SAND2010-6712P

RapTOR Data Analysis and Knowledge Discovery (DAKD)

Joe Schoeniger, MD, PhD

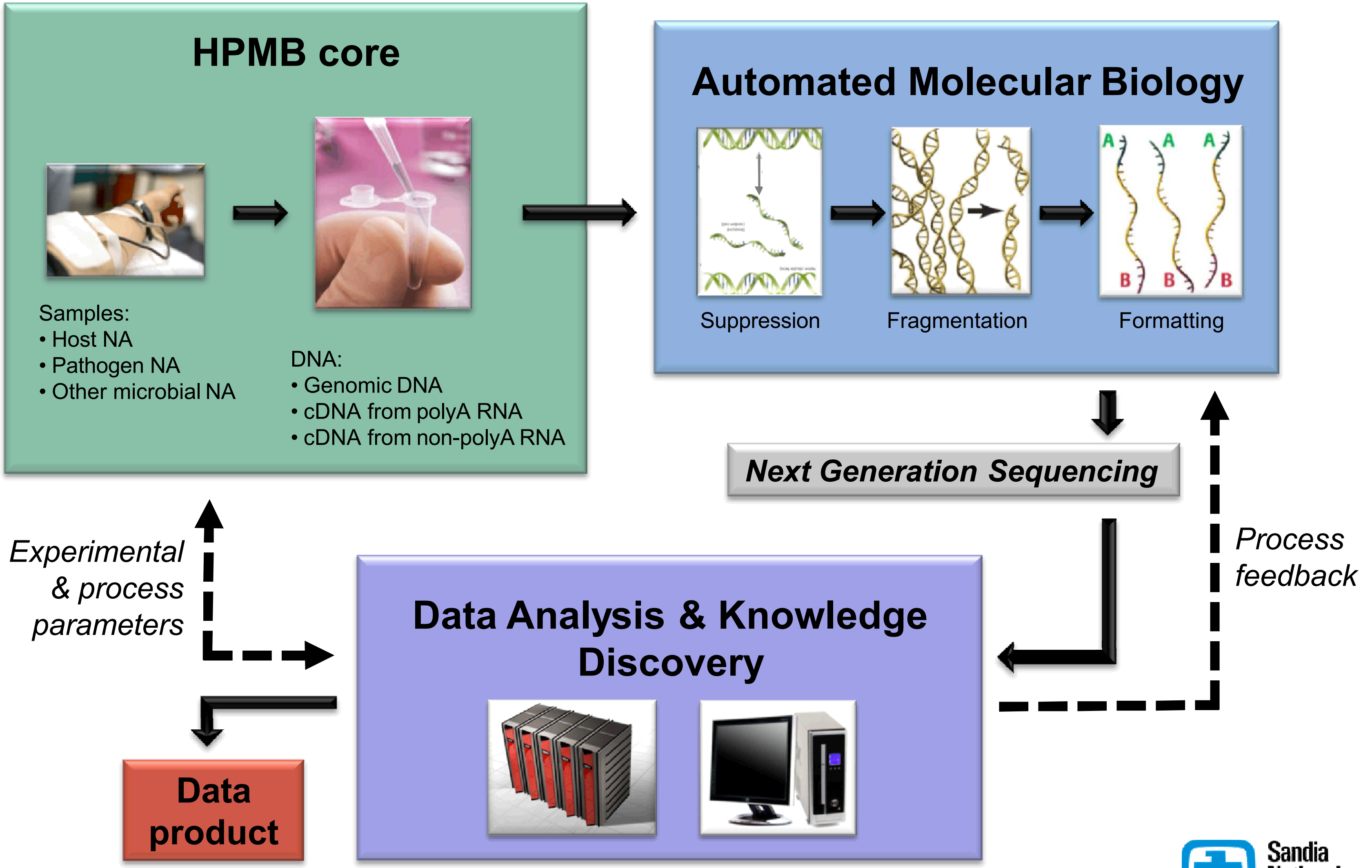
DAKD Team Lead

Rapid Threat Organism Recognition LDRD Grand Challenge

Sept 14, 2010

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.

RapTOR system concept



HPMB core

Samples:
- Host NA
- Pathogen NA
- Other microbial NA

DNA:
- Genomic DNA
- cDNA from polyA RNA
- cDNA from non-polyA RNA

Automated Molecular Biology

Suppression

Fragmentation

Formatting

Next Generation Sequencing

Process feedback

Experimental & process parameters

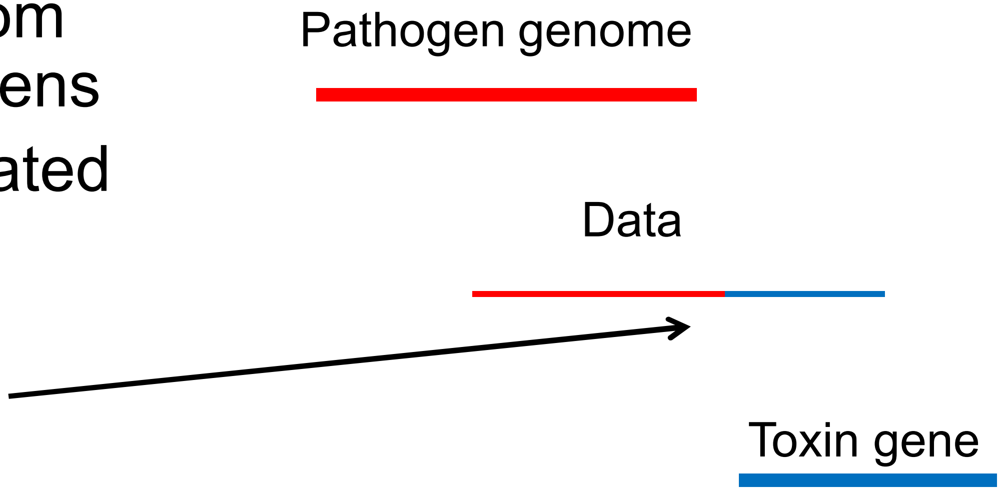
Data Analysis & Knowledge Discovery

Data product

RapTOR EAB Meeting Sept. 14, 2010

DAKD Requirement #1: Detect Agent

- Identify & Characterize Agent Sequence Targets
  - Sequences from known pathogens
  - Genes associated with virulence
  - “Unusual” recombinant sequences
  - Sequences with remote homology to pathogen genomes



Pathogen genome

Data

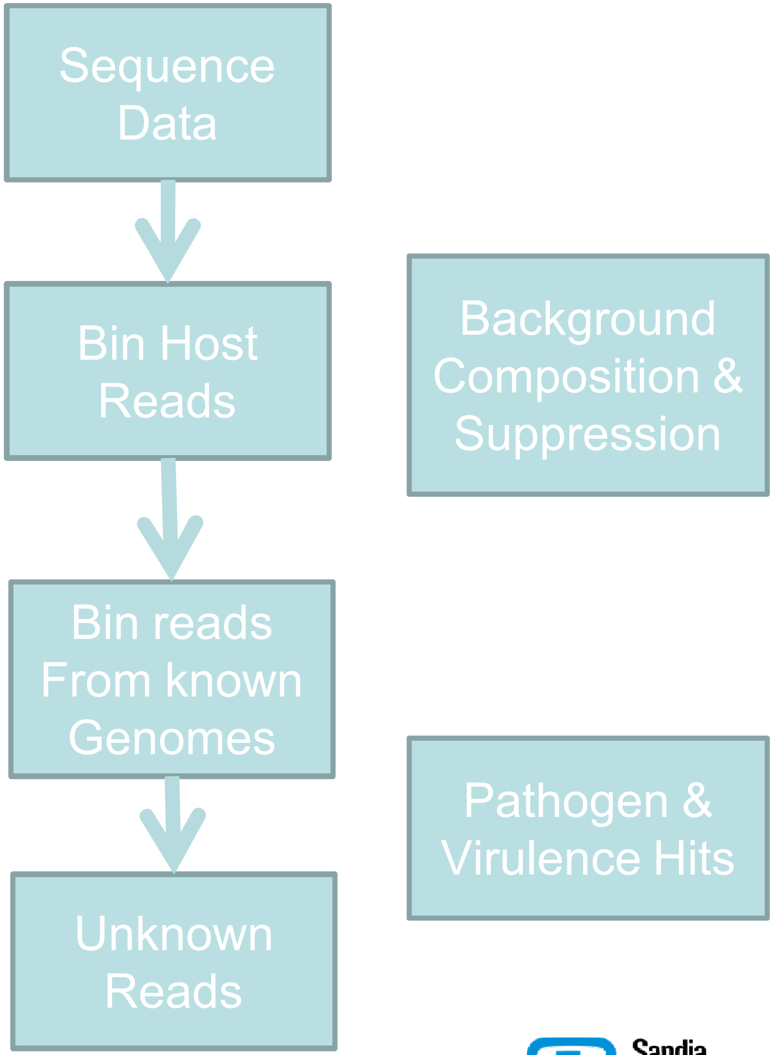
Toxin gene

Requirement #2: Diverse Samples

	Nucleic Acid Type	
	DNA	***RNA***
Sample Type		
Tissues & Cells ***e.g., PBMC/Buffy Coat***	**Genomes**	**mRNA, rRNA reg RNA, RNA virus**
Plasma/Serum	**Genomic Fragments**	**RNA virus, fragments of above**
Nasopharyngeal/Respiratory Swabs and Fluids	Genomes & Genomic Fragments	All of Above

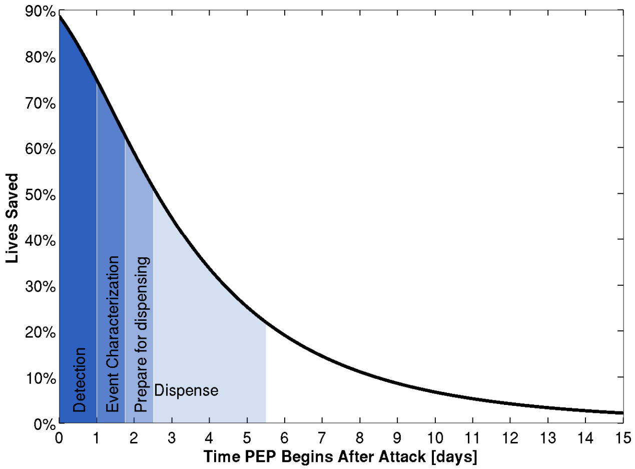
Requirement #3: Remove Background

- Background Analysis
  - Classify and Remove gibberish and background *in silico*
  - Define Characteristics of sample matrix sequence backgrounds
    - Are they “normalizable”?
    - Can we design a good probe set for suppression?
    - Normal vs. unusual abundance & composition
  - Quantitate background suppression & target enhancement methods.


Sequence Data
Bin Host Reads
Bin reads From known Genomes
Unknown Reads
Background Composition & Suppression
Pathogen & Virulence Hits

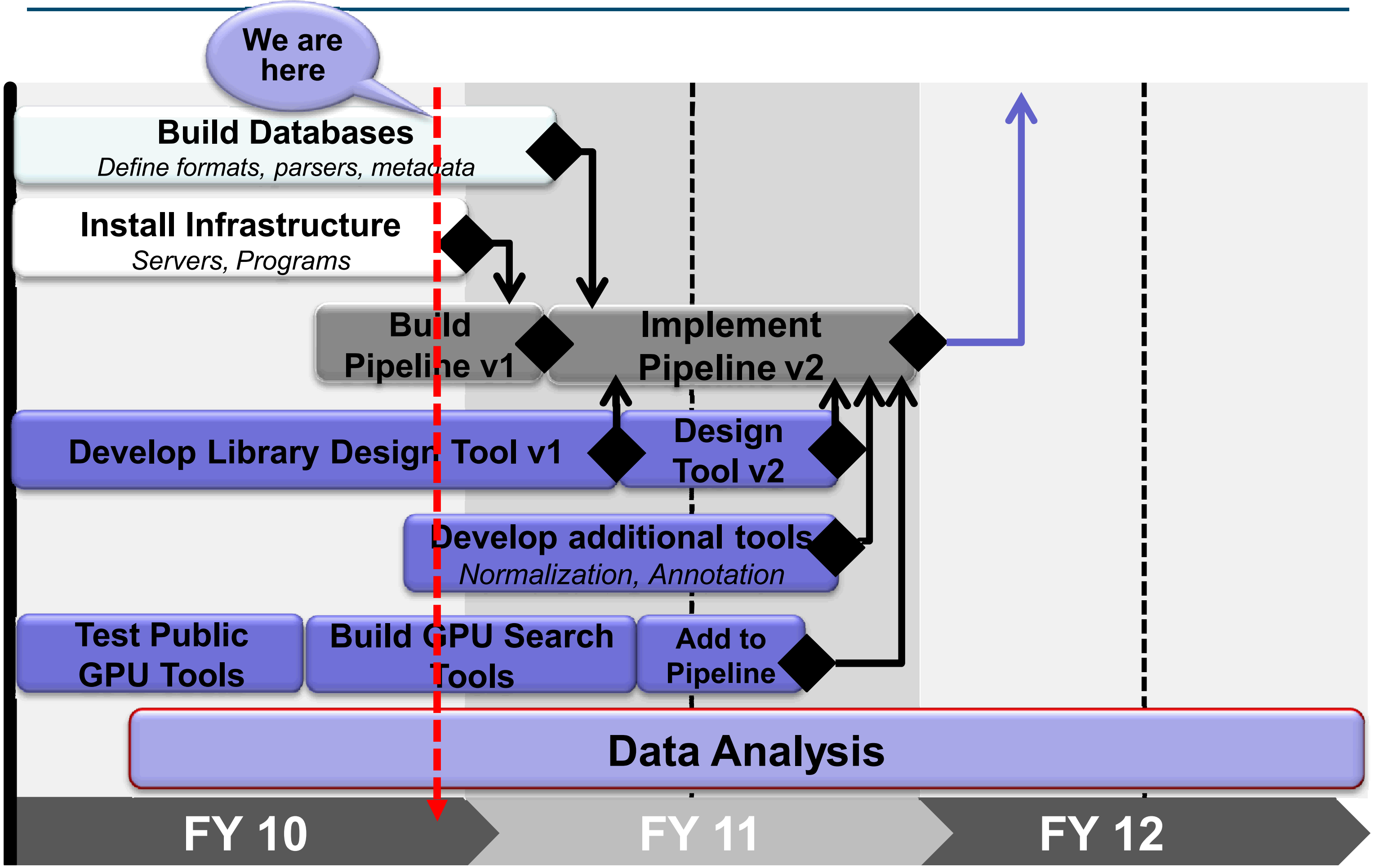
DAKD Operational Goals

- Speed
  - < 12 hours, Sequence to Results
  - 10s-100s of Samples in parallel
- Sensitivity and Reliability
  - Low False positives for
    - pathogen & virulence genes
  - False chimeras
  - Reliable suppression
- Interpretable Results
  - Associate with Metadata



Informatics must not delay initiation of response

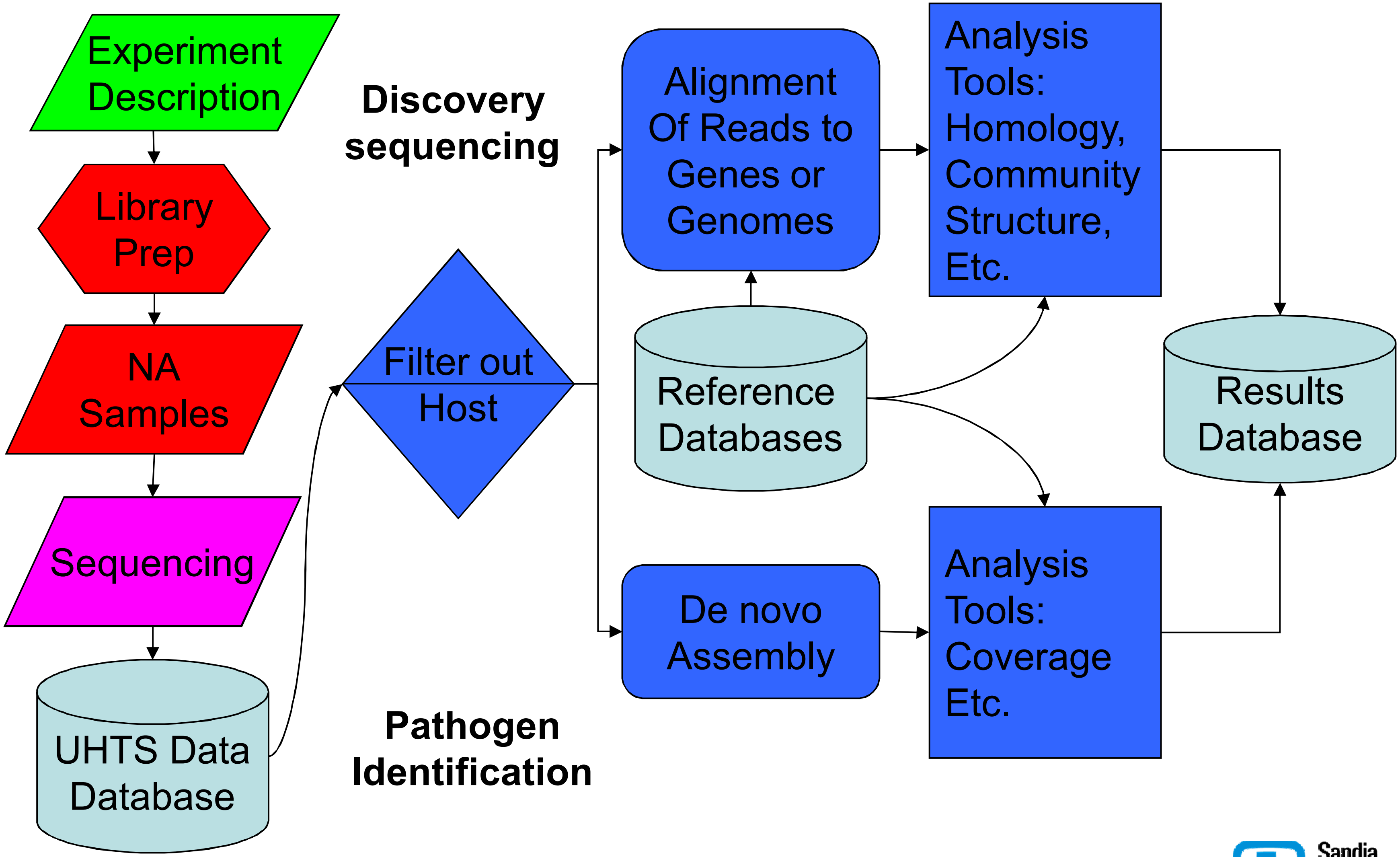
DAKD timeline with key milestones for duration of the project



End FY10 Establish Infrastructure
Mid FY11 Develop v1 Databases, GPU, Library Design Tools

End FY11 Develop Normalization, Design v2, and Annotation Tools, Implement Pipeline v2

Metagenomic Sequencing Pipelines



Experiment Description

Library Prep

NA Samples

Sequencing

UHTS Data Database

Discovery sequencing

Filter out Host

Alignment Of Reads to Genes or Genomes

Reference Databases

Analysis Tools: Homology, Community Structure, Etc.

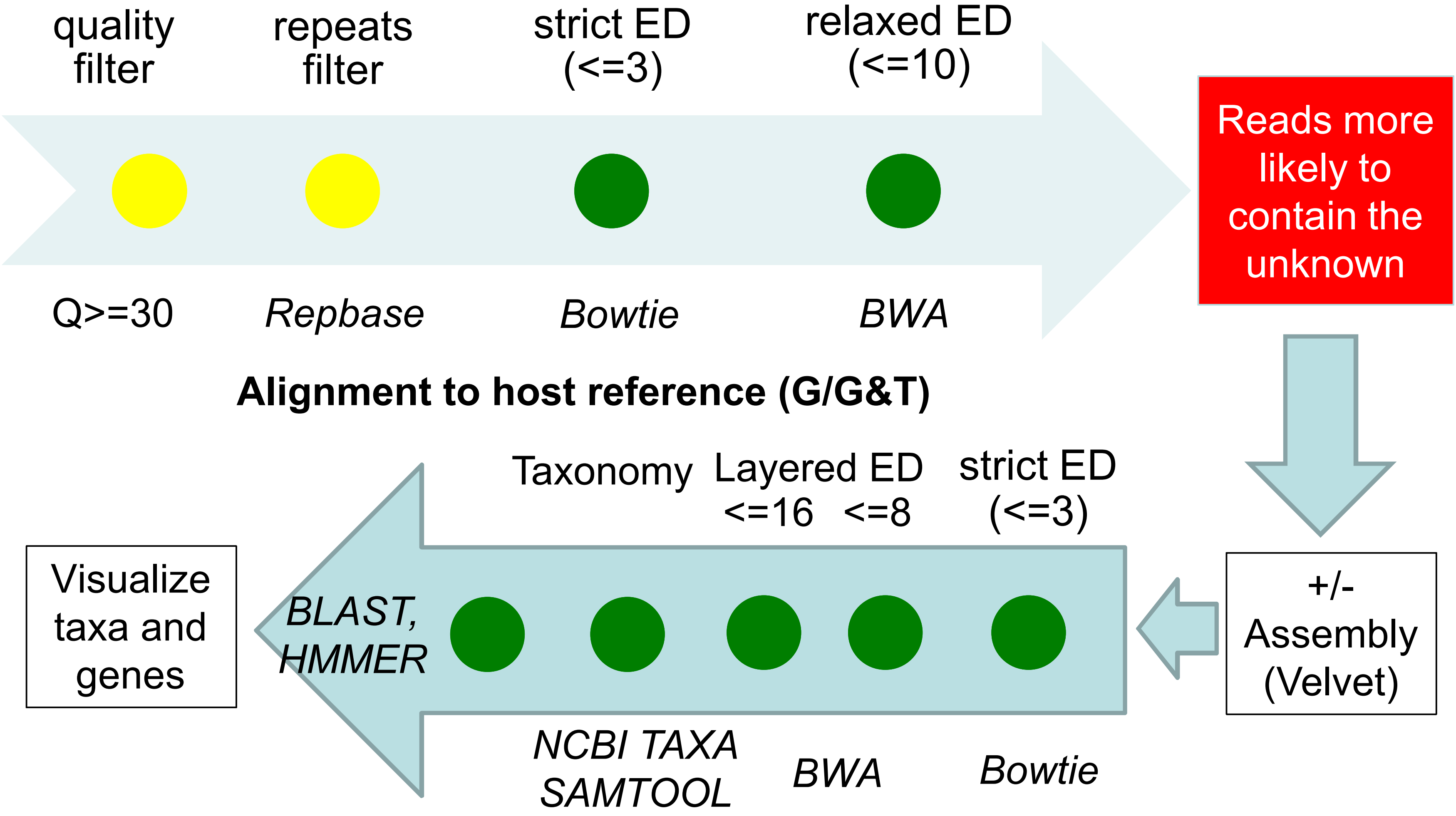
Results Database

De novo Assembly

Analysis Tools: Coverage Etc.

Pathogen Identification

“Omni-Genomic” Bioinformatics Pipeline



quality filter

repeats filter

strict ED (<=3)

relaxed ED (<=10)

Q>=30

Repbase

Bowtie

BWA

Alignment to host reference (G/G&T)

Reads more likely to contain the unknown

+/- Assembly (Velvet)

Taxonomy

Layered ED <=16 <=8

strict ED (<=3)

BLAST, HMMER

NCBI TAXA SAMTOOL

BWA

Bowtie

Visualize taxa and genes

Alignment to All Bacterial/Fungal/Viral

Software architecture

- Pipeline
  - Perl scripted pipeline
  - Accesses applications packages and algorithms primarily written in C/C++, python
  - SQL databases of reference data, raw sequence, and results under construction.
- SAM/BAM Formats for intermediate data
- Co-opting SAMtools for phylogenetic and functional analysis and vizualization
- Algorithm and Vizualization Prototyping
  - Mathematica and MatLab to C/C++

Summary of Analyses Performed

- WBC
  - DNA pipeline analysis:
    - 4 male samples with matched plasma, 2 different preps of same samples
    - Mouse BMDM infected with *F. tularensis* at MOI 1 and 100
- Plasma:
  - Pipeline Analysis of
    - DNA: 4 male, 4 female samples with deep (full lane) & barcoded (1/4 Lane) Illumina Sequencing.
    - RNA: One Sample Set
  - Preliminary Assembly Analysis
- Suppression
  - Normalization metrics for PMBC RNA spiked with *F. tularensis* mRNA

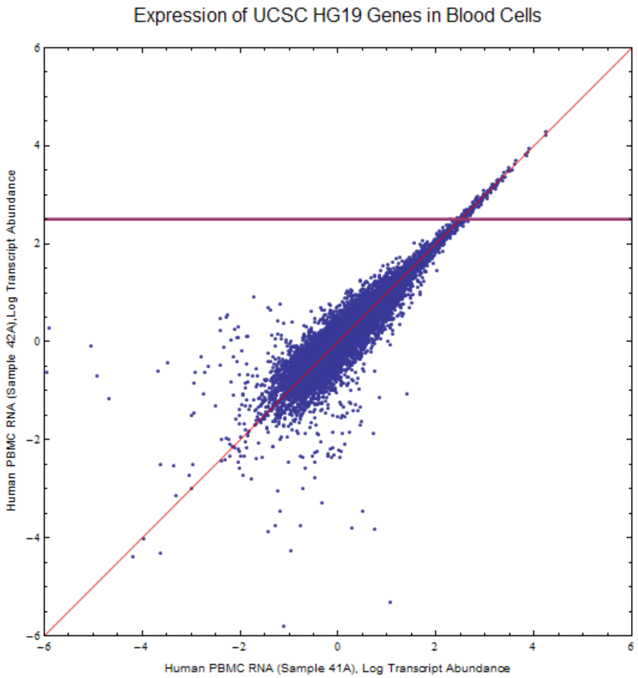
Major Findings

- Omnigenomic Pipeline time ~1 day per sample
  - No BLAST, no Amino Acid Sequence Analysis
- Can identify pathogen DNA at low MOI
- Can identify pathogen cDNA at low abundance
- Can identify very low level probable contaminant species
- Can quantitate normalization effects on
  - Abundant Host cDNA suppression
  - Pathogen cDNA enhancement
- Assembly problematic
  - Works with high-coverage species
  - Noisy, has a low yield of useful contigs so far

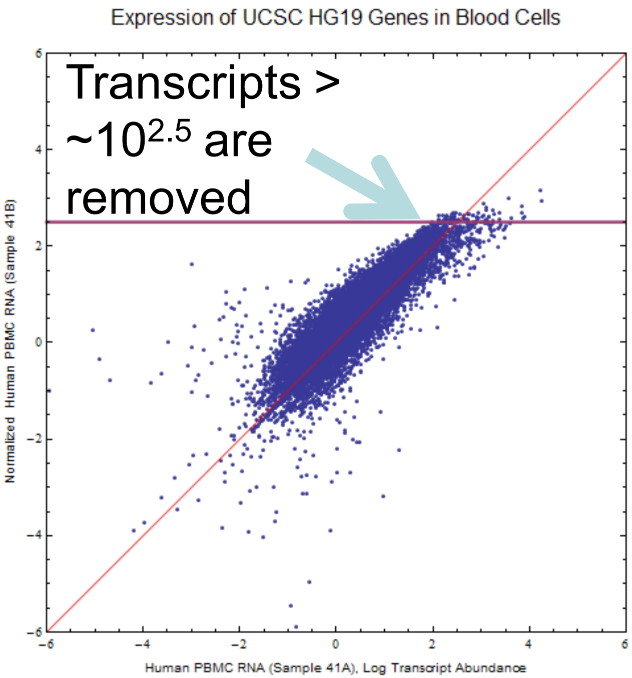
Normalization: Host Transcript Suppression

Transcriptomics Pilot:

TopHat (Bowtie-Based) -> Cufflink (Map to Exons) -> Hash & Display



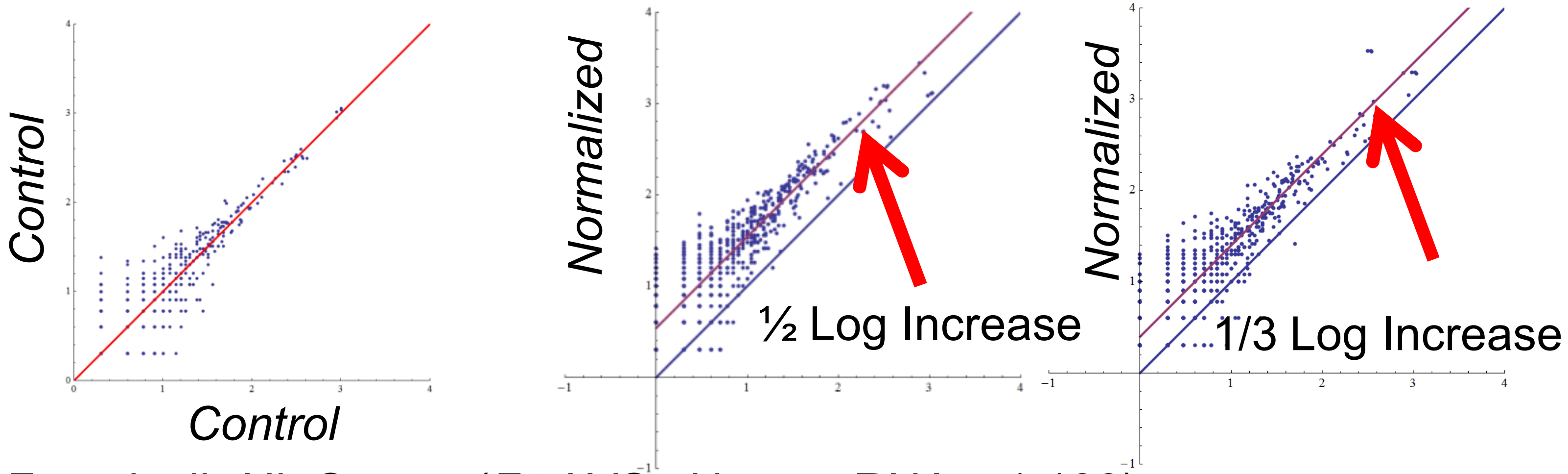
Control vs Control WBC Transcripts



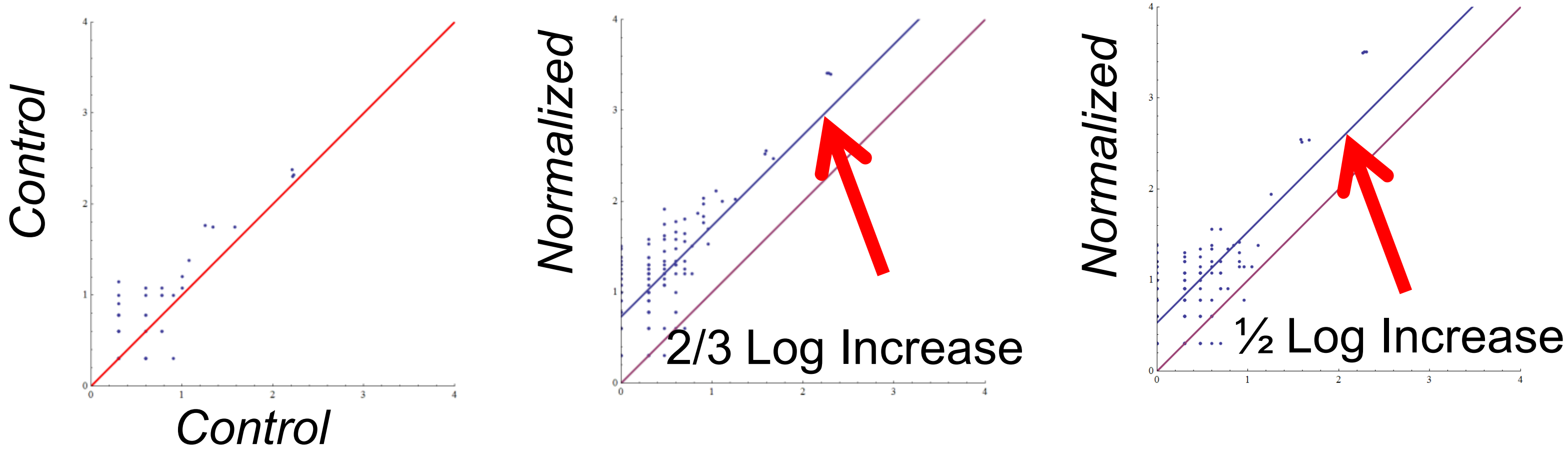
Transcripts > ~$10^{2.5}$ are removed

Control vs Normalized

Normalization: Agent Transcript Abundance

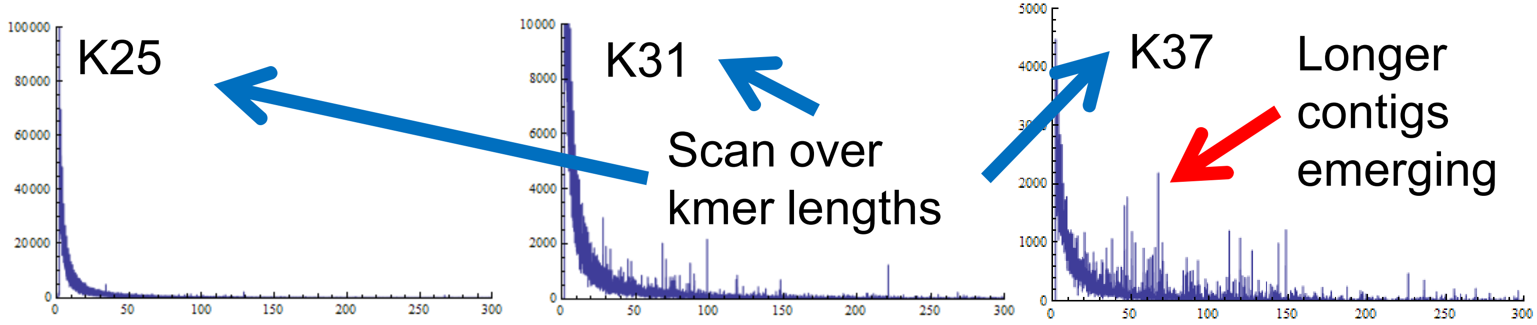


Francisella Hit Counts (*F.t.* LVS : Human RNA = 1:100)



Francisella Hit Counts (*F.t.* LVS : Human RNA = 1:10000)

Can tune assembly for complex samples

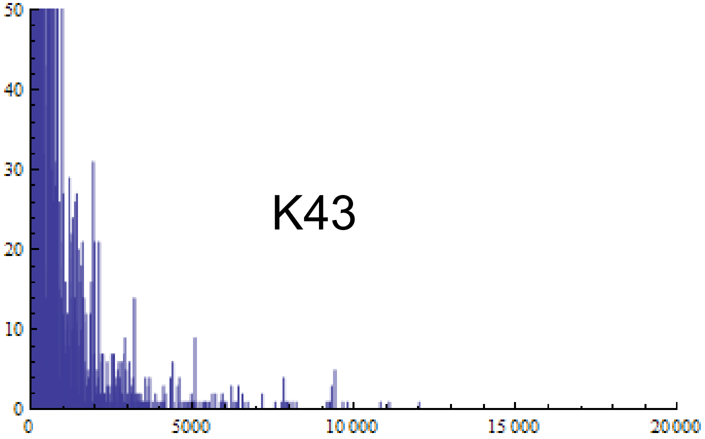


Scan over kmer lengths

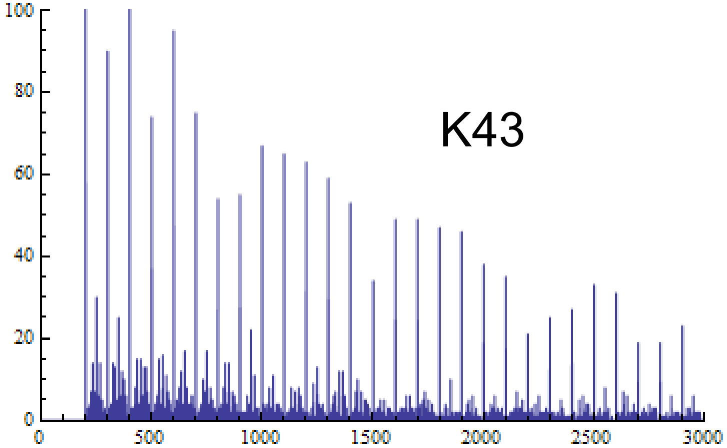
Longer contigs emerging

Length-Weighted Coverage Length-Weighted Coverage Length-Weighted Coverage

… but it produces mostly gibberish



Length-Weighted Coverage



Raw Coverage

RapTOR EAB Meeting Sept. 14, 2010

Issues

- Speed
  - Near-Exact NA sequence match screening to human host and all microbial genomes < 1 hour
    - Inexact matching takes many hours (with BWA)
    - BLASTp for millions of reads impractical
- Prep variability
  - Large variation in redundancy of reads.
  - Variation in alignment % to host of WBC data
    - ~45% for mRNA/cDNA to UCSC hg19 exon models
    - 15-80% for Human for Plasma & WBC DNA
- Assembly
  - Current generation assemblers gibber
- Scoring of alignments

Issues (continued)

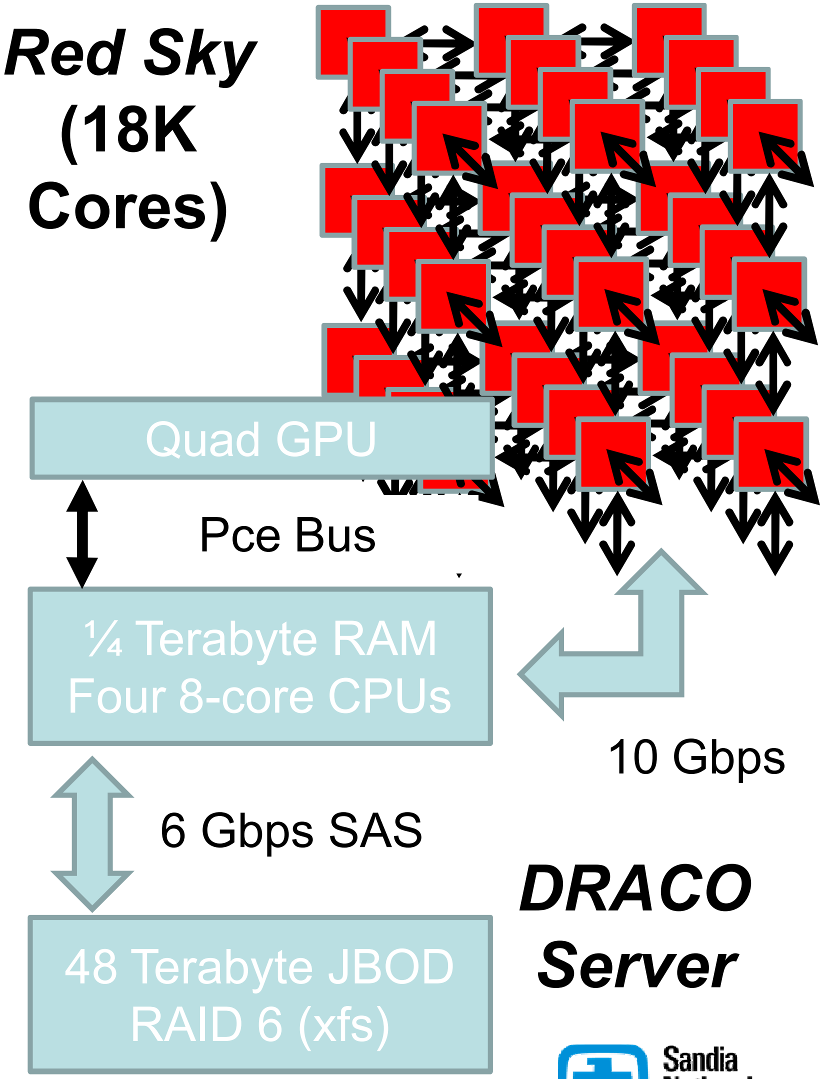
- Memory
  - Assemblers, Indexing for alignment and large data set alignments take 10s to 100s of Gb of RAM
  - Parametric analysis of assembly or transcript counting takes 100s Gb disk per data set
- Liabilities of Standard Software Tools
  - Disk I/O Bottlenecks: not optimized for large RAM
  - Fast tools mostly in NA sequence space, more needed for amino acid sequence (e.g., BLAT)
  - Normalization conventions in transcriptomics
- Data movement
  - It takes between 2 and 20 hours to move 1Tbyte over a 1 Gbps pipe.

Solutions

- Optimize Library design for both normalization and sequencing
- Improve software implementations
  - Minimize disk IO
  - GPU codes:
    - ~50x speedup for exact alignment
    - GPU BLAST & HMMER 10x
- Servers: maximize flexibility, speed & capacity
- Assembly
  - Information and complexity filtering before and after assembly
  - Templated assembly against genomes

New Server Architecture

- Memory model:
  - Large shared memory essential
  - Distributed systems can be useful if reference data segmentable.
- Compute requirements:
  - Host filtering done on workstation
  - Standard BLAST/BLAT can be done locally or on distributed systems, bandwidth allowing.
  - Intrinsic characterization and remote homology matching requires significant concurrent computing
- Data Storage:
  - Tens of terabytes, RAID 6 xfs
- Communications:
  - 10 Gbps pipe required for remote Supercomputing / Cloud computing to be useful



***Red Sky* (18K Cores)**

Quad GPU

Pce Bus

¼ Terabyte RAM
Four 8-core CPUs

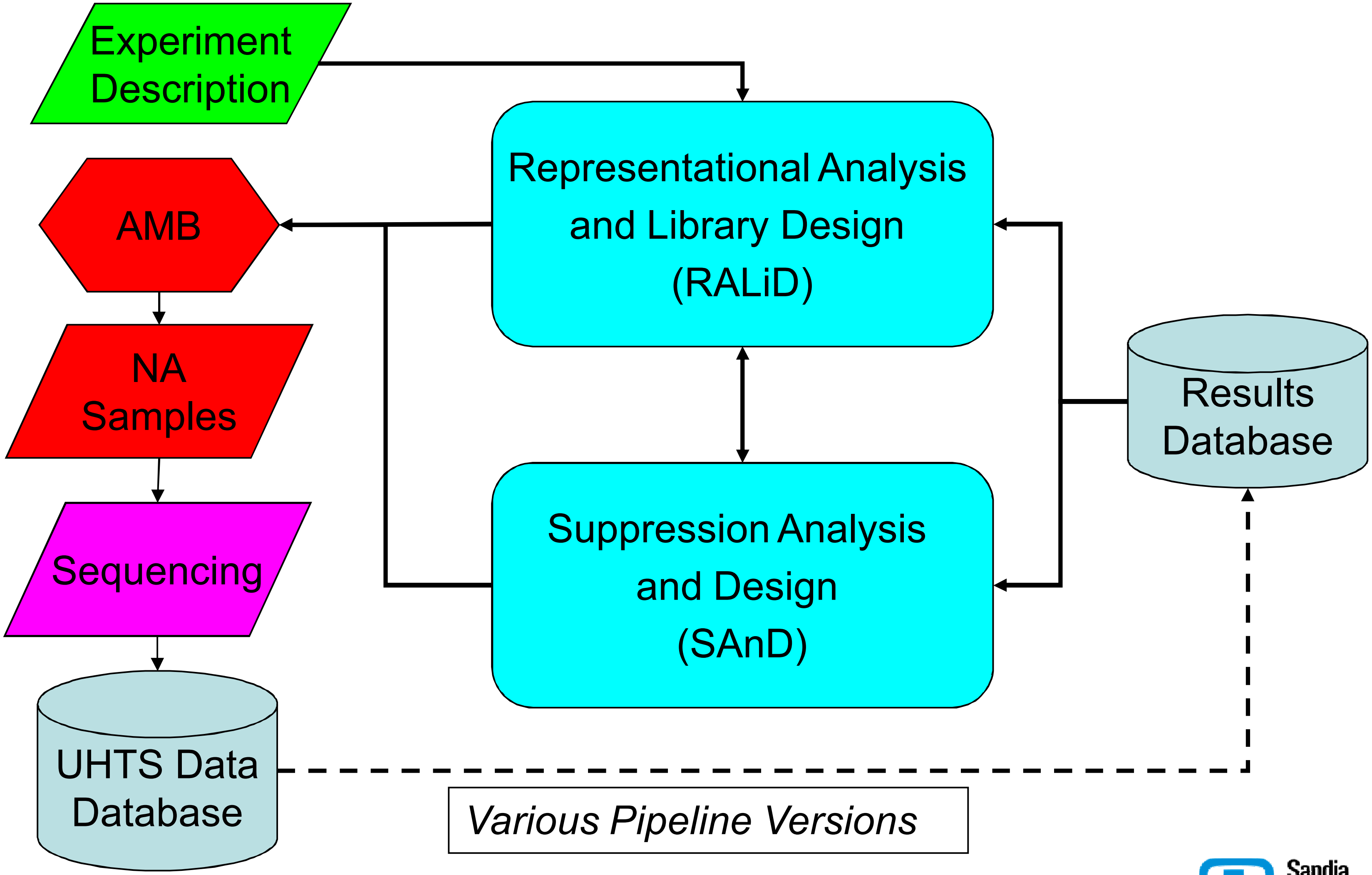
10 Gbps

6 Gbps SAS

48 Terabyte JBOD
RAID 6 (xfs)

DRACO* *Server

Suppression and Library Design



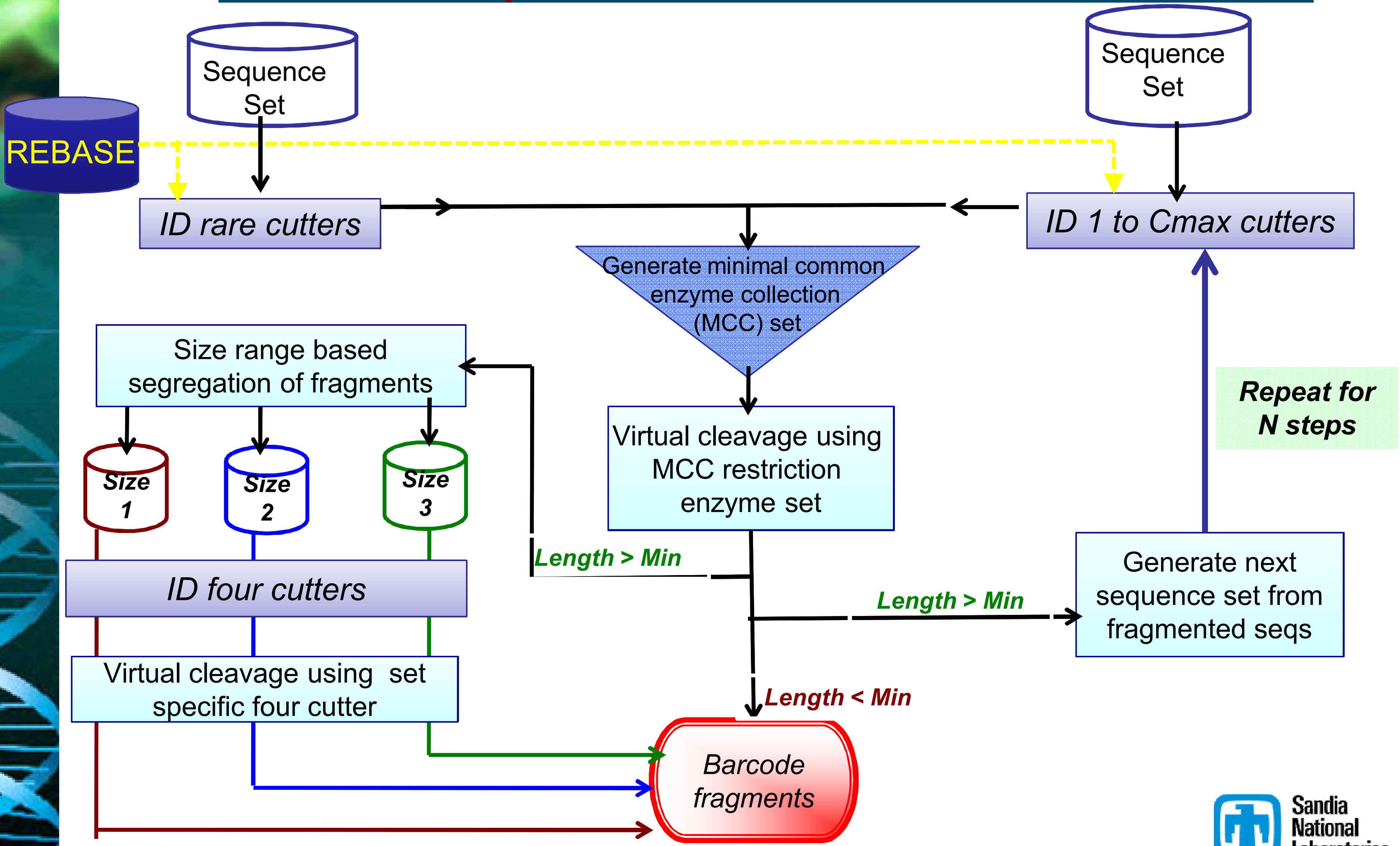
Experiment Description
AMB
NA Samples
Sequencing
UHTS Data Database
Representational Analysis and Library Design (RALiD)
Suppression Analysis and Design (SAnD)
Results Database

Various Pipeline Versions

Representational Analysis and Library Design (RALiD)

Suppression and Library Design

Restriction Enzyme-Based Size Selection Tool


Sequence Set
REBASE
ID rare cutters
Sequence Set
ID 1 to Cmax cutters
Generate minimal common enzyme collection (MCC) set
Size range based segregation of fragments
Virtual cleavage using MCC restriction enzyme set
Repeat for N steps
Size 1
Size 2
Size 3
Length > Min
ID four cutters
Length > Min
Generate next sequence set from fragmented seqs
Virtual cleavage using set specific four cutter
Length < Min
Barcode fragments

Team

- Joe Schoeniger: Team Lead, Transcriptomics, Assembly
- Milind Misra: Pipeline Prototyping
- Amy Powell: Phylogenomics, Databases
- Chi-Chi May: Library Design

Next Steps

- Backend BLAST & HMMER analysis
- Establish Local SQL Databases
- Phylogenetic and functional visualization
  - Scoring Schemes
- Improve software implementations
  - Tweak public source codes to Minimize disk IO
  - Adapt inexact string matching for aa sequence
  - GPU code for alignment (50x exact, inexact TBD)
  - Interface to Sandia HPC and cloud
- Assembly Automate pipeline interface
  - Implement information filters & Optimization
- Optimized library design tool

Questions?