

Working with the Variant Comparison Tool in NextGENe® Software

May 2010

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

Introduction

To facilitate review of multiple variant-discovery projects NextGENe includes a Variant Comparison Tool. Up to ten projects that utilized the same reference can be compared at one time. The software automatically identifies and color codes variants- green if a sample does not have the variant, blue if it does, and purple if it is a reported variant. Several filtering options are available to simplify the review process.

Methodology

A dataset (SRP001569) was downloaded from the NCBI Sequence Read Archive (<http://www.ncbi.nlm.nih.gov/sra>). This dataset contains seven samples from the *E. coli* Long-term Experimental Evolution Project representing successive generations of *E. coli* bacteria (2k, 5k, 10k, 15k, 20k, 30k, and 40k). The samples were sequenced on an Illumina® GA instrument producing millions of 36 bp reads. The fastq files were converted to fasta and quality filtered using NextGENe's format conversion tool. NextGENe's quality filtering parameters were set so that the reads had to have a median score of 20, no more than three uncalled bases, and no less than 24 called bases. If three or more bases with a score less than or equal to 16 were found at the 3' end of the reads they were trimmed off. On average 7,446,146 out of 7,581,970 reads were successfully converted (98.2%).

Each fasta file was aligned separately to the genome of *E. coli* strain B/REL606, which was derived from a variant of the bacteria used in the long-term experimental evolution experiment (*E. coli* Bc251). At least 15 bases or 70% of each read was required to match, and mutations were called if they appeared in at least 60% of reads where there was at least 3 coverage. After alignment all seven samples were loaded in the variant comparison tool (Figure 1). On average each project contained 7.3 million aligned reads (98.5% of quality filtered reads on average).

The settings menu gives access to many options for customizing the report for the experiment. The first option was used for this analysis but there are 3 main viewing options:

1. Show common and/or unique variations in all samples.
2. Show variants where at least one sample meets a minimum coverage threshold and two variants differ by a minimum amount.
3. Show only those variants with coverage below a certain threshold

The rest of the settings are the same as those found in NextGENe's mutation report and include settings for score filters, hiding synonymous mutations, and hiding reported variations. Variants are filtered out if both samples do not meet the requirement, such as a minimum mutation score. The report can be saved as a tab-delimited text file.

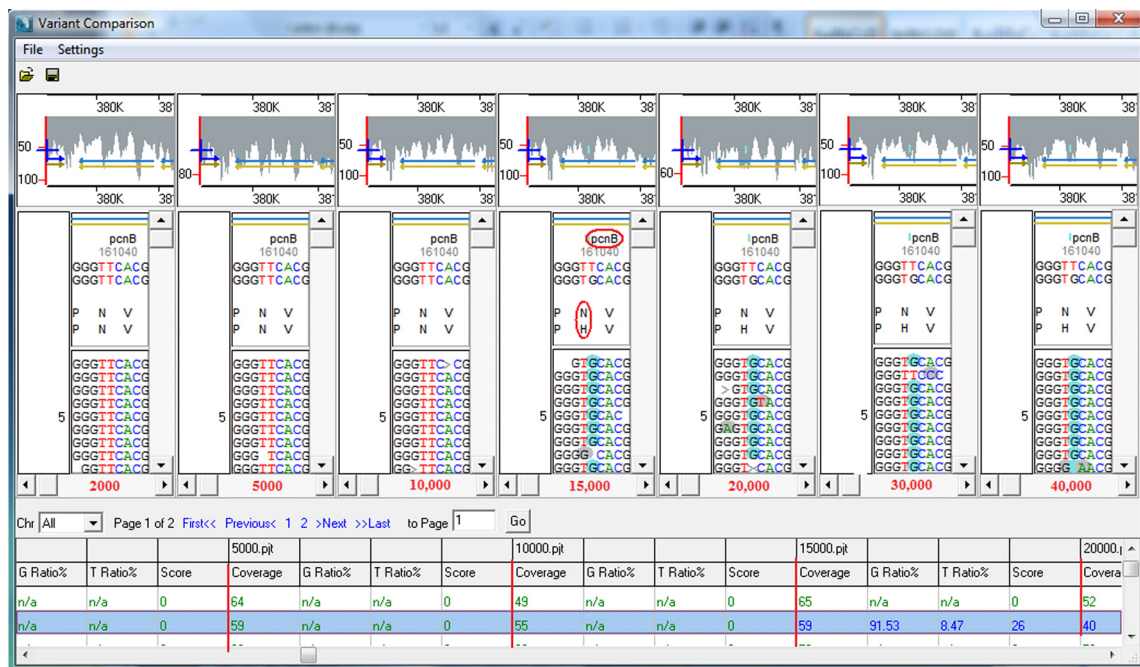



Figure 1: It is easy to see that a SNP has arisen between the 10,000th (sample 3) and 15,000th (sample 4) generations. This mutation occurred in the pcnB gene and appears to have changed an Asparagine to a Histidine.

Procedure

1. Filter and Trim reads and convert to fasta format using the Format Conversion Tool
2. Align the processed reads to the reference
3. Load all of the projects into the Variant Comparison Tool
 - A. In the NextGENe Viewer select “Variant Comparison” from the “tools” menu.
 - B. Click on the folder icon  and load .pjt files for each alignment with the “Add” button
 - C. Click “Ok”
4. Customize the report using the settings panel which can be opened from the toolbar (figure 2).
5. Save the report as a tab-delimited text file.

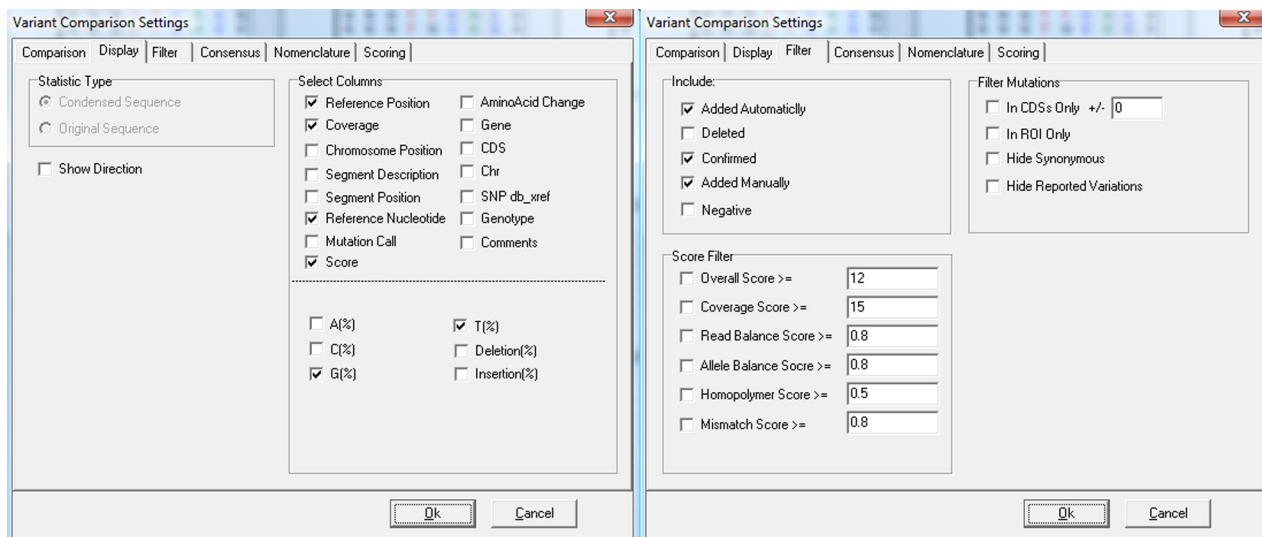


Figure 2: Some of the display and filtering settings available in the Variant Comparison tool.

Results

The variant comparison tool makes it easy to visualize in which generation a mutation first appeared. The mutations are summarized in table 1.

Generation	New Mutations (Score >= 12)	Mutations causing an amino acid change
2000	7	4
5000	7	7
10000	1	1
15000	17	13
20000	5	4
30000	268	234
40000	100	84

Table 1: New mutations are those found in each sample that were not found in any of the previous generations.

Discussion

NextGENe’s Variant Comparison Tool makes it quick and easy to compare variants detected in different samples. In this experiment samples from successive generations of bacteria were sequenced in order to monitor the appearance of new mutations. Other samples from the *E. coli* Long-term Experimental Evolution Project were found to have evolved the ability to metabolize citrate. It is hypothesized that mutations must have occurred in separate generations for this to occur [1]. By analyzing many more samples it would be possible to find the potentiating mutation(s) that occurred before 20,000 generations and the citrate-using mutation(s) that occurred after 30,000 generations.

There are many other experiments where this tool can be valuable including time-sampling studies and case-control studies. Large sequencing projects may have thousands of variations between the control sample and the reference. Variations shared between a sample and the control can be hidden in order to filter them out and focus on the potentially important variations.

NextGENE is a powerful tool for variant analysis using sequence data produced by massively parallel genome analyzers. With the Variant Comparison tool several samples can be rapidly and accurately compared in order to find important mutations.

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. Blount, Z.D., Borland, C.Z. & Lenski, R.E. Historical contingency and the evolution of a key innovation in an experimental population of *Escherichia coli*. *Proceedings of the National Academy of Sciences* 105, 7899-7906 (2008).

Trademarks are Property of their Respective Owners.