nature genetics

Complex SNP-related sequence variation in segmental genome duplications

David Fredman[1], Stefan J White[2], Susanna Potter[1], Evan E Eichler[3], Johan T Den Dunnen[2] & Anthony J Brookes[1,4]

There is uncertainty about the true nature of predicted single-nucleotide polymorphisms (SNPs) in segmental duplications (duplicons) and whether these markers genuinely exist at increased density as indicated in public databases. We explored these issues by genotyping 157 predicted SNPs in duplicons and control regions in normal diploid genomes and fully homozygous complete hydatidiform moles. Our data identified many true SNPs in duplicon regions and few paralogous sequence variants. Twenty-eight percent of the polymorphic duplicon sequences we tested involved multisite variation, a new type of polymorphism representing the sum of the signals from many individual duplicon copies that vary in sequence content due to duplication, deletion or gene conversion. Multisite variations can masquerade as normal SNPs when genotyped. Given that duplicons comprise at least 5% of the genome and many are yet to be annotated in the genome draft, effective strategies to identify multisite variation must be established and deployed.

Duplicons defined as being >1 kb with >90% similarity between copies comprise at least 5% of the human genome[1,2]. Their minimal extent has been defined[3], but the public human genome draft portrays duplicons neither accurately nor completely[4–6]. SNP databases report that SNPs are over-represented by a factor of ~2 in duplicon regions[3,7,8]. This is a minimum value, as SNP discovery efforts discard predicted variants from regions where densities are high or a duplicon is suspected[9,10]. Many or most duplicon SNPs may be nothing more than paralogous sequence variants (PSVs)[3,7,8]. Alternatively, gene conversion in duplicons may generate allelic diversity and SNP content[11,12]. Additionally, reduced selective pressure in duplicons may allow new mutations to increase in frequency more easily [13].

Initially, we undertook an *in silico* study of SNPs in duplicons to search for informative features. We noted an increased gene density in duplicons and observed that validated SNPs (65.2% of the dbSNP version used) were under-represented in duplicons compared with non-validated SNPs. Specifically, 3.7% (5.6% by two hit–two allele, 3.4% by cluster, 1.9% by frequency) of valid SNPs versus 13.1% of nonvalidated SNPs reside in the 4.5% of the genome comprised of duplicons. This could imply that duplicon SNPs are mostly PSVs, or it could reflect the difficulty of doing experiments with nonunique sequences.

We therefore devised an experiment to resolve PSVs from real SNPs. We used dynamic allele-specific hybridization (DASH)[14], which generates a DNA melting curve by heating an oligonucleotide probe duplexed with a PCR amplicon. Negative derivatives of these curves allow for direct comparisons of allele ratios in heterozygotes. Sample DNAs were from 16 normal Swedish females and 8 pathologically confirmed monospermic complete hydatidiform moles (CHMs)[15]. CHMs are fully homozygous genomes that allow distinction between true SNP alleles at a single genome locus (genotypes will always show single alleles) and PSV signals originating from multiple sites (genotypes will be 'heterozygote-like', including both alleles). The tested samples gave 98% power to detect alleles of 10% frequency[16]. We targeted 17 duplicons (**Table 1**) that fell into four broad classes according to their representation in the public genome assembly, their degree of sequence similarity and whether they seemed to be multicopy by analysis of whole-genome shotgun sequencing data (WSSD)[3]. We also included two genome regions known to be unique. For each tested region, we genotyped eight predicted SNPs that were outside known repeats as detected by RepeatMasker[17], as well as five other previously validated true SNPs of random location.

We knew that DASH would convert 90–95% of all true SNPs to useable assays[14], and we assumed that most copies of the duplicon targets would be amplified in the PCR (given the high sequence similarities of the tested duplicons). The derived results comprised various melting-curve patterns (**Fig. 1b**) that correspond to specific genetic structures (**Fig. 1a**). Overall, 107 markers were polymorphic and useable for our investigation, including 13 control markers that gave genotypes consistent with single-copy true SNPs (**Fig. 2a**). The 15 markers in duplicons that lacked WSSD support likewise produced signals consistent with true SNPs (**Fig. 2a**). This indicates that these unique genome regions were inappropriately assembled, leaving them as apparent duplicons in the public draft. It is estimated that >50% of duplicons represented in the genome draft are not real[3]. As illustrated by our data, SNP genotyping can provide an efficient means to identify these for targeted resolution.

[1]Center for Genomics and Bioinformatics, Karolinska Institute, Berzelius väg 35, S-171 77 Stockholm, Sweden. [2]Human and Clinical Genetics, Leiden University Medical Center, Wassenaarseweg 72, 2333 AL Leiden, the Netherlands. [3]Department of Genetics, Case Western Reserve University, 10900 Euclid Avenue, Cleveland, Ohio 44106, USA. [4]Present address: Department of Genetics, University of Leicester, University Road, Leicester LE1 7RH, UK. Correspondence should be addressed to A.J.B. (anthony.brookes@cgb.ki.se).

Published online 11 July 2004; doi:10.1038/ng1401

Table 1 Target regions

Region	WSSD	NCBI	Chrom	ChromStart (bp)	ChromEnd (bp)	Size (bp)	Name	Dispersal
A	Dup	Unique	1	85,402,915	85,427,399	24,485	–	Unknown
B	Dup	Unique	2	89,796,158	89,812,623	16,466	–	Unknown
C	Dup	Unique	16	18,167,513	18,191,332	23,820	–	Unknown
D	Dup	Unique	16	69,832,810	69,854,823	22,013	–	Unknown
E	Dup	Dup <98%	7	75,865,780	75,891,118	25,339	–	Intra
F	Dup	Dup <98%	9	85,988,721	86,012,093	23,373	–	Inter
G	Dup	Dup <98%	10	46,657,428	46,672,624	15,197	–	Intra
H	Dup	Dup <98%	11	88,972,901	88,996,892	23,992	–	Intra
I	Dup	Dup <98%	16	32,022,851	32,039,556	16,706	–	Inter
J	Dup	Dup >98%	8	7,161,589	7,293,710	132,121	8p23	Intra
K	Dup	Dup >98%	15	20,852,650	20,890,966	38,316	HERC2	Intra
L	Dup	Dup >98%	15	30,161,462	30,293,362	131,900	CHRNA7	Intra
M	Dup	Dup >98%	16	16,603,367	16,682,029	78,662	LCR16a	Intra
N	Dup	Dup >98%	17	44,072,366	44,126,506	54,140	MS	Intra
O	Unique	Dup >98%	1	57,845,958	57,856,075	10,117	–	Intra
P	Unique	Dup >98%	11	133,555,034	133,578,684	23,650	–	Intra
Q	Unique	Dup >98%	12	51,307,117	51,382,529	75,412	–	Intra
R	Unique	Unique	16	21,560,883	21,636,826	75,943	–	Unique
S	Unique	Unique	22	20,825,861	20,875,861	50,000	–	Unique
T	Unique	Unique	Various	Random validated SNPs	–	Unique		

Coordinates are from the July 2003 NCBI assembly. These comprise 17 duplicons and additional controls, covering a total of 1 Mb, taken from 12 different chromosomes. The target regions were grouped into four broad classes: A–D, domains that are present uniquely in the NCBI assembly but that are indicated to be duplicons by WSSD; E–I, duplicated domains in the NCBI assembly having 90–98% sequence similarity and WSSD support; J–N, duplicated domains in the assembly with >98% similarity and WSSD support; O–Q, duplicated domains in the assembly with >98% similarity but no WSSD support. Regions R–T are unique control sequences.

Behavior of markers in WSSD-positive regions was substantially different from that of those in control regions (**Fig. 2a**,**b**). A full 91% (72 of 79) of duplicon assays gave apparent heterozygote signals in at least one CHM. To interpret the various genotype patterns, we established a classification schema (**Table 2**). Many duplicon markers behaved as real SNPs, residing either in unique sequence (7 of 79, 8.9%) or in one copy of a duplicon (32 of 79, 41%). This total (50%) equates to a SNP density that is equivalent to the genome average, as duplicons are enriched for predicted SNPs by a factor of 2 in public databases[3,7,8]. In addition, and contrary to previous evidence[3,7,8], only 23% (18 of 79) of duplicon markers behaved as PSVs. The remaining 28% (22 of 79) of predicted SNPs in duplicons were neither PSVs nor SNPs but gave complex genotyping patterns that have not been described before. We called this new form of polymorphism multisite variation (MSV).

When we assessed MSVs in CHMs, they generated either homozygous genotypes, indicative of SNPs, or apparently heterozygous signals, indicative of PSVs, (**Fig. 1b**). Two such signals are combined in diploid DNAs, and so MSVs gave genotypes in normal samples that

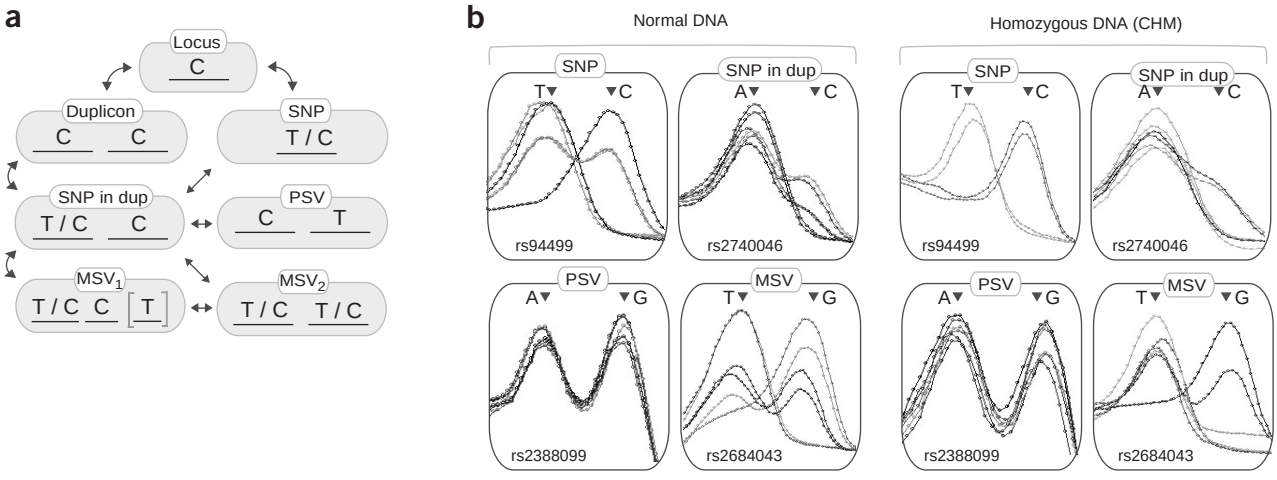

Figure 1 Genotyping patterns identifying evolutionary sequence states. (**a**) Evolutionary sequence changes from a monomorphic base to a polymorphic MSV. Arrows depict processes such as mutation, fixation, duplication, deletion and gene conversion. Most events are reversible. (**b**) Representative DASH genotyping patterns observed in normal and CHM samples for the corresponding structures in **a**. Each line shows the negative derivative of the melting curve of a probe-target duplex for one DNA sample. The temperature on the *x* axis ranges from 45 to 75 °C. Peaks marked by arrowheads indicate the presence of each particular allele as marked, with peak heights indicating the relative amount of each allele present in the tested DNA. Dup, duplicon.

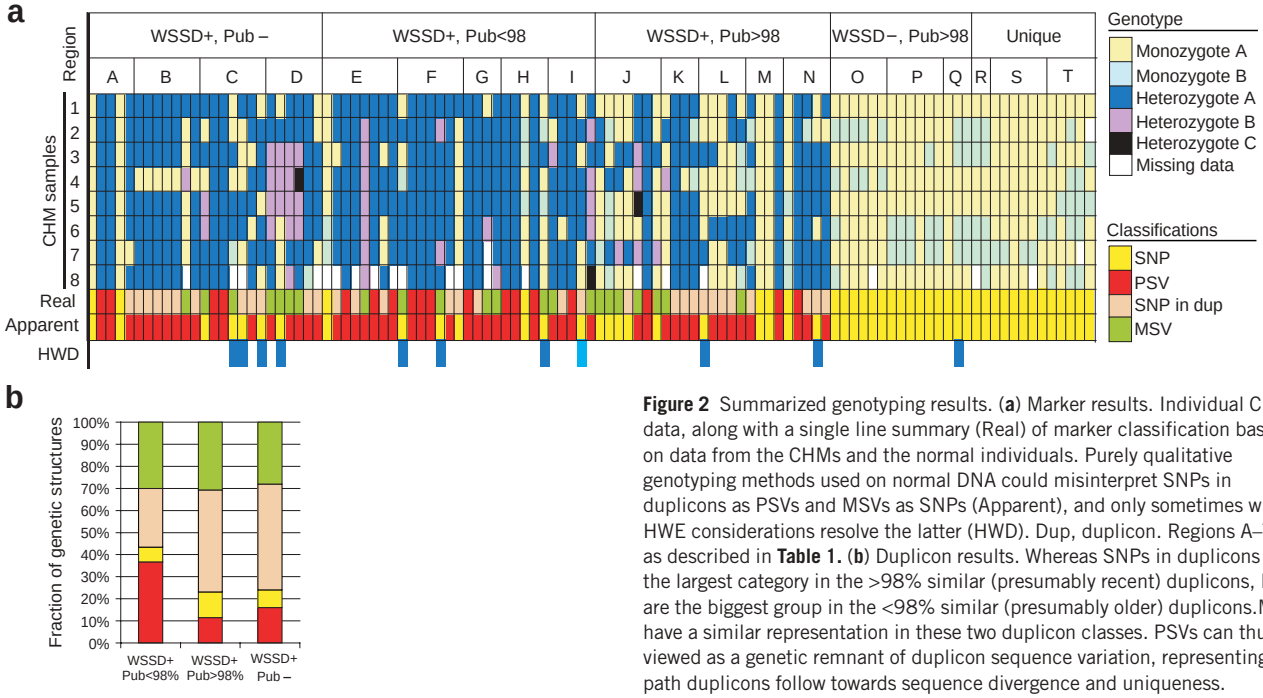

Figure 2 Summarized genotyping results. (**a**) Marker results. Individual CHM data, along with a single line summary (Real) of marker classification based on data from the CHMs and the normal individuals. Purely qualitative genotyping methods used on normal DNA could misinterpret SNPs in duplicons as PSVs and MSVs as SNPs (Apparent), and only sometimes will HWE considerations resolve the latter (HWD). Dup, duplicon. Regions A–T are as described in **Table 1.** (**b**) Duplicon results. Whereas SNPs in duplicons are the largest category in the >98% similar (presumably recent) duplicons, PSVs are the biggest group in the <98% similar (presumably older) duplicons.MSVs have a similar representation in these two duplicon classes. PSVs can thus be viewed as a genetic remnant of duplicon sequence variation, representing the path duplicons follow towards sequence divergence and uniqueness.

masqueraded as typical SNPs, but with variable allele ratios across individuals. These patterns may be explained as the sum of individual genotyping signals from various similar-sequence duplicon copies, with those duplicons themselves varying in the population. This variation may be due to (i) duplicon copy-number differences that lead to an increase, decrease or elimination of signals from different alleles that reside on the inserted or deleted duplicon copies (**Fig. 1a** ; MSV_1 pattern) or (ii) gene conversion events that lead to dispersion, mixing and perhaps homogenization of single-base alternatives across the various copies of a duplicon (**Fig. 1a**; MSV_2 pattern).

There is considerable evidence that gene conversion[18,19] and copy number variation[20,21] are active in subsets of duplicons. To evaluate the generality of these processes, we assessed sequences adjacent to 16 discovered MSVs (in nine duplicons) and two control SNPs for copy-number variation using multiplex ligation-dependent probe amplification (MLPA)[22,23]. We used another six control sequences for normalization. No CHM had more than about ten copies of any interrogated sequence (**Supplementary Fig. 1** online), and there was considerable evidence for copy-number variation in 50% (8 of 16) of cases (**Table 3**). Furthermore, sequences close to MSVs with a larger number of different allele ratios (as assessed by DASH) tended to report greater copy-number variability (**Supplementary Fig. 2** online). Thus, MSVs are a consequence (at least in part) of widespread duplicon copy-number variation. This interpretation is supported by Fosmid end-mapping data (E.E.E., unpublished results) and studies of copy-number differences related to disease[6,20,21,24]. Only some closely spaced markers showed correlated MLPA ratios (**Fig. 3**), however, indicating that there is substantial within-duplicon heterogeneity in this phenomenon.

Counting SNPs and MSVs together, at least two-thirds of predicted duplicon SNPs in public databases are polymorphic rather than PSVs. The one-third of these that are MSVs produce genotype patterns in diploid samples very similar to those of SNPs, other than having (sometimes subtle) allele ratio variability in heterozygotes. Genotyping technologies will need to detect this allele ratio variability to reliably identify MSVs. This raises a concern regarding whole-genome amplification procedures, which may distort these allele ratios. In pooled

Table 2 Identification of genomic structures by analysis of DASH genotypes for CHMs and normal DNA

Genetic structure	Material	Number of alleles	Genotypes	Het. allele ratios	Constraints
SNP	DNA	1 or 2	M, H, m	Fixed ratio	–
	CHM	1 or 2	M, m	–	–
SNP in duplication	DNA	1 or 2	M, H	2 different ratios	One DNA H ratio must match CHM ratio
	CHM	1 or 2	M, H	Fixed ratio	
PSV	DNA	2	H	Fixed ratio	Same H ratio in DNA and CHM
	CHM	2	H	Fixed ratio	
MSV	DNA	1 or 2	M, H, m	Variable ratio	–
	CHM	1 or 2	M, H, m	Variable ratio	–

Samples are either homozygous with respect to one allele (M or m) or apparently heterozygous (H). Single-locus SNPs produce consistent homozygous and heterozygous signals in normal individuals, and no heterozygotes in CHMs. For a true SNP present in one copy of a duplicon (SNP in duplicon), one of the alleles is additionally represented at the other duplicon version(s), generating a heterozygote signal in one or more CHM. In normal DNA, these completely lack one homozygote pattern and generate two distinctive heterozygote patterns with different allele ratios. PSVs render heterozygote signals of identical allele ratios in all tested samples. MSVs produce two or more heterozygote types in CHMs, three or more heterozygote types in normal DNA, or both homozygotes combined with at least one type of heterozygote in CHMs.

Table 3 MLPA analysis of 16 MSVs and two single-copy reference sequences

Nearest rs ID	Dup. region	Normalized MLPA ratios (triplicate means) CHM1	CHM2	CHM3	CHM4	CHM5	CHM6	CHM7	CHM8	s.d.	Copy-number variation
–	Unique	–	0.87	1.12	1.11	0.85	0.93	1.03	0.92	0.11	No
–	Unique	0.93	0.89	1.1	1.09	0.93	0.98	1.03	1.06	0.08	No
394595	B	1.16	1.05	0.97	0.63	0.91	1.01	–	1.04	0.18	Yes
2910545	C	1.13	1.01	1.01	1.00	0.94	0.93	0.93	1.04	0.07	No
1057729	D	1.28	1.22	0.85	0.85	0.77	0.86	1.17	1.02	0.2	Yes
2868008	D	1.28	1.17	0.83	0.92	0.73	0.89	–	0.96	0.19	Yes
2868007	D	1.35	1.18	0.89	0.78	0.74	0.93	1.00	1.14	0.21	Yes
2690641	E	1.04	0.94	1.09	1.16	0.88	0.91	–	0.82	0.12	No
505235	F	1.03	1.02	1.04	0.98	0.96	0.96	0.94	1.06	0.04	No
1836885	H	1.01	0.98	0.94	0.96	1.11	0.93	0.92	1.16	0.09	No
964055	I	1.05	1.18	0.95	1.01	1.01	1.18	0.72	–	0.16	Yes
2939843	I	1.04	1.05	0.92	1.11	1.07	0.94	0.85	1.03	0.09	No
2684043	J	1.15	1.1	1.02	1.16	0.79	0.97	0.92	0.89	0.13	No
2740736	J	1.17	1.1	1.21	1.24	0.7	0.82	1.03	0.74	0.22	Yes
2740083	J	1.03	1.1	1.01	1.11	0.91	0.89	0.97	0.98	0.08	No
746659	J	1.37	1.3	1.00	–	0.73	0.83	0.95	0.78	0.25	Yes
296349	K	0.99	1.00	1.12	1.06	0.89	1.02	0.81	1.1	0.1	No
380880	K	0.75	1.26	1.05	0.86	1.00	1.08	0.93	1.08	0.15	Yes

Half of the MSV sequences show substantial evidence of copy-number variation. The remainder, including the two reference sequences, either have a fixed number of sequence copies or have a relative difference below the threshold of detection (s.d. < 0.15 across the eight CHMs).

DNAs, because individual allele ratio information is lost, it will be impossible to identify MSVs. To detect MSVs in routine practice, CHMs or haploid genomes could be included in upstream assay validation routines. Mendelian inheritance tests might assist but will not be effective for MSVs involving intrachromosomal duplicons. Consideration of Hardy-Weinberg equilibrium (HWE) may help, but analysis will not be fool-proof if the 'single allele' and 'two allele' haploid signals for MSVs are consistent with HWE in the overall population. Beyond MSVs, SNPs residing in one copy of a duplicon may also be mis-scored, because the additional signal component from the non-polymorphic duplicon would make one of the two homozygotes appear to be a heterozygote.

How duplicon markers might be scored disregarding heterozygote allele ratio differences (which many methods tend to do) and without using CHMs is an important question. To explore this, we re-examined our total data set, ignoring these two pieces of evidence. This analysis incorrectly indicated an abundance of PSVs in duplicons (**Fig. 2a**; consistent with previous interpretations[3,7,8]), with only half of the apparent SNPs that were truly MSVs deviating from HWE (32 chromosomes; $P < 0.01$). Consistent with this, as of April 2004, four of the MSV markers we report are classified as experimentally validated SNPs with genotype data in dbSNP. Additionally, one PSV is described in current HapMap data, where it is listed as a monomorphic SNP.

In light of these considerations, we reviewed recent genotyping data from our production facility, which uses DASH. We considered almost 800 markers from different studies that used various SNP selection criteria, leaving 45 targets in duplicons. The initial validation (assessing 16–96 control individuals and considering HWE), identified 15 monomorphic single-allele signals and classified the remaining 30 markers as follows: 12 (40%) unique SNPs, 8 (27%) SNPs in one copy of a duplicon, 4 (13%) PSVs and 6 (20%) MSVs. Five of the unique SNPs had been used for production genotyping of 1,600–2,000 individuals, and only after observing several tens of heterozygote-like signals did it become clear that two of these were actually MSVs and another was a SNP in a duplicon. For the two MSVs, if samples that reported two alleles had been scored as heterozygotes (regardless of allele ratios), then the total genotype data were in complete HWE ($P = 0.115$ and 0.357).

In conclusion, our study identifies MSVs as a new form of genome polymorphism. Careful laboratory practice should often recognize MSVs as aberrant markers, and MSVs may underlie the considerable fraction of markers that fail HWE. But some MSVs are probably being interpreted and used as unique SNPs, and HWE will not always identify these, even if large sample numbers are used. More generally, MSVs (or rather duplicon copy-number variation and duplicon gene

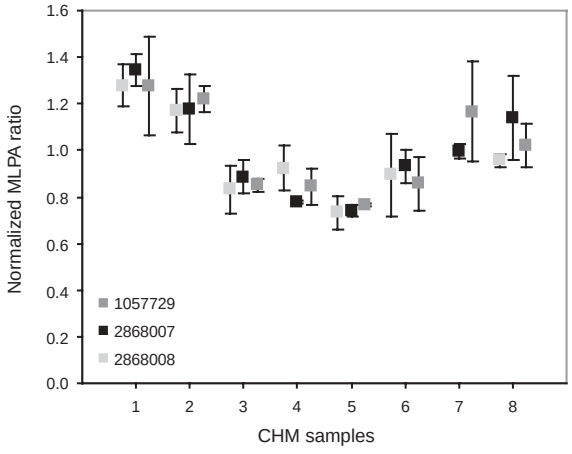

Figure 3 MLPA data for eight CHMs across three consecutive loci. These span 3.4 kb on chromosome 16 (**Table 1**). The graph shows mean ± 2 s.e.m. values across replicate experiments. For all three probes, CHMs 1 and 2 have ratios ~50% higher than those of CHMs 3–6 (a 3:2 relative copy-number difference). CHMs 7 and 8 are harder to classify because of a wider spread between replicates, but they seem to overlap mostly with CHMs 1 and 2. This result is in full agreement with observed genotyping data, in that the MLPA ratios correlate with the observed DASH heterozygote classes.

conversion processes) might underlie some common phenotypic differences between individuals. We therefore suggest that MSVs should be specifically targeted for evaluation in disease and pharmacogenomics research.

METHODS

***In silico* detection of SNP and duplicate region overlap.** Duplicon regions were as previously defined[3], derived from alignments of sequence fragments from the National Center for Biotechnology Information (NCBI) human genome assembly[2] combined with sequence read depth analysis of WSSD from the Celera human genome assembly[1]. We downloaded duplication sequence and June 2002 NCBI assembly locations from the human paralogy database. We used the most complete SNP list available with June 2002 NCBI assembly locations (dbSNP[25] build 112; 2,337,575 SNPs) and updated the annotation with data from dbSNP build 119. We downloaded gene lists from Ensembl[26]. We loaded the locations into a MySQL database and identified overlaps of chromosomal locations through SQL queries issued from a set of Perl scripts. Total counts were nonredundant so that each SNP was counted only once in our analysis, even if it mapped to multiple genome locations (duplicon paralogs).

We searched for any dbSNP annotations that might uniquely characterize duplicon SNPs. We tested the following factors: (i) validation (by cluster, 'SNP discovered by at least two different methods'; by two hit–two allele, 'SNP must be observed twice, in two different DNA samples which must have produced two alleles'; by frequency, 'allele frequency data available for SNP'); (ii) source (which discovery effort generated the SNP); and (iii) frequency of minor allele. Map weight was excluded from consideration, as these SNPs are, by definition, in repetitive sequence, and for any SNP in a duplicon with a map weight <2, the map weight is due to the difference in alignment methods and scoring thresholds between duplicon detection and SNP mapping.

DASH. We carried out DASH experiments, designed with DFold[27] software, using standard protocols as previously described[14]. Oligonucleotide sequences for all assays are available on request. We carried out PCR reactions in 20-μl volumes, containing 25–250 pg μl^{-1} of genomic DNA. We used DASH software (Thermo Hybaid) to visualize denaturation events by plotting the negative derivative of the fluorescence versus temperature profile. Genotypes were scored manually and blindly. We reviewed independent duplicate experiments for 25% of assays as a control for assay reproducibility and found scoring to be consistent across runs. We assessed deviation from HWE for individual markers using the χ^2 statistic ($P < 0.01$). We excluded 32% of assays across all regions from analysis; 3.2% (5 of 157) assays produced no PCR product, and 29% (13 of 45) of those in nonduplicon regions (control regions plus falsely predicted duplicons with support only from the public assembly) and 18% (20 of 112) of those in real duplicons gave no indication of polymorphism. These percentages were evenly distributed between different sources of SNPs (data not shown) and are consistent with what is generally found for public database SNPs[28]. Further, 4.4% (2 of 45) of assays in nonduplicon regions and 8.9% (10 of 112) of those in real duplicons were of low quality, and many gave three distinct allele signals. This is probably due to additional but uncharacterized sequence variants in the probe hybridization region at positions other than that being tested. This left 107 informative polymorphic assays covering all tested regions. Complete genotyping information is available on request.

The number of tested DNA samples affects the certainty of classification. Also, misclassifications may arise if a PCR does not amplify multiple duplicon copies with similar or equal efficiency. We cannot estimate the cumulative size of these biases, but both will tend to cause an overestimation of the number of PSVs at the expense of MSVs and suggest monomorphic sites over SNPs, SNPs over SNPs in duplicons and SNPs in duplicons over MSVs. Therefore, our PSV estimate must be considered a maximum, and our MSV estimate a minimum.

MLPA. We designed MLPA probes based on consensus sequences derived from global alignments of duplicated segments. Probes were localized in regions immediately flanking MSV variants identified by the DASH experiment. To avoid allelic discrimination and ensure specificity, no polymorphism or sequence differences between duplicon copies were allowed within 6 bp on either side of the ligation site (sequences available on request). The specific priming sequences in the 5′ ends of the half-probes allowed multiplex amplification with either the MLPA primers[23] or the MAPH primers[29]. Resulting PCR products had a minimal size difference of 2 bp, with the products ranging in size from 80 bp to 125 bp. The forward primer of each pair was fluorescently labeled (MLPAF-FAM or MAPHF-HEX), allowing probes to be distinguished also on the basis of color. Each color set included three control probes from known single-copy regions, for normalization purposes, and we added two other single-copy probes to one of the sets as controls for copy-number variation. All oligonucleotides were combined in a single mix at a final concentration of 4 fmol μl^{-1}.

We carried out the MLPA reaction essentially as described[23]. We heated 100 ng of DNA at 98 °C for 5 min. After cooling to 25 °C, we added 1.5 μl of probe mix and 1.5 μl of SALSA hybridization buffer to each sample, denatured them at 95 °C for 2 min and then hybridized them for 16 h at 60 °C. Ligation was done at 54 °C by adding 32 μl of ligation mix. After 10–15 min, we stopped the reaction by heat inactivation at 95 °C for 5 min. We carried out PCR amplification for 30 cycles in a final volume of 25 μl. In addition to the reagents described[23], we added MAPH-F and MAPH-R to each PCR reaction to a final concentration of 100 nM. From each PCR reaction, we mixed 1–2 μl of product with 10 μl (Hi Di) of formamide and 0.1 μl of ROX 500 size standard (Applied Biosystems) in a 96-well plate. We separated products by capillary electrophoresis on the ABI 3700 DNA sequencer (Applied Biosystems).

MLPA data analysis. We retrieved peak data using GeneScan (Applied Biosystems) and exported it to Excel (Microsoft) and SPSS 10 (SPSS) for further analysis. We obtained signals for 84% (16 of 19) of designed assays. We obtained a ratio for each of the working probes by dividing the height of the corresponding peak by the sum of the heights of three control peaks of the same color. We did three replicate experiments across all CHM samples, calculated the average value of the three ratios and discarded the results if the s.d. was >20%. This eliminated 6 of 144 measurements (4.2%). We then normalized the data for each probe around 1.0 by dividing by the average of the remaining values.

URLs. The Human Paralogy Server is available at http://humanparalogy.gene.cwru.edu/. The NCBI dbSNP is available at http://www.ncbi.nlm.nih.gov/SNP/. The International HapMap Project is available at http://www.hapmap.org/.

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

ACKNOWLEDGMENTS

We thank R.J. Fisher and M. Seckl for CHM DNA samples and R.A. Clark, S. Sawyer and C. Lagerberg for technical assistance. Funding was provided by Pfizer Corporation and Stiftelsen för Kompetens-och Kunskapsutveckling (to D.F. and A.J.B.) and by the US National Institutes of Health (to E.E.E.).

COMPETING INTERESTS STATEMENT

The authors declare competing financial interests (see the *Nature Genetics* website for details).

Received 23 April; accepted 22 June 2004
Published online at http://www.nature.com/naturegenetics/

1. Venter, J.C. *et al.* The sequence of the human genome. *Science* **291**, 1304–1351 (2001).
2. Lander, E.S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
3. Bailey, J.A. *et al.* Recent segmental duplications in the human genome. *Science* **297**, 1003–1007 (2002).
4. Bailey, J.A., Yavor, A.M., Massa, H.F., Trask, B.J. & Eichler, E.E. Segmental duplications: organization and impact within the current human genome project assembly. *Genome Res.* **11**, 1005–1017 (2001).
5. Istrail, S. *et al.* Whole-genome shotgun assembly and comparison of human genome assemblies. *Proc. Natl. Acad. Sci. USA* **101**, 1916–1921 (2004).
6. Shaw, C.J. & Lupski, J.R. Implications of human genome architecture for rearrangement-based disorders: the genomic basis of disease. *Hum. Mol. Genet.* **13**, R57–R64 (2004).
7. Estivill, X. *et al.* Chromosomal regions containing high-density and ambiguously mapped putative single nucleotide polymorphisms (SNPs) correlate with segmental duplications in the human genome. *Hum. Mol. Genet.* **11**, 1987–1995 (2002).
8. Cheung, J. *et al.* Genome-wide detection of segmental duplications and potential assembly errors in the human genome sequence. *Genome Biol.* **4**, R25 (2003).

9. Sachidanandam, R. *et al.* A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409**, 928–933 (2001).
10. Tsui, C. *et al.* Single nucleotide polymorphisms (SNPs) that map to gaps in the human SNP map. *Nucleic Acids Res.* **31**, 4910–4916 (2003).
11. Hurles, M.E. Gene conversion homogenizes the CMT1A paralogous repeats. *BMC Genomics* **2**, 11 (2001).
12. Hurles, M. Are 100,000 "SNPs" useless? *Science* **298**, 1509 (2002).
13. Conant, G.C. & Wagner, A. Asymmetric sequence divergence of duplicate genes. *Genome Res.* **13**, 2052–2058 (2003).
14. Prince, J.A. *et al.* Robust and accurate single nucleotide polymorphism genotyping by dynamic allele-specific hybridization (DASH): design criteria and assay validation. *Genome Res.* **11**, 152–162. (2001).
15. Sebire, N.J., Fisher, R.A. & Rees, H.C. Histopathological diagnosis of partial and complete hydatidiform mole in the first trimester of pregnancy. *Pediatr. Dev. Pathol.* **6**, 69–77 (2003).
16. Kruglyak, L. & Nickerson, D.A. Variation is the spice of life. *Nat. Genet.* **27**, 234–236 (2001).
17. Smit, A.F. Interspersed repeats and other mementos of transposable elements in mammalian genomes. *Curr. Opin. Genet. Dev.* **9**, 657–663 (1999).
18. Jeffreys, A.J. & May, C.A. Intense and highly localized gene conversion activity in human meiotic crossover hot spots. *Nat. Genet.* **36**, 151–156 (2004).
19. Rozen, S. *et al.* Abundant gene conversion between arms of palindromes in human and ape Y chromosomes. *Nature* **423**, 873–876 (2003).
20. Hollox, E.J., Armour, J.A. & Barber, J.C. Extensive normal copy number variation of a beta-defensin antimicrobial-gene cluster. *Am. J. Hum. Genet.* **73**, 591–600 (2003).
21. Locke, D.P. *et al.* BAC microarray analysis of 15q11-q13 rearrangements and the impact of segmental duplications. *J. Med. Genet.* **41**, 175–182 (2004).
22. White, S.J. *et al.* Two-colour MLPA; detecting genomic rearrangements in hereditary multiple exostoses. *Hum. Mutat.* **24**, 86–92 (2004).
23. Schouten, J.P. *et al.* Relative quantification of 40 nucleic acid sequences by multiplex ligation-dependent probe amplification. *Nucleic Acids Res.* **30**, e57 (2002).
24. Lucito, R. *et al.* Representational oligonucleotide microarray analysis: a high-resolution method to detect genome copy number variation. *Genome Res.* **13**, 2291–2305 (2003).
25. Sherry, S.T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* **29**, 308–311 (2001).
26. Birney, E. *et al.* Ensembl 2004. *Nucleic Acids Res.* **32**, D468–D470 (2004).
27. Fredman, D., Jobs, M., Stromqvist, L. & Brookes, A.J. DFold: PCR design that minimizes secondary structure and optimizes downstream genotyping applications. *Hum. Mutat.* **24**, 1–8 (2004).
28. Carlson, C.S. *et al.* Additional SNPs and linkage-disequilibrium analyses are necessary for whole-genome association studies in humans. *Nat. Genet.* **33**, 518–521 (2003).
29. White, S. *et al.* Comprehensive detection of genomic duplications and deletions in the DMD gene, by use of multiplex amplifiable probe hybridization. *Am. J. Hum. Genet.* **71**, 365–374 (2002).