Assessing Population Based Differences of Average Total Depth of Coverage in Next Generation Sequencing

ASSESSING POPULATION BASED DIFFERENCES OF AVERAGE TOTAL DEPTH OF COVERAGE IN NEXT GENERATION SEQUENCING

BY:

LATRICE LANDRY

A Dissertation Submitted to the Faculty of Harvard Medical School
In Partial Fulfillment of the Requirements for the Degree of Masters of Medical Science in Biomedical Informatics at Harvard Medical School

Boston, MA

Submitted in the Year Two Thousand and Fifteen

This thesis is presented in conjunction with the committee members:

Dr. Peter Tonellato (Chair), Dr. Sek Won Kong and Dr. Robert Green

In partnership with the Laboratory for Personalized Medicine in the Department of Biomedical Informatics at Harvard Medical School and Supported by the National Library of Medicine's Training Grant with Harvard Medical School, Principal Investigator Dr. Alexa McCray

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

“They say it takes a village to raise a child…I’d say it takes at least that to write a thesis!”

My Village

[The members of the Laboratory for Personalized Medicine- special thanks to Yassine Souilimi and Sheida Nabavi, Dr. Peter Tonellato- my advisor and committee chair, Drs. Sek Won Kong and Robert Green –committee members and mentors on this project, Drs. Alexa McCray and Isaac Kohane –Directors of the Biomedical Informatics Research Training Fellowship, The entire ~~Center~~ Department of Biomedical Informatics –special thanks to Katherine Flannery, Marisa Disarno and Aimee Doe, Deans Jeffrey Flier, Joan Reede and Ellen McCarthy for your support through the Dean’s Post-Doctoral Fellowship, and my family and friends who supported me throughout this process]—

Thank you all, for being a part of my village!

TABLE CAPTIONS

Table 1. Population Specific Average Total Depth of Coverage across all chromosomes from both an African and European population in the 1000 Genomes Phase 1 Pilot Study. Population specific Average Total Depth of Coverage for Yoruban, African (YRI) and Central European (CEU) populations in the Exon Capture and Low Coverage Sequencing projects in the 1000 Genome Project Pilot 1 Study (1KGP). These values are calculated from available files where total depth of coverage was summed across all variants in the population. In this study, we present average total depth of coverage (ADC), which is the TDC divided by the number of individuals in the population. The average values displayed in **table 1** are based on the ADC.

CEU= The Central European population identified in the 1KGP; YRI= The Yoruban Nigerian population identified in the 1KGP. Chr= chromosome; N= Total number of individuals in each population used for analysis; # of Variants = total number of variants in the population that were detected per chromosome; average coverage is based on the total depth of coverage of the variants reported in each population.

Table 2. The association between average total depth of coverage and population group from the exon capture and low-coverage whole genome phase 1 pilot study datasets in the 1000 Genomes Project. The presented comparisons are the result of T-test analysis and the Kolmogorov-Smirnov (Ksa) non-parametric statistical test.

CEU= The Central European population identified in the 1KGP dataset; YRI= The Yoruban Nigerian population identified in the 1KGP dataset. Chr= chromosome; N= sample size; Ksa; # of Variants = total number of variants in the population that were detected at each chromosome; average coverage is based on the mean total depth of coverage from all individuals within the population; [a] Results are statistically significant using a p-value of 0.002 which corrects for multiple comparisons at an alpha level of 0.05 adjusted for 22 autosomes. [b] Results are marginally significant using a p-value of 0.05.

ABSTRACT

Purpose. Next generation sequencing (NGS) is increasingly important to the development and advancement of Precision Medicine. However, there is limited data to support the establishment of a technical validation process sensitive to the complexities of genomes across populations. This validation is hinged upon key metrics used in assessing the quality control (QC) of NGS based tests. Total depth of coverage, a key QC parameter in several NGS analysis steps, is an example of one such metric that has not been assessed for systematic bias across populations.

Objective. To assess differences in average total depth of coverage between an African population and a European population from the 1000 Genomes Project (1KGP) exon capture dataset and the low-coverage whole genome sequencing (WGS) dataset.

Methods. Using previously called variant call format files (VCFs) from the 1KGP, we compared average total depth of sequencing coverage using exon capture, and WGS data in a Yoruban, Nigerian African population (YRI, N=119) and a Central European population (CEU, N=91). Additionally, we compared mean total depth of sequence coverage from a low-coverage WGS dataset (target depth of 2-6x) in the same populations. Comparisons were made using T-tests, and confirmed using Kolmogorov-Smirnov where normal distributions were questioned.

Results. We found a higher average total depth of coverage in the exon capture dataset for CEU when compared to YRI. These data suggest on average there are eighteen more reads capturing a variant in the CEU exomes compared to YRI exomes. The low-coverage data showed no meaningful difference in total depth of coverage between the two groups.

Conclusion and Significance. Given the prominence of NGS technologies in the development of precision medicine, it is imperative to understand key population differences that may affect the ability to detect genomic variation precisely and accurately. The data used in this investigation were taken from publicly available repositories and represent a consensus of different approaches to sequencing and variant calling. Thus it is not clear if these findings represent real differences or an artifact of the different approaches. Artifacts are a potential concern as ‘batch effects’ are a well-known issue for NGS analysis. Additionally, artifacts are of concern as the 1KGP study design includes many different approaches to sequencing and calling variants with a subsequent application of post-hoc filters, which are not consistent between the exon-capture and low-coverage whole genome sequencing projects. It is important to follow-up with additional analyses, where variants are called through a single pipeline with all parameters known and controlled for. Additionally, this is a preliminary step toward the much needed robust testing of NGS in preparation for technical validation and wide-spread clinical use.

Introduction

Clinical sequencing has the potential to revolutionize our understanding of human disease and the practice of modern medicine. Specifically, the use of high throughput technology for sequencing, also referred to as next generation sequencing (NGS) is changing the landscape of genetic research, diagnostic testing, and pharmaceutical development (IOM, 2012; Palmer, 2015). NGS provides fast, relatively inexpensive and accurate genomic information (Metzker, 2010). As a result, NGS technologies are regarded as a key ingredient for precision medicine. Specifically, it is proposed that NGS will dominate genetic research by accelerating research in gene discovery and genetic associations with disease, revolutionizing molecular diagnostics by providing whole genome and exome screening for disease markers and identifying molecular targets for pharmaceutical development (IOM, 2012; Palmer, 2015). The value of NGS also comes from its use as a multi-staged, multi-gene assay, which allows for rapid scaling in number and type of molecular targets. As opposed to the conventional single-gene assay, NGS can provide results for multiple genes or single nucleotide variants (SNVs) as well as information regarding copy number variation (CNV) in addition to the potential for detecting translocations and other structural variation at chromosomal levels (Jia, 2012; Talkowski, 2011; Wang, 2014). Given the attention that NGS is receiving, serious questions have arisen regarding the lack of quality regulation, protocol standardization and clinical validation for NGS sequencing and subsequent processing of variants (Izquierdo, 2011; Mattocks, 2010; Rehm, 2013; Westgard, 2003). Technical validation of clinical laboratory tests includes requirements for consistency and reproducibility as well as accuracy, which have not been fully worked out in the case of NGS. A technical validation process for each potential use of NGS is needed prior to wide-spread clinical implementation (Biesecker, 2010).

The questions raised about wide-scale implementation of NGS for the use of clinical diagnostics are complimentary to the suggested need for 1) enhanced infrastructure around data processing and storage, 2) the development of national, state, local and hospital-based policies on NGS-based diagnostic testing, and 3) the need for standard practices and a quality control (QC) verification process for all NGS protocols. To support this notion of increased testing and standardization together with improved knowledge, the U.S. Food and Drug Administration (FDA) has launched a new "NGS Informatics Community Initiative" (Litwack, 2015). The goal of this initiative is to accelerate the FDA's ability to regulate the use of NGS in diagnostic testing. One key focus of this initiative is the standardization and QC testing of the NGS bioinformatic pipelines for clinical and research-based NGS laboratories.

Typically, an NGS pipeline includes tools for mapping, alignment, quality control, filtering, variant calling and annotation (Altmann, 2012; DePriesto, 2011; GATK, 2014). Each of these steps is important for the analysis and interpretation of NGS data. The scale of whole genome and exome sequencing (WGS and WES) suggests that even small error rates can produce a large number of false positives and negatives (Rehm, 2013). One potential source of error is due to highly variable sequencing coverage (Sims, 2014). NGS sequencing coverage varies across the genome. Accurate variant detection is problematic in regions with low coverage. Insufficient coverage results in limited evidence, defined by the number of reads, which limits the ability to confirm whether a variant detected by sequencing is 'true' or a sequencing artifact (Wheeler, 2008). Additionally, these sequencing errors or 'artifacts' can be propagated through downstream analysis resulting in inaccurate variant interpretations (Sims, 2014). Current practices assign 30x as the *defacto* standard for sufficient coverage. This number is a result of early studies that showed while almost all homozygous SNVs could be detected at 15x, a depth of 30x was necessary to capture all heterozygous SNVs (Bentley, 2008).

Coverage is influenced by many biologic and technical factors. Biologic factors influencing coverage include "GC" content, the presence of structural variants, such as repetitive elements, inserts and deletions, as well as low frequency variation, and CNV events (Bentley, 2008; Cohen, 2015; Curtis, 2015; IHGSC, 2004). Additionally, technical factors such as DNA fragmentation from DNA isolation, sequencing errors and quality of sequence, mapability, the assembly algorithms used and read length also influence coverage (Ekblom, 2014; Sims, 2014; Smith, 2015). Variants without adequate coverage are often filtered out through preset thresholds in variant calling pipelines (Auwera, 2013; GATK, 2014). These filtered variants, some of which may be both 'true' and clinically important, are not included in downstream interpretations. Additionally, in clinical laboratories, variants which 'pass' the coverage threshold for the automated processing can still be filtered out upon manual review for insufficient coverage, which can be based on a different threshold (Sims, 2014). These 'true' variants that are filtered out of the diagnostic process can be problematic. In the case of pathogenic variants, lack of detection could have critical clinical implications in patient care potentially resulting in a missed diagnosis or prognostic indicator or inappropriate therapeutics. The importance of coverage in the use of NGS necessitates research regarding the determinants of coverage and approaches to detecting true variants in the context of low coverage, as well as strategic approaches to enhancing coverage in problematic areas. The first toward this comprehensive understanding is to understand factors, which contribute to insufficient coverage.

One of the known factors contributing to insufficient coverage -- complexity of the sample genome—is defined by the presence of structural variants, repetitive elements, indels and CNVs (Cohen, 2015; Curtis, 2015; IHGSC, 2004). These biological elements affect the quality of sequencing, the alignment of reads, and the mapping of reads to a reference genome. Therefore, complex genomes are logically associated with potentially lower coverage when compared to less complex genomes (Sims, 2014). African genomes are known to be more complex (Witherspoon, 2012). There is more genetic variation, including structural variation in African populations than there is in other populations (Witherspoon, 2007; Thomas, 2014). Therefore it is reasonable to infer that African genomes would likely have lower coverage than European genomes. This population-based difference, if present, is extremely problematic for the future of precision medicine and next generation sequencing. If true, our current approaches to variant calling create systemic bias resulting in a higher frequency of false negative results and subsequently less accurate variant prediction in African genomes. Here we examine the average total depth of coverage using previously called variants in the available variant calling format (VCF) data files for the Central European (CEU) and Yoruban, Nigerian African (YRI) populations of the1000 Genomes Project (1KGP) Phase 1 Pilot study Exome dataset. Additionally, we expanded our analysis to include a replication study of low-coverage genomes sequenced at 2-6x in the same populations from the 1KGP Low-Coverage sequencing dataset.

Methods

Sample Description

The 1000 Genomes Project is an international study on human genetic variation (The 1000 Genomes Project Consortium, 2010). The intended goal of the study was to provide a catalogue of human variation that could be subsequently used for association studies. In the pilot study phase of the project samples were collected from the HapMap collection. Additionally the project included genomic data from volunteers across the globe. Within the pilot phase were three projects 1) the high-coverage sequencing project on two trios, one from a Yoruban, Nigerian (YRI) population and one from a Northern and Central European (CEU) population; 2) the low-coverage sequencing project with a total of 179 individuals from 4 populations, YRI, CEU and CHB (Han Chinese) and JPT (Japanese); and 3) the Exon capture targeted sequencing project of 697 individuals from seven different populations, CEU, YRI, CHB, TSI (Tuscan Italians), LWK (Luhuye Kenyans), CHD (Chinese from Denver) (Durbin, 2010). Of the 697 individuals in the exome sequencing project, 90 are classified as CEU, and 112 classified as YRI. Of those in the low-coverage dataset 36 are classified as CEU and 25 as YRI. In this paper, YRI will be used interchangeably with African and CEU will be used interchangeably with European.

Sequencing and Variant Calling

DNA was extracted from a lymphoblastoid cell line DNA at the Coriell Institute. Sequencing of DNA is described in full elsewhere (The 1000 Genomes Consortium, 2010). In brief, DNA was sheared and prepared for sequencing libraries. The Agilent bioanalyzer 2100 and quantitative PCR were used to assess library insert size and concentrations. As a result of the number of collaborating labs, as well as the need to test the various approaches while achieving the goals of the project, the analysis plan is complex. Each of the labs contributing to the project developed its own approach.

Exon Capture Data Set Preparation

Exon targeting was achieved by hybridization capture with a targeted sequencing depth of 50x. Multiple platforms were used to capture a set of 1000 genes. Only the SNPs captured by all four data producing centers were used in the consensus call set. These genes included 980 randomly chosen genes and 20 ENCODE genes from HapMap 3. The resulting dataset included 8,140 exons from 906 genes. The total target length was 1.43Mb. Both the NimbleGen 385K capture chip and the biotinylated UTP primers from the Agilent microarray capture method were used to capture exon targets for sequencing. Following capture, exons were sequenced using a combination of single end 454 GS FLX/ Titanium machines and the Illumina GA II machines. A detailed description of the exon capture sequencing is provided by the 1000 Genomes Project Consortium (1000 Genomes Project Consortium, 2010).

Variant calling for the exon capture project was completed using the MOSAIK readmapper, as well as the GigaBayes SNP caller using the Boston College's 1000 Genome pipeline. In this pipeline, SNPs were called using the 697 samples from the 7 populations in the exon sequencing project of the 1000 genomes pilot study. Per-population call sets were derived from sites that segregated in that population. Samples were also called through the Broad Institute Pipeline. This pipeline included Mapping Assembly with Quality software (MAQ) and SSAHA2 and the Genome Analysis Tool Kit (GATK) Unified Genotyper. In the Broad pipeline, variants were called within each of the 7 populations. The SNP containing sites that segregated in the 697 samples were identified in the population-specific VCF files. The publicly released VCF includes variants from the consensus of the various exon capture and variant calling methods. The 1000 Genomes report suggests a 73% match between Boston College and Broad Institute pipeline variants for the CEU population and a 52% match for the YRI population.

Low-Coverage Whole Genome Sequencing Dataset Preparation

In the low-coverage project, DNA was sequenced at 2-6x depth using the Illumina platform with 35bp reads supplemented with 51-54bp reads. Sequencing was completed at the Broad Institute, the Michigan Genome Institute and the Wellcome Trust Sanger Institute. Low-coverage genomes underwent extensive processing by three different pipelines at these research centers. Population specific priors were used to identify possible polymorphic sites, which were then evaluated using a method that accounts for likely haplotype sharing patterns. The generation of population specific priors was used in the recalibration of quality scores. The steps for recalibration were 1) map a sample of the reads, 2) select all reads that map with a quality >= 40, 3) counting the number of matches and mismatches at genomic loci not in dbSNP for each raw quality score and ready cycle, 4) provide a posterior Bayesian estimate of true base call quality based on the raw score, and 5) application to the raw values for the entire lane. Additionally, these analyses were conducted separately for each population.

Following population specific variant calling, an extensive quality control and consensus evaluation was used to compile the resulting dataset, which was then stratified by population groups. The 1000 Genomes Consortium reports the consensus calls resulted in 30% fewer errors than the individual call sets. The complete details are described elsewhere (Le, 2009; The 1000 Genomes Consortium, 2010a).

Total Depth of Coverage

The dependent variable of focus in this analysis is the sequencing statistic, total depth of coverage (TDC). TDC at each variant site was calculated using the multicovariance function in the BedTools package. This function counts the alignments from multiple BAM files that have been indexed and sorted according to their position and overlap with BED file intervals. The resulting statistic is a report of a separate count of overlapping alignments from each BAM file at the relative BED interval (Quinlan, 2009-2015). For the low-coverage whole genome dataset, in cases where the BED interval was not available, TDC was calculated for the base immediate to the 5' event (1000Genomes Consortium, 2010b, 2014). The resulting statistic in the VCF file is the total number of reads overlapping that site that were included in analysis (1000 Genomes Consortium, 2010c). In the exon capture dataset TDC represents the reported statistic from the MOSAIK BAM file derived from the Boston College Pipeline. The coverage from the MAQ&SSHA Broad Pipeline is not reported. Additionally, the TDC reported for the low-coverage whole genomes includes all variants called in all populations. The TDC reported for the variant calling process in the exon capture dataset represents polymorphic sites that were called within the populations-specific calls and predicted based on Bayesian models to be likely true. In both exon capture and low-coverage whole genome datasets the reported coverage is a combined depth of coverage across all individuals in the population, representing the total number of reads at a specific variant site within the entire population. Statistics for this analysis are based on the calculated average depth of coverage per individual. As an approximation of average depth of coverage per individual within that population, we divided the reported total depth of coverage by the total number of individuals in the population. In this manuscript the term coverage will refer to average total depth of coverage.

Calculation of TDC:

ADC = TDC/ N

Where ADC is the average total depth of coverage and TDC is the total depth of coverage. In the exome dataset: $N_{YRI} = 25$ and $N_{CEU} = 36$. In the low-coverage whole genome dataset: $N_{YRI} = 119$ and $N_{CEU} = 91$.

Statistical Analysis

We compared mean and variances of TDC between the YRI and CEU populations for both the exon capture and low-coverage genome sequencing projects in Phase 1 of the 1KGP. Using statistical software packages R and SAS, we examined the distribution of depth of coverage using density plots and statistical differences using T-tests and non-parametric analysis of variance (ANOVA). Distributions were examined for normality. In absence of a normal distribution, the Kolmogorov Smirnov non-parametric test was used to assess differences in distributions between the YRI and CEU genome TDCs. Differences in ADC were assessed by chromosome, resulting in 22 comparisons of ADC between YRI and CEU for exomes (50x) and 22 comparisons for low-coverage (4-6x) whole genome. We established significant differences using a p-value cutoff of 0.05 with a Bonferroni correction (0.002) for multiple comparisons.

Results

Overview

In our examination of mean differences in ADC in YRI and CEU genomes using whole exon capture and whole genome sequencing data from the 1KGP, we found discordant results. For low-coverage data we observed higher ADC for YRI genomes in comparison to CEU genomes. In the 1KGP phase 1 exon capture dataset, we observed the opposite, a higher ADC between CEU exomes when compared to YRI exomes. The exon capture data, which focuses on population-specific variants (variants that segregated with a population), shows not only higher ADC for Europeans when compared to Africans, but also demonstrates greater variability around the mean ADC resulting in broader range of coverage in CEU genomes compared to YRI genomes. This may be a result of the difference in number of variants between the two populations. The low-coverage whole genome dataset, which focuses on all variants called within the population showed higher mean coverage for variants called in African genomes when compared to those called in European genomes. Additionally, in both the exon capture and whole genome datasets, we observed more variants in each chromosome in the YRI genomes compared to the CEU genomes -- which may have affected the distributions.

Exon capture

We detected consistent statistical differences between ADC in Africans and Europeans, with T-tests showing higher ADC in European exomes when compared to African exomes. The mean difference in ADC between Europeans and Africans was 18. Suggesting that on average variants in European exomes had 18 more reads than those of African exomes. The ADC ranged from 9.2 additional reads per individual in chromosome 20, to 25 additional reads per individual in chromosome 5 in the European genomes when compared to the African genomes. In all but three chromosomes, these differences remained statistically significant after bonferroni adjustment. These findings were consistent with the Ksa results, which were used as a result of possible differences in distributions between the two populations.

The highest ADC for exon capture in Europeans was 96.5 in chromosome 11, with the lowest ADC being 56.6 in chromosome 20. In Africans, the highest average total depth of coverage for exon capture was 75.2, for chromosome 11, while the lowest was 41.7 in chromosome 22. In the YRI population, there were four chromosomes that did not reach the targeted sequencing depth of 50x, chromosomes 16,19,20 and 22. The targeted sequencing depth of 50x was reached in all twenty-two chromosomes of in the CEU population. Additionally, in the CEU population, sixteen of the twenty-two chromosomes achieved an average total depth of coverage of 70x or higher with eight of these reaching 80x or higher and two with an average coverage of 90x. In the YRI population 70 x coverage was reached for only two chromosomes and there were no chromosomes with an average total depth of coverage of 80x or greater.

Low-coverage whole genomes

We tested the low-coverage genome dataset and determined that the TDC were normally distributed. Our findings suggest consistent differences in total depth of coverage between Africans and Europeans in all called variants in the 1KGP low-coverage dataset. These genome-wide differences were detected in each of 22 chromosomes, with higher coverage in African genomes when compared to European low-coverage genomes. These results remained statistically significant after application of the Bonferroni correction (Table 2). The mean difference ranged from 0.02 in chromosome 4 to 1.14 in chromosome 4, resulting in zero to one additional reads in African genomes compared with European genomes. Chromosomes fifteen through twenty-two had a mean difference in average depth of coverage between .5 and 1.5, resulting in variants detected in African genomes having one more read than those detected in European genomes. Additionally, we observed a trend towards higher mean difference in coverage in the larger chromosomes compared to the smaller chromosomes.

The average total depth of coverage for variants captured in European genomes was 7.8, while the average total depth of coverage for African genomes was 8.2. The lowest average total depth of coverage was 7.3 on chromosome 19 for Europeans and 7.8 on chromosome 4 for African low-coverage genomes. The average total depth of coverage for both populations exceeded the targeted coverage of 4-6x with the maximum coverage achieved in African genomes being 8.7 and 8.0 in European genomes. There were no clear differences in coverage based on chromosome size in the YRI or the CEU for the low-coverage whole genomes.

Number of Variants

As the unit of measure on which coverage is based, we also provide a short description of the number of variants in chromosomes between the two populations for each dataset. A paucity of variants in any one chromosome could be suggestive of insufficient sample size for statistical analysis. In our study this was not the case. However, we did observe that African genomes consistently contained more variants than did their European counterparts. This is consistent with the literature that suggests more variation in the African genome when compared to other population groups (Jorde, 2004). In our study, we found the total number of variants detected using exon capture in European exomes ranged from 23 in chromosome 21 to 344 in chromosome 1. For Europeans, both chromosomes 13 and 22 had less than 100 variants detected in European exomes (59 and 67 respectively). For African exomes, the number of variants detected ranged from 42 in chromosome 21 to 505 in chromosome 1. Only chromosomes 21 and 13 had fewer than 100 variants detected in Africans (there were 64 variants detected in chromosome 13). In chromosome 19, known to have high gene density, 210 variants were detected in Europeans, and 310 variants in Africans. In all chromosomes the number of variants detected in African exomes was greater than that of European exomes.

The number of variants detected in the low-coverage dataset ranged from 101,568 in chromosome 22 to 3,325,487 in chromosome 1 for Europeans and 137,859 in chromosome 22 to 6,935,828 in chromosome 1 for Africans genomes. We detected the greatest difference in variants detected between African and European low-coverage genomes in chromosome 17, with 196,327 variants detected in Europeans and 815,928 variants detected in Africans.

Discussion

In our assessment of population-based differences in average depth of coverage for next generation sequencing using the 1KGP exome dataset and the 1KGP low-Coverage dataset, we found ***meaningful*** significant differences in ADC ONLY for population specific variants in African and European exomes sequenced in the exon capture project. The data from the exon capture project show on average variants that segregated in the CEU (European) population had 18 more reads than variants that segregated in the YRI (African) population. These differences were consistent across all chromosomes and ranged from an additional 9 to 25 average reads in European genomes.

Given the importance of total depth of coverage as a quality control metric in the processing and filtering of genetic variants these results warrant further attention. Systematic differences in total depth of coverage between population groups would result in less evidence and subsequently less confidence in variant calls and their interpretation for African compared to European patients. Less confidence in variant interpretation could result in desperate care between these two groups. Lower coverage and thus lower quality of variant calls in African genomes could also result in 'true' variants being miscalled (type II errors) in the typical NGS analysis and variant calling pipeline. For example, the Genome Analysis Toolkit (GATK) pipeline uses coverage as a parameter in calculating variant quality (Auwera, 2013; GATK, 2014). Consequently, variants with low coverage have a greater likelihood of falling below the variant quality threshold, and therefore higher likelihood of rejection. The low coverage African-bias phenomenon could therefore potentially result in a higher false negative rate in African genomes and thus translate into a decrease ability to both 1) identify and 2) interpret variants in African genomes (Sims, 2014). If true, the complexity of African genomes as demonstrated by generally lower ADP can therefore result in health disparity if not accounted for in NGS and it's application in precision medicine.

Typically in clinical laboratories exomes and exome-based targeted panels are sequenced at 150x or more (ACMG, 2013; Aziz, 2014). It is not clear as to whether or not the differences in coverage in our study would be observed at such depth. The data presented here, suggests on average a 9-25 read deficit in ADC in the YRI African population when sequencing with a targeted depth of 50x. These differences calculated, translate to a 20-50% reduction in average coverage for variants in YRI genomes. If sequencing at 'clinical' coverage of 150x, a 50% loss may be less meaningful. However, even at 150x a 50% average reduction in coverage could be problematic in some cases. In solid Tumors, for example, NGS analysis typically results from DNA from paraffin embedded tissue samples as opposed to DNA from blood. Parrafin is a DNA denaturant, which causes DNA degradation (Aziz, 2014). This adds an additional obstacle to achieving high quality sequences for PCR amplification. Furthermore, DNA extraction from tumor specimens typically includes DNA from both constitutional DNA as well as somatic DNA. This results in sequencing reads for both constitutional and somatic DNA. In many cancer laboratories, the goal is to identify somatic mutations that may provide decision support for therapeutics and diagnostics. In a case where the tumor percentage is 30% of the DNA extracted, the percent of total reads that are somatic would be 30% of the total read depth at any variant. A potential 50 % reduction in coverage for African genomes on top of the loss of coverage due to tumor heterogeneity could result in an unpalatable disparity in variant calling and interpretation of NGS-based molecular diagnostics for African genomes.

This study did not address potential causes of disparate coverage. As previously mentioned complex genomic architecture contributes to insufficient sequencing (Sims, 2014). Complexity, in the context of genomic architecture can be explained by the presence of repetitive elements, nucleotide inserts and deletions, increased GC content, hypermutation and hyper-variability, as well as a greater presence of homologous sequences (Cohen, 2015; Curtis, 2015; IHGSC, 2004). There are many examples of structural variation, but essentially these variants cause issues for DNA 1) sequencing and PCR amplification, and 2) informatic processing which includes alignment with other reads in formation of a contig, mapping to the reference genome and variant calling (Landers, 1988; DePristo, 2011). Genomes of African ancestry are known to have more structural variants (Jorde, 2004). Therefore, the increased presence of each of these types of structural variation could cause a problem for anyone of these steps in NGS analysis. To determine if discrepancy in coverage is associated with structural variation in the 1KGP exome dataset additional studies would be required.

Another potential focus for future research would be to address the differences between exon capture and low-coverage whole genome sequencing. Our analysis of differences in ADC of low- coverage whole genome sequencing between YRI and CEU populations showed different results than those found for exon capture. Although we found statistically significant differences in ADC between African and European whole genomes sequenced at 2-6x, these data suggest that Africans have approximately one additional read when compared to Europeans. As a result of the number of variants in the low-coverage whole genome sequencing dataset -- in the millions for chromosome 1-- we had the power to detect extremely small differences in coverage between the two populations. However, given the size of the differences, the functional and clinical relevance is questionable. In addition to the size of the difference, the direction of the difference was opposite that of the exon capture. Again, in our comparison of low-coverage whole genome sequencing and exon capture of YRI and CEU genomes, the mean difference was less than one read. Although, there are cases where one read may make a difference especially in the case of low-coverage sequences, the significance of this difference is not clear. Sequencing with a targeted coverage depth of 2-6x would not provide the needed level of evidence for variant assessment in a clinical laboratory.

If lower coverage in African genomes is 'real', but only in the case of exon capture, a robust investigation into the causes for this divergence between WGS and WES by exon capture is prudent. Of interest would be, whether differences in coverage between YRI and CEU genomes are mitigated by the variation in intronic sequences. Intronic regions of DNA contain many homologous regions (Curtis, 2014; 1000 Genomes Consortium, 2010). Homologous sequences create challenges in alignment and mapping of reads. Therefore, it is possible that the sequencing of intronic regions is equally challenging in YRI and CEU genomes. Additionally, coding regions of DNA, exons, account for only 2% of DNA. Therefore, it is also possible that the inclusion of the other 98% of DNA provides sufficient noise, that signal detection - *in this case the difference in coverage between YRI and CEU genomes* - could not be detected. The non-coding regions of the genome contain alternative splicing locations, which can result in hyper-mutable regions, as well as long stretches of repetitive elements (Li, 2010; Jarrow, 2010). If the presence of these structural variants in non-coding regions in statistically greater in African genomes as compared to European genomes, and these differences are greater in non-coding regions as opposed to coding regions, this may explain this difference. Also, it is possible that the difference observed in coverage is an artifact of the sequencing itself. 'Batch effects' are a well- documented problem in NGS sequencing and analysis. It is possible that the African samples were batched separately from the European samples and that coverage is a reflection of a random 'batch affect' (Robasky, 2014).

These findings represent a first step in the assessment of differences in average total depth of coverage between African and European genomes. As such, there are many limitations that should be addressed in future studies. Our study was based on publicly available variant call format files from 1KGP. As such, we could not control the lack of uniformity and consistency of the data processing steps. The 1KGP VCFs represent the concordance between BAM files that were called at different sites with different pipelines. As a result of the need for concordance between two different pipelines, there are potentially variants that were filtered out of the QC process that may have been 'true' variants. Although study investigators suggested the consensus calls resulted in 30% reduction in error (1000 Genome Project Consortium), it would be important to assess the coverage of these variants to see if they are somehow different than those called. There is limited information regarding the specific parameters in the algorithms used in the pipelines for 1KGP. Therefore reproducing the VCF datasets would be challenging. Also, the exon capture dataset focuses on variants that segregated to YRI or CEU populations, whereas the low-coverage dataset includes ALL variants detected in each population.

As a result of the limitations of secondary data analysis, follow-up studies should focus on calling the publicly available BAM files available for the YRI and CEU populations. Additionally, if follow-up studies show similar findings, it is not clear if these differences are generalizable to other African and European populations, to African-Americans or to other admixed populations, or even the other populations included in the 1KGP dataset. Given the availability of these data in the 1KGP repository it is important to include additional populations in future analysis. It is not clear if the differences observed in average depth of coverage between YRI and CEU populations are 'true', or an artifact of the processing undergone for 1KGP. However, it is clear that future research addressing the aforementioned topics is necessary. With the Presidential Precision Medicine Initiative announced in January of 2015, precision medicine has gained great momentum. Complementing the initiative, many academic medical centers have announced their own initiatives around precision medicine (DFCI, 2015; Shaw, 2015). With this increasing momentum, it is proposed that precision medicine will advance healthcare in America. However, before wide-scale implementation can be advised it is important that we thoroughly understand the tools that are propelling this momentum and ensure that these tools and precision medicine are enabling us to advance healthcare for all people.