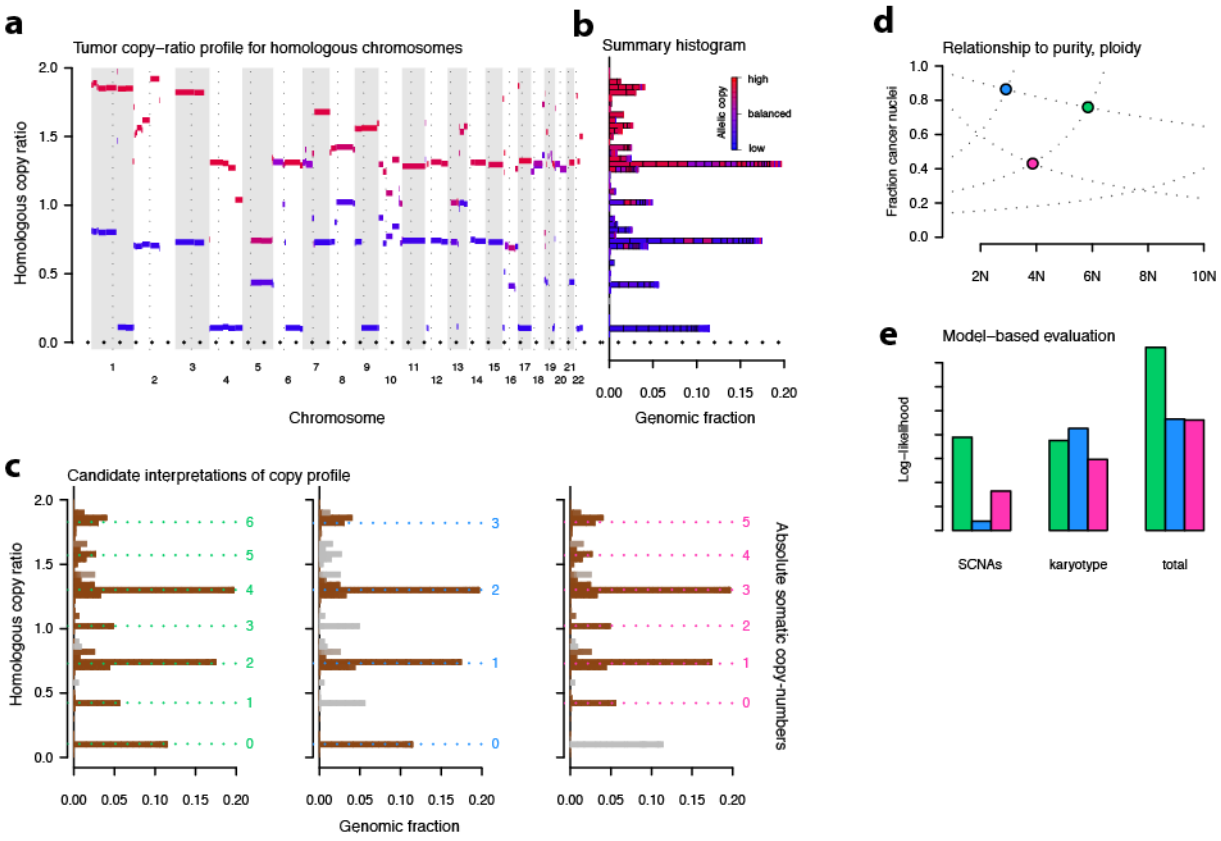
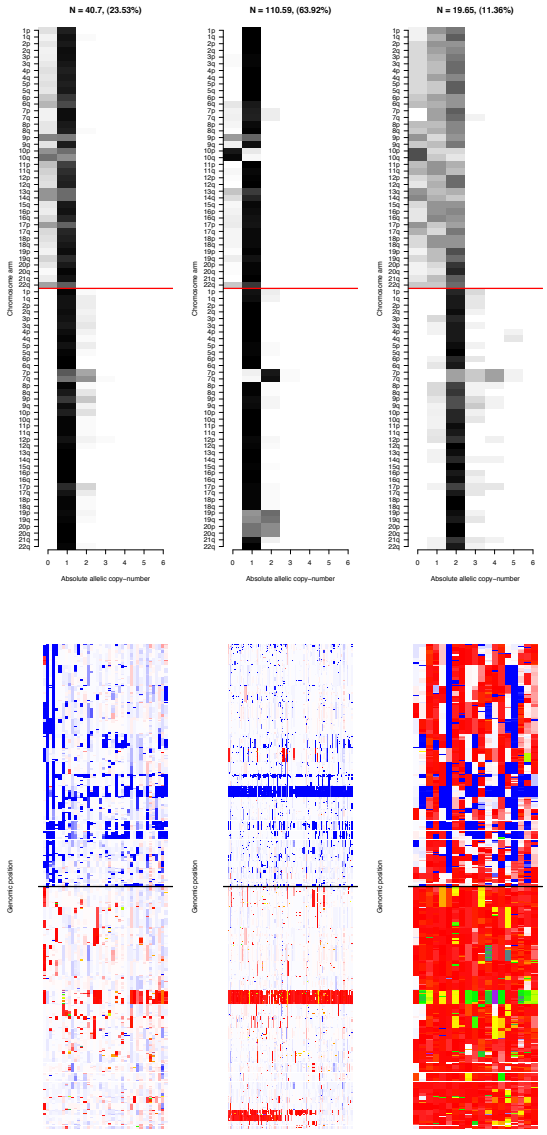
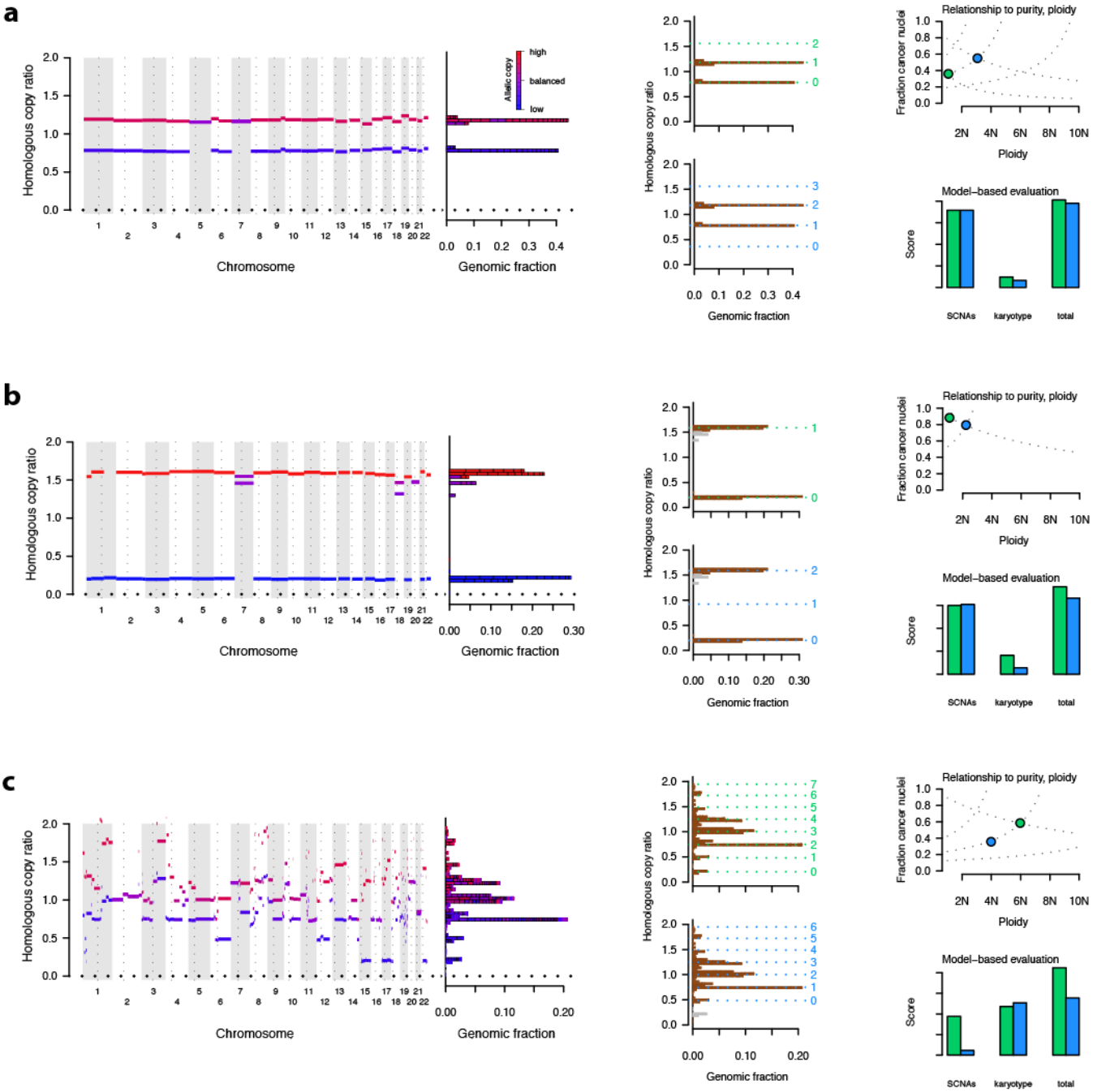


Supplementary Figure 1 | Example of ABSOLUTE analysis
a-e, An HGS-OvCa sample processed using ABSOLUTE, as in **Figure 1b-g**. In this sample, the excellent SCNA-fit score is strongly opposed to a karyotype score indicating a simpler solution, and dominates the final ranking.



Supplementary Figure 2 | Karyotype models for GBM, constructed by clustering tumor sets by absolute homolgous copy-number profile.
(top) Fit of the multivariate multinomial mixture model (Online Mehtods Eq. 8) using 3 components. Grey-scale values denote the probability of each homologous chromosome-arm (y-axis) attaining copy-numbers 0-6 (x-axis). **(bottom)** individual GBM samples are shown grouped according to their cluster membership in (top). Color-scale indicates absolute copy-number (blue – 0, white – 1, red – 2, yellow- 3, green -4. Note – low-copy HSCRs are shown on top, high-numbers on the bottom.

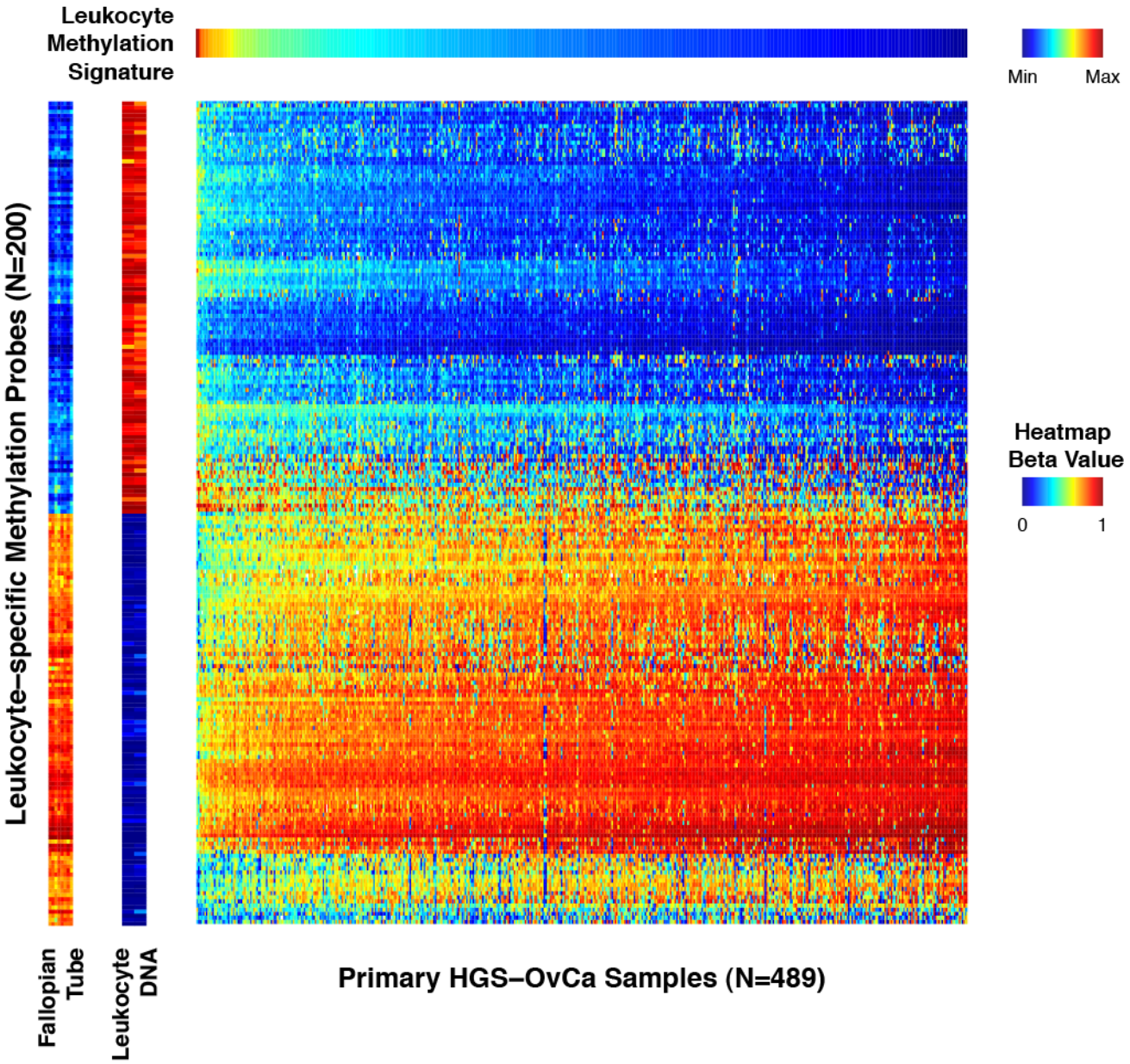
Nature Biotechnology: doi:10.1038/nbt.2203



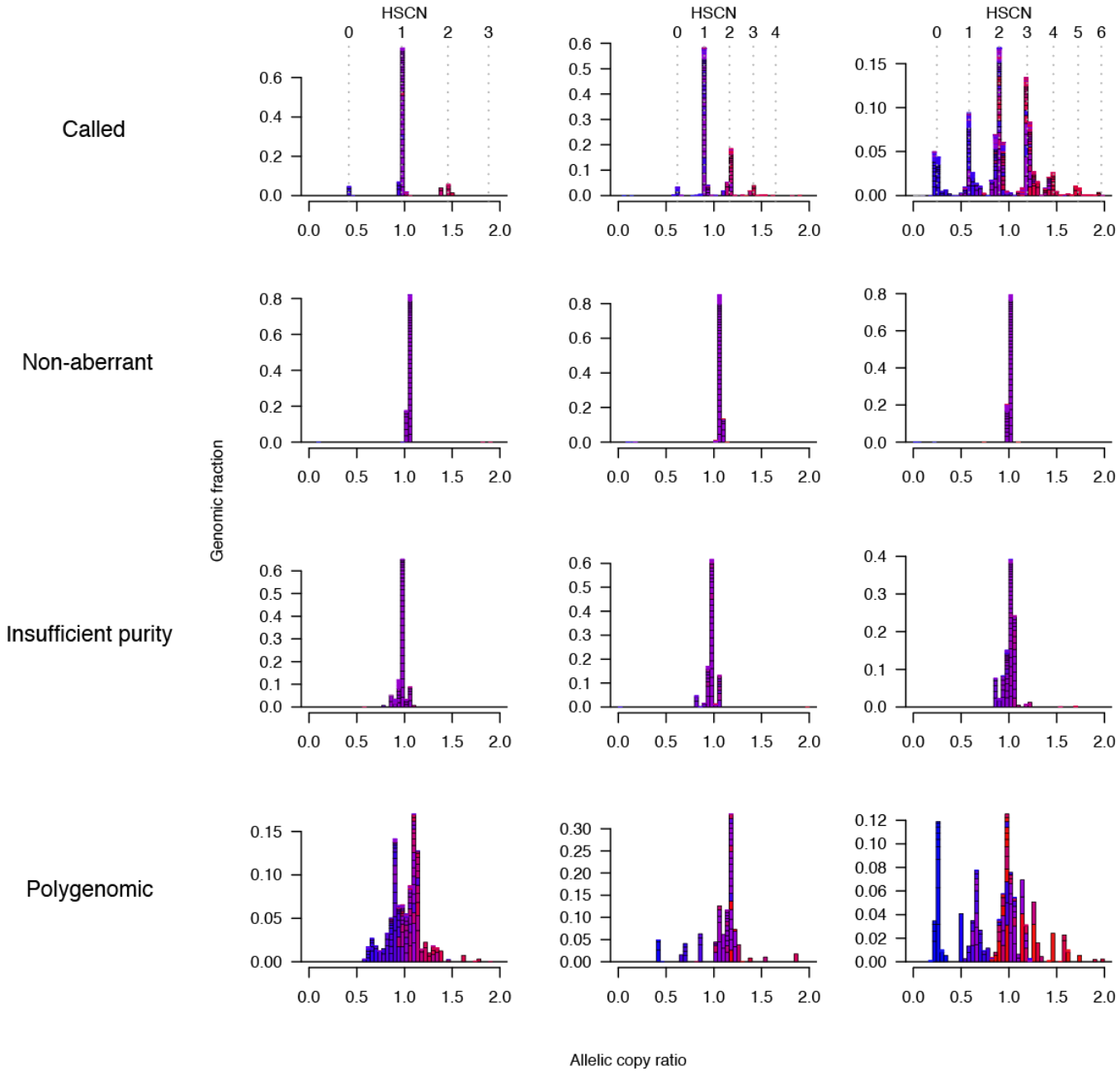
Nature Biotechnology: doi:10.1038/nbt.2203

Supplementary Figure 3 | Unusual karyotypes identified using ABSOLUTE.
a-d, Analysis of homologous copy-ratio profiles with ABSOLUTE, as in **Figure 1**.
a,b, Near-haploid genomes
c,d, Hyperaneuploid genomes
a, Lung adenocarcinoma sample SM-11ZY. Purity = 0.36, ploidy = 1.12.
b, Glioma sample 'glioma 612'. Purity = 0.88, ploidy = 1.14.
c, HGS-OvCa sample TCGA-25-1320-01A-01D-0452-01. Purity = 0.58, ploidy = 6.03.

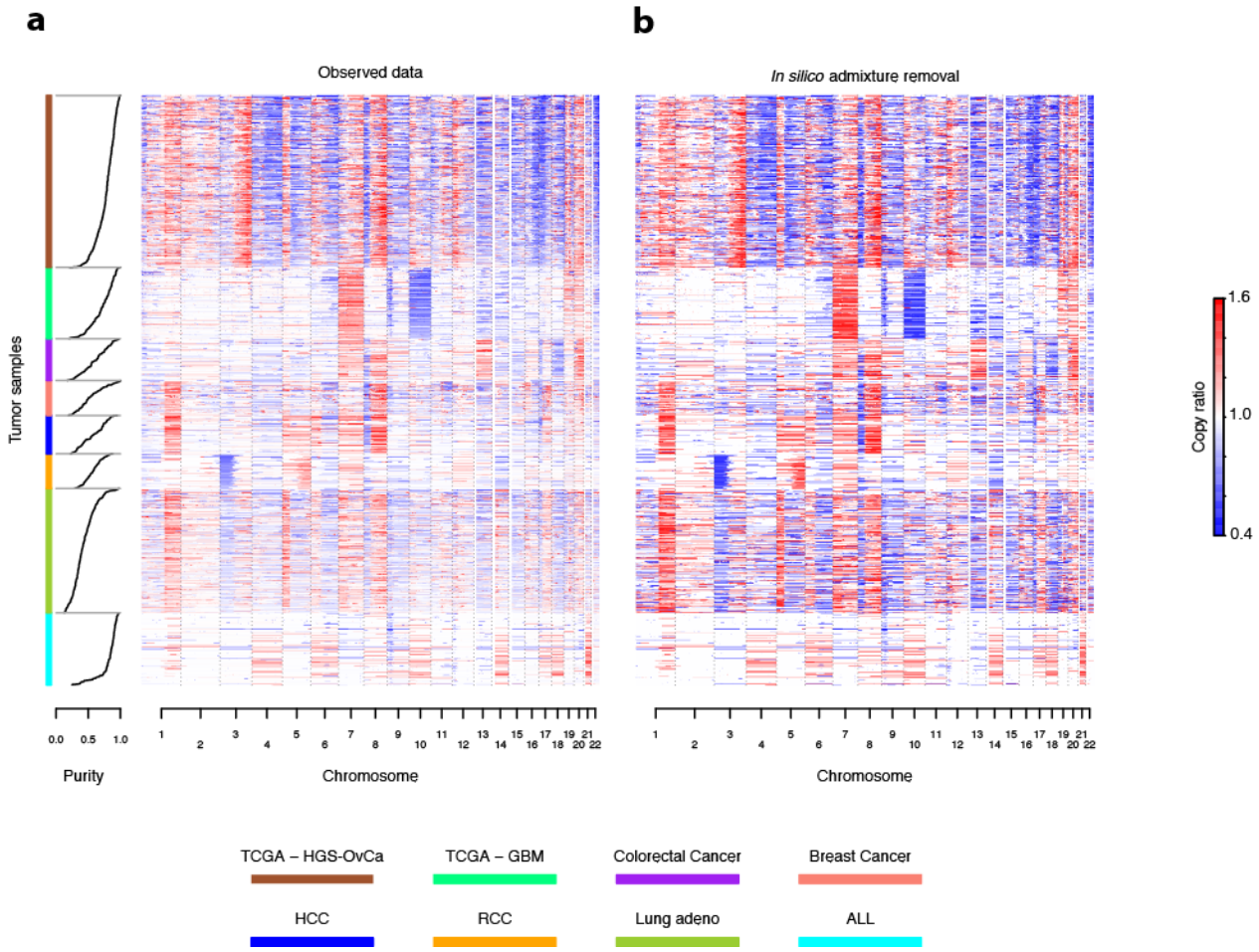
Nature Biotechnology: doi:10.1038/nbt.2203



Supplementary Figure 4 | Identification of a leukocyte methylation signature in HGS-OvCa samples. Leukocyte methylation signature (Online Methods).



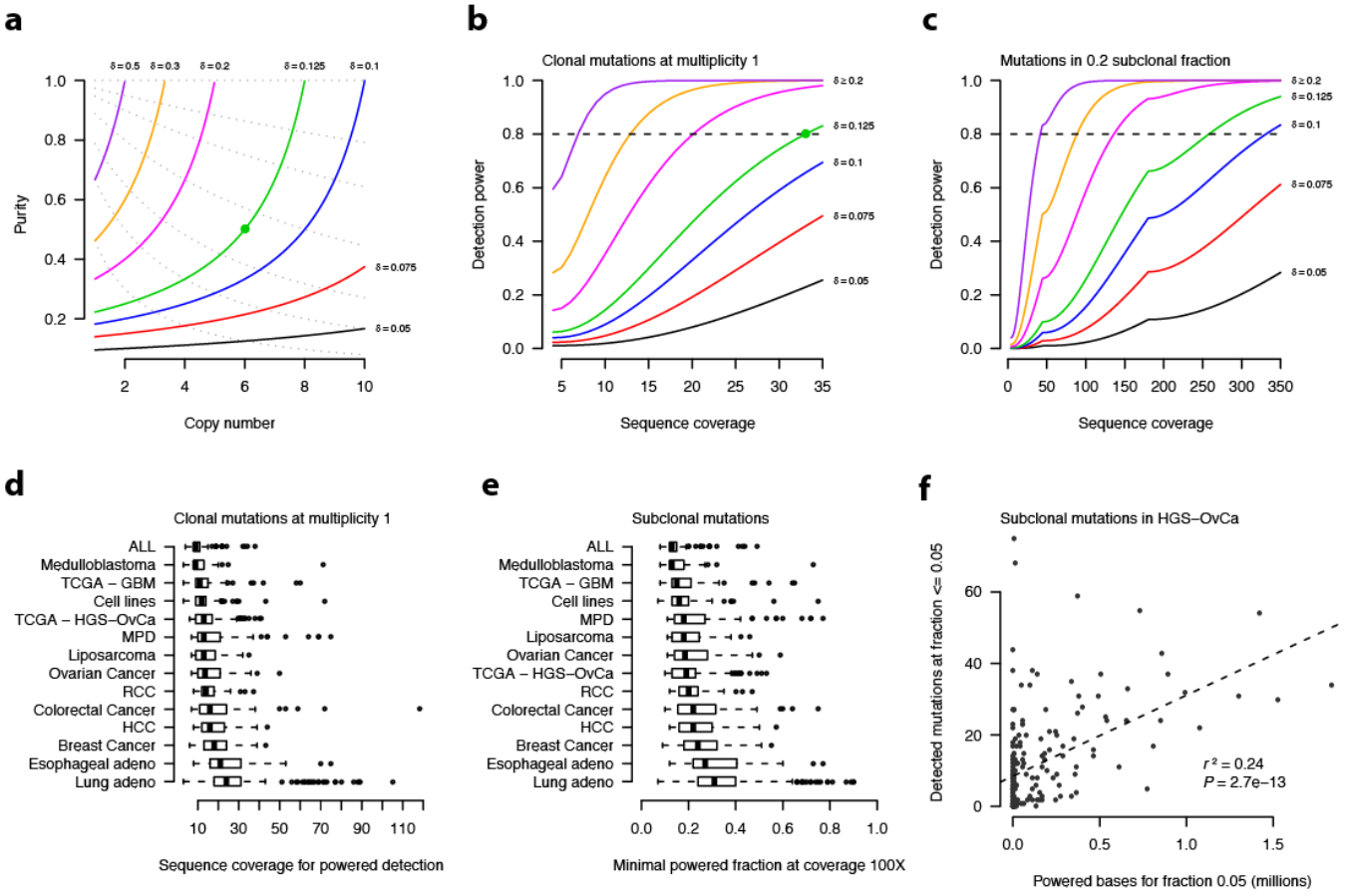
Supplementary figure 5 | Homologous copy-ratio profiles of called and uncalled cancer samples
a-d, Copy profiles for 12 cancer samples. Rows correspond to called samples (**a**), and three distinct failure modes (**b**-**d**), as indicated (left-hand labels). Homologous copy profiles are colored according to allelic imbalance, as in **Fig. 1c**. The distribution of each outcome is shown for each tumor type in **Fig. 3a**.
a, Called samples annotated with fit homologue-specific copy numbers (HSCN; top-axis).
b, The most common failure-mode designation was “non-aberrant” (9.1% of samples), indicating that the copy profile was indistinguishable from that of a normal (diploid) sample.
c, An additional 7.3% of samples failed due to insufficient purity. In these samples, copy aberrations were apparent, but so attenuated as to obscure the pattern of integer copy states.
d, The remaining 6.9% of failed samples were designated as “polygenomic”; although clear SCNAs were observed, they were not consistent with a single dominant copy profile (Eq. 1).



Supplementary figure 6 | Distribution of estimated purity in various tumor types, and effect on observed copy-ratios

a,**b,** Data are shown for each indicated sample-type, with samples sorted by tumor purity (left). Total copy ratios in were median-centered at 1 for plotting.
a, Genome wide copy-ratio profiles are shown for several tumor cohorts. The effect of tumor contamination is evident as the 'fading' of copy ratios towards 1 as tumor purity decreases.
b, The data from **a** are shown after '*in silico'* removal of tumor contamination by dividing total absolute copy-number by sample ploidy. This transformation enables more direct comparison between datasets, as samples without copy alterations may be clearly distinguished from heavily contaminated tumors.

Nature Biotechnology: doi:10.1038/nbt.2203



Supplementary figure 7 | Effect of tumor purity and ploidy on power for detection of somatic point-mutations by sequencing

a, Combinations of tumor purity and ploidy which imply equal values of $\delta=\alpha/D$ (Eqs. 1,8), the concentration-ratio of molecules present at a single copy per cancer cell (multiplicity 1) in the tumor population. Dotted lines indicate equal values of $2(1-\alpha)/D$ (Eq. 1).

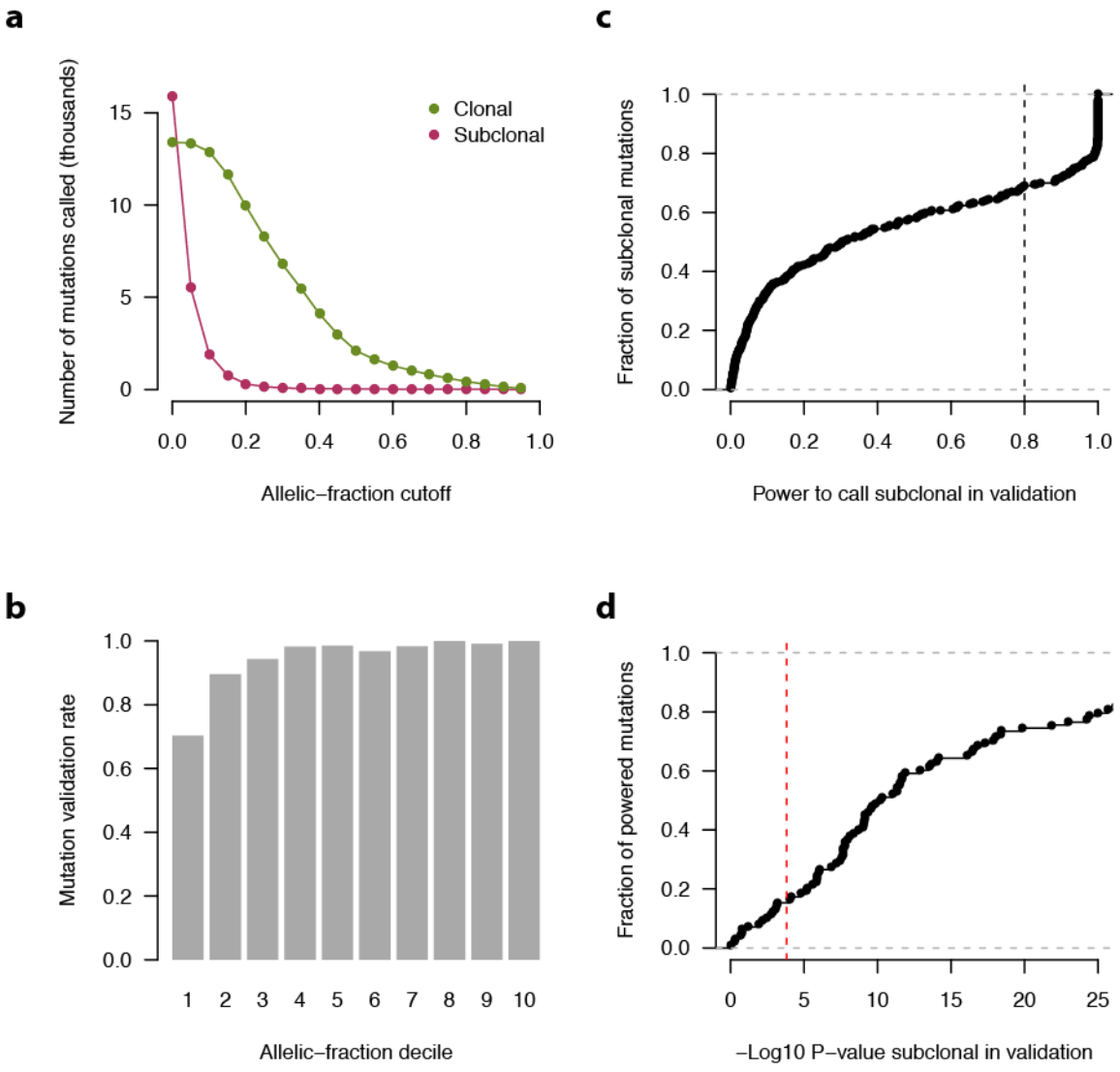
b, Theoretical power to detect clonal somatic point mutations present at multiplicity 1, as a function of sequence coverage, for various values of δ (shown in **a**). Power was calculated for FPR <= 5×10^{-7}, assuming a uniformly random sequencing error-rate of 10^{-3} (Online Methods, Eq. 9).

a,b, Green dots correspond to an example case of a region present at 6 copies, with 1 copy mutated, and purity 50%. Only 1 of 8 alleles at this locus carry the mutation (**a**), and 33 reads are required to detect the mutation with 80% power (**b**).

c, Theoretical power (as in **b**) to detect subclonal somatic point mutations present at multiplicity 1 in 0.2 of cancer cells.

d,e, Values were calculated for each tumor using the purity and ploidy estimates obtained from ABSOLUTE. Samples were considered powered if their detection power (as in **b**, **c**) exceeded 0.8.

f, Tumor purity and local absolute copy-number estimates were used to calculate the number of bases powered for the detection of subclonal mutations present at fraction 0.05 (*x*-axis, Online Methods, Eq. 9), and to compute the number of subclonal mutations at cancer-cell fraction <= 0.05 in each sample (*y*-axis). The dashed line, r^2, and *P*-value refer to a linear regression fit of the data points.



Supplementary figure 8 | Validation of subclonal mutations in ovarian cancer

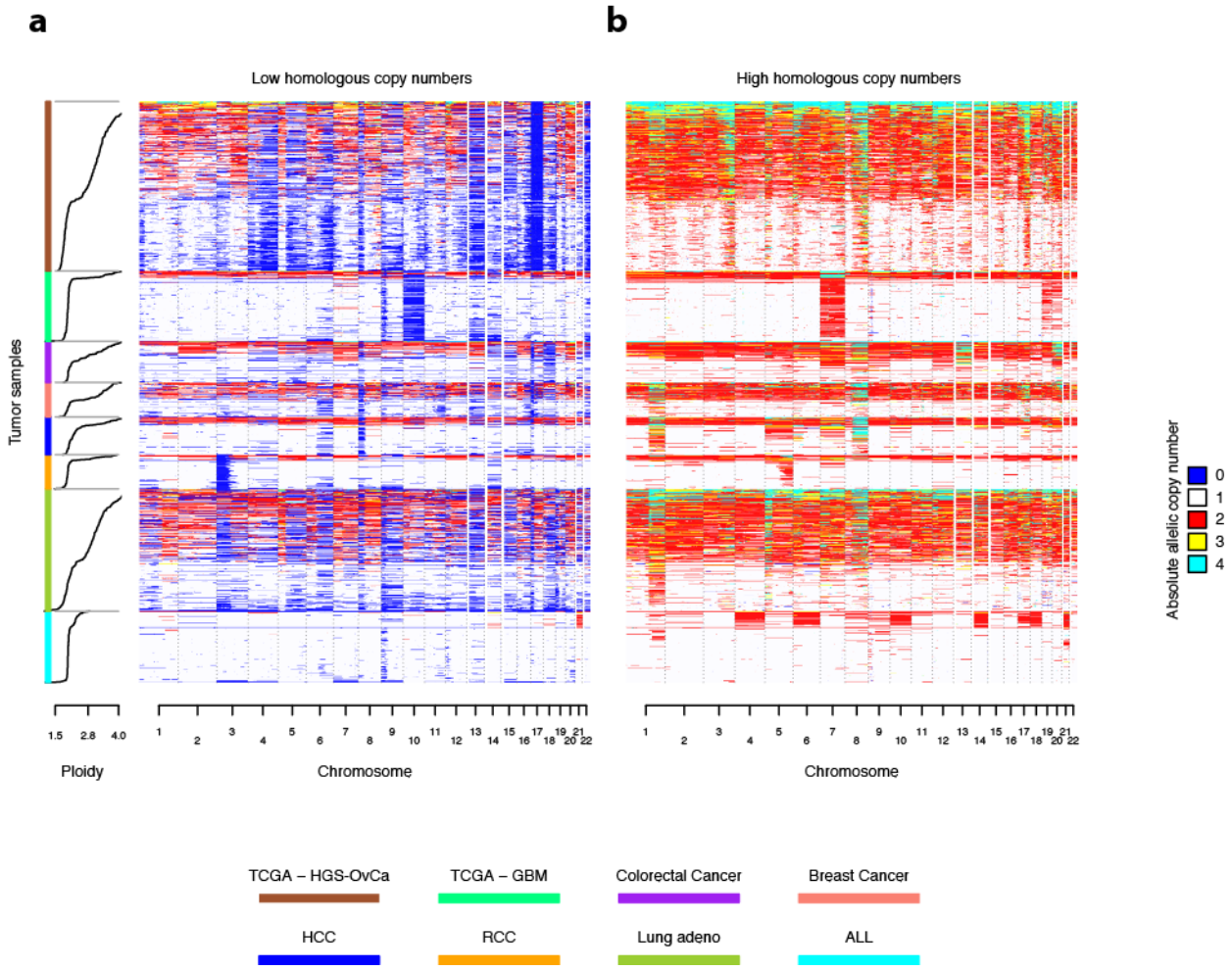
a, **b**, **Validation of somatic mutations in independent sequencing data**. A total of 6050 of the 29,268 somatic mutations called in this study (**a**) were targeted for Illumina capture sequencing of genomic DNA from an independent whole genome amplification (WGA) reaction. Of these, 3901 were powered at ≥ 0.8 for detection in the validation sequencing (Eq. 9, Online Methods), of which 3713 (95%) were detected (**b**).
a, Number of clonal and subclonal mutations called above an allelic-fraction cutoff (note that all mutations were used in the analysis for **Fig. 4**).
b, Validation rate as a function of allelic fraction decile (*x*-axis; bin 1 = allelic fraction 0 to 0.1, etc).
c, **d**, **Confirmation of sublclonal allelic fraction for validated mutations**. A total of 316 putative subclonal mutations were detected in the validation data. Of these, 98 had power > 0.8 for discrimination of subclonal status in the validation data (**c**), given the read depth (*N*), clonal allelic fraction (f_c), and subclonal allelic fraction (f_o, as estimated in the original dataset). Of these 98, 83 were confirmed subclonal in the validation sequencing with strict Bonferroni corrected *P*-values < 0.05 / 316 (**d**).

c, Cumulative distribution of estimated discrimination power. Power was calculated by solving $\alpha = \int_0^y \mathrm{Beta}\left(f \mid f_c N + 1, (1 - f_c N) + 1\right) df$ for *y*, and then

$\mathrm{Pow}\left(y, \hat{f}_o, N, \alpha\right) = \int_0^y \mathrm{Beta}\left(f \mid \hat{f}_o N + 1, (1 - \hat{f}_o) N + 1\right) df$, where α was set to the significance level 0.05 / 316.

d, Cumulative distribution of $-\log_{10}$ *P*-value for rejection of the clonal hypothesis. Dashed vertical line indicates the Bonferroni significance level 0.05 / 316.

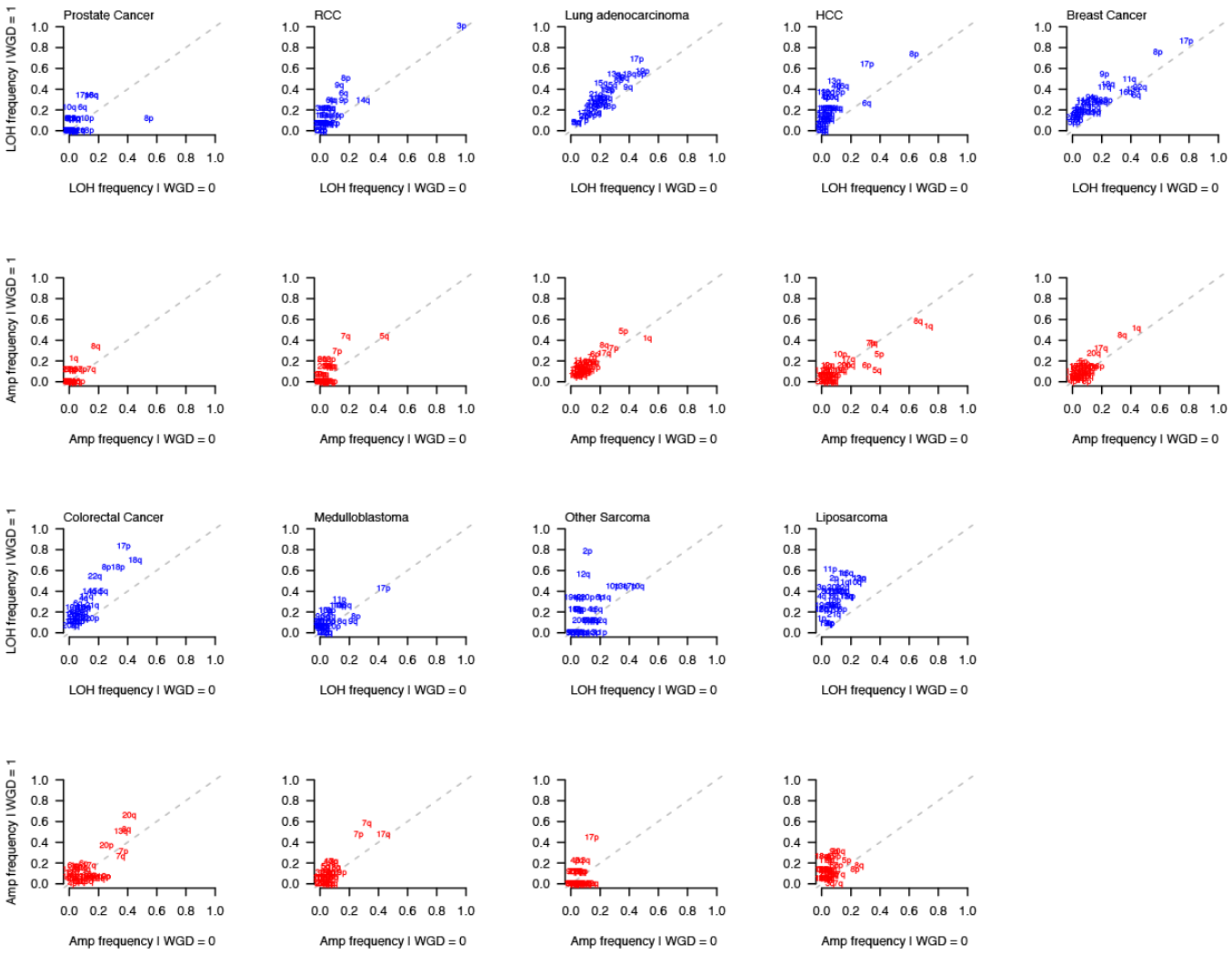
Nature Biotechnology: doi:10.1038/nbt.2203



Supplementary figure 9 | Distribution of estimated ploidy in various tumor types, and correspondence with absolute homologous copy-numbers

a,b, Data are shown for each indicated sample-type, with samples sorted by cancer-genome ploidy (left). Absolute copy-numbers of genomic segments for each sample are partitioned into low (**a**) and high (**b**) homologous copy values. Genome doubling events are discernable as inflections in the ploidy distributions of each cancer (**a; left**), corresponding with an increase in genome-wide homologous copy-number (**a,b**).

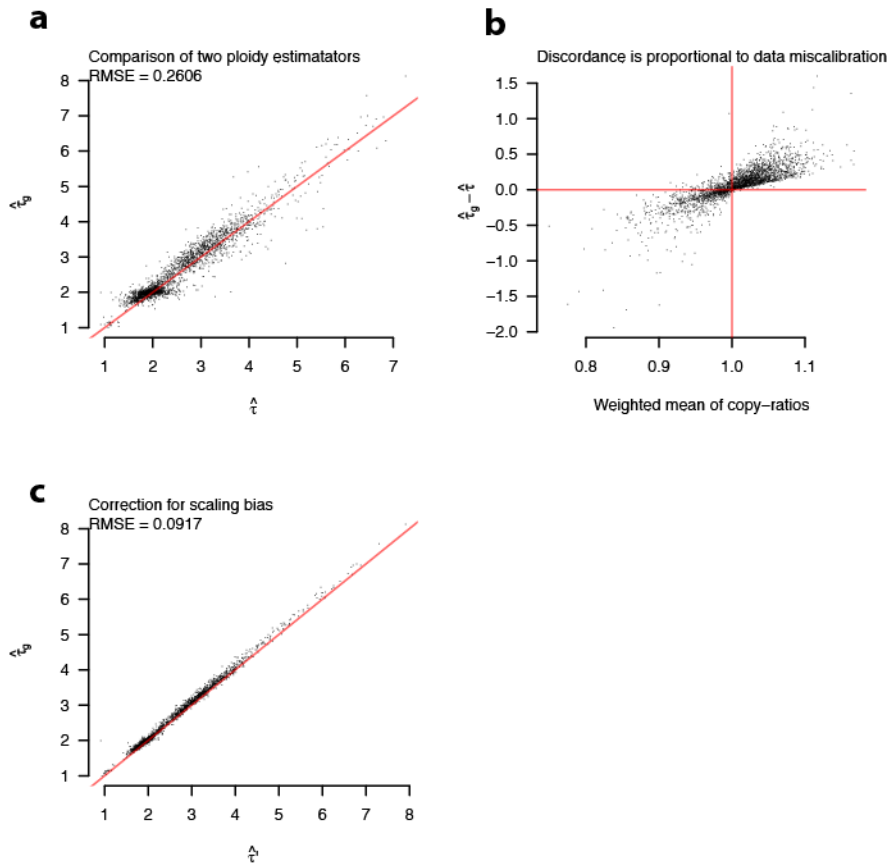
Nature Biotechnology: doi:10.1038/nbt.2203



Supplementary figure 10 | Incidence of chromosome arm-level SCNAs in non genome-doubled (WGD=0) vs. genome doubled (WGD=1) samples from 10 cancer types.

As in **Fig. 6d**. LOH (loss of heterozygosity) was defined as 0 allelic copies. Amplification was defined as > 1 allelic copy for samples with 0 genome doublings, and as > 2 allelic copies for those with 1 genome doubling. Calls were made based on the modal allelic copy numbers of each chromosome arm. Dashed lines indicate *y*=*x*.

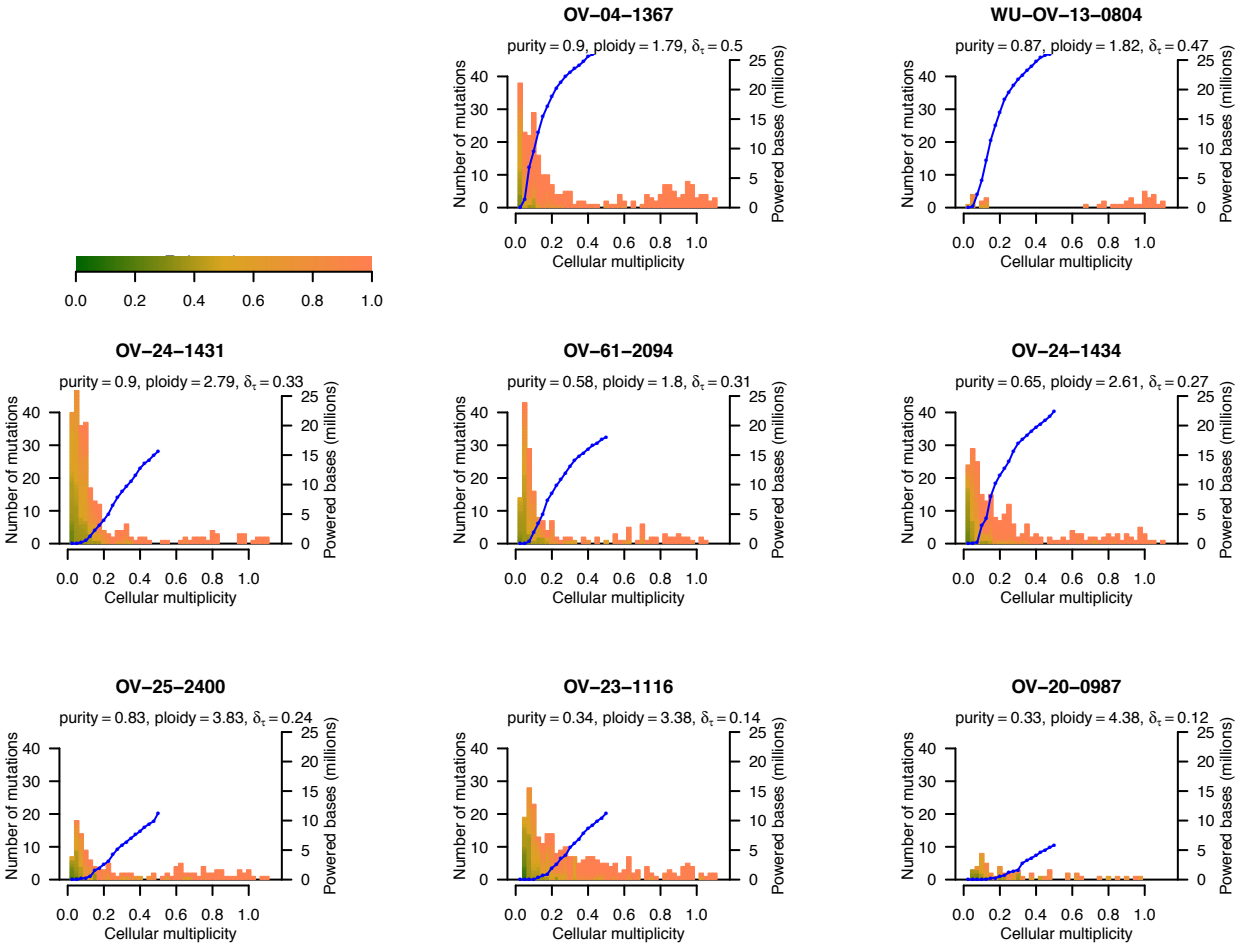
Nature Biotechnology: doi:10.1038/nbt.2203



Supplementary figure 11 | Robustness of purity / ploidy estimation to experimental variability

a, Two estimates of cancer-genome ploidy: $\hat{\tau}$ (eq. 1), and $\hat{\tau}_g$ (eq. 6). The estimator $\hat{\tau}$ is derived from the observed copy-ratio locations, whereas $\hat{\tau}_g$ is derived from the expectation of discrete integer copy-states, summed over the genome.
b, The discordance between the estimators in **a** (*y*-axis) is proportional to the mean of segmented copy-ratio data across the genome. In a perfectly calibrated experiment, this mean must equal 1, since a constant mass of DNA is input.
c, Correction of ploidy estimator $\hat{\tau}$ by $\hat{\tau} = (\hat{S} \cdot \hat{b} - 1)/(\hat{S} \cdot \hat{\delta}_\tau)$ (*x*-axis), where $\hat{\delta}_\tau = \hat{\alpha} / \hat{D}$, $\hat{b} = 2(1-\hat{\alpha}) / \hat{D}$, with $\hat{\alpha}$ and $\hat{D}$ fit from the data (eqn. 1). This correction assumes that the data miscalibration is due to a *scale S,* which is estimated for each sample as 1 / the weighted mean of segmented copy-ratios (**b**; *x*-axis). Since this correction removes most of the discordance between $\hat{\tau}$ and $\hat{\tau}_g$, we assume that the miscalibration in **b** is indeed primarily due to scaling variability. Noting that tumor purity $\alpha = \delta_\tau / (\delta_\tau + b)$, the implied correction for the purity estimate is: $\hat{\alpha}' = \hat{S} \cdot \hat{\delta}_\tau / (\hat{S} \cdot \hat{\delta}_\tau + \hat{S} \cdot \hat{b}) = \hat{\alpha}$. In other words, estimation of tumor purity is unaffected by moderate scaling variability in the data. This analysis also implies that centering of copy-ratio data, by rescaling (but not by translating) to mean = 1, is a valid preprocessing step (this was not performed). RMSE: root mean squared error

Nature Biotechnology: doi:10.1038/nbt.2203



Supplementary figure 12 | Power for detection of subclonal somatic mutations in 8 HGS-OvCa samples.

Histograms of cellular multiplicity point-estimates are shown for individual tumor samples, in order of decreasing δ_τ, which determines the average detection power in each sample (**Supplementary Fig. 7e**). Individual point mutations are colored according to their estimated detection power, calculated using their observed allelic fraction, coverage, tumor purity, and local absolute copy-number (Online Methods, Eq. 9). Coverage values and absolute copy-numbers were used to calculate theoretical detection power at each base for subclonal fractions between 0.025 and 0.5 (Online Methods, Eq. 9). The number of bases with power >= 0.8 are shown for each fraction (blue curve, right axis).

Nature Biotechnology: doi:10.1038/nbt.2203

Supplementary Note 1

ASCAT was performed using version 2.1 downloaded from http://heim.ifi.uio.no/bioinf/Projects/ASCAT/ on 01/29/2012. As is recommended for Affymetrix SNP data, the PennCNV-Affy protocol was used to convert probe-level data from CEL files to LRR and BAF values (http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2045149/). A batch-specific canonical genotype cluster file was created via PennCNV-Affy steps 1.1 through 1.3 (http://www.openbioinformatics.org/penncnv/penncnv_tutorial_affy_gw6.htm) using CEL files from all normal samples in that batch. Steps 1.1 and 1.2 were repeated on all CEL files, and step 1.4 was performed using the batch-specific canonical genotype cluster file to generate LRR and BAF values for the samples. Separate PennCNV-Affy protocols were used depending on the array type. The resulting LRR and BAF files were input into ASCAT as indicated on the software website.