Identifying Extracellular Vesicles of Placenta and Uterine Origin 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

October 5, 2015

Introduction: EVs

Extracellular Vesicles (EVs) are

- small (40-1000nm) compartments
- found in blood, saliva, and urine.
- originate from the plasma membrane or multivesicular endosome
- from many cell types
- contain proteins, lipids, mRNA, miRNA, and other ncRNAs.
- remote sensor of the transcriptome of inaccessible tissues

Identifying source tissue of EVs

1. We want to remotely sense the transcriptome of tissues
2. For each EV, identify the source tissue
3. So, identify miRNA and mRNA markers which are characteristic of multiple organs and tissues
   - Use publicly-available RNAseq transcriptomes
   - Identify characteristic markers by calculating a tissue specificity index for each gene g (τ_g):

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

   - 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

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

Source of EVs — Machine Learning

- High tissue specificity might not be the best discriminator
- Use machine learning techniques!
- Train a Support Vector Machine (SVM)
- Use genes which are highly expressed and have either
  - High Tissue Specific Index (τ_g)
  - High Entropy

Methods

- 2.14×10^{10} RNAseq reads (7.42×10^{7} average per sample)
- 289 samples (supplemented with FPKMs from 2978 additional samples without raw data in the Sequence Read Archive (SRA))
- 126 different tissues
- Samples from Roadmap Epigenomics [1], GTEx [2], uterine (GSE50599, [3]) and placenta (GSE66622, [4])
- aligned to GRCh38 using STAR [5]
- quantified using Cufflinks 2 v2.0.10 [6].
- Mean uniquely mapped reads: 72.72%
- Mean total mapping percentage: 88.19%.
- RPKMs from publicly available result files were used for samples where reads were not present in SRA.

Genes Characteristic of Uterus

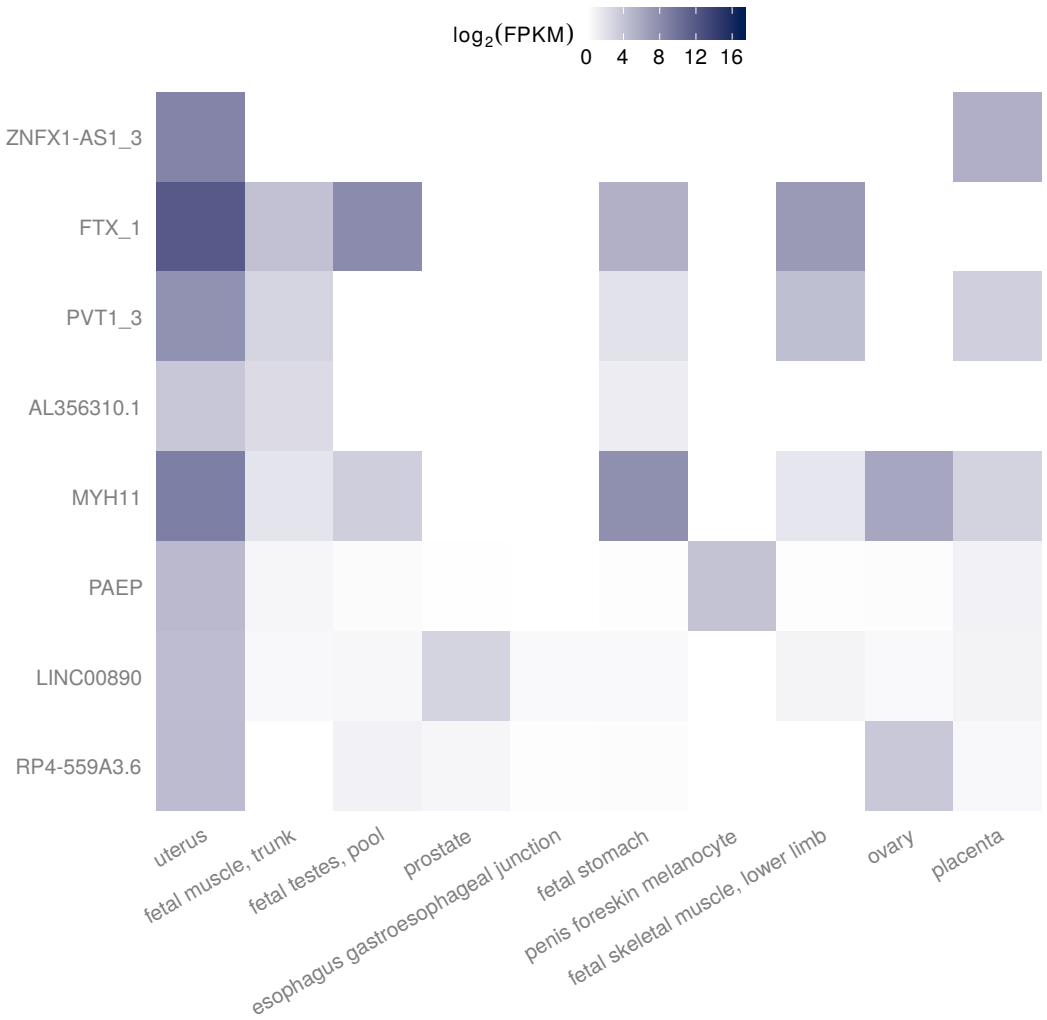


Figure: Expression of genes characteristic of uterus ($\tau_g \geq 0.98$) in uterus and all tissues with the second highest expression for any of the uterus-characterizing genes ordered by their tissue specificity. The known uterus-characterizing gene PAEP and short RNAs ZNFX1-AS1_3 (87bp) and AL356310.1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

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

Genes Characteristic of Placenta

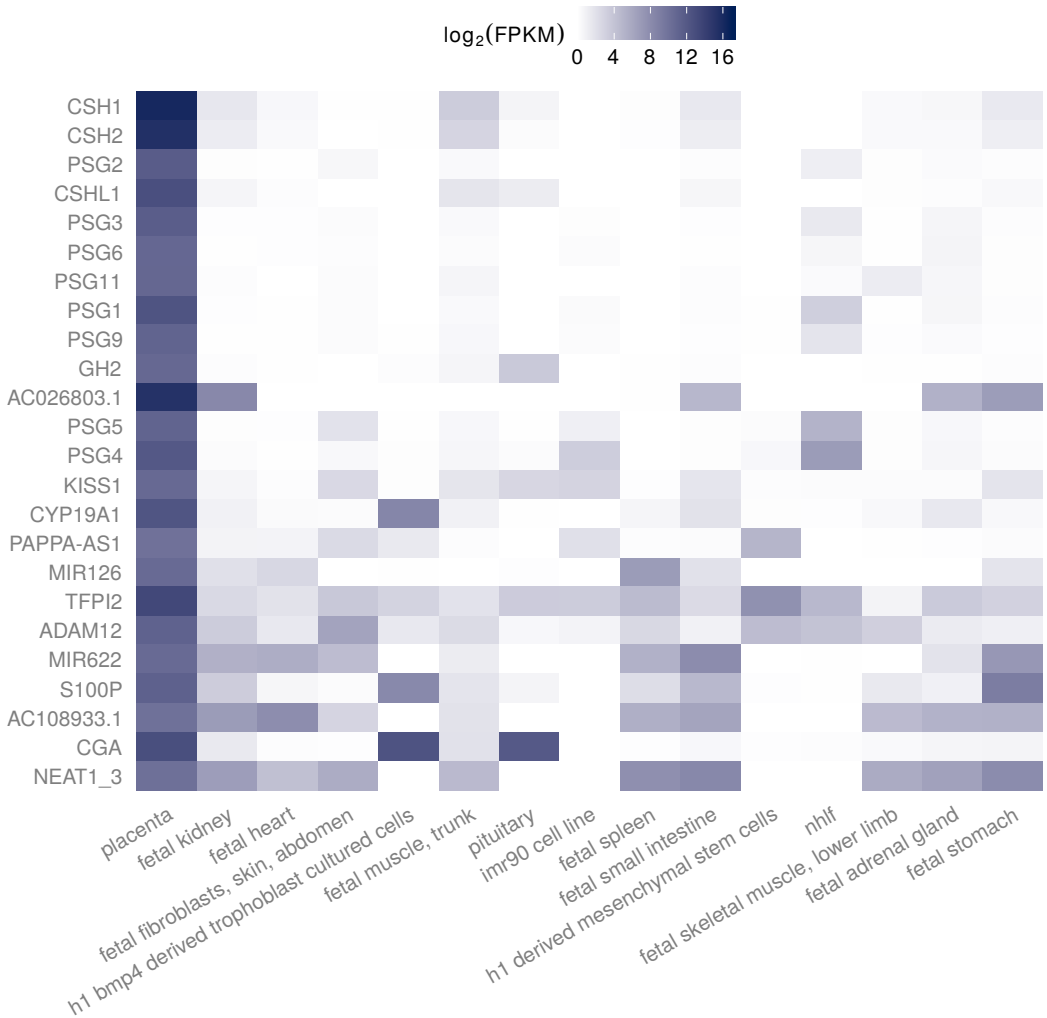


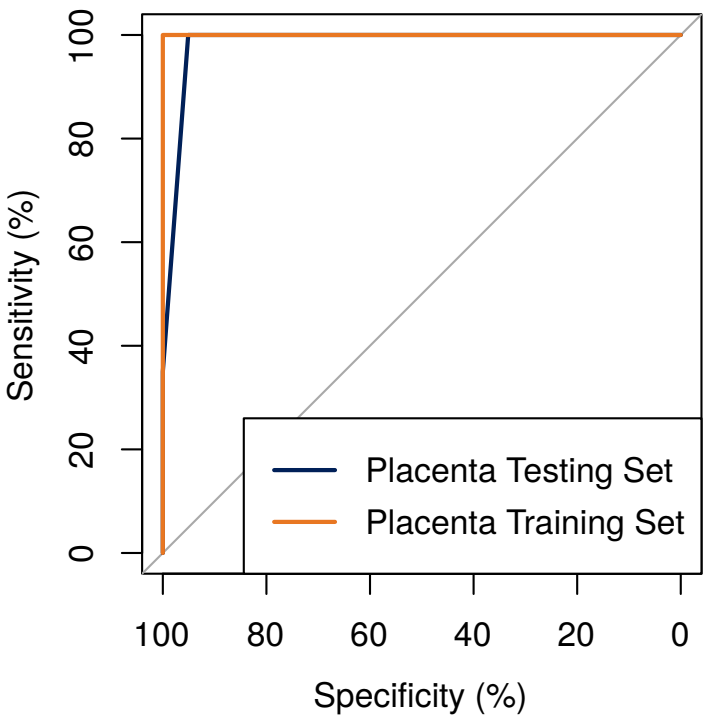
Figure: Expression of genes characteristic of placenta ($\tau_g \geq 0.98$) with high expression ($\mathrm{FPKM} \geq 1024$) in placenta and all tissues with the second highest expression for any placenta-characterizing gene ordered by their specificity to placenta. The previously known placenta-characterizing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

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

Generating SVM

- SVM was generated using caret [7] and svmRadial [8]
- 10 cycles of 10 repeats and a tune length of 8
- 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,
- remaining samples making up a testing set (≈ 820 samples)

Placenta and Uterus SVM



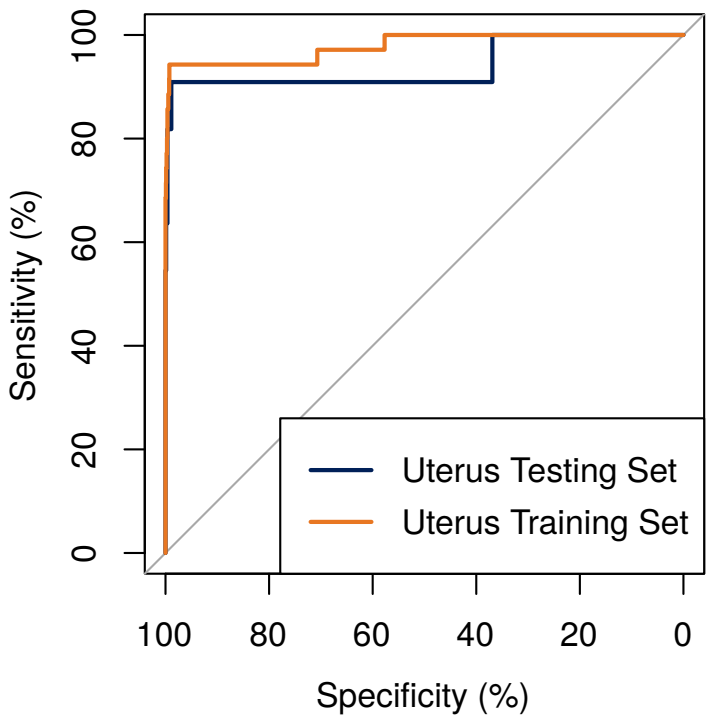


Figure: ROC curve of the placenta SVM (left) and uterus SVM right) . Both SVM are highly effective at distinguishing between samples of uterus origin and non-uterus origin.

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

Reducing Reads

- Reads sub-sampled from each sample
- Weighted by the FPKM measured
- Run using the same trained SVM shown previously

Identification of tissue in reduced reads

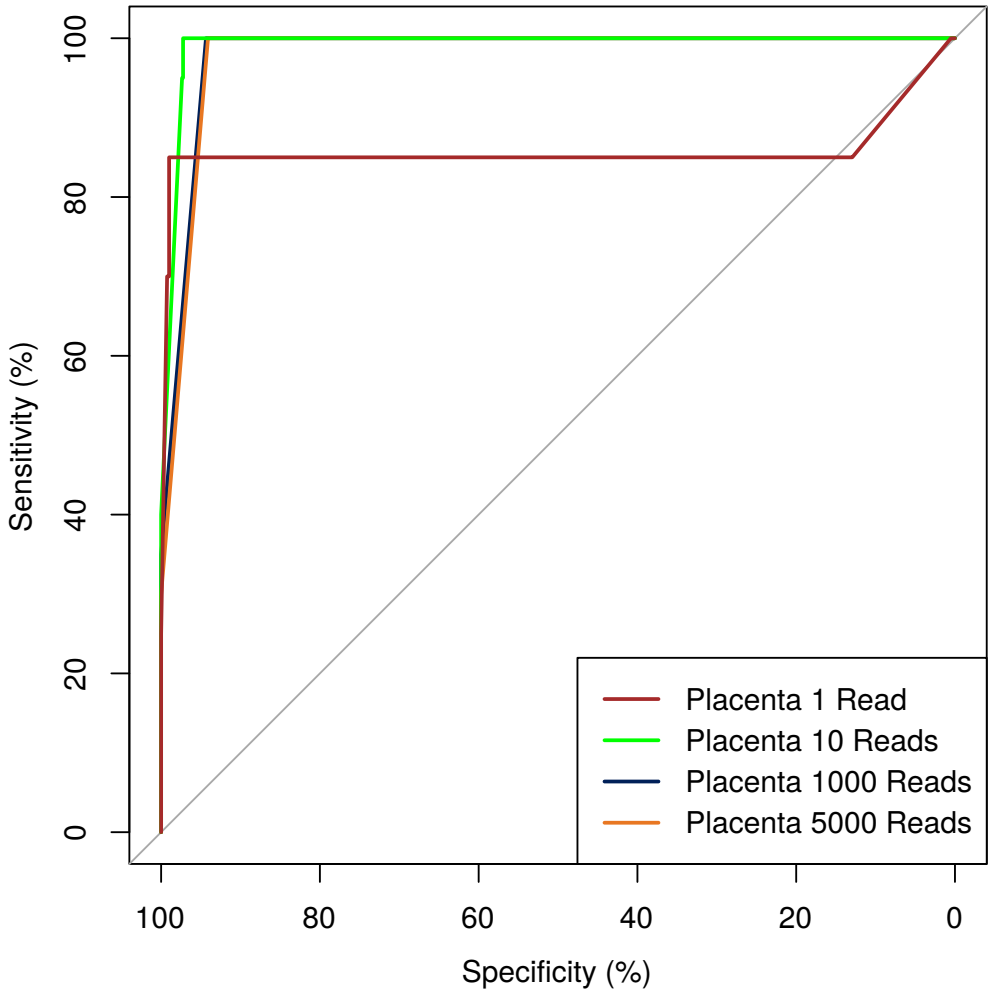


Figure: ROC curve of reduced reads in Placenta with the indicated number of total reads. The placenta SVM is fairly accurate at distinguishing between samples of placental origin and non-placental origin even when the number of reads is very low (10).

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

Reduced Reads: Uterus

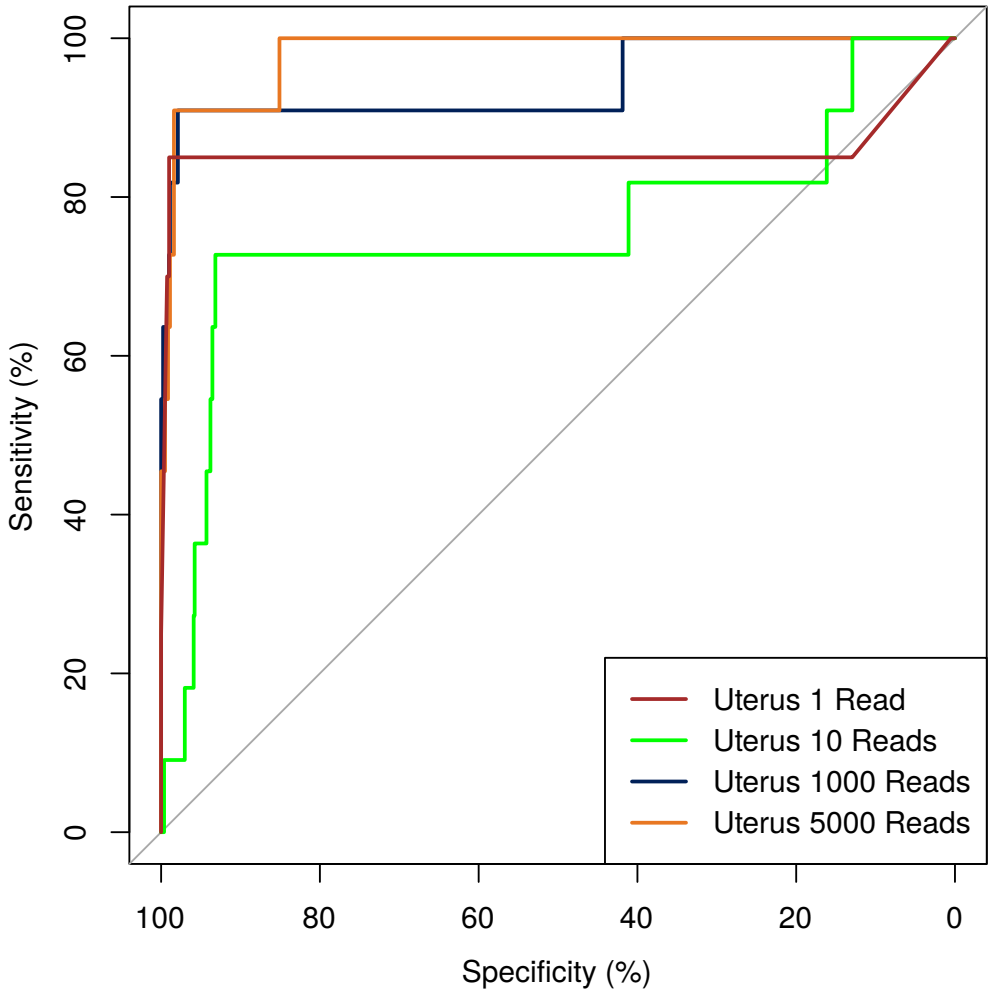


Figure: ROC curve of reduced reads in Placenta and Uterus samples with the indicated number of total reads. The uterus SVM requires many more reads in order to achieve similar accuracy. This is likely due to the much smaller number of uterus specific genes in comparison to placenta.

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

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.

Acknowledgments

- Derek Wildman
- Monica Uddin

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

http://dla2.us/p/em2015

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

References


Bernstein, B. E. *et al.* The NIH Roadmap Epigenomics Mapping Consortium. *Nat. Biotechnol.* **28,** 1045–1048 (Oct. 2010).


The GTEX Consortium. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348,** 648–660 (May 2015).


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).


Hughes, D. A. *et al.* Evaluating intra- and inter-individual variation in the human placental transcriptome. *Genome Biol.* **16,** 54 (2015).


Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29,** 15–21 (Jan. 2013).


Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25,** 1105–1111 (May 2009).


Kuhn, M. Building Predictive Models in R Using the caret Package. *J Stat Soft* **28,** 1–26. ISSN: 1548-7660 (Nov. 10, 2008).


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).


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).