

Analysis of Publicly Available Expressed Sequence Tags (ESTs) in the *Finch*TM-Server

Todd M. Smith¹, Burak Eroglu², Joe Slagel¹, Dave Campbell¹, Peter Nelson²

1. Geospiza, Inc. 3939 NW Leary Way St. Seattle, WA 98107, USA, (206) 633-4403 www.geospiza.com

2. Fred Hutchinson Cancer Research Center 1100 Fairview Ave N. Seattle, WA 98109.

Visit us at Booth 516

Abstract

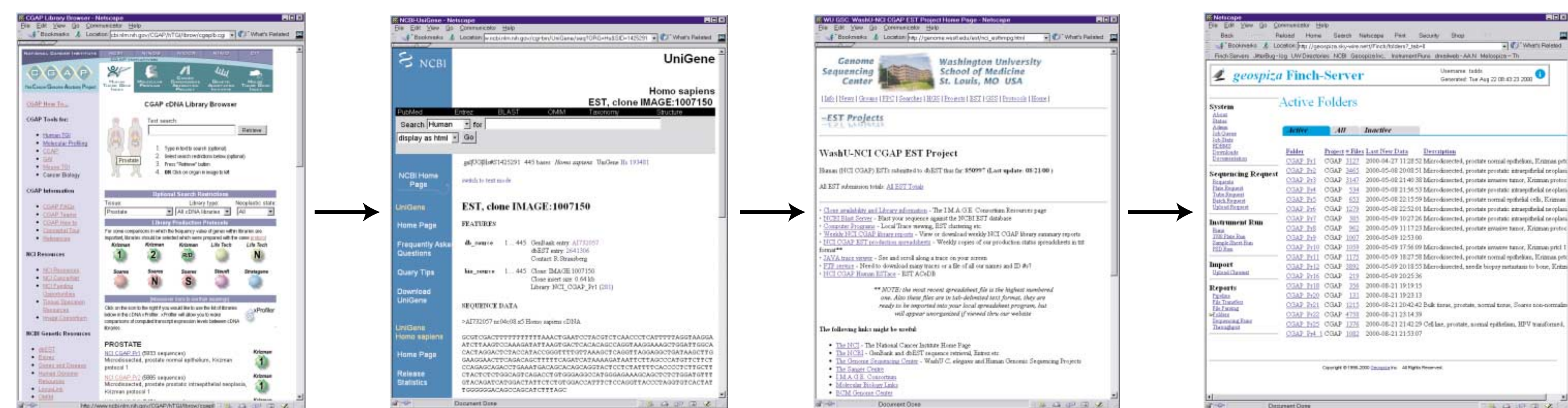
The human genome is estimated to contain between 28,000 and 100,000 genes. Most estimates are based on categorizing Expressed Sequence Tags (ESTs) by clustering and annotation. EST analysis also plays important roles in understanding normal homeostasis and diseases processes. EST clustering involves assembling large data sets of sequences and is complicated by the variability and complexity of sequences that originate from highly similar genes, have alternate forms due to RNA splicing, and sequence data quality. A complete solution for EST clustering requires software components that organize, assemble, validate, view and annotate individual sequences and sets of data. It also depends on the ability to incorporate quality information into comparative analyses to distinguish biological variation from experimental error and detect alternatively spliced messages. Finally, the system needs to incorporate data from multiple sources often stored in different file formats. Below is a detailed quality analysis of a subset of prostate sequences obtained from publicly available resources.

Methods

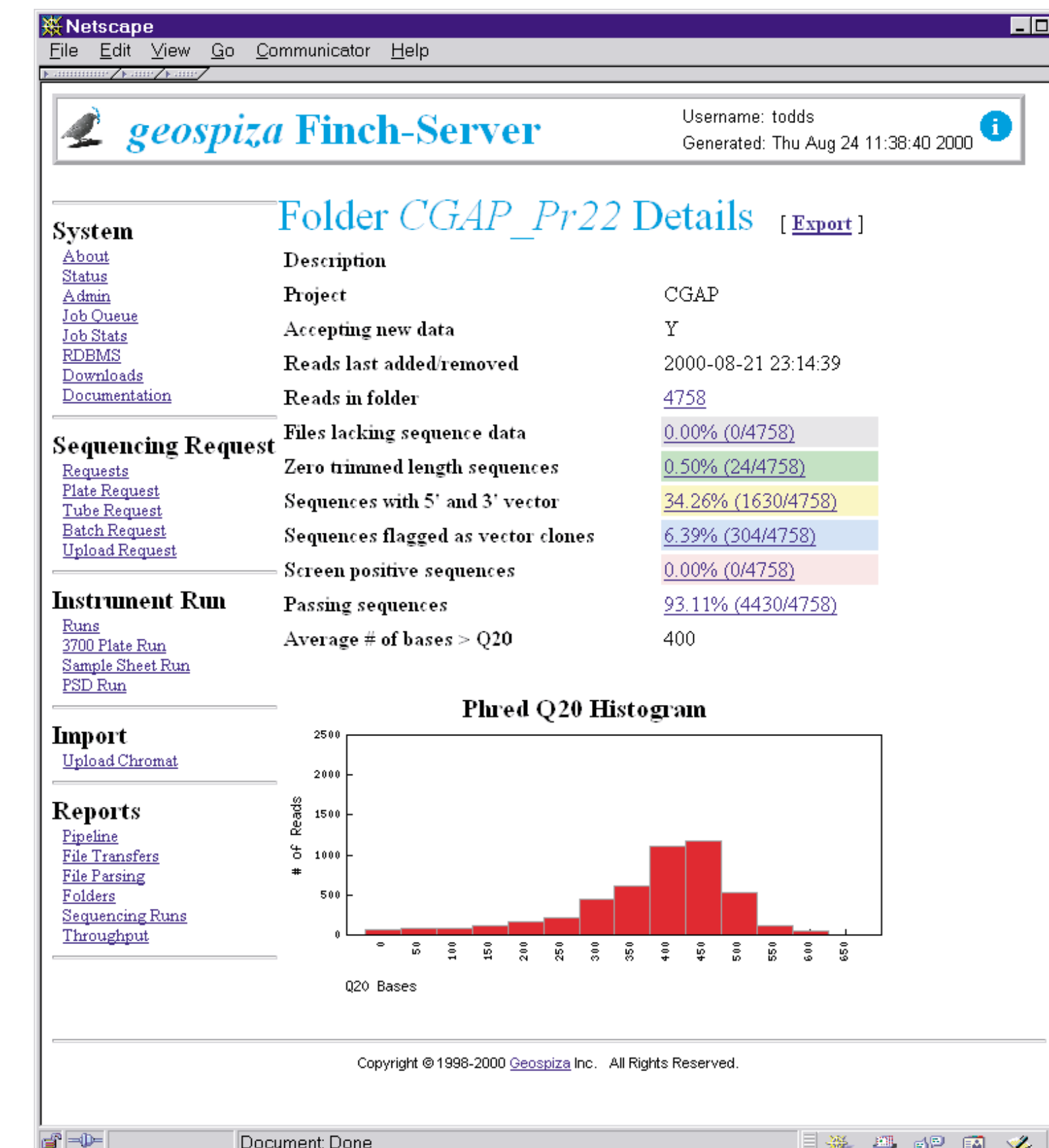
Chromatogram files corresponding to prostate ESTs were obtained as follow:

- EST libraries, organized by tissue were selected from the Human Tumor Gene Index (hTGI) at the CGAP web site (<http://www.ncbi.nlm.nih.gov/ncicgap/>).
- Each library's link was used to obtain a list of GenBank accession numbers, which were then used to retrieve sequences from Unigene (<http://www.ncbi.nlm.nih.gov/cgi-bin/UniGene/>).
- The chromatogram ID was obtained from the sequence record and used to retrieve the corresponding SCF (Staden compressed format) file from the Washington University FTP site (<ftp://genome.wustl.edu/pub>).
- SCF data were loaded into the Geospiza *Finch-Server* and automatically analyzed with Phred (Ewing et al. 1998) and *Cross_Match* (P. Green Unpublished) to obtain quality values and mask vector sequences.
- Assembly sets were created for different subsets of data and assembled with Phrap (P. Green, Unpublished).

WWW Data Flow



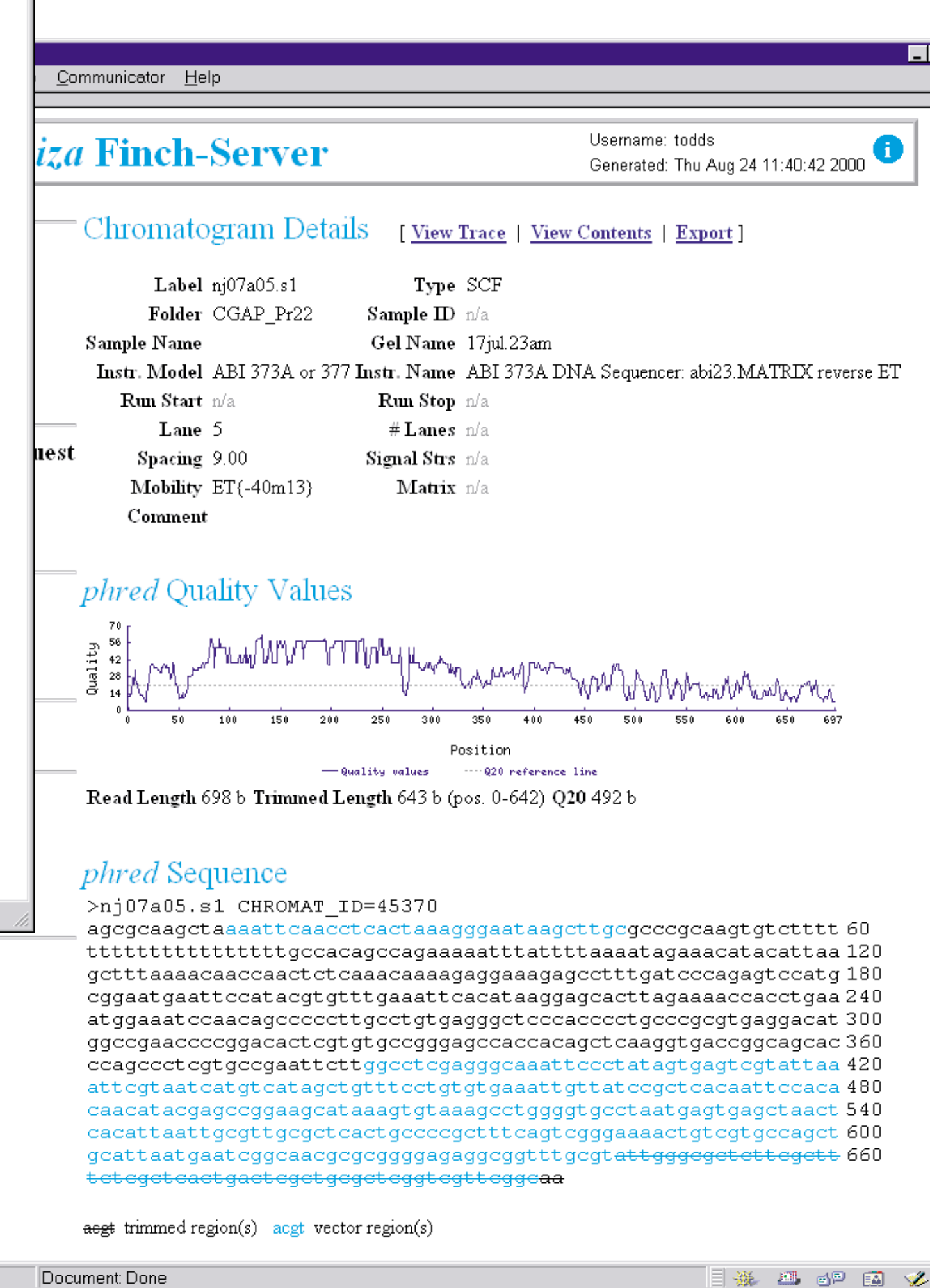
Library Report



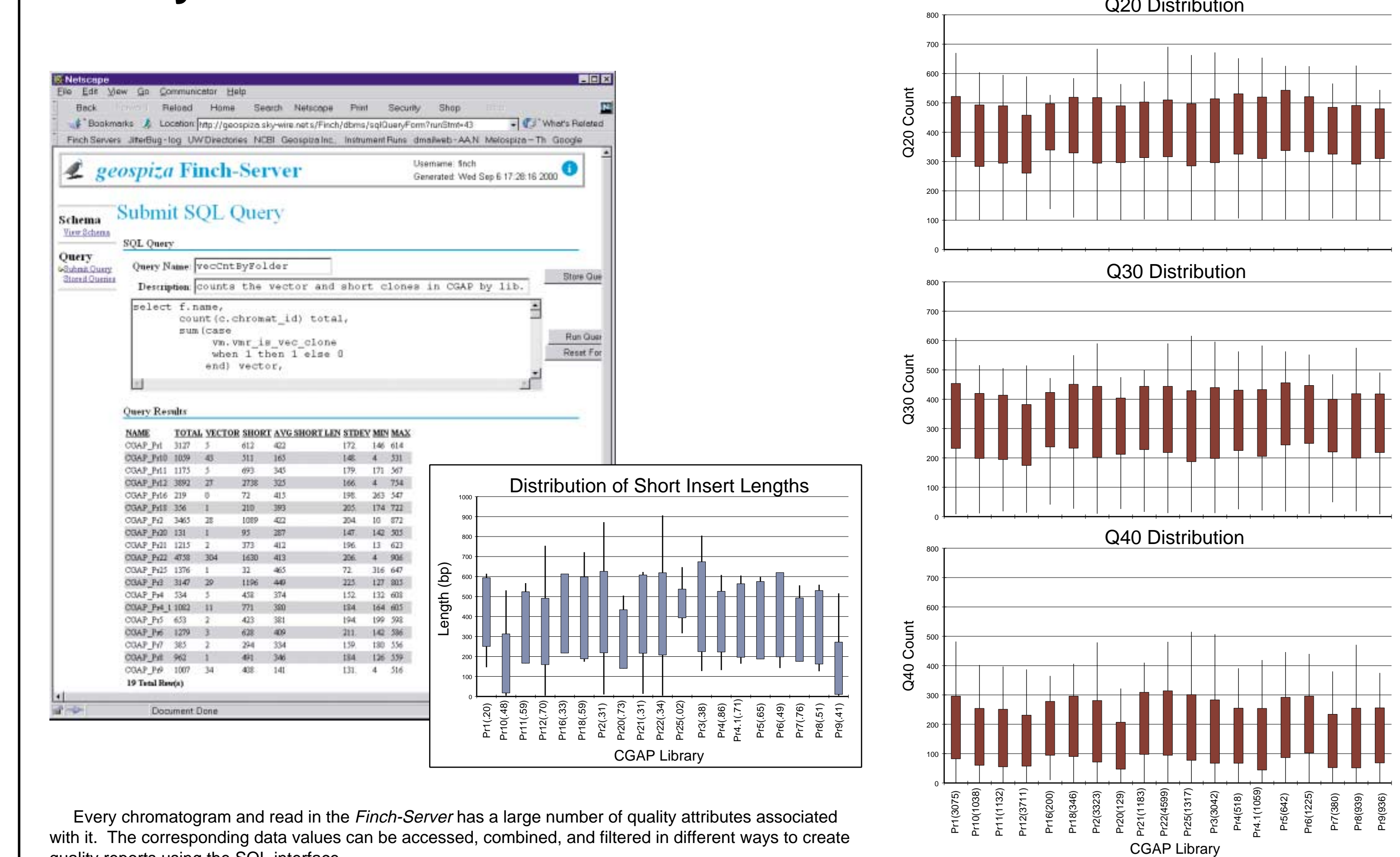
Folder detail reports (above) summarize the data processing results for chromatograms and reads. A table describes the results of processing and a histogram of Phred Q20 values shows the quality distribution.

Drill-down tables are accessed through the links in this page. The chromatogram details report (right) presents information contained in the chromatogram file, a plot of Phred qualities along the read length, and the sequence of the Phred read. Colored letters and strike through font are used to show the vector matching regions and regions of low quality, respectively. Links in this page lead to drill-down views of the trace and other information.

Chromatogram Report



Quality Statistics

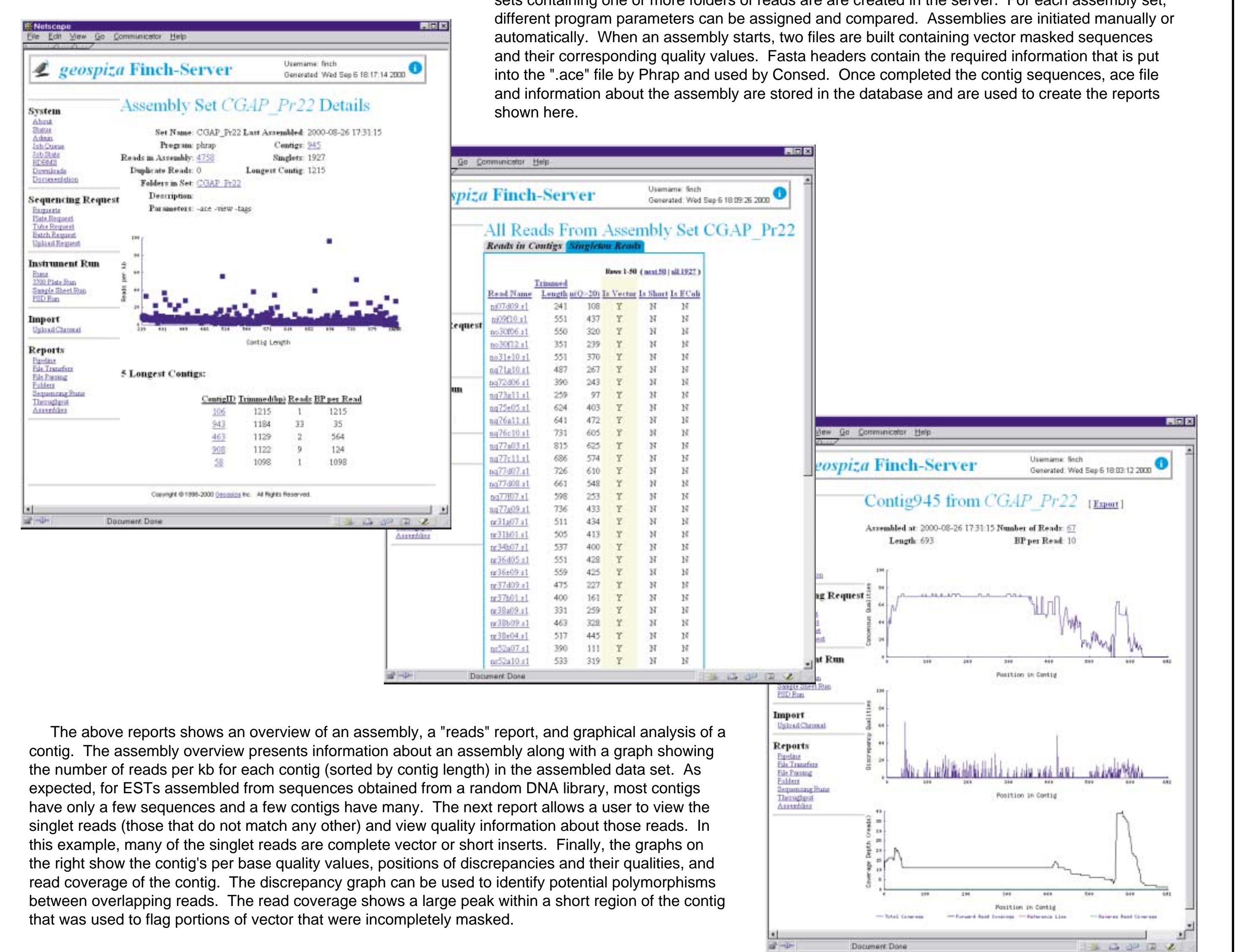


Every chromatogram and read in the *Finch-Server* has a large number of quality attributes associated with it. The corresponding data values can be accessed, combined, and filtered in different ways to create quality reports using the SQL interface.

The above screen shot shows the SQL interface and a vector report for the CGAP data. In this report, the vector analysis with *Cross_Match* is explored to determine the number of vector containing and short inserts in each library. A short insert is defined as any read that contains a 5' and 3' vector match which indicates that the insert is completely sequenced.

The graphs above to the right were produced with Excel. An SQL report was created on the *Finch-Server* and the data returned were downloaded to a local computer as a tab delimited file. The data table was then imported into Excel and formatted for graphing. The fraction of short inserts in the library is given in parentheses after the library name in the above graph.

Assembly Reports



Data are assembled in the *Finch-Server* with Phrap or SPS Phrap (www.spsoft.com). Assembly sets containing one or more folders of reads are created in the server. For each assembly set, different program parameters can be assigned and compared. Assemblies are initiated manually or automatically. When an assembly starts, two files are built containing vector masked sequences and their corresponding quality values. Fasta headers contain the required information that is put into the ".ace" file by Phrap and used by Consed. Once completed the contig sequences, ace file and information about the assembly are stored in the database and are used to create the reports shown here.

Summary

29,822 SCF files containing sequences from 19 different CGAP libraries were retrieved from the Internet and loaded into the *Finch-Server*. Each file was analyzed with the Phred basecalling algorithm and the resulting sequence read was compared to known vector sequences with *Cross_Match*. Subsets of data were then assembled with Phrap to determine groupings and identify possible polymorphisms.

The data presented, demonstrate the value of integrating data management with data analysis. A large number of cDNA clones were identified as containing short inserts and the fraction of short inserts was highly variable with some libraries containing up to 86% short inserts. Short inserts are a dead end for mapping and contig building and strategies that rely on sequencing both ends of a clone suffer from a high number of short inserts. Thus, detecting these artifacts is worthwhile.

Quality analysis with Phred allows one to observe over all sequencing quality. For detecting sequence variation and polymorphisms, a high fraction of Q30 and Q40 bases in each read is desired. Analysis of the Phred qualities shows that in many cases a good Q20 values do not correlate with high Q30 or Q40 values and each threshold needs to be examined individually. When data are assembled by Phrap, Phred quality values are used to distinguish between experimental error and biological variation. Quality values can also be used to determine a score for discrepancies. Plotting discrepancy quality by base position in a contig can be used to identify potential polymorphisms and possibly sequence fragments assembled from reads originating from highly similar genes.

Acknowledgement

This work was funded in part by SBIR grant R43HG02063-01 from the National Institutes of Health

