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] & Lawrence A Loeb[2]

The detection of minority variants in mixed samples requires methods for enrichment and accurate sequencing of small genomic intervals. We describe an efficient approach based on sequential rounds of hybridization with biotinylated oligonucleotides that enables more than 1-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 subclonal populations[1]. Although 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, subclonal 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); sequencing at great depth is impractical for targets of this size owing to the large amount of sequencing capacity that would be required. These approaches 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 subclonal variants[4].

Detection of subclonal and random mutations in a target gene also requires extremely accurate sequencing. High-throughput 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, owing to this high error rate it is not feasible to reliably detect mutations present in fewer than 5% of cells. Molecular tagging of ssDNA before amplification[7–9] can reduce the frequency of erroneously called variants—but only by approximately 20-fold, as 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 scored only if they occur at the same position on both DNA strands, whereas amplification and sequencing errors, which appear in only one strand, are not scored.

As a demonstration, we attempted to detect rare mutations in the *ABL1* gene that confer resistance to imatinib (Gleevec) therapy of chronic myeloid leukemia[11]. We synthesized 5′-biotinylated DNA oligonucleotides corresponding to exons 4–7 of *ABL1* (**Supplementary Table 1**). Duplex sequencing adaptors 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 *ABL1*-targeting oligonucleotides, and hybridization was followed by recovery with streptavidin beads. Elution and sequencing revealed a 50,000-fold enrichment of the target; however, owing 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 *ABL1* 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 25,000×. In contrast, with equivalent use of sequencing capacity, double capture gave up to 1,000,000× depth, with average and minimum depths of 830,000× and 250,000×, respectively. The duplex tags were then used to collapse into consensus sequences the PCR duplicates for which the two strands of individual DNA molecules were perfectly complementary. This yielded an average of more than 1,000 unique DNA molecules sampled at every nucleotide position within the *ABL1* target (**Supplementary Fig. 1**).

[1]Department of Medicine, Divisions of Hematology and Medical Oncology, University of Washington, Seattle, Washington, USA. [2]Department of Pathology, University of Washington, Seattle, Washington, USA. [3]Clinical Research Division, Fred Hutchinson Cancer Research Center, Seattle, Washington, USA. Correspondence should be addressed to L.A.L. (laloeb@uw.edu).

RECEIVED 19 NOVEMBER 2014; ACCEPTED 19 MARCH 2015; PUBLISHED ONLINE 6 APRIL 2015; DOI:10.1038/NMETH.3351

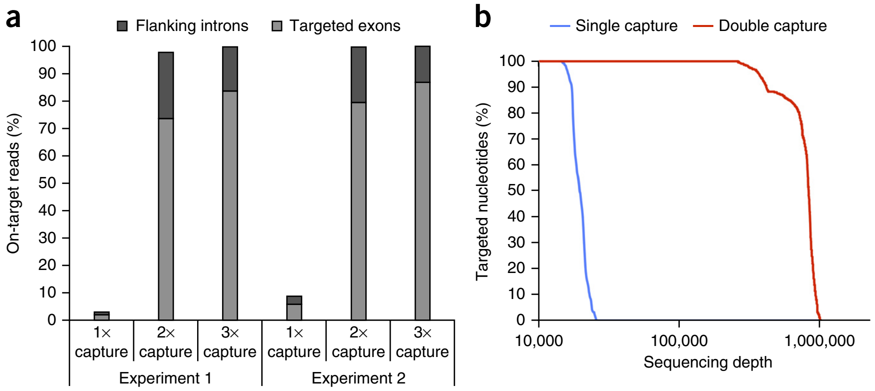

Figure 1 | High on-target recovery with sequential rounds of capture. (**a**) Human genomic DNA captured with biotinylated probes targeting *ABL1* exons 4–7 results in low on-target recovery after one round of capture, whereas two rounds result in >97% of reads mapping to the targeted gene. Experiment 1 was carried out with conventional blocking oligonucleotides mws60 and mws61; experiment 2 used chemically modified high-affinity blocking oligonucleotides mws58 and mws59 (**Supplementary Table 1**). (**b**) Percentage of targeted nucleotides covered at a given sequencing depth after 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).

We used our protocol to sequence the *ABL1* gene from an individual with chronic myeloid leukemia who relapsed after treatment with the targeted therapy imatinib. Conventional high-throughput 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 ssDNA linked together by concatemerization[12] or tagged with a molecular identifier sequence[8,9]. These approaches are inherently more error prone than duplex sequencing because 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 reanalyzed 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 *ABL1* 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 five human replicative DNA polymerases[13] (total target size, 19.4 kb) 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. Owing to the high level of enrichment, significant 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-nt reads, and thus hundreds of samples can be sequenced simultaneously on a single lane of an Illumina HiSeq 2500. The approach is therefore 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

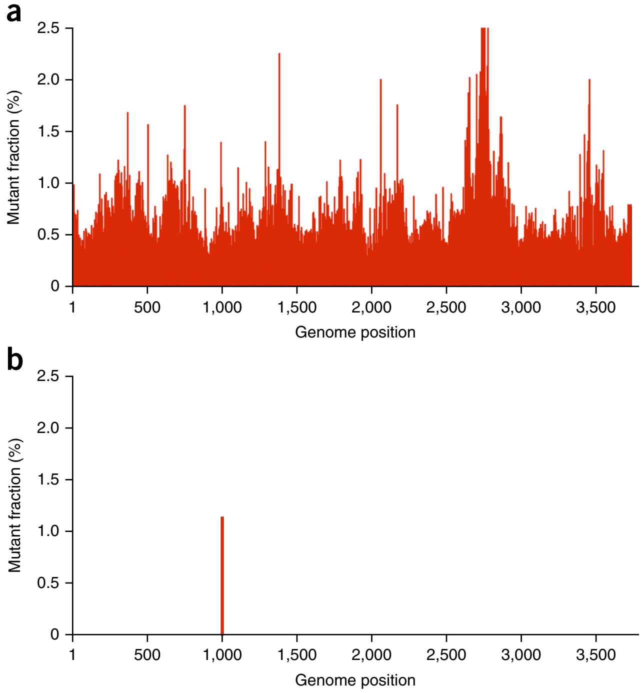

Figure 2 | Removal of sequencing artifacts by duplex sequencing. (**a**) Exons in *ABL1* spanning the active site of the enzyme were enriched by the double-capture protocol and sequenced conventionally on an Illumina HiSeq 2500. Despite extremely stringent quality filtering (minimum Phred score = 50) and removal of end-repair artifacts by 5-nt 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 *ABL1* that confers imatinib resistance. The mutation was verified by reverse-transcription PCR and Sanger sequencing.

conditions, the full exome from 100 individual cells would require approximately 2 × 10^{11} nt of sequence capacity, which is within the output range of currently available sequencers.

Our *ABL1* results indicate that it is possible to assay for the presence of preexisting subclones encoding resistance to targeted cancer therapies, which would be expected to clonally expand in the presence of corresponding inhibitors. Armed with this knowledge, physicians could treat patients 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], the detection of minimal residual disease in hematologic malignancies[17], confirming candidate subclonal 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 1× 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.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. NCBI BioProject: PRJNA275267.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

Research reported in this publication was supported by the US 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 thank T. Walsh and M. Lee for assistance with DNA sequencing.

AUTHOR CONTRIBUTIONS

M.W.S., E.J.F., M.J.P., K.S.R.-B., 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.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the online version of the paper.

1. Schmitt, M.W., Prindle, M.J. & Loeb, L.A. *Ann. NY Acad. Sci.* **1267**, 110–116 (2012).
2. Mamanova, L. *et al. Nat. Methods* **7**, 111–118 (2010).
3. Hardenbol, P. *et al. Nat. Biotechnol.* **21**, 673–678 (2003).
4. Kanagawa, T. *J. Biosci. Bioeng.* **96**, 317–323 (2003).
5. Fox, E.J., Reid-Bayliss, K.S., Emond, M.J. & Loeb, L.A. *Next Gener. Seq. Appl.* **1**, 106–109 (2014).
6. Glenn, T.C. *Mol. Ecol. Resour.* **11**, 759–769 (2011).
7. Jabara, C.B., Jones, C.D., Roach, J., Anderson, J.A. & Swanstrom, R. *Proc. Natl. Acad. Sci. USA* **108**, 20166–20171 (2011).
8. Kinde, I., Wu, J., Papadopoulos, N., Kinzler, K.W. & Vogelstein, B. *Proc. Natl. Acad. Sci. USA* **108**, 9530–9535 (2011).
9. Hiatt, J.B., Pritchard, C.C., Salipante, S.J., O'Roak, B.J. & Shendure, J. *Genome Res.* **23**, 843–854 (2013).
10. Schmitt, M.W. *et al. Proc. Natl. Acad. Sci. USA* **109**, 14508–14513 (2012).
11. Soverini, S. *et al. Leuk. Res.* **38**, 10–20 (2014).
12. Lou, D.I. *et al. Proc. Natl. Acad. Sci. USA* **110**, 19872–19877 (2013).
13. Sweasy, J.B., Lauper, J.M. & Eckert, K.A. *Radiat. Res.* **166**, 693–714 (2006).
14. Albertini, R.J., Nicklas, J.A., O'Neill, J.P. & Robison, S.H. *Annu. Rev. Genet.* **24**, 305–326 (1990).
15. Kunkel, T.A. *J. Biol. Chem.* **279**, 16895–16898 (2004).
16. Esposito, A. *et al. Cancer Treat. Rev.* **40**, 648–655 (2014).
17. Buckley, S.A., Appelbaum, F.R. & Walter, R.B. *Bone Marrow Transplant.* **48**, 630–641 (2013).
18. Alexandrov, L.B. *et al. Nature* **500**, 415–421 (2013).

ONLINE METHODS

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.

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 adaptors. Duplex sequencing was initially described with use of A-tailed adaptors[10,19]; we have since found that T-tailed adaptors result in improved ligation efficiency, and we have published a detailed protocol for their synthesis and use[20]. In brief, DNA was sheared, end repaired, A tailed and then ligated to T-tailed duplex sequencing adaptors using a 20× molar excess of adaptors relative to A-tailed DNA molecules. Following reaction cleanup with 1.0 volumes of Ampure XP beads (Agencourt), the adaptor-ligated DNA was PCR amplified for five cycles with the KAPA Biosystems hot-start high-fidelity kit using primers mws13 and mws20 (**Supplementary Table 1**). 240 ng of input DNA were used in each 100-µl PCR reaction, with 2–8 PCR reactions performed per sample. Owing 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 µl.

Targeted capture. One-third of the total amount of adaptor-ligated DNA generated by PCR was combined with 5 µg of Cot-I DNA (Invitrogen) and 1 nmol each of blocking oligonucleotides mws60 and mws61. The mixture was completely lyophilized and then resuspended in 2.5 µl water, 7.5 µl NimbleGen 2× hybridization buffer and 3 µl NimbleGen hybridization component A. The mixture was heated to 95 °C for 10 min, the temperature was adjusted to 65 °C and 3 pmol of pooled biotinylated oligonucleotides were added (Integrated DNA Technologies). After 4 h, 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 µl of beads per 50-µl PCR reaction) can result in PCR inhibition, so the quantity of beads was decreased to 75 µl per reaction, and the PCR reaction volume increased to 100 µl. 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 µg Cot-I DNA and 500 pmol each of oligonucleotides mws60 and mws61, and a second round of capture[21] was performed with 1.5 pmol of pooled biotinylated oligonucleotides. A final PCR reaction was carried out for 8–10 cycles with primers mws13 and mws21, which contain 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. Software for duplex sequencing is available at https://github.com/loeblab/Duplex-Sequencing/. Data from this paper have been uploaded to the Sequence Read Archive under BioProject ID: PRJNA275267.

Reverse-transcription PCR of the *ABL1* 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 4–9 (codons 199–507) of the *ABL1* kinase domain, and bidirectional Sanger sequencing of the PCR product was performed, as previously described[23].

19. Kennedy, S.R., Salk, J.J., Schmitt, M.W. & Loeb, L.A. *PLoS Genet.* **9**, e1003794 (2013).
20. Kennedy, S.R. *et al. Nat. Protoc.* **9**, 2586–2606 (2014).
21. Burgess, D. *et al.* SeqCap EZ Library: Technical Note http://www.nimblegen.com/products/lit/06870406001_NG_SeqCapEZ_DoubleCaptureSR_20Aug2012.pdf (Roche NimbleGen, (2012)).
22. Robinson, J.T. *et al. Nat. Biotechnol.* **29**, 24–26 (2011).
23. Egan, D.N., Beppu, L. & Radich, J.P. *Biol. Blood Marrow Transplant.* **21**, 184–189 (2014).

doi:10.1038/nmeth.3351