The Tangent copy-number inference pipeline for cancer genome analyses

by

Coyin Oh

Harvard-M.I.T. Division of Health Sciences and Technology

Submitted in Partial Fulfillment of the Requirements for the M.D. Degree
with Honors in a Special Field at Harvard Medical School

February 2020

Area of Concentration:	Cancer Genomics
Project Advisor:	Dr. Rameen Beroukhim
Prior Degrees:	B.S. (Biological Engineering)

I have reviewed this thesis. It represents work done by the author under my guidance/supervision.

Thesis Advisor

Table of Contents

Abstract

Somatic copy-number alterations (SCNAs) play an important role in the development of cancer. In cancer genome analyses, accurate profiling of SCNAs is often deterred by the presence of noise in microarray and next-generation sequencing data. Here, we present the Tangent copy-number inference pipeline, which employs a novel normalization approach to reducing systematic noise in copy-number profiles. Tangent normalization first constructs a noise model using the weighted sum of noise profiles from a collection of normal samples that most closely matches the tumor noise profile. The estimated noise profile is then subtracted from the tumor signal to generate copy-number data for the tumor. The performance of Tangent thus hinges on having adequate representation of noise profiles in the normal reference collection. Since there are often practical limitations to obtaining sufficient number of normal samples, here we also describe the Pseudo-Tangent pipeline, which is an adaptation of Tangent to generate signal-subtracted tumor profiles that can be used to augment the reference collection. We applied Tangent to single-nucleotide polymorphism (SNP) array data and whole-exome sequencing data in The Cancer Genome Atlas (TCGA). Tangent normalization offers significant reduction in noise and improvement in signal-to-noise ratio compared to conventional normalization approaches. In the case of limited normal samples, Pseudo-Tangent also provides substantial reduction in systematic noise compared to Tangent alone and other conventional approaches. Tangent and Pseudo-Tangent are broadly applicable to multiple sequencing technologies for more accurate inference of SCNAs in the cancer genome. As part of TCGA, we have published the copy-number results for SNP array data of over 10,000 tumor-normal pairs that were analyzed with the Tangent pipeline. We have also made Tangent and Pseudo-Tangent publicly available for downloads and implementation.

Acknowledgements

I would like to take this opportunity to thank Dr. Rameen Beroukhim, my thesis advisor and research mentor during my time at Harvard Medical School. From our very first meeting, he has started dispensing life and career advice to me in his unique sense of humor, benign sarcasm, and his big heart of caring for the next generation of trainees. Without him, I would not have had the chance to join the Tangent project and continue the effort of those who laid the groundwork before me. My initial research goal in HST was to gain experience in the development of algorithms and methods to supplement my training as a computational biologist. Working with Rameen has given me the perfect opportunity to learn these skills, but I have learnt so much more than copy-number algorithms. I have been time and again inspired by his success in wearing multiple hats—a physician-scientist, an amazing mentor, and a wonderful role model. His mentorship has continued long past the summer and fall of 2017, when the HST thesis project theoretically ended. I would like to express my gratitude for the endless Skype/Hangouts Tangent update calls at odd times over the weekends and holidays as he worked around my rigid schedule during MS3 and MS4 years, instead of the other way around. He has gone above and beyond in supporting me in my work, my residency applications and in expressing concerns for my personal welfare—more so than one could imagine from a summer research mentor. I thank his continued patience and kindness, and I can only aspire to do it half as well for my future students.

The pioneering work for Tangent began circa 2010 in the Cancer Genome Analysis group at the Broad Institute and the Beroukhim Lab. The Tangent pipeline was originally developed for copy-number analyses of Affymetrix SNP Array 6.0 data for TCGA, and the analysis data from Tangent was published in the GDC portal. In 2017, I picked up the project to develop the Tangent pipeline this time for copy-number analyses of whole-exome sequencing data. I also worked with Galen Gao, another member of the lab who led the work of Pseudo-Tangent as I began clerkship year. I have decided to include all aspects of Tangent in this thesis as it provides the context for my work and the ongoing development of the pipeline as we prepare the manuscript for submission. A long list of people were involved in the development of Tangent, and I would like to thank some of them for their contributions to the project and for graciously helping me in my work:

- Barbara Tabak for being the master mathematician behind Tangent normalization, and for leading much of the initial—perhaps more critical—work on Tangent;
- Gordon Saksena for never failing to miss a conference call despite being halfway across the country, and for always ready to offer a solution at lightning speed;
- Galen Gao for being my almost- partner in crime in this work, for leading the development of Pseudo-Tangent, for tackling manuscript resubmissions from Texas, and for being an all-around wonderful peer colleague to work with and learn from;
- Steve Schumacher for being a huge help in getting me started with the work when I first joined the lab, for Pipette (oh the joy and occasional frustrations!), and for his wonderful company in our shared office before he moved to rural Maine;
- Lindsay Westlake for all of her help in problem-solving particularly the less glamorous bits and pieces of Tangent, and for being an extremely reliable colleague with a great personality;
- Ashton C. Berger for also helping me get started with the work during my first few months and for teaching me the intricacies of FireCloud;
- Andy Cherniack who has been giving me invaluable advice on my work in computational genomics since 2015;
- Other co-authors, including Barbara A. Hill, Michael Reich, and Scott Carter;
- Senior authors, including Dr. Gad Getz, Dr. Matthew Meyerson (and Dr. Rameen Beroukhim) for leading the direction of this work;
- Other members of the Beroukhim lab, particularly Ofer Shapira and Jeff Meng for their enlightening conversations about work and life in the office we shared on 75 Ames Street in 2017, and for dealing with my xkcd comic fever; as well as Mimi Bandopadhayay for her generous advice since 2016 and for once reminding me that my training as a physician should always come first from here on;
- Colleagues at the Genomics Platform and colleagues from the TCGA project, for their helpful discussions and analysis support; and
- Colleagues at Broad Information Technology Services who tirelessly answered all of my questions about software support and computing infrastructure, especially

when they are on call over the weekends (which incidentally were when I had time to work on Tangent during clerkship year).

I would also like to thank HST for their continued support—Dr. Rick Mitchell, Dr. Matt Frosch, Dr. Junne Kamihara, Patty Cunningham, Karrol Altarejos, and Dr. Paul Dieffenbach—to name a few. Your doors have been open to us every step of the way since we started at HMS, thank you for your dedication to training physician-scientists, your support in our scholarly pursuits, your ability to solve any problems that stand in the way of our training, and thank you for pushing us to exceed our personal and professional growths at HMS. This work would also not have been possible without the funding support from the HMS Scholars in Medicine office.

Finally, I would like to thank my family in Malaysia and friends around the world for supporting me during medical school and this journey of training to be a physician. My family—especially both of my parents Kim Seng and Geok Leong, my brother Coyan, and my sister Coyun—have shown me unconditional love, trust, and support in the past ten years despite the 9,000 miles that separate us. I am also extremely fortunate to have many friends who are so generous with their time, encouragement, and company—classmates in medical school, CCLF colleagues-turned-friends at the Broad, college friends from MIT, and my Malaysian community in Boston (who regularly shower me with home-cooked meals and climbing partners to keep me sane in medical school). Thank you for all your love.

* Please note that this thesis has been adapted from a co-first author manuscript that as of February 2020 is being prepared for resubmission to *Bioinformatics*:

Tabak B†, Saksena G†, Oh C†, Gao G†, Hill BA, Reich M, Schumacher SE, Westlake LC, Berger AC, Carter SL, Cherniack AD, Meyerson M‡, Beroukhim R‡, Getz G‡. The Tangent copy-number inference pipeline for cancer genome analyses. 2019.

† Equal contributors.

‡ Equal contributors.

A copy of the preprint is also available on bioRxiv:
<https://www.biorxiv.org/content/10.1101/566505v1>

Glossary

BAM	Binary Alignment Map
CBS	Circular binary segmentation
CESC	Cervical squamous cell carcinoma and endocervical adenocarcinoma
CGH	Comparative genomic hybridization
CNV	Copy-number variation; copy-number variant
DNA	Deoxyribonucleic acid
GATK	Genome analysis toolkit
GBM	Glioblastoma multiforme
GC	Guanine-cytosine
GDC	Genomic Data Commons
GISTIC	Genomic Identification of Significant Targets in Cancer
IGV	Integrative Genomics Viewer
KIRC	Kidney renal clear cell carcinoma
LAML	Acute myeloid leukemia
LGG	Brain lower grade glioma
LUSC	Lung squamous cell carcinoma
NGS	Next-generation sequencing
PCF	Piecewise constant fitting
PCR	Polymerase chain reaction
PRAD	Prostate adenocarcinoma
READ	Rectum adenocarcinoma
SCNA	Somatic copy-number alteration
SNP	Single nucleotide polymorphism
SNR	Signal-to-noise ratio
STAD	Stomach adenocarcinoma
SVD	Singular value decomposition
TCGA	The Cancer Genome Atlas
UCEC	Uterine corpus endometrial carcinoma
WES	Whole-exome sequencing
WGS	Whole-genome sequencing

Introduction

Cancer is often a result of the accumulation of somatic alterations in the genome. These alterations most commonly include point mutations, copy-number alterations, and structural rearrangements (Weir *et al.* 2004). Somatic copy-number alterations (SCNAs) in particular can have significant impact in driving the development of cancer by activating oncogenes or inactivating tumor suppressor genes (Beroukhim *et al.* 2010; Zack *et al.* 2013). In 2006, the National Cancer Institute and the National Human Genome Research Institute launched The Cancer Genome Atlas (TCGA) to comprehensively characterize the genomic and molecular profiles of cancer (McLendon *et al.* 2008; Weinstein *et al.* 2013). As part of TCGA, tumor samples from more than 11,000 cancer patients across 33 cancer types were collected and processed with single nucleotide polymorphism (SNP) arrays and whole-exome sequencing (WES) technologies. The advancement in high-resolution microarrays and next-generation sequencing (NGS) has been quintessential in supporting the vision of profiling these cancer genomes to accelerate the discovery of novel cancer driver genes (Korn *et al.* 2008; McLendon *et al.* 2008; Weinstein *et al.* 2013; Zack *et al.* 2013).

Currently, the standard approaches in characterizing somatic copy-number profiles centers around comparing the DNA content at various loci across the genome in tumor samples to that in normal samples. For example, in the use of array comparative genomic hybridization (CGH) or SNP arrays, the DNA content at each locus is derived from the signal intensities of the DNA probes that target these genomic loci (LaFramboise 2009). Similarly, in next-generation sequencing, the DNA content at each locus is determined by the coverage levels of loci across the genome (Yoon *et al. 2009)*. The

detection of SCNAs is then accomplished by determining the ratios of DNA content in tumor samples to that in normal samples at the corresponding genomic loci. These ratios become a way of normalizing the different affinities of array probes or sequencing primers at each locus. For the purpose of this comparison and normalization, the normal samples used is typically a matched normal sample from the same patient or a computed average among a collection of normal samples.

Such copy-number analyses are often confounded by at least three sources of noise. First, stochastic variations may result in random deviations between the measured DNA content and true DNA content. We can overcome random noise by averaging measurements across adjacent loci or by sequencing to a greater average depth. A common solution of averaging such measurements is to apply segmentation algorithms such as Circular Binary Segmentation (CBS) or Piecewise Constant Fitting (PCF) after a gross copy-number ratio is determined at each locus (Nilsen *et al.* 2012; Venkatraman & Olshen 2007). Second, false positives may result from germline copy-number variations (CNVs) being misinterpreted as SCNAs. We can overcome these false positives by using a blacklist of common germline CNVs or by comparing tumor and normal samples collected from the same patient. Third, systematic noise may result from various factors in experimental conditions that occur during the generation of sequencing data, which can affect probe affinities and coverage at particular loci.

Despite significant advancement made in sequencing technologies and computational analysis tools to tackle the issue of noise in copy-number data, filtering out systematic noise in high-resolution microarray and NGS data remains a huge challenge. In the past ten years, a variety of copy-number analysis pipelines such as Control-FREEC, ExomeCNV, VarScan2, and CNVkit have been published in literature (Boeva *et al.* 2012; Koboldt *et al.* 2017; Rieber *et al.* 2017; Sathirapongsasuti *et al.* 2011;

Talevich *et al.* 2016; Zare *et al.* 2017). However, many of these pipelines apply similar approaches in reducing systematic noise, which can be classified into the use of matched case-control samples, the use of a computed average from a collection of normal samples, or correction of bias in guanine-cytosine (GC) content (Zhao *et al.* 2013). While the noise in matched normal samples can often approximate their corresponding tumors' noise profiles, normal samples may not always be available either from practical limitations in clinical settings or from bench research that only utilizes isolated cancer cell lines. Even the normal tissues that are obtained may also be contaminated with tumor cells. Furthermore, when normal samples are readily available, they may undergo different processing conditions during the sequencing process from their corresponding tumors and therefore may exhibit different noise profiles (Grizzle *et al.* 2011). In the average normal approach, the computed average normal may be inadequately representing the various contributions of normal samples to the tumor noise profile. While GC correction may be helpful in reducing a large portion of noise in copy-number data, it certainly does not target all potential sources of systematic noise. Other sources of systematic noise include mappability variations across the genome, experimental variations during PCR amplification, cross-hybridization, and steps in sample and library preparation. Therefore, the presently available copy-number 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 in copy-number analysis by applying a novel methodology that constructs noise profiles from a subset of normal samples to target all potential sources of systematic noise. The normal samples used in Tangent ideally will have undergone identical experimental processing conditions as the tumor samples, but may not necessarily have to be from the same patients as the tumors. Given the previously highlighted challenges

in obtaining matched normal samples, a potential bottleneck of Tangent is thus the ability to obtain sufficient normal samples to form a large and diverse collection that adequately represents tumor noise profile. To address this bottleneck, we present Pseudo-Tangent as an extended approach of Tangent that converts tumor copy-number profiles into pseudo-normal copy-number profiles by subtracting inferred SCNAs to augment the reference normal collection in the pipeline.

Methods

Overview of the Tangent pipeline

The Tangent pipeline can be applied to both high-resolution microarray data and WES data for somatic copy-number analyses. This pipeline was the basis for the analyses of all Affymetrix SNP Array 6.0 data in TCGA. To run Tangent on SNP array data, we used raw probe-level intensity data in the form of .CEL files as input. These .CEL files are also available on the Genomic Data Commons (GDC) portal <https://portal.gdc.cancer.gov/legacy-archive/> as TCGA Level 1 Data (Supplementary Figure S1; Grossman *et al.* 2016). The Tangent pipeline for processing these Affymetrix SNP Array 6.0 data consists of six primary modules:

1. Generation of probeset intensities: The Affymetrix SNP Array 6.0 contain 906,600 SNPs that are associated with single nucleotide polymorphisms and 946,000 copy-number markers at other locations <http://tools.thermofisher.com/content/sfs/brochures/cn_snp_variation_technote.pdf>. Each genomic locus and SNP allele is represented by multiple probes on these arrays, which are collectively termed a probeset. The Affymetrix GeneChip Command Console Software generate probe-level intensities at each targeted locus and represent the intensity data in the form of .CEL files. We use SNPFileCreator, a Java implementation of the dChip signal intensity determination algorithm to normalize and merge intensity values for each probeset (Li & Wong 2001a, 2001b). Probe intensities across a sample are first scaled to achieve a median brightness value of 1000 and are subsequently subjected to quantile normalization. The normalized probe intensities of each sample are then mapped to a reference sample using model-

based expression indices. Median polish is applied to each probeset across samples to produce one value per probeset for each sample. We applied SNPFileCreator to all arrays in a batch, as defined by the joint PCR amplification step. The output of SNPFileCreator is available on the GDC portal as TCGA Level 2 Data (Supplementary Figure S1; Grossman *et al.* 2016).

2. Calibration of probesets to infer copy-number: In this module, probeset intensities are mapped to copy-number levels on a batch-by-batch basis, assuming a linear relationship between signal intensity and copy-number. This probeset calibration step is determined by two parameters: the background signal intensity and a scale factor that specifies the change in intensity resulting from each added copy of DNA. For SNP loci, Birdseed is used to calibrate probesets for each allele, using intensity data collected from normal samples, allele-specific background, and scale parameters (Korn *et al.* 2008). The resulting copy-numbers for the two alleles are summed to obtain total copy-number estimates. For CNV loci, we apply a copy-number inference step where the copy-number markers are calibrated on a two-step modeling approach based on SNP Array 6.0 data of an X-dosage experiment. This experiment was performed on 46 samples from five cell lines with , known CNVs on the X-chromosome. We first calibrate each probeset on chromosome X by applying linear regression to the experimental data to fit the parameters β_{i0} and β_{i1} of the following model (i):

(i) $$I_i = \beta_{i0} + \beta_{i1}\, C_i$$

whereby I_i represents the intensity of probeset i, C_i represents the copy-number level at probeset i, and β_{i0} and β_{i1} correspond to the background signal intensity and the scale factor that specifies the change in intensity resulting from each added copy

of DNA, respectively. We extend the resulting calibration for the probesets on chromosome X to the 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 summarized in model (ii) below:

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

for $k \in \{0, 1\}$. The variable GC_i represents the GC content of probeset i, $FL_i^{(sty)}$ and $FL_i^{(nsp)}$ represent the lengths of the Sty and Nsp fragments of probeset i 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. In the second step, we apply linear regression again 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 $k = 2$, this regression is performed collectively on the complete set of X-chromosome data, $\{\beta_{ik}, GC_i, FL_i^{(sty)}, FL_i^{(nsp)}, I_i^m\}_i \in \zeta_X$ where ζ_X is the collection of indices for the probesets on chromosome X. 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 SNP Array 6.0 data. While GC content and fragment lengths do not vary with the batch, the median intensity must be computed separately for each batch. The output files of this module are also available on the GDC portal as TCGA Level 2 Data (Supplementary Figure S1; Grossman *et al.* 2016).

3. Reduction of biological noise from germline variations: We remove probesets of regions that most likely contain germline CNVs, which are regions exhibiting extensive copy-number variations across the panel of normal samples (Supplementary Table 1).
4. Reduction of systematic noise using Tangent normalization: This module is described in detail in the following section. The standard reference plane used here was constructed using 3,146 TCGA SNP Array 6.0 normal blood samples that passed certain levels of quality control.
5. Reduction of random noise through segmentation: We apply the CBS algorithm to minimize the presence of random noise in the data (Seshan & Olshen 2019; Venkatraman & Olshen 2007). This module outputs segmented copy-number data in the form of .seg files. The output files are available on the GDC portal as TCGA Level 3 Data (Supplementary Figure S1; Grossman *et al.* 2016).
6. GISTIC 2.0: As part of the analyses, we also identify genes that are amplified or deleted more than expected by chance in the cohort using GISTIC 2.0 (Mermel *et al.* 2011). The results of this module are available on the GDC portal as TCGA Level 3 Data (Supplementary Figure S1; Grossman *et al.* 2016).

To run Tangent on NGS data, we use TCGA WES BAM files as input. We first apply the Genome Analysis Toolkit (GATK) 3 DepthOfCoverage tool on these BAM files to assess coverage information across the genome (Depristo *et al.* 2011). DepthOfCoverage processes BAM files to produce a .sample_interval_summary file that contains average coverage values for various genomic loci representing the hybrid capture targets, also termed intervals. Using these coverage data files as input, we run Tangent normalization (detailed in the section below) to reduce systematic noise, followed by CBS to reduce random noise in the data (Seshan & Olshen 2019;

Venkatraman & Olshen 2007). The flow charts for each type of input data are presented in Supplementary Figure S1A-B.

Tangent Normalization

A key assumption in Tangent normalization is that the distribution of log-transformed systematic noise follows a similar additive pattern in tumor samples as in normal samples. We made this assumption based on previous internal data that the observed noise in copy-number data is mostly multiplicative. Hence, in Tangent normalization, we can compute estimated noise profiles for tumors based on normal samples, and then subtract these estimated noise profiles from the input tumor copy-number data to minimize the presence of systematic noise. More specifically, we can model the space of log2 copy-ratios as:

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

where the noise space, $\boldsymbol{N}$, is a lower-dimensional subspace spanned by the coverage profiles of a collection of normal samples, as described in detail below, and its orthogonal complement, $\boldsymbol{N}^{\perp}$, is considered the signal space. Each tumor, T_j can be uniquely represented as:

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

where $noise(T_j)$ is the noise profile of the tumor T_j and is determined by the projection of T_j to $\boldsymbol{N}$, and signal(T_j) is a vector in $\boldsymbol{N}^{\perp}$. The $\boldsymbol{N}$ subspace is individually calculated for each tumor, and noise(T_j) can be computed through the use of pseudoinverse, as described below. Signal(T_j) is the residual of the difference between

T_j and noise(T_j), and is considered as the Tangent-normalized coverage profile of the tumor T_j.

In detail, 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 log2 copy-ratio intensities in genomic order, with each coordinate corresponding to one of the non-CNV probes. The noise space, $\boldsymbol{N}$, is defined as the (n_N - 1)-dimensional plane containing the vectors $\{N_1, N_2, N_3, \ldots, N_n\}$. Note that $n_N - 1 \ll M$, where M equals the dimension of the ambient log2 copy-ratio coordinate space or equivalently, the number of markers not excluded as poor quality or potential CNVs. For $j \in \{1, 2, 3, \ldots, n_T\}$, where n_T is the number of tumor samples, the j^{th} tumor sample is represented as vector T_j under a similar concept as N_i. To construct each unique normal profile such that it most closely matches the tumor noise profile, we first identify the point in $\boldsymbol{N}$ that is closest to T_j using a Euclidean metric, i.e. the projection of T_j on $\boldsymbol{N}$. We can compute $p(T_j)$ using standard linear algebra. A rigid transformation of Euclidean marker space prior to normalization does not change the resulting normalization of T_j. An appropriate translation of the Euclidean space ensures that $\boldsymbol{N}$ passes through the origin and forms a vector subspace of the Euclidean space, in which the normal vectors now reflect the deviation from the typical normal, which is what we consider as the noise in T_j. We can define the noise profile for each tumor sample as a linear combination of (n_N- 1) translated normal vectors after projection to $\boldsymbol{N}$:

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

where $\boldsymbol{N}$ is an array of columns corresponding to (n_N- 1) normal samples that span $\boldsymbol{N}$, and N_{pi} is the pseudo inverse of $\boldsymbol{N}$. The resulting normalization of T_j is then set to the residual, $T_j - p(T_j)$, which is the resulting Tangent-normalized tumor copy-number data.

Another assumption of Tangent normalization is that the normal samples provided to define the reference subspace do not contain SCNAs. In reality however, this may not actually be the case as normal samples obtained under clinical or experimental conditions may contain tumor cells and therefore may exhibit SCNAs. To address this potential issue, we designate a normal ceiling parameter for Tangent as the threshold by which normal samples are considered to be suspect normals, i.e. samples that contain tumor cells or SCNAs. We first calculate disruption scores for each normal sample, where a disruption score is defined as the mean absolute moving average of log2 copy-ratios across the genome. The normal ceiling parameter is thus a disruption score threshold such that samples whose disruption scores fall within the tail of the distribution are excluded from the constructed noise model in Tangent. In our experience, we find 0.073 to be a fair conservative threshold for TCGA SNP Array 6.0 data, and the range of 0.18 to 0.60 to be reasonable for TCGA WES data, taking into account that this value may vary based on the sequencing quality of normal samples.

Tangent normalization currently only includes the X chromosome and not the Y chromosome. Normalization of the X chromosome may be potentially complicated by two factors: (1) the distance from a tumor to a normal on the X chromosome should reflect the true difference in noise, rather than false positives from differences in sex; (2) since $p(T_j)$ may be a weighted average of copy-ratios from both male and female samples in our reference plane, the normalization of T_j could affect the apparent copy-number of the X chromosome. To address these potential issues, our reference plane includes a theoretical normal with copy-number precisely two throughout the autosomes and one throughout the X chromosome. The presence of this theoretical normal ensures that Tangent normalization can adjust the copy-number of X for any sample, regardless of the

input sex, to a mean level of ~2 copies. Doing so allows the detection of focal SCNAs within the X chromosome but ignores changes on a whole-chromosome level. To recover whole-chromosome SCNAs on the X chromosome, one could potentially run Tangent normalization using gender-matched normal samples, the results of which are not presented here.

We also encountered computational challenges from the large number of reference normal samples (~3,000 samples) as the projection matrix relies 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, by decomposing the reference plane into orthogonal blocks such 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 consists of approximately 250 normal samples from multiple sequencing batches to reflect diversity in the reference plane. 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.

Pseudo-Tangent

The Pseudo-Tangent pipeline is depicted in the flowchart in Supplementary Figure S1C. The first step of Pseudo-Tangent is running Tangent normalization on tumors using a small collection of normal samples that define the reference subspace. This step outputs a tentative copy-number profile for each tumor. We then subtract these tentative profiles from their original log-transformed tumor profiles to create a corresponding pseudo-normal copy-number profile for each tumor. These pseudo-normal profiles can

be used to define the reference subspace in the next step of Pseudo-Tangent. Specifically, the tumors are partitioned into n approximately equal subsets, and Tangent normalization is run on each separate subset with a reference subspace that is defined by pseudo-normal profiles in the complement of that subset. The partition parameter n is inversely related to the cardinality of each subset. We designate this partition parameter as the Nsplit parameter in the pipeline. Finally, similar to the original Tangent pipeline, the Pseudo-Tangent-normalized data is segmented using CBS to generate segmented copy-number profiles in the form of log2 copy-ratios (Seshan & Olshen 2019; Venkatraman & Olshen 2007).

A possible concern for Pseudo-Tangent is the potential to overfit when too many pseudo-normal profiles are provided, and thus resulting in false negatives during the detection of SCNAs. We provide an optional step in the Pseudo-Tangent pipeline to perform truncated singular value decomposition (tSVD) on the entire collection of pseudo-normal profiles prior to the partition step. The parameter Evects in the pipeline defines the number of SVD components for the decomposition of the pseudo-normal reference subspace. This step limits the dimensionality of the pseudo-normal reference subspace by setting the number of eigenvectors used to describe the noise distribution in the pseudo-normal profiles.

Other normalization approaches for performance comparison

We compared the performance of Tangent with several other conventional normalization approaches: (1) matched normal; (2) five nearest normals; (3) average normal; and (4) GC correction. In these comparison analyses, we opted to exclude the sex chromosomes such that differences in their handling of the sex chromosomes would not

affect the comparisons. 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 1).

The matched normal approach involves using a normal sample that is obtained from the same patient as the tumor. In this approach, we subtract the $\log_2$ copy-ratios of each matched normal from the corresponding tumor. For tumors with more than one matched normal, such as the case of a normal peripheral blood sample and a normal tissue sample, the matched blood sample is preferred over the tissue sample. The five nearest normals approach involves identifying five normals that are closest to the tumor based on Euclidean distance (Beroukhim *et al.* 2007). We then subtract the mean log2 copy-ratios of the five nearest normals from the tumor. In the average normal approach, we first generate a standard average normal by averaging the coverage at each targeted interval across the entire collection of normal samples. We then subtract the log2 copy-ratios of this computed average normal from each tumor. The approach of correcting for GC bias is performed based on the GC content normalization algorithm in HMMcopy (Ha *et* al. 2012; Lai *et al.* 2016). We then apply Tangent, the matched normal approach, or the average normal approach as previously described to the GC-corrected tumor and normal data.

We compare the performance of these normalization approaches in detecting SCNAs based on preservation of signal intensity, reduction in noise levels, improvement in signal-to-noise ratio (SNR), and level of hypersegmentation. When applicable, we also directly visualize gross log2 copy-ratios in Integrative Genomics Viewer (IGV) (Robinson *et al.* 2011; Thorvaldsdóttir *et al.* 2012). We define signal as the standard deviation of median signal intensities among all chromosomal arms, and noise as the median absolute difference between log2 copy-ratios of adjacent intervals or probes. The level of

hypersegmentation is estimated by the number of segments per sample after CBS segmentation on post-normalization data. We presume that greater noise in normalized copy-number data is reflected by increased fluctuations in true copy-number values, and thus after processing by a segmentation algorithm will result in larger number of segments per sample. In other words, we consider the performance of a normalization approach to be more superior if the average number of segments per sample is lower.

Tangent analysis on HCC1143 replicates data

We first studied the pattern of systematic noise that is present within and across batches in microarray data by processing HCC1143-BL cell lines on 118 arrays spanning 110 batches and HCC1143 tumor cell lines on 138 arrays across 128 batches. HCC1143 is a commercially available breast tumor cell line, whereas HCC1143-BL are immortalized non-cancerous peripheral blood lymphocytes originally obtained from the same patient, which we used as normal replicates in this study. We applied CBS segmentation on pre-normalized copy-number data and Tangent-normalized data for both the normal and tumor cell lines (Seshan & Olshen 2019; Venkatraman & Olshen 2007). We then visualized the segmented log2 copy-ratios on IGV as shown in Figure 1A and Supplementary Figure S2 (Robinson *et al.* 2011; Thorvaldsdóttir *et al.* 2012).

Tangent analysis on TCGA microarray data

In the analysis of TCGA SNP array data (the results of which are shown in Figure 2), we applied Tangent on 497 TCGA glioblastoma multiforme (GBM) tumors and 451 matched normal samples processed under similar experimental conditions as these tumors. We used 0.073 as the normal ceiling threshold, and constructed the reference plane using the reduced set of 451 matched normal samples instead of the collection of

over 3,000 normal samples as was used in the Tangent analysis of Affymetrix SNP Array 6.0 data for TCGA (see previous section). This reduced set of normal samples is to enable fair comparisons with other normalization approaches. We compared the performance of Tangent on this SNP array dataset to the matched normal approach and the five nearest normals approach, as described above.

We also evaluated the effect of size of reference plane on the performance of Tangent. We applied Tangent normalization on the same set of 497 GBM tumors using reference planes that were constructed from 0 (i.e. no use of Tangent) to 3,146 TCGA normal samples spread across 13 microarray batches. The median number of normal samples per batch was 255, with a range of 102 to 281 samples per batch. This collection of normal samples were normal blood leukocyte samples obtained from patients of various cancer types.

Finally, we investigated whether the normal samples that made up the reference plane should be processed under the same experimental conditions as the tumors. To achieve this goal, we studied the performance of Tangent using two reference planes: one was constructed from normal samples in the same array batches as the tumors, and the other was from normal samples across batches that may not necessarily be the same as the tumors. We compared the ratio of noise in post-normalization data to noise in pre-normalization data for performance analysis.

Tangent analysis on TCGA whole-exome data across cancer types

In the analysis of TCGA WES data (the results of which are shown in Figure 3), we applied Tangent on 252 TCGA WES samples across four distinct cancer types, i.e. stomach adenocarcinoma (STAD), lung squamous cell carcinoma (LUSC), lower grade glioma (LGG), and prostate adenocarcinoma (PRAD). The samples were made up of 123

tumors and 129 matched normal samples, including 123 normal blood and 6 normal tissue samples. The number of samples used in each of the four cancer types are detailed in Supplementary Table 2. We selected these samples and cancer types based on the availability of samples present on the same array plate in their SNP array counterparts and the representation across multiple TCGA sequencing centers. In this analysis, we used 0.23 as the normal ceiling threshold, and defined the reference subspace using the reduced set of 129 normal samples to enable fair comparisons with other normalization approaches. We compared the performance of Tangent normalization on this dataset with the matched normal approach, the average normal approach, and GC correction, as described above.

We also applied Tangent normalization on these 123 tumors using reference planes of various sizes. These reference planes were constructed from 10 to 1000 matched and non-matched normal samples in TCGA across ten different cancer types: cervical squamous cell carcinoma and endocervical adenocarcinoma (CESC), GBM, kidney renal clear cell carcinoma (KIRC), acute myeloid leukemia (LAML), LGG, LUSC, PRAD, rectum adenocarcinoma (READ), STAD, and uterine corpus endometrial carcinoma (UCEC). Each of the ten cancer types contributed equal number of samples to each reference plane that was constructed. The samples were also sequenced in five distinct sequencing centers to ensure diversity in noise profiles.

Pseudo-Tangent analysis on TCGA whole-exome CESC data

We introduced Pseudo-Tangent as an adaptation of Tangent to compensate for insufficient number of normal samples during the normalization step of copy-number profiling. In the Pseudo-Tangent analysis of TCGA WES data (results of which are shown

in Figure 4), we applied Pseudo-Tangent to 305 TCGA CESC primary tumor samples that were sequenced with whole-exome sequencing. We imposed an artificial constraint of only providing five matched normal samples for this analysis. The first step of Pseudo-Tangent was running regular Tangent normalization on these tumors using the five matched normal samples to define the reference subspace. With the tentative tumor copy-number profiles that were generated from the first step, we then created 305 signal-subtracted copy-number profiles, which we called pseudo-normal profiles. Using Nsplit of 3, we divided the tumors into three subsets, and applied Tangent on each subset using reference pseudo-normals in two other complementary subsets to generate the final tumor copy-number profiles.

To evaluate the performance of Pseudo-Tangent, we decided to compare the results of Pseudo-Tangent to gold-standard absolute allelic copy-number profiles that had been previously generated as an internal dataset at the Broad Institute (Taylor *et al*. 2018). This dataset was generated by applying the regular Tangent pipeline followed by the ABSOLUTE algorithm to Affymetrix SNP Array 6.0 data for these 305 CESC tumors and 3,146 matched and non-matched normal samples in TCGA (Carter *et al.* 2012). Note that the ABSOLUTE algorithm also uses mutation calls from corresponding WES data for these samples to optimize its tumor purity estimates in the generation of allelic copy-number and assessment of aneuploidy. Here, we employed a new definition of noise for this data, which was different from how we defined noise in the regular Tangent analysis as described previously. In our Pseudo-Tangent analysis, we measure noise as the average distance of the final copy-number profile for each segment and its nearest gold-standard absolute total copy-number for that particular segment. We specifically compared the performance of Pseudo-Tangent with the performance of regular Tangent

using only five matched normal samples as reference normals, which was also the first step for the Pseudo-Tangent pipeline.

Given the potential issue with overfitting, we next investigated the effect of the dimensions of the pseudo-normal reference subspace on the performance of Pseudo-Tangent. We performed tSVD of the pseudo-normal reference subspace to obtain 305 eigenvectors with their corresponding eigenvalues. Then, we applied Pseudo-Tangent on the tumors using reduced pseudo-normal reference subspaces that were constructed from 10 to 305 (i.e. the entire pseudo-normal reference subspace) eigenvectors selected by the greatest eigenvalues. The performance of Pseudo-Tangent with these various reference subspace dimensions is evaluated by comparing the noise levels in Pseudo-Tangent normalized data to those in data normalized with regular Tangent against five matched normal samples. We defined the ratio between the two datasets as the fraction of noise eliminated by Pseudo-Tangent.

Results

Principles of Tangent Normalization

The ultimate goal of this work is to develop a copy-number analysis pipeline to detect SCNAs in the cancer genome for TCGA. A major barrier to accurate detection of SCNAs is the presence of noise in copy-number data. As highlighted previously, there are three major sources of noise: random noise, germline CNVs, and systematic noise (see Introduction). In our initial empirical studies of systematic noise using HCC1143 and HCC1143-BL cell lines, we found that systematic noise, unlike random noise, tended to produce more consistent patterns of variation in the data both within and across microarray batches. A comparison of log2 copy-ratios for the HCC1143-BL normal replicates revealed several distinct patterns of CNV, with each CNV pattern being evident in multiple cell line replicates processed across many batches (Figure 1A). Similarly, we observed false positive patterns of SCNAs recurring across the tumor cell line replicates that were processed in different batches (Supplementary Figure S2A). These false positives can theoretically be eliminated by normalizing the tumor data against matched normal samples that were profiled under identical experimental conditions. However, in practice it is almost impossible to identify the nuances of many of these experimental conditions, let alone measure them. Given the need of a more accurate control to reduce systematic biases, we developed the Tangent normalization method to build a subset of normal controls that can most accurately represent the tumor noise profile.

We postulate that systematic variations in signal intensity or coverage profiles are due to factors in experimental conditions. Therefore, normal samples that are processed

in the same batch or sequencing center as the tumors should have similar noise characteristics as their tumor counterparts. Tangent hypothesizes that a large collection of diploid normal samples should be sufficient to encompass a subspace, $\boldsymbol{N}$, of the larger landscape of all possible variations in signal intensity or coverage profiles. As such, Tangent uses this collection of normal samples with presumed similar noise profiles as the tumors to construct the reference subspace $\boldsymbol{N}$ from the space that spans all linear combinations of normal profiles. For any copy-number profile T_j of a tumor sample j, the point in subspace $\boldsymbol{N}$ that is most similar to T_j should represent the normal sample that is processed under similar conditions as the tumor j. We can now identify SCNAs as the difference between T_j and the nearest point in subspace $\boldsymbol{N}$ (Figure 1B). In other words, we can project T_j onto the reference subspace $\boldsymbol{N}$, and then estimate SCNAs from the residual difference between T_j and the projection of T_j (Figure 1C). Further details of Tangent normalization are described in the Methods section.

In addition to addressing systematic noise, our pipeline also included a blacklist of germline CNVs (Supplementary Table 1), and a CBS segmentation step to reduce random noise (Seshan & Olshen 2019; Venkatraman & Olshen 2007). We found a combination of the above approaches in the pipeline to be sufficient in reducing the majority of noise that is present in copy-number data. The overall copy-number analysis pipeline for Tangent is depicted in Supplementary Figure S1A-B.

Tangent analysis on microarray data

We assessed the performance of Tangent in inferring copy-number profiles from microarray data by applying Tangent to a TCGA SNP array dataset of GBM tumors (see Methods). We compared the performance of Tangent to two other normalization

approaches: the matched normal approach and the five nearest normals approach (Beroukhim *et al.* 2007).

We found that the signal integrity in copy-number profiles was preserved for all three normalization approaches (Figure 2A). However, when compared to the other two approaches, only Tangent normalization consistently led to lower noise levels in the normalized data, thus exhibiting the highest signal-to-noise ratio. The improvement in noise levels for the five nearest normals approach was negligible, whereas the matched normal approach resulted in greater noise level post-normalization (Figure 2B-C). We also assessed the level of hypersegmentation in segmented normalized copy-number profiles, and found the mean number of segments for Tangent normalization to be 23% less than that of the five nearest normals approach (Supplementary Figure S3).

Next, we studied the effect of reference plane size on the performance of Tangent in noise reduction. We applied Tangent to the same set of GBM tumors but this time using reference planes of increasing sizes to a maximum of 3,146 normal samples. We found a steady reduction in median noise levels with increasing number of normal samples, but the improvement decreased asymptotically and offered negligible benefits after approximately four batches, which at a median of 255 samples per batch corresponded to approximately 1,000 normal samples (Figure 2D).

Finally, we investigated the effect of reference plane composition on the performance of Tangent in noise reduction. Specifically, we asked the question of whether the normal samples provided should be processed in the same or different batches of arrays as the tumors. We found that in both cases—for normal samples within or across batches—the average amount of noise in Tangent-normalized data is lower than that pre-normalization. However, the degree of noise reduction is greater when the

reference plane was constructed from the entire set of normal samples across batches, as compared to the case of normal samples within the same batch as the tumors (Figure 2E).

Tangent analysis on whole-exome sequencing data

In the case of whole-exome sequencing data, we evaluated the performance of Tangent by applying Tangent normalization to a TCGA WES dataset of 123 tumors across four tumor types (see Methods; Supplementary Table 2). We compared the performance of Tangent to two other normalization approaches: the matched normal approach and the average normal approach. We also combined each of the three approaches with correction for GC biases to understand whether GC correction provides significant improvements beyond Tangent normalization (Ha *et al.* 2012).

We found that Tangent outperformed these conventional normalization approaches. The amount of noise present in Tangent-normalized data was the lowest when compared to the matched normal and average normal approach (Figure 3A). Similarly, Tangent normalization also led to the highest signal-to-noise ratio (Figure 3B). When GC correction is applied, the correction reduced the amount of noise present in data that was normalized with the matched normal approach and the average normal approach. However, the improvement in noise levels by GC correction is negligible for Tangent normalization (Figure 3A). These trends remained when we assessed for signal-to-noise ratio, suggesting that the benefit of correcting for GC biases in addition to Tangent normalization may be marginal (Figure 3B).

Next, we again studied the effect of reference plane size on the performance of Tangent in handling WES data. We applied Tangent to the same 123 tumors using reference planes constructed from up to 1,000 normal samples across ten cancer types

and sequenced in multiple distinct centers. We found a steady asymptotic decrease in median noise levels with increasing number of normal samples, and the performance of Tangent plateaued at approximately 500 normal samples (Figure 3C-D).

Pseudo-Tangent: an approach to compensate for insufficient normal data

The success of Tangent normalization hinges upon the presence of a sufficiently large and diverse collection of normal samples to define the reference subspace. However, due to various aforementioned reasons, in reality it is often challenging to sequence a large collection of normal samples to provide a wide landscape of noise profiles matching that of the tumor samples. To compensate for this issue, we developed Pseudo-Tangent as an adaptation of the Tangent pipeline to generate signal-subtracted tumor profiles, also known as pseudo-normal profiles, as a way of augmenting the reference subspace. Details of the Pseudo-Tangent pipeline are described in the Methods section (Supplementary Figure S1C).

In our analysis, we applied Pseudo-Tangent on 305 WES CESC primary tumors in TCGA, and created an artificial constraint of only using five matched normal samples. We benchmarked the performance of Pseudo-Tangent against gold-standard copy-number data inferred from TCGA microarrays using regular Tangent normalization and the ABSOLUTE algorithm (Carter *et al.* 2012). Noise is defined as the difference between our results and the gold-standard dataset. We compared the level of noise in Pseudo-Tangent-normalized data and Tangent-normalized data using only five normal samples to define the reference subspace. We found that the noise levels in all of the tumors were

lower after Pseudo-Tangent normalization than after Tangent normalization with only five matched normal samples (Figure 4A).

Next, due to concerns of overfitting resulting in loss of signal, we investigated whether reducing the dimensions of the pseudo-normal reference subspace could improve the performance of Pseudo-Tangent. To answer this question, we applied Pseudo-Tangent on the same set of WES CESC tumors but this time with varying number of eigenvectors that limit the dimensions of the pseudo-normal reference subspace. We discovered that there was an optimal number of eigenvectors one could use in defining the pseudo-normal reference subspace, and this number was highly dependent on the underlying tumor copy-number profile. In our dataset, we found the fraction of noise eliminated by Pseudo-Tangent to be the highest when the number of eigenvectors used was 150 (Figure 4B). Given that the eigenvectors selected were among those with the greatest eigenvalues, they captured approximately 98% of the variance of the entire pseudo-normal reference subspace. However, when we looked into individual tumor data, we found that the optimal number of eigenvectors varied across tumors. For instance, the optimal number of eigenvectors for tumors containing the greatest amount of noise in their copy-number data was closer to ten rather than 150 or other higher numbers (Figure 4C). This finding suggests that the optimal dimensionality of the pseudo-normal reference subspace may vary based on the underlying noise levels in the tumor copy-number data.

Availability and implementation of Tangent

The Tangent pipeline is the basis for the copy-number analyses of all Affymetrix SNP Array 6.0 data in TCGA. The results of these copy-number analyses are available as

CNV data in the GDC Data Portal <https://portal.gdc.cancer.gov/repository> (Grossman *et al.* 2016). Tangent normalization is the foundation for copy-number normalization in the GATK4 CNV Discovery workflow that is offered by the Broad Institute (McKenna *et al.* 2010). We have also made Tangent publicly available for downloads and implementation both through Github <https://github.com/broadinstitute/tangent> and as a Docker image <https://hub.docker.com/r/coyin/tangent>.

Discussions, Conclusions and Future Work

In this thesis, we presented Tangent as a copy-number inference pipeline to accurately detect SCNAs in the cancer genome. The Tangent pipeline highlighted the use of Tangent normalization as a novel approach in reducing the presence of systematic noise from various sources in copy-number data. By creating a subset of normal samples with a similar noise profile as the tumor in question, we can normalize the tumor copy-number data against this reference profile and reduce the amount of systematic noise. Our initial studies of noise patterns through HCC1143-BL normal replicates provided evidence that a noise model built from normal samples processed across many batches would best represent the diversity in systematic noise and thus facilitate optimal noise reduction. This principle became the foundation for normalization with the Tangent reference subspace, which is defined by the subset of normal samples that best represents the noise profile most similar to the tumor.

Tangent was originally developed for SNP array data as part of the copy-number analyses work in TCGA. In our analysis of TCGA SNP array data using GBM samples, the performance of Tangent was found to be superior to other normalization approaches, i.e. the matched normal approach and the five nearest normals approach. Tangent normalization consistently resulted in the lowest amount of noise in post-normalization data without compromising significantly in signal intensity, therefore providing the highest signal-to-noise ratio. Tangent normalization also led to the least amount of hypersegmentation, which could be a proxy for the degree of local fluctuations in normalized data. Our analysis showed that the performance of Tangent was dependent

on two factors: the size and diversity of the normal reference subspace. To optimize the performance of Tangent, one can increase the size of the reference plane by including more normal samples. One can also increase the reference diversity by including normal samples processed under a larger variety of conditions. Given the success of Tangent normalization, we implemented the Tangent pipeline as a means of inferring copy-number for all SNP Array data for over 10,000 tumors in TCGA.

With the advancement in next-generation sequencing, we then further developed the Tangent pipeline to be applied on WES data in TCGA. In our analysis of TCGA WES data using samples from multiple cancer types, the performance of Tangent in noise reduction was again superior to the normalization approaches of using a matched normal and an average normal. Meanwhile, while the implementation of GC correction improved the performance of the latter two, GC correction did not provide substantial benefit in noise reduction when applied in addition to Tangent normalization. The superior performance of Tangent likely stemmed from its ability to provide a reference normal that most appropriately encompasses the landscape of noise present in the tumor data as compared to other approaches of normalization. The benefit of GC correction with Tangent is marginal probably because the normalization step in Tangent has already corrected for systematic noise from GC biases. With WES data, the performance of Tangent also improved with the use of larger reference planes.

Although Tangent was initially developed for use with SNP array data, we have now shown that Tangent can be extended to WES data for copy-number analyses. In fact, by its basic principle of normalization, Tangent can be applied to any source of copy-number data that measures DNA dosage with varying signal intensity or depth of coverage, such as whole-genome sequencing (WGS) or microarray-based CGH. Future

work could focus on optimizing Tangent for the use of WGS data, which has become increasingly common with the decreasing cost of NGS technologies.

Regardless of the type of sequencing technologies, accurate determination of SCNAs heavily hinges upon the presence of normal controls that have been processed in identical fashion to the tumors. Since it is almost impossible in reality to obtain such normal control samples, Tangent normalization provides a computational approach to construct a reference control from a collection of normal samples to provide an estimated noise profile resembling that of the tumor. Therefore, the presence of a sufficiently large and diverse normal reference subspace is critical to the success of Tangent. In this thesis, we presented Pseudo-Tangent as an adaptation of the Tangent pipeline to augment the existing reference subspace. Pseudo-Tangent aims to enrich the Tangent reference subspace through the use of pseudo-normal profiles, which are signal-subtracted tumor profiles that no longer contain SCNAs.

We tested our Pseudo-Tangent pipeline on TCGA WES CESC tumors by imposing the constraint of having limited normal samples. In our analysis, the use of pseudo-normal profiles resulted in lower noise levels than regular Tangent normalization with the five normal samples provided. While these results were encouraging, we were concerned about potential issues with overfitting resulting in false negatives. The issue of overfitting is more likely to occur if SCNAs in the tumors were not adequately subtracted during the generation of pseudo-normal profiles. Further analysis work with Pseudo-Tangent showed that we could reduce the dimensions of the pseudo-normal reference subspace to an optimal number that would preserve the integrity of the reference subspace and maximize the performance of Pseudo-Tangent. We also found that this number is affected by the underlying noise levels in the input tumor copy-number data, and thus designated it as a parameter in the Pseudo-Tangent pipeline. We

emphasize that Pseudo-Tangent is most suited for analyses in which there are insufficient number of normals to form a diverse reference subspace. In situations where sufficient true normal samples are available, extensive profiling of these normal samples to define the reference subspace is preferable to the computational generation of pseudo-normal profiles.

In addition to Pseudo-Tangent, there are a few other potential ways to improve the performance of Tangent. While Tangent has primarily been applied to detecting SCNAs from genomic data, in the space of gene expression analysis, Fehrmann et al. developed a similar method to detect SCNAs from transcriptomic profiling data. Specifically, they removed principal components describing known transcriptional states to enrich for transcriptional changes that reflected the underlying SCNAs (Fehrmann *et al.* 2015). A few other published works in literature have also demonstrated improvements in copy-number profiling through novel algorithms that determine differences in absolute copy-number rather than relative copy-numbers by incorporating information about tumor purity and ploidy (Carter *et al.* 2012; Ha *et al.* 2014; Van Loo *et al.* 2010). However, as all of these algorithms require normalized copy-number ratios as inputs, we believe the use of these algorithms may still benefit greatly from Tangent normalization. Finally, another way of improving the accuracy of Tangent in detecting SCNAs may be to incorporate information about copy-number breakpoints through the detection of rearrangements by WGS (Drier *et al.* 2013; Layer *et al.* 2014; Rausch *et al.* 2012; Wala *et al.* 2018).

In summary, Tangent normalization is a novel and effective approach in reducing systematic noise for copy-number analyses of the cancer genomes. The Tangent pipeline incorporates Tangent normalization to accurately infer somatic copy-number alterations in cancer. This pipeline can be applied to both microarray data and whole-exome

sequencing data, and has the potential to be adapted for other sources of sequencing data to meet a variety of research and clinical needs. In situations where sufficient number of normal samples may not be readily available, Pseudo-Tangent can be applied to computationally generate pseudo-normal profiles to enrich the reference subspace. Future work in improving the performance of Tangent by incorporating alternative sources of information such as genomic rearrangements, tumor purity and ploidy can be considered. Currently, copy-number data from over 10,000 tumors in TCGA have been analyzed using the Tangent pipeline and published on the GDC data portal (Grossman *et al.* 2016). We have also made the Tangent normalization approach publicly available via GATK4 and the Tangent whole-exome sequencing pipeline available on Github (McKenna *et al.* 2010).

References

Beroukhim R, Getz G, Nghiemphu L, Barretina J, Hsueh T, Linhart D, et al. Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma. Proc Natl Acad Sci U S A. 2007 Dec;104(50):20007–12.

Beroukhim R, Mermel CH, Porter D, Wei G, Raychaudhuri S, Donovan J, et al. The landscape of somatic copy-number alteration across human cancers. Nature. 2010;463(7283):899–905.

Boeva V, Popova T, Bleakley K, Chiche P, Cappo J, Schleiermacher G, et al. Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data. Bioinformatics. 2012 Feb;28(3):423–5.

Carter SL, Cibulskis K, Helman E, McKenna A, Shen H, Zack T, et al. Absolute quantification of somatic DNA alterations in human cancer. Nat Biotechnol. 2012 May;30(5):413–21.

DePristo MA, Banks E, Poplin R, Garimella K V, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011 May;43(5):491–8.

Drier Y, Lawrence MS, Carter SL, Stewart C, Gabriel SB, Lander ES, et al. Somatic rearrangements across cancer reveal classes of samples with distinct patterns of DNA breakage and rearrangement-induced hypermutability. Genome Res. 2013 Feb;23(2):228–35.

Fehrmann RSN, Karjalainen JM, Krajewska M, Westra H-J, Maloney D, Simeonov A, et al. Gene expression analysis identifies global gene dosage sensitivity in cancer. Nat Genet. 2015 Feb;47(2):115–25.

Grizzle WE, Bell WC, Sexton KC. Issues in collecting, processing and storing human tissues and associated information to support biomedical research. Cancer Biomark. 2010;9(1–6):531–49.

Grossman RL, Heath AP, Ferretti V, Varmus HE, Lowy DR, Kibbe WA, et al. Toward a Shared Vision for Cancer Genomic Data. N Engl J Med. 2016 Sep;375(12):1109–12.

Ha G, Roth A, Khattra J, Ho J, Yap D, Prentice LM, et al. TITAN: inference of copy number architectures in clonal cell populations from tumor whole-genome sequence data. Genome Res. 2014 Nov;24(11):1881–93.

Ha G, Roth A, Lai D, Bashashati A, Ding J, Goya R, et al. Integrative analysis of genome-wide loss of heterozygosity and monoallelic expression at nucleotide resolution reveals disrupted pathways in triple-negative breast cancer. Genome Res. 2012 Oct;22(10):1995–2007.

Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res. 2012 Mar;22(3):568–76.

Korn JM, Kuruvilla FG, McCarroll SA, Wysoker A, Nemesh J, Cawley S, et al. Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. Nat Genet. 2008 Oct;40(10):1253–60.

LaFramboise T. Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances. Nucleic Acids Res. 2009 Jul;37(13):4181–93.

Lai D, Ha G, Shah S. HMMcopy: Copy number prediction with correction for GC and mappability bias for HTS data. 2019. R package version 1.28.0.

Layer RM, Chiang C, Quinlan AR, Hall IM. LUMPY: a probabilistic framework for structural variant discovery. Genome Biol. 2014;15(6):R84.

Li C, Hung Wong W. Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application. Genome Biol. 2001a;2(8):RESEARCH0032.

Li C, Wong WH. Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. Proc Natl Acad Sci U S A. 2001b Jan;98(1):31–6.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010 Sep;20(9):1297–303.

Mermel CH, Schumacher SE, Hill B, Meyerson ML, Beroukhim R, Getz G. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. Genome Biol. 2011;12(4):R41.

Network TCGAR. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature. 2008 Oct;455(7216):1061–8.

Nilsen G, Liestøl K, Van Loo P, Moen Vollan HK, Eide MB, Rueda OM, et al. Copynumber: Efficient algorithms for single- and multi-track copy number segmentation. BMC Genomics. 2012 Nov 4;13:591.

Rausch T, Zichner T, Schlattl A, Stütz AM, Benes V, Korbel JO. DELLY: structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics. 2012 Sep 15;28(18):i333–9.

Rieber N, Bohnert R, Ziehm U, Jansen G. Reliability of algorithmic somatic copy number alteration detection from targeted capture data. Bioinformatics. 2017 Sep;33(18):2791–8.

Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, et al. Integrative genomics viewer. Nat Biotechnol. 2011;29(1):24–6.

Sathirapongsasuti JF, Lee H, Horst BAJ, Brunner G, Cochran AJ, Binder S, et al. Exome sequencing-based copy-number variation and loss of heterozygosity detection: ExomeCNV. Bioinformatics. 2011 Oct 1;27(19):2648–54.

Seshan V, Olshen A. DNAcopy: DNA copy number data analysis. 2019. R package version 1.60.0.

Talevich E, Shain AH, Botton T, Bastian BC. CNVkit: Genome-Wide Copy Number Detection and Visualization from Targeted DNA Sequencing. PLoS Comput Biol. 2016 Apr;12(4):e1004873.

Taylor AM, Shih J, Ha G, Gao GF, Zhang X, Berger AC, et al. Genomic and Functional Approaches to Understanding Cancer Aneuploidy. Cancer Cell. 2018 Apr;33(4):676-689.e3.

Thorvaldsdóttir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform. 2012 Apr 19;14(2):178–92.

Van Loo P, Nordgard SH, Lingjaerde OC, Russnes HG, Rye IH, Sun W, et al. Allele-specific copy number analysis of tumors. Proc Natl Acad Sci U S A. 2010 Sep;107(39):16910–5.

Venkatraman ES, Olshen AB. A faster circular binary segmentation algorithm for the analysis of array CGH data. Bioinformatics. 2007 Mar;23(6):657–63.

Wala JA, Bandopadhayay P, Greenwald NF, O'Rourke R, Sharpe T, Stewart C, et al. SvABA: genome-wide detection of structural variants and indels by local assembly. Genome Res. 2018 Apr;28(4):581–91.

Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, Ellrott K, et al. The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet. 2013 Oct;45(10):1113–20.

Weir B, Zhao X, Meyerson M. Somatic alterations in the human cancer genome. Cancer Cell. 2004 Nov;6(5):433–8.

Yoon S, Xuan Z, Makarov V, Ye K, Sebat J. Sensitive and accurate detection of copy number variants using read depth of coverage. Genome Res. 2009 Sep;19(9):1586-92.

Zack TI, Schumacher SE, Carter SL, Cherniack AD, Saksena G, Tabak B, et al. Pan-cancer patterns of somatic copy number alteration. Nat Genet. 2013 Oct;45(10):1134–40.

Zare F, Dow M, Monteleone N, Hosny A, Nabavi S. An evaluation of copy number variation detection tools for cancer using whole exome sequencing data. BMC Bioinformatics. 2017 May 31;18(1):286.

Zhao M, Wang Q, Wang Q, Jia P, Zhao Z. Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives. BMC Bioinformatics. 2013;14(11):S1.

Figures

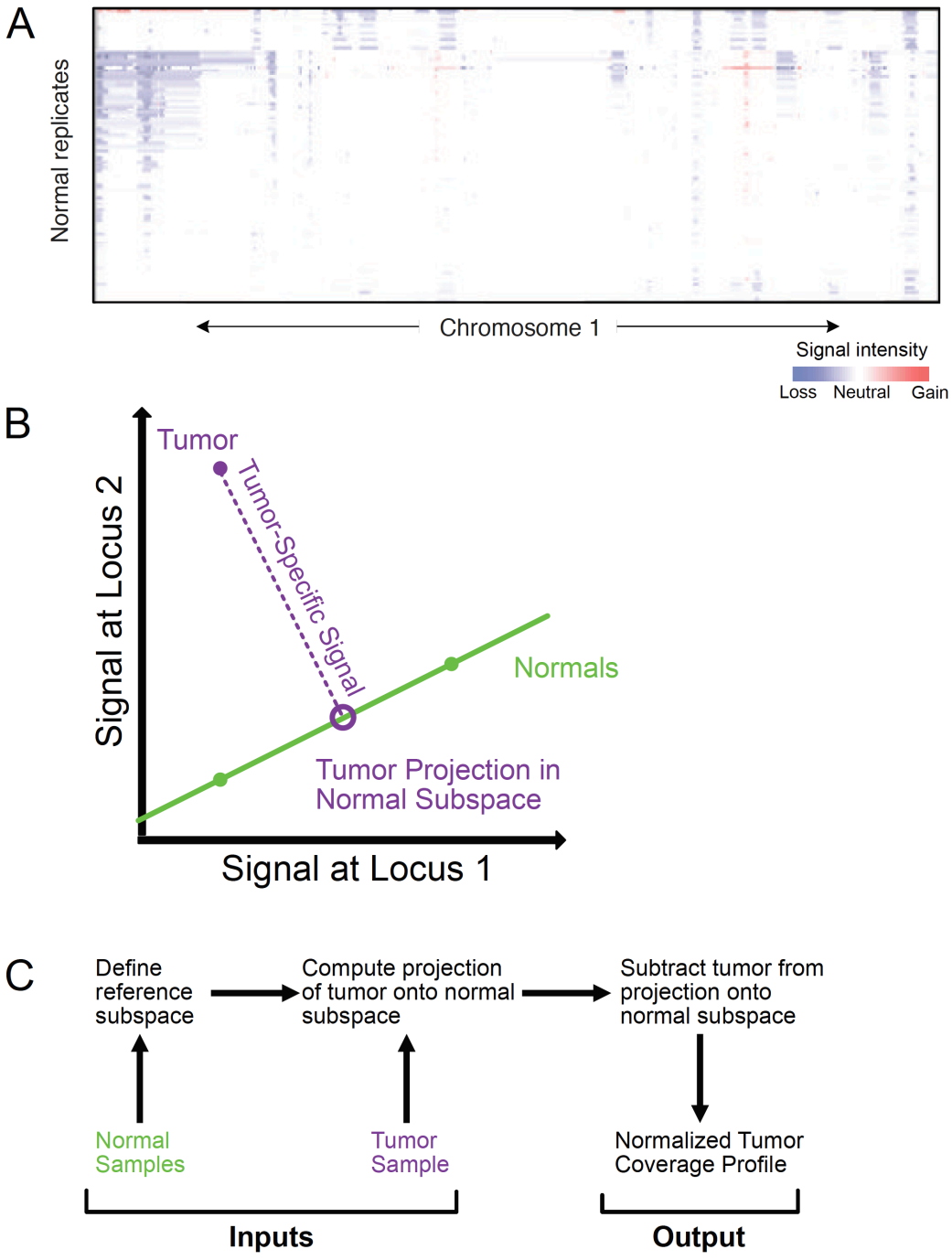


Figure 1. Overview of problem and method. (a) Variations in segmented, pre-normalized log2 copy-number ratios on chromosome 1 (as visualized in IGV) of HCC1143-BL normal replicates across 110 batches. (b) A reduced, two-dimensional representation of the

Tangent normalization approach. For each tumor (purple), we compute its projection onto a lower-dimensional subspace defined by normal samples (green) profiled in parallel with the tumors. True signal for SCNAs is computed as the difference between the tumor and its projection onto the normal space. (c) Flowchart depicting the steps in Tangent normalization.

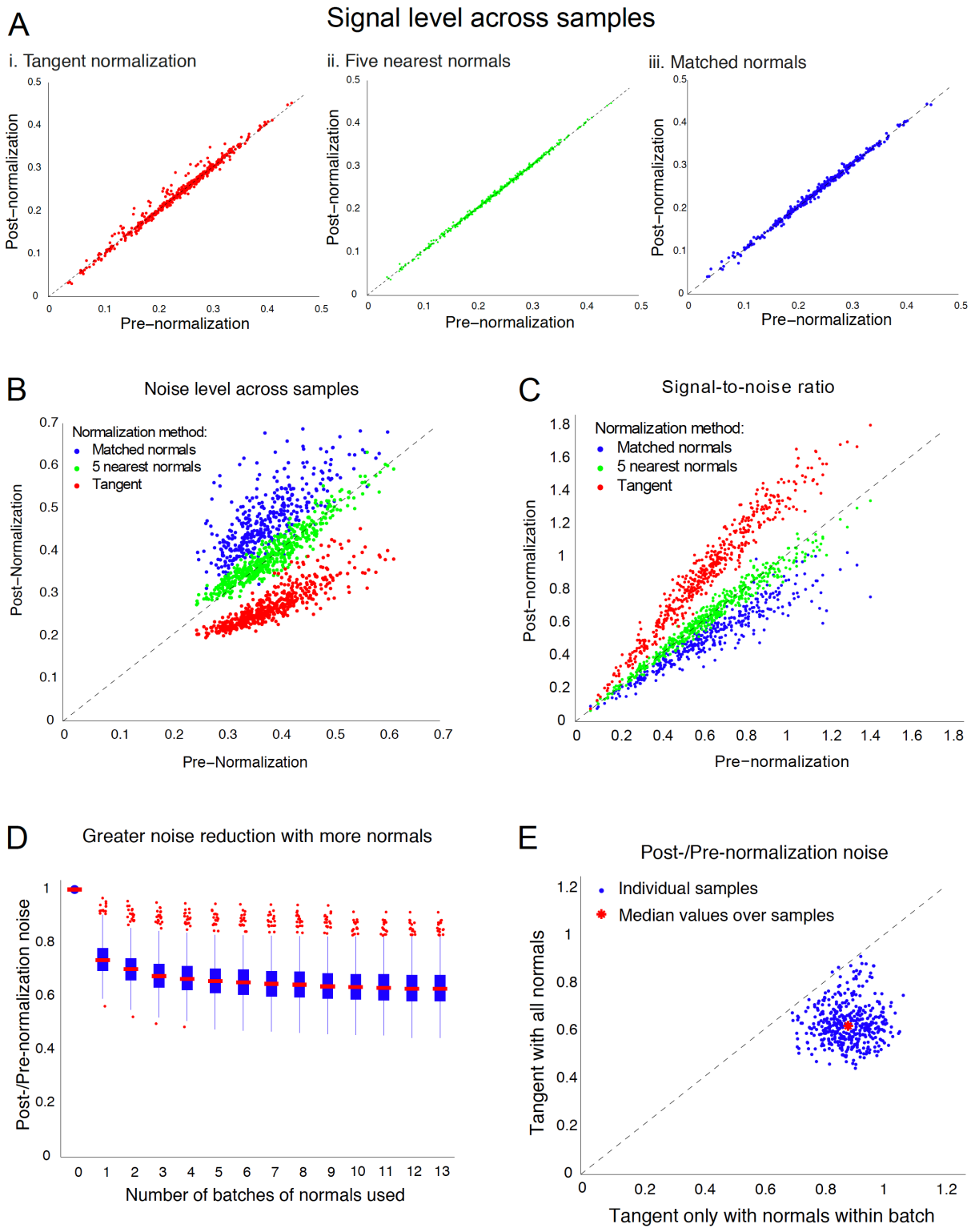


Figure 2. Performance of Tangent on SNP array data from 497 TCGA glioblastoma tumors. Scatterplots in panels A-C indicate post-normalization vs. pre-normalization (a)

signal, (b) noise level, and (c) signal-to-noise ratios for each of the three normalization approaches: Tangent (red), five nearest normals (green), and matched normals (blue). (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 per batch). (e) Noise ratio (post-normalization to pre-normalization noise) for glioblastoma samples following Tangent normalization using the entire reference plane vs. Tangent normalization using only the normal samples processed in the same batch as the tumor.

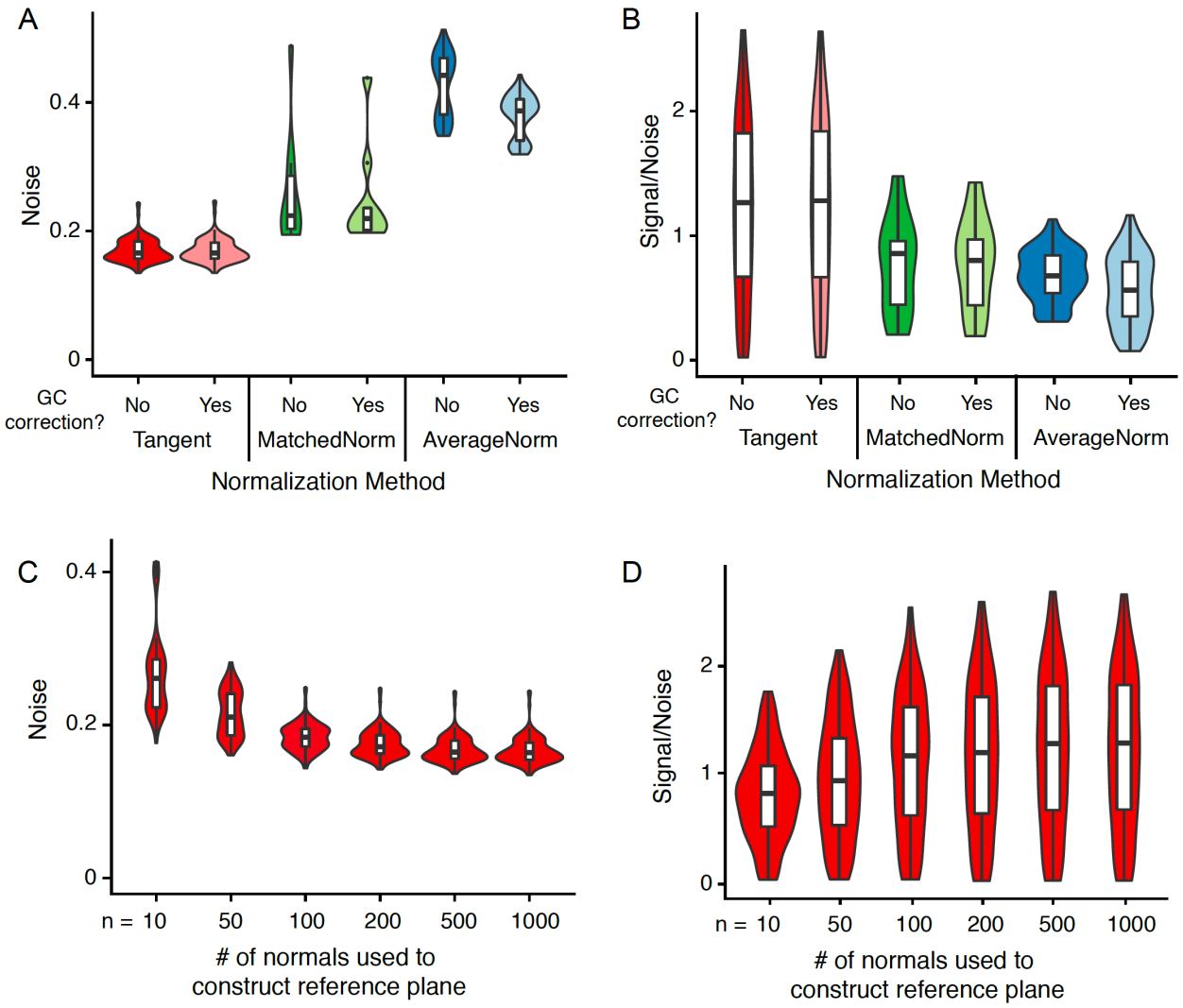


Figure 3. Performance of Tangent on WES data from 252 TCGA tumors across multiple cancer types. (a) Noise levels and (b) signal-to-noise ratios for each of the three normalization approaches: Tangent (red), matched normals (green; MatchedNorm), average normal (blue; AverageNorm). Also shown here is each normalization approach with and without additional GC correction. (c) Noise and (d) signal-to-noise ratios following Tangent normalization with increasing numbers of normal samples in the reference subspace.

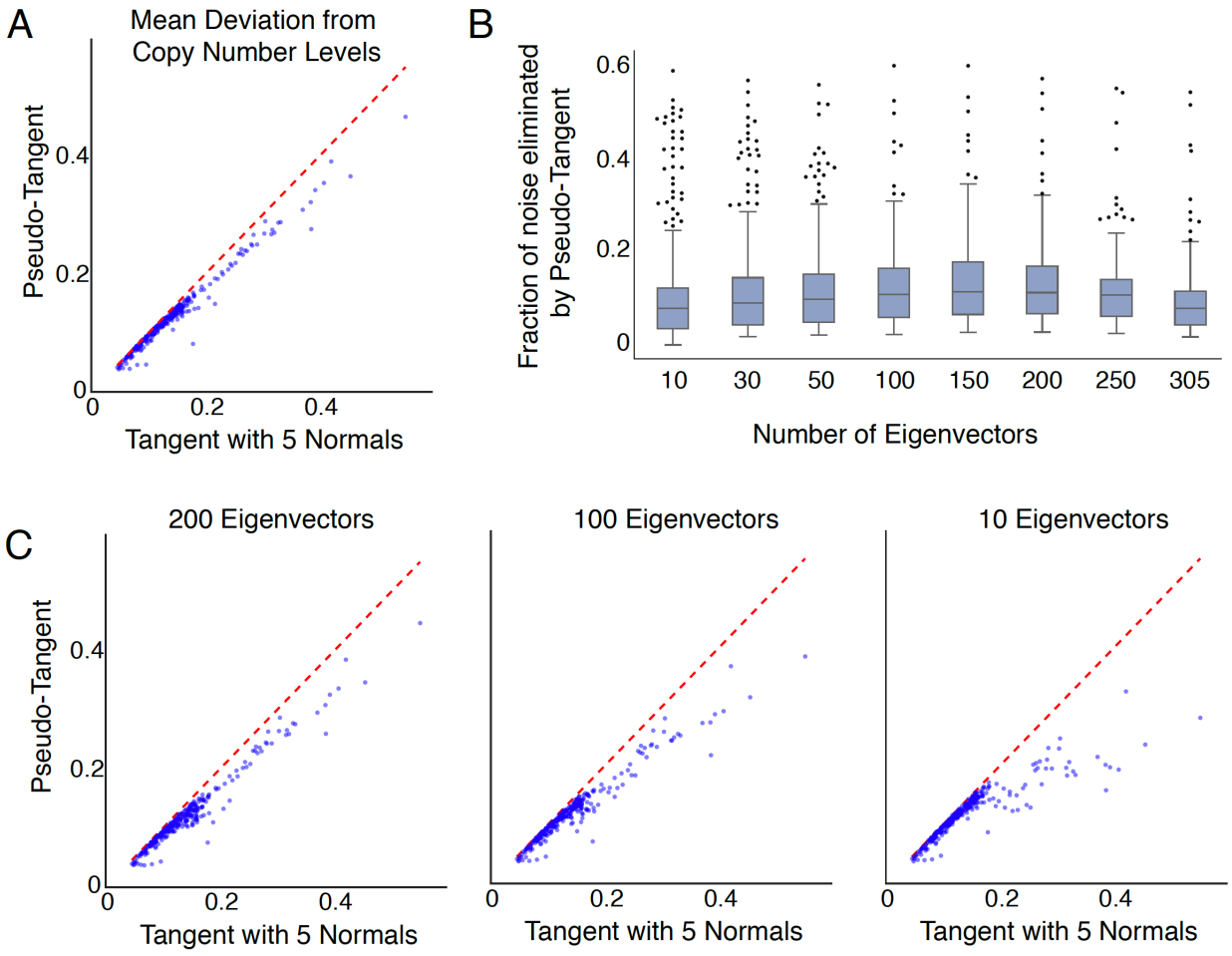
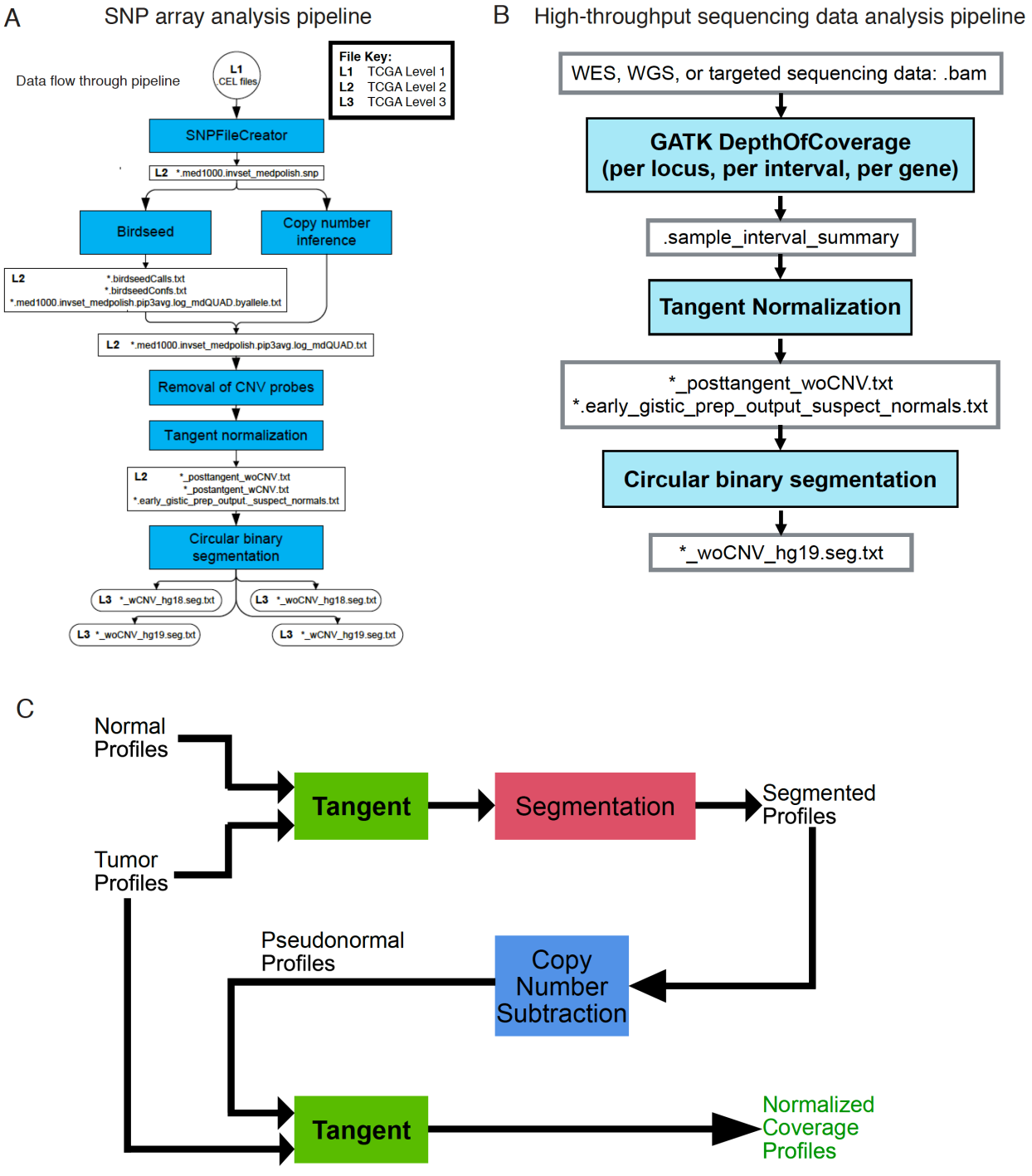


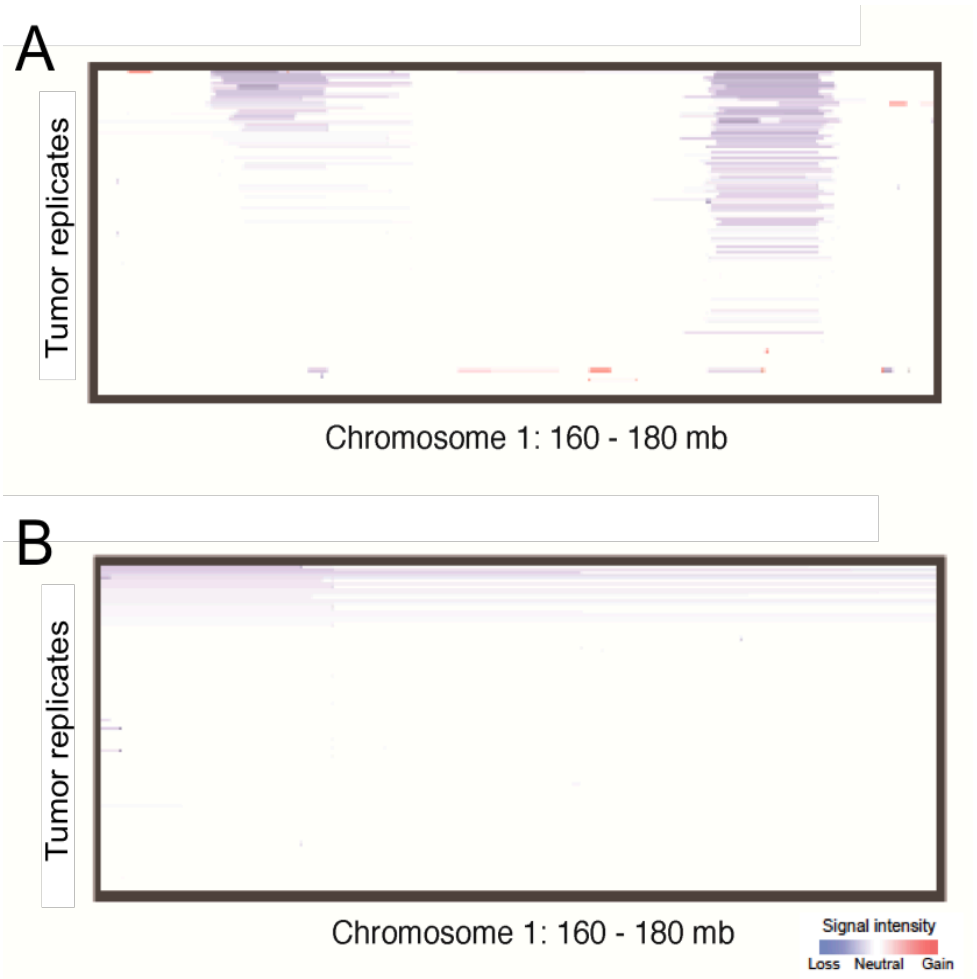
Figure 4. Performance of Pseudo-Tangent on WES data from 306 TCGA CESC tumors. (a) Average deviation from gold-standard ABSOLUTE-estimated copy-number levels, also defined as noise, following Pseudo-Tangent vs. Tangent alone with five matched normals. (b) Fraction of noise eliminated by Pseudo-Tangent vs. regular Tangent after the use of varying numbers of eigenvectors obtained from the decomposition of the pseudo-normal reference subspace. (c) Noise levels following Pseudo-Tangent vs. Tangent alone after the use of 200 eigenvectors (left panel), 100 eigenvectors (middle panel), and 10 eigenvectors (right panel).

Supplementary Figures

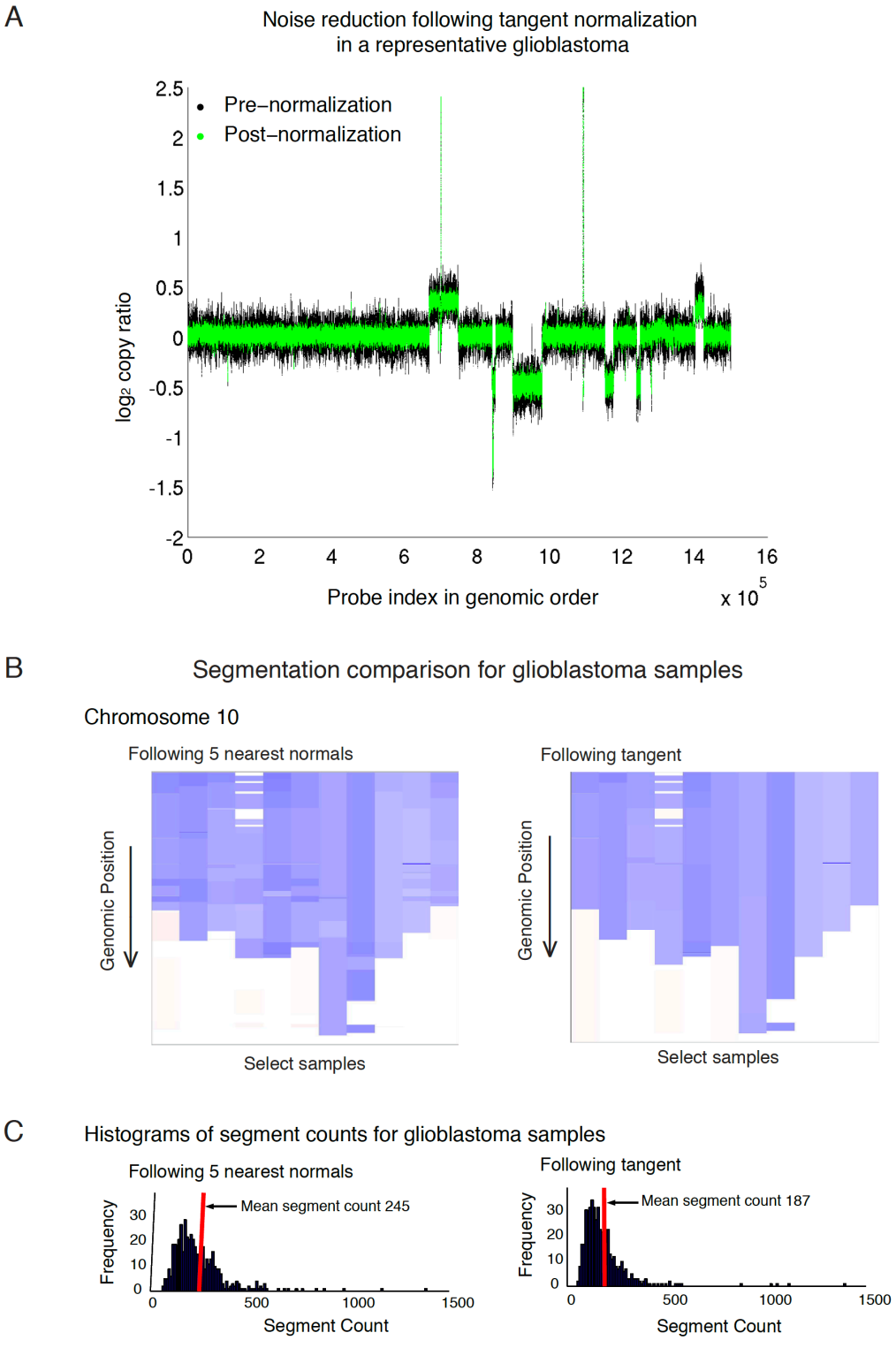


Supplementary Figure S1. (a) Overview of the Tangent copy-number inference pipeline for SNP array data. Schematic view of pipeline with key output files for pipeline modules

and assigned TCGA levels included. (b) Overview of Tangent copy-number inference pipeline for next-generation sequencing data. (c) Overview of the Pseudo-Tangent pipeline as an adaptation of the Tangent pipeline.



Supplementary Figure S2. (a) Segmented, pre-normalized log2 copy-number ratios in a region on chromosome 1 containing 160-180 megabases (as visualized in IGV) of HCC1143 tumor replicates across 128 batches. (b) Segmented Tangent-normalized log2 copy-number ratios in the same region on chromosome 1 for the same tumor replicates. The systematic artifacts present in (a) are no longer observed.



Supplementary Figure S3. Performance of Tangent on SNP array data for representative GBM tumors. (a) Log2 copy-number ratios, shown here as a 100-marker moving average

across the autosomes of the genome, following Tangent normalization (green) vs. pre-Tangent normalization (black) in a representative glioblastoma tumor. (b) Segmented copy-number ratios (as visualized in IGV) for selected glioblastoma samples following two normalization approaches: five nearest normals (left panel) and Tangent (right panel). White is copy-neutral, whereas blue indicates a deletion with intensity of color increasing as copy-number decreases. (c) Levels of hypersegmentation in resulting copy-number data, as shown by the histograms of total segment counts per sample, for 497 TCGA GBM tumor samples when CBS follows two normalization approaches: five nearest normals (left panel) and Tangent (right panel).

Supplementary Tables

Supplementary Table 1. Regions excluded from copy-number analyses.

Region ID	Chromosome	Start	End	Flanking start	Flanking end
CNV.1	1	61735	3214732	61735	3218329
CNV.2	1	3318977	3326796	3318714	3326803
CNV.3	1	3785461	3792719	3781042	3792801
CNV.4	1	4291287	4302161	4286684	4308903
CNV.5	1	4422005	4424098	4421072	4424145
CNV.6	1	4737693	4746636	4737605	4746774
CNV.7	1	4756432	4759428	4754057	4760040
CNV.8	1	4794114	4799166	4793580	4799200
CNV.9	1	4889825	4896261	4889604	4897567
CNV.10	1	5035188	5038051	5031869	5039093
CNV.11	1	5112057	5174017	5102144	5179806
CNV.12	1	5352492	5403585	5347439	5408889
CNV.13	1	5826898	5830967	5820669	5831118
CNV.14	1	5898729	5902850	5898323	5903958
CNV.15	1	6018391	6056168	6014563	6056729
CNV.16	1	6434139	6445278	6433129	6449638
CNV.17	1	6771569	6773752	6771523	6773839
CNV.18	1	7557152	7567037	7554982	7571598
CNV.19	1	7684188	7684488	7684167	7688059
CNV.20	1	8182588	8192626	8176626	8192925
CNV.21	1	8363248	8370513	8357221	8372536
CNV.22	1	8695768	8712280	8690903	8718911
CNV.23	1	8764655	8767914	8764522	8769680
CNV.24	1	8966334	8983759	8951046	8983768
CNV.25	1	9228182	9233936	9225509	9234143
CNV.26	1	9326658	9402403	9315847	9404693
CNV.27	1	9589413	9604873	9588109	9608796
CNV.28	1	9779503	9786617	9774718	9789850
CNV.29	1	9994339	10003320	9994020	10013014
CNV.30	1	10097020	10103011	10091661	10106775
CNV.31	1	10237177	10256355	10236491	10258806
CNV.32	1	10367344	10377983	10367209	10381772
CNV.33	1	10543836	10545046	10542121	10548862
CNV.34	1	10617272	10627542	10617191	10632582
CNV.35	1	10669098	10682210	10668896	10686274
CNV.36	1	12770320	12775886	12769894	12776134
CNV.37	1	12835868	13717772	12834580	13717824
CNV.38	1	13739221	13793392	13739165	13794825
CNV.39	1	14078163	14091579	14071845	14096068

CNV.40	1	14331267	14337937	14331244	14337994
CNV.41	1	14423009	14435530	14420140	14439042
CNV.42	1	15056292	15061203	15056096	15062626
CNV.43	1	15485314	15495834	15485139	15495905
CNV.44	1	15690527	15696059	15689993	15699117
CNV.45	1	15785719	15801495	15783771	15802534
CNV.46	1	16013985	16080171	16012657	16087824
CNV.47	1	16147063	16170460	16147032	16171036
CNV.48	1	16363940	16389566	16345255	16390402
CNV.49	1	16566819	16593369	16563948	16594994
CNV.50	1	16803776	17324445	16797756	17325955
CNV.51	1	17534310	17541949	17531582	17542448
CNV.52	1	17593406	17620009	17593316	17622654
CNV.53	1	19373605	19376017	19373588	19376494
CNV.54	1	19585291	19618629	19584004	19622270
CNV.55	1	20882944	20884941	20882523	20886767
CNV.56	1	21624787	21640395	21622169	21640501
CNV.57	1	21687854	21778260	21687757	21778508
CNV.58	1	22147279	22160257	22147077	22164164
CNV.59	1	22300465	22431777	22298285	22435147
CNV.60	1	22501716	22509206	22496703	22514236
CNV.61	1	22622289	22629112	22619342	22629188
CNV.62	1	22649994	22658727	22649487	22664774
CNV.63	1	24139321	24147505	24135174	24148988
CNV.64	1	24897315	24902026	24878791	24902076
CNV.65	1	24977969	24990435	24977510	24994421
CNV.66	1	25180379	25184151	25178331	25185197
CNV.67	1	25329991	25333416	25329778	25333899
CNV.68	1	25416761	25424322	25415682	25424889
CNV.69	1	25445482	25464820	25445020	25465850
CNV.70	1	25536088	25688276	25532484	25696602
CNV.71	1	25711056	25719599	25707024	25723250
CNV.72	1	26622607	26625414	26616513	26632781
CNV.73	1	26813075	26848386	26805583	26848624
CNV.74	1	26863735	26871337	26861106	26873245
CNV.75	1	27361219	27388384	27348943	27398714
CNV.76	1	27809578	27838650	27807618	27844993
CNV.77	1	28393650	28421153	28391482	28424840
CNV.78	1	28503603	28512505	28503037	28512807
CNV.79	1	28830865	28836387	28829902	28840149
CNV.80	1	29453859	29461133	29442980	29461718
CNV.81	1	30485510	30488561	30479770	30490594
CNV.82	1	30658606	30665075	30655527	30665519
CNV.83	1	30711935	30716585	30711165	30717033
CNV.84	1	30726693	30741950	30726574	30745210
CNV.85	1	30923423	30925285	30923171	30925560
CNV.86	1	31143035	31251698	31139947	31262005

CNV.87	1	31972489	31979319	31968315	31980298
CNV.88	1	32896119	32914644	32894381	32917583
CNV.89	1	33405472	33425934	33400380	33426670
CNV.90	1	34222304	34229130	34221810	34231217
CNV.91	1	34656303	34664770	34656179	34665219
CNV.92	1	35087805	35116561	35087725	35117925
CNV.93	1	36857141	36860800	36856673	36862254
CNV.94	1	37625763	37628812	37625062	37632836
CNV.95	1	37773586	37776298	37773050	37777759
CNV.96	1	37851713	37903112	37848793	37920683
CNV.97	1	38582858	38587146	38572323	38587171
CNV.98	1	39132431	39135188	39132430	39135250
CNV.99	1	39451857	39458881	39451244	39459560
CNV.100	1	40104584	40166859	40104461	40168587
CNV.101	1	40785263	40792578	40781038	40795628
CNV.102	1	40847475	40870545	40838636	40872356
CNV.103	1	40966183	40969664	40963304	40974799
CNV.104	1	41021036	41028210	41020940	41028926
CNV.105	1	41340619	41376500	41340069	41383724
CNV.106	1	44602422	44614291	44602382	44615176
CNV.107	1	45072407	45076660	45071196	45079220
CNV.108	1	45247079	45269642	45235976	45278094
CNV.109	1	46242332	46255004	46240049	46256890
CNV.110	1	46363880	46590135	46360054	46594836
CNV.111	1	46902399	46906144	46901546	46907340
CNV.112	1	47414647	47418426	47411933	47420328
CNV.113	1	47586717	47593574	47578811	47594583
CNV.114	1	47827181	47837404	47824641	47840129
CNV.115	1	48071156	48071985	48065338	48072529
CNV.116	1	48519912	48524180	48519515	48527534
CNV.117	1	48685435	48690229	48682034	48698352
CNV.118	1	48715195	48722378	48713713	48725096
CNV.119	1	48841815	48857599	48837337	48860907
CNV.120	1	49914932	49997863	49914123	49999266
CNV.121	1	50023956	50025534	50012288	50035049
CNV.122	1	50415126	50423898	50411377	50424973
CNV.123	1	51239101	51379813	51238182	51380860
CNV.124	1	53590630	53598126	53590423	53599033
CNV.125	1	54042650	54045180	54042336	54045548
CNV.126	1	54435043	54458866	54434610	54461661
CNV.127	1	55085183	55096450	55085141	55100865
CNV.128	1	55207754	55274519	55207494	55274652
CNV.129	1	55307580	55308237	55305457	55308390
CNV.130	1	55724023	55724299	55723441	55729368
CNV.131	1	55774784	55775665	55774397	55779387
CNV.132	1	56495368	56502088	56495121	56502240
CNV.133	1	57650959	57658204	57647465	57660475

CNV.134	1	58432348	58437666	58430196	58438657
CNV.135	1	58448884	58451457	58448704	58452314
CNV.136	1	58482316	58483220	58482029	58484277
CNV.137	1	58723558	58733326	58719909	58742470
CNV.138	1	58783290	58784150	58775321	58786654
CNV.139	1	59877747	59891290	59874107	59891522
CNV.140	1	59946056	59951022	59944683	59951705
CNV.141	1	60004604	60052066	59992659	60058279
CNV.142	1	60105463	60111159	60101459	60111987
CNV.143	1	60977302	60983782	60975806	60985386
CNV.144	1	61408539	61411107	61407352	61414438
CNV.145	1	62114006	62119978	62113102	62120564
CNV.146	1	62279594	62285367	62275920	62288311
CNV.147	1	62439231	62450535	62436943	62454672
CNV.148	1	62765461	62767968	62765049	62772200
CNV.149	1	64701504	64704921	64701477	64704928
CNV.150	1	64842727	64854605	64837281	64855419
CNV.151	1	65116084	65119271	65115812	65125113
CNV.152	1	65429215	65429974	65427606	65435063
CNV.153	1	65691578	65698884	65688461	65699203
CNV.154	1	65854855	65955725	65853649	65956938
CNV.155	1	66670804	66671284	66670768	66671850
CNV.156	1	69249132	69266103	69245790	69267189
CNV.157	1	69886314	69899166	69881261	69899201
CNV.158	1	70689215	70799077	70685580	70802794
CNV.159	1	70902206	70919559	70901335	70919676
CNV.160	1	70963368	70976361	70963189	70985173
CNV.161	1	70997442	70998145	70997152	70998786
CNV.162	1	71065311	71068143	71060331	71069107
CNV.163	1	71302563	71314457	71299622	71315832
CNV.164	1	71320313	71340070	71318089	71349408
CNV.165	1	71415135	71458858	71414308	71458973
CNV.166	1	71606450	71620043	71605781	71620679
CNV.167	1	71712223	71731104	71709205	71731762
CNV.168	1	71777263	71784460	71774984	71788907
CNV.169	1	71985989	72024819	71985957	72025248
CNV.170	1	72283802	72311975	72283499	72312132
CNV.171	1	72740734	72818779	72740073	72818890
CNV.172	1	72820027	72827978	72818890	72828312
CNV.173	1	73011920	73088158	73008167	73097100
CNV.174	1	73173337	73174188	73168073	73178659
CNV.175	1	73768693	73776696	73758117	73777347
CNV.176	1	73981290	73984366	73976842	73988566
CNV.177	1	74485995	74520957	74483797	74522280
CNV.178	1	74770634	74812588	74767844	74813265
CNV.179	1	74876513	74907302	74872611	74907316
CNV.180	1	75177284	75203583	75176907	75203908

CNV.181	1	75232857	75253144	75230548	75254606
CNV.182	1	75273866	75313578	75271072	75313639
CNV.183	1	76050174	76059959	76049897	76064057
CNV.184	1	76109705	76114664	76109573	76116062
CNV.185	1	76478305	76480439	76472638	76484827
CNV.186	1	77128429	77141838	77126822	77147389
CNV.187	1	77169583	77191549	77168568	77193659
CNV.188	1	77254137	77258826	77253722	77261824
CNV.189	1	77391700	77430751	77388804	77440520
CNV.190	1	77775014	77779836	77774872	77781670
CNV.191	1	78783161	78807671	78782363	78809058
CNV.192	1	79320311	79555327	79319800	79555605
CNV.193	1	80210710	80222753	80207884	80226910
CNV.194	1	80757261	80775574	80755155	80781601
CNV.195	1	81118171	81131278	81118102	81131706
CNV.196	1	81329765	81336886	81326992	81338600
CNV.197	1	81563249	81700742	81558915	81704476
CNV.198	1	81888605	81888868	81888602	81888943
CNV.199	1	83187476	83196421	83185639	83203837
CNV.200	1	83214304	83227438	83213109	83227697
CNV.201	1	83740658	83751974	83740524	83755433
CNV.202	1	83815304	83889370	83812787	83891295
CNV.203	1	85016559	85021277	85004345	85022495
CNV.204	1	85539127	85542620	85537890	85545680
CNV.205	1	86161398	86181186	86156737	86186379
CNV.206	1	86763999	86778266	86763817	86780082
CNV.207	1	87028669	87039945	87026924	87040567
CNV.208	1	87141222	87173982	87140644	87175529
CNV.209	1	88257854	88260648	88256046	88261043
CNV.210	1	88843795	88857668	88843632	88862772
CNV.211	1	89265631	89303226	89260091	89307966
CNV.212	1	89475230	89485298	89475135	89487657
CNV.213	1	89899878	89911433	89899391	89914724
CNV.214	1	90550699	90556980	90544417	90559217
CNV.215	1	91094904	91102904	91091581	91103184
CNV.216	1	91272896	91275431	91269543	91275645
CNV.217	1	92151331	92208764	92144933	92210610
CNV.218	1	92227848	92232153	92226887	92233820
CNV.219	1	92241731	92248440	92234553	92250485
CNV.220	1	94334924	94341724	94334724	94343023
CNV.221	1	94349077	94349835	94347446	94356251
CNV.222	1	94569756	94573046	94569504	94573519
CNV.223	1	94701469	94711524	94700878	94714054
CNV.224	1	95130678	95152850	95126701	95155385
CNV.225	1	95657912	95658323	95650521	95660815
CNV.226	1	96154717	96156066	96152354	96156216
CNV.227	1	97684328	97691522	97678081	97692317

CNV.228	1	98601247	98640131	98592630	98643696
CNV.229	1	101271811	101276245	101269449	101276435
CNV.230	1	102088317	102101223	102085808	102104059
CNV.231	1	102656193	102852850	102655684	102858846
CNV.232	1	104067245	104326489	104067000	104335637
CNV.233	1	104902499	104931238	104902360	104931842
CNV.234	1	105039720	105051014	105039557	105054917
CNV.235	1	105254068	105254973	105252205	105257300
CNV.236	1	105321596	105386262	105311293	105393481
CNV.237	1	105492132	105505952	105491927	105507848
CNV.238	1	105622940	105637208	105622533	105641121
CNV.239	1	106000943	106087275	106000399	106090920
CNV.240	1	106135644	106136913	106134245	106139463
CNV.241	1	106163190	106227167	106163088	106229491
CNV.242	1	106309138	106315854	106306488	106320956
CNV.243	1	106807734	106810965	106804283	106811312
CNV.244	1	107119820	107125343	107119787	107141678
CNV.245	1	107222150	107223482	107221718	107225176
CNV.246	1	107654220	107663495	107653985	107665433
CNV.247	1	107940396	107944721	107938505	107944955
CNV.248	1	108074435	108074435	108073527	108074631
CNV.249	1	108267548	108281260	108266970	108281419
CNV.250	1	108287065	108294496	108286971	108296508
CNV.251	1	108296958	108302031	108296956	108304301
CNV.252	1	108322535	108327875	108321943	108328378
CNV.253	1	108513484	108514477	108511501	108515347
CNV.254	1	108546121	108568011	108545066	108568620
CNV.255	1	108721006	108868815	108719572	108871119
CNV.256	1	108879304	108886965	108878937	108888604
CNV.257	1	108901366	109013908	108900763	109014308
CNV.258	1	109486269	109498145	109483850	109503656
CNV.259	1	109574526	109606367	109571177	109610359
CNV.260	1	109639687	109666920	109623117	109672215
CNV.261	1	109778237	109786911	109774026	109787389
CNV.262	1	110178212	110253130	110175980	110256624
CNV.263	1	110431422	110435937	110431020	110436506
CNV.264	1	111295688	111296965	111295674	111301623
CNV.265	1	111331000	111412235	111330896	111412279
CNV.266	1	111827816	111835498	111827749	111839878
CNV.267	1	111929137	111933231	111928394	111935668
CNV.268	1	112051780	112070735	112046557	112074061
CNV.269	1	112083640	112095169	112083083	112095646
CNV.270	1	112176476	112199235	112175999	112201328
CNV.271	1	112293032	112310686	112292706	112311751
CNV.272	1	112312211	112314565	112312154	112314822
CNV.273	1	112690211	112706198	112688311	112706205
CNV.274	1	112991815	112996758	112991680	112996781

CNV.275	1	113145839	113158260	113144767	113163438
CNV.276	1	113506466	113513304	113505624	113514192
CNV.277	1	113652648	113657538	113652604	113658887
CNV.278	1	114535644	114543108	114535293	114545724
CNV.279	1	114828839	114833832	114828815	114833845
CNV.280	1	117194086	117200404	117194070	117206336
CNV.281	1	117213375	117219944	117208400	117220085
CNV.282	1	117445243	117453755	117443368	117456936
CNV.283	1	118607446	118613488	118602518	118616060
CNV.284	1	119194305	119220496	119192441	119224286
CNV.285	1	119240453	119256513	119236151	119256519
CNV.286	1	120021071	120022175	120020358	120023888
CNV.287	1	120102884	120150683	120102558	120150949
CNV.288	1	120280734	120305989	120280554	120308275
CNV.289	1	120310292	120328527	120309868	120330380
CNV.290	1	120540804	149876147	120527395	149877903
CNV.291	1	150104412	150114083	150104039	150119685
CNV.292	1	150270209	150280590	150266852	150286434
CNV.293	1	150350232	150465897	150349768	150466767
CNV.294	1	150673468	150674447	150673364	150676281
CNV.295	1	150741605	150764632	150740775	150770989
CNV.296	1	150800484	150801159	150798280	150801176
CNV.297	1	150803027	150811028	150801176	150811261
CNV.298	1	150994508	150999737	150992693	151003600
CNV.299	1	151336699	151397878	151332017	151408736
CNV.300	1	151553672	151577221	151552672	151583801
CNV.301	1	152184534	152201217	152184149	152203270
CNV.302	1	152295532	152312764	152292453	152313798
CNV.303	1	152416494	152417572	152414833	152424151
CNV.304	1	152493154	152507348	152489742	152508137
CNV.305	1	152526643	152590750	152519209	152590962
CNV.306	1	152603522	152606110	152603473	152606122
CNV.307	1	152626170	152638281	152626061	152639194
CNV.308	1	152643406	152660074	152641701	152660249
CNV.309	1	152693599	152698054	152693043	152698168
CNV.310	1	152751237	152778526	152750800	152778558
CNV.311	1	152995724	153000714	152991060	153000978
CNV.312	1	153193484	153200987	153193178	153202934
CNV.313	1	153324574	153364130	153322574	153365548
CNV.314	1	153504135	153531204	153503777	153531282
CNV.315	1	153673674	153728095	153672247	153735369
CNV.316	1	155189160	155201064	155180466	155211667
CNV.317	1	155655348	155667102	155652156	155667221
CNV.318	1	157143893	157161237	157138969	157168071
CNV.319	1	157478020	157484575	157477996	157484698
CNV.320	1	157625513	157626153	157624481	157628355
CNV.321	1	158365697	158368309	158365615	158368637

CNV.322	1	158485137	158487538	158484876	158487608
CNV.323	1	158489459	158496029	158487860	158498522
CNV.324	1	158506083	158511569	158503350	158511659
CNV.325	1	158516996	158545435	158515085	158549892
CNV.326	1	158681743	158713712	158681444	158718684
CNV.327	1	158745653	158747492	158744665	158748017
CNV.328	1	158778467	158804673	158774640	158806207
CNV.329	1	158854768	158887208	158852953	158887302
CNV.330	1	158956596	158967046	158956412	158967392
CNV.331	1	159923111	159933261	159920719	159936822
CNV.332	1	159984454	159987920	159984431	159988120
CNV.333	1	160146494	160148671	160146173	160150874
CNV.334	1	160630193	160646411	160629686	160652730
CNV.335	1	160937020	160940085	160933768	160942342
CNV.336	1	161353363	161355728	161344497	161358966
CNV.337	1	161446164	161656369	161446131	161656498
CNV.338	1	161841016	161842207	161840872	161844106
CNV.339	1	162866194	162867776	162864623	162872112
CNV.340	1	163205092	163217668	163199914	163219889
CNV.341	1	163752093	163802306	163751433	163802529
CNV.342	1	164264687	164275602	164256271	164276414
CNV.343	1	165421804	165424170	165421658	165424610
CNV.344	1	165835194	165835690	165835144	165837978
CNV.345	1	166184481	166194370	166176043	166195702
CNV.346	1	166552201	166553095	166548373	166553495
CNV.347	1	167713078	167715397	167705632	167715741
CNV.348	1	167764835	167768513	167763804	167769672
CNV.349	1	168986301	168987211	168985815	168989386
CNV.350	1	168989939	169007842	168989936	169008512
CNV.351	1	169233974	169257750	169216937	169258047
CNV.352	1	170367736	170387205	170367650	170387331
CNV.353	1	170480292	170483093	170480129	170488609
CNV.354	1	171013931	171024368	171011187	171024874
CNV.355	1	171937720	171942783	171936722	171944418
CNV.356	1	174579072	174596973	174571709	174606076
CNV.357	1	174796266	174802451	174795938	174802841
CNV.358	1	174803606	174803773	174803056	174804979
CNV.359	1	175430855	175760752	175430808	175763022
CNV.360	1	176232997	176276387	176220257	176285547
CNV.361	1	177809560	177811371	177808463	177811499
CNV.362	1	178659933	178674188	178658817	178676866
CNV.363	1	179120292	179120676	179120202	179120743
CNV.364	1	179166557	179170217	179161751	179172956
CNV.365	1	179323738	179339206	179319816	179339262
CNV.366	1	179446133	179493141	179439630	179496779
CNV.367	1	181251314	181272338	181244093	181275760
CNV.368	1	182084976	182087336	182082972	182087977

CNV.369	1	182777405	182781050	182775884	182783894
CNV.370	1	184757698	184809599	184755346	184810238
CNV.371	1	184961804	184970511	184959979	184977021
CNV.372	1	185096576	185129437	185096251	185134066
CNV.373	1	185413885	185420724	185413782	185422027
CNV.374	1	186437112	186442072	186433229	186444919
CNV.375	1	187715179	187728991	187714607	187729962
CNV.376	1	188664700	188672106	188659420	188672537
CNV.377	1	188763750	188769865	188758604	188772740
CNV.378	1	189027223	189031361	189022973	189032480
CNV.379	1	189084533	189087000	189084365	189089417
CNV.380	1	189172412	189182858	189168241	189189911
CNV.381	1	189258909	189537331	189258813	189546970
CNV.382	1	189551970	189552871	189549597	189554067
CNV.383	1	189647198	189656102	189646913	189657100
CNV.384	1	189774645	189775972	189770705	189776217
CNV.385	1	189807684	189809410	189795013	189810172
CNV.386	1	189970368	189991289	189963457	189992067
CNV.387	1	190008884	190013622	190005570	190020202
CNV.388	1	190466822	190473405	190465989	190476185
CNV.389	1	190825644	190836583	190825536	190836833
CNV.390	1	190864130	190893617	190862150	190900121
CNV.391	1	191017858	191023684	191017280	191026391
CNV.392	1	191045247	191046442	191041732	191047341
CNV.393	1	191059432	191075485	191058384	191080491
CNV.394	1	191716715	191726158	191713968	191727949
CNV.395	1	191829069	191867711	191825854	191871216
CNV.396	1	192026255	192059386	192015909	192060662
CNV.397	1	192673525	192677326	192672781	192679697
CNV.398	1	192761328	192764161	192760537	192767917
CNV.399	1	192873898	192892192	192870791	192894544
CNV.400	1	193155512	193159723	193155455	193161058
CNV.401	1	193213210	193213474	193213168	193221915
CNV.402	1	194430832	194461499	194423947	194464722
CNV.403	1	194938079	194943654	194932473	194947254
CNV.404	1	195013975	195019299	195012880	195022967
CNV.405	1	195085778	195089776	195081209	195089819
CNV.406	1	195221220	195385150	195219630	195386897
CNV.407	1	195825382	195874907	195820890	195876060
CNV.408	1	195938400	195948955	195938326	195949631
CNV.409	1	196698299	196941748	196698272	196942455
CNV.410	1	196943638	196952274	196942455	196952401
CNV.411	1	198304536	198311466	198304274	198311960
CNV.412	1	198438284	198447659	198437796	198448718
CNV.413	1	198871076	198876858	198870597	198878347
CNV.414	1	199047795	199054140	199041557	199054666
CNV.415	1	199072722	199093876	199054666	199094550

CNV.416	1	199107653	199114238	199107596	199114390
CNV.417	1	199542563	199545705	199539063	199546187
CNV.418	1	199854177	199920030	199850952	199924396
CNV.419	1	200107430	200110256	200107425	200112670
CNV.420	1	200788831	200840662	200784806	200843768
CNV.421	1	202042872	202052416	202042364	202053829
CNV.422	1	202567272	202568542	202566780	202568735
CNV.423	1	203908485	203922495	203906128	203923193
CNV.424	1	203936783	203945398	203934734	203945501
CNV.425	1	204187207	204205570	204185584	204209625
CNV.426	1	206783022	206787292	206779328	206788458
CNV.427	1	207255654	207257303	207255619	207260606
CNV.428	1	207595453	207606400	207594943	207607073
CNV.429	1	208438984	208457709	208437821	208457920
CNV.430	1	208489481	208490895	208486789	208492097
CNV.431	1	209123864	209136831	209123767	209141842
CNV.432	1	209490527	209510777	209488564	209511828
CNV.433	1	210068997	210093571	210068954	210093649
CNV.434	1	210604746	210615292	210598837	210615293
CNV.435	1	213005012	213010436	213004455	213014453
CNV.436	1	213817311	213830825	213814808	213830901
CNV.437	1	214007190	214014827	214003718	214019082
CNV.438	1	214637126	214645307	214637100	214645320
CNV.439	1	215370570	215381265	215367780	215384239
CNV.440	1	215493469	215499104	215491496	215501031
CNV.441	1	215858193	215862105	215855994	215866009
CNV.442	1	216153695	216157829	216153363	216157955
CNV.443	1	216205610	216237728	216205281	216238518
CNV.444	1	217157012	217206765	217155126	217215465
CNV.445	1	217668330	217721335	217663563	217731548
CNV.446	1	218727990	218742974	218724451	218744152
CNV.447	1	220925838	220954633	220925241	220958959
CNV.448	1	220970384	220970593	220967284	220971048
CNV.449	1	221287861	221300012	221281469	221300058
CNV.450	1	221746050	221753609	221742928	221756538
CNV.451	1	222365644	222389246	222364124	222389324
CNV.452	1	223017325	223026717	223016406	223027548
CNV.453	1	223204628	223214455	223204001	223217060
CNV.454	1	223628190	223646781	223619734	223646839
CNV.455	1	225390933	225465117	225389490	225466649
CNV.456	1	225659217	225684196	225657275	225690842
CNV.457	1	225990372	226015299	225989115	226015641
CNV.458	1	226027659	226058639	226025175	226061006
CNV.459	1	226064176	226067799	226063581	226067862
CNV.460	1	227103413	227106223	227102354	227107984
CNV.461	1	227424455	227432923	227422862	227434376
CNV.462	1	228507780	228741147	228500181	228791784

CNV.463	1	229461673	229463035	229455458	229464432
CNV.464	1	229811169	229820456	229809407	229820485
CNV.465	1	230954497	230955097	230951898	230955129
CNV.466	1	231324477	231327024	231324444	231327666
CNV.467	1	231712569	231813318	231706899	231816137
CNV.468	1	231834901	231887497	231833718	231893172
CNV.469	1	231898591	231957336	231896958	231963337
CNV.470	1	232456472	232463715	232455417	232463759
CNV.471	1	232692293	232705138	232688992	232706083
CNV.472	1	232853462	232871259	232847858	232875638
CNV.473	1	232895309	232907549	232892748	232908469
CNV.474	1	233264421	233271200	233261189	233271643
CNV.475	1	234254400	234260086	234250861	234262397
CNV.476	1	234656884	234659807	234656596	234662742
CNV.477	1	234679850	234697146	234677737	234699236
CNV.478	1	234703598	234712115	234699875	234712612
CNV.479	1	234767241	234792708	234766604	234793677
CNV.480	1	234844840	234845737	234842932	234846508
CNV.481	1	234995365	235001174	234990504	235005122
CNV.482	1	235151488	235165684	235138348	235243184
CNV.483	1	235251071	235262385	235248412	235263993
CNV.484	1	235273194	235316688	235273179	235316710
CNV.485	1	235746016	235746426	235743992	235747640
CNV.486	1	236233996	236263115	236231854	236264766
CNV.487	1	237274019	237279154	237267370	237280205
CNV.488	1	238262266	238262981	238261969	238264117
CNV.489	1	238506546	238512614	238505192	238514420
CNV.490	1	238751620	238752463	238750109	238752511
CNV.491	1	238816125	238817940	238816009	238818102
CNV.492	1	239087028	239096030	239086386	239097333
CNV.493	1	239426683	239435377	239425033	239435635
CNV.494	1	239465591	239466347	239457218	239468556
CNV.495	1	239618222	239625309	239614862	239626654
CNV.496	1	240052422	240053444	240051026	240053707
CNV.497	1	240192359	240204140	240191654	240204329
CNV.498	1	240224720	240228959	240224481	240229781
CNV.499	1	240394015	240396105	240391740	240398028
CNV.500	1	241130966	241136033	241124489	241138961
CNV.501	1	241203127	241210309	241200049	241210468
CNV.502	1	241362781	241366987	241352194	241375666
CNV.503	1	241822182	241852265	241821181	241858139
CNV.504	1	242072649	242186614	242069044	242187836
CNV.505	1	242420888	242440611	242420839	242450015
CNV.506	1	243044970	243353851	243044307	243355461
CNV.507	1	243860924	243861061	243854196	243863108
CNV.508	1	244642948	244853474	244636961	244854821
CNV.509	1	245236935	245257556	245236112	245263105

CNV.510	1	245346971	245366662	245334533	245367073
CNV.511	1	245637039	245649651	245636410	245650429
CNV.512	1	245793993	245799473	245793665	245803677
CNV.513	1	245946822	245954786	245944890	245957828
CNV.514	1	245971712	245989007	245965551	245989211
CNV.515	1	246114142	246116086	246113980	246117191
CNV.516	1	246190117	246595982	246190002	246598346
CNV.517	1	246601395	247043464	246599746	247044033
CNV.518	1	247302182	247332511	247296884	247335735
CNV.519	1	247625790	247636015	247624713	247636350
CNV.520	1	247815781	249224388	247814286	249224388
CNV.521	2	12784	479375	12784	480597
CNV.522	2	591113	594324	590698	594741
CNV.523	2	694759	1916577	687775	1916630
CNV.524	2	2009887	2022106	2006274	2022348
CNV.525	2	2613419	2614174	2612229	2617204
CNV.526	2	2628842	2633047	2627251	2636413
CNV.527	2	2842932	2845241	2839594	2845467
CNV.528	2	3147861	3294815	3147136	3297006
CNV.529	2	3722486	3736538	3721821	3742071
CNV.530	2	3828109	3841739	3828093	3849512
CNV.531	2	4196485	4238857	4192934	4240306
CNV.532	2	4409073	4414066	4409059	4414342
CNV.533	2	4624710	4639380	4621133	4641000
CNV.534	2	4655985	4662656	4654887	4662688
CNV.535	2	4913700	4917612	4913643	4917689
CNV.536	2	4917897	4918142	4917809	4923866
CNV.537	2	4951491	4952545	4950617	4952797
CNV.538	2	4962364	4972157	4957344	4973678
CNV.539	2	5295889	5304015	5295633	5306069
CNV.540	2	5390962	5392211	5390932	5392247
CNV.541	2	5526601	5532841	5522488	5535020
CNV.542	2	5736321	5737785	5736039	5737817
CNV.543	2	5884206	5901926	5881184	5902605
CNV.544	2	6268752	6284873	6268252	6287168
CNV.545	2	6298140	6311063	6298134	6315388
CNV.546	2	6401909	6405061	6400019	6406740
CNV.547	2	6433173	6435188	6429691	6441231
CNV.548	2	6694519	6706459	6692000	6708722
CNV.549	2	6734644	6745066	6734032	6750693
CNV.550	2	6795995	6804213	6795606	6805776
CNV.551	2	7015016	7018881	7014670	7021878
CNV.552	2	7233524	7234476	7229829	7236150
CNV.553	2	7237774	7238284	7237504	7238335
CNV.554	2	7290380	7311136	7289759	7314628
CNV.555	2	7321992	7327008	7321937	7327114
CNV.556	2	7417816	7423402	7417304	7423488

CNV.557	2	7524467	7526903	7521526	7527017
CNV.558	2	7580944	7682371	7576098	7686319
CNV.559	2	7693398	7715663	7693271	7715804
CNV.560	2	8001352	8002872	7998979	8003547
CNV.561	2	8094632	8153116	8094445	8155559
CNV.562	2	8157394	8160926	8156251	8161538
CNV.563	2	8171098	8180673	8166518	8183929
CNV.564	2	8566687	8567632	8565622	8568779
CNV.565	2	9177204	9191031	9174380	9191332
CNV.566	2	9292001	9299046	9290357	9301065
CNV.567	2	9977737	9989442	9975557	9989982
CNV.568	2	10048102	10051946	10047828	10052211
CNV.569	2	10150523	10159436	10141191	10160082
CNV.570	2	10701044	10713932	10700239	10714198
CNV.571	2	11031993	11033567	11031838	11033580
CNV.572	2	11284768	11287294	11284415	11289928
CNV.573	2	12272548	12273286	12271846	12274022
CNV.574	2	12304073	12304315	12304045	12304371
CNV.575	2	12685369	12696236	12681835	12697043
CNV.576	2	12788014	12794978	12787968	12795600
CNV.577	2	13088551	13093966	13084077	13094700
CNV.578	2	13193220	13288025	13192848	13288926
CNV.579	2	13536934	13542222	13534976	13549672
CNV.580	2	13708820	13720048	13707079	13721711
CNV.581	2	13737239	13744513	13731985	13750565
CNV.582	2	13767654	13781387	13763674	13786624
CNV.583	2	13983638	13986304	13980757	13987873
CNV.584	2	14201739	14205189	14201628	14205319
CNV.585	2	14263542	14274078	14262553	14275242
CNV.586	2	14303639	14313160	14303429	14316914
CNV.587	2	14352324	14379327	14348914	14379495
CNV.588	2	14700683	14710364	14700629	14710487
CNV.589	2	15012057	15064586	15009071	15071092
CNV.590	2	15205821	15215202	15203160	15215500
CNV.591	2	15345345	15351500	15343530	15351651
CNV.592	2	16033471	16039796	16033427	16047628
CNV.593	2	17217850	17237961	17217075	17242537
CNV.594	2	17672439	17680727	17669991	17681666
CNV.595	2	18163569	18193264	18162068	18200134
CNV.596	2	18328339	18343939	18320586	18352405
CNV.597	2	18479749	18492559	18478749	18493236
CNV.598	2	18724302	18726434	18723225	18727332
CNV.599	2	18970968	18973863	18970472	18978576
CNV.600	2	20179800	20201418	20179220	20201458
CNV.601	2	20817493	20823219	20815964	20823252
CNV.602	2	21229860	21301892	21229003	21304432
CNV.603	2	21828618	21834033	21825935	21835926

CNV.604	2	21956461	21958162	21953992	21958745
CNV.605	2	22323262	22431122	22320133	22443343
CNV.606	2	24596812	24627234	24595328	24630099
CNV.607	2	24670492	24710635	24667823	24717928
CNV.608	2	25218716	25237499	25218019	25239264
CNV.609	2	25253672	25259520	25252648	25262560
CNV.610	2	25724781	25730439	25723780	25737187
CNV.611	2	26226114	26239580	26225725	26240974
CNV.612	2	27035978	27061253	27035944	27073640
CNV.613	2	27439751	27441904	27438212	27442009
CNV.614	2	27765308	27779518	27761000	27784428
CNV.615	2	27811406	27825107	27808154	27827161
CNV.616	2	28878741	28892544	28877585	28892768
CNV.617	2	28951698	28951698	28950818	28952773
CNV.618	2	28988465	29222800	28986350	29226559
CNV.619	2	29237786	29240005	29231421	29240060
CNV.620	2	29641874	29644937	29641200	29646424
CNV.621	2	30253853	30256351	30253699	30256393
CNV.622	2	30427310	30439774	30426596	30445026
CNV.623	2	30615651	30627654	30614883	30629876
CNV.624	2	30818847	30826948	30817685	30831808
CNV.625	2	32630548	33331778	32627997	33333871
CNV.626	2	33413193	33428307	33413047	33428653
CNV.627	2	33762963	33770267	33762021	33770317
CNV.628	2	34098692	34124115	34098224	34127078
CNV.629	2	34403301	34404938	34403165	34405804
CNV.630	2	34420363	34427646	34416184	34428936
CNV.631	2	34518872	34529100	34517813	34529244
CNV.632	2	34648698	34743036	34646409	34746457
CNV.633	2	34828677	34829116	34820764	34829378
CNV.634	2	34837639	34860226	34830059	34865070
CNV.635	2	34923003	34923268	34922341	34927885
CNV.636	2	34954042	34959421	34949368	34960942
CNV.637	2	34987031	34993043	34982759	34994158
CNV.638	2	35001138	35028209	35000446	35034732
CNV.639	2	35045615	35052108	35043996	35056247
CNV.640	2	35064073	35066670	35057525	35067745
CNV.641	2	35080863	35090870	35076002	35093293
CNV.642	2	35122206	35124027	35122161	35127305
CNV.643	2	35156295	35156295	35155781	35161672
CNV.644	2	35404383	35408759	35396761	35411425
CNV.645	2	35581507	35615675	35581151	35621916
CNV.646	2	35636223	35644367	35636183	35645115
CNV.647	2	35648158	35688570	35645712	35697479
CNV.648	2	35708846	35712218	35701553	35712222
CNV.649	2	35737252	35744382	35735771	35744997
CNV.650	2	35818934	36089929	35815793	36091829

CNV.651	2	36293922	36302005	36293840	36302217
CNV.652	2	36335595	36352666	36332258	36353114
CNV.653	2	36390773	36422846	36390715	36423276
CNV.654	2	36475533	36503774	36474979	36503818
CNV.655	2	37585776	38268325	37584643	38272342
CNV.656	2	38371476	38376029	38367734	38376569
CNV.657	2	38461551	38469133	38460559	38469908
CNV.658	2	38954495	38980922	38954419	38981047
CNV.659	2	39564841	39578783	39563239	39580377
CNV.660	2	40130032	40245464	40129165	40254811
CNV.661	2	40332851	40336640	40330275	40338067
CNV.662	2	40396253	40406679	40396078	40406745
CNV.663	2	40416373	40433420	40415673	40435325
CNV.664	2	40581070	40596524	40580918	40596947
CNV.665	2	40612784	40617820	40608371	40621060
CNV.666	2	40750107	40757655	40749996	40757791
CNV.667	2	40809991	40822146	40805954	40822227
CNV.668	2	40927375	40951516	40926729	40951667
CNV.669	2	40987849	41010688	40987767	41011338
CNV.670	2	41238064	41251724	41238048	41251754
CNV.671	2	41291584	41322344	41281133	41327687
CNV.672	2	41380447	41582472	41372634	41584777
CNV.673	2	41617674	41618171	41616164	41618337
CNV.674	2	41893994	41902542	41892171	41903118
CNV.675	2	42010287	42018118	42006280	42018443
CNV.676	2	42122150	42125292	42122048	42126403
CNV.677	2	42218545	42221789	42218334	42224021
CNV.678	2	42308536	42314455	42307621	42317341
CNV.679	2	42447357	42452613	42445808	42454185
CNV.680	2	43024676	43032241	43024439	43035049
CNV.681	2	43059121	43061322	43058278	43069415
CNV.682	2	43521167	43577615	43512243	43586902
CNV.683	2	43607811	43613236	43603152	43616589
CNV.684	2	44274768	44301625	44273710	44304047
CNV.685	2	44339612	44348768	44337754	44350362
CNV.686	2	44365835	44371813	44365780	44372909
CNV.687	2	44517190	44543447	44514122	44546016
CNV.688	2	44742517	44761066	44739872	44764999
CNV.689	2	44805053	44820764	44803885	44822149
CNV.690	2	45312362	45328303	45301922	45331043
CNV.691	2	45331904	45352049	45331043	45352508
CNV.692	2	45725855	45735916	45724088	45738397
CNV.693	2	46120390	46122164	46117914	46122806
CNV.694	2	46488186	46492272	46487373	46492682
CNV.695	2	46724887	46729409	46724721	46730140
CNV.696	2	46887084	46895324	46883739	46898352
CNV.697	2	46970686	46978878	46963287	46982564

CNV.698	2	46988073	47003118	46986781	47008575
CNV.699	2	47350728	47355276	47346568	47359666
CNV.700	2	48159326	48175759	48151785	48182006
CNV.701	2	48849066	48861904	48848515	48862117
CNV.702	2	49031819	49053711	49026928	49056967
CNV.703	2	49129890	49188314	49129316	49188343
CNV.704	2	49217916	49278098	49213601	49281642
CNV.705	2	49522885	49541190	49519654	49545041
CNV.706	2	49660211	49660756	49659120	49661335
CNV.707	2	50597915	50619624	50596607	50626023
CNV.708	2	50868551	51192815	50863662	51193105
CNV.709	2	51198972	51336055	51193105	51337487
CNV.710	2	51338052	51355493	51337596	51357702
CNV.711	2	51546721	51547046	51540445	51547151
CNV.712	2	51734686	51744233	51734676	51745254
CNV.713	2	51870503	51871521	51868178	51874541
CNV.714	2	51936812	51940718	51936592	51943774
CNV.715	2	51967692	51987720	51962230	51988634
CNV.716	2	52255235	52264456	52251936	52265663
CNV.717	2	52273900	52311667	52266907	52317229
CNV.718	2	52413486	52584090	52400156	52585286
CNV.719	2	52602640	52656384	52600309	52657944
CNV.720	2	52735148	52844226	52731979	52851573
CNV.721	2	52880751	52901101	52878016	52908309
CNV.722	2	52978194	52999337	52975226	53000765
CNV.723	2	53055759	53058038	53055692	53058194
CNV.724	2	53219203	53219203	53218853	53219904
CNV.725	2	53248330	53254100	53248305	53254951
CNV.726	2	53363506	53460843	53361731	53460939
CNV.727	2	53652716	53658392	53648450	53659941
CNV.728	2	53843747	53846547	53840021	53847282
CNV.729	2	53857398	53859017	53857361	53859598
CNV.730	2	54220049	54700855	54217050	54702374
CNV.731	2	54766158	54829596	54765174	54832262
CNV.732	2	54971581	54979182	54968395	54979759
CNV.733	2	55016061	55017837	55015764	55018910
CNV.734	2	55027887	55029949	55022709	55036376
CNV.735	2	55225281	55236546	55219799	55238565
CNV.736	2	55324288	55330350	55322895	55340812
CNV.737	2	55621744	55632245	55618957	55636755
CNV.738	2	55908939	55938747	55907463	55941440
CNV.739	2	56393591	56399611	56388213	56402415
CNV.740	2	56613379	56630137	56603985	56632551
CNV.741	2	56698061	56751574	56697359	56759264
CNV.742	2	56961156	56984599	56957466	56984615
CNV.743	2	57008881	57009102	57002137	57011803
CNV.744	2	57035091	57058271	57034299	57060184

CNV.745	2	57085733	57088328	57085705	57090186
CNV.746	2	57160532	57194748	57158169	57195116
CNV.747	2	57394902	57445471	57394885	57447398
CNV.748	2	57499057	57513858	57498209	57516048
CNV.749	2	57664237	57681549	57663776	57685889
CNV.750	2	57843046	57911155	57842075	57911755
CNV.751	2	59544127	59553428	59542149	59555544
CNV.752	2	59622482	59638256	59622422	59643498
CNV.753	2	59749350	59756978	59747694	59759139
CNV.754	2	60484909	60485593	60484680	60486247
CNV.755	2	60595163	60635766	60593634	60637656
CNV.756	2	62197878	62230970	62195826	62232300
CNV.757	2	64038711	64053562	64038356	64054777
CNV.758	2	64342469	64511598	64341723	64512286
CNV.759	2	64516258	64523758	64512286	64527360
CNV.760	2	65716616	65719941	65715444	65723956
CNV.761	2	66306389	66310012	66305945	66310093
CNV.762	2	66783811	66786726	66783680	66786769
CNV.763	2	66845379	66854106	66844120	66861385
CNV.764	2	67400600	67410112	67398540	67410623
CNV.765	2	68096284	68165637	68088399	68166080
CNV.766	2	68179098	68185919	68178609	68186622
CNV.767	2	68680879	68903422	68672721	68908150
CNV.768	2	68920828	68949738	68918682	68950318
CNV.769	2	69208834	69215397	69206549	69215619
CNV.770	2	71346563	71347392	71342451	71350043
CNV.771	2	72132392	72145857	72131775	72148446
CNV.772	2	72249991	72278960	72241504	72279463
CNV.773	2	73854923	73928886	73852537	73933527
CNV.774	2	75155963	75219391	75155355	75226252
CNV.775	2	75366190	75366607	75364299	75366686
CNV.776	2	75381677	75381796	75375179	75381837
CNV.777	2	75443590	75517910	75442644	75523472
CNV.778	2	75821502	75914374	75820691	75919255
CNV.779	2	76544466	76586711	76543105	76591844
CNV.780	2	76769363	76778258	76769272	76779798
CNV.781	2	76845173	76845426	76844466	76847586
CNV.782	2	76936031	76949564	76929056	76956494
CNV.783	2	77178484	77210873	77177328	77210992
CNV.784	2	77221320	77225981	77217310	77226087
CNV.785	2	77232633	77234750	77231346	77234853
CNV.786	2	77900422	77917297	77898110	77917893
CNV.787	2	77973053	78001308	77965521	78001419
CNV.788	2	78410008	78482343	78404361	78483687
CNV.789	2	78702076	78721902	78700957	78725531
CNV.790	2	78887291	78913359	78883975	78917764
CNV.791	2	78992869	79689920	78984048	79697771

CNV.792	2	80089224	80089819	80088940	80091474
CNV.793	2	81507571	81559177	81507306	81562056
CNV.794	2	81661303	81682824	81653225	81687766
CNV.795	2	81812181	81816294	81810572	81817139
CNV.796	2	81818446	81827393	81818331	81828061
CNV.797	2	81856926	81875505	81855725	81876472
CNV.798	2	82196980	82314620	82196945	82317856
CNV.799	2	83064171	83069329	83062834	83071937
CNV.800	2	86285956	86506144	86284286	86513102
CNV.801	2	86936838	86975318	86934745	86978895
CNV.802	2	87109272	88191907	87108211	88193611
CNV.803	2	88429147	88484323	88428754	88492058
CNV.804	2	89133124	95341388	89131067	95341550
CNV.805	2	96082049	96496891	96081973	96498624
CNV.806	2	96671363	96677724	96664037	96679238
CNV.807	2	97059865	97097107	97059029	97101834
CNV.808	2	97315806	97340885	97308792	97343799
CNV.809	2	97717272	98273631	97715154	98274527
CNV.810	2	98960331	98970419	98958691	98971401
CNV.811	2	99540101	99563298	99539884	99564428
CNV.812	2	99840143	99924555	99830510	99926946
CNV.813	2	100548791	100554945	100547759	100555175
CNV.814	2	101961951	101970319	101959688	101970854
CNV.815	2	102661735	102673492	102655344	102673775
CNV.816	2	102676923	102692511	102676129	102698360
CNV.817	2	102840107	102855298	102839755	102861949
CNV.818	2	104463192	104467941	104462512	104471726
CNV.819	2	104851479	104862568	104850784	104864108
CNV.820	2	106135637	106140282	106132543	106142178
CNV.821	2	106348112	106351211	106347216	106354138
CNV.822	2	106384517	106391338	106384498	106391913
CNV.823	2	106872736	106887408	106866680	106887870
CNV.824	2	106898380	106901931	106898283	106909238
CNV.825	2	107211681	107218740	107211045	107218831
CNV.826	2	107261144	107301223	107258754	107304040
CNV.827	2	107826777	107864069	107824929	107864737
CNV.828	2	107941819	108022271	107936757	108032914
CNV.829	2	108177053	108280904	108175761	108285043
CNV.830	2	108440290	108508052	108440138	108520376
CNV.831	2	108701607	108702571	108701310	108703546
CNV.832	2	109163093	109169103	109162706	109178504
CNV.833	2	109578513	109586437	109576838	109588356
CNV.834	2	109692425	109695769	109688972	109698148
CNV.835	2	110467473	111396765	110462496	111397799
CNV.836	2	111979218	112396181	111972750	112396722
CNV.837	2	112444132	112451387	112442561	112453204
CNV.838	2	112502453	112515328	112500035	112516882

CNV.839	2	112868426	112917028	112866758	112920343
CNV.840	2	113872051	113889652	113871822	113891842
CNV.841	2	114060553	114069021	114060118	114071171
CNV.842	2	114149060	114186957	114148996	114236536
CNV.843	2	114325089	114384353	114324463	114389463
CNV.844	2	114429573	114453589	114429211	114453880
CNV.845	2	115164459	115172820	115164298	115173745
CNV.846	2	115396421	115400999	115386847	115403214
CNV.847	2	115526908	115542394	115513600	115542505
CNV.848	2	115794769	115801249	115794474	115802456
CNV.849	2	115891527	115894284	115891114	115897370
CNV.850	2	116532520	116675350	116530428	116677020
CNV.851	2	116974422	117013066	116971659	117017559
CNV.852	2	117476453	117932723	117476014	117933023
CNV.853	2	118096131	118243133	118094523	118244990
CNV.854	2	118315232	118329461	118315191	118331829
CNV.855	2	118390355	118404852	118389642	118416015
CNV.856	2	118937683	118964860	118936392	118966616
CNV.857	2	120175056	120180566	120174136	120180986
CNV.858	2	121049152	121049152	121048878	121049166
CNV.859	2	121311000	121343284	121310139	121348224
CNV.860	2	123375961	123385193	123371222	123386820
CNV.861	2	123472179	123483668	123472119	123483921
CNV.862	2	124060508	124163193	124056024	124163281
CNV.863	2	124243806	124360519	124242210	124360539
CNV.864	2	125027970	126071265	125022866	126076783
CNV.865	2	126202023	126452049	126193512	126453382
CNV.866	2	126871605	126895981	126865662	126897465
CNV.867	2	127278793	127281547	127276098	127281778
CNV.868	2	127719856	127726891	127714880	127727338
CNV.869	2	127832684	127890986	127830299	127896263
CNV.870	2	128220756	128226708	128220429	128227111
CNV.871	2	129016580	129026749	129016127	129028891
CNV.872	2	129638478	129640329	129638078	129646448
CNV.873	2	129784159	129789108	129781214	129789335
CNV.874	2	130114143	130120748	130113561	130123649
CNV.875	2	130196099	130215281	130195492	130216043
CNV.876	2	130242500	130255110	130241617	130255181
CNV.877	2	130355506	130360682	130354239	130361020
CNV.878	2	130373789	130382299	130372333	130384554
CNV.879	2	130549275	130558377	130545565	130559029
CNV.880	2	130762892	131207769	130761944	131208117
CNV.881	2	131244208	131481632	131212846	131483965
CNV.882	2	131950542	132279325	131946313	132280836
CNV.883	2	132509673	132560272	132509119	132560381
CNV.884	2	132704084	133358589	132703277	133361070
CNV.885	2	133900979	133936908	133900128	133945444

CNV.886	2	133949948	133960852	133945444	133964488
CNV.887	2	134056154	134154923	134055974	134156074
CNV.888	2	134230217	134236942	134229008	134237077
CNV.889	2	135047919	135078574	135047888	135079580
CNV.890	2	136462661	136467164	136462554	136467218
CNV.891	2	137942769	137955456	137942464	137956006
CNV.892	2	137983647	137984261	137982148	137985285
CNV.893	2	138384826	139065889	138382750	139068616
CNV.894	2	139802461	139844786	139802172	139848110
CNV.895	2	140065236	140069761	140064644	140071563
CNV.896	2	140086004	140086004	140085417	140087181
CNV.897	2	140252510	140287264	140245380	140287352
CNV.898	2	140361372	140380328	140360216	140380671
CNV.899	2	140711999	140779944	140709142	140782556
CNV.900	2	140991487	140995770	140984797	141000302
CNV.901	2	141043970	141058834	141043143	141059230
CNV.902	2	141434484	141443398	141434354	141446431
CNV.903	2	141547533	141558025	141546920	141560690
CNV.904	2	142078003	142078226	142077884	142078322
CNV.905	2	142106831	142108901	142101663	142109093
CNV.906	2	142342699	142344619	142333874	142344650
CNV.907	2	142726145	142731390	142723425	142732467
CNV.908	2	143360629	143363158	143358557	143363171
CNV.909	2	144192472	144202272	144190929	144212698
CNV.910	2	145589091	145598977	145588358	145599194
CNV.911	2	146006157	146055981	146002153	146056286
CNV.912	2	146309983	146315024	146309463	146317392
CNV.913	2	146641347	146643868	146641073	146644471
CNV.914	2	146858851	146881053	146858369	146881099
CNV.915	2	147941342	147946882	147934790	147948679
CNV.916	2	148901257	148922704	148900148	148929937
CNV.917	2	149008402	149022674	149004714	149024595
CNV.918	2	150751726	150755192	150745883	150760021
CNV.919	2	151031451	151037732	151030802	151039208
CNV.920	2	151144891	151150543	151143010	151150611
CNV.921	2	151990393	152056920	151986397	152059484
CNV.922	2	152430227	152434786	152429679	152434825
CNV.923	2	152490390	152499136	152487983	152499374
CNV.924	2	153715962	153721471	153715382	153727760
CNV.925	2	153771503	153800774	153771334	153801268
CNV.926	2	153949686	153961349	153945439	153964194
CNV.927	2	154478253	154520320	154471215	154525542
CNV.928	2	154808546	154810786	154806692	154812333
CNV.929	2	154883942	154902505	154883112	154907564
CNV.930	2	155685078	155686043	155684173	155688875
CNV.931	2	155694839	155703266	155692796	155703558
CNV.932	2	155887451	155927441	155882777	155937872

CNV.933	2	156374770	156398624	156366936	156400540
CNV.934	2	157729401	157758027	157725698	157763818
CNV.935	2	158549742	158576191	158543078	158577630
CNV.936	2	158801474	158802326	158800532	158803785
CNV.937	2	159653823	159670449	159653721	159670493
CNV.938	2	159956896	159963306	159956520	159964196
CNV.939	2	160658895	160661107	160658798	160661278
CNV.940	2	160669255	160673425	160666409	160674855
CNV.941	2	161205420	161214857	161196868	161216721
CNV.942	2	161410542	161416709	161409668	161416792
CNV.943	2	161542328	161557056	161541313	161559511
CNV.944	2	162278238	162289146	162277301	162290072
CNV.945	2	162462671	162470769	162462535	162471392
CNV.946	2	163600553	163631047	163596659	163631302
CNV.947	2	163681969	163689398	163675541	163696185
CNV.948	2	163820126	163839108	163819925	163848347
CNV.949	2	164643488	164645126	164636852	164646642
CNV.950	2	164725276	164737668	164724148	164738679
CNV.951	2	165686419	165710142	165681579	165713672
CNV.952	2	168008976	168016789	168005727	168022374
CNV.953	2	168172932	168181440	168163730	168187928
CNV.954	2	169798005	169830515	169797678	169833310
CNV.955	2	169952102	169964551	169949329	169967174
CNV.956	2	172984359	172992806	172984009	172994091
CNV.957	2	173550168	173559068	173550013	173562251
CNV.958	2	174589940	174597568	174588058	174597656
CNV.959	2	175386051	175395232	175385771	175398387
CNV.960	2	177249726	177277336	177247835	177278437
CNV.961	2	177370261	177382589	177370189	177382699
CNV.962	2	178550061	178577002	178546919	178584008
CNV.963	2	180056011	180083629	180055857	180085736
CNV.964	2	180409483	180429164	180409470	180430069
CNV.965	2	180926228	180929966	180925205	180930698
CNV.966	2	181127404	181131811	181124724	181133435
CNV.967	2	181921723	181932870	181920338	181935498
CNV.968	2	183113130	183116652	183112782	183116900
CNV.969	2	184069020	184093998	184068573	184094262
CNV.970	2	184261432	184339696	184258772	184341553
CNV.971	2	184369241	184378863	184368648	184379416
CNV.972	2	184571310	184580939	184569070	184580950
CNV.973	2	184602735	184603704	184597076	184607919
CNV.974	2	184669897	184697939	184663103	184700796
CNV.975	2	184793800	184808794	184793232	184810947
CNV.976	2	184814480	184831241	184811670	184846794
CNV.977	2	185014650	185064842	185013409	185072376
CNV.978	2	185100535	185136918	185094677	185141033
CNV.979	2	185157484	185170450	185154896	185171057

CNV.980	2	185213629	185235861	185213481	185238581
CNV.981	2	185243929	185244043	185243899	185244063
CNV.982	2	185257053	185264965	185244063	185265274
CNV.983	2	185316880	185325519	185316676	185325786
CNV.984	2	185431570	185462424	185430182	185467914
CNV.985	2	185753993	185774139	185753080	185775157
CNV.986	2	186042609	186157627	186040111	186162731
CNV.987	2	186299721	186884137	186298456	186888968
CNV.988	2	188065554	188069276	188053476	188069538
CNV.989	2	188094542	188142085	188093586	188153257
CNV.990	2	188662464	188782216	188660929	188783773
CNV.991	2	188935831	189027247	188934309	189032187
CNV.992	2	189112811	189129179	189112399	189139076
CNV.993	2	189182885	189244556	189182307	189253820
CNV.994	2	189567956	189581242	189567787	189581417
CNV.995	2	190231976	190240405	190231962	190248651
CNV.996	2	190505766	190514273	190504787	190514461
CNV.997	2	191545147	191551209	191544557	191552632
CNV.998	2	191659147	191672128	191657773	191672765

Supplementary Table 2. TCGA datasets used in Tangent WES analysis.

TCGA Study Abbreviation	TCGA Study Name	Number of normal samples	Number of tumor samples
STAD	Stomach Adenocarcinoma	57	51
LUSC	Lung Squamous Cell Carcinoma	48	48
LGG	Lower Grade Glioma	14	14
PRAD	Prostate Adenocarcinoma	10	10