Final Technical Report for DE-FG02-98ER62099

The work on this grant was detailed in our submission for the Cooperative Agreement entitled “Co-development of high throughput sequencing system with Joint Genome Institute”, # DE-FG02-98ER62637.

We have attached the pertinent section.
DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, make any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.
DISCLAIMER

Portions of this document may be illegible in electronic image products. Images are produced from the best available original document.
2.2 Automation and Informatics: Sequencing

Whitehead began a genomic sequencing effort in May 1996. In support of this work, we have developed a set of automation systems for high-throughput sequencing. The systems have been used in routine production at Whitehead for more than a year to produce 9 Mb of assembled genomic sequence, including 5 Mb of finished sequence to date.

2.2.1 Automation. Our major automation system is designed around articulated robotic arms operating within a 5' x 5' workspace, performing DNA purification and sequencing reactions (Figure 3). Nicknamed Sequatrons, these systems integrate various devices including XYZ pipetting workstations, liquid dispensers, thermocyclers, shakers, plate washers, etc. They are described in the accompanying reference (Hawkins et al., 1997). In addition, we employ various other automated and semi-automated devices.

The system has been very effective for the pilot stage, but we have also identified issues that should be addressed in constructing a next-generation system appropriate for a true factory setting. Rather than simply transfer the existing system to JGI, the purpose of this proposal is to undertake a coherent co-development program with JGI, drawing on our experiences over the past year.

In preparation for re-design of a next-generation system, we have undertaken a critical review of our current production systems. We describe the systems below, noting both advantages and shortcomings.

(i) Clone Picking. We have used a commercial PBA Flexys robot, designed for picking M13 plaques and bacterial colonies with a 24-pin tool. The robot picks at a rate of eight plates per hour and it has an autoloader designed to allow many plates to be picked in a run.

Clones are grown overnight, a technician manually centrifuges the plates, and a robot transfers the supernatant and adds SDS (in the case of M13 phage) or removes the supernatant and resuspends (in the case of bacteria).

Advantages/Shortcomings: The principal advantage of the system is that it is an off-the-shelf commercial product. However, there are a number of
significant shortcomings. The system fails to hit plaques at a low but
significant rate: M13 is not transferred to approximately 2.6% of wells. The
autoloader device is unreliable, making it impossible to set up runs with
many plates. The system is also rather expensive at nearly $150,000.

Beyond these mechanical issues, there is also a design issue. M13
growth is affected by the condition of the host cells and by growth conditions;
these variables should be rigorously standardized. However, long runs on
most plaque picking robots result in variation in host cell state at the time of
inoculation and variation in time after inoculation before transfer to 37°
for growth. We find that these variables affect consistency of yield. We
therefore perform small runs with standardized cells. This is acceptable at
low throughput, but would be inconvenient at high throughput.

Some years ago, Washington University developed a simple
alternative: this picking robot “stabs” each plaque with a small length of
surgical tubing, which is then chopped off into a well of an empty deepwell
microtiter plate. The tubing with agar can sit for up to 48 hours before
bacterial culture is added. After an entire day’s run is completed, all plates
can be simultaneously filled with standardized bacterial culture and grown for a standardized time. The design is effective, but the implementation is slow (3 plates per hour) and allows only small runs (10 plates per run).

The design and accompanying image-recognition software are freely available. As we discuss below, we propose a straightforward redesign of the Washington University system to increase speed and run size.

(ii) DNA Preparation: M13. DNA purification is performed by magnetic bead prep originally developed at the Sanger Centre and subsequently re-designed at Whitehead (Hawkins et al., 1994). The redesigned protocol is referred to as Solid Phase Reversible Immobilization (SPRI). SPRI relies on DNA precipitation on the surface of the magnetic beads in the presence of PEG, permitting thorough washing followed by release upon addition of Tris.

We have optimized SPRI for several different applications. For M13 purification, the protocol involves: starting with 100 microliters of phage supernatant; adding 0.1 mg of beads (estap@ Superparamagnetic Microspheres, 0.96 micron diameter, encapsulated, Bangs Laboratories) and 40 microliters of 26% PEG-8000, 10 mM MgCl₂, shaking vigorously (600 rpm at 1 cm diameter) for 3 min to allow the DNA to deposit on the surface of the beads; placing the plate on a strong magnetic block to retain the DNA-bead complex for 5 min; washing the plate with 70% ethanol in a plate washer; adding 40 microliters 10 mM Tris (pH 7.8), 1 mM EDTA and shaking for 5 mins to release the DNA.

The protocol is well-suited to automation. The prep is performed by a custom-built 'purification' Sequatron. This Sequatron has a throughput of one microtiter plate every 6 minutes, corresponding to 240 plates/day. We routinely process ~30 microtiter plates per day, which is sufficient to load our 19 ABI 377s three times daily.

Advantages/Shortcomings: Magnetic bead preps have long been recognized to have the advantage of being readily automated, since they simply involve reagent addition with no centrifugation. A variety of magnetic bead preps have been described over the past five years, involving a variety of approaches including both general capture (as in SPRI) and sequence-specific capture with biotinylated oligonucleotides (forming either DNA duplexes and triplexes with target sequences) and streptavidin-coated beads.

Initially, magnetic bead preps proved inconsistent in many hands and were abandoned by many groups. The reasons are now clear, being related to inconsistency in the beads themselves and in the purification protocol. The magnetic beads are now available in an encapsulated form designed for compatibility with molecular biology. In addition, consistency depends on a number of factors: the PEG solution is important and must be optimized for the application (26% PEG, 10 mM MgCl₂ for M13 vs. 20% PEG, 2.5 mM NaCl₂ for PCR products); vigorous shaking is required to mix the DNA in the PEG solution; and very strong rare earth magnets must be used to pull beads from the viscous PEG solution (we have found only a single supplier (Amag) that
sells sufficiently strong magnets). Because of the consistency required in shaking, incubation and washing, the SPRI protocol works much better on an automated workstation than in a manual format.

The purified DNA is very clean (readily cut by restriction enzymes) and provides a good sequencing template. Sequencing reactions on SPRI templates provide ~750 bases of alignable sequence and ~550 bases with Phred quality score ≥ 20, in 5.5 hour runs on ABI377s. (These numbers are comparable or better than results for other preps run under similar gel conditions.) A histogram of recent data is shown in Figure 4. In addition, the prep is inexpensive: the magnetic beads cost 2.4¢ per sample and the overall prep costs ~5¢ per sample.

![Proportion of Reads](image)

Figure 4. Quality scores for M13 templates produced by SPRI purification, sequenced with dye-primers and electrophoresed for 5.5 hours on an ABI 377. Data was analyzed by the ABI signal processing software and by the Phred base calling software. Histogram shows the proportion of templates with specified number of bases having Phred scores ≥ 20. Median of overall distribution = 553. Median of second peak (excluding 'failures') = 571.

The robotic implementation of the SPRI protocol in the Sequatron system has been quite successful. The entire process operates in a stand-alone mode and has been used to produce all M13 templates produced at Whitehead. Nonetheless, we are not completely satisfied by the mechanical performance. We have measured a 2.1% failure rate — attributable entirely to pipetting failure (almost always failure to dispense the magnetic beads). We propose to address this in the comprehensive redesign by using a much more reliable positive-displacement pipetting system, as on the Genomatron.

In addition, the current purification robot does not routinely quantitate the DNA yield in each sample. DNA yield is the most significant variable.
affecting sequencing success. Low DNA yield (≤ 50 ng/ul) may result from incorrect picking, poor M13 growth, or purification failure. DNA quantitation would provide a valuable QC tool. It would also allow robotic replacement of wells with low DNA yield (from successful samples on a source plate) before proceeding to the most expensive step of sequencing reactions. We have recently initiated DNA quantitation on a pilot basis using a CytoFluor II fluorescence plate reader (Perceptive BioSystems), using Oligreen and Hoechst 33258. We plan to include such quantitation (to be uploaded to the central database) in the comprehensive redesign.

(iii) DNA Preparation: Plasmids. Recently, we have expanded our shotgun sequencing from exclusively M13 clones to a mixture of both M13 and plasmid clones. The plasmid provide convenient forward-reverse links and appear to cover some regions poorly covered by M13. Several sequencing groups are experimenting to find the optimal mixture. We currently sequence 75% M13: 25% plasmid.

We are still settling on the most suitable plasmid prep. Three approaches have been used:

- PCR provides an very effective and convenient plasmid prep. We are currently using it as our preferred method. The protocol has been optimized as follows: An overnight growth of bacterial culture is diluted 1:3 in water and 2 ul are added to a 40 ul PCR reaction, using both Taq and Pfu polymerases (16:1 ratio) and ‘hot-start’ conditions. (The addition of Pfu and the hot start greatly improve the success rate, leading to greater amplification and eliminating primer dimer.) The protocol is readily performed on our robotic workstation. With the optimized protocol, a single strong product is produced in 91% of samples.

- We also employ traditional manual alkaline lysis, using commercial filtration-based kits (ATGC) These produce satisfactory templates in ~95% of samples.

- We have also used a hybrid protocol in which the second half of the alkaline lysis protocol (filtration through a filter, isopropanol precipitation with 20 min centrifugation, and wash with 70% ethanol) is replaced by SPRI purification; this approach was described in Hawkins et al. (1994).

Advantages/Shortcomings. The PCR prep has the advantages that it is extremely easy to automate (the bacterial culture does not even need to be pelleted), provides high yields of good templates in the vast majority of cases and is relatively inexpensive (10¢ per sample). The disadvantage is that ~9% of samples do not yield a suitable template fragment. We do not currently eliminate these samples before sequencing, resulting in wasted sequencing reactions and empty lanes. With the availability of on-line DNA quantitation, one could eliminate these samples before performing sequencing reactions.

The commercial filtration-based alkaline lysis preps have a lower failure rate (~5%), but are labor intensive and more expensive (51 per sample with ATGC or Qiagen). Use of SPRI purification eliminates the most
expensive component (the filtration plate) and the lengthy isopropanol precipitation, although it would not fully automate the protocol.

As discussed below, the optimal production line for plasmid preps is still being discussed by the Co-Development Program. In addition to the above approaches, Paula McCready of the JGI has developed a new prep based on an ion-exchange resin in solution; it may be possible to automate this prep using an inexpensive filtration system. We return to these issues below.

(iv) Sequencing Reaction Set Up. Sequencing reactions are set up by a Sequatron robot feeding plates to a Packard workstation.

The Sequatron is also designed to load the plates into a Techne thermocycler and remove them after thermocycling. We do not use this feature at present, because there are occasional problems with plate removal (plates warp slightly when thermocycled) and our current throughput is sufficiently modest that it is no trouble to load plates by hand. The problem seems to be solved by the use of new, stiffer 384-well plates (Advanced Biotechnologies) in conjunction with MJ thermocyclers treated regularly with Teflon spray. However, we have not yet introduced this change into regular production.

Advantages/Shortcomings. Robotic setup of sequencing reactions is straightforward and a reliable interface with the thermocycler appears to be in hand.

The main shortcoming is the limited accuracy and reliability of standard liquid handling robots. Although they can transfer volumes of 2 ul, they will not be able to handle substantially smaller volumes (which would be advantageous in decreasing sequencing costs). In addition, the liquid handling robots have a small but significant failure rate in properly depositing small volumes; the rate may be about 1-2%.

It would be advantageous to employ lower volume, higher precision, higher reliability transfer devices.

(v) Finishing Reactions. The initial phase of finishing involves performing directed reverse reads and directed terminator reactions on selected templates. As discussed below, our assembly software automatically selects clones for such reactions. The software downloads a file to a Packard workstation specifying the clones to be selected. A technician then loads the appropriate source plates on the workspace and the Packard re-arrays the selected clones into plates for subsequent sequencing.

Advantages/ Shortcomings. The system is straightforward and works well. Its major shortcoming is that the current workspace is limited to 8 plates per run. This is easily addressed in the proposed production line.

Overall evaluation of current system. The current production system has been quite effective. The entire production process from picking through sequencing reactions is carried out by 4 laboratory aides, who process 30 microtiter plates per day. (A larger staff is required for gel cleaning, preparation and loading than for all aspects of the microbiology and biochemistry.)
Still, we do not believe that it would be optimal to simply replicate the system at JGI. Regular use in daily production has identified ways to redesign the system to be more appropriate for a true factory operation. Accordingly, the Co-Development Program between Whitehead and JGI aims to re-design and re-implement to reach a higher level of performance.

Gel Electrophoresis. Finally, we briefly discuss the interface with the ABI 377. We employ 8-channel Hamilton syringe loaders, which are rapid and convenient for loading 48-lane gels. We perform 'staggered' loading, which aids in lane tracking.

We are currently extending this approach to 96-lane gels. We have designed a motorized 48-channel loader (to be used for two staggered loads in alternate lanes). The device is based on the existing 48-channel pipettor used on the Genomatron, and will employ positive displacement pipetting to deliver 0.5 - 2 ul. The loading tips are held in rigid array by a cross bar. We propose to develop this simple device under the Co-Development Program.

The approach provides an alternative to a gel loading system at Washington University, in which microtitre plates are placed in a hermetically sealed box and positive pressure is used to carry the samples through capillaries into the wells. This system also has attractive features, but it is too early to know the performance and reliability of either system in a production setting. Accordingly, we believe that it is sensible to develop this alternative approach.

2.2.2. Informatics. Our sequencing informatics infrastructure consists of a heterogeneous collection of software packages, scripts and data stores, developed at Whitehead and imported from elsewhere. We briefly describe some key components

(i) Analytical software packages.
   • Lane tracking. Gel files are uploaded from the Macintoshes on each ABI377 to a UNIX system. Lane tracking is performed by the Bass/Grace package (written by Lincoln Stein at Whitehead, based on an earlier program by Anthony Bemo at Stanford). Tracked gels are stored in a queue and must be manually reviewed (and corrected, if necessary) before being passed on to the next step.
   • Five-color lane tracking. Lane tracking should ideally be based on a known internal standard, rather than on the unknown experimental signal. We developed such an approach for ABI-based genotyping, and are now adapting it to sequencing.

In ABI-based genotyping, polymorphic fragments labelled in three colors are measured against an internal size standard labelled in a fourth color. Two years ago, we modified this approach by placing different size standards in alternating lanes (pattern A+B in odd lanes, A+C in even lanes, providing some constant and some differing bands). We (Mark Daly and Eric Lander) developed lane-tracking software to use the known alternating pattern of size standards to identify lanes. If the expected pattern is found in every lane, the gels are automatically passed to the next step. If the expected
pattern is not found in every lane, the computer presents the image for human inspection and specifically flags the problematic lanes. The software has been in regular use for 18 months. Compared to current lane tracking software for sequencing gels, this system is more accurate, requires less human review, and aids in sample tracking (as even and odd lanes cannot be confused).

The same approach could be directly applied to sequencing gels, using internal standards in a fifth color. Accordingly, we began a collaboration with ABI to develop and apply five-color signal-collection software. ABI has just completed the five-color detection software (as of Oct 28, 1997). After we have tested the software, we will adapt our lane-tracking software to sequencing gels (using a set of four different internal standards labelled in Sy2, whose emission peak lies sufficiently below the dye sets used for sequencing).

The software should prove especially valuable for tracking 96-lane gels.

- **Signal Processing/Base Calling.** Traces extracted from each lane are passed to TROUT (written by Mark Daly and Eric Lander at Whitehead), which has two modules: TROUT-1, a signal processor that performs color correction, background subtraction, mobility correction, etc. and TROUT-2, a base caller. TROUT also provides a quality score, based on the traditional EBA score (reflecting proportional signal in the called base).

A major motivation for writing TROUT was to eliminate the need to return to the Macintosh platform to manually run ABI's signal processor and base caller.

Phil Green's PHRED program provides a UNIX-based base caller, but still requires the output of the Macintosh-based ABI signal processor.

Recently, Green and Ewing have been developing a UNIX-based signal processor (called PLAN) which is currently in beta-testing. We are collaborating with Green and Ewing to study the results of these signal processing and base calling systems (PLAN/PHRED and TROUT-1/TROUT-2), with the aim of identifying possible improvements.

Both systems outperform the ABI software. Each has certain advantages and disadvantages depending on the data type. (N.B. Base callers and signal processors are intimately tied together; the former is closely "tuned" to the output from the latter. For example, PHRED is tuned for the highly smoothed data from the ABI signal processor but returns poor quality scores when presented with the less-smoothed output of TROUT-1.)

- **Assembly and Finishing.** Shotgun assembly is performed by the ALEWIFE package (Mark Daly and Eric Lander). ALEWIFE detects overlap primarily based on co-occurrence of perfect 25-mer matches among reads (exploiting the fact that 25-mers occurring only once in a project typically represent sequencing errors, while 25-mers occurring too frequently typically represent repeats). (Similar principles are used in the latest version of TIGR Assembler (M. Adams, pers. comm.).) ALEWIFE creates contigs, automatically selects clones for the "prefinishing" stage (directed reverse reads, directed terminators), and passes the assembly to the GAP4 program (Bonfield and Staden, Cambridge) for viewing. In addition to ALEWIFE, we are also
incorporating Phil Green's Phrap assembler—thus employing two independent assemblers to identify potential problems.

Sequence editing is performed first by the WI/AutoEditor (substantially modified from the AutoEditor program by Richard Mott at the Sanger Centre) and then by human finishers. Human finishers manage finishing projects with GAP4—adding data, making joins and continuing to edit. Once a project is apparently finished, final sequence verification is performed by the BigBrother package (written at Whitehead). This program identifies all regions with conflicts or single-stranded coverage and also calculates expected restriction digest patterns for checking against the experimentally-observed pattern.

(ii) Data stores and scripts. Reflecting its evolution over time, the overall informatics system employs a wide array of persistent data stores and scripts. This diversity has been a strength for the pilot phase, but is also a weakness for scale-up.

The variety of persistent data stores include:

- an object-oriented database system for workflow management (LabBase, developed at Whitehead);
- flat-file databases and program logs (including the Boulder IO data exchange format, developed by Lincoln Stein at Whitehead);
- the UNIX file system functioning as a finite-state machine, in which files are moved to new directories after processing; and
- GAP databases for managing individual sequence assemblies

Processing is driven by a "pipeline" consisting of a large collection of Perl scripts that run as UNIX cron jobs or as initiated by human operators. The pipelines are responsible for automatic initiation of lane tracking, base-calling, initial assembly, daily deposit of data in public databases, etc. Various tools query the system to perform sample and process tracking. The TASKMASTER application, for example, interfaces to multiple datastores to track sequencing status—providing various reports about clones in process, shotgun sequencing status, QC reports, work in finishing queue, etc.

(iii) Evaluation of System. The Whitehead sequencing informatics system (like those at other genome centers) is an impressive but ungainly assemblage of diverse components.

The system has been very successful at supporting pilot levels of sequencing (up to ~20 Mb), substantially decreasing staffing requirements Nonetheless, a serious evaluation of current sequencing informatics systems (at Whitehead and elsewhere) must that conclude they are not suitable for a high-throughput production factory.

Data Storage. The problems with existing data storage systems include the following:

- Enhancements become increasingly difficult in systems cobbled together from diverse components. Relevant data (e.g., for process tracking, QC or analysis) may be scattered across various different data stores. Important parameters (e.g., values of variables and orders of operations) may be embedded deep within applications and not readily accessible.
Data storage systems relying on flat files have serious problems with regard to concurrency control. Concurrency control is enforced only through the UNIX file system, which can employ a 'lock out' mechanism to prevent more than one user from accessing a file at a time. This approach is sufficient for mediating among a few human users, but does not accommodate to the situation of high speed data streams with multiple processors. Industrial-strength databases solve this problem by attaching concurrency control to individual 'data atoms' rather than to entire files.

Locally-developed data storage systems (e.g., LabBase, Acedb, etc.) tend to suffer from size limitations—because of either absolute bounds embedded in the software or performance degradation with increasing size. As a result, genome centers have tended to maintain many small databases. Industrial-strength relational databases have been engineered to accommodate much larger datasets (by at least two orders of magnitude).

It should be emphasized that current sequencing informatics systems have performed admirably. However, we recognize that they are not well-suited for scale up. The systems should be re-implemented in an industrial-strength relational database (such as Sybase), to provide a firm foundation for scale up. RFA 97-17 provides the first focused opportunity to carry out such a coherent re-implementation.

Two members of the Co-Development Team—Ken Fasman from Whitehead and Tom Slezak from JGI—have collaborated together for several years on the design and implementation of such database systems and will jointly oversee the work.