Supplementary Methods

Calibration of the CN Markers

Our calibration of the CN markers relies on a two-step modeling approach ("copy-number inference") based on SNP6.0 array data of an X-dosage experiment performed on 46 samples from 5 cell lines with known variation of copy number of the X-chromosome from 1 to 5. We first calibrate each probeset on X by applying linear regression to the experimental data to fit the parameters β_{i0} and β_{i1} of the model

$$\text{(i)} \qquad I_i = \beta_{i0} + \beta_{i1} C_i.$$

Here variable I_i represents the intensity of probeset i, variable C_i represents the copy level at probeset i, and parameters β_{i0} and β_{i1} correspond respectively to the background signal intensity and the scale factor that specifies the change in intensity resulting from each added copy of DNA.

We extend the resulting calibration of the X probesets to a calibration for all probesets across the genome by modeling the background signal intensity and the scale factor as functions of local sequence features and median intensity across samples as follows:

$$\text{(ii)} \qquad \beta_{ik} = \alpha_{k0} + \alpha_{k1} GC_i + \alpha_{k2} FL^{(sty)}_i + \alpha_{k3} FL^{(nsp)}_i + \alpha_{k4} I_i^m + \alpha_{k5} (I_i^m)^2$$

for $k \in \{0,1\}$. The variable GC_i represents the GC content of the i^{th} probeset, $FL^{(sty)}_i$ and $FL^{(nsp)}_i$ represent the fragment lengths of the STY and NSP fragments of the i^{th} probeset respectively, and I_i^m represents the probeset median intensity across the samples. The linear dependence on GC content and fragment lengths and quadratic dependence on median intensity provides a good fit to our X-dosage array data. We apply linear regression again, this time to find the parameters, α_{kl} for $k \in \{0,1\}$ and $l \in \{0,1,2,3,4,5\}$ that best fit model (ii). The parameters α_{kl} are independent of the probeset and, for k = 1 or 2, this regression is performed collectively on the complete set of X-chromosome data, $\{\beta_{ik}, GC_i, FL^{(sty)}_i, FL^{(nsp)}_i, I_i^m\}_{i \in \eta_X}$ where η_X is the collection of indices for the X probesets. Each of the parameters for both models (i) and (ii) described above is computed once based on the results of the X-dosage experiment. Model (ii) is then used to predict the background and scale factor across the genome for each new batch of SNP6.0 array data. While GC content and fragment lengths do not vary with the batch, the median intensity must be computed separately for each batch.
The code and data used generate the model parameters are available in the supplementary file generate_snp6_cn_probe_calibration.tar.gz.

Quality Control

A key feature of our pipeline is automated quality control, which is integrated into the pipeline at various stages. Level 1 Quality Control (QC) is an assessment of the DNA quality of normal samples and is used to improve the accuracy of our calibration of SNP probesets. Normal samples must also pass additional quality control to be included in the CNV-detection analysis and the tangent noise model. This ensures that tumors mislabeled as normals and contaminated normal samples do not compromise our noise reduction algorithms. Level 2 and Level 3 QC provide noise assessments of copy number calls before and after segmentation respectively. Data from tumor samples that do not pass Levels 2 and 3 QC will not be input into GISTIC2.0.

Thresholds for Level 1 QC are based on recommendations from Affymetrix. Thresholds for all other QC measures were determined empirically based on the corresponding distributions for thousands of samples.

Level 1 QC: Level 1 QC consists of two components, the FQC all-call-rate and the Birdseed call-rate.

The FQC all-call-rate, which is computed using Affymetrix Power Tools software, is based on the intensities resulting from a small number of probesets whose configurations follow that of the Affymetrix 500K platform. While most probesets on the SNP6.0 platform consist of 3 identical probes for each allele with the SNP centered within each probe, these select probesets have approximately 10 probes per allele and position the SNP at a different location within each probe. The FQC-all-call-rate is a measure of self-consistency for these probesets and must exceed the threshold of 86 for a sample to pass.

The Birdseed call-rate is an assessment of Birdseed genotype calls for a normal sample. For each SNP probeset, genotyping is based on cluster assignments for the collection of normal samples from the same batch along with a set of historical prior arrays. If the intensities for the two alleles lie too far from the clusters determined by the other arrays, then no call will be made for the genotype. A call must be made for at least 95% of the SNP probesets in order for a sample to pass Level 1 QC.

Tumor Detector: Both our CNV-detection algorithm and our tangent noise model depend on the assumption that normal samples do not contain SCNAs. In order to weed out samples that bear evidence of SCNAs, we compute a disruption score for each normal sample, defined as the mean absolute moving average of $\log_2$ copy ratios across the genome. Samples whose disruption scores fall within the tail of the distribution are excluded from the noise model, $\mathcal{N}$, constructed during tangent normalization. The quantity of available data allows us to set the threshold at a conservative level, 0.073.

Level 2 QC: Level 2 QC is a noise assessment for each sample prior to segmentation. The acceptable noise level, measured as the genomic median of the absolute difference of $\log_2$ copy ratios for adjacent probes, is capped at 0.60.

Level 3 QC: Noisy samples at Level 2 can result in hypersegmentation at Level 3, with large numbers of adjacent segments whose copy-ratios differ by less than would be expected from absolute copy-number changes of one or more. Currently, samples with a segment count exceeding 2000 fail Level 3 QC. This threshold may be lowered as improvements to our pipeline upstream of segmentation have resulted in lower segment counts overall

GenePattern Module: CopyNumberInferencePipeline

The pipeline described here is available through the GenePattern module, CopyNumberInferencePipeline, at genepattern.broadinstitute.org/gp. The pipeline accepts a set of raw tumor and normal Affymetrix SNP6.0 CEL files along with a sample info file (SIF) as inputs and generates segmented copy-number calls for each sample using either the human reference genome HG18 or HG19, at the user's choosing. The user may also opt to have copy-number calls for the CNV probes included in the output, although these probes will be excluded from tangent normalization. Segmented copy-number calls are specified as $\log_2$ copy-number ratios, and normalized so that each sample appears diploid.

Supplementary Figure S1 displays the flow of data through the pipeline and indicates the key output files of the pipeline modules. The contents of the required input files and key output files are summarized below.

Further documentation is available at
broadinstitute.org/cancer/cga/copynumber_pipeline .

Required inputs

For the best calibration results, CEL files submitted together should be associated to chips that were processed within the same PCR batch and include 20 or more diploid normals. A minimum of 10 normals is required. Normals need not be matched pairs for the tumor samples. Tissue-adjacent normals and cell lines may lead to misleading results. Including more normal samples within the batch will improve the noise reduction during tangent normalization.

The SIF is a tab-delimited table with the following per-sample information:

> Array: CEL file name, but without the .CEL extension.
> Gender: M, F, or NoCall.
> Tumor/Normal: Tumor or Normal.
> Birdseed_normals: <blank> or Y. Y is for indicating that the given normal is believed to be diploid and the array measurments have low noise.

Matched_Normal_Array: not currently used.

Output Files

SNPFileCreator:

*.med1000.invset_medpolish.snp: (TCGA Level 2) One value is output per probe. Values are centered on 1000 for each sample, and are in linear space.

Calibration:

*.birdseedCalls.txt: (TCGA Level 2) The genotype per SNP probe: 0=AA, 1=AB, 2=BB, where A is the major allele.

*.birdseedConfs.txt: (TCGA Level 2) Confidence of call for each SNP probe.

*.med1000.invset_medpolish.pip3avg.log_mdQUAD.byallele.txt: (TCGA Level 2) One line is output per SNP probeset, and each line has the copy-number calls for the A and B SNP probesets at that location. The values are in linear space, and are approximately diploid. A fixed set of poorly performing SNP probesets have been dropped.

*.med100.invset_medpolish.pip3avg.log_mdQUAD.txt: (TCGA Level 2) One line is output per SNP or CN probeset. The values are in linear space, and are approximately diploid. For SNP probesets, the values for the two alleles are summed to represent total copy number. A fixed set of poorly performing SNP and CN probesets have been dropped.

Tangent normalization:

*_posttangent_woCNV.txt: (TCGA Level 2) One line is output per SNP or CN probeset. The values are in linear space, and are centered on diploid. CNV probes and the Y chromosome are excluded. A fixed set of poorly performing SNP and CN probesets have been dropped, and probes identified as outliers in the given sample have been set to NaN. These outputs are mapped to the HG19 reference genome.

*_posttangent_wCNV.txt: (TCGA Level 2) Similar to *_posttangent_woCNV.txt, but includes (unnormalized) values for CNV probes and Y.

*.early_gistic_prep_output._suspect_normals.txt: List of submitted normal arrays excluded from the reference plane during tangent normalization due to failure to pass Tumor Detector QC. (See Quality Control above.)

Circular Binary segmentation:

*_woCNV_hg18.seg, *_woCNV.hg19.seg, *_wCNV_hg18.seg, *_wCNV_hg19.seg: (TCGA Level 3) One line per segment. The values are in log space, centered on 0.

Tangent Normalization Overview

The calibrated copy number exhibits systematic variation related to the genomic location, the sample, and the batch. This is most easily observed by comparing diploid regions of various normal samples (i.e. outside of CNV regions), or by comparing replicates of one sample (Supplementary Figure S2). Similar variation is also observed in tumor samples. This systematic variation can be mistaken for many recurrent SCNAs appearing in many samples. Since we wish to focus on SCNAs, we want to remove forms of variation that we also see in normals.

Tangent Normalization models this systematic variation as a linear combination of a large panel of diploid normals from many batches, using log space. Ordinary Least Squares (OLS) is applied to each sample to determine this linear combination. This weighted sum of normals lies within a subspace (the "tangent plane") of the space containing all possible copy-number profiles; the weighted sum used for a particular tumor is that tumor's projection into this subspace. Further details are provided in Supplementary Methods.
The modeled systematic variation is subtracted from each sample, and then the sample is offset to center the probes on a nominal diploid copy number.

Tangent Normalization is applied to both tumor and normal samples, taking care that the normal is not itself included in the panel of normals. Common regions of germline CNVs are excluded during this process, and subsequently reinserted.

Recurrent CNVs within normals are identified through an adapted, abbreviated version of the overall pipeline. The data for each normal sample are normalized and segmented using tangent normalization and Circular Binary Segmentation (Venkatraman and Olshen, 2007). GISTIC G-scores (Mermel et al., 2011) representing the summed level of amplifications or deletions across samples are then computed at each locus. Probesets within the tails of the G-score distribution (representing approximately 15% of all probesets) are identified as within recurrent CNVs and removed from further analyses.

We include male and female samples in our panel of normals. To account for the variation in the number of copies of X, we include in our panel a theoretical normal, the 'ideal man,' with copy-number precisely two throughout the autosomes and one throughout the X chromosome. The resulting normalized data will adjust the copy-number profile of X for any sample to a mean value of ~2. Use of tangent normalization in this way discounts whole-chromosome (but not focal) changes in X. Use of gender-matched normals can enable recovery of these SCNAs.

The performance of tangent normalization depends on adequate representation of the noise profiles within the model to characterize all tumor samples being analyzed. Potential sources of systematic noise include variability of conditions during PCR amplification, cross-hybridization, and variability of GC-content across the genome. Therefore, we populate our noise model using normal samples that reflect all of the experimental conditions that generated our data. The noise profile for a tumor is usually best matched by noise profiles of normal samples that were processed in the same batch and shared experimental conditions. However, some tumors are found to have noise profiles that are best matched by normal samples from other batches, and we find that signal-to-noise ratios improve due to increasing noise reduction as the size of the pool of normal samples increases (Supplementary Figure S7c). Noise reduction can improve with a larger normal pool even for those tumors with noise profiles closely resembling the normal samples of the same batch (Supplementary Figure S6). Therefore, we construct a reference plane for tangent from the entire collection of TCGA blood-normal samples that pass levels 1 and 2 quality control. (See Quality Control below.) The current plane is based on 3154 TCGA blood-normal samples that pass quality control.

Tangent Normalization Algorithm Details

For $i \in \{1,2,3,\ldots n_N\}$ where n_N is the number of normal samples, the i^{th} normal sample is represented as a vector, N_i, of $\log_2$ copy-ratio intensities in genomic order, with each coordinate corresponding to one of the non-CNV probes. We use $\log_2$ copy ratios because we have found that this representation works well for noise reduction, suggesting that much of the observed noise is multiplicative. The noise space, $\mathcal{N}$, is defined as the $(n_N - 1)$-dimensional plane containing the vectors $\{N_1, N_2, N_3, \ldots, N_{n_N}\}$. Note that $n_N - 1 << M$, where M equals the dimension of the ambient ($\log_2$ copy-ratio) coordinate space or equivalently, the number of markers not excluded as poor quality or potential CNVs. Similarly, for $j \in \{1,2,3,\ldots n_T\}$ and n_T equal to the number of tumor samples, T_j represents the j^{th} tumor sample in the same format as N_i. The noise profile for a tumor, T_j, is determined as the point in $\mathcal{N}$ that is closest to T_j using a Euclidean metric, i.e. the projection, $p(T_j)$, of T_j on $\mathcal{N}$. The resulting normalization of T_j is set to the residual, $T_j - p(T_j)$.

The projection $p(T_j)$ can be computed directly using standard linear algebra techniques. A rigid transformation of Euclidean marker space prior to normalization does not alter the resulting normalization of T_j. In particular, an appropriate translation of Euclidean space ensures that $\mathcal{N}$ passes through the origin and forms a vector subspace of Euclidean space. It follows that

(iii) $$p(T_j) = N * N_{pi} * T_j$$

after translation, where N is the array whose columns correspond to $n_N - 1$ normal samples that span $\mathcal{N}$ and N_{pi} is the pseudoinverse of N.

We include both male and female normal samples, which differ in the number of copies of X. The inclusion of the X chromosome in tangent normalization requires special treatment to ensure that the distance from a tumor to a normal reflects noise differences, without being artificially inflated due to gender difference. Additionally, we must take into account that the normalization, T_j - p(T_j), of T_j could potentially alter the apparent chromosomal copy number of X, due to the fact that p(T_j) is a weighted average of copy ratios from both male and female samples. To address these issues, we include in our reference plane a theoretical normal, the 'ideal man,' with copy-number precisely two throughout the autosomes and one throughout the X chromosome. Tangent normalization against this expanded collection of normal samples will adjust the copy-profile of X for any sample, regardless of gender, to a mean level with ~2 copies of X. The ensuing analysis can detect focal SCNAs within X, but discounts whole-chromosome changes of X. Currently, the Y chromosome is excluded from tangent normalization. Use of gender-matched normals may enable recovery of whole-chromosome SCNAs involving X.

The large number of reference normal samples presents computational challenges as the projection matrix depends on the computation of the pseudo-inverse of an M x n_N matrix (~$1.5 \times 10^6 \times 3000$). To address this issue, we mimic Gram-Schmidt orthogonalization, but on a blockwise level, and decompose the reference plane into orthogonal blocks so that the projection, $p(T_j)$, can be computed on a block-by-block basis with only one block in memory at a time. Each block of data represents approximately 250 normal samples, typically from multiple batches. The orthogonalization process replaces the i^{th} block of normal data by its tangent normalization against blocks 1 through $i-1$. When a new batch is processed, an additional block is added using the normal samples from the batch at hand, which are themselves first normalized against the reference normal samples. We are somewhat less stringent in our quality control for the current batch to allow tissue normal samples as well as blood normal in order to ensure adequate representation of the noise profiles for the batch at hand. Our current reference plane consists of 13 blocks; we periodically expand our reference plane as additional TCGA data becomes available.

Tangent Analysis on TCGA Glioblastoma Data

The glioblastoma analysis was based on 497 TCGA tumors and 451 TCGA normal samples that were processed with these tumors. This is a reduced collection of normal samples compared to our standard analyses involving over 3000 normals. We used this reduced set to enable fair comparisons to other normalization techniques (Supplementary Figures S4, S5, and S7a-b). The X and Y chromosomes were excluded from these normalizations so that differences in the handling of the sex chromosomes would not contribute to the comparisons (this is not required for tangent normalization). The CNV probes were also excluded for the same reason. The preprocessing prior to normalization was identical for all three normalization

techniques. The representation of each sample as a vector of $\log_2$ copy-ratio intensities is identical to that described in the section Tangent Normalization above.

For matched normals, the normalization of a tumor consists of subtracting from its log2 ratios those of its matched normal. Only the 386 gliblastoma tumors with a matched normal could be normalized, demonstrating an additional limitation of this approach as compared to tangent or five nearest normals. For five nearest normals, normalization consists of subtracting the mean of the five normals closest to the tumor based on a Euclidean metric.

Supplementary Figures S4 and S7a-b demonstrate that the signal is preserved with all three normalization techniques, but only tangent normalization consistently reduces noise, thereby increasing the signal-to-noise ratio. The impact of five nearest normals on noise is quite small while normalization by matched normals tends to increase noise and decrease the signal-to-noise ratio.

The glioblastoma analyses specific to tangent were performed with the larger reference plane of 3154 TCGA normals obtained from patients with multiple cancer types. This reference plane is also used by the tangent GenePattern module. In order to investigate the effect of the size of the normal reference pool on noise reduction, data was saved after each step in the block normalization procedure. Supplementary Figure S7c demonstrates that each additional block of normals added to the reference plane further reduces noise, although the greatest impact was achieved following the first 4-5 blocks, which collectively contain 1000-1250 samples. For the comparison in Supplementary Figure 6, tangent was also performed on each tumor using only the normal samples from the same batch. Whether the entire reference plane is used or only the normal samples in the same batch, Supplementary Figure 6 shows that tangent reduces noise as the post-normalization to pre-normalization noise ratio is consistently below 1. However, this scatter plot of these resulting noise ratios following tangent with the entire reference plane vs. tangent with batch normals reveals greater noise reduction for almost every tumor sample when the entire reference plane is included.

Tangent Analysis on HCC1143 Replicate Data

We further examined systematic noise within and across batches by way of HCC1143 blood normal samples that were processed on 118 arrays across 110 batches and HCC1143 breast tumor samples that were processed on 138 arrays across 128 batches. Systematic noise tends to produce consistent patterns of variation in the data across samples, while random noise does not. A comparison of $\log_2$ copy-number ratios for the normal replicates prior to normalization revealed several distinct patterns of copy-number variation, each of which was evident for multiple replicates across many batches (Supplementary Figure S2a). Further, samples with similar patterns on one chromosome tended to be similar across the genome (data not shown). We then examined 20 megabases of chromosome 1 for which the copy number of the HCC1143 tumor cell line is constant. Variations in the

pre-normalized $\log_2$ copy-number ratios across this genomic region for the tumors and the diploid normal were the result of noise. Supplementary Figures S2b-c demonstrate that tangent normalization substantially eliminates this noise from the tumor samples.

We then examined the normal replicates that were processed in the 8 batches containing multiple HCC1143 normal samples. Sample replicate pairs with correlated noise were found both within batches and across batches, as were pairs with uncorrelated noise (data not shown).

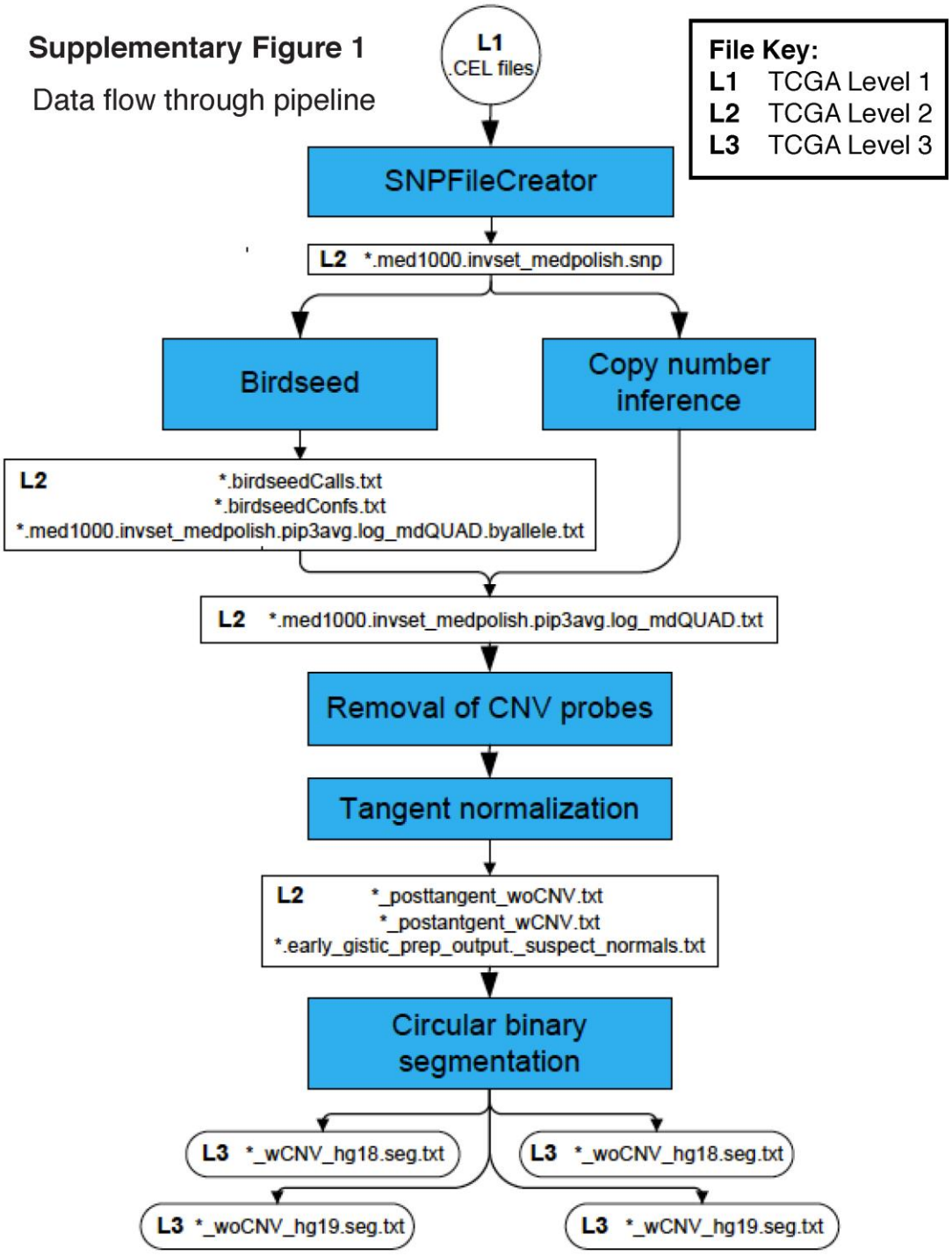
Overall, these results provide further evidence that a noise model built from normal samples processed across many batches will best represent systematic noise and facilitate noise reduction.

Tangent Normalization Discussion

The use of tangent normalization leads to substantially improved signal-to-noise ratios relative to prior methods (Supplementary Figure S7a). We estimate signal as the standard deviation of median signal intensities among all chromosome arms, and noise as the median absolute difference between $\log_2$ copy-number ratios of adjacent probes. The improvements in signal-to-noise ratios are the result of reduced noise (Supplementary Figure S7b, Supplementary Figure S3); signal remains essentially unchanged (Supplementary Figure S4). As a result, segmented copy-number profiles generated after tangent normalization exhibit less hyper-segmentation and systematic biases than profiles generated using other methods (Supplementary Figure S5).

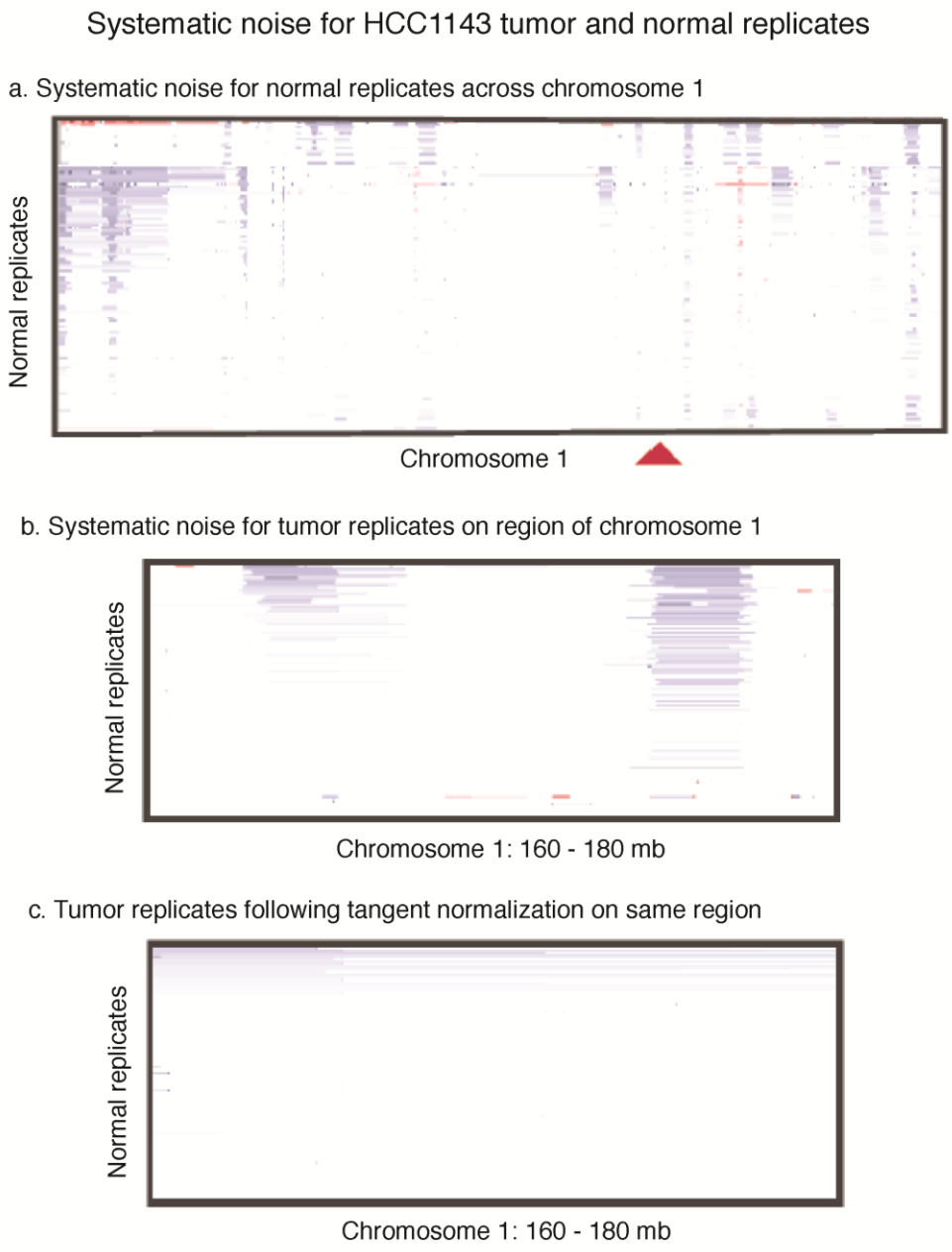
A tumor's noise profile is usually matched best by normal samples from the same batch, but some tumors exhibit noise profiles that are best matched by normals from other batches (Supplementary Figures S2, S6). As a result, signal-to-noise ratios improve as we expand the pool of normal samples (Supplementary Figure S7c). For TCGA, we use all TCGA normal samples obtained from blood that pass levels 1 and 2 of quality control.

Although this pipeline was developed for use with Affymetrix SNP array data, it can be extended to other SNP and comparative genomic hybridization (CGH) platforms. Moreover, we have had success applying similar concepts to Whole Exome and Whole Genome next-gen sequencing data. Further improvements to signal-to-noise ratios are likely to be obtained through use of such data, from improved methods to calibrate and normalize those data, and from algorithms that determine differences in absolute rather than relative copy-numbers (Carter *et al.*, 2012; Van Loo and Nordgard, 2010).



Supplementary Figure S1. Data flow through pipeline. Schematic view of pipeline as exhibited in Figure 1, with key output files for pipeline modules and assigned TCGA levels included.

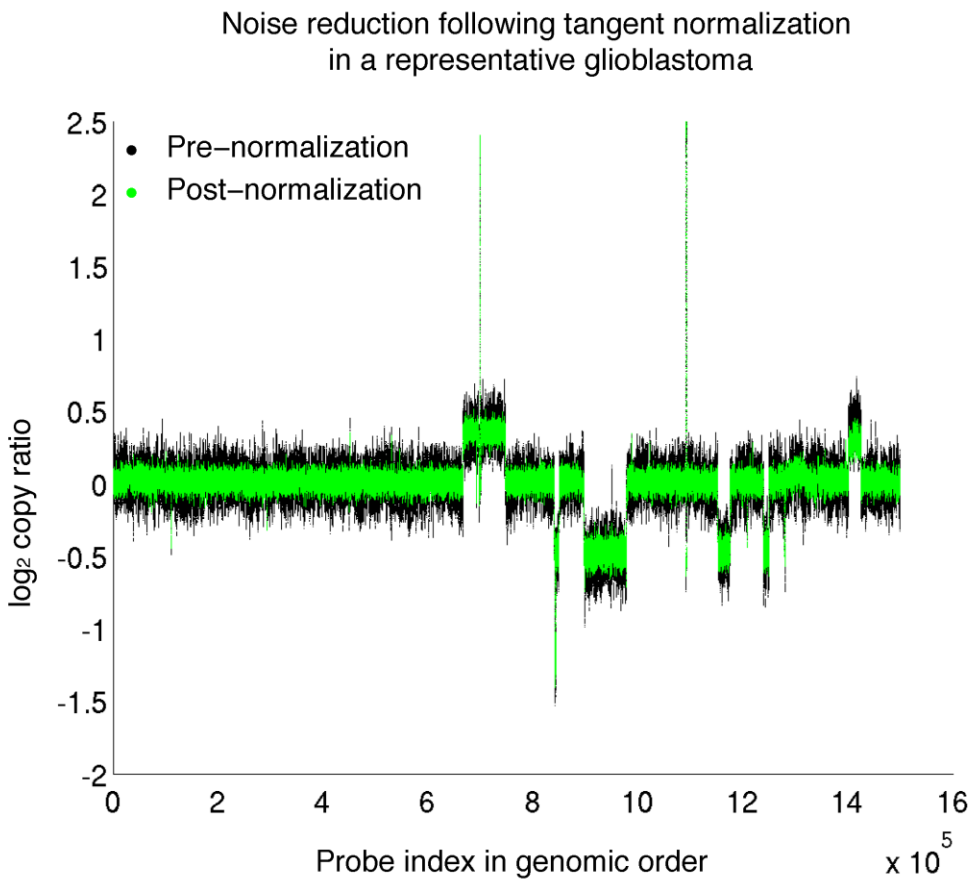
Supplementary Figure 2



Supplementary Figure S2. Systematic noise for HCC1143 tumor and normal replicates. Segmented pre-normalized log2 copy-number ratios (low and high ratios indicated by blue and red, respectively) on (a) replicates of DNA from the HCC1143BL immortalized lymphocyte (non-cancer) line across 110 batches, chromosome 1, (b) replicates of the HCC1143 tumor cell line across 128 batches, chromosome 1, 160-180 megabases. Comparisons between normal replicates

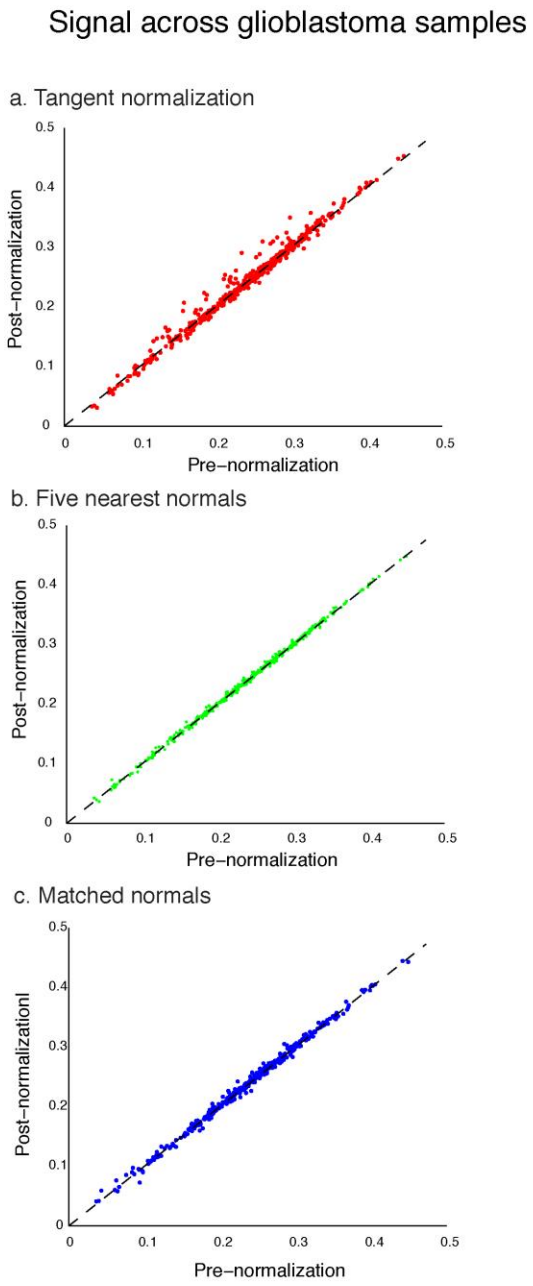
indicate correlated variations across genomic regions, which are also observed in tumor replicates. As these variations are observed in the same DNA, they represent artifact. (c) Segmented tangent-normalized data for tumor replicates, chromosome 1, 160-180 megabases. The systematic artifacts present in (a)-(b) are no longer observed. Figures were generated with IGV.

Supplementary Figure 3



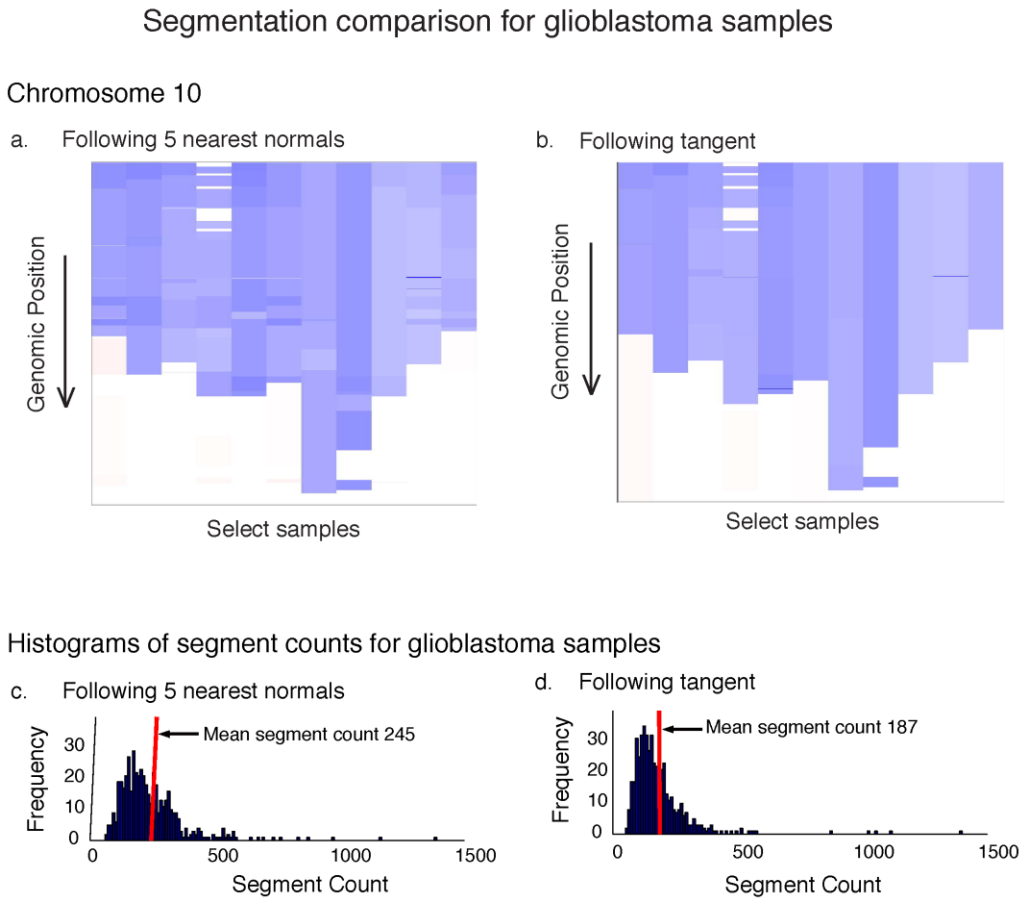
Supplementary Figure S3. Noise reduction following tangent normalization in a representative glioblastoma tumor. 100-marker moving average of log2 copy-ratios for a representative glioblastoma sample across the autosomes before (black) and after (green) tangent normalization. A moving average is employed for visualization purposes only.

Supplementary Figure 4



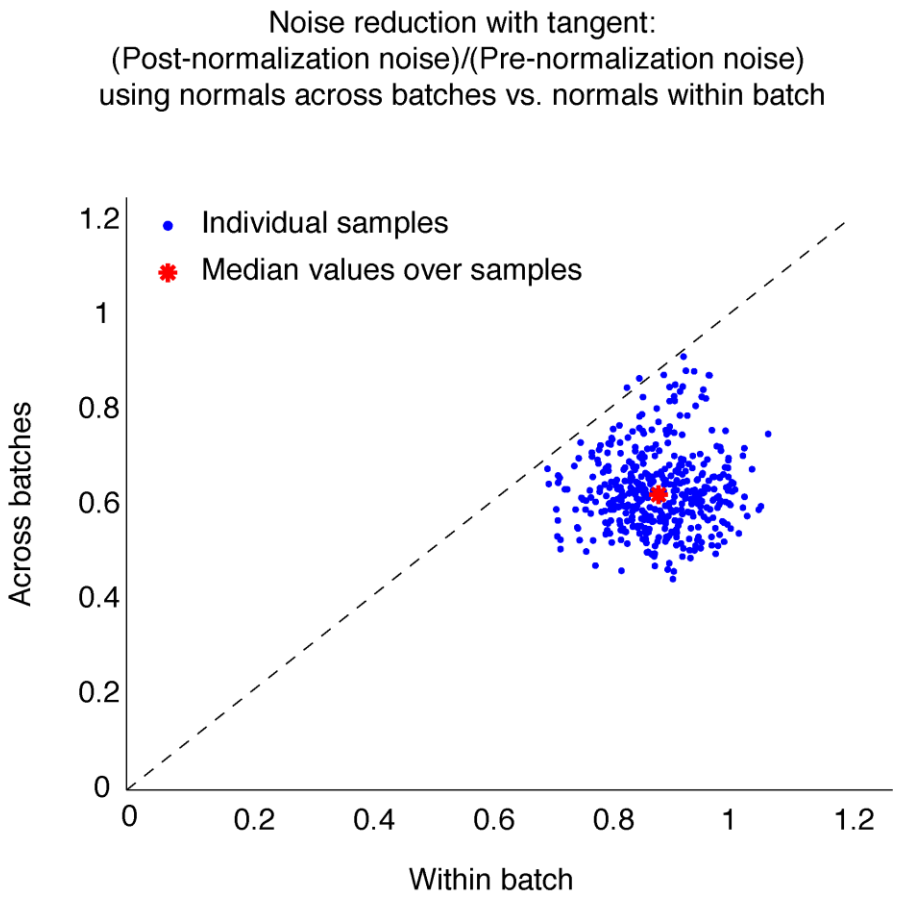
Supplementary Figure S4. Signal across 497 TCGA glioblastoma tumors. Scatter plot of post-normalization vs. pre-normalization signal for three normalization methods, (a) tangent (red), (b) five nearest normals (green) and (c) matched normals (blue), demonstrates that the signal level is largely unchanged by normalization for each of these three methods.

Supplementary Figure 5



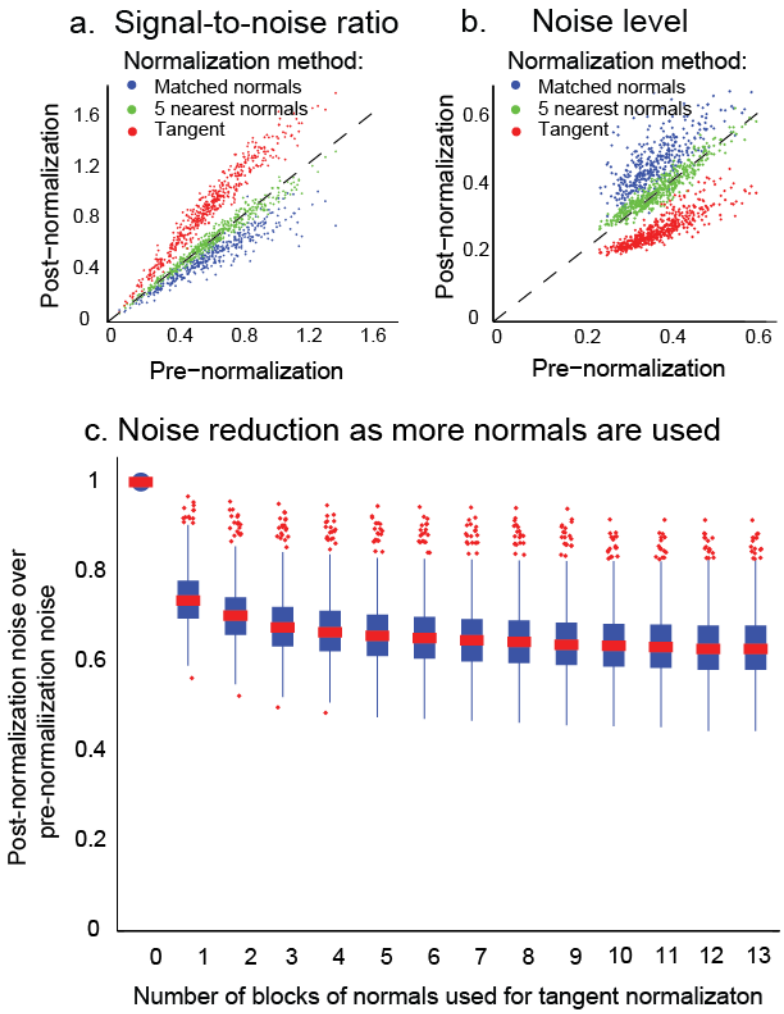
Supplementary Figure S5. Segmentation comparison for glioblastoma samples. (a-b) Post-segmentation results for selected glioblastoma samples following (a) 5NN and (b) tangent. White is copy-neutral. Blue indicates a deletion with intensity of color increasing as the copy-number decreases. Samples are displayed in the same order in both panels. Less hypersegmentation is observed when CBS is applied to tangent-normalized data. Figures were generated with the Broad Integrative Genomics Viewer (IGV) (Robinson *et al.*, 2011; Thorvaldsdottir *et al.*, 2012) (c-d) Histograms of segment counts for 497 TCGA glioblastoma tumors when CBS follows (c) 5NN and (d) tangent. Decreased segments counts for tangent normalized data is consistent with decreased hypersegmentation.

Supplementary Figure 6



Supplementary Figure S6. Noise reduction with tangent using normal samples across batches vs. normal samples within batch. Noise ratio (post-normalization over pre-normalization noise) for glioblastoma samples following tangent normalization using our pipeline's reference plane vs. tangent normalization using only the normal samples processed in the same batch as a tumor. Almost all samples lie below the diagonal (x = y) indicating that there is greater noise reduction with the full reference plane.

Supplementary Figure S7



Supplementary Figure S7: Tangent Normalization of 497 TCGA glioblastomas. (a,b) Comparison of tangent to two other normalization techniques. (c) Box plot of post-normalization noise as a fraction of pre-normalization noise, following tangent normalization with each block of approximately 250 normals in sequence. Block 0 corresponds to pre-normalization.