MICROBIAL GENOMICS
Bases to Biology

6 Month Review

Microbial Genomics: The First Six Months

Microbial Genomics (MGen) was created to be the go-to journal for microbial genomics research, and the 24 papers published in the journal's first six months provides a great snapshot into where the field is at in January 2016.

One third of the articles are cutting edge genomic epidemiology in action, using genomics to investigate the evolution and transmission of a range of pathogens, from anthrax to dysentery and food poisoning.

Between them, these eight studies generated and analysed over 2000 genomes. Much of this comes from two very large studies: >1000 shigatoxigenic *E. coli* from a study by Public Health England (summarised by the MGen Editors in Chief here) and 450 *Streptococcus pneumoniae* from the Pneumococcal African Genomics Consortium.

In line with MGen's open data policy, all of these genomes are available in public sequence databases, and metadata and analysis results are available as supplementary data. However the sharing of microbial genome data for public health and diagnostic applications still poses challenges, which Sobia Raza and Leila Luheshi of the PGH Foundation explore in their article "Big data or bust - realising the microbial genomics revolution".

Interestingly all of these studies used similar methodologies, reflecting the maturity of this field of research.

Common approaches include: (i) sequencing large numbers of isolates using high-throughput Illumina platforms; (ii) the identification of SNPs (single nucleotide polymorphisms) using read mapping approaches (with *BWA*, *SMALT*, *SAMtools* and *GATK2* being popular tools); and (iii) uniform use of *RAxML* for generating maximum likelihood phylogenies.

Some also used *BEAST* to estimate mutation rates, divergence dates and phylogeographical patterns. Interestingly, half of these papers utilised MLST (multi-locus sequence typing) to identify clades, showing that this sub-genomic approach based on capillary sequencing of ~7 gene fragments is still considered useful by many genomic researchers.

Almost all of the genomic epidemiology studies took steps to remove SNPs introduced via recombination, in order to capture the underlying signals of vertical inheritance that are so important for transmission studies. Popular tools were *BratNextGen*, *Gubbins* and *ClonalFrameML*, which were all published within the last 3 years.

Methods used in genomic epi papers

Recombination and the pan genome shape microbial evolution

Of course, recombination and other forms of horizontal gene transfer are of great interest in their own right, and genomic studies offer unprecedented insight into the role of recombination and pan genome variation in microbial evolution.

Novel insights into the impact of recombination include:

• Pekka Marttinen *et al* introduce a model for the evolution of microbial populations that incorporates mutation, homologous recombination and horizontal gene transfer. They demonstrate that the model fits well to a real population sample of >600 *Streptococcus pneumonaiae* genomes and provide R code for constructing the model.
• van Haal *et al* explore recombination among 132 *Enterococcus faecium* isolated from a single hospital in Australia. Remarkably, they found that 70% of genes in the core genome were affected by recombination, and 80% of recombination events had a likely donor strain amongst sequenced hospital population.
• Laura Spoor *et al* show how that *S. aureus* ST71 has a hybrid genome that evolved through recombination, during which novel genes linked to immune evasion and matrix adherence were introduced, creating a highly pathogenic strain that has since become widespread.

Two papers exploring pan genome variation within *Yersinia* show how gene flow has contributed to the evolution of distinct species and sublineages of various pathogens within the genus, including the creation of ecologically separated subpopulations of the foodborne pathogen *Y. enterocolitica*, and the

differentiation of *Y. pestis* and *Y. pseudotuberculosis* which could provide novel diagnostic markers for distinguishing the two pathogens.

Velvet and *SPAdes* were the most popular tools for bacterial genome assembly, with *Prokka* and *Prodigal* for gene annotation, and LS-BSR and related approaches being commonly used to cluster orthologous groups of genes.

Antibiotic Resistance

One third of MGen articles have been concerned with antibiotic resistance, which is now recognised as a critically important global health threat. The problem is one of evolution in action: microbes exposed to antibiotics during treatment can mutate to become resistant to the drugs.

It is hoped that by understanding these evolutionary processes, we may be able to design better treatment strategies that pre-empt or prevent the emergence of resistance. To that end, two articles from the Stinear lab in Australia explore the evolution of multidrug resistance *in vivo* during persistent infections in individual patients – over four months of *S. aureus* infection and 21 years of infection with *Mycobacterium tuberculosis.*

Systems Microbiology

Systems Microbiology has been a very popular category, with several articles using functional genomics to investigate microbial gene regulation (in *E. coli*, *Acinetobacter baumannii* and a transcription factor family across species). In "Genome-wide analysis of the response to nitric oxide in uropathogenic *Escherichia coli* CFT073", Mehta *et al* used RNA-Seq to show that several virulence-associated genes are upregulated by nitric oxide (NO), and Chip-Seq to identify 49 binding sites for the NO-sensitive repressor NsrR.

Comparative genomics approaches were used in two articles exploring the evolution of particular gene families: Sheridan *et al* looked at carbohydrate active enzymes (CAZymes) in gut dwelling bacteria, and Booher *et al* looked secreted effectors in the rice pathogen *Xanthomonas oryzae*, including two complete genomes generated using PacBio sequencing.

MGen's first microbiome paper was a study of the microbes inhabiting the rumen of moose from Vermont, Alaska and Norway. The differences in microbial composition provide clues to what the different moose feed on: Alaskan moose had higher rates of archaeal species that are associated with high starch diets; while Vermont moose had more microbes associated with forage diets, consistent with high fibre intake.

Methods Papers

While microbial genomic epidemiology is now well established, microbial genomics remains a developing field and methods development is critical to drive research forward. In its first six months MGen has published four Methods Papers, which are all available as free and open source software.

CRISPR was arguably the hottest topic in the science world in 2015. The design of guide RNA (gRNA) is critical, and MGen published Peng & Tarleton's paper on their Eukaryotic Pathogen gRNA Design Tool (*EuPaGDT*), which "identifies gRNA in input genes to guide users in arriving at well-informed and appropriate gRNA design for many eukaryotic pathogens" (available at http://grna.ctegd.uga.edu).

Other methods papers include *K-Pax-2*, an *R* package for Bayesian identification of cluster-defining amino acid positions in large sequence datasets (https://github.com/alberto-p/kpax2); *Shetti*, for parsing and manipulating large sets of sequences without using the command line (https://sourceforge.net/projects/shetti); and *SimBac*, for simulating whole bacterial genomes with homologous recombination, which is crucial for the development and comparison of recombination analysis methods (http://github.com/tbrown91/SimBac).

Kat Holt
***Microbial Genomics*, Senior Editor**