Supplementary Figure 1: Scatter plot of percentage sequence identity for reciprocal best hits (RBH) for coding sequences (CDS) in Pba1043 to Dda3937 by reciprocal best FASTA analysis (protein, x-axis) and reciprocal best BLASTN analysis (nucleotide, y-axis). CDS that do not have a RBH with either method are assigned a sequence identity of zero. Four classes of Pba1043 CDS can be distinguished: 451/4450 make RBH at both protein and nucleotide levels; 7/4450 make RBH only at the nucleotide level; 2369/4450 make RBH only at the protein level; and 1623/4450 make no reciprocal best hits. The majority of Pba1043 CDS have an RBH with Dda3937, but approximately one third of the genome does not. There are very few sequences that are similar at the nucleotide, but not the protein level (suggestive of positive selection), but many that are similar at the protein, and not the nucleotide level (suggestive of neutral drift). The ‘core’ conserved set of CDS comprises only approximately 10% of the Pba1043 genome.