Super Partition: fast, flexible, and interpretable large-scale data reduction in R

Katelyn J. Queen[1], Malcolm Barrett[2] and Joshua Millstein[1]

[1] Department of Population and Public Health Sciences, University of Southern California, Los Angeles, California, United States
[2] Department of Health Policy, Stanford University, Stanford, California, United States

ABSTRACT

Motivation: As data sets increase in size and complexity with advancing technology, flexible and interpretable data reduction methods that quantify information preservation become increasingly important.

Results: Super Partition is a large-scale approximation of the original Partition data reduction algorithm that allows the user to flexibly specify the minimum amount of information captured for each input feature. In an initial step, Genie, a fast, hierarchical clustering algorithm, forms a super-partition, thereby increasing the computational tractability by allowing Partition to be applied to the subsets. Applications to high dimensional data sets show scalability to hundreds of thousands of features with reasonable computation times.

Availability and implementation: Super Partition is a new function within the partition R package, available on the CRAN repository (https://cran.r-project.org/web/packages/partition/index.html).

Subjects Bioinformatics, Genetics, Genomics, Computational Science, Data Science
Keywords Data reduction, Clustering, Big data

BACKGROUND

Dimensionality reduction methods reduce complexity in data with the goal to reduce dataset dimensionality while preserving information and global relationships between features, leading to minimal information loss to allow for better data visualization and more efficient data processing. Partition is an agglomerative approach that divides data into sets of related features, where each set can be reduced to a single feature that captures a user-specified minimum amount of information. It's a surjective approach in that each original feature maps to one and only one new feature, facilitating interpretability. The metric for information capture can be selected or designed by the user, *e.g.*, mutual information for non-normal distributions or the intraclass correlation coefficient (ICC) for more normally distributed data. It's also flexible in how features are reduced within a set, *e.g.*, mean, first principal component, *etc*. Under some conditions likely to be common in real data, described in detail in *Millstein et al. (2019)*, reduction with Partition can increase power to detect true associations as compared to principal component analysis (PCA), non-negative matrix factorization (NNMF), or no reduction. In comparison to algorithms like PCA, Partition preserves local data structure.

Submitted 29 April 2024
Accepted 4 November 2024
Published 27 January 2025

Corresponding author
Katelyn J. Queen, kjqueen@usc.edu

Academic editor
Pedro Ferreira

Additional Information and Declarations can be found on page 7

DOI **10.7717/peerj.18580**

However, the original implementation is limited by memory and run-time, due to the requirement that a dissimilarity matrix be computed that includes all feature pairs. Thus, the maximum size dataset is restricted to tens of thousands of features, with exact size limitations dependent upon the user's computing environment. In contrast, the Genie hierarchical clustering algorithm incorporates single linkage (*Gagolewski, Bartoszuk & Cena, 2016*), meaning that distances are computed between clusters rather than single features. Therefore, it is not subject to these constraints, allowing it to scale up to over one million features in reasonable run-time (*Gagolewski, 2021*). Although Genie was not designed as a data reduction method, it can be used to form super-partitions in a high-dimensional setting, allowing Partition to be applied to the components, yielding a computationally tractable reduction approach. We developed an approximation to Partition using Genie that scales the algorithm to big data, with run-time benchmarking for datasets up to almost one-half million features.

SOFTWARE IMPLEMENTATION

Portions of this text were previously published as part of a thesis (https://impa.usc.edu/asset-management/2A3BF1MGJ5LQF?FR_=1&W=1159&H=707).

In an initial step, Genie is used to form super-partitions of the data composed of $\lceil \frac{N}{c} \rceil$ components or clusters, where N is the number of features in the full dataset, and c is the user-defined maximum cluster size (default value = 4,000). For initial super-partitions with size greater than c, Genie is applied in an iterative process to constrain cluster sizes to below c. Genie allows users to supply what they refer to as the Gini index, which is an economic inequity measure (range: [0, 1]). This measure aims to prevent formation of clusters with imbalanced sizes. We choose a threshold of 0.05 to highly penalize the formation of small clusters. However, for clusters of size $N_k > c$, small clusters may still form if features are highly correlated. Thus, if a large cluster is broken into smaller parts, and the smallest is of size 50 or less, we reject the Genie partition at this step and instead use k-means (*MacQueen, 1967*) with $\lceil \frac{N_k}{c} \rceil$ centroids to reduce the large cluster into more evenly sized parts. Finally, the Partition algorithm is applied to each super-partition cluster. A consequence of the super-partition step is that the minimum number of features in the final dataset is $\lceil \frac{N}{c} \rceil$ or the initial number of super-partitions.

This process preserves the Partition capability of constraining information loss to a user-specified threshold on a local level. That is, information loss is constrained with respect to a limited number of related features within a cluster rather than the entire dataset. In contrast, if the user were to apply PCA and select the top principal components that explain some percentage of the variance of the data, information loss would be constrained on a global level.

APPLICATION

The computational efficiency of Super Partition was assessed and compared to Partition in several datasets, including transcriptome-wide microarray data in whole blood from the ABRIDGE and CAMP asthma cohorts (n = 865, number of features (f) = 19,428) (*Covar et al., 2012*; *Torgerson et al., 2011*), transcriptome-wide RNA-seq data in colon

adenocarcinoma tumor tissue from The Cancer Genome Atlas (TCGA) (n = 481, f = 60,660) (*Muzny et al., 2012*), and Illumina 450K DNA methylation (DNAm) array data from the ABRIDGE cohort. The DNAm data were analyzed separately by chromosomes 1, 2, and 3 (n = 705, f = 98,216) and the full methylome (n = 705, f = 449,580).

CAMP and ABRIDGE expression profiles were measured *via* the Illumina Human HT-12 v4 array. All gene expression profile data were normalized *via* a $\log_2$-transformation and quantile-normalization within tissue. Expression profiles showed no major differences by sequencing date or clinical site, indicating an absence of technical batch effects. As previously described by the TCGA Network, TCGA expression profiles were measured *via* the Custom Agilent 244K Gene Expression Microarray (*Muzny et al., 2012*), and the data underwent extensive quality control checks, with minimal evidence of batch effects. The data were Loess normalized and $\log_2$ transformed. Finally, the DNAm data were measured on the Infinium HumanMethylation450 BeadChip array. Samples with poor intensity, unexpected beta value distribution, or with more than 5% of sites with a detection p-value greater than 0.05, were excluded. Normalization was performed on raw beta values using methods detailed in *Liu & Siegmund (2016)*.

Partition and Super Partition were run 21 times each on the ABRIDGE whole blood gene expression data, using information loss criteria (the information loss criterion (ILC) threshold is the minimum ICC for any accepted cluster) in the range [0, 1] by 0.05 increments to allow for direct comparison of the methods. Changes in computation time and dimensionality reduction with varying maximum cluster sizes were measured by performing the Super Partition algorithm on the ABRIDGE whole blood gene expression data with an ILC of 0.10 and 0.60 at maximum cluster sizes of 250 and 500 to 4,500 in increments of 500. To test the scalability of the algorithm, Super Partition was applied to each of the three aforementioned datasets at the same 21 ILCs, all with a maximum cluster size of 4,000. Additionally, to test the large-scale capability of the algorithm, Super Partition was run across the full methylome (n = 705, f = 449,580) of the DNAm data again with a maximum cluster size of 4,000. All analyses were completed using the University of Southern California Center for Advanced Research Computing (CARC) cluster on a single Xeon-2640v3 2.60 GHz node with a single CPU and 16 GB of memory. Analyses were conducted using R version 4.2.3.

In a direct comparison to the base Partition algorithm across 19,428 features in the ABRIDGE whole blood gene expression dataset, Super Partition substantially outperformed Partition in computation time (Fig. 1A) across all thresholds. The time differences hits a maximum at an ILC of 0.35 and a minimum at the largest ILC values ($\text{ILC} \in [0.90, 1]$), where minimal dimension reduction occurs. The number of features in the reduced data for each algorithm are very similar for the entire range of correlation thresholds (Fig. 1B), showcasing that the approximation of the original Partition algorithm does not significantly alter the results. Using an ILC of 0.6 and a maximum cluster size of 4,000 (the function default values), Super Partition yields 15,348 features and Partition yields 15,253. Between the methods, 14,872, or 96.90%, of the features are the exact same, showcasing the similarity in the methods.

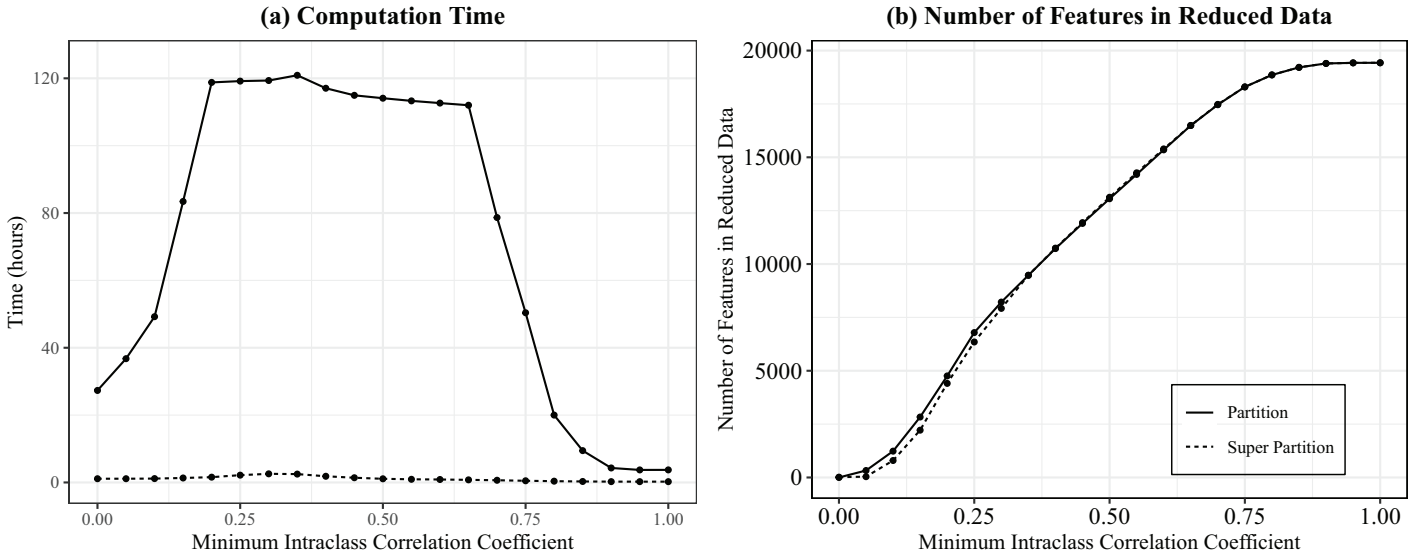

Figure 1 Comparing computation time and number of features in reduced data for partition and super partition. Plot showing (A) time in hours and (B) number of clusters identified after partitioning (solid line) and super partitioning (dashed line) data across varying thresholds for minimum intraclass correlation coefficients in the ABRIDGE whole blood gene expression data (n = 865, number of features = 19,428). The default super partition maximum cluster size of 4,000 was used.
Full-size DOI: 10.7717/peerj.18580/fig-1

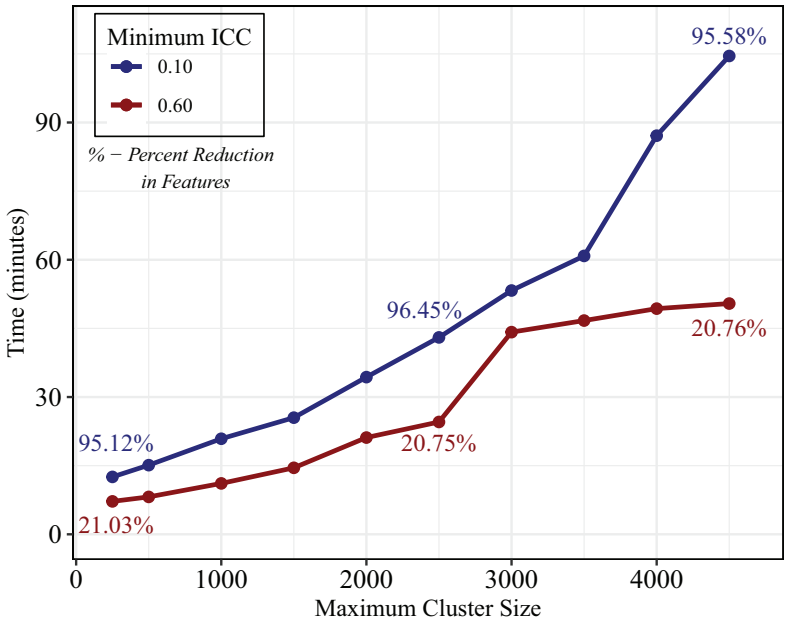

Figure 2 Time in minutes to partition ABRIDGE whole blood gene expression with varying maximum cluster sizes. Plot showing time in hours to complete super partition algorithm with an information loss criterion of 0.60 (red) and an information loss criterion of 0.10 (blue) for varying maximum cluster sizes in the ABRIDGE whole blood gene expression dataset. Percent reduction in features for the reduced data are displayed for selected maximum cluster sizes (250, 2,500, 4,500).
Full-size DOI: 10.7717/peerj.18580/fig-2

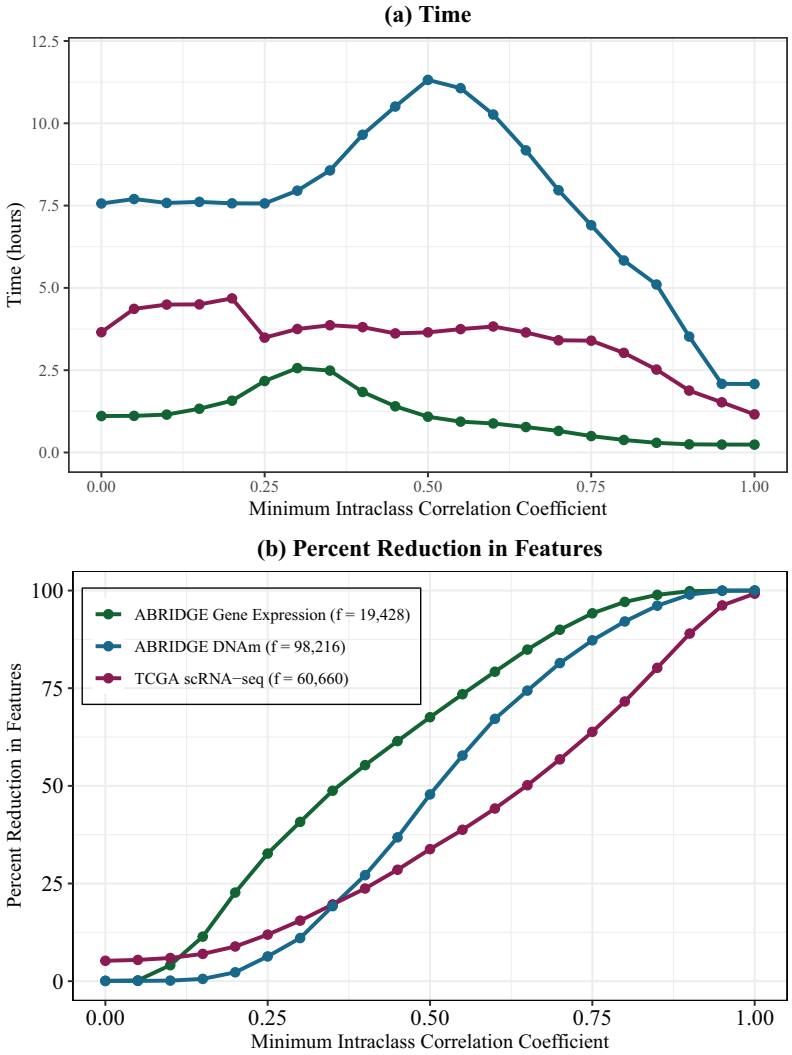

Figure 3 Time and data reduction of super partition in varying datasets. Plots showing (A) time in hours to complete the Partition algorithm and (B) percent reduction in features after partitioning data across varying thresholds for minimum intraclass correlation coefficients. ABRIDGE whole blood gene expression data is shown in green, TCGA scRNA-seq data in pink, and ABRIDGE whole blood 450 k DNA methylation in blue. Full-size DOI: 10.7717/peerj.18580/fig-3

Figure 2 shows the time in minutes to Super Partition the ABRIDGE whole blood gene expression data at ILCs of 0.10 and 0.60 for varying maximum cluster sizes (function default value = 4,000). Reasonably, as the maximum cluster size increases and thus the complexity of each super-partition, so does the computation time for both ILCs. The number of features in the reduced data for these cluster sizes range from 949 to 652 or 95.12% reduction in the number of features to 96.67% for an ILC of 0.10, and 15,342 to 15,396 or 21.03% reduction in features to 20.75% for an ILC of 0.60.

Figure 3A shows the time in hours to reduce the three datasets across the full range ([0, 1]) of ILCs. As expected, increasing numbers of features result in longer analysis times. Mid-range ILCs may be computationally expensive due to a large number of clusters and

large sizes of clusters being formed, requiring more algorithm steps. Figure 3B shows the percent reduction in the number of features after using Super Partition across the range of ILCs.

Additionally, the entire ABRIDGE 450k DNA methylation data (n = 705, number of features = 449,580) was partitioned with an ILC of 0.50, resulting in 62.75% reduction in features. This analysis took 4,374.40 min or just under 73 h to complete and required 45 GB of memory with a single CPU.

CONCLUSION

A limitation of the approach is that some sets of features may not be fully reduced. Feature reduction is only considered within each super-partition cluster, meaning that even if the information loss for combining super-partitions is below the user-specified information loss threshold, these clusters will not be combined. Additionally, given that current strategy to recombine results after applying Partition to each super-partition cluster requires sequential information, it is not feasible to run Super Partition in a multiprocessing environment. Future directions include a rework of this strategy to allow for parallel processing which could further reduce computational time.

The proposed solution approximates the original Partition method. That is, Genie is also an agglomerative algorithm, and it is solely used for an initial super-partition step, thus differences are likely to be modest. Information loss constraint and distance metric flexibility, feature mapping, and all other strengths and options of Partition are preserved in Super Partition. Application to real data shows that Super Partition greatly reduces computation time when directly compared to the original Partition algorithm and exponentially increases the number of features the algorithm is capable of handling.

We showed that maximum cluster size is inversely related to computation time, and as stated, there is a lower bound on the number of features which is dependent on maximum cluster size. Across the full range of maximum cluster sizes tested, there is a five-fold and eight-fold increase in computation times for ILCs of 0.60 and 0.10, respectively. However, the largest change in percent reduction in features occurs between a maximum cluster size of 250 and 2,500, and in this region, there is a 350% increase in computation time and less than a half-percent increase in percent reduction in features for an ILC of 0.60. Comparatively, for the analysis using an ILC of 0.10, there is also a 350% increase in computation time and a 27.29% increase in feature reduction. Therefore, we note the trade-off between dimension reduction and computation time, particularly for small ILCs. Users should consider analysis priorities when choosing these parameters.

The Super Partition algorithm is a fast, flexible way to reduce datasets while preserving information and interpretability of results, now feasible for datasets with hundreds of thousands of features.

ACKNOWLEDGEMENTS

The authors would like to thank the study participants.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

This work was supported by the National Institute of Environmental Health Sciences (T32ES013678 to Katelyn J. Queen); the National Heart Lung Blood Institute (R01HL118455 to Joshua Millstein); the National Institute on Aging (P01AG055367 to Joshua Millstein); the National Institute of Digestive and Kidney Disease (R01DK110793 to Joshua Millstein); and the National Cancer Institute (P01CA196569 to Joshua Millstein). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
National Institute of Environmental Health Sciences: T32ES013678.
National Heart Lung Blood Institute: R01HL118455.
National Institute on Aging: P01AG055367.
National Institute of Digestive and Kidney Disease: R01DK110793.
National Cancer Institute: P01CA196569.

Competing Interests

The authors declare that they have no competing interests.

Author Contributions

- Katelyn J. Queen conceived and designed the experiments, performed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.
- Malcolm Barrett analyzed the data, authored or reviewed drafts of the article, and approved the final draft.
- Joshua Millstein conceived and designed the experiments, authored or reviewed drafts of the article, and approved the final draft.

Data Availability

The following information was supplied regarding data availability:

The example code is available in the Supplemental File. The Super Partition algorithm code is available at GitHub and CRAN:

- https://github.com/USCbiostats/partition/blob/master/R/super_partition.R.
- 10.32614/CRAN.package.partition.

The third-party datasets are available at:

- GEO: GSE22324.
- TCGA: https://portal.gdc.cancer.gov/projects/TCGA-COAD.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj.18580#supplemental-information.

REFERENCES

Covar RA, Fuhlbrigge AL, Williams P, Kelly HW, Childhood Asthma Management Program Research Group. 2012. The childhood asthma management program (CAMP): contributions to the understanding of therapy and the natural history of childhood asthma. *Current Respiratory Care Reports* **1(4)**:243–250 DOI 10.1007/s13665-012-0026-9.

Gagolewski M. 2021. genieclust: fast and robust hierarchical clustering. *SoftwareX* **15(26)**:100722 DOI 10.1016/j.softx.2021.100722.

Gagolewski M, Bartoszuk M, Cena A. 2016. Genie: a new, fast, and outlier-resistant hierarchical clustering algorithm. *Information Sciences* **363**:8–23 DOI 10.1016/j.ins.2016.05.003.

Liu J, Siegmund KD. 2016. An evaluation of processing methods for HumanMethylation450 BeadChip data. *BMC Genomics* **17**:469 DOI 10.1186/s12864-016-2819-7.

MacQueen J. 1967. Some methods for classification and analysis of multivariate observations. In: *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*. Vol. 1. Oakland, CA, USA, 281–297.

Millstein J, Battaglin F, Barrett M, Cao S, Zhang W, Stintzing S, Heinemann V, Lenz H-J. 2019. Partition: a surjective mapping approach for dimensionality reduction. *Bioinformatics* **36(3)**:676–681 DOI 10.1093/bioinformatics/btz661.

Muzny DM, Bainbridge MN, Chang K, Dinh HH, Drummond JA, Fowler G, Kovar CL, Lewis LR, Morgan MB, Newsham IF, Reid JG, Santibanez J, Shinbrot E, Trevino LR, Wu Y-Q, Wang M, Gunaratne P, Donehower LA, Creighton CJ, Wheeler DA, Gibbs RA, Lawrence MS, Voet D, Jing R, Cibulskis K, Sivachenko A, Stojanov P, McKenna A, Lander ES, Gabriel S, Getz G, Ding L, Fulton RS, Koboldt DC, Wylie T, Walker J, Dooling DJ, Fulton L, Delehaunty KD, Fronick CC, Demeter R, Mardis ER, Wilson RK, Chu A, Chun H-JE, Mungall AJ, Pleasance E, Robertson AG, Stoll D, Balasundaram M, Birol I, Butterfield YSN, Chuah E, Coope RJN, Dhalla N, Guin R, Hirst C, Hirst M, Holt RA, Lee D, Li HI, Mayo M, Moore RA, Schein JE, Slobodan JR, Tam A, Thiessen N, Varhol R, Zeng T, Zhao Y, Jones SJM, Marra MA, Bass AJ, Ramos AH, Saksena G, Cherniack AD, Schumacher SE, Tabak B, Carter SL, Pho NH, Nguyen H, Onofrio RC, Crenshaw A, Ardlie K, Beroukhim R, Winckler W, Getz G, Meyerson M, Protopopov A, Zhang J, Hadjipanayis A, Lee E, Xi R, Yang L, Ren X, Zhang H, Sathiamoorthy N, Shukla S, Chen P-C, Haseley P, Xiao Y, Lee S, Seidman J, Chin L, Park PJ, Kucherlapati R, Auman JT, Hoadley KA, Du Y, Wilkerson MD, Shi Y, Liquori C, Meng S, Li L, Turman YJ, Topal MD, Tan D, Waring S, Buda E, Walsh J, Jones CD, Mieczkowski PA, Singh D, Wu J, Gulabani A, Dolina P, Bodenheimer T, Hoyle AP, Simons JV, Soloway M, Mose LE, Jefferys SR, Balu S, O'Connor BD, Prins JF, Chiang DY, Hayes DN, Perou CM, Hinoue T, Weisenberger DJ, Maglinte DT, Pan F, Berman BP, Berg DJVD, Shen H Jr, Triche T, Baylin SB, Laird PW, Getz G, Noble M, Voet D, Saksena G, Gehlenborg N, DiCara D, Zhang J, Zhang H, Wu C-J, Liu SY, Shukla S, Lawrence MS, Zhou L, Sivachenko A, Lin P, Stojanov P, Jing R, Park RW, Nazaire M-D, Robinson J, Thorvaldsdottir H, Mesirov J, Park PJ, Chin L, Thorsson V, Reynolds SM, Bernard B, Kreisberg R, Lin J, Iype L, Bressler R, Erkkilä T, Gundapuneni M, Liu Y, Norberg A, Robinson T, Yang D, Zhang W, Shmulevich I, Ronde JJd, Schultz N, Cerami E, Ciriello G, Goldberg AP, Gross B, Jacobsen A, Gao J, Kaczkowski B, Sinha R, Aksoy BA, Antipin Y, Reva B, et al. 2012. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487**:330–337 DOI 10.1038/nature11252.

Torgerson DG, Ampleford EJ, Chiu GY, Gauderman WJ, Gignoux CR, Graves PE, Himes BE, Levin AM, Mathias RA, Hancock DB, Baurley JW, Eng C, Stern DA, Celedón JC, Rafaels N, Capurso D, Conti DV, Roth LA, Soto-Quiros M, Togias A, Li X, Myers RA, Romieu I, Berg DJVD, Hu D, Hansel NN, Hernandez RD, Israel E, Salam MT, Galanter J, Avila PC,

Avila L, Rodriquez-Santana JR, Chapela R, Rodriguez-Cintron W, Diette GB, Adkinson NF, Abel RA, Ross KD, Shi M, Faruque MU, Dunston GM, Watson HR, Mantese VJ, Ezurum SC, Liang L, Ruczinski I, Ford JG, Huntsman S, Chung KF, Vora H, Li X, Calhoun WJ, Castro M, Sienra-Monge JJ, Rio-Navarro Bd, Deichmann KA, Heinzmann A, Wenzel SE, Busse WW, Gern JE, Lemanske RF, Beaty TH, Bleecker ER, Raby BA, Meyers DA, London SJ, Mexico City Childhood Asthma Study (MCAAS), Gilliland FD, Children's Health Study (CHS) and HARBORS Study, Burchard EG, Genetics of Asthma in Latino Americans (GALA) Study, Study of Genes-Environment and Admixture in Latino Americans (GALA2) and Study of African Americans, Asthma, Genes & Environments (SAGE), Martinez FD, Childhood Asthma Research and Education (CARE) Network, Weiss ST, Childhood Asthma Management Program (CAMP), Williams LK, Study of Asthma Phenotypes and Pharmacogenomic Interactions by Race-Ethnicity (SAPPHIRE), Barnes KC, Genetic Research on Asthma in African Diaspora (GRAAD) Study, Ober C, Nicolae DL. 2011. Meta-analysis of genome-wide association studies of asthma in ethnically diverse North American populations. *Nature Genetics* **43(9)**:887–892 DOI 10.1038/ng.888.