# Copy number-aware deconvolution of tumor-normal DNA methylation profiles

Elizabeth Larose Cadieux[1,2], Miljana Tanić[2,3], Gareth A. Wilson[1,2], Toby Baker[1], Michelle Dietzen[1,2], Pawan Dhami[2], Heli Vaikkinen[2], Thomas B. K. Watkins[1], Nnennaya Kanu[2], Selvaraju Veeriah[2], Mariam Jamal-Hanjani[2], Nicholas McGranahan[2], Andrew Feber[2], Charles Swanton[1,4] on behalf of the TRACERx consortium, Stephan Beck[2], Jonas Demeulemeester[1,5,#], Peter Van Loo[1,#,*]

[1]The Francis Crick Institute, London, NW1 1AT, UK; [2]University College London Cancer Institute, London, WC1E 6BT, UK; [3]Institute for Oncology and Radiology of Serbia, 11000 Belgrade, Serbia; [4]University College London Cancer Institute, Cancer Research UK Lung Cancer Centre of Excellence, London, WC1E 6DD, UK; [5]Department of Human Genetics, University of Leuven, B-3000 Leuven, Belgium.

[#]These authors contributed equally

[*]To whom correspondence may be addressed:

Peter Van Loo, The Francis Crick Institute, 1 Midland Road, London, NW1 1AT, United Kingdom. Tel: +44 (0) 20 3796 1719, e-mail: Peter.VanLoo@crick.ac.uk.

## SUMMARY

Aberrant DNA methylation is a hallmark of cancer development. Bisulfite sequencing of tumor samples can be used to study changes in the cancer methylome, but data interpretation is confounded by admixed normal cells and copy number changes. Here, we introduce CAMDAC, a novel method for Copy number-Aware Methylation Deconvolution Analysis of Cancers, enabling tumor purity, allele-specific copy number and deconvolved tumor methylation rate profiling from bulk tumor samples. We apply CAMDAC to 122 multi-region samples from 38 non-small cell lung cancers profiled by reduced representation bisulfite sequencing as part of TRACERx study. CAMDAC-derived purity and copy number profiles confirm those derived from matched whole-genome and whole-exome sequencing. Purified tumor methylation rates greatly enhance the accuracy of tumor-normal and tumor-tumor differential methylation calling and, in contrast to bulk signals, capture phylogenetic relationships between tumor clones. CAMDAC directly links genetic mutations and epigenetic changes, yields insights into allele-specific methylation and reveals (epi)genetic heterogeneity in solid tumors.

# INTRODUCTION

Like DNA mutations, epigenetic alterations, such as disruption of DNA methylation and chromatin architecture, are now acknowledged as a universal feature of tumorigenesis (Feinberg et al., 2016; Flavahan et al., 2017; Shen and Laird, 2013). These changes accumulate in normal cells throughout their lifetime (Hannum et al., 2013; Hao et al., 2016; Horvath, 2013) and while most have no effect, a small subset may provide a selective advantage for the cell (Feinberg et al., 2006). Through successive gains of hallmark cancer cellular capabilities (Hanahan and Weinberg, 2011), cells may ultimately acquire a fully malignant phenotype, while continuing to evolve in response to environmental pressures (Quail and Joyce, 2013).

DNA methylation is a mitotically heritable covalent DNA modification whereby cytosines (C) are methylated to form 5-methylcytosines (5mC), mostly in a CpG context (Bird, 2002; Smith and Meissner, 2013). In vertebrates, DNA methylation occurs throughout the entire genome and may aid to suppress cryptic transcription (Costello et al., 2010) and keep mobile genetic elements in check (Deniz et al., 2019). Active regulatory regions often contain unmethylated CpG-islands, irrespective of the expression level of associated gene(s) (Deaton and Bird, 2011; Eckhardt et al., 2006; Ziller et al., 2013), although methylation of these CpG-rich features is usually anti-correlated with transcription factor binding. Genomic alterations removing binding sites enable DNA methyltransferase activity, depleting gene expression (Smith and Meissner, 2013); *vice versa*, aberrations creating new binding sites in proximity to transcription start sites may stimulate transcription.

The cancer methylome displays characteristics from its cell of origin in addition to somatic DNA methylation changes (Greger et al., 1989; Wen et al., 2009). As methylation profiling provides insight into (disease) cell states and other aspects of biology, with a readout that is not reliant on live cells nor large amounts of high quality material, they are potentially powerful biomarkers (Feber et al., 2017; Heyn and Esteller, 2012; Koch et al., 2018).

Massively parallel sequencing efforts by the Cancer Genome Atlas (TCGA) and the International Cancer Genome Consortium (ICGC) have revealed somatic alterations across

thousands of cancer genomes (ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium, 2020; Hutter and Zenklusen, 2018). Although whole-genome and reduced-representation bisulfite sequencing (RRBS) are gaining in popularity (Hansen et al., 2011; Meissner, 2005; Pfister et al., 2014; Sun et al., 2015; Ziller et al., 2013), the cancer methylome is considerably less well charted. Interpretation of bulk tumor data is hampered by admixture of normal cells and the occurrence of somatic copy number alterations (CNAs) in most cancers. While a number of methods have been developed to assess normal cell contributions (Barrett et al., 2017; Guo et al., 2017; Teschendorff et al., 2017; Zheng et al., 2014), research focus has generally been restricted to high-purity samples and CNA-quiet cancer types such as Ewing sarcoma (Sheffield et al., 2017), and malignancies of central nervous system (Capper et al., 2018) or lymphoid origins (Kulis et al., 2012; Landau et al., 2014; Nordlund et al., 2013; Oakes et al., 2016). In the absence of an approach to purify bulk tumor methylation estimates, downstream differential methylation analysis pipelines have so far ignored the combined effect of CNAs and normal contamination on tumor-normal methylation differences.

When querying intra-tumor heterogeneity and reconstructing tumor evolutionary histories, accurate estimates of tumor purity and copy number are critical (Dentro et al., 2017). Intra-tumor DNA methylation heterogeneity has been widely reported (Brocks et al., 2014; Klughammer et al., 2018; Mazor et al., 2015), but its exact role in tumor evolution and its impact on patient outcome are still unclear. Current metrics to quantify tumor heterogeneity from bisulfite sequencing data (Landan et al., 2012; Landau et al., 2014; Li et al., 2016; Sheffield et al., 2017) assume high tumor content and/or the absence of CNAs.

To address these issues, we developed a tool for Copy number-Aware Methylation Deconvolution Analysis of Cancers (CAMDAC). CAMDAC provides accurate allele-specific copy number and purity estimates from tumor RRBS data. Formalizing the relationship between methylation rates, copy number and tumor purity, CAMDAC extracts purified tumor methylomes from bulk tumor and tissue-matched normal bisulfite sequencing data. The corrected tumor methylation rates allow for accurate quantification of differential methylation, both between tumor and normal cells and between different

tumors or sampled regions. CAMDAC deconvolved tumor methylation profiles reveal intra-tumor subclonal relationships and allow to quantify intermediate allele-specific methylation signals.

## RESULTS

### Bulk tumor methylation rates are confounded by tumor purity and copy number

To study epigenetic inter- and intra-tumor heterogeneity in non-small cell lung cancer, we performed multi-region RRBS of surgically resected primary tumors for 24 lung adenocarcinoma and 14 squamous cell lung carcinoma patients (**Table S1**) from the TRACERx 100 cohort, profiling 122 tumor regions in total (range 2-7 per patient). As a normal reference, we included samples of adjacent normal tissue for all patients.

We hypothesize that normal cell admixture (tumor purity) and somatic copy number alterations affect the methylation rate of bulk tumor samples (**Figure 1A**). We set out to assess this effect knowing that previous work using whole-exome sequencing (WES) of the same samples revealed considerable variability in purity and copy number (Jamal-Hanjani et al., 2017). We selected CpG loci that are confidently unmethylated in the adjacent normal sample (posterior 99% highest density interval on methylation rate $HDI_{99}$ [0, 0.2], **Methods**), stratified them by allele-specific copy number state in each of three tumor regions with differing purity, and evaluated their bulk methylation rates (**Figure 1B**). The majority of these loci exhibit bulk methylation rates close to 0 (88% with $HDI_{99}$ [0, 0.2]), suggesting that most sites are not differentially methylated in the tumor. A second population of CpG sites is visible across all samples, the modal methylation rate of which shifts with tumor purity and copy number. We observed a similar effect evaluating CpG loci that are confidently methylated in the adjacent normal sample (**Figure S1**). These observations confirm that normal cell admixture and copy number changes need to be accounted for when interpreting tumor methylation signals. We therefore set out to infer tumor purity and copy number from bulk tumor RRBS data, and to develop a computational

model to correct methylation profiles for these confounders to extract the tumor methylation rates, unpolluted with signals from non-tumor cells.

**Allele-specific copy number analysis from tumor bisulfite sequencing data**

We modeled our approach of purity and copy number inference on existing methods that simultaneously obtain allele-specific copy number and purity estimates from array or sequencing data (Carter et al., 2012; Van Loo et al., 2010; Nik-Zainal et al., 2012). These approaches rely on measures of coverage (LogR) and allelic imbalance (B-allele frequency, BAF) at single nucleotide polymorphisms (SNPs). As RRBS is enriched for SNP-rich CpG islands (Zhou et al., 2015), we anticipated the data to be highly amenable to allele-specific copy number analysis.

The LogR at SNP loci is readily computed from the normalized read coverage of matched tumor and normal RRBS data. However, this metric of total copy number suffers from a number of biases as a function of the RRBS protocol (**Figure S2**). (i) *MspI* digestion used during library preparation results in a heterogeneous insert size distribution, with fragments ranging from just a few base pairs to hundreds of base pairs in length, depending on the distance between two CCGG recognition sequences (Sun et al., 2015). (ii) Bisulfite conversion alters the GC content of sequences, breaking standard GC-correction for the biases introduced during PCR amplification and Illumina sequencing (Benjamini and Speed, 2012). (iii) Replication timing differs across the genome and between cell types (Ryba et al., 2010). Sequences that replicate early during S-phase tend to have higher coverage than those that replicate later. We correct our LogR estimates for each of these three biases (**Methods**, **Figure S2**).

Due to the bisulfite conversion, compiling reference and alternate allele read counts at SNP loci to obtain BAF values is challenging. During library preparation, unmethylated cytosines are converted to thymine, yielding four possible bisulfite DNA strands: (complementary to) original top and (complementary to) original bottom (**Figure 2A**). However, most bisulfite sequencing protocols are directional, reducing the sequencing

output to the original top and bottom strands. In addition, as ~50% of the SNPs captured by RRBS are found at CpG dinucleotides, methylation further complicates BAF calculations.

Addressing these issues, we propose a set of allele counting and BAF calculation rules for all SNP types. We illustrate their derivation for C>A and C>T SNPs. Bases are reported as original top (+) or complementary to original bottom strand (-) (**Figure 2A-C**). For a C>A SNP at a CpG, all reads reporting A on the + or - strand, i.e. A(+) and A(-), can be uniquely attributed to the alternate allele. Likewise, C(-) derives from the reference allele, together with C(+) and T(+) reads from the methylated and unmethylated cytosine, respectively. As a result, $BAF_{C>A}$ can be calculated as $\frac{A}{A + C + T(+)}$ (**Figure 2B**). The situation is more complex for C>T SNPs, since the + strand does not allow to distinguish between the alternate allele and the bisulfite converted unmethylated reference, as both yield T(+) (**Figure 2C**). However, reads from the - strand do distinguish the (un)methylated reference, C(-), from the alternate allele, T(-). Therefore, it is still possible to quantify $BAF_{C>T}$ as $\frac{T(-)}{T(-) + C(-)}$. Through similar reasoning, we propose a set of BAF calculation rules for all types of SNPs (**Figure 2D)**. The same principles above are generalizable to non-CpG cytosine methylation, known to occur at low percentages in the mammalian genome (Schultz et al., 2015). Therefore, our BAF calculation rules are robust to both CpG and non-CpG cytosine methylation. Note that, since only one strand is informative with regard to the allelic imbalance at C>T, T>C, A>G and G>A SNPs, BAF estimates at these sites are based on lower effective sequencing depth.

We validated our approach by comparing genotype calls and BAF estimates at SNP loci on a subset of adjacent normal samples subjected to both RRBS and whole-genome sequencing (WGS, **Methods**). We calculated false positive rates (FPR) and false negative rates (FNR) for SNP calling on RRBS data using the WGS-derived genotypes as ground truth. The average FPR across all SNP types and samples was 0.3% whilst the mean FNR was 25% (**Figure S3)**. Polymorphic CCGGs perturbing or creating an *MspI* recognition motif, resulting in allele-specific fragments during RRBS library preparation, and thereby

skewing allelic coverage, were the main cause of false negatives (49%) (**Figure S4**). Within the false negative calls, we saw a bias towards SNPs erroneously called homozygous reference (72%) compared to homozygous alternate (28%) (**Figure S3B**), which points towards an alignment bias, likely due to limited mappability of short *MspI* fragments with alternate alleles.

Leveraging the BAF and LogR values, we infer tumor purity and allele-specific copy number profiles. We call this CAMDAC module ASCAT.m, as our approach relies on ASCAT (Van Loo et al., 2010), but alternative methods such as Battenberg (Nik-Zainal et al., 2012) or ABSOLUTE (Carter et al., 2012) may equally be used. We evaluated ASCAT.m copy number calling by comparing segmented BAF and LogR data, and final allele-specific copy number calls to data obtained through WGS of the same tumor samples. Direct comparison of segmented BAF and LogR tracks derived from RRBS and WGS confirms the accuracy and robustness of ASCAT.m's estimates (**Figure 2E**).

We compared the ASCAT.m tumor purity ($\rho$) and ploidy ($\psi$) values with those inferred from whole-genome sequencing (WGS, mean coverage 67x) and high-coverage exome sequencing (WES, mean coverage 464x) performed on the same samples (**Table S3**, Jamal-Hanjani et al., 2017). RRBS-derived estimates show excellent agreement (**Figure 2F-G**) with both WGS ($cor_\rho = 0.996$, $cor_\psi = 0.991$) and WES outputs ($cor_\rho = 0.984$, $cor_\psi = 0.978$). In line with previous work (Jamal-Hanjani et al., 2017), tumor purity varies widely between sampled regions of a given primary tumor, whilst ploidy is typically similar (**Figure S5**). Whole genome doubling (**Figures 2H**), is widespread across lung adenocarcinoma (LUAD, 72%) and squamous cell lung carcinoma (LUSC, 89%). Genomic instability is a defining feature of both cancer types. Whilst some copy number gains and losses are observed at a high frequency across histological subtypes, such as +8q and -8p, other patterns are strikingly different between subgroups. For example, +3q is found in virtually all LUAD samples and is absent from LUSC (**Figure 3I**).

In conclusion, we have shown that CAMDAC's ASCAT.m module allows accurate allele-specific copy number profiling and tumor purity estimation from methylation data, obviating the need to perform separate copy number profiling experiments.

**SNP-independent methylation rate estimation**

Similar to methylation at CpG SNPs affecting BAF calculations, polymorphisms confound methylation rate estimation. Polymorphisms at CpGs account for 49.9% of SNPs, 83.3% of which are CpG>TpG polymorphisms, and methylation rates at these positions using default approaches show markedly different distributions (**Figure 3A-C**). In tumor samples, this effect shows copy number dependence (**Figure 3D-F**).

To address these biases, we developed an approach to obtain methylation rates independent of SNP status (**Figure 3G-I**, **Methods**). A number of considerations need to be made for this. (i) The original top and original bottom strand in directional protocols encode methylation information for the first and second position of a CpG, respectively. As a result, in case of a CpG>TpG SNP, the unmethylated reference and alternate alleles are distinguishable only on the bottom strand. Likewise, at CpG>CpA SNPs, only the top strand may be used. (ii) Computing the methylation rate at the non-polymorphic position in the CpG, only dinucleotides allow to separate the respective contributions from the unmethylated CpG allele and the alternative allele. (iii) To ensure copy number independence of methylation rate estimates, we define them as the average methylation rate per CpG allele (**Methods**). This enables the methylation rate at a heterozygous CpG to vary between 0 and 1, rather than between 0 and 0.5 in a diploid sample, and ensures further independence between methylation rate and copy number estimates. Methylation rates per CpG are thus computed by aggregating strand-specific dinucleotides as follows: $m = \frac{\mathrm{CG}}{\mathrm{CG} + \mathrm{TG}(+) + \mathrm{CA}(-)}$, except at C>T and G>A SNPs, where only reads from the bottom strand and top strand, respectively, contribute to the estimates (**Figure 3G-I**).

When accounting for these confounders, the observed biases disappear (**Figure 3A-F**, **Figure S6**). These results confirm that, contrary to other approaches, CAMDAC CpG methylation rate calculation is robust to the presence of polymorphisms.

**Deconvolving tumor methylation profiles**

Building upon the allele-specific copy number, tumor purity and methylation rate obtained with CAMDAC, we aim to separate tumor and normal methylation signals. When a tumor clone is methylated at a CpG site that is unmethylated in the admixed normal cells (**Figures S1A-G**), the fraction of methylated reads at that locus observed from bulk bisulfite sequencing increases along with tumor purity and copy number (**Figure 1B**). *Vice versa*, if the tumor clone lost methylation compared to the normal, the observed methylation rate will decrease with increasing purity and copy number (**Figure S1H**). We formally modeled this effect, realizing that the bulk tumor methylation rate ($m_b$) is the sum of a tumor component (with methylation rate $m_t$) and a normal component (with methylation rate $m_n$), weighted by their relative DNA proportions, which can be calculated as the product of purity and copy number ($\rho$ and $1-\rho$, and $n_t$ and $n_n$ for tumor and normal, respectively):

$$m_b = \frac{\rho n_t m_t + n_n m_n (1-\rho)}{\rho n_t + n_n (1-\rho)} \qquad Eq(1)$$

Applying this equation to our lung cancer RRBS data, using the adjacent normal samples as a proxy for the admixed normal cells, enables interpretation of the observed bulk methylation signals. Indeed, the positions of peaks of clonal bi-allelic tumor-normal differential methylation align closely with those predicted under this model, for different values of tumor purity and copy number (**Figures 1B, S1H**), and this relationship holds across all tumor samples (**Figures S7A-B**).

While these data show that our CAMDAC model can account for copy number and tumor purity, not all CpG methylation values appear at the expected peaks. Allele-specific methylation is known to be widespread, not only on the inactive X chromosome in females and imprinted genes, but also at polymorphic regulatory sequences (Kerkel et al., 2008; Reik and Walter, 2001), and this signal is also apparent in the overall methylation rates (**Figures 1B, S1G-H**). In addition, intra-tumor heterogeneity is expected to contribute intermediate methylation rate signals.

Considering the above, (i) we have now formalized the relationship between the bulk and purified tumor methylation rates; (ii) we can directly estimate tumor purity and copy number for each CpG from the bisulfite converted data; and (iii) an adjacent normal sample can be used as a reasonable proxy for the methylation rate of the admixed normal cells. We therefore have all the variables in hand and can solve *Eq(1)* for the purified tumor methylation rate:

$$m_t = \frac{m_b\big(\rho n_t + n_n(1-\rho)\big) - n_n m_n (1-\rho)}{\rho n_t} \qquad Eq(2)$$

Taken together, bulk tumor methylation rates are confounded by sample purity and local tumor copy number in a way that can be modeled. Formalizing these effects, and inferring purity and copy number from bisulfite sequencing data, CAMDAC corrects bulk methylation rates to yield deconvolved tumor methylation rate estimates.

We next applied CAMDAC to the TRACERx lung cancer multi-region RRBS dataset and assessed the purified methylomes. Post-deconvolution, we would expect CpGs which have become methylated on all chromosome copies in all tumor cells to have purified tumor methylation rates close to $m_t = 1$. *Vice versa*, tumor-normal hypomethylated loci should approach $m_t = 0$. We see a high correlation between the expected and the observed purified tumor methylation rates at these clonal differentially methylated loci present on all copies (**Figure S7C**), with noise proportional to tumor sample purity (**Figures S7D-E**).

We evaluated the purified tumor methylation profiles by comparing Pearson distances between pairs of methylation profiles. Compared to bulk signals, CAMDAC deconvolved tumor methylation profiles show increased distances between patients and to matched normals, while different tumor regions of the same patient remain similar (**Figure 4A**). This confirms that shared normal signals are effectively removed from the bulk tumor, while tumor-specific signals are retained.

Leveraging single nucleotide variant (SNV) calls from newly obtained WGS data and previously published WES (Jamal-Hanjani et al., 2017), we validate CAMDAC purified tumor methylation rates by phasing CpG methylation to clonal SNVs present on all copies

in regions with loss of heterozygosity. At these sites, any read reporting the variant allele can directly be assigned to the tumor cells, and methylation rates obtained from this subset of reads should be an unbiased estimate of the purified tumor methylation rate. Overall, RRBS-derived VAF estimates of somatic SNVs (computed analogously to BAF values of germline SNPs, **Figure 2D**) were highly correlated with matched WGS/WES data (Pearson correlation = 0.86, **Figure 4B**). In total, we obtained phased methylation estimates at 4,485 CpG loci across our dataset and observed high correlation between these SNV deconvoluted $m_t$ values and CAMDAC estimates (Pearson correlation = 0.97, **Figure 4C**). These results confirm that CAMDAC can accurately deconvolute tumor methylation rates from bulk RRBS data.

**Inferring differential methylation from CAMDAC methylation rates**

We next set out to formally identify differentially methylated CpGs. For tumor-normal comparisons, we directly compute the probabilities $P(m_t > m_n)$ and $P(m_t < m_n)$ of hyper- and hypomethylation, respectively. Substituting $m_t$ by *Eq(2)* and modelling the observed bulk and normal methylated read counts in a Bayesian fashion using beta-binomial distributions, the resulting probability density is a scaled difference of two beta posteriors ($m_n$ and $m_b \sim Beta(\#reads_{meth}, \#reads_{unmeth})$), which we can compute exactly (**Methods**):

$$P(m_t > m_n) = P(C \times (m_n - m_b) < 0) \quad \text{with } C = \frac{\rho n_t + n_n(1-\rho)}{\rho n_t} \qquad Eq(3)$$

Equations *2* and *3* reveal how coverage, copy number and tumor purity affect the power to infer tumor–normal differentially methylated positions (DMPs). Directly, the variance of $m_n$ and $m_b$ decreases with increasing normal and tumor coverage, respectively. Indirectly, increasing tumor copy number results in higher local coverage but also, together with increasing purity, shifts $m_b$ away from $m_n$ at DMPs. To identify tumor–tumor DMPs, we similarly compute $P(m_{t1} > m_{t2})$ and $P(m_{t1} < m_{t2})$ by resampling $m_t$ from the posterior distributions of $m_n$ and $m_b$ in *Eq(2)* (**Methods**).

For differential methylation analysis, it is customary to set a minimal effect size threshold, which we set to $\Delta m > 0.2$. Simulations based on our observed data (**Methods**) reveal that, in contrast to $m_b$, setting this threshold using $m_t$ removes dependence on purity and copy number, with true positive bi-allelic and mono-allelic DMPs (in balanced regions) showing absolute $m_t$-$m_n$ methylation rate difference near 1 and 0.5 respectively (**Figures 5A and S8A)**. Thresholding on $m_t$ was found to greatly reduce the number of false negative tumor–normal DMP calls in impure tumors and at loci with lower copy number states, while retaining a low false positive rate (**Figures 5B-C and S8B-C**). Similarly, failure to adjust $m_b$ for purity and copy number can readily inundate tumor-tumor differential methylation results with false positive DMPs. Use of CAMDAC $m_t$ values considerably reduces false positives in tumor-tumor comparisons while retaining a similarly low rate of false negatives (**Figures 5D and S8D**).

To gauge CAMDAC performance on real data, we compared intra-tumor DMPs called using $m_b$ or $m_t$ for patient CRUK0062, selected for its large number of samples with varying tumor purity (**Figure 5E)**. In this setting, most CAMDAC $m_t$-based calls are also identified using $m_b$ and the effect of tumor purity on power can readily be seen as an increase in the number of DMPs identified with sample purity. Note however that, using $m_t$, more DMPs are called when both samples are high purity, *i.e.* statistical power is the highest. In contrast, when using $m_b$, more DMPs are called when two samples differ more in purity. These findings are in line with our simulation results and suggest that also on real data, controlling methylation rates for tumor purity and copy number greatly reduces the number of false positive DMP calls, while maintaining low false negative rates.

To get a global overview of the performance of CAMDAC $m_t$ and bulk $m_b$ for tumor-tumor differential methylation analysis, we randomly selected CpG loci from samples of low or high tumor purities both within and between patients and obtained DMP calls (**Methods**). As expected, results showed a greater number of DMPs for inter-patient comparisons than between samples of shared clonal origins (**Figure 5F**). Furthermore, DMP calls unique to the bulk tumor were frequent between samples of differing purities

taken from the same patient, suggesting a high false-positive rate of DMP calling without deconvolution.

When investigating disease-related DNA methylation changes, researchers commonly look for differentially methylated regions (DMRs) as opposed to individual CpGs (Robinson et al., 2014). Building on DMP calls, CAMDAC identifies DMRs by binning CpGs into neighborhoods and identifying DMP hotspots within these clusters (**Methods**).

Taken together, analyses of both simulated and observed data show that CAMDAC enables accurate calling of both tumor-normal and tumor-tumor differential methylation from RRBS data, accounting for both purity and copy number.

**Quantifying allele-specific methylation**

While most CpGs are either fully methylated or fully unmethylated in the tumor, others show intermediate $m_t$ values (**Figures 1B, S1G**). We hypothesize that in addition to intra-tumor epigenetic heterogeneity, part of this is due to allele- or copy-specific differential methylation, and we set out to quantify this phenomenon. Note that, by construction, allele-specific methylation at heterozygous SNPs does not contribute to this intermediate $m_t$ signal.

To assess CAMDAC sensitivity for intermediate methylation signals, we first evaluated allele-specific methylation signals in adjacent normal samples from females. Dosage compensation in females by X-chromosome inactivation involves extensive DNA methylation of promoter-associated CpGs on the inactive copy (Lyon, 1962). Unexpectedly, intermediate methylation values on X were consistently below 0.5 for the 13 normal female samples in this cohort (**Figure 6A**). In addition, we observed decreased sequencing coverage on X compared to the rest of the genome in females (**Figure 6B**). We hypothesize that DNA extraction biases against the tightly packed Barr body lead to reduced coverage of the inactive copy. BAF of heterozygous SNP loci did not show any deviation (**Figure 6C**), suggesting that X inactivation is random at the scale of our adjacent normal lung samples.

Next, we assessed our ability to detect intermediate methylation due to germline imprinting. As expected, the normal methylation rate around the imprinting control region of the *IGf2/H19* locus (Frevel et al., 1999) is confidently around $m_n$= 0.5, whilst the CpG island shore is fully methylated (**Figure 6D**). As imprinting results in allele-specific methylation, we aimed to validate this through phasing CpGs to nearby germline heterozygous SNPs. In 31/37 (84%) normal samples, one or more heterozygous SNPs overlapping the imprinted region was found, for which on average 15 imprinted CpGs where close enough to be phaseable. In each of these, we were able to confirm that methylation in the imprinted region was indeed allele-specific (**Figure 6D**). Interestingly, in 5 of the tumors, we found that imprinting was completely lost, and both parental alleles were unmethylated (**Figure 6D,E**).

We next set out to quantify allele-specific differential methylation in the tumor samples. We selected CpG loci in regions of copy number $n_t$=2 in tumor samples that are confidently unmethylated in the adjacent normal sample ($HDI_{99}$ [0, 0.2], **Methods**). We then separated them by allele-specific copy number and plotted their purified tumor methylation rates distribution, which show DMPs clustering in two peaks corresponding to epimutation copy numbers 1 and 2 (**Figure 6F-G**). From this, we define 2 classes of DNA methylation changes: (i) gene regulatory differential methylation and (ii) allele- (and copy-) specific stochastic methylation changes. Regulatory changes typically affect all copies, while stochastic methylation changes are likely passenger epimutations due to errors in DNA methylation maintenance machinery, aberrant TET activity or spontaneous deamination of methylated cytosines (Coulondre et al., 1978; Goyal, 2006; Kohli and Zhang, 2013). Assuming faithful replication thereafter, these changes become discernable after clonal expansion, generating intermediate methylation signal(s).

**Interplay of somatic mutations and methylation changes**

Next, we performed DMR calling on previously described SNV purified tumor methylation rate estimates (**Methods**). In total, 5,727 methylation bins could be phased to SNVs, of which 602 were differentially methylated (**Figure 7A**). Of these, in regions without loss of

the wild type allele, 258 and 43 were specific to the mutated allele (in *cis*) and the wild-type allele (in *trans*), respectively. This suggests that somatic mutations occasionally result in methylation changes, most of which occur in *cis* (p $=1.67 \times 10^{-33}$ Binomial test), as expected.

In-*trans* methylation may signify bi-allelic inactivation of tumor suppressors due to deleterious mutation of one allele and hypermethylation of the other allele. Alternatively, suppression of the wild type transcript may be favorable at oncogenes with activating mutations. In any case, trans-DMRs were usually hypermethylated (p $= 2.19 \times 10^{-3}$, Binomial test). For example, *MDGA1* was found to be clonally hypermethylated on one copy of the wild type allele in all CRUK0071 regions (**Figure 7B**).

In-*cis* methylation, where somatic mutations may directly cause methylation changes, is more frequent. DMRs in *cis* were equally likely to be hypo- or hypermethylated, likely reflecting creation and ablation of a transcription factor binding sites, respectively. For example, patient CRUK0082 harbors a clonal *CDKN2A* promoter mutation, where, in regions with phased methylation information, only the mutant allele is hypermethylated (**Figure 7C**).

In regions where the wild type allele is lost, it is not possible to determine whether the DMR was in-*cis* or on both alleles. We observe clonal loss of the wild type allele, combined with a missense mutation and hypomethylation of surrounding CpGs at the *CDH10* gene locus in patient CRUK0072 (**Figure 7D**).

Taken together, CAMDAC deconvolution and phasing enables deeper understanding of the interplay between aberrant DNA methylation and genetic mutations. In our lung cancer cohort, we observed frequent phasing of the hypermethylated allele to the mutant SNV allele, potentially through ablation of adjacent transcription factor binding sites.

**Relationships between purified methylation profiles**

To gain insight into similarities and differences between samples, we clustered methylation profiles. We focused on tumor-normal promoter DMRs, selected for being enriched in gene

regulatory driven differential methylation, and performed uniform manifold approximation and projection (UMAP) of deconvolved tumor and normal methylation rate profiles (**Figures 8A,B and S9A,B**). Hierarchical clustering analysis identified four clear clusters: two clusters of cancer samples and two clusters of adjacent normal samples. Normal lung epithelium samples clustered by sex (**Figure 8A**), indicating that differences are dominated by female-to-male differential methylation. These sex-based differences were less dominant in the tumor samples, in part because many female samples showed complete (34%) or at least partial (52%) LOH of chromosome X. Tumor samples in contrast mostly separated by histology (**Figure 8B**), pointing towards large methylation differences between lung adenocarcinoma and squamous cell carcinoma, possibly reflecting a different cell of origin. Indeed, squamous cell carcinoma is often found in the trachea and is thought to arise from aberrant basal cells, whereas adenocarcinoma often derives from the bronchioalveolar duct junction, likely originating from the alveolar type I/II and Clara progenitor cells (Sutherland and Berns, 2010). Interestingly, different tumor samples from the same patient consistently clustered together (**Figure 8C,D**), reflecting their close relationships.

**Purified methylation profiles reveal phylogenetic relationships**

We next focused more closely on the relationship between methylation profiles of different samples from the same patient. For each patient, we select all CpG sites that are differentially methylated compared to the normal in at least one sample. At these sites, we expect the tumor methylation rate to be anti-correlated with the normal. However, due to normal cell contamination, the bulk methylation rate at DMPs across tumor regions is dependent on tumor purity and as such is highly correlated with the matched normal methylation rate in lower purity samples (**Figures 8E and S9C**). After deconvolution, tumor signals are no longer correlated with adjacent normal methylation rates, and tumor purity is no longer the main driver of inter-sample correlations (**Figures 8E and S9C**).

Different subclones in the same tumor are phylogenetically related, and these relationships can be inferred through exome or genome sequencing (Dentro et al., 2017; Nik-Zainal et

al., 2012; Tarabichi et al., *in press*). We reasoned that methylation profiles should similarly encode evolutionary relationships between samples. To illustrate this, we focus on patient CRUK0062, for which 7 different tumor regions are sampled with varying purity. Hierarchical clustering of deconvolved tumor methylation rates could accurately replicate phylogenetic relationships between samples, as inferred from exome sequencing (**Figure 8F,G**). In contrast, methylation profiles before deconvolution were unable to reproduce phylogenetic relationships, and clustering was mainly driven by tumor purity. Similar results were obtained for other patients (**Figure S9D,E**). We therefore conclude that CAMDAC provides a unique window of opportunity to study intra-tumor heterogeneity in solid tumors, unconfounded by signals from admixed normal cells.

## DISCUSSION

Bulk tumor methylation signals are confounded by copy number aberrations and normal cell admixture. CAMDAC determines tumor purity and allele-specific copy number from bulk tumor-normal matched RRBS data and uses this to reconstruct the true tumor methylation rate from the bulk tumor and matched normal rates. Use of these purified methylomes increases the accuracy of tumor-normal and tumor-tumor differential methylation calling and can reveal the phylogenetic relationships between samples. In addition, they provide insight into tumor biology and the taxonomy of cancer.

While a few studies have considered either the effect of sample purity (Hua et al., 2020; Zheng et al., 2017) or copy number (Martin-Trujillo et al., 2017) on bulk tumor methylation rates, CAMDAC is – to our knowledge – the first approach to estimate and leverage both to accurately reconstruct the tumor methylation signal. We demonstrate that tumor-normal differential methylation produces predictable intermediate methylation values given local copy number states. This explains, at least in part, why bulk cancer methylomes exhibit an increased proportion of intermediately methylated regions compared to normal tissues (Lister et al., 2009).

After purification, most differentially methylated CpGs are either fully methylated or unmethylated. Nevertheless, a significant subset remains intermediately methylated. Assuming contributions from tumor subclones, imprinting, and allele-specific regulatory variation in the cell of origin are limited, we posit this signal mainly reflects from copy- and allele-specific methylation derived from stochastic epimutation events which have subsequently been faithfully replicated.

Hierarchical clustering of CAMDAC methylation profiles recapitulates the phylogenetic relationships between samples, exemplifying how CAMDAC can unlock a wealth of information on intra-tumor heterogeneity. We envisage further development will enable reconstruction of subclones and accurate clone trees (as opposed to sample trees, Alves et al., 2017) from bisulfite sequencing data, akin to current methods leveraging single-nucleotide variants (Dentro et al., 2017; Tarabichi et al., *in press*).

Application of CAMDAC will help elucidate the role of epimutations in the evolutionary histories of different cancer types. As in previous studies (Gaiti et al., 2019; Hua et al., 2020), we infer differential methylation compared to adjacent normal tissue and use this as a proxy for the admixed normal cells in the tumor sample. Where matched normal tissue samples are not available, a tissue-matched normal reference may be constructed for use with CAMDAC, as demonstrated for CRUK0047 (**Methods**). Outside our cohort, the DNA methylation data flowing from large-scale initiatives such as the International Human Epigenome Consortium (IHEC, Stunnenberg et al., 2016) and the ENCODE (Feingold et al., 2004) and BLUEPRINT (Fernández et al., 2016) projects, enable construction of reference profiles for a wide range of tissues.

In low purity samples and in regions with lower tumor copy number, fewer reads report on the methylation state of the tumor. CAMDAC incorporates this greater uncertainty in its estimates for the tumor methylation rates in such cases, greatly reducing the number of false positive hits when inferring differential methylation. Bulk tumor methylomes of low purity samples are dominated by signals from the admixed normal cells, as evidenced by the purity-driven clustering of bulk profiles. In contrast, the CAMDAC-deconvoluted tumor methylation profiles revealed intra-tumor sample relationships.

DNA methylation heterogeneity has been linked to clonal progression in several tumor types (Hao et al., 2016; Mazor et al., 2015). Various metrics have been conceived to quantify intra-tumor methylation heterogeneity from bulk data and some have shown prognostic value in leukemia and Ewing sarcoma (Landau et al., 2014; Li et al., 2016; Sheffield et al., 2017), but not in *e.g.* glioblastoma (Klughammer et al., 2018). While they are likely accurate in tumor types for which high purity samples are readily obtained, we hypothesize these metrics are similarly confounded by normal cell admixture and copy number aberrations in most other cases. Applying the principles of CAMDAC to adjust these metrics accordingly may yield powerful prognosticators and shed further light on the role of DNA methylation heterogeneity in solid tumors.

In summary, CAMDAC enables us to exploit the wealth of information contained in bisulfite sequencing data: combining purity and copy number profiling with accurate differential methylation analysis based on the purified tumor methylation rate estimates. CAMDAC uniquely unlocks insights into intra-tumor heterogeneity and the biology and taxonomy of cancer, from bulk bisulfite sequencing of solid tumors. We expect CAMDAC will further our understanding of the interplay between genetic and epigenetic mutations and their roles during tumor evolution.

## REFERENCES


Alves, J.M., Prieto, T., and Posada, D. (2017). Multiregional Tumor Trees Are Not Phylogenies. Trends in Cancer.

Auton, A., Abecasis, G.R., Altshuler, D.M., Durbin, R.M., Bentley, D.R., Chakravarti, A., Clark, A.G., Donnelly, P., Eichler, E.E., Flicek, P., et al. (2015). A global reference for human genetic variation. Nature *526*, 68–74.

Barrett, J.E., Feber, A., Herrero, J., Tanic, M., Wilson, G.A., Swanton, C., and Beck, S. (2017). Quantification of tumour evolution and heterogeneity via Bayesian epiallele detection. BMC Bioinformatics *18*, 354.

Benjamini, Y., and Speed, T.P. (2012). Summarizing and correcting the GC content bias in high-throughput sequencing. Nucleic Acids Res. *40*, e72–e72.

Bird, A. (2002). DNA methylation patterns and epigenetic memory. Genes Dev. *16*, 6–21.

Brocks, D., Assenov, Y., Minner, S., Bogatyrova, O., Simon, R., Koop, C., Oakes, C., Zucknick, M., Lipka, D.B., Weischenfeldt, J., et al. (2014). Intratumor DNA Methylation Heterogeneity Reflects Clonal Evolution in Aggressive Prostate Cancer. Cell Rep. *8*, 798–806.

C. Dentro, S., Wedge, D.C., and Van Loo, P. (2017). Principles of Reconstructing the Subclonal Architecture of Cancers. In Cancer Evolution, C. Swanton, A. Bardelli, K. Polyak, S. Shah, and T. Graham, eds. (Cold Spring Harbour Laboratory Press), p.

Capper, D., Jones, D.T.W., Sill, M., Hovestadt, V., Schrimpf, D., Sturm, D., Koelsche, C., Sahm, F., Chavez, L., Reuss, D.E., et al. (2018). DNA methylation-based classification of central nervous system tumours. Nature *555*, 469–474.

Carter, S.L., Cibulskis, K., Helman, E., McKenna, A., Shen, H., Zack, T., Laird, P.W., Onofrio, R.C., Winckler, W., Weir, B.A., et al. (2012). Absolute quantification of somatic DNA alterations in human cancer. Nat. Biotechnol. *30*, 413–421.

Costello, J.F., Shchors, K., Fiore, C., Zhao, Y., Bilenky, M., Hirst, M., Maunakea, A.K., Ballinger, T.J., Jones, S.J.M., Nagarajan, R.P., et al. (2010). Conserved role of intragenic

DNA methylation in regulating alternative promoters. Nature *466*, 253–257.

Coulondre, C., Miller, J.H., Farabaugh, P.J., and Gilbert, W. (1978). Molecular basis of base substitution hotspots in Escherichia coli. Nature.

Deaton, A.M., and Bird, A. (2011). CGIs and the regulation of transcription. Genes Dev *25*, 1010–1022.

Deniz, Ö., Frost, J.M., and Branco, M.R. (2019). Regulation of transposable elements by DNA modifications. Nat. Rev. Genet. *20*, 417–431.

Dentro, S.C., Wedge, D.C., and Van Loo, P. (2017). Principles of Reconstructing the Subclonal Architecture of Cancers. Cold Spring Harb. Perspect. Med. *7*, a026625.

Dolzhenko, E., and Smith, A.D. (2014). Using beta-binomial regression for high-precision differential methylation analysis in multifactor whole-genome bisulfite sequencing experiments. BMC Bioinformatics *15*, 215.

Eckhardt, F., Lewin, J., Cortese, R., Rakyan, V.K., Attwood, J., Burger, M., Burton, J., Cox, T. V, Davies, R., Down, T.A., et al. (2006). DNA methylation profiling of human chromosomes 6, 20 and 22. Nat. Genet. *38*, 1378–1385.

Feber, A., Dhami, P., Dong, L., de Winter, P., Tan, W.S., Martínez-Fernández, M., Paul, D.S., Hynes-Allen, A., Rezaee, S., Gurung, P., et al. (2017). UroMark—a urinary biomarker assay for the detection of bladder cancer. Clin. Epigenetics *9*, 8.

Feinberg, A.P., Ohlsson, R., and Henikoff, S. (2006). The epigenetic progenitor origin of human cancer. Nat. Rev. Genet. *7*, 21–33.

Feinberg, A.P., Koldobskiy, M.A., and Göndör, A. (2016). Epigenetic modulators, modifiers and mediators in cancer aetiology and progression. Nat. Rev. Genet. *17*, 284–299.

Feingold, E.A., Good, P.J., Guyer, M.S., Kamholz, S., Liefer, L., Wetterstrand, K., Collins, F.S., Gingeras, T.R., Kampa, D., Sekinger, E.A., et al. (2004). The ENCODE (ENCyclopedia of DNA Elements) Project. Science (80-. ).

Fernández, J.M., de la Torre, V., Richardson, D., Royo, R., Puiggròs, M., Moncunill, V.,

Fragkogianni, S., Clarke, L., Flicek, P., Rico, D., et al. (2016). The BLUEPRINT Data Analysis Portal. Cell Syst.

Flavahan, W.A., Gaskell, E., and Bernstein, B.E. (2017). Epigenetic plasticity and the hallmarks of cancer. Science (80-. ). *357*, eaal2380.

Frevel, M.A.E., Sowerby, S.J., Petersen, G.B., and Reeve, A.E. (1999). Methylation sequencing analysis refines the region of H19 epimutation in Wilms tumor. J. Biol. Chem.

Gaiti, F., Chaligne, R., Gu, H., Brand, R.M., Kothen-Hill, S., Schulman, R.C., Grigorev, K., Risso, D., Kim, K.-T., Pastore, A., et al. (2019). Epigenetic evolution and lineage histories of chronic lymphocytic leukaemia. Nature *569*, 576–580.

Goyal, R. (2006). Accuracy of DNA methylation pattern preservation by the Dnmt1 methyltransferase. Nucleic Acids Res. *34*, 1182–1188.

Greger, V., Passarge, E., Höpping, W., Messmer, E., and Horsthemke, B. (1989). Epigenetic changes may contribute to the formation and spontaneous regression of retinoblastoma. Hum. Genet. *83*, 155–158.

Guo, S., Diep, D., Plongthongkum, N., Fung, H.-L., Zhang, K., and Zhang, K. (2017). Identification of methylation haplotype blocks aids in deconvolution of heterogeneous tissue samples and tumor tissue-of-origin mapping from plasma DNA. Nat. Genet. *49*, 635–642.

Hanahan, D., and Weinberg, R.A. (2011). Hallmarks of Cancer: The Next Generation. Cell *144*, 646–674.

Hannum, G., Guinney, J., Zhao, L., Zhang, L., Hughes, G., Sadda, S., Klotzle, B., Bibikova, M., Fan, J.-B., Gao, Y., et al. (2013). Genome-wide Methylation Profiles Reveal Quantitative Views of Human Aging Rates. Mol. Cell *49*, 359–367.

Hansen, K.D., Timp, W., Bravo, H.C., Sabunciyan, S., Langmead, B., McDonald, O.G., Wen, B., Wu, H., Liu, Y., Diep, D., et al. (2011). Increased methylation variation in epigenetic domains across cancer types. Nat. Genet. *43*, 768–775.

Hansen, K.D., Langmead, B., and Irizarry, R.A. (2012). BSmooth: from whole genome

bisulfite sequencing reads to differentially methylated regions. Genome Biol. *13*, R83.

Hao, J.-J., Lin, D.-C., Dinh, H.Q., Mayakonda, A., Jiang, Y.-Y., Chang, C., Jiang, Y., Lu, C.-C., Shi, Z.-Z., Xu, X., et al. (2016). Spatial intratumoral heterogeneity and temporal clonal evolution in esophageal squamous cell carcinoma. Nat. Genet. *48*, 1500–1507.

Heyn, H., and Esteller, M. (2012). DNA methylation profiling in the clinic: applications and challenges. Nat. Rev. Genet. *13*, 679–692.

Horvath, S. (2013). DNA methylation age of human tissues and cell types. Genome Biol. *14*, R115.

Hua, X., Zhao, W., Pesatori, A.C., Consonni, D., Caporaso, N.E., Zhang, T., Zhu, B., Wang, M., Jones, K., Hicks, B., et al. (2020). Genetic and epigenetic intratumor heterogeneity impacts prognosis of lung adenocarcinoma. Nat. Commun. *11*, 2459.

Hutter, C., and Zenklusen, J.C. (2018). The Cancer Genome Atlas: Creating Lasting Value beyond Its Data. Cell *173*, 283–285.

ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium, ., Campbell, P.J., Getz, G., Korbel, J.O., Stuart, J.M., Jennings, J.L., Stein, L.D., Perry, M.D., Nahal-Bose, H.K., Ouellette, B.F.F., et al. (2020). Pan-cancer analysis of whole genomes. Nature.

Jamal-Hanjani, M., Wilson, G.A., McGranahan, N., Birkbak, N.J., Watkins, T.B.K., Veeriah, S., Shafi, S., Johnson, D.H., Mitter, R., Rosenthal, R., et al. (2017). Tracking the Evolution of Non–Small-Cell Lung Cancer. N. Engl. J. Med. *376*, 2109–2121.

Kerkel, K., Spadola, A., Yuan, E., Kosek, J., Jiang, L., Hod, E., Li, K., Murty, V. V., Schupf, N., Vilain, E., et al. (2008). Genomic surveys by methylation-sensitive SNP analysis identify sequence-dependent allele-specific DNA methylation. Nat. Genet. *40*, 904–908.

Klein, H.U., and Hebestreit, K. (2016). An evaluation of methods to test predefined genomic regions for differential methylation in bisulfite sequencing data. Brief. Bioinform. *17*, 796–807.

Klughammer, J., Kiesel, B., Roetzer, T., Fortelny, N., Nemc, A., Nenning, K.-H., Furtner,

J., Sheffield, N.C., Datlinger, P., Peter, N., et al. (2018). The DNA methylation landscape of glioblastoma disease progression shows extensive heterogeneity in time and space. Nat. Med. *24*, 1611–1624.

Koch, A., Joosten, S.C., Feng, Z., de Ruijter, T.C., Draht, M.X., Melotte, V., Smits, K.M., Veeck, J., Herman, J.G., Van Neste, L., et al. (2018). Analysis of DNA methylation in cancer: location revisited. Nat. Rev. Clin. Oncol. *15*, 459–466.

Kohli, R.M., and Zhang, Y. (2013). TET enzymes, TDG and the dynamics of DNA demethylation. Nature *502*, 472–479.

Krueger, F., and Andrews, S.R. (2011). Bismark: A flexible aligner and methylation caller for Bisulfite-Seq applications. Bioinformatics *27*, 1571–1572.

Kulis, M., Heath, S., Bibikova, M., Queirós, A.C., Navarro, A., Clot, G., Martínez-Trillos, A., Castellano, G., Brun-Heath, I., Pinyol, M., et al. (2012). Epigenomic analysis detects widespread gene-body DNA hypomethylation in chronic lymphocytic leukemia. Nat. Genet. *44*, 1236–1242.

Landan, G., Cohen, N.M., Mukamel, Z., Bar, A., Molchadsky, A., Brosh, R., Horn-Saban, S., Zalcenstein, D.A., Goldfinger, N., Zundelevich, A., et al. (2012). Epigenetic polymorphism and the stochastic formation of differentially methylated regions in normal and cancerous tissues. Nat. Genet.

Landau, D.A., Clement, K., Ziller, M.J., Boyle, P., Fan, J., Gu, H., Stevenson, K., Sougnez, C., Wang, L., Li, S., et al. (2014). Locally Disordered Methylation Forms the Basis of Intratumor Methylome Variation in Chronic Lymphocytic Leukemia. Cancer Cell *26*, 813–825.

Li, S., Garrett-Bakelman, F.E., Chung, S.S., Sanders, M.A., Hricik, T., Rapaport, F., Patel, J., Dillon, R., Vijay, P., Brown, A.L., et al. (2016). Distinct evolution and dynamics of epigenetic and genetic heterogeneity in acute myeloid leukemia. Nat. Med. *22*, 792–799.

Lister, R., Pelizzola, M., Dowen, R.H., Hawkins, R.D., Hon, G., Tonti-Filippini, J., Nery, J.R., Lee, L., Ye, Z., Ngo, Q.-M., et al. (2009). Human DNA methylomes at base resolution show widespread epigenomic differences. Nature *462*, 315–322.

Van Loo, P., Nordgard, S.H., Lingjaerde, O.C., Russnes, H.G., Rye, I.H., Sun, W., Weigman, V.J., Marynen, P., Zetterberg, A., Naume, B., et al. (2010). Allele-specific copy number analysis of tumors. Proc. Natl. Acad. Sci. *107*, 16910–16915.

Lyon, M.F. (1962). Sex chromatin and gene action in the mammalian X-chromosome. Am. J. Hum. Genet. *14*, 135–148.

Martin-Trujillo, A., Vidal, E., Monteagudo-Sánchez, A., Sanchez-Delgado, M., Moran, S., Hernandez Mora, J.R., Heyn, H., Guitart, M., Esteller, M., and Monk, D. (2017). Copy number rather than epigenetic alterations are the major dictator of imprinted methylation in tumors. Nat. Commun. *8*, 467.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.Journal.

Mazor, T., Pankov, A., Johnson, B.E., Hong, C., Hamilton, E.G., Bell, R.J.A., Smirnov, I. V, Reis, G.F., Phillips, J.J., Barnes, M.J., et al. (2015). DNA Methylation and Somatic Mutations Converge on the Cell Cycle and Define Similar Evolutionary Histories in Brain Tumors. Cancer Cell *28*, 307–317.

Meissner, A. (2005). Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. Nucleic Acids Res. *33*, 5868–5877.

Nik-Zainal, S., Van Loo, P., Wedge, D.C., Alexandrov, L.B., Greenman, C.D., Lau, K.W., Raine, K., Jones, D., Marshall, J., Ramakrishna, M., et al. (2012). The life history of 21 breast cancers. Cell *149*, 994–1007.

Nordlund, J., Bäcklin, C.L., Wahlberg, P., Busche, S., Berglund, E.C., Eloranta, M.-L., Flaegstad, T., Forestier, E., Frost, B.-M., Harila-Saari, A., et al. (2013). Genome-wide signatures of differential DNA methylation in pediatric acute lymphoblastic leukemia. Genome Biol. *14*, r105.

Oakes, C.C., Seifert, M., Assenov, Y., Gu, L., Przekopowitz, M., Ruppert, A.S., Wang, Q., Imbusch, C.D., Serva, A., Koser, S.D., et al. (2016). DNA methylation dynamics during B cell maturation underlie a continuum of disease phenotypes in chronic lymphocytic leukemia. Nat. Genet. *48*, 253–264.

Pfister, S.M., Eils, J., Taylor, M.D., Yaspo, M.-L., Pietsch, T., Bunt, J., Stachurski, K., Korshunov, A., Warnatz, H.-J., Koster, J., et al. (2014). Decoding the regulatory landscape of medulloblastoma using DNA methylation sequencing. Nature *510*, 537–541.

Quail, D.F., and Joyce, J.A. (2013). Microenvironmental regulation of tumor progression and metastasis. Nat. Med. *19*, 1423–1437.

Raineri, E., Dabad, M., and Heath, S. (2014). A Note on Exact Differences between Beta Distributions in Genomic (Methylation) Studies. PLoS One *9*, e97349.

Reik, W., and Walter, J. (2001). Genomic imprinting: parental influence on the genome. Nat. Rev. Genet. *2*, 21–32.

Robinson, D. (2017). Introduction to Empirical Bayes: Examples from Baseball Statistics.

Robinson, M.D., Kahraman, A., Law, C.W., Lindsay, H., Nowicka, M., Weber, L.M., and Zhou, X. (2014). Statistical methods for detecting differentially methylated loci and regions. Front. Genet. *5*, 1–7.

Rosenthal, R., Larose Cadieux, E., Salgado, R., Bakir, M. Al, Moore, D.A., Hiley, C.T., Lund, T., Tanić, M., Reading, J.L., Joshi, K., et al. (2019). Neoantigen-directed immune escape in lung cancer evolution. Nature *567*, 479–485.

Ryba, T., Hiratani, I., Lu, J., Itoh, M., Kulik, M., Zhang, J., Schulz, T.C., Robins, A.J., Dalton, S., and Gilbert, D.M. (2010). Evolutionarily conserved replication timing profiles predict long-range chromatin interactions and distinguish closely related cell types. Genome Res. *20*, 761–770.

Schultz, M.D., He, Y., Whitaker, J.W., Hariharan, M., Mukamel, E.A., Leung, D., Rajagopal, N., Nery, J.R., Urich, M.A., Chen, H., et al. (2015). Human body epigenome maps reveal noncanonical DNA methylation variation. Nature *523*, 212–216.

Sheffield, N.C., Pierron, G., Klughammer, J., Datlinger, P., Schönegger, A., Schuster, M., Hadler, J., Surdez, D., Guillemot, D., Lapouble, E., et al. (2017). DNA methylation heterogeneity defines a disease spectrum in Ewing sarcoma. Nat. Med. *23*, 386–395.

Shen, H., and Laird, P.W. (2013). Interplay between the Cancer Genome and Epigenome.

Cell *153*, 38–55.

Smith, Z.D., and Meissner, A. (2013). DNA methylation: Roles in mammalian development. Nat. Rev. Genet. *14*, 204–220.

Stunnenberg, H.G., Abrignani, S., Adams, D., de Almeida, M., Altucci, L., Amin, V., Amit, I., Antonarakis, S.E., Aparicio, S., Arima, T., et al. (2016). The International Human Epigenome Consortium: A Blueprint for Scientific Collaboration and Discovery. Cell.

Sun, Z., Cunningham, J., Slager, S., and Kocher, J.P. (2015). Base resolution methylome profiling: Considerations in platform selection, data preprocessing and analysis. Epigenomics *7*, 813–828.

Sutherland, K.D., and Berns, A. (2010). Cell of origin of lung cancer. Mol. Oncol. *4*, 397–403.

Tarabichi, M., Salcedo, A., Deshwar, A.G., Ni Leathlobhair, M., Wintersinger, J., Wedge, D.C., Van Loo, P., Morris, Q.D., and Boutros, P.C. A practical guide to cancer subclonal reconstruction from DNA sequencing. Nat Methods in press.

Teschendorff, A.E., Breeze, C.E., Zheng, S.C., and Beck, S. (2017). A comparison of reference-based algorithms for correcting cell-type heterogeneity in Epigenome-Wide Association Studies. BMC Bioinformatics *18*, 105.

Wen, B., Montano, C., Wu, Z., Gabo, K., Potash, J.B., Ji, H., Sabunciyan, S., Rongione, M., Webster, M., Irizarry, R.A., et al. (2009). The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. Nat. Genet. *41*, 178–186.

Wu, H., Xu, T., Feng, H., Chen, L., Li, B., Yao, B., Qin, Z., Jin, P., and Conneely, K.N. (2015). Detection of differentially methylated regions from whole-genome bisulfite sequencing data without replicates. Nucleic Acids Res. *43*, 141.

Zheng, X., Zhao, Q., Wu, H.-J., Li, W., Wang, H., Meyer, C.A., Qin, Q.A., Xu, H., Zang, C., Jiang, P., et al. (2014). MethylPurify: tumor purity deconvolution and differential methylation detection from single tumor DNA methylomes. Genome Biol. *15*, 419.

Zheng, X., Zhang, N., Wu, H.-J., and Wu, H. (2017). Estimating and accounting for tumor purity in the analysis of DNA methylation data from cancer studies. Genome Biol. *18*, 17.

Zhou, D., Li, Z., Yu, D., Wan, L., Zhu, Y., Lai, M., and Zhang, D. (2015). Polymorphisms involving gain or loss of CpG sites are significantly enriched in trait-associated SNPs. Oncotarget *6*, 39995–40004.

Ziller, M.J., Gu, H., Müller, F., Donaghey, J., Tsai, L.T.-Y., Kohlbacher, O., De Jager, P.L., Rosen, E.D., Bennett, D.A., Bernstein, B.E., et al. (2013). Charting a dynamic DNA methylation landscape of the human genome. Nature *500*, 477–481.

## ACKNOWLEDGEMENTS

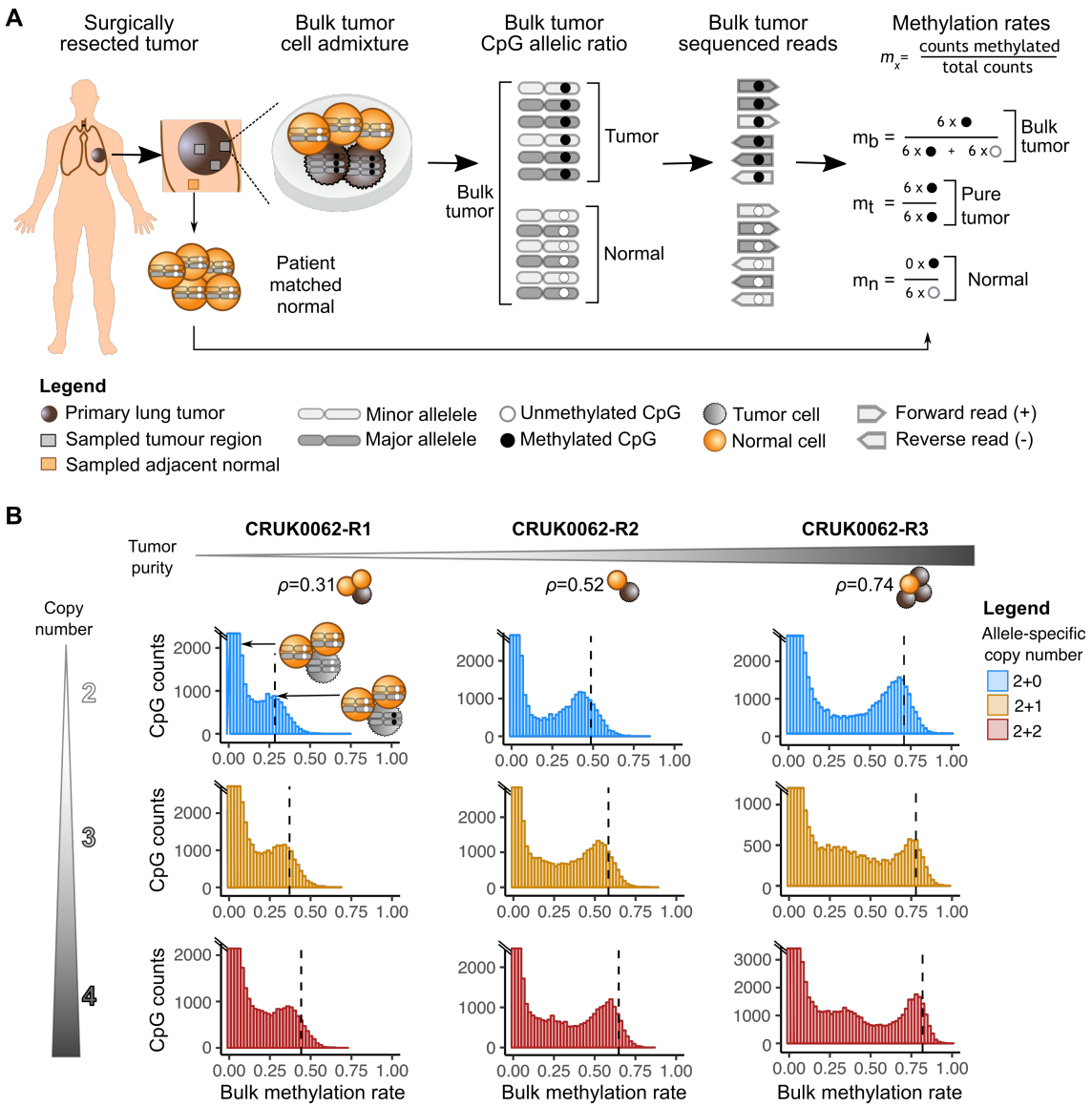
This work was supported by the Francis Crick Institute, which receives its core funding from Cancer Research UK (FC001202), the UK Medical Research Council (FC001202), and the Wellcome Trust (FC001202). We acknowledge technical support from the CRUK–UCL Centre-funded Genomics and Genome Engineering Core Facility of the UCL Cancer Institute and grant support from the NIHR BRC (BRC275/CN/SB/101330) and the Wellcome Trust (218274/Z/19/Z). JD is a postdoctoral fellow supported by the European Union's Horizon 2020 research and innovation program (Marie Skłodowska-Curie Grant Agreement No. 703594-DECODE). JD is a postdoctoral fellow of the FWO. M.T. was supported by the People Programme Marie Curie Actions (FP7/2007-2013/WHRI-ACADEMY-608765) and the Danish Council for Strategic Research (1309-00006B). C.S. is Royal Society Napier Research Professor (RP150154). His work is supported by the Francis Crick Institute, which receives its core funding from Cancer Research UK (FC001169), the UK Medical Research Council (FC001169), and the Wellcome Trust (FC001169). C.S. is funded by Cancer Research UK (TRACERx, PEACE and CRUK Cancer Immunotherapy Catalyst Network), Cancer Research UK Lung Cancer Centre of Excellence, the Rosetrees Trust, Butterfield and Stoneygate Trusts, NovoNordisk Foundation (ID16584), a Royal Society Research Professorship Enhancement Award (RP/EA/180007), the NIHR BRC at University College London Hospitals, the CRUK-UCL Centre, Experimental Cancer Medicine Centre and the Breast Cancer Research Foundation, USA (BCRF). His research is supported by a Stand Up To Cancer-LUNGevity-American Lung Association Lung Cancer Interception Dream Team Translational Research Grant (SU2C-AACR-DT23-17). C.S. also receives funding from the European Research Council (FP7-THESEUS-617844, FP7-PloidyNet 607722, PROTEUS 835297 and Chromavision 665233). PVL is a Winton Group Leader in recognition of the Winton Charitable Foundation's support towards the establishment of The Francis Crick Institute.

## DECLARATION OF INTERESTS

G.A.W. has consulted for and has stock options in Achilles Therapeutics. C.S. acknowledges grant support from Pfizer, AstraZeneca, Bristol Myers Squibb, Roche-Ventana, Boehringer-Ingelheim, Archer Dx Inc (collaboration in minimal residual disease sequencing technologies) and Ono Pharmaceutical, is an AstraZeneca Advisory Board member and Chief Investigator for the MeRmaiD1 clinical trial, has consulted for Pfizer, Novartis, GlaxoSmithKline, MSD, Bristol Myers Squibb, Celgene, AstraZeneca, Illumina, Genentech, Roche-Ventana, GRAIL, Medicxi, Bicycle Therapeutics, and the Sarah Cannon Research Institute, has stock options in Apogen Biotechnologies, Epic Bioscience, GRAIL, and has stock options and is co-founder of Achilles Therapeutics. All other authors declare no competing interests.

# FIGURES

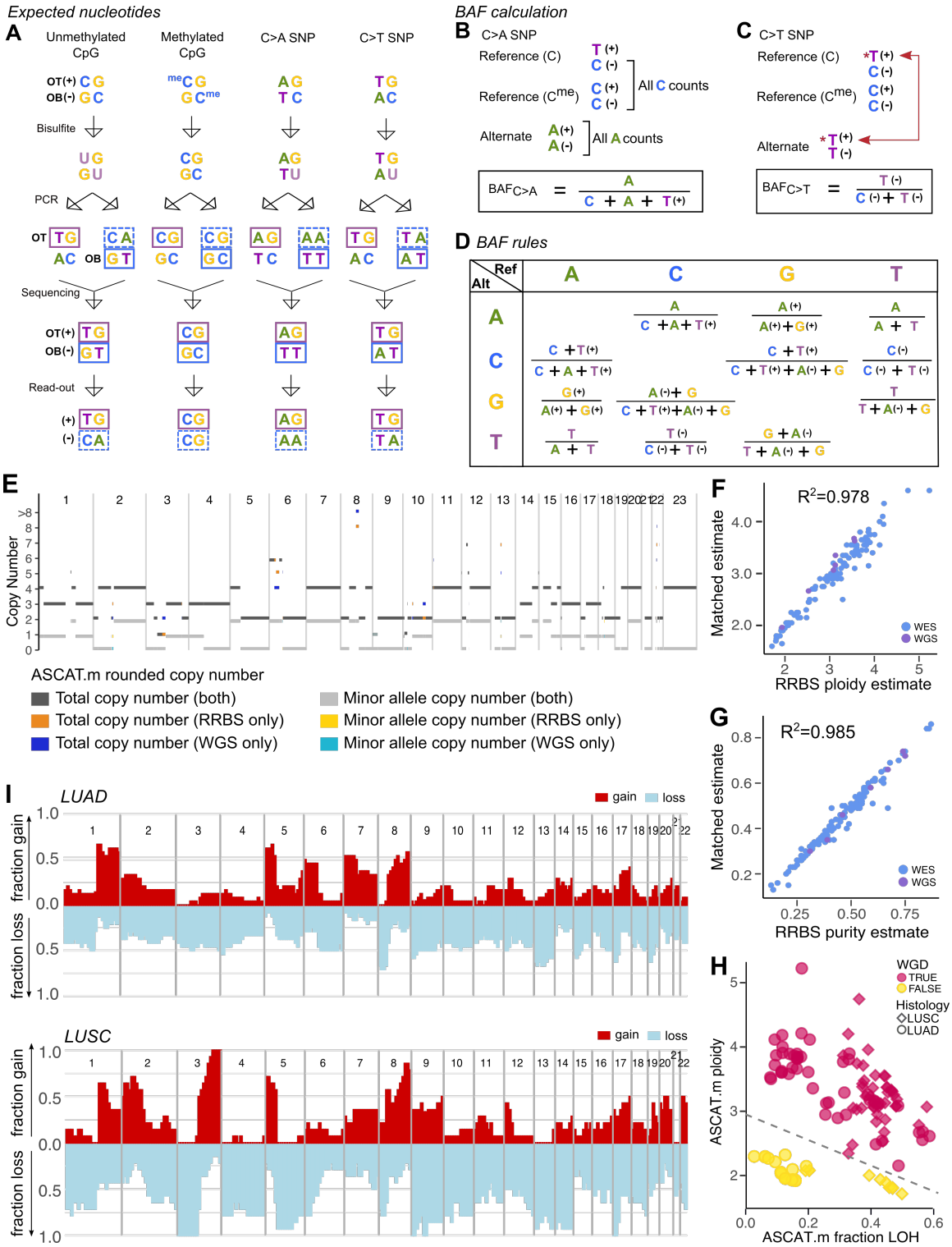
## Figure 1. Tumor purity and copy number affect methylation rates.




(A) Multi-region RRBS was performed on surgically resected non-small cell lung cancers and adjacent normal lung tissue. Sampled regions may differ both in tumor purity and copy number. The observed methylation rate of a differentially methylated CpG is expected to be influenced by purity, copy number and the methylation rate of the normal contaminating cells. (B) Bulk methylation rate histograms for tumor regions 1-3 of patient CRUK0062, for CpGs which are confidently unmethylated in the adjacent normal sample. CpGs are

stratified by copy number. A dashed line indicates the expected mode of the methylation rate peak corresponding to clonal differentially methylated CpGs on all copies ($\boldsymbol{m_t = 1}$).

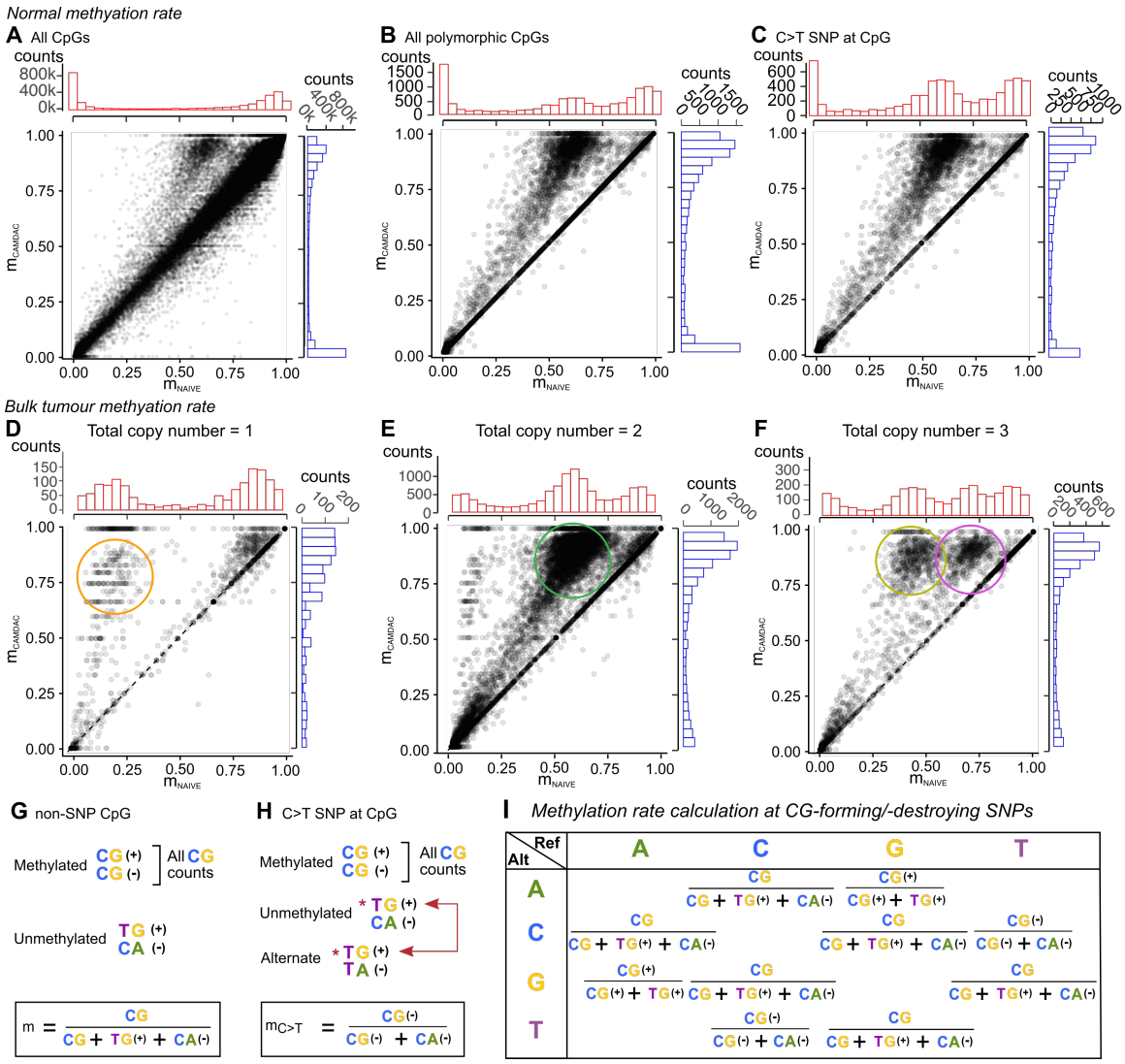
**Figure 2. Allele-Specific copy number profiling of RRBS data.**




(A) Transformation of (un)methylated reference CpG, and alternate ApG and TpG allele

dinucleotides during bisulfite sequencing. (B-C) Derivation of BAF rules from strand specific base counts C>A (B) and C>T (C) SNP loci. (D) BAF formulae for all SNP types. (E) Direct comparison of allele-specific copy number estimates derived by ASCAT.m from RRBS and WGS for sample CRUK0069-R1. (F-G) ASCAT.m RRBS-derived ploidy (F) and purity (G) estimates compared with matched exome sequencing (blue) and, where available, whole genome sequencing data (purple). (H) Tumor ploidy and the fraction of the genome with loss of heterozygosity define whole genome doubling status. (I) Frequency of gains (red) and losses (blue) across the genome in lung adenocarcinoma (LUAD) and squamous cell lung cancer (LUSC).

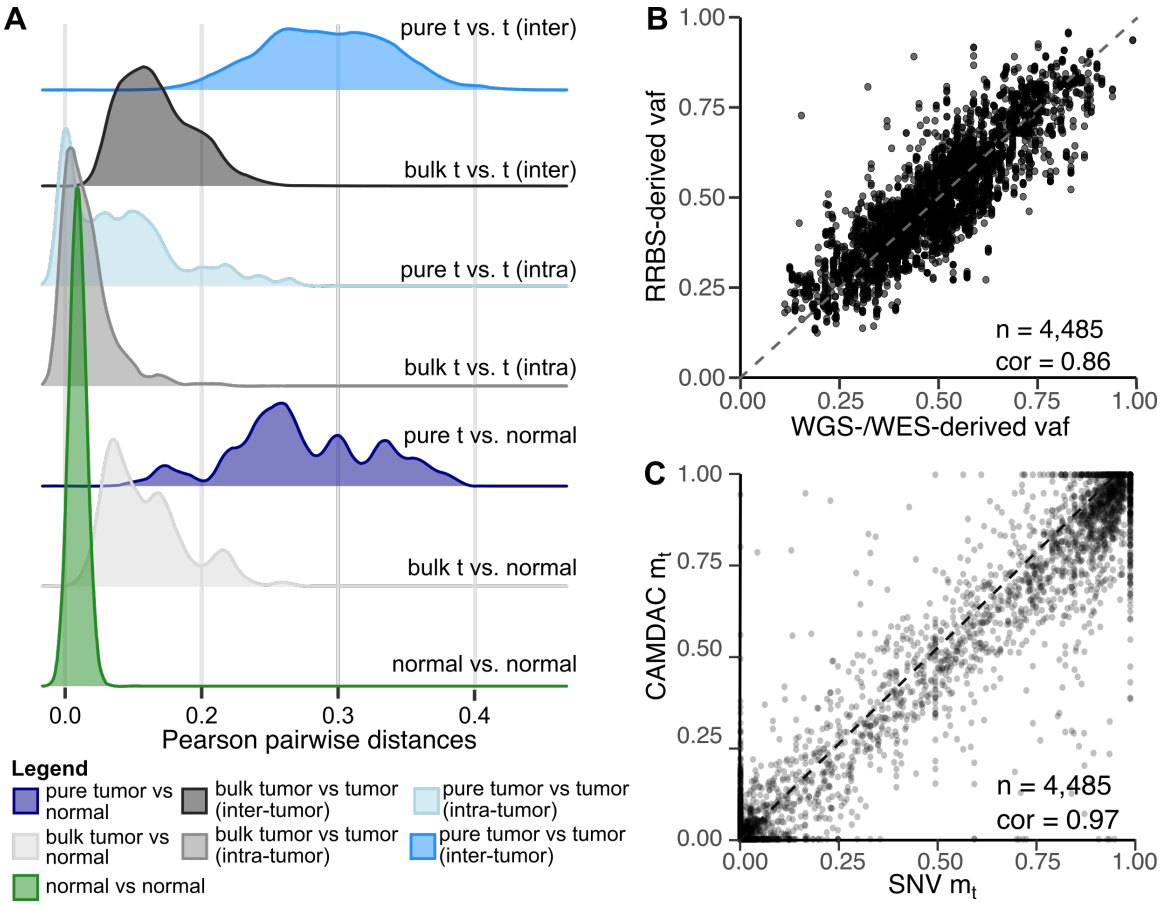
**Figure 3. Calculating polymorphism independent methylation rates with CAMDAC.**




(A-C) Naïve versus CAMDAC normal methylation rate estimates at (A) all CpGs, (B) polymorphic CpGs and (C) CpG>TpG SNPs, selected from a random sample of 3,000,000 CpGs from this cohorts 37 normal lung samples. (D-F) Naïve versus CAMDAC bulk tumor methylation rate estimates for CpG>TpG SNPs in segments with total copy number (D) 1, (E) 2 and (F) 3, pooling the data from all 3 sampled regions from patient CRUK0084 of near-equal and high tumor purity (range 0.85-0.87). The data points highlighted by the orange, green and yellow circles indicate heterozygous C>T SNPs with CpG allele copy number 1 and the pink circle CpGs with copy number 2. (G) Derivation of methylation rate

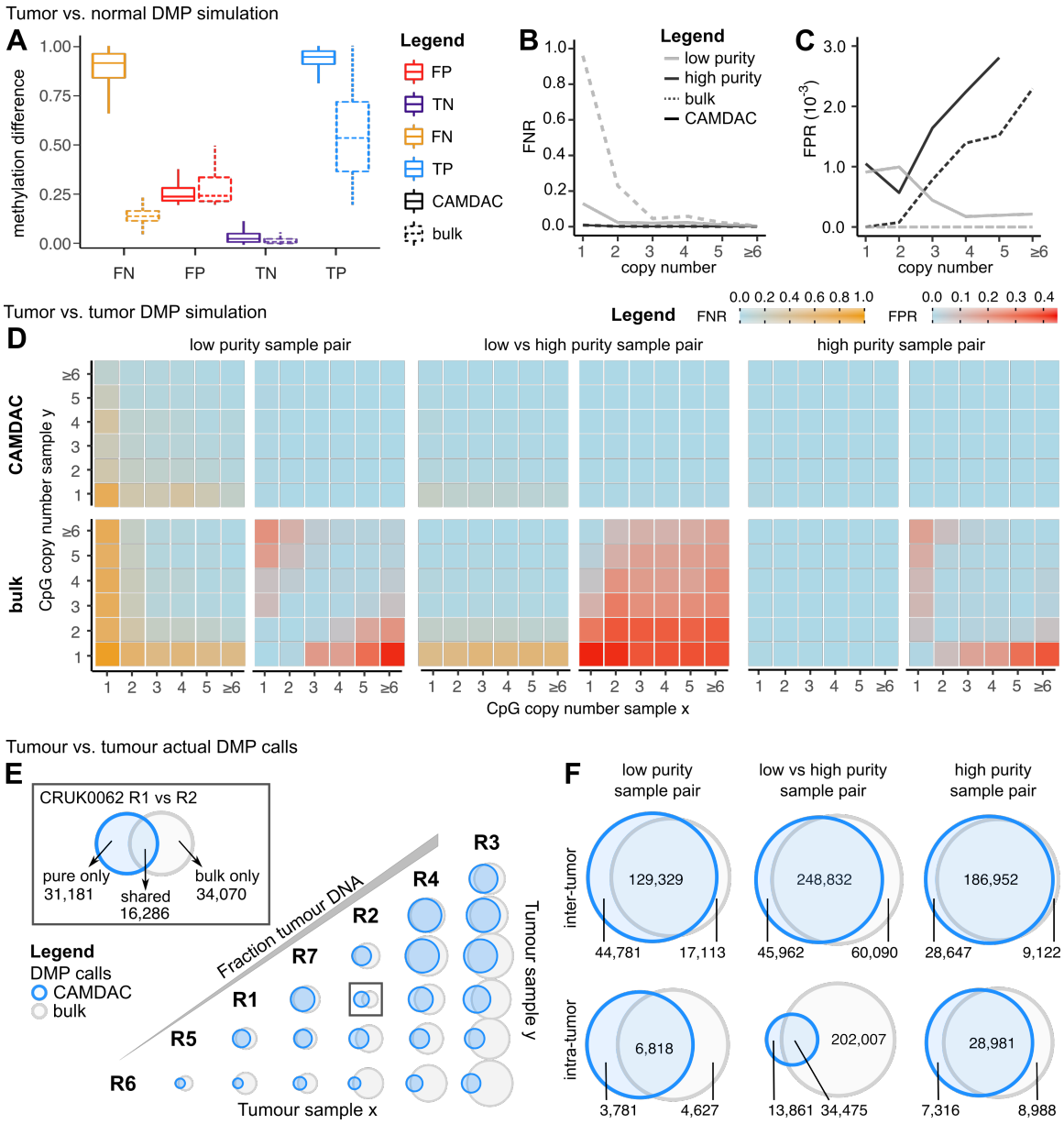
estimates at non-polymorphic CpGs. (H) Derivation of the CG-forming allele-specific methylation rate at a CpG>TpG SNP. (I) Methylation rate formulae for all possible polymorphic CpGs.

**Figure 4. CAMDAC purified methylation profiles.**




(A) Correlation between bulk tumor, CAMDAC deconvolved tumor, and adjacent normal methylation profiles. (B) Comparison of variant allele frequencies of single-nucleotide variants derived from RRBS and whole genome/exome sequencing in regions of LOH. (C) Validation of CAMDAC methylation rates through SNV-based phasing in LOH regions.

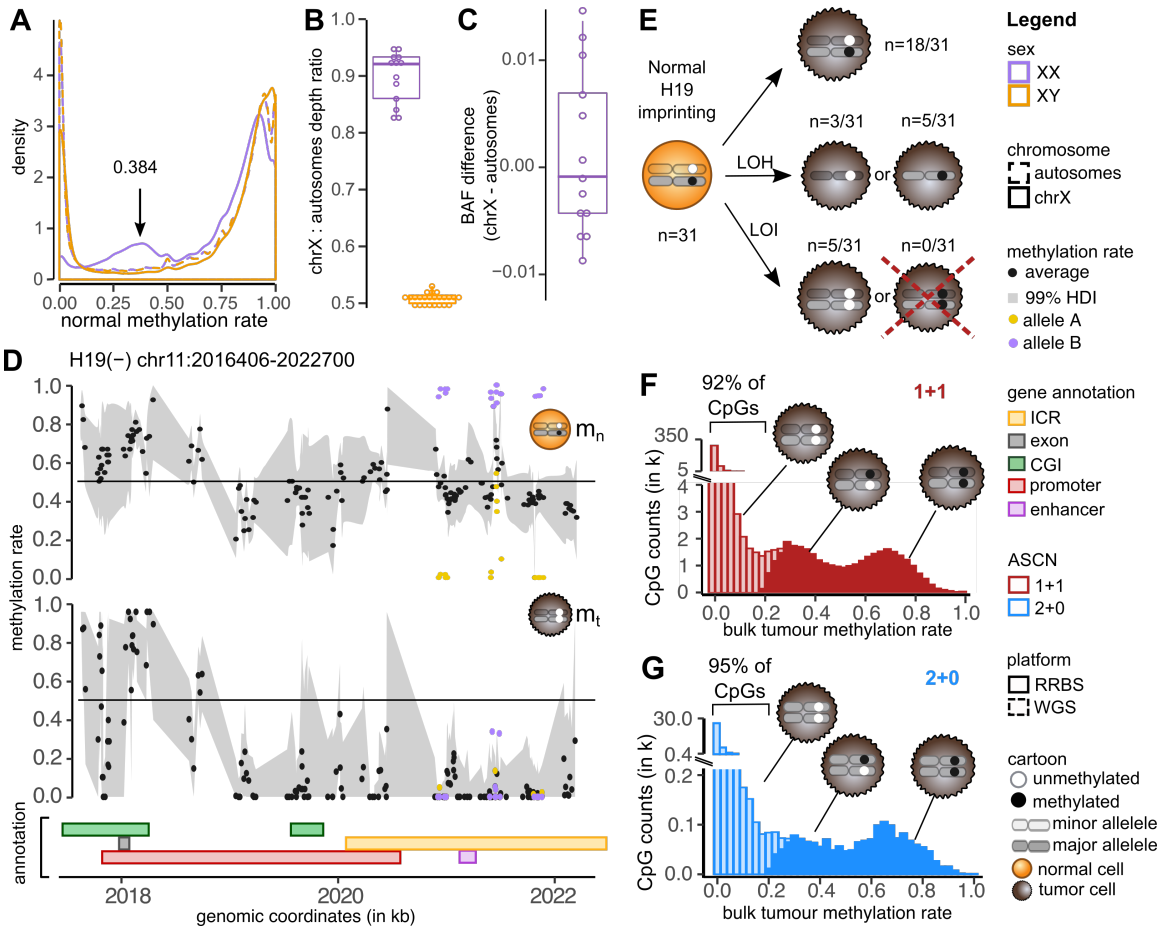
**Figure 5. Differential methylation calling using CAMDAC.**




(A-C) Tumor-normal DMP simulation results. Average methylation difference (A), false negative (B) and false positive rates (C), comparing low versus high simulated tumor purities. (D-G) Tumor-tumor DMP simulation results. (D) False negative and false positive rates as a function of the copy number state, averaged across simulated sample pairs of low (left panel), low versus high (middle panel) and high (right panel) tumor purities with CAMDAC deconvolved (top row) and bulk (bottom row) tumor methylation rates. (E)

Comparing observed tumor-tumor differential methylation calls between CRUK0062 regions from bulk (grey) and CAMDAC (blue) approaches. Samples are ordered by fraction tumor DNA content. (F) Venn diagrams showing the overlap of tumor-tumor DMPs between bulk and CAMDAC calls for intra- and inter-tumor DMPs.

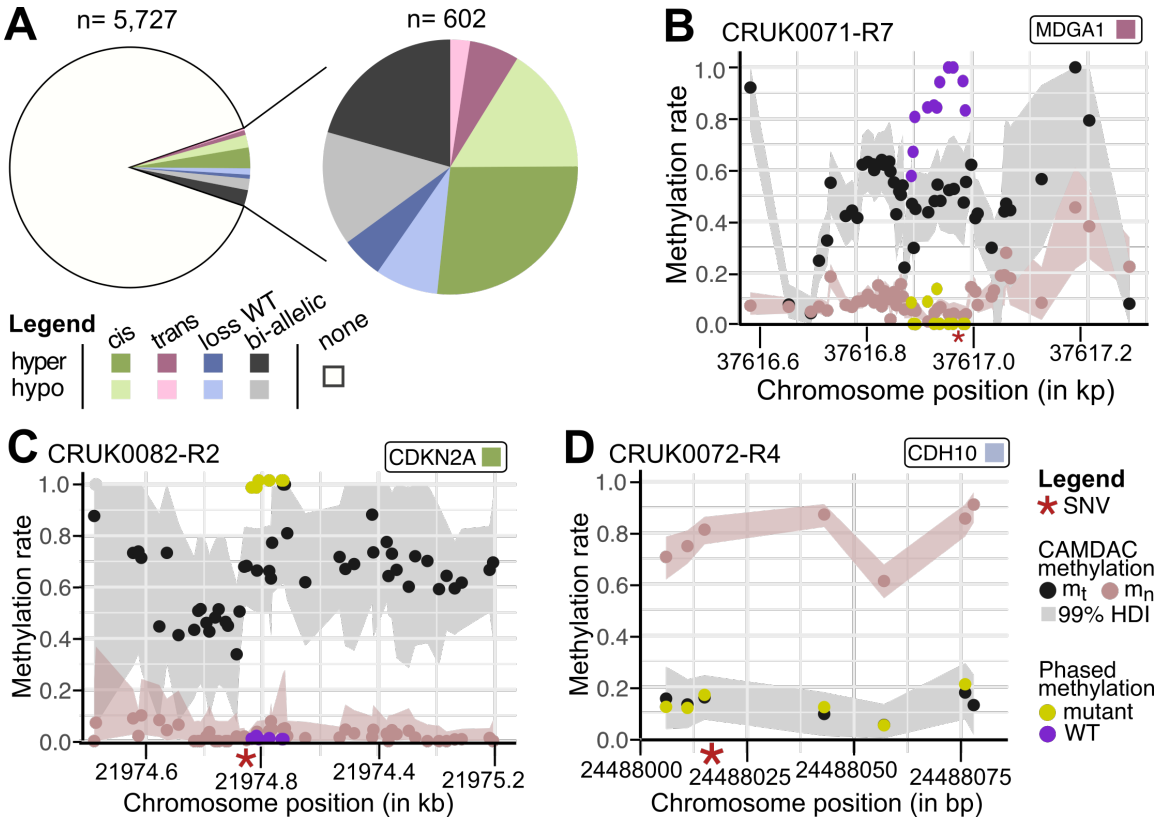
**Figure 6. Quantifying allele-specific methylation.**




(A) Allele-specific methylation signals on the X chromosome in adjacent normal lung samples across female and male patients. (B) Comparison of X chromosome sequencing depth to that on autosomes in males and females. (C) Difference between the mean BAF at heterozygous SNPs on the X chromosome and that on autosomes in females. (D) CRUK0073 normal (CRUK0073-N, top) and CAMDAC purified (CRUK0073-R2, middle) CpG methylation rate along the H19 locus with methylation rate point estimates shown as black dots, $HDI_{99}$ displayed as a grey ribbon and genomic annotations showing the transcript and promoter region as well as neighboring CpG islands, enhancers and the *IGf2/H19* imprinting control region (ICR, bottom panel). Phased methylation rates to both parental alleles are shown (purple and yellow). (E) Overview of somatic changes in the *IGf2/H19* ICR. (F-G) Deconvolved tumor methylation rate histograms for a diploid tumor sample, CRUK0062-R4, including only CpGs which are confidently unmethylated in the

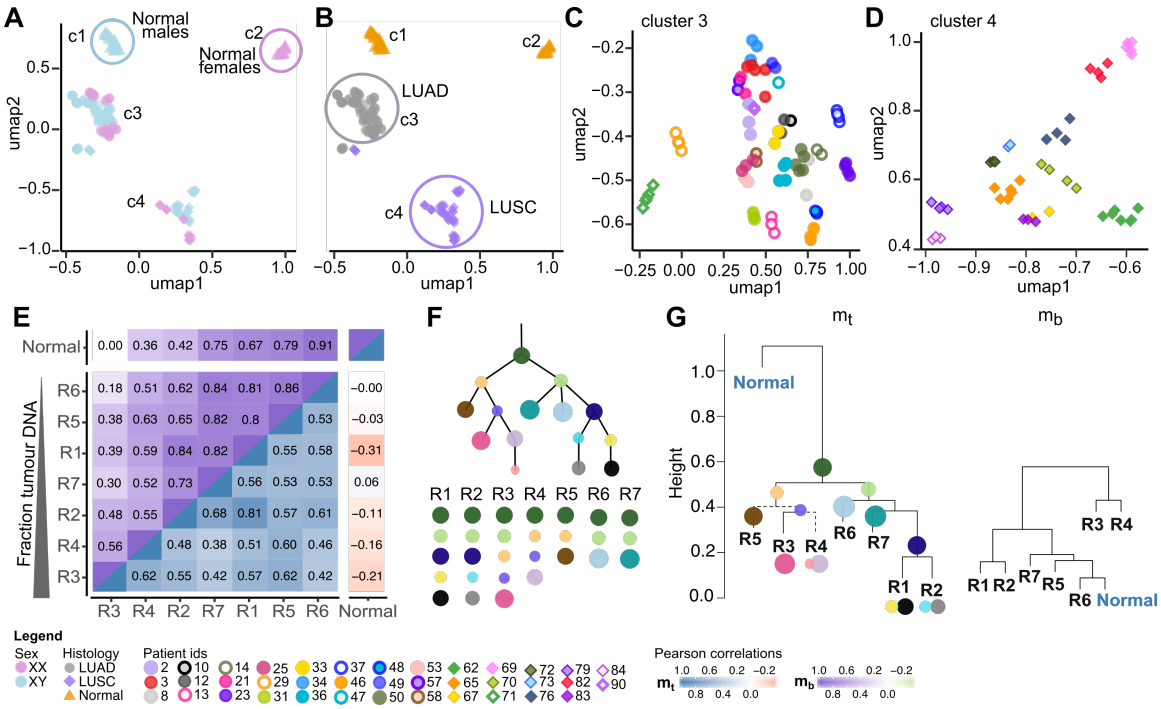
matched adjacent normal sample for 1+1 (F) and 2+0 (G) allele-specific copy number segments. Differentially methylated CpGs are shown as solid bars.

**Figure 7. The interplay of somatic mutations and methylation changes.**




(A) Overview of phasing results of 5,727 CpG bins to somatic SNVs. DMRs are annotated as either in-*cis* or in-*trans* to the SNV, on both alleles, or showing loss of the wild type (WT) allele. Each DMR category is further sub-divided into hyper- and hypomethylated subgroups. (B-D) Examples of SNV-phased, CAMDAC deconvolved tumor and matched normal methylation rate values across genes harboring SNVs and DMRs in-*trans* (B), in-*cis* (C) and with loss of the WT allele (D).

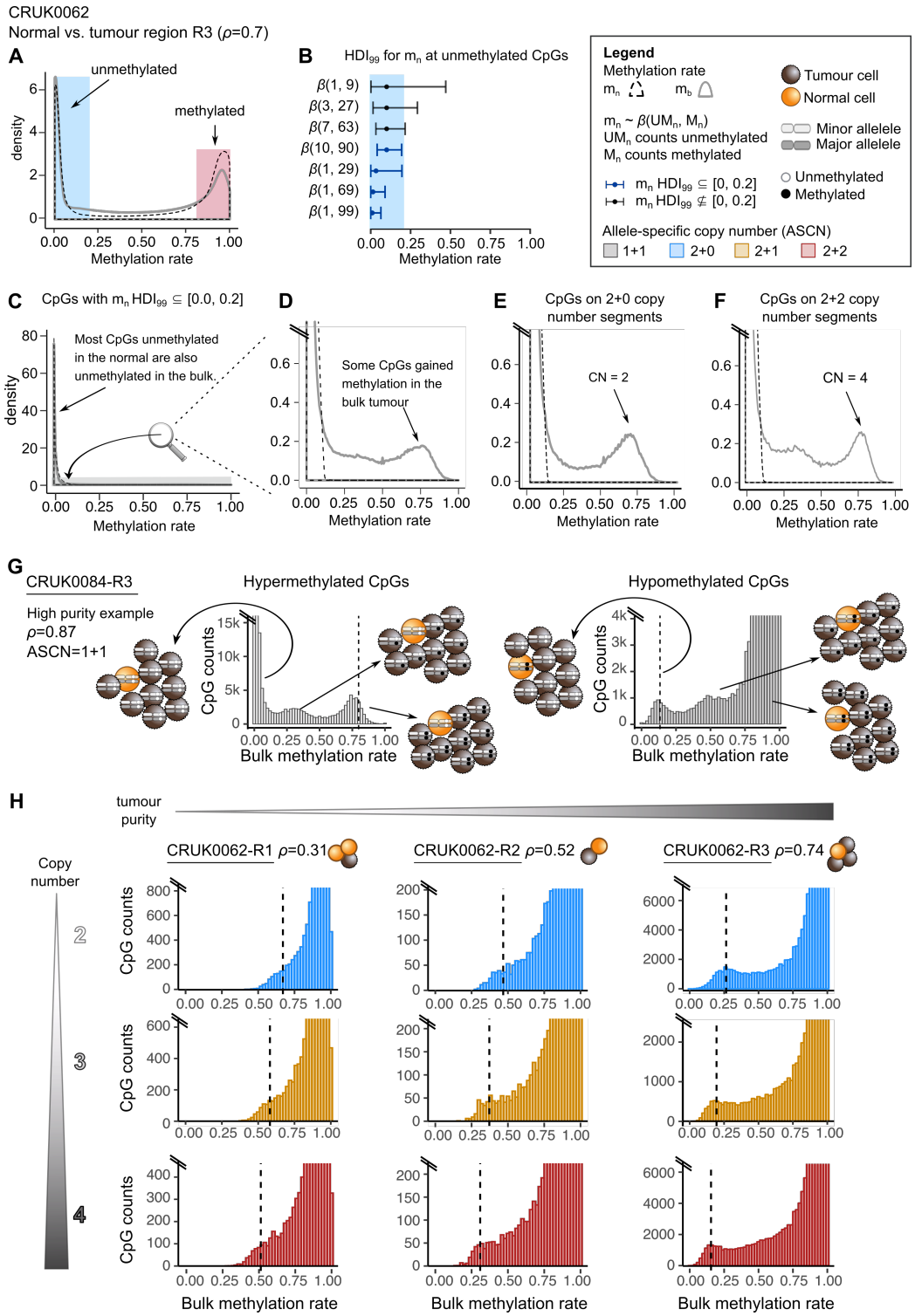
**Figure 8. CAMDAC deconvolved tumor methylation rates inform clustering and intra-tumor heterogeneity analyses.**




(A-B) UMAP analysis of the mean purified tumor methylation rates at CpGs in tumor-normal promoter DMRs, (A-B) annotated with sex (A) and sample histology (B). (C-D) Zoom in of UMAP analysis clusters 3 (C) and 4 (D), annotated with patient IDs. (E) Correlations between bulk (upper triangle) and purified methylation rates (lower triangle) and matched normal (upmost and right-most columns) using all tumor-normal DMPs in one or more of the seven tumor samples from CRUK0062. Samples are ordered by tumor purity. (F) SNV-based phylogenetic tree for CRUK0062 taken from Jamal-Hanjani et *al.* (Jamal-Hanjani et al., 2017). (G) Hierarchical clustering of the samples based on CAMDAC deconvolved and bulk methylation profiles from (E).

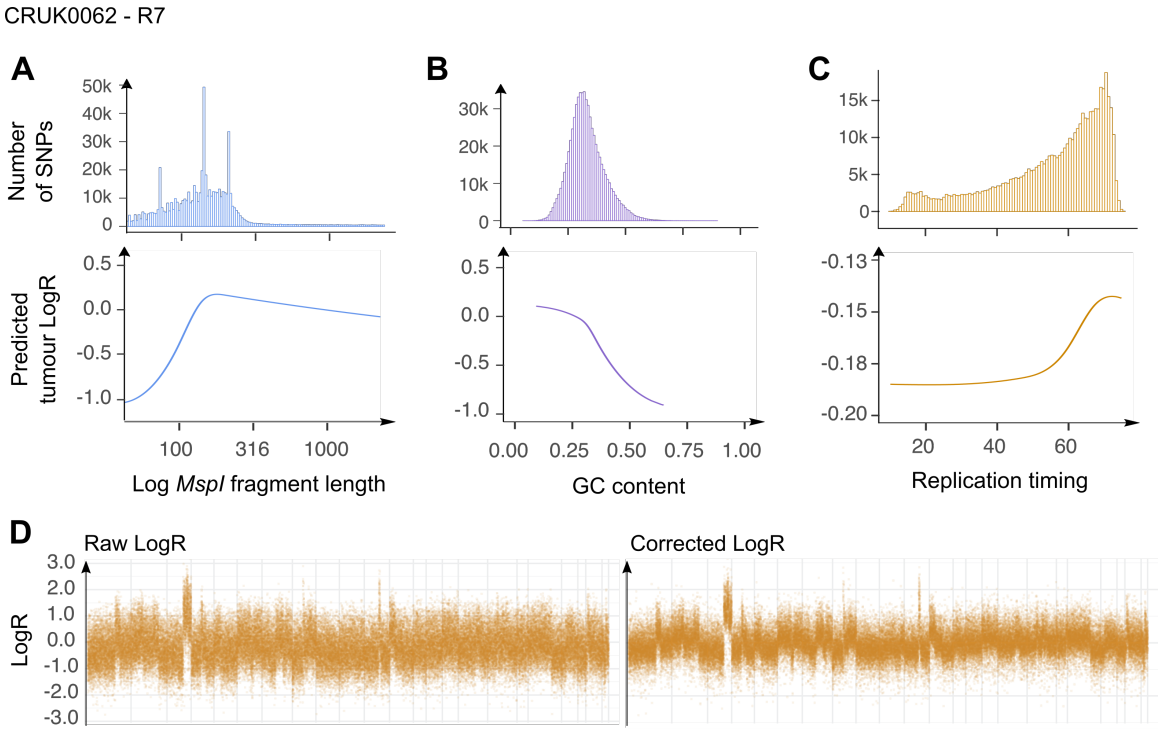
# SUPPLEMENTARY FIGURES

## Figure S1. Tumor purity and copy number affect methylation rates.



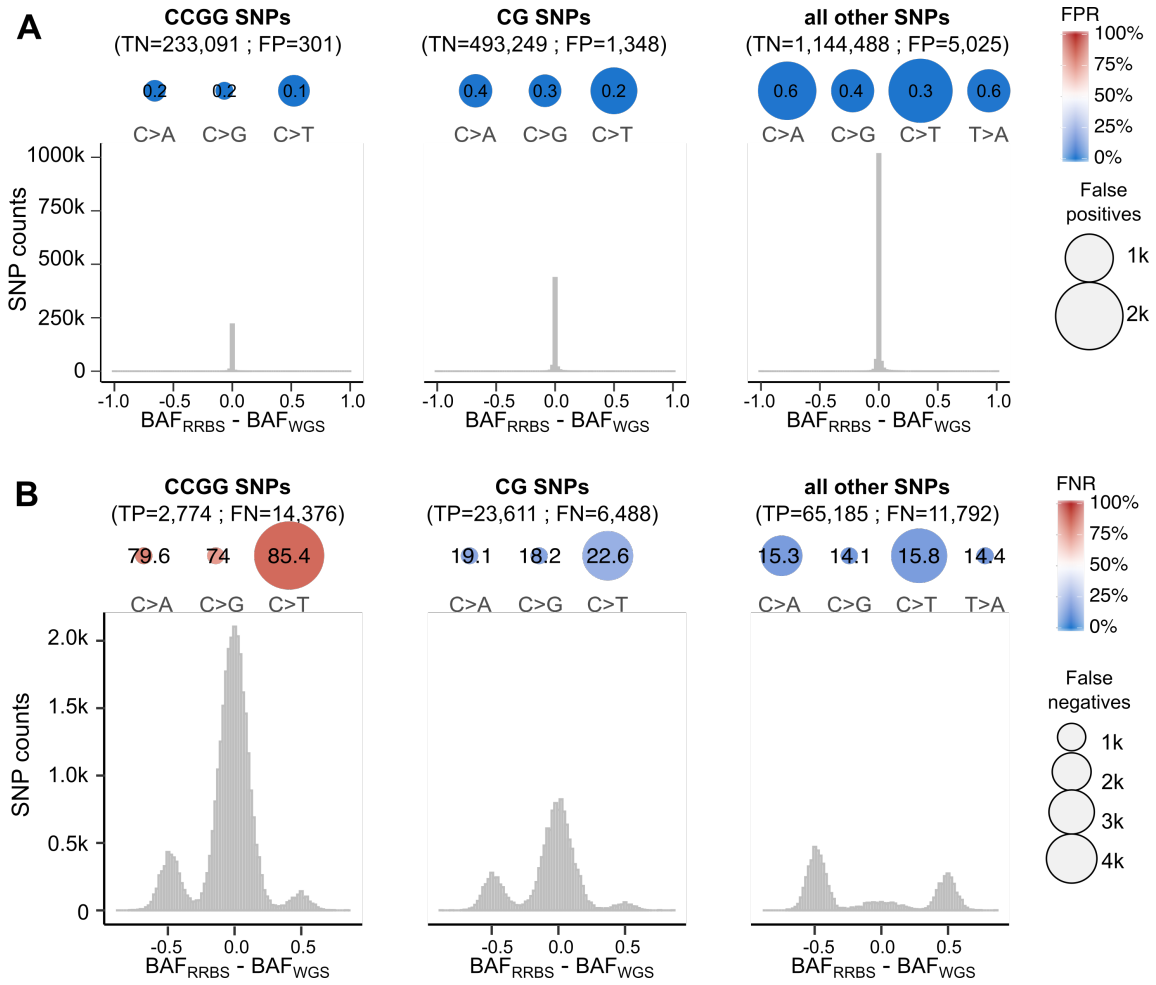
(A) Normal and bulk tumor methylation rate density for sample CRUK0062-R3 and adjacent patient-matched normal showing most CpGs are either fully methylated or unmethylated. An increased number of CpGs with intermediate methylation levels is found in the bulk tumor. (B) Simulated 99% Highest Density Interval ($HDI_{99}$) at example unmethylated CpGs with variable number of methylated and unmethylated read counts and total CpG coverage. (C-F) Normal and bulk tumor methylation rate ($m_n$ and $m_b$, respectively) histogram for CpGs labelled as confidently unmethylated in the normal ($HDI_{99} \subseteq [0,0.2]$). The majority of these loci are also unmethylated in the bulk tumor. (D-F) Zoom-in, highlighting CpGs with gained methylation in the tumor cells across all copy numbers (D), and in regions of copy number states 2+0 (E) and 2+2 (F). (G) Bulk tumor methylation rate histograms of CRUK0084 tumor region 3, having an estimated tumor purity ($\rho$) of 90%, for CpGs in diploid regions confidently unmethylated (left) and methylated (right) in the patient-matched adjacent normal. Hyper- and hypomethylated CpGs can be found in peaks that vary with the mutation copy number. (H) Bulk methylation rate histograms of CRUK0062 tumor regions 1-3, for CpGs which are confidently methylated in the adjacent normal sample, stratified by allele-specific copy number state. A dashed line indicates the expected mode of the methylation rate peak corresponding to clonal differentially methylated CpGs on all copies.

**Figure S2. RRBS-derived tumor LogR biases and correction.**



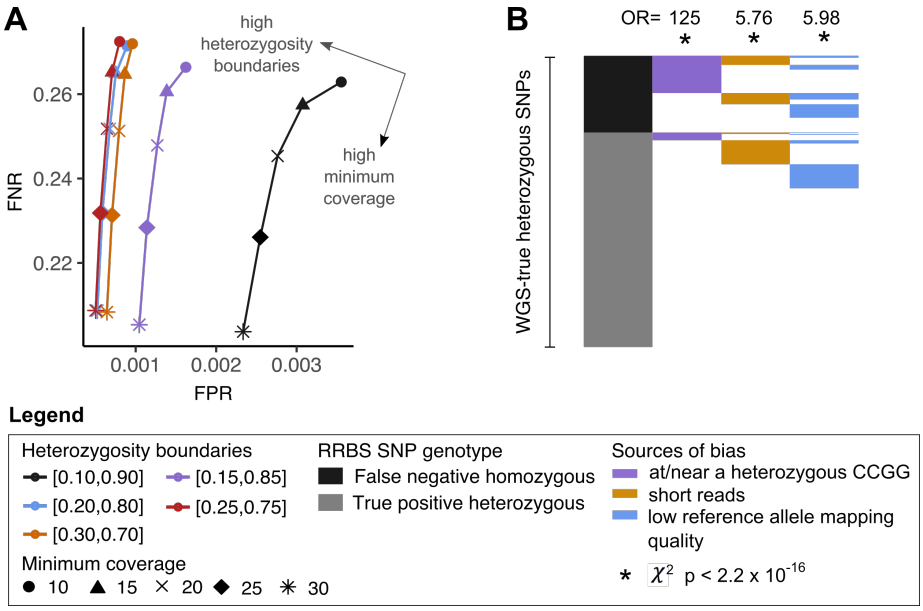

(A-C) *MspI* fragment length (A), GC content (B) and replication timing (C) affect tumor LogR. (D) A linear combination of three natural splines, modelling the effect of each of the three biases described in (A-C) is used to correct the raw LogR.

**Figure S3. Validation of CAMDAC allelic imbalance for germline heterozygous SNPs.**



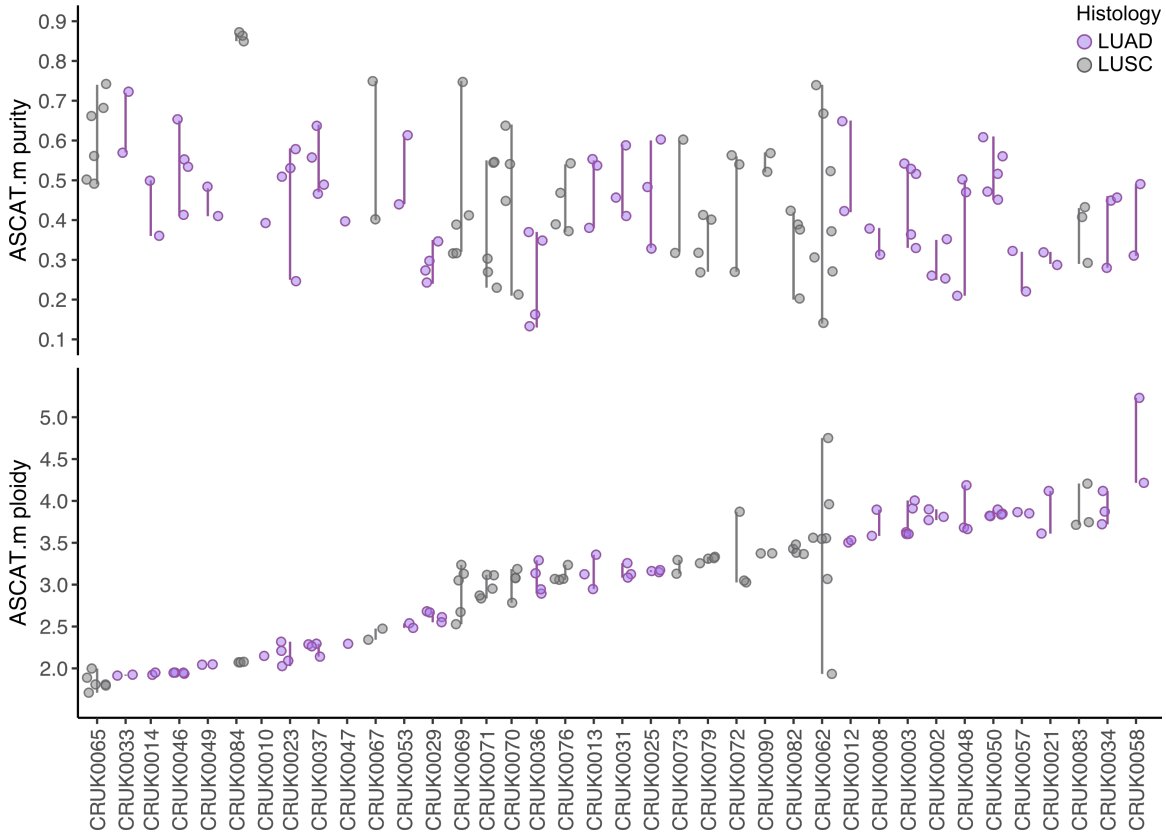

(A-B) False positive (A) and false negative (B) heterozygous call rates (%) for all SNP types (top) and histograms showing the distribution of CAMDAC BAF estimate errors and noise (bottom). SNPs are considered heterozygous when $0.1 \leq BAF \leq 0.9$.

**Figure S4. Sources of bias affecting heterozygous SNP calling from RRBS data.**



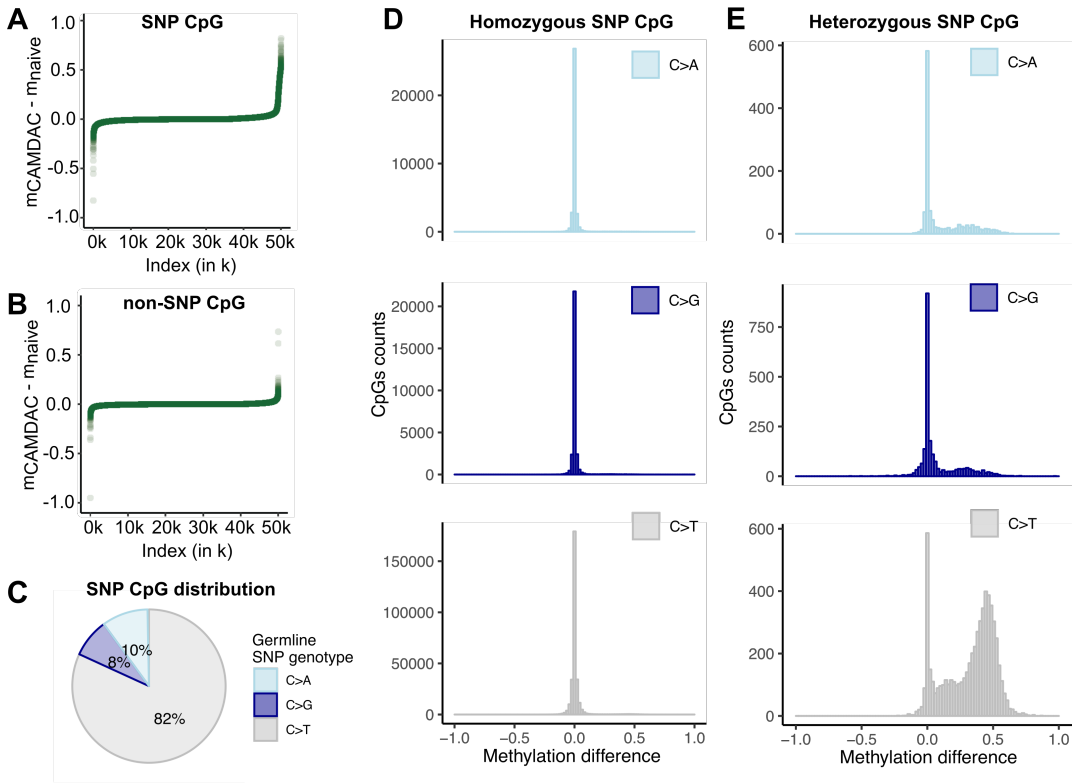

(A) Performance of heterozygous SNP identification from normal tissue RRBS data at different sequencing coverages and BAF heterozygosity boundaries. (B) Fractions of true positive and false negative heterozygous SNP calls stratified by their location (1) at/or within one read length of a heterozygous CCGG, (2) covered by only short reads (< 67bp) and (3) by reads with low reference read mapping quality. $\chi^2$ tests shows enrichment of false positives at each of these three types of regions. The data shown is from the germline of patient CRUK0062.

**Figure S5. Summary of ASCAT.m purity and ploidy results across the cohort.**



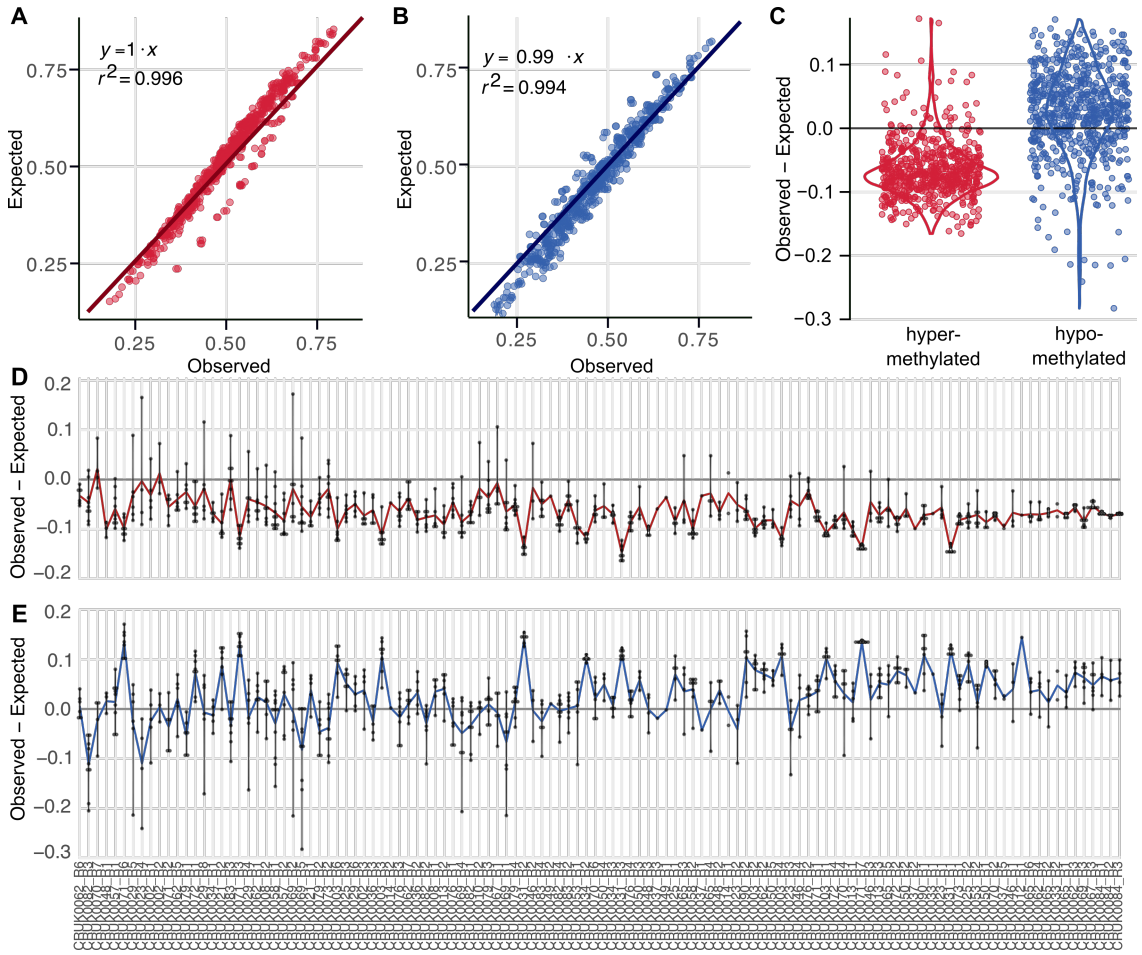

Patient-wise ASCAT.m tumor purity (top panel) and ploidy (bottom panel) distribution, as inferred by ASCAT.m.

**Figure S6. Comparison of CAMDAC and naïve bulk CpG methylation rate.**



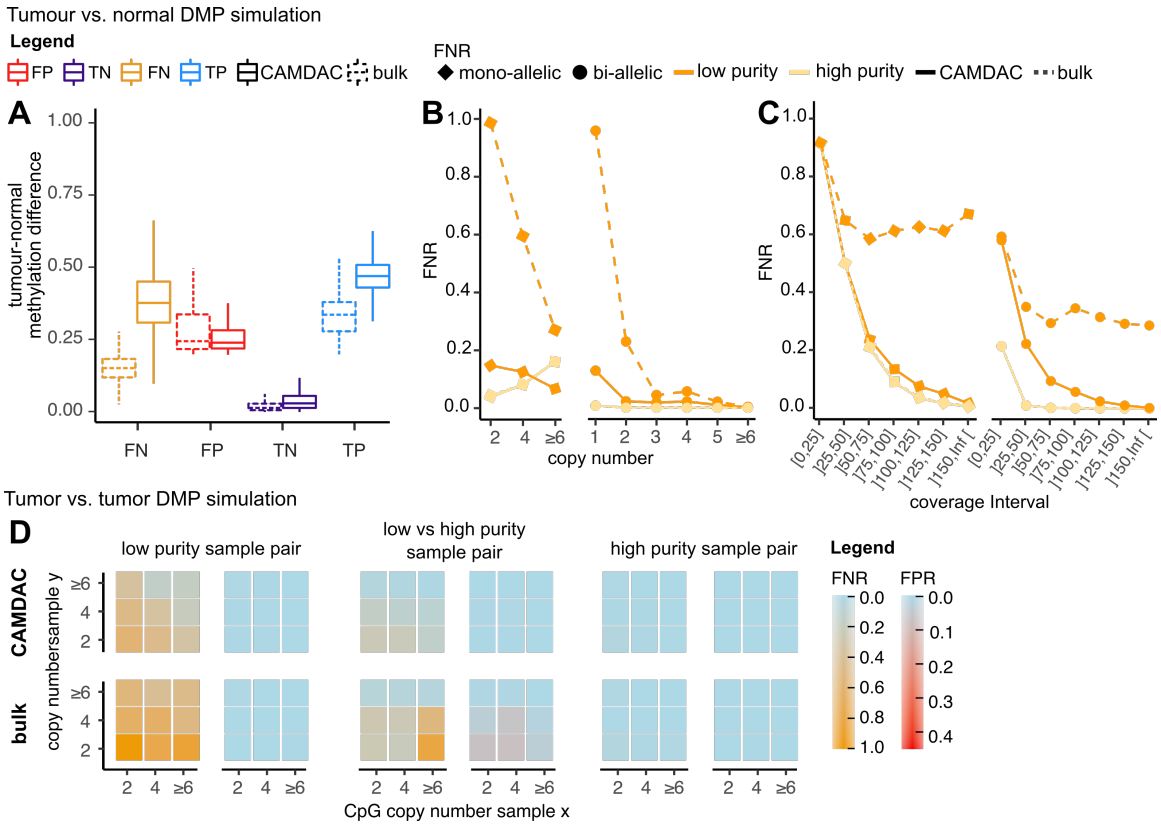

(A-B) Difference between CAMDAC and naïve bulk methylation rate estimates at 50,000 random (A) heterozygous CpGs and (B) non-polymorphic CpGs. (C) Distribution of SNP types at CpGs. (D-E) Difference between CAMDAC and naïve bulk methylation rate values at (D) homozygous *vs.* (E) heterozygous CpG, stratified by type as in (C). All data was taken from the adjacent normal sample of patient CRUK0069, which constitutes a representative example of normal lung methylation in this cohort.

**Figure S7. Validation of CAMDAC equations as a function of purity and copy number state.**



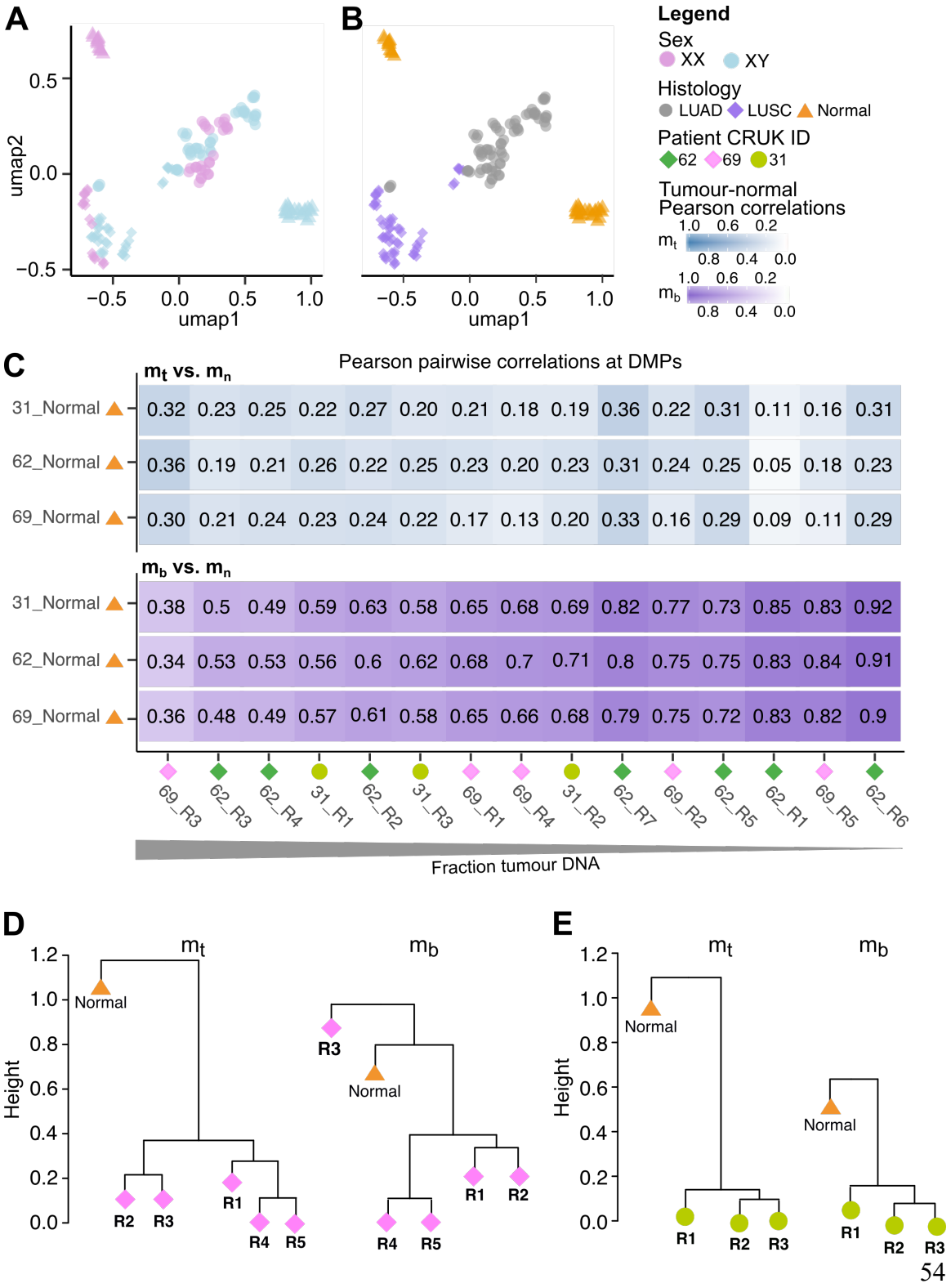

(A) Comparison of the theoretically predicted methylation rate under the CAMDAC equations as a function of tumor purity and copy number state, and the observed peak position for clonal bi-allelic tumor-normal hypermethylated CpGs that are unmethylated in the adjacent normal sample. (B) As in A, but for unmethylated CpGs that are methylated in the adjacent normal sample. (C) Observed deconvolved tumor methylation rate at clonal hypo- (blue) and hypermethylated (red) CpGs on all tumor chromosome copies compared with the expected, taken as the mean of the adjacent normal beta distribution peak for unmethylated and methylated CpGs respectively. (D-E) Observed minus expected hypermethylated deconvolved tumor methylation rate at clonal hyper- (D) and hypomethylated (E) CpGs across copy number states and tumor samples.

**Figure S8. Results of tumor-normal and tumor-tumor mono-allelic DMP simulations.**




(A-C) Results of tumor-normal DMP simulations. (A) Mean methylation difference across false negatives (FN), false positives (FP), true negatives (TN) and true positives (TP) at mono-allelic differentially methylated and non-differentially methylated CpGs. (B-C) False negative rates as a function of tumor copy number (B) and coverage (C) for the purified (solid line) and bulk tumor (dashed line) for low *vs.* high tumor purity samples (pale and dark orange lines respectively) at mono- (left panel) and bi-allelic (right panel) epimutations. (D) Results of tumor-tumor DMP simulations. False negative and false positives rates as a function of tumor copy number for low (left panel), low versus high (middle panel) and high purity (right panel) sample pairs at mono-allelic epimutations.

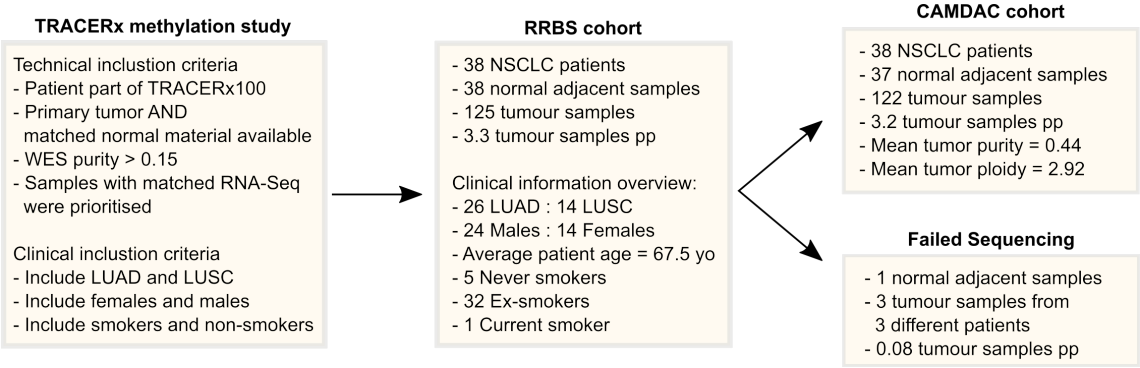
**Figure S9. CAMDAC deconvolved tumor methylation profiles are independent of purity and inform clustering analyses.**



(A-B) Results of UMAP analysis of bulk tumor methylation rates across tumor-normal promoter DMRs. (C) Pearson pairwise correlations at tumor-normal DMPs for the purified tumor (top) and the bulk (bottom), ordered by tumor purity. Tumor histological subtypes and patient IDs are annotated as well as the sampled tumor regions. (D-E) Clustering of tumor-normal DMPs from purified and bulk methylation rates for patients CRUK0069 (D) and CRUK0031 (E).

# METHODS

## TRACERx methylation study




Samples from the first 100 patients of the TRACERx lung cancer study were selected for reduced representation bisulfite sequencing (RRBS). Patients with sufficient material remaining from 2 or more tumor regions and the adjacent matched normal were considered for bisulfite sequencing. Samples with purity below 15% were discarded with the exception of CRUK0062-R6 included for comparison with the other 6 tumor regions sampled for this patient. Patients with tumor samples of high purity were prioritized as well as those with matched RNA-Seq data (Rosenthal et al., 2019). We also chose samples from both lung adenocarcinoma (LUAD) and squamous cell carcinoma (LUSC) subtypes, genders and smoking status.

### Reduced representation bisulfite sequencing

RRBS was obtained for roughly a third of the NSCLC patients from the TRACERx 100 cohort (122/327 tumor regions from 38/100 patients, each with matched normal). The NuGEN Ovation RRBS Methyl-Seq System, adapted by the manufacturer for automation on Agilent Bravo liquid handling robot, was used to generate sequencing libraries by enzymatically digesting 100 ng of gDNA using *MspI*, which recognises 5′-CCGG-3′ sequences and cleaves the phosphodiester bonds upstream of CpG di-nucleotide, leaving a 2bp overhang suitable for adaptor ligation and then a final end repair step. Generated libraries were bisulfite converted using Qiagen's EpiTect Fast DNA Bisulfite Kit.

Converted libraries were amplified by PCR using 12 cycles and purified using Agencourt® RNAClean® XP magnetic beads. Purified libraries were quantified by Qubit dsDNA HS Assay (Invitrogen) and quality was evaluated using Agilent Bioanalyzer High Sensitivity DNA Assay (Agilent Technologies). Eight samples were multiplexed per flow cell and sequenced on HiSeq2500 system using HiSeq SBS Kit v4 in paired-end 100bp runs for CRUK0062 and single end 100 bp runs for the others yielding an average of 150M raw sequencing reads per sample. Sequencing output was quality checked with FastQC (Babraham Institute, https://www.babraham.ac.uk/) and adapter sequences were trimmed with Cutadapt (Martin, 2011) and reads are aligned to the UCSC hg19 reference assembly using Bismark v0.14.4 (Krueger and Andrews, 2011). Read deduplication was carried out using NuDup, leveraging NuGEN's molecular tagging technology (NuGEN, https://github.com/nugentechnologies/nudup).

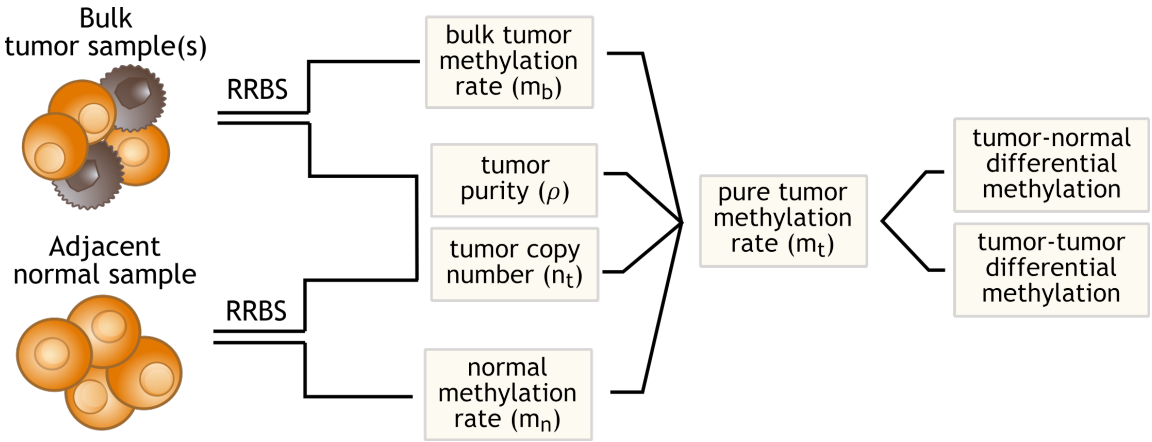
On average 1.88 x $10^8$ reads per sample remained post-processing and alignment (**Table S2**), resulting in an average of 4.5 million CpGs being supported by at least one read in any one sample. The RRBS data reported here were in part deposited into the European Genome-phenome Archive (EGA) under accession number EGAS00001003484 as part of Rosenthal *et al.* (Rosenthal et al., 2019) and the full cohort is now available at EGAS0000XXXXXXX.

**Whole genome sequencing**

Whole genome sequencing was performed on 8 samples from 3 patients included in the TRACERx100 cohort and the epiTRACERx cohort. The whole genome sequencing experiments were conducted by Edinburgh Genomics. Samples were sequencing on Illumina HiSeq in paired-end 100bp runs. Outputs sequencing reads were quality checked with FastQC v0.11.5 (Babraham Institute, https://www.babraham.ac.uk/) and then aligned to the UCSC hg19 reference assembly using BWA-MEM v0.7.15 (http://bio-bwa.sourceforge.net). Binary alignment map files (.bam) were sorted and indexed with SAMtools v0.1.19 (http://github.com/samtools/).

**Copy-number aware methylation deconvolution analysis of cancers (CAMDAC)**

As depicted below, CAMDAC requires RRBS data prepared from bulk tumor and matched adjacent normal samples. CAMDAC is only compatible with human (directional) RRBS data. The input must be quality and adapter trimmed with PCR duplicates removed and subsequently aligned to hg19 (hg38, GRCH37 and GRHCH38 formats also compatible). The tumor and matched normal data must be provided in .bam file format. The files must be sorted and indexed using SAMtools (http://github.com/samtools/). The input tumor and matched normal sequencing data is used to compute tumor purity and allele-specific copy numbers as well as bulk tumor and normal methylation rates. From this, we extract purified tumor methylation rates and carry out tumor-normal and tumor-tumor differential methylation analysis.




We created a reference RRBS genome listing all genomic regions supported by at least 5 reads in at least one of the 37 germlines from this study. Genomic regions which fail to reach this criterion are unlikely to achieve minimum depth thresholds required for copy number and differential methylation analysis and can be excluded, ultimately to speed up computation.

**Tumor copy number profiling from RRBS data**

First, we used CAMDAC get_allele_counts() command (build="hg19", mq=0, n_cores=12) to compile base counts for SNP loci described by the 1000 Genomes project

(1000g) (Auton et al., 2015) which overlap with the above-mentioned list of genomic regions. By default, no mapping quality threshold is set to avoid creating a bias against the alternate allele at SNPs. Allele counts are obtained for each of these SNP to calculate the BAF and LogR.

The LogR value at SNP $i$ represents the base-2 logarithm of the ratio of the coverage in the tumor ($cov_{t,i}$) to that of the normal at that SNP ($cov_{n,i}$), divided by the average ratio. This is easily computed from RRBS data, as shown in Equation (4).

$$\log R = \log_2\left(\frac{cov_{t,i}/cov_{n,i}}{\frac{\sum_{j=1}^{n}(cov_{t,j}/cov_{n,j})}{n}}\right) \qquad Eq(4)$$

Normalization with the adjacent normal SNP coverage, removes germline copy number variants. Next, we correct these germline normalized LogR values for fragment length, GC content and replication timing biases (**Figure S2**). This is a necessary process post-normalization as technical and biological biases affecting the sequencing output can vary between the normal and bulk tumor data. In the first instance, we reconstruct the underlying *MspI* fragment distribution from the sequencing output for each patient. Assuming, and confirming, complete enzymatic digestion, *MspI* fragment ends can be found at CCGG motifs, taking into account SNP forming/destroying cleavage sites. A CCGG is confirmed when supported with a 5' read starting with one of the following trinucleotides: CGG(+), TGG(+), CCG(-) or CCA(-). Having identified the *MspI* fragments, we can readily annotate the fragment length. Then, we compute the replication timing estimate at every 1000g SNPs leveraging Repli-seq data from 15 cell lines produced by the ENCODE project (Dataset GEO Accession: GSE34399), generating 15 reference profiles. The profile which best fits the LogR data in a given tumor sample is then used to calculate the replication timing for each *MspI* fragment in that sample. To compute the GC content of all fragments, we leverage information from the bulk tumor reads, adjusting the reference GC content for methylation rate and bisulfite conversion. Where the fragments are longer than 2x read length, we assume that all reference cytosines are converted to thymines unless they are at CpG where we assume a probability equal to the mean of all informative CpGs loci in the

fragment. Finally, we fit the observed LogR to a linear combination of the natural splines (df = 5) of *MspI* fragment length, replication timing and GC content. The model residuals provide corrected LogR values.

The BAF value at the ith 1000g SNP is the ratio of the alternate allele counts to the total counts at the ith SNP locus, $BAF = \frac{alternate\ allele\ counts}{reference\ allele\ counts + alternate\ allele\ counts}$. The strand-specific BAF rules for every SNP type are shown in **Figure 2D**. Here, we make the reasonable assumption that all reference cytosines are converted into thymines (under-conversion rate <0.5%). Note that where a variant makes up more than 5% of the total allele counts at a given SNP, it is excluded from downstream analyses in order to ensure that SNVs or mis-alignments do not affect BAF estimates and subsequent copy number profiling.

Where allelic imbalance is such that two distinct band are generated, CAMDAC can assign heterozygous SNPs ($0.3 \leq BAF \leq 0.7$) to either the gained or lost alleles. SNP phasing information can then be shared between tumor samples from the same patient. This can be used to identify mirrored subclonal allelic imbalance (Jamal-Hanjani et al., 2017), but most importantly, it allows us to rescue signal in tumors where there was no clear separation before haplotyping.

With partially phased BAF and corrected logR estimates in hand, CAMDAC then uses ASCAT (Van Loo et al., 2010) piece-wise constant segmentation (penalty = 200) leveraging germline heterozygous SNPs ($0.3 \leq BAF \leq 0.7$) and copy number fitting functions (gamma = 1) to obtain allele-specific copy number profiles and purity estimates for each tumor region. Note that we set the minimum germline coverage for SNP inclusion to 10 whilst one read is deemed sufficient in the tumor. To remove outliers and further reduce noise without introducing a bias against homozygous deletions, we remove singletons with coverage below 10 in the tumor if and only if their nearest neighbors within a 10kb moving window all pass the above threshold.

There were 6 patients which showed large intra-tumor differences in their CAMDAC ploidy estimates. In each of those cases, we looked for evidence supporting an alternative

ploidy solution. If such a solution existed and was better suited the overall tumor ploidy profile, the ASCAT copy number fitting step was re-run forcing the solution to the ploidy towards a diploid or tetraploid solution (see MINPLOIDY and MAXPLOIDY variables, from the runASCAT() function in ascat.R script; available from https://github.com/Crick-CancerGenomics/ascat). We refit 7 tumor copy number profiles in total. This is an improvement from the 17 samples which were manually curated and re-run in the matched exome sequencing data (Jamal-Hanjani et al., 2017).

**Comparing RRBS- and WGS-derived genotypes**

First, we compiled allele counts at all 1000g SNPs with coverage in the matched RRBS data using alleleCount (http://cancerit.github.io/alleleCount/). We compared CAMDAC RRBS-derived genotypes with the output from standard alleleCount (http://cancerit.github.io/alleleCount/) pipeline on WGS data for the 3 germline samples of patients CRUK0031, CRUK0062 and CRUK0069. Positive calls are heterozygous SNPs ($0.1 \leq BAF \leq 0.9$) and WGS is considered to provide ground truth with a minimum SNP read depth of 10 required on both platforms for inclusion. We compute the average false positive rate (FPR) as the sum of all false positives over the sum of false positives and true negatives across samples. Similarly, the mean false negative rate (FNR) is calculated as the sum of all false negatives divided by the sum of false negatives and true positives. The FPR allows us to evaluate allele-specific biases introduced by the approach, while the FNR assesses the influence of methylation status on BAF calculation, as well as biases introduced during the RRBS protocol. For all SNP types, we consider three different contexts: (i) SNPs at CCGG, the recognition motif of the *MspI* restriction enzyme used during library preparation, (ii) SNPs at CpGs (excluding CCGG motifs), and (iii) all other. A chi-square analysis is performed to test whether or not the FNR is context dependent. Next, we recompute FNR and FPR aggregating FP, TP, FN and TN across all samples for a range of heterozygosity boundaries and minimum coverage thresholds (**Figure S4A**). Next, we measure the number of false negatives due to (i) allele-specific *Msp1* fragments,

(ii) additional reference mismatches from neighboring SNPs, and (iii) short fragments yielding invalid alignments given 2 or less reference mismatches (**Figure S4B**).

The mean CAMDAC FPR across all SNPs and samples was 0.3% and was always below 2%. The average FNR was 25% and was highly context dependent ($\chi^2$ test, p-val < 2.2 x $10^{-16}$, $FNR_{CCGG}$ = 83%, $FNR_{CG}$ = 20%, $FNR_{other}$ = 15%). Higher sequencing coverage results in only modest improvements in FNR: by requiring a minimum SNP coverage of 30 instead of 10 reads, the average FNR dropped to 20% (**Figure S4A**). As expected, the majority of heterozygous SNPs perturbing or creating an *MspI* recognition motif are missed (49% of all false negatives**, Figure S4B**). These polymorphisms result in allele-specific fragments during RRBS library preparation that skew allelic coverage. Heterozygous SNPs on the same *MspI* fragment as a heterozygous CCGG are similarly affected. Finally, within the false negative calls, we see a bias towards SNPs erroneously called homozygous reference (72%) compared to homozygous alternate (28%), and this effect is present in all three contexts, although less pronounced at CCGGs. This points towards an alignment bias, possibly explained by limited mappability of short *MspI* fragments with alternate alleles. Supporting this theory, SNPs with lower reference allele read mapping quality scores were more likely to yield homozygous reference false negatives (Wilcoxon test, p < 2.2 x $10^{-16}$), especially outside CCGG context.

**Tumor copy-number profiling from whole genome sequencing data**

First, we compiled allele counts at all 1000g SNPs with coverage in the matched RRBS data using alleleCount (http://cancerit.github.io/alleleCount/). We then we ran ASCAT GC and replication timing LogR correction steps, fed the output into ASCAT piece-wise constant segmentation (penalty = 200) and obtained copy number and purity (gamma = 1). Comparing WGS and RRBS-derived profiles, even with the higher FNR on SNP calling in RRBS data, CAMDAC BAF estimates enable ASCAT to correctly identify separation of the BAF bands on several occasions where the WGS BAF does not seem to provide sufficient power. Likewise, comparison of the LogR tracks indicates that RRBS protocol-induced biases in the coverage are adequately modelled and removed by CAMDAC.

Unsurprisingly, final allele-specific copy number segments are virtually indistinguishable between the two data types.

**Methylation rate calculation from RRBS data**

The bulk tumor and matched normal methylation rate is readily computed by taking the ratio of methylated CpG read counts to the sum of methylated and unmethylated read counts, $m = \frac{methylated\ read\ counts}{methylated\ read\ counts + unmethylated\ read\ counts}$. CAMDAC uses strand-specific dinucleotide counts to distinguish between methylated and unmethylated CpGs, m $= \frac{\text{CG}}{\text{CG + TG(+) + CA(–)}}$. At C>T (T>C) and G>A (A>G) SNPs, only reads from the bottom strand and top strand, respectively, contribute to CAMDAC methylation rate estimates. In addition, only the CG-forming allele contributes to the methylation rate at polymorphic CpGs. This enables the methylation rate at a heterozygous CpG to vary between 0 and 1, rather than between 0 and 0.5 in a diploid sample, and ensures further independence between methylation rate and copy number estimates.

We compiled bulk tumor and normal methylation rates for all CpGs which fell within the above-mentioned reference RRBS genomic regions list. For each patient, we discarded all CpGs that failed to reach a minimum coverage of 10 in the matched normal RRBS data. CpGs that had less than 3 reads in a given tumor sample were also filtered out from that sample.

**CAMDAC-purified tumor methylation rates from RRBS data**

Bulk tumor methylation rate ($m_b$) could be expressed as a function of the methylation rate in the tumor cells ($m_t$) and contaminating normal cells ($m_n$), scaled for the purity of the sample ($\rho$) and the local copy number state (i.e. $n_n$ = 2 in the normal and $n_t$ in the tumor):

$$m_b = \frac{\rho n_t m_t + n_n m_n (1-\rho)}{\rho n_t + n_n (1-\rho)} \qquad Eq(1)$$

Taking estimates of the total copy number at each CpG and tumor purity from the RRBS copy number profile and assuming the matched normal is a reasonable proxy for the contaminating normal, we can calculate the deconvolved tumor methylation rate $m_t$:

$$m_t = \frac{m_b\big(\rho n_t + n_n(1-\rho)\big) - n_n m_n (1-\rho)}{\rho n_t} \qquad Eq(2)$$

An important distinction is made at heterozygous polymorphic CpGs, where only one copy in the diploid normal contains methylation information (i.e. $n_n = 1$). Similarly, if the SNP is CG-destroying and its BAF is smaller than 0.5 or likewise if the SNP is CG-forming and BAF is above 0.5 the major copy informs the methylation (i.e. $n_t = n_{major}$). *Vice versa*, if the tumor BAF is larger than 0.5 and the SNP is CG-destroying or the BAF lower than 0.5 and the SNP is CG-forming, the minor allele only informs the methylation rate (i.e. $n_t = n_{minor}$).

Finally, $m_t$ must be between 0 and 1. However, due to both technical and biological noise, values can fall outside of these boundaries. For downstream analyses, we set negative methylation rates up to 0 and values sitting above the upper boundary are rounded down to 1. In virtually all cases, the $m_t$ 99% HDI overlaps with allowed values.

**SNV-phased methylation rate estimates**

Leveraging SNV calls from newly obtained WGS data and previously published WES (Jamal-Hanjani et al., 2017), we phased CpG methylation to all SNVs, excluding loci with $\mathrm{VAF} \leq 0.1$ in a tumor sample or $\mathrm{VAF} > 0$ in the patient-matched adjacent normal tissue. Allele-specific methylation counts were compiled for all reads that could be phased to exactly one SNV. Phased methylation rates were obtained for 32,874 CpGs and 6,529 SNVs across samples (14,514 and 2,984 unique CpGs and SNVs respectively across patients). The VAF was derived from the mutant (mut) and wild type (WT) reads counts: $\mathrm{VAF} = \frac{counts_{mut}}{counts_{mut}+counts_{WT}}$. RRBS-derived VAF estimates were compared with those obtained from the WES/WGS (Pearson correlation = 0.86, **Figure 4B**). The mutation copy number, $n_{mut}$, was then computed as a function of the variant allele frequency, tumor

purity and copy number: $n_{mut} = \frac{1}{\rho} \times VAF \times \rho n_t + n_n(1-\rho)$. The wild type allele copy number, $n_{WT}$, is obtained by subtracting $n_{mut}$ from $n_t$: $n_{WT} = n_t - n_{mut}$. The mutant allele methylation rate, $m_{mut}$, is extracted by taking the counts methylated ($M_{mut}$) and unmethylated ($UM_{mut}$) divided by all counts phased to the variant allele: $m_{mut} = M_{mut}/(M_{mut} + UM_{mut})$. The wild type allele methylation rate, $m_{WT}$, is confounded by signal from normal contaminating cells and must be deconvolved. For this, we use a modified version of the CAMDAC equations *2* and *3*, where the tumor methylation rate and copy number are expressed in terms of the mutant and wild type alleles.

$$m_b = \frac{\rho(n_{mut}m_{mut} + n_{WT}m_{WT}) + n_n m_n(1-\rho)}{\rho(n_{mut} + n_{WT}) + n_n(1-\rho)} \qquad Eq(5)$$

$$m_{WT} = \frac{m_b\big((n_{mut} + n_{WT}) + n_n(1-\rho)\big) - n_n m_n(1-\rho) - \rho n_{mut} m_{mut}}{\rho n_{WT}} \qquad Eq(6)$$

We validate CAMDAC $m_t$ by comparison with methylation estimates phased to clonal SNVs present on all copies in regions with loss of heterozygosity across our cohort ($n_{mut} = n_t$). At these sites, all reads reporting the variant allele can directly be assigned to the tumor cells, and methylation rates obtained from this subset of reads should be an unbiased estimate of the purified tumor methylation rate (i.e. $n_{mut} = 0$ ). Overall, 4,485 CpG loci met these criteria. A high correlation was observed between these SNV deconvoluted $m_t$ values and CAMDAC estimates (Pearson correlation = 0.97, **Figure 4C**).

**Tumor-normal differential methylation analysis**

First, to call DMPs between tumor and normal, we designed a statistical test to robustly identify DMPs between tumor and normal. The number of methylated reads at a CpG is often described as a Beta-Binomial distribution where the mean of the probability distribution, the methylation rate, follows a Beta distribution. We can therefore model the observed bulk tumor and matched normal methylation rate at CpG locus i as follows:

$$m_{x,i} \sim B\big(\alpha_{x,i}, \beta_{x,i}\big) \qquad Eq(7)$$

where: $m_{x,i}$ is the methylation rate at the ith CpG locus

$\alpha_{x,i}$ is the counts methylated at the ith CpG locus

$\beta_{x,i}$ is the counts unmethylated at the ith CpG locus

$x$ is $b$ for the bulk tumor, $n$ for the normal and $t$ for the purified tumor

Our statistical design aims to test whether or not the Beta distribution of the tumor methylation rate at the ith CpG locus, $m_{t,i}$, is different from the normal methylation rate, $m_{n,i}$. This is depicted in Equations *8* and *9*.

$$m_{n,i} = m_{t,i} \qquad Eq(8)$$

$$\Delta B = m_{t,i} - m_{n,i} \qquad Eq(9)$$

We recall from *Eq(2)* that the tumor methylation rate, $m_t$, can be expressed as a function of $m_b$ and $m_n$. By substituting *Eq(2)* into *Eq(9)*, we obtain:

$$\Delta B = \frac{m_{b,i}\left(\rho n_{t,i} + n_{n,i}(1-\rho)\right) - n_{n,i}m_{n,i}(1-\rho)}{\rho n_{t,i}} - m_{n,i} \qquad Eq(10)$$

Which can be simplified to:

$$\Delta B = \left(m_{b,i} - m_{n,i}\right) \times C \qquad where \quad C = \frac{\rho n_{t,i} + n_{n,i}(1-\rho)}{\rho n_{t,i}} \qquad Eq(11)$$

Expression (11) demonstrates that our statistical model is independent of copy number and tumor purity, which are constant for the i$^{th}$ CpG site. However, the power to call DMPs will intrinsically depend on purity, local copy number and CpG coverage. The former two will affect the magnitude of the difference between $m_b$ and $m_n$ whilst the latter will affect the width of each beta distribution. Since there is a closed-form solution for testing $P(m_{b,i} > m_{n,i})$, but not for $P(m_{b,i} = m_{n,i})$ (Robinson, 2017), the null and alternate hypotheses are written as follows:

$H_0: m_{b,i} = m_{n,i}$ The methylation rate of the i$^{th}$ CpG is identical in normal and tumor

$H_1: m_{b,i} > m_{n,i}$ The tumor is hypermethylated at this locus

$H_2: m_{b,i} < m_{n,i}$ The tumor is hypomethylated at this locus

where $P(m_{b,i} > m_{n,i}) = \sum_{j=0}^{(\alpha_{n,i}-1)} \frac{B(\alpha_{b,i}+j,\beta_{b,i}+\beta_{n,i})}{(\beta_{n,i}+j)+B(1+j,\beta_{n,i})+B(\alpha_{b,i},\beta_{b,i})}$ $Eq(12)$

and $P(m_{b,i} < m_{n,i}) = \sum_{j=0}^{(\alpha_{b,i}-1)} \frac{B(\alpha_{n,i}+j,\beta_{n,i}+\beta_{b,i})}{(\beta_{b,i}+j)+B(1+j,\beta_{b,i})+B(\alpha_{n,i},\beta_{n,i})}$

$$= 1 - P(m_{b,i} > m_{n,i}) \qquad Eq(13)$$

*Eq(13)* can be rewritten as follows, incorporating our $Beta(0.5, 0.5)$ prior to each methylation count variable. The prior informs on the underlying methylation rate distribution and ensures finite logarithms:

$$P(m_{b,i} > m_{n,i}) = \sum_{j=0}^{(\alpha_{n,i}-1)} \log B(\alpha_{b,i} + j + 0.5, \beta_{b,i} + \beta_{n,i} + 1) - \log(\beta_{n,i} + j + 0.5) - \log B(1 + j, \beta_{n,i} + 0.5) - \log B(\alpha_{b,i} + 0.5, \beta_{b,i} + 0.5) \qquad Eq(14)$$

We compute the probability that a CpG site is hypo- or hypermethylated and for easier interpretation, we express these probabilities as their complement ($C$ = 1-$P$), which is a measure the overlap between $m_{b,i}$ and $m_{n,i}$. If $C(m_{b,i} > m_{n,i})$ or $C(m_{b,i} < m_{n,i}) \leq 0.01$ we accept $H_1$ or $H_2$ respectively.

Differential methylation analyses incorporating a beta distribution model reportedly show higher true positive and lower false discovery rate are obtained compared with Fisher's and z-score methods (Raineri et al., 2014). Nevertheless, given high enough coverage, even a small difference in methylation can become statistically significant. In order to focus our analysis of biologically significant methylation changes, we require a minimum effect size of 0.2 between the purified tumor methylation rate, $m_{t,i}$, and the matched normal, $m_{n,i}$, for DMP calling. In theory, this allows for subclonal or allele specific changes to be picked up whilst removing spurious signal. We obtained a second set of DMPs by applying the minimum effect size threshold on the difference between the bulk tumor methylation rate, $m_{b,i}$, and the normal methylation rate, $m_{n,i}$ as is customary (Dolzhenko and Smith, 2014; Hansen et al., 2012; Klein and Hebestreit, 2016; Robinson et al., 2014; Wu et al., 2015), and compared the output with CAMDAC calls. The threshold of 0.2 was deemed

sufficiently low to capture most mono-allelic aberrations whilst be sufficiently high to remove false positives and filter noise from the heterogeneous normal contaminating cells.

**Tumor-tumor differential methylation analysis**

The purified tumor methylation rate is not a beta distribution but rather the difference between two betas, the bulk tumor and normal methylation rate, scaled for tumor purity and copy number. As such, there is no exact solution to compute the highest density interval for $m_{t,i}$. To address this, we simulate a credible 99% HDI for $m_{t,i}$ at every CpG. We use the tumor purity and CpG copy number and simulate 2000 data points for the bulk tumor ($m_{b,i}$) and matched normal methylation rate ($m_{n,i}$), given $m_{x,i} \sim B(\alpha_{x,i}, \beta_{x,i})$. Substituting these into *Eq(2)*, we obtain a vector of values for $m_{t,i}$ and readily extract the 99% HDI. If the purified tumor methylation rate HDI does not overlap between any two tumor regions at a given CpG and the minimum effect size, 0.2, is reached, a tumor-tumor DMP is identified.

**Tumor-normal and tumor-tumor differential methylation rate simulation framework**

To appreciate the effect of tumor purity on both tumor-normal and tumor-tumor differential methylation analysis, we extracted the 20 lowest and highest purity samples in our cohort, $\rho < 0.3$ and $\rho > 0.58$, respectively. We combined the methylation information at all overlapping autosomal CpGs from these samples including the patient-matched normal of selected samples. The normal methylation rate at confidently unmethylated and methylated CpGs was extracted. Confidently unmethylated and methylated CpGs are respectively defined as having their methylation rate 99% highest density interval (HDI) boundaries in the 0.0-0.2 and 0.8-1.0 intervals (**Figure S1A-B**). This vector of values incorporates information for both distributions such as mean methylation rate and deviation as well as their respective contribution to the overall bimodal normal methylation rate profile. Sampling from this vector will yield our simulation priors.

Next, we randomly selected intra- and inter-tumor CpG pairs from samples within or across purity categories and of equal or differing total copy numbers (range 1-6). For each 168 possible combination of these simulation parameters, we sampled 10,000 loci from each tumor samples randomly assigned as $x$ and $y$ as well as their matched normal.

For each locus, we begin by obtaining the coverage information from $i^{th}$ selected CpGs in sample $s = x$ or $y$. We know the bulk tumor coverage ($cov_{b,s,i}$), the tumor copy number ($n_{t,s,i}$), the normal copy number ($n_{n,s,i} = 2$) and the global tumor purity ($\rho_{s,i}$) and so the tumor DNA fraction ($f_{t,s,i}$) is $f_{t,s,i} = \frac{n_{t,s,i} \times \rho_{s,i}}{n_{t,s,i} \times \rho_{s,i} + (2 \times (1 - \rho_{s,i})}$. We can work out the purified tumor coverage as $cov_{t,s,i} = cov_{b,s,i} \times f_{t,s,i}$ where $cov_{t,s,i} \sim Binom(cov_{bs,i}\ ,\ f_{t,s,i})$. The matched normal coverage ($cov_{n,s,i}$) is therefore $cov_{n,s,i} = cov_{b,s,i} - cov_{t,s,i}$.

We then sample a normal methylation rate prior ($p_{n,i}$) from the confidently unmethylated ($v_{unmeth}$) and methylated CpGs ($v_{meth}$) from the matched normal data which is used to simulate the normal methylation rates of normal contaminating cells from both tumor samples $x$ and $y$. For each sample, we obtain the counts methylated ($M_{n,s,i}$) and unmethylated ($UM_{n,s,i}$): $M_{n,s,i} \sim Binom(cov_{n,s,i}, p_{n,i})$ and $UM_{n,s,i} = cov_{n,s,i} - M_{n,s,i}$.

The purified tumor methylation rate is obtained by randomly selecting from the same vector as the matched normal ($\{p_{n,i}, p_{t,i}\} \in v_{unmeth}$ or $\{p_{n,i}, p_{t,i}\} \in v_{meth}$) or from opposite vector states ($p_{t,i} \in v_{unmeth}$ and $p_{n,i} \in v_{meth}$ or $p_{t,i} \in v_{meth}$ and $p_{n,i} \in v_{unmeth}$). We obtain the counts methylated ($M_{t,s,i}$) and unmethylated ($UM_{t,s,i}$) as $M_{t,s,i} \sim Binom(cov_{t,s,i}, p_{t,s,i})$ and $UM_{t,s,i} = cov_{t,s,i} - M_{t,s,i}$. The bulk methylation counts for both samples are easily calculated by adding the tumor and normal counts: $M_{bs,i} = M_{n,s,i} + M_{t,s,i}$ and $UM_{b,s,i} = UM_{n,s,i} + UM_{t,s,i}$. In balanced copy number regions, we also simulate allele-specific DMPs whereby one allele is in the normal ground state and the other is differentially methylated. We obtain the counts methylated from the minor allele, allele A, and for the major allele, allele B, and combine them to obtain total counts methylated and unmethylated.

The expected absolute tumor-normal methylation difference at simulated bi-allelic DMPs is $|m_t - m_n| \approx 0.95$ (**Figure 5A**). In the bulk, the magnitude of the difference depends on sample tumor purity and CpG copy number. The purified methylation rate at mono-allelic DMPs usually depends on the copy number of the mutated allele, however, in balanced copy number regions, where $n_A = n_B$ and given that one copy is clonally differentially methylated and the other is in ground, $m_t = \frac{M_{t,A} + M_{t,B}}{(M_{t,A} + UM_{t,A}) + (M_{t,B} + UM_{t,B})} \approx 0.5$. The expected tumor-normal difference at simulated mono-allelic DMPs is thus $|m_t - m_n| \approx 0.5$ (**Figure S8A**).

Tumor-normal and tumor-tumor differential methylation calls were made using CAMDAC differential methylation analysis and the output compared with the ground truth. False negative and positive rates are obtained for the bulk and deconvolved tumor simulated data (**Figures 5B-D and S8B-D**).

**Tumor-tumor differential methylation analysis on real data**

To evaluate the impact of CAMDAC bulk tumor methylation deconvolution, we compared deconvolved with bulk tumor-tumor DMP calls between all 21 sample pairs from CRUK0062, taken for having the most tumor regions. Both DMP call sets were obtained by calculating the exact overlap between the 99% HDI of the methylation rate at the $i^{th}$ CpG ($m_{b,i}$) between sample pairs and applying a minimum effect size of 0.2.

To appreciate the effect of tumor purity on both inter- and intra-tumor-tumor differential methylation, we extracted the bulk and CAMDAC deconvolved tumor methylation profiles from the 20 lowest and highest purity samples in our cohort, $\rho < 0.3$ and $\rho > 0.58$, respectively. We randomly selected 1 million CpG loci from samples with the same or different purity categorical assignments both within and between patients and obtained tumor-tumor DMP calls from $m_b$ and CAMDAC $m_t$.

**Quantifying allele-specific methylation**

We obtained allele-specific methylation rate estimates at the imprinting control region (ICR) of the *H19/IGf2* by phasing the germline heterozygous SNPs. The normal methylation rate for the reference allele, $m_{n,ref}$, is computed from the methylated ($M_{n,ref}$) and unmethylated ($UM_{n,ref}$) reads phased to the reference allele: $m_{n,ref} = M_{n,ref}/(M_{n,ref} + UM_{n,ref})$. *Vice versa*, the alternate allele normal methylation rate, $m_{n,alt}$, is calculated as: $m_{n,alt} = M_{n,alt}/(M_{n,alt} + UM_{n,alt})$. Overall, 31/37 patients had polymorphisms at the *H19/IGf2* ICR. On average, we could phase 15 CpGs per sample to these heterozygous SNPs.

Obtaining bulk tumor methylation rates at the reference and alternate allele, $m_{b,ref}$ and $m_{b,alt}$, respectively, follows the same principle. To obtain the purified tumor methylation rate, we first assign clonal copy numbers to each allele using the BAF at the heterozygous and the allele-specific copy number from ASCAT.m. If BAF < 0.5, then major allele, $n_A$, is the reference allele copy number, $n_{ref} = n_A$ and the minor allele, $n_B$, is the alternate allele, $n_{alt} = n_B$. If BAF > 0.5, $n_{alt} = n_A$ and $n_{ref} = n_B$. Both the reference and alternate allele methylation rates are confounded by signal from normal contaminating cells and must be deconvolved. We modified CAMDAC equations *2* and *3* for this purpose, where $x$ is either the reference or alternate allele:

$$m_{b,x} = \frac{\rho n_x m_{t,x} + n_{n,x} m_{n,x}(1-\rho)}{\rho n_x m_{t,x} + n_{n,x}(1-\rho)} \qquad Eq(15)$$

$$m_{t,x} = \frac{m_{b,x}\big(\rho n_x m_{t,x} + n_{n,x}(1-\rho)\big) - n_{n,x} m_{n,x}(1-\rho)}{\rho n_{n,x}} \qquad Eq(16)$$

Using $m_{t,x}$, we classified tumor samples has having retained imprinting when absolute differences between the reference and alternate allele phased methylation rates at 1 or more CpGs across the ICR were larger than 0.7: $|m_{t,ref} - m_{t,alt}| > 0.7$. Assignments were then manually validated. Tumors with loss of heterozygosity were classified as having lost either the inactive or active copies by taking the average methylation rate across CpGs within the ICR which were phased to the remaining allele. Similarly, in tumors with loss

of imprinting, we took the average methylation level across CpGs at the ICR phased to either alleles to determine whether gain or loss of methylation had occurred.

**Identifying tumor-normal differential methylated regions**

First, CpGs are grouped into bins. CpGs which fall within 100bp of one another are grouped together. For each bin, we compute the number of consecutive DMPs with effect size 0.2 and $p < 0.01$ and the total number of DMPs. Genomic bins with 4 or more consecutive DMPs and at least 5 DMPs in total are deemed differentially methylated. Each bin is annotated for downstream analyses. Annotated gene features include CpG Islands, shores and shelves, exons, introns, 5'UTR, 3'UTR, promoters and enhancers. CpG islands, intragenic features and enhancers genomic coordinates are pulled from Ensembl via biomaRt. CpG island shores and shelves are respectively defined as the regions 0-2kb and 2kb-5kb either side of a CpG island. Intragenic annotations are simplified for each gene to the GENCODE basic transcript set. Each gene transcript promoter is defined as starting 2.5kb upstream and ending 250bp downstream of the transcription start site. For this work, we used CAMDAC with hg19 annotation set, but hg38 is also available. Note that any given CpG cluster can be associated to several features.

To validate our DMR calls, we leverage SNV purified tumor methylation rates. First, we perform tumor-normal DMP calling following the same logic as in $Eq(14)$, but this time computing the probability that $P(m_{mut} > m_n)$, $P(m_{WT} > m_n)$ and their complements. We feed the DMP calls into CAMDAC DMR calling as described above. We then classified DMRs as either occurring on both alleles, only the mutant allele (*in-cis*) or only the wild type allele (*in-trans*). DMRs in regions with loss of wild type allele are categorized separately as it is not possible to determine whether the DMR was *in-cis, in-trans* or on both alleles prior to the loss.

**Clustering analysis**

We performed hierarchical clustering at the single patient level across 3 patients to investigate inter-sample clonal relationships. Clustering was carried out on the purified tumor methylation rate at all CpGs that were DMPs in at least one of the tumor samples. The clustering output nicely fit the SNV phylogenetic trees derived from patient-matched multi-region WES data. This analysis was repeated on the bulk tumor methylation rate and the outputs were compared. Bootstrap hierarchical clustering was performed using the R package pvclust (http://www.sigmath.es.osaka-u.ac.jp/shimo-lab/prog/pvclust/) with the hierarchical clustering method set to the average and using a Pearson distance matrix. For each clustering, we ran 100 bootstrap iterations with r-values going from 0.5 to 1.5.

Next, we compiled a list of all CpGs which fell within a promoter-associated tumor-normal DMR in one or more samples based on CAMDAC purified methylomes (totaling 8,570 gene promoters). We then calculated the mean methylation rate for each of those genomic regions across this cohort's 159 samples (122 tumor and 37 normal lung samples) and performed UMAP clustering analyses. We repeated this analysis selecting promoter regions based on the bulk tumor-normal DMR calls (totaling 8,387 gene promoters) feeding the mean bulk tumor methylation rates into the UMAP.

**CAMDAC with tissue-matched normal**

We were unable to sequence a patient-matched tumor-adjacent normal sample for CRUK0047. Leveraging this cohort's 37 normal lung samples, we were able to build a healthy lung tissue reference RRBS profile with the median normal SNP coverage and methylation counts which we fed into CAMDAC alongside tumor sample CRUK0047-R2. The ploidy and purity values ($\psi_{RRBS} = 2.15$, $\rho_{RRBS} = 0.40$) obtained are in agreement with the matched exome data ($\psi_{WES} = 2.29$, $\rho_{WES} = 0.31$). After deconvolution, >99% of CpGs had purified tumor methylation rates HDI overlapping [0,1]. Both hierarchical and UMAP clustering analyses performed on CAMDAC purified tumor methylomes found CRUK0047-R2 in the same cluster as other LUAD samples. We conclude that genome-

wide RRBS relative SNP coverage and methylation levels in normal lung are consistent between different patients and that tissue-matched normal RRBS data is a suitable alternative reference for CAMDAC when patient-matched data is unavailable.