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

Sandia National Laboratories
Todd Lane, Steve Branda, Bryan Carson, Julie Kaiser,
Robert Meagher, Milind Misra, Ken Patel

Problem

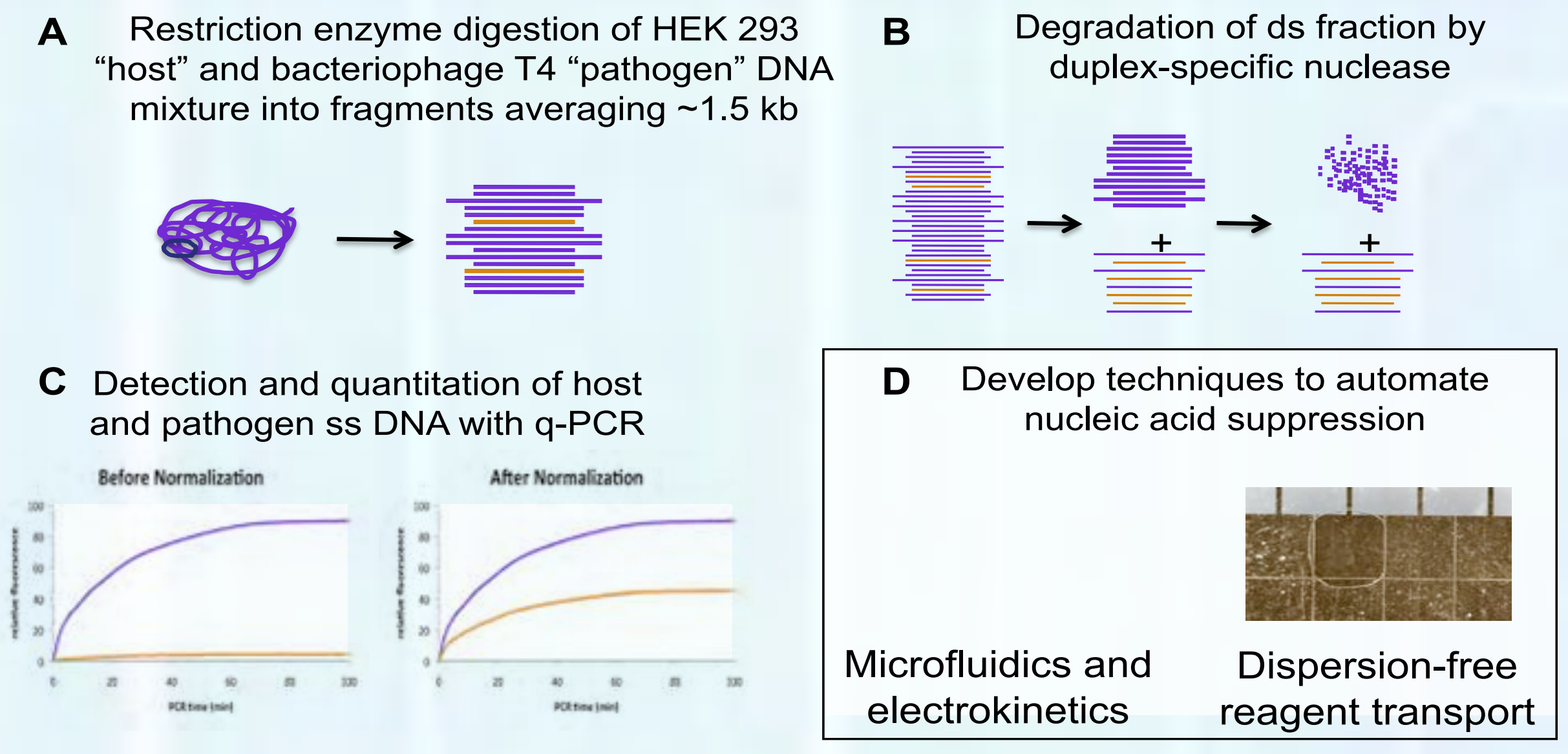
- Our nation's biodefense and public health infrastructure are geared toward detecting threats from known pathogenic agents.
- Advances in biotechnology as well as global travel networks make it ever more likely that we will face a threat from an unknown pathogen.
  - An engineered pathogen designed specifically to elude detection by conventional means is a particularly grave threat.
- Modern ultrahigh throughput sequencing (UHTS) techniques allow analysis of the pathogens at the whole genome level, without prior knowledge of protein markers or genetic signatures
- However, a novel pathogen might be present at very low levels, with a very high background of human DNA.
  - Not just a "needle in a haystack" problem – the "needles" and "hay" are made of chemically identical building blocks.
  - Requires sophisticated sequence analysis (bioinformatics) to sort host, non-host background, and pathogen sequences.

Approach

Development of nucleic acid normalization

Objective: Develop technique for selective destruction of host-derived DNA in presence of pathogen DNA.

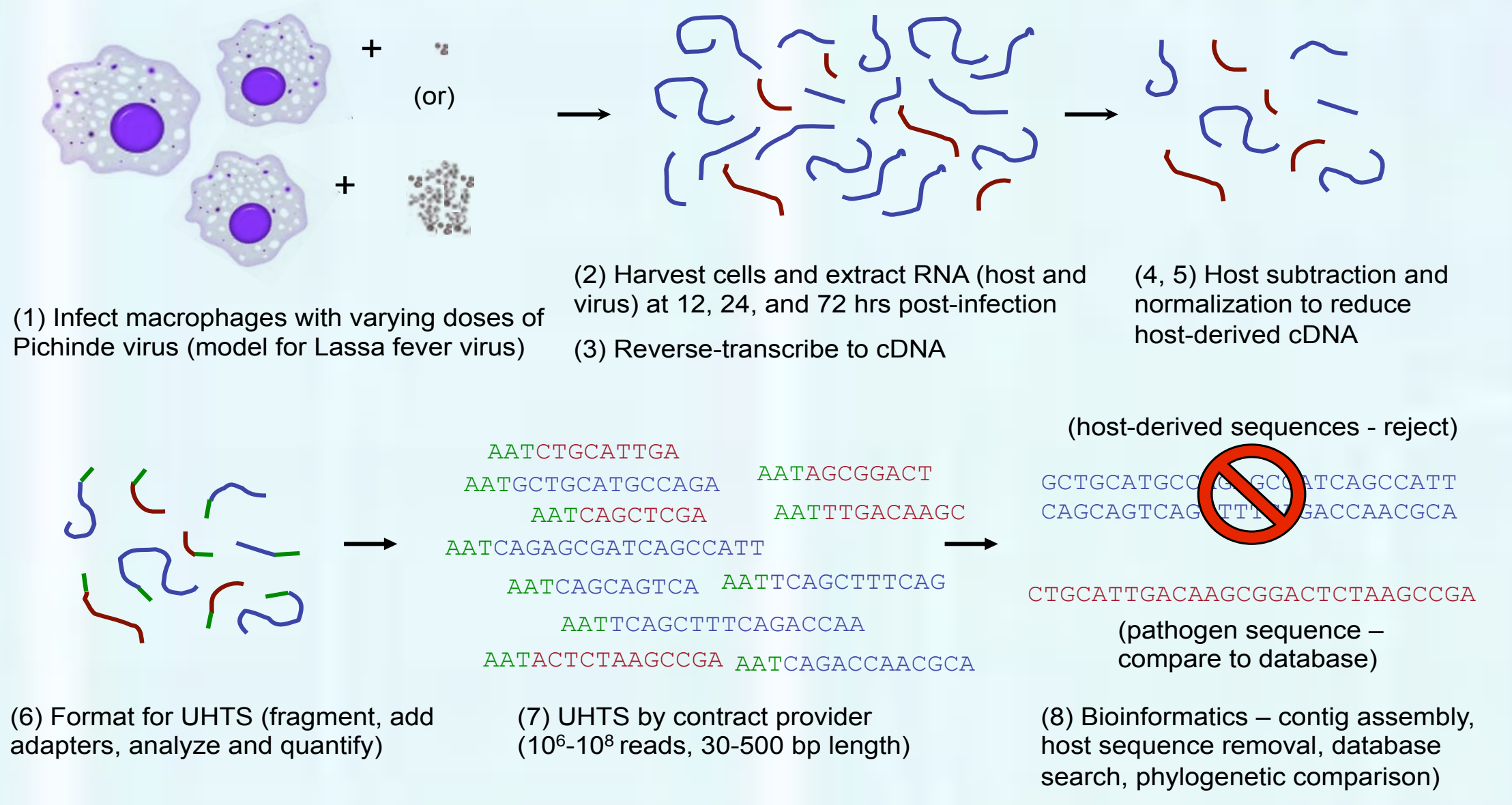
Status: Experiments ongoing; q-PCR method development underway


A
Restriction enzyme digestion of HEK 293 "host" and bacteriophage T4 "pathogen" DNA mixture into fragments averaging ~1.5 kb
B
Degradation of ds fraction by duplex-specific nuclease
+
+
C
Detection and quantitation of host and pathogen ss DNA with q-PCR
Before Normalization
After Normalization
D
Develop techniques to automate nucleic acid suppression
Microfluidics and electrokinetics
Dispersion-free reagent transport

Pathogen detection using UHTSFeasibility and Sensitivity Study

Objective: determine ability of UHTS to detect viral pathogen sequences present at known levels in host cells following subtraction and normalization

Status: Experiments underway


+
(or)
+
(1) Infect macrophages with varying doses of Pichinde virus (model for Lassa fever virus)
(2) Harvest cells and extract RNA (host and virus) at 12, 24, and 72 hrs post-infection
(3) Reverse-transcribe to cDNA
(4, 5) Host subtraction and normalization to reduce host-derived cDNA
(host-derived sequences - reject)
AATCTGCATTGA
AATGCTGCATGCCAGA
AATAGCGGACT
AATCAGCTCGA
AATTTGACAAGC
AATCAGAGCGATCAGCCATT
AATCAGCAGTCA
AATTCAGCTTTCAG
AATTCAGCTTTCAGACCAA
AATACTCTAAGCCGA
AATCAGACCAACGCA
GCTGCATGCC
ATCAGCCATT
CAGCAGTCAG
GACCAACGCA
CTGCATTGACAAGCGGACTCTAAGCCGA
(pathogen sequence – compare to database)
(6) Format for UHTS (fragment, add adapters, analyze and quantify)
(7) UHTS by contract provider (10^6-10^8 reads, 30-500 bp length)
(8) Bioinformatics – contig assembly, host sequence removal, database search, phylogenetic comparison)

Results

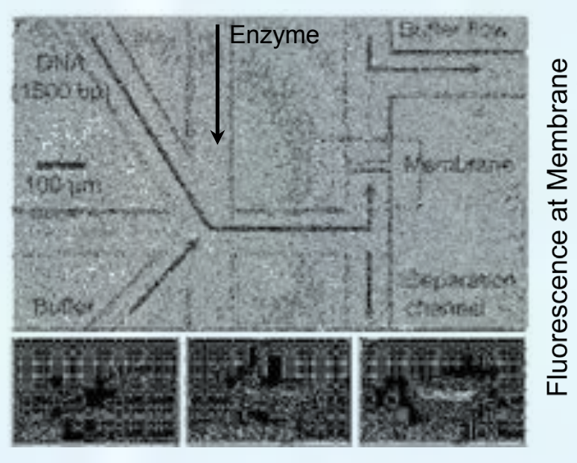
Microscale manipulation of DNA

Objective: accelerate manipulations with DNA by:

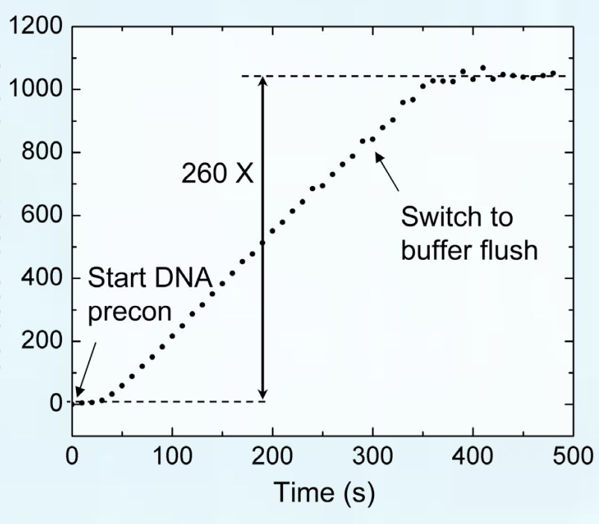
- Concentrating into a small reaction volume adjacent to a charged, nanoporous membrane
- Actively control electric field to achieve mixing and separation
- Integrate reaction with size-based separation of products

Builds capability for automated DNA analysis and multi-step operations such as normalization.

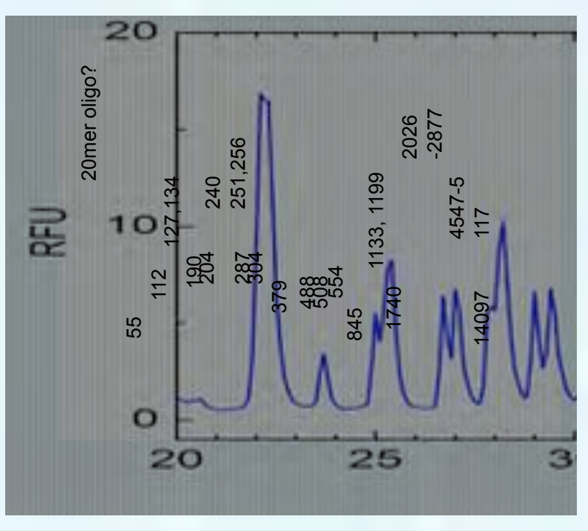
Status: Concentration of DNA and enzymes into nanovolumes, and medium-resolution fractionation of DNA have been demonstrated, enzymatic reactions are underway.



Images illustrating concentration of dye-labeled dsDNA at a nanoporous polyacrylamide membrane. Field can also be cycled to achieve mixing.



Significant concentration of DNA can be achieved in a matter of minutes, without apparent drop-off in concentration rate.



Chip electrophoretic separation of DNA ladder following membrane concentration

Results (cont.)

Research Directions

Bioinformatics challenges with mapping short UHTS reads

- Computational limitations (speed, memory)
- Unknown reference genome requires de novo assembly
- Repetitive structure of genome (~20% repetitive for 32 bp reads)
  - Paired-end reads may assist assembly
- Technical challenges with UHTS
  - Read errors are major assembly challenge.
  - Sequencer differences (e.g., longer 454 reads require different tools, SOLiD uses color space that needs to be converted to base space)

Bioinformatics Goals and approaches

- Identify hardware architectures suited to this problem
- Assess existing algorithms and data pipelines for the challenge of rare-sequence identification with short read UHTS
- Perform in silico experiments with simulated UHTS data to develop bioinformatics pipeline.
  - Develop capability and identify technical challenges before we collect our first UHTS data set.

Pipelines for short reads alignment and assembly

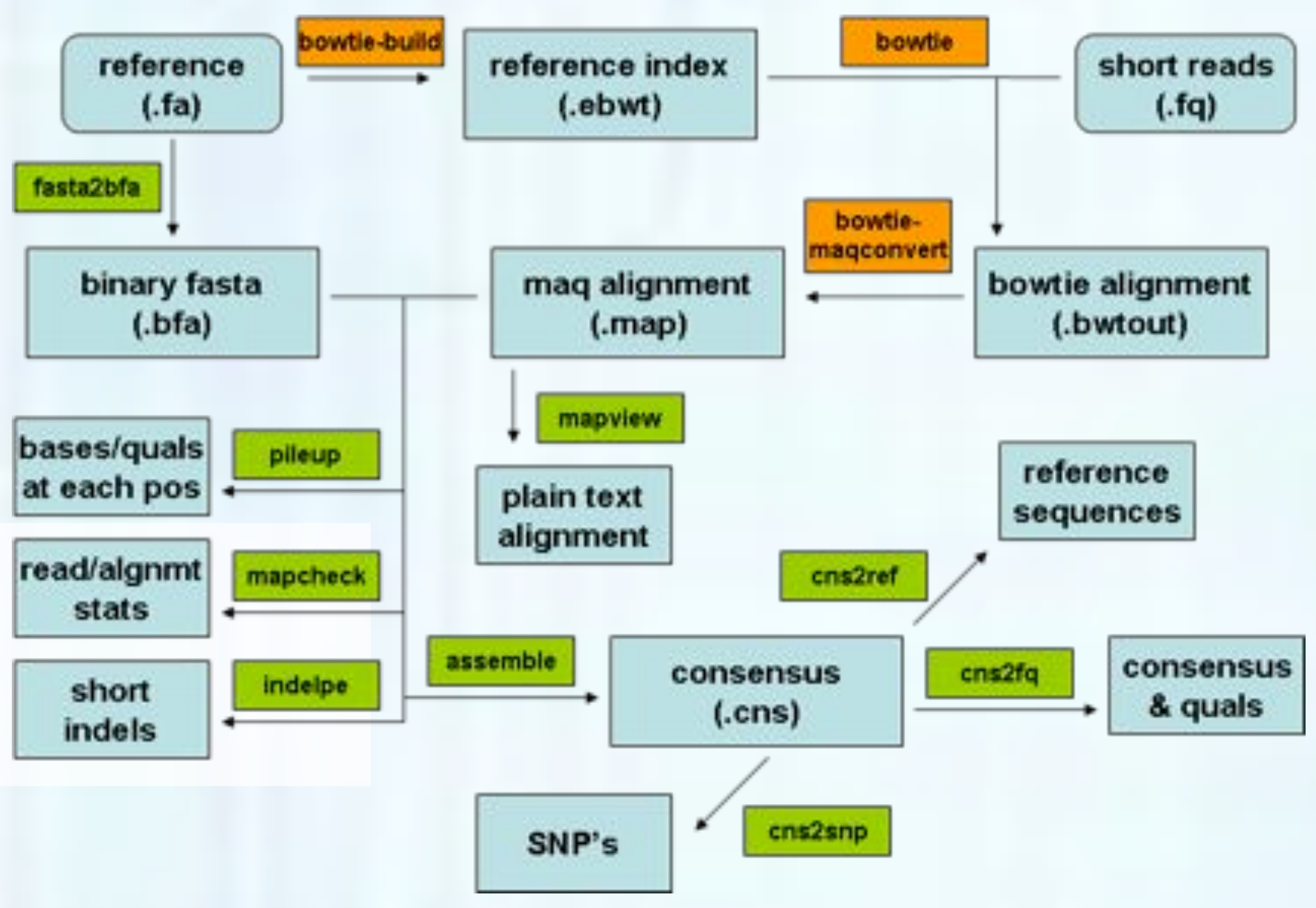
Bowtie/MAQ

- More mature
- Fast alignment and consensus generation
- Small memory footprint (1.3 GB for the human genome)
- Paired-end able
- SNP, indel calling
- Bowtie does not support SOLiD, Helicos
- No gapped alignment

BWA/SAMtools

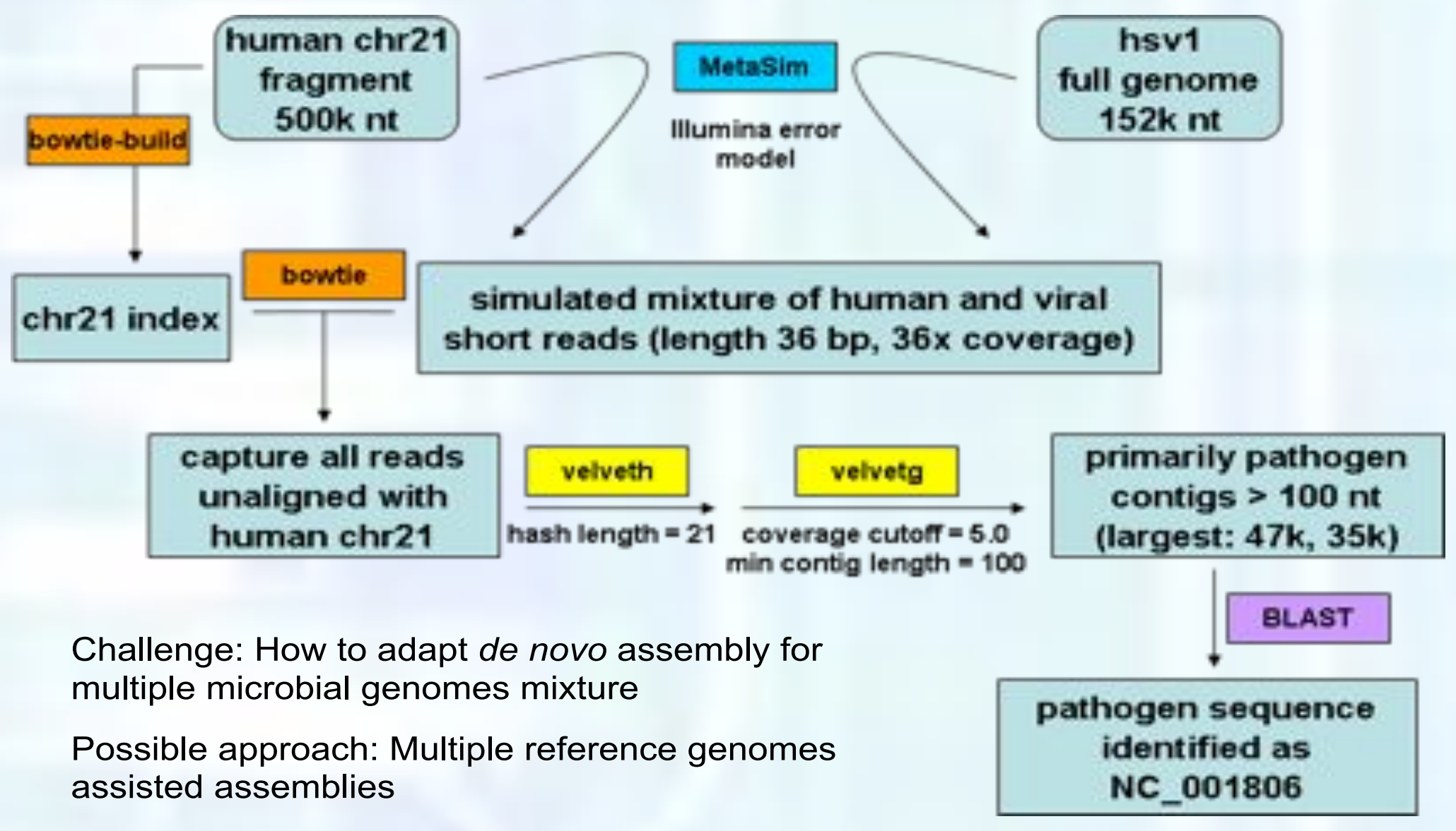
- Newer (BWA, like bowtie uses Burrows Wheeler Transform; SAM, or sequence and alignment map format, may become a standard)
- Improved short indel caller
- Gapped alignment

Bowtie/MAQ
workflow



In silico infection and "unknown" pathogen identification

- Introduce mutations in host (human chromosome 21 fragment) and pathogen (HSV1) reference sequences
- Simulate short reads for mutated sequences
- Align all simulated reads to reference chr21 sequence
- Perform de novo assembly of all unaligned reads
- BLAST obtained contigs to NCBI's database of reference sequences


human chr21 fragment 500k nt
MetaSim
Illumina error model
hsv1 full genome 152k nt
bowtie-build
chr21 index
bowtie
simulated mixture of human and viral short reads (length 36 bp, 36x coverage)
capture all reads unaligned with human chr21
velveth
hash length = 21
velvetg
coverage cutoff = 5.0
min contig length = 100
primarily pathogen contigs > 100 nt (largest: 47k, 35k)
BLAST
pathogen sequence identified as NC_001806

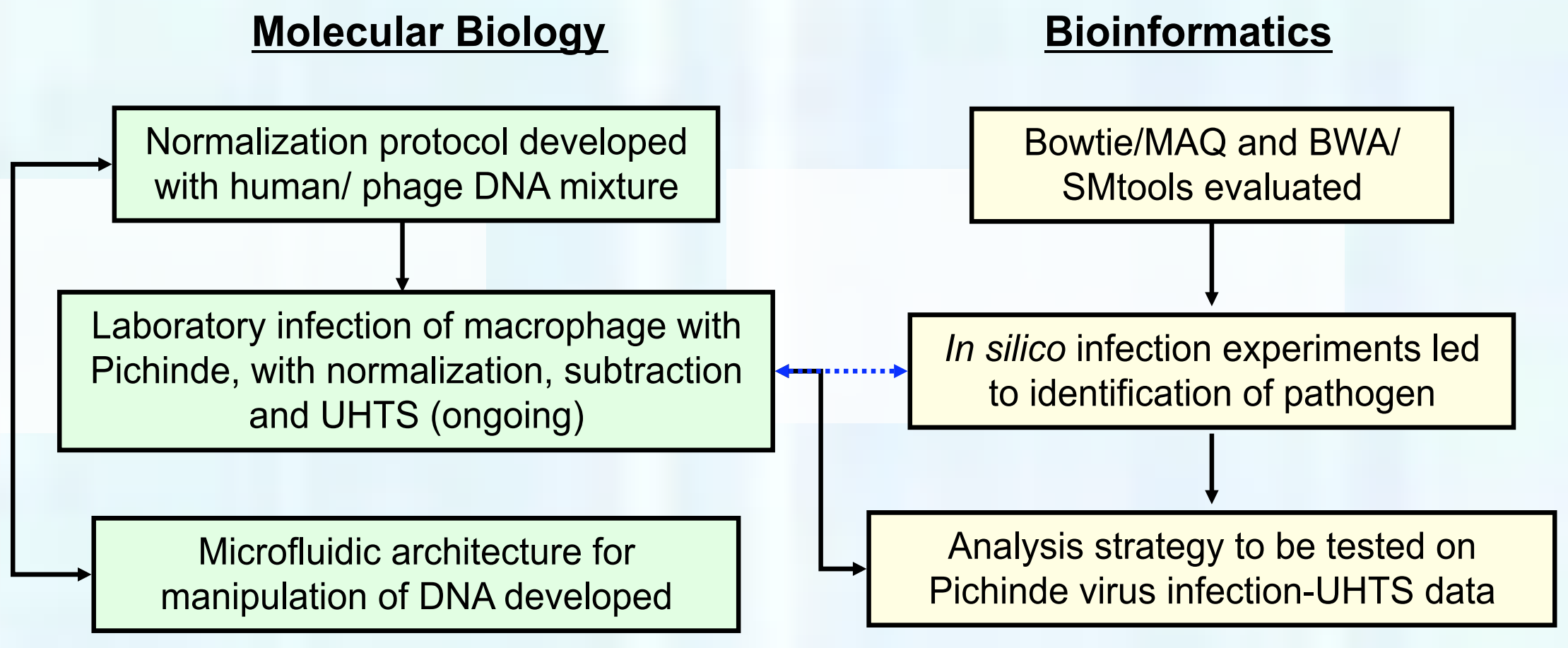
Challenge: How to adapt *de novo* assembly for multiple microbial genomes mixture

Possible approach: Multiple reference genomes assisted assemblies

Significance

In 4 months of this late-start LDRD, we have laid the groundwork for pathogen detection by UHTS, which is the most promising method available for detection and defense against unknown or engineered pathogens.

"Wet" experiments are still ongoing, with initial results expected soon. The Bioinformatics effort has identified key tools and strategies for establishing a data analysis pipeline, with in silico experiments to validate the approach.


Molecular Biology
Bioinformatics
Normalization protocol developed with human/ phage DNA mixture
Laboratory infection of macrophage with Pichinde, with normalization, subtraction and UHTS (ongoing)
Microfluidic architecture for manipulation of DNA developed
Bowtie/MAQ and BWA/ SMtools evaluated
In silico infection experiments led to identification of pathogen
Analysis strategy to be tested on Pichinde virus infection-UHTS data

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. SAND2009-5566P