BMC Genomics

BioMed Central

Research article

Open Access

Toward accurate high-throughput SNP genotyping in the presence of inherited copy number variation

Laura E MacConaill[1,2], Micheala A Aldred[3,4], Xincheng Lu[4] and Thomas LaFramboise*[4]

Address: [1]Dana-Farber Cancer Institute, 44 Binney Street, Boston, Massachusetts 02116, USA, [2]The Broad Institute of Harvard and MIT, 7 Cambridge Center, Cambridge, Massachusetts 02141, USA, [3]Genomic Medicine Institute, Lerner Research Institute, Cleveland Clinic, 9500 Euclid Avenue, Cleveland Ohio 44195, USA and [4]Department of Genetics, Case Western Reserve University, 10900 Euclid Avenue, Cleveland Ohio 44106, USA

Email: Laura E MacConaill - laura_macconaill@dfci.harvard.edu; Micheala A Aldred - aldredm@ccf.org; Xincheng Lu - xxl16@case.edu; Thomas LaFramboise* - thomas.laframboise@case.edu

* Corresponding author

Published: 3 July 2007

BMC Genomics 2007, **8**:211 doi:10.1186/1471-2164-8-211

Received: 9 March 2007
Accepted: 3 July 2007

This article is available from: http://www.biomedcentral.com/1471-2164/8/211

Abstract

Background: The recent discovery of widespread copy number variation in humans has forced a shift away from the assumption of two copies per locus per cell throughout the autosomal genome. In particular, a SNP site can no longer always be accurately assigned one of three genotypes in an individual. In the presence of copy number variability, the individual may theoretically harbor any number of copies of each of the two SNP alleles.

Results: To address this issue, we have developed a method to infer a "generalized genotype" from raw SNP microarray data. Here we apply our approach to data from 48 individuals and uncover thousands of aberrant SNPs, most in regions that were previously unreported as copy number variants. We show that our allele-specific copy numbers follow Mendelian inheritance patterns that would be obscured in the absence of SNP allele information. The interplay between duplication and point mutation in our data shed light on the relative frequencies of these events in human history, showing that at least some of the duplication events were recurrent.

Conclusion: This new multi-allelic view of SNPs has a complicated role in disease association studies, and further work will be necessary in order to accurately assess its importance. Software to perform generalized genotyping from SNP array data is freely available online [1].

Background

A copy number variant (CNV) is defined as a chromosomal segment, at least 1 kb in length, whose (germline) copy number varies across individuals in the human population [2]. As the importance of these duplications and deletions in the study of a variety of diseases [3-6] is being realized, cataloging them and assessing their frequencies has become an important goal. Toward this end, two recent studies [7,8] have exploited erroneous SNP genotype calls, inferring germline deletions at clusters of calls that violate Mendelian inheritance or other conditions. The violations occur, however, as a result of the (diallelic) assumption of three possible genotypes (*e.g.* GG, GT, or TT) at each SNP site. If this assumption of two copies at each locus were relaxed, one could consider a generalized genotype whereby the SNP is multi-allelic when consider-

ing both base residue and copy number. An individual could carry, for example, a GGT (duplication), G – (hemizygous deletion), or – (homozygous deletion) genotype at a SNP locus. As most recent estimates put the proportion of the genome harboring CNVs at at least 12% [9], allowing for more general genotypes is crucial for the accuracy of SNP typing in disease studies. Such direct and accurate typing would, of course, reveal CNVs automatically.

The GeneChip Human Mapping Array Set [10] is a popular platform for high throughput SNP genotyping. We use data from the version of the platform – herein referred to as the SNP array – that interrogates over 500,000 SNP sites. Since 85% of the genome is within 10 kb of at least one SNP on the array [10], many of the duplications and deletions of the size that have been reported thus far should contain several of the SNP sites represented on the array. Indeed, 58,353 of its 490,032 autosomal SNPs are contained in at least one of the CNVs that have been reported in the literature to date and catalogued in the Database of Genomic Variants [11]. In an earlier study [12] we utilized the array data to detect, in an allele-specific manner, somatic copy number changes in cancer samples, also demonstrating an extremely high genotyping accuracy (> 99.7%) in the diallelic setting. We therefore endeavored in the present study to adapt this approach to SNP array data from "phenotypically normal" individuals in an effort to provide a generalized genotype, allowing for germline CNVs as described above. The approach is somewhat akin to array CGH [13] methods to detect CNVs, but with at least two advantages. First, it is difficult in array CGH analysis to determine whether an apparent deviation from copy number two is the result of a CNV in the test sample or in the reference sample. In our approach, we exploit a large reference panel of individuals, ensuring that the reference signal is essentially two-copy, except potentially in regions with very common CNVs. Second, while the array CGH platforms lack allele-specific information, the SNP array is comprised of oligonucleotide probes that can distinguish between each of the two alleles of each SNP. Our method models these probes' intensities as a function of allele-specific copy number, which directly determines the generalized genotype. The copy numbers are inferred from the SNP array data by applying statistical model-fitting procedures.

Results and discussion

Detected aberrations

We analyzed SNP array data from a collection of 48 individuals of various ethnic backgrounds. For a reference panel, we used 16 unrelated individuals of African, African-American, Asian, European, and Hispanic ethnicities (see Methods). An ethnically diverse panel of reference samples minimizes the likelihood of a recurrent CNV skewing the "copy number two" reference signal. Applying our algorithm to all automsomal SNPs on the array, we found 21,568 SNP loci that demonstrated aberrant genotypes (Additional data file 1). Of these, 17,390 were detected as duplications (total copy number > 2), 5,051 as hemizygous deletions (total copy number 1), and 214 as homozygous deletions (total copy number 0). There were 881 sites that were detected as both duplications and deletions. The 21,568 SNPs can be grouped into 5,622 regions of consecutive duplicated SNPs and 1,130 regions of consecutive deleted SNPs Regarding recurrence, 3,721 (17.3%) SNPs were aberrant in more than one individual (Figure 1a), with one SNP (rs1842908) showing a non-diploid genotype in 24 of the 48 individuals.

Experimental validation

In order to validate our discoveries using independent experimental means, we performed quantitative real-time PCR (qPCR) experiments for 30 of the regions containing putatively aberrant SNPs, using DNA from individuals in our sample set. A comparison of the qPCR results with our genotype inferences is given in Table 1. Overall, 26

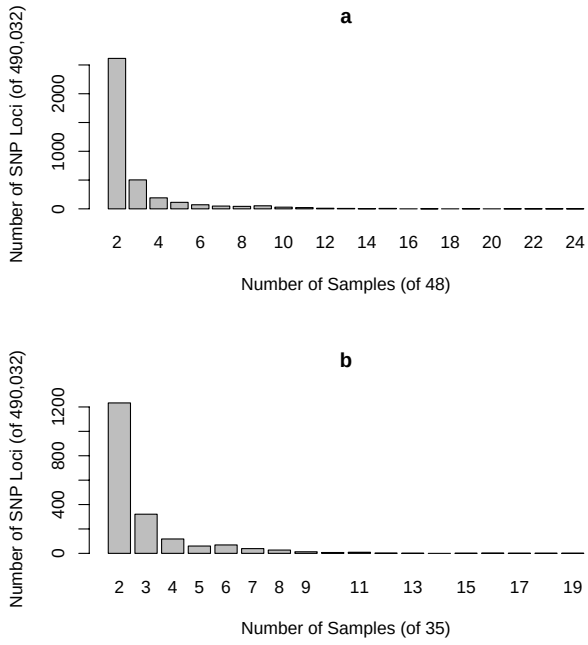

Figure 1

Frequencies of aberrant SNPs in our study. For each SNP, the number of samples aberrant at that locus was counted. We constructed histograms of frequencies for SNPs that were aberrant in more than one sample. (*a*) For each count *c* on the horizontal axis, the height of the bar indicates the number of SNPs that were aberrant in *c* samples out of 48. (*b*) Same as *a*, but with all trio offspring removed so that only 35 unrelated samples are considered.

(86.7%) of the qPCR results agreed with the presence and type (duplication or deletion) of CNV predicted by our *in silico* approach. There are a variety of potential reasons for the four nonconcordant loci. Since the PCR primers are explicitly designed to avoid the SNP sites (yet be near the aberrant SNPs), and since CNVs can be quite focal, it is possible for the PCR-amplified region to miss the aberrant locus entirely. A detected deletion could also be an artifact of the SNP array assay, since the deletion or duplication elsewhere in the restriction fragment may result in its length moving outside the range of lengths that the assay's PCR step would amplify [10]. As mentioned above, common CNVs in the human population are another potential source of error, since it is implicitly assumed that the reference panel primarily harbors two copies per cell at each locus. To provide further independent validation, we also performed multiplex ligation-dependent probe amplification (MLPA) [14] experiments on 17 other putative CNVs (Figure 2 and Additional data file 2). The concordance with the *in silico* genotypes was similar to that of the qPCR results.

Mendelian inheritance considerations

It is well-established [9,15,16] that there are genomic regions harboring germline CNVs in the form of both duplications and deletions in different individuals. Indeed, the non-allelic homologous recombination model of the formation of the variants generates both duplications and deletions, simultaneously at the same locus. It follows that, as these variants segregate through the population, there will be individuals carrying both a gain and a loss at the same locus, though on different parental chromosomes. The observable patterns of Mendelian inheritance in the aberrant genotype setting are different from those in diallelic SNP genotyping or even in aggregate copy number measurement, and ignoring the presence of CNVs can result in the false appearance of non-Mendelian inheritance as a result of genotyping errors (Figure 3a and Additional data file 3). Truly non-Mendelian inheritance, *e.g. de novo* events (Figure 3b and Additional data file 3) can be masked as well. These errors will also occur when only aggregate copy number is measured but allelic information is ignored. Misinterpreting

Table 1: Comparison of *in silico* and *in vitro* results for 30 putatively aberrant SNPs. Here, the diploid genotype refers to that provided by the SNP array's default software [24], under the assumption of two copies of the SNP. The errors shown are typical in the presence of CNVs. The aberrant genotype here is our algorithm's call. We consider the putative CNV to be validated if the rounded qPCR copy number is less than 2 (for deletions) or greater than 2 (for duplications).

Sample	rs ID	Diploid genotype	Aberrant genotype	Frequency (unrelated samples)	qPCR copy number	Validated?
NA10851	rs17525374	GG	G -	1	1.17	Yes
NA10851	rs9542207	TT	T -	1	0.86	Yes
NA10851	rs6601728	GG	G -	1	1.40	Yes
NA10851	rs17133566	CC	C -	1	0.98	Yes
NA10863	rs5751296	CT	CTT	1	2.74	Yes
NA10863	rs17577094	TT	TTT	5	4.27	Yes
NA12707	rs1565516	GG	G -	1	1.95	No
NA12707	rs2013317	GG	GGG	4	6.14	Yes
NA12801	rs2304717	CC	CCC	1	4.29	Yes
NA12801	rs12697975	CC	C -	1	0.84	Yes
NA12801	rs2889833	GG	G -	1	1.24	Yes
NA10851	rs11780672	AA	A -	1	0.76	Yes
NA10863	rs3828886	AA	AAA	8	2.12	No
NA10863	rs3858489	AG	AGG	5	3.40	Yes
NA10863	rs17662235	TT	TTT	7	3.36	Yes
NA10863	rs6994627	AA	A -	1	0.88	Yes
NA10863	rs2604357	GG	G -	1	1.54	No
NA10863	rs10110189	CC	C -	1	0.85	Yes
NA10863	rs2721243	CC	C -	1	1.01	Yes
NA10863	rs737714	No Call	A -	1	0.00	Yes
NA10863	rs7833963	AA	A -	1	1.05	Yes
NA10863	rs4831667	AA	A -	1	0.82	Yes
NA10863	rs7464441	AA	A -	1	1.36	Yes
NA12707	rs1018685	AA	AAA	1	1.59	No
NA12707	rs16993280	No Call	- -	3	0.00	Yes
NA12707	rs361901	AA	A -	4	1.40	Yes
NA12707	rs12170791	CC	C -	2	0.98	Yes
NA12707	rs2532292	AA	AAA	7	4.51	Yes
NA12707	rs2732675	No Call	TTT	10	2.53	Yes
NA12707	rs4822622	CC	C -	1	1.06	Yes

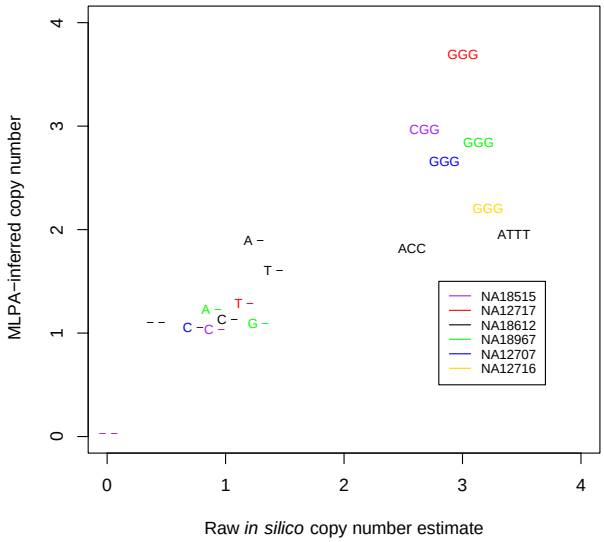

Figure 2
Concordance between *in silico* and MLPA results. We plotted the MLPA-inferred dosage against our raw total copy number inferences (see Methods) from the SNP array data. The plotting symbol for each SNP site is its genotype inferred by our procedure, and the color indicates the sample. The concordance is very strong with the exception of one sample, NA18612 (black), which could be due to either noisy array data or experimental difficulties for that particluar sample.

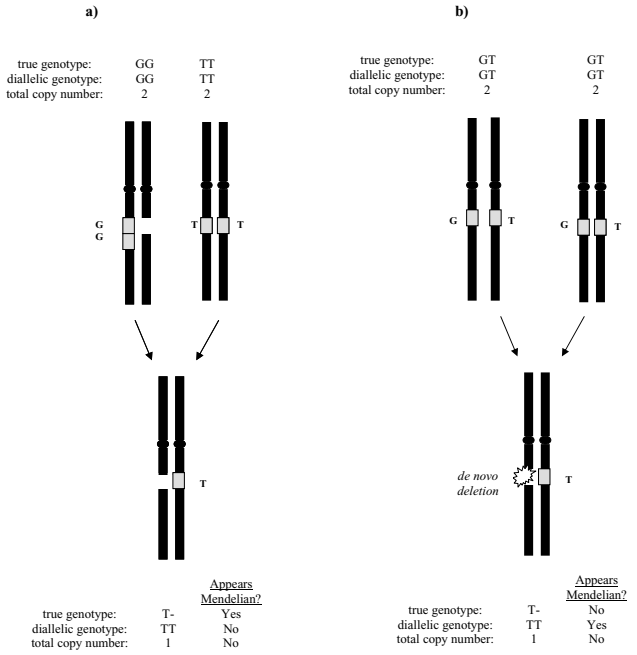

Figure 3
Mendelian and non-Mendelian inheritance patterns in the presence of CNVs. (*a*) Although the parent on the left has copy number two at the locus, this is the result of a duplication on one chromsome and a deletion on the other. Standard (diallelic) genotyping methods would incorrectly identify a non-Mendelian pattern at the locus in this trio, as would total copy number information alone. However, accurate genotyping, taking CNVs into account, reveals that the pattern is Mendelian. (*b*) The *de novo* deletion is obscured when diallelic genotyping alone is considered, though copy number information reveals the event.

germline CNVs as *de novo*, or *vice versa*, can have important implications, particularly in clinical settings where such variants in an affected individual are considered less relevant when inherited from an unaffected parent [17]. With multi-allelic SNP genotypes, we should be able to distinguish between the two cases. We were able to check aberrant SNPs for Mendelian inheritance in the 13 mother-father-child trios in our data set. Accounting for both SNP and copy number variation, 1,535 of the 1,771 instances of putatively aberrant SNPs in the four individuals (86.7%) demonstrated Mendelian inheritance. Possible explanations for non-Mendelian events include *de novo* CNVs or uniparental disomy, both of which have been detected using the SNP array platform [18-20]. Alternatively, detected CNVs may be instead artifacts of cell line culture, as observed in [9].

SNP alleles in duplications

Examination of the SNP allelic composition of duplicated regions can provide insight into the history of copy number variation in the human population. In the case of a SNP site that is duplicated with (haploid) copy number two, there are theoretically five possible haploid genotypes for that SNP: AA, AB, BB, A, and B, where A and B are the two base residues for the SNP site. The presence or absence of each of these five haploid genotypes sheds light on recurrence and temporal order of both the duplication event at that locus as well as the point mutation that resulted in the SNP. For example, the presence of both AA and BB would imply a recurrent duplication, occurring on both the A and B SNP background. To investigate empirically, we examined 496 aberrant loci in detail (see Methods). We found no evidence of a single chromosome with different base residues on each copy of a duplication (i.e. an AB chromosome) (Figure 4a). It is therefore unlikely that the "SNP in duplicon" phenomenon noted recently [21] for segmental duplications is common in CNVs. This is also consistent with the conclusion of the HapMap consortium [22] that the point mutations leading to SNPs are largely non-recurrent. In the vast majority of cases, only one of the SNP alleles was duplicated in our sample set. Still, six (1.2%) of the SNP sites had evidence

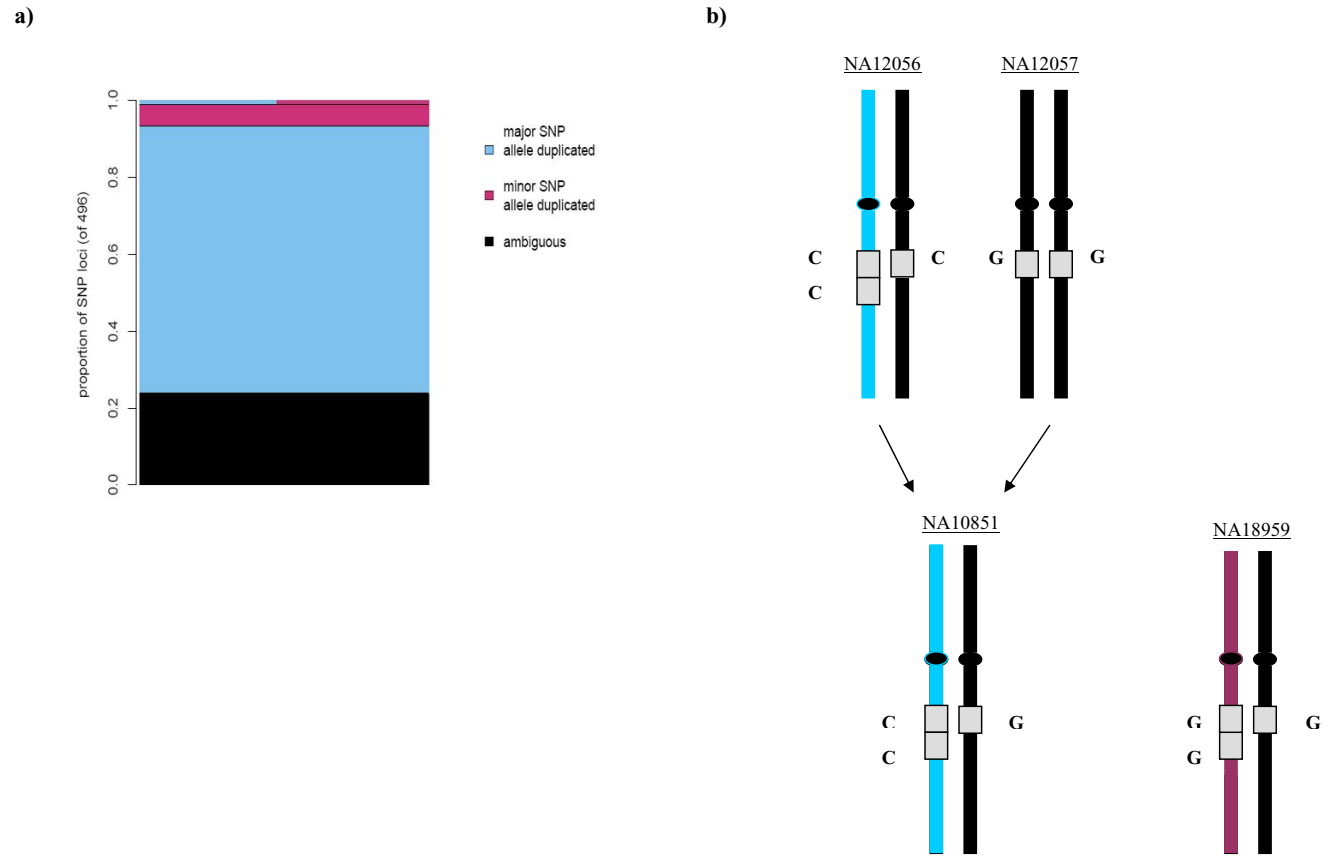

Figure 4

Interplay between duplication status and SNP allele. (*a*) Of 496 interrogated duplication loci, we observed 6 cases (1.2%) with both AA and BB chromosomes, and none with AB chromosomes. When only one SNP allele was unambiguously duplicated, 92.5% of the time it was the major allele in our sample set. (*b*). An example of a Caucasian trio and a Japanese individual harboring different SNP alleles in the duplicon.

of both AA and BB chromosomes. The presence of both types show that at least some duplication events were most likely recurrent in human history (alternative explanations seem unlikely, particularly given the complete absence of evidence for any AB chromosomes). An example is rs7895458 on chromosome 10 (Figure 4b), which is contained in a previously known recurrently duplicated region [9]. In our data set, one Caucasian family – NA12056 (father) and NA10851 (child) – harbors the duplication with the C allele at the SNP in both copies. Japanese individual NA18959, on the other hand, harbors the duplication with the G allele at the SNP in both copies. Interestingly, the dbSNP database [23] lists the allele frequency for C as 65% in the Caucasian population, but only 15% for the Japanese population. These duplications were observed independently in [9] in all three of these individuals, though the SNP allelic information was ignored in that study. Similarly, the SNP array manufacturer's [24] diallelic calls, CC for NA10851 and GG for NA18959, were erroneous though consistent with what was expected given their two-copy assumption. This example points to the insights that can be gained from consideration of both copy number and SNP residue simultaneously.

Comparison with previous work

Since our detected aberrant SNPs automatically indicate the presence of CNVs, we compared the genomic coordinates of these loci to CNVs previously reported in the literature. Of our 21,568 SNP loci, 5537 (25.7%) are contained in regions catalogued as CNVs in the Database of Genomic Variants, while only 11.9% of the 490,032 autosomal SNPs on the SNP array are contained in these regions. Though this demonstrates more overlap than would be expected by chance, the majority of loci we have uncovered are novel. However, one would expect that recurrently variant SNPs in our data set would often exist at a higher frequency in the general population, and

would thus be more likely to have been previously discovered. Indeed, if we restrict our attention to SNPs that are aberrant in at least two unrelated individuals in our sample set, we find that 1062 of 1905 (55.7%) of these lie in previously reported CNVs. As the aberrance rate of a SNP in our sample increases, its likelihood of having been previously reported continues to rise – over 95% of the 351 SNPs that are aberrant in at least 10% of our unrelated sample set are contained in previously reported CNVs. Therefore, although we report thousands of new CNVs, our results are in some sense concordant with what has been revealed on a population level. Since it has been shown that sequence between intrachromosomal segmental duplications are predisposed to CNVs [25], we also checked for enrichment of these "hotspots" [16] in our set of CNVs. Only 2,336 (10.8%) of our aberrant SNP loci were contained in these regions reported in the Structural Variation Database, which is only a slightly higher proportion than that of all autosomal SNPs on the array (8.1%).

Genes in regions of copy number variation

Many genes previously confirmed as polymorphic have functions in metabolism and immunity, and likely act as mediators of normal human variation as well as genomic disease. We compiled a list of transcribed CNVs in our data, along with the Gene Ontology [26] (GO) terms that were associated to these transcripts. We examined our list for GO terms that were present at a statistically higher rate compared to all genes containing SNPs represented on the array (see Methods). This allowed the identification of several interesting categories of genes involved in cell surface structure, glutamate metabolism and signaling, and genes with metabolic, enzymatic and neurological functions. In concordance with previous studies, we confirmed the presence of CNVs in genes such as DUSP22, NCAM2 [27], and NF1 [16]. Also present in our list are genes that are known to influence "normal" human phenotypes, such as the copy number-polymorphic olfactory receptor genes [28], and the neuropeptide-Y4 receptor PPYR1 [27] that is directly involved in the regulation of food intake and body weight [29]. A number of "environmental sensor" genes involved in immune system response were also observed, in categories such as granulocyte differentiation, receptor-mediated endocytosis, antibiotic response, regulation of IgG/IgE isotype switching, regulation of NK cell activity, IL-4 receptor binding and MHC class 1 receptor activity. In addition, a large proportion of the CNV-enriched classes have receptor and/or signaling functions. It is important to note that a number of genes previously reported to be copy number variant, such as the glutathione S-transferase genes GSTM1 and GSTT1, are not represented on the SNP array, perhaps due to the fact that they would give ambiguous genotype calls.

Conclusion

We have presented the first computational method for genotyping SNPs from microarray data in the general case where the individual is not restricted to two copies of the SNP locus per cell. Our work highlights, with several examples, the relevance of considering both copy number and SNP allelic information simultaneously. We have uncovered tens of thousands of SNPs with aberrant genotypes in humans of various ethnicities, corresponding to thousands of novel CNVs. It is likely that our results actually drastically underrepresent the prevalence of aberrant SNPs in the population, as the array's manufacturer deliberately excluded SNPs that violated Hardy-Weinberg equilibrium, Mendelian inheritance, and other quality control requirements [10] that would naturally not be met in the presence of CNVs. Moreover, our own requirement that at least three consecutive SNPs show the CNV is very conservative, and will by definition omit more focal events (in practice, our method can be tuned in this way to control the false positive/false negative tradeoff, as desired). It follows that the number and frequency of these multi-allelic SNPs segregating through the population is likely to be much more substantial than previously suspected, and therefore generalized genotyping of the sort we have described here is crucial in studies using SNPs as markers. Our work is a step in that direction, though the goal should be to attain the high rates of accuracy (> 99%) that are assumed in the diallelic setting. Such highly accurate genotyping would automatically provide information regarding the presence or absence of CNVs. Given the difficulty in precisely mapping the boundaries of these germline CNVs (with recent evidence that the boundaries actually differ between individuals in the population [30]), and given the density of SNPs on the genome, we propose aberrant SNP genotyping as an alternative to other methods of categorizing CNVs from SNP array data [31]. Such genotypes map to precise genomic locations, and provide information on both copy number and base residue. The resulting multi-allelicism will be an aid to disease association studies, whether these more accurately ascertained SNP alleles are actually causal inherited variants, or are simply used as markers. Since there are many thousands of SNP array samples extant, these multi-allelic SNPs segregating through the population will be identified and their frequencies ascertained. We have developed software, freely available at our web site, with which users can scan their arrays for aberrant SNPs, using reference panels of their choice. As the platforms increase in throughput and decrease in cost, accurate multi-allelic genotyping will be even more crucial.

Methods

SNP array data and biological samples

The raw .cel files from the 48 individuals – NA10851, NA10855, NA10863, NA11831, NA11832, NA12056,

NA12057, NA12234, NA12264, NA12707, NA12716, NA12717, NA12801, NA12812, NA12813, NA18503, NA18504, NA18505, NA18506, NA18507, NA18508, NA18515, NA18516, NA18517, NA18532, NA18545, NA18558, NA18605, NA18612, NA18959, NA18967, NA18969, NA18997, NA19137, NA19138, NA19139, NA19152, NA19153, NA19154, NE00088, NE00090, NE00091, NE00375, NE00403, NE00598, NE00963, NE01118, and NE01119 – in the Mapping 500 K Sample Data Set were downloaded from the Affymetrix web site [32]. These individuals are of African (15), European American (15), Han Chinese (5), Japanese (4), African American (3), Asian American (3), and Hispanic American (3) ethnicities. DNA and cell lines from 20 of these individuals was obtained from the Coriell Cell Repositories for qPCR and MLPA experiments.

Generalized genotyping and candidate CNVs

We used 16 individuals – NA11831, NA12057, NA18505, NA18507, NA18517, NA18532, NA18545, NA18558, NA18959, NA18967, NA18969, NA19138, NA19152, NE00090, NE00403, and NE01119 – as our reference panel, selected because they are unrelated and are from a variety of ethnic backgrounds. Using the data from these 16, we trained the PLASQ [12] model parameters as described. We then used PLASQ to infer the "raw" allele-specific copy numbers (ASCNs) in our test samples, restricting our attention to the autosomes. The pairwise sums of the raw ASCNs yielded raw total copy numbers, which were rounded to the nearest integer for total copy numbers. Calls with total copy number deviating from two provided a preliminary list of aberrant SNPs. These were converted to the generalized genotypes by assigning the whole-number portions of the total copy number to each allele so that the (nearest integer) raw ASCNs were retained as much as possible. In order to enrich our candidate set for true positives, we restricted our attention to SNPs in runs of at least three independent aberrant SNPs with aberrations in the same direction (duplication or deletion). In this context, we consider adjacent SNPs to be independent only if they reside on different restriction enzyme fragments, since fragment-specific artifacts arising during the PCR step of the SNP array protocol [10] would presumably affect all SNPs on the fragment.

PCR validation of CNVs

Relative gene copy numbers were determined by quantitative real-time PCR using a PRISM 7900HT Sequence Detection System (384 well) (Applied Biosystems, Foster City, CA). Real-time PCR was performed in 12.5-μl (384 well) reactions with 2 ng of template DNA. A QuantiTect SYBR Green PCR kit (Qiagen Inc., Valencia, CA) was used for the PCR reaction. PCR conditions were as follows: 2 min at 50°C, 15 min at 95°C, followed by 40 three-step cycles of (20 s at 95°C, 20 s at 58°C, and 30 s at 72°C). Primers were designed using Primer 3 [33] and synthesized by Integrated DNA Technologies (IDT; Coralville, IA). Primer sequences are available upon request. Quantification was based on standard curves from a serial dilution of human normal genomic DNA. The standard curve method was used to calculate target DNA copy number in each DNA sample normalized to a repetitive element Line-1 and normal reference DNA. For our reference sample, we used female genomic DNA pools, derived from multiple anonymous donors (Promega, Madison, WI), since combining DNA from multiple individuals should dilute out all but the most common copy number variants.

MLPA validation of CNVs

Custom MLPA probes were designed to match suitable sequences within 300 bp of the original SNP location. Control probes were drawn from other chromosomal locations and have previously been used to analyze more than 100 individuals without evidence for copy number variation [34]. Oligonucleotides were synthesized by IDT, with 5'-phosphorylation of each downstream probe and tagged with common PCR primer sequences [14]. Probes were hybridized with 100 ng aliquots of DNA using MLPA reagents (part number EK5, MRC-Holland BV, Amsterdam, The Netherlands) according to the recommended protocol. Samples were then diluted 10-fold and analyzed on a 3730xl sequencer with GeneMapper software (Applied Biosystems). We used male and female genomic DNA pools, derived from multiple anonymous donors (Promega, Madison, WI). Furthermore, peak height ratios were normalized to the mean of the entire data set, rather than to the controls alone, with subsequent elimination of outlier samples from the calculation of the mean. Our experience with well-characterized deletions shows that this approach gives equivalent results to normalizing against controls alone, provided that samples with altered copy number are in the minority (data not shown).

Analysis of SNP alleles in duplications

Determining which SNP alleles are harbored in a duplication is subject to the same phasing difficulties as SNP haplotype determination. In order to maximize our ability to determine correct phasing, we considered only SNP sites that were duplicated to total copy number three in at least one trio offspring, with one parent having copy number three and the other either two or three. To avoid the possibility of a deletion on one chromosome, we omitted from consideration all loci contained in a deleted regions in any individual (either in our data or in the Database of Genomic Variants). This left us with 496 SNPs for which we sought to detect the presence of the AA, AB, BB, A, and B chromosomes in our sample set, through a combination of phasing and examination of individual genotypes (see Additional data file 4). Note that lack of detection does

not necessarily indicate absence, but could instead be the result of ambiguous phase.

Comparison with previously-published CNVs and segmental duplications

The previously published CNVs were those catalogued (as of December 4, 2006) in the Database of Genomic Variants (Build 35 coordinates). The "rearrangement hotspots" [16] are regions, between 50 kb and 10 Mb in length, flanked by segmental duplications at least 10 kb in length with at least 95% sequence identity. The Build 35 coordinates of these segmental duplications were downloaded from the Segmental Duplication Database [35] on December 4, 2006.

Statistical analysis of GO associations in CNVs

We mapped all SNPs on the array to their genomic positions using the UCSC Genome Browser (Build 35). The 11,944 genes whose transcribed regions contain at least one of the (autosomal) SNPs on the array comprised our "gene universe". Our duplicated genes and deleted genes were those with transcripts containing SNP sites that were duplicated or deleted, respectively, in at least one of our 48 samples. We made use of the R [36] software package GOstats [37] to test our duplicated and deleted genes for statistical enrichment in certain GO terms (as annotated by the hgu133plus2 package) [38]. Briefly, for a fixed GO term, the software performs a Fisher's exact test for the null hypothesis of no association between duplication or deletion status and annotation to that term, using all genes in our gene universe.

Authors' contributions

LEM carried out all PCR experiments, helped refine the computational method, and drafted parts of the manuscript. MAA carried out all MLPA experiments and drafted parts of the manuscript. XL contributed expertise in the handling and extraction of DNA. TL conceived of the study, designed the statistical methodology, and wrote the manuscript. All authors read and approved the final manuscript.

Additional material

Additional file 1

All 21568 duplicated and deleted SNPs. *This table lists all 21568 detected SNPs. Included in the table are SNP ID, genomic coordinates, frequency of deletion/duplication, and the gene harboring the SNP.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-8-211-S1.xls]

Additional file 2

Comparison of in silico results with MLPA results. *This table lists sample, SNP ID, genomic coordinates, MLPA copy number, and in silico copy number and generalized genotype for all 17 SNP loci that were compared.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-8-211-S2.xls]

Additional file 3

Mendelian and non-Mendelian inheritance patterns identified by generalized genotype. *This figure shows two hypothetical cases in which the generalized genotype accurately assesses Mendelian inheritance. a) Under the assumption of a diallelic genotype, the inheritance appears to be non-Mendelian. When copy number variation is taken into account, Mendelian inheritance is revealed. b) The* de novo *duplication is obscured when total copy number alone is considered. However, the true genotype uncovers this event, since allelic information is taken into account.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-8-211-S3.pdf]

Additional file 4

Supplementary Methods. *This file describes how we were able to "phase" the alleles in duplicated SNPs.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-8-211-S4.pdf]

Acknowledgements

We thank Joseph Nadeau for valuable discussion. We are also grateful to the Genomics Core Facility, Lerner Research Institute, for efficient genotyping services.

References

1. **CNVgeno R package** [http://genetics.case.edu/LaFramboise/CNVgeno/]
2. Feuk L, Carson AR, Scherer SW: **Structural variation in the human genome.** *Nat Rev Genet* 2006, **7(2):**85-97.
3. Pollack JR, Sorlie T, Perou CM, Rees CA, Jeffrey SS, Lonning PE, Tibshirani R, Botstein D, Borresen-Dale AL, Brown PO: **Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors.** *Proc Natl Acad Sci USA* 2002, **99(20):**12963-12968.
4. Lucito R, Healy J, Alexander J, Reiner A, Esposito D, Chi M, Rodgers L, Brady A, Sebat J, Troge J, West JA, Rostan S, Nguyen KCQ, Powers S, Ye KQ, Olshen A, Venkatraman E, Norton L, Wigler M: **Representational oligonucleotide microarray analysis: a high-resolution method to detect genome copy number variation.** *Genome Res* 2003, **13(10):**2291-2305.
5. LaFramboise T, Weir BA, Zhao X, Beroukhim R, Li C, Harrington D, Sellers WR, Meyerson M: **Allele-specific amplification in cancer revealed by SNP array analysis.** *PLoS Comput Biol* 2005, **1(6):**e65.
6. Freeman JL, Perry GH, Feuk L, Redon R, McCarroll SA, Altshuler DM, Aburatani H, Jones KW, Tyler-Smith C, Hurles ME, Carter NP, Scherer SW, Lee C: **Copy number variation: new insights in genome diversity.** *Genome Res* 2006, **16(8):**949-961.
7. Conrad DF, Andrews TD, Carter NP, Hurles ME, Pritchard JK: **A high-resolution survey of deletion polymorphism in the human genome.** *Nat Genet* 2006, **38:**75-81.
8. McCarroll SA, Hadnott TN, Perry GH, Sabeti PC, Zody MC, Barrett JC, Dallaire S, Gabriel SB, Lee C, Daly MJ, Altshuler DM: **Common deletion polymorphisms in the human genome.** *Nat Genet* 2006, **38:**86-92.
9. Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, Fiegler H, Shapero MH, Carson AR, Chen W, Cho EK, Dallaire S, Free-

man JL, Gonzalez JR, Gratacos M, Huang J, Kalaitzopoulos D, Komura D, MacDonald JR, Marshall CR, Mei R, Montgomery L, Nishimura K, Okamura K, Shen F, Somerville MJ, Tchinda J, Valsesia A, Woodwark C, Yang F, Zhang J, Zerjal T, Zhang J, Armengol L, Conrad DF, Estivill X, Tyler-Smith C, Carter NP, Aburatani H, Lee C, Jones KW, Scherer SW, Hurles ME: **Global variation in copy number in the human genome.** *Nature* 2006, **444(7118):**444-454.

10. Affymetrix: *GeneChip Human Mapping 500 K Array Set Data Sheet* Santa Clara (California): Affymetrix, Inc; 2005.
11. **The Database of Genomic Variants** [http://projects.tcag.ca/variation]
12. LaFramboise T, Harrington D, Weir BA: **PLASQ: A generalized linear model-based procedure to determine allelic dosage in cancer cells from SNP array data.** *Biostatistics* 2007, **8(2):**323-336.
13. Snijders AM, Nowak N, Segraves R, Blackwood S, Brown N, Conroy J, Hamilton G, Hindle AK, Huey B, Kimura K, Law S, Myambo K, Palmer J, Ylstra B, Yue JP, Gray JW, Jain AN, Pinkel D, Albertson DG: **Assembly of microarrays for genome-wide measurement of DNA copy number.** *Nat Genet* 2001, **29(3):**263-264.
14. Schouten JP, McElgunn CJ, Waaijer R, Zwijnenburg D, Diepvens F, Pals G: **Relative quantification of 40 nucleic acid sequences by multiplex ligation-dependent probe amplification.** *Nucleic Acids Res* 2002, **30(12):**e57.
15. Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, Qi Y, Scherer SW, Lee C: **Detection of large-scale variation in the human genome.** *Nat Genet* 2004, **36(9):**949-951.
16. Sharp AJ, Locke DP, McGrath SD, Cheng Z, Bailey JA, Vallente RU, Pertz LM, Clark RA, Schwartz S, Segraves R, Oseroff VV, Albertson DG, Pinkel D, Eichler EE: **Segmental duplications and copy-number variation in the human genome.** *Am J Hum Genet* 2005, **77:**78-88.
17. de Vries BBA, Pfundt R, Leisink M, Koolen DA, Vissers LELM, Janssen IM, Reijmersdal Sv, Nillesen WM, Huys EHLPG, Leeuw Nd, Smeets D, Sistermans EA, Feuth T, van Ravenswaaij-Arts CMA, van Kessel AG, Schoenmakers EFPM, Brunner HG, Veltman JA: **Diagnostic genome profiling in mental retardation.** *Am J Hum Genet* 2005, **77(4):**606-616.
18. Bruce S, Leinonen R, Lindgren CM, Kivinen K, Dahlman-Wright K, Lipsanen-Nyman M, Hannula-Jouppi K, Kere J: **Global analysis of uniparental disomy using high density genotyping arrays.** *J Med Genet* 2005, **42(11):**847-851.
19. Altug-Teber O, Dufke A, Poths S, Mau-Holzmann UA, Bastepe M, Colleaux L, Cormier-Daire V, Eggermann T, Gillessen-Kaesbach G, Bonin M, Riess O: **A rapid microarray based whole genome analysis for detection of uniparental disomy.** *Hum Mutat* 2005, **26(2):**153-159.
20. Friedman JM, Baross A, Delaney AD, Ally A, Arbour L, Armstrong L, Asano J, Bailey DK, Barber S, Birch P, Brown-John M, Cao M, Chan S, Charest DL, Farnoud N, Fernandes N, Flibotte S, Go A, Gibson WT, Holt RA, Jones SJM, Kennedy GC, Krzywinski M, Langlois S, Li HI, McGillivray BC, Nayar T, Pugh TJ, Rajcan-Separovic E, Schein JE, Schnerch A, Siddiqui A, Van Allen MI, Wilson G, Yong SL, Zahir F, Eydoux P, Marra MA: **Oligonucleotide microarray analysis of genomic imbalance in children with mental retardation.** *Am J Hum Genet* 2006, **79(3):**500-513.
21. Fredman D, White SJ, Potter S, Eichler EE, Den Dunnen JT, Brookes AJ: **Complex SNP-related sequence variation in segmental genome duplications.** *Nat Genet* 2004, **36(8):**861-866.
22. International HapMap Consortium: **A haplotype map of the human genome.** *Nature* 2005, **437(7063):**1299-1320.
23. **The Single Nucleotide Polymorphism Database** [http://www.ncbi.nlm.nih.gov/projects/SNP/]
24. Di X, Matsuzaki H, Webster TA, Hubbell E, Liu G, Dong S, Bartell D, Huang J, Chiles R, Yang G, Shen Mm, Kulp D, Kennedy GC, Mei R, Jones KW, Cawley S: **Dynamic model based algorithms for screening and genotyping over 100 K SNPs on oligonucleotide microarrays.** *Bioinformatics* 2005, **21(9):**1958-1963.
25. Shaw CJ, Bi W, Lupski JR: **Genetic proof of unequal meiotic crossovers in reciprocal deletion and duplication of 17p11.2.** *Am J Hum Genet* 2002, **71(5):**1072-1081.
26. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.** *Nat Genet* 2000, **25:**25-29.
27. Sebat J, Lakshmi B, Troge J, Alexander J, Young J, Lundin P, Maner S, Massa H, Walker M, Chi M, Navin N, Lucito R, Healy J, Hicks J, Ye K, Reiner A, Gilliam TC, Trask B, Patterson N, Zetterberg A, Wigler M: **Large-scale copy number polymorphism in the human genome.** *Science* 2004, **305(5683):**525-528.
28. Trask BJ, Friedman C, Martin-Gallardo A, Rowen L, Akinbami C, Blankenship J, Collins C, Giorgi D, Iadonato S, Johnson F, Kuo WL, Massa H, Morrish T, Naylor S, Nguyen OT, Rouquier S, Smith T, Wong DJ, Youngblom J, van den Engh G: **Members of the olfactory receptor gene family are contained in large blocks of DNA duplicated polymorphically near the ends of human chromosomes.** *Hum Mol Genet* 1998, **7:**13-26.
29. Sainsbury A, Schwarzer C, Couzens M, Jenkins A, Oakes SR, Ormandy CJ, Herzog H: **Y4 receptor knockout rescues fertility in ob/ob mice.** *Genes Dev* 2002, **16(9):**1077-1088.
30. Goidts V, Cooper DN, Armengol L, Schempp W, Conroy J, Estivill X, Nowak N, Hameister H, Kehrer-Sawatzki H: **Complex patterns of copy number variation at sites of segmental duplications: an important category of structural variation in the human genome.** *Hum Genet* 2006, **120(2):**270-284.
31. Komura D, Shen F, Ishikawa S, Fitch KR, Chen W, Zhang J, Liu G, Ihara S, Nakamura H, Hurles ME, Lee C, Scherer SW, Jones KW, Shapero MH, Huang J, Aburatani H: **Genome-wide detection of human copy number variations using high-density DNA oligonucleotide arrays.** *Genome Res* 2006, **16(12):**1575-1584.
32. **Affymetrix Web Site** [http://www.affymetrix.com]
33. **Primer 3** [http://frodo.wi.mit.edu/cgi-bin/primer3/primer3_www.cgi]
34. Aldred MA, Vijayakrishnan J, James V, Soubrier F, Gomez-Sanchez MA, Martensson G, Galie N, Manes A, Corris P, Simonneau G, Humbert M, Morrell NW, Trembath RC: **BMPR2 gene rearrangements account for a significant proportion of mutations in familial and idiopathic pulmonary arterial hypertension.** *Hum Mutat* 2006, **27(2):**212-213.
35. **The Segmental Duplication Database** [http://humanparalogy.gs.washington.edu]
36. R Development Core Team: *R: A Language and Environment for Statistical Computing* 2006 [http://www.R-project.org]. R Foundation for Statistical Computing, Vienna, Austria
37. Falcon S, Gentleman R: **Using GOstats to test gene lists for GO term association.** *Bioinformatics* 2007, **23(2):**257-258.
38. **Bioconductor** [http://www.bioconductor.org]