A Comprehensive Analysis of Common Copy-Number Variations in the Human Genome

Kendy K. Wong,* Ronald J. deLeeuw,* Nirpjit S. Dosanjh, Lindsey R. Kimm, Ze Cheng, Douglas E. Horsman, Calum MacAulay, Raymond T. Ng, Carolyn J. Brown, Evan E. Eichler, and Wan L. Lam

Segmental copy-number variations (CNVs) in the human genome are associated with developmental disorders and susceptibility to diseases. More importantly, CNVs may represent a major genetic component of our phenotypic diversity. In this study, using a whole-genome array comparative genomic hybridization assay, we identified 3,654 autosomal segmental CNVs, 800 of which appeared at a frequency of at least 3%. Of these frequent CNVs, 77% are novel. In the 95 individuals analyzed, the two most diverse genomes differed by at least 9 Mb in size or varied by at least 266 loci in content. Approximately 68% of the 800 polymorphic regions overlap with genes, which may reflect human diversity in senses (smell, hearing, taste, and sight), rhesus phenotype, metabolism, and disease susceptibility. Intriguingly, 14 polymorphic regions harbor 21 of the known human microRNAs, raising the possibility of the contribution of microRNAs to phenotypic diversity in humans. This in-depth survey of CNVs across the human genome provides a valuable baseline for studies involving human genetics.

Genetic variation in the human genome exists in different forms. Recent studies have shown that variations exist in the human genome at various levels: the single base pair,[1] the kilobase pair,[2–4] and tens to thousands of kilobase pairs.[5–8] Extensive studies, including the recently published haplotype map from HapMap,[1] have identified millions of SNPs in the human genome. Three recent studies that used the SNP data each identified several hundred sites of deletion in the human population; however, gains could not be deduced from this data set.[2–4] By use of a fosmid paired-end sequence analysis, a comprehensive comparison between two genomes quantified 241 sites of insertion or deletion.[8] By use of array comparative genomic hybridization (array CGH) techniques, large-scale copy-number variations (CNVs) were demonstrated in a fraction of the human genome.[5,6] Each of these studies added to our knowledge about CNVs in the human population, but with little overlap in findings.[9] Thus, many characteristics of CNVs in the human population remain unknown, such as the total number, genomic positions, gene content, frequency spectrum, and patterns of linkage disequilibrium with one another. Understanding CNVs is critical for the proper study of disease-associated changes because segmental CNVs have been demonstrated in developmental disorders and susceptibility to disease.[10] Therefore, analysis of CNVs at the whole-genome level is required to create a baseline of human genomic variation. In this study, using a whole-genome tiling-path BAC array CGH approach,[11] we measured large scale (>40 kb) segmental gains and losses in >100 individuals to expand our knowledge about CNVs and to estimate the extent of this form of variation in the human population.

Material and Methods

DNA Samples

Samples were collected and were rendered anonymous. These samples included 16 from healthy blood donors, 51 from a British Columbia Cancer Agency (BCCA) screening program, and 26 B-lymphoblast DNA samples encompassing 16 distinct ethnic groups from the Human Variation Collection and 14 CEPH pedigree samples from the Coriell Cell Repository (National Institute of General Medical Sciences, Camden, NJ). The DNA samples from cell lines were included to represent diverse ethnic populations. The 51 samples from the BCCA screening program included 19 from a breast cancer screening program and 32 from a colon cancer screening program. These were constitutional DNA samples obtained from blood that did not contain any neoplastic cells, and none showed CNV association with *BRCA1* (MIM 113705), *BRCA2* (MIM 600185), *APC* (MIM 175100), *MSH2* (MIM 609309), or *MSH6* (MIM 600678). Only 2 of the 32 samples from the colon cancer screening program showed CNV association with *MLH1* (MIM 120436). In addition, no CNVs were associated with a specific sample type or source, which suggests no obvious selection bias. In total, 105 DNA samples (from 44 males and 61 females) were included in this study (table 1), 95 of which were used for CNV discovery. DNA from the four grandparents of the CEPH pedigree were included in the CNV discovery sample set, whereas DNA from 10 other members of the family were included only for clustering and inheritance analysis. A donor sample was used as the male reference, and a single female sample was used

From the Departments of Cancer Genetics and Developmental Biology (K.K.W.; R.J.d.; N.S.D.; L.R.K.; W.L.L.) and Cancer Imaging (C.M.), British Columbia Cancer Research Centre, Department of Pathology, British Columbia Cancer Agency (D.E.H.), and Departments of Computer Science (R.T.N.) and Medical Genetics (C.J.B.), University of British Columbia, Vancouver; and Department of Genome Sciences, University of Washington School of Medicine, and Howard Hughes Medical Institute, Seattle (Z.C.; E.E.E.)

Received August 1, 2006; accepted for publication November 2, 2006; electronically published December 5, 2006.

Address for correspondence and reprints: Dr. Kendy K. Wong, 675 West 10th Avenue, Vancouver, BC, V5Z 1L3, Canada. E-mail: kwong@bccrc.ca

* These two authors contributed equally to this work.

Am. J. Hum. Genet. 2007;80:91–104. 0002-9297/2007/8001-0010$15.00

Table 1. Samples Used in This Study

Sample	Sample Source[a]	Sex
S1	Coriell (NA17755), Han of L.A.	M
S2	Coriell (NA10975), Mayan	M
S3	Coriell (NA17392), Mexican Indian	M
S4	Coriell (NA17075), Puerto Rican	M
S5	Coriell (NA15724), Czechoslovakian	M
S6	Coriell (NA15760), Iceland	M
S7	Coriell (NA17384), African North of Sahara	M
S8	Coriell (NA10469), Biaka	M
S9	Coriell (NA10492), Mbuti	M
S10	Coriell (NA17361), Ashkenazi Jewish	M
S11	Coriell (NA11522), Druze	M
S12	Coriell (NA13613), Taiwan Ami tribe	M
S13	Coriell (NA13611), Taiwan Ami tribe	M
S14	Coriell (NA13603), Taiwan Atayal tribe	M
S15	Coriell (NA13606), Taiwan Atayal tribe	M
S16	Coriell (NA11587), Japanese	M
S17	Coriell (NA10540), Melanesian	M
S18	Screening program, ethnicity unknown	M
S19	Screening program, ethnicity unknown	M
S20	Screening program, ethnicity unknown	M
S21	Screening program, ethnicity unknown	M
S22	Screening program, ethnicity unknown	M
S23	Screening program, ethnicity unknown	M
S24	Screening program, ethnicity unknown	M
S25	Screening program, ethnicity unknown	M
S26	Screening program, ethnicity unknown	M
S27	Screening program, ethnicity unknown	M
S28	Screening program, ethnicity unknown	M
S29	Screening program, ethnicity unknown	M
S30	Screening program, ethnicity unknown	M
S31	Screening program, ethnicity unknown	M
S32	Screening program, ethnicity unknown	M
S33	Donor, ethnicity unknown	M
S34	Donor, ethnicity unknown	M
S35	Donor, ethnicity unknown	M
S36	Donor, ethnicity unknown	M
S37	Coriell (NA17766), Han of Los Angeles	F
S38	Coriell (NA17076), Puerto Rican	F
S39	Coriell (NA15729), Czechoslovakian	F
S40	Coriell (NA15766), Icelandic	F
S41	Coriell (NA17348), African South of Sahara	F
S42	Coriell (NA10471), Biaka	F
S43	Coriell (NA11521), Druze	F
S44	Coriell (NA10539), Melanesian	F
S45	Screening program, ethnicity unknown	F
S46	Screening program, ethnicity unknown	F
S47	Screening program, ethnicity unknown	F
S48	Screening program, ethnicity unknown	F
S49	Screening program, ethnicity unknown	F
S50	Screening program, ethnicity unknown	F
S51	Screening program, ethnicity unknown	F
S52	Screening program, ethnicity unknown	F
S53	Screening program, ethnicity unknown	F
S54	Screening program, ethnicity unknown	F
S55	Screening program, ethnicity unknown	F
S56	Screening program, ethnicity unknown	F
S57	Screening program, ethnicity unknown	F
S58	Screening program, ethnicity unknown	F
S59	Screening program, ethnicity unknown	F
S60	Screening program, ethnicity unknown	F
S61	Screening program, ethnicity unknown	F
S62	Screening program, ethnicity unknown	F
S63	Screening program, ethnicity unknown	F
S64	Screening program, ethnicity unknown	F
S65	Screening program, ethnicity unknown	F
S66	Screening program, ethnicity unknown	F
S67	Screening program, ethnicity unknown	F
S68	Donor, ethnicity unknown	F
S69	Donor, ethnicity unknown	F
S70	Donor, ethnicity unknown	F
S71	Donor, ethnicity unknown	F
S72	Donor, ethnicity unknown	F
S73	Donor, ethnicity unknown	F
S74	Donor, ethnicity unknown	M
S75	Screening program, ethnicity unknown	F
S76	Screening program, ethnicity unknown	F
S77	Coriell (NA17393), Mexican Indian	F
S78	Donor, ethnicity unknown	F
S79	Donor, ethnicity unknown	F
S80	Donor, ethnicity unknown	F
S81	Screening program, ethnicity unknown	F
S82	Screening program, ethnicity unknown	F
S83	Screening program, ethnicity unknown	F
S84	Screening program, ethnicity unknown	F
S85	Screening program, ethnicity unknown	F
S86	Screening program, ethnicity unknown	F
S87	Screening program, ethnicity unknown	F
S88	Screening program, ethnicity unknown	F
S89	Screening program, ethnicity unknown	F
S90	Screening program, ethnicity unknown	F
S91	Screening program, ethnicity unknown	F
F1	Coriell (NA11917, paternal grandfather), Utah	M
F2	Coriell (NA11918, paternal grandmother), Utah	F
F3	Coriell (NA11919, maternal grandfather), Utah	M
F4	Coriell (NA11920, maternal grandmother), Utah	F
F5[b]	Coriell (NA10842, dad), Utah	M
F6[b]	Coriell (NA10843, mom), Utah	F
F7[b]	Coriell (NA11909, son), Utah	M
F8[b]	Coriell (NA11910, daughter), Utah	F
F9[b]	Coriell (NA11911, daughter), Utah	F
F10[b]	Coriell (NA11912, son), Utah	M
F11[b]	Coriell (NA11913, son), Utah	M
F12[b]	Coriell (NA11915, daughter), Utah	F
F13[b]	Coriell (NA11916, son), Utah	M
F14[b]	Coriell (NA11921, daughter), Utah	F

[a] Coriell Cell Repository (Coriell) sample numbers are shown in parentheses.

[b] These 10 CEPH family samples were not included in the CNV discovery set of 95.

only in control experiments. Genomic DNA from donors was extracted from whole blood by use of the QIAamp DNA Blood Maxi Kit (QIAGEN) and was quantified by spectrophotometry (ND-1000 [NanoDrop]).

BAC Array CGH Analysis

DNA labeling and hybridization was performed as described elsewhere,[11] with slight modifications. In brief, 200 ng of sample and reference DNA were differentially labeled with Cyanine 3–dCTP and Cyanine 5–dCTP (Perkin Elmer Life Sciences). The random priming reaction was incubated in the dark at 37°C for 16–18 h. DNA samples were then combined, and unincorporated nucleotides were removed using microcon YM-30 columns (Millipore). Purified samples were mixed with 100 μg of human Cot-1 DNA

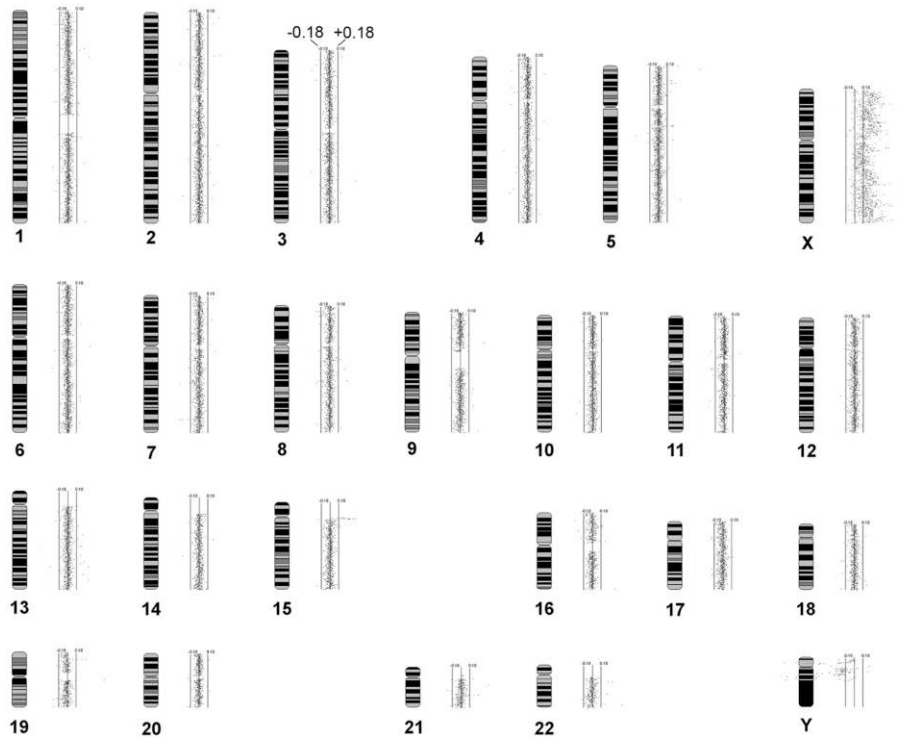

Figure 1. Example of a karyogram from a hybridization experiment in this study. Custom SeeGH software was used to visualize normalized data as $\log_2$ ratio plots.[13] The figure illustrates an example of a hybridization of a female sample versus the male reference. The $\log_2$ ratios of the data are shown as dots; the left and right vertical lines represent threshold lines for this experiment at $\log_2$ ratios of −0.18 and 0.18, respectively.

(Invitrogen) and were precipitated. DNA pellets were resuspended in 45 μl of DIG Easy hybridization solution (Roche) containing 20 mg/ml sheared herring sperm DNA and 10 mg/ml yeast tRNA. Sample mixture was denatured at 85°C for 10 min, and repetitive sequences were blocked at 45°C for 1 h before hybridization. The mixture was then applied onto BAC arrays containing 26,363 clones spotted in duplicate (53,856 elements with controls) on single slides. (These clones were selected from the SMRT clone set, to optimize tiling coverage of the genome; the clone list is available at the SMRT Array Web site.[11]) Hybridization was performed in the dark at 45°C for ~36 h inside a hybridization chamber, followed by washing three times for 3 min each with agitation in 0.1 × saline sodium citrate (SSC) and 0.1% SDS at 45°C. Arrays were then rinsed three times for 3 min each in 0.1 × SSC at room temperature and were dried by an air stream before imaging. Slides were scanned using a charge-coupled device–based imaging system (arrayWoRx eAuto [Applied Precision]) and were analyzed with the SoftWoRx Tracker Spot Analysis software (Applied Precision). The $\log_2$ ratios of the Cyanine 3 to Cyanine 5 intensities for each spot were assessed. To remove systematic effects from nonbiological sources that introduce bias, the ratios were then normalized using a stepwise normalization technique.[12] Custom SeeGH software was used to visualize normalized data as $\log_2$ ratio plots (fig. 1).[13]

CNV-Detection Algorithm

For each experiment, 1,398 clones from chromosomes X and Y were removed, and the remaining data were median normalized to remove bias introduced because of any sex-mismatched hybridization. In addition, 573 clones were removed from analysis because of printing anomalies or their shift in $\log_2$ ratios, possibly due to homology with the X or Y chromosome, leaving a total of 24,392 reliable clones for analysis (see the tab-delimited ASCII file, which can be imported into a spreadsheet, of data set 1 [online only]). Experimental SDs ($SD_{autosome}$) were calculated for each experiment on the basis of the $\log_2$ ratios of the 24,392 reliable clones minus the clones removed because of low signal-to-noise ratio (SNR) or high SD of replicate clone measures (SD_{clone}). Thresholds for determining CNV clones were set at a multiple of the $SD_{autosome}$ value. For each experiment, clones were annotated as uninformative if they were filtered via SNR or SD_{clone}, as a CNV loss if the $\log_2$ ratio was less than the negative threshold, as unchanged if the $\log_2$ ratio was between the negative and positive thresholds, and as a CNV gain if the $\log_2$ ratio was above the positive threshold.

To determine the optimal values for SNR, SD_{clone}, and the $SD_{autosome}$ multiplier, eight hybridization experiments (four repeat experiments of male reference versus the single female DNA and four experiments between those two DNAs and two additional DNA pools) were used. On the basis of the possible combinations of copy-number status in the four DNA samples used, we determined the expected CNV patterns in the eight hybridization experiments (table 2). The three parameters were recursively varied until the highest proportion of CNV clones match the expected patterns (table 2); this resulted in the filter settings of $SD_{clone} > 0.15$, $SNR < 3$, and a stringent $SD_{autosome}$ multiplier of 3.3 × . On

Table 2. Expected CNV Patterns of Eight Hybridizations between Four DNA Samples

CNV Combinations[a]				Expected CNV Patterns[b]							
MR	FS	MP	FP	MR vs. FS	MR vs. FS	MR vs. FS	MR vs. FS	MR vs. MP	FS vs. MP	MR vs. FP	FS vs. FP
Loss	Loss	Loss								−	−
Loss	Loss	Loss	Gain							−	−
Loss	Loss		Loss					−	−		
Loss	Loss							−	−	−	−
Loss	Loss		Gain					−	−	−	−
Loss	Loss	Gain	Loss					−	−		
Loss	Loss	Gain						−	−	−	−
Loss	Loss	Gain	Gain					−	−	−	−
Loss		Loss	Loss	−	−	−	−		+		+
Loss		Loss		−	−	−	−		+	−	
Loss		Loss	Gain	−	−	−	−		+	−	−
Loss			Loss	−	−	−	−	−			+
Loss				−	−	−	−	−		−	
Loss			Gain	−	−	−	−	−		−	−
Loss		Gain	Loss	−	−	−	−	−	−		+
Loss		Gain		−	−	−	−	−	−	−	
Loss		Gain	Gain	−	−	−	−	−	−	−	−
Loss	Gain	Loss	Loss	−	−	−	−		+		+
Loss	Gain	Loss		−	−	−	−		+	−	+
Loss	Gain	Loss	Gain	−	−	−	−		+	−	
Loss	Gain		Loss	−	−	−	−	−	+		+
Loss	Gain			−	−	−	−	−	+	−	+
Loss	Gain		Gain	−	−	−	−	−	+	−	
Loss	Gain	Gain	Loss	−	−	−	−	−			+
Loss	Gain	Gain		−	−	−	−	−		−	+
Loss	Gain	Gain	Gain	−	−	−	−	−		−	
	Loss	Loss	Loss	+	+	+	+	+		+	
	Loss	Loss		+	+	+	+	+			−
	Loss	Loss	Gain	+	+	+	+	+		−	−
	Loss		Loss	+	+	+	+		−	+	
	Loss			+	+	+	+		−		−
	Loss		Gain	+	+	+	+		−	−	−
	Loss	Gain	Loss	+	+	+	+	+	−	+	
	Loss	Gain		+	+	+	+	+	−		−
	Loss	Gain	Gain	+	+	+	+	+	−	−	−
		Loss	Loss					+	+	+	+
		Loss						+	+		
		Loss	Gain					+	+	−	−
			Loss							+	+
			Gain							−	−
		Gain	Loss					−	−	+	+
		Gain						−	−		
		Gain	Gain					−	−	−	−
	Gain	Loss	Loss	−	−	−	−	+	+	+	+
	Gain	Loss		−	−	−	−	+	+		+
	Gain	Loss	Gain	−	−	−	−	+	+	−	
	Gain		Loss	−	−	−	−		+	+	+
	Gain			−	−	−	−		+		+
	Gain		Gain	−	−	−	−		+	−	
	Gain	Gain	Loss	−	−	−	−	−		+	+
	Gain	Gain		−	−	−	−	−			+
	Gain	Gain	Gain	−	−	−	−	−		−	
Gain	Loss	Loss	Loss	+	+	+	+	+		+	
Gain	Loss	Loss		+	+	+	+	+		+	−
Gain	Loss	Loss	Gain	+	+	+	+	+			−
Gain	Loss		Loss	+	+	+	+	+	−	+	
Gain	Loss			+	+	+	+	+	−	+	−
Gain	Loss		Gain	+	+	+	+	+	−		−
Gain	Loss	Gain	Loss	+	+	+	+		−	+	
Gain	Loss	Gain		+	+	+	+		−	+	−
Gain	Loss	Gain	Gain	+	+	+	+		−		−
Gain		Loss	Loss	+	+	+	+	+	+	+	+
Gain		Loss		+	+	+	+	+	+	+	
Gain		Loss	Gain	+	+	+	+	+	+		−
Gain			Loss	+	+	+	+	+		+	+
Gain				+	+	+	+	+		+	
Gain			Gain	+	+	+	+	+			−
Gain		Gain	Loss	+	+	+	+		−	+	+
Gain		Gain		+	+	+	+		−	+	
Gain		Gain	Gain	+	+	+	+		−		−
Gain	Gain	Loss	Loss					+	+	+	+
Gain	Gain	Loss						+	+	+	+
Gain	Gain	Loss	Gain					+	+		
Gain	Gain		Loss					+	+	+	+
Gain	Gain							+	+	+	+
Gain	Gain		Gain					+	+		
Gain	Gain	Gain	Loss							+	+
Gain	Gain	Gain								+	+

NOTE.—FP = female pool; FS = single female sample; MP = male pool; MR = single male reference.

[a] Possible combinations of copy-number status in the four DNA samples. Blank cells indicate no copy-number change. Gain = copy-number gain; Loss = copy-number loss.

[b] Expected CNV patterns of eight hybridizations between the four DNA samples. Observed experimental data were compared against these expected patterns. In each hybridization, the first sample is expected to have a net gain in copy-number (+), a net loss in copy-number (−), or the same copy number (blank cell) as the second sample for a CNV with the particular combination of copy-number status shown on the left.

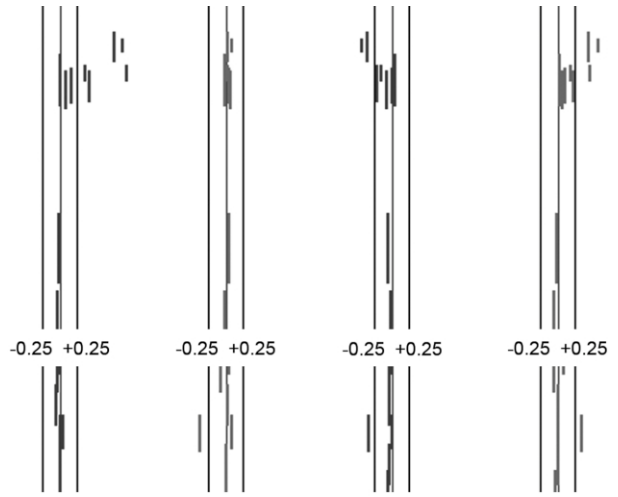

Figure 2. Detection of CNVs. The upper part illustrates a region of CNV at 19p13.2 among four individuals. Each short line represents the average fluorescent intensity ratio between sample and reference DNA for an individual BAC clone spotted on the array. The left and right vertical lines represent the average threshold for the hybridizations shown, at $\log_2$ ratios of -0.25 and 0.25. A ratio to the right of the positive threshold line represents a copy-number gain, whereas a ratio to the left of the negative threshold represents a copy-number loss. Equal, greater, and fewer copies relative to the reference DNA are shown. The lower part illustrates a single BAC clone CNV at 7q32.1 among the four individuals; the clone (RP11-636E12) overlaps with the *IMPDH1* gene, a mutation in which was shown to cause retinitis pigmentosa.

the basis of six self-versus-self hybridizations to calibrate array performance, experiments with >10% uninformative data points or with an $SD_{autosome} > 0.12$ were repeated. Normalized $\log_2$ ratio profiles were generated for the 105 individuals from hybridization of sample DNA versus a single male reference DNA. Data points that did not meet our SD_{clone} or SNR criteria were annotated as uninformative, whereas those whose average ratio exceeded the $3.3 \times SD_{autosome}$ were identified as CNV clones (see the tab-delimited ASCII file of data set 2 [online only]). CNV clones that overlapped in genomic coverage were considered to represent the same CNV loci. A custom track file for uploading the identified CNV clones to the University of California–Santa Cruz (UCSC) Human Genome Browser is available on request. After submission of the custom track file, clones displayed in blue, red, green, and black represented CNVs seen once or twice, three times, four or five times, and six or more times, respectively.

Determination of False-Positive and False-Negative Rates

To estimate our false-positive and false-negative rates in this study, six repeat experiments (of the single female vs. the male reference) were analyzed per our CNV algorithm (see above). In total, 803 CNV calls were made, with 340 seen only once, 50 twice, 46 three times, 15 four times, 15 five times, and 15 six times. Given that our false-positive results cannot exceed the total number of calls (i.e., 803), our maximum false-positive rate is 0.5487% (803/24,392 measures × 6 experiments). By use of this maximum false-positive rate of 0.5487%, the binomial probability, p, of detecting the same clone twice within six experiments by random chance is $p = 0.000445$. Therefore, we concluded that any clone detected twice or more was a true CNV in these six repeat experiments. In theory, we expected to detect 141 true CNVs (i.e., 50 calls seen twice, 46 seen three times, and 15 each seen four, five, and six times) in each of the six experiments (846 calls). In practice, 463 were detected, yielding an estimated false-negative rate of 45.3%. Although statistically a fraction of the single-occurrence calls (those seen only once) represent true CNVs, we conservatively considered all 340 as false-positive results, resulting in a false-positive rate of 0.2323% (340 calls/24,392 measures × 6 experiments). In short, we tolerated this high false-negative rate of 45.3% to achieve our very low false-positive rate for confidence in CNV discovery.

On the basis of the false-positive and false-negative rates calculated above, in a repeat of the same hybridization experiment, one would expect to see 134 calls (803 calls/6 experiments), of which 57 would be false-positive results (0.2323% × 24,392 measures) and 77 would be true CNVs. On the basis of our false-negative rate, we would have missed 64 true CNVs (of 141 true CNVs). Therefore, of a total of 141 true CNVs, the probability of obtaining the same true CNVs in a repeat hybridization should be 54.7% (77 of 141), and the probability of seeing those same CNVs in a second repeat hybridization would be 54.7% × 54.7% (42 of the 141 true CNVs). This represents 84 calls (2 × 42 CNVs) of the 268 expected total calls (134 × 2) (a 31.3% overlap). To verify our calculated rates, three repeat hybridization experiments were performed using the same samples. The observed overlaps of CNV calls between the three possible comparisons were 31.3%, 28.6%, and 31.2%, which is in complete agreement with the expected value. The above calculations are summarized in figure A1 (online only). Additionally, 20 samples (F1, F2, F3, S1, S3, S4, S7, S8, S10, S11, S12, S14, S16, S17, S33, S38, S39, S40, S41, and S44) from the discovery set were each repeated once with a fluorochrome reversal. The overlapping calls between repeats ranged from 21% to 46%, with an average of 30%, again consistent with the expected value from our false-positive and false-negative rates.

Furthermore, we employed an additional platform to verify our CNV calls. We recognize that oligonucleotide arrays are generally not designed for measuring CNVs in certain loci, since many

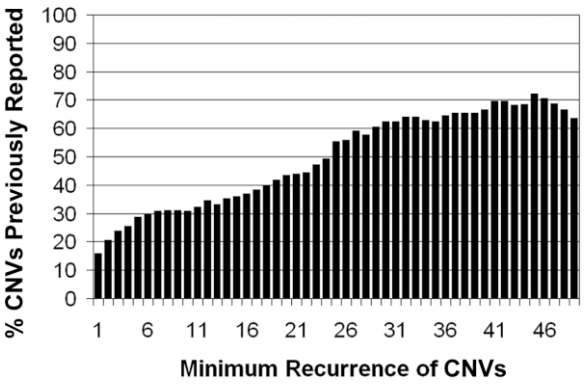

Figure 3. Distribution of overlapped CNVs at different recurrence levels. The percentage of our CNV loci that overlapped with previously reported CNVs were plotted against minimum recurrence levels of CNVs from 1 to 50 within our sample set of 95.

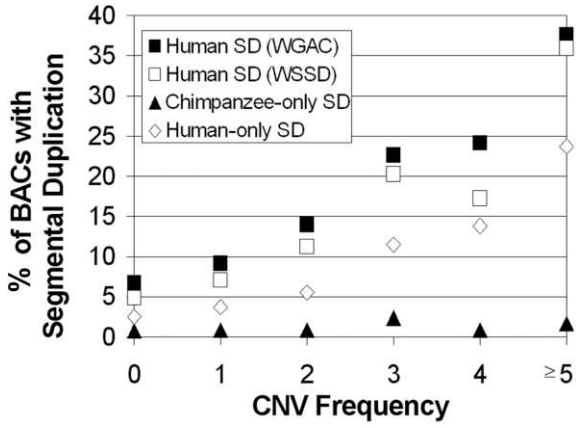

Figure 4. Overlap of CNVs with segmental duplications (SD). The percentage of BACs that contain segmental duplications (>10 kb) is graphed against the frequency of the CNV (0 = no variation) for two measures of human segmental duplication (WSSD and WGAC; see the "Material and Methods" section). Segmental duplications unique to human or chimpanzee are further distinguished.[19]

segmental duplications and repeat sequences are excluded from array design, and thus we constructed a custom oligonucleotide array (NimbleGen Systems) covering our 3,654 CNV loci with 389,027 elements (~2 kb spacing between elements). Five samples (S70, S71, S72, S73, and S80) were assayed using this custom platform. Each of these DNA samples were hybridized against the same single male reference DNA used for BAC array analysis onto the oligonucleotide array. As described elsewhere,[14] to identify gains or losses from the oligonucleotide array, thresholds of 2 SDs of the mean $\log_2$ ratio for all elements in the hybridization were used. On the basis of the detection sensitivity of BAC array CGH,[15] a moving window size of 19 elements (for a total of ~40 kb, with ~2 kb spacing between elements) was applied. In each window, the number of elements reporting a loss (beyond the threshold) was subtracted from the number of elements reporting a gain. The difference was then divided by 19—the total number of elements in the window. Gains or losses were scored for results at >0.1 or <−0.1, respectively. Calls from the oligonucleotide array were then directly compared with CNVs detected by BAC array analysis. To confirm a BAC CNV gain (or loss), at least 10 gains (or losses) were required from the oligonucleotide probe calls covering the same BAC.

CNV Association

To obtain the genomic loci of our identified copy-number–altered clones, we used UCSC May 2004 mapping annotations from BACPAC Resources. For comparison, locations of previously identified CNVs obtained from the Database of Genomic Variants and from various publications were also anchored to the UCSC May 2004 assembly (from UCSC Genome Bioinformatics).[2–4] These were then converted to elements (i.e., clones) within our clone set by comparison of chromosome number, base-pair start position, and base-pair end position.

RefSeq gene information was downloaded from the UCSC May 2004 assembly and was viewed in relation to our CNVs. A gene with any overlap across a CNV boundary was considered to be associated with the CNV. Genes overlapping our CNVs were then used to match genes downloaded from the Online Mendelian Inheritance in Man (OMIM) Morbid Map. The locations of human microRNAs were downloaded from the Sanger miRBase database, were converted to the UCSC May 2004 mapping annotations, and were viewed in relation to our CNVs as described above.[16]

Duplication Analysis

BAC clones and segmental duplication data were mapped to the UCSC May 2004 assembly. CNV loci were assessed for duplication content on the basis of whole-genome assembly comparison (WGAC) and whole-genome shotgun sequence detection (WSSD) analyses of human and chimpanzee genome assemblies.[17–20] We required >10 kb of duplicated sequence to consider a BAC as duplicated. Lineage-specific duplications were distinguished on the basis of human and chimpanzee-only comparisons,[19] available at the Segmental Duplication Database.

Clustering Analysis

A total of 105 individuals were clustered on the basis of our CNV clones, including 14 members of a CEPH pedigree: 4 grandparents (already part of our 95-sample CNV discovery set), 2 parents, and 8 offspring. All clones with copy-number gains and losses were annotated as +1 and −1, respectively. Uninformative measures were left blank, whereas the remaining cells were annotated as 0. Hierarchical clustering of the samples with single linkage was performed using Cluster and was visualized using Treeview[21] (Eisen Lab: Software Web site).

Sample Diversity

The diversity between every possible pair of individuals was calculated by enumerating the number of CNVs (observed at least three times among the 95 samples) with differing status. The pair with the largest value was taken to be the most diverse.

Variation in genome size was determined by first enumerating the net gain or net loss of clones (observed at least three times among the 95 samples) within each individual compared with our reference. The maximum variation was calculated by adding the lowest net loss and the highest net gain. To convert this difference in net clones to genomic size, the number of clones was multiplied by the minimum detection sensitivity of BAC array technology, previously shown to be 40 kb for the average-sized BAC clone.[15]

Quantitative PCR

The iQ SYBR Green Supermix system (Bio-Rad) was used for quantitative PCR (qPCR). Primers were designed using Primer3,[22] and the primers tested are summarized in the tab-delimited ASCII file of data set 3 (online only). In brief, 10 ng genomic DNA was used in a 25-μl reaction with a test or reference primer pair at 600 nM. Reactions were performed in triplicate and were repeated on different days by use of a Bio-Rad iCycler Optical Module (at 95°C for 10 min, then 40 cycles at 95°C for 15 s and 60°C for 1 min, followed by final extension 55°C for 1 min and a melting-curve analysis). Standard curves for each primer pair were generated using a 10-fold dilution series ranging from 0.1 ng to 100 ng. Data analysis was performed as described by Weksberg et al.[23]

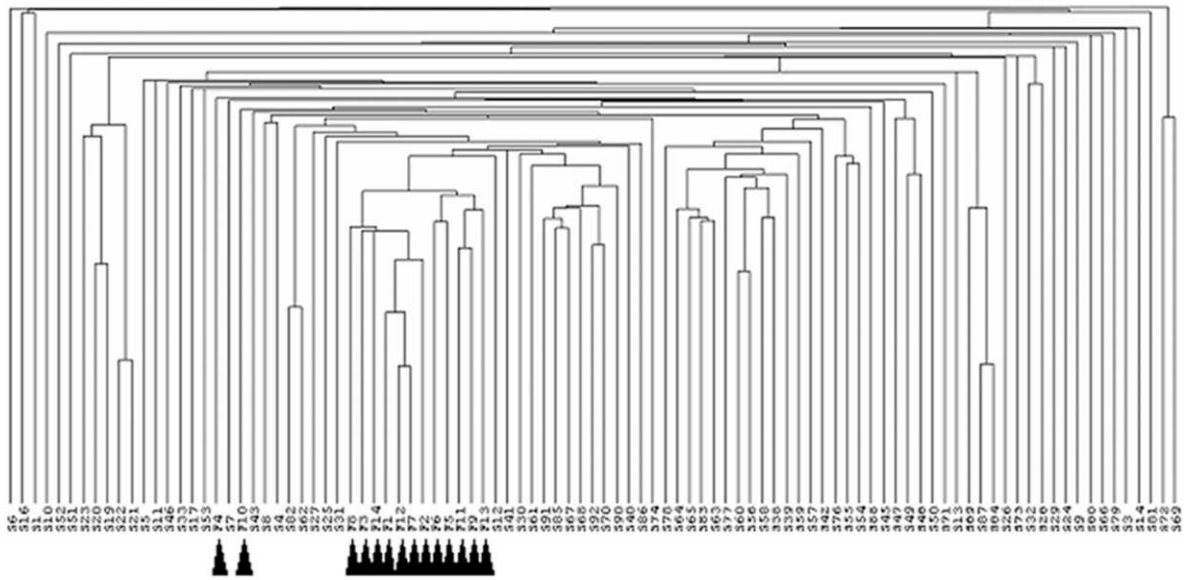

Figure 5. Cluster analysis by use of a CEPH pedigree. Clustering of 105 individuals was based on the high-frequency CNV clones. The 14 CEPH pedigree members are indicated by triangles.

Results

Identification of CNVs

By application of a whole-genome tiling-path BAC array CGH technique, pairwise comparison of DNA samples from 95 unrelated individuals against a single reference DNA sample identified a total of 14,711 CNV BAC clones, averaging 155 per individual (array CGH data for all hybridization experiments have been made publicly available at the Gene Expression Omnibus [series accession number GSE5442]). This resulted in 5,132 unique clones that span 3,654 loci throughout the mapped autosomes (fig. 2 and the tab-delimited ASCII file of data set 2 [online only]). To determine a confidence level for our CNVs, we first calculated the probability of an event occurring repeatedly within our sample set. On the basis of our false-positive rate of 0.23%, calculated from repeat hybridization experiments, the probability of a random false-positive event occurring twice or three times by chance within our sample set of 95 was calculated ($p = 0.02089$ and $p = 0.001479$, respectively). A detailed description of the false-positive rate calculation is given in the "Material and Methods" section.

Second, we examined the amount of overlap with previously reported CNVs[2–8,24] (fig. 3). To facilitate the comparison of our CNVs with previously reported CNVs, the locations of all published CNVs were anchored to the same human genome assembly and were mapped to elements in our clone set. As the minimum recurrence of our CNVs increased, so did the proportion that overlapped with previously reported CNVs (fig. 3). Below a recurrence of 3, little overlap was seen between our study and previous studies. This is likely because of false-positive events or very rare CNVs. Between recurrences of 3 and 30, a steadily increasing overlap with previous studies was observed. This may reflect that the more frequent the CNV in the population, the more likely it will be observed in any given study. Beyond a recurrence of 30, no significant increase in overlap was observed. This may reflect the differences in the composition of each study's population.

Twenty of the 95 experiments were repeated using fluorochrome reversal. In both the original and the repeat experiments, 771 CNV calls were observed. Of the repeated calls, 81% appeared at least three times in the original CNV discovery sample set of 95. This observation increased confidence for CNVs that were detected three or more times within our sample set. qPCR was performed as a quality check on a small number of loci but was not used for large-scale validation because of the limited throughput of single-locus analysis (see the tab-delimited ASCII file of data set 3 [online only]). For further verification of our calls, five separate hybridizations were repeated using a custom-designed oligonucleotide array covering our 3,654 loci with 389,027 elements (~2 kb spacing between elements) (see the "Material and Methods" section). In the five experiments, 265 CNV calls were confirmed by the oligonucleotide array analysis. Of these CNV calls, 83% were among CNVs detected three or more times in the original CNV discovery set of 95.

We next assessed whether our CNVs coincided with segmental duplications in the genome. To achieve this, we evaluated the segmental-duplication content of the CNVs detected in this study, comparing it against both human and chimpanzee sequences, since there is a significant correlation between contemporary human genome structural variation and historical segmental duplications[6–8] (fig. 4). As the frequency of the CNV increased, so did the en-

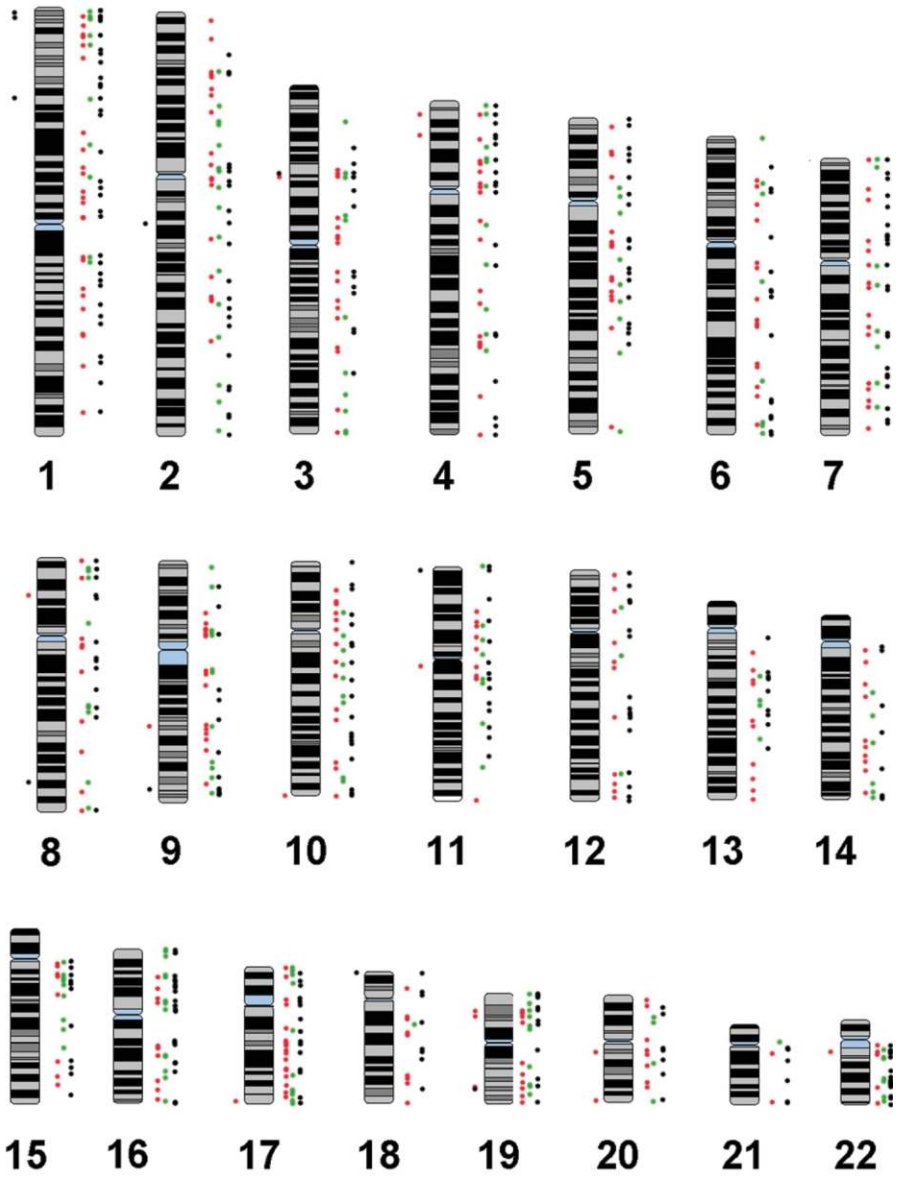

Figure 6. Distribution of CNV clones. High-frequency CNV clones are shown as dots to the right of each chromosome; red, green, and black dots represent presence in three, four or five, and six or more individuals, respectively. Dots to the left of the chromosomes represent locations of CNVs that overlap microRNAs (*red dots*) and select cancer genes (*black dots*).

richment with segmental duplication. This trend increased confidence for CNVs that were observed three or more times in this sample set. We calculated a 5.7-fold duplication enrichment for the most common variants ($\geq$5 occurrences in the 95 individuals), which is similar to previous estimates.[7,8] Interestingly, the effect was most dramatic (a 12.1-fold increase) for duplications that arose specifically within human.[19] In contrast, no enrichment was observed among chimpanzee-only segmental duplications (fig. 4). Elsewhere, we reported an apparent asymmetry with respect to deletion and de novo duplication; 65% of duplications found only in chimpanzee appeared to arise as the result of de novo duplication in the human lineage, as opposed to deletion of a shared duplication in a common ancestor of human and chimpanzee.[19] As a result, chimpanzee-only duplications were not expected to be polymorphic in the human lineage.

We also used clustering analysis to assess our CNV calls. We identified the CNVs present within a CEPH family. Clustering of these samples in combination with our original data set samples showed clear grouping of the CEPH family (fig. 5).

The results from the multiple approaches described above collectively support the presence of novel CNV loci. In addition, the overlaps with previously reported CNVs and segmental duplications, the repeated CNV calls from

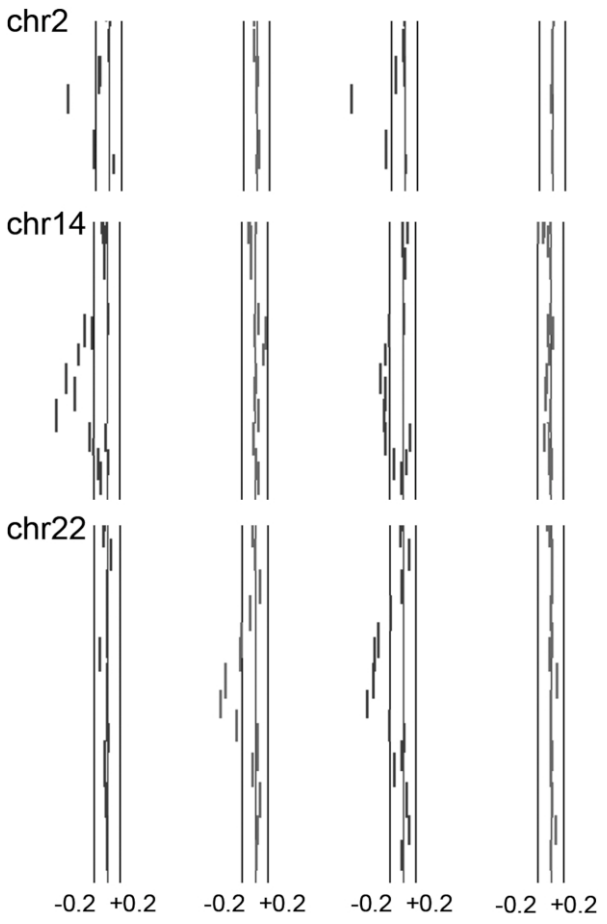

Figure 7. Detection of immunoglobulin variations. The three parts illustrate expected CNVs associated with the immunoglobulin loci at 2p11.2, 14q32.33, and 22q11.22 (*top, middle,* and *bottom,* respectively). The left and right vertical lines represent the average threshold for the hybridizations shown, at $\log_2$ ratios of −0.2 and 0.2. An equal intensity ratio falls on the middle line ($\log_2$ ratio of 0), a ratio to the right of the positive threshold line represents a copy-number gain, and a ratio to the left of the negative threshold represents a copy-number loss. chr = Chromosome.

replicate BAC array CGH experiments and oligonucleotide array hybridizations, the clustering of related individuals on the basis of their CNVs, and the qPCR verification of CNV loci sampled further support their existence. However, the prevalence of these CNVs in the human population can be confirmed only by their presence in multiple individuals. We placed the highest level of confidence in their prevalence when multiple occurrences were observed—for example, 800 loci appeared three or more times in our sample set of 95 individuals. We do not rule out the possibility that true CNVs exist among the loci that we observed at only single and double occurrences in our sample set, since they may represent infrequent events, and a larger sample size will be required to confirm their frequency in the population.

We focused on the high-frequency CNVs (i.e., those found in at least 3 of 95 individuals) for further analysis. There were a total of 9,848 high-frequency CNVs annotated in the 95 individuals analyzed, averaging ~104 per individual. These represent 800 unique loci in the human genome (fig. 6). Strikingly, when these 800 loci are compared with known CNVs, 23% overlap with previously reported CNVs and 77% are novel. The genomic distribution of the 800 CNVs showed no apparent correlation with GC content, imprinted regions, recombination rates, or gene density. Nonrandom somatic alterations—such as the three CNVs associated with immunoglobulin gene rearrangement at chromosomal subbands 2p11.2, 14q32.33, and 22q11.22 (fig. 7)—were detected and removed from further analysis, whereas random somatic alterations not reflecting germline status are not expected to appear recurrently.

Genomic Diversity within the Sample Population

We next examined the genomic diversity within our sample set. The 800 high-frequency CNV loci (or 1,005 BAC clones) were calculated to span a minimum of 40 Mb of DNA (calculated on the basis of BAC array CGH minimum detection sensitivity of 40 kb per clone[15]). This equates to ~1.5% of the mapped human autosomes[25] that were able to withstand CNV within our sample set. This did not take into account the percentage of single- and double-occurrence loci that represented true CNVs. The two most diverse samples were S73 and S83. They differed at 266 of the high-frequency CNV loci. Then, we asked the question, What is the greatest difference in genome size between two samples within our set? S55 has the highest net gain of CNV clones, at 97, whereas S83 has the highest net loss of CNV clone, at −131. Comparison of these genomes revealed a difference of 228 clones, representing a difference of at least 9 Mb in genomic size between these two individuals.

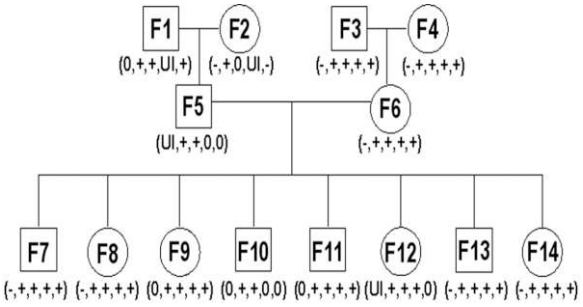

Figure 8. Inheritance of CNVs at five olfactory receptor loci in 14 members of a CEPH pedigree. The five loci (and clones), in the order shown, are *OR2A1* (RP11-466J6), *OR2Z1* (RP11-367L15 and RP11-282G19), *OR4K1* (RP11-449I24 and CTD-2024K23), *OR4M1* (RP11-597A11), and *OR4Q3* (RP11-490A23). − = Copy-number loss; + = copy-number gain; 0 = no copy-number change; UI = uninformative. Male and female family members are shown as squares and circles, respectively.

Table 3. Sensory-Related Genes Associated with CNVs

Chromosome Band	Gains and Losses[a]	Gene(s)[b]	Product[c]	Disease[c]	Clone(s) in Locus[d]
1p36.31	25	*TAS1R1*	Sweet taste receptor T1r isoform a,b,c,d	…	RP11-58A11, RP11-719E21
3p21.31	18	*GNAT1*	Guanine nucleotide binding protein, alpha	Night blindness, congenital stationary	RP11-787O14
7q32.1	5	*IMPDH1*	Inosine monophosphate dehydrogenase 1 isoform a,b	Retinitis pigmentosa-10	RP11-636E12
7q32.1	3	*OPN1SW*	Opsin 1 (cone pigments), short-wave-sensitive	Colorblindness, tritan	RP11-638M14
7q35	54	*OR2A12, OR2A14, OR2A2, OR2A25, OR2A5, OR2A1, OR2A42, OR2A7*	Olfactory receptor, family 2, subfamily A	…	RP11-703N5, RP11-466J6
8p23.3	5	*OR4F21, OR4F29*	Olfactory receptor, family 4, subfamily F	…	RP11-418D21
11q11	8	*OR4C6, OR4P4, OR4S2, OR5D13*	Olfactory receptor, family 4, subfamily C,P,S,D	…	RP11-626N6
11q12.3	3	*ROM1*	Retinal outer segment membrane protein 1	Retinitis pigmentosa, digenic	RP11-484M5
12p13.2	3	*TAS2R14, TAS2R44, TAS2R48, TAS2R49, TAS2R50*	Taste receptor, type 2, member 14,44,48,49,50	…	RP11-202N1
12q13.2	3	*OR6C2, OR6C4, OR6C68, OR6C70*	Olfactory receptor, family 6, subfamily C	…	RP11-222A15
14q11.2	61	*OR4M1, OR4Q3, OR4K1, OR4K2, OR4K5, OR4N2, OR4K13, OR4K14, OR4K15*	Olfactory receptor, family 4, subfamily M,Q,K,N	…	RP11-597A11, RP11-490A23, RP11-449I24, CTD-2024K23
15q11.2	26	*OR4M2, OR4N4*	Olfactory receptor, family 4, subfamily M,N	…	RP11-281J20
16p13.3	7	*OR1F1*	Olfactory receptor, family 1, subfamily F	…	RP11-680M24
17q25.3	18	*ACTG1, FSCN2*	Actin, gamma 1 propeptide; fascin 2	Deafness, autosomal dominant 20/26; retinitis pigmentosa-30	RP11-730A9, RP13-550B21
19p13.2	62	*OR2Z1*	Olfactory receptor, family 2, subfamily Z	…	RP11-282G19, RP11-367L15
22q11.1	15	*OR11H1*	Olfactory receptor, family 11, subfamily H	…	RP11-561P7
22q12.3	5	*MYH9*	Myosin, heavy polypeptide 9, nonmuscle	Deafness, autosomal dominant 17	RP11-108P21

[a] Total number of copy-number gains and losses observed for a CNV locus.
[b] Sensory-related gene(s) overlapping a CNV locus.
[c] Gene product(s) and associated disease(s) according to ReqSeq of the UCSC May 2004 assembly and the OMIM Morbid Map.
[d] Clone or overlapping clones in a CNV locus.

CNV-Associated Genes

We next identified candidate genes whose dosage may be affected by the 800 CNV loci (fig. 6 and the tab-delimited ASCII file of data set 2 [online only]). In total, 1,673 RefSeq-annotated genes overlapped 546 of the 800 CNV loci. First, we looked for the CNV containing the *AMY1A-AMY2A* (MIM 104700; MIM 104650) amylase locus, which was a frequently observed copy-number polymorphism.[5] This clone was found to be gained in seven individuals and to be lost in five individuals in our sample set (see the tab-delimited ASCII file of data set 2 [online only]). Intriguingly, many genes possibly involved in the senses were found to associate with our CNVs, including a large group of olfactory receptor genes (table 3). In fact, the CNVs associated with olfactory receptor loci segregated in a Mendelian manner in the CEPH family (fig. 8). We also observed genes associated with taste (*TAS2R* and *TAS1R1* [MIM 606225], encoding taste receptors), hearing (*ACTG1* [MIM 102560] and *MYH9* [MIM 160775]), and sight (*OPN1SW* [MIM 190900], encoding the short-wave–sensitive cone pigment; *GNAT1* [MIM 139330], related to night blindness; and *FSCN2* [MIM 607643], *IMPDH1* [MIM 146690], and *ROM1* [MIM 180721], linked to retinitis pigmentosa) (table 3). In addition, the genes encoding rhesus blood group and defensins were also observed within these common CNVs (see the tab-delimited ASCII file of data set 2 [online only]).

Surprisingly, many genes associated with disease and susceptibility to disease were also found to have CNV among our sample population. For example, a 630-kb re-

Table 4. Select Examples of CNVs Associated with Cancer-Related Genes

Chromosome Band	Gains and Losses[a]	Gene(s)[b]	Product[c]	Clone(s) in Locus[d]
1p36.33	49	*SKI*	V-ski sarcoma viral oncogene homolog	RP11-83K22, RP11-181G12
1p36.32	12	*TP73*	Tumor protein p73	RP11-631K6
1p36.31	16	*TNFRSF25*	Tumor necrosis factor receptor superfamily,	RP11-58A11
1p32.3	32	*RAB3B*	RAB3B, member RAS oncogene family	RP11-469M21, RP11-91A18
1p13.3	6	*VAV3*	Vav 3 oncogene	RP11-480L11
2q14.2	18	*RALB*	V-ral simian leukemia viral oncogene homolog B	RP11-818M2
2q37.3	6	*BOK*	BCL2-related ovarian killer	RP11-343P10
3p21.31	20	*NAT6, TUSC2, TUSC4*	Putative tumor suppressor FUS2, tumor suppressor candidates 2 & 4	RP11-787O14, RP13-487A19
4q31.1	3	*RAB33B*	RAB33B, member RAS oncogene family	RP11-124P22
6q21	3	*C6orf210*	Candidate tumor suppressor protein	RP11-601O12
6q25.1	20	*ESR1*	Estrogen receptor 1	RP11-655H19
7p22.3	19	*MAFK*	V-maf musculoaponeurotic fibrosarcoma oncogene	RP11-16P10
7p22.3	6	*MAD1L1*	MAD1-like 1	RP11-325O9
8q24.21	4	*MYC*	V-myc myelocytomatosis viral oncogene homolog	CTD-2034C18
9q34.2	22	*VAV2*	Vav 2 oncogene	RP11-352K12, RP11-651E2
10p11.23	11	*MAP3K8*	Mitogen-activated protein kinase kinase kinase	RP11-350D11
11p15.4	15	*CDKN1C*	Cyclin-dependent kinase inhibitor 1C	RP11-494F4
11p13	3	*WT1, WIT-1*	Wilms tumor 1 isoform A/B/C/D, Wilms tumor associated protein	RP11-710L2
11p11.2	3	*C1QTNF4*	C1q and tumor necrosis factor related protein 4	RP11-425G10
11q13.1	3	*MEN1*	Menin isoform 1	RP11-485O9
11q13.3	6	*CCND1, ORAOV1*	Cyclin D1, oral cancer overexpressed 1	RP11-124K14
12q13.12	4	*MLL2*	Myeloid/lymphoid or mixed-lineage leukemia 2	RP11-66M13
13q31.1	4	*C13orf10*	Cutaneous T-cell lymphoma tumor antigen se70-2	RP11-86D5
14q32.32	3	*TNFAIP2*	Tumor necrosis factor, alpha-induced protein 2	RP11-455L5
16p13.3	19	*AXIN1*	Axin 1 isoform a/b	RP11-598I20
16q22.3	3	*BCAR1*	Breast cancer anti-estrogen resistance 1	RP11-109K6
17p13.2	6	*TAX1BP3*	Tax1 (human T-cell leukemia virus type I)	RP11-753P16
17q11.2	6	*NF1*	Neurofibromin	RP11-518B17
17q21.32	3	*PHB*	Prohibitin	RP11-472H5
17q25.3	17	*MAFG*	V-maf musculoaponeurotic fibrosarcoma oncogene	RP11-634L10, RP11-712H22
17q25.3	6	*C1QTNF1*	C1q and tumor necrosis factor related protein 1	RP11-167N2
18p11.32	15	*YES1*	Viral oncogene yes-1 homolog 1	RP11-806L2
18q21.1	8	*DCC*	Deleted in colorectal carcinoma	RP11-346H17
19p13.3	6	*SH3GL1*	SH3-domain GRB2-like 1	RP11-406I1
19p13.3	4	*TNFSF9, TNFSF7, TNFSF14*	Tumor necrosis factor (ligand) superfamily, members	RP11-526C20
19p13.3	4	*VAV1*	Vav 1 oncogene	CTD-2200O16
19p13.11	16	*RAB3A*	RAB3A, member RAS oncogene family	RP11-512B16
19q13.33	15	*PTOV1*	Prostate tumor overexpressed gene 1	RP11-597G9
19q13.33	7	*BAX*	BCL2-associated X protein isoform sigma/gamma/epsilon/delta/beta/alpha	CTD-2017J20
19q13.33	8	*RRAS*	Related RAS viral (r-ras) oncogene homolog	RP11-264M8, RP11-808J4
20q13.13	3	*BCAS4*	Breast carcinoma amplified sequence 4 isoform a/b	RP11-124P7
22q11.21	3	*HIC2*	Hypermethylated in cancer 2	CTD-2245I11

[a] Total number of copy-number gains and losses observed for a CNV locus.
[b] Gene associated with cancer, according to ReqSeq of the UCSC May 2004 assembly and the OMIM Morbid Map, overlapping a CNV locus.
[c] Product encoded by the gene.
[d] Clone or overlapping clones in a CNV locus.

gion on chromosome 3p21.3 shown to be deleted in lung cancer was observed to be associated with copy-number loss in 20 individuals in this study.[26] This region encompasses the putative tumor-suppressor genes *TUSC2* (MIM 607052), *TUSC4* (MIM 607072), and *NAT6* (MIM 607073) (fig. 6, table 4, and the tab-delimited ASCII file of data set 2 [online only]). Many other putative oncogenes and tumor-suppressor genes were also associated with CNVs, such as the *VAV2* (MIM 600428) oncogene; *RAB3B* (MIM 179510), of the RAS oncogene family; *TNFRSF25* (MIM 603366); and *CDKN1C* (MIM 600856) (table 4 and the tab-delimited ASCII file of data set 2 [online only]). In addition to cancer-related genes, CNVs also overlapped genes associated with a bleeding disorder (*TBXA2R* [MIM 188070]), diabetes mellitus (*GCK* [MIM 138079]), and spinal muscular atrophy (*BSCL2* [MIM 606158], *SMA3* [MIM 253400], *SMA4* [MIM 271150], and *SMN1* [MIM 600354]), as well as with susceptibility to Alzheimer disease (*A2M* [MIM 103950]), coronary artery disease (*LPA* [MIM 152200]), and schizophrenia (*COMT* [MIM 116790])

Table 5. Select CNVs Overlapping Genes Associated with Diseases or Disease Susceptibility

Chromosome Band	Gains and Losses[a]	Gene(s)[b]	Product(s)[c]	Disease[d]	Clone(s) in Locus[e]
1p36.11	7	*NR0B2*	Short heterodimer partner	Obesity, mild, early-onset	RP11-492E20
2q31.2	7	*TTN*	Titin isoform N2-A, N2-B; isoform novex-1,2,3	Muscular dystrophy, limb-girdle, type 2J	RP11-95I17
4q11	3	*SGCB*	Sarcoglycan, beta (43kDa dystrophin-associated)	Muscular dystrophy, limb-girdle, type 2E	RP11-61F5
5q13.2	60	*SMA3, SMA4*	SMA3, SMA4	Spinal muscular atrophy-2,-1	RP11-313J5, RP11-155O16
5q13.2	6	*SMN1*	Survival of motor neuron 1, telomeric isoform d	Spinal muscular atrophy-4	RP11-195E2
6q25.3	34	*LPA*	Lipoprotein, Lp(a)	Coronary artery disease, susceptibility to	CTD-2310B5
6q26	5	*PARK2*	Parkin isoform 1, 2, 3	Parkinson disease, juvenile, type 2	CTD-2019O18
7p13	10	*GCK*	Glucokinase isoform 2,3	Diabetes mellitus, neonatal-onset	RP11-808H7
9q22.33	4	*GPR51*	G protein-coupled receptor 51	Nicotine dependence, susceptibility to	RP11-786E15
11q12.3	3	*BSCL2*	Seipin	Spinal muscular atrophy, distal, type V	RP11-484M5
12p13.31	79	*A2M*	Alpha-2-macroglobulin precursor	Alzheimer disease, susceptibility to	RP11-536M6
19p13.3	29	*TBXA2R*	Thromboxane A2 receptor isoform 2	Bleeding disorder due to defective thromboxane A2 receptor	RP11-584K12
19q13.32	3	*FKRP*	Fukutin-related protein	Muscular dystrophy, limb-girdle, type 2I	RP11-422M7
22q11.21	6	*COMT*	Catechol-O-methyltransferase isoform S-COMT	Schizophrenia, susceptibility to	RP11-651A4

[a] Total number of gains and losses observed for a CNV locus.
[b] Gene associated with disease or disease susceptibility, according to ReqSeq of the UCSC May 2004 assembly and the OMIM Morbid Map, overlapping a CNV locus.
[c] Product encoded by the gene.
[d] Disease or disease susceptibility associated with the gene, according to the OMIM Morbid Map.
[e] Clone or overlapping clones in a CNV locus.

(table 5). Furthermore, we found 21 human microRNAs that reside within 14 of the high-frequency CNV loci (fig. 6 and table 6).

Discussion

The existence of large segmental duplications and deletions in the human genome has long been observed through conventional cytogenetic analyses that use light microscopy.[27] More recent genomewide analyses with increased resolutions have revealed that CNVs are present throughout the entire human genome[2–6]; however, limited genomic coverage of the arrays or the limitations of the various techniques has restricted the discovery of CNVs present in the sample populations. It is currently hypothesized that several thousand CNVs exist within the human genome and thus that most are yet to be discovered.[9,28] Here, we used a whole-genome tiling BAC array CGH approach and identified both segmental gains and segmental losses throughout the entire human genome. With complete genome coverage and the tiling nature of our array, we were able to identify a large number of candidate CNVs (3,654). With a focus on only the 800 frequently occurring loci, this study has significantly expanded our knowledge of CNVs. A large proportion (77%) of these high-frequency CNVs are novel; the lack of complete overlap between our CNVs and previously reported CNVs is consistent with the current hypothesis that thousands of CNVs exist in the human population.

In our data set, the net difference in genomic size between two individuals could vary widely, by at least 9 Mb in the two most diverse, representing a difference of 228 distinct CNV clones. In addition, pairwise comparison of the high-frequency CNVs among the 95 individuals revealed that the genomes of the two most diverse individuals differed at 266 loci. These data demonstrate that a significant fraction of the human genome can vary in copy number. On the basis of our high-frequency CNV data set and a minimum detection sensitivity for BAC array CGH of 40 kb, at least 1.5% of the mapped human autosomes is tolerant to CNV. This is an underestimate because the percentage of single- and double-occurrence loci that may represent true CNVs was not taken into account.

Over 1,500 genes were found to overlap the high-frequency CNVs detected in this study. Several of these CNV-associated genes are related to the senses, including a group of olfactory receptor genes, multiple taste-receptor genes, and several genes related to sight or hearing. Genes that are well-known to have variable copy number—such as those encoding rhesus blood group, amylases, and defensins—were also observed within our common CNVs. These associations suggest that CNVs may contribute to phenotypic diversity in humans. Elsewhere, segmental copy-number gains or losses have been demonstrated to associate with developmental disorders and susceptibility to human disease.[10] Many genes associated with disease and susceptibility to disease were found to show CNV among the individuals within our study. These include genes associated with diabetes mellitus or a bleeding disorder; cancer-related genes, such as putative oncogenes

Table 6. MicroRNAs Overlapping CNVs

Chromosome Band	Gains and Losses[a]	microRNA(s)[b]	Clone(s) in Locus[c]
3p21.2	7	hsa-let-7g, hsa-mir-135a-1	RP11-185J5, RP11-258D4
4p16.1	15	hsa-mir-95	CTD-2104N3, RP11-512D9
4p15.31	27	hsa-mir-218-1	RP11-644J20
8p21.3	9	hsa-mir-320	RP11-13A10
9q22.32	18	hsa-let-7a-1, hsa-let-7d, hsa-let-7f-1	RP11-519D15
10q26.3	21	hsa-mir-202	RP11-319M21, RP11-466F21, RP13-520O22
11q12.1	3	hsa-mir-130a	RP11-781C10
17q25.3	13	hsa-mir-338	RP11-149I9
19p13.2	13	hsa-mir-199a-1	RP11-20N24, RP11-751C24
19p13.13	4	hsa-mir-181c, hsa-mir-181d, hsa-mir-23a, hsa-mir-24-2, hsa-mir-27a	RP11-423F4
19q13.33	25	hsa-mir-150	RP11-21O13
20q11.22	3	hsa-mir-499	RP11-638P17
20q13.33	74	hsa-mir-124a-3	CTD-2240P21, RP11-543D7
22q11.21	6	hsa-mir-185	RP11-651A4

[a] Total number of gains and losses observed for a CNV locus.
[b] Human microRNA(s) downloaded from the Sanger miRBase database overlapping a CNV locus.
[c] Clone or overlapping clones in a CNV locus.

and tumor-suppressor genes; and genes associated with susceptibility to coronary artery disease or Alzheimer disease. Like other aspects of human genetic variation, understanding of CNVs is critical for studying disease-associated changes correctly, as illustrated in the genome profiling of patients with mental retardation.[24] Clinically relevant alterations in copy number need to be separated from a baseline of CNVs for gene discovery. Therefore, it is of utmost importance when genetic association studies of diseases are conducted that they be interpreted in the context of baseline segmental copy-number status; CNVs identified in this study provide a source of information for such a baseline. Interestingly, several of our CNV loci were also found to overlap with microRNAs. Although the functions of microRNAs are largely unknown, they may play a role in the regulation of various biological processes, such as the control of development, differentiation, cell proliferation, and apoptosis, and they have also been linked to human diseases.[29–31] Recent studies have shown a global downregulation of microRNAs in tumors compared with in normal tissues and an upregulation of microRNA expression via copy-number changes in lymphoma.[32,33] Our data raise the possibility that CNVs encompassing microRNAs contribute to human diversity and disease susceptibility.

This comprehensive whole-genome study, identifying both segmental gains and losses in the human population, has significantly expanded our knowledge of CNVs. Remarkably, the genomes of the two most diverse individuals within this study differed by at least 9 Mb in size, or 266 loci in content. In addition, on the basis of our high-frequency CNV data set, at least 1.5% of the human genome is tolerant of CNV. However, with the lack of complete overlap between our CNVs and those identified elsewhere and the hypothesis that thousands of CNVs exist in the human genome, this comprehensive study is still an early step toward a more complete understanding of CNVs within the human population, and more studies are needed to examine the functional roles of CNVs.

Acknowledgments

We thank Media Farshchi and Wendy Peng for computational analysis, Andy Lam and Eric Lee for technical assistance, Sharon Gee for sample collection, the Lam Lab array CGH group for array production, Drs. Carlos Alvarez and Ford Doolittle and members of the Lam Lab for helpful discussions, and especially all donors. This work was supported by funds from Genome Canada/British Columbia, the Canadian Institute of Health Research (to W.L.L. and C.J.B.), the National Institutes of Health (NIH) National Institute of Dental and Cranial Research (to W.L.L.), a Michael Smith Foundation for Health Research scholarship (to R.J.D.), a National Sciences and Engineering Research Council of Canada scholarship (to R.J.D.), and an NIH grant (to E.E.E.). E.E.E. is an Investigator of the Howard Hughes Medical Institute.

Web Resources

The URLs for data presented herein are as follows:

BACPAC Resources, http://bacpac.chori.org/genomicRearrays.php (for UCSC May 2004 mapping annotations)
Database of Genomic Variants, http://projects.tcag.ca/variation/
Eisen Lab: Software, http://rana.lbl.gov/EisenSoftware.htm (for Cluster and Treeview)
Gene Expression Omnibus (GEO), http://www.ncbi.nlm.nih.gov/geo/
miRBase, http://microrna.sanger.ac.uk/sequences/
OMIM, http://www.ncbi.nlm.nih.gov/Omim/ (for *BRCA1, BRCA2, APC, MSH2, MSH6, MLH1, AMY1A, AMY2A, TAS1R1, ACTG1, MYH9, OPN1SW, GNAT1, FSCN2, IMPDH1, ROM1, TUSC2, TUSC4, NAT6, VAV2, RAB3B, TNFRSF25, CDKN1C, TBXA2R, GCK, BSCL2, SMA3, SMA4, SMN1, A2M, LPA,* and *COMT*)
OMIM Morbid Map, ftp://ftp.ncbi.nih.gov/repository/OMIM/morbidmap
Segmental Duplication Database, http://humanparalogy.gs.washington.edu

SMRT Array, http://www.bccrc.ca/arraycgh/
UCSC Genome Bioinformatics, http://genome.ucsc.edu/ (for May 2004 assembly)
UCSC Human Genome Browser, http://genome.ucsc.edu/cgi-bin/hgGateway

References

1. Altshuler D, Brooks LD, Chakravarti A, Collins FS, Daly MJ, Donnelly P (2005) A haplotype map of the human genome. Nature 437:1299–1320
2. Conrad DF, Andrews TD, Carter NP, Hurles ME, Pritchard JK (2006) A high-resolution survey of deletion polymorphism in the human genome. Nat Genet 38:75–81
3. Hinds DA, Kloek AP, Jen M, Chen X, Frazer KA (2006) Common deletions and SNPs are in linkage disequilibrium in the human genome. Nat Genet 38:82–85
4. McCarroll SA, Hadnott TN, Perry GH, Sabeti PC, Zody MC, Barrett JC, Dallaire S, Gabriel SB, Lee C, Daly MJ, et al (2006) Common deletion polymorphisms in the human genome. Nat Genet 38:86–92
5. Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, Qi Y, Scherer SW, Lee C (2004) Detection of large-scale variation in the human genome. Nat Genet 36:949–951
6. Sebat J, Lakshmi B, Troge J, Alexander J, Young J, Lundin P, Maner S, Massa H, Walker M, Chi M, et al (2004) Large-scale copy number polymorphism in the human genome. Science 305:525–528
7. Sharp AJ, Locke DP, McGrath SD, Cheng Z, Bailey JA, Vallente RU, Pertz LM, Clark RA, Schwartz S, Segraves R, et al (2005) Segmental duplications and copy-number variation in the human genome. Am J Hum Genet 77:78–88
8. Tuzun E, Sharp AJ, Bailey JA, Kaul R, Morrison VA, Pertz LM, Haugen E, Hayden H, Albertson D, Pinkel D, et al (2005) Fine-scale structural variation of the human genome. Nat Genet 37:727–732
9. Eichler EE (2006) Widening the spectrum of human genetic variation. Nat Genet 38:9–11
10. Inoue K, Lupski JR (2002) Molecular mechanisms for genomic disorders. Annu Rev Genomics Hum Genet 3:199–242
11. Ishkanian AS, Malloff CA, Watson SK, DeLeeuw RJ, Chi B, Coe BP, Snijders A, Albertson DG, Pinkel D, Marra MA, et al (2004) A tiling resolution DNA microarray with complete coverage of the human genome. Nat Genet 36:299–303
12. Khojasteh M, Lam WL, Ward RK, MacAulay C (2005) A stepwise framework for the normalization of array CGH data. BMC Bioinformatics 6:274
13. Chi B, deLeeuw RJ, Coe BP, MacAulay C, Lam WL (2004) SeeGH—a software tool for visualization of whole genome array comparative genomic hybridization data. BMC Bioinformatics 5:13
14. Locke DP, Sharp AJ, McCarroll SA, McGrath SD, Newman TL, Cheng Z, Schwartz S, Albertson DG, Pinkel D, Altshuler DM, et al (2006) Linkage disequilibrium and heritability of copy-number polymorphisms within duplicated regions of the human genome. Am J Hum Genet 79:275–290
15. Pinkel D, Segraves R, Sudar D, Clark S, Poole I, Kowbel D, Collins C, Kuo WL, Chen C, Zhai Y, et al (1998) High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. Nat Genet 20: 207–211
16. Griffiths-Jones S (2004) The microRNA Registry. Nucleic Acids Res 32:D109–D111
17. Bailey JA, Gu Z, Clark RA, Reinert K, Samonte RV, Schwartz S, Adams MD, Myers EW, Li PW, Eichler EE (2002) Recent segmental duplications in the human genome. Science 297: 1003–1007
18. Bailey JA, Yavor AM, Massa HF, Trask BJ, Eichler EE (2001) Segmental duplications: organization and impact within the current human genome project assembly. Genome Res 11: 1005–1017
19. Cheng Z, Ventura M, She X, Khaitovich P, Graves T, Osoegawa K, Church D, DeJong P, Wilson RK, Paabo S, et al (2005) A genome-wide comparison of recent chimpanzee and human segmental duplications. Nature 437:88–93
20. She X, Jiang Z, Clark RA, Liu G, Cheng Z, Tuzun E, Church DM, Sutton G, Halpern AL, Eichler EE (2004) Shotgun sequence assembly and recent segmental duplications within the human genome. Nature 431:927–930
21. Eisen MB, Spellman PT, Brown PO, Botstein D (1998) Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci USA 95:14863–14868
22. Rozen S, Skaletsky H (2000) Primer3 on the WWW for general users and for biologist programmers. Methods Mol Biol 132: 365–386
23. Weksberg R, Hughes S, Moldovan L, Bassett AS, Chow EW, Squire JA (2005) A method for accurate detection of genomic microdeletions using real-time quantitative PCR. BMC Genomics 6:180
24. de Vries BB, Pfundt R, Leisink M, Koolen DA, Vissers LE, Janssen IM, Reijmersdal S, Nillesen WM, Huys EH, Leeuw N, et al (2005) Diagnostic genome profiling in mental retardation. Am J Hum Genet 77:606–616
25. International Human Genome Sequencing Consortium (2004) Finishing the euchromatic sequence of the human genome. Nature 431:931–945
26. Lerman MI, Minna JD, for The International Lung Cancer Chromosome 3p21.3 Tumor Suppressor Gene Consortium (2000) The 630-kb lung cancer homozygous deletion region on human chromosome 3p21.3: identification and evaluation of the resident candidate tumor suppressor genes. Cancer Res 60:6116–6133
27. Seabright M (1971) A rapid banding technique for human chromosomes. Lancet 2:971–972
28. Lee C (2005) Vive la difference! Nat Genet 37:660–661
29. Alvarez-Garcia I, Miska EA (2005) MicroRNA functions in animal development and human disease. Development 132: 4653–4662
30. Bartel DP (2004) MicroRNAs: genomics, biogenesis, mechanism, and function. Cell 116:281–297
31. Wienholds E, Plasterk RH (2005) MicroRNA function in animal development. FEBS Lett 579:5911–5922
32. Lu J, Getz G, Miska EA, Alvarez-Saavedra E, Lamb J, Peck D, Sweet-Cordero A, Ebert BL, Mak RH, Ferrando AA, et al (2005) MicroRNA expression profiles classify human cancers. Nature 435:834–838
33. Tagawa H, Seto M (2005) A microRNA cluster as a target of genomic amplification in malignant lymphoma. Leukemia 19:2013–2016