Fast Annotation-Agnostic Differential Expression Analysis

Leonardo Collado-Torres[1, 2], Alyssa C. Frazee[1], Michael I. Love[3], Rafael A. Irizarry[3], Andrew E. Jaffe[2, 1, 4], Jeffrey T. Leek[1, 4]

[1]Department of Biostatistics, The Johns Hopkins University Bloomberg School of Public Health,
[2]Lieber Institute for Brain Development, Johns Hopkins Medical Campus,
[3]Department of Biostatistics, Dana-Farber Cancer Institute and Harvard School of Public Health,
[4]Center for Computational Biology, McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine

Introduction

Since the development of high-throughput technologies for probing the genome researchers have been interested in finding differences across groups that could potentially explain the observable phenotypic differences. The first step being developing methods for large-scale hypothesis generation. The traditional tools have focused on the known transcriptome and are highly dependent on existing annotation. Frazee et al (Biostatistics 2014) developed a statistical framework to find candidate Differentially Expressed Regions (DERs) without relying on annotation that produced sensible results. We have implemented a faster version of this approach in order to handle larger data sets: up to a few hundred samples. The software can also quickly produce gene/exon table counts for differential expression analysis at feature resolution using DESeq, edgeR, and other similar packages.

DERfinder

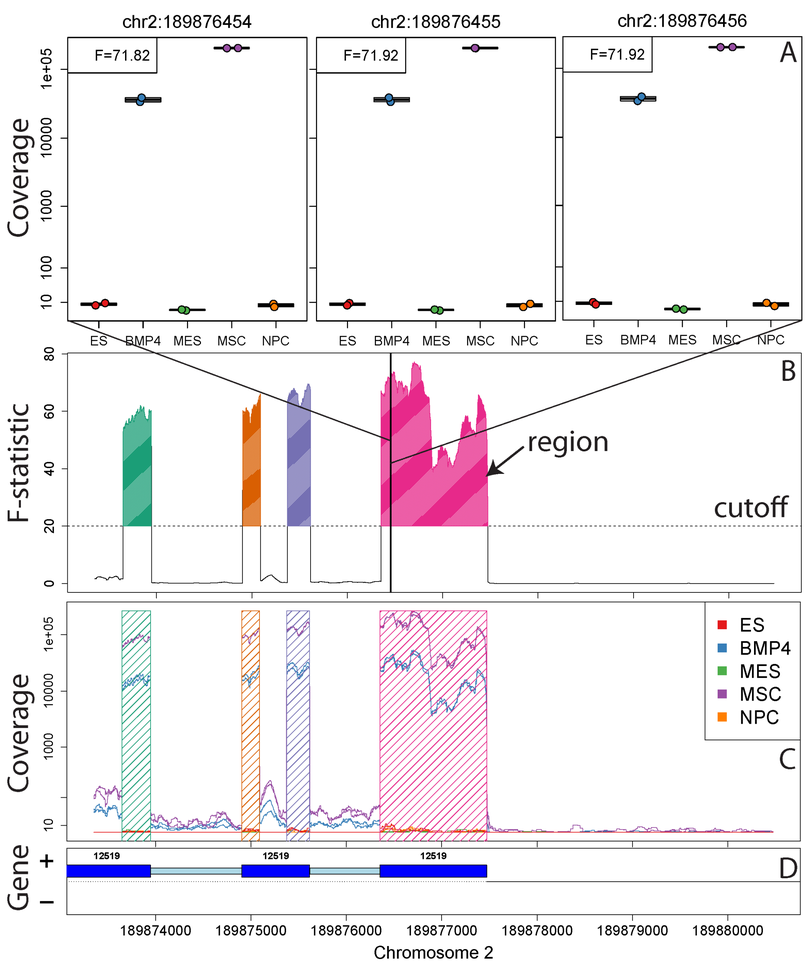

Figure 1. (A) Coverage boxplots for 3 different base pairs. (B) F-statistics curve with candidate DERs in color. (C) Coverage curves for the 5 sample groups. (D) Known annotation.

F-statistics are calculated at each base pair. Contiguous base pairs with F-statistics above a cutoff are considered a candidate differentially expressed region (DER).

derfinder

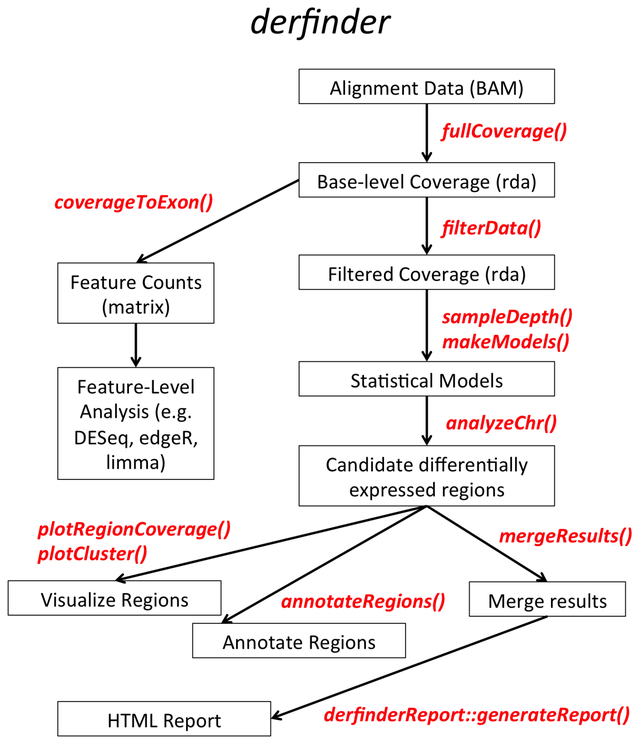

Figure 2. Relationships between the main functions in derfinder and results produced.

For base pairs passing the filter, two nested models are fitted adjusting for confounders and batch effects. A F-statistic testing for significance of the coefficients of interest (group in this case) is calculated. Candidate DERs are defined (Fig 1B) and their area is compared against areas from null regions (obtained via permutations) to calculate empirical p-values. Q-values, which control the FDR, are used to determine significance.

$$Y_{ij} = \log_2\left(\text{Coverage}_{ij} + 32\right)$$

$$i = 1, \ldots, 760e6; \quad j = 1, ..., \text{Nsamples}$$

$$Y_{ij} = \beta_0 + \beta_1 \text{CoverageAdjustment}_j$$

$$Y_{ij} = \beta_0 + \beta_1 \text{CoverageAdjustment}_j + \beta_2 \text{Group}_j$$

$$k = 1, \ldots, nDERs; \quad W_k := \text{width of region } k$$

$$\text{Area}_k = \sum_{l=1}^{W_k} \text{F-statistic}_l$$

$$M = \text{number of null regions across all chrs}$$

$$\text{p-value}_k = \frac{\sum_{m=1}^{M} \text{I}\left(\text{NullArea}_m > \text{Area}_k\right) + 1}{M + 1}$$

Results

Public data sets with high (*Stem*) and low (*Hippo*) group differences, and a time course data set (*Snyder*) were used to demonstrate **derfinder**.

Data set	% Genome w/data	# Candidate DERs	# Significant candidate DERs
Hippo	1.2	28902	2595
Snyder	9.8	20145	1304
Stem	11.2	2626	2491

Table 1. Number of candidate DERs found.

Wall time (hrs.) x cores	Memory (GB) / cores	Software	Data set
5.6	38.8	derfinder	Hippo
15.2	13.5	HTSeq	Hippo
7.1	1.8	summOv	Hippo
12.6	53.1	derfinder	Snyder
164	9	HTSeq	Snyder
26.9	6.3	summOv	Snyder
16.1	48.4	derfinder	Stem
124.4	55.2	HTSeq	Stem
47.3	12.6	summOv	Stem

Table 2. Wall time and memory used to produce gene count tables from BAM files by derfinder, HTSeq, and summarizeOverlaps from GenomicRanges. Resources used were adjusted by the number of cores used. Annotation used: UCSC hg19 knownGene.

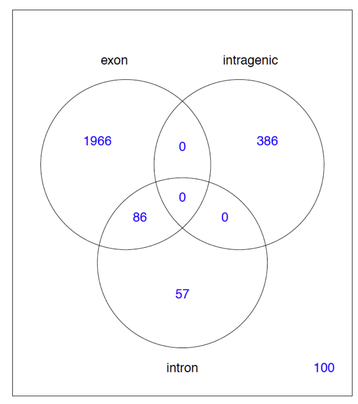

Figure 3. For Hippo data set, overlap (minimum 20 bp) between the 2595 significant candidate DERs and UCSC hg19 known annotation basic features categorized into exons, introns, or intergenic features.

Conclusions

derfinder can readily handle different types of data sets with sample sizes up to several hundred. The number of candidate DERs is sensible to the F-statistics cutoff used (Fig 1 B), yet **derfinder** finds similar numbers of significant candidate DERs (Table 1). The latter ones have a tendency to overlap known exons (Fig 3), with variability due to the underlying biological mechanism under study.

derfinder produces gene/exon count tables much faster than the most commonly used competitors, at the expense of higher (yet feasible) memory requirements (Table 2).

References

1. A. C. Frazee, S. Sabunciyan, K. D. Hansen, R. A. Irizarry, and J. T. Leek (2014). Differential expression analysis of RNA-seq data at single base resolution, *Biostatistics*.
2. L. Collado-Torres, A.C. Frazee, M. I. Love, R. A. Irizarry, A. E. Jaffe, J. T. Leek (2014). Manuscript *submitted*.
3. https://github.com/lcolladotor/derfinder

LCT is supported by CONACyT México and LIBD.