White Rose Research Online

eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk
Universities of Leeds, Sheffield and York

Deposited via The University of Leeds.

White Rose Research Online URL for this paper:
https://eprints.whiterose.ac.uk/id/eprint/124598/

Version: Accepted Version

Article:
Droop, A, Bruns, A, Tanner, G et al. (2018) How to analyse the spatiotemporal tumour samples needed to investigate cancer evolution: A case study using paired primary and recurrent glioblastoma. International Journal of Cancer, 142 (8). pp. 1620-1626. ISSN: 0020-7136

https://doi.org/10.1002/ijc.31184

(c) 2017 UICC. This is the peer reviewed version of the following article: Droop, A , Bruns, A, Tanner, G et al. (2017) How to Analyse The Spatiotemporal Tumour Samples Needed To Investigate Cancer Evolution: A Case Study using Paired Primary and Recurrent Glioblastoma. International Journal of Cancer , which has been published in final form at https://doi.org/10.1002/ijc.31184. This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Self-Archiving.

Reuse
Items deposited in White Rose Research Online are protected by copyright, with all rights reserved unless indicated otherwise. They may be downloaded and/or printed for private study, or other acts as permitted by national copyright laws. The publisher or other rights holders may allow further reproduction and re-use of the full text version. This is indicated by the licence information on the White Rose Research Online record for the item.

Takedown
If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

UNIVERSITY OF LEEDS
University of Sheffield
UNIVERSITY of York

How to Analyse The Spatiotemporal Tumour Samples Needed To Investigate Cancer Evolution: A Case Study using Paired Primary and Recurrent Glioblastoma

Alastair Droop[1], Alexander Bruns[2¥], Georgette Tanner[2], Nora Rippaus[2], Ruth Morton[2], Sally Harrison[2], Henry King[2], Katherine Ashton[3], Khaja Syed[4], Michael D. Jenkinson[4,5], Andrew Brodbelt[4], Aruna Chakrabarty[6], Azzam Ismail[6], Susan Short[2] and Lucy F. Stead[2,*]

[1] MRC Medical Bioinformatics Centre, University of Leeds, Leeds, LS2 9NL, UK; [2] Leeds Institute of Cancer and Pathology, University of Leeds, Leeds, LS9 7TF, UK; [3] Lancashire Teaching Hospitals NHS Trust, Royal Preston Hospital, Preston, PR2 9HT, UK; [4] Walton Centre NHS Trust, Liverpool, L9 7LJ, UK; [5] Institute of Translational Medicine, University of Liverpool, L9 7LJ, UK; [6] Leeds Teaching Hospitals NHS Trust, St James's University Hospital, Leeds, LS9 7TF, UK

* To whom correspondence should be addressed. Tel: +44 113 343 8410 Email: l.f.stead@leeds.ac.uk.

¥ Present Address: Alexander Bruns - Leeds Institute of Cardiovascular and Metabolic Medicine, University of Leeds, Leeds, LS2 9JT, UK.

Keywords: Somatic mutation; Variant calling; Intratumour heterogeneity; Spatiotemporal; Duplicates; Tumour evolution

Abbreviations:

BAF: B-allele frequency
FFPE: Formalin fixed, paraffin embedded
GBM: Glioblastoma
H&E: Hematoxylin and eosin
SNP: Single nucleotide polymorphism
Somatic TP: Somatic true positive
VAF: Variant allele frequency

Article Type: Short Report

Novelty and Impact: We present a new two-stage approach to identifying somatic mutations that are shared across multiple tumour samples or datasets (eg RNA and DNA sequenced separately) from the same patient, and test it in three independent cohorts of paired primary and recurrent glioblastoma samples. Our results show that our approach more sensitively detects shared genetic variants, which are candidate drivers of tumour progression.

ABSTRACT

Many traits of cancer progression (e.g. development of metastases or resistance to therapy) are facilitated by tumour evolution: Darwinian selection of subclones with distinct genotypes or phenotypes that enable such progression. Characterising these subclones provides an opportunity to develop drugs to better target their specific properties but requires the accurate identification of somatic mutations shared across multiple spatiotemporal tumours from the same patient. Current best practices for calling somatic mutations are optimised for single samples, and risk being too conservative to identify shared mutations with low prevalence in some samples. We reasoned that datasets from multiple matched tumours can be used for mutual validation and thus propose an adapted two-stage approach: 1) low-stringency mutation calling to identify mutations shared across samples irrespective of the weight of evidence in a single sample; 2) high-stringency mutation calling to further characterise mutations present in a single sample. We applied our approach to three independent cohorts of paired primary and recurrent glioblastoma tumours, two of which have previously been analysed using existing approaches, and found that it significantly increased the amount of biologically-relevant shared somatic mutations identified. We also found that duplicate removal was detrimental when identifying shared somatic mutations. Our approach is also applicable when multiple datasets e.g. DNA and RNA are available for the same tumour.

INTRODUCTION

Analysing multiple tumours from the same patient provides novel insights into cancer evolution[1-3]. Genomic subclones shared across spatiotemporal samples highlight candidate drivers of progressive behaviours, such as metastasis (spatially separated samples) and recurrence (temporally separated samples)[4, 5]. Using high-coverage DNA sequencing to characterise somatic mutations in all samples is the first step to identifying shared subclones. Best practices for somatic mutation calling in sequencing data were developed for application to single tumour samples and aim to reduce false positive calls caused by the relatively high error rates in high-throughput sequencers[6-8]. However when analysing multiple tumours, the most biologically relevant mutations are arguably those present in small subclones in one sample but expanded in others. Analysis must, therefore, maximise the chance of capturing such situations, ensuring shared low-allelic fraction mutations are not filtered out from the sample where they are less prevalent. This is especially important for formalin fixed and paraffin embedded (FFPE) samples because this process is known to introduce artefacts at low-allelic fractions, and where multiple samples exist from the same patient it is likely that at least some will be in FFPE[9, 10]. In considering this problem, we reasoned that multiple samples from the same patient provide internal and mutual validation for mutations that may have otherwise been more difficult to assign correctly. We therefore propose a new approach to somatic mutation calling across multiple matched samples:

1) A first round of low-stringency mutation calling to identify tumour-specific variants that self-validate i.e. are present in more than one dataset irrespective of the strength of evidence of any one call. We denote these Somatic TPs (true positives);
2) A second round of stringent mutation calling to additionally identify variants found only in one sample. We denote these Somatic Unknowns.

MATERIALS AND METHODS

More detailed methods are given in Supplementary Materials and Methods

Samples

We identified three independent cohorts of paired patient GBM samples (surgical tissue from primary GBM and subsequent recurrent samples). Clinical information is given in Supplementary Table S1.

Stead Cohort: Eight patients from three tissue banks (Leeds, Liverpool and Preston) with tumours in paraffin blocks and matched blood samples available. Ethical approval was acquired (REC 13/SC/0509). DNA and RNA were extracted simultaneously from tumours (>60% cancer cells), and DNA from blood, using appropriate Qiagen kits (Qiagen, Sussex, UK). PE100 exome libraries (tumour and blood DNA) were made using the SureSelectXT V5 kit (Agilent). PE100 strand-specific whole transcriptome libraries were prepared using the NEBNext Ultra Directional RNA Library Prep Kit for Illumina (New England BioLabs, UK), following rRNA depletion with NEBNext rRNA Depletion Kit or Ribo-Zero Gold. Libraries were sequenced on a HiSeq2500.

Rabadan Cohort: Ten patients from Wang, et al. [11] with exome and whole transcriptome sequencing data for paired tumours, and exome data for matched blood, downloadable from the sequencing read archive (accession SRP074425).

Verhaak Cohort: Four patients from Kim, et al. [12] with high coverage exome (tumour and blood) and poly-A transcriptome sequencing alignment data (tumour) was acquired, and converted to raw fastq format, following application to the European Genome-Phenome Archive (accession EGAS00001001033).

Sequencing Data Processing

Quality processed exome sequencing data was aligned to human reference genome hg19 using BWA mem (v0.7.15)[13]. Two bam files were produced per sample: one with duplicates removed and one with them retained (Picard tools (v2.6.0). All bam files underwent base recalibration and indel realignment (GATK v3.4-46)[14]. RNAseq data was processed as previously described[15].

Variant Calling

Variants were called in all DNA and RNA datasets using Varscan2 (v2.3.9). Briefly: samtools mpileup was run with low mapping and base quality threshold (Phred>=1) and duplicates ignored where required; Varscan2 is then run twice in somatic mode, once with the primary tumour and matched blood, and once with the recurrent tumour and matched blood (minimum coverage: 10X; minimum variant allele frequency [VAF] 3.5%); Varscan2 processSomatic (max VAF in the blood 0.5%) somaticFilter commands are run; finally a customised perl script iterates through the low confidence somatic calls in the primary tumour and re-annotates them as high confidence if they were also called

as somatic (either high or low confidence) in the recurrent sample and then repeats this for the low confidence somatic calls in the recurrent tumour via iterative inspection in the primary tumour somatic calls. Variant consequences were assigned using the Ensembl (release 86) Variant Effect Predictor[16]. All of our wraparound scripts are specific to the variant calling pipeline we have established in house but are available upon request, and guidance in the adaption of existing pipelines is also available via the corresponding author.

Assessment of Variant Calls

Three tables of annotated variation data were created per patient: Germline variants (found in either tumour DNA and in the blood DNA), Somatic TP (true-positive somatic mutations: found in more than one tumour dataset and not in the blood) and Somatic Unknown (found only in the DNA of one tumour only and not in the blood).

Comparative Analysis

To compare the number of Somatic TPs identified in our approach using paired versus unpaired samples, the somatic mutations in each primary tumour were also compared with three unpaired recurrences i.e. random selection of the same number of mutations that were in the paired recurrence from three unpaired recurrences from the same cohort. Somatic TPs identified by our approach were also compared with those from the original analyses (listed in supplementary tables of both published papers[11, 12])

SNP arrays

80ng DNA from three Stead cohort tumours underwent the OncoScan™ FFPE SNP array assay. B-allele frequencies (BAFs) in the raw_snps.txt files were compared with those from variant calling in the exome data.

RESULTS

To test our two-stage approach (Fig.1) we acquired high-coverage exome and RNA sequencing data from three independent cohorts of longitudinal glioblastoma (GBM) samples: the Verhaak cohort (four

patients from Kim, et al.[12]); the Rabadan cohort (ten patients from Wang, et al.[11]); the Stead cohort (our own six patients). These were first diagnosis GBM samples acquired from an initial surgical resection (denoted the primary sample) and post-treatment recurrences (recurrent sample) from a second surgical resection. The Verhaak and Rabadan samples had mutations called, and published, using best practices and validated somatic mutation calling pipelines[11, 12]. Stead tumours were FFPE; Rabadan and Verhaak were snap frozen. Clinical information and sequencing metrics for all samples are in Supplementary Tables S1-3. Supplementary Table S4 shows how many Somatic TPs were validated in the DNA of the remaining tumour and how many, instead, in the RNA of either tumour. The ability to validate within RNA was varied (8±14% of TPs were validated this way) but indicates the applicability of this approach when a single tumour is being analysed but using more than one sequencing dataset.

Our Approach Identifies Additional Shared Variants that are Biologically Relevant

A paired sample analysis identifies significantly more biologically relevant shared mutations. Our approach assumes that observing a mutation in more than one dataset from the same patient validates its existence, irrespective of the strength of evidence in any single dataset. To assess this assumption, in contrast to the possibility than the same mutations may be observed in different tumours owing to technical errors biased towards certain genomic loci, or by chance because relaxed filters identify so many variant loci, we also inspected the number of Somatic TPs acquired when our approach was applied to unpaired primary and recurrent tumours i.e. from different patients. We repeated our analysis three times per primary tumour, randomly selecting the same number of mutations found in its paired recurrence from the mutations called in a different patient's recurrence (same cohort). On average, there were 98±1% (97±5% with duplicates retained) fewer Somatic TPs in unpaired snap frozen samples versus the paired analyses, and 92±6% (93±5% with duplicates retained) fewer in FFPE samples. This indicates that our approach identifies variants that are shared for biological rather than technical reasons.

Comparison with the original Verhaak cohort analysis. Variants called in both our and the original Verhaak cohort (n=4) analysis are in Supplementary Table S5. 241 Somatic TPs were identified in both studies and for these the VAF correlation was 1.00 for both primary and recurrent tumour

samples (0.99 when we retained duplicates). The previous analysis identified one Somatic TP that we called germline as there were reads supporting the variant in the blood according to our alignment. We, however, identified 583 protein-altering Somatic TPs not previously published, likely because they were filtered out during independent tumour variant calling. These were in 517 genes enriched in members of Signalling Pathways in Glioblastoma (Wikipathways WP2261, hypergeometric adjusted p=0.036) including: a *PTEN* splice site mutation previously observed in glioma (COSM39456) and predicted to be pathogenic (fathmm score of 0.99); a *NF1* splice site mutation; an *EGFR* missense mutation, predicted to be damaging (PolyPhen2 p=0.997), only identified in the recurrence in the original analysis. Within the 60 Somatic TPs, uniquely identified by our approach, with a VAF increase of 5% or more from primary to recurrence (i.e. potentially located within clones that not only survived but expanded following therapy), 29 were predicted to be damaging by SIFT, PolyPhen2 and/or fathmm including several in genes previously associated with gliomagenesis e.g. *EXT1*, *NOTCH1* and *TRAF1*[17-19].

Comparison with the original Rabadan cohort analysis. Variants called in both our and the original Rabadan cohort (n=10) study are in Supplementary Table S6. 357 Somatic TPs were identified in both analyses; 7 that were experimentally validated and all 14 known GBM driver mutations. The VAF correlation was 0.99 for both tumours (0.95 and 0.96 with duplicates retained). The previous analysis identified 25 unclustered Somatic TPs that our approach did not: we called 24% germline, 60% only in one tumour and did not observe 16%. We missed one experimentally validated mutation in the primary tumour as it was below our VAF threshold. However, we identified 6416 protein-altering Somatic TPs not previously published. The 4667 genes containing these are: significantly expressed in brain (normal and tumour) and epithelial tissue; enriched for genes involved in nervous system and neuron development and in Signalling Pathways in Glioblastoma (Wikipathways WP2261, Table 1) (hypergeometric, $q<0.05$); contain a significant number of the 75 GBM mutational driver genes listed in the Integrative Onco-Genomics database (n=35, chi-squared p=0.04). The uniquely identified genes in which VAF increased by 5% or more were enriched for members of MAPK and Wnt signalling (Wikipathways WP382 and WP399, hypergeometric $q<0.05$), both strongly associated with gliomagenesis[20, 21].

Duplicate Removal Can Reduce Biological Information

Detecting low VAF (potentially subclonal) variants requires high sequencing coverage[7]. Most analysis pipelines remove duplicates before variant calling for fear that these are PCR artefacts that will amplify errors[6]. However, duplicate removal programmes define duplicates as reads sharing start and end alignment coordinates, ignoring actual sequence. As coverage increases, the chance of two independent reads sharing alignment coordinates increases; if such reads span the position of a low allelic fraction variant, the evidence for it will likely be removed as the programme selects one 'duplicate' at random (or the best according to the summation of base qualities) to retain. We inspected how retaining duplicates affects the number and VAF of each type of variant (Fig.2). Retaining duplicates increases the number of both types of internally validated variation: Germline and Somatic TPs. However, there is a disproportionate increase (note the log scale, Fig.2B) in the number of Somatic Unknowns (comprising false and true positives). The VAF correlation between duplicate removed versus duplicate retained data is always >0.8, though a reduction in the correlation coefficient is observed as the proportion of duplicates increases (Fig.2C). To recap from above, retaining duplicates also did not i) increase the Somatic TPs found in unpaired samples, ii) reduce the VAF between ours and previously published (duplicate removed analyses) Somatic TPs.

We then compared results from SNP microarrays to those of sequencing data (10-27% duplicates) for three Stead cohort samples. In all cases (duplicate removed and retained), the BAF correlated significantly. However, duplicate retention increased the number of variants that could be used in the comparison by 2-5%.

DISCUSSION

Best practice analysis pipelines aim to maximise both sensitivity (detection of real events) and specificity (avoidance of non-real events) and standardise approaches for better cross-dataset comparison. Their use must be with the understanding that each analysis is unique (different data, different questions) and even best-practice cannot reveal the whole truth. For identifying somatic mutations in tumours from sequencing data, best practices were developed for application to single tumour samples, with matched normal (most often blood DNA) providing a germline reference. Commonly studies now require somatic mutation calling across multiple tumours, or regions, from the same patient. We propose that these analyses would benefit from an adapted two-stage approach (Fig.1) that exploits mutual validation across samples to increase the sensitivity of shared mutations

223 detection; mutations of particular interest as they are candidates for conferring clinically relevant
224 phenotypes e.g. the ability metastasize or resist therapy. We recognise, however, that this attempt to
225 increase sensitivity could reduce specificity; low-stringency mutations could appear shared between
226 samples owing to the repeated introduction of technical artefacts or FFPE-induced mutations. We
227 tested this by assessing the number of Somatic TPs identified when primary tumours were analysed
228 with *unpaired* recurrences i.e. where shared variation is due to artefacts at the same position in both
229 samples or independently arising real mutations, which cannot be ruled out but could also be the case
230 in paired samples owing to convergent evolution. We found a large (>90%) reduction in Somatic TPs
231 in unpaired versus paired samples, indicating that our approach identifies real, biological mutations
232 even in FFPE samples. Alternatively, our approach is mis-calling germline variants as shared somatic
233 mutations. This is unlikely as all mutations are called in parallel to a matched blood, also sequenced
234 to a high coverage (167±54X or 213±72X in duplicate removed and retained data) with minimum 10X
235 is required at variant loci. Furthermore, Somatic TPs identified uniquely by our approach in the
236 Verhaak and Rabadan cohorts are enriched in genes in biologically relevant pathways; germline
237 variants and artefacts would occur randomly throughout the genome whereas somatic mutations
238 occur more often in genes activated in the diseased tissue owing to DNA exposure upon
239 transcription[22]. If sequencing is being done to detect specific mutations that may be driver events,
240 therapeutic targets or useful in clinical diagnosis, specificity is key and false positives are intolerable.
241 However, if the aim is to better understand patterns of tumour evolution across numerous patients as
242 part of basic scientific discovery e.g. in order to assess gene networks, signalling pathways or
243 biological processes enriched in clonally expanded populations, it is arguably worth risking noise in
244 the data to ensure true signal is detected above such background. Findings then form the basis of
245 hypotheses to be thoroughly tested in the laboratory. More sensitive detection of shared variation will
246 also improve our detection of pan-genome mutational signatures which can: indicate cancer aetiology;
247 inform on modes of evolution[23]; more accurately indicate therapy-driven mutational load[24].
248 We inspected the effect of duplicate removal on shared variant calling and found that the first round of
249 low-stringency variant calling benefitted from duplicate retention but the second more stringent round
250 of variant calling should be in duplicate-removed data.
251 Numerous variant callers exist and benchmarking studies show they often give very different results[25].
252 Such studies are, however, challenging owing to the difference in parameter defaults for each caller,

and the need to account for external variables e.g. sequencing depth and tumour purity. We herein used Varscan2, which we previously found to accurately identify low VAF somatic mutations[7]. The Verhaak and Rabadan cohort studies used different callers: MuTect and SAVI2 respectively (with different versions of BWA for the initial alignment). Despite this, we identified a high percentage of the same Somatic TPs (99.6% and 93.5% for each study respectively) but our approach additionally identified many more biologically relevant mutations. We suggest, therefore, that it is worth adapting existing pipelines, irrespective of the variant caller employed, to incorporate a reduced stringency first round of mutation calling and subsequent identification of mutually validating shared variants.

AUTHOR CONTRIBUTIONS LFS devised the project. LFS and SS acquired funding. KA, KS, MDJ, AB and AA sourced samples and provided clinical annotation. AB and NR (supervised by LFS) and RM and HK (supervised by SS) processed samples following annotation and diagnostic confirmation from AA and AI. SH optimised the sequencing of FFPE samples. AD, GT and LFS performed data analysis and interpretation. LFS wrote the manuscript, which was reviewed and approved by all authors.

ACKNOWLEDGEMENTS We thank: Professor Roel Verhaak and the legal officers at the M.D Anderson Cancer Centre for access to dataset EGAS00001001033 and feedback on our manuscript; Drs Marica Eoli and Gaetano Finocchiaro for providing clinical information for the Rabadan cohort and for feedback on our manuscript; Agilent for performing the OncoScan™ FFPE SNP array assay and subsequent analysis. This work was supported by the Leeds Teaching Hospitals Charitable Foundation [grant number 9R11/14-11 to LFS]; the Wellcome Trust and the University of Leeds [grant numbers RGCALA101195, 11061187 personal fellowships to LFS]; Leeds MRC Medical Bioinformatics Centre and Cancer Research UK Leeds Centre [grant number MR/LO1629X and infrastructure award C37059/A18080 used to fund AD]. Funding for open access charge: University of Leeds Academic Fellowship. The authors declare that they have no conflicts of interest

REFERENCES

1. Gerlinger M, Horswell S, Larkin J, Rowan AJ, Salm MP, Varela I, Fisher R, McGranahan N, Matthews N, Santos CR, Martinez P, Phillimore B, et al. Genomic architecture and evolution of clear cell renal cell carcinomas defined by multiregion sequencing. *Nat Genet* 2014;**46**: 225-33.

2. Sottoriva A, Spiteri I, Piccirillo SGM, Touloumis A, Collins VP, Marioni JC, Curtis C, Watts C, Tavare S. Intratumor heterogeneity in human glioblastoma reflects cancer evolutionary dynamics. *Proceedings of the National Academy of Sciences* 2013;**110**: 4009-14.

3. Wood HM, Conway C, Daly C, Chalkley R, Berri S, Senguven B, Stead L, Ross L, Egan P, Chengot P, Graham J, Sethi N, et al. The clonal relationships between pre-cancer and cancer revealed by ultra-deep sequencing. *The Journal of Pathology* 2015;**237**: 296-306.

4. Fisher R, Pusztai L, Swanton C. Cancer heterogeneity: implications for targeted therapeutics. *Br J Cancer* 2013;**108**: 479-85.

5. Kim J, Lee I-H, Cho Hee J, Park C-K, Jung Y-S, Kim Y, Nam So H, Kim Byung S, Johnson Mark D, Kong D-S, Seol Ho J, Lee J-I, et al. Spatiotemporal Evolution of the Primary Glioblastoma Genome. *Cancer Cell* 2015;**28**: 318-28.

6. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell TJ, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* 2011;**43**: 491-8.

7. Stead LF, Sutton KM, Taylor GR, Quirke P, Rabbitts P. Accurately identifying low-allelic fraction variants in single samples with next-generation sequencing; applications in tumour subclone resolution. *Human Mutation* 2013;**34**: 1432-8.

8. Wang Q, Jia P, Li F, Chen H, Ji H, Hucks D, Dahlman KB, Pao W, Zhao Z. Detecting somatic point mutations in cancer genome sequencing data: a comparison of mutation callers. *Genome Medicine* 2013;**5**: 91-.

9. Wong SQ, Li J, Tan AYC, Vedururu R, Pang J-MB, Do H, Ellul J, Doig K, Bell A, McArthur GA, Fox SB, Thomas DM, et al. Sequence artefacts in a prospective series of formalin-fixed tumours tested for mutations in hotspot regions by massively parallel sequencing. *BMC Medical Genomics* 2014;**7**: 23.

10. Yost SE, Smith EN, Schwab RB, Bao L, Jung H, Wang X, Voest E, Pierce JP, Messer K, Parker BA, Harismendy O, Frazer KA. Identification of high-confidence somatic mutations in whole genome sequence of formalin-fixed breast cancer specimens. *Nucleic Acids Research* 2012;**40**: e107-e.

11. Wang J, Cazzato E, Ladewig E, Frattini V, Rosenbloom DIS, Zairis S, Abate F, Liu Z, Elliott O, Shin Y-J, Lee J-K, Lee I-H, et al. Clonal evolution of glioblastoma under therapy. *Nat Genet* 2016;**48**: 768-76.

12. Kim H, Zheng S, Amini SS, Virk SM, Mikkelsen T, Brat DJ, Grimsby J, Sougnez C, Muller F, Hu J, Sloan AE, Cohen ML, et al. Whole-genome and multisector exome sequencing of primary and post-treatment glioblastoma reveals patterns of tumor evolution. *Genome Research* 2015;**25**: 316-27.

13. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 2009;**25**: 1754-60.

14. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research* 2010;**20**: 1297-303.

15. Conway C, Graham JL, Chengot P, Daly C, Chalkley R, Ross L, Droop A, Rabbitts P, Stead LF. Elucidating drivers of oral epithelial dysplasia formation and malignant transformation to cancer using RNAseq. *Oncotarget; Vol 6, No 37* 2015.

16. Aken BL, Ayling S, Barrell D, Clarke L, Curwen V, Fairley S, Fernandez Banet J, Billis K, García Girón C, Hourlier T, Howe K, Kähäri A, et al. The Ensembl gene annotation system. *Database* 2016;**2016**: baw093-baw.

17. Xiong A, Kundu S, Forsberg-Nilsson K. Heparan sulfate in the regulation of neural differentiation and glioma development. *FEBS Journal* 2014;**281**: 4993-5008.

18. Somasundaram K, Reddy SP, Vinnakota K, Britto R, Subbarayan M, Nambiar S, Hebbar A, Samuel C, Shetty M, Sreepathi HK, Santosh V, Hegde AS, et al. Upregulation of ASCL1 and inhibition of Notch signaling pathway characterize progressive astrocytoma. *Oncogene* 2005;**24**: 7073-83.

19. Lee J, Hoxha E, Song H-R. A novel NFIA-NFκB feed-forward loop contributes to glioblastoma cell survival. *Neuro-Oncology* 2017;**19**: 524-34.

20. Pandey V, Bhaskara VK, Babu PP. Implications of mitogen-activated protein kinase signaling in glioma. *Journal of Neuroscience Research* 2016;**94**: 114-27.

21. Lee Y, Lee J-K, Ahn SH, Lee J, Nam D-H. WNT signaling in glioblastoma and therapeutic opportunities. *Lab Invest* 2016;**96**: 137-50.

22. Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A, Carter SL, Stewart C, Mermel CH, Roberts SA, Kiezun A, Hammerman PS, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* 2013;**499**: 214-8.

23. Williams MJ, Werner B, Barnes CP, Graham TA, Sottoriva A. Identification of neutral tumor evolution across cancer types. *Nat Genet* 2016;**48**: 238-44.

24. Johnson BE, Mazor T, Hong C, Barnes M, Aihara K, McLean CY, Fouse SD, Yamamoto S, Ueda H, Tatsuno K, Asthana S, Jalbert LE, et al. Mutational Analysis Reveals the Origin and Therapy-Driven Evolution of Recurrent Glioma. *Science* 2014;**343**: 189-93.

25. Krøigård AB, Thomassen M, Lænkholm A-V, Kruse TA, Larsen MJ. Evaluation of Nine Somatic Variant Callers for Detection of Somatic Mutations in Exome and Targeted Deep Sequencing Data. *PLOS ONE* 2016;**11**: e0151664.

FIGURE LEGENDS

Fig.1. An overview of our two-stage approach to identifying somatic variants across multiple tumour samples or datasets from the same patient.

Fig.2. Assessing the effect of duplicate removal on variant calling in multiple glioblastoma (GBM) tumour samples. A) The fraction of reads marked as duplicates (± SD). B) The effect of retaining duplicates on the number of different types of mutation called (± SD). C) Scatterplot showing how the fraction of duplicates alters the correlation between allelic frequencies in variants identified in duplicate-removed versus duplicate-retained sequencing data. Verhaak, Rabadan and Stead are three independent cohorts of samples trios (blood, primary GBM and recurrent GBM). See Methods for the definition of Germline, Somatic TP and Somatic Unknown variants.