SUPPLEMENTAL METHODS

A method for rapid, targeted CNV genotyping identifies rare variants associated with neurocognitive disease

Heather C. Mefford[1,2,6], Gregory M. Cooper[2,6], Troy Zerr[2,6], Joshua D. Smith[2], Carl Baker[2], Neil Shafer[2], Erik C. Thorland[3], Cindy Skinner[4], Charles E. Schwartz[4], Deborah A. Nickerson[2] and Evan E. Eichler[2,5,7] 2009. *Genome Res.* **19**(9):1579-85.

[1] Department of Pediatrics, University of Washington, Seattle, Washington 98195, USA;
[2] Department of Genome Sciences, University of Washington, Seattle, Washington 98195, USA;
[3] Department of Laboratory Medicine and Pathology, Mayo Clinic, Rochester, Minnesota 55905, USA;
[4] Greenwood Genetics Center, Greenwood, South Carolina 29646, USA;
[5] Howard Hughes Medical Institute, University of Washington, Seattle, Washington 98195, USA
[6] These authors contributed equally to this work.
[7] Corresponding author

Statistical Assumptions

SCOUT analyzes fluorescence data and SNP genotype calls reported by BeadStudio (Illumina) and produces, for each sample and each site, a score indicating the degree of deviation of the observed data from the population mean. Informally, SCOUT is based on the following statistical assumptions:

(1) The copy number of each sample is the same for every probe in a single interrogated interval; any apparent inconsistency in copy number between probes is due to measurement error.

(2) Null samples, if present, form a cluster near the origin.

(3) For each probe, log-transformed A-allele fluorescence measurements (*x*-coordinate values) for A-allele homozygotes, and B-allele fluorescence measurements (*y*-coordinate values) for B-allele homozygotes, are normally distributed with equal variance (but not necessarily equal mean). Samples with measurements unusually close to the origin are more likely to harbor deletions than those with measurements near the cluster center; conversely, samples with measurements unusually far from the origin are more likely to harbor duplications.

(4) For any probe, fluorescence measurements for SNP heterozygotes ('AB' samples) are bivariate normally distributed. Samples with measurements unusually far from the origin and unusually distant from to the line connecting the origin to the cluster center (i.e. samples with allelic states 'AAB' and 'ABB') are

more likely to harbor duplications than those with measurements near the cluster center.

Our previous work, SCIMM (Cooper 2008), is based on similar assumptions, but differs in how these assumptions are used: SCIMM assumes that there exist three classes of samples (null, haploid and diploid) and uses mixture-likelihood based clustering (Dempster 1977) to find the parameters maximizing the likelihood of the observed data, whereas SCOUT assumes that all samples have the same copy number (diploid) and attempts to identify 'outlier' samples likely to violate this assumption (Hawkins 1980).

In the discussion below, observed fluorescence data for sample i ($i = 1 .. n$) at probe j ($j = 1 .. m$) are represented by (x_{ij} , y_{ij}), and observed SNP genotype calls are represented by indicator variables

$s_{ij1} = 1$ if sample i has SNP genotype call 'AA' at probe j

$s_{ij2} = 1$ if sample i has SNP genotype call 'BB' at probe j

$s_{ij3} = 1$ if sample i has SNP genotype call 'AB' at probe j.

Scoring:

Samples near the origin are filtered by an initial round of mixture-likelihood clustering, and remaining samples with 'no call' SNP genotypes are assigned a SNP genotype of either 'AA', 'AB', or 'BB', in the same manner as SCIMM (with the exception that SCOUT is less likely than SCIMM to treat a 'no call' sample as a sample with a homozygous SNP genotype).

Per-probe scores for SNP homozygotes are determined as follows: Observed data are log-transformed

$$x'_{ij} = \log(x_{ij} + \varepsilon) \qquad (\varepsilon = 10^{-10}).$$
$$y'_{ij} = \log(y_{ij} + \varepsilon)$$

and mean and variance parameters are estimated separately for each probe

$$\mu_{j1} = \sum_i x'_{ij} s_{ij1} \Big/ \sum_i s_{ij1}$$
$$\mu_{j2} = \sum_i y'_{ij} s_{ij2} \Big/ \sum_i s_{ij2}$$
$$\sigma_j^2 = \left(\sum_i s_{ij1} (x'_{ij} - \mu_{j1})^2 + \sum_i s_{ij2} (y'_{ij} - \mu_{j2})^2 \right) \Big/ \sum_i (s_{ij1} + s_{ij2}) \ .$$

Scores are then calculated as

$$z_{ij} = \begin{cases} (x'_{ij} - \mu_{j1}) \Big/ \sigma_j^2 & \text{if } s_{ij1} = 1 \\ (y'_{ij} - \mu_{j2}) \Big/ \sigma_j^2 & \text{if } s_{ij2} = 1 \end{cases}.$$

To calculate per-probe scores for SNP heterozygotes, data are translated and rotated

$$\bar{x}_j = \sum_i x_{ij} s_{ij3} \Big/ \sum_i s_{ij3}$$
$$\bar{y}_j = \sum_i y_{ij} s_{ij3} \Big/ \sum_i s_{ij3}$$

$$v_j = \bar{x}_j \Big/ \sqrt{\bar{x}_j^2 + \bar{y}_j^2}$$
$$w_j = \bar{y}_j \Big/ \sqrt{\bar{x}_j^2 + \bar{y}_j^2}$$

$$a_{ij} = v_j (x_{ij} - \bar{x}) - w_j (y_{ij} - \bar{y})$$
$$b_{ij} = w_j (x_{ij} - \bar{x}) + v_j (y_{ij} - \bar{y})$$

and scaled

$$a'_{ij} = a_{ij} \big/ \text{mad}_j(a_{ij})$$
$$b'_{ij} = b_{ij} \big/ \text{mad}_j(b_{ij})$$

('mad' represents median absolute deviation) so that the transformed data (a'_{ij}, b'_{ij}) have mean zero and variance approximately one. a'_{ij} represents the difference between the observed data and the mean attributable to variability in overall intensity, and b'_{ij} represents the difference attributable to variability in allelic ratio.

At this point we assume that (a'_{ij}, b'_{ij}) are observations of two independent, normally distributed random variables (A_j, B_j) and calculate

$$\begin{aligned} z_{ij} &= \text{qnorm}(\, P(a_{ij} > A_j, |\, b_{ij}\, | > |\, B_j\, |) \,) \\ &= \text{qnorm}(\, 2P(a_{ij} > A_j) P(|\, b_{ij}\, | > B_j) \,) \\ &= \text{qnorm}(\, 2(1 - \text{pnorm}(a_{ij}))(1 - \text{pnorm}(|\, b_{ij}\, |)) \,) \end{aligned}$$

where 'qnorm' and 'pnorm' denote the quantile and distribution functions of $\boldsymbol{N}(0,1)$ respectively.

Per-site scores are determined by summation of per-probe scores:

$$z_i = \sum_j z_{ij} \Big/ \sqrt{m} .$$

Sample quality control:

The SCOUT scoring scheme alone is unable to reliably distinguish between samples harboring duplications and deletions (which we expect to generate high scores only at specific sites) and samples of low quality (which we expect to generate anomalous fluorescence intensity and allelic ratio measurements throughout the genome). Moreover, the presence of low-quality samples leads to increased estimates of probe noisiness (e.g. σ_j^2 above) and correspondingly lower scores for high-quality samples than would otherwise be obtained.

To provide robustness against the presence of low-quality samples, SCOUT performs an initial quality control pass, independently generating per-probe scores z_{ij} and discarding samples with an assay-wide excess of extreme scores. A second pass the calculates per-site scores using the remaining data as described above. For the present study, we discarded all samples where $| z_{ij} | > 2.5$ for at least 10% of all probes.

Implementation:

The implementation of SCOUT consists a front-end PERL script used to parse the input BeadStudio report and a back-end R script used to perform quality control, generate scatterplots, and calculate per-site scores as described above.

References:

(Illumina) Illumina Incorporated, San Diego, California, http://www.illumina.com

(Cooper 2008) Cooper GM, Zerr T, Kidd JM, Eichler EE, Nickerson DA 2008. Systematic assessment of copy number variant detection via genome-wide SNP genotyping. *Nat Genet.* **40**(10):1199-203.

(Dempster 1977) Dempster AP, Laird NM, Rubin DB 1977. Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B. Methodological* **39**: 1–38.

(Hawkins 1980) Hawkins D. Identification of Outliers. Chapman and Hall. London, 1980.