RESEARCH

True Positive Rate of Germline Variant Calling in Formalin-fixed Paraffin-embedded Tumours

Shyong Quin Yap[1] and Aly Karsan[1,2*]

*Correspondence:
akarsan@bcgsc.ca
[1]British Columbia Cancer Research Centre, 675 West 10th Ave, V5Z 1L3 Vancouver, BC, Canada
Full list of author information is available at the end of the article

Abstract

Background: The tumour genome contains germline information that may have clinical implications for patients and their families. Therefore, clinical laboratories must be equipped to confirm potential germline variants identified by routine tumour sequencing through downstream germline testing. The most common source of DNA in clinical laboratories is formalin-fixed paraffin-embedded (FFPE) tissues. However, fixation with formalin is known to cause DNA fragmentation and sequence artifacts that impose technical challenges in DNA sequencing.

Using amplicon sequencing data from FFPE tumours with matched blood specimens, we compared efficiency of amplicon generation and sequencing results between specimen types to characterize formalin-induced DNA damages. We determined the true positive rate of detecting germline variants in FFPE tumours by assessing the concordance of germline variants between blood and FFPE specimens. We also compared sensitivity of germline variant calling in FFPE tumours to blood at various variant allele frequency thresholds and established a cut-off to achieve high positive predictive value for referral of potential germline variants for confirmatory germline testing without compromising sensitivity of detection.

Results:

Conclusions:

Keywords: Germline variants; Formalin-fixed paraffin-embedded tumours; Amplicon-based targeted next-generation sequencing

Background

The application of next-generation sequencing (NGS) technologies for tumour profiling has been increasingly integrated into oncologic care to detect targetable somatic mutations and personalize treatments for cancer patients. Although analysis of tumour-normal paired samples is required to accurately discriminate between somatic and germline variants, most clinical laboratories only sequence tumour samples to minimize cost and turnaround time [1]. However, genomic analyses of tumours can also reveal secondary genomic findings, which are germline information that may have clinical implications for patients and their family members [1]. In fact, several studies demonstrated that a germline cancer-predisposing variant is present in 3-10% of patients undergoing tumour-normal sequencing [1, 2, 3, 4]. Therefore, clinical laboratories providing tumour genomic testing must be equipped to perform germline confirmatory testing on potential germline variants or be prepared to refer such cases to external services.

Because the tumour genome contains germline information, clinical laboratories can leverage tumour genomic testing to perform initial screening for clinically relevant germline variants such as variants in pharmacogenomic (PGx) genes. Subsequently, a similar framework for validating secondary germline findings can be applied, in which only patients with potential germline PGx variants are subjected to downstream germline testing. This procedure for germline PGx testing is more cost-effective because it does not require processing, sequencing, and analysis of normal DNA for every patient. The ability to implement germline PGx testing at a reduced cost can significantly benefit patient care because these variants cause functional changes in drug targets and drug disposition proteins (proteins involved in drug metabolism and transport), thereby contributing to inter-patient differences in chemotherapeutic response [5]. Hence, such genomic information can be used to guide the selection of chemotherapeutic drugs and optimization of drug dosage for cancer patients, leading to improved safety and efficacy of treatment and reduced risk of toxicity [5].

Detection of genomic alterations in tumour DNA is also faced with technical challenges conferred by formalin-fixed paraffin-embedded (FFPE) tumour specimens [6, 7]. Tumour biopsies are often formalin-fixed to preserve tissue morphology for histological examination and to enable storage at room temperature; however, formalin fixation causes DNA fragmentation and base modifications, which pose difficulties in using DNA extracted from FFPE tumours for clinical genomic testing [6, 7]. Fragmentation damage caused by formalin fixation leads to reduced template DNA for PCR amplification, thereby affecting the efficiency of amplicon-based NGS testing [6, 7]. Furthermore, the degree of DNA fragmentation was shown to be higher in tissues from older FFPE blocks and tissues fixed with formalin of lower pH [6]. Formalin fixation is also problematic because it gives rise to depurination, which generates abasic sites, and cytosine deamination resulting in C>T/G>A transitions [6]. These forms of formalin-induced DNA damage contribute to the presence of sequence artifacts in FFPE specimens, which can be inaccurately identified as real genomic alterations.

In this study, we assessed the concordance of germline PGx variants between tumour and matched normal DNA by analyzing amplicon-based targeted NGS data from 213 patients with tumour-normal paired samples. While matched normal DNA was derived from peripheral blood, tumour DNA was extracted from FFPE tumour blocks; thus, we compared the quality metrics of sequencing data between tumour and blood specimens and evaluated the prevalence of formalin-induced DNA damages to address the impact of formalin fixation on amplicon-based NGS testing. We demonstrated that germline PGx variants can be identified with high sensitivity and precision in FFPE tumour DNA using a clinical targeted sequencing panel.

Methods

Patient Samples

Blood and FFPE tumour specimens were acquired from 213 patients who provided informed consent for The OncoPanel Pilot (TOP) study, a pilot study to optimize the OncoPanel, which is an amplicon-based targeted NGS panel for solid tumours, and assess its application for guiding disease management and therapeutic intervention. Patients in the TOP study are those with advanced cancers including colorectal

cancer, lung cancer, melanoma, gastrointestinal stromal tumour (GIST), and other cancers (Table 1). The age of paraffin block for tumour specimens ranges from 18 to 5356 days with a median of 274 days.

Table 1 Distribution of cancer types in the TOP cohort.

Cancer Type	Number of Cases	Percentage (%)
Colorectal	97	46
Lung	59	28
Melanoma	18	8
Other*	17	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, lung sarcomatoid carcinoma, Fallopian tube, gastric, endometrial, squamous cell carcinoma, anal, salivary gland, peritoneal epithelial mesothelioma, adenoid cystic carcinoma, pancreas, breast, gall bladder, parotid epithelial myoepithelial carcinoma, and small bowel cancers.

OncoPanel (Solid Tumour NGS Panel)

The OncoPanel is offered by the British Columbia Cancer Agency (BCCA) for clinical genomic testing of coding exons and clinically relevant hotpots of 20 cancer-related genes and six PGx genes: *DPYD*, *GSTP1*, *MTHFR*, *TYMP*, *TYMS*, and *UGT1A1*. Full list of genes and gene reference models for the OncoPanel is presented in Table 8. Primers were designed by RainDance Technologies (Lexington, MA) using the GRCh37/hg19 reference sequence to generate 416 amplicons between 100 bp and 250 bp in size, which interrogate ~ 20 kb of target bases. Target regions of the six PGx genes in the OncoPanel were assayed by 49 out of 416 amplicons. Complete lists of primers and amplicons are provided in the Supplemental Materials.

Sample Preparation, Library Construction, and Illumina Sequencing

Genomic DNA was extracted from blood and FFPE tumour specimens 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 [8] to attain 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).

Variant Calling Pipeline

Reads that passed the Illumina Chastity filter were aligned to the GRCh37/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). Variant calling was performed with the SAMtools `mpileup` function (`samtools mpileup -BA -d 500000 -L 500000 -q 1`) to generate pileup files for all target bases followed by the VarScan2 `mpileup2cns` (version 2.3.6) function with parameter thresholds of variant allele frequency $\geq 10\%$

and Phred-scaled base quality score ≥20 (`--min-var-freq 0.1 --p-value 0.01 --strand-filter 0 --output-vcf --variants --min-avg-qual 20`). Variant calls were filtered using the VarScan2 `fpfilter` function with fraction of variant reads from each strand ≥ 0.1 and default thresholds for other parameters. SnpEff (version 4.2) was used for variant annotation and effect prediction whereas the SnpSift package in SnpEff was used to annotate variants with databases such as dbSNP (b138), COSMIC (version 70), 1000 Genomes Project, ClinVar, and ExAC (release 0.3) for interpretation.

Data Analysis
More details . . .

Results

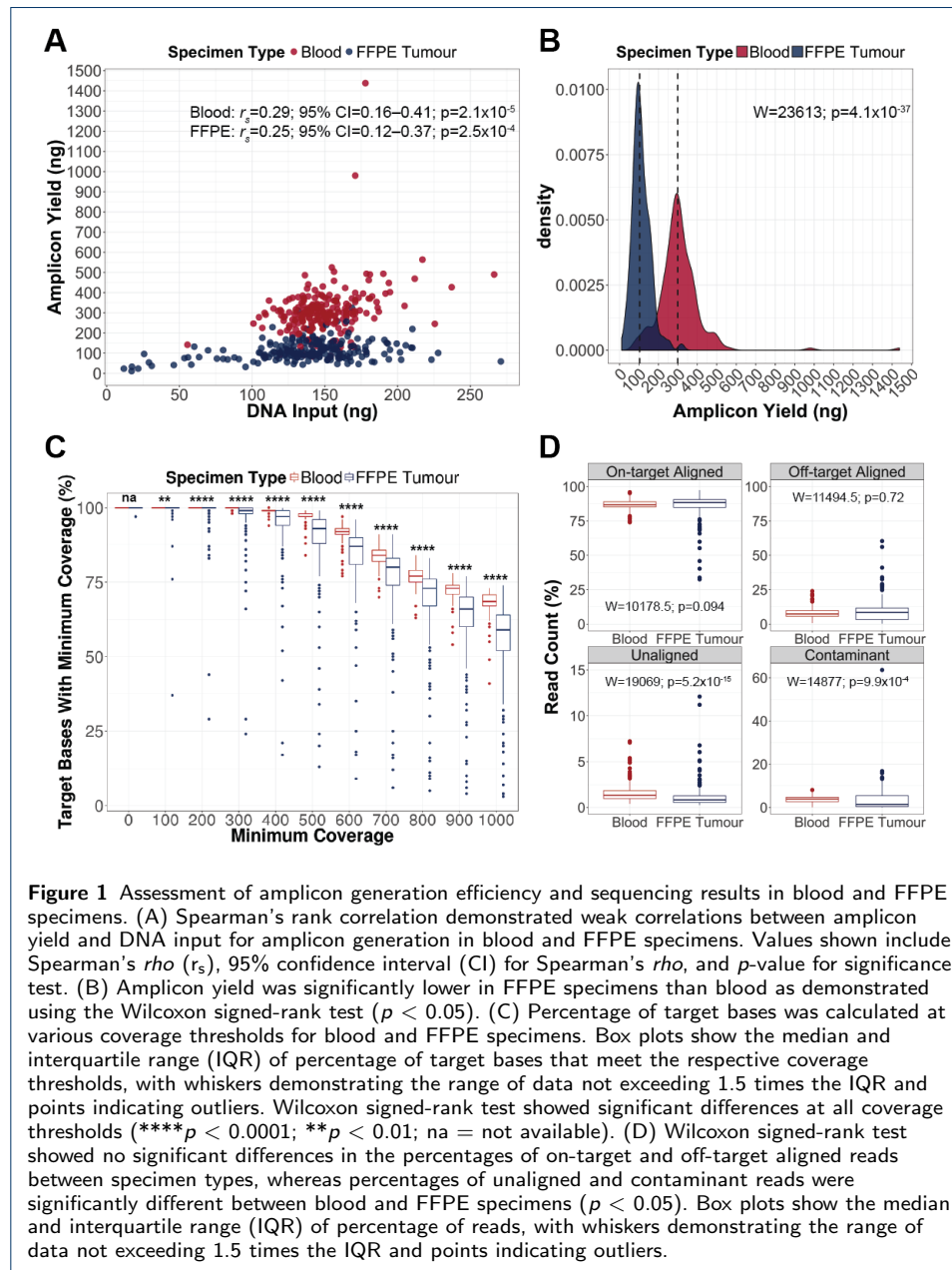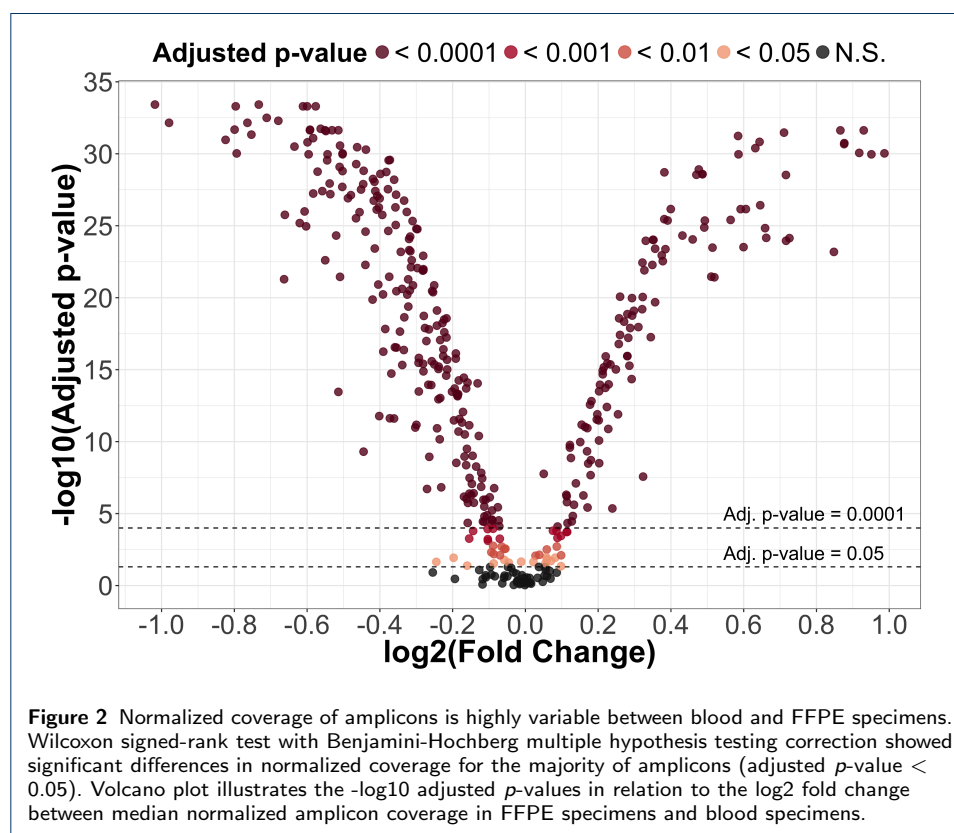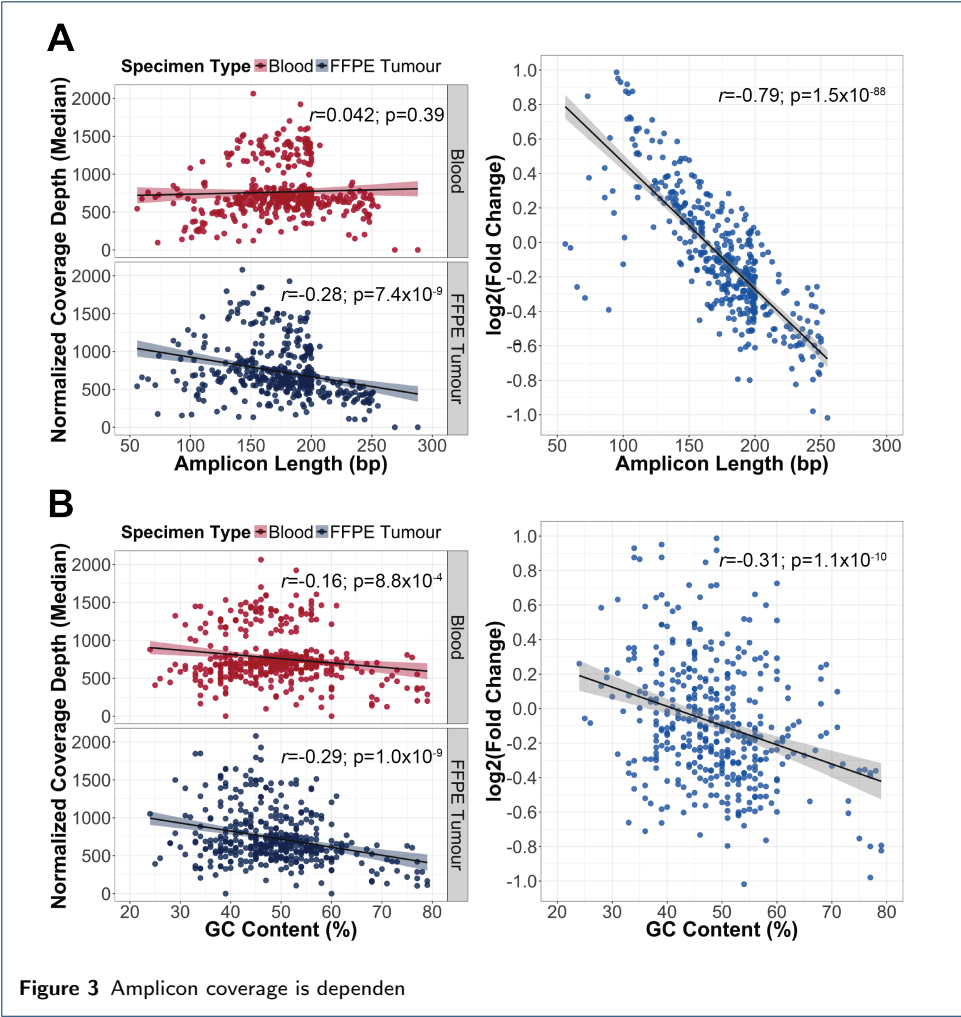Comparison of amplicon generation and sequencing metrics between blood and FFPE specimens

Figure 1 Assessment of amplicon generation efficiency and sequencing results in blood and FFPE specimens. (A) Spearman's rank correlation demonstrated weak correlations between amplicon yield and DNA input for amplicon generation in blood and FFPE specimens. Values shown include Spearman's *rho* (r_s), 95% confidence interval (CI) for Spearman's *rho*, and *p*-value for significance test. (B) Amplicon yield was significantly lower in FFPE specimens than blood as demonstrated using the Wilcoxon signed-rank test ($p < 0.05$). (C) Percentage of target bases was calculated at various coverage thresholds for blood and FFPE specimens. Box plots show the median and interquartile range (IQR) of percentage of target bases that meet the respective coverage thresholds, with whiskers demonstrating the range of data not exceeding 1.5 times the IQR and points indicating outliers. Wilcoxon signed-rank test showed significant differences at all coverage thresholds (****$p < 0.0001$; **$p < 0.01$; na = not available). (D) Wilcoxon signed-rank test showed no significant differences in the percentages of on-target and off-target aligned reads between specimen types, whereas percentages of unaligned and contaminant reads were significantly different between blood and FFPE specimens ($p < 0.05$). Box plots show the median and interquartile range (IQR) of percentage of reads, with whiskers demonstrating the range of data not exceeding 1.5 times the IQR and points indicating outliers.

Table 2 Comparison of amplicon generation and sequencing results between specimen types using Wilcoxon signed-rank test.

Parameter	Blood Median	Range	FFPE Tumour Median	Range	p value
DNA Input (ng)	147.8	55.5–266.4	140.9	11.8–271.0	3.5×10^{-4}
Amplicon Yield (ng)	299.2	84.0–1438.0	103.6	11.6–325.5	4.1×10^{-37}
Average Per Base Normalized Coverage	1270	950–1519	1194	283–1405	1.7×10^{-22}
$\geq$ 0x Target Bases (%)	100.0	100.0-100.0	100.0	97.0-100.0	–
$\geq$ 100x Target Bases (%)	100.0	100.0-100.0	100.0	37.0–100.0	1.5×10^{-3}
$\geq$ 200x Target Bases (%)	100.0	100.0-100.0	100.0	29.0–100.0	1.4×10^{-7}
$\geq$ 300x Target Bases (%)	100.0	98.0-100.0	99.0	24.0–100.0	2.3×10^{-16}
$\geq$ 400x Target Bases (%)	99.0	94.0-100.0	97.0	17.0–100.0	2.0×10^{-23}
$\geq$ 500x Target Bases (%)	97.0	84.0-99.0	89.5	13.0–99.0	5.1×10^{-29}
$\geq$ 600x Target Bases (%)	92.0	77.0-97.0	87.0	9.0–96.0	3.0×10^{-26}
$\geq$ 700x Target Bases (%)	84.0	70.0-91.0	80.0	6.0–91.0	2.3×10^{-21}
$\geq$ 800x Target Bases (%)	77.0	63.0-84.0	73.0	5.0–83.0	6.0×10^{-23}
$\geq$ 900x Target Bases (%)	73.0	54.0-78.0	66.0	4.0–77.0	9.8×10^{-30}
$\geq$ 1000x Target Bases (%)	68.5	41.0-73.0	59.0	3.0-74.0	1.0×10^{-30}
On-target Aligned Reads (%)	86.8	74.0–95.9	88.4	32.5–97.4	9.4×10^{-2}
Off-target Aligned Reads (%)	7.3	0.8-24.0	8.4	0.4–60.4	0.7
Unaligned Reads (%)	1.3	0.4–7.2	0.8	0.3–12.1	5.2×10^{-15}
Contaminant (%)	3.9	0.1–8.1	1.4	0.03–63.6	9.9×10^{-4}

Reduced amplicon coverage in FFPE specimens is more profound for increased amplicon length

Figure 2 Normalized coverage of amplicons is highly variable between blood and FFPE specimens. Wilcoxon signed-rank test with Benjamini-Hochberg multiple hypothesis testing correction showed significant differences in normalized coverage for the majority of amplicons (adjusted *p*-value < 0.05). Volcano plot illustrates the –log10 adjusted *p*-values in relation to the log2 fold change between median normalized amplicon coverage in FFPE specimens and blood specimens.

Figure 3 Amplicon coverage is dependen

Table 3

Variable	Unstandardized Coefficient	Standard Error	Standardized Coefficient	p-value
Length (bp)	-7.24×10^{-3}	2.54×10^{-4}	-7.75×10^{-1}	2.47×10^{-99}
GC Content (%)	-9.92×10^{-3}	9.77×10^{-4}	-2.77×10^{-1}	8.70×10^{-22}
			Intercept = 1.66, Adjusted R^2 = 0.695	
			$F(2, 411) = 471$, p-value $= 4.65 \times 10^{-107}$	

Deamination effects lead to increased C>T/G>A transitions at low allele frequency in FFPE specimens
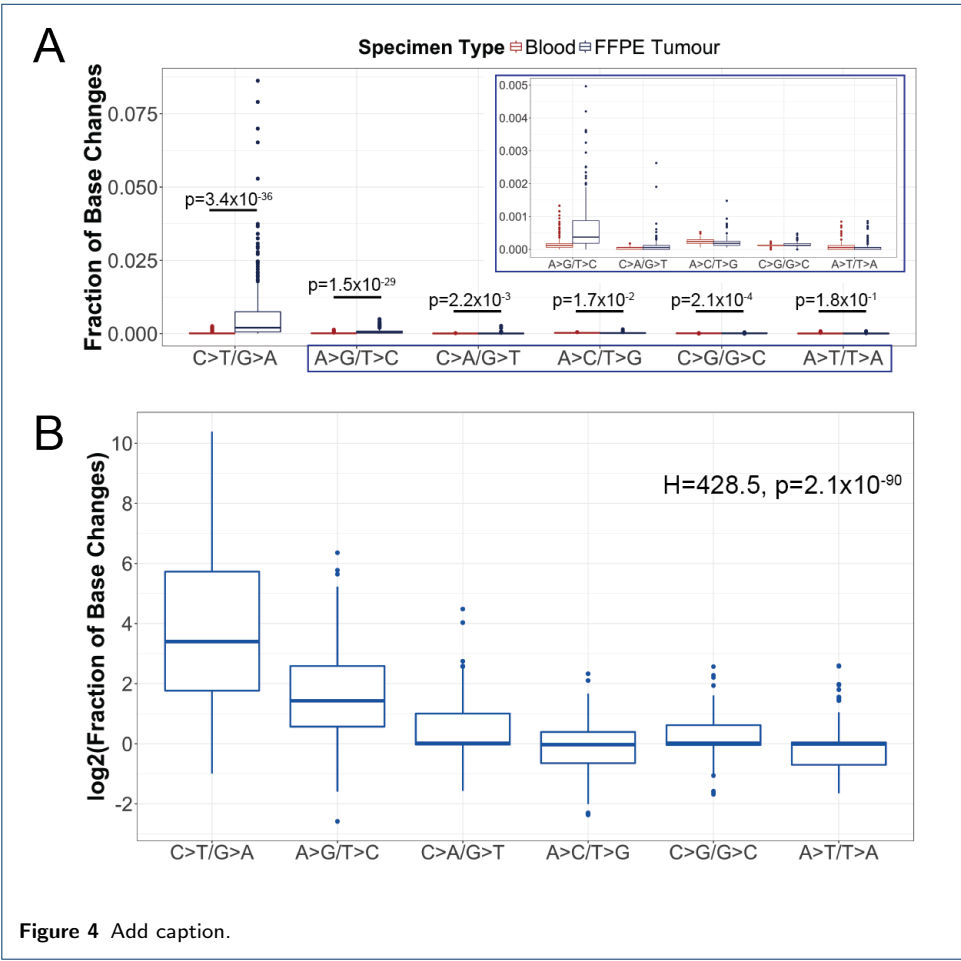
Figure 4 Add caption.

Table 4 Multiple pairwise comparison of log2 fold change of fraction of base changes in FFPE specimen over blood using Dunn's test with Benjamini-Hochberg multiple hypothesis testing correction. Top values represent Dunn's pairwise z statistics, whereas bottom values represent adjusted p-value. Asterisk(*) indicates significance level of adjusted p-value < 0.05.

Type of Base Changes	A>C/T>G	A>G/T>C	A>T/T>A	C>A/G>T	C>G/G>C
A>G/T>C	-11.7 $4.15 \times 10^{-31}*$				
A>T/T>A	-0.399 3.45×10^{-1}	9.57 $1.31 \times 10^{-21}*$			
C>A/G>T	-3.46 $4.00 \times 10^{-4}*$	6.39 $1.52 \times 10^{-10}*$	-2.73 $3.99 \times 10^{-3}*$		
C>G/G>C	-3.02 $1.73 \times 10^{-3}*$	8.63 $6.76 \times 10^{-18}*$	-2.17 $1.71 \times 10^{-2}*$	0.918 1.92×10^{-1}	
C>T/G>A	-17.1 $7.78 \times 10^{-65}*$	-5.60 $1.76 \times 10^{-8}*$	-14.3 $5.10 \times 10^{-46}*$	-11.1 $1.32 \times 10^{-28}*$	-14.1 $6.46 \times 10^{-45}*$

Increased age of paraffin block results in reduced amplicon yield and elevated events of C>T/G>A sequence artifacts

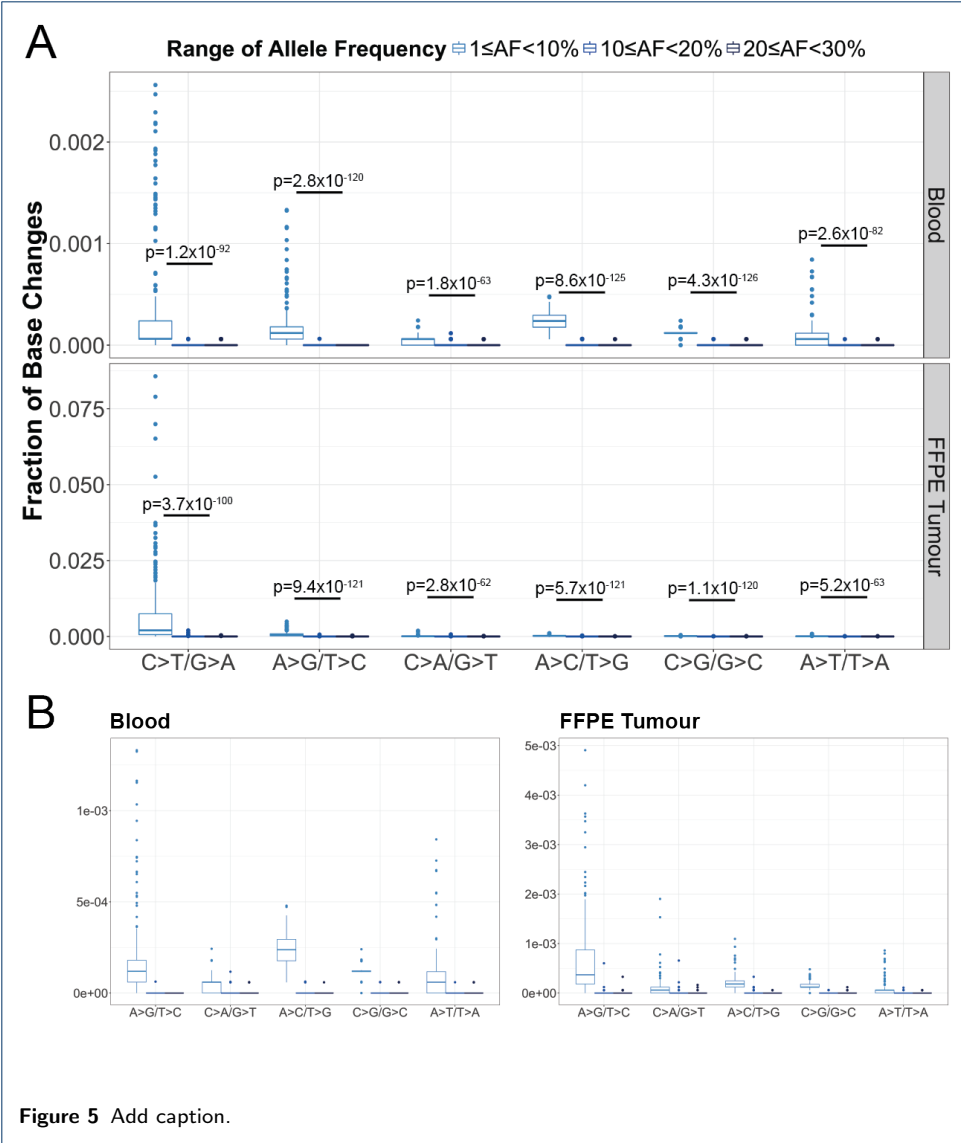Frequency of germline and somatic variants

Figure 5 Add caption.

Table 5 Determination of correlation between pre-sequencing variables and sequencing results using Spearman's correlation. Top values represent Spearman's *rho* and 95% confidence interval in brackets, whereas bottom values represent *p*-value. Asterisk(*) indicates significance level of *p*-value < 0.05.

Variable	Amplicon Yield (ng)	Age of Paraffin Block (Day)	Fraction of C>T/G>A	Average Per Base Normalized Coverage
Age of Paraffin Block (Day)	-0.42 (-0.52– -0.30) 5.2×10^{-7}*			
Fraction of C>T/G>A	-0.72 (-0.77– -0.65) 1.9×10^{-11}*	0.54 (0.61–0.75) 6.3×10^{-35}*		
Average Per Base Normalized Coverage	0.69 (0.61–0.75) 8.5×10^{-20}*	-0.47 (-0.57– -0.36) 4.7×10^{-7}*	-0.80 (-0.84– -0.75) 7.5×10^{-17}*	
On-target Aligned Reads (%)	0.58 (0.48–0.66) 2.1×10^{-13}*	-0.35 (-0.46– -0.23) 8.2×10^{-3}*	-0.57 (-0.65– -0.47) 4.2×10^{-8}*	0.73 (0.66–0.79) 3.1×10^{-58}*

Germline variants are highly concordant between blood and FFPE specimens

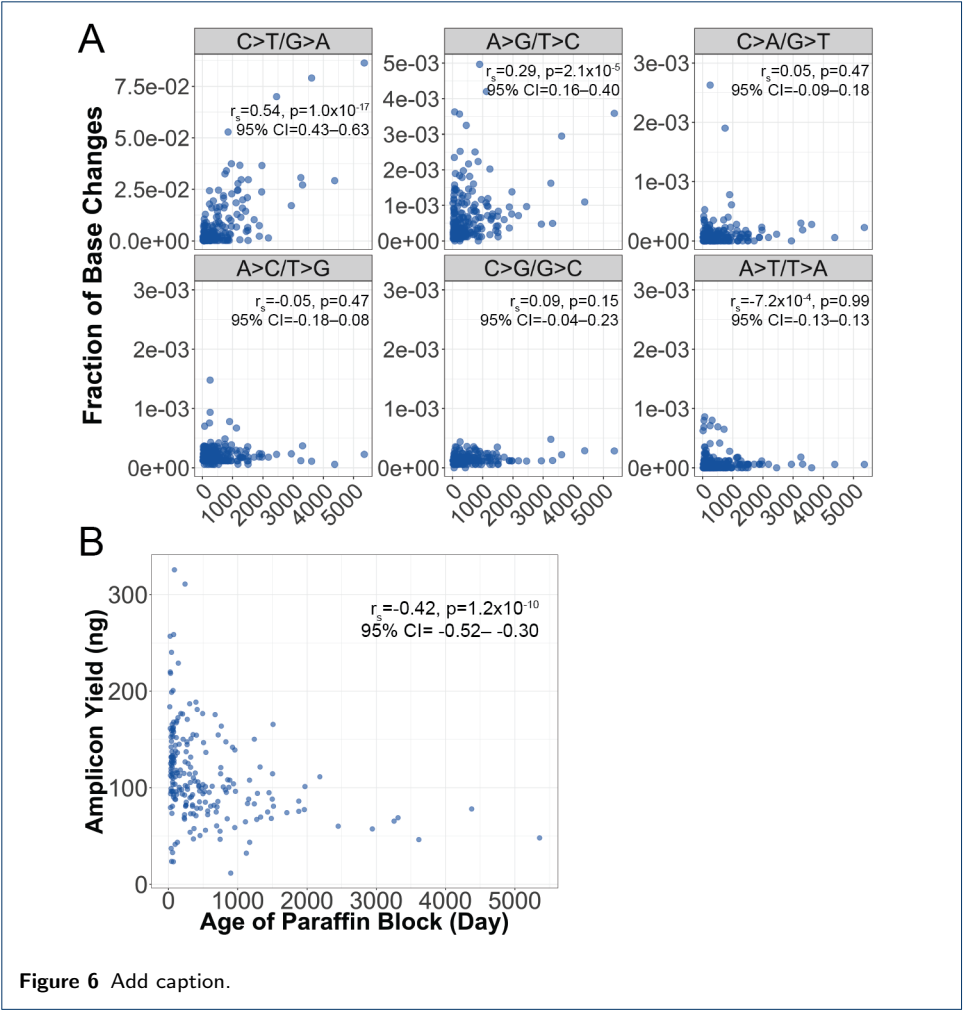Reduced sensitivity is observed for detection of germline variants in FFPE specimens compared to blood

Figure 6 Add caption.

Table 6 Sensitivity of detecting germline variants in blood and FFPE specimens at various variant allele frequency thresholds.

	Blood				FFPE Tumour			
VAF (%)	FN*	TP**	Sensitivity	95% CI	FN*	TP**	Sensitivity	95% CI
10	0	2461	1.0	1.0–1.0	0	2428	1.0	1.0–1.0
15	2	2459	1.0	1.0-1.0	12	2416	1.0	0.99–1.0
20	3	2458	1.0	1.0–1.0	48	2380	0.98	0.97–0.99
25	15	2446	0.99	0.99–1.00	79	2349	0.97	0.96–0.97
30	20	2441	0.99	0.99–1.00	121	2307	0.95	0.94–0.96
35	33	2428	0.99	0.98–0.99	197	2231	0.92	0.91–0.93
40	107	2354	0.96	0.95–0.96	328	2100	0.86	0.85–0.88
45	234	2227	0.90	0.89–0.92	470	1958	0.81	0.79–0.82

High positive predictive value can be achieved for referral of potential germline variants to downstream confirmatory testing
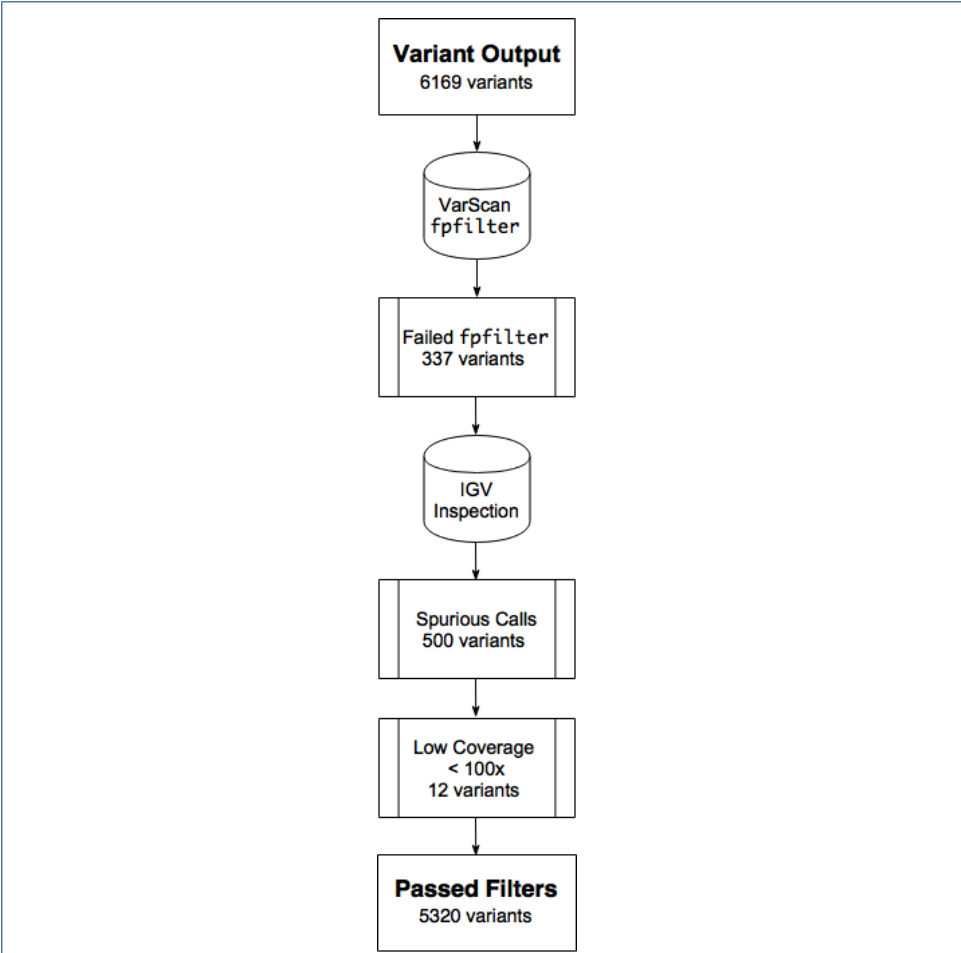
Figure 7 Add caption.

Table 7 Positive predictive value for referral of potential germline variants for downstream confirmatory testing.

VAF (%)	False Positive	True Positive	Total Calls	Positive Predictive Value	95% CI
10	431	2428	2859	0.85	0.84–0.86
15	319	2416	2735	0.88	0.87–0.90
20	273	2380	2653	0.90	0.88–0.91
25	245	2349	2594	0.91	0.89–0.92
30	203	2307	2510	0.92	0.91–0.93
35	178	2231	2409	0.93	0.91–0.94
40	146	2100	2246	0.93	0.92–0.94
45	118	1958	2076	0.94	0.93–0.95

Discussion
Conclusions

**Figure 8** Add caption.

Figure 9 Add caption.
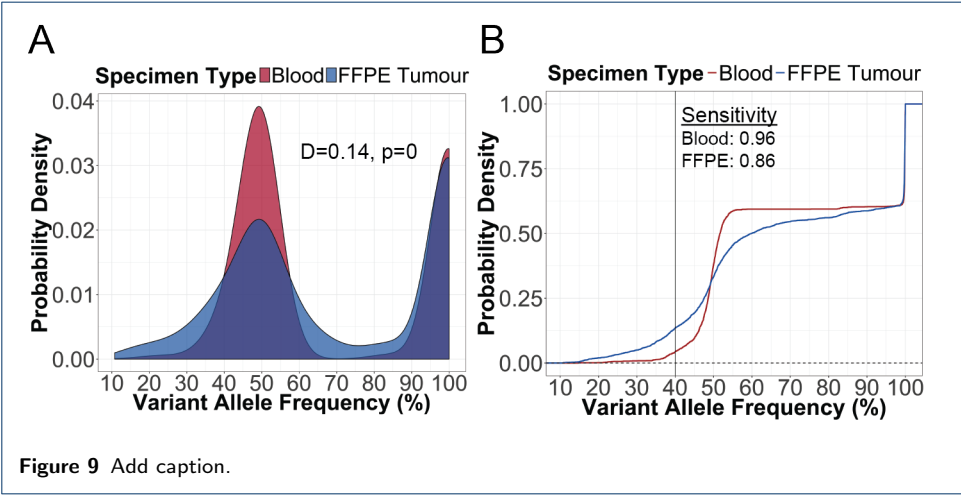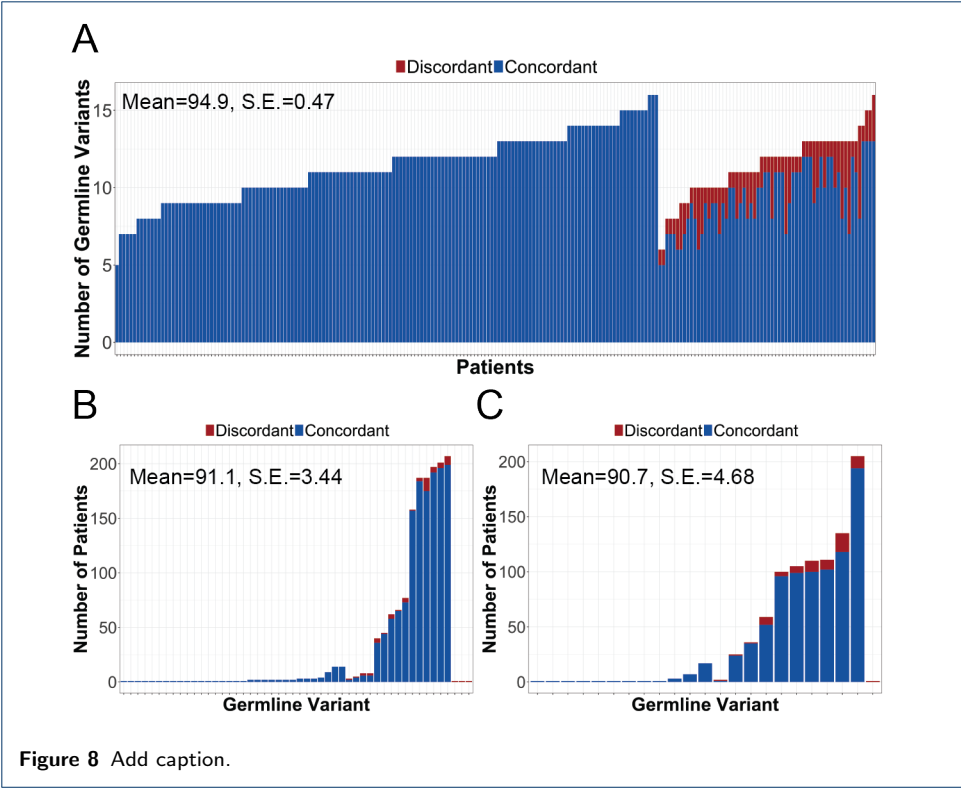
Author details
[1]British Columbia Cancer Research Centre, 675 West 10th Ave, V5Z 1L3 Vancouver, BC, Canada. [2]Department of Pathology and Laboratory Medicine, University of British Columbia, Random Street, Random Post Code Vancouver, BC, Canada.

References
1. Raymond, V.M., Gray, S.W., Roychowdhury, S., Joffe, S., Chinnaiyan, A.M., Parsons, D.W., Plon, S.E.: Germline findings in tumor-only sequencing: Points to consider for clinicians and laboratories. Journal of the National Cancer Institute **108**(4), 1–5 (2016). doi:10.1093/jnci/djv351
2. Meric-Bernstam, F., Brusco, L., Daniels, M., Wathoo, C., Bailey, A.M., Strong, L., Shaw, K., Lu, K., Qi, Y., Zhao, H., Lara-Guerra, H., Litton, J., Arun, B., Eterovic, A.K., Aytac, U., Routbort, M., Subbiah, V., Janku, F., Davies, M.A., Kopetz, S., Mendelsohn, J., Mills, G.B., Chen, K.: Incidental germline variants in 1000 advanced cancers on a prospective somatic genomic profiling protocol. Annals of Oncology **27**(5), 795–800 (2016). doi:10.1093/annonc/mdw018
3. Schrader, K.A., Cheng, D.T., Joseph, V., Prasad, M., Walsh, M., Zehir, A., Ni, A., Thomas, T., Benayed, R.,

Figure 10 Add caption.
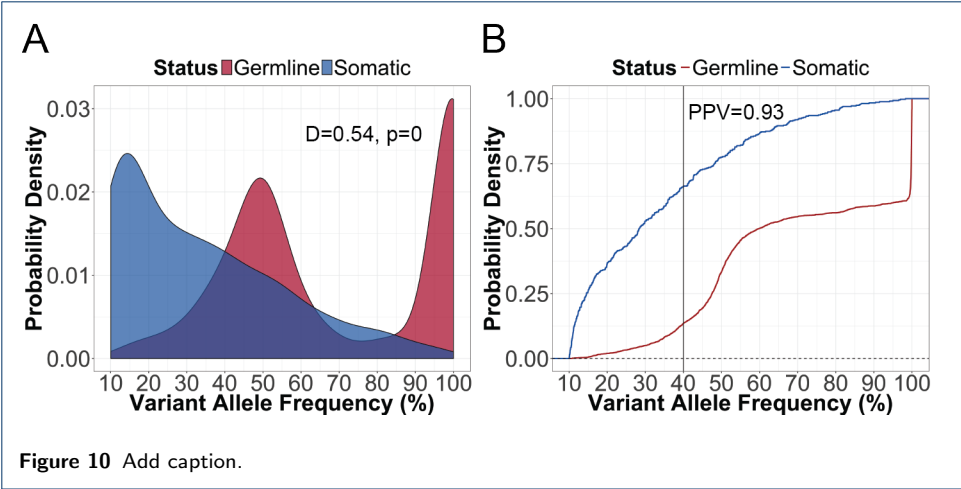
Ashraf, A., Lincoln, A., Arcila, M., Stadler, Z., Solit, D., Hyman, D., Zhang, L., Klimstra, D., Ladanyi, M., Offit, K., Berger, M., Robson, M.: Germline Variants in Targeted Tumor Sequencing Using Matched Normal DNA. JAMA oncology **2**(1), 1–8 (2015). doi:10.1001/jamaoncol.2015.5208

4. Jones, S., Anagnostou, V., Lytle, K., Parpart-li, S., Nesselbush, M., Riley, D.R., Shukla, M., Chesnick, B., Kadan, M., Papp, E., Galens, K.G., Murphy, D., Zhang, T., Kann, L., Sausen, M., Angiuoli, S.V., Jr, L.A.D., Velculescu, V.E.: Personalized genomic analyses for cancer mutation discovery and interpretation. Science Translational Medicine **7**(283), 283–53 (2015). doi:10.1126/scitranslmed.aaa7161

5. McLeod, H.L.: Cancer Pharmacogenomics: Early Promise, But Concerted Effort Needed. Science **339**(March), 1563–1566 (2013). doi:10.1126/science.1234139

6. Do, H., Dobrovic, A.: Sequence artifacts in DNA from formalin-fixed tissues: Causes and strategies for minimization. Clinical Chemistry **61**(1), 64–71 (2015). doi:10.1373/clinchem.2014.223040

7. Wong, S.Q., Li, J., Y-C Tan, A., Vedururu, R., Pang, J.-M.B., Do, H., Ellul, J., Doig, K., Bell, A., MacArthur, G.A., Fox, S.B., Thomas, D.M., Fellowes, A., Parisot, J.P., Dobrovic, A.: Sequence artefacts in a prospective series of formalin-fixed tumours tested for mutations in hotspot regions by massively parallel sequencing. BMC Medical Genomics **7**(1), 1–10 (2014). doi:10.1186/1755-8794-7-23

8. Bosdet, I.E., Docking, T.R., Butterfield, Y.S., Mungall, A.J., Zeng, T., Coope, R.J., Yorida, E., Chow, K., Bala, M., Young, S.S., Hirst, M., Birol, I., Moore, R.A., Jones, S.J., Marra, M.A., Holt, R., Karsan, A.: A clinically validated diagnostic second-generation sequencing assay for detection of hereditary BRCA1 and BRCA2 mutations. Journal of Molecular Diagnostics **15**(6), 796–809 (2013). doi:10.1016/j.jmoldx.2013.07.004

Figures

Figure 11 Sample figure title.

Figure 12 Sample figure title. Figure legend text.

Supplementary Tables

Table 8 Gene Reference Models for Genes in the OncoPanel.

Gene	Protein	Reference Model
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
DPYD	Dihydropyrimidine dehydrogenase	NM_000110.3
EGFR	Epidermal growth factor receptor	NM_005228.3
ERBB2	Receptor tyrosine-protein kinase erbB-2	NM_001005862.1
GSTP1	Glutathione S-rransferase pi 1	NM_000852.3
HRAS	GTPase HRas	NM_005343.2
IDH1	Isocitrate dehydrogenase 1	NM_005896.2
IDH2	Isocitrate dehydrogenase 2	NM_002168.2
KIT	Tyrosine-protein kinase Kit	NM_000222.2
KRAS	KRas proto-oncogene GTPase	NM_033360.2
MAPK1	Mitogen-activated protein kinase 1	NM_002745.4
MAP2K1	Mitogen-activated protein kinase kinase 1	NM_002755.3
MTHFR	Methylenetetrahydrofolate reductase	NM_005957.4
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
TYMP	Thymidine phosphorylase	NM_001113755.2
TYMS	Thymidylate synthetase	NM_001071.2
UGT1A1	Uridine diphosphate (UDP)-glucuronosyl transferase 1A1	NM_000463.2

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.