Appendix 16

First and joint first author primary research publications

---

1) Abelson, S., Collord G., et al. (2018). "Prediction of acute myeloid leukaemia risk in healthy individuals." *Nature* 559 (7714): 400-404. [PMID: 29988082]

2) Collord, G, et al. (2018). "An integrated genomic analysis of anaplastic meningioma identifies prognostic molecular signatures." *Sci Rep* 8(1): 13537. [PMID: 30202034]

3) Caesar R, Collord G, et al. (2018). "Targeting MEK in vemurafenib-resistant hairy cell leukemia." *Leukemia*. [PMID: 30341394]

4) Wegert J, Vokuhl C, Collord G, et al. (2018). "Recurrent intragenic rearrangements of EGFR and BRAF in soft tissue tumors of infants." *Nat Commun* 9(1): 2378. [PMID: 29915264]

5) Collord G, et al. (2018). "Recurrent histone mutations in T-cell acute lymphoblastic leukaemia." *Br J Haematol*. [PMID: 29602208]

6) Collord G, et al. (2017). "Clonal haematopoiesis is not prevalent in survivors of childhood cancer." *Br J Haematol*. [PMID: 28369776]

LETTER

https://doi.org/10.1038/s41586-018-0317-6

Prediction of acute myeloid leukaemia risk in healthy individuals

Sagi Abelson[1,46], Grace Collord[2,3,46], Stanley W. K. Ng[4], Omer Weissbrod[5], Netta Mendelson Cohen[5], Elisabeth Niemeyer[6], Noam Barda[7], Philip C. Zuzarte[8], Lawrence Heisler[8], Yogi Sundaravadanam[8], Robert Luben[9], Shabina Hayat[9], Ting Ting Wang[1,10], Zhen Zhao[1], Iulia Cirlan[1], Trevor J. Pugh[1,8,10], David Soave[8], Karen Ng[8], Calli Latimer[2], Claire Hardy[2], Keiran Raine[2], David Jones[2], Diana Hoult[11], Abigail Britten[11], John D. McPherson[8], Mattias Johansson[12], Faridah Mbabaali[8], Jenna Eagles[8], Jessica K. Miller[8], Danielle Pasternack[8], Lee Timms[8], Paul Krzyzanowski[8], Philip Awadalla[8], Rui Costa[13], Eran Segal[5], Scott V. Bratman[1,8,14], Philip Beer[2], Sam Behjati[2,3], Inigo Martincorena[2], Jean C. Y. Wang[1,15,16], Kristian M. Bowles[17,18], J. Ramón Quirós[19], Anna Karakatsani[20,21], Carlo La Vecchia[20,22], Antonia Trichopoulou[20], Elena Salamanca-Fernández[23,24], José M. Huerta[24,25], Aurelio Barricarte[24,26,27], Ruth C. Travis[28], Rosario Tumino[29], Giovanna Masala[30], Heiner Boeing[31], Salvatore Panico[32], Rudolf Kaaks[33], Alwin Krämer[34], Sabina Sieri[35], Elio Riboli[36], Paolo Vineis[36], Matthieu Foll[12], James McKay[12], Silvia Polidoro[37], Núria Sala[38], Kay-Tee Khaw[39], Roel Vermeulen[40], Peter J. Campbell[2,41], Elli Papaemmanuil[2,42], Mark D. Minden[1,10,15,16], Amos Tanay[5], Ran D. Balicer[7], Nicholas J. Wareham[11], Moritz Gerstung[2,13,47]*, John E. Dick[1,43,47]*, Paul Brennan[12,47]*, George S. Vassiliou[2,41,44,47]* & Liran I. Shlush[1,6,45,47]*

The incidence of acute myeloid leukaemia (AML) increases with age and mortality exceeds 90% when diagnosed after age 65. Most cases arise without any detectable early symptoms and patients usually present with the acute complications of bone marrow failure[1]. The onset of such de novo AML cases is typically preceded by the accumulation of somatic mutations in preleukaemic haematopoietic stem and progenitor cells (HSPCs) that undergo clonal expansion[2,3]. However, recurrent AML mutations also accumulate in HSPCs during ageing of healthy individuals who do not develop AML, a phenomenon referred to as age-related clonal haematopoiesis (ARCH)[4–8]. Here we use deep sequencing to analyse genes that are recurrently mutated in AML to distinguish between individuals who have a high risk of developing AML and those with benign ARCH. We analysed peripheral blood cells from 95 individuals that were obtained on average 6.3 years before AML diagnosis (pre-AML group), together with 414 unselected age- and gender-matched individuals (control group). Pre-AML cases were distinct from controls and had more mutations per sample, higher variant allele frequencies, indicating greater clonal expansion, and showed enrichment of mutations in specific genes. Genetic parameters were used to derive a model that accurately predicted AML-free survival; this model was validated in an independent cohort of 29 pre-AML cases and 262 controls. Because AML is rare, we also developed an AML predictive model using a large electronic health record database that identified individuals at greater risk. Collectively our findings provide proof-of-concept that it is possible to discriminate ARCH from pre-AML many years before malignant transformation. This could in future enable earlier detection and monitoring, and may help to inform intervention.

To examine the occurrence of somatic mutations before the development of AML, we carried out deep error-corrected targeted sequencing of AML-associated genes in a discovery cohort of 95 pre-AML cases and 414 age- and gender-matched controls (Supplementary Table 1). A validation cohort comprising 29 pre-AML cases and 262 controls (Supplementary Table 1) was analysed using deep sequencing with an overlapping gene panel. Taking both cohorts together, ARCH, defined on the basis of putative driver mutations (ARCH-PD), was found in 73.4% of the pre-AML cases at a median of 7.6 years before diagnosis. By contrast, ARCH-PD was observed in 36.7% of controls ($P < 2.2 \times 10^{-16}$, two-sided Fisher's exact test; Fig. 1a), consistent with data from a study of more than 2,000 unselected individuals assayed using a similarly sensitive method[9,10]. Additionally, 39% of pre-AML cases above the age of 50 had a driver mutation with a variant allele frequency (VAF) of more than 10%, compared to only 4% of controls,

[1]Princess Margaret Cancer Centre, University Health Network (UHN), Toronto, Ontario, Canada. [2]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, UK. [3]Department of Paediatrics, University of Cambridge, Cambridge, UK. [4]Institute of Biomaterials and Biomedical Engineering, University of Toronto, Toronto, Ontario, Canada. [5]Department of Computer Science and Applied Mathematics, Weizmann Institute of Science, Rehovot, Israel. [6]Department of Immunology, Weizmann Institute of Science, Rehovot, Israel. [7]Clalit Research Institute, Tel Aviv, Israel. [8]Ontario Institute for Cancer Research, Toronto, Ontario, Canada. [9]Department of Public Health and Primary Care, Institute of Public Health, University of Cambridge School of Clinical Medicine, Cambridge, UK. [10]Department of Medical Biophysics, University of Toronto, Toronto, Ontario, Canada. [11]MRC Epidemiology Unit, University of Cambridge, Cambridge, UK. [12]International Agency for Research on Cancer, World Health Organization, Lyon, France. [13]European Molecular Biology Laboratory, European Bioinformatics Institute EMBL-EBI, Wellcome Genome Campus, Hinxton, UK. [14]Department of Radiation Oncology, University of Toronto, Toronto, Ontario, Canada. [15]Department of Medicine, University of Toronto, Toronto, Ontario, Canada. [16]Division of Medical Oncology and Hematology, University Health Network, Toronto, Ontario, Canada. [17]Department of Molecular Haematology, Norwich Medical School, The University of East Anglia, Norwich, UK. [18]Department of Haematology, Norfolk and Norwich University Hospitals NHS Trust, Norwich, UK. [19]Public Health Directorate, Asturias, Spain. [20]Hellenic Health Foundation, Athens, Greece. [21]2nd Pulmonary Medicine Department, School of Medicine, National and Kapodistrian University of Athens, "ATTIKON" University Hospital, Haidari, Athens, Greece. [22]Department of Clinical Sciences and Community Health, Università degli Studi di Milano, Milan, Italy. [23]Escuela Andaluza de Salud Pública, Instituto de Investigación Biosanitaria ibs.GRANADA, Hospitales Universitarios de Granada/Universidad de Granada, Granada, Spain. [24]CIBER Epidemiology and Public Health CIBERESP, Madrid, Spain. [25]Department of Epidemiology, Murcia Regional Health Council, IMIB-Arrixaca, Murcia, Spain. [26]Navarra Public Health Institute, Pamplona, Spain. [27]Navarra Institute for Health Research, Pamplona, Spain. [28]Cancer Epidemiology Unit, Nuffield Department of Population Health, University of Oxford, Oxford, UK. [29]Cancer Registry and Histopathology Department, Civic-M. P. Arezzo Hospital, Azienda Sanitaria Provinciale, Ragusa, Italy. [30]Cancer Risk Factors and Life-Style Epidemiology Unit, Cancer Research and Prevention Institute – ISPO, Florence, Italy. [31]Department of Epidemiology, German Institute of Human Nutrition (DIfE), Potsdam-Rehbrücke, Germany. [32]Dipartimento Di Medicina Clinica E Chirurgia, Federico II University, Naples, Italy. [33]Division of Cancer Epidemiology, German Cancer Research Center (DKFZ), Heidelberg, Germany. [34]Clinical Cooperation Unit Molecular Hematology/Oncology, German Cancer Research Center (DKFZ) and Department of Internal Medicine V, University of Heidelberg, Heidelberg, Germany. [35]Epidemiology and Prevention Unit, Fondazione IRCCS Istituto Nazionale dei Tumori, Milano, Italy. [36]Department of Epidemiology and Biostatistics, School of Public Health, Imperial College London, London, UK. [37]Italian Institute for Genomic Medicine, Torino, Italy. [38]Unit of Nutrition and Cancer, Cancer Epidemiology Research Program and Translational Research Laboratory, Catalan Institute of Oncology, ICO-IDIBELL, Barcelona, Spain. [39]University of Cambridge, Cambridge, UK. [40]Division of Environmental Epidemiology and Veterinary Public Health, Institute for Risk Assessment Sciences, Utrecht University, Utrecht, The Netherlands. [41]Department of Haematology, University of Cambridge, Cambridge, UK. [42]Center for Molecular Oncology and Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center, New York, NY, USA. [43]Department of Molecular Genetics, University of Toronto, Toronto, Ontario, Canada. [44]Wellcome Trust–Medical Research Council Cambridge Stem Cell Institute, University of Cambridge, Cambridge, UK. [45]Division of Hematology, Rambam Healthcare Campus, Haifa, Israel. [46]These authors contributed equally: Sagi Abelson, Grace Collord. [47]These authors jointly supervised this work: Moritz Gerstung, John E. Dick, Paul Brennan, George S. Vassiliou, Liran I. Shlush. *e-mail: moritz.gerstung@ebi.ac.uk; John.Dick@uhnresearch.ca; BrennanP@iarc.fr; gsv20@sanger.ac.uk; liranshlush3@gmail.com

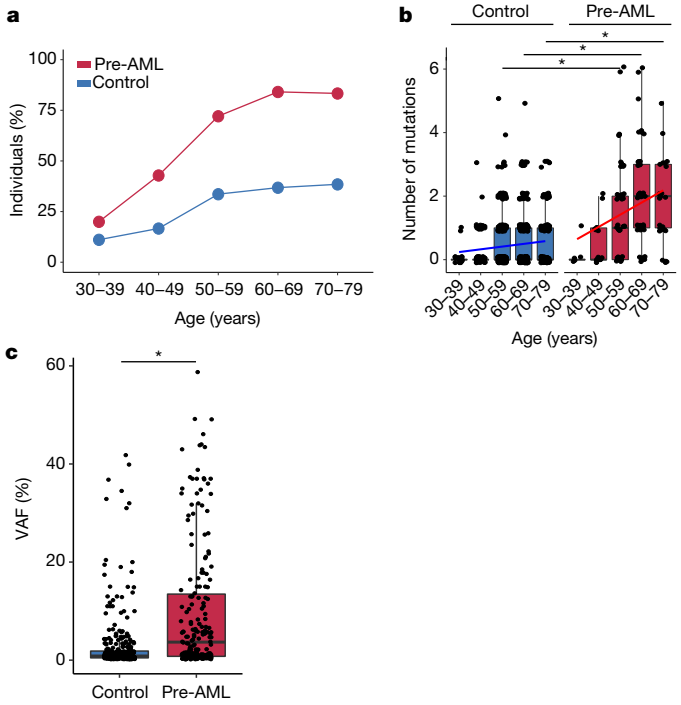

Fig. 1 | Prevalence of ARCH, number of mutations and clone size in individuals who developed AML. a, Prevalence of ARCH-PD among pre-AML cases (red) and controls (blue). **b**, The number of ARCH-PD mutations detected in cases and controls according to age. Box plot centres, hinges and whiskers represent the median, first and third quartiles and 1.5× interquartile range, respectively. Individual values are indicated as dots. **c**, VAF of ARCH-PD mutations. *$P < 0.0005$, two-sided Wilcoxon rank-sum test with Bonferroni multiple testing correction. All panels show data for $n = 800$ biologically independent samples.

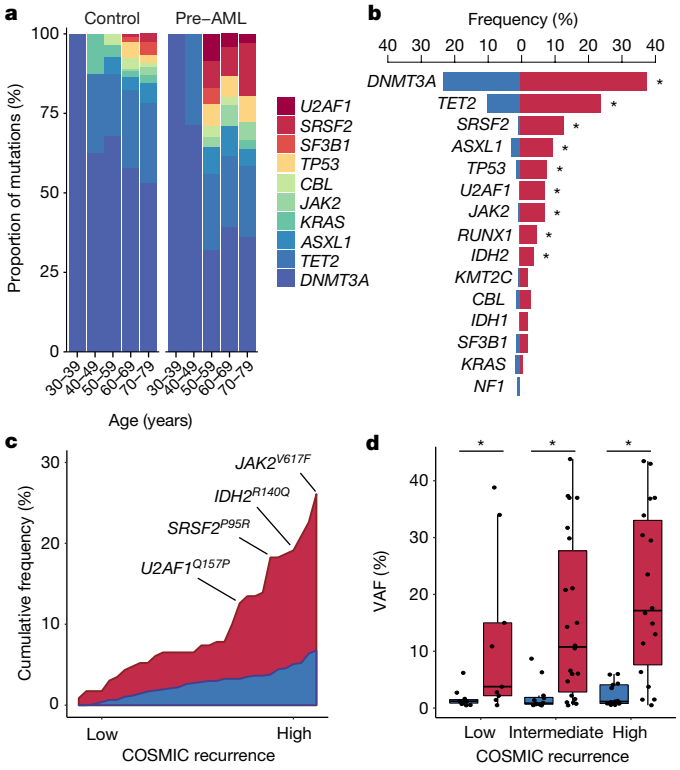

Fig. 2 | Accumulation of specific recurrent AML mutations in healthy individuals at a young age is associated with progression to AML. a, Relative frequency of mutations in the indicated genes according to age group for pre-AML cases and controls. **b**, Proportion of pre-AML cases (red) and controls (blue) who had ARCH-PD mutations in recurrently mutated genes. *$P < 0.05$, Fisher's exact test with Bonferroni multiple testing correction. **c**, The cumulative frequency of recurrent AML mutations (reported in >5 specimens in COSMIC) in pre-AML cases and controls. ARCH-PD mutations are ranked from left to right along the *x* axis from low to high recurrence. **d**, VAF of recurrent mutations in pre-AML cases and controls. Low, intermediate and highly recurrent COSMIC mutations are defined as those reported in 5–19 samples, 20–300 samples and >300 samples, respectively. Box plots indicate median, first and third quartiles and 1.5× interquartile range. *$P < 0.05$, two-sided Wilcoxon rank-sum test with Bonferroni multiple testing correction. All panels show data for $n = 800$ unique individuals.

a prevalence that is in line with the largest studies of ARCH in the general population[4] ($P < 2.2 \times 10^{-16}$, two-sided Fisher's exact test; Extended Data Fig. 1).

The median number of ARCH-PD mutations per individual increased with age and was significantly higher in the pre-AML group relative to controls (Fig. 1b and Supplementary Table 2). Furthermore, examination of ARCH-PD VAF distribution revealed significantly larger clones among the pre-AML cases ($P = 1.2 \times 10^{-13}$, two-sided Wilcoxon rank-sum test; Fig. 1c). To gain insight into clonal growth dynamics, we examined serially collected samples that were available for a subset of the validation cohort. We did not find significant differences in clonal expansion rates between pre-AML cases and controls (Extended Data Fig. 2a, b), although this may in part reflect the shorter follow-up of pre-AML cases, small sample size and large variance in growth rates (Extended Data Fig. 2c). The observed differences between pre-AML cases and controls may arise through cell-intrinsic or -extrinsic factors. Although these variables have not been adequately studied in ARCH, a number of observations in different contexts, such as aplasia, advanced age and after chemotherapy, have shown that increased clonal fitness is associated with distinct mutations depending on context[10–12]. Notably, mutations in splicing factor genes were significantly enriched among the pre-AML cases relative to the controls (odds ratio, 17.5; 95% confidence interval, 8.1–40.4; $P = 5.2 \times 10^{-16}$, two-sided Fisher's exact test) and were present in significantly younger individuals (median age 60.3 compared to 77.3 years, $P = 1.7 \times 10^{-4}$, two-sided Wilcoxon rank-sum test; Fig. 2a). Previous work suggests that spliceosome mutations appear to confer a competitive advantage in the context of ageing[10]. Therefore, it is possible that the significantly higher prevalence of such clones in younger pre-AML cases may reflect extrinsic selection pressures rather than earlier mutation acquisition.

In line with previous reports[5,6], we found that *DNMT3A* and *TET2* were the most commonly mutated genes in both groups (Fig. 2b). We could not identify any canonical *NPM1* mutations nor any *FLT3*-internal tandem duplication mutations, consistent with these arising late in leukaemogenesis[10,13]. Recurrent *CEBPA* mutations, which are implicated in around 10% of de novo AML[14], were also absent, suggesting that driver events in this gene may also be late events in AML evolution. In order to quantify the effect of different mutations on the likelihood of progression to AML, we ranked ARCH-PD mutations based on the number of times that they have been reported in Catalogue of Somatic Mutations in Cancer (COSMIC) database among individuals with haematological malignancies[15]. We found that mutations that are highly recurrent in cancer specimens were more common in pre-AML cases than in controls with ARCH-PD, whereas driver events in the controls tended to affect loci that are less frequently mutated in haematological malignancies and occurred at significantly lower VAF (Fig. 2c, d). Overall, these findings demonstrate notable differences in the mutational landscape of ARCH and pre-AML. Moreover, this work, in conjunction with recent insights into the origins of AML relapse[16], suggests that AML progression typically occurs over many years through clonal evolution of pre-leukaemic HSPCs before acquisition of late mutations leads to overt malignant transformation.

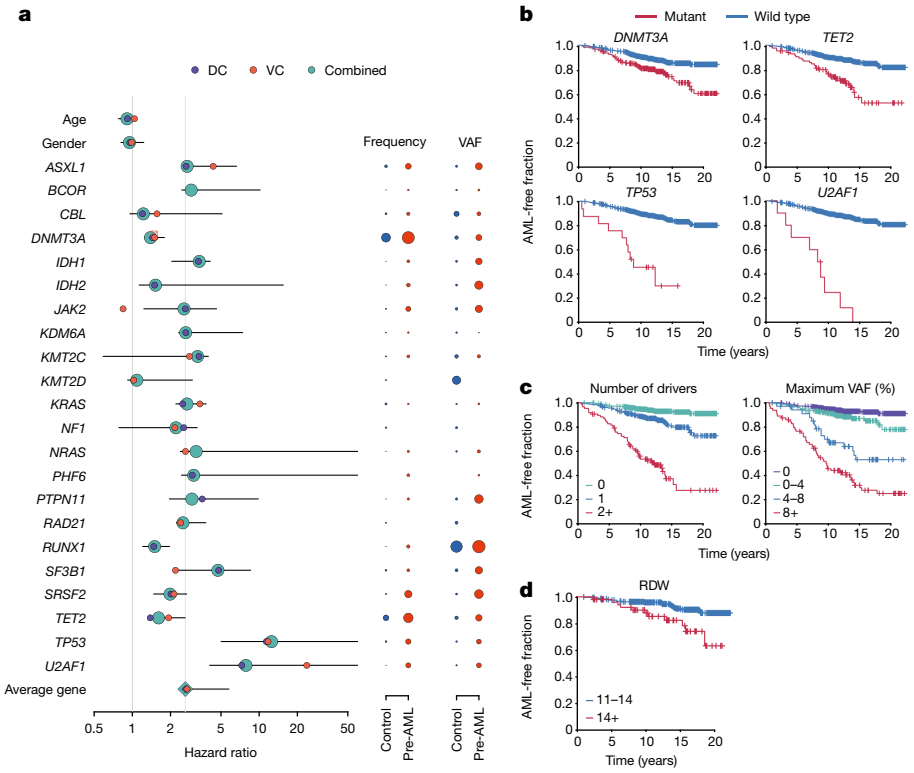

Fig. 3 | Model of future risk of AML. a, Forest plot of the risk of AML. Purple, orange and green circles indicate hazard ratios for the discovery (DC), validation (VC) and combined cohort, respectively. The horizontal lines denote 95% confidence intervals for the combined cohort. For each gene, the indicated hazard ratio applies to the 10-year risk of AML development conferred by each 5% increase in mutation VAF. The green vertical line indicates the mean hazard ratio across all genes. The hazard ratio for *RUNX1* must be interpreted with caution owing to the relatively high prevalence of deleterious germline variants in this gene, which may not be readily distinguishable from somatic mutations in unmatched sequencing assays (see Methods). The proportion of individuals with mutations in each gene and the average VAF are indicated to the right of the forest plot; red and blue circles represent pre-AML cases and controls, respectively, with circle sizes scaled to reflect mutation frequency and VAF. **b**–**d**, Kaplan–Meier curves of AML-free survival, defined as the time between sample collection and AML diagnosis, death or last follow-up. Survival curves are stratified according to mutation status for selected genes (**b**), number of driver mutations per individual and largest clone detected (**c**) and RDW (**d**). Data for $n = 796$ unique individuals (**a**–**c**); $n = 299$ individuals for whom RDW measurements were available (**d**).

On the basis of these findings, we next developed an approach to quantify the relative contributions of driver mutations and clone sizes to the risk of progressing to AML. We tested different regularised logistic and Cox proportional hazards regression approaches, which achieved similar performance in both the discovery cohort (concordance (C) $= 0.77 \pm 0.03$) and the validation cohort ($C = 0.84 \pm 0.05$; Extended Data Figs. 3, 4 and Supplementary Table 3). Models that were only trained on data from the discovery or validation cohort had similar coefficients (Fig. 3a). We therefore combined the datasets for a more accurate analysis of the contributions of mutations in individual genes to risk ($C = 0.77 \pm 0.05$; area under curve, 0.79; Supplementary Table 3). Quantitatively, we found that driver mutations in most genes conferred an approximately twofold increased risk of developing AML per 5% increase in clone size (Fig. 3a and Supplementary Table 3). Notable exceptions to this trend are the most frequently mutated ARCH genes, *DNMT3A* and *TET2*, which confer a lower risk of progression to AML (Fig. 3a, b and Supplementary Table 3). By contrast, a larger effect size was apparent for *TP53* (hazard ratio, 12.5; 95% confidence interval, 5.0–160.5) and *U2AF1* (hazard ratio, 7.9; 95% confidence interval, 4.1–192.2) mutations (Fig. 3a, b). However, we note that other ARCH-PD genes, such as *SRSF2*, can contribute a similar relative risk owing to their presence at a higher VAF in pre-AML cases (Fig. 3a, Extended Data Fig. 5a and Supplementary Note). Of note, mutations in *TP53* and spliceosome genes (including *U2AF1*) are also associated with a poorer prognosis in AML[14]. Because the effect of each ARCH-PD mutation is deleterious and the effect of multiple mutations that are present in the same individual is multiplicative, a higher number of mutations is predicted to increase the risk of progression to AML (Fig. 3c). Similarly, the size of the largest driver clone was also strongly associated with the risk of progression to AML, in agreement with the risk of individual mutations generally being proportional to VAF (Fig. 3c). Collectively, although the VAF and the number of mutations confer much of the predictive value, this model does demonstrate distinct gene-level risk factors, and is able to quantify the cumulative impact of multiple mutations and clonal size on the likelihood of progression to AML.

Although our predictive model performs well in identifying those at risk of developing AML in our experimental cohorts, AML incidence rates in the general population are low (4:100,000)[1], and thus millions of individuals would need to be screened to identify the few pre-AML cases, with many false positives. We therefore sought to determine whether routinely available clinical information could improve prediction accuracy or identify a high-risk population for targeted genetic screening. We first analysed complete blood count and biochemistry data that were available for 37 of the pre-AML cases and 262 controls. As reported previously[5,10,17], ARCH-PD was overwhelmingly associated with normal blood counts and this was also the case for pre-AML cases, indicating that these did not represent undiagnosed myelodysplastic syndrome[18]. We identified a significant association between higher red blood cell distribution width (RDW) and risk of progression to AML ($P = 0.0016$, Wald test with Bonferroni multiple-testing correction, Fig. 3d). Although traditionally used in the evaluation of anaemia, raised RDW has been correlated with inflammation, ineffective erythropoiesis, cardiovascular disease and adverse outcomes in several inflammatory and malignant conditions[19]. The correlation between RDW and risk of AML development remained highly significant when controls without ARCH-PD were excluded

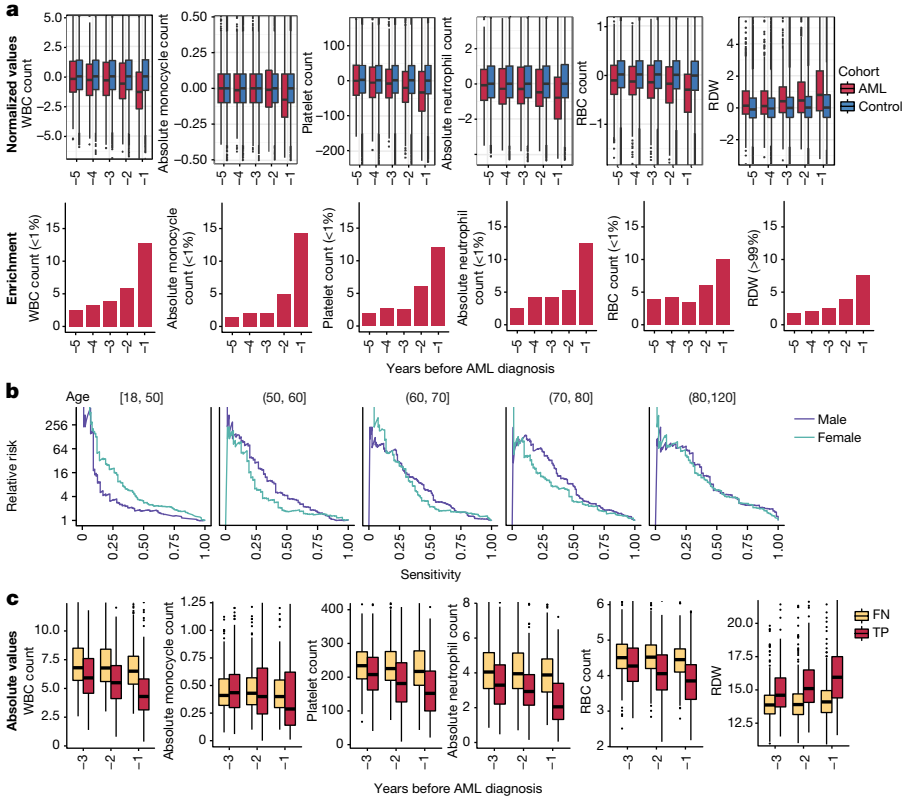

Fig. 4 | Increased risk of AML development inferred from electronic health records. a, Box plot of normalized laboratory measurements. Increased RDW, reduction in monocyte, platelet, red blood cell (RBC) and white blood cell (WBC) counts (top) show a high association (bottom) with a higher risk of AML development and differed at least a year before AML diagnosis. **b**, Model performance stratification by age and gender. Age ranges are indicated above each graph. **c**, Absolute laboratory values for true positive (TP) and false negative (FN) predictions. Box plots indicate median, first and third quartiles and 1.5× interquartile range.

from the analysis ($P = 3.5 \times 10^{-6}$, Wald test with Bonferroni multiple testing correction; Extended Data Fig. 5b). Higher RDW has previously been associated with ARCH and overall mortality[5], but has never been shown to distinguish ARCH from pre-leukaemia. In order to verify RDW as a predictive factor and determine whether additional clinical parameters are associated with risk of AML development, we studied the Clalit database[20], which contains electronic health records that include an average of 3.45 million individuals per year and data that were collected over a 15-year period[21]. We identified 875 cases with AML using stringent criteria based on diagnostic codes and treatment records (Extended Data Fig. 6 and Supplementary Table 4). Analysis of RDW trends revealed significantly raised measurements several years before AML diagnosis relative to age and sex-matched controls (Fig. 4a). Additional parameters that correlated with risk of AML development included reductions in monocyte, platelet, red blood cell and white blood cell counts, albeit usually remaining above the thresholds for clinically relevant cytopenias[18] (Fig. 4a and Extended Data Fig. 7). These findings suggest that evolving de novo AML may sometimes have a considerable prodrome with subtle but discernible clinical manifestations. We next applied a machine-learning approach to construct an AML prediction model based entirely on variables that are routinely documented in electronic health records (Extended Data Fig. 8 and Supplementary Table 4). This model was able to predict AML 6–12 months before diagnosis with a sensitivity of 25.7% and overall specificity of 98.2%. The model performed consistently across different age groups with an increased relative risk of 28 and 24 for males and females, respectively, between the age of 60 and 70 years (Fig. 4b). To better understand which patients are most likely to be accurately classified by this model, we compared absolute laboratory values for true positives and false negatives. We found that 35.5% of false-negative predictions were for patients for whom infrequent blood count data were available (Extended Data Fig. 9). Some of the true-positive cases had mildly abnormal blood counts that would not initiate a diagnostic work-up (Fig. 4c), and cytopenias that would be compatible with undiagnosed myelodysplastic syndrome[18] were uncommon.

Collectively, our findings provide new insights into the pre-clinical evolution of AML and support the hypothesis that individuals at high risk of AML development can be identified years before they develop overt disease. To this end, we present two distinct models for the prediction of de novo AML: one based on somatic point mutations and the other on routinely documented clinical information. We find that basic clinical and laboratory data can identify a high-risk subgroup 6–12 months before AML presentation, while genetic information can identify a substantial fraction of cases several years to more than a decade before diagnosis. By characterizing features that distinguish benign ARCH from pre-leukaemia, our models give valuable insights into leukaemogenesis. It is evident from the current study, together with our recent analysis of mutation acquisition from pre-leukaemic development through to relapse[16], that long-term pre-leukaemic HSPCs frequently carry mutations and undergo considerable clonal expansion while retaining differentiation capacity for years before AML diagnosis. Furthermore, it is clear that some mutations, particularly those affecting *TP53* and *U2AF1*, impart a relatively high risk of subsequent AML, whereas mutations in other genes, for example *DNMT3A* and *TET2*, confer a lesser risk of malignant transformation. Previous studies suggest that oncogenic mutations in *TP53* and spliceosome genes confer little or no competitive advantage in the absence of particular selective pressures[11,22], indicating that cell-extrinsic factors may be important determinants of clonal trajectory.

Cancer predictive models have enabled successful early detection and intervention programmes for several solid tumours[23–25]. However, screening tests are unavailable for the sub-clinical stages of most haematological malignancies. Our study provides proof-of-concept for the feasibility of early detection of healthy individuals at high risk

of developing AML, and is a first step in the design of future clinical studies to investigate the potential benefits of early interventions in this deadly disease. However, the infrequency of AML necessitates that future screening tests provide high sensitivity and specificity. Our findings suggest that basic clinical data may identify a higher risk population that might benefit from targeted genetic screening. Equally, combining clinical and genetic information in a single model and including structural driver events is likely to improve model accuracy further. Nevertheless, establishing the utility of such a tandem approach will require extensive clinical and genetic analysis on the same population cohort, in a prospective setting. Furthermore, ARCH is associated with several non-malignant conditions[4,5], and may have a causal role in cardiovascular disease[26,27]. Therefore, genetic testing for ARCH may also prove useful in the management of common age-related diseases. Moreover, this study has broader implications for cancer screening and early intervention beyond AML. Advances in sequencing technologies have revealed a remarkable degree of somatic genetic diversity in normal ageing tissues, often characterized by the presence of clones that have canonical oncogenic mutations[28]. The degree to which clones at high risk of malignant transformation can be reliably distinguished from their indolent counterparts is an important biological question with compelling clinical ramifications. Understanding the selective pressures and cell-intrinsic mechanisms governing clonal fate is the next important step in developing strategies to predict and prevent progression to overt malignancy.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0317-6

Received: 9 July 2017; Accepted: 3 May 2018;
Published online 9 July 2018.

1. Deschler, B. & Lübbert, M. Acute myeloid leukemia: epidemiology and etiology. *Cancer* **107**, 2099–2107 (2006).
2. Corces-Zimmerman, M. R., Hong, W. J., Weissman, I. L., Medeiros, B. C. & Majeti, R. Preleukemic mutations in human acute myeloid leukemia affect epigenetic regulators and persist in remission. *Proc. Natl Acad. Sci. USA* **111**, 2548–2553 (2014).
3. Shlush, L. I. et al. Identification of pre-leukaemic haematopoietic stem cells in acute leukaemia. *Nature* **506**, 328–333 (2014).
4. Genovese, G. et al. Clonal hematopoiesis and blood-cancer risk inferred from blood DNA sequence. *N. Engl. J. Med.* **371**, 2477–2487 (2014).
5. Jaiswal, S. et al. Age-related clonal hematopoiesis associated with adverse outcomes. *N. Engl. J. Med.* **371**, 2488–2498 (2014).
6. Xie, M. et al. Age-related mutations associated with clonal hematopoietic expansion and malignancies. *Nat. Med.* **20**, 1472–1478 (2014).
7. Busque, L. et al. Nonrandom X-inactivation patterns in normal females: lyonization ratios vary with age. *Blood* **88**, 59–65 (1996).
8. Shlush, L. I. Age-related clonal hematopoiesis. *Blood* **131**, 496–504 (2018).
9. Acuna-Hidalgo, R. et al. Ultra-sensitive sequencing identifies high prevalence of clonal hematopoiesis-associated mutations throughout adult life. *Am. J. Hum. Genet.* **101**, 50–64 (2017).
10. McKerrell, T. et al. Leukemia-associated somatic mutations drive distinct patterns of age-related clonal hemopoiesis. *Cell Rep.* **10**, 1239–1245 (2015).
11. Wong, T. N., et al. Role of *TP53* mutations in the origin and evolution of therapy-related acute myeloid leukaemia. *Nature* **518**, 552–555 (2015).
12. Yoshizato, T. et al. Somatic mutations and clonal hematopoiesis in aplastic anemia. *N. Engl. J. Med.* **373**, 35–47 (2015).
13. Krönke, J. et al. Clonal evolution in relapsed *NPM1*-mutated acute myeloid leukemia. *Blood* **122**, 100–108 (2013).
14. Papaemmanuil, E. et al. Genomic classification and prognosis in acute myeloid leukemia. *N. Engl. J. Med.* **374**, 2209–2221 (2016).
15. Forbes, S. A. et al. COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Res.* **45**, D777–D783 (2017).
16. Shlush, L. I. et al. Tracing the origins of relapse in acute myeloid leukaemia to stem cells. *Nature* **547**, 104–108 (2017).
17. Buscarlet, M. et al. *DNMT3A* and *TET2* dominate clonal hematopoiesis and demonstrate benign phenotypes and different genetic predispositions. *Blood* **130**, 753–762 (2017).
18. Arber, D. A. et al. The 2016 revision to the World Health Organization classification of myeloid neoplasms and acute leukemia. *Blood* **127**, 2391–2405 (2016).
19. Hu, L. et al. Prognostic value of RDW in cancers: a systematic review and meta-analysis. *Oncotarget* **8**, 16027–16035 (2017).
20. Balicer, R. D. & Afek, A. Digital health nation: Israel's global big data innovation hub. *Lancet* **389**, 2451–2453 (2017).
21. Dagan, N., Cohen-Stavi, C., Leventer-Roberts, M. & Balicer, R. D. External validation and comparison of three prediction tools for risk of osteoporotic fractures using data from population based electronic health records: retrospective cohort study. *Br. Med. J.* **356**, i6755 (2017).
22. McKerrell, T. & Vassiliou, G. S. Aging as a driver of leukemogenesis. *Sci. Transl. Med.* **7**, 306fs38 (2015).
23. Vickers, A. J. Prediction models in cancer care. *CA Cancer J. Clin.* **61**, 315–326 (2011).
24. Cassidy, A. et al. The LLP risk model: an individual risk prediction model for lung cancer. *Br. J. Cancer* **98**, 270–276 (2008).
25. Wang, X., Oldani, M. J., Zhao, X., Huang, X. & Qian, D. A review of cancer risk prediction models with genetic variants. *Cancer Inform.* **13**, 19–28 (2014).
26. Fuster, J. J. et al. Clonal hematopoiesis associated with TET2 deficiency accelerates atherosclerosis development in mice. *Science* **355**, 842–847 (2017).
27. Jaiswal, S. et al. Clonal hematopoiesis and risk of atherosclerotic cardiovascular disease. *N. Engl. J. Med.* **377**, 111–121 (2017).
28. Martincorena, I. & Campbell, P. J. Somatic mutation in cancer and normal cells. *Science* **349**, 1483–1489 (2015).

Acknowledgements This work was supported by a Quest for Cure grant to L.I.S., J.C.Y.W. and M.D.M. from the Leukemia and Lymphoma Society, and the following grants to L.I.S from: ERC Horizon 2020 MAMLE, Abisch-Frenkel foundation and an American Society of Hematology Scholar Award. Further funding to J.E.D. was provided by the Canada Research Chair Program, Ontario Institute for Cancer Research, the province of Ontario, Canadian Cancer Society, the Canadian Institutes for Health Research and the Ontario Ministry of Health and Long Term Care to UHN, whose views are not expressed here. Work conducted at the Sanger Institute was supported by the Wellcome Trust and UK Medical Research Council. S.A. was personally funded by the Benjamin Pearl fellowship from the McEwen Centre for Regenerative Medicine, G.C. by a Wellcome Trust Clinical PhD Fellowship (WT098051); G.S.V. by a Wellcome Trust Senior Fellowship in Clinical Science (WT095663MA) and a Cancer Research UK Senior Cancer Research Fellowship (C22324/A23015). G.S.V.'s laboratory is also funded by the Kay Kendall Leukaemia Fund and Bloodwise. We thank A. Mitchell and all members of the Dick and Shlush laboratories for comments and T. Hudson for early study planning; G. Barabash for organising the Clalit dataset collaboration. The EPIC study centres were supported by the Hellenic Health Foundation, Regional Government of Asturias, the Regional Government of Murcia (no. 6236), the Spanish Ministry of Health network RTICCC (ISCIII RD12/0036/0018), FEDER funds/European Regional Development Fund (ERDF), "a way to build Europe", Generalitat de Catalunya, AGAUR 2014SGR726; EPIC Ragusa in Italy-Aire-Onlus Ragusa; Epic Italy-Associazione Italiana per la Ricerca sul Cancro (AIRC) Milan, Italy. S.V.B. and T.J.P. are supported by the Gattuso-Slaight Personalized Cancer Medicine Fund at the Princess Margaret Cancer Centre.

Reviewer information *Nature* thanks R. Levine, P. Van Loo and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions S.W.K.N., O.W., N.M.C. and E.N. contributed equally to the work. S.A. performed error-corrected sequencing, analysed sequencing data, performed statistical analyses, contributed to genetic predictive model derivation and wrote the manuscript. G.C. performed variant calling, statistical analyses, derived genetic predictive models and wrote the manuscript. M.G., S.W.K.N., O.W. and R.C. derived genetic predictive models. N.M.C., E.N. and N.B. derived the clinical prediction model. P.C.Z., Z.Z., I.C., K.N., C.L., C.H., D.H., F.M., J.E., J.K.M., D.P., L.T., P.K., S.V.B. and A.Br. and A.Ba. provided sequencing and technical support and enabled sample acquisition. L.H., Y.S., T.T.W., T.J.P., K.R. and D.J. provided bioinformatics support. R.L., S.H., M.J., K.M.B., A.Kr. and N.J.W. enabled sample acquisition, clinical data curation and/or provided clinical expertise. D.S., J.D.M., P.A., E.S., S.B., P.Be., M.D.M and I.M. contributed to data analysis and interpretation. P.J.C. and E.P. contributed to data interpretation and designed the targeted sequencing assay for the validation cohort. J.C.Y.W. revised the manuscript. J.R.Q., A.Ka., C.L.V., A.T., E.S.-F., J.M.H., R.C.T., R.T., G.M., H.B., S.Pa., R.K., S.S., S.Po., N.J.W., N.S., K.-T.K., M.F., J.M.K., E.R., P.V. and R.V. enabled sample acquisition (EPIC). A.T. and R.D.B. analysed Clalit data and derived the clinical prediction model. M.G. derived predictive genetic models, contributed to sequencing data analysis and manuscript writing. J.E.D. contributed to funding applications, study supervision and manuscript writing. P.Br. supervised sample acquisition from all EPIC centres. G.S.V. and L.I.S. designed and supervised all aspects of the study and wrote the manuscript.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0317-6.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0317-6.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to M.G., J.E.D., P.B., G.S.V. or L.I.S.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Data reporting. No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

Study participants. Samples for both the discovery and validation cohort were obtained from participants in the EPIC study[29]. All relevant ethical regulations were followed. Written informed consent was obtained from all participants in accordance with the Declaration of Helsinki and protocols were approved by the relevant ethics committees (IARC Ethics Committee approval #14-31, the Weizmann Institute of Science Ethics board approval #60-1 and East of England–Cambridgeshire and Hertfordshire Research Ethics Committee reference number 98CN01). Patients with AML were identified based on the following ICD9 codes: 9861/3, 9860/3, 9801/3, 9866/3, 9891/3, 9867/3, 9874/3, 9840/3, 9872/3, 9895/3, 9873/3, which included only cases of de novo AML, and no secondary AML. All patients provided peripheral blood samples for which the buffy coat fractions were separated and aliquoted for long-term storage in liquid nitrogen before DNA extraction.

Discovery cohort. In total, 509 DNA samples were collected from individuals upon enrolment into the EPIC study between 1993 and 1998 across 17 different centres[29] (Supplementary Table 1). Altogether, 95 individuals who developed AML an average of 6.3 years (interquartile range (IQR) = 4.8 years) after the sample was collected were included in the pre-AML group. For the control group, 414 age- and gender-matched individuals were selected, as they did not develop any haematological disorders during the average follow-up period of 11.6 years (IQR = 2.1 years). The median age at recruitment was 56.7 years (range, 36.08–74.42). In order to minimize any possible demographic biases, an approximate 1:4.5 pre-AML to control ratio was maintained across the different centres.

Validation cohort. Samples were obtained from individuals enrolled in the EPIC-Norfolk longitudinal cohort study between 1994 and 2010. Samples and clinical metadata were available from 37 patients with AML (of which 8 were already included in the discovery cohort) and 262 age- and gender-matched controls without a history of cancer or any haematological conditions. The average time between the first blood sampling and AML diagnosis was 10.5 years (IQR = 8.3 years). The average follow-up period for the control cohort was 17.5 years (IQR = 3.8). For 12 individuals in the pre-AML cohort, 2–3 blood specimens were available, taken a median of 3.4 years apart. Of the 262 controls, 141 had multiple blood samples available, spanning a median of 10.5 years. Blood counts and other clinical parameters were available for all study participants (Supplementary Table 1).

Targeted sequencing. *Discovery cohort sequencing.* Targeted deep sequencing was performed using error-corrected sequencing as follows.

Shearing of genomic DNA, preparation of pre-capture sequencing libraries, hybridization-based enrichment, assessment of the libraries quality and enrichment following hybridization were performed as previously described[30]. In brief, 100 ng of genomic DNA was sheared before library construction (KAPA Hyper Prep Kit KK8504, Kapa Biosystems) with a Covaris E220 instrument using the recommended settings for 250-bp fragments. Following end repair and A-tailing, adaptor ligation was performed using 100-fold molar excess of Molecular Index Adaptor. Library clean-up was performed with Agencourt AMPure XP beads (Beckman-Coulter) and the ligated fragments were then amplified for eight cycles using 0.5 μM Illumina universal and indexing primers.

Targeted capture was carried out on pools containing three indexed libraries. Each pool of adaptor-ligated DNA was combined with 5 μl of 1 mg ml^{-1} Cot-I DNA (Invitrogen), and 1 nmol each of xGEN Universal Blocking Oligo, TS-p5, and xGen Universal Blocking Oligo, TS-p7 (8 nucleotides). The mixture was dried using a SpeedVac and then re-suspended in 1.1 μl water, 8.5 μl NimbleGen 2× hybridization buffer and 3.4 μl NimbleGen hybridization component A. The mixture was heat denatured at 95 °C for 10 min before addition of 4 μl of xGen Lockdown Probes (xGen AML Cancer Panel v.1.0, 3 pmol). Each pool was then hybridized at 47 °C for 72 h. Washing and recovery of the captured DNA was performed according to the manufacturer's specifications. In brief, 100 μl of clean streptavidin beads was added to each capture. Following separation and removal of the supernatant using a magnet, 200 μl 1× Stringent Wash Buffer was added and the reaction was incubated at 65 °C for 5 min. The supernatant containing unbound DNA was removed before repeating the high stringency wash one additional time. Then, the bound DNA was washed as follows: (1) 200 μl 1× Wash Buffer I and separation of the supernatants by magnetic separation; (2) 200 μl 1× Wash Buffer II after magnetic separation; (3) 200 μl 1× Wash Buffer III and removal of the supernatants using magnetic separation. The captured DNA on beads was resuspended in 40 μl of Nuclease-Free water before dividing the total volume into two PCR tubes and subjecting the libraries to 10 cycles of post-capture amplification (manufacturer-recommended conditions; Kapa Biosystems). Before sequencing, libraries were spiked with 2% PhiX.

Validation cohort sequencing. Targeted sequencing was performed using a custom complementary RNA bait set (SureSelect, Agilent, ELID 0537771) designed complementary to all coding exons of 111 genes that have been implicated in myeloid leukaemogenesis (Extended Data Table 1). Genomic DNA was extracted from peripheral whole blood and sheared using the Covaris M220. Equimolar pools of 10 libraries were prepared and sequenced on the Illumina HiSeq 2000 using 75-bp paired-end sequencing as per Illumina and Agilent SureSelect protocols.

Variant calling. *Discovery cohort variant calling and error correction.* The 126-bp paired-end reads sequencing data from the Illumina platform were converted to FASTQ format, the 2-bp molecular barcode information at each read of the pair was trimmed and was written in the reads' name. The thymine nucleotide required for ligation was removed from the sequences. Burrows–Wheeler aligner (BWA-mem)[31] was used for alignment of the processed FASTQ files to the reference hg19 genome, after realignment of insertions and deletions (indels) using GATK[32]. An in-house algorithm was written to collapse read families that share the same molecular barcode sequence, the left-most genomic position of where each read of the pair maps to the reference and the CIGAR string. Families that consisted of at least two reads were used to generate consensus reads and a consensus base was called when there was at least 70% agreement. When a consensus base was called, it was assigned with the maximum base quality score observed in its corresponding pre-collapsed reads. Furthermore, when possible, duplex reads[33] were generated from two consensus reads, from a singleton read and a consensus read, or from two singleton reads. For each sequenced sample, we generated two BAM files, called BAM1 and BAM2. BAM1 consisted of duplex reads, consensus reads and singleton reads, thereby including some error-corrected and non-error corrected reads, while still containing all the genomic information encoded in the data in the form of unique DNA molecules. BAM2 consisted of duplex reads and consensus reads but not singleton reads. Both files were then analysed to detect single nucleotide variants (SNVs) and small indels using Varscan2[34]. To further remove sequencing artefacts and improve sensitivity, we applied a two-step polishing statistical approach that models the error rate for each sequenced genomic position. For both steps, BAM1 was used and all samples except the sample that was investigated were included for error rate modelling. At step one, as previously described[30], the error rates were modelled by fitting Weibull distribution curves to the non-reference allele fractions. SNVs with allele fractions that were statistically distinguishable from the background error rates ($P = 0$) were further analysed. At step 2, the coverage of the non-reference allele fractions was considered using linear line fitting that describes the negative correlation that exist between the log(non-reference allele fraction) and the corresponding log(coverage) values. This allowed us to estimate different error rates at different coverage depths. Because indel errors are rare and cannot be appropriately modelled by the same statistical framework, they were called using barcode-mediated error correction alone. At least 10 consensus reads, 5 supporting reads on the forward strand, 5 supporting reads on the reverse strand and 2 duplex reads were required to call an indel. Additional post-processing steps applied to data from both the discovery cohort and validation cohort are detailed in 'Additional post-processing filters applied to discovery and validation cohort data'. Variants were annotated using Annovar[35].

Validation cohort variant calling. Sequencing reads were aligned to the reference genome (GRCh37d5) using the Burrows–Wheeler aligner (BWA-aln)[31]. Unmapped reads, PCR duplicates and reads mapping to regions outside the target regions (merged exonic regions and 10 bp either side of each exon) were excluded from analysis. Sequencing depth at each base was assessed using Bedtools coverage v.2.24.0[36].

Somatic SNVs were called using shearwater, an algorithm developed for detecting subclonal mutations in deep-sequencing experiments (https://github.com/gerstung-lab/deepSNV v.1.21.5)[37–39] considering only reads with minimum nucleotide and mapping quality of 25 and 40, respectively. This algorithm models the error rate at individual loci using information from multiple unrelated samples. Additionally, allele counts at the recurrent AML mutation hotspots listed in 'Curation of oncogenic variants' were generated using an in-house script (https://github.com/cancerit/alleleCount) and manually inspected in the Jbrowse genome browser[40]. To further complement our SNV calling approach, we applied an extensively validated in-house version of CaVEMan v.1.11.2 (Cancer variants through expectation maximization)[41]. CaVEMan compares sequencing reads between study and nominated normal samples and uses a naive Bayesian model and expectation-maximization approach to calculate the probability of a somatic variant at each base (https://github.com/cancerit/CaVEMan).

Post-processing filters required that the following criteria were met for CaVEMan to call a somatic substitution. (1) If coverage of the mutant allele was less than 8, at least one mutant allele was detected in the first two-thirds of the read. (2) Less than 3% of the mutant alleles with base quality ≥15 were found in the nominated normal sample. (3) Mean mapping quality of the mutant allele reads was ≥21. (4) The mutation does not fall in a simple repeat or centromeric region. (5) Fewer than 10% of the reads covering the position contained an indel according to mapping. (6) Less than 80% of the reads report the mutant allele at the same read position. (7) At least a third of the reads calling the variant had a base quality

of 25 or higher. (8) Not all mutant alleles reported in the second half of the read. (9) Position does not fall within a germline insertion or deletion.

The following additional post-processing criteria were applied to all SNV calls. (1) Minimum VAF = 0.5% with a minimum of five bidirectional calls reporting the mutant allele (with at least two reads in forward and reverse directions). (2) No indel called within a read length (75 bp) of the putative substitution.

Small indels were sought using two complementary bioinformatics approaches. First, an in-house version of Pindel v.2.2[42] (https://github.com/cancerit/cgpPindel) was applied. We additionally used the aforementioned deepSNV algorithm in order to increase sensitivity for indels present at low VAF. VAF correction was performed using an in-house script (https://github.com/cancerit/vafCorrect).

Post-processing filters required that the following criteria were met for a variant to be called. (1) A minimum of five reads supporting the variant with a minimum of two reads in each direction. For Pindel, the total read count was based on the union of the BWA and Pindel reads reporting the mutant allele. (2) VAF $\geq$ 0.5%. (3) Variant not present within an unmatched normal panel of approximately 400 samples. (4) No reads supporting the variant identified in the nominated normal sample.

Mutations were annotated according to ENSEMBL v.58 using VAGrENT[43] for transcript and protein effects (https://github.com/cancerit/VAGrENT) and Annovar[35] for additional functional annotation.

Additional post-processing filters applied to discovery and validation cohort data. The following variants were flagged for additional inspection for potential artefacts, germline contamination or index-jumping event. (1) Any mutant allele reported within 75 bp of another variant. (2) Any mutant allele with a population allele frequency >1 in 1,000 according to any of five large polymorphism databases (ExAC, 1000 Genomes Project, ESP6500, CG46 and Kaviar) that is not a canonical hotspot driver mutation with COSMIC recurrence >100. (3) Mutations that were present in >10% of the control cohort but not recurrent in COSMIC were flagged as potential germline variants or sequencing artefacts. (4) As artefactual indels tend to be recurrent, any indels occurring in >2 samples were flagged as for additional inspection.

Curation of oncogenic variants. Putative oncogenic variants were identified according to evidence for functional relevance in AML as previously described and used to define ARCH-PD[14].

Variants were annotated as likely driver events if they fulfilled any of the following criteria. (1) Truncating mutations (nonsense, essential splice site or frameshift indel) in the following genes implicated in AML pathogenesis by loss-of-function: *NF1*, *DNMT3A*, *TET2*, *IKZF1*, *RAD21*, *WT1*, *KMT2D*, *SH2B3*, *TP53*, *CEBPA*, *ASXL1*, *RUNX1*, *BCOR*, *KDM6A*, *STAG2*, *PHF6* and *KMT2C*. (2) Truncating variants in *CALR* exon 9. (3) *JAK2*^{V617F}. (4) *FLT3* internal tandem duplication. (5) Non-synonymous variants at the following hotspot residues: *CBL* E366, L380, C384, C404, R420 and C396; *DNMT3A* R882; *FLT3* D835; *IDH1* R132; *IDH2* R172 and R140; *KIT* W557, V559 and D816; *KRAS* A146, Q61, G13 and G12; *MPL* W515; *NRAS* Q61, G12 and G13; *SF3B1* K700 and K666; *SRSF2* P95; *U2AF1* Q157, R156 and S34. (6) Non-synonymous variants reported at least 10 times in COSMIC with VAF <42% and population allele frequency <0.003. (7) Non-synonymous variants clustering within a functionally validated locus or within four amino acids of a hotspot variant with population allele frequency <0.003 and VAF <42%. (8) Non-synonymous variants reported in COSMIC >100 times with population allele frequency <0.003 regardless of VAF.

Our driver curation strategy inevitably runs a small risk of including germline variants in familial AML genes. We feel that in the real world, where a matched constitutional DNA sample would be unavailable, this is the best approach.

Statistical analysis. All statistical analyses were performed in the R statistical programming environment. A two-sided Wilcoxon rank-sum test was used to assign significance level for differences in the median number of somatic mutations among the pre-AML and control groups, the median VAF of mutations among groups. and the age of individuals with spliceosome mutations. Fisher's exact test was used to assess the significance of differences in the prevalence of ARCH among the groups and spliceosome mutations in the pre-AML group.

Predictive modelling. *Cox proportional hazards model with random effects.* We used a Cox proportional hazards regression to model AML progression-free survival as previously described[14,38]. We used random effects for the Cox proportional hazards model in the CoxHD R package (http://github.com/gerstung-lab/CoxHD). A key strength of this approach is the ability to include many variables in one model while shrinking estimated effects for parameters with weak support in the data, thus controlling for overfitting. We used weighting to minimize the biases introduced by the artificial case–control ratio[44,45] and calculated hazard ratios relative to the (approximate) true cumulative incidence of about 1–3/1,000 in the given age range over a follow up of 10–20 years. The observed driver mutation frequency and VAF in pre-AML cases closely resembled values expected based on the estimated risks, indicating that risk model and driver prevalence are well aligned (Extended Data Fig. 4). Full details of model derivation and comparisons with alternative methods are included in the accompanying code (Supplementary Note, also available at https://github.com/gerstung-lab/preAML). In brief, variables comprised age, gender and the VAF of putative driver mutations (see 'Curation of oncogenic variants' for details of variant curation). We performed agnostic imputation of missing variables by mean and linear rescaling of gene variables by a power of 10 to a magnitude of 1. The model was first trained separately on the discovery cohort and validation cohort. For each of these two models, we evaluated the following measures of predictive accuracy before and after leave-one-out cross-validation (LOOCV): concordance (C)[46] and time-dependent area under the receiver-operating characteristic curve (AUC)[47]. The models trained on the validation and discovery cohorts were then cross-validated using the data from the other cohort. In view of the cross-validation results and close correlation between coefficients (Supplementary Table 3), we derived a model on the combined cohorts using both cohorts in order to achieve greater accuracy on the individual effects. Confidence intervals were calculated using 100 bootstrap samples. The coefficients and performance metrics for each iteration of the model are included in Supplementary Table 3.

Concordance measures were obtained using the survConcordance() function implemented in the survival R package[45]. Dynamic AUC was calculated with AUC.uno() implemented in the survAUC package. Time-independent AUCs were calculated using the performance function implemented in the ROCR package. The expected incidence of AML was calculated from the UK office of national statistics, available at http://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/leukaemia-aml/incidence. All-cause mortality data was obtained from the office of national statistics (https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/lifeexpectancies/datasets/nationallifetablesunitedkingdomreferencetables).

Ridge-regularized logistic regression. Using the same covariates as in 'Cox proportional hazards model with random effects', we fitted a ridge-regularized logistic regression model to dichotomised outcome data. While logistic regression is a common choice for case–control analyses, a downside of this approach is the inability to explicitly use time-dependent covariates. The penalty parameter was chosen using LOOCV on the full cohort; this value was then used on the discovery cohort and validation cohort to yield the same scaling of coefficients. Confidence intervals were calculated using 100 bootstrap samples. Fitting was performed using the glmnet R package. AUC as the primary performance metric was calculated using the ROCR R package.

Additional regression models. Two alternative predictive models were developed. Model 1 performs logistic-regression-based predictions using four types of features: gender, age at blood sampling, the sum of the VAFs ARCH-PD reported in COSMIC v.80 to be recurrent (at least two case reports in haematopoietic and lymphoid tissues) and somatic mutation burden of selected genes, where each gene was represented by the sum of the VAFs corresponding to ARCH-PD mutations in that gene. We measured the predictive performance of each gene via the AUC obtained in a fivefold cross-validation when using only the gene as a predictive feature, and only retained genes with AUC > 55% in the final model.

For model 2 we applied LASSO regression as implemented in the glmnet R package, while enabling LOOCV to fit a Cox regression model. A minimal subset of ARCH-PD variants was selected for which the respective weighted combined VAFs were highly predictive of AML development in the training set. Scores were calculated for each patient as a linear combination of VAF of mutations weighted by regression coefficients that were estimated from the training data. As most scores were zero in the training subset, non-zero scores were discretized to take on a value of 1 that corresponds to AML prediction.

Models 1 and 2 were trained on the discovery cohort and tested for their association with AML development using the validation cohort data. Survival analysis was performed using the Kaplan–Meier and Cox proportional hazards models. Wald's test was used to evaluate the significance of hazard ratios. Logistic regression models were used with the positive predictive value metric to determine the ability of various mutations and other patient parameters to predict AML development. The rms R package was used for logistic regression analysis, and the pROC 1.8 R package was used for receiver-operating characteristic curve analysis.

AML-predictive model based on electronic health records. *Clalit database.* The Clalit database includes information from patients covered by the Clalit health services in Israel[20] during the years 2002–2017. The Clalit training-set data, contains the electronic health records (EHR) of 3.45 million individuals per year on average. All data was anonymized through hashing of personal identifiers and addresses and randomization of dates by sampling a random number of weeks for each patient and adding it to all dates in the patient diagnoses, laboratory and medication records. This approach maintained differential data analysis per patient. Diagnoses codes were acquired from both primary care and hospitalization records, and were mapped to the ICD-9 coding system for historical reasons, with few exceptions that used a partial ICD-10 coding system. Laboratory records were normalized for age and gender by subtracting raw test values from the median

levels observed among all test values with matching gender and age (using a bin size of five years). We observed some chronological biases in laboratory ranges, but avoid normalizing these and instead insured case and controls are matched for chronological distributions.

Defining AML cases. We screened for all active patients ($18 < \text{age} < 100$) who were diagnosed with AML (ICD-9 code 205.0*) between the years 2003 and 2016. We then excluded cases based on the following criteria. (1) We excluded patients with prior myeloid malignancies to omit secondary AML, consistent with the case selection for the genetic model. The following diagnosis were excluded if documented within five years before the diagnosis of AML: essential thrombocythemia (ICD-9 238.71), low-grade myelodysplastic syndrome (MDS) (ICD-9 238.72); high-grade MDS lesions (ICD-9 238.73); MDS with 5q deletion (ICD-9 238.74); MDS, unspecified (ICD-9 238.75); polycythemia vera (ICD-9 238.4); myelofibrosis (ICD-9 289.83); chronic myelomonocytic leukaemia (ICD-9 206.10-206.22).

(2) Patients that had any procedures performed on bone marrow or spleen (ICD-10 code Z41) in the five-year period before first mention of AML diagnosis code in their record. These patients were presumed to have an inaccurate AML diagnosis date or misdiagnosis recorded.

(3) Patients that received medications suggestive of an alternative diagnosis of chronic myeloid leukaemia, lymphoid malignancy or acute promyelocytic leukaemia (APL). At any time before diagnosis: imatinib, dasatinib, anagrelide, hydroxycarbamide, asparaginase, pegaspargase or arsenic trioxide. At any time after diagnosis: imatinib, dasatinib, methotrexate, tretinoin or arsenic trioxide. At any time after diagnosis, along with any acute lymphoblastic leukaemia diagnosis (ICD-9 204) or more than single dose: mercaptopurine. APL cases were excluded as early diagnosis of APL will most probably not change its outcome, as treatment is successful already.

(4) Patients without a hospitalization record within three months before or after the onset diagnosis. This parameter was used as it is unlikely that a patient with AML will not be hospitalized close to diagnosis. This filter reduced false-positive cases and better defined the onset date.

We refined the estimated time of onset using the earliest time at which any of the following diagnosis appeared in the patient's history: amyloidosis (ICD-9 277.3), lymphoid leukaemia (ICD-9 204), myeloid leukaemia (ICD-9 205), leukaemia of unspecified cell type (ICD-9 208).

This strategy retained 875 AML cases in the training set for further analysis. These were further validated by manual expert inspection of the complete records of 8% of the cases.

To define the control set, we included all Clalit individuals that were not cases. Since our analysis was aggregating data from a historical time window of 15 years, we associated each control with a randomized time point for evaluation. Using this approach, both cases and controls represented a specific time point in the historical record of a patient, with matching calendric, age and gender distributions. Through this strategy 5,238,528 controls were used.

Defining features for construction of a predictive a score. We extracted the following features for discriminative analysis of cases and controls (this procedure was applied repeatedly in cross-validation as discussed below). (1) Age (in years) at time point. (2) Gender. (3) Laboratory features. Out of 2,770 different types of laboratory tests, we selected the top 50 most frequent laboratory tests (Supplementary Table 4). For each laboratory measurement, we used median age- and gender-normalized test values per patient in three time windows for 6–12 months before onset, 1–2 years before onset and 2–3 years before onset. In addition, we compute the slope of the normalized laboratory measurements for the 6–12 month time window using a linear regression model. (4) Diagnosis features. Of the 1780 different major ICD-9 diagnosis codes, we selected only diagnoses that were previously observed in at least 10 different cases and have an increased relative risk for AML >twofold (as observed in the training set, Supplementary Table 4). For each diagnosis code, we mark whether it appeared in each of the patients in time intervals of 6 months to 3 years, and 3–5 years before onset. (5) BMI features. For each patient in the cohort, we extracted median BMI, weight and height as measured in time intervals of 6 months to 2 years, and 2–3 years before onset.

Gradient boosting. We used the R package xgboost to infer parameters for a classifier given cases and controls. Objective was set to binary:logistic, the evaluation metric to AUC. We set nrounds = 5000, eta = 0.001, gamma = 0.1, lambda = 0.01, alpha = 0.01, max_depth = 6, min_child_weight = 2, subsample = 0.7 and colsample_bytree = 0.7. The boosting algorithm reports a function f that computes a predictive score given the features. Given a threshold T the expression f(patient features) $> T$ defines a classifier. To standardise thresholds we estimate quantiles for the scores on the training set $T(p) = \text{quantile}(f(\text{train}),p)$ and define the classifier for specificity level p as f(patient features) $> T(p)$ (Supplementary Table 4).

Cross-validation and relative risk evaluation. To evaluate the predictive value of the classification scheme while considering the strong age and gender biases in the incidence of AML, we performed fivefold cross-validation after splitting the cases and controls into five age- and gender-matched groups. For each fold, we sampled 100,000 controls and combined with the cases, constructed the feature set and trained the model. The model was then tested on the fold cases along with 200,000 sampled controls. We used standardized classifier parameters and standardized thresholds that were inferred based on each training set to generate a series of classifications on each test set and merged these based on the control quantiles in the test as described above. Given a threshold p to define high and low prediction score, we counted for each bin b that defines a patient in a specific age (<40, 40–50, 50–60, 60–70, 70–80, >80) and gender group: the number of cases in bin b (N^b_{case}) and the number of controls in bin b (N^b_{control}) where N^b is the number of patients in bin b (entire database minus recall controls that are only a sample of the cohort). $N^b(\text{case, high score}) = N^b_{\text{TP}}$ indicates the number of true positives (TP); $N^b(\text{case, low score}) = N^b_{\text{FN}}$ indicates the number of false negatives (FN); $N^b(\text{control, high score}) = N^b_{\text{FP}}$ indicates the number of false positives (FP); $N^b(\text{control, low score}) = N^b_{\text{TN}}$ indicates number of true negatives (TN).

For each age and gender group, the absolute risk for AML in the bin is computed by $r^b_{\text{abs}} = N^b_{\text{case}}/N^b$. The absolute risk given a high score is estimated as $r^b_{\text{abs,high}} = N^b_{\text{TP}}/(N^b_{\text{FP}} + N^b_{\text{TP}})$. The relative risk in the bin is defined by $\text{rr}^b = r^b_{\text{abs,high}}/r^b_{\text{abs}}$ where the sensitivity level for the classifier threshold level is defined as $\text{sense}^b = N^b_{\text{TP}}/N^b_{\text{case}}$.

$$\text{rr} = \frac{\frac{\text{TP} \times \text{cases}}{(\text{TP}+\text{FN})}}{\frac{\text{TP} \times \text{cases}}{(\text{TP}+\text{FN})} + \frac{\text{FP} \times \text{controls}}{(\text{FP}+\text{TN})}} \Big/ \frac{\text{cases}}{\text{cases}+\text{controls}}$$

Clonal growth rate calculation. Individual clones were defined by different mutations in different study participants. Per clone we calculated α according to the following equation:

$$a = \log(V/V_0)\ /(T - T_0)$$

where T and T_0 indicate the age of the individual at the two measurement time points. V and V_0 correspond to the VAF at T and T_0, respectively.

Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

Code availability. Code for derivation of the prediction model is publically available on Github (https://github.com/gerstung-lab/preAML). Code for the analysis of error-corrected sequencing is available from the Shlush lab upon request.

Data availability. Targeted sequencing data for the discovery cohort are deposited as BAM files at the European Genome-phenome Archive (http://www.ebi.ac.uk/ega/) under accession number EGAD00001003583. All other data are available from the corresponding authors upon reasonable request. Sequencing data for the validation cohort are deposited at the European Genome-phenome Archive with accession number EGAD00001003703.

29. Riboli, E. et al. European Prospective Investigation into Cancer and Nutrition (EPIC): study populations and data collection. *Public Health Nutr.* **5**, 1113–1124 (2002).
30. Newman, A. M. et al. An ultrasensitive method for quantitating circulating tumor DNA with broad patient coverage. *Nat. Med.* **20**, 548–554 (2014).
31. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).
32. McKenna, A. et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
33. Kennedy, S. R. et al. Detecting ultralow-frequency mutations by Duplex Sequencing. *Nat. Protoc.* **9**, 2586–2606 (2014).
34. Koboldt, D. C. et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.* **22**, 568–576 (2012).
35. Yang, H. & Wang, K. Genomic variant annotation and prioritization with ANNOVAR and wANNOVAR. *Nat. Protoc.* **10**, 1556–1566 (2015).
36. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
37. Gerstung, M. et al. Reliable detection of subclonal single-nucleotide variants in tumour cell populations. *Nat. Commun.* **3**, 811 (2012).
38. Gerstung, M. et al. Precision oncology for acute myeloid leukemia using a knowledge bank approach. *Nat. Genet.* **49**, 332–340 (2017).
39. Martincorena, I. et al. High burden and pervasive positive selection of somatic mutations in normal human skin. *Science* **348**, 880–886 (2015).
40. Buels, R. et al. JBrowse: a dynamic web platform for genome visualization and analysis. *Genome Biol.* **17**, 66 (2016).
41. Stephens, P. J. et al. The landscape of cancer genes and mutational processes in breast cancer. *Nature* **486**, 400–404 (2012).
42. Raine, K. M. et al. cgpPindel: identifying somatically acquired insertion and deletion events from paired end sequencing. *Curr. Protoc. Bioinformatics* **52**, 15.17.1–15.7.12 (2015).
43. Menzies, A. et al. VAGrENT: Variation Annotation Generator. *Curr. Protoc. Bioinformatics* **52**, 15.18.1–15.18.11 (2015).

44. Antoniou, A. C. et al. A weighted cohort approach for analysing factors modifying disease risks in carriers of high-risk susceptibility genes. *Genet. Epidemiol.* **29**, 1–11 (2005).
45. Therneau, T. & Grambsch P. M. *Modeling Survival Data: Extending the Cox Model* 1st edn (Springer-Verlag, New York, 2000).
46. Harrell, F. E. Jr, Lee, K. L. & Mark, D. B. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat. Med.* **15**, 361–387 (1996).
47. O'Quigley, J., Xu, R. & Stare, J. Explained randomness in proportional hazards models. *Stat. Med.* **24**, 479–489 (2005).

SCIENTIFIC REPORTS

OPEN

An integrated genomic analysis of anaplastic meningioma identifies prognostic molecular signatures

Received: 30 April 2018
Accepted: 16 August 2018
Published online: 10 September 2018

Grace Collord[1,2], Patrick Tarpey[1], Natalja Kurbatova[3], Inigo Martincorena[1], Sebastian Moran[4], Manuel Castro[4], Tibor Nagy[1], Graham Bignell[1], Francesco Maura[1,5,6], Matthew D. Young[1], Jorge Berna[7], Jose M. C. Tubio[7], Chris E. McMurran[8], Adam M. H. Young[8], Mathijs Sanders[1,21], Imran Noorani[1,8], Stephen J. Price[8], Colin Watts[9], Elke Leipnitz[10], Matthias Kirsch[10], Gabriele Schackert[10], Danita Pearson[11], Abel Devadass[11], Zvi Ram[17], V. Peter Collins[11], Kieren Allinson[11], Michael D. Jenkinson[12,20], Rasheed Zakaria[12,13], Khaja Syed[12,13], C. Oliver Hanemann[14], Jemma Dunn[14], Michael W. McDermott[15], Ramez W. Kirollos[8], George S. Vassiliou[1,16], Manel Esteller[4,18,19], Sam Behjati[1,2], Alvis Brazma[3], Thomas Santarius[8] & Ultan McDermott[1,20,22]

Anaplastic meningioma is a rare and aggressive brain tumor characterised by intractable recurrences and dismal outcomes. Here, we present an integrated analysis of the whole genome, transcriptome and methylation profiles of primary and recurrent anaplastic meningioma. A key finding was the delineation of distinct molecular subgroups that were associated with diametrically opposed survival outcomes. Relative to lower grade meningiomas, anaplastic tumors harbored frequent driver mutations in SWI/SNF complex genes, which were confined to the poor prognosis subgroup. Aggressive disease was further characterised by transcriptional evidence of increased PRC2 activity, stemness and epithelial-to-mesenchymal transition. Our analyses discern biologically distinct variants of anaplastic meningioma with prognostic and therapeutic significance.

[1]Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, CB10 1SA, UK. [2]Department of Paediatrics, University of Cambridge, Cambridge Biomedical Campus, Cambridge, CB2 0QQ, UK. [3]European Molecular Biology Laboratory, European Bioinformatics Institute, EMBL-EBI, Wellcome Trust Genome Campus, Hinxton, CB10 1SD, UK. [4]Cancer Epigenetics and Biology Program (PEBC), Bellvitge Biomedical Research Institute (IDIBELL), L'Hospitalet de Llobregat, Barcelona, Catalonia, Spain. [5]Department of Oncology and Hemato-Oncology, University of Milan, Milan, Italy. [6]Department of Hematology, Fondazione IRCCS Istituto Nazionale dei Tumori, Milan, Italy. [7]Mobile Genomes and Disease, Molecular Medicine and Chronic diseases Centre (CIMUS), Universidade de Santiago de Compostela, Santiago de Compostela, 15706, Spain. [8]Department of Neurosurgery, Department of Clinical Neuroscience, Cambridge University Hospitals NHS Foundation Trust, Cambridge, CB2 0QQ, UK. [9]Department of Neurosurgery, Institute of Cancer and Genomic Sciences, University of Birmingham, Birmingham, UK. [10]Klinik und Poliklink für Neurochirurgie, "Carl Gustav Carus" Universitätsklinikum, Technische Universität Dresden, Fetscherstrasse 74, 01307, Dresden, Germany. [11]Department of Pathology, Cambridge University Hospital, CB2 0QQ, Cambridge, UK. [12]Department of Neurosurgery, The Walton Centre, Liverpool, L9 7LJ, UK. [13]Institute of Integrative Biology, University of Liverpool, Liverpool, L9 7LJ, UK. [14]Institute of Translational and Stratified Medicine, Plymouth University Peninsula Schools of Medicine and Dentistry, Plymouth University, Plymouth, Devon, PL4 8AA, UK. [15]Department of Neurosurgery, UCSF Medical Center, San Francisco, CA, 94143-0112, USA. [16]Department of Haematology, Cambridge University Hospitals NHS Trust, Cambridge, CB2 0QQ, UK. [17]Department of Neurosurgery, Tel-Aviv Medical Center, Tel-Aviv, Israel. [18]Physiological Sciences Department, School of Medicine and Health Sciences, University of Barcelona (UB), Catalonia, Spain. [19]Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Catalonia, Spain. [20]Institute of Translational Medicine, University of Liverpool, Liverpool, L9 7LJ, UK. [21]Erasmus University Medical Center, Department of Hematology, Rotterdam, The Netherlands. [22]Present address: AstraZeneca, CRUK Cambridge Institute, Robinson Way, Cambridge, CB2 0RE, UK. Grace Collord and Patrick Tarpey contributed equally. Correspondence and requests for materials should be addressed to T.S. (email: ts381@cam.ac.uk) or U.M. (email: um1@sanger.ac.uk)

Meningiomas arise from arachnoidal cells of the meninges and are classified as grade I (80% of cases), grade II (10–20%) or grade III (1–3%). Grade III meningiomas comprise papillary, rhabdoid and anaplastic histological subtypes, with anaplastic tumors accounting for the vast majority of grade III diagnoses[1,2]. Nearly half of anaplastic meningiomas represent progression of a previously resected lower grade tumor, whereas the remainder arise *de novo*[3,4]. Recurrence rates are 5–20% and 20–40%, respectively, for grade I and II tumors[2,5]. By contrast, the majority of anaplastic meningioma patients suffer from inexorable recurrences with progressively diminishing benefit from repeated surgery and radiotherapy and 5-year overall survival of 30–60%[4,6].

A recent study of 775 grade I and grade II meningiomas identified five molecular subgroups defined by driver mutation profile[7]. In keeping with previous smaller studies, mutually exclusive mutations in *NF2* and *TRAF7* were the most frequent driver events, followed by mutations affecting key mediators of PI3K and Hedgehog signalling[7,8]. Recurrent hotspot mutations were also identified in the catalytic unit of RNA polymerase II (*POLR2A*) in 6% of grade I tumors[7]. More recently, a study comparing benign versus *de novo* atypical (grade II) meningiomas found the latter to be significantly associated with *NF2* and *SMARCB1* mutations[9]. Atypical meningiomas were further defined by DNA and chromatin methylation patterns consistent with upregulated PRC2 activity, aberrant Homeobox domain methylation and transcriptional dysregulation of pathways involved in proliferation and differentiation[9].

Despite the high mortality rate of anaplastic meningiomas, efforts to identify adjuvant treatment strategies have been hampered by a limited understanding of the distinctive molecular features of this aggressive subtype. A recent analysis of meningioma methylation profiles identified distinct subgroups within Grade III tumors predictive of survival outcomes, though the biology underpinning these differences and any therapeutic implications remain unknown[10]. Here, we present an analysis of the genomic, transcriptional and DNA methylation patterns defining anaplastic meningioma. Our results reveal molecular hallmarks of aggressive disease and suggest novel approaches to risk stratification and targeted therapy.

Results

Overview of the genomic landscape of primary and recurrent anaplastic meningioma. We performed whole genome sequencing (WGS) on a discovery set of 19 anaplastic meningiomas resected at first presentation ('primary'). A subsequent validation cohort comprised 31 primary tumors characterised by targeted sequencing of 366 cancer genes. We integrated genomic findings with RNA sequencing and methylation array profiling in a subset of samples (Supplementary Table S1). Somatic copy number alterations and rearrangements were derived from whole genome sequencing reads, with RNA sequences providing corroborating evidence for gene fusions. Given the propensity of anaplastic meningioma to recur, we studied by whole genome sequencing 13 recurrences from 7 patients.

Excluding a hypermutated tumor (PD23359a, see Supplementary Discussion), the somatic point mutation burden of primary anaplastic meningioma was low with a median of 28 somatic coding mutations per tumor (range 11 to 71; mean sequencing coverage 66X) (Supplementary Fig. S1). Mutational signatures analysis of substitutions identified in whole genome sequences revealed the age-related, ubiquitous processes 1 and 5 as the predominant source of substitutions (Supplementary Fig. S2)[11]. The rearrangement landscape was also relatively quiet, with a median of 12 structural rearrangements (range 0–79) in the 18 primary tumor genomes (Supplementary Fig. S3, Table S3). Somatic retrotransposition events, a significant source of structural variants in over half of human cancers, were scarce (Supplementary Fig. S4, Table S4)[12]. Analysis of expressed gene fusions did not reveal any recurrent events involving putative cancer genes (Supplementary Table S5).

Recurrent large copy number changes were in keeping with known patterns in aggressive meningiomas, notably frequent deletions affecting chromosomes 1p, 6q, 14 and 22q (Fig. 1b, Supplementary Table S6)[7,9,13].

Driver genes do not delineate subgroups of anaplastic meningioma. Over 80% of low grade meningiomas segregate into 5 distinct subgroups based on driver mutation profile[7,9]. In anaplastic meningioma, however, we found a more uniform driver landscape dominated by deleterious mutations in *NF2* (Fig. 1a). A key feature distinguishing anaplastic meningioma from its lower grade counterparts were driver events in genes of the SWI/SNF chromatin regulatory complex (Fig. 1a; Supplementary Fig. S7). The SWI/SNF (mSWI/SNF or BAF) complex is the most commonly mutated chromatin-regulatory complex in cancer[14,15], and acts as a tumor suppressor in many cell types by antagonising the chromatin modifying PRC2[16–18]. The most frequently mutated SWI/SNF component was *ARID1A*, which harbored at least one deleterious somatic change in 12% of our cohort of 50 primary tumors (Supplementary Table S1). *ARID1A* has not been implicated as a driver in grade I or grade II meningiomas[7,9]. Single variants in *SMARCB1*, *SMARCA4* and *PBRM1* were also detected in three tumors (Supplementary Fig. S7). In total, 16% of anaplastic meningiomas contained a damaging SWI/SNF gene mutation. By contrast, SWI/SNF genes are mutated in $<5\%$ of benign and atypical meningiomas[7,9].

In the combined cohort of 50 primary tumors, we found at least one driver mutation in *NF2* in 70%, similar to the prevalence reported in atypical meningiomas and more than twice that found in grade I tumors[7,9]. As observed in other cancer types, it is possible that non-mutational mechanisms may contribute to *NF2* loss of function in a proportion of anaplastic meningiomas[19,20]. We considered promoter hypermethylation as a source of additional *NF2* inactivation, but found no evidence of this (Supplementary Table S7). There was no significant difference in NF2 expression between *NF2* mutant and wild-type tumors (p-value 0.960; Supplementary Fig. S8), suggesting that a truncated dysfunctional protein may be expressed.

Other driver genes commonly implicated in low grade tumors were not mutated, or very infrequently (Fig. 1a). Furthermore, and consistent with the most recent reports[7,9], we did not observe an increased frequency of *TERT* promoter mutations, previously associated with progressive or high grade tumors[21]. Notably[13], methylation analysis revealed *CDKN2A* and *PTEN* promoter hypermethylation in 17% and 11% of primary tumors, respectively (Fig. 1a). We did not find evidence of novel cancer genes in our cohort, applying established methods

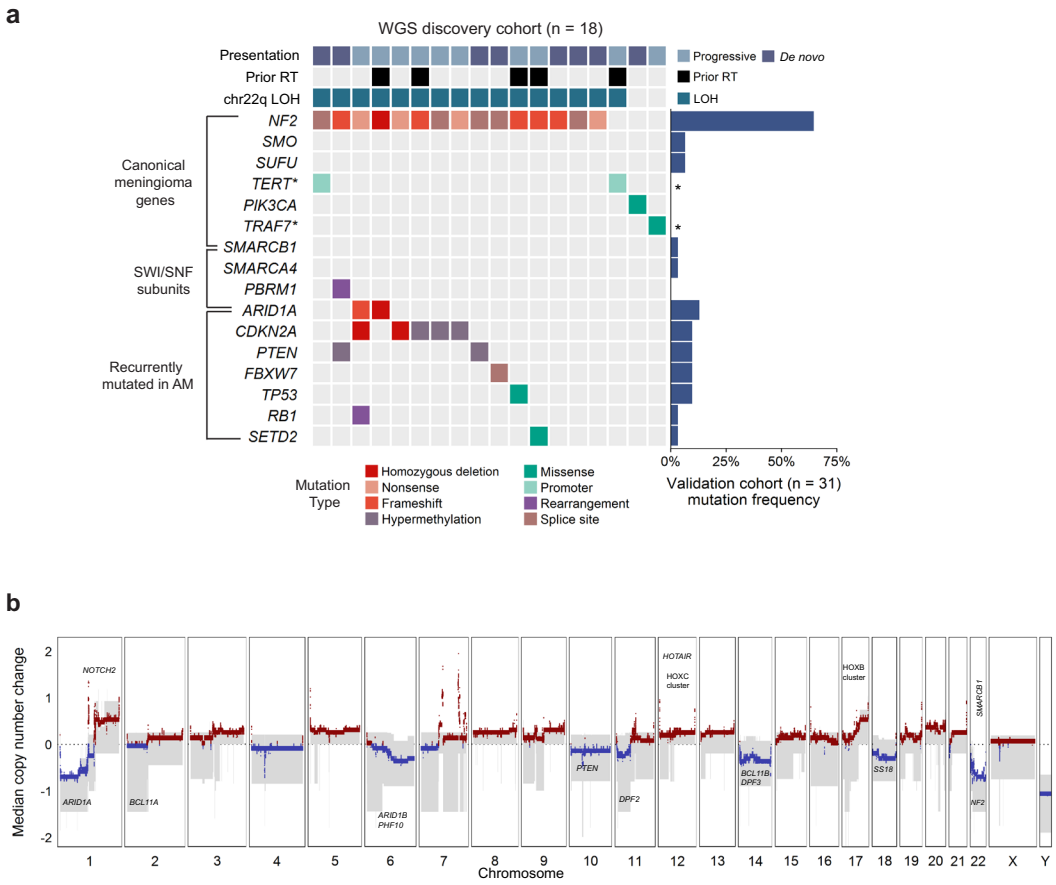

Figure 1. The landscape of driver mutations and copy number alterations in anaplastic meningioma. (**a**) The landscape of somatic driver variants in primary anaplastic meningioma. Somatic mutation and promoter methylation data is shown for a discovery cohort of 18 primary tumors characterised by whole genome sequencing. Mutations in recurrently altered genes, established meningioma genes and SWI/SNF complex subunits are included. Samples are annotated for chromosome 22q LOH, prior radiotherapy exposure, and clinical presentation (*de novo* verus progression from a lower grade meningioma). The bar plot to the right indicates mutation frequency in a validation cohort of 31 primary tumors sequenced with a 366 cancer gene panel. Asterisks indicate genes not included in the targeted sequencing assay. (**b**) Aggregate copy number profile of primary anaplastic meningioma. For the 18 tumors characterized by whole genome sequencing, the median relative copy number change was calculated across the genome in 10 kilobase segments, adjusting for ploidy. The grey shaded area indicates the first and third quantile of copy number for each genomic segment. The solid red and blue lines represent the median relative copy number gain and loss, respectively, with zero indicating no copy number change. X-axis: Chromosomal position. Y-axis: median relative copy number change. Potential target genes are noted. AM, anaplastic meningioma; LOH, loss of heterozygosity; RT, radiotherapy.

to search for enrichment of non-synonymous mutations[22]. The full driver landscape of anaplastic meningioma, considering point mutations, structural variants with resulting copy number changes and promoter hypermethylation is presented in Supplementary Fig. S7.

The genomic landscape of recurrent tumors was largely static both with respect to driver mutations and structural variation. Driver mutations differed between primary and recurrent tumors for only two of eleven patients with serial resections available. For seven sets of recurrent tumors studied by whole genome sequencing, only two demonstrated any discrepancies in large copy number variants (PD23344 and PD23346; Supplementary Fig. S5). Similarly, matched primary and recurrent samples clustered closely together by PCA of transcriptome data, suggesting minimal phenotypic evolution (Supplementary Fig. S6).

Differential gene expression defines anaplastic meningioma subgroups with prognostic and biological significance. We performed messenger RNA (mRNA) sequencing of 31 anaplastic meningioma samples from a total of 28 patients (26 primary tumors and 5 recurrences). Gene expression variability within the cohort did not correlate with clinical parameters including prior radiotherapy, anatomical location or clinical presentation (*de novo* versus progressive tumor) (Supplementary Fig. S6). However, unsupervised hierarchical clustering demonstrated segregation of tumors into two main groups, hereafter referred to as C1 and C2 (Fig. 2a). These groups were recapitulated by principal component analysis (PCA) of normalised transcript counts (Fig. 2b), which delineated C1 as a well-demarcated cluster clearly defined by the first two principal components

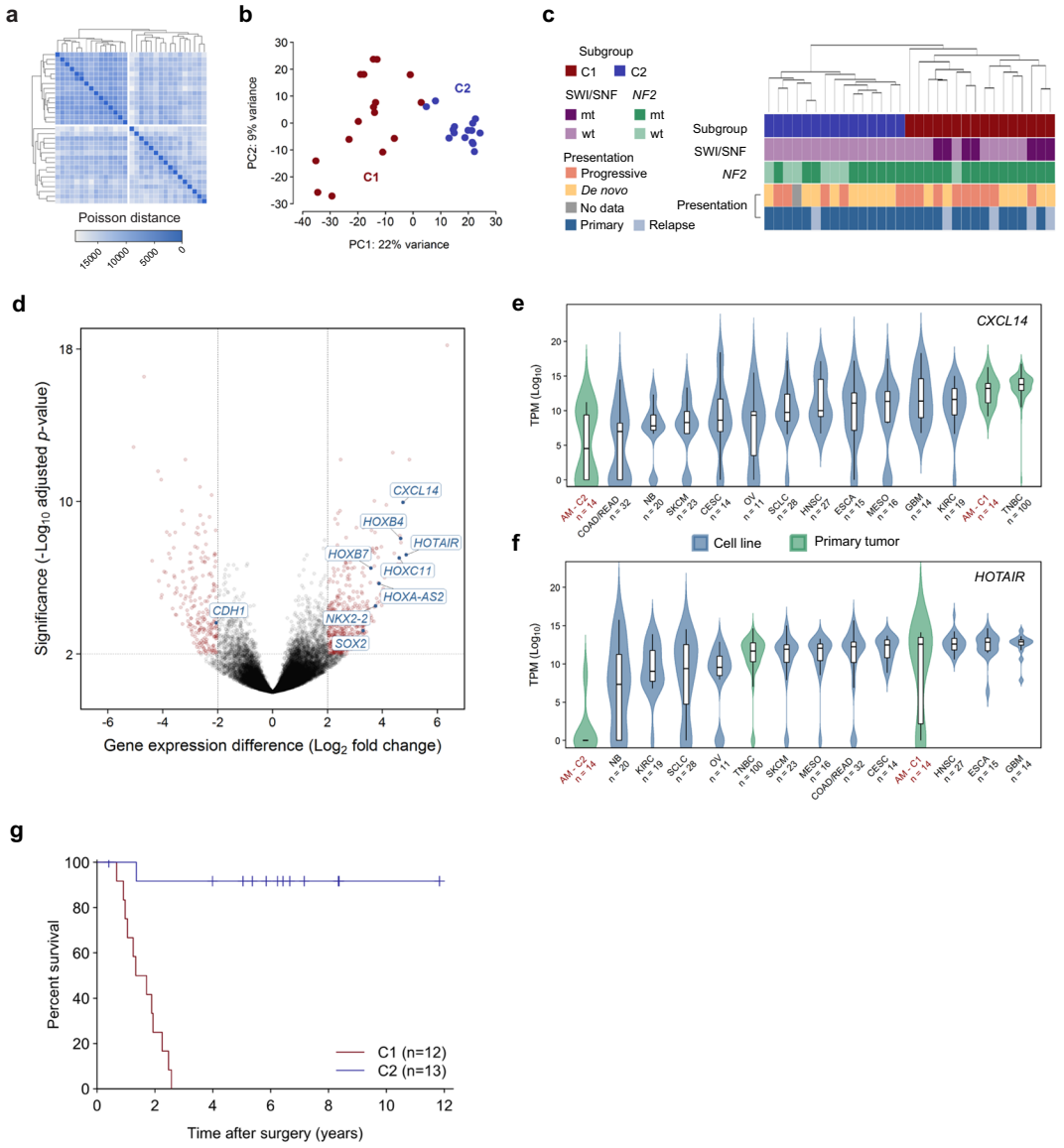

Figure 2. Transcriptomic classification of anaplastic meningioma. (**a**) Unsupervised hierarchical clustering and (**b**) principal component analysis of anaplastic meningioma gene expression revealed two subgroups (denoted C1 and C2). (**c**) Dendrogram obtained by unsupervised clustering annotated with clinical and genomic features. (**d**) Volcano plot depicting genes differentially expressed between C1 versus C2 anaplastic meningioma samples. X-axis, $\log_2$ fold change; y-axis, $-\log_{10}$ adjusted *P*-value. Genes with an adjusted *P*-value < 0.01 and absolute $\log_2$ fold change >2 are highlighted in red. (**e**,**f**) Box plots of (**e**) CXLC14 and (**f**) HOTAIR expression across 31 anaplastic meningomas classified into C1 and C2 subgroups, 100 primary breast tumors, and 219 cancer cell lines from 11 tumor types. Upper and lower box hinges correspond to first and third quartiles, horizontal line and whiskers indicate the median and 1.5-fold the interquartile range, respectively. Underlying violin plots show data distribution and are color-coded according to specimen source (blue, cell line; green, primary tumor). X-axis indicates tumor type and number of samples; y-axis shows $\log_{10}$ TPM values. (**g**) Kaplan-Meier curves showing overall survival for 25 anaplastic meningioma patients in C1 and C2 subgroups for whom follow-up data was available. Dashes indicate timepoints at which subjects were censored at time of last follow-up. TPM, transcripts per kilobase million; AM, anaplastic meningioma; TNBC, triple negative breast carcinoma; wt, wild-type; mt, mutated; PC, principal component.

(PC). Of note, all SWI/SNF mutations were confined to the poor prognosis (C1) subgroup (Fig. 2c). C1 constituted a more diffuse group on PCA, distinguished from C2 mainly along the first principal component. We next retrospectively sought follow-up survival data from the time of first surgery, which was available for 25 of the 28 patients included in the transcriptome analysis (12 patients in C1, 13 in C2; mean follow-up of 1,403 days from surgery). We observed a significantly worse overall survival outcome in C1 compared to C2 ($P < 0.0001$; hazard ratio 17.0, 95% CI 5.2–56.0) (Fig. 2g; Supplementary Table S8). The subgroups were well balanced with respect

to potential confounding features such as gender, age, radiotherapy, anatomical location and amount of residual tumor remaining after surgery (Supplementary Table S9).

Recent work has demonstrated that anaplastic meningiomas segregate into 2–3 prognostically significant subgroups on the basis of methylation profile[10]. Unsupervised hierarchical clustering using methylation data available for a subset of the cohort (n = 19) demonstrated segregation into two main groups largely overlapping the subgroups delineated on the basis of gene expression profile, though correlation with survival outcomes was less marked (Supplementary Fig. S8).

Transcriptional programs segregating indolent and aggressive anaplastic meningioma. Nineteen hundred genes underpinned the differentiation of anaplastic meningioma into subgroups C1 and C2, which could be reduced to only 6 transcripts selected on the basis of PCA coefficient and differential expression analysis (see Methods; Supplementary Tables S10 and S11, Fig. S9). Pathway enrichment analysis was most significant for evidence of epithelial-mesenchymal transition (EMT) in the C1 tumors, with concordant loss of E-cadherin (*CDH1*) and upregulation of *CXCL14*, both prognostic biomarkers in diverse other cancers (Supplementary Table S12, Fig. 2d–f)[23–25]. EMT, which involves reprogramming of adherent epithelial cells into migratory mesenchymal cells, is critical for embryogenesis and tissue plasticity, and can play an important role in malignant progression, metastasis and therapy resistance[24,26]. Interestingly, NF2 and the closely related cytoskeletal protein ezrin normally help maintain E-cadherin expression at adherence junctions, whereas *HOXB7* and *HOXB9*, both overexpressed in C1 tumors, suppress *CDH1* expression[27–29]. It is increasingly recognised that CXCL14 and other EMT mediators are often derived from cancer-associated fibroblasts (CAFs) and function in a paracrine manner[25,30,31]. It is hence possible that some of the gene expression patterns we observed may reflect differences in the tumor stromal compartment, itself an increasingly recognised therapeutic target[30,32,33].

The C1 tumors were further characterised by upregulation of transcriptional programs associated with increased proliferation, PRC2 activity and stem cell phenotype (Supplementary Table S13). Hox genes constituted a notable proportion of the transcripts distinguishing the two anaplastic meningioma subgroups, largely underpinning the significance of pathways involved in tissue morphogenesis. Furthermore, differentially methylated genes were also significantly enriched for Hox genes, with pathway analysis results corroborating the main biological themes apparent from the transcriptome (Supplementary Tables S14 and S15). Given the transcriptional evidence of increased PRC2 activity in the C1 subgroup, is noteworthy that SWI/SNF gene mutations occurred exclusively in C1 tumors ($P = 0.016$, Fisher's exact test).

Comparison of the anaplastic and benign meningioma transcriptome. Previous studies investigating the relationship between meningioma WHO grade and gene expression profiles have included few anaplastic tumors[34,35]. We therefore extended our analysis to include published RNA sequences from 19 benign grade I meningiomas. External data was processed using our in-house pipeline with additional measures taken to minimise batch effects (Methods, Supplementary Tables S16 and S17). Unsupervised hierarchical clustering and principal component analysis demonstrated clear tumor segregation by histological grade (Fig. 3a,b). In keeping with previous reports, the anaplastic tumors demonstrated marked upregulation of major growth factor receptor and kinase circuits implicated in meningioma pathogenesis, notably epidermal growth factor receptor (EGFR), insulin-like growth factor (IGFR), vascular endothelial growth factor receptor (VEGFR) and mTOR complex 1 (mTORC1) kinase complex[36–41].

Consistent with there being a coherent biological trend across histological grades and anaplastic meningioma subgroups, we noted significant overlap between genes differentially expressed between grades and between C1 and C2 tumors (hypergeometric distribution $P = 5.08 \times 10^{-9}$). In keeping with this finding, formal pathway analysis identified significant dysregulation of stemness, proliferation, EMT and PRC2 activity (Supplementary Tables S18 and S19). The most significantly dysregulated pathways also included TGF-beta, Wnt and integrin signalling, mediators of invasion and mesenchymal differentiation that are normally in part controlled by NF2 and other Hippo pathway members[20,24,42]. Yes-associated protein 1 (Yap1), a cornerstone of oncogenic Hippo signalling, is frequently overexpressed in cancer and synergises with Wnt signalling to induce EMT[43,44]. *YAP1* was upregulated in anaplastic tumors along with *MYL9*, a key downstream effector essential for Yap1-mediated stromal reprogramming (Fig. 3c)[43].

Discussion

Meningiomas constitute a common, yet diverse tumor type with few therapeutic options[6,7,9,45]. Efforts to improve clinical outcomes have been hampered by limited understanding of the molecular determinants of aggressive disease. Here, we explored genomic, epigenetic and transcriptional features of anaplastic meningioma, the most lethal meningioma subtype[4].

Frequent somatic changes in SWI/SNF complex genes, predominantly *ARID1A*, constitute the main genomic distinction between anaplastic and lower grade meningiomas[7,9]. SWI/SNF inactivation is associated with aberrant PRC2 activation, stem cell-like phenotype and poor outcomes in diverse cancer types[46–48].

Although anaplastic tumors resist comprehensive classification based on driver mutation patterns, transcriptional profiling revealed two biologically distinct subgroups with dramatically divergent survival outcomes. This finding is emblematic of the limitations of histopathological grading as a risk stratification system for meningioma[2,4,10,45,49]. All SWI/SNF mutations were confined to the poor prognosis (C1) subgroup, which was further characterised by transcriptional signatures of PRC2 target activation, stemness, proliferation and mesenchymal differentiation. These findings were in part underpinned by differential expression of Hox genes. Acquisition of invasive capacity and stem cell traits are frequently co-ordinately dysregulated in cancer, often through subversion of Hox gene programs integral to normal tissue morphogenesis[50–52]. Hox genes have a central role in orchestrating vertebrate development and act as highly context-dependent oncogenes and tumor suppressors in cancer[51,53].

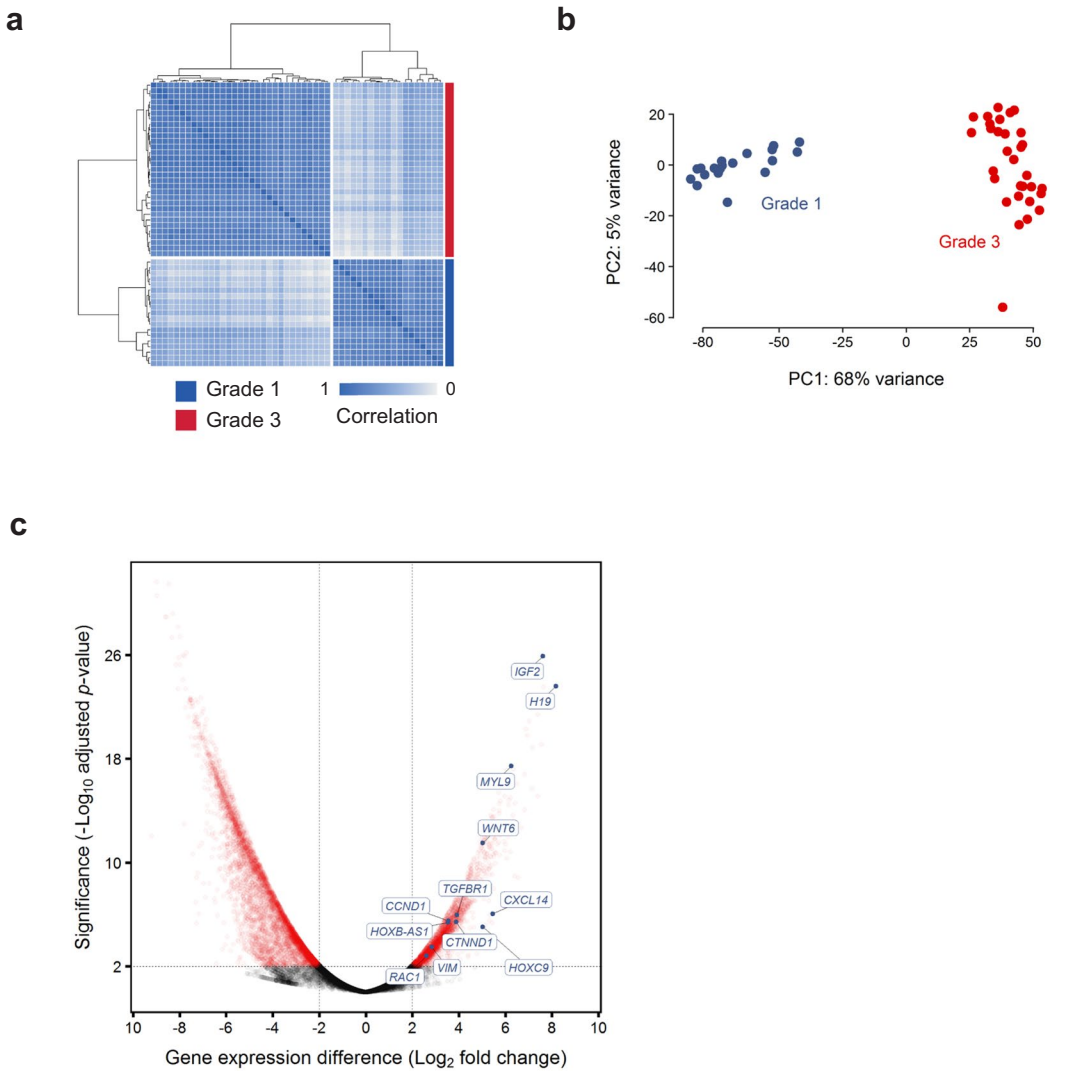

Figure 3. Differences in gene expression profile between grade I and anaplastic meningomas. (**a**,**b**) Normalised transcript counts from grade I and anaplastic meningioma samples clustered by (**a**) Pearson's correlation coefficient and (**b**) principal component analysis. (**c**) Volcano plot illustrating differences in gene expression between anaplastic versus grade I meningiomas with selected genes indicated. The horizontal axis shows the $\log_2$ fold change and the vertical axis indicates the $-\log_{10}$ adjusted *P*-value. Genes with an adjusted *P*-value < 0.01 and absolute $\log_2$ fold change >2 are highlighted in red. PC, principal component.

Several of the most starkly upregulated Hox genes in the C1 tumors consistently function as oncogenes across a range of solid and haematological malignancies, including *HOTAIR*, *HOXB7*, *HOXA4*, *HOXA-AS2*, *HOXC11*, and *NKX2-2*[28,29,51,54–62]. Like many other long non-coding RNAs (lncRNA), *HOTAIR* and *HOXA-AS2* modulate gene expression primarily by interacting directly with chromatin remodelling complexes, exerting oncogenic activity by recruiting PRC2 to target genes[54,56,61–65]. *HOXA-AS2* has been shown to mediate transcriptional repression of the tumor suppressor gene *CDKN2A* ($p16^{INK4A}$), deletion of which is associated with poor meningioma survival[54,61,62,66,67]. Given the antagonistic relationship between the SWI/SNF and PRC2 chromatin regulators, deleterious SWI/SNF mutations and overexpression of lncRNAs known to mediate PRC2 activity emerge as potentially convergent mechanisms underpinning the differences between C1 and C2 tumors[68]. Further endorsing a link between transcriptional subgroups and chromatin dysregulation, 15 of the differentially expressed transcripts delineating C1 and C2 subgroups (absolute $\log_2$ fold change >2 and FDR < 0.01) are among the 50 genes most often associated with frequently bivalent chromatin segments (FBS) in cancer, including 11 transcripts from the HOXB cluster on chromosome 17[69]. This overlap was highly statistically significant (hypergeometric distribution $P = 1.98 \times 10^{-11}$). Bivalent, or epigenetically 'poised', chromatin is characterised by finely balanced activating (H3K4me1/H3K4me3) and repressive (H3K27me3) histone marks and pre-loaded DNA polymerase II poised to transcribe in response to modest epigenetic changes[70]. Bivalent chromatin most often marks genes involved in developmental reprogramming, in particular Hox cluster genes and homeotic non-coding transcripts, and is a frequent target of aberrant chromatin modification in cancer[65,69,71].

In the context of recent studies of lower grade meningiomas, our findings raise the possibility that the balance between PRC2 and SWI/SNF activity may have broader relevance to meningioma pathogenesis. Compared to grade I tumors, atypical meningiomas are more likely to harbor *SMARCB1* mutations and large deletions encompassing chromosomes 1q, 6q and 14q. Notably, these genomic regions encompass *ARID1A* and several other SWI/SNF subunit genes. Both *SMARCB1* mutations and the aforementioned copy number changes were associated with epigenetic evidence of increased PRC2 activity, differential Homeobox domain methylation, and upregulation of proliferation and stemness programs in atypical grade II meningiomas[9].

The extent to which SWI/SNF depletion plays a role in meningioma development may be therapeutically relevant. Diverse SWI/SNF mutated cancers exhibit dependence on both catalytic and non-catalytic functions of EZH2, a core subunit of PRC2[72–74]. Several EZH2 inhibitors are in development with promising initial clinical results[75]. Other modulators of PRC2 activity, including *HOTAIR*, may also be relevant therapeutic targets[76,77]. Furthermore, growing recognition of the relationship between EMT and resistance to conventional and targeted anti-cancer agents has profound implications for rational integration of treatment approaches[32,33]. Notably, EGFR inhibition has yielded disappointing response rates in meningioma despite high EGFR expression[37,78]. A mesenchymal phenotype is strongly associated with resistance to EGFR inhibitors in lung and colorectal cancer[32,33,79–81]. Combining agents that abrogate EMT with other therapies is a promising strategy for addressing cell-autonomous and extrinsic determinants of disease progression and may warrant further investigation in meningioma[32,33].

This study has revealed biologically and prognostically significant anaplastic meningioma subgroups and identified potentially actionable alternations in SWI/SNF genes, PRC2 activity and EMT regulatory networks. However, a substantially larger series of tumors, ideally nested in a prospective multicentre observational study, will be required to expand upon our main findings and explore mechanistic and therapeutic ramifications of meningioma diversity.

Methods

Sample selection. DNA was extracted from 70 anaplastic meningiomas; 51 samples at first resection ('primary') and 19 from subsequent recurrences. Matched normal DNA was derived from peripheral blood lymphocytes. Written informed consent was obtained for sample collection and DNA sequencing from all patients in accordance with the Declaration of Helsinki and protocols approved by the NREC/Health Research Authority (REC reference 7/YH/0101) and Ethics Committee at University Hospital Carl Gustav Carus, Technische Universität Dresden, Germany (EK 323122008). Samples underwent independent specialist pathology review (V.P.C and K.A). DNA extracted from fresh-frozen material was submitted for whole genome sequencing whereas that derived from formalin-fixed paraffin-embedded (FFPE) material underwent deep targeted sequencing of 366 cancer genes.

One tumor sample PD23348 (and two subsequent recurrences) separated from the main study samples in a principal components analysis of transcriptomic data (Supplementary Fig. S10). Analysis of WGS and RNA sequencing data identified an expressed gene fusion, *NAB2-STAT6*. This fusion is pathognomonic of meningeal hemangiopericytoma, now classified as a separate entity, solitary fibrous tumors[82–84]. We therefore excluded three samples from this tumor from further study. A second sample (PD23354a), diagnosed as an anaplastic meningioma with papillary features, was found to have a strong APOBEC mutational signature as well as an *EML4-ALK* gene fusion (exon 6 EML4, exon 19 ALK) (Supplementary Fig. S11)[85]. Therefore this sample was also removed as a likely metastasis from a primary lung adenocarcinoma. The hypermutator sample PD23359a underwent additional pathological review to confirm the diagnosis of anaplastic meningioma (K.A., Department of Histopathology, Cambridge University Hospital, Cambridge, UK).

RNA was extracted from fresh-frozen material from 34 primary and recurrent tumors, 3 of which were from PD23348 and were subsequently excluded from final analyses (Supplementary Table S1).

Whole genome sequencing. Short insert 500 bp genomic libraries were constructed, flowcells prepared and sequencing clusters generated according to Illumina library protocols[86]. 108 base/100 base (genomic), or 75 base (transcriptomic) paired-end sequencing were performed on Illumina X10 genome analyzers in accordance with the Illumina Genome Analyzer operating manual. The average sequence coverage was 65.8X for tumor samples and 33.8X for matched normal samples (Supplementary Table S1).

Targeted genomic sequencing. For targeted sequencing we used a custom cRNA bait set (Agilent) to enrich for all coding exons of 366 cancer genes (Supplementary Table S20). Short insert libraries (150 bp) were prepared and sequenced on the Illumina HiSeq 2000 using 75 base paired-end sequencing as per Illumina protocol. The average sequence coverage was 469X for the tumor samples.

RNA sequencing and data processing. For transcriptome sequencing, 350 bp poly-A selected RNA libraries were prepared on the Agilent Bravo platform using the Stranded mRNA library prep kit from KAPA Biosystems. Processing steps were unchanged from those specified in the KAPA manual except for use of an in-house indexing set. Reads were mapped to the GRCh37 reference genome using STAR (v2.5.0c)[87]. Mean sequence coverage was 128X. Read counts per gene, based on the union of all exons from all possible transcripts, were then extracted BAM files using HTseq (v0.6.1)[88]. Transcripts Per kilobase per Million reads (TPM) were generated using an in-house python script (https://github.com/TravisCG/SI_scripts/blob/master/tpm.py)[87,88]. We downloaded archived RNA sequencing FASTQ files for 19 grade I meningioma samples representing the major mutational groups (*NF2*/chr22 loss, *POLR2A*, *KLF4/TRAF7*, *PI3K* mutant) (ArrayExpress: GSE85133)[7]. Reads were then processed using STAR and HTseq as described above. Cancer cell line (n = 252) and triple-negative breast cancer (n = 100) RNA sequencing data was generated in-house by the aforementioned sequencing and bioinformatic pipeline.

Expressed gene fusions were sought using an in-house pipeline incorporating three algorithms: TopHat-Fusion (v2.1.0), STAR-Fusion (v0.1.1) and deFuse (v0.7.0) (https://github.com/cancerit/cgpRna)[87,89,90]. Fusions identified by one or two algorithms or also detected in the matched normal sample were flagged as likely artefacts. Fusions were further annotated according to whether they involved a kinase or known oncogene and whether they occurred near known fragile sites or rearrangement break points[91] (Supplementary Table S5).

The C1 and C2 subgroups were defined by unsupervised hierarchical clustering using Poisson distance between samples[92,93]. Poisson distance was calculated using the PoissonDistance function implemented in the 'PoiClaClu' R package[92] and unsupervised hierarchical clustering performed with the stats::hclust() function using the 250 transcripts with the most variable expression across tumors. Silhouette information was computed using the cluster::silhouette() function. The highest mean silhouette score was consistently achieved with two clusters.

Differential gene expression and pathway enrichment analysis. The DESeq2 R package was used for all differential gene expression analyses[94,95]. DESeq2 uses shrinkage estimation of dispersion for the sample-specific count normalization and subsequently applies a linear regression method to identify differentially expressed genes (DEGs)[94,95].

Preliminary comparison of anaplastic and externally-generated grade I meningioma data revealed evidence of laboratory batch effects, which we mitigated with two batch-correction methods: RUVg and PEER[96,97]. RUVg estimates the factor attributed to spurious variation using control genes that are assumed to have constant expression across samples[98–100]. We selected control genes (*RPL37A*, *EIF2B1*, *CASC3*, *IPO8*, *MRPL19*, *PGK1* and *POP4*) on the basis of previous studies of suitable control genes for transcript-based assays in meningioma[101]. PEER ('probabilistic estimation of expression residuals') is based on factor analysis methods that infer broad variance components in the measurements. PEER can find hidden factors that are orthogonal to the known covariates. We applied this feature of PEER to remove additional hidden effect biases. The final fitted linear regression model consists of the factor identified by RUVg method that represents the unwanted laboratory batch effect and 13 additional hidden factors found by PEER that are orthogonal to the estimated laboratory batch effect. Using this approach we were able to reduce the number of DEGs from more than 18000 to 8930, of which <4,000 are predicted to be protein-coding.

To identify biological pathways differentially expressed between meningioma grades and anaplastic meningioma subgroups we applied a functional class scoring algorithm using a collection of 461 published gene sets mapped to 10 canonical cancer hallmarks (Supplementary Table S21)[50,102–106]. We further corroborated these findings with a more general Gene Ontology (GO) pathway analysis[107].

Identification of 6 transcripts recapitulating anaplastic meningioma clusters. Mapped RNA sequencing reads were normalised using the regularised logarithm (rlog) function implemented by the DESeq2 package[94,95]. PCA was performed using the top 500 most variably expressed transcripts and the R stats::prcomp function[108]. Given that primary component 1 (PC1) was the vector most clearly distinguishing the closely clustered C2 subgroup from the more diffusely clustered C1 (Fig. 3a), we extracted the top 50 transcripts with the highest absolute PC1 coefficients. We then identified the subset that overlapped with the most significantly differentially expressed genes (absolute $\log_2$ fold change >4 and adjust p-value < 0.0001) between i) the C1 and C2 anaplastic meningioma subgroups and ii) the C1 anaplastic meningiomas and the 19 grade I tumors (Supplementary Tables S10 and S17). Iteratively reducing the number of PC1 components identified the minimum number of transcripts that recapitulated segregation of C1 and C2 tumors upon unsupervised hierarchical clustering and PCA (Supplementary Table S11, Fig. S9).

Processing of genomic sequencing data. Genomic reads were aligned to the reference human genome (GRCh37) using the Burrows-Wheeler Aligner, BWA (v0.5.9)[109]. CaVEMan (Cancer Variants Through Expectation Maximization: http://cancerit.github.io/CaVEMan/) was used for calling somatic substitutions. Small insertions and deletions (indels) in tumor and normal reads were called using a modified Pindel version 2.0. (http://cancerit.github.io/cgpPindel/) on the NCBI37 genome build[110,111]. Annotation was according to ENSEMBL version 58. Structural variants were called using a bespoke algorithm, BRASS (BReakpoint AnalySiS) (https://github.com/cancerit/BRASS) as previously described[112].

The ascatNGS algorithm was used to estimate tumor purity and ploidy and to construct copy number profiles from whole genome data[113].

Identification of cancer genes based on the impact of coding mutations. To identify recurrently mutated driver genes, we applied an established dN/dS method that considers the mutation spectrum, the sequence of each gene, the impact of coding substitutions (synonymous, missense, nonsense, splice site) and the variation of the mutation rate across genes[22].

Identification of driver mutations in known cancer genes. Non-synonymous coding variants detected by Caveman and Pindel algorithms were flagged as putative driver mutations according to set criteria and further curated following manual inspection in the Jbrowse genome browser[114]. Variants were screened against lists of somatic mutations identified by a recent study of 11,119 human tumors encompassing 41 cancer types and also against a database of validated somatic drivers identified in cancer sequencing studies at the Wellcome Trust Sanger Institute (Supplementary Tables S22 and S23)[115].

Copy number data was analysed for homozygous deletions encompassing tumor suppressor genes and for oncogene amplifications exceeding 5 or 9 copies for diploid and tetraploid genomes, respectively. Only focal (<1 Mb) copy number variants meeting these criteria were considered potential drivers. Additional truncating events (disruptive rearrangement break points, nonsense point mutations, essential splice site mutations and

frameshift indels) in established tumor suppressors were also flagged as potential drivers. Only rearrangements with breakpoints able to be reassembled at base pair resolution are included in this dataset.

TraFiC pipeline for retrotransposon integration detection. For the identification of putative solo-L1 and L1-transduction integration sites, we used the TraFiC (Transposome Finder in Cancer) algorithm[12]. TraFiC uses paired-end sequencing data for the detection of somatic insertions of transposable elements (TEs) and exogenous viruses. The identification of somatic TEs (solo-L1, Alu, SINE, and ERV) is performed in three steps: (i) selection of candidate reads, (ii) transposable element masking, (iii) clustering and prediction of TE integration sites and (iv) filtering of germline events[12].

Methylation arrays and analysis. We performed quantitative methylation analysis of 850,000 CpG sites in 25 anaplastic meningiomas. Bisulfite-converted DNA (bs-DNA) was hybridized on the Ilumina Infinium HumanMethylationEPIC BeadChip array following the manufacturer's instructions. All patient DNA samples were assessed for integrity, quantity and purity by electrophoresis in a 1.3% agarose gel, picogreen quantification and Nanodrop measurements. Bisulfite conversion of 500 ng of genomic DNA was done using the EZ DNA Methylation Kit (Zymo Research), following the manufacturer's instructions. Resulting raw intensity data (IDATs) were normalized using the Illumina normalization method developed under the minfi R package (v1.19.10). Normalized intensities were then used to calculate DNA methylation levels (beta values). We then excluded from the analysis the positions with background signal levels in methylated and unmethylated channels ($p > 0.01$). Finally we removed probes with one or more single nucleotide polymorphisms (SNPs) with a minor allele frequency (MAF) >1% in the first 10 bp of the interrogated CpG, as well as the probes related to X and Y chromosomes. From the filtered positions, we selected only CpG sites present both in promoter regions (TSS, 5′UTR and 1st exon) and CpG islands (UCSC database, genome version hg19).

For the supervised analysis of the probes, CpG sites were selected by applying an ANOVA test to identify statistically significant CpG positions (FDR adjusted p-value < 0.01) that were differentially methylated among the compared groups ($\Delta\beta > 0.2$). Selected CpG sites were later clustered based on the Manhattan distances aggregated by ward's linkage. Finally, the genes corresponding to the selected CpGs were used to perform a Gene Set Enrichment Analysis (GSEA) with curated gene sets in the Molecular Signatures Database[116]. The gene sets used were: H: hallmark gene sets, BP: GO biological process, CC: GO cellular component, MF: GO molecular function and C3: motif gene sets (http://software.broadinstitute.org/gsea/msigdb/collections.jsp). The gene clusters resulting from the hypergeometric test with a FDR adjusted p-value < 0.05 were finally considered. We observed high levels of methylation for *CREBBP* in the majority of tumor samples, however, similar patterns were manifest in normal tissue controls, hence *CREBBP* hypermethyation does not appear to be a feature of oncogenesis in these samples.

For principal component analysis, we used the R function prcomp to calculate the Singular Value Decomposition of the beta value matrix after removing the CpGs without methylation information. We plotted the first two principal components which contain most variation by using the ggbiplot R package (http://github.com/vqv/ggbiplot). For each group we plotted a normal data ellipse with size defined as a normal probability equal to 0.68. Unsupervised hierarchical clustering was performed with the stats::hclust() function using the 75 probes with the highest variance in methylation beta values.

Mutational signature analysis. Mutational signature extraction was performed using the nonnegative matrix factorization (NNMF) algorithm[11]. Briefly, the algorithm identifies a minimal set of mutational signatures that optimally explains the proportions of mutation types found across a given mutational catalogue and then estimates the contribution of each identified signature to the mutation spectra of each sample.

Patient survival analysis. The Kaplan-Meier method was used to analyze survival outcomes by the log-rank Mantel-Cox test, with hazard ratio and two-sided 95% confidence intervals calculated using the Mantel_Haenszel test (GraphPad Prism, ver 7.02). Overall survival data from time of first surgery for each anaplastic meningioma within gene-expression defined subgroups C1 and C2 was collected and used to plot a Kaplan-Meier survival curve.

Supplementary Discussion

A hypermutator anaplastic meningioma with a haploid genome. One primary anaplastic meningioma resected from an 85-year old female (PD23359a) had a hypermutator phenotype, with 27,332 point mutations and LOH across nearly its entire genome (Supplementary Fig. S12, Table S24). Independent pathological review confirmed the original diagnosis of anaplastic meningioma, and transcriptome analysis demonstrated that this tumor clustered appropriately with the rest of the cohort (Fig. 3a,b). The majority of the mutations were substitutions, 72% of which were C > T transitions. We identified two deleterious mutations in DNA damage repair mediators: a *TP53* p.R248Q missense mutation and a homozygous truncating variant in the mismatch repair gene *MSH6* (p.L1330Vfs*9). Despite the latter finding, mutational signatures analysis was dominated by signature 1, with no evidence of signatures typically associated with defects in homologous recombination, mismatch repair or *POLE* activity (signatures 3, 6, 10, 15, 20 or 26). The copy number profile is most consistent with this tumor having first undergone haploidization of its genome, with the exception of chromosomes 7, 19 and 20, followed by whole genome duplication (Supplementary Fig. S12). Of note, several important oncogenes are located on chromosome 7, including *EGFR*, *MET* and *BRAF*. Widespread LOH has been described in a significant proportion of oncocytic follicular thyroid cancers where preservation of chromosome 7 heterozygosity has also been observed[117].

Data Availability

All sequencing data that support the findings of this study have been deposited in the European Genome-Phenome Archive and are accessible through the accession numbers EGAS00001000377, EGAS00001000828, EGAS00001000859, EGAS00001001155 and EGAS00001001873. All other relevant data are available from the corresponding author on request.

References

1. Mawrin, C. & Perry, A. Pathological classification and molecular genetics of meningiomas. *J Neurooncol* **99**, 379–391, https://doi.org/10.1007/s11060-010-0342-2 (2010).
2. Rogers, C. L. *et al.* Pathology concordance levels for meningioma classification and grading in NRG Oncology RTOGTrial 0539. *Neuro Oncol* **18**, 565–574, https://doi.org/10.1093/neuonc/nov247 (2016).
3. Moliterno, J. *et al.* Survival in patients treated for anaplastic meningioma. *Journal of neurosurgery* **123**, 23–30, https://doi.org/10.3171/2014.10.JNS14502 (2015).
4. Champeaux, C., Wilson, E., Brandner, S., Shieff, C. & Thorne, L., World Health Organization. grade III meningiomas. A retrospective study for outcome and prognostic factors assessment. *Br J Neurosurg* **29**, 693–698, https://doi.org/10.3109/02688697.2015.1054350 (2015).
5. Durand, A. *et al.* WHO grade II and III meningiomas: a study of prognostic factors. *J Neurooncol* **95**, 367–375, https://doi.org/10.1007/s11060-009-9934-0 (2009).
6. Buttrick, S., Shah, A. H., Komotar, R. J. & Ivan, M. E. Management of Atypical and Anaplastic Meningiomas. *Neurosurg Clin N Am* **27**, 239–247, https://doi.org/10.1016/j.nec.2015.11.003 (2016).
7. Clark, V. E. *et al.* Recurrent somatic mutations in POLR2A define a distinct subset of meningiomas. *Nat Genet* **48**, 1253–1259, https://doi.org/10.1038/ng.3651 http://www.nature.com/ng/journal/vaop/ncurrent/abs/ng.3651.html - supplementary-information (2016).
8. Clark, V. E. *et al.* Genomic analysis of non-NF2 meningiomas reveals mutations in TRAF7, KLF4, AKT1, and SMO. *Science* **339**, 1077–1080, https://doi.org/10.1126/science.1233009 (2013).
9. Harmanci, A. S. *et al.* Integrated genomic analyses of de novo pathways underlying atypical meningiomas. *Nat Commun* **8**, 14433, https://doi.org/10.1038/ncomms14433 (2017).
10. Sahm, F. *et al.* DNA methylation-based classification and grading system for meningioma: a multicentre, retrospective analysis. *Lancet Oncol* **18**, 682–694, https://doi.org/10.1016/S1470-2045(17)30155-9 (2017).
11. Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415–421, https://doi.org/10.1038/nature12477 (2013).
12. Tubio, J. M. *et al.* Mobile DNA in cancer. Extensive transduction of nonrepetitive DNA mediated by L1 retrotransposition in cancer genomes. *Science* **345**, 1251343, https://doi.org/10.1126/science.1251343 (2014).
13. Galani, V. *et al.* Genetic and epigenetic alterations in meningiomas. *Clinical neurology and neurosurgery* **158**, 119–125, https://doi.org/10.1016/j.clineuro.2017.05.002 (2017).
14. Kadoch, C. *et al.* Proteomic and bioinformatic analysis of mammalian SWI/SNF complexes identifies extensive roles in human malignancy. *Nat Genet* **45**, 592–601, https://doi.org/10.1038/ng.2628 (2013).
15. Shain, A. H. & Pollack, J. R. The spectrum of SWI/SNF mutations, ubiquitous in human cancers. *PLoS One* **8**, e55119, https://doi.org/10.1371/journal.pone.0055119 (2013).
16. Wu, J. I., Lessard, J. & Crabtree, G. R. Understanding the words of chromatin regulation. *Cell* **136**, 200–206, https://doi.org/10.1016/j.cell.2009.01.009 (2009).
17. Kia, S. K., Gorski, M. M., Giannakopoulos, S. & Verrijzer, C. P. SWI/SNF mediates polycomb eviction and epigenetic reprogramming of the INK4b-ARF-INK4a locus. *Mol Cell Biol* **28**, 3457–3464, https://doi.org/10.1128/MCB.02019-07 (2008).
18. Wilson, B. G. & Roberts, C. W. SWI/SNF nucleosome remodellers and cancer. *Nat Rev Cancer* **11**, 481–492, https://doi.org/10.1038/nrc3068 (2011).
19. Morales, F. C., Molina, J. R., Hayashi, Y. & Georgescu, M. M. Overexpression of ezrin inactivates NF2 tumor suppressor in glioblastoma. *Neuro Oncol* **12**, 528–539, https://doi.org/10.1093/neuonc/nop060 (2010).
20. Petrilli, A. M. & Fernandez-Valle, C. Role of Merlin/NF2 inactivation in tumor biology. *Oncogene* **35**, 537–548, https://doi.org/10.1038/onc.2015.125 (2016).
21. Goutagny, S. *et al.* High incidence of activating TERT promoter mutations in meningiomas undergoing malignant progression. *Brain Pathol* **24**, 184–189, https://doi.org/10.1111/bpa.12110 (2014).
22. Martincorena, I. *et al.* Universal Patterns of Selection in Cancer and Somatic Tissues. *Cell* **171**, 1029–1041.e1021, https://doi.org/10.1016/j.cell.2017.09.042 (2017).
23. Berx, G. & van Roy, F. Involvement of members of the cadherin superfamily in cancer. *Cold Spring Harbor perspectives in biology* **1**, a003129, https://doi.org/10.1101/cshperspect.a003129 (2009).
24. De Craene, B. & Berx, G. Regulatory networks defining EMT during cancer initiation and progression. *Nat Rev Cancer* **13**, 97–110, https://doi.org/10.1038/nrc3447 (2013).
25. Benarafa, C. & Wolf, M. CXCL14: the Swiss army knife chemokine. *Oncotarget* **6**, 34065–34066, https://doi.org/10.18632/oncotarget.6040 (2015).
26. Polyak, K. & Weinberg, R. A. Transitions between epithelial and mesenchymal states: acquisition of malignant and stem cell traits. *Nat Rev Cancer* **9**, 265–273, https://doi.org/10.1038/nrc2620 (2009).
27. Pujuguet, P., Del Maestro, L., Gautreau, A., Louvard, D. & Arpin, M. Ezrin regulates E-cadherin-dependent adherens junction assembly through Rac1 activation. *Mol Biol Cell* **14**, 2181–2191, https://doi.org/10.1091/mbc.E02-07-0410 (2003).
28. Hayashida, T. *et al.* HOXB9, a gene overexpressed in breast cancer, promotes tumorigenicity and lung metastasis. *Proc Natl Acad Sci USA* **107**, 1100–1105, https://doi.org/10.1073/pnas.0912710107 (2010).
29. Wu, X. *et al.* HOXB7, a homeodomain protein, is overexpressed in breast cancer and confers epithelial-mesenchymal transition. *Cancer Res* **66**, 9527–9534, https://doi.org/10.1158/0008-5472.CAN-05-4470 (2006).
30. Kalluri, R. The biology and function of fibroblasts in cancer. *Nat Rev Cancer* **16**, 582–598, https://doi.org/10.1038/nrc.2016.73 (2016).
31. Sjoberg, E., Augsten, M., Bergh, J., Jirstrom, K. & Ostman, A. Expression of the chemokine CXCL14 in the tumour stroma is an independent marker of survival in breast cancer. *Br J Cancer* **114**, 1117–1124, https://doi.org/10.1038/bjc.2016.104 (2016).
32. Gotwals, P. *et al.* Prospects for combining targeted and conventional cancer therapy with immunotherapy. *Nat Rev Cancer* advance online publication, https://doi.org/10.1038/nrc.2017.17 (2017).
33. Marcucci, F., Stassi, G. & De Maria, R. Epithelial-mesenchymal transition: a new target in anticancer drug discovery. *Nat Rev Drug Discov* **15**, 311–325, https://doi.org/10.1038/nrd.2015.13 (2016).
34. Watson, M. A. *et al.* Molecular characterization of human meningiomas by gene expression profiling using high-density oligonucleotide microarrays. *Am J Pathol* **161**, 665–672, https://doi.org/10.1016/s0002-9440(10)64222-8 (2002).
35. Wrobel, G. *et al.* Microarray-based gene expression profiling of benign, atypical and anaplastic meningiomas identifies novel genes associated with meningioma progression. *Int J Cancer* **114**, 249–256, https://doi.org/10.1002/ijc.20733 (2005).

36. Mawrin, C. *et al*. Different activation of mitogen-activated protein kinase and Akt signaling is associated with aggressive phenotype of human meningiomas. *Clin Cancer Res* **11**, 4074–4082, https://doi.org/10.1158/1078-0432.ccr-04-2550 (2005).
37. Mawrin, C., Chung, C. & Preusser, M. Biology and clinical management challenges in meningioma. *American Society of Clinical Oncology educational book. American Society of Clinical Oncology. Meeting*, e106–115, https://doi.org/10.14694/EdBook_AM.2015.35.e106 (2015).
38. Lopez-Lago, M. A., Okada, T., Murillo, M. M., Socci, N. & Giancotti, F. G. Loss of the tumor suppressor gene NF2, encoding merlin, constitutively activates integrin-dependent mTORC1 signaling. *Mol Cell Biol* **29**, 4235–4249, https://doi.org/10.1128/mcb.01578-08 (2009).
39. James, M. F. *et al*. NF2/merlin is a novel negative regulator of mTOR complex 1, and activation of mTORC1 is associated with meningioma and schwannoma growth. *Mol Cell Biol* **29**, 4250–4261, https://doi.org/10.1128/mcb.01581-08 (2009).
40. Johnson, M. D., Okedli, E., Woodard, A., Toms, S. A. & Allen, G. S. Evidence for phosphatidylinositol 3-kinase-Akt-p7S6K pathway activation and transduction of mitogenic signals by platelet-derived growth factor in meningioma cells. *Journal of neurosurgery* **97**, 668–675, https://doi.org/10.3171/jns.2002.97.3.0668 (2002).
41. Weisman, A. S., Raguet, S. S. & Kelly, P. A. Characterization of the epidermal growth factor receptor in human meningioma. *Cancer Res* **47**, 2172–2176 (1987).
42. Harvey, K. F., Zhang, X. & Thomas, D. M. The Hippo pathway and human cancer. *Nat Rev Cancer* **13**, 246–257, https://doi.org/10.1038/nrc3458 (2013).
43. Calvo, F. *et al*. Mechanotransduction and YAP-dependent matrix remodelling is required for the generation and maintenance of cancer-associated fibroblasts. *Nat Cell Biol* **15**, 637–646, https://doi.org/10.1038/ncb2756 (2013).
44. Rosenbluh, J. *et al*. beta-Catenin-driven cancers require a YAP1 transcriptional complex for survival and tumorigenesis. *Cell* **151**, 1457–1473, https://doi.org/10.1016/j.cell.2012.11.026 (2012).
45. Louis, D. N. *et al*. The2016 World Health Organization Classification of Tumors of the Central Nervous System: a summary. *Acta Neuropathol* **131**, 803–820, https://doi.org/10.1007/s00401-016-1545-1 (2016).
46. Le Loarer, F. *et al*. SMARCA4 inactivation defines a group of undifferentiated thoracic malignancies transcriptionally related to BAF-deficient sarcomas. *Nat Genet* **47**, 1200–1205, https://doi.org/10.1038/ng.3399 (2015).
47. Luchini, C. *et al*. Prognostic role and implications of mutation status of tumor suppressor gene ARID1A in cancer: a systematic review and meta-analysis. *Oncotarget* **6**, 39088–39097, https://doi.org/10.18632/oncotarget.5142 (2015).
48. Lu, C. & Allis, C. D. SWI/SNF complex in cancer. *Nat Genet* **49**, 178–179, https://doi.org/10.1038/ng.3779 (2017).
49. Goldbrunner, R. *et al*. EANO guidelines for the diagnosis and treatment of meningiomas. *Lancet Oncol* **17**, e383–391, https://doi.org/10.1016/S1470-2045(16)30321-7 (2016).
50. Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: the next generation. *Cell* **144**, 646–674, https://doi.org/10.1016/j.cell.2011.02.013 (2011).
51. Shah, N. & Sukumar, S. The Hox genes and their roles in oncogenesis. *Nat Rev Cancer* **10**, 361–371, https://doi.org/10.1038/nrc2826 (2010).
52. Xu, Q. *et al*. Long non-coding RNA regulation of epithelial-mesenchymal transition in cancer metastasis. *Cell Death Dis* **7**, e2254, https://doi.org/10.1038/cddis.2016.149 (2016).
53. Krumlauf, R. Hox genes in vertebrate development. *Cell* **78**, 191–201 (1994).
54. Xie, M. *et al*. Long noncoding RNA HOXA-AS2 promotes gastric cancer proliferation by epigenetically silencing P21/PLK3/DDIT3 expression. *Oncotarget* **6**, 33587–33601, https://doi.org/10.18632/oncotarget.5599 (2015).
55. Bao, X. *et al*. Knockdown of long non-coding RNA HOTAIR increases miR-454-3p by targeting Stat3 and Atg12 to inhibit chondrosarcoma growth. *Cell Death Dis* **8**, e2605, https://doi.org/10.1038/cddis.2017.31 (2017).
56. Gupta, R. A. *et al*. Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. *Nature* **464**, 1071–1076, https://doi.org/10.1038/nature08975 (2010).
57. Kim, K. *et al*. HOTAIR is a negative prognostic factor and exhibits pro-oncogenic activity in pancreatic cancer. *Oncogene* **32**, 1616–1625, http://www.nature.com/onc/journal/v32/n13/suppinfo/onc2012193s1.html (2013).
58. Li, X. *et al*. Long non-coding RNA HOTAIR, a driver of malignancy, predicts negative prognosis and exhibits oncogenic activity in oesophageal squamous cell carcinoma. *Br J Cancer* **109**, 2266–2278, https://doi.org/10.1038/bjc.2013.548 (2013).
59. Ozes, A. R. *et al*. NF-kappaB-HOTAIR axis links DNA damage response, chemoresistance and cellular senescence in ovarian cancer. *Oncogene* **35**, 5350–5361, https://doi.org/10.1038/onc.2016.75 (2016).
60. Shi, J. *et al*. Long non-coding RNA in glioma: signaling pathways. *Oncotarget*. https://doi.org/10.18632/oncotarget.15175 (2017).
61. Ding, J. *et al*. Long noncoding RNA HOXA-AS2 represses P21 and KLF2 expression transcription by binding with EZH2, LSD1 in colorectal cancer. *Oncogenesis* **6**, e288, https://doi.org/10.1038/oncsis.2016.84 (2017).
62. Zhao, H., Zhang, X., Frazao, J. B., Condino-Neto, A. & Newburger, P. E. HOX antisense lincRNA HOXA-AS2 is an apoptosis repressor in all trans retinoic acid treated NB4 promyelocytic leukemia cells. *J Cell Biochem* **114**, 2375–2383, https://doi.org/10.1002/jcb.24586 (2013).
63. Ponting, C. P., Oliver, P. L. & Reik, W. Evolution and functions of long noncoding RNAs. *Cell* **136**, 629–641, https://doi.org/10.1016/j.cell.2009.02.006 (2009).
64. Rinn, J. L. *et al*. Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell* **129**, 1311–1323, https://doi.org/10.1016/j.cell.2007.05.022 (2007).
65. Khalil, A. M. *et al*. Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proc Natl Acad Sci USA* **106**, 11667–11672, https://doi.org/10.1073/pnas.0904715106 (2009).
66. Goutagny, S. *et al*. Genomic profiling reveals alternative genetic pathways of meningioma malignant progression dependent on the underlying NF2 status. *Clin Cancer Res* **16**, 4155–4164, https://doi.org/10.1158/1078-0432.CCR-10-0891 (2010).
67. Bostrom, J. *et al*. Alterations of the tumor suppressor genes CDKN2A (p16(INK4a)), p14(ARF), CDKN2B (p15(INK4b)), and CDKN2C (p18(INK4c)) in atypical and anaplastic meningiomas. *Am J Pathol* **159**, 661–669, https://doi.org/10.1016/S0002-9440(10)61737-3 (2001).
68. Kadoch, C. & Crabtree, G. R. Mammalian SWI/SNF chromatin remodeling complexes and cancer: Mechanistic insights gained from human genomics. *Sci Adv* **1**, e1500447, https://doi.org/10.1126/sciadv.1500447 (2015).
69. Bernhart, S. H. *et al*. Changes of bivalent chromatin coincide with increased expression of developmental genes in cancer. *Sci Rep* **6**, 37393, https://doi.org/10.1038/srep37393 (2016).
70. Voigt, P., Tee, W. W. & Reinberg, D. A double take on bivalent promoters. *Genes Dev* **27**, 1318–1338, https://doi.org/10.1101/gad.219626.113 (2013).
71. Bernstein, B. E. *et al*. A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell* **125**, 315–326, https://doi.org/10.1016/j.cell.2006.02.041 (2006).
72. Helming, K. C., Wang, X. & Roberts, C. W. Vulnerabilities of mutant SWI/SNF complexes in cancer. *Cancer Cell* **26**, 309–317, https://doi.org/10.1016/j.ccr.2014.07.018 (2014).
73. Kim, K. H. *et al*. SWI/SNF-mutant cancers depend on catalytic and non-catalytic activity of EZH2. *Nat Med* **21**, 1491–1496, https://doi.org/10.1038/nm.3968 (2015).
74. Bitler, B. G. *et al*. Synthetic lethality by targeting EZH2 methyltransferase activity in ARID1A-mutated cancers. *Nat Med* **21**, 231–238, https://doi.org/10.1038/nm.3799 http://www.nature.com/nm/journal/v21/n3/abs/nm.3799.html - supplementary-information (2015).

75. Kim, K. H. & Roberts, C. W. Targeting EZH2 in cancer. *Nat Med* **22**, 128–134, https://doi.org/10.1038/nm.4036 (2016).
76. Ozes, A. R. *et al*. Therapeutic targeting using tumor specific peptides inhibits long non-coding RNA HOTAIR activity in ovarian and breast cancer. *Scientific reports* **7**, 894, https://doi.org/10.1038/s41598-017-00966-3 (2017).
77. Pfister, S. X. & Ashworth, A. Marked for death: targeting epigenetic changes in cancer. *Nature reviews. Drug discovery* **16**, 241–263, https://doi.org/10.1038/nrd.2016.256 (2017).
78. Norden, A. D. *et al*. Phase II trials of erlotinib or gefitinib in patients with recurrent meningioma. *J Neurooncol* **96**, 211–217, https://doi.org/10.1007/s11060-009-9948-7 (2010).
79. Byers, L. A. *et al*. An epithelial-mesenchymal transition gene signature predicts resistance to EGFR and PI3K inhibitors and identifies Axl as a therapeutic target for overcoming EGFR inhibitor resistance. *Clinical cancer research: an official journal of the American Association for Cancer Research* **19**, 279–290, https://doi.org/10.1158/1078-0432.ccr-12-1558 (2013).
80. Buonato, J. M. & Lazzara, M. J. ERK1/2 blockade prevents epithelial-mesenchymal transition in lung cancer cells and promotes their sensitivity to EGFR inhibition. *Cancer Res* **74**, 309–319, https://doi.org/10.1158/0008-5472.can-12-4721 (2014).
81. Thomson, S., Petti, F., Sujka-Kwok, I., Epstein, D. & Haley, J. D. Kinase switching in mesenchymal-like non-small cell lung cancer lines contributes to EGFR inhibitor resistance through pathway redundancy. *Clinical & experimental metastasis* **25**, 843–854, https://doi.org/10.1007/s10585-008-9200-4 (2008).
82. Chmielecki, J. *et al*. Whole-exome sequencing identifies a recurrent NAB2-STAT6 fusion in solitary fibrous tumors. *Nat Genet* **45**, 131–132, https://doi.org/10.1038/ng.2522 (2013).
83. Gao, F. *et al*. Inversion-mediated gene fusions involving NAB2-STAT6 in an unusual malignant meningioma. *Br J Cancer* **109**, 1051–1055, https://doi.org/10.1038/bjc.2013.395 (2013).
84. Schweizer, L. *et al*. Meningeal hemangiopericytoma and solitary fibrous tumors carry the NAB2-STAT6 fusion and can be diagnosed by nuclear expression of STAT6 protein. *Acta Neuropathol* **125**, 651–658, https://doi.org/10.1007/s00401-013-1117-6 (2013).
85. Soda, M. *et al*. Identification of the transforming EML4-ALK fusion gene in non-small-cell lung cancer. *Nature* **448**, 561–566, https://doi.org/10.1038/nature05945 (2007).
86. Kozarewa, I. *et al*. Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. *Nature methods* **6**, 291–295, https://doi.org/10.1038/nmeth.1311 (2009).
87. Dobin, A. *et al*. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics (Oxford, England)* **29**, 15–21, https://doi.org/10.1093/bioinformatics/bts635 (2013).
88. Anders, S., Pyl, P. T. & Huber, W. HTSeq–a Python framework to work with high-throughput sequencing data. *Bioinformatics (Oxford, England)* **31**, 166–169, https://doi.org/10.1093/bioinformatics/btu638 (2015).
89. McPherson, A. *et al*. deFuse: an algorithm for gene fusion discovery in tumor RNA-Seq data. *PLoS computational biology* **7**, e1001138, https://doi.org/10.1371/journal.pcbi.1001138 (2011).
90. Kim, D. & Salzberg, S. L. TopHat-Fusion: an algorithm for discovery of novel fusion transcripts. *Genome biology* **12**, R72, https://doi.org/10.1186/gb-2011-12-8-r72 (2011).
91. Bignell, G. R. *et al*. Signatures of mutation and selection in the cancer genome. *Nature* **463**, 893–898, https://doi.org/10.1038/nature08768 (2010).
92. M. Witten, D. Witten, D. M.: *Classification and clustering of sequencing data using a Poisson model*. *Ann. Appl. Stat.* *5*(4), 2493–2518 Vol. 5 (2012).
93. Reeb, P. D., Bramardi, S. J. & Steibel, J. P. Assessing Dissimilarity Measures for Sample-Based Hierarchical Clustering of RNA Sequencing Data Using Plasmode Datasets. *PLoS One* **10**, e0132310, https://doi.org/10.1371/journal.pone.0132310 (2015).
94. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome biology* **11**, R106, https://doi.org/10.1186/gb-2010-11-10-r106 (2010).
95. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq. 2. *Genome biology* **15**, 550, https://doi.org/10.1186/s13059-014-0550-8 (2014).
96. Peixoto, L. *et al*. How data analysis affects power, reproducibility and biological insight of RNA-seq studies in complex datasets. *Nucleic acids research* **43**, 7664–7674, https://doi.org/10.1093/nar/gkv736 (2015).
97. Stegle, O., Parts, L., Piipari, M., Winn, J. & Durbin, R. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nat Protoc* **7**, 500–507, https://doi.org/10.1038/nprot.2011.457 (2012).
98. Wang, X. *et al*. Analysis of gene expression profiling in meningioma: deregulated signaling pathways associated with meningioma and EGFL6 overexpression in benign meningioma tissue and serum. *PLoS One* **7**, e52707, https://doi.org/10.1371/journal.pone.0052707 (2012).
99. Savvidis, C. & Koutsilieris, M. Circadian rhythm disruption in cancer biology. *Mol Med* **18**, 1249–1260, https://doi.org/10.2119/molmed.2012.00077 (2012).
100. Sharma, S., Ray, S., Moiyadi, A., Sridhar, E. & Srivastava, S. Quantitative proteomic analysis of meningiomas for the identification of surrogate protein markers. *Sci Rep* **4**, 7140, https://doi.org/10.1038/srep07140 (2014).
101. Pfister, C., Tatabiga, M. S. & Roser, F. Selection of suitable reference genes for quantitative real-time polymerase chain reaction in human meningiomas and arachnoidea. *BMC Res Notes* **4**, 275, https://doi.org/10.1186/1756-0500-4-275 (2011).
102. Luo, W., Friedman, M. S., Shedden, K., Hankenson, K. D. & Woolf, P. J. GAGE: generally applicable gene set enrichment for pathway analysis. *BMC Bioinformatics* **10**, 161, https://doi.org/10.1186/1471-2105-10-161 (2009).
103. Iorio, F. *et al*. Population-level characterization of pathway alterations with SLAPenrich dissects heterogeneity of cancer hallmark acquisition. *bioRxiv*. https://doi.org/10.1101/077701 (2016).
104. Ben-Porath, I. *et al*. An embryonic stem cell-like gene expression signature in poorly differentiated aggressive human tumors. *Nat Genet* **40**, 499–507, https://doi.org/10.1038/ng.127 (2008).
105. Sarrio, D. *et al*. Epithelial-mesenchymal transition in breast cancer relates to the basal-like phenotype. *Cancer Res* **68**, 989–997, https://doi.org/10.1158/0008-5472.can-07-2017 (2008).
106. Wong, D. J. *et al*. Module map of stem cell genes guides creation of epithelial cancer stem cells. *Cell stem cell* **2**, 333–344, https://doi.org/10.1016/j.stem.2008.02.009 (2008).
107. The Gene Ontology Consortium. Gene Ontology Consortium: going forward. *Nucleic acids research* **43**, D1049–D1056, https://doi.org/10.1093/nar/gku1179 (2015).
108. R: A language and environment for statistical computing (R Foundation for Statistical Computing, Vienna, Austria, 2016).
109. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)* **25**, 1754–1760, https://doi.org/10.1093/bioinformatics/btp324 (2009).
110. Ye, K., Schulz, M. H., Long, Q., Apweiler, R. & Ning, Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics (Oxford, England)* **25**, 2865–2871, https://doi.org/10.1093/bioinformatics/btp394 (2009).
111. Raine, K. M. *et al*. cgpPindel: Identifying Somatically Acquired Insertion and Deletion Events from Paired End Sequencing. *Current protocols in bioinformatics* **52**, 15.17.11–12, https://doi.org/10.1002/0471250953.bi1507s52 (2015).
112. Nik-Zainal, S. *et al*. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature* **534**, 47–54, https://doi.org/10.1038/nature17676 (2016).

113. Raine, K. M. *et al*. ascatNgs: Identifying Somatically Acquired Copy-Number Alterations from Whole-Genome Sequencing Data. *Current protocols in bioinformatics* **56**, 15.19.11–15.19.17, https://doi.org/10.1002/cpbi.17 (2016).
114. Buels, R. *et al*. JBrowse: a dynamic web platform for genome visualization and analysis. *Genome biology* **17**, 66, https://doi.org/10.1186/s13059-016-0924-1 (2016).
115. Chang, M. T. *et al*. Identifying recurrent mutations in cancer reveals widespread lineage diversity and mutational specificity. *Nat Biotechnol* **34**, 155–163, https://doi.org/10.1038/nbt.3391 (2016).
116. Subramanian, A. *et al*. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA* **102**, 15545–15550, https://doi.org/10.1073/pnas.0506580102 (2005).
117. Corver, W. E. *et al*. Genome haploidisation with chromosome 7 retention in oncocytic follicular thyroid carcinoma. *PLoS One* **7**, e38287, https://doi.org/10.1371/journal.pone.0038287 (2012).

Acknowledgements

This work was supported by the Wellcome Trust, Cancer Research UK, Meningioma UK and Tadhg and Marie-Louise Flood. U.M. was personally supported by a Cancer Research UK Clinician Scientist Fellowship; G.C. by a Wellcome Trust Clinical PhD Fellowship (WT098051); F.M. by A.I.L. (Associazione Italiana Contro le Leucemie-Linfomi e Mieloma ONLUS) and by S.I.E.S. (Società Italiana di Ematologia Sperimentale); S.B. was funded by a Wellcome Trust Intermediate Clinical Research Fellowship and a St. Baldrick's Foundation Robert J. Arceci Innovation Award. J.M.C.T. is supported by ERC Starting Grant StG-2016_716290_SCUBA CANCERS and by MINECO Grants RYC 2014 14999 and SAF2015-66368-P. J.B. was funded by the charity Brain Tumour Research. The samples were received from the tissue banks from Cambridge (UK), Dresden (Germany), Liverpool (UK), Plymouth (UK) and Tel Aviv (Israel). The Human Research Tissue Bank is supported by the NIHR Cambridge Biomedical Research Centre. We are grateful to the patients who enabled this study and to the clinical teams coordinating their care.

Author Contributions

G.C. and N.K. performed mRNA expression analysis. G.C. and P.T. analysed whole genome and targeted sequencing data. I.M. performed statistical analyses to detect novel driver mutations. S.M. analysed methylation array data. F.M. generated mutational signatures analysis. J.M.C.T. and M.C. performed retrotransposon analysis. C.O.H. and J.D. performed protein expression analysis. A.B., S.B. and M.Y. contributed to data analysis strategy. A.Y., T.N., G.R.B. and J.T. provided informatic support. T.S., R.W.K., M.K., G.S., D.P., A.D., C.E.M., A.Y., I.N., S.J.P., C.W., Z.R., M.D.J., R.Z. and K. S. provided samples and clinical data. S.B., G.S.V., I.N. and M.W.M. provided conceptual advice. V.P.C. and K.A. carried out central pathology review. U.M. and T.S. devised and supervised the project. G.C. wrote the manuscript with input from U.M., S.B., T.S., P.T., and G.S.V. All authors approved the manuscript.

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-018-31659-0.

Competing Interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

nature COMMUNICATIONS

ARTICLE

DOI: 10.1038/s41467-018-04650-6 OPEN

Recurrent intragenic rearrangements of *EGFR* and *BRAF* in soft tissue tumors of infants

Jenny Wegert[1], Christian Vokuhl[2], Grace Collord[3,4], Martin Del Castillo Velasco-Herrera[3], Sarah J. Farndon[3,5], Charlotte Guzzo[3], Mette Jorgensen[6], John Anderson[5,6], Olga Slater[6], Catriona Duncan[6], Sabrina Bausenwein[1], Heike Streitenberger[1], Barbara Ziegler[1], Rhoikos Furtwängler[7], Norbert Graf[7], Michael R. Stratton[3], Peter J. Campbell[3], David TW Jones[8,9], Christian Koelsche[10,11,12], Stefan M. Pfister[8,9,13], William Mifsud[6], Neil Sebire[5,6], Monika Sparber-Sauer[14], Ewa Koscielniak[14,15], Andreas Rosenwald[16,17], Manfred Gessler[1,17] & Sam Behjati[3,4]

Soft tissue tumors of infancy encompass an overlapping spectrum of diseases that pose unique diagnostic and clinical challenges. We studied genomes and transcriptomes of cryptogenic congenital mesoblastic nephroma (CMN), and extended our findings to five anatomically or histologically related soft tissue tumors: infantile fibrosarcoma (IFS), nephroblastomatosis, Wilms tumor, malignant rhabdoid tumor, and clear cell sarcoma of the kidney. A key finding is recurrent mutation of *EGFR* in CMN by internal tandem duplication of the kinase domain, thus delineating CMN from other childhood renal tumors. Furthermore, we identify *BRAF* intragenic rearrangements in CMN and IFS. Collectively these findings reveal novel diagnostic markers and therapeutic strategies and highlight a prominent role of isolated intragenic rearrangements as drivers of infant tumors.

[1] Theodor-Boveri-Institute/Biocenter, Developmental Biochemistry, University of Wuerzburg, 97074 Wuerzburg, Germany. [2] Kiel Pediatric Tumor Registry, Section of Pediatric Pathology, Department of Pathology, Christian Albrechts University, 24105 Kiel, Germany. [3] Wellcome Trust Sanger Institute, Hinxton, CB10 1SA, UK. [4] Department of Paediatrics, University of Cambridge, Cambridge, CB2 0QQ, UK. [5] UCL Great Ormond Street Institute of Child Health, London, WC1N 1EH, UK. [6] Great Ormond Street Hospital for Children NHS Foundation Trust, London, WC1N 3JH, UK. [7] Department of Pediatric Oncology and Hematology, Saarland University Hospital, 66421 Homburg, Germany. [8] Hopp Children's Cancer Center at the NCT Heidelberg (KiTZ), 69120 Heidelberg, Germany. [9] Department of Pediatric Neurooncology, German Cancer Research Center (DKFZ) and German Cancer Consortium (DKTK), 69120 Heidelberg, Germany. [10] Clinical Cooperation Unit Neuropathology, German Cancer Research Center (DKFZ), 69120 Heidelberg, Germany. [11] Department of Neuropathology, Institute of Pathology, Heidelberg University Hospital, 69120 Heidelberg, Germany. [12] Department of General Pathology, Institute of Pathology, Heidelberg University Hospital, 69120 Heidelberg, Germany. [13] Department of Pediatric Hematology and Oncology, Heidelberg University Hospital, 69120 Heidelberg, Germany. [14] Klinikum Stuttgart—Olgahospital, Stuttgart Cancer Center, Zentrum für Kinder-, Jugend- und Frauenmedizin, Pediatrics 5 (Oncology, Hematology, Immunology), 70174 Stuttgart, Germany. [15] Department of Pediatric Hematology and Oncology, Children's Hospital, 72076 Tübingen, Germany. [16] Institute of Pathology, University of Wuerzburg, 97080 Wuerzburg, Germany. [17] Comprehensive Cancer Center Mainfranken, University of Wuerzburg, 97078 Wuerzburg, Germany. These authors contributed equally: Jenny Wegert, Christian Vokuhl, Grace Collord, Martin Del Castillo Velasco-Herrera. These authors jointly supervised this work: Manfred Gessler, Sam Behjati. Correspondence and requests for materials should be addressed to M.G. (email: gessler@biozentrum.uni-wuerzburg.de) or to S.B. (email: sb31@sanger.ac.uk)

Many childhood tumors show a predilection for specific developmental stages. Tumors that predominantly occur in infancy include congenital mesoblastic nephroma (CMN), which accounts for 4% of all childhood renal malignancies and the majority of those diagnosed in children under 6 months of age[1,2]. CMN is classified histologically into classical, cellular, and mixed subtypes based primarily on degree of cellularity and mitotic activity[3]. The cellular variant is characterized by a sarcoma-like diffuse hypercellular morphology, whereas classical CMN is composed of less proliferative spindle cells[3]. Cellular CMN is driven by rearrangements involving the tropomyosin receptor kinase (TRK) gene *NTRK3*, most commonly a t(12;15)(p13;q25) reciprocal translocation with the *ETV6* transcription factor[4,5]. Less frequent somatic aberrations include trisomies of chromosomes 8, 11, 17, and 20[6,7] and rarer TRK fusions, involving *NTRK1*, *NTRK2*, or *NTRK3*[8]. By contrast, the genetic changes underpinning the classical variant, accounting for >30% of cases, are unknown[9]. Cellular CMN shares its genetic and morphological hallmarks with infantile fibrosarcoma (IFS), a spindle cell tumor typically arising in the soft tissues of the extremities or abdomen[5,9,10].

Standard treatment for CMN and IFS is complete surgical resection[9–11]. In the case of IFS, local control frequently requires cytotoxic chemotherapy[10,11]. The role for up-front chemotherapy in CMN is less clear[9]. Recently, a phase I/II clinical trial of a selective TRK inhibitor, larotrectinib, reported high response rates in diverse tumor types harboring TRK gene fusions, including IFS and other soft tissue tumors of infancy[12]. Morbidity and infrequent death result from tumor recurrence or from treatment-related complications[9–11].

Here, we investigated the genetic basis of CMN and IFS lacking the canonical *NTRK3-ETV6* fusion gene. We identify oncogenic rearrangements in MAPK signaling genes across all cases interrogated by unbiased sequencing, notably therapeutically tractable intragenic rearrangements in *EGFR* and *BRAF*.

Results

Overview of the genomic landscape of CMN. To identify the genetic basis of cryptogenic CMN, we first applied whole genome and transcriptome sequencing to a discovery cohort of ten classical CMN lacking an *NTRK3* fusion (Supplementary Data 1). Somatic variants were identified by comparing tumor and matched peripheral blood sequences (see Methods). The genomic landscape was universally quiet, with a low burden of point mutations (median of 45 substitutions and 9 insertions or deletions per genome; Supplementary Data 2). The predominant mutational signatures, as defined by the trinucleotide context of substitutions, were the ubiquitous signatures 1 and 5[13]

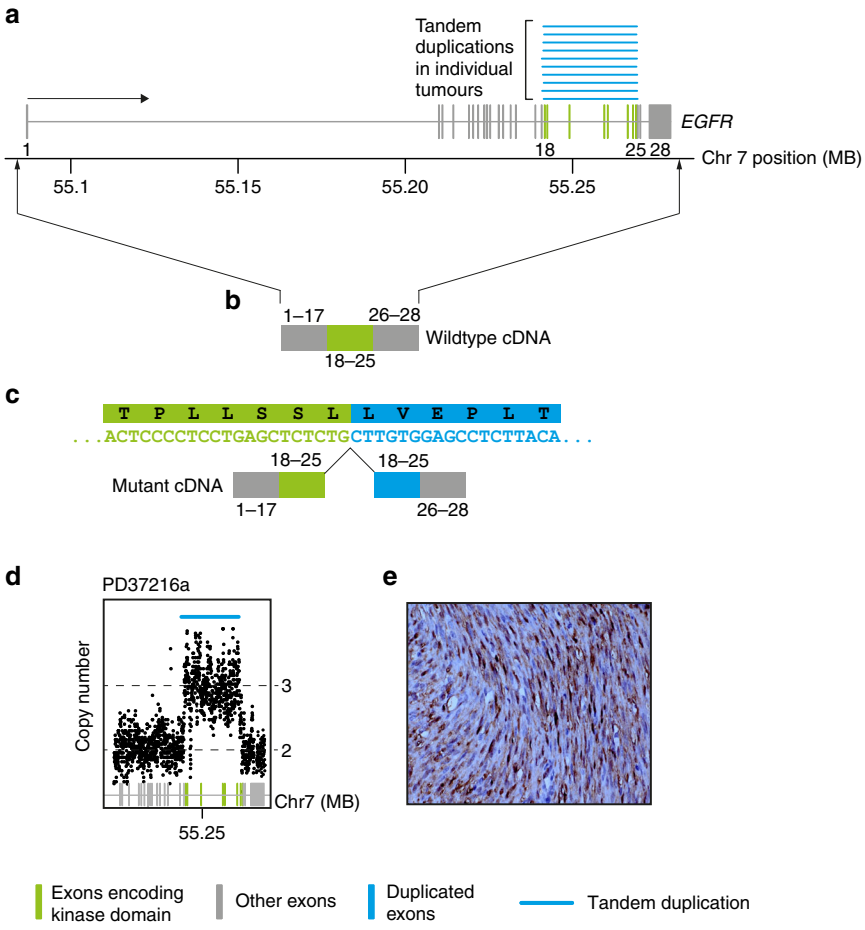

Fig. 1 *EGFR* internal tandem duplication. **a** The genomic footprint of *EGFR* is depicted with exons represented by gray and green vertical lines. Green exons encode the kinase domain. Blue lines superiorly show the tandem duplications found in the discovery cohort of ten congenital mesoblastic nephroma of classical histology. **b** Schematic of the wild-type transcript. **c** Schematic of the fusion transcript annotated with cDNA sequence of rearrangements (sense orientation) and protein translation. **d** Intragenic copy number of *EGFR* showing focal amplification over the kinase domain (*x*-axis: genomic coordinate; *y*-axis: copy number derived from coverage). **e** Representative phospo-ERK immunohistochemistry

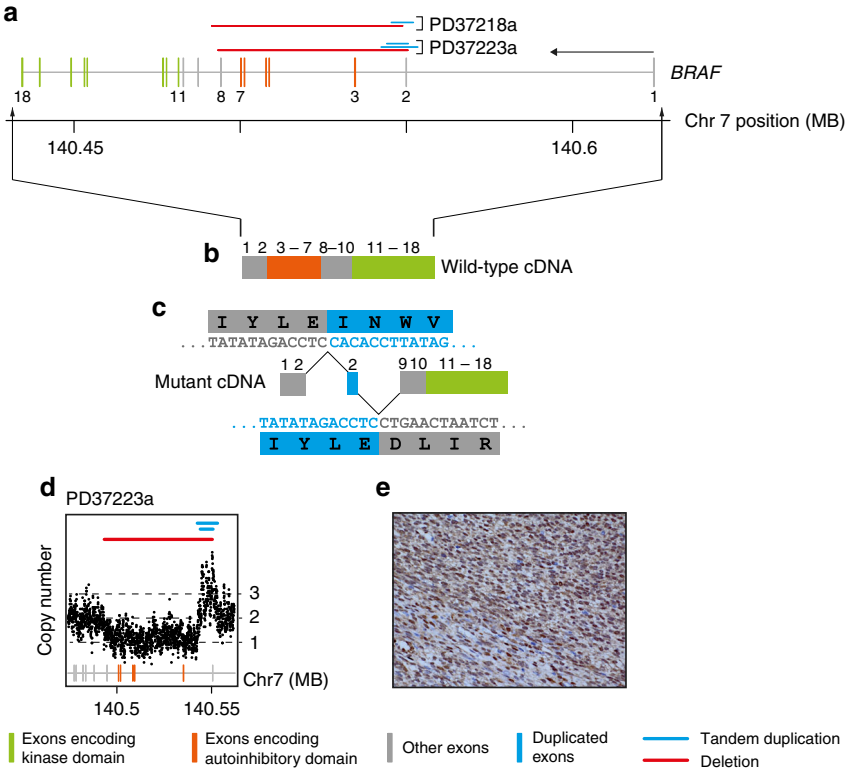

Fig. 2 Internal *BRAF* deletion. **a** The genomic footprint of *BRAF* is depicted with exons represented by gray, green, and orange vertical lines. Green and orange exons encode the kinase domain and conserved region 1, respectively. Horizontal lines above exons demarcate rearrangements (blue: tandem duplication; red: deletion). **b** Outline of wild-type transcript. **c** Outline of fusion transcript with cDNA sequence of rearrangements (sense orientation) with translation. **d** Intragenic copy number of *BRAF* (*x*-axis: genomic coordinate; *y*-axis: copy number derived from coverage). **e** Representative phospho-ERK immunohistochemistry

(Supplementary Fig. 1). Copy number changes and structural rearrangements were likewise scarce (Supplementary Fig. 2).

Internal tandem duplication of the *EGFR* kinase domain in CMN. Annotating all cases for potential oncogenic variants revealed a single intragenic, in-frame internal tandem duplication (ITD) of the *EGFR* kinase domain in all ten tumors (Table 1; Fig. 1; Supplementary Data 3). The breakpoints clustered in a narrow genomic window around the kinase domain of *EGFR* encoded in exons 18−25 (Fig. 1a). This rearrangement is rarely observed in several other tumor types including in glioma and in lung adenocarcinoma, and confers sensitivity to a targeted EGFR inhibitor, afatinib[14]. We validated all rearrangements by genomic copy number analysis and reconstruction of cDNA reads spanning the breakpoint junction (Fig. 1; see Methods). Of note, the same mutant cDNA junction sequence was found in every case, irrespective of the genomic location of breakpoints. A search for additional known or novel driver variants revealed no further plausible candidates in any of the *EGFR*-mutant tumors. We next extended this investigation to seven non-classical CMN lacking an *NTRK3* fusion, including four mixed cellularity cases and three cellular tumors (Table 1; Supplementary Data 1). Two of the four mixed cellularity tumors surveyed also harbored an *EGFR*-ITD. Of note, for one child with *EGFR*-ITD-positive mixed cellularity CMN (PD37214), both primary tumor and recurrence were studied, with no additional driver events apparent at relapse.

***BRAF* rearrangements in CMN and IFS**. A further striking finding was the discovery of mutations in the *BRAF* oncogene in 2/3 cellular histology CMNs. *BRAF* fusions have been implicated in a minority of IFS but not in CMN[15]. In both cases the *BRAF* rearrangement involved a compound deletion of conserved region 1 (CR1) and tandem duplication of exon 2 (Fig. 2; Table 1; Supplementary Data 3). CR1 encompasses the negative regulatory Ras-binding domain (RBD), loss of which is predicted to generate a constitutively active form of BRAF[16,17]. Mutated tumors displayed intense staining of phosphorylated ERK by immunohistochemistry, consistent with activated signaling downstream of BRAF (Figs. 1e and 2e). A further tumor harbored the *KIAA1549-BRAF* fusion, a molecular hallmark of a childhood brain tumor, pilocytic astrocytoma[18,19]. This fusion likewise results in loss of the N-terminal portion of the BRAF protein containing the RBD[17,18].

Other TRK fusions in CMN. The remaining two cases of CMN interrogated by whole genome and transcriptome sequencing were accounted for by gene fusions involving *NTRK1*, an alternate kinase of the TRK family of protein kinases: *TPR-NTRK1* and *LMNA-NTRK1*. Both of these fusions have been observed in IFS and rarely in adult cancers, but not, to our knowledge, in CMN[20–23] (Table 1). Hence, every cryptogenic CMN interrogated by whole-genome sequencing contained an oncogenic rearrangement in *BRAF*, *EGFR*, or *NTRK1*, all of which encode kinases involved in MAPK signaling and are amenable to inhibition with existing drugs[9,12,14,17,24].

***EGFR-ITD* distinguishes CMN from other childhood renal tumors**. To validate and extend our findings, we screened IFS and a range of childhood renal tumors for *EGFR*-ITD, *BRAF*-ID, and *ETV6-NTRK3* using PCR. Tumor types included additional cases of CMN ($n = 63$), IFS ($n = 26$), Wilms tumor ($n = 208$), clear cell sarcoma of the kidney without *BCOR* rearrangements ($n = 20$), malignant rhabdoid tumor ($n = 3$), and nephroblastomatosis

Table 1 Rearrangements in infant soft tissue tumors

Assay	Tumor type	Subtype	Total	*EGFR*-ITD	*BRAF*-ID	*BRAF*-ID + *ETV6-NTRK3*	*ETV6-NTRK3*	*KIAA1549-BRAF*	*LMNA-NTRK1*	*EML4-NTRK3*	*TPR-NTRK1*
WGS + mRNA sequencing	CMN	Cellular	3	0	2	0	0	0	1	0	0
		Classical	10	10	0	0	0	0	0	0	0
		Mixed	4	2	0	0	0	1	0	0	1
	IFS	–	1	0	0	0	0	0	0	1	0
PCR for *EGFR*-ITD, *BRAF*-ID and *ETV6-NTRK3*	CMN	Cellular	17	2	0	0	13	-	-	-	-
		Classical	35	20	0	0	0	-	-	-	-
		Mixed	11	9	0	0	0	-	-	-	-
	IFS	-	26	0	1	2	16	-	-	-	-
	WT	-	208	0	0	0	0	-	-	-	-
	CCSK[a]	-	20	0	0	0	0	-	-	-	-
	MRT	-	3	0	0	0	0	-	-	-	-
	NB	-	12	0	0	0	0	-	-	-	-

CMN congenital mesoblastic nephroma, *IFS* infantile fibrosarcoma, *WT* Wilms tumor, *CCSK* clear cell sarcoma of the kidney, *MRT* malignant rhabdoid tumor, *NB* nephroblastomatosis, *WGS* whole genome sequencing, *mRNA* messenger RNA, *PCR* polymerase chain reaction
[a]Negative for BCOR rearrangement

($n = 12$; Table 1; Supplementary Data 1). *EGFR*-ITD was most prevalent in classical and mixed cellularity CMN, though was also found in cellular CMN (2/17 cases). The frequency of *EGFR* rearrangement in classical tumors was lower in the validation cohort (20/35 cases) than in the initial discovery cohort (10/10 cases). None of the IFS cases, nor other childhood kidney tumors, harbored *EGFR*-ITD. However, we encountered three cases of IFS with intragenic *BRAF* deletions. Remarkably, in two cases *BRAF*-ID co-occurred with *NTRK3* fusions, the disease-defining mutation of IFS. We were unable to accurately estimate relative allele frequencies by nested PCR (see Methods). Hence, it is possible that both fusions co-exist within the same clone or represent independent clones that evolved in parallel within the same tumor.

Discussion

In this exploration of infant tumors we identify ITD of the *EGFR* kinase domain that delineates a genetic subgroup of CMN transcending histological subtypes. Additionally, we report a novel rearrangement of *BRAF* present in both cellular CMN and IFS. These mutations represent diagnostic markers that can be readily integrated into routine clinical practice. Furthermore, EGFR and BRAF emerge as therapeutic targets, which may be exploited in certain clinical situations, e.g., large surgically intractable tumors, disease recurrence or metastases.

It is noteworthy that an oncogenic mutation was identified in every tumor that we studied by whole-genome sequencing. Of these, 78% harbored either *EGFR*-ITD or *BRAF*-ID, while the remaining 22% presented with non-canonical mutations involving *BRAF*, *NTRK1*, or *NTRK3*. This suggests that less recurrent rearrangement variants, albeit implicated in the same signaling circuity, may elude detection by targeted diagnostic assays. Moreover, our results indicate that a subset of tumors harbor multiple drivers with important implications for targeted therapy efforts. The finding of co-mutation of *NTRK3* and *BRAF* in IFS raises the possibility of intrinsic resistance of some tumors to TRK inhibition, regardless of whether these mutations occur in the same clone or in independent competing clones. This finding is pertinent to clinical trials of TRK inhibitors in CMN and IFS[12]. In this vein a structurally similar *BRAF* fusion transcript, albeit without duplication of exon 2, has recently been implicated as a mechanism of resistance to certain BRAF/MEK inhibitors[16,17]. These considerations underscore the need for adequate genomic profiling in order to match patients to the most appropriate basket studies and to enable meaningful interpretation of treatment responses. Therefore, we would advocate extending the diagnostic work-up of refractory or relapsed CMN and IFS to whole genome sequencing, particularly in the context of clinical trials.

Biologically our findings draw further parallels between CMN and IFS. We identify *BRAF* and *NTRK1* as additional cancer genes operative in both malignancies, substantiating the view that these diagnoses represent variants on the same disease spectrum converging on aberrant RAS-RAF-MEK-ERK signaling[5,8,9]. Furthermore, in the wider context of the childhood cancer genome, our findings add to the growing body of studies that identify short distance intragenic rearrangements as a dominant source of oncogenic mutations in otherwise quiet genomes. We note the parallel between CMN, clear cell sarcoma of the kidney and low-grade glioma that are in large part driven by ITDs often involving kinase domains, mostly as isolated driver events[18,25–29]. Furthermore, even in acute myeloid leukemia, where *FLT3*-ITD is a recurrent driver event in adult disease, childhood AML demonstrates a distinct structural variant profile enriched for focal chromosomal gains and losses[30]. We can only speculate on the biological significance of this parallel which may allude to specific mutational mechanisms operative during discrete stages of human development.

Methods

Patient samples. All tissue samples were obtained after gaining written informed consent for tumor banking and future research from the patient (or their guardian) in accordance with the Declaration of Helsinki and appropriate national and local ethical review processes. German tissue samples were obtained from the following studies: SIOP93-01/GPOH and SIOP2001/GPOH (Ethikkommission der Ärztekammer des Saarlandes reference numbers 23.4.93/Ls and 136/01), the PTT2.0 study (Medical Faculty Heidelberg ethics reference number S-546/2016), the CWS trials CWS-96 and CWS-2002P (Universitätsklinikum Tübingen Medizinische Fakultät ethics approval, reference numbers 105/95 and 51/2003) and the SoTiSaR registry (ethics approval reference 158/2009B02). UK patients were enrolled under ethics approval from National Research Ethics Service Committee East of England, Cambridge Central (reference 16/EE/0394). Use of UK archival material was approved by the National Research Ethics Service Committee London Brent (reference 16/LO/0960). Additional tissue was obtained from the UK Children's Cancer and Leukaemia Group tissue bank.

Sequencing. Tumor DNA and RNA were extracted from fresh frozen tissue that had been reviewed by reference pathologists. Normal tissue DNA was derived from blood samples. Whole genome sequencing was performed by 150-bp paired-end sequencing on the Illumina HiSeq X platform. We followed the Illumina no-PCR library protocol to construct short insert libraries, prepare flowcells, and generate clusters. Coverage was at least 30×. Messenger RNA was enriched by polyA-

selection and sequenced on an Illumina HiSeq 2000 (paired end, 75-bp read length). DNA and RNA sequencing reads were aligned to the GRCh 37d5 reference genome using the Burrows–Wheeler transform (BWA-MEM)[31] and STAR (2.0.42)[32], respectively.

Variant detection. The Cancer Genome Project (Wellcome Trust Sanger Institute) variant calling pipeline was used to call somatic mutation and includes the following algorithms: CaVEMan (1.11.0)[33] for substitutions, an in-house version of Pindel (2.2.2; github.com/cancerit/cgpPindel)[34] for indels, BRASS (5.3.3; github.com/cancerit/BRASS) for rearrangements, and ASCAT NGS (4.0.0) for copy number aberrations[35]. RNA sequences were analyzed with an in-house pipeline (github.com/cancerit/cgpRna/wiki) which uses HTSeq[36] for gene feature counts, and a combination of TopHat-Fusion (v2.1.0)[37], STAR-fusion (v0.1.1)[32] and DeFuse (v0.7.0)[38] to detect expressed gene fusions. In addition to filters inherent to the CaVEMan algorithm, we used the following post-processing filtering criteria for substitutions: a minimum of two reads in each direction reporting the mutant allele, at least tenfold coverage at the mutant allele locus, minimum variant allele fraction 5%; no insertion or deletion called within a read length (150 bp) of the putative substitution, no soft-clipped reads reporting the mutant allele, and a median BWA alignment score of the reads reporting the mutant allele ≥140. The following variants were flagged for additional inspection for potential artifacts, germline contamination or index-jumping event: any mutant allele reported within 150 bp of another variant, any mutant allele with a population allele frequency >1 in 1000 according to any of five large polymorphism databases (ExAC, 1000 Genomes Project, ESP6500, CG46, Kaviar), variant reported in more than 10% of the tumor samples and mutant allele reported in >1% of the matched normal reads. For indels, the inbuilt filters of the Pindel algorithm, as implemented in our pipeline, were used. In addition, recurrent indels occurring in >2 samples were flagged for additional inspection.

Mutational signatures were derived using principal component analysis and non-negative matrix factorization as implemented in the SomaticSignatures R package[39].

Variant validation. The Cancer Genome Project (Wellcome Trust Sanger Institute) variant calling pipeline has been continually validated and bench-marked[40,41]. We confirmed variant calling quality through manual visual inspection of raw sequencing read for 8% of all variants called. All rearrangements reported were validated by reconstruction at base pair resolution and by cDNA reads spanning the breakpoint junction.

Analysis of mutations in cancer genes. We considered variants as potential drivers if they presented in established cancer genes[42]. Tumor suppressor coding variants were considered if they were annotated as functionally deleterious by an in-house version of VAGrENT (http://cancerit.github.io/VAGrENT/)[43] or were disruptive rearrangement breakpoints or focal (<1 Mb) homozygous deletions. Mutations in oncogenes were considered driver events if they were located at previously reported canonical hot spots (point mutations) or amplified the intact gene. Amplifications also had to be focal (<1 Mb) and increase the copy number of oncogenes to a minimum of five copies for a diploid genome. To search for driver variants in novel cancer genes or in non-coding regions, we employed previously developed statistical methods that identify significant enrichment of mutations, taking into account various confounders such as overall mutation burden and local variation in the mutability of the genomic region[44].

Targeted mutation screening. RNA from frozen tumors (1 µg) or corresponding to approximately 5 cm^2 of 10 µm FFPE sections was reverse transcribed using oligo-dT or random hexamer primers (RevertAid first strand cDNA synthesis kit, ThermoFisher). PCR screening was performed using primer combinations that allow amplification of candidate alterations as well as additional control fragments from the unaffected allele to assess cDNA quality. Amplified fragments were sequenced by Sanger sequencing (GATC, Konstanz, Germany) using primers detailed in Supplementary Table 1.

Immunohistochemistry. Immunohistochemical staining for phospho-ERK1/2 (Cell Signaling Technology, clone D13.14.4E) was performed according to standard protocol (dilution 1:800, pre-treatment with target retrieval TR6.1, Dako). Results were scored in a semi-quantitative fashion (negative, weak, moderate, strong).

Code availability. The algorithms used to analyze sequencing data are available at http://cancerit.github.io/.

Data availability. All data supporting the findings of this study are available within the article and its supplementary files or from the corresponding author on reasonable request. Sequencing data have been deposited at the European Genome-Phenome Archive (http://www.ebi.ac.uk/ega/) that is hosted by the European Bioinformatics Institute (accession numbers EGAS00001002534 and EGAS00001002171).

Received: 15 December 2017 Accepted: 14 May 2018
Published online: 18 June 2018

References

1. Marsden, H. B. & Lawler, W. Primary renal tumours in the first year of life. A population based review. *Virchows Arch. A. Pathol. Anat. Histopathol.* **399**, 1–9 (1983).
2. Glick, R. D. et al. Renal tumors in infants less than 6 months of age. *J. Pediatr. Surg.* **39**, 522–525 (2004).
3. Charles, A. K., Vujanic, G. M. & Berry, P. J. Renal tumours of childhood. *Histopathology* **32**, 293–309 (1998).
4. Rubin, B. P. et al. Congenital mesoblastic nephroma t(12;15) is associated with ETV6-NTRK3 gene fusion: cytogenetic and molecular relationship to congenital (infantile) fibrosarcoma. *Am. J. Pathol.* **153**, 1451–1458 (1998).
5. Knezevich, S. R. et al. ETV6-NTRK3 gene fusions and trisomy 11 establish a histogenetic link between mesoblastic nephroma and congenital fibrosarcoma. *Cancer Res.* **58**, 5046–5048 (1998).
6. Adam, L. R., Davison, E. V., Malcolm, A. J., Pearson, A. D. & Craft, A. W. Cytogenetic analysis of a congenital fibrosarcoma. *Cancer Genet. Cytogenet.* **52**, 37–41 (1991).
7. Schofield, D. E., Yunis, E. J. & Fletcher, J. A. Chromosome aberrations in mesoblastic nephroma. *Am. J. Pathol.* **143**, 714–724 (1993).
8. Church, A. J. et al. Recurrent EML4-NTRK3 fusions in infantile fibrosarcoma and congenital mesoblastic nephroma suggest a revised testing strategy. *Mod. Pathol.* **31**, 463–473 (2018).
9. Gooskens, S. L. et al. Congenital mesoblastic nephroma 50 years after its recognition: a narrative review. *Pediatr. Blood Cancer* **64**, e26437 (2017).
10. Orbach, D. et al. Infantile fibrosarcoma: management based on the European experience. *J. Clin. Oncol.: Off. J. Am. Soc. Clin. Oncol.* **28**, 318–323 (2010).
11. Soule, E. H. & Pritchard, D. J. Fibrosarcoma in infants and children: a review of 110 cases. *Cancer* **40**, 1711–1721 (1977).
12. Drilon, A. et al. Efficacy of larotrectinib in TRK fusion-positive cancers in adults and children. *N. Engl. J. Med.* **378**, 731–739 (2018).
13. Alexandrov, L. B. et al. Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).
14. Gallant, J. N. et al. EGFR kinase domain duplication (EGFR-KDD) is a novel oncogenic driver in lung cancer that is clinically responsive to afatinib. *Cancer Discov.* **5**, 1155–1163 (2015).
15. Kao, Y. C. et al. Recurrent BRAF gene fusions in a subset of pediatric spindle cell sarcomas: expanding the genetic spectrum of tumors with overlapping features with infantile fibrosarcoma. *Am. J. Surg. Pathol.* **42**, 28-38 (2018).
16. Johnson, D. B. et al. BRAF internal deletions and resistance to BRAF/MEK inhibitor therapy. *Pigment Cell Melanoma Res.* **31**, 432-436 (2018).
17. Karoulia, Z., Gavathiotis, E. & Poulikakos, P. I. New perspectives for targeting RAF kinase in human cancer. *Nat. Rev. Cancer* **17**, 676–691 (2017).
18. Jones, D. T. et al. Tandem duplication producing a novel oncogenic BRAF fusion gene defines the majority of pilocytic astrocytomas. *Cancer Res.* **68**, 8673–8677 (2008).
19. Ross, J. S. et al. The distribution of BRAF gene fusions in solid tumors and response to targeted therapy. *Int. J. Cancer* **138**, 881–890 (2016).
20. Wong V. et al. Evaluation of a congenital infantile fibrosarcoma by comprehensive genomic profiling reveals an LMNA-NTRK1 gene fusion responsive to crizotinib. *J. Natl Cancer Inst.* **108**, djv307 (2016).
21. Davis, J. L. et al. Infantile NTRK-associated Mesenchymal Tumors. *Pediatr. Dev. Pathol.* **21**, 68–78 (2018).
22. Sartore-Bianchi, A. et al. Sensitivity to entrectinib associated with a novel LMNA-NTRK1 gene fusion in metastatic colorectal cancer. *J. Natl Cancer Inst.* **108**, djv306 (2016).
23. Doebele, R. C. et al. An oncogenic NTRK fusion in a patient with soft-tissue sarcoma with response to the tropomyosin-related kinase inhibitor LOXO-101. *Cancer Discov.* **5**, 1049–1057 (2015).
24. Cook, P. J. et al. Somatic chromosomal engineering identifies BCAN-NTRK1 as a potent glioma driver and therapeutic target. *Nat. Commun.* **8**, 15987 (2017).
25. Roy, A. et al. Recurrent internal tandem duplications of BCOR in clear cell sarcoma of the kidney. *Nat. Commun.* **6**, 8891 (2015).
26. Zhang, J. et al. Whole-genome sequencing identifies genetic alterations in pediatric low-grade gliomas. *Nat. Genet.* **45**, 602–612 (2013).
27. Jones, D. T. et al. Oncogenic RAF1 rearrangement and a novel BRAF mutation as alternatives to KIAA1549:BRAF fusion in activating the MAPK pathway in pilocytic astrocytoma. *Oncogene* **28**, 2119–2123 (2009).
28. Jones, D. T. et al. Recurrent somatic alterations of FGFR1 and NTRK2 in pilocytic astrocytoma. *Nat. Genet.* **45**, 927–932 (2013).
29. Paugh, B. S. et al. Genome-wide analyses identify recurrent amplifications of receptor tyrosine kinases and cell-cycle regulatory genes in diffuse intrinsic

pontine glioma. *J. Clin. Oncol.: Off. J. Am. Soc. Clin. Oncol.* **29**, 3999–4006 (2011).
30. Bolouri, H. et al. The molecular landscape of pediatric acute myeloid leukemia reveals recurrent structural alterations and age-specific mutational interactions. *Nat. Med.* **24**, 103–112 (2018).
31. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows −Wheeler transform. *Bioinformatics (Oxford, England)* **26**, 589–595 (2010).
32. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics (Oxford, England)* **29**, 15–21 (2013).
33. Jones, D. et al. cgpCaVEManWrapper: simple execution of CaVEMan in order to detect somatic sngle nucleotide variants in NGS data. *Curr. Protoc. Bioinforma.* **56**, 15.10.11–15.10.18 (2016).
34. Raine, K. M. et al. cgpPindel: identifying somatically acquired insertion and dletion events from paired end sequencing. *Curr. Protoc. Bioinform.* **52**, 15.17.11–15.17.12 (2015).
35. Raine, K. M. et al. ascatNgs: identifying somatically acquired copy-number alterations from whole-genome sequencing data. *Curr. Protoc. Bioinforma.* **56**, 15.19.11–15.19.17 (2016).
36. Anders, S., Pyl, P. T. & Huber, W. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics (Oxford, England)* **31**, 166–169 (2015).
37. Kim, D. & Salzberg, S. L. TopHat-Fusion: an algorithm for discovery of novel fusion transcripts. *Genome Biol.* **12**, R72 (2011).
38. McPherson, A. et al. deFuse: an algorithm for gene fusion discovery in tumor RNA-Seq data. *PLoS Comput. Biol.* **7**, e1001138 (2011).
39. Gehring, J. S., Fischer, B., Lawrence, M. & Huber, W. SomaticSignatures: inferring mutational signatures from single-nucleotide variants. *Bioinformatics (Oxford, England)* **31**, 3673–3675 (2015).
40. Nik-Zainal, S. et al. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature* **534**, 47–54 (2016).
41. Behjati, S. et al. Recurrent mutation of IGF signalling genes and distinct patterns of genomic rearrangement in osteosarcoma. *Nat. Commun.* **8**, 15936 (2017).
42. Forbes, S. A. et al. COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Res.* **45**, D777–d783 (2017).
43. Menzies, A. et al. VAGrENT: Variation Annotation Generator. *Curr. Protoc. Bioinform.* **52**, 15.8.1–15.8.11 (2015).
44. Martincorena, I. et al. Universal patterns of selection in cancer and somatic tissues. *Cell* **171**, 1029–1041.e1021 (2017).

Acknowledgements

This work was supported by funding by the Wellcome Trust, St. Baldrick's Foundation, the Deutsche Forschungsgemeinschaft (GE 539/13-1), the Deutsche Krebshilfe (50-2709-Gr2, T9/96/TrI, 50-2721-Tr2) and NIHR GOSH BRC. G.C., S.B., C.G., P.J.C., and M.R.S. received personal fellowships from the Wellcome Trust. The Cooperative Weichteilsarkom Studiengruppe (CWS) was additionally supported by the Deutsche Kinderkrebsstiftung (SoTiSaR, A2007/13DKS2009.08) and by the Förderkreis Krebskranke Kinder e.V. Stuttgart, Germany. The SIOP-RTSG/GPOH-nephroblastoma study group is supported by the charity "Elterninitiative krebskranker Kinder im Saarland e.V.". We thank children and their families for participating in our research and the clinical teams involved in their care. We thank Sabine Roth and Sharna Lunn for expert technical assistance.

Author contributions

J.W., G.C., M.D.C.V.H., and C.G. analyzed sequencing data. C.V. performed histological analyses. S.Ba., H.S., and B.Z. provided technical assistance. S.J.F., M.J., J.A., O.S., C.D., R.F., N.G., D.T.W.J., C.K., S.M.P., W.M., E.K., N.S., A.R. and M.S.-S. curated and reviewed the samples, clinical data, and/or provided clinical expertise. M.R.S. and P.J.C. contributed to discussions. M.G. and S.B. directed this research and wrote the manuscript, with contributions from G.C., J.W., and M.D.C.V.H.

Additional information

Supplementary Information accompanies this paper at https://doi.org/10.1038/s41467-018-04650-6.

Competing interests: The authors declare no competing interests.

Recurrent histone mutations in T-cell acute lymphoblastic leukaemia

Mutations affecting key modifiable histone type 3 (H3; Table SI) residues are frequent oncogenic events in certain solid tumours (Feinberg *et al*, 2016), and have also recently been implicated in a subset of acute myeloid leukaemia (AML) (Lehnertz *et al*, 2017). Here, we systematically reviewed the somatic mutations in >20 000 cancer specimens to identify tumours harbouring H3 mutations. In a subset of T-cell acute lymphoblastic leukaemia (T-ALL) we identified non-methionine mutations of the key modifiable H3 residues, lysine (K) 27 and 36.

The starting point of our investigation was a search for H3 hotspot mutations in 1020 human cancer cell lines (Table SII). In two cell lines, both derived from T-ALL, we found lysine-to-arginine mutations at H3K27 and H3K36 (Table I). One of the cell lines, LOUCY, is derived from a *NOTCH1* wild-type adult T-ALL (Ben-Bassat *et al*, 1990). The second, CML-T1, was derived from the T-lymphoblastic blast crisis of chronic myeloid leukaemia (Kuriyama *et al*, 1989). Ten further T-ALL cell lines lacked coding H3 mutations (Table SIII). In solid tumours, H3K27 and H3K36 are typically mutated to methionine (Fig 1) (Feinberg *et al*, 2016). However, recent functional studies of H3 lysine-to-isoleucine mutations in AML demonstrate that the latter also dramatically alter global H3 methylation and acetylation patterns (Lehnertz *et al*, 2017). Therefore, we speculated that lysine-to-non-methionine mutations may also be drivers of a subset of T-ALL.

We next searched for canonical H3 mutations in a published targeted sequencing study of 633 epigenetic regulator genes in >1000 childhood tumours encompassing 21 cancer subtypes (Huether *et al*, 2014). Amongst 91 T-ALL specimens, there were two cases with canonical H3 mutations: *H3F3A* p.K27R and *H3F3A* p.K36R (Table I). Both mutations were clonal, with a variant allele fraction (VAF) of 38% and 55%, respectively. Among the 37 tumours with H3K mutations, lysine-to-arginine mutations were restricted to T-ALL ($P = 0{\cdot}001502$; Fisher's exact test).

We then extended our screen for H3 mutations to 18 704 tumours, encompassing >60 cancer types other than T-ALL (Tables SIV and SV). This dataset comprised 8764 internally sequenced specimens and 9940 TCGA samples re-analysed using an in-house variant calling pipeline as previously described (Martincorena *et al*, 2017). We identified only one neomorphic H3 mutation in an acute leukaemia specimen: a previously reported *HIST1H3D* p.K27M mutation in an adult AML case (TCGA-AB2927-03) (Lehnertz *et al*, 2017).

Finally, we examined an additional T-ALL cohort by capillary sequencing of recurrently mutated modifiable residues K27, G34, and K36 across four frequently mutated H3 genes (Tables SVI and SVII). The cohort comprised 38 T-ALL cases described in detail previously (Maser *et al*, 2007). One specimen from a 30-year-old patient harboured a *H3F3A* p.K27N mutation (Figure S1). Interestingly, a *H3F3A* p.K27N mutation and a *H3F3A* p.K27T variant were previously identified in a T-ALL RNA sequencing study ($n = 31$) (Atak *et al*, 2013). Collectively, our findings indicate that H3K27 and H3K36 mutations are recurrent in T-ALL, a result we were able to reproduce across multiple different cohorts encompassing adult and paediatric cases.

This finding is congruent with the fact that mutations in *SETD2* and *EZH2*, methyltransferases that catalyse trimethylation (me3) of H3K36 and H3K27, respectively, are frequent T-ALL drivers (Belver & Ferrando, 2016). Disruptive *SETD2* alterations occur in 7·8% of early T cell precursor acute lymphoblastic leukaemia (ETP-ALL), an aggressive subtype with stem cell-like features (Belver & Ferrando, 2016). Interestingly, both T-ALL specimens with H3K36R mutations originated from ETP-ALL (Table I). Notably, mutually exclusive *SETD2* and H3K36/H3K34 mutations are reported in paediatric high grade glioma, where both result in reduced H3K36me3 mediated by *SETD2* (Feinberg *et al*, 2016). It is unclear whether a similar co-mutation pattern exists in T-ALL, as H3 genes have not been included in targeted sequencing panels used by the largest T-ALL genomic studies (Belver & Ferrando, 2016).

The role of H3K27 modifications in T-ALL pathogenesis is complex (Belver & Ferrando, 2016). It is plausible that mutations affecting this residue could impact the activity of several histone modifiers with established roles in T-ALL pathogenesis. Loss-of-function mutations in *EZH2* or other core components of Polycomb repressive complex 2 (PRC2) are found in 42% of ETP-ALL and 25% of T-ALL overall (Belver & Ferrando, 2016). Impaired PRC2 catalytic activity in T-ALL is associated with reduced H3K27me3, stemness and poor prognosis (Belver & Ferrando, 2016). *H3F3A* p.K27M mutations appear to act predominantly by blocking H3K27 di- and trimethylation and increasing H3K27 acetylation (Feinberg *et al*, 2016). Recent work demonstrates that H3K27I mutations in AML are associated with similar changes in H3 modification patterns (Lehnertz *et al*, 2017), suggesting that other non-methionine mutations at modifiable H3 residues may influence the activity of PRC2 and

doi: 10.1111/bjh.15155

Table I. Type 3 histone mutations in T cell leukaemia.

Sample name	Sample type	Donor age (years)	Donor sex	H3 mutation
LOUCY	Cell line derived from ETP-ALL	38	Female	*HIST1H3G* p.K36R
CML-T1	Cell line derived from the acute T-lympoblastic blast crisis of CML	36	Female	*H3F3A* p.K27R
SJTALL174	Primary ETP-ALL specimen	Unknown (paediatric)	Unknown	*H3F3A* p.K36R
SJTALL080	Primary T-ALL specimen	Unknown (paediatric)	Unknown	*H3F3A* p.K27R
PD2752a	Primary T-ALL specimen	30	Male	*H3F3A* p.K27N

Out of 141 T cell leukaemia specimens screened (12 cell lines and 129 primary samples), 5 (3·5%) harboured a missense mutation at a modifiable lysine residues K27 or K36. CML, chronic myeloid leukaemia; ETP-ALL, early T cell precursor acute lymphoblastic leukaemia; T-ALL, T cell acute lymphoblastic leukaemia.

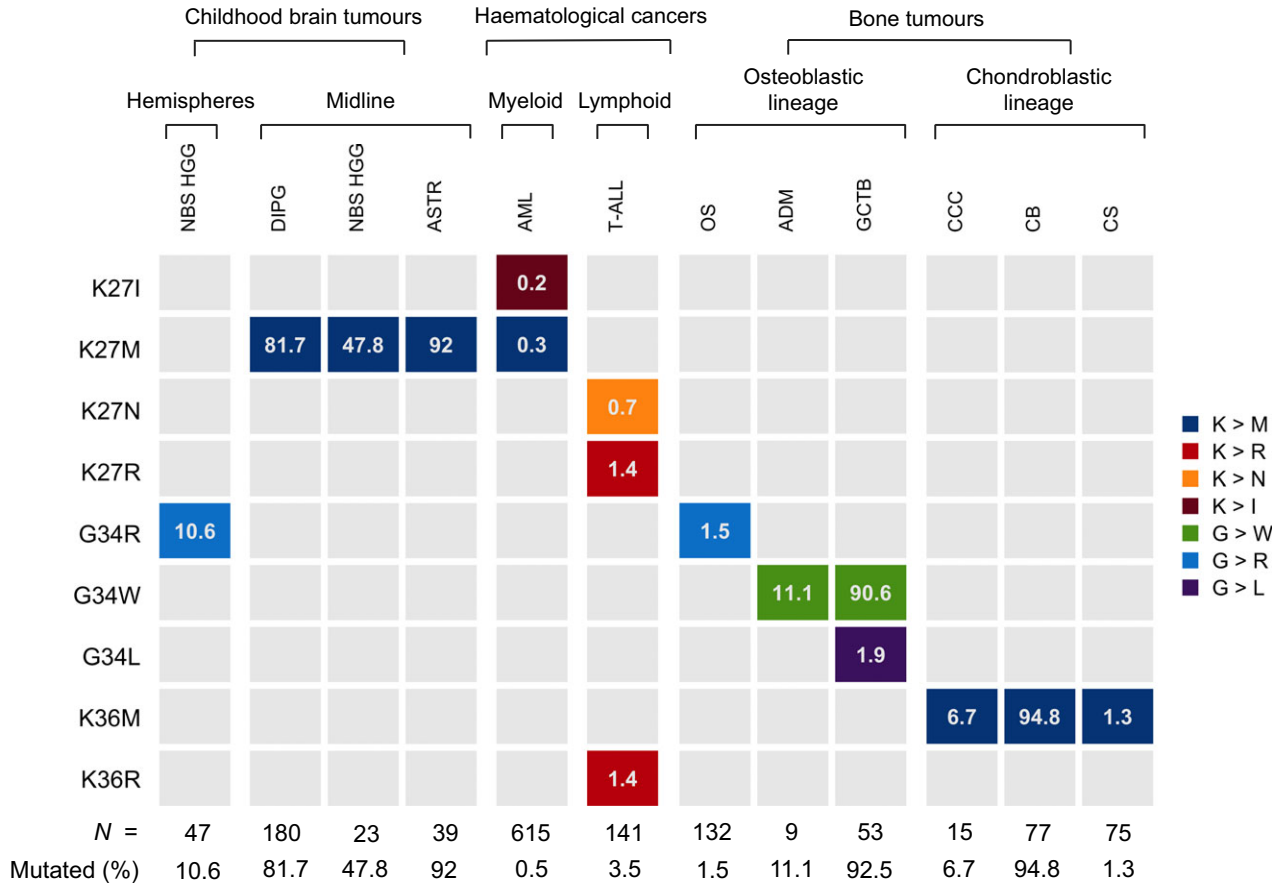

Fig 1. Prevalence and amino acid specificity of type 3 histone mutations in different cancer types. Columns indicate cancer types and rows show key histone type 3 regulatory residues. Tiles are coloured according to amino acid substitution. The percentage of each tumour type affected by the given class of histone mutation is indicated within the tiles and the overall prevalence of histone mutations is summarised at the bottom of each column. NBS HGG, non-brain stem high grade glioma; DIPG, diffuse intrinsic pontine glioma; ASTR, astrocytoma; AML, acute myeloid leukaemia; T-ALL, T cell acute lymphoblastic leukaemia; OS, osteosarcoma; ADM, adamantinoma; GCTB, giant cell tumour of bone; CCC, clear cell chondrosarcoma; CB, chondroblastoma; CS, chondrosarcoma.

other histone modifying enzymes. The lysine-specific demethylases *JMJD3* and *UTX* are further important regulators of H3K27me3 distribution in T-ALL (Belver & Ferrando, 2016), and it is conceivable that these enzymes may also be affected by H3K27 or H3K36 mutations.

A feature of H3 mutations in solid cancers is their exquisite tumour type specificity (Fig 1) (Feinberg *et al*, 2016). In this context, it is notable that 5/5 H3 mutations in T-ALL identified by this survey are lysine-to-non-methionine mutations, and 4/5 are lysine-to-arginine mutations. Out of the >20 000 tumour specimens screened for H3 variants, only two other samples harboured H3 lysine-to-arginine mutations, both at low VAF and in tumours with relatively high coding mutation burdens (TCGA-BT-A20Q-01 and TCGA-

AN-A0FW-01). Hence, it is possible that lysine-to-arginine mutations confer particular selective advantage in the context of T cell leukaemogenesis.

In summary, ~3% of T-ALL harbour non-methionine variants in H3 genes at key modifiable lysine residues. Given the role of dysregulated H3K27/H3K36 modification in T-ALL pathogenesis and the established prognostic significance of mutations in lysine-specific histone modifiers (Belver & Ferrando, 2016), this finding warrants further investigation of the prevalence, clinical and functional significance of H3 mutations in T-ALL. In light of the recent discovery of oncogenic H3K37 mutations in AML (Lehnertz *et al*, 2017), our findings suggest a broader role for histone mutations in acute leukaemias and clearly justify incorporation of H3 genes into haematological cancer sequencing panels.

Acknowledgments

This work was supported by the Wellcome Trust. S.B. was funded by a Wellcome Trust Intermediate Clinical Research Fellowship and a St. Baldrick's Foundation Robert J. Arceci Innovation Award; G.C. by a Wellcome Trust Clinical PhD Fellowship (WT098051); N.B. by AIRC (Associazione Italiana per la Ricerca sul Cancro) through a MFAG (n.17658); G.S.V by a Wellcome Trust Senior Fellowship in Clinical Science (WT095663MA) and P.J.C. by a Wellcome Trust Senior Clinical Research Fellowship (WT088340MA). We thank Professor Adele Fielding for providing samples.

Authorship

S.B., M.R.S. and P.J.C. conceived and designed the study. G.C. and S.B. performed analysis with input from M.Y., I.M. and N.B. L.F. contributed materials. G.C. and S.B. wrote the manuscript with contributions from G.S.V. and P.J.C.

Conflict of interest

The authors have no competing financial interests to declare.

Grace Collord[1,2]
Inigo Martincorena[1]
Matthew D. Young[1]
Letizia Foroni[3,4]
Niccolo Bolli[5,6]
Michael R. Stratton[1]
George S. Vassiliou[1,7]
Peter J. Campbell[1,7]
Sam Behjati[1,2]

[1]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire, [2]Department of Paediatrics, University of Cambridge, Cambridge, [3]Centre for Haematology, Faculty of Medicine, Imperial College London, [4]Clinical Haematology, Imperial College Healthcare NHS Trust, London,UK, [5]Department of Oncology and Haemato-Oncology, University of Milan, [6]Department of Oncology and Haematology, Fondazione IRCCS Istituto Nazionale dei Tumori, Milan,Italy and [7]Department of Haematology, University of Cambridge, Cambridge, UK.
E-mails: pc8@sanger.ac.uk; sb31@sanger.ac.uk

Keywords: acute leukaemia, cancer genetics, aetiology, haematological malignancy

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Figure S1. Histone 3 mutation in T-ALL validation cohort.

Table SI. Type 3 histone genes.

Table SII. COSMIC version 81 cell lines screened for type 3 histone mutations.

Table SIII. T-cell leukaemia lines screened for type 3 histone mutations.

Table SIV. Internal database screened for histone 3 mutations.

Table SV. TCGA cohort screened for histone 3 mutations.

Table SVI. Validation cohort of 38 primary human T-ALL specimens screened by Sanger sequencing of histone 3 genes.

Table SVII. Primers used to Sanger sequence hotspot residues in histone 3 genes.

References

Atak, Z.K., Gianfelici, V., Hulselmans, G., De Keersmaecker, K., Devasia, A.G., Geerdens, E., Mentens, N., Chiaretti, S., Durinck, K., Uyttebroeck, A., Vandenberghe, P., Wlodarska, I., Cloos, J., Foa, R., Speleman, F., Cools, J. & Aerts, S. (2013) Comprehensive analysis of transcriptome variation uncovers known and novel driver events in T-cell acute lymphoblastic leukemia. *PLoS Genetics*, **9**, e1003997.

Belver, L. & Ferrando, A. (2016) The genetics and mechanisms of T cell acute lymphoblastic leukaemia. *Nature Reviews Cancer*, **16**, 494–507.

Ben-Bassat, H., Shlomai, Z., Kohn, G. & Prokocimer, M. (1990) Establishment of a human T-acute lymphoblastic leukemia cell line with a (16;20) chromosome translocation. *Cancer Genetics and Cytogenetics*, **49**, 241–248.

Feinberg, A.P., Koldobskiy, M.A. & Gondor, A. (2016) Epigenetic modulators, modifiers and mediators in cancer aetiology and progression. *Nature Reviews Genetics*, **17**, 284–299.

Huether, R., Dong, L., Chen, X., Wu, G., Parker, M., Wei, L., Ma, J., Edmonson, M.N., Hedlund, E.K., Rusch, M.C., Shurtleff, S.A., Mulder, H.L., Boggs, K., Vadordaria, B., Cheng, J., Yergeau, D., Song, G., Becksfort, J., Lemmon, G., Weber, C., Cai, Z., Dang, J., Walsh, M., Gedman, A.L., Faber, Z., Easton, J., Gruber, T., Kriwacki, R.W., Partridge, J.F., Ding, L., Wilson, R.K., Mardis, E.R., Mullighan, C.G., Gilbertson, R.J., Baker, S.J., Zambetti, G., Ellison, D.W., Zhang, J. & Downing, J.R. (2014) The landscape of somatic mutations in epigenetic regulators across 1,000

paediatric cancer genomes. *Nature Communications*, **5**, 3630.

Kuriyama, K., Gale, R.P., Tomonaga, M., Ikeda, S., Yao, E., Klisak, I., Whelan, K., Yakir, H., Ichimaru, M., Sparkes, R.S. (1989) CML-T1: a cell line derived from T-lymphocyte acute phase of chronic myelogenous leukemia. *Blood*, **74**, 1381–1387.

Lehnertz, B., Zhang, Y.W., Boivin, I., Mayotte, N., Tomellini, E., Chagraoui, J., Lavallee, V.P., Hebert, J. & Sauvageau, G. (2017) H3(K27M/I) mutations promote context-dependent transformation in acute myeloid leukemia with RUNX1 alterations. *Blood*, **130**, 2204–2214.

Martincorena, I., Raine, K.M., Gerstung, M., Dawson, K.J., Haase, K., Van Loo, P., Davies, H., Stratton, M.R. & Campbell, P.J. (2017) Universal patterns of selection in cancer and somatic tissues. *Cell*, **171**, e1021.

Maser, R.S., Choudhury, B., Campbell, P.J., Feng, B., Wong, K.K., Protopopov, A., O'Neil, J., Gutierrez, A., Ivanova, E., Perna, I., Lin, E., Mani, V., Jiang, S., McNamara, K., Zaghlul, S., Edkins, S., Stevens, C., Brennan, C., Martin, E.S., Wiedemeyer, R., Kabbarah, O., Nogueira, C., Histen, G., Aster, J., Mansour, M., Duke, V., Foroni, L., Fielding, A.K., Goldstone, A.H., Rowe, J.M., Wang, Y.A., Look, A.T., Stratton, M.R., Chin, L., Futreal, P.A. & DePinho, R.A. (2007) Chromosomally unstable mouse tumours have genomic alterations similar to diverse human cancers. *Nature*, **447**, 966–971.

Leukemia
https://doi.org/10.1038/s41375-018-0270-2

BRIEF COMMUNICATION

Mechanisms of resistance

Targeting MEK in vemurafenib-resistant hairy cell leukemia

Rebecca Caeser[1,2] · Grace Collord[3,4] · Wen-Qing Yao[5] · Zi Chen[5] · George S. Vassiliou[1,3] · Philip A. Beer[3] · Ming-Qing Du[5] · Mike A. Scott[6] · George A. Follows[7] · Daniel J. Hodson[1,2,7]

Received: 17 July 2018 / Revised: 13 August 2018 / Accepted: 20 August 2018

Hairy cell leukemia (HCL) is a chronic, incurable B cell malignancy that presents with splenomegaly, bone marrow infiltration, and cytopenias [1]. Long remissions are typically achieved with purine analogs such as cladribine, but most cases will relapse and require further therapy. The discovery of the *BRAF* V600E mutation in almost all cases of HCL [2] has led to the widespread adoption of the BRAF inhibitor vemurafenib for treatment of patients relapsing after cladribine [3–5]. Impressive responses are reported; however, relapse is inevitable [5, 6] and hematologists are now faced with a growing number of patients with vemurafenib-resistant HCL. The optimal management of these patients remains unclear.

The molecular basis of vemurafenib resistance has been extensively investigated in recent years in patients with *BRAF* mutant solid organ malignancies such as melanoma and colorectal cancer [7]. Resistance to vemurafenib in melanoma frequently results from reactivation of ERK pathway signaling by a variety of genetic mechanisms that include activating mutations of *NRAS* or *KRAS*, amplification of mutant *BRAF*, aberrant splicing of *BRAF*, and activating mutation of *MAP2K1*, which encodes the MEK1 protein [7, 8]. ERK-independent mechanisms are less frequent and include activation of PI3K signaling [7]. The predominance of genetic mechanisms converging on ERK reactivation has led to the successful use of dual MEK/BRAF inhibition in melanoma [9]. In colorectal cancer however, different mechanisms apply; here primary resistance usually results from reduced feedback inhibition upon upstream receptor tyrosine kinase activity leading to restoration of ERK activity [10]. In this scenario, combined BRAF and MEK inhibition has not proved effective [11].

In contrast to our comprehensive understanding in solid organ cancer, very little is known about the mechanistic basis of vemurafenib resistance in HCL. Given the diversity of resistance mechanisms established in other cancers, it is unclear which, if any, of these might predominate in HCL. Two acquired subclonal, activating *KRAS* mutations were previously reported in a single patient with vemurafenib resistance [5]. Deletions of *NF1* and *NF2* have been proposed as an alternative mechanism in another case of primary resistance [12]. The use of MEK inhibition has been suggested as a logical therapeutic strategy in patients who have reactivated ERK signaling. However, the use of MEK inhibition has never previously been reported in a patient with HCL and at present there is no consensus on the optimal management of patients relapsing on vemurafenib.

A 74-year-old patient with HCL had been treated at our institution with splenectomy, cladribine, and pentostatin. We previously reported his initial response to vemurafenib at a dose of 240 mg twice daily [4]. This dose was lower than used in the initial phase II trial [5], but has since been shown in several reports to be an effective dosing strategy for HCL [3, 13, 14]. Vemurafenib was initially stopped after 58 days; however, this was associated with rapid return of marrow infiltration and thrombocytopenia. Vemurafenib was restarted at the same dose and cytopenias rapidly

These authors contributed equally: Rebecca Caeser, Grace Collord.

Electronic supplementary material The online version of this article (https://doi.org/10.1038/s41375-018-0270-2) contains supplementary material, which is available to authorized users.

✉ Daniel J. Hodson
djh1002@cam.ac.uk

1 Department of Haematology, University of Cambridge, Cambridge, UK

2 Wellcome-MRC Cambridge Stem Cell Institute, Cambridge, UK

3 Wellcome Sanger Institute, Hinxton, UK

4 Department of Paediatrics, University of Cambridge, Cambridge, UK

5 Division of Molecular Histopathology, University of Cambridge, Cambridge, UK

6 Haematopathology & Oncology Diagnostic Service, Cambridge University Hospitals, Cambridge, UK

7 Department of Haematology, Cambridge University Hospitals, Cambridge, UK

Published online: 19 October 2018

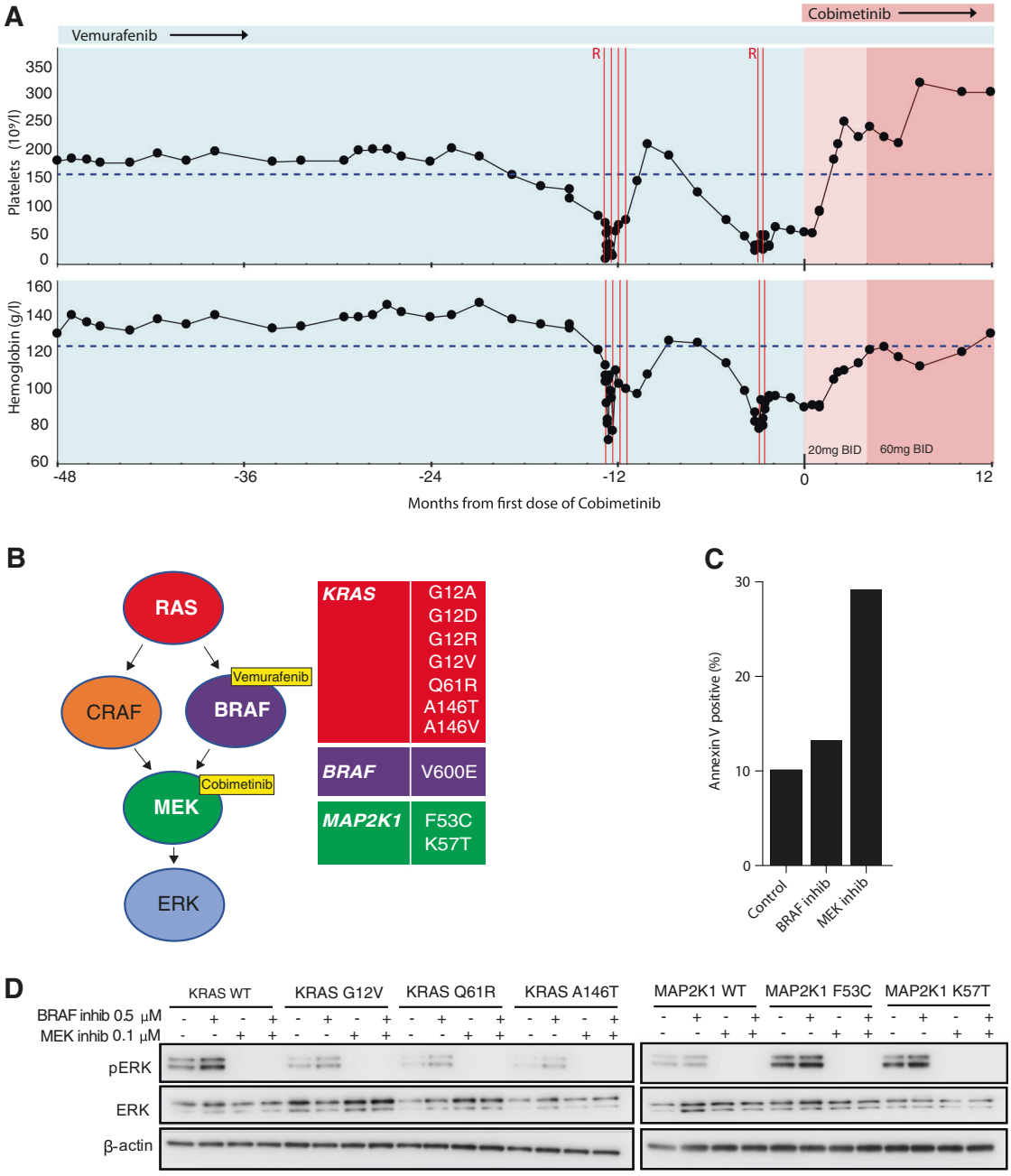

Fig. 1 **a** The patient's peripheral blood indices are shown over time relative to the first dose of the MEK inhibitor cobimetinib. Vertical red lines indicate the timing of rituximab dosing. Blue shading indicates vemurafenib monotherapy 240 mg twice daily (vem mono). Pale pink shading indicates vemurafenib with cobimetinib 20 mg daily (cobi-20). Darker pink indicates vemurafenib with cobimetinib 60 mg daily (21/28 days) (cobi-60). The lower limits of normal reference values are indicated by horizontal dashed lines. **b** Schematic of the MEK-ERK signaling pathway with mutations identified in purified tumor cells after emergence of resistance to vemurafenib. **c** Annexin V staining was used to quantify the induction of apoptosis in tumor cells purified from the patient and incubated for 48 h ex vivo with inhibitors of BRAF (vemurafenib) or MEK (trametinib). Apoptosis is induced by MEK inhibition but not by BRAF inhibition. **d** Immunoblots of a lymphoma cell line transduced with the indicated *KRAS* or *MAP2K1* constructs and incubated with inhibitors of BRAF or MEK. Complete suppression of ERK activity is seen with MEK inhibition but not with BRAF inhibition

resolved. Continuous low-dose vemurafenib continued to sustain his remission for over 3 years, attesting to the efficacy of this dosing schedule. However, 38 months after restarting vemurafenib, his blood indices deteriorated, and he required platelet transfusion (Fig. 1a). Bone marrow trephine biopsy confirmed relapse of HCL. A trial of rituximab with continued vemurafenib led to transient recovery of hematological indices. However, bone marrow infiltration did not improve over the next 4 months, and the patient became anemic, thrombocytopenic, and required

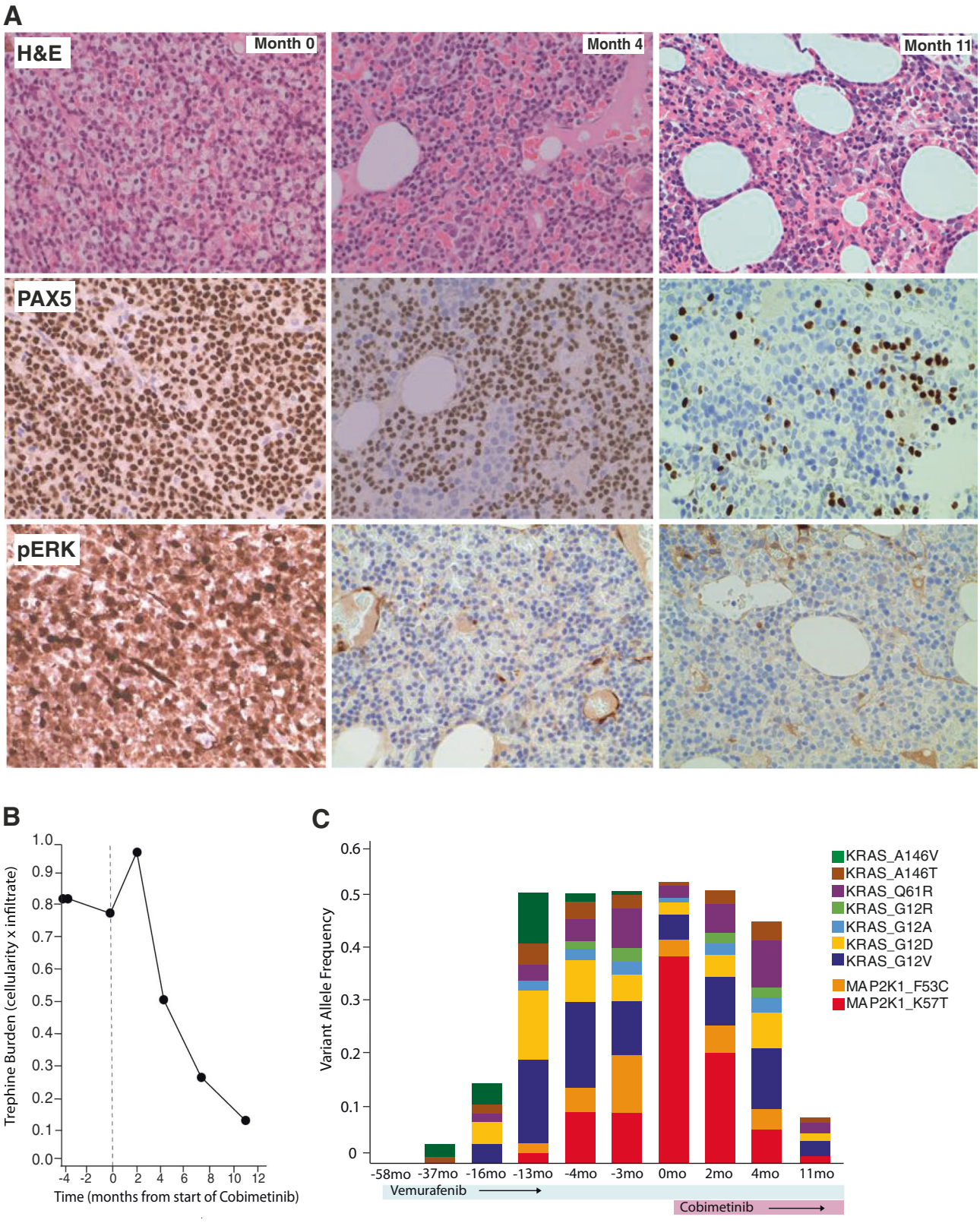

Fig. 2 **a** Bone marrow trephine biopsies stained with H&E (top) or PAX5 antibody (middle) or pERK (lower) taken at the indicated time points relative to start of cobimetinib. **b** Leukemic burden prior to and after starting cobimetinib therapy was calculated as the product of bone marrow trephine cellularity and leukemic cell infiltrate. **c** Mutant allele frequency for the indicated *KRAS* and *MAP2K1* mutations quantified by targeted amplicon sequencing at multiple time point relative to treatment

further platelet transfusion. A second trial of two doses of rituximab produced a minimal improvement of platelet count to 30×10^9/l. The patient became systemically unwell with B symptoms. Bone marrow trephine biopsy confirmed 99% infiltration with HCL.

To elucidate the mechanism of his resistance we performed whole-genome and deep-targeted sequencing of 292 genes (Supplementary Table 1) of DNA from purified tumor cells collected prior to starting vemurafenib and again at relapse. Samples were used with informed written patient consent in accordance with the Declaration of Helsinki and appropriate institutional ethical approvals. Sequencing studies revealed the presence of the known *BRAF* V600E mutation and chromosome 7q deletion. Remarkably, we also identified seven distinct activating mutations in *KRAS* and two mutations in *MAP2K1* (encoding MEK1) (Fig. 1b and Supplementary Table 2). These were detectable at relapse but were not detectable prior to vemurafenib exposure. Allele frequencies were consistent with the parallel, convergent evolution of multiple clones. Deep-targeted amplicon sequencing at multiple time points showed how *KRAS* mutations developed early, initially with codon 146 mutations, followed by emergence of the more classical activating mutations of codons 12 and 61 [15]. *MAP2K1* mutations appeared later with *MAP2K1* p.K57T expanding to become the dominant clone (Fig. 2c and Supplementary Table 2). The convergent nature of these mutations strongly implicated reactivation of MEK-ERK signaling as the likely mechanism of resistance. Indeed, immunohistochemistry confirmed strong pERK activity in all tumor cells (Fig. 2a). We looked for other mechanisms of resistance reported in melanoma. Specifically, we found no genetic or protein evidence of either increased pAKT activity or altered BRAF splicing (Supplementary Figure 1A & B).

We concluded that reactivation of MEK/ERK activity was the most likely driver of relapse and hypothesized that MEK inhibition might be a successful therapeutic strategy. Expression of the KRAS and MAP2K1 mutants in a lymphoid cell line showed that while each mutation was able to activate ERK in the presence of vemurafenib, all mutations remained sensitive to MEK inhibition (Fig. 1d). Exposure of the patient's purified tumor cells to vemurafenib ex vivo failed to completely suppress ERK activity and did not induce apoptosis. In contrast, ERK activity was completely suppressed and apoptosis induced by exposure to a MEK inhibitor (Supplementary Figure 1C and Fig. 1c).

Based on our in vitro experiments, we treated the patient with the MEK inhibitor cobimetinib, initially at 20 mg daily combined with vemurafenib 240 mg twice daily. Remarkably, B symptoms resolved within 1 week, followed by rapid platelet count recovery. A bone marrow biopsy at 4 months showed complete suppression of ERK activity (Fig. 2a). However, HCL marrow infiltration persisted, and therefore cobimetinib dose was increased to 60 mg daily (taken 21 out of 28 days). The dose was well tolerated and was associated with further resolution of cytopenias, a substantial reduction in bone marrow cellularity and HCL infiltration, ongoing suppression of ERK activity and restoration of normal hematopoiesis (Fig. 2a, b). Ongoing treatment was also associated with suppression of mutant allele frequencies for *BRAF*, *KRAS*, and *MAP2K1* mutations (Fig. 2c). At 12 months, the patient remains well and asymptomatic with continued combination therapy.

The finding of nine activating mutations, all converging upon the activation of RAS/RAF/MEK/ERK signaling, underscores the centrality of this pathway in HCL and its reactivation as a potent mechanism of resistance to vemurafenib in this disease. This report provides a detailed analysis of the molecular basis for acquired vemurafenib resistance in HCL. It is the first reported use of a MEK inhibitor to treat vemurafenib-resistant HCL. It proposes a new therapeutic option for such patients and provides impetus for testing in a formal trial setting.

Acknowledgements DJH was personally supported by a Clinician Scientist Fellowship from the Medical Research Council (MR/M008584/1), GC by a Wellcome Clinical PhD Fellowship (WT098051). W-QY was supported by an International Collaboration Award from the Pathological Society of UK and Ireland. Research in M.D. lab was supported by grants from Bloodwise. Core support was received from the Cancer Research UK, Cambridge Cancer Centre. We thank Joanna Baxter and Cambridge Blood and Stem Cell Bank for sample collection and storage, and Calli Latimer and Claire Hardy for expert technical assistance.

Author contributions RC designed and performed the experiments. GC analyzed the whole-genome sequencing and cRNA bait pulldown data. PAB designed the targeted gene pulldown panel. W-QY and ZC designed and conducted the targeted amplicon sequencing and analyzed the results. MSA, GAF, M-QD, GSV, and PAB provided clinical and diagnostic expertise and contributed to data interpretation. DJH designed the experiments, provided clinical expertise, directed the research, and wrote the manuscript with contributions from RC and GC.

Compliance with ethical standards

Conflict of interest DJH research funding: Gilead Sciences. GAF honoraria: Bayer AG, Roche, Gilead Sciences, Janssen Pharmaceuticals, and AbbVie. Consulting or advisory role: Bayer AG, Roche, Gilead Sciences, Janssen Pharmaceuticals, AbbVie. Speakers' bureau: Bayer AG, Roche, Gilead Sciences, Janssen Pharmaceuticals, and AbbVie. PAB employment: Karus.

References

1. Falini B, Martelli MP, Tiacci E. BRAF V600E mutation in hairy cell leukemia: from bench to bedside. Blood. 2016;128:1918–27.
2. Tiacci E, Trifonov V, Schiavoni G, Holmes A, Kern W, Martelli MP, et al. BRAF mutations in hairy-cell leukemia. N Engl J Med. 2011;364:2305–15.
3. Dietrich S, Glimm H, Andrulis M, von Kalle C, Ho AD, Zenz T. BRAF inhibition in refractory hairy-cell leukemia. N Engl J Med. 2012;366:2038–40.
4. Follows GA, Sims H, Bloxham DM, Zenz T, Hopper MA, Liu H, et al. Rapid response of biallelic BRAF V600E mutated hairy cell leukaemia to low dose vemurafenib. Br J Haematol. 2013;161:150–3.
5. Tiacci E, Park JH, De Carolis L, Chung SS, Broccoli A, Scott S, et al. Targeting mutant BRAF in relapsed or refractory hairy-cell leukemia. N Engl J Med. 2015;373:1733–47.
6. Holderfield M, Deuker MM, McCormick F, McMahon M. Targeting RAF kinases for cancer therapy: BRAF-mutated melanoma and beyond. Nat Rev Cancer. 2014;14:455–67.
7. Shi H, Hugo W, Kong X, Hong A, Koya RC, Moriceau G, et al. Acquired resistance and clonal evolution in melanoma during BRAF inhibitor therapy. Cancer Discov. 2014;4:80–93.
8. Poulikakos PI, Persaud Y, Janakiraman M, Kong X, Ng C, Moriceau G, et al. RAF inhibitor resistance is mediated by dimerization of aberrantly spliced BRAF(V600E). Nature. 2011;480:
387–90.
9. Larkin J, Ascierto PA, Dreno B, Atkinson V, Liszkay G, Maio M, et al. Combined vemurafenib and cobimetinib in BRAF-mutated melanoma. N Engl J Med. 2014;371:1867–76.
10. Corcoran RB, Ebi H, Turke AB, Coffee EM, Nishino M, Cogdill AP, et al. EGFR-mediated re-activation of MAPK signaling contributes to insensitivity of BRAF mutant colorectal cancers to RAF inhibition with vemurafenib. Cancer Discov. 2012;2:227–35.
11. Kopetz S, Desai J, Chan E, Hecht JR, O'Dwyer PJ, Maru D, et al. Phase II pilot study of vemurafenib in patients with metastatic BRAF-mutated colorectal cancer. J Clin Oncol. 2015;33:4032–8.
12. Durham BH, Getta B, Dietrich S, Taylor J, Won H, Bogenberger JM, et al. Genomic analysis of hairy cell leukemia identifies novel recurrent genetic alterations. Blood. 2017;130:1644–8.
13. Peyrade F, Re D, Ginet C, Gastaud L, Allegra M, Ballotti R, et al. Low-dose vemurafenib induces complete remission in a case of hairy-cell leukemia with a V600E mutation. Haematologica. 2013;98:e20–2.
14. Dietrich S, Pircher A, Endris V, Peyrade F, Wendtner CM, Follows GA, et al. BRAF inhibition in hairy cell leukemia with low-dose vemurafenib. Blood. 2016;127:2847–55.
15. Janakiraman M, Vakiani E, Zeng Z, Pratilas CA, Taylor BS, Chitale D, et al. Genomic and biological characterization of exon 4 KRAS mutations in human cancer. Cancer Res. 2010;70:5901–11.

Clonal haematopoiesis is not prevalent in survivors of childhood cancer

Clonal haematopoiesis driven by leukaemia-associated somatic mutations is a common feature of ageing (Link & Walter, 2016). This phenomenon, termed clonal haematopoiesis of indeterminate potential (CHIP), is associated with an increased risk of haematological malignancies and all-cause mortality (Link & Walter, 2016). Recent empirical evidence and computational models suggest that mutation acquisition may not be the major rate-limiting factor in the emergence of CHIP (Altrock *et al*, 2015; McKerrell *et al*, 2015; Link & Walter, 2016; Young *et al*, 2016). Instead, clonal expansion of mutant haematopoietic stem cells (HSCs) probably reflects the interaction between the effects of driver mutations and selection pressures prevailing in the bone marrow microenvironment (Link & Walter, 2016). Notably, cytotoxic therapies have been shown to favour expansion of pre-malignant haematopoietic clones (Link & Walter, 2016). Furthermore, both adult and paediatric cancer patients treated with intensive chemoradiotherapy display an earlier onset of ageing-associated morbidities and an elevated risk of therapy-related myeloid neoplasms (t-MN) and other secondary malignancies (Rowland & Bellizzi, 2014). A recent study in adult cancer patients found that CHIP was more prevalent than in the general population and was strongly associated with t-MN and overall mortality (Gibson *et al*, 2017). Although CHIP is extremely rare in healthy young individuals, its prevalence and prognostic significance in paediatric cancer patients has not been studied. We therefore performed targeted deep sequencing of peripheral blood DNA from 84 childhood cancer survivors to search for subclonal mutations common in t-MN and adult clonal haematopoiesis. No individuals with somatic variants at these loci were identified. Whilst our findings could be explained by a rarity of driver mutations, the fact that human HSCs accrue somatic variants from the first decade of life (Welch *et al*, 2012) proposes the alternative possibility that such mutations may not confer clonal advantage in the young.

We obtained peripheral blood DNA samples from patients enrolled on long-term follow-up after treatment for a paediatric malignancy and from three age-matched controls with no cancer history. Written informed consent was obtained for sample collection and DNA sequencing from all patients or their guardian in accordance with the Declaration of Helsinki and protocols approved by the relevant institutional ethics committees (approval numbers 09REG2015, 1-09/12/2015). The median age at cancer diagnosis was 4·5 years, and the commonest malignancies were acute lymphoblastic leukaemia (n = 21), neuroblastoma (n = 17) and non-Hodgkin lymphoma (n = 10). Nineteen patients had received a HSC transplant (8 allogeneic and 11 autologous). The median interval between completion of cancer treatment and blood sampling was 6 years (range 2–25). Patient characteristics are summarized in Table SI.

We performed targeted next generation sequencing (NGS) using multiplex polymerase chain reaction to amplify 32 regions of 14 genes frequently mutated in CHIP or t-MN (Table I) (McKerrell *et al*, 2015; Link & Walter, 2016; Gibson *et al*, 2017). For this we extended a previously validated assay that detected clonal haemopoiesis in 2·6% of unselected adults (McKerrell *et al*, 2015), to include all coding exons of *TP53* and *PPM1D*, genes implicated in t-MN pathogenesis (Rowland & Bellizzi, 2014; Link & Walter, 2016; Gibson *et al*, 2017). Primer design and sequencing was performed as described previously (McKerrell *et al*, 2015); see Table SII for primer sequences. Reads were aligned to human genome build GRCh37 using the Burrows-Wheeler Aligner (Li & Durbin, 2010) and analysed for somatic single nucleotide variants. Allele counts were generated using an in-house script (https://github.com/cancerit/alleleCount), considering only loci with ≥1000 reads and minimum base and mapping quality of 25 and 35, respectively. Somatic mutations with

Table I. Genomic regions sequenced.

Gene	Chromosome	Target codon/exon
NRAS	1	p.G12
SF3B1	2	p.K666; p.K700
DNMT3A	2	p.R882
IDH1	2	p.R132
KIT	4	exon 17
NPM1	5	exon 12
JAK2	9	p.V617
KRAS	12	p.G12
IDH2	15	p.R140; p.R172
PPM1D	17	exons 1–6
TP53	17	exons 1–12
SRSF2	17	p.P95
ASXL1	20	exon 12
U2AF1	21	p.S34; p.Q157

doi: 10.1111/bjh.14630

variant allele frequency (VAF) ≥0·008 (McKerrell *et al*, 2015) were sought and examined visually and by interrogation with the Shearwater algorithm (https://github.com/mg14/deepSNV) (Gerstung *et al*, 2014).

The median sequencing depth across regions of interest was $5{\cdot}3 \times 10^3$. No somatic mutations with VAF ≥ 0·008 were observed in any of our patients or controls, demonstrating that CHIP driven by mutations at these loci is not prevalent in young individuals who have received cytotoxic treatment. By contrast, Gibson *et al* (2017) identified post-chemotherapy CHIP (VAF > 0·02) in 29·9% of 401 adult lymphoma patients. Notably, mutations in *PPM1D*, a regulator of *TP53*, were the commonest CHIP drivers (Gibson *et al*, 2017). Similarly, several smaller studies have demonstrated clonal expansion in older patients undergoing chemoradiotherapy for other cancers (Link & Walter, 2016). An investigation of haematopoietic clonal dynamics in 15 adult acute myeloid leukaemia patients found that, after induction chemotherapy, five had marked expansion of clones unrelated to their leukaemia (Link & Walter, 2016). Most clones carried canonical leukaemia mutations and continued to expand years after remission (Link & Walter, 2016). In a study exploring the clonal origins of t-MN, *TP53*-mutated clones expanded dramatically after cytotoxic treatment, whereas the same mutations demonstrated very modest clonal advantage in healthy individuals (Link & Walter, 2016). In light of the above, our findings have two plausible explanations: (i) that somatic driver mutations are very uncommon in young individuals even after exposure to chemotherapy or (ii) that accrual of such mutations is insufficient to trigger clonal expansion in this age group. The latter is supported by findings that oncogenic mutations begin accumulating early in life (Welch *et al*, 2012) and that cancer-associated mutations are less able to drive clonal expansion in young compared to old stem cells (Zhu *et al*, 2016). The fact that bona-fide driver mutations do not always lead to haematopoietic clonal expansion, even after several years, was highlighted by Young *et al* (2016), using ultra-sensitive sequencing. Therefore our results should not be taken to reflect absence of potentially oncogenic HSC mutations in young cancer survivors. Rather, it is possible that even canonical leukaemogenic mutations may not commonly drive clonal outgrowth in children and young adults despite exposure to extreme haematopoietic stress, implicating age-related changes in HSCs and/or their microenvironment as key determinants of relative fitness. More sensitive DNA sequencing methods may enable detection of very rare cells harbouring known CHIP drivers mutations in similar patient cohorts, which would lend support to this hypothesis. Studies of larger numbers of paediatric cancer survivors are needed to identify rare individuals with CHIP after chemoradiotherapy, whose particular characteristics may offer insights into factors facilitating clonal outgrowth of mutated HSCs. Furthermore, in view of the shifting patterns of mutations driving CHIP in different adult age groups (McKerrell *et al*, 2015), selective pressures particular to a less mature bone marrow environment may confer clonal advantage on a distinct spectrum of somatic variants in the very young. Although a much broader screening approach is required to identify such mutations, the potential role for CHIP as a biomarker for patient risk-stratification (Gibson *et al*, 2017) may render this a worthwhile endeavour.

Acknowledgements

This project was funded by the Wellcome Trust Sanger Institute (grant number WT098051). G.S.V. is funded by a Wellcome Trust Senior Fellowship in Clinical Science (WT095663MA). F.F. is funded by Compagnia di San Paolo Grant: "Le cellule staminali del sangue nei guariti di leucemia" Codice SIME 2013-0958 (codice ROL 4201). I.V is funded by the Spanish Ministerio de Economía y Competitividad, Programa Ramón y Cajal.

Author contributions

GSV, GC and FF conceived and designed the study. NH designed sequencing assays. GC performed experiments and bioinformatics analysis. GC and GSV wrote the manuscript with input from FF. DJ and IV wrote scripts and contributed to analysis strategy. FF, MP, MD and DC contributed to sample acquisition and patient recruitment.

Grace Collord[1,2]
Naomi Park[1]
Marina Podestà[3]
Monica Dagnino[3]
Daniela Cilloni[4]
David Jones[1]
Ignacio Varela[5]
Francesco Frassoni[3,*]
George S. Vassiliou[1,6,7,*]

[1]Wellcome Trust Sanger Institute, [2]Department of Paediatrics, Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK, [3]Laboratorio Cellule Staminali e Terapie Cellulari, Istituto Giannina Gaslini IRCCS, Genova, [4]Department of Clinical and Biological Sciences, University of Turin, Turin, Italy, [5]Instituto de Biomedicina y Biotecnología de Cantabria, Cantabria, Spain, [6]Department of Haematology, University of Cambridge, and [7]Department of Haematology, Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK

E-mails: gsv20@sanger.ac.uk, francesco.l.frassoni@gmail.com

**Contributed equally.*

Keywords: haematopoiesis, late effects of therapy, haematopoietic stem cells, paediatric cancer, clonal evolution

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Table SI. Patient characteristics
Table SII. Primer sequences

References

Altrock, P.M., Liu, L.L. & Michor, F. (2015) The mathematics of cancer: integrating quantitative models. *Nature Reviews Cancer*, **15**, 730–745.

Gerstung, M., Papaemmanuil, E. & Campbell, P.J. (2014) Subclonal variant calling with multiple samples and prior knowledge. *Bioinformatics (Oxford, England)*, **30**, 1198–1204.

Gibson, C.J., Lindsley, R.C., Tchekmedyian, V., Mar, B.G., Shi, J., Jaiswal, S., Bosworth, A., Francisco, L., He, J., Bansal, A., Morgan, E.A., Lacasce, A.S., Freedman, A.S., Fisher, D.C., Jacobsen, E., Armand, P., Alyea, E.P., Koreth, J., Ho, V., Soiffer, R.J., Antin, J.H., Ritz, J., Nikiforow, S., Forman, S.J., Michor, F., Neuberg, D., Bhatia, R., Bhatia, S. & Ebert, B.L. (2017) Clonal hematopoiesis associated with adverse outcomes after autologous stem-cell transplantation for lymphoma. *Journal of Clinical Oncology*, JCO2016716712. [Epub ahead of print]

Li, H. & Durbin, R. (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)*, **26**, 589–595.

Link, D.C. & Walter, M.J. (2016) 'CHIP'ping away at clonal hematopoiesis. *Leukemia*, **30**, 1633–1635.

McKerrell, T., Park, N., Moreno, T., Grove, C.S., Ponstingl, H., Stephens, J., Understanding Society Scientific Group, Crawley, C., Craig, J., Scott, M.A., Hodkinson, C., Baxter, J., Rad, R., Forsyth, D.R., Quail, M.A., Zeggini, E., Ouwehand, W., Varela, I. & Vassiliou, G.S. (2015) Leukemia-associated somatic mutations drive distinct patterns of age-related clonal hemopoiesis. *Cell Reports*, **10**, 1239–1245.

Rowland, J.H. & Bellizzi, K.M. (2014) Cancer survivorship issues: life after treatment and implications for an aging population. *Journal of Clinical Oncology*, **32**, 2662–2668.

Welch, J.S., Ley, T.J., Link, D.C., Miller, C.A., Larson, D.E., Koboldt, D.C., Wartman, L.D., Lamprecht, T.L., Liu, F., Xia, J., Kandoth, C., Fulton, R.S., McLellan, M.D., Dooling, D.J., Wallis, J.W., Chen, K., Harris, C.C., Schmidt, H.K., Kalicki-Veizer, J.M., Lu, C., Zhang, Q., Lin, L., O'Laughlin, M.D., McMichael, J.F., Delehaunty, K.D., Fulton, L.A., Magrini, V.J., McGrath, S.D., Demeter, R.T., Vickery, T.L., Hundal, J., Cook, L.L., Swift, G.W., Reed, J.P., Alldredge, P.A., Wylie, T.N., Walker, J.R., Watson, M.A., Heath, S.E., Shannon, W.D., Varghese, N., Nagarajan, R., Payton, J.E., Baty, J.D., Kulkarni, S., Klco, J.M., Tomasson, M.H., Westervelt, P., Walter, M.J., Graubert, T.A., DiPersio, J.F., Ding, L., Mardis, E.R. & Wilson, R.K. (2012) The origin and evolution of mutations in acute myeloid leukemia. *Cell*, **150**, 264–278.

Young, A.L., Challen, G.A., Birmann, B.M. & Druley, T.E. (2016) Clonal haematopoiesis harbouring AML-associated mutations is ubiquitous in healthy adults. *Nature Communications*, **7**, 12484.

Zhu, L., Finkelstein, D., Gao, C., Shi, L., Wang, Y., Lopez-Terrada, D., Wang, K., Utley, S., Pounds, S., Neale, G., Ellison, D., Onar-Thomas, A. & Gilbertson, R.J. (2016) Multi-organ mapping of cancer risk. *Cell*, **166**, 1132–1146.e7.

