SUPPLEMENTAL METHODS, RESULTS, FIGURES AND TABLES

Multiplex padlock capture and sequencing reveal human hypermutable CpG variations

Jin Billy Li[1,6,8], Yuan Gao[2,6], John Aach[1,6], Kun Zhang[3,6], Gregory V. Kryukov[4,6], Bin Xie[2], Annika Ahlford[1,7], Jung-Ki Yoon[1,7], Abraham M. Rosenbaum[1], Alexander Wait Zaranek[1], Emily LeProust[5], Shamil R. Sunyaev[4], George M. Church[1,8]

1. Department of Genetics, Harvard Medical School, Boston, MA 02115

2. Center for the Study of Biological Complexity, Virginia Commonwealth University, Richmond, VA 23284

3. Department of Bioengineering, University of California, San Diego, CA 92093

4. Division of Genetics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA 02115

5. Genomics Solution Unit, Agilent Technologies Inc., 5301 Stevens Creek Blvd., Santa Clara, CA 95051

6. These authors contributed equally to this work.

7. Present address: Department of Medical Sciences, Uppsala University, Sweden (A.A.); College of Medicine, Seoul National University, Seoul, Korea (J.-K.Y.)

8. Corresponding Author. E-mail: http://arep.med.harvard.edu/gmc/email.html (G.M.C.) and jli@genetics.med.harvard.edu (J.B.L.); fax: (617) 432-6513.

SUPPLEMENTAL METHODS AND RESULTS

Illumina sequence processing

The seven sequencing lanes comprised one lane for single padlock probe libraries for each of PGP1, PGP2, PGP3, PGP9, and PGP10, and lanes for each of two separately prepared padlock probe capture and sequence library replicates for NA10835 (NA10835-1 and NA10835-2). For reasons described in the text and below, the same libraries were sequenced using Illumina Genome Analyzer 1 ("run 1") and Genome Analyzer 2 ("run 2").

The processing pipeline for read sequence processing for each subject is depicted in Figure S2. After initial filtering of reads, the initial 36nt of each read was aligned against a database of reference sequence ligation arms and polymerase extension regions for each target using a simple custom-built dynamic programming software (Aach et al. 2009). The software finds the best-scoring target sequence match for the full 36nt length of each read, where mismatches and gaps earn a -1 point penalty, and matches score 0 points. After examining the impact of different filtering options on SNP calls and predicted heterozygosity (see **Genotyping performance characteristics** below), the options were set to accept only those reads with 0 or 1 mismatches or gaps against the best matching target and for which no other target matched with a score within 2 points of that of the best match. The number of reads that passed these filters is given in Table S1.

Target coverage – read coverage relationship

Of the 53777 targets for which probes were designed, the number of targets that were covered by at least one filtered read in each sample for each of the two sequencing runs is shown plotted against total filtered read coverage in Figure S3. The points in this Figure appeared to fall along a common curve despite the fact that the two sequencing runs were on different versions of the Illumina Genome Analyzer. We estimated the relationship between these two variables by a sum of exponential functions using the Curve Fitting Toolbox in MatLab (The Mathworks: Waltham, MA) (Figure S3). Using this functional fit, we estimate that the percentage of targets covered by at least one read should range between ~90.8% and 94.0% when the total number of filtered reads ranges between 2e6 and 3e6 (see Figure S3).

Genotype determination

In this study, we use terms "locus" and "site" to denote single base or (for CpGs) dinucleotide positions within individual probe Target10 regions and not the entire probe target regions, and use "genotype" to denote the allele pair possessed by a subject for any such site (vs. only for previously identified SNPs within these regions). Genotypes were computed for the Target10 region (see main text) of each read. Target10 occupies read positions 26-35, with positions 26-27 containing a target CpG, while positions 1-25 comprise the ligation arm synthesized for the padlock probe. Genotypes were computed using an algorithm that integrates Illumina quality scores for each base call, base-specific error rate information that is computed from the sequence data themselves, and a specification of Bayesian priors. Examples of quality score distributions from the two sequencing runs are provided in Figure S5. Briefly, the algorithm computes the probability that each genotype obtains given the set of base calls and quality scores for a target position, and predicts the genotype to be the one which has the highest probability. The algorithm also generates a score = $-\log_{10}(1-P)$, where P is the probability computed for the genotype. The value $1-P$ is the probability that the genotype at a site was other than the one called by the algorithm given the base call and quality score data, and so can be considered a probability that the genotype was miscalled according the algorithm's error model. The score value $-\log_{10}(1-P)$ is therefore a measure of confidence in the call. Unless otherwise noted, all analyses of computed genotypes below

only consider genotypes with scores ≥ 5, which indicates a probability of error of 1e–5. The Bayesian prior used by the algorithm assumes that the probability of heterozygosity is .001, so that each of the 6 possible heterozygous genotypes AC, AG, AT, CG, CT, and GT have probability .001/6, and each of the 4 possible homozygous genotypes AA, CC, GG, TT, have probability .999/4. Details on the algorithm will be published subsequently (Aach et al. 2009).

Computational assessment of reproducibility

Reproducibility of results was assessed by comparison of results obtained for the two NA10835 padlock probe capture and sequencing library replicates NA10835-1 and NA10835-2. Comparisons were confined to results obtained for the individual sequencing runs vs. the intersection: Not only is this more conservative than comparison of NA10835-1 and NA10835-2 in the intersections of the two runs, but it allows analysis of the reproducibility of variables that are not available in the intersection. For instance, genotype scores are available from genotyping calculations based on the individual sequencing runs, but not available for the intersection, which constructs the set of genotypes computed with score ≥ 5 from each that match across the runs, but does not generate an aggregate genotype score from the two individual scores.

Results of the computational comparison are as follows: (a) Read coverages (see Figure S6) are highly correlated ($r > 0.98$) between NA10835-1 and NA10835-2 in each of run 1 and run 2, indicating that padlock probes have the same propensity for target capture across multiple reactions. Read coverages are also highly correlated across sequencing runs: For NA10835-1, $r = 0.954$ for run 1 vs. run 2 read coverages, while for NA10835-2, $r = 0.973$ for run 1 vs. run 2 read coverages. (b) Within individual sequencing runs, the percentage of sites for which identical genotypes were computed in both replicate libraries regardless of genotype score (i.e., including sites whose genotype scores were < 5) was 99.78% (486320 of 487376 sites genotyped in both replicates) for run 1, and 99.91% (504591 of 505051 sites genotyped in both replicates) for run 2. (c) Spearman rank and Kendall tau correlation coefficients were computed for genotype scores between NA10835-1 and NA10835-2 replicates within the individual sequencing runs for sites with exactly matching genotypes in the two replicates, again including sites whose genotype scores were < 5. For run 1, the correlation coefficients were 0.923 and 0.794, respectively, across the 486320 sites with exactly matching genotypes. For run 2, the correlation coefficients were 0.937 and 0.836, respectively, across the 504591 sites with exactly matching genotypes. (b) and (c) indicate that the genotyping algorithm yields highly concordant results when applied to sequence obtained from replicate padlock probe capture reactions. To streamline processing, Kendall tau was calculated only for a random subset of 50000 identically genotyped sites. Because the genotype score $-\log_{10}(1-P)$ ranges between 0 and infinity (represented in output data by the highest real number computable by MatLab) genotype scores are best correlated via rank vs. product moment (i.e., Pearson) correlations.

Other factors that affect capturing efficiency

Our results indicate that better probe design and optimized experimental protocol have significantly improved the sensitivity, reproducibility and uniformity of target capture. To identify characteristics that may lead to further improvement, we analyzed the relationship between coverage (i.e., the number of reads per site) and sequence features inherent in the probe design. First, we hypothesized that the first base on the ligation arm is critical to prevent the polymerase from displacing the ligation arm. The ligation event would be disrupted when a single base was displaced even if the Tm of the ligation arm is high. We observed that the average coverage of a probe with G at the end of the ligation arm proximal to the target is ~2 times more than that of a probe with A or C and 7.7 times more than that of a probe with T, whereas no significant difference was observed at the distal end of the ligation arm nor on either

end of the extension arm (Figure S7A). This suggests that in addition to the difference between C+G and A+T clamping at the ligation arm that leads to differential strand displacement, the ligase might be biased in favor of G or A (purines) to be ligated to the CpG at the other side of the ligation junction. Secondly, we found that the G+C content of the target regions, either too high or too low, greatly decreased the coverage (Figure S7B). This bias was likely derived from amplification at the library construction stage or cluster generation with bridge PCR during Illumina Genome Analyzer sequencing (Bentley et al. 2008; Hillier et al. 2008). The effect of G+C content on the extension and ligation arms was less pronounced, consistent with the narrow range of G+C content and Tm designed into the probes (Table S1 and Figures S1, S7C). Thirdly, the uniqueness of the extension and ligation arms, within the preset threshold we chose, did not seem to affect the capture efficiency (Figure S7D).

Known SNP analysis

The UCSC Genome Browser (Kent et al. 2002; Karolchik et al. 2008) was used to download all SNPs of molType "single" and locType "exact" that fell within the Target10 region of each target site. To verify correct positioning of the SNPs, FASTA sequences for these SNPs were downloaded from dbSNP (Build 129) (Sherry et al. 2001) and aligned using Blast-2-Sequences (Tatusova and Madden 1999) to the 41bp of reference sequence comprising the ligation arm and Target10 region for corresponding padlock probe. dbSNP SNPs were accepted for subsequent processing only when the first BLAST hit between the SNP FASTA and padlock probe reference sequences had an E-value <1e–6, for which the corresponding alignment contained both the C from the CpG target of the padlock (position 26) and the SNP, and for which no indels appeared between these two locations in the alignment. Of 5056 dbSNP SNPs for which both UCSC and dbSNP FASTA information was available, 48 failed this location verification check while 5008 were confirmed for location. The positions of these 5008 SNPs along the padlock probe reference sequences are summarized in Table S2. These 5008 SNPs were used to assess genotyping performance (see **Genotyping performance characteristics**).

Of the 5008 location-confirmed known SNPs, 4044 (80.8%) were detected in at least one sample (Table S2). A strong correlation obtains between sample read coverage and the number of known SNPs detected in a sample (Figure S8, left panel).

Genotyping performance characteristics

We used genotypes determined for NA10835 by the HapMap project (The International HapMap Consortium 2003) and for the other subjects by the Personal Genome Project (http://www.personalgenomes.org/) to assess the performance of our padlock-based targeted sequencing method and genotyping algorithm. HapMap genotypes for chromosome 21 were downloaded from http://ftp.hapmap.org/genotypes/latest/forward/non-redundant/ (data set time stamp: 01-Apr-2008 08:56). Of the 5008 position-verified dbSNP SNPs in Target10 regions, 2025 were genotyped in the HapMap data set for NA10835, and of these 2025, 1995 (98.52%) had exactly matching HapMap and Target10 genotypes. Smaller numbers (mean = 232.2) of independently assayed SNPs were available for the five other subjects, and of these, an average of 99.58% had exactly matching PGP and Target 10 genotypes. The high level of agreement between our Target10 genotypes and those obtained from these independent data sources both validates our methods and also confirms that our samples came from the designated subjects and were free from contamination.

In addition to the comparison of computed and HapMap-generated genotypes for NA10835, we considered the consistency of computed genotypes at the dbSNP loci detected in all subjects with the alleles identified in dbSNP. Here "consistency" is the condition that a computed genotype contains only the alleles indicated for the locus in dbSNP. We found > 99.54% consistency for all subjects (Figure S9).

Known SNPs detected in our samples that were annotated as validated in dbSNP exhibited statistically significantly higher consistency than known SNPs whose validation annotation was blank or stated as "unknown" (Figure S9). Consistency varied slightly with read position in a manner that was highly reproducible across all samples (Figure S9).

Heterozygosity analysis

We calculated the heterozygosity of computed genotypes over all Target10 regions in individual read positions and position ranges. Here, heterozygosity was calculated as the fraction of sites identified as heterozygous compared to the total number of sites for which genotypes could be calculated. Heterozygosity as computed here differs from but roughly estimates heterozygosity and nucleotide diversity (π) as standardly computed, as the latter require defined populations and presume Hardy-Weinberg equilibrium. Human nucleotide diversity (π) due to single nucleotide variations is roughly estimated as 0.00075 overall and 0.00052 for chromosome 21 (Sachidanandam et al. 2001). As CpG dinucleotides have a high variation rate, and each of our 53777 chromosome 21 padlock probes targets a CpG at read positions 26-27, we should expect an elevated heterozygosity at those positions compared to all the others. In fact, we found the aggregate heterozygosity of positions 26-27 to be 0.00599 and that of the rest of the Target10 region (positions 28-35) to be 0.00067 (Figure S10, Table S5). The heterozygosity of the target CpG position is ~9x greater than that of the rest of the Target10 region, and this region's heterozygosity is ~1.3x greater than the overall chromosome 21 estimated.

Trends in Target 10 region heterozygosity exhibited in Figure S10 and Table S5 likely reflect both biological and assay-related factors.

1. A rise in heterozygosity is seen from position 26 to position 27 in all six samples. As CpGs are palindromic, variation rates in the two positions should be equivalent; thus this almost certainly represents an artifact. A related observation that considers CpG polymorphic allele fractions across the Target10 region is described below (see **CpG polymorphic allele fractions**), and we discuss hypothetical mechanisms that might account for these observations in **Observations related to possible biases in padlock probe circularization**.
2. The 0.00067 heterozygosity measured for positions 28-35 partly reflects the presence of additional CpGs in the Target10 region beyond the target CpG in positions 26-27 as padlock probes cover many CpG-rich regions of chromosome 21. In fact, among the 44364 probes for which Target10 position 26 was genotyped in at least one of the six subjects, the fraction of CpGs starting in positions 28-35 relative to all dinucleotides in these positions is ~0.016, about 49% higher than the chromosome 21-wide CpG fraction of 0.0108, but comparable with the elevated CpG fractions found in genic and perigenic-regions (Saxonov et al. 2006). We can roughly adjust this heterozygosity value for chromosome 21's actual CpG concentration by using the 0.0060 heterozygosity of position 26-27 to estimate CpG heterozygosity generally, and expressing non-CpG heterozygosity as 0.006/*x*, and then solving 0.97*(0.0060/x)+0.03*0.0060 = 0.00067 for *x* to get the relative decrease in heterozygosity of non-CpG compared to CpG sites. Here 0.03 is the calculated fraction of bases in positions 28-35 in the 44364 probes above that are in CpG dinucleotides, and 0.00067 is our observed rate of heterozygosity in positions 28-35. This yields a value of x of ~11.87. Factoring back the 0.0108 fraction of CpG dinucleotides over chromosome 21 generally, we obtain an adjusted heterozygosity of 0.000623, about 7% less than our directly measured 0.00067 and about 19.8% higher than the HapMap chromosome 21 value for π.
3. PGP10 exhibits ~35% more heterozygosity compared to the other 5 subjects (Table S5), and, consistent with this, a greater number of candidate novel SNPs are found for PGP10 compared to the other subjects (see **Identification of candidate novel SNPs**). This may reflect PGP10's African-

American ancestry, which may confer on PGP10 a higher level of genetic variation compared to the other subjects, who are all of European American ancestry. This 35% increase is quite close to the 32% increase in heterozygosity (0.1484 vs. 0.1126) seen in 15 African-Americans compared to 20 European-Americans computed over 39440 coding SNPS in (Lohmueller et al. 2008) (heterozygosity computed according to our method from Table S1; when only the 20893 synonymous SNPs are considered, the increase is 33% (0.1632 vs. 0.1225)). It is also close to the ~30% increase in autosomal π reported for the genome of a West African individual (Bentley et al. 2008) when compared to the autosomal HapMap π (9.94×10^{-4} vs. 7.65×10^{-4}, where 7.65×10^{-4} was determined by considering only the autosomes in Table 2 of (Sachidanandam et al. 2001)).

Identification of candidate novel SNPs

In addition to identifying known SNPs in the Target10 regions, we also identified all non-dbSNP locations at which a heterozygous genotype was detected in at least one subject, finding in total 489 such locations. Additionally, we found 13 locations that were detected as homozygous in each subject, but for which the homozygous genotypes differed between subjects. These 502 locations may represent candidate novel SNPs (see Table S4). Like known SNPs, the number of non-dbSNP locations for which a heterozygous genotype is detected in a sample correlates with sample read coverage (see Figure S8, right panel); however, this correlation only obtains when sample PGP10 is removed. PGP10 exhibits considerably more non-dbSNP heterozygous sites than any other subject (215 vs. ≤ 87 for the other five subjects). Like the 35% increased heterozygosity exhibited by PGP10 compared to the other five subjects, this relative abundance of candidate novel SNPs in PGP10 may reflect PGP10's African-American ancestry vs. the European ancestry of the other subjects. Taking the correlation with read coverage that obtains for the other subjects into account, PGP10 exhibits 2.94x more non-dbSNP heterozygous sites than the other subjects.

As a computational test of the potential validity of the status of these non-dbSNP heterozygous sites as previously unrecognized sites of genetic variation, we computed, for each subject, the mean number of other subjects that shared heterozygous genotypes (MNOSH) detected in the original subject, and compared this statistic between known SNPs annotated in dbSNP (MNOSH:dbSNP) and non-dbSNP heterozygous sites (MNOSH:non-dbSNP). To illustrate this calculation, supposing we have detected a heterozygous genotype such as GT at a given site in NA10835, we counted the number of the other five subjects who share a GT genotype at this site, and average these counts over all sites and heterozygous genotypes found in NA10835; and also performed the same calculation for the other five subjects. We reasoned that as sites in dbSNP mainly represent sites of common variation, non-dbSNP heterozygous sites should represent rarer or less common sites of variation, so that the MNOSH:non-dbSNP should be lower than the MNOSH:dbSNP. This expectation is upheld in our analysis based on the intersection of the two sequencing runs (see Figure S11). Figure S11 also shows that the large number of non-dbSNP heterozygous sites found in PGP10 are also shared to a much lower degree with other subjects, compared to any of the other subjects.

As noted above, as many as 17 of the 502 loci (~3.4%) that we identified as candidate novel SNPs may represent known SNPs in dbSNP that failed our conservative location verification criteria (see **Known SNP analysis**). Additionally, our criteria for identifying candidate novel SNPs considered only loci that are heterozygous in at least one subject, or homozygous genotypes that differ across subjects. A more common way of identifying SNPs is to look for any genotype differing from the reference genome sequence. The only potential for difference between these criteria would be loci that have identical homozygous non-reference genotypes across all subjects in which they are detected. A search for such loci turned up a single example: position 26 of probe chr21.36773626. This locus is not counted among

our 502 candidate loci and is not included in our analysis. It does not correspond to a known SNP in dbSNP, and was detected in PGP10 only. It corresponds to the C of a target CpG, and PGP10 was found to have TT genotype vs. the CC of the reference genome, and thus represents a bi-allelic CpG→TpG variation.

Need for intersection in determination of non-dbSNP heterozygous sites

While genotypes based on the intersection of the two sequencing runs conform to expectations relative to genotype sharing, we found this not to be true for genotypes based on individual sequencing runs. Indeed, initially, when we analyzed genotypes detected in our original sequencing run (run 1), we found that non-dbSNP heterozygous genotypes were shared among subjects to a *greater* extent than dbSNP heterozygous genotypes (Figure S12). This phenomenon was limited to positions 28-35, but in these positions, MNOSH:non-dbSNP was not only greater than MNOSH:dbSNP, but this difference was highly statistically significant for four of the six subjects. The extent of sharing was such that for 14 non-dbSNP heterozygous sites, the heterozygous genotype was shared by all 6 subjects (vs. 2 dbSNP sites whose heterozygous genotypes were shared by all 6 subjects). For positions 26-27 (the target CpG location), MNOSH:dbSNP > MNOSH:non-dbSNP, as expected, and this difference was highly statistically significant. Thus, the unexpected high degree of sharing of heterozygous genotypes was greater in later positions of the read, another apparently anomalous result.

These observations led us to consider whether aspects of the padlock probe definition, library preparation protocols, unrecognized close repeats of target sequence in the human genome, sequencing, sequence processing, or genotyping algorithm might be introducing biases that would led to incorrect or biased results later in the reads. Because base call quality in run 1 degraded over the course of the read (see Figure S5, left panels), we resequenced the existing libraries. This second sequencing run (run 2) was also performed on a later version of the Illumina Genome Analyzer (see **Illumina sequence processing**) which promised more precision than the version used in the original run (run 1). We found that the genotypes obtained from run 2 still exhibited the same anomaly (see Figure S13).

However, we noticed that the apparently highly shared non-dbSNP heterozygous sites that were yielded by run 1 were generally distinct from those that were yielded by run 2. When we took the intersection of the two sequencing runs, the anomaly was eliminated (Figure S11); by contrast, MNOSH analysis of the heterozygous loci that were filtered out of the intersection indicated very high degrees of sharing (Figure S14). For a locus to be filtered out of the intersection, all genotypes appearing for the locus in one run must be different from the genotypes found in the same subjects in the other. For a candidate SNP to be filtered out of the intersection, either the locus as a whole must be filtered out, or all genotype calls that were heterozygous in a subject (or homozygous and different between subjects) must be discrepant in the two sequencing runs. Examples of highly shared non-dbSNP heterozygous sites that were filtered out in these ways are provided in Figure S15. These findings led us to use the intersection of the two sequencing runs for all analysis of genotypes in this study.

Because the padlock probes, sequencing libraries, sequence processing, and genotyping algorithms were the same for the two sequencing runs, these observations suggest that the source of the anomalously shared heterozygous sequences lies within the sequencer itself or the protocols of its operation. A mechanism which might account for this phenomenon is, however, obscure. We found no significant features of sequences flanking the sites of anomaly. Above we note high correlations > 0.95 in read coverage across sequencing runs, and we also find very high overall concordance >99.9% between genotypes at loci detected in both runs (Table S1). These observations suggest that the unknown

influence that may give rise to apparently incorrect reads affects only a small subset of reads and a small subset of targets.

CpG polymorphic allele fraction

We estimated variation rates in CpGs and non-CpG contexts by computing, for each base, the fraction of times Target10 bases given in subject genotype calls differed from the corresponding bases in the human genome reference sequence used to design the padlock probes, where all counts were allele-wise (see below). Results are given graphically in Figure S16 and numerically in Table S6a.

Allele-wise counts were computed as follows: For reference sequence CpG dinucleotides, all instances in which the corresponding pair of genotypes in consecutive positions subject appeared as CT or TT in the first position, and GG in the second position, were counted as one or two CpG-to-TpG alleles, respectively; similarly, if the first position's genotype was CC and the second position's was AG or AA, one or two CpG-to-CpA alleles were counted, respectively. Total sums of each of these allele counts, aggregated over all subjects, were divided by the twice the total number of CpG sites for which the consecutive genotype calls were any of the above combinations or also the combination CC and GG (which represents no variation), to yield estimates of CpG-to-TpG and CpG-to-CpA polymorphic allele fractions, respectively ("cg>tg", and "cg>ca" below and in Figure S16 and Table S6). These cg>tg and cg>ca allele fractions were calculated for CpGs entirely contained in read positions 26-27 (the padlock probe "target CpGs"), and for CpGs entirely contained within positions 28-35. These CpG polymorphic allele fractions reflect actual CpG mutation rates in humans but only roughly estimate the latter, as accurate estimation depends on using ancestral vs. reference sequence CpGs (see text). Also, these allele fractions do not count small numbers of CpGs that may be found to correspond to other combinations of consecutive genotypes in a subject, such as AC and AG. The reason such other combinations were not counted is that we relied on the results of the genotype algorithm described above, and this algorithm does not, by itself, phase genotypes; thus, given consecutive genotypes AC and AG, it is not possible to tell if this should be counted as an unchanged CpG allele accompanied by a CpG-to-ApA allele (*no* CpG-to-CpA variations), or as a CpG-to-CpA allele accompanied by a CpG-to-ApG allele (*one* CpG-to-CpA variation).

We also computed polymorphic allele fractions for non-CpG contexts. Here the genotype corresponding to any reference genome sequence position that contained an A or T was examined, and any genotype base call that differed from this reference sequence base was counted as an allele for the other base. As above, allele counts were aggregated over all subjects, and divided by twice the number of As or Ts for which genotypes were called, to yield allele fractions for A and T variations to other bases (denoted as "a>c", "a>g", "a>t", and "t>a", "t>c", and "t>g" in Figure S16 and Table S6). Similar counts and fractions were computed for reference genome sequence positions containing a C or a G where the C was changed to A or G, or the G to a C or T, to yield "c>a" and "c>g", and "g>c" and "g>t" allele fractions. However, we restricted counting of C-to-T and of G-to-A variations to only those reference sequence positions for which the C or G was not part of a CpG, and divided by twice the total number of such non-CpG positions for which genotypes were called. These are denoted "non-CpG:c>t" and "non-CpG:g>a" variation rates below and in Figure S16 and Table S6.

Figure S16 shows that cg>tg and cg>ca allele fractions are very much greater (~$43x10^{-4}$ to ~$56x10^{-4}$; Table S6a) than all other allele fractions (~$1x10^{-4}$ to ~$5x10^{-4}$; Table S6a), and that non-CpG:c>t and non-CpG:g>a allele fractions are comparable to all other non-CpG nucleotide fractions. Table S6b quantifies the excess of cg>tg and cg>ca fractions over their non-CpG:c>t and non-CpG:g>a counterparts as between 11.5- and 15-fold, depending on which pairs of allele fractions are considered. These findings

are consistent with numerous prior findings of substantially elevated variation rates for CpG dinucleotides compared to all other substitutions.

Meanwhile, comparison of cg>tg and cg>ca polymorphic allele fractions across positions indicates that cg>tg and cg>ca fractions are virtually identical at ~$43x10^{-4}$ in position 28-35 (Table S6a), differing by < 1% (Table S6b), in conformity with the symmetry of CpG dinucleotides. However, position 26-27 cg>tg and cg>ca fractions differ not only from this common $43x10^{-4}$ value but from each other. The position 26-27 cg>ca fraction is considerably higher than all others, being ~31% higher than the cg>ca fraction in positions 28-35 (Table S6b), but even the less elevated position 26-27 cg>tg allele fraction(~6% more than cg>tg variation in 28-35; Table S6b) is contrary to expectation, as the CpG allele fraction should not depend on position in the Target10 sequence nor should cg>tg and cg>ca fractions differ at any given position. These findings relate to the observations above of highly elevated heterozygosity seen in position 27 (see **Heterozygosity analysis** and Figure S10) and are discussed below (see **Observations related to possible biases in padlock probe circularization** and Figure S17).

Meanwhile, the nearly identical cg>tg and cg>ca allele fractions for positions 28-35, which are free from indications of bias, suggests that $43x10^{-4}$ provides a good benchmark for cg>tg and cg>ca polymorphic allele fractions generally, and that these CpG allele fractions are ~12.5-fold increased over corresponding non-CpG:c>t and non-CpG:g>a fractions (where 12.5 is approximately the average of the 13.3- and 11.5-fold increases of cg>tg to non-CpG:c>t and cg>ca to non-CpG:cg>ca rates in positions 28-35; Table S6b). As noted above, the CpG polymorphic allele fraction should roughly reflect CpG variation rates in humans, and it is, in fact, close to the 13.7 CpG transition rate computed using ancestral alleles described in the text.

Observations related to possible biases in padlock probe circularization

The unequal polymorphism rate between positions 26 (C) and 27 (G) suggests that positions 26 and 27 are subject to an artifact that does not affect positions 28-35. Here we briefly discuss hypotheses related to these observations. As positions 26 and 27 are adjacent to the ligation arm terminus of the padlock probe, and all padlock probes were designed so that a CpG is found in the reference human genome sequence at these positions, it is natural to consider whether padlock probe circularization might be biased by variations away from the reference CpGs found in subject genomes. One hypothesis is that common CpG-to-TpG mutations in position 26-27, in which the C in position 26 is replaced by a T, may bias against padlock probe circularization by destabilizing clamping at the ligation junction, thereby depressing the fraction of circles formed. This hypothesis is suggested immediately by the observation in Figure S10 that heterozygosity in position 26 is lower than that in position 27, as the consequence of a bias against padlock probe circularization due to changes in position 26 would be to lower heterozyogosity relative to position 27.

However, Figure S16 and Table S6, and the discussion in **CpG polymorphic allele fraction** above, suggest that padlock probe formation is not depressed by CpG-to-TpG variations in position 26, but rather that is elevated by CpG-to-CpA variations in position 27 that amplify the number of padlock probes seen carrying this variation relative to variation in position 26. Note that CpG-to-CpA would also tend to destabilize clamping at the ligation arm, but presumably to a lesser extent than variation at position 26. Thus, if destabilization of clamping were to directly cause diminished circle formation, we would still expect to see lower variation at position 26 than at position 27. We must therefore consider more complex mechanisms. By considering the effect of variations on polymerase extension as well as ligation, it is possible to postulate mechanisms that result in circle formation biases that are congruent

to the results given in Figure S16. However, as discussed in the caption to Figure S17, these mechanisms are difficult to reconcile with the low rates of genotype miscall errors for heterozygotic loci (Table S3).

Thus, at this time, no mechanism premised on biased padlock probe circularization presents itself that would account for all features of the heterozygosity, CpG polymorphic allele fraction, and genotyping accuracy currently observed. However, we will explore the possibility of circularization bias in future work by designing padlock probes where target Cs in CpGs are at different distances away from the ligation arm terminus.

Nevertheless, the impact of the bias between position 26 and 27 is small compared to the magnitude of the CpG polymorphism that was our main objective of study. For instance, the variation in heterozygosity between positions 26 and 27 is only ~1.25 compared to a 9-fold change between these positions and 28-35 (Table S5), while the maximum difference between CpG polymorphic allele fractions across all positions is ~1.31 compared to a 12.5-fold difference between CpG and non-CpG allele fractions.

SUPPLEMENTAL FIGURES

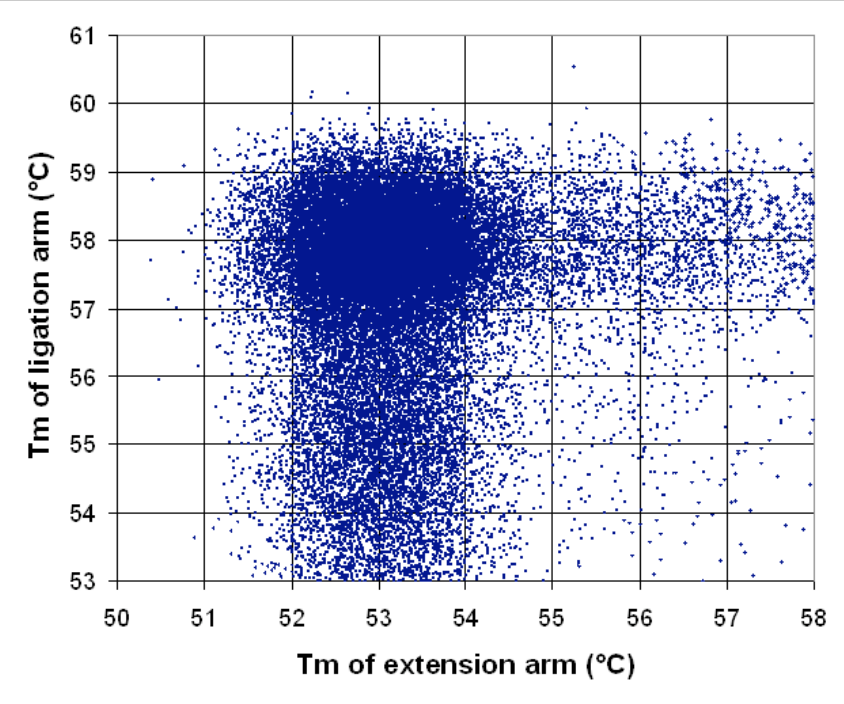


Figure S1. The paired melting temperature (Tm) of the extension and ligation arm in the padlock probe. The Tm of the extension arm ranges from 50-58°C, and the ligation arm from 53-61°C. The optimal Tm of extension and ligation arms is 53°C and 58°C, respectively, which explains the clustering of data points around (53°C, 58°C).

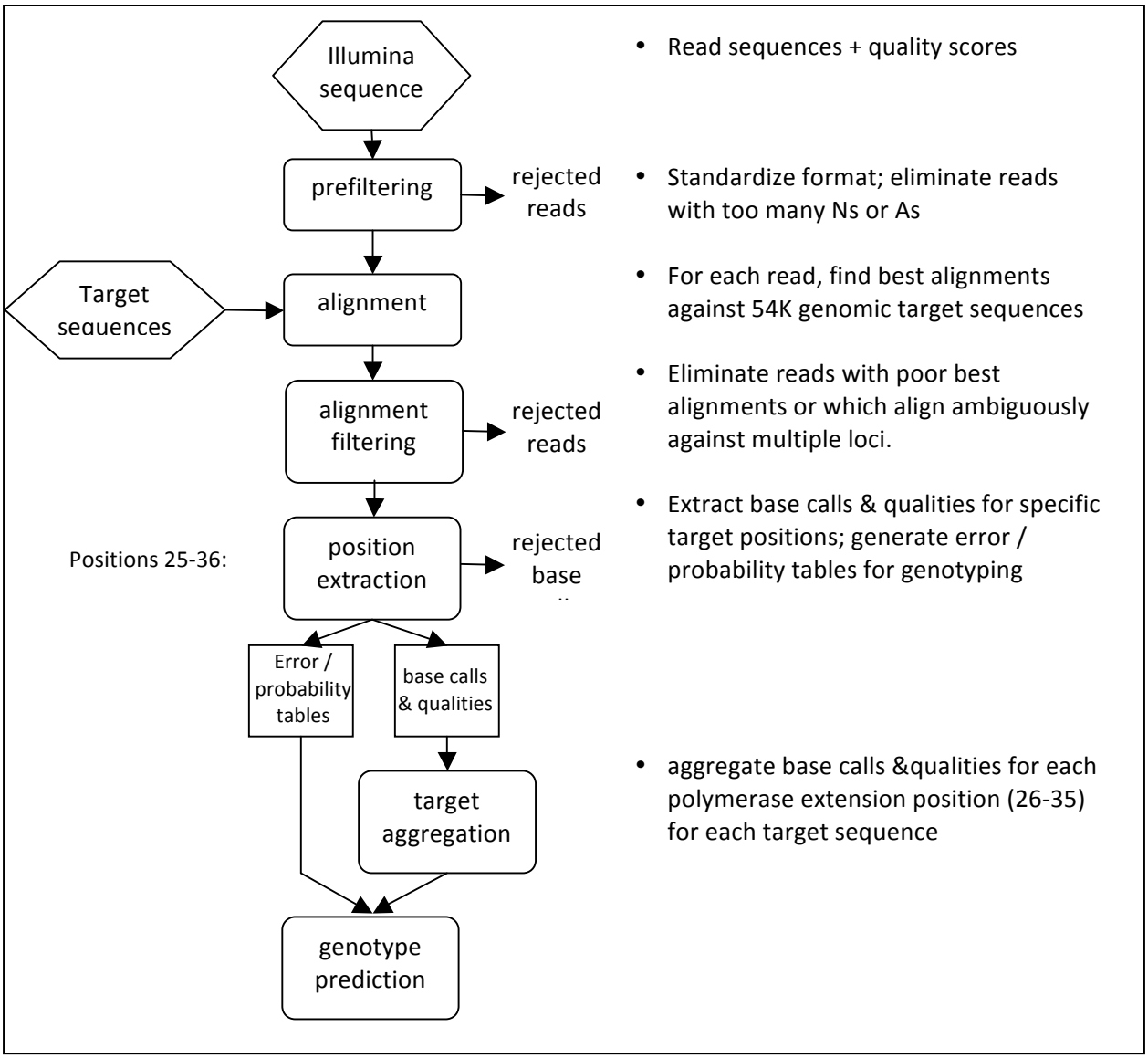


Figure S2. Process flow for analysis of Illumina sequence for individual sequencing runs.

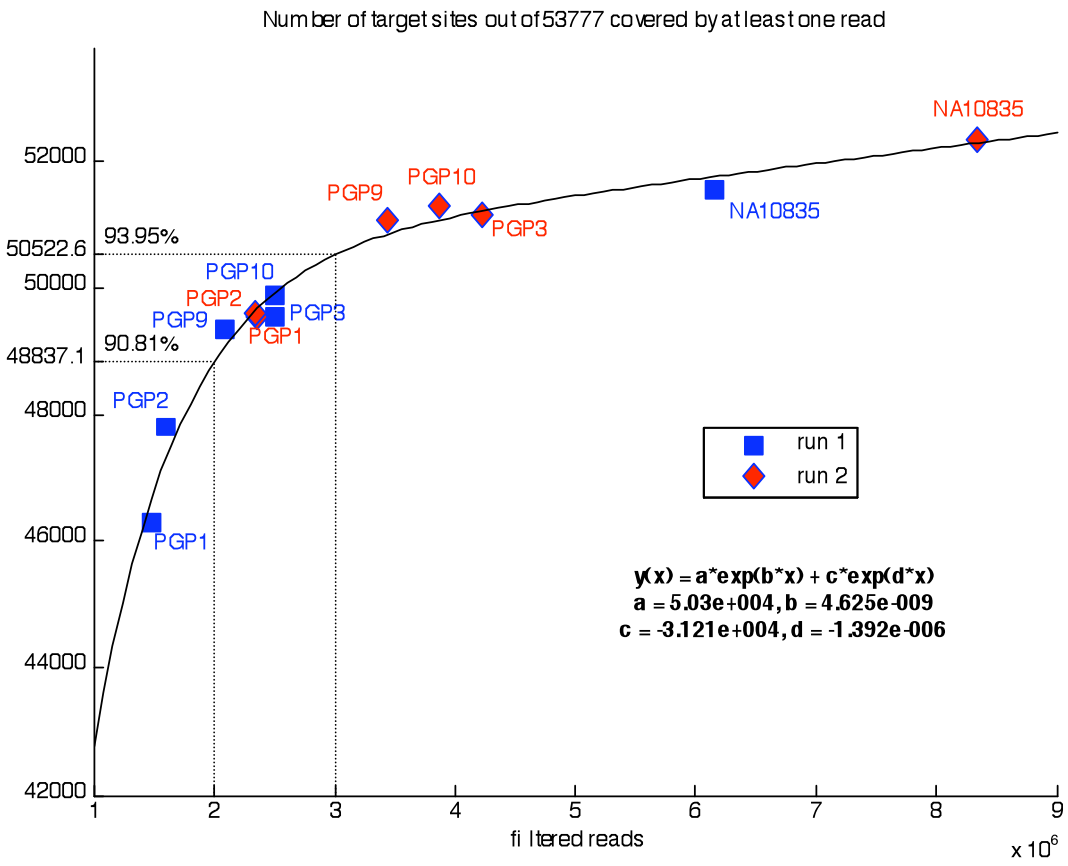


Figure S3. Relationship between total read coverage and number of target sites covered by at least one read. See **Target coverage – read coverage relationship** for details. Note that the NA10835 points represent the aggregates of the sequences obtained for the two NA10835 library replicates (see **Illumina sequence processing**).

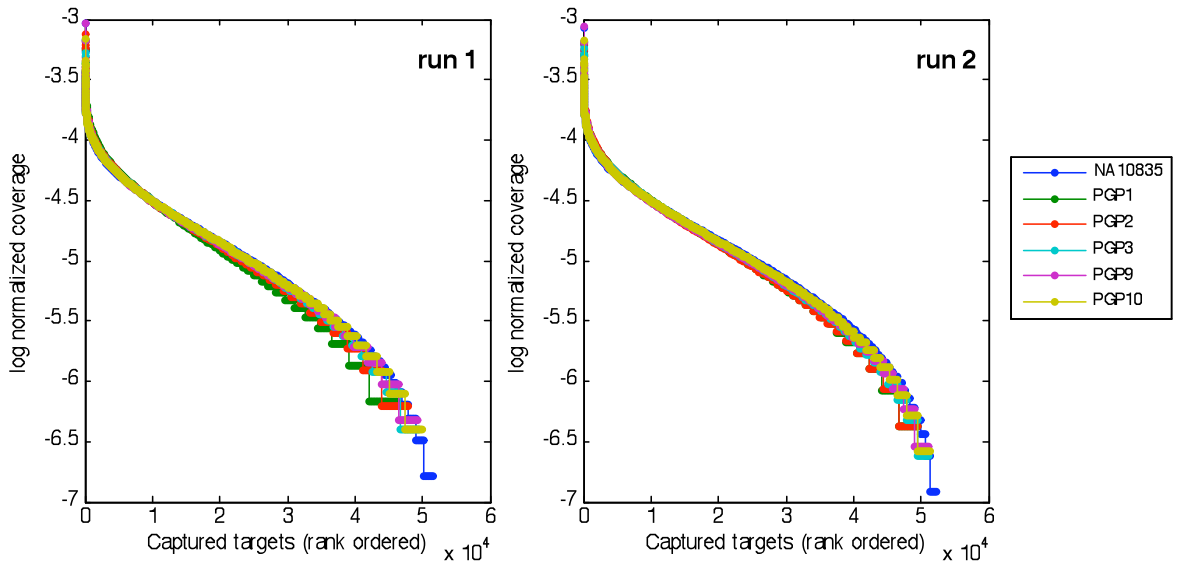


Figure S4. Uniformity of target sites. For each sample, log normalized coverage levels from sequencing of padlock probe reaction products were computed for each captured target as the $\log_{10}$ of the number of target-mapped, filtered reads dividing by the total number of mapped, filtered reads from the reaction. Targets were then ranked for each sample from highest to lowest numbers of mapped, filtered reads and plotted. Except at the extremes, curves exhibit a gradually decreasing slope, indicating that a large number of targets have coverage levels within two orders of magnitude. The plot on the left depicts sequencing run 1 and is the same as Figure 3 of our manuscript. The plot on the right depicts sequencing run 2, which is very similar. For both sequencing runs, over all samples, between 54.4% and 56.5% of all captured targets had coverage levels within a 10-fold range, and between 87.2% and 92.7% had coverages within a 100-fold range. For details on sequencing runs and read mapping and filtering, see **Illumina sequence processing**.

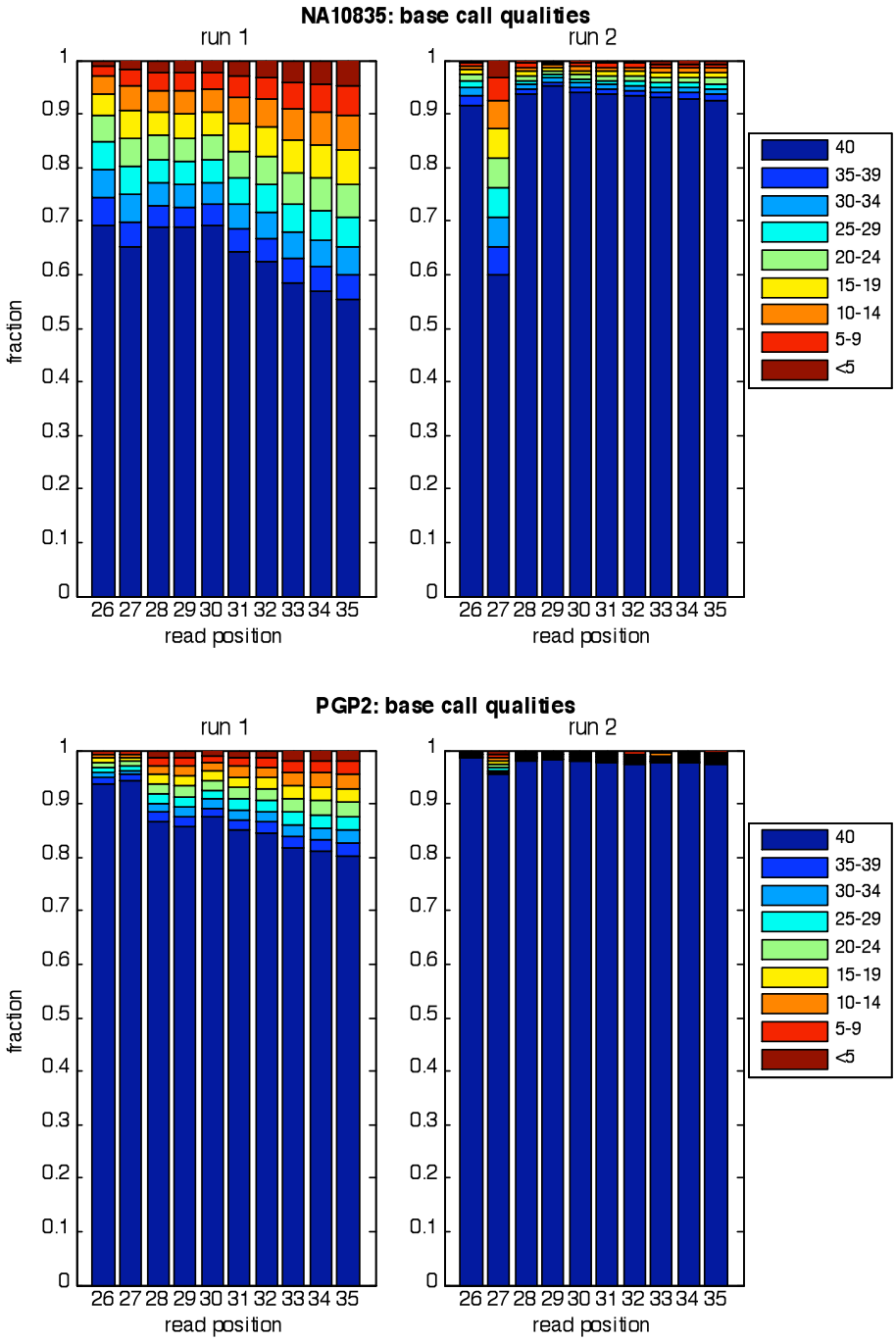


Figure S5. Examples of Illumina base call quality score distributions for the two sequencing runs. The first sequencing run, performed on Illumina Genome Analyzer 1, generally exhibited gradually declining quality scores as read position advanced beyond position 27. In the second sequencing run, performed on Illumina Genome Analyzer 2, high quality was generally maintained throughout the read except for position 27. Although graphs are only shown for NA10835 and PGP2, these trends were seen in all samples.

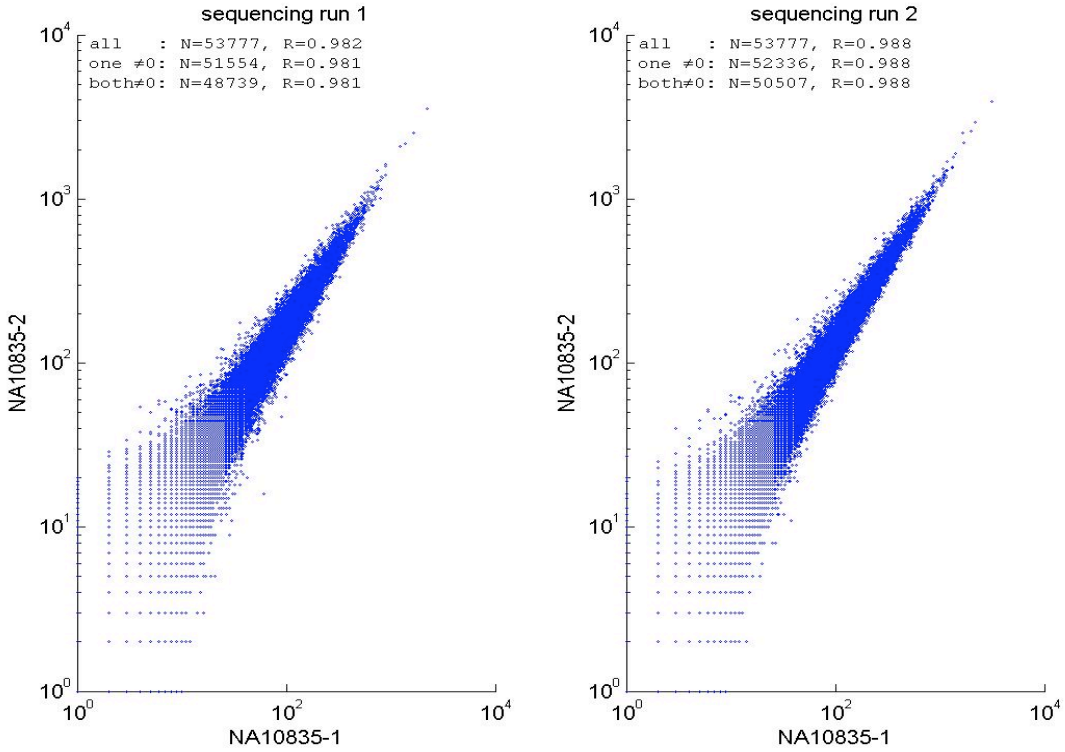


Figure S6. Scatter plot of read coverages in both sequencing runs of the replicate libraries sequenced for NA10835. Read counts were determined from coverages of target position 26 of all reads. Pearson correlation coefficients (R) between read counts are provided for all target sites regardless of whether they have zero coverage (all), for which one of the replicates has non-zero coverage (one ≠0), and for which both replicates have non-zero coverage (both ≠0). Of the 53777 padlock probe targets, 2223 (4.1%) in sequencing run 1 and 1441 (2.7%) in sequencing run 2 had zero coverage in both replicate libraries were therefore not detected in the NA10835 sample at all. All Pearson correlation coefficients are > 98.1% The scatter plot is presented on a log-log scale and therefore only contains points corresponding to targets in the "both ≠0" set.

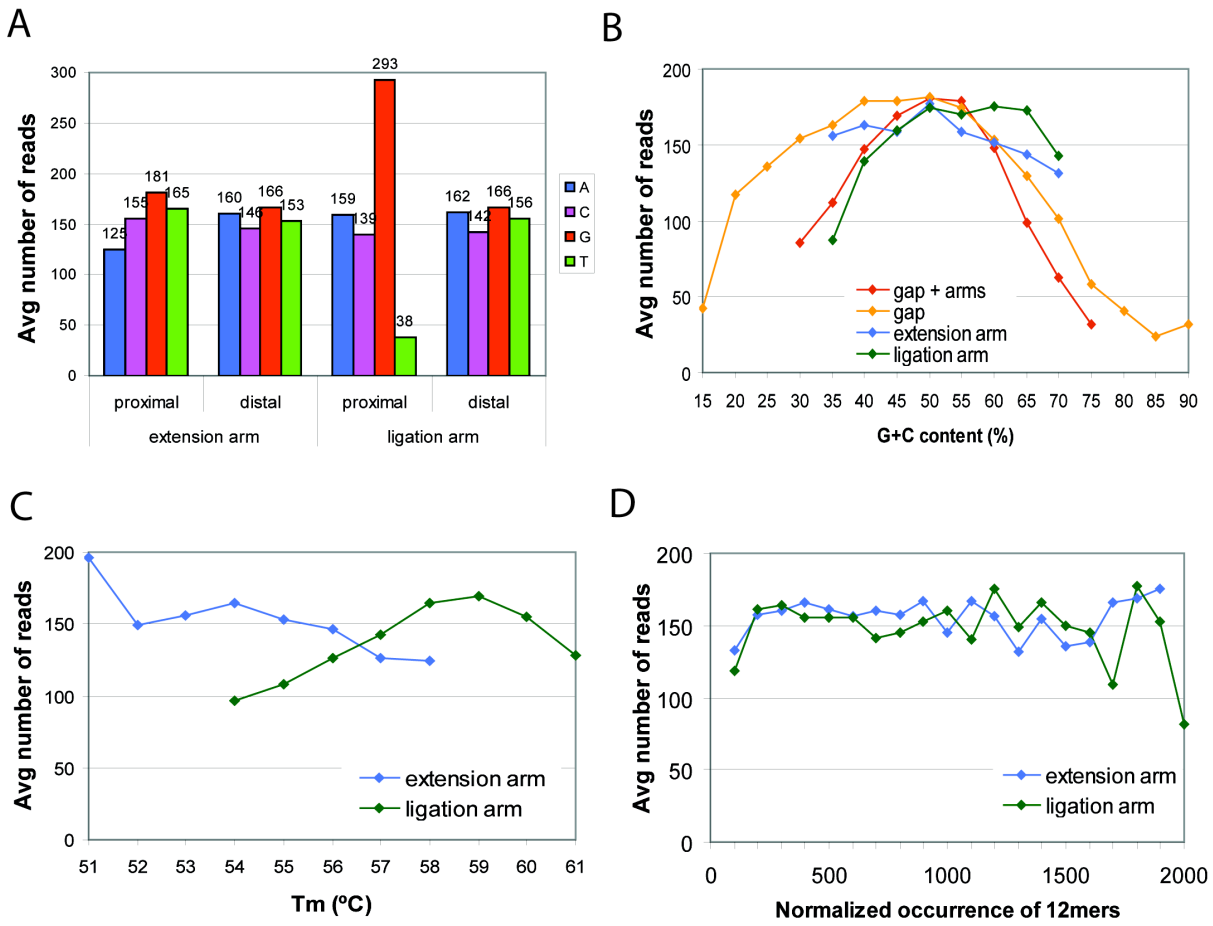


Figure S7. Analysis of uniformity bias of padlock capturing. (A) The average number of reads per site was plotted against the base at the proximal and distal end of both extension and ligation arms, with the proximal ends being immediately adjacent to the target fill-in gap, and distal ends being the other end of the arms distal to the gap. The actual numbers of reads are printed above the bars. The median numbers were also calculated, with similar results (data not shown). (B) The average number of site occurrence was affected by the G+C content of the target region. Although the effect of the G+C content of the targeted gap on the uniformity was significant, only <9% of targets fell in the G+C range of <30% or >70%. (C) The average number of site occurrences was not heavily affected by the Tm of extension and ligation arms, which were tightly controlled in design. (D) The average number of site occurrences was relatively constant regardless of the uniqueness of the arms in the defined range. The uniqueness is measured by the sum of the 12mer occurrences in the human genome normalized by the arm length.

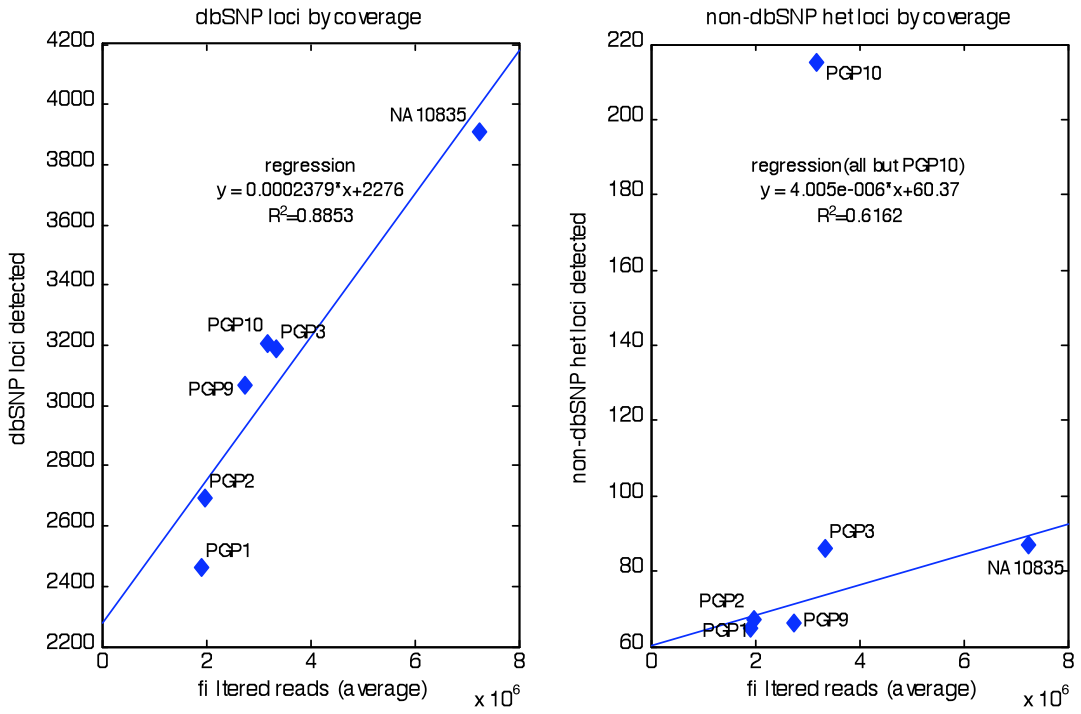


Figure S8. Correlations between read coverage and detected dbSNP loci (left), and detected non-dbSNP heterozygous loci (right). Read coverage is estimated as the average of the number of filtered reads for the two sequencing runs given in Table S1, while dbSNP loci detected and non-dbSNP heterozygous loci detected are directly from Table S1. See **Illumina sequence processing** and **Genotype determination** for discussion of the meaning of "detection" in use in this figure. Note also that "dbSNP loci detected" refers only to the presence of reads sufficient to enable computation of a genotype for a dbSNP locus meeting the criteria for "detection," regardless of whether that genotype is heterozygous or not. By contrast, "non-dbSNP heterozygous loci detected" refers to the number of heterozygous loci actually detected among genotypes, where these loci are not in dbSNP locations (see **Identification of candidate novel SNPs**).

Left: A very strong Pearson correlation (0.94) between the number of known SNPs detected and read coverage indicates that higher coverage results in the ability to genotype more SNP sites.

Right: Similar to detection of dbSNP loci, a strong correlation is found between read coverage and detection of non-dbSNP heterozygous loci, but only when PGP10 is removed from the regression. See **Heterzygosity analysis** and **Identification of candidate novel SNPs** for further discussion. Note that a small number of dbSNP loci annotated as being within Target10 regions failed stringent sequence annotation-based location verification (see **Known SNP analysis**), so that a fraction of the loci identified here as "non-dbSNP heterozygous" could actually be dbSNP loci that failed this verification.

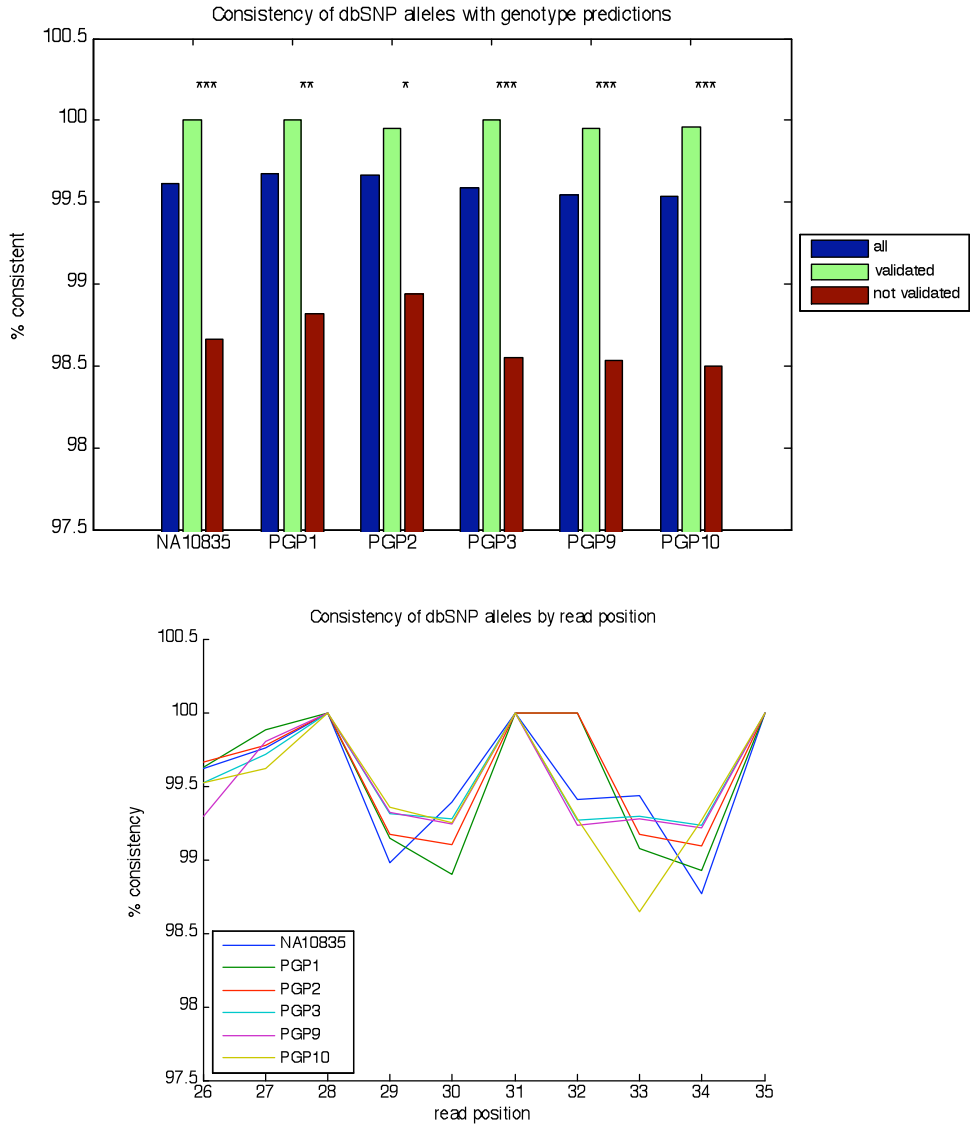
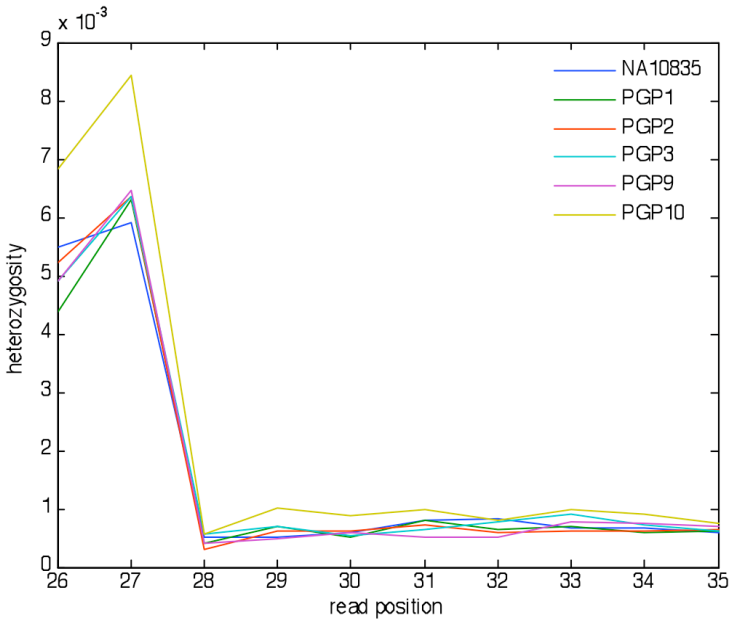


Figure S9. Top: Consistency of genotypes computed for dbSNP loci with dbSNP-indicated alleles for the loci, for all subjects over the entire Target10 region. Percent consistency is reported for all detected dbSNP loci (denoted "all"), and for the subsets of these loci which were annotated in dbSNP as having been validated by some method (denoted "validated") vs. those whose validation was marked as "unknown" (denoted "not validated"). For the meaning of "detection" of a genotype, see **Illumina sequence processing** and **Genotype determination** above. All percentages reported are > 98.49%. The P-values (non-multiple hypothesis-corrected) of two-tailed Z-tests for equality of the fractions of consistent "validated" vs. consistent "not validated" loci are also indicated: * for P<0.05, ** for P<0.005, *** for P<0.0005. The hypothesis of equality is rejected in all cases even after correction for multiple hypotheses: Thus, even though the fractions of dbSNP-consistent genotypes are very high for all categories, the fraction of genotypes consistent with dbSNP "validated" SNP allele annotations is statistically significantly greater than that for genotypes consistent with dbSNP "not validated" SNP annotations. **Bottom:** Consistency of genotypes computed for all subjects with dbSNP allele annotations, for all dbSNP loci ("validated" or "not validated") meeting the conditions above. Consistency is high (> 98.6%) over all read positions, but some read positions exhibit lower consistency than others for all subjects.



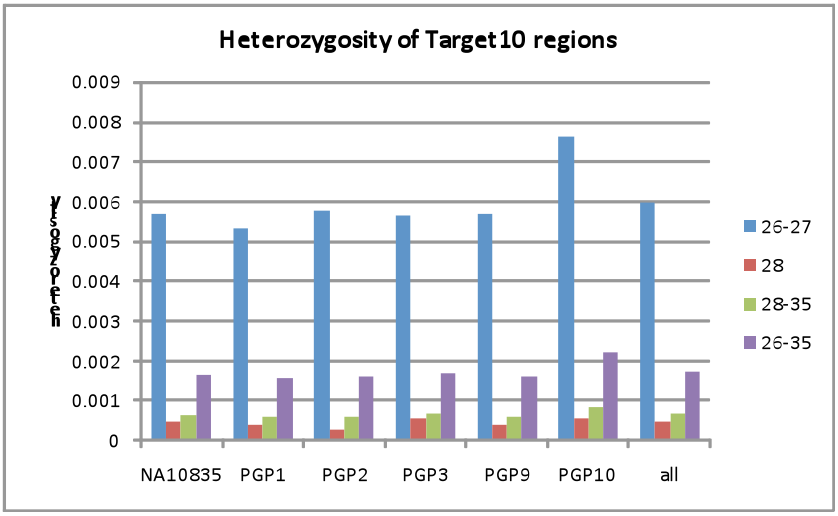


Figure S10. Top: Heterozygosity of the Target10 region, computed as the number of heterozygous genotypes over the total number of genotypes, determined for each subject and shown for each read position in the region. (See Table S5 for numeric values.) The high values in positions 26-27 compared to the rest of the region reflect the increased mutability of the CpG dinucleotides at those positions; however, the inequality between the position 26 and 27 heterozygosity likely represents an artifact. Subject PGP10 exhibits 35% higher heterozygosity than the other subjects, which may reflect the different ancestry of PGP10 compared to the others. **Bottom:** Heterozygosity of computed genotypes over Target10 regions aggregated over the target CpG positions (28-26) and the rest of the Target10 region (positions 28-35). Heterozygosity is computed as the number of heterozygous computed genotypes over the total number of computed genotypes over all detected genotypes (see **Heterozygosity analysis**). For the meaning of "detection" of a genotype, see **Illumina sequence processing** and **Genotype determination** above. Numeric data corresponding to charts above are provided in Table S5.

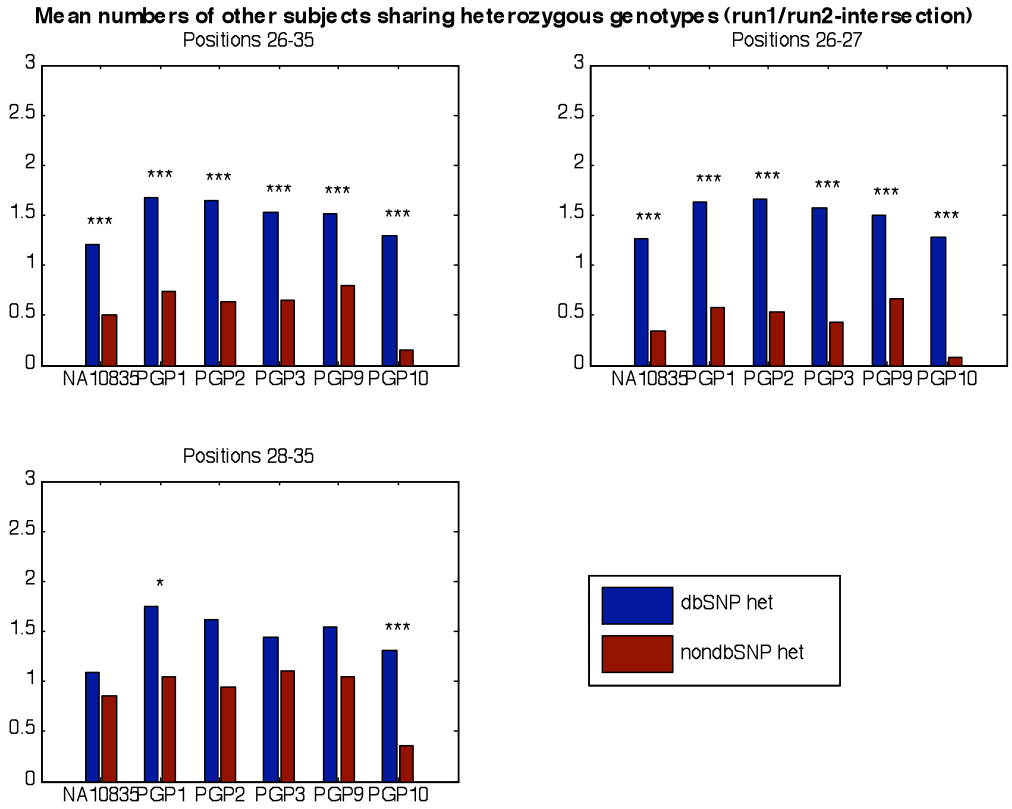


Figure S11. Mean numbers of other subjects sharing heterozygous genotypes (MNOSH) for dbSNP (MNOSH:dbSNP) and non-dbSNP heterozygotic loci (MNOSH:non-dbSNP) for genotypes detected in the intersection of sequencing runs 1 and 2. For the meanings of "detection" of a genotype, and or "intersection," see **Illumina sequence processing** and **Genotype determination** above. For both the target CpG locations in read positions 26-27 (upper right), and the remainder of the target 10 region (positions 28-35; lower left), dbSNP heterozygotes are shared among subjects more than non-dbSNP heterozygotes, as would be expected from the presumed status of dbSNP locations as sites of common variation and non-dbSNP heterozygotic locations as rare or less common variations. This expectation was not met for genotypes computed from sequencing run 1 or sequencing run 2 individually (see Figures S12 and S13). See Figure S15 for examples of highly shared heterozygous genotypes that were filtered out by taking the intersection of the two runs. In the figure above, ***, **, and * signify that MNOSH:dbSNP is statistically significantly different (2-tailed Z test) from MNOSH:non-dbSNP at $P<0.0005$, <0.005, and <0.05, respectively.

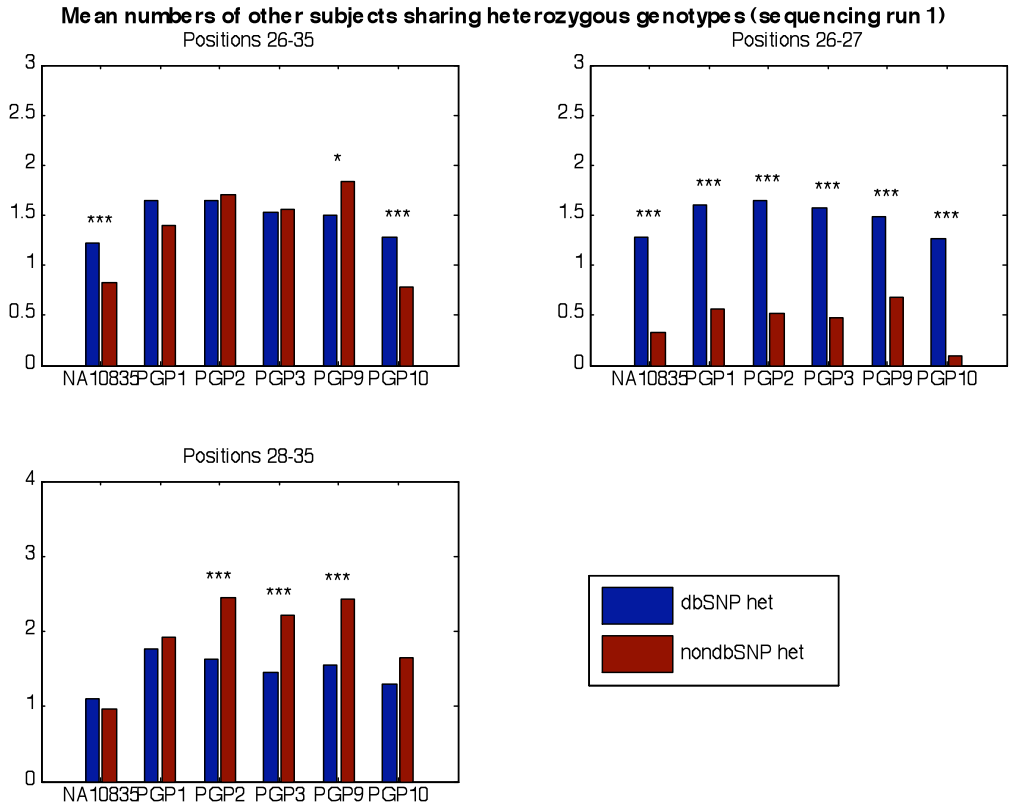


Figure S12. Mean numbers of other subjects sharing heterozygous genotypes (MNOSH) for dbSNP (MNOSH:dbSNP) and non-dbSNP heterozygotic loci (MNOSH:non-dbSNP) for genotypes detected in sequencing run 1 only. For the meaning of "detection" of a genotype, see **Illumina sequence processing** and **Genotype determination** above. While in the target CpG locations in read positions 26-27 (upper right), MNOSH:dbSNP > MNOSH:non-dbSNP with high statistical significance, this relationship is not seen for the remainder of the Target10 region (positions 28-35; lower left). This observation is anomalous given the expectation that dbSNP locations represent sites of common variation whose heterozygotic genotypes should therefore be shared more than non-dbSNP heterozygotic genotypes, which presumably represent rare or less common variations. See **Need for intersection in determination of non-dbSNP heterozygous sites** for details. ***, **, *: see caption to Figure S11.

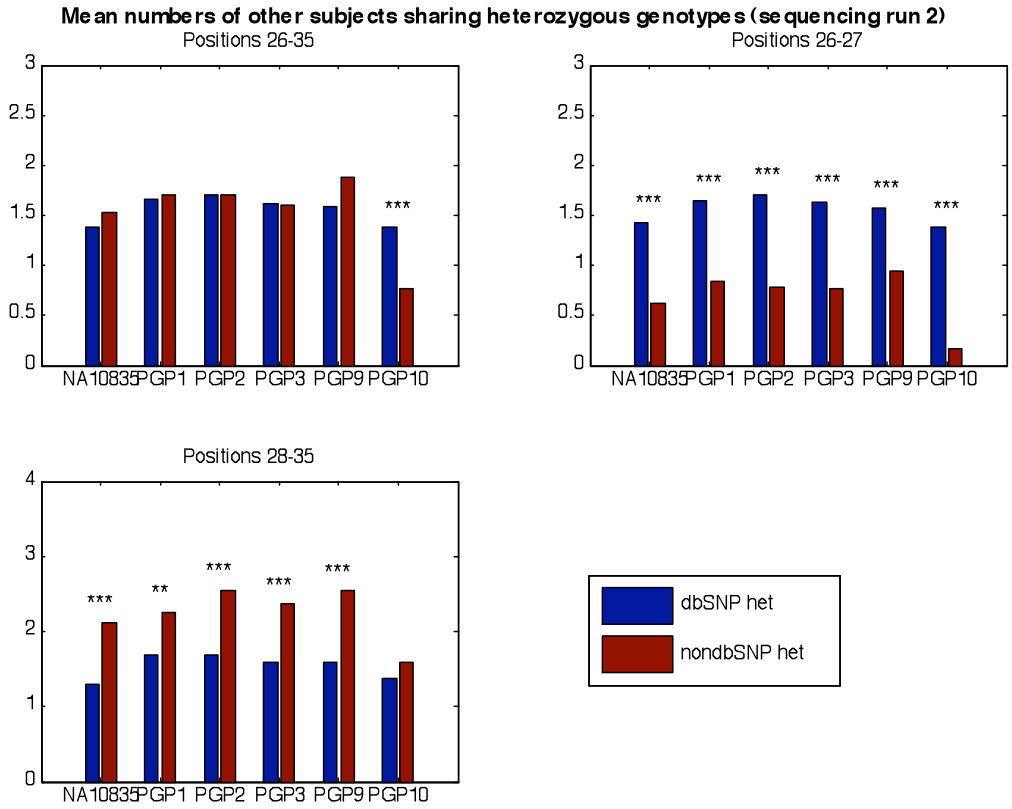


Figure S13. Mean numbers of other subjects sharing heterozygous genotypes (MNOSH) for dbSNP (MNOSH:dbSNP) and non-dbSNP heterozygotic loci (MNOSH:non-dbSNP) for genotypes detected in sequencing run 2 only. See Figure S12, which presents these data for sequencing run 1, for explanation of the figure and **Need for intersection in determination of non-dbSNP heterozygous sites** for further discussion.

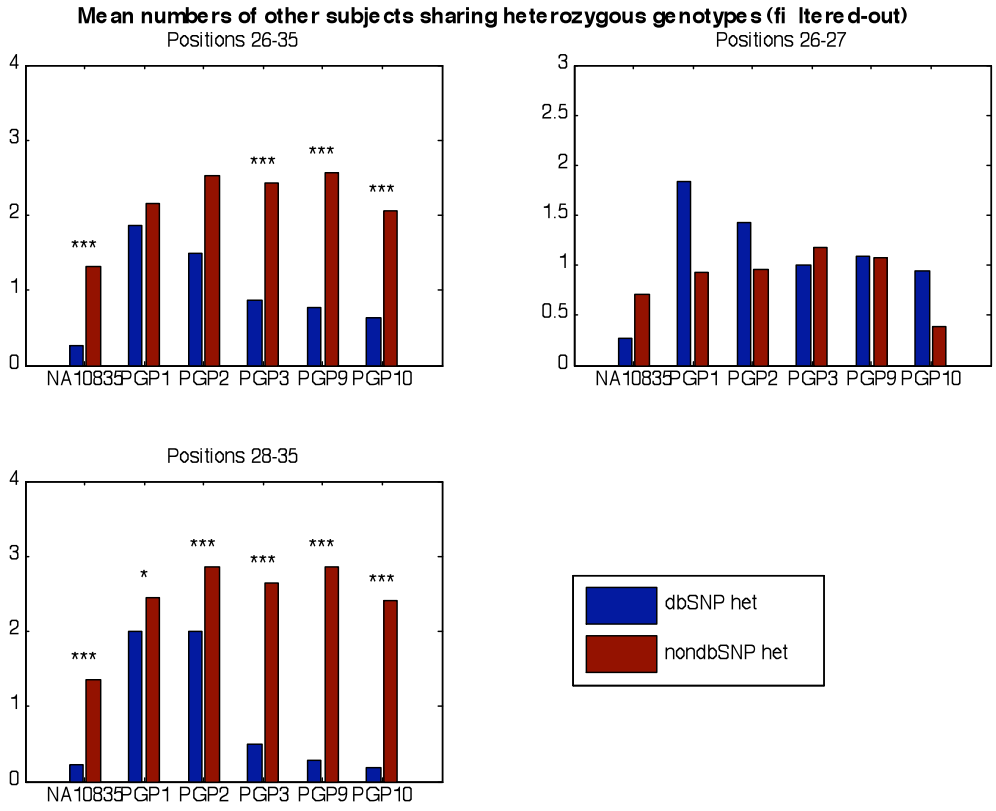


Figure S14. Mean numbers of other subjects sharing heterozygous genotypes (MNOSH) for dbSNP (MNOSH:dbSNP) and non-dbSNP heterozygotic loci (MNOSH:non-dbSNP) for genotypes detected in one or the other sequencing run but not in the intersection between the two sequencing runs. See Figure S12, which presents these data for sequencing run 1, for explanation of the figure and **Need for intersection in determination of non-dbSNP heterozygous sites** for further discussion.

Sequencing run	Highly shared genotype	Fraction of subjects with genotype	NA10835	PGP1	PGP2	PGP3	PGP9	PGP10
			Genotype					
			Reads					
			A:C:G:T					
Target: chr21.31055963, read position 33								
1	GT	6/6	GT	GT	GT	GT	GT	GT
			164	26	89	80	100	86
			0:0:68:96	0:0:17:9	0:0:42:47	0:0:28:52	0:0:45:55	0:0:32:54
2	GG	4/4	GG	GG	gt	GG	gg	GG
			281	57	71	126	122	119
			0:0:239:42	0:0:52:5	0:0:56:15	0:0:102:24	0:0:95:27	0:0:98:21
intersect	XX	0/0	XX	XX	XX	XX	XX	XX
Target: chr21.33541093, read position 27								
1	GG	6/6	GG	GG	GG	GG	GG	GG
			328	78	99	134	119	140
			0:0:320:8	0:0:77:1	0:0:99:0	0:0:134:0	0:0:118:1	0:0:140:0
2	GT	5/6	GT	GT	GT	GT	GT	GG
			543	139	143	232	191	228
			1:0:333:209	0:0:101:38	0:0:110:33	0:0:161:71	0:0:138:53	0:0:174:54
intersect	GG	1/1	XX	XX	XX	XX	XX	GG

Figure S15. Examples of highly shared heterozygous loci detected in sequencing runs 1 and 2 that were filtered out in all subjects from the intersection of the two sequencing runs. Loci are specified as read position within targets, where the target location is the chromosome 21 location of a target CpG. For the meanings of "detection" of a genotype, and of the "intersection" of the sequencing runs, see **Illumina sequence processing** and **Genotype determination** above.

In the first example, chr21.31055963, read position 33, the heterozygous genotype GT was detected for all subjects in run 1, whereas in run 2 GG genotypes were detected in 4 subjects while two subjects had genotypes that did not meet the score threshold of 5 (indicated by lower case letters). As a result, this locus was entirely eliminated from the intersection (indicated by XX genotypes for all subjects). Additionally, this locus was Sanger sequenced in all subjects and all genotypes were found to be GG, confirming that the run 1-based genotypes were all in error.

In the second example, chr21.33541093, read position 27, the heterozygous genotype GT was detected for five out of six subjects in run 2, while run 1 genotypes were all GG. Because the genotype for PGP10 in run 2 matched the GG in run 1, this locus was *not* entirely eliminated from the intersection (indicated by a GG for PGP10 and XXs for all other subjects). However, as the locus no longer exhibited any heterozygous genotypes in any subject, it was no longer considered a candidate novel SNP in the intersection, as it had been in run 2.

Note that neither of these loci appears in dbSNP.

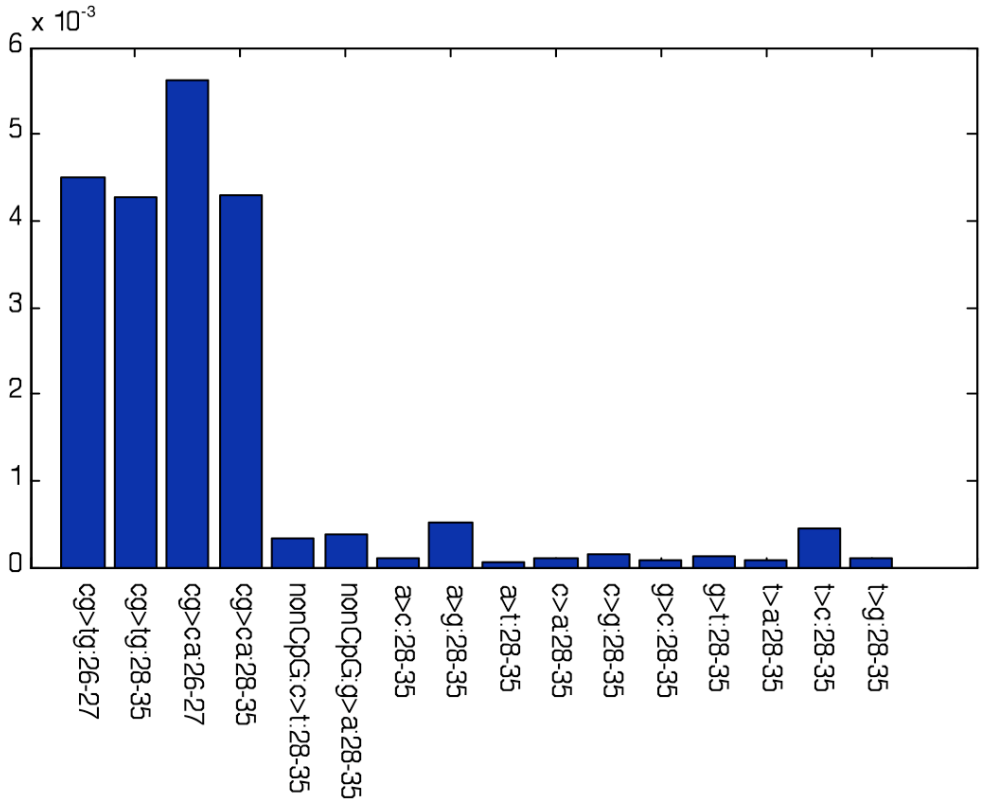


Figure S16. CpG polymorphic allele fractions compared with non-CpG polymorphic allele fractions. CpG polymorphic allele fractions were computed by counting TpG and CpA alleles in genotypes for loci that are CpGs in the human reference sequence; non-CpG polymorphic allele fractions are computed similarly. These allele fractions reflect CpG variation rates but accurate estimation of the latter requires using ancestral vs. reference sequence CpGs. CpG-to-TpG (cg>tg) and CpG-to-CpA (cg>ca) allele fractions are estimated for the target CpGs in positions 26-27, and also for positions 28-35. All other allele fractions are given for positions 28-35 only. CpG-to-TpG (cg>tg) and CpG-to-CpA allele fractions are equal in positions 28-35 where they are ~12.5-fold higher than corresponding non-CpG fractions. CpG allele fractions are higher in positions 26-27 than in positions 28-35, likely reflecting the same artifact that affected position 26-27 heterozygosity in Figure S10. See Table S6 for numerical values presented in this figure.

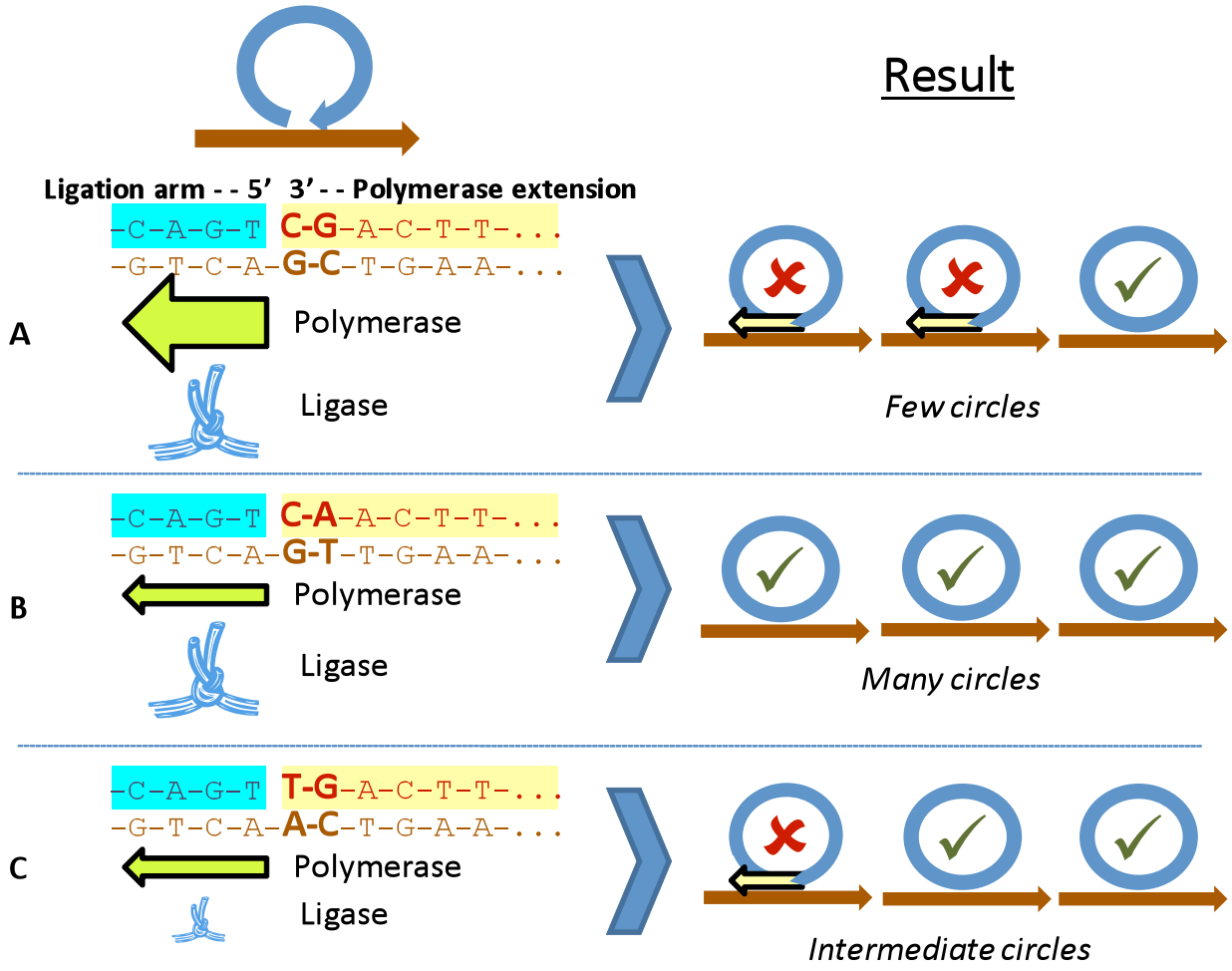


Figure S17. Hypothetical mechanism explaining elevated variation rates in positions 26-27 as artifacts of biased padlock probe circularization. The left side of the panel shows a padlock probe that has been extended by the polymerase up to the end of the ligation arm, a point at which polymerase and ligase activities compete. The right side depicts the result of the competition: If the polymerase wins the competition, it will continue to extend and displace the ligation arm, which will result in failure of the padlock probe to circularize and capture the target. If the ligase wins, a circle will be generated. For this hypothetical mechanism, it is assumed that tight clamping of the polymerase extension by GC base pairing in the two positions adjacent to the ligation arm will very strongly encourage further extension, and that weakening of either of these base pairs will substantially reduce this activity. It is also assumed that ligase activity is strong where there is a GC base pair adjacent to the ligation arm terminus, and that this is weakened by the weaker clamping induced by an AT base pair. There are three scenarios:

A. Given a CpG at the target position, further polymerase extension is strongly induced (large yellow arrow), even though the ligase is also strong at this position. The result is that there will be few circles formed. Thus, when there is no variation at the target CpG, circle generation is suboptimal.

B. If the CpG has mutated to a CpA, the polymerase's tendency to extend is greatly reduced (narrow yellow arrow) while ligase activity remains strong. This results in many circles.

C. If the CpG has mutated to a TpG, ligase activity is reduced (small 'knot' image), but as polymerase activity is also greatly reduced (narrow yellow arrow), there is still more tendency to form circles than in situation A. The result is an intermediate number of circles.

Thus, if both alleles are CpGs in the target position, there will be suboptimal generation of circles from each allele, but there will be no bias because both alleles are identical. However, if one allele is CpG while the other is CpA, there will be more circles generated from the CpA allele than from the CpG allele, resulting in greater sensitivity of detection of the CpA, which, under suitable conditions, will result in an increased detection rate of CpG→CpA variations. If one allele is CpG and the other is TpG, there will be a larger number of circles produced from the TpG allele than the CpG allele, but not as large as the number that would be generated by a CpA. This yields an increased detection rate of CpG→TpG variations, but less than the increase that would be detected for CpA alleles.

The hypothesis above gives results congruent to our findings in Figure S16 in which the highest CpG polymorphic allele fraction is CpG→CpA in position 27, an intermediate fraction is found for CpG→TpG in position 26, and there is no inequality between these fractions elsewhere in the Target10 region. However, the actual prediction of this mechanism is only that in CpG / CpA and CpG / TpG heterozygotes in the target CpG position, the variant allele will generate a larger number of circles than the unchanged CpG. But to move from a hypothesis of circularization bias to an explanation of Target10 position-dependent variation rates requires that the genotypes of the heterozygotic loci subject to the bias must be miscalled as a result of unequal circle formation from the alleles, and so requires the apparent discrepancy in rates to be within the genotype call error rate for heterozygotic loci. However, our comparisons with independently assayed SNPs indicate at most a 3% miscall rate of heterozygous SNPs (Table S3), too small to account for the observed differences in variation rates in positions 26, 27, and 28-35.

SUPPLEMENTAL TABLES

	sequencing run 1			sequencing run 2			intersection of run 1 and 2			
subject	all reads	filtered reads	genotypes	all reads	filtered reads	genotypes	concordance (%)	matching genotypes	dbSNP loci	non-dbSNP het
NA10835	8018986	6161215	441762	8996520	8345133	470072	99.94	436527	3910	87
PGP1	2661906	1476110	291103	2513926	2355095	359611	99.97	284577	2461	65
PGP2	1935688	1605124	316020	2494420	2342450	359256	99.97	307761	2693	67
PGP3	2914546	2497545	366098	4538958	4224097	417536	99.97	362358	3186	86
PGP9	2474571	2094992	353174	3696791	3437129	405522	99.97	348340	3066	66
PGP10	2979256	2511078	372173	4127705	3862489	417082	99.97	366503	3205	215
NA10835-1	3486935	2325087	362598	3908843	3589478	413489	99.96	356176	3133	69
NA10835-2	4532051	3836128	410420	5087677	4755655	436589	99.95	401528	3548	79

Table S1: Aggregate read and genotype statistics by subject**.** Values are given for all six subjects and for the two NA10835 replicates (NA10835-1 and NA10835-2) that were combined for analysis of subject NA10835.

Reads = numbers of Illumina reads, both total (all) and after alignment and filtering (filtered) (see Figure S.1 and text).

As noted in the text, all libraries were sequenced twice (sequencing runs 1 and 2) and analysis of dbSNP loci and candidate novel SNPs were analyzed in the intersection of the two runs. See **Illumina sequence processing** and **Genotype determination** for definition of the terms “detection” and “intersection.”

Concordance = the fraction of Target10 sites for which genotypes were detected in both sequencing runs, for which the genotypes matched identically.

dbSNP loci = number of known dbSNP single nucleotide changes in the Target10 regions in the intersection of the two sequencing runs. See **Known SNP analysis** above for details.

Non-dbSNP het = number of loci in the Target10 regions in the intersection of the two sequencing runs were heterozygous and which did not correspond to dbSNP locations (see **Identification of candidate novel SNPs**).

position	known SNP	SNPs detected			% detected
		run 1	run 2	intersection	
26	1657	1376	1467	1351	81.5
27	1566	1317	1397	1289	82.3
28	213	171	185	167	78.4
29	237	200	209	198	83.5
30	225	177	191	175	77.8
31	233	183	201	182	78.1
32	219	176	195	174	79.5
33	238	184	201	181	76.1
34	220	173	191	169	76.8
35	200	161	172	157	78.5
all	5008	4119	4410	4044	80.8

Table S2. Distribution of locations of known SNPs from dbSNP (see **Known SNP analysis**, above) in the Target10 regions (read positions 26-35). See **Illumina sequence processing** and **Genotype determination** for definition of the terms "detection" and "intersection." As positions 26 and 27 correspond to CpG dinucleotides targeted by the padlock probes, and CpGs are known to have a higher polymorphism rate, the high number of known SNPs in these positions compared to others is expected. Detection of SNPs ranges between 76.1% and 83.5% and falls off with advancing read position. Among subjects, SNP detection is highly correlated with coverage (see Figure S8).

		NA10835	PGP1	PGP2	PGP3	PGP9	PGP10
Independently assayed SNPs	N-all	2025	217	221	246	237	245
	N-het	412	53	56	51	53	63
	N-hom	1613	164	165	195	184	182
	%-het	20.35	24.42	25.34	20.73	22.36	25.71
	%-hom	79.65	75.58	74.66	79.27	77.64	74.29
	source	HapMap	PGP	PGP	PGP	PGP	PGP
exact matches	N-all	1995	217	220	245	236	243
	%-all	***98.52***	***100.00***	***99.55***	***99.59***	***99.58***	***99.18***
	N-het	400	53	55	51	52	62
	N-hom	1595	164	165	194	184	181
	TP-het (%)	97.09	100.00	98.21	100.00	98.11	98.41
	TP-hom (%)	98.88	100.00	100.00	99.49	100.00	99.45
mismatches	N-all	30	0	1	1	1	2
	%-all	1.48	0.00	0.45	0.41	0.42	0.82
	N-het(*)	12	0	1	0	1	1
	N-hom	18	0	0	1	0	1
	%-het-miscall	2.91	0.00	1.79	0.00	1.89	1.59
	%-hom-miscall	1.12	0.00	0.00	0.51	0.00	0.55

Table S3. Accuracy of genotypes computed for Target10 regions based on comparisons with independently obtained genotype data. Genotypes for varying numbers of known SNP sites in Target10 regions were obtained from the HapMap project (The International HapMap Consortium 2003) for NA10835 and the Personal Genome Project (PGP; http://www.personalgenomes.org/) for the other five subjects. In the table above, the designations "het" and "hom" refer to genotype status as given by the independently assayed genotypes. *Independently assayed SNPs:* N-all, N-het, N-hom, %-het, %-hom give total numbers and percentages of SNPs in the Target10 regions provided by the independent data source. *Exact matches:* N-all, N-het, N-hom give the total numbers of SNPs for which the genotypes determined in our study were identical to those given by the independent data source. The bolded line %-all gives the overall accuracy of genotypes determined from our padlock probe sequencing. TP-het and TP-hom give the fractions of the independent source's heterozygous and homozygous genotypes (respectively) that were correctly determined from our padlock probe-based genotype determinations for these SNP classes. *Mismatches:* N-all, N-het, N-hom give the total numbers of SNPs for which the genotypes determined in our study did not match those given by the independent source. %-het-miscall and %-hom-miscall give the fractions of the independent source's heterozygous and homozygous genotypes (respectively) that were incorrectly determined from our padlock probe-based genotype determinations for these SNP classes . For details, see **Genotyping performance characteristics** above.

(*) All N-*het mismatches* for all subjects were such that the genotypes determined by our study were homozygous on one of the alleles given by the independent data source. Therefore, the only miscalls of heterozygous genotypes in our study were failures to detect one of the alleles.

read position	NA10835	PGP1	PGP2	PGP3	PGP9	PGP10	all	hom diff
26	24	11	19	24	13	63	133	3
27	35	32	30	34	29	97	221	5
28	2	1	0	1	1	7	11	0
29	4	3	4	5	3	5	17	0
30	1	1	1	2	5	5	14	2
31	3	1	2	4	2	10	19	1
32	5	5	2	6	2	8	21	1
33	6	4	3	6	3	6	16	1
34	1	4	3	0	3	6	16	0
35	6	3	3	4	5	8	21	0
26-27	59	43	49	58	42	160	354	8
28-35	28	22	18	28	24	55	135	5
26-35	87	65	67	86	66	215	489	13

Table S4. Candidate novel SNPs identified in the Target10 regions in different positions and position bins. See **Identification of candidate novel SNPs** for further discussion. As noted in **Known SNP analysis** and **Identification of candidate novel SNPs**, dbSNP loci were verified for location in the analysis of known SNPs, and a small number of dbSNP loci that were identified as being in the Target10 regions failed this location test. Therefore it is possible that a small number of these candidate novel SNPs might, in fact, already be in dbSNP, but are near indels or have other features that might have caused the location test to fail.

all: Numbers of non-dbSNP loci identified as heterozygous in any of the samples.

hom diff: Numbers of non-dbSNP loci for which all samples for which genotypes could be computed were homozygous, but where at least two distinct homozygous genotypes were identified.

read position	NA10835	PGP1	PGP2	PGP3	PGP9	PGP10
26	0.005488	0.004387	0.005221	0.004911	0.004903	0.006839
27	0.00592	0.006303	0.006348	0.006362	0.006473	0.008437
28	0.000502	0.000419	0.000292	0.000578	0.000401	0.000572
29	0.000526	0.000697	0.000616	0.000688	0.000487	0.001006
30	0.000595	0.000526	0.000616	0.000551	0.000602	0.000872
31	0.0008	0.000806	0.000713	0.000634	0.000516	0.000981
32	0.000826	0.000633	0.000585	0.000774	0.000517	0.000793
33	0.000666	0.000707	0.000618	0.000913	0.000777	0.000984
34	0.000668	0.000605	0.00062	0.00072	0.000749	0.000904
35	0.000599	0.000606	0.000653	0.00061	0.000693	0.00074

read position or ratio	NA10835	PGP1	PGP2	PGP3	PGP9	PGP10	all	all but PGP10	PGP10 to others
26-27	0.0057	0.00534	0.00578	0.00564	0.00569	0.00764	0.00599	0.00564	1.35
28-35	0.00065	0.00062	0.00059	0.00068	0.00059	0.00086	0.00067	0.00063	1.36
26-35	0.00166	0.00157	0.00163	0.00168	0.00161	0.00222	0.00174	0.00163	1.36
27-35	0.00123	0.00126	0.00123	0.00132	0.00125	0.0017	0.00133	0.00126	1.35
(26-27)/(28-35)	8.81	8.55	9.82	8.25	9.60	8.92	8.95	8.96	1.00
(28-35)/.00052	1.25	1.20	1.13	1.31	1.14	1.65	1.29	1.21	1.36
26/(28-35)	8.48	7.02	8.87	7.19	8.28	7.98	7.98	7.97	1.00
27/(28-35)	9.14	10.09	10.78	9.31	10.93	9.85	9.92	9.94	0.99

Table S5. Heterozygosity by read position for all subjects. Heterozygosity is computed as described in **Heterozygosity analysis** above and the captions to Figure S10. **Top:** Heterozygosity at individual read positions. **Bottom:** Heterzogosity aggregated over multiple positions (above dashed lines), and selected ratios of heterozygosities (below dashed line). The column to the right of the double line gives the ratio of PGP10 heterozygosity to that of all other subjects. (28–35)/.00052: Heterozygosity of Target10 region outside of the Target10 region compared to overall 0.00052 nucleotide diversity (π) estimated for chromosome 21 (Sachidanandam et al. 2001).

variation	position	variation rate x 10^{-4}
cg>tg	26-27	45.09
cg>tg	28-35	42.70
cg>ca	26-27	56.15
cg>ca	28-35	42.93
c>t (non-CpG)	28-35	3.21
g>a (non-CpG)		3.72
a>c		1.00
a>g		5.12
a>t		0.49
c>a		1.03
c>g		1.57
g>c		0.84
g>t		1.20
t>a		0.83
t>c		4.49
t>g		0.96

Table S6a. CpG and non-CpG polymorphic allele fractions. See **CpG polymorphic allele fraction** and Figure S16 for details.

variation rate ratio	value
cg>tg(26-27) / cg>tg(28-35)	1.06
cg>ca(26-27) / cg>ca(28-35)	1.31
cg>tg(26-27) / c>t(non-CpG,28-35)	14.05
cg>tg(28-35) / c>t(non-CpG,28-35)	13.30
cg>ca(26-27) / g>a(non-CpG,28-35)	15.10
cg>ca(28-35) / g>a(non-CpG,28-35)	11.54
cg>tg(26-27) / cg>ca(26-27)	0.80
cg>tg(28-35) / cg>ca(28-35)	0.99

Table S6b. CpG-to-TpG and CpG-to-CpA polymorphic allele fractions relative to non-CpG C-to-T and G-to-A polymorphic allele fractions. See **CpG polymorphic allele fraction** and Figure S16 for details.

REFERENCES

Aach, J., Li, J.B., Gao, Y., Porreca, G., Levanon, E., and Church, G.M. 2009. Genotype computation framework that incorporates base-specific error profiles and coverage distributions. *in preparation*.

Bentley, D.R., Balasubramanian, S., Swerdlow, H.P., Smith, G.P., Milton, J., Brown, C.G., Hall, K.P., Evers, D.J., Barnes, C.L., Bignell, H.R., et al. 2008. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**(7218): 53-59.

Karolchik, D., Kuhn, R.M., Baertsch, R., Barber, G.P., Clawson, H., Diekhans, M., Giardine, B., Harte, R.A., Hinrichs, A.S., Hsu, F., et al. 2008. The UCSC Genome Browser Database: 2008 update. *Nucleic Acids Res* **36**(Database issue): D773-779.

Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., and Haussler, D. 2002. The human genome browser at UCSC. *Genome Res* **12**(6): 996-1006.

Lohmueller, K.E., Indap, A.R., Schmidt, S., Boyko, A.R., Hernandez, R.D., Hubisz, M.J., Sninsky, J.J., White, T.J., Sunyaev, S.R., Nielsen, R., et al. 2008. Proportionally more deleterious genetic variation in European than in African populations. *Nature* **451**(7181): 994-997.

Sachidanandam, R., Weissman, D., Schmidt, S.C., Kakol, J.M., Stein, L.D., Marth, G., Sherry, S., Mullikin, J.C., Mortimore, B.J., Willey, D.L., et al. 2001. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409**(6822): 928-933.

Saxonov, S., Berg, P., and Brutlag, D.L. 2006. A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proc Natl Acad Sci U S A* **103**(5): 1412-1417.

Sherry, S.T., Ward, M.H., Kholodov, M., Baker, J., Phan, L., Smigielski, E.M., and Sirotkin, K. 2001. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* **29**(1): 308-311.

Tatusova, T.A. and Madden, T.L. 1999. BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences. *FEMS Microbiol Lett* **174**(2): 247-250.

The International HapMap Consortium. 2003. The International HapMap Project. *Nature* **426**(6968): 789-796.