

The Cancer Genome Atlas Pilot Project
Stanford University School of Medicine
Raw Data Description

Platform: Illumina Infinium 550K SNP Array

Raw Data

Among the raw data files that we will provide are the IDAT files, which are the binary data files produced by the Illumina scanner, one for each color channel of each sample, that contain the average intensity data for each SNP averaged over >20 beads. These files can be read by the Illumina BeadStudio analysis software to produce all the other data files.

We will provide all the genotype calls for each sample, as well as a cluster file that defines the genotype cluster positions for each SNP that we generated from the TCGA samples. We will also export from the Illumina Beadstudio software the raw intensity values and genotyping quality scores.

The Cancer Genome Atlas Pilot Project
Stanford University School of Medicine
Normalized Data Description

Platform: Illumina Infinium 550K SNP Array

Normalized Data

The Illumina Beadstudio software will be used to generate normalized intensity values for each allele of every SNP, the logR (log of total intensity, summed over both alleles) and B allele frequency values for each SNP, as well the differences in logR and B allele frequency between each pair of tumor and normal samples. Such pairwise differences will be the basis of inferring copy number changes.

We will also provide additional normalized logR and B allele frequency data files that result from our custom normalization procedures. We developed these procedures to correct for additional sources of noise or bias, such as sample-specific, bead pool-specific, and SNP-specific effects in the intensity data that have not been adequately removed by Illumina's genotyping software.

The Cancer Genome Atlas Pilot Project
Stanford University School of Medicine
Data Descriptions

Platform: Illumina Infinium 550K SNP Array

Segmented Data

We will provide summary tables of CNVs discovered for each sample based on our segmentation software. Each line of summary will contain the sample ID, the beginning and ending SNPs of the CNV, their positions, gene(s) affected, and the estimated nature of variation (deletion, loss of heterozygosity, or amplification, etc).