Bioinformatics, 2017, 0–0
doi: 10.1093/bioinformatics/xxxxx
Advance Access Publication Date: DD Month YYYY
Application Notes

Genome analysis

Integrative pipeline for profiling DNA copy number and inferring tumor phylogeny

Yuchao Jiang[1,2,*], Hao Chen[3], Zilu Zhou[4] and Nancy R Zhang[5]

[1] Department of Biostatistics, University of North Carolina, Chapel Hill, NC 27599, USA, [2] Department of Genetics, University of North Carolina, Chapel Hill, NC 27599, USA, [3] Department of Statistics, University of California, Davis, Davis, CA 95616, USA, [4] Genomics and Computational Biology Graduate Program, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA, [5] Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104, USA.

[*] To whom correspondence should be addressed.
Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Summary: Copy number variation is an important and abundant source of variation in the human genome, which has been associated with a number of diseases, especially cancer. Massively parallel next-generation sequencing allows copy number profiling with fine resolution. Such efforts, however, have met with mixed successes, with setbacks arising partly from the lack of reliable analytical methods to meet the diverse and unique challenges arising from the myriad experimental designs and study goals in genetic studies. In cancer genomics, detection of somatic copy number changes and profiling of allele-specific copy number (ASCN) are complicated by experimental biases and artifacts as well as normal cell contamination and cancer subclone admixture. Furthermore, careful statistical modeling is warranted to reconstruct tumor phylogeny by both somatic ASCN changes and single nucleotide variants. Here we describe a flexible computational pipeline, integrating multiple related statistical software, for copy number profiling and downstream analyses in disease genetic studies.

Availability and implementation: CODEX, CODEX2, FALCON, FALCON-X, iCNV, and Canopy are publicly available R-packages in CRAN, Bioconductor, and GitHub.
Contact: yuchaoj@email.unc.edu
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Copy number variations (CNVs) refer to duplications and deletions that lead to gains and losses of large genomic segments. CNV is pervasive in the human genome and plays a causal role in genetic diseases. With the dramatic growth of sequencing capacity and the accompanying drop in cost, massively parallel next-generation sequencing (NGS) offers appealing platforms for genome-wide CNV detection. In this note, we describe an analysis pipeline that integrates multiple aspects of CNV analysis, which can flexibly adapt to diverse study designs and research goals.

Despite the rapid technological development, CNV detection by high-throughput sequencing still faces analytical challenges due to the rampant biases and artifacts that are introduced during library preparation and sequencing. Proper data normalization is crucial for sensitive and robust CNV detection, regardless of sequencing design and library preparation protocol. In whole-exome sequencing (WES) and targeted sequencing, where technical biases are usually magnitudes larger than CNV signals, data normalization is usually the pivotal step in affecting detection accuracy. The pipeline that we describe start with data normalization using the statistical software CODEX (Jiang, et al., 2015) and CODEX2 (unpublished), which allow full-spectrum CNV profiling and are sensitive to both common and rare variants. CODEX2 is applicable to study designs with and without negative controls.

Many large-scale genetic studies involve samples that have previously collected microarray data. It is currently unclear how microarray data can be used to improve sensitivity and robustness of the sequencing-based analyses. The pipeline proposed here seamlessly combine CODEX-normalized sequencing data with array-based log-ratio and B-allele-frequency measurements, through the statistical software iCNV (unpublished).

Germline and somatic copy number changes are common in cancer and are associated with tumorigenesis and metastasis. Characterization of these

DNA copy number changes is important for both the basic understanding of cancer and for its diagnosis. In addition to the detection of total copy number changes, sequencing data give, at germline heterozygous loci, reads containing both alleles allowing the disambiguation of allele-specific copy number (ASCN), which quantifies the number of somatic copies of each inherited allele. Compared to total copy number analysis, ASCN analysis gives a more complete picture of the copy number states, including copy-neutral loss of heterozygosity (LOH), which cannot be detected by total copy number analysis. Here we show how the software FALCON (Chen, et al., 2015) and FALCON-X (Chen, et al., 2017) for estimating ASCN integrate with CODEX and CODEX2, for application to WES data from tumor and matched normal, with adjustment of site-specific background noise and allelic bias.

Tumors are heterogeneous, genetically related populations of cells undergoing constant evolution. Much effort has been devoted to the reconstruction of the evolutionary phylogeny of tumors from bulk DNA sequencing data (see reviews in Kuipers, et al., 2017). Proper data normalization and careful statistical modeling are important for accurate ASCN profiling, which can then be used for phylogenetic inference. In addition to somatic ASCN changes, single nucleotide variants (SNVs) also provide valuable information for the reconstruction of the tumor phylogeny and assessment intratumor heterogeneity. We show that the Canopy statistical framework (Jiang, et al., 2016) for reconstruction of longitudinal and spatial clonal evolutionary history can be applied to the outputs from FALCON and FALCON-X in an integrative analysis. This enables researchers to get both total and allele-specific DNA copy number calls and tumor phylogeny directly from BAM files.

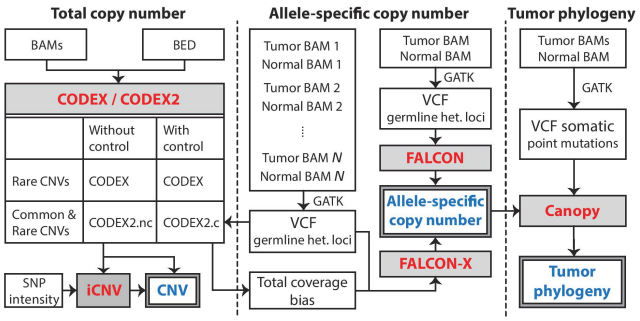

Figure 1. A flowchart outlining the procedures for profiling CNV, ASCN, and reconstructing tumor phylogeny. CNVs with common and rare population frequencies can be profiled by CODEX and CODEX2, with and without negative control samples. iCNV integrates sequencing and microarray data for CNV detection. ASCNs can be profiled by FALCON and FALCON-X using allelic read counts at germline heterozygous loci. Canopy infers tumor phylogeny using somatic SNVs and ASCNs.

Analysis Scenario			Pipeline
Goal	Study Design	Sample requirements	
Total copy number analysis of normal*	WES/WGS with or without matched array	>20 samples	CODEX2.nc → iCNV
Total copy number analysis of tumor	WES/WGS /targeted sequencing	>20 normal controls	CODEX2.c
Tumor allele-specific copy number	WGS	1 tumor with matched normal sample	FALCON
Tumor allele-specific copy number	WES/WGS tumor with matched normal	1 or more tumor samples, multiple (>20) normal controls (no need to be matched)	CODEX2.c → FALCONX
Tumor phylogeny analysis	WGS	Multiple spatially or temporally separated tumor samples with matched	FALCON → CANOPY
Tumor phylogeny analysis	WES/WGS tumor with matched normal	Multiple spatially or temporally separated tumor samples, multiple (>20) normal controls (no need to be matched)	CODEX2.c → FALCONX → CANOPY

Table 1. Analysis scenarios and pipeline design. The last column shows the sequencing of software that should be used for each analysis scenario. *By "normal" we mean samples that are not derived from tumor tissue. Broad copy number changes are not expected to cover the genome in the normal controls.

2 Methods

The possible analysis scenarios are listed in Table 1. Figure 1 gives an outline for the relationship between the software: CODEX and CODEX2 perform read depth normalization for total copy number profiling; read depth normalized by CODEX/CODEX2 is received by iCNV, which combines it with allele-specific read counts and microarray data to detect CNVs; FALCON and FALCON-X perform ASCN analysis; and Canopy receives input from FALCON/FALCON-X to perform tumor phylogeny reconstruction. All methods are compiled as R packages publicly available in CRAN, Bioconductor, and GitHub (see Supplement for details).

CODEX (Jiang, et al., 2015) adopts a Poisson latent factor model for normalization to remove biases due to GC content, exon capture and amplification efficiency, and latent systemic artifacts. CODEX also includes a Poisson-likelihood based recursive segmentation procedure that explicitly models the count-based sequencing data. CODEX2 (unpublished) builds on CODEX with a significant improvement of sensitivity for both rare and common variants, thus allowing full-spectrum CNV detection. CODEX2 can be applied to two scenarios: the case control scenario (CODEX2.c in Figure 1) where the goal is to detect CNVs that are enriched in the case samples; and the scenario where control samples are not available (CODEX2.nc in Figure 1) and the goal is simply to profile all CNVs. CODEX and CODEX2 take as input assembled BAM files as well as bed files specifying targets for WES and targeted sequencing. Outputs are tab-delimited text files with copy number calls, which can be loaded into IGV (Robinson, et al., 2011) for visualization.

iCNV (unpublished) uses the normalized coverage from CODEX/CODEX2, and makes use of sequenced reads at inherited single nucleotide polymorphism (SNP) positions for CNV detection. These heterozygous loci are shown to be valuable in improving detection and genotyping accuracy. If microarray (e.g., Illumina or Affymetrix arrays) data are available, iCNV also integrates log-ratios and B-allele frequencies from these platforms to boost accuracy and enable CNV detection in intronic regions. iCNV takes as input normalized coverage by CODEX/CODEX2, allelic frequency at inherited SNP positions from sequencing, and log-ratio and B-allele frequency from SNP array. An integrated hidden Markov model is implemented with parameters estimated by Baum-Welch algorithm. iCNV generates CNV calls with quality scores and infers integer copy number if necessary.

For ASCN estimation in a matched tumor-normal setting, FALCON (Chen, et al., 2015) is based on a change-point model on a process of a mixture of two bivariate Binomial distributions, which explicitly models the copy numbers of the two chromosomal haplotypes and corrects for local allele-specific coverage biases. FALCON takes as input allelic read counts at germline heterozygous loci called by the Genome Analysis Toolkit (GATK) and outputs ASCN estimates with genome segmentations. FALCON is in general sufficient for whole-genome sequencing (WGS) data. However, for WES data, due to the additional amplification step for targeted exons, biases and artifacts cannot be fully captured by comparing the tumor sample to the matched normal sample. FALCON-X (Chen, et al., 2017) extends upon FALCON, where it further utilizes site-specific bias estimated from cross-sample normalization by CODEX2.c as additional input (Figure 1) to correct for biases and artifacts. That is, FALCON-X takes as inputs allelic read counts at germline heterozygous loci called by GATK and total coverage biases for each of these loci estimated by CONDEX2.c and outputs ASCN estimates with genome segmentations.

Canopy (Jiang, et al., 2016) identifies subclones within a tumor, determines the mutational profiles of these subclones, and infers the tumor's phylogenetic history by NGS data from temporally and/or spatially separated tumor resections from the same patient. Canopy jointly models so-

matic copy number changes and SNVs in a similar fashion to non-negative matrix factorization and adopts a Bayesian framework to reconstruct phylogeny with posterior confidence assessment, which is useful when the data support multiple tree configurations. Canopy takes as input both somatic ASCN changes returned by FALCON/FALCON-X as well as somatic SNVs and outputs tumor phylogenetic trees with somatic mutations placed along tree branches and subclones placed at the leaves.

3 Results

The proposed pipeline adapts to different study designs and research goals (Table 1). For population genetic and disease association studies, one would start with read depth normalization using CODEX/CODEX2, followed by CNV calling using iCNV. Pipeline for iCNV analysis with CODEX's normalization is included in the Supplement. For cancer genomics studies where the goal is to obtain ASCNs and reconstruct tumor clonal history, one would start with read depth normalization using CODEX/CODEX2, followed by ASCN profiling using FALCON/FALCON-X and clonal history analysis using Canopy. Example R-scripts performing FALCON-X analysis using the output from CODEX2 and Canopy analysis using the output from FALCON are given in the Supplement.

In this short note, we demonstrate a cancer phylogenetic study, the pipeline for which is more complicated than the rest. We adopt a WGS dataset of normal, primary tumor, and relapse genome of a neuroblastoma patient PASGAP (Eleveld, et al., 2015). Since this is WGS, and not WES data, we can simply apply FALCON to profile ASCN. If this were WES data, one can substantially improve accuracy by first applying CODEX2 to perform normalization, followed by ASCN analysis using FALCON-X.

On this data set, the ASCN estimated by FALCON are shown to be consistent with the output from Sequenza (Favero, et al., 2015) (Supplementary Figure S1-S2). Sequenza simultaneously estimates purity and ploidy of the tumor, and outputs integer absolute copy number of each allele assuming all copy number changes are clonal events. The caveat is that purity and ploidy are often not identifiable from the data (Supplementary Figure S3) and estimation of integer-valued ASCN requires the assumption of clonal copy number alternations, which is usually not known prior to knowing the phylogenetic history. FALCON has less stringent assumptions and returns fractional ASCNs with standard deviations estimated by a bootstrap-based method (Supplementary Table S2). We focus on chr4p loss, chr6p gain, chr14 LOH, and chr20p loss, of which the primary and relapse have distinct profiles, to reconstruct the tumor phylogeny.

Somatic SNVs are called by GATK and MuTect and further annotated by ANNOVAR (Wang, et al., 2010). Stringent quality control procedures are adopted to remove possible germline mutations, low-quality indels, variants with missing genotypes, variants within segmental duplication regions, and variants with low coverage, etc. We also focus on variants that are annotated as deleterious by at least one scoring metric, resulting in 32 high-confidence SNVs – 7 are unique to the primary, 21 are unique to the relapse, and 4 are shared (Supplementary Table S1, Supplementary Figure S4). Notably, a relapse-specific SNV in the gene *CORIN* lies within the heterozygous deletion in chr4 and has a variant allele frequency of 68.8% in the relapse.

The SNVs, as well as the ASCNs, are used as input for Canopy, which returns a most likely tree with four leaves, including one leaf representing the normal cells (Supplementary Figure S4). The purity estimates are largely concordant with those returned by Sequenza, which selects posterior modes with ploidy equal to two (Supplementary Figure S3). The heterozygous deletion in chr4p occurs on the first tree branch, followed by the point mutation in gene *CORIN* on the remaining allele, together with the other relapse-specific SNVs. These SNVs exist in clone 2 and 3, which have cellular fractions of 41.5% and 30.2%, respectively. Chr6p duplication separates clone 3 from clone 2, which are merged if the tree is only allowed to have three leaves.

ACKNOWLEDGEMENTS

The authors thank Drs. Derek A. Oldridge, Sharon J. Diskin, and John M. Maris for providing the neuroblastoma dataset.

Funding: This work was supported by the National Institutes of Health (NIH) grant P01 CA142538 to YJ and R01 HG006137 to NRZ.

Conflict of Interest: none declared.

References

Chen, H., *et al.* Allele-specific copy number profiling by next-generation DNA sequencing. *Nucleic Acids Res* 2015;43(4):e23.

Chen, H., *et al.* Allele-specific copy number estimation by whole exome sequencing. *The Annals of Applied Statistics* 2017;11(2):1169-1192.

Eleveld, T.F., *et al.* Relapsed neuroblastomas show frequent RAS-MAPK pathway mutations. *Nat Genet* 2015;47(8):864-871.

Favero, F., *et al.* Sequenza: allele-specific copy number and mutation profiles from tumor sequencing data. *Ann Oncol* 2015;26(1):64-70.

Jiang, Y., *et al.* CODEX: a normalization and copy number variation detection method for whole exome sequencing. *Nucleic Acids Res* 2015;43(6):e39.

Jiang, Y., *et al.* Assessing intratumor heterogeneity and tracking longitudinal and spatial clonal evolutionary history by next-generation sequencing. *Proc Natl Acad Sci U S A* 2016;113(37):E5528-5537.

Kuipers, J., Jahn, K. and Beerenwinkel, N. Advances in understanding tumour evolution through single-cell sequencing. *Biochim Biophys Acta* 2017;1867(2):127-138.

Robinson, J.T., *et al.* Integrative genomics viewer. *Nat Biotechnol* 2011;29(1):24-26.

Wang, K., Li, M. and Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* 2010;38(16):e164.