Downloaded from genome.cshlp.org on July 16, 2026 . Published by Cold Spring Harbor Laboratory Press

Single-step capture and sequencing of natural DNA for detection of *BRCA1* mutations

John F Thompson, Jeffrey G Reifenberger, Eldar Giladi, et al.

Genome Res. published online July 15, 2011
Access the most recent version at doi:10.1101/gr.122192.111

P<P Published online July 15, 2011 in advance of the print journal.

Accepted Manuscript Peer-reviewed and accepted for publication but not copyedited or typeset; accepted manuscript is likely to differ from the final, published version.

License

Email Alerting Service Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or **click here.**

To subscribe to *Genome Research* go to:
https://genome.cshlp.org/subscriptions

Copyright © 2011, Cold Spring Harbor Laboratory Press

Downloaded from genome.cshlp.org on July 16, 2026 . Published by Cold Spring Harbor Laboratory Press

Single-Step Capture and Sequencing of Natural DNA for Detection of *BRCA1* Mutations

John F. Thompson*, Jeffrey G. Reifenberger*, Eldar Giladi, Kristen Kerouac, Jaime Gill, Erik Hansen, Avak Kahvejian, Philip Kapranov, Travis Knope, Doron Lipson, Kathleen E. Steinmann, and Patrice M. Milos

Helicos BioSciences Corporation
One Kendall Square, Building 200
Cambridge, MA 02139

*These authors contributed equally to the publication.

Corresponding author:
John F. Thompson
jthompson@helicosbio.com
Phone: 617-264-1888 FAX: 617-264-1700

Running title: *BRCA1* mutations via single-step capture/sequencing

Keywords: Next Generation Sequencing, cancer, targeted capture

Downloaded from genome.cshlp.org on July 16, 2026 . Published by Cold Spring Harbor Laboratory Press

Abstract

Genetic testing for disease risk is an increasingly important component of medical care. However, testing can be expensive which can lead to patients and physicians having limited access to the genetic information needed for medical decisions. To simplify DNA sample preparation and lower costs, we have developed a system in which any gene can be captured and sequenced directly from human genomic DNA without amplification, using no proteins or enzymes prior to sequencing. Extracted whole-genome DNA is acoustically sheared and loaded in a flow cell channel for single-molecule sequencing. Gene isolation, amplification, or ligation is not necessary. Accurate and low cost detection of DNA sequence variants is demonstrated for the *BRCA1* gene. Disease-causing mutations as well as common variants from well-characterized samples are identified. Single-molecule sequencing generates very reproducible coverage patterns and these can be used to detect any size insertion or deletion directly, unlike PCR-based methods which require additional assays. Because no gene isolation or amplification is required for sequencing, the exceptionally low costs of sample preparation and analysis could make genetic tests more accessible to those who wish to know their own disease susceptibility. Additionally, this approach has applications for sequencing integration sites for gene therapy vectors, transposons, retroviruses, and other mobile DNA elements in a more facile manner than possible with other methods.

Downloaded from genome.cshlp.org on July 16, 2026 . Published by Cold Spring Harbor Laboratory Press

Introduction

Genetic testing has become increasingly common for diagnosing or determining risks for a variety of conditions (Andermann and Blancquaert 2010). For example, mutations in *BRCA1* and *BRCA2* play a major role in familial breast and ovarian cancer (Miki et al. 1994; King et al. 2003) and account for a large fraction of the incidence of such cancers in women less than 50 years old. Affected and potentially affected women have a high level of interest in having these genes sequenced to help them make appropriate medical decisions (Lerman et al. 1995). New technologies such as next-generation sequencing have the potential to further accelerate knowledge in this area and provide prodigious amounts of data from a wide range of samples (Kahvejian et al. 2008). Though sequencing costs have been dramatically lowered, these technologies have not quickly transitioned to the molecular diagnostics arena partially because the complicated sample preparation and complex informatic analyses makes logistics and per sample costs relatively high (ten Bosch and Grody 2008; Milos 2009; Tucker et al. 2009). Targeted resequencing is often carried out for determination of genetic risk. However, the sample preparation burden and cost are escalated by the need for hybrid enrichment and/or PCR for gene targeting prior to sequencing. Gene targeting is currently carried out using multiple approaches for selective capture of specific DNA sequences, but all involve complex sample manipulations and require many steps including at least one and sometimes multiple rounds of amplification (Hedges et al. 2011).

One of the most common tests for determination of genetic risk, detection of mutations in *BRCA1* and *BRCA2*, uses PCR amplification prior to classical Sanger sequencing. Other

Downloaded from genome.cshlp.org on July 16, 2026 . Published by Cold Spring Harbor Laboratory Press

methods of sequencing such as oligonucleotide arrays or next-generation sequencing technologies, have also been used (Morgan et al. 2010; Schroeder et al. 2010; Walsh et al. 2010), but there is extensive sample preparation required, raising the costs and increasing the opportunity for sample handling errors in a clinical setting. Additionally, many tests have become mired in intellectual property issues due to gene patents (Cho 2010).

We have employed a single-molecule sequencing technology (Harris et al. 2008; Bowers et al. 2009) with a custom flow cell that allows natural DNA to be sequenced directly after being extracted and sheared with no amplification at any point. Instead of a flow cell with oligo-dT50 primers typically used in a Helicos Genetic Analysis System, a gene-specific flow cell was constructed with capture primers for defined genomic regions. These generate reads on both strands that allow sequence variant calling. The entire gene can thus be sequenced without copying the DNA or using any enzymes prior to sequencing.

Results

Single-molecule sequencing requires neither gene isolation nor amplification during sample preparation or during the sequencing reaction. The standard Helicos single-molecule sequencing protocol, described previously for genomic(Pushkarev et al. 2009), ChIP(Goren et al. 2010), and cDNA (Lipson et al. 2009), includes tailing DNA with terminal transferase and dATP followed by hybridizing to a flow cell with surface-attached dT50 oligonucleotides. This allows all nucleic acids in a sample to be sequenced. However, if only a subset of DNA regions is of interest, the standard dT50

Downloaded from genome.cshlp.org on July 16, 2026 . Published by Cold Spring Harbor Laboratory Press

oligonucleotides can be substituted with specific primers to capture and sequence the desired regions. This has the added benefit of simplifying sample preparation by eliminating the need for polyA tailing. Additionally, tailed samples require a blocking step in which the 3' end of the DNA is inactivated. This is unnecessary when using flow cells with a defined sequence because most of the randomly sheared fragments hybridizing to capture primers are longer than the primer and thus have single-stranded 3' overhangs (Figure 1 and Supplementary Figure 1). Because there is no template across from the 3' end of overhanging genomic DNA, the polymerase does not add extra bases to that end (Supplementary Figure 2). Incoming DNA with $\geq$20nt overhangs hybridizes to the surface as well as molecules with no overhangs. The upper limit for allowable overhangs has not been determined and is likely to be sequence dependent (data not shown). Secondary structures that involve the targeted region can also affect hybridization efficiency. Although some sources of DNA require shearing, DNA that is already fragmented requires absolutely no sample preparation and can truly be sequenced directly.

Flow cells were made that included 564 oligonucleotides (24-41nt) for *BRCA1* (Supplementary Table 1). The target sequences were chosen based on the exons and adjacent intronic sequences known to harbor clinically significant variants as listed in the Breast Cancer Information Core database (BIC, http://research.nhgri.nih.gov/projects/bic/), an international collaborative effort maintained by NHGRI. This targeted regions included all coding sequences including exons 2, 3, and 5-24 as well as 20 bp of intronic sequence at each acceptor splice site and

Downloaded from genome.cshlp.org on July 16, 2026 . Published by Cold Spring Harbor Laboratory Press

10 bp of intronic sequence at each donor splice site. 5' C12-amine modified oligonucleotides were covalently attached to an epoxide-containing surface. Additional oligonucleotides can be added to the same flow cells for sequencing of *BRCA2* and other genes.

To demonstrate that extracted DNA could be sequenced directly with no prior amplification, selection, or purification; DNAs with known mutations (shaded in Table 1) were obtained from the Coriell Institute for Medical Research biorepository, sheared to 200 bp, and used with no additional sample preparation. Typically, 1 µg of genomic DNA was hybridized overnight in each channel of a 25-channel flow cell but ranges of 0.1-20 µg genomic DNA and shorter hybridization times have also been used successfully. Higher DNA amounts yield more sequence reads from *BRCA1* though it is possible to use substantially less DNA, if necessary. DNAs were sequenced for 120 cycles of nucleotide addition. The sequence reads were filtered for artifacts caused by DNA molecules spaced too closely to be resolved by the camera and for length, keeping all reads of $\geq$25nt. Reads were then aligned (Giladi et al. 2010) to the *BRCA1* reference sequence. With the highest DNA amounts, >20% of the resulting sequences aligned to the *BRCA1* gene. Since *BRCA1* exons represent 0.0002% of the human genome, this is an enrichment of >100,000-fold. Of the reads that do not align to *BRCA1*, >90% align sporadically to the remainder of the genome. A typical alignment of sequence reads to the reference is shown in Figure 2. Variants detected in 22 samples are shown in Table 1.

Downloaded from genome.cshlp.org on July 16, 2026 . Published by Cold Spring Harbor Laboratory Press

To call substitutions and insertion/deletions of less than 5 bp, the Helicos open-source SNP caller was used with modifications specific for the direct capture method. In addition to the standard methods for calling variants, coverage of all regions was compared among samples. Because the hybridization and sequence yield is very consistent for a given set of hybridization conditions, variations in coverage are diagnostic for larger insertions and deletions. Thus, sequencing can be used to directly detect large rearrangements, in contrast to Sanger methods which do not distinguish large insertions/deletions from normal DNA. Since large indels are estimated to occur in 12% of *BRCA* high risk patients (Walsh et al. 2006), this approach has a significant advantage in detecting a wider range of clinically significant variants without additional assays.

The coverage profile for the *BRCA1* region for 19 samples run simultaneously on the same flow cell is shown in Supplementary Figure 3. The median coverage at each position for this sample set was 47x. Though there are coverage differences across this region caused by differing hybridization efficiencies of each oligonucleotide, the reproducibility at a given location is very high from sample to sample, making it possible to assess variations in coverage caused by large insertions and deletions. For example, there is a drop in coverage in exon 11 of NA13707 which contains a 40 bp deletion. The coverage for the 500 bp surrounding the deletion is provided for 19 DNAs with the lower coverage of the highlighted NA13707 DNA standing out (Figure 3). Furthermore, the exact boundaries of the deletion can be determined by examining reads that align to two separate segments of the reference (Figure 4). In this region, reads corresponding to both

Downloaded from genome.cshlp.org on July 16, 2026 . Published by Cold Spring Harbor Laboratory Press

the reference allele and the deleted allele are found. The novel sequence junction arising from the deletion can be easily observed.

While this method can yield complete coverage of *BRCA1*, problems can arise if coverage is not complete due to the presence of repetitive regions or regions that do not hybridize well. Unless primers are placed close together, there are regions of low coverage because reads from one primer may end before the next starts. Also, some repetitive regions are inaccessible due to the inability to place unique primers within those regions. To circumvent both problems, we have carried out a short, on-surface extension of the capture primer after hybridization but prior to sequencing. By including low concentrations of natural deoxynucleotides and polymerase, primers on the surface are extended by variable amounts into the captured DNA, allowing a randomization of the sequencing start sites and enabling reads within repetitive regions. Such random starting, which we refer to as JumpStart™ Sequencing, improves evenness of coverage (Figures 2 and S4) relative to standard Direct Capture Sequencing (with no extension). Flow cells using this approach can also be generated with fewer, more broadly spaced primers.

Discussion

The new method of capturing and sequencing natural DNA in a single step provides a much simpler approach to targeted sequencing and, as shown in Table 1, allows detection of all the previously identified variants in the targeted regions in the tested samples.

Downloaded from genome.cshlp.org on July 16, 2026 . Published by Cold Spring Harbor Laboratory Press

Though the current flow cell is directed at the entire coding sequence and 20nt of adjacent intronic sequence, it is straightforward to incorporate any other sequences that might be found to be medically relevant. Repetitive sequences that are more than twice as long as the read length and hence inaccessible to standard direct capture sequencing are observable using JumpStart Sequencing. Indeed, by generating very long extensions to primers prior to sequencing, repetitive sequences could be analyzed by capturing DNA with unique sequences just outside the repetitive regions and then extending into the repeat sequences prior to starting the sequence process. All reported clinically associated variants in *BRCA1* and *BRCA2* are within sequences that are amenable to this technique though there will be repetitive regions of the genome that will be difficult to cover.

A significant advantage of this approach relative to Sanger and amplification-based approaches is the ability to more readily detect large insertions and deletions because of the quantitative nature of the sequencing. Primers differ in their hybridization efficiency relative to each other due to sequence context but each primer is consistent across multiple samples, allowing coverage at any given position to be compared across many DNAs. The lack of amplification allows heterozygous changes in copy number caused by insertions or deletions to be readily detected. Deletion of whole exons or multiple exons will show extended blocks of reduced coverage. Smaller, intra-exonic deletions will show shorter regions of reduced coverage, but these will be further supported by sequence reads that span the new junction sequences caused by the deletion (Figure 4). Because insertions have the potential to introduce novel sequence, the fixed set of surface primers will only allow full sequencing of short inserts. However, the presence of a long

Downloaded from genome.cshlp.org on July 16, 2026 . Published by Cold Spring Harbor Laboratory Press

insertion will make it likely that a pathogenic variation has been introduced and this could be followed up for a more complete sequence determination using extended JumpStart sequencing conditions.

Though DNA extracted from cell lines and blood has been used thus far, other DNA sources should be suitable substrates as well. For example, many clinical samples are stored as formalin fixed paraffin embedded blocks and the DNA from such sources is frequently degraded during processing. In such cases, no sample preparation would be necessary for the extracted DNA. The DNA could be loaded directly on the flow cell and sequenced. Similarly, many buccal samples produce fragmented DNA and thus are sequenceable with no sample processing.

The novel approach described here holds broad potential not only for *BRCA1* but also for any other gene for which accurate, quantitative sequencing is medically important. The minimal sample handling and lack of amplification translate directly to lower sample preparation costs and reduced sample handling errors critical for molecular diagnostic laboratories. Additionally, informatic analysis is simplified due to the high enrichment for the region of interest. Beyond focusing on particular genes of interest, this selective capture approach can also be used for other purposes. For example, whole exome studies typically yield sequence on only ~80% of the originally targeted sequences (Hedges et al. 2011). Regions difficult to target or amplify using other selective capture methods are reproducible and could be addressed with this technology. Also, sequences present at low quantities in complex mixtures such as from a bacterial or viral infection or a tumor

Downloaded from genome.cshlp.org on July 16, 2026 . Published by Cold Spring Harbor Laboratory Press

sample could be applied to custom flow cells in order to detect DNA present at low quantities. Additionally, it is frequently useful to be able to determine integration sites for gene therapy vectors, transposons, retroviruses, and other mobile elements in genomic DNA (Gawronski et al. 2009; Hackett et al. 2010). This method of single-step capture and sequencing make such studies much more facile even if the sequence of only a single end of the mobile element is known (data not shown). Thus, direct sequencing technology offers multiple possibilities for expanding the use of next-generation sequencing as well as for making it feasible for much larger numbers of interested people to determine their genetic risk for a variety of diseases.

Materials and Methods

Flow Cell Design

Surface capture oligonucleotides were designed complementary to both strands of *BRCA1*. Attempts were made to keep GC content within 20-80%, make all T_ms > 65ºC, and minimize 3' and total self complementarity within a capture primer through the use of the program BatchPrimer3. When possible, oligonucleotides were spaced 20-30nt apart so adjacent sequence reads would overlap. While these criteria could be met for most of the sequence, there were regions where it was not possible so some criteria were relaxed to ensure that there was no gap of >35nt. Oligonucleotides were compared to all human genome sequences to ensure that the 3' terminal 20nt were not perfect matches to any other location in the genome. Where high levels of homology were present, efforts were made to place mismatches as close to the 3' end as possible while retaining reasonable spacing. For example, the distal end of exon 2 and the targeted intronic

Downloaded from genome.cshlp.org on July 16, 2026 . Published by Cold Spring Harbor Laboratory Press

sequence have 182 of 195 nt that are identical with a pseudogene so sequences had to be carefully selected to take advantage of the *BRCA1*-specific differences to ensure that sequence was read from exon 2 and not from the pseudogene. In addition, the gene-specific oligonucleotides were spiked with a 29nt oligonucleotide (5'ACTTATCCTTGCATCCATCCTCTGCCCTG 3'). This is included so that quality control oligonucleotides could be included without interfering with the *BRCA1* hybridizations and to provide a sufficient number of actively sequencing DNA molecules to maintain image registration during sequencing cycles. HPLC-purified, 5'amine-C12 oligonucleotides were ordered from either IDT or Operon and deposited on epoxide-modified glass at 37ºC at individual concentrations of 10 pM for 1 hr. Glass was assembled into 25 channel flow cells for use on the HeliScope Genetic Analysis System. Flow cells with oligonucleotides for other specific DNA sequences of interest can be ordered on a custom basis from Helicos.

DNA preparation and Hybridization

HapMap genomic DNA was ordered from the Coriell Institute for Medical Research (Camden, NJ) and K562 DNA from Biochain (Hayward, CA). The DNA was sheared to various sizes, generally to an average of 200 bp, using the Covaris Acoustic Shearing method as described by Covaris (Woburn, MA). Up to 25 µg of DNA was sheared in a single 120 µL shearing reaction. After shearing, 0.1-20 µg of genomic DNA was denatured at 95ºC for 5 min and snap cooled on ice. This was then added to 10 µl of 2x hybridization buffer. A variety of standard buffers (Hutton 1977; Tsai et al. 2005) was tried and most provided similar results. Most data presented were hybridized with 5x SSC and varying concentrations of formamide with either 0.02% TritonX-100 or 0.1%

Downloaded from genome.cshlp.org on July 16, 2026 . Published by Cold Spring Harbor Laboratory Press

SDS present. Hybridization temperatures between 37 and 60ºC were assessed with optimal results varying depending on the formamide concentration (1 - 25%). Hybridizations were carried out for 1-24 hrs in a humidified incubation oven set to the desired temperature. Flow cells were then placed in the HeliScope Genetic Analysis System and sequenced as described previously (Bowers et al. 2009). Most runs were 550 fields of view rather than the standard 1100 fields of view to allow faster data collection. Further reductions in fields of view are possible because significantly more sequence was being generated than was necessary for variant calling.

Read Filtering and Alignment

Prior to alignment, reads were filtered to remove sequences that would not be useful. Reads shorter than a specified minimum length, generally 25nt, were removed. Also, artifactual sequences arise when DNA molecules are too close on the surface or if nucleotides tend to stick to a particular spot on the flow cell. Both of these situations lead to an apparent over-incorporation of nucleotides so reads that closely resemble the base addition order are removed. Filtered reads were aligned to a set of sequences corresponding to the expected target sequences of the capture probes in a left-justified setting (i.e. alignments were penalized for starting away from the first base of the reference) for standard sequencing and to the whole gene reference for JumpStart Sequencing as well as the sequences corresponding to the spiked oligonucleotides. In parallel, reads were also aligned to the entire human genome and any read which had a higher-scoring alignment to the untargeted genome than the target sequence was discarded. Aligned reads were then optimally positioned with respect to the target *BRCA1* gene sequence.

Downloaded from genome.cshlp.org on July 16, 2026 . Published by Cold Spring Harbor Laboratory Press

Sequence Variant Detection

For detection of sequence variants in the aligned read data, we used SnpSniffer, a Helicos-developed algorithm. SnpSniffer calls SNPs and short indels using a multi-stage process. At first, the nucleotide counts at each position are tallied, based on aligned reads. The nucleotide counts are compared to those expected by chance due to machine error using a binomial model, and any positions that differ from the reference in a statistically significant fashion are flagged. The next stage identifies the most likely alleles at each candidate location and realigns the reads to the candidate alleles. This step reduces alignment errors substantially in locations with true sequence variants, and avoids spurious allele calls. Following the re-alignment stage mutations are called again using the realigned reads.

Homopolymer length mutations were determined by comparing the distribution of homopolymer lengths indicated by reads that span a homopolymer to the expected distributions. The expected distribution is computed using the machine deletion and insertion rates. The models to which the empirical data is compared correspond to both homozygous and heterozygous length mutations. Detected sequence variants were then filtered by minimal frequency, statistical significance, and average read length of sequence reads. Variants that did not pass these filters were considered artifacts and were ignored.

Insertions and deletions of 5nt or longer were detected by comparing the relative coverage of each individual sample to the median of all samples on the same run.

Downloaded from genome.cshlp.org on July 16, 2026 . Published by Cold Spring Harbor Laboratory Press

Regions that had significantly different coverage profiles were flagged and then all possible deletions examined for consistency with the observed sequences. Exact sequences for deletions can be determined in this manner. Insertions can be detected in this way as well but, if the insertion was much longer than the read length, determination of their exact sequence would require a longer JumpStart extension in order to span the sequence. Knowledge that a long insertion had occurred at a particular site would generally be sufficient to implicate that variant as clinically significant.

Acknowledgements

This publication was made possible by grant RC2 HG005598 from the National Human Genetics Research Institute (NHGRI) at the National Institutes of Health. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of NHGRI.

Downloaded from genome.cshlp.org on July 16, 2026 . Published by Cold Spring Harbor Laboratory Press

References

Andermann A, Blancquaert I. 2010. Genetic screening: A primer for primary care. *Can Fam Physician* **56**(4): 333-339.

Bowers J, Mitchell J, Beer E, Buzby PR, Causey M, Efcavitch JW, Jarosz M, Krzymanska-Olejnik E, Kung L, Lipson D et al. 2009. Virtual terminator nucleotides for next-generation DNA sequencing. *Nat Methods* **6**(8): 593-595.

Cho M. 2010. Patently unpatentable: implications of the Myriad court decision on genetic diagnostics. *Trends Biotechnol* **28**(11): 548-551.

Gawronski JD, Wong SM, Giannoukos G, Ward DV, Akerley BJ. 2009. Tracking insertion mutants within libraries by deep sequencing and a genome-wide screen for Haemophilus genes required in the lung. *Proc Natl Acad Sci U S A* **106**(38): 16422-16427.

Giladi E, Healy J, Myers G, Hart C, Kapranov P, Lipson D, Roels S, Thayer E, Letovsky S. 2010. Error tolerant indexing and alignment of short reads with covering template families. *J Comput Biol* **17**(10): 1279-1293.

Goren A, Ozsolak F, Shoresh N, Ku M, Adli M, Hart C, Gymrek M, Zuk O, Regev A, Milos PM et al. 2010. Chromatin profiling by directly sequencing small quantities of immunoprecipitated DNA. *Nat Methods* **7**(1): 47-49.

Hackett PB, Largaespada DA, Cooper LJ. 2010. A transposon and transposase system for human application. *Mol Ther* **18**(4): 674-683.

Harris TD, Buzby PR, Babcock H, Beer E, Bowers J, Braslavsky I, Causey M, Colonell J, Dimeo J, Efcavitch JW et al. 2008. Single-molecule DNA sequencing of a viral genome. *Science* **320**(5872): 106-109.

Hedges DJ, Guettouche T, Yang S, Bademci G, Diaz A, Andersen A, Hulme WF, Linker S, Mehta A, Edwards YJ et al. 2011. Comparison of Three Targeted Enrichment Strategies on the SOLiD Sequencing Platform. *PLoS One* **6**(4): e18595.

Hutton JR. 1977. Renaturation kinetics and thermal stability of DNA in aqueous solutions of formamide and urea. *Nucleic Acids Res* **4**(10): 3537-3555.

Kahvejian A, Quackenbush J, Thompson JF. 2008. What would you do if you could sequence everything? *Nat Biotechnol* **26**(10): 1125-1133.

King MC, Marks JH, Mandell JB. 2003. Breast and ovarian cancer risks due to inherited mutations in BRCA1 and BRCA2. *Science* **302**(5645): 643-646.

Lerman C, Seay J, Balshem A, Audrain J. 1995. Interest in genetic testing among first-degree relatives of breast cancer patients. *Am J Med Genet* **57**(3): 385-392.

Lipson D, Raz T, Kieu A, Jones DR, Giladi E, Thayer E, Thompson JF, Letovsky S, Milos P, Causey M. 2009. Quantification of the yeast transcriptome by single-molecule sequencing. *Nat Biotechnol* **27**(7): 652-658.

Miki Y, Swensen J, Shattuck-Eidens D, Futreal PA, Harshman K, Tavtigian S, Liu Q, Cochran C, Bennett LM, Ding W et al. 1994. A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. *Science* **266**(5182): 66-71.

Milos PM. 2009. Emergence of single-molecule sequencing and potential for molecular diagnostic applications. *Expert Rev Mol Diagn* **9**(7): 659-666.

Morgan JE, Carr IM, Sheridan E, Chu CE, Hayward B, Camm N, Lindsay HA, Mattocks CJ, Markham AF, Bonthron DT et al. 2010. Genetic diagnosis of familial breast cancer using clonal sequencing. *Hum Mutat* **31**(4): 484-491.

Downloaded from genome.cshlp.org on July 16, 2026 . Published by Cold Spring Harbor Laboratory Press

Pushkarev D, Neff NF, Quake SR. 2009. Single-molecule sequencing of an individual human genome. *Nat Biotechnol* **27**(9): 847-852.

Schroeder C, Stutzmann F, Weber BH, Riess O, Bonin M. 2010. High-throughput resequencing in the diagnosis of BRCA1/2 mutations using oligonucleotide resequencing microarrays. *Breast Cancer Res Treat* **122**(1): 287-297.

ten Bosch JR, Grody WW. 2008. Keeping up with the next generation: massively parallel sequencing in clinical diagnostics. *J Mol Diagn* **10**(6): 484-492.

Tsai MH, Yan H, Chen X, Chandramouli GV, Zhao S, Coffin D, Coleman CN, Mitchell JB, Chuang EY. 2005. Evaluation of hybridization conditions for spotted oligonucleotide-based DNA microarrays. *Mol Biotechnol* **29**(3): 221-224.

Tucker T, Marra M, Friedman JM. 2009. Massively parallel sequencing: the next big thing in genetic medicine. *Am J Hum Genet* **85**(2): 142-154.

Walsh T, Casadei S, Coats KH, Swisher E, Stray SM, Higgins J, Roach KC, Mandell J, Lee MK, Ciernikova S et al. 2006. Spectrum of mutations in BRCA1, BRCA2, CHEK2, and TP53 in families at high risk of breast cancer. *JAMA* **295**(12): 1379-1388.

Walsh T, Lee MK, Casadei S, Thornton AM, Stray SM, Pennil C, Nord AS, Mandell JB, Swisher EM, King MC. 2010. Detection of inherited mutations for breast and ovarian cancer using genomic capture and massively parallel sequencing. *Proc Natl Acad Sci U S A* **107**(28): 12629-12633.

Downloaded from genome.cshlp.org on July 16, 2026 . Published by Cold Spring Harbor Laboratory Press

Legends

Figure 1 Single –Molecule Sequencing Protocols

The standard single-molecule sequencing protocol using tailed DNA and oligo dT flow cells is shown on the left and the novel protocol using direct capture on the right. The diagram on the left shows three steps in the standard protocol while the diagram on the right shows the hybridization step only for the direct capture protocol with three different possible types of hybridizing molecules, a DNA that is only partially complementary to the capture primer (underhang), a DNA that is perfectly complementary to the capture primer, and a DNA that has a 3' overhang. All of these molecules are sequenceable.

Figure 2 Sequence reads from representative region using Direct Capture and JumpStart Sequencing

Representative sequence reads from one region of *BRCA1* exon 16 in the forward (black) and reverse (blue) directions are shown for the NA14096 sample. In the top half of the figure with Direct Capture Sequencing, horizontal black and blue arrows show where sequencing begins after each capture primer. In the lower half of the figure with JumpStart Sequencing, start sites were randomized by extension with polymerase prior to sequencing. A histogram for start sites is shown in Supplemental Figure S4. Sequence reads are aligned to the reference in the center (bold) and differences noted in red for substitutions, with dashes for deletions, and with bolded, underlined bases for insertions. Gaps are included between sequences so that individual reads can be distinguished. Random errors occur throughout the reads but these are easily distinguished from real variants in which differences from the reference occur much more often (downward red

Downloaded from genome.cshlp.org on July 16, 2026 . Published by Cold Spring Harbor Laboratory Press

arrow). This position corresponds to dbSNP variant rs1799966, a common substitution SNP. Reads shown are a sampling of those obtained for this region.

Figure 3: Sequence coverage

Coverage for 19 of the DNAs listed in Table 2 that were hybridized and sequenced on the same flow cell is shown for 500 bp of exon 11. Coverage bins of 5 bp were normalized to achieve the same amount of total coverage for each sample by multiplying by the appropriate factor (range = 0.70 to 2.7). DNA sample NA13707, which contains a heterozygous 40 bp deletion in this region, is shown with a heavy black line while all others are shown with thin gray lines. Numbering for the region is as found in the BIC database. The lowered coverage over a larger region than 40 bp and not centered on the deletion is caused by the presence of primers located entirely within or overlapping the deletion that generate no reads from those primers for the deletion allele.

Figure 4: Deletion sequences

The exact coordinates of the 40 bp deletion in NA13707 can be determined by examining reads that map to the reference sequence in the region of interest in an interrupted manner. The reference sequence is shown in the center with the known 40 bp deletion region underlined. From 928 reads aligning to this region, 28 selected reads with different end points are shown. Forward reads (left to right) are in black and reverse reads (right to left) are in blue. Deletions relative to the reference are shown by dashes and insertions by bolding and underling the adjacent base.

Downloaded from genome.cshlp.org on July 16, 2026 . Published by Cold Spring Harbor Laboratory Press

Table 1

Homozygous and heterozygous variations detected in 22 Coriell samples (13705, 13707, 13710, 13711, 13712, 13713, 13714, 13715, 14090, 14091, 14092, 14095, 14096, 14097, 14622, 14623, 14637, 14638, 14639, 14684, 14788 and 14805) are listed. No variations were detected in *BRCA1* for NA14788. The variant location as defined by the Breast Cancer Information Core database (http://research.nhgri.nih.gov/projects/bic/) and dbSNP identifiers are shown in columns 1 and 2. In column 1, numbers refer to cDNA positions unless preceded by a letter which indicates the amino acid and the position. Each of the clinically significant mutations that had been previously characterized is shaded. Five of the tested samples have no clinically significant mutations in *BRCA1*. Samples that were sequenced using both Direct Capture and JumpStart Sequencing are in white font. Data from these samples has been deposited in the Sequence Read Archive with accession number SRP007097.

Table 1: Sequence variations identified in *BRCA1*

BIC	dbSNP	exon	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
185delAG	rs80357713	2									AG/-													
C61G	rs28897672	5														G/T								
IVS5-11T>G	rs80358061	5-6 IVS																		T/G				
917delTT	rs80357724	11																				TT/-		
1186	rs1799950	11				A/G								A/G		A/G						A/G		
1294del40	rs80359874	11		40/-																				
2201	rs1799949	11	C/T	C/T	C/T		C/T	C/T	C/T	C/T	T/T	C/T	C/T		T/T		C/T	C/T		C/T	T/T			T/T
2276insA	rs80357715	11					InsA																	
2430	rs16940	11	C/T	C/T	C/T		C/T	C/T	C/T	C/T	C/C	C/T	C/T		C/C		C/T	C/T		C/T	C/C			C/C
2434	rs80357467	11		C/T																				
2731	rs799917	11	C/T	C/T	C/T		C/T	C/T	C/T	C/T	C/T	C/T	C/T		T/T		C/T	C/T		C/T	C/T			T/T
3232	rs16941	11	A/G	A/G	A/G		A/G	A/G	A/G	A/G	G/G	A/G	A/G		G/G		A/G	A/G		A/G	G/G			G/G
S1040N	rs4986852	11				A/G								A/G										
3600del11	rs80357877	11													11/-									
3667	rs16942	11	A/G	A/G	A/G		A/G	A/G	A/G	A/G	G/G	A/G	A/G		G/G		A/G	A/G		A/G	G/G			G/G
E1250X	rs28897686	11						G/T																
3875del4	rs80357868	11	GTCT/-																					
R1347G	rs28897689	11																				A/G		
4427	rs1060915	13	C/T	C/T	C/T		C/T	C/T	C/T	C/T	C/C	C/T	C/T		C/C		C/T	C/T		C/T	C/C			C/C
R1443X	rs41293455	13			C/G														C/T					
Y1563X	rs80357433	16																						
4956	rs1799966	16	A/G	A/G	A/G		A/G	A/G	A/G	A/G	G/G	A/G	A/G		G/G		A/G	A/G		A/G	G/G			G/G
5075	rs1799967	16											A/G								A/G			
5256delG	rs80357997	18												G/-										
V1713A	rs80357132	18											C/T											
5382insC	rs80357906	20								insC		insC												
5438insC	rs80357823	21							insC															

Downloaded from genome.cshlp.org on July 16, 2026 - Published by Cold Spring Harbor Laboratory Press

Downloaded from genome.cshlp.org on July 16, 2026 . Published by Cold Spring Harbor Laboratory Press

Supplementary Material

Supplementary Figure S1

Because acoustical shearing of DNA generates random ends that will have variable degrees of complementarity to the primers on the flow cell surface, a series of oligonucleotides was synthesized to test hybridization properties. A 29-mer was attached to an epoxide-containing flow cell surface via a 5' C12 amino link. The 3' end of this molecule can serve as a primer to read the sequence of any molecule that stably hybridizes to it. Thirty oligonucleotides were synthesized with unique 5' ends that could be distinguished by sequencing and connected to variable 3' ends that were complementary to 9-29nt of the surface-bound primer. Oligonucleotides were made complementary to 9, 14, and 19-29nt (corresponding to gaps of 0 to 20nt) of the primer. Additionally, oligonucleotides were made with perfect 29nt matches to the primer and 3'-overhangs of 1-10, 15 and 20nt. Duplicate oligonucleotides were made for -20, -15, -10, -5, 0, +5, +10, +15, and +20 relative to the primer (sequences provided in Table S2). All oligonucleotides except those containing only 9 and 14nt of complementarity provided significant amounts of sequence data with standard hybridization conditions and 1% formamide (dashed lines, open circles). At 25% formamide (solid lines, filled circles), the yield of oligos with only 19-22nt complementarity drops relative to lower formamide. In the presence of 1 μg of sheared human genomic DNA, this effect is magnified as the oligos show more specificity in the background of non-specific DNA (data not shown). Though there is an obvious length dependence of the hybridization, this effect is not monotonic as longer stretches of complementarity sometimes generate lower yield. In some cases, this appears to be due to internal secondary structure in the incoming

Downloaded from genome.cshlp.org on July 16, 2026 . Published by Cold Spring Harbor Laboratory Press

oligonucleotides as some have structures that might be stable in these conditions and inhibit binding to the surface. For example, the oligonucleotide with 23nt complementarity has a stable intramolecular structure that includes the 3' end which likely reduces its hybridization efficiency. Interestingly, these differences tend to disappear at 25% formamide, suggesting that the intramolecular structures melt and the entire sequence is available for intermolecular interactions. In addition, oligonucleotides with 3' overhangs of up to 20nt (the longest that was tested) that go beyond the 5' end of the primer hybridize as well or better than oligonucleotides that have no overhang. Interestingly, the overhang sequence affects the yield of reads as all the oligonucleotides whose overhang starts with the sequence AACCT yield more reads than the two that start with TTAGG. This may be due to interactions with the surface or the hybrid stability could be affected by the ease with which the single-stranded DNA can turn to avoid the surface or stack with the double-stranded DNA. This behavior is mirrored in the *BRCA1* sequences where it is found that the sequence just beyond the capture primer has an effect on hybridization efficiency. On average, the highest yielding capture primers have the sequence TT just beyond the 5' end or, AA on the incoming DNA, mirroring the effect of the synthetic oligonucleotides.

Supplementary Figure S2

When standard sequencing using tailed DNA on an oligo dT50 flow cell is carried out on the HeliScope Sequencing system, the DNA to be sequenced is blocked at the 3' end so that the only end available for sequence extension is the 3' end of surface-attached primer. Oligonucleotides were used to study this in more detail. One set of oligonucleotides was

Downloaded from genome.cshlp.org on July 16, 2026 . Published by Cold Spring Harbor Laboratory Press

blocked at the 3' end with a phosphate so they could not extend. A fraction of this mix was phosphatased to provide free 3' ends that could theoretically be extended during the sequencing reaction and yield sequence signal from both DNA strands, causing apparent insertions. Increased errors are observed with underhang oligonucleotides that are not fully complementary to the capture primer. The insertion (solid black squares) and substitution (gray triangles) rate is slightly higher for the molecules with a 3' OH relative to a 3' PO_4. The deletion rate (open diamonds) is unaffected. In contrast, oligonucleotides with a 3' overhang have exactly the same error rate whether they contain a 3' OH or PO_4. Thus, when no template is present, there is no need to block the 3' end. Since extended molecules hybridize efficiently and are likely to make up the majority of hybridizing molecules when using randomly sheared DNA, blocking the 3' end of samples is not necessary.

Supplementary Figure S3

Average coverage for 19 samples across the target region, demonstrating high reproducibility of probe capture efficiency among different samples. Coverage was calculated in bin sizes of 50nt. Exon numbering is shown. Intronic size is not to scale.

Supplementary Figure S4

The start site for observed sequence reads is shown for the standard direct capture (solid line) and with primer extension (JumpStart Sequencing, dashed line) prior to sequencing. Most reads start within one base of the expected start site for standard sequencing but

Downloaded from genome.cshlp.org on July 16, 2026 . Published by Cold Spring Harbor Laboratory Press

vary by more than 20nt when extended prior to sequencing. The length of this variable start can be changed by altering polymerization time or nucleotide concentration.

Supplementary Table 1

Sequences for the oligonucleotides used in the construction of the flow cell are listed as well as the exons for which they were designed.

Supplementary Table 2

Sequences for the oligonucleotides used for testing hybridization efficiency are shown with the length of under/overhang in each.

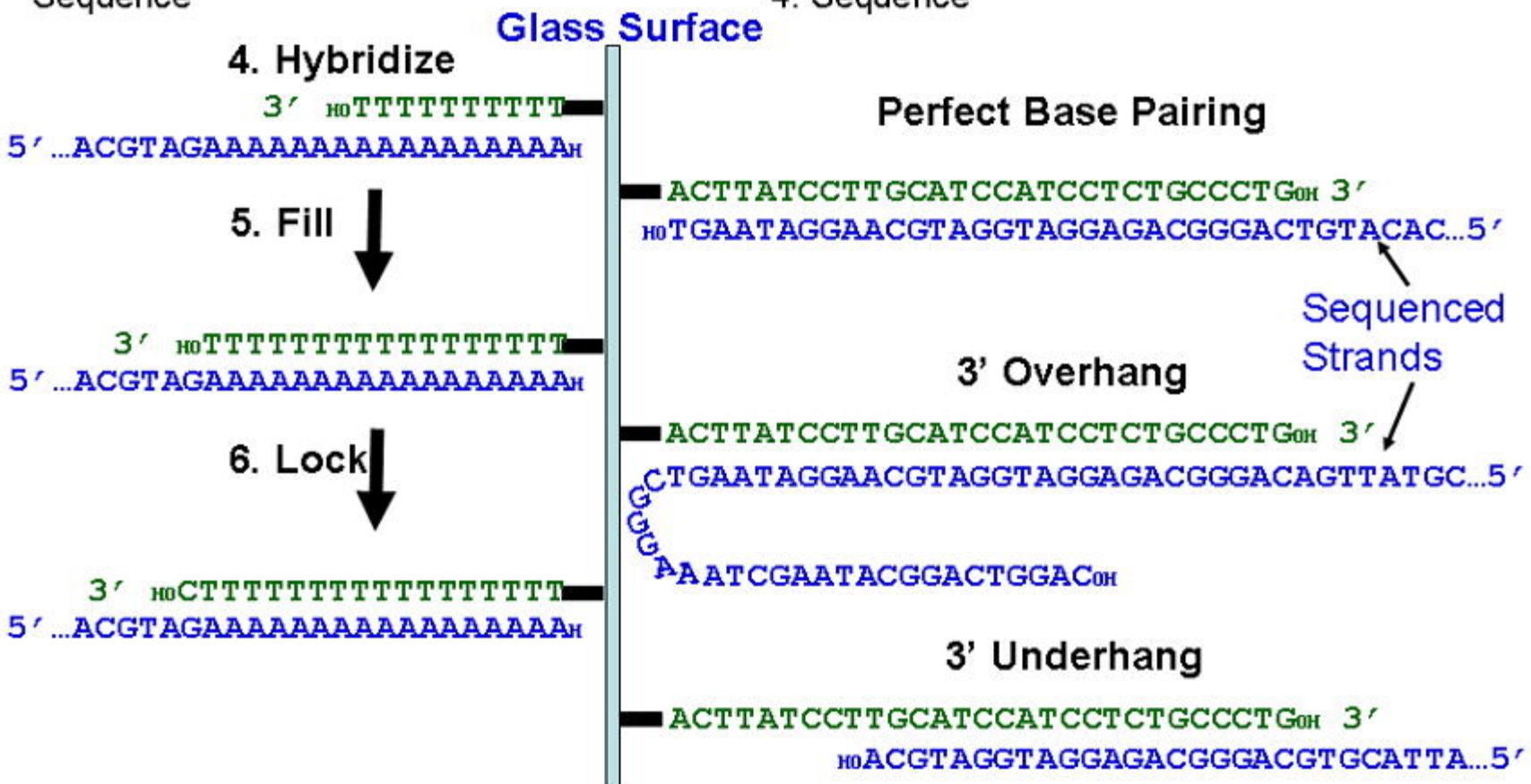
Single-Molecule Sequencing Protocols

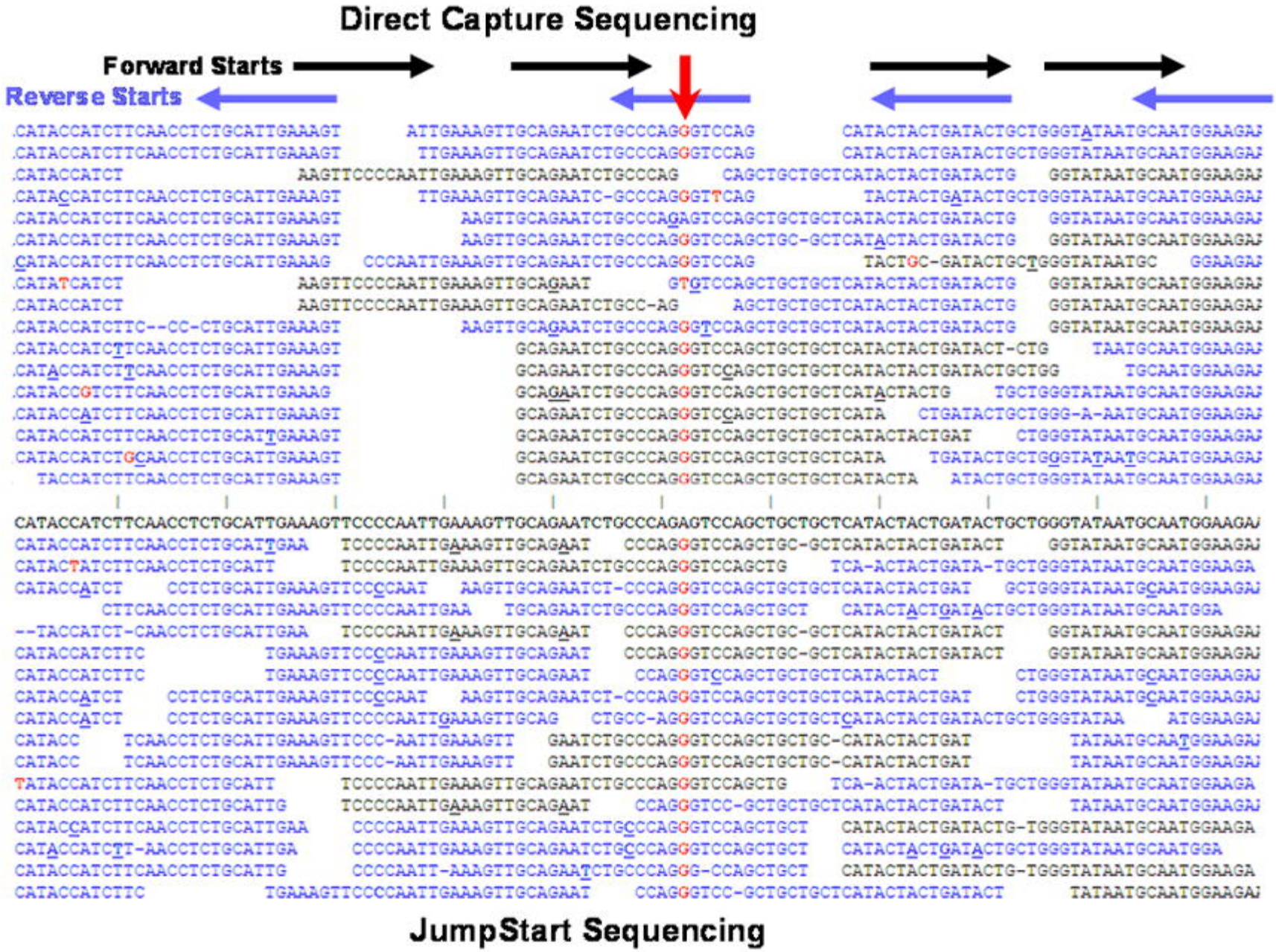
Standard Single-molecule Sequencing

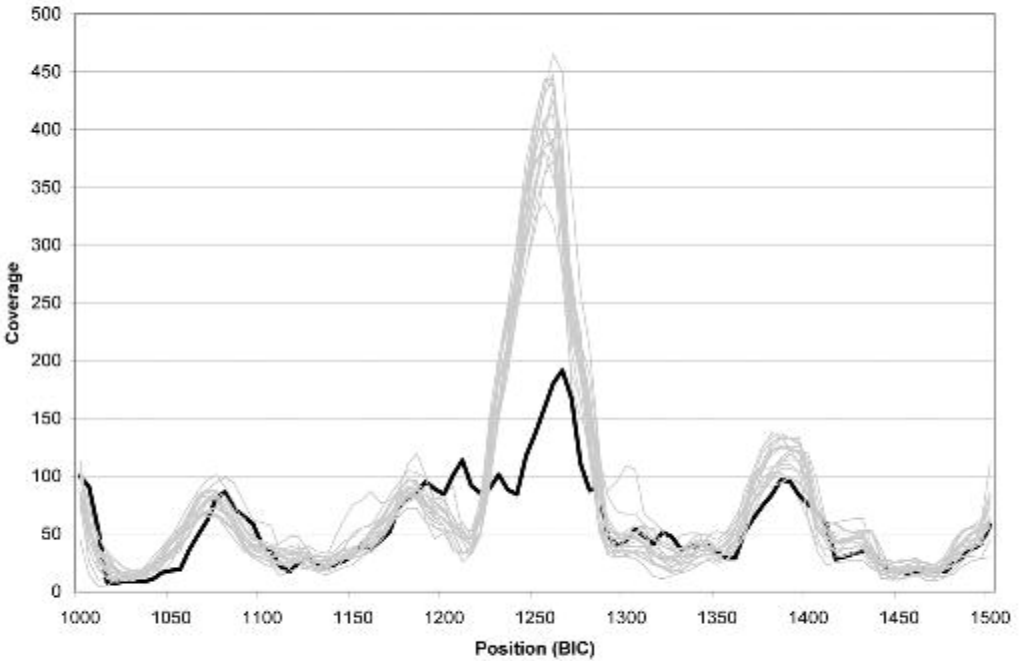
1. Shear DNA
2. Poly dA tail DNA
3. Block 3' end with dideoxynucleotide
4. Hybridize to oligo dT flow cell
5. Fill excess dAs in tail with TTP
6. Lock in place with fluorescent nucleotide
7. Sequence

Direct Capture Sequencing

1. Shear DNA
2. Hybridize to custom sequence flow cell
3. Visualize with fluorescent nucleotide
4. Sequence




Direct Capture Sequencing
Forward Starts
Reverse Starts
JumpStart Sequencing


500
450
400
350
300
250
200
150
100
50
0
Coverage
1000
1050
1100
1150
1200
1250
1300
1350
1400
1450
1500
Position (BIC)


GTTTTCCAGAAGTGATGAAC----------AAATGCC
TGAGTGGTTTTCCAGAAGTGATGAAC----------AAATGCCAAAGT
CCAGAAGTGATGAAC----------AAATGCCAAAGT
TTTCCAGAAGTGATGAAC----------AAATGCCAAAGT
GTTTTCCAGAAGTGATGAAC----------AAATGCCAAAGT
TTTTCCAGAAGTGATGAAC----------AAATGCCAAAGT
AAGTTAATGAGTGGTTTTCCAGAAGTGATGAAC----------AAATG
TCCAGAAGTGATGAAC----------AAATGCCAAA
TCCAGAAGTGATGAAC----------AAATGCCAAAGT
TTCCAGAAGTGATGAAC----------AAATGCCAAAGT
TAATGAGTGGTTTTCCAGAAGTGATGAAC----------AAATGC
AAGTTAATGAGTGGTTTTCCAGAAGTGATGAACTGTTAGGTTCTGATGACTCACATGATGGGGAGTCTGAATCAAATGCCAAAGTAGCTGATGTATTGGACGTTCTAA
TTAGGTTCTGATGACTCACATGATGGGGAGTCTGAAT
CAAATGCCAAAGTAGCTGATGTATTGGACGTTCTAA
TGATGACTCACATGATGGGGAGTCTGAAT
CAGAAGTGATGAACTGTTAGGTTCTGATG
TTCTGATGACTCACATGATGGGGAGTCTGA
GATGACTCACATGATGGGGAGTCTGAA
GATGAACTGTTAGGTTCTGATGACTCACATGATG
ATGAACTGTTAGGTTCTGATGACTCACATGATGGGGAGTCTGA
GGGAGTCTGAATCAAATGCCAAAGT
TGAA-CAAATGCCAAAGTAGCTGATGTAGTTGGACGTT
CAGAAGTGATGAACTGTTAGGTTCTGATG
GGAGTCTGAACTCAAATGCCAAAGT
CTGTTAGGTTCTGATGACTCACATGATGGGG
AAGTGATGAACTGTTAGGTTCTGATGAC
ATGAACTGTTAGGTTCTGATGACTCACATGATGG
ACTGTTAGGTTCTGATGACTCACATGATGGGGAGTCTGA
AGTGATGAACTGTTAGGTTCTGATGACTCAC