Poly-Enrich: Count-based Methods for Gene Set Enrichment Testing with Genomic Regions and Updates to ChIP-Enrich

Christopher T Lee[1], Raymond G Cavalcante[2,*], Chee Lee[2,^], Tingting Qin[2], Snehal Patil[2], Shuze Wang[2], Zing TY Tsai[§], 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 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.

Introduction

Regulatory genomics facilitates understanding 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 are still open questions and different interpretations may improve results on downstream tests such as gene set enrichment.

Gene Set Enrichment (GSE) is a category of tests that 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 the most widely used gene set databases. Although originally developed for gene expression data, GSE testing is now often used to help interpret ChIP-seq peak sets and other sets of genomic regions. Existing methods for general GSE tests include Fisher's exact test, random sets, logistic regression like LRPath [7], and GSEA-type tests [8]. GSE methods specifically for ChIP-seq data include Genomic Regions Enrichment of Annotations Tool (GREAT) [9], ChIP-Enrich [10], and Broad-Enrich [11]. With so many different tests, one may wonder which test is optimal to use, but there is no single recommendation across data types. Different tests are needed for different types of genomic regions as properties such as peak widths, number of peaks, and location relative to genes all make a difference. Thus GSE testing for genomic regions should not be a one-size-fits-all test; some methods work better than others in specific scenarios. For example, Cavalcante et al. showed that Broad-Enrich is more powerful than ChIP-Enrich for broad regions, but lacks power for narrow regions [11]. As another example, GREAT does not account for variability among genes, so it is best used in situations where the probability of a peak is constant

across genomic space (e.g. per kb), as opposed to clustered near transcription start sites or displaying variability among gene loci.

Our previous method, ChIP-Enrich, uses a binary score to classify a gene as having at least one peak. We saw that ChIP-Enrich tends to underperform when nearly all genes have at least one genomic region associated with them; in this case, ChIP-Enrich will not yield meaningful results. We hypothesized that a count-based method that captures the frequency of binding would perform better in those situations. In this paper, we introduce such a method, Poly-Enrich to expand our available methods to be suitable for any type of narrow genomic regions including those that tend to saturate genes. ChIP-Enrich has the hypothesis that a single binding site is sufficient for regulation, whereas Poly-Enrich assumes that regulation is incremental, i.e. more peaks correspond to stronger or more likely regulation. To identify under which situations one is more appropriate than the other, we performed a comparison of Poly-Enrich and ChIP-Enrich using a set of 90 transcription factor (TF) ChIP-seq datasets from the Encyclopedia of DNA Elements (ENCODE) [12]. We also introduce a hybrid test that combines information from both ChIP-Enrich and Poly-Enrich, which can be used when there is no optimal test for an entire dataset.

To illustrate the usage of Poly-Enrich, we apply it to sets of repetitive elements in the human Alu and LINE1 families, revealing for the first time a comprehensive view of the processes and functions enriched or depleted with these repetitive elements in the human genome. Finally, we describe several updates to our ChIP-Enrich website and *chipenrich* Bioconductor package, including additional methods for assigning genomic regions to target genes, new gene set databases, and more supported species.

Results

Motivation of Poly-Enrich

The motivation for our new methods comes from situations observed with real sets of genomic regions, often with ChIP-seq peak datasets, but also from other sources, such as families of repetitive elements or large sets of DNA polymorphisms such as those different between closely related species or sub-species. Although our original method, ChIP-Enrich, performs extremely well for most transcription factor (TF) ChIP-seq datasets (Figure 1a), but because it uses a simple binary score for each gene, there are some scenarios where this simplification has a significant loss of information. For example, ChIP-Enrich models a gene with many peaks the same as a gene with only one peak, even though gene regulation may be affected by additional peaks (Figure 1b). Alternatively, if nearly every gene is assigned at least one peak, ChIP-Enrich would be unable to distinguish among them and thus unable to detect any gene set enrichment (Figure 1c). We therefore developed Poly-Enrich as a count-based method that addresses both of these scenarios.

Although GREAT is also a count-based gene set enrichment method, Poly-Enrich differs significantly from it in two respects. Firstly, whereas GREAT counts the number of peaks in an entire gene set, Poly-Enrich counts them per gene. By separating counts per gene, we are also able to adjust for each gene's locus length and the variability in peak count across genes, which we previously showed was an important adjustment to control for Type I error [10]. Secondly, the binomial model used by GREAT assumes that the background probability of a peak is constant across the genome. Poly-Enrich uses a more flexible, empirical approach to this that provides for a range of different assumptions about peak distribution.

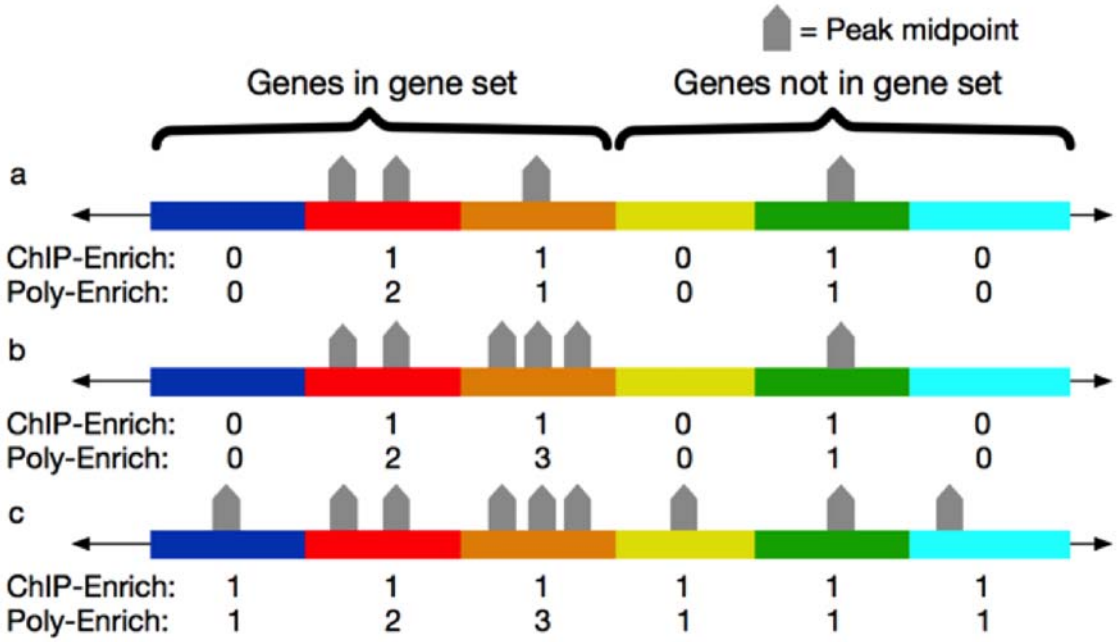

Figure 1: Three scenarios of ChIP-seq peak distributions illustrating how ChIP-Enrich and Poly-Enrich perform. Each color represents a different gene locus; the left three are in a gene set and the right three are not. (a.) Peaks are relatively evenly distributed, with a small number across a subset of genes. Given this situation, ChIP-Enrich evaluates 2/3 vs 1/3 while Poly-Enrich evaluates 2+1 vs 1; both methods perform well. (b.) Some genes contain significantly more peaks than others, such that information is to be gained from the number per gene. ChIP-Enrich evaluates 2/3 vs 1/3, Poly-Enrich evaluates 2+3 vs 1; ChIP-Enrich performs adequately, but Poly-Enrich is optimal. (c.) Nearly all genes have at least one peak, with some having significantly more than others. ChIP-Enrich evaluates 3/3 vs 3/3, Poly-Enrich evaluates 1+2+3 vs 1+1+1; ChIP-Enrich would not detect any enrichment, while Poly-Enrich can still detect gene sets enriched with more peaks.

Assigning peaks to genes

Genomic regions can be assigned to genes in different ways, so that regulation from different types of regions (e.g., promoters, introns, or regions distal to TSSs) can be

studied. For example, ChIP-Enrich, GREAT, and Poly-Enrich all use a peak's midpoint to define the location of the peak. We define a gene's locus definition as the region on the genome such that peaks in that region are assigned to the gene. These loci are defined using properties of the gene, such as within 5kb of a gene's transcription start site (TSS), or simply by assigning each region to the nearest TSS (Figure 2). In the new version of our website and Bioconductor package, we offer several 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.

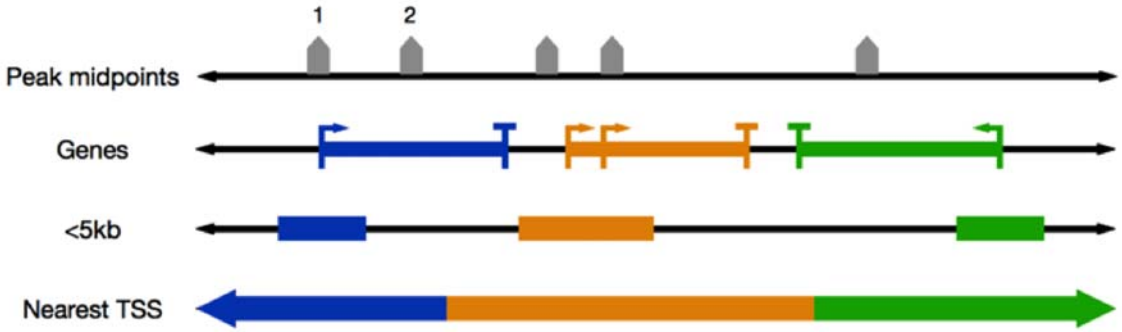

Figure 2: Overview of peak-to-gene assignments given gene locus definitions. Examples shown are: "<5kb", peaks within 5 kb of a gene's TSS are assigned to the gene; "Nearest TSS", peaks are assigned to the gene with the closest TSS. A gene's locus length is defined by the number of base pairs assigned 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 is an indicator of gene set inclusion, 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).

Testing Type 1 error and power

We first tested the type I error rate of the count-based method under the null hypothesis of no enrichment signal. By permuting the genes in the peak-to-gene assignment pairs and breaking the peak-gene relationships, we simulated three scenarios of no enrichment: *i)* the "*complete*" randomization was done by shuffling the gene IDs in the whole dataset; *ii)* the "*bylength*" randomization first groups the

genes together into bins of similar locus length, and then randomizes genes within those bins to preserve the locus length relationship; *iii)* the "*bylocation*" randomization groups genes together by their physical location on the chromosomes, and then randomizes genes within those bins. (See *Methods* for more detail.) We ran the randomizations on our 90 selected ChIP-Seq datasets from ENCODE (see *Methods*), 10 times each, and the proportion of p-values < 0.05 and < 0.001 for each dataset were plotted (Supplementary Fig 1). We see that the test is properly controlled at an acceptable level for Type 1 error in all cases.

To characterize the statistical power of Poly-Enrich under different situations, we permutated data while simulating enrichment of a gene set, and compared the results with those from ChIP-Enrich. We used three datasets with a small, medium, and large number of peaks, and two GO terms with a small and large number of genes. Three types of enrichment were simulated: one that biases towards ChIP-Enrich (CEBias), one that biases towards Poly-Enrich (PEBias), and one that is balanced. For each type of enrichment, we simulated four levels of enrichment: 0.05, 0.1, 0.2, and 0.3, where a higher number indicates a larger simulated enrichment. (See *Methods* for more detail.) Finally, we chose two different levels of significance: $\alpha = 0.05$ and 0.001, as our cutoffs.

As expected, a larger gene set and higher simulated enrichment results in higher power. Simulations on larger datasets artificially reduced power because randomized larger datasets include more noise. However, in real experiments such as in ChIP-seq, we expect larger datasets (more peaks) to be more powerful, since the majority of the peaks in the dataset are not noise given that the experiment can successfully capture the real regulatory regions. Overall, we see that ChIP-Enrich has more power than Poly-Enrich in simulations that enrich a gene set by adding peaks to genes without them, and Poly-Enrich has more power in simulations that enrich a gene set by increasing the number of peaks per gene. Finally, the *Balanced* simulation results in the two methods having similar power in most cases (Figure 3).

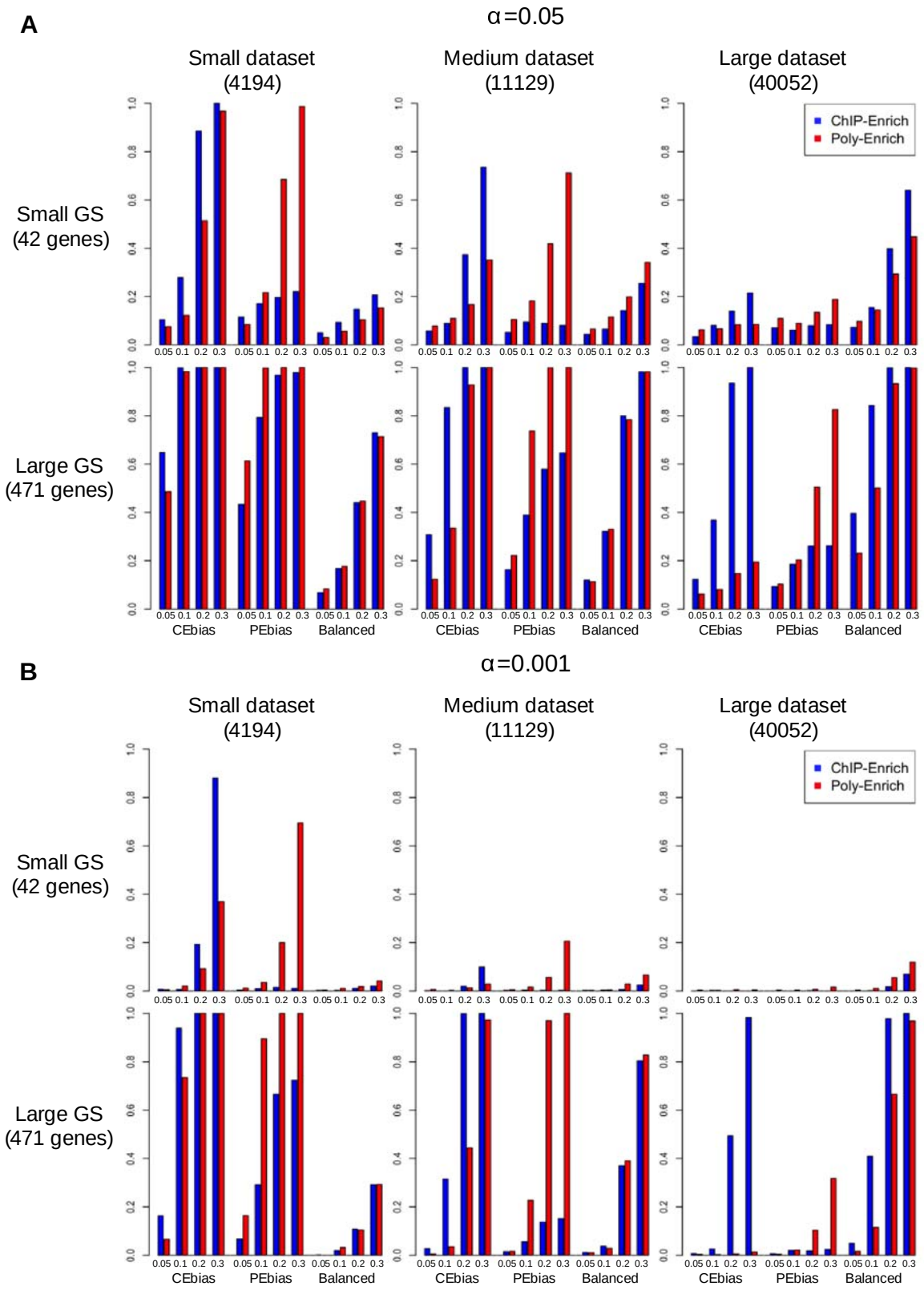

Figure 3: Statistical power comparisons between ChIP-Enrich (blue) and Poly-Enrich (red) for datasets with three different sizes (i.e. number of peaks: small,

medium, and large) and two gene set sizes (small and large GS), under two significance levels: $\alpha = 0.05$ (A) and 0.001 (B). 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.

Poly-Enrich with weighted genomic regions

The height and confidence of peaks in a ChIP-seq experiment can vary dramatically, and we reasoned that incorporating this additional information should improve the ability to pinpoint the truly enriched pathways. Although the most apparent motivation for weighting genomic regions is to account for ChIP-seq peak strength, there are other situations where each peak or genomic region may be assigned a unique score (e.g. confidence or strength). Therefore, we added the option to weight regions by signal value (see *Methods* for details), and examined the extent to which adjusting for peak strength improves enrichment results by comparing the $-\log_{10}$ p-values per gene set. To also ensure that no enrichment result swapped from enriched to depleted or vice versa, we used a signed $-\log_{10}$ p-value, where values for depleted gene sets were negative, and values for enriched gene sets were positive. We noticed for 25% of the experiments, most enriched gene sets were more significant with weighting, thus as we hypothesized, binding events near genes in enriched GO terms were stronger than those near other genes (Figure 4A, 4B).

In another 20% of the experiments, the enrichment p-values were split between the two methods (Figure 4C). Interestingly, the distribution of log signal values for these experiments showed a bimodal pattern (Figure 4D). This suggests that some gene sets tend to have genes with significantly stronger binding peaks than others. For the remaining 55% of experiments tested, weighting made little or no difference on the results.

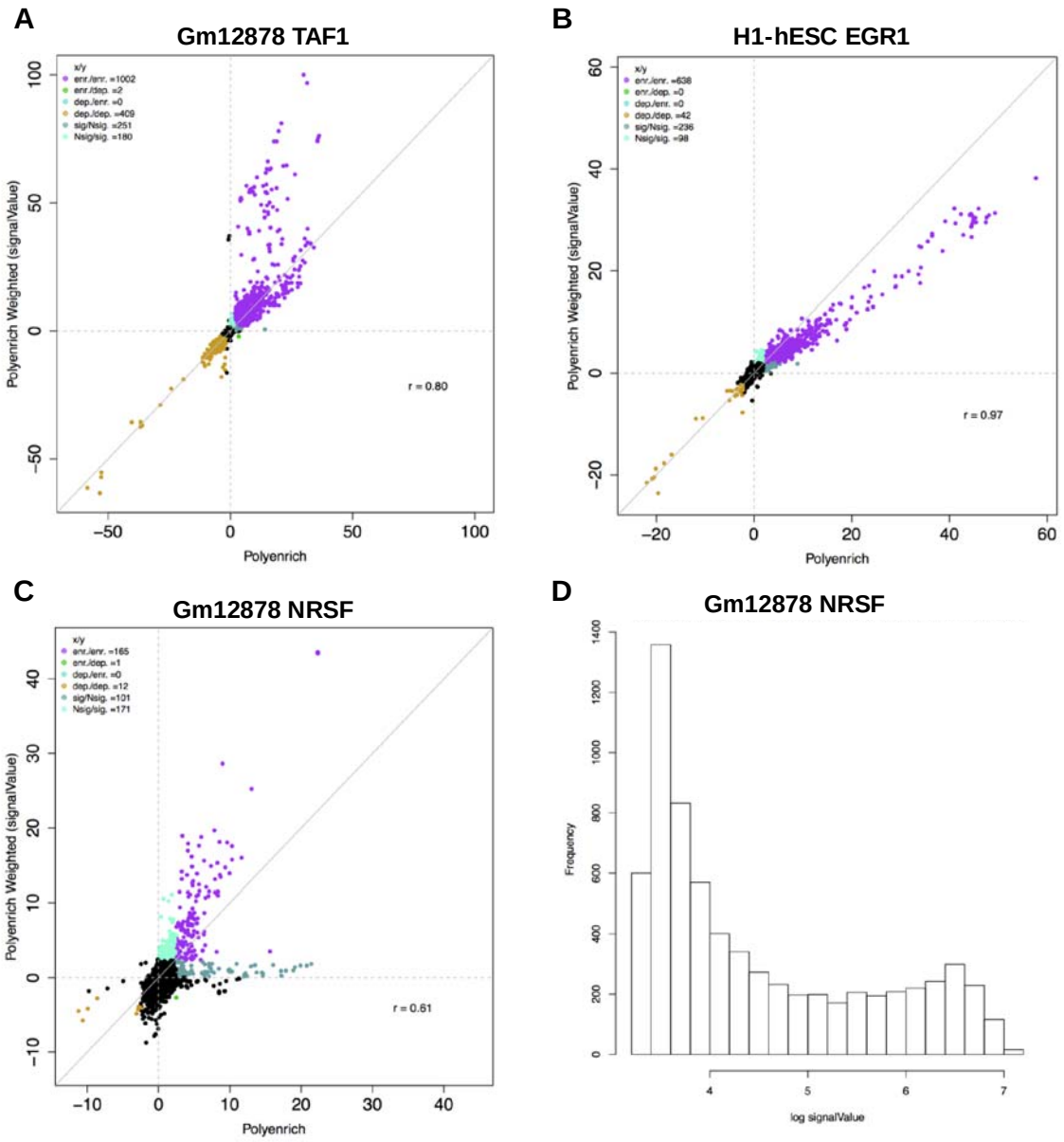

Figure 4: Comparison of GO term enrichment results between standard Poly-Enrich and its weighted version using signal values to weight. Each point is a GO term's $-\log_{10}$ p-value of the two methods, signed positive for enriched, negative for depleted. (A) Using weighting results in more significant enrichment in many GO terms in the Gm12878 TAF1 ChIP-Seq experiment. (B) Using weighting results in less significant enrichment in many GO terms in the H1-hESC EGR1 ChIP-Seq experiment. (C) Using weighting on the Gm12878 NRSF experiment results in several more significant GO terms as well as several less significant ones. (D) The histogram of log signal values from the NRSF experiment. There is a bimodal pattern in the weights, suggesting that there are some GO terms with genes that tend to have stronger or weaker binding.

Since the gene set, rather than the experiment, was a stronger determinant of the more appropriate method, in many cases we are unable to recommend either Poly-Enrich or ChIP-Enrich for an entire experiment; the one exception is that Poly-Enrich is recommended for experiments with a very large number (>100k) of peaks. We therefore developed a hybrid test that uses information from both ChIP-Enrich and Poly-Enrich.

Comparison of the count-based (Poly-Enrich) versus binary (ChIP-Enrich) model of enrichment

Our initial hypothesis was that some experiments would be clearly modeled better by one method or the other (i.e. dependent on the transcription factor). However, the results strongly suggest that the optimal method is more dependent on the gene set than the TF. This can be visualized by a split in the significance levels of GO terms between the binary and count-based methods (Figure 5A), and suggested that a single transcription factor may regulate genes differently depending on the function of the gene. Thus, we sought to understand this further.

The binary model used by ChIP-Enrich assumes that a single binding event (i.e. a single genomic region) is sufficient for regulation, while the Poly-Enrich count-based model assumes that strength of regulation is incremental with the number of binding sites. Based on the results above, we asked what kinds of genes were more consistent with either of those assumptions. To answer this, we first created a set of true positives comprised of GO term-TF pairs by using the GO term biological process (BP) assignments for the gene encoding the transcription factor (e.g. the gene encoding for JunD is assigned to the GO term, "cell death"). This gold standard makes the reasonable assumption that TFs tend to regulate genes in the same biological processes in which they are active. Observing the enrichment results using the 5kb locus definition for these true positive GO term-TF pairs, we used clustering to identify patterns of TFs and GO terms that are optimal with one of the methods. We found that the method that worked better was most often determined by the GO term (Figure 5B). For example, GO terms related to "cell cycle" clustered together and displayed greater power with ChIP-Enrich. Conversely, GO terms involving "negative regulation" tended to do better with Poly-Enrich except for those involving cell cycle (Figure 5C,5D). The results using the Nearest TSS locus definition were similar (Supplementary Figure 2).

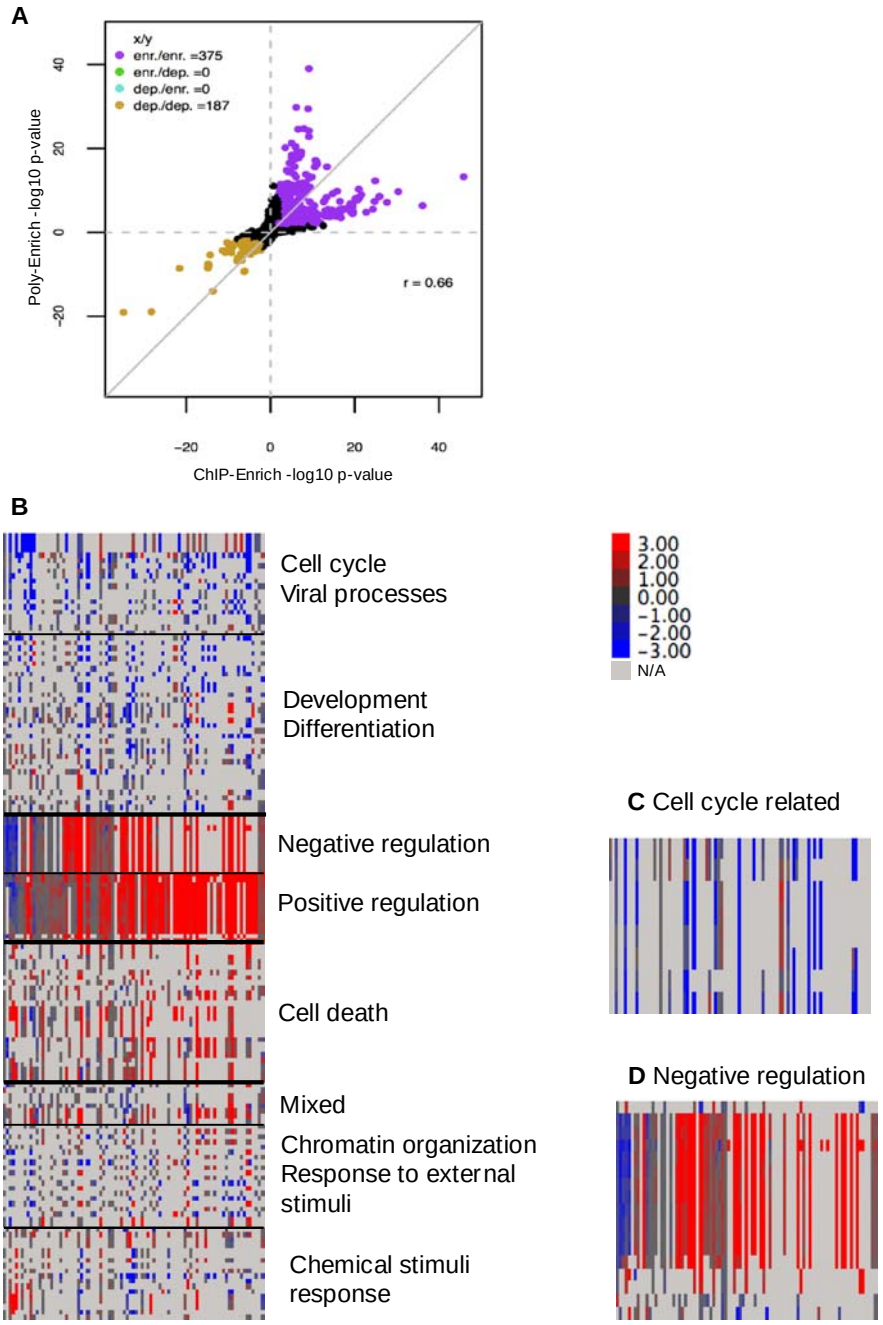

Figure 5: (A) Comparison of GO term significance levels between ChIP-Enrich and Poly-Enrich. Each point is the $-\log_{10}$ p-value of a GO term from the two methods, signed positive for enriched or negative for depleted. Several gene sets are much more significant using ChIP-Enrich and several are much more significant using Poly-Enrich, however 32% of the datasets showed a split pattern like shown. (B) Heatmap of $-\log_{10}$ p-value differences between Poly-Enrich and ChIP-Enrich for GO terms and ChIP-seq experiments, where each row is a GO term and each column is a ChIP-seq experiment. Shown are GO terms where more than 15% of the experiments had a $-\log_{10}$ p-value difference of 2 or larger. Red indicates Poly-Enrich

was more significant, and blue indicates ChIP-Enrich was more significant. Light grey indicates the transcription factor used in the experiment was not assigned to the GO term and is omitted in the clustering. Representative GO terms are shown for each cluster. (C) GO terms related to cell cycle are mostly blue, indicating that a binary score provides a more appropriate model. (D) GO terms containing "negative regulation" are mostly red, indicating that a count score provides a more appropriate model.

Hybrid test

To obtain the best results across all types of GO terms and datasets, we developed a hybrid test that incorporates both the binary and count-based models. After performing the two models, the hybrid p-value of the two tests is defined as: $p_{hybrid} = 2 \times \min(p_{CE}, p_{PE})$ [13], where p_{CE} and p_{PE} are the p-values given by ChIP-Enrich and Poly-Enrich, respectively. This is similar to a Bonferroni-adjusted p-value for two tests. This hybrid has been shown to be beneficial if the two tests are sufficiently different, but loses power and is conservative if the tests are identical or nearly identical [13]. While the hybrid test is not as powerful as the better method between ChIP-Enrich and Poly-Enrich, it is dramatically more powerful than using the worse method, making it the optimal method to use across all GO terms (Figure 6). While this hybrid test only accommodates ChIP and Poly-Enrich, we can extend this to accommodate several additional gene set enrichment tests.

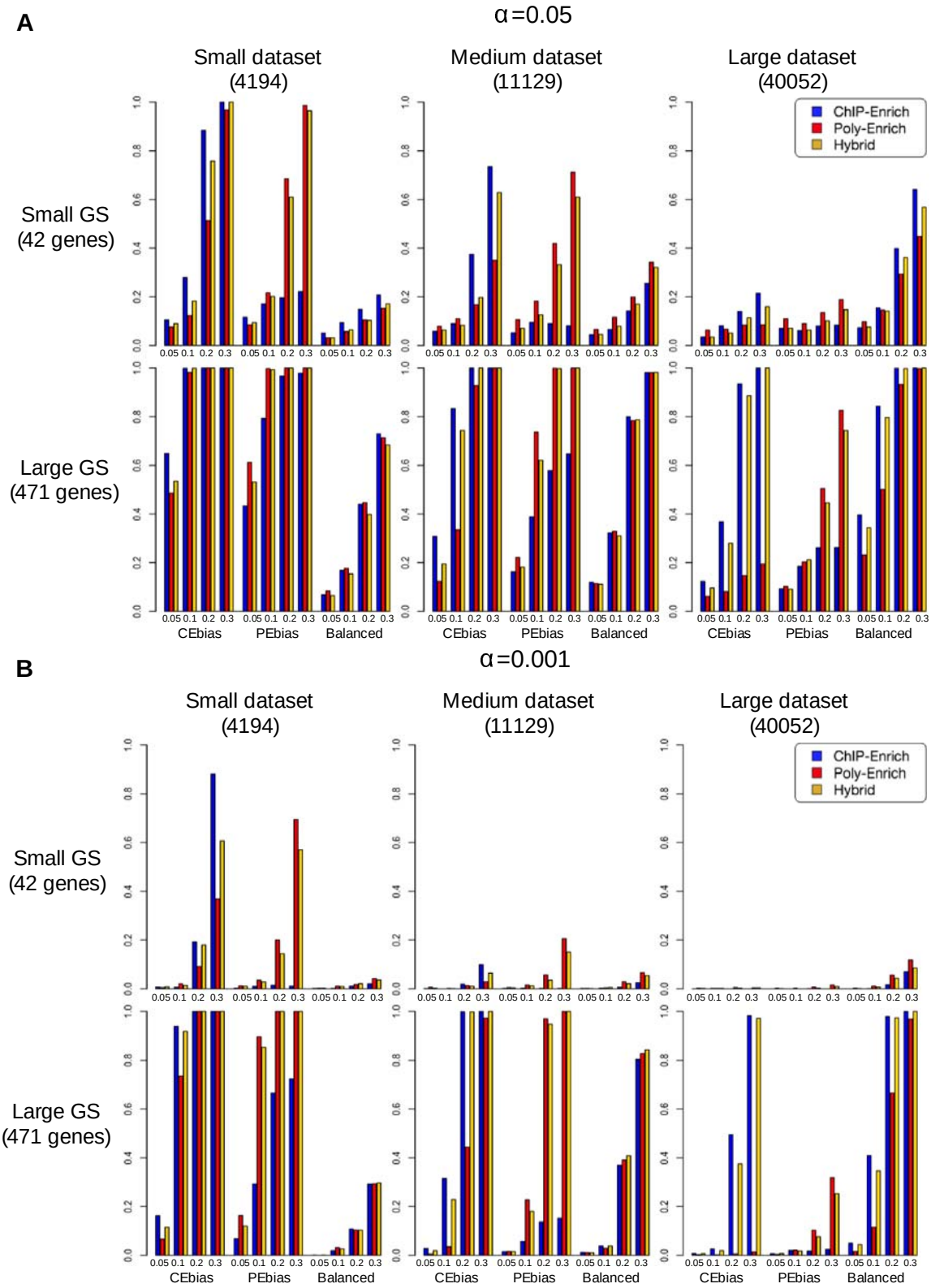

Figure 6: Statistical power comparisons for Poly-Enrich (red), ChIP-Enrich (blue), and the hybrid test (gold) for datasets with three different sizes (i.e. number of peaks: small, medium, and large) and two gene set sizes (small and large GS), under

two significance levels: $\alpha = 0.05$ (A) and 0.001 (B). The values on the X-axis indicate the percent of extra peaks added to simulate enrichment; a higher value simulates stronger enrichment. The hybrid test is shown to have much more power than the wrong method, but only a bit less power than the correct method.

Identifying biological processes enriched with or depleted in repetitive element families using Poly-Enrich

To further illustrate the utility of Poly-Enrich, we used it to test sets of repetitive element regions. We asked whether we could identify gene sets that tended to be either enriched or depleted for certain types of repetitive elements. Significant enrichment of the repetitive elements in the promoter regions of genes, for example, can sequester the transcription factors that will inhibit activities at another transcription factor binding site or other regulatory motif [14]. Some of these mobile elements remain active with new insertions having neutral, detrimental, or beneficial effects. Although repetitive element families have been well studied for over 30 years, little is yet known about the biological processes that they have adapted to help regulate or that they can too easily disrupt and thus are negatively selected against [15]. Using the database of human repetitive elements from the UCSC Table Browser (RepeatMasker 3.0) [16], we performed GSE testing on repetitive element families. Certain families of repetitive elements have over a million occurrences across the human genome, and thus virtually all genes have at least one nearby instance. This is an example where ChIP-Enrich performs poorly, as nearly all genes in all pathways have at least one insertion. Thus, in this situation, modeling the number of insertions per gene is critical to identify differences.

We examined two of the most abundant types of repetitive elements: the *Alu* and LINE1 (L1) elements, which make up an estimated 11% and 17% of the human genome, respectively [17, 18]. We also chose four gene locus definitions: Nearest TSS, <5kb (promoter regions), >5kb (distal regions), and Intron. We tested GO Biological Processes, and used hierarchical clustering of the resulting GSE significance levels to identify related groups of biological processes enriched with or depleted of the repetitive elements (Figure 7). We found strong enrichment of Alu's in GO terms describing metabolic processes, most significantly "ATP metabolic process" and "rRNA metabolic process", especially in promoter regions, which is consistent with an analysis of *Alu* distribution in chromosomes 21 and 22 that showed *Alu* elements on these chromosomes were enriched in or near metabolism and signaling genes [19]. Conversely, Alu elements were sharply depleted in the promoter regions of many development and morphogenesis gene sets, with the strongest depletions in "cell fate commitment" and "connective tissue development". Interestingly, depletions were also seen in the introns of genes in these gene sets, but not in regions >5kb upstream, suggesting the negative selection (depletion) is limited to the regions that are more commonly regulatory.

Novel insertions of L1 elements into or near key genes are known to be associated with neurological diseases [20]. Consistent with this, we found that all neuro-related GO terms in Figure 7 were depleted for L1 (but not for Alu) (Supplementary Figure

3), which suggests that L1's evolutionarily have been selected against occurring in the regulatory regions of neurological genes; when they are inserted into the introns or promoters of these genes, the inserted elements may cause disease.

In general, we observed that the significance of the distal upstream regions (>5kb locus definition) was much lower than the other three locus definitions (with the exception of some enrichments for Alu elements) (Supplemental Table 2), implying that most repetitive element negative (or positive) selection has occurred in the

promoter regions or introns of genes. Alternatively, the gene distal enriched and depleted regions may be limited to a specific set of enhancer regions, the signal from which could have been diluted in our analysis. Interesting additional findings are that L1 elements are enriched in chemical stimulus detection such as "detection of chemical stimulus" and "sensory perception of chemical stimulus", while Alu elements are depleted in the genes in these processes. We also find that both Alu and L1 are enriched in centrosome-related GO terms, which were only made possible with recent advancements in genome mapping near the centromeres [21], and is consistent with previous findings [22] . Additionally, both Alu and L1 elements are significantly depleted in genes in GO terms relating to development and morphogenesis.

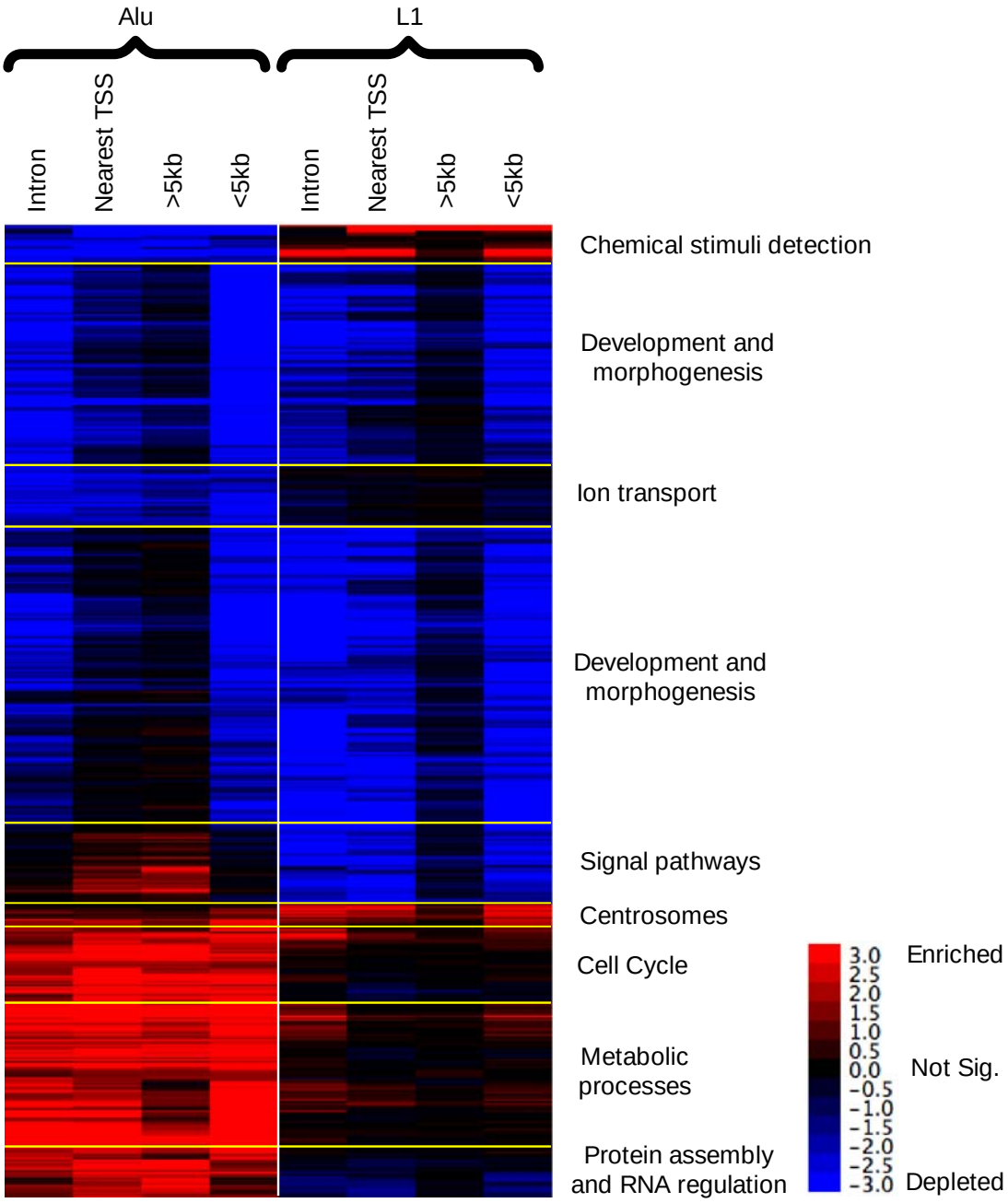

Figure 7: Poly-Enrich results for Alu (first four columns) and L1 (last four columns) repetitive element families using four different gene locus definitions. Shown are signed $-\log_{10}$ FDR, where positive values (red) indicate enrichment and negative values (blue) indicate depletion. Shown are GO terms that were significant for at least 3 columns at the FDR = 0.05 level. We identified nine clusters of GO terms with similar enrichment patterns.

Availability, usage, and updates

Poly-Enrich is available in the *chipenrich* Bioconductor package and as a web interface at http://chip-enrich.med.umich.edu. Several additional gene set

databases and gene locus definitions have been added for the user to choose (see http://chip-enrich.med.umich.edu/data/ChipenrichMethods.pdf).

To perform gene set enrichment analysis, the user first needs a file of genomic regions, which may be a narrowPeak, BED, or text file with chromosome, start, and end positions for each region. The user then selects a species, one or more gene set databases, a gene locus definition, and the test method (ChIP-Enrich, Poly-Enrich, Hybrid, or Fisher's exact test); the gene set and locus definition can be built-in or user-defined. The user can then also choose to add weight based on a peak-specific score, and a number of other options, such as adjustment for read mappability.

The enrichment function outputs five files:

- opts: The options that the user input into the function.
- peaks: A peak-level summary showing the peak-to-gene assignments for each peak.
- peaks-per-gene: A gene-level summary showing gene locus lengths and the number of peaks assigned to them.
- results: The results of the GSE tests. Lists the tested gene sets along with their descriptions, the test effect, odds ratio, enrichment status, p-value, and FDR. Also included is the list of gene IDs with contributing signal for each enrichment test.
- qcplot: A plot of the gene locus lengths with a fitted smoothing spline.

Discussion

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

inverse distance to a gene, or by the probability that the region is in an open chromatin region in a particular cell type.

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

When applying Poly-Enrich to repetitive element families, we both reconfirmed known associations and also identified novel findings. Poly-Enrich confirmed a known fact that Alu elements tend to regulate genes for metabolism and signaling by finding enrichment for related GO terms. Additionally, we know that L1 insertions into or near certain neurological-related genes are associated with neurological diseases [24]. We find that L1 is depleted in neuro-related GO terms, implying there normally are fewer L1 elements in the regulatory regions of these genes, which is consistent with neurological diseases being associated with L1 element insertions near these genes. We also find that there is little enrichment or depletion in the distal regulatory regions of genes, suggesting that repetitive elements may not have as large of an affect there due to mitigated regulatory activity at larger distances from transcription start sites. Poly-Enrich also detected some novel associations between repetitive element families and biological pathways. Both Alu and L1 elements were significantly depleted in genes in GO terms relating to development and morphogenesis, such as “connective tissue development” and “skeletal system morphogenesis”, suggesting that it is especially critical to have developmental regulatory regions free from potentially disruptive repetitive elements during early growth.

One shortcoming of our current methods (as well as current alternatives) is that they rely on associating each genomic region with the nearest gene. However, it is estimated that 79-95% of DNAse I hypersensitive sites, markers for enhancer regions, actually regulate a different, distal target gene [24, 25]. We are currently developing a set of enhancer locus definitions that identify and assign enhancer regions to their appropriate target genes, so peaks in enhancer regions will be correctly assigned and false positive peaks in nonfunctional intergenic regions will be filtered out. We believe this will improve all future gene enrichment analyses.

Methods

Datasets:
All ChIP-Seq data were obtained from Encyclopedia of DNA Elements (ENCODE) at University of Califonia, Santa Cruz [12]. We chose a total of 90 experiments over the three Tier 1 cell lines (Gm12878, H1hesc, and K562), and all 35 transcription

factors that had available ChIP-seq data for at least two of the three Tier 1 cell lines. (Supplementary Table 1)

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

Assigning regions to genes

The UCSC knownGene database for hg19 was used to define the transcription start sites across the genome [25]. Each locus definition (e.g. nearest TSS, <5kb) was generated as a table containing the columns: chromosome, Start, End, gene ID. All genomic regions whose midpoint was between a gene's start and end values would be assigned to that gene. It is possible for some locus definitions to have many disjoint regions for a certain gene.

Poly-Enrich model: a GLM with a Negative Binomial family

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

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

where for each gene i, GS is an indicator for whether the gene is in the gene set of interest or not (=1 if in the gene set; 0 otherwise), μ is the mean of the negative binomial distribution for the number of genomic regions assigned to each gene, and the overdispersion parameter θ is estimated so that $Var(Y|GS) = \mu + \theta\mu^2$, where Y is the number of genomic regions for the gene. The function f is a negative binomial cubic smoothing spline that adjusts for the gene's locus length and optionally adjusts for m, the mappability of the gene's locus. Details about mappability can be found in the ChIP-Enrich manuscript [10]. We use the *gam* function in the *mgcv* R package to fit the model, which uses a penalized likelihood maximization, and the smoothing spline penalty is a squared second derivative penalty [26].

A Wald test on the coefficient for the gene set is used: the test statistic is defined as $W^2 = \widehat{\beta_1}/\hat{V}(\widehat{\beta_1})$, which follows a χ_1^2 distribution under the null hypothesis that there is no association between gene set and number of peaks.

Poly-Enrich with weighting based on genomic region scores

In certain cases, each genomic region in a dataset may be associated with a numeric score. For example, ChIP-seq results often include a value denoting the strength of a peak, (e.g. signalValue in ENCODE dataset results or -10*$\log_{10}$(p-value) in MACS2 results). Poly-Enrich weights based on these scores by giving each peak a weight proportional to its signal value (or other score) and normalizing such that the mean of all peak weights is equal to 1. For every peak assigned to a gene, we then sum all the weights and substitute the weighted sum in place of the original number of peaks. The same model is used, except assuming a quasi-negative binomial family to

accommodate for non-whole number data. The calculations can be carried out identically to the standard negative binomial family.

Comparing p-values between methods

To compare p-values between methods, we use a scatterplot, plotting a signed $-\log_{10}$ p-value per gene set. If a gene set is enriched, the sign is positive, and if the gene is depleted, the sign is negative. This allows us to detect if there are any cases where two methods may contradict each other's conclusions.

Spline approximation

With a library of over 20,000 genes and most gene sets being less than 1000 genes, the cubic smoothing spline estimate changes very little between gene sets. Thus we can reasonably assume that the spline is approximately equal for any gene set of interest, including the spline with no gene set (Supplementary Figure 4A).

We first run the same model except without the gene set (*GS*) term: $\log(\mu_i) = \beta_0 + f(LL_i)$. We then extract the fitted spline using the *predict* function with *type="terms"* from the *mgcv* package to obtain a spline-adjusted locus length for each gene. This new value is then input in the model for every gene set, which allows us to fit a spline only once instead of once for each gene set. This saves a significant amount of time when testing a large number of gene sets (approximately 75% time saved when testing 4000 gene sets). Compared to the original model, we find that the p-values from the spline approximation model are nearly identical (Supplementary Figure 4B, 4C).

Score test

One of the alternatives for the Wald test is the Score test [27]. We can calculate the score test statistic for ChIP-Enrich as:

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

where for each gene i, *GS* is an indicator for whether the gene is in the gene set of interest or not, W_i is an indicator for whether the gene has at least one genomic region (=1 if true), and $\hat{\pi}$ is the predicted probability of the gene having at least one genomic region obtained using the *fitted* function with *type="terms"* in the *mgcv* package.

The score test statistic for Poly-Enrich is:

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

where for each gene i, *GS* is an indicator for whether the gene is in the gene set of interest or not, Y_i is the number of genomic regions assigned to the gene, $\hat{\mu}$ is the

predicted number of genomic regions assigned to the gene obtained using the *fitted* function with *type="terms"* in the *mgcv* package, and its variance is estimated empirically, assuming $GS_i = 0$ in the variance calculation.

The advantage of using the score test is that all the required parameters are already estimated during the spline approximation, so all subsequent calculations will no longer require fitting of a GLM. This reduces the runtime of the enrichments significantly further (approximately 95% time saved when testing 4000 gene sets). However, there are some scenarios where the Score test differs from the Wald test by a substantial amount, mostly notably for depleted gene sets, so the default option is set to the Wald test (Supplementary Figure 5). However, if results are desired quickly, the Score test is offered (method="chipapprox" or "polyapprox") and can serve as a convenient approximation.

Testing Type I error

The null hypothesis of Poly-Enrich is that there is no true biological enrichment. To test the Type-I error, we randomly permute the genes to simulate a scenario where there is no association between genes and the number of peaks. However, to ensure that the results are not biased by gene locus length or gene location, we performed two additional permutations: one permutes genes within bins of similar locus length, and one permutes within bins of chromosomal locations. In both cases, the genes are sorted by the variable of interest (locus length or location), and then assigned to consecutive bins of 100 genes each.

For each of the 90 TF peak data sets chosen, after assigning the peaks to genes, we permute the gene IDs using the randomization of interest, and then perform enrichment tests against GO biological processes. We ran a total of 10 trials and took the median p-value per gene set as the randomization p-value. Then, the proportion of p-values less than a defined confidence level was determined per experiment to calculate the overall Type I error. We then plotted all 90 overall Type I errors for each experiment in a box plot to convey overall Type I error.

Testing power

To test statistical power, we chose three TF peak data sets of varying size (4194, 11129, 40052 peaks) and two gene sets of varying size (42 and 471 genes) as our base scenarios. After assigning the peaks to genes, we randomized the genes in bins of locus length to remove all true gene set enrichment signal while keeping locus length association, and then randomly added peaks into the gene set to simulate enrichment. We chose three scenarios of enrichment, each with varying levels (x%) of enrichment:

1. CEbias: Enriched to closely satisfy the assumptions of the binary (ChIP-Enrich) model. We added peaks to x% of the remaining genes in the gene set without a peak. This increases the proportion of genes with a peak, without causing a large increase in the mean number of peaks per gene.

2. PEbias: Enriched to closely satisfy the assumption of the count-based (Poly-Enrich) model. We added a number of peaks, equal to *x*% of the number of peaks in the gene set, to a fraction of the genes in the gene set. This increases the mean number of peaks per gene, with little effect on the proportion of genes with a peak.

3. Balanced: We added a number of peaks, equal to *x*% of the number of peaks in the gene set, into the gene set weighted by gene locus length. This increases both the proportion of genes with a peak and the mean number of peaks per gene by a similar degree.

Defining the true positive transcription factor-GO term pairs

For each transcription factor, we identified the gene that codes for it, and then identified every GO biological process that gene is assigned to. This set of GO terms, along with all of its ancestors, is what we use as the true positive set. .

Hybrid test

Given n tests that test for the same hypothesis, the same Type I error rate, and converted to p-values $p_1, \ldots, p_n$, the Hybrid p-value is computed as: $p_{hybrid} = n \times \min(p_1, \ldots, p_n)$. This hybrid test will have at most the same Type I error rate as the n tests, and if at least one test is consistent (power converges to 1 as sample size reaches infinity), the hybrid test will also be consistent. Proofs and simulations of the test in general were done by Zhang et. al [13]. Here, we've implemented the hybrid test for users to use with two methods (n = 2): ChIP-Enrich and Poly-Enrich. Users may also choose any two results files and

Clustering and heatmaps

For every GO term, we calculated the difference in $-\log_{10}$ p-value for each of the 90 experiments between ChIP-Enrich and Poly-Enrich, with positive values indicating a more significant result for Poly-Enrich. We then focused on GO terms where > 10% of the experiments had an absolute $\log_{10}$ p-value difference greater than 2. Clustering was performed using uncentered correlation as the similarity metric and average linkage as the clustering method. Using Java TreeView, we extracted specific groups of GO terms that contain certain strings such as "cell cycle" or "positive regul."

Repetitive elements

Data was obtained from the UCSC Table Browser with RepeatMasker 3.0 on the hg19 genome. We chose the two most abundant families in the dataset: Alu and L1, as well as four gene locus definitions: Intron, Nearest TSS, >5kb, and <5kb. Poly-Enrich was then used to perform gene set enrichment. Before clustering for the heatmap, we filtered out GO terms where there were 2 or fewer significant FDR values among the 8 categories. The clustering method was the same as mentioned above.

Website and Bioconductor updates

The Chip-Enrich website (http://chip-enrich.med.umich.edu) updated the *chipenrich* package version from 1.7.2 version to 2.5.0. (from https://github.com/sartorlab/chipenrich, on Aug 8th, 2018). We have added the following reference genomes: human (hg38), rat (rn5,rn6), Drosophilla melanogaster (dm6) and zebrafish (danRer10) species.
We also added the following databases from MSigDB (Version 6.0): Hallmark, Immunologic, MicroRNA, Transcription Factors and Oncogenic [6, 28]. We also added sets of genes that are known to be affected by particular environmental toxins from Comparative Toxicogenomics Database (CTD).

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

Acknowledgements
This work was partially funded by National Institutes of Health grants R01 CA158286 and P30 ES017885.

References

1. Gotea V, Visel A, Westlund JM, Nobrega MA, Pennacchio LA, Ovcharenko I. Homotypic clusters of transcription factor binding sites are a key component of human promoters and enhancers. Genome Res. 2010;20(5):565-77. Epub 2010/04/02. doi: 10.1101/gr.104471.109. PubMed PMID: 20363979; PubMed Central PMCID: PMCPMC2860159.
2. Pennacchio LA, Bickmore W, Dean A, Nobrega MA, Bejerano G. Enhancers: five essential questions. Nat Rev Genet. 2013;14(4):288-95. doi: 10.1038/nrg3458. PubMed PMID: 23503198; PubMed Central PMCID: PMCPMC4445073.
3. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000;25(1):25-9. doi: 10.1038/75556. PubMed PMID: 10802651; PubMed Central PMCID: PMCPMC3037419.
4. Fabregat A, Jupe S, Matthews L, Sidiropoulos K, Gillespie M, Garapati P, et al. The Reactome Pathway Knowledgebase. Nucleic Acids Res. 2018;46(D1):D649-D55. doi: 10.1093/nar/gkx1132. PubMed PMID: 29145629; PubMed Central PMCID: PMCPMC5753187.
5. Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res. 2017;45(D1):D353-D61. Epub 2016/11/28. doi: 10.1093/nar/gkw1092. PubMed PMID: 27899662; PubMed Central PMCID: PMCPMC5210567.
6. Liberzon A, Subramanian A, Pinchback R, Thorvaldsdóttir H, Tamayo P, Mesirov JP. Molecular signatures database (MSigDB) 3.0. Bioinformatics. 2011;27(12):1739-40. Epub 2011/05/05. doi: 10.1093/bioinformatics/btr260. PubMed PMID: 21546393; PubMed Central PMCID: PMCPMC3106198.
7. Sartor MA, Leikauf GD, Medvedovic M. LRpath: a logistic regression approach for identifying enriched biological groups in gene expression data. Bioinformatics.

2009;25(2):211-7. Epub 2008/11/27. doi: 10.1093/bioinformatics/btn592. PubMed PMID: 19038984; PubMed Central PMCID: PMCPMC2639007.
8. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005;102(43):15545-50. Epub 2005/09/30. doi: 10.1073/pnas.0506580102. PubMed PMID: 16199517; PubMed Central PMCID: PMCPMC1239896.
9. McLean CY, Bristor D, Hiller M, Clarke SL, Schaar BT, Lowe CB, et al. GREAT improves functional interpretation of cis-regulatory regions. Nat Biotechnol. 2010;28(5):495-501. Epub 2010/05/02. doi: 10.1038/nbt.1630. PubMed PMID: 20436461; PubMed Central PMCID: PMCPMC4840234.
10. Welch RP, Lee C, Imbriano PM, Patil S, Weymouth TE, Smith RA, et al. ChIP-Enrich: gene set enrichment testing for ChIP-seq data. Nucleic Acids Res. 2014;42(13):e105. Epub 2014/05/30. doi: 10.1093/nar/gku463. PubMed PMID: 24878920; PubMed Central PMCID: PMCPMC4117744.
11. Cavalcante RG, Lee C, Welch RP, Patil S, Weymouth T, Scott LJ, et al. Broad-Enrich: functional interpretation of large sets of broad genomic regions. Bioinformatics. 2014;30(17):i393-400. doi: 10.1093/bioinformatics/btu444. PubMed PMID: 25161225; PubMed Central PMCID: PMCPMC4147897.
12. Rosenbloom KR, Dreszer TR, Pheasant M, Barber GP, Meyer LR, Pohl A, et al. ENCODE whole-genome data in the UCSC Genome Browser. Nucleic Acids Res. 2010;38(Database issue):D620-5. Epub 2009/11/17. doi: 10.1093/nar/gkp961. PubMed PMID: 19920125; PubMed Central PMCID: PMCPMC2808953.
13. Zhang S, Okhrin O, Zhou QM, Song PX. Goodness-of-fit test for specification of semiparametric copula dependence models. Journal of Econometrics. 2016;193(1):215-33.
14. Liu X, Wu B, Szary J, Kofoed EM, Schaufele F. Functional sequestration of transcription factor activity by repetitive DNA. J Biol Chem. 2007;282(29):20868-76. Epub 2007/05/25. doi: 10.1074/jbc.M702547200. PubMed PMID: 17526489; PubMed Central PMCID: PMCPMC3812952.
15. Brunner AM, Schimenti JC, Duncan CH. Dual evolutionary modes in the bovine globin locus. Biochemistry. 1986;25(18):5028-35. PubMed PMID: 3768329.
16. Tarailo-Graovac M, Chen N. Using RepeatMasker to identify repetitive elements in genomic sequences. Curr Protoc Bioinformatics. 2009;Chapter 4:Unit 4.10. doi: 10.1002/0471250953.bi0410s25. PubMed PMID: 19274634.
17. Roy-Engel AM, Carroll ML, Vogel E, Garber RK, Nguyen SV, Salem AH, et al. Alu insertion polymorphisms for the study of human genomic diversity. Genetics. 2001;159(1):279-90. PubMed PMID: 11560904; PubMed Central PMCID: PMCPMC1461783.
18. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. Nature. 2001;409(6822):860-921. doi: 10.1038/35057062. PubMed PMID: 11237011.
19. Wanichnopparat W, Suwanwongse K, Pin-On P, Aporntewan C, Mutirangura A. Genes associated with the cis-regulatory functions of intragenic LINE-1 elements. BMC Genomics. 2013;14:205. Epub 2013/03/27. doi: 10.1186/1471-2164-14-205. PubMed PMID: 23530910; PubMed Central PMCID: PMCPMC3643820.

20. Solyom S, Kazazian HH. Mobile elements in the human genome: implications for disease. Genome Med. 2012;4(2):12. Epub 2012/02/24. doi: 10.1186/gm311. PubMed PMID: 22364178; PubMed Central PMCID: PMCPMC3392758.
21. Aldrup-Macdonald ME, Sullivan BA. The past, present, and future of human centromere genomics. Genes (Basel). 2014;5(1):33-50. PubMed PMID: 24683489; PubMed Central PMCID: PMCPMC3966626.
22. de Sotero-Caio CG, Cabral-de-Mello DC, Calixto MDS, Valente GT, Martins C, Loreto V, et al. Centromeric enrichment of LINE-1 retrotransposons and its significance for the chromosome evolution of Phyllostomid bats. Chromosome Res. 2017;25(3-4):313-25. Epub 2017/09/15. doi: 10.1007/s10577-017-9565-9. PubMed PMID: 28916913.
23. Giorgetti L, Siggers T, Tiana G, Caprara G, Notarbartolo S, Corona T, et al. Noncooperative interactions between transcription factors and clustered DNA binding sites enable graded transcriptional responses to environmental inputs. Mol Cell. 2010;37(3):418-28. doi: 10.1016/j.molcel.2010.01.016. PubMed PMID: 20159560.
24. Thomas CA, Paquola AC, Muotri AR. LINE-1 retrotransposition in the nervous system. Annu Rev Cell Dev Biol. 2012;28:555-73. doi: 10.1146/annurev-cellbio-101011-155822. PubMed PMID: 23057747.
25. 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.
26. Wood SN, Goude Y, Shaw S. Generalized additive models for large data sets. Journal of the Royal Statistical Society. 2015;64(1):139-55.
27. 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.
28. 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.