International Journal of Population Data Science

Journal Website: www.ijpds.org

Linking Genomic Data with Phenotypes Derived from Electronic Health Records

Appleby, Phil¹

¹University of Dundee

Objectives

To build a searchable database for SNP array data from the Go-DARTS data set, in which a combined view of genotype data derived from multiple assay platforms can be extracted for both candidate gene and GWA studies and to combine this with a database of phenotype descriptors which are saved as shareable, reusable database objects and which persist beyond the lifetime of any analysis script.

To build databases and software solutions which can be made readily available to laboratories and academic institutions which may not have the resources to adopt one of the larger Genotype / Phenotype integration solutions.

Electronic Health Records are saved as reusable, shareable and modifiable database objects, allowing rapid retrieval of phenotype data.

Conclusion

The ability to access Genomic data from multiple assay platforms and to use this in conjunction with shareable libraries of phenotype objects allows rapid access to data for analysis using both Genomic SNP Array data and linked Electronic Health Records. Analysis on data extracted from our linked databases should proceed more rapidly and should be more easily reproducible.

Approach

Two databases were built. The first is a hybrid Genomics one in which variant and study subject data are stored in a database with variant detail data retained in Variant Call Format (VCF) files. The second database saves phenotype descriptors as shareable, modifiable database objects alongside a table of events derived from the set of available Electronic Health Records (EHRs). All detail from the EHRs is also retained in the database which is delivered on a project by project basis using virtual machines.

Results

Traditionally the process of deriving genotype and phenotype data for epidemiological studies can be a laborious one with genotype data being retrieved from large, flat data files and phenotypes being defined by codes in flat EHR records which are tested and filtered in scripts, written for analysis in a statistical package such as Stata, SPSS or R.

In our solution, genotype data can be retrieved in seconds and delivered to the users' desktops. Similarly lists of cases and controls can be downloaded based on saved or transient phenotype descriptors. Phenotypes descriptors derived from codes in

*Corresponding Author:

Email Address: pappleby@dundee.ac.uk (P. Appleby)