

# Working with Sequence Capture Data in NextGENe® Software

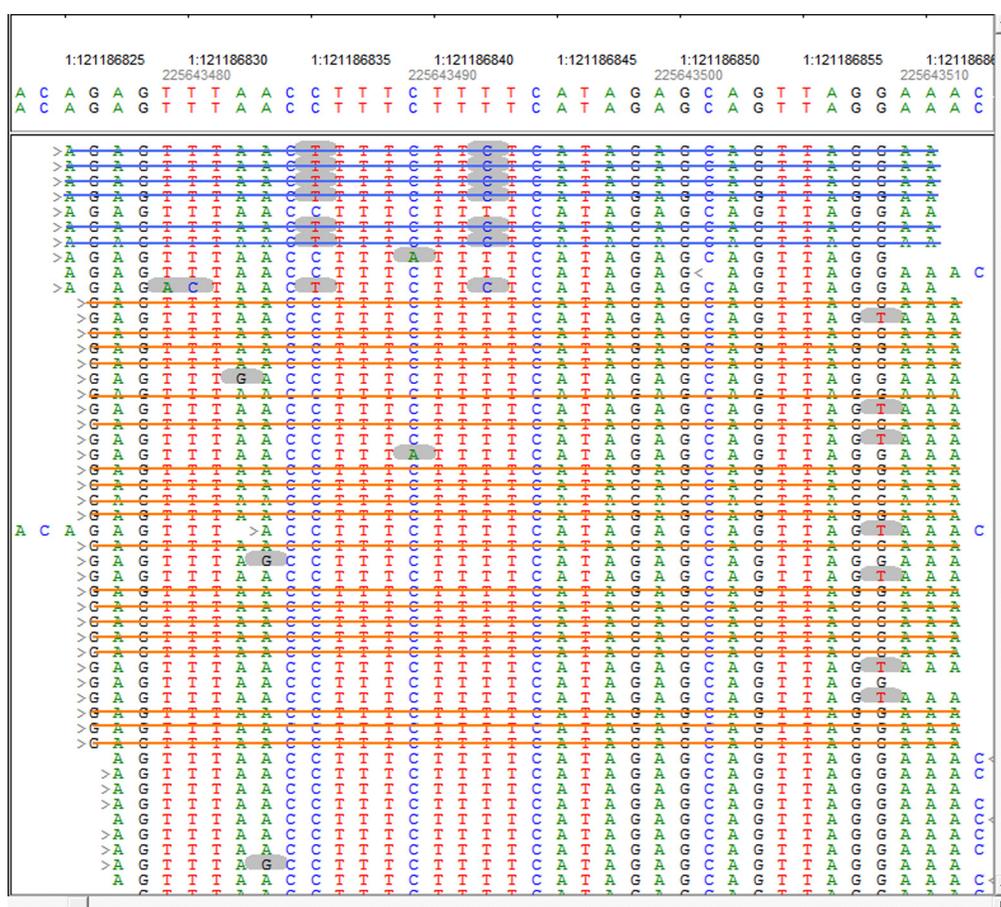
May 2010

John McGuigan, Megan Manion, Kevin LeVan, CS Jonathan Liu

## Introduction

Due to the size and complexity of the human genome, several issues arise when whole genome sequencing is performed. The short reads generated by next generation sequencing systems are susceptible to misalignment because of repetitive sequences. A massive amount of sequencing data needs to be generated in order to obtain sufficient coverage at regions of interest, requiring a significant amount of time and money. Targeted enrichment and sequencing of specific genomic regions is used to limit the amount of data that needs to be generated and processed while still obtaining a high depth of coverage for regions of interest. This allows for increased allele calling accuracy and thus increased accuracy in SNP and structural variation detection.

There are currently several commercially available capture systems. NimbleGen Sequence Capture technology from Roche NimbleGen, Inc. [1,3] and The HybSelect™ system from FeBit Biomed, GmbH [4] both use microarrays. The SureSelect Target Enrichment System from Agilent Technologies [2] uses “hybrid selection” that combines microarray RNA oligomer synthesis and bead-based capture in solution. RainDance’s Sequence Enrichment application uses microdroplet PCR amplification of targeted regions [5].



**Figure 1:** A portion of chromosome 1 from the alignment of the original data. Two groups of duplicate reads have been highlighted. Nearly all of the reads align in the same direction as indicated by “>” or “<” placed at the ends of the reads.

While these capture methods use different techniques, they all differ from whole-genome sequence data in similar ways. All four techniques employ PCR amplification, which can result in inaccurate mutation percentages. Reads with normal alleles that generate more copies can cause false negatives (the mutation percentage for the mutant allele in other reads is too low) and errors in the reads may cause false positives. Some capture methods use PCR a second time which can introduce duplicates of PCR errors as well. An example of these duplicate reads can be seen in figure 1. The reads often align in one direction instead of in both forward and reverse directions (see figure 1). It is important to only analyze data from the captured regions. None of these systems have perfect specificity and any sequence information aligned outside of the regions of interest may be misaligned or inadvertently captured sequence.

## Methods

In this analysis publicly available data from the 1000 Genomes Project Pilot 3 was downloaded from the NCBI Sequence Read Archive (<http://www.ncbi.nlm.nih.gov/sra>). The sample (SRR014820) is hybrid capture of targeted human exons followed by Illumina® sequencing. The data was aligned to the whole human genome and reports were created for the regions of interest (ROI).

## Procedure

1. Use the Format Conversion Tool to convert the fastq files to fasta format while filtering the data based on quality.
2. Use the Condensation Tool to remove duplicate reads.
  - A. Open the “Advanced Settings” and select “One index/read”.
  - B. Reads that match perfectly will be combined into one read
    - i. Reads must have the same number of bases with no mismatches- a read that had one additional base on the end is not considered a duplicate.
3. Align the reads to the whole-genome reference.
4. Generate an Expression Report for the ROI and use it to calculate specificity
  - A. In the NextGENe Viewer open “Expression Report” from the “Reports” menu
  - B. Select “Input Region of Interest”, load a .bed file (figure 2), and click Ok
  - C. Sum the number of reads from each segment (shown in the “Read Counts” column) in order to get the total number of reads aligned in the ROI
  - D. Divide this number by the total number of aligned reads in order to find the specificity.

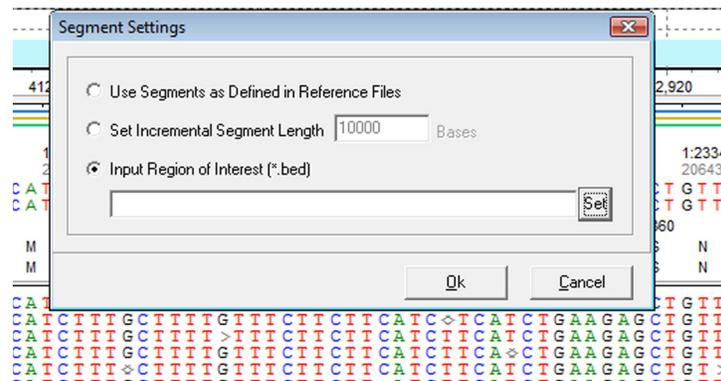


Figure 2: Specifying the Region of Interest for the Expression Report

5. Generate a Mutation Report for the ROI
  - A. Adjust the mutation report settings as desired
    - i. Open the mutation report settings with the  button
  - B. Select “Output the Points of Interest” from the “Mutation Report” menu
  - C. Select a properly formatted .bed or .txt file and a save directory (figure 3)
    - i. Selecting “Filter by the Mutation Report Settings” will apply any filters you set up in the mutation report settings window.

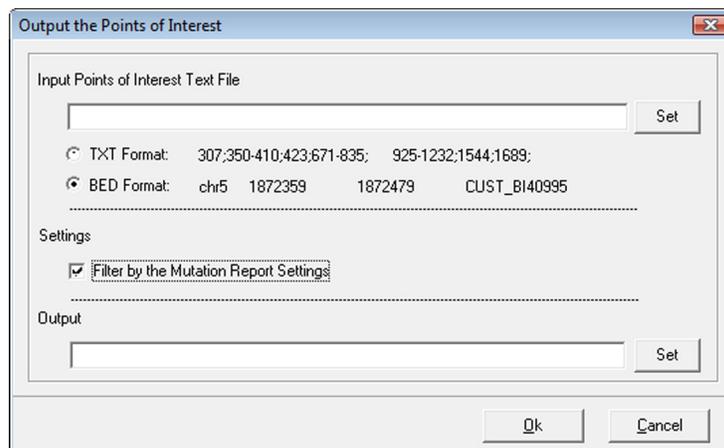


Figure 3: Specifying the Region of Interest for the Mutation Report

# Results

	Alignment of the Original Data	Alignment of the Data after Removing Duplicate Reads
Total Reads	34,390,420	34,390,420
Format Conversion Reads Removed	8,947,228	8,947,228
Duplicates Removed	-	14,285,998 (58.4%)
Remaining Reads	25,443,192	10,175,404*
Matched Reads	22,791,862	8,548,726
Matched %	89.6%	84.0%
Reads in ROI	4,086,237	1,331,408
% Reads in ROI	17.93%	15.57%
Average # of Reads	481.0	156.7
ROI with 0 Reads	114 (1.34%)	122 (1.44%)
ROI with <1/2 Avg # of Reads	4498 (52.94%)	6140 (72.27%)

**Table 1:** Alignment Results summary

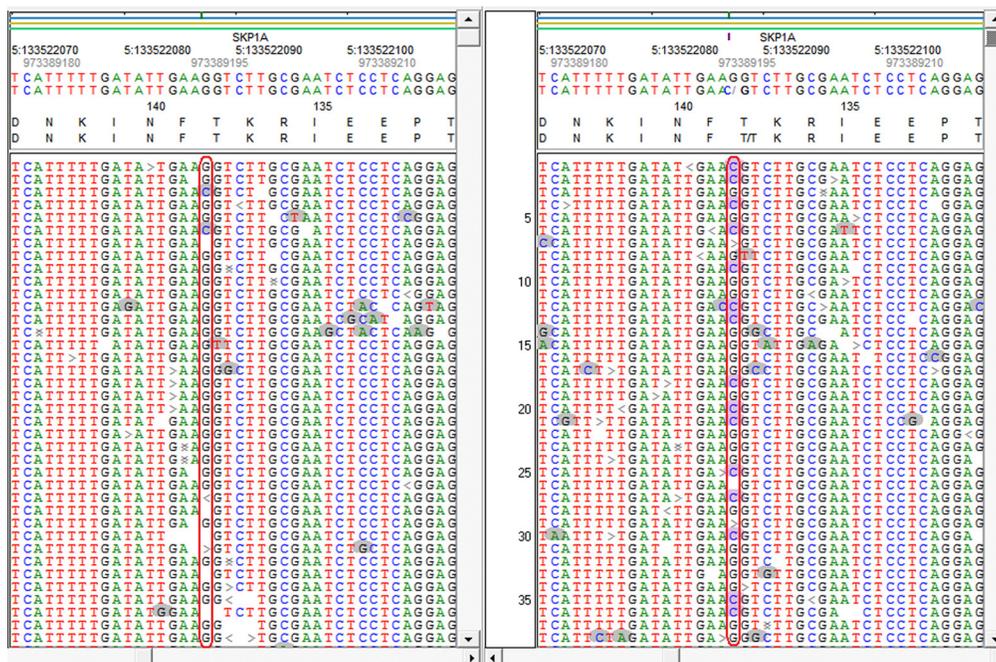
\*981,790 reads were not used in the duplicate read removal step and are mostly repetitive sequences

	Alignment of the Original Data	Alignment of the Data after Removing Duplicate Reads
Called Mutations	100,320	37,548
Called Mutations in ROI	1,047	1,108
% in ROI	1.04%	2.95%

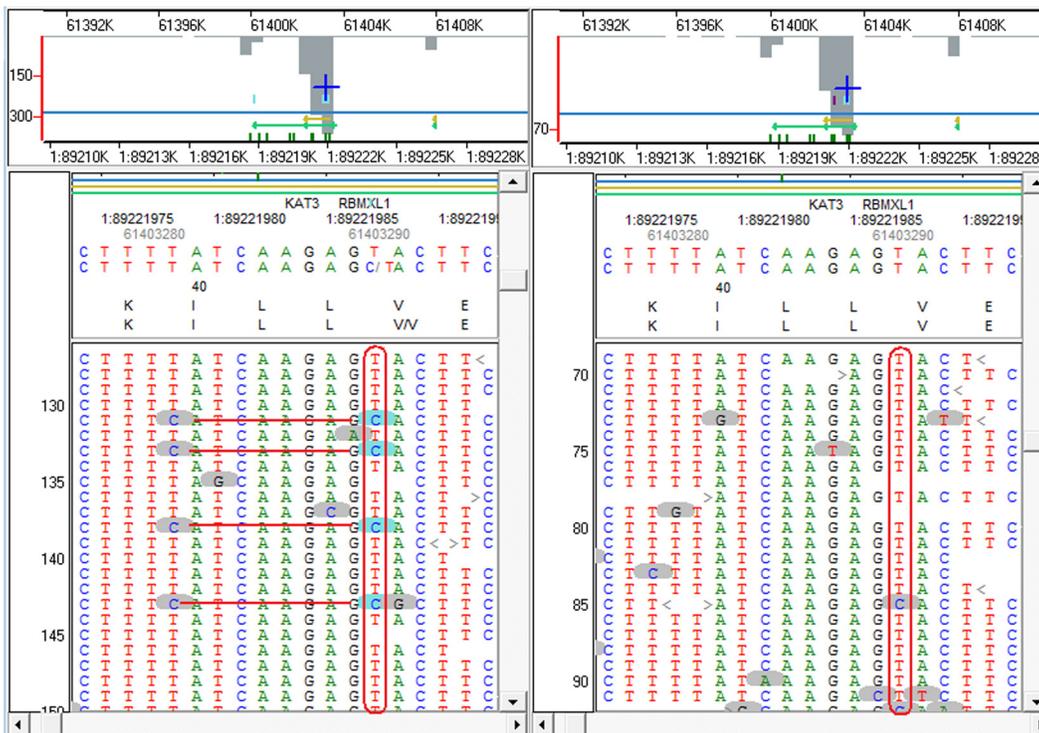
**Table 2:** Number of called mutations in each alignment and in the ROI

Most mutations outside the ROI have low coverage (less than 10x) and a mutation score less than 5, indicating that they are almost all false positives.

At the position shown in figure 4 the original alignment showed a G>C mutation in 56 out of 362 reads (15.47%) while 39 out of 103 reads (37.86%) showed the SNP after removing duplicate reads. Because the mutation filter was set to 20%, the SNP was not detected in the original alignment. Figure 5 shows a position where a mutation was called in the original alignment and not called after removing duplicate reads. Every read with the SNP either had a mismatch nine bases upstream or did not cover that position. It is possible that these reads were captured from a different area of the genome with a similar sequence. Most reads aligned at this position were duplicates- after removing them the coverage was reduced from 538 to 99 and the SNP was found less often (11.11% vs 24.16%). Removing duplicate reads will always lower the overall coverage, but it will also remove much of the potential PCR bias.



**Figure 4:** (chr5, pos. 133,522,087) - a SNP was detected after duplicate reads were removed (right) but not when the original reads were aligned (left)



**Figure 5:** (chr1, pos. 89,221,987) – A SNP was called in the original alignment but not after duplicate reads were removed. The SNP appears to be associated with a mismatch 9 bases upstream. The presence of multiple mismatches strongly associated with one another indicates potential misalignment.

## Discussion

NextGENE has several tools for dealing with the problems unique to capture data. The “One index/read” setting in the condensation tool can be used to remove duplicate reads. This reduces problems that can arise from the use of PCR. For example, a sample may contain a heterozygous mutant allele. Due to different efficiencies of PCR, one allele may be amplified more than the other resulting in an incorrect mutation percentage. The mutation will not be called if the normal allele is amplified significantly more. If a mutation occurs only in one read after removing duplicates then it isn’t likely to be real.

NextGENE is able to use a bed file to filter many of the reports that it produces. This is especially important for the mutation report- any mutations called outside of the captured region may be false positives. As seen in table 2, removing duplicate reads resulted in a significant reduction in the number of mutations called outside the ROI. The bed file input for the expression report can be used to find the number of reads aligned to each captured region. The specificity can be calculated using this information and used to determine if there was a problem with the sequence capture process. It is important to align to the entire genome and filter the results because aligning to a reference containing only the captured regions may force reads to map incorrectly. A read may match perfectly outside the ROI because it was derived from that position but it might also map within the ROI with some mismatches.

NextGENE is a powerful tool for variant analysis using sequence data produced by massively parallel genome analyzers. It has several tools that are useful for processing capture data. NextGENE is a versatile software package designed for the new era of high through-put genome sequencing, supporting the Illumina® Genome Analyzer, the Applied Biosystems SOLiD™ System and the Genome Sequencer FLX System from Roche Applied Science. Similar software packages are available such as GS Reference Mapper Software from Roche Diagnostics Corp and Genomics Workbench from CLC bio. Some of NextGENE’s advantages include accurate results for indel and SNP detection with whole genome alignment and an easy-to-use graphical user interface.

NextGENE includes software applications for a variety of application types including expression studies like Digital Gene Expression, transcriptome analysis, microRNA studies and SAGE, as well as de novo assembly, SNP and indel detection, and ChIP-Seq.

- ## References
1. Albert, T.J. et al. Direct selection of human genomic loci by microarray hybridization. *Nat. Methods* 4, 903-905(2007).
  2. Gnirke, A. et al. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat Biotech* 27, 182-189(2009).
  3. Okou, D.T. et al. Microarray-based genomic selection for high-throughput resequencing. *Nat Meth* 4, 907-909(2007).
  4. Summerer, D. et al. Microarray-based multicycle-enrichment of genomic subsets for targeted next-generation sequencing. *Genome Res* 19, 1616-1621(2009).
  5. Tewhey, R. et al. Microdroplet-based PCR enrichment for large-scale targeted sequencing. *Nat Biotech* 27, 1025-1031(2009).

Trademarks are Property of their Respective Owners.