Validation of single-nucleotide polymorphisms: A population genetics limitation for the observed success rates.

[1,3]Zhiyong Yang, [1,3]Gane Ka-Shu Wong, [2,3]Michael A. Eberle, [1]Miho Kibukawa, [1]Douglas A. Passey, [1]William R. Hughes, [2]Leonid Kruglyak, and [1,4]Jun Yu.

[1]Human Genome Center. Department of Medicine. University of Washington. Seattle, WA 98195. [2]Division of Human Biology. Fred Hutchinson Cancer Research Center. Seattle, WA 98109. [3]These authors contributed equally to this work.

[4]Corresponding author. E-MAIL junyu@u.washington.edu; FAX (206) 685-7344.

Single-nucleotide-polymorphisms (SNPs) are pouring into the public databases from several large-scale discovery projects. Available resources allow only a small fraction of these SNPs to be validated in a second population. It is therefore imperative that we understand what factors govern validation success. By analyzing a validation experiment performed on 436 putative SNPs identified from the EST databases, we demonstrate that population genetics is at least as important as data quality in predicting validation success. For this experiment, we designed PCR primers to amplify the SNP loci from genomic DNA, and re-sequenced a second population to determine if both alleles could be observed again. We were able to replicate only 314 of these 436 SNPs, even though the EST data were selected for the highest possible sequence quality. We present a population genetics model to explain these observed validation rates, and provide a simple rule-of-thumb for future experiments.

Several major single-nucleotide polymorphism (SNP) discovery projects have been launched in both the public and private sectors (1,2). These projects are motivated by the idea that, if we score a sufficiently large number of polymorphisms, association studies may have the power to detect genetic variations underlying common diseases (3). SNPs are the preferred type of polymorphism for these studies because of their abundance, 1 in 10^3 base pairs (4), and the growing number of high-throughput genotyping techniques (5,6). Of the 26287 SNPs presently in GenBank, 10223 are from the National Cancer Institute's Genetic Annotation Initiative (7). Another 7396 are from the SNP Consortium (2); although, as we were submitting this article, 102719 new SNPs were released through their web site. However, almost all of these SNPs are putative, in the sense that only a small fraction of them have ever been replicated in a second population sample.

In the SNP community, it is commonly assumed that raw sequence quality is the most important determinant of whether or not a site is polymorphic (*i.e.* most high-quality sequence differences are assumed to be true SNPs). We will show that population genetics also requires consideration. Population genetics predicts the existence of a large class of SNPs for which the minor allele is rare (8). As a class, these SNPs comprise a surprisingly large fraction of the segregating sites in a discovery sample. The probability that any one of these sites will be polymorphic in a second sample is very small, and they will typically not be validated, even though they are polymorphic in the first sample, and in the population. The precise fraction of validated SNPs will depend on the detailed population history of the discovery and validation samples. However, the bulk of this effect can be explained by simple models of human demographic history, because most genetic variations occur within populations (9), rather than between them.

To examine the validation problem empirically, we surveyed the expressed-sequence-tag (EST) databases (10) for high-quality discrepancies (11,12), designed PCR primers to amplify these putative SNP loci from genomic DNA, re-sequenced a genetically diverse population (13), and then compared the observed validation rates with our population genetics model. We applied stringent data quality restrictions on the high-quality discrepancies (HQDs) chosen for validation, to minimize the number of artifactual SNPs. Except in a small handful of cases, all selected HQDs had two or more high-quality EST reads in both alleles, and the overall *Phred* quality, as defined in Table 1, was Q30 or better (*i.e.*, the expected error rate is 10^{-3} or better). We inspected all of the relevant EST traces visually, before proceeding with the PCR primer design. After PCR testing, 436 putative SNPs were available. For each SNP, we re-sequenced up to 24 individuals. All three genotypes (AA, Aa, aa) were observed in 213 of these SNPs. Two genotypes (AA, Aa) were observed in 97 of these SNPs, of which, 83 were supported by more than one heterozygote

observation. In 4 SNPs, both homozygotes were observed (AA, aa), but not the heterozygote. For the remaining 122 of these SNPs, we observed a single genotype – a homozygote, but one whose sequence agreed precisely with the initial EST assembly.

To understand why a large fraction of these SNPs were not validated, we examined how the validation rates vary as a function of either the EST data quality or the number of cDNA libraries in the HQD, as defined in Table 1. The observed rates were only weakly dependent on the data quality (Figures 1a, 2a), but they were highly correlated with the number of libraries in the HQD (Figures 1b, 2b). When the number of libraries dropped from two to one, the validation rates fell from 81% to 46%. Artifacts from cDNA cloning are not a plausible explanation because 114 of the 122 un-validated SNPs had at least two high-quality EST reads in both alleles. Differences in the ethnic makeup of the discovery and validation samples are potentially relevant, particularly since the cDNA libraries were not constructed to be genetically diverse. However, this effect is expected to be small because population-specific polymorphisms comprise less than 15% of all variations (9).

To understand the validation experiments, we consider two simple, but very different, models of human population history. The first model assumes a constant effective population size, which appears to fit the data from two recent SNP discovery projects (14,15). The second model adds a period of rapid expansion, following the "out-of-Africa" dispersal of modern humans 100,000 years ago – a model for which there is considerable evidence (8). Our base model with expansion (16) assumes a pre-expansion size of 5000 people, and then expands exponentially to a present day size of 5 billion, over 5000 generations. We also varied these parameters to account for uncertainties in the historical details. Although neither model can fully capture the complex demographic history of modern humans, the two models together provide a useful starting point for exploring the full range of population genetic effects on SNP validation. In particular, with significant expansion, the phylogeny tends to be more "star-shaped", and there is a significantly higher proportion of rare SNPs (17), as we show in Figure 3.

The predicted validation rates, with and without expansion, are shown in Figure 2. When the minor allele is observed in two or more libraries, the predicted rates are the same for both models. The two models differ only when the minor allele is observed in one library, with the expanded population model predicting a much lower validation rate that agrees better with the observed data. With 3 or more libraries, both models predict essentially 100% validation. The observed residual failure rate, with 3 or more libraries, is 6.3%. This may be due to ethnic differences in the discovery and validation samples. Focusing only on those SNPs where the minor allele was

observed in a single library, and taking into account this 6.3% residual failure rate (that is, raising the observed rate by this amount), we find that the predicted validation rate from the expanded population model is not significantly different from the observed data (66% versus 52%, $\chi^2 = 2.8$, $p = 0.09$), while the predicted validation rate from the constant-size population model is strikingly different (85% versus 52%, $\chi^2 = 16.5$, $p = 0.00005$).

To understand how the predictions change for different expansion histories, we performed a series of simulations in which we changed the pre-expansion population size by a factor of two in either direction, and advanced or delayed the start of the expansion by 2500 generations. In general, the validation rates increased with larger pre-expansion populations and decreased with longer times since the start of expansion. When the minor allele is observed in only one library, the expected validation rates ranged from 59% to 72%, resulting in a χ^2 of 0.6 to 6.2, or p-values from 0.01 to 0.45. Note that, in constant population models, the validation rates are independent of population size because the allele frequency distributions are independent of population size (18).

In conclusion, we validated 314 of 436 putative SNPs identified from the EST databases, for a success rate of 72%. Rather than being a product of poor data quality, this low validation rate was primarily due to those SNPs for which one of the alleles was initially observed in a single cDNA library, representing a single individual. This result is expected from population genetics – many of these sites are members of a large class of rare SNPs that appear frequently because they are so common as a class. The observed validation rates were consistent with an expanding population model, but not a constant-sized population model. Although we find good agreement with the data using a simple model of population expansion, this result does not rule out other expansion scenarios. Indeed, very different population models can produce very similar allele frequency distributions (19), and hence very similar validation rates.

The lesson for the SNP community is that many real SNPs discovered in regions of high quality will fail to replicate for population genetic reasons. This is true for almost all discovery strategies, including those used by the SNP Consortium (2). In another SNP validation study (20) that tried to estimate the effects of sequence data quality, the actual validation rate of 56% was significantly smaller than the predicted rate of 78%. However, the ratio of these two numbers is 72%, which is essentially identical to our observed population-genetics-limited rate. We would suggest that, if the objective is to find common SNPs, for use as genetic markers, the validation rates can be improved by screening for high-quality discrepancies where both alleles are observed twice in the discovery sample (for more details, see Eberle and Kruglyak, submitted).

All of our 436 putative SNPs, validated or not, are available in an Excel spreadsheet, at the web site http://www.genome.washington.edu/UWGC/web_publications/default.htm.

Methods

Putative SNPs from EST databases

Putative SNPs were derived from an analysis of 526,711 human ESTs from the WUSTL/Merck EST project (10), using the same basic strategy as other recent similar projects (7,21). SNPs were identified by searching for high-quality-discrepancies (HQDs) in the 23507 *Phrap* contigs (22) with 5 or more ESTs, each of *Phred* quality Q20 or better (11,12). To minimize spurious cloning artifacts, we focused on polymorphisms with multiple EST reads in both alleles and avoided all SNPs within 50-bp of the poly-A tail. 8063 putative SNPs were found this way.

The gene identity of each EST contig was established by a Blast search against the nucleotide and protein databases. Certain genes were avoided: (a) highly polymorphic genes, like HLA and *Ig* genes, (b) highly redundant genes, like ribosomal and cytoskeleton genes, and (c) highly studied disease genes, like *BRCA1* and dystrohpin. We favored oncogenes and tumor suppressor genes, membrane-associated proteins, transcription factors, and genes with functional implications from studies of model organisms. We were not able to assign a gene identity to 60% of the ESTs. A small fraction of these were included, nevertheless, for an unbiased sampling. Of the 335 putative SNPs (out of 436) with an assigned gene identity, 6 were in the 5'-UTR, 36 were synonymous changes, 14 were non-synonymous amino acid changes, and 279 were in the 3'-UTR. There was no correlation between validation rate and lack of gene identity.

DNA amplification and sequencing

We anticipated that primer design for this experiment would be challenging, because we did not have genomic sequences, and had to design primers based on exonic sequences. Therefore, we aimed for targets of size 100 to 130-bp, to reduce the likelihood that we would straddle an intron. The fact that many putative SNPs were on 3'-UTRs was an advantage, as their exons tend to be larger. UTRs are also more divergent than coding regions, so the likelihood of paralog confusion is minimized. Candidate primers were selected with the assistance of *Primer3* (23). Altogether, we designed 643 pairs of primers. These primer pairs were tested against four DNA samples. 522 gave visible PCR products of the expected size. 49 yielded larger PCR products than expected, presumably due to intron interruptions. 72 failed completely. After sequencing the 522 good PCR

products, another 86 primer pairs were rejected because they did not give interpretable sequence traces. Overall, 436 of 643 primer pairs survived, for a success rate of 68%.

DNA samples for the re-sequencing experiments were derived from the Human Diversity Panels distributed by the Coriell Institute for Medical Research (13). The samples were amplified in a total volume of 10 μl, with 7 ng genomic DNA template, 0.5 μM of each primer, 100 μM of each deoxynucleotide triphosphate, and 0.25 U Taq polymerase. Unincorporated primers and dNTPs were removed, before the sequencing reaction, by incubation with 1 μl of exonuclease I (10 U/μl) and 1 μl of shrimp alkaline phosphatase (2 U/μl). To this mix, 1 μl of 3 mM sequencing primer, 3.8 μl of 5X reaction buffer (400 mM Tris-HCl, pH 9.0, 10 mM $MgCl_2$), and 2.2 μl of Big-Dye sequencing mixture were added and thermal-cycled. Since our PCR products were so small, less than 200-bp, we used 6% polyacrylamide gel exclusively, to improve resolution.

Sequencing traces were assembled into contigs with *Phred* (11,12) and *Phrap* (22). The results were viewed in *Consed* (24). All genotypes were interpreted manually, and tagged within *Consed*. As a visual aid, each directory was seeded with an artificial read encoding the expected sequence, based on the original EST contigs, and the position of the putative SNP. Software was developed to export the assembled sequences and tagged SNPs, back to an Access database, for comparison against the original ESTs. Submissions to GenBank were made under the handle UWGC.

Population genetics model details

Genetic trees were calculated using the coalescent model (25). For the expanded population, we started with a constant population and added a period of exponential expansion to arrive at the present population (16). For each population history, we tracked 1000 chromosomes back to their most recent common ancestor. Allele frequency distributions were then calculated by placing mutations at each branch of the tree, weighting the resultant mutation frequencies by the branch length, and normalizing by the total tree length. The distributions from one million trees were averaged. Using these allele frequency distributions, we calculated the probability of validating each of the 436 putative SNPs in a second sample of 25 chromosomes. To minimize costs, the actual validation experiment employed a stopping procedure that halted sequencing as soon as a SNP was validated. To be conservative, we based our model parameters on the average number of samples successfully re-sequenced for the un-validated SNPs (*i.e.*, when the stopping rule was not used). This number was 12.4 individuals, or 24.8 chromosomes.

Because of the inherent ambiguity in determining the actual number of sampled chromosomes in an EST assembly, we used a weighted-average to calculate the probability of validating a SNP.

The ambiguity arises from the fact that we only know the actual number of chromosomes from a particular library if either both alleles are observed or if the library is sampled only once. For example, the assembly in Table 1 has two interpretations: (1) library *c* is heterozygous and allele-2 occurs three times in an effective sample size of 6 chromosomes, *i.e.* (6,3) where we define (*n*,*k*) as a sample of size *n* with *k* minor alleles; or (2) library *c* is homozygous and allele-2 occurs four times in a effective sample size of 6 chromosomes, *i.e.* (6,2) where allele 1 is now the minor allele in the sample. In computing the effective sample size for our statistical analysis, we count both alleles of libraries *a* and *c*, because they were sampled multiple times, but only one allele of libraries *b* and *d*, because they were only sampled once. We weight the validation rate for each scenario by the *a priori* probability of (*n*,*k*), and the probability of the sample observations given (*n*,*k*). For a given allele frequency, the *a priori* probability of (*n*,*k*) is calculated by taking the product of the probability of the frequency (as described above), and the probability that, for the given mutation frequency, the mutant allele is observed *k* times (or equivalently *n*-*k* times) in *n* chromosomes. Summing over all frequencies gives the probability of (*n*,*k*). To calculate the probability of the sample observations, we exclude the "known" alleles and calculate the probability of the "unknown" alleles for all possible (*n*,*k*)'s. For the given example, the only unknown is whether or not library *c* is homozygous. If library *c* is homozygous, *i.e.* (6,2), two samples of the chromosomes that represent library c will always show the same allele; If library c is heterozygous, *i.e.* (6,3), there is only a 25% chance of two samples from library c showing allele-2.

Acknowledgments

We thank Phil Green and Colin Wilson for their assistance in constructing the EST contigs. We thank Maynard Olson and Debbie Nickerson for numerous suggestions. This work was supported by two grants from the National Institutes of Health (1 RO1 ES09909 and 5 RO1 MH59520-02). L.K. is a James S. McDonnell Centennial Fellow.

References

1. Collins, F.S., Guyer, M.S., and Charkravarti, A. Variations on a theme: cataloging human DNA sequence variation. *Science* **278**: 1580-1581 (1997).

2. The official web site for the SNP Consortium, http://snp.cshl.org/about/, contains a detailed description of their experimental protocols.

3. Risch, N., and Merikangas, K. The future of genetic studies of complex human diseases. *Science* **273**: 1516-1517 (1996). Genetic analysis of complex diseases. *Science* **275**: 1327-1330 (1997).

4. Cooper, D.N., Smith, B.A., Cooke, H.J., Niemann, S., and Schmidtke, J. An estimate of unique DNA sequence heterozygosity in the human genome. *Hum. Genet.* **69**: 201-205 (1985).

5. Wang, D.G., Fan, J.B., Siao, C.J., Berno, A., Young, P., Sapolsky, R., Ghandour, G., Perkins, N., Winchester, E., Spencer, J., Kruglyak, L., Stein, L., Hsie, L., Topaloglou, T., Hubbell, E., Robinson, E., Mittmann, M., Morris, MS., Shen, N., Kilburn, D., Rioux, J., Nusbaum, C., Rozen, S., Hudson, T.J., Lander, E.S., *et al*. Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science* **280**: 1077-1082 (1998).

6. Landegren, U., Nilsson, M., and Kwok, P.-Y. Reading bits of genetic information: methods for single-nucleotide polymorphism analysis. *Genome Res.* **8**: 769-776 (1998).

7. Buetow, K.H., Edmonson, M.N., and Cassidy, A.B. Reliable identification of large numbers of candidate SNPs from public EST data. *Nat. Genet.* **21**: 323-325 (1999).

8. Harpending, H.C., Batzer, M.A., Gurven, M., Jorde, L.B., Rogers, A.R., and Sherry, S.T. Genetic traces of ancient demography. *Proc. Natl. Acad. Sci. USA* **95**: 1961-1967 (1998).

9. Barbujani, G., Magagni, A., Minch, E., and Cavalli-Sforza, L.L. An apportionment of human DNA diversity. *Proc. Natl. Acad. Sci. USA* **94**: 4516-4519 (1997).

10. Hillier, L.D., Lennon, G., Becker, M., Bonaldo, M.F., Chiapelli, B., Chissoe, S., Dietrich, N., DuBuque, T., Favello, A., Gish, W., Hawkins, M., Hultman, M., Kucaba, T., Lacy, M., Le, M., Le, N., Mardis, E., Moore, B., Morris, M., Parsons, J., Prange, C., Rifkin, L., Rohlfing, T., Schellenberg, K., Marra, M., et al. Generation and analysis of 280,000 human expressed sequence tags. *Genome Res.* **6**: 807-828 (1996).

11. Ewing, B., Hillier, L., Wendl, M.C., and Green, P. Base-calling of automated sequencer traces using *Phred*. I. Accuracy assessment. *Genome Res.* **8**: 175-185 (1998).

12. Ewing, B., and Green, P. Base-calling of automated sequencer traces using *Phred*. II. Error probabilities. *Genome Res.* **8**: 186-194 (1998).

13. Collins, F.S., Brooks, L.D., and Chakravarti, A. A DNA polymorphism discovery resource for research on human genetic variation. *Genome Res.* **8**: 1229-1231 (1998).

14. Cargill, M., Altshuler, D., Ireland, J., Sklar, P., Ardlie, K., Patil, N., Lane, C.R., Lim, E.P., Kalayanaraman, N., Nemesh, J., Ziaugra, L., Friedland, L., Rolfe, A., Warrington, J., Lipshutz, R., Daley, G.Q., and Lander, E.S. Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nat. Genet.* **2**2: 231-238 (1999).

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

16. Kruglyak, L. Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nat. Genet.* **22**: 139-144 (1999).

17. Slatkin, M., and Hudson, R.R. Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. *Genetics* **129**: 555-562 (1991).

18. Ewens, W.J. Mathematical Population Genetics. *Biomathematics, Volume 9*. Springer-Verlag, Berlin (1979).

19. Reich, D.E., Feldman, M.W., and Goldstein, D.B. Statistical properties of two tests that use multilocus data sets to detect population expansion. *Mol. Biol. Evol.* **16**: 453-466 (1999).

20. Marth, G.T., Korf, I., Yandell, M.D., Yeh, R.T., Gu, Z., Zakeri, H., Stitziel, N.O., Hillier, L., Kwok, P.Y., and Gish, W.R. A general approach to single-nucleotide polymorphism discovery. *Nat. Genet.* **23**: 452-456 (1999).

21. Picoult-Newberg, L., Ideker, T.E., Pohl, M.G., Taylor, S.L., Donaldson, M.A., Nickerson, D.A., and Boyce-Jacino, M. Mining SNPs from EST databases. *Genome Res.* **9**: 167-174 (1999).

22. Green, P. Information on *Phrap* can be found at the author's web site http://bozeman.mbt.washington.edu/phrap.docs/phrap.html.

23. Rozen, S., and Skaletsky, H.J. Information on *Primer3* can be found at the authors' web site http://www-genome.wi.mit.edu/genome_software/other/primer3.html.

24. Gordon, D., Abajian, C., and Green, P. *Consed*: A graphical tool for sequence finishing. *Genome Res.* **8**: 195-202 (1998).

25. Hudson, R.R. Properties of a neutral allele model with intragenic recombination. *Theoret. Pop. Biol.* **23**: 183-201 (1983).

Tables

Table 1: Different measures for the reliability of a high-quality-discrepancy (HQD). We depict here an EST contig with 9 reads. The cDNA libraries from which these EST reads are derived are identified by letters *a,b,c,d*. Different libraries are assumed to have been derived from different individuals, with the exception of the Soares adult brain N2b4HB55Y, adult brain N2b5HB55Y, retina N2b4HR, and retina N2b5HR libraries, all of which were derived from some 55 year old male. Every library has two unique chromosomes, but seeing two reads from the same library in a particular allele, as with libraries *a* and *c*, does not imply that we have seen both chromosomes. Only with library *a* are we certain that we have seen both chromosomes because library *a* is seen in both alleles. Any model must address this inherent ambiguity in the EST data. To assess the quality of the EST sequence itself, we compute the minimum across both alleles of the maximum *Phred* score within each allele.

	Allele 1	**Allele 2**	**HQD**
Library identifiers	*a,a,a,a,b*	*a,c,c,d*	
# of EST sequences	5	4	4
Per read qualities	20-30	20-35	30
# of cDNA libraries	2	3	2
# of chromosomes	2	3-4	2

Supplemental Table for Reviewers: Expected validation rates, for different population histories, of those SNPs that are observed in only one library.

Present Population	**Time of Expansion (Generations)**	**Pre-expansion Population**	**Validation Rates**	**Chi-squared**	**p-Value**
5e9	2500	5,000	0.7227	6.20	0.013
5e9	5000	5,000	0.6577	2.80	0.094
5e9	7500	5,000	0.6153	1.30	0.254
5e9	5000	2,500	0.5850	0.58	0.447
5e9	5000	10,000	0.7228	6.21	0.013

Figures

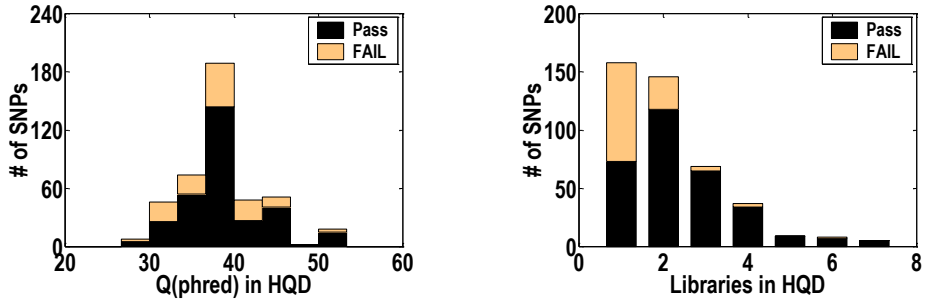

Figure 1: Distribution of sequenced SNPs, plotted against the *Phred* quality (a), and the number of libraries in the HQD (b), as defined in Table 1. Validated (Pass) and un-validated (Fail) SNPs are distinguished by different colors. All of the 436 putative SNPs are depicted (314 validated and 122 un-validated).

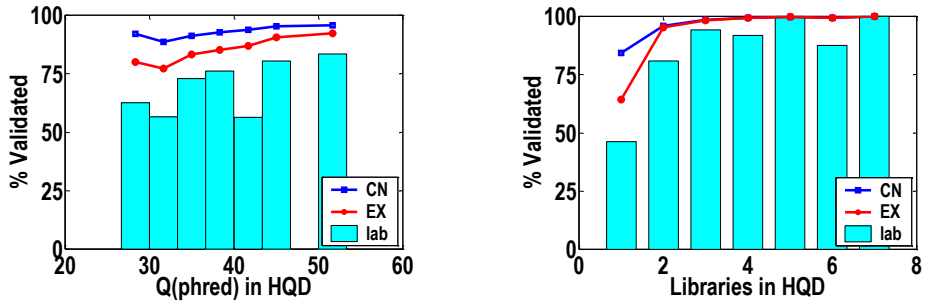

Figure 2: Distribution of SNP validation success rates, plotted against the *Phred* quality (a), and the number of libraries in the HQD (b), as defined in Table 1. The solid lines show the predicted validation rates for the constant (CN) and expanding (EX) population models.

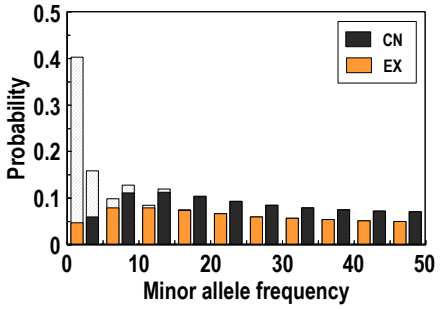

Figure 3: Allele frequency distribution for SNPs discovered in a sample size of 10 chromosomes, under the constant (CN) and expanding (EX) population models. Each bin represents a 5% range of allele frequencies, from 0-5% to 45-50%. The lightly-colored hashing indicates that portion which is predicted to fail validation if 25 chromosomes are re-sequenced. For the putative SNPs that failed validation in our experiments, there was an average of 9.1 unique chromosomes per EST contig, and 24.8 re-sequenced chromosomes.