LETTERS

nature genetics

Systematic assessment of copy number variant detection via genome-wide SNP genotyping

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

SNP genotyping has emerged as a technology to incorporate copy number variants (CNVs) into genetic analyses of human traits. However, the extent to which SNP platforms accurately capture CNVs remains unclear. Using independent, sequence-based CNV maps, we find that commonly used SNP platforms have limited or no probe coverage for a large fraction of CNVs. Despite this, in 9 samples we inferred 368 CNVs using Illumina SNP genotyping data and experimentally validated over two-thirds of these. We also developed a method (SNP-Conditional Mixture Modeling, SCIMM) to robustly genotype deletions using as few as two SNP probes. We find that HapMap SNPs are strongly correlated with 82% of common deletions, but the newest SNP platforms effectively tag about 50%. We conclude that currently available genome-wide SNP assays can capture CNVs accurately, but improvements in array designs, particularly in duplicated sequences, are necessary to facilitate more comprehensive analyses of genomic variation.

Copy number variants (CNVs) occur commonly in the human genome[1–4], often affect genes, contribute to genomic evolution and genetic diversity (reviewed in ref. 5) and influence a number of human traits[6–10]. Considering these observations, it is likely that future genetic studies would benefit from analyzing CNVs in addition to SNPs[11]. However, relative to SNPs, CNVs are a priori likely to have larger phenotypic effects[5], and the mutation rate generating them in some regions of the genome is substantially higher[12]. Consequently, an important technological goal is the development of a platform capable of discovering rare CNVs in addition to genotyping common variants, which are related but distinct challenges. One promising solution is to leverage commercially available genome-wide SNP platforms, which have been and will continue to be widely applied in association studies[13]. These assays can indirectly interrogate CNVs via linkage disequilibrium (LD)[3,14–16] and directly quantify copy number for some variants[17–20].

Because of an absence of high-resolution, independently generated maps of variation, the extent to which commercial SNP platforms accurately capture CNVs remains largely unknown. To address this, we leveraged genome-wide fosmid end-sequence-pair (ESP) maps recently developed for nine humans[2,4]. We found that even newer platforms miss a large fraction of the CNVs present in any given individual. However, using a hidden Markov model (HMM) approach, we show that many CNVs can be discovered within a given sample and systematically validated. We also develop a novel algorithm, known as SNP-Conditional Mixture Modeling (SCIMM), to robustly genotype common variants directly in large collections of individuals and evaluate their correlations with neighboring SNPs. Our results have implications for retrospective analysis of existing genome-wide SNP data as well as for future assay designs.

We first assessed the probe coverage for commonly used SNP arrays within variants identified systematically in nine human genomes by fosmid ESP mapping and validated by orthogonal approaches[4]. Using breakpoints inferred from high-density oligonucleotide array–comparative genomic hybridization (CGH) experiments for 500 deletions larger than 1 kb, we found that older genome-wide platforms (Illumina HumanHap 300 and Affymetrix 500K) lack probes within ~75% of deletions, and fewer than 20% harbor multiple probes (**Fig. 1**). Newer arrays (Illumina Human 1M and Affymetrix 6.0) show improved coverage, but ~20% of deletions harbor zero probes and most span fewer than 5. We obtained similar results when we considered deletions annotated by complete fosmid sequencing and alignment to the reference assembly, with ~30% missed even on newer platforms (**Supplementary Table 1** online).

To discover CNVs within a given sample using Illumina Infinium[17] data, we applied a simple HMM-based approach using HMMSeg[21] (see Methods). The procedure simultaneously analyzes both the normalized total intensity ('LogR ratio') and allelic intensity ratios ('B-allele frequency')[17] to detect regions of homozygous deletion, hemizygous deletion, or amplification. In the nine samples for which a fosmid library was available[4] we identified a total of 368 events greater than 1 kb in length (258 deletions, 110 amplifications; **Supplementary Table 2** online). We found that 116 of 258 (~45%) predicted deletions overlap a deletion discovered by fosmid ESP mapping, with a strong correlation in estimated sizes ($R^2 = 0.79$; **Fig. 2**). We observed that a substantial, albeit smaller, fraction of the inferred amplifications map to previously defined insertion events (15 of 110 amplifications). The vast majority of the nonvalidated amplifications are large (81% are >40 kb) and heavily enriched for

[1]Department of Genome Sciences, University of Washington, Seattle, Washington 98195, USA. [2]Howard Hughes Medical Institute. [3]These authors contributed equally to this work. Correspondence should be addressed to G.M.C. (coopergm@u.washington.edu).

Received 29 January; accepted 18 August; published online 7 September 2008; doi:10.1038/ng.236

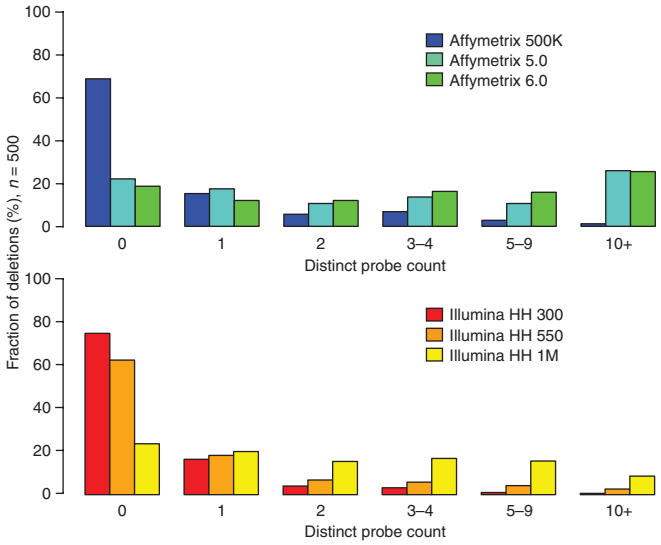

Figure 1 Probe-coverage histogram for 500 nonredundant deletion events greater than 1 kb in size identified in nine human samples by fosmid ESP placements and refined using oligonucleotide array-CGH experiments[4]. We analyzed three array platforms available from Affymetrix (top) and Illumina (bottom). 'Distinct' probes correspond to each distinct location in the genome (hg17) physically represented on the array and internal to the annotated deletion breakpoints (physically redundant probes for a given location are not counted).

segmental duplications in the reference assembly (72% of events, 70% of nucleotides), reflecting the known enrichment for CNVs in duplication-rich regions of the reference assembly[1,3,22]. ESP mapping, however, has reduced sensitivity to both large insertions and variants within duplication-rich sequence[2]; thus, the lower rate of validation for amplification events is expected.

We subsequently leveraged the entire fosmid ESP maps available for these samples (~900,000 clones per individual; see URLs section in Methods) to determine if the nonvalidated predictions were false positives or previously missed variants. Many of the deletions inferred here were below the size thresholds used previously[4], and by relaxing thresholds we found support for 18 (~7%) additional deletions (**Supplementary Methods** and **Supplementary Table 3** online). We also sought to overcome the inability to validate large amplifications through fosmid ESP placements using two approaches. First, we hypothesized that if a sequence unique in the reference assembly is tandemly duplicated in a given sample, then a clone that spans the duplication breakpoint will align to the reference genome such that the reads orient away from the center of the clone (**Fig. 3**). We found that an additional 18 (16%) amplifications inferred using SNP data overlap a cluster of such clones. Second, we considered variants annotated within eight of these nine samples by a combination of SNP array and BAC-CGH analysis[3], the latter of which has better power to detect large variants in duplication-rich regions. We found that an additional 25 (23%) inferred amplification events overlap a previously defined 'gain' within the same sample. Combining these and other analyses of fosmid ESP placements (**Supplementary Tables 4** and **5** and **Supplementary Fig. 1** online), we conclude that at least 67% and 64% of the inferred deletion and amplification events, respectively, correspond to experimentally validated variants. The actual true-positive rate is likely higher than this (**Supplementary Table 5**).

Genome-wide CNV discovery must be conducted per-sample to detect rare events and must also account for the low prior probability that any given probe is inside a CNV. Therefore, high specificity was our primary goal. However, we also considered the extent to which known variants in these samples had been missed. Restricting our analysis to defined 'detectable' (spanning enough probes) deletions, we found that our sensitivity was ~47% (7/12 sequence-defined deletions, 30/66 CGH-defined deletions). We note that many of the deletions that were missed correspond to duplication-rich sites: ~67% of the nucleotides within the false negatives are within a segmental duplication, representing a 13-fold enrichment over the genomic average[23]. Thus, most of the missing deletions correspond to sequences present in multiple copies in the reference assembly.

In contrast to discovery, targeted genotyping can leverage the knowledge that a CNV exists at a particular location and, for common variants, borrow information across samples, reducing the number of probes required for analysis. We therefore implemented a strategy, denoted SCIMM, for genotyping polymorphic insertion/deletion variants spanning as few as 2 probes. SCIMM uses mixture-likelihood–based clustering[24], motivated by the observation that hemizygous or homozygously deleted samples often manifest as distinct clusters in the fluorescence intensity data for SNP probes inside common deletions (**Fig. 4**; see Methods). A second algorithm, SCIMM-Search, identifies copy-number informative probes within known deletions (**Supplementary Methods**).

To validate this approach, we analyzed data generated by the Illumina Human 1M assay for 126 samples (125 HapMap samples plus NA15510; see Methods), including 28 parent-child trios. We compared insertion/deletion genotypes produced by our algorithm for 18 common, autosomal deletions that have been independently genotyped using quantitative PCR and GoldenGate fluorescence data[4,25] (**Supplementary Tables 6** and **7** online). SCIMM-Search identified informative probe sets for 13 of these sites, and generated genotypes with correlation (r^2) to the reference genotypes exceeding

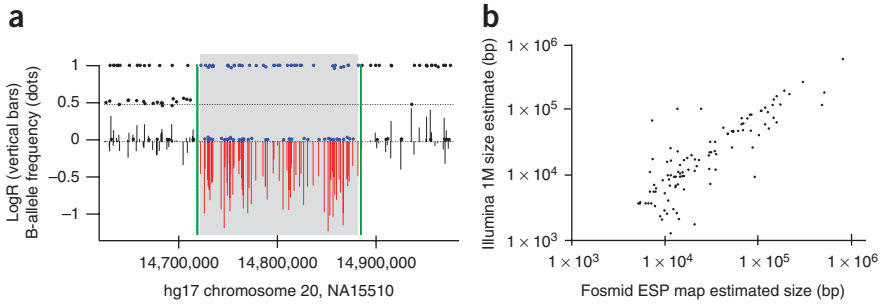

Figure 2 Deletion predictions validated by fosmid ESP placement data. (**a**) Example of a deletion event inferred using Illumina Human 1M data. Intensity data for all probes in the indicated genomic interval (*x* axis) for sample NA15510 (aka 'G248') are plotted. 'LogR ratio' and 'B-allele frequency'[17] are plotted as vertical bars and filled dots, respectively. The gray box indicates the deletion span inferred by segmentation of the SNP data; probes internal to this box are colored red (LogR ratio) or blue (B-allele frequency). Green vertical bars indicate the deletion borders defined by complete fosmid re-sequencing. (**b**) Correlation in size estimates between deletions inferred from SNP genotyping data (*y* axis) that overlap deletions annotated by fosmid ESP mapping (*x* axis). Both axes are log-scaled.

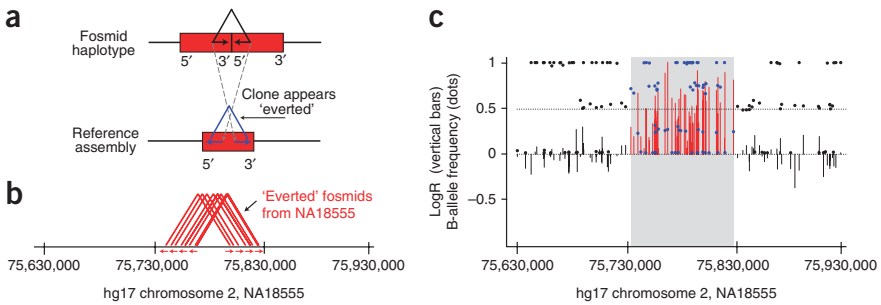

Figure 3 Amplification events validated by clusters of 'everted' fosmid ESP placements. (**a**) Consider a block of sequence (red bar) that is unique in the reference assembly (bottom portion) but tandemly duplicated in the haplotype of interest (top portion). In principle, for clones that span the breakpoint of this duplication, the end sequences will be everted when aligned to the reference assembly such that they are oriented away from the center of the clone. Note that this should occur at all such duplication breakpoints, regardless of duplication size. (**b**) We identified 233 sites in the nine fosmid libraries harboring multiple overlapping everted clones, one of which is shown here. This site is supported by eight distinct clones (red triangles), each of which has reads oriented outwards as indicated by the underlying red arrows. (**c**) Illumina Human 1M data for the same region in the same sample (*x* axes are identical) is shown, with logR and B-allele frequency plotted as vertical bars and dots, respectively (similar to **Fig. 2a**). The gray box corresponds to the duplication interval inferred by segmentation of the SNP data.

80% at all sites and >97% genotype concordance at 12 sites (**Supplementary Table 6**).

We subsequently applied SCIMM-Search to 252 non-overlapping, independently defined autosomal deletions spanning two or more probes on the Illumina Human 1M array, identifying informative probe sets for 136 of these sites (**Supplementary Fig. 2** and **Supplementary Table 8** online). Of the 130 sites passing subsequent manual review, 126 are polymorphic (allele frequency >1%), with only six mendelian inconsistencies across the 3,640 trio offspring genotypes (**Supplementary Tables 9** and **10** online). We also applied SCIMM to data produced for 120 HapMap samples using the Illumina HumanHap 550 assay. We found that deletions spanned by two or more HumanHap 550 probes yield highly concordant genotypes (99.8% identical). This demonstrates high technical reproducibility and the applicability of reduced probe sets to lower-density data. However, single-probe genotypes were more prone to discordancy (**Supplementary Fig. 3** online), indicating that multiple probes are required for accurate genotypes.

Previous studies have evaluated LD between SNPs and insertion/deletion polymorphisms, suggesting that associations between CNVs and phenotypes may be detected through correlation with SNP genotypes[3,15,16]. We searched for tag SNPs for each of the successfully genotyped deletion polymorphisms (126 total) using SNP genotype data from Phase II HapMap data[26] and from four genome-wide SNP datasets (**Table 1**). We found that 82% (69/84) of the common deletions (worldwide frequency >5%) were strongly correlated to a HapMap SNP (worldwide $r^2 > 0.8$); in contrast, each high-density genome-wide SNP platform effectively tagged only about half (48–54%) of the common deletions (**Table 1** and **Supplementary Table 8**) by the same criterion.

Although it has previously been shown that SNP-array data can be used to infer the presence of intermediate size CNVs[17–20], the reliability of the resulting annotations has never been systematically validated. Exploiting data from the Illumina Human 1M BeadArray, we accurately predicted the identity and size of 368 intermediate-size CNVs in nine samples, at least two-thirds of which were validated with independent experimental data. Our validation expands previous uses of fosmid ESP mapping information and includes a novel technique to confirm the presence of large duplication events. This technique circumvents a previously recognized limitation of the approach (inability to identify large insertions) and may prove useful in future applications of high-throughput clone ESP data[27,28].

We also developed a novel genotyping algorithm that accurately infers genotypes for polymorphic deletions using as few as two probes. We found that 82% of the genotyped sites can be tagged by SNPs near the CNV; however, the best available platforms only tag ~50% (**Table 1**). We note that the set of deletion events that we successfully genotyped is not a random sample. In particular, segmental duplications in the reference assembly are under-represented on all genome-wide SNP platforms (not shown), and even when probes are present, cross-hybridization of paralogous sequences can confound deletion genotyping (**Supplementary Methods**). However, the mutation rate for events in and around clusters of segmental duplications is substantially higher than the background mutation rate[8,12]. Thus, the regions that might be prone to recurrent deletion generation, and that would in turn not be in strong LD with neighboring variants, are probably enriched in the set of events for which we did not obtain genotypes. Our estimate that 18% of common deletion variants are not strongly correlated with any known SNP is thus likely to be a lower-bound estimate, and underscores the need for independent experimental information to evaluate CNV detection[11,27].

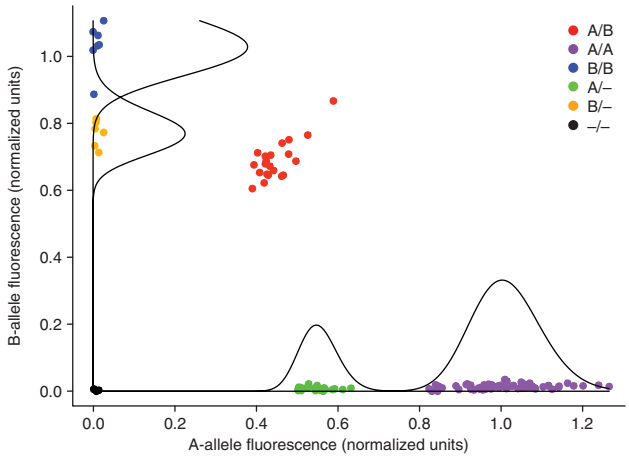

Figure 4 Example of fluorescence intensity measurements for each of 126 samples for a single SNP probe (rs10076425). These data are used by SCIMM to automatically determine insertion/deletion genotypes (see Methods). The genotype for each sample is denoted by color as indicated in the legend. Mixture component distributions are represented by superimposed curves on both the *x* and *y* axes. In this case, insertion/deletion status is computed by analyzing this probe in conjunction with four additional probes within a deleted region at chr. 4:10070382–10076653 (**Supplementary Table 8**).

Table 1 Pairwise correlation estimates between SNP and deletion genotypes

SNP Dataset	Fraction of sites with at least one tag SNP (%), $r^2>0.8$	Fraction of sites with at least one tag SNP (%), $r^2>0.7$	Mean (max r^2), all sites
HapMap 2.0	82	88	0.88
Illumina Human 1M	54	70	0.77
Affymetrix 6.0	51	65	0.73
Illumina HumanHap 650Y	48	64	0.74
Illumina HumanHap 550	48	61	0.71

Direct interrogation of copy number variation by genome-wide SNP platforms is limited by probe coverage. Even on the newest genome-wide SNP platforms, at least 20% of all intermediate-size deletion events span zero, and most fewer than five, probes (**Fig. 1**). Older array designs have particularly poor coverage and retrospective mining of these data is likely to be of limited utility. This finding is in contrast to the higher coverage estimates one would obtain with lower-resolution, and generally inflated, CNV annotations, such as those derived from BAC-CGH experiments[4,29] (**Supplementary Fig. 4** online). We note that differences in assay chemistry, probe specificity and physical redundancy strongly influence dynamic range; thus, probe count alone does not provide sufficient information for comparison of different platforms.

We show that SNP arrays can be used to infer the presence of many individually rare CNVs with reasonable specificity given a considerable probe count, and can furthermore be used to robustly genotype common deletions using as few as two probes. However, when considering balanced events (for example, inversions), novel insertion sequences not represented in the reference assembly[4], and the bias against segmental duplications in array designs contrasted with the enrichment for CNVs both within and flanking duplicated sequences[1,3,22], we conclude that a large fraction of genomic variation cannot be captured by existing genome-wide SNP platforms. Significant improvements to array designs, perhaps in the form of a targeted CNV genotyping platform, may ultimately be necessary. In any case, it will be important to continue to benchmark such efforts against high-resolution, ultimately sequence-based maps of variation to accurately assess both successes and failures. These analyses should lay the framework for more comprehensive assessments of human genomic variation.

METHODS

Genome-wide SNP genotyping. We obtained SNP genotyping data generated by the Illumina Human 1M and HumanHap 550K platforms directly from Illumina (courtesy of D. Peiffer, Illumina). The 1M data include within-sample normalized fluorescence ('x' and 'y'), between-sample normalized fluorescence ('Log R ratio' and 'B-allele frequency'), and SNP calls for 125 HapMap samples (**Supplementary Methods**), including eight samples for which fosmid libraries have been generated[4]. We supplemented the genotyping data with one additional sample, NA15510 (also known as 'G248'), which was previously analyzed by fosmid ESP analysis[2]. SNP genotyping was done in accordance with manufacturer's protocols[17]. Note that only 120 HapMap samples (a subset of the 126 described above) were available on the Illumina HumanHap 550K array.

CNV discovery using Illumina Human 1M genotyping data. Large CNV discovery was accomplished by using HMMSeg[21], considering both the 'LogR ratio' and 'B-allele frequency' data for each sample simultaneously, based essentially on a previously established approach[17]. We used a four-state model, one each for null (homozygous deletion), hemizygous deletion, diploid and amplification. Initial segmentation results were merged and filtered, requiring all variants to be larger than 1 kb in length and to span at least 10 probes for amplifications or hemizygous deletions, or 3 probes for homozygous deletions. Additional details on data normalization, model specifications and implementation can be found in **Supplementary Methods**.

CNV validation using whole-genome fosmid ESP placement analysis. We validated our CNV predictions by comparing their locations and sizes with the locations and sizes of variants that had been previously annotated by analysis of fosmid ESP placements for the same nine individuals; all these data are available in the supplementary information of ref. 4 and at the Human Genome Structural Variation Project web site (see URLs section below). We considered any amount of overlap between the CNV maps as validation, given the restriction that the event is in the same sample and in the same direction (that is, only ESP deletions are used to validate predicted deletions). We performed additional validation by considering all fosmid ESP placement information, borrowing information across samples, and leveraging information on clone placements that were previously excluded as a result of size, alignment score or other quality-control thresholds (**Supplementary Methods**).

Probe coverage analysis. Genomic locations of SNP probes were obtained from the Affymetrix and Illumina web sites for the SNP genotyping products provided by the respective companies. We mapped coordinates as appropriate from hg18 to hg17 using the 'Liftover' tool at the UCSC Genome Bioinformatics web site. Locations of CNVs identified through analysis of fosmid ESP mapping were obtained directly from the supplementary data provided in ref. 4. An overview of the datasets used here can be found in **Supplementary Methods**.

Insertion-deletion genotyping. SCIMM is a clustering algorithm which, given a set of probes, produces a classification of each sample as 'null', 'haploid' or 'diploid'. Two rounds of mixture-likelihood–based clustering implemented by the expectation-maximization algorithm[24] are used; the first operates on per-sample summary intensity values to identify null samples, and the second operates directly on two-channel fluorescence data to classify the remaining samples as either 'haploid' or 'diploid'. SNP genotypes are used for direct inference of copy number (heterozygosity is used as evidence of diploidy) and for model fitting. During the second round of clustering, a single-component (copy-number-invariant) model is also fit to the data to produce a score for the probe set using the Bayesian information criterion (BIC)[30].

SCIMM-Search is an iterative search algorithm which, given the coordinates of a region spanning a deletion and the identity of a sample known to carry the variant, determines the set of probes used to genotype the deletion variant carried by the reference sample. SCIMM-Search uses the BIC to evaluate alternate probe sets. It is not assumed that all probes within the annotated region are informative for copy number or that informative probes are contiguous. SCIMM-Search allows specification of constraints on genotype consistency between probes and cluster separation for each probe in the probe set. A more comprehensive description along with detailed model specifications and thresholds for both SCIMM and SCIMM-Search are available in **Supplementary Methods**.

Taggability. For each polymorphic insertion/deletion site, we extracted all Phase II HapMap SNP genotypes within 200 kb of the deletion interval and calculated r^2 between each SNP and the SCIMM-generated insertion/deletion samples for 90 unrelated HapMap individuals (**Supplementary Table 8**). We combined data across populations to obtain a single estimate of correlation (within-population correlation estimates were similar; data not shown) and ignored sites with calls for fewer than 75% of the unrelated samples. We repeated this process for the Affymetrix 6.0, Illumina Human 1M, Illumina 650Y and Illumina HumanHap 550 assays, using SNP genotypes provided by the manufacturers of each assay.

URLs. Human Genome Structural Variation Project, http://hgsv.washington.edu.

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

We thank D. Peiffer and colleagues at Illumina for sharing Human 1M and HumanHap 550K genotyping data. We apologize to all colleagues whose work we could not cite because of space constraints. G.M.C. is supported by a Merck, Jane Coffin Childs Memorial Fund Postdoctoral Fellowship. T.Z. acknowledges support from the National Human Genome Research Institute (NHGRI) Interdisciplinary Training in Genomic Sciences grant T32 HG00035. J.M.K. is supported by a National Science Foundation graduate fellowship. This work was supported by the National Heart, Lung, and Blood Institute Programs for Genomic Applications grant HL066682 to D.A.N. and NHGRI grant HG004120 to E.E.E. E.E.E. is an investigator of the Howard Hughes Medical Institute.

Published online at http://www.nature.com/naturegenetics/
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/

1. Sebat, J. *et al.* Large-scale copy number polymorphism in the human genome. *Science* **305**, 525–528 (2004).
2. Tuzun, E. *et al.* Fine-scale structural variation of the human genome. *Nat. Genet.* **37**, 727–732 (2005).
3. Redon, R. *et al.* Global variation in copy number in the human genome. *Nature* **444**, 444–454 (2006).
4. Kidd, J.M. *et al.* Mapping and sequencing of structural variation from eight human genomes. *Nature* **453**, 56–64 (2008).
5. Cooper, G.M., Nickerson, D.A. & Eichler, E.E. Mutational and selective effects on copy-number variants in the human genome. *Nat. Genet.* **39**, S22–S29 (2007).
6. Singleton, A.B. *et al.* alpha-Synuclein locus triplication causes Parkinson's disease. *Science* **302**, 841 (2003).
7. Gonzalez, E. *et al.* The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility. *Science* **307**, 1434–1440 (2005).
8. Sharp, A.J. *et al.* Discovery of previously unidentified genomic disorders from the duplication architecture of the human genome. *Nat. Genet.* **38**, 1038–1042 (2006).
9. Perry, G.H. *et al.* Diet and the evolution of human amylase gene copy number variation. *Nat. Genet.* **39**, 1256–1260 (2007).
10. Walsh, T. *et al.* Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. *Science* **320**, 539–543 (2008).
11. Estivill, X. & Armengol, L. Copy number variants and common disorders: filling the gaps and exploring complexity in genome-wide association studies. *PLoS Genet.* **3**, 1787–1799 (2007).
12. Shaffer, L.G. & Lupski, J.R. Molecular mechanisms for constitutional chromosomal rearrangements in humans. *Annu. Rev. Genet.* **34**, 297–329 (2000).
13. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–678 (2007).
14. Conrad, D.F., Andrews, T.D., Carter, N.P., Hurles, M.E. & Pritchard, J.K. A high-resolution survey of deletion polymorphism in the human genome. *Nat. Genet.* **38**, 75–81 (2006).
15. Locke, D.P. *et al.* Linkage disequilibrium and heritability of copy-number polymorphisms within duplicated regions of the human genome. *Am. J. Hum. Genet.* **79**, 275–290 (2006).
16. McCarroll, S.A. *et al.* Common deletion polymorphisms in the human genome. *Nat. Genet.* **38**, 86–92 (2006).
17. Peiffer, D.A. *et al.* High-resolution genomic profiling of chromosomal aberrations using Infinium whole-genome genotyping. *Genome Res.* **16**, 1136–1148 (2006).
18. Komura, D. *et al.* Genome-wide detection of human copy number variations using high-density DNA oligonucleotide arrays. *Genome Res.* **16**, 1575–1584 (2006).
19. Colella, S. *et al.* QuantiSNP: an objective Bayes hidden-Markov model to detect and accurately map copy number variation using SNP genotyping data. *Nucleic Acids Res.* **35**, 2013–2025 (2007).
20. Wang, K. *et al.* PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res.* **17**, 1665–1674 (2007).
21. Day, N., Hemmaplardh, A., Thurman, R.E., Stamatoyannopoulos, J.A. & Noble, W.S. Unsupervised segmentation of continuous genomic data. *Bioinformatics* **23**, 1424–1426 (2007).
22. Sharp, A.J. *et al.* Segmental duplications and copy-number variation in the human genome. *Am. J. Hum. Genet.* **77**, 78–88 (2005).
23. She, X. *et al.* Shotgun sequence assembly and recent segmental duplications within the human genome. *Nature* **431**, 927–930 (2004).
24. Dempster, A.P., Laird, N.M. & Rubin, D.B. Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B. Methodological* **39**, 1–38 (1977).
25. Newman, T.L. *et al.* High-throughput genotyping of intermediate-size structural variation. *Hum. Mol. Genet.* **15**, 1159–1167 (2006).
26. International HapMap Consortium. A haplotype map of the human genome. *Nature* **437**, 1299–1320 (2005).
27. Eichler, E.E. *et al.* Completing the map of human genetic variation. *Nature* **447**, 161–165 (2007).
28. Korbel, J.O. *et al.* Paired-end mapping reveals extensive structural variation in the human genome. *Science* **318**, 420–426 (2007).
29. de Smith, A.J. *et al.* Array CGH analysis of copy number variation identifies 1284 new genes variant in healthy white males: implications for association studies of complex diseases. *Hum. Mol. Genet.* **16**, 2783–2794 (2007).
30. Schwarz, G. Estimating the dimension of a model. *Annals of Statistics* **6**, 461–464 (1978).