PhD Project Update: Photorhabdus Virulence Cassettes

Joe Healey

Department of Chemistry (MOAC),
University of Warwick, UK

J.R.J.Healey@warwick.ac.uk

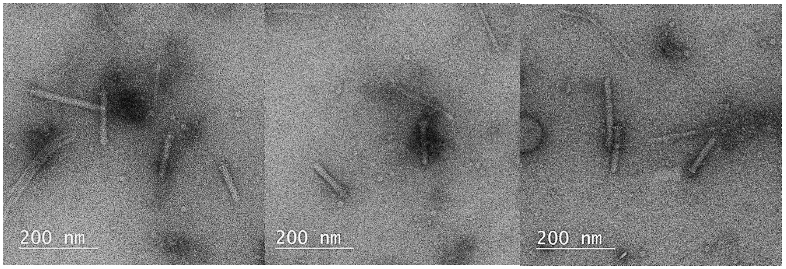
www.warwick.ac.uk/go/gibsongroup

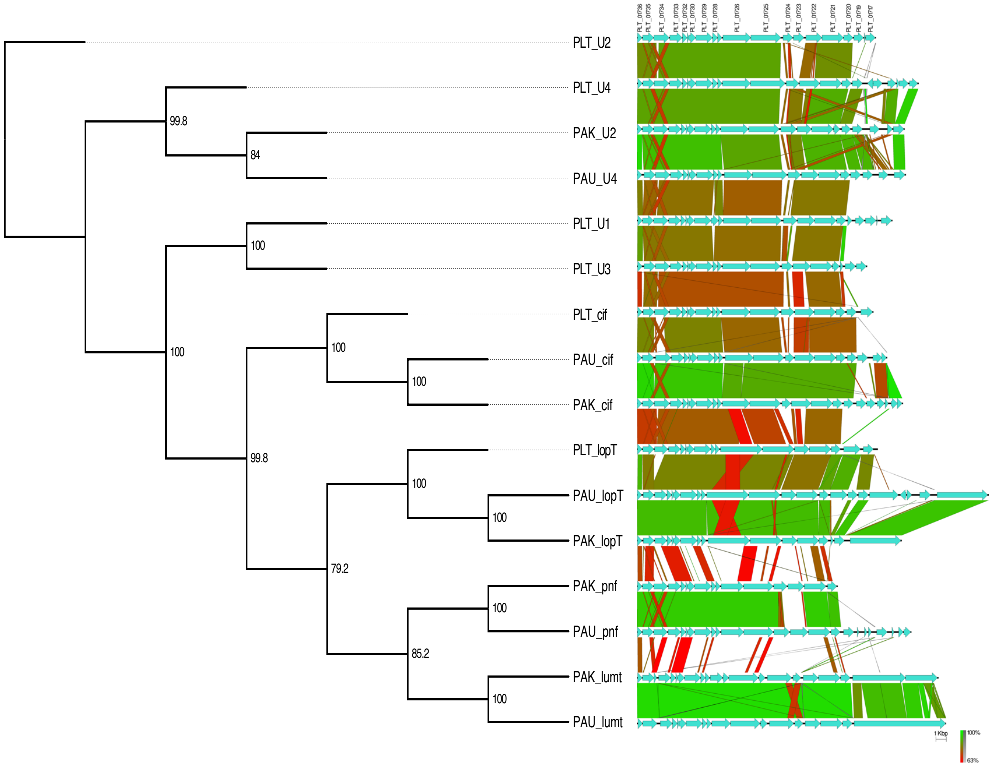
@JRJHealey

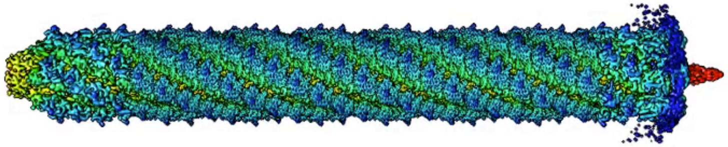
Group Presentation 8/2/2017

Quick Refresher

PhD based on trying to understand and exploit a toxin delivery mechanism created by the insect/human pathogen *Photorhabdus* – the *Photorhabdus* Virulence Cassette (PVC)

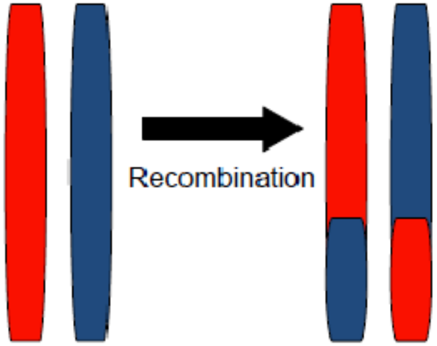






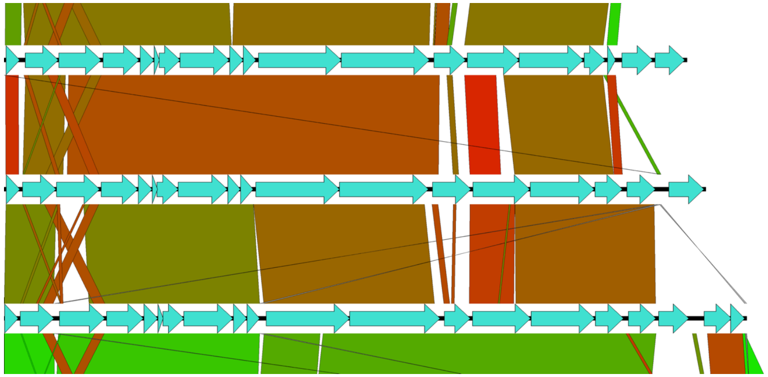
Yang, G., *et al*. J Bacteriol, 2006. **188**(6): p. 2254-61

Chapter 1: Understanding recombination in the PVC operons



PVCs are ‘the same but different’. They have diverse sequences but perform the same jobs – we think this is vital to their function and why there are multiple forms.

‘Usual’ methods of determining recombination are high resolution, but only work for high identity seqs (e.g. ClonalFrame).

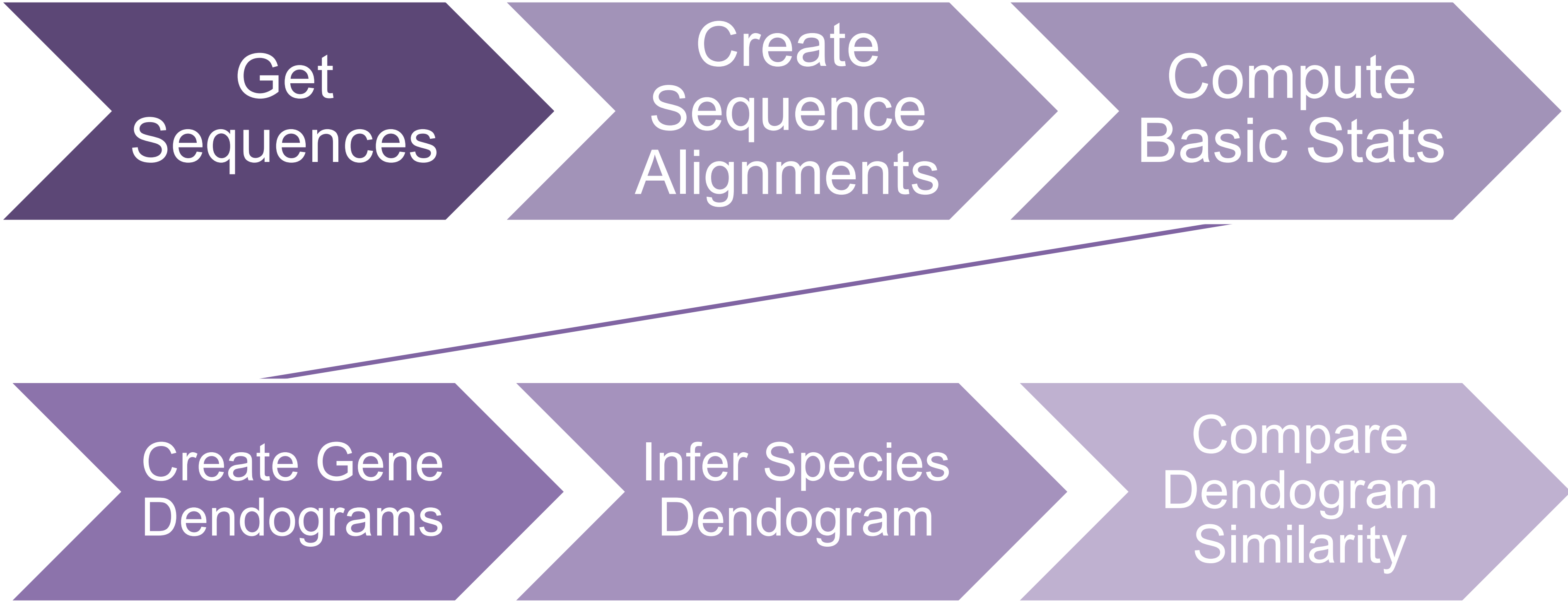


We can look on a ’gene-by-gene’ basis though.

Key terms:

- **Recombination** (interchanging segments of DNA, usually requiring some sequence similarity)
- **Operon** (A cluster of genes with different products, but are functionally linked)
- **Identity** (How similar 2 stretches of a gene/protein are. Identity =/= homology.

Basic workflow



Get Sequences

Create Sequence Alignments

Compute Basic Stats

Create Gene Dendograms

Infer Species Dendogram

Compare Dendogram Similarity

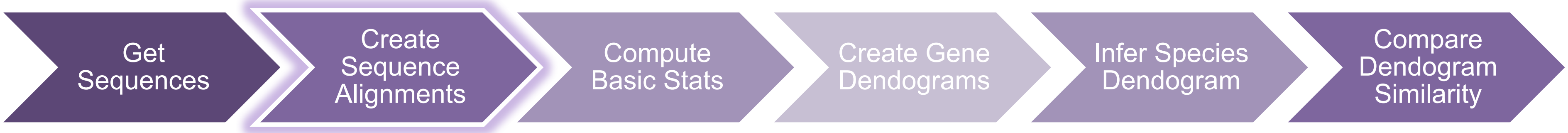
Creating Sequence Alignments

Pairwise or Multiple Sequence Alignment (pairwise limited to 2 seqs), but can be more informative

Loads of aligners. Each has strengths and weaknesses:

- e.g. MUSCLE (accurate, good for medium datasets and proteins)
- e.g. MAFFT (fast, good for medium to large datasets)
- e.g. T-Coffee (accurate, with error correction, small datasets)
- e.g. Clustal (User friendly, lots of algorithm options)

Once you have an MSA, you can calculate %ID, and do many downstream analyses.



Get Sequences

Create Sequence Alignments

Compute Basic Stats

Create Gene Dendograms

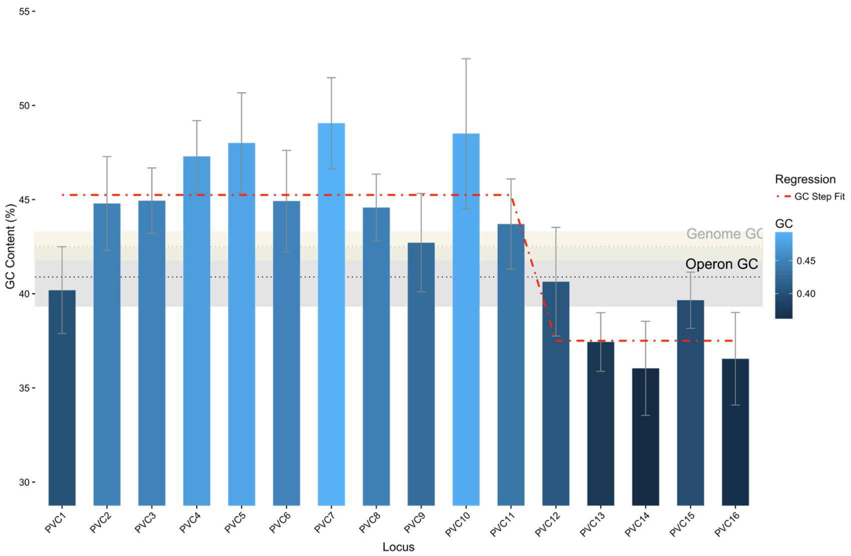
Infer Species Dendogram

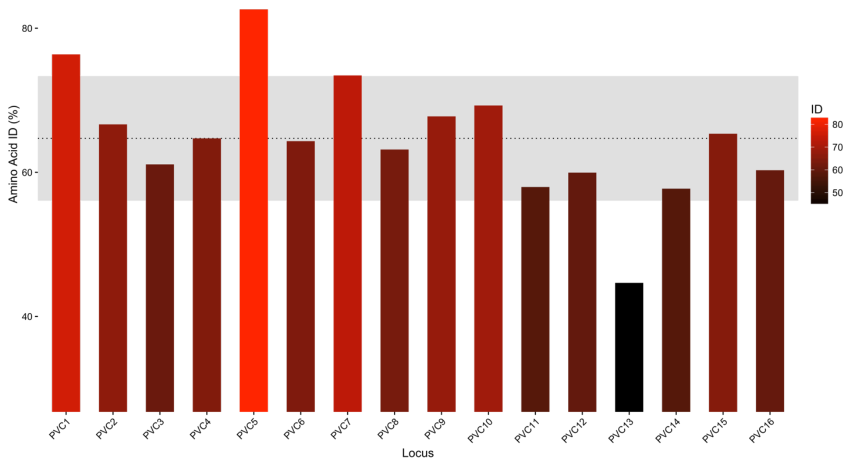
Compare Dendogram Similarity

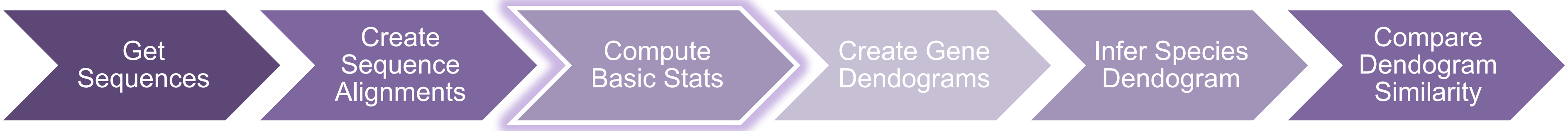
http://www.ebi.ac.uk/Tools/msa/ or http://www.ebi.ac.uk/Tools/psa/

Basic Sequence Stats

Once you have the sequences you need, with simple scripts (~12 lines of python) or online tools we can get basic info which can be surprisingly informative:







Get Sequences

Create Sequence Alignments

Compute Basic Stats

Create Gene Dendograms

Infer Species Dendogram

Compare Dendogram Similarity

http://www.ebi.ac.uk/services - loads of good little web tools

Comparing sequences

From a MSA, you create a hierarchy of relatedness, AKA a dendogram or phylogeny. This is repeated for every (structural) gene along the operon.

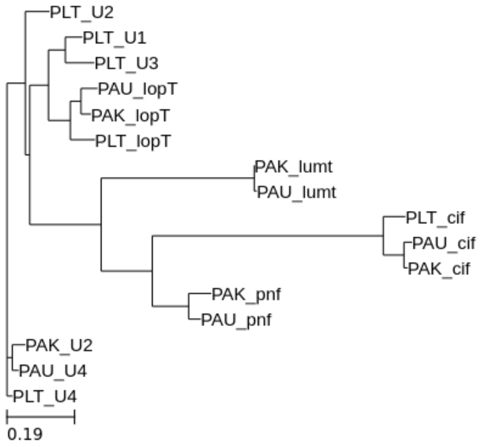
- e.g. RAxML (accurate, pretty fast, horrible to use)
- e.g. Fasttree (uses a weird algorithm, but fast and easy)

Visit http://phylo.cs.mcgill.ca/ and thank me later (I take no responsibility for unfinished theses in the event you get hooked).

Multiple Sequence Alignment

```
PAU_U4    MSTTPEQIAV EYPIPTYRFV NSVSGLDISH...
PLT_U4    MSTTPEQIAV EYPIPTYRFV NSVSGLDISH...
PAK_U2    MSTTPEQIAV EYPIPTYRFV NSVSGLDISH...
PLT_U2    MSTTPEQIAV EYPIPTYRFV NSVSGLDISH...
PAK_lopT  MTTTT----V DYPIPAYRFV NNVSGLDITY...
PAU_lopT  MATTT----V DYPIPAYRFV NSVSGLDITY...
PLT_lopT  MSVTTEQIAV DYPIPTYRFV NNVSGLDITY...
```

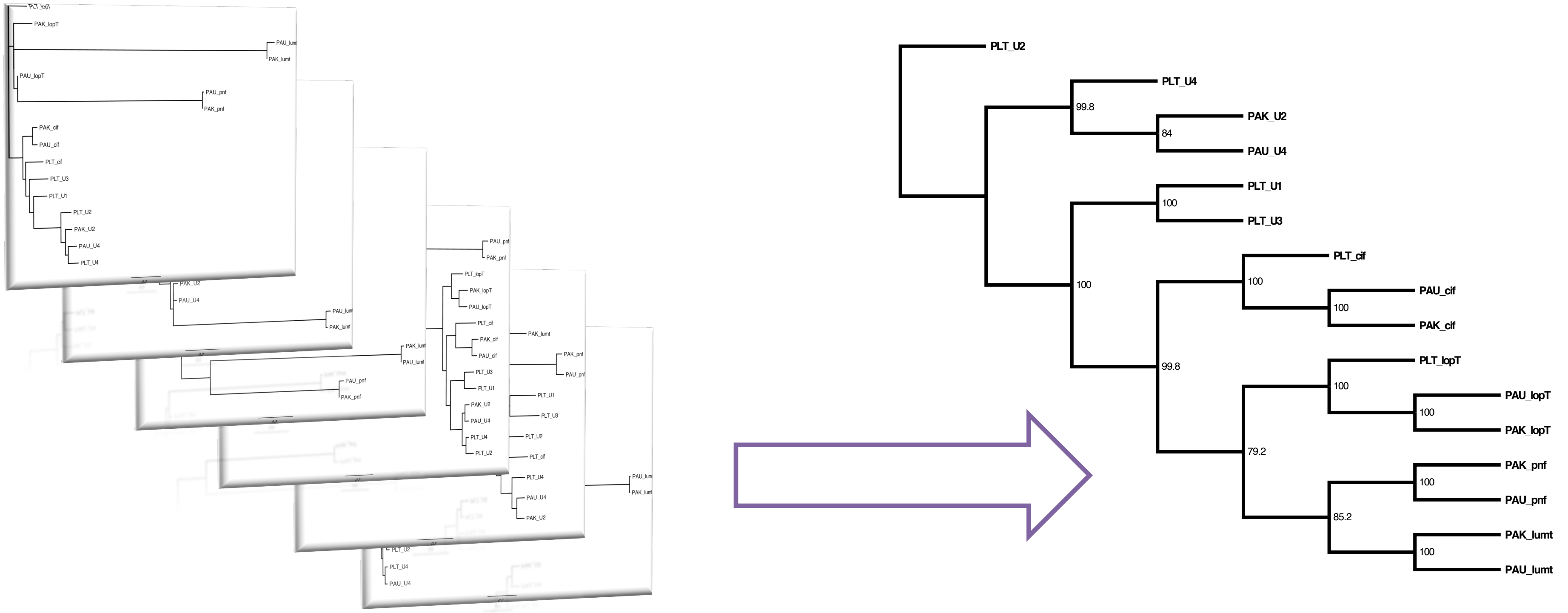
Make a Distance Matrix

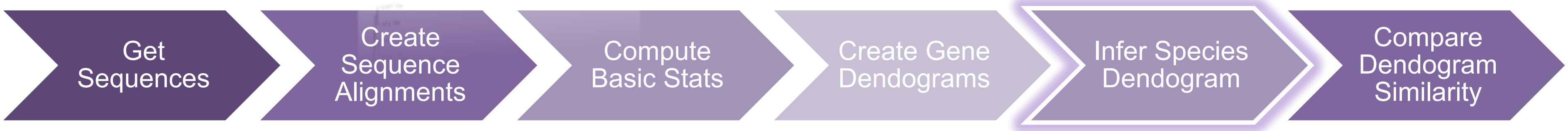


Get Sequences → Create Sequence Alignments → Compute Basic Stats → Create Gene Dendograms → Infer Species Dendogram → Compare Dendogram Similarity

Comparing phylogenetic patterns

With trees for every single gene, they can be compared for consensus, and a simulated tree for the species is inferred by a program called ASTRAL





Get Sequences

Create Sequence Alignments

Compute Basic Stats

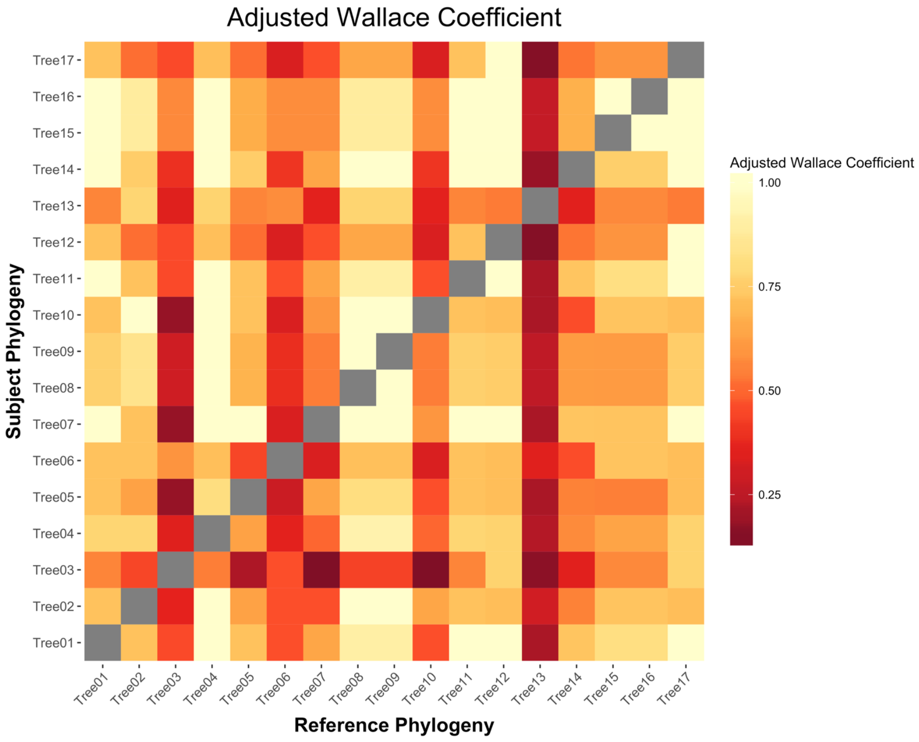
Create Gene Dendograms

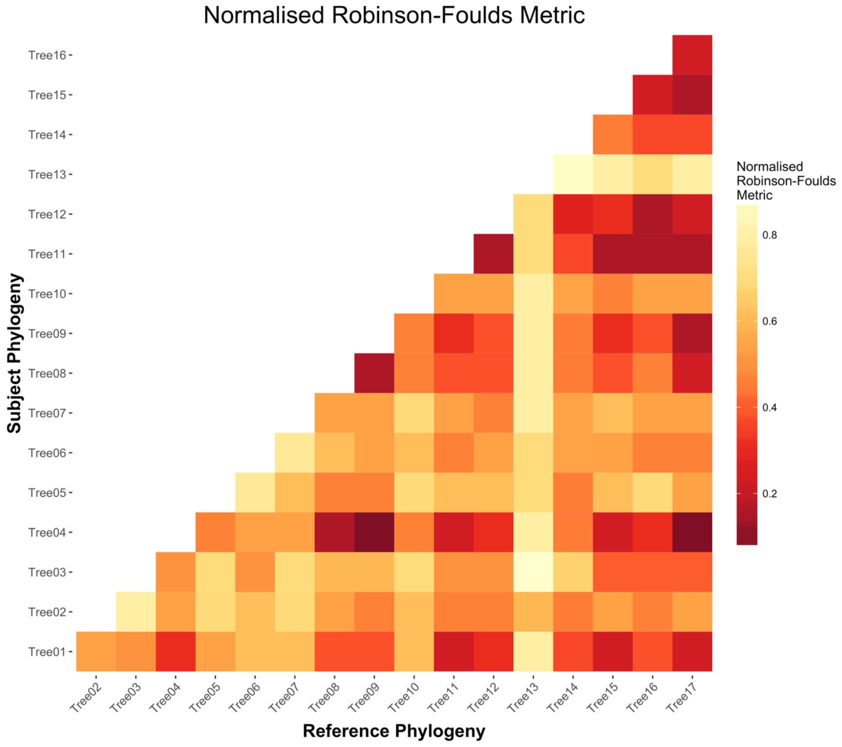
Infer Species Dendogram

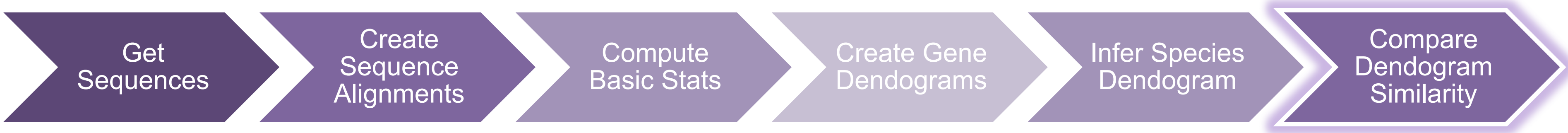
Compare Dendogram Similarity

Visualising Phylogenetic Patterns

Lastly, there are a number of metrics of tree similarity (“congruence”). We can calculate and then visualise these to look for patterns.






Get Sequences

Create Sequence Alignments

Compute Basic Stats

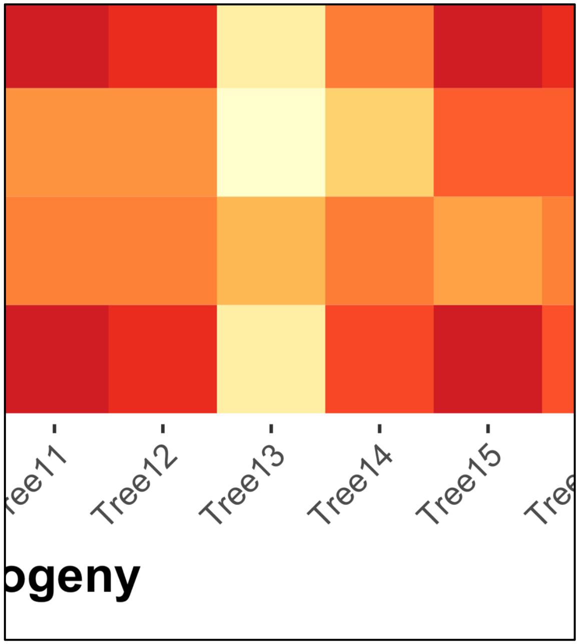
Create Gene Dendograms

Infer Species Dendogram

Compare Dendogram Similarity

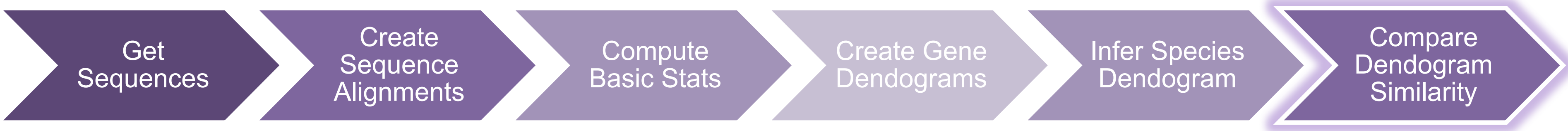
So what next?

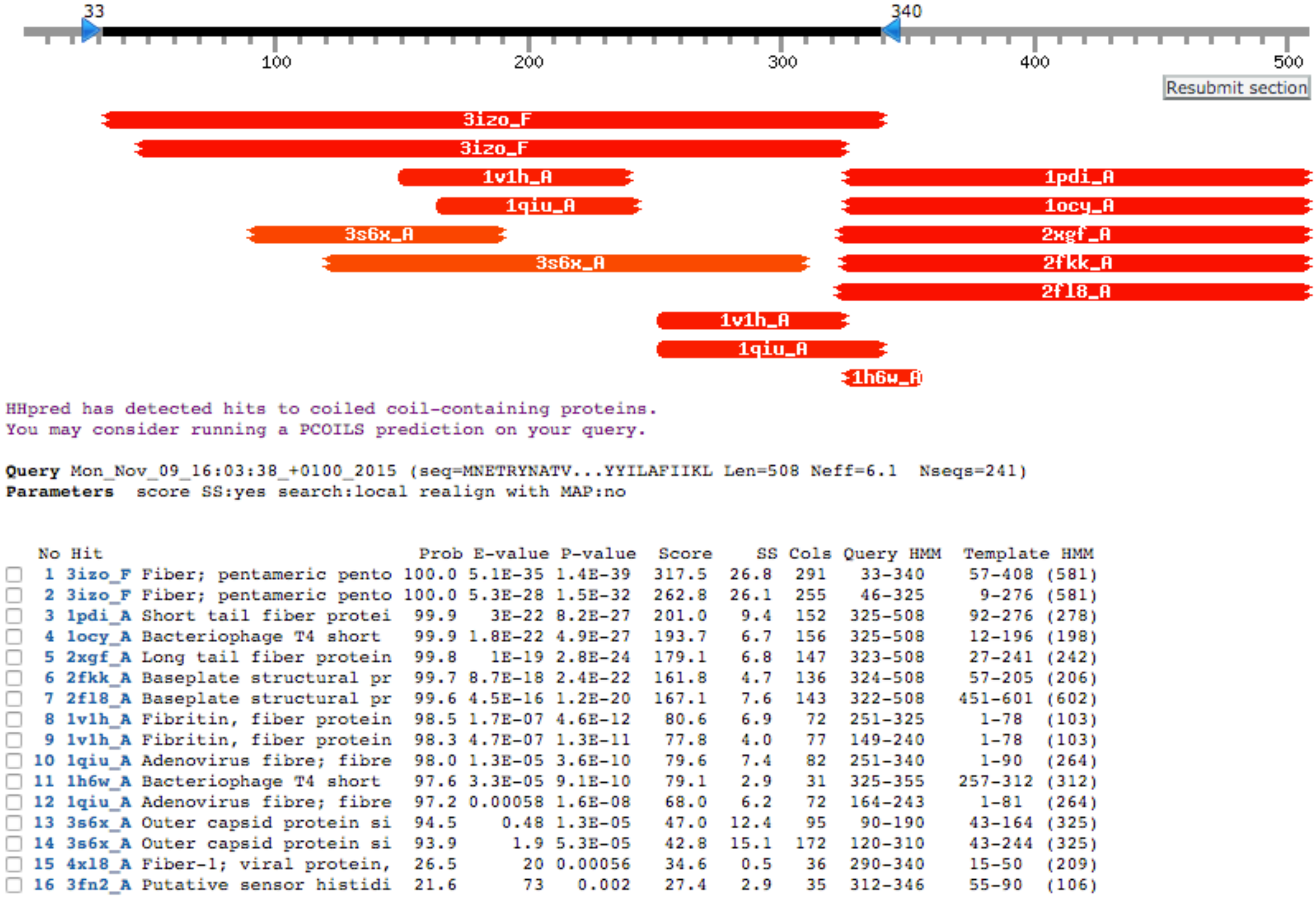
There is a stand out trend in the data, no matter which metric you assess it with – Tree 13. PVC13 is the tail-fibre binding protein that belongs to the PVC (the same proteins Laura talked about in her presentation 2 weeks ago)



Based on their variability and proposed functions, they are great targets for cloning. (Thesis chapter 2 is structural characterisation of 2 of these proteins).

There are tools that give you functional information (with caveats). Most famous is BLAST, but there are more sensitive tools (HHpred).





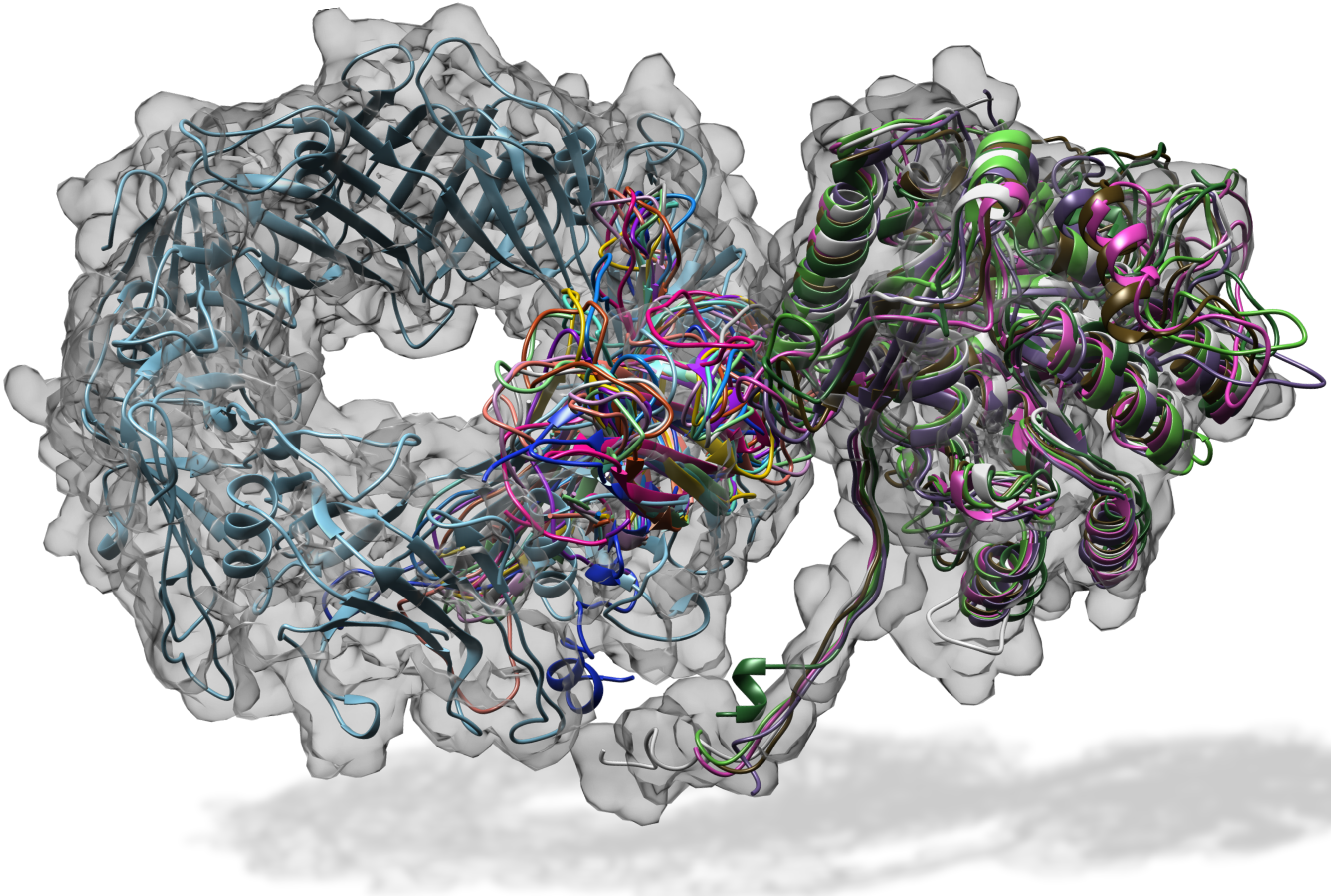
```
HHpred has detected hits to coiled coil-containing proteins.
You may consider running a PCOILS prediction on your query.

Query Mon_Nov_09_16:03:38_+0100_2015 (seq=MNETRYNATV...YYILAFIIKL Len=508 Neff=6.1  Nseqs=241)
Parameters  score SS:yes search:local realign with MAP:no


   No Hit                             Prob E-value P-value  Score    SS Cols Query HMM  Template HMM
□   1 3izo_F Fiber; pentameric pento 100.0 5.1E-35 1.4E-39 317.5  26.8  291   33-340    57-408 (581)
□   2 3izo_F Fiber; pentameric pento 100.0 5.3E-28 1.5E-32 262.8  26.1  255   46-325     9-276 (581)
□   3 1pdi_A Short tail fiber protei  99.9   3E-22 8.2E-27 201.0   9.4  152  325-508    92-276 (278)
□   4 1ocy_A Bacteriophage T4 short   99.9 1.8E-22 4.9E-27 193.7   6.7  156  325-508    12-196 (198)
□   5 2xgf_A Long tail fiber protein  99.8   1E-19 2.8E-24 179.1   6.8  147  323-508    27-241 (242)
□   6 2fkk_A Baseplate structural pr  99.7 8.7E-18 2.4E-22 161.8   4.7  136  324-508    57-205 (206)
□   7 2f18_A Baseplate structural pr  99.6 4.5E-16 1.2E-20 167.1   7.6  143  322-508   451-601 (602)
□   8 1v1h_A Fibritin, fiber protein  98.5 1.7E-07 4.6E-12  80.6   6.9   72  251-325     1-78  (103)
□   9 1v1h_A Fibritin, fiber protein  98.3 4.7E-07 1.3E-11  77.8   4.0   77  149-240     1-78  (103)
□  10 1qiu_A Adenovirus fibre; fibre  98.0 1.3E-05 3.6E-10  79.6   7.4   82  251-340     1-90  (264)
□  11 1h6w_A Bacteriophage T4 short   97.6 3.3E-05 9.1E-10  79.1   2.9   31  325-355   257-312 (312)
□  12 1qiu_A Adenovirus fibre; fibre  97.2 0.00058 1.6E-08  68.0   6.2   72  164-243     1-81  (264)
□  13 3s6x_A Outer capsid protein si  94.5    0.48 1.3E-05  47.0  12.4   95   90-190    43-164 (325)
□  14 3s6x_A Outer capsid protein si  93.9     1.9 5.3E-05  42.8  15.1  172  120-310    43-244 (325)
□  15 4x18_A Fiber-1; viral protein,  26.5      20 0.00056  34.6   0.5   36  290-340    15-50  (209)
□  16 3fn2_A Putative sensor histidi  21.6      73   0.002  27.4   2.9   35  312-346    55-90  (106)
```

https://toolkit.tuebingen.mpg.de/hhpred

If you're lucky, your protein (or at least domains of it) will already be known and you'll get useful structural info for free.

You may even be able to simulate some of them…

Again, EBI has *loads* of tools for analysing all sorts of things: e.g. presence of repeats (RADAR), ontology (InterProScan), superfamily identification, signal peptide detection, transmembrane domain detection - *ad infinitum.*



http://www.ebi.ac.uk/Tools/pfa/ (Protein Functional Analyses)

Some additional thoughts

What tools you need and what analysis is open to you very much depends on the question typically…however:

InterProScan is a great suite of tools (and a good place to start) if you want a big report on pretty much everything your protein could be doing.

HHpred is our favourite for more sensitive sequence analysis. BLAST is probably still king for nucleotide searches though.

Many of the tools I talked about are commandline based, but lots of them have GUIs (e.g. clustal) or webserver interfaces, if you are commandline averse (most biologists are).

Life is easier if you learn a ‘proper’ programming language (more than happy to help!).

Acknowledgements

Gibson Group, **2017**

Post-docs

- *Dr. Sarah-Jane Richards*
- *Dr. Caroline Biggs*
- *Dr. Collette Guy*
- *Dr. Lucienne Otten*
- *Kathryn Styles*
- *Dr. Muhammad Hasan*

Undergrad Students

- *Nick Vail*

PhD Students

- *Sang Ho Won*
- *Lewis Blackman*
- *Benjamin Martyn*
- *Joseph Healey*
- *Nick Alcaraz*
- *Chris Stubbs*
- *Ben Graham*
- *Trisha Bailey*
- *Laura Wilkins*
- *Maria Grypioti*
- *Julia Lipecki*
- *Vinko Varas*
- *Gabriel Erni-Cassolla*
- *Robyn Wright*
- *Alice Faytor*
- *Iain Galpin*

Funders £$€

erc
European Research Council
Established by the European Commission

EPSRC
Pioneering research and skills

The Leverhulme Trust

THE UNIVERSITY OF WARWICK

BBSRC
bioscience for the future

Noreen Murray Philanthropy Award

BBINTERNATIONAL

BASF
The Chemical Company

Iceni Diagnostics

ingenza

THE ROYAL SOCIETY

MRC Medical Research Council

investing in **your** future
European Regional Development Fund
European Union

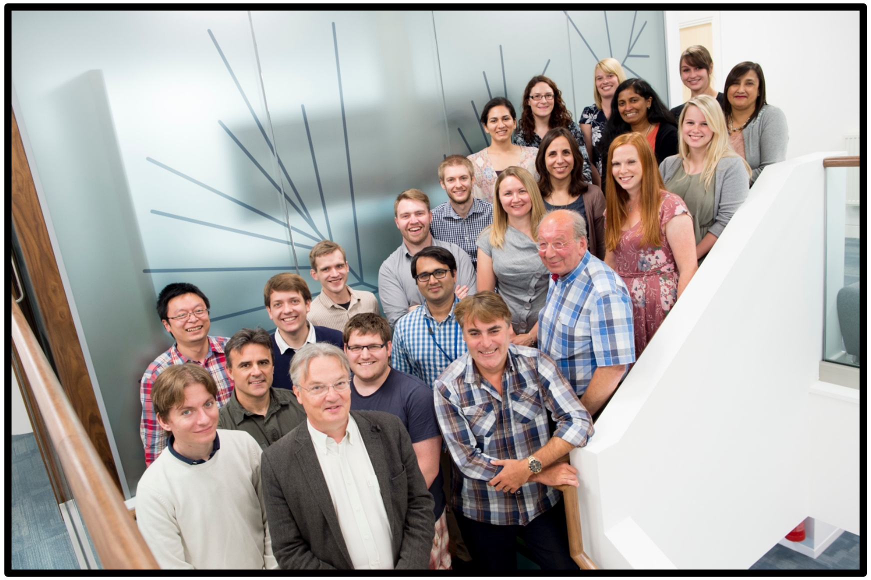
Advantage West Midlands
www.advantagewm.co.uk

Acknowledgements

Waterfield Group

Post-docs
- Dr. Alexia Hapeshi

PhD Students
- Tom Brooker



Microbiology and Infection Group

Funding/Studentship

Joe Healey

MOAC DTC, University of Warwick, UK

J.R.J.Healey@warwick.ac.uk

www.warwick.ac.uk/go/gibsongroup

www.warwick.ac.uk/waterfieldlab

@JRJHealey @LabGibson @Nick_Waterfield