NIHR | Health Protection Research Unit in Gastrointestinal Infections at University of Liverpool

Ancestral inference and encoding of bacterial nonsense mutations for phylogenetic genome-wide association studies and convergence analysis

Matthew P. Moore[1,2], Aditya K. Lankapalli[1], Xavier Didelot[1,2]
1 University of Warwick 2 Health Protection Research Unit in Gastrointestinal Infections

Background

Adaptation by loss of function has been observed in many human-adapted, pathogenic bacteria[1–3]. In the *Shigella* spp., convergent loss of function is well characterised and the smaller genome size, compared with *E. coli,* is likely the result of genome streamlining[4]. Across *Shigella dysenteriae, Shigella flexneri* and *Shigella sonnei* parallel loss of function events have resulted in convergent metabolic capability reduction and accounted for 47% of within-species metabolic gene variation[4]. A study of *Salmonella enterica* serotype Typhyimurium found that pseudogenisation was linked to host adaptation including, for example, in the *sseL* gene linked, via inactivation, with systemic dissemination. Further, pseudogenisation of *sseK2*, *sseK3*, *avrA* and *sseL* involved in activating the pro-inflammatory response during infection were observed[5].

Bacterial genome-wide association studies (GWAS) have successfully been used to make associative and correlative inferences about the genetic basis of phenotypes of interest[6–8]. Confounding any GWAS inferences are the clonality of bacterial population structures[9], genetic content variability[10], imperfect heritability[11] and homologous recombination[12]. Corrections for confounding population structure have been developed based on phenotype rearrangement, clustering and dimension-reduction[13–19]. Methods have further been developed that take advantage of the population structure, providing phylogenetic approaches. Inferences can be made between pseudogenisation events at the gene level, though possibly resulting from many independent nonsense mutations, by collapsing to a binary genotype as with gene presence/absence data. Additionally, correlations may be inferred between an ancestral change in both genotype and phenotype. However, providing binary genotype data of LoF per gene, from independent but phylogenetically clustered nonsense mutations could incorrectly appear as a single ancestral event.

Methods

For the GWAS analysis a dataset of avian *E. coli* was selected, comprised of both infectious and asymptomatic carriage isolates from chickens, with some additional environmental isolates[20]. Variants were called by Snippy[21] against the reference genome (APECO78), providing SnpEff[22] formatted variant call file (*.vcf*) output of variant impact on gene function. Variants categorised as 'HIGH' impact by SnpEff were included in the analysis of loss of function. Genome-wide associations were performed for each phylogroup, as in the original study: A (n=71), B1 (n=85), B2 (n=152) and G/ST-117 (n=220) with TreeWAS2. Core SNP trees were generated for each phylogroup. Core genome alignments were generated with snippy-multi[21]. IQ-TREE[23] with the GTR+I+G substitution model and 1000 ultrafast bootstraps (UFBoot2)[24] was used to generate an initial phylogeny and ClonalFrameML[25] was used to detect regions of homologous recombination and adjust branch lengths to non-recombinant SNPs only.

All shared mutations (same position&variant) were subjected to tests of ancestry. The first test, whether the nonsense mutation is monophyletic is performed with the ETE 3 Toolkit[26]. The most recent common ancestor (MRCA) of all genomes with the mutation is determined and if they comprise all the children of that node, ancestry is asserted. If polyphyly is observed, ancestral state reconstruction (ASR) is conducted for every internal node back to the nonsense mutation MRCA with PastML[27]. All variants at the reference position are provided as mutated reference codons unless there is no variation, whereby the reference codon is provided. Where nonsense mutations are nested within an ancestral nonsense mutation clade, the ancestral nonsense mutation is considered to be canonical. Whether mutations are canonical will determine the pattern of ancestry that is reported for that LoF event; ignoring those that are nested.

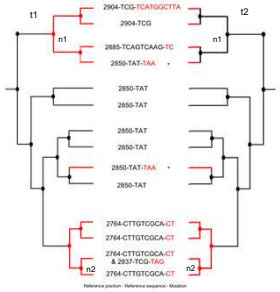

Figure 1. The imagined core genome phylogeny of 14 genomes (t1&t2) with some nonsense alleles in a single gene. In t1 branches are coloured red based on the ASR of binary nonsense allele states (t1) or retention of true states with NonsenseCanon (t2). MRCA node n2, is inferred to carry the ancestral-state of the frame-shifting deletion at reference position 2764. However, one genome in the clade also has a premature stop codon. The ancestral mutation is taken as canonical. (*) Shows a parallel mutation

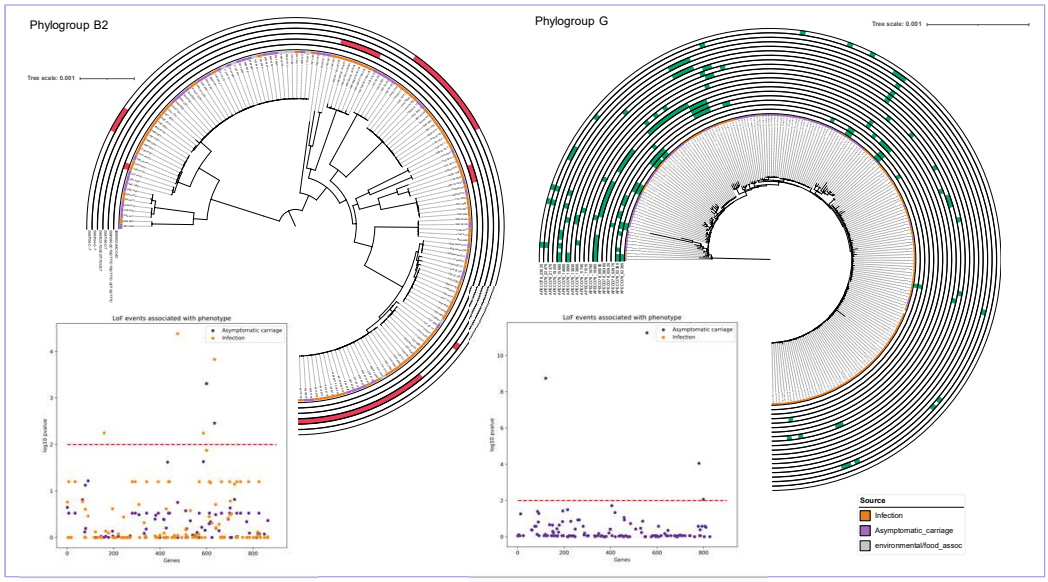

Figure 2. (Left) Core genome phylogeny of *E. coli* phylogroup B2 isolates. Colour rings correspond to phenotypes (inner) and *cirA* catecholate siderophore receptor gene nonsense mutations (outer rings). Each colour ring moving outwards corresponds to a different nonsense mutation (labelled). (Scatterplot, left) shows the p-values (log10) where nonsense mutations are associated with the 'Infection' or 'Asymptomatic carriage' phenotype. Above the red dashed line represents a p-value of <.01. (Scatterplot, right) displays significant values for phylogroup G. (Right) Core genome phylogeny of *E. coli* phylogroup G isolates. Colour rings correspond to phenotypes (inner) and moving outwards, a selection of genes with the largest number of independent nonsense events. Green blocks show where a genome carries (any) inactivating mutation in the gene (labelled)

Table 1. Group B2 and G genes with >4 parallel inactivation events, hypothetical genes not included

Phylogroup	Locus ID	No. parallel nonsense events	Product
GroupG	APECO78_05430	21	Diguanylate cyclase ***adrA***
	APECO78_02390	21	Predicted dinucleotide-binding enzymes
	APECO78_02385	15	Transposase and inactivated derivatives
	APECO78_19150	13	G:T/U mismatch-specific DNA glycosylase ***ygjF***
	APECO78_05920	9	bifunctional UDP-sugar hydrolase/5'-nucleotidase periplasmic precursor ***ushA***
	APECO78_22475	8	Beta-glucosidase/6-phospho-beta-glucosidase/beta-galactosidase ***bglB***
	APECO78_10385	8	4-aminobutyrate aminotransferase and related aminotransferases ***puuE/goaG***
	APECO78_14690	8	Site-specific recombinases, DNA invertase Pin homologs
	APECO78_11515	7	Putative tail fiber chaperone; Qin prophage ***tfaQ***
	APECO78_17410	6	CRISPR-associated helicase Cas3
	APECO78_23025	6	Mg-dependent Dnase ***yigX***
	APECO78_11740	6	Anaerobic dehydrogenases, typically selenocysteine-containing ***ynfF***
	APECO78_09000	6	Glucose-1-phosphatase/inositol phosphatase
	APECO78_19255	6	Inner membrane protein ***ygjV*** family
	APECO78_13800	6	Predicted transporter component
	APECO78_15680	6	Cellulase M and related proteins ***ypdE***
	APECO78_06385	6	Glycosyltransferases, probably involved in cell wall biogenesis ***nfrB***
	APECO78_13990	5	Nucleoside-diphosphate-sugar epimerases ***fcl***
	APECO78_10790	5	NAD-dependent aldehyde dehydrogenases ***aldA***
	APECO78_13835	5	Histidinol dehydrogenase
	APECO78_03975	5	Type II secretory pathway, ATPase PulE/Tfp pilus assembly pathway, ATPase ***PilB***
	APECO78_21370	5	AraC-type DNA-binding domain-containing proteins ***gadX***
GroupB2	APECO78_08030	10	Nitroreductase A ***nsfA***
	APECO78_14490	7	Catecholate siderophore receptor ***cirA***
	APECO78_02375	6	Transposase and inactivate derivatives
	APECO78_20405	6	Putative peptidoglycan-binding domain-containing protein (T3SS) ***gspA***
	APECO78_10515	5	Hydrolase A ***ycjY***
	APECO78_05920	5	bifunctional UDP-sugar hydrolase/5'-nucleotidase periplasmic precursor ***ushA***
	APECO78_14995	5	Type V secretory pathway, adhesin ***aidA***
	APECO78_22465	5	Enterochelin yieL

Results

It was possible to automatically determine nonsense ancestry patterns for most genes with nonsense mutation(s) in >1 genome. For the largest two phylogroups, groupB2 and groupG: 97.95% and 98.57% of genes ancestries were inferred, respectively. Table 1 displays the genes with the largest number of parallel LoF events in groupG & B2 genomes. Group G is thought to be an avian-host specialising lineage and as such convergence on LoF may indicate adaptive trends both towards and away from infectivity/pathogenicity. As groupG is specialised to the avian host many genes observed to have high levels of nonsense convergence in other groups are already missing or with a high proportion of inactivation in group G. Others are observed to be convergent across lineages such as *ushA,* which has convergently been inactivated across groups B1 (9 events), B2 (5 events) and G (9 events). An *ushA* homolog has additionally been reported to be adaptively inactivated in *Salmonella enterica.* Others such as *bglB* (8 inactivation events in groupG) is a known (strain-specific) pseudogene in *Shigella flexneri.*

Determination of the ancestral patterns also allowed a full, phylogenetic GWAS in order to infer genes, that by their inactivation, are associated with phenotypic switches. GroupG had 4 significant nonsense-genes (<.01), all associated with asymptomatic carriage. In group B2, 4 genes were observed to be associated, with infection with another 2 associated with asymptomatic carriage.

Discussion

Across all lineages the nonsense ancestry patterns were determined for >98% of genes with nonsense mutations in >1 genome (278-388 genes). However, genes were later filtered due to uncertainty at some nodes during ASR. Development is ongoing on how to treat these sites. Development is also ongoing into breaking ancestry due to differential patterns of homologous recombination. Finally, work is ongoing into determining the encoding of very-similar indels, compensatory mutations, reversals and determining the most appropriate ASR models to infer ancestral phenotypes. At present, runtime ranged from 8-21 minutes. However, the rate–limiting step of ASR with pastML is readily parallelisable

1. Sokurenko, E. V., Hasty, D. L. & Dykhuizen, D. E. Pathoadaptive mutations: Gene loss and variation in bacterial pathogens. *Trends in Microbiology* (1999). doi:10.1016/S0966-842X(99)01493-6
2. Didelot, X., Walker, A. S., Peto, T. E., Crook, D. W. & Wilson, D. J. Within-host evolution of bacterial pathogens. *Nature Reviews Microbiology* (2016). doi:10.1038/nrmicro.2015.13
3. Hottes, A. K. *et al.* Bacterial Adaptation through Loss of Function. *PLoS Genet.* **9**, (2013).
4. Hawley, J., Monk, J. M., Billman-Jacobe, H., Palsson, B. & Holt, K. E. Impact of insertion sequences on convergent evolution of Shigella species. *PLoS Genet.* (2020). doi:10.1371/journal.pgen.1008931
5. Bawn, M. *et al.* Evolution of salmonella enterica serotype typhimurium driven by anthropogenic selection and niche adaptation. *PLoS Genet.* (2020). doi:10.1371/journal.pgen.1008850
6. Read, T. D. & Massey, R. C. Characterizing the genetic basis of bacterial phenotypes using genome-wide association studies: A new direction for bacteriology. *Genome Medicine* (2014). doi:10.1186/s13073-014-0109-z
7. Chen, P. E. & Shapiro, B. J. The advent of genome-wide association studies for bacteria. *Curr. Opin. Microbiol.* **25**, 17–24 (2015).
8. Power, R. A., Parkhill, J. & De Oliveira, T. Microbial genome-wide association studies: lessons from human GWAS. *Nature Reviews Genetics* (2016). doi: 10.1038/nrg.2016.132
9. Didelot, X., Lawson, D., Darling, A. & Falush, D. Inference of homologous recombination in bacteria using whole-genome sequences. *Genetics* (2010). doi:10.1534/genetics.110.120121
10. Medini, D., Donati, C., Tettelin, H., Masignani, V. & Rappuoli, R. The microbial pan-genome. *Current Opinion in Genetics and Development* (2005). doi:10.1016/j.gde.2005.09.006
11. Asim Ansari, M. & Didelot, X. Bayesian inference of the evolution of a phenotype distribution on a phylogenetic tree. *Genetics* (2016). doi:10.1534/genetics.116.190496
12. Didelot, X. & Maiden, M. C. J. Impact of recombination on bacterial evolution. *Trends in Microbiology* (2010). doi:10.1016/j.tim.2010.04.002
13. Jombart, T., Devillard, S. & Balloux, F. Discriminant analysis of principal components: A new method for the analysis of genetically structured populations. *BMC Genet.* (2010). doi:10.1186/1471-2156-11-94
14. Weinert, L. A. *et al.* Genomic signatures of human and animal disease in the zoonotic pathogen Streptococcus suis. *Nat. Commun.* (2015). doi:10.1038/ncomms9740
15. Chewapreecha, C. *et al.* Comprehensive Identification of Single Nucleotide Polymorphisms Associated with Beta-lactam Resistance within Pneumococcal Mosaic Genes. *PLoS Genet.* **10**, (2014).
16. Howell, K. J. *et al.* The use of genome wide association methods to investigate pathogenicity, population structure and serovar in Haemophilus parasuis. *BMC Genomics* (2014). doi:10.1186/1471-2164-15-1179
17. Lees, J. A. *et al.* Sequence element enrichment analysis to determine the genetic basis of bacterial phenotypes. *Nat. Commun.* (2016). doi:10.1038/ncomms12797
18. Earle, S. G. *et al.* Identifying lineage effects when controlling for population structure improves power in bacterial association studies. *Nat. Microbiol.* (2016). doi:10.1038/nmicrobiol.2016.41
19. Brynildsrud, O., Bohlin, J., Scheffer, L. & Eldholm, V. Rapid scoring of genes in microbial pan-genome-wide association studies with Scoary. *Genome Biol.* **17**, (2016).
20. Mageiros, L. *et al.* Genome evolution and the emergence of pathogenicity in avian Escherichia coli. *Nat. Commun.* (2021). doi:10.1038/s41467-021-20988-w
21. Seeman, T. Snippy-2.0. Available at: http://www.vicbioinformatics.com/software.snippy.shtml.
22. Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *80–92* (2012).
23. Nguyen, L. T., Schmidt, H. A., Von Haeseler, A. & Minh, B. Q. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).
24. Hoang, D. T., Chernomor, O., Von Haeseler, A., Minh, B. Q. & Vinh, L. S. UFBoot2: Improving the ultrafast bootstrap approximation. *Mol. Biol. Evol.* (2018). doi:10.1093/molbev/msx281
25. Didelot, X. & Wilson, D. J. ClonalFrameML: Efficient Inference of Recombination in Whole Bacterial Genomes. *PLoS Comput. Biol.* (2015). doi:10.1371/journal.pcbi.1004041
26. Huerta-Cepas, J., Serra, F. & Bork, P. ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. *Mol. Biol. Evol.* **33**, 1635–1638 (2016).
27. Ishikawa, S. A., Zhukova, A., Iwasaki, W., Gascuel, O. & Popko, T. A Fast Likelihood Method to Reconstruct and Visualize Ancestral Scenarios. *Mol. Biol. Evol.* (2019). doi:10.1093/molbev/msz131

FUNDED BY

The research was funded by the National Institute for Health Research Protection Research Unit (NIHR HPRU) in Gastrointestinal Infections at University of Liverpool in partnership with the UK Health Security Agency (UKHSA), in collaboration with University of Warwick. The views expressed are those of the author(s) and not necessarily the NIHR, the Department of Health and Social Care or the UK Health Security Agency.