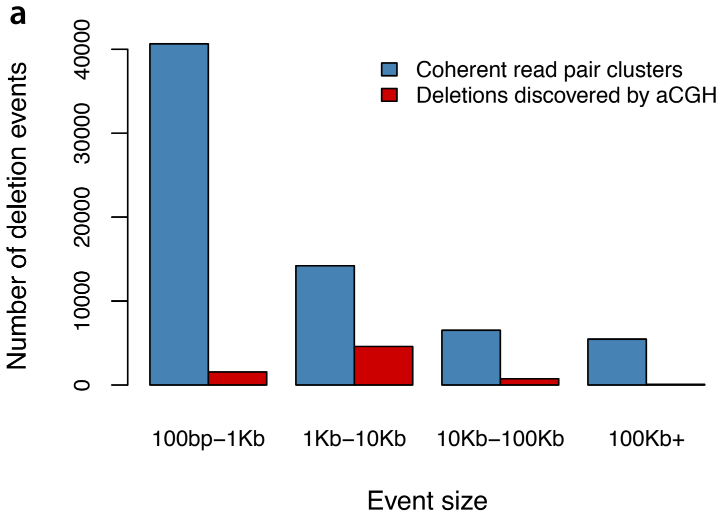
Discovery and genotyping of genome structural polymorphism by sequencing on a population scale

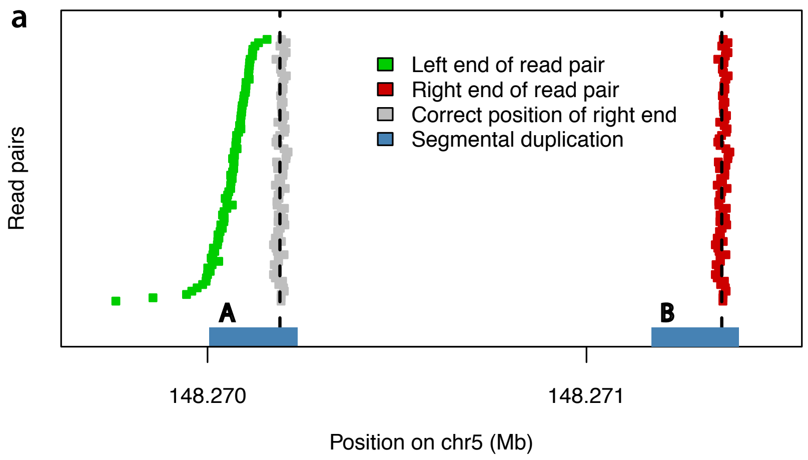
Robert E. Handsaker, Joshua M. Korn, James Nemesh, and Steven A. McCarroll

Supplementary Figure S1



Comparison of read pair clusters to deletions discovered by tiling-resolution array CGH
(a) Before population-genetic criteria were applied, the number of putative deletions from sequence data (genomic clusters of aberrantly spaced read pairs, indicated by height of blue bars) greatly exceeded a reasonable expectation given the number of deletions discovered by tiling-resolution array CGH (red bars) in the same 40 genomes in the study by Conrad *et al.* [1]. The Conrad *et al.* study, which employed tiling-resolution arrays (42M probes), had ample technical power to discover almost all deletions larger than 10 kb and the great majority of deletions larger than 1 kb in these 40 genomes. This indicated (as was ultimately confirmed) that a very large fraction of these aberrant read pair clusters (blue bars) arise not from real deletions but rather from molecular and alignment artifacts.

Supplementary Figure S2



b

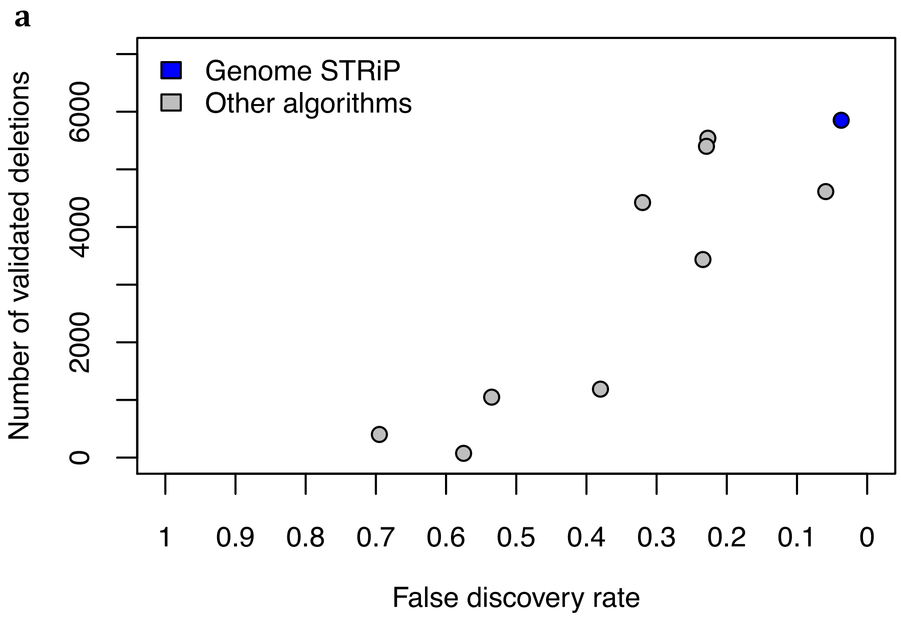
```
Copy A (ref)  GCTGACAAAACGCTGTTTTGGAAAGACTCATCTGGCAG CA CACACCAAGGCTGGTGGGGCA
Copy A (del)  GCTGACAAAACGCTGTTTTGGAAAGACTCATCTGGCAG -- CACACCAAGGCTGGTGGGGCA
Copy B        GCTGACAAAACACTGTTTTGGAAAGACTCATCTGGCAG -- CACACCAAGGATGGTGGGGCA

              gctgacaaaacgctgttttggaaagactcatccggcag    cacaccaa
               ctgacaaaacgctgttttggaaagactcatctggcag    cacaccaa
                tgacaaaacgctgttttggaaagactcatctggcag    cacaccaaggctg
                       acgctgttttggaacgcctcatctggcag    cacacca
                        cgctgttttgcaaaaactcatctggcag    cacaccaa
                         gctgttttggaaagactcatctggcag    cacaccaag
                           tgttttggaaagactcatctggcag    cacaccaaggctggtg
                            gttttggaaagactcatctggcag    cacaccaaggctg
```

Example of an artifactual read pair cluster that does not arise from a true deletion polymorphism

(a) Many aberrantly spaced read pairs map to the same locus on chr5. Green and red dots indicate the MAQ-aligned positions of the left ends (green) and right ends (red) of each sequence library insert. A segmental duplication is present at this locus (blue rectangles, with the two copies labeled A and B). An alternative potential location for the right ends, not selected by the aligner (but consistent with the insert size distribution of each library) is shown in gray.

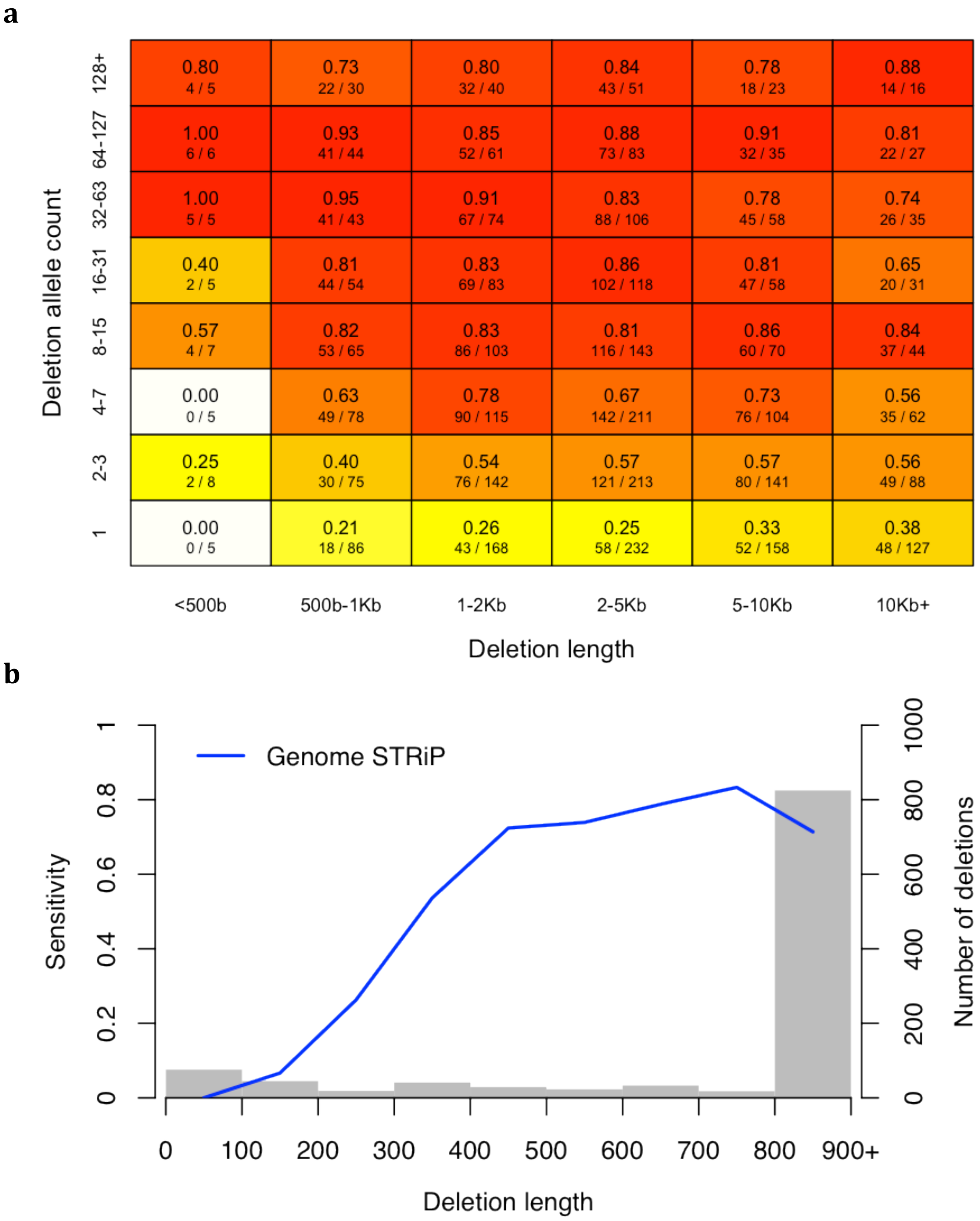
(b) The apparent misalignments appear to be caused by a indel (CA/–) polymorphism (rs35979103**,** red) which is polymorphic in Copy A. Reads arising from the "–" allele of Copy A are mis-aligned to copy B, because most current alignment algorithms impose large penalties for gaps. The correct alignment for the reads can be verified in this case by noting mismatches at two other sites (blue) that differ between Copy A and Copy B.

Supplementary Figure S3



Sensitivity and specificity of Genome STRiP and nine other deletion discovery algorithms
(a) Among ten algorithms used for deletion discovery in the 1000 Genomes Project [2] on population-scale, low-coverage sequence data, Genome STRiP (blue) had the lowest false discovery rate and simultaneously predicted the largest number of deletions that were subsequently validated using either PCR, array data or local assembly of short reads.

Supplementary Figure S4



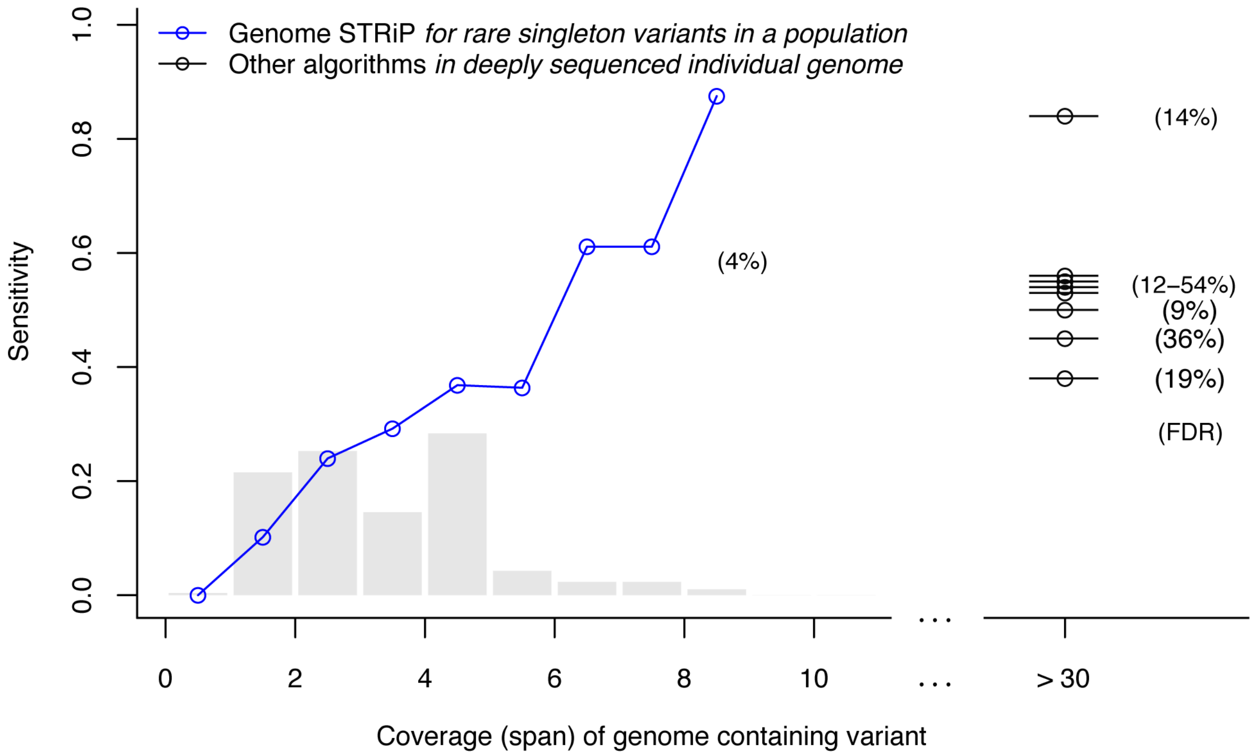
Deletion allele count	<500b	500b-1Kb	1-2Kb	2-5Kb	5-10Kb	10Kb+
128+	0.80 4 / 5	0.73 22 / 30	0.80 32 / 40	0.84 43 / 51	0.78 18 / 23	0.88 14 / 16
64-127	1.00 6 / 6	0.93 41 / 44	0.85 52 / 61	0.88 73 / 83	0.91 32 / 35	0.81 22 / 27
32-63	1.00 5 / 5	0.95 41 / 43	0.91 67 / 74	0.83 88 / 106	0.78 45 / 58	0.74 26 / 35
16-31	0.40 2 / 5	0.81 44 / 54	0.83 69 / 83	0.86 102 / 118	0.81 47 / 58	0.65 20 / 31
8-15	0.57 4 / 7	0.82 53 / 65	0.83 86 / 103	0.81 116 / 143	0.86 60 / 70	0.84 37 / 44
4-7	0.00 0 / 5	0.63 49 / 78	0.78 90 / 115	0.67 142 / 211	0.73 76 / 104	0.56 35 / 62
2-3	0.25 2 / 8	0.40 30 / 75	0.54 76 / 142	0.57 121 / 213	0.57 80 / 141	0.56 49 / 88
1	0.00 0 / 5	0.21 18 / 86	0.26 43 / 168	0.25 58 / 232	0.33 52 / 158	0.38 48 / 127

Sensitivity of Genome STRiP on rare and short deletions

(a) Sensitivity of Genome STRiP is shown as a function of both deletion length (x-axis) and frequency (y-axis) compared to deletions from Conrad *et al.* [1]. Sensitivity is reported both as a decimal (top number) and as a fraction (bottom numbers). Frequency is measured in 144 genomes overlapping between the two data sets; one base-pair overlap was used to calculate sensitivity.
(b) Sensitivity of deletion discovery for short deletions measured against a "gold standard" set [3, 4] that includes 150 shorter deletions ascertained from Sanger sequencing in sample NA12156 (blue line). Gray bars indicate the number of deletions in each bin (right hand scale). One base-pair overlap was used to calculate sensitivity.

Supplementary Figure S5

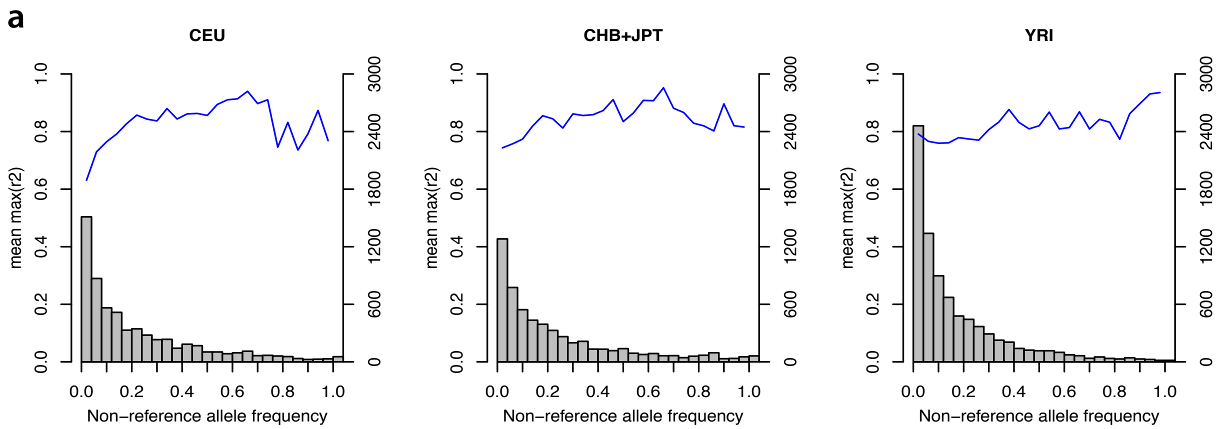
a



Relationship between coverage and sensitivity for discovery of rare, singleton deletion variants in a population by Genome STRiP

(a) The low-coverage genomes sequenced in the 1000 Genomes Project vary in their levels of sequencing coverage, making it possible to evaluate the relationship between sensitivity and coverage of each individual genome. We evaluated this relationship for the most challenging application – the discovery of rare, singleton variants that are private to individual members of a large sequenced cohort. Singleton variants were identified using genotype data from Conrad *et al.* [1]. Coverage is here parameterized as "span" coverage, the extent to which the genome is covered by molecular inserts flanked by aligned sequence reads; span coverage is the coverage statistic most predictive of sensitivity for SV discovery algorithms that make use of paired-end analysis. (For algorithms that rely exclusively on read depth, the number of reads is most predictive of power; for SNP discovery, traditional sequence coverage measures are most predictive of power.) Across the 1000 Genomes samples, span coverage is approximately equal to sequence coverage though it can be greater or less in individual genomes due to variation in the insert sizes of the libraries sequenced. Gray bars indicate the fraction of singleton deletions in each bin. On the right are sensitivities of seven algorithms for deletion discovery in deeply sequenced individual genomes; the sensitivity measurements were made by the 1000 Genomes Project in the NA12878 genome and are reported separately in that study [2]. False discovery rates (FDRs) of each algorithm, as estimated by the 1000 Genomes Project validation analyses, are shown in parentheses.

Supplementary Figure S6



Linkage disequilibrium between genotyped deletions and nearby SNPs

(a) Linkage disequilibrium (r^2) between each deletion and the best Hapmap3 tag SNP is measured in each population and the deletions are binned by alternate allele frequency. Higher frequency deletions are better tagged by Hapmap SNPs (blue line). The number of deletions in each frequency bin is also shown (gray bars, right hand scale).

Supplementary Table 1

Algorithms evaluated by the 1000 Genomes Project for discovery of deletion polymorphism in population-scale, low-coverage sequence data.

Callset origin	Algorithm name	Method type
Broad	Genome STRiP	Integrative
Boston College	Spanner	Paired End
WTSI	N/A	Paired End
Leiden/WTSI	Pindel	Split Read
UCSD	Event-wise testing	Read Depth
WashU	BreakDancer	Paired End
EMBL/Yale	PEMer	Paired End
AECOM	N/A	Read Depth
Yale	N/A	Split Read
Yale	CNVnator	Read Depth

Supplementary Note

In this supplementary note we provide additional details on the methods underlying Genome STRiP, including notes on data requirements and algorithm performance.

Contents

Calculation of most likely deletion length 1

Calculation of incoherence metric 2

Evaluation of candidate variants 2

Evaluation of accuracy (specificity) 3

Gaussian mixture model for read depth genotyping 4

Data requirements 6

Algorithm performance 7

Calculation of most likely deletion length. To evaluate a cluster of read pairs with excessive spacing, we calculated the most likely deletion length (d_{opt}) that would explain the spacing and location of the read pair alignments, based on their insert size distributions.

$$d_{opt} = \arg\max_{d} L_C(d)$$

where

$$L_C(d) = \sum_{p \in pairs(C)} \log(\xi_p(isize(p) - d))$$

Here, C is the cluster of read pairs, ξ_p is the empirical insert size distribution of the library of molecules from which read pair p is drawn, and *isize(p)* is the nominal insert size of read pair p computed from the alignments to the reference genome. d_{opt} maximizes the likelihood that all of the read pairs were drawn from their respective insert size distributions under the model of a deletion of length d. Note that d_{opt} is the most likely value for the difference in length between the reference allele and the deletion allele and so will over-estimate the length of missing reference sequence when non-template sequence is present on the deletion allele. This is illustrated in Fig. 2b, where d_{opt} corresponds to the 316 base pair shift in the read pair spacing that is induced by the deletion and is the most likely value for the difference in length between the reference and alternate alleles.

Calculation of incoherence metric. In Genome STRiP, read pair clustering entails an implicit degree of coherence between the read pairs in each cluster, but because we use a connected components algorithm for clustering, the resulting clusters will be pair-wise coherent but not necessarily coherent as a whole. We evaluate coherence for the cluster as a whole using an "incoherence" metric $F_C(d_{opt})$ where

$$F_C(d) = \frac{1}{N} \sum_{p \in pairs(C)} \log(\int_{x=d}^{\infty} \xi_p(x))$$

This metric evaluates the weight of the right tails of the insert size distributions under the model that the cluster arises from a deletion of length d_{opt}. Coherent clusters should have few pairs where the predicted insert size under the model of a true deletion is in the extreme right tail, whereas for clusters that are a collection of read pairs randomly spaced across the genomic locus, this statistic should approximately follow the null distribution

$$F_{NULL} = \frac{1}{N} \sum_{N} \log(u)$$

where u is uniformly distributed.

Evaluation of candidate variants. Genome STRiP considers many potential structurally variant loci across the genome and generates as primary output a list of numerous candidate SVs along with annotations that must be used for filtering to select a final call set.

The filtering criteria we chose for the 1000 Genomes Project were based on an empirical evaluation of the candidate deletions using known positive control SVs and evidence from SNP genotyping at the candidate deletion loci and the properties of the 1000 Genomes data set. Our specific choices may not be optimal for other data sets, particularly as longer sequence reads increase the contribution of "split", breakpoint-spanning reads to the heterogeneity and coherence tests. The optimal choice of thresholds will also depend upon the number of genomes sequenced and average depth of coverage in a study.

In general, we recommend that the thresholds (for coherence, heterogeneity, substitution) be selected based on calibration to a gold-standard reference data set, aiming for efficient ascertainment of known variants and a realistic number of novel variants. In future data sets, a far-larger number of gold-standard, positive-control SVs will be available and we expect it will be possible to choose selection criteria computationally by fitting to the profiles of known SVs, taking into account the sensitivity/specificity tradeoff most appropriate to the study being performed. In the immediate future, an optimal set of SVs to use for this calibration is the set released in the 1000 Genomes Project pilot [2].

Evaluation of accuracy (specificity). As described in [2], experimental validation of SV discovery data sets was performed by a group of investigators for the 1000 Genomes Project using two approaches: (i) analysis of array-based copy-number data and (ii) PCR of a randomly selected subset of 100 putative deletions from each callset.

The array-based analysis integrated data from three array platforms: Illumina 1M, Affymetrix 6.0, and Nimblegen (genome-wide custom CGH array with 5M probes), which were combined into a single data set (the "super-array"). The array-based data have only partial power to definitively identify true deletions (especially small deletions that cross only one or two probes) but an FDR for a deletion "call set" can be estimated from an overall distribution of p-values. Because array platforms differ in their quantitative response to underlying copy number, the following non-parametric test was used. For each probe, the 168 genomes were ranked in intensity. Each putative deletion consisted of a genomic segment (chr, start, end) and a list of genomes predicted to carry the putative deletion. For each putatively deleted segment, intensity ranks were compared between two sets of genomes: (i) the genomes predicted to carry a deletion of that segment, and (ii) the genomes analyzed but not predicted to carry the deletion. A Wilcoxon rank sum *p*-value (for the directional, one-sided comparison of these two sets of genomes) was calculated. Since the null distribution of p-values (from randomly selected genomic loci) under this test was symmetrical and approximately uniform from 0 to 1 (as predicted by theory), an estimate of the false discovery rate is twice the area of the right half of the distribution (two times the fraction of p-values that are greater than 0.5).

For PCR validation, a random set of 100 putative deletions was selected from each callset, flanking primers were designed, and a relevant genome tested for whether or not the size of the resulting PCR product matched expectation given the predicted deletion. For almost all call sets that could be evaluated by both approaches (array-based and PCR), FDRs estimated using the two approaches agreed to within the statistical sampling noise expected from the PCR sampling strategy [2]. In the 1000 Genomes analysis, the array- and PCR-based FDRs were collapsed into a single, overall FDR estimate by applying the PCR-based FDR to all of the deletion calls for which the array data were uninformative. The resulting FDRs are reported in detail in Supplementary Table 8 of [2].

Gaussian mixture model for read depth genotyping. To incorporate read depth information into the genotypes produced by Genome STRiP, we fit the counts of observed DNA fragments to a constrained Gaussian mixture model to estimate the copy number at each locus.

The mixture model can be restricted to exactly three copy-number classes (copy number 0, 1 and 2, based on the model of a pure deletion), as was done for this study, or it can accommodate additional copy number classes. The model allows for differential sequencing depth of each sample by incorporating an estimate, e_i, of the number of reads expected at the locus for sample i corresponding to copy number one. We estimated the e_i from the total genome-wide read depth and the effective length of the deletion locus (the number of confidently alignable bases). The expected means μ_{ij} and variances ν_{ij} for sample i in copy-number class j are modeled as

$$\mu_{i,j} = m_1 \cdot a_j \cdot e_i$$
$$\nu_{i,j} = m_2 \cdot b_j \cdot e_i$$

where m_1 and m_2 are parameters to be fitted and the constant vectors $a = <0,1,2,\ldots>$ and $b = <k,1,2,\ldots>$ represent an assumed linear relationship between read depth and copy number. The value k is a constant that allows for non-zero variance in the copy number zero class, which makes the model more tolerant of reads mapped to homozygous deleted regions that can arise from mis-alignment, sequencing error or incorrectly estimated breakpoint boundaries. Using concordance with array-based genotypes[11], we empirically estimated $k=0.2$ to provide optimal results.

The model follows the typical form of a Gaussian mixture, with the addition of sample-specific means and variances, where the expectation that sample i is in copy number class j is

$$z_{ij} = \frac{w_j \cdot \mathcal{N}(x_i, \mu_{ij}, \nu_{ij})}{\sum_j w_j \cdot \mathcal{N}(x_i, \mu_{ij}, \nu_{ij})}$$

Here x_i is the number of observed reads for sample i, $\mathcal{N}(x_i, \mu_{ij}, \nu_{ij})$ is the normal distribution evaluated at x_i with mean μ_{ij} and variance ν_{ij}, and w_i represents the weight assigned to cluster j, corresponding to the copy-number frequencies.

We estimated the model parameters m_1, m_2 and the w_j using an expectation-maximization (EM) algorithm. We used three different sets of initial conditions, initializing both scaling factors m_1 and m_2 to 0.9, 1.0 and 1.1, and initializing the values of w_j to equal frequencies $1/C$ where C is the total number of copy-number classes in the model. Using the derivation of a standard EM for a Gaussian mixture model, but substituting our definitions of μ_{ij} and ν_{ij}, we can derive expressions for computing updated values of the model parameters in the maximization step (here N is the total number of samples):

$$w_j = \frac{1}{N}\sum_i z_{ij}$$

$$m_1 = \frac{\sum_{i,j} \frac{z_{ij} \cdot e_i \cdot a_j}{b_j}}{\sum_{i,j} \frac{z_{ij} \cdot e_i \cdot a_j^2}{b_j}}$$

$$m_2 = \frac{\sum_{i,j} \frac{z_{ij}(x_i - \mu_{ij})^2}{e_i \cdot b_j}}{\sum_{i,j} z_{ij}}$$

The expectation and maximization steps are iterated alternately until convergence.

The estimated model parameters are combined with the observed read depths to calculate the relative likelihood of each copy number class for genome *i*. The estimated model parameters also provide information about the quality of the genotyping. The w_i provide estimates of the allele frequency of the locus (for unrelated samples, these can be tested for Hardy-Weinberg equilibrium) and the m_1 parameter provides an indication of whether the normalization was effective (m_1 should be close to 1).

Data requirements. For variant discovery, Genome STRiP uses as input aligned, paired-end sequence data (or a mixture of paired-end and single-end sequencing data) in BAM format (http://samtools.sourceforge.net/SAM1.pdf). For genotyping, Genome STRiP requires as input aligned, paired-end and/or single-end sequencing data in BAM format. The structural variants to be genotyped are supplied in VCF format (http://vcftools.sourceforge.net/specs.html). Because the VCF format can include a molecular description of the alternative alleles at each locus, a separate library of breakpoint sequences, as was used in the pilot phase of the 1000 Genomes Project, is no longer necessary.

Algorithm performance. Genome STRiP is designed to allow parallel computation on large compute clusters with modest memory requirements (4GB/node) and we have successfully run the algorithm using LSF (Load Sharing Facility by Platform Computing, www.platform.com). Although we have not done formal benchmarking on the runtime performance of Genome STRiP, we estimate that a full analysis of the 1000 Genomes pilot data (168 individuals at roughly 4x sequencing coverage), including data preprocessing, BWA realignment, deletion discovery and genotyping of approximately 22,000 deletions, required roughly 600 CPU-hours in total, spread across the available CPUs on our compute cluster.

References

1. Conrad, D.F., et al., *Origins and functional impact of copy number variation in the human genome.* Nature, 2010. 464(7289): p. 704-12.
2. Durbin, R.M., et al., *A map of human genome variation from population-scale sequencing.* Nature, 2010. 467(7319): p. 1061-73.
3. Mills, R.E., *Mapping copy number variation by population scale sequencing.* Nature, *in press*.
4. Mills, R.E., et al., *An initial map of insertion and deletion (INDEL) variation in the human genome.* Genome Res, 2006. 16(9): p. 1182-90.