Genomic Approaches to Molecular Epidemiology and Typing

King Jordan

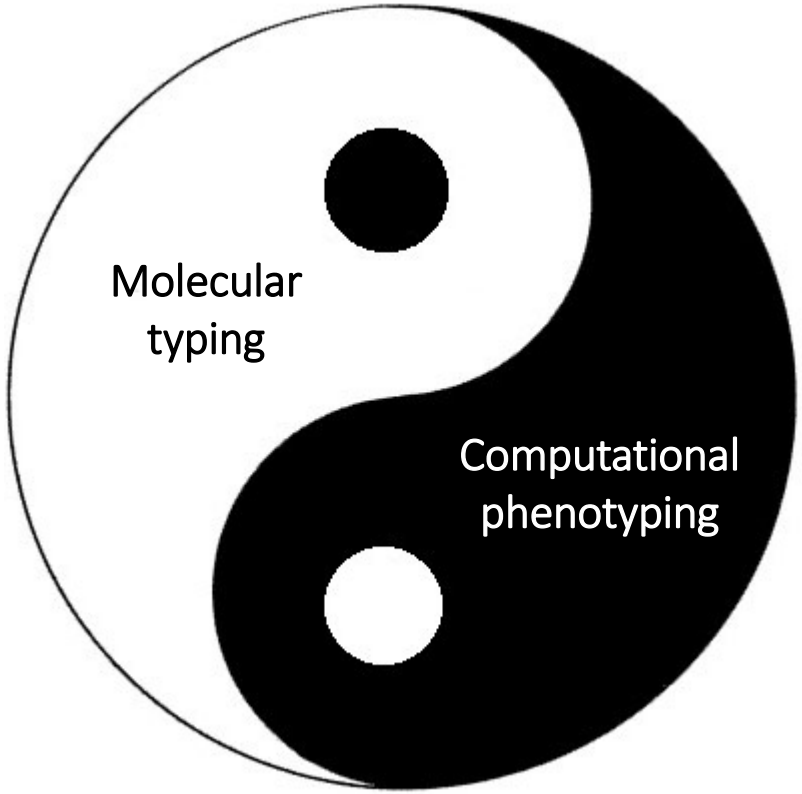
Bioinformatics Graduate Program

Applied Bioinformatics Laboratory (ABiL)

Georgia Institute of Technology

Ying & Yang of Computational Genomics

- *Molecular typing* – DNA sequences as passive (markers) of evolutionary lineages
  - Its about evolution!
- *Computational phenotyping* – determination of the genomic basis of phenotype (e.g. virulence, antimicrobial resistance)
  - Its about function!



Learning Objectives

1. Consider your work in this class in the larger context of the course history
2. Understand the fundamentals of molecular epidemiology and typing
3. Understand pre-NGS molecular typing
4. Understand implications of NGS revolution for molecular typing
5. Familiarity with specific NGS-based methods for molecular typing
6. Sense of what the future may hold for molecular epidemiology and typing

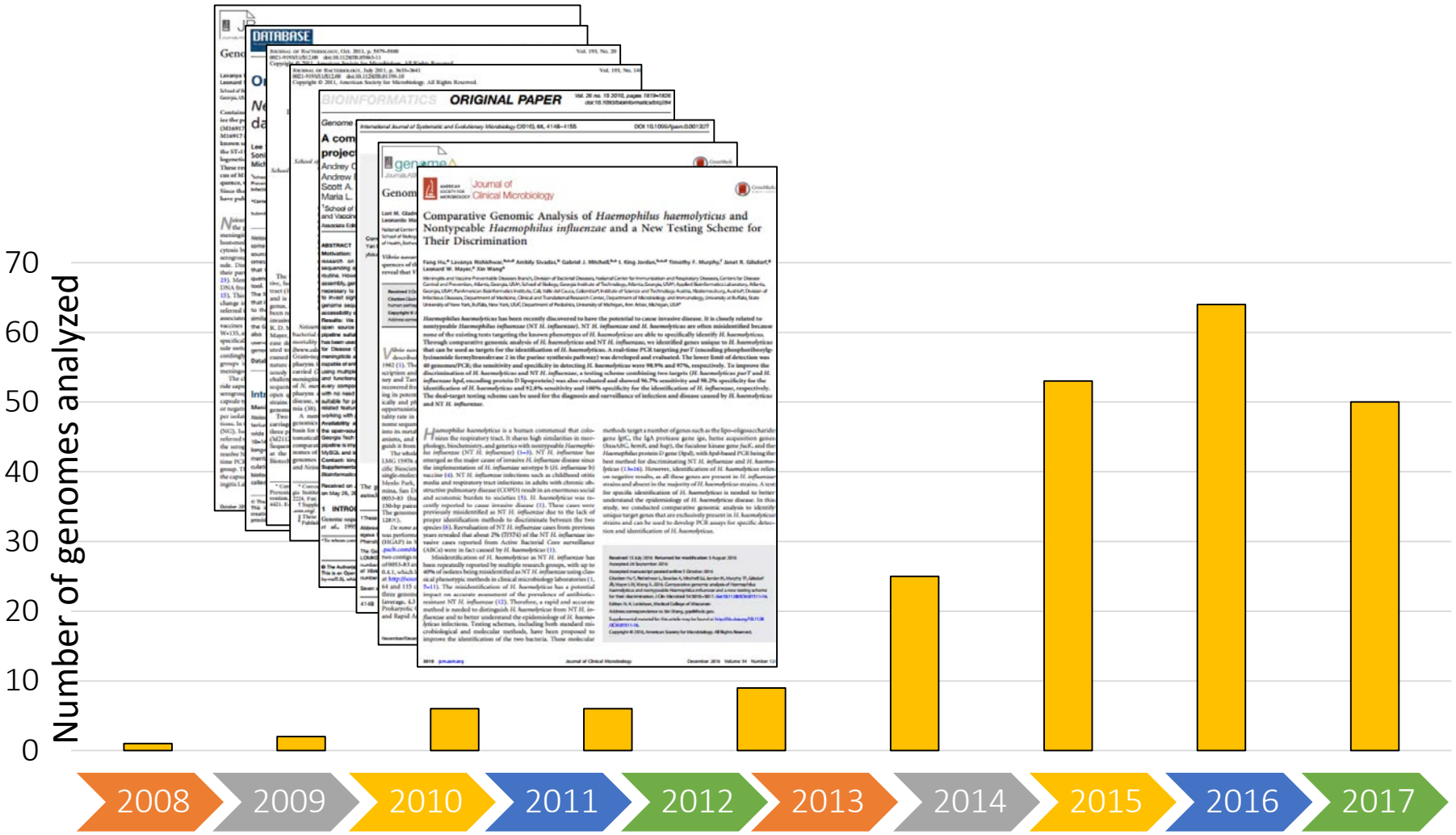
Outline

- Computational genomics class: goals and accomplishments
- Molecular epidemiology & typing in the NGS era
- Bacterial sequence typing
- Implications of NGS for molecular epidemiology & typing
- NGS-based typing methods: stringMLST, STing, others
- Future vision

Outline

- Computational genomics class: goals and accomplishments
- Molecular epidemiology & typing in the NGS era
- Bacterial sequence typing
- Implications of NGS for molecular epidemiology & typing
- NGS-based typing methods: stringMLST, STing, others
- Future vision

Computational genomics education @ Georgia Tech



Georgia Tech Bioinformatics Interdisciplinary Graduate Program

ABiL Applied BioInformatics Laboratory

Past Course Accomplishments

Year	Organism(s)	#Genome(s)	Platform	Goal(s)
2009	*Neisseria meningitidis*	1	454	• Fully assembled and annotated genome • Develop genome analysis pipeline and protocol
2010	*Neisseria meningitidis*	2	454	• Fully assembled and annotated genomes • Develop genome analysis pipeline and protocol • Develop multiple-strain database
2018	*Klebsiell spp.* (could be scores of different species)	**252**	Illumina	• Distinguish between susceptible and heteroresistant strains/species • Discover genomic determinants of antibiotic resistance • Development of a predictive webserver: data in -> knowledge out
	Haemophilus haemolyticus, Haemophilus influenzae			genotype • Identify recombination within Hi *cap* locus • Discover the mechanism of NTHi
2016	*Haemophilus haemolyticus, Haemophilus influenzae*	64	Illumina	• Develop a typing scheme for various strains of non-typeable *Haemophilus influenzae* (NTHi) • Determine the evolutionary relationship between typeable and non-typeable Hi.
2017	*Salmonella enterica*	50	Illumina	• Devised an approach to distinguish sporadic from outbreak strains • Develop automated methods to discover genomic determinants of virulence

Student Presentation at CDC

Partners in Public Health

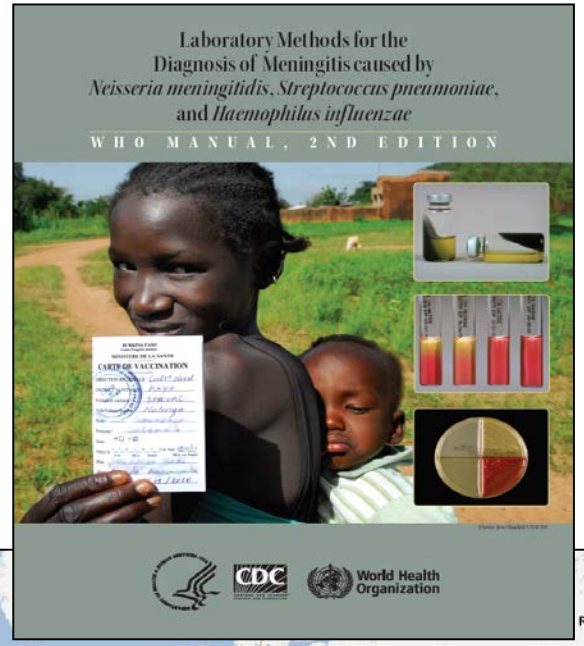
Molecular epidemiology and typing: *Neisseria*

W606–W611 Nucleic Acids Research, 2009, Vol. 37, Web Server issue *Published online 25 May 2009*
doi:10.1093/nar/gkp288

Meningococcus genome informatics platform: a system for analyzing multilocus sequence typing data

Lee S. Katz[1,*], Chris R. Bolen[1], Brian H. Harcourt[2], Susanna Schmink[2], Xin Wang[2], Andrey Kislyuk[1], Robert T. Taylor[1], Leonard W. Mayer[2,*] and I. King Jordan[1]

[1]School of Biology, Georgia Institute of Technology, Atlanta, GA 30332 and [2]Meningitis and Vaccine Preventable Diseases Branch, Centers for Disease Control and Prevention, Atlanta, GA 30333, USA



Applications of MGIP for molecular epidemiology

Contents lists available at ScienceDirect
Vaccine
journal homepage: www.elsevier.com/locate/vaccine

Prevalence and genetic diversity of candidate vaccine antigens among invasive *Neisseria meningitidis* isolates in the United States

Xin Wang[a,*], Amanda Cohn[a], Maurizio Comanducci[b], Lubomira Andrew[c], Xin Zhao[a], Jessica R. MacNeil[a], Susanna Schmink[a], Alessandro Muzzi[b], Stefania Bambini[b], Rino Rappuoli[b], Mariagrazia Pizza[b], Ellen Murphy[c], Susan K. Hoiseth[c], Kathrin U. Jansen[c], Annaliesa S. Anderson[c], Lee H. Harrison[d], Thomas A. Clark[a], Nancy E. Messonnier[a], Leonard W. Mayer[a]

[a] *Meningitis and Vaccine Preventable Disease Branch, Division of Bacterial Diseases, National Center of Immunization and Respiratory Diseases, Centers for Disease Control and Prevention, Atlanta, GA 30333, United States*
[b] *Novartis Vaccines, Siena, Italy*
[c] *Pfizer Vaccine Research, Pearl River, NY, United States*
[d] *Infectious Diseases Epidemiology Research Unit, University of Pittsburgh School of Medicine and Graduate School of Public Health, Pittsburgh, PA, United States*

OPEN ACCESS Freely available online
PLoS one

sodC-Based Real-Time PCR for Detection of *Neisseria meningitidis*

Jennifer Dolan Thomas[1]*, Cynthia P. Hatcher[1], Dara A. Satterfield[1,7], M. Jordan Theodore[1], Michelle C. Bach[1,7], Kristin B. Linscott[1,7], Xin Zhao[1], Xin Wang[1], Raydel Mair[1], Susanna Schmink[1], Kathryn E. Arnold[2,4], David S. Stephens[3,4,5], Lee H. Harrison[8], Rosemary A. Hollick[8], Ana Lucia Andrade[9], Juliana Lamaro-Cardoso[9], Ana Paula S. de Lemos[10], Jenna Gritzfeld[11], Stephen Gordon[11], Ahmet Soysal[12], Mustafa Bakir[12], Dolly Sharma[3,6], Shabnam Jain[3,6], Sarah W. Satola[3,4,5], Nancy E. Messonnier[1], Leonard W. Mayer[1]

1 Meningitis and Vaccine Preventable Diseases Branch, Division of Bacterial Diseases, National Center for Immunization and Respiratory Diseases, Centers for Disease Control and Prevention, Atlanta, Georgia, United States of America, **2** Division of Public Health, Georgia Department of Community Health, Atlanta, Georgia, United States of America, **3** Emory University School of Medicine, Atlanta, Georgia, United States of America, **4** Georgia Emerging Infections Program, Atlanta, Georgia, United States of America, **5** Veterans Affairs Medical Center, Atlanta, Georgia, United States of America, **6** Children's Healthcare of Atlanta, Atlanta, Georgia, United States of America, **7** Biology Department, Agnes Scott College, Decatur, Georgia, United States of America, **8** Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland, United States of America, **9** Instituto de Patologia Tropical e Saúde Pública, Universidade Federal de Goiás, Goiânia, Goiás, Brazil, **10** Instituto Adolfo Lutz, São Paulo, Brazil, **11** Respiratory Infection, Clinical Group, Liverpool School of Tropical Medicine, Liverpool, United Kingdom, **12** Division of Pediatric Infectious Diseases, Marmara University School of Medicine, Istanbul, Turkey

Contents lists available at ScienceDirect
Vaccine
journal homepage: www.elsevier.com/locate/vaccine

Diversity of factor H-binding protein in *Neisseria meningitidis* carriage isolates

Jane W. Marsh[a,*], Kathleen A. Shutt[a], Rolando Pajon[b], Mary M. Tulenko[a], Stephen Liu[a], Rosemary A. Hollick[c], Julia A. Kiehlbauch[d], Thomas A. Clark[e], David S. Stephens[f,g], Kathryn E. Arnold[h], Robert A. Myers[d], Leonard W. Mayer[e], Lee H. Harrison[a,c]

[a] *Infectious Diseases Epidemiology Research Unit, University of Pittsburgh School of Medicine and Graduate School of Public Health, Pittsburgh, PA, United States*
[b] *Center for Immunobiology and Vaccine Development, Children's Hospital Oakland Research Institute, Oakland, CA, United States*
[c] *Department of International Health, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, United States*
[d] *Laboratories Administration, Maryland Department of Health and Mental Hygiene, Baltimore, MD, United States*
[e] *Meningitis and Vaccine Preventable Diseases Branch, Division of Bacterial Diseases, National Center for Immunization and Respiratory Diseases, Centers for Disease Control and Prevention, Atlanta, GA, United States*
[f] *Emory University, Robert W. Woodruff Health Sciences Center, Atlanta, GA, United States*
[g] *Medical Research Service, VA Medical Center, Atlanta, GA, United States*
[h] *Georgia Division of Public Health and Emerging Infections Program, Atlanta, GA, United States*

OPEN ACCESS Freely available online
PLOS ONE

Molecular Characterization of Invasive Meningococcal Isolates from Countries in the African Meningitis Belt before Introduction of a Serogroup A Conjugate Vaccine

Dominique A. Caugant[1,2]*, Paul A. Kristiansen[1], Xin Wang[3], Leonard W. Mayer[3], Muhamed-Kheir Taha[4], Rasmata Ouédraogo[5], Denis Kandolo[6], Flabou Bougoudogo[7], Samba Sow[8], Laurence Bonte[9]

1 WHO Collaborating Centre for Reference and Research on Meningococci, Norwegian Institute of Public Health, Oslo, Norway, **2** Faculty of Medicine, University of Oslo, Oslo, Norway, **3** WHO Collaborating Center for Prevention and Control of Epidemic Meningitis, Centers for Disease Control and Prevention, Atlanta, Georgia, United States of America, **4** WHO Collaborating Centre for Reference and Research on Meningococci, Institut Pasteur, Paris, France, **5** Laboratoire de Référence Méningite, Centre Hospitalier Universitaire Pédiatrique Charles de Gaulle, Ouagadougou, Burkina Faso, **6** WHO Inter country Support Team for West Africa, Ouagadougou, Burkina Faso, **7** Institut National de Recherche en Santé Publique (INRSP), Bamako, Mali, **8** Centre pour les Vaccins en Développement (CVD), Bamako, Mali, **9** Support Logistique, Médecins Sans Frontières, Paris, France

Bacterial Strain Typing

Duncan MacCannell, PhD

KEYWORDS
• Bacterial typing techniques • Molecular epidemiology • Multilocus sequence typing
• DNA sequence analysis • Pulsed-field gel electrophoresis • Genomics

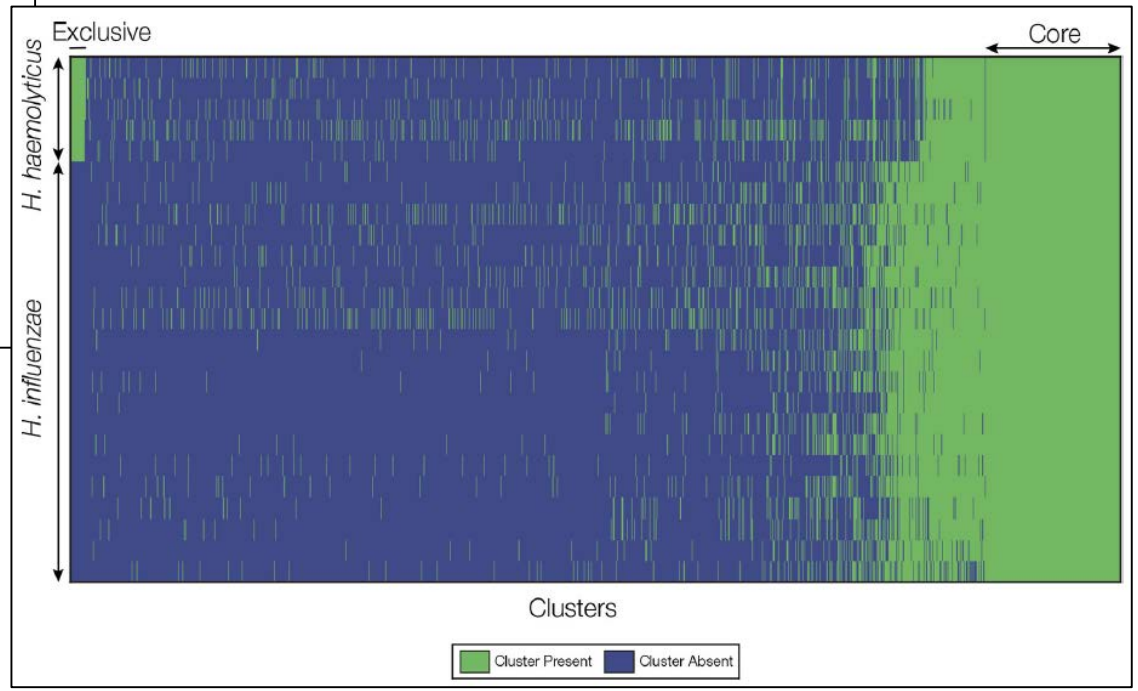
Molecular epidemiology and typing: *Haemophilus*

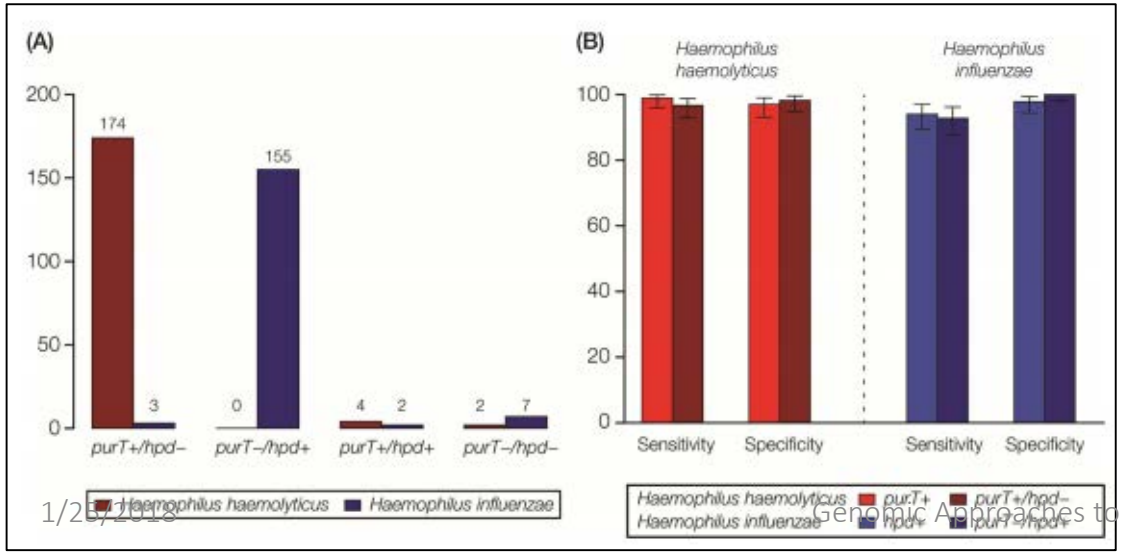
AMERICAN SOCIETY FOR MICROBIOLOGY | Journal of Clinical Microbiology

Comparative Genomic Analysis of *Haemophilus haemolyticus* and Nontypeable *Haemophilus influenzae* and a New Testing Scheme for Their Discrimination

Fang Hu,[a] Lavanya Rishishwar,[b,c,d] Ambily Sivadas,[b] Gabriel J. Mitchell,[b,e] I. King Jordan,[b,c,d] Timothy F. Murphy,[f] Janet R. Gilsdorf,[g] Leonard W. Mayer,[a] Xin Wang[a]

Meningitis and Vaccine Preventable Diseases Branch, Division of Bacterial Diseases, National Center for Immunization and Respiratory Diseases, Centers for Disease Control and Prevention, Atlanta, Georgia, USA[a]; School of Biology, Georgia Institute of Technology, Atlanta, Georgia, USA[b]; Applied Bioinformatics Laboratory, Atlanta, Georgia, USA[c]; PanAmerican Bioinformatics Institute, Cali, Valle del Cauca, Colombia[d]; Institute of Science and Technology Austria, Klosterneuburg, Austria[e]; Division of Infectious Diseases, Department of Medicine, Clinical and Translational Research Center, Department of Microbiology and Immunology, University at Buffalo, State University of New York, Buffalo, New York, USA[f]; Department of Pediatrics, University of Michigan, Ann Arbor, Michigan, USA[g]





Outline

- Computational genomics class: goals and accomplishments
- Molecular epidemiology & typing in the NGS era
- Bacterial sequence typing
- Implications of NGS for molecular epidemiology & typing
- NGS-based typing methods: stringMLST, STing, others
- Future vision

Molecular Epidemiology

- **Molecular** – "use of molecular biology techniques"

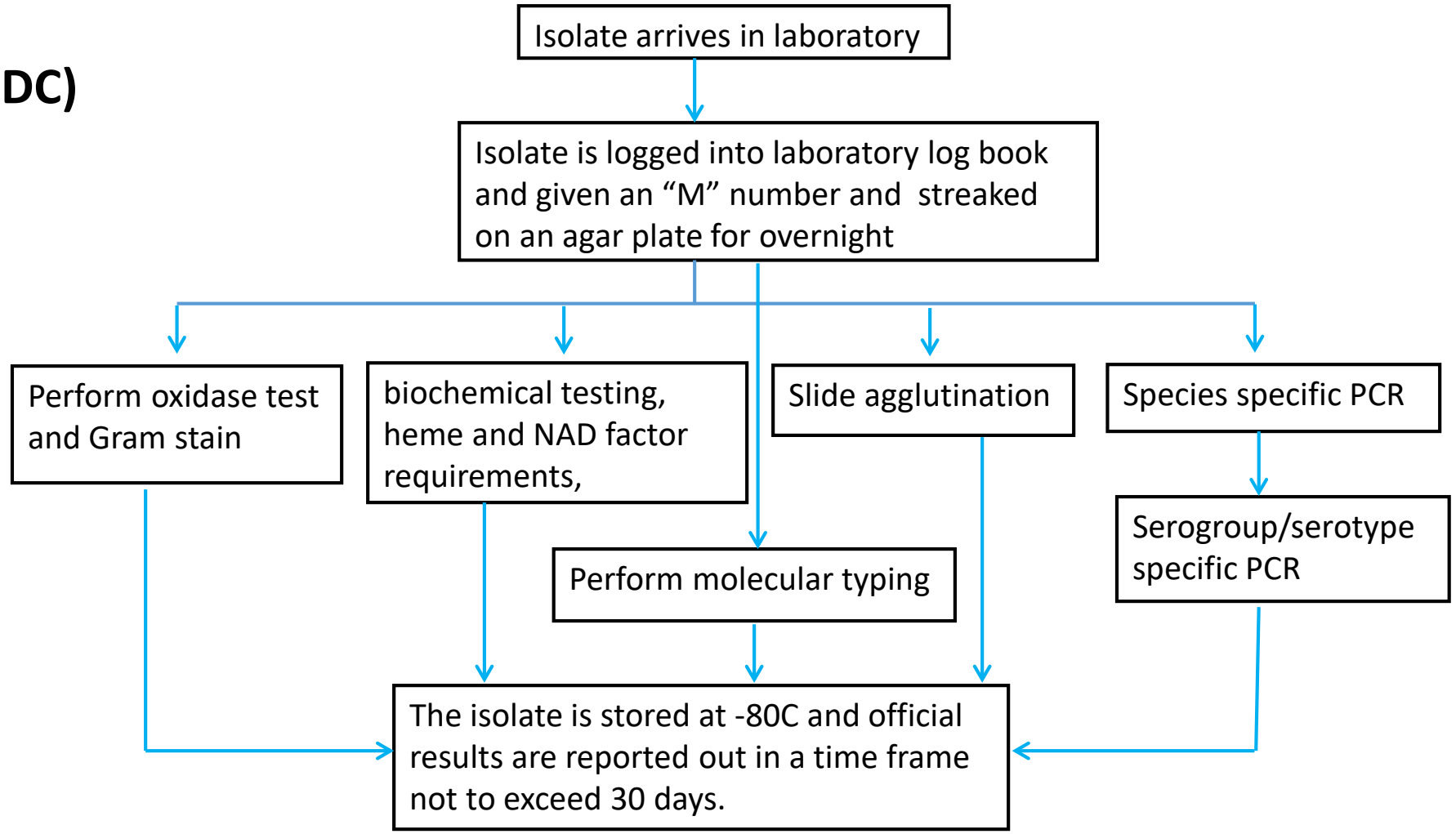
- **Epidemiology** – "the study of the distribution and determinants of disease occurrence in human populations"

Application of Molecular Techniques to Epidemiology

Application	Method	Technique
Identification	Conventional	Culture
		Enzyme-linked immunosorbent assay (ELISA), Enzyme immunosorbent assay (EIA)
		Monoclonal antibodies
	Nucleic acid based	DNA hybridization for known genes, Direct sequencing of one or more regions
		Multilocus sequence typing (MLST)
	PCR* based	Amplification of a single target specific to a pathogen, Ligase chain reaction (LCR)
	Protein based	Western blot or immunoblotting
Typing	Conventional	Serotype
		Antibiotic susceptibilities
	Nucleic acid based	Plasmid profiles
		Restriction fragment length polymorphism (RFLP)
		Pulsed field gel electrophoresis (PFGE)
		Segmented RNA gel electrophoresis, Ribosomal RNA gel electrophoresis
		Direct sequencing of one or more regions
		Multilocus sequence typing (MLST)
	PCR based	Amplification of a single target specific to a pathogen
		Targeting known repetitive sequences (enterobacterial repetitive intergenic consensus sequences (ERIC), repetitive extragenic palindromic sequences (REP), double repetitive element (DRE), BOX, insertional sequence (IS), polymorphic guanine/cytosine-rich repetitive sequences (PGRS))
		Random primers (randomly amplified polymorphic DNA (RAPD), arbitrary primed PCR (AP-PCR))
		Restriction endonuclease of a single amplified product
		Amplified fragment length polymorphism (AFLP)
	Protein based	Multilocus enzyme electrophoresis (MLEE)
	Gene expression	Reverse transcriptase PCR
		Microarray technologies

Traditional Testing Flow

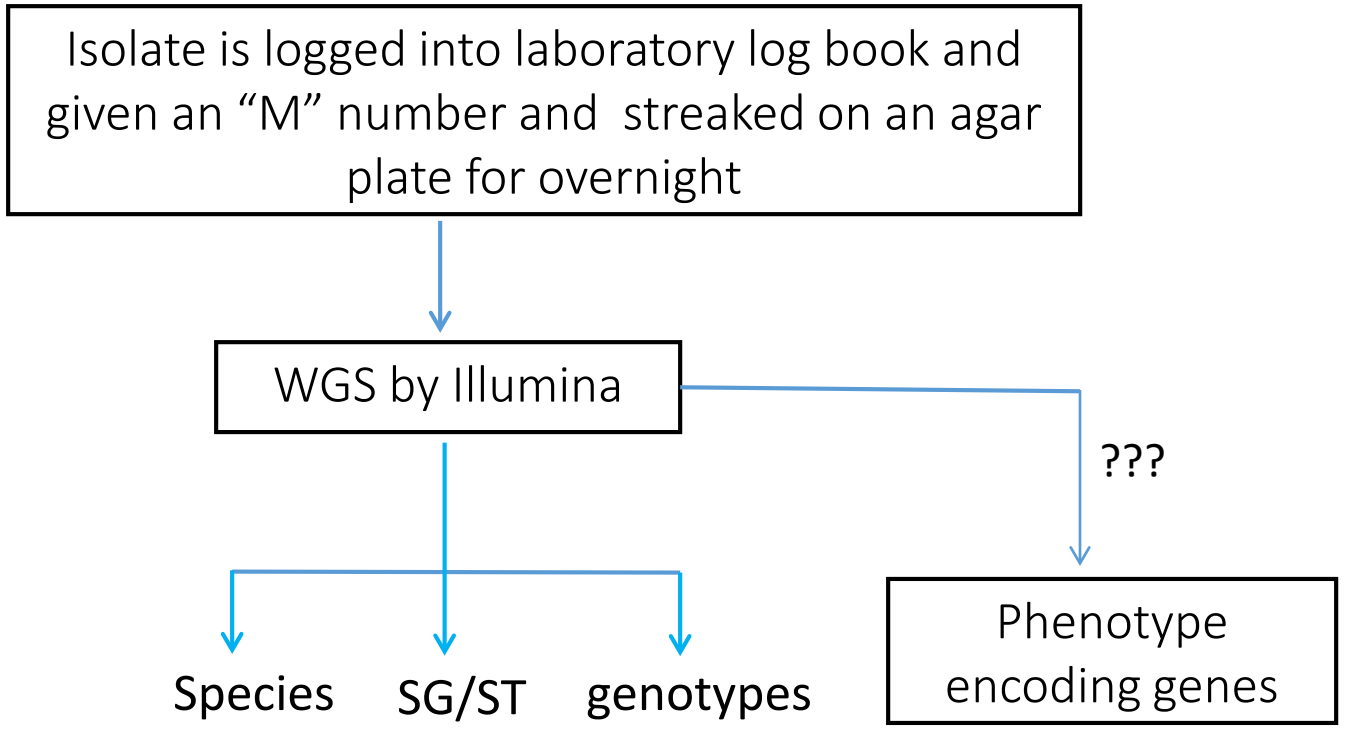
From
Xin Wang (CDC)



One laboratorian 1-2 weeks

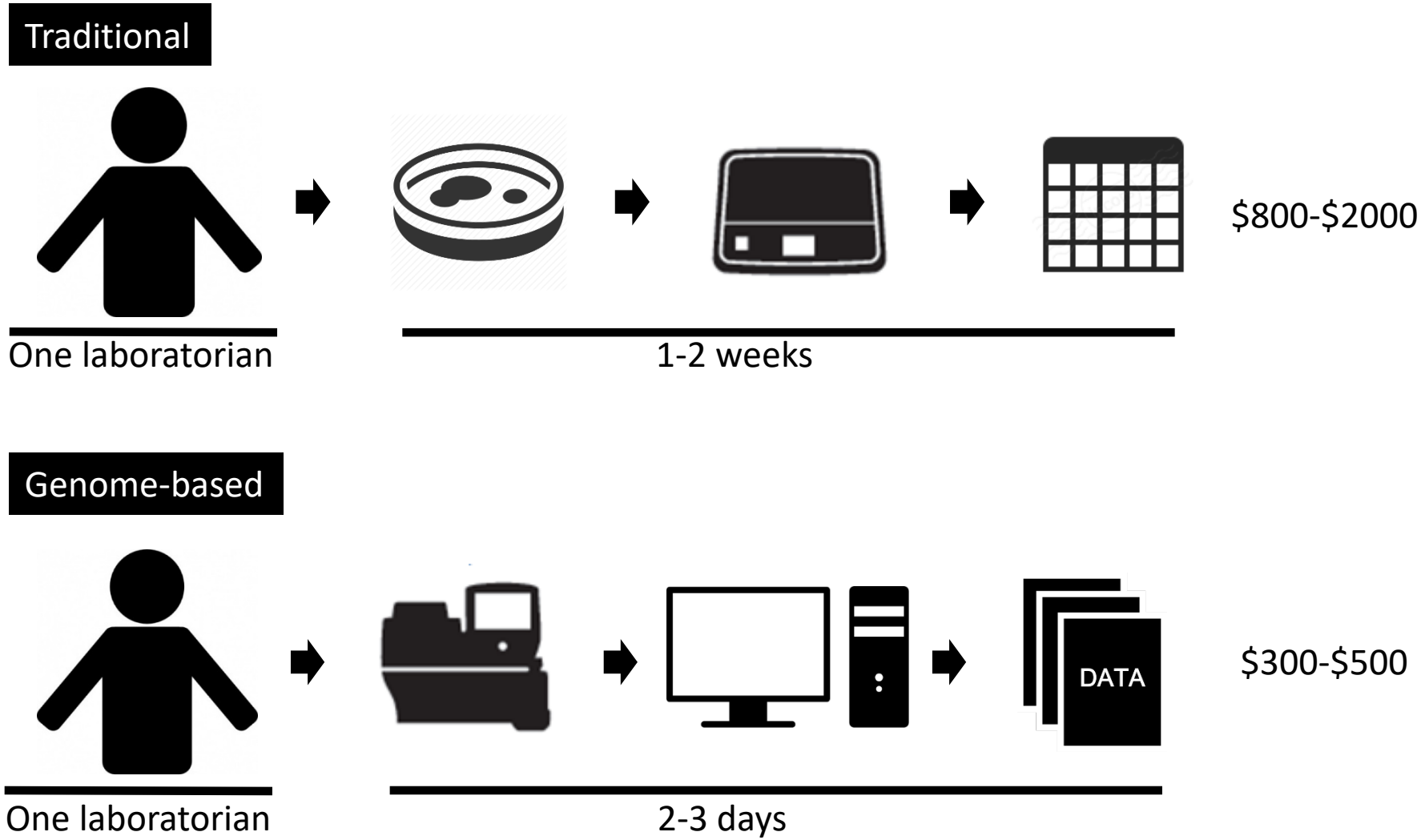
WGS based workflow

From
Xin Wang (CDC)



One laboratorian; 2-3 days

Comparison of Cost



*Estimation based on 8 PCR tests for 20 individual genes and median income for CDC laboratory technologist of $40,213 .

Outline

- Computational genomics class: goals and accomplishments
- Molecular epidemiology & typing in the NGS era
- Bacterial sequence typing
- Implications of NGS for molecular epidemiology & typing
- NGS-based typing methods: stringMLST, STing, others
- Future vision

Sequence Typing

- Sequence typing - identifying different types of organisms within a species
- Critical for epidemiological surveillance and outbreak control
- Human pathogens of one species can comprise very diverse set of organisms
- Critically important that the typing technique have enough discriminatory power needed to distinguish all epidemiologically unrelated isolates.

Desirable properties of any bacterial typing system

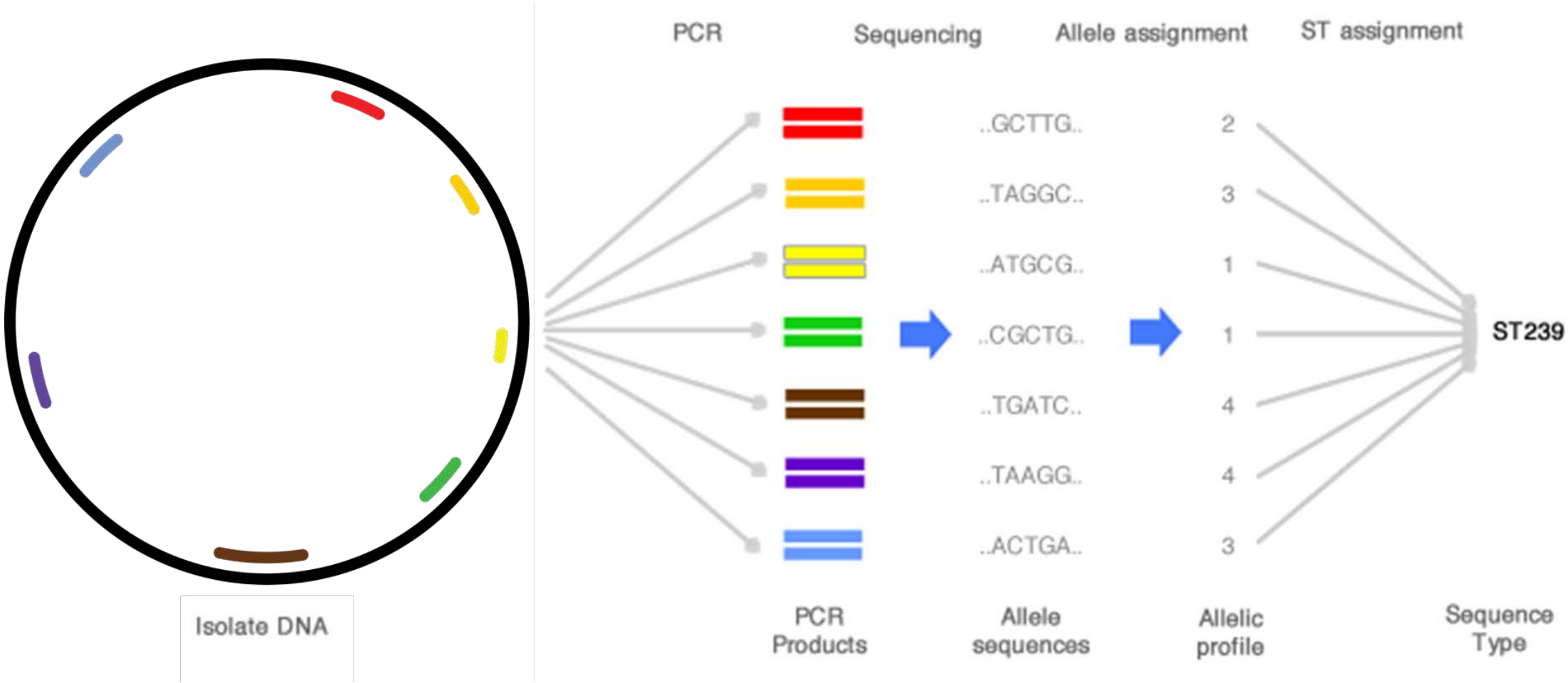
- **Universal** - applicable to all bacteria
- **Natural** - reflecting genealogical relationships while retaining the capacity to describe closely related organisms with distinct properties
- **Understandable**
- **Expandable** – account for incomplete knowledge and flexible to changes in the knowledge
- **Portable**
- **Technology independent**
- **Readily available** to the entire community
- **Scalable**
- Able to **accommodate** a wide range of variation
- **Broadly accepted** by those who use them and open to contributions
- **Backwards compatible**, where possible (locus-based=granular)

Maiden et al. *Nat Rev Microbiol*. 2013. 11(10):728-36

Multilocus Sequence Typing (MLST)

- MLST is a conventional gene-based approach for sequence typing
- Entails sequencing of 7 housekeeping genes from around the genome to accurately type the bacteria
- Popular and widely used in the pre-NGS era due to its good trade-off between sequencing and resolution
- Limited resolution for NGS era
- Older method – but has a lot legacy sequence type information and is still being used in many public health labs

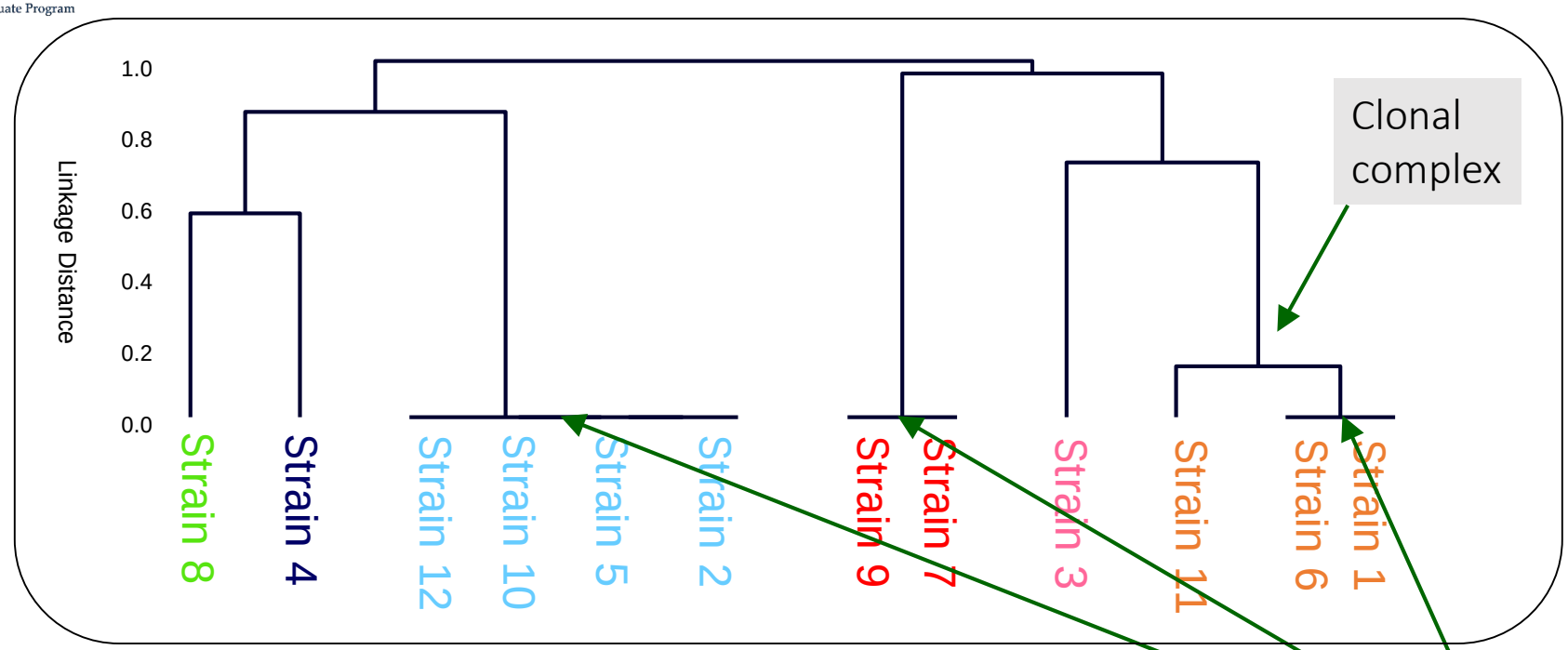
Multilocus Sequence Typing (MLST)



http://beta.mlst.net/Instructions/default.html

Allele Profiles

- Alleles with good power of discrimination are used for devising the MLST technique
- Good power of discrimination: Similar enough to group same strains together, distant enough to distinguish different strains



Strain 1	1	5	3	7	2	8	4
Strain 2	11	15	30	12	22	11	20
Strain 3	6	5	9	1	2	13	17
Strain 4	9	10	1	19	12	18	14
Strain 5	11	15	30	12	22	11	20
Strain 6	1	5	3	7	2	8	4
Strain 7	20	11	29	9	21	13	11
Strain 8	9	15	13	27	22	18	14
Strain 9	20	11	29	9	21	13	11
Strain 10	11	15	30	12	22	11	20
Strain 11	1	5	3	5	2	8	4
Strain 12	11	15	30	12	22	11	20

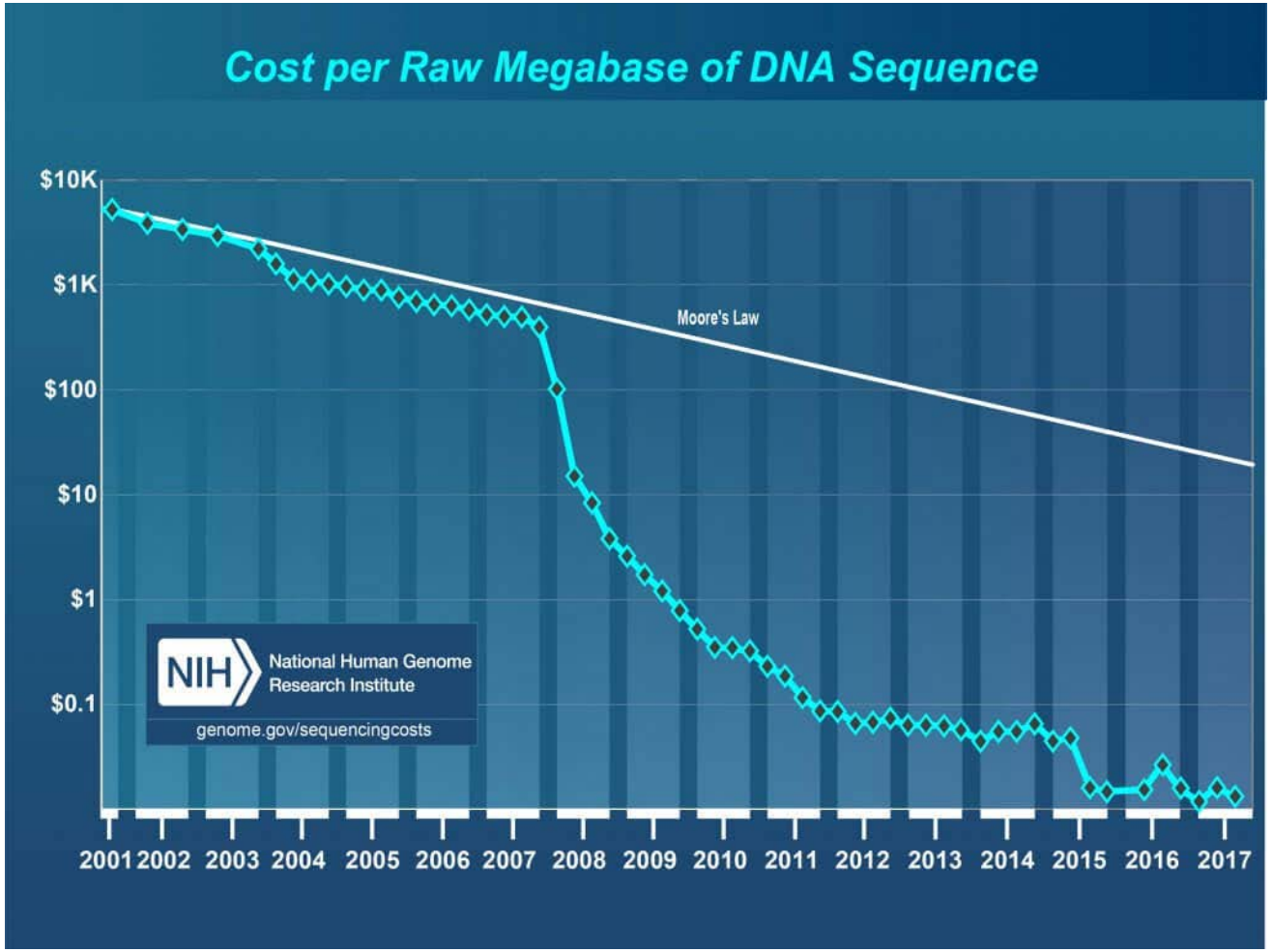
Clone

Sequence Type

Outline

- Computational genomics class: goals and accomplishments
- Molecular epidemiology & typing in the NGS era
- Bacterial sequence typing
- Implications of NGS for molecular epidemiology & typing
- NGS-based typing methods: stringMLST, STing, others
- Future vision

Cost Reduction of Genome Sequencing



Computational Genomics Pipeline

BIOINFORMATICS ORIGINAL PAPER
Vol. 26 no. 15 2010, pages 1819–1826
doi:10.1093/bioinformatics/btq284

Genome analysis
Advance Access publication June 2, 2010

A computational genomics pipeline for prokaryotic sequencing projects

Andrey O. Kislyuk[1], Lee S. Katz[1], Sonia Agrawal[1], Matthew S. Hagen[1], Andrew B. Conley[1], Pushkala Jayaraman[1], Viswateja Nelakuditi[1], Jay C. Humphrey[1], Scott A. Sammons[2], Dhwani Govil[2], Raydel D. Mair[3], Kathleen M. Tatti[3], Maria L. Tondella[3], Brian H. Harcourt[3], Leonard W. Mayer[3] and I. King Jordan[1,*]

[1]School of Biology, Georgia Institute of Technology, Atlanta, GA 30332, [2]Core Biotechnology Facility and [3]Meningitis and Vaccine Preventable Diseases Branch, Centers for Disease Control and Prevention, Atlanta, GA 30333, USA

Associate Editor: Alex Bateman

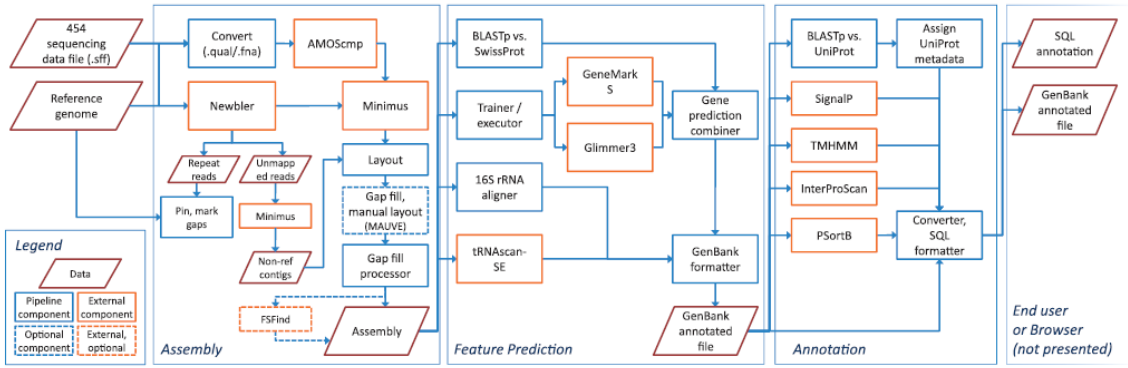
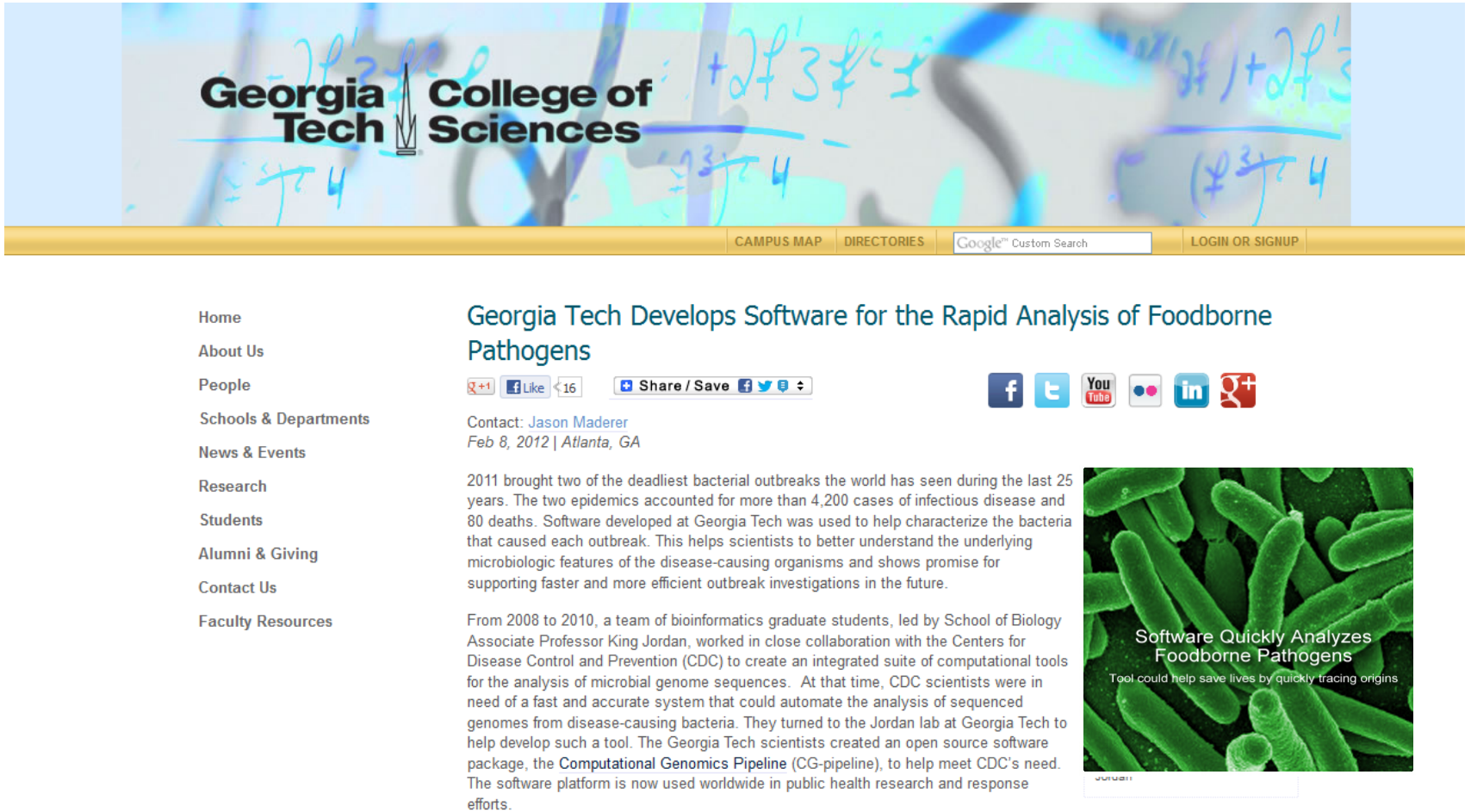


Fig. 1. Chart of data flow, major components and subsystems in the pipeline. Three subsystems are presented: genome assembly, feature prediction and functional annotation. Each subsystem consists of a top-level execution script managing the input, output, format conversion and combination of results for a number of components. A hierarchy of scripts and external programs then performs the tasks required to complete each stage. The legend for the flowchart indicates the identities of the distinct pipeline components: data, pipeline component, optional component, external component and external, optional component.

Computational Genomics Pipeline



Computational Genomics Pipeline

abcNEWS HOME | VIDEO | U.S. | WORLD | POLITICS | ENTERTAINMENT | TECH | HE

NOW 9-11 ANNIVERSARY • OSCAR PISTORIUS VERDICT • TED CRUZ BOOED • ROGER G

776 Like 336 share 129 Tweet 0 +1

0 Comments

E. Coli Outbreak in Europe Is One of the Deadliest in History

June 4, 2011

By OLIVIA KATRANDJIAN

ABCNEWS.com NEXT VIDEO » Massive E. Coli Outbreak Hits Europe

The rapidly developing **European E. coli outbreak** that has killed 19 people and sickened thousands, including four suspected cases in the United States, has become one of the deadliest outbreaks of E. coli in modern history.

Where exactly people are being infected with the disease is still unknown, although 17 people fell ill after eating in the northern German city of Luebeck in May, according to the local media. Researchers from Germany's national disease control center are inspecting the restaurant in question.

Other health experts suspect the disease first spread last month at a festival in the northern German city of Hamburg that was visited by 1.5 million people. But as of yet, there is no concrete proof that either site is the cause of the outbreak.

The New York Times Business Day

WORLD | U.S. | N.Y. / REGION | BUSINESS | TECHNOLOGY | SCIENCE | HEALTH | SPORTS | OPINION

Search | Global | DealBook | Markets | Economy | Energy | Media | Person

Listeria Outbreak Traced to Cantaloupe Packing Shed

Ed Andrieski/Associated Press

The Food and Drug Administration recalled 300,000 cases of melons from Jensen Farms in Colorado following a listeria outbreak.

By WILLIAM NEUMAN

Published: October 19, 2011

A nationwide listeria outbreak that has killed 25 people who ate tainted cantaloupe was probably caused by unsanitary conditions in the packing shed of the Colorado farm where the melons were grown, federal officials said Wednesday.

TWITTER LINKEDIN PRINT REPRINTS

Computational Genomics Pipeline

- Rapid, automated annotation on big machines
- Tight integration of multiple layers & components
- Still too slow and expensive for molecular epidemiology

Genomic Methods Must Scale

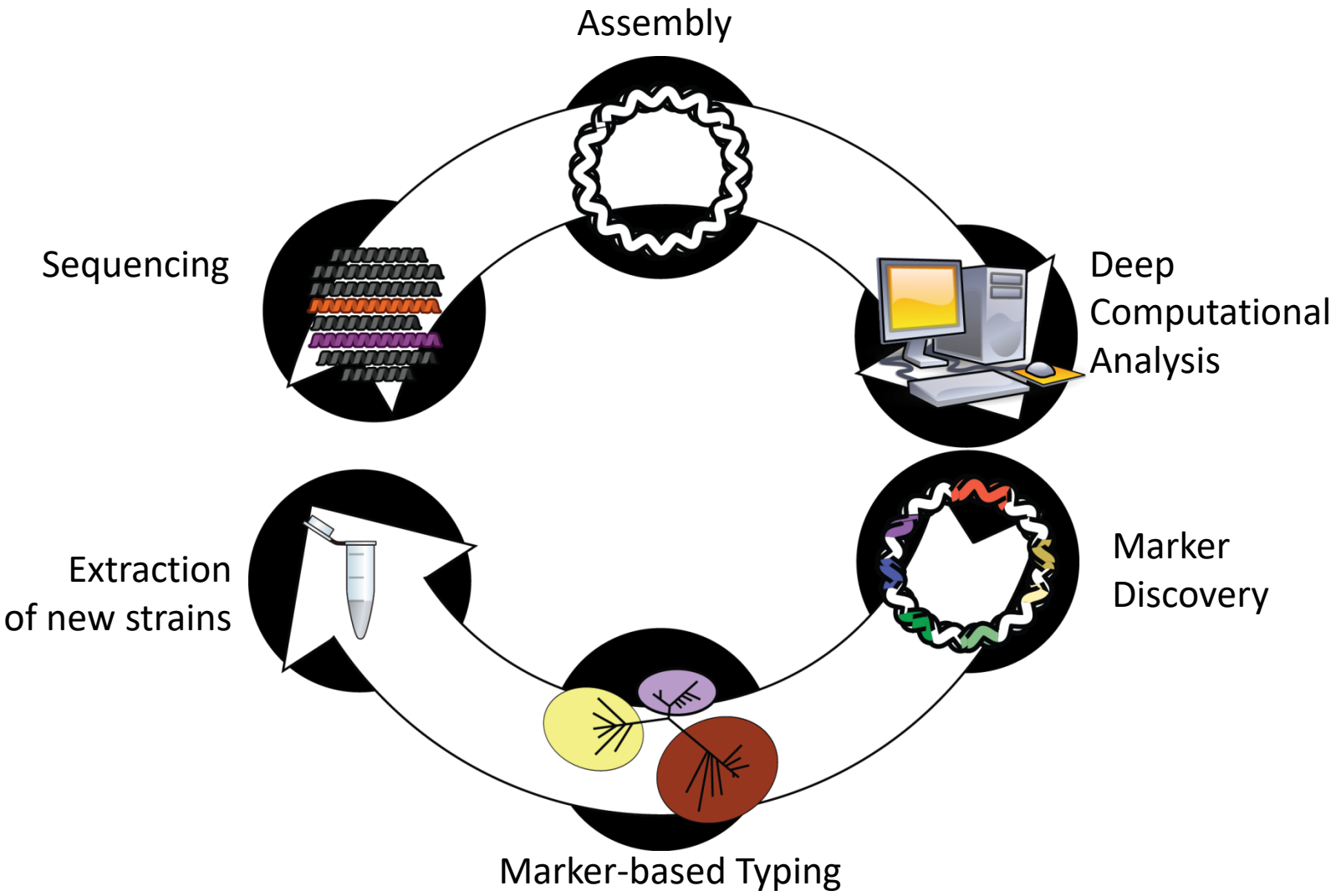
- Genomic approaches must scale to tens, hundreds, thousands of samples entailed by outbreak & epidemiological studies

- Must understand difference between research grade bioinformatics/genomics and what is needed for molecular typing & epidemiology

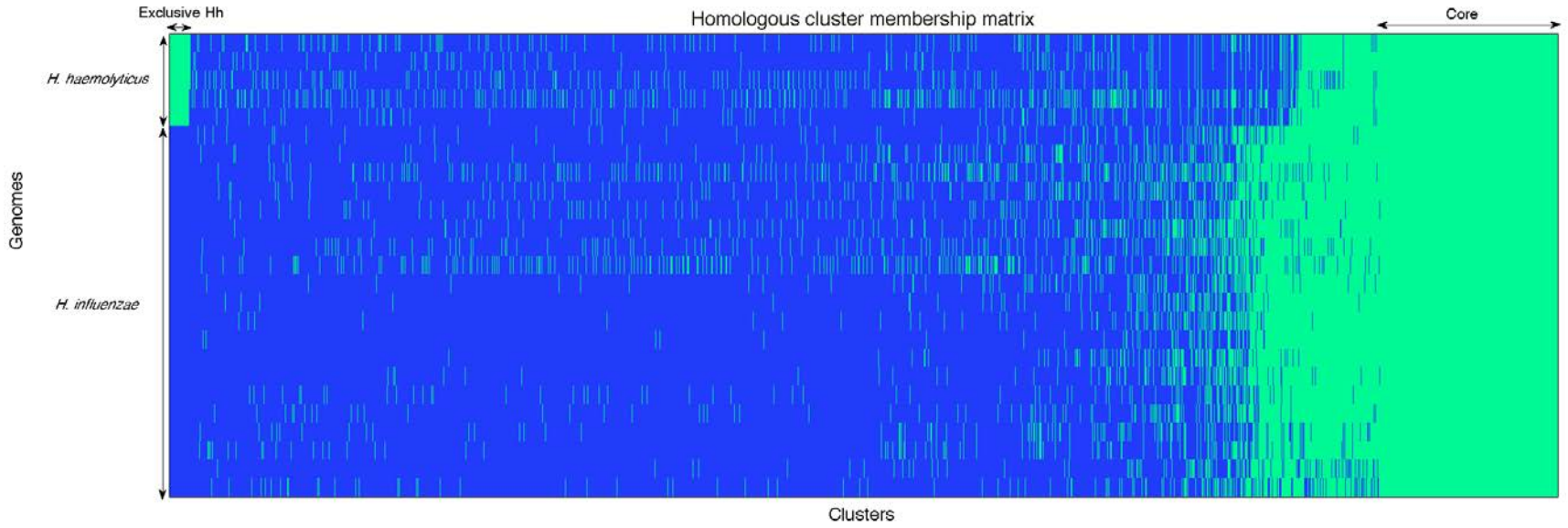
Genomic Research Methods Do Not Scale

- Methods based on genome assembly, gene prediction & functional annotation pipelines do not scale (e.g. CG-pipeline)
- Methods based on ortholog detection do not scale (e.g. ANI)
- Remain powerful and highly useful research tools
- But should not be shoehorned into molecular typing & epidemiology

Genomics-to-Marker Discovery Cycle

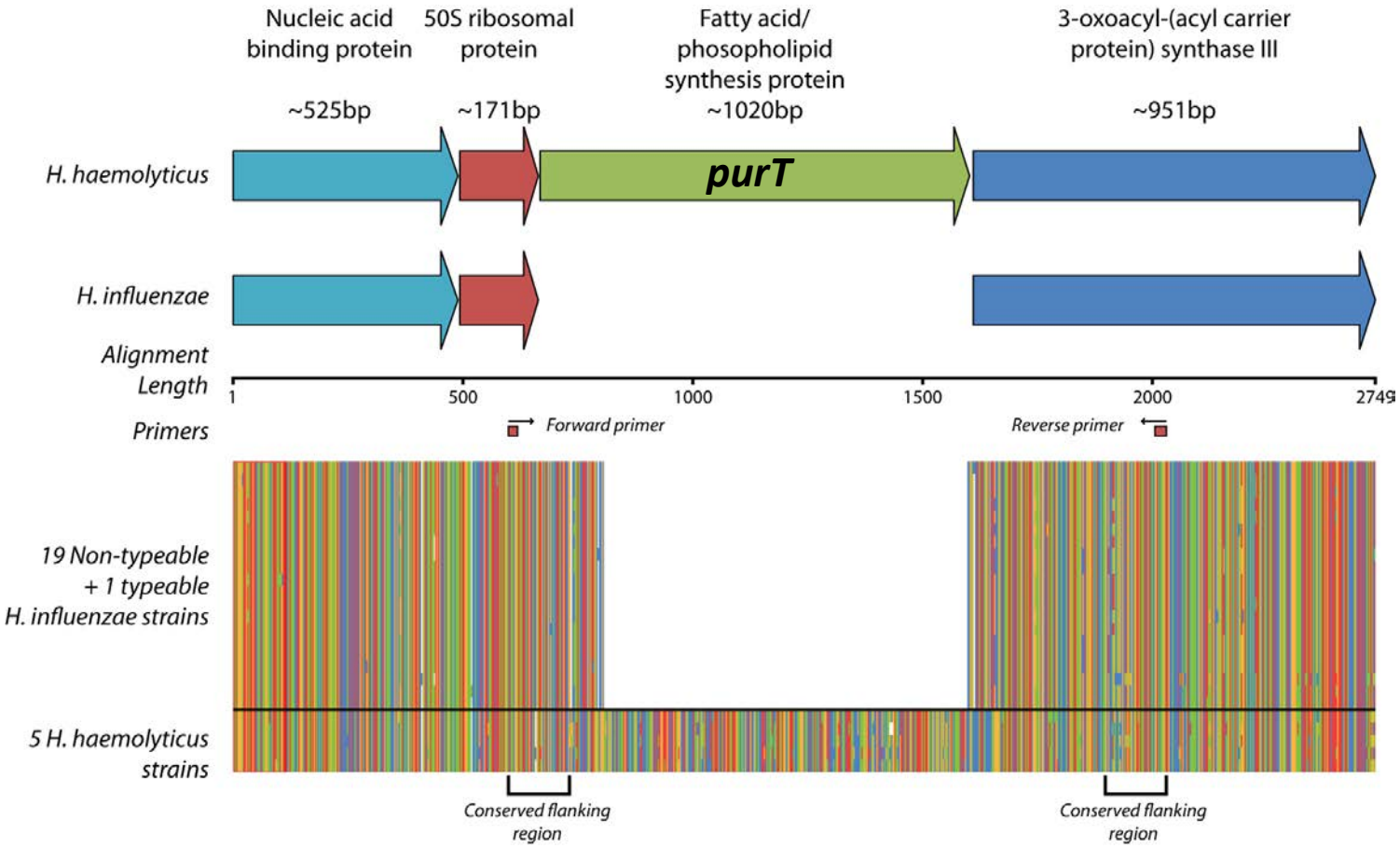


Genome based marker discovery with *H. haemolyticus*



All-against-all genome comparison among 26 Haemophilus genomes and identification of core genes exclusive to *H. haemolyticus*

Present / absent gene flanked by conserved regions



PCR based validation of marker discriminatory power

	No. of isolates/No. of total (%)			
purT	+	-	+	-
hpd	-	+	+	-
H. haemolyticus	174/180 (96.7%)	0/180 (0.00%)	4/180 (0.02%)	2/180 (0.01%)
H. influenzae	3/167 (0.02%)	155/ 167 (92.8%)	2/167 (0.01%)	7/167 (0.04%)

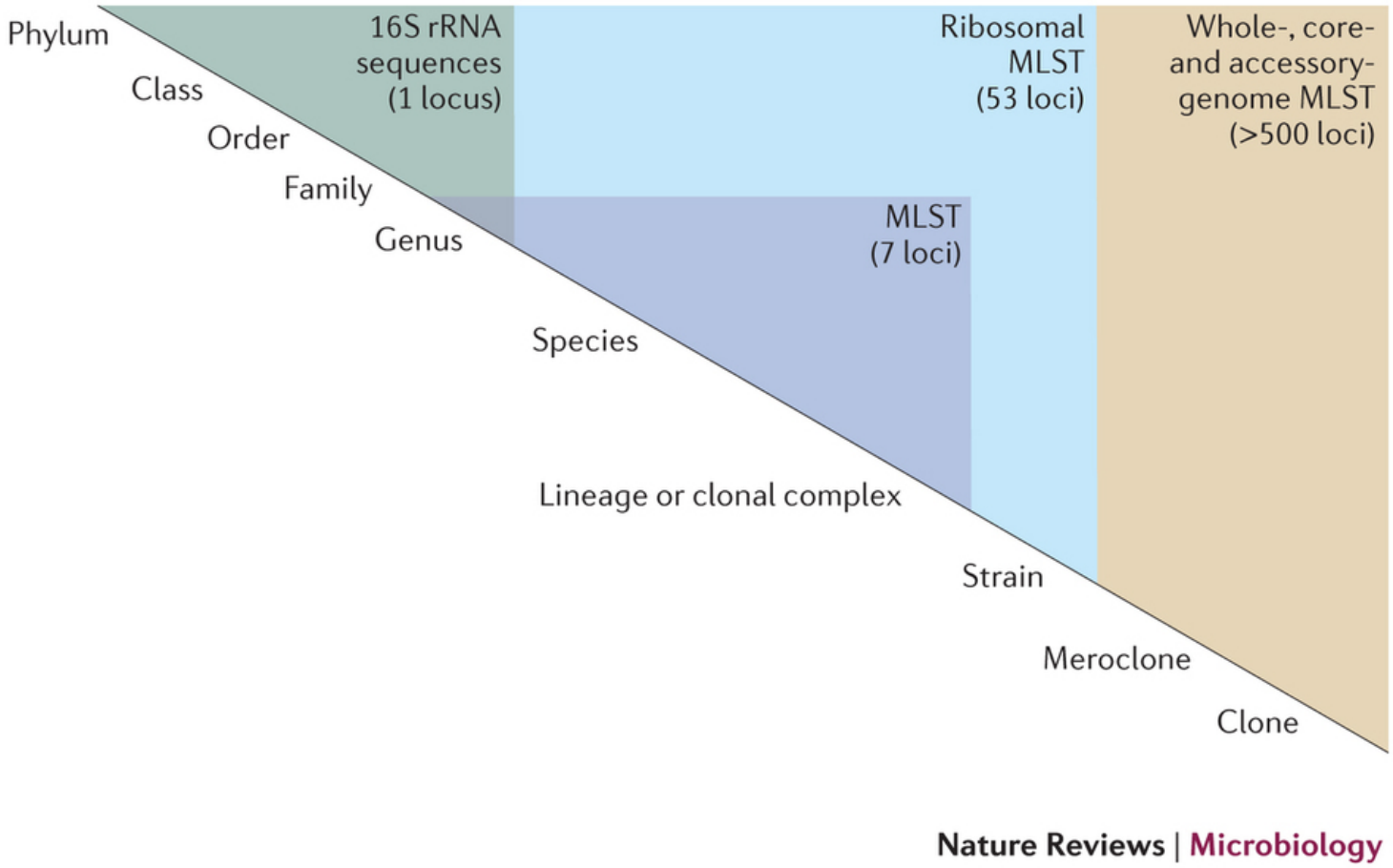
Genomic Methods Must Find Their Level

- Tools and approaches need to be tuned to:
  - The organism being studied
  - The level of relatedness being assessed

- Binary MLST scheme designed for *N. meningitidis* as it evolves faster via recombination than via point mutation

- SNP typing will outperform K-mer based methods for highly clonal organisms (*B. anthracis*) & for distinguishing close levels of relatedness

Typing Schemes at Different Levels

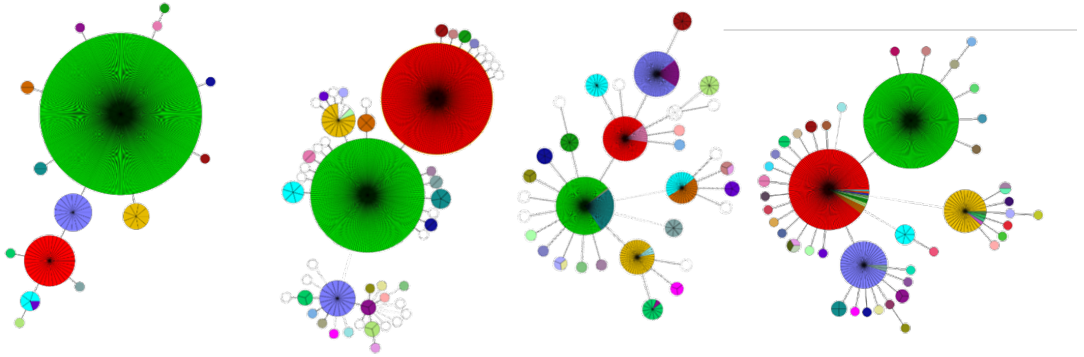


Maiden et al. 2013. *Nat Rev Microbiol*. 11:728-36

Other MLST-like Typing Schemes

- MLST is based on the 7 house keeping loci
- They may not always provide enough discriminatory power – e.g., if they are very conserved
- Other methods in existence include:
  - Ribosomal MLST (rMLST) – based on 53 loci
  - Antigen MLST (aMLST) – based on few antigen loci (3-4)

Comparative results of different MLST schemes



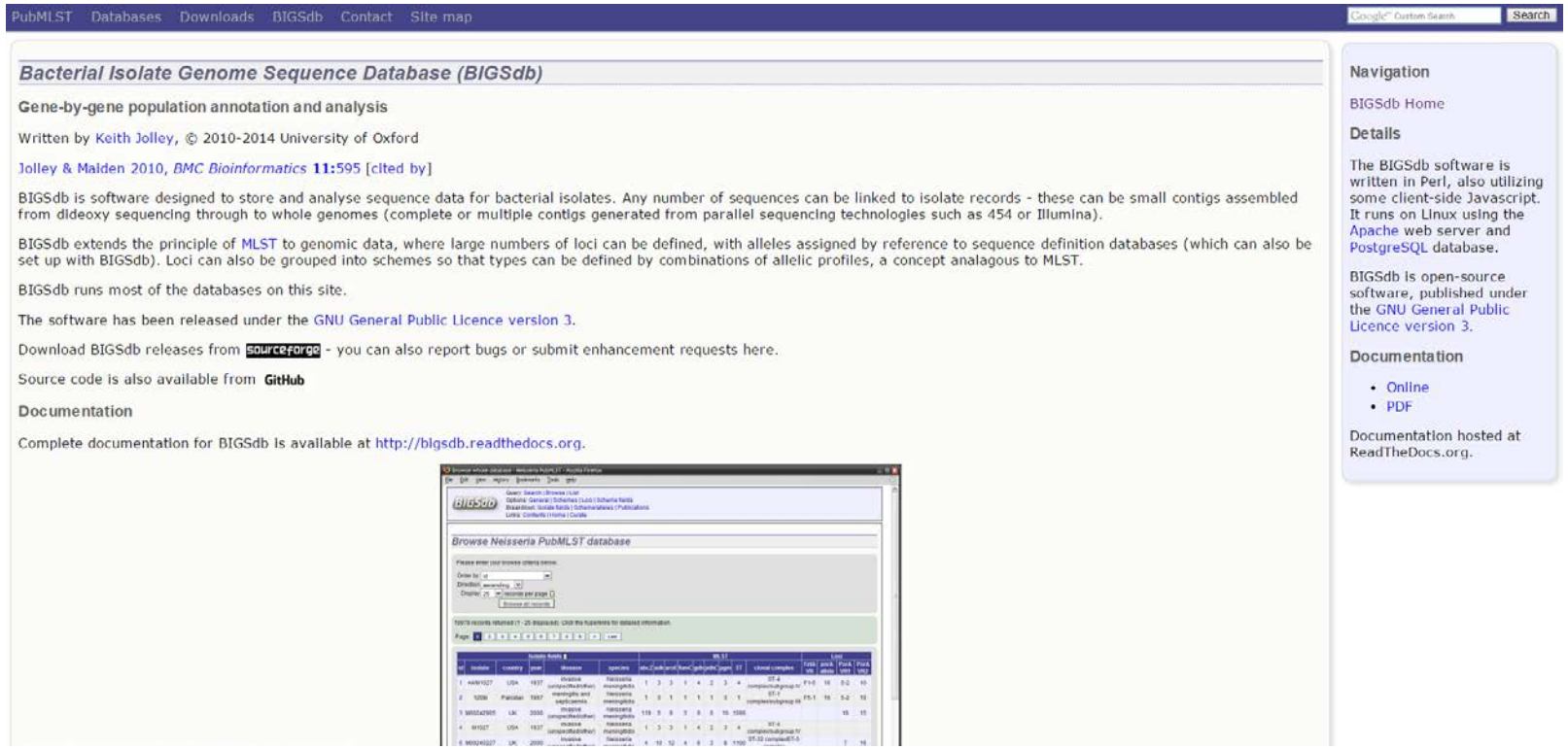
	hMLST	rMLST	aMLST	gtMLST
# of Groups	15	50	32	**45**
# of STs	16	52	41	**60**
# in the largest group	357	173	88	177
# of identical loci	2 (28.6%)	24 (45.3%)	0 (00.0%)	0 (00.0%)
# of variable loci	2 (28.6%)	13 (24.5%)	3 (75.0%)	4 (57.1%)
# of truncated loci	3 (42.9%)	16 (30.2%)	1 (25.0%)	3 (42.9%)
# of loci used	**7**	**53**	**4**	**7**
Resolution	Low	Medium	High	Highest

Whole genome analysis and MLST

- Whole genome sequence analysis and comparison can help with the selection of loci to be used in MLST schemes in order to yield maximum discriminatory power
- This approach can also be used to tune the loci selection to the particular question that is being asked – i.e. to customize the level of resolution
- Many groups have successfully developed MLST-like typing schemes using similar approaches

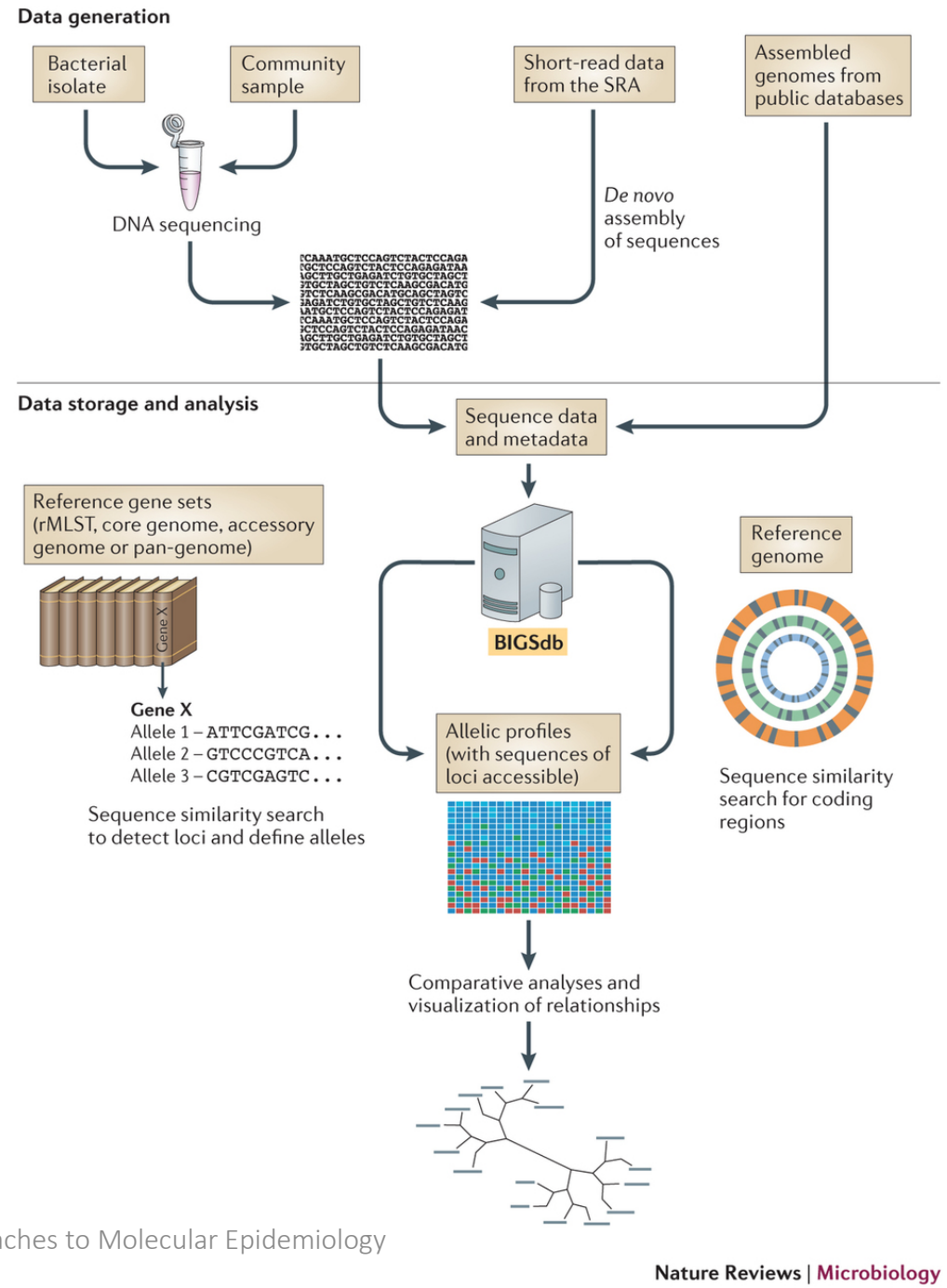
Building Custom Typing Scheme

- BIGSdb is a great resource that can help you develop typing schemes considerably fast



Building Custom Typing Scheme

Maiden et al. 2013. *Nat Rev Microbiol*. 11:728-36



Utility of whole genome typing approaches

Approach	Gene-by-gene based	SNP based	WG alignment based (ANI)
Linking to historical & epidemiological data	Yes	No; Varies with reference genome	Yes
Approach standardization	Yes	No; # of SNP sites can change	Yes
Portable	Yes	No; Varies with reference genome	Yes
Scalable	Yes	Yes	Limited
Between species comparison	Limited	Limited	Yes
Existing resources	Yes	No	No
Expertise required	Low	High	Medium

Shared **advantages** of the three approaches:

- Universal
- Natural
- Understandable
- High resolution

Shared **disadvantages** of the three approaches:

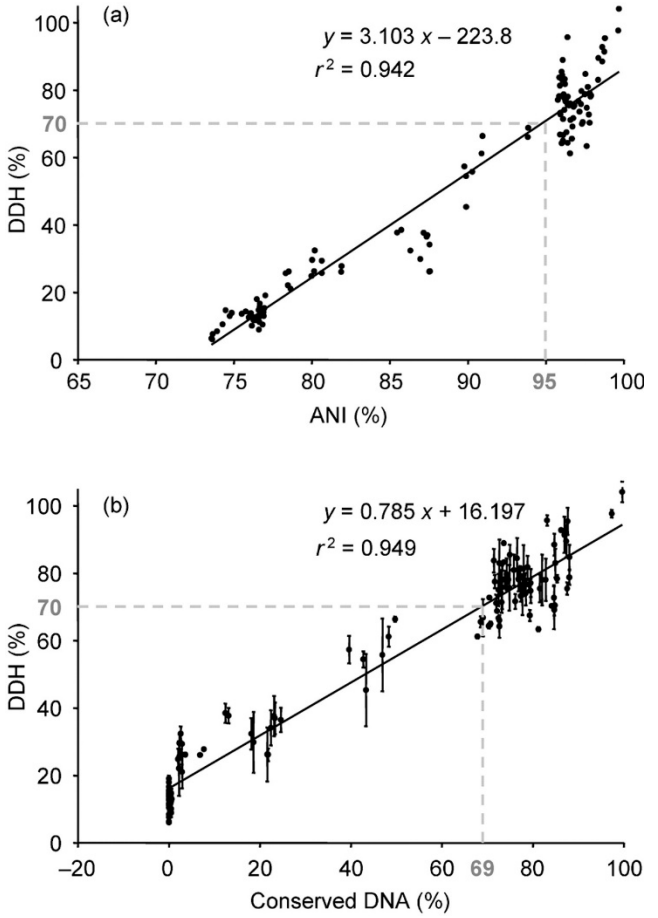
- Requires some form of manual curation
- Existing methods require substantial computational processing

Outline

- Computational genomics class: goals and accomplishments
- Molecular epidemiology & typing in the NGS era
- Bacterial sequence typing
- Implications of NGS for molecular epidemiology & typing
- NGS-based typing methods: stringMLST, STing, others
- Future vision

Average Nucleotide Identity

- ANI was first introduced in 2005 by Konstantinidis & Tiedje as the ANI between gene pairs
- Later in 2007 Goris et al. provided a more practical approach for computing ANI
- They demonstrated correlation between DDH and ANI by comparing 28 strains with 124 DDH values
- They showed that the 70% DDH cut off for species delineation corresponds to 95% ANI cut off
- ANI has since been employed in numerous research publications to compare genomes of different prokaryotes

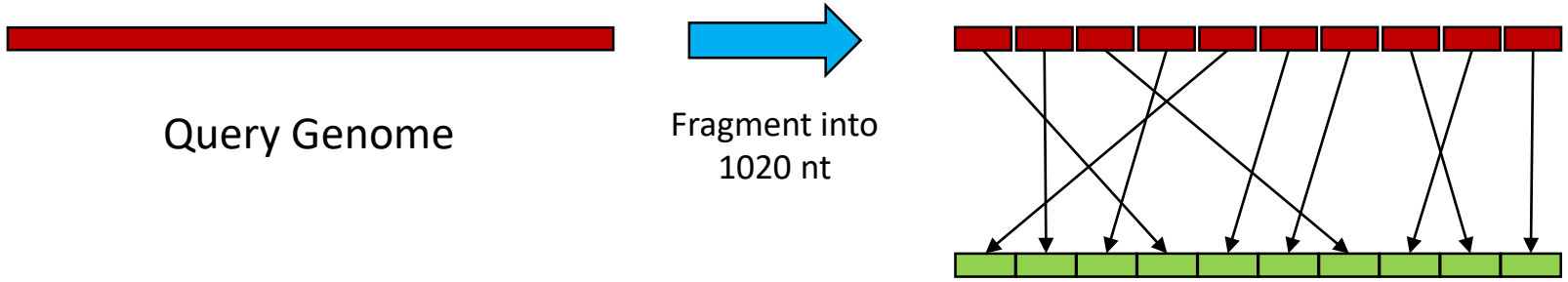


Goris et al. 2007. *Int J Syst Evol Microbiol.* 57:81-91

Average Nucleotide Identity – Goris et al.

- Using the genome sequence, there are two methods by which the ANI can be computed:
  - ANIb
  - ANIm
- ANIb was the original method proposed by Goris et al.
- The method tries to mimic DDH as much as possible
- It fragments the query genome first into 1020 nt long sequences and then computes the similarity against the subject genome using BLAST, hence the name ANIb
- Fragments showing less than 30% sequence identity over 70% of their length were discarded

Average Nucleotide Identity – Goris et al.



Query	Subject	Identity	Coverage	
Region_i1	Region_j1	100%	100%	✓
Region_i2	Region_j2	80%	90%	✓
Region_i3	Region_j3	20%	60%	✗

Average

Average Nucleotide Identity (ANI) value

Solving the Issue of Scale

- Two approaches to resolving the issue of scale

- Novel genome analysis methods
  - Methods based on reference assemblies & SNP calling can scale
  - K-mer (word) based methods can scale

- Cyclical genome analysis approach
  - Research grade genome analysis to identify novel markers
  - More standard (higher throughput) marker based methods

In silico DNA aptamers

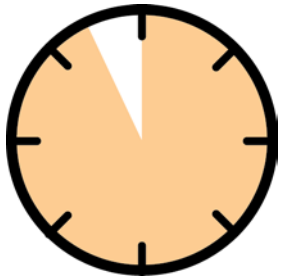
- Also known as … K-mers, substrings, l-tuples, n-mers, n-grams etc.
- K-mers are sequence substrings of length k
- e.g., DNA 1-mers (monomers) are A, C, G, T; 2-mers (dimers) are AA, AC, AG, AT, CA, CC, CG, CT, GA, GC, GG, GT, TA, TC, TG, TT etc …
- These are used in numerous tools and application areas in bioinformatics
- Well known k-mer applications – BLAST, genome assemblers, genome mappers, kraken (metagenomics)

Methods for Identifying an Isolate's MLST

Traditional Methods

- Traditional methods for performing MLST entails performing Sanger sequencing on each loci followed by allele number designation

- With 2nd generation sequencing technologies, it is cheaper to sequence the whole genome over sequencing each loci



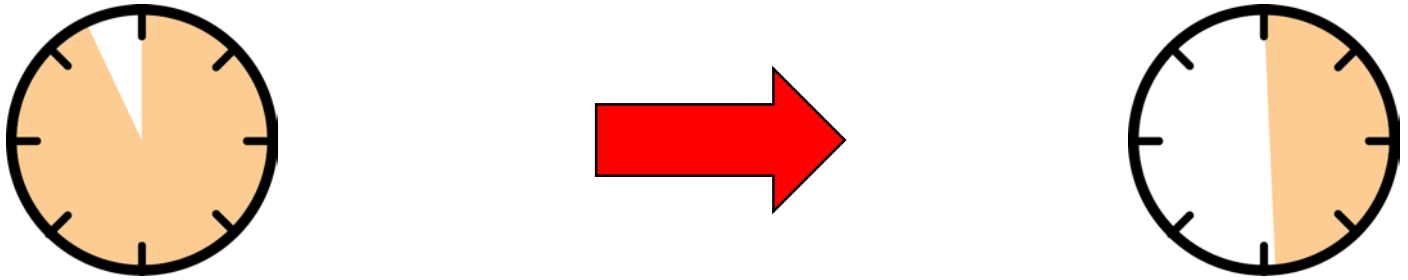
Methods for Identifying an Isolate's MLST

Traditional Methods

- Traditional methods for performing MLST entails performing Sanger sequencing on each loci followed by allele number designation
- With 2nd generation sequencing technologies, it is cheaper to sequence the whole genome over sequencing each loci

Contemporary Methods

- Once the genome has been sequenced, it undergoes quality control and genome assembly followed by sequence alignment and allele identification
- Depending on the sequencing depth, this can take a few hours per sample

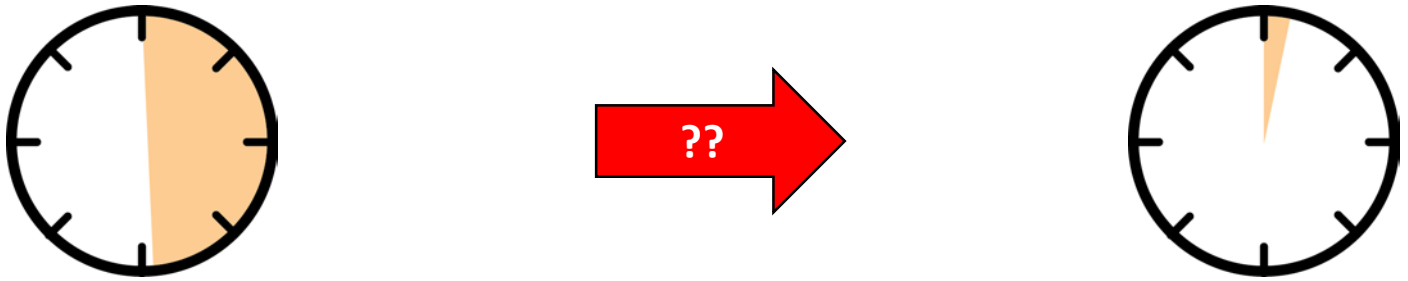


Next generation MLST methods

- The contemporary process is not ideal for an outbreak scenario
  - Process runtime unwieldy for large number of samples
  - Requires certain level of computational familiarity with the process – process bottleneck

- A more desirable method will yield MLST directly from sequence reads

- i.e., a method that can directly identify the ST from the reads with minimum expertise and time



stringMLST

- stringMLST is designed to address this specific need – accurate ST detection directly from sequence reads
- stringMLST works on a k-mer based approach that entails direct string matching from reads
- It is an accurate and rapid typing method with minimum dependence and computational familiarity
- *Works directly from NGS sequence reads: no QC, no assembly, no mapping, no alignment needed*

stringMLST

Bioinformatics, 2016, 1–3
doi: 10.1093/bioinformatics/btw586
Advance Access Publication Date: 7 September 2016
Application Note

OXFORD

Genome analysis

stringMLST: a fast k-mer based tool for multilocus sequence typing

Anuj Gupta[1,2], I. King Jordan[1,2,3] and Lavanya Rishishwar[1,2,3,*]

[1]School of Biology, Georgia Institute of Technology, Atlanta, GA 30332, USA, [2]Applied Bioinformatics Laboratory, Atlanta, GA 30332, USA and [3]PanAmerican Bioinformatics Institute, Cali, Valle del Cauca 760043, Colombia

*To whom correspondence should be addressed.

Associate Editor: John Hancock

Received on April 22, 2016; revised on August 17, 2016; accepted on September 5, 2016

Anuj Gupta

Lavanya Rishishwar

stringMLST

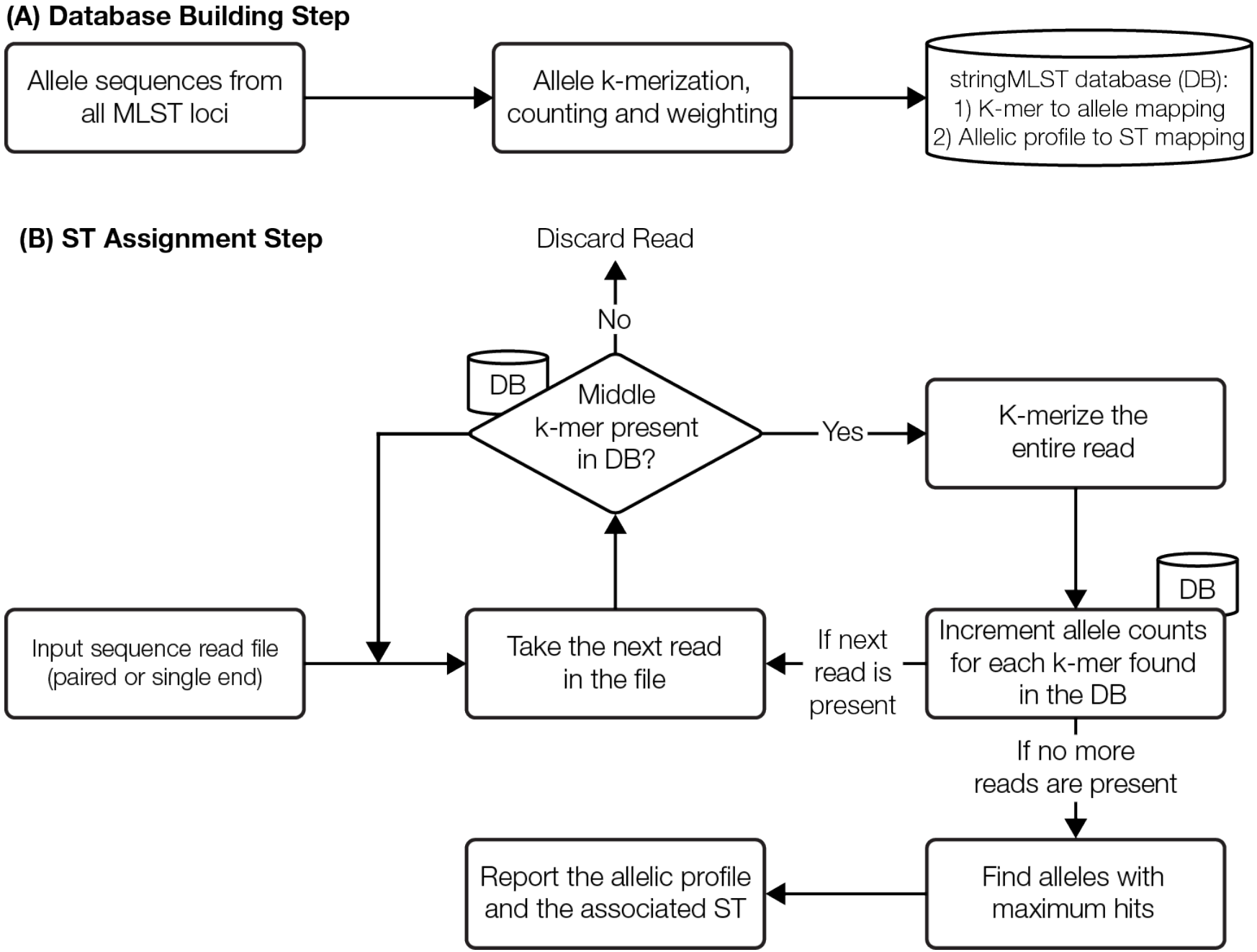
- stringMLST algorithm is based on two simple concepts:
  - Direct k-mer matching
  - Lookup tables

- It also incorporates a few smart steps to:
  - Maximize accuracy
  - Increase speed
  - Minimize user input

- **End result:** We can type an isolate from its read in under 60 **seconds** with ~**100% accuracy**

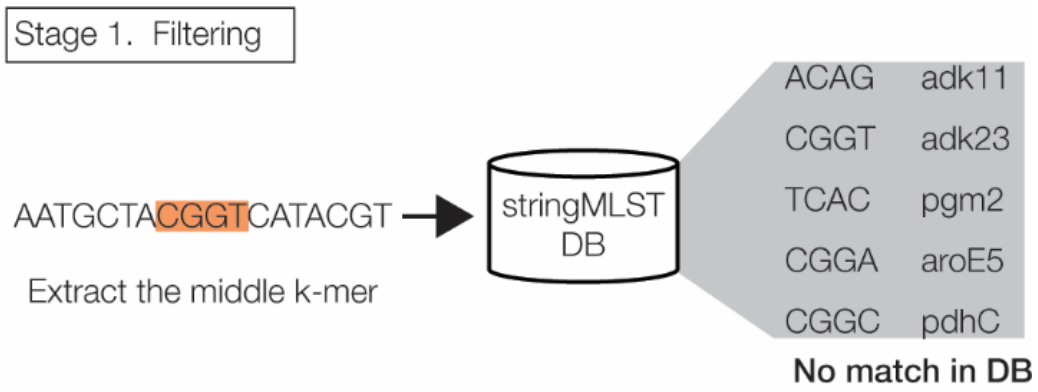
Georgia Tech Bioinformatics
Interdisciplinary Graduate Program

Algorithm overview

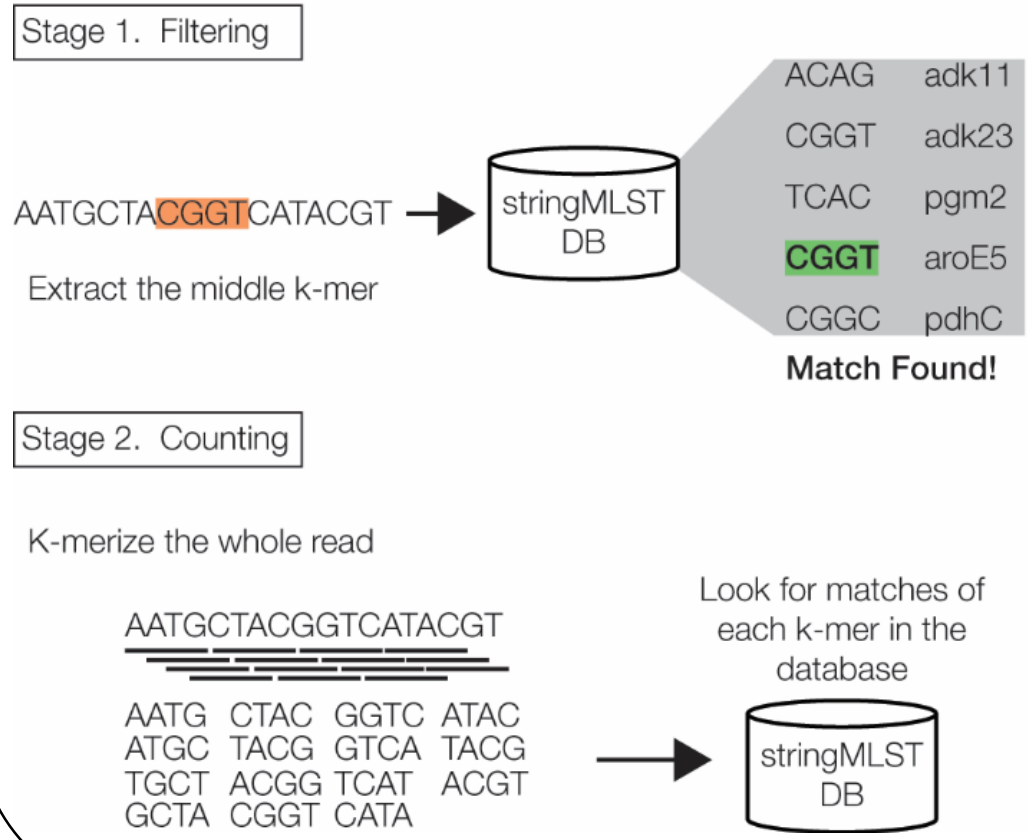


Algorithm overview

No k-mer match found



k-mer match found



Our stringMLST performance evaluation

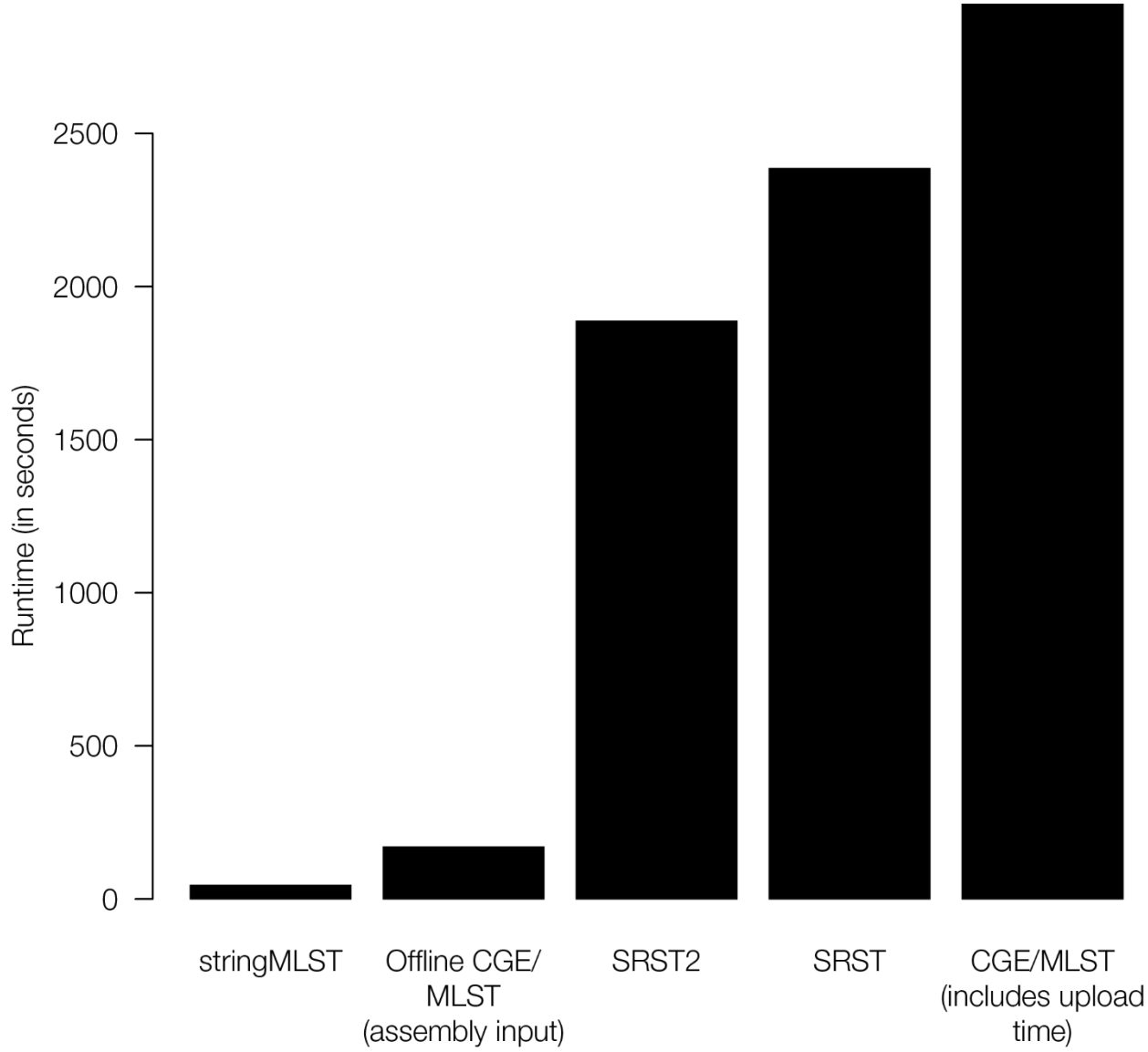
- We ran the following tests were performed to evaluate stringMLST's performance:
    - Multi-species **comparative test** – ST prediction of 10 known isolates from 4 species (=40 isolates) compared against four other tools (CGE/MLST, offline CGE/MLST, SRST, SRST2)
    - Large-scale **accuracy test** – ST prediction of known 1,002 *Neisseria meningitidis* isolates
- For each test, the number of correctly predicted alleles, STs and runtime was recorded
- 418 different STs evaluated

stringMLST performance results

Comparative Test					
Tool Name	Type	Input	% Correct		Run time (sec)
			Alleles	STs	
stringMLST	K-mer	Reads	100.0	100.0	45
CGE/MLST	BLAST	Reads	99.6	97.5	2,922
SRST2	Mapping	Reads	98.6	92.5	1,887
SRST	BLAST	Assembly	95.0	77.5	2,386
Offline CGE	BLAST	Assembly	96.1	80.0	170

Accuracy Test (stringMLST; k=35)					
#Isolates3	#Alleles	#Correctly Predicted		Run time (sec)	Memory (GB)
		STs	Alleles		
1,002	7,014	1,000 (99.8%)	7,012 (99.97%)	40.7	0.67

Comparative test results (MLST scheme)



Independent performance evaluation: User testimonials

Ryan Wick <rrwick@gmail.com> Sep 14 (9 days ago)

to lavanya.rishis., Kathryn

Hi, I've been trying out stringMLST and it seems to work well, so thanks! I hope to make more use of it in the future.

Philip Mabon <philip.mabon@phac-aspc.gc.ca> Sep 21 (2 days ago)

to Lavanya

Hello Lava,

I have added some issues to the repo which already have been address by Anuj. My testing so far has found that for MLST it has been 100% accuracy for 186/186 of Streptococcus pneumoniae strains and 39/39 for Listeria monocytogenes. Also was successfully in finding MCR resistance gene ,http://www.ncbi.nlm.nih.gov/pubmed/26603172 , in our PCR confirmed strains.

How much time and resources are you planning to allocate to improving stringMLST? From my point of view, it is worth continue testing and have plans to incorporate stringMLST into our local Galaxy instance ,https://galaxyproject.org/ . I would make the tool available on the public repository for others to install and try out. For that to happen more smoothly, it requires a few small changes to the codebase. It can be installed as is, however it would make life easier if they could be made.

I am looking followed to your reply.

Cheers,

Philip Mabon

Philip Mabon <philip.mabon@phac-asp Oct 24 (7 days ago)

to Lavanya

Hello Lava,

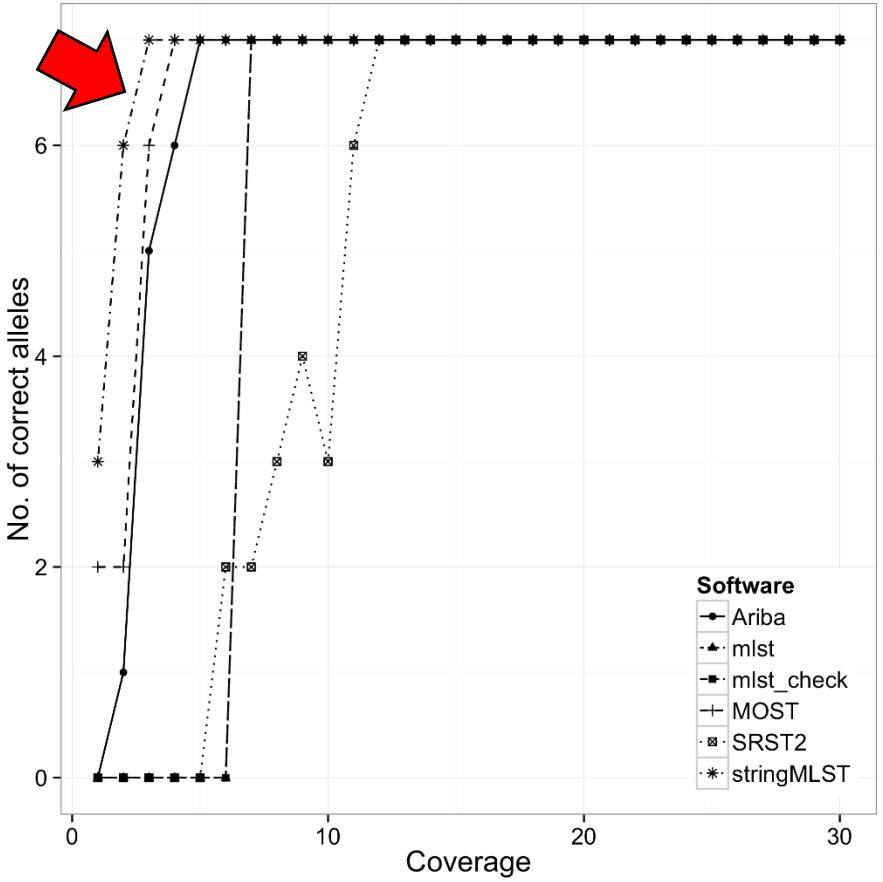
I published my initial Galaxy Wrapper on the public repository. Wrappers are normally divided into two separate repos. The binary itself https://toolshed.g2.bx.psu.edu/view/nml/package_stringmlst_2_1/75d83e0939c5 followed by the wrapper https://toolshed.g2.bx.psu.edu/view/nml/stringmlst/fc0f15ca12e0 .

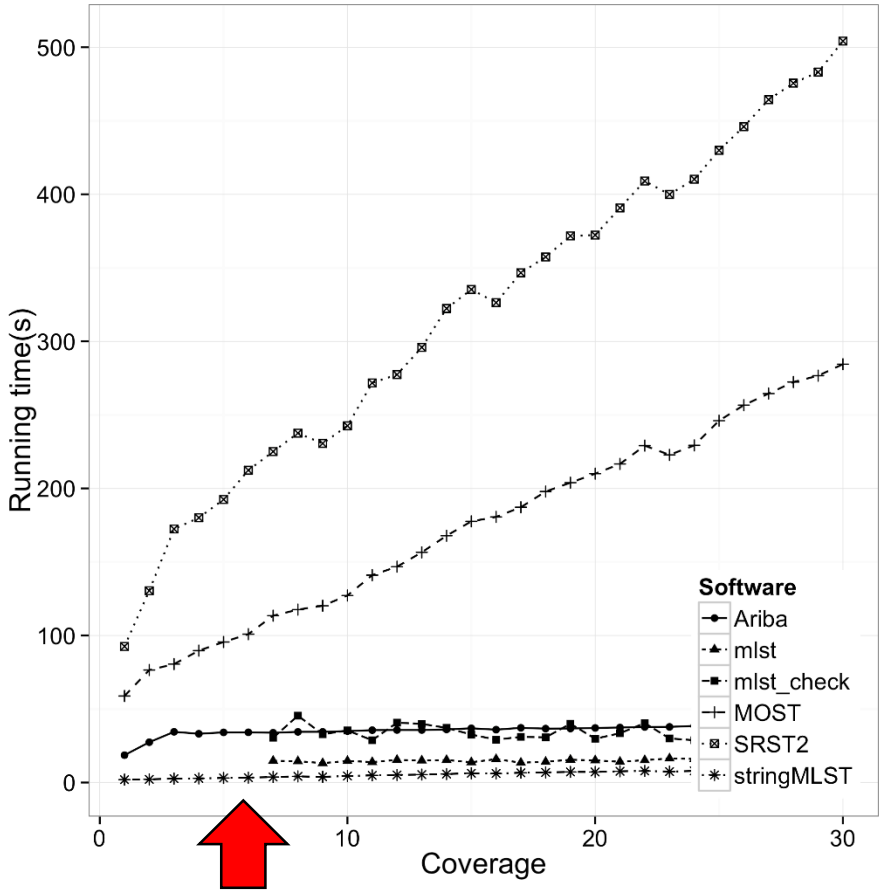
If you want any changes, please feel free to give me a shout.

Cheers,

Philip Mabon

Independent performance evaluation: Tool comparison





https://github.com/andrewjpage/docker_mlst/tree/master/results/coverage

Limited resolution of MLST

- Standard MLST still remains a popular approach in molecular epidemiology and genomic typing
- The approach suffers from the drawback of limited resolution provided from 7 genes
- 7 genes based typing methods were a good trade-off in sequencing and resolution. Sequencing is no longer a limiting factor.
- Consequently, a number of newer typing methods have been proposed such as “super-MLST” (rMLST, cgMLST, wgMLST), SNP typing, whole genome alignment and comparison

stringMLST for higher resolution typing schemes

- Thus we asked can stringMLST scale?

- *i.e.*, Can stringMLST work on bigger gene sets? 53 genes (rMLST)? Or ~1600 genes (cgMLST)?

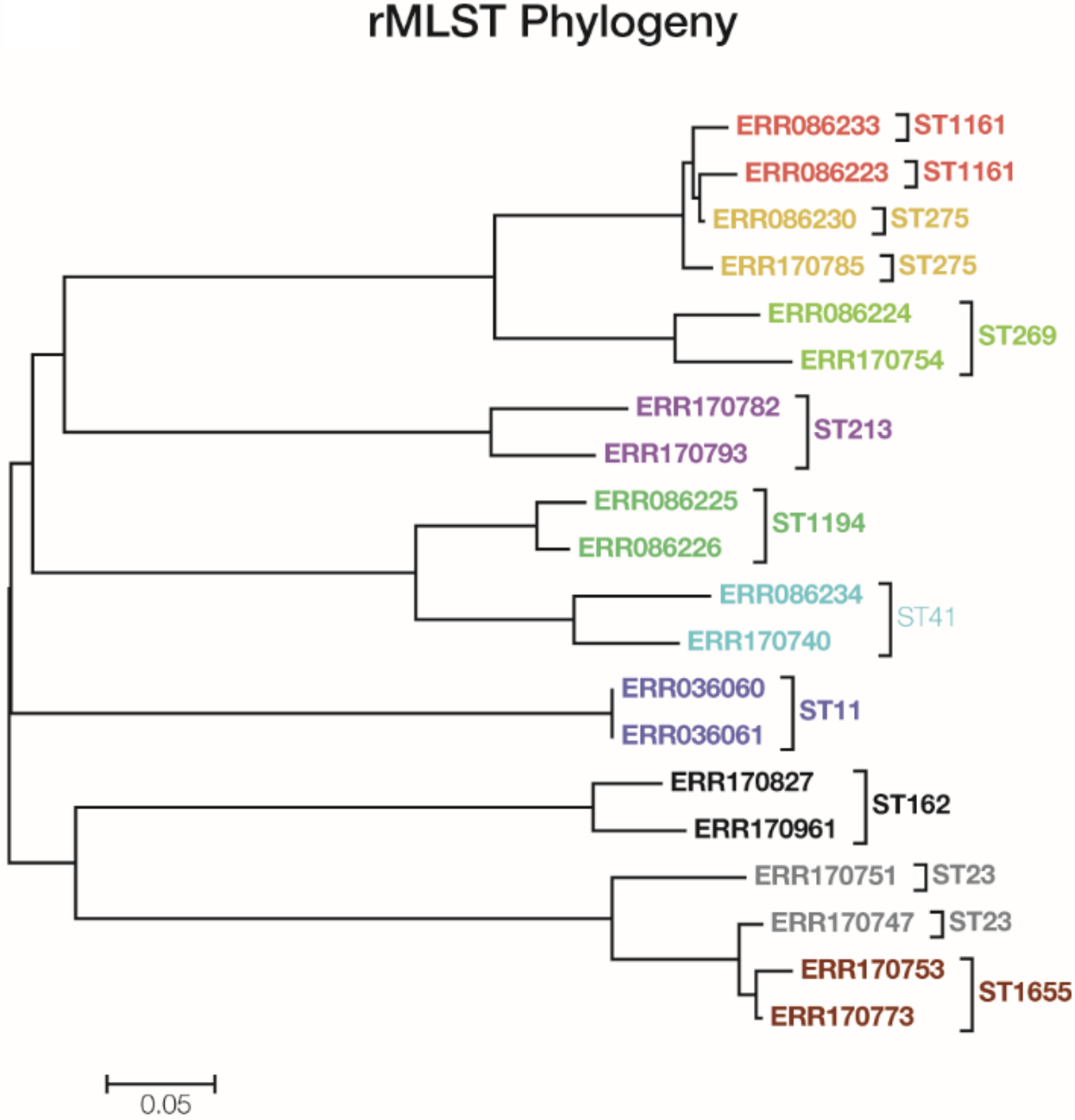
Performance on higher resolution typing schemes

Larger-scale Schemes (stringMLST vs BLAST)					
#Isolates	#Alleles	#Correctly Predicted		Processing Rate (Kb/sec)	Scheme
		Alleles	%		
20	1,060	1,009	95.2	516.7	rMLST
20	31,919	28,976	90.8	43.0	cgMLST

Slight loss of accuracy – but this can be expected and is tolerated with many loci

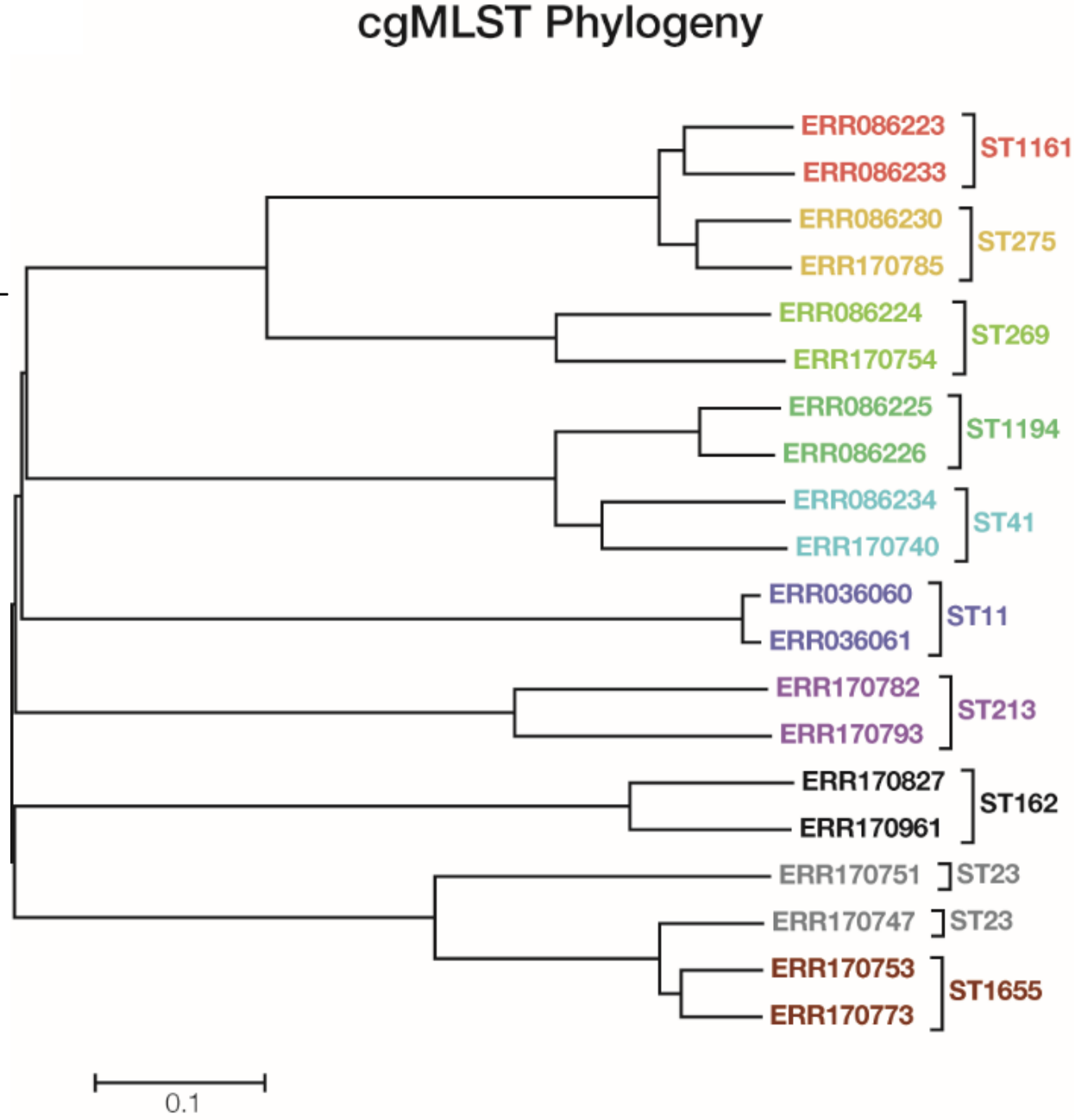
stringMLST scaling issues

- rMLST and cgMLST phylogenies from stringMLST are concordant with MLST results
- Implying, stringMLST can work for larger loci schemes.
- The biggest issue with stringMLST was the time and memory consumption (=> 100GB!!)
- stringMLST was developed with smaller scheme in mind and it wasn't optimized for larger scheme testing



stringMLST scaling issues

- rMLST and cgMLST phylogenies from stringMLST are concordant with MLST results
- Implying, stringMLST can work for larger loci schemes.
- The biggest issue with stringMLST was the time and memory consumption (=> 100GB!!)
- stringMLST was developed with smaller scheme in mind and it wasn't optimized for larger scheme testing



stringMLST - Conclusions

- stringMLST helped establish a proof of concept for exact k-mer matching in molecular typing
- The method is fast, lightweight and scalable in terms of the number of genomes to be analyzed
- Caveat – stringMLST is only as good as the database its running on … must be current
- However, it's scalability, in terms of the number of loci, is limited to a small set
- The algorithm in its current state, scales in an exponential manner which is not good

Georgia Tech Bioinformatics
Interdisciplinary Graduate Program

Aroon Chande

Lavanya Rishishwar

stringMLST public repository

PubMLST database is downloaded and bundled once a week

Constantly updated and maintained

Personal Open source Business Explore Pricing | This repository Search | Sign in or Sign up

jordanlab / stringMLST

Watch 5 | Star 5 | Fork 6

Code | Issues 0 | Pull requests 0 | Projects 0 | Pulse | Graphs

Fast k-mer based tool for multi locus sequence typing (MLST)

45 commits | 3 branches | 9 releases | 5 contributors

Branch: master | New pull request | Find file | Clone or download

ar0ch Update github URL, typos — Latest commit 26f8e63 3 days ago

datasets @ cfa5087	Datasets to submodule	4 days ago
tests/fastqs	Deleted fastq files in 'example read files' folder. Replace them by a...	5 months ago
.gitignore	Version bump for PyPi release	4 days ago
.gitmodules	Datasets to submodule	4 days ago
LICENSE.txt	Version bump for PyPi release	4 days ago
License.txt	Create License.txt	5 months ago
README.md	Update github URL, typos	3 days ago
README.rst	Version bump for PyPi release	4 days ago
download_example_reads.sh	Deleted fastq files in 'example read files' folder. Replace them by a...	5 months ago
setup.cfg	Version bump for PyPi release	4 days ago
setup.py	Update github URL, typos	3 days ago
stringMLST.py	Clean up path resolution and close #25 and push updated to PyPi	3 days ago

STing – **S**equence **T**yp**ing**

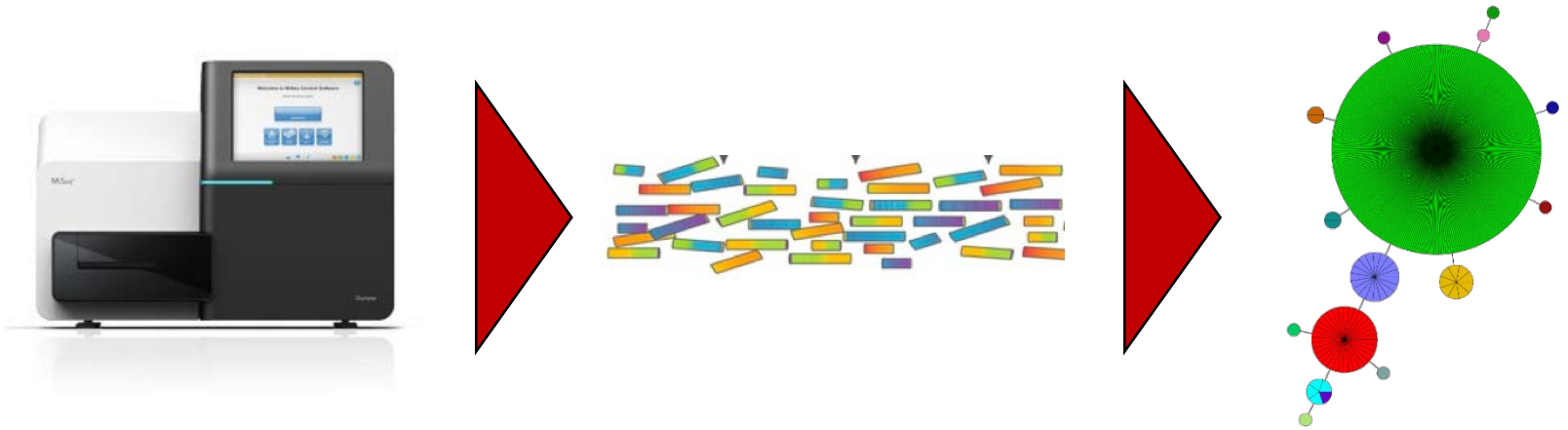
STing – Sequence **T**yp**ing**

- To address the specific shortcomings of stringMLST
- The core algorithm data structure was changed from hash tables (lookup tables) to suffix trees
- Suffix trees are data structures that help in quickly determining the membership of an input string
- Database size and search time are substantially reduced (polynomial vs. exponential)
- They are used in some of the popular bioinformatics tools – BWA, MUMmer

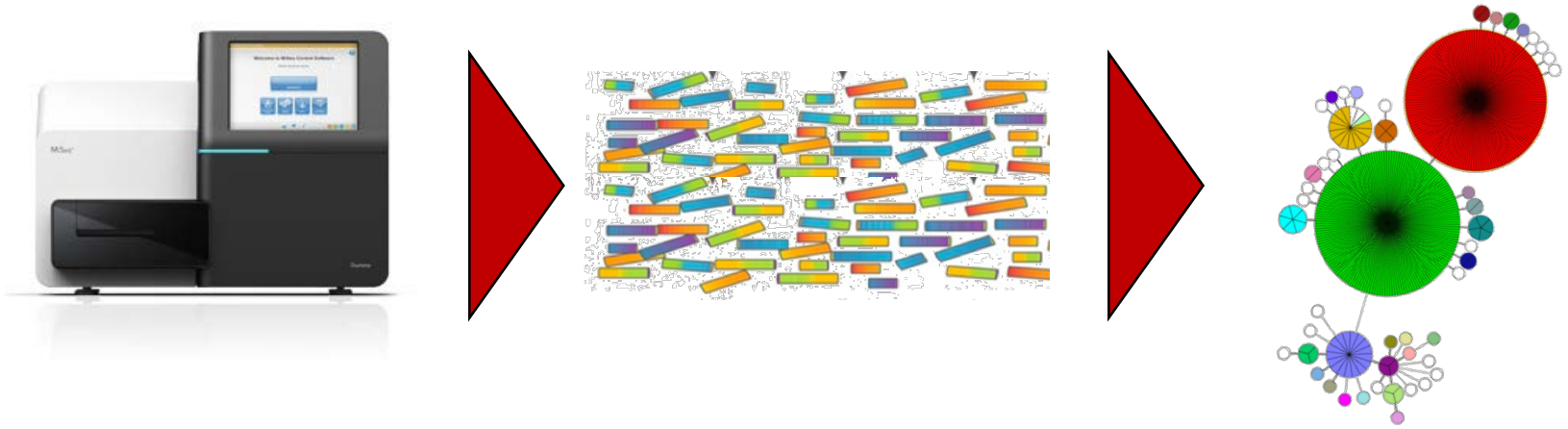
Outline

- Computational genomics class: goals and accomplishments
- Molecular epidemiology & typing in the NGS era
- Bacterial sequence typing
- Implications of NGS for molecular epidemiology & typing
- NGS-based typing methods: stringMLST, STing, others
- Future vision

Possibility of real-time, read-based molecular typing

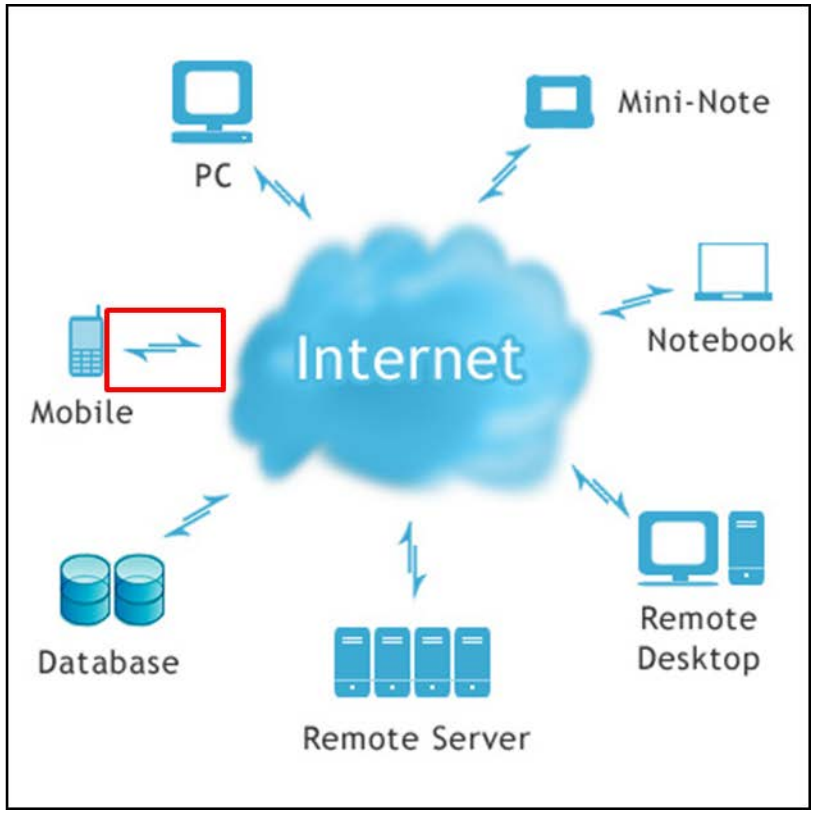


Possibility of real-time, read-based molecular typing



Possibility of real-time, read-based molecular typing

Portable genome based typing with cloud analytics

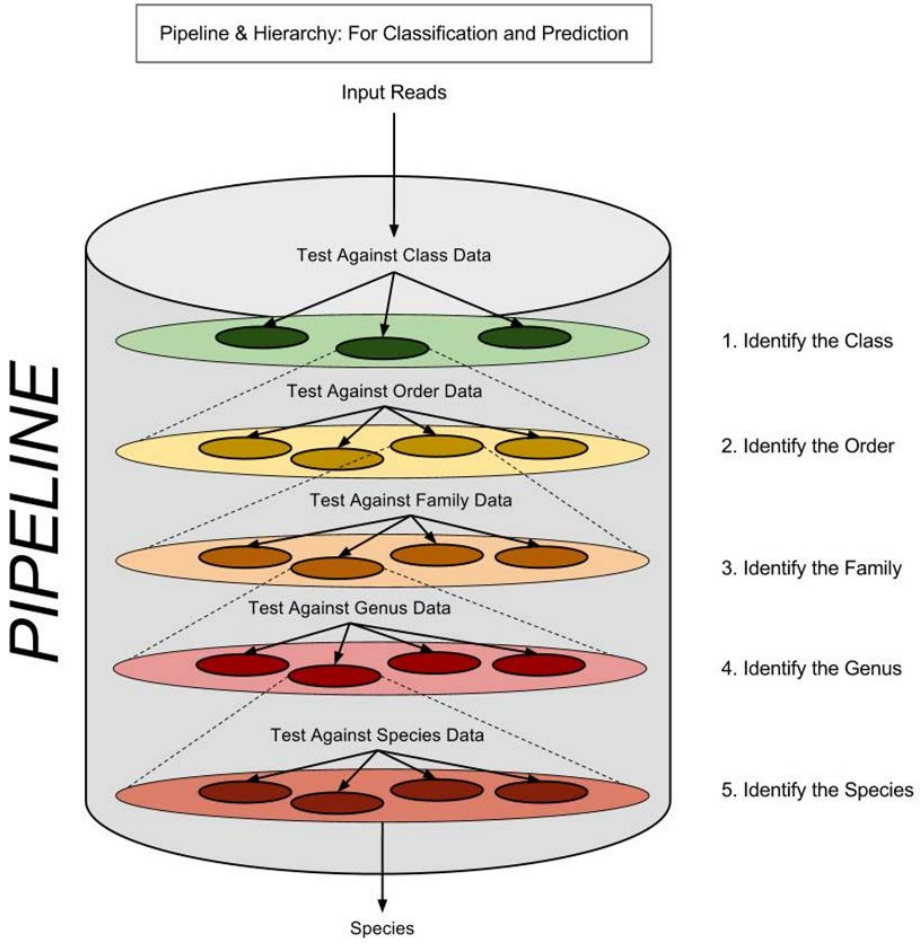


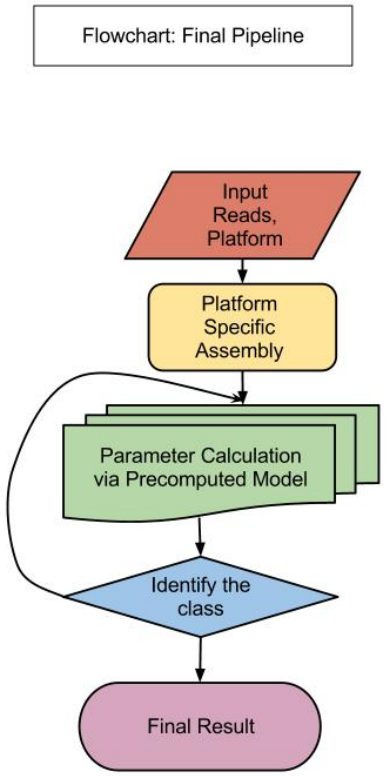
Additional Slides

Genome Based Typing with *H. haemolyticus*

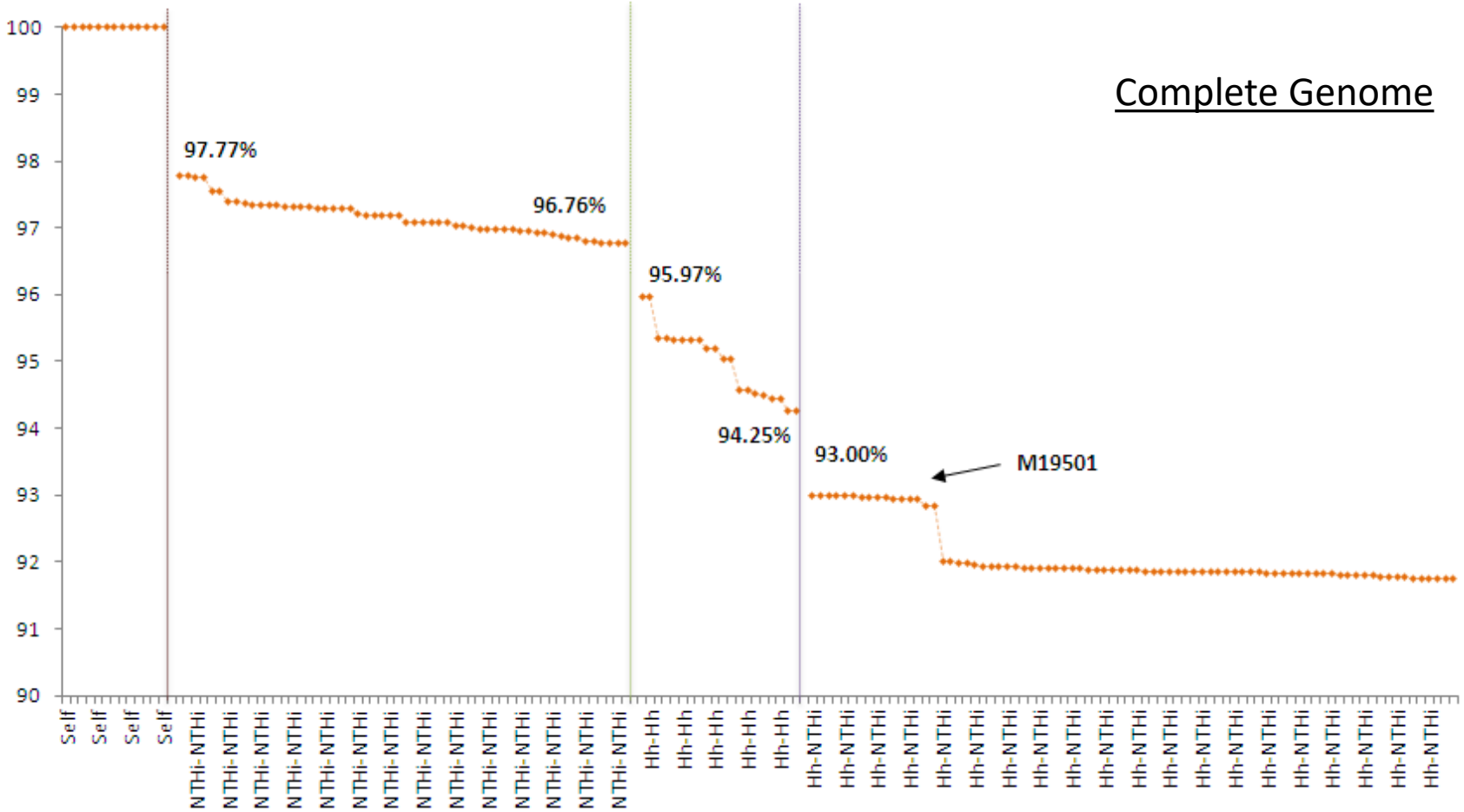
- CDC received a few reports of *H. haemolyticus* (Hhae) associated disease
- Could represent a case of an emerging pathogen
- Difficult to distinguish from non-typeable *H. influenzae* (NTHi) using conventional methods
- Want to develop typing scheme to see if previous disease cases associated with NTHi can actually be attributed to Hhae

Analytical framework for *H. haemolyticus* typing

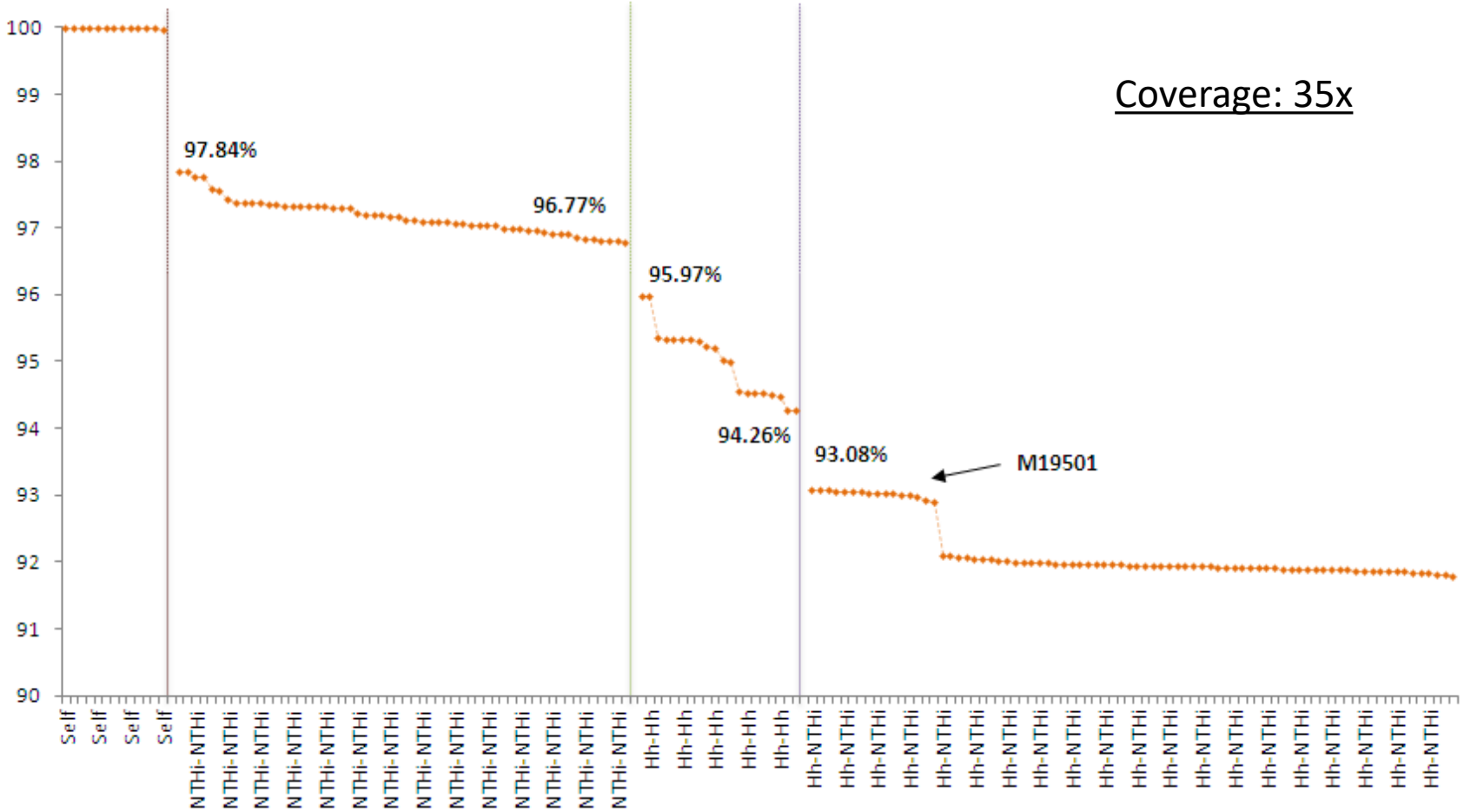




Rapid species / strain discrimination at low genomic coverage

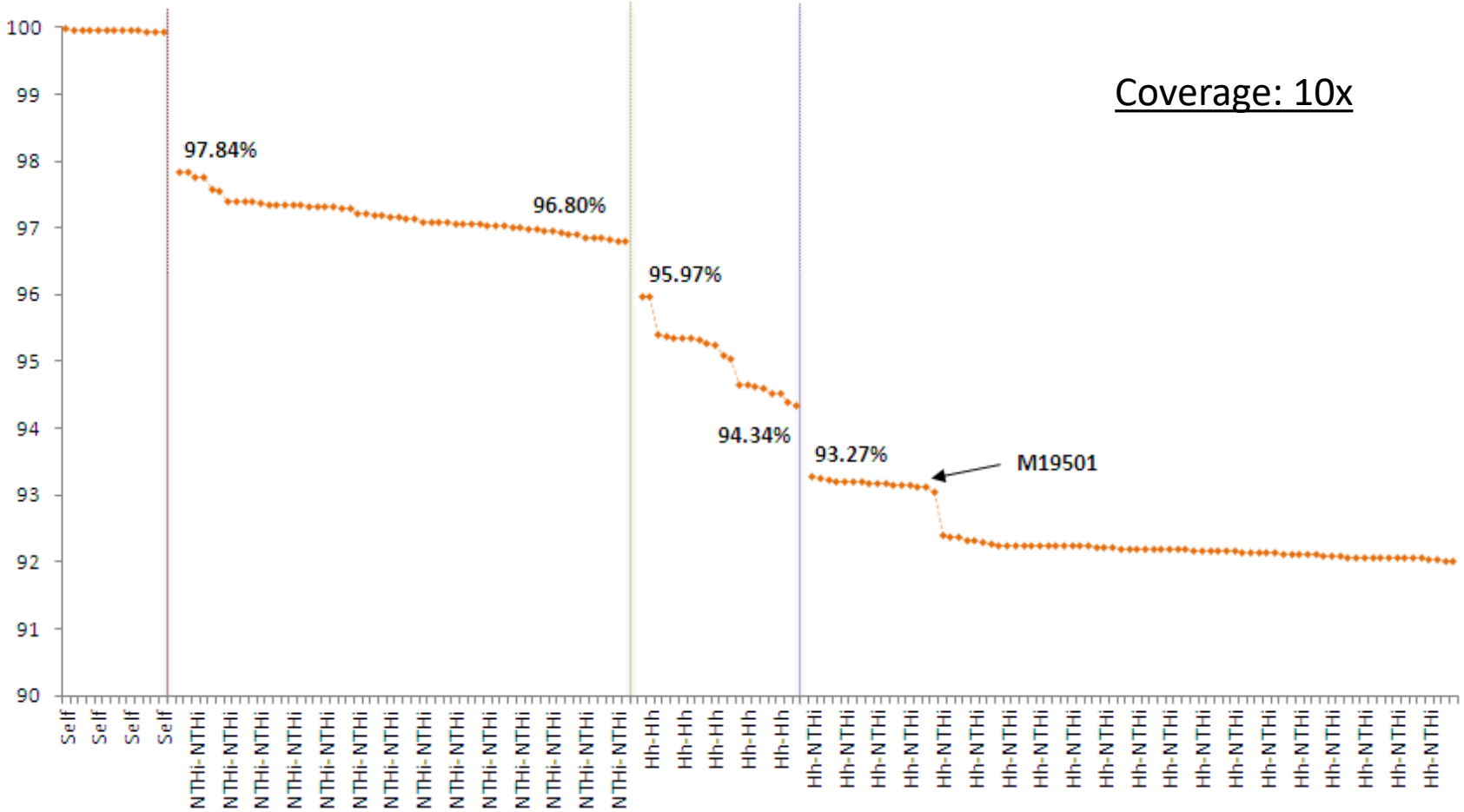


Rapid species / strain discrimination at low genomic coverage



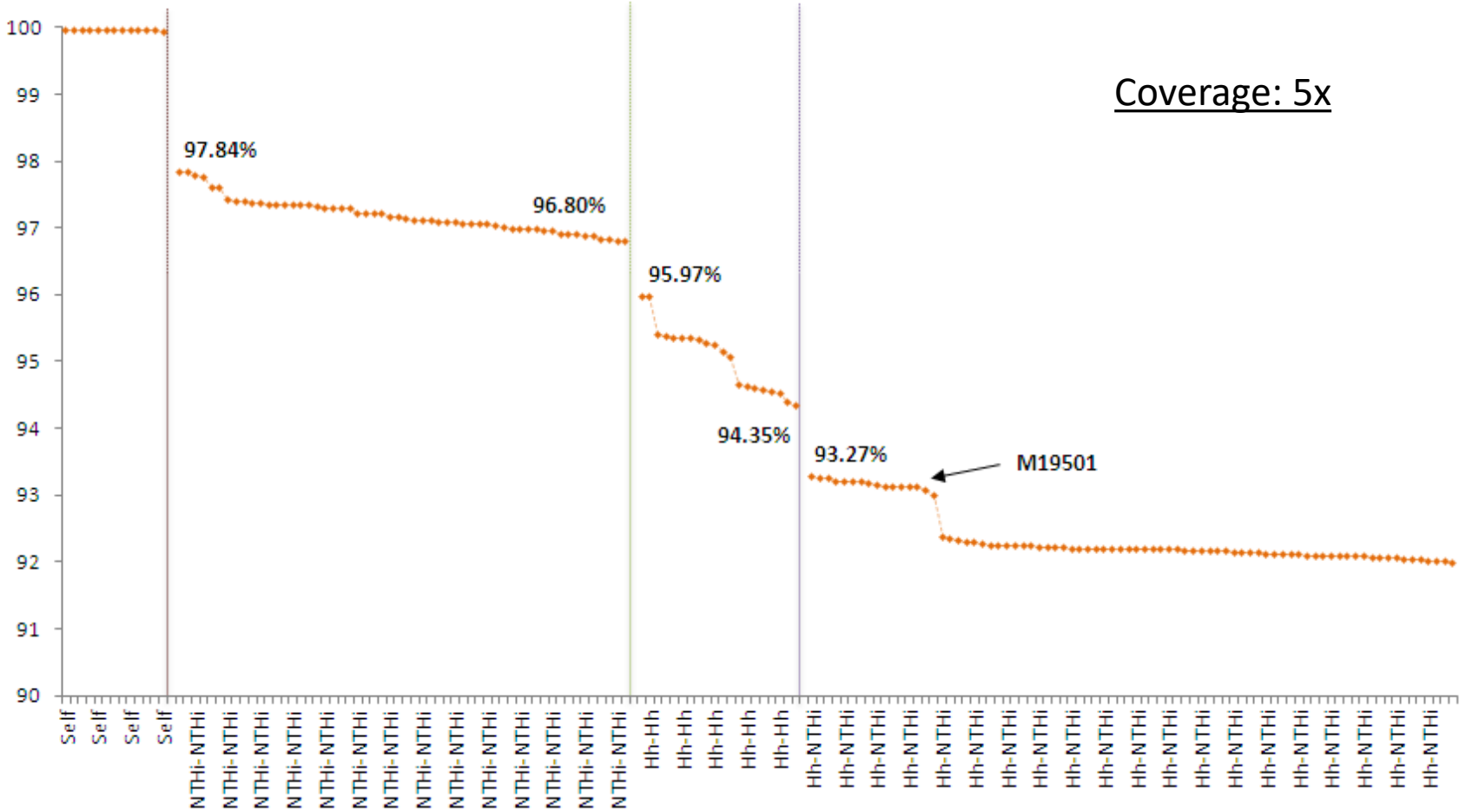
Coverage: 35x

Rapid species / strain discrimination at low genomic coverage



Coverage: 10x

Rapid species / strain discrimination at low genomic coverage



Rapid species / strain discrimination at low genomic coverage