Trawling for genetic switches in bacterial genomes

Paul Swift[1], Peter Power[1], Milo Thurston[2], Stewart Houton[3], Dawn Field[3]

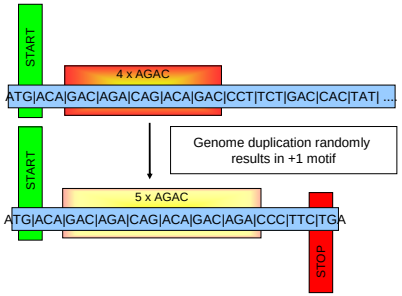
[1] Dept. of Paediatrics, University of Oxford, OX3 9DU, UK

[2] Dept. of Atmospheric, Oceanic and Planetary Physics, University of Oxford, OX1 3PU, UK

[3] CEH Oxford, Mansfield Rd, Oxford, OX1 3SR, UK

Background

Microsatellites are short DNA repeats (eg. ATATAT). They are found in the vast majority of sequenced genomes, both eukaryotic and prokaryotic. An important property is that they randomly change in length during genome replication. If a microsatellite is present in a gene, the resulting frameshift can switch gene translation on or off. Genes containing these genetic switches are known as contingency loci.



Since they are rapidly and reversibly switching on and off, contingency loci guarantee diversity across a bacterial population. For example, they can ensure a diverse set of epitopes (cell-surface exposed proteins recognised by the immune system), maximising survival. This mechanism has been well-characterised in *Haemophilus influenzae*[1] and *Neisseria meningitidis*[2], pathogens responsible for meningitis and respiratory tract infections.

Contingency loci have only been found in a small number of pathogenic bacteria.

Aim

To develop a bioinformatic workflow which will identify potential contingency loci. This will give an indication of the rarity of contingency loci, and whether they are restricted to bacterial pathogens.

Methods

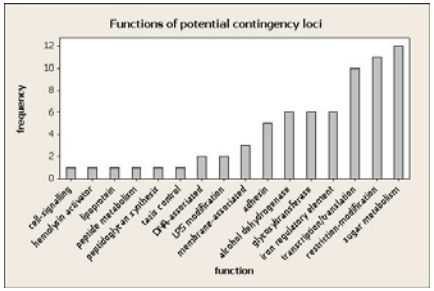
- Downloaded 500 bacterial genomes from the NCBI and recorded their taxonomy
- Markov model used to determine the minimum number of motifs required for each genome and motif length.
- Ran Msatfinder[3] to identify microsatellites using these thresholds (A).
- Removed microsatellites whose motif lengths were divisible by three (eg. ATG), since they cannot produce frameshifts (B).
- Identified microsatellites present in a proposed gene using Glimmer[4] (C).
- Filtered for motif number greater than 13 (D), microsatellite being within the first 23% of the proposed gene (E), and the microsatellite occupying <60% of the gene (F). These filters were decided upon using a test set of known contingency loci.
- Carried out Blast[5] search against the Swissprot[6] database, with score >50 and e-value <e^{-6} (G)
- Checked that BLAST results were for characterised genes (H).

Results

From 22,562 microsatellites found by Msatfinder, our workflow identifed 69 as being potential contingency loci. Of these, 87% came from genera with contingency loci recorded in the literature. All but one of the species containing potential contingency loci are commensals or pathogens.

Species without known contingency loci:
1. The plant pathogen *Xylella fastidiosa*. While no plant pathogen is known to contain contingency loci, the proposed gene returns BLAST hits for adhesins, which fits with the accepted model.
2. The extremophile *Salinibacter ruber*. Proposed gene appears to be involved with peptide metabolism. It is intriguing that while *Salinibacter* is not a pathogen, it is exposed to a similarly hostile environment.

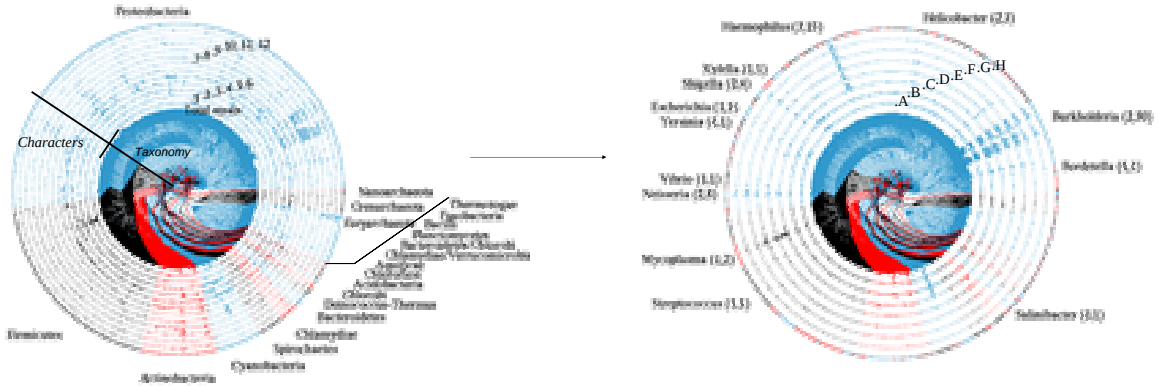
Functions of potential contingency loci from BLAST searches: While the majority are well-recognised and relate to immune system avoidance or defence against predators, others, such as alcohol dehydrogenase (found predominantly in *Burkholderia*) are novel in this context.



Fruity plots

These plots map characters onto a taxonomy. Our taxonomy here covers kingdom (bacteria /archaea), division, order, family, genus, species and genome. On the outer rings of the plot characters are represented by short lines of varying lengths. The length of a line is the character taken as a percentage of the maximum found. The colouration of these plots divides the genomes by division.

Tool online at
http://www.genomics.ceh.ac.uk/cgi-bin/msatfinder/frooty/input.cgi



A. *Total number of microsatellites:* The first of the character rings, closest to the centre, is the total number of microsatellites found. Moving outwards are twelve rings showing the number of microsatellites with motifs of length 1 (eg. AAAA) to 12. The ubiquity of microsatellites and the diversity in motif length preference are apparent.

B. *Filtering to identify potential contingency loci:* The innermost character ring (A) represents the total number of microsatellites. These are progressively pared down by filtering (B-G – see methods) to identify repeats similar to known *H. influenzae* contingency loci. Genera containing such repeats are indicated, with the number of species and microsatellites in parentheses.

Conclusions

- Contingency loci, as a mechanism, have appeared independently many times
- Nevertheless they appear to be extremely rare.
- We have identified potential contingency loci with novel functions.
- The vast majority of our potential contingency loci are in commensal/pathogenic bacteria, there is some evidence that contingency loci may be present in environmental bacteria.

References

1. De Bolle, X., Bayliss, C.D., Field, D., van de Ven, T., Saunders, N.J., Hood, D.W. and Moxon, E.R. (2000). The length of a tetranucleotide repeat tract in *Haemophilus influenzae* determines the phase variation of a gene with homology to type III DNA methyltransferases. *Molecular Microbiology* 35, 211-222.

2. Bayliss, C.D., Field, D. and Moxon, E.R. (2001). The simple sequence contingency loci of *Haemophilus influenzae* and *Neisseria meningitidis*. *J. Clin. Invest* 107, 657-666.

3. Thurston, M.I, Swift, P. and Field, D. (2005). Msatfinder: detection and characterisation of microsatellites. Distributed by the authors at http://www.genomics.ceh.ac.uk/msatfinder.

4. Delcher, A.L., Harmon, D., Kasif, S., White, O. and Salzberg, S.L. (1999) Improved microbial gene identification with GLIMMER. *Nuc. Acids Res.* 27, 4636-4641.

5. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.* 215, 403-410.

6. Bairoch, A., Boeckmann, B., Ferro, S., and Gasteiger, E. (2004). Swiss-Prot: Juggling between evolution and stability. *Brief. Bioinform.* 5, 39-55.