Detection of Mobile Genetic Elements (MGEs) in Bacterial Genomes

PhD student: Zheng WANG

Supervisor: Professor Margaret IP

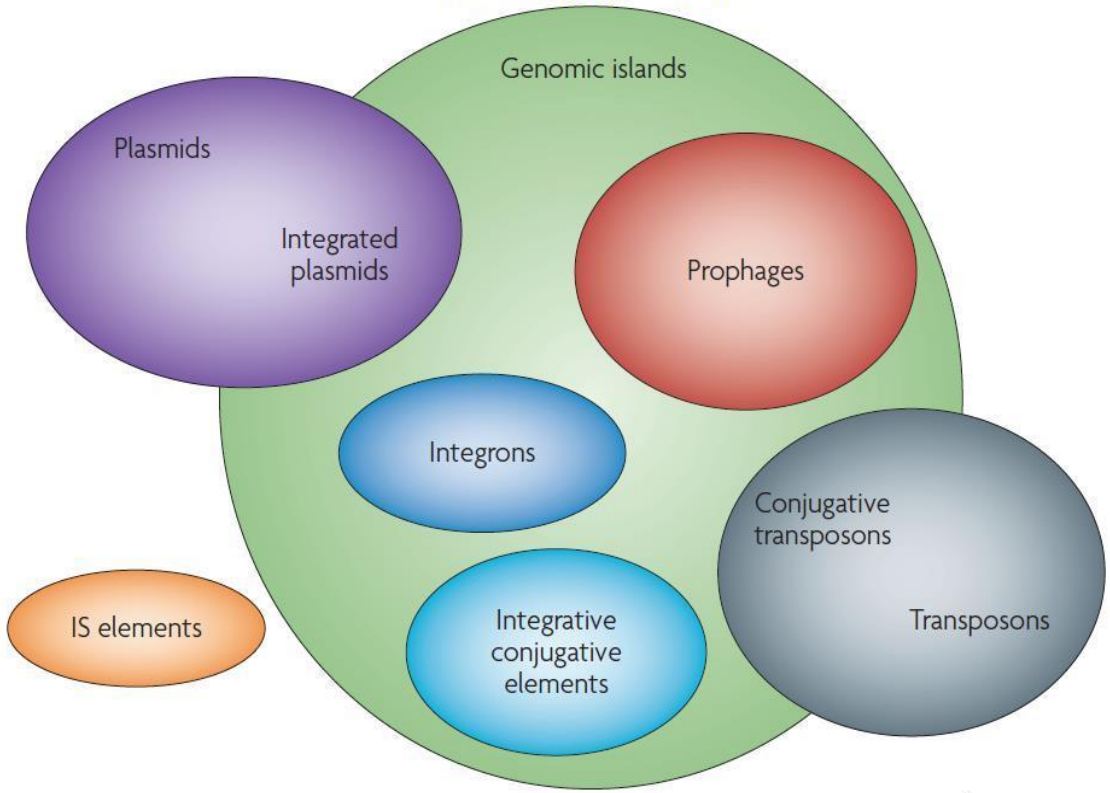
Department of Microbiology, CUHK

Date: 3rd Dec, 2013

Contents

Introduction

Mobile genetic elements



Mobile genetic element

Any sequence of DNA that is physically moved within an organism genome or between different organisms.

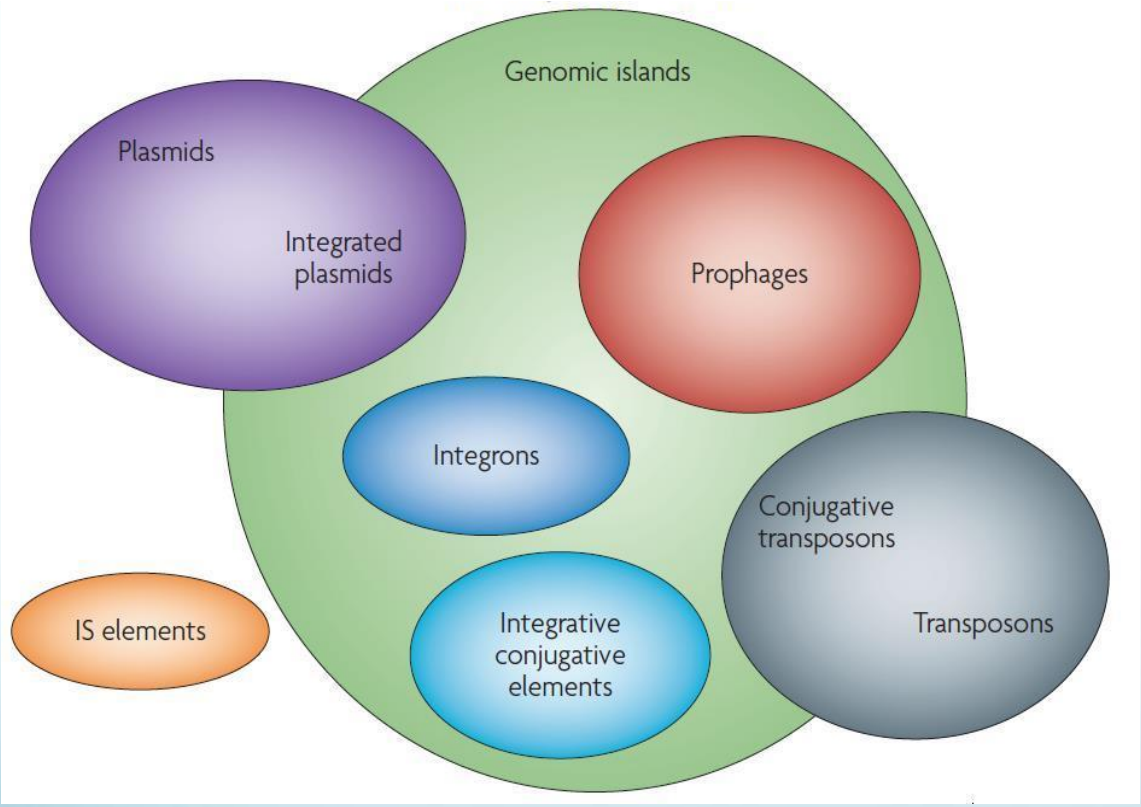
10% - 20% of the Bacterial genome consists of MGEs

Horizontal gene transfer

Transfer of genetic material from one organism to another organism that is not its offspring

Introduction

Mobile genetic elements



Genomic island

In a bacterial genome, a cluster of genes for which there is evidence of horizontal origins.

- **Prophage**
- **Integron**
- **Integrative conjugative element**
- **Conjugative transposon**
- **Integrated plasmids**

Introduction

Importance (X 4)

- 1. Frequently associated with microbial adaptations that are of medical and environmental (or industrial) interest;

  Metal resistance

  Antimicrobial resistance

  Secondary Metabolic properties

- 2. Known virulence factors are over-represented in GIs. The selective loss and regain of GIs could provide an additional means to modulate pathogenicity

Introduction

Importance (X 4)

- 3. The spontaneous excision of PAIs has been observed in various pathogens ;
  results in distinct pathogenic phenotypes

- 4.Had a substantial impact on bacterial evolution.

Methodology

The **Bioinformatics Approaches** for predicting MGEs (especially GIs) with genome sequencing data fall into two broad categories:

- **Sequence composition**
- *SIGI-HMM* (Hidden Markov Model)
- *PAI-IDA.*
- *Centroid.*
- *Alien_Hunter.*
- *PredictBias*.
- *PHAST*

- **Comparative genomics.**
- *IslandPick*
- *MobilomeFINDER*
- *Whole genome alignment*

In fact, there are also some **wet-lab methods** to detect MGEs. However, here we just focus on the above well -developed bioinformatic methods.

Methodology

All of the above methods are based on whole genome sequencing data ;

Most of the methods are designed base on GIs sequence and structural Features.

- ***Sporadic distribution***
  only found in some isolates of a given specie;
  gene phyletic patterns different with host genome;
- ***Sequence composition bias***
  oligonucleotides of various lengths ;
  GC content; (Traditional Methods)
- ***Large size (>8 kb)***
- ***Mobility, phage and virulence genes***
  Over-representation of certain classes of genes and unknown function genes
- ***Neighbouring tRNA genes ; direct repeats***

Methods associated with different Features

Feature	Methods for detection	Benefits and pitfalls when used for GI prediction
Sporadic distribution, instability and an ability to excise spontaneously	Comparative genomics to identify unique (versus shared) genomic regions	Multiple closely related sequenced genomes are required for comparison
Sequence composition bias	Various methods	False-positive results are obtained owing to a bias in highly expressed genes, and false-negative results are obtained owing to the sequence composition being similar to that of the host genome (which is sometimes the result of amelioration)
Size (usually > 8 kb)	Comparative genomics to identify large insertions or features such as sequence composition bias in a region over a certain length	Large horizontally acquired regions are easier to predict than regions containing a single gene
Adjacent to a tRNA gene	Detection of full or partial tRNA genes using BLAST or tRNAscan-SE	Many GIs are not inserted in or near tRNA genes
Flanked by direct repeats	Use of repeat finders such as REPuter	Not all GIs are flanked by direct repeats, and the identification of relevant repeats can be difficult owing to their small size
Over-representation of certain classes of genes such as mobility genes, genes encoding virulence factors, phage-related genes and genes encoding proteins of unknown function	Use of existing genome annotations or searching for similarity to functional databases such as COG or PFAM	Can be used as supporting evidence for GI prediction, and can allow further subclassification of GIs into other MGEs such as prophages or integrated plasmids; but some GIs might have lost all mobility genes, or these genes can be missed because they are not identified by the particular search used

Overview of genomic island prediction programs

Program	Description	Accuracy* and limitations
SIGI-HMM	Measures the codon adaptation index and removes ribosomal regions	• Precision: 92% • Recall: 33% • Accuracy: 86% • The most precise and most accurate program, along with IslandPath-DIMOB
PAI-IDA	Measures percentage GC content and dinucleotide and codon usage	• Precision: 68% • Recall: 32% • Accuracy: 84%
Centroid	Allows various options, but pentamers are the default	• Precision: 61% • Recall: 28% • Accuracy: 82%
Alien_Hunter	Uses variable-length *k*-mers	• Precision: 38% • Recall: 77% • Accuracy: 71% • The program with the highest recall, but at the expense of precision
PredictBias	Measures percentage GC content and dinucleotide and codon bias, and predicts PAIs using similarity to a database of virulence genes	• Accuracy measurements could not be calculated, as the entire dataset was not available for download
IslandPick	Automatically 'picks' default comparison genomes for use in whole-genome alignments	• The highest agreement with a data set of GIs that have been reported in the literature • Requires related genomes for use
MobilomeFINDER[42]	Uses tRNA gene locations and whole-genome alignments to identify GIs	• Limited to only identifying GIs in tRNA genes • Comparison genomes cannot be automatically selected

Applications

Application example 1

Identification and characterization of ϕH111-1: A novel myovirus with broad activity against clinical isolates of *Burkholderia cenocepacia*. **(Lynch, K. H.,et al,2013)**

- **Prophage identification (One of the most important GIs)**
  Using the PHAST method (prophage-finding program Phage Search Tool) to identify prophages in the B. cenocepacia strain H111 genome sequence
- **Confirmation of the characterization with laboratory experiments**

Applications

Methods Selection

Target Genome status:

B. cenocepacia strain H111 only have Draft Genome (gaps unclosed)

PHAST

This program accepts either raw reas data or contigs data, however, like all the other GI predict programs, to get a better result, complete genome data are recommended .

Input: 71 available H111 contigs.

Applications

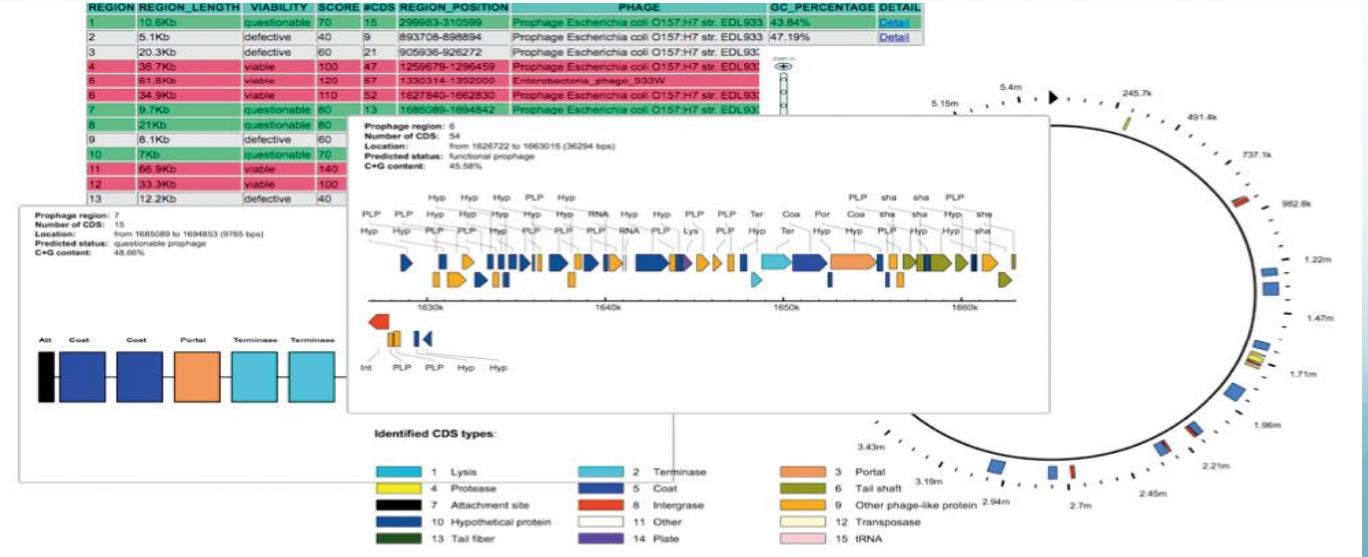
PHAST procedures

- Genome-scale ORF prediction/translation (by GLIMMER)
- Protein identification (by BLAST matching ; annotation by homology)
- Phage sequence identification (byBLAST matching to a phage-specific database)
- tRNA identification
- Attachment site recognition ;
- Gene clustering density measurements (using density-based spatial clustering; DBSCAN)
- Evaluates the completeness of the prophage (give a Score)

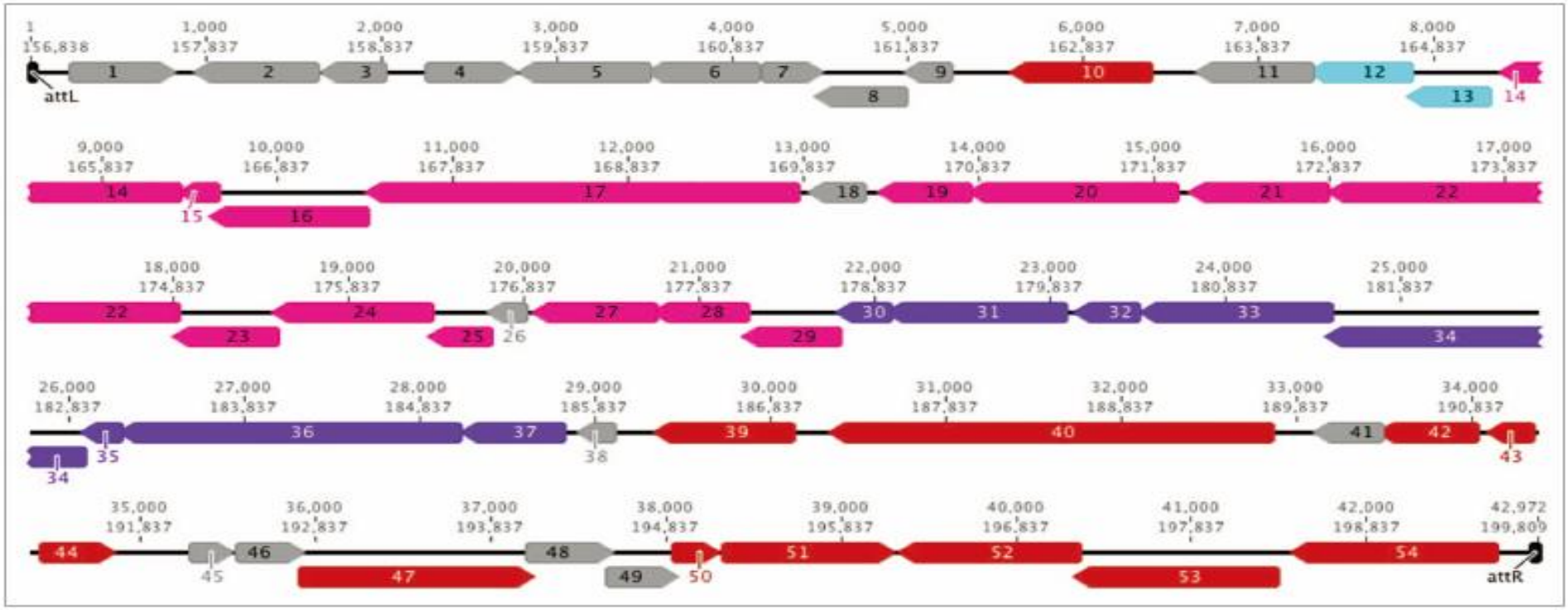
Applications

PHAST Results

- **GC_PERCENTAGE; COMPLETENESS**: (intact or incomplete, according to **SCORE**); **REGION_LENGTH and POSITION; CDS** ;
- In this case, This program identified potential intact prophages (Score >120 ; total score 150) in contig 43 ;
- GC content 62% (lower than the H111 GC content of 67%) ;



Results



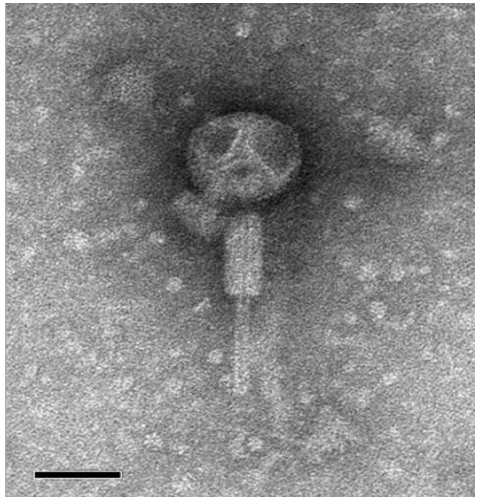
Map of the φH111-1 prophage; the position in the C43 and the CDS;

No putative toxin genes were identified.

Confirmation with laboratory experiments

- Transmission electron microscope analysis
- Phage isolation and analysis
- shotgun cloning;

(Lynch, K. H.,et al,2013)



Applications

Application example 2

Insight into the specific virulence related genes and toxin-antitoxin virulent pathogenicity islands in swine streptococcosis pathogen *Streptococcus equi ssp. zooepidemicus* strain ATCC35246

(Ma, Z. et al,2013)

- **Identification of GIs by Comparative genomics and Sequence composition related methods**

Applications

Target strain: *S. zooepidemicus* strain ATCC35246

NGS: Complete Genome ; 454 Platform.

Comparative Genomics

- 3 Reference genomes : *S. zooepidemicus* MGCS10565 and H70
  S. equi 4047. (All Complete Genomes)
- identify clusters of genes in target genome that are not present (or scattered)in closely related other 3 Reference genomes
- identify important mobility genes, such as integrases, transposases were present at the boundaries of the region
- GC content (different with the average of whole genome)

Confirm with IslandViewer

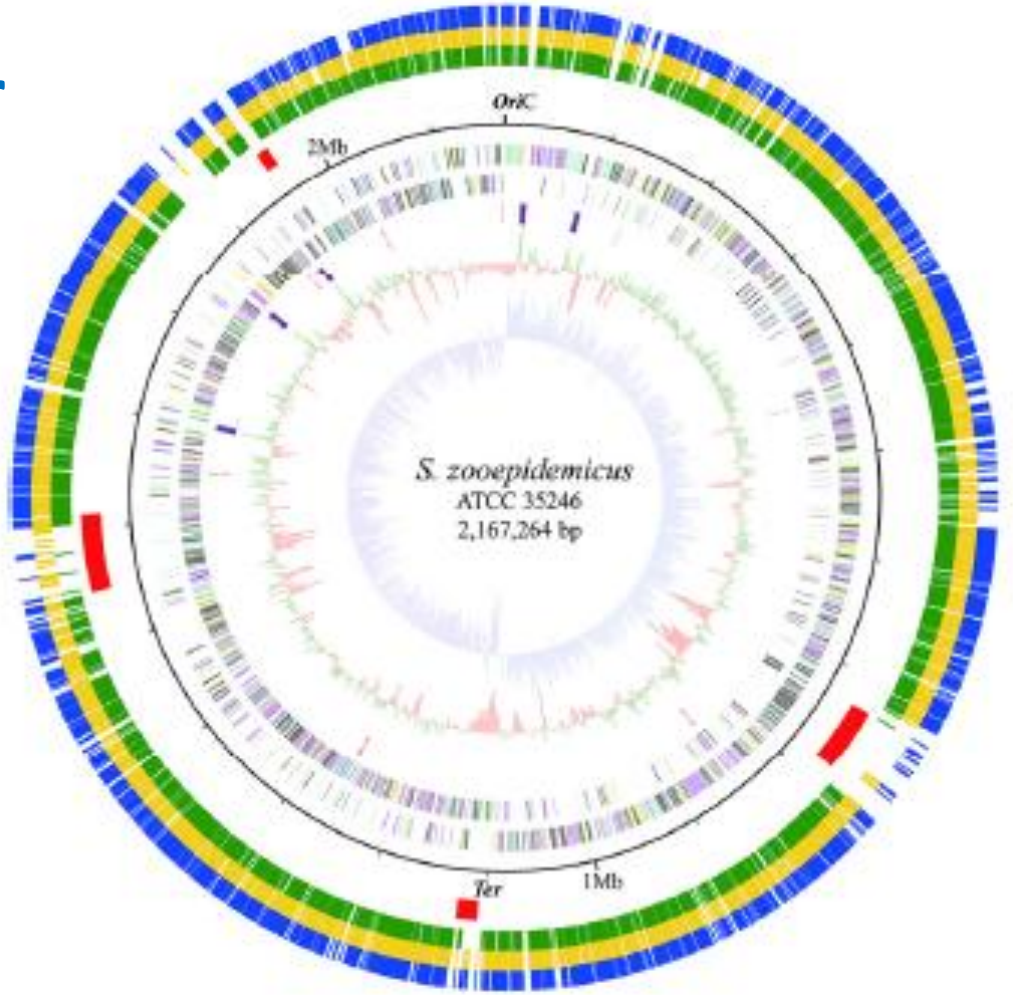
An genomic island predictor that integrates 3methods:

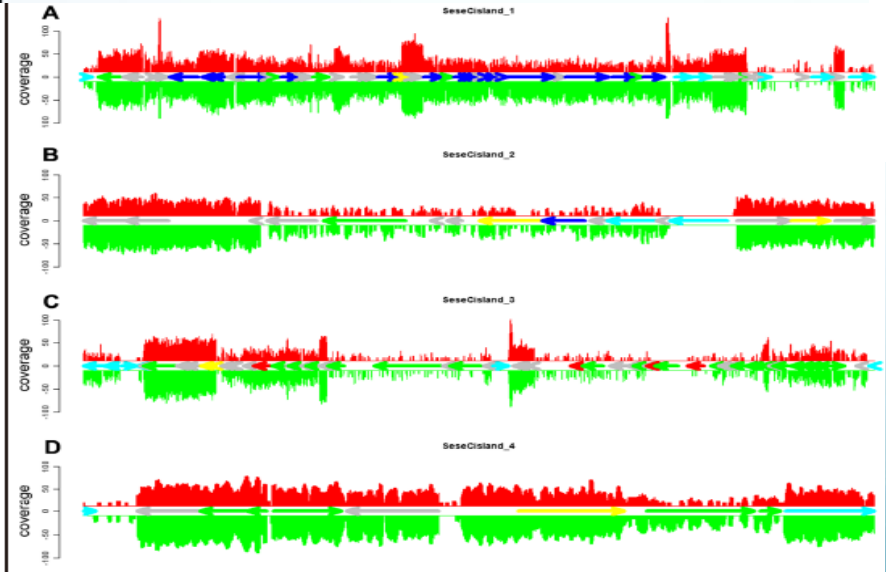
IslandPick,

IslandPath-DIMOB,

SIGI-HMM

GIs which identified by at least 2 methods were marked.





Total 4 GIs associated with pathogenicity and virulence were confirmed

(Ma, Z. et al,2013)

Future Improvements

- **Difficulties :How to Handle un-assembled millions of raw reads .**

  An increasing proportion of microbial genome sequences are the result of unfinished/unclosed genome sequences
  Shorter reads might not provide enough signals for sequence composition.

- **Trends :The integration of the strengths of previously developed methods coupled with increased genomic database of bacteria and phages.**

References

- 1. Dhillon, B. K., Chiu, T. A., Laird, M. R., Langille, M. G. & Brinkman, F. S. IslandViewer update: Improved genomic island discovery and visualization. *Nucleic Acids Res.* 41, W129-32 (2013).
- 2. Lynch, K. H., Liang, Y., Eberl, L., Wishart, D. S. & Dennis, J. J. Identification and characterization of varphiH111-1: A novel myovirus with broad activity against clinical isolates of. *Bacteriophage* 3, e26649 (2013).
- 3. Ma, Z. *et al*. Insight into the specific virulence related genes and toxin-antitoxin virulent pathogenicity islands in swine streptococcosis pathogen Streptococcus equi ssp. zooepidemicus strain ATCC35246. *BMC Genomics* 14, 377-2164-14-377 (2013).
- 4. Zhou, Y., Liang, Y., Lynch, K. H., Dennis, J. J. & Wishart, D. S. PHAST: a fast phage search tool. *Nucleic Acids Res.* 39, W347-52 (2011).
- 5. Langille, M. G., Hsiao, W. W. & Brinkman, F. S. Detecting genomic islands using bioinformatics approaches. *Nat. Rev. Microbiol.* 8, 373-382 (2010).
- 6. Boyd, E. F., Almagro-Moreno, S. & Parent, M. A. Genomic islands are dynamic, ancient integrative elements in bacterial evolution. *Trends Microbiol.* 17, 47–53 (2009).
- 7. Winstanley, C. *et al.* Newly introduced genomic prophage islands are critical determinants of *in vivo* competitiveness in the Liverpool Epidemic Strain of *Pseudomonas aeruginosa*. *Genome Res.* 19, 12–23 (2008).
- 8. Langille, M. G. & Brinkman, F. S. IslandViewer: an integrated interface for computational identification and visualization of genomic islands. *Bioinformatics* 25, 664–665 (2009).
- 9.Chen, J. & Novick, R. P. Phage-mediated intergeneric transfer of toxin genes. *Science* 323, 139–141 (2009).

Thank You !