SANDIA REPORT
SAND2009-7369
Unlimited Release
Printed November 2009

Host Suppression and Bioinformatics for Sequence-based Characterization of Unknown Pathogens

Mllind Misra Julia N. Kaiser, Robert Meagher, Steve Branda, Kamlesh Patel, and Todd W. Lane

Prepared by
Sandia National Laboratories
Albuquerque, New Mexico 87185 and Livermore, California 94550

Sandia is a multiprogram laboratory operated by Sandia Corporation,
a Lockheed Martin Company, for the United States Department of Energy's
National Nuclear Security Administration under Contract DE-AC04-94AL85000.

Issued by Sandia National Laboratories, operated for the United States Department of Energy by Sandia Corporation.

Available to DOE and DOE contractors from
U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831

Telephone: (865) 576-8401
Facsimile: (865) 576-5728
E-Mail: reports@adonis.osti.gov
Online ordering: http://www.osti.gov/bridge

Available to the public from
U.S. Department of Commerce
National Technical Information Service
5285 Port Royal Rd.
Springfield, VA 22161

Telephone: (800) 553-6847
Facsimile: (703) 605-6900
E-Mail: orders@ntis.fedworld.gov
Online order: http://www.ntis.gov/help/ordermethods.asp?loc=7-4-0#online

SAND2009-7369
Unlimited Release
Printed November 2009

Host Suppression and Bioinformatics for Sequence-based Characterization of Unknown Pathogens

Mllind Misra Julia N. Kaiser, Robert Meagher, Steve Branda, Kamlesh Patel, and Todd W. Lane

Biosystems Research Department
Sandia National Laboratories
PO Box 969
Livermore, CA 94550-0969

This page intentionally left blank

Contents

List of Figures

List of Tables

Executive Summary

Bioweapons and emerging infectious diseases pose formidable and growing threats to our national security. Rapid advances in biotechnology and the increasing efficiency of global transportation networks virtually guarantee that the United States will face potentially devastating infectious disease outbreaks caused by novel ("unknown") pathogens either intentionally or accidentally introduced into the population. Unfortunately, our nation's biodefense and public health infrastructure is primarily designed to handle previously characterized ("known") pathogens. While modern DNA assays can identify *known* pathogens quickly, identifying *unknown* pathogens currently depends upon slow, classical microbiological methods of isolation and culture that can take weeks to produce actionable information. In many scenarios that delay would be costly, in terms of casualties and economic damage; indeed, it can mean the difference between a manageable public health incident and a full-blown epidemic.

To close this gap in our nation's biodefense capability, we will develop, validate, and optimize a system to extract nucleic acids from unknown pathogens present in clinical samples drawn from infected patients. This system will extract nucleic acids from a clinical sample, amplify pathogen and specific host response nucleic acid sequences. These sequences will then be suitable for ultra-high-throughput sequencing (UHTS) carried out by a third party. The data generated from UHTS will then be processed through a new data assimilation and Bioinformatic analysis pipeline that will allow us to characterize an unknown pathogen in hours to days instead of weeks to months. Our methods will require no *a priori* knowledge of the pathogen, and no isolation or culturing; therefore it will circumvent many of the major roadblocks confronting a clinical microbiologist or virologist when presented with an unknown or engineered pathogen.

1. Molecular Biology

Introduction

The goal of the molecular biology portion of this research was to develop and demonstrate protocols that are able to selectively enrich/suppress nucleic acid (NA) fragments specific to a pathogen or to the host response to a pathogen. Such subtractive techniques are advanced molecular biology research methods that can amplify, in an undirected way, pathogen-derived NAs from a complex background of host NAs. We have prepared mock clinical samples on which to carry out this development and evaluation. We have used the human 293 cell line as the model host and the bacteriophage T4 as a mock pathogen in these samples. These models were selected because they are both Risk Group 1, and can be safely handled in the laboratories at Sandia. In addition, we are able to easily generate high titer stocks of T4 in-house and obtain commercially available 293 genomic DNA (Genscript Corporation, M00094). We set out to develop benchtop methods to detect and identify small amounts of T4 nucleic acids in a background containing large amounts of contaminating host 293 DNA. In order to characterize the success of a given subtractive technique, we must be able to quantitate levels of host and pathogen NA in a given sample. We tested two different methods of nucleic acid detection: DNA blots probed with fluorescent or biotinylated probes and quantitative PCR using Taq-man probes.

DNA Southern Blotting

We loosely modeled our DNA blotting methods after a previously available commercial kit, the QuantiBlot Human DNA Quantitation kit (Applied Biosystems, Inc). This product was commonly used for detecting and quantifying small amounts (supposedly as little as 0.15 ng) of human DNA in forensic applications, until newer, more rapid methods replaced it. Our blotting method involves immobilizing human 293 or T4 DNA on a nylon membrane via a slot blot apparatus, and then hybridizing fluorescently labeled or biotinylated probes complementary to human or T4 genes to the DNA on the membrane. The fluorescent probes were directly detected on a Typhoon phosphorimaging machine. The samples incubated with the biotinylated probes underwent detection by enhanced chemiluminescence (ECL) using streptavidin-horseradish peroxidase (Perkin Elmer) and ECL plus western blotting reagents (Amersham). The quantity of DNA is determined by comparing the intensities of the resultant bands to the intensities of the bands from a serial dilution of standard DNA ranging from 0.15 to 10.0 ng. In addition to the Quantiblot D17Z1 human DNA probe, which is complementary to a primate specific alpha satellite DNA sequence on chromosome 17, we designed and tested a series of other human and T4 specific probes, described in
Table 1. A detailed outline of our blotting method follows.

Materials:

Biodyne B Membranes (Pall, 60209)
D17Z1 biotinylated probe (5'- TAG AAG CAT TCT CAG AAA CTA CTT TGT GAT GAT TGC ATT C -3')
human and T4 specific biotinylated probes (IDT)
human genomic DNA standard (Genscript M00094)
T4 genomic DNA standard (prepped in-house with Qiagen lambda kit)
HRP-SA (Thermo/Pierce, PI-21130)
30% H_2O_2 (Sigma, 216763-100 mL)
Citrate Buffer: 0.1 M Sodium Citrate, pH 5.0 (Teknova, S0231)
Hybridization Solution: 5X SSPE, 0.5% w/v SDS
for 1 liter: 250 mL 20X SSPE
700 mL H_2O
50 mL 10% SDS
Pre-Wetting Solution: 0.4N NaOH, 25 mM EDTA
for 1 liter: 100 mL 4 N NaOH
850 mL H_2O
50 mL 0.5 M EDTA
Spotting Solution: 0.4 N NaOH, 25 mM EDTA, 0.00008% Bromothymol Blue
for 100mL: 100 mL pre-wetting solution
200 uL 0.04% bromothymol blue solution
Wash Solution: 1.5X SSPE, 0.5% w/v SDS
for 1 liter: 75 mL 20X SSPE
875 mL H_2O
50 mL 10% SDS

Procedure:

Preparation of Human DNA Standards

1. Label seven microcentrifuge tubes A through G.
2. Vortex DNA Standard and quick spin.
3. Pipette 14 ul DNA Standard into tube labeled A.
4. Pipette 7 ul TE, pH 8.0 into each of the remaining tubes.
5. Do a serial dilution by pipetting 7 ul from tube A into tube B.
6. Mix.
7. Pipette 7 ul from tube B into tube C.
8. Mix.
9. Continue dilutions through tube G.

DNA Standard Concentration (ng/ul)	Quantity of DNA in 5ul
A 2	10
B 1	5
C 0.5	2.5

D 0.25	1.25
E 0.125	0.625
F 0.0625	0.3125
G 0.03125	0.15625

Slot Blotting / DNA Immobilization

1. Determine the number of tubes needed for standards, blanks and samples. Label microcentrifuge tubes according to placement on the membrane (i.e. A1, A2, B1, B2, etc.).
2. Pipette 150 ul of Spotting Solution into each tube.
3. Vortex Human DNA standards. Quick spin. Pipette 5 ul of DNA solution into the appropriate tube.
4. Vortex DNA samples. Quick spin. Pipette 1 to 20 ul of sample DNA into the appropriate tube.
5. While wearing clean gloves, cut a piece of Biodyne B membrane 11.0 cm by 8 cm. Label in the upper right-hand corner to mark orientation. Place the membrane in a tray containing an adequate amount of Pre-Wetting Solution to wet the membrane. Leave for 1 to 30 minutes.
6. Using forceps, place the wetted membrane on top of 3 sheets of pre-wetted filter paper in the slot blotting apparatus (BioRad Biodot SF), according to Figure 1**.** Tighten the sealing screws finger-tight, in a diagonal pattern.
7. Rinse the tray with DI water.
8. Slowly pipette each sample into the center of the appropriate well of the slot blot apparatus. Take care not to get bubbles on the membrane in the well(s).
9. Slowly turn on the sample vacuum. After all samples have been drawn through the membrane inspect each slot for a uniform blue band. Turn off the sample vacuum.
10. Turn off the vacuum source. Disassemble the slot blot apparatus and remove the membrane. Proceed immediately to hybridization or store the membrane in 5X SSPE at 4°C for up to 24 hours. Do not allow the membrane to dry out.
11. Wash the slot blot apparatus with 0.1% SDS. Rinse apparatus with an excess of DI water and allow to air dry. Do not use bleach.

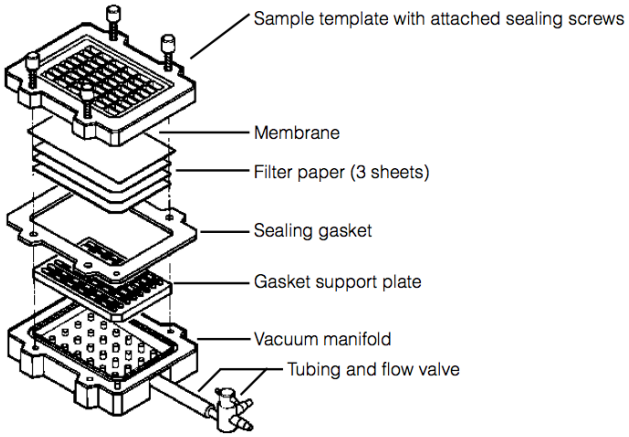

Figure 1: Proper assembly of Biodot SF slot blotting apparatus.

Hybridization

1. Do not allow the membrane to dry out during any of the following steps. Warm the Hybridization Solution and the Wash Solution to 37° to 50°C before use. All solids must be in solution before use. Turn on the hybridization oven and warm to 50°C.
2. Place the membrane into the hybridization tube. Add 60 ml of pre-warmed Hybridization Solution. Add 3 ml of 30% H2O2. Cap the tube, and place into a clamp in the hyb oven. Rotate at 50-60 rpm for 15 minutes. Pour off solution.
3. Add 30 ml of pre-warmed Hybridization Solution to the tube with the membrane. Add 20 ul of QuantiBlot® D17Z1 Probe into the Hybridization Solution (final concentration = 5-10 ng/mL). Place the lid on and rotate at 50-60 rpm and 50°C for 20 minutes. Pour off the solution.
4. Rinse the membrane in 60 ml of pre-warmed Wash Solution briefly. Pour off the solution.
5. Add 30 ml of pre-warmed Wash Solution to the tray. Tilt tray to one side and pipette 180 ul of HRP-SA into the solution. Rotate at 50-60 rpm and 50°C for 10 minutes. Pour off solution.
6. Rinse with 60 ml of Wash Solution for 1 minute at room temperature. Pour off solution. Repeat.
7. Add 60 ml of Wash Solution. Rotate at 100-120 rpm at room temperature on an orbital shaker for 15 minutes. Pour off solution.
8. Rinse the membrane briefly in 60 ml of Citrate Buffer. Pour off the solution.

Detection

1. Remove ECL plus detection reagents from fridge, and allow to come to room temperature. Mix solution A with solution B at a ratio of 40:1 (2 mL A + 50 uL B).
2. Drain excess wash buffer from the membrane and place it, DNA side up, on a piece of saran wrap. Pipette the A/B mixture onto the membrane.
3. Incubate for 5 minutes at room temperature.
4. Drain off excess detection reagents by touching the edge of the membrane to a piece of paper towel. Place membrane, DNA side down, on a clean piece of saran wrap, and smooth out any air bubbles.
5. Photograph the blot on the Fluorchem imager using chemiluminescence filter. Or, scan the blot on the typhoon imager using the blue fluorescence/chemifluorescence mode (100 microns, PMT between 650-1000 v).

Probe Name	Sequence, 5'-3'	Specificity
uvsX	TTTGGTATTACTCCTGCTTATTTGCGGTCTATGGG	T4
gene44	ATCGTGGTGCTATTGATGATGTTCTTGAGTCTCTC	T4
gene32	CATTGATTTTATCCCAGATTTTCTTACCAAAGCGG	T4
ipl	CTGCTAAAAAGGATGGTGCTACAATCGTTATCTCTCC	T4
gene8	GTTACGCCTAATACAGAATCGGTTGGATAAGGTGG	T4
dda.1	GCAATCTTCCATTCAAAAAATACAGCAGGACGATG	T4
D17z1	TAGAAGCATTCTCAGAAACTACTTTGTGATGATTGCATTC	Human
RPL13A	ATTGAAAAGAAGAGGGGACGAGTAACAGACGAACG	Human
B2M	ACATTGACAGAGTAACATTTTAGCAGGGAAAGAAGAATCC	Human
PGK-1	TACTTTGTTAGGAAGGGTGAGAATAGAATCTTGAGGAACG	Human
HPRT-1	GCAGTATTATGTAGTGGCTGAAAGTGTGAGTTCCG	Human
BACT	AGGTTTTGTCAAGAAAGGGTGTAACGCAACTAAGTCATAG	Human

Table 1: Sequences of human and T4 specific probes used in the DNA blotting and hybridization detection method described herein. Fluorescent probes were labeled with Tye665 (comparable to Cye5) at their 5' ends; biotinylated probes were labeled with biotin at their 5' ends.

In testing the sensitivity of our blotting method, we first blotted a serial dilution of the D17Z1 probe (39-5000 ng) directly onto the membranes and detected them directly to determine the detection limits of the probe alone, in the absence of potentially complicated hybridization conditions. These results are shown in Figure 2

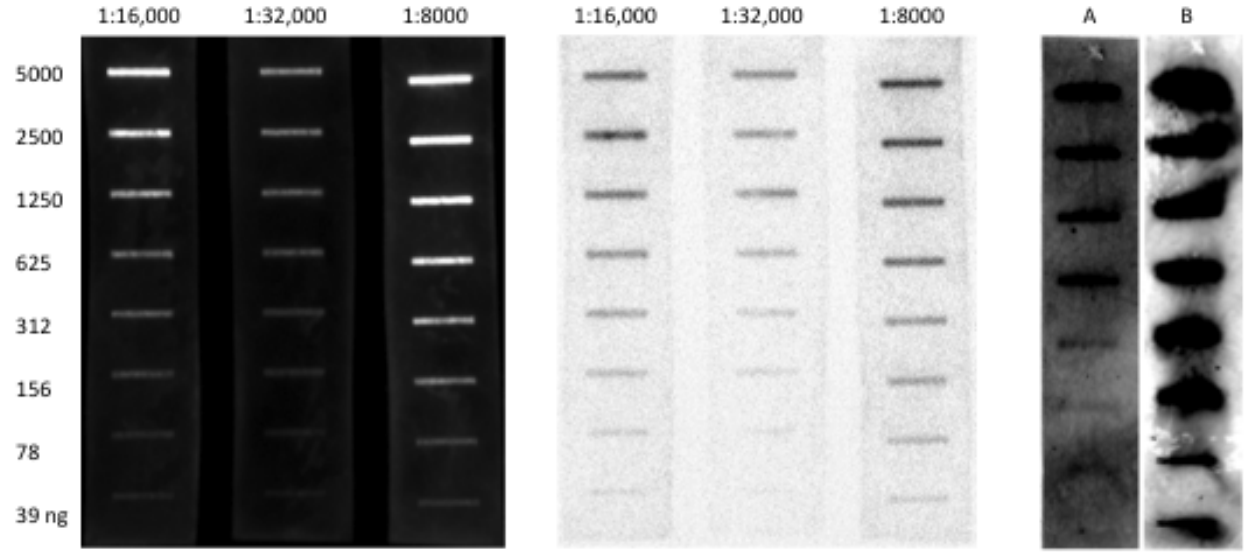

Figure 2: Detection of serial dilution of the biotinylated D17Z1 probe via ECL, on the Fluorchem imager (left) and on the Typhoon phosphorimager (middle). Three different dilutions of horseradish peroxidase-streptavidin were tested: 1:8000, 1:16000, and 1:32000. Detection of the Tye-665 labeled D17Z1 probe on the Typhoon under two different settings (far right). A: 600V, normal, excitation 532, emission 670 bp30, pixel 25; B: 600V, normal, excitation 633, emission 670 bp30, pixel 100.

Next we blotted a serial dilution of 293 DNA onto the membrane and hybridized Tye-665 labeled probes to the membranes to determine the assay sensitivity with a hybridization step. We included all six human probes, as well as all six T4 probes, to verify that the

T4 probes did not cross-hybridize with human DNA. We tested the Tye-665 labeled probes, because detection of the fluorescent probes appeared to be more sensitive than detection of biotinylated probes via ECL. These blots are shown Figure 3

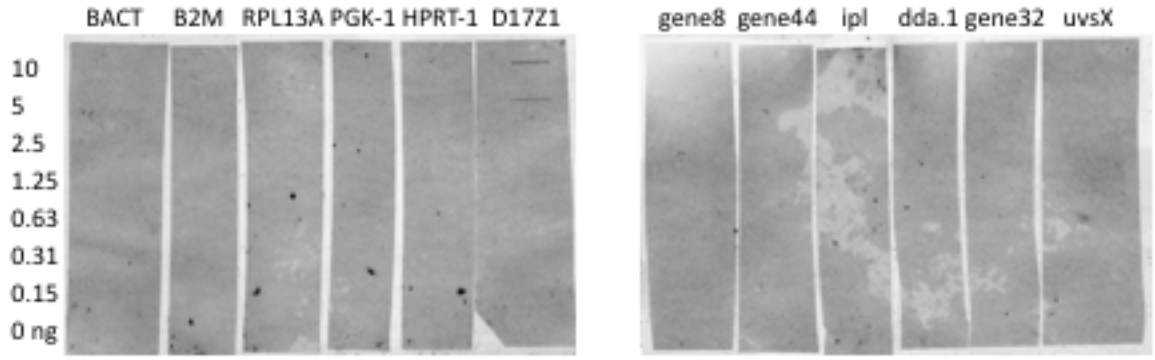

Figure 3: Detection of human 293 DNA serial dilutions with human (left 6) and T4 (right 6) specific Tye-665 labeled probes.

Unfortunately, only one of the probes gave any signal at all, and the smallest amount of human DNA we could detect with this probe (D17Z1) was 5 ng, over 30-fold more than the expected 0.15 ng. This probe has multiple targets to which to hybridize within the D17Z1 locus, so it is not unexpected that none of our single copy gene probes gave no signal at all. Repeating the experiment and adjusting the power and sensitivity parameters on the Typhoon imager did not lead to better sensitivity. Therefore, we opted to probe the blotted DNA with a cocktail of probes. In this experiment, we blotted both human 293 DNA and a lysate of T4 particles onto the membranes, and probed both types of DNA with two different probe mixtures: 1. all 5 human probes, except D17Z1, and 2. all 6 T4 probes. This result is summarized in Figure 4. Using a mixture of probes increased the sensitivity of the assay only slightly, to around 2 ng. Repeating the assay with freshly prepared reagents did not improve the results, so we opted to discontinue optimization of the DNA blotting protocol and explore other methods of nucleic acid detection.

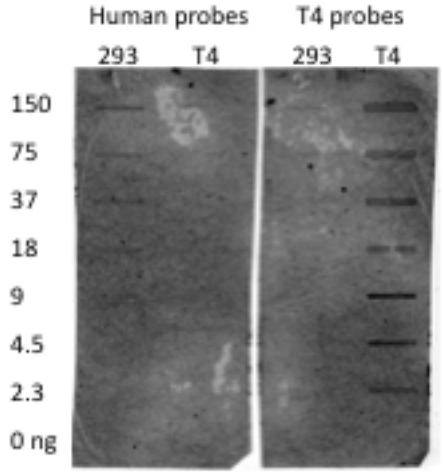

Figure 4: Human 293 and T4 DNA hybridized with a cocktail of human DNA probes or a cocktail of T4 DNA probes.

Quantitative PCR

We sought to use real time quantitative polymerase chain reaction (q-PCR) with Taq-Man style probes as a more sensitive and powerful nucleic acid detection technique. This method relies on a set of PCR primers complementary to the sequence of interest, in addition to a complementary dual-labeled probe dual-labeled probe with a 5' FAM fluorophore and a 3' dark, non-fluorescent quencher. During cycling of a reaction, the primers and probe bind to their complements, and the sequences are extended by the polymerases. After reaching the probe, the 3' exonuclease activity of the polymerase cleaves the quenching group from the probe, resulting in a detectable fluorescent signal. This signal can only come about when the amplicon containing the probe is amplified, eliminating problematic background signals from nonspecific amplification. We designed a series of primer/probe sets specific to human DNA or T4 DNA; two human and two T4 probe/primer sets were purchased from IDT Technologies in the form of PrimeTime mini qPCR assay kits for testing and optimization. These are summarized in Table 2

Probe/Primer Name	Sequence 5'-3'	Specificity
PRB.B2M	/56-FAM/CCT GCC TTG ATC TAC ACC CAT CTG A/3IABlk_FQ/	human
REV.B2M	GTT TCC ACC CCT TCC ATT	
FOR.B2M	CTG GGC AAT GGA ATG AGA	
PRB.GAPD	/56-FAM/AGC TGA GTC ATG GGT AGT TGG AAA AGG /3IABlk_FQ/	human
REV.GAPD	CGG TGA CAT TTA CAG CCT	
FOR.GAPD	GAA GTG GGT TTA TGG AGG TC	
PRB.GENE12	/56-FAM/AAA GAC CAC GCA TGT CAG GCA ATC /3IABlk_FQ/	T4
REV.GENE12	TGA GAA CCA CGA CCA GAA	
FOR.GENE12	GGC GGA AAC CCA TCA AAT	
PRB.NRDG	/56-FAM/TGG ATA TAC GAT TTG CTT AAA TGG GAC GC/3IABlk_FQ/	T4
REV.NRDG	GGA TGT AGG GTC GTT CTC TT	
FOR.NRDG	GGT TCT GTG GAT AAA GTG GG	

Table 2: Sequences of human and T4 specific primer/probe sets used for q-PCR studies. Probes were labeled with a 5' FAM fluorophore and a 3' non-fluorescent quencher Iowa Black.

The instrument we used for these studies was the DNA Engine Opticon 2 system from Bio-Rad. Initial attempts to analyze 100 ng of human DNA with the B2M probe/primer set using the iQ Supermix reagents from Bio-Rad failed, and we were not able to identify any successful cycling conditions for these samples with this reagent. We next obtained the FailSafe Probes real-time PCR kit from Epicentre, with the Probe 4 PCR Premix, and followed the accompanying protocol. We tested twelve different annealing temperatures across a gradient of 52-62°C with this new reagent, in combination with all four probe/primer sets, and were finally able to see a significant increase in fluorescence over 50 cycles of PCR for some samples (Figure 5). The reaction cycling conditions used are as follows: 1. 95°C, 2 minutes; 2. 95°C, 30 seconds; 3. 52-62°C gradient across 12 wells, 30 seconds; 4. 68°C, 30 seconds; 5. read fluorescence; 6. Go to step 2, repeat 49 times; 7. hold temperature at 4°C. We did not use a wax or oil

overlay, and used the heated lid instead, to prevent condensation of the reaction mix on the lids of the tubes. The probe/primer sets which worked properly were GAPD (human specific) and gene12 (T4 specific), and the optimal annealing temperature ranged from 52.3-54.8°C. The blue GAPD curve in Figure 5 with a much higher signal than all the other curves was an outlier caused by pipetting error.

Because the Epicentre reagents worked with a subset of our primer/probes, we sought to re-test all four sets, with the FailSafe Probes real-time PCR optimization kit. This kit provides reagents to test a series of defined buffer conditions simultaneously, allowing the user to quickly select the best conditions for a particular template and primer/probe pair. We had hoped to validate the B2M and nrdG primer/probe sets using this kit, but as shown in Figure 6, we still did not observe any signal amplification with these reagents and any of the buffer mixes supplied with the optimization kit. The cycling conditions were as described above, but with an annealing temperature of 53°C for all samples. These primer/probes will need to be re-designed before they can be used in a quantitation assay. Other sets may be used with the buffer 4 set described in Figure 5.

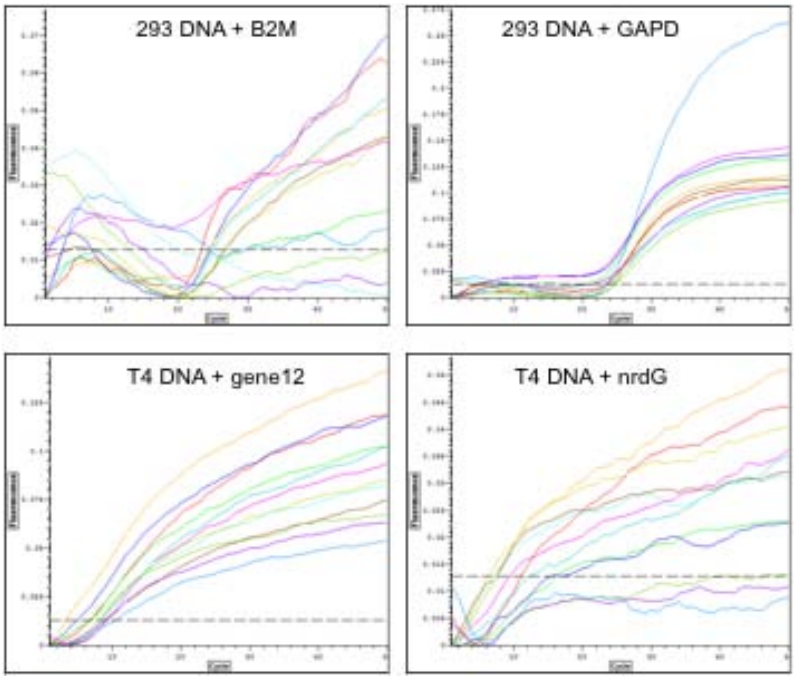

Figure 5: Fluorescence vs. cycle number plots for q-PCR reactions with Epicentre FailSafe Probes reagents and human or T4 DNA and corresponding primer/probe sets. Each plot contains 12 curves, representing a different annealing temperature for each of the 12 reactions.

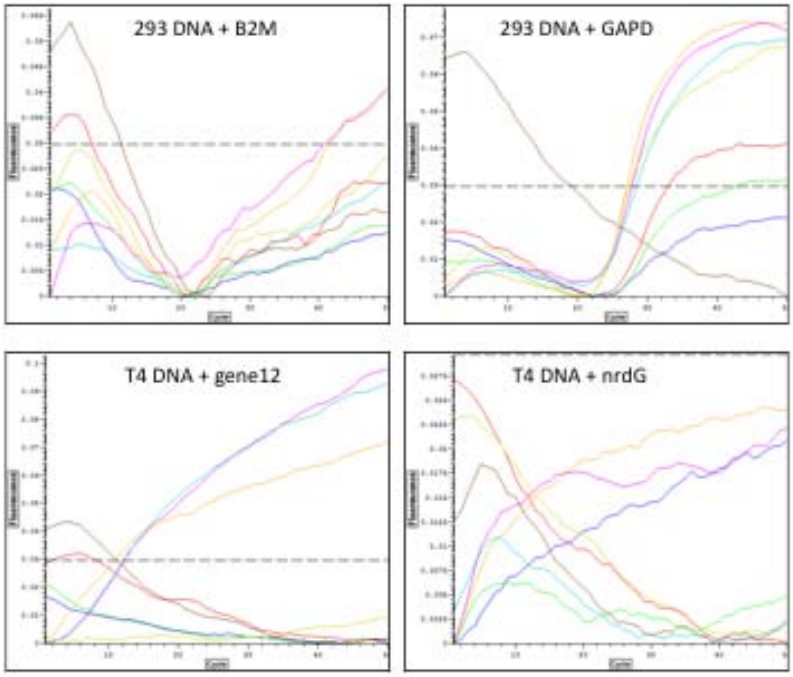

Figure 6: Fluorescence vs. cycle number plots for q-PCR reactions with Epicentre FailSafe Probes reagents and human or T4 DNA and corresponding primer/probe sets. Each plot contains 12 curves, representing a different buffer mix for each of the 12 reactions.

DNA Normalization

We devised a DNA normalization strategy for suppressing abundant host nucleic acids in a mixed sample of host and pathogen, thereby enriching the sample for pathogen nucleic acids. We were not able to test the method yet, but a summary is outlined in Figure 8. Basically, a mixture of host and pathogen nucleic acids is digested with restriction enzymes into fragments averaging about 1.5 kb in size. A sequence analysis of the T4 bacteriophage genome and human chromosome 20 demonstrates that the restriction enzymes SspI or DraI would be good candidates for this process, leading to the desired fragment sizes (Figure 7). This mixture is then denatured by incubation at a high temperature, and then the DNA is allowed to reassociate. The kinetics of this hybridization of single stranded DNA to its complement leads to an enrichment of pathogen DNA in the single stranded fraction, and an abundance of host sequences in the double stranded fraction. The single stranded fraction can be isolated by degrading the double stranded fraction with a commercially available duplex-specific nuclease. Then, if pathogens are known, as in a mock clinical sample, we can use quantitative real time PCR with host and pathogen specific primer/probe sets to determine the level of pathogen enrichment in the sample. We anticipate this method to be tested in other projects currently ongoing at Sandia.

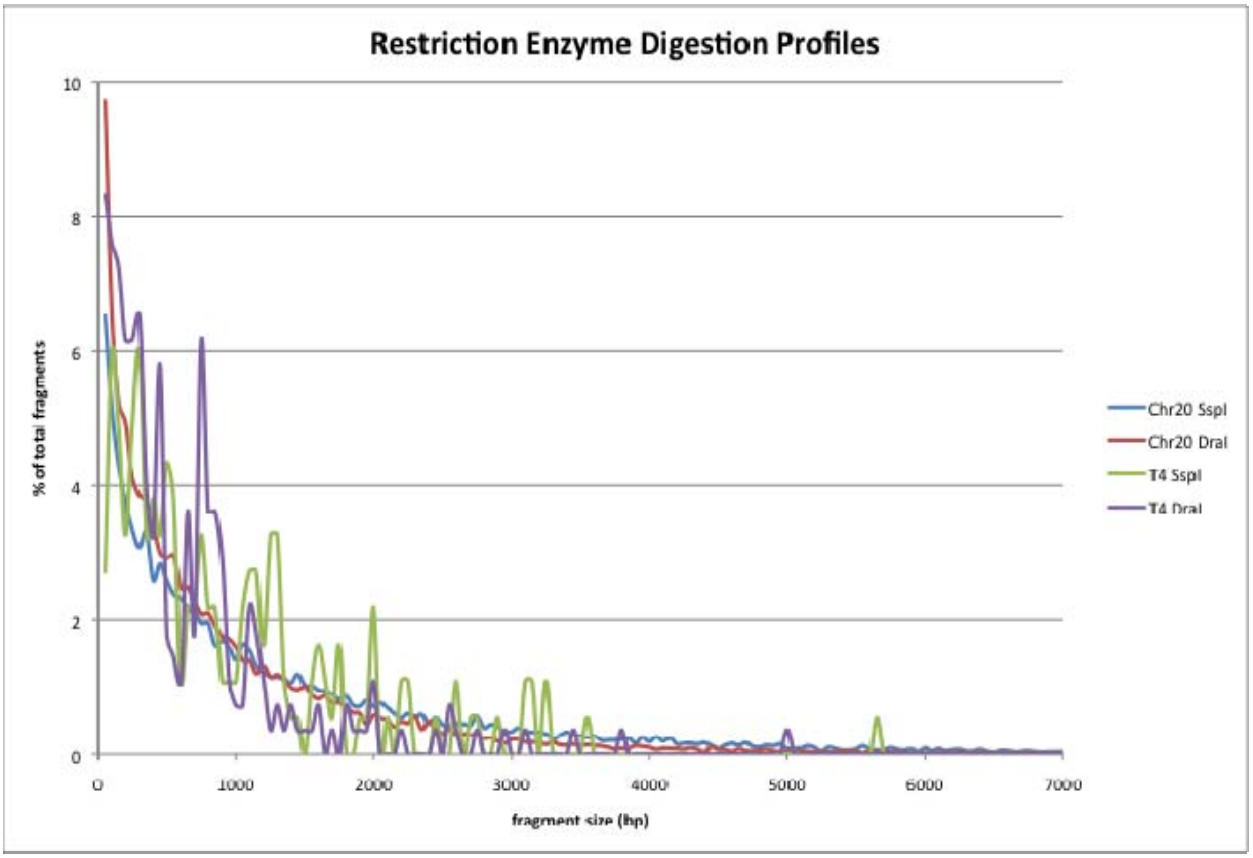

Figure 7: Restriction digest profiles for human chromosome 20 and bacteriophage T4 (complete genome), showing fragment sizes as a percent of total fragments after digestion with either enzyme.

Figure 8: Overview of DNA normalization strategy for enriching mixed samples for less abundant pathogen nucleic acids.

2. Bioinformatics

Overview

Current commercial UHTS sequencers produce a much higher throughput at a much lower cost than traditional sequencing techniques. The output varies according to read length (35-300 nt), number of reads/coverage, error class, and base quality/fidelity scoring techniques. Table 1 lists the size of genomic data generated by popular sequencers.

Table 1: Output Specification for Various Sequencers

Sequencer	Read Length	Time / Run	# Reads or Tags / Run	Output / Day	Base Calls with Q>30	Accuracy / Base	% Perfect Reads
Illumina	1x35	2.5d	138-168m	1.8-2.4 GB	70-85%	>99%	>90%
	2x35	5.5d		1.9-2.3 GB	70-85%	>99%	>90%
	2x50	6.5d		2.0-2.5 GB	70-85%	>98.5%	>80%
	2x75	9.5d		2.1-2.6 GB	>70%	>98.5%	>70%
454 Life Sciences	1x250	10h	??	0.4-0.6 GB		99.5%	
	2x250	7.5h	??	0.1 GB		99.5%	
Applied Biosystems	1x50	6-7d	200-300m	10-15 GB		99.94%	
	2x50	12-14d	400-600m	20-30 GB		99.94%	

Table 3

The critical informatics step in processing UHTS data is the large-scale text searching for possibly ambiguous patterns. This problem spawns both hardware and software challenges for data analysis.

From a hardware perspective, the problem is categorized as “embarrassingly parallel” because the overall job can be partitioned and fed independently into many processors with little communication necessary between the processors for any temporary results. The computation itself is frequently integer based and input/output intensive. Such problems typically do not require supercomputer level infrastructure. Instead, the huge amount of UHTS data generated in *each* sequencing run necessitates enhanced

investment in and access optimization of random access memory and data storage disks.

From a software perspective, the problem is that of accurate and efficient fragment assembly followed by downstream data annotation and analysis. Fragment assembly is the inference of a consensus genomic sequence from short reads that are the output of current UHTS sequencers. Typically, this is achieved by the overlap-layout-consensus paradigm: identify overlapping regions in all reads, place the reads such that the overlaps align, and ascertain from this layout the most common base at each position in the overall sequence. During assembly, the reads are either mapped on to a reference sequence, if available, or assembled *de novo* otherwise, with the particular circumstance being reflected in the choice of specific assembly algorithms.

In practice, anticipating and attenuating the effects of sequencing errors and other complications is crucial for successful UHTS data analysis. There can be four different types of complications: 1) sequencing errors including: polymorphisms (insertions, deletions, substitutions, and transpositions), chimeric fragments, and contamination by foreign sequence; 2) unknown orientation of reads that makes it impossible to tell which DNA strand they belong to; 3) repeated regions that may appear two or more times in the target sequence; and 4) insufficient coverage due to random sampling of the original sequence that leads to formation of contigs.

Achievements

The key R&D goals of the project were to 1) identify candidate bioinformatics tools, 2) construct a bioinformatics pipeline, and 3) demonstrate and refine the effectiveness of the pipeline using data typical of UHTS.

Identification of candidate bioinformatics tools

Rapid and accurate mapping of short reads to a reference genome and contig assembly are important components of UHTS data processing. Typically the data management requirements and approach for the reference genome are different from those for the short reads. In both cases, however, the sequences are usually indexed and hashed before analysis.

Indexing a reference genome such as the human genome is memory expensive: the ~3 GB human genome may need ~15 GB for storing reference sequences and index tables. A number of programs (such as MUMmer {Delcher et al, 2002}, SOAP {Li et al, 2008a}, and Exonerate {Slater and Birney, 2005}) use a variety of ways to handle reference genome data.

Standard BLAST-like algorithms are not designed to handle short read data and associated higher errors per base. Given a set of reference genomes and a set of short

reads, most aligners a) generate a memory-resident indexing data structure for either one of the input sets, which then facilitates fast random access to that set; and b) scan the other input set against the indexed set, produce alignment seeds, extend the seeds with BLAST-like algorithms, and finally join adjacent extended seeds using Smith-Waterman type dynamic programming after considering read errors, SNPs, and/or indels. Others like Maq {Li et al, 2008b}, ELAND (Solexa), RMAP {Smith et al, 2008}, Qpalma {Bona et al, 2008}, and Slider {Malhis et al, 2009}, also consider the base quality/fidelity scores to improve accuracy and hash tables for better efficiency. For instance, SOAP converts the short reads and reference genome to a numeric data type using two bits per base coding. It loads the reference genome into memory, allows for two mismatches or 1-3 base pair continuous gap, and is 300 (gapped) to 1200 (ungapped) times faster than BLAST. Other innovative uses of existing graph theoretical or data compression algorithms for genomic data analysis include using de Bruijn graphs (e.g., Euler {Pevzner et al, 2001} and Velvet {Zerbino and Birney, 2008}) or Burrows-Wheeler indexing (e.g., Bowtie {Langmead et al, 2009 } and BWA {Li and Durbin, 2009}).

In this project, several open source algorithms were evaluated for short reads alignment and assembly. Three of these were identified for pipeline development: Maq, Bowtie, and Velvet.

a) Maq (mapping and assembly with quality) is a widely used short reads alignment program {Li et al, 2008b}. It was developed for Illumina's sequencing data but is now also capable of handling Applied Biosystems' SOLiD data. The algorithm first aligns short reads to a reference sequence and then determines the consensus. At the mapping stage, it performs ungapped alignment. For single-end reads, it finds all hits with up to 2 or 3 mismatches while for paired-end reads, it finds all paired hits with one of the two reads containing up to 1 mismatch. At the assembly stage, Maq finds the consensus sequence based on a statistical model. It also assigns a mapping quality (MQ) to each read alignment that considers: the repeat structure of the reference; the base quality of the read; the sensitivity of the alignment algorithm; and single or paired-end read type. Conventionally, a MQ of 30 or above indicates that the overall base quality of the read is good and that the best alignment for that read has few mismatches.

b) Bowtie is an ultrafast, memory-efficient aligner that is designed to align short reads to a reference genome {Langmead et al, 2009}. It can align 35 bp reads to the human genome at the rate of over 25 million reads per hour on a typical workstation and achieves greater speed with multiple processor cores. It keeps its memory footprint small by indexing the reference genome with a Burrows-Wheeler index (typically about 2.2 GB for the human genome; 2.9 GB for paired-end data). It is designed to be extremely fast for data sets many reads have at least one good alignment to the reference, are of high quality, and the number of alignments reported per read is small. Bowtie does not currently support SOLiD data.

c) Velvet is a de novo genomic assembler that was specifically designed for assembly "from scratch" of Illumina or 454 Life Sciences short read data in the absence of a reference sequence {Zerbino and Birney, 2008}. In graph theoretical terms, the overlap-layout-consensus approach represents each read as a node and each overlap as an edge. The use of de Bruijn graphs for representing short read data presents a different approach. In this representation, the elements are not organized around reads but rather around word of *k* nucleotides or *k*-mers. Reads are mapped as paths through the graph and proceed from one word to the next in a predetermined order. Since the basic data structure in the de Bruijn graph is based on *k*-mers not reads, high redundancy can be handled without affecting the number of nodes. This approach is attractive for identifying repeats and avoiding mis-assembly errors common to the overlap-layout-consensus approach. Velvet consists of a set of algorithms that manipulate de Bruijn graphs for sequence assembly. It performs various error corrections on the structure of the de Bruijn graphs such as removing at a node fragments that are disconnected on one end; removing "bubbles" or loops with the Tour Bus algorithm; and removing faulty connections.

Besides these three programs, a few others were also identified that are newer and include metrics and file formats that may become the standard in this field. These include the Burrows Wheeler Alignment tool (BWA) {Li and Durbin, 2009} and SAMtools {Li et al, 2009} which are a set of utilities for manipulating alignments in the SAM format including sorting, merging, indexing, and generating alignments in a per-position format. The SAM (or Sequence Alignment/Map) format is a generic format for storing large sequence alignments and aims to be easily reproducible and interconvertible.

Construction of bioinformatics pipelines

(i) Overlap-layout-consensus assembly (where a reference sequence is available). For this approach, Maq is one of the most widely used open source aligners and is accompanied by a set of utilities for format conversion, and for constructing, viewing, and merging alignments. It converts reference sequences and short reads to binary formats to speed up file input/output and to reduce disk space usage. The output is also stored in compressed binary format with human readable information only extracted when necessary. Although it can be used by itself to perform alignments, in cases where huge amounts of data are involved (or if the short reads are greater than 63 bp), it is more efficient to use the ultrafast Bowtie (which can support read lengths up to 1024 bp) for the initial alignment. The output of Bowtie can be easily converted to Maq format which enables the use of Maq's analysis utilities. Figure 9 shows the combined Bowtie/Maq pipeline with associated file extensions at various points. The consensus (.cns) file allows the extraction of SNP information with the output including at each position the reference base, the consensus base, phred-like consensus quality, read depth, average coverage, consensus quality, and the second and third best base calls.

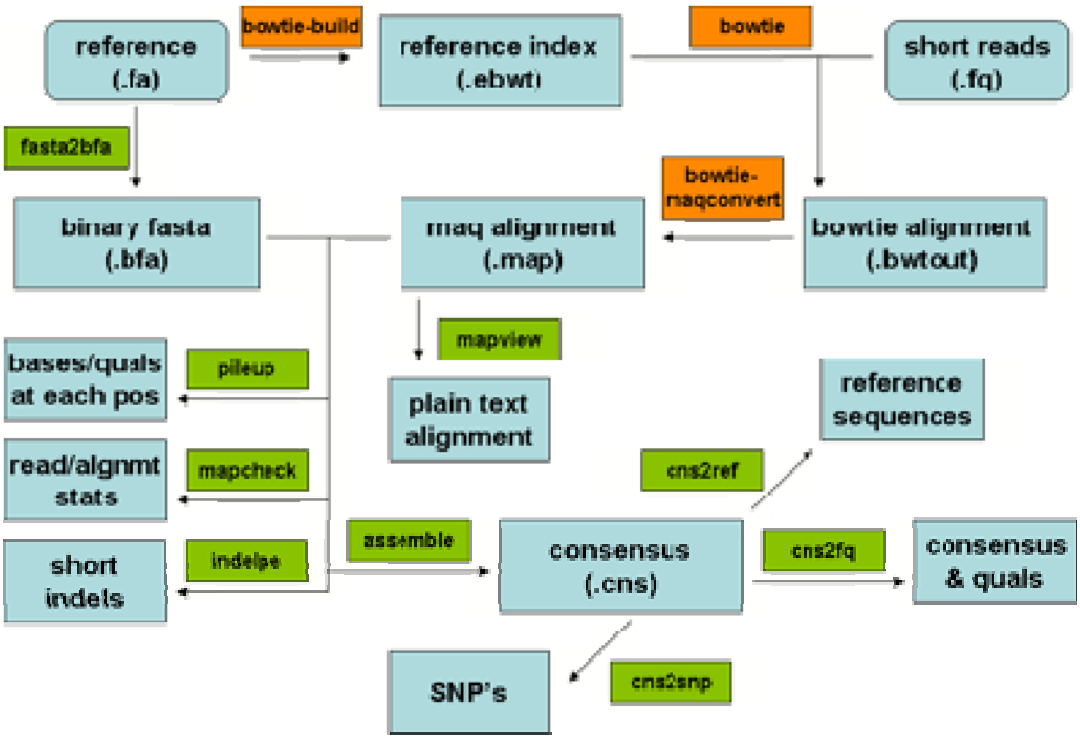

Figure 9 The Bowtie/Maq pipeline for sequence assembly and related analyses

.

(ii) *De novo* assembly (where a reference sequence is unavailable). The Velvet suite was selected and evaluated for *de novo* assembly problems. Among its main programs, *velveth* construct the input dataset that is used subsequently by *velvetg* and indicates what each sequence file represents. It takes in several sequence files, produces a hashtable, and outputs files necessary for *velvetg*. A hash length (or *k*-mer length) corresponding to the size of the words being hashed has to be supplied to *velveth*. The hash length must be an odd number (to avoid palindromes), must be less than 32 (because it is stored on 64 bits), and must be less than the read length (to allow overlap observation). The choice of the hash length is an important step in the program flow: longer *k*-mers lead to better overlap but reduce coverage. The core of the program suite is *velvetg*, which builds and manipulates the de Bruijn graphs. Several important parameters may be specified as input. These include the coverage cutoff (the threshold below which primarily short, low coverage nodes are likely to be found) and the desired minimum contig length. In the presence of mated pair information, it may be possible to use resolve ambiguous contigs even for a low coverage cutoff.

Demonstration of pipelines using UHTS data

Real and simulated UHTS data was used for *in silico* experiments that have demonstrated the bioinformatics pipelines developed; helped identify technical

challenges in UHTS data analysis; and led to development of capability for short reads analysis and contig assembly. The experiments included assembling 20 million paired-end bacterial short reads and identifying an unknown viral pathogen in a simulated mixture of viral and human short reads. The bioinformatics platform developed in this project can be suitably modified to answer metagenomic questions in more complex microbiome related studies.

(i) Short reads assembly of *E. coli* using Bowtie/Maq pipeline. Data available from NCBI's Short Reads Archive was used to assemble 20 million Illumina paired-end reads (with 200 bp inserts) of *E. coli* K-12 MG1655 (accession number NC_000913.2; genomic size 4.6 million bases). *bowtie-build* was used to index the bacterial reference genomic sequence (NC_000913.2). A multithreaded (on eight processors) *bowtie* command was used using paired-end Illumina-specific parameters to align the reads. The output alignment was converted to Maq format using *bowtie-maqconvert*. The reference sequence was also converted to Maq's binary format to allow faster processing. The results showed that reads covered 99.8% of the non-gapped regions in the reference sequence. Six possible SNP sites were indicated. Approximately 93.9% of the reads had MQ >= 70 (i.e., 1 in every 10 million reads could align incorrectly) while about 98.9% of the reads had MQ >= 50 (i.e., 1 in every 100,000 reads could align incorrectly). In addition, 86.2% of the reads had no mismatches, 96.0% of the reads had up to 1 mismatch, and 98.5% of the reads had up to 2 mismatches.

(ii) Unknown pathogen detection using Velvet. In this experiment, short reads were simulated using the Illumina error model for both the host and the pathogen sequences. The host sequence was a 500 kb fragment of human chromosome 21 while the pathogen sequence was the HSV1 viral genome. The MetaSim program {Richter et al, 2008} was used to simulate 36 bp short reads at 36x coverage (i.e., the number of reads generated was equal to the number of bases in the original sequences). The substitution rate in the total number of processed bases was ~1.6%. Bowtie was used to build the reference genome index and to align against it the mixture of simulated host and pathogen reads. All reads that did not align to the host (~1.6% host reads and 100% pathogen reads) were subjected to *de novo* assembly using Velvet. The parameters used are displayed in Figure 10, which provides the scheme for this simulation. These included hash length of 21, coverage cutoff of 5.0, and minimum contig length of 100 bp. The result included 26 contigs with length >100 bp. The three largest contigs were 47 kb, 35 kb, and 8 kb long. A BLAST search against NCBI's non-redundant nucleotide collection correctly identified the Velvet contigs as fragments of HSV1 with the accession NC_001806.

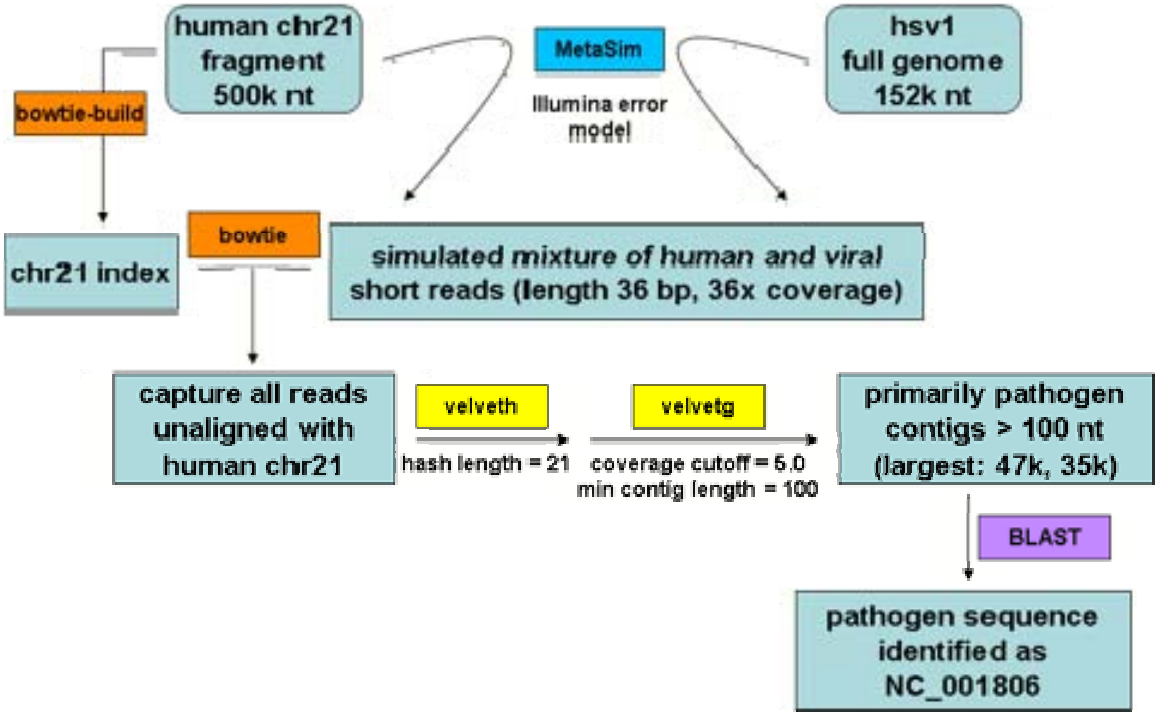

Figure 10 “Unknown” pathogen identification using *de novo* assembly.

References

- Bona et al, (2008) “Optimal spliced alignments of short sequence reads”, *Bioinformatics*, **24**:i174.
- Delcher et al, (2002) “Fast algorithms for large-scale genome alignment and comparison”, *Nucleic Acids Res.*, **30**:2478.
- Gentleman et al, (2004) “Bioconductor: Open software development for computational biology and bioinformatics”, *Genome Biology*, **5**:R80.
- Langmead et al, (2009) “Ultrafast and memory-efficient alignment of short DNA sequences to the human genome”, *Genome Biology*, **10**:R25.
- Li and Durbin, (2009) “Fast and accurate short read alignment with Burrows-Wheeler transform”, *Bioinformatics*, **25**:1754.
- Li et al, (2008b) “Mapping short DNA sequencing reads and calling variants using mapping quality scores”, *Genome Research*, **18**:1851.
- Li et al, (2008a) “SOAP: Short oligonucleotide alignment program”, *Bioinformatics*, **24**:713.
- Li et al, (2009) “The Sequence alignment/map (SAM) format and SAMtools”, *Bioinformatics*, **25**:2078.
- Malhis et al, (2009) “Slider—maximum use of probability information for alignment of short sequence reads and SNP detection”, *Bioinformatics*, **25**:6.
- Pevzner et al, (2001) “An Eulerian path approach to DNA fragment assembly”, *Proc. Natl. Acad. Sci. U.S.A.*, **98**:9748.

- Richter et al, (2008) “MetaSim – Sequencing Simulator for Genomics and Metagenomics”, *PLoS ONE*, **3**:e3373.
- Slater and Birney, (2005) “Automated generation of heuristics for biological sequence comparison”, *BMC Bioinformatics*, **6**:31.
- Smith et al, (2008) “Using quality scores and longer reads improves accuracy of Solexa read mapping”, *BMC Bioinformatics*, **9**:128.
- Zerbino and Birney, (2008) “Velvet: Algorithms for de novo short read assembly using de Bruijn graphs”, *Genome Research*, **18**:821.

3. Microfluidics

Abstract

The feasibility of performing DNA manipulations using a microfluidic platform was tested. A thin, nanoporous, negatively charged membrane was photopolymerized within a microchannel, serving as a "filter" or "trap" against which DNA could be trapped, *e.g.* for enzymatic or hybridization reactions. The DNA trap could be seamlessly coupled to a microchannel for electrophoretic sieving for online analysis of product size.

Introduction

Preparing DNA samples for ultra-high throughput sequencing, including suppression of host-derived background, involves a series of molecular manipulations of DNA, including a series of enzymatic reactions (ligations, endo- and exonuclease digestions, *etc*.) as well as hybridizations, with intermediate steps of separation and purification. Our aim was to demonstrate that these steps could be automated and accelerated, to both simplify and speed up the rather laborious and lengthy bench-scale protocol.

To this end, we have developed a microfluidic platform capable of concentrating both a DNA sample and enzyme in a small (~nL) volume, coupled with a separation channel. Initial development has focused on characterizing the efficiency of DNA concentration, along with optimizing the efficiency of the downstream electrophoretic analysis.

Materials and Methods

Microfluidic chips consisting of 30-µm deep channels etched in glass substrates were purchased from Caliper. Microchannels were pretreated with 1M NaOH for 10 minutes, followed by rinsing with DI water and methanol. The channel surface was then derivatized with methacrylate groups by flushing the channels with a 2:3:5 (v:v:v) mixture of 3-methacryloxypropyltrimethoxysilane, glacial acetic acid, and water for 30 minutes. The channels were rinsed with 30% acetic acid, water, and methanol, and filled with a membrane precursor solution consisting of acrylamide and bisacrylamide (45%T, 12%C) with 100 mM acrylic acid and 5 mg/mL VA-086 photoinitiator. A membrane (~50 µm thick) was fabricated in the microchannel using a 355-nm laser (frequency-tripled Nd:YAG), projected through a slit and a cylindrical lens. The monomer solution was flushed from the channels, and replaced with a 5% acrylamide solution with 5 mg/mL VA-086, and exposed to 365-nm UV for 10 minutes, resulting in a wall coating of linear polyacrylamide throughout the remainder of the device. The channels were flushed with water, and the device was stored at 4 °C until further use.

The microfluidic chip was mounted in a Delrin manifold. The separation channel portion of the device was filled with a polymer sieving matrix (1.2 wt% hydroxyethylcellulose, M_w

~ 720,000) in 89 mM Tris, 89 mM borate (1X TB) buffer. The remainder of the device was filled with 1X TB buffer with no polymer. DNA sample was added at 1:50 dilution in the sample reservoir of the device. Two DNA samples were tested: (1) a series of PCR products (500, 1000, and 1500 bp in length), amplified from a cloning vector, using forward and/or reverse primers labeled at the 5' end with TAMRA, and (2) A Genescan size standard from Applied Biosystems (essentially a restriction digest modified with a TAMRA-labeled oligonucleotide).

For imaging experiments, the chip was mounted on the stage of an inverted microscope, using epifluorescence illumination (mercury lamp, 10X objective, Chroma filter set 31008 (Cy3/TMR), with images captured by a CoolSnap HQ CCD camera. For separation experiments, the chip was mounted in a laser-induced fluorescence detection setup consisting of a 532-nm laser, an optical chopper, a 560 long-pass dichroic reflector, a 40X achromat objective lens, a 500-µm pinhole, a 570-610 nm bandpass emission filter, and a photomultiplier module (Hamamatsu 5384-20), with the signal demodulated using a lock-in amplifier, and digitized using a National Instruments DAQpad. The chip was positioned such that the laser focus was located at the centerline of the separation channel, at a distance ranging from 5-14 mm downstream of the injection point.

DNA was introduced to the membrane by applying a bias of 100-1000V between the DNA sample well and the membrane. After concentrating DNA against the membrane for anywhere from 1-10 minutes, the field was switched to apply a bias from a well containing buffer only for 1-2 minutes, electrophoretically flushing any DNA remaining in the channels against the membrane. Finally, the field was again switched to flush the DNA away from the membrane, and down the separation channel.

Results and Discussion

The charged membrane was found to be impermeable to all sizes of DNA tested, at voltages across the membrane up to 500V, including both the 500-1500 bp dsDNA PCR products, as well as the entire range of DNA sizes in the Genescan digest (~55-14000 bp, plus an additional product that may be a 20mer oligonucleotide used in production of the standard). In separate experiments, the charged membrane has also been found to be impermeable to DNA oligos in the 18-25 base size range. The inclusion of negatively charged groups (*e.g.* acrylic acid) appears to be critical for exclusion of DNA, as tests performed with the Genescan digest using a neutral membrane indicated both decreased signal intensity relative to a charged membrane, and a bias toward larger fragments, indicating that smaller fragments were passed by the membrane.

Preconcentration of the 1500 bp PCR product was observed by imaging during electrophoresis, with integration of the fluorescent signal in a rectangular region of interest adjacent to the membrane. No DNA was observed to enter or pass the membrane. The rate of preconcentration of the 1500 bp PCR product was found to be linear over the course of a 5-minute preconcentration plus subsequent buffer flush, as

shown in Figure 11. The linearity indicates that no fundamental limit of preconcentration has been reached during this time. At higher fields (*e.g.* 300-500 V) the preconcentration current was found to drop somewhat during the first minute of preconcentration, indicating some tendency for concentration polarization of counterions near the membrane; this effect indicates that there is no advantage to very high preconcentration voltages.

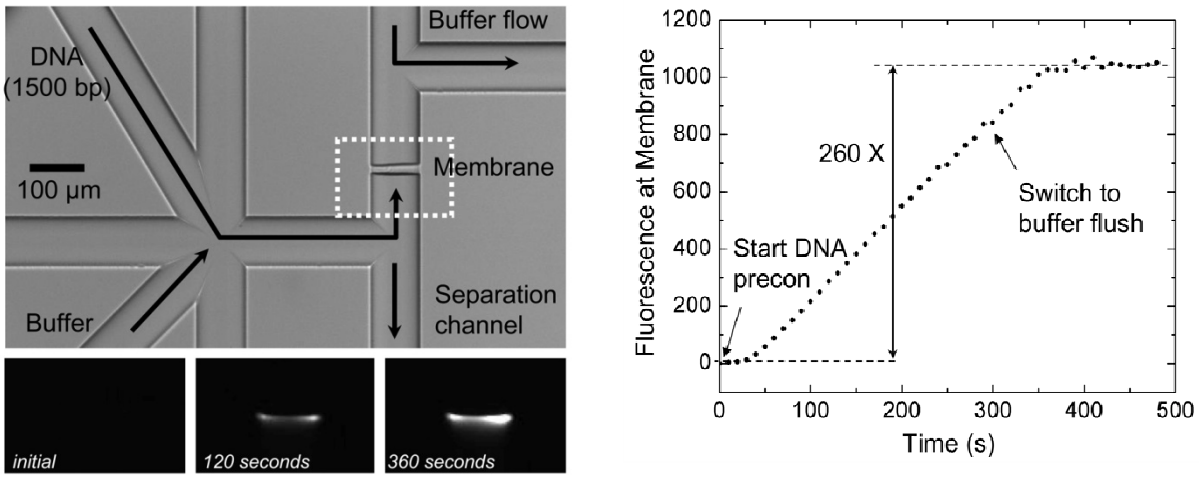

Figure 11 Electrophoretic concentration of DNA at a photopatterned polyacrylamide/acrylic acid membrane. A field of ~30 V/cm was used to concentrate the dye-labeled fragment at the membrane for 300 seconds, followed by an electrophoretic buffer flush. The white rectangle indicates the area of detail for fluorescence micrographs (bottom images). The total volume of the concentrated band is approximately 0.5 nL.

A representative separation of the Genescan digest following preconcentration at the membrane is shown in Figure 12 below. Approximately 20 of the 27 peaks in the digest are either partially or fully resolved. Although not a complete separation (as can be obtained by separation in a 30-50 cm long capillary in 20-30 minutes), the result is impressive for a 14 mm separation length, and less than one minute of separation time. It is likely that performance could be improved by optimizing the polymer sieving matrix (either a different polymer or including polymer of different molecular weight), and by using a microchannel with a somewhat longer separation difference (increasing the separation distance from 10 to 14 mm substantially improved the quality of the separation; a minimum separation distance is always necessary to overcome the initial width of the injected plug).

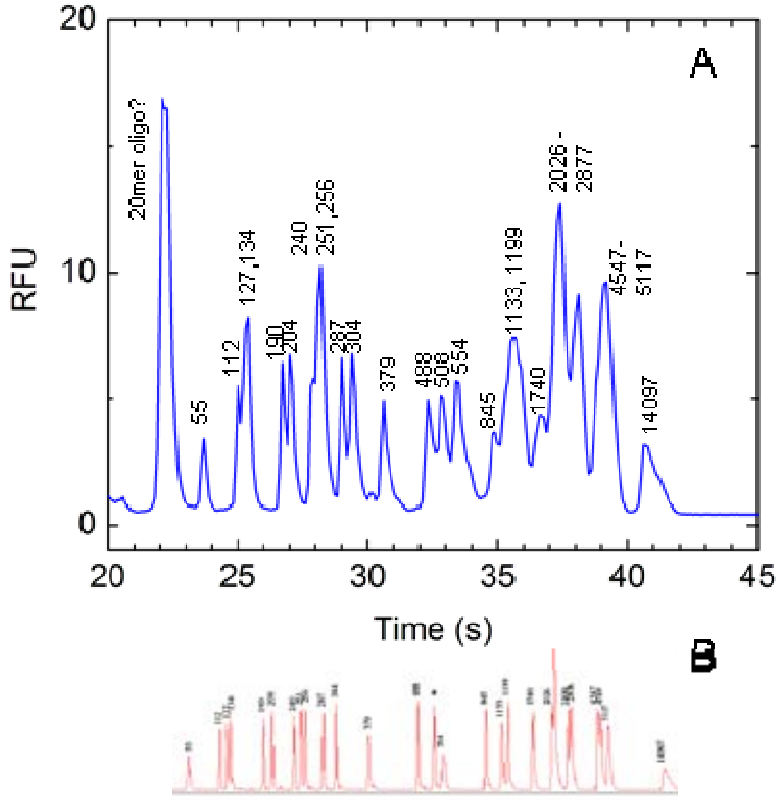

Figure 12 (A) Chip separation of Genescan 2500-TAMRA size standard (a modified λ-phage PstI digest labeled with TMR). The separation length = 14 mm. Buffer = 1X TB, 1.2 wt% HEC Mw = 720 kd. Detection by LIF (532 nm excitation, 560 DRLP dichroic, 570-610 nm BP emission filter). Preconcentration for 120s at 100V (~60x precon based on Figure 11). Separation field strength is ~300 V/cm. (B) Separation of Genescan digest in a capillary (from ABI Genescan manual).

Conclusion

The tests with DNA preconcentration have demonstrated that a negatively charged membrane effectively excludes DNA, and allows >100X concentration of a dilute DNA sample within a few minutes. The preconcentration membrane can be coupled to a separation channel, allowing online electrophoretic analysis of products from an enzymatic reaction. The preconcentration membranes have previously been demonstrated to be useful for concentration of proteins as well, and further testing is planned to determine whether concentrating DNA and enzyme together within the same nanoliter volume, along with electrophoretic “mixing” is useful for accelerating DNA manipulations.

Distribution

1	MS9292	Todd Lane,	8621 (Electronic Copy)
1	MS9671	Julia Kaiser,	8621 (Electronic Copy)
1	MS9292	Joseph Schoeniger,	8621 (Electronic Copy)
1	MS9292	Robert Meagher,	8621 (Electronic Copy)
1	MS9292	Kamlesh Patel	8621 (Electronic Copy)
1	MS1322	Milind Misra,	1435 (Electronic Copy)
1	MS9291	Anup Singh,	8621 (Electronic Copy)
1	MS9291	Malin Young,	8620 (Electronic Copy)
1	MS9405	Glenn Kubiak,	8600 (Electronic Copy)
1	MS9004	Duane Lindner	8120 Electronic Copy)
1	MS0899	Technical Library,	4536 (Electronic Copy)
2	MS9018	Central Technical Files,	8944
2	MS0899	Technical Library,	9536(Electronic Copy)
1	MS0123	D. Chavez, LDRD Office,	1011