To ‘Finish’ or not to ‘Finish’-the $64K question in Genomics - data mining using 454 draft sequences

Shanmuga Sozhamannan, Ph. D

Genomics Department

Biological Defense Research Directorate

Naval Medical Research Center

Silver Spring, MD

Disclaimer

The views expressed in this presentation are those of the authors and do not necessarily reflect the official policy or position of the Navy, Department of Defense, nor the U.S. Government.

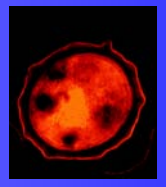
Research Focus

- Understand the biology of biodefense pathogens and near neigbors using genomic / metagenomic approaches such as *rapid DNA sequencing*

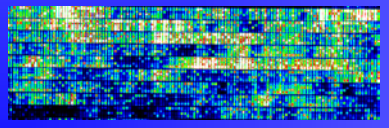
  and *gene/phenotypic expression profiling*
- Use this knowledge for *rapid, definitive identification of these* biothreat agents, and *identification of targets* for detection, diagnostics and therapeutics for these agents

The Digital Strain Collection Concept

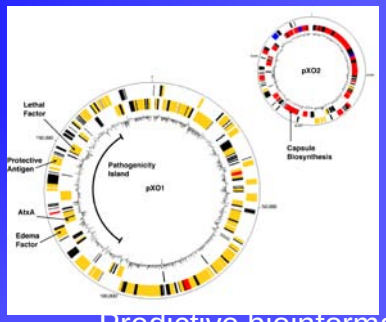
strains



microarray screen



454 sequencer



Predictive bioinformatics

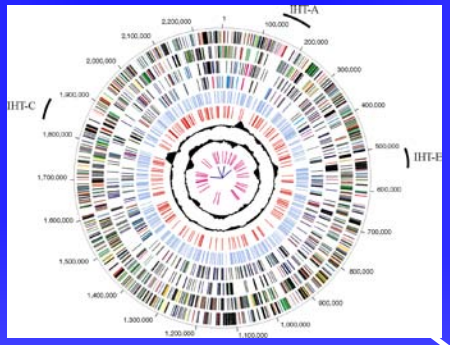
QuickTime™ and a
TIFF (Uncompressed) decompressor
are needed to see this picture.

phenotypic arrays

Population Genomics: the gateway to multiple research avenues

New drug targets

Vaccine development



Detection/ diagnosis

Genetically modified threats

Virulence mechanisms

DNA sequencing of bacterial genomes

- What we have done: all sequencing to draft std
  - *Bacillus cereus- 92*
  - *Yersiniae* spp- 60
  - *Francisellae - 6*
  - *Burkholderia - 15*
  - *Yersinia pestis* KIM deletion mutants -6
  - *Sterne* variants-12
  - *Rickettsiae* strains - 6
  - *Acenitobacter - 3*

To ‘Finish’ or not to ‘Finish’ Genomes

The answer is: it depends on what you want to get out of genome sequences?

From a biodefense standpoint, you need to know whether a biothreat agent or an outbreak pathogen is

A Known or Unknown organism

Has Known or unknown phenotypes

A Natural variant or genetically engineered or synthetic

An Unculturable organism

Need to know quickly to institute counter measures to prevent fatalities !!

Identifying a known threat organism rapidly is relatively easy given the various

diagnostics/ detection methods available

Nature of the agent: bacteria, Virus, toxins, and other infectious agents

Methods

Microbiological:

Nucleic acid based: PCR, qPCR, microarray etc.,

Immuno based

Time and cost varies with technologies

Not all field deployable

Why and when we need sequencing based ID?

Known organism with a new phenotype

A BD agent with a new phenotype such as a new toxin or Cipro resistant *B. anthracis*

A new organism with a new phenotype

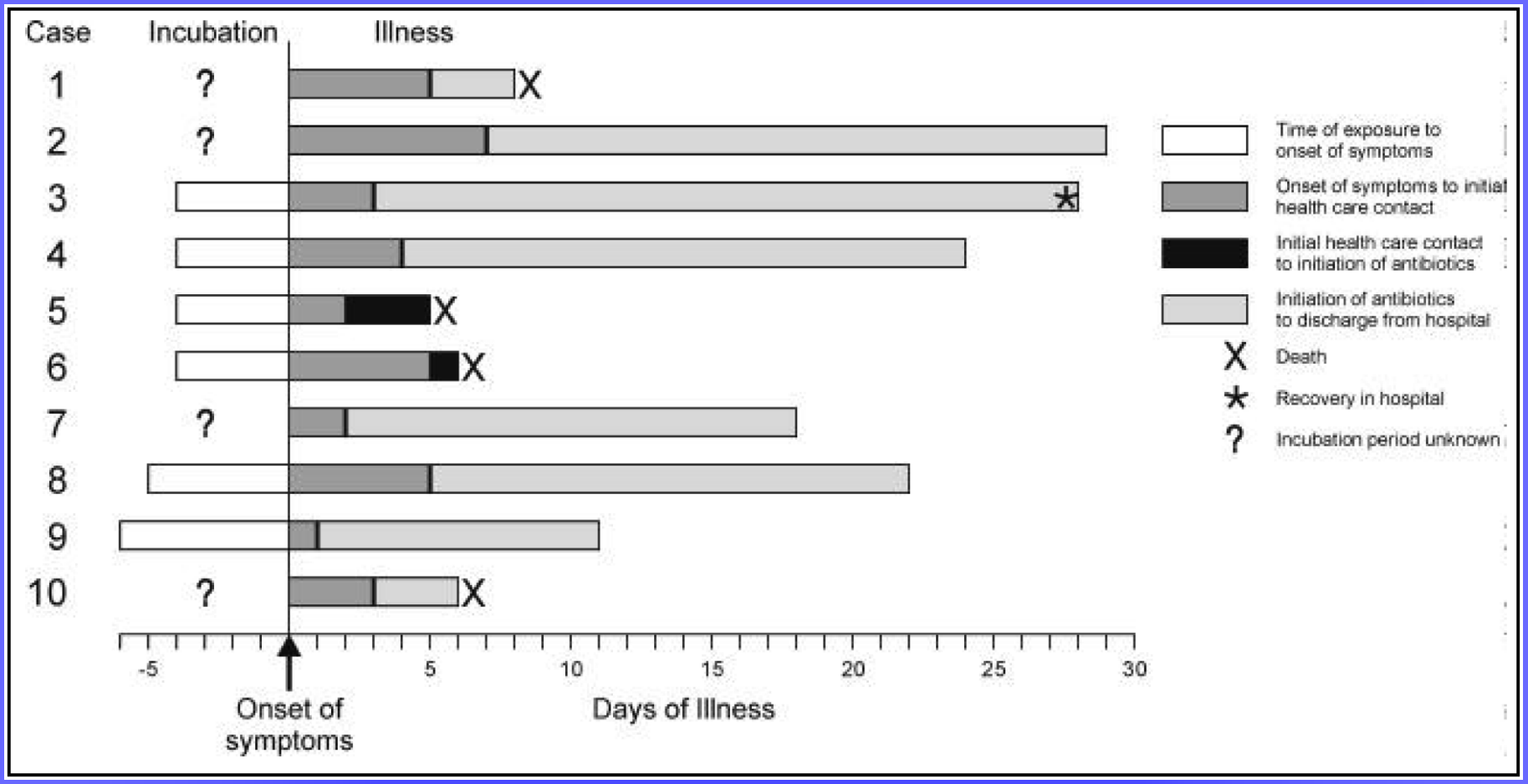
Naturally occuring B. cereus causing anthrax like disease (*B. cereus* G9241, Côte d'Ivoire (termed *B. anthracis* CI) and Cameroon (termed *B. anthracis* CA)

Genetically engineered ???

Anthrax

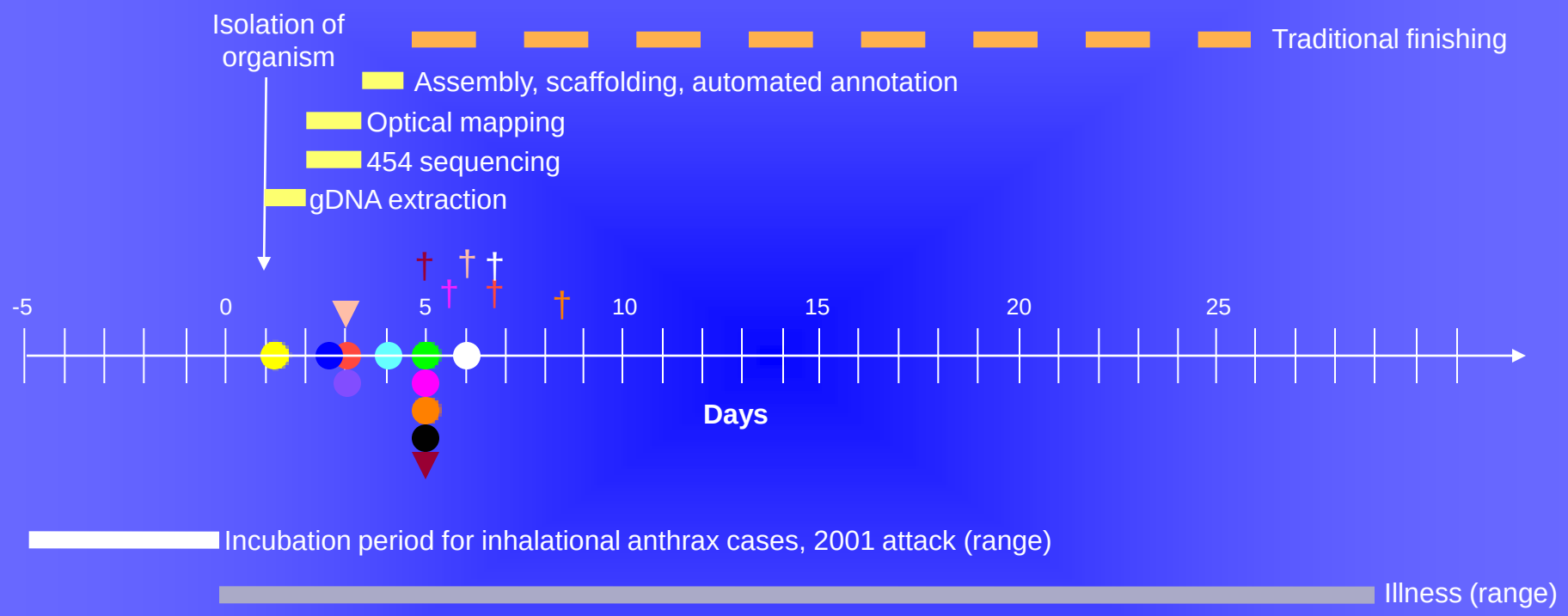
- Inhalational anthrax generally occurs after an incubation period of 1 to 6 days
- Estimated ID50 of 8,000-10,000 spores
- Death is universal in untreated cases and may occur in as many as 95% of treated cases if therapy is begun more than 48 hours after the onset of symptoms.
- Post exposure therapies and vaccinations
- Many experts consider ciprofloxacin (400 mg intravenously (i.v.) q 12 h) the drug of choice for treating victims of terrorism or warfare.
- Doxycycline (100 mg i.v. q 12 h) is an acceptable alternative
- Vaccines are available

Anthrax infection and disease progression time line



Jernigan *et al* Nov-Dec 2001. Bioterrorism-Related Inhalational Anthrax: The First 10 Cases Reported in the United States Emerging Infectious Diseases. Vol. 7, No (6): 933-944.

Timelines of Genome Sequencing and Finishing in relation to anthrax disease progression



- Colored circles indicate when individual cases of inhalational anthrax from 2001 attack were started on antibiotic therapy (Jernigan et al 2001. Emerging Infectious Diseases 7: 933-944)
- Colored triangles indicate when 1996 Bacillus cereus fatal pneumonia cases resembling anthrax were started on antibiotic therapy (Miller et al 1997. J Clin Micro 35:504-507)
- Colored crosses (†) indicate when anthrax or B. cereus patients succumbed to infection

What level of genome finishing you need to do in order to understand the nature of a biothreat agent?

Level of Genome finishing: until we can attribute the genotype to a phenotype

In a bioterror attack do we have the time to finish a genome?

Is draft sequence good enough to identify a genetically modified organism?

Case studies to test this idea using *B. anthracis* as a model

Insertion of a drug marker?

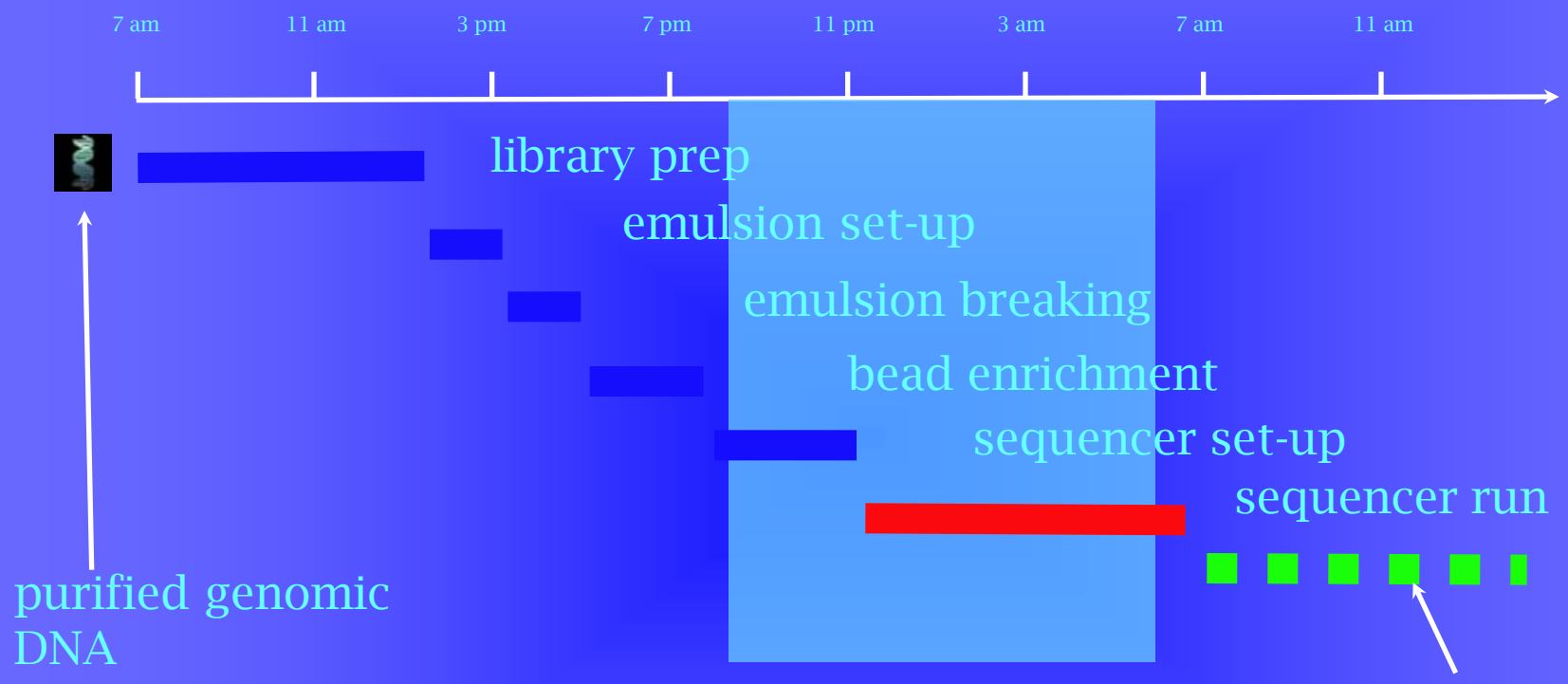
How about a SNP?

How about a phenotype of unkown genotype?

Known Knowns (known pathogen-known change)

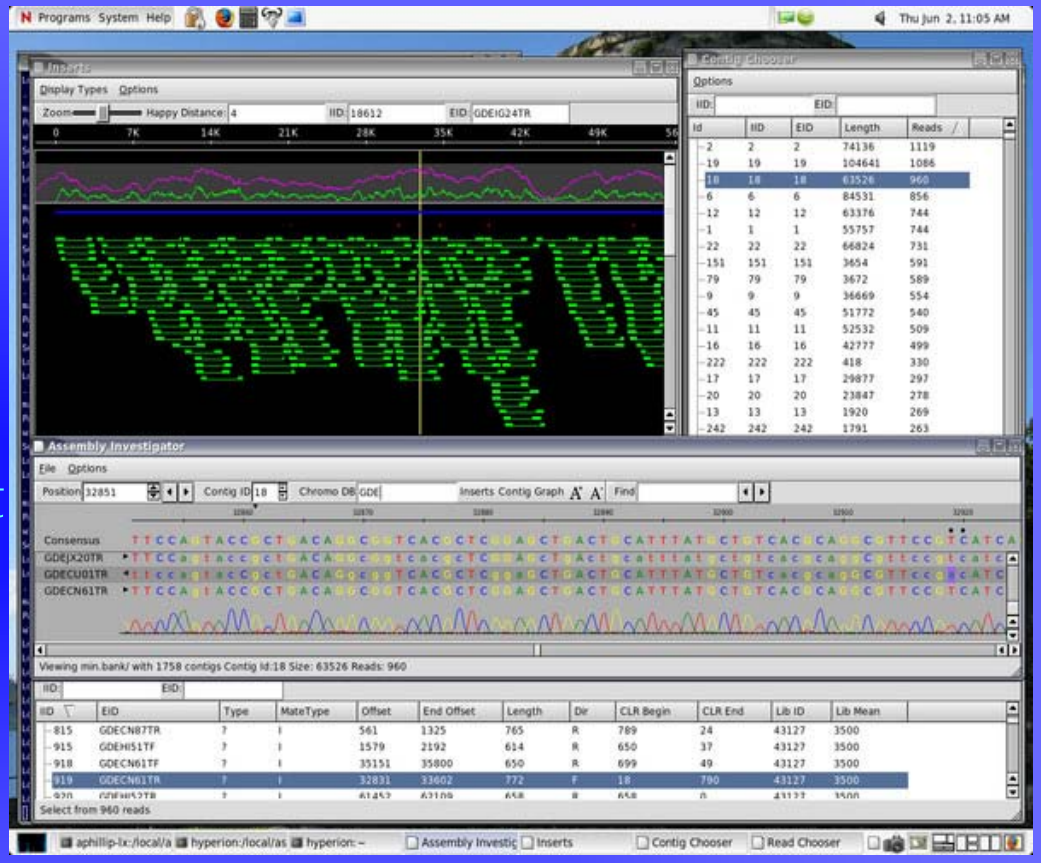
- Can we identify a known change in a known pathogen? Examples of engineered threats
  - *Bacillus anthracis Sterne X* mutant
    - How long will it take to identify an '*insertion*' in the genome?
    - In 24 hrs?

Sequencing *B. anthracis* in 24 hours



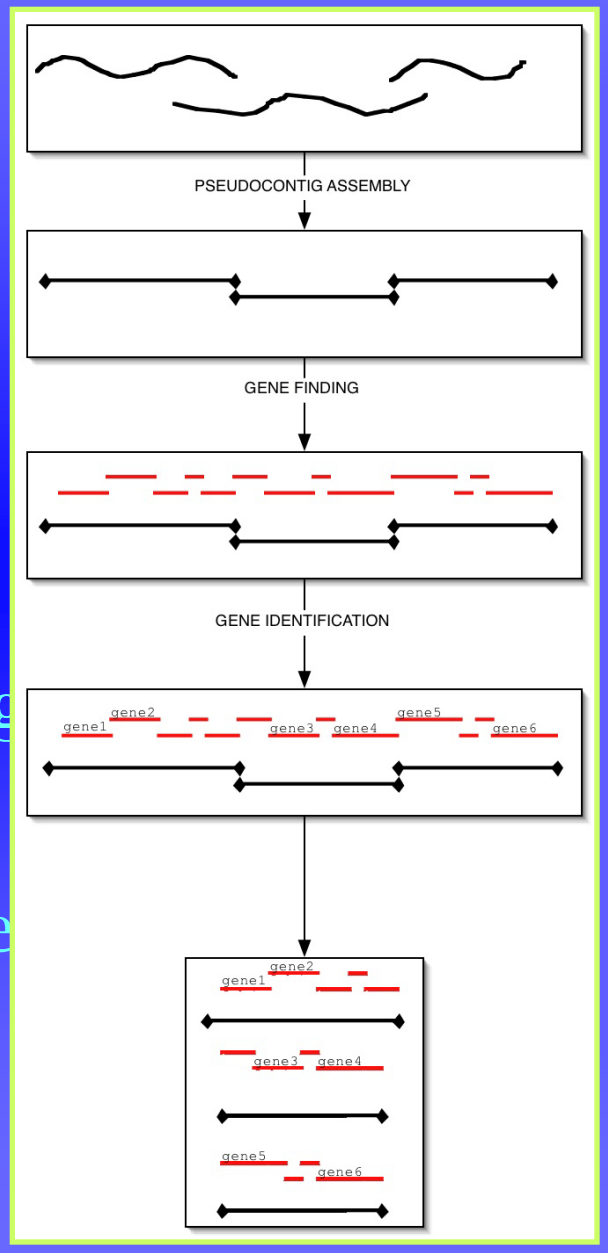
Post assembly processing

- 2 runs -709242 reads, 188,502,961 nt.
- *newbler* assembly: 30 mins
- 72 contigs, largest 889,166 nt
- 5,353,630 total length of contigs



DIYA Pipeline

- In-house bioperl based pipeline
- modular structure - can add new annotation programs
- gene ID based on blast comparisons to reference genome and NCBI protein cluster database
- run time used to be 5 - 6 hours per genome depending on modules used (with a computer cluster 45')
- With a computer cluster take about 45'

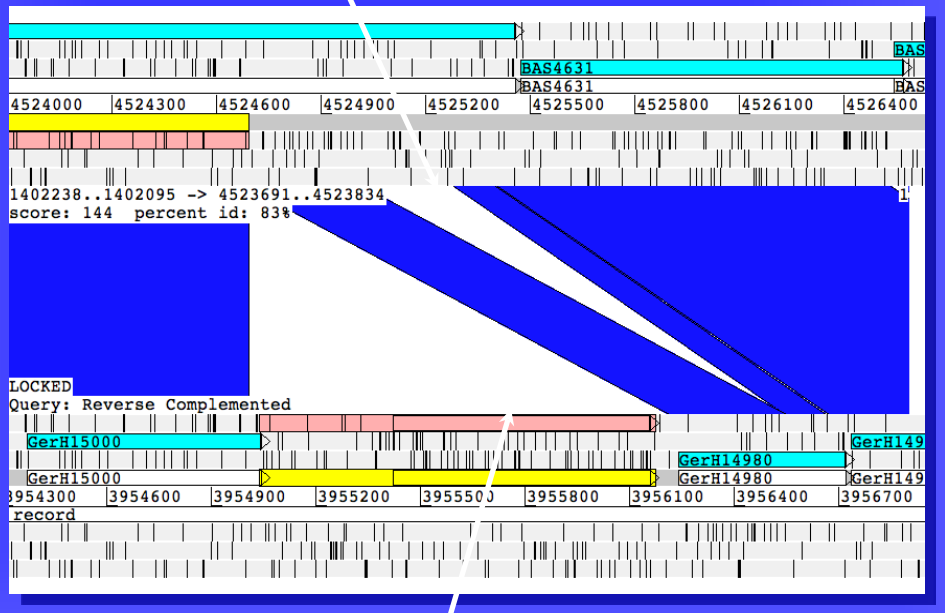


Sequenced *B. anthracis* has genetic modification

'Artemis Comparison Tool (ACT) screen shot'

deletion at the terminus of *gerH*

Ba *Sterne*



Ba Strain X

insertion of erythromycin gene

The culprit

- *gerH* null mutant - spore doesn't germinate
- used for vaccine, macrophage challenges, making antispore antiserum
- model for sequencing studies

JOURNAL OF BACTERIOLOGY, Feb. 2003, p. 1462–1464 Vol. 185, No. 4
0021-9193/03/$08.00+0 DOI: 10.1128/JB.185.4.1462–1464.2003
Copyright © 2003, American Society for Microbiology. All Rights Reserved.

Identification and Characterization of the *gerH* Operon of *Bacillus anthracis* Endospores: a Differential Role for Purine Nucleosides in Germination

Matthew A. Weiner,[1] Timothy D. Read,[2] and Philip C. Hanna[1]*

Department of Microbiology and Immunology, University of Michigan Medical School, Ann Arbor, Michigan,[1] and The Institute for Genomic Research, Rockville, Maryland[2]

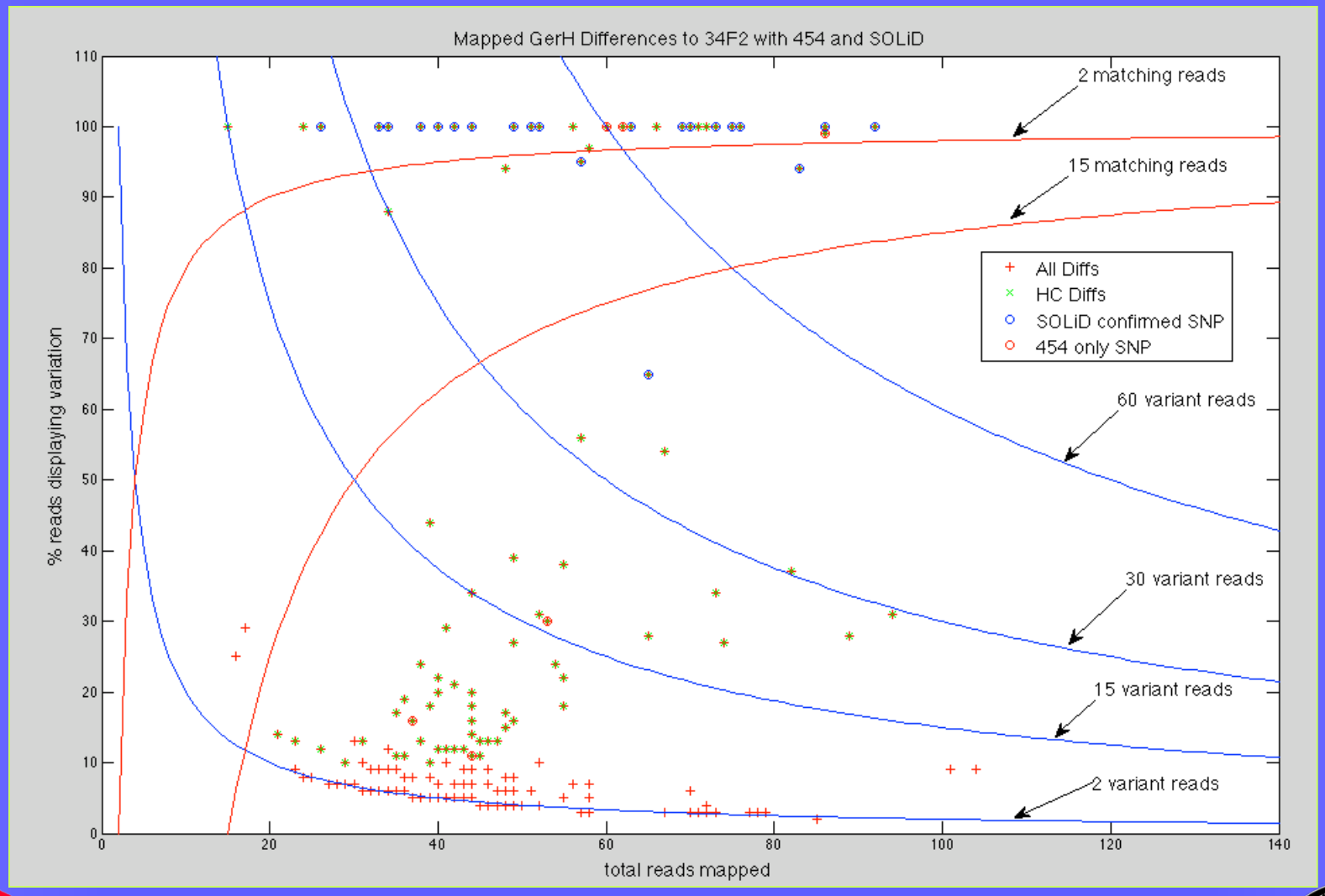
Received 19 August 2002/Accepted 18 November 2002

We identified a tri-cistronic operon, *gerH*, in *Bacillus anthracis* that is important for endospore germination triggered by two distinct germination response pathways termed inosine-His and purine-Ala. Together, the two pathways allow *B. anthracis* endospores a broader recognition of purines and amino acids that may be important for host-mediated germination.

Bacterial endospores are metabolically inactive and are capable of surviving extended periods of time under harsh environmental conditions but germinate rapidly in the presence of small molecules termed germinants (2). Work by Hachisuka demonstrated that the addition of exogenous adenosine and L-alanine was required for in vivo germination of *Bacillus anthracis* in the rat peritoneal cavity, thus establishing nucleosides as potentially contributing to in vivo germination (5). Germination is characterized by the hydration of the core and the breakup of the endospore cortex, though the molecular mechanisms underlying these activities remain undetermined (2, 9). Endospore germination results in the expulsion of Ca^{2+}

which was cloned into pUC19 (New England Biolabs) and maintained in *Escherichia coli* XL1 Blue (American Type Culture Collection). A central section of *gerHA* (Δ*gerHA*) was deleted and replaced by an erythromycin cassette, which had been amplified from pDG641 (*Bacillus* Genetic Stock Center), at the *gerHA ClaI* restriction site with appropriate primers (5′-CCATCGATGGGCGGTGTAGATGTTGATGA-3′, 5′-CCATCGATGGACATGCTACACCTCCGGATA-3′) (3). The resulting construct was transferred into the gram-positive shuttle vector pKSV7, creating pKSV7:Δ*gerHA:Erm*, and maintained in *E. coli* GM272 (*Bacillus* Genetic Stock Center) (10). Electroporation of *B. anthracis* Sterne 34F2 was performed

Surprises in *gerH* mutant !



Confirmation of very high confidence 454 diffs by SOLiD

Table5 SNP for Ger, reference: B_ANTHRACIS_STR_STERNE.fasta

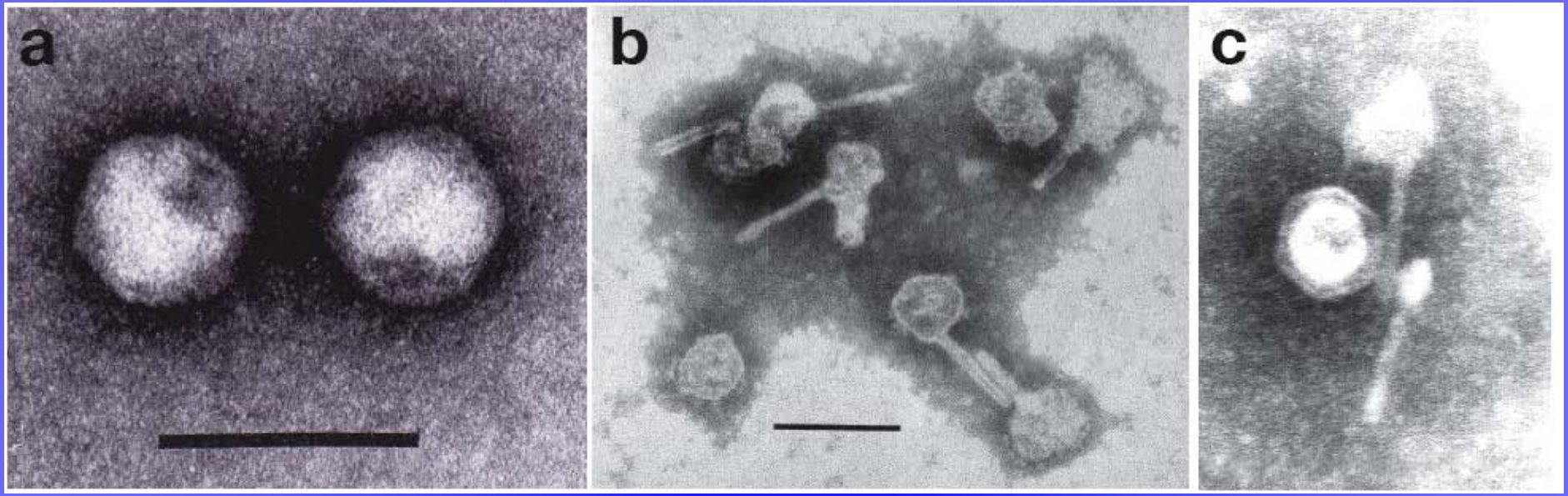
# cov	ref	consen	score	confi	Single	Paired	score	conf	Single	Paired	coord
26	G	C	0	0	0/0	0/0	1	0.9296	26/17	0/0	226566
35	T	C	0	0	0/0	0/0	0.8865	0.9333	31/18	0/0	629739
30	G	A	0	0	0/0	0/0	0.8665	0.912	26/16	0/0	1747138
29	T	C	0	0	0/0	0/0	0.9343	0.9366	27/20	0/0	2076699
20	T	C	0	0	0/0	0/0	1	0.9257	20/17	0/0	2099226
31	T	C	0	0	0/0	0/0	0.8726	0.8817	27/17	0/0	3509997
73	G	A	0	0	0/0	0/0	0.9461	0.9012	69/34	0/0	3657937
41	A	G	0	0	0/0	0/0	0.9325	0.9027	38/21	0/0	3658084
22	G	A	0	0	0/0	0/0	0.9529	0.9323	21/13	0/0	4371853
18	T	W	0.5806	0.8878	10/7	0/0	0.204	0.7799	4/3	0/0	5011559
6	T	G	0.1643	0.8289	1/1	0/0	0.6707	0.846	4/4	0/0	876972

BAS2094-heat shock protein, hsp20 family

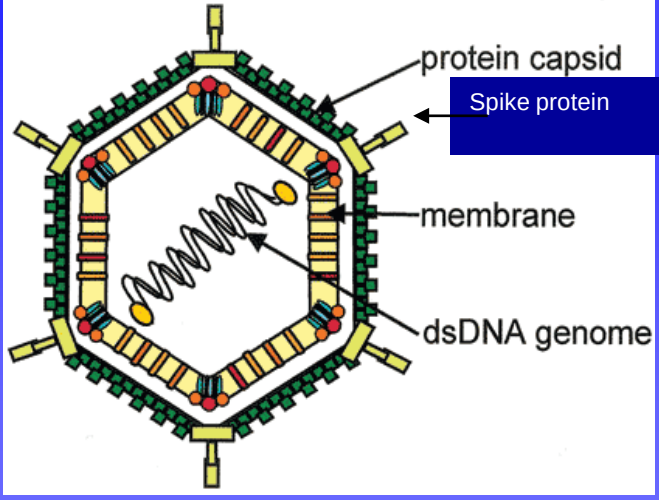
Understanding the Mechanism of phage resistance using whole genome sequencing approach- *Bacillus anthracis* phage AP50 paradigm

Known Unknown

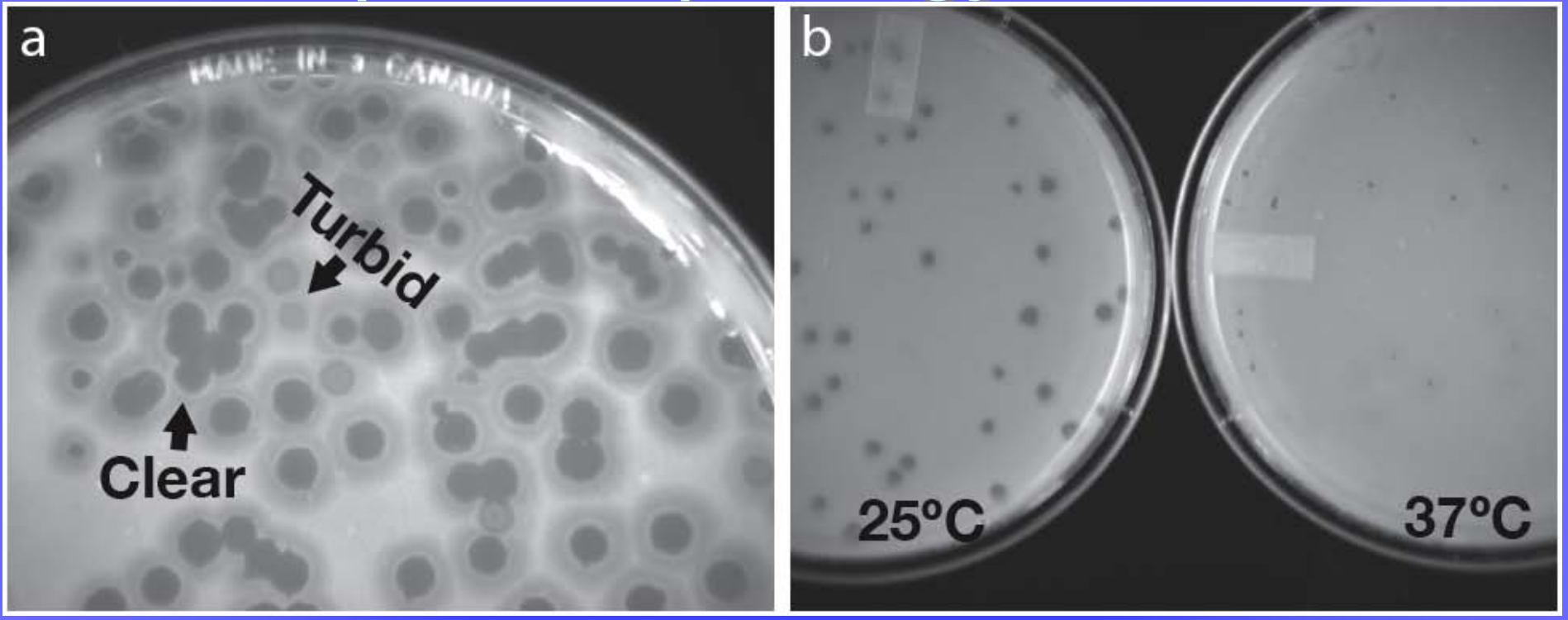
Morphology of Phage AP50



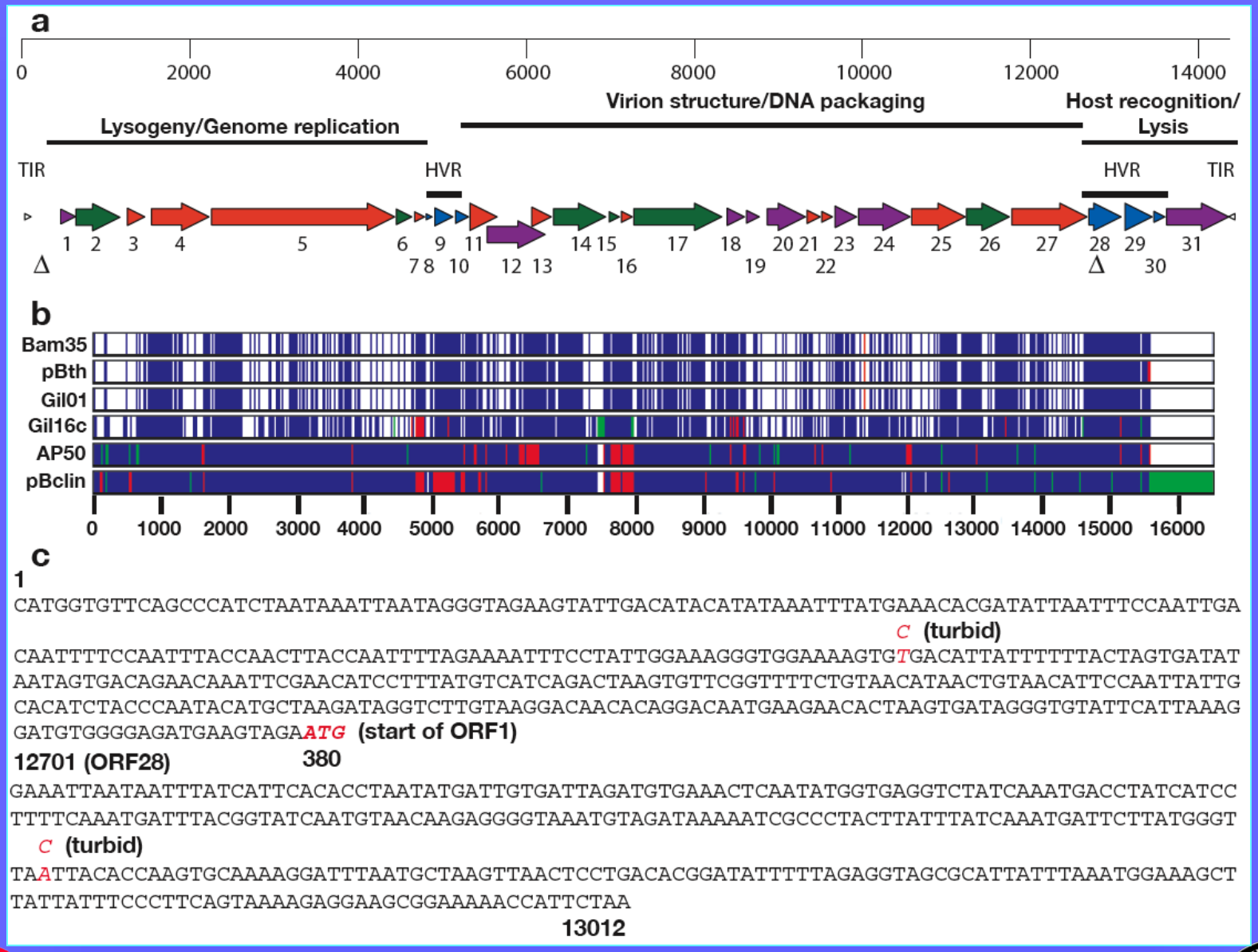
Schematic diagram of PRD-1, a Gram-negative *Tectiviridae* phage



Plaque Morphology of AP50



Appl. Environ. Microbiol. (2008) 74: 6792-6796

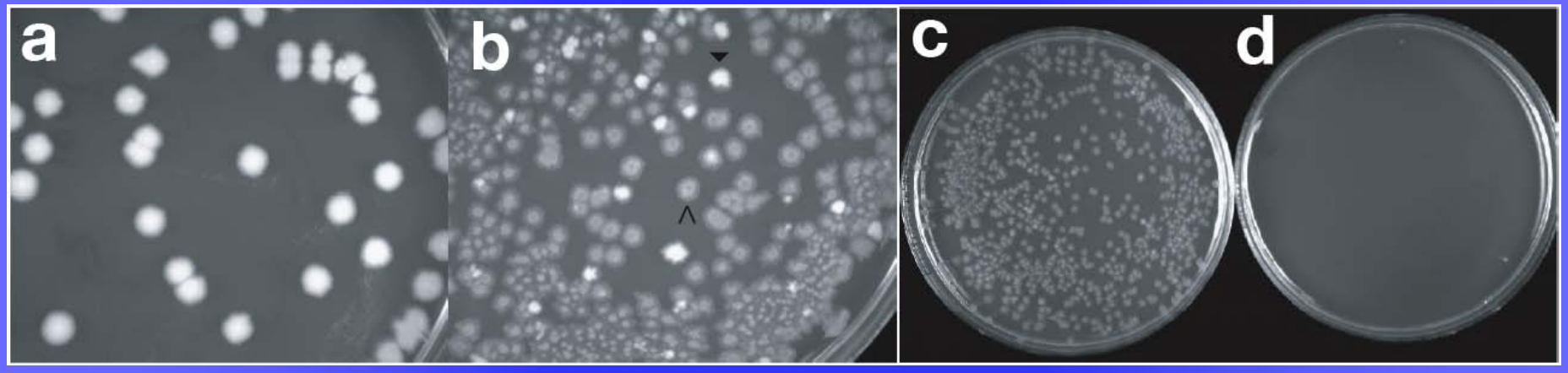


Appl. Environ. Microbiol. (2008) 74: 6792-6796

AP50 resistance mechanisms

1. Lysogeny - the first line of defense
2. Spontaneous resistance-masking the receptor
3. Mutations affecting adsorption, entry and replication of phage genome

Survivors of AP50t/AP50c infection



a - uninfected; b - AP50t infection
c - uninfected; d - AP50c infection

S. Sozhamannan

NAVAL MEDICAL RESEARCH CENTER
U.S. NAVY

TRANSFORMATIONAL MEDICAL TECHNOLOGIES INITIATIVE
Protecting the Warfighter & the Nation

Genome Sequencing of phage AP50^{R} mutants by 454

454 sequencing of Bacillus anthracis Sterne 34F2 and its AP50R derivatives

NS id	description	#runs	run type	bp total	Q39 score	read length	# large contigs	# all contig	est.average coverage (X)	mean coverage (X)	median coverage (X)	range of coverage (X)	454 HC vars (x-some)
2438	34F2	4	GS-20	121,977,190	0.20%	103.1	31	43	23	22	21	99	55
3918	34F2-AP50R1	1	FLX	62,081,067	0.45%	244.7	142	168	12	12	11	136	61
5597	34F2-AP50R 2	2	FLX	247,734,395	0.03%	270.7	60	151	46	27	25	293	59
5598	34F2 -AP50R 3	1	FLX	150,986,592	0.06%	282.8	83	150	28	28	27	377	59
5599	34F2 -AP50R 4	1	FLX	112,791,407	0.07%	282.2	89	127	21	21	20	259	61
5600	34F2 -AP50R 6	1	FLX	142,088,452	0.08%	276.9	90	146	27	26	25	287	62

loc	Ref Sterne	NS2438 34F2	NS3918 R-1	NS5597 R-2	NS5598 R-3	NS5599 R-4	NS5600 R-6	Codon wt (aa)>new codon (aa) / locus ID/ annotation
210679	T	T	T	G	T	T	T	N\A->N\A BAS0211 5' flanking gene: hypothetical protein BAS0211
226566	G	C	C	C	C	C	C	N\A->N\A BAS0225 5' flanking gene: hypothetical protein BAS0225
629739	T	C	C	C	C	C	C	GTA (V)->GTG (V) BAS0581 iron compound ABC transporter, iron compound-binding protein
876972	T	G	G	G	G	G	G	TTT (F)->TTG (L) BAS0827 hydrolase, haloacid dehalogenase-like family
894906	-	-	G	-	-	-	-	GGG (G)->GGG (G) BAS0840 CsaB protein
895064	A	A	A	A	A	-	A	ATA (I)->AT- (null) BAS0840 CsaB protein
895082	-	-	-	-	-	-	AGCTT	TTA (L)->TTA (L) BAS0840 CsaB protein
895092	-	-	-	G	G	-	-	GGG (G)->GGG (G) BAS0840 CsaB protein
1373973	-	T	-	T	T	T	T	TTT (F)->TTT (F) BAS1340 proton/glutamate symporter protein, N-terminus
1423605	TGAG	TGAG	TGAG	TGAG	GTGGA	TGAG	TGAG	N\A->N\A BAS1391 5' flanking gene: hypothetical protein BAS1391
1423608	G	G	G	-	G	G	G	N\A->N\A BAS1391 5' flanking gene: hypothetical protein BAS1391
1727727	G	G	A	A	A	A	A	N\A->N\A BAS1707 5' flanking gene: dehydrogenase, putative
1747138	G	A	A	A	A	A	A	GAG (E)->GAA (E) BAS1726 acetyl-CoA hydrolase/transferase family protein
1962551	G	-	-	-	-	-	-	N\A->N\A BAS1953 5' flanking gene: hypothetical protein BAS1953
2076699	T	C	C	C	C	C	C	TTT (F)->TTC (F) BAS2066 alcohol dehydrogenase, iron-containing
2324125	T	T	-	T	T	T	T	AAA (K)->AA- (null) BAS2325 penicillin-binding protein
3014210	A	G	A	A	A	A	G	N\A->N\A BAS3040 5' flanking gene: hypothetical protein BAS3040
3014222	T	C	T	T	T	T	T	N\A->N\A BAS3040 5' flanking gene: hypothetical protein BAS3040
3014231	A	G	G	A	A	A	A	N\A->N\A BAS3040 5' flanking gene: hypothetical protein BAS3040
3014239	T	A	A	T	T	T	T	N\A->N\A BAS3040 5' flanking gene: hypothetical protein BAS3040
3014266	-	-	T	-	-	-	-	N\A->N\A BAS3040 5' flanking gene: hypothetical protein BAS3040
3014308	TTAAA	TTAAA	CTAAG	TTAAA	TTAAA	TTAAA	TTAAA	N\A->N\A BAS3040 5' flanking gene: hypothetical protein BAS3040
3014323	TA	TA	-	TA	TA	TA	TA	N\A->N\A BAS3040 5' flanking gene: hypothetical protein BAS3040
3014328	TT	TT	GTTC	TT	TT	TT	TT	N\A->N\A BAS3040 5' flanking gene: hypothetical protein BAS3040
3014345	A	A	G	A	A	A	A	N\A->N\A BAS3040 5' flanking gene: hypothetical protein BAS3040
3509997	T	C	C	C	C	C	C	TTA (L)->TTG (L) BAS3551 hypothetical protein BAS3551
3657937	G	A	A	A	A	A	A	AGC (S)->AGT (S) BAS3692 tRNA (guanine-N(1)-)-methyltransferase
3658084	A	G	G	G	G	G	G	ATG (M)->ACG (T) BAS3693 16S rRNA-processing protein
3894626	G	G	A	A	A	A	A	GGC (G)->AGC (S) BAS3946 ribulose bisphosphate carboxylase, putative
4303653	-	-	-	-	-	-	CA	GTT (V)->GGT (G) BAS4390 bacterial extracellular solute-binding protein, family 5
4371853	G	A	A	A	A	A	A	GGT (G)->GAT (D) BAS4458 small acid-soluble spore protein SspI
Unique to strain		0	7	2	2	1	2	

Color codes

- Unique to AP50cR strains — Blue letters: Phenotype confirmed by deletion of the gene (A. Fouet data)
- Different from Ref sequence-most likely errors in ref seq- 8 SNPs
- Like yellow but missing in one strain
- Unique to 34F2 lineage
- Strain specific
- Unique to two strains and in one region; Red letters: all potantial variations in one region

```
>BAS0840 CsaB gene wild type and mutant DNA sequences
                                                G
                                                ∇
GTGCGGTTAGTTTTATCAGGATATTATGGTTTTTATAATGTTGGGG ACGAAGCAATTTTGCAATCAATTATTAAAGCGCTACATGAAGAGGATCCT
ACTCTTGAGCTTGTTGTACTTTCAAATGACCCCGACTATACAAGAAAAATGTACGGTGTAGAGGCAGTAAATCGCTGGGATATAAGAGCAATTTATA
                        AGCTT            G
         Δ               ∇               ∇
AAGAAATAAAGAGAAGTAATGGCTT AATTAGCGGAGGGGG AAGCCTCCTTCAAGATAAAACGAGTATTAAAAGTATTTTGTACTATACAGGCATT
ATGCGAATTGCTCGTTTTTTGAAGAAGCCGTATTATATTTATGCGCAAGGAATTGGTCCAATTACGAAAAGACAAAATCGTTTATTAGTGAAATGGC
AAGTATCAAAAGCAGAATATATATCAGTTCGTGATGAAGATTCATTTTTGTATTTAAAAGAAATTGGTATTAAGAAGGATATTGAACTAGTTCCAGA
TCCAGTATTGGCATGTCAACCAGAAGGAATGAAGTCGGAGTGGTTGCAAAAACATTCGATTCAAGGGAAAGTAATTGCAGTTAGTGTGAGGTATTGG
GACGCAAAAGAAGATTATATGAAGAAGCTAGCAGATACGCTGAAACAATTAAAGCGTGATGGATACCATATTTTATTCGTACCAATGCATGGACCAT
TTGATCAAAATGCATCACGGGATATTATTAATTTAATGGGAGAAGAAGCTCATATGCTTCCGTATAAGCTTGATATTCATGAGAAAATTTCAATTTT
GTCGGAATGTTCGCTTCTAATCGGTATGAGACTTCATGCGCTCATACTATCTGCCGTTGCTAATATCCCTATGGTCGGTATTTCATATGATCCAAAG
ATCGATTCATTTTTACAACAAGTAAATCAGCCGATTATCGGAAACGTAGATGGTGATTGGACAGCTGAAACTTTGTATAATGTCGCAACGAAGCAAC
TAGAACAAAAAGAATATGTACAAGAAACATTGGAACAAAGAGTAGAAGAATTACGAGAACAAATTTCAACAGCGTCTCGATATATCATAAGTGACTT
GAATTCAAAAGAGTTCAAAAAAAGAGGAATGGGATCTTAA


>BAS0840 CsaB protein wild type and mutant aa sequences
Wild type      ↓                                                    ↓      ↓     ↓
MRLVLSGYYGFYNVG DEAILQSIIKALHEEDPTLELVVLSNDPDYTRKMYGVEAVNRWDIRAIYKEI KRSNGL ISGGG SLLQDKTSIKSILYY
TGIMRIARFLKKPYYIYAQGIGPITKRQNRLLVKWQVSKAEYISVRDEDSFLYLKEIGIKKDIELVPDPVLACQPEGMKSEWLQKHSIQGKVIAVSV
RYWDAKEDYMKKLADTLKQLKRDGYHILFVPMHGPFDQNASRDIINLMGEEAHMLPYKLDIHEKISILSECSLLIGMRLHALILSAVANIPMVGISY
DPKIDSFLQQVNQPIIGNVDGDWTAETLYNVATKQLEQKEYVQETLEQRVEELREQISTASRYIISDLNSKEFKKRGMGS.
Mutant 1
MRLVLSGYYGFYNVGGRSNFAINY.
Mutant 2
MRLVLSGYYGFYNVGDEAILQSIIKALHEEDPTLELVVLSNDPDYTRKMYGVEAVNRWDIRAIYKEIREVMA.
Mutant 3
MRLVLSGYYGFYNVGDEAILQSIIKALHEEDPTLELVVLSNDPDYTRKMYGVEAVNRWDIRAIYKEIKRSNGLA.
Mutant 4
MRLVLSGYYGFYNVGDEAILQSIIKALHEEDPTLELVVLSNDPDYTRKMYGVEAVNRWDIRAIYKEIKRSNGLISGGGKPPSR.
```

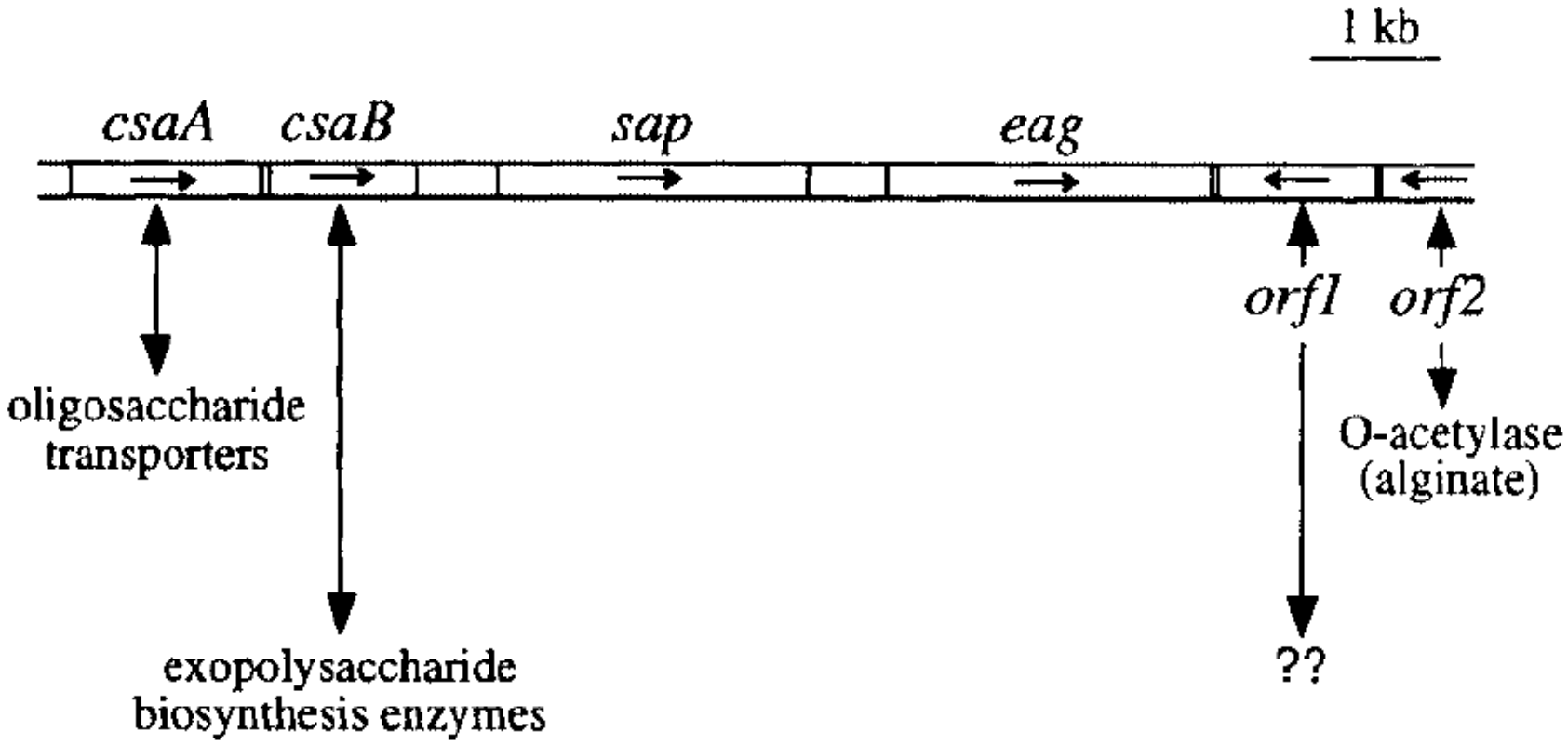


Fig. 1. Schematic diagram of the organization of the chromosomal region containing the *B.anthracis* S-layer genes *sap* and *eag*. Putative functions encoded by the ORFs flanking the S-layer genes, based on sequence comparison with protein databases, are shown.

Masnage et al 2000 EMBO J 19: 4473-4484

Agnes Fouet Lab

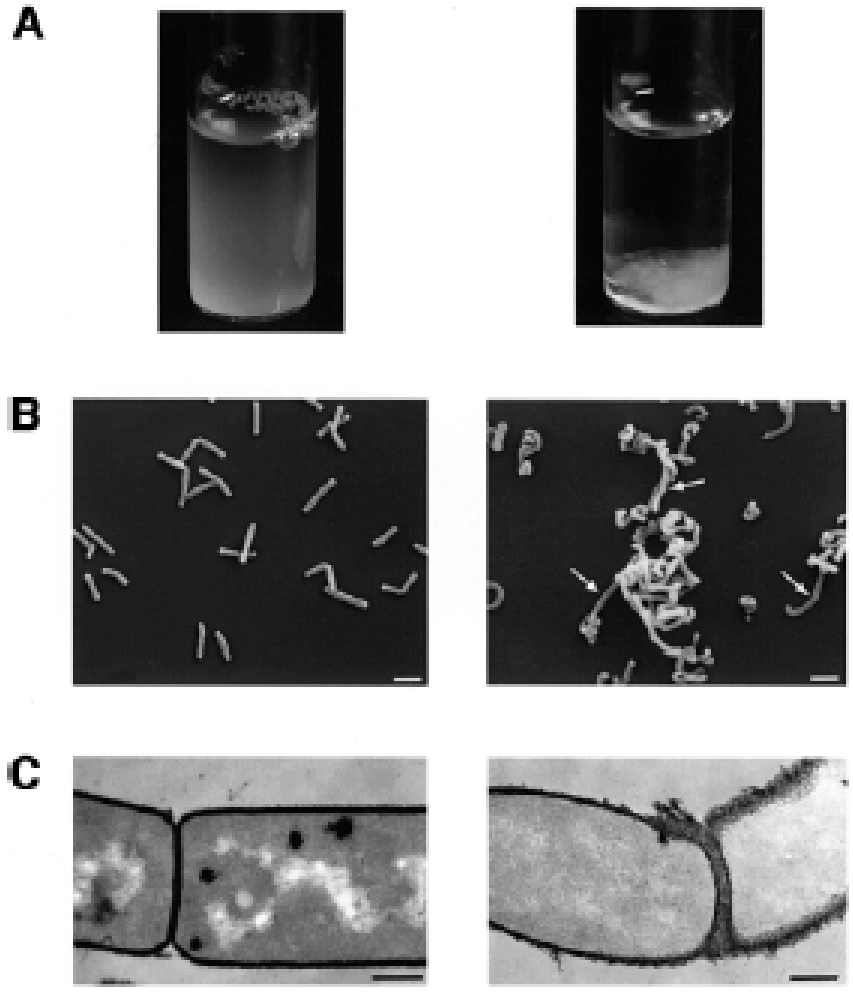


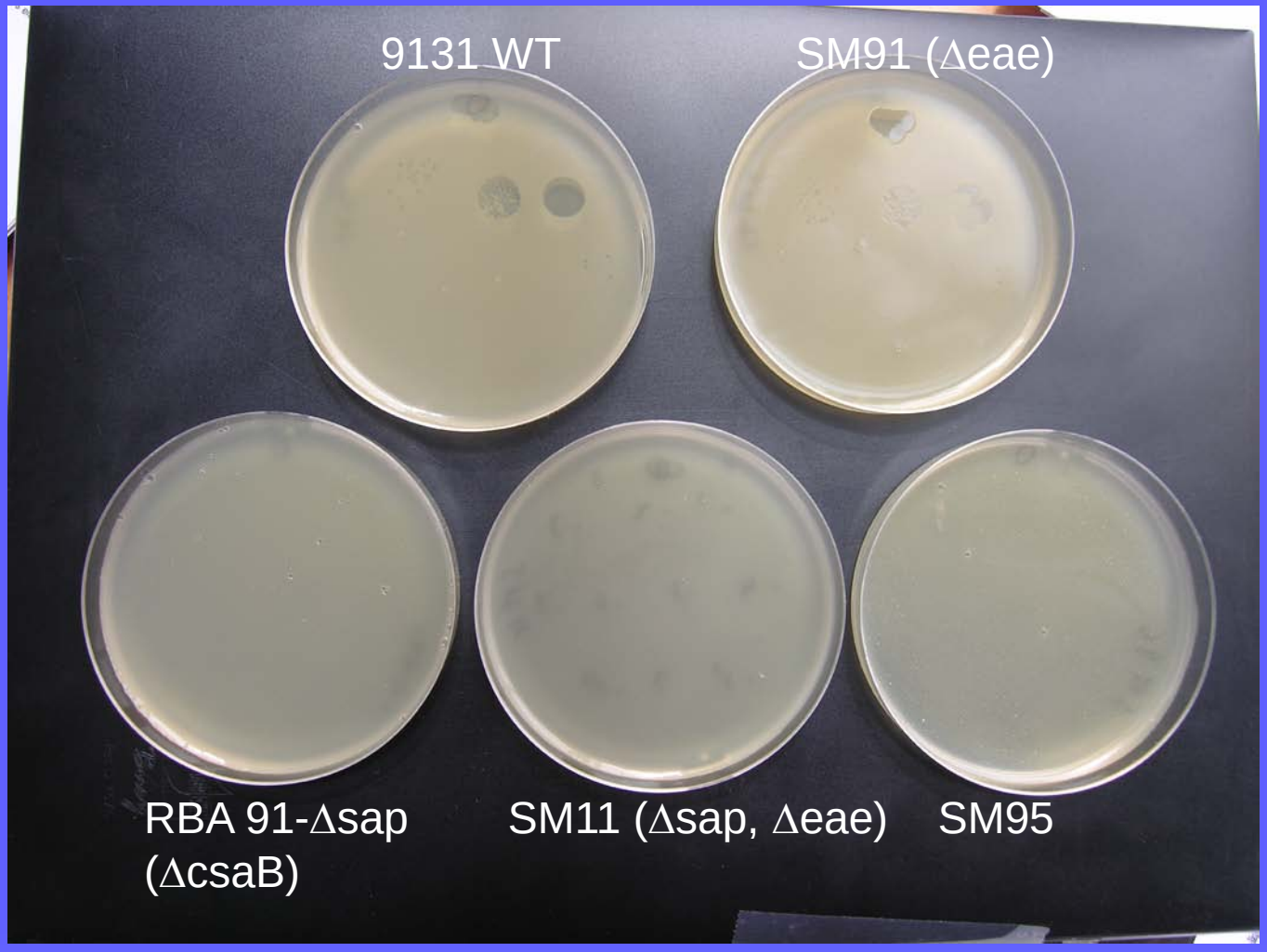
Fig. 3. SLH-anchoring deficient mutants show aberrant morphology. (A) Sedimentation profile of liquid-grown parental (left panel) and Δ*csaB* (right panel) strains. (B) Scanning electron microscopy of parental bacilli (left panel) and long twisted, septate (see arrows) filaments formed by the Δ*csaB* mutant (right panel). (C) Thin sections showed that more of cell wall material accumulated in the Δ*csaB* mutant (right panel) than in the parental strain (left panel), suggesting an autolysis defect. Bars represent 2.5 and 0.35 μm in (A) and (B), respectively.

csaB deletions have curly morphology, and accumulate cell wall material as do the spontaneous phage resistant mutants

Masnage et al 2000 EMBO J 19: 4473-4484

Agnes Fouet Lab

Deletion mutants of *sap* and *csaB* are phage AP50 resistant



A. Fouet

CONCLUSIONS

Genome sequence finishing may be necessary for understanding the absolute genome structure and content and ultimate forensic purposes

Draft sequence may be sufficient to decipher genome modifications such as indels and subtle variations (SNPs and small indels)-both natural and genetically modified

Genome sequencing may be a rapid means of mapping mutations circumventing time-consuming and laborious conventional genetic techniques such as transduction and transposon mapping

While we are trying to close the gap between finished and unfinished genomes using various strategies, we also should ask ourselves

What type of questions we can answer with draft sequences and what type of information we can extract

What type of questions we can answer only when we finish the genome or what interesting information we will miss by not finishing..

ACKNOWLEDGEMENTS

Dr. Al Mateczun-Director BDRD

Tim Read- Former Head of Genomics

Sequencing Team:

Kim Bishop-Lilly

Shannon Lentz

Kristin Willner

Amy Butani

Shakia Dorsey

Bioinformatics Team

Arya Akmal

Peter Chen

Andy Stewart

Chris Cook

Collaborators

Richard Calendar- UC Berkeley & Agnes Fouet- Inst. Pasteur, France

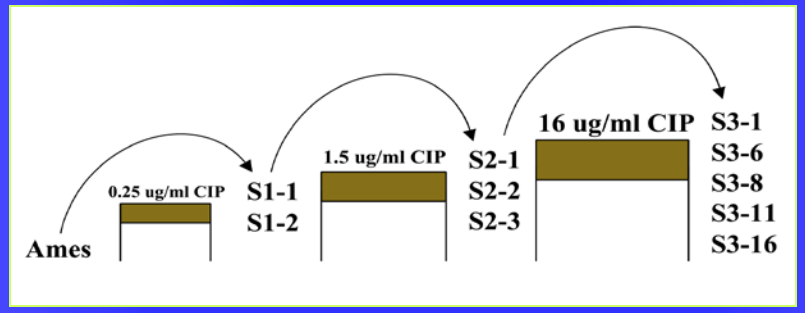
Funding- DOD/DTRA-TMTI
Special thanks to

Mr. Dave Klaasse and Dr. Paula Imbro @ TMTI

Cipro Resistance (low and medium level) in *B. anthracis*

- Study Design: - dANR (pX01$^-$, pX02$^-$) *B. anthracis* strain used as reference sequence

- low, medium, high-level cipro-resistant mutants



Phenotype	Gene containing candidate casual variant	Physical position	Reference allele	dANR allele	Low-level allele	Med-level allele	High-level allele
Low-level resistance	Gyrase A	6848	C	-	T	T	T
Med-level resistance	ParC DNA Topoisomerase	3364760	G	-	-	A	A
High-level resistance	None found						

Cipro Resistance (low and medium level) in *B. anthracis*

loc	ref	NS3907	NS3908	Annotation
6848	C	T	T	TCA (S)->TTA (L) GBAA0006 DNA gyrase subunit A
843104	C	C	G	N\A->N\A GBAA0834 5' flanking gene: TetR family transcriptional regulator
1251351	T	-	T	N\A->N\A GBAA1302 5' flanking gene: transporter, putative
1430279	A	-	A	AGC (S)->-GC (null) GBAA1510 adaptor protein
2388335	T	T	-	N\A->N\A GBAA2566 5' flanking gene: acetyltransferase
2667924	T	-	T	AAA (K)->AA- (null) GBAA2880 hypothetical protein GBAA2880
2772518	T	T	-	ACT (T)->-CT (null) GBAA3004 hypothetical protein GBAA3004
3227575	-	A	-	N\A->N\A GBAA3514 5' flanking gene: metallo-beta-lactamase family protein
3364887	G	T	A	TCC (S)->TTC (F) GBAA3656 DNA topoisomerase IV subunit A
4009233	G	G	A	CGT (R)->TGT (C) GBAA4394 stage 0 sporulation protein A
4373941	C	C	A	CTA (L)->ATA (I) GBAA4810 hypothetical protein GBAA4810
4500491	G	A	A	N\A->N\A GBAA4955 5' flanking gene: 2'-5' RNA ligase family protein, putative
4779457	G	T	T	GAT (D)->TAT (Y) GBAA5275 DNA-binding response regulator
5052174	A	-	A	TTT (F)->TT- (null) GBAA5566 hypothetical protein GBAA5566