HHS Public Access

Author manuscript

Nat Methods. Author manuscript; available in PMC 2015 November 01.

Published in final edited form as:

Nat Methods. 2015 May ; 12(5): 423–425. doi:10.1038/nmeth.3351.

Sequencing small genomic targets with high efficiency and extreme accuracy

Michael W. Schmitt[1,3], **Edward J. Fox**[2], **Marc J. Prindle**[2], **Kate S. Reid-Bayliss**[2], **Lawrence D. True**[2], **Jerald P. Radich**[3], and **Lawrence A. Loeb**[2]

[1]Department of Medicine, Divisions of Hematology and Medical Oncology, University of Washington, Seattle, WA, USA

[2]Department of Pathology, University of Washington, Seattle, WA, USA

[3]Clinical Research Division, Fred Hutchinson Cancer Research Center, Seattle, WA, USA

Abstract

The detection of minority variants in mixed samples demands methods for enrichment and accurate sequencing of small genomic intervals. We describe an efficient approach based on sequential rounds of hybridization with biotinylated oligonucleotides, enabling more than one-million fold enrichment of genomic regions of interest. In conjunction with error correcting double-stranded molecular tags, our approach enables the quantification of mutations in individual DNA molecules.

Diseases such as cancer or viral infections do not manifest as a single population of cells, but rather, as a heterogeneous mixture of sub-clonal populations[1]. While massively parallel sequencing has made it feasible to scan whole genomes for clonal nucleotide variations, this approach cannot readily delineate the heterogeneity of mutations within a cell population. In order to detect rare, sub-clonal mutations, sequencing must be carried out to depths that can be prohibitively expensive, and at low frequencies it becomes difficult or impossible to distinguish sequencing-related errors from true variation. We overcome these challenges by coupling extensive purification of targeted sequences with highly accurate DNA sequencing.

Targeted capture approaches[2] sequence large genomic regions (typically hundreds of kilobases to several megabases), limiting the depth that can be obtained. These approaches

Corresponding author: Lawrence A. Loeb, laloeb@uw.edu.

Accession codes:
Data from this paper have been uploaded to the Sequence Read Archive under BioProject ID: PRJNA275267.

Competing Financial Interests:
The University of Washington has filed a patent application regarding Duplex Sequencing.

Author contributions:
M.W.S., E.J.F, M.J.P, K.S.R., L.D.T., J.P.R., and L.A.L. contributed to experimental design. M.W.S., E.J.F, and M.J.P. performed the experiments in the paper and analyzed data. E.J.F., L.D.T., and J.P.R. contributed patient samples. M.W.S. and L.A.L. wrote the manuscript.

Human subjects approval:
Use of human samples was approved by the Institutional Review Board at the University of Washington. Informed consent was obtained from patients who contributed samples.

do not scale to small targets (< 50 kb) and typically result in recovery of targeted DNA sequences of 5% or less. Small targets can be amplified by methods such as PCR or molecular inversion probes[3]; however these methods are error-prone and generate artifactual mutations that overwhelm the detection of sub-clonal variants[4].

Detection of sub-clonal and random mutations in a target gene also requires extremely accurate sequencing. Next-generation sequencing has a high error rate of 0.1%-1%, averaging one artifactual mutation in every sequencing read. Thus, millions of sequencing errors occur in every sequenced genome [5,6]. These errors can be averaged to obtain a single consensus sequence for a population of cells; however due to this high error rate it is not feasible to reliably detect mutations present in fewer than 5% of cells. Molecular tagging of single-stranded DNA prior to amplification[7,8,9] can reduce the frequency of erroneously called variants, but only by approximately 20-fold, since it cannot correct errors that occur in the first round of amplification and are propagated to subsequent copies[10].

To overcome these limitations, we developed an alternative approach based on sequential rounds of capture with individual biotinylated DNA oligonucleotides in conjunction with Duplex Sequencing, which uses double-stranded, complementary molecular tags to separately label and amplify each of the two strands of individual duplex DNA molecules[10]. In Duplex Sequencing, mutations are only scored if they occur at the same position on both DNA strands, whereas amplification and sequencing errors, which only appear in one strand, are not scored.

As a demonstration, we attempted to detect rare mutations in the *Abl* gene that confer resistance to imatinib (Gleevec) therapy of chronic myeloid leukemia (CML)[11]. We synthesized 5′ biotinylated DNA oligonucleotides corresponding to exons 4–7 of *Abl* (Supplementary Table 1). Duplex Sequencing adapters containing complementary molecular tag sequences that identify each of the two strands of individual DNA molecules were ligated to sheared human genomic DNA (Online Methods). The product was then PCR amplified and hybridized to the pooled *Abl*-targeting oligonucleotides followed by recovery with streptavidin beads. Elution and sequencing revealed a 50,000-fold enrichment of the target; however due to the small size of the target, this enrichment resulted in only 2%–5% of reads being on-target (Fig. 1a). The recovered DNA was then subjected to iterative rounds of PCR and capture. In two independent experiments, two rounds of capture resulted in >97% of reads mapping to the *Abl* gene. A third round of capture provided no further improvement (Fig. 1a).

The double-capture approach resulted in extremely high depth and uniformity of coverage (Fig. 1b). Conventional capture yielded a maximum on-target depth of 25,000x. In contrast, with equivalent use of sequencing capacity, double capture gave up to 1,000,000x depth, with an average and minimum depth of 830,000x and 250,000x, respectively. The Duplex Tags were then used to collapse PCR duplicates for which the two strands of individual DNA molecules were perfectly complementary into consensus sequences. This yielded an average of more than 1,000 unique DNA molecules sampled at every nucleotide position within the *Abl* target (Supplementary Fig. 1).

We used our protocol to sequence the *Abl* gene from an individual with CML at early relapse, following treatment with the targeted therapy imatinib. Conventional next-generation sequencing was unable to resolve any mutations in the sample (Fig. 2a). Even stringent quality filtering (requiring a minimum Phred quality score of 50) was unable to remove background errors, as many sequencing errors occur during PCR amplification and thus cannot be removed by quality filtering[10]. In contrast, Duplex Sequencing revealed a single mutation with a mutant fraction of 1% (Fig. 2b). This mutation, E279K, is known to confer imatinib resistance[11].

Alternative methods to detect subclonal mutations have been described that result in multiple copies of single-stranded DNA, linked together by concatemerization[12] or tagged with a molecular identifier sequence[8,9]. These approaches are inherently more error-prone than Duplex Sequencing, since they use information from only one of the two DNA strands and thus have less capability for error correction. To compare our double-stranded tagging approach to these methods, we re-analyzed our data using information from only one of the two tagged strands, which we refer to as "Single Strand Consensus Sequences"[10]. This analysis resulted in mutations at hundreds of positions in the *Abl* target (Supplementary Fig. 2), in contrast to the one true mutation that was found by Duplex Sequencing. The discrepancy indicates that >99% of mutations identified by the single-stranded tagging approach are artifacts.

We next determined whether our approach could scale to multiple targets. We obtained biotinylated oligonucleotides corresponding to the coding exons of the 5 human replicative DNA polymerases[13] (19.4 kb total target size) and applied the double-capture approach to DNA extracted from histologically normal human prostate and colon. More than 90% of reads mapped to the targeted genes, revealing mutation frequencies of 1×10^{-7} to 4×10^{-7} (Supplementary Table 2). Among the mutations, six were in introns and two changed the coding sequence of DNA polymerase epsilon (Supplementary Table 3). The frequency of mutations is in accord with prior estimates[14,15] of the spontaneous mutation rate in human cells, and thus could be the result of multiple rounds of cell division and endogenous mutagenic processes. Alternatively, these mutations could represent artifacts in our assay. However, the error frequency of Duplex Sequencing has been estimated to be $< 4 \times 10^{-10}$, as complementary errors would need to occur in both strands to be scored[10].

Our approach allows for the study of small genomic regions, such as individual human exons or viral sequences present at low levels in human samples. Due to the high level of enrichment, substantial depth can be obtained with modest sequencing. For example, a 1 kb target can be sequenced to 100,000-fold depth with 4×10^5 paired-end 125 nucleotide reads, and thus hundreds of samples can be sequenced simultaneously on a single lane of an Illumina HiSeq 2500. Thus the approach is highly scalable and cost-effective for sequencing small targets. Duplex Sequencing on larger targets, such as whole exomes, is also possible in principle with a greater use of sequencing capacity. For example, under optimized conditions, the full exome from 100 individual cells would require approximately 2×10^{11} nucleotides of sequence capacity, which is within the output range of currently available sequencers.

Our *Abl* results indicate that it is possible to assay for the presence of pre-existing sub-clones encoding resistance to targeted cancer therapies, which would be expected to clonally expand in the presence of corresponding inhibitors. Armed with this knowledge, patients could be treated with drugs chosen for their lack of any detectable resistance. Targeted, high-accuracy capture has additional applications in a wide range of fields, including the detection of tumor-specific circulating DNA as a biomarker for cancer treatment[16], detection of minimal residual disease in hematologic malignancies[17], confirming candidate sub-clonal mutations that are found by conventional sequencing, analysis of mutational processes in cancer[18], and testing for low-level resistance mutations in viral populations. Moreover, as the extreme accuracy of the approach results in a theoretical need of only 1x coverage of a genome to obtain an accurate sequence, we anticipate applications in settings where sample availability is extremely limited, such as paleogenomics, forensics, and the study of circulating tumor cells.

Online Methods

DNA isolation

Genomic DNA was extracted from peripheral blood mononuclear cells or tissue by high-salt extraction, using the Agilent DNA extraction kit #200600.

Ligation of Duplex Sequencing adapters

Duplex Sequencing was initially described with use of A-tailed adapters[10,19]; we have since found that T-tailed adapters result in improved ligation efficiency, and have published a detailed protocol for their synthesis and use[20]. In brief, DNA was sheared, end-repaired, and A-tailed, then ligated to T-tailed Duplex Sequencing adapters using a 20x molar excess of adapters relative to A-tailed DNA molecules. Following reaction cleanup with 1.0 volumes of Ampure XP beads (Agencourt), the adapter-ligated DNA was PCR amplified for 5 cycles with the KAPA Biosystems hot start high-fidelity kit, using primers mws13 and mws20. 240 nanograms of input DNA was used in each 100 microliter PCR reaction, with two to eight PCR reactions performed per sample. Due to the small amount of on-target DNA present in the starting sample, multiple PCR reactions are needed to amplify sufficient on-target DNA for capture. Each PCR reaction results in sequence data representing approximately 500 independent genomes; the number of PCR reactions performed can be adjusted depending on the sequencing coverage desired. The products from all reactions were pooled and purified with 1.2 volumes of Ampure XP beads, with a final elution volume of 50 microliters.

Targeted capture

One-third of the total amount of adapter-ligated DNA generated by PCR was combined with 5 micrograms of Cot-I DNA (Invitrogen) and 1 nanomole each of blocking oligonucleotides mws60 and mws61. The mixture was completely lyophilized, then resuspended in 2.5 microliters water, 7.5 microliters Nimblegen 2x hybridization buffer, and 3 microliters Nimblegen hybridization component A. The mixture was heated to 95°C for 10 minutes, the temperature was adjusted to 65°C, and 3 pmol of pooled 120nt biotinylated oligonucleotides were added (Integrated DNA Technologies). After 4 hours, M-270 streptavidin beads (Life

Technologies) were added and washes were performed according to the IDT xGen lockdown probe protocol version 2.0. We found that the standard quantity of streptavidin beads (the IDT protocol calls for 100 microliters of beads per 50 microliter PCR reaction) can result in PCR inhibition, so the quantity of beads was decreased to 75 microliters per reaction, and the PCR reaction volume increased to 100 microliters. The product was PCR amplified for 16 cycles with primers mws13 and mws20, and purified with 1.2 volumes of Ampure XP beads. The purified DNA was combined with 2.5 micrograms Cot-I DNA and 500 picomoles each of oligonucleotides mws60 and mws61, and a second round of capture[21] was performed with 1.5 picomoles of pooled biotinylated oligonucleotides. A final PCR reaction was carried out for 8–10 cycles with primers mws13 and mws21, which contains a fixed index sequence for multiplexing. After cleanup with 1.2 volumes of Ampure XP beads, the product was sequenced on an Illumina HiSeq 2500.

Data processing

Processing of Duplex Sequencing data was performed essentially as previously described[20]. Mutations identified by Duplex Sequencing were individually inspected in the Integrated Genome Viewer[22] to verify that they were not affected by alignment errors.

Code availability

Software for Duplex Sequencing is available at https://github.com/loeblab/Duplex-Sequencing

Reverse transcription PCR of the *Abl* gene

Total RNA was extracted from peripheral blood using TRIzol reagent (Invitrogen). An initial RT-PCR step with nested PCR was used to amplify exons 4 to 9 (codons 199 to 507) of the *Abl* kinase domain, and bidirectional Sanger sequencing of the PCR product was performed, as previously described[23].

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Research reported in this publication was supported by the National Institutes of Health under award numbers NCI P01-CA77852, R01-CA160674 and R33-CA181771 to L.A.L. and NCI U10-CA180861, P01-CA018029, R01-CA175008, and R01-CA175215 to J.P.R. We would like to thank Tom Walsh and Ming Lee for assistance with DNA sequencing.

References

1. Schmitt MW, Prindle MJ, Loeb LA. Ann N Y Acad Sci. 2012; 1267:110–116. [PubMed: 22954224]
2. Mamanova L, et al. Nat Methods. 2010; 7:111–118. [PubMed: 20111037]
3. Hardenbol P, et al. Nat Biotechnol. 2003; 21:673–678. [PubMed: 12730666]
4. Kanagawa T. J Biosci Bioeng. 2003; 96:317–323.10.1016/S1389-1723(03)90130-7 [PubMed: 16233530]
5. Fox EJ, Bayliss-Reid KS, Emond MJ, Loeb LA. Next Generat Sequenc & Applic. 2014; 1:106–109.
6. Glenn TC. Mol Ecol Resour. 2011; 11:759–769. [PubMed: 21592312]

7. Jabara CB, Jones CD, Roach J, Anderson JA, Swanstrom R. Proc Natl Acad Sci USA. 2011; 108:20166–20171. [PubMed: 22135472]
8. Kinde I, Wu J, Papadopoulos N, Kinzler KW, Vogelstein B. Proc Natl Acad Sci USA. 2011; 108:9530–9535. [PubMed: 21586637]
9. Hiatt JB, Pritchard CC, Salipante SJ, O'Roak BJ, Shendure J. Genome Res. 2013; 23:843–854. [PubMed: 23382536]
10. Schmitt MW, et al. Proc Natl Acad Sci USA. 2012; 109:14508–14513. [PubMed: 22853953]
11. Soverini S, et al. Leuk Res. 2014; 38:10–20. [PubMed: 24131888]
12. Lou DI, et al. Proc Natl Acad Sci USA. 2013; 110
13. Sweasy JB, Lauper JM, Eckert KA. Radiat Res. 2006; 166:693–714. [PubMed: 17067213]
14. Albertini RJ, Nicklas JA, O'Neill JP, Robison SH. Annu Rev Genet. 1990; 24:305–326. [PubMed: 2088171]
15. Kunkel TA. J Biol Chem. 2004; 279:16895–16898. [PubMed: 14988392]
16. Esposito A, et al. Cancer Treat Rev. 2014; 40:648–655. [PubMed: 24184333]
17. Buckley SA, Appelbaum FR, Walter RB. Bone Marrow Transplant. 2013; 48:630–641. [PubMed: 22825427]
18. Alexandrov LB, et al. Nature. 2013; 500:415–421. [PubMed: 23945592]
19. Kennedy SR, Salk JJ, Schmitt MW, Loeb LA. PLoS Genet. 2013; 9:e1004794.
20. Kennedy SR, et al. Nat Protoc. 2014; 9:2586–2606. [PubMed: 25299156]
21. Green, DJ.; Wendt, C.; Kashuk, M.; Brockman, M.; Burgess, D. SeqCap EZ Library: Technical notes. Roche NimbleGen, Inc; 2012. p. 1-12.
22. Robinson JT, et al. Nat Biotechnol. 2011; 29:24–26. [PubMed: 21221095]
23. Egan DN, Beppu L, Radich JP. Biol Blood Marrow Transplant. 2014; 14:567–569.

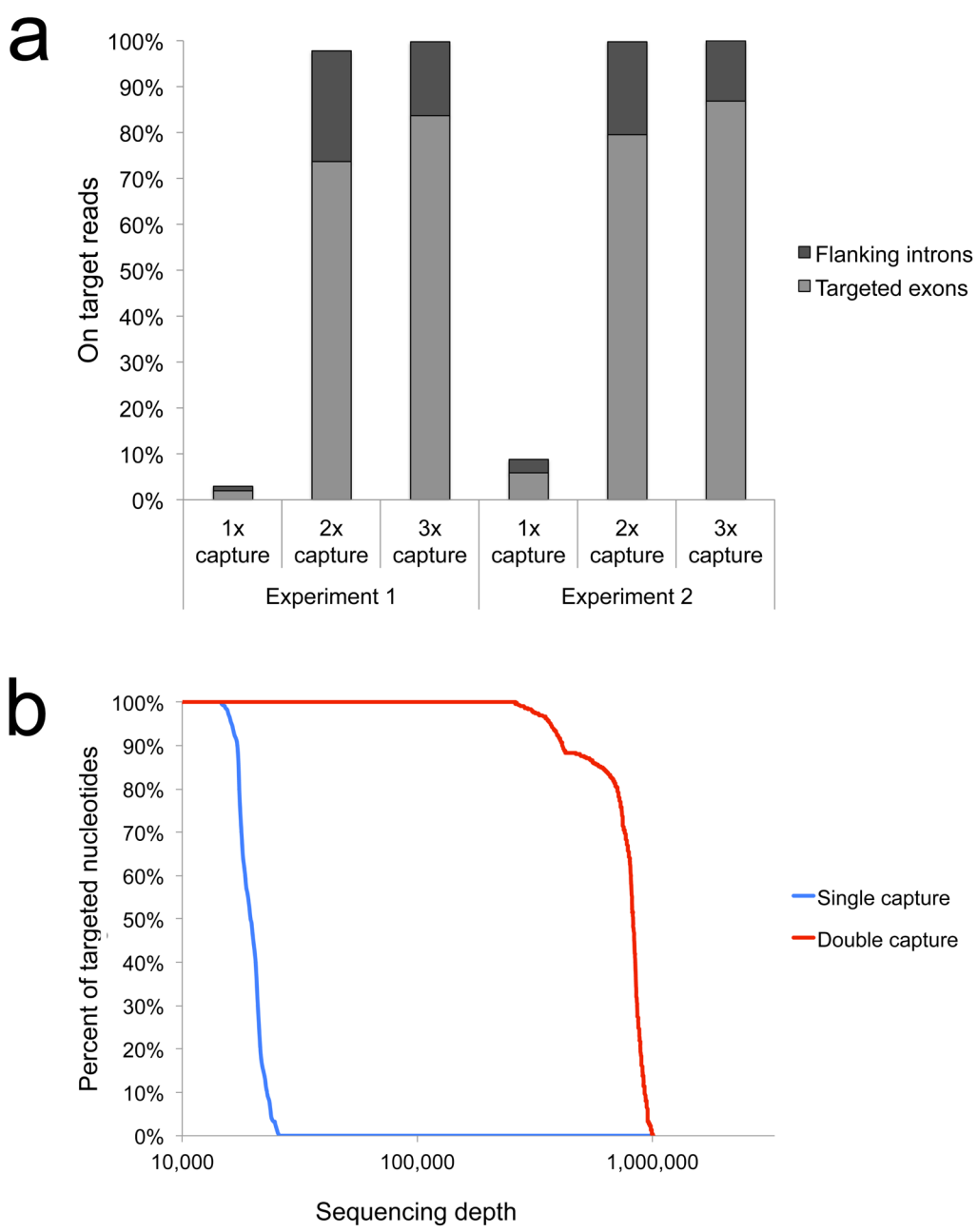

Figure 1.

High on-target recovery with sequential rounds of capture. (a) Human genomic DNA captured with biotinylated probes targeting *Abl* exons 4–7 results in low on-target recovery following one round of capture, while two rounds result in >97% of reads mapping to the targeted gene. Experiment 1 was carried out with conventional blocking oligonucleotides mws60 and mws61; experiment 2 used chemically modified high-affinity blocking oligonucleotides mws58 and mws59. (b) Percent of targeted nucleotides covered at a given sequencing depth following single and double capture. Both samples were sequenced on an

equivalent fraction of a HiSeq 2500 lane (5×10^6 paired-end reads, corresponding to 3% of a single lane).

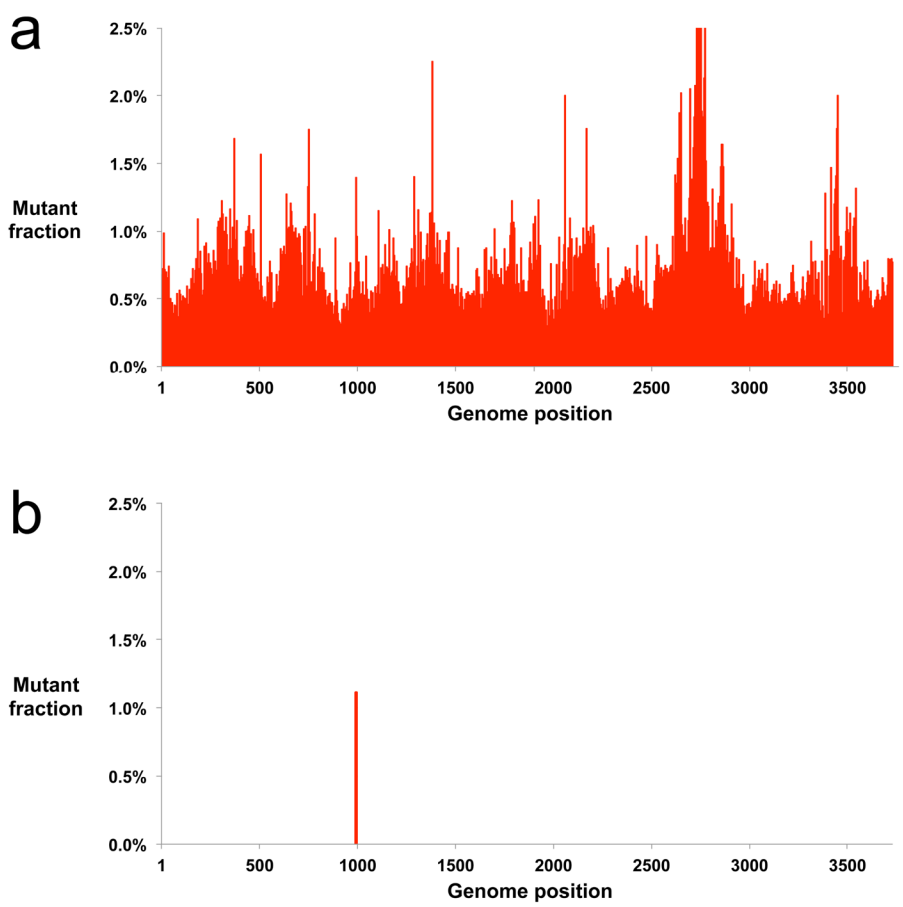

Figure 2.

Removal of sequencing artifacts by Duplex Sequencing. (a) Exons in *Abl* spanning the active site of the enzyme were enriched by the double-capture protocol and sequenced conventionally on an Illumina HiSeq 2500. Despite extremely stringent quality filtering (minimum Phred score 50), and removal of end-repair artifacts by 5-nucleotide trimming from read ends, true mutations cannot be discerned among the thousands of sequencing errors that persist. (b) Duplex Sequencing of the same sample reveals a single point mutation in *Abl* which confers imatinib resistance. The mutation was verified by RT-PCR and Sanger sequencing.