Visualising genome data

Scott Beatson

Australian Infectious Diseases Research Centre
School of Chemistry and Molecular Biosciences
University of Queensland
Acknowledgements: Beatson group

Kirstin  Bryan  Mitchell SC  Nouri
Nathan  Nabil  Mitchell S
Beatson microbial genomics group

• The Australian Infectious Diseases research centre (AID) links > 50 groups in molecular microbiological and clinical expertise from the UQ Faculties of Science and Health Sciences, and UQCCR, QCMRI, IMB, AIBN, the Diamantina Institute and QIMR.

• Microbial genomics is a key research strength that benefits from closer links between clinicians and molecular microbiologists.

• My group uses sequencing technologies to better understand bacterial pathogenesis (pathogenomics), virulence factor and antibiotic resistance mobilization, and the spread of bacterial infectious diseases (genomic epidemiology).
Microbial genomics in the Beatson group


**Easyfig:** easy preparation of scaled genetic loci images for bacterial genome comparisons. Sullivan et al., Bioinformatics. 2011 Apr 1;27(7):1009-10.

First genome sequence for the globally disseminated *E. coli* ST131 clone (454). Totsika et al., PLoS One. 2011;6(10):e26578

Puerperal sepsis (Streptococcus pyogenes) outbreak investigation with Illumina sequencing Ben Zakour et al., J Clin Microbiol. 2012 Jul;50(7):2224-8
Genomics visualisation: the near future

- Rich, dynamic visualisation within the modern web browser
D3.js: Data Driven Documents

D3 is a JavaScript library for manipulating documents based on data. D3 helps you bring data to life using HTML, SVG and CSS. D3’s emphasis on web standards gives you the full capabilities of modern browsers without tying yourself to a proprietary framework, combining powerful visualization components and a data-driven approach to DOM manipulation.

http://d3js.org/

D3: Data-Driven Documents
Michael Bostock, Vadim Ogievetsky, Jeffrey Heer
(Bad) Bioinformatician: Here are your VCF files. It contains information on the variants we detected from you NGS data:

```
##fileformat=VCFv4.0
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=1000GenomesPilot-NCBI36
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">

#CHROM POS     ID        REF ALT    QUAL FILTER INFO                              FORMAT      NA00001        NA00002        NA00003
20     14370   rs6054257 G      A       29   PASS   NS=3;DP=14;AF=0.5;DB;H2           GT:GQ:DP:HQ 0|0:48:1:51,51 1|0:48:8:51,51 1/1:43:5:.,.
20     17330   rs6040355 A      G,T     67   PASS   NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27 2/2:0:18,2  2/2:35:4
21     1230237 .         T      .       47   PASS   NS=3;DP=13;AA=T                   GT:GQ:DP    0/1:35:4       0/2:17:2       1/1:40:3

(Excited) Biologist: Thanks. I will load it into Excel
(a few weeks later)

(Depressed, Good) Biologist: Can you give me the SNPs that satisfy this <xyz> criteria
(a few hours later)

Biologist: ... and this this <xyz> criteria

Why use data visualisation?
Alternatives: choose your language

- **Shiny** | Easy web applications in R | [http://www.rstudio.com/shiny/](http://www.rstudio.com/shiny/)

**Matplotlib WebAgg**
Render matplotlib plots directly to the web browser.

In current development branch

**Bokeh**
Python interactive visualization library for large datasets that natively uses the latest web technologies

[http://github.com/ContinuumIO/Bokeh](http://github.com/ContinuumIO/Bokeh)
Large genome viewers

Small genome viewers

- Artemis & Artemis Comparison Tool (Sanger Institute)
  - http://www.sanger.ac.uk/resources/software/artemis/
  - http://www.sanger.ac.uk/resources/software/act/
E. coli O25b-ST131 clone
– the new global face of UPEC

- Pandemic
  - Since 2008 simultaneous spread and high prevalence in multiple countries on several continents (Europe, Asia, Africa, North America and recently Australia) (Nicolas-Chanoine et al 2008; Coque et al 2008; Clermont et al 2008; Lau et al 2008)

Rogers et al 2011
BRIG (Blast Ring Image Generator)
Alikhan, Petty, Ben Zakour, Beatson
BMC Genomics. 2011 12:402

Totsika et al 2011 PLoS ONE
BRIG implementation

• BRIG is cross-platform and is written and requires JAVA 1.6.
• BRIG uses BLAST for genome alignments.
• JDOM is used for the internal data structure and CGView for Image rendering. Both are bundled in the package.
• Screenshots are from BRIG in Vista, it looks a little different on Linux and Mac.
Step 1: specify input files

Reference sequence appears in the centre of ring, FASTA or Genbank/EMBL

Pool of sequences to use as queries

BLAST options e.g. number of cores, filter on/off
Step 1: specify ring settings

Legend text

Image title

BLAST type

Add custom annotations

Sequences shown on this ring

Other settings

Sequence pool
Step 3: Submit and wait

Submit image to render
Coverage

Contig boundaries (alternating red/blue)

GC Content

Custom annotations

Legend showing colour gradient for % similarity

GC Skew

2 mbp
Comparison of five M28 isolates Illumina raw reads mapped onto MGAS 6180

(RIG, Alikhan et al. BMC Microbiology 2011)
• Phage 6180.2 encodes for SpeK and Sla
• ICEpPS_008 encodes for multidrug efflux proteins and a putative lipoprotein
• Phage pPS_008 encodes for several hypothetical proteins
BRIG used to survey individual genes in raw reads

Complete assembled genomes:

Raw reads:
Group software: SeqFindR


- **SeqDB**: A multifasta file of virulence factors
  - This is built/provided by the user.
    - >identifier, gene id, annotation, organism [class]
    - SEQUENCE
    - >APEC01_O1CoBM73, tsh, Tsh, Escherichia coli O1:K1:H7 (APEC) [Autotransporters]
    - ATGAACAGAATTTATTCTCTTCGCTACAGCGCTGTGG...

- **TOL**: A hit acceptance tolerance/cuttoff (0.95 default)
  - Defined as: hsp.identities/record.query_length >= cutoff

- **ASS**: Directory containing assemblies

- **CONS**: Directory containing consensus sequence from mapping reads to VFDB (optional)

- **INDEX**: A text file containing a pre-defined order (optional)
  - With this option, no clustering is performed

(Stanton-Cook, Beatson unpublished)
SeqFinder: independent of assembly

Output order determined by user:
- input order
- hierarchical clustering

Match consensus and mapping ≥ 95%
No match or match < 95%
Match assembly only ≥ 95%
Match mapping only ≥ 95%
SeqFindR

User adjustable hit thresholds
SeqFindR

Plot many categories simultaneously
SeqFindR

Cluster rows by similarity or order according to phylogenetic analysis
Show characteristic regions
Summary: BRIG

- comparative circular images showing conservation compared to a reference genome
- platform independent
- whole genomes or groups of genes/genomes can be used as reference
- raw reads, assembled genomes can be queried with BLAST
- custom graphs can be plotted, including coverage from BAM file

Nabil Alikhan


http://sourceforge.net/projects/brig/
Summary: SeqFindR

- comparative grid images showing conservation compared to reference genes
- currently command-line only (web-version in d3.js coming soon)
- more scalable than BRIG; produces generic matrix suitable for other image rendering software
- “traffic light” format suitable for alleles differing by one SNV

https://github.com/mscook

SeqFindR manuscript in preparation.

See also Banzai: high-throughput QC, assembly, mapping, reporting and phylogenomics for large groups of bacterial genomes
Summary: EasyFig

- Linear genetic loci images
  - from one gene to whole genomes
- BLAST comparisons similar to Artemis Comparison Images
- Custom graphs can be plotted; i.e. suitable for RNASeq mapping figures

Mitchell Sullivan
http://sourceforge.net/projects/easyfig/

Minh-Duh Phan Scott Beatson, Mark Shembri et al., submitted.
Summary: Ongoing work

- Database/software development tailored to large-scale bacterial genome sequencing efforts
- “Clinic ready” reports from raw sequencing data from infectious disease bacteria
  - E.g. See Walker & Beatson, Science, Epidemiology: Outsmarting Outbreaks (2012)

- Data driven documents interacting with sequence data in cloud
  - Enable anyone to rapidly access precomputed genome analyses via a variety of graphical interfaces.

Identification: *Klebsiella pneumoniae*

Phylogenomic comparison with recent *K. pneumoniae* isolates:

Outbreak status: confirmed

Patient trace data:

<table>
<thead>
<tr>
<th>Patient C, isolate 1</th>
<th>Identification: <em>Klebsiella pneumoniae</em></th>
</tr>
</thead>
<tbody>
<tr>
<td>Phylogenomic comparison with recent <em>K. pneumoniae</em> isolates:</td>
<td></td>
</tr>
<tr>
<td>Outbreak status: confirmed</td>
<td></td>
</tr>
<tr>
<td>Patient trace data:</td>
<td></td>
</tr>
<tr>
<td>Gene profile:</td>
<td></td>
</tr>
<tr>
<td>Antibiotic profile: Amp^R^ Cef^R^ Mer^R^ Gen^S^ Tig^R^ (carbapenemase positive)</td>
<td></td>
</tr>
<tr>
<td>Virulence profile: Positive: tox1, tox3; tox4 Negative: tox2, tox5</td>
<td></td>
</tr>
<tr>
<td>Comments: Tigecycline resistance detected</td>
<td></td>
</tr>
</tbody>
</table>
Acknowledgements

Beatson group
Nouri Ben Zakour
Mitchell Stanton-Cook
Brian Forde
Nathan Bachmann
Kirstin Hanks
Nabil Alikhan
Mitchell Sullivan
Bryan Wee
Nicola Petty (now UTS)
Elizabeth Skippington

Schembri group (UQ)
Mark Schembri
Minh-Duy Phan
Kate Peters
Sohinee Sarkar
Luke Allsopp
Maud Achard
Danilo Moriel
Makrina Totsika

University of Manchester
Mathew Upton

SCMB, UQ
Mark Walker

Funding:
AID microbial genomics
Australian Research Council
Australian National Health & Medical Research Council

Contact: s.beatson@uq.edu.au
http://github.com/BeatsonLab-MicrobialGenomics