Poly-Enrich: Count-based Methods for Gene Set Enrichment Testing with Genomic Regions

Christopher T Lee[1], Raymond G Cavalcante[2,*], Chee Lee[2,^], Tingting Qin[2], Snehal Patil[2], Shuze Wang[2], Zing TY Tsai[2,§], Alan P Boyle[2], Maureen A Sartor[1,2]

[1] Biostatistics Department, University of Michigan

[2] Department of Computational Medicine and Bioinformatics, University of Michigan

*** Currently at Epigenomics BRCF Core, University of Michigan**

^ Currently at LA Care Health Plan, Los Angeles, CA

§ Currently at Illumina Inc, San Diego, CA

Abstract

Gene set enrichment (GSE) testing enhances the biological interpretation of ChIP-seq data and other large sets of genomic regions. Our group has previously introduced two GSE methods for genomic regions: ChIP-Enrich for narrow regions and Broad-Enrich for broad genomic regions, such as histone modifications. Here, we introduce new methods and extensions that more appropriately analyze sets of genomic regions with vastly different properties. First, we introduce Poly-Enrich,

which models the number of peaks assigned to a gene using a generalized additive model with a negative binomial family to determine gene set enrichment, while adjusting for gene locus length (#bps associated with each gene). This is the first method that controls for locus length while accounting for the number of peaks per gene and variability among genes. We also introduce a flexible weighting approach to incorporate region scores, a hybrid enrichment approach, and support for new gene set databases and reference genomes/species.

As opposed to ChIP-Enrich, Poly-Enrich works well even when nearly all genes have a peak. To illustrate this, we used Poly-Enrich to characterize the pathways and types of genic regions (introns, promoters, etc) enriched with different families of repetitive elements. By comparing ChIP-Enrich and Poly-Enrich results from ENCODE ChIP-seq data, we found that the optimal test depends more on the pathway being regulated than on the transcription factor or other properties of the dataset. Using known transcription factor functions, we discovered clusters of related biological processes consistently better modeled with either the binary score method (ChIP-Enrich) or count based method (Poly-Enrich). This suggests that the regulation of certain processes is more often modified by multiple binding events (count-based), while others tend to require only one (binary). Our new hybrid method handles this by automatically choosing the optimal method, with correct FDR-adjustment.

Author Summary

Although every cell in our body contains the same DNA, our cells perform vastly different functions due to differences in how our genes are regulated. Certain regions of the genome are bound by DNA binding proteins (transcription factors), which regulate the expression of nearby genes. After an experiment to identify a large set of these regions, we can then model the association of these regions with various cellular pathways and biological processes. This analysis helps understand the overall biological effect that the binding events have on the cells. For example, if genes relating to apoptosis tend to have the transcription factor, Bcl-2, bind more often nearby, then Bcl-2 is likely to have a vital role in regulating apoptosis. The specifics of how to best perform this analysis is still being researched and depends on properties of the set of genomic regions. Here, we introduce a new, more flexible method that counts the number of occurrences per gene and models that in a sophisticated statistical test, and compare it to a previous method. We show that the optimal method depends on multiple factors, and the new method, Poly-Enrich, allows interesting findings in scenarios where the previous method failed.

Introduction

Regulatory genomics experiments help us understand how cells use more than their genetic sequence to carry out a vast repertoire of cellular programs. Common regulatory genomics methods include chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-Seq) and ATAC-seq, which identify transcription factor (TF) binding sites and open chromatin regions, respectively, across the

genome. Other types of data, such as DNA Methylation assays, copy number alterations, repetitive element families and groups of SNPs, also lead to large sets of genomic regions that potentially play a specific role in regulatory genomics, with each type having notably different properties in terms of the number, size, and location of genomic regions.

Proteins that bind near a gene can regulate it in ways such as improving structural properties or physically blocking other proteins, often positively or negatively regulating the gene's expression, respectively. Additionally, some proteins bind DNA several times in a clustered region [1], or in distant enhancer regions that interact with the same or distinct proteins bound in promoter regions [2]. Binding sites also differ in strength; a protein may bind in only a portion of cells in a sample at the time of immunoprecipitation, either due to weak binding or due to varying chromatin accessibility among the cell types in the sample. These binding sites along the genome are interpreted as peaks of varying strengths, depending on the signal-to-noise ratio or significance level of the peak. In general, interpreting each peak's target gene(s) and effects remains an active area of research, which over time may improve results on downstream tests such as gene set enrichment.

Gene Set Enrichment (GSE) is an approach to test for over (or under) representation of genes in a set of genes with similar functionalities. Gene Ontology [3], Reactome [4], KEGG pathways [5], and MsigDB [6] are widely used gene set databases. Although originally developed for gene expression data, GSE testing is now often

87 used to help interpret ChIP-seq peak sets and other sets of genomic regions. Existing
88 methods for general GSE tests include Fisher's exact test, random sets, logistic
89 regression like LRPath [7], and GSEA-type tests [8]. GSE methods specifically for
90 ChIP-seq data include Genomic Regions Enrichment of Annotations Tool (GREAT)
91 [9], ChIP-Enrich [10], and Broad-Enrich [11]. With so many different tests, one may
92 wonder which test is optimal to use, but there is no single recommendation across
93 data types. Different tests are needed for different types of genomic regions as
94 properties such as peak widths, number of peaks, and location relative to genes all
95 make a difference. Thus GSE testing for genomic regions should not be a one-size-
96 fits-all test; some methods work better than others in specific scenarios. For
97 example, Cavalcante et al. showed that Broad-Enrich is more powerful than ChIP-
98 Enrich for broad regions, but lacks power for narrow regions [11]. As another
99 example, GREAT does not account for variability among genes, so it is best used in
100 situations where the probability of a peak is constant across genomic space (e.g. per
101 kb), as opposed to clustered near transcription start sites or displaying variability
102 among gene loci.
103
104 Our previous method, ChIP-Enrich, uses a binary score to classify a gene as having at
105 least one peak. We saw that ChIP-Enrich tends to underperform when nearly all
106 genes have at least one associated genomic region; in this case, ChIP-Enrich will not
107 yield meaningful results. We hypothesized that a count-based method that captures
108 the frequency of binding would perform better in those situations. In this paper, we
109 introduce such a method, Poly-Enrich to expand our available methods to be

110 suitable for any set of narrow genomic regions including those that tend to saturate
111 genes. Whereas ChIP-Enrich has the hypothesis that a single binding site is
112 sufficient for regulation, Poly-Enrich allows for regulation that is incremental, i.e.
113 more genomic regions correspond to stronger or more likely regulation. To identify
114 under which situations one is more appropriate than the other, we performed a
115 comparison of Poly-Enrich and ChIP-Enrich using a set of 90 transcription factor
116 (TF) ChIP-seq datasets from the Encyclopedia of DNA Elements (ENCODE) [12]. We
117 also introduce a hybrid test that combines information from both ChIP-Enrich and
118 Poly-Enrich.
119
120 To illustrate the usage of Poly-Enrich, we apply it to sets of repetitive elements in
121 the human Alu and LINE1 families, revealing for the first time a comprehensive view
122 of the processes and functions enriched or depleted with these repetitive elements
123 in the human genome. Finally, we describe several updates to our ChIP-Enrich
124 website and *chipenrich* Bioconductor package, including additional methods for
125 assigning genomic regions to target genes, new gene set databases, and more
126 supported species.
127

128 Results

129 **Motivation for development of Poly-Enrich**

130 The motivation for our new methods comes from situations observed with real sets
131 of genomic regions, often with ChIP-seq peak datasets, but also from other sources,

132 such as families of repetitive elements or large sets of DNA polymorphisms such as
133 those different between closely related species or sub-species. Although our original
134 method, ChIP-Enrich, performs extremely well for most transcription factor (TF)
135 ChIP-seq datasets (Figure 1a), because it uses a simple binary score for each gene,
136 there are some scenarios where this simplification has a significant loss of
137 information. For example, ChIP-Enrich models a gene with many peaks the same as
138 a gene with only one peak, even though gene regulation may be affected by
139 additional peaks (Figure 1b). Alternatively, if nearly every gene is assigned at least
140 one peak, ChIP-Enrich would be unable to distinguish among them and thus unable
141 to detect any gene set enrichment (Figure 1c). We therefore developed Poly-Enrich
142 as a count-based method that addresses both of these scenarios.
143
144 **Figure 1:** Three scenarios of ChIP-seq peak distributions illustrating how ChIP-
145 Enrich and Poly-Enrich perform. Each color represents a different gene locus; the
146 left three are in a gene set and the right three are not. (a.) Peaks are relatively evenly
147 distributed, with a small number across a subset of genes. Given this situation, ChIP-
148 Enrich evaluates 2/3 vs 1/3 while Poly-Enrich evaluates {0,2,1} vs {0,1,0}; both
149 methods perform well. (b.) Some genes contain significantly more peaks than
150 others, such that information is to be gained from the number per gene. ChIP-Enrich
151 evaluates 2/3 vs 1/3, Poly-Enrich evaluates {0,2,3} vs {0,1,0}; ChIP-Enrich performs
152 adequately, but Poly-Enrich is optimal. (c.) Nearly all genes have at least one peak,
153 with some having significantly more than others. ChIP-Enrich evaluates 3/3 vs 3/3,

154 Poly-Enrich evaluates {1,2,3} vs {1,1,1}; ChIP-Enrich would not detect any
155 enrichment, while Poly-Enrich can still detect gene sets enriched with more peaks.
156
157 Although GREAT is also a count-based gene set enrichment method, Poly-Enrich
158 differs significantly from it in two respects. Firstly, whereas GREAT counts the
159 number of peaks in an entire gene set, Poly-Enrich counts them per gene. By
160 separating counts per gene, we are able to adjust for each gene's locus length and
161 the variability in peak count across genes, which we previously showed was an
162 important adjustment to control for Type I error [10]. Secondly, the binomial model
163 used by GREAT assumes that the background probability of a peak is constant
164 across the genome. Poly-Enrich uses a more flexible, empirical approach to this that
165 provides for a range of different assumptions about peak distribution.
166
167 **Assigning peaks to genes**
168 Genomic regions can be assigned to genes in different ways, so that regulation from
169 different types of regions (e.g., promoters, introns, or regions distal to TSSs) can be
170 studied. For example, ChIP-Enrich, GREAT, and Poly-Enrich all use a peak's midpoint
171 to define the location of the peak. We define a gene's locus definition as the region
172 on the genome such that peaks in that region are assigned to the gene. These loci are
173 defined using properties of the gene, such as within 5kb of a gene's transcription
174 start site (TSS), or simply by assigning each region to the nearest TSS (Figure 2). In
175 the new version of our website and Bioconductor package, we offer several
176 additional choices, including exons, introns, and distal regions only (>10kb

upstream from a TSS). Users can also upload their own custom locus definition, such as open chromatin regions for a specific cell type, or known enhancers and their target genes.

Figure 2: Overview of peak-to-gene assignments given gene locus definitions. Examples shown are: "<5kb", peaks within 5 kilobases of a gene's TSS are assigned to the gene; "Nearest TSS", peaks are assigned to the gene with the closest TSS. A gene's locus length is defined by the number of base pairs assigned to the gene. In this toy example, peak 1 would be assigned to the blue gene in the <5kb locus definition and for Nearest TSS, while peak 2 would not be assigned to any gene in the <5kb locus definition.

Poly-Enrich model

We model the number of peaks per gene using a negative binomial generalized linear regression as a function of gene set membership, and with a cubic smoothing spline to empirically model the relationship with gene locus length:

$$\log(\mu_i) = \beta_0 + \beta_1 GS_i + f(LL_i)$$

where for gene *i*, μ_i is the expected mean number of genomic regions assigned to the gene, GS_i=1 if the gene is a member of the gene set of interest and 0 otherwise, and $f(LL_i)$ is a negative binomial cubic smoothing spline to adjust for the gene's locus length. We then look at the sign and significance of β_1 to test for enrichment, where a positive β_1 indicates enrichment, and a negative value indicates depletion (fewer regions than expected at random).

199

200 **Testing Type 1 error and power**

201 We tested the type I error rate of the count-based method under the null hypothesis
202 of no enrichment signal. By permuting the genes in the peak-to-gene assignment
203 pairs and breaking the peak-gene relationships, we simulated three scenarios of no
204 enrichment: *i)* the "*complete*" randomization was done by shuffling the gene IDs in
205 the whole dataset; *ii)* the "*bylength*" randomization first groups the genes together
206 into bins of similar locus length, and then randomizes genes within those bins to
207 preserve the locus length relationship; *iii)* the "*bylocation*" randomization groups
208 genes together by their physical location on the chromosomes, and then randomizes
209 genes within those bins. (See *Methods* for more detail.) We ran the randomizations
210 on our 90 selected ChIP-Seq datasets from ENCODE (see *Methods*), 10 times each,
211 and the proportion of p-values < 0.05 and < 0.001 for each dataset were plotted
212 (Supplementary Figure 1). We see that the test is properly controlled at an
213 acceptable level for Type 1 error in all cases.

214

215 To characterize the statistical power of Poly-Enrich under different situations, we
216 permutated data while simulating enrichment of a gene set, and compared the
217 results with those from ChIP-Enrich. We used three datasets with a small, medium,
218 and large number of peaks, and two GO terms with a small and large number of
219 genes. Three types of enrichment were simulated: one that adds peaks mainly
220 according to the regulatory assumptions of ChIP-Enrich (CEBias), one that adds
221 peaks mainly according to the assumptions of Poly-Enrich (PEBias), and one that is

222 balanced. For each type of enrichment, we simulated four levels of enrichment: 0.05,
223 0.1, 0.2, and 0.3, which indicates the proportion of additional peaks added to the
224 gene set. (See *Methods* for more detail.) Finally, we chose two different levels of
225 significance: $\alpha = 0.05$ and 0.001, as our cutoffs.
226
227 As expected, a larger gene set and higher simulated enrichment results in higher
228 power. Overall, we see that Poly-Enrich has more power in simulations that enrich a
229 gene set by increasing the number of peaks per gene, and ChIP-Enrich has more
230 power than Poly-Enrich in simulations that enrich a gene set by adding peaks to
231 genes without them. Finally, the *Balanced* simulation results in the two methods
232 having similar power in most cases (Supplementary Figure 2). Simulations on larger
233 datasets artificially reduced power because randomized larger datasets include
234 more noise. However, in real experiments, we expect larger datasets (more peaks)
235 to be more powerful, since the majority of the peaks in the dataset are not expected
236 to be noise.
237
238 **Poly-Enrich with weighted genomic regions**
239 The height and confidence of peaks in a ChIP-seq experiment can vary dramatically,
240 and we reasoned that incorporating this additional information would improve the
241 ability to pinpoint the truly enriched pathways. Although the most apparent
242 motivation for weighting genomic regions is to account for ChIP-seq peak strength,
243 there are other situations where each peak or genomic region may be assigned a
244 unique score (e.g. confidence or strength). Therefore, we added the option to weight

245 regions by signal value (see *Methods* for details), and examined the extent to which
246 adjusting for peak strength improves enrichment results using 90 ENCODE ChIP-seq
247 datasets by comparing the $-\log_{10}$ p-values per gene set. To also ensure that no
248 enrichment result flipped from enriched to depleted or vice versa, we used a signed
249 $-\log_{10}$ p-value, where values for depleted gene sets were negative, and values for
250 enriched gene sets were positive.
251
252 We noticed for 25% of the experiments, most enriched gene sets were more
253 significant with weighting, thus as we hypothesized, binding events near genes in
254 enriched GO terms were stronger than those near other genes (Figure 3A, 3B). In
255 another 20% of the experiments, the enrichment p-values were split between the
256 two methods (Figure 3C). Interestingly, the distribution of log signal values for these
257 experiments showed a bimodal pattern (Figure 3D). This suggests that some gene
258 sets tend to have genes with significantly stronger binding peaks than others, and
259 that both sets may be biologically interesting. For the remaining 55% of
260 experiments tested, weighting made little difference on the results.
261
262 **Figure 3:** Comparison of GO term enrichment results between standard Poly-Enrich
263 and its weighted version using signal values as weights. Each point is a GO term's –
264 $\log_{10}$ p-value of the two methods, signed positive for enriched, negative for depleted.
265 (A) Using weighting results in more significant enrichment in many GO terms in the
266 Gm12878 TAF1 ChIP-Seq experiment. (B) Using weighting results in less significant
267 enrichment in many GO terms in the H1-hESC EGR1 ChIP-Seq experiment. (C) Using

268 weighting on the Gm12878 NRSF experiment results in several more significant GO
269 terms as well as several less significant ones. (D) The histogram of log signal values
270 from the NRSF experiment. There is a bimodal pattern in the weights, suggesting
271 that there are some GO terms with genes that tend to have stronger or weaker
272 binding.
273
274 **Comparison of the count-based (Poly-Enrich) versus binary (ChIP-Enrich)**
275 **model of enrichment**
276 We next compared signed log10 p-values from Poly-Enrich versus ChIP-Enrich on
277 the same set of 90 ENCODE ChIP-seq datasets. Our initial hypothesis was that some
278 experiments would be clearly modeled better by one method or the other (i.e.
279 dependent on the transcription factor). However, our results strongly suggest that
280 the optimal method is more dependent on the gene set than the TF. This can be
281 visualized by a split in the significance levels of GO terms between the binary and
282 count-based methods (Figure 4A), and suggests that a single transcription factor
283 may regulate genes differently depending on the function of the gene. Thus, we
284 sought to understand this further.
285
286 **Figure 4:** (A) Comparison of GO term significance levels between ChIP-Enrich and
287 Poly-Enrich. Each point is the $-\log_{10}$ p-value of a GO term from the two methods,
288 signed positive for enriched or negative for depleted. Several gene sets are much
289 more significant using ChIP-Enrich and several are much more significant using
290 Poly-Enrich, however 32% of the datasets showed a split pattern like shown. (B)

291 Heatmap of $-\log_{10}$ p-value differences between Poly-Enrich and ChIP-Enrich for GO
292 terms and ChIP-seq experiments, where each row is a GO term and each column is a
293 ChIP-seq experiment. Shown are GO terms where more than 15% of the
294 experiments had a $-\log_{10}$ p-value difference of 2 or larger. Red indicates Poly-Enrich
295 was more significant, and blue indicates ChIP-Enrich was more significant. Light
296 grey indicates the transcription factor used in the experiment was not assigned to
297 the GO term and is omitted in the clustering. Representative GO terms are shown for
298 each cluster. (C) GO terms related to cell cycle are mostly blue, indicating that a
299 binary score provides a more appropriate model. (D) GO terms containing "positive
300 regulation" are mostly red, indicating that a count score provides a more
301 appropriate model.
302
303 The binary model used by ChIP-Enrich assumes that a single binding event (i.e. a
304 single genomic region) is sufficient for regulation, while the Poly-Enrich count-
305 based model assumes that strength of regulation is incremental with the number of
306 binding sites. Based on the results above, we asked what kinds of genes were more
307 consistent with either of those assumptions. To answer this, we first created a set of
308 true positives comprised of GO term-TF pairs by using the GO term biological
309 process (BP) assignments for the gene encoding the transcription factor (e.g. the
310 gene encoding for JunD is assigned to the GO term, "cell death"). This gold standard
311 makes the reasonable assumption that TFs tend to regulate genes in the same
312 biological processes in which they are active. Observing the enrichment results
313 using the 5kb locus definition for these true positive GO term-TF pairs, we used

314 clustering to identify patterns of TFs and GO terms that are optimal with one of the
315 methods. We found that the method that worked better was most often determined
316 by the GO term (Figure 4B). For example, GO terms related to "cell cycle" clustered
317 together and displayed greater power with ChIP-Enrich. Conversely, GO terms
318 involving "positive regulation" tended to do better with Poly-Enrich except for those
319 involving cell cycle (Figure 4C,4D). Parallel results using the Nearest TSS locus
320 definition were similar (Supplementary Figure 3).
321
322 Since the gene set, rather than the experiment, was a stronger determinant of the
323 more appropriate method, in many cases we are unable to recommend either Poly-
324 Enrich or ChIP-Enrich for an entire experiment; one exception is that Poly-Enrich is
325 recommended for experiments with a very large number (>100k) of peaks. We
326 therefore developed a hybrid test that uses information from both ChIP-Enrich and
327 Poly-Enrich.
328
329 **Hybrid test**
330 To obtain the best results across all types of GO terms and datasets, we developed a
331 hybrid test that incorporates both the binary and count-based models. After
332 performing the two models, the hybrid p-value of the two tests is defined as:
333 $p_{hybrid} = 2 \times \min(p_{CE}, p_{PE})$ [13], where p_{CE} and p_{PE} are the p-values given by ChIP-
334 Enrich and Poly-Enrich, respectively. This is essentially a Bonferroni-adjusted p-
335 value for two tests. This hybrid has been shown to be beneficial if the two tests are
336 sufficiently different, but loses power and is conservative if the tests are identical or

337 nearly identical [13]. While the hybrid test is not as powerful as the better method
338 between ChIP-Enrich and Poly-Enrich, it is dramatically more powerful than using
339 the worse method, making it the optimal method to use across all GO terms (Figure
340 5). While this hybrid test currently only accommodates ChIP and Poly-Enrich, we
341 can extend this to accommodate several additional gene set enrichment tests.
342
343 **Figure 5:** Statistical power comparisons for Poly-Enrich (red), ChIP-Enrich (blue),
344 and the hybrid test (gold) for datasets with three different sizes (i.e. number of
345 peaks: small, medium, and large) and two gene set sizes (small and large GS), under
346 two significance levels: $\alpha = 0.05$ (A) and 0.001 (B), and three different methods of
347 simulated enrichment (CEbias: add peaks according to the regulatory assumptions
348 of ChIP-Enrich, PEbias: add peaks mainly according to the assumptions of Poly-
349 Enrich, Balanced: add peaks proportional to each gene's locus length). The values on
350 the X-axis indicate the percent of extra peaks added to simulate enrichment; a
351 higher value simulates stronger enrichment. The hybrid test is shown to have much
352 more power than the wrong method, but only a bit less power than the correct
353 method.
354
355 **Identifying biological processes enriched with or depleted in repetitive**
356 **element families using Poly-Enrich**
357 To further illustrate the utility of Poly-Enrich, we used it to test sets of repetitive
358 element regions. We asked whether we could identify gene sets that tended to be
359 either enriched or depleted for certain types of repetitive elements. Significant

360 enrichment of the repetitive elements in the promoter regions of genes, for example,
361 can sequester the transcription factors that will inhibit activities at another
362 transcription factor binding site or other regulatory motif [14]. Some of these
363 mobile elements remain active with new insertions having neutral, detrimental, or
364 beneficial effects. Although repetitive element families have been well studied for
365 over 30 years, little is yet known about the biological processes that they have
366 adapted to help regulate or that they can too easily disrupt and thus are negatively
367 selected against [15]. Using the database of human repetitive elements from the
368 UCSC Table Browser (RepeatMasker 3.0) [16], we performed GSE testing on
369 repetitive element families. Certain families of repetitive elements have over a
370 million occurrences across the human genome, and thus virtually all genes have at
371 least one nearby instance, making this an example where ChIP-Enrich performs
372 poorly. Thus, in this situation, modeling the number of insertions per gene is critical
373 to identify differences.
374
375 We examined two of the most abundant types of repetitive elements: the *Alu* and
376 LINE1 (L1) elements, which make up an estimated 11% and 17% of the human
377 genome, respectively [17, 18]. We also chose four gene locus definitions: Nearest
378 TSS, <5kb (promoter regions), >5kb (distal regions), and Intron. We tested GO
379 Biological Processes, and used hierarchical clustering of the resulting GSE
380 significance levels to identify related groups of biological processes enriched with or
381 depleted of the repetitive elements (Figure 6). We found strong enrichment of Alu's
382 in GO terms describing metabolic processes, most significantly "ATP metabolic

383 process" and "rRNA metabolic process", especially in promoter regions, which is
384 consistent with an analysis of *Alu* distribution in chromosomes 21 and 22 that
385 showed *Alu* elements on these chromosomes were enriched in or near metabolism
386 and signaling genes [19]. Conversely, Alu elements were sharply depleted in the
387 promoter regions of many development and morphogenesis gene sets, with the
388 strongest depletions in "cell fate commitment" and "connective tissue development".
389 Interestingly, depletions were also seen in the introns of genes in these gene sets,
390 but not in regions >5kb upstream, suggesting the negative selection (depletion) is
391 limited to the regions that are more commonly regulatory.
392
393 **Figure 6:** Enrichment results using Poly-Enrich for Alu (first four columns) and L1
394 (last four columns) repetitive element families using four different peak-to-gene
395 assignments. Shown are signed $-\log_{10}$ FDR, where positive values (red) indicate
396 enrichment and negative values (blue) indicate depletion. Only GO terms that were
397 significant for at least 3 columns at the FDR = 0.05 level are displayed. We identified
398 nine clusters of GO terms with similar enrichment patterns. Representative GO
399 terms are used to label each cluster.
400
401 Novel insertions of L1 elements into or near key genes are known to be associated
402 with neurological diseases [20]. Consistent with this, we found that all neuro-related
403 GO terms in Figure 6 were depleted for L1 (but not for Alu) (Supplementary Figure
404 4), which suggests that L1's evolutionarily have been selected against occurring in

405 the regulatory regions of neurological genes; when they are inserted into the introns
406 or promoters of these genes, the inserted elements may cause disease.
407
408 In general, we observed that the significance of the distal upstream regions (>5kb
409 locus definition) was lower than the other three locus definitions (with the
410 exception of some enrichments for Alu elements) (Supplementary Table 1),
411 implying that most repetitive element negative (or positive) selection has occurred
412 in the promoter regions or introns of genes. Alternatively, the gene distal enriched
413 and depleted regions may be limited to a specific set of enhancer regions, the signal
414 from which could have been diluted in our analysis. Interesting additional findings
415 are that L1 elements are enriched in chemical stimulus detection such as "detection
416 of chemical stimulus" and "sensory perception of chemical stimulus", while Alu
417 elements are depleted in the genes in these processes. We also find that both Alu
418 and L1 are enriched in centrosome-related GO terms, which were only made
419 possible with recent advancements in genome mapping near the centromeres [21],
420 and is consistent with previous findings [22] . Finally, both Alu and L1 elements are
421 significantly depleted in genes in GO terms relating to development and
422 morphogenesis.
423
424 **Availability, usage, and updates**
425 Poly-Enrich is available in the *chipenrich* Bioconductor package and as a web
426 interface at http://chip-enrich.med.umich.edu. Several additional gene set

427 databases and gene locus definitions have been added for the user to choose (see
428 http://chip-enrich.med.umich.edu/data/ChipenrichMethods.pdf).
429
430 To perform gene set enrichment analysis, the user first needs a file of genomic
431 regions, which may be a narrowPeak, BED, or text file with chromosome, start, and
432 end positions for each region. The user then selects a species, one or more gene set
433 databases, a gene locus definition, and the test method (ChIP-Enrich, Poly-Enrich,
434 Hybrid, or Fisher's exact test); the gene set and locus definition can be built-in or
435 user-defined. The user can then also choose to add weight based on a peak-specific
436 score, and a number of other options, such as adjustment for read mappability.
437
438 The enrichment function outputs five files:
439 - opts: The options that the user input into the function.
440 - peaks: A peak-level summary showing the peak-to-gene assignments for each
441 peak.
442 - peaks-per-gene: A gene-level summary showing gene locus lengths and the
443 number of peaks assigned to them.
444 - results: The results of the GSE tests. Lists the tested gene sets along with
445 their descriptions, the test effect, odds ratio, enrichment status, p-value, and
446 FDR. Also included is the list of gene IDs with contributing signal for each
447 enrichment test.
448 - qcplot: A plot of the gene locus lengths with a fitted smoothing spline.
449

450 # Discussion

451 Gene set enrichment testing methods for genomic regions should take into account
452 the differing properties of the input datasets, including the widths and number of
453 genomic regions, and where they tend to occur relative to genes. However, no single
454 method is appropriate for all types, and therefore no single GSE method should be
455 recommended for all sets of genomic regions. Although our previously developed
456 ChIP-Enrich method for gene set enrichment with genomic regions performs well
457 for most transcription factor ChIP-seq datasets [10], above we described some
458 common situations where it does not. Such cases include when nearly all genes are
459 assigned at least one genomic region, and when the strength or likelihood of
460 regulation increases incrementally with the number of genomic regions. As an
461 example, the transcription factor NF-kappaB is known to regulate the gene NFKBIA
462 by binding to a few or even many motif positions in the promoter [23], with gene
463 expression correlated with the number of bound factors. Thus, motivated by specific
464 examples of regulatory mechanisms, we developed Poly-Enrich, a method that
465 models the number of regions per gene, empirically adjusts for each gene's locus
466 length, and takes into account variability among genes in each gene set. Poly-Enrich
467 is also flexible, in that it easily allows for weighting of each genomic region by any
468 score of interest. We used the example of weighting by peak strength, but other
469 examples include weighting by SNP significance in a GWAS analysis, by the inverse
470 distance to a gene, or by the probability that the region is in an open chromatin
471 region in a particular cell type.

472

473 We showed that our count-based method, Poly-Enrich, works well when almost all
474 genes are assigned a peak, whereas ChIP-Enrich does not. In comparing when each
475 is most appropriate, we discovered that it is mostly dependent on the gene set,
476 rather than the transcription factor. Because in many cases we can not recommend
477 a single best method to test all gene sets for an experiment, we developed and
478 implemented a hybrid test that uses information from both methods and performs
479 better than either test across GO terms for most datasets.

480

481 When applying Poly-Enrich to repetitive element families, we both reconfirmed
482 known associations and also identified novel findings. Poly-Enrich confirmed that
483 Alu elements tend to regulate genes for metabolism and signaling by finding
484 enrichment for related GO terms. Additionally, we know that L1 insertions into or
485 near certain neurological-related genes are associated with neurological diseases
486 [24]. We find that L1 is depleted in neuro-related GO terms, implying there normally
487 are fewer L1 elements in the regulatory regions of these genes, which is consistent
488 with neurological diseases being associated with L1 element insertions near these
489 genes. We also find that there is little enrichment or depletion in the distal
490 regulatory regions of genes, suggesting that repetitive elements may not have as
491 large of an affect there due to mitigated regulatory activity at larger distances from
492 transcription start sites. Poly-Enrich also detected some novel associations between
493 repetitive element families and biological pathways. Both Alu and L1 elements were
494 significantly depleted in genes in GO terms relating to development and

495 morphogenesis, such as "connective tissue development" and "skeletal system
496 morphogenesis", suggesting that it is especially critical to have developmental
497 regulatory regions free from potentially disruptive repetitive elements during early
498 growth.
499
500 One shortcoming of our current methods (as well as current alternatives) is that
501 they rely on associating each genomic region with the nearest gene. However, it is
502 estimated that 79-95% of DNAse I hypersensitive sites, markers for enhancer
503 regions, actually regulate a different, distal target gene [24, 25]. We are currently
504 developing a set of enhancer locus definitions that identify and assign enhancer
505 regions to their appropriate target genes, as was recently introduced by Chicco D, et
506 al [25], so peaks in enhancer regions will be correctly assigned and false positive
507 peaks in nonfunctional intergenic regions will be filtered out. We believe this will
508 improve all future gene enrichment analyses.
509

510 Methods

511 **Datasets:**

512 All ChIP-Seq data were obtained from Encyclopedia of DNA Elements (ENCODE) at
513 University of Califonia, Santa Cruz [12]. We chose a total of 90 experiments over
514 the three Tier 1 cell lines (Gm12878, H1-hESC, and K562), and all 35 transcription
515 factors that had available ChIP-seq data for at least two of the three Tier 1 cell lines.
516 (Supplementary Table 2)

517

518 The gene sets used were from Gene Ontology: Biological Processes (GOBP) ver. 3.4.2
519 [3]. We filtered out gene sets with less than 15 genes or more than 2000 genes as
520 gene sets with too few genes generally have insufficient power and may not satisfy
521 the assumptions of the statistical model, and gene sets with too many genes are too
522 vague to be biologically informative. In total, there were 5015 gene sets.

523

524 **Assigning regions to genes**

525 The UCSC knownGene database for hg19 was used to define the transcription start
526 sites across the genome [26]. Each locus definition (e.g. nearest TSS, <5kb) was
527 generated as a table containing the columns: chromosome, Start, End, gene ID. All
528 genomic regions whose midpoint was between a gene's start and end values would
529 be assigned to that gene. It is possible for a locus definition to have several disjoint
530 regions for a certain gene.

531

532 **Poly-Enrich model: a generalized linear model with a negative binomial family**

533 We model the number of genomic regions assigned to each gene with a generalized
534 linear model (GLM) with a negative binomial (NB) family. The model is:

$$\log(\mu_i) = \beta_0 + \beta_1 GS_i + f(\log(LL_i * m + 1)$$

535 where for each gene i, GS is an indicator for whether the gene is in the gene set of
536 interest or not (=1 if in the gene set; 0 otherwise), μ is the mean of the negative
537 binomial distribution for the number of genomic regions assigned to each gene, and
538 the overdispersion parameter θ is estimated so that $Var(Y|GS) = \mu + \theta\mu^2$, where Y

539 is the number of genomic regions for the gene. The function f is a negative binomial
540 cubic smoothing spline that adjusts for the gene's locus length and optionally
541 adjusts for *m*, the mappability of the gene's locus. Details about mappability can be
542 found in the ChIP-Enrich manuscript [10]. We use the *gam* function in the *mgcv* R
543 package to fit the model, which uses a penalized likelihood maximization, and the
544 smoothing spline penalty is a squared second derivative penalty [27]. Use of a cubic
545 smoothing spline to adjust for the genes' locus lengths was first introduced in ChIP-
546 Enrich, as a powerful, flexible way to model this relationship.
547
548 A Wald test on the coefficient for the gene set is used to test for enrichment (or
549 depletion) of each gene set: the test statistic is defined as $W^2 = \widehat{\beta_1}/\hat{V}(\widehat{\beta_1})$, which
550 follows a χ_1^2 distribution under the null hypothesis that there is no association
551 between gene set membership and number of genomic regions.
552
553 **Poly-Enrich with weighting based on genomic region scores**
554 In certain cases, each genomic region in a dataset may be associated with a numeric
555 score. For example, ChIP-seq peak finding results often include a value denoting the
556 strength of a peak, (e.g. signalValue in ENCODE dataset results or -10*$\log_{10}$(p-value)
557 in MACS2 results). Poly-Enrich weights based on these scores by giving each
558 genomic region a weight proportional to its signal value (or other score) and
559 normalizing such that the mean of all weights is equal to 1. For every genomic
560 region assigned to a gene, we sum all the weights and substitute the weighted sum
561 in place of the original count. The same model is used, except assuming a quasi-

562 negative binomial family to accommodate for non-whole number data. The
563 calculations can be carried out identically to the standard negative binomial family.
564
565 **Comparing p-values between methods**
566 To compare p-values between methods, we use a scatterplot, plotting a signed
567 -$\log_{10}$ p-value per gene set. If a gene set is enriched, the sign is positive, and if the
568 gene is depleted, the sign is negative. This allows us to detect if there are any cases
569 where two methods may contradict each other's conclusions.
570
571 **Spline approximation for Poly-Enrich**
572 With a library of over 20,000 genes and most gene sets being less than 1000 genes,
573 the cubic smoothing spline estimate changes very little between gene sets. Thus, we
574 have confirmed we can reasonably assume that the spline is approximately equal for
575 any gene set of interest, including the spline with no gene set (Supplementary
576 Figure 5A).
577
578 We first run the same model except without the gene set (GS) term:
579 $\log(\mu_i) = \beta_0 + f(LL_i)$. We then extract the fitted spline using the *predict* function
580 with *type="terms"* from the *mgcv* R package to obtain a spline-adjusted locus length
581 for each gene. This new value is then input in the model for every gene set, which
582 allows us to fit a spline only once instead of once for each gene set. This saves a
583 significant amount of time when testing a large number of gene sets (approximately
584 75% time saved when testing 4000 gene sets). Compared to the original model, we

585 find that the p-values from the spline approximation model are nearly identical

586 (Supplementary Figure 5B, 5C).

587

588 **Use of Score test for quick approximations**

589 One of the alternatives for the Wald test is the Score test [28]. To obtain quick,

590 approximate enrichment results for many sets of genomic regions (e.g. a large group

591 of ChIP-seq peak datasets) using either the Poly-Enrich or ChIP-Enrich approach, we

592 now provide the Score test approximation for convenience. We can calculate the

593 score test statistic for ChIP-Enrich as:

594

595 $$S_{CE}^2 = \frac{\sum_{i=1}^{n} GS_i(W_i - \hat{\pi}_i)}{\sum_{i=1}^{n} GS_i\, \hat{\pi}_i\, (1-\hat{\pi}_i)}$$

596 where for each gene i, GS_i=1 if the gene is in the gene set of interest and 0

597 otherwise, W_i=1 if the gene has at least one genomic region and 0 otherwise, and $\hat{\pi}_i$

598 is the predicted probability of the gene having at least one genomic region obtained

599 using the *fitted* function with *type="terms"* in the *mgcv* package, assuming the null

600 hypothesis of no association (i.e. β_1=0 in the model).

601

602 The score test statistic for Poly-Enrich is:

603 $$S_{PE}^2 = \frac{\sum_{i=1}^{n} GS_i(Y_i - \hat{\mu}_i)}{\sum_{i=1}^{n} GS_i\, \hat{V}(\hat{\mu}_i)}$$

604

605 where for each gene i, GS_i=1 if the gene is in the gene set of interest and 0
606 otherwise, Y_i is the number of genomic regions assigned to the gene, $\hat{\mu}$ is the
607 predicted number of genomic regions assigned to the gene obtained using the *fitted*
608 function with *type="terms"* in the *mgcv* package, assuming the null hypothesis of no
609 association, and its variance is estimated empirically, assuming $GS_i = 0$ in the
610 variance calculation.
611
612 The advantage of using the score test is that all the required parameters are already
613 estimated during the spline approximation, so all subsequent calculations will no
614 longer require fitting of a GLM. This reduces the runtime of the enrichments
615 significantly further (approximately 95% time saved when testing 4000 gene sets).
616 However, there are some scenarios where the Score test differs from the Wald test
617 by a substantial amount, mostly notably for depleted gene sets, so the default option
618 is set to the Wald test (Supplementary Fig 6). If results are desired quickly, the Score
619 test is offered (method="chipapprox" or "polyapprox") and can serve as a
620 convenient approximation.
621
622 **Testing Type I error**
623 The null hypothesis of Poly-Enrich is that there is no true biological enrichment. To
624 test the Type-I error, we randomly permute the genes to simulate a scenario where
625 there is no association between genes and the number of peaks. However, to ensure
626 that the results are not biased by gene locus length or gene location, we performed
627 two additional permutations: one permutes genes within bins of similar locus

628 length, while the other permutes within bins of chromosomal locations. In both
629 cases, the genes are sorted by the variable of interest (locus length or location), and
630 then assigned to consecutive bins of 100 genes each.
631
632 For each of the 90 TF peak datasets chosen, after assigning the peaks to genes, we
633 permute the gene IDs using the randomization of interest, and then perform
634 enrichment tests against GO biological processes. We ran a total of 10 trials and took
635 the median p-value per gene set as the randomization p-value. Then, the proportion
636 of p-values less than a defined confidence level was determined per experiment to
637 calculate the overall Type I error. We then plotted all 90 overall Type I errors for
638 each experiment in a box plot to convey overall Type I error.
639
640 **Testing power**
641 To test statistical power, we chose three TF peak data sets of varying size (4194,
642 11129, and 40052 peaks) and two gene sets of varying size (42 and 471 genes) as
643 our base scenarios. After assigning the peaks to genes, we randomized the genes in
644 bins of locus length to remove all true gene set enrichment signal while keeping
645 locus length association, and then randomly added peaks into the gene set to
646 simulate enrichment. We chose three scenarios of enrichment, each with varying
647 levels (x%=5,10,20, or 30) of enrichment:
648
649 1. CEbias: Enriched to closely satisfy the assumptions of the binary (ChIP-Enrich)
650 model. We added peaks to x% of the remaining genes in the gene set without a peak.

651 This increases the proportion of genes with a peak, without causing a large increase
652 in the mean number of peaks per gene.
653
654 2. PEbias: Enriched to closely satisfy the assumption of the count-based (Poly-
655 Enrich) model. We added a number of peaks, equal to x% of the number of peaks in
656 the gene set, to a fraction of the genes in the gene set. This increases the mean
657 number of peaks per gene, with little effect on the proportion of genes with a peak.
658
659 3. Balanced: We added a number of peaks, equal to x% of the number of peaks in the
660 gene set, into the gene set weighted by gene locus length. This increases both the
661 proportion of genes with a peak and the mean number of peaks per gene by a
662 similar degree.
663
664 **Defining the true positive transcription factor-GO term pairs**
665 For each transcription factor, we identified the gene that codes for it, and then
666 identified every GO biological process that gene is assigned to. This set of GO terms,
667 along with all of its ancestors, is what we use as the true positive set.
668
669 **Hybrid test**
670 Given n tests that test for the same hypothesis, the same Type I error rate, and
671 converted to p-values $p_1, \ldots, p_n$, the Hybrid p-value is computed as:
672 $p_{hybrid} = n \times \min(p_1, \ldots, p_n)$. This hybrid test will have at most the same Type I
673 error rate as the n tests, and if at least one test is consistent (power converges to 1

674 as sample size reaches infinity), the hybrid test will also be consistent. Proofs and
675 simulations of the test in general were done by Zhang et. al [13]. Here, we've
676 implemented the hybrid test for users to use two methods (n = 2): ChIP-Enrich and
677 Poly-Enrich. Users may also choose any two results files and run a hybrid test based
678 on those.
679
680 **Clustering and heatmaps**
681 For every GO term, we calculated the difference in $-\log_{10}$ p-value for each of the 90
682 experiments between ChIP-Enrich and Poly-Enrich, with positive values indicating a
683 more significant result for Poly-Enrich. We then focused on GO terms where > 10%
684 of the experiments had an absolute $\log_{10}$ p-value difference greater than 2.
685 Clustering was performed using uncentered correlation as the similarity metric and
686 average linkage as the clustering method. Using Java TreeView, we extracted specific
687 groups of GO terms that contain certain strings such as "cell cycle" or "positive
688 regul."
689
690 **Repetitive elements**
691 Data was obtained from the UCSC Table Browser with RepeatMasker 3.0 on the
692 hg19 genome. We chose the two most abundant families in the dataset: Alu and L1,
693 as well as four methods of peak-to-gene assignments: Intron, Nearest TSS, >5kb, and
694 <5kb. Poly-Enrich was then used to perform gene set enrichment. Before clustering
695 for the heatmap, we filtered out GO terms where there were 2 or fewer significant
696 FDR values among the 8 categories. The clustering method was the same as

697 mentioned in the previous section.

698

699 **Website and Bioconductor updates**

700 The Chip-Enrich website (http://chip-enrich.med.umich.edu) updated from the
701 *chipenrich* package version 1.7.2 to version 2.5.0. (from
702 https://github.com/sartorlab/chipenrich, on Aug 8th, 2018). We have added the
703 following reference genomes: human (hg38), rat (rn5, rn6), *Drosophilla*
704 *melanogaster* (dm6) and zebrafish (danRer10) species. We also added the following
705 databases from MSigDB (Version 6.0): Hallmark, Immunologic, MicroRNA,
706 Transcription Factors and Oncogenic [6, 29]. Finally, we added sets of genes that are
707 known to be affected by particular environmental toxins from the Comparative
708 Toxicogenomics Database (CTD) [30].

709

710 In addition to the previous locus definitions: 'nearest TSS', 'nearest gene', '≤1 kb
711 from TSS' and '≤5 kb from TSS', we also now support gene locus definitions for
712 regions <10 kb from a TSS and gene distal regions (>10kb upstream of a TSS).

713

714 **Acknowledgements**

715 This work was partially funded by National Institutes of Health grants R01
716 CA158286 and the Michigan Lifestage Environmental Exposures and Disease (M-
717 LEEaD) Center P30 ES017885.

718

Supplementary Figures and Tables

Supplementary Figure 1: Poly-Enrich Type I error rate plots using the <5kb and Nearest TSS gene locus definitions. Shown are the proportion of significant gene sets out of 50,150 randomized gene sets for each of 90 ENCODE ChIP-Seq dataset. Type I error rates are acceptable at the 0.05 and 0.001 level. There is some inflation for the *bylocation* randomization.

Supplementary Figure 2: Statistical power comparisons between ChIP-Enrich (blue) and Poly-Enrich (red) for datasets with three different sizes (i.e. number of peaks: small, medium, and large) and two gene set sizes (small and large GS), under two significance levels: $\alpha = 0.05$ (A) and 0.001 (B), and three different methods of simulated enrichment (CEbias: add peaks according to the regulatory assumptions of ChIP-Enrich, PEbias: add peaks mainly according to the assumptions of Poly-Enrich, Balanced: adding peaks in proportion to each gene's locus length). The values on the X-axis indicate the percent of extra peaks added to simulate enrichment; a higher value simulates stronger enrichment. A stricter significance level results in less power, a larger gene set results in more power, and a larger dataset (more noise) results in less power. In actuality, larger real data sets should have more power.

Supplementary Figure 3: Using the nearest TSS method of assigning peaks to genes, a heatmap of $-\log_{10}$ p-value differences between Poly-Enrich and ChIP-Enrich

741 for GO terms and experiments, where each row is a GO term and each column is a
742 ChIP-seq experiment. Red indicates Poly-Enrich was more significant, and blue
743 indicates ChIP-Enrich was more significant. Light grey indicates the transcription
744 factor used in the experiment was not assigned to the GO term and is omitted in the
745 clustering. Representative GO terms are used to label each cluster.
746
747 **Supplementary Figure 4:** Enrichment results using Poly-Enrich for Alu (first four
748 columns) and L1 (last four columns) repetitive element families using four different
749 peak-to-gene assignments. Shown are signed –log10 FDR, where positive values
750 (red) indicate enrichment and negative values (blue) indicate depletion. GO terms
751 with "neuro" are almost all being significantly depleted.
752
753 **Supplementary Figure 5:** (A) A comparison of the spline estimates with and
754 without the Gene Set (GS) covariate for a gene set with ~1900 genes. There is very
755 little difference despite this gene set being larger than 97% of all gene sets. (B)
756 Signed -log10 P-value comparisons between ChIP-Enrich and its faster counterpart
757 which usees a spline approximation. (C) Signed -log10 P-value comparisons
758 between Poly-Enrich and its faster counterpart which uses a spline approximation.
759 As the gene set specific spline is almost similar to the approximated spline, we see
760 that using the spline approximation does not change the results by much, with an
761 almost perfect identity trend with $r^2 = 1.00$.
762

763 **Supplementary Figure 6:** Signed -$\log_{10}$ p-value comparison plots between ChIP-
764 Enrich (A) and Poly-Enrich (B) vs their Score test counterparts. The gene sets that
765 are enriched have similar significance, but those for the gene sets that are depleted
766 may vary by several magnitudes.
767
768 **Supplementary Table 1:** Complete table of signed –$\log_{10}$ p-values for all
769 combinations of GO terms, locus definitions (peak-to-gene assignments), and type of
770 repeated elements. Column names are in the format of [repeated element]-[locus
771 definition].
772
773 **Supplementary Table 2:** All 90 ENCODE datasets used for gene set enrichment
774 testing. Downloaded from ENCODE Analysis data at UCSC:
775 http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeAwgTfbs
776 Uniform/
777
778
779

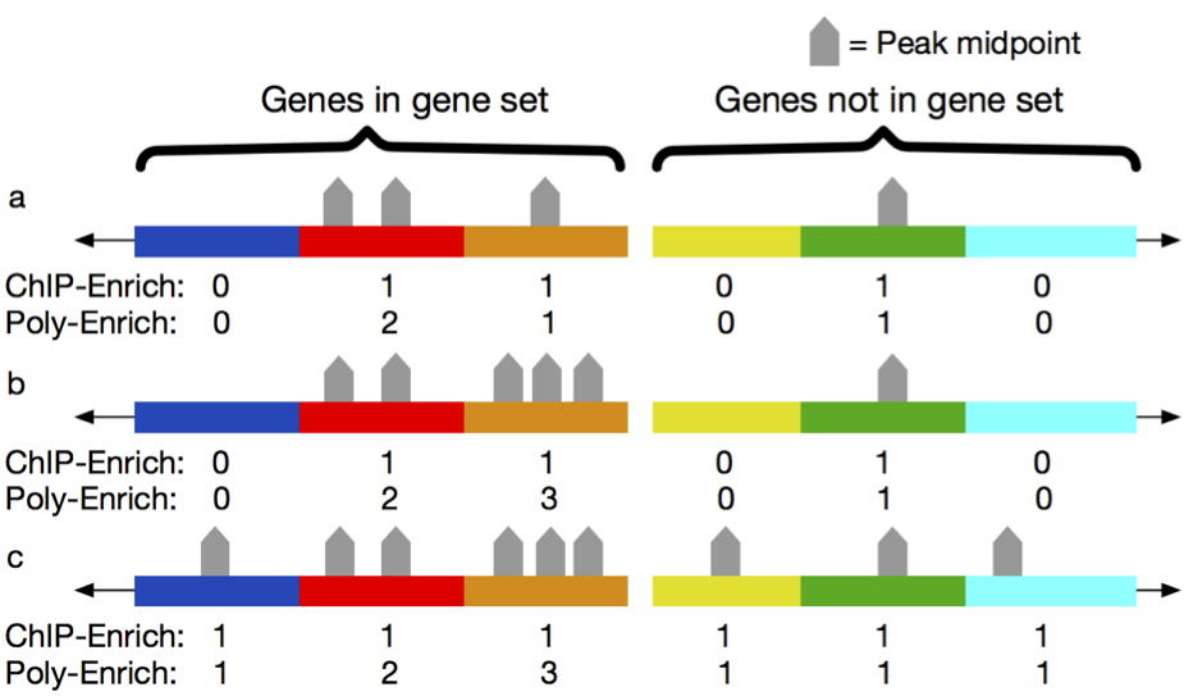
780 # References

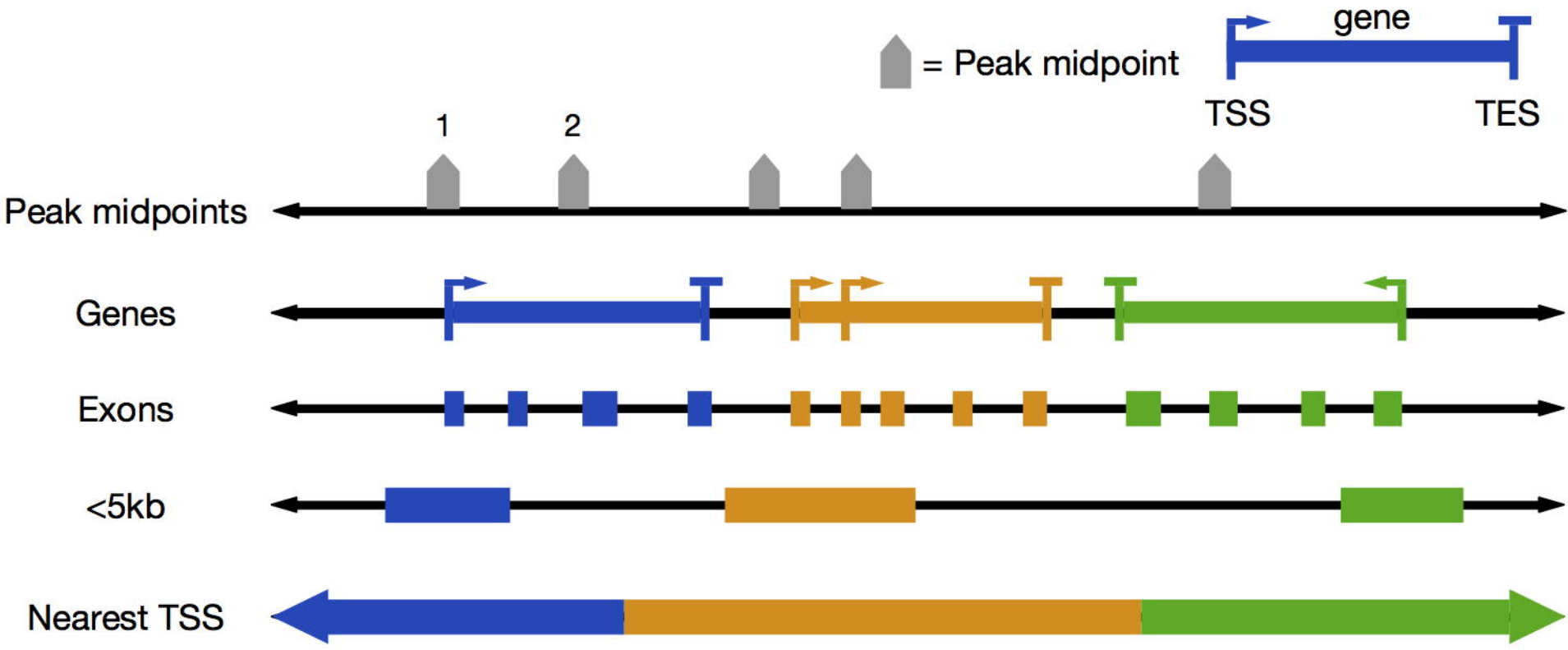
781 1. Gotea V, Visel A, Westlund JM, Nobrega MA, Pennacchio LA, Ovcharenko I.
782 Homotypic clusters of transcription factor binding sites are a key component of
783 human promoters and enhancers. Genome Res. 2010;20(5):565-77. Epub
784 2010/04/02. doi: 10.1101/gr.104471.109. PubMed PMID: 20363979; PubMed
785 Central PMCID: PMCPMC2860159.

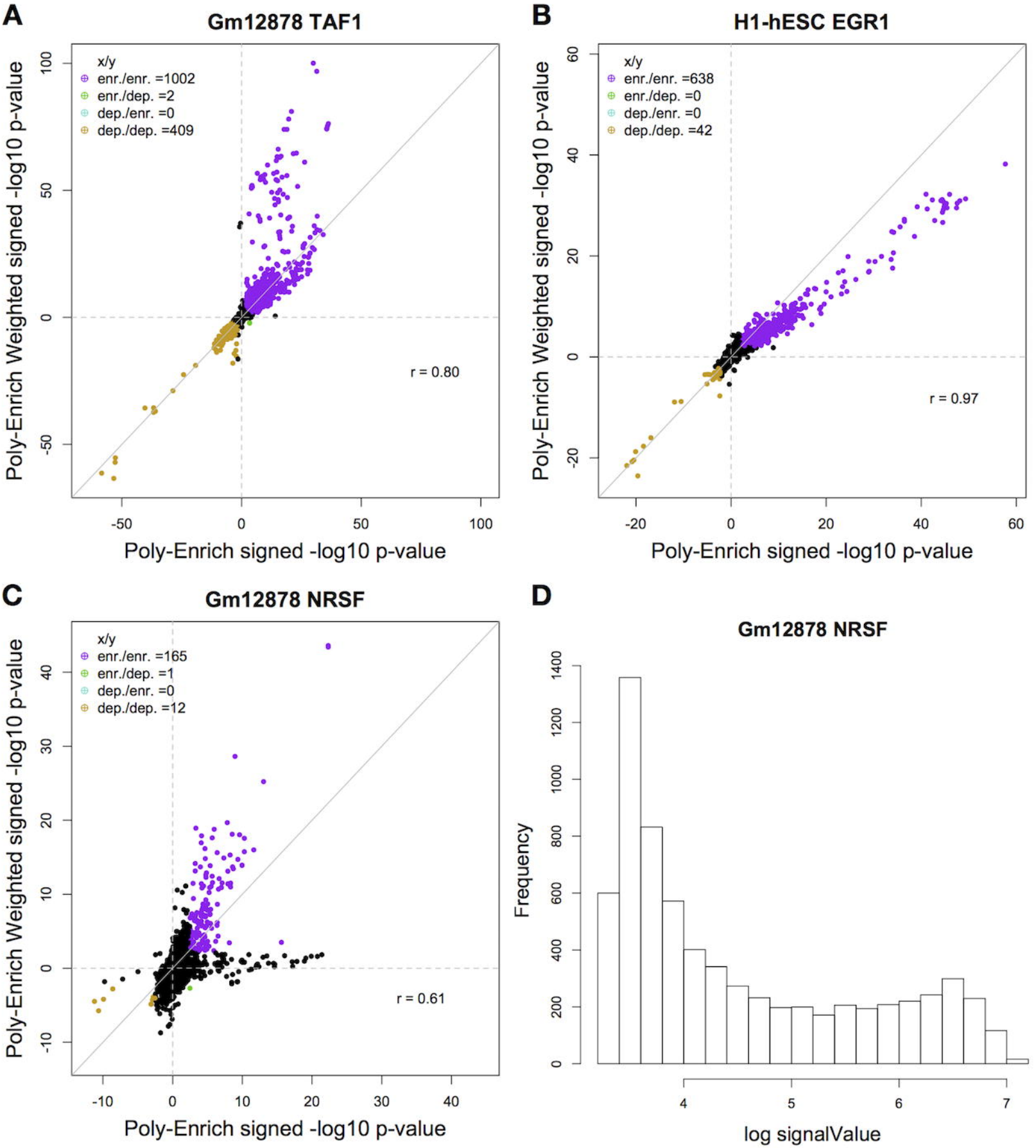
786 2. Pennacchio LA, Bickmore W, Dean A, Nobrega MA, Bejerano G. Enhancers:
787 five essential questions. Nat Rev Genet. 2013;14(4):288-95. doi: 10.1038/nrg3458.
788 PubMed PMID: 23503198; PubMed Central PMCID: PMCPMC4445073.
789 3. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene
790 ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat
791 Genet. 2000;25(1):25-9. doi: 10.1038/75556. PubMed PMID: 10802651; PubMed
792 Central PMCID: PMCPMC3037419.
793 4. Fabregat A, Jupe S, Matthews L, Sidiropoulos K, Gillespie M, Garapati P, et al.
794 The Reactome Pathway Knowledgebase. Nucleic Acids Res. 2018;46(D1):D649-D55.
795 doi: 10.1093/nar/gkx1132. PubMed PMID: 29145629; PubMed Central PMCID:
796 PMCPMC5753187.
797 5. Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: new
798 perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res.
799 2017;45(D1):D353-D61. Epub 2016/11/28. doi: 10.1093/nar/gkw1092. PubMed
800 PMID: 27899662; PubMed Central PMCID: PMCPMC5210567.
801 6. Liberzon A, Subramanian A, Pinchback R, Thorvaldsdóttir H, Tamayo P,
802 Mesirov JP. Molecular signatures database (MSigDB) 3.0. Bioinformatics.
803 2011;27(12):1739-40. Epub 2011/05/05. doi: 10.1093/bioinformatics/btr260.
804 PubMed PMID: 21546393; PubMed Central PMCID: PMCPMC3106198.
805 7. Sartor MA, Leikauf GD, Medvedovic M. LRpath: a logistic regression approach
806 for identifying enriched biological groups in gene expression data. Bioinformatics.
807 2009;25(2):211-7. Epub 2008/11/27. doi: 10.1093/bioinformatics/btn592.
808 PubMed PMID: 19038984; PubMed Central PMCID: PMCPMC2639007.
809 8. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et
810 al. Gene set enrichment analysis: a knowledge-based approach for interpreting
811 genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005;102(43):15545-50.
812 Epub 2005/09/30. doi: 10.1073/pnas.0506580102. PubMed PMID: 16199517;
813 PubMed Central PMCID: PMCPMC1239896.
814 9. McLean CY, Bristor D, Hiller M, Clarke SL, Schaar BT, Lowe CB, et al. GREAT
815 improves functional interpretation of cis-regulatory regions. Nat Biotechnol.
816 2010;28(5):495-501. Epub 2010/05/02. doi: 10.1038/nbt.1630. PubMed PMID:
817 20436461; PubMed Central PMCID: PMCPMC4840234.
818 10. Welch RP, Lee C, Imbriano PM, Patil S, Weymouth TE, Smith RA, et al. ChIP-
819 Enrich: gene set enrichment testing for ChIP-seq data. Nucleic Acids Res.
820 2014;42(13):e105. Epub 2014/05/30. doi: 10.1093/nar/gku463. PubMed PMID:
821 24878920; PubMed Central PMCID: PMCPMC4117744.
822 11. Cavalcante RG, Lee C, Welch RP, Patil S, Weymouth T, Scott LJ, et al. Broad-
823 Enrich: functional interpretation of large sets of broad genomic regions.
824 Bioinformatics. 2014;30(17):i393-400. doi: 10.1093/bioinformatics/btu444.
825 PubMed PMID: 25161225; PubMed Central PMCID: PMCPMC4147897.
826 12. Rosenbloom KR, Dreszer TR, Pheasant M, Barber GP, Meyer LR, Pohl A, et al.
827 ENCODE whole-genome data in the UCSC Genome Browser. Nucleic Acids Res.
828 2010;38(Database issue):D620-5. Epub 2009/11/17. doi: 10.1093/nar/gkp961.
829 PubMed PMID: 19920125; PubMed Central PMCID: PMCPMC2808953.

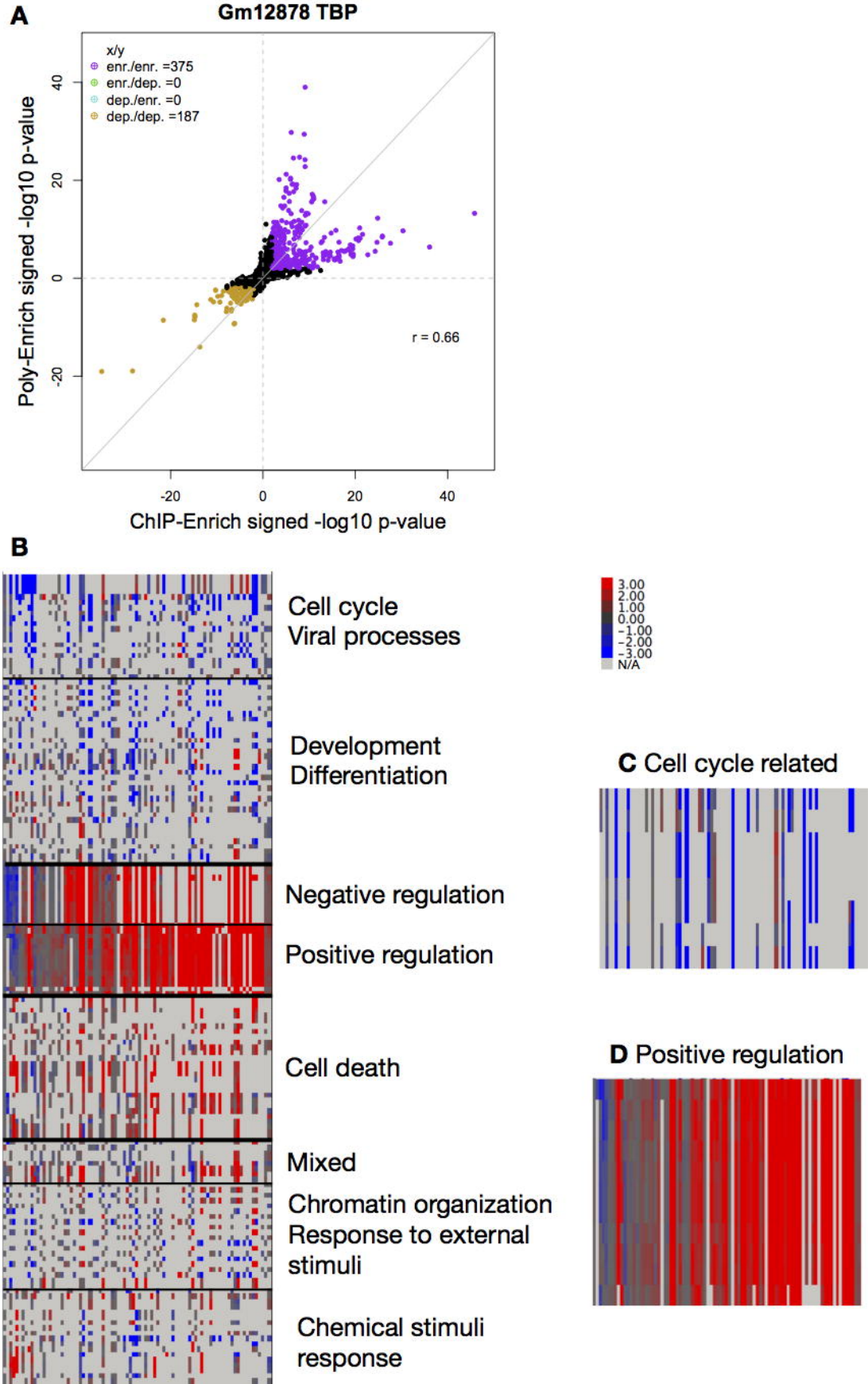
830 13. Zhang S, Okhrin O, Zhou QM, Song PX. Goodness-of-fit test for specification of
831 semiparametric copula dependence models. Journal of Econometrics.
832 2016;193(1):215-33.
833 14. Liu X, Wu B, Szary J, Kofoed EM, Schaufele F. Functional sequestration of
834 transcription factor activity by repetitive DNA. J Biol Chem. 2007;282(29):20868-
835 76. Epub 2007/05/25. doi: 10.1074/jbc.M702547200. PubMed PMID: 17526489;
836 PubMed Central PMCID: PMCPMC3812952.
837 15. Brunner AM, Schimenti JC, Duncan CH. Dual evolutionary modes in the
838 bovine globin locus. Biochemistry. 1986;25(18):5028-35. PubMed PMID: 3768329.
839 16. Tarailo-Graovac M, Chen N. Using RepeatMasker to identify repetitive
840 elements in genomic sequences. Curr Protoc Bioinformatics. 2009;Chapter 4:Unit
841 4.10. doi: 10.1002/0471250953.bi0410s25. PubMed PMID: 19274634.
842 17. Roy-Engel AM, Carroll ML, Vogel E, Garber RK, Nguyen SV, Salem AH, et al.
843 Alu insertion polymorphisms for the study of human genomic diversity. Genetics.
844 2001;159(1):279-90. PubMed PMID: 11560904; PubMed Central PMCID:
845 PMCPMC1461783.
846 18. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial
847 sequencing and analysis of the human genome. Nature. 2001;409(6822):860-921.
848 doi: 10.1038/35057062. PubMed PMID: 11237011.
849 19. Wanichnopparat W, Suwanwongse K, Pin-On P, Aporntewan C, Mutirangura
850 A. Genes associated with the cis-regulatory functions of intragenic LINE-1 elements.
851 BMC Genomics. 2013;14:205. Epub 2013/03/27. doi: 10.1186/1471-2164-14-205.
852 PubMed PMID: 23530910; PubMed Central PMCID: PMCPMC3643820.
853 20. Solyom S, Kazazian HH. Mobile elements in the human genome: implications
854 for disease. Genome Med. 2012;4(2):12. Epub 2012/02/24. doi: 10.1186/gm311.
855 PubMed PMID: 22364178; PubMed Central PMCID: PMCPMC3392758.
856 21. Aldrup-Macdonald ME, Sullivan BA. The past, present, and future of human
857 centromere genomics. Genes (Basel). 2014;5(1):33-50. PubMed PMID: 24683489;
858 PubMed Central PMCID: PMCPMC3966626.
859 22. de Sotero-Caio CG, Cabral-de-Mello DC, Calixto MDS, Valente GT, Martins C,
860 Loreto V, et al. Centromeric enrichment of LINE-1 retrotransposons and its
861 significance for the chromosome evolution of Phyllostomid bats. Chromosome Res.
862 2017;25(3-4):313-25. Epub 2017/09/15. doi: 10.1007/s10577-017-9565-9.
863 PubMed PMID: 28916913.
864 23. Giorgetti L, Siggers T, Tiana G, Caprara G, Notarbartolo S, Corona T, et al.
865 Noncooperative interactions between transcription factors and clustered DNA
866 binding sites enable graded transcriptional responses to environmental inputs. Mol
867 Cell. 2010;37(3):418-28. doi: 10.1016/j.molcel.2010.01.016. PubMed PMID:
868 20159560.
869 24. Thomas CA, Paquola AC, Muotri AR. LINE-1 retrotransposition in the nervous
870 system. Annu Rev Cell Dev Biol. 2012;28:555-73. doi: 10.1146/annurev-cellbio-
871 101011-155822. PubMed PMID: 23057747.
872 25. Chicco D, Bi HS, Reimand J, Hoffman MM. BEHST: genomic set enrichment
873 analysis enhanced through integration of chromatin
874 long-range interactions. bioRxiv. 2019;(168427).
875 doi: https://doi.org/10.1101/168427.

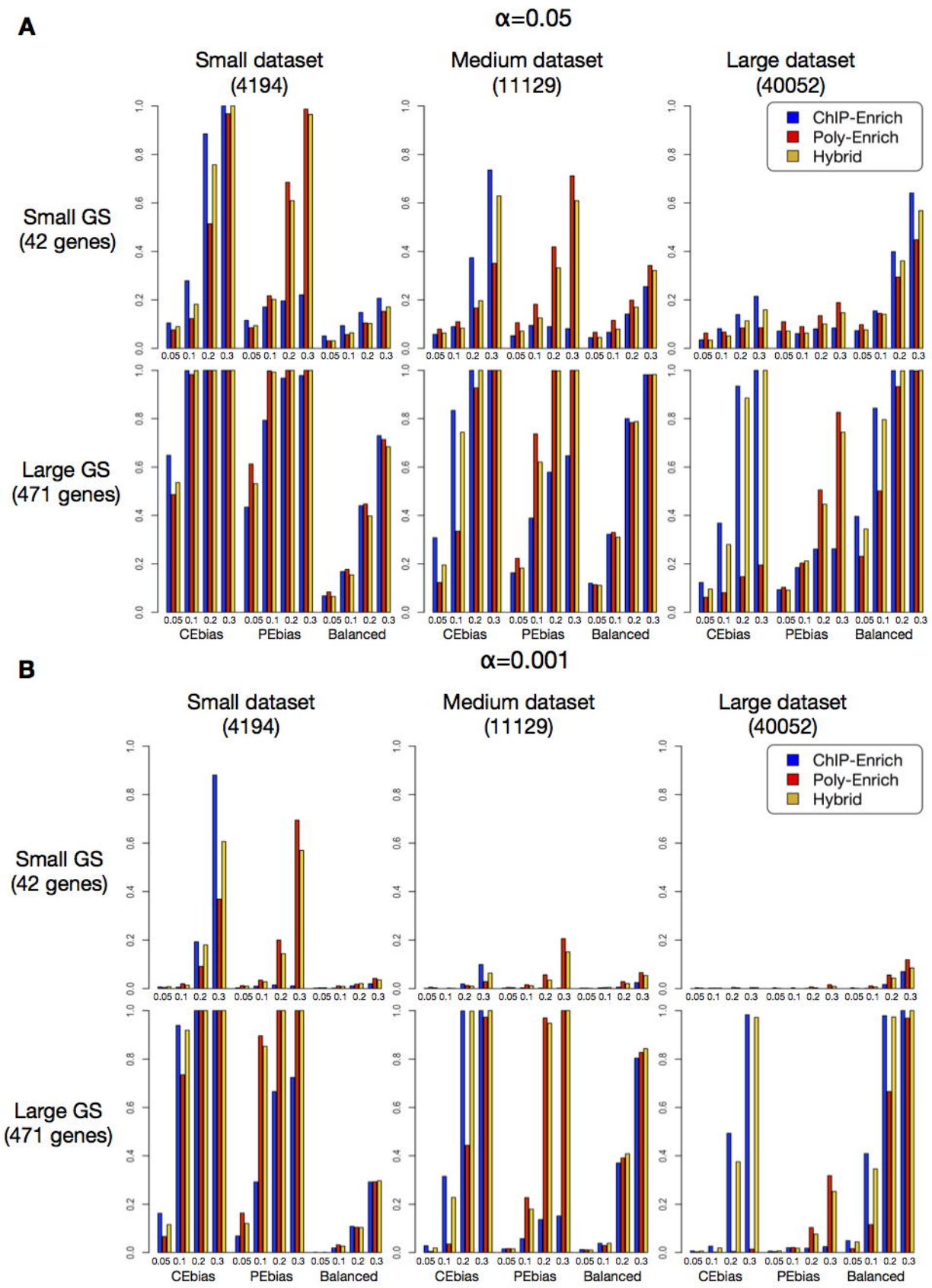
26. Hsu F, Kent WJ, Clawson H, Kuhn RM, Diekhans M, Haussler D. The UCSC Known Genes. Bioinformatics. 2006;22(9):1036-46. Epub 2006/02/24. doi: 10.1093/bioinformatics/btl048. PubMed PMID: 16500937.
27. Wood SN, Goude Y, Shaw S. Generalized additive models for large data sets. Journal of the Royal Statistical Society. 2015;64(1):139-55.
28. Rao CR. Large Sample Tests of Statistical Hypotheses Concerning Several Parameters with Applications to Problems of Estimation. Proceedings of the Cambridge Philosophical Society1948. p. 44, 50-7.
29. Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P. The Molecular Signatures Database (MSigDB) hallmark gene set collection. Cell Syst. 2015;1(6):417-25. doi: 10.1016/j.cels.2015.12.004. PubMed PMID: 26771021; PubMed Central PMCID: PMCPMC4707969.
30. Davis AP, Grondin CJ, Johnson RJ, Sciaky D, King BL, McMorran R, et al. The Comparative Toxicogenomics Database: update 2017. Nucleic Acids Res. 2017;45(D1):D972-D8. Epub 2016/09/19. doi: 10.1093/nar/gkw838. PubMed PMID: 27651457; PubMed Central PMCID: PMCPMC5210612.

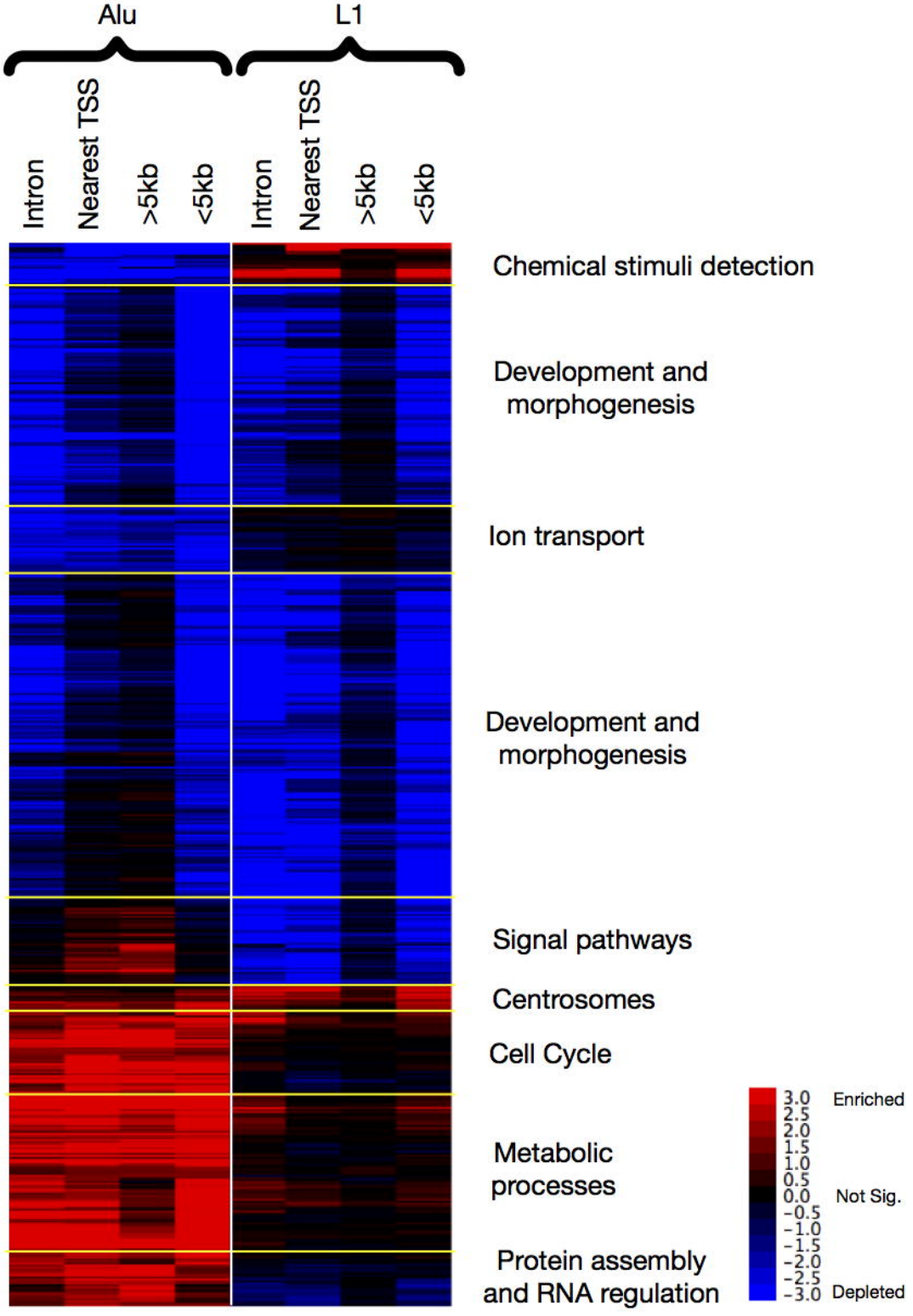

= Peak midpoint
Genes in gene set
Genes not in gene set
a
ChIP-Enrich: 0 1 1 0 1 0
Poly-Enrich: 0 2 1 0 1 0
b
ChIP-Enrich: 0 1 1 0 1 0
Poly-Enrich: 0 2 3 0 1 0
c
ChIP-Enrich: 1 1 1 1 1 1
Poly-Enrich: 1 2 3 1 1 1


= Peak midpoint
gene
TSS
TES
1
2
Peak midpoints
Genes
Exons
<5kb
Nearest TSS


A
Gm12878 TAF1
x/y
enr./enr. =1002
enr./dep. =2
dep./enr. =0
dep./dep. =409
r = 0.80
Poly-Enrich Weighted signed -log10 p-value
Poly-Enrich signed -log10 p-value
B
H1-hESC EGR1
x/y
enr./enr. =638
enr./dep. =0
dep./enr. =0
dep./dep. =42
r = 0.97
Poly-Enrich Weighted signed -log10 p-value
Poly-Enrich signed -log10 p-value
C
Gm12878 NRSF
x/y
enr./enr. =165
enr./dep. =1
dep./enr. =0
dep./dep. =12
r = 0.61
Poly-Enrich Weighted signed -log10 p-value
Poly-Enrich signed -log10 p-value
D
Gm12878 NRSF
Frequency
log signalValue


A
Gm12878 TBP
x/y
enr./enr. =375
enr./dep. =0
dep./enr. =0
dep./dep. =187
r = 0.66
Poly-Enrich signed -log10 p-value
ChIP-Enrich signed -log10 p-value
B
Cell cycle
Viral processes
Development
Differentiation
Negative regulation
Positive regulation
Cell death
Mixed
Chromatin organization
Response to external stimuli
Chemical stimuli response
3.00
2.00
1.00
0.00
−1.00
−2.00
−3.00
N/A
C Cell cycle related
D Positive regulation


A
α=0.05
Small dataset
(4194)
Medium dataset
(11129)
Large dataset
(40052)
ChIP-Enrich
Poly-Enrich
Hybrid
Small GS
(42 genes)
Large GS
(471 genes)
0.0
0.2
0.4
0.6
0.8
1.0
0.05 0.1 0.2 0.3
CEbias
PEbias
Balanced
B
α=0.001
Small dataset
(4194)
Medium dataset
(11129)
Large dataset
(40052)
ChIP-Enrich
Poly-Enrich
Hybrid
Small GS
(42 genes)
Large GS
(471 genes)
CEbias
PEbias
Balanced


Alu
L1
Intron
Nearest TSS
>5kb
<5kb
Intron
Nearest TSS
>5kb
<5kb
Chemical stimuli detection
Development and morphogenesis
Ion transport
Development and morphogenesis
Signal pathways
Centrosomes
Cell Cycle
Metabolic processes
Protein assembly and RNA regulation
3.0
2.5
2.0
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
-2.0
-2.5
-3.0
Enriched
Not Sig.
Depleted