J Hum Genet (2002) 47:605–610

ORIGINAL ARTICLE

Hisanori Haga · Ryo Yamada · Yozo Ohnishi
Yusuke Nakamura · Toshihiro Tanaka

Gene-based SNP discovery as part of the Japanese Millennium Genome Project: identification of 190 562 genetic variations in the human genome

Received: August 22, 2002 / Accepted: August 27, 2002

Abstract To construct an infrastructure for genome-wide association studies of common diseases or drug sensitivities, we have been systematically exploring common variants by resequencing genomic regions containing genes in DNA from 24 Japanese individuals. We have analyzed a total of 154 Mb, corresponding to approximately 5% of the human genome, and so far have identified 174 269 single-nucleotide polymorphisms and 16 293 insertion/deletion polymorphisms within gene regions, i.e., one polymorphism in 807 bp on average. Our data are freely available via our web site (http://snp.ims.u-tokyo.ac.jp) and will facilitate studies to identify genes associated with susceptibility to common diseases and genes involved in sensitivity to therapeutic drugs.

Key words Single-nucleotide polymorphism (SNP) · Japanese population · Genetic marker · Association study · Common diseases

Introduction

During the past several years, knowledge of the human genomic sequence has accumulated to the point that millions of single-nucleotide polymorphisms (SNPs) are now registered in publically available databases such as dbSNP (Sherry et al. 2001). However, the quality of the dbSNP database has been questioned because a considerable portion of the archived information does not truly reflect polymorphism, but simply represents sequencing errors; SNPs in the database are often defined as sequence differences in clusters of overlapping expressed sequence tags, or differences among large-insert clones such as bacterial artificial chromosomes assembled by computer algorithms (Altshuler et al. 2000; Sachidanandam et al. 2001; Venter et al. 2001). Moreover, elements deposited in the dbSNP database are not, for the most part, located within genomic regions that contain genes. Although extragenetic SNPs can serve as markers to identify loci for susceptibility to common diseases or loci defining drug sensitivity, they might not be appropriate for such studies if the goal is ultimately to identify the genes involved.

There are two approaches to construction of SNP databases; one is genome-wide screening, and the other is gene-based screening. The former, the "whole-genome shotgun method," involves sequencing of genomic clones prepared from a number of individuals; this approach does not require synthesis of oligonucleotide primers for polymerase chain reaction (PCR) amplification or knowledge of genomic sequence (Altshuler et al. 2000; Sachidanandam et al. 2001; Venter et al. 2001). However, to avoid spurious SNPs, the quality of results depends heavily on the accuracy of detection algorithms; such accuracy is technically difficult to achieve, and, furthermore, most SNPs identified in this way inevitably will be located in intergenic regions. The second method is based on locus-specific PCR amplification and direct sequencing of the products, but it requires millions of oligonucleotide primers and extensive genomic sequence information to cover the whole genome. However, the latter approach can focus on specific regions of interest to explore for variations, i.e., promoter regions or coding elements. Because gene-based variations should have greater usefulness for identifying loci containing genes of medical or biological importance by means of association studies, we have adopted this strategy and describe here a large-scale discovery of genetic variations within gene-containing regions as one of the Japanese Millennium Genome projects.

H. Haga · R. Yamada · Y. Ohnishi · Y. Nakamura (✉) · T. Tanaka
Laboratory of Molecular Medicine, Human Genome Center,
Institute of Medical Science, University of Tokyo, 4-6-1
Shirokanedai, Minato-ku, Tokyo 108-8639, Japan
Tel. +81-3-5449-5372; Fax +81-3-5449-5433
e-mail: yusuke@ims.u-tokyo.ac.jp

Subjects and methods

Selection of gene-containing regions

Because we decided to concentrate on finding SNPs present in genes or their promoters, we defined "gene regions" by four approaches. First, we extracted appropriate genomic sequences from the GenBank DNA database (Benson et al. 2002) by selecting entries containing annotations of exons and promoters. Second, we applied an exon-prediction program (GENSCAN; Burge and Karlin 1997) and searched for homologies using BLAST ver.2 (Altschul et al. 1997), partly with sim4 (Florea et al. 1998), between mRNA/cDNA/Unigene records (Wheeler et al. 2002) and genomic sequence records. Third, we determined exonic elements by comparing mRNA records with the genomic contig records in the NCBI Reference Sequence Project (RefSeq; Pruitt and Maglott 2001) using our original computer software. Last, we defined promoter regions as encompassing 2.5 kb of DNA upstream of the first exon. To avoid identifying controversial SNPs in repetitive elements, we eliminated such regions by means of the RepeatMasker program (http://ftp.genome.washington.edu/RM/RepeatMasker.html), using the option of "Do not mask simple repeats and low-complexity DNA," and we did not amplify regions in which more than 20% of the DNA was repetitive. PCR primers were selected to amplify extracted exons, introns, and promoter regions using the computer program Primer 3.0 (Rozen and Skaletsky 2000); the size of each PCR product was designed to be approximately 0.6 or 1.2 kb. Internal and nested primers were also selected for sequencing when necessary.

DNA samples

After obtaining written informed consent, genomic DNA was extracted from the peripheral blood of 24 unrelated female Japanese volunteers. At first we had included samples from male volunteers as well; however, pseudoautosomal regions can confound assignment of variations to sex chromosomes, and we decided not to investigate the Y chromosome further. No identifying or phenotypic information was recorded in relation to any specimen, so that links to individual donors were irreversibly broken. The Ethics Committee at the Institute of Medical Science, University of Tokyo had given us permission to analyze these samples.

Locus-specific PCR amplification and direct sequencing of PCR products

In the interest of cost-effectiveness, we mixed equal amounts of DNA from three unrelated individuals because preliminary data had indicated that one variant in six chromosomes could be effectively detected in this way (Ohnishi et al. 2000; Yamada et al. 2000). Each PCR experiment was performed with 20 ng of pooled genomic DNA. Amplification was carried out under the conditions described in our web site (http://snp.ims.u-tokyo.ac.jp), and for liquid handling we used Biomek2000 (Beckman Coulter, Fullerton, CA, USA) and Multimek96 (Beckman Coulter) robotics. PCR products were purified on Multiscreen384-PCR plates (Millipore, Bedford, MA, USA) according to the manufacturer's instructions. The purified products were sequenced directly, with sets of two PCR primers for both strands. When the target size of a PCR product was 1.2 kb, two internal primers were also used to cover the whole amplified region with both strands. In case a PCR product showed faint nonspecific band(s) after electrophoresis on an agarose gel, two nested primers rather than the PCR primers themselves were used for sequencing. Samples were sequenced using Big-Dye terminator chemistry on ABI 3700 capillary sequencers (Applied Biosystems, Foster City, CA, USA). To perform large-scale sequencing, we introduced 22 capillary sequencers. All procedures were administered by means of bar-code systems to avoid human errors.

SNP detection

The Polyphred Computer program (Nickerson et al. 1997) was used to assemble fragments and indicate candidate SNPs. For verification we inspected most of the amplified regions and all candidate SNPs visually. When all eight of the pooled samples showed the same allelic ratio on an electropherogram, we did not consider the candidate site to be a polymorphism, because noise and/or the presence of homologous regions can present risks of false verification.

SNP database

Our database was constructed as described previously (Hirakawa et al. 2002). For statistical purposes, exons and genomic regions were defined as sequences in the RefSeq (build 29) whose accession format began with XM or NM, or with NT, respectively.

Results and discussion

We designed PCR primers for 260 448 gene-containing fragments from throughout the human genome, and we were able to analyze 84% of them for polymorphisms. The unsuccessful PCR amplifications may have been due to errors in the reference sequences in the genomic databases.

We screened 153 774 997 nonredundant nucleotides, this number corresponding to approximately 5% of the human genome, and identified 190 562 genetic variations; 174 269 were SNPs and 16 293 were insertion/deletion polymorphisms, including 2439 repeat-length polymorphisms such as microsatellites (Table 1). Approximately 70% of the SNPs were transitions, this type occurring 2.3 times more often than transversions, an observation consistent with previous results and theory (Ohnishi et al. 2000; Dawson

et al. 2001; Venter et al. 2001). Interestingly, the number of insertion/deletion polymorphisms decreased exponentially according to their sizes (Fig. 1). Because our screening method should not affect the discovery rate regardless of the sizes of insertion/deletions, this result probably reflects a genome-wide distribution of this kind of polymorphism.

The distribution of polymorphisms with respect to genetic structures is summarized in Table 2. Approximately 3% of the SNPs we identified were mapped to multiple genomic regions, mainly as a consequence of redundancy in the RefSeq (build 29) database. However, 10% of our SNPs could not be mapped to any of the sequences deposited in databases, even though we had synthesized the PCR primers using GenBank data. We think it is clear that the RefSeq database is far from complete, because draft sequences still cover 35% of its genomic contigs. Table 2 will surely be in need of modification whenever this database is changed to reflect new data.

Among 12830 biallelic SNPs identified within regions judged to contain coding elements on the basis of annotations in RefSeq (153 SNPs were located on annotated "CDS," which did not have long open reading frames), approximately half were nonsynonymous substitutions and most of those were missense (Table 3). When we classified these SNPs further on the basis of the positions of variant bases within codons, half were located at the third nucleotide (Table 4), supporting the theory that many deleterious alleles encoding nonsynonymous substitutions at the first or second site probably have been eliminated during human evolution.

The calculated length of all exonic sequences analyzed in this study was 27015998 nucleotides. Because by our estimates the combined length of all exons in the human genome is 67214630 nucleotides, the exons we screened for genetic variations corresponded to approximately 40% of the entire exonic sequence of the genome. The estimate of 67.2Mb is consistent with previous estimates that were based on average exon length and the number of genes in the genome (Sachidanandam et al. 2001; Venter et al. 2001). We went on to investigate the coverage of our data set against known genes. When we used the set of 46118 XM or NM accession formats (RefSeq build 29) as known genes, 48% (22155) were found to contain one or more of the SNPs identified by this study. Furthermore, as summarized in Table 5, the density of our SNPs in autosomes appeared to range from 6.6 to 34.4kb. We conclude that our SNP collection should be useful for genome-wide case control association studies, especially when those studies focus on regions containing specific genes.

We also determined allelic frequencies for 78570 of the SNPs identified in our study using our own high-throughput genotyping method (Ohnishi et al. 2001). The distribution of allelic frequencies was largely even, with an average minor-allelic frequency of 24% (Fig. 2 and http://snp.ims.u-tokyo.ac.jp). These frequency data will serve as controls

Table 1. Summary of polymorphisms identified in this study

SNP		174269 (91.45%)
Transition	A/G (C/T)	121466 (63.74%)
Transversion		52697 (27.65%)
	A/C (G/T)	26377 (13.84%)
	A/T	8881 (4.66%)
	C/G	17438 (9.15%)
Triallelic		107 (0.06%)
	A/C/G	33 (0.02%)
	C/G/T	27 (0.01%)
	A/C/T	24 (0.01%)
	A/G/T	23 (0.01%)
Insertion/deletion		16293 (8.55%)
Total		190562 (100%)

SNP, Single-nucleotide, polymorphism

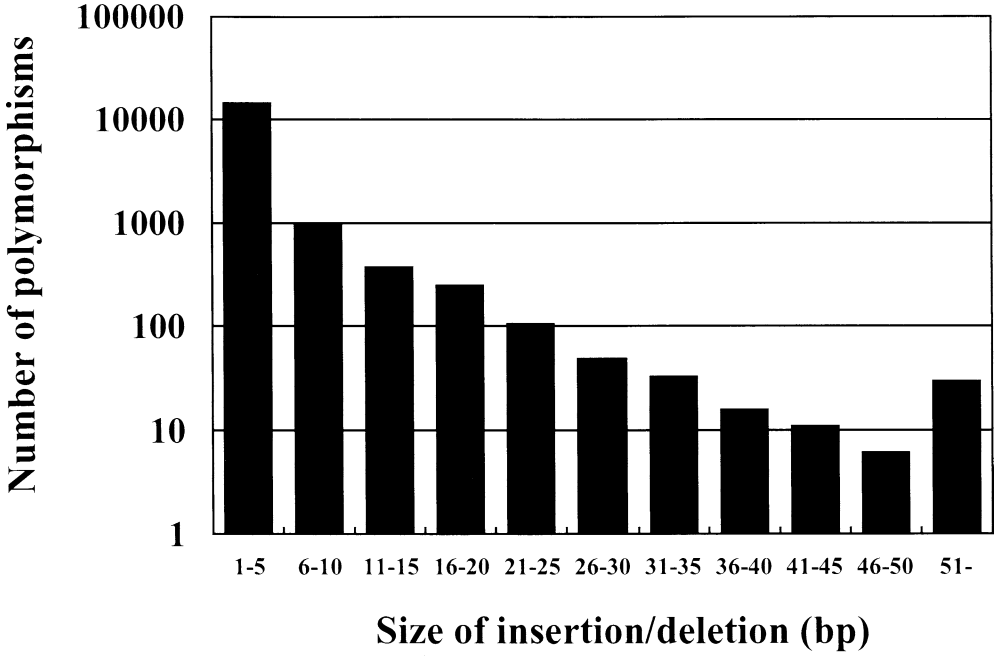

Fig. 1. Size distribution of insertion/deletion polymorphisms

Table 2. Polymorphism distribution by gene structure using RefSeq (build 29)

	SNP	Insertion/deletion	Total polymorphisms	Base pairs screened	Frequency of polymorphism (bp)
Promoter region	14142	1227	15369	13228503	861
5′ UTR	2933	166	3099	2805473	905
CDS	12983	179	13162	17084908	1298
3′ UTR	7216	826	8042	7125617	886
Intron	99609	8350	107959	91444581	847
3′ Flanking region	4512	545	5057	4900896	969
Other[a]	12305	932	13237	13540215	1023
Mapped to multiple regions[b]	6174	458	6632	n.a.	n.a.
Unmapped[c]	14395	3610	18005	3644804	202
Total	174269	16293	190562	153774997	807

SNP, Single-nucleotide polymorphism; UTR, untranslated region
[a] Not mapped within gene regions
[b] Due to the redundancy of NT sequences in RefSeq (build 29)
[c] Could not be mapped to RefSeq (build 29)

Table 3. SNP distribution by amino acid substitution

	SNPs
Synonymous	6704 (52%)
Nonsynonymous	6126 (48%)
Nonsense	111 (1%)
Missense	6015 (47%)
Conservative	3526 (28%)
Nonconservative	2489 (19%)

SNP, Single-nucleotide polymorphism

Table 4. Distribution of SNPs according to position of nucleotides within codons

	SNPs
First base of codon	3192 (25%)
Second base of codon	2772 (22%)
Third base of codon	6866 (54%)

SNPs, Single-nucleotide polymorphisms

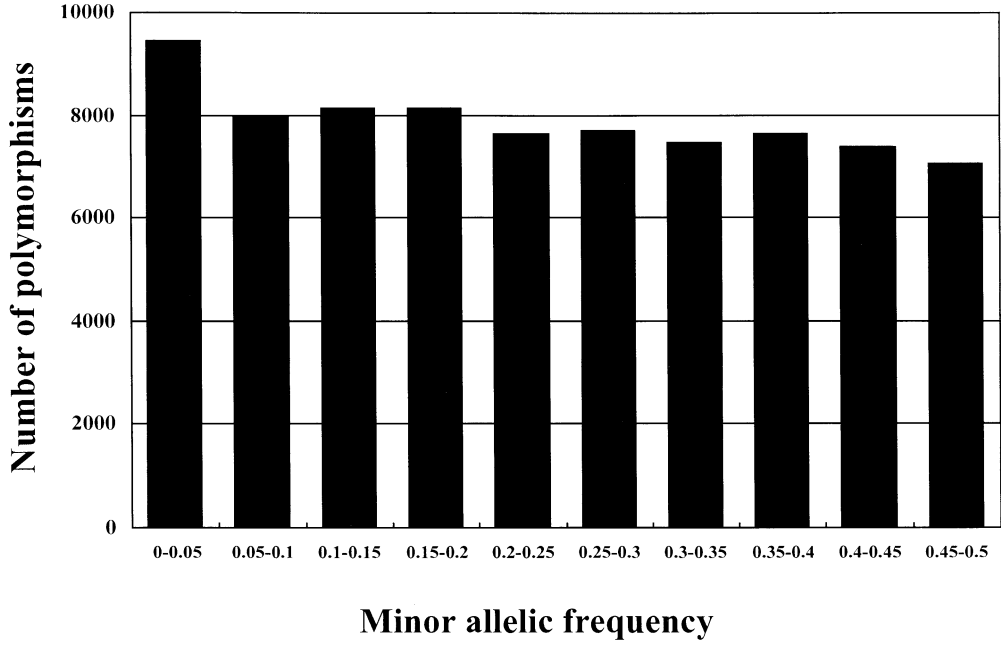

Fig. 2. Distribution of allelic frequencies for genotyped SNPs

when case-control association studies are performed in the Japanese population.

We found, on average, one polymorphism in every 807 bp, a lower density than that found in other studies, where a variation was reported to occur approximately every 300 bp (Cargill et al. 1999; Halushka et al. 1999; Stephens et al. 2001). In the traditional definition of genetics, to be called a polymorphism, a variation at a given site must be present in at least 1% of the population; therefore, an assumption of classical neutral theory of population genetics would suggest that polymorphisms are present at 11 million sites in the human genome (Kruglyak and Nickerson 2001). Under that criterion, a variation should be found every 272 bp, but our two pilot studies (Ohnishi et al. 2000; Yamada et al. 2000) identified one gene-based SNP in every 641 and 638 bp, respectively. A reason for this

Table 5. Distribution of variations by chromosome

Chromosome	SNP	Insertion /deletion	Total variations	Screened bases	Frequency of variation (bp)	Length of chromosome[a] (Mb)	Density of identified variations (bp)
1	15166	1181	16347	14344460	877	220	13458
2	12772	1126	13898	11608613	835	240	17269
3	9052	805	9857	8813579	894	200	20290
4	6336	638	6974	5826029	835	186	26670
5	8284	777	9061	7648904	844	182	20086
6	11609	1023	12632	11252798	891	172	13616
7	11969	931	12900	11528580	894	146	11318
8	4545	364	4909	3887257	792	146	29741
9	5561	441	6002	4802482	800	113	18827
10	5819	470	6289	5014774	797	130	20671
11	5995	430	6425	6109367	951	132	20545
12	7108	570	7678	7190880	937	134	17452
13	2632	242	2874	2236249	778	99	34447
14	5827	436	6263	5231516	835	87	13891
15	4118	324	4442	4001428	901	80	18010
16	5154	314	5468	5517492	1009	75	13716
17	5673	434	6107	6403592	1049	78	12772
18	2263	203	2466	1947367	790	79	32036
19	6952	425	7377	6301849	854	58	7862
20	5075	363	5438	4688878	862	61	11217
21	3452	220	3672	2970778	809	33	8987
22	5198	261	5459	5814927	1065	36	6595
X	3105	246	3351	6169474	1841	128	38198
Y	24	0	24	556613	23192	19	791667

SNP, Single-nucleotide polymorphism
[a] Data from Venter et al. 2001

discrepancy might be ethnic difference; the Japanese population seems more homogeneous than others. Other discrepancies may be related to study designs, particularly in regard to sizes of the population samples and whether individual DNA samples are mixed. Three previous studies analyzed more than 50 individuals each, among whom a considerable percentage of the SNPs identified had low minor-allelic frequencies: Halushka et al. (1999) identified SNPs with an average minor-allelic frequency of 11%, Cargill et al. (1999) reported that about 70% of the SNPs they identified had minor-allelic frequencies of <15%, and Stephens et al. (2001) observed 38% of their SNPs only once, in single heterozygotes. Because we analyzed only 24 individuals and mixed three DNAs in one tube (i.e., eight analytes), it is possible that we did not detect variations having low minor-allelic frequencies, as Fig. 2 indicates. However, we do not think this will diminish the value of our database, because we consider that SNPs with low frequencies of minor alleles are not necessary at present. Detecting significant association of any SNP with a minor-allelic frequency of <5% with common diseases through a genome-wide approach would require a sample size of up to 10000 people. It is simply not practical for such a large group to be genotyped for 100000 SNP loci within a reasonable period of time using systems that are available at present.

All of our data and our methods, including primer sequences and PCR conditions, are freely available from our web site (http://snp.ims.u-tokyo.ac.jp). We believe that this information constitutes a potent infrastructure for the next step toward personalized medicine, i.e., whole-genome association studies of common diseases or drug sensitivities.

Acknowledgments We are grateful to the members of our SNP discovery team for providing us with a high degree of experimental expertise. This work was supported by a grant from the Japanese Millennium Project.

References

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402

Altshuler D, Pollara VJ, Cowles CR, Van Etten WJ, Baldwin J, Linton L, Lander ES (2000) An SNP map of the human genome generated by reduced representation shotgun sequencing. Nature 407:513–516

Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Rapp BA, Wheeler DL (2002) GenBank. Nucleic Acids Res 30:17–20

Burge C, Karlin S (1997) Prediction of complete gene structures in human genomic DNA. J Mol Biol 268:78–94

Cargill M, Altshuler D, Ireland J, Sklar P, Ardlie K, Patil N, Shaw N, Lane CR, Lim EP, Kalyanaraman N, Nemesh J, Ziaugra L, Friedland L, Rolfe A, Warrington J, Lipshutz R, Daley GQ, Lander ES (1999) Characterization of single-nucleotide polymorphisms in coding regions of human genes. Nat Genet 22:231–238

Dawson E, Chen Y, Hunt S, Smink LJ, Hunt A, Rice K, Livingston S, Bumpstead S, Bruskiewich R, Sham P, Ganske R, Adams M, Kawasaki K, Shimizu N, Minoshima S, Roe B, Bentley D, Dunham I (2001) A SNP resource for human chromosome 22: extracting dense clusters of SNPs from the genomic sequence. Genome Res 11:170–178

Florea L, Hartzell G, Zhang Z, Rubin GM, Miller W (1998) A computer program for aligning a cDNA sequence with a genomic DNA sequence. Genome Res 8:967–974

Halushka MK, Fan JB, Bentley K, Hsie L, Shen N, Weder A, Cooper R, Lipshutz R, Chakravarti A (1999) Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. Nat Genet 22:239–247

Hirakawa M, Tanaka T, Hashimoto Y, Kuroda M, Takagi T, Nakamura Y (2002) JSNP: a database of common gene variations in the Japanese population. Nucleic Acids Res 30:158–162

Kruglyak L, Nickerson DA (2001) Variation is the spice of life. Nat Genet 27:234–236

Nickerson DA, Tobe VO, Taylor SL (1997) PolyPhred: automating the detection and genotyping of single nucleotide substitutions using fluorescence-based resequencing. Nucleic Acids Res 25:2745–2751

Ohnishi Y, Tanaka T, Yamada R, Suematsu K, Minami M, Fujii K, Hoki N, Kodama K, Nagata S, Hayashi T, Kinoshita N, Sato H, Sato H, Kuzuya T, Takeda H, Hori M, Nakamura Y (2000) Identification of 187 single nucleotide polymorphisms (SNPs) among 41 candidate genes for ischemic heart disease in the Japanese population. Hum Genet 106:288–292

Ohnishi Y, Tanaka T, Ozaki K, Yamada R, Suzuki H, Nakamura Y (2001) A high-throughput SNP typing system for genome-wide association studies. J Hum Genet 46:471–477

Pruitt KD, Maglott DR (2001) RefSeq and LocusLink: NCBI gene-centered resources. Nucleic Acids Res 29:137–140

Rozen S, Skaletsky H (2000) Primer3 on the WWW for general users and for biologist programmers. Methods Mol Biol 132:365–386

Sachidanandam R, Weissman D, Schmidt SC, Kakol JM, Stein LD, Marth G, Sherry S, Mullikin JC, Mortimore BJ, Willey DL, Hunt SE, Cole CG, Coggill PC, Rice CM, Ning Z, Rogers J, Bentley DR, Kwok PY, Mardis ER, Yeh RT, Schultz B, Cook L, Davenport R, Dante M, Fulton L, Hillier L, Waterston RH, McPherson JD, Gilman B, Schaffner S, Van Etten WJ, Reich D, Higgins J, Daly MJ, Blumenstiel B, Baldwin J, Stange-Thomann N, Zody MC, Linton L, Lander ES, Altshuler D; The International SNP Map Working Group (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature 409:928–933

Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K (2001) dbSNP: the NCBI database of genetic variation. Nucleic Acids Res 29:308–311

Stephens JC, Schneider JA, Tanguay DA, Choi J, Acharya T, Stanley SE, Jiang R, Messer CJ, Chew A, Han JH, Duan J, Carr JL, Lee MS, Koshy B, Kumar AM, Zhang G, Newell WR, Windemuth A, Xu C, Kalbfleisch TS, Shaner SL, Arnold K, Schulz V, Drysdale CM, Nandabalan K, Judson RS, Ruano G, Vovis GF (2001) Haplotype variation and linkage disequilibrium in 313 human genes. Science 293:489–493

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson DH, Wortman JR, Zhang Q, Kodira CD, Zheng XH, Chen L, Skupski M, Subramanian G, Thomas PD, Zhang J, Gabor Miklos GL, Nelson C, Broder S, Clark AG, Nadeau J, McKusick VA, Zinder N, Levine AJ, Roberts RJ, Simon M, Slayman C, Hunkapiller M, Bolanos R, Delcher A, Dew I, Fasulo D, Flanigan M, Florea L, Halpern A, Hannenhalli S, Kravitz S, Levy S, Mobarry C, Reinert K, Remington K, Abu-Threideh J, Beasley E, Biddick K, Bonazzi V, Brandon R, Cargill M, Chandramouliswaran I, Charlab R, Chaturvedi K, Deng Z, Di Francesco V, Dunn P, Eilbeck K, Evangelista C, Gabrielian AE, Gan W, Ge W, Gong F, Gu Z, Guan P, Heiman TJ, Higgins ME, Ji RR, Ke Z, Ketchum KA, Lai Z, Lei Y, Li Z, Li J, Liang Y, Lin X, Lu F, Merkulov GV, Milshina N, Moore HM, Naik AK, Narayan VA, Neelam B, Nusskern D, Rusch DB, Salzberg S, Shao W, Shue B, Sun J, Wang Z, Wang A, Wang X, Wang J, Wei M, Wides R, Xiao C, Yan C, Yao A, Ye J, Zhan M, Zhang W, Zhang H, Zhao Q, Zheng L, Zhong F, Zhong W, Zhu S, Zhao S, Gilbert D, Baumhueter S, Spier G, Carter C, Cravchik A, Woodage T, Ali F, An H, Awe A, Baldwin D, Baden H, Barnstead M, Barrow I, Beeson K, Busam D, Carver A, Center A, Cheng ML, Curry L, Danaher S, Davenport L, Desilets R, Dietz S, Dodson K, Doup L, Ferriera S, Garg N, Gluecksmann A, Hart B, Haynes J, Haynes C, Heiner C, Hladun S, Hostin D, Houck J, Howland T, Ibegwam C, Johnson J, Kalush F, Kline L, Koduru S, Love A, Mann F, May D, McCawley S, McIntosh T, McMullen I, Moy M, Moy L, Murphy B, Nelson K, Pfannkoch C, Pratts E, Puri V, Qureshi H, Reardon M, Rodriguez R, Rogers YH, Romblad D, Ruhfel B, Scott R, Sitter C, Smallwood M, Stewart E, Strong R, Suh E, Thomas R, Tint NN, Tse S, Vech C, Wang G, Wetter J, Williams S, Williams M, Windsor S, Winn-Deen E, Wolfe K, Zaveri J, Zaveri K, Abril JF, Guigo R, Campbell MJ, Sjolander KV, Karlak B, Kejariwal A, Mi H, Lazareva B, Hatton T, Narechania A, Diemer K, Muruganujan A, Guo N, Sato S, Bafna V, Istrail S, Lippert R, Schwartz R, Walenz B, Yooseph S, Allen D, Basu A, Baxendale J, Blick L, Caminha M, Carnes-Stine J, Caulk P, Chiang YH, Coyne M, Dahlke C, Mays A, Dombroski M, Donnelly M, Ely D, Esparham S, Fosler C, Gire H, Glanowski S, Glasser K, Glodek A, Gorokhov M, Graham K, Gropman B, Harris M, Heil J, Henderson S, Hoover J, Jennings D, Jordan C, Jordan J, Kasha J, Kagan L, Kraft C, Levitsky A, Lewis M, Liu X, Lopez J, Ma D, Majoros W, McDaniel J, Murphy S, Newman M, Nguyen T, Nguyen N, Nodell M, Pan S, Peck J, Peterson M, Rowe W, Sanders R, Scott J, Simpson M, Smith T, Sprague A, Stockwell T, Turner R, Venter E, Wang M, Wen M, Wu D, Wu M, Xia A, Zandieh A, Zhu X (2001) The sequence of the human genome. Science 291:1304–1351

Wheeler DL, Church DM, Lash AE, Leipe DD, Madden TL, Pontius JU, Schuler GD, Schriml LM, Tatusova TA, Wagner L, Rapp BA (2002) Database resources of the National Center for Biotechnology Information: 2002 update. Nucleic Acids Res 30:13–16

Yamada R, Tanaka T, Ohnishi Y, Suematsu K, Minami M, Seki T, Yukioka M, Maeda A, Murata N, Saiki O, Teshima R, Kudo O, Ishikawa K, Ueyosi A, Tateishi H, Inaba M, Goto H, Nishizawa Y, Tohma S, Ochi T, Yamamoto K, Nakamura Y (2000) Identification of 142 single nucleotide polymorphisms in 41 candidate genes for rheumatoid arthritis in the Japanese population. Hum Genet 106:293–297