Differential genome display

Genomes that have been sequenced completely provide us with the challenge of explaining the differences between the phenotypes of species by their different genomic content. A specific example is the difference between a pathogenic species, such as *Haemophilus influenzae*, and a rather closely related, relatively benign species, such as *Escherichia coli*. We show here how differential genome analysis is well-suited for the rapid identification of special features of organisms and apply it to identify potential virulence factors in *H. influenzae*. By 'subtracting' from *H. influenzae* the genes that have a homolog in *E. coli*, we identify a set of genes that are potentially responsible for the pathogenicity of *H. influenzae* and can serve as *H. influenzae*-specific or virulence-specific drug targets. The positional context of these genes can be visualized by a 'differential genome display' (Fig. 1).

There are 116 genes in *H. influenzae* that have no homolog in *E. coli* but have a known function or have a significant similarity to a gene of another genome (in the limits of the methods used; see Fig. 1 legend). Out of these, 28 appear to be derived from bacteriophages. They are mainly located in two clusters, the largest of which has a significantly higher GC content than the overall composition of *H. influenzae* (Fig. 1, this has also been noted in Ref. 1). In the remainder (88) we search for proteins that are involved in *H. influenzae*'s

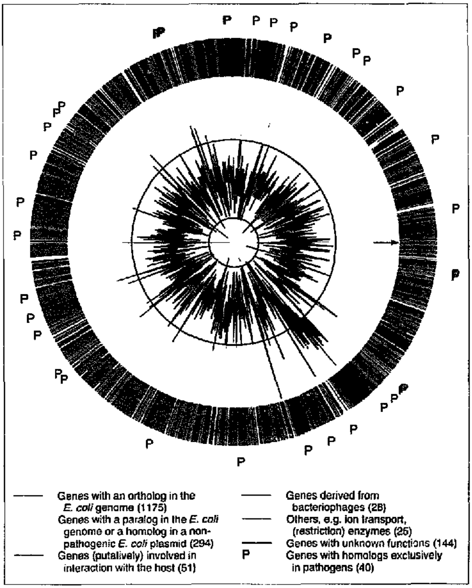

FIGURE 1. Differential genome display of *Haemophilus influenzae* versus *Escherichia coli*. Genes are depicted according to their relative position on the *H. influenzae* genome. The origin of replication is at the right (arrow), the ordering of the genes is counter-clockwise. The terms 'ortholog' and 'paralog' are used as defined in Ref. 6 (i.e. orthologs are genes that have the same function in various species and that have arisen by speciation, paralogs are other members of multigene families). Each gene was assigned to a hierarchical set (detailed below). 'Set 1' was subtracted from the entire genome, then 'set 2' was subtracted from the remainder and so on. Set 1 (green): genes with a putative ortholog in *E. coli* (based on relative similarities with *E. coli* genes, positioning in *H. influenzae* and *E. coli* next to another orthologous pair, and relative similarities of the *H. influenzae* gene with the *E. coli* ortholog 'candidate' and with genes from more distantly related, completely sequenced, genomes (see Ref. 2 for an extensive description of a similar approach). Set 2 (yellow): genes with a paralog in *E. coli* or a homolog in an *E. coli* plasmid not associated with pathogenicity. Set 3 (blue): genes with a homolog in a bacteriophage. Set 4 (red): host-interaction factors in *H. influenzae* or homologs of genes found exclusively in pathogens. Set 5 (magenta): remaining genes with a (putative) function. Set 6 (black): genes without a putative function. In set 4 we observed eight clusters (genes are considered to be part of a cluster if they are separated by five or fewer other genes), six of the eight clusters in the figure can easily be observed from the overlapping 'P's (each P denotes a gene with homologs that have only been observed in pathogens). The other two are, starting from the origin, located at the 5th P and between the 7th and the 8th P. The lines in the center of the picture depict the GC content of third coding positions in polar coordinates (where the inner and outer rings are two standard deviations from the average GC content for the third coding position). The scale from the center of the circle to the inner edge of the 'genes circle' is 10–60% G+C. For genes that have a GC content in the third coding position that falls outside two standard deviations of the average, the GC contents are depicted in the color of the set to which they belong. The Smith–Waterman algorithm[7] as implemented in the s-search program[8] was used to determine homology between genes in *H. influenzae* (Refs 1, 9) and *E. coli* (Refs 10, 11) [using an E-value (expected function of false positives) cut-off of 0.02]. The remaining genes in *H. influenzae* were compared with the protein coding potential in the *E. coli* DNA sequence to find hits with unannotated genes in *E. coli*. Finally, the remaining genes in *H. influenzae* were subjected to a more thorough analysis (see Ref. 12 for details). An annotated list of the proteins in *H. influenzae* that do not have a homolog in *E. coli* is available (http://www.bork.embl-heidelberg.de/Genome/HIEC).

PII: S0168-9525(97)01255-9

interaction with its host. More than half of them (51) fulfill the criteria of being exclusively found in pathogens (40) and/or having a virulence factor-associated function (41) (e.g. surface proteins or proteins involved in the biosynthesis of toxins). These results extend the qualitative analysis of a comparison of *H. influenzae* with a subset of the *E. coli* proteins in Ref. 2. Virulence factors have often been observed to be clustered in pathogenicity islands (PAIS) (e.g. in pathogenic *E. coli*[3] and in *Salmonella typhimurium*[4]) with a GC content that differs from that of the rest of the genome. In *H. influenzae*, 19 out of the 51 'non-*E. coli* homolog' host-interaction factors lie in eight clusters, whereas seven have a GC content in third coding positions that falls outside two standard deviations of the average. Although the set of genes without a homolog in *E. coli* contains a high fraction of potential virulence factors, there are also a few genes among the homologs of *E. coli* proteins that contribute to *H. influenzae*'s pathogenicity. A case in point is the urease operon, that has been implicated in virulence in *Helicobacter pylori* and other pathogens[5]. Three out of a total of seven genes in the *H. influenzae* urease operon have homologs in the *E. coli* chromosome. These three are, however, more closely related to genes from the urease operon in more distantly related, pathogenic, species. The fraction of potential host-interaction factors in the *H. influenzae* sequences that have a homolog in *E. coli* is relatively small. Using the same criteria as we used to find potential host-interaction factors above, we observed 12 potential host-interaction factors in a random sample of 100 annotated *H. influenzae* proteins (in http://www.TIGR.ORG/) that have a homolog in *E. coli*. Only one of these 12 (HI0964), was explicitly annotated as a putative virulence factor.

Although our case study has focused mainly on differences between the genomes correlated with the pathogenicity of *H. influenzae*, one can of course also focus on other differences, such as secondary metabolism. Provided that the genomes of two reasonably related species that have a different phenotype are available, the approach described here will filter the genes within hours.

Martijn A. Huynen
huynen@embl-heidelberg.de

Yolande Diaz-Lazcoz
ydiaz@embl-heidelberg.de

Peer Bork
bork@embl-heidelberg.de

European Molecular Biology Laboratory, Meyerhofstrasse 1, 69012 Heidelberg, Germany and the Max-Delbrück-Centrum for Molecular Medicine, 13122 Berlin-Buch, Germany.

References

1 Fleishmann, R. *et al.* (1995) *Science* 269, 496–512
2 Tatusov, R.L. *et al.* (1996) *Curr. Biol.* 6, 279–291
3 Blum, G. *et al.* (1994) *Infect. Immun.* 62, 606–614
4 Shea, J.E., Hensel, M., Gleeson, C. and Holden, D.W. (1996) *Proc. Natl. Acad. Sci. U. S. A.* 93, 2593–2597
5 Mobley, H.L., Island, M.D. and Hausinger, R.P. (1995) *Microb. Rev.* 59, 451–480
6 Fitch, W.M. (1970) *Syst. Zool.* 19, 99–110
7 Smith, T. and Waterman, M.S. (1981) *J. Mol. Biol.* 147, 195–197
8 Pearson, W.R. (1991) *Genomics* 11, 635–650
9 Sequences from ncbi.nlm.nih.gov
10 Blattner, F. *et al.* (1997) *Science* 277, 1453–1462
11 Sequences from ftp.genetics.wisc.edu
12 Bork, P. and Gibson, T. (1996) *Methods Enzymol.* 266, 162–184