Genotyping Rare Variants

A simulated analysis achieves high call rates and low error rates from loci containing rare variants.

Introduction

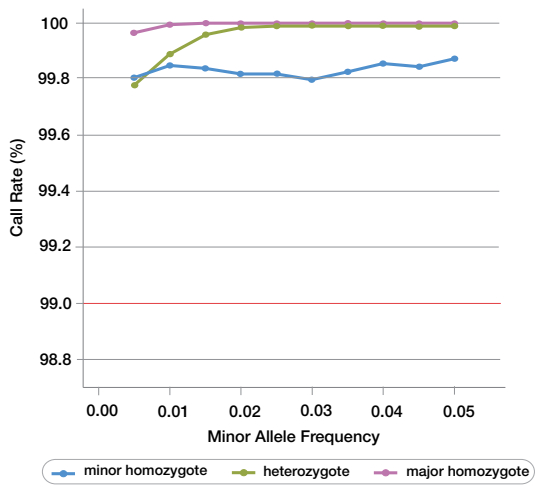
Over the last decade, the International HapMap project has been the major resource of validated SNPs for the development of whole-genome genotyping arrays. SNPs chosen for the HapMap database and used in commercial array products were selected to preferentially include common (minor allele frequency (MAF) > 5%) variants and provide only a small subset of less common (MAF < 5%) variants. As a result, the arrays designed from this data set have primarily targeted common variants.

By sequencing thousands of individuals across many populations, the 1000 Genomes Project is seeking to identify SNPs down to 1% MAF, dramatically expanding the catalog of known human variants. This data set is expected to provide a more comprehensive understanding of the true spectrum of variation in human populations and provide a more accurate picture of the linkage disequilibrium (LD) structure of the human genome, enabling researchers to explore the role of rarer variants in traits and diseases. Illumina is leveraging these data to develop the next generation of whole-genome genotyping arrays, improving genomic coverage and fueling new discoveries.

As the target MAF of the next-generation microarrays decreases, it is essential that the quality of these rarer variant genotyping calls remain high. Genotype calls are normally made using a training data set, where the assay signal is referenced against the expected cluster positions of the three genotype classes: homozygous major, homozygous minor, and heterozygous. However, as the MAF for a given SNP becomes increasingly rare, the training data sets will contain fewer examples of the homozygous rare and heterozygous genotypes. This can impact the reliability of genotype calls made using this limited data set. For common variants, it is well understood that re-clustering the experimental data set using GenomeStudio® software can significantly improve both call rates and reproducibility for a given study, but there is limited data showing how this re-clustering approach will perform for rare variants.

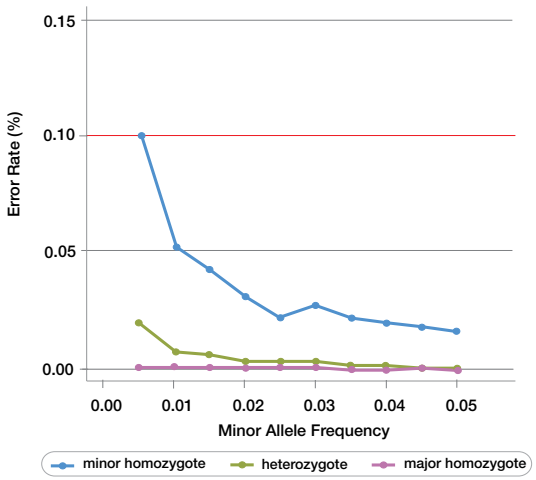
One challenge with determining the quality of rare variant genotyping calls is the lack of a “gold standard” reference source for the true genotypes of a given SNP in a given sample. One approach is to calculate transmission rates and Mendelian inheritance errors from a large data set of related individuals; however, data sets of this type are not widely available. An alternative approach is to simulate rare variants by down sampling from common, high-quality, and well-accepted SNP genotype data and comparing the new call rate and genotype information with the original genotype calls. This technical note presents the results of such a rare variant genotyping simulation and demonstrates that these variants can be called with high accuracy.

Figure 1: Call Rates for Simulated Genotype Calling



A simulated analysis shows that high call rates (> 99%) can be achieved for alleles that exhibit low minor allele frequencies. The red line indicates the Illumina product specification, 99%.

Figure 2: Error Rates for Simulated Genotype Calling



Low error rates (< 0.1%) can be achieved for all three genotype classes for alleles that exhibit very low minor allele frequencies. The red line indicates the Illumina product specification, 0.10%.

AGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAAC
TCAACGTACCGTAACGAACGTATCATTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGACGAAAAGAATGATAACAGTAACACACTTCTGTTAACCTT
CGACGAAAAGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACC
AACGTACCATTAAGAGCTACCGTGCAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGACGAAAAGAATGATAA
AGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAAC
GATTACTTGATCCACTGATTCAACGTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAGCAACGAGC
CGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGAAAAGAATGATAACAGTA

Simulated Rare Variant Calling

Analysis was based on a set of 2,000 samples genotyped using the Illumina Human1M-Duo BeadChip. The data set was reduced to focus on 10,000 common loci that have at least 125 observations of each of the three possible diploid genotypes.

For the simulation, one of the two alleles at a given locus was randomly selected to be the rare variant. Observations were simulated for that locus by randomly selecting 1000 samples from the total sample set using an appropriately parameterized multinomial distribution, where individuals who are homozygous for the rare allele will be present with probability p^2, heterozygotes with probability 2pq, and common homozygotes with probability q^2. This sampling approach ensures that the simulated data set conforms to Hardy-Weinberg Equilibrium. The desired number of individuals was sampled from each class to construct the final sample set for each simulated loci. Finally, the no-call rate from each common locus was used to randomly replace samples in the corresponding simulated locus with the same proportion of no-called samples (i.e. samples that are not near any of the three clusters, and therefore are not assigned a genotype).

These simulated rare loci were then clustered with the GenTrain algorithm in GenomeStudio software. Genotypes were called for each sample and error rates were calculated for each of the three genotype classes. An error was defined as a sample that produced different genotype calls for the original common locus and the simulated locus. For each locus, the rare allele was simulated at incremental minor allele frequencies, from 0.5% up to 5%. The simulations were repeated for each MAF until 500,000 examples of the minor homozygote had been simulated.

As shown in Figure 1, at very low minor allele frequencies (< 0.1 to 0.5), the call rate for all three genotype classes exceeds 99% (Illumina's minimum average call rate specification for whole-genome Infinium® arrays). As expected, the major homozygote exhibits very high call rates (> 99.9%), but even the minor homozygote exceeds the 99% threshold. These results indicate that high call rates can be achieved for rare alleles. Over the same MAF range, errors rates remain remarkably low, demonstrating that accurate genotype calls are possible for rare alleles (Figure 2). The Infinium product specification for error rate is < 0.1%. Figure 2 shows that all simulations exceed this specification, with the exception of MAF 0.5% minor homozygote, but even that allele is virtually on the specification line.

Summary

As researchers begin to leverage 1000 Genomes Project data to explore the role of rare variants in human traits and diseases, it will be essential to maintain high-quality genotype calls. The lack of a standard reference of known genotype information for rare variants creates a challenge for assessing call accuracy. Using simulated rare variant loci created by down sampling from a high-quality common variant data set, this analysis demonstrated that accurate genotype calls and low error rates can be achieved with very low MAF alleles.

Illumina, Inc. • 9885 Towne Centre Drive, San Diego, CA 92121 USA • 1.800.809.4566 toll-free • 1.858.202.4566 tel • techsupport@illumina.com • illumina.com

FOR RESEARCH USE ONLY

© 2010 Illumina, Inc. All rights reserved.
Illumina, illumina*Dx*, Solexa, Making Sense Out of Life, Oligator, Sentrix, GoldenGate, GoldenGate Indexing, DASL, BeadArray, Array of Arrays, Infinium, BeadXpress, VeraCode, IntelliHyb, iSelect, CSPro, GenomeStudio, Genetic Energy, HiSeq, and HiScan are registered trademarks or trademarks of Illumina, Inc. All other brands and names contained herein are the property of their respective owners. Pub. No. 370-2010-008 Current as of 14 April 2010

illumina®

AGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAAC
TCAACGTACCGTAACGAACGTATCATTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGACGAAAAGAATGATAACAGTAACACACTTCTGTTAACCTT
CGACGAAAAGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACC
AACGTACCATTAAGAGCTACCGTGCAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGACGAAAAGAATGATAA
AGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAAC
GATTACTTGATCCACTGATTCAACGTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAGCAACGAGC
CGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGAAAAGAATGATAACAGTA