COMPRESSING HAPMAP DATA USING SUFFICIENT SUBPOPULATION MARKOV CHAIN STATISTICS

Kevin Cummiskey, Caleb Lareau, Matthew Ploenzke

Contact: caleblareau@g.harvard.edu

ABSTRACT

As the cost of sequencing has significantly dropped in recent years, scientists have observed a corresponding explosion of data explaining the variation in DNA, especially in humans. To address the need for an efficient storage algorithm, we evoke the theory of sufficient statistics and the theoretical basis of population-level imputation to develop a novel storage algorithm. We hypothesize that generating Markov Chain summary statistics would 1) provide an efficient storage approach and 2) retain the population-level properties of our sequencing data. To access the efficacy of our novel paradigm, we analyze data from chromosome 22 of the HapMap project. While our initial results demonstrate a substantial benefit in the storage and efficiency of processing these data, our initial results suggest that simple transition probabilities are inadequate to fully regenerate population substructure.

BACKGROUND

Recent advances in sequencing technology from companies like Illumnia have significantly reduced the cost inferring the variation in the human genome. While sequencing the original human genome cost nearly $300 million dollars, just 15 years later, this cost has been cut by orders of magnitude to only ~$1,000. Consequently, a new problem has emerged where computational scientists must develop novel approaches to storing and handling the volume of data generated from these experiments.

The innovation presented in this product is motivated the principle of sufficient statistics. At a high level, we note that sufficient statistics allow for one to "throw away" the raw data and retain only summary statistics that enables the reconstruction of the distributional data. In our case, we hypothesized that we could make inferences about the subpopulations defined in HapMap using only summary statistics of the variants, which would result in a considerably smaller data structure to be stored, loaded, and shared.

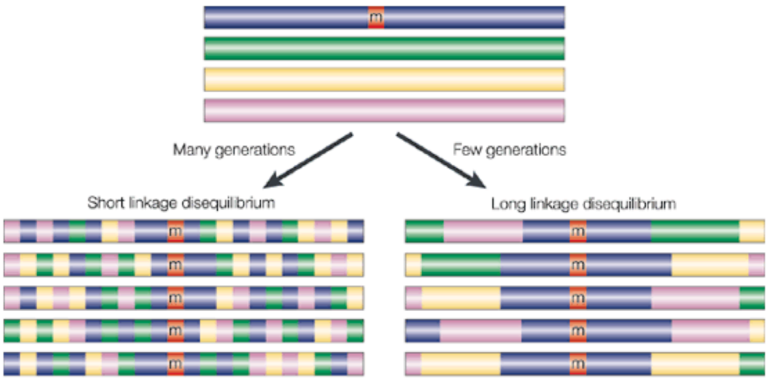

FIGURE 1. Schematic overview of linkage disequilibrium. As genetic recombination in certain regions occurs non-randomly, present individuals are observed with haplotypes that are a composition of ancestral genomes. We motive our storage algorithm with this biological observation that also serves as the basis for the IMPUTE software (Marchini et al.; Ostrer).

Though sufficient statistics should in theory provide an efficient solution to this computational problem, the challenge is to determine what statistics would in fact be sufficient to determine the distributional parameters of genetic data. To attempt to define these, we considered the principles of genetic recombination and linkage disequilibrium. As FIGURE 1 shows, later generations of individuals tend to be combinations of ancestral regions of DNA. The popular software IMPUTE uses this theory to suggest that individuals with only a small proportion of variants tagged can recover much of the missing data by using phased haplotype data from a reference panel. The theory behind this software is that the individuals with missing data are some combination of the reference haplotypes that are mixed through this process of genetic recombination. The mixture from the reference panels is modeled as a Markov Chain process where at each variant; the individuals with missing data have a non-zero probability of transitioning from one haplotype to a different haplotype in the reference panel. Under this framework, the transitions between variants seem to be sufficient in determining the full genetic data given a large enough reference panel. Thus, for our algorithm, we hypothesized that transition probabilities themselves would be sufficient to infer the genetic data that we aimed to compress.

METHODS

Our algorithm is implemented in both R and Julia and supporting code for principle component analysis is implemented in python. All code is made publically available (see `github` repository under the **AVAILABILITY** section). We've coined our approach as the "Sufficient Markov Chain Transition" (SMCT) algorithm, which consists of two main steps. First, we compute transition probabilities for each variant within each ethnicity, which compresses the data by a factor of k/m, where k is the number of subpopulations and m is the number of individuals. Second, we regenerate the data at a distributional level using these Markov Chain probabilities. Since SMCT needs to pass through the data once to establish the transition probabilities, our algorithm has a runtime complexity of *O*(n). Further details of the implementation of our algorithm are provided in our presentation and the other authors' contributions.

RESULTS

As the primary computational cost is in loading and storing the variant matrix, we report results from the Julia implementation, which has a more efficient interface for working with this type of problem than R. Our simulated simulation analysis was executed on a 1.7 GHz Intel Dual-Core i7 Processor with 8 GB RAM. While the raw test file exceeded 3.6GB unzipped, the burden of the compressed version was dramatically less than the raw data, 2% (72.4MB). In the absence of "thinning" (which we define as collapsing/averaging the transition probabilities), our reduced file structure achieved (146.6MB). While thinning to every other transition probability halved the file size, we observed a 140% runtime increase. Runtime was around a 10-fold increase to that of standard zip compression. Though our algorithm has a non-trivial cost of compression, the resulting structure can be easily loaded and handled in secondary analyses.

To determine the validity of our hypothesis, we tested whether we could regenerate a principle component mapping of our subpopulations that could be derived from the raw data. FIGURE 2 shows the results of our principle components analysis to test the validity of our hypothetical sufficient statistics. These plots suggest that, unfortunately, our initial attempt at determining sufficient statistics was inadequate.

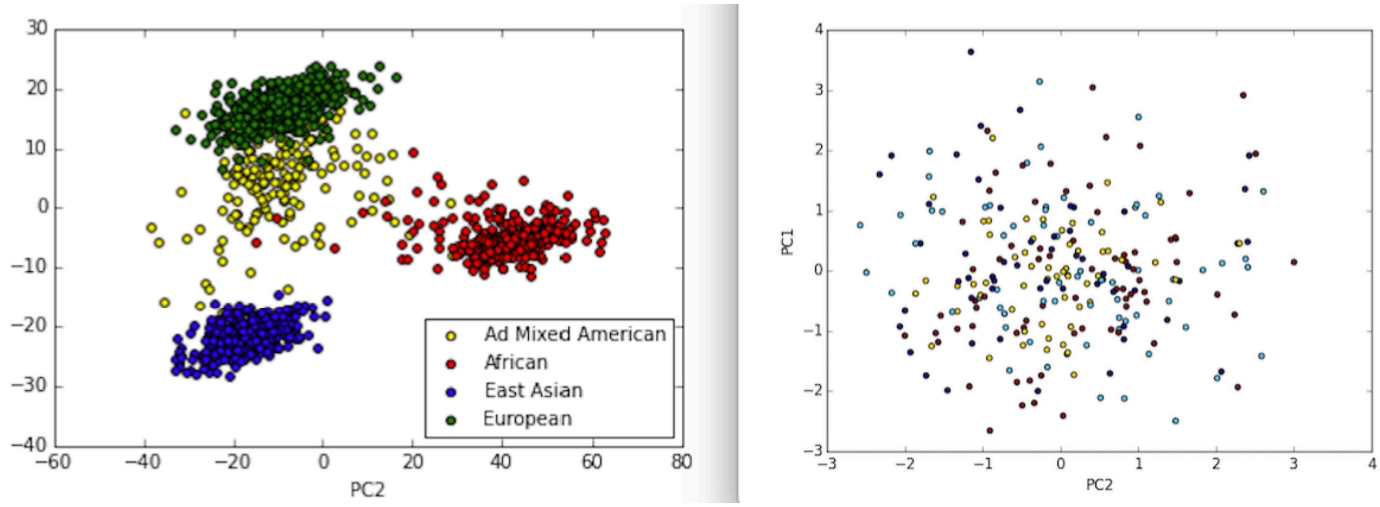

FIGURE 2. A comparison of the principle component plots from (A) raw data and (B) Markov Chain Statistics. While the original data was able to define clear clusters in the superpopulations, our algorithm failed to recapture this structure in our initial analysis.

CONCLUSIONS

- Our initial attempt to summarize the HapMap data using sufficient demonstrated a marked improvement in storage efficiency and computational burden.
- However, the determined transition probabilities were not sufficient to recreate population substructure.
- Our population-level summaries enable the easier sharing of genetic data as individual-level information is no longer a feature. Thus, our novel structure has considerable implications in the ability to share genetic data.

Future Directions

- Trying variants of the transition probability statistics (e.g. conditional probabilities) to determine truly sufficient statistics.
- Implementation in GWAS data to regional associations with disease using probability weighting metrics.
- Retaining a few "marker" variants that aren't sufficient to identify an individual but may boost power to infer the haplotype structure (similar to IMPUTE).

References

- Casella G, Berger RL. Statistical inference. Pacific Grove, CA: Duxbury; 2002.
- Golub, Gene H., and Charles F. Van Loan. Matrix computations. Vol. 3. JHU Press, 2012.
- Jolliffe, Ian. Principal component analysis. John Wiley & Sons, Ltd, 2002.
- Marchini J, Howie B, Myers S, McVean G, Donnelly P. A new multipoint method for genome-wide association studies by imputation of genotypes. Nature genetics. 2007 Jul 1;39(7):906-13
- Ostrer H. A genetic profile of contemporary Jewish populations. Nature Reviews Genetics. 2001 Nov 1;2(11):891-8.

Availability

`https://github.com/caleblareau/BIO234FinalProject`

Notes

My primary responsibility for the group was developing the background structure. Hence, my report emphasizes this section as well as the implications of our algorithm, should we create a successful sufficient implementation. From working with Matt and Kevin, we're optimistic that we could find sufficient statistics in this framework and achieve our goals of recreating population-level inferences if we had more time to play with the project.