Supplementary Tables

Supplementary Table S1: Comparison of Goldmine with two existing tools for annotating genomic ranges.

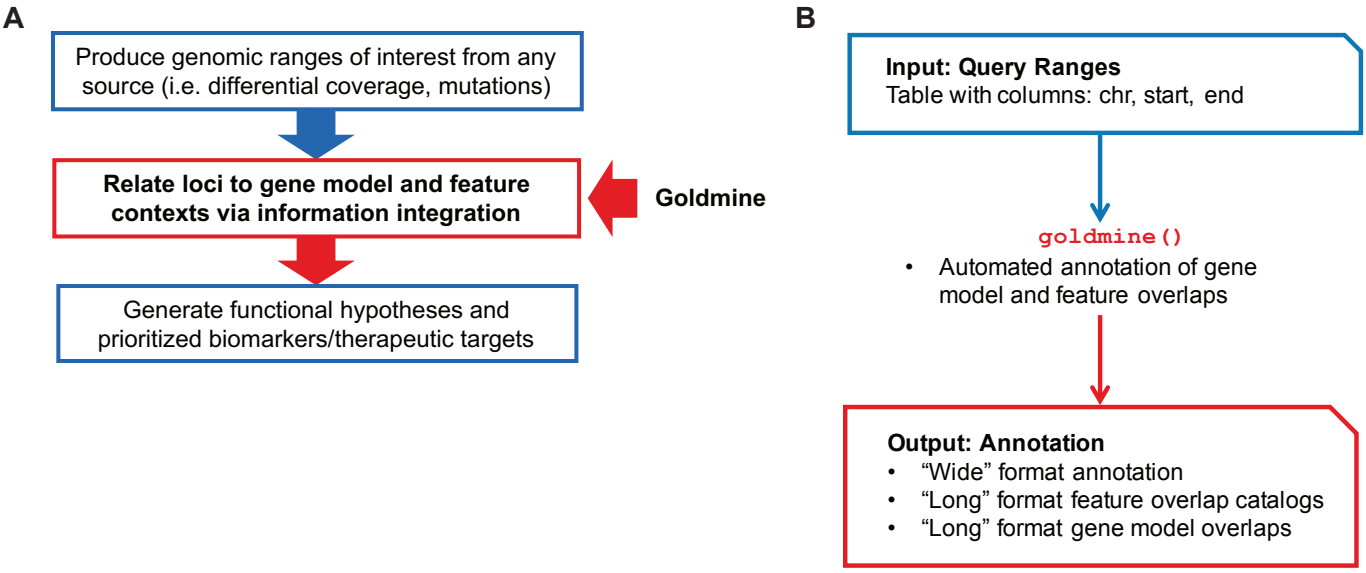
	Goldmine	HOMER*	ChIPpeakAnno
Provides simple gene model annotation?	Yes	Yes	Yes
Provides splice-isoform level detail?	Yes	No	No
Annotates intergenic ranges with functional elements?	Yes	No	No
Synchronizes and utilizes multiple gene databases (UCSC, ENSEMBL, RefSeq) and feature databases (CpG islands, GWAS catalog, ENCODE supertracks, and all other UCSC feature tables)?	Yes	No	No
Implementation	R package	Perl program	R package

*annotatePeaks.pl

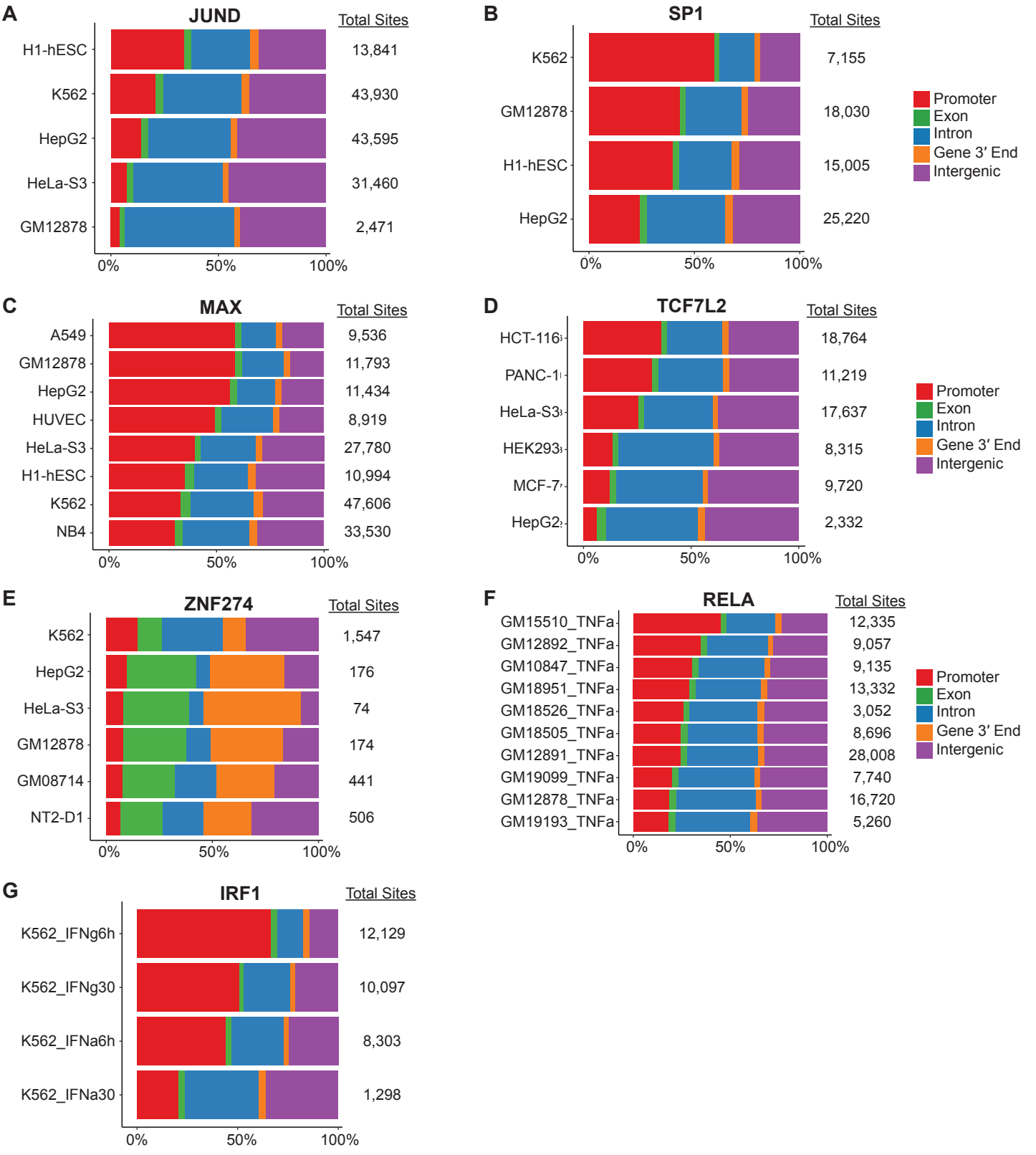
Supplementary Table S2: Fraction of binding sites in gene model contexts for all ENCODE ChIP-seq factors (XLSX File). Rows are in the same order as figure 1b and contain the factor names and proportion of all ENCODE supertrack binding sites that are assigned by Goldmine to each context. The gene set and gene model contexts can be user-adjusted when running Goldmine, and this table was produced using the UCSC gene set and default promoter (-1000 bp to +500 bp) and 3' end (-1000 bp to + 1000 bp) settings. The “majority” column gives the context call with the highest proportion. Rows are ordered descending by promoter percentage.

Supplementary Table S3 (XLSX File): Table of CD4+ and CD8+ DMRs after annotation using Goldmine. On the “DMR List” sheet, each row contains data corresponding to an individual DMR between CD4+ and CD8+ T-cells. The genomic range coordinates are provided for each DMR (columns A through C), in addition to p-value, pattern, and fold change (columns E though G). Next, the annotation columns provided by Goldmine indicating overlap with gene model and feature contexts are provided (columns I through AA). A complete description of all columns is provided in the “Column Descriptions” sheet of the Microsoft Excel file.

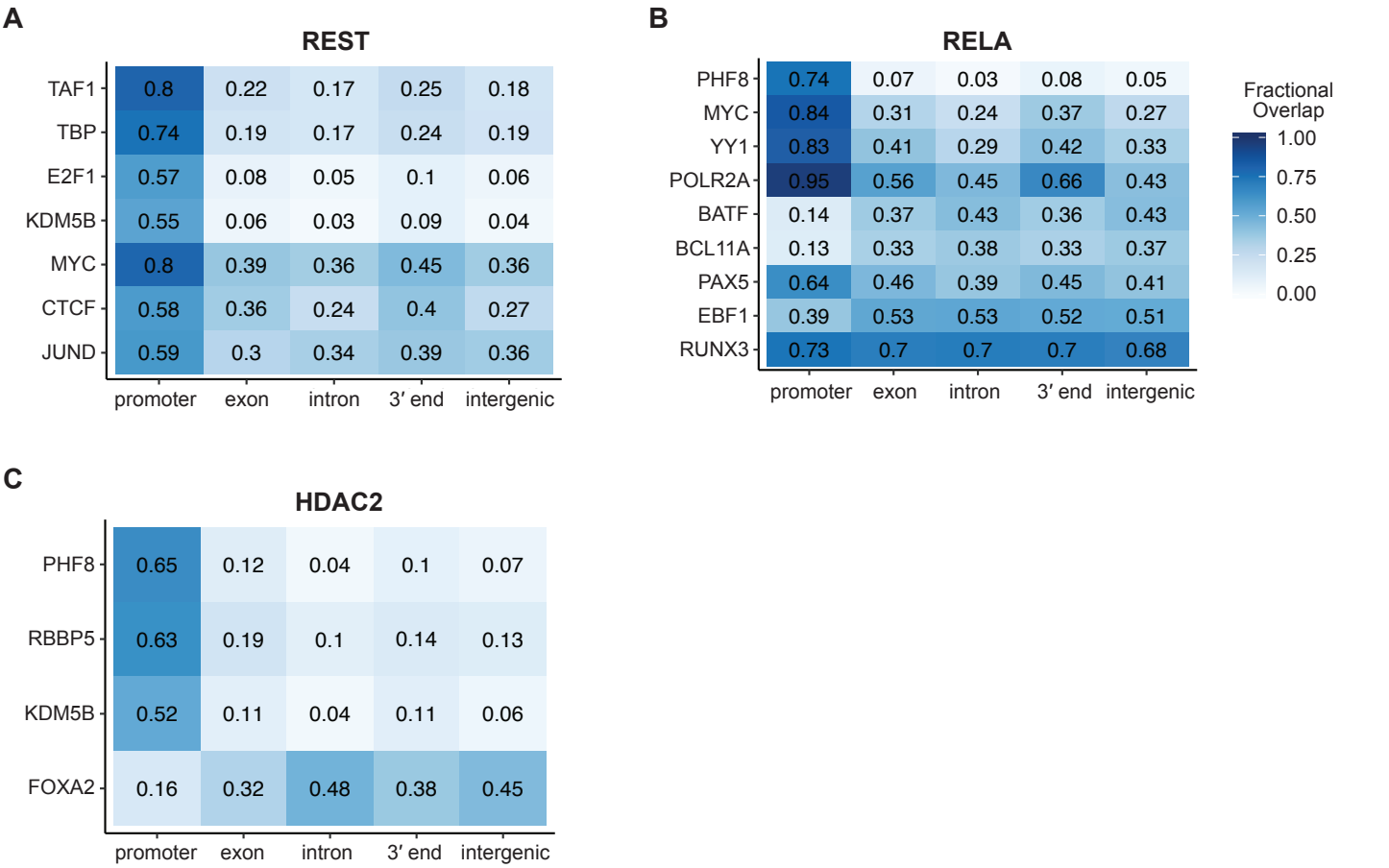
Supplementary Figure S1



Supplementary Figure S1: Use of Goldmine in bioinformatics workflows. (A) Goldmine addresses the need for annotation to bridge the gap between un-interpretable genomic range sets and novel functional hypotheses about the biology of these range sets. (B) Input and output of Goldmine's core function. The "wide" format annotations retain each query genomic range as a row, and add columns describing the spatial relationship of the individual range with each gene model and feature context sets of interest. The "long" format have one row for each pair of query range to subject row overlaps. There is one table produced for gene model contexts detailing promoter, specific intron/exons by number, and 3' end overlaps. There is also one table produced for each feature set, and the columns associated with each feature are joined to the columns of the row to facilitate cross-comparison. For example, the "wide" format could enable quick summarization of which query ranges overlap with GWAS SNPs (when annotating using the "gwasCatalog" table), and the "long" format would identify the specific phenotypes and source publications annotated with each individual overlapping SNP.



Supplementary Figure 2: Gene model contexts annotated by Goldmine are dynamic across biological conditions. Each plot shows the proportion of ChIP-seq peaks across the named cell lines (A-F) or treatment conditions (G) within each Goldmine gene model context. The total number of peaks for the factor in a cell line is given in the column next to the graph.



Supplementary Figure S3: Co-binding modularity has specificity to genomic context. Each heatmap square is valued with the fraction of binding sites for the named factor that overlap with each co-binding partner given on the heatmap rows. Fractional overlaps are computed between the unions of all peaks across all available cell lines in ENCODE for each factor. Each column stratifies this relationship across the Goldmine genomic contexts. Rows were selected for display from the set of all pairwise fractions to contain examples of promoter-biased co-occurrences (for example, REST with TAF), non-promoter-biased co-occurrences (for example, HDAC2 with FOXA2), and for contrast, examples with relatively even co-occurrence frequencies cross context (for example, RELA with RUNX3).