**MCF-7 as a model for functional analysis of breast cancer risk variants**

Alix Booms1, Gerhard A Coetzee\*1, Steven E Pierce\*1

1Center for Neurodegenerative Science, Van Andel Research Institute, Grand Rapids, MI, United States

\*These authors co-directed this work.

**Running title:** Functional analysis of breast cancer risk variants in MCF-7

**Corresponding authors:** email addresses: [gerry.coetzee@vai.org](mailto:gerry.coetzee@vai.org), [steven.pierce@vai.org](mailto:steven.pierce@vai.org)

Van Andel Research Institute

333 Bostwick Ave N.E.

Grand Rapids, MI 49503

Voice: 616-234-5305

**Legends to Supplemental Data**

**Supplemental Table 1.** H3K27ac peaks passing IDR filter score of 850. Annotated peaks for Coetzee, annotated peaks for union of Coetzee with ENCODE1 and ENCODE2, annotated peaks for intersection of Coetzee with ENCODE2, annotated peaks for intersection of Coetzee with ENCODE1 and ENCODE2, annotated peaks for HMEC, output from IDR (>850) for Coetzee, peaks for ENCODE1, peaks for ENCODE2. CTCF IDR filtered peaks for all four datasets.

**Supplemental Table 2.** BCa GWAS SNPs from Michailidou, et. al. CRVS, Manhattan plot set, LD SNPS, LD SNPS in bed format, most significant CRVs, CRV risk spans, and enlarged regions around CRVs used for enrichment calculations. Overlap of risk SNPs with Enhancers in MCF-7 and HMEC

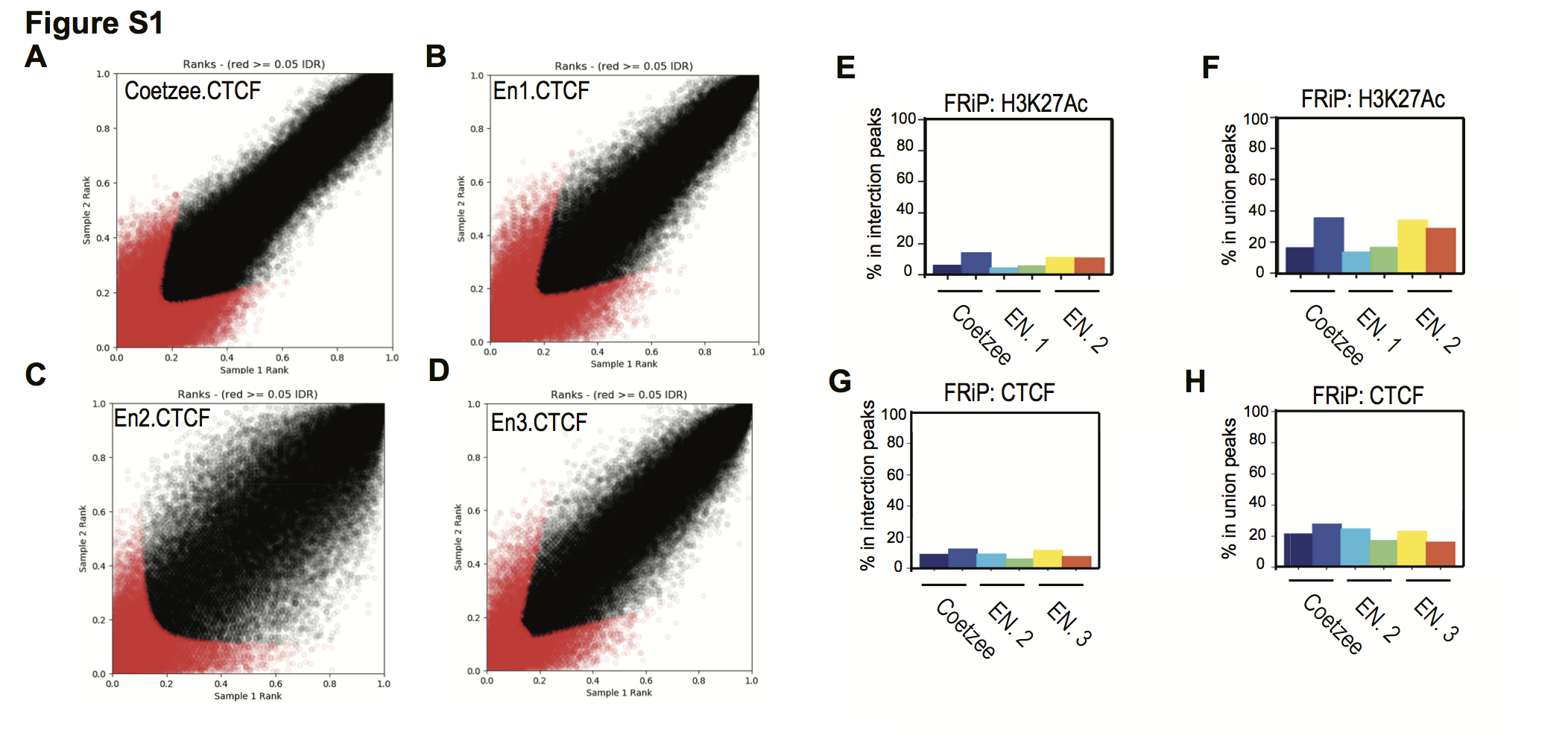
**Supplemental Table 3.** Raw data table 1, raw data used for figure 3, and data in figure 3 ordered as in figure.

**Supplemental Table 4.** CRV overlapping Coetzee CTCF peaks in MCF-7, CRV overlapping CTCF coverage present in all 4 datasets, motifbreakR analysis showing CRVs which overlap MCF-7 CTCF peaks and that disrupt CTCF binding.

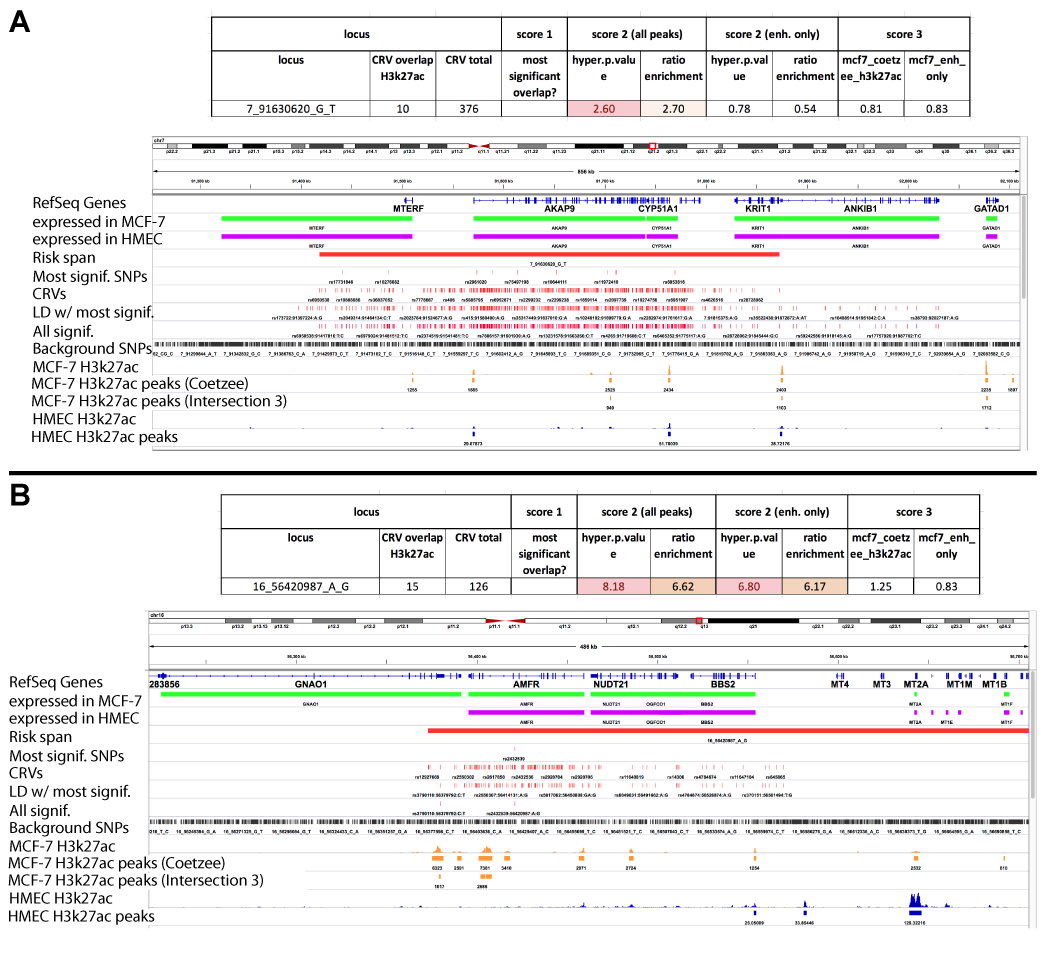
**Supplemental Table 5.** RNA-seq results for MCF-7 (comparing 2 batches generated in the present study), RNA-seq results comparing combined for MCF-7 vs. combined HMEC, bed file of genes expressed in MCF-7, and bed file of DE genes in MCF-7 relative to HMEC, GO enrichment of DE genes upregulated in MCF7, GO enrichment of DE genes upregulated in HMEC.

**Supplemental Table 6.** Set of genes within 500 KB of MCF-7 risk enhancer peak (Coetzee) and expressed by > 1CPM, GO enrichment of MCF-7 risk genes, set of genes within 500 KB of HMEC risk enhancer peak and expressed by > 1CPM, GO enrichment of HMEC risk genes.

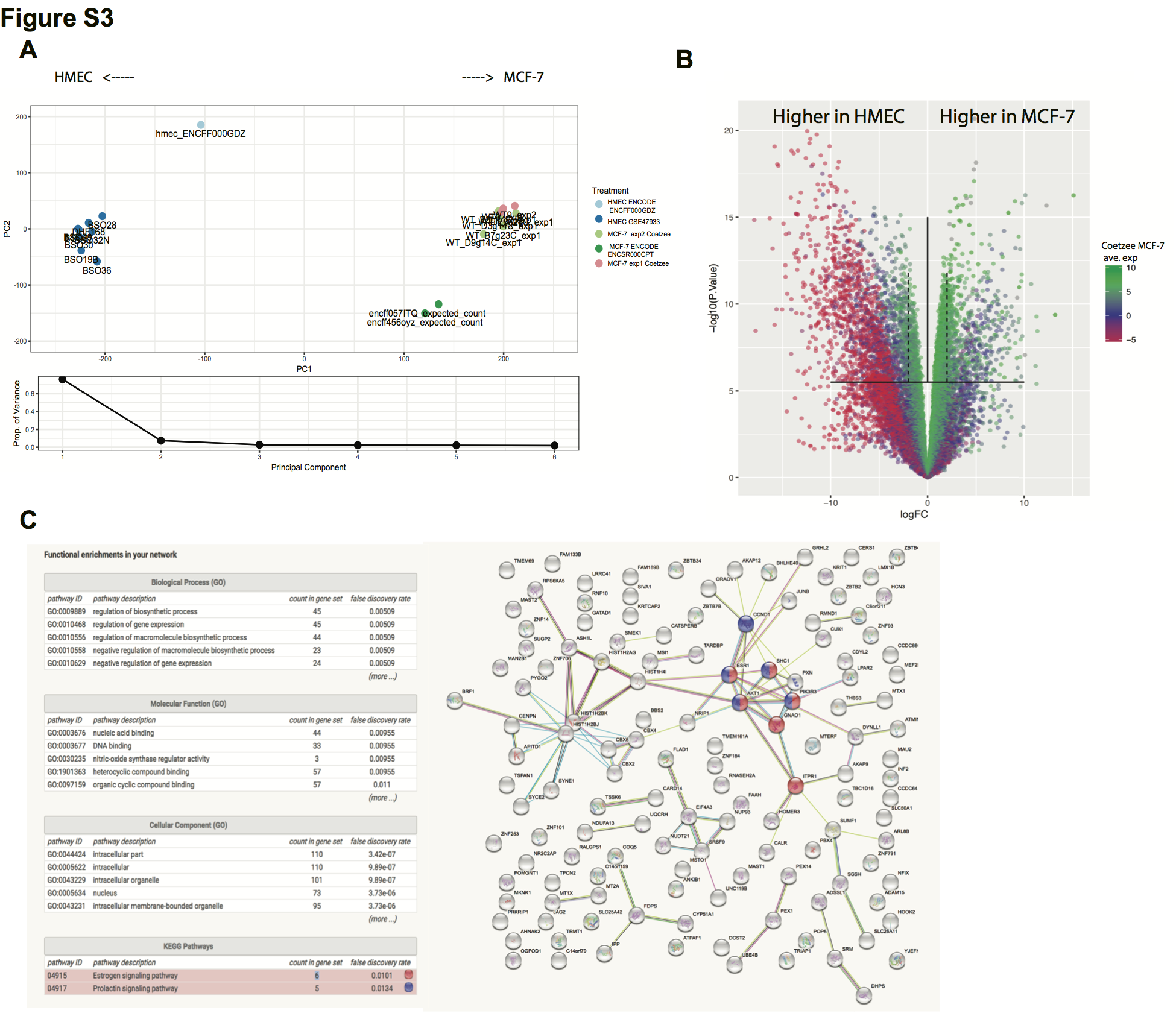
**Supplemental Table 7.** 69 genes with eQTL for breast tissue with SNPS that reside in risk H3k27ac peaks, 48 egenes which are also expressed in MCF-7. List of differentially expressed genes near (<500kb) MCF-7-only risk enhancers, list of all expressed MCF-7 genes near MCF-7- only risk enhancers, GO enrichment for DE genes near MCF-7-only risk enhancers.



**Supplemental Figure 1.** Characterization of ChIP-Seq data. A) IDR plot showing the correspondence between replicates 1 and 2 and the threshold for filtering (in black) for our CTCF ChIP-Seq (Coetzee) B) for Encode dataset CTCF 2, and D) for ENCODE CTCF dataset 3. E) Frequency of read in peak (FRiP) plot showing overlap of H3K27Ac alignment data with the intersection or F) union of peaks. G) FRiP plot of overlap of CTCF alignment with intersection or H) union of peaks.



**Supplemental Figure 2.** Genomic views of 2 loci where enrichment (score 2) indicates a good risk target but risk span analysis (score 3) suggests ambiguity of identification. In both cases a widely spaced set of risk SNPs overlaps multiple enhancers. Score 2 (enrichment of sample/background) has been calculated for both all H3K27Ac peaks, and the subset of H3K27Ac peaks defined as enhancers.

­­

**Supplemental Figure 3.** RNA-seq comparing gene expression of MCF-7 and HMEC. A) PCA plot, B) volcano plot of gene expression changes. Dotted lines at two-fold change and horizontal solid line at significance threshold. C) Network diagram of differentially expressed genes expressed in MCF-7 and with 500 Kb of MCF-7-specific (not active in HMEC) enhancers.