Analytical pipeline for next-generation exome sequencing data

Fischer M[1], Stocker G[1], Trajanoski Z[1,2]

[1] Institute for Genomics and Bioinformatics, Technical University, Graz, Austria
[2] Biocenter, Division of Bioinformatics, Medical University, Innsbruck, Austria

Bioinformatics Integration Network

Introduction

Exome sequencing is a promising new approach to find mutations which spur and maintain human diseases [1, 2]. The thereby generated amount of raw sequence data demands for well designed bioinformatics tools able to handle these large datasets in an efficient way. Until now many algorithms have emerged, each of them addressing a different task in the downstream analysis of next-generation sequencing (NGS) data. We have combined these algorithms into an analysis pipeline which greatly facilitates the identification of potential "disease driver" mutations as it connects all necessary analysis steps, manages the ensuing additional data handling, and distributes computationally expensive jobs on a HPC cluster.

Methods

Exome Capturing and Exome Sequencing

Exome capturing was done with NibmleGen's Sequence Capture 2.1M Human Exome Array which is capable of capturing about 180.000 human coding exons defined by the Consensus Coding Sequence project (CCDS) [3]. The resulting sequences were then sequenced by the Illumina GS II platform generating reads at a length of 50 bp and different fold coverages.

Exome Sequencing Analysis Pipeline

The pipeline is designed to integrate several in-house developed as well as open source analysis tools (see list 1) into one single pipeline while handling all required task management on a cluster. The pipeline itself is implemented in Java using the internally developed JClusterService for communication with the cluster.
The analysis pipeline performs quality statistics, filtering and trimming of sequence reads, and aligns them to a reference genome. Post alignment analysis includes the calculation of alignment statistics, region filtering, and the detection of variants resulting in a list of potential "disease driver" candidates (see Figure 1).

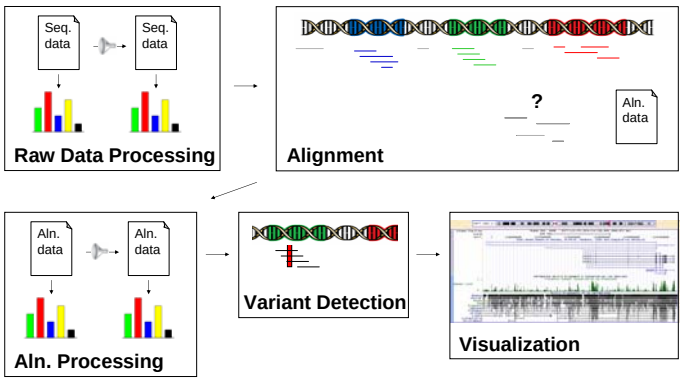


Figure 1: Exome Sequencing Pipeline *schematic overview of the exome analysis pipeline workflow.*

Figure 2 illustrates the software architecture which was introduced to provide an efficient system to analyze and visualize the exome sequencing data. After the pipeline ensured that the user is authorized to access the cluster, all required data is transferred via the JClusterService. A cluster-queuing system (in our case Sun Grid Engine) internally handles task scheduling and resource management between different analysis jobs.

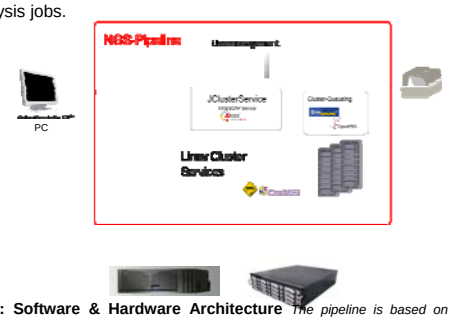


Figure 2: Software & Hardware Architecture *The pipeline is based on the Java 2 Enterprise Edition (J2EE) three-tier architecture and uses JClusterService to communicate with the HPC cluster. The currently used hardware architecture consists of two X4600 Sun Servers (32 CPUs, 160 GB RAM) attached to SAS storage with 16 TB (and extendable to max. 256 TB) via GBit Ethernet interconnect.*

	In-house development	Open Source
Pipeline Organization	Java, JClusterService	
Raw Data Processing	R, Perl, C++	
Alignment		BWA [4]
Alignment Processing	Java	SAMtools [5], Picard
Variant Detection		SAMtools [5]
Visualization		SAMtools [5], SeqMonk

List 1: Analysis software *incorporated into the pipeline listed by analysis tasks.*

Results

Pipeline Analysis

The pipeline was used to analyze three preliminary test samples with two technical replicas each. ~ 4 M, 4.5 M and 1.8 M reads were sequenced for sample 1, 2, and 3 respectively. After filtering reads containing > 2 unidentified nucleotides 3.4 M and 2.8 M (sample 1), 3.7 M and 3.1 M (sample 2), and 1.5M and 1.2 M (sample 3) reads remained for alignment.
Quality value (QV) statistics performed on raw and filtered sequence data illustrate the positive effects of filtering error prone reads (see Figure 3).

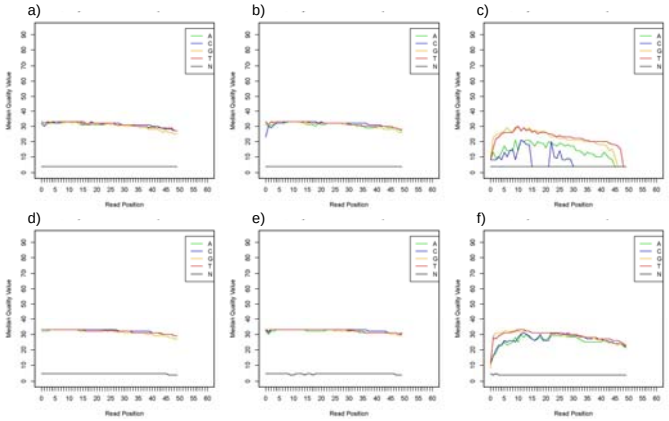
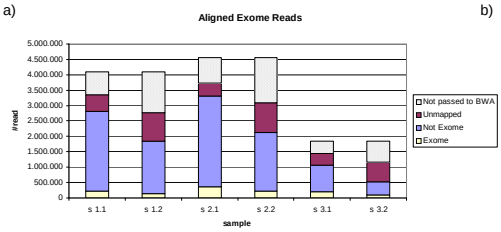


Figure 3: Median QV of Base Calls per Read Position *before (a, b, c) and after (d, e, f) application of N filter (N < 3) for sample 1.2 (a, d), sample 2.2 (b, e), and sample 3.2 (c, f). A slight decrease in QVs in higher read positions can be observed outlining the sequencer's small decay in sensitivity over base calling cycles. It is expected to see clear differences between QVs of unidentified and identified calls (as seen in all graphs except c)). Before filtering, Sample 3 shows very poor median quality values especially for Cytosine calls. QVs drastically improve after the elimination of error-prone reads.*

2.8 M and 1.8 M, 3.3 M and 2.1 M, and 1 M and 0.5 M reads from sample 1, 2, and 3 could be mapped to the human reference sequence NCBI Build 36.1 using the alignment program BWA [5]. From these 215 k and 136 k, 354 k and 222 k, and 194 k and 91 k reads respectively origin from exon regions. Between 2 k and 16 k SNPs and indels were detected in exons (see Figure 4 a and b).

a)



b)

Sample	# SNP	# Indel
s 1.1	15.542	924
s 1.2	10.357	503
s 2.1	15.024	1.038
s 2.2	10.596	535
s 3.1	4.257	444
s 3.2	2.241	128

c)

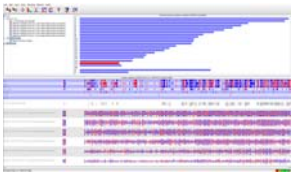


Figure 4: Final Pipeline Results

a) Total number of sequence reads for sample 1, 2, and 3. Banded, dark-red, violet, and yellow areas illustrate the amount of reads not passed to alignment, unmapped, not exome and exome reads, respectively.
b) Final amount of detected SNPs and Indels in exon regions for each sample.
c) Overall visualization of mapped reads by SeqMonk.

Conclusions

We have developed a Java based pipeline capable of performing the computationally expensive downstream analysis of exome sequencing data. By applying the pipeline to preliminary test data we could proof that the pipeline is able to evaluate the quality of sequencing runs and prepare reads for alignment to a reference genome. Additional filtering of mapped reads results in the set of reads originating from exon regions. These reads are further analyzed to detect potential "disease driver" variants which then need to be verified by additional experiments.

References

1. Shendure *et al.* Targeted capture and massively parallel sequencing of 12 human exomes. Nature. 2009 Sep 10; 461(7261):272-6.
2. Maher. Exome sequencing takes centre stage in cancer profiling. Nature. 2009 May 14; 459(7244):146-7.
3. Lipman *et al.* The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes. Genome Res. 2009 Jul;19(7):1316-23.
4. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009 Jul 15; 25(14):1754-60.
5. Li *et al.* The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009 Aug 15;25(16):2078-9.

Acknowledgements

This work was supported by the Austrian Ministry of Science and Research GEN-AU project "Bioinformatics Integration Network".