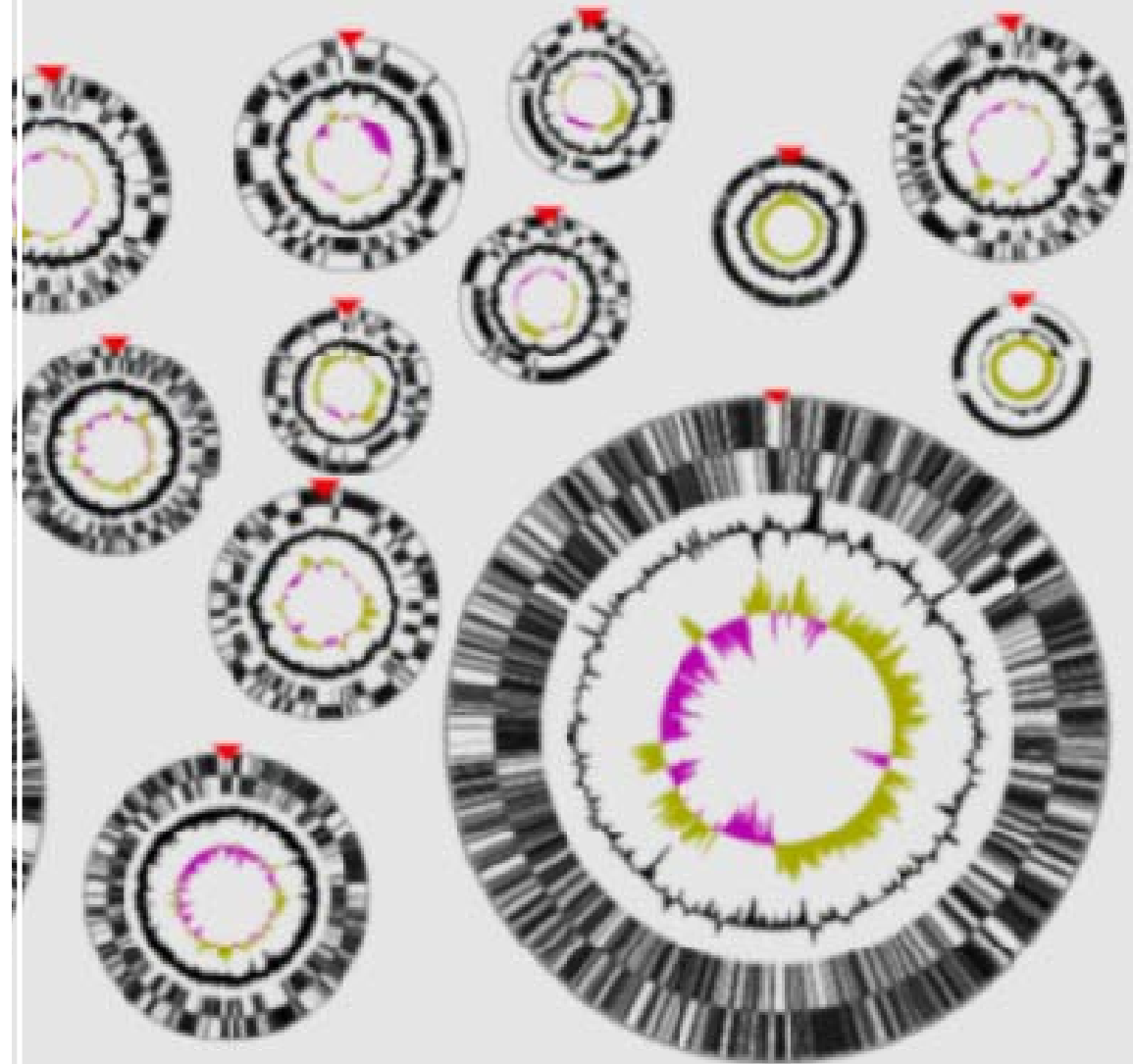




Extracting Biologically Meaningful Patterns from Bacterial Genome Data

Pattanapon Kayansamruaj,
PhD

*Department of Aquaculture,
Faculty of Fisheries, Kasetsart
University*



Overview

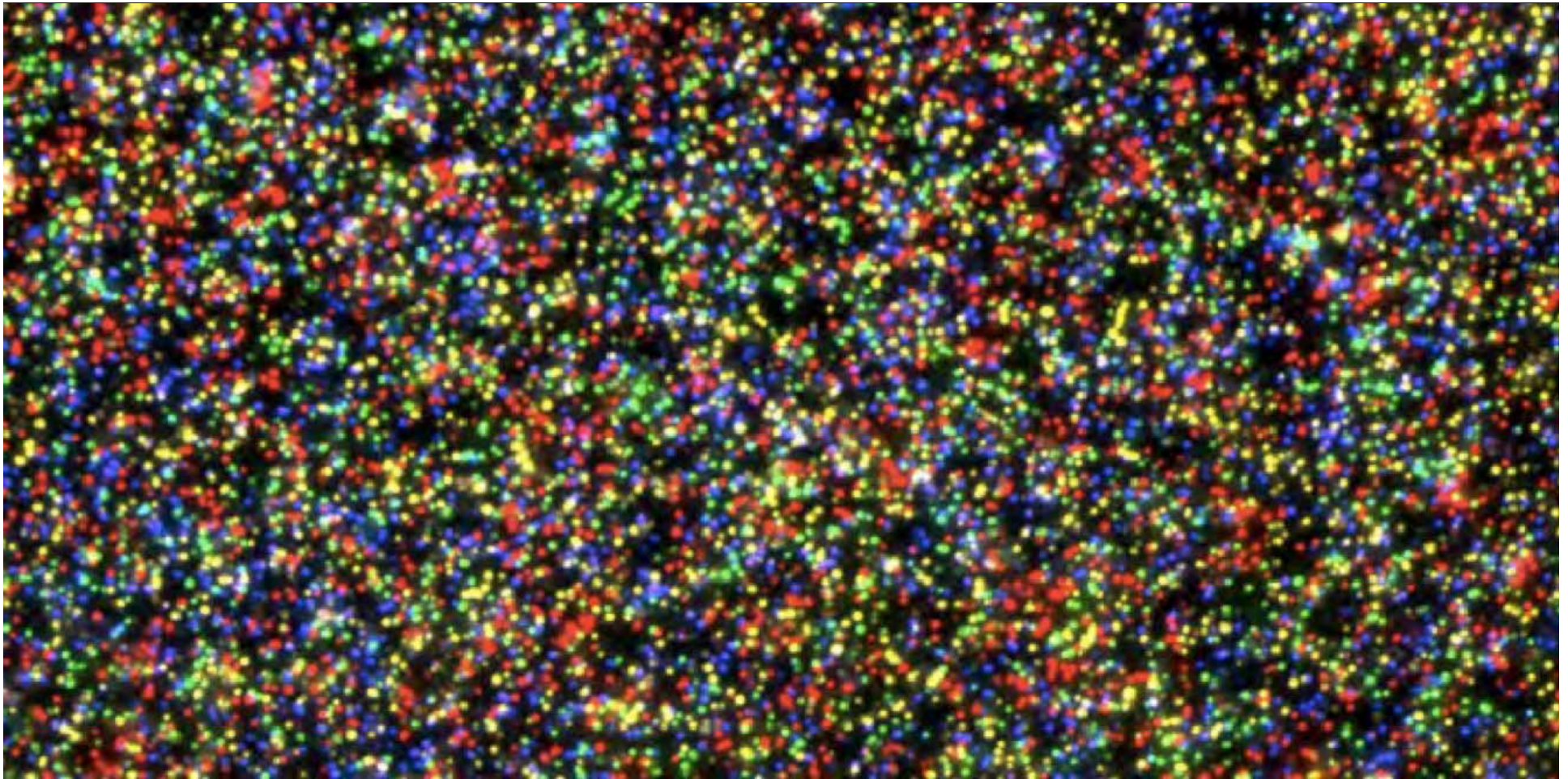
- Impact of Next-Generation sequencing to science
- Generic workflow in bacterial genome analysis
- High-throughput screening of bacterial genomes
- Comparative genomics and Phylogenomics
- Getting published

One reviewer's comment to a manuscript conducting comparative genomics study

“The work is **very descriptive** and **does not have a real hypothesis** that can be tested...”

And the authors replied...

“...thought this manuscript **has no hypothesis** to be tested and that this observation somehow diminished the value of the study. We can simply point out that for “genome” analyses **that is universally true**. Such studies are **often hypothesis generating** and that perhaps is one of **their primary purposes**.”



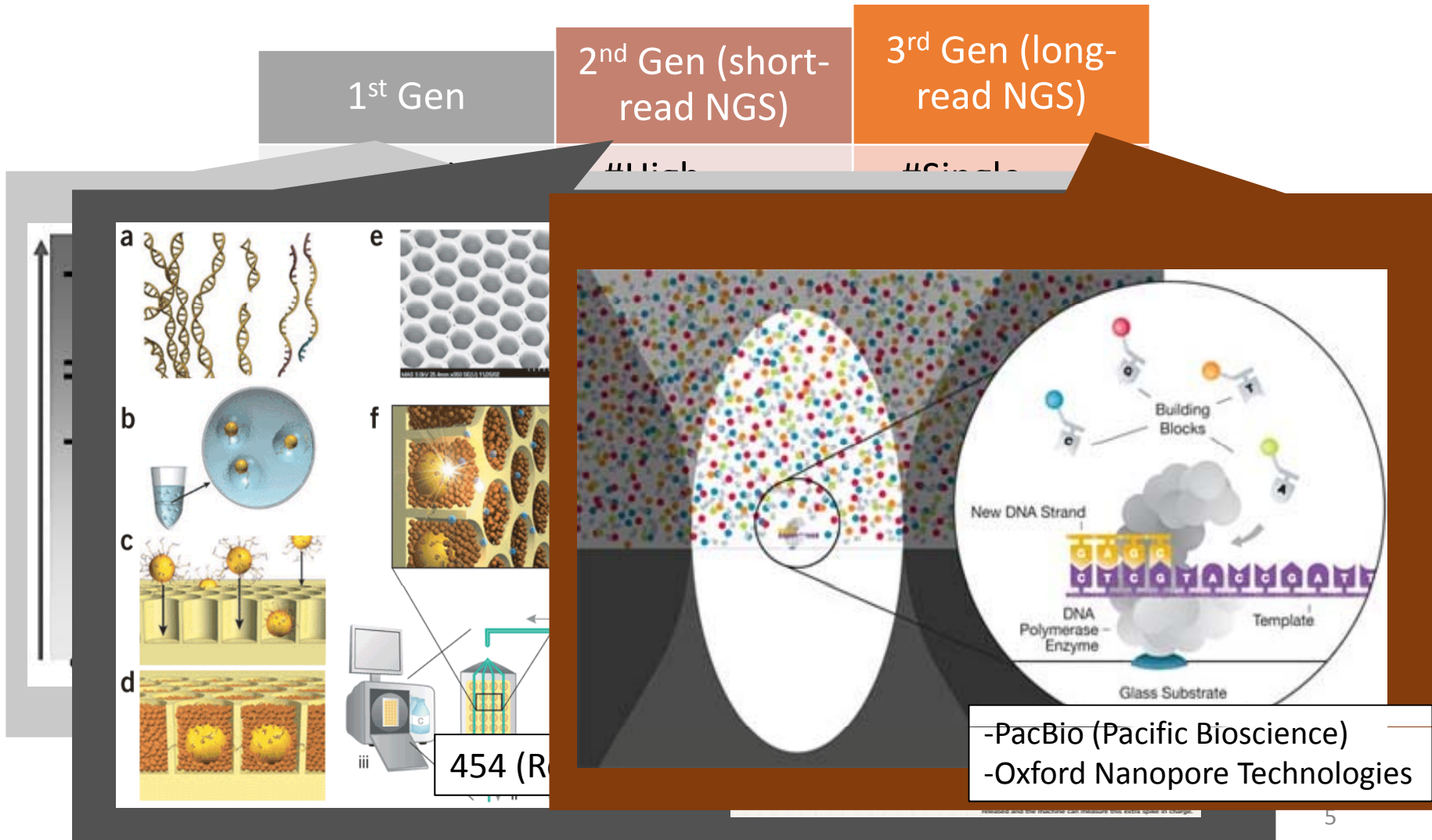
Impact of Next-Generation sequencing to science |

What is Next-Generation sequencing?

1st Gen

2nd Gen (short-read NGS)

3rd Gen (long-read NGS)



NGS Platform available in the market

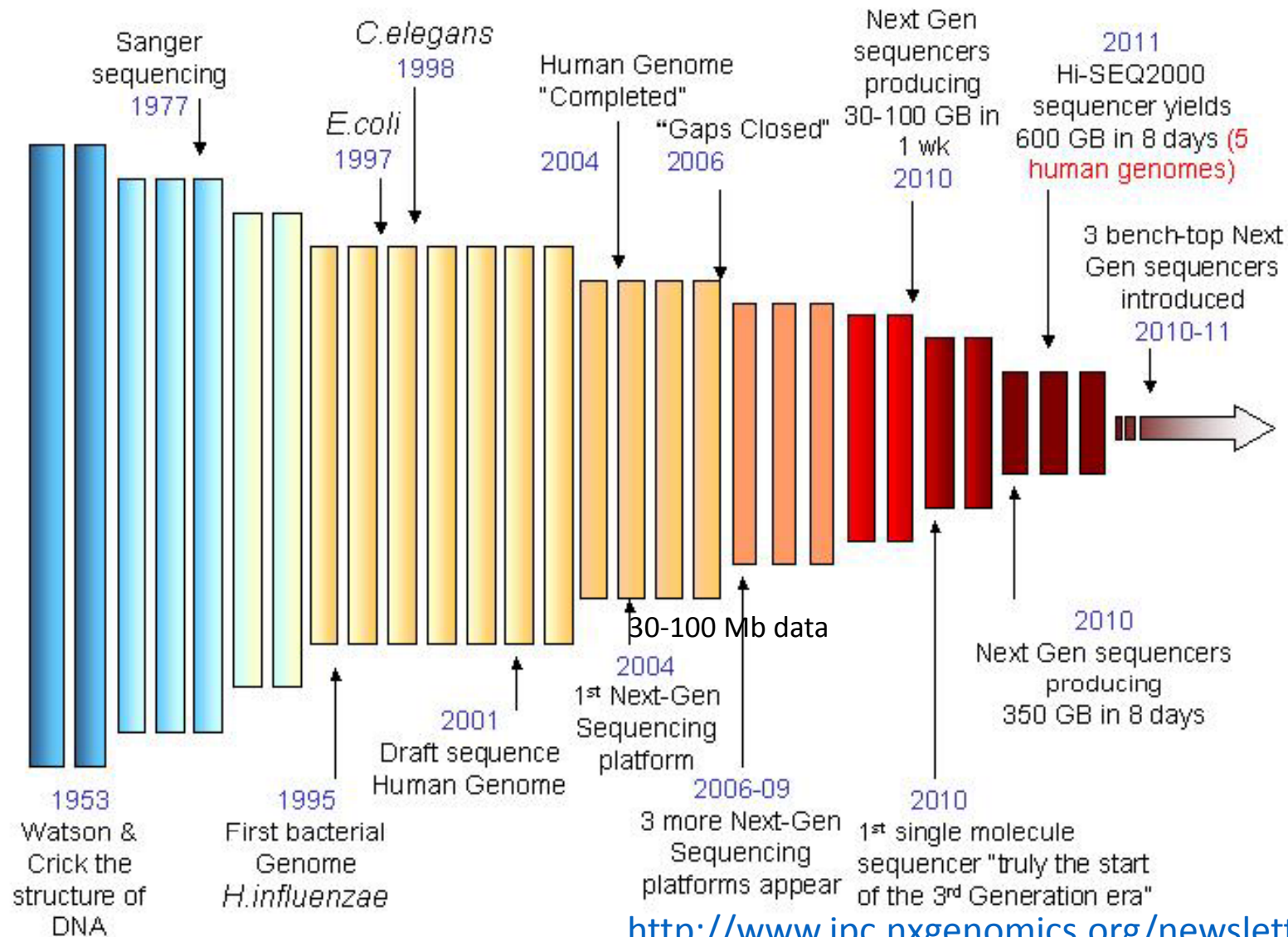
Short-read NGS

- 454 (Roche)
- SOLiD (Thermo Fisher)
- Complete Genomics (BGI)
- Illumina
- Ion Torrent (Thermo Fisher)

Long-read NGS

- PacBio (Pacific Biosciences)
- Oxford Nanopore Technologies

Sequencing technology timeline

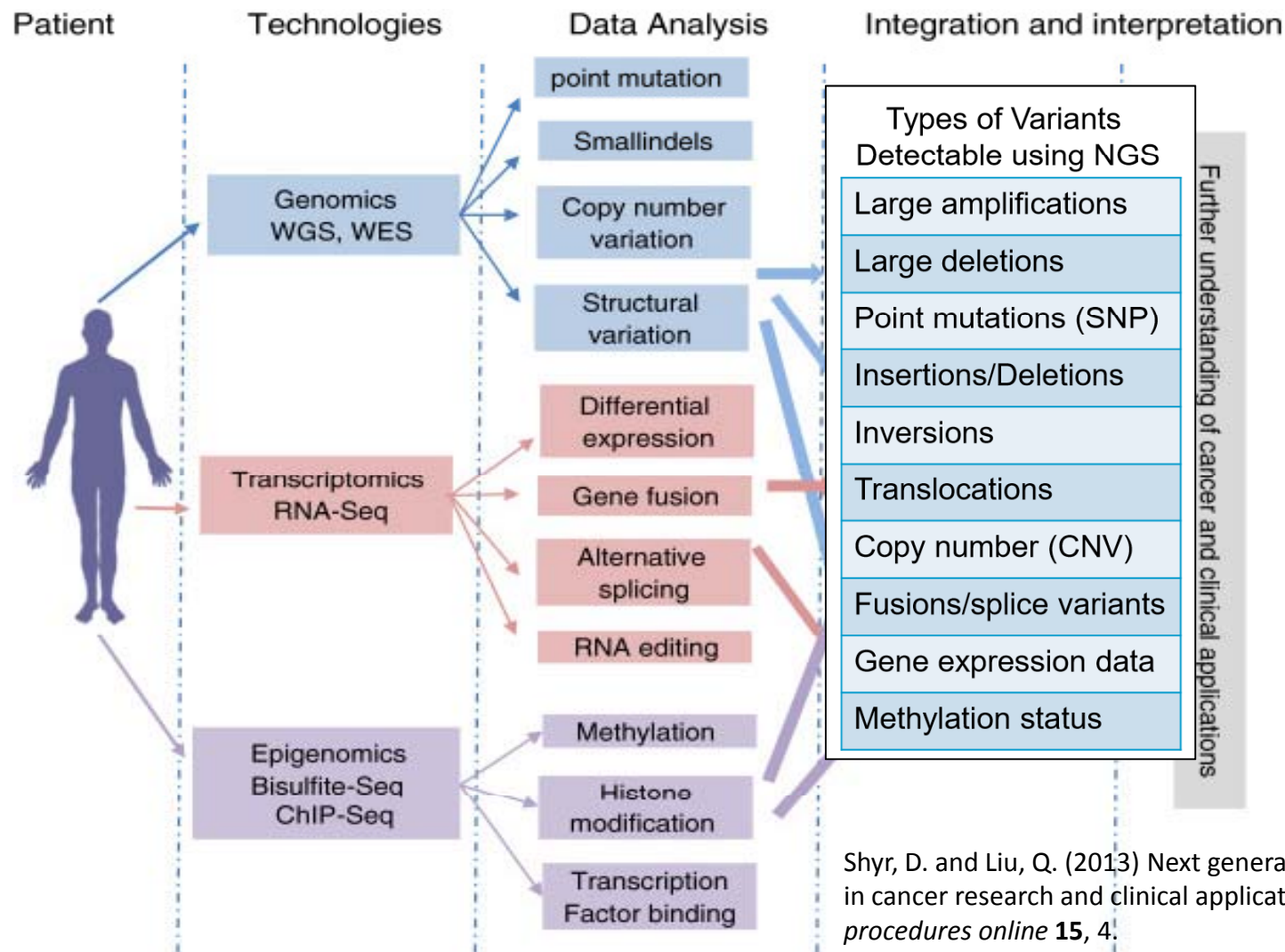


<http://www.ipc.nxgenomics.org/newsletter.htm>



- The first milestone for high-throughput analysis
- Initiated ~1990
- First draft of human genome was published in 2001 (92% of the genome)
- BAC clone and genome walking technique
- Multiple bioinformatic tools have been invented and polished to hasten analytical process
 - Sequence assembly
 - Genome annotation (gene finding)
 - Distributed computing
- Nowadays, the cost of sequencing a human genome down to around **US\$1,000** (finished within a week)

Application for basic and clinical research



Shyr, D. and Liu, Q. (2013) Next generation sequencing in cancer research and clinical application. *Biological procedures online* **15**, 4.



Generic workflow in bacterial genome analysis

Working with NGS data

Is there a perfect OS for NGS analysis



Windows



Mac



Linux



<https://my.vmware.com/web/vmware/downloads>

- *Program compatibility:* **Linux > MacOS >>>>>> Windows**
- Most of NGS data analysis packages run on Unix-based OS (Linux, Mac)
- For Windows user ...
 - Dual installation of Linux (Ubuntu is recommended)
 - Use virtual machine program (VMware) → limited computing performance

Working with NGS data

Operate computational analysis by...

Local computing

- Most versatile, a lot of experimentation
- Time-consuming, not applicable for large dataset



Computer networking (server)

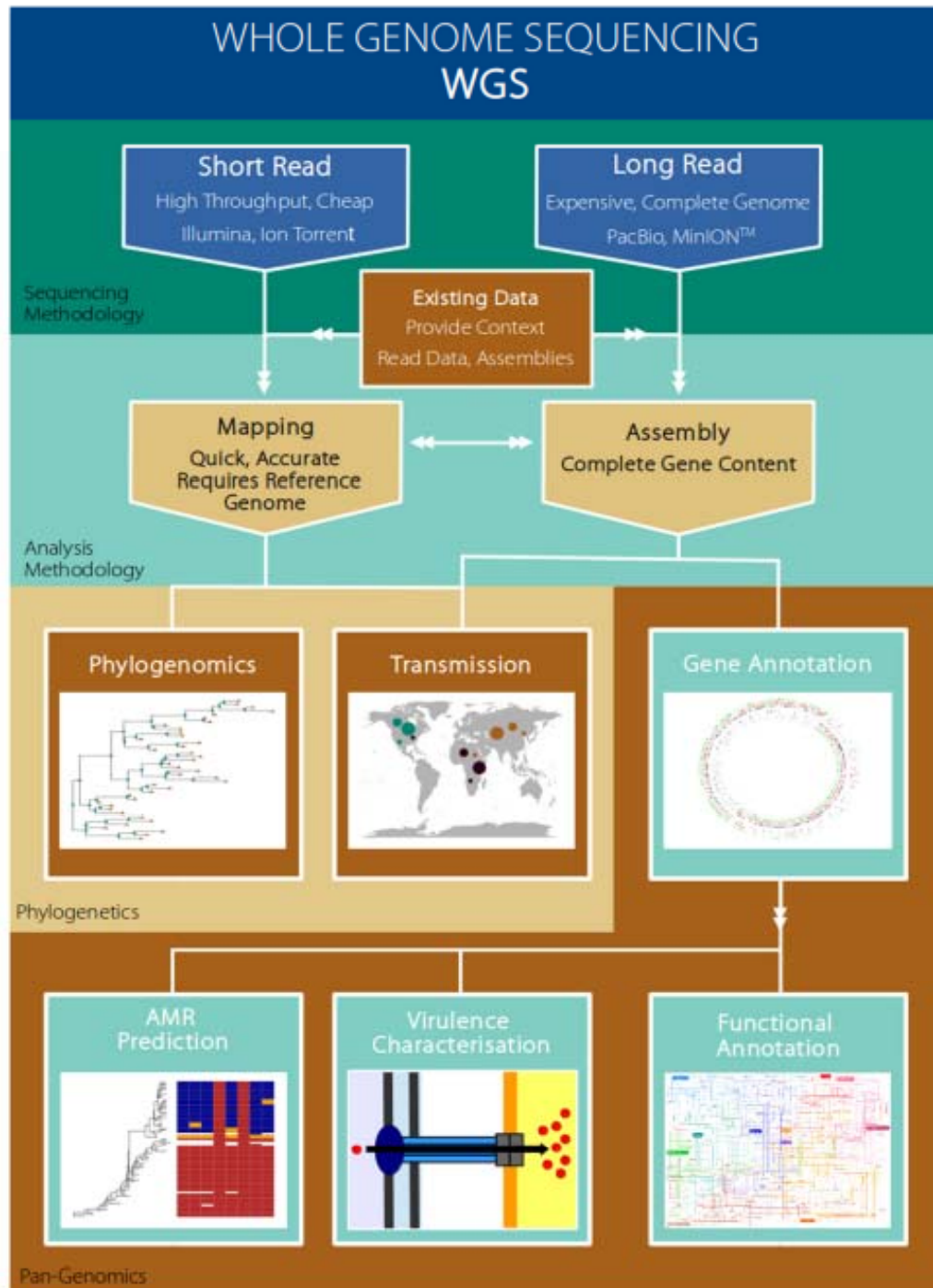
- High performance computing
- Limited programs to use
- *Authorization?*



Web-based analysis (via web browser)

- Fit for basic users
- OS independent
- Does not require hi-end desktop
- Large variety of application and computational power





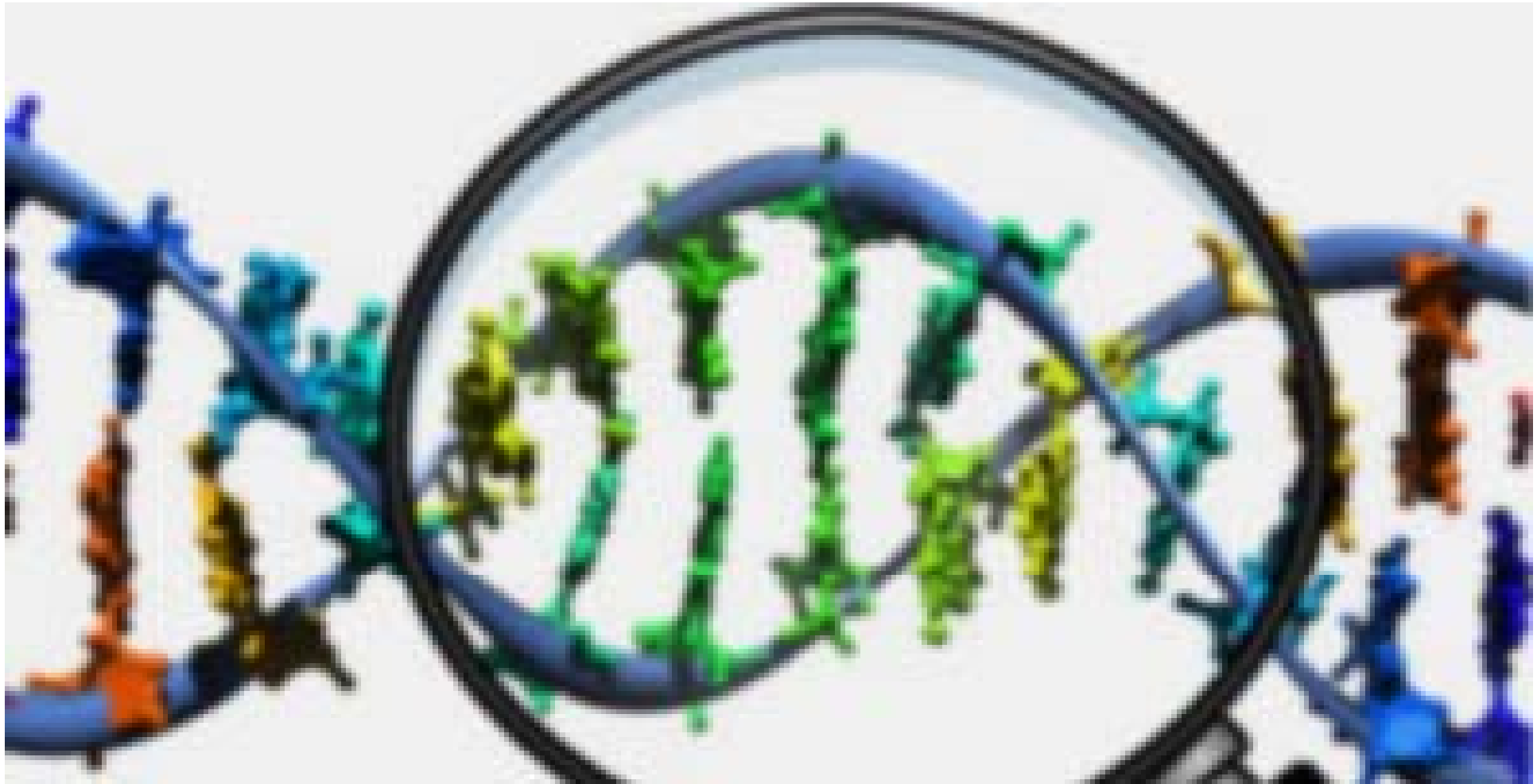
Pre-assembly analysis

- **Input:** Raw reads (.fastq)
- **Output :** Contigs, scaffolds, complete genome (.fasta)

Post-assembly analysis

- Data mining, Data interpreting, Visualization
- **Input:** .fasta
- **Output :** Biologically meaningful result

Bayliss, S.C., et al. (2017) The Promise of Whole Genome Pathogen Sequencing for the Molecular Epidemiology of Emerging Aquaculture Pathogens. *Frontiers in microbiology* **8**, 121.



High-throughput screening of bacterial genomes |

Identification gene of interest from bacterial genome

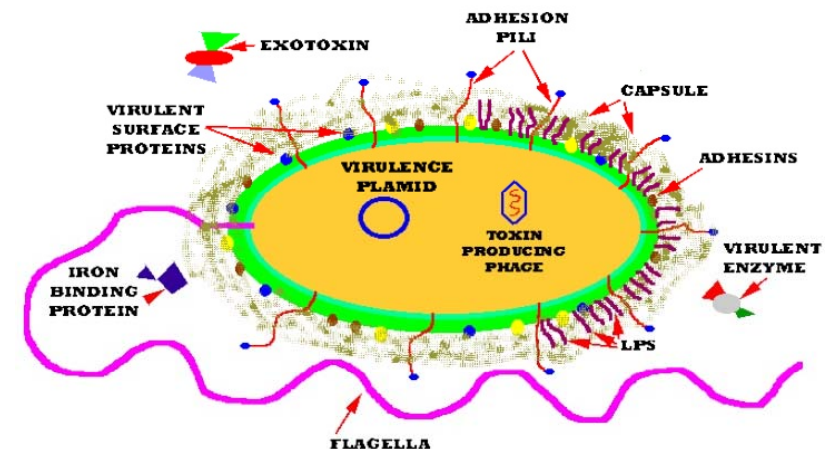
- Important elements in bacterial genome
 - Virulence factors
 - Antibiotic resistance genes
 - Genomic island
 - Bacteriophage
 - CRISPR
 - Prokaryotic Regulatory Proteins

Public databases enlisting important bacterial elements

Virulence factor

- **Virulence factors** refer to the **properties (i.e., gene products) that enable a microorganism to establish itself on or within a host** of a particular species and enhance its potential to cause disease.
- *e.g.* bacterial toxins, surface proteins mediated attachment, cell surface carbohydrates and proteins that protect a bacterium, and hydrolytic enzymes

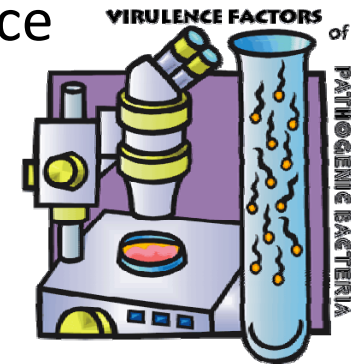
Bacterial Virulence Factors



Public databases enlisting important bacterial elements

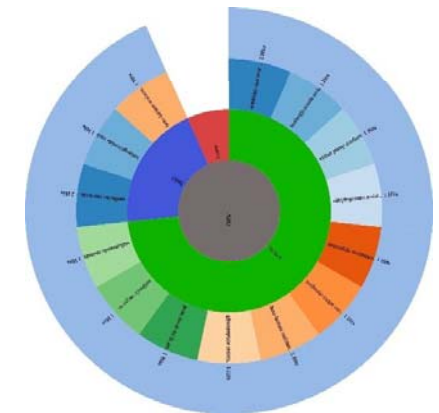
Virulence factor

- Virulence Factor Database (**VFDB**) → online resource for curating information about virulence factors of bacterial pathogens (<http://www.mgc.ac.cn/VFs/>)



Antibiotic resistance genes

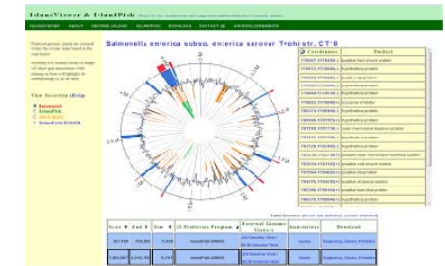
- The Comprehensive Antibiotic Resistance Database (**CARD**): <http://arpcard.mcmaster.ca/>
- Antibiotic Resistance Genes Database (**ARDB**): <https://ardb.cbcb.umd.edu/>



Public databases enlisting important bacterial elements

Genomic island

- IslandViewer(www.pathogenomics.sfu.ca/islandviewer/)
 - Curated database for GIs present in bacterial genomes
 - Predict GI from submitted genome using integrated tools



Bacteriophage

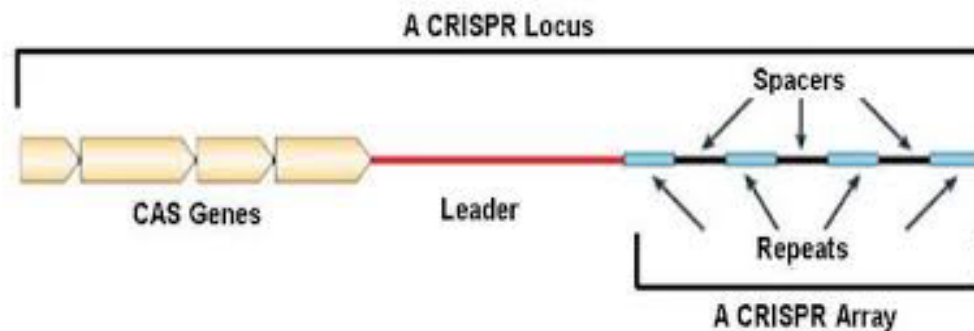
- Database
 - PhageDB (<http://phagesdb.org/>)
- Prediction tool
 - PHASTER (PHAge Search Tool - <http://phaster.ca/>)
 - → web server for the rapid identification and annotation of prophage sequences within bacterial genomes and plasmids.



Public databases enlisting important bacterial elements

CRISPR

- Clustered Regularly Interspaced Short Palindromic Repeats
- A prokaryotic immune system that confers resistance to foreign genetic elements
- CRISPRFinder (<http://crispr.i2bc.paris-saclay.fr/>) → automated tool for identification of CRISPR array present in bacterial genome

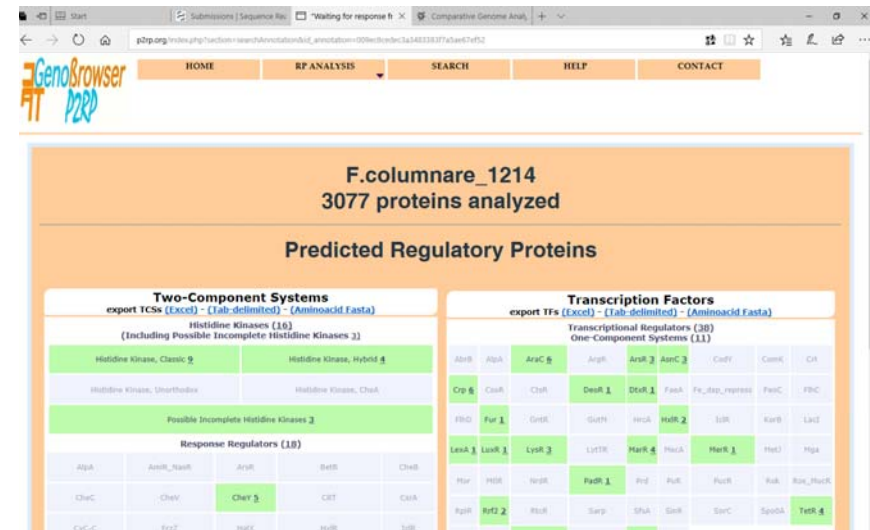


The screenshot shows the CRISPRFinder program interface. It includes a navigation panel on the left, a main display area showing the results of a search, and a detailed view of a CRISPR candidate. The interface is annotated with numbered circles (1-5) indicating key features: 1. The main search results table; 2. The 'Upload done with success' message; 3. The 'CRISPRs found in the submitted genomic sequence' section; 4. The detailed view of a CRISPR candidate; 5. A circular diagram representing the CRISPR array structure.

Public databases enlisting important bacterial elements

Prokaryotic regulatory proteins

- transcription factors (TFs) and two-component systems (TCSs)
- **P2RP** (Predicted Prokaryotic Regulatory Proteins)
 - <http://www.p2rp.org/>
 - Freely accessible web server, to provide a RP database and a platform for RPs prediction

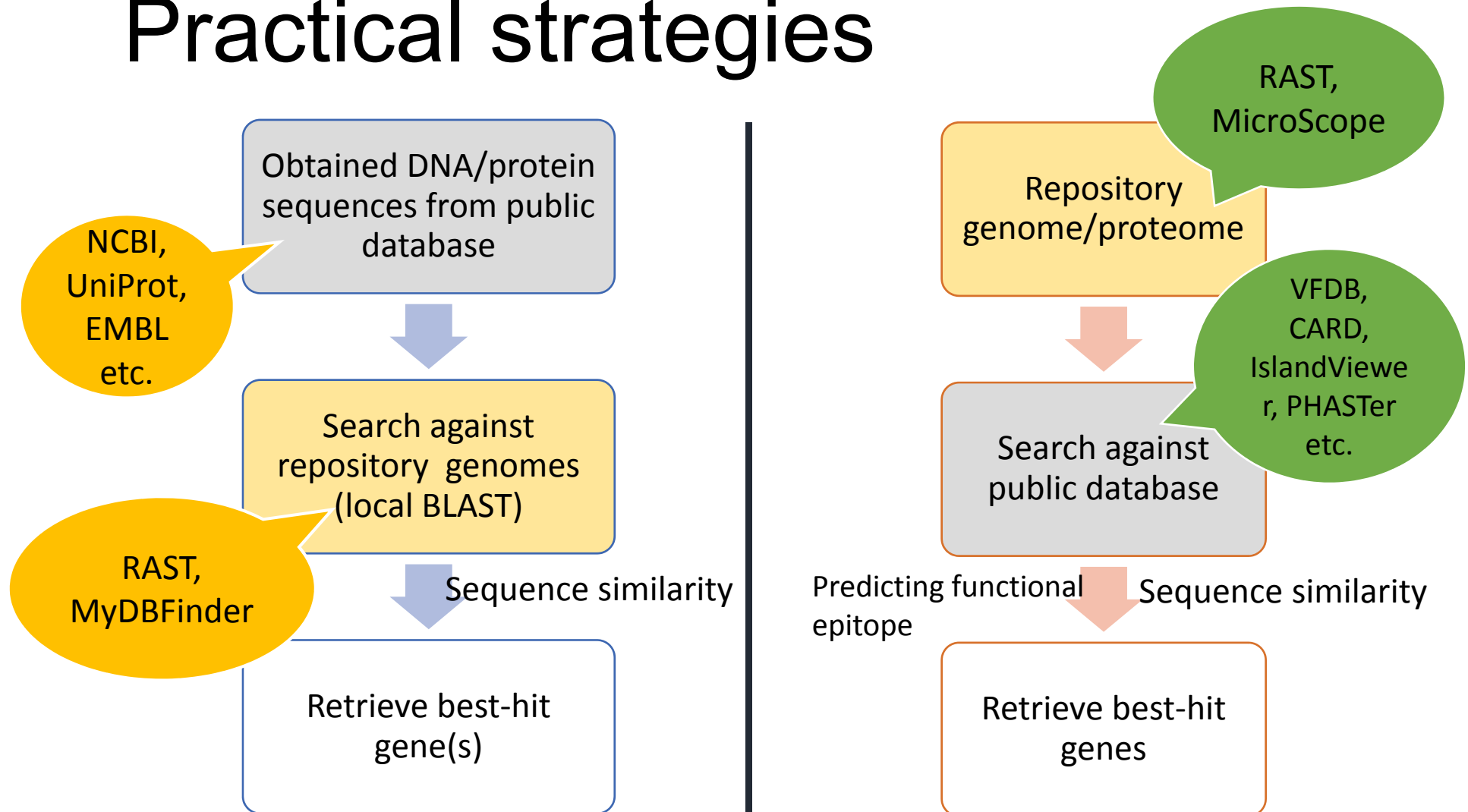


The screenshot displays the P2RP web interface for the analysis of *F. columnare_1214*, where 3077 proteins were analyzed. The interface is divided into two main sections: Two-Component Systems and Transcription Factors. The Two-Component Systems section includes a table for Histidine Kinases (16) and Response Regulators (18). The Transcription Factors section includes a table for Transcriptional Regulators (38) and One-Component Systems (11). The tables are color-coded to highlight specific proteins.

Two-Component Systems					
export TCSs (Excel) - (Tab-delimited) - (Aminoacid Fasta)					
Histidine Kinases (16)					
(Including Possible Incomplete Histidine Kinases 3)					
Histidine Kinase, Classic 9			Histidine Kinase, Hybrid 4		
Histidine Kinase, Short tandem					
Histidine Kinase, CheA					
Possible Incomplete Histidine Kinases 3					
Response Regulators (18)					
AtpA	AwhR_Nahr	ArxR	BdrR	ChdR	
ChcC	CheV	CheY 3	CET	CusA	
Cyc-C	FlyZ	HuX	HuR	IsrR	

Transcription Factors										
export TFs (Excel) - (Tab-delimited) - (Aminoacid Fasta)										
Transcriptional Regulators (38)										
One-Component Systems (11)										
AbrB	AlpA	ArxR 6	ArgR	ArxR 3	ArcC 3	CadY	CowR	CA		
Crp 6	CasR	ChfR	DeaR 1	DctR 1	FusA	Fu_Rap_repress	FlaC	FliC		
FliG	Fur 1	GntR	GutR	HrcA	HuR 2	IsrR	KarB	LuxI		
LexA 1	LoxR 1	LysR 3	LysR	MarR 4	NusA	MarR 1	MecI	Nga		
Hsr	HsrR	HsrR	FlyR 1	Phj	PuE	PucR	Rak	Rok_MacK		
PprR	Rfz 2	RuR	SarP	SNA	SarR	SarC	SpaA	TdkR 4		

Searching gene of interest: Practical strategies



Identification gene of interest from bacterial genome

- Download and extract specific CDS from a genome curated in RAST server (<http://rast.nmpdr.org/>)



RAST Rapid Annotation using Subsystem Technology
Rapid Annotation using Subsystem Technology version 1.0

The NMPDR, SEED-based, prokaryotic genome annotation service.
For more information about The SEED please visit theSEED.org.

Logout 48-hour

Jobs Overview





The overview below list all genomes currently processed and the progress on the annotation. To get a more detailed report on an annotation job, please click on the progress bar graphic in the overview.

Your personal jobs:

You currently have no jobs.

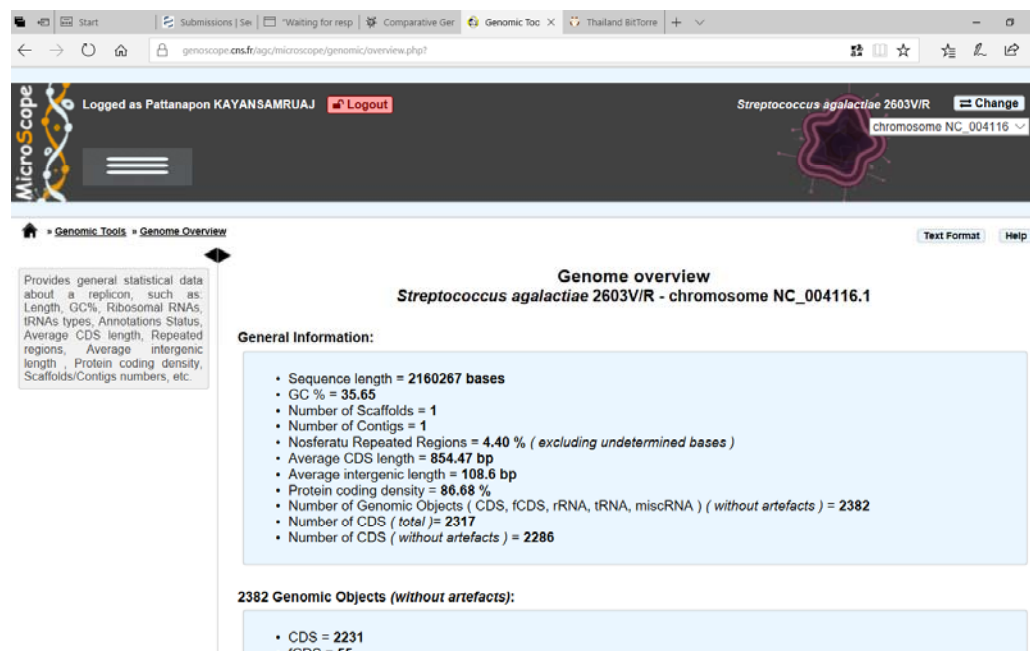
» [Upload a new genome](#)

Jobs of your Organization:

Job ▲ ▼	User ▲ ▼	Genome ID ▲ ▼	Genome Name ▲ ▼	Annotation Progress ▲ ▼
#24	olson	83331.3	Mycobacterium tuberculosis CDC1551	
#25	olson	150340.4	Vibrio sp. Ex25	
#26	olson	160490.3	Streptococcus pyogenes M1 GAS	
#28	olson	192222.3	Campylobacter jejuni subsp. jejuni NCTC 11169	

Identification gene of interest from bacterial genome

- **MicroScope** platform: a web-based platform for microbial comparative genome analysis and manual functional annotation (<http://www.genoscope.cns.fr/agc/microscope/home/index.php>)



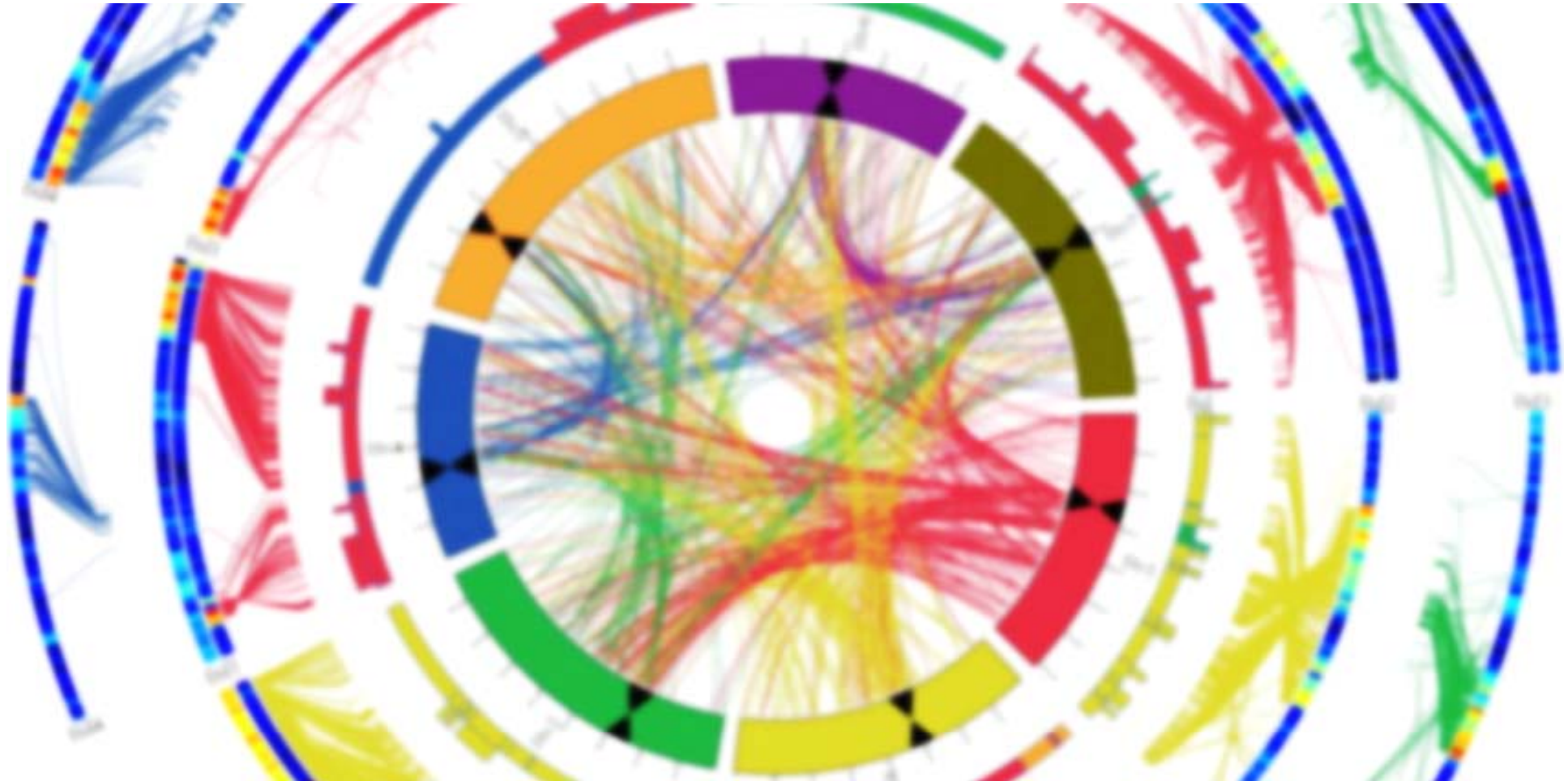
The screenshot displays the MicroScope web interface. The browser address bar shows the URL <http://www.genoscope.cns.fr/agc/microscope/genomic/overview.php?>. The user is logged in as Pattanapon KAYANSAMRUJ. The current view is for *Streptococcus agalactiae* 2603V/R, chromosome NC_004116.1. The page title is "Genome overview" and the subtitle is "Streptococcus agalactiae 2603V/R - chromosome NC_004116.1".

General Information:

- Sequence length = 2160267 bases
- GC % = 35.65
- Number of Scaffolds = 1
- Number of Contigs = 1
- Nosferatu Repeated Regions = 4.40 % (excluding undetermined bases)
- Average CDS length = 854.47 bp
- Average intergenic length = 108.6 bp
- Protein coding density = 86.68 %
- Number of Genomic Objects (CDS, fCDS, rRNA, tRNA, miscRNA) (without artefacts) = 2382
- Number of CDS (total) = 2317
- Number of CDS (without artefacts) = 2286

2382 Genomic Objects (without artefacts):

- CDS = 2231
- rCDS = 55



Comparative genomics and Phylogenomics |



COMPARATIVE GENOMICS

NHGRI FACT SHEETS

genome.gov

Researchers choose the appropriate time-scale of evolutionary conservation for the question being addressed.



Common features of different organisms such as humans and fish are often encoded within the DNA evolutionarily conserved between them.

By carefully comparing characteristics between various organisms → can **pinpoint regions** of similarity and difference.

Looking at **closely related species** such as humans and chimpanzees shows which genomic elements are unique to each.

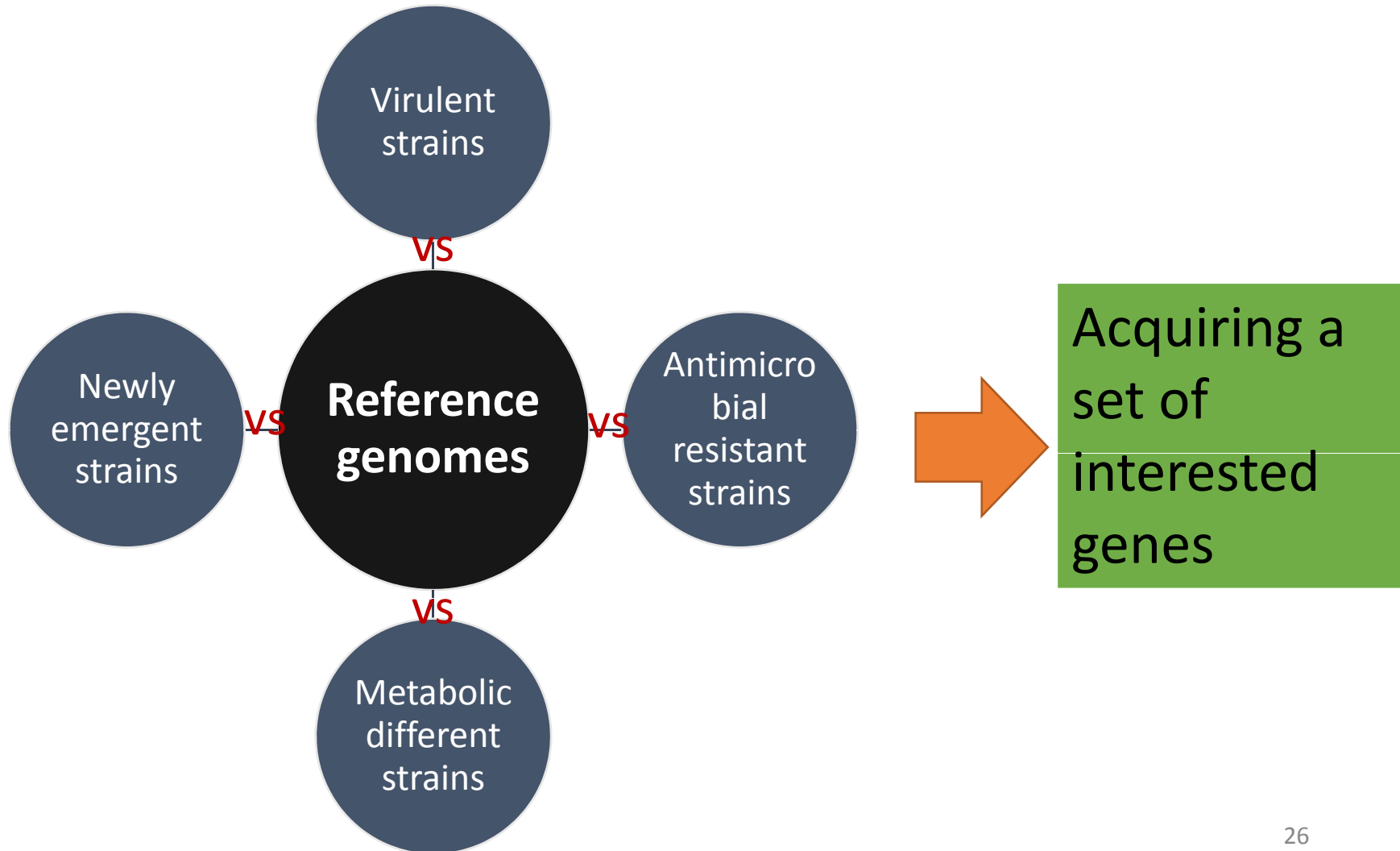


Genetic differences **within one species** such as our own can reveal variants with a role in disease.



NIH National Human Genome Research Institute

The benefits of comparative genomics

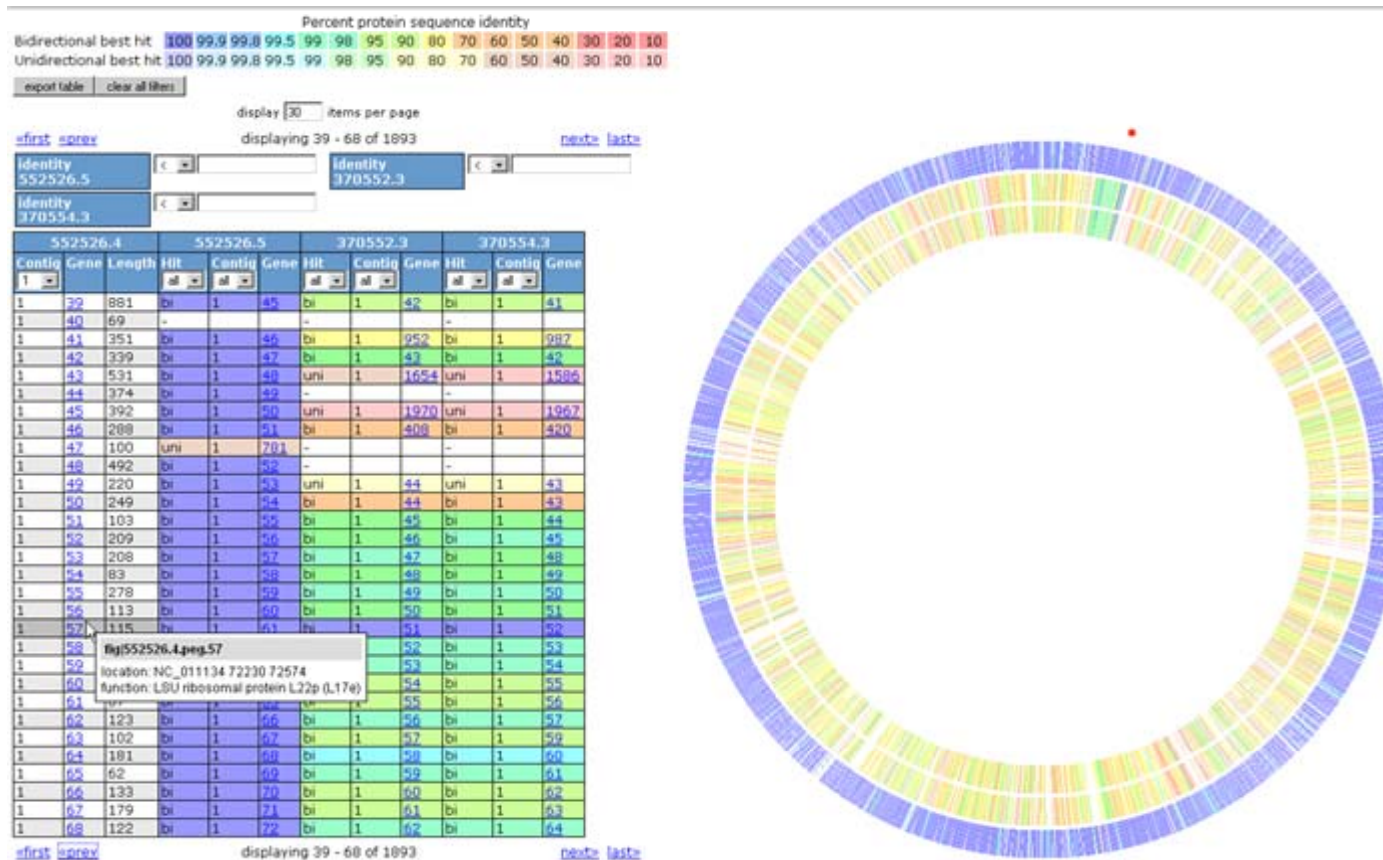


How to conduct comparative genome analysis: relevant bioinformatic tools

- Whole genome alignment (DNA-based)
- Ortholog clustering (protein-based)

Whole genome alignment

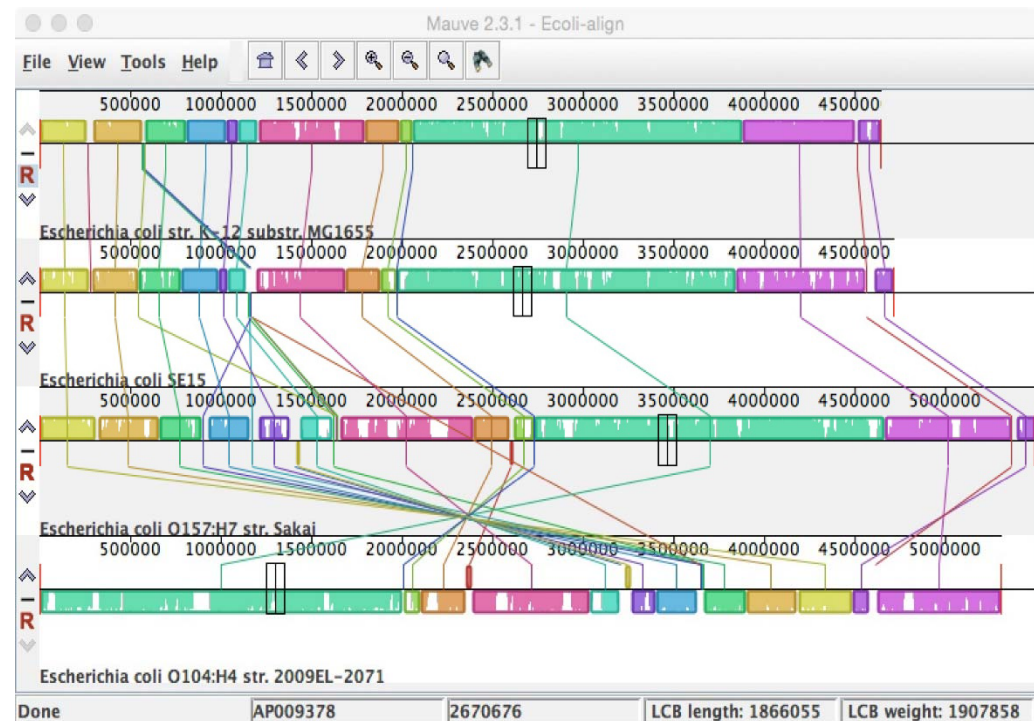
RAST & SEED viewer (<http://rast.nmpdr.org>)

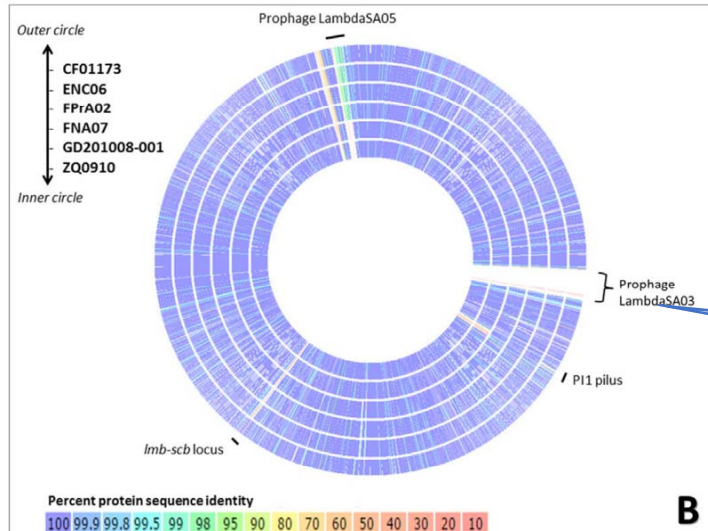


Whole genome alignment

MAUVE

- Input:
 - DNA fasta
 - GenBank file (.gbk)
- Visualizing genome rearrangement
- Java program
- <http://darlinglab.org/mauve/mauve.html>



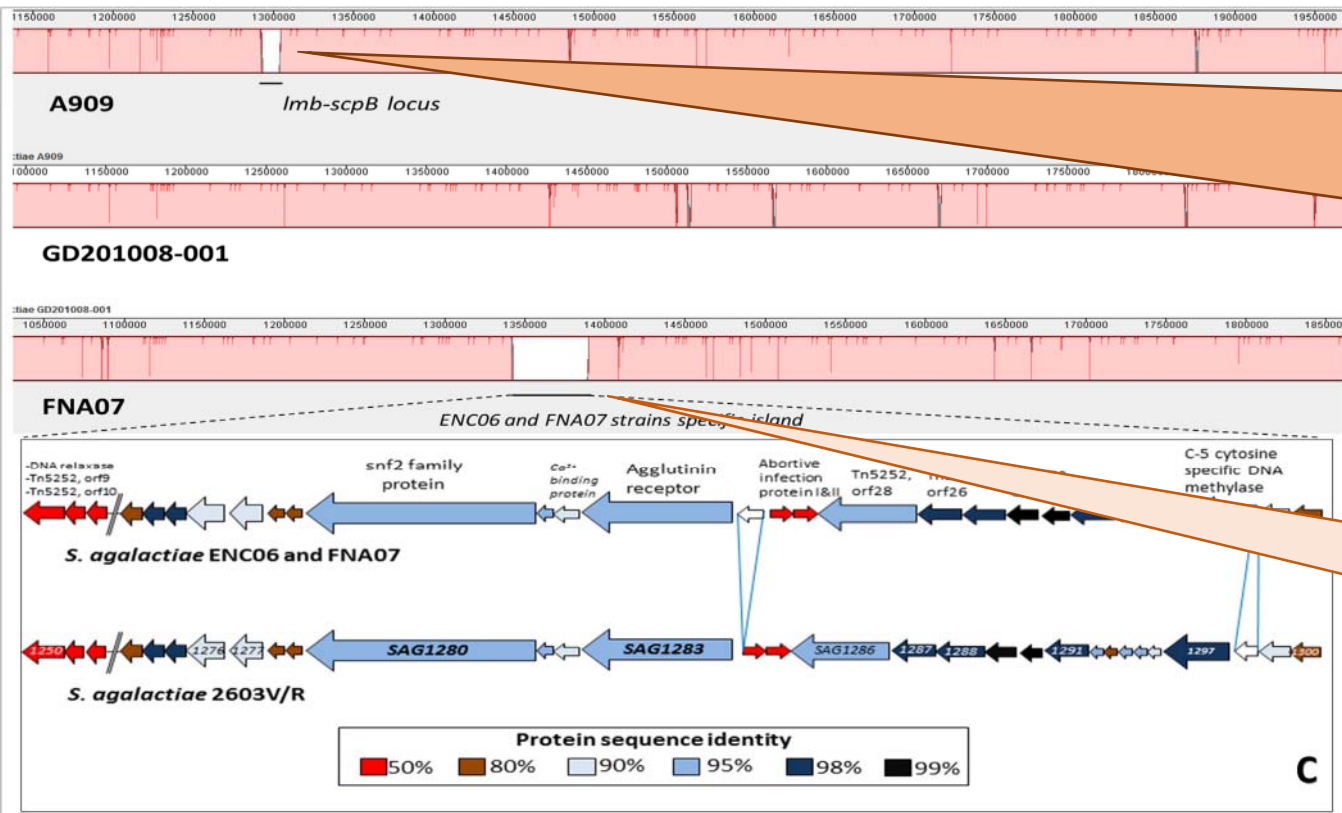


Genomic comparison between pathogenic *Streptococcus agalactiae* isolated from Nile tilapia in Thailand and fish-derived ST7 strains

Pattanapon Kayansamruaj ^a, Nopadon Pirarat ^b, Hidehiro Kondo ^c, Ikuo Hirono ^c, Channarong Rodkhum ^{a,*}

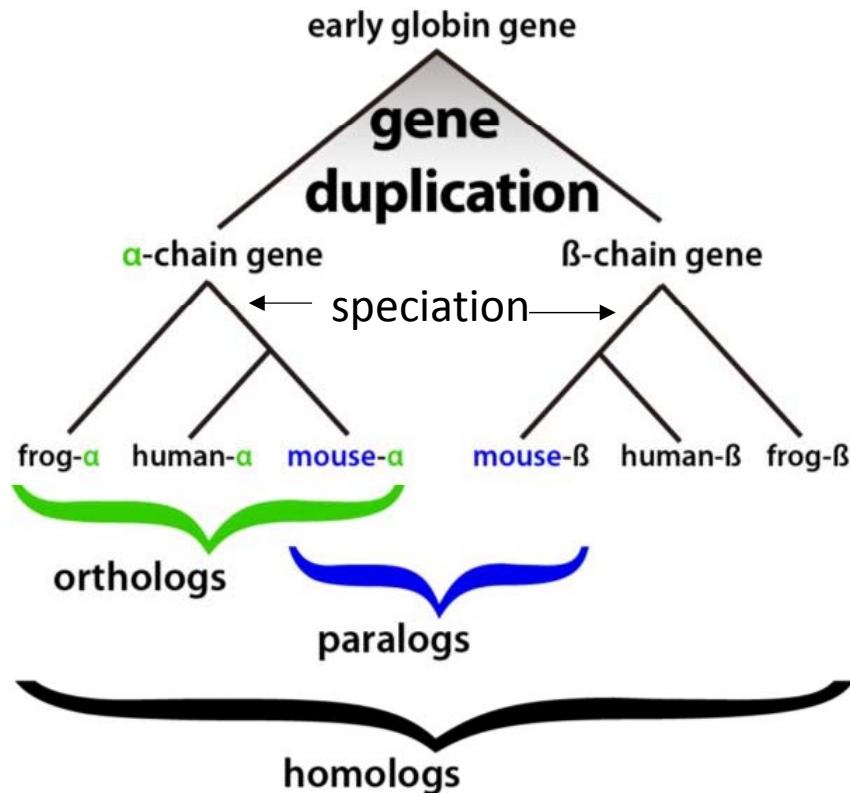
Prophage sequence is deleted in piscine *S. agalactiae* strains comparing to human strains

- Laminin binding protein and C5a peptidase genes were found in human strains but not in fish strains
- Deletion involved in niche adaptation



Genomic island specifically present in *S. agalactiae* Thai isolate

Ortholog clustering



Group orthologous protein together

- All vs All protein comparison
- Information used: sequence similarity, synteny
- Difficulty increases with taxa distance
- ***DEPEND ON ANNOTATION QUALITY!!***

Software: OrthoMCL, cd-hit, uclust etc.

Ortholog clustering

(almost) all-in-one comparative genomics tool

MicroScope

<http://www.genoscope.cns.fr/agc/microscope/home/index.php>

The screenshot displays the MicroScope web interface with the following components:

- Navigation Menu:** MaGe, Genomic Tools, Comparative Genomics (highlighted), Metabolism, Search/Export, Transcriptomics, Variant Discovery, User Panel, About.
- Left Sidebar:** Gene Phyloprofile, Regions of Genomic Plasticity, LinePlot, Fusion / Fission, PKGBS Synteny Statistics, RefSeq Synteny Statistics, Pan/Core-Genome, Resistome, Virulome.
- LinePlot:** A plot showing synteny between chromosomes BRADO and BBTA of *Bradyrhizobium* sp. ORS278.
- Resistome:** Antibiotic Resistance for *Acinetobacter baumannii* AYE. Includes result statistics:
 - Number of CDS (Total) = 55
 - Number of CDS (Homolog Model) = 52
 - Number of CDS (Variant Model) = 3
 - Number of CDS by replicon:
 - ABAYE (chromosome) : 55Buttons for "CARD Proteins Homologs" and "CARD Proteins Variants" are visible.
- Regions of Genomic Plasticity:** A circular plot showing genomic plasticity across a 3500 kbp region.
- Gene phyloprofile:** A table of gene profiles across different species.

Gene ID	Label	Begin	End	Enzyme	Gene	Product	Lactobacillus rhamnosus ATCC 19835	Lactobacillus rhamnosus DSM 20173	Lactobacillus rhamnosus ATCC 21332	Lactobacillus rhamnosus DSM 20173	Lactobacillus rhamnosus ATCC 21332
0819	12045	14882	automatic/finished	shakA	shikimate Hcp70, ac-shikimate w/o DnaJ		No HI	No HI	No HI	No HI	No HI
0820	24653	27489	automatic/finished	hds	histidinol-5-phosphate hydrolase		No HI	No HI	No HI	No HI	No HI
0824	29638	29729	automatic/finished	rhcC	Regulator of ribonucleotide hydrolase 2 (rhc-1)		No HI	No HI	No HI	No HI	No HI
0841	33444	36665	automatic/finished	carB	carbamoyl-phosphate synthase large subunit		Yes HI	Yes HI	Yes HI	Yes HI	Yes HI
0873	68305	68064	automatic/finished	andC	L-lysine 6-phosphate 4-epimerase		Yes HI	Yes HI	Yes HI	Yes HI	Yes HI
0876	99209	102011	automatic/finished	hnsA	RNA-dependent polynucleotide transferase		No HI	No HI	No HI	No HI	No HI
08113	109132	110283	automatic/finished	hslZ	GTP binding hslZ-like hot domain protein		Yes HI	Yes HI	Yes HI	Yes HI	Yes HI
08116	112108	114811	automatic/finished	hscA	proteasome hscA-like subunit, ATPase		No HI	No HI	No HI	No HI	No HI
08195	180794	180492	automatic/finished	pts	5'-methylthioadenosine-S-adenosylmethionine synthetase		No HI	No HI	No HI	No HI	No HI
08204	199974	200897	automatic/finished	_	30S ribosomal protein S2		Yes HI	Yes HI	Yes HI	Yes HI	Yes HI
- Pan/core Genome:** A Venn diagram showing the overlap of genes across five *Lactobacillus rhamnosus* strains: RD011, LAKE2.1, ATCC 21602, HN001, and Lc 705. The central intersection contains 2097 genes.

Comparative Genomics and Transcriptional Analysis of *Flavobacterium columnare* Strain ATCC 49512

Hasan C. Tekedar¹, Attila Karsi¹, Joseph S. Reddy², Seong W. Nho¹, Safak Kalindamar¹ and Mark L. Lawrence^{1*}

the strain ATCC 49512 genome to four other *Flavobacterium* genomes. In this analysis, we identified predicted proteins whose functions indicate *F. columnare* is capable of denitrification, which would enable anaerobic growth in aquatic pond sediments.

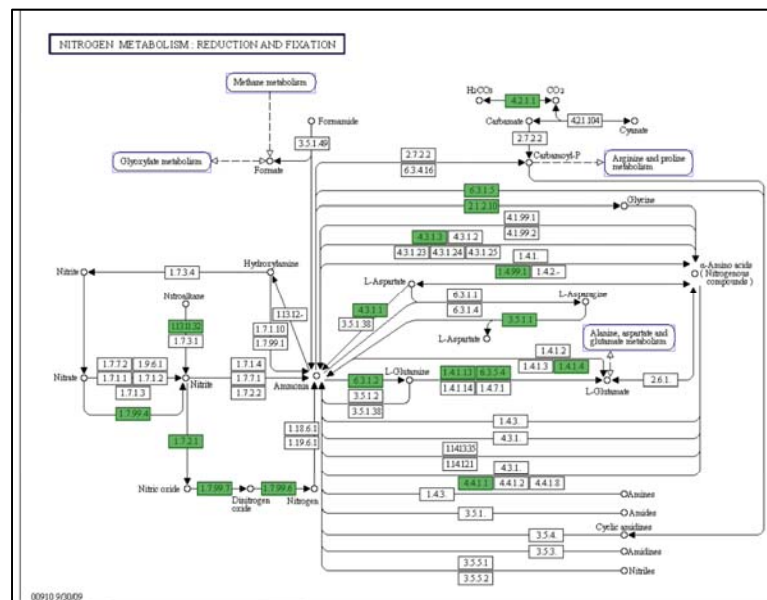
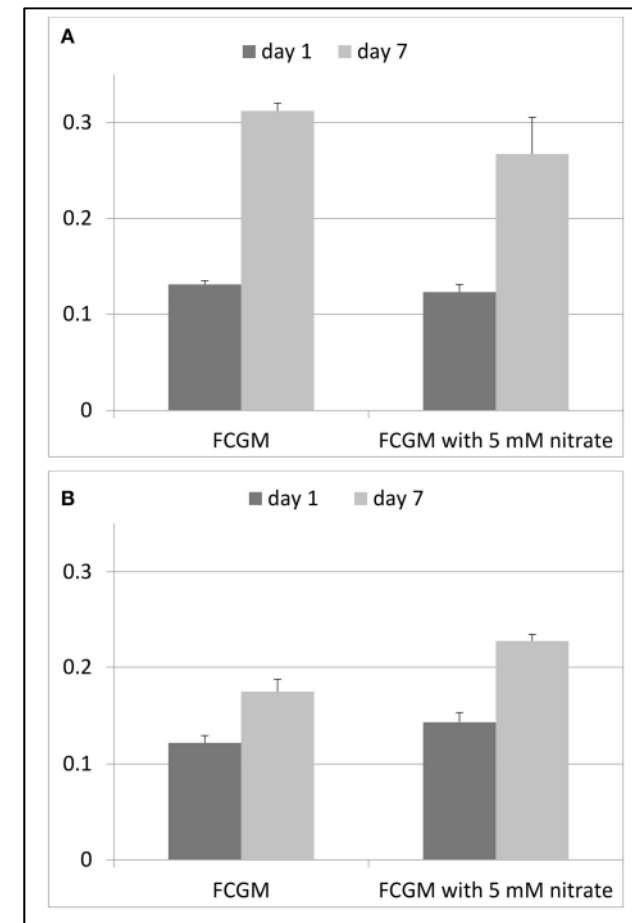


FIGURE 5 | Growth of *F. columnare* ATCC 49512 in FCGM and FCGM supplemented with 5 mM nitrate. (A) Growth under aerobic conditions and (B) growth under anaerobic conditions.



Taxonomic classification based on NGS data

Standard parameters for taxonomic revision in the genomic era

I. Average nucleotide identity (ANI)

- The ANI values between genomes of the same species are typically **above 95%**
- <https://www.ezbiocloud.net/tools/ani>
- <http://ani.mypathogen.cn/>

II. Digital DNA-DNA hybridization (dDDH)

- Result can be interpreted in the same scale as conventional DDH
- Cut-off: 70% (same species), 80% (same subspecies)
- <http://ggdc.dsmz.de/ggdc.php>

Inference of taxonomic position of unknown bacterium must be made by comparison with TYPE STRAIN only!!!

Taxonomic classification based on NGS data

Standard parameters for taxonomic revision in the genomic era

- Currently, **International Journal of Systematic and Evolutionary Microbiology** accepts the WGS as genotypic tool for species demarcation
- Can be used *alternative* to conventional **DDH**

However, polyphasic approach (**phenotype**) is still **mandatory** to define nov. sp.

INTERNATIONAL
JOURNAL OF SYSTEMATIC
AND EVOLUTIONARY
MICROBIOLOGY

TAXONOMIC DESCRIPTION
Du et al., *Int. J. Syst. Evol. Microbiol.*
DOI 10.1099/ijsem.0.002388



Oceanibaculum nanhaiense sp. nov., isolated from surface seawater

Yaping Du, Xiupian Liu, Qiliang Lai, Weiwei Li, Fengqin Sun and Zongze Shao*

International Journal of Systematic and Evolutionary Microbiology (2014), **64**, 316–324

DOI 10.1099/ijse.0.054171-0

Review

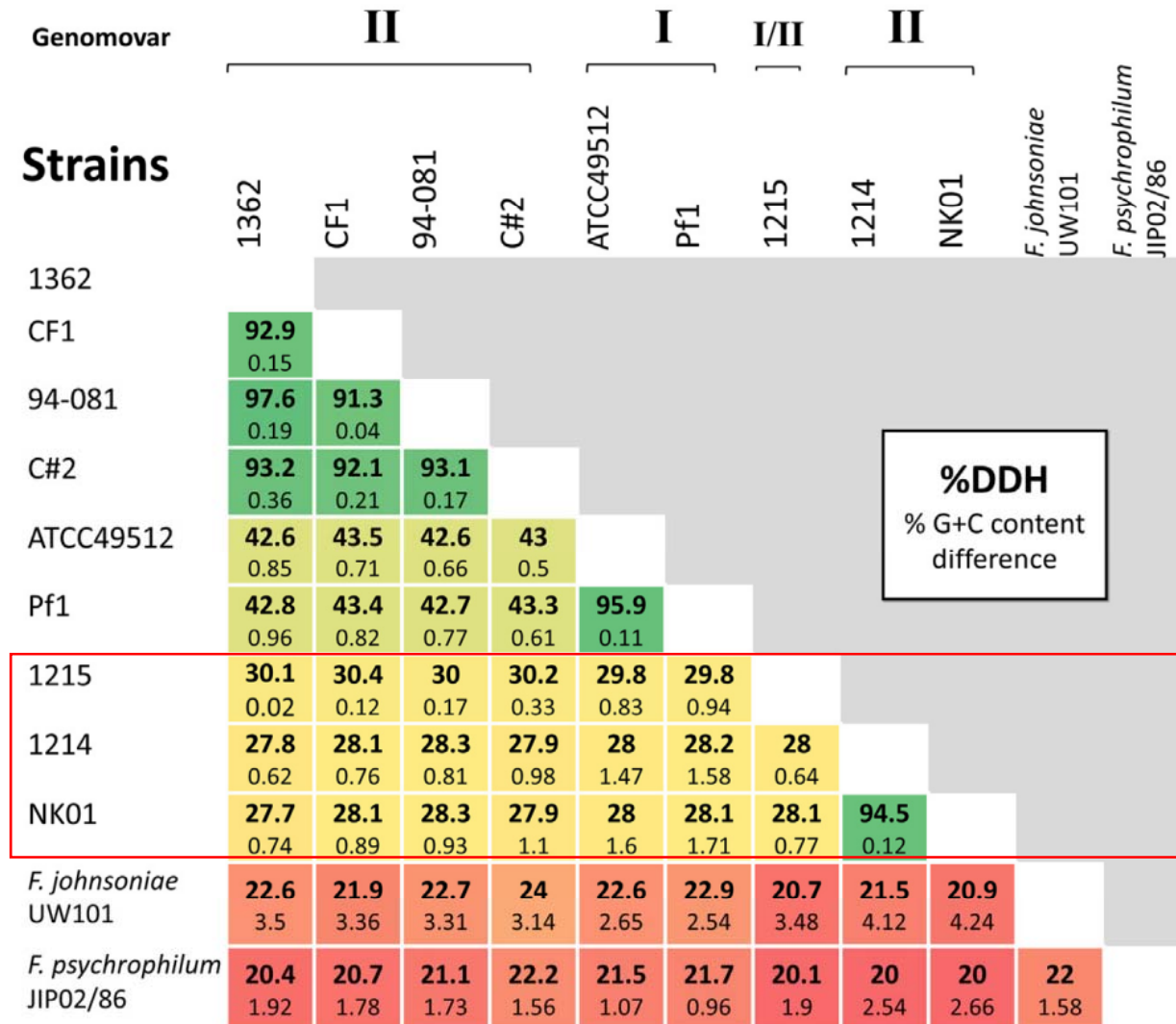
Integrating genomics into the taxonomy and systematics of the *Bacteria* and *Archaea*

Jongsik Chun¹ and Fred A. Rainey²

Comparative genome analysis of fish pathogen *Flavobacterium columnare* reveals extensive sequence diversity within the species



Pattanapon Kayansamruaj ^{a,b,*}, Ha Thanh Dong ^c, Ikuo Hirono ^d, Hidehiro Kondo ^d, Saengchan Senapin ^e, Channarong Rodkhum ^{a,**}



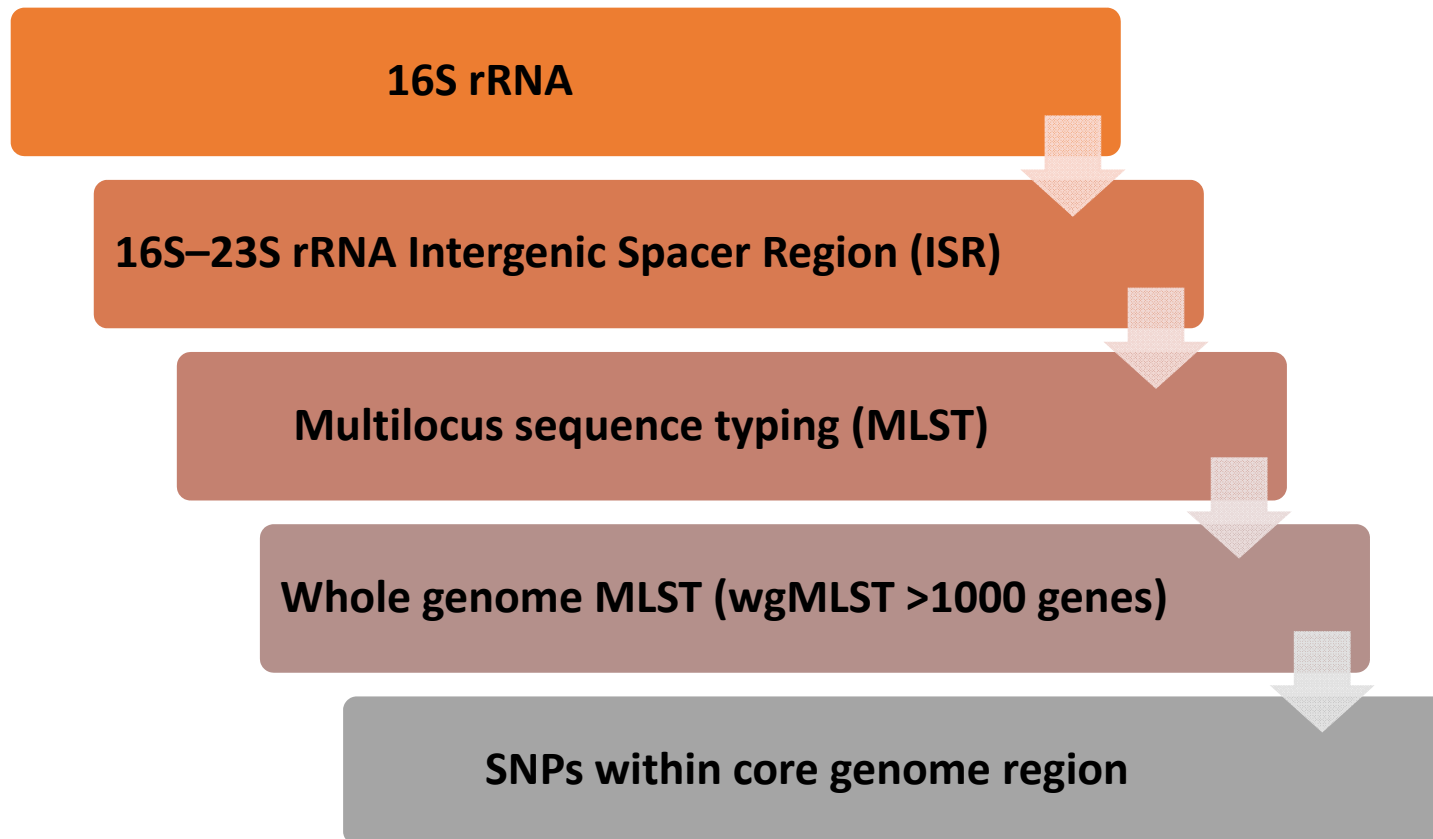
Phylogenomics

What is “phylogenomics”

- Evolution + Genomics
- **Phylogenomics** draws information by comparing **entire genomes**, or at least **large portions of genomes**, whereas phylogenetic compares only a small number of genes
- Can infer evolutionary relationships, gene family evolution (duplication, deletion, recombination), lateral gene transfer, taxonomy etc.

Genetic markers for reconstructing evolutionary history

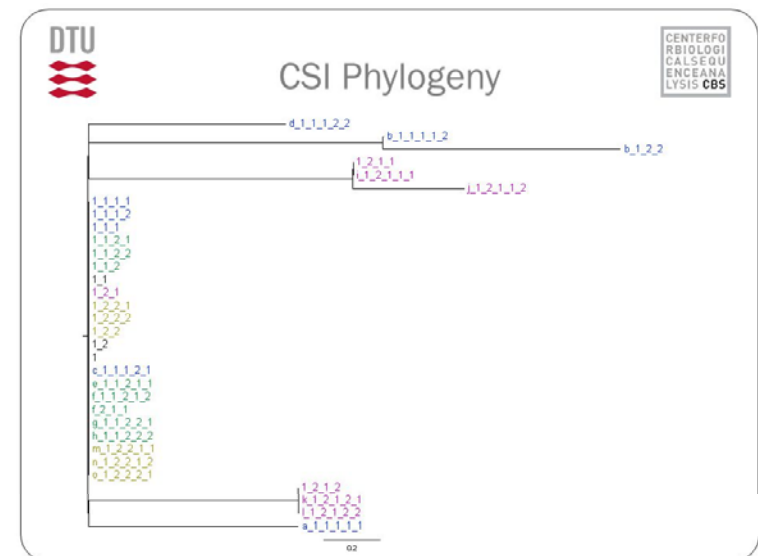
Phylogeny reconstruction based on sequence alignment



Genetic markers for reconstructing evolutionary history

Extracting concatenated SNP sequence from core genome region

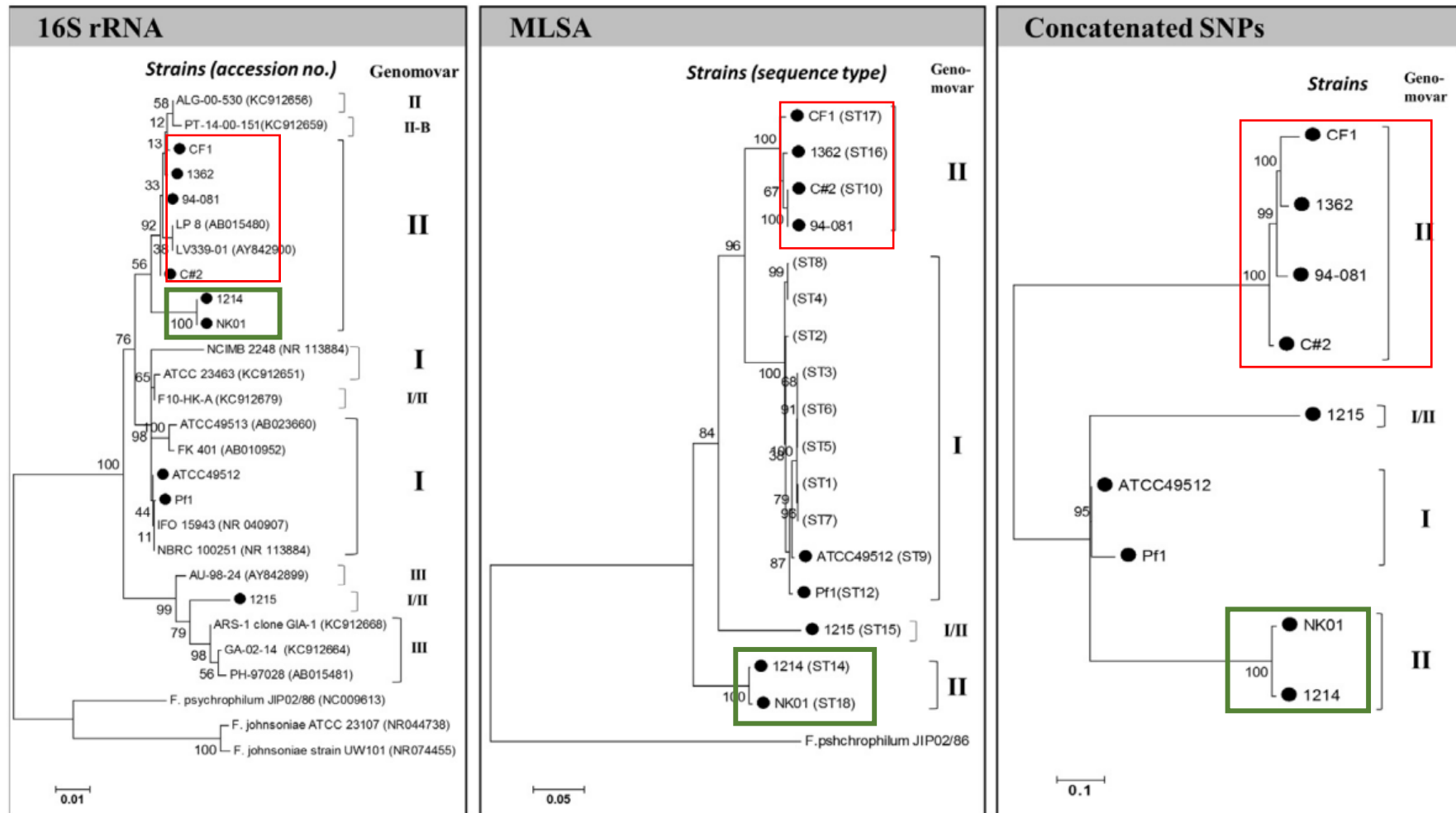
- CSI phylogeny
- <https://cge.cbs.dtu.dk/services/CSIPhylogeny/>
- Automated web-based tool
- Obtained multiple alignment file to construct phylogenetic tree using secondary program



Comparative genome analysis of fish pathogen *Flavobacterium columnare* reveals extensive sequence diversity within the species

Pattanapon Kayansamruaj^{a,b,*}, Ha Thanh Dong^c, Ikuo Hirono^d, Hidehiro Kondo^d, Saengchan Senapin^e, Channarong Rodkhum^{a,**}

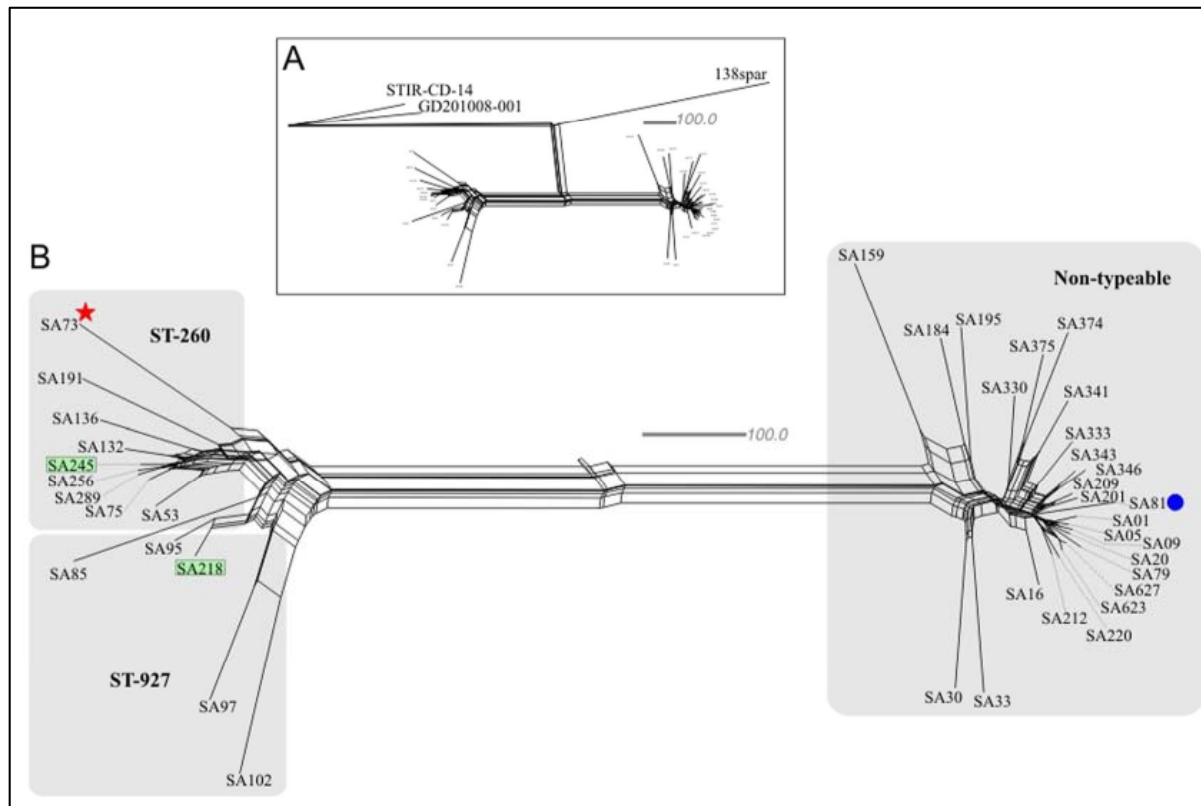
Phylogenetic relationship of *F. columnare* strains generated from different genetic markers



Genetic markers for reconstructing evolutionary history

- Revisited *PubMLST*

- Phylogenetic network generated from wgMLST data

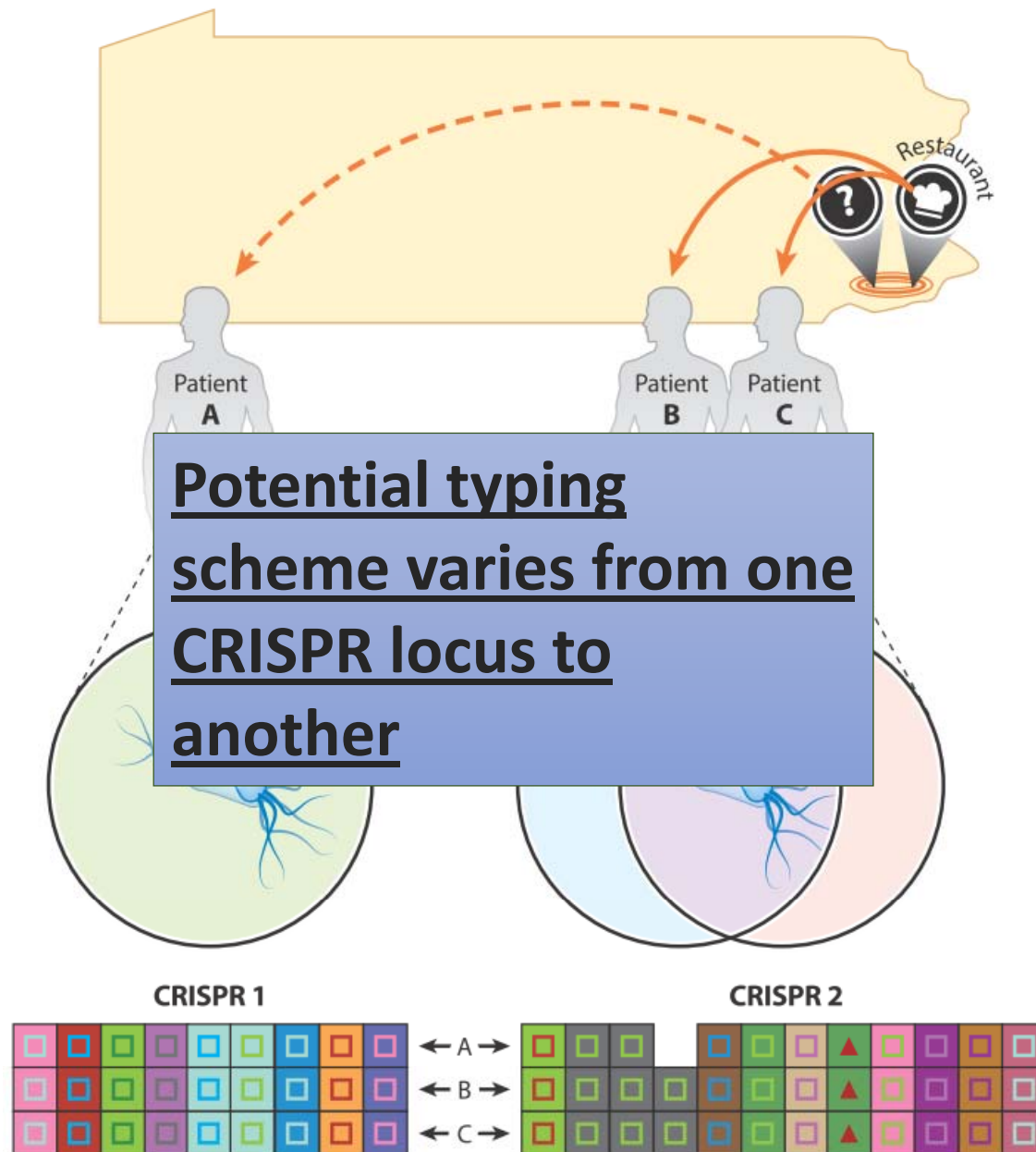


Phylogenomic NeighborNet network of whole genome MLST (wgMLST) data inferred that at least 2 population of fish pathogenic *S. agalactiae* were present in Brazil

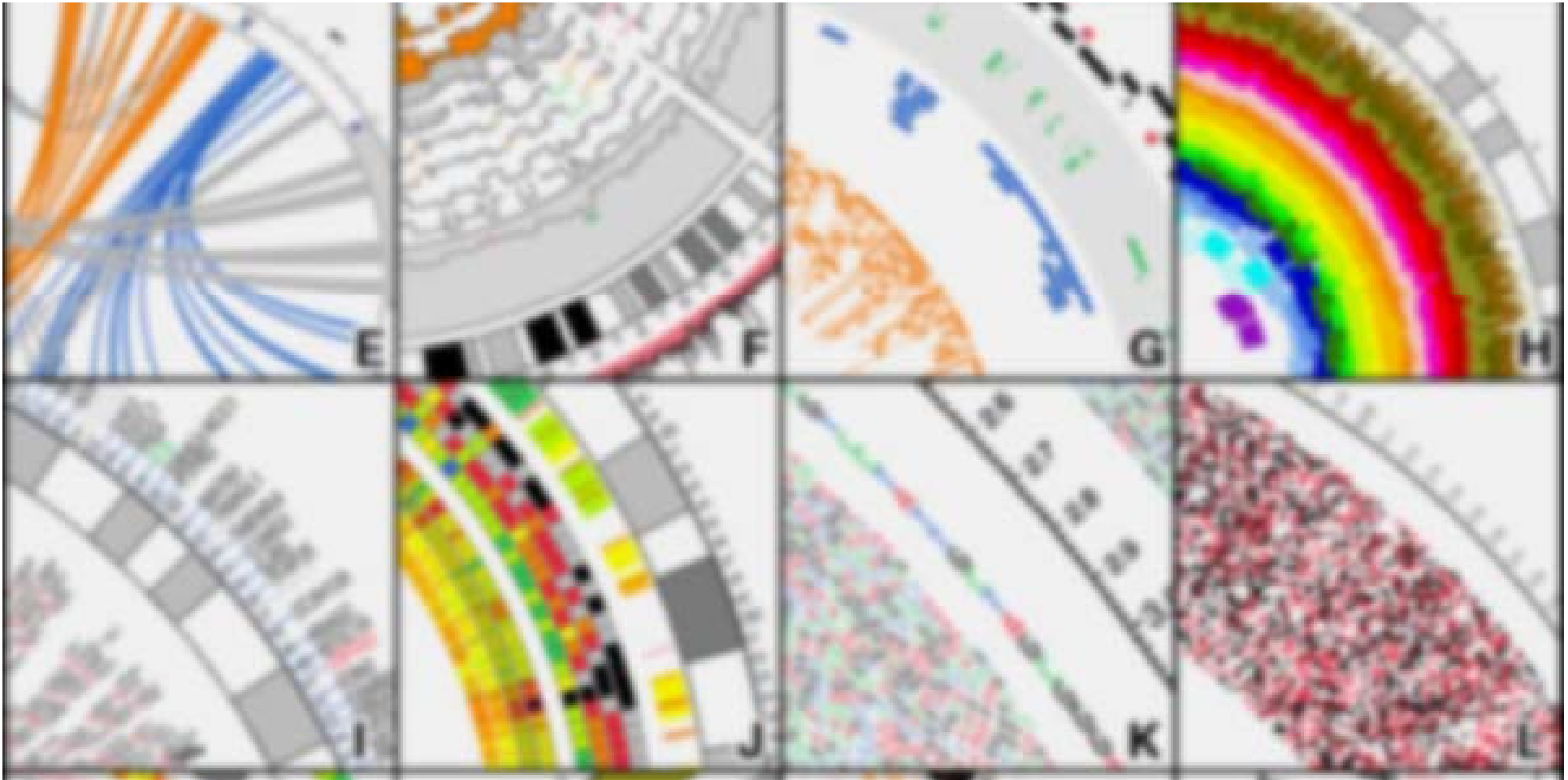
Large-scale genomic analyses reveal the population structure and evolutionary trends of *Streptococcus agalactiae* strains in Brazilian fish farms

CRISPR-Based Typing and Next-Generation Tracking Technologies

Rodolphe Barrangou^{1,2,*} and Edward G. Dudley²



- Three *Salmonella* Enteritidis isolates were
- received during the **same time frame**
- All three are **same pulsotype** (same PFGE pattern)
- Pattern B and C ate at the same restaurant
- No epidemiologic data linked to patient A
- CRISPR sequencing revealed a **one-spacer difference**
- **Two separate incidents** were responsible for these illnesses



Getting
published |

List of published genomic studies featured fish pathogenic flavobacteria (2016-2017)

- Tekedar HC, Karsi A, Reddy JS, Nho SW, Kalindamar S, Lawrence ML. **Comparative Genomics and Transcriptional Analysis of *Flavobacterium columnare* Strain ATCC 49512**. *Frontiers in Microbiology*. 2017;8:588. doi:10.3389/fmicb.2017.00588.
- Kumru S, Tekedar HC, Gulsoy N, Waldbieser GC, Lawrence ML and Karsi A (2017) **Comparative Analysis of the *Flavobacterium columnare* Strain ATCC 49512**. *Infection, Genetics and Evolution* 54: 7-17.
- Castillo D, Christiansen RH, Dalsgaard I, Madsen L, Espejo R & Middelboe M (2016) **Comparative genome analysis provides insights into the pathogenicity of *Flavobacterium psychrophilum***. *PLOS ONE* 11: e0152515.
- Kayansamruaj P, Dong HT, Hirono I, Kondo H, Senapin S & Rodkhum C (2017) **Comparative genome analysis of fish pathogen *Flavobacterium columnare* reveals extensive sequence diversity within the species**. *Infection, Genetics and Evolution* 54: 7-17.

All these are **descriptive studies** and **not hypothesis testing** type of research



Is it possible to publish paper based on WGS data only?



If you have **only genome sequences** without further analysis

- Submit data to *NCBI*

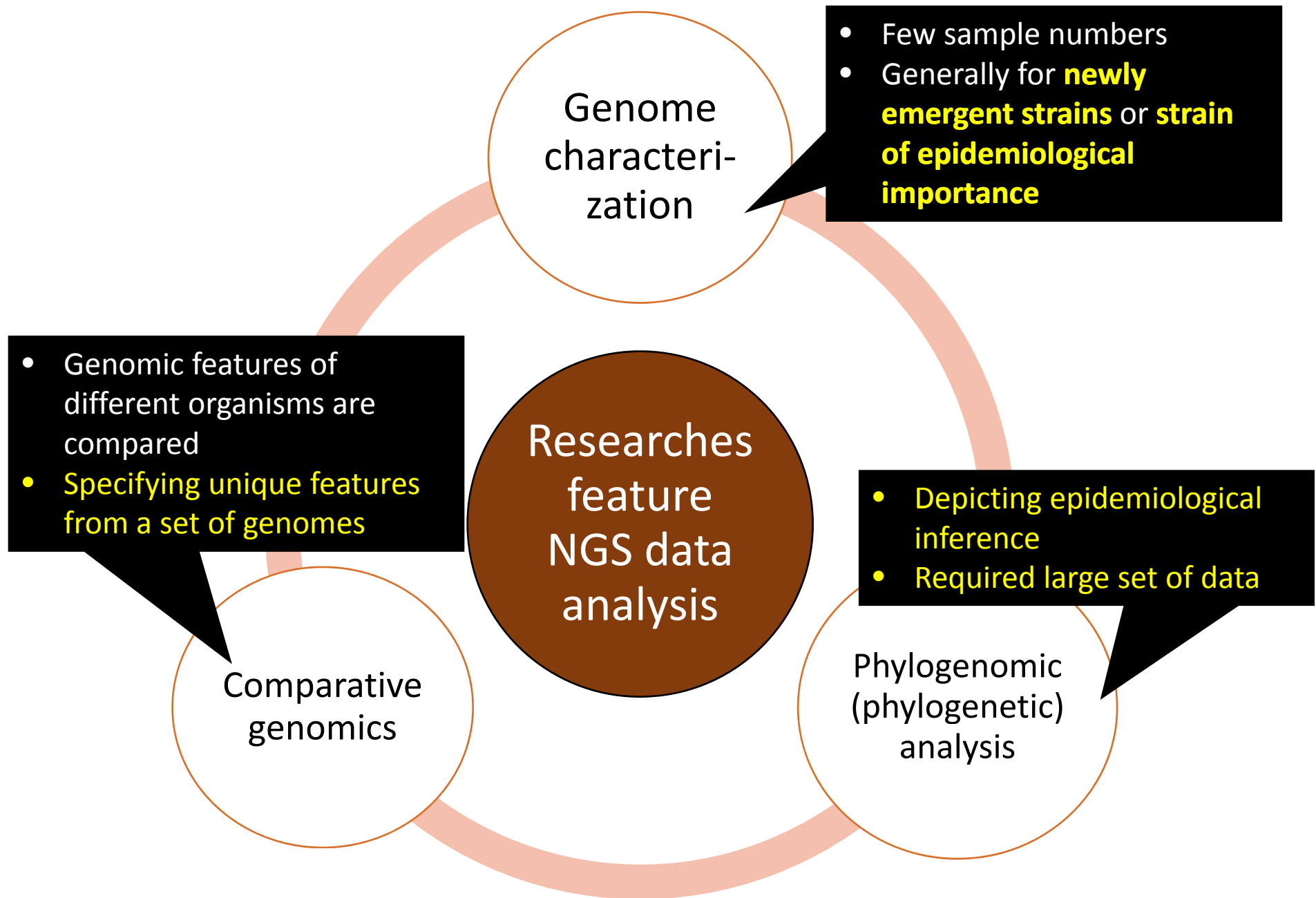


If you provide **genome sequences** with **basic annotation**

- *Genome Announcements*
- *Standard in Genomic Sciences*

If you provide **genome sequences** with **specific analyses**. Study must be **constructive** and **informative**

- *Frontiers, PLOS ONE, BMC, Genome Research, MEEGID etc.*



EXTENDED GENOME REPORT

Open Access



Draft genome sequence and characterization of commensal *Escherichia coli* strain BG1 isolated from bovine gastrointestinal tract

Audrey Segura^{1††} , Pauline Auffret^{1†}, Christophe Klopp², Yolande Bertin¹ and Evelyne Forano¹

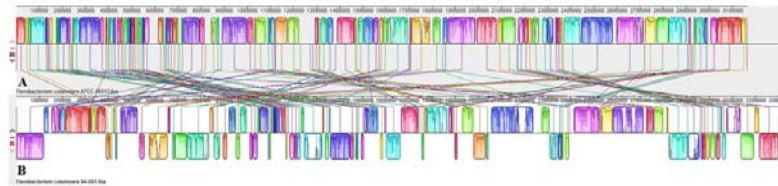
Table 6 Adherence systems encoded by the *E. coli* BG1 genome

Adherence system	Gene or genes cluster	Pathotype ^a	in vitro cell adherence ^b	Receptor
Curli fimbriae	<i>csgDEFG, csgBA</i>	EHEC, ETEC, aEPEC, APEC	T84	Matrix, plasma proteins
EhaA autotransporter	<i>ehaA</i>	EHEC, EAEC, ETEC, AIEC, EPEC	Primary bovine epithelial cells (terminal rectum)	Unknown
EhaB autotransporter	<i>ehaB</i>	EHEC, UPEC, ETEC, EIEC, EAEC	NA ^c	Collagen I, laminin
EhaC autotransporter	<i>ehaC (yfaL)</i>	EHEC, UPEC	Unknown	Unknown
ECP (<i>E. coli</i> Common Pilus)	<i>ecpRABCDE</i>	ETEC, EHEC, NMEC, EAEC, aEPEC, septicemia	HT29, Hep-2, HeLa, HTB-4	Arabinosyl residues
ELF (<i>E. coli</i> Laminin-binding Fimbriae)	<i>ycbQRST</i>	EHEC, aEPEC	HT29, Hep-2, MDBK	Laminin
F9 Fimbriae	<i>z2200-z2206</i>	EHEC, UPEC, APEC, AIEC, EAEC, EPEC	EBL	Bovine fibronectin, Galβ1-3GlcNAc

Table 4 Number of genes associated with general COGs functional categories

Code	Value	% age ^a	Description
J	250	6.55	Translation, ribosomal structure and biogenesis
A	2	0.05	RNA processing and modification
K	293	7.68	Transcription
L	154	4.04	Replication, recombination and repair
B	0	0.00	Chromatin structure and dynamics
D	41	1.07	Cell cycle control, cell division, chromosome partitioning
V	93	2.44	Defense mechanisms
T	176	4.61	Signal transduction mechanisms
M	271	7.11	Cell wall/membrane/envelope biogenesis
N	156	4.09	Cell motility
U	60	1.57	Intracellular trafficking, secretion, and vesicular transport
O	153	4.01	Post-translational modifications, protein turnover, chaperones
C	282	7.39	Energy production and conversion
G	381	9.99	Carbohydrate transport and metabolism
E	335	8.78	Amino acid transport and metabolism
F	101	2.65	Nucleotide transport and metabolism
H	169	4.43	Coenzyme transport and metabolism
I	119	3.12	Lipid transport and metabolism
P	190	4.98	Inorganic ion transport and metabolism
	53	1.39	Secondary metabolites biosynthesis, transport and catabolism
	211	5.53	General function prediction only
	238	6.24	Function unknown
	750	7.43	Not in COGs

Complete genome sequence of *Streptococcus agalactiae* strain SA20-06, a fish pathogen associated to meningoencephalitis outbreaks

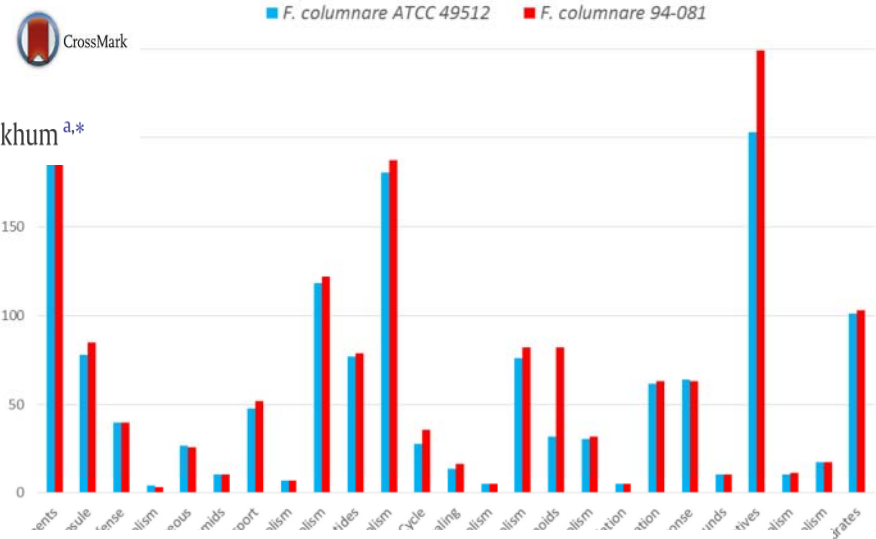
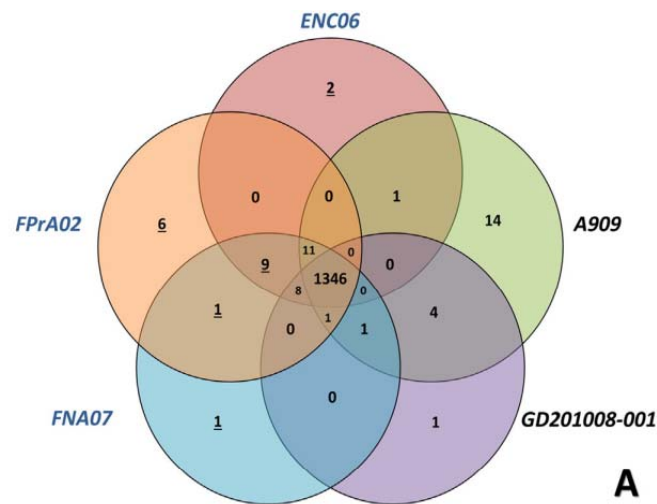


Comparative Analysis of the *Flavobacterium columnare* Genomovar I and II Genomes

Salih Kumru^{††}, Hasan C. Tekedar^{††}, Nagihan Gulsoy², Geoffrey C. Waldbieser³, Mark L. Lawrence^{1*} and Attila Karsi^{1*}

Genomic comparison between pathogenic *Streptococcus agalactiae* isolated from Nile tilapia in Thailand and fish-derived ST7 strains

Pattanapon Kayansamruaj^a, Nopadon Pirarat^b, Hidehiro Kondo^c, Ikuo Hirano^c, Channarong Rodkhum^{a,*}



Journal of Fish Diseases 2014

doi:10.1111/jfd.12319

Cor. **Genomic comparison of virulent and non-virulent *Streptococcus agalactiae* in fish**

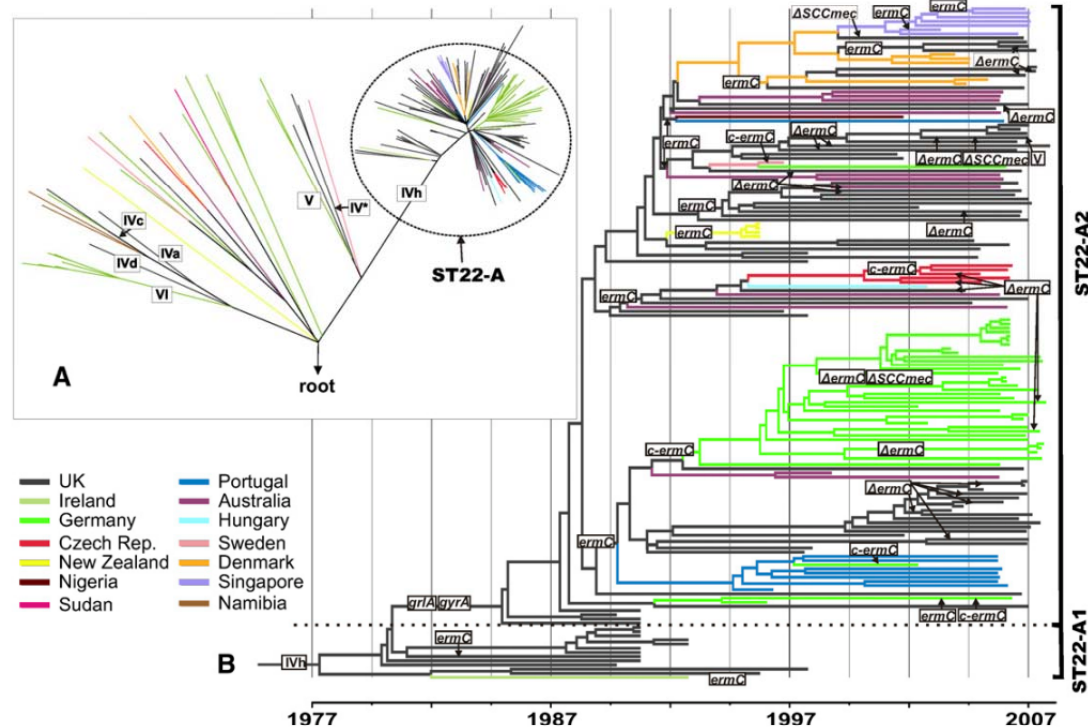
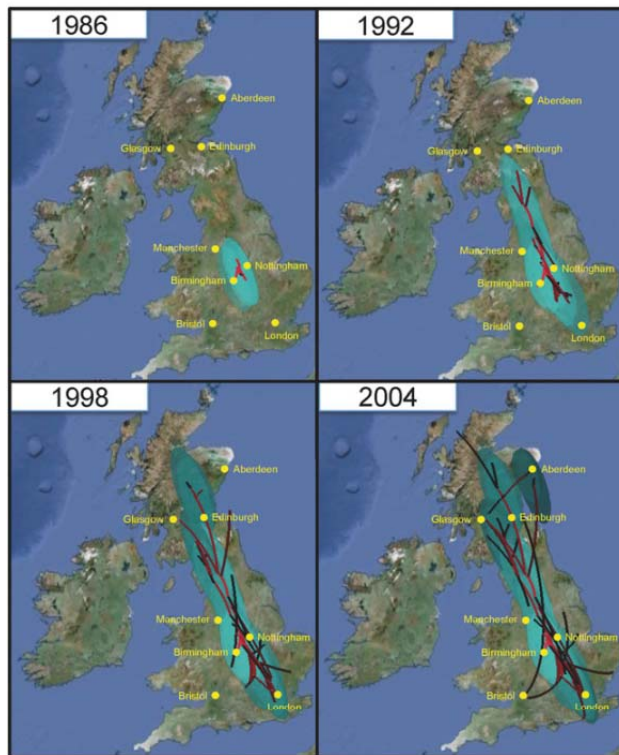
C M J Delannoy^{1,2}, R N Zadoks^{2,3}, M Crumlish¹, D Rodgers², F A Lainson², H W Ferguson⁴, J Turnbull¹ and M C Fontaine²



A genomic portrait of the emergence, evolution, and global spread of a methicillin-resistant *Staphylococcus aureus* pandemic

Matthew T.G. Holden,^{1,25} Li-Yang Hsu,^{1,2} Kevin Kurt,³ Lucy A. Weinert,^{4,22}

122 isolates of *S. aureus* ST22 were sequenced



SPREAD: Spatial Phylogenetic Reconstruction of Evolutionary Dynamics

BEAST analysis based on core genome sequence

Today's topics

- Impact of Next-Generation sequencing to science
- Generic workflow in bacterial genome analysis
- High-throughput screening of bacterial genomes
- Comparative genomics and Phylogenomics
- Getting published

**THANK YOU
FOR
YOUR
ATTENTION!
ANY QUESTIONS?**



I have to leave early today.
Can you run the gel for me?...please!

Running the gel?
I'm 100 percent
computational my friend!



www.biocomicals.com

Genetic markers for reconstructing evolutionary history

- **Revisited *PubMLST***

- Comparison of bacterial genotypes based on several typing schemes:
 - *MLST, capsule genes, antibiotic resistance genes, antigen typing, metabolic pathways typing, whole genome MLST (wgMLST) and much more...*
- Locus allocation of bacterial isolates can be accomplished automatically
 - Submit your genome assembly to **BIGSdb**
- <https://pubmlst.org/>

Genetic markers for reconstructing evolutionary history

Genome submission to BIGSdb

<https://pubmlst.org/>

Streptococcus agalactiae (group B streptococcus, GBS) MLST Databases

The *Streptococcus agalactiae* MLST website contains two linked databases - one for allelic profiles and sequences, the other for isolate information. This structure offers advantages over a single database system. Further details can be found [here](#).

- Information
 - Primers and protocol used for amplification and sequencing
- Access main databases
 - Sequence and profile definitions
 - Isolates**
- Policy document
- Submission of data
- Submission history
- BIGSdb software
- Recent publications using MLST in Streptococcus research

Please note we are now using a new submission system -

This MLST scheme was developed by Nicola Jones (Nuffield Clinical Laboratory Services, John Radcliffe Hospital, Oxford) and Brian Spratt (now at the Department of Infectious Disease Epidemiology, Faculty of Medicine, Imperial College, London) with Derrick Crook (Nuffield Department of Clinical Laboratory Science, John Radcliffe Hospital, Oxford, UK) and Man-Su...

Citing the database

The preferred format for citing this website in publications is:

This publication made use of the *Streptococcus agalactiae*

[Submission template.xls](#)

Streptococcus agalactiae isolates database

Log in

- Query database
 - Search or browse database
 - Search by combinations of loci (profiles)
- Option settings
 - Set general options - including isolate table field handling.
 - Set display and query options for locus, schemes or scheme fields.
- Submissions**
 - Manage submissions**
- General information
 - Isolates: 4124
 - Last updated: 2017-11-06
 - Defined field values
 - Update history
 - About BIGSdb

Breakdown

- Single field
- Two field
- Unique combinations
- Scheme and alleles
- Publications
- Sequence bin

Export

- Export dataset
- Contigs
- Sequences - XMFA / concatenated FASTA formats

Analysis

- Codon usage
- Presence/absence status of loci
- Genome comparator
- BLAST
- PhyloViz

Miscellaneous

- Description of database fields

Genetic markers for reconstructing evolutionary history

Output

Analysis against defined loci

All loci

Allele numbers are used where these have been defined, otherwise sequences will be marked as 'New#1', 'New#2' etc. Missing alleles are marked as 'X'. Incomplete alleles (located at end of contig) are marked as 'I'.

Locus	4469 (ERR1672933)	4470 (ERR1672935)	4645 (SBVN)	4646 (3896VN)
BACT000001 (rpsA)	3211	102	new#1	102
BACT000002 (rpsB)	816	111	112	112
BACT000003 (rpsC)	754	66	69	69
BACT000004 (rpsD)	76	76	77	77
BACT000005 (rpsE)	56	55	56	56
BACT000006 (rpsF)	22	63	22	22
BACT000007 (rpsG)	1951	59	59	59
BACT000008 (rpsH)	49	49	49	49
BACT000009 (rpsI)	4093	676	67	67
BACT000010 (rpsJ)	61	62	63	63
BACT000011 (rpsK)	50	50	50	50
BACT000012 (rpsL)	61	61	61	61
BACT000013 (rpsM)	47	47	47	47
BACT000014 (rpsN)	45:217	45:217	45:217	45:217
BACT000015 (rpsO)	46	46	47	47
BACT000016 (rpsP)	43	43	43	43
BACT000017 (rpsQ)	47	47	47	47
BACT000018 (rpsR)	42	42	42	42
BACT000019 (rpsS)	37	37	37	37
BACT000020 (rpsT)	59	59	59	59
BACT000021 (rpsU)	34	34	34	34
BACT000030 (rplA)	85	84	85	85
BACT000031 (rplB)	846	74	75	75
BACT000032 (rplC)	75	4463	75	75
BACT000033 (rplD)	57	56	57	57
BACT000034 (rplE)	61	61	61	61
BACT000035 (rplF)	2066	68	69	69
BACT000036 (rplL)	new#1	58	58	58
BACT000038 (rplI)	141	141	142	142
BACT000039 (rplJ)	810	59	59	59
BACT000040 (rplK)	73	74	4327	4327
BACT000042 (rplM)	792	792	53	53
BACT000043 (rplN)	50	50	50	50
BACT000044 (rplO)	59	786	59	59
BACT000045 (rplP)	1747	49	49	49
BACT000046 (rplQ)	42	42	42	42
BACT000047 (rplR)	54	54	54	54
BACT000048 (rplS)	814	814	90	90
BACT000049 (rplT)	68	69	68	68
BACT000050 (rplU)	new#1	54	54	54
BACT000051 (rplV)	759	54	50	50

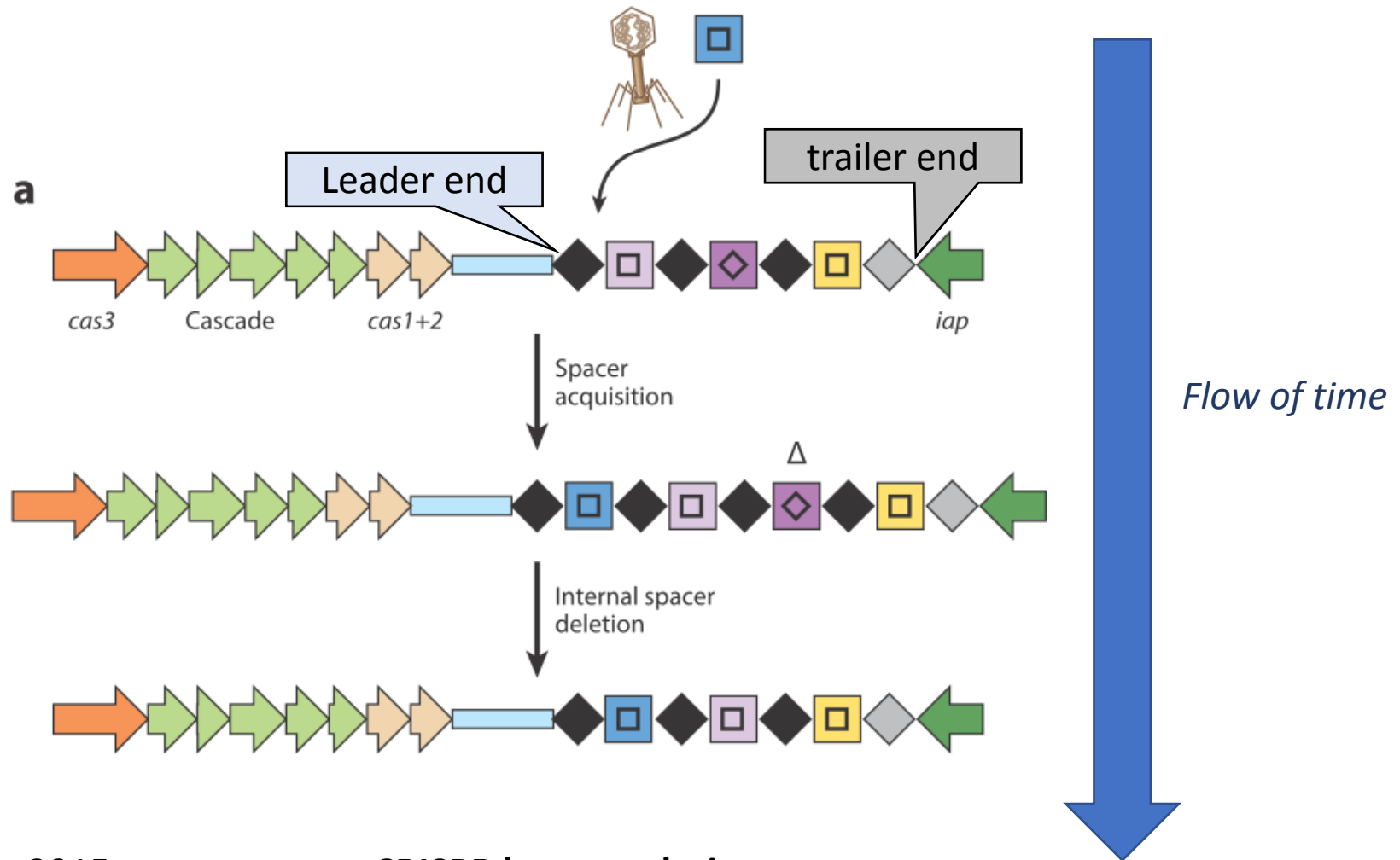
BIGSdb: Genome
Comparator output

Genetic markers for reconstructing evolutionary history

High-resolution identification and typing using **CRISPR array**

- CRISPR array acquires novel *spacer* in an ordinal manner → infer the chronological record of bacterium
- Spacer can be identified using CRISPRFinder tool

Genetic markers for reconstructing evolutionary history



Lier et al., 2015

CRISPR locus evolution