SUPPLEMENTAL METHODS

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 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

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

and mean and variance parameters are estimated separately for each probe

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

Scores are then calculated as

$$z_{ij} = \left[\begin{array}{ll} (x'_{ij} - \mu_{j1}) \big/ \sigma_j^2 & \text{if } s_{ij1} = 1 \\ (y'_{ij} - \mu_{j2}) \big/ \sigma_j^2 & \text{if } s_{ij2} = 1 \end{array} \right. .$$

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

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

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

$$\begin{aligned} a_{ij} &= v_j (x_{ij} - \bar{x}_j) - w_j (y_{ij} - \bar{y}_j) \\ b_{ij} &= w_j (x_{ij} - \bar{x}_j) + v_j (y_{ij} - \bar{y}_j) \end{aligned}$$

and scaled

$$\begin{aligned} a'_{ij} &= a_{ij} \big/ \mathrm{mad}_j(a_{ij}) \\ b'_{ij} &= b_{ij} \big/ \mathrm{mad}_j(b_{ij}) \end{aligned}$$

('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

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

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:

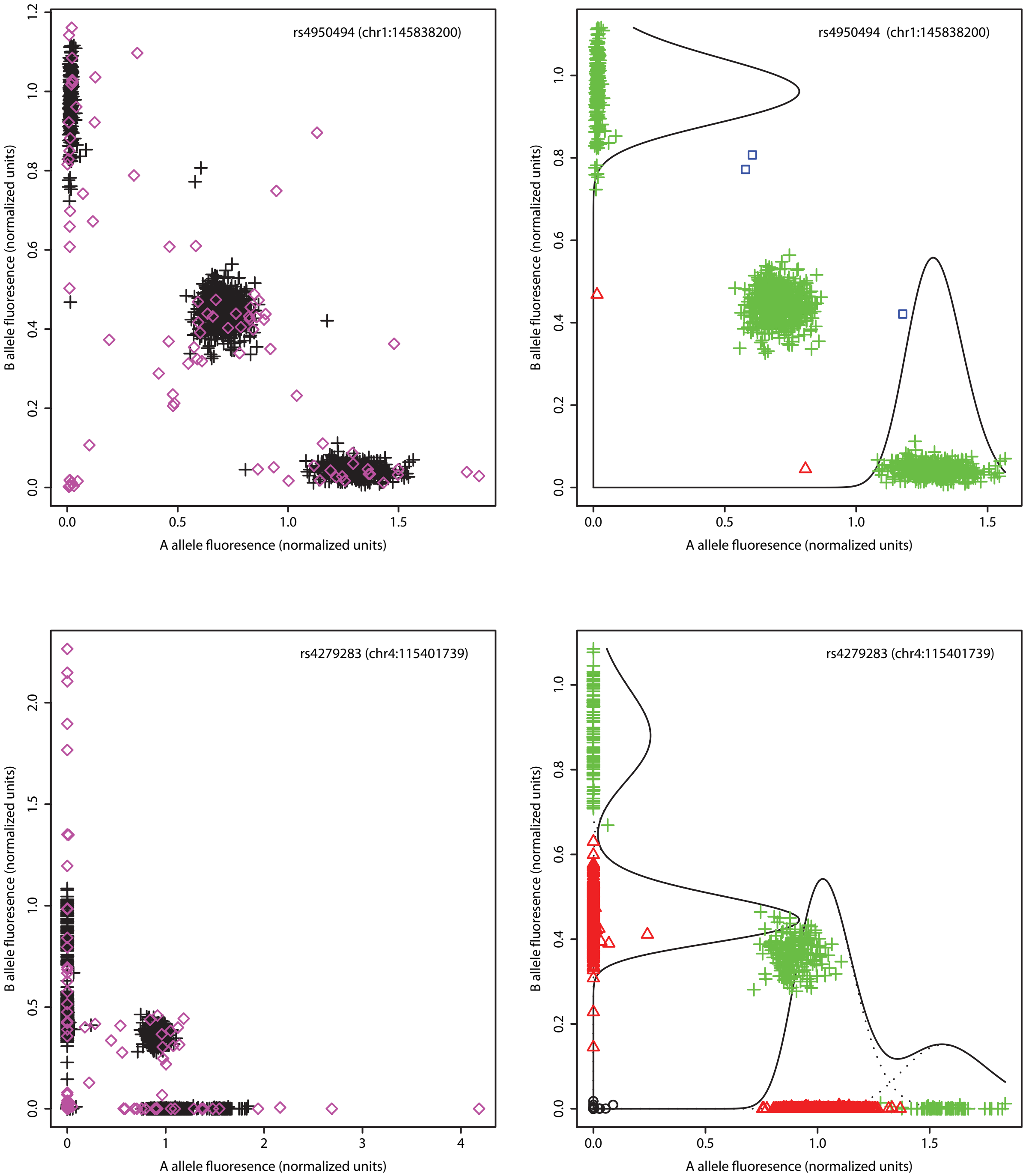
(Illumina) Illumina Incorporated, San Diego, California, http://www.illumina.com

(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.

Supplemental Figure 1. Sample quality control for rare and common copy number variant probes. Upper Left: Rare variant probe (rs4950494, chr1:145838200), including low-data-quality samples (represented by purple diamonds). Upper Right: SCOUT output, with low-data-quality samples excluded. Red triangles, green crosses, and blue squares represent putative copy-number-1, copy-number-2, and copy-number-3 samples, respectively. Curves indicate estimated distribution for A-allele and B-allele homozygote fluorescence intensity. Lower Left: Common variant probe (rs4950494, chr1:145838200), including low-data-quality samples. Lower Right: SCIMM output, with low-data-quality samples excluded. Black circles, red triangles, green crosses represent putative copy-number-0, copy-number-1, and copy-number-2 samples, respectively. Curves indicate estimated two component mixture distribution for A-allele and B-allele homozygote fluorescence intensity. For this image, we display all samples analyzed. However, for analysis, each plate was analyzed independently (see Methods).


rs4950494 (chr1:145838200)
B allele fluoresence (normalized units)
A allele fluoresence (normalized units)
0.0
0.2
0.4
0.6
0.8
1.0
1.2
0.0
0.5
1.0
1.5
rs4950494 (chr1:145838200)
B allele fluoresence (normalized units)
A allele fluoresence (normalized units)
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.5
1.0
1.5
rs4279283 (chr4:115401739)
B allele fluoresence (normalized units)
A allele fluoresence (normalized units)
0.0
0.5
1.0
1.5
2.0
0
1
2
3
4
rs4279283 (chr4:115401739)
B allele fluoresence (normalized units)
A allele fluoresence (normalized units)
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.5
1.0
1.5

Supplemental Table 2. Control samples with known CNVs in targeted hotspot regions

Sample	Region Name	Chr	Start	Stop	Type	Score	# probes
1	17q12	chr17	31818094	33424757	dup	7.73	9
2	17q12	chr17	31818094	33424757	dup	9.57	9
3	17q12	chr17	31818094	33424757	dup	11.25	9
4	1q21.1	chr1	144743482	147025354	del	-12.96	8
5	10q23	chr10	81129804	89119386	partial_dup	4.50	4
6	15q13 BP4-BP5	chr15	28722450	30505878	del	-12.05	9
7	15q13 BP4-BP5	chr15	28722450	30505878	del	-14.21	9
8	15q13 BP4-BP5	chr15	28722450	30505878	del	-15.37	9
9	1q21.1 TAR	chr1	144004163	144340000	del	-6.87	2
9	1q21.1 TAR poly	chr1	144340001	144450000	del	-6.12	2
9	1q21.1	chr1	144743482	147025354	del	-13.91	8
10	16p11.2	chr16	29350923	30209759	del	-9.25	5
11	15q11 BP1-BP2	chr15	20306549	20691555	dup	5.38	3
11	15q11 BP2-BP3	chr15	21100000	26374583	dup	5.19	4
12	17q23	chr17	55005024	55434745	partial_dup	5.92	4
13	16p13.11 BP1-BP2	chr16	15118883	16343188	del	-8.29	7
14	17q12	chr17	31818094	33424757	del	-10.80	9
15	15q25.2	chr15	80506941	83597519	del	-12.09	7
16	1q21.1	chr1	144743482	147025354	dup	12.41	8
17	15q13 BP3-BP4	chr15	26741971	28229170	dup	9.53	5
17	15q13 BP4-BP5	chr15	28722450	30505878	dup	7.53	9
18	17q21.3	chr17	40949281	41500000	del	-6.60	4
19	16p13.11 BP1-BP2	chr16	15118883	16343188	dup	8.15	7
20	15q24 BP0-BP1	chr15	70698861	72197406	del	-10.43	5
20	15q24 BP1-BP2	chr15	72197407	73384192	del	-8.70	3
20	15q24	chr15	73384193	73580000	del	-9.99	5
21	17q21.3	chr17	40949281	41500000	del	-8.56	4
22	16p13.11 BP1-BP2	chr16	15118883	16343188	del	-8.99	7
22	7q11.23 (WBS)	chr7	72004122	74938688	del	-4.59	2
23	22q11 VCFS CR	chr22	17037082	18989585	del	-8.60	4
23	22q11 VCFS distal	chr22	18989586	19899201	del	-5.79	4
24	22q11 VCFS CR	chr22	17037082	18989585	dup	5.26	4
24	22q11 VCFS distal	chr22	18989586	19899201	dup	6.78	4
25	1q21.1 TAR poly	chr1	144340001	144450000	dup	3.67	2
25	7q11.23 (WBS)	chr7	72004122	74938688	del	-4.80	2
26	17p11 SMS CR	chr17	15435335	18161156	partial_del	-3.26	3
26	17p11 SMS distal	chr17	18616157	20500000	del	-8.96	5
27	15q11 BP2-BP3	chr15	21100000	26374583	del	-9.44	4
28	15q11 BP1-BP2	chr15	20306549	20691555	dup	5.76	3
28	15q11 BP2-BP3	chr15	21100000	26374583	dup	6.46	4
29	16p11.2	chr16	29350923	30209759	del	-9.01	5
30	15q13 BP4-BP5	chr15	28722450	30505878	dup	5.85	9
31	15q13 BP4-BP5	chr15	28722450	30505878	del	-11.45	9
32	16p13.11 BP1-BP2	chr16	15118883	16343188	dup	9.62	7
33	17p11 SMS CR	chr17	15435335	18161156	partial_del	-5.64	3
33	17p11 SMS distal	chr17	18616157	20500000	del	-8.88	5
34	15q13 BP3-BP4	chr15	26741971	28229170	del	-7.52	5
34	15q13 BP4-BP5	chr15	28722450	30505878	del	-10.71	9
35	1q21.1 TAR	chr1	144004163	144340000	del	-7.29	2
35	1q21.1 TAR poly	chr1	144340001	144450000	del	-6.96	2
36	15q13 BP3-BP4	chr15	26741971	28229170	del	-8.12	5
37	15q11 BP1-BP2	chr15	20306549	20691555	dup	8.04	3
37	15q11 BP2-BP3	chr15	21100000	26374583	dup	7.73	4
37	15q13 BP3-BP4	chr15	26741971	28229170	dup	6.55	5
37	15q13 BP4-BP5	chr15	28722450	30505878	dup	11.64	9
38	17q12	chr17	31818094	33424757	del	-10.50	9
39	16p13.11 BP2-BP3	chr16	16343189	18355459	del	-8.85	4

TAR, thrombocytopenia-absent radius; poly, polymorphism; WBS, Williams-Beurens syndrome; SMS, Smith-Magenis syndrome; CR, critical region; VCFS, velocardiofacial syndrome; distal, refers to distal part of a larger deletion or duplication, usually part of the common event but does not include the critical region.