SUPPLEMENT
Targeted Interrogation of Copy Number Variation using SCIMMkit
Troy Zerr, Gregory M. Cooper, Evan E. Eichler, and Deborah A. Nickerson

CONTENTS

SUPPLEMENTARY METHODS

Introduction

We performed a cross-platform, genome-wide analysis of common diallelic deletion sites, comparing genotypes produced by SCIMMkit using data from the Illumina 1M-DuoV3 SNP array with previously published genotypes produced by Birdsuite using the Affymetrix 6.0 SNP array for 269 HapMap samples (McCarroll and Kuruvilla *et al.* 2008, Korn and Kuruvilla *et al.* 2008). The results of this analysis include a list of sites and probe sets producing highly consistent genotypes in both assays, suitable for retrospective analysis of genome-wide SNP data.

Data Sources

Probe names, probe coordinates, and normalized A-allele and B-allele intensity ('X' and 'Y') and SNP genotype data for 269 HapMap samples from the Illumina 1M-DuoV3 SNP array were obtained from Illumina Corporation:

ftp://ftp.illumina.com/Whole Genome Genotyping Files/Human1M-Duo_v3_product_files/Human1M-duoV3_SNP_List.txt

ftp://ftp.illumina.com/Whole Genome Genotyping Files/Human1M-Duo_v3_product_files/1M-DuoV3_FullCall_Reports/1M-DuoV3_CEU_Final_Call_Report.csv.gz

ftp://ftp.illumina.com/Whole Genome Genotyping Files/Human1M-Duo_v3_product_files/1M-DuoV3_FullCall_Reports/1M-DuoV3_JPT+CHB_Final_Call_Report.csv.gz

ftp://ftp.illumina.com/Whole Genome Genotyping Files/Human1M-Duo_v3_product_files/1M-DuoV3_FullCall_Reports/1M-DuoV3_YRI_Final_Call_Report.csv.gz

CNV coordinates and copy number genotypes for these samples, generated by BirdSuite using data produced with the Affymetrix 6.0 genome-wide SNP array, were obtained from supplementary tables S2 and S3 of McCarroll and Kuruvilla *et al.* 2008.

http://www.nature.com/ng/journal/v40/n10/suppinfo/ng.238_S1.html

Sample NA12236 is present in the Affymetrix data set but is not present in the Illumina data set, and is not included in the following analysis.

SCIMMKit build 10Sep2009, with default parameter settings, was used for all analysis of Illumina 1MDuo-V3 data. No information regarding parental relationships, gender, or genotypes of reference samples was used by SCIMM or SCIMM-Search.

Common Autosomal Deletions

We used SCIMM-Search to search for copy-number informative probe sets for 272 common autosomal deletion sites listed in McCarroll Table S2 (all sites for which no copy number genotype other than 0, 1, or 2 is reported in McCarroll Table S3, and which have a sample deletion allele frequency exceeding 5%). The coordinates provided by McCarroll Table S2 are internal to each deleted interval; to accommodate the possibility that 1M-DuoV3 probes specific to the deleted sequence map outside the reported coordinates, we expanded the search to all probes mapping within 5,000 bp of each deleted interval.

The Illumina 1M-DuoV3 SNP array contains at least two probes (the minimum number of probes required by SCIMM-Search) within the search region for 94% (256/272) of these sites. 42% (113/272) of these sites are SCIMM-Search positive. The number of 1M-DuoV3 probes within each search region is significantly higher for SCIMM-Search positive sites (median for positive sites = 9, median for negative sites = 4, Mann-Whitney $U = 13927.5$, $n_1 = 113$, $n_2 = 159$, $P = 8.66 \cdot 10^{-15}$) (Figure S3). Sample allele frequency (computed from Birdsuite genotypes) for SCIMM-Search positive sites is shifted modestly downward relative to SCIMM-Search negative sites (median for positive sites = 0.15, median for negative sites = 0.22, Mann-Whitney $U = 6530$, $n_1 = 113$, $n_2 = 159$, $P = 1.25 \cdot 10^{-4}$). 85% (96/113) of all SCIMM-Search positive sites have per-site concordance with McCarroll genotypes exceeding 99%.

Uncommon Autosomal Deletions

We also applied SCIMM-Search to 694 lower frequency (<5% sample deletion allele frequency) sites listed in McCarroll Table S2. Each site was expanded by 5,000 bp in each direction, as with the common deletions. 691 of these regions span at least two probes, and 56% (392/694) of these regions are SCIMM-Search positive. As with the common deletions, probe coverage is significantly higher for SCIMM-Search positive sites (median probe coverage for positive sites = 10, median for negative sites = 5, Mann-Whitney $U = 96036.5$, $n_1 = 392$, $n_2 = 302$, $P < 2.2 \cdot 10^{-16}$) (Figure S3). Sample allele frequency (computed from Birdsuite genotypes) for SCIMM-Search positive sites is shifted modestly upward relative to SCIMM-Search negative sites (median allele frequency for positive sites = 0.0094, median for negative sites = 0.0077, Mann-Whitney $U = 67255$, $n_1 = 392$, $n_2 = 302$, $P = 0.0021$).

Since per-site concordance rate is less informative as a measure of correlation for rare variant genotypes (*e.g.* high concordance rates can be trivially achieved by labeling all

samples as homozygous for the major allele), we adopted more stringent consistency criteria for lower-frequency deletions. We denote the predictive positive value (PPV) of the SCIMM genotypes, assuming the Birdsuite genotypes are correct, as

$$\mathrm{PPV}_S = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

$$= \frac{\text{Samples with SCIMM genotype 0 or 1 and Birdsuite genotype 0 or 1}}{\text{Samples with SCIMM genotype 0 or 1}}$$

and conversely denote the positive predictive value of the Birdsuite genotypes, assuming the SCIMM genotypes are correct, as

$$\mathrm{PPV}_B = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

$$= \frac{\text{Samples with SCIMM genotype 0 or 1 and Birdsuite genotype 0 or 1}}{\text{Samples with Birdsuite genotype 0 or 1}}$$

We find that 89% (347/392) sites satisfy both $\mathrm{PPV}_S \geq 0.8$ and $\mathrm{PPV}_B \geq 0.8$; all sites satisfying these criteria have per-site genotype concordance exceeding 99%. 28 sites have concordance exceeding 99% with $\mathrm{PPV}_S < 0.8$ or $\mathrm{PPV}_B < 0.8$, while the remaining 17 sites have concordance below 99% with $\mathrm{PPV}_S < 0.8$ or $\mathrm{PPV}_B < 0.8$.

X-Linked Deletions

McCarroll Table S3 also contains 14 X-linked diallelic deletions. 6 of these span two or more 1MDuoV3 probes; all 6 are SCIMM-Search positive and have genotype concordance exceeding 98.5%. (BirdSuite and SCIMM report genotypes for X-linked sites in a similar manner; male samples are expected to have genotype ‘0’ if carrying a deletion allele, and are expected to have genotype ‘1’ otherwise.)

Tabulation of Results

The complete list of targets is presented as supplementary Table S1; the list of all genotyped targets, including per-site concordance statistics, is presented as supplementary Table S2.

Tables of SCIMM genotyping targets (containing hg18 genomic coordinates, McCarroll CNP ids, and the Illumina 1MDuoV3 probe sets generated by SCIMM-Search), in SCIMMkit target file format, are provided on the SCIMMkit web site:

http://droog.gs.washington.edu/scimmkit/data/targets_1MDuoV3_validated_common_autosomal_25Sep2009.csv

SCIMM genotyping targets for 96 highly concordant common autosomal deletions (sites with deletion allele frequency exceeding 99% and concordance with Birdsuite genotypes exceeding 99%).

http://droog.gs.washington.edu/scimmkit/data/targets_1MDuoV3_validated_all_25Sep2009.csv

SCIMM genotyping targets for 96 highly concordant autosomal deletions, 347 highly concordant ($\mathrm{PPV}_S \geq 0.8$ and $\mathrm{PPV}_B \geq 0.8$) uncommon autosomal deletions, and 6 X-linked sites.

SCIMM genotypes for these targets are available at:
http://droog.gs.washington.edu/scimmkit/data/genotypes_1MDuoV3_validated_all_25Sep2009.csv .

REFERENCES

Korn J.M. and Kuruvilla F.G. *et al.* (2008) Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. *Nat Genet.* **40**(10):1253-60.

McCarroll S.A. and Kuruvilla F.G. *et al.* (2008) Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nat Genet.* **40**(10):1166-74.

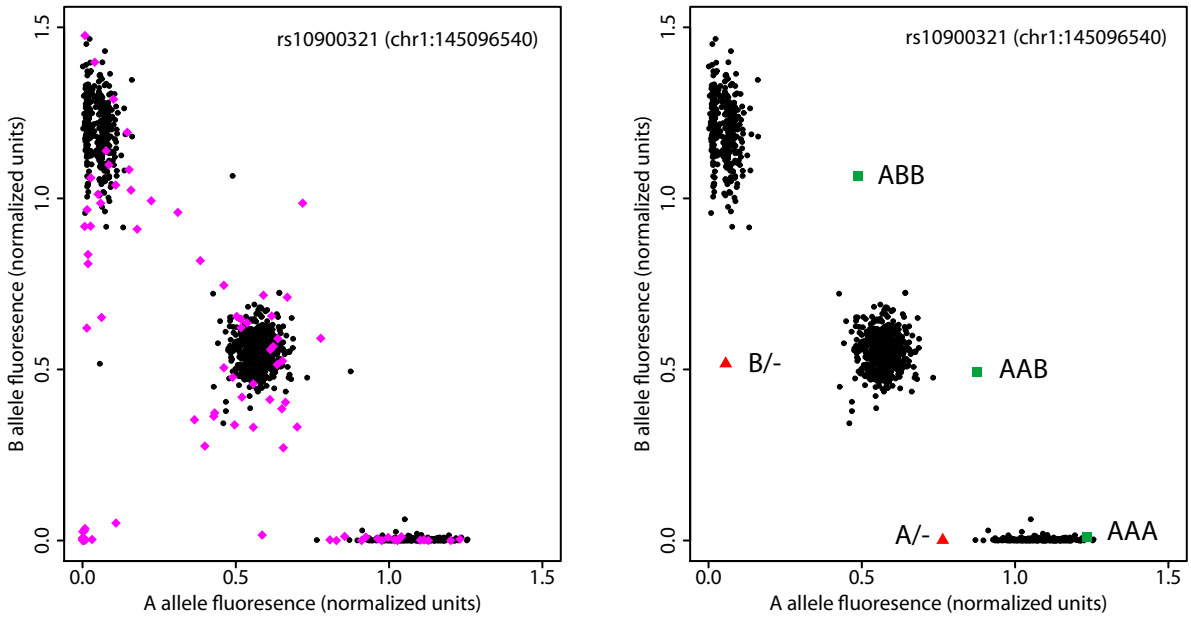

FIGURE S1. Fluorescence data for 1233 samples at a single SNP probe (rs10900321) within a rare reciprocal deletion/duplication site (chr1:144743482-147025354). *Left:* SCOUT initial quality control pass. Diamonds indicate 87 samples failing quality control. *Right:* Rare CNVs detected by SCOUT. Red triangles indicate samples with deletions (A/-, B/-); green squares indicate samples with duplications (AAA, AAB, ABB). Gains and losses were called by SCOUT using 8 SNP probes and confirmed by array-CGH (Mefford *et al.* 2009).

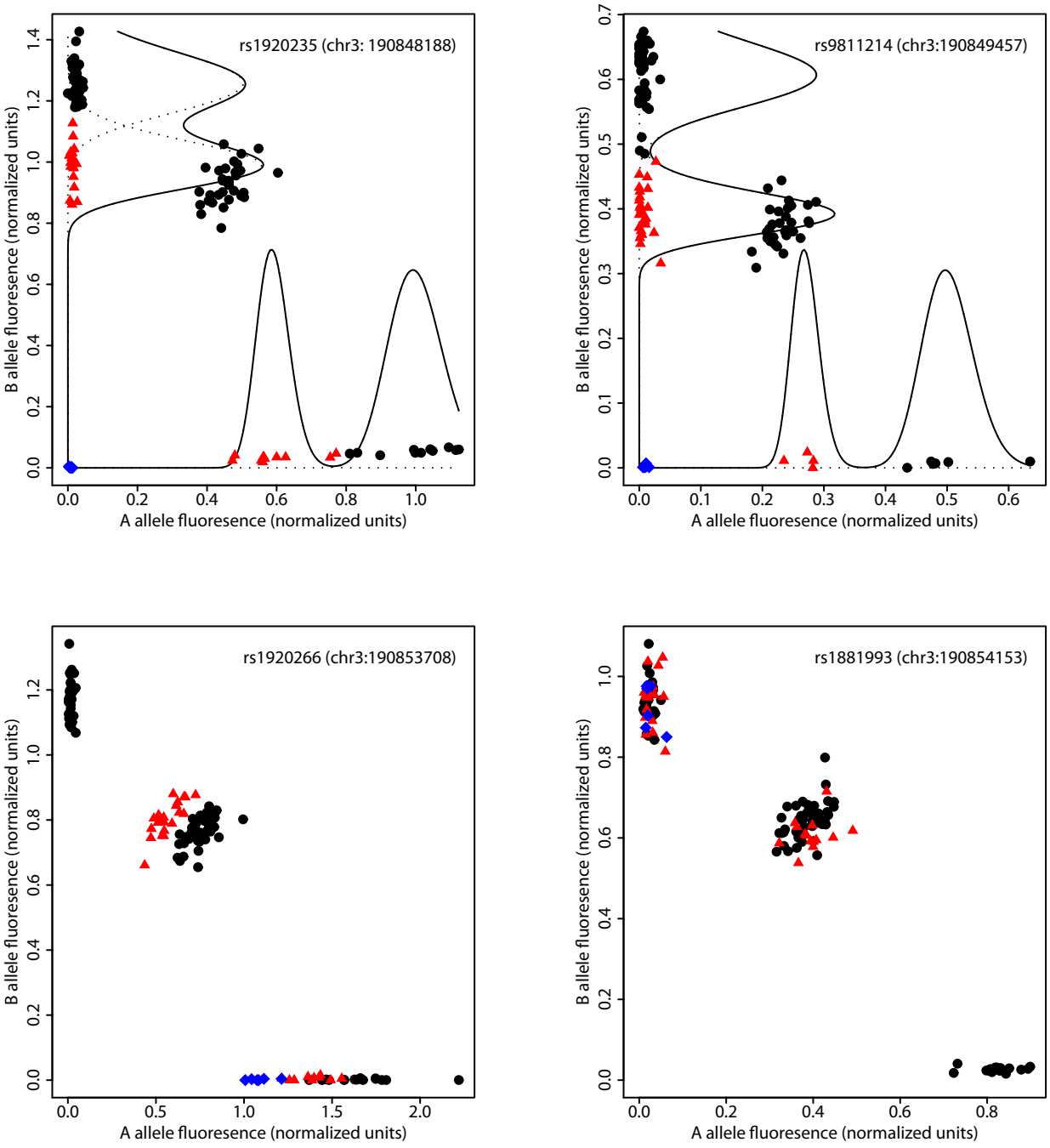

FIGURE S2. Fluorescence data for four probes mapping within a common deletion polymorphism (chr3:190841855-190855757). Copy number genotypes for 125 HapMap samples (Cooper and Zerr *et al.* 2008) were computed using a set of 6 Illumina Human1M probes selected automatically by SCIMM-Search. Copy number genotypes were computed by SCIMM (blue diamonds, copy number 0; red triangles, copy number 1; black circles, copy number 2). *Upper left, upper right:* insertion-allele-specific probes included in the probe set; superimposed curves describe components of the mixture distribution estimated by SCIMM. *Lower left, lower right:* non-insertion-allele-specific probes excluded from the probe set.

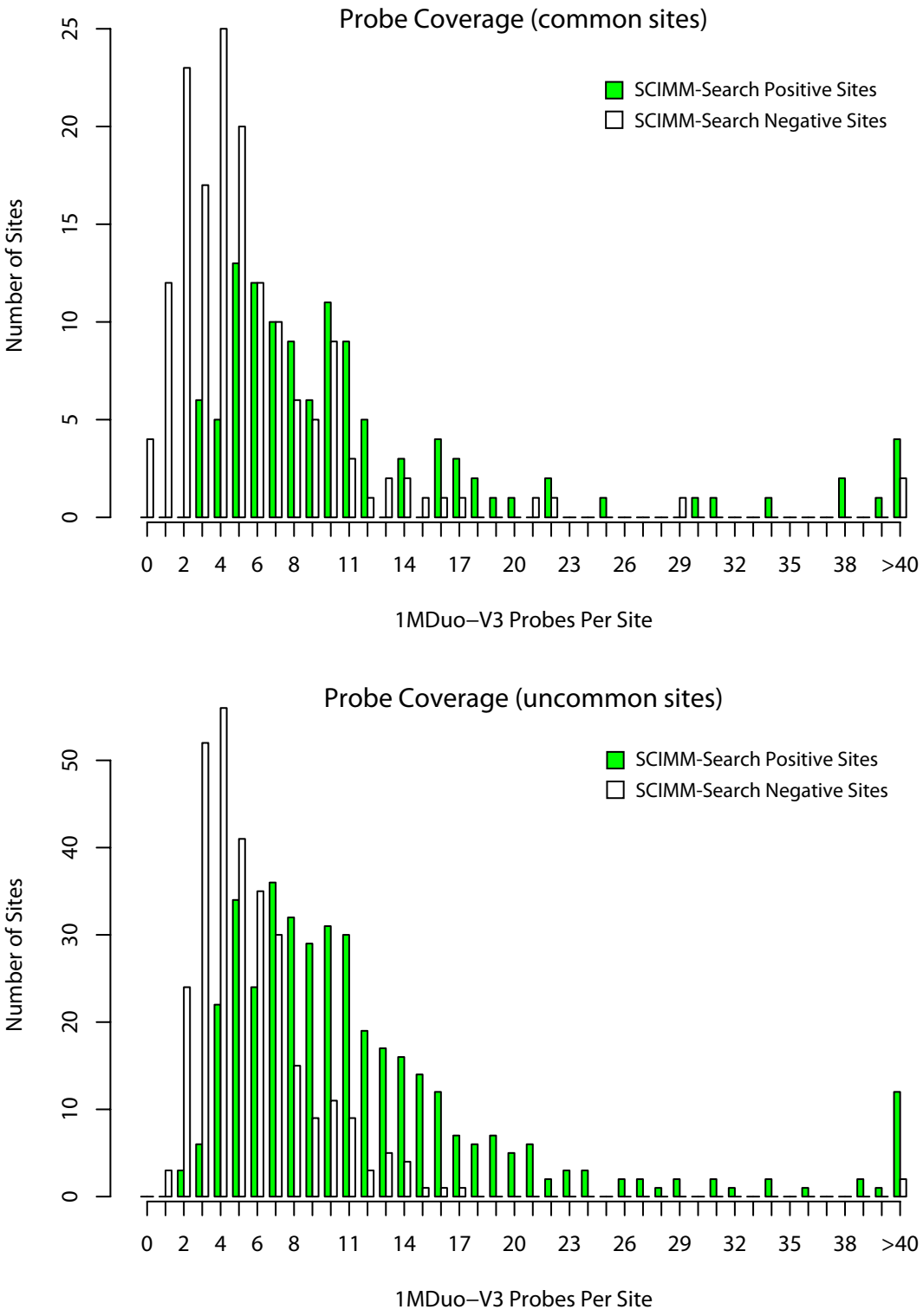

FIGURE S3. Probe coverage for SCIMM-Search positive and SCIMM-Search negative sites. We used SCIMM-Search to search for copy-number probes for 966 previously reported biallelic deletion sites (McCarroll and Kuruvilla *et al*. 2008). *Top:* 272 sites with deletion allele frequency at least 5%; (113 SCIMM-Search positive, 159 SCIMM-Search negative). *Bottom:* 694 sites with deletion allele frequency below 5% (392 SCIMM-Search positive, 302 SCIMM-Search negative).