Published in final edited form as:

Hum Mutat. 2006 March ; 27(3): 249–254.

Distribution of Human SNPs and Its Effect on High-Throughput Genotyping

Daniel C. Koboldt¹, Raymond D. Miller¹, and Pui-Yan Kwok^{2,*}

¹ Department of Genetics, Washington University, St. Louis, Missouri

² Cardiovascular Research Institute, University of California–San Francisco, San Francisco, California

Abstract

Utilizing the results of extensive single nucleotide polymorphism (SNP) studies in humans, stimulated by the International HapMap Project, we present evidence that SNPs are not randomly spaced across the genome, but are somewhat clustered. This observation has important consequences for assay design, since hidden variants in primer sites can affect the accuracy of data. Indeed, using data from the calibration exercises of the HapMap Project, we found instances in which primer site mutations caused allele dropout and other genotyping failures. Given the dynamic nature of SNP discovery, it was inevitable that SNPs would be identified in the primer sites of many assays used for HapMap genotyping. We found that assays with such primer site mutations were correlated with elevated rates of genotype failure and allele dropout. This suggests that taking nearby SNPs into account is important for optimal genotyping assay design.

Keywords

SNPs; primer site variants; HapMap Project; genotyping; assay design; allele dropout

INTRODUCTION

The identification of human SNPs has proceeded rapidly in recent years and has particularly been stimulated by the International HapMap project. Prior to 2002, the majority of SNPs in the public domain were discovered by comparison of overlapping BAC sequences and by the SNP Consortium (TSC), a public/private collaboration [Marth et al., 2001; International SNP Map Working Group, 2001]. Subsequent discovery efforts by the International HapMap Consortium, Perlegen Sciences, and other groups yielded an enormous increase in the number of human SNPs in the public database (dbSNP, www.ncbi.nlm.nih.gov/SNP), which grew from 2.2 million SNPs at the time of the first HapMap strategy meeting in October 2002 to 10.1 million (build 123) at the end of the first phase of the HapMap Project in early 2005.

The characterization of allele frequencies in human SNPs has also proceeded rapidly [Gabriel et al., 2002; Hinds et al., 2005; Miller et al., 2005]. The HapMap Project has greatly stimulated this effort. At the conclusion of phase I in early 2005, genotypes for 1.1 million SNPs (build 16c) throughout the genome had been obtained in samples from each of the four HapMap populations. The HapMap data (freely available at www.hapmap.org) have provided the most extensive view to date of SNPs in humans, and serve as a powerful resource for testing

*Correspondence to: Pui-Yan Kwok, M.D., Ph.D., University of California–San Francisco, 513 Parnassus Ave., Box 0793, HSW-901G, San Francisco, CA 94143-0793. E-mail: Pui.Kwok@ucsf.edu.

Grant sponsor: National Human Genome Research Institute; Grant number: R01 HG-1720.

Ann-Christine Syvänen

hypotheses about the patterns of human variation. One idea is that mutations arose randomly with respect to genomic locations in humans, and consequently the distances between adjacent SNPs should reflect this. In this paper we use the HapMap data to examine this hypothesis. We also investigate the role of unknown primer site mutations in genotyping discrepancies observed in the HapMap calibration exercise, and we compute the effect of such mutations on the seven genotyping platforms that were used in phase I of the HapMap Project.

MATERIALS AND METHODS

HapMap Project

Most of the human DNA sequence variants originally arose from single historical mutation events. Because the mutation rate in humans is very low, a new mutation is initially associated with the alleles of nearby variants on the ancestral chromosome where the mutation occurred. The tendency of certain alleles of nearby variants to be inherited together creates an association between them called linkage disequilibrium [International HapMap Consortium, 2003]. This association makes it possible to use a relatively small set of variants, called a haplotype, to capture much of the variation across a chromosomal region. By genotyping millions of SNPs distributed throughout the genome, members of the International HapMap Project Consortium hoped to construct a haplotype map describing the patterns of common human genetic variation.

Samples for the HapMap Project were collected only after consultations with the communities who provided them, and following approval from the appropriate institutions [International HapMap Consortium, 2004]. Lymphoblasts were cultured from donor blood samples, immortalized, and grown, and DNA was isolated by the Coriell Institute. The resulting DNA panels were distributed to each of nine major genotyping centers that would contribute to the first phase of the project. Seven genotyping platforms were used: Affymetrix (www.affymetrix.com, Santa Clara, CA), BeadArray (www.illumina.com, San Diego, CA), FP-TDI (www.perkinelmer.com, Wellesley, MA), Invader (www.twt.com, Madison, WI), MIP (www.parallelebio.com, San Francisco, CA), Perlegen (www.perlegen.com, Mountain View, CA), and Sequenom (www.sequenom.com, San Diego, CA). Details on the assay design protocols for each platform are available through the HapMap Project website at www.hapmap.org/downloads/assay-design-protocols.html.

All assay and genotype data were downloaded from the public database of the International HapMap Project (www.hapmap.org). The Data Control Center (DCC) also provided SNP "allocation" files to the genotyping centers to assist with primer design. These allocation files provided 1,000 bases of flanking sequence in each direction with repetitive regions, and nearby polymorphisms were marked. Because of the dynamic nature of SNP discovery, however, all groups attempted to perform some genotyping using primers that could be subsequently shown to be in regions now known to contain SNPs. Unfortunately, the frequency and effect of these "SNPs-in-primers" were generally not known.

Early in the course of the HapMap Project, all genotyping centers participated in two large calibration exercises using a variety of technologies to verify the accuracy and quality of the data. While the concordance among centers was very high, one center reported discrepant results with some assays. The most common experimental discrepancy was that some genotypes that were scored as heterozygous according to the consensus data were recorded as homozygous by one center or technology. Rigorous investigations of these discrepancies revealed instances in which primer site variants (SNPs) had affected the genotype calling.

Nearby SNP Bin Calculations

To calculate the frequency of SNPs with nearby polymorphisms, we examined the SNPs in the latest allocation files for chromosomes 1–22 (November 2004). We tested each SNP to see if it had an ambiguity code within any of the six distance bins of 1–5 bp, 6–10 bp, 11–15 bp, 16–20 bp, 21–25 bp, and 26–30 bp. We then compared the number of SNPs that had an ambiguity code in each bin to the total number of autosomal SNPs examined. To test whether the distribution of the SNPs was random, we calculated the expected random distribution of the same number of SNPs over a sequenced part of the human genome. This analysis was repeated separately for the X chromosome.

Investigation of Calibration Exercise Discrepancies

Near the end of phase I, the DCC provided a list of the HapMap data in which SNPs had been characterized by multiple genotyping centers. These data are a subset the “redundant-unfiltered” folders on the HapMap FTP site (Gutmundur Arni Thorisson, personal communication). Analysis of the overlapping genotype sets revealed some instances in which data from one center did not concur with those of another. To resolve such discrepancies, certain samples were resequenced on an ABI 3300 machine. The sequence data were used both to determine the correct SNP genotype and to identify the possible causes of data discordance.

Detecting SNP-in-Primer Assays

To obtain the target SNP and primer sequences of HapMap genotyping assays, we downloaded and parsed the publicly available assay files from HapMap data release 16 (March 2005). We also downloaded the latest SNP allocation files (November 2004), which contain 1,000 bp of annotated flanking sequence in each direction for assayed SNPs. We compared primer sequences with the updated flanking sequence for the target SNP to find assays in which ambiguity codes indicated a putative SNP in a primer site. For molecular inversion probe (MIP) assays, we searched for ambiguity codes within 20 bp of the assayed SNP, since it was difficult to parse the homologous segments out of probe sequences. According to the literature, the homology sequences of MIP probes are centered on the target SNP and average 40.4 bp in length [Hardenbol et al., 2005]. Assays with SNP-in-3'-tail had an ambiguity code within 5 bp of the 3' end of the primer.

A table of HapMap genotyping assays with putative SNPs in primer sites is available for download on our website at <http://snp.wustl.edu/bio-informatics/snp-in-primer/>.

Genotype Failure Rate Determination

We downloaded genotype data for the CEU panel, which was donated by U.S. residents in Utah with northern and western European ancestry, from HapMap release 16. For the passed-data analysis, we used redundant-filtered genotypes and calculated the blank rate as the proportion of blank genotypes (NN) compared to the total number of genotypes attempted. To calculate the blank rate with all data, we used the redundant-unfiltered genotypes, which contained both passed and failed data. We considered any genotype in a set flagged as blank, duplicate, Mendel, Hardy-Weinberg, or other failure to be a blank even if it was called. A chi-squared test was used to show the correlation between SNP-in-primer assays and higher genotype failure rates.

Analysis of Allele Dropout

We cross-referenced this list of overlapping genotype sets with our collection of known SNP-in-primers. Next, we isolated all overlapping genotype sets in which homozygote/heterozygote discrepancies indicated the probable loss of an allele in scoring. A chi-squared test was used to test the association between SNP-in-primer assays and increased allele dropout. We also

identified a subset of genotype sets that were scored as monomorphic by one center, and as having at least one heterozygote by another center. These examples of apparent complete loss of heterozygosity (LOH) were also examined for a correlation with primer site SNPs.

RESULTS

Relative Distribution of SNPs in the Human Genome

Motivated by some results from the HapMap calibrations exercises described below, we tested the hypothesis that mutations arose randomly with respect to genomic locations in humans, and that consequently the distances between adjacent SNPs should reflect this. We felt that the actual distribution of variants was important because it affects the probability that an allocated SNP had a nearby polymorphism that might affect primer hybridization. To estimate the actual spacing of SNPs across human chromosomes, we used the latest HapMap SNP allocation files (November 2004) and calculated, for each SNP, the distance in bp to the nearest neighboring polymorphism (Fig. 1a). According to our findings, some 25.4% of SNPs have a neighboring SNP within 25 bp. Our second analysis took advantage of the annotated flanking sequences provided in HapMap allocation files, in which nearby SNPs were marked with ambiguity codes. We constructed an algorithm to tabulate ambiguity codes within 30 bp in either direction of each allocated SNP. On both autosomes and the X chromosome, we found that the observed distribution of SNPs did not fit the random distribution hypothesis, but instead were clustered. The proportion of adjacent SNPs was sharply increased if they were very near each other compared to pairs that were further apart. For allocated SNPs on autosomes, some 7.5% had an SNP within 1–5 bp, but only 6.0% had an SNP within 6–10 bp, and 3.3% had an SNP within 26–30 bp (Fig. 1b). The percentages are doubled compared to a unidirectional traverse of the chromosome, since two SNPs within 5 bp of each other would be counted twice. For the X chromosome, a similar pattern was observed, except that the percentage of nearby SNPs fell even more rapidly compared to autosomes as the distance between the target SNP and bin increased. Clearly, SNPs are not positioned randomly throughout the genome, but are clustered. An important consequence of the tendency of variants to be clustered together rather than evenly spaced is that the region surrounding a target SNP (most likely to be used for genotyping) is more likely to contain a neighboring variant.

Discrepancies in the HapMap Calibration Exercises

In the rigorous investigation of HapMap calibration exercise discrepancies, resequencing data clearly showed the effect of flanking SNPs on genotyping results across multiple platforms. When the calibration exercise SNP rs227854 was genotyped, the University of California–San Francisco and Washington University (UCSF-WU) submitted a heterozygous genotype for five samples for which the consensus genotype was homozygous. Our group had used a reverse-orientation extension primer that annealed downstream of the SNP. Resequencing was successful in four of the five samples, and revealed not only that the UCSF-WU genotypes were correct, but that an unknown SNP upstream of rs227854 had caused allele dropout in the consensus data (Fig. 2a). Another calibration exercise SNP, rs3025246, was called monomorphic by eight centers, but there were discrepancies in the data submitted by four centers. Resequencing efforts by UCSF-WU for this SNP in the discrepant samples showed that the monomorphic calls were correct. Furthermore, the sequence data revealed that all discrepant samples were homozygous for the minor allele (A) of a nearby SNP, rs16569, which had been in dbSNP since 1999. The effect of this flanking SNP was not allele dropout, but simply genotyping error (i.e., overaggressive calling by a few centers erroneously indicated a heterozygote; Fig. 2b). Thus, although the calibration exercises were designed to demonstrate concordance among the genotyping centers, they also provided experimental verification of rare genotyping anomalies caused by unknown SNPs in primer sites.

Primer Site Variants and Genotyping Failure Rate

A mean sequence divergence of just 1% has a detectable effect on hybridization, and for chip-based hybridizations one disruptive variant in the primer sequence will cause the dropout of an allele [Gilad et al., 2005; Leibel et al., 2003]. If an SNP-in-primer is in high linkage disequilibrium with the assayed target SNP, many individual DNAs that are heterozygous for the target SNP will be scored as homozygous due to allele dropout, and potentially the individual DNAs that are homozygous for the disruptive variant of the SNP-in-primer will be scored the same as the no-DNA control (i.e., as a genotyping failure or blank). Thus, if 1) there is a substantial rate of SNP-in-primer, 2) the SNP-in-primer is in linkage disequilibrium with the target SNP, and 3) the disruptive allele of the SNP-in-primer causes allele dropout, we would expect the blank rate to be increased in assays with SNP-in-primer, regardless of the genotyping platform.

To examine this prediction, we analyzed HapMap data from all genotyping groups to determine the rate of genotyping blanks with regard to primer sequences now known to contain an SNP-in-primer vs. those without. We compared primer sequences from assays submitted by genotyping centers to the annotated flanking sequence provided by the HapMap Data Control Coordinator. The DCC group annotated identified SNPs in the flanking sequence by using single letter ambiguity codes. During phase I of the project, ongoing SNP discovery efforts frequently revealed SNPs in the primer sites of assays that had already been used for genotyping. For each platform we were able to identify a substantial portion of HapMap assays with SNPs in at least one primer site by searching for primer sequences that contained ambiguity codes in later allocation files (Table 1). Some 1.7 million assays were investigated, reflecting not only data displayed on the HapMap web site but also redundant data and data with failure flags. The proportion of assays that had an SNP-in-primer reflects both the length of primers for the platform and the timing in the SNP discovery process when individual groups designed assays. The proportion of assays with SNP-in-primer with the SNP within 5 bp of the 3' end varied with the technologies used, due to the average lengths of the primers.

Given a genotyping assay with a primer-site variant, the SNP could conceivably have a very rare minor allele frequency, not be in linkage disequilibrium with the target SNP, and/or have no effect on hybridization of primers (and thus have no consistent effect on the genotyping failure rate). On the other hand, if the SNP-in-primer had a high minor allele frequency, was in linkage disequilibrium with the target SNP, and did effect the hybridization of primers, then the genotyping failure rate should be increased. We examined the passed genotype data from the March 2005 freeze for CEU. We found a modest but statistically higher blank rate in assays with SNP-in-primer compared to assays with no known SNP-in-primer, for all platforms (Fig. 3a). For genotyping data to be passed by DCC, the call rate had to be $\geq 80\%$. We also analyzed the "unfiltered" data set, which contained all genotype data, including failed genotype sets. In this calculation a failed assay was counted as blank for each of the 90 independent DNAs. Using the unfiltered data set, we found a more noticeable correlation of SNP-in-primer assays with increased genotyping failure rates (Fig. 3b). An examination of both passed and failed genotype sets revealed that the blank rate was consistently higher in the data generated with SNP-in-primer assays. The one exception was the BeadArray platform, where the correlation between SNP-in-primer and increased genotyping failure rate was less obvious, regardless of data set filtration. However, a chi-squared test of the blank distribution showed P-values of <0.001 for all platforms (Table 2). With the possible exception of one platform, we found that an unrecognized SNP-in-primer does indeed have a substantial effect on genotyping success as measured by the blank rate.

The lore of the field suggests that a mismatch near the 3' end of a primer may have a larger effect than a mismatch at other locations. We calculated the effect of the SNP-in-primer if the SNP was in bases 1–5 of the 3' end (Fig. 3b). The results are quite platform-specific and likely

influenced by the particular assay design parameters used. For example, an SNP in the 3' tail (compared to any SNP-in-primer) increased the blank rate for Bead Array, but actually decreased the blank rate for FP-TDI and Invader. The results suggest that in general for assay design it is better to avoid any SNP-in-primer—not just those in the 3' end.

Primer Site Variants and Allele Dropout

The HapMap Project's preparations for phase II and the collaborative approach to genotyping near the completion of phase I inevitably resulted in some submissions of genotypes for the same SNP population combination by more than one group. At the end of phase I there were 173,924 of these redundant genotype sets, permitting further analysis of the results of an SNP-in-primer. We cross-referenced these redundant data with the collection of SNP-in-primer assays we had tabulated. Of the 5,742 SNP-in-primer assays for which redundant data were available, we found that 11.6% were associated with allele dropout in at least one DNA sample in a genotype set (scored as a homozygote by the technology with an SNP-in-primer, and as a heterozygote by the other technology), compared to 7.5% of the 196,029 assays without a known SNP in the primer site. While no allele dropout was observed in 5,076 of the SNP-in-primer assays for which redundant data were available, a chi-squared test showed that the P-value for the increased allele dropout among SNP-in-primer assays was significant ($P < 0.001$). In a second analysis, we identified 905 discordant genotype sets that showed the most severe form of allele dropout: complete LOH. Although just 2.85% of the assays with redundant data had an SNP in the primer site, they accounted for nearly 4% (36 cases) of LOH. A chi-squared test showed this trend to be significant as well ($P < 0.001$).

Masking of SNPs in the Human Genome Assembly

Given the importance of taking nearby SNPs into account when designing assays, we believed that a version of the human genome assembly with SNP positions marked would make an excellent resource for the community. We downloaded the genomic sequence for each human chromosome from the FTP server at UCSC's Golden Path [Karolchik et al., 2003; International Human Genome Sequencing Consortium, 2001]. We marked the position of each non-indel SNP in public databases using IUPAC ambiguity codes. The SNP-masked files are freely available at <http://snp.wustl.edu/bio-informatics/human-snp-annotation.html>.

DISCUSSION

We tested the hypothesis that mutations in humans arose randomly with respect to genomic locations, and that consequently the distances between adjacent SNPs should be comparable to those of SNPs placed at random. Surprisingly, the results did not support the hypothesis. For autosomes the SNPs are in fact clustered, and for the X chromosome SNPs are even more clustered. The reasons for this SNP clustering are unknown, but might include a greater susceptibility of local regions to mutation due to DNA packaging and/or generation of multiple mutations during episodes of error-prone repair of DNA damage.

It has been estimated that when any two human chromosomes are compared, the level of heterozygosity is 7.6×10^{-4} for autosomes and only 4.7×10^{-4} for the X chromosome [International SNP Map Working Group, 2001]. We found that the incidence of SNPs within 1–5 bp of a target SNP was nearly the same for autosomes and the X chromosome, but the incidence of nearby SNPs decreased more rapidly for X than for autosomes. The reduced heterozygosity of the X chromosome compared to autosomes is due to the more rapid drop-off of nearby SNPs.

The ability of an SNP in the primer site to disrupt a genotyping assay depends on the allele frequency of the disruptive SNP. Because most SNPs are biallelic, only one allele of a primer

site (disruptive) SNP results in a mismatch. The sequencing data collected to resolve calibration exercise discrepancies showed that an SNP in the primer site does indeed cause allele dropout. The pattern of complete allele dropout in genotype sets generated with an SNP-in-primer assay suggests that the disruptive SNP is typically in high linkage disequilibrium with the genotyped SNP. This apparent correlation indicates that while only a fraction of assays with a putative SNP in the primer site are associated with LOH, affected genotype sets appear as homozygotes. Our sequencing results also revealed that individuals with two copies of a disruptive SNP-in-primer may be assigned incorrect genotypes when the results are called too aggressively. Such instances of allele dropout or erroneous genotypes not only distort the true variance of an SNP, but can be disastrous if not accounted for in tag SNP selection.

While all genotyping platforms in our study were affected by SNPs-in-primers, the measurable effect was not identical among them. Obviously, quality control measures and assay stringency were factors in these differences. Since the HapMap Project lasted 3 years, timing was a critical element as well. Because many SNPs were discovered after genotyping had begun, it was inevitable that polymorphisms would be discovered in primers from previously-performed assays. As a result, early participants in HapMap genotyping were more prone to design assays with polymorphic primer sites.

The observation that SNPs are clustered rather than randomly spaced increases the challenge of designing assays for high-throughput genotyping. Clearly, SNP-in-primer has not been a problem for most genotyping assays. However, given the importance to the scientific and medical communities of obtaining a high-resolution haplotype map, it is critical that the data used to generate it be as accurate as possible. Some groups have already taken steps to incorporate the effect of SNPs in primer sites into their genotyping quality control. In a recent publication, scientists at Perlegen indicated that the location of known SNPs within 25 bp was used to score their ability to genotype an SNP [Hinds et al., 2005]. Obviously, it is not possible to avoid a potential primer site that contains an unknown SNP, but retroactively identifying assays with high-frequency SNPs in primer sites could be used to screen existing data.

The high level of detail provided by the International HapMap Project's catalog of human variation, coupled with the broad spectrum of technologies involved, has created a powerful resource for scientists worldwide. One advantage of such a large data set is that it provides sufficient examples of the SNP-in-primer phenomenon for it to be analyzed in detail. Our analysis revealed that unknown SNPs in primer sites can and do increase the rate of genotyping failure in all of the HapMap platforms.

Acknowledgements

The data used in this research were generated by the International HapMap Consortium (lists of participants and affiliations are available at www.hapmap.org). We thank Rachel Donaldson and Shenghui Duan for excellent technical assistance; Patricia Taillon-Miller for quality control; Denise Lind and Nicholas Addleman for performing resequencing; Lisa Brooks for her suggestion to look for unexpected features in the data; David Cutler for observations about allele dropout; Nancy Saccone, Ryan Christensen, and Gonçalo Abecasis for guidance on statistics; Gutmundur Thorisson and Albert Smith for continued help with HapMap data at DCC; and two reviewers for comments on the manuscript. This research was supported by a grant from the National Human Genome Research Institute (R01 HG-1720 to P.-Y.K.).

References

- Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M, Liu-Cordero SN, Rotimi C, Adeyemo A, Cooper R, Ward R, Lander ES, Daly MJ, Altshuler D. The structure of haplotype blocks in the human genome. *Science* 2002;296:2225–2229. [PubMed: 12029063]

- Gilad Y, Rifkin SA, Bertone P, Gerstein M, White KP. Multi-species microarrays reveal the effect of sequence divergence on gene expression profiles. *Genome Res* 2005;15:674–680. [PubMed: 15867429]
- Hardenbol P, Yu F, Belmont J, Mackenzie J, Bruckner C, Brundage T, Boudreau A, Chow S, Eberle J, Erbilgin A, Falkowski M, Fitzgerald R, Ghose S, Iartchouk O, Jain M, Karlin-Neumann G, Lu X, Miao X, Moore B, Moorhead M, Namsaraev E, Pasternak S, Prakash E, Tran K, Wang Z, Jones HB, Davis RW, Willis TD, Gibbs RA. Highly multiplexed molecular inversion probe genotyping: over 10,000 targeted SNPs genotyped in a single tube assay. *Genome Res* 2005;15:269–275. [PubMed: 15687290]
- Hinds DA, Stuve LL, Nilsen GB, Halperin E, Eskin E, Ballinger DG, Frazer KA, Cox DR. Whole-genome patterns of common DNA variation in three human populations. *Science* 2005;307:1072–1079. [PubMed: 15718463]
- International HapMap Consortium. The International HapMap Project. *Nature* 2003;426:789–796. [PubMed: 14685227]
- International HapMap Consortium. Integrating ethics and science in the International HapMap Project. *Nat Rev Genet* 2004;5:467–475. [PubMed: 15153999]
- International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* 2001;409:860–921. [PubMed: 11237011]
- International SNP Map Working Group. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* 2001;409:928–933. [PubMed: 11237013]
- Karolchik D, Baertsch R, Diekhans M, Furey TS, Hinrichs A, Lu YT, Roskin KM, Schwartz M, Sugnet CW, Thomas DJ, Weber RJ, Haussler D, Kent WJ. The UCSC Genome Browser Database. *Nucl Acids Res* 2003;31:51–54. [PubMed: 12519945]
- Leibelt C, Budowle B, Collins P, Daoudi Y, Moretti T, Nunn G, Reeder D, Roby R. Identification of a D8S1179 primer binding site mutation and the validation of a primer designed to recover null alleles. *Forensic Sci Int* 2003;133:220–227. [PubMed: 12787655]
- Marth G, Yeh R, Minton M, Donaldson R, Li Q, Duan S, Davenport R, Miller RD, Kwok PY. Single-nucleotide polymorphisms in the public domain: how useful are they? *Nat Genet* 2001;27:371–372. [PubMed: 11279516]
- Miller RD, Phillips MS, Jo I, Donaldson MA, Studebaker JF, Addleman N, Alfisi SV, Ankener WM, Bhatti HA, Callahan CE, Carey BJ, Conley CL, Cyr JM, Derohannessian V, Donaldson RA, Elosua C, Ford SE, Forman AM, Gelfand CA, Grecco NM, Gutendorf SM, Hock CR, Hozza MJ, Hur S, In SM, Jackson DL, Jo SA, Jung SC, Kim S, Kimm K, Kloss EF, Koboldt DC, Kuebler JM, Kuo FS, Lathrop JA, Lee JK, Leis KL, Livingston SA, Lovins EG, Lundy ML, Maggan S, Minton M, Mockler MA, Morris DW, Nachtman EP, Oh B, Park C, Park CW, Pavelka N, Perkins AB, Restine SL, Sachidanandam R, Reinhart AJ, Scott KE, Shah GJ, Tate JM, Varde SA, Walters A, White JR, Yoo YK, Lee JE, Boyce-Jacino MT, Kwok PY. SNP Consortium Allele Frequency Project. High-density single-nucleotide polymorphism maps of the human genome. *Genomics* 2005;86:117–126. [PubMed: 15961272]

FIGURE 1.

SNP spacing in the human genome. **A:** The distance to the nearest neighboring polymorphism was obtained for every SNP in the November 2004 HapMap allocation. **B:** Allocated SNPs whose flanking sequence contained ambiguity codes within 30 bp were categorized based upon the location of the ambiguity code. Data for autosomes and chromosome X are shown with 95% CI. The same calculation for a random distribution of the same number of SNPs is plotted for comparison.

FIGURE 2.

Resequencing of discrepant samples from the first calibration exercise. **A:** Sequence data from five samples with discrepancies for *rs227854* (black arrow) showed that the consensus erroneously called the SNP as homozygous. Allele dropout was caused by an upstream SNP (gray arrow). **B:** Sequence data in both directions are shown for two samples with discrepancies for the SNP *rs3025246* (black arrow). The gray arrow indicates nearby known SNP *rs16569*, for which all discrepant individuals carried two copies of the noncanonical allele (AA).

FIGURE 3.

Effects of SNP-in-primer on genotyping failure rates. The failure rates of genotypes for assays without SNP-in-primer, assays with SNP-in-primer, and assays with an SNP in the 5-bp 3' tail of the primer are shown. **A:** Passed HapMap data from release 16. **B:** Passed and failed HapMap data from release 16. Platform abbreviations: Affy = Affymetrix, Bead = BeadArray, FP = FP-TDI, Inv = Invader, MIP = MIP, Perl = Perlegen, Seq = Sequenom. The data were sequentially assigned to three bins, and the error bars represent the SEM of these groups.

TABLE 1**SNP-in-Primer in HapMap Assays***

Platform	Assays	SNP in primer	
		All (% assays)	3'tail (% all)
Affymetrix	108,799	1.78	46.0
BeadArray	771,530	1.94	23.0
FP-TDI	14,449	9.25	22.4
Invader	641,990	3.13	21.7
MIP	58,950	6.21	26.2
Perlegen	5,902	1.36	12.5
Sequenom	106,892	5.75	23.1
Total	1,708,512		

* Primer sequences from HapMap data release 16 were compared to updated flanking sequence of the SNP assayed (last allocation November 2004), in which putative SNPs are marked with ambiguity codes. Assays for each platform were identified as assay-sequence-OK (no known SNP in primer sites) or SNP-in-primer. SNP-in-primer assays were further categorized as SNP-in-3'-tail if the SNP is within 5 base pairs of the 3' end of the primer.

TABLE 2

The Effect of SNPs in Primers on Genotyping Successes and Failures*

Platform	Primer-Seq-OK			SNP-In-Primer			Chi sum
	Genotypes	Blanks	Blank rate	Genotypes	Blanks	Blank rate	
Affymetrix	9,393,974	735,686	7.26%	167,349	15,716	8.58%	465.37
Beadarray	49,743,261	5,422,479	9.83%	1,005,283	108,307	9.73%	13.20
FP-TDI	1,947,775	234,565	10.75%	183,641	27,259	12.93%	935.25
Invader	17,845,840	3,297,265	15.59%	539,725	159,570	22.82%	26515.15
MIP	4,455,956	477,489	9.68%	246,037	45,613	15.64%	10861.25
Perlegen	540,805	10,385	1.88%	7,408	192	2.53%	16.65
Sequenom	4,354,343	1,116,422	20.41%	255,256	75,439	22.81%	1105.17

* For each platform, the total number of genotypes and total number of blanks were tabulated for assays with or without a known SNP-in-primer. Genotypes in failed gt-sets were all considered blank.