OXFORD

Genome analysis

Tangent normalization for somatic copy-number inference in cancer genome analysis

Galen F. Gao [1,†]**, Coyin Oh** [1,2,3,†]**, Gordon Saksena**[1,†]**, Davy Deng**[1,3,4]**,
Lindsay C. Westlake**[1]**, Barbara A. Hill**[1]**, Michael Reich**[1,5]**, Steven E. Schumacher**[1,3]**,
Ashton C. Berger**[1,3]**, Scott L. Carter**[1,2,3]**, Andrew D. Cherniack**[1]**,
Matthew Meyerson**[1,3,6,†]**, Barbara Tabak**[1,3,†]**, Rameen Beroukhim**[1,3,7,*,†] **and
Gad Getz** [1,8,9,*,†]

[1]Cancer Program, Broad Institute of MIT and Harvard, Cambridge, MA, USA, [2]Harvard-MIT Division of Health Sciences and Technology, Harvard Medical School, Boston, MA, USA, [3]Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA, USA, [4]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Harvard University, Boston, MA, USA, [5]Department of Medicine, Division of Medical Genetics, University of California, San Diego, La, Jolla, CA, USA, [6]Department of Genetics, Harvard Medical School, Boston, MA, USA, [7]Department of Medicine, Harvard Medical School, Boston, MA, USA, [8]Department of Pathology, Harvard Medical School, Boston, MA, USA and [9]Department of Pathology, Massachusetts General Hospital, Boston, MA, USA

*To whom correspondence should be addressed.

[†]The authors wish it to be known that, in their opinion, the first three authors and last four authors should be regarded as Joint Authors.

Associate Editor: Russell Schwartz

Abstract

Motivation: Somatic copy-number alterations (SCNAs) play an important role in cancer development. Systematic noise in sequencing and array data present a significant challenge to the inference of SCNAs for cancer genome analyses. As part of The Cancer Genome Atlas, the Broad Institute Genome Characterization Center developed the Tangent normalization method to generate copy-number profiles using data from single-nucleotide polymorphism (SNP) arrays and whole-exome sequencing (WES) technologies for over 10 000 pairs of tumors and matched normal samples. Here, we describe the Tangent method, which uses a unique linear combination of normal samples as a reference for each tumor sample, to subtract systematic errors that vary across samples. We also describe a modification of Tangent, called Pseudo-Tangent, which enables denoising through comparisons between tumor profiles when few normal samples are available.

Results: Tangent normalization substantially increases signal-to-noise ratios (SNRs) compared to conventional normalization methods in both SNP array and WES analyses. Tangent and Pseudo-Tangent normalizations improve the SNR by reducing noise with minimal effect on signal and exceed the contribution of other steps in the analysis such as choice of segmentation algorithm. Tangent and Pseudo-Tangent are broadly applicable and enable more accurate inference of SCNAs from DNA sequencing and array data.

Availability and implementation: Tangent is available at https://github.com/broadinstitute/tangent and as a Docker image (https://hub.docker.com/r/broadinstitute/tangent). Tangent is also the normalization method for the copy-number pipeline in Genome Analysis Toolkit 4 (GATK4).

Contact: rameen@broadinstitute.org or gadgetz@broadinstitute.org

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1. Introduction

Cancer often arises from the accumulation of somatic alterations in the genome, including point mutations, structural rearrangements and copy-number alterations (Weir *et al.*, 2004). Somatic copy-number alterations (SCNAs) can have significant impact in activating onco-genes or inactivating tumor suppressor genes to drive the development of cancer (Beroukhim *et al.*, 2010; Zack *et al.*, 2013). In 2006, the NCI and NHGRI launched The Cancer Genome Atlas (TCGA) project to

comprehensively characterize the genomic and molecular features of different cancer types (The Cancer Genome Atlas Network *et al.,* 2013). TCGA collected samples from more than 11 000 cancer patients across 33 tumor types. The use of next-generation sequencing (NGS) and high-resolution microarrays allowed us to finely characterize SCNAs in cancer genomes and facilitate the discovery of novel genes that drive cancer (Korn *et al.,* 2008; The Cancer Genome Atlas Network *et al.,* 2013; Zack *et al.,* 2013).

Standard approaches to detect somatic copy-number profiles involve determining DNA content at various sites across the genome in tumor samples, and comparing this tumor DNA content to that in normal samples. For example, in array comparative genomic hybridization (CGH) or single-nucleotide polymorphism (SNP) arrays, signal intensities of DNA probes for various genomic loci scale with sample DNA content at each locus (LaFramboise, 2009). Similarly, high-throughput sequencing enables determination of coverage levels at loci across the genome, also reflecting sample DNA content (Yoon *et al.,* 2009). Detection of somatic copy-number alterations (SCNAs) typically relies on determining the ratios between DNA content in tumor versus normal samples across these loci, which aims to normalize the different affinities (either of probes or sequencing) associated with each locus.

Such analyses can be confounded by at least three sources of noise. First, stochastic variations result in random deviations between measurements of DNA content and true DNA content. This can be overcome by averaging measurements across adjacent loci (e.g. using segmentation algorithms; Venkatraman and Olshen, 2007) or by sequencing to greater average depth. Second, germline copy-number variations (CNVs) can be misinterpreted as SCNAs. This can be overcome by comparing tumor DNA to normal DNA from the same patient, or by masking common CNVs. Third, systematic errors can result from subtle differences between experimental conditions that were applied when generating sequencing or microarray data from tumors and their normal controls, which can affect the locus-specific affinities.

Despite rapid advancement in sequencing technologies and improvements in copy-number tools that attempt to combat systematic noise, such as Control-FREEC, ExomeCNV, VarScan2 and CNVkit (Boeva *et al.,* 2012; Koboldt *et al.,* 2012; Rieber *et al.,* 2017; Sathirapongsasuti *et al.,* 2011; Talevich *et al.,* 2016; Zare, 2017), filtering out systematic noise present in NGS and microarray data remains a significant challenge. Many of these tools use similar approaches to reduce systematic noise, either with matched case–control samples or with GC correction (Zhao *et al.,* 2013). While matched normal samples can sometimes approximate their tumors' noise profiles, they are not always available, and during the sequencing process, many of them may be processed under conditions different from those of their corresponding tumors and therefore may not have similar noise profiles. And while GC-content bias constitutes a large component of systematic noise, GC correction does not target all sources of noise present in copy-number data. Other potential sources of systematic noise include mappability biases across the genome and variability in experimental conditions during PCR amplification, cross-hybridization, or sample and library preparation. Thus, currently available tools do not adequately address these gaps in copy-number analysis.

Here, we present Tangent, a copy-number inference pipeline that aims to address these gaps by constructing noise profiles using a subset of normal samples to target all potential sources of systematic noise. The normal samples used for Tangent will ideally have been processed using the same experimental conditions as the tumor samples, but do not have to be from the same patients as the tumors. Our pipeline begins with either a whole-exome sequencing (WES) BAM file or raw probe-level intensity data and concludes with segmented copy-number calls, processing data with special attention to noise reduction, artifact removal, germline CNV removal and quality control. The Tangent pipeline can be applied to both WES and Affymetrix SNP Array 6.0, both of which have been the basis for data analyses in TCGA. Tangent can also be extended to other sequencing platforms. Additionally, we describe Pseudo-Tangent, an approach that uses signal-subtracted tumor data to augment

standard normal data in the Tangent pipeline. Pseudo-Tangent is particularly useful when there is a limited number of normal samples that can be used for denoising. Tangent is the basis for copy-number normalization in the GATK4 CNV workflow available within Genome Analysis Toolkit 4 (GATK4; McKenna *et al.,* 2010) and is available through Github and Docker.

2. Materials and methods

2.1 Generation of raw coverage data
As input to Tangent, we generated raw coverage data from either Affymetrix SNP arrays or from WES. For SNP arrays, the procedure to generate raw coverage data is described in Supplementary Methods. This procedure includes an initial quantile-normalization pre-processing step that ensures the total signal across all SNP loci is uniform between all samples in each cohort analyzed. For WES data, we used the GATK DepthOfCoverage tool on input .bam files to assess coverage from the input .bam file (Depristo *et al.,* 2011). DepthOfCoverage outputs values for a set of genomic loci ('intervals') representing the hybrid capture targets. Interval files are available in Firecloud/Terra from the broad-firecloud-tutorials/ Broad_MutationCalling_QC_Workflow_BestPractice workspace. Flow charts for each type of input data are presented in Supplementary Figure S1A and B.

2.2 Tangent normalization
Tangent assumes that systematic noise, after log-transformation, is distributed according to a similar additive pattern in tumor samples as in normal samples. (We use $\log_2$ copy ratios because we have found that this representation works well for noise reduction [data not shown], suggesting that much of the observed noise is multiplicative.) We model the space of $\log_2$ copy ratios as:

$$N \oplus N^{\perp},$$

where the noise space, N, is determined empirically based on the collection of normal samples, as described in detail below, and the signal space is its orthogonal complement, $N^{\perp}$. As a consequence of linear algebra:

1. Each tumor, T_j can be represented uniquely as

$$T_j = noise(T_j) + signal(T_j)$$

where $noise(T_j)$ is a vector in N, and $signal(T_j)$ is a vector in $N^{\perp}$.

2. The noise vector, $noise(T_j)$, equals the projection of T_j to N and can be explicitly computed by way of the pseudoinverse, as described below. (Note that, in principle, the variation across normal samples may not be entirely orthogonal to the signal of somatic copy-number alterations in a specific tumor—raising concerns that the projection of T_j to N will include true signal. However, in practice, if N is of a much lower dimension than the overall space, such 'overfitting' is likely to be minimal and the signal in the tumor will not be substantially reduced. We evaluate the extent to which the signal is reduced in practice in Section 3.2 below.)

3. The signal vector, $signal(T_j)$, equals the residual, $T_j - noise(T_j)$.

In summary, we define the noise profile of a tumor to be $noise(T_j)$, the projection of the tumor to a lower-dimensional subspace spanned by the coverage profiles of the normal samples. This is individually calculated for each tumor using data from normal samples. To minimize systematic noise, we then subtract that projection from the raw $\log_2$-transformed copy-number profile of the tumor. This difference, the signal, is the Tangent-normalized coverage profile for that tumor.

In detail, for $i \in \{1, 2, 3, \dots n_N\}$ where n_N is the number of normal samples, the ith 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. The noise space, 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 $n_N - 1 \ll M$, 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 $j \in \{1, 2, 3, \ldots, n_T\}$, where n_T is the number of tumor samples, T_j represents the jth tumor sample in the same format as N_i. A constructed normal profile that closely matches the noise profile for a tumor T_j is determined as the point in N that is closest to T_j using a Euclidean metric, i.e. the projection, $p(T_j)$, of T_j on 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 the Euclidean space ensures that N passes through the origin and forms a vector subspace of Euclidean space, in which the normal vectors now reflect the deviation from the typical normal (i.e. the noise). After projection to N, the noise profile for each sample can be expressed as a linear combination of $n_N - 1$ translated normal vectors,

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

after translation, where N is the array whose columns correspond to $n_N - 1$ normal samples that span N and N_{pi} is the pseudoinverse of N. This noise profile, that is closest to the tumor, is then subtracted from the tumor signal to obtain a 'clean' signal.

The sex chromosomes have traditionally presented a challenge in the normalization step of copy-number calling. Theoretically, we can address this issue by having separate reference planes for male and female samples, and normalizing tumor samples against their sex-matched reference plane. However, since the performance of Tangent improves with the number of normal samples in the reference plane, here, we include both male and female normal samples in our reference plane for Tangent normalization. The inclusion of the X chromosome requires special treatment to ensure that the distance from a tumor to a normal reflects noise differences, without being artificially inflated due to difference in sex. 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 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 sex, 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 sex-matched normals should enable recovery of whole-chromosome SCNAs in the sex chromosomes.

The large number of reference normal samples presents computational challenges as the projection matrix depends on the computation of the pseudoinverse of an $M \times n_N$ matrix (~1.5e6 × 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 ith 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.

2.3 Pseudo-Tangent

Amassing a sufficiently large collection of normal samples that spans the range of systematic noise types in the tumor samples is often difficult and sometimes infeasible. We therefore developed Pseudo-Tangent as an adaptation of Tangent that uses signal-subtracted tumor profiles (i.e. 'pseudo-normal profiles') to populate its reference subspace in addition to the standard normal profiles used by Tangent.

In the first step of Pseudo-Tangent, we use Tangent with a small set of normals to define the reference subspace and then circular binary segmentation (CBS) to generate a tentative copy-number profile for each tumor (Venkatraman and Olshen, 2007). We then subtract these tentative profiles from their original log-transformed tumor profiles in order to generate a corresponding pseudo-normal profile for each tumor input (keeping only deviations from the CBS segment values). In the next step of Pseudo-Tangent, the tumors are partitioned into n approximately equal subsets, with each subset then Tangent-normalized against a reference subspace of pseudo-normals in that subset's complement. The partition parameter n is a user-controlled parameter that is inversely related to the cardinality of each subset. Finally, CBS is used to convert the resulting Pseudo-Tangent-normalized coverage profiles into segmented copy-number calls in the form of $\log_2$ copy ratios (Supplementary Fig. S1C).

It is possible that with sufficient numbers of pseudo-normal samples, true SCNAs in a tumor may be normalized away due to overfitting. We therefore explored implementation of an additional step to limit the number of dimensions of the pseudo-normal space to improve Pseudo-Tangent's overall performance. We first performed truncated singular value decomposition (tSVD) on the entire collection of pseudo-normal profiles. We next retained a subset of the singular vectors with the greatest singular values to construct a reduced subspace spanned by these singular vectors before proceeding with the remaining steps of Pseudo-Tangent by normalizing our tumors against this reduced subspace. This tSVD step limited the dimensionality of our pseudo-normal reference subspace and constrained us to a smaller number of singular vectors to describe our pseudo-normal noise distribution.

2.4 Comparisons against other normalization methods

When comparing Tangent normalization to other normalization methods, we opted to exclude the X and Y chromosomes from our analyses so that differences in their handling of the sex chromosomes would not affect their performances. For similar reasons, we excluded CNV probes that map to known germline copy-number polymorphisms or other regions where, due to errors in the experimental platform, data across normals vary widely (Supplementary Table S1). In our data, we identified such regions by running GISTIC on the normal samples, to identify loci whose estimated copy-number changes in normals rose to levels that might lead to false positive GISTIC peaks. Specifically, we excluded loci whose G-scores across normals were >20% of the threshold for being called a significantly recurrent SCNA by GISTIC in the corresponding tumor cohort. To normalize using matched normals, we subtracted the $\log_2$ ratios of each matched normal from its corresponding tumor. For tumors with more than one matched normal (blood or normal tissue sample), the matched blood sample was preferred over the matched normal tissue sample. To normalize using the five nearest normals approach, we subtracted from each tumor the mean of the five normals closest to it based on Euclidean distance (Beroukhim et al., 2007).

To normalize using the average-normal method with WES data, we first averaged the coverage at each interval across the entire panel of normal samples to produce a standard average-normal. We then subtracted the $\log_2$ ratios of this computed average-normal from each tumor. Normalization using GC correction was performed based on the GC content normalization algorithm in HMMcopy (Lai et al., 2016).
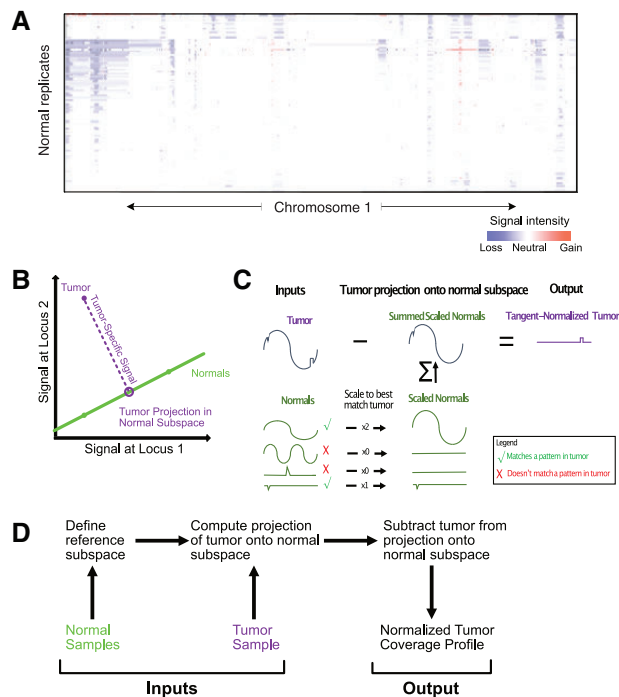
Fig. 1. Overview of problem and method. (**A**) Segmented pre-normalized $\log_2$ copy-number ratios on replicates of DNA from the HCC1143BL immortalized lymphocyte (non-cancer) line across 110 batches in chromosome 1. As these variations are observed in the same DNA, they represent experimental artifacts. (**B**) Reduced, 2D representation of the Tangent methodology. For each tumor we compute its projection onto a lower-dimensional subspace defined by normal samples profiled in parallel with the tumors. Signal representing somatic copy-number alterations is contained within the difference between the tumor and its projection. (**C**) Concept of Tangent normalization: by using a linear combination of normal samples as a reference, Tangent can compare each tumor sample to a reference constructed to match its noise components. (**D**) Flowchart describing the steps of Tangent normalization

3. Results

3.1 Tangent method overview

We have found systematic biases to be prevalent in both array- and sequencing-based data, both within and across batches, and found that these biases can generate widespread false positive SCNAs that can recur across samples (Fig. 1A). In principle, these biases can be overcome by normalizing tumor data only against normal control samples that have been profiled under identical experimental conditions. In practice, many of these experimental conditions are neither known nor measured. We developed the Tangent method to reconstruct normal controls that most accurately represent the tumor noise profile, so as to overcome these tumor-specific biases.

Tangent assumes that variations in experimental conditions can introduce variations in signal intensity or coverage profiles, such that normal samples that represent a single diploid state can produce signal intensity or coverage profiles encompassing a subspace N of the space of all possible coverage profiles. By accruing a collection of normal samples from the same batch/center as the tumors and with similar noise characteristics, Tangent attempts to construct this reference subspace N as the space that spans all linear combinations of normal profiles. Tangent then assumes that, for any copy-number profile T from a tumor sample, the point in subspace N that is most similar to T represents the profile of a normal sample characterized under similar conditions as T. SCNAs are then represented as the difference between T and that nearest point in subspace N (Fig. 1B–D; see Section 2).

3.2 Tangent analysis on microarray data

To assess Tangent's performance on copy-number profiles generated from microarray data, we applied it to data comprising 497

glioblastomas and 451 normal samples profiled by TCGA using SNP 6.0 arrays. We benchmarked Tangent against two other normalization methods: use of matched normal samples from the same patient (which was possible for only 386 of the GBMs), and use of the five normal samples with noise profiles closest to those in the tumor (Beroukhim *et al.*, 2007). We compared the performance of these normalization approaches in detecting SCNAs based on preservation of signal intensity, reduction in 'empirical noise', and improvement in signal-to-noise ratio (SNR). We measured the 'empirical signal' as the standard deviation of median signal intensities among all chromosome arms and the 'empirical noise' as the median absolute difference between $\log_2$ copy ratios of adjacent intervals or probes.

All three normalization methods described preserved signal integrity, but only Tangent normalization consistently reduced systematic empirical noise and thus increased SNRs (Fig. 2A–C; Supplementary Fig. S2). Normalization using the five nearest normals improves empirical noise levels negligibly, and normalization by matched normals tends actually to increase empirical noise levels and decrease SNRs relative to data that had not been normalized. As a result, segmented copy-number profiles generated after Tangent normalization exhibited less hyper-segmentation than profiles generated using other methods (Supplementary Fig. S2).

We next investigated the effects of increasing the size of the normal reference pool used by Tangent on reducing noise. We re-applied Tangent to our set of glioblastomas while incrementing the numbers of normal samples used to define our reference subspace from 0 (i.e. no use of Tangent) to 3146 samples across 13 batches (median number of normal samples per batch 255, range 102 to 281). These normal samples represented data generated by TCGA from normal blood leukocytes obtained from patients with a variety of cancers. We observed a monotonic reduction in median empirical noise levels with increasing numbers of normal samples, although this improvement decreased asymptotically and offered negligible benefits after approximately 1000 normal samples (four batches; Fig. 2D; note that the number of normals varied across TCGA batches). It is possible that the optimal number of normals depends upon the resolution of the underlying data. We would anticipate that higher-resolution data, measured at more genomic loci, might benefit more from large panels of normal samples.

We also investigated the effects of altering the composition of our normal reference pool, and specifically the utility of including normal samples that had been profiled in the same versus different batches of arrays as the tumor under study. We observed greater empirical noise reduction when utilizing the entire set of normal samples across batches than we did when applying Tangent using a reference subspace containing only normal samples from the same batch (Fig. 2E). Nevertheless, whether Tangent utilizes the entire reference subspace or it uses only a subset of normal samples from the same batch, both methods consistently yield lower levels of post-normalization empirical noise compared to pre-normalization noise for all tumors.

3.3 Tangent analysis on whole-exome sequencing data

We next evaluated Tangent's performance on sequencing data, by applying it to WES data generated by TCGA. We initially evaluated data from 123 tumors and 129 matched normal samples across four tumor types: low grade gliomas, lung squamous cell carcinomas, prostate adenocarcinomas and stomach adenocarcinomas ('WES set 1'; see Supplementary Table S3). We compared the performance of Tangent with normalizing against matched normals, an average normal from a panel of normals. We also combined each approach with a method that corrects for variations in local GC content (Ha *et al.*, 2014) to determine whether Tangent provides improvements beyond GC correction.

We found that Tangent outperforms these conventional normalization methods. Specifically, the average empirical noise in post-Tangent-normalized data is 35% lower than post-normalization against matched normals and 26% lower than post-normalization
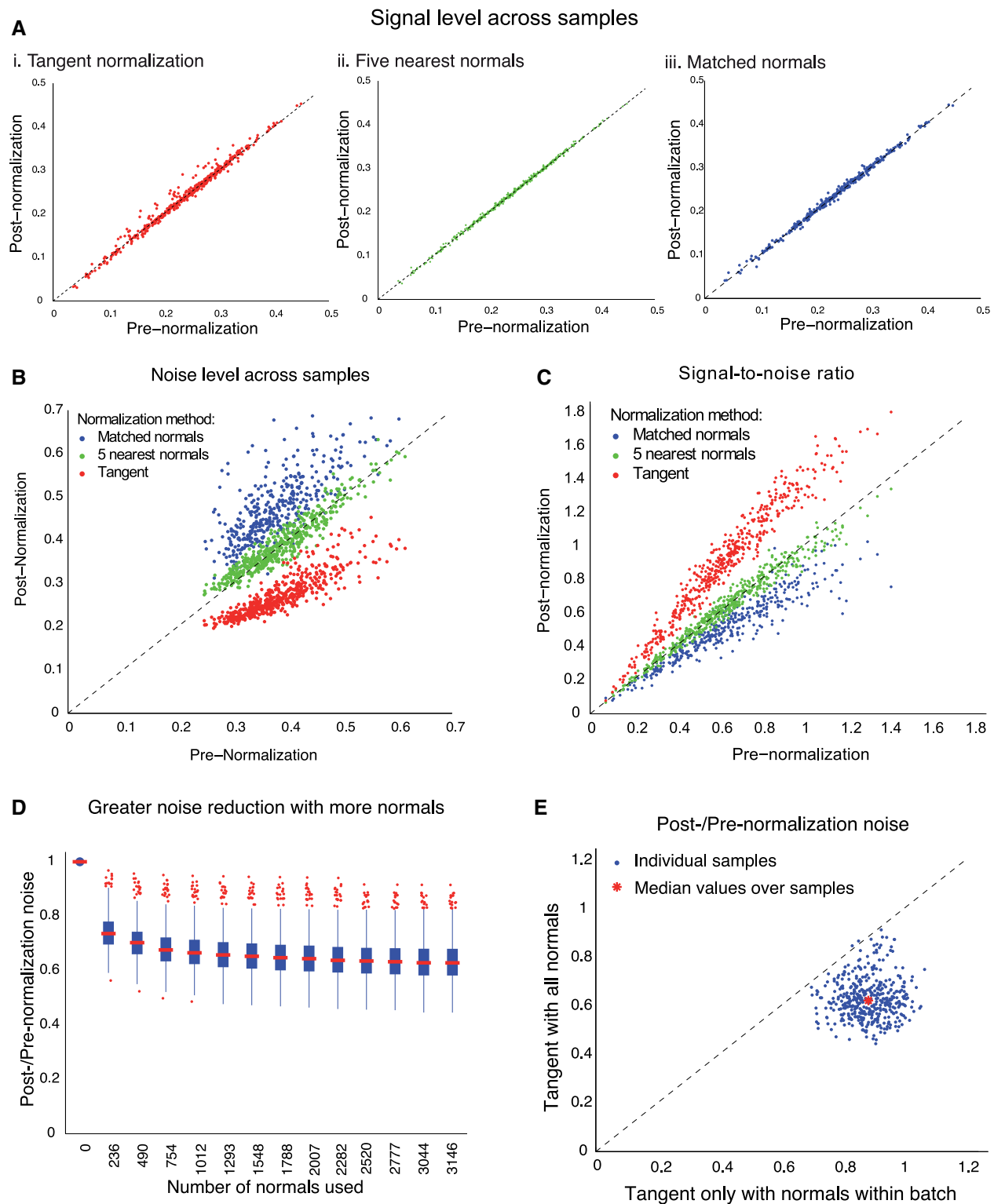
Fig. 2. Normalization of SNP array data from 497 TCGA glioblastomas. Scatter plots indicate post-normalization versus pre-normalization (**A**) signal, (**B**) noise level and (**C**) signal-to-noise ratios for the normalization methods: Tangent, five nearest normals and matched normals. (**D**) Box plot of post-normalization noise as a fraction of pre-normalization noise, following tangent normalization with increasing numbers of normal samples (approximately 250 normal samples were added in each batch). (**E**) Noise ratio (post-normalization over pre-normalization noise) for glioblastoma samples following tangent normalization using the entire reference plane versus tangent normalization using only the normal samples processed in the same batch as a tumor. Almost all samples lie below $y = x$, indicating that there is greater noise reduction with the full reference plane

against an average normal (Fig. 3A). This level of noise reduction is attained without significant compromise on signal. The average SNR in post-Tangent-normalized data is 58% higher than that post-normalization against matched normals, and 78% higher than post-normalization against an average normal (Fig. 3B). Adding GC correction to the other two normalization methods does not enable
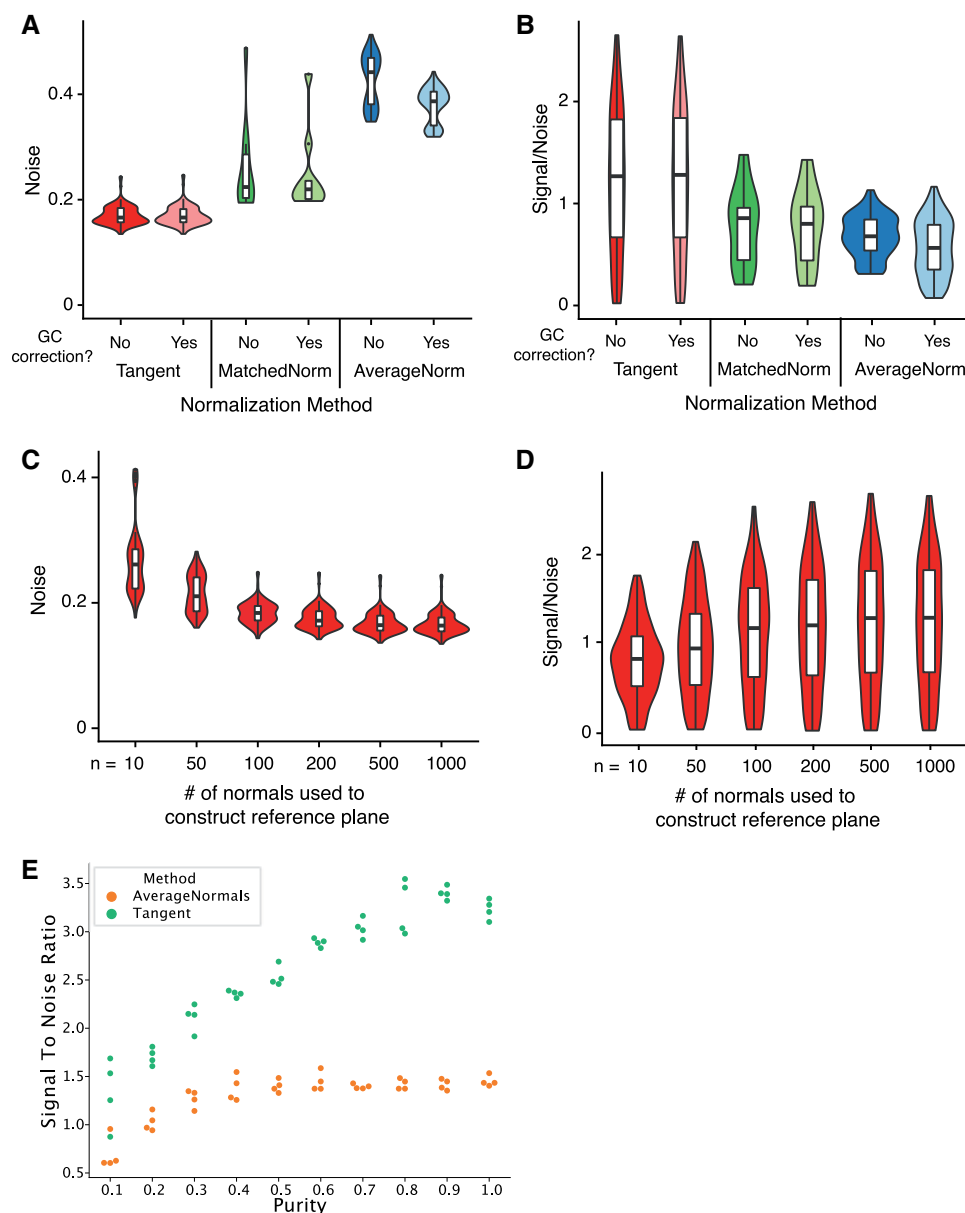
Fig. 3. Performance of Tangent on WES data. (**A**) Noise levels and (**B**) signal-to-noise ratios for Tangent-normalized data (tangent); data normalized against corresponding matched normals (MatchedNorm); and data normalized against an average across a panel of normals (AverageNorm), both with and without additional GC correction. (**C**) Noise and (**D**) signal-to-noise ratios plotted against the number of normal samples in the reference subspace. (**E**) Signal-to-noise ratios for average-normal normalized profiles and Tangent-normalized profiles plotted against purity for 40 simulated tumor sample generated at 10 different levels of purity

them to reach the performance of Tangent. Use of Tangent results in 31% lower empirical noise and 57% higher SNR on average than use of matched normals and 55% lower empirical noise and 115% higher SNR on average than use of an average normal. Application of GC-correction to Tangent-normalized data provided only marginal benefit relative to Tangent alone (Fig. 3A and B).

These analyses suggest that Tangent normalization largely corrects for noise resulting from variations in GC content across the genome. We also assessed the extent to which GC content, as well as replication timing, contribute to the estimated noise profiles (the projection of the tumor vector onto the noise subspace N) that Tangent subtracts. We first determined the estimated noise profile of each tumor in WES set 1 by computing its projection into the subspace of matched normal samples. We then determined the correlation between each of these estimated noise profiles with genome-wide vectors for GC content and replication timing (The ICGC/TCGA Pan-Cancer Analysis of

Whole Genomes Consortium, 2020). Overall, we observed only weak correlations (average Spearman ρ of 0.00614 and 0.0223, respectively; Supplementary Fig. S3A and B).

Tangent also performs better in detecting copy-number breakpoints from WES data. To establish ground-truth breakpoints, we applied SvABA to call rearrangement loci in 249 TCGA samples that had whole-genome sequencing (WGS) data available (Wala *et al.*, 2018). By using the SvABA rearrangement calls as gold standard, we compared copy-number breakpoints detected in WES analyses to these loci from WGS data, and calculated the distances between these breakpoints and rearrangement loci. We specifically compared the performance of Tangent normalization to the average-normal method. The Tangent pipeline resulted in more copy-number breakpoints that were adjacent to rearrangement loci from WGS data compared to the average-normal pipeline (41% versus 13%). Among breakpoints that were adjacent to a rearranged locus, the median distance between breakpoints and rearranged loci

was lower in the Tangent analyses than the average-normal analyses (34.1 kb versus 203 kb; Wilcoxon Rank-Sum $P = 2.2 \times 10^{-16}$) (Supplementary Fig. S3C and D).

Moreover, tumor purity and ploidy estimates improve with use of Tangent. We applied ABSOLUTE on the segmented copy-number calls generated by the Tangent and average-normal pipelines on the samples with WGS data (Carter *et al.*, 2012). The resulting purity and ploidy estimates were compared to the ground truth that was defined by the ABSOLUTE purity and ploidy calls available on Genomic Data Commons (Taylor *et al.*, 2018). We found that all of the Tangent-normalized samples received ABSOLUTE calls, whereas 25% of the samples in the average-normal pipeline were issued a 'no call'. The average difference in purity estimates in Tangent-normalized data compared to the ground truth was lower than that in the average-normal pipeline (0.057 versus 0.103). To assess the accuracy of the ploidy estimates, we calculated the $\log_2$ ratio of the estimated ploidy to the ground truth ploidy. We found that 64% of Tangent-normalized samples had the correct ploidy calls, as compared to 37% in the case of the average-normal pipeline. Among samples with incorrect ploidy calls, in the setting of Tangent normalization, 25% could be explained by a discrepancy in genome doubling calling, compared to 17% in the setting of the average-normal pipeline (Supplementary Fig. S3E–G).

Furthermore, we demonstrated that Tangent improves the accuracy of purity and ploidy estimates by comparing these estimates with 'real' measures of purity and ploidy from FACS-sorting analyses on 36 ovarian carcinomas in TCGA (Carter *et al.*, 2012). ABSOLUTE purity and ploidy calls were generated after either Tangent or average-normal normalization, and the absolute differences in purity and ploidy from the FACS data were computed. Tangent significantly outperformed Average-Normal normalization in estimating ploidies relative to the FACS-sorting estimates (Mann–Whitney $P = 1.33e{-}3$; Supplementary Fig. S3H). Estimated purities from FACS-based estimates were not significantly closer to either normalization method (Mann–Whitney $P = 0.479$; Supplementary Fig. S3I).

Similar to our experience with SNP arrays, we found that increased numbers of normal samples in the reference pool improved empirical noise profiles after Tangent normalization. Using WES set 1, we applied Tangent using between 10 and 1000 reference normal samples sequenced by TCGA, including normal samples from patients with the four tumor types under study and six other tumor types (see Supplementary Material). We found that Tangent's performance plateaued at approximately 200 normal samples (Fig. 3C and D).

In order to understand how Tangent improves SCNA calling under different tumor purity conditions, we assessed Tangent's performance relative to average-normal normalization on a set of simulated sequencing runs. Using EAGLE, a whole genome simulator developed by Illumina to mimic their sequencers' performance including biases and errors, we generated 40 simulated tumor samples (4 at each of 10 purity levels ranging from 10% to 100%), as well as 40 simulated normal samples. After running Tangent on these simulated samples, we noticed that SNRs improve as underlying tumor purity increases (Fig. 3E). We estimated signal here as the standard deviation of signal intensities of all segments after CBS and empirical noise as the median absolute difference between $\log_2$ copy ratios of adjacent intervals. In particular, signal, noise and SNR all remain significantly improved with Tangent relative to Average Normals Normalization at all tumor purities down to 10% (Fig. 3E and Supplementary Fig. S4A and B). We conclude that Tangent normalization improves SNR compared to average-normal normalization regardless of purity.

One concern in using denoising algorithms such as Tangent is that overfitting of tumors may eliminate real tumor signal. One consequence could be to decrease estimated tumor purity. To address this concern, we ran ABSOLUTE on 10 tumors within WES set 1 after normalizing their profiles using between 10 and 1000 reference normal samples. The average difference in the purity estimated using 10 samples versus 1000 samples was only 0.006 (Supplementary Fig. S4C and D).

In addition to running Tangent on these samples in TCGA, the vast majority of which were fresh frozen samples, we also compared the performance of Tangent to average-normal normalization using a cohort of 37 Formalin-Fixed-Paraffin-Embedded (FFPE) CESC samples in TCGA. Traditionally, FFPE samples have represented a greater challenge in copy-number calling and often result in highly noisy copy-number profiles (McSherry *et al.*, 2007). We demonstrated substantially improved copy-number calling when using Tangent versus average-normal normalization, as quantified by reduction of hypersegmentation, improving from a median of 378 segments to 131 segments (Wilcoxon Rank-Sum $P = 2.72 \times 10^{-10}$), regardless of segmentation algorithm used (Supplementary Fig. S4E and F). These results demonstrate that Tangent improves copy-number determination in FFPE samples, as well as in fresh frozen samples.

We also found Tangent to run efficiently. When applying Tangent to exomes representing 40 tumor and 40 normal samples, Tangent consumed 2 min and 3 s of CPU on one Intel Skylake core.

3.4 The relative importance of normalization to other pipeline components

We performed additional benchmarking analyses on a cohort of TCGA CESC samples (306 WES tumors and 295 corresponding SNP array tumors, detailed in Supplementary Methods) to evaluate the magnitude of improvements in copy-number calling as a result of Tangent normalization relative to differences between segmentation algorithms. We generated copy-number calls using four copy-number pipelines: first by comparing Tangent to average-normal normalization, then CBS to piecewise constant fitting (PCF) segmentation post-normalization (Nilsen *et al.*, 2012; Venkatraman and Olshen, 2007). We performed comparisons according to four metrics: (i) manual review of copy-number profiles; (ii) average number of copy-number segments; (iii) distance between copy-number profiles generated from WES versus SNP array data; and (iv) recapitulation of known copy-number driver events using GISTIC2.0.

Visualization of the copy-number profiles from all 306 WES samples demonstrated marked improvement from the average-normal pipelines to Tangent and comparatively modest gains in performance using CBS versus PCF as the segmentation algorithm of choice (Supplementary Fig. S5A). The profiles generated from the average-normal pipelines showed markedly increased and more inconsistent copy-number events and empirical noise compared to the cleaner profiles observed with Tangent.

We observed similar improvements in Tangent from average-normal normalization when assessing noise in terms of level of hypersegmentation. Analyses with extensive noise tend to result in hypersegmentation, which can be reflected in higher number of copy-number segments per sample. Following this principle, we found that Tangent consistently generated fewer segments than average-normal normalization regardless of our choice in segmentation algorithms and across different experimental platforms (Supplementary Fig. S5B).

When comparing copy-number pipelines, one would expect the best pipeline to provide the greatest concordance between results from different experimental platforms. We used distance correlation to evaluate the concordance across experimental platforms. We found that once again regardless of the choice of segmentation algorithm, Tangent normalization led to higher distance correlation between SNP array and whole-exome data when compared to the average-normal pipelines (Supplementary Fig. S5C).

To compare the ability of the pipelines in recapitulating known driver SCNAs, we first applied GISTIC2.0 on WES data to generate a list of significant copy-number events in the form of GISTIC peaks (Mermel *et al.*, 2011). We then compared these events to a published list of 17 known gene amplifications and 3 known gene deletions in CESC (The Cancer Genome Atlas Research Network, 2017). Copy-number data processed using the Tangent pipelines detected more known driver events than the average-normal pipelines, as evident in both cases of CBS (19/20 versus 2/20) and PCF segmentation (16/20 versus 6/20). In the Tangent pipeline with CBS segmentation,

19 of the 20 driver events were detected, with the exception being KLF5 amplification (Supplementary Table S2). We suspect the KLF5 amplification peak is a result of local super-enhancer duplication, and since WES technology likely would not have captured the non-coding super-enhancer region, this event was not detected in any of the four pipelines (Zhang *et al.*, 2018). However, the overall greater sensitivity in detecting copy-number driver genes from Tangent-normalized data compared to data processed with the average-normal pipelines strongly suggests that Tangent normalization offers a substantial improvement in copy-number calling, more so than the choice of segmentation algorithm.

In all of these analyses, we found that Tangent normalization meaningfully resulted in higher accuracy in copy-number calling when compared to average-normal normalization. These improvements in accuracy also exceeded the differences between CBS and PCF segmentation, which affirms the important role of Tangent in accurately profiling SCNAs.

3.5 Pseudo-Tangent: a method to compensate for insufficient normal data

Tangent assumes that systematic noise distributions in tumors are identical to those in normal samples. However, it is often impossible to collect a sufficiently large collection of normal samples to encompass the range of systematic noise types spanned by the tumor samples. In light of this limitation, we developed Pseudo-Tangent as an adaptation of the Tangent pipeline that utilizes a reference subspace composed of signal-subtracted tumor profiles (i.e. 'pseudo-normal profiles') instead of the standard normals used in Tangent. In brief, the method first estimates SCNAs for each tumor using standard Tangent with a limited number of normal samples. Pseudo-Tangent then applies Tangent again to detect SCNAs for each tumor, using a reference subspace comprising other tumors from which the initially detected SCNAs had been subtracted (see Section 2).

We applied Pseudo-Tangent to TCGA WES data from 306 CESC primary tumors ('WES set 2'). We initially normalized these data against WES data from five normal samples obtained from blood and used these to generate 306 corresponding pseudo-normal profiles. We then divided the tumors and their matching pseudo-normal profiles into three batches, and normalized each tumor in each batch against the pseudo-normal profiles in the other two batches. (The number of batches is a modifiable parameter.)

We then compared these results to previously generated gold-standard absolute allelic copy-number profiles (Taylor *et al.*, 2018). The gold-standard profiles were generated by applying the standard Tangent pipeline and the ABSOLUTE algorithm (see Section 2; Carter *et al.*, 2012) to primarily SNP array data from these 306 tumors and 3154 normal samples. (ABSOLUTE did, however, use mutation calls from WES data to optimize its tumor purity estimates.) We selected gold-standard profiles based upon a different experimental platform (SNP array data) to minimize cross-contamination of artifacts in the WES data used by Pseudo-Tangent. We measured empirical noise as the average distance of each probe in the Pseudo-Tangent-generated coverage profile from its nearest estimated absolute total copy-number level. We found that all 278 CESC tumors with ABSOLUTE solutions displayed lower empirical noise levels after undergoing Pseudo-Tangent normalization than they did after just the initial round of Tangent normalization using only the 5 true normal samples (Fig. 4A).

One concern with applying Pseudo-Tangent is that with sufficient numbers of pseudo-normal samples, true SCNAs in a tumor may be normalized away due to overfitting. We therefore explored whether limiting the number of dimensions of the pseudo-normal space could improve Pseudo-Tangent's overall performance. Specifically, we performed singular value decomposition of the pseudo-normal reference subspace, retained between 10 and 306 of the singular vectors with the greatest singular values, and normalized our tumors against a reduced subspace spanned by these singular vectors. We then determined the number of singular vectors that provided optimal results, as indicated by generating copy-number profiles with the smallest deviations from the results of our gold-standard ABSOLUTE runs on the same tumors.

We found that the median difference between copy-number levels generated after Pseudo-Tangent and those generated by the gold-standard ABSOLUTE pipeline was lowest when we used the 150 singular vectors with the greatest singular values (Fig. 4B), which captured 98% of the variance of the entire pseudo-normal reference subspace. However, the optimal number of singular vectors varied across the different tumors. In particular, the noisiest tumors seemed to have greater empirical noise reductions when 10 singular vectors were used rather than larger numbers of singular vectors (Fig. 4C). This behavior suggests that optimal use of Pseudo-Tangent might take into account the noise level of the tumor being normalized when determining the number of singular vectors to retain.

In addition to running Pseudo-Tangent on these 306 fresh-frozen CESC samples, we also investigated Pseudo-Tangent's effect on noisier FFPE tumors. We ran Pseudo-Tangent on a set of 37 FFPE TCGA tumors, using 5 corresponding blood normal samples for initial normalization. Upon comparison with gold-standard absolute allelic copy-number profiles as described above, the Pseudo-Tangent-normalized copy-number profiles again exhibited lower empirical noise levels as compared to the profiles of the same samples after only the initial round of Tangent normalization with only five normal samples (Fig. 4D). Furthermore, when we attempted Pseudo-Tangent normalization using a reduced subspace spanned by the top 10 pseudo-normal singular vectors, we noted an additional reduction in empirical noise, again primarily driven by the noisiest tumors (Fig. 4E).

4. Discussion

Although Tangent was developed for use with SNP array data, we have extended its use to WES data, and in principle it can be applied to any source of copy-number data that measures DNA dosage with varying signal intensity or depth of coverage, such as WGS or CGH. For example, we have applied Tangent to targeted NGS cancer panels to identify somatic copy-number alterations (Brastianos *et al.*, 2013). Indeed, Fehrmann *et al.* (2015) developed a similar method to detect SCNAs from transcriptomic profiling data, in which they remove principal components reflecting different transcriptional states to enrich for transcriptional changes reflecting underlying SCNAs. PEER (Stegle *et al.*, 2010) also applies a similar approach to tangent—factor analysis—to remove common expression patterns in transcriptomic analyses.

Several additional approaches can further improve copy-number estimates. For example, integrating rearrangement data collected from WGS (Drier *et al.*, 2013; Layer, 2014; Rausch *et al.*, 2012; Wala *et al.*, 2018) can provide information about copy-number breakpoints, thereby further improving accuracy of SCNA profiles. Further improvements to SNRs can also be obtained from algorithms that determine differences in absolute rather than relative copy-numbers (Carter *et al.*, 2012; Van Loo *et al.*, 2010), including methods such as Sequenza (Favero *et al.*, 2015), Accucopy (Fan *et al.*, 2021) and Sclust (Cun *et al.*, 2018), which integrate tumor purity and ploidy estimation with normalization and segmentation to further improve accuracy. However, these algorithms require normalized copy-number ratios as inputs or calculate them internally, and therefore are likely to benefit from the improved normalization Tangent provides.

Accurate SCNA determination relies on having normal control samples that have been processed in identical fashion to the tumors. For example, SCNA profiling of tumors obtained from a large variety of institutional sources—such as may occur when profiling tumors studied in multi-institutional clinical trials or in clinical laboratories—would ideally make use of normal tissue obtained from each institution contributing tumors. Unfortunately, this is often difficult or impossible in practice. Likewise, tumor tissue obtained through careful surgical resection in which the tumor is separated from its blood supply for an extended period may not be adequately reflected by normal DNA from blood samples. Pseudo-Tangent may
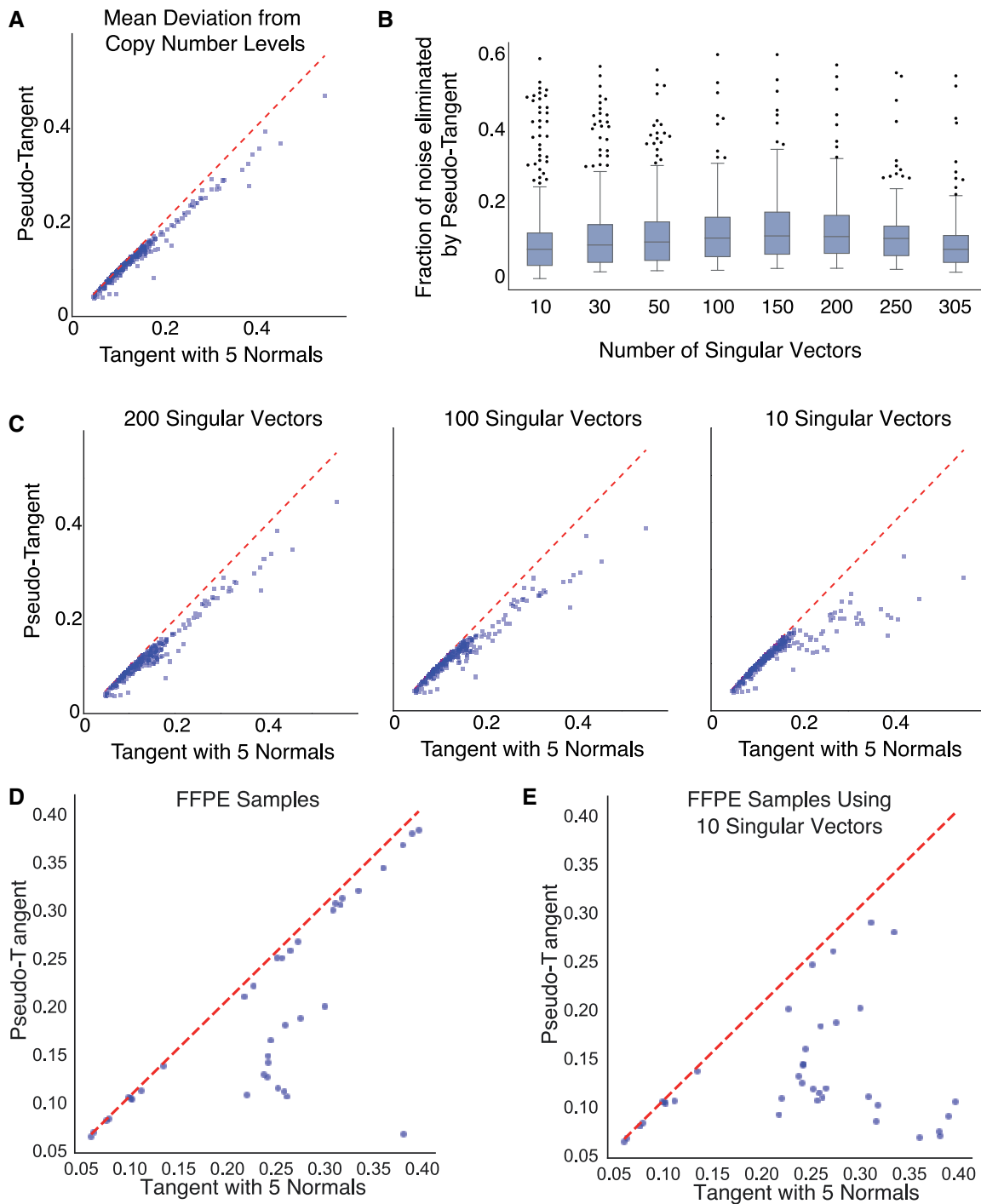
Fig. 4. Pseudo-Tangent decreases noise in resulting copy-number profiles compared to standard tangent, as measured by deviation from ABSOLUTE-estimated copy-number levels. (**A**) Average deviation from ABSOLUTE-estimated copy-number levels after Pseudo-Tangent (vertical axis) versus Tangent alone (horizontal axis). (**B**) Improvement in the deviation from ABSOLUTE-estimated copy-number levels after use of Pseudo-Tangent, as a fraction of the deviation after only standard Tangent had been used (vertical axis), against the number of singular vectors used for Pseudo-Tangent (horizontal axis). The median improvement was greatest when 150 singular vectors were used. (**C**) Average deviation from ABSOLUTE-estimated copy-number levels after Pseudo-Tangent (vertical axis) versus Tangent alone (horizontal axis) as in (A), after use of (left) 200 singular vectors, (middle) 100 singular vectors and (right) 10 singular vectors. Although median levels of deviation from ABSOLUTE-estimated copy-number levels increased when fewer than 150 singular vectors were used, the noisiest tumors saw the greatest improvements when only 10 singular vectors were used. (**D**) Average deviation from ABSOLUTE-estimated copy-number levels after Pseudo-Tangent (vertical axis) versus Tangent alone (horizontal axis) for 37 FFPE samples. (**E**) Average deviation from ABSOLUTE-estimated copy-number levels after Pseudo-Tangent with use of 10 singular vectors (vertical axis) versus Tangent alone (horizontal axis) for 37 FFPE samples

help remove the effects of systematic noise in these situations by generating pseudo-normals from tumors that were processed in similar fashion to each other. However, application of Pseudo-Tangent carries risk of overfitting and loss of signal, particularly if SCNAs are not adequately removed while generating pseudo-normals from

tumor samples. This is also a concern with Tangent itself, if some of the 'normals' it receives as input in fact contain substantial signal from tumors. In situations where true normals are available, extensive profiling of these normals as controls for the tumors is preferable to computational generation of pseudo-normals as described

here. Moreover, we recommend performing quality control on these normals to ensure that they contain no tumor-in-normal contamination—with special consideration to the possibility that tumors are mislabeled as normals, and to tumor-adjacent normals that might actually include infiltrating tumor.

The Tangent pipeline we describe here was the basis for copy-number determination across TCGA. Through improved denoising of raw coverage data, Tangent permits more accurate detection of SCNAs compared to previous methods. With this improved detection comes potential increased ability to identify subclonal SCNAs, which have traditionally been and remain rather difficult to detect. Additionally, both Tangent and Pseudo-Tangent are widely applicable to a large variety of research and clinical applications and copy-number profiling platforms, and can be integrated with further improvements in SCNA detection that make use of alternative sources of information such as rearrangement locations and tumor purity and ploidy.

Data availability

The data underlying this article were accessed from the Genomic Data Commons Data Portal (https://portal.gdc.cancer.gov/) and Firecloud/Terra (https://firecloud.terra.bio/). The derived data generated in this research will be shared on reasonable request to the corresponding author.

References

Beroukhim,R. *et al.* (2007) Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma. *Proc. Natl. Acad. Sci. USA*, 104, 20007–20012.

Beroukhim,R. *et al.* (2010) The landscape of somatic copy-number alteration across human cancers. *Nature*, 463, 899–905.

Boeva,V. *et al.* (2012) Control-FREEC: a tool for assessing copy number and allelic content using next generation sequencing data. *Bioinformatics*, 28, 423–425.

Brastianos,P.K. *et al.* (2013) Genomic sequencing of meningiomas identifies oncogenic SMO and AKT1 mutations. *Nat. Genet.*, 45, 285–289.

Carter,S.L. *et al.* (2012) Absolute quantification of somatic DNA alterations in human cancer. *Nat. Biotechnol.*, 30, 413–421.

Cun,Y. *et al.* (2018) Copy-number analysis and inference of subclonal populations in cancer genomes using sclust. *Nat. Protoc.*, 13, 1488–1501.

DePristo,M. *et al.* (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.*, 43, 491–498.

Drier,Y. *et al.* (2013) Somatic rearrangements across cancer reveal classes of samples with distinct patterns of DNA breakage and rearrangement-induced hypermutability. *Genome Res.*, 23, 228–235.

Fan,X. *et al.* (2021) Accucopy: accurate and fast inference of allele-specific copy number alterations from low-coverage low-purity tumor sequencing data. *BMC Bioinformatics*, 22, 23.

Favero,F. *et al.* (2015) Sequenza: allele-specific copy number and mutation profiles from tumor sequencing data. *Ann. Oncol.*, 26, 64–70.

Fehrmann,R.S.N. *et al.* (2015) Gene expression analysis identifies global gene dosage sensitivity in cancer. *Nat. Genet.*, 47, 115–125.

Ha,G. *et al.* (2014) TITAN: inference of copy number architectures in clonal cell populations from tumor whole-genome sequence data. *Genome Res.*, 24, 1881–1893.

Koboldt,D.C. *et al.* (2012) VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.*, 22, 568–576.

Korn,J.M. *et al.* (2008) Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. *Nat. Genet.*, 40, 1253–1260.

LaFramboise,T. (2009) Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances. *Nucleic Acids Res.*, 37, 4181–4193.

Lai,D. *et al.* (2016) HMMcopy: copy number prediction with correction for GC and mappability bias for HTS data. R package version 1.22.0.

Layer,R.M. *et al.* (2014) LUMPY: A probabilistic framework for structural variant discovery. *Genome Biol.*, 15, R84.

McKenna,A. *et al.* (2010) The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.*, 20, 1297–1303.

McSherry,E.A. *et al.* (2007) Formalin-fixed paraffin-embedded clinical tissues show spurious copy number changes in array-CGH profiles. *Clin. Genet.*, 72, 441–447.

Mermel,C.H. *et al.* (2011) GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol.*, 12, R41.

Nilsen,G. *et al.* (2012) Copynumber: efficient algorithms for single- and multi-track copy number segmentation. *BMC Genomics.*, 13, 591.

Rausch,T. *et al.* (2012) DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics*, 28, i333–i339.

Rieber,N. *et al.* (2017) Reliability of algorithmic somatic copy number alteration detection from targeted capture data. *Bioinformatics*, 33, 2791–2798.

Sathirapongsasuti,J.F. *et al.* (2011) Exome sequencing-based copy-number variation and loss of heterozygosity detection: exomeCNV. *Bioinformatics*, 27, 2648–2654.

Stegle,O. *et al.* (2012) Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nat. Protoc.*, 7, 500–507.

Talevich,E. *et al.* (2016) CNVkit: genome-wide copy number detection and visualization from targeted DNA sequencing. *PLoS Comput. Biol.*, 12, e1004873.

Taylor,A.M. *et al.*; Cancer Genome Atlas Research Network (2018) Genomic and functional approaches to understanding cancer aneuploidy. *Cancer Cell*, 33, 676–689.e3.

The Cancer Genome Atlas Network (2008) Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, 455, 1061–1068.

The Cancer Genome Atlas Network et al. (2013) The Cancer Genome Atlas Pan-Cancer analysis project. *Nat Genet*, 45, 1113–1120.

The Cancer Genome Atlas Research Network. (2017) Integrated genomic and molecular characterization of cervical cancer. *Nature*, 543, 378–384.

The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium (2020) Pan-cancer analysis of whole genomes. *Nature*, 578, 82–93.

Van Loo,P. *et al.* (2010) Allele-specific copy number analysis of tumors. *Proc. Natl. Acad. Sci. USA*, 107, 16910–16915.

Venkatraman,E.S. and Olshen,A.B. (2007) A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics*, 23, 657–663.

Wala,J.A. *et al.* (2018) SvABA: genome-wide detection of structural variants and indels by local assembly. *Genome Res.*, 28, 581–591.

Weir,B. *et al.* (2004) Somatic alterations in the human cancer genome. *Cancer Cell*, 6, 433–438.

Yoon,S. *et al.* (2009) Sensitive and accurate detection of copy number variants using read depth of coverage. *Genome Res.*, 19, 1586–1592.

Zack,T.I. *et al.* (2013) Pan-cancer patterns of somatic copy number alteration. *Nat. Genet.*, 45, 1134–1140.

Zare,F. *et al.* (2017) An evaluation of copy number variation detection tools for cancer using whole exome sequencing data. *BMC Bioinformatics*, 18, 286.

Zhang,X. *et al.* (2018) Somatic superenhancer duplications and hotspot mutations lead to oncogenic activation of the KLF5 transcription factor. *Cancer Discov.*, 8, 108–125.

Zhao,M. *et al.* (2013) Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives. *BMC Bioinformatics*, 14, S1.