Clinical application of tumour in normal contamination assessment from whole genome sequencing

Jonathan Mitchell (Genomics England, UK),
Jack Bartram (Department of Haematology, Great Ormond Street Hospital for Children, UK)
Susan Walker (Genomics England, UK)
Jane Chalker (SIHMDS-AG, Great Ormond Street Hospital for Children, UK)
Magdalena Zarowiecki (Genomics England, UK)
Salvatore Milite (University of Trieste)
Genomics England Research Consortium (Genomics England, UK)
Alona Sosinsky (Genomics England, UK)
Giulio Caravagna (University of Trieste, Italy & The Institute of Cancer Research, UK)

Corresponding: (AS) alona.sosinsky@genomicsengland.co.uk; (GC) gcaravagna@units.it.

Abstract. Bioinformatics pipelines designed for paired analysis of tumour and matched normal samples use germline variants detected from normal tissue samples to identify *bona fide* tumour somatic mutations. Unexpected contamination of normal samples with tumour cells can lead to the subtraction of genuine somatic variants, reducing variant detection sensitivity and compromising downstream analyses. Leveraging the power of whole-genome sequencing data available from the 100,000 Genomes Project at Genomics England, we developed the Tumour in Normal Contamination (TINC) tool for TIN contamination assessment. TINC was validated *in silico* using patient-derived whole-genome samples ($R^2 = 0.95$; $p < 2.2 \times 10^{-16}$ and $R^2 = 0.85$; $p < 3.3 \times 10^{-13}$ for expected versus observed contamination in two tumour types), and against orthogonal minimal residual disease (MRD) testing with 70 leukaemia samples. From a systematic review of whole-genome data available for 771 patients with haematological malignancies and sarcomas, we found TIN contamination (>1%) across a range of cancer clinical indications and DNA sources. The highest prevalence was among saliva samples from acute myeloid leukaemia (AML) patients (43/114 samples, 38%), and sorted CD3+ T cells from myeloproliferative neoplasms (MPN) patients (22/24 samples, 91%). Further exploration of genomic data for AML and MPN patients has revealed 108 hotspot mutations in genes with high prevalence of somatic alterations in haematological cancers (*JAK2, FLT3, DNMT3A, TP53, KIT, NRAS* and *IDH2*), 27 of which supported by at least 5% of reads in the normal sample and therefore at risk of being subtracted by standard pipeline for somatic variant calling. Our work highlights the importance of contamination assessment for accurate detection of somatic variants in

both research and clinical settings. TINC allows accurate and highly sensitive quantification of contamination levels, giving confidence in the validity of the whole genome assessment of the tumour. We propose this to become an essential metric in clinic practice alongside other routine quality control measures, especially as large-scale pan-cancer sequencing projects are adopted to deliver accurate data from which to make clinical decisions for patient care.

Introduction

In popular experimental designs for cancer bulk DNA sequencing - whole-exome sequencing (WES) or whole-genome sequencing (WGS) - individual tumour samples are matched with a reference “normal” sample from the same patient, usually obtained from peripheral blood, saliva, or a skin biopsy (Mwenifumbo and Marra 2013; Gagan and Van Allen 2015). Analysis is performed by first detecting variants with respect to the human reference genome in both normal and tumour samples, followed by subtraction of the patient-specific variants from the tumour to select variants that are private to the tumour sample (Barnell et al. 2019). However, this experimental design is only effective if the matched normal sample is free of contaminating tumour cells, an assumption not often emphasised (Figure 1a,b) (Kim et al. 2018; ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium 2020). When DNA derived from the normal sample is contaminated by tumour DNA, standard bioinformatics pipelines can mistakenly subtract genuine somatic mutations from the set of mutations identified in the tumour sample due to evidence for a mutation being present in the normal sample, resulting in a reduction in sensitivity for true somatic mutations (i.e. higher number of false negatives). Based on the clonal evolution model (Nowell 1976), the probability of a false negative increases with the fraction of cells harbouring the mutation (due to a higher likelihood that the variant will be present in the sequencing data for the normal sample), resulting in a bias towards erroneous subtraction of somatic variants at high allele frequency in the tumour. High allele frequency variants are of the greatest importance because they triggered tumour formation and determined the subsequent clonal evolution patterns (Levine, Jenkins, and Copeland 2019).

More recently, in order to adapt the canonical bioinformatics somatic analysis approach to accommodate tumour in normal (TIN) contamination, a model for normal contamination was introduced into somatic variant calling algorithms (Kim et al. 2018; Nangali 2021). However, these tools require an estimate of the TINC level as an input parameter, and only support low levels of contamination. Computationally, only one tool is available to assess tumour contamination of normal samples (Taylor-Weiner et al. 2018). At high levels of TIN, however, the only alternative to a canonical

tumour-matched-normal design is currently a tumour-only pipeline, using population germline frequency databases to filter likely germline variants from the set of putative somatic mutations. This approach attempts to minimise the rate of false negatives, a fundamental requirement for clinical reporting where the failure to identify actionable somatic mutations can be detrimental to patient care, but leaves many ultra-rare germline variants variants of indeterminate origin

Here we present TINC, a computational method to assess the level of tumour in normal contamination applicable for tumour and matched normal pairs, leveraging on state of the art tools for measuring tumour clonal evolution from WGS data. We have demonstrated the performance of TINC with simulated data and by comparison with orthogonal minimal residual disease (MRD) test data for 70 leukaemia patients. By applying TINC to sequencing data from 771 participants in the Genomics England 100,000 Genome Project (Turnbull et al. 2018), including 617 patients with haematological malignancies, we detected tumour contamination in normal samples from different germline sources across several tumour types.

Assessment of sample quality is essential in clinical reporting where treatment decisions are based on the analysis of genomic data. The level of contamination predicted by TINC provides reassurance for clinicians and clinical scientists for the accuracy of variant detection in the somatic analysis, and can highlight the risk that clinically relevant variants may not have been reliably detected. Therefore, we propose that TIN is an essential metric for clinical analysis and reporting of WGS data.

Results

The TINC method

We have developed TINC, a new approach for tumour in normal contamination assessment, leveraging the concept of tumour clonal evolution. It is freely available as an open-source R package (Data Availability). TINC uses the variant allele frequencies (VAFs) of somatic single nucleotide variants (SNVs) detected in tumour and normal samples to identify clonal somatic mutations (i.e. those detected in all tumour cells sampled) and evaluate their level in the normal sample. From the observed VAFs, TINC determines scores for the percentage of tumour cells in the tumour (usually referred to as tumour purity) and in the normal sample, which are called Tumour in Tumour (TIT) and Tumour in Normal (TIN) scores respectively. These scores can be expressed as units of read fractions (i.e., percentage of reads in the sequencing data for the normal sample that are derived from the tumour).

TINC first identifies high confidence clonal somatic mutations in the tumour sample (grey, yellow and red lineages in Figure 1c-e), which are then used to estimate tumour purity (TIT score). Support for these clonal variants is then assessed in the normal sample to determine the level of tumour in normal contamination (TIN score, Figure 1e). Specifically, TINC targets the ancestor of the tumour cells found in both the tumour and normal samples, i.e., the most recent common ancestor (MRCA) of the sequenced cells (yellow and grey lineages in Figure 1c-e) which can not be directly sampled or sequenced. Due to sampling differences, private mutations are anticipated in the different tumour cell lineages present in the tumour and the normal samples (red and blue lineages in Figure 1c-e). Note that clonal somatic mutations can be mistakenly labelled as private tumour mutations due to the difference in sequencing read depth between tumour (about 100x) and normal (about 30x), as well as sequencing noise that obscure support for clonal somatic mutations in normal samples. Therefore, assessment of clonal mutations in the normal sample is impacted by the tumour architecture and data quality at variant sites. Building from the expected tumour architecture, we are able to model the anticipated data distribution in both samples for somatic variants of the various cell lineages (Figure 1e).

Using TINC, clonal mutations are identified with the MOBSTER machine learning model for subclonal deconvolution from WGS (Caravagna, Heide, et al. 2020). MOBSTER integrates population genetics and machine learning to cluster somatic variants based on their VAF, decoupling clones that undergo positive selection from neutral mutations. Read counts for high confidence clonal somatic mutations identified by MOBSTER are then fitted to a Binomial mixture in the matched normal sample. From the analysis of the tumour and normal samples, TINC obtains the information to compute TIT and TIN scores (Online Methods).

TINC can also utilise allele-specific tumour Copy Number Alteration (CNA) calls to retain SNVs in a subset of genomic intervals with the copy number state (e.g. heterozygous diploid or tetraploid) spanning the largest proportion of the tumour genome. By incorporating CNAs in the logic, TINC normalises the observed VAFs in the tumour sample for chromosome copy number and therefore is resilient to confounding effects of tumour aneuploidy. Incorporation of copy number data is only performed for tumours for which the most extensive copy number state is one of 1:0 (loss of heterozygosity, LOH), 1:1 (heterozygous diploid), 2:0 (copy-neutral LOH), 2:1 (triploid) or 2:2 (tetraploid genome-doubled), representing the majority of copy number states observed for cancer genomes (Househam et al. 2021). For such cases, only SNVs residing within regions of the most prevalent copy number state are used in TIN

estimation. The SNV-only analysis (i.e. not incorporating CNA data) requires that the tumour genome does not harbour a high number of CNAs as otherwise the value of VAF is cofounded by copy number variations (Househam et al. 2021).

In silico validation of TINC performance

The performance of TINC was assessed using WGS data generated by Genomics England for participants recruited as a part of 100,000 Genomes Project (Turnbull et al. 2018). Synthetic test data were generated by artificially contaminating normal BAM files with sequencing reads sampled from the corresponding tumour BAM file (Methods). Variant calling of the synthetic samples was performed using the Genomics England bioinformatics pipeline and the generated SNV and CNA calls analysed with the TINC package (Figure 2a).

Synthetic TINC test data were generated using high quality WGS data derived from seven patients diagnosed with haematological malignancies (acute lymphoid leukaemia (ALL), acute myeloid leukaemia (AML) or multiple myeloma (MM)) and five patients diagnosed with lung cancer (adenocarcinoma or squamous cell carcinoma), for whom the normal sample sequenced was not affected by tumour contamination. In total, artificially contaminated WGS data were generated for thirty-nine haematological tumour-normal pairs and thirty lung cancer tumour-normal pairs, with a range of TIN contamination from 0% to 25%.

TINC successfully estimated the correct level of contamination for the majority of synthetic WGS samples, in both haematological and lung cancers (Figure 2b,c, $R^2 = 0.95;\ p < 2.2 \times 10^{-16}$ and $R^2 = 0.85;\ p < 3.3 \times 10^{-13}$, correspondingly). These data demonstrated the benefit of incorporating CNAs, particularly for cancers with high chromosomal instability, where failing to account for the presence of CNAs can result in overestimation of TIN contamination (Supplementary Figure S1a,b). For a number of cases, the level of contamination was underestimated at higher levels of TIN contamination. This decrease in performance was attributed to the impact of high levels of TIN contamination on somatic variant detection. At high levels of TIN, the somatic variants with the highest VAF in the normal sample are likely to be subtracted from the somatic variant set. This results in an underestimate of the TIN score since the variant set used for the estimation of TIN is biassed towards variants with lower VAF in the normal sample. This effect is illustrated in Figure 2b and 2c by the gradient colour representing the decreasing fraction of clonal mutations used in TIN estimation for increasing levels of TIN. This limitation is unlikely to impact clinical reporting in practice, as the effect is only significant for samples with a level of TIN far higher than would be

considered acceptable for clinical-grade analyses. At extremely high levels of TIN the majority of clonal somatic variants are subtracted during the somatic variant calling, leaving only subclonal variants which can lead to a strong underestimation of TIT. However, implementation of appropriate quality control thresholds can be used to prevent this scenario resulting in high TIN contamination going undetected

To assess further the performance of TINC, we compared outcomes with the only alternative method available to assess tumour contamination of normal samples, DeTiN (Taylor-Weiner et al. 2018). DeTiN was found to perform well with WGS data, generating results similar to TINC (R^2 scores are exceeding 0.9 for both cohorts, note that R^2score is calculated for TIN/TIT ratio) (Figure 2d,e and Supplementary Figure S1). However, TINC is more flexible as it can process copy number segments regardless of which CNA caller has been used. In contrast DeTiN requires mean allele fraction of minor parental allele at each segment and two centred segment copy ratio data. In addition, TINC generates absolute TIN and TIT values (as opposed to the TIN/TIT ratio in DeTIN) which are directly compatible with orthogonal minimal residual disease (MRD) tests used in clinical practice.

Experimental validation of TINC performance

To further validate TINC, contamination estimates for a cohort of participants in the 100,000 Genomes Project with haematological cancers (see Contamination analysis of clinical samples) were compared either with molecular Minimal Residual Disease (MRD) test using real-time PCR assessment of rearranged immunoglobulin/T-cell receptor genes or flow cytometry test for leukaemia-associated immunophenotype (Figure 3). For 53 ALL and 10 AML samples assessed (Figure 3), the estimates from TINC and MRD were consistent, with the same four of 63 samples considered as being contaminated (i.e. TIN contamination >1%). In order to extend experimental validation to additional samples with TIN, we included in our validation cohort seven additional ALL samples not meeting the 100,000 Genomes Project sample collection criteria (Figure 3). Consistent with observations from the *in silico* validation experiment, the TIN value was underestimated when compared with experimental data but still very significant for samples with high TIN.

TINC implementation in a high throughput bioinformatics pipeline

TINC has been implemented as an essential quality control step for contamination assessment in the high-throughput bioinformatics pipeline at Genomics England (Figure 4a). Using variant call format (VCF) files generated in the somatic SNV and CNA

detection components of the analysis pipeline, TINC classifies normal samples as “PASS”, “FAIL” or “Cannot estimate TINC reliably”, based on the level of TIN and TIT detected. For all analysed samples, TIT and TIN scores are presented as read-fractions (RF), i.e. the fraction of reads in a given sample originating from the tumour. Samples with TIN >1% are classified as “FAIL” indicating an alternative analysis not reliant on the matched normal sample should be conducted (for example, using an unmatched normal sample from another individual). As deconvolution of the VAF distribution is typically unreliable for tumour samples with low tumour purity, samples with a TIT score estimated at <25% are considered as not eligible for TIN estimation (cluster of samples with clonal/total mutations ratio = 1 on Figure 4b) and classified as “Can’t estimate TIN reliably”. This cutoff has been estimated using a cohort of 617 whole-genome samples (see Contamination analysis of clinical samples), and accounts for i) samples with genuinely low tumour purity and for ii) samples for which TIT is underestimated due to high TIN (i.e., extreme high TIN causing catastrophic subtraction of clonal somatic SNVs).

Contamination analysis of clinical samples

To assess the clinical impact of implementing TINC in a bioinformatics pipeline, TINC was used to analyse 771 tumour-normal pairs from participants in the 100,000 Genomes Project with either haematological cancer (n=617) or sarcoma (n=154). All samples were re-analysed through the Genomics England Pipeline 2.0 (Methods). Since TIN contamination is not expected to be a frequent occurrence for sarcoma samples, these samples were included as a control group for comparison. Normal DNA used for WGS was derived from blood, cultured fibroblast, saliva or skin biopsy samples. As previously described, TIN scores were determined only for the samples with TIT score above 25%.

The haematological malignancy samples covered a wide range of the most common diagnoses: AML (n=168), chronic lymphocytic leukaemia (CLL, n=158), MM (n=87), ALL (n=90), myeloproliferative neoplasm (MPN, n=58), chronic myeloid leukaemia (CML, n=57), diffuse large B-cell lymphoma (DLCBL, n=23), low and moderate grade non-Hodgkin B-cell lymphoma (Low/mid grade NHL, n=20) and high-risk myelodysplastic syndrome (High-risk MDS, n=20). Normal sample collection was performed according to criteria established for the 100,000 Genomes Project (Supplementary Table 1).

Of the 771 cases assessed, CNA and SNV data were incorporated for 758 cases, with TIN estimated using SNV data only (due to sample ploidy not matching TINC criteria)

for 13 cases. The proportion of cases for which TIN contamination was observed (tumour read fraction in normal sample >1%) varied across the cancer subtypes (Figure 5a,c). As expected, no normal samples derived from the cultured fibroblasts had TIN >1%. Notably, the two cancer subtypes with the highest fraction of contaminated samples were MPN and AML. 22 of 24 (91%) of sorted CD3+ T cell-derived normal samples for MPN cases and 43 of 114 (38%) saliva-derived normal samples for AML cases were found to be contaminated, consistently with previous reports (Garcia-Gisbert et al. 2020; Tenedini et al. 2014; Felix and Lukens 1986). It's worth to note that saliva samples were accepted as a normal sample in myeloid malignancies only if sufficient treatment has been given to remove all circulating myeloid cells from the peripheral blood, e.g. after administration of anthracycline chemotherapy in patients receiving intensive induction in AML.

In contrast, among the sarcoma samples, only 2% were found to be contaminated (4 out of 154, 3 out 4 with TINC below 2%). This can be explained by the fact that two of the contaminated normal samples for sarcoma patients were derived from fresh frozen tissue, which can present a higher risk for contamination than blood in patients with solid cancers. In an additional group of ALL cases for whom the sample collection procedures did not meet the criteria specified for the 100,000 Genomes Project, 11 of 46 (24%) showed contamination. A full description of the samples tested and the TINC results are provided in Supplementary Figures S3-5.

To demonstrate the abundance of essential somatic variants that could be subtracted in a tumour-normal pipeline, we computed read support for hotspot substitutions in tumour and normal genomes in the AML and MPN cohorts. We present results for genes with highest prevalence of somatic mutations in haematopoietic and lymphoid tissue in COSMIC (Tate et al. 2019): *JAK2, FLT3, DNMT3A, TP53, KIT, NRAS* and *IDH2*. For these genes, we considered hotspots supported by at least 100 COSMIC samples. Out of 108 high-confidence hotspot mutations with VAF in tumour >5%, we identified 51 mutations with normal VAF >1% and 27 mutations with normal VAF >5% (Supplementary Figure S6). Among those, we found hotspot mutations with normal VAF >10% in 6 AML cases with *DNMT3A p.R882X*, 5 with *IDH2 p.R140Q*, 3 with *JAK2 p.V617F and* 1 with *TP53 p.R273X, IDH2 p.R172K* and *NRAS p.Q61X*, and in 4 MPN cases with JAK2 p.V617F. Notably, all these variants were identified correctly upon running these samples through the tumour-only pipeline following TINC test analysis.

The availability of WGS allows us to examine the extent of contamination for all tumour clonal mutations, going beyond hotspot ones (which are usually clonal). To this extent, in FIgure 6 and Supplementary Figures S7-S8 we report two example cases of AML

patients from the 100,000 Genomes Project with tumour in normal contamination. In the first case, 982 diploid SNVs are analysed by the TINC test (see Figure 6a-c for VAF distribution in tumour and normal samples), of which 378 are identified as clonal mutations by deconvolution analysis (teal dots on Figure 6e). Deconvolution analysis identified two clusters of subclonal mutations in addition to the clonal cluster (green and blue peaks in Figure 6d, top). The allele fraction in the normal sample for the clonal variants peaks at ~8% (Figure 6d, bottom), indicating that approximately 16% of cells sampled in the normal sample are of tumour origin (Figure 6f). In this case, the analysis performed during the 100,000 Genomes Project (Genomics England pipeline 1.0, Online Methods) failed to detect a driver somatic hotspot variant in the *JAK2* gene (ENST00000381652:p.V617F) due to TIN contamination (variant supported by 50 out of 110 paired reads in the tumour, and 4 out of 34 in the normal; Figure 6a). Additionally, a frameshift deletion in *TP53* (ENST00000269305:c.594delA) was observed at 28% VAF in the normal sample raising uncertainty as to whether this variant is germline or somatic (variant supported by 41 out of 90 paired reads in the tumour, and 6 out of 21 in the normal), a classification that is important for clinical reporting. This variant has a higher VAF in the normal sample than all other somatic variants (Figure 6a) and one can hypothesise that it is a marker of clonal hematopoiesis of indeterminate potential (CHIP). In the second case, 358 clonal SNVs are used to estimate TIN contamination, again at approximately 16% tumour cells in the normal sample (Figure 6g-i). Analogous to somatic small variant calling, somatic structural variant calling also requires the subtraction of germline variants and therefore TIN can result in false negatives. In this case, a diagnostic somatic *PML-RARA* fusion was not detected due to TIN contamination (variant supported by 25 out of 150 paired reads in the tumour, and 6 out of 52 in the normal), which would affect patient diagnostic classification (Arber et al. 2016). Further examples of the determination of TINC status using tumour-normal pairs from participants with MPNs are shown in Supplementary Figures S9-S10.

Discussion

Typical tumour-normal analyses utilise a normal sample to identify the patient's germline variants, which are then subtracted from the variants identified in the tumour to define the tumour-specific somatic mutations. This approach is successful when high quality normal samples are available, but somatic variant detection is significantly impacted by poor quality normal samples, particularly those in which there is contaminating tumour DNA. The incidence of contaminated normal samples has been understudied, thus some tumour type and normal source combinations may have an unforeseen prevalence of TIN contamination, making them unsuitable for a canonical tumour-normal analysis.

Our novel computational technology to assess TIN contamination exploits tumour evolutionary principles to quantify the proportion of contaminating tumour cells in a normal sample, using WGS data for paired tumour and normal samples. The score generated by TINC is simple to interpret and can resist confounding factors such as tumour aneuploidy and tumour sample purity. Furthermore, it can be computed automatically by integrating different data types, regardless of the variant calling methodologies used. The TIN score therefore allows an informed decision-making process to either proceed with variant interpretation and reporting, or whether alternative variant calling procedures are required, such as tumour only analysis.

By applying our new method to WGS data from a large cohort of tumour-normal pairs, we have performed a thorough investigation of TIN contamination in the most common types of haematological cancers and sources of normal DNA. Haematological cancers were the ideal candidates for this assessment due to the natural spread of tumour cells in the bloodstream. Previous studies have demonstrated that saliva DNA from MPN patients can be positive for *JAK2 p.V617F* mutations (Barnholt et al. 2012; Tenedini et al. 2014). A more recent study also reported contamination in saliva samples of MPN patients due to leucocyte presence in the oral mucosa, and suggested the use of CD3+ T cells as a source of normal sample (Garcia-Gisbert et al. 2020). Strikingly, we found that normal DNA derived from sorted CD3+ T cells for MPN patients show clear signs of contamination and exhibit JAK2 mutations with high read support.

We also found a high prevalence of TINC in saliva samples for AML patients. Similarly to our finding, other studies report the oral cavity to present the first clinical manifestations of leukaemia (Quispe et al. 2021), with gingival infiltration of AML cells demonstrated by biopsy (Fernandes et al. 2018; Lupi et al. 2018) and most commonly seen in acute monocytic leukaemia and acute myelomonocytic leukaemia (Felix and Lukens 1986; McKenna 2000). Our findings are therefore consistent with previous literature, with the adjoint value of presenting also a methodology for contamination assessment that goes beyond inspecting the frequency of specific mutations in the normal, but rather leverages the full power of modern WGS assays.

Given the incidence of these tumour types and current standard practises for normal sample collection, findings with our TINC tool are crucial for improving best practice guidelines for sample handling and highlight the importance of thorough quality control processes, particularly when genomic data are used to inform clinical decisions. Failure to identify TIN contamination increases the potential for false negative somatic variants, particularly those that occurred within the earliest stages of tumorigenesis and thus

have the highest representation in the sequencing data. These mutations could determine disease course (Levine, Jenkins, and Copeland 2019), stratify patients with respect to treatment response (McGranahan and Swanton 2017), and inform targets for therapy (McGranahan et al. 2016), and so failure to detect them carries a high risk for clinical reporting, familial screening, and basic cancer research.

Reassuringly, our findings demonstrated the validity of specific normal DNA sources (e.g., cultured fibroblasts) for haematological cancers, and the overall lack of contamination in a class of solid tumours (sarcomas). This analysis could be extended to a broader range of cancer types, especially focusing on those that have begun to spread (such as in late-stage patients) and might present contamination. However, the high incidence and level of contamination observed for other tumour and DNA source types highlights the importance of bioinformatics pipelines that can accommodate TIN contamination. Some variant calling algorithms are now attempting to mitigate the impact of contaminated normal samples, (Nangali 2021), but this is more difficult for complex and structural variants for which variant calling from short read data is more challenging and less well established.

The high proportion and high level of contamination observed for some haematological cancer subtypes demonstrates the importance of careful consideration of normal sample source and collection protocols. It is paramount that bioinformatics pipelines are capable of detecting and reporting TIN contamination and subsequently mitigating the impact on somatic variant detection. Thus we recommend that assessment of tumour in normal contamination using a tool such as TINC becomes part of standard quality control procedures for tumour-normal matched pair analyses, especially in clinical settings where data are being used in patient care.

Data Availability

The TINC method is available as an open source R package hosted at GitHub, with manual and RMarkdown vignettes that present TINC analyses.

- https://caravagnalab.github.io/TINC/

Somatic variant data from the Genomics England data set were analysed within the Genomics England Research Environment secure data portal, under Research Registry project code RR-360, and exported from the Research Environment following data transfer request 100000000985 and 1000000009912 on 17 February 2021.

The Genomics England data set can be accessed by joining the community of academic and clinical scientists via the Genomics England Clinical Interpretation Partnership (GeCIP), https://www.genomicsengland.co.uk/about-gecip/. To join a GeCIP domain, the following steps have to be taken:

1. Your institution has to sign the GeCIP Participation Agreement, which outlines the key principles that members of each institution must adhere to, including our Intellectual Property and Publication Policy.

2. Submit your application using the relevant form found at the bottom of the page (https://www.genomicsengland.co.uk/join-a-gecip-domain/).

3. The domain lead will review your application, and your institution will verify your identity for Genomics England and communicate confirmation directly to Genomics England.

4. Your user account will be created.

5. You will be sent an email containing a link to complete Information Governance training and sign the GeCIP rules (https://www.genomicsengland.co.uk/wp-content/uploads/2019/07/GeCIP-Rules_29-08-2018.pdf). Completing the training and signing the GeCIP Rules are requirements for you to access the data. After you have completed the training and signed the rules, you will need to wait for your access to the Research Environment to be granted.

6. This will generally take up to one working day. You will then receive an email letting you know your account has been given access to the environment, and instructions for logging in (for more detail, see: https://www.genomicsengland.co.uk/join-a-gecip-domain/). Details of the data access agreement can be retrieved from https://figshare.com/articles/GenomicEnglandProtocol_pdf/4530893/5.

All requests will be evaluated by the Genomics England Access Review Committee taking into consideration patient data protection, compliance with legal and regulatory requirements, resource availability and facilitation of high-quality research. All analysis of the data must take place within the Genomics England Research Environment secure data portal, https://www.genomicsengland.co.uk/understanding-genomics/data/ and exported following approval of a data transfer request.

All other data supporting the findings of this study are available from the corresponding author upon request.

Acknowledgments

This research was made possible through access to the data and findings generated by the 100,000 Genomes Project. The 100,000 Genomes Project is managed by Genomics England Limited (a wholly owned company of the Department of Health and Social Care). The 100,000 Genomes Project is funded by the National Institute for Health Research and NHS England. The Wellcome Trust, Cancer Research UK and the Medical Research Council have also funded research infrastructure. The 100,000 Genomes Project uses data provided by patients and collected by the National Health Service as part of their care and support. The research leading to these results has received funding from AIRC under MFAG 2020 - ID. 24913 project – P.I. Caravagna Giulio.

We wish to thank Jude Fitzgibbon and Trevor A Graham for helpful discussions on a preliminary version of this manuscript.

Authors contribution

GC, JM and AS conceived TINC, which GC formalised and implemented with support from SM. MZ generated contaminated BAM files for synthetic tests. JM ran all analyses and synthetic tests; JM, GC and AS analysed the results. JC and JB performed experimental validation. GC, AS, SW, JB and JM drafted the manuscript, which all authors finalised and approved in its final version.

Competing interests

The authors declare no competing interests.

Online Methods

The TINC method

TINC tracks putative clonal somatic SNVs in the tumour and normal samples to determine the overall purity of the tumour (TIT score), and the contamination of tumour cells in the normal biopsy (TIN score). As input it requires the read counts of somatic SNVs (e.g. Strelka2 VCF file) and, if available, the CNA segments (e.g. Canvas VCF file).

TIT and TIN scores are computed as fractions (0 to 1). These can be represented either as the fraction of tumour cells or the fraction of reads originating from tumour DNA in the normal sample. These values are equivalent if the tumour genome is diploid. The conversion between the two values (cell fraction and read fraction) requires the knowledge of tumour CNAs in order to normalise observed VAF values for the tumour ploidy profile.

Bulk tumour deconvolution. The subset of clonal SNVs is derived from the input set of somatic variants using the MOBSTER mixture model (Caravagna, Heide, et al. 2020; Caravagna, Sanguinetti, et al. 2020). For data x, the likelihood is that of a $(k+1)$-dimensional Dirichlet mixture

$$f(x \mid \theta, \pi) = \pi_1 \mathrm{PL}(x \mid \theta) + \sum_{i=2}^{k} \pi_i \mathrm{Beta}(x \mid \theta)\,.$$

Here θ are model parameters considering the density function $PL(x|\theta)$ of a Pareto Type-I power law, and $Beta(x|\theta)$ for a Beta distribution; π are mixing proportions representing the proportion of mutations assigned to each cluster.

The fit is carried out to select the best possible model, using a score function based on the principles of the integrated classification likelihood (ICL), which extends the likelihood $f(x|\theta,\pi)$ with a regularisation for the complexity of the model - measured by $|\theta| + |\pi|$ - akin to the Bayesian information criterion, and include the separation of the clusters through an entropy term defined over latent variables $z_{n,k}$ (clustering responsibilities)

$$\mathrm{ICL} = -2\log f(x|\theta,\pi) + (|\theta| + |\pi|)\log(n) + \sum_{z_{n,k}} z_{n,k}\log(z_{n,k})$$

Through a gradient-based procedure MOBSTER optimises the value of k and the presence of the tail in the data as defined by the PL density, improving largely over standard methods for subclonal deconvolution (Caravagna, Heide, et al. 2020). In general, MOBSTER can be used to study the full tumour architecture, a more broad and complicated problem than just determining clonal mutations, as we perform here; nonetheless, the tool is fast and provides precise evolutionary information that TINC can use to estimate contamination.

From MOBSTER fits we obtain a set of vector-valued latent variables $z_{n,k}$ reporting the probability of each input mutation to be part of one of a set of clusters; these contains the normalised posterior densities

$$z_{n,k} = \frac{\pi_k g(x \mid \theta_k)}{\sum_i \pi_i g(x \mid \theta_i)}$$

were $g(\cdot)$ are the density functions used in the mixture (power-law and Beta, depending on the component index i).

By design MOBSTER labels with "C1" (or C_1) the set of mutations with the highest VAF. These should be the clonal mutations we need in TINC, unless there are CNA events in the data and TINC is run without input CNAs. If that is the case, e.g. the tumour bears some miscalled large loss of heterozygosity (LOH), then C_1 might be artifactually related to the CNA event. Mistaking such a cluster for clonal mutations would inflate TIT and TIN estimates; therefore TINC maps mutations in C_1 to chromosomes, and tests for their enrichment when it is run without CNA data.

Mutations mapping is done by a function $map(\cdot)$ that returns the chromosome counts were a set of mutations map onto the genome. Cluster C1 is rejected - as the putative clonal cluster - by using an empirical 60/20 rule: if more than 60% of mutations in C_1 map to less than 20% of the chromosomes (we define this to be the case in which $map(C_1, 0.6) < 0.2$). The 60/20 cutoffs are determined from analysis of pan-cancer WGS data at Genomics England.

When C1 is rejected, TINC performs a recursive test for clusters with progressively lower Beta means, stopping when a suitable cluster is found or all clusters are rejected; in the latter case TINC determines it is impossible to assess reliable clonal mutations. In practice, the TINC algorithm selects C_i such that

$$map(C_i, 0.6) > 0.2 \ \wedge \ \forall j < i.\ map(C_j, 0.6) < 0.2$$

When TINC is run with CNA data this test is not required since we are already filtering mutations by the tumour's most prevalent karyotype. This choice allows control for CNAs that confound the VAF distribution (Household et al. 2021).

Note that by running TINC with CNA one elicits the assumption that CNA segments are correct. We also note that, using CNAs for tumour with high aneuploidy (i.e., cases with copy neutral LOH, triploid or tetraploid genomes etc.), the mutations associated with the cluster with highest VAF are those that happened before the copy number event. In other words, if we work with a tumour that is prevalently triploid with two and one copies of the major and minor alleles, the putative subset of clonal mutations that we find in the high-VAF cluster have happened before the amplification of the major allele.

When TINC has identified a cluster w of putative clonal mutations, it selects the ones with assignment probability above a threshold $z_+ > 0$, disregarding all others. I.e., it defines

$$C = \{n \mid z_{n,w} > z_+\}$$

If there are not enough such mutations, z is decreased until we include a predefined number of mutations that the user can decide. For instance, TINC can be parameterized to search for all mutations with at least 90% probability to be assigned to the clonal cluster, requiring at least $n = 150$ mutations back. With its dynamic-cutoff strategy, TINC might determine that, in order to select 150 clonal mutations, z_+ must be decreased to 80%.

Bulk normal deconvolution. Read counts for putative clonal mutations are collected from the normal biopsy, and then used by TINC to fit a Binomial mixture model. This is available in the open-source R package BMix (Caravagna, Heide, et al. 2020), which provides univariate Binomial and Beta-Binomial mixtures. For a set of clonal mutations with associated sequencing depth $d > 0$, and $r \geq 0$ reads with the alternative allele, we use the likelihood

$$f(x \mid \theta, \pi) = \sum_{i=1}^{k} \pi_i \mathrm{Bin}(r \mid d, \theta) \,.$$

Here we have a mixture of $k \geq 1$ components, where b is a Binomial density function for d trials with r successes, and unknown success probability p (parameter in θ). In this case BMix optimises the value of k scoring models similarly to MOBSTER, and then returns clustering assignments, latent variables, and Binomial parameters $p_1,\ldots,p_k$ similar to MOBSTER. We omit the formulation and refer to (Caravagna, Heide, et al. 2020).

TIT and TIN scores. The computation of TIN and TIT scores is done after deconvolution of tumour and normal bulks. In both cases TINC uses the same principle to normalise for the tumour genome karyotype, which requires knowing copy number data or assuming a known state (e.g., diploid).

We use our CNAqc open-source R package for WGS to make these conversions (Household et al. 2021). For mutations mapping on a genome segment with m copies of the minor allele, M of the major, allelic frequency v and mutation multiplicity μ (i.e., number of copies of the mutation on the genome), the tumour purity ρ is determined by

$$[\mu + v(2 - m - M)]\rho = 2v$$

Solving for ρ we determine the percentage of tumour cells in the biopsy.

When TINC is run without CNA data, the model assumes that the above equation is to be solved for a diploid tumour ($m = M = \mu = 1$).

This equation is used for both TIT and TIN estimation, using different values for v in the tumour, and in the normal.

- For TIT, v is the mean VAF of clonal mutations identified by TINC (mapping to cluster w in MOBSTER analysis)

  $$v = mean\{VAF_x \mid x \in w\}$$

- For TIN, v is expressed as a linear combination of the Binomial peaks and cluster sizes reported by BMix, i.e., a mean of $p_1,\ldots,p_k$ weighted by the mixing proportions $\pi_1,\ldots,\pi_k$,

  $$v = \sum_{i=1}^{k} \pi_i p_i$$

For the normal biopsy, we use all the mutations determined in the tumour as clonal (red MRCA in Figure 1b). One could be tempted to subtract those with VAF above 0 from that set, which are theoretically those accruing from the MRCA of both the cells in the tumour and the normal (yellow MRCA in Figure 1b). However, we are expecting to work with quite low VAF values corresponding to a normal with contained levels of contamination. Given the median coverage of WGS assays available in Genomics England for normal biopsies (30x), we cannot neglect a strong effect of Binomial sampling (sequencing observational model) on the observed counts. For this reason, retaining all highly-confident clonal mutations in the tumour is a reasonable conservative choice. Their effect on TINC estimation is weighted by their proportions, as obtained from BMix clustering.

TIT and TIN scores can be finally converted to units of read counts by solving for another equation. If we denote with λ the tumour cellular fraction, the read counts fraction η is

$$\eta = \frac{\lambda(m+M)}{\lambda(m+M)+2-2\lambda}$$

This equation follows from simple arguments about allele multiplicity (Dentro, Wedge, and Van Loo 2017; Van Loo et al. 2010; Househam et al. 2021).

Complementary multivariate analysis. A variational Binomial mixture model is used jointly on read counts data of both the tumour and normal biopsies; this is available through the VIBER open-source R package (Caravagna, Heide, et al. 2020), which is designed to implement multi-dimensional mixtures with arbitrary dimensions..

This type of mixture is semi-parametric and determines the number of clusters in an automatic fashion via variational inference, using the likelihood function

$$f(x|\theta,\pi) = \sum_{i=1}^{k} \pi_i \mathrm{Bin}(r_t|d_t,\theta_{i,t})\mathrm{Bin}(r_n|d_n,\theta_{i,n})$$

where the counts are considered for both the normal (d_n and r_n) and tumour (d_t and r_t) assays. Here we are assuming that the counts are independent, as obtained from two sequencing runs.

From VIBER outputs TINC checks the position of the clonal mutations identified for tracking. When these associate a single cluster - as one might expect - we can obtain an alternative TIT and TIN set of scores; TINC uses these to confirm the original estimates obtained by using MOBSTER and BMix, and reports this further evidence to the user.

Mapping/variant calling pipeline

Genomics England Pipeline 2.0 (implemented in November 2020). All samples were sequenced on HiSeq platform to an average coverage of 100x for tumour and 30x for normal. Illumina's North Star pipeline (version 2.6.53.23) was used for primary WGS analysis. Read alignment against the human reference genome GRCh38+Decoy+EBV was performed with DRAGEN software (version 3.2.22). Small variant calling together with tumour-normal subtraction was performed using Strelka (version 2.9.9). In addition to default Strelka filters we applied the following additional filters in order to reduce the false positive rate in the set of somatic variants: i) Variants with a population germline allele frequency above 1% in the Genomics England or gnomAD datasets; ii) Recurrent somatic variants with frequency above 5% in the Genomics England dataset; iii) Variants overlapping simple repeats as defined by Tandem Repeats Finder; iv) Small indels in regions with high levels of sequencing noise where at least 10% of the basecalls in a window extending 50 bases to either side of the indel's call have been filtered out by Strelka due to the poor quality; v) SNVs resulting from systematic mapping and calling artefacts. We tested whether the ratio of tumour allele depths at each somatic SNV site were significantly different to the ratio of allele depths at this site in a panel of normals (PoN) using Fisher's exact test. The PoN was composed of a cohort of 7000 non-tumour genomes from the Genomics England dataset, and at each genomic site only individuals not carrying the relevant alternate allele were included in the count of allele depths. The mpileup function in bcftools v1.9 was used to count allele depths in the PoN, and to replicate Strelka filters duplicate reads were removed and quality thresholds set at mapping quality >= 5 and base quality >= 5. All somatic SNVs with a Fisher's exact test phred score < 50 were filtered. Copy number aberrations were identified with Canvas 1.39.

Genomics England Pipeline 1.0. Bioinformatics pipeline used by Genomics England in the 100,000 Genomes Project. This pipeline utilised Strelka version 2.4.7 for somatic small variant detection which is not adapted for variant detection at any level of TIN contamination.

In silico contamination data generation

Tumour in normal contamination was generated *in silico* for cohorts of haematological and lung cancer tumour-normal pairs, with samples selected based on a Ccube tumour purity estimate > 30% and baseline normal samples checked to ensure they were not contaminated (Yuan et al. 2018). Various levels of tumour in normal contamination were created for each tumour-normal pair by using samtools to combine fractions of the normal and tumour BAM files. The level of contamination was calculated accounting for the purity of the tumour sample predetermined with Ccube (Yuan et al. 2018). Using the *in silico* contaminated normal bam files TINC R package input was generated using the Genomics England pipeline outlined above.

Samples and data collection

All samples as well as sample metadata were collected as part of Genomics England 100,000 Genomes Project (Turnbull et al. 2018). Sample Specifications for haematological cancers are described in Supplementary Table 1.

References

Arber, Daniel A., Attilio Orazi, Robert Hasserjian, Jürgen Thiele, Michael J. Borowitz, Michelle M. Le Beau, Clara D. Bloomfield, Mario Cazzola, and James W. Vardiman. 2016. “The 2016 Revision to the World Health Organization Classification of Myeloid Neoplasms and Acute Leukemia.” *Blood* 127 (20): 2391–2405.

Barnell, Erica K., Peter Ronning, Katie M. Campbell, Kilannin Krysiak, Benjamin J. Ainscough, Lana M. Sheta, Shahil P. Pema, et al. 2019. “Standard Operating Procedure for Somatic Variant Refinement of Sequencing Data with Paired Tumor and Normal Samples.” *Genetics in Medicine: Official Journal of the American College of Medical Genetics* 21 (4): 972–81.

Barnholt, Kimberly E., David A. Hinds, Amy K. Kiefer, Chuong B. Do, Nicholas Eriksson, Joanna L. Mountain, Uta Francke, et al. 2012. “Estimation of JAK2 V617F Prevalence by Detection of the Mutation in Saliva Samples From Online MPN and General Population Cohorts.” *Blood* 120 (21): 1737.

Caravagna, Giulio, Timon Heide, Marc J. Williams, Luis Zapata, Daniel Nichol, Ketevan Chkhaidze, William Cross, et al. 2020. “Subclonal Reconstruction of Tumors by Using Machine Learning and Population Genetics.” *Nature Genetics* 52 (9): 898–907.

Caravagna, Giulio, Guido Sanguinetti, Trevor A. Graham, and Andrea Sottoriva. 2020. “The MOBSTER R Package for Tumour Subclonal Deconvolution from Bulk DNA Whole-Genome Sequencing Data.” *BMC Bioinformatics* 21 (1): 531.

Dentro, Stefan C., David C. Wedge, and Peter Van Loo. 2017. “Principles of Reconstructing the Subclonal Architecture of Cancers.” *Cold Spring Harbor Perspectives in Medicine* 7 (8). https://doi.org/10.1101/cshperspect.a026625.

Felix, D. E., and J. Lukens. 1986. “Oral Symptoms as a Chief Sign of Acute Monoblastic Leukemia: Report of Case.” *Journal of the American Dental Association* 113 (6): 899–900.

Fernandes, Karin Sá, Marina Gallottini, Talita Castro, Mauricio Flamínio Amato, Juvani Saturno Lago, and Paulo Henrique Braz-Silva. 2018. “Gingival Leukemic Infiltration as the First

Manifestation of Acute Myeloid Leukemia." *Special Care in Dentistry: Official Publication of the American Association of Hospital Dentists, the Academy of Dentistry for the Handicapped, and the American Society for Geriatric Dentistry* 38 (3): 160–62.

Gagan, Jeffrey, and Eliezer M. Van Allen. 2015. "Next-Generation Sequencing to Guide Cancer Therapy." *Genome Medicine* 7 (1): 80.

Garcia-Gisbert, Nieves, Laura Camacho, Lierni Fernández-Ibarrondo, Concepcion Fernández-Rodriguez, Raquel Longarón, Joan Gibert, Anna Angona, et al. 2020. "Analysis of Saliva Samples and Cluster of Differentiation 3 (CD3)+ Lymphocytes as a Source of Germline DNA in Myeloproliferative Neoplasms." *British Journal of Haematology* 189 (5): e204–7.

Household, Jacob, Riccardo Bergamin, Salvatore Milite, William C. H. Cross, and Giulio Caravagna. 2021. "Integrated Quality Control of Allele-Specific Copy Numbers, Mutations and Tumour Purity from Cancer Whole Genome Sequencing Assays." *biorXiv*. https://doi.org/10.1101/2021.02.13.429885.

ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. 2020. "Pan-Cancer Analysis of Whole Genomes." *Nature* 578 (7793): 82–93.

Kim, Sangtae, Konrad Scheffler, Aaron L. Halpern, Mitchell A. Bekritsky, Eunho Noh, Morten Källberg, Xiaoyu Chen, et al. 2018. "Strelka2: Fast and Accurate Calling of Germline and Somatic Variants." *Nature Methods* 15 (8): 591–94.

Levine, Arnold J., Nancy A. Jenkins, and Neal G. Copeland. 2019. "The Roles of Initiating Truncal Mutations in Human Cancers: The Order of Mutations and Tumor Cell Type Matters." *Cancer Cell* 35 (1): 10–15.

Lupi, Saturnino Marco, Arianna Rodriguez Y Baena, Gabriele Cervino, Claudia Todaro, and Silvana Rizzo. 2018. "Long-Term Effects of Acute Myeloid Leukemia Treatment on the Oral System in a Pediatric Patient." *The Open Dentistry Journal* 12 (March): 230–37.

McKenna, S. J. 2000. "Leukemia." *Oral Surgery, Oral Medicine, Oral Pathology, Oral Radiology, and Endodontics* 89 (2): 137–39.

Mwenifumbo, Jill C., and Marco A. Marra. 2013. "Cancer Genome-Sequencing Study Design." *Nature Reviews. Genetics* 14 (5): 321–32.

Nangali, Jyoti. 2021. "Personal Communication."

Nowell, P. C. 1976. "The Clonal Evolution of Tumor Cell Populations." *Science* 194 (4260): 23–28.

Quispe, Reyna Aguilar, Elizabeth Marques Aguiar, Claudia Teresa de Oliveira, Ana Cristina Xavier Neves, and Paulo Sérgio da Silva Santos. 2021. "Oral Manifestations of Leukemia as Part of Early Diagnosis." *Hematology, Transfusion and Cell Therapy*, November. https://doi.org/10.1016/j.htct.2021.08.006.

Tate, John G., Sally Bamford, Harry C. Jubb, Zbyslaw Sondka, David M. Beare, Nidhi Bindal, Harry Boutselakis, et al. 2019. "COSMIC: The Catalogue Of Somatic Mutations In Cancer." *Nucleic Acids Research* 47 (D1): D941–47.

Taylor-Weiner, Amaro, Chip Stewart, Thomas Giordano, Mendy Miller, Mara Rosenberg, Alyssa Macbeth, Niall Lennon, et al. 2018. "DeTiN: Overcoming Tumor-in-Normal Contamination." *Nature Methods* 15 (7): 531–34.

Tenedini, E., I. Bernardis, V. Artusi, L. Artuso, E. Roncaglia, P. Guglielmelli, L. Pieri, et al. 2014. "Targeted Cancer Exome Sequencing Reveals Recurrent Mutations in Myeloproliferative Neoplasms." *Leukemia* 28 (5): 1052–59.

Turnbull, Clare, Richard H. Scott, Ellen Thomas, Louise Jones, Nirupa Murugaesu, Freya Boardman Pretty, Dina Halai, et al. 2018. "The 100 000 Genomes Project: Bringing Whole Genome Sequencing to the NHS." *BMJ* 361 (April): k1687.

Van Loo, Peter, Silje H. Nordgard, Ole Christian Lingjærde, Hege G. Russnes, Inga H. Rye, Wei Sun, Victor J. Weigman, et al. 2010. “Allele-Specific Copy Number Analysis of Tumors.” *Proceedings of the National Academy of Sciences of the United States of America* 107 (39): 16910–15.

Yuan, Ke, Geoff Macintyre, Wei Liu, PCAWG-11 working group, and Florian Markowetz. 2018. “Ccube: A Fast and Robust Method for Estimating Cancer Cell Fractions.” *biorXiv*. https://doi.org/10.1101/484402.

Main Text Figures

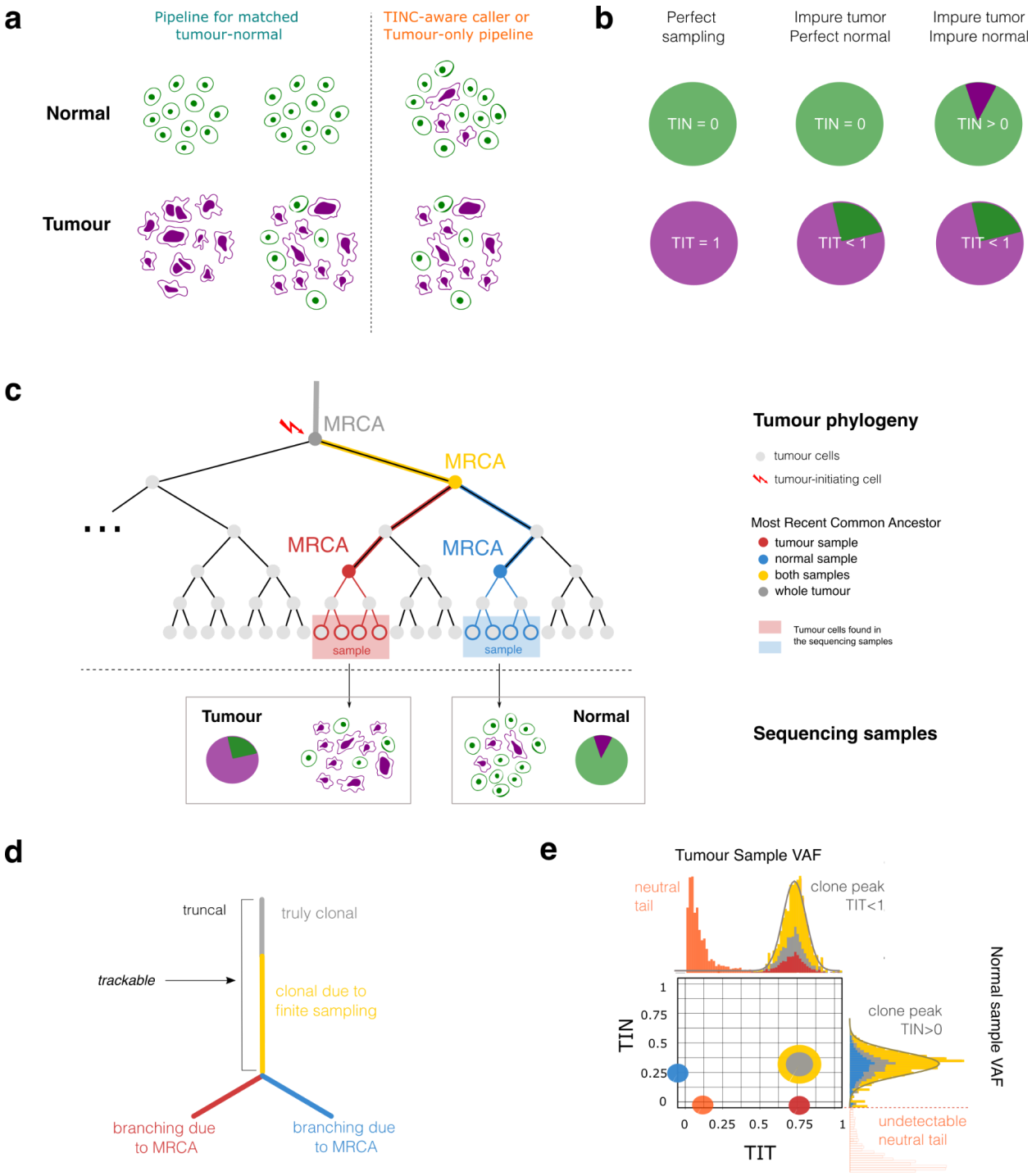


Figure 1. a) Cellular composition of a tumour bulk sample and a normal sample (e.g. peripheral blood, saliva, or skin biopsy). With perfect sampling there would be no cross-contamination between tumour and normal samples (green and purple cells show perfect separation). In

reality, all tumour samples contain normal cells. For a paired analysis, challenges in somatic variant detection arise when the normal sample is contaminated with tumour cells, resulting in subtraction of true somatic variants and a decrease in variant detection sensitivity. b) The level of contamination of a bulk sample can be defined as the fraction of tumour cells observed in the sample. With perfect sampling tumour purity (TIT score) equals 1, and tumour in normal contamination (TIN score) equals 0; for most real-life samples, TIT < 1. When there is tumour in normal contamination, TIN > 0. c) Tumour cell phylogeny showing cell divisions as a tree representing the evolutionary relationship between sampled tumour cells. Colours represent distinct Most Recent Common Ancestors (MRCAs) of the tumour cells, according to sampling. With TIN contamination, phylogenetically related tumour cells are found in both samples (yellow and grey trunk). Tumour cells found in the tumour and the normal carry common (yellow/grey trunk) as well as private mutations (red branch for the tumour sample and blue branch for the normal sample). TIN and TIT are determined using the mutations accrued up to the yellow MRCA, an ancestral cell common to the tumour cells present in both samples. The MRCA is never directly sequenced. d) Summary phylogenetic tree for the cell divisions in panel (c) shows a branching effect that describes a lineage division and spatial sampling bias. e) Expected cell fraction distribution for tumour cells in tumour and normal samples carrying ancestral (yellow and grey) and private (blue and red) mutations for a case with 75% tumour purity (TIT), and 25% tumour in normal contamination (TIN). Somatic mutations common to tumour cells found in both samples, which are frequently subtracted in tumour-normal analysis, are the yellow and grey cluster. These mutations can include the key tumour truncal driver mutations. Mutations only found in the tumour cells within the normal sample (shown in blue) have no read support in the tumour BAM files and are not considered by standard somatic variant callers.

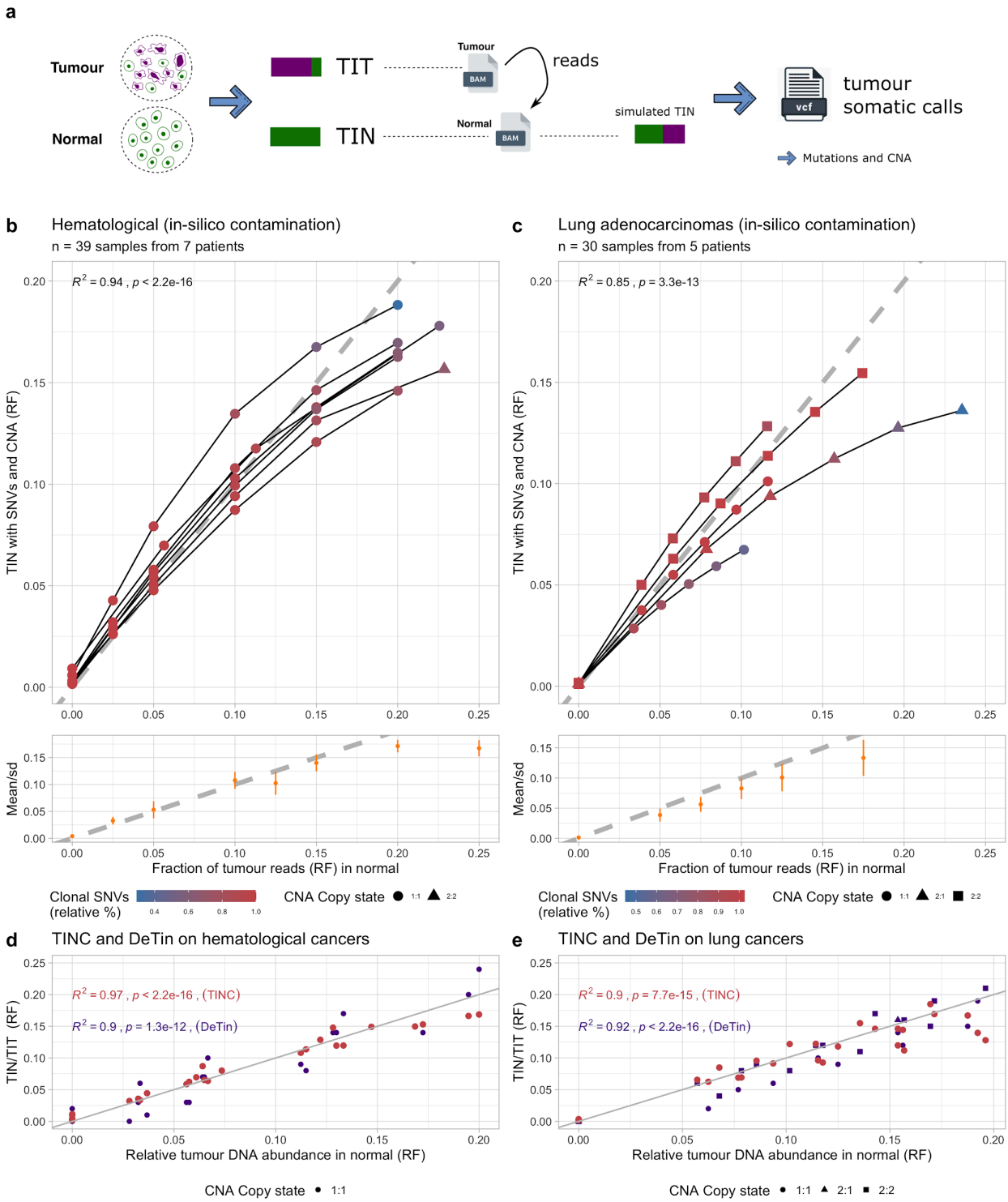


Figure 2. a) Generation of test data by *in silico* contamination of patient WGS datasets. Synthetic data to represent a range of TIN levels were generated from tumour and normal BAM files, injecting reads from the tumour BAM in the normal BAM, according to the tumour purity and the required level of TIN contamination. Somatic variant calling of small variants and CNAs

was performed with the original tumour BAM paired with the *in silico* contaminated normal, and the resulting variant calls used for TINC analysis. b) Performance of TINC with the *in silico* contaminated haematological cancer samples. Using data from cancer participants in the 100,000 Genomes Project with haematological malignancies, 39 datasets were generated with a range of TIN contamination levels. The upper scatter plot compares the expected level of TIN contamination (based on *in silico* contamination) to TINC estimates and the Pearson correlation coefficient. Both axes report the score in read fractions for the tumour (RF). To reflect signal strength, each point is coloured by the percentage of clonal mutations used by TINC, relative to the original uncontaminated sample. The fraction of clonal mutations decreases with increasing contamination, due to the limitations of variant callers that fail to report genuine somatic variants (false negatives). With few clonal mutations, identifying clonal peaks is more difficult; in this case clonal variants are also biased towards those with lower support in the normal sample. The lower plot reports overall TINC mean and standard error, binning RF values using a width of 2.5% (shown in the x-axis); bins with a single observation have undefined standard deviation and are removed. c) Performance of TINC with lung cancer samples contaminated *in silico*. The same information available in panel (b) is provided, using lung adenocarcinoma patient samples (total 30 datasets) available at Genomics England. These tumours have higher aneuploidy compared with haematological cancers, and, e.g., “2:2” refers to genome-doubled cancers with two copies of the major and minor alleles. d) and e) Performance of DeTiN and TINC on the same haematological and lung cancer samples shown in panels (b) and (c). Consistent with the definition of DeTin, the relative tumour DNA abundance in the normal and tumour samples is shown on the x-axis. This plot is restricted to cases with a maximum ratio of 20%, which includes samples within the anticipated contamination range for use in clinical reporting (full plot, Supplementary Figure S1). The y-axis shows the ratio between TIN and TIT scores returned by the two tools. The Pearson correlation coefficient shows strong agreement between expected and calculated levels of TINC for both methods.

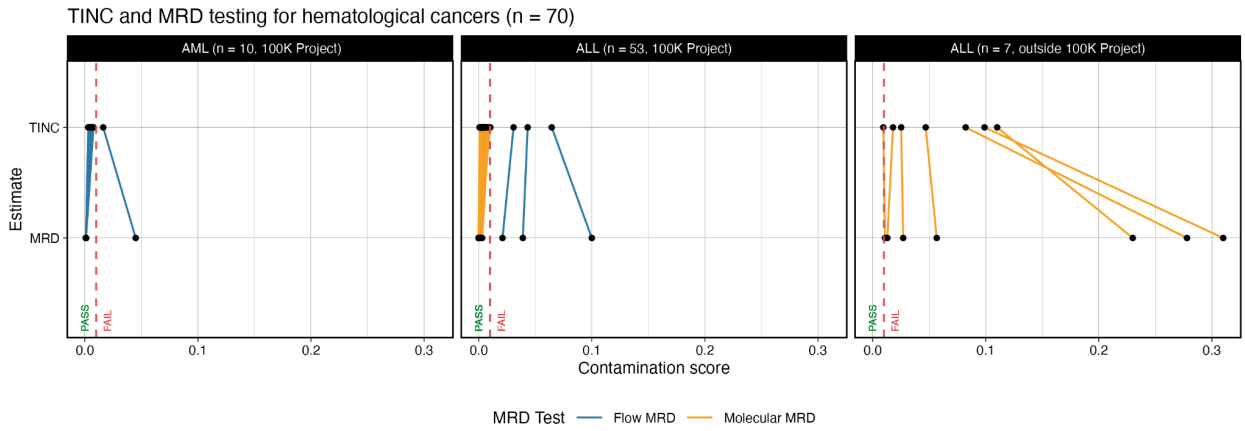


Figure 3. Validation of TINC by comparison with orthogonal test data generated either by flow cytometry or molecular MRD for a cohort of n=70 cases from ALL and AML samples. Here n=63 are from samples recruited in the 100,000 Genomes Project (10 AMLs and 53 ALLs), while 7 ALLs are not (criteria for project enrollment reported in Supplementary Table 1). The threshold for TIN contamination (>1% TIN) is shown with the dashed vertical line.

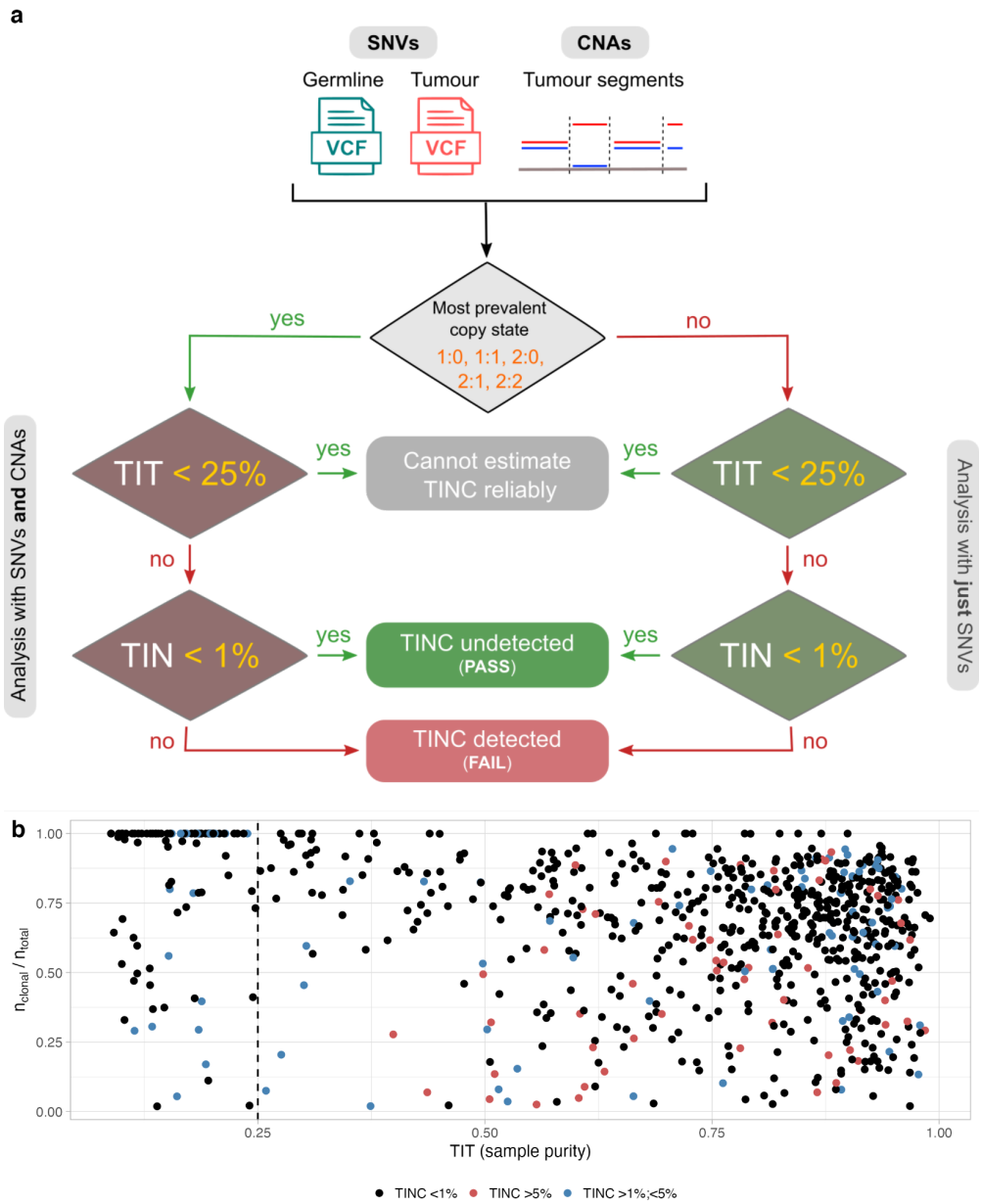


Figure 4. a) Somatic SNVs are used in TIN assessment; by default all variants are used. If run with CNA integration, only SNVs mapping to the most prevalent copy state are used. The supported copy states are 1:0 (loss of heterozygosity, LOH), 1:1 (heterozygous diploid), 2:0 (copy-neutral LOH), 2:1 (triploid) or 2:2 (tetraploid genome-doubled) TIN contamination is estimated for samples with tumour purity (TIT score) >25%; lower purity samples have generally not enough clonal somatic mutations to perform reliable clonal deconvolution. Samples that can be analysed are assigned a TINC score, which can be converted in tumour read fractions (RF) detected in the normal sample, and used to determine a final status for the presence or absence of contamination. The threshold implemented at Genomics England to determine PASS status

(TINC undetected) versus FAIL (TINC detected), is set to 1% RF. b) Scatter reporting the ratio between the number of clonal mutations over total mutational burden, against estimated sample purity (TIT) for 617 WGS samples of haematological cancers. When clonal/total mutations ratio = 1, TINC did not separate clonal somatic variants from subclonal variants and TIN estimates are less reliable. The majority of samples with ratio = 1 are clustered with TIT score < 25%. The colour of each point represents the sample contamination as estimated by our method; the vertical dashed line represents the 25% purity cutoff for TINC analysis adopted in GEL. Further details on this cohort and contamination assessment are shown in Figure 5.

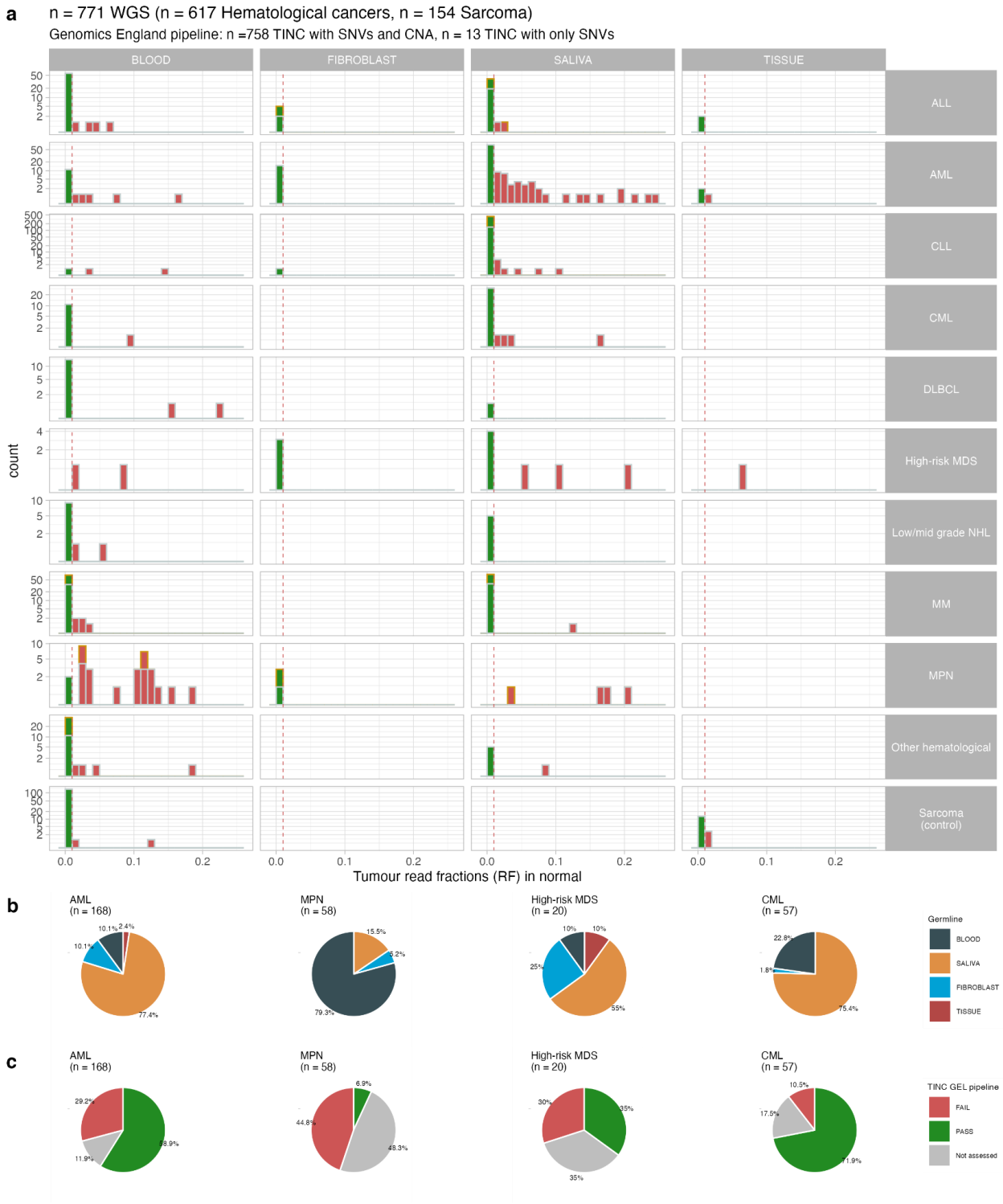


Figure 5. a) Distribution of the estimated level of tumour in normal contamination for 771 tumour-normal pairs derived from participants in the 100,000 Genomes Project (n=617

haematological cancers, n=154 sarcomas). Data are shown for haematological cancers of the subtypes: Acute Lymphoblastic Leukaemia (ALL), Acute Myeloid Leukaemia (AML), Chronic Lymphocytic Leukaemia (CLL), Chronic Myeloid Leukaemia (CML), Diffuse Large B-cell Lymphoma (DLCBL), High-risk Myelodysplastic Syndrome (High-risk MDS), Low and moderate grade Non-Hodgkin B-cell Lymphoma (Low/mid grade NHL), Multiple Myeloma (MM) and Myeloproliferative Neoplasm (MPN). Green bars represent normal samples with <1% TIN score expressed in read fractions, and red bars represent normal samples with >1% TIN. b) Distribution of normal sample source for haematological cancers. The fraction of normal samples for which the DNA was derived from blood, saliva, fibroblasts or tissue samples is shown for haematological cancers of different subtypes (AML, MPN, High-risk MDS and CML). c) The proportion of normal samples determined to have a PASS or FAIL status by TINC (1% read fraction threshold) is shown in red and green respectively for AML, MPN, High-risk MDS and CML cancers. The proportion of cases that could not be analysed by our pipeline (i.e. tumour purity estimated to be below 25%) is shown in grey.

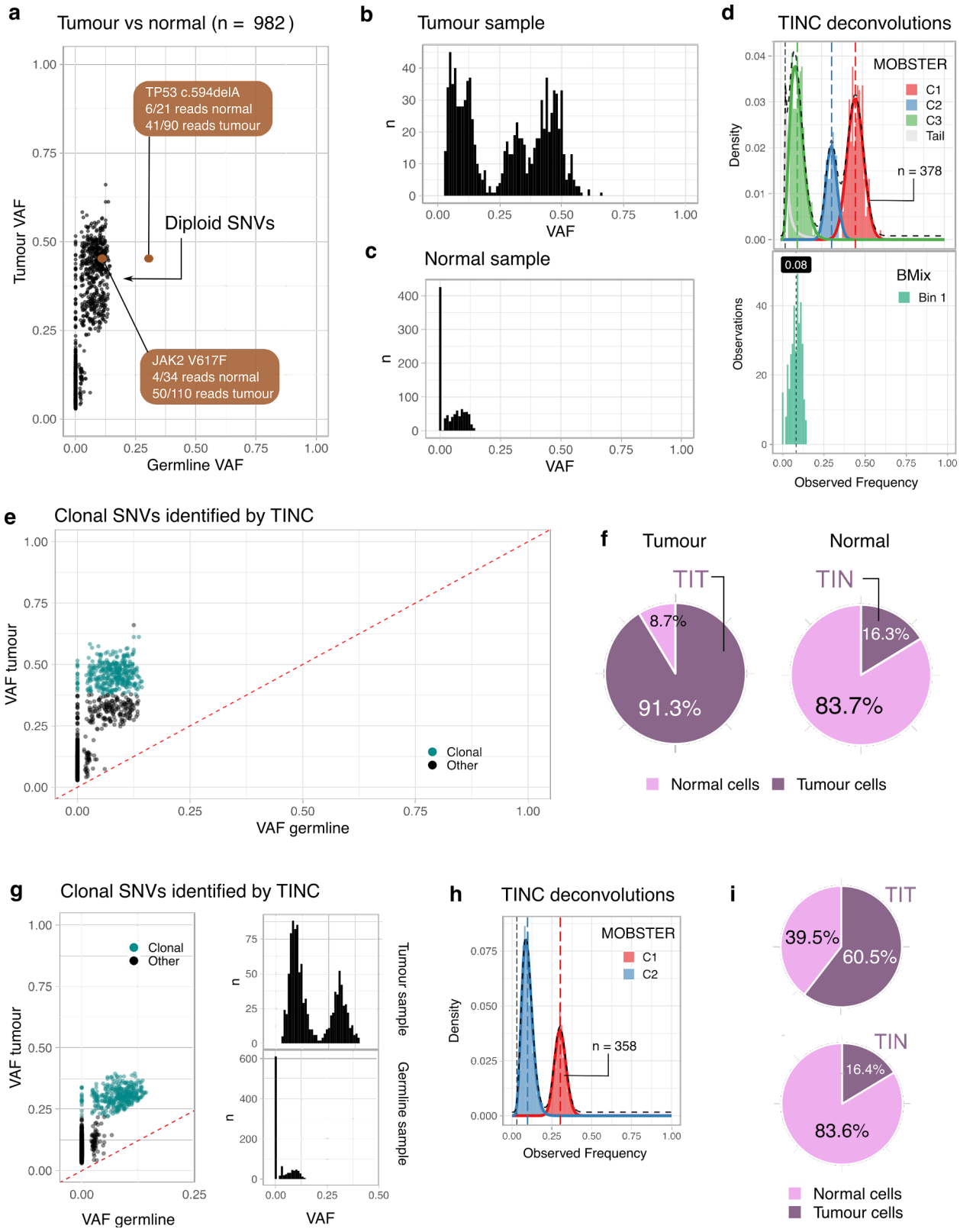
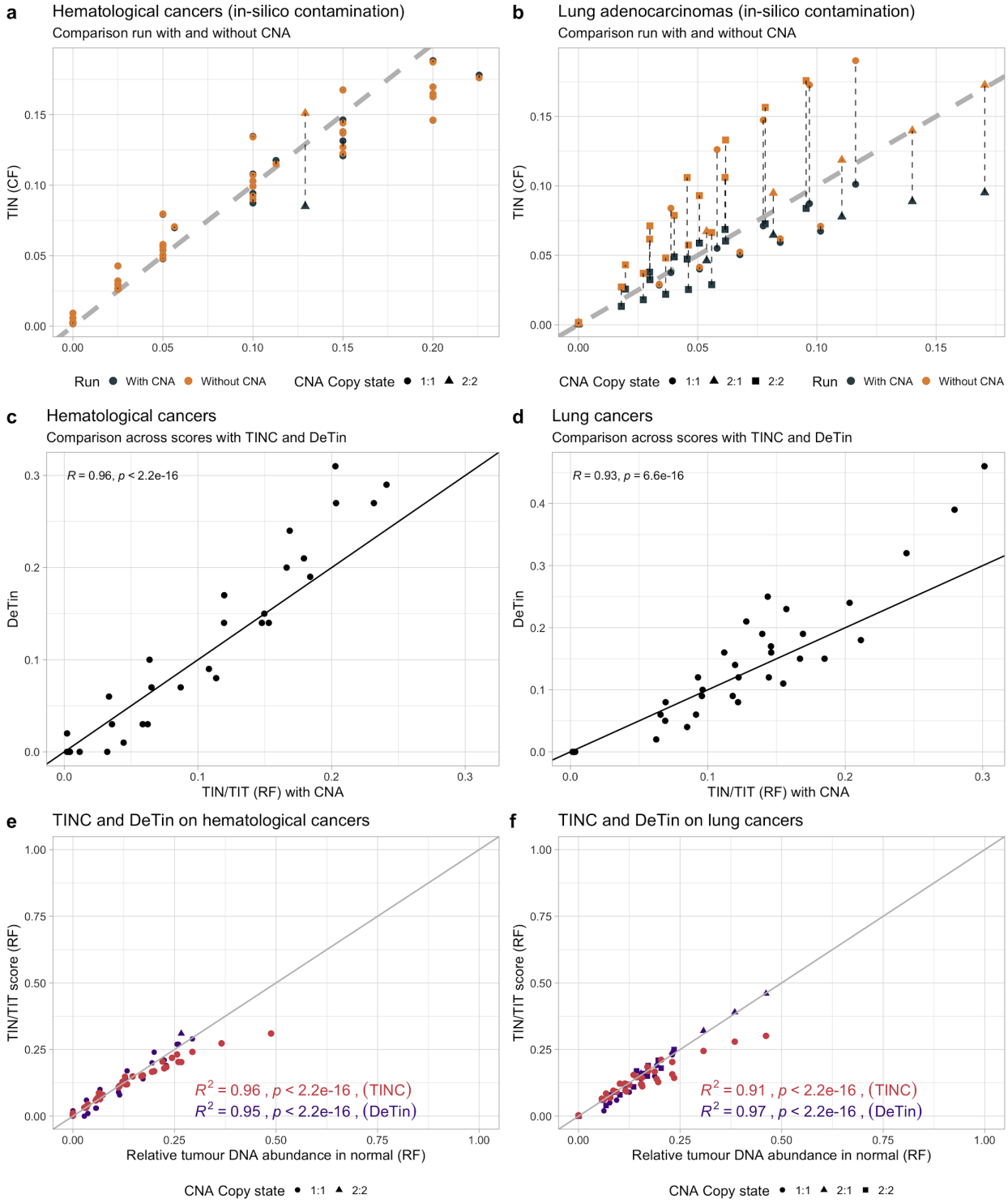
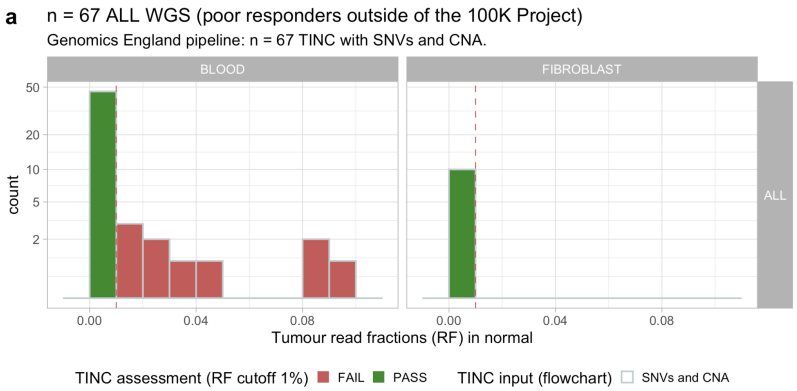

a
Tumour vs normal (n = 982)
TP53 c.594delA
6/21 reads normal
41/90 reads tumour
Diploid SNVs
JAK2 V617F
4/34 reads normal
50/110 reads tumour
Tumour VAF
Germline VAF
b
Tumour sample
n
VAF
c
Normal sample
n
VAF
d
TINC deconvolutions
MOBSTER
C1
C2
C3
Tail
n = 378
Density
0.08
BMix
Bin 1
Observations
Observed Frequency
e
Clonal SNVs identified by TINC
VAF tumour
VAF germline
Clonal
Other
f
Tumour
Normal
TIT
TIN
8.7%
91.3%
16.3%
83.7%
Normal cells
Tumour cells
g
Clonal SNVs identified by TINC
Clonal
Other
VAF tumour
VAF germline
n
Tumour sample
Germline sample
VAF
h
TINC deconvolutions
MOBSTER
C1
C2
n = 358
Density
Observed Frequency
i
TIT
39.5%
60.5%
TIN
16.4%
83.6%
Normal cells
Tumour cells

Figure 6. a) Scatter distribution of somatic mutation VAF in tumour and normal samples (case 1). VAF is shown for n=982 mutations detected from WGS data which reside within heterozygous diploid regions in the tumour genome (1:1). Two variants of clinical significance are highlighted; a *TP53* frameshift deletion (c.594delA) supported by 6 of 21 reads in the normal and 41 of 90 reads in the tumour sample, and a *JAK2* V617F mutation supported by 4 of 34 reads in the normal and 50 of 110 reads in the tumour sample. Neither mutation would be detected using a standard tumour-normal calling pipeline, due to the tumour contamination in the normal. b) and c) Histograms of VAF values for tumour and normal samples in panel (a). d) Deconvolution analysis with TINC (case 1). n=378 clonal mutations were identified in the tumour using MOBSTER (upper panel) with mean VAF ~45% (cluster C1). Subsequent deconvolution determines one cluster in the normal sample for the corresponding mutations with a VAF peak at about ~8% (lower panel). e) Representation of somatic mutation VAF in tumour and normal samples after deconvolution (case 1). After deconvolution of somatic mutations (shown in d), clonality status can be attributed to the mutations shown in panel (a) - clonal mutations are shown with teal dots. f) Representation of contamination in tumour and normal samples (case 1). TIT and TIN scores can be determined from the parameters fit by the deconvolution methods, accounting for the copy state of somatic SNVs. In this case, using heterozygous diploid regions with TINC, when considered in terms of cells percentages in each sample, the data indicate an overall tumour purity of 90% (TIT score, high-purity tumour sample) and tumour in normal contamination level of ~16% (TIN score). g) Representation of somatic mutation VAF in tumour and normal samples (case 2) as in panels (a, b and c). VAF is shown in tumour and normal samples for n=356 somatic mutations. For this case, a translocation resulting in a PML-RARA fusion was not detected (not shown) using a standard tumour-normal analysis pipeline. h) Deconvolution analysis with the TINC method (case 2). Deconvolution identifies a cluster of clonal somatic mutations of n=358 SNVs (cluster C1) with VAF ~30%. i) Representation of contamination in tumour and normal samples (case 2). TIT and TIN scores determined by TINC, expressed in cellular proportions and adjusted for copy number states, show a tumour purity of ~60% (TIT), and tumour contamination of the normal sample of ~16% (TIN).

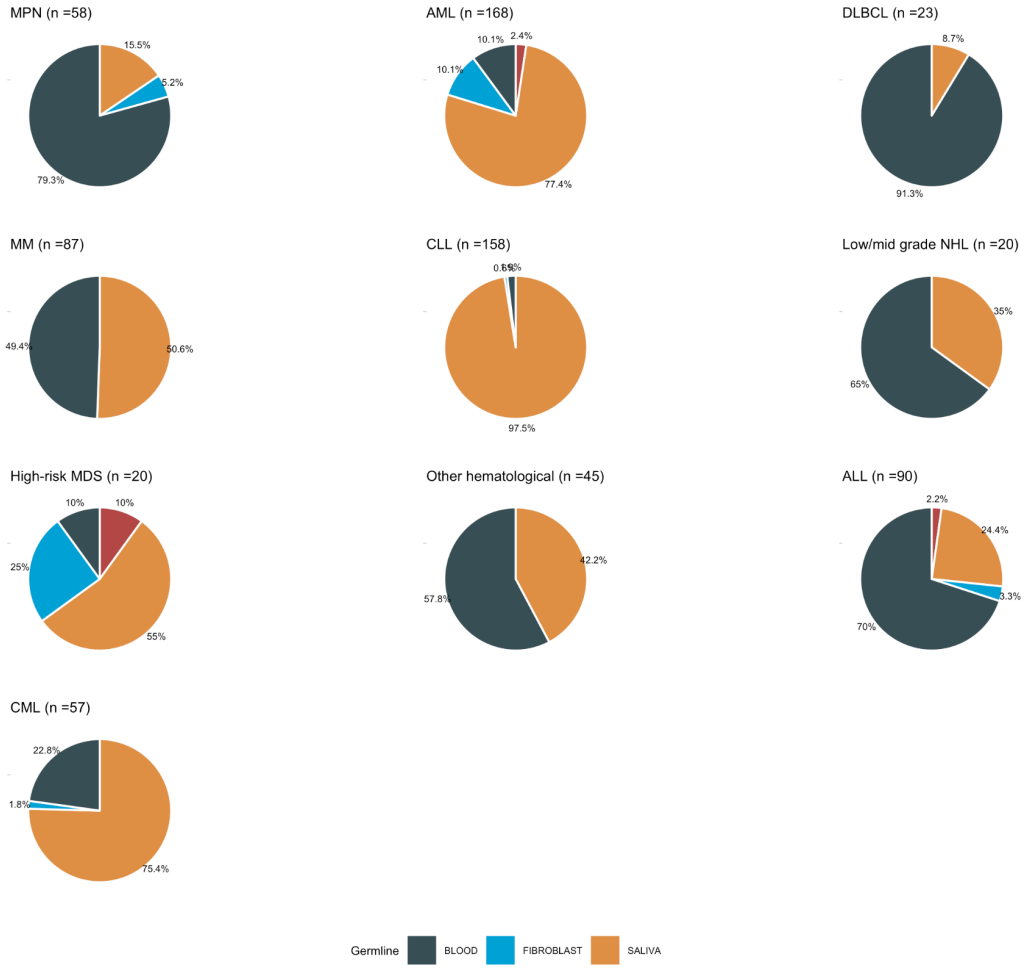
Supplementary Figures



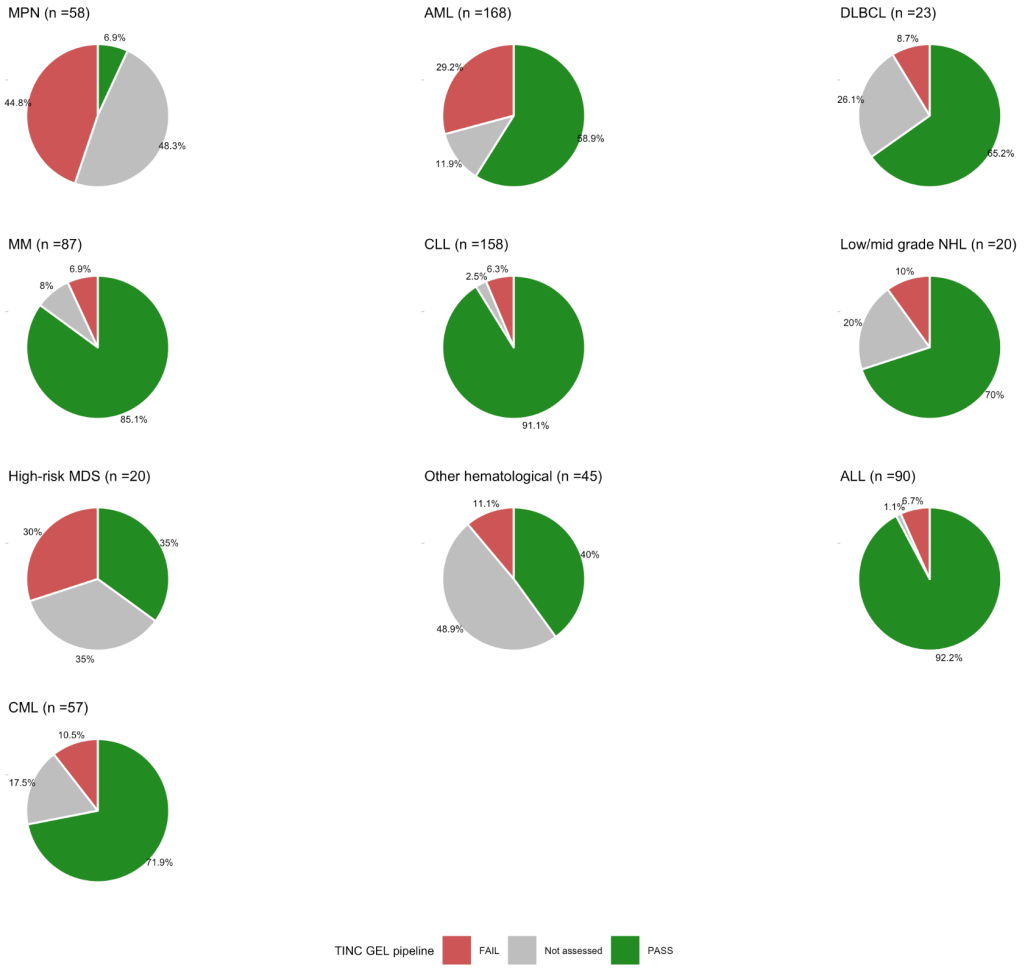
Supplementary Figure S1. **a,b.** In-silico tests with TINC run with and without CNA data, for the two cohorts presented in the Main Text. The x-axis reports contamination in units of cell fractions (CF). The solutions are connected by vertical lines. **c,d.** Correlation between DeTin and TINC scores, computed as ratios to follow DeTin convention. **e.f.** Extended version of Figure 2d,e, to include cases with very high contamination levels - beyond the parameters that are acceptable for the flowchart in Figure 4.



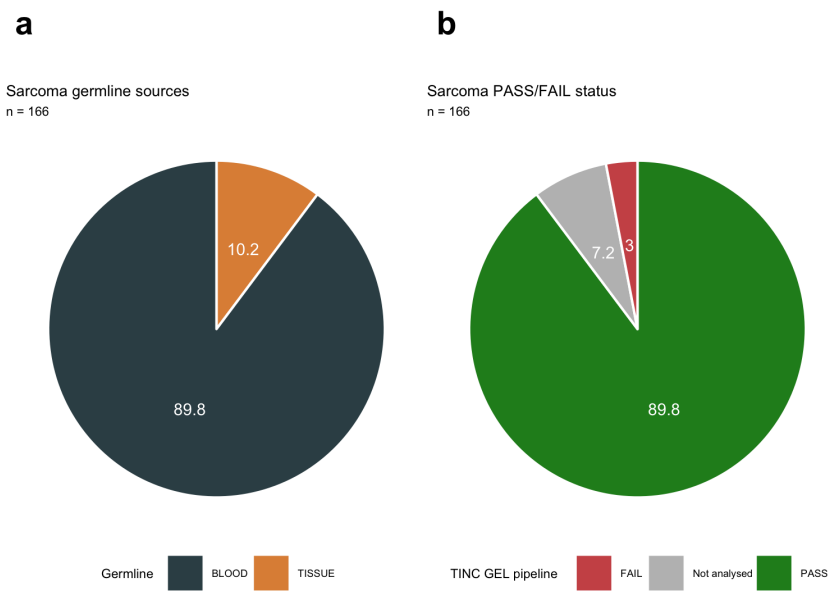
Supplementary Figure S2. **a.** Contamination in tumour read fractions determined in $n = 67$ ALL patients whose normal samples do not fulfil recruitment standards for the 100,000 Genome Project, as reported in Supplementary Table 1. The samples are shown as in Figure 5. Note that blood normal samples from these cohorts show contamination in 11 out 46 cases. All 10 cases with fibroblasts as a source of normal samples show no sign of contamination.



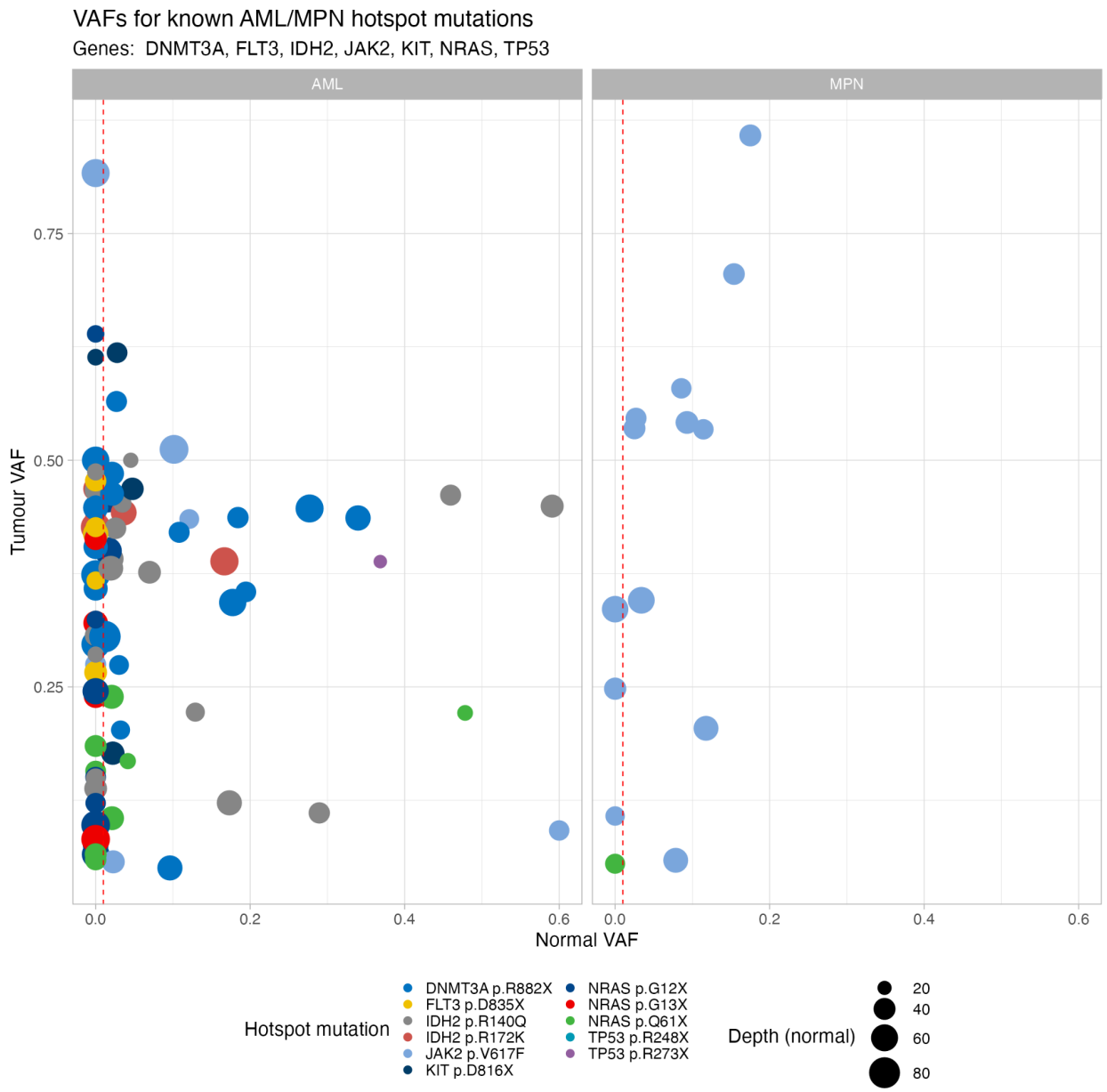
Supplementary Figure S3. Normal sample source proportions for the hematological cohort shown in Figure 5 of the Main Text.



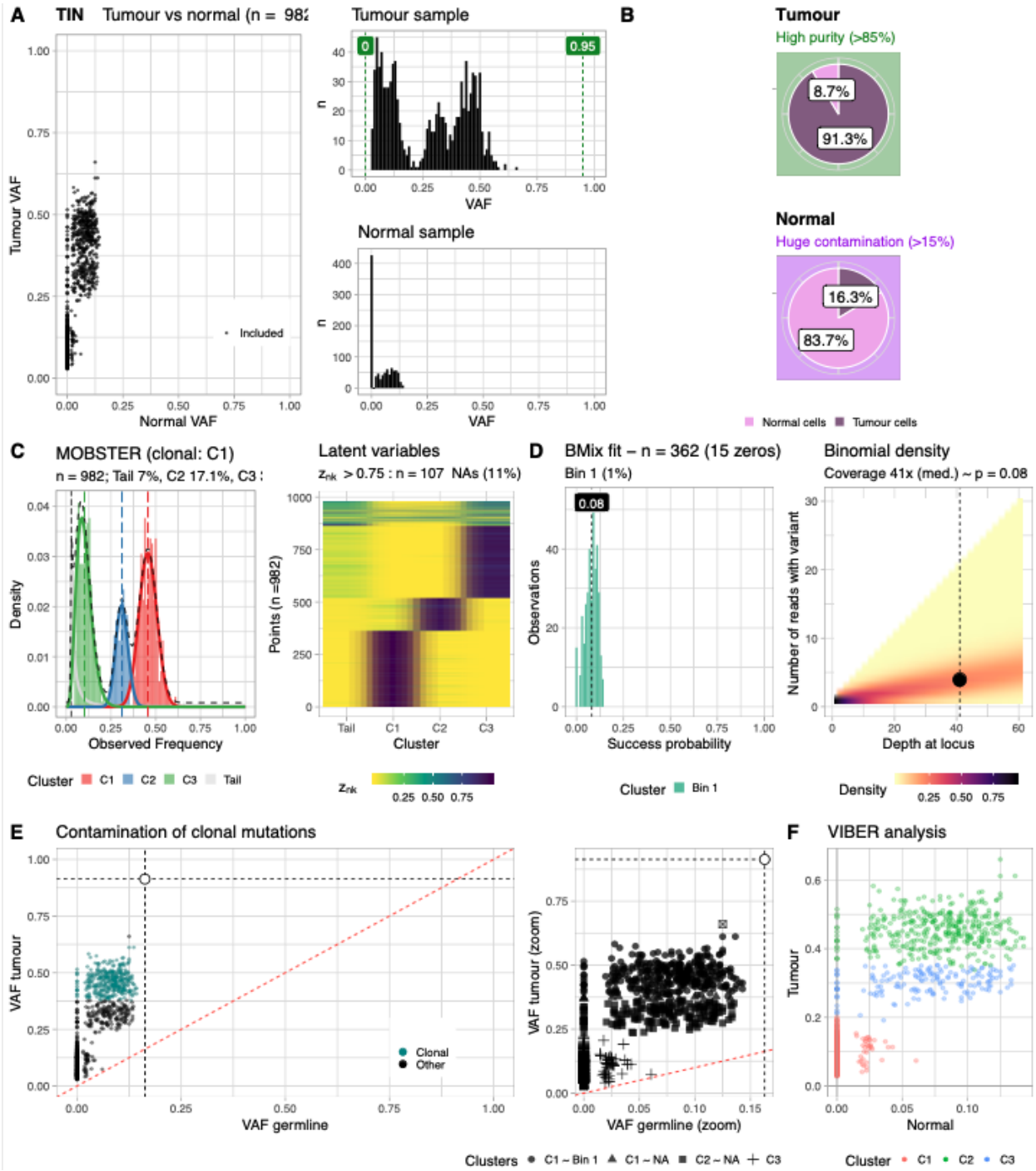
Supplementary Figure S4. PASS and FAIL status for the hematological cohort shown in Figure 5 of the Main Text.



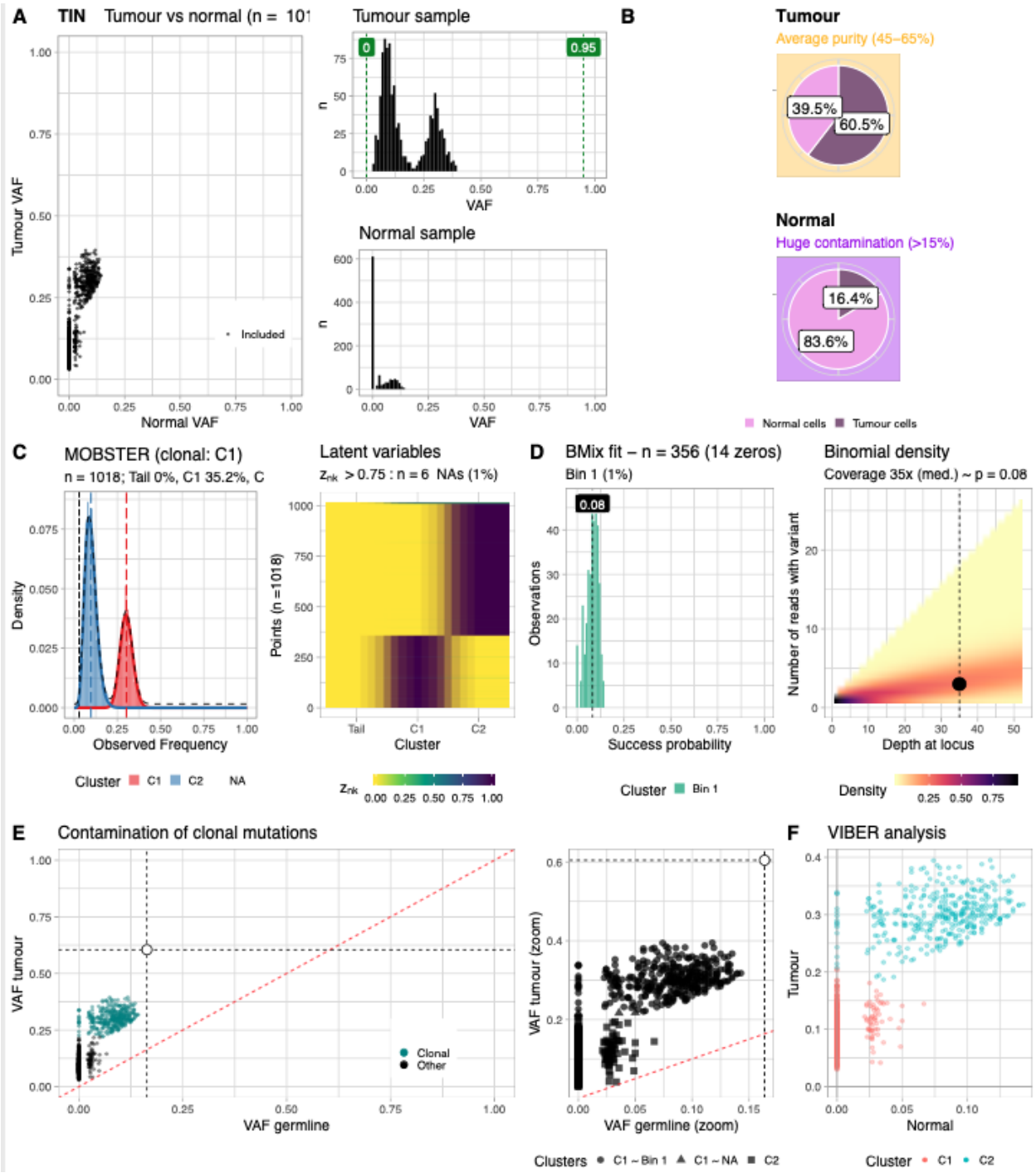
Supplementary Figure S5. **a.** Normal sample source proportions for the sarcoma cohort shown in Figure 5 of the Main Text. **b.** Results from the Genomics England pipeline for the sarcoma cohort. Criteria used to PASS or FAIL samples are as in Figure 5.



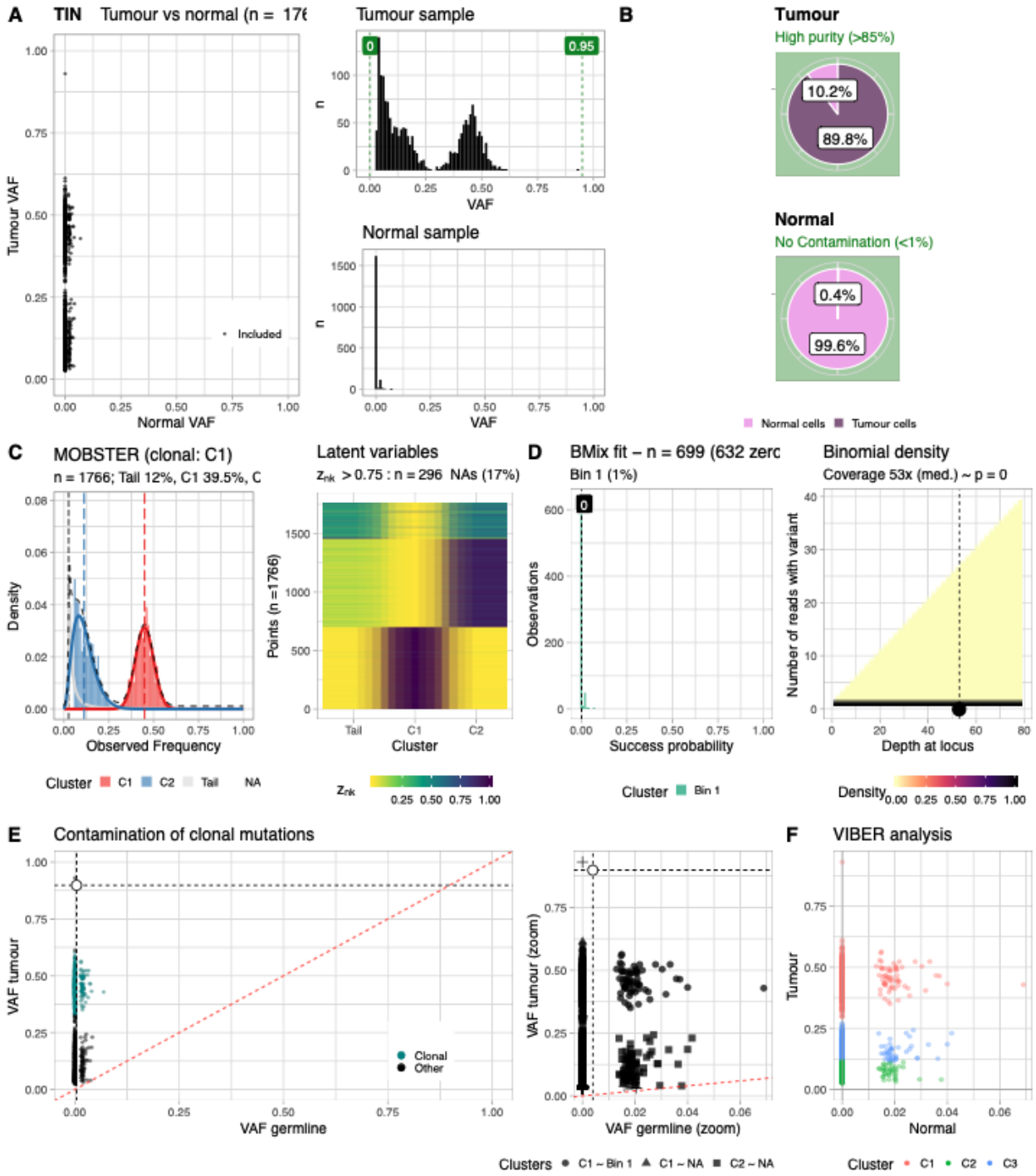
Supplementary Figure S6. Normal and tumour VAFs for a set of hotspot mutations in AML and MPN cohorts: JAK2 (p.V617F), FLT3 (p.D835X), DNMT3A (p.R882X), TP53 (p.R248X, p.R273X), KIT (p.D816X), NRAS (p.G12X, p.G13X, p.Q61X) and IDH2 (p.R140Q, p.R172K). The vertical dashed line denotes a 1% VAF in the normal.



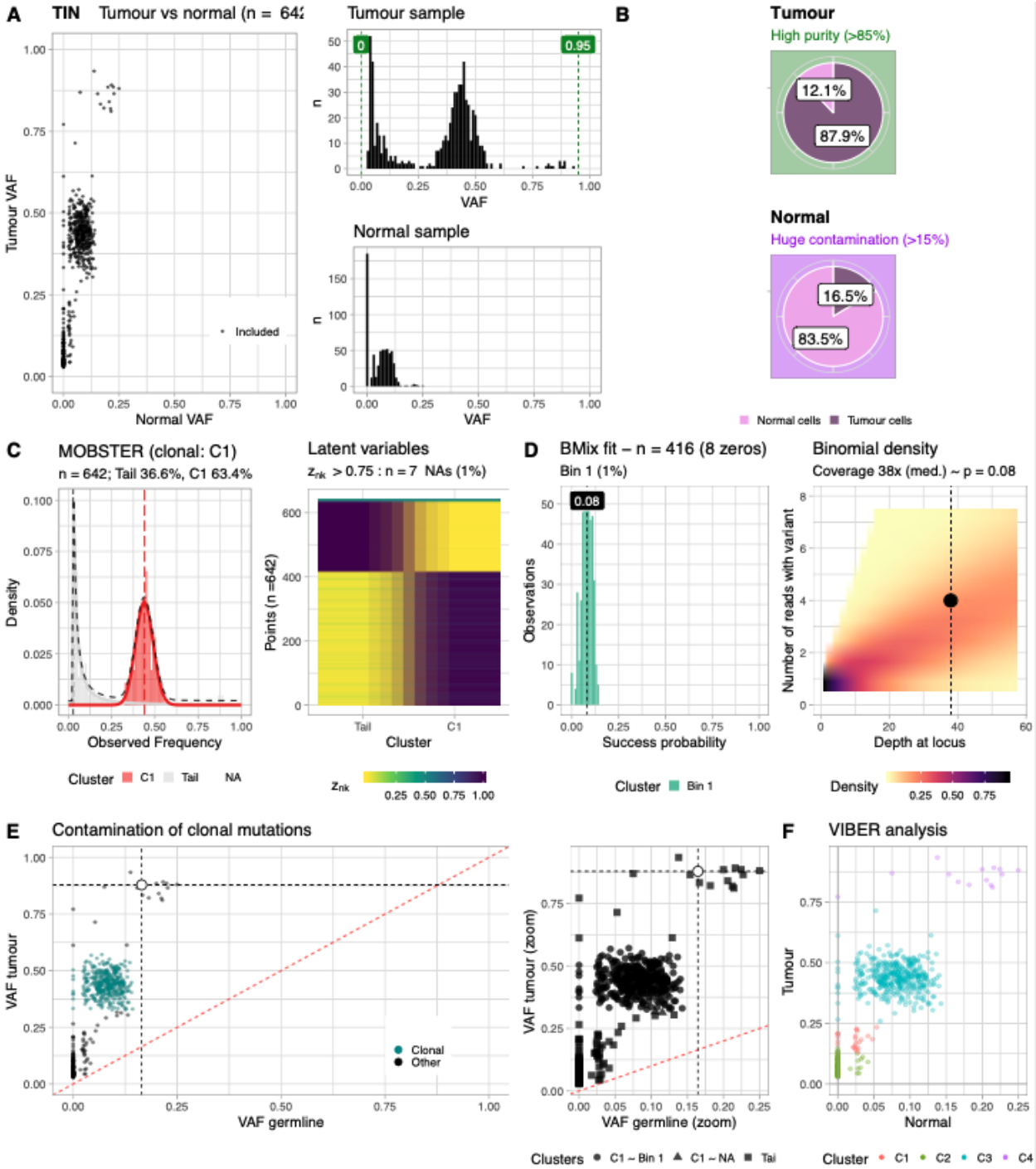
Supplementary Figure S7. Full report generated by TINC analysis of the sample discussed in Figure 6a-f. **a,b.** Data distribution for diploid SNVs, and TIT/TIN scores. **c.** MOBSTER deconvolution and latent variables reporting the probability of each SNVs to be assigned to each one of the detected clusters. **d.** BMix deconvolution and Binomial density at the observed coverage. **e.** Final assignments of clonal status to the input SNVs, with zoom for low-frequency variants. **f.** Complementary VIBER multivariate clustering.



Supplementary Figure S8. Full report generated by TINC analysis of the sample discussed in Figure 6g-i. **a,b.** Data distribution for diploid SNVs, and TIT/TIN scores. **c.** MOBSTER deconvolution and latent variables reporting the probability of each SNVs to be assigned to each one of the detected clusters. **d.** BMix deconvolution and Binomial density at the observed coverage. **e.** Final assignments of clonal status to the input SNVs, with zoom for low-frequency variants. **f.** Complementary VIBER multivariate clustering.



Supplementary Figure S9. Full report generated by TINC analysis of a GEL sample that shows no contamination of the normal sample. **a,b.** Data distribution for diploid SNVs, and TIT/TIN scores. **c.** MOBSTER deconvolution and latent variables reporting the probability of each SNVs to be assigned to each one of the detected clusters. **d.** BMix deconvolution and Binomial density at the observed coverage. **e.** Final assignments of clonal status to the input SNVs, with zoom for low-frequency variants. **f.** Complementary VIBER multivariate clustering.



Supplementary Figure S10. Full report generated by TINC analysis of a GEL sample that shows contamination of the normal sample similar to samples in Figure 6. **a,b.** Data distribution for diploid SNVs, and TIT/TIN scores. **c.** MOBSTER deconvolution and latent variables reporting the probability of each SNVs to be assigned to each one of the detected clusters. **d.** BMix deconvolution and Binomial density at the observed coverage. **e.** Final assignments of clonal status to the input SNVs, with zoom for low-frequency variants. **f.** Complementary VIBER multivariate clustering.

Supplementary Tables

Supplementary Table 1.

Requirements for the collection of samples from haematological malignancies

Clinical Indication	Normal sample source	Tumour sample source
Acute Myeloid Leukaemia	Saliva[a] Cultured fibroblasts Others[b]	Bone marrow aspirate or peripheral blood containing >=20% blasts morphologically *or* any blast percentage if there is an AML-defining genetic abnormality as per WHO 2016 Guidelines
Myelodysplastic Syndrome	Saliva[a] Cultured fibroblasts Others[b]	Bone marrow aspirate or peripheral blood containing >=5% blasts morphologically
Chronic Myeloid Leukaemia ● Extreme 'Good' Responders[c]	Saliva Peripheral Blood[d]	Bone marrow aspirate or peripheral blood (no minimum blast cell percentage required)
Chronic Myeloid Leukaemia ● Extreme 'Poor' Responders[e] ● Additional Cytogenetic Abnormality[f] ● Accelerated or Blast Phase[g]	Cultured fibroblasts Others[b]	Bone marrow aspirate or peripheral blood (no minimum blast cell percentage required)
Unclassified Haematological Malignancies[h]	Saliva[a] Cultured fibroblasts Others[b]	Bone marrow aspirate or peripheral blood (no minimum blast cell percentage required)
Acute Lymphoblastic Leukaemia	Saliva[a] Cultured fibroblasts MRD negative peripheral blood / bone marrow	Bone marrow aspirate or peripheral blood containing >=40% blasts

	aspirate[i]	morphologically
Lymphoproliferative Disorders[j]	Saliva Peripheral blood[k]	Fresh frozen tissue (i.e. biopsy or resection) Bone marrow aspirate or peripheral blood containing >=40% malignant cell nuclei[l] Other liquid sample containing >=40% malignant cell nuclei[m]
Multiple Myeloma	Saliva Peripheral blood[n]	CD138+ sorted cells with a purity of >=40%[o]
Chronic Lymphocytic Leukaemia	Saliva[q]	Bone marrow aspirate or peripheral blood[r] containing >=40% malignant cell nuclei

Table notes

[a] Saliva is acceptable as a normal sample in myeloid malignancies only if sufficient treatment has been given to remove all circulating myeloid cells from the peripheral blood e.g. on day 5 after administration of two doses of anthracycline chemotherapy (or equivalent) in patients receiving intensive induction in Acute Myeloid Leukaemia.

[b] Alternative normal options are being pursued in the disease types indicated to facilitate recruitment to the programme including sorted CD3+ cells (T cells) and uncultured skin biopsies. If and when these normal sample types are acceptable, supplementary guidance will be issued detailing specific requirements.

[c] Extreme 'Good' Responders in Chronic Myeloid Leukaemia are defined as those patients who, after 3 months of treatment with a tyrosine kinase inhibitor, have achieved a *BCR-ABL* transcript level (by RQ-PCR) of <1% using International Standards.

[d] Peripheral blood is an acceptable source of normal DNA for patients who are classified as Chronic Myeloid Leukaemia Extreme 'Good' Responders providing the *BCR-ABL* transcript level (by RQ-PCR) using International Standards is <0.1%.

[e] Extreme 'Poor' Responders in Chronic Myeloid Leukaemia are defined as those patients who, after 3 months of treatment with a tyrosine kinase inhibitor have a *BCR-ABL* transcript level (by

RQ-PCR) of >10% using International Standards.

[f] Refers to any additional cytogenetic abnormality detected using karyotyping in the diagnostic sample in Chronic Myeloid Leukaemia other than a variant *BCR-ABL* transcript.

[g] Patients either presenting in Accelerated or Blast Phase Chronic Myeloid Leukaemia or progressing to Accelerated or Blast Phase Chronic Myeloid Leukaemia are eligible for recruitment.

[h] The definition of this category is broad but includes disorders such as 'Triple negative' Myeloproliferative Neoplasms (defined as no variant detected in *JAK2* exon 12, exon 14 (codon 617), *CALR* exon 9 or *MPL* exon 10), and Myelodysplastic/Myeloproliferative Overlap Syndromes.

[i] In Acute Lymphoblastic Leukaemia peripheral blood or bone marrow aspirate samples which are either negative for or have a diagnostic MRD marker (e.g. BCR or TCR gene rearrangement or *BCR-ABL* transcript) detectable at a level of <0.1% are suitable for use as the source of normal DNA.

[j] Any patient with a Lymphoproliferative Disorder (high or low grade) for which treatment is planned is eligible for recruitment to the project.

[k] Peripheral blood is suitable for use as the source of normal DNA in Lymphoproliferative Disorders providing there are no circulating tumour cells in the peripheral blood. Please note it is not necessary to undertake anything beyond normal standard of care assessments to demonstrate the absence of circulating tumour cells.

[l] Peripheral blood or bone marrow aspirate samples could be used as a source of tumour DNA in Lymphoproliferative Disorders providing the malignant lymphoid cells constitute >=40% of the nucleated cells in the sample.

[m] It is appreciated that there may be situations where malignant lymphoid cells constitute >=40% of the nucleated cells in a sample of a different body fluid e.g. pleural fluid; in these rare situations these would be an acceptable source of tumour DNA.

[n] Peripheral blood is an acceptable source of normal DNA in Myeloma providing there are no circulating plasma cells in the peripheral blood.

[o] All myeloma samples should undergo enrichment for CD138+ cells even if the starting plasma cell percentage of the bone marrow aspirate smear is >=40% in order to obtain the highest possible purity of plasma cells. Laboratories carrying out CD138+ cell enrichment / sorting will need to supply verification of the sorting technique and the CD138+ sorting checklist (Part 5: Appendix D) prior to commencement.

[p] It is appreciated that most myeloma samples will not have sufficient cells for tumour lysate

collection for subsequent RNA extraction.

[q] Saliva collection in Chronic Lymphocytic Leukaemia should be postponed until such a time as the peripheral blood lymphocyte count is <25x109/L.

[r] The lymphocyte count of peripheral blood samples to be used as the source of the tumour DNA in Chronic Lymphocytic Leukaemia should be >25x109/L.