Aims\\

Update on current, and motivation

Issues we could solve, with what technique

specific details of technique, buts, issues we could solve

Method\\

What we implement

How our technique differs from present literatures

Data \\

Data, techniques

What we did with data

What kind result we got

Results \\

Result we found

Some problems we encountered

What our technique will enable people to do

Conclusion\\

91 Advances in sequencing technologies have significantly lowered the barriers to collecting large amounts of human DNA data for association studies, including rarer variation not usually genotyped with prior methods. (Update on current, and motivation) Identifying population structure is an essential part of these studies to minimise confounding, for which principal components analysis (PCA) is a common technique. (Issues we could solve, with what technique) Rare variants have been thought to characterise finer-scale population structure, but classical PCA methods are unsuitable for the large matrix sizes collected and the optimal weighting of rare variants is not clear. (specific details of technique, buts, issues we could solve) In this thesis, we investigate these issues.

58 We implement two efficient stochastic algorithms and one streaming algorithm for performing PCA in R, and compare these to classical algorithms in terms of accuracy and speed. What we implement Accuracy comparisons assess the resulting PCAs' similarity to results from the classical methods in terms of both overall subspace and individual dimensions, whereas present literature only examine one or the other. How our technique differs from present literatures

104 To investigate the effect of rare variants on population structure estimates from PCA, we use three weightings for individual variants (unit-variance, uniform, and Beta distribution) on a simulated DNA sequence and data from the 1000 Genomes Project. Data, techniques The ability of each weighting to recover the true population structure using either the entire sequence or subsets of rare or common variants is compared. What we did with data Each set of principal components are used as covariates in association tests of a phenotype based on population membership and a set of variants, and the effectiveness of these at reducing confounding is compared to using true population membership as a covariate. What kind result we got

86 We find the stochastic and streaming algorithms provide significant reductions in the computational burden of PCA compared with classical algorithms while obtaining essentially identical results given appropriate parameters. Result we found Recent demographic events are unable to be identified with common variants alone in simulations. Some problems we encountered However, we find rare variants distort principal components extracted from the 1000 Genomes dataset while enabling the identification of an outlier. The inclusion of rare variants and using unit-variance or Beta distribution weights provide covariates closest to using true population membership in association tests. What our technique will enable people to do

339