Structural Variation Baseline Genome Set: Data Format Description

October 2011

Introduction

Complete Genomics Structural Variation (SV) analysis identifies junctions between regions of the genome being sequenced that are not adjacent on the reference genomes. Such regions serve as evidence for the presence of structural variation events such as deletions, inversions, and translocations. Reported junctions are associated with various annotations, allowing researchers to quickly filter them based on confidence, impact on gene function, and overlap with repetitive elements and known SVs in dbSNP. Additionally, for each junction detected, Complete Genomics annotates the frequency with which the specific junction also appears in a set of baseline genomes. This baseline genome set is comprised of 52 genomes from normal, disease-free individuals. Frequency measures provide a powerful approach for filtering out junctions that represent technical artifacts or ancient events that are unlikely to be relevant to the disease under study.

Important: SV analysis from the Complete Genomics Assembly Pipeline versions prior to 2.0 used a different set of baseline genomes. The frequency of SVs called in the baseline data described in this document will differ from frequencies reported in files produced by previous versions of the Assembly Pipeline software.

The SV Baseline Genome Data Set

Given its utility in the Complete Genomics SV analysis pipeline, we are providing the composite files of the underlying data for the 52 baseline genomes used for the SV portion of the Assembly Pipeline version 2.0. These files, ***B36baseline-junctions.tsv*** and ***B37baseline-junctions.tsv***, are distinguished

by the human genome reference build used (NCBI Build 36 and 37, respectively) and can be downloaded from:

ftp://ftp2.completegenomics.com/Baseline_Genome_Set/SVBaseline/

The files may be useful for a number of reasons, including:

- Updating *FrequencyInBaselineGenomeSet* annotation in junction files from earlier versions of the Assembly Pipeline with frequency information provided by new baseline set.
- Running CGA™ Tools junctiondiff with user-specified parameters on merged junctions in the baseline file and a Complete Genomics junctions file to filter for junctions that are absent from the baseline genomes.

The rest of this document lists the genomes used in the SV baseline set and briefly describes the processing steps involved in generating the baseline.

Methods of Baseline Generation

The 52 baseline genome set is comprised of genomes included in the Complete Genomics Diversity Panel and represents a range of ethnicities to better capture the representation of junctions in the population. Complete data for the assembled genome of the baseline set—including called variants and reads and mappings—can be downloaded from:

www.completegenomics.com/sequence-data/download-data/

Figure 1 lists the genomes used in the SV baseline genome set and their ethnicities.

Figure 1: Coriell Identifiers and Ethnicity Information for the SV Baseline Genome Set

YRI: Yoruba in Ibadan, Nigeria

NA18501 NA18502 NA18504 NA18505 NA18508 NA18517 NA19129

ASW: African ancestry in Southwest USA

NA19700 NA19701 NA19703 NA19704 NA19834

CEU: Utah residents with Northern and Western European ancestry from the CEPH collection

NA06985 NA06994 NA07357 NA10851 NA12004

CEPH/Utah Pedigree 1463

NA12889, NA12890 NA12891 NA12892

CHB: Han Chinese in Beijing, China

NA18537 NA18555 NA18558 NA18526

JPT: Japanese in Tokyo, Japan

NA18940 NA18942 NA18947 NA18956

LWK: Luhya in Webuye, Kenya

NA19020 NA19025 NA19026 NA19017

MXL: Mexican ancestry in Los Angeles, CA

NA19648 NA19649 NA19669 NA19670 NA19735

TSI: Toscans in Italy

NA20502 NA20509 NA20510 NA20511

GIH: Gujarati India in Houston, Texas

NA20845 NA20846 NA20847 NA20850

MKK: Maasai in Kinyawa, Kenya

NA21732 NA21733 NA21737 NA21767

PUR: Puerto Rican in Puerto Rico

HG00731 HG00732

Junction Merging Procedure

Junctions from ***allJunctionsBeta*** files of all baseline genomes were used to build location clusters. The latter were created from left and right sides of the junctions based on the following compatibility criteria:

1. Two junction sides (either from the same or from different junctions) are considered compatible if they belong to the same chromosome and same strand, and the shortest distance between these sides in reference coordinates does not exceed 500 bases.
2. Two different clusters are merged together if their locations come from the same chromosomes and strands, and if the shortest distance between the clusters in reference coordinates does not exceed 500 bases.
3. The clustering procedure is repeated iteratively until there remains no junction sides or clusters that can be merged together.
4. From the final set of clusters, we select cluster pairs that are compatible with any of the original junctions from the baseline genome set. A cluster pair is said to be compatible with a junction if one side of the junction is contained within the first cluster, and the other side within the second cluster. Some cluster pairs may correspond to more than one junction. For each cluster the Frequency in Baseline parameter is calculated as (number of baseline genomes containing corresponding junctions) / (total number of baseline genomes).
5. The resulting cluster pairs are saved as the baseline-junctions file.

The underlying data for the merged junctions across the baseline genome set are captured in two files, ***B36baseline-junctions.tsv*** and ***B37baseline-junctions.tsv***. The files are distinguished by the human genome reference build used, either NCBI Build 36 or 37.

Example *B36baseline-junctions.tsv*

>LeftChr	LeftBegin	LeftEnd	LeftStrand	RightChr	RightBegin	RightEnd	RightStrand	FrequencyInBaselineGenomeSet
chr1	816736	817118	+	chr1	2.48E+08	2.48E+08	+	0.173077
chr1	817493	817881	-	chrY	9974480	9974516	-	0.037037
chr1	818295	818333	+	chr1	5725867	5726689	+	0.019231
chr1	822672	823160	-	chr9	1.41E+08	1.41E+08	-	0.346154
chr1	824619	826317	+	chr1	5725867	5726689	+	0.442308
chr1	824619	826317	+	chr1	5727346	5729724	+	1
chr1	824619	826317	+	chr1	1.43E+08	1.43E+08	-	0.019231
chr1	824619	826317	+	chr6	58642363	58642433	-	0.038462

Header Description *B36baseline-junctions.tsv*

Key	Description	Allowed Values
#SOFTWARE_VERSION	CGI pipeline build number.	Two or more digits separated by periods. For example "2.0.0.1".
#GENERATED_BY	Assembly pipeline component that generated the output.	Alpha-numeric string.
#GENERATED_AT	Date and time of the assembly.	Year-Month-Day Time. For example "2010-Sep-08 20:27:52.457773".
#FORMAT_VERSION	Version number of the file format.	Two or more digits separated by periods. For example, "0.6".
#GENOME_REFERENCE	Human genome build used for assembly.	"NCBI build XX" where X's are digits.
#TYPE	Indicates the type of data contained in the file.	"BASELINE_JUNCTIONS": information on detection junctions and their frequencies across the baseline genome set.

Content Description *B36baseline-junctions.tsv*

	Column Name	Description
1	LeftChr	Chromosome name in text: `chr1`, `chr2`,..., `chr22`, `chrX`, `chrY` of the left side of the merged junction. The mitochondrion is represented as `chrM`. The pseudoautosomal regions within the sex chromosomes X and Y are reported at their coordinates on chromosome X.
2	LeftBegin	Zero-based position of the left-most mate read in the cluster, on the left side of the merged junction.
3	LeftEnd	Zero-based position of the right-most mate read in the cluster, on the left side of the merged junction.
4	LeftStrand	Strand of the left side of the merged junction ("+" or "−").
5	RightChr	Chromosome name in text: `chr1`, `chr2`,..., `chr22`, `chrX`, `chrY` of the right side of the merged junction. The mitochondrion is represented as `chrM`, though this may be absent from SV analyses. The pseudoautosomal regions within the sex chromosomes X and Y are reported at their coordinates on chromosome X.
6	RightBegin	Zero-based position of the left-most mate read in the cluster, on the right side of the merged junction.
7	RightEnd	Zero-based position of the right-most mate read in the cluster, on the right side of the merged junction.
8	RightStrand	Strand of the right side of the merged junction ("+" or "−").
9	FrequencyInBaselineGenomeSet	Frequency with which the junction is detected in set of baseline genomes.