CNV analysis from genotyping data

Introduction and examples

Lucia Trastulla, PhD student

Max Planck Institute for Psychiatry,
Ziller lab

1 Introduction

The aim of our karyotyping analysis is to detect de novo copy number abnormalities arising in cultured cell lines.
In particular, our goal is to determine differences between cell lines and the starting material from which they were derived in term of copy number variation. Hence, the genomic screening for chromosomal abnormalities is used as quality control to establish and maintain stem cell lines.

2 Methods

The analysis is divided in two steps:

1. Sample hybridization using Illumina Global Screening Array-24 v1.0 (GSA) [1] and Genotyping (GT) module of GenomeStudio software to call the genotypes. For each probe, GT module estimates Log R ratio (LRR) and B-allele frequency (BAF) for each sample using a clustering module applied to the distribution of signal intensities.

2. Pairwise comparison between cell line (e.g. iPSC) and starting material (e.g. fibroblast) using GT module estimates of LRR and BAF to detect copy number variations (CNVs) through two-sample Hidden Markov Model (HMM) [2].

3 Workflow

1. **LRR and BAF estimates:**

   - Genotypes are called by comparing the generated data with those in the Illumina supplied clustering file.
   - Log R ratio and B-allele frequency for each sample are estimated through the GT module.
   - Samples quality is evaluated and quality control of genotype calls are curated. In particular, SNPs statistics from GenomeStudio are reviewed and filtered based on the protocols [3], [4].
   - Sample mismatches or mislabelling are checked using the genotype matches, i.e. lines from same donor should show genotype match around 99%.

2. **CNV detection:**

   - Copy number differences between starting material and derived lines from the same donor are inferred using a HMM algorithm implemented in BCFtools (cnv command) [2]. The program is run in the pairwise mode with the default parameters.
   - Copy number aberration (CNA) calls are filtered to reduce the number of false calls (see Methods section [5]) and only CNA higher or equal than 0.2 Mb in length are retained.

4 Example

1. First, SNPs are filtered following the QC protocol and sample call rates are computed. Fig.1 shows

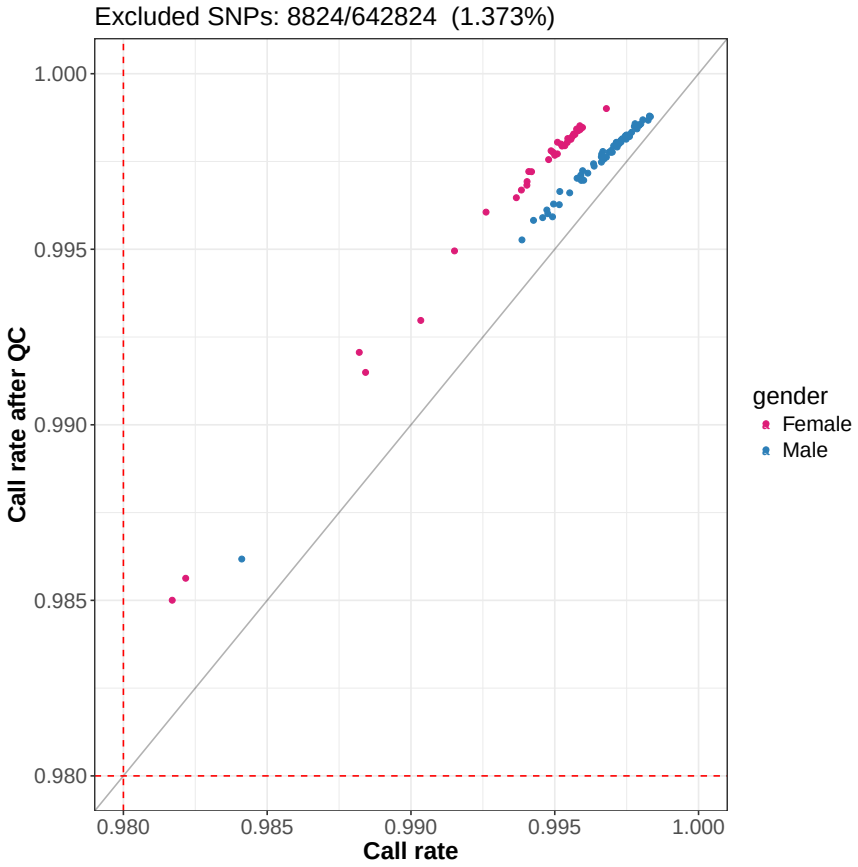

Fig. 1: Quality control preprocessing

call rate for each sample before and after SNPs filtering, together with the number of SNPs that did not pass the QC. Sample call rate should be higher than 0.98.

2. Pairwise genotype match (GT match) frequencies are computed for each line pair combination in the array (see Fig.2). Lines from the same donor show a GT match higher than 0.99, while unrelated lines present a GT match around 0.75.

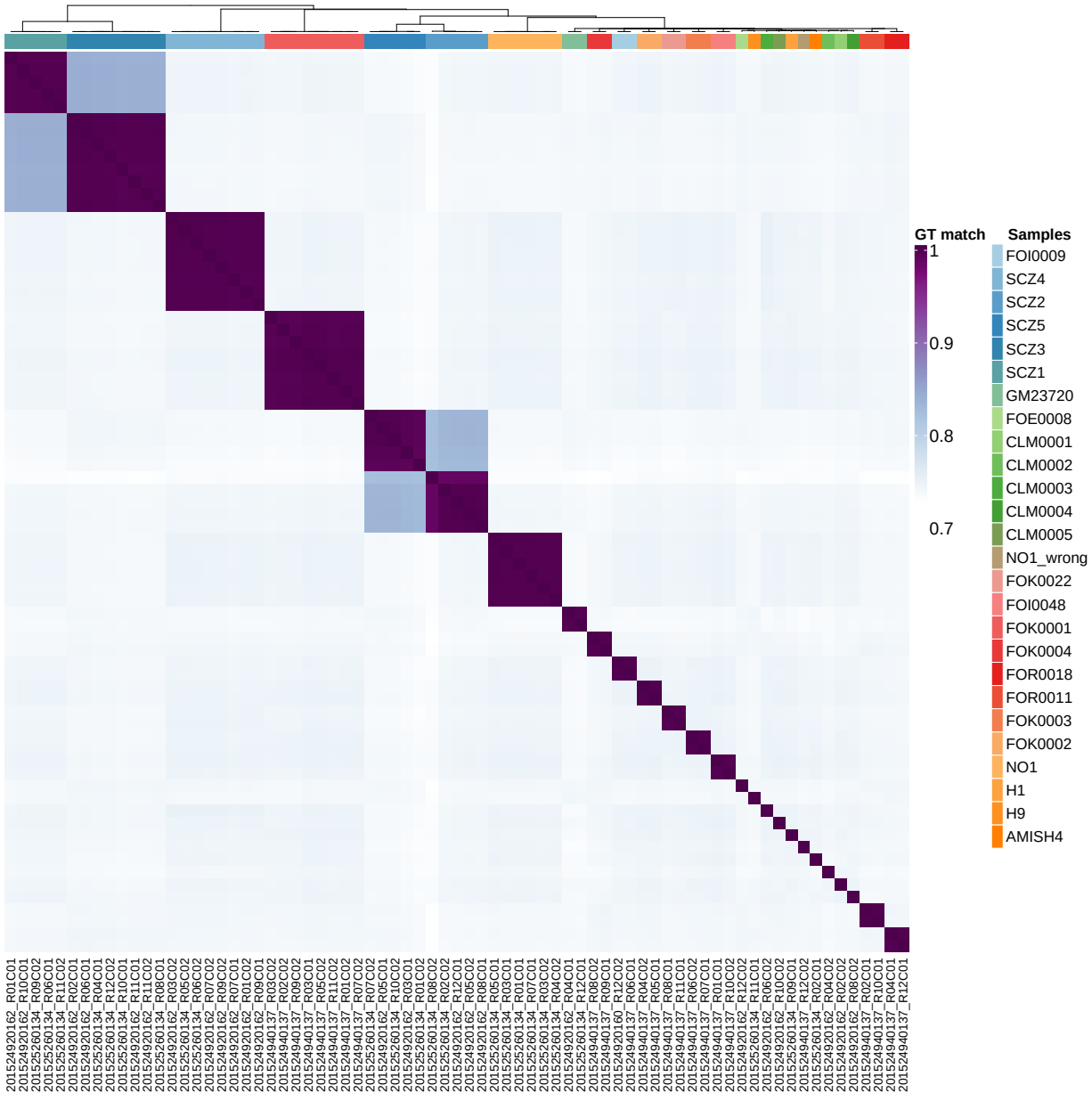

Fig. 2: Genotype match frequencies for each line pair combination

3. CNVs for starting material (fibroblast) and derived line (iPSC) from the same donor are detected and compared. An example is shown in Fig.3, the comparison is performed for each chromosome. Fig.3

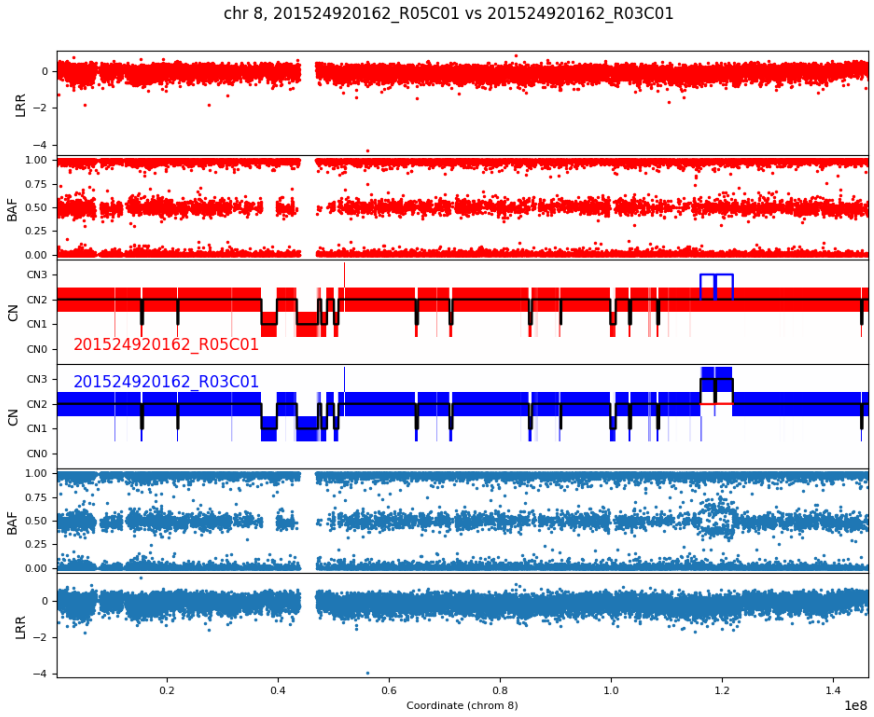

Fig. 3: CNV detection and comparison (pairwise comparison)

shows the results of the algorithm for the chromosome 8 together with the LRR and BAF used to infer the CNV status. The red part of the plot refers to the fibroblast, the blue part to the iPSC line.

4. More than one iPSC line can be derived from the same fibroblast, the summurized results are shown in Fig.4. The first plot with a bar-code name (SentrixBarcode and SentrixPosition form Illumina chip) refers to the fibroblast, the other bar-code named plots refer to the iPSC derived lines. The analysis is always performed in a pairwise mode (fibr-iPSC comparison). Hence, the plot shows the predicted CN status for each sample in the performed comparison with the colour code shown in the legend. The fibroblast plot contains also the colour red lines referring to a portion of the chromosome for which the fibroblast CN status is predicted in a different way in different comparisons. See following Tab.1.

Chr region		
	Fibr-iPSC1 comparison	Fibr-iPSC2 comparison
Fibr	CN = 1	CN = 2
iPSC1	CN = 1	
iPSC2		CN = 2

Tab. 1: Different CN status prediction of fibroblast for different comparisons.

The comparison line plot in Fig.4 shows whether there is any difference in the predicted CN status for at least one comparison in a specific region (e.g. fibr. CN=2 and iPSC1 CN=3 in fibr-iPSC1 comparison).

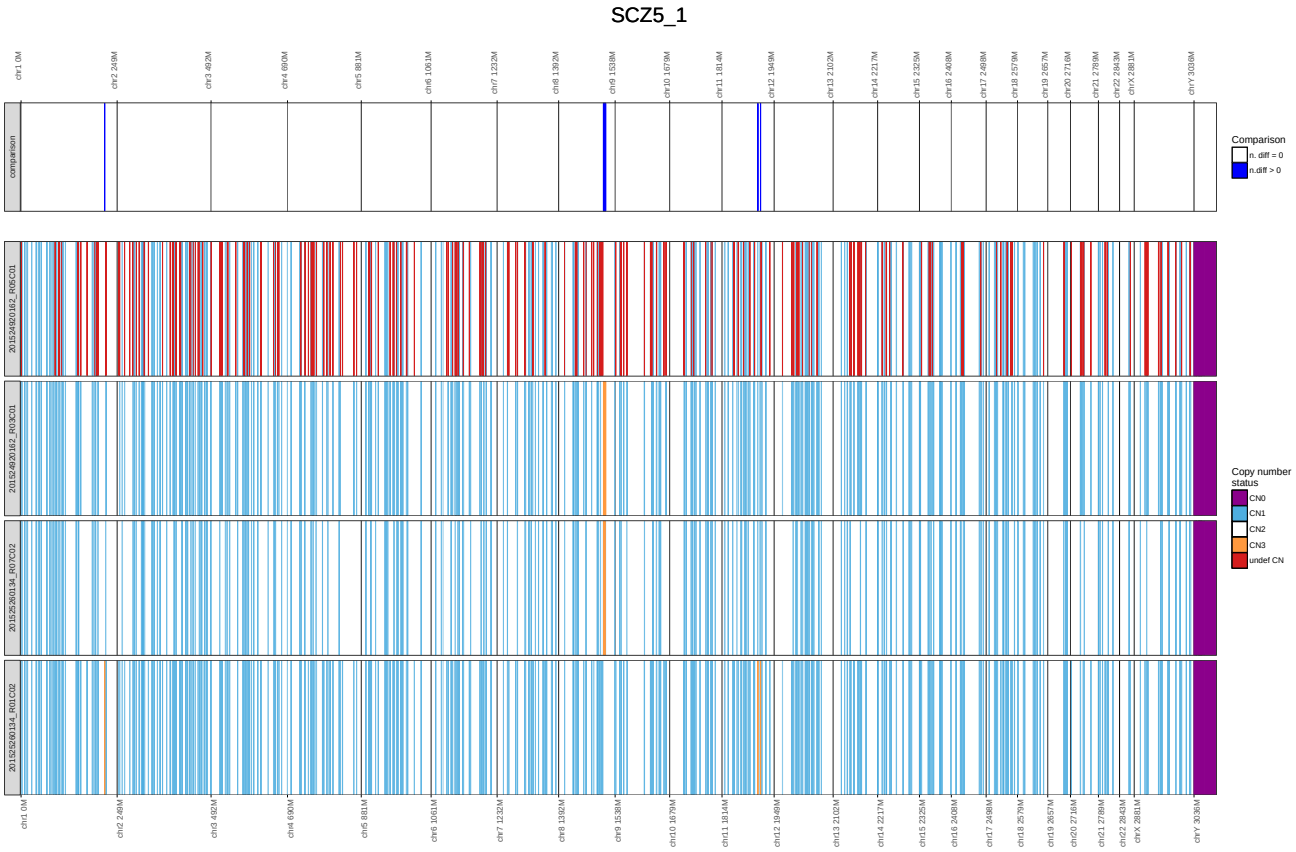

Fig. 4: CNV detection and comparison (donor lines overview)

5. Finally, to better visualize the number and the size of regions with a different CN status prediction between the fibroblast and the iPSC derived lines, the plot as shown in Fig.5 is generated.

   Each box refers to a derived iPSC lines and is divided per chromosome.

6. To summarize the result, a txt file is produced for each donor containing the prediction of CN status from each fibr-iPSC comparison for the entire genome.

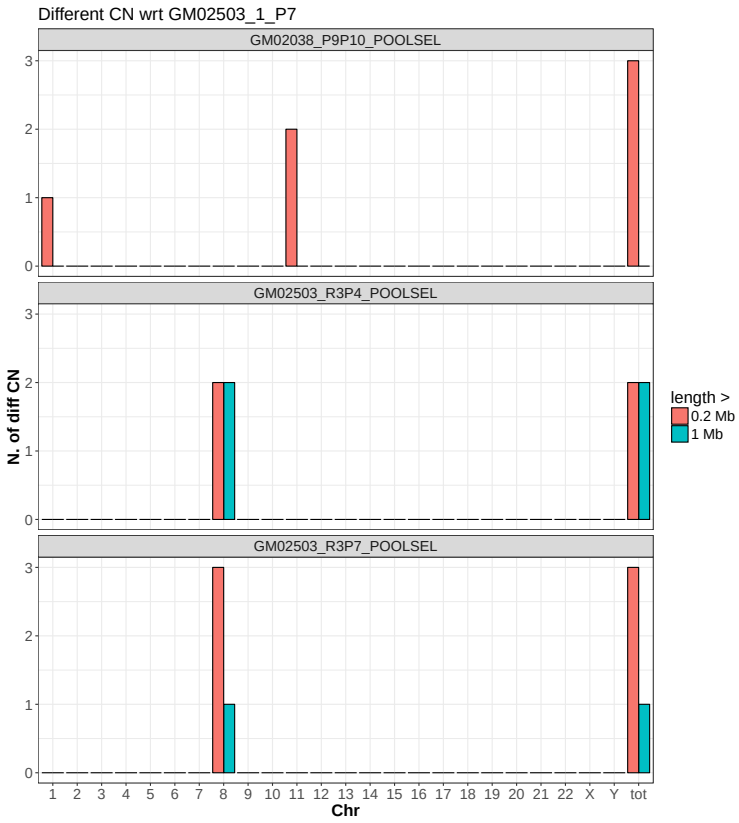

Fig. 5: Number of different predicted CN status in each comparison divided per chromosome

5 Note:

The algorithm can be also used in the single analysis option, in order to detect the CNV for a sample (e.g iPSC of fibroblast only). In this case, a single-sample HMM is used.

References

[1] Infinium Global Screening Array-24 v1.0, A powerful, high-quality, cost-effective array for population-scale genetic studies. [Technical Data Sheet]. http://www.illumina.com/content/dam/illumina-marketing/documents/products/datasheets/infinium-commercial-gsa-data-sheet-370-2016-016.pdf

[2] Danecek P., McCarthy S.A., HipSci C. and Durbin R., A Method for Checking Genomic Integrity in Cultured Cell Lines from SNP Genotyping Data. *PLoS One* 11 (2016).

[3] Illumina (2016). Infinium Genotyping Data Analysis. [Technical Data Sheet]. http://www.illumina.com/Documents/products/technotes/technote_infinium_genotyping_data_analysis.pdf

[4] Guo Y., He J., Zhao S., Wu H., Zhong X., Sheng Q., Samuels D.C., Shyr Y., Long J., Illumina human exome genotyping array clustering and quality control. *Nature protocols* (11) 2643-2662 (2014).

[5] Kilpinen H., Goncalves A., Leha A., Afzal V., Alasoo K., Ashford S., Bala S., Bensaddek D., Casale F. P., Culley O. J., Danecek P., Faulconbridge A., Harrison P. W., Kathuria A., McCarthy D., McCarthy S. A., Meleckyte R., Memari Y., Moens N., Soares F., Mann A., Streeter I., Agu C. A., Alderton

A., Nelson R., Harper S., Patel M., White A., Patel S. R., Clarke L., Halai R., Kirton C. M., Kolb-Kokocinski A., Beales P., Birney E., Danovi D., Lamond A. I., Ouwehand W. H., Vallier L., Watt F. M., Durbin R., Stegle O., Gaffney D. J. Common genetic variation drives molecular heterogeneity in human iPSCs, *Nature* 546 (2017)