Identifying the Tissue of Origin of Extracellular Vesicles Using RNA Expression Signatures

Don L. Armstrong, Monica Uddin, Derek Wildman

Institute for Genomic Biology, Computing Genomes for Reproductive Health, University of Illinois, Urbana-Champaign

CARL R. WOESE
INSTITUTE FOR
GENOMIC BIOLOGY
Where Science Meets Society

Introduction

Extracellular Vesicles (EVs) are small (40-1000nm) compartments that can be found in blood, saliva, and urine. EVs originate from the plasma membrane or multivesicular endosome of many cell types and contain proteins, lipids, mRNA, miRNA, and many other non-coding RNAs. Because EVs contain mRNA from the cell of origin and can be obtained without invasive procedures, they are a promising remote sensor of the transcriptome of tissues that would otherwise be inaccessible.

One requirement for remotely sensing the transcriptome is identifying the source tissue of a specific EV. To identify the source tissue of a specific EV, we have identified miRNA and mRNA markers which are characteristic of multiple organs and tissues from existing publicly-available RNAseq transcriptomes by calculating a tissue specificity index.

The tissue specificity index for gene g (τ_g) is

$$\tau_g = \frac{\sum_{i=1}^{i=N}\left(1 - \frac{X_{i,g}}{\max(X_g)}\right)}{N-1} \qquad (1)$$

where N is the number of tissues, $X_{i,g}$ is the mean expression of gene g in tissue i and $\max(X_g)$ is the maximal expression of gene g across all tissues.

We also trained a multiclass Support Vector Machine (SVM) to accurately identify the tissue of origin, and present the results of the identification of reads from uterine and placental origin here. Tables showing tissue specific genes from all 126 tissues can be found on `http://dla2.us/p/em2015`.

Methods

2.14×10^{10} RNAseq reads from 289 samples (supplemented with FPKMs from 2978 additional samples without raw data in the Sequence Read Archive (SRA)) corresponding to 126 different tissues were obtained from the Roadmap Epigenomics [1] and GTEx [2] projects, as well as uterine (GSE50599, [3]) and placenta (GSE66622, [4]) samples from the SRA.

Reads were aligned to *Homo sapiens* reference genome GRCh38 with Ensembl annotation release 80 using STAR v2.4.2a [5] and called using Tophat 2 v2.0.10 [6]. The average percentage of uniquely mapped reads was 72.72% and the average total mapping percentage was 88.19%. For samples where the reads were not publicly available (GTEx and some Roadmap Epigenomics samples), the RPKM were taken from publicly available result files (see code for details).

The SVM was generated using caret [7] and svmRadial [8] with 10 cycles of 10 repeats and a tune length of 8 with a training set representing $\approx 75\%$ of the samples (≈ 2450 samples total, including ≈ 35 uterine or ≈ 60 placenta) in each of the groups chosen at random, with the remaining samples making up a testing set (≈ 820 samples).

Genes Characteristic of Uterus

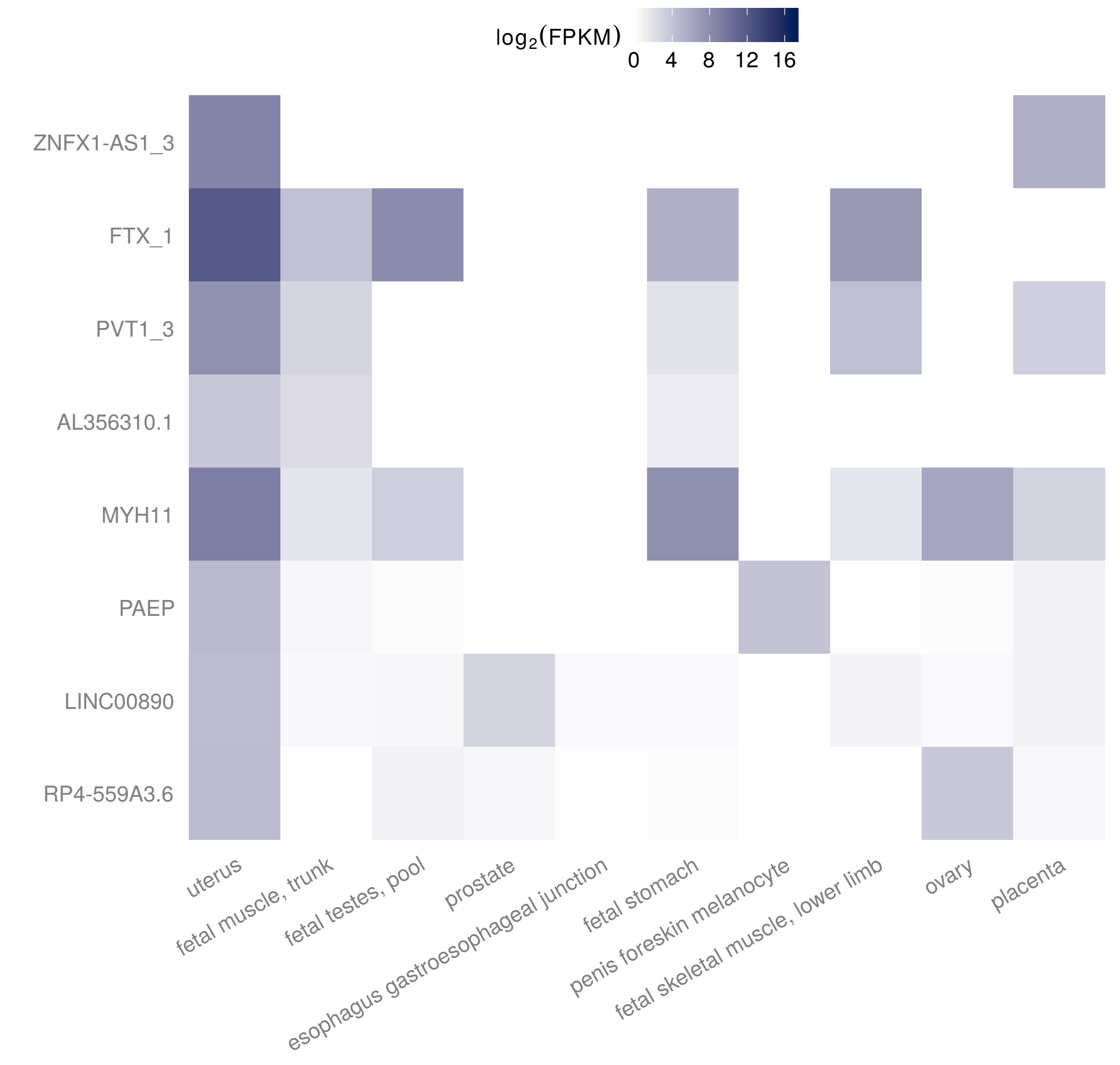

Figure 1: Expression of genes characteristic of uterus ($\tau_g \geq 0.98$) in uterus and all tissues with the second highest expression for any of the uterus-characterizing genes ordered by their tissue specificity. The known uterus-characterizing gene PAEP and short RNAs ZNFX1-AS1_3 (87bp) and AL356310.1 (94bp) are highly characteristic of uterus.

Genes Characteristic of Placenta

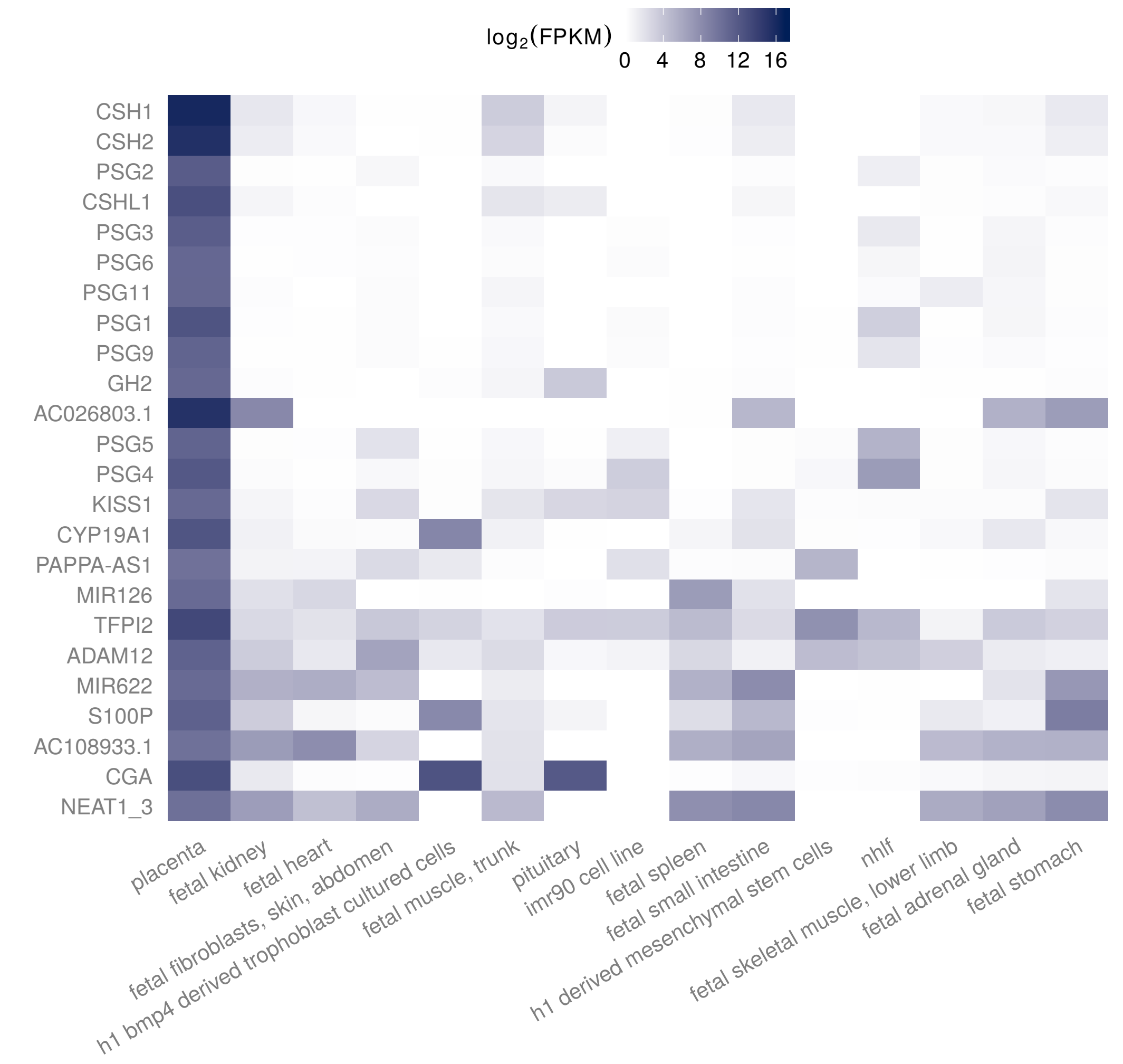

Figure 2: Expression of genes characteristic of placenta ($\tau_g \geq 0.98$) with high expression (FPKM ≥ 1024) in placenta and all tissues with the second highest expression for any placenta-characterizing gene ordered by their specificity to placenta. The previously known placenta-characterizing PSG genes and CSH genes are nearly unique to placenta and the short RNAs mir-126, mir-622, and PAPPA-AS1 are less so.

SVM to distinguish tissues

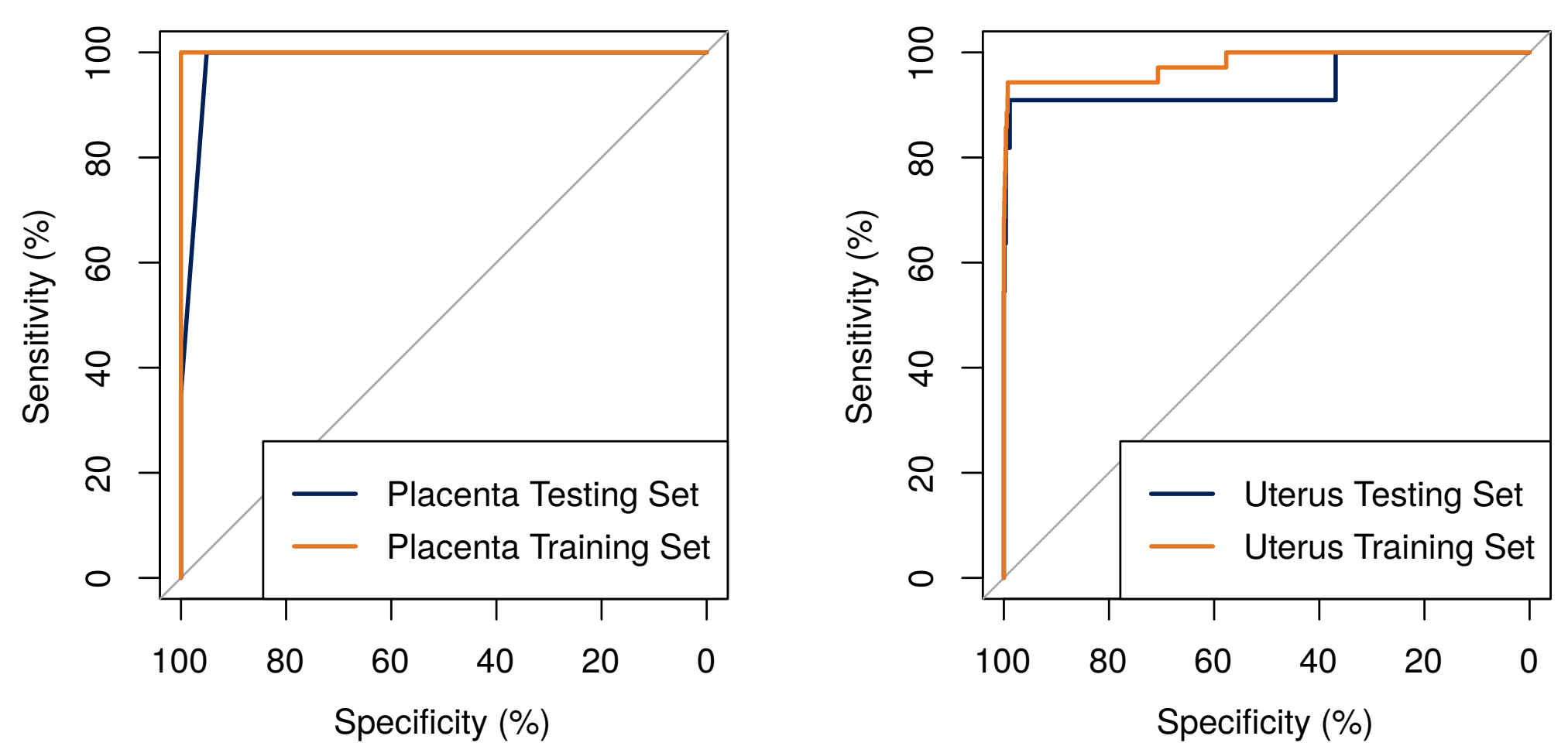

Figure 3: ROC curve of the placenta SVM (left) and uterus SVM right) . Both SVM are highly effective at distinguishing between samples of uterus origin and non-uterus origin.

Identification of tissue in reduced reads

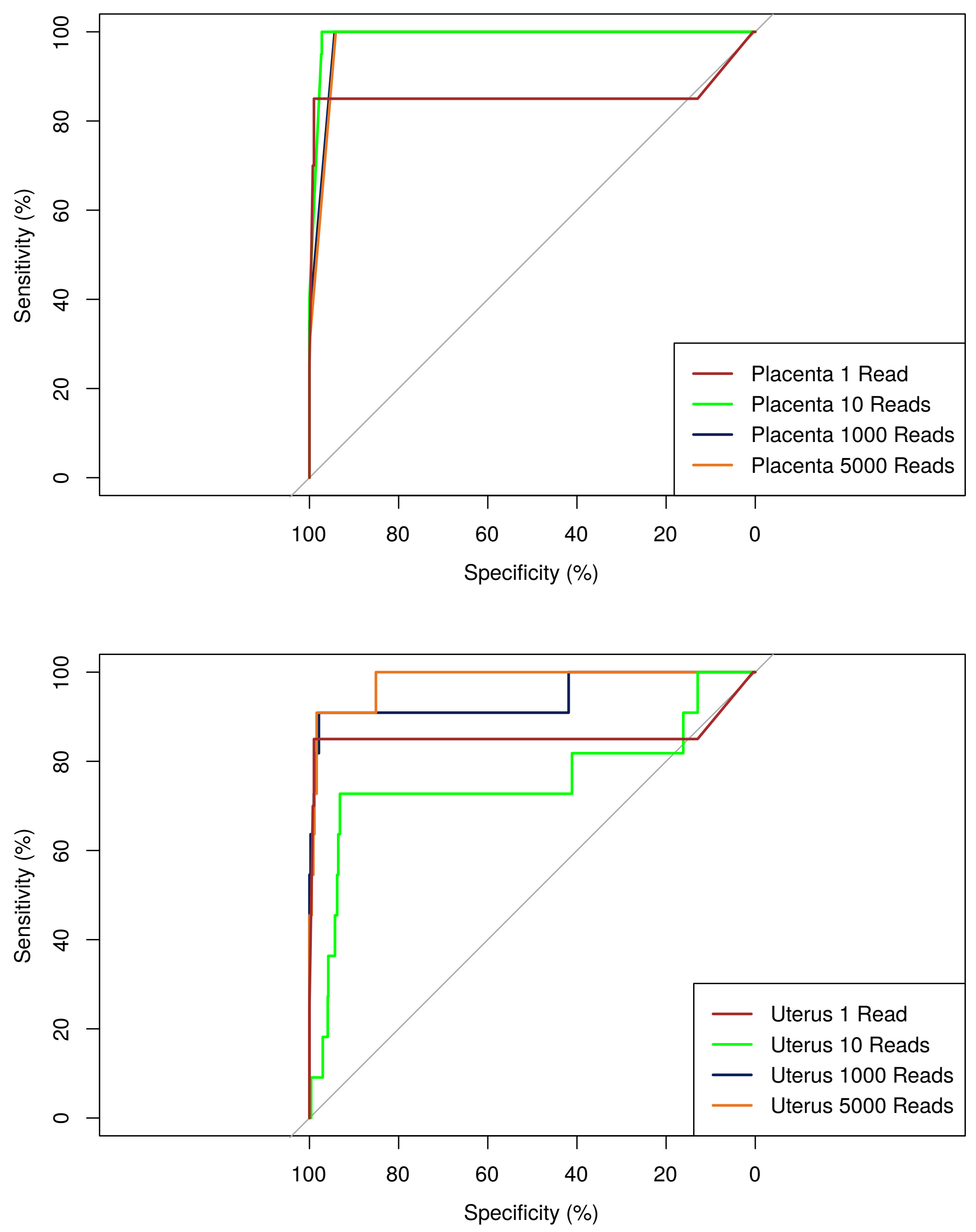

Figure 4: ROC curve of reduced reads in Placenta and Uterus samples with the indicated number of total reads. Reads were reduced by sampling reads for each sample weighted by the FPKM measured for each sample, and run using the same trained SVM depicted in Figure 3. The placenta SVM is fairly accurate at distinguishing between samples of placental origin and non-placental origin even when the number of reads is very low (10). The uterus SVM requires many more reads in order to achieve similar accuracy. This is likely due to the much smaller number of uterus specific genes in comparison to placenta.

Conclusions

- Pregnancy-specific glycoproteins (PSG1, 2, 3 et al.) are characteristic of placental tissues, and PAEP is characteristic of uterus, as previously described.
- SVM are able to accurately identify the tissue of origin for uterus and placenta.
- Very few sequenced reads are required to accurately identify tissues of Placental origin. We are able to detect 100% of the placenta tissues with 80% specificity with just 10 reads.
- We can easily identify exosomes originating from the placenta, as even the smallest exosomes can contain 70 RNA molecules [9].
- Non-maternal variants may also be able to distinguish between placental reads and maternal reads.

References

1. Bernstein, B. E. *et al.* The NIH Roadmap Epigenomics Mapping Consortium. *Nat. Biotechnol.* **28,** 1045–1048 (Oct. 2010).
2. The GTEX Consortium. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348,** 648–660 (May 2015).
3. Chan, Y.-W., van den Berg, H. A., Moore, J. D., Quenby, S. & Blanks, A. M. Assessment of myometrial transcriptome changes associated with spontaneous human labour by high-throughput RNA-seq. *Exp. Physiol.* **99,** 510–524 (Mar. 2014).
4. Hughes, D. A. *et al.* Evaluating intra- and inter-individual variation in the human placental transcriptome. *Genome Biol.* **16,** 54 (2015).
5. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29,** 15–21 (Jan. 2013).
6. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25,** 1105–1111 (May 2009).
7. Kuhn, M. Building Predictive Models in R Using the caret Package. *J Stat Soft* **28,** 1–26. ISSN: 1548-7660 (Nov. 10, 2008).
8. Karatzoglou, A., Smola, A., Hornik, K. & Zeileis, A. kernlab - An S4 Package for Kernel Methods in R. *J Stat Soft* **11,** 1–20. ISSN: 1548-7660 (Nov. 2, 2004).
9. Li, M. *et al.* Analysis of the RNA content of the exosomes derived from blood serum and urine and its potential as biomarkers. *Philos. Trans. R. Soc. Lond., B, Biol. Sci.* **369,** (Sept. 2014).

Code and Datasets

The code and underlying data for this poster can be found at

`http://dla2.us/p/em2015`

I L L I N O I S
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN
1867

LaTeX TikZposter