A general approach to single-nucleotide polymorphism discovery

Gabor T. Marth[1], Ian Korf[1], Mark D. Yandell[1], Raymond T. Yeh[1], Zhijie Gu[2], Hamideh Zakeri[2], Nathan O. Stitziel[1], LaDeana Hillier[1], Pui-Yan Kwok[2] & Warren R. Gish[1]

Single-nucleotide polymorphisms (SNPs) are the most abundant form of human genetic variation and a resource for mapping complex genetic traits[1]. The large volume of data produced by high-throughput sequencing projects is a rich and largely untapped source of SNPs (refs 2–5). We present here a unified approach to the discovery of variations in genetic sequence data of arbitrary DNA sources. We propose to use the rapidly emerging genomic sequence[6,7] as a template on which to layer often unmapped, fragmentary sequence data[8–11] and to use base quality values[12] to discern true allelic variations from sequencing errors. By taking advantage of the genomic sequence we are able to use simpler yet more accurate methods for sequence organization: fragment clustering, paralogue identification and multiple alignment. We analyse these sequences with a novel, Bayesian inference engine, POLYBAYES, to calculate the probability that a given site is polymorphic. Rigorous treatment of base quality permits completely automated evaluation of the full length of all sequences, without limitations on alignment depth. We demonstrate this approach by accurate SNP predictions in human ESTs aligned to finished and working-draft quality genomic sequences, a data set representative of the typical challenges of sequence-based SNP discovery.

We started with 1,268,211 bp finished (less than 1 error per 10,000 bp) human reference sequence of 10 genomic clones, with EST content typical of gene-bearing clones. To initiate the analysis procedure (Fig. 1) to identify human ESTs that originated from these clones, we performed a database search against the public EST set (dbEST) and recovered 1,954 hits (representing potentially multiple exons of 1,365 unique ESTs) for which chromatograms were available. Sequence clusters were constructed as groups of overlapping alignments (147 clusters). Sequence traces were re-processed with the PHRED base-calling program[13,14] to obtain base quality values. Subsequent analyses used the full length of the ESTs, including low-quality portions. Cluster members were multiply aligned with an anchored alignment technique. Unlike traditional algorithms, this method rapidly produces correct multiple alignments even in the presence of abundantly expressed or alternatively spliced transcripts. In total, EST clusters represented 80,469 bp of expressed genomic sequence, 38% of this in regions of single EST coverage and 81% in regions covered by 8 or fewer ESTs (Table 1).

Inclusion of sequences representing highly similar regions duplicated elsewhere in the genome may give rise to false SNP predictions, and the presence of such sequence paralogues points to difficulties during marker development. We devised a Bayesian[15]

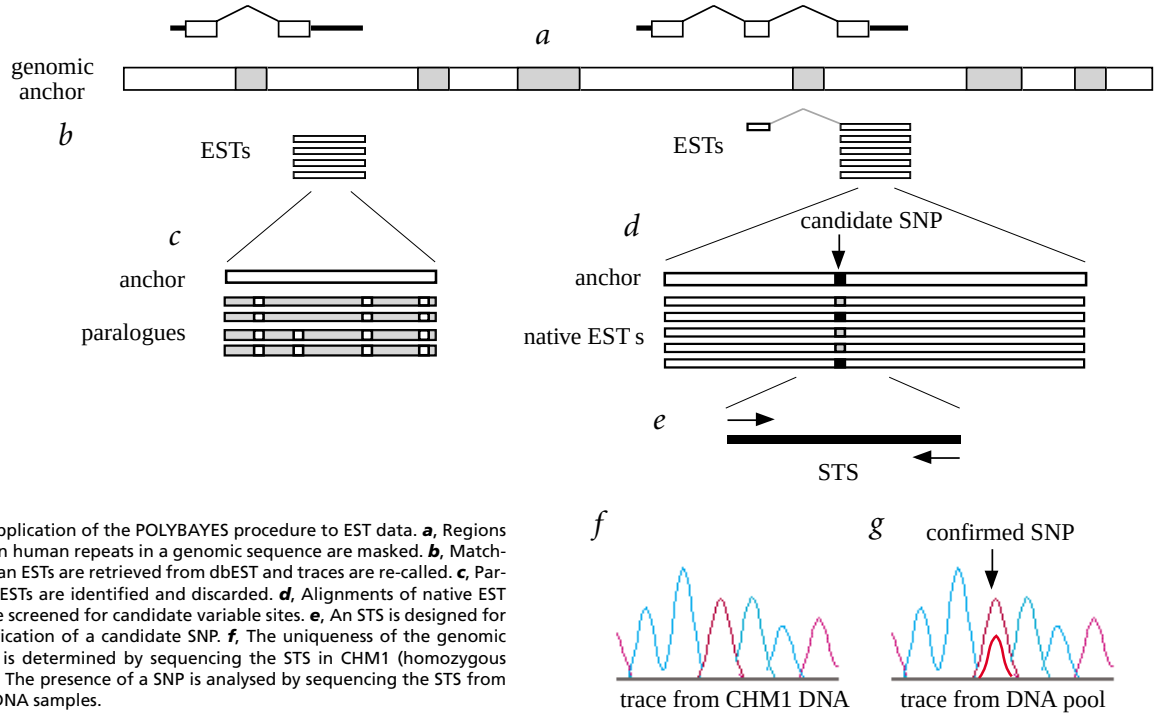

Fig. 1 Application of the POLYBAYES procedure to EST data. ***a***, Regions of known human repeats in a genomic sequence are masked. ***b***, Matching human ESTs are retrieved from dbEST and traces are re-called. ***c***, Paralogous ESTs are identified and discarded. ***d***, Alignments of native EST reads are screened for candidate variable sites. ***e***, An STS is designed for the verification of a candidate SNP. ***f***, The uniqueness of the genomic location is determined by sequencing the STS in CHM1 (homozygous DNA). ***g***, The presence of a SNP is analysed by sequencing the STS from pooled DNA samples.

Washington University [1]Department of Genetics and Genome Sequencing Center and [2]Division of Dermatology, St. Louis, Missouri, USA. Correspondence should be addressed to G.T.M. (e-mail: gmarth@watson.wustl.edu) or P.-Y.K. (e-mail: kwok@im.wustl.edu).

Table 1 • SNP discovery in EST alignments of varying coverage

	No. of clusters		No. of aligned sites			Distribution of SNPs		
Depth[a]	before paralogue filtering[b]	after paralogue filtering[c]	before paralogue filtering[d]	after paralogue filtering[e]	Candidate[f]	analysed[g]	confirmed[h]	Confirmation rate[i]
1	47 (32.0%)	40 (32.0%)	30,828 (38.3%)	26,275 (37.7%)	12 (22.2%)	6 (16.7%)	5 (25.0%)	83%
2	25 (17.0%)	24 (19.2%)	15,771 (19.6%)	15,072 (21.6%)	8 (14.8%)	7 (19.4%)	2 (10.0%)	29%
3,4	23 (15.6%)	21 (16.8%)	12,478 (15.5%)	9,937 (14.2%)	17 (31.5%)	8 (22.2%)	5 (25.0%)	63%
5–8	17 (11.6%)	14 (11.2%)	6,627 (8.2%)	5,467 (7.8%)	7 (13.0%)	7 (19.4%)	1 (5.0%)	14%
9–16	14 (9.5%)	8 (6.4%)	7,704 (9.6%)	6,383 (9.2%)	3 (5.5%)	3 (8.4%)	3 (15.0%)	100%
17 or more	21 (14.3%)	18 (14.4%)	7,061 (8.8%)	6,662 (9.5%)	7 (13%)	5 (13.9%)	4 (20.0%)	80%
Total	147 (100%)	125 (100%)	80,469 (100%)	69,756 (100%)	54 (100%)	36 (100%)	20 (100%)	Overall 56%

[a]Depth of coverage (or cluster size), not including the genomic reference sequence. [b]Number of clusters of given cluster size before removal of paralogous ESTs. [c]Number of clusters of given cluster size after removal of paralogous ESTs. [d]Number of sites of given alignment depth in multiple alignments before removal of paralogous ESTs. [e]Number of sites of given alignment depth in multiple alignments after removal of paralogous ESTs. [f]Number of candidate SNPs found at sites of given alignment depth. [g]Number of unambiguously analysed candidate SNPs. [h]Number of SNPs confirmed in at least one of four population pools. [i]SNP confirmation rate. [b–i]Numbers in parentheses indicate percentages of relevant total.

discrimination algorithm (Fig. 2*a*) that takes into account base quality values to calculate the probability, P_{NAT}, that a cluster member is native to (derived from) the given genomic region. The bimodal distribution of these probability values (Fig. 2*b*) indicates that we can distinguish between less accurate sequences that nevertheless originate from the same underlying genomic location, and more accurate sequences with high-quality discrepancies that are likely to be paralogous. Using a conservative threshold value, $P_{NAT,MIN}$, of 0.75, 23% of cluster members were declared paralogous and removed from further consideration, leaving 69,756 sites of native EST coverage.

Once a proper data set is organized, the key to reliable detection of SNPs is the ability to discern true allelic variation from sequencing error. To this end, we have developed a Bayesian-statistical model for the mathematically rigorous treatment of sequence differences within a multiple alignment that takes into account the depth of coverage, the base quality values of the sequences and the a priori expected rate of polymorphic sites in the region. For each site within a multiple alignment of native sequences, the POLYBAYES algorithm calculates the probability, P_{SNP}, that the site is polymorphic, as opposed to monomorphic. The distribution of probability scores (Fig. 3*a*) exhibits a high level of specificity: most sites (99.83%) produce scores below 0.1. They represent sites either with no disagreements between aligned sequences or with low-quality discrepancies that are likely the result of sequencing errors or possibly very rare SNPs. By marking a site as a candidate SNP if the corresponding SNP probability exceeded a threshold value, $P_{SNP,MIN}$, of 0.40, we extracted 97 candidates. Of these, 38 were located in adenine-rich regions of the genomic clones matching the 3′ ends of ESTs. Subsequent negative verification results are consistent with the hypothesis[16] that these sites result from internal priming events during cDNA library construction and that the adenine allele is contributed by the reverse transcription primer rather than the RNA template.

We validated candidate sites with a pooled sequencing approach[17] that allowed us to confirm true positives, provided the minor allele frequency was above 10%. We eliminated five candidates that did not fulfil this requirement. An additional 18 sites could not be analysed for lack of unique amplification (9 candidates in regions of low complexity or repetitive sequence, 4 candidates for unknown reasons and in 5 cases, the homozygous control genome[18] indicated the presence of paralogues absent in the EST set). Of the remaining 36 sites, 20 were confirmed in at least 1 of 4 populations screened (13 transitions, 7 transversions), yielding a 56% overall confirmation rate.

The confirmation rate is somewhat lower than the average SNP score of 0.78. Some of this difference may be due to systematic base-calling errors (compressions) and reverse transcriptase errors introduced during cDNA library construction. Several of the candidate sites may be true polymorphisms specific to the donors of the cDNA samples but absent in the population pools used in verification. Although precise calibration of the SNP probability values would require analysing the genomic source of

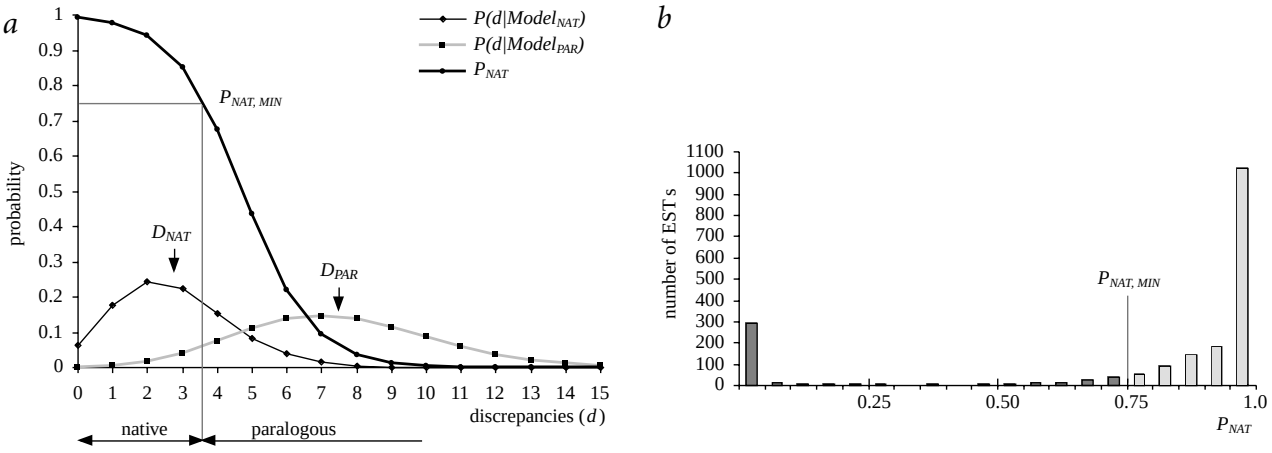

Fig. 2 Paralogue discrimination. ***a***, Example probability distributions for a matching sequence with (hypothetical) uniform base quality values of 20, in pair-wise alignment with base perfect genomic anchor sequence (quality values 40), over a length of 250 bp. $P_{POLY,2}$ = 0.001, P_{PAR} = 0.02, E=2.525, D_{NAT} = 2.775 and D_{PAR} = 7.525. If the posterior probability, P_{NAT}, is higher than $P_{NAT,MIN}$, the EST is considered native; otherwise, it is considered paralogous. ***b***, Distribution of the posterior probability values, P_{NAT}, calculated for 1,954 cluster members anchored to ten genomic clone sequences.

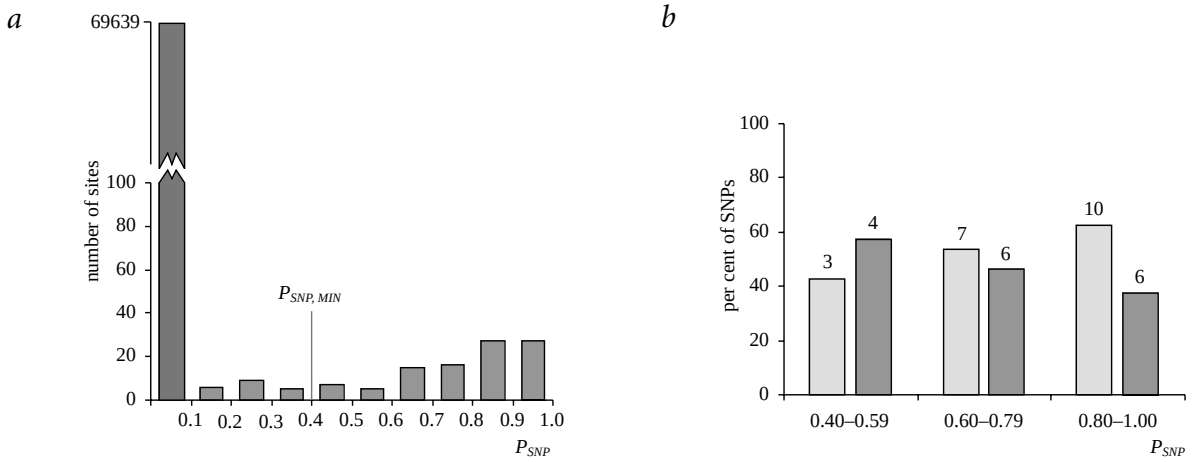

Fig. 3 SNP probability scores. ***a***, Distribution of the posterior probability value that a site is polymorphic, P_{SNP}, for 69,756 sites in multiple alignments of native ESTs. ***b***, Correlation between P_{SNP} score and confirmation rate. The fraction of confirmed candidate SNPs (striped bars) and the fraction of candidate SNPs that were not detected in population-specific DNA pools (shaded bars) are shown. The absolute number of SNPs is shown above each bar.

each EST, an undertaking beyond the scope of this study, higher SNP probability scores correspond to higher confirmation rates (Fig. 3*b*). This is the true significance of the SNP score: it enables one to strike a balance between true positive rates and the recovery of low-frequency alleles. Using a higher detection threshold reduces the number of false positives, but also discards more true polymorphic sites. Conversely, the recovery of rare SNPs requires a lower threshold, which in turn increases the false-positive rate, reflective of the fact that rare alleles or alleles in low-quality sequence are indistinguishable from sequencing error. The sensitivity of the algorithm as a function of allele frequency, sequence quality, alignment depth and SNP probability threshold are reported (Fig. 4). The algorithm successfully detected variations in clusters containing a single EST aligned to the reference sequence (five confirmed sites), indicating that POLYBAYES is effective even in very shallow alignments (Table 1). For the same reasons, our mining efficiency (1 candidate per 25 ESTs and 1 confirmed SNP per 68 ESTs analysed) compares favourably with recently published results[4,5].

During verification of candidates, we found only two novel SNPs in 11,455 bp of STS sequence. One SNP was outside an EST cluster and could not have been found in the data set. The other one was a rare variation present in one of four sampled populations, but not within the EST cluster members. The dearth of novel SNPs unique to the population pools suggests that the ESTs contained most common variations in the analysed regions, and that POLYBAYES successfully detected them.

We evaluated the performance of POLYBAYES with assembled shotgun, 'working-draft' quality genomic reference sequence. To this end, we simulated clone sequences of 2–6-fold shotgun coverage by reassembling random subsets of the original shotgun reads for 5 of 10 clones with the PHRAP (P. Green, unpublished data) fragment assembler. Using the resulting contig sequences as a reference, we repeated the subsequent SNP analysis with unchanged parameters (Fig. 5). Even at threefold shotgun coverage, an average 94% of ESTs were identified and 81% of confirmed SNPs detected (respectively, 98% and 94% at fivefold coverage), indicating that POLYBAYES does not require base-perfect reference sequence to be effective and will work well with draft-quality sequences that have begun to dominate sequence production[19].

Because expressed regions comprise but a small fraction of the genome, polymorphic sites recovered from ESTs alone, however valuable, are

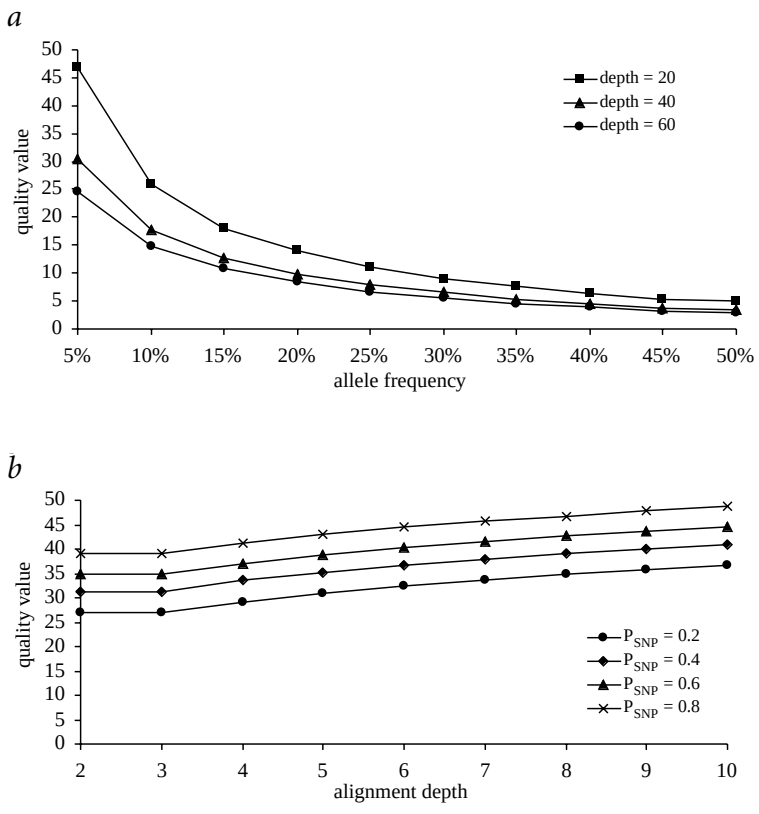

Fig. 4 Sensitivity of the SNP detection algorithm. ***a***, Minimum base quality requirement for the detection of minor alleles of a given frequency, in alignments of depth N=20, 40, 60, at a detection threshold value $P_{SNP,MIN}$ = 0.40. ***b***, Base quality requirement for the detection of a single minor allele in alignments of depth N = 2,...,10, and SNP probability threshold values $P_{SNP,MIN}$ = 0.20, 0.40, 0.60 and 0.80. In (*a*,*b*), the quality value for each base was assumed to be uniform.

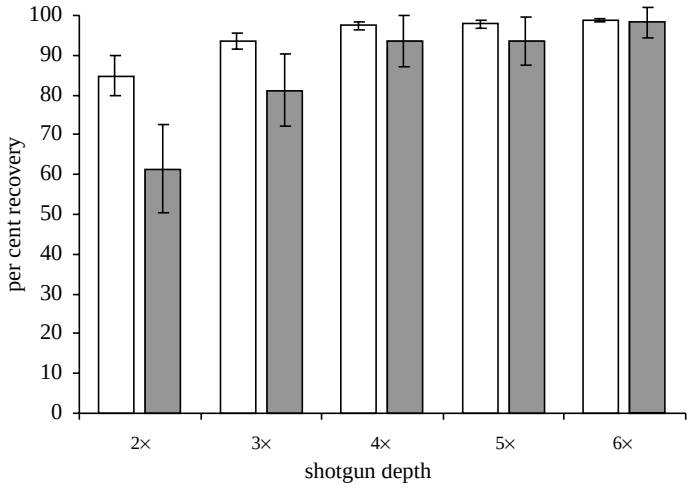

Fig. 5 SNP detection with assembled shotgun genomic reference sequence. Fractions of ESTs recovered (white bars) and SNPs recovered (grey bars) are shown. Percentages were based on the 733 ESTs anchored by 5 of 10 genomic clones in the primary experiment, and the 14 confirmed SNPs detected among these sequences. Error bars indicate standard deviation among 20 consecutive experiments.

unlikely to produce a SNP marker map of uniformly high density. Through the coordinated efforts of large-scale sequencing efforts worldwide, the nearly complete sequence of the human genome will soon be available, augmented by the generation of a staggering amount of fragmentary sequences. Our study demonstrates that through precise treatment of the data, combined with objective evaluation of data quality, it is possible to discover variations in these sequences with great efficiency, contributing to the creation of valuable resources[20–22] with which to analyse complex genetic traits and further our understanding of human origins.

Methods

Data organization. Known human repeats in the genomic sequences were masked with RepeatMasker (A.F.A. Smit and P. Green, unpublished data) and searched against dbEST with WU-BLAST (W.R.G., http://blast.wustl.edu) with parameters: M=5, *N*=–11, Q=11, *R*=11, S=170, gapS2=150, filter=seg (*P*-value cutoff 10^{-50}). Sequence traces that were available at the Washington University ftp site (ftp://genome.wustl.edu/pub/gsc1/est) were processed with the PHRED base-calling program; the full length of each sequence, together with base quality values (expressing the likelihood that the called nucleotide is incorrect), was used in the subsequent analysis. Distinct groups of matching ESTs were registered as clusters. Each cluster member was first pair-wise aligned to the genomic anchor sequence with CROSS_MATCH (P. Green, unpublished data). We then produced a multiple alignment by propagating gaps and insertions in the pair-wise alignments into all remaining sequences, a procedure known as 'sequence padding'. The computational complexity of the algorithm grows linearly with the length and number of sequences.

Paralogue identification. We identified paralogous sequences by determining if the number of mismatches observed between the genomic reference sequence and a matching EST was consistent with polymorphic variation as opposed to sequence difference between duplicated chromosomal locations, taking into account sequence quality. On the basis of our annotation experience of over 40 Mb of genomic sequence, we stipulated that most 'paralogous' sequences exhibit a pair-wise dissimilarity rate higher than $P_{PAR} = 0.02$ (2%) compared with the average pair-wise polymorphism rate, $P_{POLY,2} = 0.001$ (0.1%). In a pair-wise match of length L, we expect $L \times P_{POLY,2}$ mismatches due to polymorphism, versus $L \times P_{PAR}$ mismatches due to paralogous difference. In both cases, an additional number, E, of mismatches are expected to arise from sequencing errors, approximated as the sum of base error probabilities of both sequences (calculated from the base quality values) along the pair-wise alignment. We considered two models: an EST is either native ($Model_{NAT}$) and we expect $D_{NAT} = L \times P_{POLY,2} + E$ discrepancies, or it is paralogous ($Model_{PAR}$) and we expect $D_{PAR} = L \times P_{PAR} + E$ mismatches. The probability of observing d discrepancies in the pair-wise alignment is approximated by a Poisson distribution, with parameter $\lambda = D_{NAT}$ for $Model_{NAT}$ and $\lambda = D_{PAR}$ for $Model_{PAR}$. In absence of reliable a priori knowledge of the expected proportions of native versus paralogous ESTs, we used uninformed (flat) priors. The posterior probability, $P_{NAT} = P(Model_{NAT}|d)$, that the EST represents native sequence was determined as:

$$P(Model_{NAT} \mid d) = \frac{1}{1 + e^{(D_{NAT} - D_{PAR})} \cdot \left(D_{PAR} / D_{PAR} \right)}$$

ESTs that scored above a cutoff value, $P_{NAT,MIN}$, were considered native; sequences scoring below the threshold were declared paralogous.

SNP detection in multiple alignments. The algorithm identifies polymorphic locations by evaluating the likelihood of nucleotide heterogeneity within cross-sections of a multiple alignment. Each of the nucleotides, $S_1,\ldots,S_N$, in such a cross-section of N sequences, $R_1,\ldots,R_N$, can be any one of the four DNA bases, for a total of 4^N nucleotide permutations. The likelihood, $P(S_i|R_i)$, that a nucleotide, S_i, is A, C, G or T is estimated from the error probability, $P_{Error,i}$, obtained from the base quality value. We assign $(1–P_{Error,i})$ to the called base and $(P_{Error,i}/3)$ to each of the three uncalled bases. In the absence of likelihood estimates, insertions and deletions are not considered. Each heterogeneous (polymorphic) permutation is classified according to its nucleotide multiplicity, the specific variation and the distribution of alleles. We used the value P_{POLY}=0.003 (1 polymorphic site in 333 bp) as the total a priori probability that a site is polymorphic[21,22] (1/1,000 polymorphism rate between any pair of sequences). This value was distributed to assign a prior probability, $P_{Prior}(S_1,\ldots,S_N)$, to each permutation. Permutations of higher nucleotide multiplicities received exponentially lower shares, in accordance with a random allele generation model. In this study, we assigned equal shares to different variation types (although unequal shares can be specified in the software to account for a higher rate of transitions compared to transversions). A prior value of $(1–P_{POLY})/4$ was assigned to each of the four non-polymorphic permutations, corresponding to a uniform base composition, $P_{Prior}(S_i)$. The Bayesian posterior probability of a particular nucleotide permutation was calculated through another application of Bayesian inference, considering the 4^N different permutations as the set of conflicting models:

$$P(S_1,\ldots,S_N \mid R_1,\ldots,R_N) = \frac{\frac{P(S_1 \mid R_1)}{P_{Prior}(S_1)} \cdot \ldots \cdot \frac{P(S_N \mid R_N)}{P_{Prior}(S_N)} \cdot P_{Prior}(S_1,\ldots,S_N)}{\sum_{every\ (S_{i1},\ldots,S_{iN})} \frac{P(S_{i1} \mid R_1)}{P_{Prior}(S_{i1})} \cdot \ldots \cdot \frac{P(S_{iN} \mid R_1)}{P_{Prior}(S_{iN})} \cdot P_{Prior}(S_{i1},\ldots,S_{iN})}$$

The Bayesian posterior probability of a SNP, P_{SNP}, is the sum of posterior probabilities of all heterogeneous permutations. The computation is performed with an efficient, recursive algorithm. A site within a multiple alignment is reported as a candidate SNP if the corresponding posterior probability exceeds a set threshold value, $P_{SNP,MIN}$. We examined the sensitivity of the detection algorithm under the simplifying assumption of uniform base quality. We determined the relationship between observed minor allele frequency and base quality to produce a SNP probability score $P_{SNP} = 0.4$ (the threshold value used in this study), in alignments of various depths of coverage. We also determined the minimum base quality value required for detecting a single minor allele in alignments of ten or fewer sequences at various threshold values.

Software. POLYBAYES was developed in a UNIX environment and runs efficiently on a conventional workstation. Sequence clustering is performed with custom scripts. The anchored alignment, paralogue filtering and SNP detection are accessed through a single program. SNP locations and probabilities are reported in text files or as a database compatible with the CONSED sequence editor[23] to enable viewing the multiple alignments,

quality values, sequence traces and annotated SNPs *in toto*. Instructions for obtaining POLYBAYES are available (at no cost for non-profit use, see http://genome.wustl.edu/gsc/polybayes).

Accession numbers. 127H14, NID:g2439515; DJ0604G05, NID:g3006227; DJ0777O23, NID:g3242763; DJ327A19, NID:g2341021; GS345D13, NID:g2078461; GS541B18, NID:g2781380; GS542D18, NID:g2388554; RG085C05, NID:g1669367; RG104I04, NID:g1809226; RG119C02, NID:g3004572. Confirmed SNPs were submitted to dbSNP, NCBI assay ID 4277–4281, 4618–4628, 4643–4648.

Acknowledgements

We thank T. Blackwell and S. Eddy for informative discussions during the development of the mathematical framework of the technique. This work was supported by NIH grants P50HG01458 (L.H. and W.R.G.), R01HG1720 (P.-Y.K.) and T32AR07284 (Z.G.), and an equipment loan from Compaq Computer Corporation.

Received 17 August; accepted 18 October 1999.

1. Collins, F.S., Guyer, M.S. & Chakravarti, A. Variations on a theme: cataloging human DNA sequence variation. *Science* **278**, 1580–1581 (1997).
2. Wang, D.G. *et al*. Large-scale identification, mapping, and genotyping of single nucleotide polymorphisms in the human genome. *Science* **280**,1077–1082 (1998).
3. Taillon-Miller, P., Gu, Z., Hillier, L. & Kwok, P.-Y. Overlapping genomic sequences: a treasure trove of single-nucleotide polymorphisms. *Genome Res.* **8**, 748–754 (1998).
4. Picoult-Newberg, L. *et al*. Mining SNPs from EST databases. *Genome Res.* **9**, 167–174 (1999).
5. Buetow, K.H., Edmondson, M.N. & Cassidy, A.B. Reliable identification of large numbers of candidate SNPs from public EST data. *Nature Genet.* **21**, 323–325 (1999).
6. The Sanger Centre & The Washington University Genome Sequencing Center. Toward a complete human genome sequence. *Genome Res.* **8**, 1097–1108 (1998).
7. Venter, J.C. *et al*. Shotgun sequencing of the human genome. *Science* **280**, 1540–1542 (1998).
8. Hillier, L. *et al*. Generation and analysis of 280,000 human expressed sequence tags. *Genome Res.* **6**, 807–828 (1996).
9. Adams, M.D., Soares, M.B., Kerlavage, A.R., Fields, C. & Venter, J.C. Rapid cDNA sequencing (expressed sequence tags) from a directionally cloned human infant brain cDNA library. *Nature Genet.* **4**, 373–380 (1993).
10. Hudson, T.J. *et al*. An STS-based map of the human genome. *Science* **270**, 1945–1954 (1995).
11. Marra, M., Weinstock, L.A. & Mardis, E.R. End sequence determination from large insert clones using energy transfer fluorescent primers. *Genome Res.* **6**, 1118–1122 (1996).
12. Durbin, R. & Dear, S. Base qualities help sequencing software. *Genome Res.* **8**, 161–162 (1998).
13. Ewing, B., Hillier, L., Wendl, M.C. & Green, P. Base-calling of automated traces using Phred. I. Accuracy assessment. *Genome Res.* **8**, 175–185 (1998).
14. Ewing, B. & Green, P. Base-calling of automated traces using Phred. II. Error probabilities. *Genome Res.* **8**, 186–194 (1998).
15. Bayes, T. An essay towards solving a problem in the doctrine of chances. *Philos. Trans. R. Soc.* **53**, 370–418 (1763). Reprinted in *Biometrika* **45**, 293–315 (1958).
16. Aaronson, J. *et al*. Toward the development of a gene index to the human genome: an assessment of the nature of high-throughput EST sequence data. *Genome Res.* **6**, 829–845 (1996).
17. Kwok, P.-Y., Carlson, C., Yager, T., Ankener, W. & Nickerson, D.A. Comparative analysis of human DNA variations by fluorescence-based sequencing of PCR products. *Genomics* **23**, 138–144 (1994).
18. Taillon-Miller, P. *et al*. The homozygous complete hydatidiform mole: a unique resource for genome studies. *Genomics* **46**, 307–310 (1997).
19. Collins, F.S. *et al*. New goals for the U.S. Human Genome Project: 1998–2003. *Science* **282**, 682–689 (1998).
20. Nickerson, D.A. *et al*. DNA sequence diversity in a 9.7-kb region of the human lipoprotein lipase gene. *Nature Genet.* **19**, 233–240 (1998).
21. Cargill, M. *et al*. Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nature Genet.* **22**, 231–238 (1999).
22. Halushka, M.K. *et al*. Patterns of single-nucleotide polymorphisms in candidate genes regulating blood-pressure homeostasis. *Nature Genet.* **22**, 239–247 (1999).
23. Gordon, D., Abajian, C. & Green, P. Consed: a graphical tool for sequence finishing. *Genome Res.* **8**, 195–202 (1998).