Supplementary Information

Systematic assessment of copy-number variant detection via genome-wide single nucleotide polymorphism genotyping

Gregory M. Cooper[1,2,4], Troy Zerr[1,2], Jeffrey M. Kidd[2], Evan E. Eichler[2,3], and Deborah A. Nickerson[2]

[1] These authors should be regarded as co-first authors.
[2] Department of Genome Sciences, University of Washington, Seattle, WA 98195-5065
[3] Howard Hughes Medical Institute

[4] Corresponding author: University of Washington School of Medicine
Department of Genome Sciences
Box 355065
Foege S433D, 1705 NE Pacific St.
Seattle, WA 98195
E-mail: coopergm@u.washington.edu

The Supplementary Methods are organized into sub-sections by topic, appearing in the order that they appear in the main text. Section 1 includes a description of all the data sets generated and/or analyzed here. Sections 2-4 describe the CNV discovery methodology and validation. Sections 5-13 describe the methods underlying SCIMM and SCIMM-Search, and also describe the validation and quality-control for the deletion genotyping results presented in the main text. Sections 14-15 describe the effects of probe density on CNV detection. Finally, Section 16 lists the identities of the HapMap samples used.

Methods

1. Primary data sets analyzed

1. 368 CNVs (258 deletions, 110 amplifications) inferred using HMMSeg on Illumina Human 1M SNP array data for the 9 samples analyzed in Kidd et al. (Kidd et al. 2008). This set constitutes our primary set of annotations generated here and used for subsequent validation, size comparison, etc, as described in the main text and below. The size thresholds used to generate these events are a minimum of 10 probes for hemizygous deletions and amplification events and 3 probes for homozygous deletions.

2. 906 CNVs (691 deletions, 215 amplifications) inferred using HMMSeg for the same 9 samples using minimal size criteria of 5 probes for hemizygous deletions and amplifications and 2 probes for nulls.

3. Fosmid End-Sequence-Pair (ESP) maps generated and described in Kidd et al. 2008. These data consist of both insertion and deletion events defined as described below. These data also include the map locations for ~900,000 fosmid ESPs for each genome, some of which can be used to support the presence of a CNV even if a particular CNV was not previously annotated. All of these data are publicly available at http://hgsv.washington.edu

4. 500 deletion loci originally identified by fosmid ESP maps that were subsequently validated by oligonucleotide array-CGH in Kidd et al. The array-CGH experiments used two independent sets of custom arrays (from Agilent and Nimblegen) with probes spaced 100-300 bp apart, offering both validation and refined breakpoint information. To identify these oligonucleotide-array-CGH defined deletion events from the supplementary information in Kidd et al., we extracted all those events discovered as a result of 'large' fosmids (type = "B", indicating a deletion) that were marked as 'validated' and that had refined breakpoints defined either by Nimblegen array-CGH or Agilent array-CGH. In cases where both CGH experiments generated breakpoints, we used the average of the two sets of coordinates. This identifies 812 deletion events. For purposes of estimating sensitivity, we compared our SNP-based deletion predictions to 'detectable' deletion events, which we defined as those deletion annotations that spanned 10 or more probes on the Illumina Human 1M array. For the purpose of probe coverage estimation, we collapsed all overlapping deletions to allow for breakpoint uncertainty, yielding 500 non-redundant deletion loci.

5. 119 deletions originally identified by fosmid ESP maps that were validated by complete fosmid re-sequencing followed by alignment back to the reference assembly (hg17/ncbi35). To obtain these sequence-confirmed events from the supplementary information in Kidd et al., we extracted all those events annotated based on sequence analysis to be simple deletions (sequence validation = "D") with defined sequence breakpoints. We filtered out events that were determined

by sequence to be smaller than 1 kb and one event that was determined to be identical in 2 samples, yielding a total of 119 distinct deletion loci. Note that 2 of these loci overlap, but since they have distinct breakpoints and the breakpoints are not ambiguous, we treat them as distinct events. As above, for purposes of estimating sensitivity, we compared our SNP-based deletion predictions to 'detectable' deletion events, which we defined as the deletion annotations that spanned 10 or more probes on the Illumina Human 1M array.

6. Reference genotypes for 34 polymorphic sites of sequenced structural variation, generated using custom Illumina Goldengate assays and quantitative PCR (Kidd et al. 2008, Supplementary Information section 6.1 and Table S12). Of 34 sites, 18 of these sites are autosomal deletions spanning two or more Illumina 1M probes; 13 of these sites were genotypable by SCIMM using the Illumina 1M assay data. SCIMM-generated genotypes and reference genotypes are presented in Supplementary Table 7.

7. Genotypes for 126 deletion sites determined by SCIMM to be polymorphic, based on analysis of 126 HapMap samples (including 7 replicates) using Illumina Human 1M data. These events constitute a subset of the deletions identified by fosmid ESP mapping and subsequently validated by sequencing and/or array-CGH. Note that as a result of this usage of independent experimental data, we are confident that all of the sites genotyped by SCIMM do, in fact, correspond to deletion events and not potential artifacts of SNP array data.

8. Genotypes for 78 deletion sites genotyped by SCIMM using Illumina HumanHap 550 data. These sites are a subset of the 126 deletion sites above.

2. CNV discovery using 1M genotyping data

Large CNV discovery was accomplished using a simple Hidden Markov Model (HMM), implemented using the HMMSeg software package (Day et al. 2007), that analyzes both 'LogR Ratio' and 'B-allele Frequency' data simultaneously. We employed a 4-state model, corresponding to null (homozygous deletion), hemizygous, diploid, and greater than diploid states, and assumed a Gaussian distribution for both standard-normal transformed LogR and square root-transformed deviations from ideal B-allele Frequency. LogR values were transformed to a standard normal distribution on a per-sample, per-chromosome basis. Deviations from 'ideal' B-allele Frequency values were determined based upon the SNP call: for 'AA' homozygotes the ideal value is 0, for 'AB' heterozygotes this value is 0.5, and for 'BB' homozygotes this value is 1. For probes without a SNP call, the ideal value was set at the closest value selecting from 0, 0.5, or 1.

State definitions were established based on the methodological principles described in Peiffer et al (Peiffer et al. 2006). Homozygous deletion events are expected to have very low total intensity (LogR < -1) values and B-allele Frequency values that are randomly distributed between zero and 1 (since the ratio of the two non-existent alleles is undefined). Hemizygous deletion events are expected to have lower than normal

intensity values (LogR < 0), and be populated exclusively (or nearly so) by non-heterozygous SNP calls (B-allele frequency values of near zero or near 1). Amplification events are expected to have higher than normal intensity values (LogR > 0), and heterozygous SNPs should have an allelic imbalance such that the B-allele Frequency values are shifted away from 0.5 towards intermediate values (ie 0.3 or 0.7). Model parameters including means, variances, and transition probabilities were selected on the basis of manual analysis of results.

We merged the initial CNV predictions (i.e. sets of consecutive probes where the assigned state is not diploid) that were less than 5 probes and 10 kb apart and filtered the CNV annotations based on a number of quality-control thresholds. For homozygous deletion predictions, we required events to be at least 3 probes and 1 kb in length with an average LogR value less than -1. For hemizygous deletion events we required at least 10 probes and 1 kb in length with an average LogR value < -0.25, and the proportion of heterozygous SNP calls to be less than 10%. For amplification events we required a minimum of 10 probes and 1 kb in length, LogR values to be greater than 0.25, and B-allele frequency deviation values at heterozygous SNPs greater than 0.05. If no heterozygous SNPs existed within the variant, the latter restriction was eliminated. Probes targeting HLA regions were excluded from these analyses.

3. CNV validation using genome-wide fosmid ESP placement analysis

We used the entire fosmid ESP map (~900,000 clones) for each library, rather than relying solely on CNV annotations (clone mapping information can be obtained at http://hgsv.washington.edu). To explain our approach, it is important to consider the ESP mapping and CNV identification methodology that was previously used (Kidd et al. 2008). Clone placements were determined based on a 13-point heuristic scoring system that favors concordant placements and considers alignment length, identity, orientation, and read quality (Tuzun et al. 2005). A 'concordant' clone has an ESP placement with internally oriented reads and an *in silico* size (ie. size based upon the distance between the end placements in the reference assembly, hg17/ncbi35) that is similar to the expected average physical size for a fosmid (~40 kb). A clone is 'discordant' if the ends are placed too far apart, too close together, or in inappropriate orientations. A variant 'site' is defined by two or more overlapping clones from the same library that show the same type of discordancy. In order to contribute to a variant 'site', discordant clones were required to pass additional quality criteria not applied to concordant clones (see below). Kidd et al. required that clones must be more than 3 standard deviations larger (for deletions) or smaller (for insertions) than the average fosmid size for any given library.

Initial validation consisted of simply comparing the locations of variants inferred using the 1M data with the locations of variants inferred using fosmid ESP placements for the same sample. We compared only similar variant types (ie. 1M deletions compared only to ESP deletions, 1M amplifications compared only to ESP insertions). To compare size estimates for the validated, inferred deletions, we compared the sizes estimated for each overlapping deletion by fosmid ESP analysis to the Illumina 1M-inferred deletion length. Note that deletion size for ESP variants is determined by the discrepancy between the

observed and expected clone sizes, rather than the distance between the end-sequence placements; for example, a fosmid clone that spans 60 kb of the reference assembly is inferred to harbor a 20 kb deletion since the fosmid is assumed to be ~40 kb in actual size (Tuzun et al. 2005). In all but five cases of overlap, a given fosmid ESP deletion site overlapped with a single Illumina 1M-inferred deletion. In the five cases where multiple 1M-inferred deletions overlapped a single fosmid ESP deletion, the sizes of the 1M-inferred deletions were first summed before sizes were compared (main text Figure 2B).

We next considered relaxing size thresholds (Supplementary Table 3) to validate smaller deletions. We compared our list of non-validated deletion events to all fosmid ESP placements, and determined that an additional 18 of our deletion predictions overlap a site supported by 2 or more clones that are 2 standard deviations larger than the average fosmid, confirming that a small deletion is present.

As a further guard against false positive sites, Kidd et al. required that the discordant clones defining a variant site must pass a more stringent set of alignment criteria than a concordant one: both end-sequence alignments must be at least 400 nucleotides long and include at least 150 non-repeatmasked bases (with a 2% divergence threshold). As a result of these stringent criteria, particularly when coupled with randomness in clone coverage (each haplotype is spanned by ~5X physical coverage, but coverage varies across the genome), it is expected that some truly variant sites in any given sample will be missed due to a lack of a sufficient number of discordant clones. A deletion on chromosome 11 illustrates this phenomenon (Supplementary Figure 1). Based upon the Illumina 1M genotyping data, we inferred the presence of homozygous deletion events for the same region in G248 and ABC9. Both of these predictions are validated by fosmid ESP deletion sites. However, based on the 1M data, we also inferred the presence of overlapping (identical or nearly identical breakpoints) hemizygous deletions in ABC11, ABC12, and ABC13, yet these were not validated by corresponding fosmid ESP variant sites.

A closer examination of the fosmid placements for the three hemizygous samples reveals the presence of multiple discordant clones from these libraries: three from ABC11, two from ABC12, and one from ABC13. However, four of these six clones did not pass the more stringent criteria required for a deletion site, resulting in false negative ESP deletion predictions (Supplementary Table 4). Exploiting this rationale, we considered the presence of an ESP variant at the same location in a different sample as a source of secondary validation, and we find that 38 (~15%) deletions that are not directly supported by a fosmid ESP deletion in the same sample are supported by the presence of a fosmid ESP deletion in the same location within another sample. Thus, we conclude that at least 67% of all deletion predictions correspond to true positives.

We also sought to validate additional duplication events predicted from SNP genotyping information, as fosmid clones are incapable of spanning large (> 40 kb) insertions. We therefore considered how a tandemized amplification event of a single-copy segment of the reference assembly would appear in a fosmid ESP map. We hypothesized that clones spanning the breakpoints of a tandemization event should map to the reference assembly

with an ‘everted’ orientation, in which the end-sequences are aligned to the reference assembly with an orientation such that they point away from the center of the clone (Figure 3 in the main text). We subsequently identified clusters of overlapping clones within each sample that were ‘everted’, and established as a putative tandemization event any site supported by 2 or more such clones; 233 sites were identified within these 9 libraries, and these sites overlap 18 duplication events inferred using 1M SNP data. Finally, analogous to the rationale established above for deletions, we find that an additional 7 inferred amplification events overlap an ESP insertion event using smaller (2 standard deviation) size thresholds, and an additional 6 overlap an insertion event inferred for a different sample. Coupled with the annotations from (Redon et al. 2006), this brings the total validation rate to 64%.

We subsequently compared all the non-validated events to a complete list of fosmid ESP placements to determine if any additional predictions could be plausibly explained (Supplementary Table 5). We note that none of these categories validate a site *per se*, but may explain a lack of validation or provide suggestive validation:

1. Regions where no concordant fosmid maps. These regions of the genome are under-represented in the library and thus the lack of validation for variants in these intervals is not meaningful. These regions may alternatively correspond to homozygous sites of structural variation.

2. Regions that harbor one or more discordant clones that cannot be unambiguously assigned to a single location (termed ‘tied discordant’). A ‘tied discordant site’ is a region spanned by 2 or more such clones. These may reflect the presence of variants in duplication-rich regions.

3. ESP sites of variation (ie. harboring multiple discordant clones supporting the existence of a single variant) annotated using a less stringent sequence alignment scoring scheme which only considers alignment length and sequence identity (‘4 pt Sites’, see (Tuzun et al. 2005) for details).

4. Regions where one or more discordant clone maps, either uniquely placed (‘Best’) or mapped to multiple (‘Tied’) locations (see point 2), without regard to the more stringent quality-control filters (eg. Supplementary Table 4).

We find that 36% of the remaining non-validated deletions and 56% of the non-validated amplifications overlap one or more of the categories described above (Supplementary Table 5); this represents a substantial enrichment (3-6 fold) over the expected level of overlap assuming our predictions were distributed randomly in the genome.

We considered other plausible explanations for those sites that do not validate. First, we note that amplification events that are not tandemized will not validate through fosmid ESP placements: while the 1M assay determines that a particular segment of the genome is duplicated, if the duplicated copy exists elsewhere in the genome, the clones that capture the event will map to the insertion site rather than the original locus. Second,

some events will not validate via the BAC-CGH experiments performed in (Redon et al. 2006); these experiments did not include one of the 9 samples that we analyzed (NA15510) and furthermore utilized an individual reference sample that may itself harbor an amplification for any given region. The latter 'reference' effect is discussed in more detail in Kidd et al. Considering these observations and the overlaps shown in Supplementary Table 5, we conclude that many of the non-validated sites correspond to sites that either could not be validated or are at least suggestively validated. We therefore regard the validation rates described in the main text as a conservative estimate.

We also considered a relaxed stringency set of HMMSeg predictions by retaining all hemizygous deletions spanning five or more SNP probes and nulls of two or more. At these thresholds a total of 691 deletions and 215 amplifications are inferred (906 total events). Using the same validation process as described above, we find that this reduced stringency set has a lower overall validation rate, but independent experimental information confirms over 50% of these predictions (not shown). In terms of sensitivity, we detect a deletion overlapping ~52% (8/12 large sequence-defined deletions and 33/66 CGH-defined) of experimentally determined deletions, a marginal increase over the 47% sensitivity for the higher stringency set of predictions.

Finally, we compared the results from our *ab initio* CNV detection using HMMSeg to the results obtained from SCIMM genotyping (below) to assess both the specificity and the sensitivity of our single-sample CNV discovery approach. HMMSeg identifies 59 deletion-bearing samples at the 126 loci identified by SCIMM to be polymorphic; all of these 59 events were also classified to be deletions by SCIMM. Conversely, we find that ~80% (249 out of 308) of deletion-bearing samples identified by SCIMM were not predicted to be deletions by HMMSeg. The majority of the missing deletions occur at loci with enough probes (at least 2) to allow deletion genotyping but not enough to allow discovery (at least 3 for homozygous deletions, at least 10 for hemizygous deletions). Considering only those loci where HMMSeg infers at least one sample to harbor a deletion, for example, we find that 50% (59/118) of deletion-bearing samples are missed by HMMSeg. Furthermore, using our lower-stringency set of CNVs (5 or more probes for hemizygous deletions, 2 or more for homozygous deletions), this same estimation for the false negative discovery rate drops to 31%. Thus, relaxing size/probe count thresholds can increase discovery sensitivity, but still falls short of the accuracy desired for systematic genotyping.

5. Insertion/Deletion Genotyping: Overview

Our strategy for genotyping polymorphic deletion variants within populations of moderate size (~100-1000 samples) is implemented by two algorithms: SCIMM, a clustering algorithm which derives genotypes from a set of copy-number-informative probes, and SCIMM-Search, a search algorithm which determines a set of copy-number informative probes for each annotated deletion variant. The central idea is to automatically estimate the location and extent of the 6 clusters that appear in intensity data for SNP probes specific to the site of a common deletion: specifically, the three diploid clusters ('AA', 'AB', and 'BB'), the two hemizygous clusters ('A -' and 'B -'),

and the null (homozygous deletion) cluster ('- -'; see Figure 4 in the main text). Estimated locations are used to assign a copy-number label of 'null', 'haploid' or 'diploid' to each sample.

It is assumed by SCIMM that within each deleted site there is a set of SNP probes such that:

> (1) The copy number of each sample is the same for each probe; 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) Haploid ('A -', 'B -') samples have been given heterozygous ('AA', 'BB') SNP genotypes. (We later describe exceptions to this assumption.)
>
> (4) For any probe, log-transformed A-allele fluorescence measurements (*x*-coordinate values) for 'A-' and 'AA' samples, and B-allele fluorescence measurements (*y*-coordinate values) for 'B-' and 'BB' samples, are each normally distributed with equal variance for all four classes.

In practice, one cannot assume that all probes that map within the annotated boundaries of a deleted region satisfy assumptions (1)-(4). Inaccurate mapping of the deletion site or SNP probes may lead to the presence of probes within the provided boundaries but outside the actual deletion, and cross-hybridization to non-specific probes may lead to incorrect SNP calls and inconsistency in copy number. The goal of SCIMM-Search is to identify a set of 'informative' probes satisfying these assumptions.

6. SCIMM

SCIMM uses two-channel fluorescence and SNP genotype data for one or more probes to produce a putative classification of each sample as copy number 0, 1, or 2 ('null', 'haploid' or 'diploid') and a statistical score indicating the improvement in fit of a copy-number variant statistical model of probe fluorescence over a copy-number invariant model.

'Null' labels are assigned by an initial round of mixture-likelihood based clustering using the Expectation Maximization algorithm (Dempster et al. 1977). SCIMM first computes a univariate summary intensity value r_i for each sample. In accordance with assumptions (1) and (2) above, 'null' samples are modeled as a single univariate normal distribution centered at zero

$$\begin{aligned} P_0(i \mid \theta) &= P(r_i \mid \text{sample } i \text{ is null}, \theta) \\ &= (2\pi\sigma^2)^{-1/2} \exp(-(r_i / 2\sigma)^2) \end{aligned}$$

and the remaining samples are modeled as a mixture of two normal distributions with unconstrained location:

$$\begin{aligned} P_1(i \mid \theta) &= P(r_i \mid \text{sample } i \text{ is not null}, \theta) \\ &= (2\pi\sigma^2)^{-1/2}(\beta_1 \exp(-((r_i - \mu_1)/2\sigma)^2) + \beta_2 \exp(-((r_i - \mu_2)/2\sigma)^2)) \end{aligned}$$

where $\theta = \{\mu_1, r_1, \beta_1, \mu_2, r_2, \beta_2, \sigma\}$, ($\beta_1 + \beta_2 = 1$). The EM algorithm is used to find 'hidden' parameters $\hat{\theta}, \hat{\alpha}_1, \hat{\alpha}_2$, ($\hat{\alpha}_1 + \hat{\alpha}_2 = 1$) maximizing the mixture likelihood

$$L_0(\vec{r}, \hat{\theta}) = P(\vec{r} \mid \hat{\alpha}, \hat{\theta}) = \prod_i (\hat{\alpha}_0 P_0(i) + \hat{\alpha}_1 P_1(i)).$$

Under this model,

$$P(\text{sample } i \text{ is null} \mid \hat{\alpha}, \hat{\theta}) = \frac{\hat{\alpha}_0 P_0(i \mid \hat{\theta})}{\hat{\alpha}_0 P_0(i \mid \hat{\theta}) + \hat{\alpha}_1 P_1(i \mid \hat{\theta})}.$$

SCIMM assigns a 'null' label to any sample which is more likely to be null than non-null (i.e. $P(\text{sample } i \text{ is null} \mid \hat{\alpha}, \hat{\theta}) > 1/2$).

Following assumption (3), SCIMM assigns a 'diploid' label to all remaining samples with at least one heterozygous SNP genotype. SNP heterozygosity is not assumed to be informative of copy number for samples already labeled as 'null'; we allow the possibility that spurious SNP genotypes may have been assigned to deletion homozygotes.

Assumption (3) may be violated in cases where 'no call' SNP genotypes are provided for hemizygous samples or for CNV (non-SNP) probes. SCIMM employs a heuristic at this stage to assign SNP genotypes to samples near the *x* and *y* axes for which a 'no call' SNP genotype was provided as input. CNV probes (for which SNP genotypes are not supplied) are treated as monomorphic SNP probes.

SCIMM labels remaining samples as 'haploid' or 'diploid' by a second round of mixture-likelihood based clustering, using the EM algorithm to fit a two-component $2n$-dimensional mixture model to two-channel fluorescence data from n SNP probes. Observed fluorescence data for sample i at probe j are represented as (x_{ij}, y_{ij}), 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*

and projected, log-transformed fluorescence data are represented as $d_i = \{d_{ij}\}_{j=1..n}$, where

$$d_{ij} = \begin{cases} \log(x_{ij} + \varepsilon) & \text{if } s_{ij1} = 1 \\ \log(y_{ij} + \varepsilon) & \text{if } s_{ij2} = 1 \end{cases} \qquad (\varepsilon = 10^{-10}).$$

Following assumption (4), transformed data for samples of each copy number class

$$P(d_i \mid G_1, \theta) = P(d_i \mid \text{sample i is haploid})$$
$$P(d_i \mid G_2, \theta) = P(d_i \mid \text{sample i is diploid})$$

are modeled by the density function

$$P(d_i \mid G_c, \theta) = \prod_{j \le n, k \le 2, s_{ijk} = 1} (2\pi\sigma_j^2)^{-1/2} \exp(-((d_{ij} - \mu_{jkc})^2 / 2\sigma_j^2).$$

in terms of parameters $\theta = \{\mu_{jkc}, \sigma_j\}_{j \le n, k \le 2, c \le 2}$. In this model, μ_{j11}, μ_{j12}, μ_{j21}, and μ_{j22} represent the positions of the 'A -', 'AA', 'B -', and 'BB' clusters of probe *j*, respectively, and the log-variance σ_j^2 represents the 'noisiness' of probe *j*. In the case where none of the probes are SNP-polymorphic, this model simplifies to a homoscedastic mixture of multivariate lognormal distributions, each with a diagonal covariance matrix.

As before, the EM algorithm is used to estimate $\hat{\theta}, \hat{\alpha}_1, \hat{\alpha}_2$, $(\hat{\alpha}_1 + \hat{\alpha}_2 = 1)$ maximizing the mixture likelihood

$$L_0(\vec{d}, \hat{\theta}) = P(\vec{d} \mid \hat{\alpha}, \hat{\theta}) = \prod_i \sum_c \hat{\alpha}_c P(d_i \mid G_c, \hat{\theta})$$

and each remaining sample is labeled 'haploid' or 'diploid' accordingly.

To regularize parameter estimation, SCIMM treats samples that are SNP-heterozygous for at least one probe as pseudo-observations of diploid samples, consistent with assumptions (1) and (3). Specifically, SCIMM treats these samples in the same manner as samples of unknown copy number, but forces $P(d_i \mid G_1, \theta) = 0$, $P(d_i \mid G_2, \theta) = 1$ in the E-step of the EM algorithm.

7. Probe Set Scores

The Bayesian Information Criterion (BIC) of Schwarz (Schwarz 1978; Fraley and Raftery 2002) is used by SCIMM-Search to compute probe set scores for model

selection. Specifically, for any probe set *P,* SCIMM computes a maximum-likelihood one-component model during the 'haploid/diploid' clustering step, and computes a probe set score S_P as the difference of BIC values between the one component model and two-component model. One generally accepts the two component model (and the resulting 'haploid' and 'diploid' classifications) if $S_P > 0$.

Scores are also be used to evaluate alternate probe sets for a given deletion site; if $A \subset B$, and the set of inferred null homozygotes is the same for both A and B, $S_A - S_B$ is the difference between the BIC for the two-component model for probe set B and the BIC for a two component model with the added constraint $\mu_{jk1} = \mu_{jk2}$ for $j \in B \setminus A$. Thus, if $S_A - S_B > 0$, one prefers the model for which probes in A are informative for copy number but probes in $B \setminus A$ are not (and thus prefers probe set A over probe set B).

8. SCIMM-Search

SCIMM-Search is a search algorithm which, given the coordinates of a region spanning a deletion variant and optionally, the identity of one or more samples known to carry the deletion variant, determines a set of probes that can be used to genotype the deletion variant carried by the reference sample. It is not assumed that all probes within the annotated region are informative for copy number or that informative probes are contiguous. The algorithm first finds two probes (not necessarily adjacent) which produce the highest score (subject to the constraint that the reference sample is scored by the probe set as 'haploid' or 'null', and subject to additional constraints described below). Probes are then added incrementally; at each iteration, the probe producing the greatest score improvement (subject to similar constraints) is added. The search terminates when all probes are used, or every possible addition fails to meet a "minimum progress" criterion. The following parameters control the search algorithm:

$S_{\min}$: minimum initial probe set score. If no two probes within the annotated region produce a score exceeding $S_{\min}$, the algorithm terminates without producing a probe set.

$d_{\min\,\mathrm{init}}$: Minimum distance between the two initial probes. This criteron prevents generation of spurious genotypes due to autocorrelation of closely spaced probes.

$d_{\max\,\mathrm{init}}$, and $d_{\max}$: the maximum distance between the first two probes, and the maximum distance between incrementally added probes and the existing probe set. These serve primarily to limit the computational complexity of the search.

$\Delta S_{\min}$: Minimum progress threshold: Each probe added to the probe set must increase the probe set score by at least this amount.

We also specify constraints the cluster separation δ_p and consistency κ_P, where δ_p is the Mahalanobis distance between component means for each SNP allele, weighted by the number of SNP heterozygous samples for each allele, and κ_P is the fraction of samples classified as haploid using the two-component mixture model that would be classified as haploid by every single-probe projection of the model.

9. Implementation:

SCIMM and SCIMM-Search are implemented in **R** (http://www.r-project.org/). The following parameters and constraints were used to evaluate all regions in this study:

$S_{\min} = 60$
$\Delta S_{\min} = 30$
$d_{\min\ \text{init}} = 20$ bp
$d_{\max\ \text{init}} = d_{\max} = 10000$ bp
$3.0 \le \delta_p \le 8.0$ for all $p \in P$
$\kappa_P \ge 0.75$

Parameters and constraints were chosen empirically using cross-platform comparisons of the Human1M and HumanHap550 data sets described above. For CGH-refined deletions in which multiple reference samples carrying the deletion allele were specified, we required that at least one reference sample carry the deletion allele in the genotypes produced by each probe set.

10. Concordance with reference genotypes

We compared SCIMM-generated insertion/deletion genotypes to an independently generated set of reference genotypes for 18 autosomal deletion sites (Kidd et al. 2008). SCIMM-Search generated a probe set for 13 of these sites; thus, we estimate that the assay conversion rate for SCIMM is ~72% (13/18). Counts of concordant and discordant genotypes and correlation between genotypes (r^2, calculated as in (Carlson et al. 2004; Moorhead et al. 2006)) are shown below (Supplementary Table 6). The sum of concordant and discordant samples at each site varies due to missing genotypes in the reference set. The concordance rate between the SCIMM and reference genotypes is 98.8% (1250/1265 genotypes identical); however, we expect that neither data set will be free of error, and thus advise that the concordance rate should not be interpreted as an exact estimate of error rate.

We manually reviewed each of the five cases for which SCIMM did not yield genotypes. In three cases, only one probe was insertion-allele-specific; a result attributable to probe non-specificity, or inaccurate mapping of either the probes or the deletion detected by the fosmid reference. In one case, no probe was insertion-allele-specific, and multiple probes displayed a pattern of cross-hybridization, with evident “AAB” and “ABB” clusters present. In one case, two probes appeared by eye to be insertion-allele specific, but did not satisfy the SCIMM-Search criteria described above.

11. SCIMM-Search Sensitivity

We applied SCIMM-Search to a set of 483 non-overlapping autosomal deletion events identified by ESP mapping and refined by Nimblegen array-CGH and/or fosmid re-sequencing (Kidd et al. 2008). 126 of these sites did not span any Illumina 1M probe and 105 of these sites spanned only probe. Of the remaining 252 sites spanning two or more probes (the minimal probe coverage requirement for SCIMM-Search), the algorithm yielded probe sets in 136 cases, corresponding to an assay conversion rate of 54% (136/252). There are two potential sources of genotyping failure: the deletion may be common but fail to span 2 or more copy-number responsive probes, or the deletion allele may be at low frequency or private to a single individual. The number of sites yielding genotypes (as a function of probe count) is displayed in Supplementary Figure 2.

12. Validation of SCIMM-Search by manual inspection of quantitative SNP assay data

We manually inspected scatter plots of normalized fluorescence data for every probe within each region for which SCIMM-Search generated a probe set, including probes which were not used for estimating genotypes (1009 probes total). At most sites, multiple probes exhibited scatter plot patterns similar to that generated by GoldenGate probes within insertion/deletion regions (Carlson et al. 2006, McCarroll et al. 2006, Newman et al. 2006); however, cluster separation was generally poorer than the examples displayed in GoldenGate-based studies, and deletion allele frequencies were generally lower. Inconsistencies between manual and automatic analyses were detected at six sites:

chr12:11395556-11448902 (reference samples: NA12878, NA19240, NA12156)
Diploid (I/I) class contains copy-number-3 and copy-number-4 classes

chr12:34098698-34108391(reference sample: NA19240)
Haploid (I/D) samples are null (near origin) at rs7358760; remaining probes may not be insertion-allele specific

chr15:54454331-54683686 (reference sample: NA18507)
Probe set (chr15:54579805-54582930) partially spans region; deletion carried by NA18507 spans entire region

chr19:47998406-48237071 (reference sample: NA18956)
Informative probes are sparse throughout region

chr19:59418356-59439332 (reference sample: NA15510)
Diploid (I/I) class contains copy-number-3 and copy-number-4 classes

chr20:14718107-14887610 (reference sample: NA15510)
Probe set (chr19:14745572-14773787) partially spans region.

13. Inconsistencies with Mendelian transmission

In our analysis of 126 polymorphic insertion/deletion sites, we detected six sets of trio genotypes that were inconsistent with Mendelian transmission of the deletion allele (Supplementary Table 9). Five of these inconsistencies are directly attributable to genotyping errors; in each case one sample in the trio was assigned a genotype of I/I by SCIMM, but could be plausibly assigned genotype I/D by manual analysis (Supplementary Table 10). Genotypes thus assigned are consistent with Mendelian transmission in each case. We conclude that Mendelian inconsistencies in the SCIMM-generated genotypes are the result of genotyping errors and not an indication of novel mutational events.

14. Illumina HH 1M vs. HH 550K comparison

We investigated the applicability of SCIMM to lower-density assays by analysis of Illumina HumanHap 550 data. At each site, we used all HumanHap 550 probes within the probe set previously identified in the 1M assays. For 50/124 polymorphic sites, two or more informative probes were present in the 550 assay; 99.8% (5887/5900) of all genotypes are concordant between the two assays, with genotype correlation (r^2) exceeding 0.8 at 48/50 sites. The fraction of discordant genotypes at sites with only one informative probe in the 550 assay is much greater; only 98.3% (3247/3302) of all genotypes are concordant, with genotype correlation (r^2) exceeding 0.8 at only 21/28 sites (Supplementary Figure 3). These results demonstrate a very high degree of technical reproducibility across genotyping experiments, and indicate that informative probes identified by analysis of high-density SNP data can be applied to lower-density assays; however, insertion/deletion genotypes from single SNP probes are less accurate than those produced from multiple probes.

15. CNV resolution and SNP array probe coverage

We compared SNP array probe coverage estimates between BAC array-CGH annotations (Redon et al. 2006) and fosmid ESP variant annotations (Kidd et al. 2008). We considered only those sites that were discovered in the same location and in the same sample by the two methods, ensuring the same event is being independently annotated (~60 events). We find that there is a significant inflation in estimates of probe coverage when lower-resolution annotations are used: while the vast majority of the annotations generated by BAC-CGH span 10 or more probes on even the oldest arrays (eg. ~80% for the Affy 500K array; Supplementary Figure 4), the breakpoints offered by fosmid ESP mapping reveal that these same variants contain many fewer probes (eg. 15% of events span 10 or more probes on the Affy 500K array). Additionally, we note that fosmid ESP

coordinates (in the absence of refinement by array CGH or sequencing) are themselves somewhat inflated (generally ~10 kb on either end of the variant); thus, actual coverage for these events is even smaller than these estimates would indicate.

16. Samples genotyped on the Illumina Human 1M BeadArray:

HapMap CEU Samples:

NA06985	NA06991°	NA06993^{+}	NA06994
NA07000	NA07029°	NA07345	NA07348°
NA07357	NA10830°	NA10835°	NA10847°
NA10851°	NA10857°	NA10859°	NA10860°
NA10861°	NA10863°	NA11881	NA11882
NA11992	NA11993^{+}	NA11994	NA11995
NA12043	NA12044	NA12056	NA12057
NA12146	NA12154	NA12156*	NA12234
NA12236	NA12239	NA12248^{+}	NA12249
NA12264	NA12740°	NA12750	NA12751
NA12801°	NA12812	NA12813	NA12865°
NA12874	NA12875	NA12878°*	NA12891
NA12892			

HapMap CHB+JPT Samples:

NA18529	NA18537	NA18542	NA18547
NA18550	NA18555*	NA18558	NA18576
NA18577	NA18593	NA18608	NA18609^{+}
NA18612	NA18944	NA18949	NA18951^{+}
NA18952	NA18953	NA18956*	NA18960
NA18965	NA18968	NA18971	NA18972
NA18978	NA18987	NA18992	NA18994
NA18999	NA19003		

HapMap YRI Samples:

NA18502	NA18507*	NA18515°	NA18516
NA18517*	NA18852	NA18853	NA18854°
NA18855	NA18856^{+}	NA18857°	NA19092
NA19093	NA19094°	NA19116	NA19119
NA19120°	NA19129*	NA19137	NA19138
NA19139°	NA19140	NA19141	NA19142°
NA19143	NA19144	NA19145°	NA19159
NA19160	NA19161°	NA19171	NA19172
NA19173°	NA19192	NA19193	NA19194°
NA19238^{+}	NA19239	NA19240°*	

Non-HapMap Sample:

NA15510*

Legend:

+ : Sample genotyped in replicate (7 samples)

* : Individual with corresponding fosmid library (9 samples)

° : Offspring with both parents genotyped (28 trios)

Tables

Supplementary Table 1. SNP array probe counts within non-redundant deletions discovered by fosmid ESP mapping and subsequently validated with either complete fosmid re-sequencing (119 non-redundant sites, median size ~9.9 kbp) or high-density oligonucleotide array-CGH (500 non-redundant sites, median size ~6.3 kbp).

Non-Redundant Oligonucleotide Array-CGH Deletions, n=500

Probe Count	*Affy 500K*	*Affy 5.0*	*Affy 6.0*	*Ill 317K*	*Ill 550K*	*Ill 1M*
0	68%	22%	19%	75%	62%	24%
1	15%	17%	12%	16%	18%	20%
2	6%	11%	12%	4%	7%	15%
3-4	7%	14%	16%	3%	6%	17%
5-9	3%	11%	16%	1%	4%	16%
10+	1%	26%	25%	1%	3%	9%

Non-Redundant Sequenced Deletions, n=119

Probe Count	*Affy 500K*	*Affy 5.0*	*Affy 6.0*	*Ill 317K*	*Ill 550K*	*Ill 1M*
0	74%	30%	32%	82%	69%	29%
1	12%	8%	6%	8%	10%	15%
2	5%	3%	7%	3%	8%	16%
3-4	5%	10%	7%	3%	4%	17%
5-9	1%	3%	14%	2%	7%	13%
10+	3%	46%	34%	3%	3%	10%

Supplementary Table 2. 258 deletion events and 110 amplification events in 9 samples detected by HMM-based segmentation (Excel spreadsheet provided as Supplementary Data on Nature Genetics web site).

Supplementary Table 3. Average and standard deviations for the 9 fosmid libraries analyzed previously.

Population	Library	Mean	SD	2SD	3SD
Unk	G248	39892	2747	5494	8241
Yoruba	ABC7	37593	3877	7754	11631
Yoruba	ABC8_a	36704	3848	7696	11544
Yoruba	ABC8_b	36088	1913	3826	5739
Japan	ABC9	39512	2260	4520	6780
Yoruba	ABC10	41005	1837	3674	5511
China	ABC11	40033	1768	3536	5304
CEPH	ABC12	39752	1396	2792	4188
Yoruba	ABC13	39289	1775	3550	5325
CEPH	ABC14	39442	1727	3454	5181

Supplementary Table 4. Summary of discordant clones from ABC11, ABC12, and ABC13 overlapping a deletion on chr11 predicted using Illumina 1M genotyping data.

Clone ID	Chrm	Begin	End	Size	Reason Excluded
ABC11_47366100_F20	chr11	55,103,359	55,230,933	127,575	< 150 non-RepeatMasked bases
ABC11_48043700_I16	chr11	55,098,780	55,229,713	130,934	< 150 non-RepeatMasked bases
ABC11_49656300_L11	chr11	55,089,874	55,218,669	128,796	
ABC12_46947900_P2	chr11	55,088,290	55,218,287	129,998	alignment < 400 bases
ABC12_47837400_K3	chr11	55,102,803	55,230,190	127,388	< 150 non-RepeatMasked bases
ABC13_47487100_C8	chr11	55,099,255	55,226,463	127,209	

Supplementary Table 5. Intersection between non-validated CNV predictions and fosmid ESP maps.

Intersection Class	Event Type: Deletions+Nulls	Event Type: Duplications
Total	86	39
Regions of No Coverage	16	15
Tied Discordant Sites (3SD)	1	0
4 pt' Sites (3SD)	5	0
>=1 'Best' Discordant Clone (3SD)	19	10
>= 1 'Tied' Discordant Clone (3SD)	12	8
Number With Intersection	31	22
Percent Intersecting	**36.0%**	**56.4%**
Randomized Overlap	6.13% +/- 0.24%*	17.59% +/- 0.65%*
Enrichment	5.9	3.2

*average +/- standard error determined by 100 independent randomizations

Supplementary Table 6. Concordance between SCIMM genotypes and reference genotypes produced by manual analyses of Illumina GoldenGate and quantitative PCR data.

Deletion Coordinates	Probe Count		Deletion Allele Frequency	Concordant Samples	Discordant Samples	Concordance Rate	Correlation (r^2)
	Total	Genotyping Probe Set					
chr1:34767144-34784461	5	2	0.14	115	0	100%	1.00
chr1:149572945-149583429	3	3	0.16	116	2	98%	0.94
chr7:97039956-97047292	3	2	0.13	115	0	100%	1.00
chr7:115524423-115535024	3	3	0.02	48	0	100%	1.00
chr8:51193494-51200884	4	4	0.18	77	2	97%	0.93
chr8:144771628-144785837	2	2	0.29	76	1	99%	0.97
chr10:70950995-70961085	3	3	0.17	96	6	94%	0.82
chr11:5740206-5765860	3	2	0.20	98	0	100%	1.00
chr11:106743432-106749364	3	3	0.29	61	1	98%	0.97
chr15:32481834-32604507	23	7	0.10	108	0	100%	1.00
chr19:40542993-40553524	2	2	0.18	110	0	100%	1.00
chr19:56824337-56840796	2	2	0.31	116	2	98%	0.96
chr22:37691312-37710639	6	4	0.11	114	1	99%	0.96

Supplementary Table 7. Comparison of SCIMM-generated biallelic insertion/deletion genotypes using the Illumina Human 1M assay with reference genotypes produced by manual analyses of Illumina GoldenGate and quantitative PCR data (Excel spreadsheet provided as Supplementary Data on Nature Genetics web site).

Supplementary Table 8. Analysis of SCIMM-generated biallelic insertion/deletion genotypes using the Illumina Human 1M and HumanHap 550 assays (Excel spreadsheet provided as Supplementary Data on Nature Genetics web site).

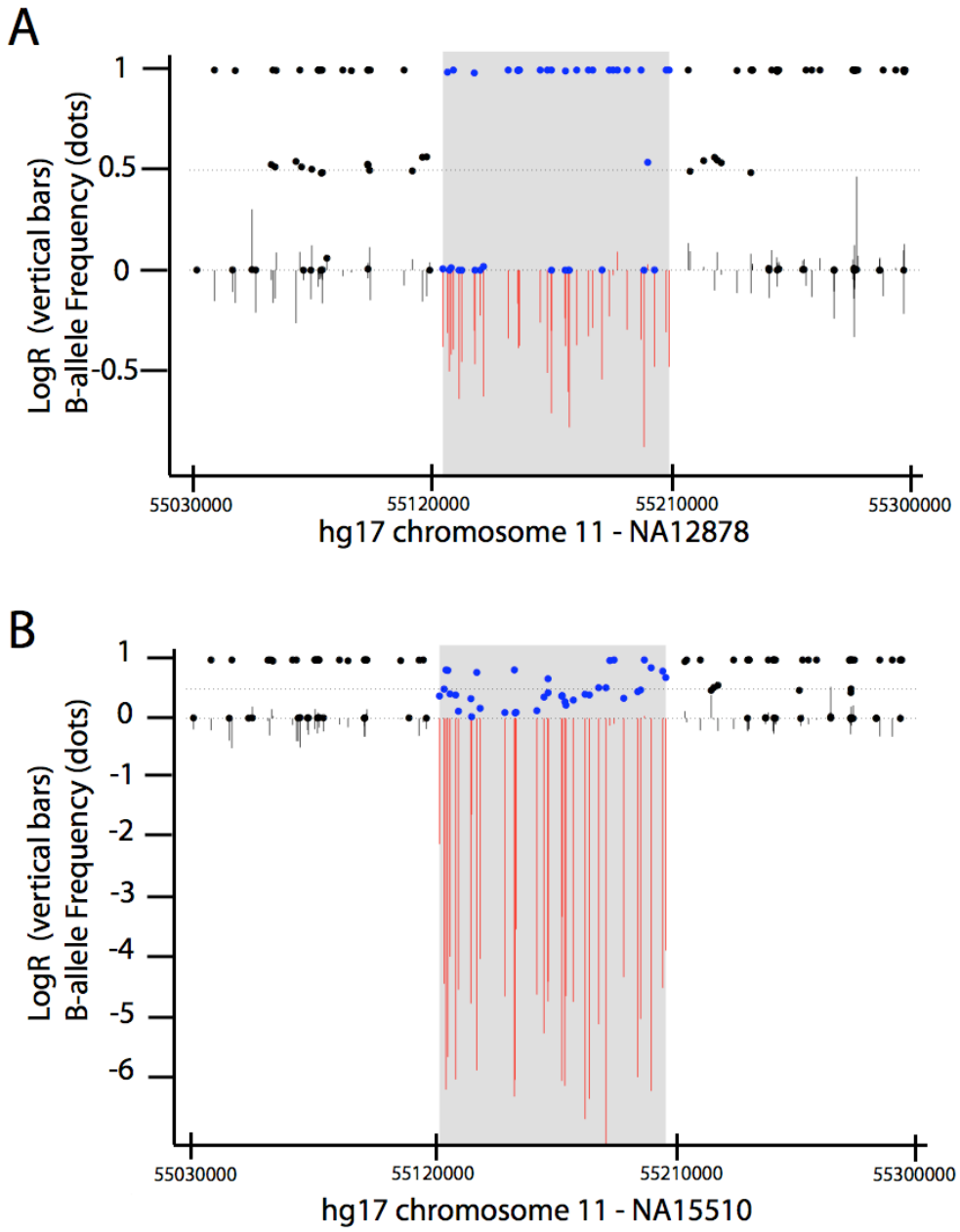
Supplementary Table 9. Trio genotypes inconsistent with Mendelian Transmission (HumanHap 1M assay, SCIMM genotype calls)

Site	Father	Paternal genotype	Mother	Maternal genotype	Offspring	Offspring genotype
chr1:149368616-149400815	NA19144	I/I	NA19143	D/D	NA19145	I/I
chr2:223690675-223697309	NA18856	I/I	NA18855	I/I	NA18857	I/D
chr8:144771628-144785837	NA18852	I/I	NA18853	I/I	NA18854	I/D
chr10:66976631-66985176	NA12248	I/I	NA12249	I/D	NA10835	D/D
chr17:32830114-32832327	NA12981	I/D	NA12892	I/I	NA12878	D/D
chr19:56824337-56840796	NA18852	I/D	NA18853	I/I	NA18854	D/D

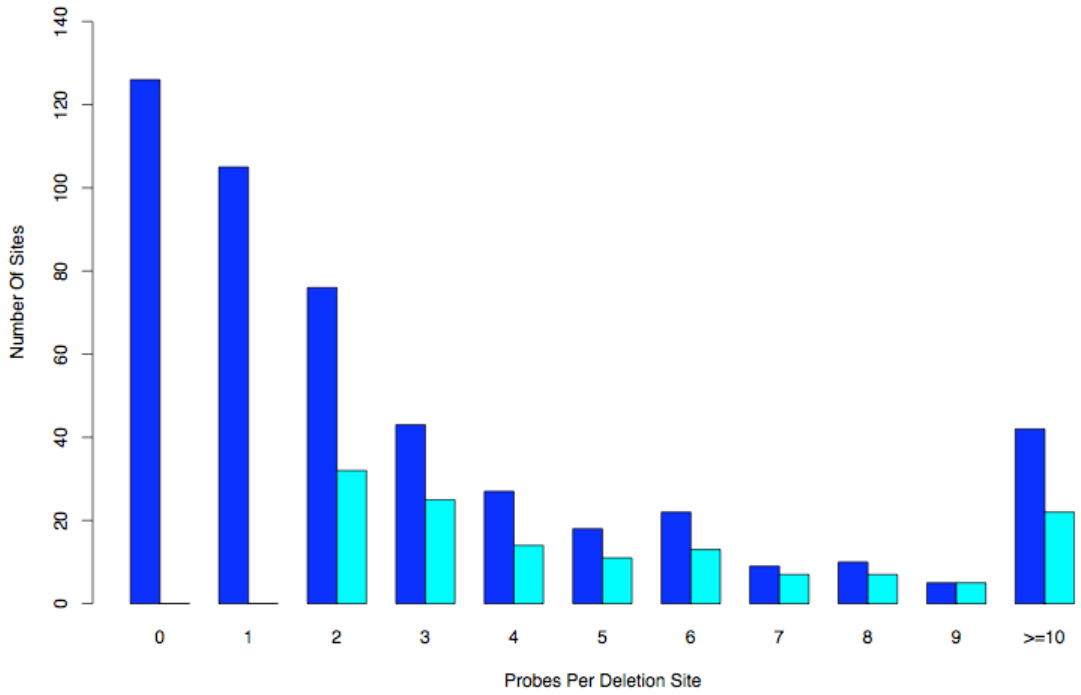
Supplementary Table 10. Manual analysis of inconsistent trios.

Site	Erroneous Sample	Manual Scatterplot Analysis
chr1:149368616-149400815	NA19145	NA19145 present at center of A/- cluster at rs12098109, center of B/B cluster at rs6668454
chr2:223690675-223697309	NA18856	NA18856 present at lower boundary of B/B cluster at rs12987860, lower boundary of A/A cluster at rs7588542 NA18856_R present at center of B/B cluster at rs12987860, center of A/- cluster at rs7588542
chr8:144771628-144785837	NA18853	NA18853 present at lower boundary of B/B cluster at rs35896889, lower boundary of A/A cluster at cnvGap_CNV_12447p62
chr10:66976631-66985176	NA12248	NA12248_R is scored by SCIMM as I/D
chr17:32830114-32832327	NA12892	NA12892 present at center of B/- cluster at rs9911273, center of B/- cluster at rs829158
chr19:56824337-56840796	*unknown*	NA18853 present at center of A/A cluster and NA18854 near origin at both probes in probe set

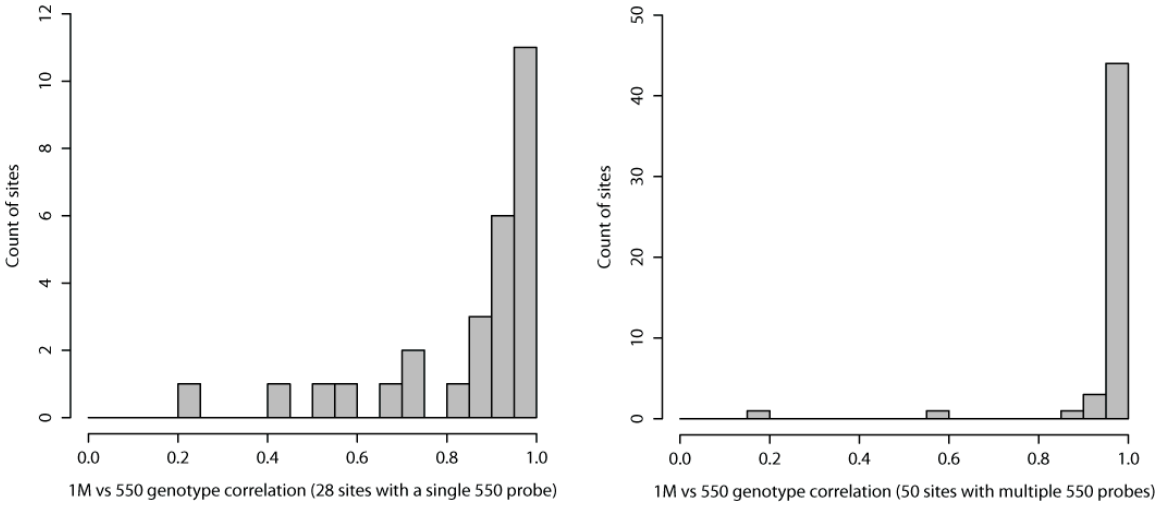
Figures



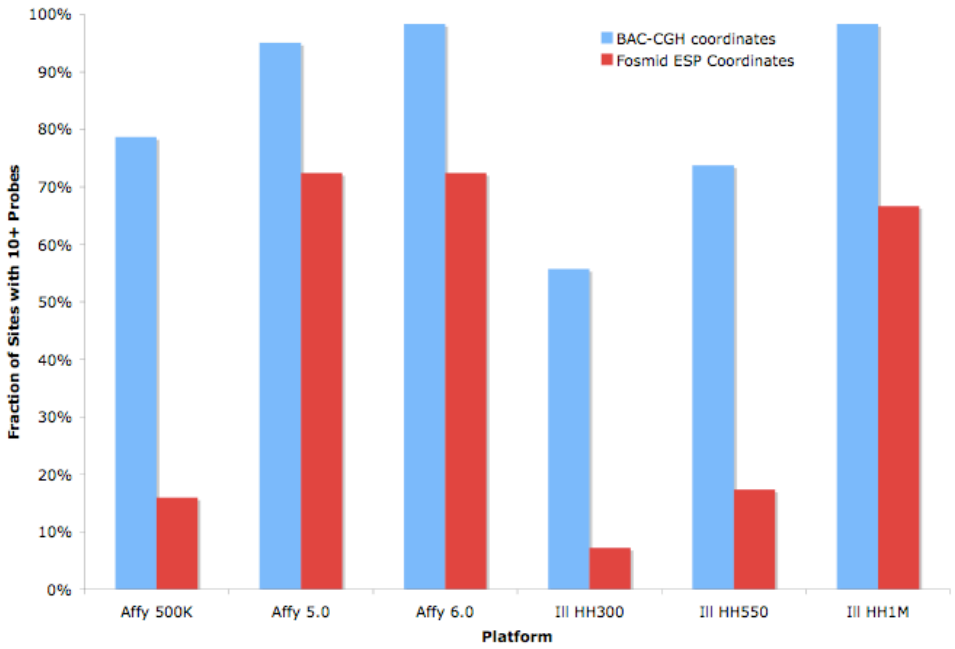
Supplementary Figure 1. Some inferred deletions that correspond to true events are not validated by a corresponding ESP deletion in the same sample. 1M SNP data for ABC11 (NA12878, panel A) and G248 (NA15510, panel B) are shown for the indicated region on chromosome 11 (X-axes are identical in both panels), with LogR plotted as vertical bars and B-allele Frequency plotted as dots (similar to Figure 2A in the main text). Gray boxes indicate the breakpoints of the deletion events inferred by HMMSEG using only the SNP data. The null event (B) is confirmed by ESP placements for the same sample, while the hemizygous deletion (A) is not. However, the correspondence in breakpoints between the two independently predicted events and the convincing visual impression strongly suggest that the deletion in A is real. In fact, discordant clones support this conclusion (Supplementary Table 4).



Supplementary Figure 2. Probe coverage histogram for autosomal deletion genotyping targets. SCIMM-Search was used to evaluate Illumina Human 1M SNP genotype data for non-overlapping deletions previously identified by fosmid ESP mapping and refined with array-CGH. 252 sites spanned 2 or more probes; 136 of these sights (light blue) yielded probe sets judged to be sufficiently high-quality for subsequent genotyping use by SCIMM.



Supplementary Figure 3. Correlation of insertion/deletion genotypes for 1M and 550 arrays. Correlation of genotypes (r^2) was calculated for each polymorphic deletion site that could be genotyped using both assays. Left: distribution of r^2 for deletions genotyped by only one 550 probes. Right: distribution of r^2 for deletion genotyped by using two or more 550 probes.



Supplementary Figure 4. BAC-CGH annotations tend to inflate probe coverage estimates for CNVs. The fraction of CNVs covered by 10 or more distinct probes (Y-axis) is plotted for each of 6 commonly used SNP genotyping platforms (X-axis), using breakpoints annotated by BAC-CGH experiments (blue) and fosmid ESP mapping analysis (red). These analyses are restricted to events independently discovered in the same sample at the same genomic location (n=60 total events in 8 HapMap samples).

Supplementary References

Carlson, C.S., M.A. Eberle, M.J. Rieder, Q. Yi, L. Kruglyak, and D.A. Nickerson. 2004. Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *Am J Hum Genet* **74:** 106-120.

Carlson, C.S., J.D. Smith, I.B. Stanaway, M.J. Rieder, and D.A. Nickerson. 2006. Direct detection of null alleles in SNP genotyping data. *Hum Mol Genet* **15:** 1931-1937.

Day, N., A. Hemmaplardh, R.E. Thurman, J.A. Stamatoyannopoulos, and W.S. Noble. 2007. Unsupervised segmentation of continuous genomic data. *Bioinformatics* **23:** 1424-1426.

Dempster, A.P., N.M. Laird, and D.B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B* **39:** 1-38.

Fraley, C. and A.E. Raftery. 2002. Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association* **97:** 611-630.

Kidd, J.M., G.M. Cooper, W.F. Donahue, H.S. Hayden, N. Sampas, T. Graves, N. Hansen, B. Teague, C. Alkan, F. Antonacci et al. 2008. Mapping and sequencing of structural variation from eight human genomes. *Nature* **453:** 56-64.

McCarroll, S.A., T.N. Hadnott, G.H. Perry, P.C. Sabeti, M.C. Zody, J.C. Barrett, S. Dallaire, S.B. Gabriel, C. Lee, M.J. Daly et al. 2006. Common deletion polymorphisms in the human genome. *Nat Genet* **38:** 86-92.

Moorhead, M., P. Hardenbol, F. Siddiqui, M. Falkowski, C. Bruckner, J. Ireland, H.B. Jones, M. Jain, T.D. Willis, and M. Faham. 2006. Optimal genotype determination in highly multiplexed SNP data. *Eur J Hum Genet* **14:** 207-215.

Newman, T.L., M.J. Rieder, V.A. Morrison, A.J. Sharp, J.D. Smith, L.J. Sprague, R. Kaul, C.S. Carlson, M.V. Olson, D.A. Nickerson et al. 2006. High-throughput genotyping of intermediate-size structural variation. *Hum Mol Genet* **15:** 1159-1167.

Peiffer, D.A., J.M. Le, F.J. Steemers, W. Chang, T. Jenniges, F. Garcia, K. Haden, J. Li, C.A. Shaw, J. Belmont et al. 2006. High-resolution genomic profiling of chromosomal aberrations using Infinium whole-genome genotyping. *Genome Res* **16:** 1136-1148.

Redon, R., S. Ishikawa, K.R. Fitch, L. Feuk, G.H. Perry, T.D. Andrews, H. Fiegler, M.H. Shapero, A.R. Carson, W. Chen et al. 2006. Global variation in copy number in the human genome. *Nature* **444:** 444-454.

Schwarz, G. 1978. Estimating the dimension of a model. *Annals of Statistics* **6:** 461-464.

Tuzun, E., A.J. Sharp, J.A. Bailey, R. Kaul, V.A. Morrison, L.M. Pertz, E. Haugen, H. Hayden, D. Albertson, D. Pinkel et al. 2005. Fine-scale structural variation of the human genome. *Nat Genet* **37:** 727-732.