Generation and Analysis of Breast Tumor Data

Tennessee TECH

Farzana Ahamed Bhuiyan, MD Bulbul Sharif

30% of newly diagnosed cancers in women will be breast cancers

women in the U.S., breast cancer death rates are higher than those for any other cancer, besides lung cancer

about 1 in 8 U.S. women (about 12.4%) will develop invasive breast cancer over the course of her lifetime

About 5-10% of breast cancers can be linked to gene mutations and about 85% of breast cancers occur in women who have no family history of breast cancer

Motivation

Precision medicine is an emerging approach for disease treatment and prevention that takes into account individual variability in genes, environment, and lifestyle. Considering the severe complication of breast cancer treatment, it is high time we engage precision medicine in breast cancer treatment to improve our ability to predict which treatments will be the most effective for specific patients and for this to work we need a clearer idea about the genomes of breast tumors.

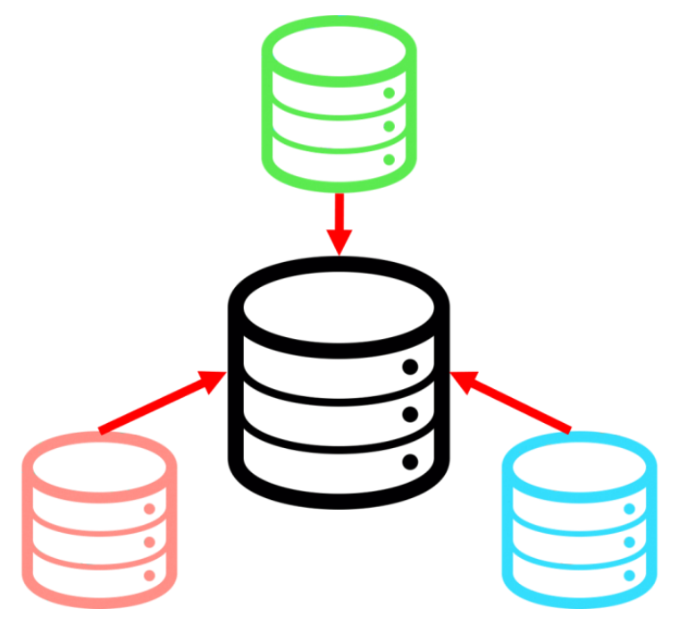
Introduction

A few decades ago, breast cancer classification systems were based on tumor response to endocrine therapy. Currently, molecular classification of breast tumors is used along with classical prognostic factors to predict tumor evolution and behavior and to select specific treatments accordingly. This development improves diagnostic accuracy and enhances the ability to individualize therapy for breast cancer, thereby leading to direct implications for patient management.

Methods

Intrinsic Gene List: Our primary goal is to derive an objective "intrinsic sub-type" classifier that can be used clinically. For this purpose, first, we collected a breast tumor intrinsic list. Intrinsic gene list is the list of genes with significantly greater variation in expression between different tumors than between paired samples from the same tumor.

Data Fusion: validation is often unconvincing because the size of the available test set is typically small. For this reason, to generate a validation set we merged four publicly available breast cancer expression datasets using Distance Weighted Discrimination (DWD)



Hierarchical Clustering: Tumor subtypes identified by the intrinsic gene set are predictive of outcome. To find the subtypes, first, we did the hierarchical clustering. To determine how many biologically relevant tumor subtypes might be present within the cluster, we used a dendrogram branching pattern and the knowledge of the reference paper.

K-means clustering: To find the tumor subtypes we tried a different method called subgroup discovery and again used our domain knowledge from the reference paper to determine the tumor subtypes.

Kaplan-Meier survival analysis: After finding the groups, we tried to find out the differences in outcomes and associations with other clinical parameters between each of the groups. We did Kaplan-Meier survival analysis on the combined set, to compare hierarchical clustering and subgroup discovery method and to find out whether the groups were predictive of Relapse-Free Survival and Overall Survival.

Cox Proportional-Hazards Model: This regression model was used for investigating the association between the survival time of patients and our intrinsic subtype classifications. We also used this analysis to compare the two methods mentioned above.

Results

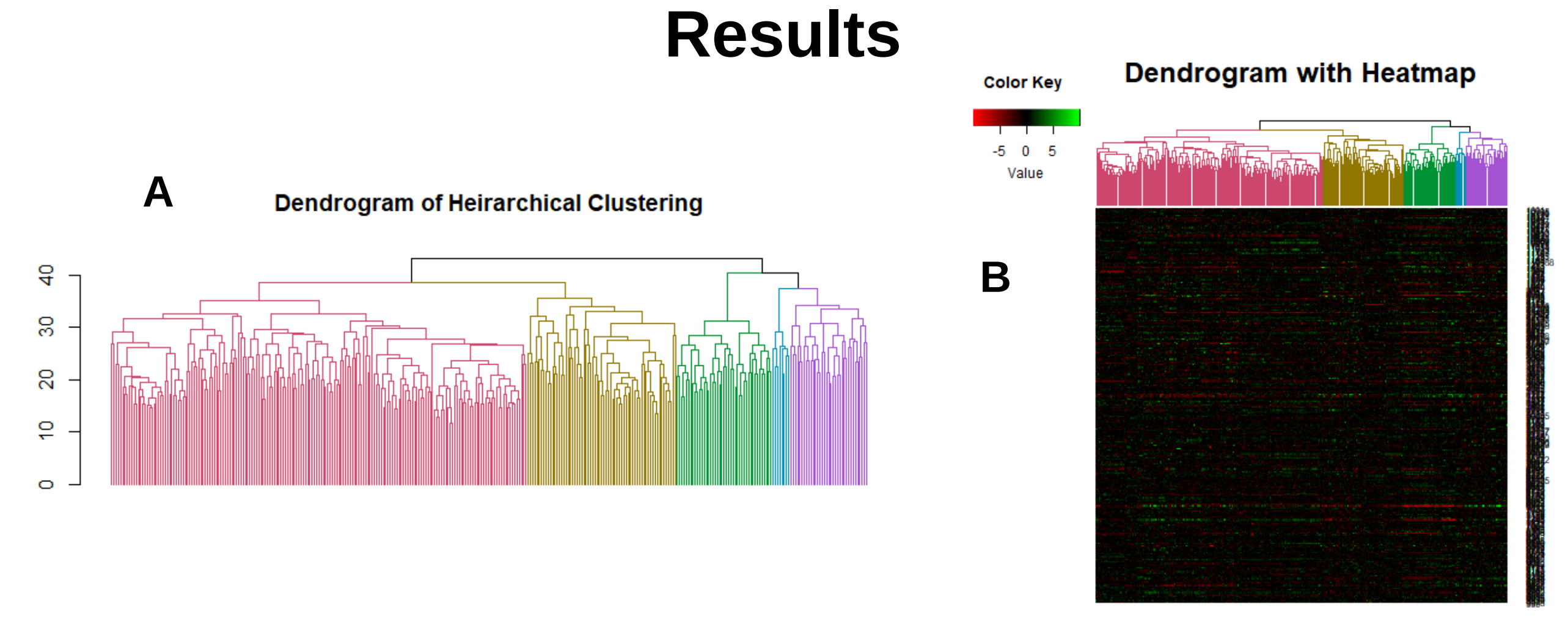


Fig 1: Hierarchical cluster analysis of the combined data set **(A)** Overview of complete cluster diagram. **(B)** Experimental sample-associated dendrogram.

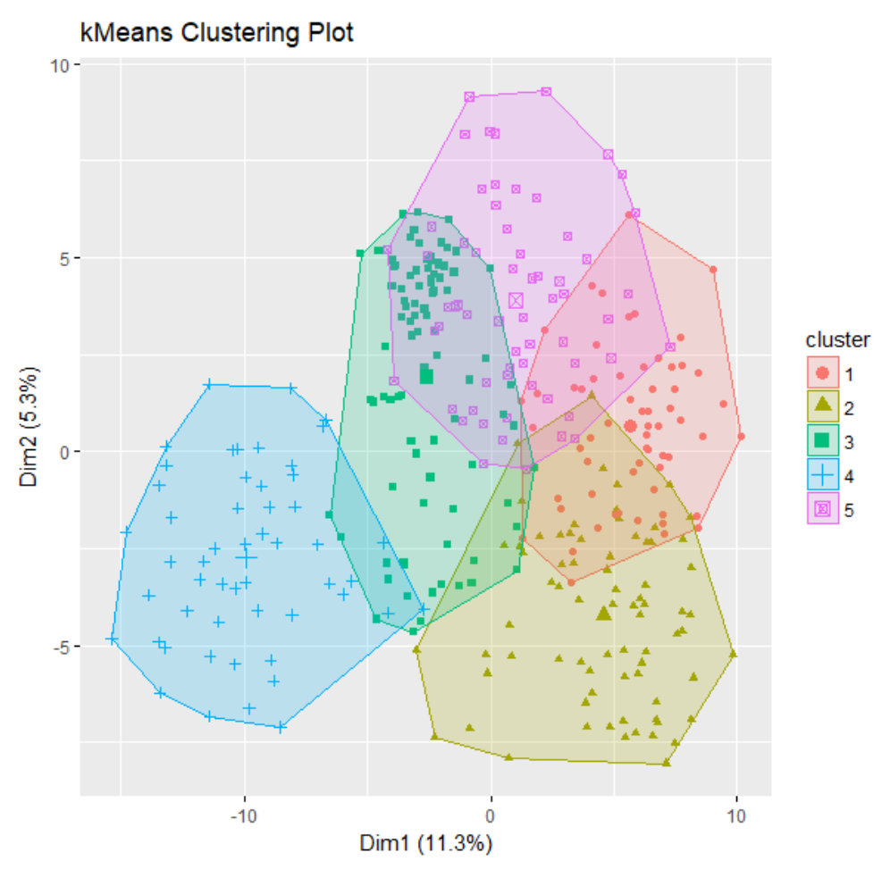


Fig 2: k-means cluster analysis of the combined data set

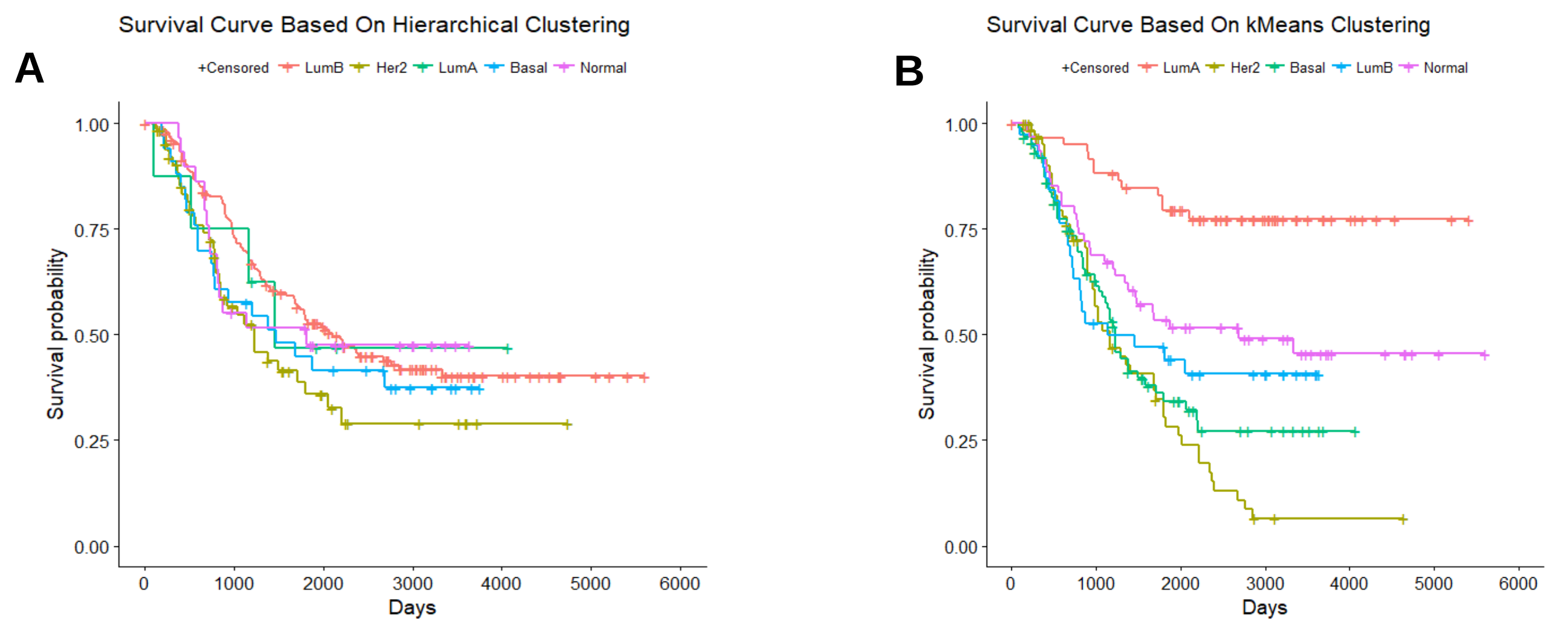


Fig 3: Kaplan-Meier survival curves of breast tumors classified by intrinsic subtype. Survival curves are shown for **(A)** the combined data set classified by hierarchical clustering and **(B)** the combined data set classified by k-means clustering

Method	coeff	Exp (coeff)	Se (coeff)	Z	P
Hierarchical Clustering	0.0429	1.0438	0.0541	0.79	0.43
k-means clustering	0.0924	1.0968	0.0516	1.79	0.074

Table 1: Cox Proportional-Hazards Model analysis for the Kaplan-Meier survival curves.

Discussion

- The development and validation of gene sets for cancer patients require significant resources because large training and test sets are required to achieve robust results and the DWD method can be a good solution of this.
- The groupings of the samples showed inter-dataset mixing and were significant predictors of outcome in univariate Kaplan-Meier and Cox analysis.
- In the Cox analysis, the covariates for k-means clustering is more significant (small p) then the hierarchical clustering. The p-value for k-means clustering is 0.0746, with a hazard ratio HR = exp(coef) = 1.09, indicating a strong relationship between the clusters and decreased risk of death.
- We could conclude that these classifications were successful in predicting outcome.

Conclusions

Using this study, we might have a much clearer picture about the genomes of breast cancer and can generate data about the intrinsic characteristics of a tumor, thereby providing useful diagnostic, prognostic and predictive information. What all this means for patients now is that ever more information is becoming available to help guide decisions about treatment.

Future Directions

Our future goal is to create a Continuous/Lifelong Learning model, which will learn continuously and adaptively from more and more gene expression sets. We are also interested in Transfer Learning, where we will use the knowledge of this work of breast tumor on a different kind of tumor. Transfer Learning model will help us to gain knowledge in an easier and faster way about different kind of tumors using the domain knowledge we get from this work.

References

[1] Zhiyuan Hu, Cheng Fan, Daniel S. Oh, J. S. Marron, Xiaping He, Bahjat F. Qaqish, Chad Livasy, Lisa A. Carey, Evangeline Reynolds, Lynn Dressler, Andrew Nobel, Joel Parker, Matthew G. Ewend, Lynda R. Sawyer, Junyuan Wu, Yudong Liu, Rita Nanda, Maria Tretiakova, Alejandra Ruiz Orrico, Donna Dreher, Juan P. Palazzo, Laurent Perreard, Edward Nelson, Mary Mone, Heidi Hansen, Michael Mullins, John F. Quackenbush, Matthew J. Ellis, Olufunmilayo I. Olopade, Philip S. Bernard, and Charles M. Perou. 2006. **The molecular portraits of breast tumors are conserved across microarray platforms**. BMC Genomics 7, (2006), 112. DOI:https://doi.org/10.1186/1471-2164-7-96.

[2] Therese Srlie, Charles M. Perou, Robert Tibshirani, Turid Aas, Stephanie Geisler, Hilde Johnsen, Trevor Hastie, Michael B. Eisen, Matt van de Rijn, Stefanie S. Jeffrey, Thor Thorsen, Hanne Quist, John C. Matese, Patrick O. Brown, David Botstein, Per Eystein Lnning, and Anne-Lise Brresen-Dale. 2001. **Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications**. Proc. Natl. Acad. Sci. 98, 19 (September 2001), 1086910874. DOI:https://doi.org/10.1073/PNAS.191367098.

[3] Christos Sotiriou, Soek-Ying -Y Neo, Lisa M McShane, Edward L Korn, Philip M Long, Amir Jazaeri, Philippe Martiat, Steve B Fox, Adrian L Harris, and Edison T Liu. 2003. **Breast cancer classification and prognosis based on gene expression profiles from a population-based study**. Proc. Natl. Acad. Sci. U. S. A. 100, 18 (2003). DOI:https://doi.org/10.1073/pnas.1732912100

[4] T. Srlie, R. Tibshirani, J. Parker, T. Hastie, J. S. Marron, A. Nobel, S. Deng, H. Johnsen, R. Pesich, S. Geisler, J. Demeter, C. M. Perou, P. E. Lonning, P. O. Brown, A.-L. Borresen-Dale, and D. Botstein. 2003. **Repeated observation of breast tumor subtypes in independent gene expression data sets**. Proc. Natl. Acad. Sci. 100, 14 (2003). DOI:https://doi.org/10.1073/pnas.0932692100

[5] Charles M Perou, Therese Srlie, Michael B Eisen, Matt Van De Rijn, Stefanie S Jeffrey, Christian A Rees, Jonathan R Pollack, Douglas T Ross, Hilde Johnsen, Lars A Akslen, I Fluge, Alexander Pergamenschikov, Cheryl Williams, Shirley X Zhu, Per E Lnning, Patrick O Brown, David Botstein, and Service Grant. 2000. **Molecular portraits of human breast tumours**. 533, May (2000), 747752. DOI:https://doi.org/10.1038/35021093.

We would like to thank Dr. William Eberle, Dr. Doug Talbert, Paul Joshua Tinker and Katherine E Brown for all of their help and support.