Big Data Analysis of Integrative Conjugative Exchange (ICE) and AMR Evolution

James Kaufman
IBM Research, Almaden

Science to Solutions @ IBM Research – Almaden

a cross-disciplinary research group focused on microbes and molecules

IBM Research

A Agarwal
K Beck
G DuBois
J Hedrick
JH Kaufman
E Kandogan
H Krishnareddy
G Nayar
N Park
V Piunova
M Roth
E Seabolt
I Terrizzano
D Zubarev

University of Minnesota, College of Veterinary Medicine

N. Noyes et al.

Institute of Bioengineering and Nanotechnology, Singapore

Yi Yan Yang et al.

Research Areas:

- Metagenomics
- Proteomics
- Genome Assembly
- Food Safety
- Antibiotic Resistance
- Cellular Engineering
- Microbiome

Expertise:

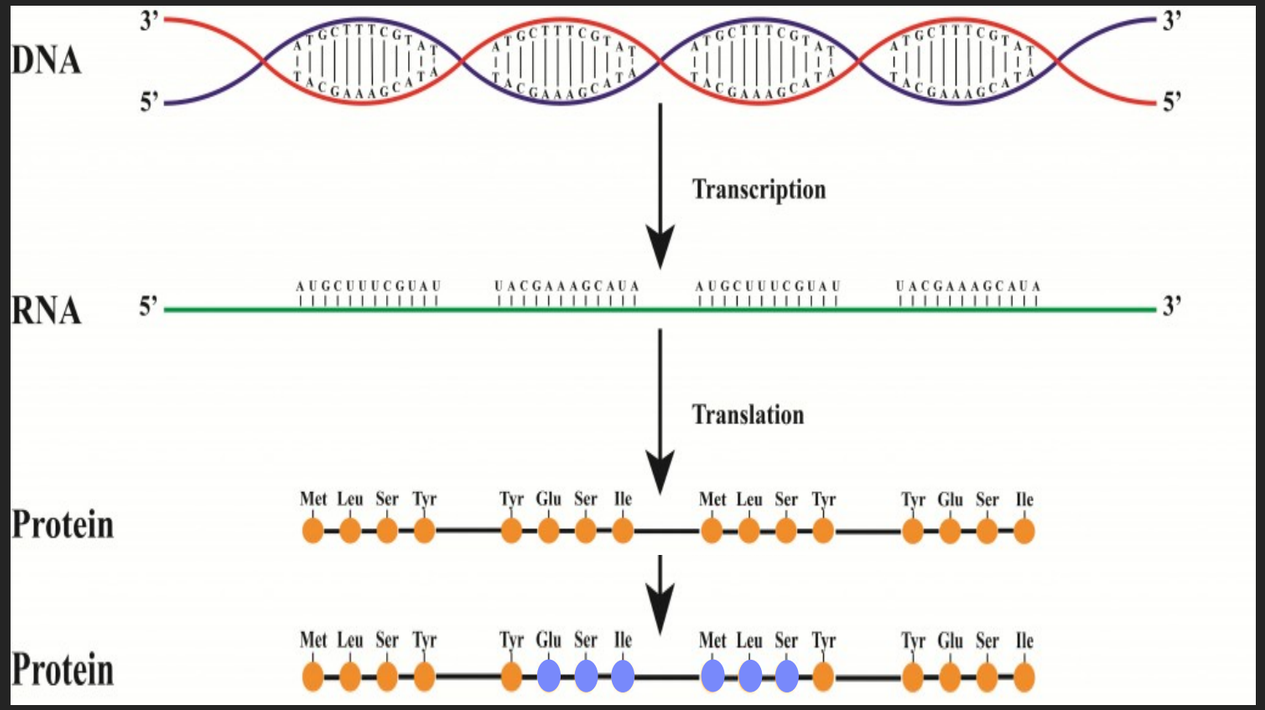
- Chemists
- Computer Scientists
- Mathematicians
- Microbiologists
- Physicists
- …

This work represents the efforts of a multi-disciplinary team of scientists at **IBM Research** working in collaboration with Prof Noyes group at the **Univ. of Minn.** And Yi Yan Yangs' group at the ***Institute of Bioengineering and Nanotechnology, Singapore***

Combating Antibiotic Resistant Bacteria Requires
Linking Genotype to Phenotype

Patient Diagnostics: Identify the best antibiotics to treat patient infection

Agriculture Biotechnology: Characterize soil and animal health



Microbiome: Compare a human microbiome to microbial reference data to diagnose cause of illness

Food Safety: Detect and assess the risk and impact of bacteria on goods for CPGs

These protein domains

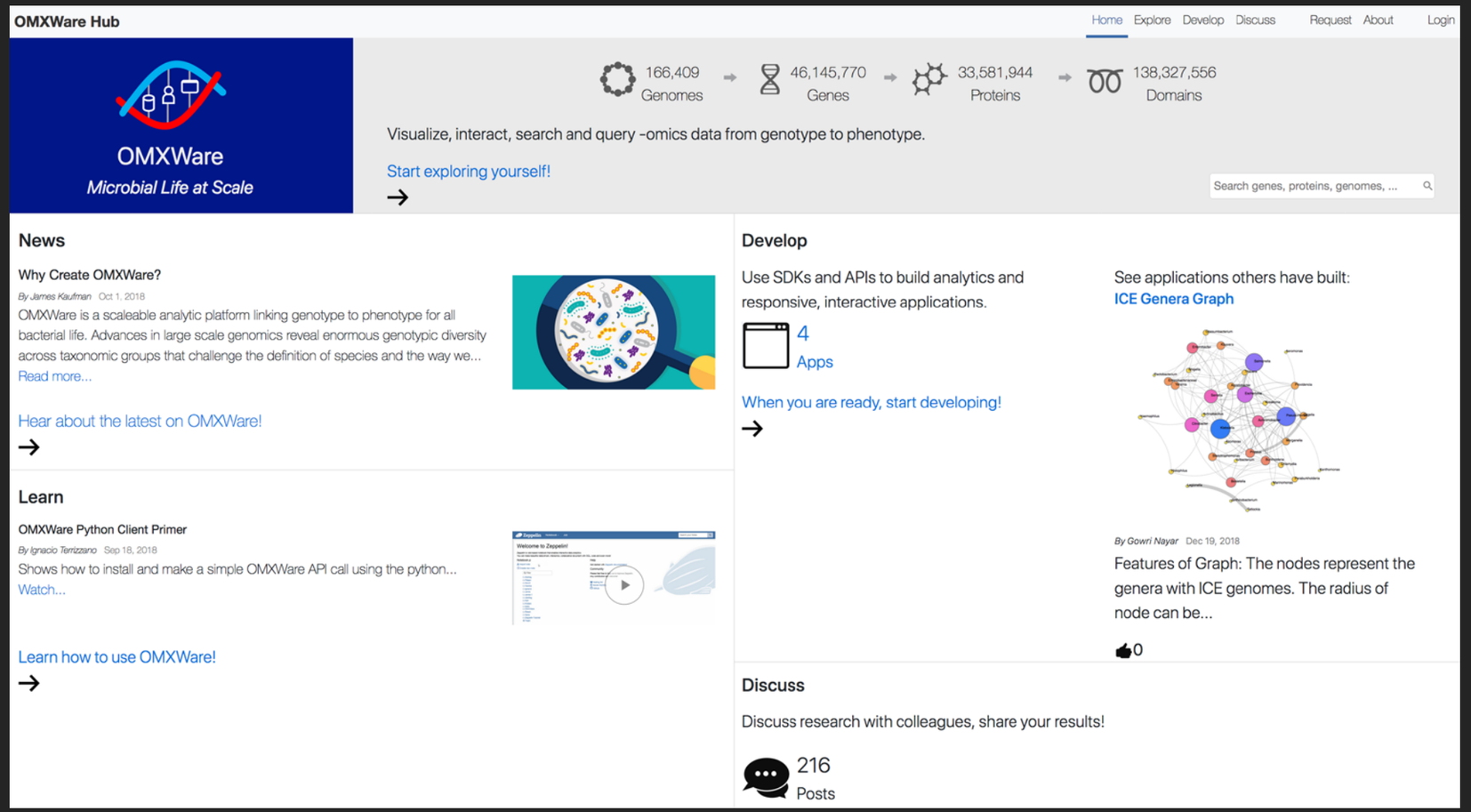
- *define phenotype of patient illness*
- *are the targets of drugs*
- *determine correct treatment*

IBM Research

Background: Understanding antibiotic resistance requires understanding phenotype not just genotype. Phenotype represents what an organisms can do. Genotype is defined by genome sequence. Most of us have heard of the so called “Central Dogma of Molecular Biology”, that DNA codes for RNA and RNA codes for Protein. But there is an important detail not as widely understood. The genes (and the proteins they encode) are not but fundamental objects. Within each protein are “Protein Domains”. These substrings of the protein are the active regions that evolve, function. Protein domains exist independently of the rest of the **protein** chain on which you find them (the same domain is often found on different genes with different names). Protein domains are fundamental objects of life. The field of proteomics has annotated many of them and assigned to them standardized codes representing molecular functions and pathways.

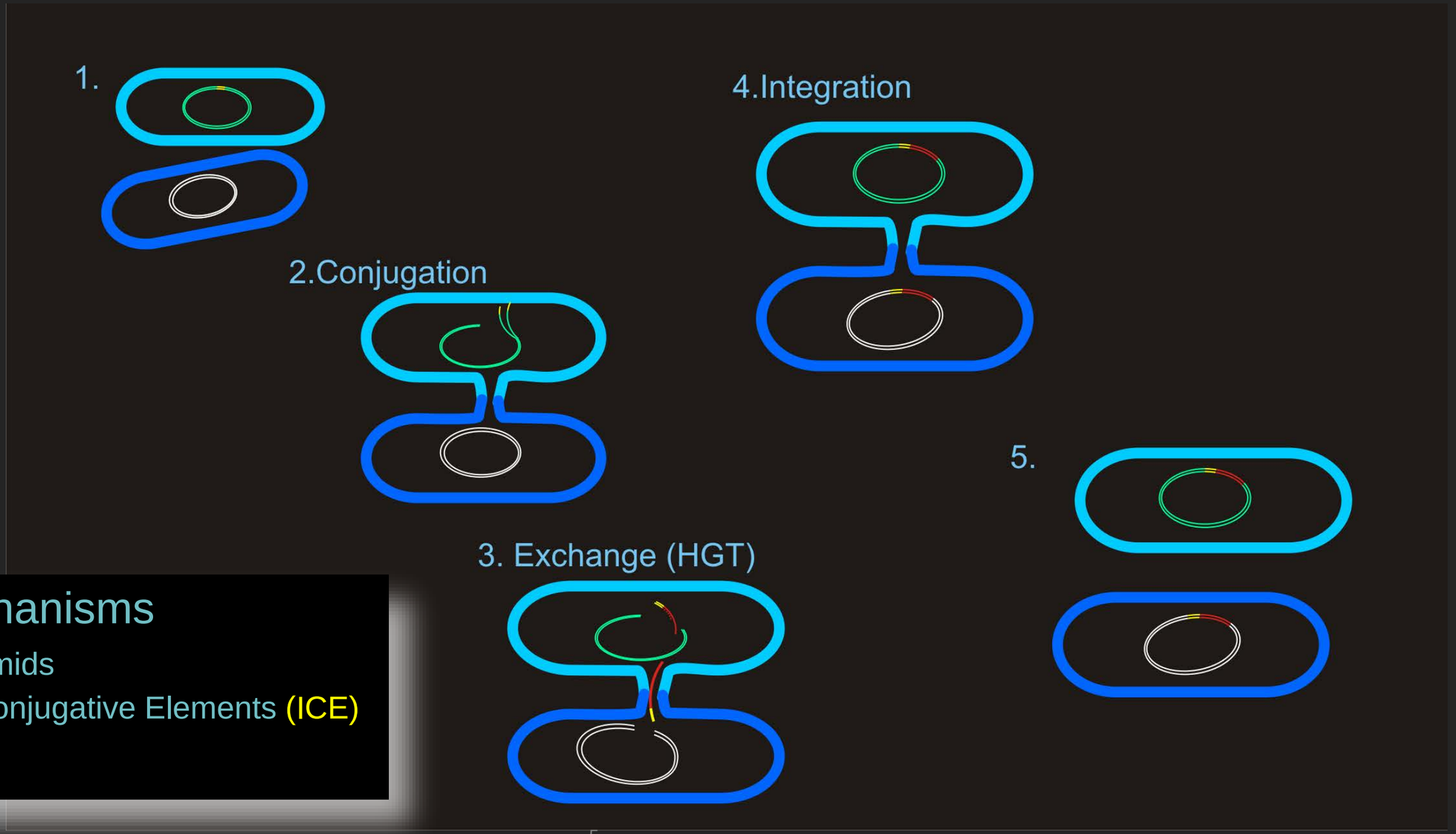
OMX Ware

- Public Data!
  - Not all of it accurate
  - Some lack info on Phenotype
  - Not queryable
- Used the cloud to assemble all bacterial genomes in the SRA
- Created OMX Ware
- With a proper database, asking important biological questions becomes *a simple query*
- Example: *Mobile Genetic Elements*



- **Public resources including NCBI are important – vital - national assets!!**
- *However*
    - Not all public genomes are high quality
    - Many meta-data errors
    - Most of the public genomes lack data on phenotype (eg resistance)
    - Today the data is not in a queriable big data form
      (i.e. not in a real database)
- IBM Research invested a few million cloud compute hours to
    - assemble all bacterial genomes in SRA
    - pick the highest quality (about 170,000 of them)
    - annotate them (finding all the genes, proteins, protein domains, etc.)
    - put them in a real database linking **genomes <=> genes <=> proteins <=> domains <=> function**
- With a proper database, asking important biological questions becomes *a simple query*
- Example: *Mobile Genetic Elements*

Integrative Conjugative Elements: an important AMR transmission mechanism

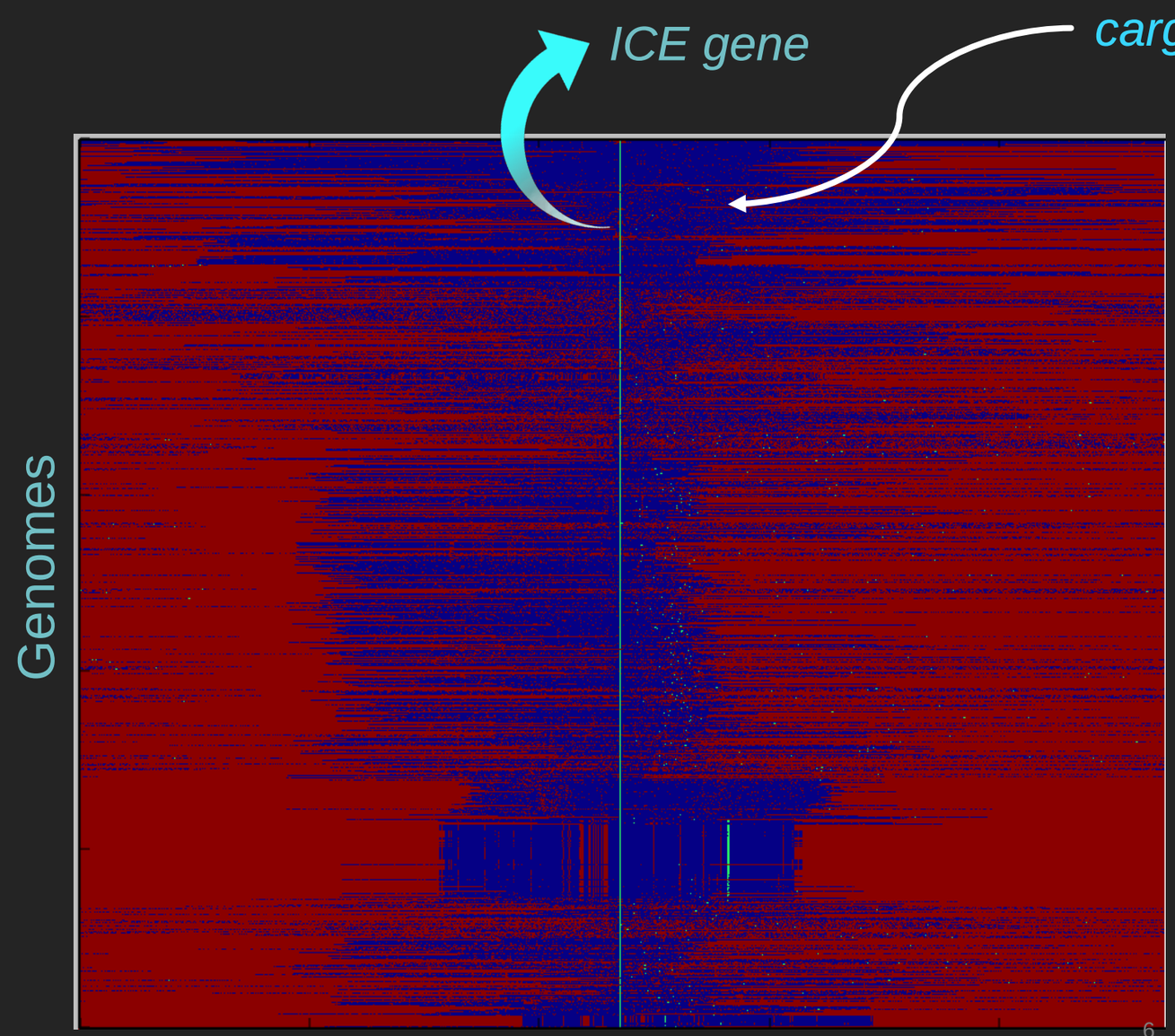


- Multiple Mechanisms
  - HGT by plasmids
  - Integrative Conjugative Elements (ICE)

This slide shows horizontal gene transfer by conjugative transposons, in particular **Integrative Conjugative Elements (ICE)**

Under stress, some bacteria will conjugate. If their genome contains an ICE genes, they may copy and transfer the ICE genes along with a large number of cargo genes. These are all integrated into the chromosome of the receiving bacteria (which may be from an entirely different taxonomic group).

Finding ICE genomes and cargo genes



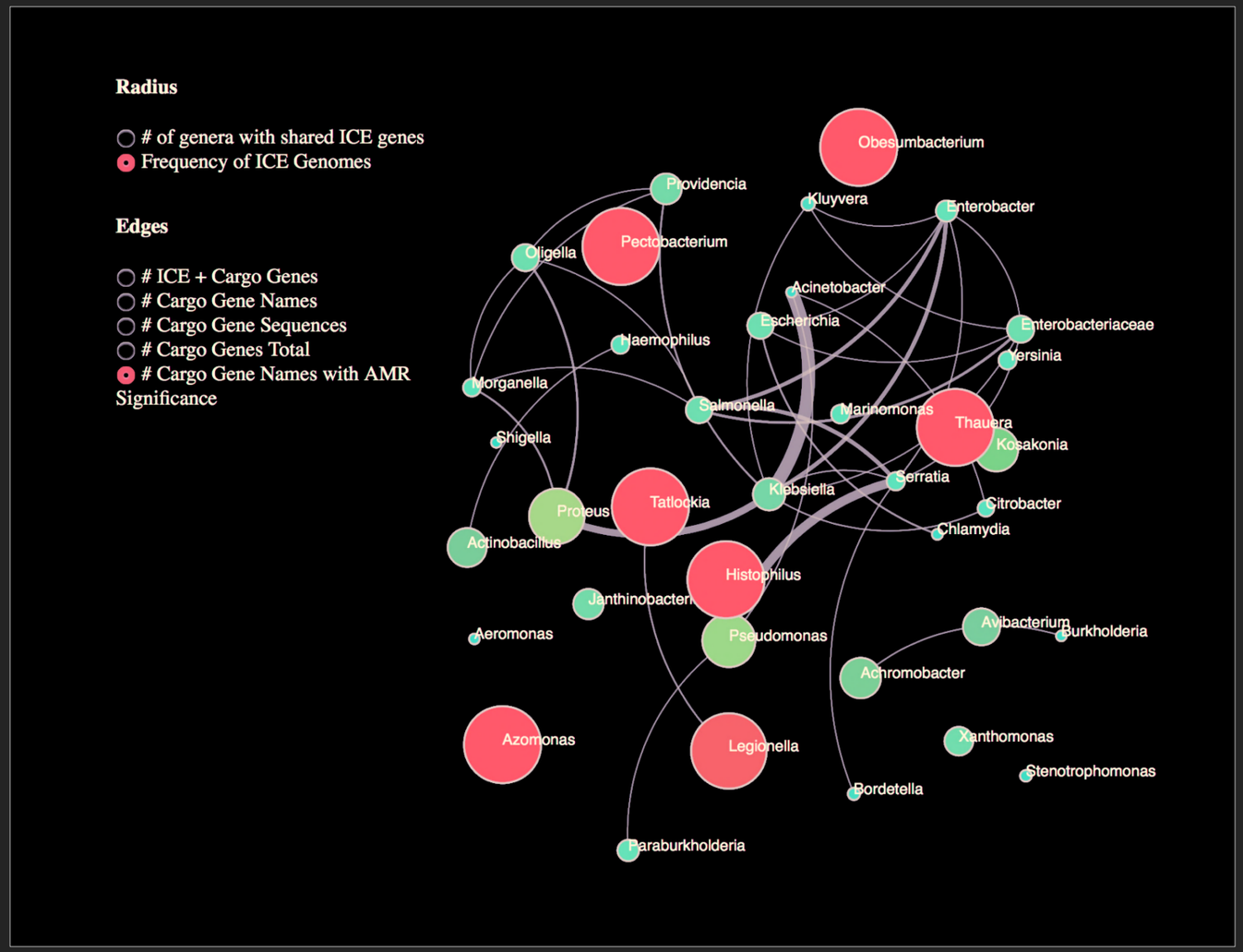
Discovering ICE and cargo genes in 170,000 Public Genomes

Description	Unique Sequences aa = amino acid, nt = nucleotide
ICE protein domains	1,750 aa sequences
ICE proteins	1,025 aa sequences
Genomes containing an ICE protein	17,176 genomes
All genes in ICE genomes	5,033,636 nt sequences
ICE proteins found in 2 or more genomes	517 aa sequences
ICE genes found in 2 or more genomes	554 nt sequences
Distinct cargo genes in 2 or more ICE genomes	1,152,267 nt sequences

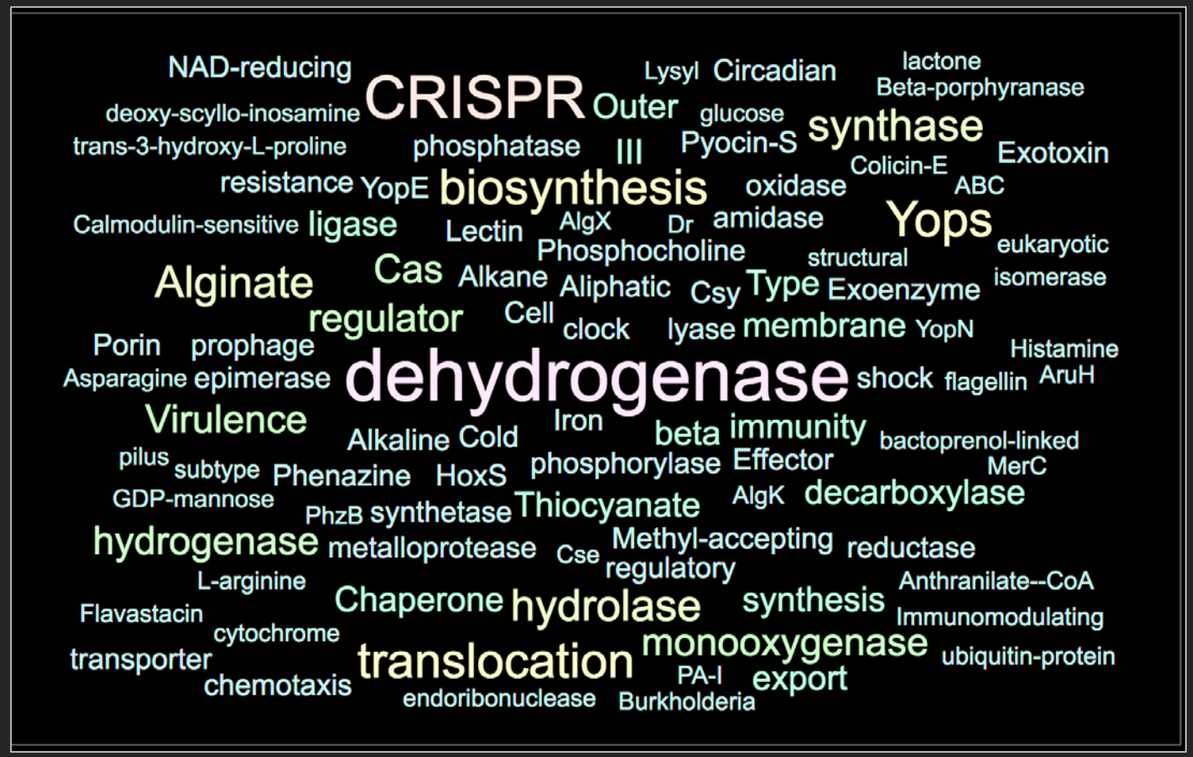
Using our big data approach, finding the all the genomes containing ICE genes and associated cargo genes becomes a simple database query. The figure shows the 1000 genomes with the largest number of cargo genes. All ICE proteins are shown in white. Cargo genes are blue and other genes red. The data for each genome was rotated (bit shifted) left to center the first ICE genes which, therefore, appears as a white vertical line. Some genomes contained more than one ICE gene. These appear as individual scattered white points to the right of the central white line (since the genomes are rotated left). The shorter white line segment observed in the lower right hand part of the figure derives from a set of 101 genomes that are all from the same NCBI BioProject (PRJEB12239). According to NCBI metadata, all of these genomes were samples of Legionella pneumophila obtained from a single site.

The cargo genes

where they are, where they go, and what they do



- *Transmission between genera*
- *Includes known AMR genes*
- *Also includes generic stress response genes that contribute to resistance !*

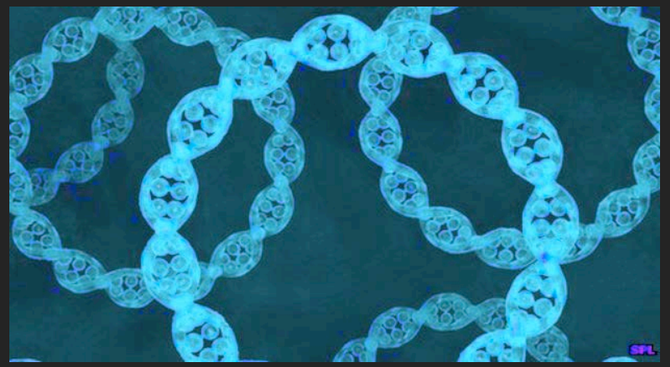


Having discovered all of the ICE genomes and all of the cargo genes we can used the database to learn what the cargo genes do – what phenotoypes they transmit. In the examples above we show each Genus as a circle with diameter based on the frequency of ICE genes. The connections between genera in the figure on the left represent the frequency of AMR gene transmission (for known AMR genes). The figure on the right is a word cloud based on the names of all cargo genes – with dehydrogenase and CRISPR as the biggest words. It is clear from the word cloud the importance of conjugative exchange as a mechanism for organisms to quickly acquire new phenotypes in response to stress.

Combating Antibiotic Resistance

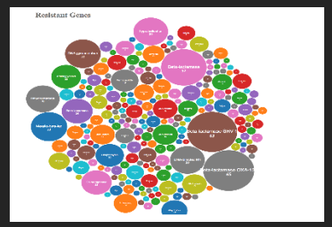
requires understanding phenotype!

Resistant Genome



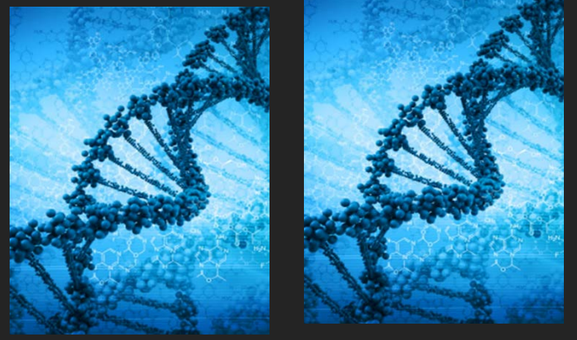
OMX Ware

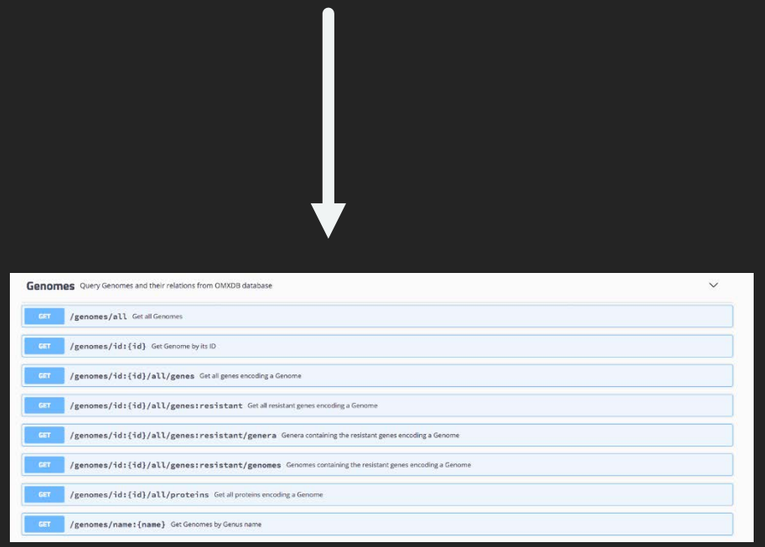
OMXWARE
Genomes
Genes



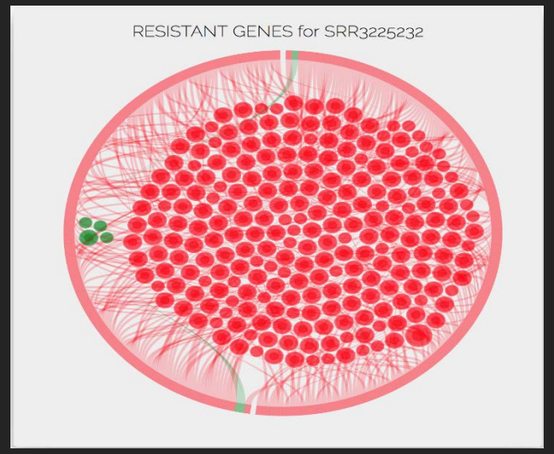
PUBLIC
RESISTANCE
DATA

RESISTANT GENES

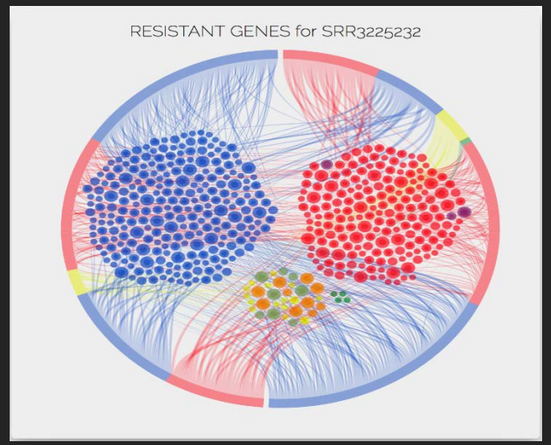




When do resistant genes lead to resistant organisms?



Some genes always
lead to resistance



Some genes sometimes lead
to resistance

IBM Research

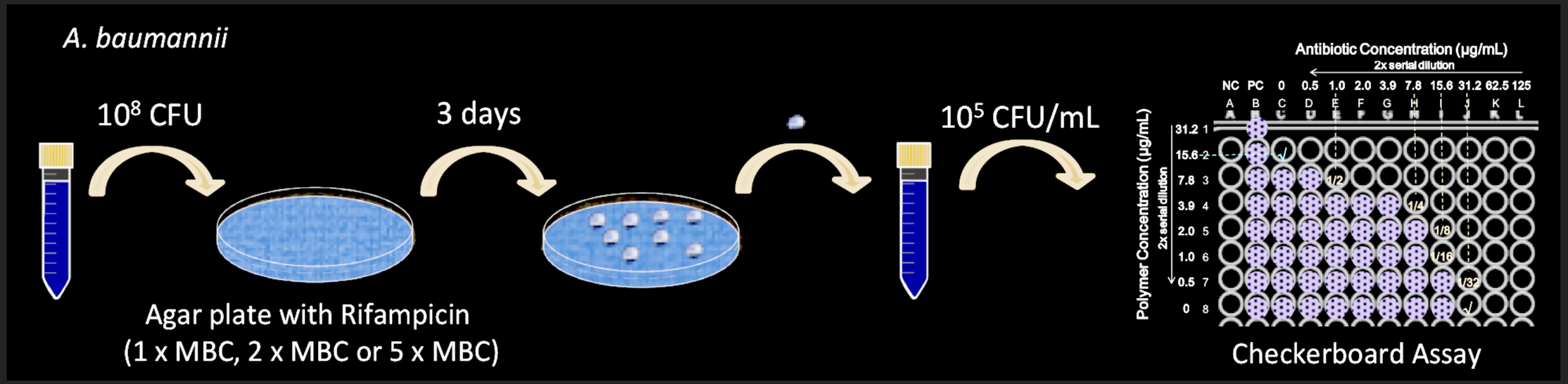
Discovering single genes (or gene names) does not always predict phenotype. The figure on the lower left shows AMR genes (in the outer ring) along with the genomes that contain those genes (as circles within the ring). Red indicated confirmed antibiotic resistance. Some AMR genes always lead to resistance while the presence of others may be necessary but not sufficient.

In Order to relate Genotype to Phenotype,

- Need for controlled experiments
- Machine learning to identify genes that respond to antibiotic pressure
- Some of these genes have high sequence homology to non-amr genes!!

Diagram on slide showing how the resistant genome is run through OMXWare, that the isolates the resistant genes in the database. The database then points to the decision mechanism of when resistant genes lead to resistant organisms.

Resistance evolves in a culture

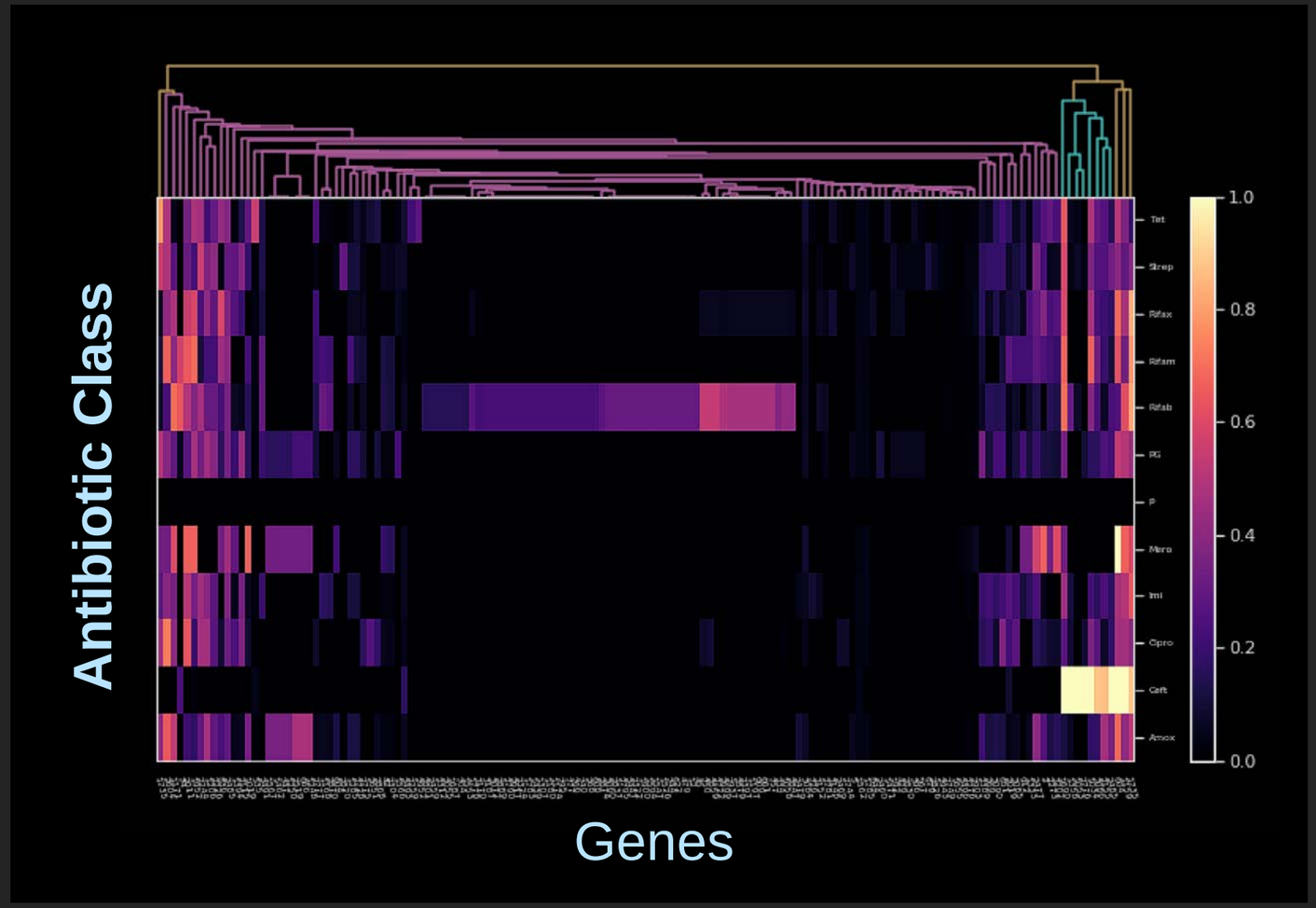


Yi Yan Yang et al., Singapore Institute of Bioengineering and Nanotechnology,

Scientists are now doing experiments where organisms are cultured with antibiotics and their genomes sequenced before and after evolution of AMR. This will reveal which point mutations contribute to evolution of resistance.

Machine Learning identifies genes that evolve under antibiotic pressure

- some changes are specific to the antibiotic mechanism
- some are general stress response genes!



IBM Research

In the figure above, each column is a gene within the cultured organism before evolution of resistance. Each row represents an average over genomes cultured with different antibiotics. The color represents how much the gene changed from the reference genome. Some genes change in response to several different antibiotic classes. These are generic stress response genes – but they play an important role in AMR. Other genes respond to specific antibiotics. Specific AMR genes often have high sequence homology to genes with different names and molecular functions not related to AMR.

Conclusions

- AMR should be understood in the context of stress response
  - specific genes target antibiotic mechanism(s)
  - generic response genes that help organisms survive
- Resistance Genes Transmit in response to antibiotic stress
  - across genera
  - between microbiomes
  - up and down the food chain
- Combating antibiotic resistant bacteria requires
  - Big Data (publicly available from resources like NCBI)
  - machine learning and AI tools to link genotype to phenotype
  - controlled experiments to
    - ✓ establish resistance ground truth
    - ✓ measure transmission within and between food chain microbiomes