PCa-Clf: A Classifier of Prostate Cancer Patients into Patients with Indolent and Aggressive Tumors Using Machine Learning

Yashwanth Karthik Kumar Mamidi[1], Tarun Karthik Kumar Mamidi[2], Md Wasi Ul Kabir[1], Jiande Wu[3], Md Tamjidul Hoque[1,*], Chindo Hicks[3,*]

[1] Department of Computer Science, University of New Orleans, New Orleans, LA 70148, USA; mamidiyashwanth20@gmail.com (Y.K.K.M.); mkabir3@uno.edu (M.W.U.K.); thoque@uno.edu (M.T.H.)
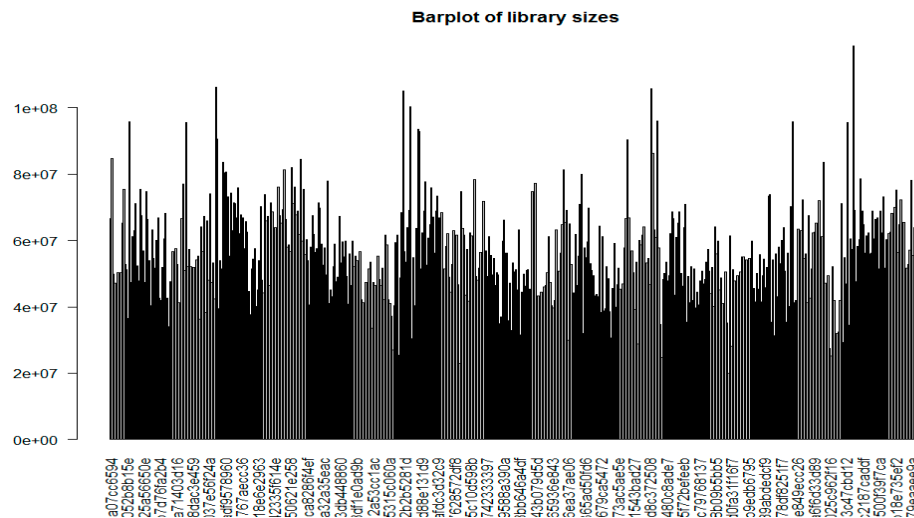
[2] Department of Genetics, Heersink School of Medicine, The University of Alabama at Birmingham, 720 20th Street S, Birmingham, AL, 35294, USA; tmamidi@uab.edu (T.K.K.M.)

[2] Center for Computational Genomics and Data Science, The University of Alabama at Birmingham, 912 18th St S, Birmingham, AL, 35233, USA; tmamidi@uab.edu (T.K.K.M.)
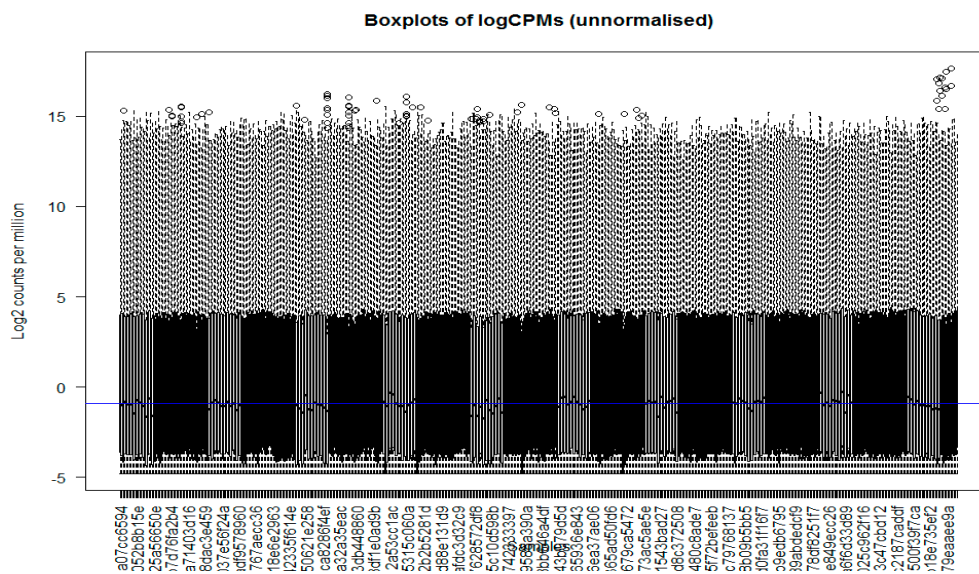
[3] Department of Genetics and the Bioinformatics and Genomics Program, Louisiana State University Health Sciences Center, School of Medicine, 533 Bolivar Street, New Orleans, LA 70112, USA; jiandewu@lsuhsc.edu (J.W.); chick3@lsuhsc.edu (C.H.)

* Correspondence: thoque@uno.edu (M.T.H.); chick3@lsuhsc.edu (C.H.)
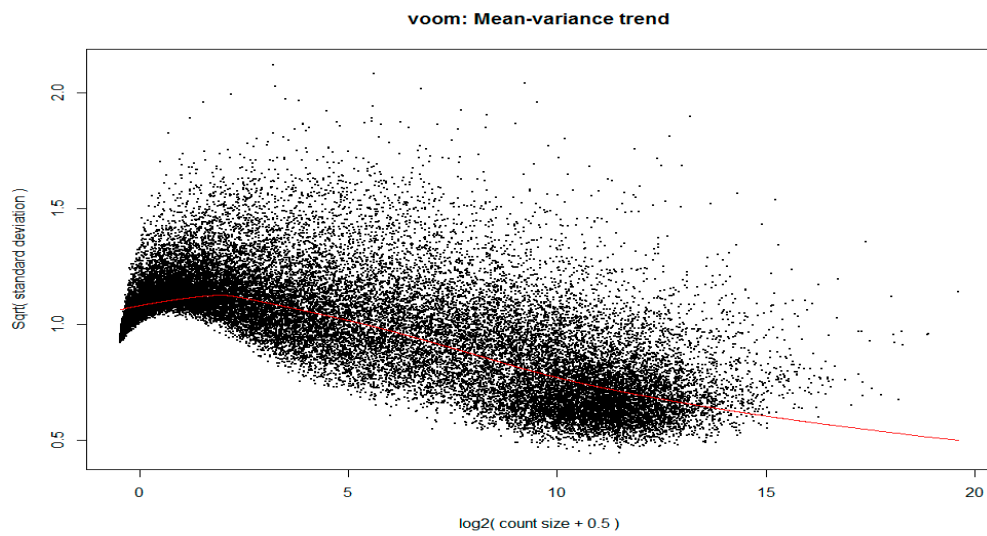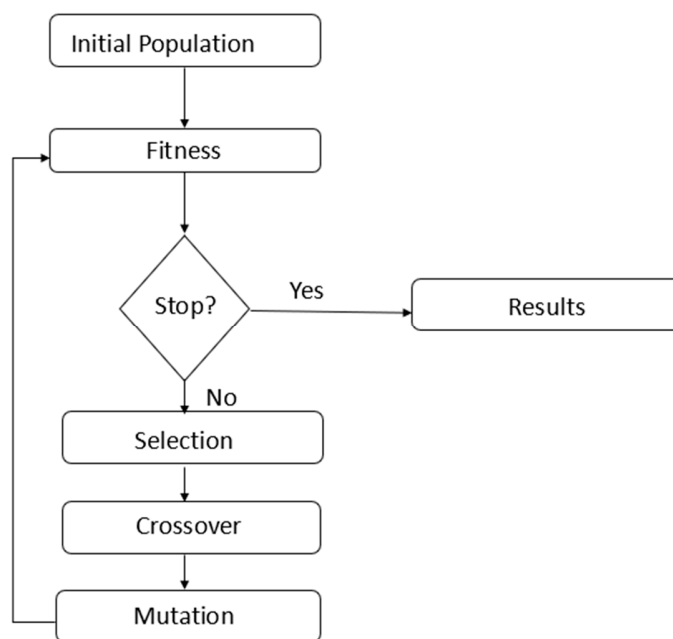
(Supplementary Material)

Supplementary Figure S1. Library sizes of all samples expressed using a bar plot constitute data quality and un-normalized library sizes.
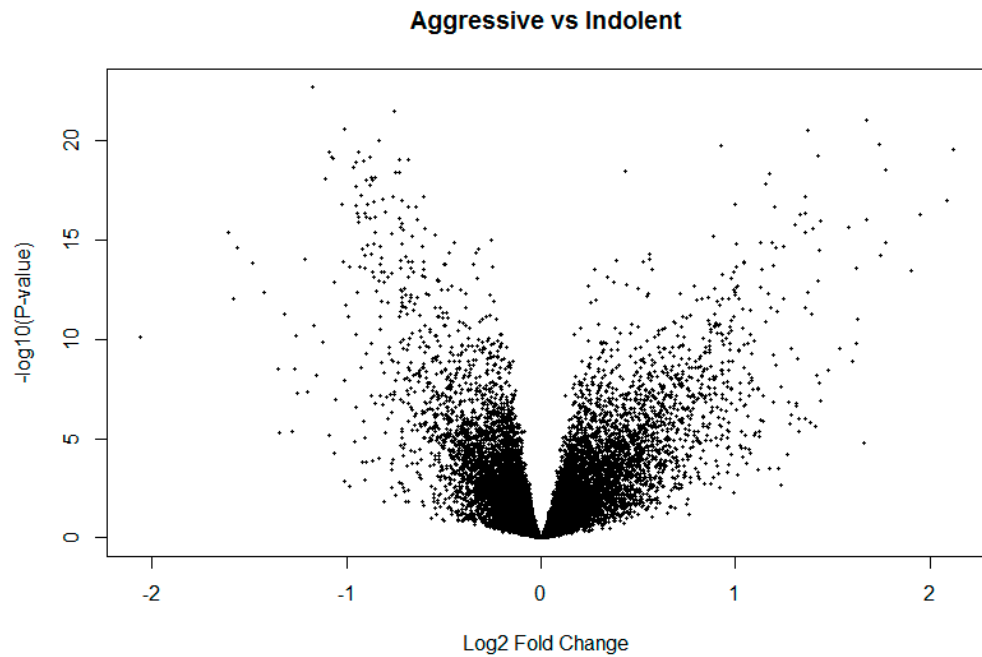

Supplementary Figure S2. Figure checks the read counts' distribution on the log2 scale of logCPM (log counts per million) before normalization.

voom: Mean-variance trend

Supplementary Figure S3. Figure showing the Voom transformation using a decision matrix and mean–variance trend.
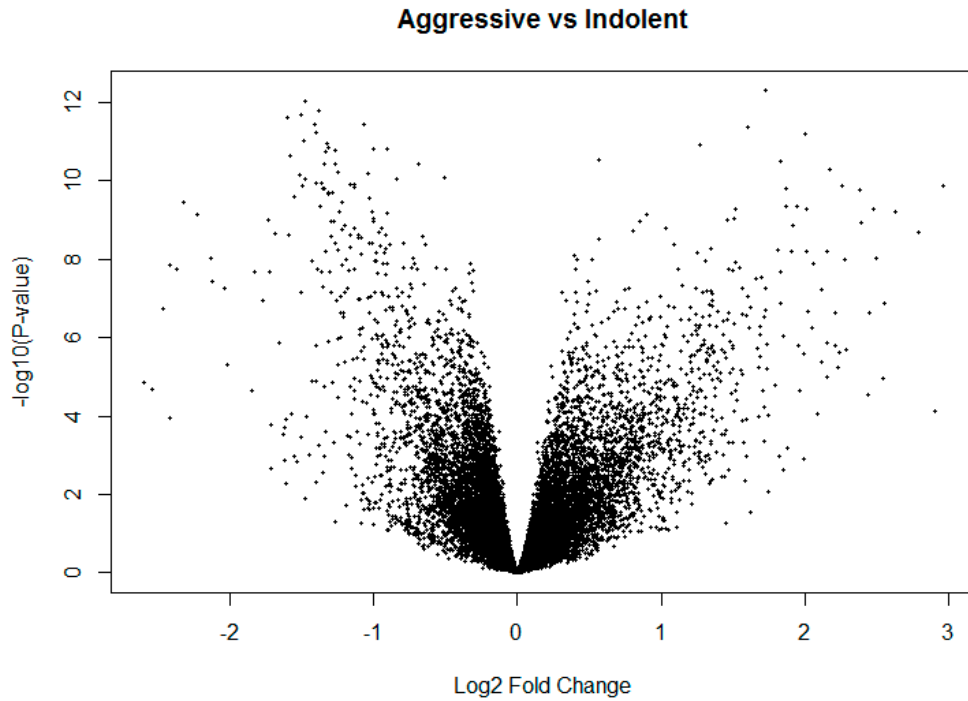

Supplementary Figure S4. The flowchart represents the implementation of the Genetic Algorithm.

Supplementary Figure S5. Volcano plot from Differential Expression Analysis using Limma package on indolent- vs. aggressive-PCa samples (Model 1).

- This initial level 1 analysis for Model 2 yielded 15,105 significantly ($p < 0.05$) and 20,712 significantly ($p < 0.05$) differentially expressed probes associated with indolent and aggressive PCas, respectively. We used a volcano plot (Figure 6) to discover a signature of genes significantly ($p < 0.05$) associated with each disease state.
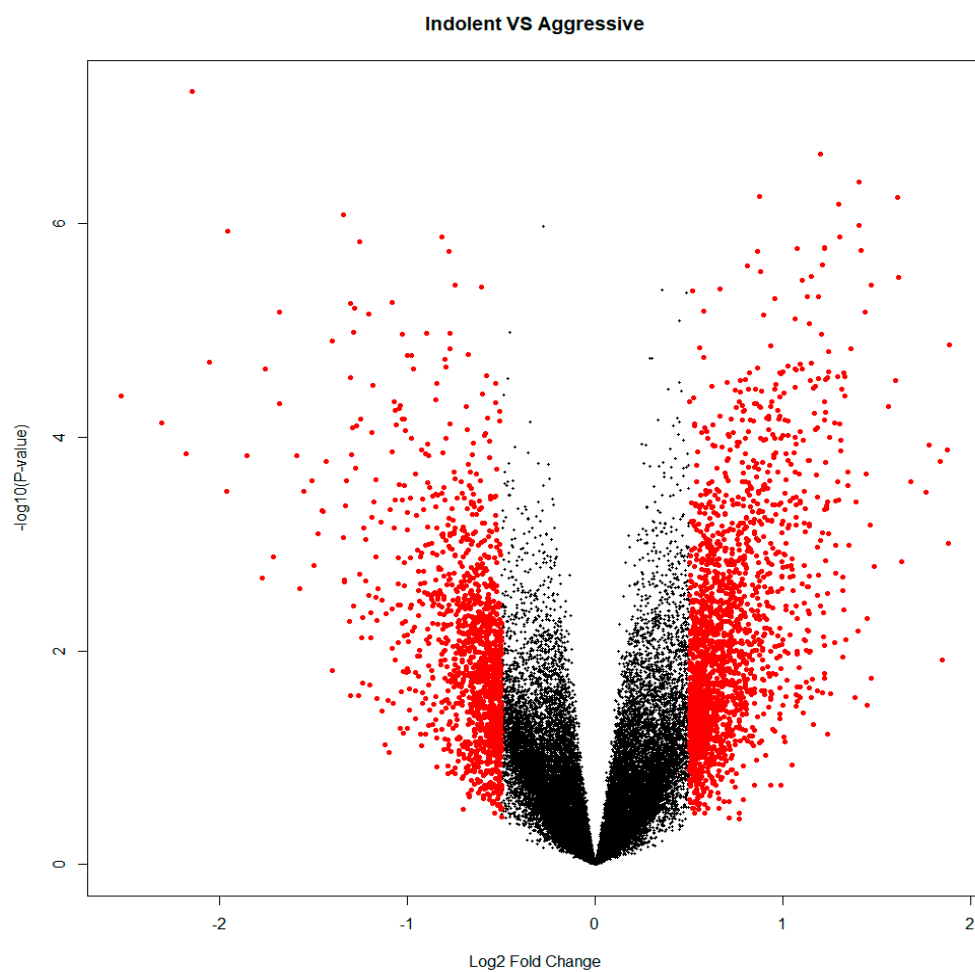
Aggressive vs Indolent

Supplementary Figure S6. Volcano plot from Differential Expression Analysis using Limma package on indolent- vs. aggressive-PCa samples (Model 2).

- The initial level-1 analysis for Model 3 yielded 5220 significantly ($p < 0.05$) and 3352 significantly ($p < 0.05$) differentially expressed probes associated with Indolent and Aggressive, respectively. We then compared indolent- and aggressive-PCa samples using the differentially expressed probes from the earlier analysis and used a volcano plot (Figure 7) to discover a signature of genes significantly ($p < 0.05$) associated with each disease state.

Supplementary Figure S7. Volcano plot from Differential Expression Analysis using Limma package on indolent- vs. aggressive-PCa samples (Model 3).