Supplemental information

Supplemental Methods - Data Analysis

Nascent strands from DNA Replication sites were sequenced on the Illumina Genome Analyzer II. Sheared genomic DNA was sequenced as a control. Reads from the samples used in this study are summarized in Supplemental Table 1.

Sequence Alignment

FASTQ formatted sequence files were aligned to the human genome reference sequence build 37.1 (hg19) using Bowtie (Langmead et al. 2009) version 0.11.3. For single read data, Bowtie was run with the following parameters:

bowtie -q --solexa1.3-quals -p 3 -n 2 -l 40 -m 1 -S --sam-nohead hg19 <FASTQ File> <SAM File>

For paired-end reads, bowtie was run with the following parameters:

bowtie -q --solexa1.3-quals -p 3 -n 2 -l 33 -I 76 -X 136 --fr -m 1 -S --sam-nohead hg19 -1 <FASTQ 1> -2 <FASTQ 2> <SAM File>

These settings allowed for up to two mismatches and returned only unique alignments. Alignments matching more than one genomic location were discarded. For paired-end reads, the alignments required alternating direction and an insert size within the range reported by the sequencing facility (76–136). The output was in SAM file format.

Sorted, indexed binary alignment files (BAM files) were then created using SAMtools (Li et al. 2009) version 0.1.7. BAM files were used both for visualization of the alignments, and for subsequent statistical analysis of replication initiation. The command line options used for the conversion, merging, indexing and sorting were:

samtools import hg19.ref_list <SAM File> <Unsorted BAM File>

samtools sort -m 4000000000 <Unsorted BAM> <Lane N BAM>

samtools merge <Lane 1 BAM> <Lane 2 BAM> . . . <Lane N BAM> <Sample BAM>

samtools index <Sample BAM>

Finally, a coverage track was produced for each BAM file for high-level visualization using igvtools version 1.4.1. (http://www.broadinstitute.org/igv/.)

igvtools count -z 4 -w 10 <BAM File> <TDF File> hg19

Replication Quantification – Enrichment Ratio

The depth of aligned reads from nascent strand sequences was used in this study to determine the level of replication activity at a given genomic location. The nascent strands isolation protocol that was used for this study, targeted strands that were approximately 400 bases long. Non-directional reads from nascent strand sequences were obtained using random primers. As reads were randomly generated within the 400 base targets, we chose to bin read counts in 100 base bins, and smooth the counts across seven bins using a Gaussian algorithm (kernel size = 7 bins, variance = 1.75). Samples were binned across the entire genome in 100 base increments, as reads per kilobase per million aligned reads (RPKM). RPKM values support normalized comparison between samples with a different number of total reads. See Supplemental Figure 4 for a view of initial reads and smoothed/binned RPKM values.

Potential biases in read coverage had to be corrected prior to analysis of coverage depth and replication activity. Next generation sequencing technologies do indeed have coverage biases that are introduced by sample preparation, sequence composition, and other factors (Dohm et al. 2008; Auerbach et al. 2009; Harismendy et al. 2009). These biases were very consistent across runs, even with different tissues of origin. We observed several regions in our data that had extremely amplified coverage in multiple cell lines. This was seen both in controls and nascent strand samples. To correct for sequencing bias, we calculated the ratio of nascent RPKM to control RPKM. See Supplemental Figure 5 for an example of a region with anomalous reads (thousands of reads in the nascent strand sample, and hundreds of reads in the control) that was corrected by taking the ratio of nascent to control reads.

When calculating the nascent strand to control ratio, extremely small control values would lead to very high ratios, even when the nascent strand RPKM was not very large. To prevent this bias, we established a minimum control value, and raised very small control values to this minimum value.

The minimum value was set to the smoothed RPKM value of a single read in a bin. In addition, negative replication initiation enrichment was not a meaningful concept in this analysis, so bins in which the ratio was < 1, were set to 1. An enrichment ratio of 1 indicated no enrichment. All subsequent analyses of replication initiation was performed using this enrichment ratio.

Another sequencing bias that is prevalent in cancer samples is copy number variation. Compared to wild-type chromosomes, regions that have been amplified in a cancer cell line produce more reads when aligned to the reference genome. Calculation of an enrichment ratio also corrected for copy number variation, because copy number variations were also reflected in the genomic control reads. See Supplemental Figure 6 for data from an amplified region of the MCF7 genome. Amplification was reflected both in nascent strand and control reads, allowing the enrichment ratio to correct for this anomaly.

The final data processing step was the production of a publicly available .bed file of high confidence replication initiation regions for K562 and MCF7. These tracks were assembled by selecting peaks in the enrichment ratio values that exceeded a significance threshold. This enrichment level was selected by a Metropolis Monte Carlo simulation that identified an appropriate threshold for a target false discovery rate (FDR). We calculated empirical FDR values by swapping the nascent and control samples to calculate the ratio of the number of false peaks that met the threshold in the control sample over the number of true peaks in the nascent sample: FDR = the number of control peaks/the number of nascent strand peaks. This approach for finding empirical FDR values has been used by many ChIP peak-selection programs (Pepke et al. 2009). After thresholds for 30% and 10% FDR had been established, we selected high-confidence peak regions using these thresholds. Adjacent peaks with a gap of two or less bins were merged. See the dark blue tracks at the bottom of Supplemental Figure 4 for examples of the 10% and 30% FDR peaks.

Reproducibility of Results

To test the reproducibility of the methods used in this study, we performed two different runs of MCF7 and K562 nascent strand sequencing. MCF7 runs were performed at different sequencing facilities. All samples were sequenced using Agilent Genome Analyzer II platforms. Reproducibility of results was measure by calculating Pearson and Spearman correlation coefficients for the binned enrichment ratios for the full genome.

Replication and Genes – Transcript Start Site

To investigate replication initiation activity in genomic regions containing protein-coding genes, we first use the sum of replication activity (binned nascent/control enrichment ratio) around the transcription start site (TSS) of all genes, covering -2000 to +2000 bases of the TSS. The enrichment ratio of each 100 base bin from -2000 to +2000 of the TSS represents the average replication activity across all genes at the specified position relative to the TSS. The human protein coding genes from NCBI Gene that have a known TSS, and one or more recorded transcripts were used for this study.

Since the pattern of replication levels across the TSS region was strongly influenced by gene expression levels, genes were separated into four groups by expression levels that were measured in previous studies of K562 and MCF7 cells using the Affymetrix U133 Plus 2 expression microarray. Expression data was GCRMA normalized and $\log_2$ transformed. The groups were defined as very low expression (< 2.3), low expression (between 2.3 and 5.3), medium expression (between 5.3 and 8.5), and high expression (> 8.5).

TSS regions contain a known sequencing bias that results in increased reads just upstream of gene transcript start sites. For example, this bias can be seen in our MCF7 control read (green line) in Supplemental Figure 8. The enrichment ratio that we calculated (yellow line) corrected for this observed TSS sequencing bias. The remaining replication initiation enrichment (approximately 400 bases downstream from the TSS) remains after removing the bias reflected in our control.

We used the enrichment ratio of nascent/control RPKM values to measure replication initiation activity (see the Replication Quantification section above). The hg19 gene

positions were obtained from NCBI Gene, and only genes with at least one transcript record in RefSeq or GenBank were used (20,063 genes). Expression levels were extracted from a previously developed dataset of NCI60 gene expression, measured with the Affymetrix U133 Plus 2 microarray. Expression values were GCRMA normalized and $\log_2$ transformed. Genes were matched with expression values using NCBI Gene IDs so that expression values could be assigned to 17,931 genes.

We sorted genes by expression values and binned the data into 20 bins with an equal number of genes in each. The first 3 bins were combined, as their expression values were essentially identical. An additional graph was prepared using box plots for each gene bin to show the mean, median, 25%, 75%, min, and max values.

Chromatin Feature Analysis

The next phase of our analysis was to look at chromatin features (e.g. CpG islands, histone modification sites, DNase hypersensitive regions) to explore the relationship between these features and DNA replication initiation. A large number of feature tracks of interest were available from the UCSC Genome Browser (Rhead et al. 2010). Since many features are currently only available in hg18 coordinates (Build 36 of the human genome), and we performed our read alignments to hg19 (Build 37), feature positions had to be translated from hg18 positions to hg19 positions. For each feature, sequence regions were extracted from hg18, and realigned to hg19. Only unique alignments with two or fewer mismatches were included. On average, we were able to **align** > 98% of the features from each track using this technique.

A custom program was developed to perform automated analysis of replication activity and features. The enrichment ratio of nascent to control RPKM was used as the measure of replication initiation activity. Enrichment ratios for bins contained within a given feature region were averaged. A size-matched control was also generated by averaging bin values of a randomly selected non-feature region with an equivalent size. Average replication activity for the feature was produced by averaging enrichment ratios of all feature regions. A p-value was then calculated by performing Welch's t-test, using the feature regions' enrichment ratios, and ratios of the size-matched, non-feature regions, to assess feature replication enrichment significance.

Please see Table 3 and the UCSC Genome Browser Human March 2006 (NCBI 36/hg18) and Feb 2009 (GRCh37/hg19) “regulation” tracks for descriptions of these feature files. The specific tracks used in this study are listed in Table 3.

We also examined whether sequences exhibiting combinations of two features (e.g. DNAse hypersensitivity and CTCF binding) exhibited higher frequencies of replication activity than was seen with each individual feature. Selected results of combined analyses are shown in Supplemental Figure 2A-C. For each pair of single-feature tracks, a new double-feature track was created that only contained intersecting regions of the original features (contiguous regions were treated as one feature region). Feature analyses were performed (as described above) with these double-feature tracks. The average enrichment ratio of the double-feature track was compared to both single-feature tracks

Supplemental Figure Legends

Supplemental Figure 1. (A) Examples of nascent strand, control, and enrichment ratio data. A screenshot from the IGV browser displaying data for the *CTCF* locus (the same region as in Figure 1D) is shown. A chromosome map is shown at the top, and the region-of-interest is delineated by a red rectangle. The analyzed region is shown underneath the ideogram, with map coordinates indicated. Experimental tracks show the distribution of sequence reads (aligned with the indicated region) obtained from massively parallel sequencing. All data are shown as reads per kilobase per million mapped reads (RPKM). From the top, K562 NS refers to values derived from a nascent strand preparation of K562 cells. K562 Gen. refers to values derived from a control sheared genomic DNA preparation of K562 cells. K562 Ratio refers to the ratio of values of nascent strands and values of control genomic DNA. MCF7 NS, MCF7 Gen. and MCF7 Ratio refers to similar values obtained from MCF7 cells. Ref-Seq genes are aligned under the nascent strand distribution. (B) Replication enrichment ratios in MCF7 cells (left) and K562 cells (right) plotted against distance from the TSS for all genes.

Supplemental Figure 2. Chromatin modifications and replication initiation events. The average replication enrichment ratio for genomic regions that contain the indicated chromatin modification features, and for genomic regions that exhibit two types of modifications.

Supplemental Figure 3. Box plots representing the distribution of data shown in Figure 6. Boxes indicate distributions of the second and third quartiles; dots indicate average values; error bars indicate the fifth and 95th percentiles. Values for gene expression (blue, left), and replication enrichment ratios (red, right) are shown.

Supplemental Figure 4. MCF7 read counts for nascent strands and genomic control reads (blue). The same data after binning (100 base bins) and smoothing (teal). The enrichment ratio of nascent to control (purple), and regions of high confidence (30% and 10% FDR) (dark blue).

Supplemental Figure 5. MCF7 samples with aberrant reads in a region of chromosome 1. A high number of reads was present in both nascent strand and control samples, so calculation of an enrichment ratio corrected this aberration.

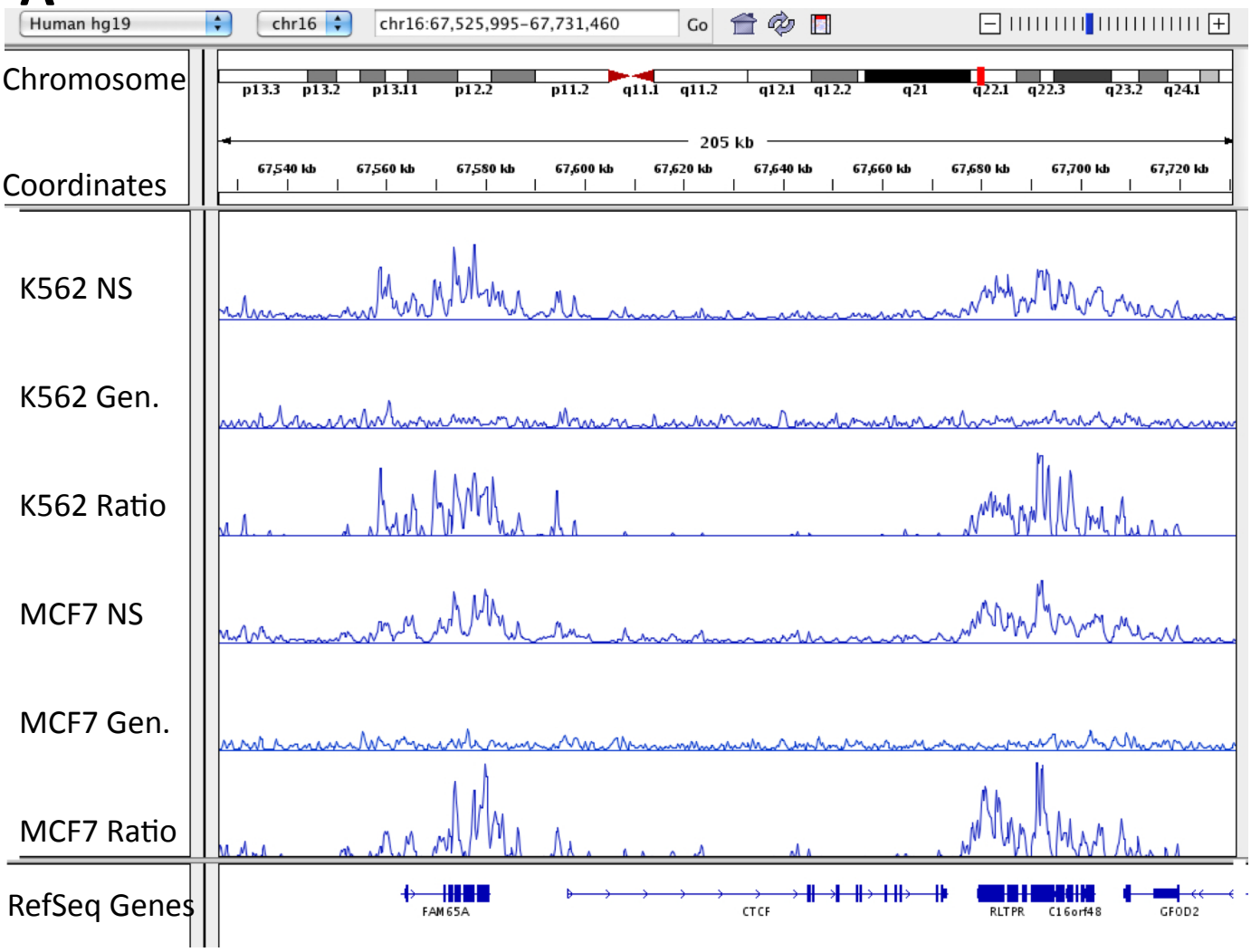
Supplemental Figure 6. An MCF7 genomic region with known amplification (identified previously using an Agilent copy number microarray). Amplification was reflected in control reads, and normalized by the enrichment ratio.

Supplemental Figure 7. Chromosome 1 Enrichment Ratios for MCF7 and a replicate run of MCF7. Also shown are K562 enrichmet ratios (green).

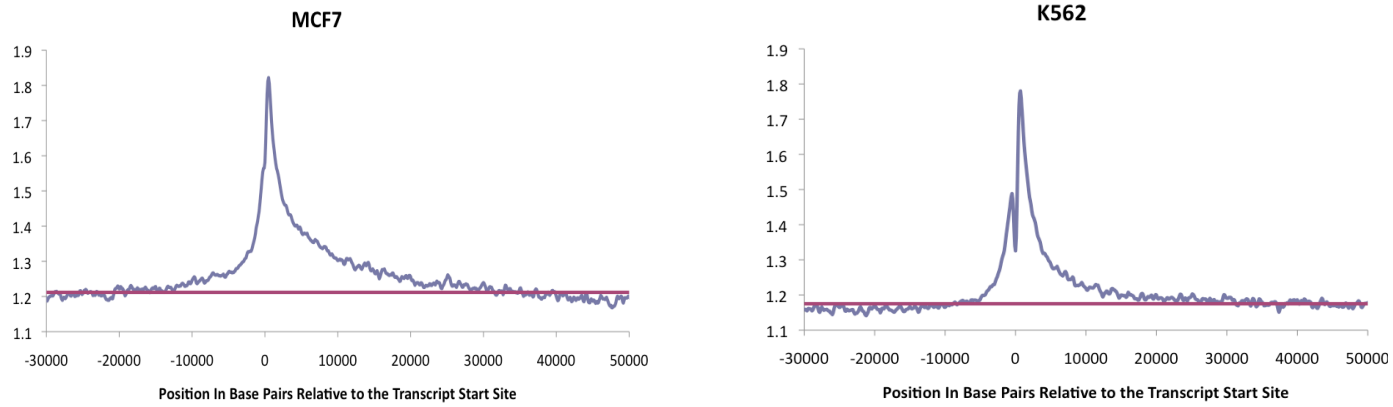
Supplemental Figure 8. Control genomic reads demonstrate a bias at the TSS (green). The TSS bias present in the control is removed when an enrichment ratio (yellow) is calculated.

Supplemental Figure 1

A

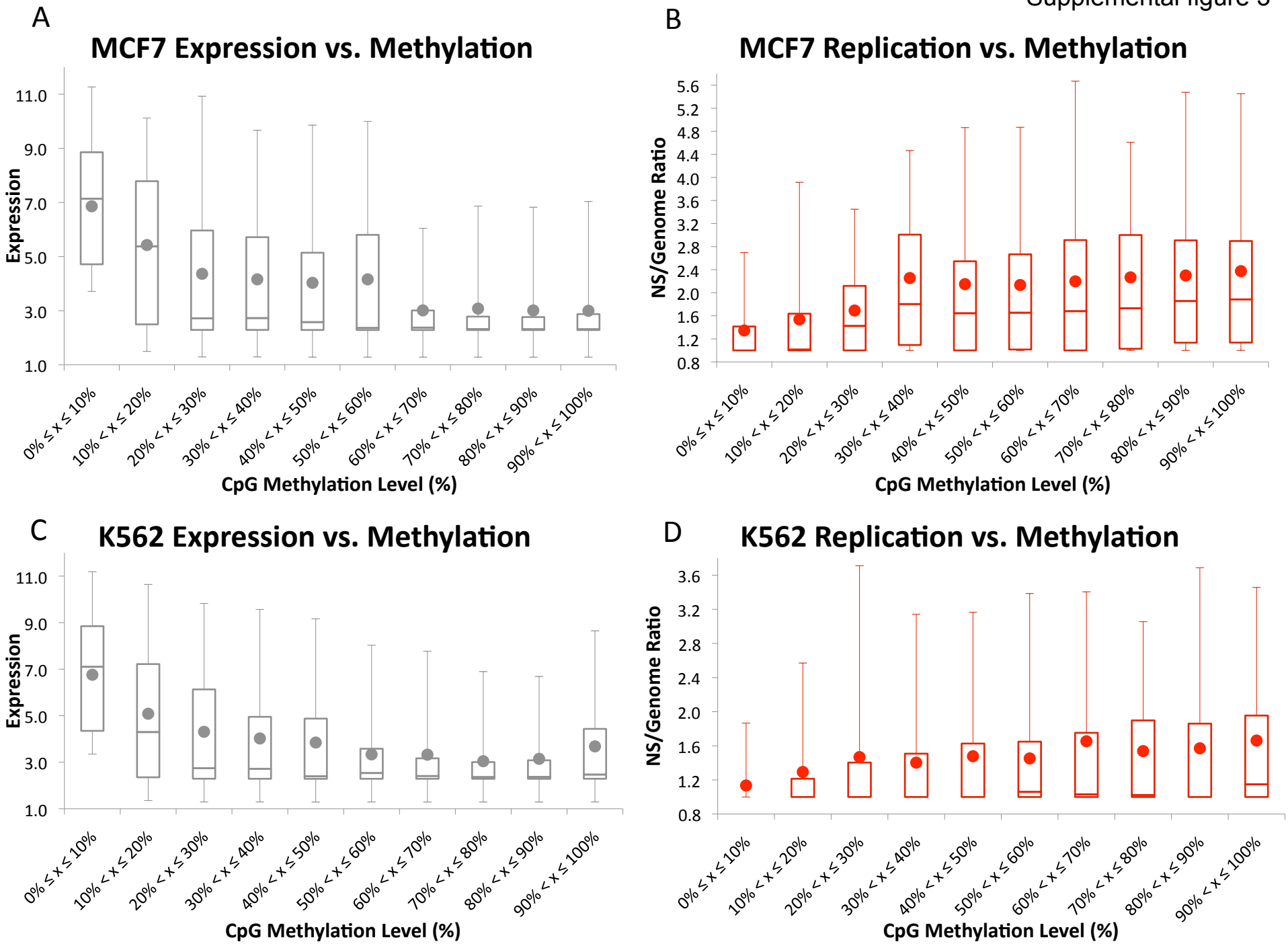


B

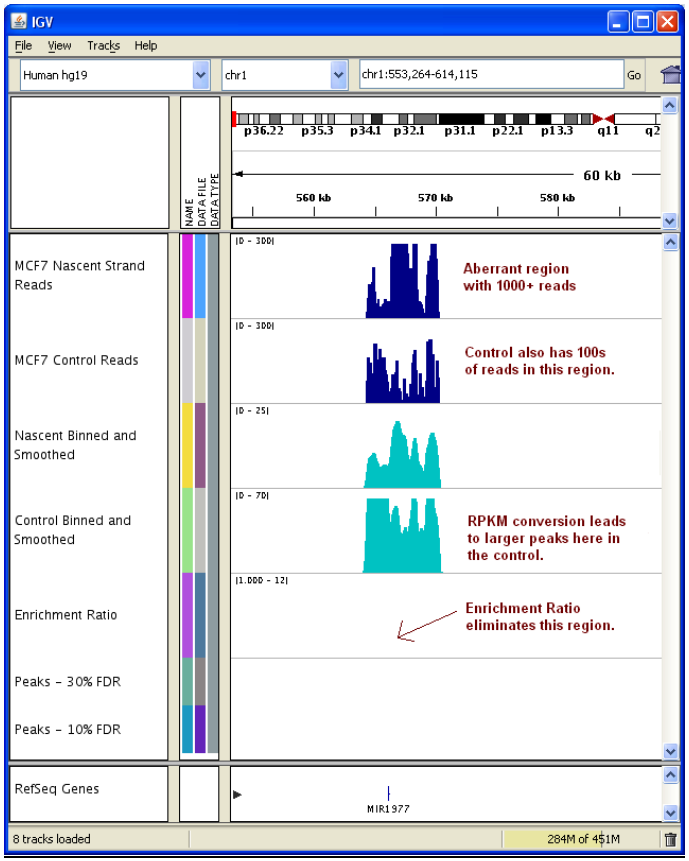


Supplemental figure 2

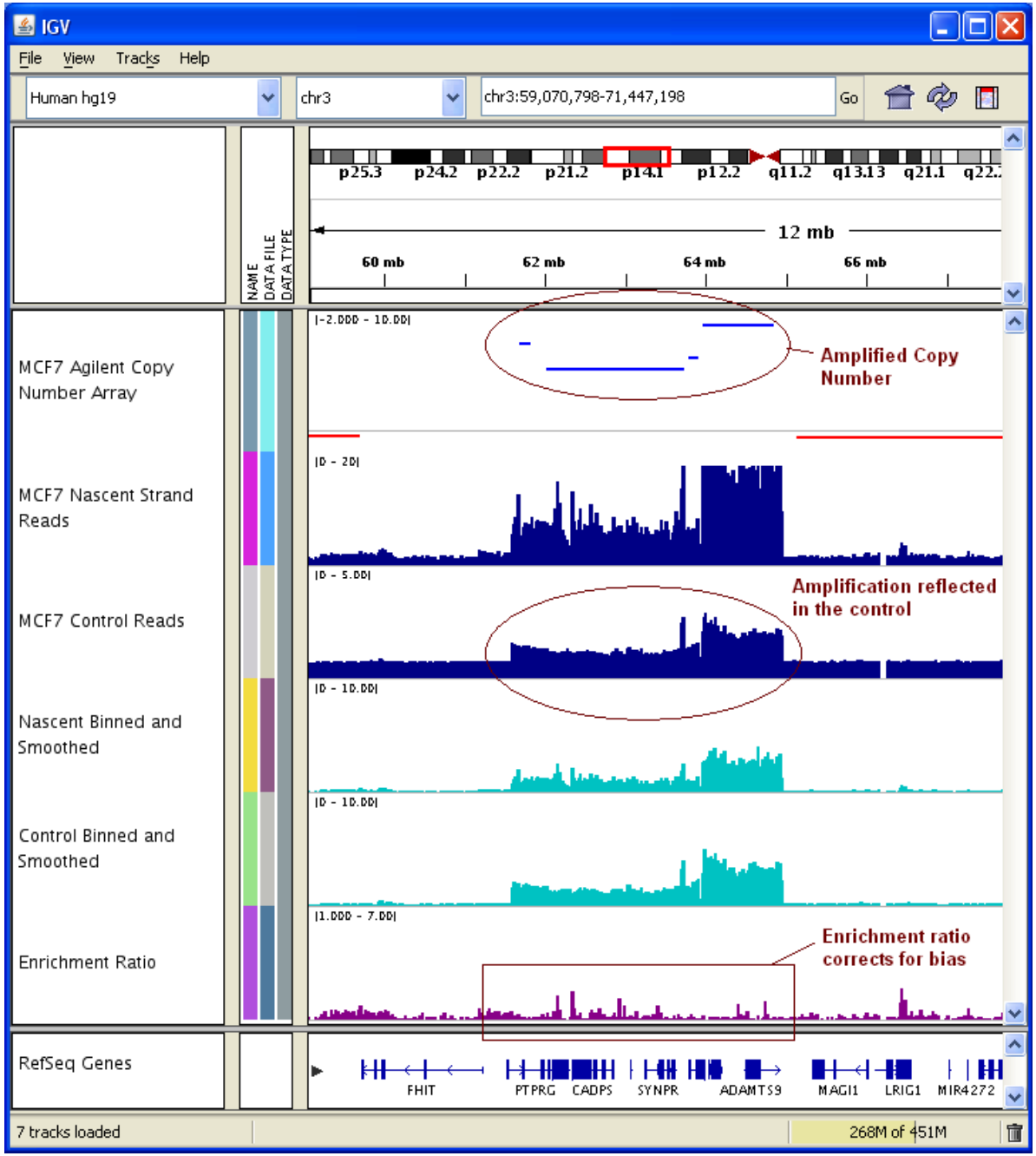
Supplemental figure 3



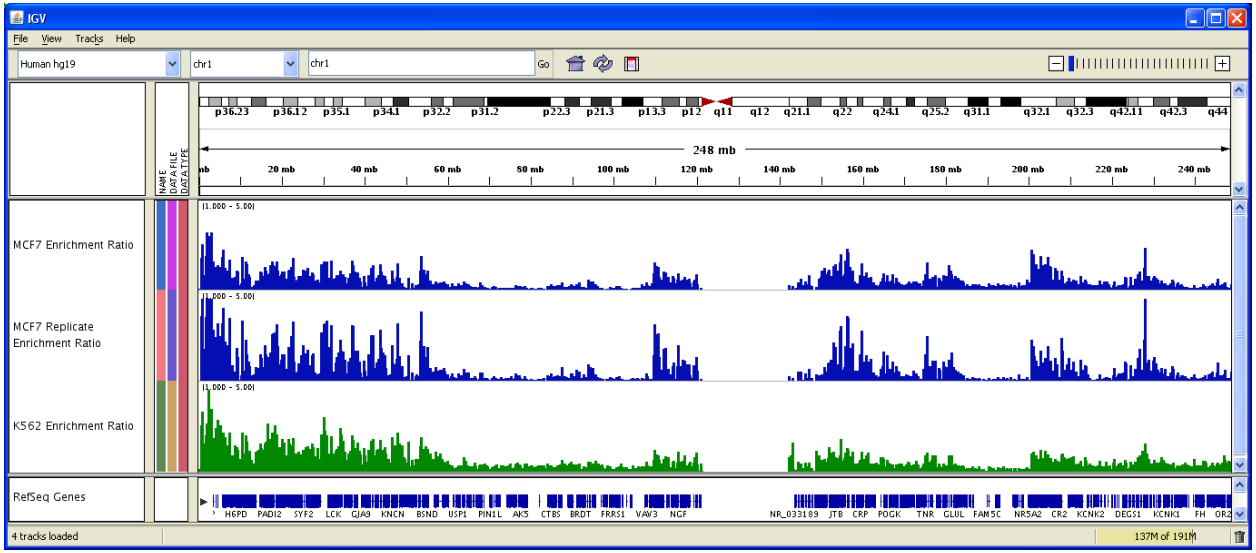
Supplemental figure 5



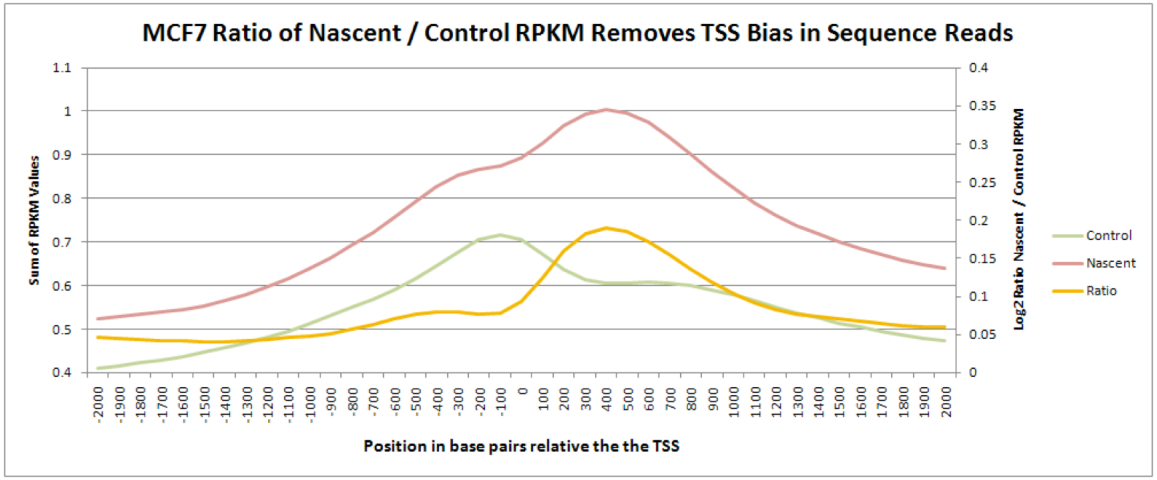
Supplemental figure 6



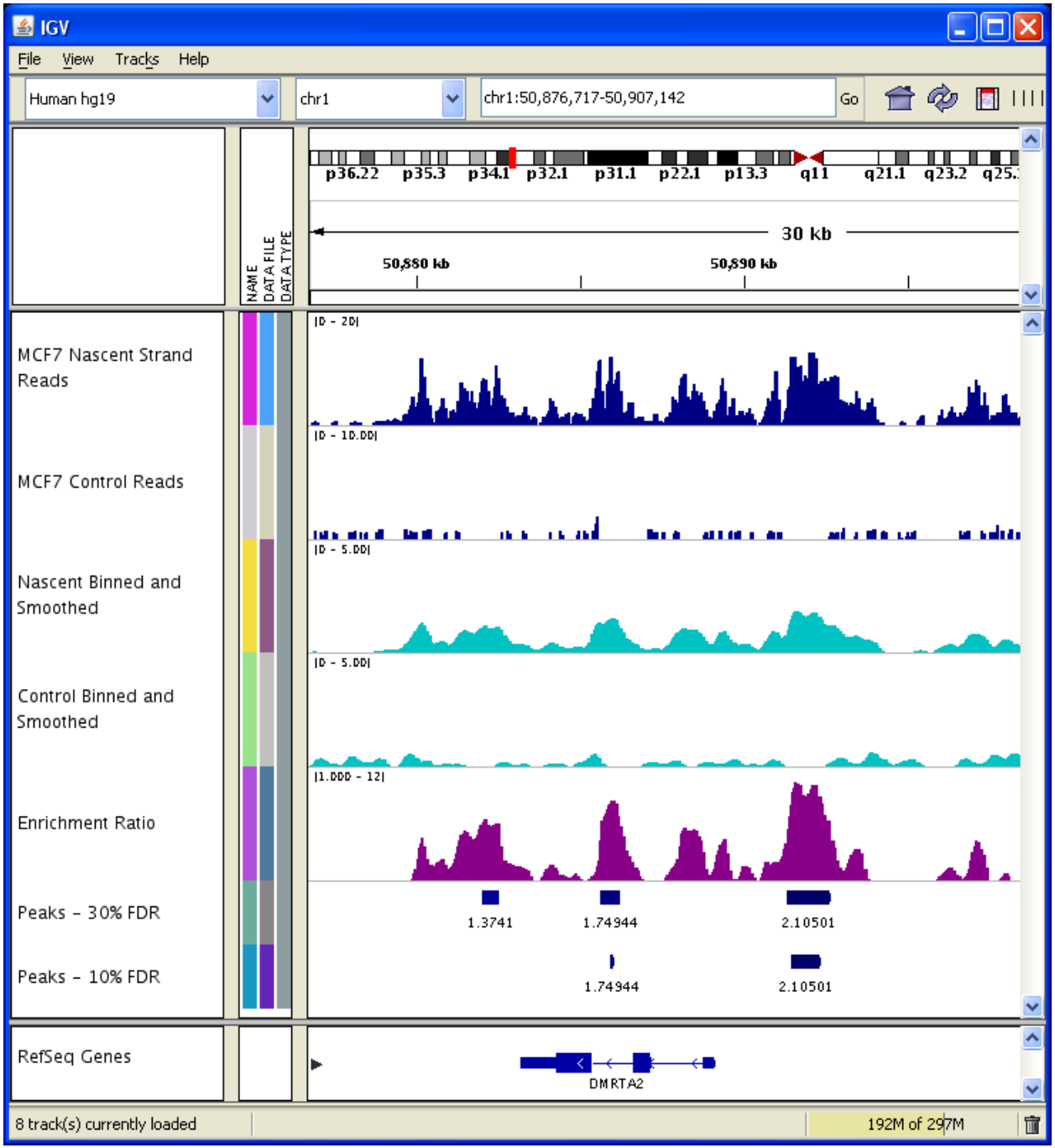
Supplemental figure 7



Supplemental figure 8



Supplemental figure 4



Supplemental Table 1: Summary of Sample Sequencing

Sample	# Lanes	Paired?	Read Length	# Reads	Aligned Reads
MCF7	7	Yes	33	226002540	162654644
MCF7 Control	2	No	30	31417391	16750928
K562	4	Yes	35	241711712	147515012
K562 Control	2	No	35	66408095	43597078
MCF7 Replicate	5	No	varied 30 to 72	87425510	44653787
K562 Replicate	4	No	40	84546209	31351386