Whole Genome, Physics-based Sequence Alignment for Pathogen Signature Design

Jason Gans[1], Wu-Chun Feng[2], Murray Wolinsky[1]

[1]Los Alamos National Laboratory

[2]Virginia Tech

Los Alamos NATIONAL LABORATORY EST.1943

The World's Greatest Science Protecting America

NNSA

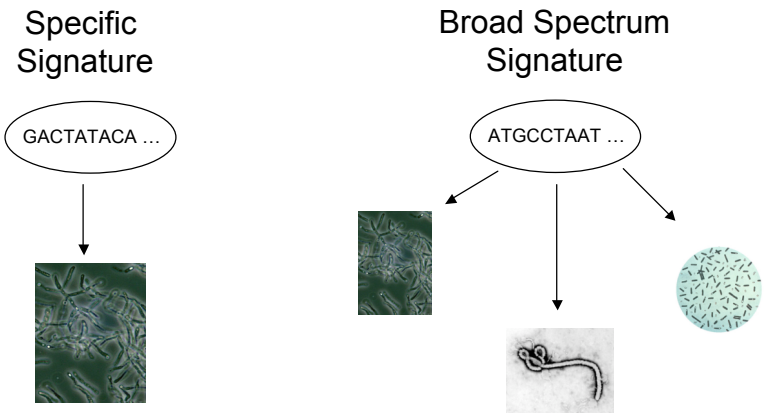
The need for pathogen detection

Los Alamos NATIONAL LABORATORY EST.1943

The World's Greatest Science Protecting America

NNSA

Specific vs broad spectrum signatures



Detecting multiple targets with a single experiment reduces cost

Los Alamos NATIONAL LABORATORY EST.1943
The World's Greatest Science Protecting America

Images courtesy of http://www.microbelibrary.org

NNSA

DNA-based detection assays

DNA signature development can exploit the growing number of sequenced genomes

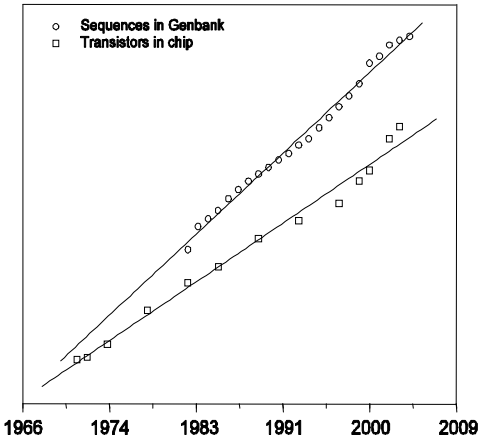
- ~ 400 bacterial genomes available
- ~ 8000 viral genomes available

Mature technologies for high throughput detection

- Gene chip, PCR

Los Alamos NATIONAL LABORATORY EST.1943
The World's Greatest Science Protecting America
NNSA

Parallel Computation (Moore's Law is not enough)



http://www.intel.com/museum/archives/history_docs/mooreslaw.htm
http://www.ncbi.nlm.nih.gov/Genbank/genbankstats.html

Signature Detection Pipeline

1. Partition DNA sequence data
2. Identify sequence fragments unique to target species (i.e. not found in background sequences)
3. Identify unique fragments that are found in multiple target species
4. Compute minimal set of unique fragments required to span target species set
5. Convert unique fragments into biological assays
6. Computationally and experimentally validate assays

How to define sequence similarity/uniqueness?

The answer depends on how we ask the question

The assay format, e.g.
- PCR amplification
- DNA chip probe hybridization
- Single base extension

defines similarity.

How to define sequence similarity/uniqueness?

The answer depends on how we ask the question

The assay format, e.g.
- PCR amplification
- DNA chip probe hybridization
- Single base extension

defines similarity.

A sequence similarity metric that supports common assay formats is DNA melting temperature, T_m

$T_m^{A,B} > T_m^H \Rightarrow$ Sequences A and B are equivalent

$T_m^{A,B} < T_m^L \Rightarrow$ Sequences A and B are dissimilar

Los Alamos NATIONAL LABORATORY EST.1943
The World's Greatest Science Protecting America
NNSA

How to define sequence similarity/uniqueness?

$$T_M = \frac{\Delta H^\circ}{\Delta S^\circ + R \ln\left(\frac{C_T}{N}\right)}$$

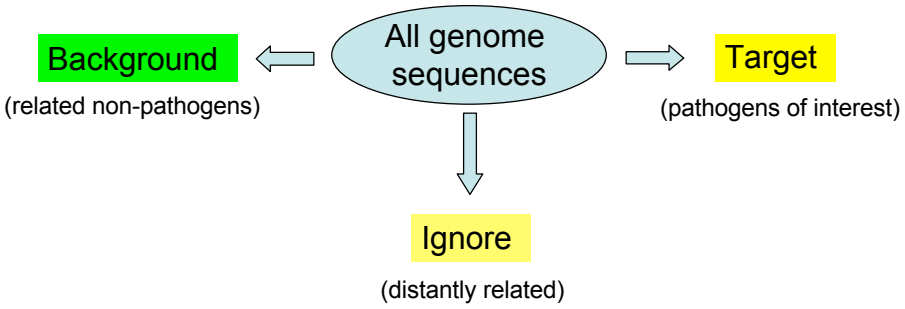
R = Gas constant
C_T = Strand concentration
N = sequence dependant constant

Nearest Neighbor Model

$$\Delta H^\circ = \Delta H_{init} + \Delta H_{sym} + \sum_{\substack{i=\{A,T,G,C\} \\ j=\{A,T,G,C\}}} n_{ij} \Delta H_{ij}$$

$$\Delta S^\circ = \Delta S_{init} + \Delta S_{sym} + \sum_{\substack{i=\{A,T,G,C\} \\ j=\{A,T,G,C\}}} n_{ij} \Delta S_{ij}$$

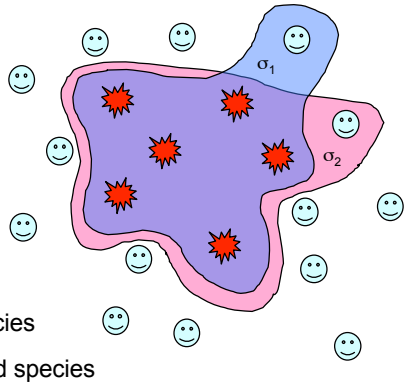
Partition Target Sequences

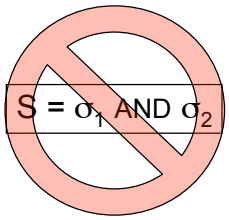


Isolate sequences unique to targets

Why search for unique sequence?

Non-unique signatures susceptible to false positives





Target species

Background species

σ_1= ATGACCATGA …

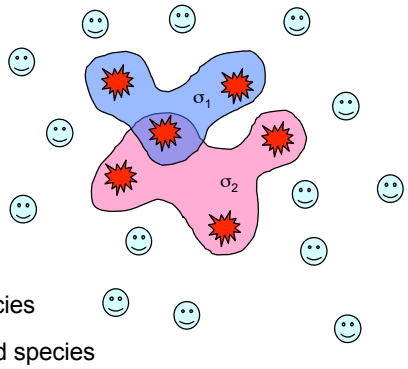
σ_2= GCATTCAGTA …

Los Alamos NATIONAL LABORATORY EST.1943

The World's Greatest Science Protecting America

NNSA

Isolate sequences unique to targets



Detection Signature

S = σ_1 OR σ_2

Target species

Background species

σ_1= ATGACCATGA …

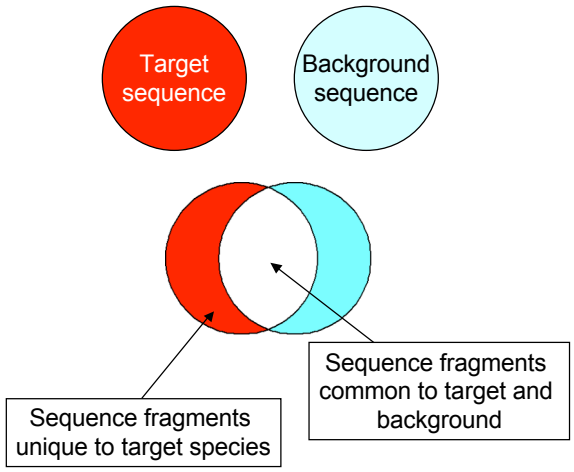
σ_2= GCATTCAGTA …

Los Alamos NATIONAL LABORATORY EST.1943

The World's Greatest Science Protecting America

NNSA

Isolate sequence fragments unique to target species



Los Alamos NATIONAL LABORATORY EST.1943
The World's Greatest Science Protecting America
NNSA

Isolate sequences unique to targets: 1st generation

Using mpiBLAST to compute the set difference.

$\tilde{\cap}$ →

1. Use mpiBLAST to identify regions of similarity between U_i and B.
2. If a sequence fragment has a T_m > 55 °C (over a ≤ 21 base window) then it is considered a match.

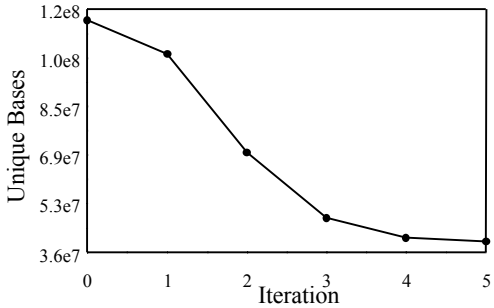
$$U = T - (T \cap B)$$

$$U \approx U_f$$

$$U_{i+1} = U_i - (U_i \tilde{\cap} B)$$

$$U_0 = T$$

$$U_f \approx U_3$$



Los Alamos NATIONAL LABORATORY EST.1943
The World's Greatest Science Protecting America
NNSA

Isolate sequences unique to targets: 1st generation

mpiBLAST

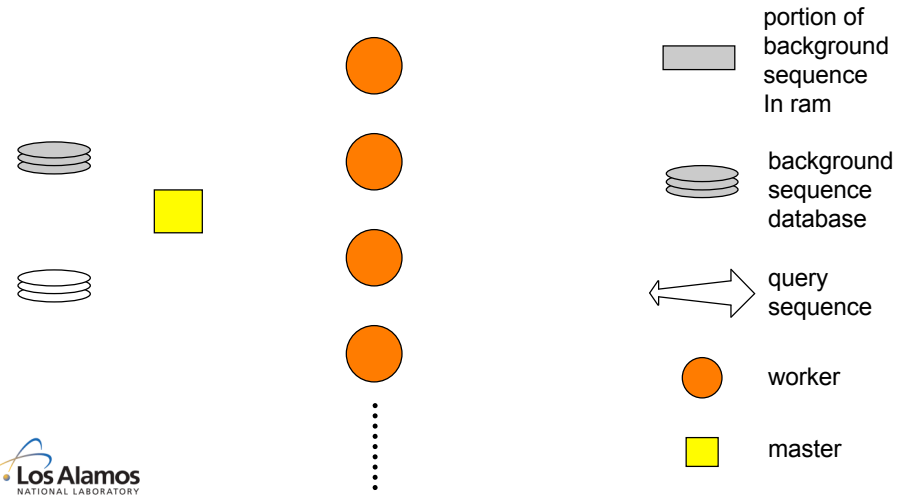
- Pro: Allowed signature design to use all available bacterial genomes
- Con: Required iteration and separate computation of set difference

2nd generation implementation

- Use high level structure of mpiBLAST (i.e. master/worker, database segmentation, query segmentation, MPI/C++)
- Directly compute set difference: *in silico* subtractive hybridization

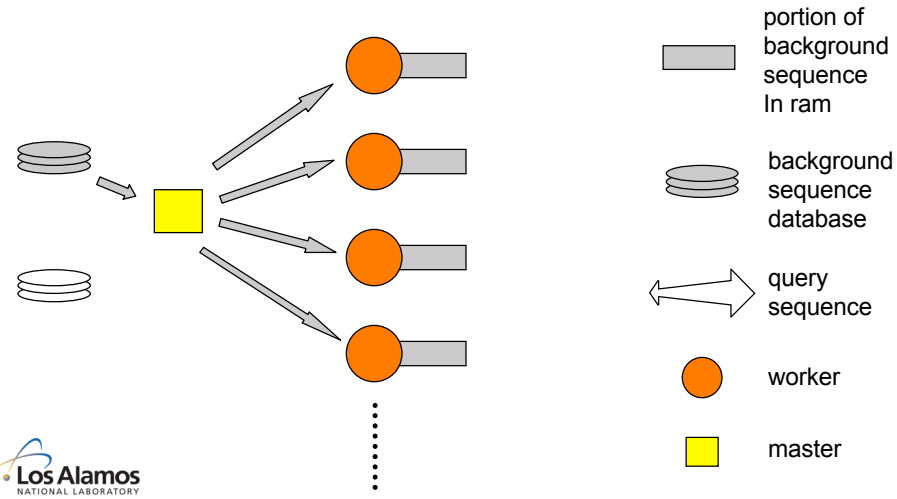
Los Alamos NATIONAL LABORATORY EST.1943
The World's Greatest Science Protecting America
NNSA

In silico subtractive hybridization



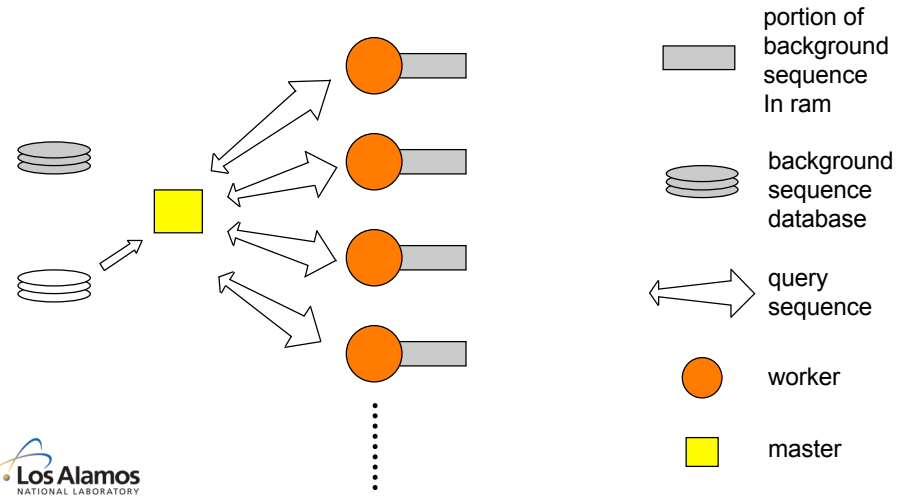
Los Alamos NATIONAL LABORATORY EST.1943
The World's Greatest Science Protecting America
NNSA

Insilico subtractive hybridization



Los Alamos NATIONAL LABORATORY EST.1943
The World's Greatest Science Protecting America
NNSA

In silico subtractive hybridization



Los Alamos NATIONAL LABORATORY EST.1943
The World's Greatest Science Protecting America
NNSA

Target fragment alignment

In silico subtractive hybridization produces fragments unique to every target species

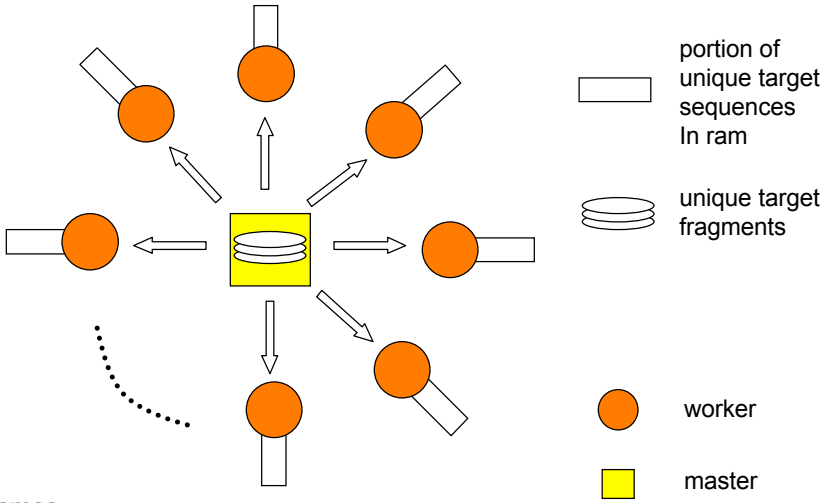
species A → {ACTGGATCGATC, GGCTGGATTCTAGG, TAGGCTTAGGCTTA, ATTCGGCCCAGATAG, …}
species B → {GCTTCTAGACAAC, ATGGCGATTAGCCA, GCTAAGCCTAGCTA, TTTGACTAGATCAC, …}
species C → {GGCTAGGCCCA, GGCTCTAGGATATAC, CGTCTAGGCTTATAT, CGGATTCGGCTTAG, …}
⋮

Need to associate every unique target fragment with one or more target genomes

CGTAGGCTTAGGATAT	AGGCTATAGCGGA	CAATAGGCTTATAGGATTA	▪▪▪▪
↓	↓	↓	
{species A, species B, speceis C}	{species B, species D, species E, species F}	{species A, species F}	

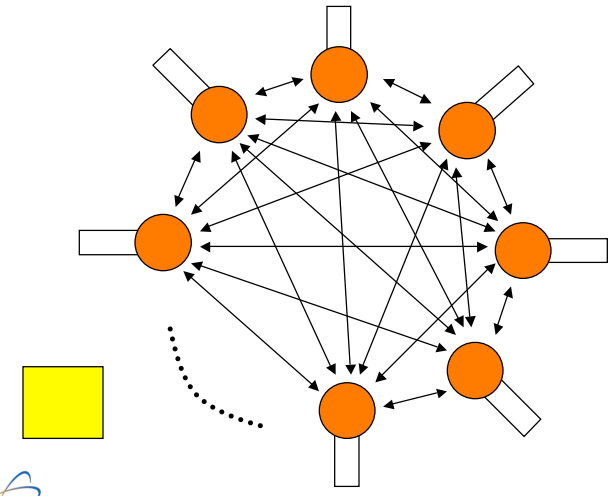
Los Alamos NATIONAL LABORATORY EST.1943
The World's Greatest Science Protecting America
NNSA

Target fragment alignment



Los Alamos NATIONAL LABORATORY EST.1943
The World's Greatest Science Protecting America
NNSA

Target fragment alignment



portion of unique target sequences In ram

unique target fragments

worker

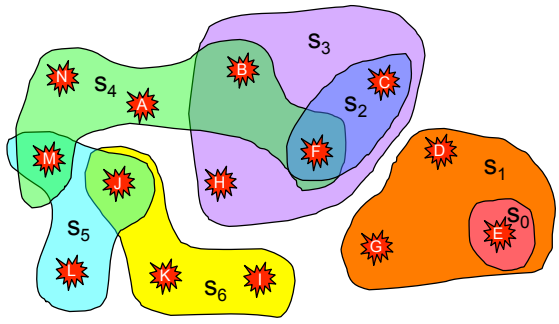
master

Los Alamos NATIONAL LABORATORY EST.1943

The World's Greatest Science Protecting America

NNSA

Find the minimal set of signatures that will detect all target species



s_i is a set of species that contain a common sequence fragment

Los Alamos NATIONAL LABORATORY EST.1943

The World's Greatest Science Protecting America

NNSA

Find the minimal set of sequence fragments that will detect all target species

Given:

$T = \{A, B, C, D, E, F, G, \ldots, \}$
$s_0 = \{E\}$
$s_1 = \{D, E, G\}$
$s_2 = \{B, C, F, H\}$
⋮

Our task is to find $G \equiv \{s_\alpha, s_\beta, s_\delta, \ldots\}$

Where $T = \bigcup_i G_i$
and
$|G|$ is a global minimum

The set coverage problem is NP hard!

Los Alamos NATIONAL LABORATORY EST.1943
The World's Greatest Science Protecting America
NNSA

Find the minimal set of sequence fragments that will detect all target taxa

Approximate solution

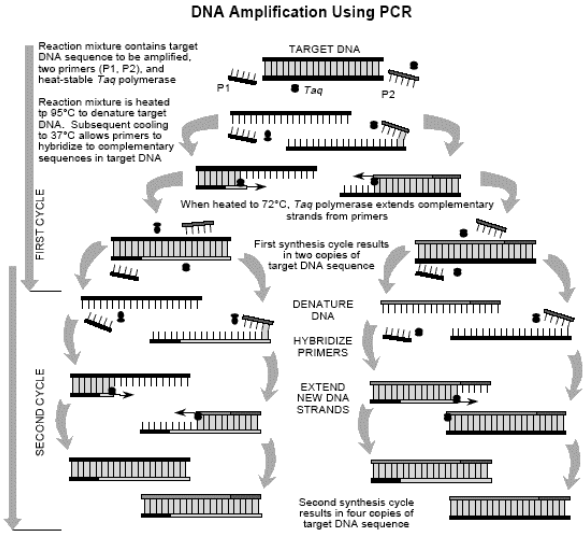
Solve for G using Metropolis Monte-Carlo driven simulated annealing

Trial move generation: $G_i \rightarrow G'_{i+1}$
Add random ([0,n]) number of sets
and
Delete random ([0,n]) number of sets

Acceptance criteria: $G_{i+1} = G'_{i+1}$
$|G'_{i+1}| \leq |G_i|$
or
$\mathcal{R} \leq \exp(-[|G'_{i+1}| - |G_i|]/T)$
$\mathcal{R} \in [0,1]$

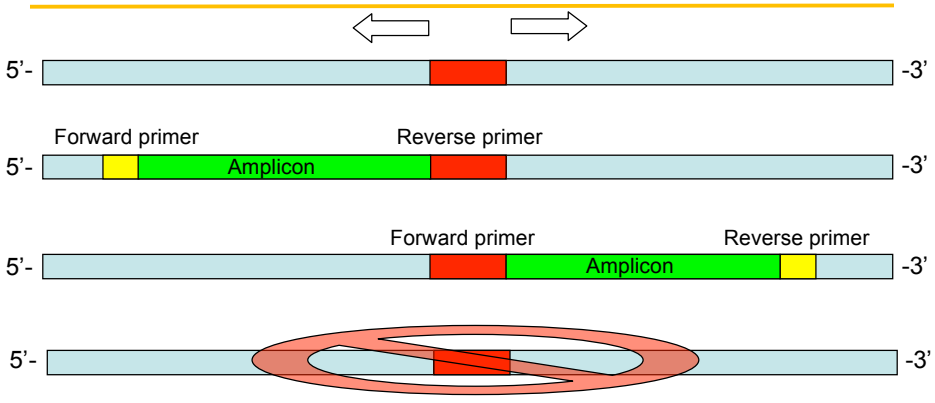
Los Alamos NATIONAL LABORATORY EST.1943
The World's Greatest Science Protecting America
NNSA

Filter signatures with experimental constraints: PCR Primers

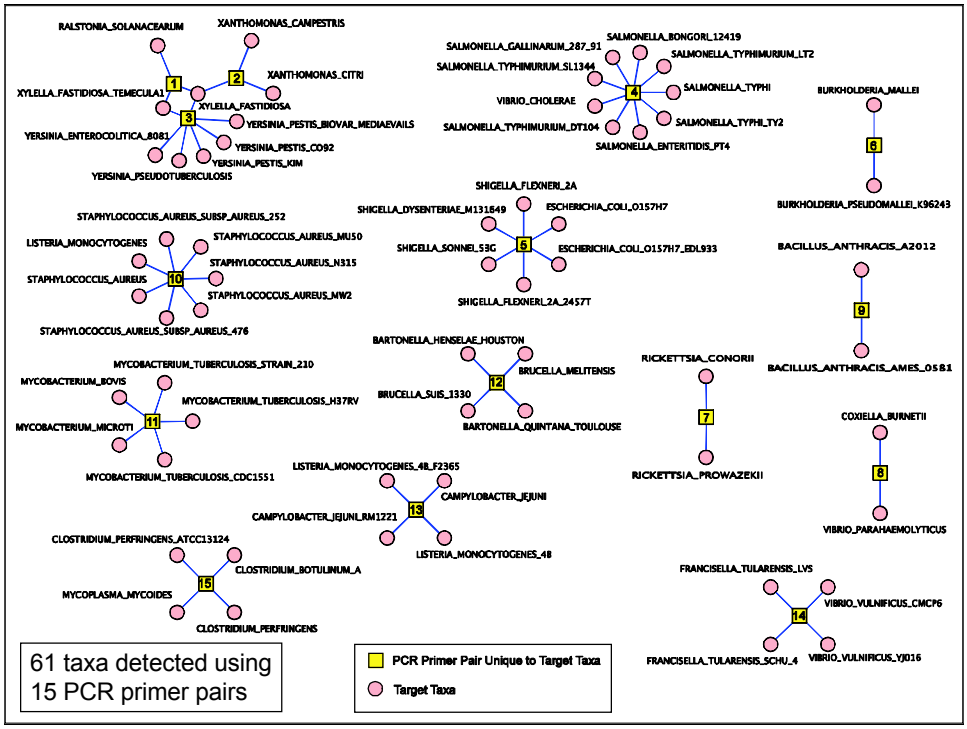


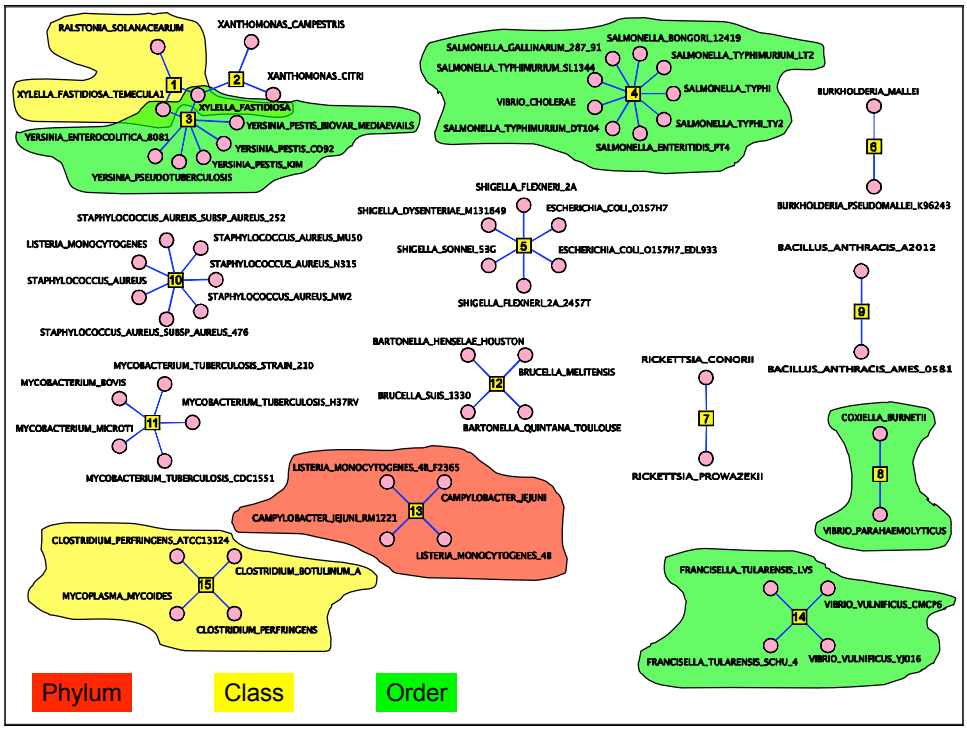
Source: *DNA Science*, From DOE Human Genome pamphlet, *Primer of Molecular Genetics*

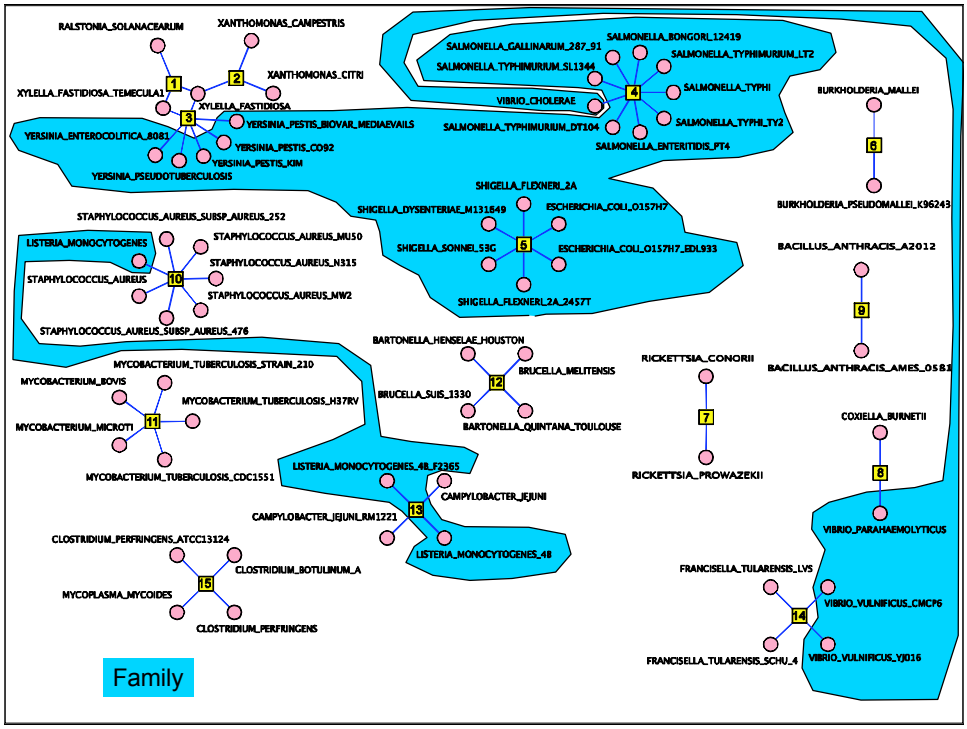
Filter signatures with experimental constraints: PCR Primers

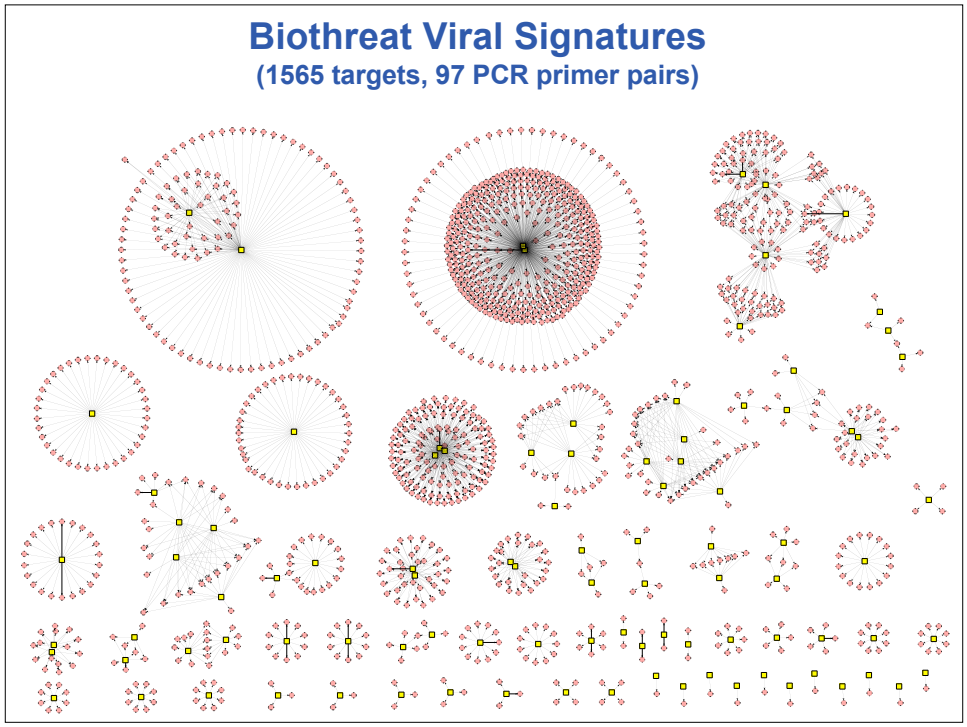


Los Alamos NATIONAL LABORATORY EST. 1943
The World's Greatest Science Protecting America
NNSA

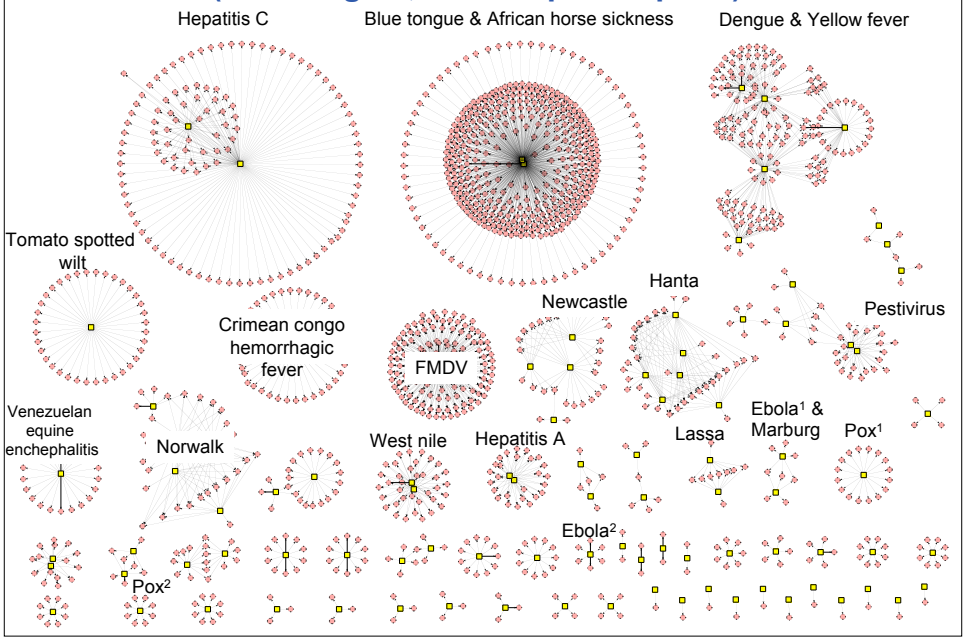

RALSTONIA_SOLANACEARUM
XANTHOMONAS_CAMPESTRIS
XANTHOMONAS_CITRI
XYLELLA_FASTIDIOSA_TEMECULA1
XYLELLA_FASTIDIOSA
YERSINIA_PESTIS_BIOVAR_MEDIAEVAILS
YERSINIA_ENTEROCOLITICA_8081
YERSINIA_PESTIS_CO92
YERSINIA_PESTIS_KIM
YERSINIA_PSEUDOTUBERCULOSIS
SALMONELLA_GALLINARUM_287_91
SALMONELLA_BONGORI_12419
SALMONELLA_TYPHIMURIUM_LT2
SALMONELLA_TYPHIMURIUM_SL1344
SALMONELLA_TYPHI
VIBRIO_CHOLERAE
SALMONELLA_TYPHIMURIUM_DT104
SALMONELLA_TYPHI_TY2
SALMONELLA_ENTERITIDIS_PT4
BURKHOLDERIA_MALLEI
BURKHOLDERIA_PSEUDOMALLEI_K96243
SHIGELLA_FLEXNERI_2A
SHIGELLA_DYSENTERIAE_M131649
ESCHERICHIA_COLI_O157H7
SHIGELLA_SONNEI_53G
ESCHERICHIA_COLI_O157H7_EDL933
SHIGELLA_FLEXNERI_2A_2457T
BACILLUS_ANTHRACIS_A2012
BACILLUS_ANTHRACIS_AMES_0581
STAPHYLOCOCCUS_AUREUS_SUBSP_AUREUS_252
LISTERIA_MONOCYTOGENES
STAPHYLOCOCCUS_AUREUS_MU50
STAPHYLOCOCCUS_AUREUS
STAPHYLOCOCCUS_AUREUS_N315
STAPHYLOCOCCUS_AUREUS_MW2
STAPHYLOCOCCUS_AUREUS_SUBSP_AUREUS_476
BARTONELLA_HENSELAE_HOUSTON
BRUCELLA_MELITENSIS
BRUCELLA_SUIS_1330
BARTONELLA_QUINTANA_TOULOUSE
RICKETTSIA_CONORII
RICKETTSIA_PROWAZEKII
COXIELLA_BURNETII
VIBRIO_PARAHAEMOLYTICUS
MYCOBACTERIUM_TUBERCULOSIS_STRAIN_210
MYCOBACTERIUM_BOVIS
MYCOBACTERIUM_TUBERCULOSIS_H37RV
MYCOBACTERIUM_MICROTI
MYCOBACTERIUM_TUBERCULOSIS_CDC1551
LISTERIA_MONOCYTOGENES_4B_F2365
CAMPYLOBACTER_JEJUNI
CAMPYLOBACTER_JEJUNI_RM1221
LISTERIA_MONOCYTOGENES_4B
CLOSTRIDIUM_PERFRINGENS_ATCC13124
CLOSTRIDIUM_BOTULINUM_A
MYCOPLASMA_MYCOIDES
CLOSTRIDIUM_PERFRINGENS
FRANCISELLA_TULARENSIS_LVS
VIBRIO_VULNIFICUS_CMCP6
FRANCISELLA_TULARENSIS_SCHU_4
VIBRIO_VULNIFICUS_YJ016
61 taxa detected using
15 PCR primer pairs
PCR Primer Pair Unique to Target Taxa
Target Taxa


Phylum
Class
Order


RALSTONIA_SOLANACEARUM
XANTHOMONAS_CAMPESTRIS
XANTHOMONAS_CITRI
XYLELLA_FASTIDIOSA_TEMECULA1
XYLELLA_FASTIDIOSA
YERSINIA_PESTIS_BIOVAR_MEDIAEVALIS
YERSINIA_ENTEROCOLITICA_8081
YERSINIA_PESTIS_CO92
YERSINIA_PESTIS_KIM
YERSINIA_PSEUDOTUBERCULOSIS
SALMONELLA_GALLINARUM_287_91
SALMONELLA_BONGORI_12419
SALMONELLA_TYPHIMURIUM_LT2
SALMONELLA_TYPHIMURIUM_SL1344
SALMONELLA_TYPHI
VIBRIO_CHOLERAE
SALMONELLA_TYPHI_TY2
SALMONELLA_TYPHIMURIUM_DT104
SALMONELLA_ENTERITIDIS_PT4
BURKHOLDERIA_MALLEI
BURKHOLDERIA_PSEUDOMALLEI_K96243
SHIGELLA_FLEXNERI_2A
SHIGELLA_DYSENTERIAE_M131649
ESCHERICHIA_COLI_O157H7
SHIGELLA_SONNEI_53G
ESCHERICHIA_COLI_O157H7_EDL933
SHIGELLA_FLEXNERI_2A_2457T
STAPHYLOCOCCUS_AUREUS_SUBSP_AUREUS_252
LISTERIA_MONOCYTOGENES
STAPHYLOCOCCUS_AUREUS_MU50
STAPHYLOCOCCUS_AUREUS_N315
STAPHYLOCOCCUS_AUREUS
STAPHYLOCOCCUS_AUREUS_MW2
STAPHYLOCOCCUS_AUREUS_SUBSP_AUREUS_476
BACILLUS_ANTHRACIS_A2012
BACILLUS_ANTHRACIS_AMES_0581
BARTONELLA_HENSELAE_HOUSTON
BRUCELLA_MELITENSIS
BRUCELLA_SUIS_1330
BARTONELLA_QUINTANA_TOULOUSE
RICKETTSIA_CONORII
RICKETTSIA_PROWAZEKII
COXIELLA_BURNETII
MYCOBACTERIUM_TUBERCULOSIS_STRAIN_210
MYCOBACTERIUM_BOVIS
MYCOBACTERIUM_TUBERCULOSIS_H37RV
MYCOBACTERIUM_MICROTI
MYCOBACTERIUM_TUBERCULOSIS_CDC1551
LISTERIA_MONOCYTOGENES_4B_F2365
CAMPYLOBACTER_JEJUNI
CAMPYLOBACTER_JEJUNI_RM1221
LISTERIA_MONOCYTOGENES_4B
VIBRIO_PARAHAEMOLYTICUS
CLOSTRIDIUM_PERFRINGENS_ATCC13124
CLOSTRIDIUM_BOTULINUM_A
MYCOPLASMA_MYCOIDES
CLOSTRIDIUM_PERFRINGENS
FRANCISELLA_TULARENSIS_LVS
VIBRIO_VULNIFICUS_CMCP6
FRANCISELLA_TULARENSIS_SCHU_4
VIBRIO_VULNIFICUS_YJ016
Family

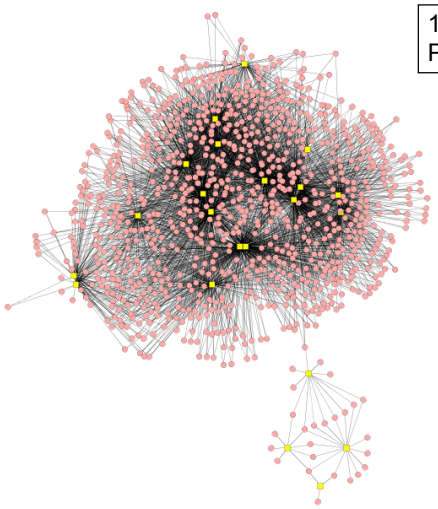

Biothreat Viral Signatures
(1565 targets, 97 PCR primer pairs)

Biothreat Viral Signatures
(1565 targets, 97 PCR primer pairs)



PCR Primers and TaqMan® Probes: HIV-1

1005 strains detected with 22 PCR primer/probe sets

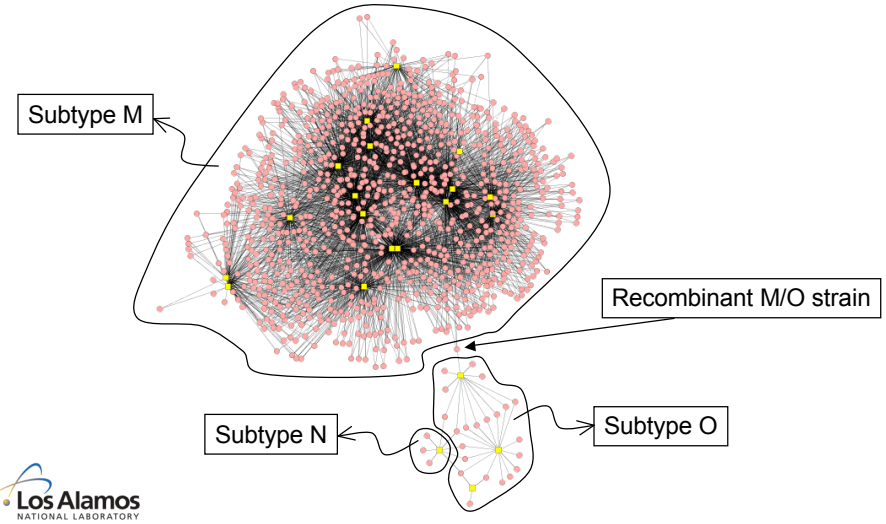


Los Alamos NATIONAL LABORATORY EST. 1943

The World's Greatest Science Protecting America

NNSA

PCR Primers and TaqMan® Probes: HIV-1



Los Alamos NATIONAL LABORATORY EST.1943
The World's Greatest Science Protecting America
NNSA

Acknowledgements

Informatics Team
Byron Delaney
Norman Dogget
Rob Leach
Jian Song
Chris Stubben
Murray Wolinsky*

mpiBLAST
Lucas Carey
Aaron Darling
Wu-Chun Feng

Experimental Validation
John Dunbar*
Lance Green
Scott White

Funding
DHS
TSWG

Los Alamos NATIONAL LABORATORY EST.1943
The World's Greatest Science Protecting America
NNSA