Course: Estimation of relatedness from genotyping by sequencing data

Content

Participants will estimate relatedness from genotyping by sequencing data using the unbiased KGD method (Dodds et al, submitted, http://dx.doi.org/10.1101/025379) using R software. Either an example data set or attendee's own data can be used. The workshop will investigate parentage checking and assignment, diversity (PCA plots) and saving the relationship matrix for downstream analyses.

Requirements

This course will be conducted with R software. Participants are expected to bring a laptop with R software (preferably the 64-bit and at least version 3.0.0) installed, and to have downloaded the course materials (see below).

Download R code, example files and instructions (for using the R code and preparing data) from: https://github.com/AgResearch/KGD by clicking the "Download ZIP" button at the bottom of the right hand panel.

Participants may bring their own GBS data, by following the guidelines in the documentation (GBScode.pdf). The datafile can be processed directly, and more efficiently, from gzip (.gz) format.

The example file consists of 96 individuals with 14,709 SNPs called with mean depth 9.3. The runtime for this dataset on a desktop PC (3.4GHz I7 processor) is around 11s, while around 120Mb of memory is used for data storage within R. Runtime is likely to be approximately linear with the number of SNPs and quadratic with the number of individuals, i.e. would be ~ 22s for 30,000 SNPs (96 individuals) and ~ 44s for 198 individuals (15,000 SNPs). Memory use is likely to be linear with both the number of individuals and number of SNPs. If participants bring their own data they should aim for a dataset that can be processed in 5-10 minutes.

Outline of coursework

Prepare directories	Unzip files Create a new directory (e.g. Rerun) to re-run the example KGD-master directory: GBSRun.R Copy to Rerun Copy input files from Example to Rerun HapMap.hmc.txt.gz Ped-GBS.csv Ped-Groups.csv
Prepare GBSRun.R	Open R Set working directory to Rerun Open GBSRun.R in editor Modify directory for sourcing code source("**..**/GBS-Chip-Gmatrix.R") Run first 2 lines of code
Calculate relatedness	Run next line of code Gfull <- calcG()
Calculate relatedness with SNP filter	GHWdgm.05 <- calcG(which(HWdis > -0.05),"HWdgm.05", npc=4) str(GHWdgm.05)
Check recorded pedigree	pedfile <- "Ped-GBS.csv"
Find parents	groupsfile <- "Ped-Groups.csv"
Check recorded pedigree / Find parents	rel.thresh <- 0.2 GCheck <- "GHWdgm.05$G5" source("../GBSPedAssign.R")
Customising	Colour points in the PCA by family Redo the PC2 vs PC1 plot outside of calcG, with different symbols, sizes etc Plot the G matrix ordered by family (using image)

Resources

Dodds *et al.* (2015) Construction of relatedness matrices using genotyping-by-sequencing data. *(BMC Genomics, submitted)*

Preprint: http://www.biorxiv.org/content/early/2015/08/24/025379

R code: https://github.com/AgResearch/KGD