Annotation-agnostic RNA-seq differential expression analysis software

JOHNS HOPKINS BLOOMBERG SCHOOL *of* PUBLIC HEALTH

LIBD LIEBER INSTITUTE *for* BRAIN DEVELOPMENT MALTZ RESEARCH LABORATORIES

Collado-Torres L[1,2,3], Frazee AC[1,3], Love MI[4], Irizarry RA[4], Jaffe AE[1,2,3,5], Leek JT[1,2,3]

[1]Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, 21205 Baltimore, USA. [2]Lieber Institute for Brain Development, Johns Hopkins Medical Campus, 21205 Baltimore, USA. [3]Center for Computational Biology, McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, 21205 Baltimore, USA. [4]Department of Biostatistics, Dana-Farber Cancer Institute and Harvard School of Public Health, 02115 Boston, USA. [5]Department of Mental Health, Johns Hopkins Bloomberg School of Public Health, 21205 Baltimore, USA.

Why annotation agnostic?

* Annotation could be incomplete
* Counting for feature-level is non-trivial
* Assembly is hard

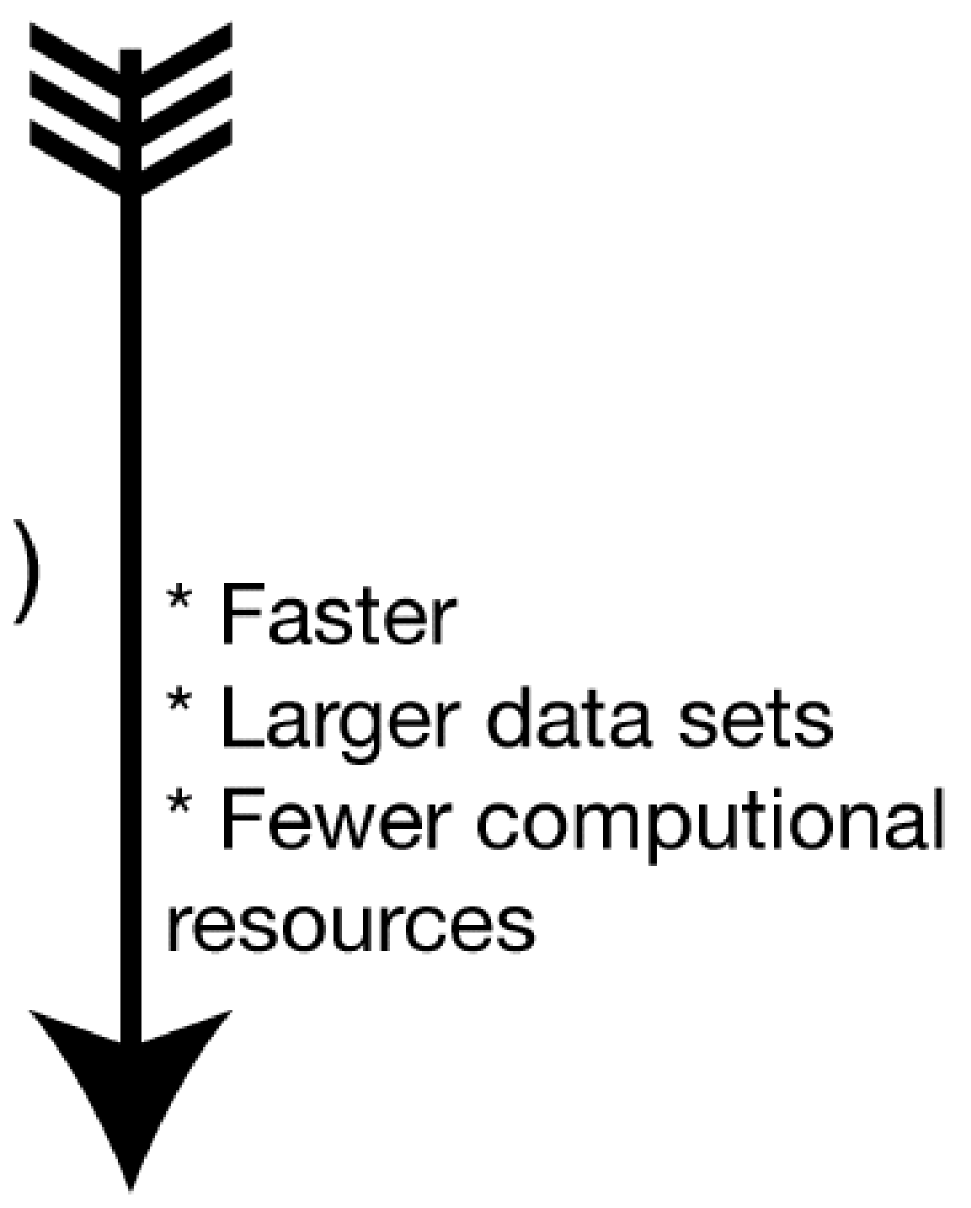
Our approach: find differentically expressed regions (DERs) then find nearest annotated feature.

DER finder versions

* HMM-based for 2 groups
github.com/leekgroup/derfinder

* F-statistics at single-base level (Figure 1)
bioconductor.org/packages/derfinder

* Expressed-regions (Figure 2)
github.com/leekgroup/derfinder2



* Faster
* Larger data sets
* Fewer computional resources

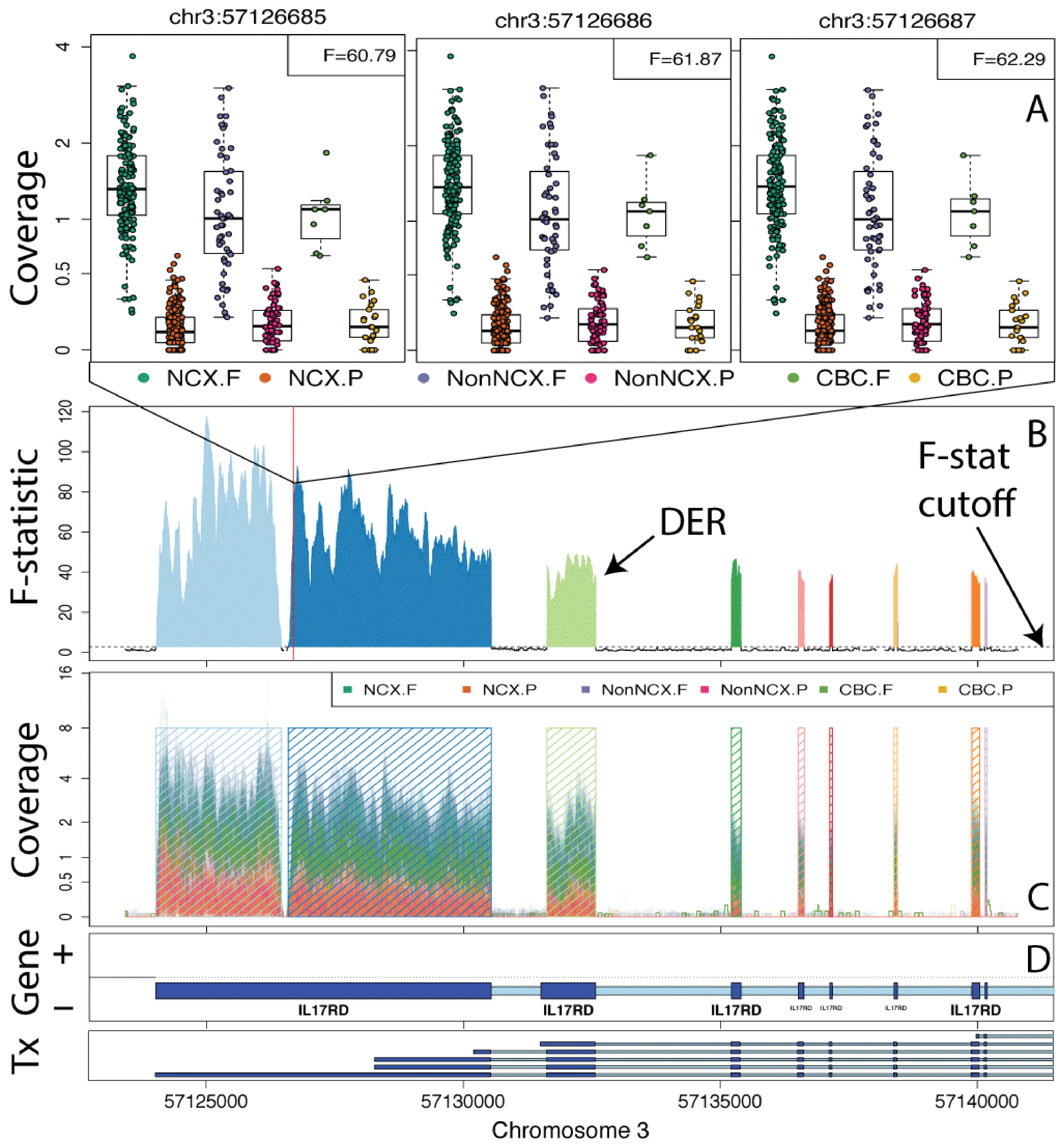


Figure 1
Finding DERs on chromosome 3 with BrainSpan data set using six groups: Neocortical regions (NCX: DFC, VFC, MFC, OFC, M1C, S1C, IPC, A1C, STC, ITC, V1C), Non-neocortical regions (NonNCX: HIP, AMY, STR, MD), and cerebellum (CBC) split by whether the sample is from a fetal (F) or postnatal (P) subject. **A** Boxplots for three specific bases. **B** F-statistics curve with regions passing the F-stat cutoff marked as candidate DERs. **C** Raw coverage curves superimposed with the candidate DERs. **D** Known exons (dark blue) and introns (light blue) by strand. The third DER matches the shorter version of the second exon shown in the *Tx* track.

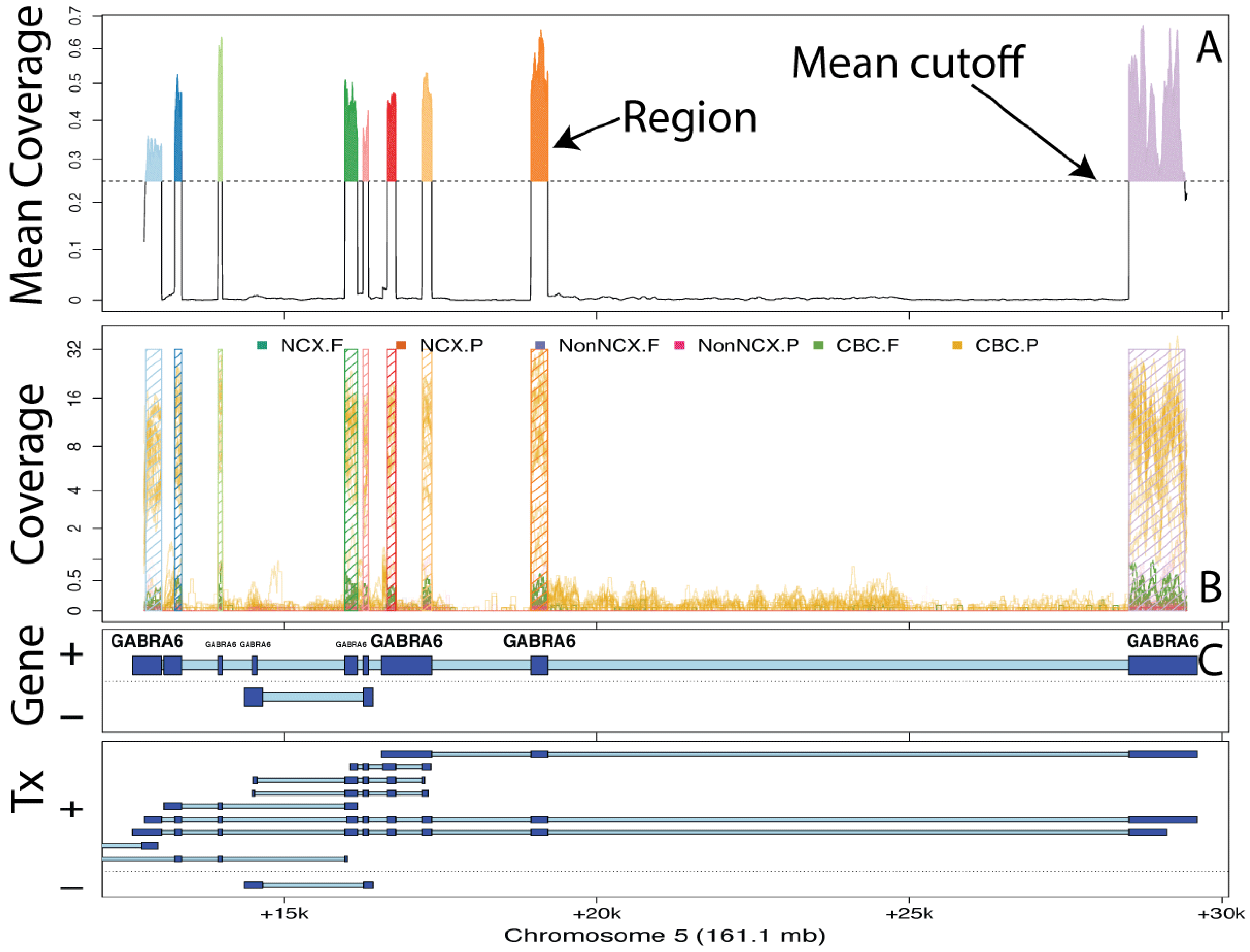


Figure 2
Finding regions via expressed-region approach on chromosome 5 with BrainSpan data set. A Mean coverage with segments passing the mean cutoff (0.25) marked as regions. **B** Raw coverage curves superimposed with the candidate regions. Coverage curves are colored by brain region and developmental stage (NCX: Neocortex: Non-NCX: Non-neocortex, CBC: cerebellum, F:fetal, P: postnatal). **C** Known exons (dark blue) and introns (light blue) by strand for genes and subsequent transcripts in the locus.)

	Original	Single base level	Expressed regions level	DESeq2 20% incomplete at random
FDR	2.1	0	4.2	12.1
FPR	2.7	0	6.3	18.9
Power	83.4	82.2	93.5	89.9

Empirical false discovery rate (FDR), false positive rate (FPR) and power from a simulated data set. Original implementation, single-base and expressed-region level analyses are compared against DESeq2. All analyses were performed controlling the FDR at 5%.

References

Frazee AC, et al (2014). Differential expression analysis of RNA-seq data at single-base resolution. Biostatistics. doi:10.1093/biostatistics/kxt053

Jaffe AE, et al (2014). Developmental regulation of human cortex transcription and its clinical relevance at single base resolution. Nat. Neurosci. doi:10.1038/nn.3898

Collado-Torres L, et al (2015). derfinder: Software for annotation-agnostic RNA-seq differential expression analysis. bioRxiv. doi:10.1101/015370

Funding

JTL was supported by NIH Grant 1R01GM105705, LCT was supported by Consejo Nacional de Ciencia y Tecnología México 351535.