Targeted Sequencing of the Human X Chromosome Exome
Kajari Mondal, Emory University
Amol Carl Shetty, Emory University
Viren Patel, Emory University
David J Cutler, Emory University
Michael Zwick, Emory University

Journal Title: Genomics
Volume: Volume 98, Number 4
Publisher: Elsevier: 12 months | 2011-10, Pages 260-265
Type of Work: Article | Post-print: After Peer Review
Publisher DOI: 10.1016/j.ygeno.2011.04.004
Permanent URL: http://pid.emory.edu/ark:/25593/f0gxm

Final published version: http://dx.doi.org/10.1016/j.ygeno.2011.04.004

Copyright information:
© 2011 Elsevier Inc. All rights reserved.
This is an Open Access work distributed under the terms of the Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License (http://creativecommons.org/licenses/by-nc-sa/3.0/).

Accessed December 25, 2019 5:39 AM EST
Targeted Sequencing of the Human X Chromosome Exome

Kajari Mondal1, Amol Carl Shetty1, Viren Patel1, David J Cutler, and Michael E. Zwick1

1Department of Human Genetics, Emory University School of Medicine, Atlanta, GA, USA

Abstract

We used a RainDance Technologies (RDT) expanded content library to enrich the human X chromosome exome (2.5 Mb) from 26 male samples followed by Illumina sequencing. Our multiplex primer library covered 98.05% of the human X chromosome exome in a single tube with 11,845 different PCR amplicons. Illumina sequencing of 24 male samples showed coverage for 97% of the targeted sequences. Sequence from 2 HapMap samples confirmed missing data rates of 2–3% at sites successfully typed by the HapMap project, with an accuracy of at least ~99.5% as compared to reported HapMap genotypes. Our demonstration that a RDT expanded content library can efficiently enrich and enable the routine sequencing of the human X chromosome exome suggests a wide variety of potential research and clinical applications for this platform.

1. Introduction

Recent technological advances have dramatically reduced costs while increasing the throughput of DNA sequencing (see reviews in [1–3]). These changes in DNA sequencing have stimulated research focused on improving methods of target DNA isolation from complex eukaryotic genomes [4–15]. One promising approach of DNA target selection builds upon advances in microdroplet-based technology that is currently being developed by RainDance Technologies (RDT) [16]. This methodology used a highly controlled system to create emulsions containing reactive compounds that allow an enormous number of defined reactions to be carried out in parallel. An initial application of this technology demonstrated the ability to perform in parallel greater than 1 million simplex PCR reactions targeting up to 4,000 distinct amplicons in order to generate target DNA for a second generation sequencing platform [11]. A recent study demonstrated the use of a library of 10,280 distinct amplicons to target 437 target genes that when mutated cause severe childhood recessive diseases [17].

While these original experiments were very successful, sequencing the entire coding content of human chromosomes may be of interest for both research and clinical applications and requires the validation of even larger RDT libraries. For example, the human X chromosome is replete with loci that when mutated can result in intellectual disability (ID) [18,19]. Autism spectrum disorders (ASDs) display a profound male excess, much like ID, the causes of which remain largely undiscovered, although there is growing evidence for a role of X-linked variation in at least some cases of ASD [20–22]. While recent resequencing studies of the human X chromosome in patients with ID or ASD have resulted in a number...
of exciting findings, their reliance on traditional PCR-based methods of target DNA amplification were limiting [21,23,24]. Mutations at the X-linked dystrophin locus, the largest gene in the human genome, result in Becker and Duchenne muscular dystrophies. Identifying the varying classes of functional mutations, at dystrophin or other clinically significant X-linked loci, remains a significant challenge for clinical diagnostic testing [25–27].

We have therefore focused on determining if RDT’s microdroplet-based technology could be used to efficiently sequence the human X chromosome exome. A project targeting the human X chromosome exome (~2.5 Mb) would require approximately a 12,000 distinct PCR amplicons. To address these larger target regions, RainDance Technologies developed expanded content libraries which allow up to 5 primer pairs within each primer droplet and maintain single plex PCR by limiting the amount of genomic DNA within the PCR reaction, with no other changes in the standard workflow. Here we validate a custom-designed RDT expanded content library with 3 sets of primers in each droplet in order to perform the efficient targeted resequencing of the human X chromosome exome. By combining the RDT expanded content library with the Illumina Genome Analyzer sequencing platform, our results show that it is possible to sequence multiple human X chromosome exomes with high uniformity and accuracy, while at the same time minimizing sequencing costs. This approach will likely prove invaluable for researchers and clinicians in a wide variety of applications that require the systematic sequencing of all the coding and non-coding exons of the human X chromosome.

2. Materials and Methods

2.1 Primer library design process

We obtained the target sequence for the human X chromosome exome from the UCSC genome browser RefSeq Genes track (hg18 build). The total reference sequence consisted of 7,427 fragments with a total size of 2,495,062 bases and included all coding and non-coding (3’ and 5’ untranslated regions) exons of human chromosome X. To enable greater than 4,000 unique primer pairs within a primer library the RainDance Primer Design Pipeline was modified to evaluate each primer pair to determine the ability to pool up to 5 primer pairs within the same primer droplet. The number of amplicons offered in an expanded content RDT primer library ranges from 4,000 to 20,000 unique primer pairs. Since the human X chromosome exome design required less than 12,000 primers, we chose to optimize 3 primer pairs per droplet. The custom primer library was designed using the manufacturer’s design parameters (RainDance Technologies, Lexington, MA, USA) and the Primer3 algorithm (http://primer3.sourceforge.net/). All SNPs from dbSNP build 129 were filtered from the primer selection region. Repeat masking was not performed on the input regions to the primer design pipeline. The primer design pipeline performed an exhaustive primer selection across all of the regions submitted.

After the primers were selected, duplicate amplicons and their associated primers were removed from the full design and only unique regions were kept in the collapsed design. A total of 11,845 unique amplicons were required to cover the entire targeted human X chromosome exome (Table 1). Filtering of these amplicons led us to reject 27 primer pairs whose design parameters were too extreme to meet the stringent primer picking criteria used by RDT. An additional 242 amplicons and their associated primer pairs were deleted from the design because they were predicted to produce more than 3 products in the human genome.

Primer pairs were pooled based on the proximity of each primer pair’s target within the genome. Each of the primer’s are evaluated for off target products using the re-PCR
algorithm for each of the 6 primer-primer interactions within each pool. Once pools of 3 primer pairs were determined, each pool was processed within the RainDance Primer Library Manufacturing process. Then each pooled primer pair aliquot was processed to create an emulsion containing an equal representation of each of the 3 primer pairs within a unique primer droplet.

The final human X chromosome exome RDT expanded content primer library consisted of 11,576 amplicons and associated primer pairs, or 97.7% of the initial design (Table 1). The full length of DNA amplified was predicted to be 5,916,297 bases. In total, 98.05% of the targeted human X chromosome exome bases (2,446,304 out of 2,495,062) were covered by at least one amplicon (Table 1). After accounting for overlapping amplicons, a reference sequence consisting of 5,748 fragments with a total size of 4,723,733 bp was generated and used for mapping.

2.2 Sample selection
We chose 24 male samples from SFARI’s Simplex Collection, New York, NY, USA. The SFARI Simplex Collection (SSC) is a core project and resource of the Simons Foundation Autism Research Initiative (SFARI). SSC has a permanent repository of genetic samples from approximately 3000 families, each of which has one child affected with Autism Spectrum Disorder (ASD), one unaffected child, and two parents unaffected with ASD. Two male HapMap samples, NA18500 and NA18503, were also enriched and sequenced in order to compare HapMap genotype calls with those from our Illumina sequencing. Prior to processing, patient and control DNA samples were quantified by measuring OD260/280 using a NanoDrop instrument. Following quantification, 100 ng of DNA, as determined by the NanoDrop, were run on a 0.8% agarose gel to verify that the DNA was of high molecular weight. A total of 26 genomic DNA samples (2 HapMap males as control and 24 Autistic patients DNA) passing quality control were sent to RDT facility and were then processed on the RDT 1000 with the RDT Sequence Enrichment Application using standard RDT procedures for genomic DNA.

2.3 Genomic DNA Fragmentation
Genomic DNA samples were fragmented using a nebulization kit (Invitrogen, Carlsbad, CA, USA, catalogue # K7025-05) following the manufacturer’s recommended protocol: 2.5 µg of genomic DNA was re-suspended in 750 µL Shearing Buffer (TE, pH 8.0, Fisher, Worcester MA, USA, catalogue # 50843207) containing 10% glycerol (Fisher, catalogue # AC15892) and was nebulized at 6 – 10 pounds per square inch (psi) for 90 seconds to produce 2–4 kb DNA fragments. Fragmentation of the genomic DNA to 2–4 kb allows for optimal template size for performing PCR in droplets. Sheared genomic DNA was precipitated by adding 80 µL 3 M sodium acetate, pH 5.2 (Fisher, catalogue # 50843081), 4 µL 20 mg/ml Mussel Glycogen (Fisher, catalogue # NC9329100) and 700 µL 100% isopropanol (Fisher, catalogue # AC14932) mixed and stored overnight at −20°C. The samples were centrifuged at the maximum speed for 15 minutes at 4°C. The supernatant was removed, 500 µL of cold 80% ethanol (Fisher, catalogue # 5739852) wash buffer was added and the DNA pellet was spun down by centrifugation at the maximum speed for 5 minutes at 4°C. The pellet was air dried and re-suspended in 10 µL 10 mM Tris-HCL, pH 8.0 (Sigma, St. Louis, MO, USA, catalogue # T2694). Fragmented genomic DNA was run on a 0.8% agarose gel to confirm that the genomic DNA was in the correct size range (2 – 4 kb).

2.4 Genomic DNA Template Mix
In order to prepare the input DNA template mixture for targeted amplification, 1.0 µg of the purified Genomic DNA Fragmentation reaction was added to 4.7 µL 10× High-Fidelity Buffer (Invitrogen, catalogue # 11304-029), 1.26 µL of MgSO4 (Invitrogen, catalogue #
11304-029), 1.71 µL 10 mM dNTP (New England Biolabs (NEB), Ipswich, MA, USA, catalogue # NO447S/L), 3.6 µL Betaine (Sigma, catalogue # B2629-50G), 3.6 µL of RDT Droplet Stabilizer (RainDance Technologies, Lexington, MA, USA, catalogue # 30-00826), 1.8 µL dimethyl sulfoxide (Sigma, catalogue # D8418-50ml) and 0.72 µL 5 units/µL of Platinum High-Fidelity Taq (Invitrogen, catalogue # 11304-029) the samples was brought to a final volume of 25 µL with Nuclease Free Water, Teknova (Fisher, catalogue # 50843418).

2.5 RDT 1000: Merge

PCR droplets were generated on the RDT1000 (RainDance Technologies, catalogue # 20-01000) using the manufacturer’s recommended protocol: To process a single sample the user placed onto the RDT1000 a single tube containing 25 µL of Genomic DNA Template Mix, a custom primer droplet library (RainDance Technologies) and a disposable microfluidic chip (RainDance Technologies). The custom primer droplet library consists of a collection of individual primer droplets where each primer droplet contains matched pairs of forward and reverse primer (5.2 µM per primer) for each amplicon that is in the primer library. The final primer concentration in the PCR reaction is 0.53 uM per primer. The RDT1000 generated each PCR droplet by pairing a single gDNA template droplet with a single primer droplet. The paired droplets flow past an electrode embedded in the chip and is instantly merged together. All of the resulting PCR droplets were automatically dispensed as an emulsion into a PCR tube and transferred to a standard thermal cycler for PCR amplification. Each single sample generated more than 1,000,000 single plex PCR droplets. After PCR Amplification the emulsion of PCR droplets were broken to release each individual amplicon from the PCR droplets and were purified over a Qiagen MinElute column. The purified PCR product was then run on the Agilent Bioanalyzer Bioanalyzer to confirm that the amplicon profile matched the predicted histogram profile (Supplemental Figure 1).

2.6 Multiplex Illumina library preparation

After PCR purification, amplified fragments for each individual were repaired to blunt ends using NEB Quick blunting kit (NEB, catalogue # E1201L, 15 minutes RT) followed by inactivating the enzyme in the blunting reaction by heating at 70°C for 10 minutes. The PCR fragments were then concatenated using NEB Quick ligation kit (NEB, catalogue # M2200L). Ligation was done overnight at 25°C. After that 5 µl of Quick T4 DNA ligase was added to the reaction was incubated at 37°C for one hour followed by inactivating the ligase at 65°C for 15 minutes. The ligated products were made into 100 µl volume by adding elution buffer and were then sheared using Covaris E210 (Duty cycle 10%, Intensity cycle 5, Cycle/Burst: 200, Time :180 sec). The sheared fragments were then purified using Qiagen QIAquick PCR purification column and was eluted in 32 µl of elution buffer. The samples then entered the standard Illumina Genome Analyzer multiplex library introduced preparation protocol. At the enrichment step, a 6 base index tag was attached to each sample using PCR following the standard Illumina protocol. Only exception is, while purifying the adaptor ligated products we have used Invitrogen E-Gel SizeSelect 2%(Invitrogen, catalogue # G6610-02) instead of using the gel purification method suggested by Illumina. The enrichment was confirmed by running a Agilent BioAnalyzer 7500 DNA chip. A quantitative qPCR was done to quantitate the library using KAPA Library quantification kit (KAPA Biosystems, Woburn, MA, USA, catalogue # KK4824).

2.7 Illumina sequencing and data analysis

Enriched DNA was denatured and diluted to a concentration of 8pM. Cluster generation and 70bp single end sequencing was performed using standard IGAI manual and version 4 kits. We performed multiplex single-end sequencing of three samples per lane of Illumina
sequencing. After sequencing, the reads were mapped and variants sites identified using EmoryMapper (Cutler and Zwick, pers. comm.) against the reference sequence consisting of 5748 fragments covering 4,723,773 bases. This region is larger than the actual targeted bases (2,446,304) as RDT included some intronic and intergenic regions to facilitate primer picking. Sequences obtained for HapMap samples NA18500 and NA18503 base calls were compared to those reported by HapMap using a custom perl script to assess the rates of data completion and accuracy. The HapMap data was assumed to be without error when estimating data accuracy.

2.8 Microarray-based Genomic Selection (MGS)

We performed MGS for the HapMap sample NA18503 twice using the methods published previously [6,10,15]. After MGS the enriched DNA samples were each sequenced in a single lane of an Illumina IGA IIX (76bp, non-multiplexed, single end). The sequence obtained for the two replicates of sample NA18503 were each compared to those reported by HapMap using a custom perl script to assess the rates of data completion and accuracy [28,29]. The HapMap data was assumed to be without error when estimating data accuracy.

3. Results

3.1 Characteristics of human X chromosome exome RDT expanded content multiplex library design

We obtained the human X chromosome exome sequence from the RefSeq Genes track using the UCSC genome browser (hg 18 build). The total region we targeted for enrichment and sequencing included the coding exons and the 3’ and 5’ untranslated regions for all annotated X chromosome genes with a total length of approximately 2.5 Mb. A RDT library, designed to enrich the entire targeted region, consisted of 11,845 different PCR amplicons. Bioinformatic filtering of these amplicons eliminated a small percentage from the final multiplex primer library (as described in the Materials and Methods). The final synthesized multiplex primer library covered 98.05% of the targeted region and included 97.7% of the total designed amplicons (Table 1).

3.2 Illumina sequencing of RDT expanded content multiplex library enriched samples

We next determined the optimal level of sample multiplexing that minimizes sequencing costs while maintaining adequate coverage of the human X chromosome exome. To do so, we enriched 4 male SSC samples using our multiplex human X chromosome exome RDT library. Illumina libraries were then constructed using a different multiplex adapter for each sample. The resulting sample libraries were then sequenced (70 basepair, single-end reads) on an Illumina Genome Analyzer IIX in 1-plex, 2-plex, 3-plex and 4-plex configurations. For all configurations, between 53% and 64% of the total reads mapped uniquely to the human X chromosome exome reference sequence (Table 2). As expected, increasing the number of samples per lane sequenced reduced the median depth across the targeted region, while variability in the number of reads per lane correlated with the observed median depth. Finally, approximately 3% of targeted bases had zero coverage across samples irrespective of the level of multiplexing (Figure 1). From our data, we concluded that the 3-plex configuration was the optimal level of sample multiplexing that minimized our sequencing costs while maintaining adequate coverage of the human X chromosome exome.

We next assessed the repeatability of the RDT enrichment and Illumina sequencing. To do so, we enriched and sequenced 18 additional male samples in a 3-plex configuration and 2 samples in a 2-plex configuration. One sample failed amplification because of insufficient genomic DNA. However, the remaining 19 samples performed as well, or better, than our original set of four test samples described previously (Supplemental Table 1). In total, we
observed that 99.57% (11,526 of 11,576 total) of the amplicons were successfully amplified and sequenced in at least one sample among the 24 samples analyzed. While the types of failures were very consistent among all samples processed, we sought to determine if amplification or sequencing failure was correlated with a specific sequence composition. The sequence composition of those amplicons that failed to be sequenced in any sample (0.43%, 50 of 11,576 total amplicons) was not significantly different from the successful amplicons (t-test, $p = 0.96$, Figure 2).

### 3.3 Data completion and accuracy

In order to evaluate the quality of data obtained, we evaluated the rates of data completion and accuracy as compared to reported genotypes for 2 HapMap samples [28,29]. We evaluated the accuracy of variant bases called in 2 male HapMap samples (NA18500 and NA18503) by comparing the HapMap genotypes with our called genotypes, assuming the HapMap call was always accurate. In both of these samples only 2–3% of the data was missing, at sites able to be called by HapMap, and the accuracy (agreement with HapMap calls) was greater than 99% (Table 3). We next compared the performance of RDT-based enrichment method with that of MGS [6,10,15]. The results shown in Table 3 demonstrate that the MGS performance was very similar in terms of data accuracy and missing data. However, the MGS capture method required between 3 to 7 fold more reads to achieve this similar level of performance as compared to RDT-based enrichment.

### 4. Discussion

Choosing the best enrichment method for targeted sequencing in research or clinical applications requires consideration of a variety of factors. These include: the ease and reliability of the assay, the extent of enrichment of targeted sequences, the uniformity of coverage along the targeted sequence, the accuracy of identifying variant genotypes, and of course, the cost effectiveness of the approach. With the continuing rapid drop in the cost of DNA sequencing, the ability to efficiently enrich targeted regions of eukaryotic genomes is likely to contribute to novel research and clinical applications.

Our results show that a RDT expanded content library can be used to effectively sequence the human X chromosome exome. Using our original RDT library design, sequencing two HapMap samples shows between 2 – 3% missing data at known segregating sites with accuracies of at least 99.5%. While we obtain similar levels of coverage and accuracy by performing Microarray-based Genomic Selection (MGS), we observed that MGS required between 3 to 7 fold more sequence per sample to achieve the same level of data accuracy and completeness as compared to the RDT methodology. Applying the RDT approach to a larger collection of 24 male samples from the SSC showed that the method was reliable and rapid. On the other hand, while the MSG protocol is somewhat slower than the RDT methodology, its main advantage lies in the ability to rapidly custom design a microarray for a small number of samples [15]. In contrast, the RDT approach is most efficient when the number of samples to be processed with a defined assay is large.

In our initial multiplex sequencing experiments, between 53% and 64% of the total reads mapped uniquely to the human X chromosome exome reference sequence. In our later experiments, up to 73% of reads mapped uniquely to the human X chromosome exome (Supplemental Table 1). One potential question that might arise is why do so many reads fail to map to the reference sequence? The explanation for this pattern is straightforward. After PCR amplification of the fragments, the standard RainDance Technologies protocol randomly concatenates the resulting products into longer fragments of DNA. These longer fragments are then physically sheared, end-repaired, and subjected to library construction in preparation for next-generation sequencing. Any resulting read that spans the concatenation
boundary will not map uniquely to the human X chromosome exome reference sequence. An additional analysis pipeline step that splits these reads and maps each portion to its respective unique location would likely be able to increase the percentage of reads mapped. However, for this analysis, we chose to use existing tools for our analysis, recognizing that we may fail to map reads spanning these concatenation boundaries.

Our demonstration that a RDT expanded content library can efficiently enrich and enable the routine sequencing of the human X chromosome exome suggests a wide variety of potential research and clinical applications of this platform. For example, routine sequencing the human X chromosome exome in human disorders with a male excess could help provide insight into the etiology of X-linked genetic disorders. Furthermore, while we have focused on the coding and non-coding exons of the X chromosomes, the size of the amplified targets could be easily expanded to include other regions of interest, such as conserved non-coding X-linked regions or panels of genes from across the genome.

Finally, as the routine use of DNA sequencing for clinical diagnostics becomes increasingly widespread, the RDT platform offers the potential to provide rapid and reliable assays, allowing a clinic to routinely amplify a large, but standardized suite of regions known to harbor disease variants of interest. One strength of the RDT platform as compared to a capture based method, MGS, is that it requires dramatically less sequencing with comparable accuracy rates and missing data (Table 3). This outcome presumably reflects the lower variance in coverage among different fragments of the RDT methodology as compared to MGS. We have previously shown that better models for selecting oligos can significantly improve the performance of capture based methods [15], and iterative selection combined with empirical validation can allow investigators to rebalance capture oligos in order to reduce the variance in capture efficiency still further. In support of this, Bell et al. 2011 observed that RDT and a different capture based method performed similarly, although their level of recurrent primer synthesis failures for RDT (~11%) was far higher than what we observed in our experiment (0.43%, 50 of 11,576 total amplicons). In total, our data suggest that for new assays, RDT may require less initial optimization as compared to a capture based approach and may require less sequencing for comparable accuracy and data completeness. Finally, a second advantage of a targeted enrichment methodology like RDT is that it can be applied for the routine targeted sequencing of large numbers of individual patients, but by standardizing the designs, the bioinformatic challenges faced by the diagnostic laboratory can be dramatically minimized, relative to whole human genome sequencing.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We would like to thank Ms. Brenda Billote, Mr. Ephrem Chin, Dr. Madhuri Hegde for their assistance in processing our research samples on the RDT 1000 platform in the Emory Genetics Laboratory. We would also like to thank Dr. Timothy Read, Mr. Chad Haase and the other members of the Emory-Georgia Research Alliance Genomic Center for their assistance in the Illumina sequencing of our enriched research samples. Finally, we would like to thank Mr. Jim Brayer and colleagues at RainDance Technologies for their assistance with the RDT expanded content libraries. This his work was supported in part by the National Institutes of Health/National Institute of Mental Health and Gift Fund [grant number MH076439] to MEZ, the Simons Foundation Autism Research Initiative [MEZ], and the PHD Grant (UL1 RR025008, KL2 RR025009 or TL1 RR025010) from the Clinical and Translational Science Award program, National Institutes of Health, National Center for Research Resources.
References


Figure 1.
Bar graph representing percentage of uniquely mapped targeted bases in four different SSC sample configurations (1-plex, 2-plex, 3-plex, 4-plex) versus the observed sequence depth.
Figure 2.
(A) UCSC genome browser view of the location of failed amplicons along chromosome X.
(B) Bar graph showing proportion of failed and successful amplicons versus the GC content of the amplicons.
Table 1
Design summary for human chromosome X exome RDT Library

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Total size of target region</td>
<td>2,495,062 bp</td>
</tr>
<tr>
<td>Size of region covered by RDT library</td>
<td>2,446,304 bp</td>
</tr>
<tr>
<td>Percent of region covered by RDT library</td>
<td>98.05%</td>
</tr>
<tr>
<td>Total number of amplicons to cover target region</td>
<td>11,845</td>
</tr>
<tr>
<td>Final number of amplicons in RDT library</td>
<td>11,576</td>
</tr>
<tr>
<td>Percent of amplicons in RDT library</td>
<td>97.7%</td>
</tr>
</tbody>
</table>
Table 2

Results of targeted sequencing of human chromosome X exome

<table>
<thead>
<tr>
<th>Sample ID</th>
<th>Plex</th>
<th>Total Number of Reads</th>
<th>Percent of Reads that Map Uniquely</th>
<th>Median Sequence Depth</th>
<th>Average Sequence Depth</th>
<th>Percent of Bases With at Least One Read</th>
</tr>
</thead>
<tbody>
<tr>
<td>SSC1</td>
<td>1</td>
<td>22,540,925</td>
<td>53.24</td>
<td>129</td>
<td>175.5</td>
<td>97.87</td>
</tr>
<tr>
<td>SSC3</td>
<td>2</td>
<td>15,060,253</td>
<td>64.37</td>
<td>100</td>
<td>142.5</td>
<td>97.52</td>
</tr>
<tr>
<td>SSC4</td>
<td>2</td>
<td>11,199,304</td>
<td>61.75</td>
<td>77</td>
<td>101.5</td>
<td>97.67</td>
</tr>
<tr>
<td>SSC2</td>
<td>3</td>
<td>5,926,188</td>
<td>53.14</td>
<td>32</td>
<td>46.1</td>
<td>96.76</td>
</tr>
<tr>
<td>SSC3</td>
<td>3</td>
<td>11,488,977</td>
<td>64.29</td>
<td>77</td>
<td>108.6</td>
<td>97.49</td>
</tr>
<tr>
<td>SSC4</td>
<td>3</td>
<td>8,410,820</td>
<td>61.64</td>
<td>58</td>
<td>76.0</td>
<td>97.59</td>
</tr>
<tr>
<td>SSC1</td>
<td>4</td>
<td>4,021,807</td>
<td>53.52</td>
<td>23</td>
<td>31.5</td>
<td>97.13</td>
</tr>
<tr>
<td>SSC2</td>
<td>4</td>
<td>4,696,801</td>
<td>52.71</td>
<td>25</td>
<td>36.3</td>
<td>96.75</td>
</tr>
<tr>
<td>SSC3</td>
<td>4</td>
<td>8,953,172</td>
<td>64.17</td>
<td>60</td>
<td>84.5</td>
<td>97.49</td>
</tr>
<tr>
<td>SSC4</td>
<td>4</td>
<td>6,008,050</td>
<td>61.48</td>
<td>41</td>
<td>54.2</td>
<td>97.56</td>
</tr>
</tbody>
</table>
Table 3

Results of comparing of two methods of target enrichment (RDT, MGS) followed by Illumina sequencing with HapMap reported genotype calls

<table>
<thead>
<tr>
<th>HapMap ID</th>
<th>Method of Enrichment</th>
<th>Number of Reads Mapping</th>
<th>Data Accuracy</th>
<th>Missing Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>NA18500</td>
<td>RDT</td>
<td>2,734,713</td>
<td>99.53%</td>
<td>2.79%</td>
</tr>
<tr>
<td>NA18503</td>
<td>RDT</td>
<td>1,966,363</td>
<td>99.44%</td>
<td>2.34%</td>
</tr>
<tr>
<td>NA18503</td>
<td>MGS</td>
<td>9,349,327</td>
<td>99.64%</td>
<td>3.92%</td>
</tr>
<tr>
<td>NA18503</td>
<td>MGS</td>
<td>13,995,254</td>
<td>99.43%</td>
<td>3.37%</td>
</tr>
</tbody>
</table>