Comparative and Functional Genomics Identifies Major Differences Between Genomic Islands in Soft Rotting Enterobacterial Plant Pathogens[1,2]

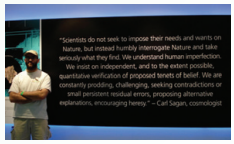
Leighton Pritchard[1], Gunnhild Takle[1,2], Hui Liu[1], Sonia Humphris[1], Lucy Moleleki[1,3], Eduard Venter[1,3], Ron Wheatley[1], Jacques Schrenzel[4], Peter Hedley[1], Paul Birch[1] and Ian Toth[1].

[1] SCRI, Errol Road, Invergowrie, Dundee DD2 5DA.

[2] Bioforsk, Norwegian Institute for Agricultural and Environmental Research, N-1432 Ås, Norway

[3] University of Johannesburg, Aukland Park, 2006, South Africa

[4] Genomic Research Laboratory, Infectious Diseases Service, Geneva University Hospitals and the University of Geneva, Switzerland



Background

The soft rotting enterobacterial plant pathogens *Pectobacterium atrosepticum* (*Pba*), *Pectobacterium carotovorum* (*Pcc*) and *Dickeya dadantii* (*Dda*) are closely-related but differ in their host ranges, geographical distributions, and survival in the environment. Each bacterium causes disease by a similar mechanism, namely the production of plant cell wall degrading enzymes, but the molecular interactions and processes that distinguish between the course of disease in each case are largely unknown. We investigated the extent of differential horizontal gene transfer in these pathogens, using computational and microarray comparative genomic hybridisation (M-CGH) techniques to identify genomic islands in *Pba* strain SCRI1043 that are absent or divergent in *Pcc* strain SCRI193 and/or *Dda* strain 3937. Such islands may make a contribution to *Pba*1043-specific phenotypes, niche adaptation or pathogenicity.

Rationale

It is anticipated that differences between the genomes of related bacteria will reflect adaptation to their particular environmental niche; differences between the genomes of *Pba*, *Pcc* and *Dda* species should reflect the differences between their infection strategies and host interactions. The *Pba*1043 and the *Dda*3937 genomes have been sequenced and annotated, permitting direct sequence comparisons, but no such sequence is available for *Pcc* species. We developed a microarray representing 4450 CDS from the *Pba*1043 genome, which enables M-CGH of unsequenced *Pcc* with the *Pba*1043 sequence by challenge of the microarray with *Pcc* genomic DNA (gDNA). This comparison reveals which sequences in the *Pba*1043 genome are likely to be present or divergent/absent in *Pcc*. We validated the predictive method against the sequenced strain *Dda*3937, and predictions for CDS of interest were validated by Southern hybridisation.

M-CGH Hybridisation

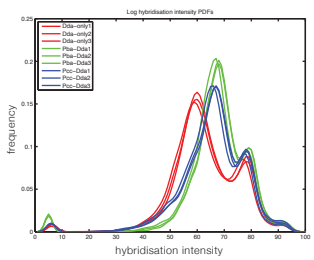


Figure 1:
Plots of hybridisation strengths for *Dda* gDNA hybridised to the *Pba* microarray, under three experimental conditions: *Dda*-only (red) and Dda with gDNA from two reference organisms (*Pba* and *Pcc*). Two major peaks are seen indicating strong and weak *Dda* hybridisation. The *Dda*-only weak hybridisation peak is shifted with respect to experiments in which reference gDNA was present.

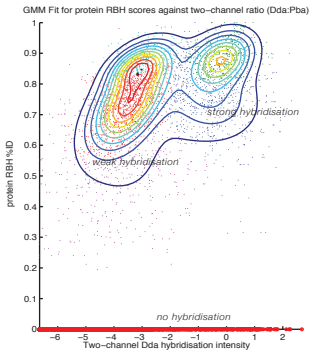


Figure 2:
Pairwise amino acid sequence identity for *Dda* coding sequences (CDS) to their most similar *Pba* homologue (reciprocal best hit: RBH) against the $\log_2$ hybridisation strength to the *Pba* array. There is no simple relationship between hybridisation strength and %identity. Two large centres of population density are identified by Gaussian mixture modelling (GMM), corresponding to strong and weak hybridisation. A third population of sequences with no significant sequence identity is plotted along the *x*-axis. Hybridisation strength is not a good predictor of sequence identity, and sequence identity is not a good predictor of hybridisation.

A HMM-Based Predictive Model

We used a combination of microarray hybridisation intensity from the *Pba*1043:*Dda*3937 array comparison, and the presence or absence of putative orthologues (RBH) from a direct genome comparison, ordered on the *Pba*1043 genome sequence, to construct a first-order hidden Markov model (HMM) to predict individual genes that are expected to be divergent between *Dda*3937 and *Pba*1043. Use of the HMM improved predictive performance, when compared to predictions of divergence based on hybridisation intensity (Table 1). This model was used to predict CDS in *Pcc*193 that are likely to be divergent from *Pba*1043 (Figure 3).

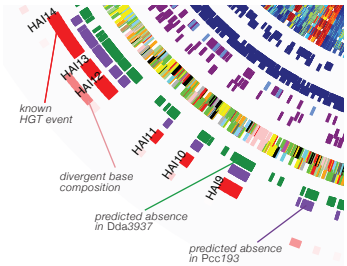


Figure 3:
GenomeDiagram[3] image of predictions of divergent CDS in *Pcc*193 (purple blocks) and *Dda*3937 (green blocks) compared to a region of the *Pba*1043 genome, using the derived HMM. Predictions correspond well to known HGT events (red blocks), and to regions of divergent base composition (as found with `alien_hunter`[4]; outer red blocks).

	hybridisation only	HMM
specificity	0.39-0.42	0.62
sensitivity	0.852-0.858	0.855
correct prediction rate	0.56-0.58	0.769
false positive rate	0.58-0.61	0.32

Table 1:
Predictive performance of a range of models based on optimal hybridisation score cutoff, and performance of the derived HMM at similar sensitivity, on the *Dda*3937 genome, using predicted RBH as the reference.

Species / strains	Host	cfa6	cfa7	1487	1488	2106	2109	ehpF	nifA	nifJ	0482
Pectobacterium atrosepticum											
1039, 1043 (Sco); 1140, 1143, 1147 (Nor); 41 (Net); 1086 (Can); 13, 31 (USA); 44 (Aus); 84, 87 (Per); 1054 (Isr); 4 (Zim)	Potato	+	+	+	+	+	+	+	+	+	+
9 (UK)	Tomato	+	+	+	+	+	+	+	+	+	+
45 (USA)	Sugar beet	+	+	+	+	+	+	+	+	+	+
1140 (Nor)	Potato	+	+	-	-	+	+	+	+	+	+
*5 (Tan)	Schizanthus	-	-	-	-	-	+	-	-	+	+
*27 (USA)	Rocket larkspur	-	-	-	-	-	+	-	-	-	+
Pectobacterium carotovorum* subsp. *carotovorum											
Prediction (*Pcc*193)		-	-	-	-	+	-	-	-	-	+
212 (UK) ; 108 (Fin); 136 (USA); 177 (Per); 112 (Jap)	Potato	-	-	-	-	-	-	-	-	-	+
205 (Mex); 120 (Uga)	Sunflower	-	-	-	-	-	-	-	-	-	+
193 (USA)	**Potato**	-	-	-	-	-	+	-	-	-	+
111 (Isr)	Tomato	-	-	-	-	-	+	-	-	-	+
290 (USA); 132 (Jap); 348 (Isr)	Carrot	-	-	-	-	-	+	-	-	-	+
116, 235 (Sco)	Swede	-	-	-	-	-	+	-	-	-	+
122 (Tan)	Tomato	-	-	-	-	+	+	-	-	-	+
101 (USA)	Tobacco	-	-	-	-	-	+	-	-	-	+
318 (Sco)	Swede	-	-	-	-	-	-	-	+	+	+
121 (Jam)	Sugar cane	-	-	+	+	-	-	-	-	-	+
***Dickeya* species**											
Prediction (*Dda*3937)		-	-	-	-	+	-	-	+	+	-
4052 (UK); 4044, 4000 (Net); 419, 4039 (Per)	Potato	-	-	-	-	-	-	-	-	-	-
403, 4018, 4071 (USA); 413 (Egy)	Maize	-	-	-	-	-	-	-	-	-	-
4033, 4064 (Jap)	Rice	-	-	-	-	-	-	-	-	-	-
4063' 4080 (EC16 - Fra)	St Paulia	-	-	-	-	-	-	-	-	-	-
4073 (UK)	Carnation	-	-	-	-	-	-	-	-	-	-
4078 (Egy)	Maize	-	-	-	-	-	-	-	+	-	-
4081 (3937 - Fra)	**St Paulia**	-	-	+	-	-	-	-	+	-	-
409 (Den)	Carnation	-	-	+	-	-	-	-	-	-	-
4074 (Ger)	Diffenbachia	+	+	-	-	-	-	-	-	-	-

Table 2:
Validation of the HMM predictor by Southern hybridisation of ten CDS in 18 strains each of *Pba*, *Pcc* and *Dickeya* species. Strains used for M-CGH are indicated in bold face. There is appreciable strain variation across each species, but for the strains used to challenge the array, three false positives and two false negatives are seen, for twenty predictions. Treating the *Pcc*193 and *Dda*3937 hybridisations as proxies for all strains of the corresponding species, the overall correct prediction rate in this experiment is 86/360=0.76. The two *Pba* strains marked with an asterisk (*) were subsequently reclassified as *Pcc*.

Summary

- The prediction of divergent or absent CDS in related organisms using M-CGH, on the basis of hybridisation strength alone, is not very reliable (table 1, figure 2). Figure 1 also indicates that the presence of gDNA from a reference organism modifies hybridisation strength in M-CGH experiments.

- The use of a first-order HMM improved the predictive abilities of the M-CGH experiments significantly, in comparison to prediction on the basis of hybridisation strength alone (table 1).

- Using the HMM constructed from *Dda*3937 hybridisation, we identified 197 islands of contiguous genes on the *Pba*1043 genome that are predicted to be divergent from *Dda*3937. Several of these islands contain genes encoding proteins with functions expected to be relevant to niche-adaptation or host specificity, such as lipopolysaccharide synthesis, coronafacic acid synthesis, sugar transport, polysaccharide synthesis and secretion, and putative phenazine antibiotic synthesis.

- We also used the HMM to predict 80 contiguous islands on the *Pba*1043 genome that are divergent in or absent from *Pcc*193. Several of these islands are also suggestive of niche- or functional-adaptive processes, mostly but not always exhibiting overlap with the *Dda*3937-divergent islands, such as coronafacic acid synthesis, iron transport, polysaccharide synthesis and export, colicin, phenazine biosynthesis, and nitrogen fixation.

- Ten candidate divergent or absent CDS from related species were chosen for validation by Southern hybridisation, and the HMM was found to perform similarly to the *in silico* estimates based on a known genome sequence (table 2).

References

[1] Pritchard et al. (in preparation) "Comparative and functional genomic analysis of major differences between genomic islands of *Pectobacterium atrosepticum* 1043, *Pectobacterium carotovorum* 193 and *Dickeya dadantii* 3937."

[2] Pritchard et al. (in preparation) "A method for HMM-based identification of horizontally-acquired islands using microarray comparative genomic hybridisation"

[3] Pritchard et al. (2006) "GenomeDiagram: a Python package for the visualization of large-scale genomic data" *Bioinformatics*, 22, 616-617

[4] Vernikos & Parkhill (2006) "Interpolated variable order motifs for identification of horizontally acquired DNA", *Bioinformatics*, 22, 2196-2203