The Detectable Genome: How much of the human genome is accessible to variant discovery by next-generation sequencing?

Heather E. Peckham[1], Yutao Fu[1], Stephen F. McLaughlin[1], Eric F. Tsung[1], Swati S. Ranade[2], 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], Bin Li[2], Alan P Blanchard[1], Joel A. Malek[3], Gina L Costa[1] and Kevin J. McKernan[1]

Applied Biosystems, [1]Beverly, MA, USA, [2]Foster City, CA, USA
[3]Weill Cornell Medical College in Qatar, Doha, Qatar

ABSTRACT

The human genome is being vigorously sequenced in an effort to understand the extent of normal human variation as well as disease causing variants. This initiative brings with it the challenge of assessing the areas of the human genome that are accessible to variant detection. We illustrate the amount of the human genome that is covered with uniquely placed single tags and uniquely placed mate pairs and demonstrate how both larger insert sizes and read lengths increase the portion of the genome that is uniquely mappable by paired-end tags. We use various human genomes (NA18507 – 10x Yoruban male, NA19240 – 26x Yoruban female) sequenced with SOLiD^{TM} sequencing to illustrate the amount of SNPs and indels that are detected at various levels of average sequence coverage. We also demonstrate the sequence and clone coverage needed to identify indels of any size between paired-end reads. We use libraries with an assortment of insert sizes to show that larger libraries increase the accessibility of the genome by spanning larger insertions. We show that the bisulfite converted human genome is less uniquely mappable than the normal human genome but significantly less signature is lost in color space than in base space. We also illustrate that a significant portion of large segmental duplications are accessible to sequence and clone coverage by paired-end reads. These principles are applicable to all next-generation sequencing platforms and are essential to comprehend the amount and location of variability in the human genome.

Sequence Coverage

Increased coverage gained from mate pair libraries demonstrates that a more comprehensive sampling of the human genome is achieved with uniquely placed mate pairs than with the unique placement of each of the tags.

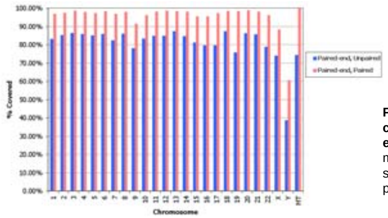


Figure 1. Sequence coverage of each chromosome of NA18507 by 2x25 paired-end reads. The coverage is separated by mate pair data treated as uniquely placed single tags and as uniquely placed mate pairs.

The mappability of the human genome with uniquely placed mate pairs increases with both insert size and tag length.

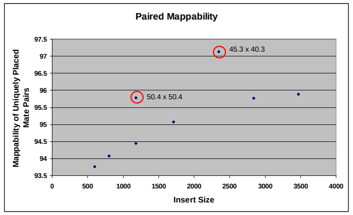
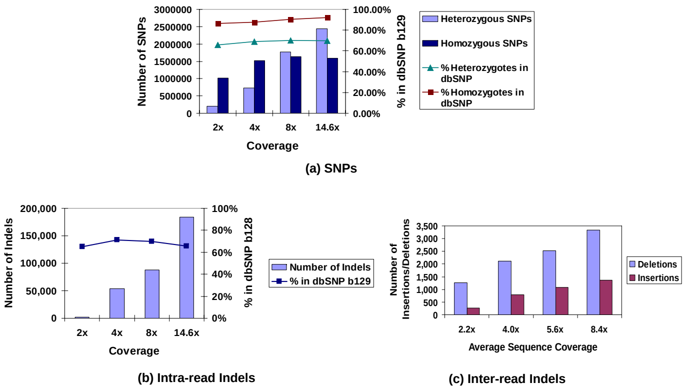


Figure 2. The mappability of uniquely placed mate pairs as a function of insert size and tag length. All data points represent the mappability of uniquely placed paired 25-mers allowing up to 2 mismatches in each tag (25.2 x 25.2) unless otherwise indicated. N's in the reference sequence are discarded in these calculations.

SNP and Indel Detection at Various Coverage Levels

Significant homozygous SNP detection can be achieved at 2X with a significant increase at 4X but only a modest increase at 8X. Heterozygote SNP detection requires more coverage and increases steadily at these coverage levels. The identification of small indels under the sequence read is fairly incomplete at these low coverage depths due to more stringent mapping requirements. The number of inter-read insertions and deletions rises steadily from 2.2x to 8.4x average coverage and indicates that further sequencing will enable more variants to be detected.



(a) SNPs

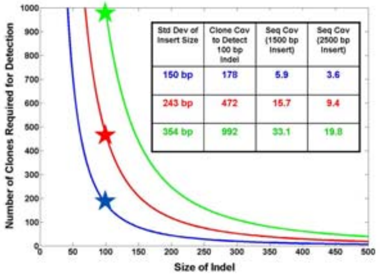
(b) Intra-read Indels

(c) Inter-read Indels

Figure 3. Subsets of SOLiD data that accumulate to various levels of average sequence coverage and assessment of how this affects the number of SNPs, intra-read indels <= 11 bp and inter-read indels >= 200 bp that are detected.

The Size Limit of Inter-read Indel Detection

A novel approach for detecting indels between paired-end reads allows the detection of variants that are smaller than the standard deviation of the insert size of the library. The size of an indel that can be detected depends on the standard deviation of the insert size of the library as well as the clone coverage at the site of the variation.



Std Dev of Insert Size	Clone Cov to Detect 100 bp Indel	Seq Cov (1500 bp Insert)	Seq Cov (2500 bp Insert)
150 bp	178	5.9	3.6
243 bp	472	15.7	9.4
354 bp	992	33.1	19.8

Figure 4. Sequence and clone coverage required to detect insertions and deletions at 6 standard deviations of significance. The data illustrates the size limit of detection of large inter-read insertions and deletions at each level of clone coverage.

Larger Insert Size Libraries Increase the Accessibility of the Human Genome by Spanning Larger Insertions

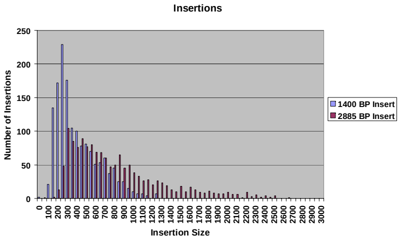


Figure 5. The distribution of insertion sizes detected with a 1,400 bp library and a 2,885 bp library. While the libraries have different levels of sequence and clone coverage so the absolute number of insertions cannot be directly compared, it is evident that the insert size of the library is the upper limit on the size of insertions that can be detected.

The Mappability of the Normal and Bisulfite Converted Human Genome

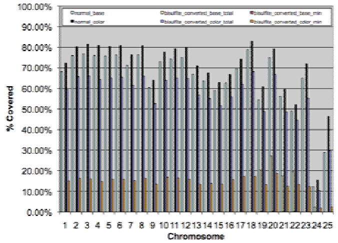


Figure 6. The bisulfite converted human genome is less uniquely mappable than the normal human genome but significantly less signature is lost in color space than in base space. All mappability calculations are for 25-mers allowing up to 2 mismatches. normal_base and normal_color are the mappability of the human genome in base space and color space, respectively. bisulfite_converted refers to the human genome is which all of the C's have been changed to T's. Total refers to unique mapping on either strand and min refers to unique mapping on both strands to afford strand-specific methylation studies. There is little mappable on both strands in base space and thus it is difficult to see on the Y axis. These values are shown only for chromosomes 19-25 as over 99% of these regions are not mappable and thus are computer intensive. The mappability of both strands in color space is far better than in base space. Chromosomes 23, 24 and 25 refer to X, Y and Mitochondria, respectively.

Large Segmental Duplications are Accessible to Sequence and Clone Coverage by Paired-end Reads.

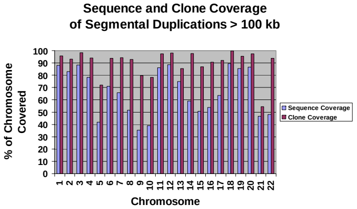


Figure 7. Sequence and clone coverage of segmental duplications > 100 kb in the autosomal human genome by paired-end tags (2x50).

For Research Use Only. Not for use in diagnostic procedures.