# A profile-based method for identifying functional divergence of orthologous genes in bacterial genomes

#### Paul Gardner

University of Canterbury Christchurch New Zealand

## In this talk...

I'll discuss a ncRNA project that went wrong...

Estimating functionally significant genetic variation



# Can we identify genetic variations more likely to be causative of phenotypic change?

 A chance meeting with Rob Kingsley: ~ 800 SNPs seperate a host restricted pigeon pathogen from a broad host-range gastroenteritis-causing pathogen.



#### Profile HMMs: a powerful homology search tool



Image provided by Sean Eddy.

# Profile HMM background

 As a general rule, profile based methods are more accurate (and faster) than sequence based methods

 I2
 3 4 5 6 7 8 9 1011

 yeast
 GUCUUCGGCAC

 fly
 GCCUUCGGAGC

 cow
 GCUUUCGUCGC

 mouse
 GCUUUCGAUGC

 human
 GCGUUCGCUGC

 chicken
 GUAUUCGUAAC

 snake
 GUGUUCGCGAC

 croc
 GUUUUCGAGAC

One HMM node per alignment column

- 3 states per node: (M) Match: emits residues
- (I) Insert: inserts extra residues
- (D) Delete: deletes residues

HMMs generate homologous sequences.





#### Refactoring an accurate homology search tool



 $bitscore = \log_2\left(\frac{P(seq|model)}{P(seq|null)}\right)$ 

# Species 2 FYFIIRKEMQLRAE

#### Accuracy of $\Delta bitscore$



# Comparing proteomes



Kingsley et al. (2013) Genome and Transcriptome Adaptation Accompanying Emergence of the Definitive Type 2 Host-Restricted Salmonella enterica Serovar Typhimurium Pathovar. mBio

#### Comparing proteomes: generalists vs extraintestinal

- S. enterica serovars divided into 2 groups:
  - Gastrointestinal and extraintestinal (or invasive)
- Characterised hypothetically disrupted coding sequences (HDCs)
- We find hypothetically attenuated coding sequences (HACs)



Nuccio & Bäumler (2014) Comparative Analysis of Salmonella Genomes Identifies a Metabolic Network for Escalating Growth in the Inflamed Gut. *mBio.* 

#### Signatures of adaptation: generalists vs extraintestinal

\*

\*

\*

\*

\*

200

А





#### Signatures of adaptation: Enteritidis vs Gallinarum



#### Signatures of adaptation: generalists vs host-restricted





outer membrane usher protein FimD precursor hypothetical zinc-dependant alcohol dehydrogenase methyl-accepting chemotaxis protein chaperone heat shock protein copper-transporting ATPase biotin sulfoxide reductase periplasmic oligopeptide-binding proteinprecursor (OppA) periplasmic beta-glucosidase precursor protein DinP (DNA damage-inducible protein) putative ABC-transporter ATP-binding protein N-methyl-L-tryptophan oxidase putative RND-family transporter protein putative efflux pump aromatic-amino-acid aminotransferase conserved hypothetical protein putative membrane protein possible monooxygenase putative hydroxypyruvate isomerase Cobvrinic acid A.C-diamide synthase. putative cation efflux pump citrate carrier deoxyribodipyrimidine photolyase putative ribosomal large subunit pseudouridinesynthase B putative hydrolase ribosomal-protein-alanine acetvltransferase putative esterase dipeptidyl carboxypeptidase II

#### Gastrointestinal

#### Extraintestinal

#### Potential application: Using delta-bitscore for the classification of organisms



#### Works on ncRNAs too!

Protein DBS







DBS

- Developed a new, simple approach for determining the significance of variation
- Noise cancels out! Consistent LoF/GoF variation amplifies
- Working well in "the field"
- More sensitive than selection measures (e.g. dN/dS) for between strain comparisons
- Developing the approach for testing ncRNAs and conserved DNA elements

Wheeler, Barquist, Kingsley & Gardner (2016) A profile-based method for identifying functional divergence of orthologous genes in bacterial genomes. *bioRxiv*.

## Thanks

- Nicole Wheeler
- Lars Barquist
- Rob Kingsley

- Honour McCann
- and many more...



Funded by: Rutherford Discovery Fellowship, BPRC and Biological Heritage: National Science Challenge.