Supplementary Appendix

This appendix has been provided by the authors to give readers additional information about their work.

Supplement to: Coorens THH, Farndon SJ, Mitchell TJ, et al. Lineage-independent tumors in bilateral neuroblastoma. N Engl J Med 2020;383:1860-5. DOI: 10.1056/NEJMoa2000962

Supplementary Appendix

This appendix has been provided by the authors to give readers additional information about their work.

Supplementary Appendix to:

Lineage-independent Tumors in Bilateral Neuroblastoma

Tim H.H. Coorens, MPhil[1]*, Sarah J. Farndon, MBBS[2]*, Tom Mitchell, BMBCh, DPhil[1,2,3], Neha Jain, MBBS[4,5], Sangjin Lee, MPhil[1], Michael Hubank, PhD[6], Neil Sebire, MBBS, PhD[4,5], John Anderson, MBBS, PhD[4,5], Sam Behjati, BMBCh, PhD[1,2,7,†]

[1]Wellcome Sanger Institute, Hinxton, CB10 1SA, UK.
[2]Cambridge University Hospitals NHS Foundation Trust, Cambridge, CB2 0QQ, UK.
[3]Department of Surgery, University of Cambridge, Cambridge, CB2 0QQ, UK.
[4]UCL Great Ormond Street Institute of Child Health, London, WC1N 1EH, UK.
[5]Great Ormond Street Hospital for Children NHS Foundation Trust, London, WC1N 3JH, UK.
[6]The Royal Marsden NHS Foundation Trust, London, SW3 6JJ, UK
[7]Department of Paediatrics, University of Cambridge, Cambridge, CB2 0QQ, UK.

*These authors contributed equally.
[†]Corresponding author. Email: sb31@sanger.ac.uk

Table of contents

Methods

Samples and sequencing

All human material was obtained from patients enrolled in the study, "Investigating how childhood tumors and congenital disease develop" (approved by a UK NHS National Research Ethics Service; reference 16/EE/0394). DNA was extracted from fresh frozen tumor samples or blood samples. Prior to extraction of DNA from blood, plasma was removed. Short insert (500bp) genomic libraries were constructed and 150 base pair paired-end sequencing clusters were generated on the Illumina HiSeq X platform according to Illumina no-PCR library protocols. An overview of samples, including the average sequence coverage is shown in **Table S1**.

Data and materials availability

Raw sequencing data have been deposited in the European Genome-phenome Archive (EGA) under study ID EGAD00001005770.

DNA sequence processing and mutation calling

DNA sequences were aligned to the GRCh37d5 reference genome by the Burrows-Wheeler algorithm (BWA-MEM)[1].

Single-nucleotide variants (SNVs) and short insertion and deletions (indels) were called against the reference genome using CaVEMan[2] and Pindel[3], respectively. Beyond the standard post-processing filters of CaVEMan, we removed variants affected mapping artefacts associated with BWA-MEM by setting the median alignment score of reads supporting a mutation as greater than or equal to 140 (ASMD>=140) and requiring that fewer than half of the reads were clipped (CLPM=0). Across all samples from one patient and their parents, we recounted the SNVs and indels that were called in either blood or tumor from the patient, using a cut-off for read mapping quality (30) and base quality (25). Germline variants were removed by filtering out any variants supported in the parents, allowing one supporting read per parent for SNVs to accommodate sequencing noise. Variants were also filtered out if they were called in a region of consistently low or high depth across all non-tumor samples from one patient. This amounted to an average depth of between 75 and 200 for PD34954, and an average depth of between 30 and 90 for autosomes, and between 15 and 45 for sex chromosomes in PD36812 (**Figure S1**).

Using a beta-binomial model of a site-specific error rate as previously employed[5], we distinguished true presence of somatic variants from support due to noise. All shared SNVs and indels were further visually inspected using the genome browser, Jbrowse[6]. Within the remaining subset of variants that were present in more than one sample from the same patient, we further distinguished between *de novo* germline variants and true somatic variants. For this we used a one-sided binomial exact test on the number of variant reads and depth present in the matched blood sample to test whether the observed counts were consistent with a true VAF of 0.5 (or 0.95 for XY chromosomes in PD36812). Resulting p-values were corrected for multiple testing with the Benjamini-Hochberg method[4] and a cut-off was

set at $q < 10^{-5}$. Please see **Table S2** and **Table S3** for indel and SNV calls of PD34954 and PD36812, respectively.

In PD34954a and PD34954c, we discovered a small region (chr12: 69Mb-70.7Mb) with many SNV and indel calls, most of which (16/22) were present in dbSNP. In addition, SNPs with informative parental genotypes in this region exhibited VAFs inconsistent with a true gain of either maternal or paternal chromosome (**Figure S2**). The MDM2 gene lies in this region, which frequently forms double minutes in human cancers and can be present in many copies[7]. Therefore, a very small contamination (~0.1%) with tumor DNA containing such a high number of MDM2 double minutes would explain this very focally observed contamination. SNVs, indels, structural variants, and copy number changes from this region were excluded from further analysis in these samples.

Copy number variants (CNVs) were called using ASCAT[8] and Battenberg[9]. In addition, the sequenced parental DNA allowed for full phasing of the copy number variants and assessing whether the maternal or paternal chromosome was affected. Structural variants (SVs) were called using BRASS[10] and SVs were retained if having a paired end read support of at least 30 or where reconstruction of the breakpoint via assembly was possible. SV calls were further validated using SvABA[11] (see **Table S4**).

Classification of SNVs and Indels

To distinguish subclonal from clonal mutations in the tumor samples, we employed a binomial mixture model to deconvolve the mutation counts into separate components. For each component, the optimal binomial probability and mixing proportion is estimated using an expectation-maximisation algorithm. The optimal number of components is determined by the Bayesian information criterion. If the binomial probability of a component approximates the expected VAF (0.5 for diploid regions) adjusted for tumor purity, the mutations assigned to that cluster are classified as clonal. If the estimated binomial probability for a component is lower, it is classified as subclonal. With the exception of regions affected by copy number changes or male sex chromosomes, a higher than expected binomial probability was not observed.

Timing of CNVs

Large-scale copy number duplications can be timed by comparing the number of mutations before and after the copy number gain. For 2:1 or 2:0 configurations, this relation takes the following form:

$$T = \frac{C_M + C_P}{\max(C_M, C_P) + \frac{P_{ND}}{P_D}}$$

Where C_M and C_P are the maternal and paternal copy number, respectively, and P_D and P_{ND} the proportion of mutations assigned to the duplicated or non-duplicated copy number. The timepoint of duplication will be a number between 0 and 1, the former representing the zygote and the latter the most recent common tumor ancestor.

We estimated the proportion of duplicated and non-duplicated mutations using the previously mentioned binomial mixture model on the variant supporting counts and total counts in the region of the gain, fitting the estimated binomial probability of the components to the expected VAFs resulting from the different copy number states (**Figure S3**). The expected VAFs were corrected for the purity of the tumor sample as reported by Battenberg.

To obtain a confidence interval around the single timepoint estimate, we employed an exact Poisson test on the rounded duplicated and non-duplicated mutation counts. These were obtained by multiplying the estimated proportion of duplicated and non-duplicated mutations with the total number of mutations in the region.

To discern the likelihood of two gains occurring at the same time, we used the Poisson test again to compare the sets of duplicated and non-duplicated counts from two different copy number gains in the same patient.

Results for this analysis can be found in **Table S5**.

Code availability

All bespoke code used and described in this paper can be found online at https://github.com/TimCoorens/BilateralNeuroblastoma.

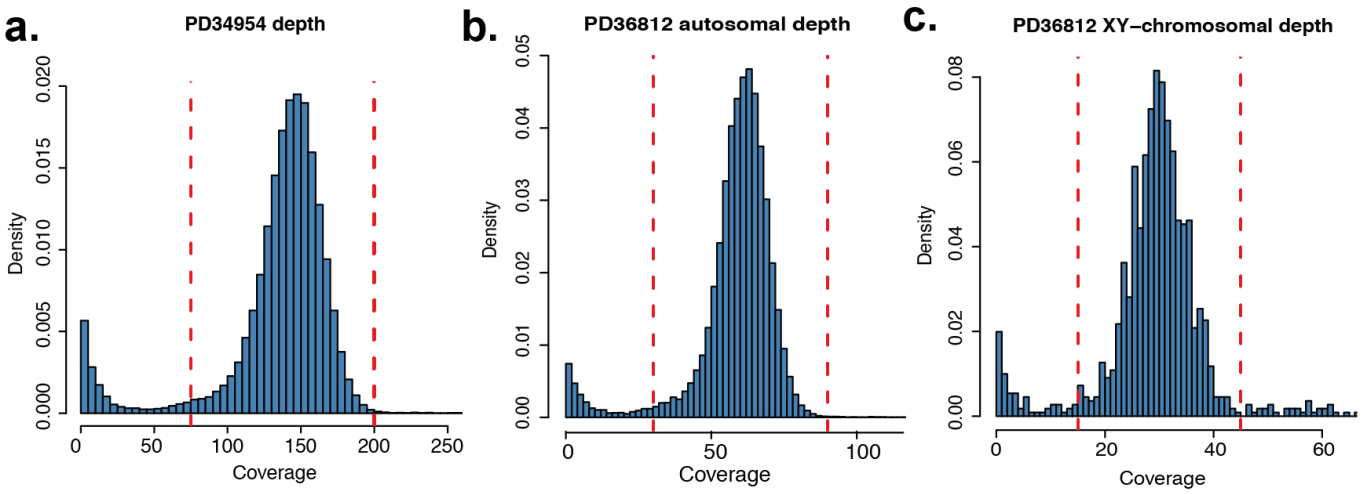

Figure S1: Exclusion of areas with low or high coverage
Histograms of the total depth of variant loci from CaVEMan calls, including rare germline variants, with the used lower and upper bound for acceptably covered sited for PD34954 (**a**), autosomes (**b**) and sex chromosomes in PD36812 (**c**).

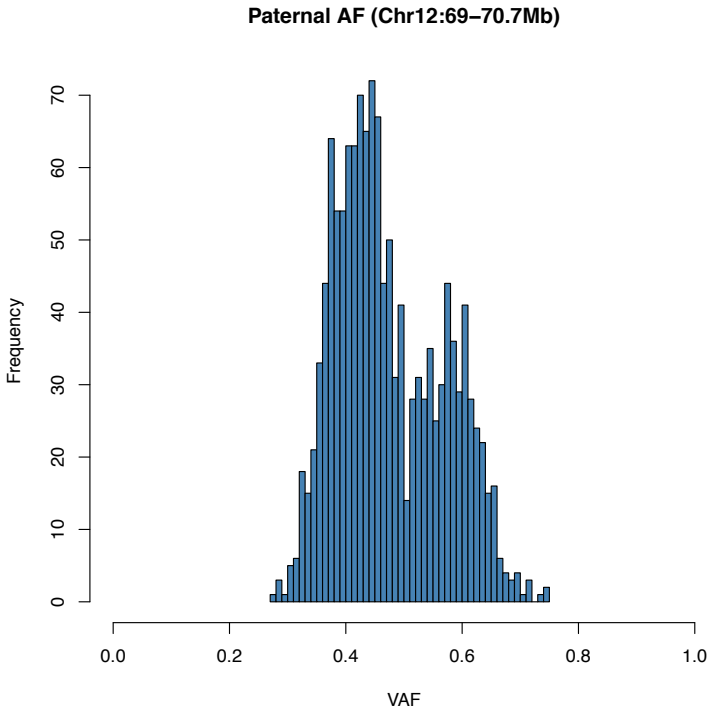

Figure S2: Allele frequency imbalance in contaminated region

Paternal allele frequency of heterozygous informative SNPs in the region affected by a focal contamination. While no copy number change was detected by Battenberg using full phasing, the parental VAF distribution shows a bimodality that is inconsistent with a true normal diploid copy number state. Instead, this indicates a contamination from a different individual, who happens to share a proportion of germline SNPs with one parent or the other, resulting in higher- and lower-than-expected VAFs. As discussed in the methods, this observation is consistent with a very low-level contamination of an independent tumor with a largely amplified double minute carrying MDM2.

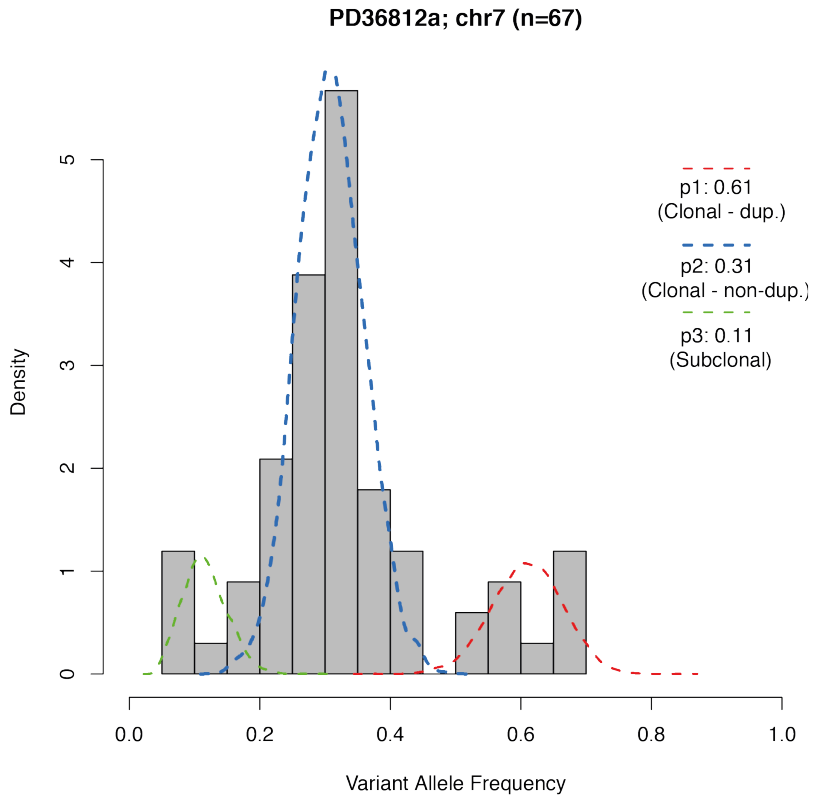

Figure S3: Proportion estimation for CNV timing

Histogram of VAF of mutations located on chromosome 7 in PD36812a. This tumor has a duplication of the maternal chromosome. Hence, the mutations pre-duplication on the duplicated allele will have a VAF of 2/3. All other mutations in this region will have a VAF of 1/3. Mutations with a lower VAF will be subclonal. These estimates will be influence by the estimated purity of the tumor, which is 0.92 in this case. The binomial mixture model extracts three components for mutations in this region: one clonal and duplicated copy number (red), one clonal and non-duplicated copy number (blue), and one subclonal component (green). The ratio between the proportion of the red and blue component is then used to time the copy number gain.

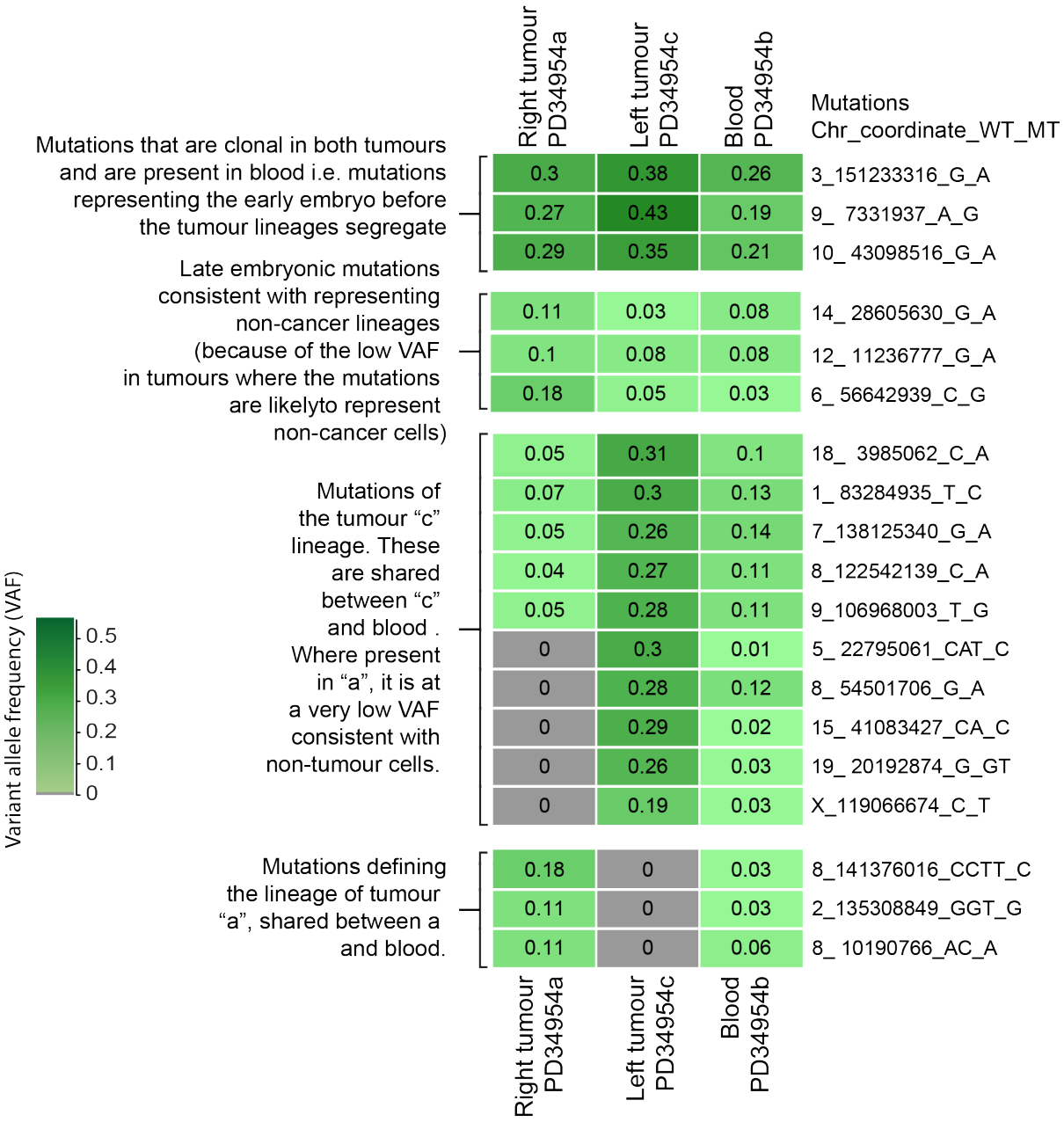

Figure S4: Early somatic variants in PD34954
Heatmap of variant allele frequencies of all shared somatic variants, i.e. variants present in more than one sample. All of these variants are present in the blood sample.

Overview of supplementary tables

Tables S1 to S5 are contained in file “Supplementary_Tables_S1-S5.xlsx”.

Table S1. Overview of study cohort.
Table S2. SNVs/Indels in PD34954
Table S3. SNVs/Indels in PD36812
Table S4. Structural variants
Table S5. Clonal Copy Number Variants and Timing of Gains

References

1. Li H, Durbin R, Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics. 2009; 25(14): 1754-1760
2. Jones DM, Raine KM, Davies H, et al. cgpCaVEManWrapper: simple execution of CaVEMan in order to detect somatic single nucleotide variants in NGS data. Current protocols in bioinformatics. 2016; 56(1): 15-10
3. Ye K, Schulz MH, Long Q, et al. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics. 2009; 25(21): 2865-2871
4. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the royal statistical society. Series B (Methodological). 1995; 289-300.
5. Coorens THH, Treger T, Al-Saadi R, et al. Embryonal precursors of Wilms tumor. Science. 2019; 366(6470):1247-1251
6. Buels R, Yao E, Diesh CM, et al. JBrowse: a dynamic web platform for genome visualization and analysis. Genome biology. 2016; 17(1): 66.
7. Oliner JD, Saiki AY, Caenepeel S. The role of MDM2 amplification and overexpression in tumorigenesis. Cold Spring Harbor perspectives in medicine. 2016; 6(6): a026336.
8. Van Loo P, Nordgard SH, Lingjærde OC, et al. Allele-specific copy number analysis of tumors. Proceedings of the National Academy of Sciences. 2010; 107(39): 16910-16915
9. Nik-Zainal S, Van Loo P, Wedge DC, et al. The life history of 21 breast cancers. Cell. 2012; 149: 994-1007
10. Nik-Zainal S, Davies H, Staaf J, et al. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. Nature. 2016; 534(7605): 47054
11. Wala JA, Bandopadhayay P, Greenwald NF, et al. SvABA: genome-wide detection of structural variants and indels by local assembly. Genome research. 2018; 28(4), 581-591.