Quality and completeness of SNP databases

Published online 24 March 2003; doi:10.1038/ng1133

To address the quality and completeness of single-nucleotide polymorphism (SNP) databases, we resequenced 173 kb (spanning 17 loci) in 150 chromosomes of west African and European ancestry. Over 88% of SNPs in the public (TSC and BAC overlap) and Celera databases were confirmed in independent resequencing. Approximately 45% of all human heterozygosity is attributable to SNPs already available from the two databases, and of SNPs with minor-allele frequencies >10%, more than half are represented.

To assess the quality and completeness of the approximately four million candidate SNPs in the public and Celera databases[1–3], we resequenced a multiethnic sample of sufficient depth to capture most common variants (>1% in frequency) across a representative sample of the genome. We studied segments drawn from 17 loci[4] with a total of 173 kb of sequence examined in each individual. We included in this analysis only those nucleotide positions that were sequenced at high quality (Phred[5] score ≥25) in at least 40 European American and 40 west African chromosomes (we surveyed an average of 76 European American and 74 west African chromosomes at each nucleotide position). In total, we identified 923 variant sites and estimated overall heterozygosity to be $\pi = 7.1 \quad 10^{-4}$, consistent with the genome-wide average (refs. 1,2; see Supplementary Note 1 online for a description of how π was calculated). Identification of SNPs by sequencing was blind to the positions of SNPs in the databases.

We first determined the fraction of SNPs in the public and Celera databases that had independently been identified in our deep resequencing. We made this assessment for (i) SNPs identified by Celera alone[2], (ii) SNPs identified by the public SNP discovery efforts (TSC and BAC overlap only) and (iii) candidate SNPs identified as differences between the Celera and public reference sequences[2]. We did not include public SNPs discovered by efforts other than The SNP Consortium and BAC overlap projects in this analysis because other SNPs were discovered using a variety of experimental approaches, depth of resequencing and analytic methods. Of 213 SNPs identified in our regions from these databases, 188 (88%) were independently observed in our resequencing study, confirming that most of the SNPs identified as above are true variant sites and thus can be useful in genetic studies (Table 1). The estimated validation rate is even higher, 94%, if one of the 17 loci in our study (*ACVR2B*) is removed from analysis. At this locus, a single BAC accession in the public database (AP000500) contributes a high fraction of the SNPs not identified by our deep resequencing; this may be due to a rare haplotype sequenced by the human genome project or an error of sequencing or genome assembly.

The 6–12% rate of non-validation of database SNPs in our survey represents the total of errors in the construction of SNP databases (false-positive SNPs), rare variants not observed again in a survey of 150 chromosomes, population-specific SNPs (not present in the European American and African American samples) and true SNPs that were missed in resequencing (false-negative SNPs). Based on this study[6] and genotyping of 3,738 TSC SNPs with a validation rate of 89% (see also ref.3), we estimate that no more than 12% of SNPs in the TSC database represent rare variants or false positives. In this issue, Carlson *et al.*[7] report a resequencing survey in which 28–35% of TSC SNPs were not identified. The higher rate of failure to detect TSC SNPs in their survey could reflect true SNPs missed by their sequencing effort or a systematic difference between the 50 genes that they studied and the 68 regions in ref. 6 and this report. These possibilities can be distinguished by using an independent method to genotype in a broad population sample the TSC SNPs not discovered in resequencing studies.

We next assessed the proportion of all genetic variation across the 173 kb that was attributable to SNPs already present in the databases. Of the overall heterozygosity identified by complete resequencing ($7.1 \quad 10^{-4}$ per bp), 45% was due to variants already present in the databases. Moreover, of SNPs with frequency ≥10%, 54% were already present in the databases (Fig. 1). These data support population genetic predictions[8] of the number of common variants in the human genome, indicating that there are only about 5 million SNPs with minor-allele frequency worldwide of >10% and about 10 million with frequency >1%. We note that even though most common human genetic variation is shared across populations[9,10], the number of SNPs with a frequency >1% in any single population should be somewhat larger than that number in a multiethnic sample, such as we have studied.

Table 1 • Quality and percentage of human variation captured in the SNP databases

Method of identification	Number of SNPs identified	Percentage validated by resequencing	Percentage of heterozygosity captured	Percentage of SNPs in the database[a,b] that validate with frequency >10%
Celera	118	95% (95%[b])	29%	75%
Public (union of TSC and BAC overlap)	105	83% (95%[b])	21%	71%
TSC	54	94% (98%[b])	12%	73%
BAC	69	74% (93%[b])	13%	69%
Sites divergent between Celera & HGP reference sequences	51	82% (95%[b])	10%	75%
All SNPs (union of above)[c]	213	88% (94%[b])	45%	70%
Celera and public double-hit SNPs (intersection of the two databases)	30	100% (100%[b])	9%	97%

[a]As a fraction of all SNPs in the database (including those that do not validate in resequencing data). [b]SNPs from the *ACVR2B* locus are removed from these calculations. [c]The 'All SNPs' category includes polymorphisms from three sources: the Celera SNP database, the public database and divergent sites between the Celera and Human Genome Project (HGP) sequences. Together the SNPs in the individual databases sum to 274, but because of overlaps, there are only 213 unique entries in the 'All SNPs' category. The percentage of SNPs that validate with frequency >10% is higher in the individual databases (71–75%) than in the 'All SNPs' category (70%). This is because the 'All SNPs' category contains a larger proportion of SNPs with frequencies less than 10%. The slightly lower rate of validation in the 'All SNPs' category is due to the fact that a more complete database will yield a higher proportion of lower frequency SNPs.

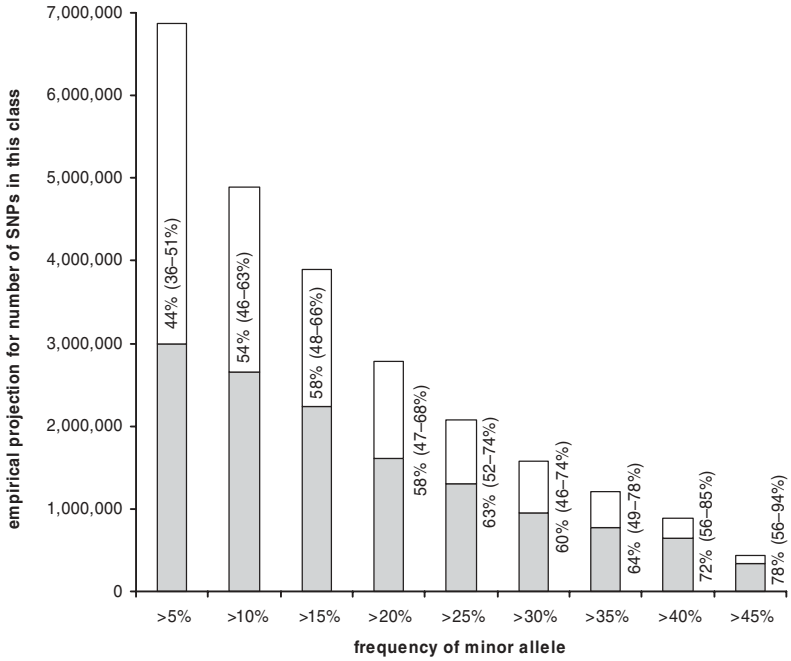

Fig. 1 Estimated number of all SNPs in the human genome that are already represented in the combined Celera and public SNP databases as a function of minor-allele frequency. Estimates for the entire genome were obtained by extrapolating from the number of SNPs observed over the 173 kb that we surveyed by deep resequencing to obtain the number expected over a genome of 3.29 billion bp[12]. The estimated percentage of SNPs in each frequency class that have already been identified by the two databases is labeled (with 95% central confidence intervals) and shaded in gray.

The planned construction of a haplotype map of the human genome will involve genotyping a high density of common SNPs to identify ancestral relationships and allelic associations[6], which is required to guide selection of markers for haplotype-based association studies. Our analysis indicates that over half of SNPs with frequency >10% are already represented in available databases and hence are a valuable resource for haplotype mapping. Despite the high quality and substantial completeness of the current SNP databases, however, there is a practical challenge in their use: at present, the databases contain limited information about population frequency. Because almost a third of all database SNPs have a minor-allele frequency <10% or are false positives (Table 1), considerable resources will have to be expended to genotype SNPs that are less useful for identifying common haplotypes because of their low frequency.

We and others have noted that the SNPs for which both alleles have been seen more than once in the process of SNP discovery ('double-hit' SNPs) have more desirable properties for initial construction of haplotype maps (see also Carlson *et al.*[7], this issue). In the current data, we identified those SNPs that were seen independently by both Celera and the public SNP discovery efforts and found that these validated at a very high rate, with 97% having a minor-allele frequency >10%. In a second, larger survey, we used mass spectrometry–based genotyping[11] to evaluate 204 double-hit SNPs in the same west African and European American populations. This analysis showed that 99% of double-hit SNPs were validated, with 91% having a minor-allele frequency >10% in the combined sample. These numbers are much more favorable than those seen with 'single-hit' SNPs (88–94% validation rate, with 70% of all candidate SNPs having allele frequencies >10%; Table 1; see also ref. 6). Thus, we suggest that an ideal SNP database for identifying common haplotypes across the genome is one in which there is a publicly available map of double-hit SNPs that covers the genome at high density (see also Carlson *et al.*[7], this issue). This will require additional effort in SNP discovery, but the cost savings in terms of marker development (as well as identification of SNPs not yet in the databases) will pay rich dividends for public and private projects to chart out patterns of variation anywhere in the genome.

URL. The SNP Consortium website can be found at http://snp.cshl.org.

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

Acknowledgments
We are grateful to D. Richter for computer support; J. Roy for assistance with database searches; B. Ferrell for the Beni samples; J. Platko, T. Lavery, A. Rachupka, T. Takahashi, G. McDonald and K. Sunter for assistance with DNA sequencing and gel scoring; and M. Daly, J. Mullikin and D. Cutler for comments on the analysis and manuscript.

Competing interests statement
The authors declare that they have no competing financial interests.

David E. Reich[1], Stacey B. Gabriel[1] & David Altshuler[1,2]

[1]Program in Medical and Population Genetics, Whitehead Institute / MIT Center for Genome Research, One Kendall Square, Cambridge, Massachusetts 02139, USA. [2]Departments of Genetics and Medicine, Harvard Medical School and Department of Molecular Biology and Diabetes Unit, Massachusetts General Hospital, Boston, Massachusetts, USA. Correspondence should be addressed to D.E.R. (e-mail: reich@genome.wi.mit.edu).

Received 9 September 2002; accepted 27 February 2003.

1. Sachidanandam, R. *et al. Nature* **409**, 928–933 (2001).
2. Venter, J.C. *et al. Science* **291**, 1304–1351 (2001).
3. Marth, G. *et al. Nat. Genet.* **27**, 371–372 (2001).
4. Reich, D.E. *et al. Nature* **411**, 199–204 (2001).
5. Nickerson, D.B., Tobe, V.O. & Taylor, S.L. *Nucleic Acids Res.* **25**, 2745–2751 (1997).
6. Gabriel, S.B. *et al. Science* **296**, 2225–2229 (2002).
7. Carlson, C.S. *et al. Nat. Genet.* **33**; advance online publication 24 March 2003; doi:10.1038/ng1128
8. Kruglyak, L. & Nickerson, D.A. *Nat. Genet.* **27**, 234–236 (2000).
9. Lewontin, R.C. *Evol. Biol.* **6**, 381–398 (1972).
10. Rosenberg, N.A. *et al. Science* **298**, 2381–2385 (2002).
11. Tang, K. *et al. Proc. Natl. Acad. Sci. USA* **96**, 10016–10020 (1999).12. Lander, E.S. *et al. Nature* **409**, 860–921 (2001).
12. Lander, E.S. *et. al. Nature* **409**, 8 60–921 (2001).