SAND2021-11932C

Illuminate the dark matter

Gene list for a high-support 22-kbp island (Eco11885.22.S) within a tRNA gene of an E. coli genome

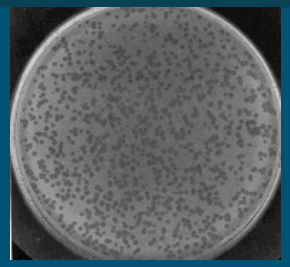
52% of genes "hypothetical" (vs. 5% for whole genome)

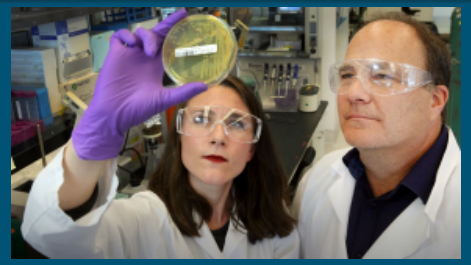
									Pfam hit
AAAAID010000002.1	Prodigal:2.6	CDS	45470	46306	.	+	0	ID=00381;product=hypothetical protein;	
AAAAID010000002.1	Prodigal:2.6	CDS	46523	46741	.	+	0	ID=00382;product=AlpA;pfam1=Phage_AlpA;	
AAAAID010000002.1	Prodigal:2.6	CDS	46823	47113	.	+	0	ID=00383;product=hypothetical protein;	
AAAAID010000002.1	Prodigal:2.6	CDS	47166	47462	.	+	0	ID=00384;product=hypothetical protein;	
AAAAID010000002.1	Prodigal:2.6	CDS	47535	47777	.	+	0	ID=00385;product=nucleoid DNA binding protein;	
AAAAID010000002.1	Prodigal:2.6	CDS	47886	48236	.	+	0	ID=00386;product=hypothetical protein;	
AAAAID010000002.1	Prodigal:2.6	CDS	48300	48704	.	-	0	ID=00387;product=DNA-binding protein H-NS	Histone_HNS
AAAAID010000002.1	Prodigal:2.6	CDS	48863	49342	.	-	0	ID=00388;product=hypothetical protein;	
AAAAID010000002.1	Prodigal:2.6	CDS	49516	50247	.	+	0	ID=00389;product=hypothetical protein;	
AAAAID010000002.1	Prodigal:2.6	CDS	50332	50913	.	-	0	ID=00390;product=hypothetical protein ;	RepL
AAAAID010000002.1	Prodigal:2.6	CDS	52267	53229	.	+	0	ID=00391;product=hypothetical protein;	
AAAAID010000002.1	Prodigal:2.6	CDS	53290	54462	.	+	0	ID=00392;product=Helix-hairpin-helix motif protein	HHH_3
AAAAID010000002.1	Prodigal:2.6	CDS	54554	54739	.	-	0	ID=00393;product=hypothetical protein;	
AAAAID010000002.1	Prodigal:2.6	CDS	54952	55425	.	-	0	ID=00394;product=hypothetical protein;	
AAAAID010000002.1	Prodigal:2.6	CDS	55858	56388	.	+	0	ID=00395;product=hypothetical protein;	
AAAAID010000002.1	Prodigal:2.6	CDS	56399	56674	.	+	0	ID=00396;product=hypothetical protein;	
AAAAID010000002.1	Prodigal:2.6	CDS	56734	57570	.	-	0	ID=00397;product=hypothetical protein;	
AAAAID010000002.1	Prodigal:2.6	CDS	57707	58216	.	+	0	ID=00398;product=hypothetical protein ;	DUF1863
AAAAID010000002.1	Prodigal:2.6	CDS	58348	59271	.	-	0	ID=00399;product=TIR domain protein	TIR
AAAAID010000002.1	Prodigal:2.6	CDS	59328	60830	.	-	0	ID=00400;product=Mobilization protein A	MobA_MobL
AAAAID010000002.1	Prodigal:2.6	CDS	61476	61781	.	+	0	ID=00401;product=Conjugal protein TraD	TraD
AAAAID010000002.1	Prodigal:2.6	CDS	62180	62737	.	+	0	ID=00402;gene=bfpA;product=Pilus precursor;	PilS;Bundlin
AAAAID010000002.1	Prodigal:2.6	CDS	62790	64181	.	+	0	ID=00403;product=Bacterial shufflon protein	Shufflon_N
AAAAID010000002.1	Prodigal:2.6	CDS	64296	66116	.	-	0	ID=00404;product=hypothetical protein;	
AAAAID010000002.1	Prodigal:2.6	CDS	66230	67450	.	-	0	ID=00405;product=integrase;	Phage_integrase

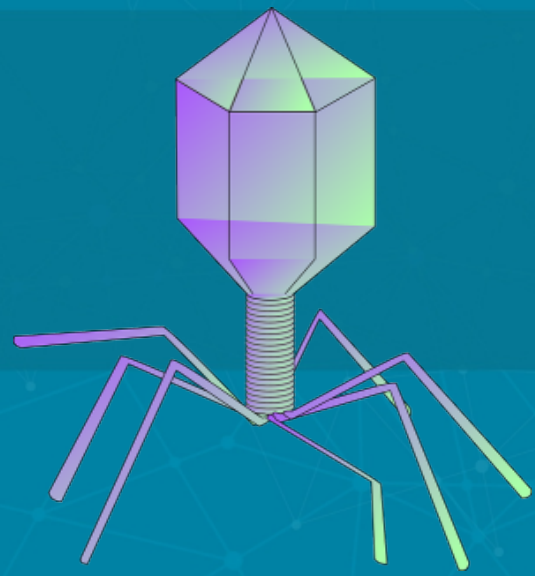
How do we turn more reds yellow?

- 9:05 Kelly Williams: finding prophages in genomes
- 9:35 Simon Roux: finding phages in metagenomes
- 10:05 Katelyn McNair: calling phage gene frames and frameshifts
- 10:15 Stephanie Malfatti: gene functions in phages
- 10:25 Break
- 10:35 Migun Shakya: finding phages in metagenomes
- 10:45 Rebecca Wattam: gene functions in phages
- 10:55 Jason Gill: gene functions in phages
- 11:25-11:45 Discussion
  - Gene Annotation
  - Phage Identification in Sequence Data
  - Catch-All

Discovery Through Precise Prophage Mapping







Kelly Williams

Sandia National Laboratories

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

So, Nat'ralists observe, a Flea
Hath smaller Fleas that on him prey,
And these have smaller yet to bite 'em,
And so proceed *ad infinitum*.

– *Swift, On Poetry: A Rapsody*

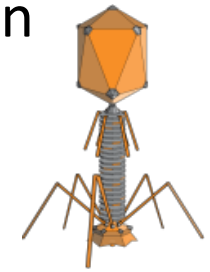
Outline

- Introduction: genomic islands, satellite/helpers, integration/excision
- Value of precise mapping of integrative DNAs
- Software that maps prophages precisely
- Discoveries
  - Integrase site specificity
  - Regulated gene integrity
  - Helper-embedded satellites
- Phage factory for therapy and energy

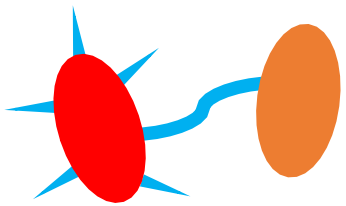
Genomic islands as part of the bacterial mobilome

Delivery vehicles

- Virion



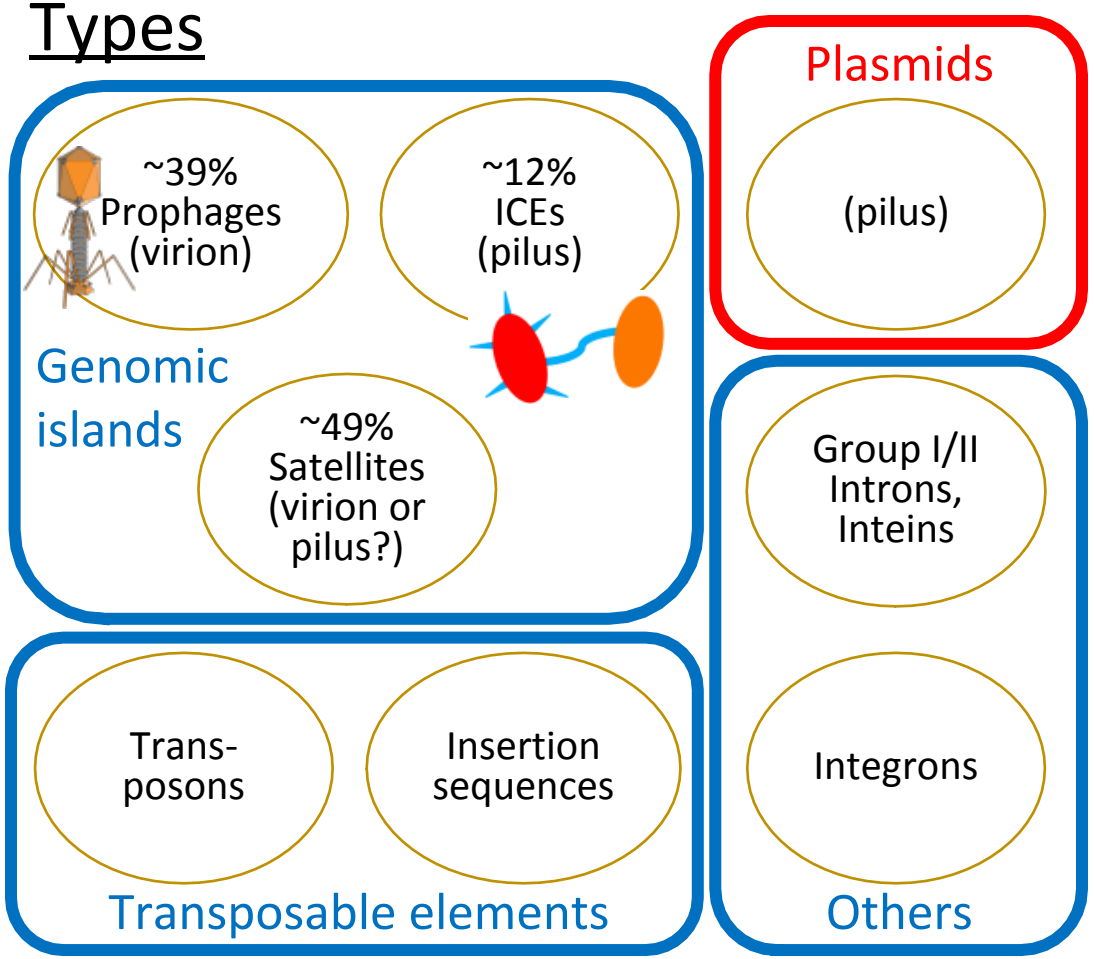
- Conjugation pilus



Stability modes

- Integration into chromosome
- Self-replicating circles

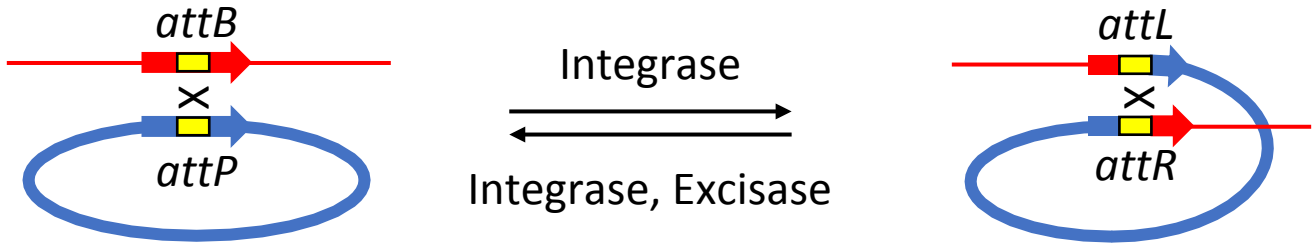
Types



Satellite/helper relationships

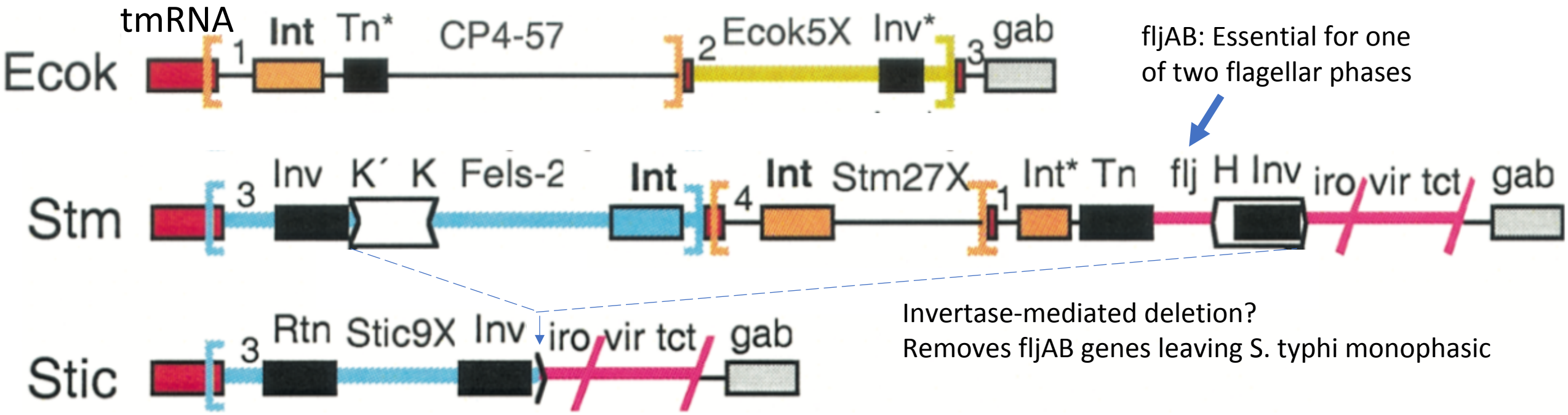
- Classic example: satellite P4 and its helper P2
- Other satellite classes: GP-PICI, GN-PICI, P
- Satellites don't fully encode their own mobility vehicle
  - Rely on helper to provide mobility gene function (virion or pilus)
  - Satellite may need a DNA site allowing virion packaging or pilus entry
- Satellites exert a reproductive cost on their helpers
- Genetic interactions can be extensive and reciprocal
  - P4 can derepress early genes and activate late genes of P2 helper, and respond in the same ways to P2-encoded regulators

Integration



- Integrases typically target a specific chromosomal site (*attB*)
  - However, off-target events can occur, and some clades are habitually site-promiscuous
- Can function even when *attB* already occupied (forming tandem islands)
- Directionality: excision typically requires a separate excisase (RDF)
- Biotech utility: more efficient at integrating big DNA than CRISPR-Cas homology directed repair
- Two main integrase protein families
  - Tyrosine integrases (Y-int)
    - Pfam Phage_integrase + idiosyncratic arm-binding domains
    - Additional requirement for host DNA-binding proteins
  - Serine integrases (S-int)
    - Domains Pfam Resolvase + Pfam Recombinase

Rearrangements at tandem islands



Outline

- Introduction: genomic islands, satellite/helpers, integration/excision
- Value of precise mapping of integrative DNAs
- Software that maps prophages precisely
- Discoveries
  - Integrase site specificity
  - Regulated gene integrity
  - Helper-embedded satellites
- Phage factory for therapy and energy

Benefits of precise prophage mapping

- Advances completion of prokaryote genome annotation
- Phage genome
  - Phage genome is complete, with one known host
  - Phage factory: perfectly mapped prophages are ready to engineer and reboot
- Integrases
  - Link each integrase protein sequence to the DNA *att* sequence it recognizes
  - Integrase evolution, biology and biotechnology
- Integration target site
  - Regulation of integration target gene
    - Some GIs break/restore key target gene by integration/excision
  - Discovery of helper-embedded satellites (HESs)

Outline

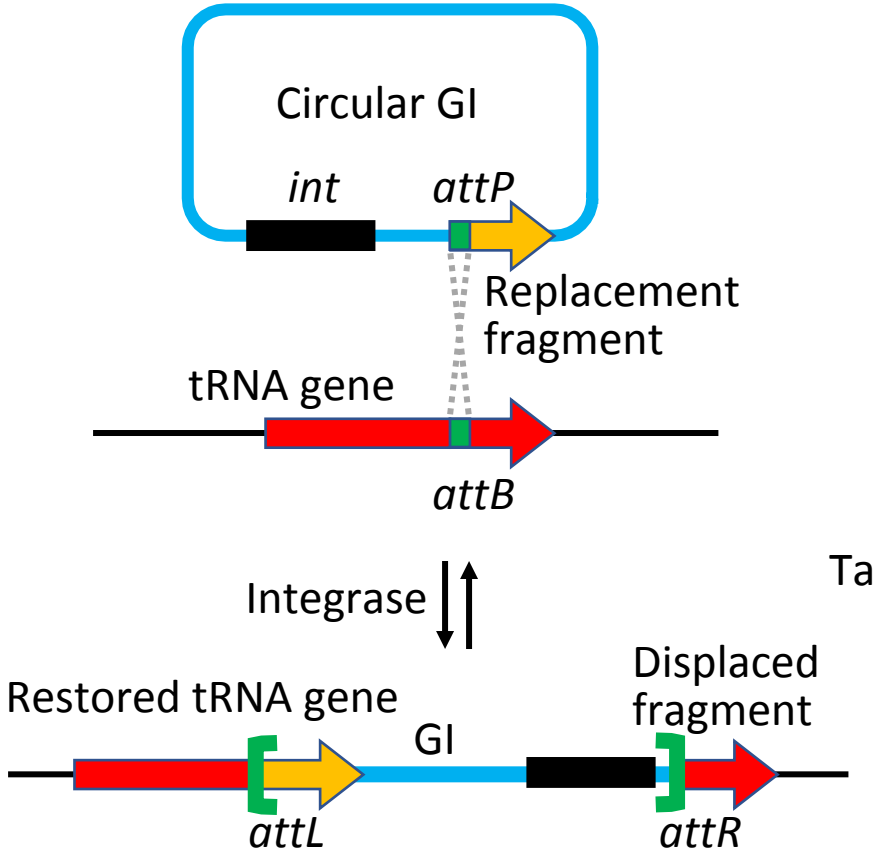
- Introduction: genomic islands, satellite/helpers, integration/excision
- Value of precise mapping of integrative DNAs
- Software that maps prophages precisely
- Discoveries
  - Integrase site specificity
  - Regulated gene integrity
  - Helper-embedded satellites
- Phage factory for therapy and energy

Software for precise GI mapping

- Islander
  - Finds islands in tRNA or tmRNA genes (~40% of Gis)
- TIGER
  - Comparative, finds reference genomes where the island site is unoccupied
  - Support value is the number of such reference genome
  - Precise mapping by ping-pong BLAST

Islander

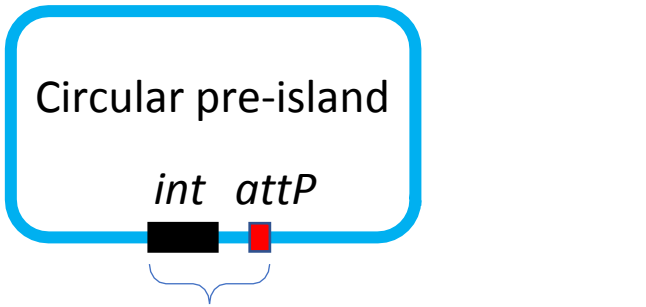
Integrase preference for t(m)RNA genes:
~40 % of islands are in t(m)RNA genes,
which account for only ~0.1% of a genome

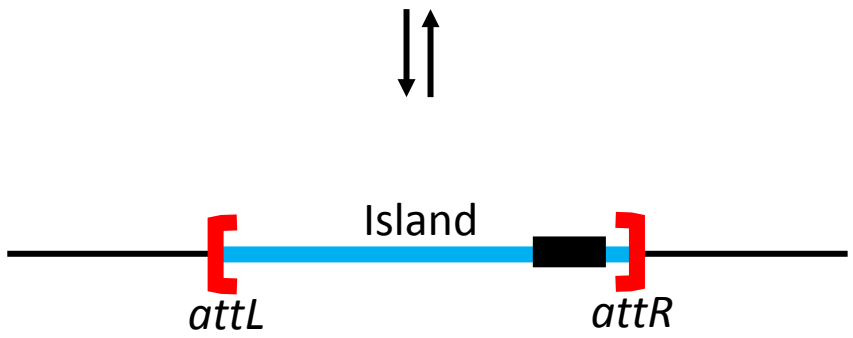


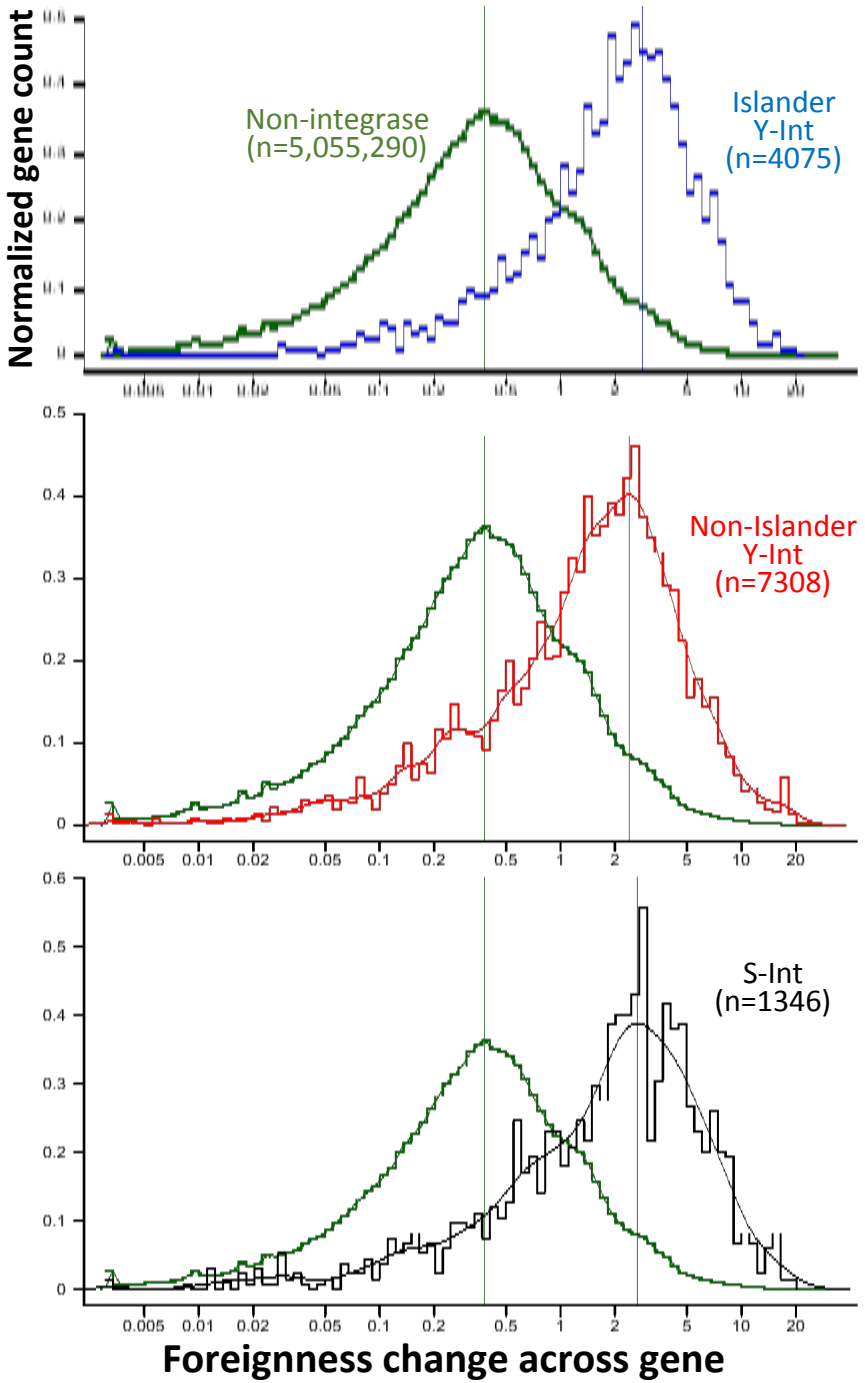
Target gene is split, yet "repaired"

Bioinformatic signature: Gene and its co-oriented fragment

Integration module cohesion: integrase gene stays near attP, therefore found at island terminus

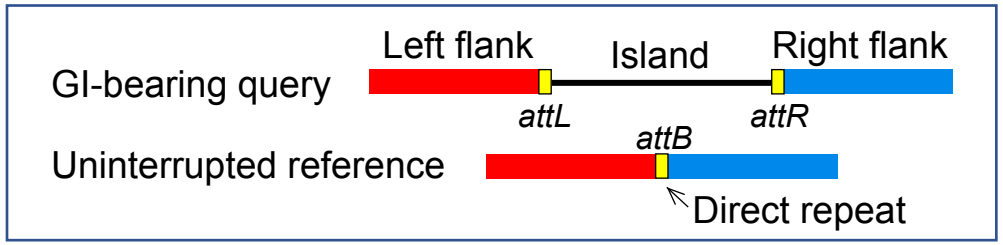




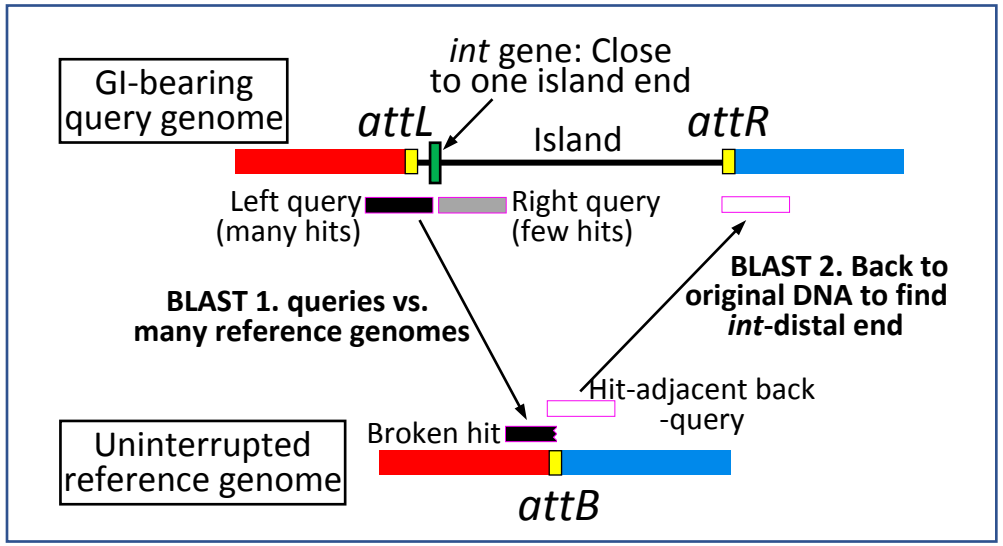


TIGER software: comparative and precise

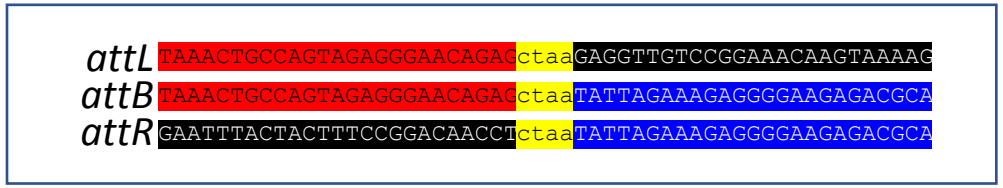
Comparison



Algorithm

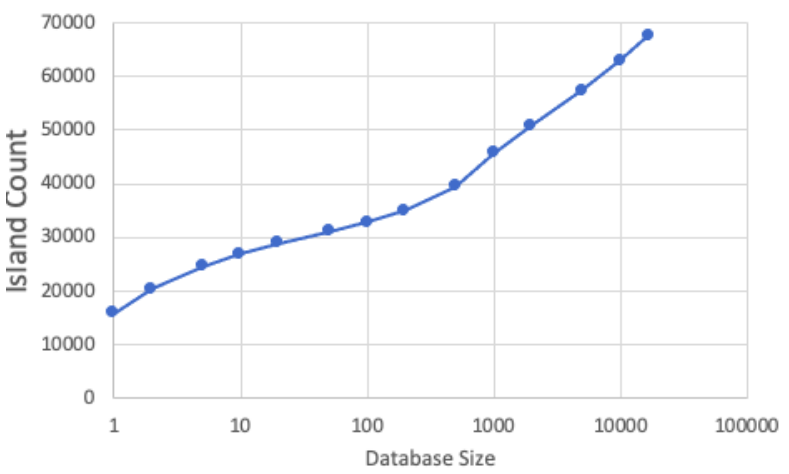


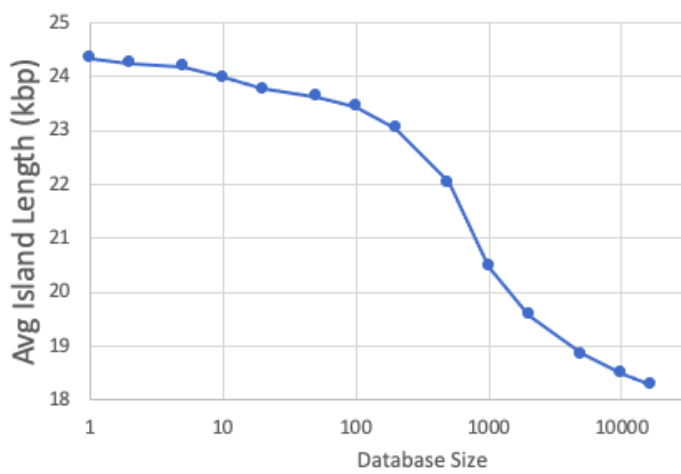
Precision

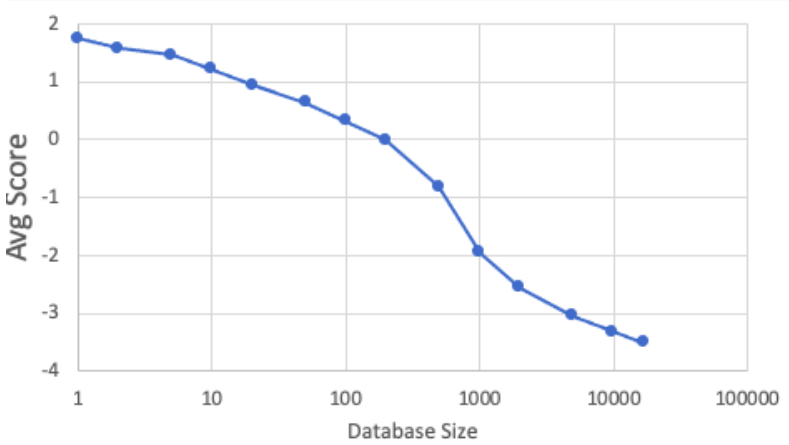


TIGER speed-up with smarter reference genome databases

- Collaboration with Noushin Ghaffari, Fatema Shormin and Bernard Nyarko (PVA&MU)
- Study set: 16790 *E. coli* genomes and their islands
- Measure pairwise distances between all genomes, as basis for scientific design of smaller genome databases
- Measure effects of smaller databases, beginning with randomly-selected memberships



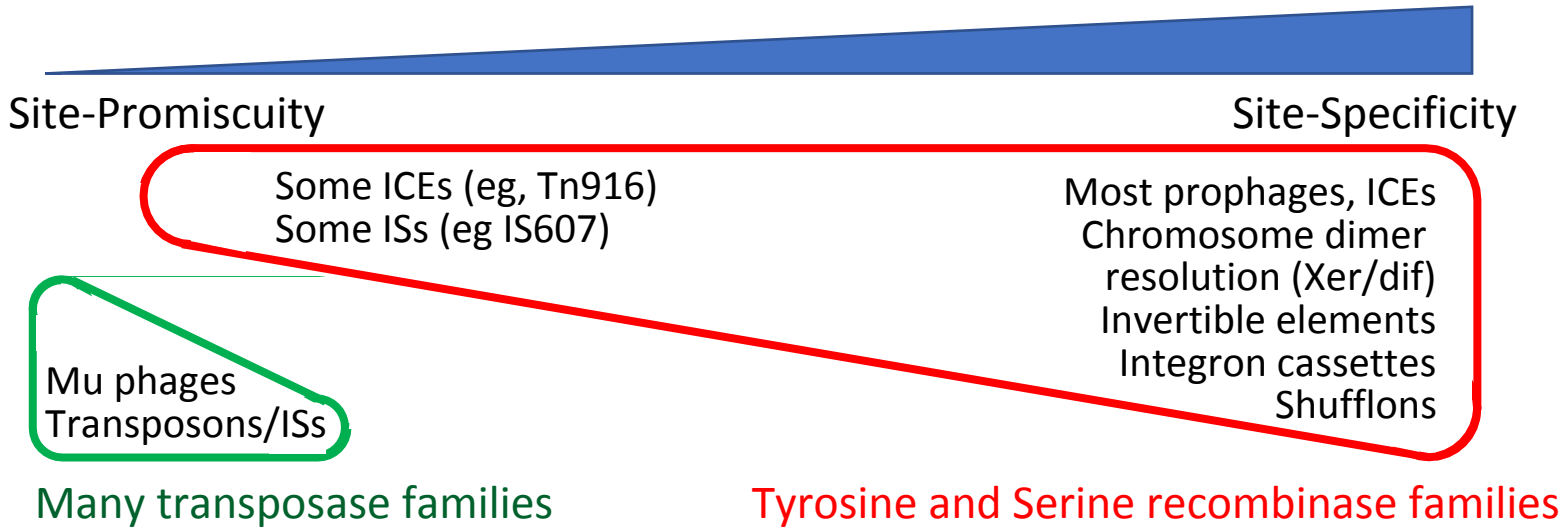




Outline

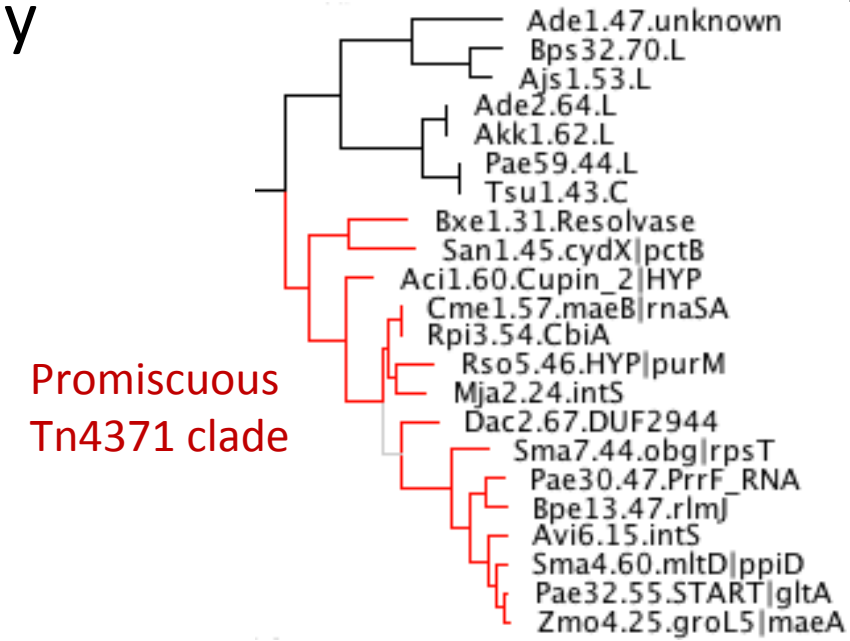
- Introduction: genomic islands, satellite/helpers, integration/excision
- Value of precise mapping of integrative DNAs
- Software that maps prophages precisely
- Discoveries
  - Integrase site specificity
  - Regulated gene integrity
  - Helper-embedded satellites
- Phage factory for therapy and energy

DNA specificity spectrum for recombinases

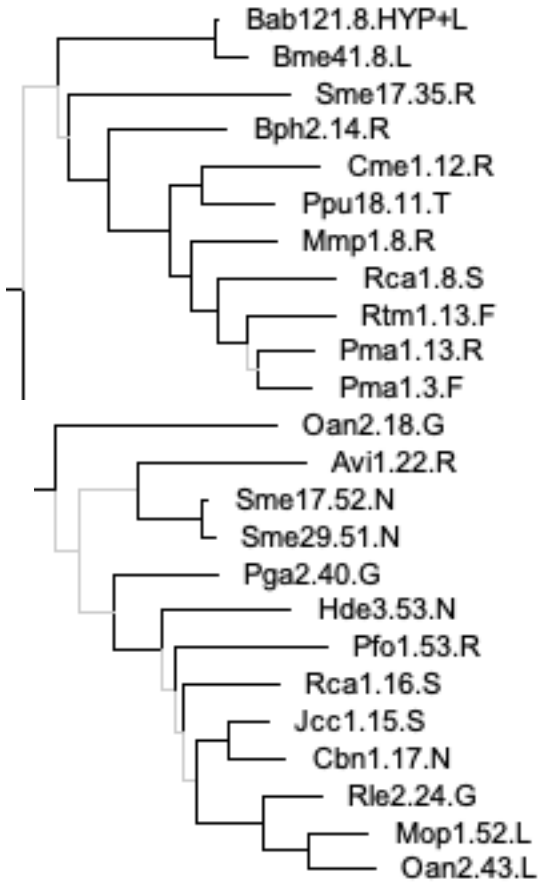


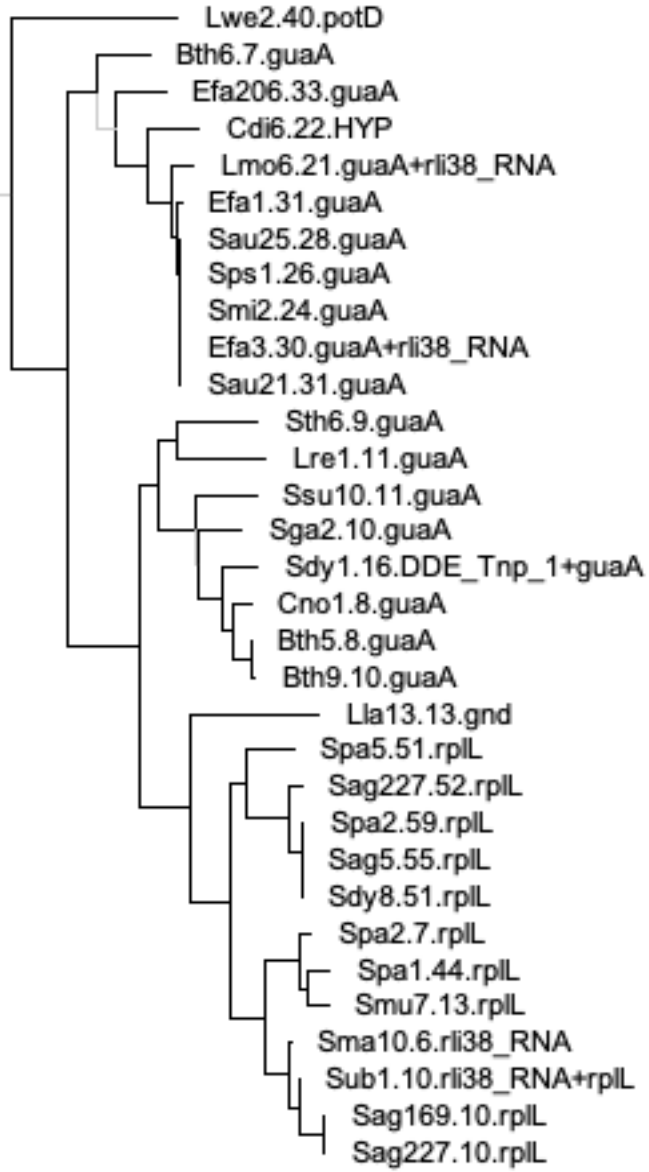
Utility of good integrase phylogeny

- Build integrase tree
  - We can do better (next slide)
- Decorate tree with information about site usage and bacterial taxonomy



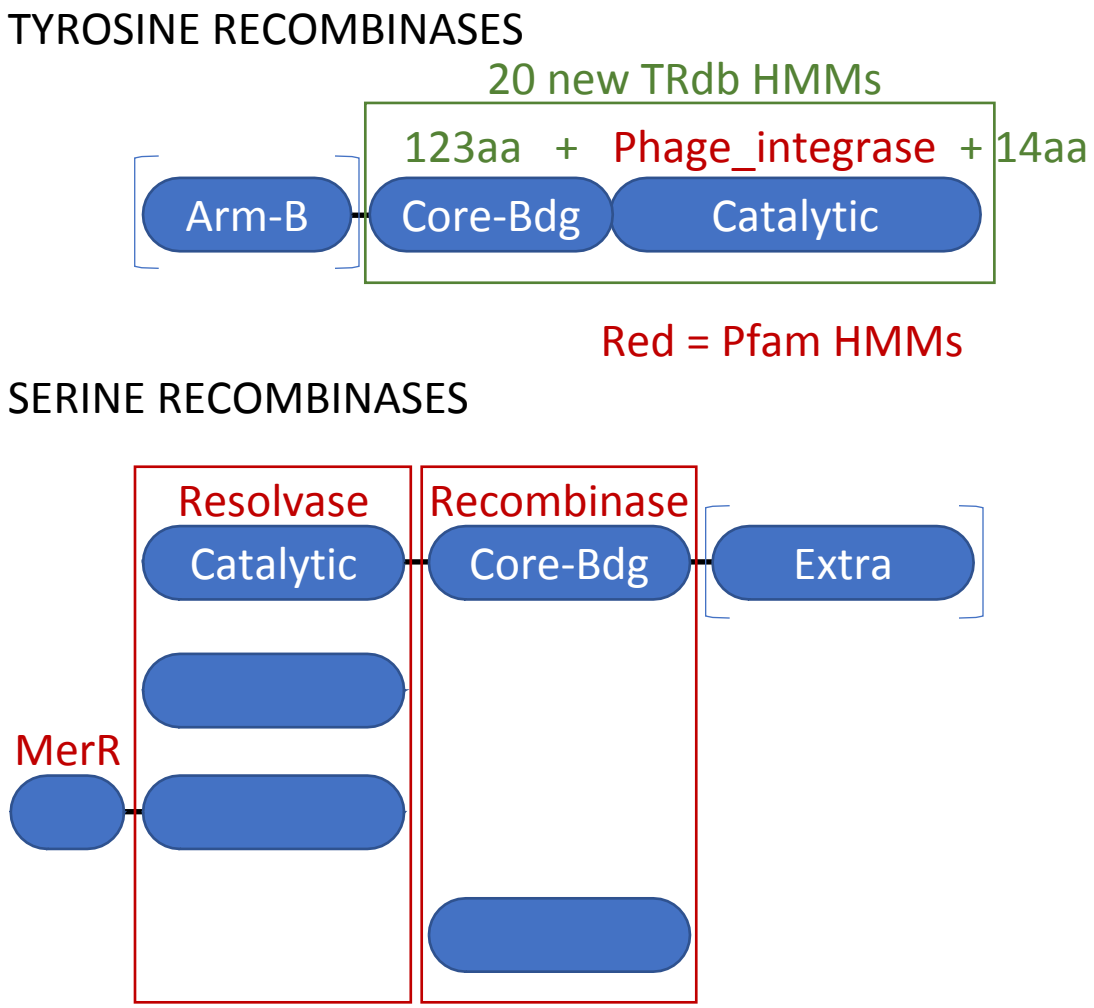
Promiscuous Tn4371 clade





Integrase trees

- Collaboration with Tandy Warnow and Paul Zaharias (UIUC)
- Three 1000-member backbone alignments now built for Tyr, Res, Rec
- Later, broaden alignments to all proteins
- Build two trees (Tyr and concatenated Res+Rec), based on nucleotide sequences

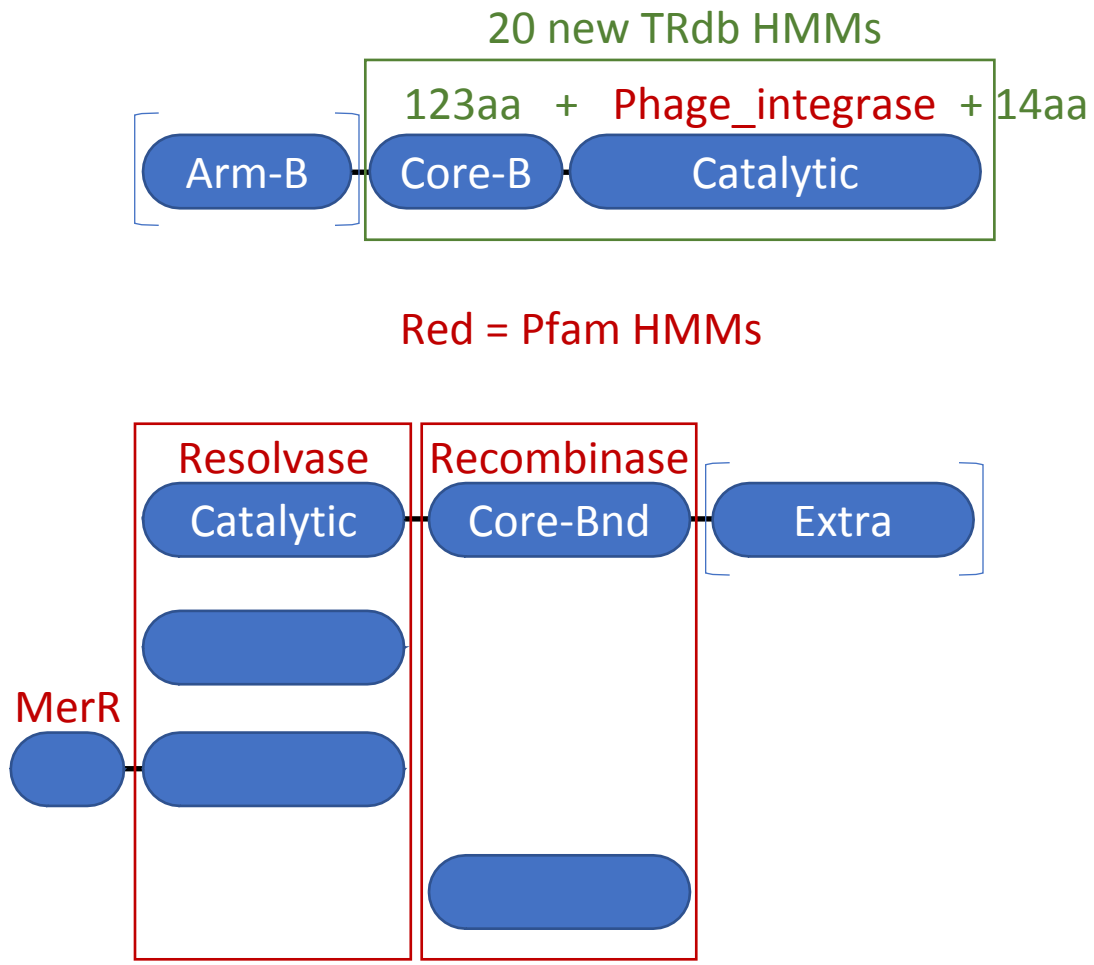


	Unique Proteins	Mapped attBs (Projected)
TYROSINE RECOMBINASES	470297	87033
SERINE RECOMBINASES	136812	
Resolvase Catalytic Core-Bdg Extra	56061	5691
MerR	68986	4790
	4387	1187
	7378	389

288K bacterial/archaeal genomes > 3.27M total int proteins > 0.61M unique proteins

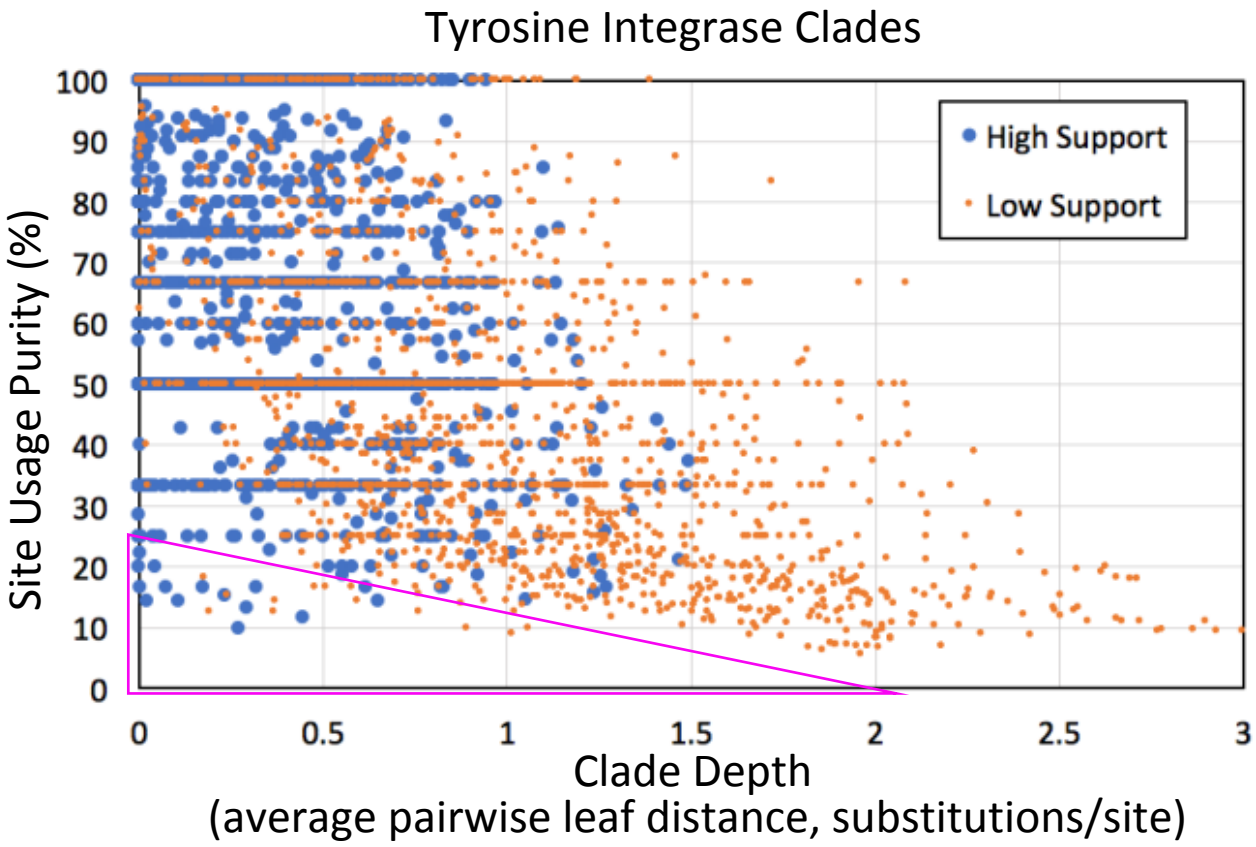
Add in: absent gold standards, absent TRdb proteins (197/866, many from phages)

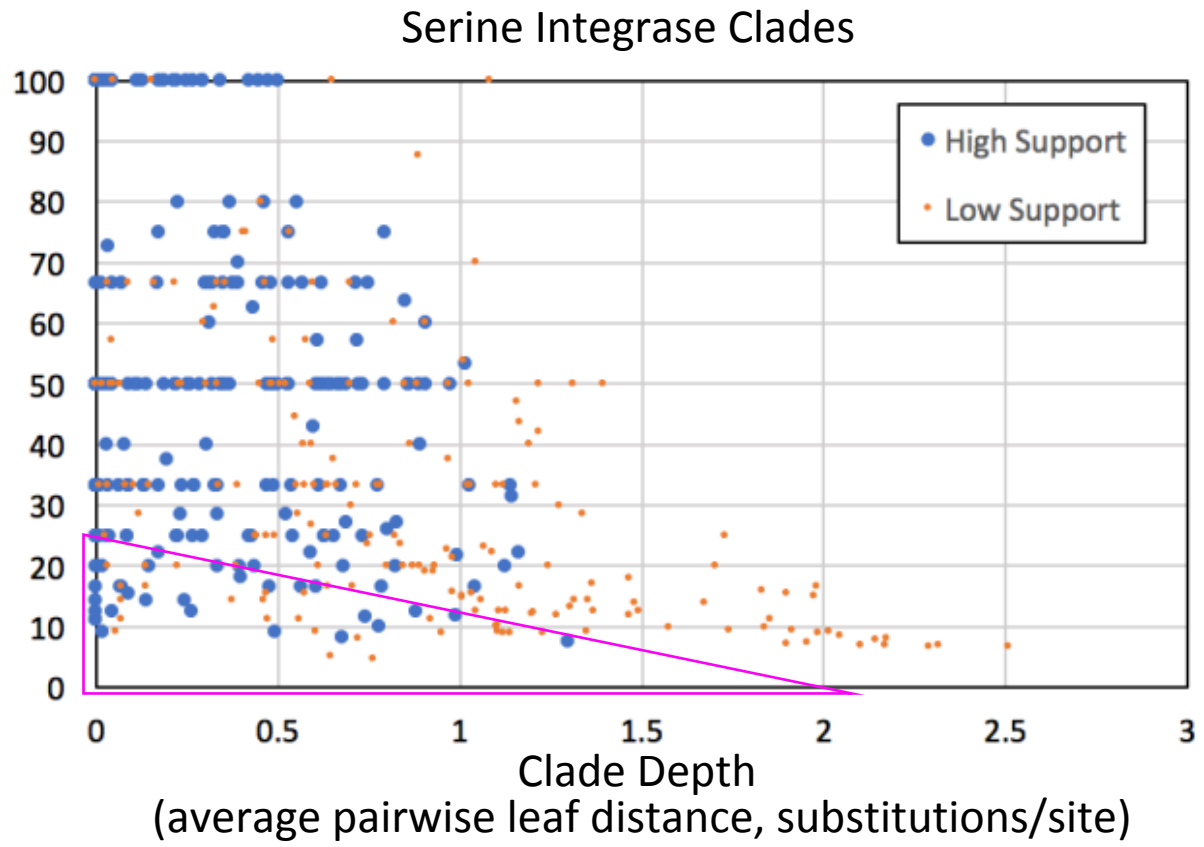
	Unique Proteins	Mapped attBs (Projected)
TYROSINE RECOMBINASES	470297	87033
SERINE RECOMBINASES	136812	
	56061	5691
	68986	4790
	4387	1187
	7378	389



3 fasta files, with HMM-hit portions in upper case: Tyr, Res and Rec

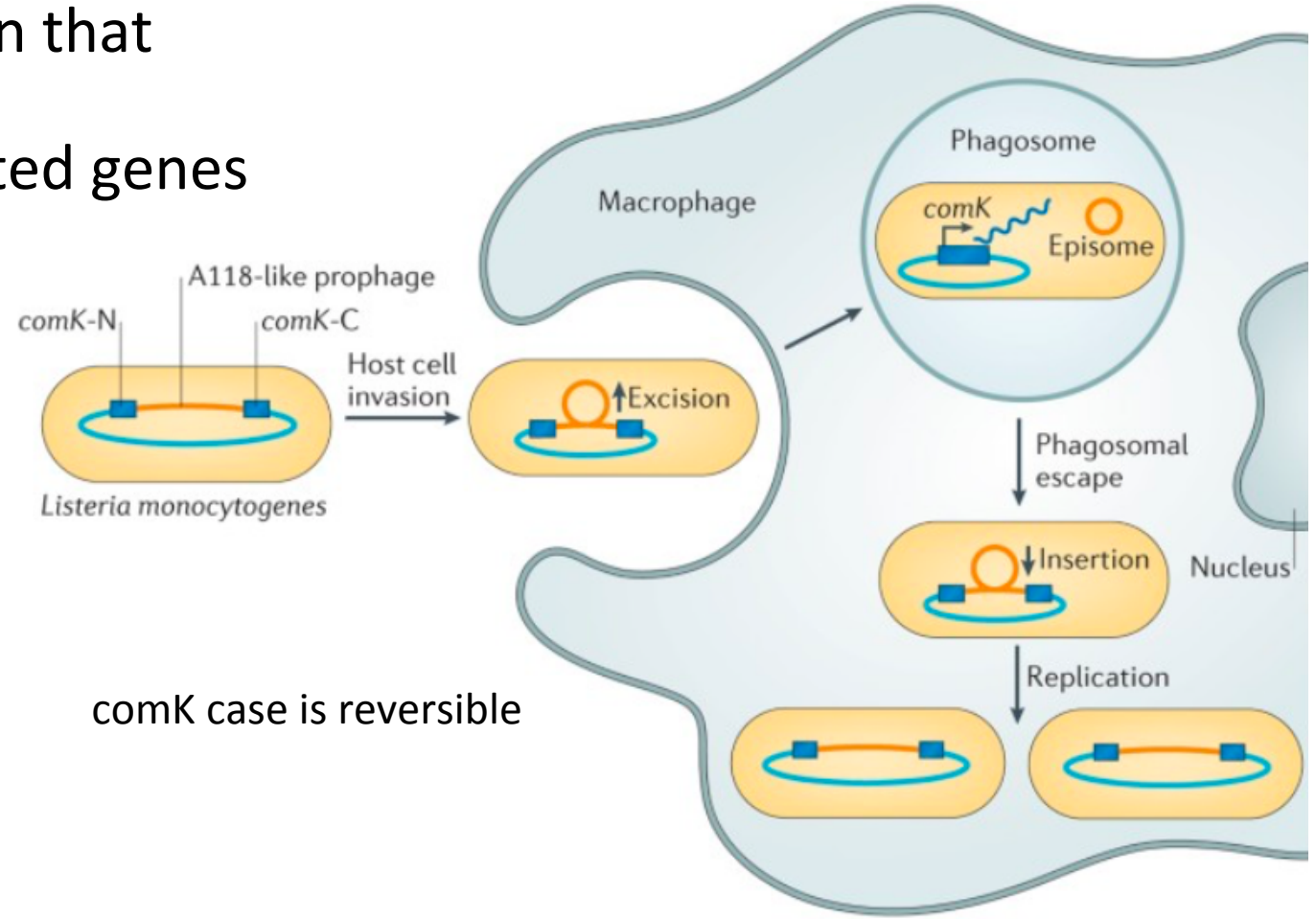
Rooting: unknown for Ser, topoisomerases for Tyr


Tyrosine Integrase Clades
Site Usage Purity (%)
100
90
80
70
60
50
40
30
20
10
0
0
0.5
1
1.5
2
2.5
3
High Support
Low Support
Clade Depth
(average pairwise leaf distance, substitutions/site)


Serine Integrase Clades
100
90
80
70
60
50
40
30
20
10
0
0
0.5
1
1.5
2
2.5
3
High Support
Low Support
Clade Depth
(average pairwise leaf distance, substitutions/site)

Island-regulated gene integrity

- Contrast with benign integration that doesn't inactivate gene
- Previously known island-regulated genes
  - Phagosome escape
    - comK
  - Spore mother cell
    - sigK, spsM, gerE
  - Heterocyst differentiation
    - nifD, fdxN, hupL
  - Mutation rate control
    - mutL
  - Curli/biofilm
    - mlrA
  - Beta-hemolysin conversion
    - hlb



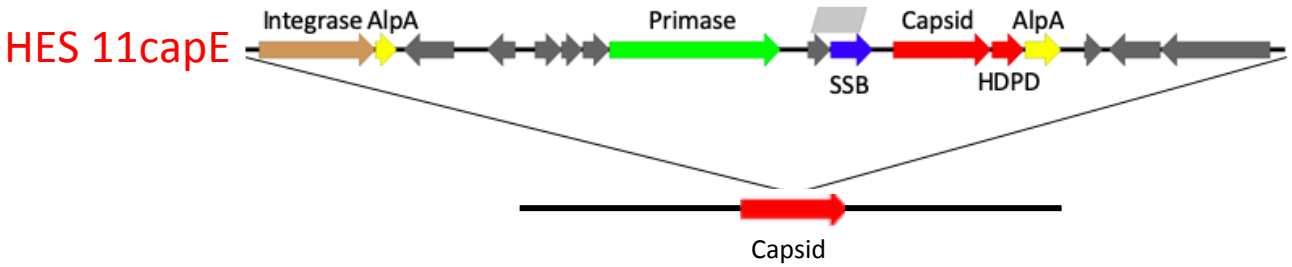
comK case is reversible

Feiner et al, 2015. Prophages as active regulatory switches of bacteria. Nat. Rev. Microbiol 13: 641–650.

Detecting Regulated Gene Integrity

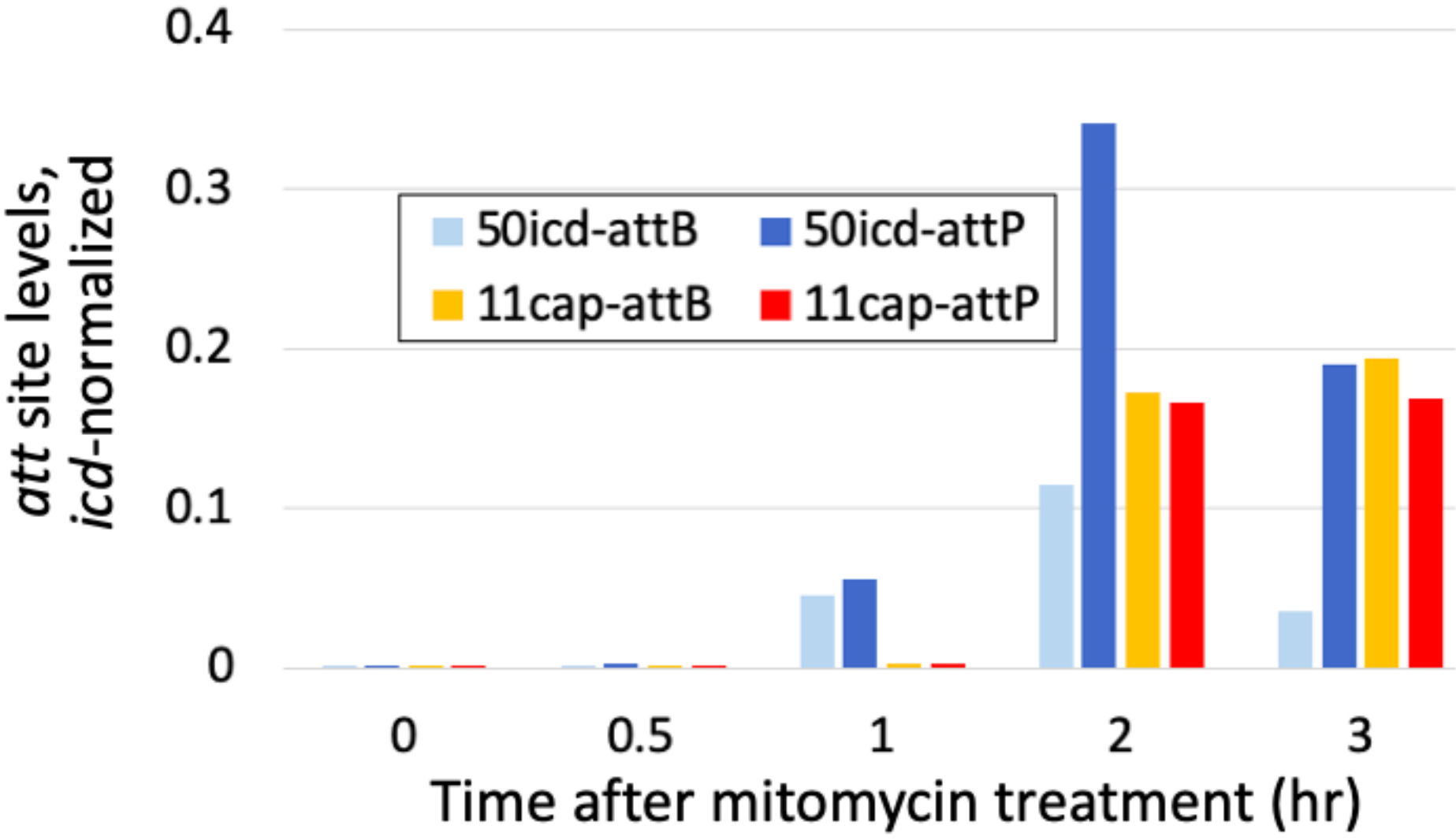
- Deduplicate GI set to remove possible vertical inheritance cases
- Stringent gene inactivation test: Pfam disruption
  - Non-tRNA GIs invade domain-coding regions only half as often as mock GIs from same genomes
  - attB:attLR Pfam bitscore ratio > 1.1 indicates disruption
  - Detects 8 of 11 previously known RGI cases
- Exclude one-off gene inactivations from promiscuous integrases or rare off-target
  - Insist on two or more deduplicated GIs at same gene, in a tight clade
- 19 new candidates for RGI: dut, eccCa1, gntT, hrpB, merA, ompN, prkA, tqsA, traG, yifB, yfaT, ynfE, and 7 others

Helper-Embedded Satellites (HESs)

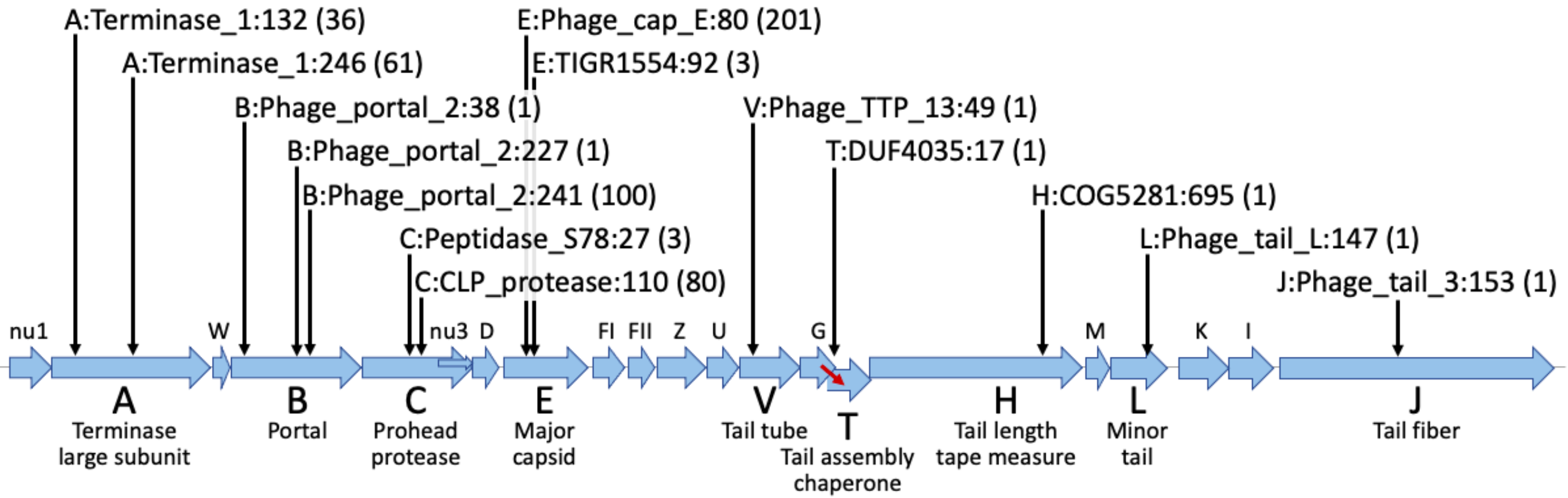


Integrase | Recombination | Lysis | Head | Other
Lysogeny | Replication | AlpA | Tail | Unknown

Induction: entire composite excises, it replicates, then the satellite excises



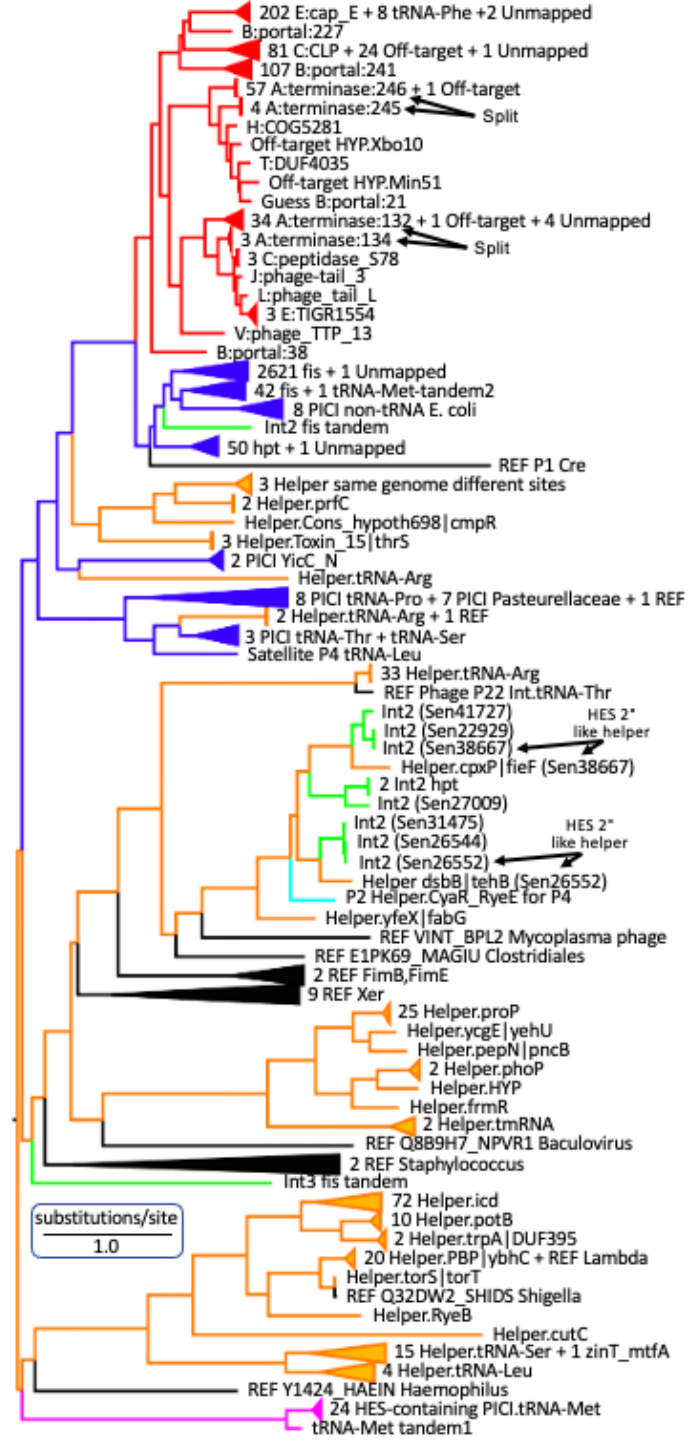
Search for integrase relatives identified 491 additional HESs, in 14 additional sites all in prophage late genes



HES sites mapped onto phage lambda late gene region

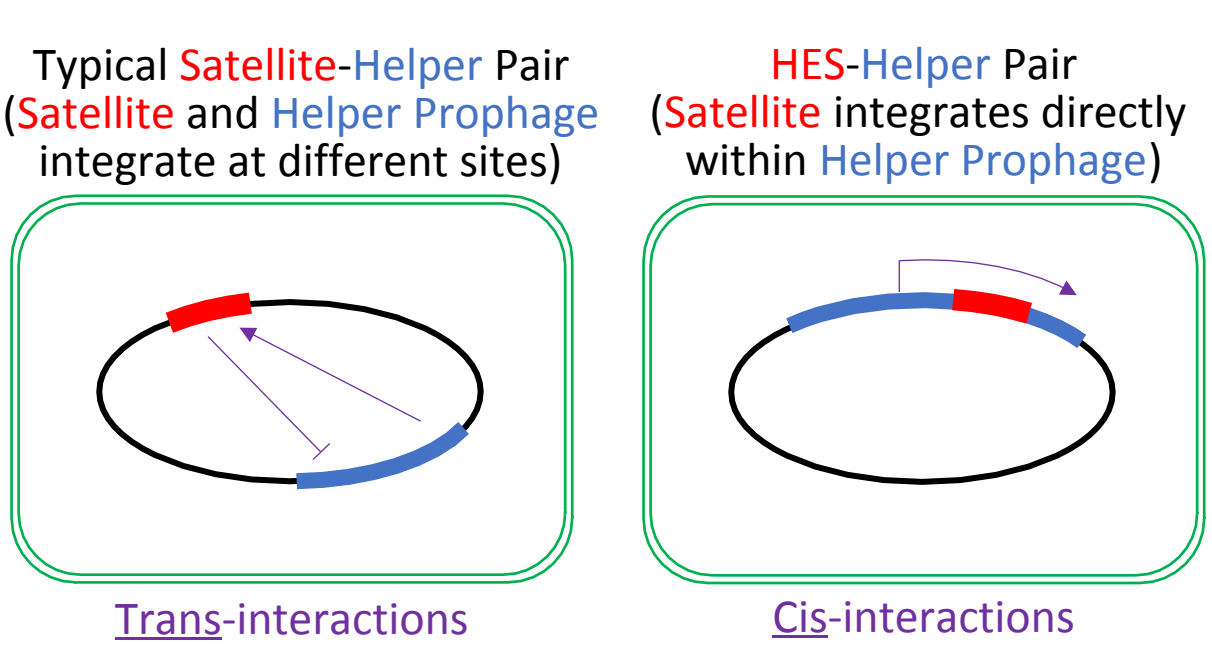
Integrase tree

- Orange and Black – from prophages
- Blue – from previously known and newly discovered GN-PICIs
- Red – HESs
  - HES int subclades perfectly respect usage of the 15 sites
  - Clade is adapted to finding new sites only in phage late genes: mechanism?!

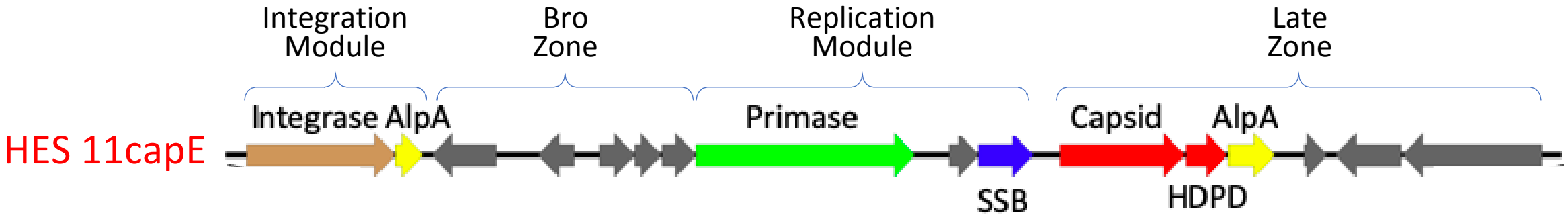


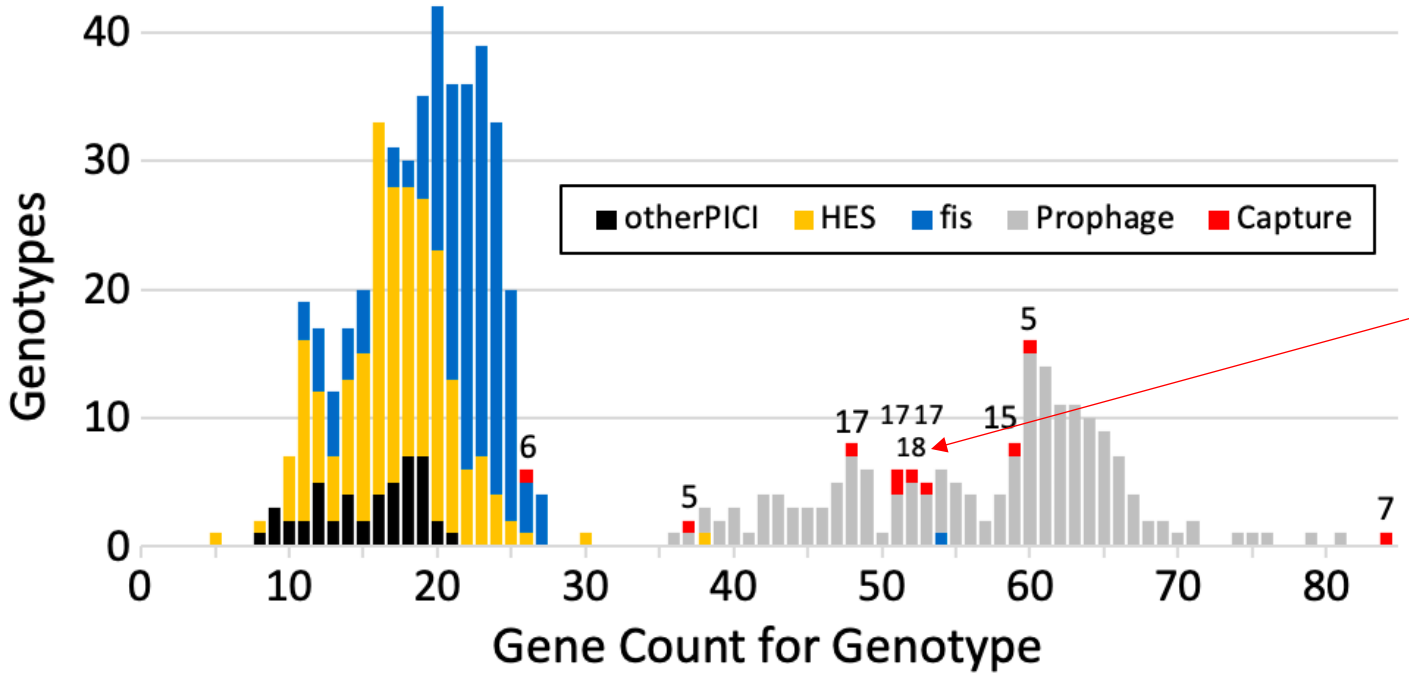
New helper cis-interactions for HESs

- HES transcription is directly coupled to helper late transcription
- HES replication is directly coupled to helper replication
- Target late gene in helper is broken until HES excises



HES/GN-PICI genome organization



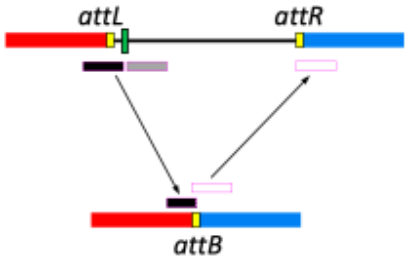


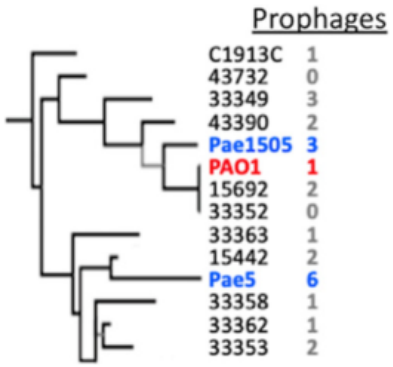
Capture in satellite late zones of clusters as long as 18 genes from known helper late gene regions

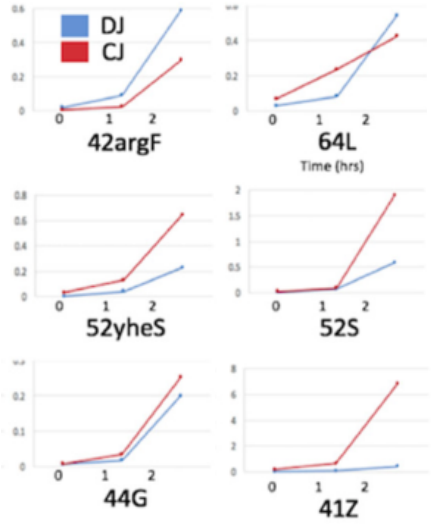
Outline

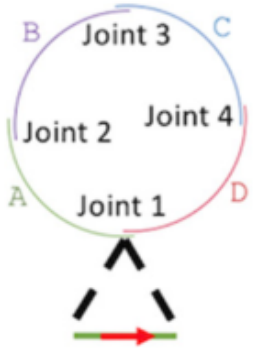
- Introduction: genomic islands, satellite/helpers, integration/excision
- Value of precise mapping of integrative DNAs
- Software that maps prophages precisely
- Discoveries
  - Integrase site specificity
  - Regulated gene integrity
  - Helper-embedded satellites
- Phage factory for therapy and energy

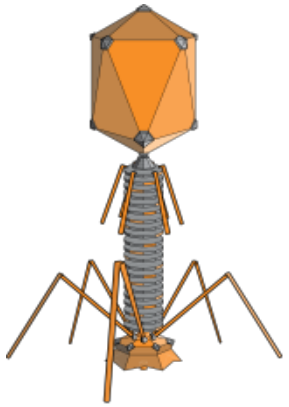
Phage factory: targeting any bacterium





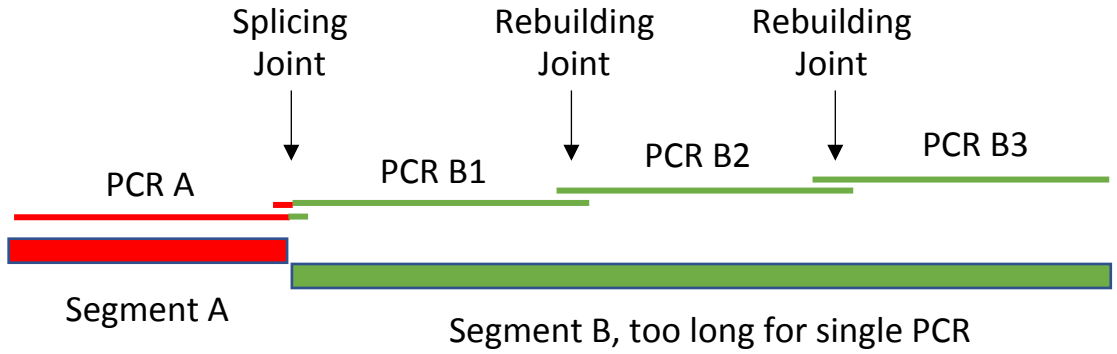




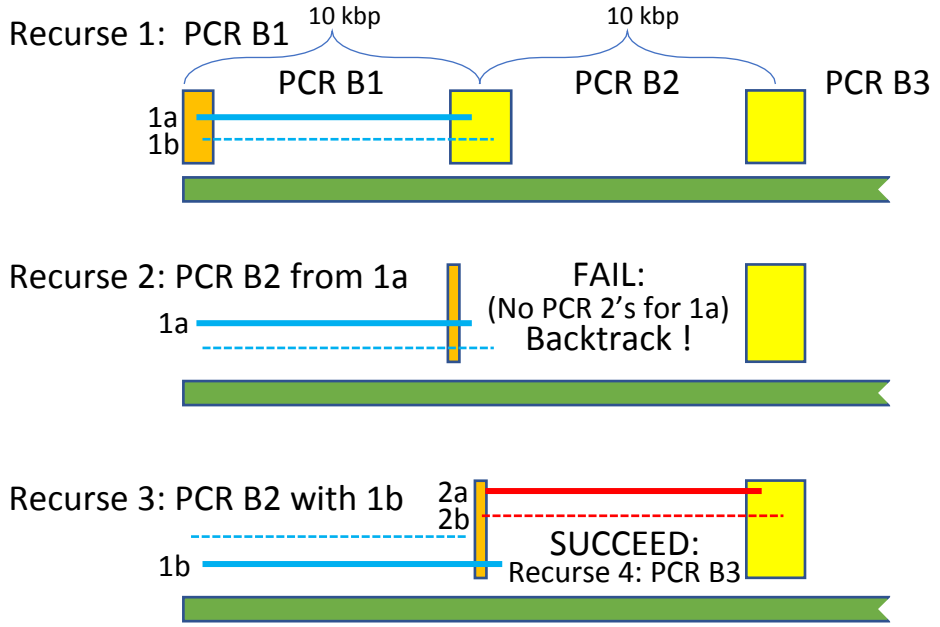


Prophage database → Identify prophage-rich close relatives of target → Validate phages → Engineer phages → Reboot

BigDNA software for designing big overlap assemblies



Solution: Recursive backtracking with PRIMER3/tntBLAST



Summary

- Precise island mapping leads to more discovery
- Huge database of integrase/attB pairs
- HESs
  - New cis-interactions with helper
  - Mechanism for HES integrase clade to only find new sites in prophage late genes
  - For utility in vectors, helpers for HESs are already known

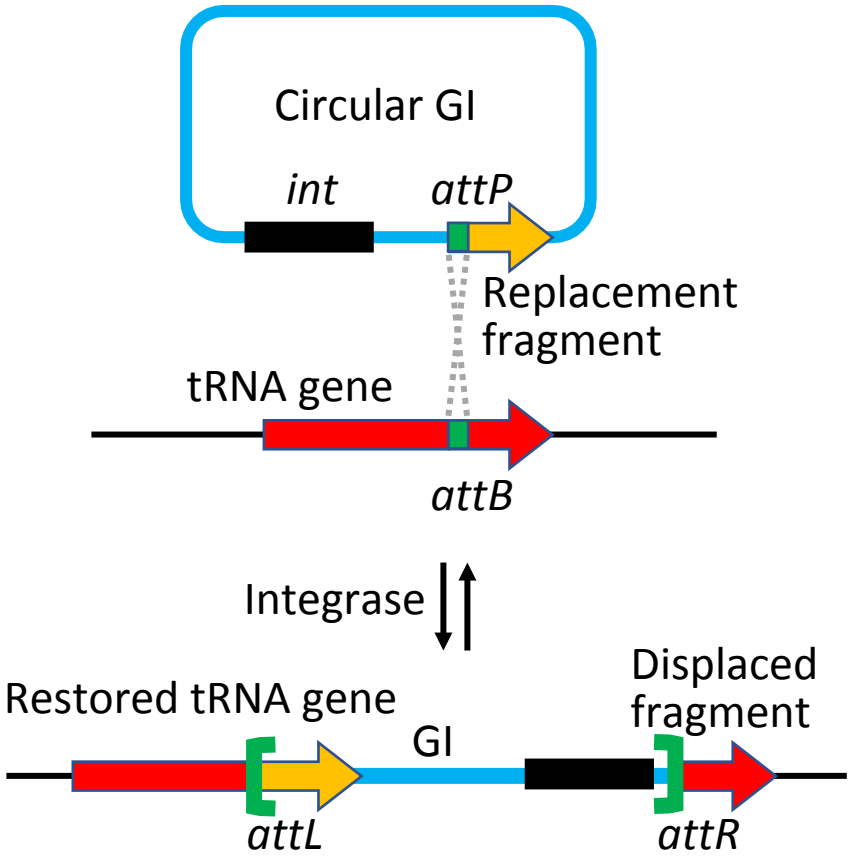
Acknowledgements

- Sandia Collaborators
  - Katie Mageeney
  - Raga Krishnakumar
  - Joe Schoeniger
  - Corey Hudson
  - Britney Lau
  - Julian Wagner
  - Ivan Vuong
  - Todd Lane
  - Anupama Sinha
- Prairie View A&M U
  - Noushin Ghaffari
  - Fatema Shormin
  - Bernard Nyarko
- U Illinois Urbana-Champaign
  - Tandy Warnow
  - Paul Zaharias
- Funding
  - Sandia Laboratory-Directed Research & Development: Bioscience and E/HS
  - DOE-BER SFA-Secure Biosystems Design (Joe Schoeniger, PI)

Sandia National Laboratories is a multi-mission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

Regulated Gene Integrity

Benign Integration into a Gene



Regulated Gene Integrity

- Island lacks Replacement fragment, so target gene is inactivated upon integration
- Some bacteria control key genes this way
  - Sporulation in Bacillus
  - Multicellular differentiation in cyanobacteria
- 10 such regulated genes had been previously reported
- We recovered most of these and discovered 19 new such genes