Quantifying ultra-rare pre-leukemic clones via targeted error-corrected sequencing

Leukemia (2015) 29, 1608–1611; doi:10.1038/leu.2015.17

The quantification of rare clonal and subclonal populations from a heterogeneous DNA sample has multiple clinical and research applications for the study and treatment of leukemia. Specifically, in the hematopoietic compartment, recent reports demonstrate the presence of subclonal variation in normal and malignant hematopoiesis,1,2 and leukemia is now recognized as an oligoclonal disease.3 Currently, clonal heterogeneity in leukemia is studied using next-generation sequencing (NGS) targeting subclone-specific mutations. With this method, detecting mutations at 2–5% variant allele fraction (VAF) requires costly and time-intensive deep sequencing and identifying lower frequency variants is impractical regardless of sequencing depth. Recently, various methods have been developed to circumvent the error rate of NGS.4,5 These methods tag individual DNA molecules with unique oligonucleotide indexes, which enable error correction after sequencing.

Here we present a direct application of error-corrected sequencing (ECS) to study clonal heterogeneity during leukemogenesis and validate the accuracy of this method with a series of benchmarks for ECS and the identification of rare pre-leukemic mutations. (a, b) DNA extracted from a diagnostic leukemia sample with known mutations in RUNX1 (a) and IDH2 (b) was serially diluted into non-cancer, unrelated human DNA. Two replicates were run per sample/dilution. The coefficient of determination (r²) between diluted tumor concentration in the sample and VAF in the generated read families was 0.9999 and 0.9991 for RUNX1 and IDH2, respectively. (c) The VAF at every nucleotide not expected to contain mutations in the error-corrected consensus sequences compared with conventional deep sequencing. A cumulative distribution function of VAF demonstrated a reduced error rate in read families relative to conventional deep sequenced reads. (d) The most frequent class of substitution seen in read families was in G to T (C to A) transversions, which was consistent with oxidative conversion of guanine to 8-oxo-guanine. (e, f) The leukemia-specific variants identified in ASXL1 and U2AF1 at diagnosis (circled) were not distinguishable from sequencing errors in the same substitution class by conventional deep sequencing. (g, h) Targeted error-corrected sequencing identified the ASXL1 variant in the 2002 banked sample at 0.004 VAF and the U2AF1 variant in the 2004 banked sample at 0.009 VAF.
experiment to assess bias during library preparation and consensus sequence (ECCS). We performed a dilution series within a read family and removed to create an error-corrected family. Sequencing errors are identified originating from the same molecule are grouped into read families. Following sequencing, sequence reads containing the same index adapters containing random indexes instead of the manufacturer’s supplied adapters and a quantitative PCR (qPCR) quantification step before sequencing (Supplementary Table 1). Following sequencing, sequence reads containing the same index and originating from the same molecule are grouped into read families. Sequencing errors are identified by comparing reads within a read family and removed to create an error-corrected consensus sequence (ECCS). We performed a dilution series experiment to assess bias during library preparation and determine the limit of detection for ECS. For this experiment, we spiked DNA from a t-AML sample into control human DNA, which was serially diluted over five orders of magnitude. The experiment was comprised of two technical replicates targeting two separate mutations (20 total independent libraries). The results demonstrate that ECS is quantitative to a VAF of 1:10,000 molecules and provides a highly reproducible digital readout of tumor DNA prevalence in a heterogeneous DNA sample (r² of 0.9999 and 0.9991, Figures 1a and b). We next characterized the error profile based on the wild-type nucleotides included in the dilution series experiment. Variant identification using the ECCSs was 99% specific at a VAF of 0.0016 versus 0.0140 for deep sequencing alone (Figure 1c). We noticed that ECCS errors were heavily biased towards G to T transversions and to a lesser degree C to T transitions (Figure 1d, Supplementary Figure 2), as previously observed.4,9 When separated by substitution type, variants identified from the ECCSs were 99% specific at a VAF of 0.0034 for G to T (C to A) mutations, 0.00020 for C to T (G to A) mutations and 0.000079 for the other eight possible substitutions. Although excess G to T mutations are a known consequence of DNA oxidation leading to 8-oxo-guanine conversion,4 the pre-treatment of samples with formamidopyrimidine-DNA glycosylase before PCR amplification did not appreciably improve the error profile of G to T mutations (Supplementary Figure 3).

As proof of principle, we applied ECS to study rare pre-leukemic clonal hematopoiesis in seven individuals who later developed t-AML/t-MDS. Leukemia/normal whole-genome sequencing at diagnosis was used to identify the leukemia-specific somatic mutations in each patient’s malignancy (Supplementary Table 2). We applied targeted ECS to query these 18 different loci in 10 cryopreserved or formalin-fixed paraffin-embedded blood and bone marrow samples that were 9–22-year old and banked up to 12 years before diagnosis (Supplementary Table 3).

We generated ~25 Gb of 150 bp paired-end reads from six Illumina (San Diego, CA, USA) MiSeq runs. We targeted 1–7 somatic mutations per individual (25 mutations spanning 5.5 kb from 15 genes in total) and identified leukemia-specific subclonal populations in four individuals up to 12 years before diagnosis.

Table 1. Patient-specific leukemia-associated somatic mutations identified by ECS

| UPN   | Sample ID | Years prior | Gene | Chr | Position | Mutation | Amino-acid change | Variant RFs | Reference RFs | VAF   |
|-------|-----------|-------------|------|-----|----------|----------|-------------------|-------------|--------------|-------|
| 446294 | 75.02     | 1           | OBSCN| 1   | 228461129| A to G   | H1857R           | 61.238      | 156.986      | 0.2806 |
| 499258 | 24.06     | 2           | TP53 | 17  | 75782721 | T to A   | H193L            | 220.551     | 110.047      | 0.6671 |
| 574214 | 26.04     | 7           | RUNX1| 21  | 36252865 | C to G   | R139P            | 486.196     | 0            |       |
| 643006 | 80.01     | 12          | ASXL1| 20  | 31022442| G to T   | G645C            | 7           | 85.781       | 0.0001 |
| 643006 | 80.01     | 12          | ASXL1| 20  | 31022442| del G    | G645fs            | 2.898       | 82.245       | 0.034  |
| 684949 | 91.01     | 5           | U2AF | 21  | 44524456 | G to T   | S34Y             | 85          | 414.613      | 0.0002 |
| 684949 | 91.01     | 5           | U2AF | 21  | 44524456 | G to T   | S34Y             | 85          | 414.613      | 0.0002 |
| 684949 | 91.01     | 5           | U2AF | 21  | 44524456 | G to T   | S34Y             | 85          | 414.613      | 0.0002 |
| 856024 | 30.02     | 1           | ASXL1| 20  | 31023112| G to T   | L866*            | 3.583       | 853.598      | 0.0042 |
| 856024 | 30.02     | 1           | IGSF8| 1   | 160062252| G to T   | P516S            | 545.3       | 514.410      | 0.0011 |
| 942008 | 33.04     | 9           | IDH2 | 15  | 90631934| G to T   | R88Q             | 23.170      | 236.587      | 0.0892 |
| 107.01 | <1        | 1           | RUNX1| 21  | 36321791| T to C   | D171G            | 40          | 253.168      | 0.0002 |

Abbreviations: ECS, error-corrected sequencing; RFs, read families; VAF, variant allele fraction. Two to seven mutations were queried per individual and the number of read families containing the variant allele or reference allele were reported and used to calculate the variant allele fraction.
targets for ECS enables the surveillance of known mutations and the simultaneous discovery of new somatic mutations. Ongoing work will directly compare gold-standard MRD methods with targeted ECS in patients with and without relapsed leukemia.

CONFLICT OF INTEREST
The authors declare no conflict of interest.

AL Young¹,², TN Wong³, AEO Hughes¹,², SE Heath³, TJ Ley¹,²
¹Department of Pediatrics, Division of Hematology and Oncology, Washington University School of Medicine, Saint Louis, MO, USA;
²Center for Genome Sciences and Systems Biology, Washington University School of Medicine, Saint Louis, MO, USA and
³Department of Medicine, Division of Oncology, Washington University School of Medicine, Saint Louis, MO, USA
E-mail: Druley_t@kids.wustl.edu

REFERENCES
1 Holstege H, Pfeiffer W, Sie D, Hulsman M, Nicholas TJ, Lee CC et al. Somatic mutations found in the healthy blood compartment of a 115-year-old woman demonstrate oligoclonal hematopoiesis. Genome Res 2014; 24: 733–742.
2 Walter MJ, Shen D, Ding L, Shao J, Koboldt DC, Chen K et al. Clonal architecture of secondary acute myeloid leukemia. N Engl J Med 2012; 366: 1090–1098.
3Welch JS, Ley TJ, Link DC, Miller CA, Larson DE, Koboldt DC et al. The Origin and Evolution of Mutations in Acute Myeloid Leukemia. Cell 2012; 150: 264–278.
4 Schmitz MW, Kennedy SR, Salk JJ, Fox EJ, Hiatt JB, Loeb LA. Detection of ultra-rare mutations by next-generation sequencing. Proc Natl Acad Sci USA 2012; 109: 14508–14513.
5 Kinde I, Wu J, Papadopoulos N, Knizler KW, Vogelstein B. Detection and quantification of rare mutations with massively parallel sequencing. Proc Natl Acad Sci USA 2011; 108: 9530–9535.
6 Godley LA, Larson RA. Therapy-related myeloid leukemia. Semin Oncol 2008; 35: 418–429.
7Wong T, Ramsingh G, Young AL, Miller CA, Touma W, Welch JS et al. The role of TP53 mutations in the origin and evolution of therapy-related AML. Nature 2015; 518: 552–555.
8 Fu GK, Xu W, Wilhelmy J, Mindinios MN, Davis RW, Xiao W et al. Molecular indexing enables quantitative targeted RNA sequencing and reveals poor efficiencies in standard library preparations. Proc Natl Acad Sci USA 2014; 111: 1891–1896.
9 Lou DI, Hussmann JA, McBee RM, Acevedo A, Andino R, Press WH et al. High-throughput DNA sequencing errors are reduced by orders of magnitude using circle sequencing. Proc Natl Acad Sci USA 2013; 110: 19872–19877.
10 Cancer Genome Atlas Research Network. Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. N Engl J Med 2013; 368: 2059–2074.
11 Salipante SJ, Fromm JR, Shendure J, Wood BL, Wu D. Detection of minimal residual disease in NPM1-mutated acute myeloid leukemia by next-generation sequencing. Mod Pathol 2014; 27: 1438–1446.
12 Kohlmann a, Nadarajah N, Alpermann T, Grossmann V, Schindela S, Dicker F et al. Monitoring of residual disease by next-generation deep-sequencing of RUNX1 mutations can identify acute myeloid leukemia patients with resistant disease. Leukemia 2014; 28: 129–137.
13 Loman NJ, Misra RV, Dallman TJ, Constantinidou C, Gharbia SE, Wain J et al. Performance comparison of benchtop high-throughput sequencing platforms. Nat Biotechnol 2012; 30: 434–439.
14 Hourigan CS, Karp JE. Minimal residual disease in acute myeloid leukaemia. Nat Rev Clin Oncol 2013; 10: 460–471.

This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/