Large-Scale SNP Detection via Ligation-based Dibase Sequencing Across Multiple HapMap Individuals: NA18507, NA19240, and NA12878

Stephen F. McLaughlin[1], Heather E. Peckham[1], Swati S. Ranade[1], Clarence C. Lee[1], Christopher R. Clouser[1], Jonathan M. Manning[1], Cynthia L. Hendrickson[1], Lei Zhang[1], Eileen T. Dimalanta[1], Tanya D. Sokolsky[1], Jeffrey K. Ichikawa[1], Jason B. Warner[1], Mike W Laptewicz[1], Brittney E Coleman[1], Fiona C Hyland[2], Jeffrey G. Reid[4], Aniko Sabo[4], Donna M. Muzny[4], Richard A. Gibbs[4], Alan P Blanchard[1], Joel A. Malek[3], Gina L Costa[1] and Kevin J. McKernan[1]

1. Applied Biosystems, 500 Cummings Center, Beverly, MA 01915, 2. Applied Biosystems, 850 Lincoln Centre Dr, Foster City, CA 94404, 3. Weill Cornell Medical College in Qatar, Doha, Qatar, 4. Baylor College of Medicine, Houston, TX 77030

ABSTRACT

The HapMap project along with next-generation sequencing technologies provides unprecedented opportunities to fully characterize whole-genome polymorphism events comprising many individuals across multiple populations. Genetic variants such as single-nucleotide-polymorphisms (SNPs), small indels, large-scale indel events on the order of several kilobases, genomic rearrangements such as inversions and translocations, and even full-scale *de novo* sequencing can be characterized rapidly and at per-base cost orders of magnitude less than the original Human Genome Project. Three HapMap samples were sequenced via the ligation–based approach utilized in the SOLiD™ sequencing system: two Yoruba samples NA18507 (18x) and NA19240 (26x), and one CEPH sample NA12878 (12x) using paired-end libraries with various insert sizes (600bp-3.5kb) as well as several fragment libraries. A total of 6.9M distinct variant-allele SNPs were detected across the three genomes via a heuristic approach which considers the number of reads per allele as well as a score which weights the SNP calls based on the error profile of the reads. The total numbers of heterozygous SNPs, homozygous SNPs, and %dbSNP v129 concordance per sample detected (presented in this order) are as follows: NA18507 (2.33M,1.53M,81%), NA19240 (2.51M, 1.54M,79.1%), and NA12878 (1.46M,1.68M,87.9%). The higher concordance of NA12878 to dbSNP may reflect a bias in dbSNP toward entries for the CEPH population. Since NA19240 was sequenced to a greater depth, there are fewer under called heterozygous SNPs in this dataset relative to the others. We present an analysis of the SNPs identified in the three samples including a greater degree of overlap between the two Yoruba samples than the CEPH sample as expected. Novel SNPs have been validated via Sanger Sequencing for NA18507 (111/112 called heterozygotes) as well as NA19240 (434/448).

INTRODUCTION

In many resequencing projects one of the most important objectives is to measure Single Nucleotide Polymorphisms (SNPs) that may be responsible for differences in phenotype. Due to the fact that each base is measured twice, a single base change in base space leads to 2 changes in color space. Any color space change which contradicts this rule is considered invalid and is likely a measurement error. This feature of color space is very powerful in aiding SNP detection as it vastly reduces the error rate and improves consensus accuracy.

Figure 1. Valid vs. Invalid Color-Space Changes applied to SNP Detection

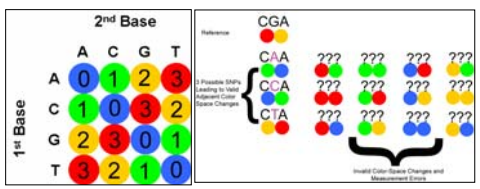

Figure 1: Each of the 16 dibase combinations is represented by one of 4 different-colored dyes which are in turn shared evenly between 4 dibase combinations. By convention, these are represented as numbers (shown above) which comprise the alignments of individual tags as illustrated in Figure 5. Because each base is interrogated twice by two different oligos, a single-base change leads to 2 color-space changes. This means that once a tag is confidently aligned to a reference sequence, only 4 dibase combinations are valid: agreement with the reference and 3 adjacent color space changes. These Color Space changes are referred to as *Valid Adjacent* changes and are indicative of a SNP (see Figure 2). On the other hand, if any of the other 12 color-space changes are observed the reference alignment makes no sense and is likely a measurement error. These are referred to as *Invalid* changes.

RESULTS

Figure 2. Cumulative Coverage for NA18507, NA19240, and NA12878

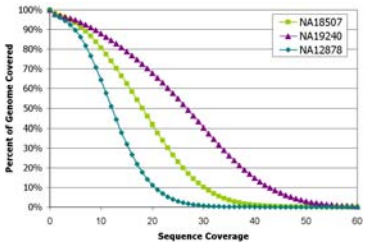

Figure2: Cumulative Sequence Coverage for NA18507, NA19240, and NA12878 Sequenced by SOLiD™ (i.e. 10% of NA12878 is sequenced to >=20x).

Table 1. Average Paired-End and Fragment Coverage per Genome

	Total Coverage	Mate Pair	Fragment	Sequenced by
Yoruba male (NA18507)	17.9x	14.9x	3x	AB
Yoruba female (NA19240)	26X	14.6x	11.4x	Baylor & AB
CEPH female (NA12878)	12.1X	12.1x	n/a	Broad

Table1: Average Sequence Coverage for NA18507, NA19240, and NA12878 Sequenced by SOLiD™. NA18507 and NA19240 were sequenced with a mixture of Fragment and Paired-end libraries, and NA12878 was sequenced with Paired-end libraries only. These coverage levels along with the cumulative coverage (Figure 2) is for non-redundant pairs and uniquely mapping fragment reads; it is these reads which were used for SNP calling these three genomes.

Figure 3. Total Heterozygous and Homozygous SNPs for NA18507, NA19240, and NA12878 per chromosome with dbSNP concordance.

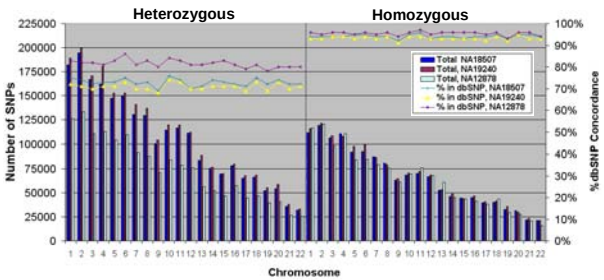

Figure 3: The total number (left y-axis) of heterozygous (left half) and homozygous (right half) SNPs discovered in NA18507 (dark blue), NA19240 (violet) and NA12878 (light blue). Also shown is the dbSNP[2] concordance (version 129) for the SNPs discovered in all three genomes. For all genomes, fewer of the heterozygous SNPs are also found in dbSNP compared to homozygous SNPs. This is likely because heterozygous SNPs are less likely to be found in dbSNP, perhaps because they are more difficult to detect. Despite the lower coverage for NA12878 (12.1x) compared to NA18507 (17.9x) and NA19240 (26x), there are a higher percentage known heterozygous SNPs likely because SNPs from the CEPH individual are more likely to be in dbSNP compared to the two Yoruba samples less Yoruba SNPs being discovered to date and consequently deposited into dbSNP.

Figure 4. Overlapping SNPs Between 6 Human Genomes

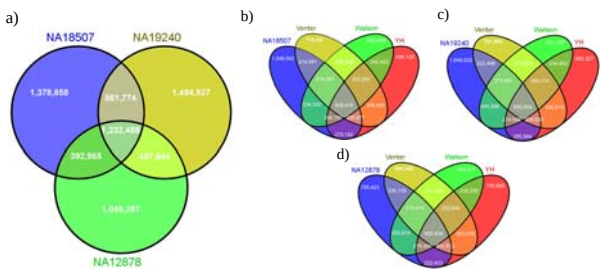

Figure 4: Venn diagrams for overlapping reference-variant SNPs across the 3 genomes sequenced with SOLiD™, NA18507, NA19240, and NA12878 (a) and each of the 3 SOLiD™-sequenced genomes compared to recently published SNPS for Venter[1], Watson[4], and the YH Asian[3] genome (NA18507, (b); NA19240, (c); and NA12878 (d)). As expected, the two Yoruba samples have the most overlap between each other. Also, the two Yoruba samples have more SNPs unique to themselves than to the Venter, Watson, and YH (b,c) samples as well as each other (a). Despite being sequenced to lower coverage, there is more overlap between NA12878 and the other published genomes (d) as expected likely because Venter, Watson, and NA12878 are all of European decent.

Figure 5. SNP Detection at Various Levels of Average Coverage

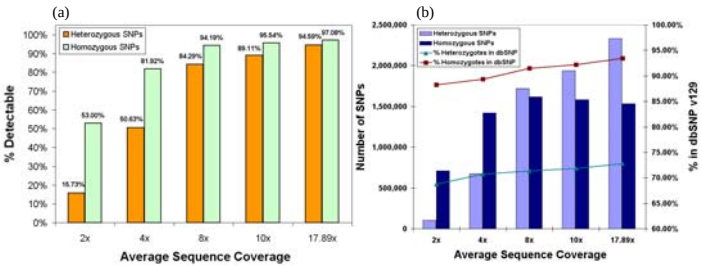

Figure 5: To assess the capabilities of SNP Detection at various average coverage levels, different SOLiD™ runs were grouped together to yield various levels of average coverage ranging from 2-17.89x for the NA18507 sample. Since we require a minimum of two reads for homozygous SNP detection and 4 reads (2 per allele) for heterozygous detection, we can calculate the upper-limit of potential SNP detection by assessing the percentage of the genome covered at >=2x (homozygous SNPs) and >=4x (heterozygous SNPs) (a). For heterozygous detection, this does not take into account the probability of sampling both alleles, but it nonetheless provides an estimate for the upper limit of the proportion of the genome which is detectable. We also show the result of running our SNP calling algorithm on these same data sets (b) and present the number of homozygous and heterozygous SNPs detected along with the dbSNP concordance (v129) for these SNPs.

Figure6. Average Paired-End and Fragment Coverage per Genome

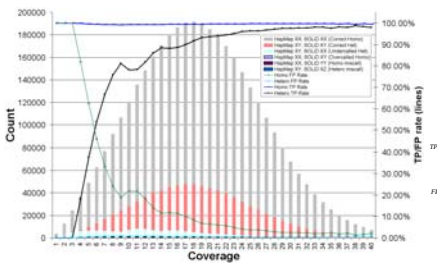

Table 2. Using HapMap to Estimate FP/TP Rate

HapMap Genotype	SOLiD™ Genotype	Het-context Verdict	Homo-context Verdict	Name
XX	XX	TN	TP	homo_saw_homo
XY	XY	TP	TN	het_saw_het
XY	XX	FN	FP	het_saw_homo
XX	XY	FP	FN	homo_saw_het

$$TPR = \frac{TP}{TP + FN} \times 100\%$$

$$\text{HeterozygousTPR} = \frac{\text{het\_saw\_het}}{\text{het\_saw\_het} + \text{het\_saw\_homo}} \times 100\%$$

$$\text{HomozygousTPR} = \frac{\text{homo\_saw\_homo}}{\text{homo\_saw\_homo} + \text{homo\_saw\_het}} \times 100\%$$

$$FPR = \frac{FP}{TP + FN} \times 100\%$$

$$\text{HeterozygousFPR} = \frac{\text{homo\_saw\_het}}{\text{homo\_saw\_het} + \text{homo\_saw\_homo}} \times 100\%$$

$$\text{HomozygousFPR} = \frac{\text{het\_saw\_homo}}{\text{het\_saw\_homo} + \text{het\_saw\_het}} \times 100\%$$

Figure6 + Table 2: HapMap genotypes can be used to estimate the FP and TP rate of our SOLiD™ genotype calls (Table 2 clarifies this method). We used our genotype calls for NA18507 and compared them to the annotated genotypes part of HapMap r26[5]. All of our calls are presented (Figure 6) with the counts shown on the left Y-axis and the TP/FP rate graphed on the right Y-axis. We binned the calls by coverage to demonstrate the effect of coverage on the success of the genotype calling. At fairly low levels of coverage, heterozygous genotypes are getting undercalled (as expected). As coverage increases, both the homozygous FP-rate decreases (meaning fewer heterozygous genotypes are getting erroneously under called as heterozygous SNPs) and the heterozygous TP-rate increases (meaning both alleles are increasingly getting sampled and accurately detected). Coverage is not an issue for the TP-rate of homozygous SNP detection (which is >99% irrespective of the coverage) because sampling of both alleles is not an issue. This also applies to the FP-rate of heterozygous SNP detection (because true homozygous SNPs are very rarely detected as heterozygous by SOLiD™ at any coverage level, except for a very small number of calls in areas of extremely high coverage (>> the mean) which are likely caused by mismapping to regions containing segmental duplications or other repeat elements.

REFERENCES

1. Levy, S., Sutton, G., Ng, P.C., Feuk, L., Halpern, A.L., Walenz, B.P., Axelrod, N., Huang, J., Kirkness, E.F., Denisov, G. et al. 2007. The diploid genome sequence of an individual human. PLoS Biol 5(10): e254.
2. ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 2001 Jan 1;29(1):308-11.
3. Wang, J., Wang, W., Li, R., Li, Y., Tian, G., Goodman, L., Fan, W., Zhang, J., Li, J., Guo, Y. et al. 2008. The diploid genome sequence of an Asian individual. Nature 456(7218): 60-65.
4. Wheeler, D.A., Srinivasan, M., Egholm, M., Shen, Y., Chen, L., McGuire, A., He, W., Chen, Y.J., Makhijani, V., Roth, G.T. et al. 2008. The complete genome of an individual by massively parallel DNA sequencing. Nature 452(7189): 872-876.
5. The International HapMap Consortium. 2003. The International HapMap Project. Nature 426:789-796.