Genome analysis

The Tangent copy-number inference pipeline for cancer genome analyses

Barbara Tabak[1,2,3,*], Gordon Saksena[1,*], Stefano Monti[1,4], Jeffrey Gentry[1], Bryan C. Hernandez[1], Michael O'Kelly[1], Marc-Danie Nazaire[1], Barbara Hill Meyers[1], Scott L. Carter[1], Andrew D. Cherniack[1], Steven E. Schumacher[1,2], Nam H. Pho[1], Travis I. Zack[1,2,5], Nicholas Stransky[1], Joshua Gould[1], David Twomey[1], Mark Nadel[1], Wendy Winckler[1], Matthew Meyerson[1,2,6,**], Rameen Beroukhim[1,2,6,**], and Gad Getz[1,7,**]

[1]Cancer Program, Broad Institute, Cambridge, MA 02142, USA

[2]Departments of Medical Oncology and Cancer Biology and Center for Cancer Genome Characterization, Dana Farber Cancer Institute, Boston, MA 02115, USA

[3]University of Massachusetts Medical School, Worcester, MA 01655

[4]Section of Computational Biomedicine, B.U. School of Medicine, Boston, MA 02118, USA

[5]Biophysics Department, Harvard University, Boston, MA 02142, USA

[6]Departments of Pathology and Medicine, Harvard Medical School, Boston MA 02115, USA

[7]Massachusetts General Hospital, Boston, MA, 02114, USA

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Summary: As part of The Cancer Genome Atlas (TCGA), the Broad Institute Genome Characterization Center has generated copy-number profiles using single nucleotide polymorphism (SNP) array data from over 10,000 pairs of cancer and matched normal DNA samples. We describe the copy-number inference pipeline, which begins with raw probe-level intensity data and concludes with the identification of genes potentially targeted by somatic copy-number alterations (SCNAs).

Availability: The pipeline is available as a GenePattern pipeline at www.broadinstitute.org/cancer/cga/copynumber_pipeline.

Contact: matthew_meyerson@dfci.harvard.edu, rameen@broadinstitute.org, gadgetz@broadinstitute.org

Supplementary information: Supplementary methods, code, and data are available at Bioinformatics online

1 INTRODUCTION

High-resolution microarrays enable fine-scale characterization of somatic copy-number alterations (SCNAs) in cancer genomes and facilitate the discovery of genes that drive cancer (Garraway et al., 2005; Weir *et al.*, 2007; TGCA Network, 2008; Beroukhim *et al.*, 2010; Northcott *et al.*, 2012; Zack *et al.*, 2013). We developed a pipeline to process such data with special attention to noise reduction, artifact removal, and quality control. Our pipeline is for use with Affymetrix SNP 6.0 arrays, containing for 906,600 "SNP markers" associated with single nucleotide polymorphisms and 946,000 "copy number markers" at other locations (www.affymetrix.com/support/technical/technotes/cn_snp_variation_technote.pdf), and has been the basis for analyses of all such data for The Cancer Genome Atlas (TCGA). However, the underlying techniques can be extended to other platforms.

We describe this pipeline here, including two techniques that have not been previously described: copy-number inference to calibrate copy-number probes, and tangent normalization to reduce systematic noise. The methods we describe are instantiated in the GenePattern pipeline CopyNumberInferencePipeline.

2 ALGORITHM

Our pipeline includes six primary modules (Figure 1) to:

(1) Normalize and combine probe intensities to produce a single intensity per probeset.

(2) Calibrate probesets to associate copy-numbers to each genomic locus.

(3) Reduce biological noise from germline copy-number variations (CNVs).

(4) Reduce systematic experimental noise.

(5) Reduce random experimental noise through segmentation.

(6) Identify genes that are amplified or deleted more than expected by chance using GISTIC2.0 (Mermel *et al.*, 2011).

[*]These authors contributed equally to this work.

[**]To whom correspondence should be addressed.

The TCGA .CEL files, inputs to our pipeline, are available at the TCGA Data Portal (tcga-nci.nih.gov/tcga) as Level 1 Data. Output files of modules (1) – (4) are available as Level 2 Data; normalized, segmented data (outputs of module (5)) are uploaded as Level 3 Data (see Supplementary Figure 1 and Supplementary Methods).

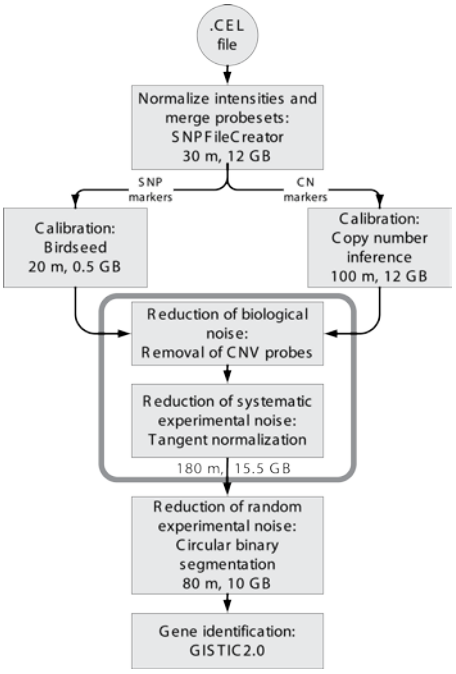

Figure 1. Flow diagram indicating key modules of the Tangent copy-number inference pipeline with compute time and memory requirements for each for processing a 96 sample batch (platform: Linux RedHat 5.5 64bit, 12 cores with AMD Opteron 6180 SE 2.5Ghz, 32GB RAM).

2.1 Normalizing and combining intensities over probesets

Each genomic locus and SNP allele is represented by multiple probes on the array (a "probeset"). Probe-level intensities are represented in .CEL files generated by Affymetrix GeneChip Command Console. SNPFileCreator, a Java implementation of the dChip signal intensity determination algorithm (Li and Wong, 2001b; 2001a), is used to normalize and merge intensity values for each probeset. Probe intensities across a sample are first scaled to achieve a median brightness value of 1000 and then are subjected to quantile normalization. The normalized probe intensities of each sample are mapped to a reference sample using model-based expression indices (MBEI). Median polish is then applied to each probeset across samples to produce one value per probeset for each sample. SNP array data are typically generated in batches (defined by joint PCR amplification step); we apply SNPFileCreator to all arrays in each batch.

2.2 Calibration and Copy Number Inference

Probeset intensities are mapped to copy-number levels ("calibrated") on a batch-by-batch basis, assuming a linear relationship between signal intensity and copy-number. Calibration of a probeset 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 (Korn *et al.*, 2008) is used to calibrate probesets for each allele, using intensity data collected from normal samples, allele-specific background, and scale parameters. The resulting copy-numbers for the two alleles are summed to obtain total copy-number estimates.

Calibration of copy-number markers relies on SNP array data we have generated from 5 cell lines with known variation of copy-number of the X-chromosome from 1 to 5. We applied linear regression to these data to determine the background and scale factor for each probeset on X, and modeled these parameters as a function of local sequence features and median intensity across samples (see Supplementary Methods). This model enables calibration of all probesets across the genome.

2.3 Reduction of Biological Noise

A major source of "noise" in somatic copy-number profiles is germline copy-number variations (CNVs) misidentified as SCNAs. We address this issue by removing probesets overlapping CNVs. We identify CNVs as genomic regions that display extensive variation across normal samples. In some cases, these may reflect noisy probes rather than true CNVs.

2.4 Reduction of Systematic Experimental Noise

We have found systematic variations in signal intensities across the genome between analyses of the same DNA using different arrays, both within and across batches (Supplementary Figure S2). These may reflect variations in experimental conditions between different arrays and can lead to the false appearance of SCNAs that recur across samples.

Tangent normalization assumes that noise (in log-transformation units) in SNP array data is distributed according to a similar pattern in cancer samples to normal samples. Therefore, to minimize noise, we subtract estimated noise profiles individually calculated for each tumor using data from all normal samples. Specifically, we determine the weighted sum of noise profiles from all normal samples that most closely matches each tumor's profile, and subtract it from that tumor. These weighted sums of normal profiles lie within a subspace (the "tangent plane") of the space containing all possible copy-number profiles; the weighted sum used for each tumor is that tumor's projection into this subspace. (See Supplementary Methods for details and results with TCGA data.)

2.5 Segmentation and GISTIC2.0

Random noise is removed by Circular Binary Segmentation (Venkatraman and Olshen, 2007). GISTIC2.0 is then applied to determine significantly amplified or deleted SCNAs.

2.6 Quality Control

Automated quality control is integrated into various stages of the pipeline. Assessment of the DNA quality of normal samples improves the accuracy of our calibration of SNP probesets. A 'tumor-detector' ensures that tumors mislabeled as normals and contaminated normal samples do not compromise CNV detection or the tangent noise model. Noise assessments before and after segmentation gate the quality of samples that are input into GISTIC2.0. (See Supplementary Methods for details.)

3 ACKNOWLEDGEMENTS

We would like to acknowledge support from our colleagues from the Broad Genomics Platform and The Cancer Genome Atlas Project.

Funding: This work was supported by the National Institutes of Health [U24CA126546(M.M., G.G., R.B.), U24CA143845 (G.G, M.M.), U24CA143867 (M.M., G.G., R.B.), and U54CA143798 (R.B.)]; and the Pediatric Low-Grade Astrocytoma Foundation (R.B.).

4 REFERENCES

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

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

Garraway,L.A. *et al.* (2005) Integrative genomic analyses identify MITF as a lineage survival oncogene amplified in malignant melanoma. *Nature*, **436**, 117–122.

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

Li,C. and Wong,W.H. (2001a) Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application. *Genome Biology*, **2**, 1–11.

Li,C. and Wong,W.H. (2001b) Model-based analysis of oligonucleotide arrays:Expression index computation and outlier detection. *Proc. Nat. Acad. Sci.*, **98**, 1–6.

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

Robinson,J.T. *et al.* (2011) Integrative genomics viewer. *Nature Biotechnology*, **29**, 24–26.

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

Thorvaldsdottir,H. *et al.* (2012) Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in Bioinformatics*, 1 -15.

Northcott,P.A. *et al.* (2012) Subgroup-specific structural variation across 1,000 medulloblastoma genomes. *Nature*, **488**, 49–56.

Van Loo,P. and Nordgard,S.H. (2010) Allele-specific copy number analysis of tumors. *Proc. Nat. Acad. Sci.*, **107**, 16910–16915.

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

Weir,B.A. *et al.* (2007) Characterizing the cancer genome in lung adenocarcinoma. *Nature*, **450**, 893–898.

Zack TI *et al.* (2013) Pan-cancer patterns of somatic copy number alteration. *Nature Genetics*, **45,** 1134-1140.