SEQUENCE AND STRUCTURAL VARIATION IN A HUMAN GENOME

Peckham, H.E.[1], McLaughlin, S.F.[1], Fu, Y.[1], Tsung, E.F.[1], Hyland, F.C.[2], Clouser, C.R.[1], Duncan, C[1]., Ichikawa, J.K.[1], Lee, C.C.[1], Zhang, Z.[2], Ranade, S.S.[2], Dimalanta, E.T.[1], Sokolsky, T.D.[1], Zhang, L.[1], Li, B.[2], Kotler, L.[1], Stuart, J.R.[1], Malek, J.A.[3], Manning, J.M.[1], Antipova, A.A.[1], Perez, D.S.[1], Hayashibara, K.C.[2], Lyons, M.R.[1], Coleman, B.E.[1], Laptewicz, M.W.[1], Bafna, V.[4], Bashir, A.[4], Eichler, E.E.[5], De La Vega, F.M.[2], Blanchard A.P[1], Costa, G.L.[1] and McKernan, K.J.[1]

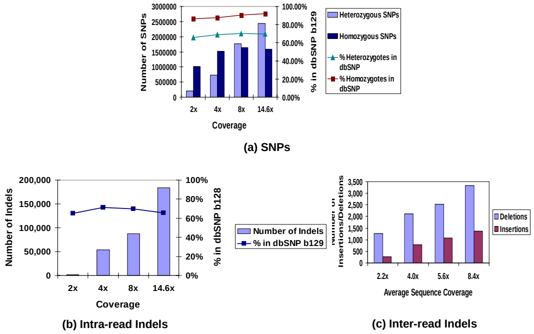
[1]Applied Biosystems, Beverly MA, 01915, [2]Foster City CA, 94404, [3]Weill Cornell Medical College in Qatar, Doha, Qatar, [4]University of San Diego, La Jolla, CA 92093, [5]School of Medicine, University of Washington, Seattle, WA 98195, USA.

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. We describe the genome sequencing of an anonymous individual of African origin using the SOLiD$^{\mathrm{TM}}$ ligation based sequencing assay that enables a unique form of error correction that improves the raw accuracy of the aligned reads to >99.9% allowing accurate SNP detection while also improving the detection of intra-read indels and the mapping of bisulfite converted reads. We use ~18X haploid coverage of aligned sequence and nearly 300X clone coverage to identify over 3.8 million SNPs (19% of which are not in dbSNP), 229,375 intra-read indels (33% of which are not in dbSNP), 5,590 indels between mate pair reads, 78 inversions and 4 gene fusions. We present a novel approach for detecting indels between mate pair reads that are smaller than the standard deviation of the insert size of the library and discover deletions in common with those detected with our intra-read approach and thus introduce the first time that a single technology has traversed the previously unattainable gap between small intra-read and large inter-read indel detection. We illustrate the additional use of split reads to achieve breakpoint resolution of detected structural variations. We explore phasing of heterozygous SNPs and illustrate that 43% of those we detect are in phase with at least one other heterozygous SNP and provide numerous examples of heterozygous SNPs in phase with heterozygous structural variations providing the potential to phase and possibly genotype structural variations with single base changes. We also address the challenge of assessing the areas of the human genome that are accessible to variant detection with short reads by illustrating the portions of the genome that gain coverage as insert sizes and read lengths are increased, the amount of the genome that meets the requirements for SNP and indel detection at various levels of average sequence coverage and how this corresponds to the number of actual variants detected, the sequence and clone coverage needed to identify indels of any size between mate pair reads and the portion of large segmental duplications that are accessible to sequence and clone coverage by mate pair reads. There is more genetic variation in the human genome still to be uncovered and these principles are essential to comprehend the amount and location of this variability.

Evaluation of Sequencing Strategies for Genetic Variation Discovery

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.

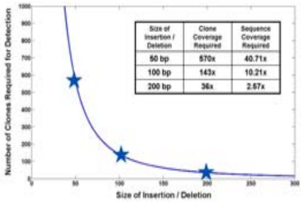


(a) SNPs

(b) Intra-read Indels

(c) Inter-read Indels

Figure 1. 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.



Size of Insertion / Deletion	Clone Coverage Required	Sequence Coverage Required
50 bp	570x	40.71x
100 bp	143x	10.21x
200 bp	36x	2.67x

Figure 2. Sequence and clone coverage required to detect insertions and deletions at 6 standard deviations of significance given a 1400 bp insert library with a standard deviation of 199 bp. The data illustrates the size limit of detection of large inter-read insertions and deletions at each level of clone coverage.

Deletions Identified by Both Intra- and Inter-read Approaches

The 2x50 mate pair technology allows us to push the limits of the size of insertions and deletions that are detectable within a read far beyond what we can do with shorter tag lengths. Using this mate pair data to detect larger insertions and deletions within a single read, we have identified 2,068 deletions ranging in size from 12 to 498 bases of which 40.7% are in dbSNP 129 and 13,604 insertions ranging in size from 4 to 21 of which 63.15% are in dbSNP 129. 193 of the deletions identified within reads are in common with the deletions identified by deviations in the average insert size between reads. Amongst these deletions, 60 of them have also been identified in the Venter, Watson and YH genomes. Figure 4 illustrates a 328 bp deletion that has been identified by both the intra- and inter-read approaches in NA18507 by SOLiD and in each of the other 3 genomes.

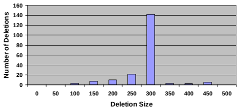


Figure 3. The distribution of the 193 deletions identified in NA18507 with SOLiD by both the intra-read and inter-read approaches.

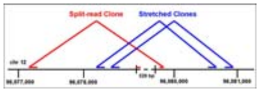


Figure 4. Illustration of a 328 bp deletion detected in NA18507 with SOLiD using both the inter- and intra-read approaches. Four non-redundant molecules identifiy the deletion with the intra-read approach while 81 clones identify the deletion with the inter-read approach. This deletion has also been found in the Venter, Watson and YH genomes.

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

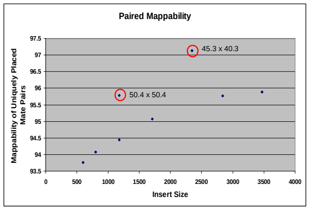


Figure 5. The mappability of uniquely placed mate pairs as a function of insert size and tag length. All data points represent the mappability of uniquely placed paired 25-mers allowing up to 2 mismatches in each tag (25.2 x 25.2) unless otherwise indicated. N's in the reference sequence are discarded in these calculations.

Large Segmental Duplications are Accessible to Sequence and Clone Coverage by Mate Pair Reads

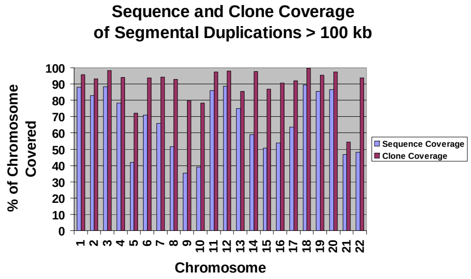


Figure 6. Sequence and clone coverage of segmental duplications > 100 kb in the autosomal human genome by mate pair tags (2x50).

Diversity amongst Human Genomes

We have compared the SNPs and structural variations that we have identified in NA18507 to those that have been found in the Venter, Watson and YH genomes. Figure 7 demonstrates the percent of the variants found in each genome that are unique to that genome as well as the percent of the variants that are found in NA18507 that are in common between all of the genomes. Over 20% of the SNPs in each genome are in each of the other genomes while another 20-40% of the SNPs in each genome are unique to that genome. Less insertions, deletions and inversions are in common amongst the four genomes and a higher proportion of them are unique to the genome in which they are identified. While it is noteworthy to compare this data it must be understood in the context of what is still yet to be uncovered in each genome. The percent of the NA18507 variants that are in common with the other genomes is a lower bound while the number of variants that are unique to each genome is an upper bound. These values will certainly shift as more variation is uncovered in each of the genomes. Since structural variations are typically more difficult to identify than SNPs with current sequencing technologies, it will be exciting to discover whether the tendency for structural variations to be more distinct to single genomes than SNPs will hold as more of the variations in each genome are revealed or whether this is an artifact of the current state of detection.

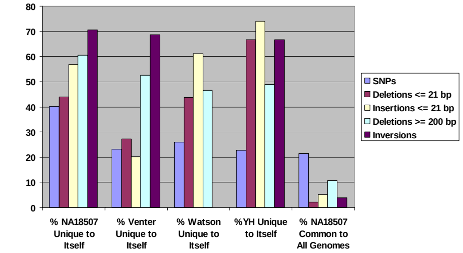


Figure 7. The percentage of identified SNPs and structural variations which are unique to each of the four genomes – NA18507, Venter, Watson and YH – as well as the percentage of the variants identified in NA18507 that are common to all of the genomes. There are no inversions published for the Watson genome so for inversions only the other 3 genomes are considered.

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