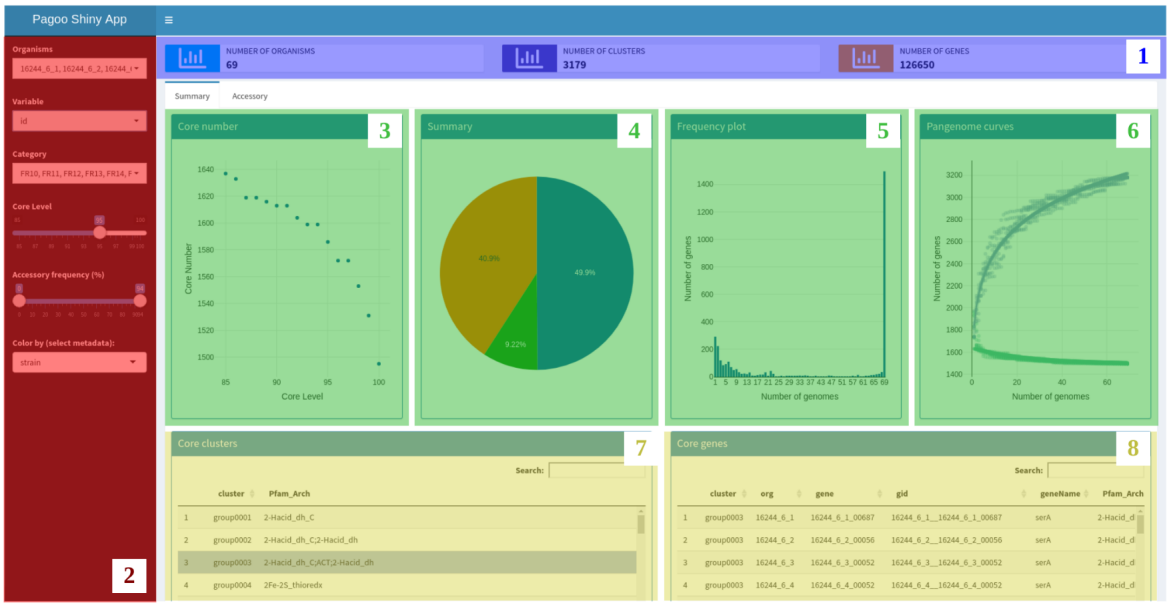
Cell Reports Methods, Volume *1*

Supplemental information

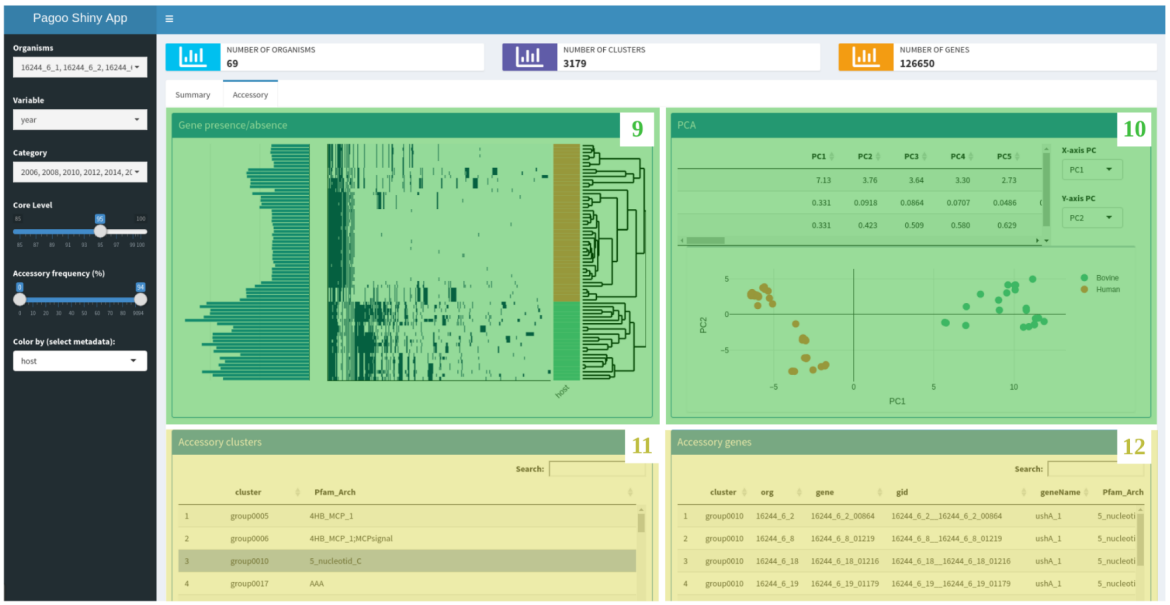
An object-oriented framework for evolutionary pangenome analysis

Ignacio Ferrés and Gregorio Iraola

A)

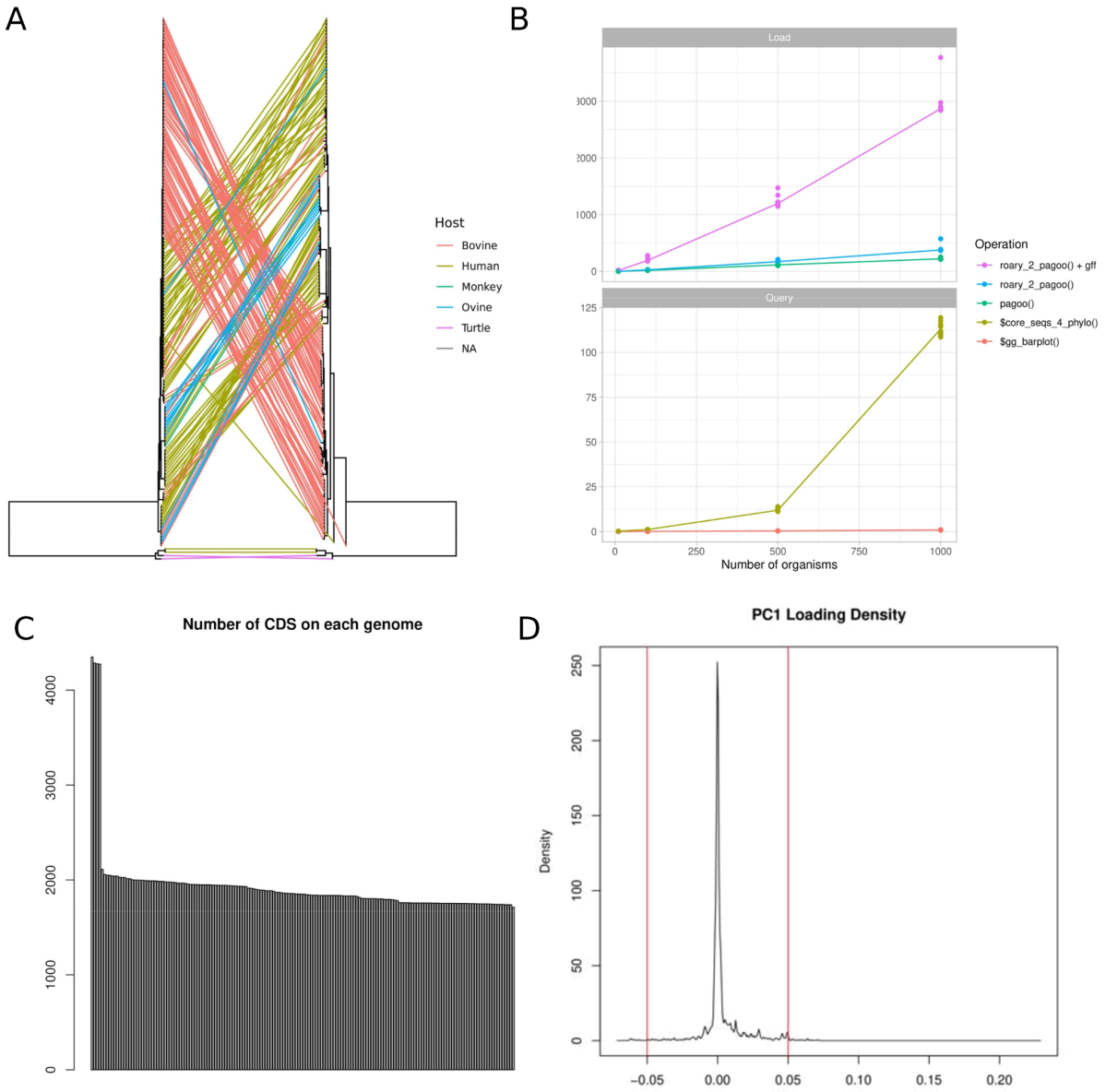


B)



Supplementary Figure 1. Overview of Pagoo Shiny Application. Related to Figure 1. A) Dashboard showing summary statistics and core genome characteristics of the dataset under analysis (Summary tab). The header is permanently displayed and shows the number of organisms, the number of orthologous clusters and the number of individual genes under analysis (panel 1). The left menu (panel 2) has several dropdown and scrollable menus that control parameters whose values affect plots and information displayed in the dashboard. The "Organisms" menu lists names for all organisms in the analysis and the user can select or deselect them. If the user deselects one or several organisms, they will be excluded from the analysis and plots will be automatically updated without that information. The "Variable" menu lists all metadata columns associated with organisms present in the pangenome object. This menu is associated with the "Category" menu which displays all possible values for any selected variable. For example, in the provided dataset, the user can select "host" in the "Variable" menu so the "Category" menu will display the host of each organism. Then, the user can filter organisms for example by keeping only those having human host by deselecting the "Bovine" value. The "Core level" slide bar controls the percentage of organisms that must contain a certain gene for this to be considered a core gene. For example, if the core level is set to 90, a certain gene needs to be present in at least 90% of organisms (in this case 62 out of 69) to be counted as a core gene. The "Core number" (panel 3) displays the number of retained core genes at

different core levels. As expected, a more stringent core level (near 100%) will result in a smaller set of core genes. Hovering over dots will display a label in the format "(95, 1586)", with the first number being the core level and the second being the number of core genes. The "Summary" pie chart (panel 4) shows the percentage of core genes defined as those who appear in a frequency higher than the core level, the percentage of shell genes defined as those accessory genes present in more than 1 genome, and the percentage of cloud genes defined as genome specific genes. Hovering over the pie chart will display a label showing the number of genes and percentages corresponding to these subsets. The "Frequency plot" (panel 5) displays the number of genes while adding genomes in the dataset. This plot typically shows a "U" shape that describes the distribution of genes in the pangenome. The "Pangenome curves" plot (panel 6) shows the cumulative number of genes as genomes are added in the dataset that tend to adjust to a power law function. This is useful to rapidly explore if a certain pangenome is open or closed. This plot also shows the number of core genes as genomes are added in the dataset that tend to adjust to an exponential decay function. The "Core clusters" scrollable menu (panel 7) lists names of all core genome clusters and displays associated metadata like cluster annotation, etc. The "Core genes" scrollable menu (panel 8) responds to the previous "Core clusters" menu. Once a certain core cluster is selected, those individual core genes belonging to the cluster are displayed in the "Core genes" menu. For each gene, information like start and end position in the genome, annotation, gene name and strand are displayed. B) Dashboard showing summary statistics and accessory genome characteristics of the dataset under analysis (Accessory tab). In the left menu, the "Accessory frequency" slide bar controls the minimum and maximum frequency for any certain accessory gene to be considered in the analyses. The "Gene presence/absence" plot (panel 9) displays the gene presence/absence matrix highlighting present accessory genes in each genome in blue and absent genes in white. The rows (organisms) are ordered according to a clustering analysis displayed on the right side that is based on the Bray-Curtis distance calculated over the presence/absence matrix. The bar plots on the left show the number of the accessory genes in each genome. A vertical colored strip highlights each genome according to any metadata that the user selects in the "Color by" menu. Gene presence/absence blocks can be explored in detail by selecting a desired area of the plot that will zoom in automatically. The "PCA" plot (panel 10) is based on the sample gene presence/absence matrix but users can select the principal components to be displayed. Also, information about the contribution of each principal component is shown at the top. Points can be colored by selecting metadata in the "Color by menu". "Accessory clusters" and "Accessory genes" panels work as explained for panels 7 and 8, respectively.



Supplementary Figure 2. Supporting analyses. Related to Figure 2. A) Tanglegram between mcp4 gene phylogeny and the core genome phylogeny. Links are colored by the host associated with each isolate. B) Scalability of Pagoo was tested by measuring the time to perform certain operations as a function of the number of genomes included in the pangenome. Operations were divided into "Load" operations (top) and "Query" operations (bottom). In the "Load" panel: time to load a Pagoo object from a roary output gene_presence_absence.csv file and including the gff3 information (violet), time to load a Pagoo object from the roary output but only the genes_presence_absence.csv file (sky-blue), time to create a Pagoo object but with all input already loaded into the R session (green). In the "Query" panel: time to retrieve the core gene sequences (yellow), and time to plot a frequency plot (orange). C) Number of CDS per genome, sorted in decreasing order. The four biggest genomes, with an abnormal number of CDS, were removed from the dataset. D) Loadings density plot of the first principal component. Based on this graph, we define as "highly discriminative" those genes whose loadings were lesser than -0.05 and greater than 0.05.

Supplementary Table S1. Related to Figure 2. Genes identified as having a highly discriminative distribution among bovine or human isolates.

group_1140	putative nicotinate-nucleotide pyrophosphorylase
barS1	[carboxylating]
cirA_2	hypothetical protein
dapH	hypothetical protein
dctA_2	Competence protein ComM
dpnM_2	Serine/threonine transporter SstT
exsA	hypothetical protein
fabG_1	hypothetical protein
fabG_2	Imidazoleglycerol-phosphate dehydratase
fmt_2	hypothetical protein
glf	Type IV secretion system protein virB8
group_1024	hypothetical protein
group_1026	hypothetical protein
group_1350	hypothetical protein
group_1149	FhuE receptor
group_1153	hypothetical protein
group_1282	hypothetical protein
group_1350	Arginine transport ATP-binding protein ArtM group_1359 hypothetical protein
group_1371	hypothetical protein
group_1565	hypothetical protein
group_1371	2-iminoacetate synthase
group_1760	hypothetical protein
group_1798	hypothetical protein
group_1799	hypothetical protein
group_1859	hypothetical protein
group_1923	4-hydroxy-tetrahydrodipicolinate synthase
group_2327	hypothetical protein
group_2328	Ribosomal RNA small subunit methyltransferase G
group_2508	Endoribonuclease YbeY
group_2880	hypothetical protein
group_3321	hypothetical protein
group_3323	hypothetical protein
group_3325	hypothetical protein
group_3336	hypothetical protein
group_3361	aminotransferase
group_3371	hypothetical protein
group_3372	hypothetical protein
group_3390	D-inositol-3-phosphate glycosyltransferase
group_446	hypothetical protein

group_447	Guanosine-5'-triphosphate,3'-diphosphate pyrophosphatase
group_449	putative acyltransferase YihG
group_544	Prophage integrase IntS
group_455	hypothetical protein
group_643	hypothetical protein
group_472	Zinc-type alcohol dehydrogenase-like protein
group_475	Holliday junction ATP-dependent DNA helicase RuvB
group_764	hypothetical protein
group_482	GTP cyclohydrolase 1 type 2
group_505	hypothetical protein
group_509	Vitamin B12 transporter BtuB
group_512	hypothetical protein
group_1153	hypothetical protein
group_593	hypothetical protein
group_543	hypothetical protein
group_544	DNA adenine methylase
group_593	Uridylate kinase
group_625	hypothetical protein
group_626	5-hydroxyisourate hydrolase
group_627	Hemin receptor
group_643	Single-stranded DNA-binding protein
group_7159	L-xylulose/3-keto-L-gulonate kinase
group_762	hypothetical protein
group_764	LexA repressor
group_766	hypothetical protein
group_830	hypothetical protein
group_832	Multidrug resistance protein MdtE
group_909	Type IV secretion system protein virB2
group_1798	hypothetical protein
group_986	3-oxoacyl-[acyl-carrier-protein] synthase 1
hldD_1	Col shock-like protein CspE
lexA_2	hypothetical protein
lexA_3	hypothetical protein
moaA_1	hypothetical protein
moaA_2	hypothetical protein
pctA	Trifunctional NAD biosynthesis/regulator protein NadR
rfbE	hypothetical protein
trmR_2	Protein RarD