Subclonal variant calling with multiple samples and prior knowledge

Moritz Gerstung1, Elli Papaemmanuil1, and Peter J. Campbell1,2,3∗

1Cancer Genome Project, Wellcome Trust Sanger Institute, Hinxton, CB10 1SA, UK; 2Department of Haematology, Addenbrooke’s Hospital, Cambridge CB2 0QQ, UK; 3Department of Haematology, University of Cambridge, Cambridge CB22XY, UK

ABSTRACT
Motivation: Targeted resequencing of cancer genes in large cohorts of patients is important to understand the biological and clinical consequences of mutations. Cancers are often clonally heterogeneous and the detection of subclonal mutations is important from a diagnostic point of view, but presents strong statistical challenges.

Results: Here we present a novel statistical approach for calling mutations from large cohorts of deeply resequenced cancer genes. These data allow for precisely estimating local error profiles and enable detecting mutations with high sensitivity and specificity. Our probabilistic method incorporates knowledge about the distribution of variants in terms of a prior probability. We show that our algorithm has a high accuracy of calling cancer mutations and demonstrate that the detected clonal and subclonal variants have important prognostic consequences.

Availability: Code is available as part of the Bioconductor package deepSNV.
Contact: {mg14,pc8}@sanger.ac.uk

1 INTRODUCTION
In recent years genome sequencing has greatly enhanced our understanding of cancer biology (Stratton, 2011). Tumors are evolving entities and display complex clonal architectures with many mutations present in only a subset of cells (Nik-Zainal et al., 2012; Yates and Campbell, 2012). Subclonal mutations provide insights into disease evolution and influence prognosis (Landau et al., 2013; Papaemmanuil et al., 2013). Subclonal variants can be detected utilizing the deep coverage of next-generation sequencing technologies, but their distinction from sequencing errors, library preparation and alignment artifacts suffers from an unfavourable signal to noise level (Gerstung et al., 2012; Schmitt et al., 2012).

A series of powerful variant callers has been developed in recent years for calling variants from genome or exome sequencing data of tumor-normal pairs (Cibulskis et al., 2013; Larson et al., 2011; Goya et al., 2010). For detecting subclonal variants, or mutations in samples with a very low purity, which are both reported by small fractions of reads only, it is mandatory to accurately quantify the abundance of sequencing artifacts, which may otherwise lead to large numbers of false positives. With increasing numbers of genomic datasets being generated, it becomes apparent that sequencing artifacts tend to occur in a systematic way and on specific sites.

Targeted resequencing experiments, in which a selected set of candidate genes is resequenced across hundreds or thousands of samples, are increasingly prepared to evaluate findings from large-scale sequencing studies. Such datasets present an opportunity to precisely estimate the distribution of sequencing artifacts by aggregating information across samples, rather than across sites as is commonly done in tumour-normal variant calling. This will help avoid artifacts and likewise enable calling more variants on sites with lower error rates.

The growing catalogs of somatic mutations in cancer make it also possible to define genomic loci more likely to be mutated. One may therefore attempt to incorporate this prior knowledge to facilitate variant calling on mutational hotspots while remaining conservative on the remaining sites. Hence a well-chosen prior will increase sensitivity at a given level of specificity.

Here we present a novel approach for detecting clonal and subclonal variants that exploits the power of a large sample set for precisely defining the local error rates and which utilises prior information to call variants with high specificity and sensitivity.

2 APPROACH
Detecting mutations in deep sequencing data is essentially a model selection problem: One compares the probability of observing a given number of reads reporting a base change under a null model specifying the distribution of sequencing artifacts to the probability in an alternative model allowing for true variants. A mutation is called if the probability under the alternative exceeds that of the null model. A probabilistic framework offers the flexibility to account for prior information, which can be useful as some genes are more likely to be mutated in particular cancers and there often exist mutational hotspots within a gene. The approach we present here for modeling the error distribution is based on the observation that sequencing artifacts are recurrent on specific loci. In a large cohort this allows to define a background error distribution on each locus, above which true variants can be called.

© The Author(s) 2014. Published by Oxford University Press. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/3.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.
allele counts. Let the coverage be denoted by $n_{ijk}$, the count of that nucleotide in forward and backward read orientations in sample $i$, and $n_{′ijk}$, respectively. For the ease of reading we omit the indexes $i, j, k$, unless a clear distinction is necessary. We model the nucleotide counts to be distributed by a beta-binomial distribution,
\[
X \sim \text{BetaBin}(n, \mu, \rho),
\]
\[
X' \sim \text{BetaBin}(n', \mu', \rho).
\]

The parameters $\mu = \mu_{ijk}$ and $\mu' = \mu'_{ijk}$ define the expected number of nucleotide counts per read,
\[
E[X] = \mu n,
\]
\[
\text{Var}[X] = n\mu(1 - \mu)(1 + (n - 1)\rho), \quad \text{for } \rho \leq 1.\]

The dispersion factor $\rho = \rho_{ijk}$ (no sample index) defines the amount of extra variance, as compared to pure sampling errors; for $\rho = 0$, the model is the usual binomial.

Variant calling is commonly performed against a matched normal. Here we construct an aggregate control sample for sample $i$ from the set of all other samples $J(i) = \{h \neq i\}$, $X_{ijk} = \sum_{h \in J(i)} X_{hjk}$ and $X'_{ijk} = \sum_{h \in J(i)} X'_{hjk}$ instead. The latter is justified if the particular variant occurs only rarely, or if the set of reference samples $J(i)$ is chosen such that they are unlikely to contain the variant, e.g., by only selecting samples with a variant allele frequency (VAF) $X_{i}/n_i$ smaller than a predefined threshold, typically around 10%. We assume that the control counts are also beta-binomially distributed with mean $\nu = \nu_{ijk}, \nu'_{ijk}$ and coverage $n = n_{ijk} = \sum_{h \in J(i)} n_{hjk}$ and $n' = n'_{ijk} = \sum_{h \in J(i)} n'_{hjk}$:
\[
X \sim \text{BetaBin}(n, \nu, \rho),
\]
\[
X' \sim \text{BetaBin}(n', \nu', \rho).
\]

This definition is consistent with the assumption that the individual samples are beta-binomially distributed, as long as the dispersion parameter $\rho$ is small. The above parameterisation is very similar to the deepSNV algorithm (Gerstung et al., 2012), but uses aggregate control counts $X, X'$ instead of a single control sample. We find that the model realistically reflects the observed distribution of nucleotide counts (Figure 1a).

We formulate calling variants as a model selection problem. A true variant will be present on both strands, $\mu = \mu'$, and at a higher frequency than both background error rates $\nu, \nu'$, because it is the sum of the true allele frequency
and the error rate. The null-model is that \( X \) and \( X' \) are distributed with the same rate as the control counts \( X \) and \( X' \) on either strand, which we assume to contain only errors but no variants. We then have the two models

\[
M_0 : \quad \mu = \nu \quad \land \quad \mu' = \nu'; \\
M_1 : \quad \mu = \mu' > \nu, \nu'.
\] (4)

3.2 Inference

Denoting the data by \( D = \{X, X', X, X'\} \) the Bayes factor \( \Pr(D | M_0) / \Pr(D | M_1) \) can be approximated using point estimates:

\[
\frac{\Pr(D | M_0)}{\Pr(D | M_1)} \approx \frac{\Pr(D | \hat{\mu}_0, \hat{\nu}_0, \hat{\nu}') + \Pr(D | \hat{\mu}_1, \hat{\nu}, \hat{\nu}') - \Pr(D | \hat{\nu}_0, \hat{\nu}_0')}{\Pr(D | \hat{\mu}, \hat{\nu}, \hat{\nu}')}
\] (5)

The three terms in the numerator arise from the OR condition of \( M_0 \), Eq. (4), and denote the probability that either the error rates in forward, the reverse, or both orientations are identical. Hence, the third term is usually small in cases where both allele frequencies \( \mu, \mu' \) are different from the error rates. Note that this approximation is rather strong, but efficient to compute and works well in real applications. The point estimates are defined in the following way, using the method of moments:

\[
\hat{\nu} = X/n', \quad \hat{\nu}' = X'/n' \\
\hat{\nu}_0 = (X + X')/(n + n'), \quad \hat{\nu}_0' = (X' + X)/(n' + n') \\
\hat{\mu}_0 = X/n, \quad \hat{\mu}'_0 = X'/n'.
\] (6)

The symbols \( \hat{\nu}_0 \) and \( \hat{\nu}'_0 \) are the error rates across all samples; \( \hat{\mu}_0 \) and \( \hat{\mu}'_0 \) are the VAF in forward and reverse orientation for each sample (Figure 1b).

The likelihood factorizes into \( \Pr(X' | \mu' \nu' \rho) \Pr(X | \mu \nu' \rho) \Pr(X' | \mu' \nu' \rho) \Pr(X | \mu \nu' \rho) \); this allows to write the Bayes factor as

\[
\frac{\Pr(D | M_0)}{\Pr(D | M_1)} = \frac{\Pr(X | \hat{\nu}_0) \Pr(X' | \hat{\nu}'_0) \Pr(X | \hat{\mu}_0) \Pr(X' | \hat{\mu}'_0)}{\Pr(X | \hat{\mu}) \Pr(X' | \hat{\mu}) \Pr(X | \hat{\nu}) \Pr(X' | \hat{\nu}')}
\] (7)

The value of the Bayes factor \( \frac{\Pr(D | M_0)}{\Pr(D | M_1)} \) as a function of \( X, X' \) is illustrated in Figure 1c-f for different error rates. For a small error rate of \( \nu = \nu' = 10^{-4} \), which is found on the majority of sites, only a few variant alleles lead to a Bayes factor small enough to call a variant.

3.3 Estimating \( \rho \)

There exists no closed-form solution to estimate \( \rho \), but it can be estimated from the variances of the VAF \( \hat{\rho} = \hat{\mu}_j \hat{\nu}_j \) and total coverage \( m_{ij} = n_{ij} + n_{ij}' \) across samples \( i \) by the method-of-moments estimator \( \hat{\rho} = \hat{\rho}_j \); 

\[
\hat{\rho}_j = \frac{N s^2}{N - \sum_{i=1}^N 1/m_{ij}}, \quad \text{with} \quad s^2 = \frac{N \sum_{i=1}^N \sum_{j=1}^N (\hat{\nu}_{0,ijk} - \hat{\rho}_j)^2}{(N-1) \sum_{i=1}^N m_{ij}},
\] (8)

As this estimator is not guaranteed to yield values in \((0, 1)\), we found it to \([10^{-6}, 0.1]\). Empirically, we found that \( \hat{\rho} \) is usually small (Figure 1b).

3.4 Prior data

The posterior probability that \( M_0 \) is true can be computed by Bayes’ formula:

\[
\Pr(M_0 | D) = \frac{(1 - \pi) \Pr(D | M_0) + \pi \Pr(D | M_1)}{1 + \pi \Pr(D | M_1)},
\] (9)

We use the probability of the null model \( M_0 \) because of its similarity to a \( P \)-value in a hypothesis testing scheme and call variants below a certain threshold \( \Pr(M_0 | D) < \pi \). The parameter \( \pi = \pi_{jk} \) denotes the prior probabilities that a variant \( k \) exists at position \( j \). The prior \( \pi \) essentially shifts the relation between the Bayes factor \( \Pr(D | M_0) / \Pr(D | M_1) \) and the posterior probability \( \Pr(M_0 | D) \). A higher prior probability results in a lower posterior probability of an artifact for a given signal as quantified by the Bayes factor (Figure 1g).

Prior information about the likelihood of an allele being mutated can be extracted for example from the COSMIC database (Forbes et al., 2011). We assume that the prior can be written as

\[
\pi_{jk} = \pi_{\text{gene}} \times \pi_{jk},
\] (10)

where the histogram

\[
\hat{\pi}_{jk} = \frac{\# \text{ mutations k at locus j}}{\# \text{ total mutations in gene}}
\] (11)

denotes the relative frequency of mutations \( k \) at site \( j \) in a given gene. The factor \( \pi_{\text{gene}} \) defines the probability of a gene being mutated. These probabilities vary greatly between genes and for the same gene also between different tumour types. As there are currently many systematic studies being performed, we expect that accurate estimates will be available soon for many cancers. For all sites not present in COSMIC, we use a constant value of \( \pi = 10^{-5} \). An example of the prior distribution obtained from COSMIC is shown in Figure 1b.

3.5 Implementation

We have implemented the algorithm in the statistical language R (R Core Team, 2012) and released code as part of the deepSNV Bioconductor package (>1.8) (Gerstung et al., 2012). We named the algorithm “shearwater” after the seabirds that fly long distances over the ocean, watching the water closely and eventually dive into the water to pick up prey, often with prior help from other fish. More information can be found in the accompanying vignette:

> library(deepSNV)

> vignette("shearwater")

The runtime of 1kb over 800 samples is approximately 1 CPU minute on a 2.2GHz AMD processor. This performance is sufficient to process a completely targeted screen with one hundred genes in a few hours on an 8-core machine and the algorithm can be parallelised easily.

4 RESULTS

We benchmark our algorithm against data from two large gene screens in hematological cancers, a subset of 738 patients with Myelodysplastic syndromes (MDS) we have published recently (Papaemmanuil et al., 2013). In these screens 111 cancer genes were sequenced using barcoded libraries prepared from whole genome amplified DNA. Samples were sequenced in batches of 96 per lane on a HiSeq2000 and reads were aligned with bwa (0.5.9-r16+rugo) (Li and Durbin, 2010) to the GRCh37 human reference genome. Technical replicates existed for 20 samples with acute myeloid leukaemia (AML) assayed by the same gene panel. Moreover we included 32 normal samples to quantify specificity. Here we focus on a subset of 43 genes with good coverage and in which we had previously found oncogenic mutations (Papaemmanuil et al., 2013; Table 1). The availability of survival data in the MDS cohort allows for evaluating the quality of variant calls by their prognostic potential, which is an orthogonal measure to technical replication.
4.1 Simulations and control data

To assess the sensitivity and specificity of shearwater we used a panel of 500 samples, including 32 normals and 2 × 20 AML replicates. The remaining samples served for defining the background error distribution and for assessing how reproducible the calls are. To analyse the sensitivity for different combinations of coverage, we simulated mutations at different variant allele frequencies using the coverage and strand bias of one of the normal samples (median 128x, 5% 13x, 95% 372x coverage). For each position j we drew a vector of variant allele frequencies for \{A, T, C, G, –\} from a Dirichlet distribution, \(\mu_j \sim \text{Dir}(1, 1, 1, 1)\). We then sampled reads \(X_j \sim \text{Mult}(\mu_j, n_j)\), \(X'_j \sim \text{Mult}(\mu_j, n'_j)\), where \(n_j\) and \(n'_j\) are the coverages on forward and reverse strand as observed in the normal sample. We ran shearwater on the cohort of 500 samples to compute the Bayes factors of each simulated variant.

**Sensitivity** The fraction of variants with a Bayes factor smaller than \(10^{-4}\) for fixed dispersion \(\rho\) is shown in Figure 2a. This cutoff corresponds to a posterior odds of 1, or a cutoff of \(P_0 = 0.5\), under a uniform prior with probability \(10^{-4}\). For a coverage of 250x, the true positive rate of a 5% variant is 70%, and that of a 10% variant is approximately 85%. Variants present in 20% can be called almost with certainty. When the dispersion is estimated from the data using all samples with VAF below 10%, the Bayes factors become larger with certainty. When the dispersion is estimated from the data using all samples with VAF below 10%, the Bayes factors become larger with certainty.

**AUC and cohort size** We evaluated the area under the ROC curve (AUC) as a global measure of predictive accuracy for different VAF frequencies as a function of cohort size (Figure 2c). Typical AUC values range from 60% for 1% variants to 98% for 50% VAF with only a mild influence of the cohort size. A small percentage of variants could not be called with the experimentally observed coverage.

**Specificity** As variants in cancer samples are typically rare and millions of loci are analysed, specificity is a major concern. We compared shearwater’s specificity on 32 normal samples against three other algorithms: Caveman, an established variant caller, which has been used in many large-scale genome and exome sequencing projects (Stephens et al., 2012; Nik-Zainal et al., 2012; Jones et al., 2013), MuTect (Cibulskis et al., 2013), and deepSNV (Gerstung et al., 2012). We ran Caveman as described against a single unmatched normal sample (Papaemmanuil et al., 2013). Similarly, we ran MuTect (v.1.1.4) with default options --cosmic b37_cosmic_v54_l20711.vcf and --dbsnp dbsnp_132_b37.leftAligned.vcf.gz against the same unmatched normal. The options of deepSNV (v.1.3.3) were combine.method='fisher' and adjust.method='BH'. After calling variants, we filtered the output by removing variants in Ensemble variation (v70) and removed unknown polymorphisms with P(VAF > 0.5) > 0.1.

In total shearwater called 5 non-polymorphic variants (Figure 2d), deepSNV, in contrast called of 32 variants, Caveman 48 and MuTect 79 variants. Hence the specificity of shearwater appears very satisfying.
whereas the first replicate did not. The Bayes factors of replicates here the second replicate underwent whole genome amplification, evaluated 20 AML samples which had been sequenced in replicates. 

To quantify the reproducibility of shearwater we evaluated 20 AML samples, of which 2,363 were unique variants (identical alleles present in multiple samples). Of these unique variants 757 were found either in Ensembl variation or in an in-house panel of 500 normal exomes. 200 variants were present in COSMIC, but not in Ensembl variation and 1,406 were new (Figure 3c).

The distribution of variant allele frequencies of known polymorphisms has two narrow peaks at 0.5 and 1, confirming the accuracy of allele frequency estimates (Figure 3d). Non-polymorphic calls have a broad distribution with typical frequencies ranging from 0 to 0.5, with slightly more mass towards lower frequencies. This is consistent with the expectation that more variability exists at lower frequencies. The distributions of COSMIC and new variants are very similar, which gives us confidence that these are indeed real. This also indicates that the prior did not lead to overcalling, which would occur specifically at low frequencies.

Comparison to other variant callers We ran Caveman, MuTect and deepSNV against an unmatched normal as described above. After filtering variants from Ensembl variation 576 variants were called by all four approaches (Figure 3e). 1,256 variants were unique to shearwater, compared to 360 for Caveman, 1,381 for deepSNV and 855 for MuTect. 405/1,256 unique variants were single base deletions, which could not be called by Caveman or MuTect. It therefore appears that shearwater achieves a good level of specificity, given that Caveman used a series of post-processing filters, whereas deepSNV and MuTect did not.

In the presence of noise variant calling amounts to balance sensitivity and specificity. We evaluated this tradeoff by comparing the ability for calling the SF3B1 K700E hotspot mutation, which is very characteristic of MDS and can thus be considered true somatic, versus the overall number of false positive calls in the normal panel as discussed in the previous section. All three variant callers detect 97 K700E variants; above this level, however, MuTect and Caveman begin to call many artifacts (Figure 3f). Shearwater calls a total of 108 variants without decreasing specificity, because of the higher prior weight \( \pi = 0.05 \) put on this variant. Yet no K700E variants were found in the normal samples, showing that shearwater does not blindly call this hotspot.
4.3 Prognostic performance

In the absence of a known ground truth and reliable methods for validating subclonal mutations that are guaranteed not to replicate systematic artifacts it is generally difficult assess the quality of one variant caller over another (Kim and Speed, 2013). An indirect measure of the quality of a predicted genotype can be the correlation with a known phenotype, such as survival. Suppose there exists a correlation \( C \) between genotype \( G \) and a quantitative trait \( Y \). In practice we don’t know the genotype with certainty, and only have estimates \( \hat{G} = G + \epsilon \), where \( \epsilon \) is the deviation of the estimate from the truth. If \( \epsilon \) is zero the observed correlation between genotype and phenotype is \( C = C \); if \( \epsilon \) is large and completely randomises \( \hat{G} \) then the observed correlation becomes zero. Conversely, a higher correlation between genotype estimates and phenotype indicates a lower average bias of the genotype estimates. This reasoning requires the error \( \epsilon \) and the phenotype \( Y \) to be uncorrelated and it appears unlikely to us that the ability to call mutations is confounded with the outcome of the patient in such a way that it leads over- and under-calling of mutations in specific sets of genes.

Marginal effects of single genes
Survival in MDS depends on the absence and presence of mutations in multiple genes. For example, we and others have shown previously that oncogenic mutations in the SF3B1 gene are associated with better prognosis (Papaemmanuil et al., 2011; Damm et al., 2012; Malcovati et al., 2011), while alterations in TP53, DNMT3A, STAG2 and other genes are indicative of a worse outcome (Papaemmanuil et al., 2013). Patients with any novel mutations should hence follow these survival trends.

Survival data was available for 517 patients. We considered a gene to be mutated if it contained at least one non-silent mutation; the endpoint was AML-free survival. Figure 4a-d shows Kaplan-Meyer curves for patients carrying mutations identified by Caveman and/or shearwater. Patients with mutations only detected by shearwater generally display the expected behaviour—that is on average better survival if SF3B1 was mutated, very poor survival if TP53 or STAG2 were mutated and a moderate change for DNMT3A.

Overall prognostic accuracy
To assess the overall prognostic power combining all mutated genes we trained Cox proportional hazards survival models with mutated genes as covariates. We used a five-fold cross validation scheme to estimate Harrel’s C-statistic (Harrell et al., 1996), measuring the correspondence of the estimated risk and the ordering of deaths, similar to an AUC statistic, on the remaining fifth. The predictive potential \( C \) increases with \( P_0 \) and the prior odds, with typical values between 0.67 and 0.68 (Figure 4e). For a prior weight of 100 and \( P_0 = 0.5 \), \( C \) starts dropping again as shearwater starts over-calling variants with a high prior probability. The maximal value of \( C = 0.682 \) was observed for a cutoff of \( P_0 = 0.5 \) and a prior odds of 1, justifying our previous parameter choices.

The C-statistics of shearwater’s competitors were slightly lower, with Caveman having \( C = 0.666 \) and permutations of the discrepant calls show that this difference is unlikely to be an artefact.
\( P = 0.03; \) Figure 4f). For Mutect we obtained \( C = 0.666 \) \( (P = 0.02) \) and deepSNV \( = 0.583 \) \( (P < 0.01) \). The higher prognostic accuracy of shearwater suggests that shearwater calls more survival-associated variants and less noise.

In a practical application one will most likely rely on a combination of variant callers to avoid the biases of a single method. Combining the genotypes of different methods by either the intersection (AND) or the union (OR) of variant calls, however, did not further increase \( C \) (Figure 4g-h). This indicates that the variants that shearwater may be missing do not have a large influence on survival.

5 DISCUSSION

In this paper we presented a statistical approach for detecting clonal and subclonal single nucleotide variants in targeted gene screens. The availability of large numbers of samples allows for precisely estimating the rate of artifacts, which is important for reliably detecting subclonal mutations that can have a very disadvantageous signal to noise level. Our model incorporates prior information on mutational hotspots, which selectively increases the sensitivity for known mutations. Shearwater automatically determines the noise levels from the data and we therefore expect it to deal well with sequencing data from other sequencing platforms and aligners.

Shearwater has both a very high specificity and good power to detect variants. The genotypes obtained by shearwater have a higher prognostic value than those from established variant callers, and are likely to contain fewer artifacts. To an extent this behaviour is expected, because of our algorithm’s ability to exploit the power of a large cohort of samples and to incorporate prior knowledge about which mutations are more likely than others.

As our algorithm uses unmatched samples, it relies on the quality of polymorphism databases such as dbSNP or Ensembl variation, which can generally be expected to become better in the future. The same holds true for the quality of databases of somatic mutations that will get richer over time and contain more precise information about the mutational patterns in each cancer type. Here we used the same probability for each gene to be mutated, but once unbiased estimates for the mutation frequencies in each cancer type exist from systematic gene screens, one will be able to further improve the accuracy of our algorithm. The idea of using a prior for recurrent mutations may also be incorporated easily into other variant callers.

Finally, our core algorithm may also be improved in many ways. For example, one could account for base qualities by a weighted counting scheme, instead of a simple phred quality threshold. One limitation of our approach is its reliance on a variant to be present on reads from both directions due to the specifics of the null model \( M_0 \). This was introduced as it greatly increases the specificity of calls, but leads to a decrease in power in regions with low coverage and also at the flanks of the target regions, where often reads in only one direction are available. Our implementation allows the user to choose an essentially strand-agnostic null model \( M_0 : \mu = \nu \land \mu' = \nu' \), but this may be less specific. To analyse matched samples one could derive the joint probability of a variant being present in only the tumour but not the normal, or simply remove the intersection of variants in tumour and matched normal.

In summary, we have presented a coherent statistical methodology and robust algorithm for calling subclonal variants in cancer samples with great specificity. As genomic sequencing is about to enter clinical diagnostics, we believe that our method will have broad applicability.

ACKNOWLEDGEMENT

Funding: This work was supported by a Specialized Center of Research grant from the Leukemia Lymphoma Society (LLS), the Kay Kendall Leukaemia Fund and the Wellcome Trust (grant reference 077012/Z/05/Z). PIC is personally funded through a Wellcome Trust Senior Clinical Research Fellowship (grant reference WT088340MA).

Data: Data are available under EGA accession EGAD00001000283.

REFERENCES

Larson DE, Harris CC, Chen K, Koboldt DC, Abbott TE, et al. (2011) SomaticSniper: Identification of Somatic Point Mutations in Whole Genome Sequencing Data. Bioinformatics