RESEARCH

A sample article title

Jane E Doe[1*†] and John RS Smith[1,2]

*Correspondence:
jane.e.doe@cambridge.co.uk
[1]Department of Zoology, Cambridge, Waterloo Road, London, UK
Full list of author information is available at the end of the article
[†]Equal contributor

Abstract

Background: Germline alterations have clinical implications for cancer patients and their families. Because the tumour genome contains both germline and somatic variants, clinical tumour sequencing presents an opportunity for pre-screening of germline variants. This framework is time- and cost-effective because only patients with potential germline variants are referred to downstream confirmatory testing. A key challenge in this framework is distinguishing between germline and somatic variants in the tumour. Tumour specimens are also commonly formalin-fixed paraffin-embedded (FFPE), which induces DNA damage that interferes with molecular testing. To determine the feasibility of leveraging tumour sequencing for identifying germline variants, the usability of FFPE DNA for germline testing must be evaluated and an approach to differentiate between germline and somatic variants must be established.

Methods:

Results:

Conclusions:

Keywords: Germline variants; Tumour-only sequencing; Formalin-fixed paraffin-embedded tumours

Background

Germline alterations have clinical implications for cancer patients and their families. Inherited predisposition to cancer can be predicted

Tumour sequencing has been rapidly integrated into clinical oncology Because the tumour genome contains both germline and somatic variants, clinical tumour sequencing presents an opportunity for pre-screening of germline variants. This framework is time- and cost-effective because only patients with potential germline variants would be referred to follow-up germline testing. Follow-up germline testing could A key challenge in this framework is distinguishing between germline and somatic variants in the tumour.

Tumour specimens are also commonly formalin-fixed paraffin-embedded (FFPE), which induces DNA damage that interferes with molecular testing.

Genomic analyses of tumours can not only reveal actionable somatic mutations, but also germline variants with clinical implications for patients and their families. Thus, clinical tumour sequencing presents an opportunity to perform initial screening for germline variants. This framework is cost-saving because only patients with potential germline variants would be referred to downstream confirmatory testing.

Methods

Patient samples

Blood and FFPE tumour samples were acquired from 213 patients who provided informed consent for The OncoPanel Pilot (TOP) study (Human Research Ethics Protocol H14-01212), a pilot study to optimize the OncoPanel, which is an amplicon-based targeted NGS panel for solid tumours. The TOP study also assessed the OncoPanel's application for guiding disease management and therapeutic intervention. One blood sample and four FFPE tumours were sequenced in duplicates, which resulted in 217 tumour-normal paired samples (434 sequencing libraries were included in our analyses). Patients in the TOP study were those with advanced cancers including CRC, lung cancer, melanoma, gastrointestinal stromal tumour (GIST), and other cancers (Table 1). The age of paraffin block for tumour samples ranged from 18 to 5356 days with a median of 274 days.

Sample preparation, library construction, and Illumina sequencing

Genomic DNA was extracted from blood and FFPE tumour samples using the Gentra Autopure LS DNA preparation platform and QIAamp DNA FFPE tissue kit (Qiagen, Hilden, Germany), respectively. The extracted DNA was sheared according to a previously described protocol [**?**] to obtain approximate sizes of 3 Kb followed by PCR primer merging, amplification of target regions, and adapter ligation using the Thunderstorm NGS Targeted Enrichment System (RainDance Technologies, Lexington, MA) as per manufacturer's protocol. Barcoded amplicons were sequenced with the Illumina MiSeq system for paired end sequencing with a v2 250-bp kit (Illumina, San Diego, CA).

OncoPanel (Amplicon-based targeted sequencing panel for solid tumours)

The OncoPanel assesses coding exons and clinically relevant hotspots of 15 cancer-related genes and six PGx genes. Germline alterations in the six PGx genes could serve as predictors of susceptibility to chemotherapy-induced toxicity. Primers were designed by RainDance Technologies (Lexington, MA) using the GRCh37/hg19 human reference genome to generate 416 amplicons between 56 bp and 288 bp in size, which interrogate $\sim$ 20 Kb of target bases. Complete list of genes and gene reference models for the OncoPanel as well as target regions and amplicons are presented in Supplemental Materials.

Variant calling pipeline

Read alignment and variant calling

Reads that passed the Illumina Chastity filter were aligned to the hg19 human reference genome using the BWA mem algorithm (version 0.5.9) with default parameters, and the alignments were processed and converted to the BAM format using SAMtools (version 0.1.18). The SAMtools `mpileup` function (`samtools mpileup -BA -d 500000 -L 500000 -q 1`) was used to generate pileup files for all target bases followed by variant calling with the VarScan2 `mpileup2cns` (version 2.3.6) function with parameter thresholds of VAF $\geq$ 0.1 and Phred-scaled BAQ score $\geq$ 20 (`--min-var-freq 0.1 --min-avg-qual 20 --strand-filter 0 --p-value 0.01 --output-vcf --variants`).

Identifying patients with homozygous minor alleles that are present the hg19 reference genome

Four genomic positions at which the hg19 human reference genome contained the minor alleles were identified (Table 3). Hence, patients homozygous for these four minor alleles would not be identified by our standard variant calling procedure. For these four genomic sites, our method for variant calling was modified to provide calls for every patient in the cohort. The VarScan2 `mpileup2cns` function was used with parameter thresholds of VAF $\geq$ 0.25, VAF to call homozygote $\geq$ 0.9, BAQ score $\geq$ 20, and fraction of variant reads from each strand $\geq$ 0.1 `(--min-var-freq 0.25 --min-freq-for-hom 0.9 --min-avg-qual 20 --strand-filter 1 --p-value 0.01 --output-vcf)`. Next, allelic statuses were re-assigned, in which wild type calls were re-assigned as homozygous variants, while homozygous variants were re-assigned as wild type calls. Corrections to the VAFs of these four genomic sites were also made to ensure that the VAFs reflected percentage of reads with the minor alleles.

Variant filtering

Variant calls were filtered using the VarScan2 `fpfilter` function with fraction of variant reads from each strand $\geq$ 0.1 and default thresholds for other parameters (Table 4). The VarScan2 `fpfilter` removed 247 low quality variants. Seventy germline variants in the blood were also excluded from our analysis because these variants in the tumours were filtered by the VarScan2 `fpfilter`. There were also 16 risk allele calls in tumour samples that did not pass the strand filter, causing the removal of 10 risk allele calls in the blood samples from our evaluation. Overall, a total of 343 calls were excluded by the VarScan2 `fpfilter` and strand filter. Manual inspection was performed for a subset of variants, including variants detected within primer regions and in PGx genes, using the Integrative Genomics Viewer (IGV, version 2.3). This resulted in the removal of 500 spurious calls, which stemmed from software bugs, sequencing artifacts, primer masking, and primer artifacts (Table 5). Eleven low coverage calls ($\leq$ 100x) were also excluded from our analysis. Implementation of this filtering pipeline reduced the raw variant output of 5288 calls from 217 paired tumour-blood samples (434 sequencing libraries) to 4434 calls (Figure 1B).

Variant annotation and interpretation

SnpEff (version 4.2) was used for effect prediction, and the SnpSift package in SnpEff was used to annotate variants with databases such as dbSNP (b138), COSMIC (version 70), 1000 Genomes Project, and ExAC (release 0.3) for interpretation. Clinical significance reported by the ClinVar database and literature review were also used for variant interpretation.

Sequence analysis

A custom Python script was used to process BAM files to quantify the number of on-target aligned (reads that map to target regions), off-target aligned (reads that map to hg19 but not target regions), and unaligned reads with a Phred-scaled mapping quality (MAPQ) score $\geq$ 10. Unaligned reads were also screened against microbial

sequences, including viruses, archaea, bacteria, and fungi, to ensure that samples did not contain significant amount of microbial contaminants. Coverage depth for target bases with MAPQ $\geq$ 1 and BAQ $\geq$ 20 were obtained using bam-readcount (https://github.com/genome/bam-readcount). To measure coverage depth of amplicons, the SAMtools `view` function was used to filter for reads with MAPQ $\geq$ 1 `(samtools view -b -q 1)` followed by the bedtools `intersect` function (version 2.25.0) to quantify the number of reads that overlap with amplicon positions `(intersect -a $AMPLICON_POSITIONS -b $BAM_FILE -f 0.85 -r -c)`.

Per-base metrics generated using bam-readcount were also used for assessment of sequence artifacts. A custom R script was used to count and categorize the different groups of base changes (i.e. C>T/G>A, A>G/T>C, C>A/G>T, A>C/T>G, C>G/G>C, and A>T/T>A). Unless stated otherwise, analysis of sequence artifacts excluded true variants identified by our VarScan2 variant calling pipeline and base changes with VAF $< 1\%$, which were considered sequencing errors. All statistical analyses and data visualization were performed using the R statistical software package (version 3.3.2) and associated open source packages.

Application of VAF thresholds to separate germline alterations from somatic mutations

Variants in the tumours that passed our filtering criteria were subjected to VAF thresholds between 10–45%. At each VAF cut-off, variants that were not filtered out were considered predicted germline variants. Given that all tumour samples have matched blood samples, true positives were identified as predicted germline variants that overlap with variants in the blood. Conversely, false negatives were identified as variants that were filtered out by the VAF cut-off (predicted as somatic), but were present in the blood samples. Sensitivity at each VAF threshold was calculated by dividing the number of true positives with the sum of true positives and false negatives. Because predicted germline variants would be referred to follow-up germline testing, positive predictive values (PPVs) were calculated at each VAF cut-off to evaluate precision of our approach. False positives were identified as predicted germline variants that were absent in the blood, and PPV was calculated by dividing the number of true positives with the sum of true positives and false positives.

Results

Comparison of efficiency in amplicon enrichment and sequencing results between blood and FFPE specimens

To determine whether FFPE DNA is viable for identifying germline variants, we characterized the effect of formalin-induced DNA damage on efficiency in amplicon enrichment and sequencing metrics. We compared these outcomes to DNA isolated from blood, which is a gold standard for germline testing. Efficiency in amplicon enrichment was measured as the $\log_2$ fold change between DNA input and amplicon yield ($\log_2$ (Amplicon Yield/DNA Input)) and compared between FFPE specimens to blood. There was a significant decrease in enrichment efficiency in FFPE specimens compared to blood (Wilcoxon signed-rank test, $W = 24754$, $Z = 12.7$, $p = 4.6 \times 10^{-57}$, $r = 0.61$; **??**). The median $\log_2$ (Amplicon Yield/DNA Input) for blood

was 1.04, whereas the median $\log_2$ (Amplicon Yield/DNA Input) for FFPE specimens was -0.332. This demonstrates that production of amplicons was less efficient in FFPE specimens compared to blood.

We also compared read alignments between FFPE specimens and blood.

Reduced coverage depth in FFPE specimens is more pronounced for longer amplicons
Deamination effects lead to increased C>T/G>A transitions in FFPE specimens
Increased age of paraffin block results in reduced amplicon yield and elevated level of C>T/G>A sequence artifacts
Frequency and interpretation of germline alterations in patients from TOP cohort
The majority of germline alterations in the blood are retained in the tumour
Application of VAF thresholds to separate germline alterations from somatic mutations in tumour-only analyses

Discussion

Thus we observe that this expected value is finite for all $v > 0$ (also see [1, 2, 3, 4, 5]).

Competing interests
The authors declare that they have no competing interests.

Author's contributions
Text for this section . . .

Acknowledgements
Text for this section . . .

Author details
[1]Department of Zoology, Cambridge, Waterloo Road, London, UK. [2]Marine Ecology Department, Institute of Marine Sciences Kiel, Düsternbrooker Weg 20, 24105 Kiel, Germany.

References

1. Koonin, E.V., Altschul, S.F., Bork, P.: Brca1 protein products: functional motifs. Nat Genet **13**, 266–267 (1996)
2. Kharitonov, S.A., Barnes, P.J.: Clinical Aspects of Exhaled Nitric Oxide. in press
3. Zvaifler, N.J., Burger, J.A., Marinova-Mutafchieva, L., Taylor, P., Maini, R.N.: Mesenchymal cells, stromal derived factor-1 and rheumatoid arthritis [abstract]. Arthritis Rheum **42**, 250 (1999)
4. Jones, X.: Zeolites and synthetic mechanisms. In: Smith, Y. (ed.) Proceedings of the First National Conference on Porous Sieves: 27-30 June 1996; Baltimore, pp. 16–27 (1996). Stoneham: Butterworth-Heinemann
5. Margulis, L.: Origin of Eukaryotic Cells. Yale University Press, New Haven (1970)
6. Orengo, C.A., Bray, J.E., Hubbard, T., LoConte, L., Sillitoe, I.: Analysis and assessment of ab initio three-dimensional prediction, secondary structure, and contacts prediction. Proteins **Suppl 3**, 149–170 (1999)
7. Schnepf, E.: From prey via endosymbiont to plastids: comparative studies in dinoflagellates. In: Lewin, R.A. (ed.) Origins of Plastids vol. 2, 2nd edn., pp. 53–76. Chapman and Hall, New York (1993)
8. Innovative Oncology
9. Smith, Y. (ed.): Proceedings of the First National Conference on Porous Sieves: 27-30 June 1996; Baltimore. Butterworth-Heinemann, Stoneham (1996)
10. Hunninghake, G.W., Gadek, J.E.: The alveloar macrophage. In: Harris, T.J.R. (ed.) Cultured Human Cells and Tissues, pp. 54–56. Academic Press, New York (1995). Stoner G (Series Editor): Methods and Perspectives in Cell Biology, vol 1
11. Advisory Committee on Genetic Modification: Annual Report. London (1999). Advisory Committee on Genetic Modification
12. Kohavi, R.: Wrappers for performance enhancement and obvious decision graphs. PhD thesis, Stanford University, Computer Science Department (1995)
13. The Mouse Tumor Biology Database. http://tumor.informatics.jax.org/cancer_links.html

Figures

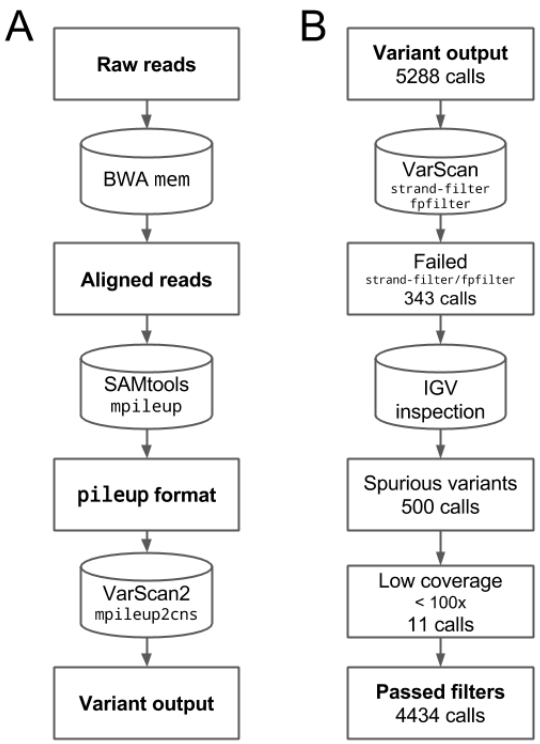

Figure 1 Pipelines for (A) variant calling and (B) filtering.

Figure 2 Sample figure title. Figure legend text.

Tables

Table 1 Distribution of cancer types in the TOP cohort.

Cancer Type	Number of Cases	Percentage (%)
Colorectal	97	46
Lung	60	28
Melanoma	18	8
Other†	16	8
GIST	7	3
Sarcoma	4	2
Neuroendocrine	4	2
Cervical	2	0.9
Ovarian	2	0.9
Breast	2	0.9
Unknown	1	0.5

†This category includes thyroid, peritoneum, Fallopian tube, gastric, endometrial, squamous cell carcinoma, anal, salivary gland, peritoneal epithelial mesothelioma, adenoid cystic carcinoma, pancreas, breast, gall bladder, parotid epithelial myoepithelial carcinoma, carcinoid, and small bowel cancers.

Table 2 Gene reference models for HGVS nomenclature of OncoPanel genes.

Gene	Protein	Reference Model
Cancer-related		
AKT1	Protein kinase B	NM_001014431.1
ALK	Anaplastic lymphoma receptor tyrosine kinase	NM_004304.3
BRAF	Serine/threonine-protein kinase B-Raf	NM_004333.4
EGFR	Epidermal growth factor receptor	NM_005228.3
HRAS	GTPase HRas	NM_005343.2
MAPK1	Mitogen-activated protein kinase 1	NM_002745.4
MAP2K1	Mitogen-activated protein kinase kinase 1	NM_002755.3
MTOR	Serine/threonine-protein kinase mTOR	NM_004958.3
NRAS	Neuroblastoma RAS viral oncogene homolog	NM_002524.3
PDGFRA	Platelet-derived growth factor receptor alpha	NM_006206.4
PIK3CA	Phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit alpha	NM_006218.2
PTEN	Phosphatase and tensin homolog	NM_000314.4
STAT1	Signal transducer and activator of transcription 1	NM_007315.3
STAT3	Signal transducer and activator of transcription 3	NM_139276.2
TP53	Tumor protein P53	NM_000546.5
Pharmacogenomic-related		
DPYD	Dihydropyrimidine dehydrogenase	NM_000110.3
GSTP1	Glutathione S-rransferase pi 1	NM_000852.3
MTHFR	Methylenetetrahydrofolate reductase	NM_005957.4
TYMP	Thymidine phosphorylase	NM_001113755.2
TYMS	Thymidylate synthetase	NM_001071.2
UGT1A1	Uridine diphosphate (UDP)-glucuronosyl transferase 1A1	NM_000463.2

Table 3: Potential minor alleles in the hg19 human reference genome within the target regions of the OncoPanel. (Supplementary)

Gene	Chr	Pos	Minor Allele	dbSNP ID	HGVS*
DPYD	chr1	98348885	C	rs1801265	p.Cys29Arg c.85T>C
MTOR	chr1	11205058	G	rs386514433;	p.Ala1577Ala
				rs1057079	c.4731A>G
	chr1	11288758	C	rs1064261	p.Asn999Asn
					c.2997T>C
TP53	chr17	7579472	C	rs1042522	p.Arg72Pro c.215G>C

*Description of sequence variants according to the HGVS recommendations.

Table 4 Thresholds for parameters of VarScan2 `fpfilter` used for filtering raw variant output. (Supplementary)

Parameter	Description	Threshold
`--min-var-count`	Min number of var-supporting reads	4
`--min-var-count-lc`	Min number of var-supporting reads when depth below somaticPdepth	2
`--min-var-freq`	Min variant allele frequency	0.1
`--max-somatic-p`	Max somatic p-value	0.05
`--max-somatic-p-depth`	Depth required to test max somatic p-value	10
`--min-ref-readpos`	Min average read position of ref-supporting reads	0.1
`--min-var-readpos`	Min average read position of var-supporting reads	0.1
`--min-ref-dist3`	Min average distance to effective 3' end of ref reads	0.1
`--min-var-dist3`	Min average distance to effective 3' end of variant reads	0.1
`--min-strandedness`	Min fraction of variant reads from each strand	0.1
`--min-strand-reads`	Min allele depth required to perform the strand tests	5
`--min-ref-basequal`	Min average base quality for ref allele	15
`--min-var-basequal`	Min average base quality for var allele	15
`--min-ref-avgrl`	Min average trimmed read length for ref allele	90
`--min-var-avgrl`	Min average trimmed read length for var allele	90
`--max-rl-diff`	Max average relative read length difference (ref - var)	0.25
`--max-ref-mmqs`	Max mismatch quality sum of ref-supporting reads	100
`--max-var-mmqs`	Max mismatch quality sum of var-supporting reads	100
`--max-mmqs-diff`	Max average mismatch quality sum (var - ref)	50
`--min-ref-mapqual`	Min average mapping quality for ref allele	15
`--min-var-mapqual`	Min average mapping quality for var allele	15
`--max-mapqual-diff`	Max average mapping quality (ref - var)	50

Table 5: Spurious variants removed by the variant filtering pipeline. (Supplementary)

Gene	Chr	Pos	Ref	Alt	Reason
KIT	chr4	55599268	C	T	Variant masked by primer in FFPE specimen
MAPK1	chr22	22162126	A	G	Variant masked by primer in FFPE specimen
MTOR	chr1	11186783	G	A	Sequencing artifact within primer region
MTOR	chr1	11190646	G	A	Variant masked by primer in FFPE specimen
TYMP	chr22	50964446	A	T	Poor target region, alignment of different sized amplicons
TYMP	chr22	50964862	A	T	Poor target region, alignment of different sized amplicons
TYMS	chr18	673449	G	C	VarScan2 bug after chr18:673443 c.*447_*452delTTAAAG
UGT1A1	chr2	234668879	CAT	C	Sequencing artifact at AT repeats in promoter
UGT1A1	chr2	234668881	T	TAC	VarScan2 bug after AT insertion in promoter

Table 6 Multiple linear regression to predict $\log_2$ fold change between amplicon coverage depth in blood and FFPE specimens ($\log_2$ (Median $\text{Coverage}_{\text{FFPE}}$/Median $\text{Coverage}_{\text{Blood}}$)) based on amplicon length and GC content.

Variable	Unstandardized Coefficient	Standard Error	Standardized Coefficient	*p*-value
Length (bp)	-6.97×10^{-3}	2.59×10^{-4}	-7.56×10^{-1}	7.45×10^{-93}
GC Content (%)	-1.03×10^{-2}	1.01×10^{-3}	-2.88×10^{-1}	4.71×10^{-22}
			Intercept = 1.63, Adjusted R^2 = 0.673	
			$F(2, 413) = 427.6$, *p*-value $= 2.41 \times 10^{-101}$	

Additional Files

Additional file 1 — Sample additional file title
Additional file descriptions text (including details of how to view the file, if it is in a non-standard format or the file extension). This might refer to a multi-page table or a figure.

Additional file 2 — Sample additional file title
Additional file descriptions text.