ScatterShot; a Java Program For Creating Cluster Plots from Affymetrix and Illumina Genotype Data

N.W. Rayner[1,2,3], N. Robertson[1,2], M.I. McCarthy[1,2,4]

1) WTCHG, University Oxford, Oxford, United Kingdom; 2) OCDEM, University Oxford, United Kingdom; 3) Wellcome Trust Sanger Institute, Hinxton, Cambridge, United Kingdom ;4) NIHR Oxford Biomedical Research Centre, Churchill Hospital, Oxford, United Kingdom

Contact: wrayner@well.ox.ac.uk

Introduction

- With the ever larger number of genomes being sequenced there is increased interest in rarer variation (<1% minor allele frequency (MAF)).
- The number of these polymorphisms appearing on the newer genotyping chips has grown dramatically. (**Figure 1**).

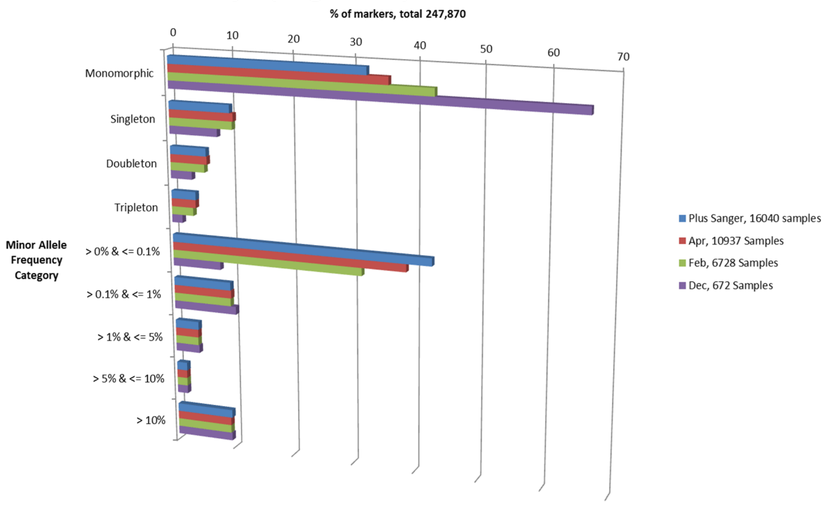

Figure 1: MAF profile of the variants from the Illumina Exome chip with up to ~16,000 UK samples.

- This growth, and the uncertainty around the efficacy of calling of these variants, has greatly expanded the number of genotype cluster plots that need to be examined.
- To date cluster plotting programs have been relatively inflexible in the file formats used, often requiring extensive reformatting before plotting or the original raw data, which may no longer be available.
- In addition the number of display options available is relatively limited.

Scattershot

- To address this issue we developed a new cluster plotting program, ScatterShot, that:
  1. Is simple and fast to run.
  2. Is flexible in the data input formats required.
  3. Has a greater range of display options available.

Simple and Fast to Run

- ScatterShot is run from the command line and has been developed in Java providing support for multiple OS platforms (Windows, Unix, Mac OSX).
- The program makes use of Java's asynchronous Input/Output (NIO) libraries, and has a novel architecture that can scale to make use of available processors.

File Formats

- A variety of input formats are supported, including standardised formats:
  1. Binary .chp files from Affymetrix's genotyping console.
  2. Final Report from Illumina's Genome Studio.
- Also supported is a generic format, which requires two separate files.
  1. Genotype calls in ped or binary ped file format.
  2. XY coordinates.
- The XY coordinate data format is very flexible.
  1. Data can be ordered as a row per SNP or row per sample.
  2. Extra leading columns and rows can be specified to be skipped. (**Figure 2**)
  3. The order of the SNPs and samples can differ between the two files.
  4. SNP and sample identifiers can differ between the two files; a mapping file can be used to link the two (**Figure 3**).
  5. Remapping can be for all or just some of the SNPs or samples.
  6. All input files can be gzipped.
- Run options are specified in a properties file (**Figure 2**)

```
xy_stem
/big2/HumanCoreExome/human_core_exome
xy_suffix txt
xy_snp_inclusion /home/snp
xy_ind_inclusion /home/samples
xy_row_offset 1
xy_column_offset 2
xy_order SNP_MAJOR
allele_stem
/big2/HumanCoreExome/human_core_exome.zcall
IOProperties.OUT_DIRECTORY
/home/clusters
```

Figure 2: An Example of the properties file used to control all aspects of the program including file location, SNP per row or column and the number of columns and rows to skip in the XY data.

7684333036_R02C02 SL00393	exm847453 rs148923672
7684333036_R03C01 SL01133	exm847457 rs200213822
7684333036_R03C02 SL00535	exm847464 rs141069497
7684333036_R04C01 SL00630	exm847469 rs199733612
7684333036_R04C02 SL01386	exm847471 rs74153530
7684333036_R05C01 SL00720	exm847475 rs145270262
7684233036_R05C02 SL01259	exm847478 rs199557224

Figure 3: Examples of Sample and SNP identifier mapping lists used to match files with differing identifiers, say after the ped file has been full processed and QC'd.

Display Options

- Plots are output using Scalable Vector Graphics (SVG) which can be viewed directly using modern HTML5 web browsers, or statically rendered into images files (JPEG, TIFF) or PDF files for printing. (**Figure 4**)

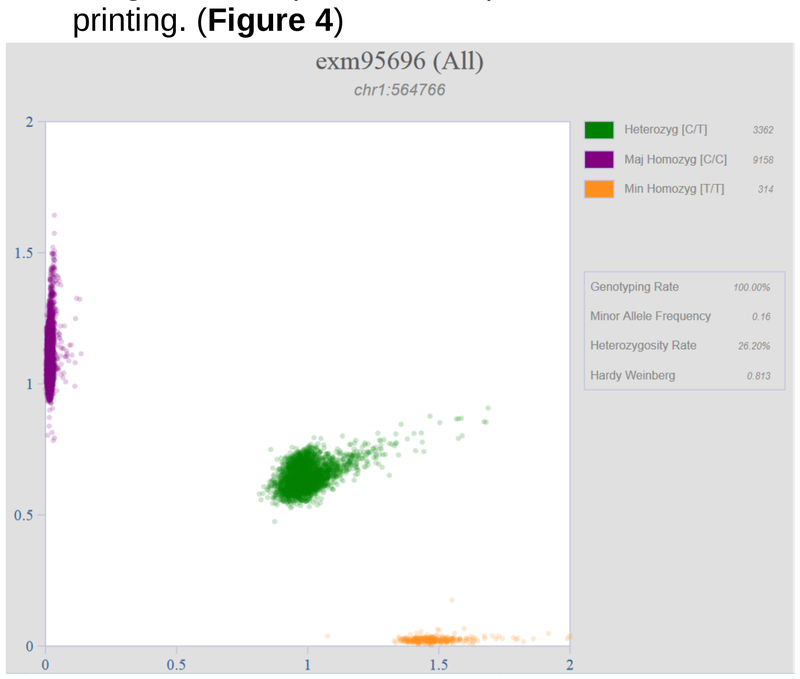

Figure 4: Example cluster plot of exome chip variant exm95696.

- The plots use a consistent colouring scheme tied to the genotype call (**Figure 5**)

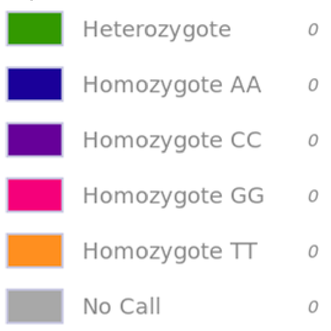

***Figure 5*:** legend from the cluster plots, Homozygote calls are green regardless of the genotype calls, fixed colours are defined for AA, CC, GG and TT calls.

- As an XML-based format, SVG offers the opportunity to embed other information such as gender and cohort. The SVG plots can then be filtered on these annotations.
- Filtering can be, for example, showing the plots for males (**Figure 6**) and females (**Figure 7**) or cases (**Figure 8**) and controls (**Figure 9**) separately.
- Other options could be removing individuals failing quality control or selecting one cohort out of many.

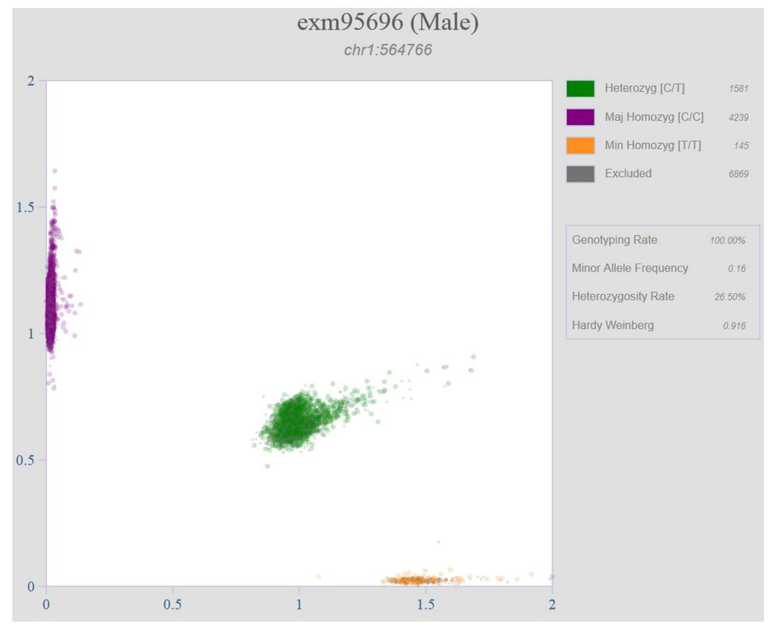

Figure 6: Cluster plot for exm95696 filtered for males only, females are marked as excluded using the smaller grey dots.

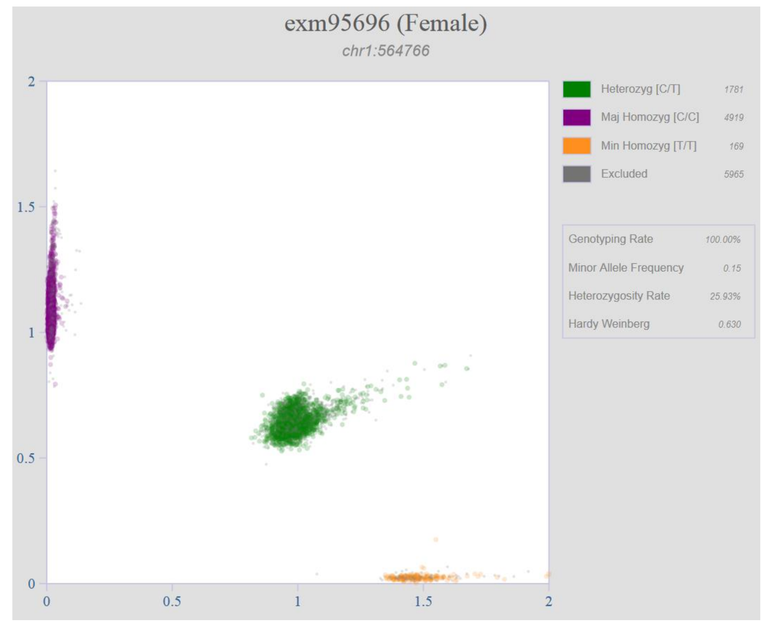

Figure 7: Cluster plot for exm95696 filtered for females only, males are marked as excluded.

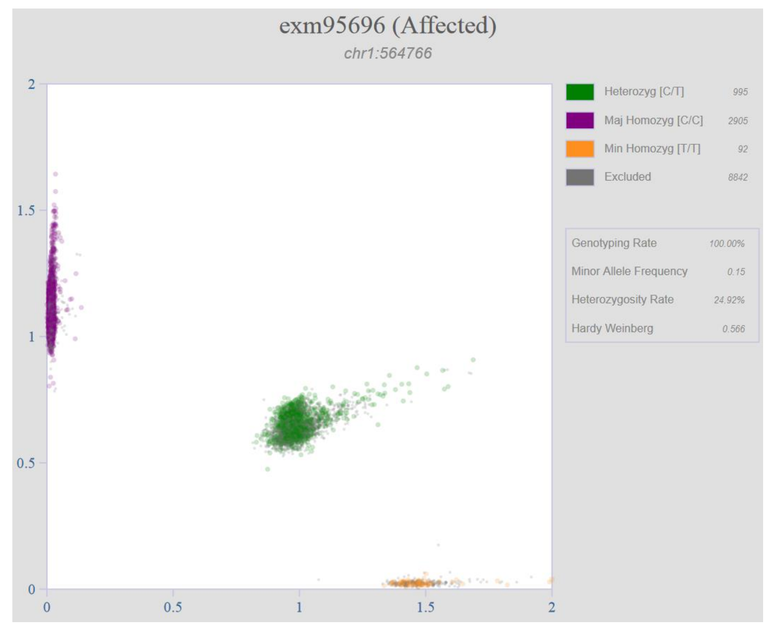

Figure 8: Cluster plot of exm95696 filtered for cases only, controls are marked as excluded.

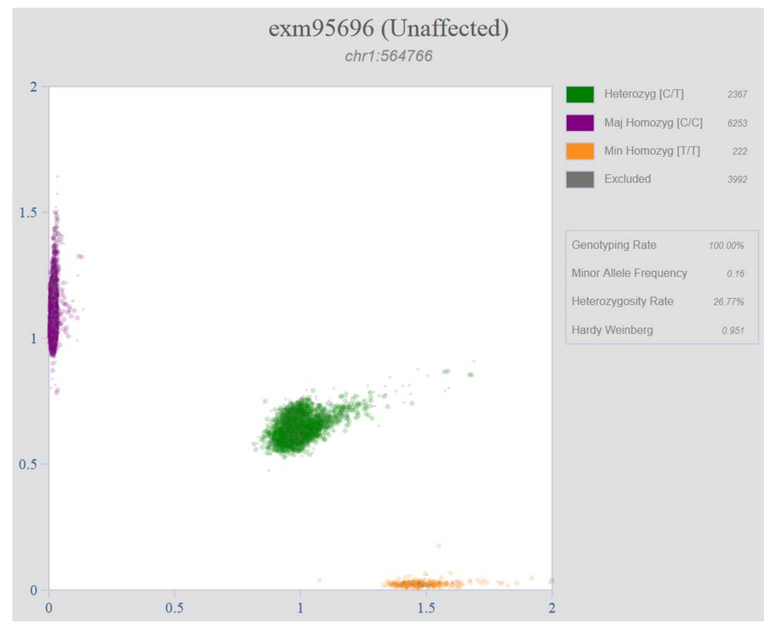

Figure 9: Cluster plot of exm95696 filtered for controls only, cases are marked as excluded.

Download

- ScatterShot is freely available to download http://www.well.ox.ac.uk/~wrayner/tools/