TIGR
The Institute for Genomic Research

Genome Sequence Analysis as a Tool for Understanding and Controlling Infectious Diseases

Eric Eisenstadt, Ph.D.
The Institute for Genomic Research

Genomics and TIGR Highlights

- 1977 Sanger, F., Nicklen, S. & Coulson, A.R.
- 1986 Human Genome Project is launched
- July 1992 TIGR starts industrial-scale sequencing operation with initial focus on ESTs
- April 1994 *H. influenza* genome begun via shotgun sequencing
- May 1995 *H. influenza* genome finished
- **2006**
  - **Hundreds of genomes**
  - **10^{11} bases of DNA sequence in databases**
  - **Multiple industrial-scale genome sequencing centers**
    - **Motivated by the power and promise of genomic analysis**
    - **Enabled by improving and affordable technologies**

J. Craig Venter Joint Technology Center

Template Production Lab

Capacity enabled by robots:

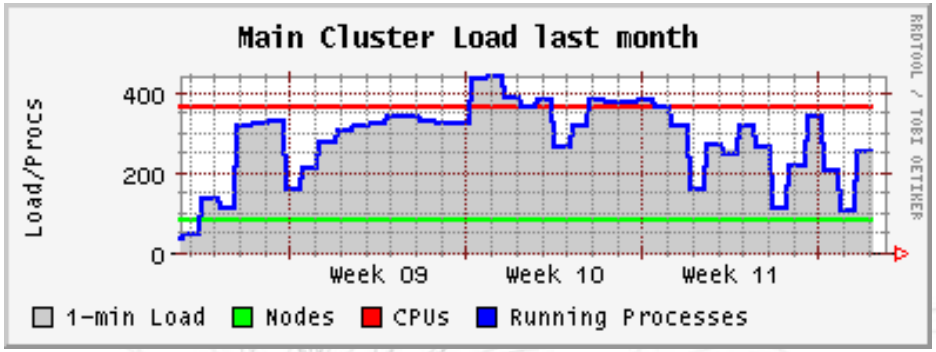
90,000 plasmids/day or 20 million plasmids/year

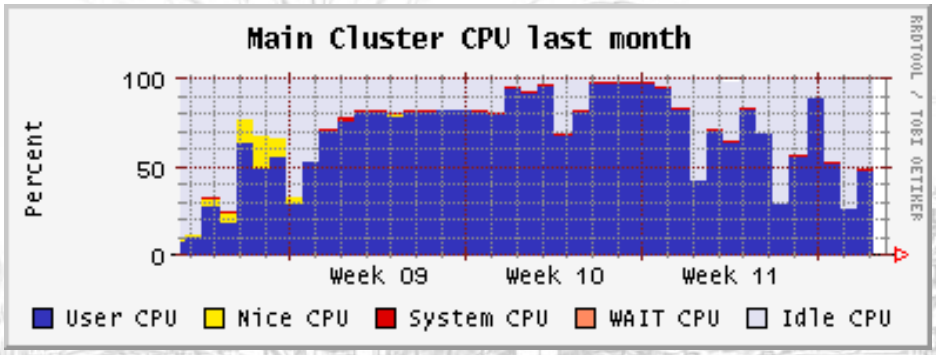
Sequencing Lab

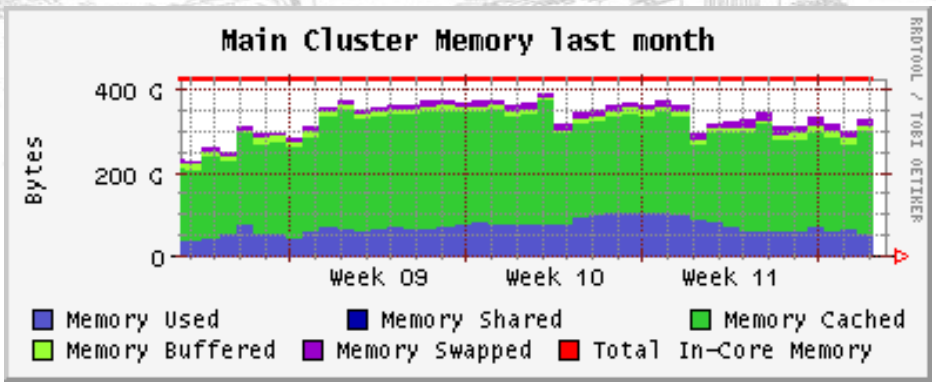
Capacity enabled by capillary sequencers:

120,000 sequences/day or 40 million lanes/year at 12 runs per day
240,000 sequences/day or 80 million lanes/year at 24 runs per day

TIGR Computational Grid







- 400 nodes of 32- and 64-bit Intel, AMD, and Sun CPUs running Linux and Solaris with aggregate memory of 400+ GB
- High Throughput (HTC) and High Performance (HPC) Computing capacity of 0.56 TFLOPS
- Access to over 50 TB of network storage
- 720 Gbits campus LAN backbone with high speed connections to the Internet and Internet2 Research Network
- **"The world is not flat; it's a point" (Smarr, UCSD Supercomputer Center)**

TIGR
THE INSTITUTE FOR GENOMIC RESEARCH

Genomics >> Sequencing

- Sequencing *per se* is the easiest part; reading the sequence is the hard part
- Critical value added is provided by subsequent computational and experimental tools
  - Sequence assembly into larger structures (scaffolds, contigs, chromosomes)
  - Annotation to predict *e.g.*, ORFs, genes, gene products
  - Database tools for bioinformatics to predict, *e.g.*, antigens, pathways, networks
  - Functional genomic reagents and assays for experimental biology (*e.g.*, clones, proteins, arrays)

IT TAKES A COMMUNITY TO PRACTICE GENOMIC ANALYSIS AND EXTRACT BIOLOGY FROM SEQUENCE

The “Drivers” for Genomics of Infectious Disease Agents

- Diagnostics
- Vaccines
- Therapeutics
- Predictive Phylogeny and Epidemiology

TRANSLATING GENOMICS RESEARCH INTO PRODUCTS AND APPLICATIONS REQUIRES CONSTANT AND ITERATIVE PULLING AND PUSHING BETWEEN BASIC AND APPLIED RESEARCH COMMUNITIES

NIAID Supported Resources at TIGR

- Microbial Sequencing Centers (TIGR + Broad Institute in Boston)
- Pathogen Functional Genomics Resource Center
- Bioinformatics Resource Centers (TIGR is one of 6 and BRC central)

Microbial Sequencing Center

http://www.tigr.org/msc/status.shtml

TIGR > MSC > Production Status

- About Us
- What's New
- Production Status
- Genome Projects
- Publications
- Resources
- Data Release Policies
- Team Members
- >> NIAID MSC

Production Status

TIGR's current MSC effort includes the following genome projects. Species/strain names link to the appropriate MSC site, if any, and other links lead to GenBank records. A 'p' indicates submission is in progress, and an 's' indicates the data has been submitted but is not yet available.

Recent status changes are listed on the What's New page. For a list of genome projects with individual MSC web sites, please see the Genome Projects page.

The projects on this page are organized into the following categories:

- Disease Vectors
- NIAID Category A Pathogens
- NIAID Category B Pathogens
- NIAID Category C Pathogens
- Other Pathogens
- Related Species

All MSC data is rapidly deposited in public databases at the NCBI

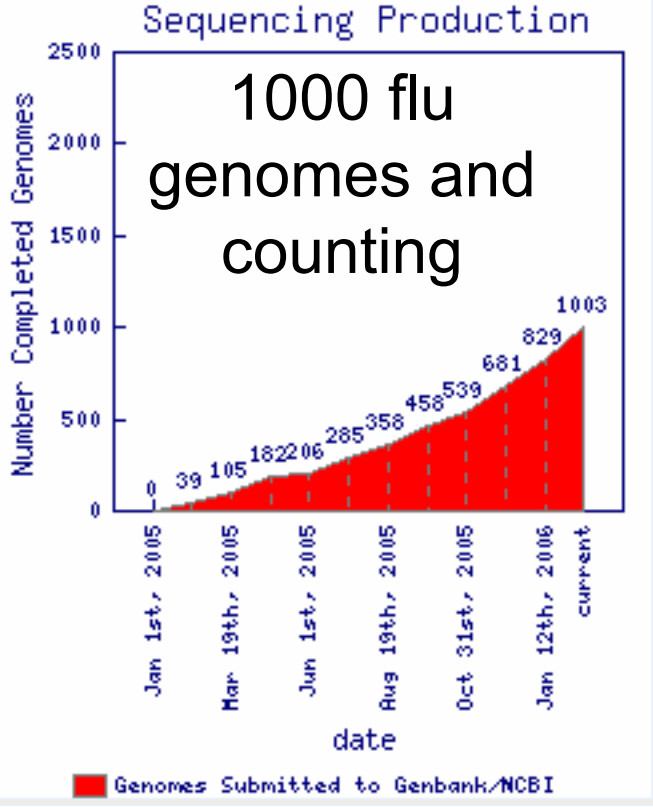
About Us
What's New
Production Status
Genome Projects
Influenza A Virus
- Collections
- Download
- Pipeline Schema
- Publications
- Contact Us

Publications
Resources
Data Release Policies
Team Members
>> NIAID MSC

Influenza A Virus Genome Project

Goals

This project aims to dramatically improve the availability of influenza genomic sequence in the public domain. We will sequence the complete genomes of a large collection of human influenza A isolates, as well as a select number of avian and other non-human influenza strains. The strains will be chosen to represent a wide geographical and chronological survey of the influenza virus. All data will be released immediately to the public domain.



http://pfgrc.tigr.org/

Resources for the Study of Pathogens

DNA Microarrays

Gateway® Cloning

Protein Expression

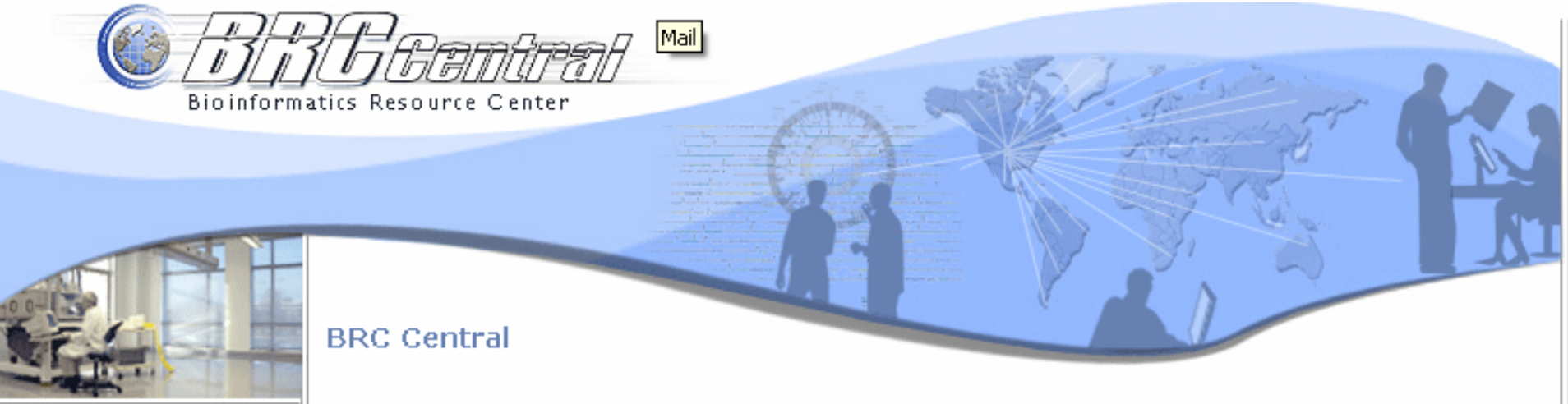
Proteomics

Comparative Genomics

Distribution to International Pathogen Research Community

Data and Databases

http://www.brc-central.org/cgi-bin/brc-central/brc_central.cgi



- Home
- Search Tools
- Conferences
- FTP
- Organisms
- Software
- Links
- Contact Us
- About Us

BRC Central

ApiDB | BioHealthBase | ERIC | NMPDR | Pathema | Patric | VBRC | VectorBase

BRC Central A repository linking to eight Biodefense Resource Centers (BRCs) sponsored by the NIAID. The BRCs are providing web-based resources to scientific community conducting basic and applied research on organisms considered potential agents of biowarfare or bioterrorism or causing emerging or re-emerging diseases.

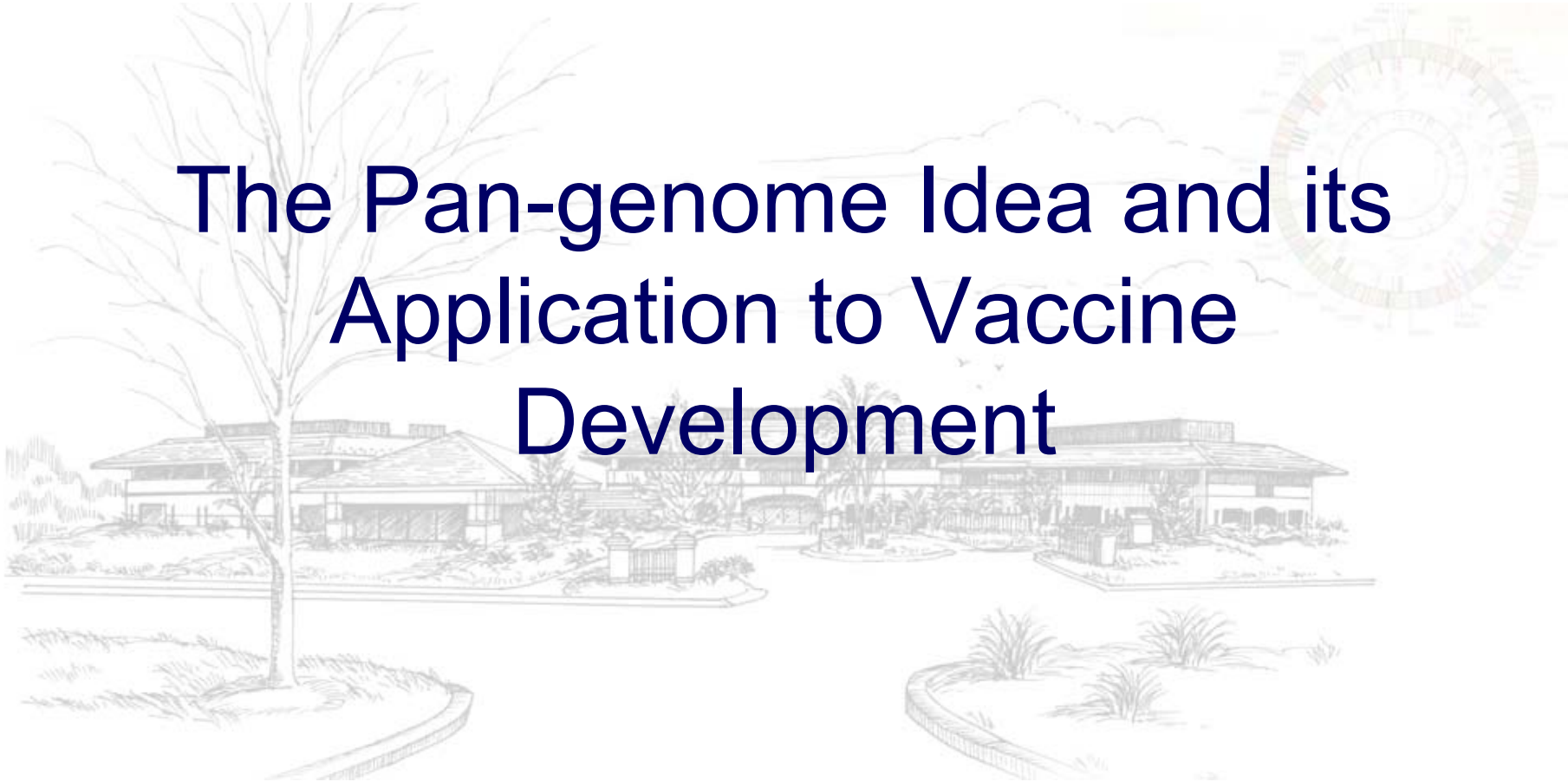
These centers support existing and newly developed techniques for bioinformatic analysis aimed at obtaining a deeper understanding of the fundamental biology of a specific set of pathogenic organisms, and efforts to counter the threats posed by these pathogens.

ApiDB -- **Apicomplexan database** portal to several sites including Toxoplasma gondii, Cryptosporidium parvum.

BioHealthBase -- **The Biodefense/Public Health DataBase** focuses on data about six priority pathogens to help fill in gaps in genomic and other data critical to scientific researchers. The six pathogens are: Giardia lamblia parasite, Mycobacterium tuberculosis, Influenza virus, Francisella Tularensis, Microsporidia parasites and Ricinus communis (castor bean).

ERIC -- **Enteropathogen Resource Integration Center** resource for five members of the family Enterobacteriaceae including: Diarrheagenic E. coli, Shigella, Salmonella, Yersinia enterocolitica, Yersinia

TIGR
THE INSTITUTE FOR GENOMIC RESEARCH



The Pan-genome Idea and its Application to Vaccine Development

Pan-genome Idea

- What is the genomic space of a bacterial species?
- Bacterial chromosomes engage in dynamic information exchange with plasmids, phage and viruses
- Only ~8% of all sequenced bacteria have had >2 isolates sequenced

Number of genomes sequenced in different bacterial species.

Species with sequenced genome(s)	Number of species (% of the total)	Number of genomes sequences per species
Streptococcus agalactiae, *Bacillus anthracis*, *Burkholderia mallei*	3 (1.2%)	8
Burkholderia pseudomallei	1 (0.4%)	7
Staphylococcus aureus, *Streptococcus pyogenes*	2 (0.8%)	6
Salmonella enterica, *Escherichia coli*, *Bacillus cereus*, *Chlamydophila pneumoniae*, *Haemophilus influenzae*, *Listeria monocytogenes*, *Xylella fastidiosa*	7 (2.8%)	5
Prochlorococcus marinus, *Buchnera aphidicola*, *Burkholderia cenocepacia*, *Ehrlichia ruminantium*, *Legionella pneumophila*, *Pseudomonas syringae*, *Streptococcus thermophilus*, *Yersinia pestis*	8 (3.2%)	3
Streptococcus pneumoniae, *Mycobacterium tuberculosis*, *Neisseria meningitidis*, *Bacillus licheniformis*, *Bifidobacterium longum*, *Campylobacter jejuni*, *Chlorobium phaeobacteroides*, *Corynebacterium glutamicum*, *Haemophilus somnus*, *Helicobacter pylori*, *Lactococcus lactis*, *Leptospira interrogans*, *Mycoplasma genitalium*, *Pseudomonas aeruginosa*, *Shigella flexneri*, *Staphylococcus epidermidis*, *Synechococcus elongates*, *Thermus thermophilus*, *Tropherymaa whipplei*, *Vibrio vulnificus*, *Xanthomonas campestris*	21 (8.3%)	2
Various species	211 (83.3%)	1

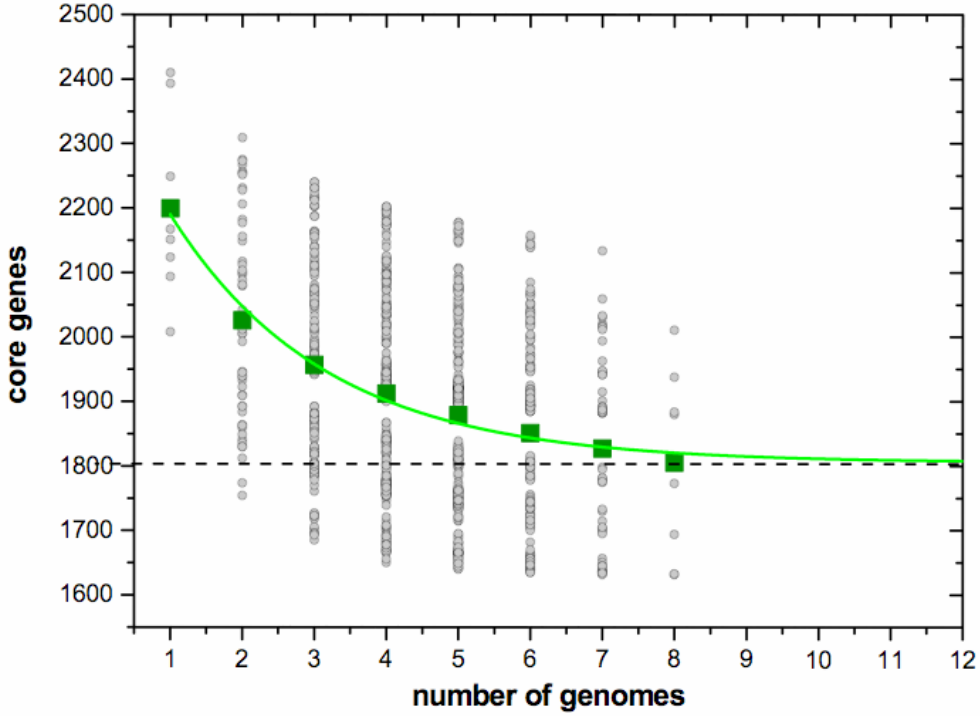
Medini et al., 2005

Streptococcus agalactiae = Group B Strep = GBS

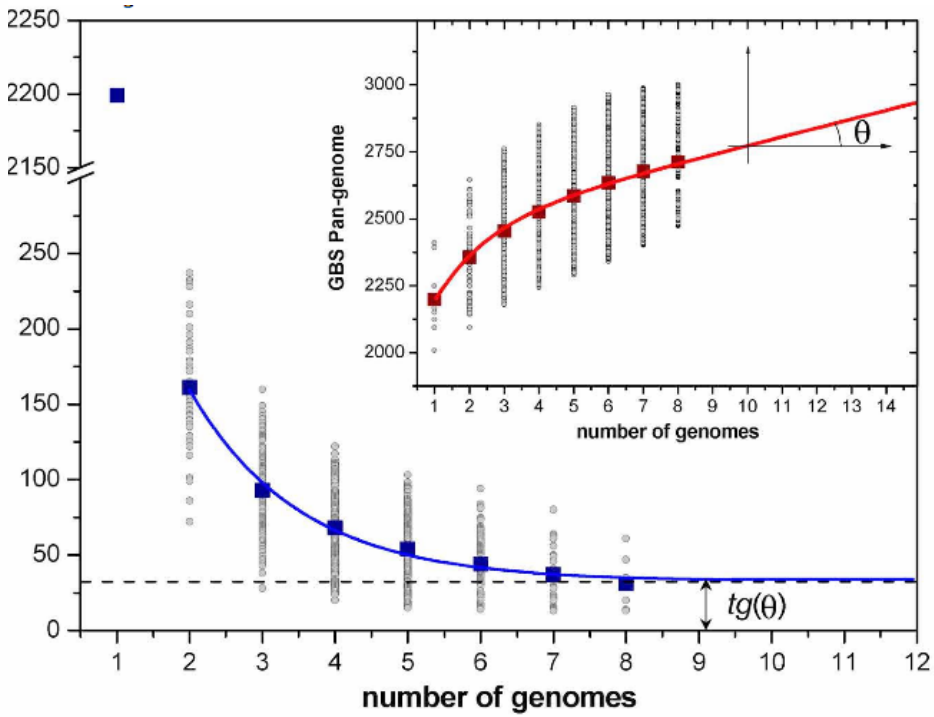
- Responsible for most meningitis in newborn infants (0 – 2 months)
- Most of the infections occur during delivery
- Capsular polysaccharides elicit protective antibodies which, if present in the mother, are transferred to the baby through the placenta and prevent infection
- 9 capsular serotypes, 5 major serotypes associated with disease

GBS Core and Pan Genomes

Genes **common** to all strains



Genes **unique** in all strains



of **shared** genes as a function of number of strains sequenced plateaus at ~ 1800

of **new** genes as a function of number of strains sequenced plateaus at ~ 33

Implications for Taxonomy and Vaccine Development

- Classical taxonomic approaches for classifying isolates rely on invariable core genome features such as 16s rRNA typing
- But the features we most care about such as virulence, serotype, and antibiotic resistance are encoded by the highly variable parts of the genome

Antigen combinations from core and pan-genome protect against 92% of tested strains

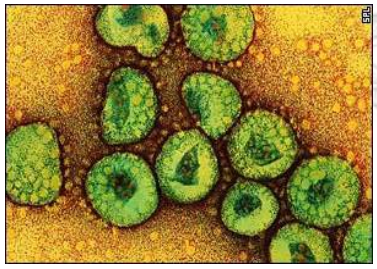
		FACS (Δ Mean)			MIX=322+80+104+67		PBS	
GBS strains	Type	GBS 80	GBS 67	GBS 322	alive/treated	% protection	alive/treated	% protection
515	Ia	0	409	227	39/40	97	6/40	15
7357b-	Ib	91	316	102	19/30	63	1/30	3
DK21	II	0	331	416	25/34	73	17/48	35
5401	II	170	618	135	35/40	87	3/37	8
3050	II	43	460	188	48/48	100	1/30	3
COH1	III	305	0	130	36/36	100	7/40	17
M781	III	65	0	224	30/40	75	4/39	10
2603	V	125	105	313	27/33	82	10/35	28
CJB111	V	370	481	63	25/28	89	4/46	9
JM9130013	VIII	597	83	143	37/39	95	5/40	12
JMU071	VIII	556	79	170	44/50	88	18/50	36
NT1169	NT	0	443	213	12/32	37	11/35	31

Immunizations:
3 doses of 15 μg of each protein in Freund adjuvant

1 antigen is from the core genome
3 antigens are from the pan-genome

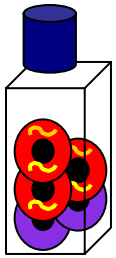
TIGR
THE INSTITUTE FOR GENOMIC RESEARCH

Tracking Influenza via Whole Genome Sequencing



TIGR
THE INSTITUTE FOR GENOMIC RESEARCH

1



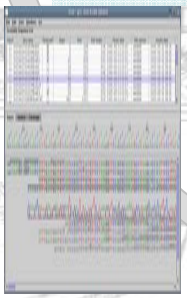
2

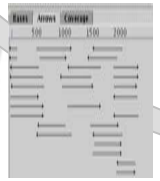
3

4

5

6





7

8

Seg1 (PB2)
Seg2 (PB1)
Seg3 (PA)
Seg4 (HA)
Seg5 (NP)
Seg6 (NA)
Seg7 (M)
Seg8 (NS)

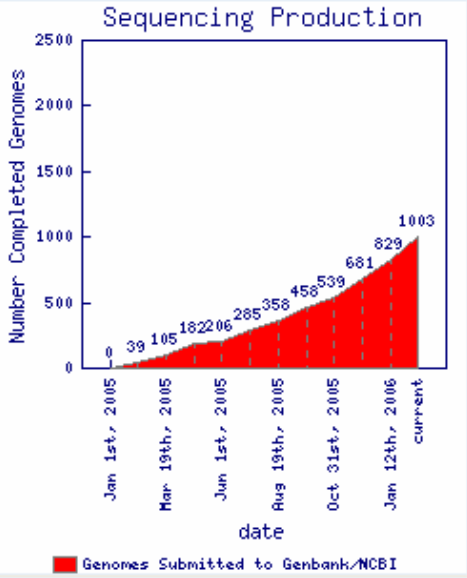
1. Collect and amplify isolates
2. Extract viral RNA
3. Ship to TIGR
4. Primer design in 96-well plate format
5. Amplify by RT-PCR tiling
6. Sequence amplicons
7. Trim sequences, assemble segments, edit nucleotides
8. Review for quality and submit to NCBI

Influenza A Virus Genome Project

Released Data

Goals

This project aims to dramatically improve the availability of influenza genomic sequence in the public domain. We will sequence the complete genomes of a large collection of human influenza A isolates, as well as a select number of avian and other non-human influenza strains. The strains will be chosen to represent a wide geographical and chronological survey of the influenza virus. All data will be released immediately to the public domain.



~ 50 genomes/week

Bulk Download all available genome assembly files
Download this data as an Excel spreadsheet
Download this data as a CSV spreadsheet

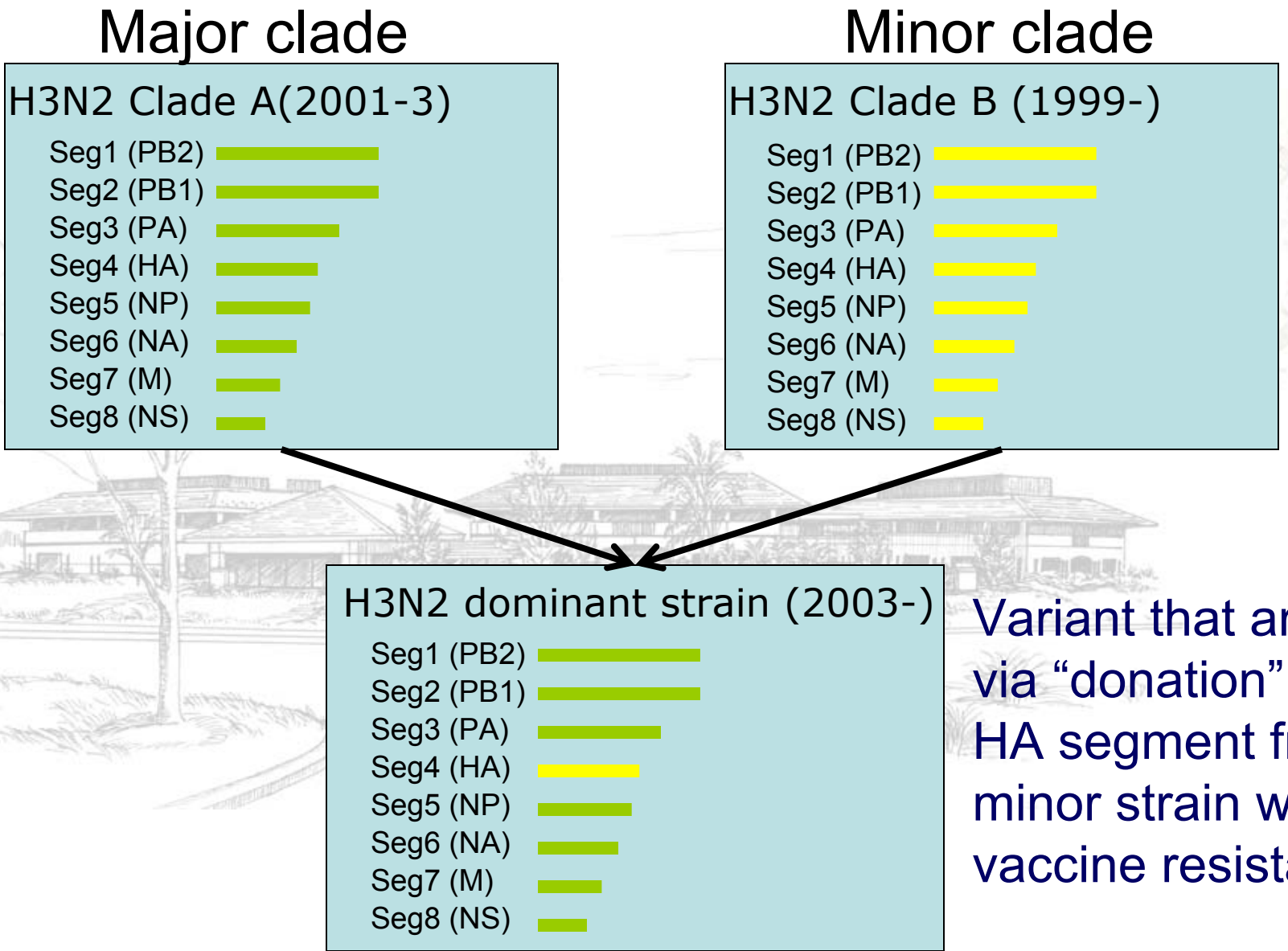
Click Here to View The Full List of All Avaliable Genome Assemblies

Downloads (Last updated Wed Mar 22 01:58:21 EST 2006)

File	Type	size (MB)
Nucleotides Fasta	Fasta	12.12
Proteins Fasta	Fasta	4.75
Submissions Catalog	Microsoft&trade Excel®	2.13
Proteins Catalog	CSV	0.91
Nucleotides Catalog	CSV	0.66
Amplicons Catalog	Microsoft&trade Excel®	3.56
Primers Catalog	Microsoft&trade Excel®	0.12

Page last updated: 21 March 2006; MSC site version: 2.09

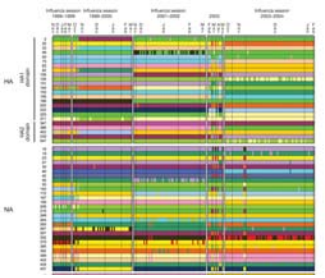
H3N2 Antigenic Shift Event in 2002-2003

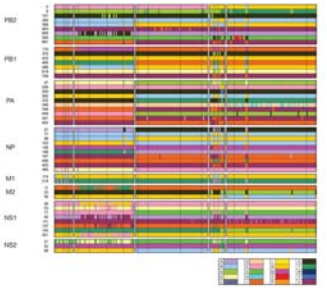


Variant that arose via “donation” of HA segment from minor strain was vaccine resistant

TIGR
THE INSTITUTE FOR GENOMIC RESEARCH

Some Conclusions





- Variation occurs in all gene segments
- Multiple lineages co-circulate within a flu season
- **Reassortment of variants of same subtypes can lead to emergence of epidemiologically relevant antigenic novelty**
- **Whole genome sequence-based sampling can reveal co-circulating strains before emergence of antigenic novelty, *i.e.* sequencing is a surveillance tool**

Open access, freely available online PLoS BIOLOGY

Whole-Genome Analysis of Human Influenza A Virus Reveals Multiple Persistent Lineages and Reassortment among Recent H3N2 Viruses

Edward C. Holmes[1], Elodie Ghedin[2], Naomi Miller[2], Jill Taylor[3], Yiming Bao[4], Kirsten St. George[3], Bryan T. Grenfell[1], Steven L. Salzberg[2], Claire M. Fraser[2], David J. Lipman[4]*, Jeffery K. Taubenberger[5]

1 Center for Infectious Disease Dynamics, Department of Biology, Pennsylvania State University, University Park, Pennsylvania, United States of America, **2** Institute for Genomic Research, Rockville, Maryland, United States of America, **3** Wadsworth Center, New York State Department of Health, Albany, New York, United States of America, **4** National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Department of Health and Human Services, Bethesda, Maryland, United States of America, **5** Department of Molecular Pathology, Armed Forces Institute of Pathology, Rockville, Maryland, United States of America

nature Vol 437|20 October 2005|doi:10.1038/nature04239

LETTERS

Large-scale sequencing of human influenza reveals the dynamic nature of viral genome evolution

Elodie Ghedin[1], Naomi A. Sengamalay[1], Martin Shumway[1], Jennifer Zaborsky[1], Tamara Feldblyum[1], Vik Subbu[1], David J. Spiro[1], Jeff Sitz[1], Hean Koo[1], Pavel Bolotov[2], Dmitry Dernovoy[2], Tatiana Tatusova[2], Yiming Bao[2], Kirsten St George[3], Jill Taylor[3], David J. Lipman[2], Claire M. Fraser[3], Jeffery K. Taubenberger[4] & Steven L. Salzberg[1,5]

Flu Project Collaborators are Numerous and Widely Distributed

Center for Bioinformatics and Computational Biology, University of Maryland

National Center for Biotechnology Information, National Institutes of Health

Clinical Virology Program, Wadsworth Center, New York State Department of Health

Armed Forces Institute of Pathology

Centers for Disease Control and Prevention

Ohio State University

Mount Sinai School of Medicine

Baylor College of Medicine

St. Jude's Children's Hospital (Memphis)

Canterbury Health Laboratories, New Zealand

Los Alamos National Laboratories

Other Virus Projects at TIGR

- Avian flu whole genome sequencing
- Other RNA virus sequencing projects (rhinovirus, coronavirus)
- Diversity, discovery and characterization of novel viruses

- **David Spiro, Ph.D. (dspiro@tigr.org) is the point of contact for further information about these projects**

Challenges for Controlling Infectious Diseases

- Limited data
- Limited technologies
- Limited understanding of biology
- Moving from basic research to products and applications

TIGR
THE INSTITUTE FOR GENOMIC RESEARCH

Opportunities

- Interdisciplinary education and training to bring math, physics and engineering into biology
  - New technologies
  - New mathematical and computational frameworks for biology
- Communication between the applied and basic research communities is a two-way street

Pan-genome Project Acknowledgements

Chiron Vaccines

Vega Masignani
Duccio Medini
Claudio Donati
John Telford
Rino Rappuoli

Harvard Medical School

Michael Cieslewicz
Michael Wessels
Larry Madoff
Dennis Kasper

TIGR

Naomi Ward
Samuel Angiuoli
Jonathan Crabtree
Scott Durkin
Claire Fraser

Influenza Project Acknowledgments

Viral Genomics Lab
Elodie Ghedin
David Spiro
Naomi Sengamalay
Vivien Dugan
Rebecca Halpin
Alex Boyne
Shiliang Wang

Closure Team
Tamara Feldblyum
Jennie Zaborsky
Vic Subbu
Jeffrey Sparenborg

Informatics
Martin Shumway
Jeff Sitz
Dan Katzel

Univ. of Maryland
Steven Salzberg

Wadsworth Center
Kirsten St. George
Jill Taylor
Sara Griesemer

AFIP
Jeff Taubenberger

Penn State
Edward Holmes

NCBI/NIH
David Lipman
Yiming Bao
Tatiana Tatusova

Funding:
NIAID/NIH
DHS HSARPA