Validation of 1000 Genomes Project SNP calls using the Affymetrix® Axiom™ Genotyping Solution

Jeremy Gollub, Tom Asbury, Jonathan Bleyhl, Yontao Lu, Gangwu Mei, Matthew Purdy, Li Weng, Yiping Zhan, Michael H. Shapero and Teresa Webster

Affymetrix, Inc., 3420 Central Expressway, Santa Clara, CA 95051

Affymetrix® Revolutionize life

ABSTRACT #1925

Background: We have validated approximately 3 million novel human SNPs discovered by the 1000 Genomes Project that were not found in the HapMap project or dbSNP (release 130), for use with the Axiom™ Genotyping Solution. We observed very high call rates and good concordance to 1000 Genomes Project results for individuals genotyped in common. Here we present analyses of coverage and linkage disequilibrium (LD), genotype concordance and accuracy, and SNP validation rates for a set of SNPs discovered on chromosome 3 by the 1000 Genomes low-coverage (2x-4x) and high-coverage (20x-60x) pilot projects, selected for expected polymorphism in the Yoruba (YRI) population.

Results: We genotyped the HapMap Yoruba population (90 individuals) on a set of screening microarrays containing novel 1000 Genomes Project SNP content. We found very high concordance to high-coverage-derived genotypes and somewhat lower concordance to low-coverage-derived results. SNP validation rates were higher for SNPs discovered in high-coverage sequencing than for SNPs discovered only by the low-coverage project. Surprisingly, we found a difference in apparent LD with previously known SNPs between novel SNPs discovered by low- and high-coverage sequencing. SNPs discovered only by low-coverage sequencing had lower LD to, and were not as well tagged by, HapMap3 SNPs.

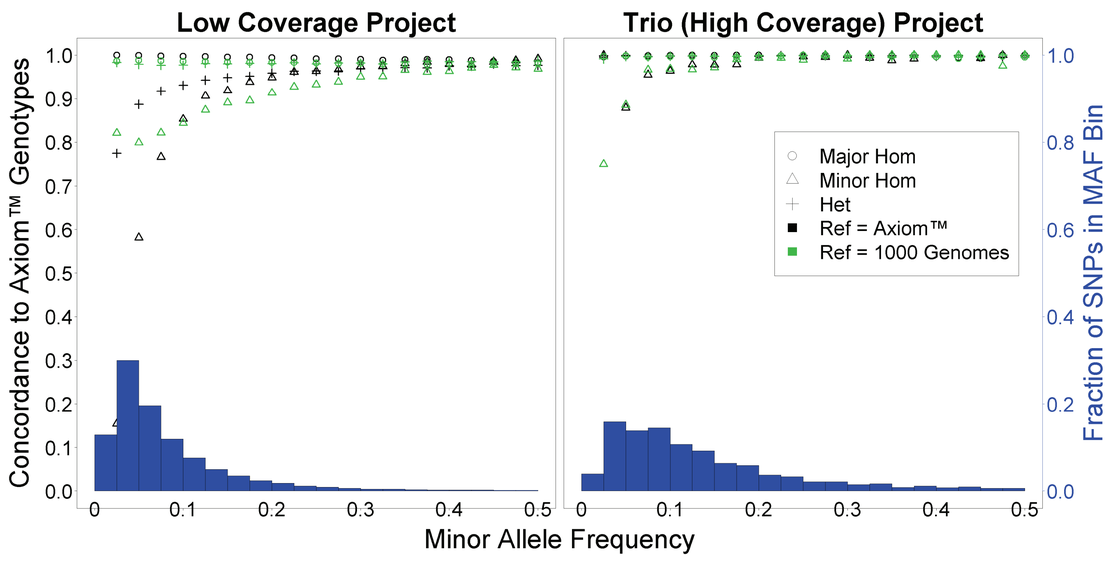
References:

- 1000 Genomes Project: www.1000genomes.org
- Axiom™ Genotyping Solution: www.affymetrix.com

We gratefully acknowledge the 1000 Genomes Project for allowing public access to, and use of, their data.

Genotype accuracy

Figure 1: Detailed genotype concordance between technologies.



Detailed concordance of genotype calls from the Axiom platform to calls from low- and high-coverage sequencing. Data are shown for SNPs discovered on chromosome 3 in the low-coverage (left panel) and trio (right panel) projects, selected for expected polymorphism in the Yoruba population and validated for use with the Axiom platform. Concordance values are the average over all SNPs in each minor allele frequency (MAF) range, considering all samples genotyped in common; MAF was calculated based on Axiom calls for 90 Yoruba individuals. Concordance was calculated taking as the reference either sequence-derived genotypes (black) or microarray-derived genotypes (green). Axiom concordance was significantly better for the high-coverage results. In particular, the low-coverage results under-called heterozygotes, as would be expected, especially at lower MAF.

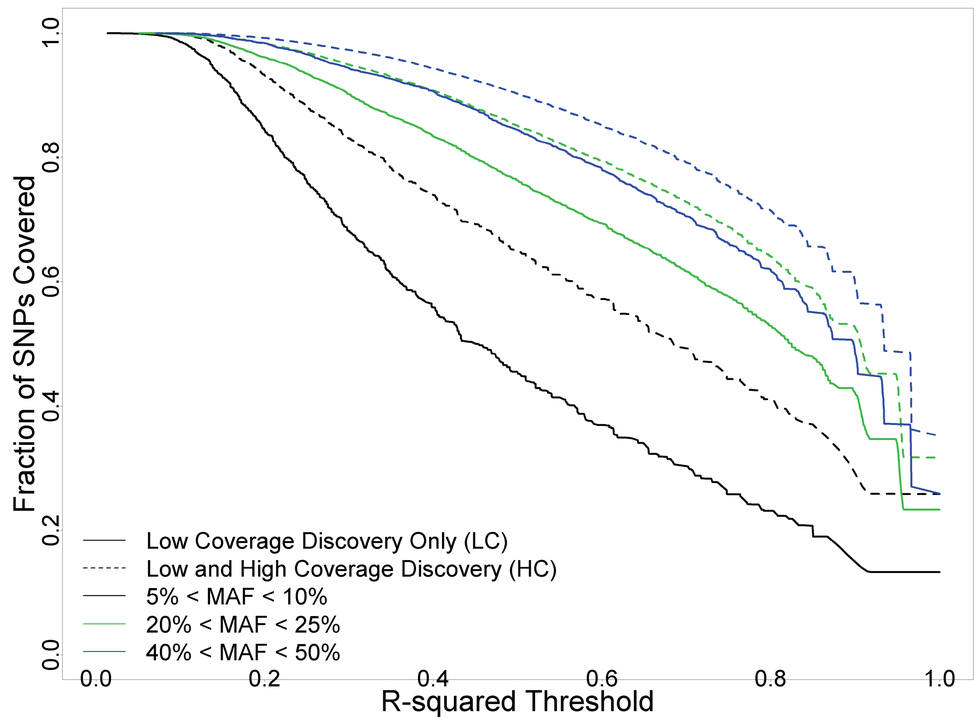
Table 1: Average genotype concordance between platforms.

	Low coverage		High coverage	
Reference	**Sequencing**	**Axiom**	**Sequencing**	**Axiom**
Overall	98.5%		99.7%	
Major homozygote	98.6%	99.7%	99.7%	99.9%
Heterozygote	97.9%	90.1%	99.7%	99.8%
Minor homozygote	88.9%	88.4%	98.3%	98.5%

Average concordance of genotype calls from the Axiom platform to calls from low- and high-coverage sequencing. Concordance was significantly better for high-coverage sequencing results; low-coverage sequencing tended to miss heterozygotes, as expected.

Coverage and linkage disequilibrium in YRI

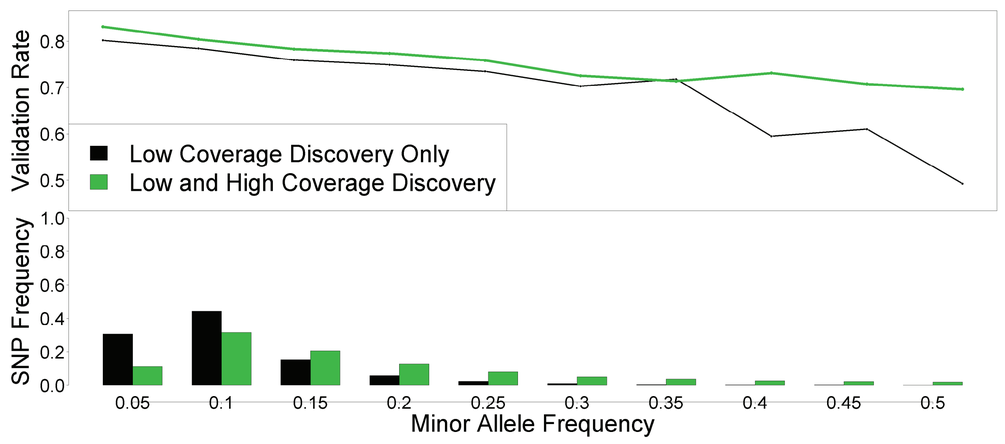
Figure 2: Genetic coverage of novel SNPs.



Coverage provided by HapMap3 SNPs. Chromosome 3 SNPs discovered in the Yoruba population in the low-coverage project were divided into three mutually exclusive groups: SNPs present in the HapMap3 database (HM); SNPs discovered only in the low-coverage project (LC); SNPs also discovered in the trio project (HC). Single-marker coverage of LC and HC, as provided by HM, is shown. LC contains a greater proportion of SNPs with low LD to HapMap3 SNPs. Some MAF ranges omitted for clarity.

Validation rates

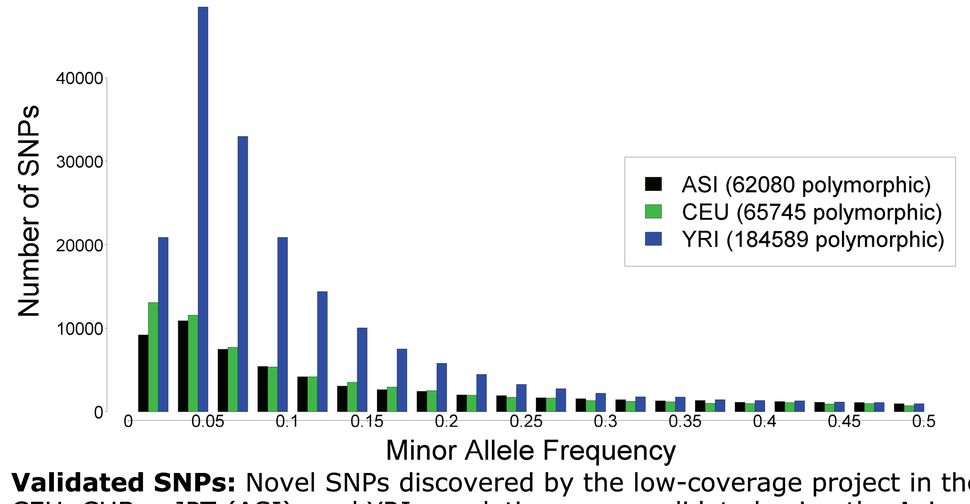
Figure 3: SNP validation rates as a function of sequence depth.



Validation rates: SNPs were validated based on cluster resolution and observed polymorphism, requiring at least three examples of the minor allele. Among SNPs satisfying the polymorphism requirement according to the low-coverage project results, SNPs also discovered in the trio project were validated at a higher rate across the MAF range. The discrepancy increases at high MAF. This may reflect false SNP discoveries in the low-coverage results; at higher MAF, the proportion of true discoveries that should also be found in a single, arbitrary individual increases.

Validated chromosome 3 SNPs

Figure 4: MAF distribution of validated SNPs.



Validated SNPs: Novel SNPs discovered by the low-coverage project in the CEU, CHB + JPT (ASI), and YRI populations were validated using the Axiom Genotyping Solution. The distributions of minor allele frequencies (MAF) for SNPs on chromosome 3 are shown.