A Pardanani[1], T Lasho[1], D Chen[2], TK Kimlinger[1], C Finke[1], D Zblewski[1], MM Patnaik[1], KK Reichard[2], E Rowinsky[3], CA Hanson[2], C Brooks[3] and A Tefferi[1]
[1]*Division of Hematology and Department of Medicine, Rochester, MN, USA;*
[2]*Division of Hematopathology and Department of Laboratory Medicine and Pathology, Mayo Clinic, Rochester, MN, USA and*
[3]*Stemline Therapeutics Inc., New York, NY, USA*
E-mail: Pardanani.animesh@mayo.edu

REFERENCES

1 Woodcock JM, Bagley CJ, Zacharakis B, Lopez AF. A single tyrosine residue in the membrane-proximal domain of the granulocyte-macrophage colony-stimulating factor, interleukin (IL)-3, and IL-5 receptor common beta-chain is necessary and sufficient for high affinity binding and signaling by all three ligands. *J Biol Chem* 1996; **271**: 25999–26006.
2 Metcalf D, Begley CG, Johnson GR, Nicola NA, Lopez AF, Williamson DJ. Effects of purified bacterially synthesized murine multi-CSF (IL-3) on hematopoiesis in normal adult mice. *Blood* 1986; **68**: 46–57.
3 Testa U, Pelosi E, Frankel A. CD 123 is a membrane biomarker and a therapeutic target in hematologic malignancies. *Biomark Res* 2014; **2**: 4.
4 Jordan CT, Upchurch D, Szilvassy SJ, Guzman ML, Howard DS, Pettigrew AL *et al.* The interleukin-3 receptor alpha chain is a unique marker for human acute myelogenous leukemia stem cells. *Leukemia* 2000; **14**: 1777–1784.
5 Florian S, Sonneck K, Hauswirth AW, Krauth MT, Schernthaner GH, Sperr WR *et al.* Detection of molecular targets on the surface of CD34+/CD38— stem cells in various myeloid malignancies. *Leuk Lymphoma* 2006; **47**: 207–222.
6 Frolova O, Benito J, Brooks C, Wang RY, Korchin B, Rowinsky EK *et al.* SL-401 and SL-501, targeted therapeutics directed at the interleukin-3 receptor, inhibit the growth of leukaemic cells and stem cells in advanced phase chronic myeloid leukaemia. *Br J Haematol* 2014; **166**: 862–874.
7 He SZ, Busfield S, Ritchie DS, Hertzberg MS, Durrant S, Lewis ID *et al.* A phase 1 study of the safety, pharmacokinetics, and anti-leukemic activity of the anti-CD123 monoclonal antibody, CSL360, in relapsed, refractory or high-risk acute myeloid leukemia (AML). *Leuk Lymphoma* 2014; e-pub ahead of print 20 November 2014; doi:10.3109/10428194.2014.956316.
8 Frankel AE, Konopleva M, Hogge D, Rizzieri D, Brooks C, Cirrito T *et al.* Activity and tolerability of SL-401, a targeted therapy directed to the interleukin-3 receptor on cancer stem cells and tumor bulk, as a single agent in patients with advanced hematologic malignancies. *J Clin Oncol* 2013; **31**: 15.
9 Frankel AE, Woo JH, Ahn C, Pemmaraju N, Medeiros BC, Carraway HE *et al.* Activity of SL-401, a targeted therapy directed to interleukin-3 receptor, in blastic plasmacytoid dendritic cell neoplasm patients. *Blood* 2014; **124**: 385–392.
10 Valent P, Besemer J, Sillaber C, Butterfield JH, Eher R, Majdic O *et al.* Failure to detect IL-3-binding sites on human mast cells. *J Immunol* 1990; **145**: 3432–3437.
11 Teodosio C, Garcia-Montero AC, Jara-Acevedo M, Sanchez-Munoz L, Alvarez-Twose I, Nunez R *et al.* Mast cells from different molecular and prognostic subtypes of systemic mastocytosis display distinct immunophenotypes. *J Allergy Clin Immunol* 2010; **125**: 719–726.
12 Moonim MT, Kossier T, van Der Walt J, Wilkins B, Harrison CN, Radia DH. CD30/CD123 expression in systemic mastocytosis does not correlate with aggressive disease. *Blood* 2012; **120**: 21.
13 Horny HP, Metcalfe DD, Bennett JM, Bain BJ, Akin C, Escribano L *et al.* Mastocytosis. In: Swerdlow SH, Campo E, Harris NL, Jaffe ES, Pileri SA, Stein H *et al.* (eds) *WHO Classification of Tumors of Hematopoietic and Lymphoid Tissues*, 4th edn. International Agency for Research and Cancer (IARC): Lyon, 2008; pp 54–63.
14 Pardanani A. Systemic mastocytosis in adults: 2013 update on diagnosis, risk stratification, and management. *Am J Hematol* 2013; **88**: 612–624.
15 Gotlib J, Kluin-Nelemans HC, George TI, Akin C, Sotlar K, Hermine O *et al.* Durable responses and improved quality of life with midostaurin (PKC412) in advanced systemic mastocytosis (SM): updated stage 1 results of the Global D2201 trial. *Blood* 2013; **122**: 21.

Supplementary Information accompanies this paper on the Leukemia website (http://www.nature.com/leu)

OPEN

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

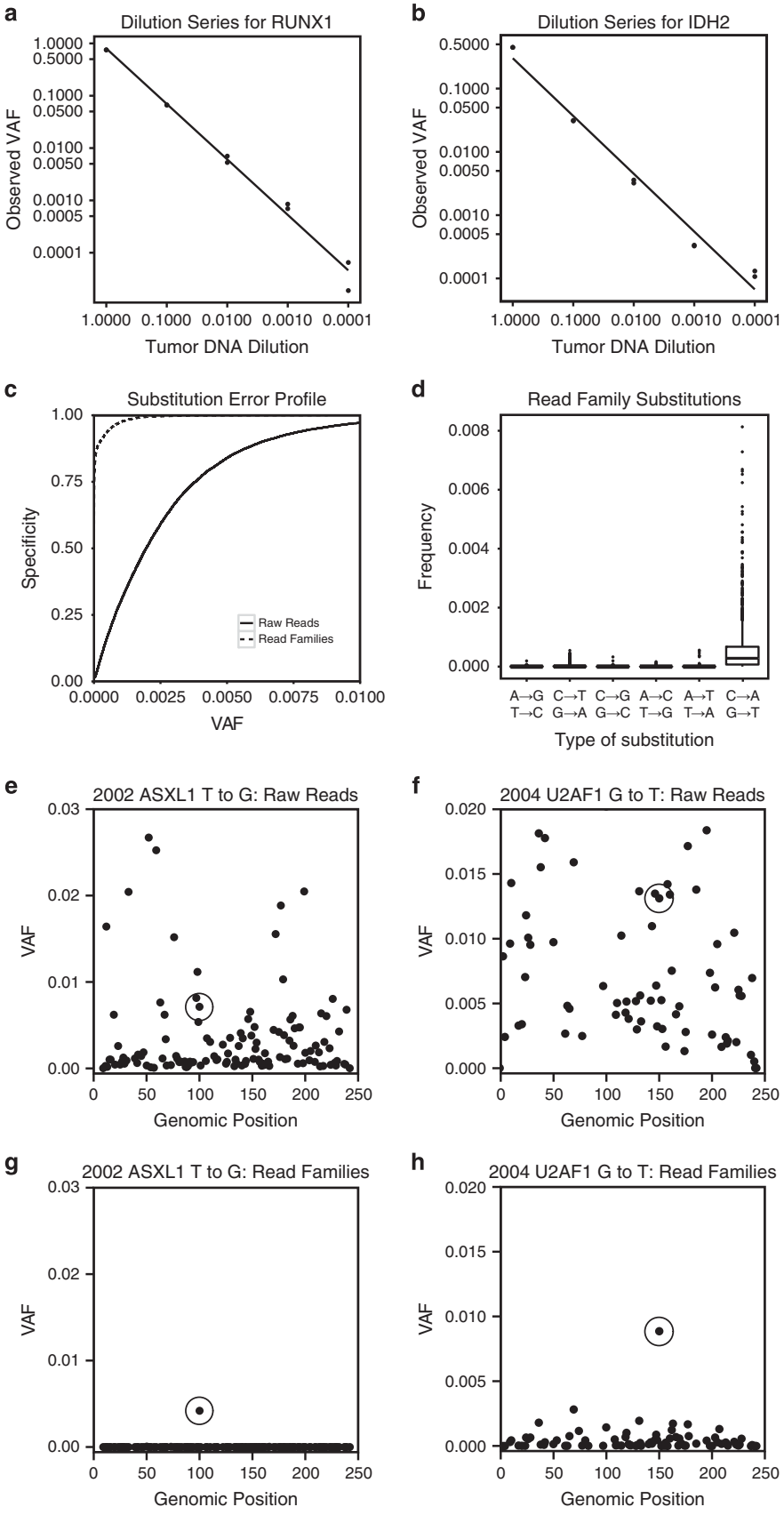
Leukemia (2015) **29,** 1608–1611; doi:10.1038/leu.2015.17

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 resequencing 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

Figure 1. Benchmarking 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 (r^2) 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 dilution series experiment were analyzed to determine the error profile of the error-corrected consensus sequences compared with conventional deep sequencing. A cumulative distribution function of VAF demonstrated a reduced error profile 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.

Accepted article preview online 3 February 2015; advance online publication, 20 February 2015



of benchmarking experiments. Specifically, we demonstrate the ability of ECS to identify leukemia-associated mutations in banked pre-leukemic blood and bone marrow from patients with either therapy-related acute myeloid leukemia (t-AML) or therapy-related myelodysplastic syndrome (t-MDS). T-AML/t-MDS occurs in 1–10% of individuals who receive alkylator- or epipodophyllotoxin-based chemotherapy or radiation to treat a primary malignancy.[6] For the seven individuals surveyed in this study, matched leukemia/normal whole-genome sequencing identified the t-AML/t-MDS-specific somatic mutations present at diagnosis. We applied our method for ECS to identify leukemia-specific mutations in four individuals from DNA extracted from blood and bone marrow samples collected years before diagnosis. In a separate study into the role of *TP53* mutations in t-AML/t-MDS leukemogenesis, this method was used to identify leukemia-associated mutations at low frequency in samples banked years before diagnosis.[7] In two cases, subclones were identified below the 1% threshold of detection governed by conventional NGS. These results highlight the ability of targeted ECS to identify clinically silent single-nucleotide variations (SNVs).

We employed ECS by tagging individual DNA molecules with adapters containing 16 bp random oligonucleotide molecular indexes in a manner similar to other reports.[4,5,8] Our implementation of ECS easily targets loci of interest by single or multiplex PCR and inserts seamlessly into the standard NGS library preparation (Supplementary Figure 1, Supplementary Methods). Our only deviations from the standard protocol are ligation of customized 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 (r^2 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*	*Mut*	*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
			TP53	17	7578271	T to A	H193L	220 551	110 047	0.6671
499258	24.06	2	RUNX1	21	36252865	C to G	R139P	2	486 196	0
574214	26.04	7	DMD	X	32827676	G to A	R187*	7	199 945	0
643006	80.01	12	ASXL1	20	31022448	G to T	G645C	7	85 781	0.0001
			ASXL1	20	31022442	del G	G645fs	2 898	82 245	0.034
			GATA2	3	128200135	del CTT	K390in_fr_del	0	4 187	0
			U2AF1	21	44524456	G to T	S34Y	85	414 613	0.0002
684949	91.01	5	ASXL1	20	31023112	T to G	L866*	3 583	853 598	0.0042
			U2AF1	21	44524456	G to T	S34Y	545	514 410	0.0011
	92.02	4	ASXL1	20	31023112	T to G	L866*	54 074	535 976	0.0916
			U2AF1	21	44524456	G to T	S34Y	11 195	355 276	0.0305
	93.01	3	ASXL1	20	31023112	T to G	L866*	17 319	573 629	0.0293
			U2AF1	21	44524456	G to T	S34Y	827	92 104	0.0089
856024	30.02	1	S100A4	1	153517192	A to G	F27L	0	211 512	0
			IGSF8	1	160062252	G to A	P516S	0	22 614	0
			PLA2R1	2	160798389	A to G	L1431P	2	338 616	0
			POU3F2	6	99282794	C to A	S15R	8	201 240	0
			ANKRD18B	9	33524645	G to A	C53Y	7	214 836	0
			ESR2	14	64701847	G to A	A416V	10	135 861	0.0001
			FBN3	19	8155081	G to A	P2029L	0	152 304	0
942008	33.04	9	IDH2	15	90631934	C to T	R88Q	23 170	236 587	0.0892
			RUNX1	21	36231791	T to C	D171G	40	253 168	0.0002
	107.01	< 1	IDH2	15	90631934	C to T	R88Q	138 180	161 371	0.4613
			RUNX1	21	36231791	T to C	D171G	368 438	50 796	0.8788

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.

(Table 1). For each sequencing library, we tagged ~2.5 million locus-specific amplicons generated from genomic DNA using high-fidelity PCR with randomly indexed custom adapters. Sequencing errors were removed to create ECCSs as described above. Each ECCS was then aligned to the reference genome for variant calling (Supplementary Figure 1).

Using conventional deep sequencing, we detected t-AML/t-MDS-specific mutations in prior banked samples at variant allele fractions between 0.03 and 0.87 (data not shown). In one individual (UPN 684949), deep sequencing alone was insufficient to distinguish known *ASXL1* and *U2AF1* mutations from the sequencing errors in samples banked 5 and 3 years before t-MDS diagnosis, respectively (Figures 1e and f). However, ECS identified the L866* nonsense mutation in *ASXL1* at a VAF of 0.004 (Figure 1g) and the S34Y missense mutation in *U2AF1* at a VAF of 0.009 (Figure 1h). In addition, ECS was able to temporally quantify these mutations from three pre-t-MDS samples banked yearly from 3 to 5 years before diagnosis (Supplementary Figures 4 and 5). In two cases (UPN643006 and UPN942008), only a subset of the variants identified at diagnosis were present in the prior banked sample (Table 1). Specifically, in the UPN643006 sample, banked 12 years before diagnosis, a single-nucleotide deletion in *ASXL1* was present at VAF 0.03. But, the G to T substitution in *ASXL1*, CTT deletion in *GATA2* and G to T substitution in *U2AF1* were not detectable in this prior banked sample.

Here we present a practical and clinically oriented application for targeted error-corrected NGS utilizing single molecule indexing. This method easily integrates into existing NGS library preparation protocols and enables the quantification of previously undetectable mutations in heterogeneous DNA samples. The only modification to the standard NGS library preparation is the replacement of the stock adapters with our randomly indexed adapters and the addition of a qPCR step before sequencing. The qPCR step limits the number of molecules sequenced, ensuring adequate coverage for each read family. With these two modifications, we achieve highly specific detection for rare mutations. The bioinformatics analysis is straightforward and does not require proprietary algorithms or tools (Supplementary Methods). Our results highlight the ability of this method to identify rare subclonal populations in a heterogeneous biological sample. As applied to t-AML/t-MDS, we show these previously undetectable mutations are present years before diagnosis and fluctuate in prevalence over time.

A clinical application of ECS is to quantify minimal residual disease (MRD). As the genomic characterization of leukemia becomes more readily available, identifying causative genetic lesions and rare therapy-resistant subclones will become increasingly useful for risk stratification, therapeutic selection and disease monitoring. Already, whole-genome sequencing of AML has demonstrated that nearly every case of AML harbors one or more somatic SNVs.[10] These SNVs are more reliable clonal markers of malignancy than cell surface markers, which can change over time. Leveraging this information, conventional NGS was implemented retrospectively to detect MRD harboring leukemia-specific insertions/deletions (indels) as rare as 0.00001 VAF in *NPM1*[11] and 0.0001 VAF in *RUNX1*.[12] This was possible because indels are only rarely generated erroneously by NGS. Unfortunately, measuring rare leukemia-associated substitutions is limited owing to the relatively high error profile of conventional NGS.[13] However, ECS can achieve the 1:10 000 limit of detection featured by conventional MRD platforms.[14] For patients whose leukemia lacks suitable markers for conventional MRD, ECS could offer an alternative with comparable sensitivity and specificity that is easy to implement in a clinical sequencing lab. Furthermore, the ability to multiplex 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 Young[1,2], TN Wong[3], AEO Hughes[1,2], SE Heath[3], TJ Ley[3], DC Link[3] and TE Druley[1,2]
[1]*Department of Pediatrics, Division of Hematology and Oncology, Washington University School of Medicine, Saint Louis, MO, USA;* [2]*Center for Genome Sciences and Systems Biology, Washington University School of Medicine, Saint Louis, MO, USA and* [3]*Department 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-yr-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.
3 Welch 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 Schmitt 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, Kinzler 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.
7 Wong 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, Mindrinos 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.

Supplementary Information accompanies this paper on the Leukemia website (http://www.nature.com/leu)