Genomic data and bioinformatics

Jason Hodgson

6/11/2017

1 Data

DNA data is easy to display visualise.
Biggest problem at the moment is that there is so mcuh data.
Up to 900 billion bases in a single experiment now.
Often presented in FASTA files:

¿Taxa1
ATCGTAGCTACGTTTACCAGAAC
GGATCATTATTCTATATGCGGGA
etc.

FASTQ files:
DNA or RNA sequence data, includes a quality score.

VCF (variant call format)
Just variable sites mapped to a genome build. Meta data including some quality information.

PLINK
SNP genotype data. Either 2 or 3 files per a dataset, a genotype file, a family file and a marker file.

2 How to analyse the data

Common software for genomic analysis in R: Genetic packages (popgen etc).
Advantages: easy to integrate with other statistical analysis. Excellent plot-

ting capabilities.
Disadvantages: Extremely memry intensive. Often not possible with very large datasets. Slow.

Because of this lots of stand alone programmes.
PLINK: SNP data, designed for GWAS(genome wide association study) Powerful, fast, supports large datasets. Basic population genetics. Many models for testing genotype, phenotype associations.

Admixture: SNP or microsatellite data, model based method for inferring population structure and admixture proportions. Cam handle large datasets.

Others: ALDER (admixture dating using LD), CHROMOPAINTER (local ancestry assignment), SAMSTOOLS (NGS data), USEARCCH (analysing metagenomic data), TREEMIX (inferring migrations) plus many more.

Work in UNIX - fast, simple and repeatable.
Script your analysis: automates it making it faster to run, makes it faster to repeat, makes a record of exactly what was done. REPEATABILITY.