

Statistical process control in DNA sequencing using the Finch™-Chromatogram Server

Chris Abajian, Joe Slagel, Todd M. Smith



Geospiza, Inc.
Turning Information into Knowledge
Software for Molecular Biology

2442 NW Market St., #344, Seattle, WA 98107a
(206) 283-9338, www.geospiza.com

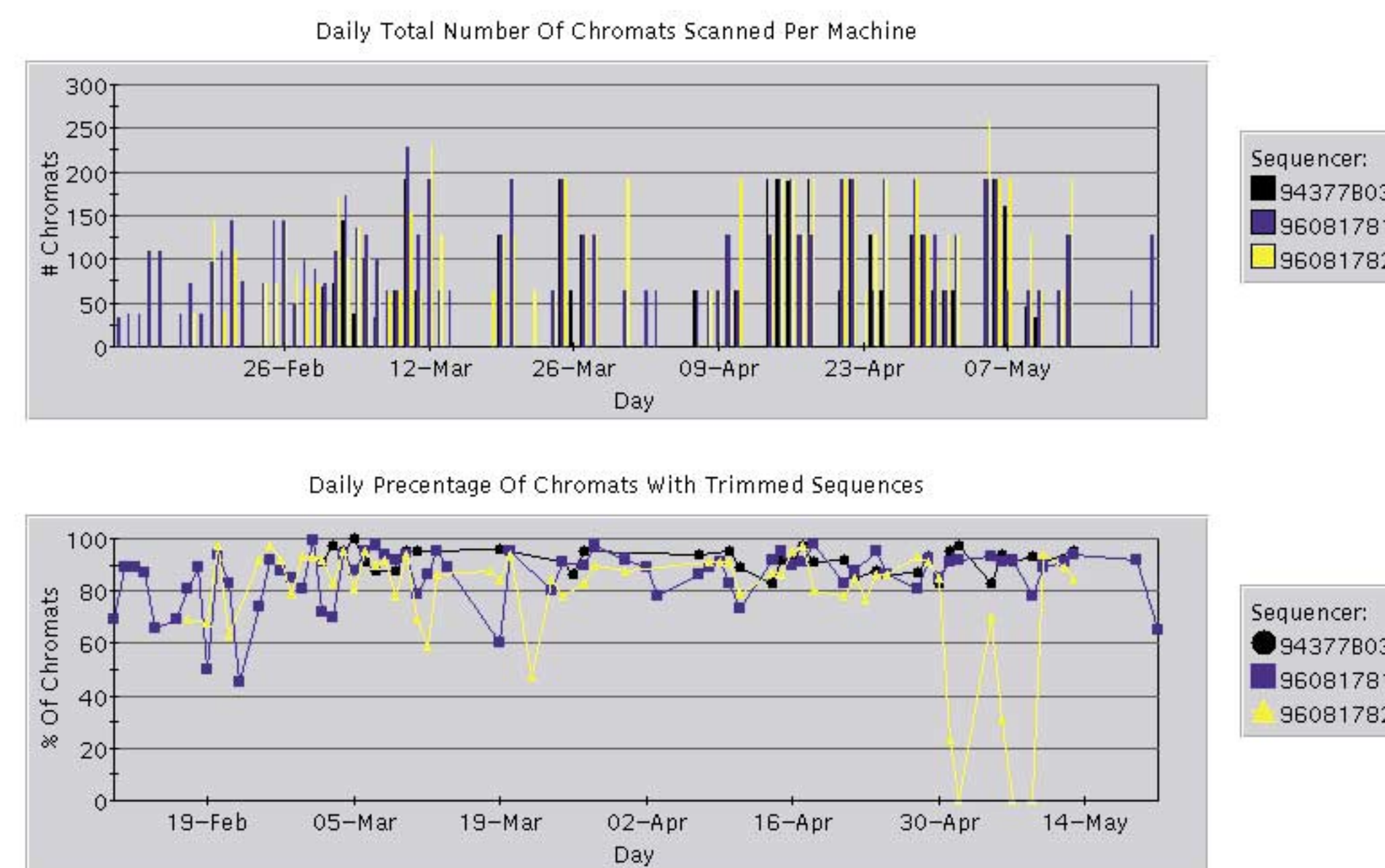
Introduction

Genome-scale DNA sequencing is a multistep process in which large numbers of template clones are propagated, purified, sequenced and analyzed on acrylamide gels. Concepts used in statistical process control (SPC, Gevirtz, C. McGraw-Hill, Inc., 1994) can be applied to high-throughput DNA sequencing to lower costs (T.M. Smith et. al. CABIOS 1997).

SPC is a quality assurance technique that is used to monitor processes that have variability and uncertainties and identify how those variations affect product quality. It is administered through process control charts that graphically compare different variables in a process. SPC is useful in reliability engineering, the analysis of failure data, sampling, process monitoring and process control.

The Finch[™]-Chromatogram Server, automates quality assurance procedures to apply SPC in DNA sequencing. The phred base calling algorithm (Ewing, B. et. al. Genome Research, 1998) is used to analyze sequence length and quality in real time. Vector, E. coli and other sequence contaminants are identified and annotated with cross_match (Green, P., unpublished). The chromatogram server monitors system performance to address issues encountered when planning for higher capacity. A web-based interface is used to control applications, enter information, and view system reports. Data demonstrating SPC concepts and applications of the Finch™-Chromatogram Server will be presented.

Throughput Reports



Library details

difficilis Finch Server Username: finch Generated: Thu Sep 10 14:09:39 1998

Library B5D

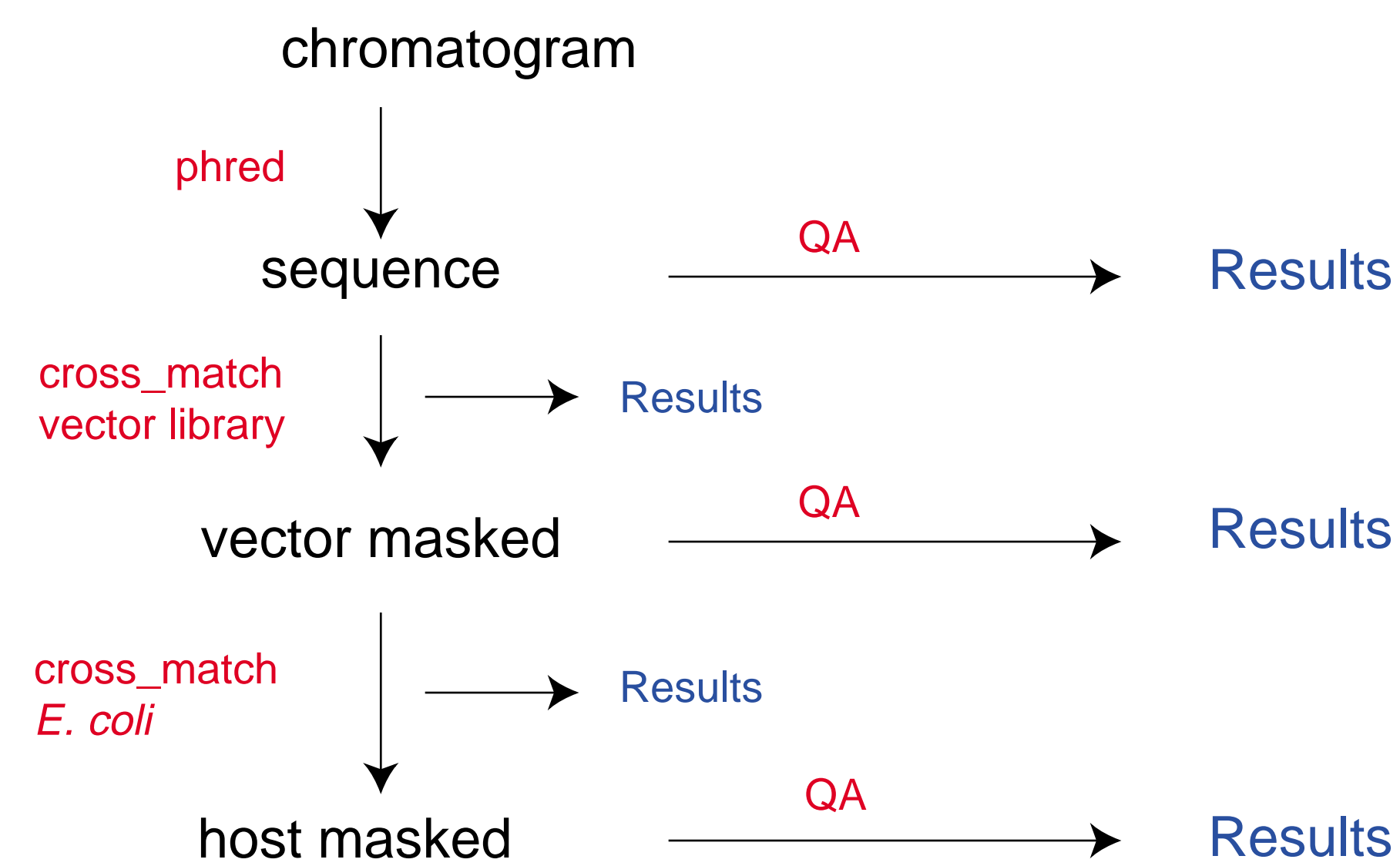
System
Description: Subclones of cosmid B5D for BRCA
Project: BRCA
Accepting new data: Y
Data last added on: 1995-01-10 18:13:01

Admin
Chromats in library: 151
Chromats with no trace data: 13995 (3/151)
Chromats with low quality data: 2856 (4/151)
Chromats with short inserts: 10569 (16/151)
Chromats flagged as vector clones: 39076 (59/151)
Chromats flagged as E.Coli: 3975 (6/151)

Chromats

Name	Gel Run Time	Machine	Trimmed Length	Is Vector	Is Short	Is E.Coli
b5d_01_g02.s1	1995-01-10 18:13:01	Genoza	396	Y	N	N
b5d_01_g03.s1	1995-01-10 18:13:01	Genoza	415	Y	Y	N
b5d_01_g04.s1	1995-01-10 18:13:01	Genoza	451	Y	N	N
b5d_01_g05.s1	1995-01-11 17:08:14	Genoza	469	Y	N	N
b5d_01_g06.s1	1995-01-11 17:08:14	Genoza	447	N	N	N
b5d_01_g07.s1	1995-01-11 17:08:14	Genoza	0	N	N	N
b5d_01_g08.s1	1995-01-11 17:08:14	Genoza	no data	N	N	N
b5d_01_g09.s1	1995-01-12 17:04:01	Genoza	453	N	N	N

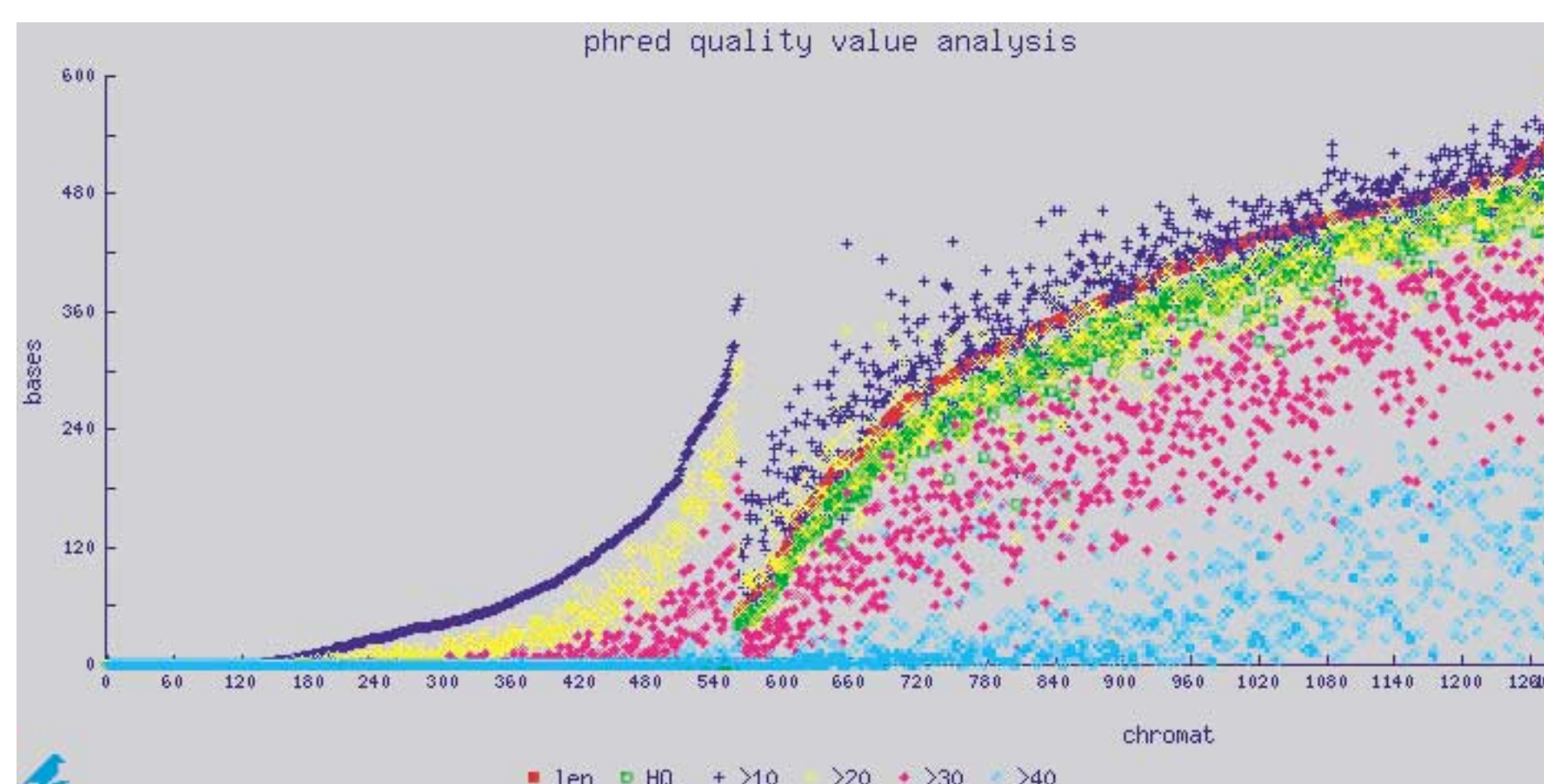
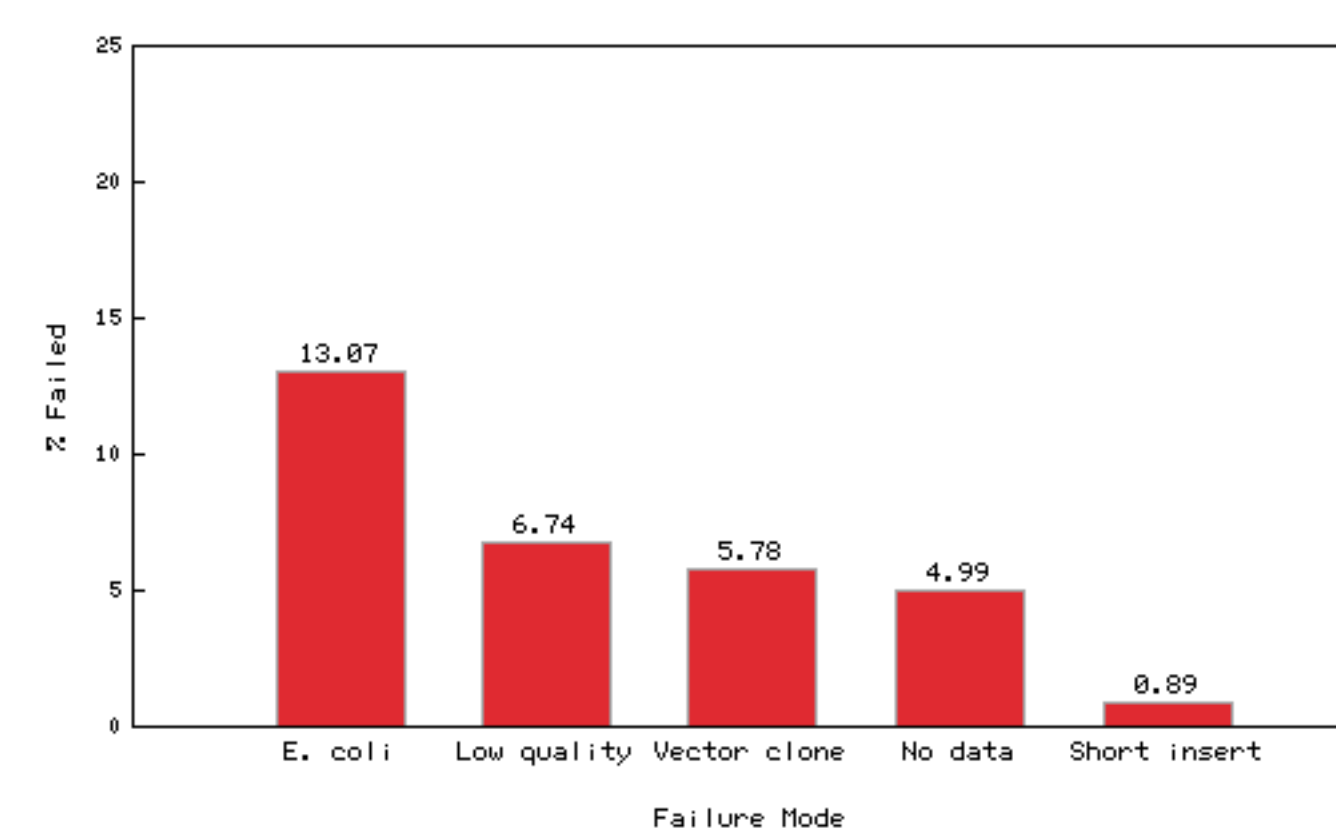
Automated Data Analysis



Example of an automated data analysis pipeline. Each chromatogram file, and resulting sequence is analyzed with different algorithms (red text). The results (blue text) are stored in an embedded relational database tables. High level summaries are created using SQL and graphically presented with CGI programs.

Real time analysis of sequence data collected on ABD-377 sequencers and managed using an Oracle RDBMS. Top. Number of sequences obtained per day per instrument (colors). Bottom. Quality analysis of the data plotted as fraction passing. Quality values were obtained with the Phred basecalling software and a passing length for each sequence was calculated with PhdClip. Graphs were created with JAVA applets using data obtained from an SQL query embedded in a CGI program (Mahairis, G., Smith, T.M., Abajian, C., Slagel, J., Hood, L., et. al. 1997).

Pareto diagrams



Chromatogram Reports

difficilis Finch Server Username: finch Generated: Fri Sep 11 16:19:06 1998

Chromat

System
Name: b5d_02_b06.s1
Orig name: B5D_02_B06.s1
Gel lane: 10
Thumbprint: 053102173114000065366334307365
Dyeset: DyePrimer(-21m13)
Direction:
Run start: 1995-01-17 17:46:30
Run stop: 1995-01-18 07:44:28
Project: BRCA
Library: B5D
Spacing: 10.62
Signal A,C,G,T: 82, 116, 78, 61

Sample Sheet
Cosmid Id: B5D
Plate Id: 02
Well Id: B06
User: MARK
% Acryl.: 4.75
Prep. Method: Sodium iodide
Cycle Method: METHODDA
enzyme: Taq
DNA Type: single

Phred Quality
Trimmed length/called: 600 / 730 (pos. 0-599) [View Trace](#)

Phred Sequence Data

```
>b5d_02_b06.s1 CHROMAT_ID=1948
tgctccaggctcgactctagaggatcccagaaggatattgggtagtntatTTTTTaaact
tgcagatttcatcctagctctccagttatctgttctcctagcactccaatgtcccaagatg
tgtaccaccaaggactctctctatttttccctgggcccctttctactgaggag
tagtggccttccatcagtagaagccgagttctgtgtccgaaattggtgggtcttgg
tctcactgactccaagaagaagtgcggaccctcaagctgagtggtacagttctaaag
atgattgtccagagttgttctctgagttcggacgtgttcagagtnaccctctct
ggtgattctgtggttctcgtggcttcaggagtgagctgcagacctttgcggtgagttt
acagctcttaaggcgcatgtctggagttgtctgttctcccgctggagttgttcatt
ctctcgtgggtctgtggtctcgtggcttcaggagtgagctgcagacctctcggctc
ggtgtttacagcagataaaatgctatgcggaccacaagagtgagcagcagcaagatttt
tcaaaagacacatgtacaaagtttcacagcgtgaaggagaccagagcgggtttctgctt
tggctcaggcagctgcatTTTTTTTTTTTatTTTTTTTTTaaagatggggctcctttca
ccaggttgat
```