**Recent Development of HipSci Custom CNVs.**

The custom CNV method that uses a direct comparison of genotype calls between the control and iPS samples, has been developed to allow inclusion of small gaps in the CNVs. This was because often the CNVs identified previously were observed to be separated by small numbers of SNPs, and so were considered one CNV event when reviewed by RD.

This was performed by using a sliding window of up to 25 SNPs, allowing up to 90% of the SNPs within the window to be gaps in the CNV (i.e. matches between the genotype calls of the control and the iPS sample). This % of SNPs being CNV gaps is referred to as % flexibility from here on.

Benchmarking of the method used the last five sets of genotyping data, although cohorts currently identified as containing a swap were excluded. To identify the optimal combination of window size and % flexibility, benchmarking used a window size of 2 to 25 SNPs combined with a flexibility of 10-90% (in increments of 10). Combinations were ignored when the minimum number of matching SNPs in the window was <2 (e.g. window size: 4, flexibility: >50%).

Results are shown in the attached graphs (samples predicted to have 0 CNVs are not included). The graphs do not show the results for window size >15 SNPs, as they did not provide useful results; the size of the CNVs stayed the same or decreased in all cases. Where a CNV size in a graph is displayed as two or more numbers (e.g. 11,8), each number gives the size of a separate CNV. The lowest optimal parameter combination is shown by a red square, i.e. all CNVs are detected and merged when using this combination. These lowest optimal combinations are also summarised in the following table.

|  |  |  |
| --- | --- | --- |
| **Cohort/sample** | **Window Size** | **% Flexibility** |
| FS03: samples 1&3 | 8 | 50 |
| FS03: sample 5 | 2 | 10 |
| FS12h: sample 1 | 3 | 40 |
| FS13: sample 2 | 3 | 40 |
| Cesj: sample 2 | 8 | 80 |
| 92255: all samples | 15 | 90 |

CNVs in the same locality progressively merge into coherent CNVs as the window size and flexibility are increased, although no parameter combination provides adequately merged CNVs for all samples. A window size of 8 combined with a flexibility of 80% would have achieved this, if cohort 92255 was excluded.

The custom CNV method has now been adapted, so that it applies all combinations of 2-15 SNPs and 10-90% flexibility, for all samples. The results are compared from the different parameter combinations, and the CNVs containing the largest overall number of SNPs are selected for each sample. Typically the larger CNVs will now look like the example shown below, where mismatch:1 indicates a difference between the control and iPS samples.

iPS Sample QC1Hip-93:

2 CNVs containing 10 SNPs (window size = 4 flexibility = 50%)

CNV #1: Chr:Y Coord:2951972 SNP:rs2534087 Mismatch:1

Coord:2970126 SNP:rs34320223 Mismatch:0

Coord:3009108 SNP:rs2534470 Mismatch:0

Coord:3126520 SNP:rs4033704 Mismatch:1

CNV #2: Chr:16 Coord:89386808 SNP:rs889574 Mismatch:1

Coord:89405827 SNP:rs2353030 Mismatch:0

Coord:89422823 SNP:rs3096324 Mismatch:0

Coord:89448663 SNP:rs3096299 Mismatch:1

Coord:89475817 SNP:rs17177108 Mismatch:1

Coord:89488587 SNP:rs2911253 Mismatch:1

Next steps are to include a *p*-value for the CNVs, and to predict gains.