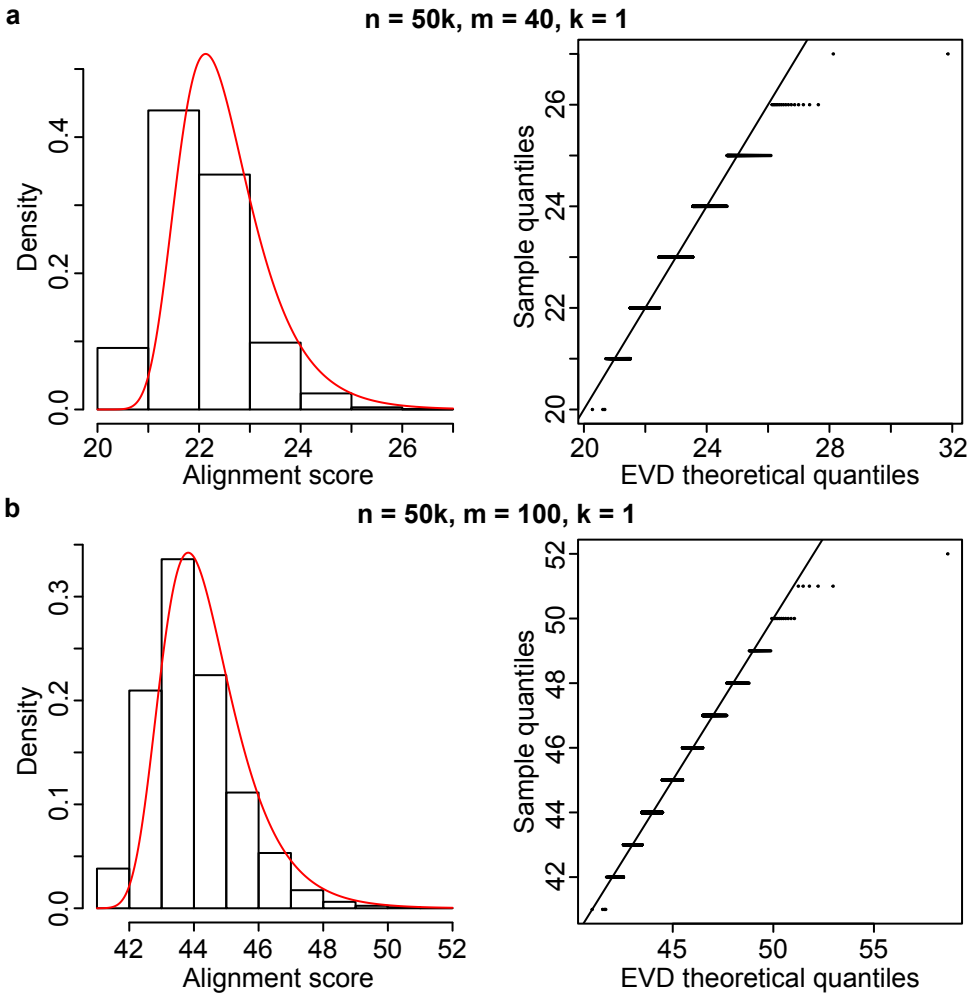


Figure 4.4: Extreme value approximation for $score_T(S, 1)$. Empirical score distribution of $score_T(S, 1)$ with a fitted EVD using the method of moments estimator. Q-Q plot comparing the theoretical and empirical distributions are shown. (a) $m = 40$. (b) $m = 100$.

Since the letters of T and S are iid, we have

$$X_j \sim binom(m, p)$$

where $p = \sum_{a=\Sigma} p_a^2$. For a large enough $m$, $X_j$ can be approximated by a normal distribution as

$$X_j \sim N(mp, mp(1-p)).$$

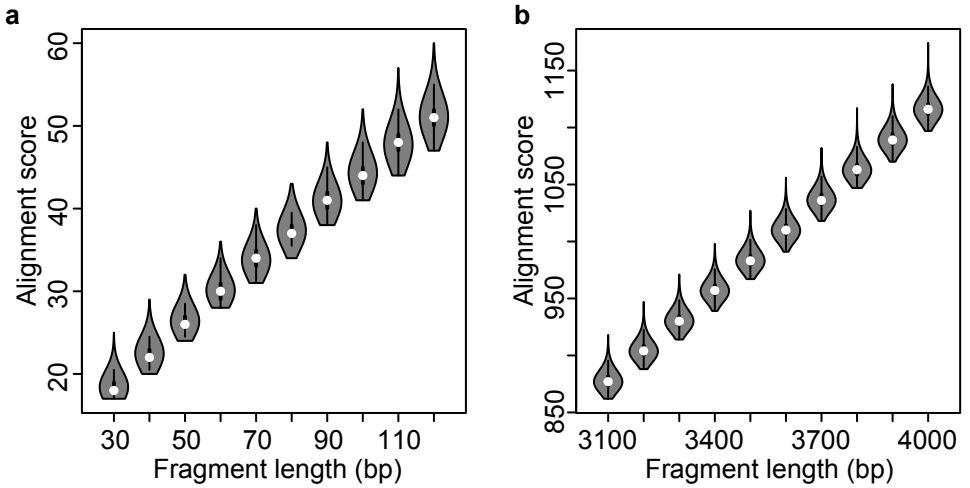


Figure 4.5: Empirical score distribution approximation for $score_T(S, 1)$. (a) Empirical score distribution for $m$ corresponding to shorter fragments. (b) Empirical score distribution form $m$ corresponding to longer fragments.

$score_T(S, 1)$ is the maximum score over all positions in $T$,

$$score_T(S, 1) = \max_{0 \leq j \leq n-m+1} X_j.$$

$score_T(S, 1)$ is a maximum of normal distributions, which follows an extreme value distribution (EVD) (Kotz and Nadarajah, 2000). Here, we have a dependence between $X_j$ and $X_k$ for $|j - k| < m$.

We verified score distribution of $score_T(S, 1)$ by generating a random genome of length 50kb from the DNA alphabet with equal probabilities, and reads of length 40 bp and 100 bp. For each read length, we determined the score distribution by aligning 10,000 reads generated at random (Fig. 4.4a, b). The parameters for the EVD was estimated using the method of moments. Further, the score distribution for increasing read lengths shows an increasing trend in the mean and standard deviation of the distribution (Fig. 4.4c, d).

### 4.7.2 Score distribution for a given fragment set

The distribution of $score_T(S, k)$ for $k > 1$ and a given $P$ is the sum of $k$ independent distributions of $score_T(S, 1)$, i.e the distribution of $score_T(S, k)$ is the sum of $k$ independent extreme value distributions

$$score_T(S, k) = \sum_{i=1}^{k} score_T(S[P_i \dots P_{i+1}], 1).$$

The independence of the distributions for each fragment is justified because it is required that $n >> m$, and the probability of two fragments to aligning to overlapping location on $T$ is extremely small.

The distribution of the sum and linear combination of extreme value distributions has been studied (Cetinkaya et al., 2001; Loaiciga and Leipnik, 1999; Marques et al., 2015; Nadarajah, 2008). In (Loaiciga and Leipnik, 1999) the exact distribution of two independent Gumbel distributions is given and in (Nadarajah, 2008) the exact distribution of the linear combination of Gumbel distributions is given. However, these distributions do not follow a "standard" distribution.

Since each the distribution of score of each fragment is independent, when the fragments are of equal lengths, the distribution of $score_T(S, k)$ is a sum of i.i.d. random variables. Thus, we can apply the central limit theorem to approximate the score distribution to a normal distribution as $k \to \infty$. To test the convergence to $score_T(S, k)$ to normal, we compared the score distribution to the normal distribution for $k$ that are typical for a SMURF-seq read (Fig. 4.6). The fragments lengths for all the comparisons were kept constant at 40 bp, and the parameters for the normal distribution was determined using the method of moments estimator.

In aligning a SMURF-seq read, we cannot expect the fragment lengths to be equal. The distribution of $score_T(S, k)$, when the fragment lengths differ, is a sum of independent, but not identical, random variables. We empirically verified that this distribution converges to normal (Fig. 4.7). For each $k$, the fragment lengths were generated from at random from a geometric distribution.

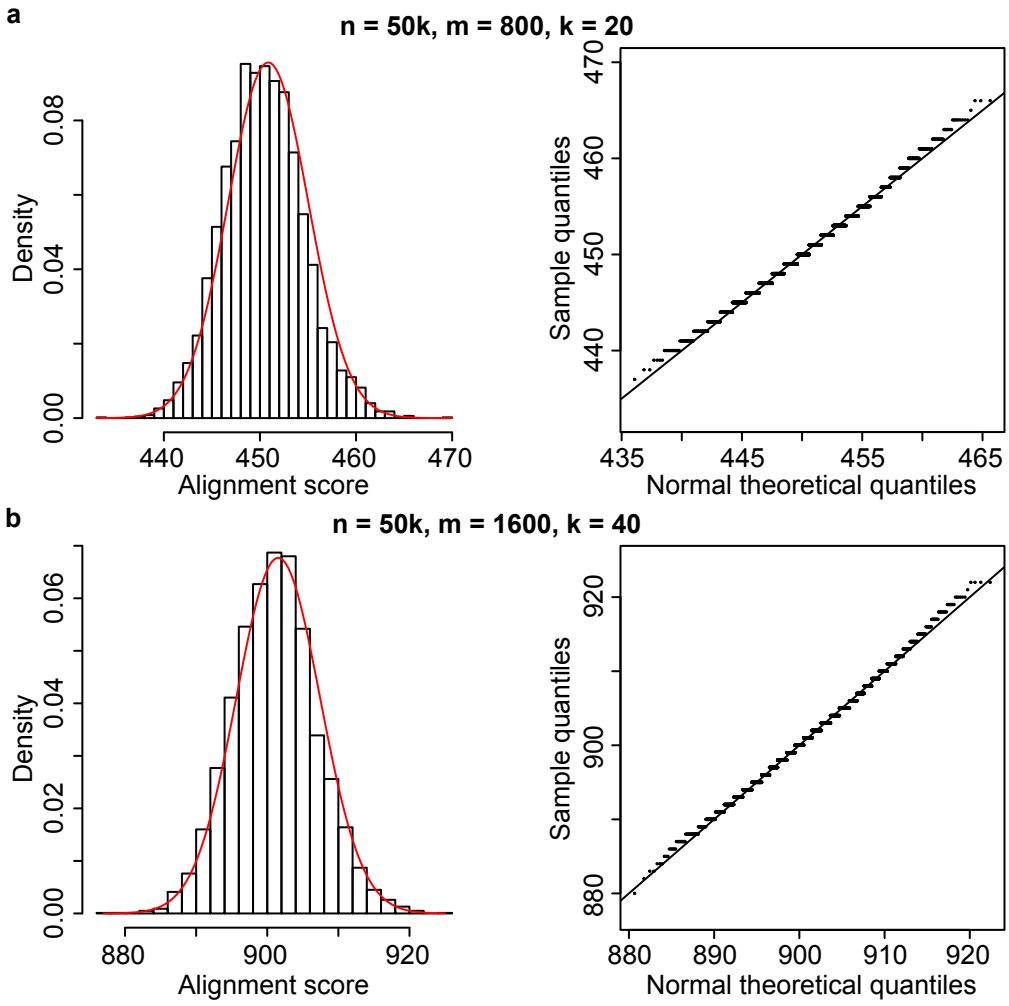


Figure 4.6: Normal approximation for $score_T(S, k)$ with equal fragment lengths. Empirical score distribution of $score_T(S, k)$ with a fitted normal using the method of moments estimator. All fragments are 40 bp. Q-Q plot comparing the theoretical and empirical distributions are shown. (a) $m = 800, k = 20$. (b) $m = 1600, k = 40$.

## 4.8 Estimating the optimal fragment set

The score distribution of aligning a random read $S_{\text{rand}}$ to a random genome $T_{\text{rand}}$ can be used to estimate the optimal $k$ for aligning a SMURF-seq read $S_{\text{SMURF}}$ to a reference genome $T_{\text{ref}}$. For a SMURF-seq read and, find the best alignment score $k_{\text{score}}$ and the fragment set $k_{\text{P}}$ for all $k$ from 1 to $m$ using algorithms given in section **??**. In practice, $k$ can be restricted to a much smaller subset of possible values, and the goal is to determine a value of $k$ that best represents the fragments that

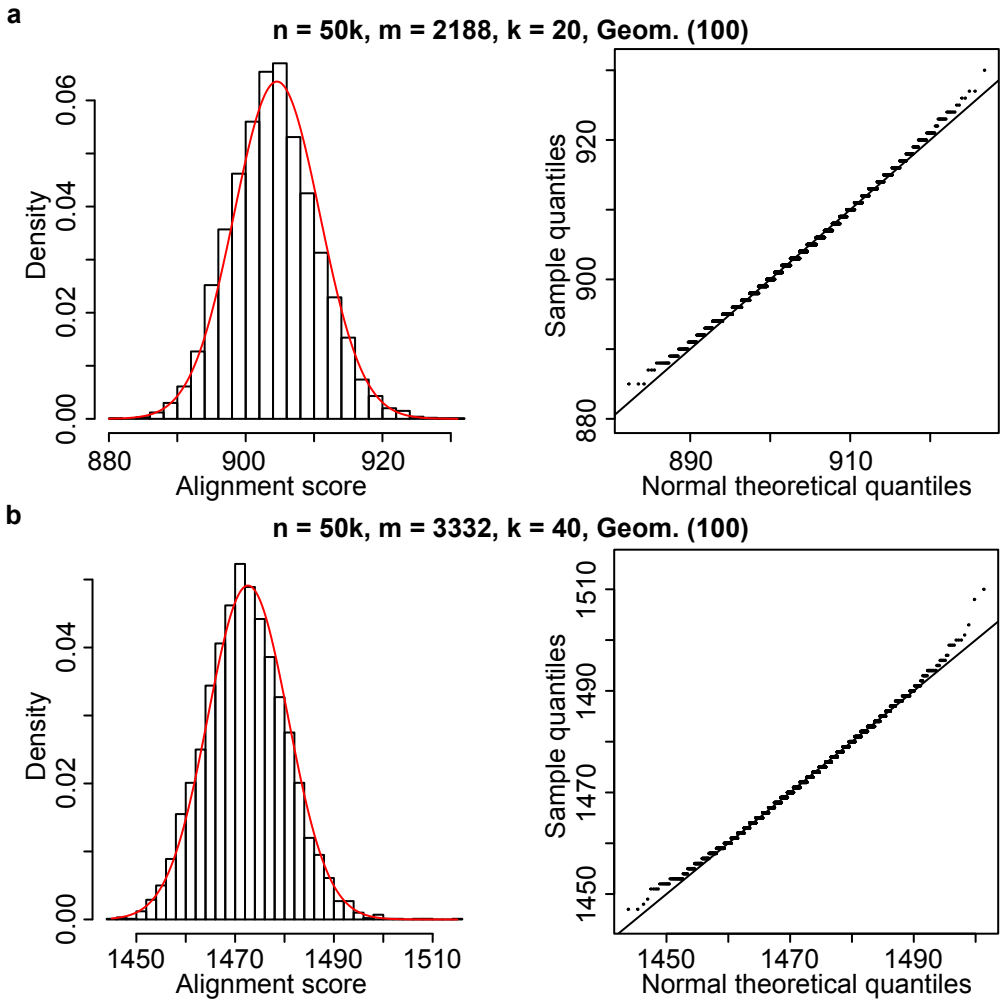


Figure 4.7: Normal approximation for $score_T(S, k)$ with random fragment lengths. Empirical score distribution of $score_T(S, k)$ with a fitted normal using the method of moments estimator. The fragment lengths are generated from a geometric distribution with mean 100. Q-Q plot comparing the theoretical and empirical distributions are shown. (a) $m = 2188, k = 20$. (b) $m = 3332, k = 40$.

were ligated to generated the read.

As described in the previous section, the null score distribution of aligning a read generated at random for $k = 1$ follows an extreme value distribution, and a normal distribution for sufficiently large $k$. For each $k$, the parameters for the null distributions can be estimated from the empirical distribution by aligning random reads with the fragment set $k_{\mathrm{P}}$. The p-value for each $k$ can be determined by finding the probability of a score greater than $k_{\mathrm{score}}$ from the null distribution. Finally,

the optimal $k$ for a read is the one with the lowest p-value (Algorithm 3).

**Algorithm 3** OptimalK $(T, S)$

$k_{\text{opt}} \leftarrow 1$
$\text{Pr}_{\text{opt}} \leftarrow 1$
**for** $k \leftarrow 1$ **to** $m - 1$ **do**
    $k_{\text{score}}, k_{\text{P}} \leftarrow FragMatch(T_{\text{ref}}, S_{\text{SMURF}}, k)$
    $k_{\text{Pr}} \leftarrow \text{Pr}(score_{T_{\text{rand}}}(S_{\text{rand}}, k_{\text{P}}) > k_{\text{score}})$
    **if** $k_{\text{Pr}} < \text{Pr}_{\text{opt}}$ **then**
        $\text{Pr}_{\text{opt}} \leftarrow k_{\text{Pr}}$
        $k_{\text{opt}} \leftarrow k$
**return** $k_{\text{opt}}$

As discussed in section **??**, as $k$ increases up to $k_{\text{opt}}$ the number of bases on a SMURF-seq that are aligned to its true location on the reference genome would increase and the bases aligned to random locations would decrease. Thus, getting further away from a random alignment with an expected decrease in p-value. At $k = k_{\text{opt}}$, all bases are aligned to their true locations, and would have the lowest p-value. As $k$ gets farther away from $k_{\text{opt}}$, fragments on a read are further split into smaller fragments with a small increase in alignment score, and eventually with fragment aligning to random locations on the reference. Thus, getting closer to a random alignment with an expected increase in p-value.

As an example, we simulated a reference genome of length 50 kb with the DNA alphabet having equal probabilities. A simulated SMURF-seq read with 20 fragments each of length 40 bp was generated from this genome, and 10% mismatch sequencing errors was introduced. This read are then mapped back to the reference genome for values of $k = 1$ to $30$ using algorithm **??** (scoring a match as 1, a mismatch as 0, and not allowing indels), yielding the fragment set that maximizes the alignment score for each $k$. These fragment sets were then used to generate the null distribution by simulating a random reference genome with the same base probabilities, and aligning 10,000 random reads with fragment start locations based on the fragment set. The p-value for each fragmentation was determined using an EVD for $k = 1$ and normal distributions for $k > 1$ with parameters estimated using the method of moments from the simulated reads (although

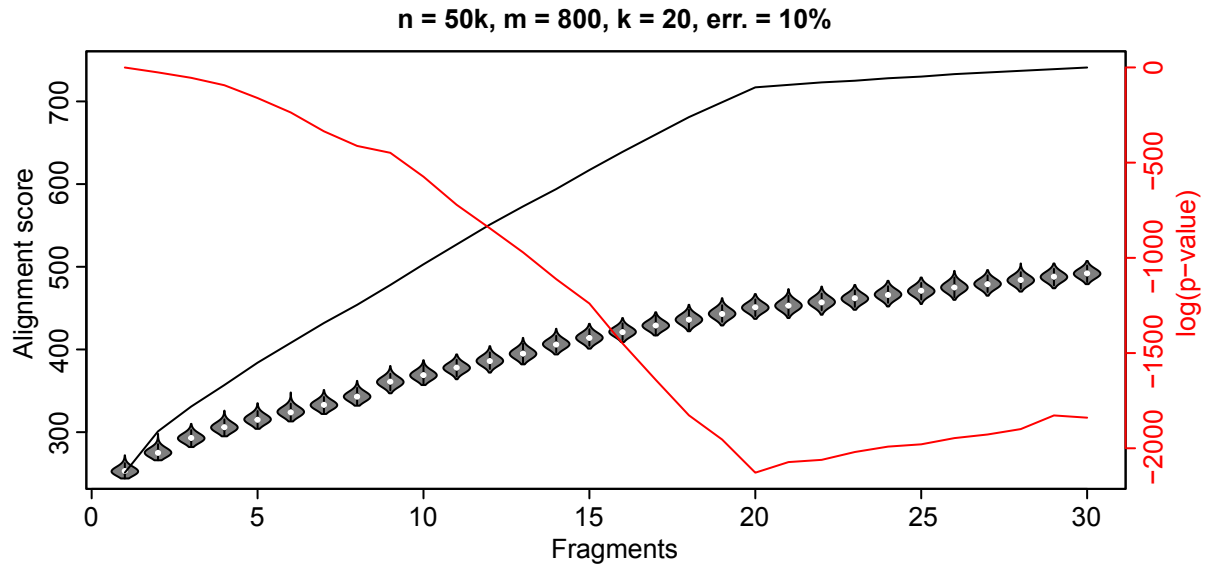


Figure 4.8: Determining the optimal fragmentation of a SMURF-seq read. The black line is the alignment score of a simulated SMURF-seq read with 20 fragments of 40 bp each, and with 10% mismatch errors. The violin plots are the empirical null distributions using fragment set corresponding the best alignment score for each $k$. The red line is the p-value for each $k$ determined from the alignment score and the null distribution. The optimal fragmentation has the lowest p-value.

the normal approximation does not hold for small values of $k$, we have included it for clarity). The fragmentation with the smallest p-value was considered as the optimal fragmentation, and as expected, $k = 20$ has the lowest p-value; With the p-value increasing on either side of $k = 20$ (Fig. 4.8).

The example considered here is simplified to illustrate the procedure to determine the optimal fragmentation of a SMURF-seq read. In aligning real SMURF-seq reads: (1) The reference genome would be significantly larger (e.g. the human genome) with different base (or dinucleotide) probabilities. (2) The fragments lengths cannot expected to be constant. (3) The sequencing error model will depend on the properties of the sequencing technology used. (4) The alignment score used would depend on the sequencing error model and would allow indels (with possibly different penalties for gap open and gap extend). (5) Parameters for the null distribution were determined

empirically, which cannot be done in practice. Some of these issues are considered in the following sections.

### 4.8.1 Fast computation of p-values

When aligning a SMURF-seq read, p-values need to calculated for a read to determine an optimal fragmentation among the potential fragmentations. It is thus important to be able to determine the p-value using an efficient procedure requiring the least computation. However, the procedure used above requires generating an empirical distribution for each fragmentation to determine the parameters for the null distribution. This process is computationally intensive and cannot be used in practice.

For any $k > 1$, the score distribution is the sum of $k$ independent extreme value distributions. An approach for fast computation of p-values is to compute the mean and standard deviation for the $k = 1$ distribution for all possible values of fragment lengths (i.e. 1 to maximum possible read length); Then the mean and standard deviation for the normal distribution, for any fragmentation with $k > 1$ fragments, can be calculated from the sum of the mean and variance of the $k$ $k = 1$ distributions correspond to the fragment set. As an example, for a fragmentation with three fragments of length 40, 100, 70 bp, the parameters for the normal distribution can be calculated as the sum of of the mean (and sd) of the distribution of aligning 40 bp, 100 bp, 70 bp fragments individually.

Thus, for a reference genome and a given score function, the parameters for the $k = 1$ determined once, and the p-value for any fragmentation can be computed by looking the table and adding the values. We tested the effectiveness of this procedure by computing the parameters for a genome of length 50 kb generated at random. Figure 4.9 compared null distribution generated by aligning a 100,000 random reads (empirical approach) and the null distribution computed using this method (analytic approach).

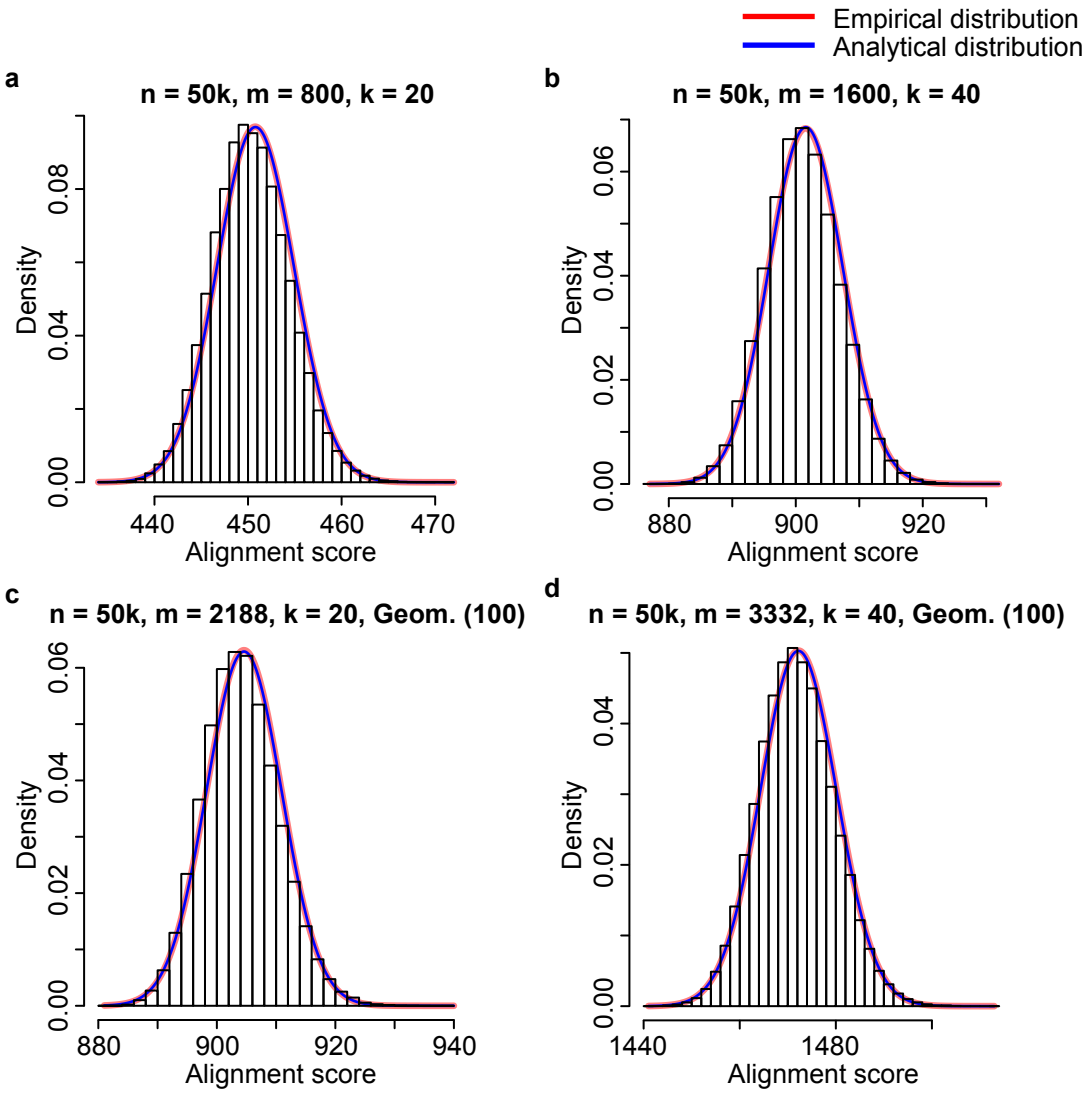


Figure 4.9: Fast computation of p-values as the as the sum of mean and variance of the $k$ $k = 1$ distributions correspond to the fragment set. (a-b) Reads with $k = 20$ and $k = 40$ fragments of constant 40 bp length. (c-d) Reads with $k = 20$ and $k = 40$ fragments of length generated at random from a $Geom.(100)$ distribution.

## 4.9 Limitations and future directions

As SMURF-seq evolves, the fragments can be expected to get shorter challenging the existing mapping tools. A crucial element of aligning SMURF-seq read is to determine the number of fragments on a SMURF-seq read and to determine the optimal fragment boundaries on a SMURF-

seq read. To this end, we defined a score function for a SMURF-seq read, suggested algorithms to find fragment boundaries, and provided a statistical procedure to estimate the number of fragments on a read. However, there are still several questions that remain to be answered in this regard, both in terms of using this procedure for real SMURF-seq reads and terms of understanding the score distribution of aligning SMURF-seq reads. Some of these are discussed in the sections below.

## Aligning with a general score function

In the analysis of aligning a SMURF-seq read and the random distributions, we used a simple score function of scoring a match as 1, a mismatch as 0, and not allowing indels. However, is a simplified score function, and in practice, it would depend on several factors such as the error profile of the sequencing technology used and the algorithms used to align these reads.

When scoring a mismatch with a negative penalty and allowing indels, $X_j$ denote the score of aligning a read $S$ to a substring of the reference $T[j \dots j+m-1]$ will not follow a binomial distribution. The extreme value distribution can be used to approximate the maximum of any independent distributions (i.e. not just the normal distribution). Thus, we hypothesized that

$$score_T(S,1) = \max_{0 \le j \le n-m+1} X_j.$$

could be approximated with an EVD irrespective of the score function used.

In the context of local alignment, the theoretical background for the score distributions for a random model were derived when indels were not allowed (); And it was shown empirically that allowing indels follows similar distributions as allowing only mismatches, although a theoretical proof does not exist ()

First, we verified empirically that the score distribution of for $k = 1$, $score_T(S,1)$, can still be approximated using an EVD when scoring a match as 1, and mismatch and indels scored as $-1$. We generated a random genome of length 50kb from the DNA alphabet with equal probabilities,

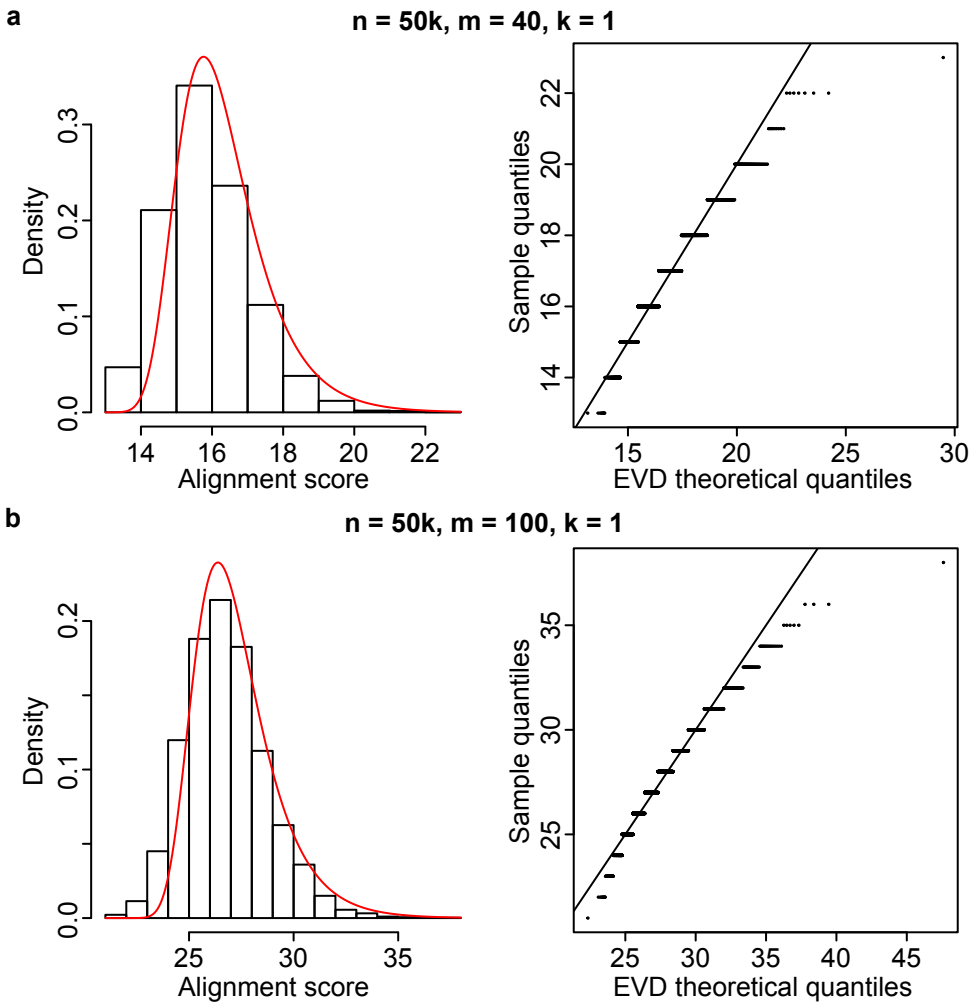


Figure 4.10: Extreme value approximation for $score_T(S, 1)$ with a general score function. A match is scored as 1, mismatch and indels as $-1$. Empirical score distribution of $score_T(S, 1)$ with a fitted EVD using the method of moments estimator. Q-Q plot comparing the theoretical and empirical distributions are shown. (a) $m = 40$. (b) $m = 100$.

and reads of length 40 bp and 100 bp. For each read length, we determined the score distribution by aligning 5,000 reads generated at random (Fig. 4.10a, b). The parameters for the EVD was estimated using the method of moments.

Then, we also verified empirically that the score distribution for $k > 1$, which is still the sum of $k$ independent extreme value distributions, is well approximated with a normal distribution. The fragments are 40 bp with 20 and 40 fragments per read (Fig. 4.11a, b).

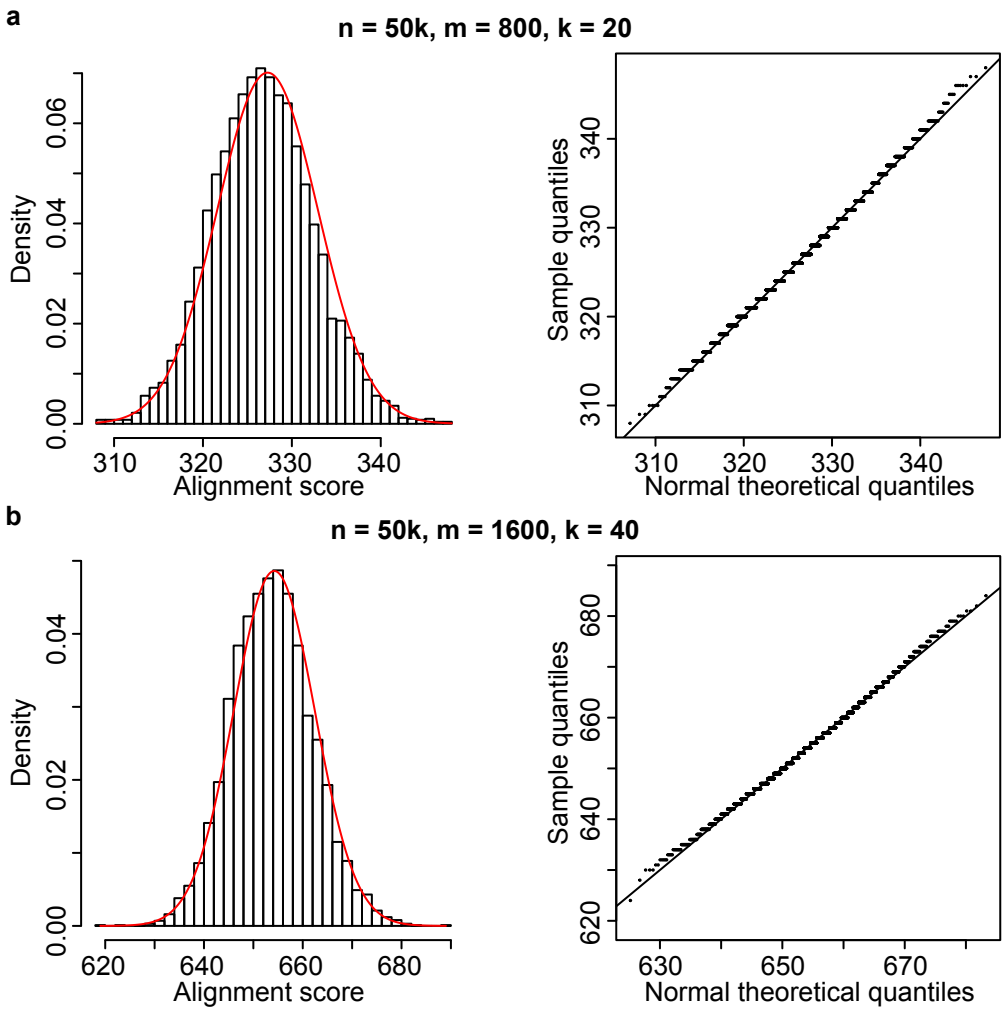


Figure 4.11: Normal approximation for $score_T(S, k)$ with a general score function. A match is scored as 1, mismatch and indels as $-1$. Empirical score distribution of $score_T(S, k)$ with a fitted normal using the method of moments estimator. All fragments are 40 bp. Q-Q plot comparing the theoretical and empirical distributions are shown. (a) $m = 800, k = 20$. (b) $m = 1600, k = 40$.

Although the extreme value and normal distributions are good approximations for the score function tested above, we have not tested other score functions. These approximations need to tested for the score function that is used to align SMURF-seq reads. Further, we have also not verified the validity of these approximation when gap open penalties are used in addition to gap extend and mismatch penalties.

## Errors in extreme value and normal approximation

Any approximation for the score distributions are approximations, and would have associated limitations. We have not assessed the limitations of these approximations.

Aligning a SMURF-seq read as one fragment is approximated with an extreme value distribution. Convergence to an EVD depends on the length of the genome, $n$. The minimum length of the genome with for convergence, and the error in approximation as a function of the genome length needed to be determined.

Aligning a SMURF-seq read with more than one fragment ($k > 1$) is approximated as a normal distribution, and the natural question is how large does $k$ have to be for a good approximation. Further, understanding the dependence of the number of fragments and the fragment lengths are also crucial for determining the effective less of calculation the p-value for a fragmentation. For example, aligning two reads with a same $k$ but one with shorter fragments and the other with longer could have to different error bounds for the normal approximation.

In aligning a real SMURF-seq read, the dependence of these approximation on the choice of score function used needs to evaluated.

## Effectiveness of p-value procedure

We determined the optimal fragmentation of a SMURF-seq read as the one with the lowest p-value. The effectiveness of this procedure for predicting the optimal fragmentation need to be evaluated. This could be done with simulated SMURF-seq reads for which the number of fragments and fragment boundaries are known. (Simulated SMURF-seq reads can be generated by sampling substrings from a reference genome and then sequencing errors could be added. They can also be generated by concatenating substring from a long read sequenced on a nanopore machine without SMURF-seq, and thus, preserving the error profile.)

Shorter fragments on a SMURF-seq read would lower the rate of increase of the alignment

score with increasing the number of fragments on the read. This makes predicting the optimal number of fragments on a read challenging. Similarly, an increase in sequencing errors on a read would also make the prediction challenging. Thus, the mispredictions of the optimal fragmentation of a read needs to evaluated based on these factors.

## Extending to a large genome

Here, we used a genome of length 50 kb generated with letters of the DNA alphabet. However, a refernece gneome for use with SMURF-seq can be expexted to be significaltly larger (e.g. the human genome). Using a larger genome

A random model used to deteremine the null distribution would have a significant effect on the p-value, and so the effectiveness of the procudure to determine the optimal number of fragments on a read. For example, in the context of local aligninments, using a ... has . Similarly, the effect of the these random models for SMURF-seq reads need to assessed.

## Theoretical proof for score distributions

## SMURF-seq read mapping tool

# Chapter 5

# Conclusions

# References


Mark Akeson, Daniel Branton, John J Kasianowicz, Eric Brandin, and David W Deamer. Microsecond time-scale discrimination among polycytidylic acid, polyadenylic acid, and polyuridylic acid as homopolymers or as segments within single rna molecules. *Biophysical journal*, 77(6):3227–3233, 1999.

Stephen F Altschul and Bruce W Erickson. Significance of nucleotide sequence alignments: a method for random sequence permutation that preserves dinucleotide and codon usage. *Molecular biology and evolution*, 2(6):526–538, 1985.

Stephen F Altschul, Warren Gish, Webb Miller, Eugene W Myers, and David J Lipman. Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–410, 1990.

Stephen F Altschul, Thomas L Madden, Alejandro A Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J Lipman. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic acids research*, 25(17):3389–3402, 1997.

Richard Arratia and Michael S Waterman. An erdös-rényi law with shifts. *Advances in mathematics*, 55(1):13–23, 1985.

Richard Arratia, Louis Gordon, Michael Waterman, et al. An extreme value theory for sequence matching. *The annals of statistics*, 14(3):971–993, 1986.

Richard Arratia, Michael S Waterman, et al. The erdös-rényi strong law for pattern matching with a given proportion of mismatches. *The Annals of Probability*, 17(3):1152–1169, 1989.

Nurit Ashkenasy, Jorge Sánchez-Quesada, Hagan Bayley, and M Reza Ghadiri. Recognizing a single base in an individual dna strand: a step toward dna sequencing in nanopores. *Angewandte Chemie International Edition*, 44(9):1401–1404, 2005.

Yann Astier, Orit Braha, and Hagan Bayley. Toward single molecule dna sequencing: direct identification of ribonucleoside and deoxyribonucleoside 5 -monophosphates by using an engineered protein nanopore equipped with a molecular adapter. *Journal of the American Chemical Society*, 128(5):1705–1710, 2006.

Timour Baslan, Jude Kendall, Linda Rodgers, Hilary Cox, Mike Riggs, Asya Stepansky, Jennifer Troge, Kandasamy Ravi, Diane Esposito, B Lakshmi, et al. Genome-wide copy number analysis of single cells. *Nature Protocols*, 7(6):1024, 2012.

Timour Baslan, Jude Kendall, Brian Ward, Hilary Cox, Anthony Leotta, Linda Rodgers, Michael Riggs, Sean D'Italia, Guoli Sun, Mao Yong, et al. Optimizing sparse sequencing of single cells for highly multiplex copy number profiling. *Genome Research*, 25(5):714–724, 2015.

Hagan Bayley. Nanopore sequencing: from imagination to reality. *Clinical chemistry*, 61(1): 25–31, 2015.

Seico Benner, Roger JA Chen, Noah A Wilson, Robin Abu-Shumays, Nicholas Hurt, Kate R Lieberman, David W Deamer, William B Dunbar, and Mark Akeson. Sequence-specific detection of individual dna polymerase complexes in real time using a nanopore. *Nature nanotechnology*, 2(11):718, 2007.

Jesse L Berry, Liya Xu, A Linn Murphree, Subramanian Krishnan, Kevin Stachelek, Emily Zolfaghari, Kathleen McGovern, Thomas C Lee, Anders Carlsson, Peter Kuhn, et al. Potential of aqueous humor as a surrogate tumor biopsy for retinoblastoma. *JAMA ophthalmology*, 135(11):1221–1230, 2017.

Rory Bowden, Robert W Davies, Andreas Heger, Alistair T Pagnamenta, Mariateresa de Cesare, Laura E Oikkonen, Duncan Parkes, Colin Freeman, Fatima Dhalla, Smita Y Patel, et al. Sequencing of human genomes with nanopore technology. *Nature communications*, 10(1):1–9, 2019.

Daniel Branton, David Deamer, Andre Marziali, Hagan Bayley, Steven Benner, Thomas Butler, Slaven Garaj, Andrew Hibbs, Xiaohua Huang, Stevan Jovanovich, Predrag Krstic, and Stuart Lindsay. The potential and challenges of nanopore. *Nature biotechnology*, 26:1146, 08 2009.

Clive G Brown and James Clarke. Nanopore development at oxford nanopore. *Nature biotechnology*, 34(8):810–811, 2016.

Tom Z Butler, Mikhail Pavlenok, Ian M Derrington, Michael Niederweis, and Jens H Gundlach. Single-molecule dna detection with an engineered mspa protein nanopore. *Proceedings of the National Academy of Sciences*, 105(52):20647–20652, 2008.

Coskun Cetinkaya, Vikram Kanodia, and Edward W Knightly. Scalable services via egress admission control. *IEEE Transactions on multimedia*, 3(1):69–81, 2001.

Mark J Chaisson and Glenn Tesler. Mapping single molecule sequencing reads using basic local alignment with successive refinement (blasr): application and theory. *BMC bioinformatics*, 13 (1):238, 2012.

Themoula Charalampous, Gemma L Kay, Hollian Richardson, Alp Aydin, Rossella Baldan, Christopher Jeanes, Duncan Rae, Sara Grundy, Daniel J Turner, John Wain, et al. Nanopore metagenomics enables rapid clinical diagnosis of bacterial lower respiratory infection. *Nature Biotechnology*, 37(7):783–792, 2019.

Yangho Chen, Tade Souaiaia, and Ting Chen. Perm: efficient mapping of short sequencing reads with periodic full sensitive spaced seeds. *Bioinformatics*, 25(19):2514–2521, 2009.

Gerald M Cherf, Kate R Lieberman, Hytham Rashid, Christopher E Lam, Kevin Karplus, and Mark Akeson. Automated forward and reverse ratcheting of dna in a nanopore at 5-å precision. *Nature biotechnology*, 30(4):344, 2012.

Derek Y Chiang, Gad Getz, David B Jaffe, Michael JT O'kelly, Xiaojun Zhao, Scott L Carter, Carsten Russ, Chad Nusbaum, Matthew Meyerson, and Eric S Lander. High-resolution mapping of copy-number alterations with massively parallel sequencing. *Nature Methods*, 6(1):99, 2009.

John Chu, Marcos González-López, Scott L Cockroft, Manuel Amorin, and M Reza Ghadiri. Real-time monitoring of dna polymerase function and stepwise single-nucleotide dna strand translocation through a protein nanopore. *Angewandte Chemie International Edition*, 49(52): 10106–10109, 2010.

James Clarke, Hai-Chen Wu, Lakmal Jayasinghe, Alpesh Patel, Stuart Reid, and Hagan Bayley. Continuous base identification for single-molecule nanopore dna sequencing. *Nature nanotechnology*, 4(4):265, 2009.

Scott L Cockroft, John Chu, Manuel Amorin, and M Reza Ghadiri. A single-molecule nanopore device detects dna polymerase activity with single-nucleotide resolution. *Journal of the American Chemical Society*, 130(3):818–820, 2008.

Angel E Dago, Asya Stepansky, Anders Carlsson, Madelyn Luttgen, Jude Kendall, Timour Baslan, Anand Kolatkar, Michael Wigler, Kelly Bethel, Mitchell E Gross, et al. Rapid phenotypic and genomic change in response to therapeutic pressure in prostate cancer inferred by high content analysis of single circulating tumor cells. *PloS ONE*, 9(8):e101777, 2014.

M Dayhoff, R Schwartz, and B Orcutt. 22 a model of evolutionary change in proteins. In *Atlas of protein sequence and structure*, volume 5, pages 345–352. National Biomedical Research Foundation Silver Spring MD, 1978.

David Deamer, Mark Akeson, and Daniel Branton. Three decades of nanopore sequencing. *Nature biotechnology*, 34(5):518, 2016.

Daniel P Depledge, Kalanghad Puthankalam Srinivas, Tomohiko Sadaoka, Devin Bready, Yasuko Mori, Dimitris G Placantonakis, Ian Mohr, and Angus C Wilson. Direct rna sequencing on nanopore arrays redefines the transcriptional complexity of a viral pathogen. *Nature communications*, 10(1):1–13, 2019.

Achilles Dugaiczyk, Herbert W Boyer, and Howard M Goodman. Ligation of ecori endonuclease-generated dna fragments into linear and circular structures. *Journal of molecular biology*, 96(1): 171–184, 1975.

Paul Erdös and Pál Révész. On the length of the longest head-run. *Topics in information theory*, 16:219–228, 1975.

Philipp Euskirchen, Franck Bielle, Karim Labreche, Wigard P Kloosterman, Shai Rosenberg, Mailys Daniau, Charlotte Schmitt, Julien Masliah-Planchon, Franck Bourdeaut, Caroline Dehais, et al. Same-day genomic and epigenomic diagnosis of brain tumors using real-time nanopore sequencing. *Acta Neuropathologica*, 134(5):691–703, 2017.

Nuno Rodrigues Faria, Ester C Sabino, Marcio RT Nunes, Luiz Carlos Junior Alcantara, Nicholas J Loman, and Oliver G Pybus. Mobile real-time surveillance of zika virus in brazil. *Genome medicine*, 8(1):97, 2016.

Paolo Ferragina and Giovanni Manzini. Opportunistic data structures with applications. In *Foundations of Computer Science, 2000. Proceedings. 41st Annual Symposium on*, pages 390–398. IEEE, 2000.

Lars Feuk, Andrew R Carson, and Stephen W Scherer. Structural variation in the human genome. *Nature Reviews Genetics*, 7(2):85–97, 2006.

Walter M Fitch. Random sequences. *Journal of Molecular Biology*, 163(2):171–176, 1983.

Jennifer L Freeman, George H Perry, Lars Feuk, Richard Redon, Steven A McCarroll, David M Altshuler, Hiroyuki Aburatani, Keith W Jones, Chris Tyler-Smith, Matthew E Hurles, et al. Copy number variation: new insights in genome diversity. *Genome research*, 16(8):949–961, 2006.

Daniel R Garalde, Elizabeth A Snell, Daniel Jachimowicz, Botond Sipos, Joseph H Lloyd, Mark Bruce, Nadia Pantic, Tigist Admassu, Phillip James, Anthony Warland, et al. Highly parallel direct rna sequencing on an array of nanopores. *Nature methods*, 15(3):201, 2018.

Erik Gerdtsson, Milind Pore, Jana-Aletta Thiele, Anna Sandström Gerdtsson, Paymaneh D Malihi, Rafael Nevarez, Anand Kolatkar, Carmen Ruiz Velasco, Sophia Wix, Mohan Singh, et al. Multiplex protein detection on circulating tumor cells from liquid biopsies using imaging mass cytometry. *Convergent Science Physical Oncology*, 4(1):015002, 2018.

Sara Goodwin, John D McPherson, and W Richard McCombie. Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, 17(6):333, 2016.

Jacqueline Goordial, Ianina Altshuler, Katherine Hindson, Kelly Chan-Yam, Evangelos Marcolefas, and Lyle G Whyte. In situ field sequencing and life detection in remote (79°26' n) canadian high arctic permafrost ice wedge microbial communities. *Frontiers in microbiology*, 8:2594, 2017.

Louis Gordon, Mark F Schilling, and Michael S Waterman. An extreme value theory for long head runs. *Probability Theory and Related Fields*, 72(2):279–287, 1986.

Brett Gyarfas, Felix Olasagasti, Seico Benner, Daniel Garalde, Kate R Lieberman, and Mark Akeson. Mapping the position of dna polymerase-bound dna templates in a nanopore at 5 å resolution. *ACS nano*, 3(6):1457–1466, 2009.

Philip J Hastings, James R Lupski, Susan M Rosenberg, and Grzegorz Ira. Mechanisms of change in gene copy number. *Nature Reviews Genetics*, 10(8):551–564, 2009.

Steven Henikoff and Jorja G Henikoff. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences*, 89(22):10915–10919, 1992.

A John Iafrate, Lars Feuk, Miguel N Rivera, Marc L Listewnik, Patricia K Donahoe, Ying Qi, Stephen W Scherer, and Charles Lee. Detection of large-scale variation in the human genome. *Nature genetics*, 36(9):949–951, 2004.

Miten Jain, Hugh E Olsen, Benedict Paten, and Mark Akeson. The oxford nanopore minion: delivery of nanopore sequencing to the genomics community. *Genome biology*, 17(1):239, 2016.

Miten Jain, Sergey Koren, Karen H Miga, Josh Quick, Arthur C Rand, Thomas A Sasani, John R Tyson, Andrew D Beggs, Alexander T Dilthey, Ian T Fiddes, et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nature Biotechnology*, 2018a.

Miten Jain, Sergey Koren, Karen H Miga, Josh Quick, Arthur C Rand, Thomas A Sasani, John R Tyson, Andrew D Beggs, Alexander T Dilthey, Ian T Fiddes, et al. Nanopore sequencing and assembly of a human genome with ultra-long reads, 2018b. URL `https://github.com/nanopore-wgs-consortium/NA12878/blob/master/Genome.md`.

Tanjina Kader, David L Goode, Stephen Q Wong, Jacquie Connaughton, Simone M Rowley, Lisa Devereux, David Byrne, Stephen B Fox, Gisela Mir Arnau, Richard W Tothill, et al. Copy number analysis by low coverage whole genome sequencing using ultra low-input DNA from formalin-fixed paraffin embedded tumor tissue. *Genome Medicine*, 8(1):121, 2016.

Samuel Karlin and Stephen F Altschul. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proceedings of the National Academy of Sciences*, 87(6):2264–2268, 1990.

Samuel Karlin, Amir Dembo, and Tsutomu Kawabata. Statistical composition of high-scoring segments from molecular sequences. *The Annals of Statistics*, 18(2):571–581, 1990.

John J Kasianowicz, Eric Brandin, Daniel Branton, and David W Deamer. Characterization of individual polynucleotide molecules using a membrane channel. *Proceedings of the National Academy of Sciences*, 93(24):13770–13773, 1996.

Jude Kendall and Alexander Krasnitz. *Computational Methods for DNA Copy-Number Analysis of Tumors*, pages 243–259. Springer New York, New York, NY, 2014.

W James Kent. Blat–the blast-like alignment tool. *Genome research*, 12(4):656–664, 2002.

Szymon M Kiełbasa, Raymond Wan, Kengo Sato, Paul Horton, and Martin C Frith. Adaptive seeds tame genomic sequence comparison. *Genome research*, 21(3):487–493, 2011.

Martin Kircher and Janet Kelso. High-throughput DNA sequencing–concepts and limitations. *Bioessays*, 32(6):524–536, 2010.

Samuel Kotz and Saralees Nadarajah. *Extreme value distributions: theory and applications*. World Scientific, 2000.

Stefan Kurtz, Adam Phillippy, Arthur L Delcher, Michael Smoot, Martin Shumway, Corina Antonescu, and Steven L Salzberg. Versatile and open software for comparing large genomes. *Genome biology*, 5(2):R12, 2004.

Ben Langmead, Cole Trapnell, Mihai Pop, and Steven L Salzberg. Ultrafast and memory-efficient alignment of short dna sequences to the human genome. *Genome biology*, 10(3):R25, 2009.

Andrew H Laszlo, Ian M Derrington, Brian C Ross, Henry Brinkerhoff, Andrew Adey, Ian C Nova, Jonathan M Craig, Kyle W Langford, Jenny Mae Samson, Riza Daza, et al. Decoding long nanopore sequencing reads of natural dna. *Nature biotechnology*, 32(8):829, 2014.

Richard M Leggett, Cristina Alcon-Giner, Darren Heavens, Shabhonam Caim, Thomas C Brook, Magdalena Kujawska, Samuel Martin, Ned Peel, Holly Acford-Palmer, Lesley Hoyles, et al. Rapid minion profiling of preterm microbiota and antimicrobial-resistant pathogens. *Nature Microbiology*, 5(3):430–442, 2020.

Heng Li. Aligning sequence reads, clone sequences and assembly contigs with bwa-mem. *arXiv preprint arXiv:1303.3997*, 2013.

Heng Li. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, 34(18):3094–3100, 2018.

Heng Li and Richard Durbin. Fast and accurate short read alignment with burrows–wheeler transform. *bioinformatics*, 25(14):1754–1760, 2009.

Heng Li and Richard Durbin. Fast and accurate long-read alignment with burrows–wheeler transform. *Bioinformatics*, 26(5):589–595, 2010.

Jian Li, Tielin Yang, Liang Wang, Han Yan, Yinping Zhang, Yan Guo, Feng Pan, Zhixin Zhang, Yumei Peng, Qi Zhou, et al. Whole genome distribution and ethnic differentiation of copy number variation in caucasian and asian populations. *PloS one*, 4(11), 2009.

Kate R Lieberman, Gerald M Cherf, Michael J Doody, Felix Olasagasti, Yvette Kolodji, and Mark Akeson. Processive replication of single dna molecules in a nanopore catalyzed by phi29 dna polymerase. *Journal of the American Chemical Society*, 132(50):17961–17972, 2010.

David J. Lipman, W. John Wilbur, Temple F. Smith, and Michael S. Waterman. On the statistical significance of nucleic add similarities. 1984.

Bo Liu, Dengfeng Guan, Mingxiang Teng, and Yadong Wang. rHAT: fast alignment of noisy long reads with regional hashing. *Bioinformatics*, 32(11):1625–1631, 2015.

Bo Liu, Yan Gao, and Yadong Wang. LAMSA: fast split read alignment with long approximate matches. *Bioinformatics*, 33(2):192–201, 2017.

Huanle Liu, Oguzhan Begik, Morghan C Lucas, Jose Miguel Ramirez, Christopher E Mason, David Wiener, Schraga Schwartz, John S Mattick, Martin A Smith, and Eva Maria Novoa. Accurate detection of $m^6A$ rna modifications in native rna sequences. *Nature communications*, 10(1):1–9, 2019.

HA Loaiciga and RB Leipnik. Analysis of extreme hydrologic events with gumbel distributions: marginal and additive cases. *Stochastic Environmental Research and Risk Assessment*, 13(4): 251–259, 1999.

Nicholas J Loman, Joshua Quick, and Jared T Simpson. A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nature methods*, 12(8):733–735, 2015.

Bin Ma, John Tromp, and Ming Li. Patternhunter: faster and more sensitive homology search. *Bioinformatics*, 18(3):440–445, 2002.

Geoff Macintyre, Teodora E Goranova, Dilrini De Silva, Darren Ennis, Anna M Piskorz, Matthew Eldridge, Daoud Sie, Liz-Anne Lewsley, Aishah Hanif, Cheryl Wilson, et al. Copy number signatures and mutational processes in ovarian carcinoma. *Nature Genetics*, 50(9):1262, 2018.

Alberto Magi, Davide Bolognini, Niccoló Bartalucci, Alessandra Mingrino, Roberto Semeraro, Luna Giovannini, Stefania Bonifacio, Daniela Parrini, Elisabetta Pelo, Francesco Mannelli, et al. Nano-gladiator: real-time detection of copy number alterations from nanopore sequencing data. *Bioinformatics*, 35(21):4213–4221, 2019.

Giovanni Maglia, Marcela Rincon Restrepo, Ellina Mikhailova, and Hagan Bayley. Enhanced translocation of single dna molecules through $\alpha$-hemolysin nanopores by manipulation of internal charge. *Proceedings of the National Academy of Sciences*, 105(50):19720–19725, 2008.

Paymaneh D Malihi, Michael Morikado, Lisa Welter, Sandy T Liu, Eric T Miller, Radu M Cadaneanu, Beatrice S Knudsen, Michael S Lewis, Anders Carlsson, Carmen Ruiz Velasco, et al. Clonal diversity revealed by morphoproteomic and copy number profiles of single prostate cancer cells at diagnosis. *Convergent Science Physical Oncology*, 4(1):015003, 2018.

Elizabeth A Manrao, Ian M Derrington, Mikhail Pavlenok, Michael Niederweis, and Jens H Gundlach. Nucleotide discrimination with dna immobilized in the mspa nanopore. *PloS one*, 6(10), 2011.

Elizabeth A Manrao, Ian M Derrington, Andrew H Laszlo, Kyle W Langford, Matthew K Hopper, Nathaniel Gillgren, Mikhail Pavlenok, Michael Niederweis, and Jens H Gundlach. Reading dna at single-nucleotide resolution with a mutant mspa nanopore and phi29 dna polymerase. *Nature biotechnology*, 30(4):349, 2012.

Filipe J Marques, Carlos A Coelho, and Miguel De Carvalho. On the distribution of linear combinations of independent gumbel random variables. *Statistics and Computing*, 25(3):683–701, 2015.

Edward M McCreight. A space-economical suffix tree construction algorithm. *Journal of the ACM (JACM)*, 23(2):262–272, 1976.

Amit Meller, Lucas Nivon, Eric Brandin, Jene Golovchenko, and Daniel Branton. Rapid nanopore discrimination between single polynucleotide molecules. *Proceedings of the National Academy of Sciences*, 97(3):1079–1084, 2000.

Eli L Moss, Dylan G Maghini, and Ami S Bhatt. Complete, closed bacterial genomes from microbiomes using nanopore sequencing. *Nature Biotechnology*, pages 1–7, 2020.

M Muthukumar. Theory of capture rate in polymer translocation. *The Journal of Chemical Physics*, 132(19):05B605, 2010.

Saralees Nadarajah. Exact distribution of the linear combination of p gumbel random variables. *International Journal of Computer Mathematics*, 85(9):1355–1362, 2008.

Nicholas Navin, Jude Kendall, Jennifer Troge, Peter Andrews, Linda Rodgers, Jeanne McIndoo, Kerry Cook, Asya Stepansky, Dan Levy, Diane Esposito, et al. Tumour evolution inferred by single-cell sequencing. *Nature*, 472(7341):90, 2011.

Saul B Needleman and Christian D Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology*, 48(3):443–453, 1970.

Adam B Olshen, ES Venkatraman, Robert Lucito, and Michael Wigler. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*, 5(4):557–572, 2004.

William R Pearson and David J Lipman. Improved tools for biological sequence comparison. *Proceedings of the National Academy of Sciences*, 85(8):2444–2448, 1988.

Robert F Purnell and Jacob J Schmidt. Discrimination of single base substitutions in a dna strand immobilized in a biological nanopore. *ACS nano*, 3(9):2533–2538, 2009.

Joshua Quick, Nicholas J Loman, Sophie Duraffour, Jared T Simpson, Ettore Severi, Lauren Cowley, Joseph Akoi Bore, Raymond Koundouno, Gytis Dudas, Amy Mikhail, et al. Real-time, portable genome sequencing for ebola surveillance. *Nature*, 530(7589):228–232, 2016.

Arthur C Rand, Miten Jain, Jordan M Eizenga, Audrey Musselman-Brown, Hugh E Olsen, Mark Akeson, and Benedict Paten. Mapping dna methylation with high-throughput nanopore sequencing. *Nature methods*, 14(4):411, 2017.

Richard Redon, Shumpei Ishikawa, Karen R Fitch, Lars Feuk, George H Perry, T Daniel Andrews, Heike Fiegler, Michael H Shapero, Andrew R Carson, Wenwei Chen, et al. Global variation in copy number in the human genome. *nature*, 444(7118):444–454, 2006.

Jonathan Sebat, B Lakshmi, Jennifer Troge, Joan Alexander, Janet Young, Pär Lundin, Susanne Månér, Hillary Massa, Megan Walker, Maoyen Chi, et al. Large-scale copy number polymorphism in the human genome. *Science*, 305(5683):525–528, 2004.

Jonathan Sebat, B Lakshmi, Dheeraj Malhotra, Jennifer Troge, Christa Lese-Martin, Tom Walsh, Boris Yamrom, Seungtai Yoon, Alex Krasnitz, Jude Kendall, et al. Strong association of de novo copy number mutations with autism. *Science*, 316(5823):445–449, 2007.

Venkatraman E Seshan, Adam B Olshen, et al. DNAcopy: a package for analyzing DNA copy data. 2010.

Jared T Simpson, Rachael E Workman, PC Zuzarte, Matei David, LJ Dursi, and Winston Timp. Detecting dna cytosine methylation using nanopore sequencing. *Nature methods*, 14(4):407, 2017.

Temple F Smith, Michael S Waterman, et al. Identification of common molecular subsequences. *Journal of molecular biology*, 147(1):195–197, 1981.

Temple F Smith, Michael S Waterman, and Christian Burks. The statistical distribution of nucleic acid similarities. *Nucleic Acids Research*, 13(2):645–656, 1985.

TF Smith, MS Waterman, and JR Sadler. Statistical characterization of nucleic acid sequence functional domains. *Nucleic acids research*, 11(7):2205–2220, 1983.

Ivan Sović, Mile Šikić, Andreas Wilm, Shannon Nicole Fenlon, Swaine Chen, and Niranjan Nagarajan. Fast and sensitive mapping of nanopore sequencing reads with GraphMap. *Nature Communications*, 7:11307, 2016.

Mircea Cretu Stancu, Markus J Van Roosmalen, Ivo Renkens, Marleen M Nieboer, Sjors Middelkamp, Joep De Ligt, Giulia Pregno, Daniela Giachino, Giorgia Mandrile, Jose Espejo Valle-Inclan, et al. Mapping and phasing of structural variation in patient genomes using nanopore sequencing. *Nature communications*, 8(1):1–13, 2017.

David Stoddart, Andrew J Heron, Ellina Mikhailova, Giovanni Maglia, and Hagan Bayley. Single-nucleotide discrimination in immobilized dna oligonucleotides with a biological nanopore. *Proceedings of the National Academy of Sciences*, 106(19):7702–7707, 2009.

John R Tyson, Nigel J O'Neil, Miten Jain, Hugh E Olsen, Philip Hieter, and Terrance P Snutch. Minion-based long-read sequencing and assembly extends the caenorhabditis elegans reference genome. *Genome Research*, 28(2):266–274, 2018.

E Van Binsbergen. Origins and breakpoint analyses of copy number variations: up close and personal. *Cytogenetic and genome research*, 135(3-4):271–276, 2011.

Meni Wanunu, Jason Sutin, Ben McNally, Andrew Chow, and Amit Meller. DNA translocation governed by interactions with solid-state nanopores. *Biophysical Journal*, 95(10):4716–4725, 2008.

David Weese, Anne-Katrin Emde, Tobias Rausch, Andreas Döring, and Knut Reinert. Razers–fast read mapping with sensitivity control. *Genome research*, 19(9):1646–1654, 2009.

Rachael E Workman, Alison D Tang, Paul S Tang, Miten Jain, John R Tyson, Roham Razaghi, Philip C Zuzarte, Timothy Gilpatrick, Alexander Payne, Joshua Quick, et al. Nanopore native rna sequencing of a human poly (a) transcriptome. *Nature methods*, 16(12):1297–1305, 2019.

Mehdi Zarrei, Jeffrey R MacDonald, Daniele Merico, and Stephen W Scherer. A copy number variation map of the human genome. *Nature reviews genetics*, 16(3):172–183, 2015.

# Appendix A

# Supplemental methods

## DNA samples

The normal diploid female DNA was purchased from Promega (Cat. no. G1521). Breast cancer cell line SK-BR-3 (American Type of Culture Collection (ATCC), Cat. no. HTB-30) was cultured in RPMI-1640 medium (Thermo Fisher Scientific, Cat. no. 11875093) supplemented with 10% fetal bovine serum (FBS) (Thermo Fisher Scientific, Cat. no. 35011CV) and was maintained at $37^{\circ}$ in a humidified chamber supplied with 5% $CO_2$ and was regularly tested for mycoplasma infection.

## Cell lysis and DNA purification

The DNA from SK-BR-3 cells was extracted and purified with the QIAamp DNA Blood Mini Kit (Qiagen, Cat. no. 51104) following the protocol for cultured cells given by the manufacturer. RNA and proteins in the cells were degraded using RNase A stock solution ($100$ mg/ml) (Qiagen, Cat. no. 19101) and Protease-K (Qiagen, Cat. no. 19133) respectively. Both purchased female diploid DNA and extracted SK-BR-3 DNA were treated with the same downstream processes.

## Fragmenting genomic DNA

2-3 µg of genomic DNA was fragmented with restriction enzyme Anza 64 SaqAI (Thermo Fisher Scientific, Cat. no. IVGN0644) for 30 min at 37°. The fragmented DNA was cleaned with the QIAquick PCR purification kit (Qiagen, Cat. no. 8106) and eluted with 34 µl nuclease-free water. The concentration of DNA was quantified on a Qubit Fluorometer v3 (Thermo Fisher Scientific, cat. no. Q33216) with the Qubit dsDNA HS assay kit (Thermo Fisher Scientific, cat. no. Q32854).

## Ligation of fragmented DNA

500 ng of fragmented DNA in 10 µl nuclease-free water was mixed with 10 µl Anza T4 DNA Ligase Master Mix (Thermo Fisher Scientific, Cat. no. IVGN210-4) and incubated for 30 min at room temperature. The ligated DNA was cleaned with 2× volume Ampure XP beads (Beckman Coulter, Cat. no. A63881) and eluted in nuclease-free water. This step was done in multiple tubes if more than 500 ng of fragmented DNA was needed to be ligated. The concentration of DNA was quantified on a Qubit Fluorometer v3 with the Qubit dsDNA HS assay kit to ensure ≥1 µg (≥ 400 ng, if the Rapid kit was used for library preparation) remained. The size of the ligated DNA molecules were assessed with 1% agarose gel electrophoresis run at 90 V for 30 min.

## Library preparation (SQK-LSK108 1D DNA by ligation)

1 µg of re-ligated DNA in 45 µl of nuclease-free water was end-repaired and dA-tailed (New England Biolabs (NEB), Cat. no. E7546), followed by elution in nuclease-free water after 1.5× volume Ampure XP beads clean-up. Sequencing adapters (AMX1D) were ligated with Blunt/TA Ligase Master Mix (NEB, Cat.no. M0367) and cleaned with 0.4× volume Ampure XP beads and eluted using 15 µl Elution Buffer (ELB) following the manufacturer's protocol (Oxford Nanopore Technologies (ONT), 1D genomic DNA by ligation protocol).

## Multiplexed library preparation (EXP-NBD103 and SQK-LSK108)

700 ng of each re-ligated sample in 45 µl of nuclease-free water was end-repaired, dA-tailed (NEB, Cat. no. E7546), cleaned with 1.5× volume Ampure XP beads and eluted in nuclease-free water. Different Native Barcodes (NB-x) for each sample was ligated with Blunt/TA Ligase Master Mix (NEB, Cat.no. M0367), cleaned with 2× volume Ampure XP beads and eluted in nuclease-free water. Equimolar amounts of each sample was pooled to have 700 ng of DNA in 50 µl water. Barcode adapters (BAM) were ligated with Quick T4 DNA Ligase (NEB, Cat. no. E6056), cleaned with 0.4× volume Ampure XP beads and eluted using 15 µl Elution Buffer (ELB) following the manufacturer's protocol (ONT, 1D native barcoding genomic DNA).

## Library preparation (SQK-RAD003 Rapid sequencing)

400 ng of re-ligated DNA was concentrated with 2× volume Ampure XP beads to 7.5 µl nuclease-free water. DNA was tagmented with Fragmentation Mix (FRA), and Rapid 1D Adapter (RPD) was attached following the manufacturer's protocol (ONT, rapid sequencing).

## MinION sequencing and base-calling

All the prepared libraries were loaded on R9.5 Flowcells following the manufacturer's protocol (ONT) and sequenced for up to 48 hours using the script specific to library preparation protocol. Base-calling and de-multiplexing barcoded reads were performed using ONT Guppy (2.3.5) with the appropriate parameters based on the library preparation kit.

## Sequencing RE digested normal diploid genome

$1\,\mu g$ of genomic DNA was fragmented with restriction enzyme Anza 64 SaqAI (Thermo Fisher Scientific, Cat. no. IVGN0644) for $30$ min at $37°$. The fragmented DNA was cleaned with the QIAquick PCR purification kit (Qiagen, Cat. no. 8106) and eluted with $31\,\mu l$ nuclease-free water. The concentration of DNA was quantified on a Qubit Fluorometer v3 (Thermo Fisher Scientific, cat. no. Q33216) with the Qubit dsDNA HS assay kit (Thermo Fisher Scientific, cat. no. Q32854).

$0.5\,\mu g$ of restriction enzyme digested DNA in $45\,\mu l$ of nuclease-free water was end-repaired and dA-tailed (New England Biolabs (NEB), Cat. no. E7546), followed by elution in nuclease-free water after $1.5\times$ volume Ampure XP beads clean-up. Sequencing adapters (AMX1D) were ligated with Blunt/TA Ligase Master Mix (NEB, Cat.no. M0367) and cleaned with $1.0\times$ volume Ampure XP beads (manufacturer's protocol uses $0.4\times$ volume XP beads, we increased to $1.0\times$ to get as many short molecules as possible) and eluted using $15\,\mu l$ Elution Buffer (ELB) following the manufacturer's protocol (Oxford Nanopore Technologies (ONT), 1D genomic DNA by ligation protocol).

The prepared library was loaded on R9.4 Flowcell following the manufacturer's protocol (ONT) and sequenced for 48 hours. Base-calling was performed using ONT Guppy (2.3.5).

## Estimation of copy number variations

CNV profiles were generated using the procedure described in (Baslan et al., 2012; Kendall and Krasnitz, 2014) with the modification employed in (Gerdtsson et al., 2018; Malihi et al., 2018). Briefly, the human reference genome (hg19) was split into 5,000 (20,000 or 50,000) bins containing an equal number of uniquely mappable locations and the bin counts were determined using uniquely mapped fragments. Bins with spuriously high counts ('bad bins', typically around centromeric and telomeric regions) were masked for downstream analysis (Kendall and Krasnitz, 2014). This procedure normalizes bin counts for biases correlated with GC content by fitting a

LOWESS curve to the GC content by bin count, and subtracting the LOWESS estimate from each bin (Kendall and Krasnitz, 2014). Circular binary segmentation (CBS) (Olshen et al., 2004), implemented in DNAcopy (Seshan et al., 2010) package, then identifies breakpoints in the normalized bin counts. Following (Gerdtsson et al., 2018; Malihi et al., 2018), after CBS, spurious segmentation calls were removed. The influence of the GC content correction can be seen in Additional file 1: Figure S14.

## Comparison with Illumina WGS of SK-BR-3 genome.

DNA from SK-BR-3 cells was used to construct WGS library with the NEBNext UltraII FS DNA Library Prep Kit (NEB, Cat. no. E7805) following the manufacturer's instructions. After library quality and quantity assessment with Qubit 3.0 HS dsDNA assay and BioAnalyzer HS dsDNA assay (Agilent), libraries were sequenced on HiSeq 2500 (Illumina) with single-end 130 cycles mode.

The reads were mapped with BWA-MEM using the default parameters, PCR duplicates were removed, and CNV profiles were generated using exactly the same method as used for SMURF-seq reads. The scatter plots and Pearson correlations comparing the CNV profiles were produced using R.

# Appendix B

# Data availability and summary of sequencing runs

| Sample | Kit | Reads | Mean length | Fragments | Accession |
|---|---|---|---|---|---|
| Diploid | SQK-LSK108 | 270.82k | 6.8 kb | 7.28M | SRX5893474 |
| Diploid | SQK-LSK108 | 497.92k | 3.7 kb | 7.55M | SRX5893475 |
| SK-BR-3 | SQK-LSK108 | 146.98k | 7.6 kb | 4.52M | SRX5893478 |
| SK-BR-3 | SQK-LSK108 | 132.64k | 7.3 kb | 4.02M | SRX5893479 |
| Diploid | SQK-RAD003 | 213.38k | 3.9 kb | 2.81M | SRX5893473 |
| Multiplexed run | EXP-NBD103 + | 442.9k | | | |
| Diploid (BC01) | SQK-LSK108 | 138.19k | 4.8 kb | 2.95M | SRX5893472 |
| SK-BR-3 (BC02) | | 144.57k | 7.7 kb | 4.97M | SRX5893476 |
| Diploid (short-read) | SQK-LSK108 | 2.58M | 630.9 bp | | SRX5893480 |
| SK-BR-3 (WGS) | Illumina WGS | 5.56M | 130 bp | | SRX5893477 |

Table B.1: Summary of sequencing run. Samples are processed with the SMURF-seq protocol, unless intdicated otherwise. Sequence data generated during the study are available in SRA with the accession number PRJNA454059.