Vignette rKOMICS
application example

Manon Geerts & Frederik Van Den Broeck

January 23, 2026

1 The rKOMICS package

The core features of the rKOMICS package include data aggregation, analyses and visualization that allows to examine, summarize and extract meaningful information from minicircle sequence alignments as obtained by KOMICS or a custom bioinformatic pipeline, and from USEARCH cluster format (UC) files as generated by USEARCH or VSEARCH. In addition to storing data files, rKOMICS stores the analyses and visualization results into single list objects that can be called by the user at a later stage.

rKOMICS incorporates multiple methods of visualizations using the ggplot2 R package to plot the foundation of graphs. By adding ggplot2 functions to the rKOMICS visualization functions, the user has direct control over the finishing touches of the graph's appearances. Our package also utilizes sample-specific metadata that allows multi-group data visualizations to facilitate exploratory analysis. The overall data set can be examined using barplots, heatmaps, PCA plots and box plots that are generated for each specified minimum percent identity. This makes it possible to visualize population structure and diversity based on minicircle sequence composition.

To show the functionality of rKOMICS, we performed an example analysis using whole-genome sequencing data from a recently published study on the history of diversification of the *Leishmania braziliensis* species complex in Peru. This species complex comprises two closely related species: the lowland and zoonotic *L. braziliensis* parasite circulating in a diverse range of wild mammals in Neotropical rainforests, and the highland anthroponotic *L. peruviana* parasite that is largely endemic to the Pacific slopes of the Peruvian Andes. A total of 67 *Leishmania* parasites from 47 localities in Peru were cultured and subjected to whole genome sequencing.

```
data(exData, package = "rKOMICS")
table(exData$species)
```

```
    hybrid L. braziliensis   L. peruviana
        13               23             31
```

2 Required R-packages

```
library(ggplot2)
library(rKOMICS)
library(ggpubr)
library(viridis)
```

3 Quality of the assembly

The `msc.quality` function allows you to examine the quality of the assembly by alignment of reads to the assembled minicircles (see https://github.com/FreBio/komics for tutorial). We found that on average 77% of all mapped reads aligned with a mapping quality larger than 20 (Figure 1) and on average 84% aligned in proper pairs. On average 93% of all CSB3-containing reads aligned against the

assembled minicircle contigs and 88.5% aligned perfectly, suggesting that KOMICS was able to retrieve a large proportion of the minicircle classes.

```
map <- msc.quality(mapstats = system.file("extdata",
                                          exData$mapstats,
                                          package = "rKOMICS"),
                   groups = exData$species)
lapply(map$proportions, mean)$MR_HQ
```

```
[1] 77.04245
```

```
map$plots$MR_HQ + labs(caption = paste0('Proportion_of_mapped_reads_with_
    high_quality,_', Sys.Date()))
```
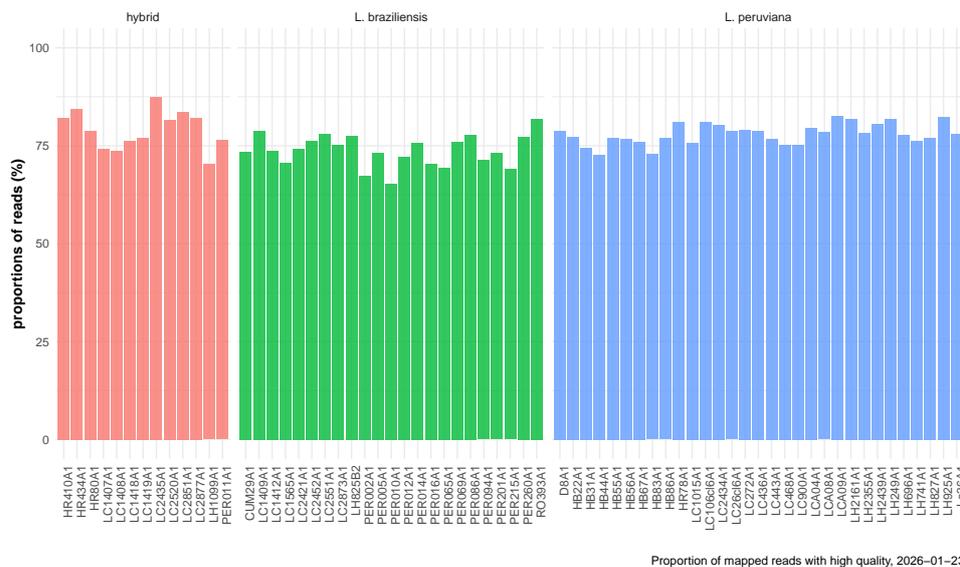

Figure 1: *Proportion of reads with a mapping quality higher than 20 per Leihmania isolate.*

4   Minicircle copy numbers

The depth statistics (see https://github.com/FreBio/komics for tutorial) include average, median, minimum and maximum per site read depth of every minicircle contig that has been assembled. The `msc.depth` function allows you to summarize those statistics (Figure 2) and to estimate minicircle copy numbers by standardizing median read depths per minicircle contig to the median genome-wide read depths.

```
depth <- msc.depth(depthstats = system.file("extdata",
                                            exData$depthstats,
                                            package = "rKOMICS"),
                   groups = exData$species,
                   HCN = exData$medGWD/2) # haploid copy number
depth$CN
```
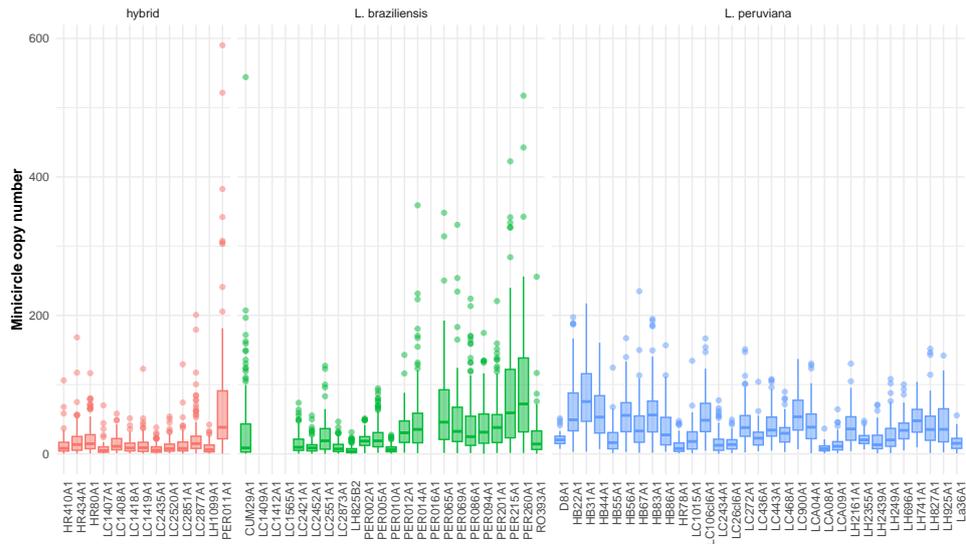
Figure 2: *Minicirle copy number distribution per Leihmania isolate.*

5 Inspect minicircle lengths

A combined total of 7,760 minicircles were assembled for 67 *Leishmania* isolates. When examining the length distribution of the circularized minicircle sequences using the function `msc.length`, we found that the majority of minicircles (95.2%) were 720-760 bp long, which is within the expected length range of minicircles in *Leishmania* parasites (Figure 3). 294 minicircle contigs (4.8%) showed twice this length (1400-1700 bp) (Figure 3), which may suggest that these are artificial minicircle dimers introduced by the assembly process, and were removed.

```
bf <- msc.length(file = system.file("extdata",
                                    "all.minicircles.fasta",
                                    package = "rKOMICS"),
                 samples = exData$samples,
                 groups = exData$subspecies)
bf$plot
```
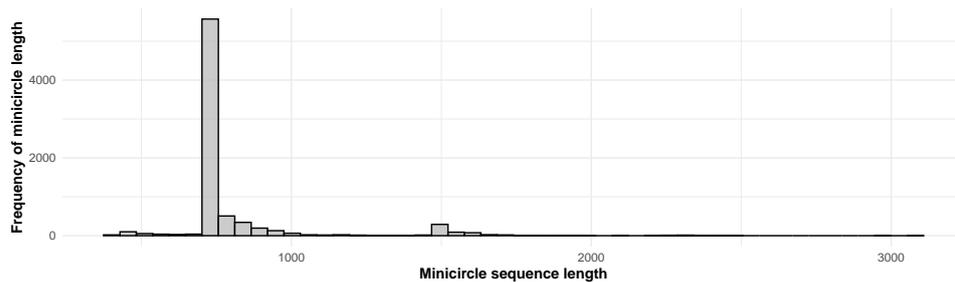

Figure 3: *Length distribution of the 7,760 assembled minicircles in 67 Leishmania isolates.*

```
c(length(bf$length),
  length(which(bf$length < 800)),
  length(which(bf$length > 1400)))
```

```
[1] 7760 6329  576
```

6 Filter minicircle sequences

For downstream analyses, we only retained the circularized minicircles of the expected length (720-760bp) using the `preprocess` function (Figure 4; coloured barplots), resulting in a final set of 5,849 minicircles.

```
pre <- preprocess(files = system.file("extdata",
                                      exData$fastafiles,
                                      package = "rKOMICS"),
                  groups = exData$species, circ = TRUE,
                  min = 500, max = 1200,
                  writeDNA = FALSE)
pre$summary
```

```
beforefiltering  afterfiltering
          7760            5849
```

```
pre$plot +
  labs(caption = paste0('N of MC sequences before and after filtering, ',
      Sys.Date()))
```
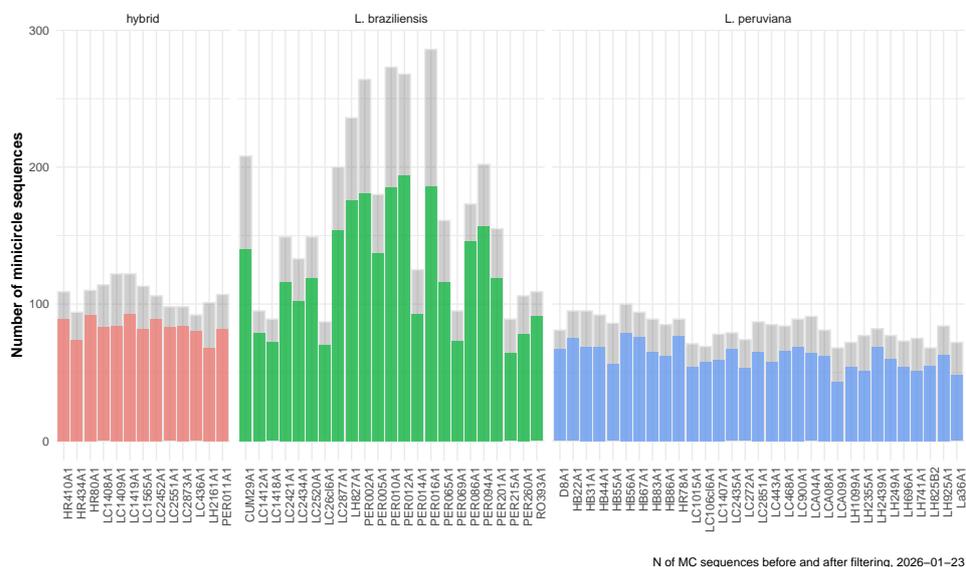

Figure 4: Gray barplots show the total number of minicircles found per Leishmania isolate, and coloured barplots indicate the number obtained after retaining only circularized minicircles of the expected length.

7 Clustering results

We used the function `msc.uc` to then examine the combined number of minicircle sequence classes (MSCs) (based on overall identity) across all 67 isolates, and we identified a total of 3,811 MSCs at 100% identity. This number decreased sharply to 918 MSCs at 97% identity and 603 MSCs at 95% identity (Figure 5). The proportion of perfectly aligned minicircle sequences (i.e. alignments without any insertion/deletion) during the clustering process decreased from 100% (only perfect alignments) at 100% identity to 79% (79% of the alignments were perfect) at 97% identity and 68% at 95% identity (Figure 5).

```
ucs <- msc.uc(files = system.file("extdata", exData$ucs, package = "rKOMICS
    "))
c(ucs$MSCs["100"],ucs$MSCs["97"],ucs$MSCs["95"])
```

```
 100   97   95
3811  918  603
```
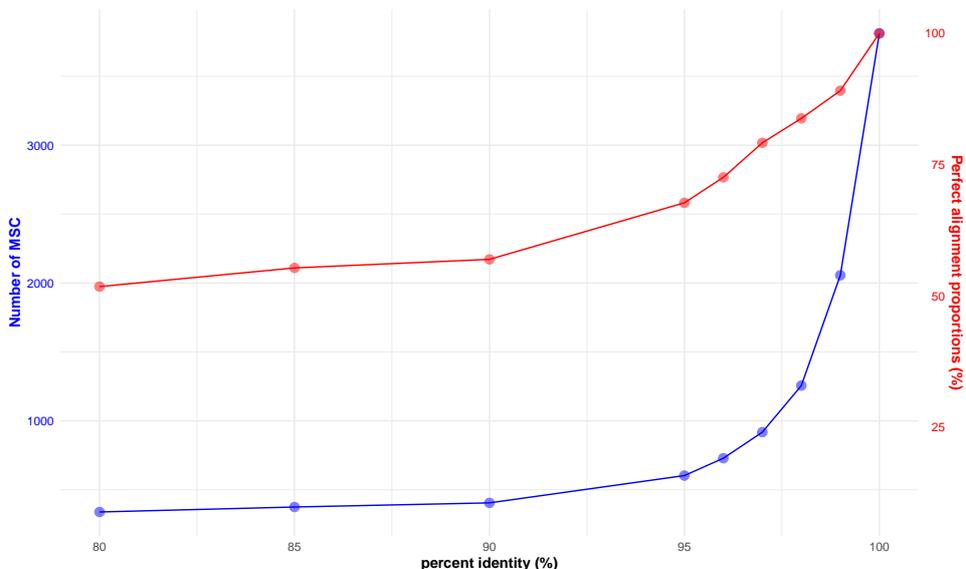
`ucs$plots$‘MSCs and perfect aligments‘`

Figure 5: Number of MSCs (blue) and proportion of perfect alignments (red) as obtained following clustering analyses for a range of percent identities.

While insertions were mostly 1 bp long (Figure 6; left), the number of insertions per alignment increased with decreasing percent identity (Figure 6; right). Most notably, below 97% identity, we found a steady increase in alignments with 3 or 4 insertions (Figure 6; right). Similar results were obtained for deletions (results not shown). Hence, we decided to focus most of our downstream analyses at the 97% identity threshold, as this would capture sufficient minicircle sequence classes (Figure 5) while minimizing the number of alignment gaps (Figure 6).
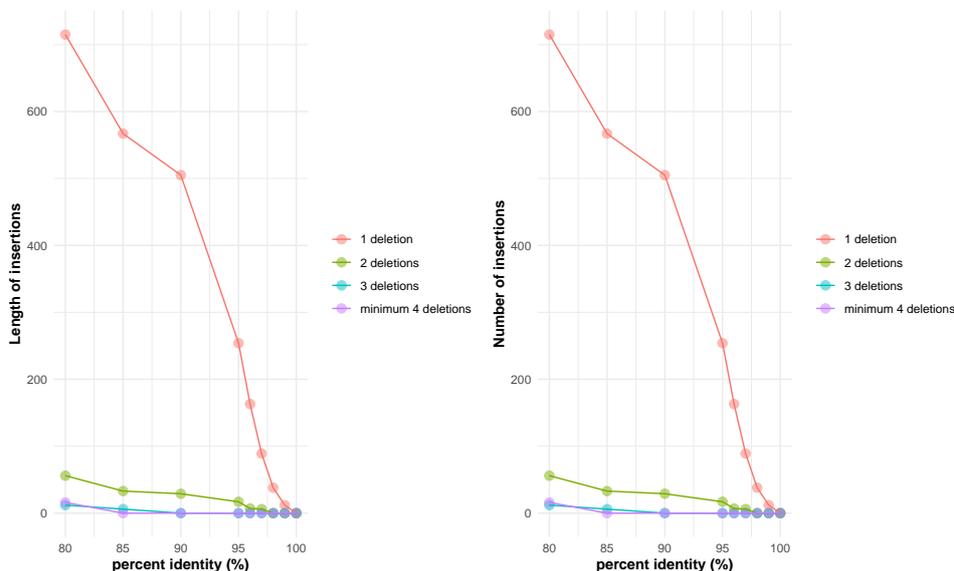
`ucs$plots$insertions`

Figure 6: Length and number of insertions in MSC alignments following clustering analysis for a range of percent identities.

8   Build MSC matrix

Clustering based on a percent identity, performed with the VSEARCH tool, will generate files in uc format. The `msc.matrix` function will transform every input file into a cluster matrix. The columns of

the matrix correspond to the samples and the rows of the matrix correspond to the minicircle sequence cluster (MSC). The absence of a MSC in a sample is indicated with the value of zero while the presence of a MSC in a sample will be indicated with a value $>= 1$.

```
matrices <- msc.matrix(files = system.file("extdata",
                                           exData$ucs,
                                           package = "rKOMICS"),
                       samples = sort(exData$samples),
                       groups = exData$species[order(exData$samples)])
### or: data(matrices, package = "rKOMICS")
# rowSums(matrices[["id97"]]) # --> frequency of MSC across all samples
colSums(matrices[["id97"]]) # --> number of MSC per sample
```

CUM29A1	D8A1	HB22A1	HB31A1	HB44A1	HB55A1	HB56A1
140	67	75	69	69	56	79
HB67A1	HB83A1	HB86A1	HR410A1	HR434A1	HR78A1	HR80A1
76	65	62	89	74	77	92
La36A1	LC1015A1	LC106cl6A1	LC1407A1	LC1408A1	LC1409A1	LC1412A1
54	58	59	83	84	79	72
LC1418A1	LC1419A1	LC1565A1	LC2421A1	LC2434A1	LC2435A1	LC2452A1
93	82	116	102	67	89	119
LC2520A1	LC2551A1	LC26cl6A1	LC272A1	LC2851A1	LC2873A1	LC2877A1
83	70	53	65	84	154	80
LC436A1	LC443A1	LC468A1	LC900A1	LCA04A1	LCA08A1	LCA09A1
58	66	69	64	62	43	54
LH1099A1	LH2161A1	LH2355A1	LH2439A1	LH249A1	LH696A1	LH741A1
68	51	69	60	54	51	55
LH825B2	LH827A1	LH925A1	PER002A1	PER005A1	PER010A1	PER011A1
176	63	48	181	137	185	82
PER012A1	PER014A1	PER016A1	PER065A1	PER069A1	PER086A1	PER094A1
194	93	186	116	73	146	157
PER201A1	PER215A1	PER260A1	RO393A1			
119	64	78	91			

The `msc.heatmap` function generates a heatmap of the input cluster matrix that summarizes the presence or absence of Minicircle Cluster Sequences (MCSs) between groups of samples.

```
msc.heatmap(clustmatrix = matrices[["id97"]],
            groups = exData$species,
            samples = exData$samples)
```
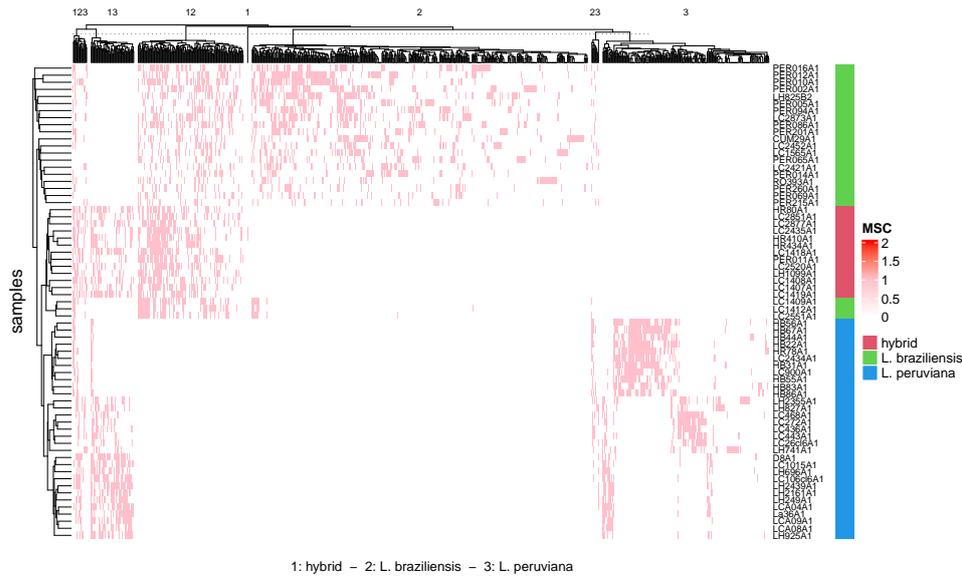
Figure 7: Heatmap of the MSC matrix at a percent identity of 97.

9   MSC Richness

Focusing on the results at 97% identity, we observed that the Andean and near-clonal *L. peruviana* parasites harbored substantially less MSCs (mean = 62 MSCs per isolate) compared to the Amazonian and recombining *L. braziliensis* parasites (mean = 124 MSCs per isolate).

```
richness <- msc.richness(matrices,
                         samples = exData$samples,
                         groups = exData$species)
apply(richness$table[which(richness$table$group=="L. peruviana"),-(1:2)],
    2, mean)
```

```
      id80      id85      id88      id89      id90      id91      id92      id93
  61.80645  61.80645  61.80645  61.80645  61.80645  61.80645  61.80645  61.80645
      id94      id95      id96      id97      id98      id99     id100
  61.80645  61.83871  61.83871  61.87097  61.87097  61.87097  61.87097
```

```
apply(richness$table[which(richness$table$group=="L. braziliensis"),-(1:2)
    ], 2, mean)
```

```
      id80      id85      id88      id89      id90      id91      id92      id93
  117.2174  119.8261  121.0000  121.1739  121.3478  121.4783  121.7826  122.3913
      id94      id95      id96      id97      id98      id99     id100
  122.7826  123.2609  123.5217  123.8261  123.8261  123.8261  123.8261
```

```
apply(richness$table[which(richness$table$group=="hybrid"),-(1:2)], 2, mean
    )
```

```
      id80      id85      id88      id89      id90      id91      id92      id93
  79.69231  80.00000  80.07692  80.46154  80.76923  81.46154  81.38462  81.53846
      id94      id95      id96      id97      id98      id99     id100
  81.84615  82.38462  83.07692  83.30769  83.30769  83.30769  83.30769
```
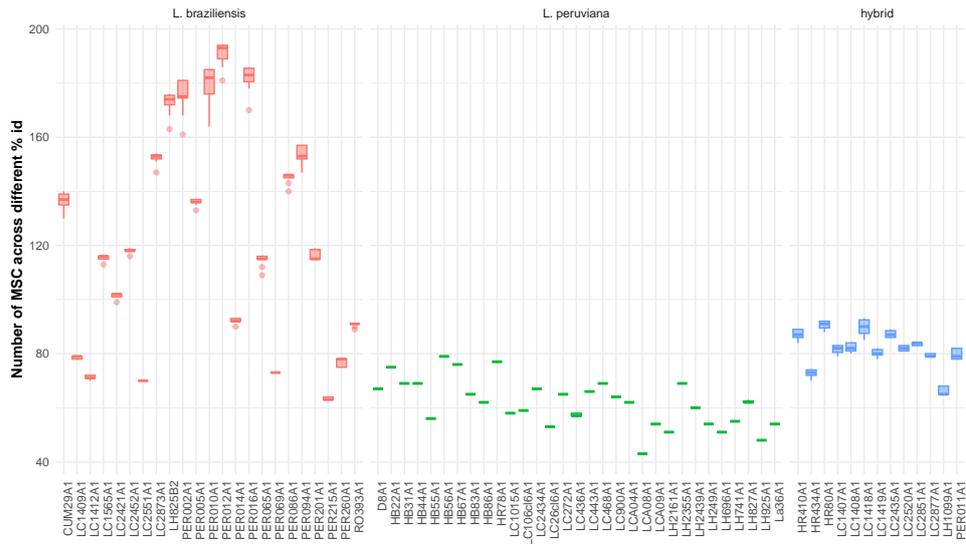
```
richness$plot
```

Figure 8: Number of MSC per isolate across different percent identities.

10 Similarity

Using the function `msc.similarity`, we found that 48.9% and 25.9% of the MSCs were unique to *L. braziliensis* and *L. peruviana*, respectively, while hybrid *L. braziliensis* x *L. peruviana* parasites shared MSCs with both parents (Figure 9). This confirms that hybrid parasites inherited minicircles from both Leishmania parental species. We also confirmed that the Andean and near-clonal *L. peruviana* parasites harbored substantially less MSCs (mean = 62 MSCs per isolate) compared to the Amazonian and recombining *L. braziliensis* parasites (mean = 124 MSCs per isolate).

```
sim <- msc.similarity(matrices,
                      samples = exData$samples,
                      groups = exData$species)
sim$relfreq.plot + scale_fill_viridis(discrete = TRUE)
```
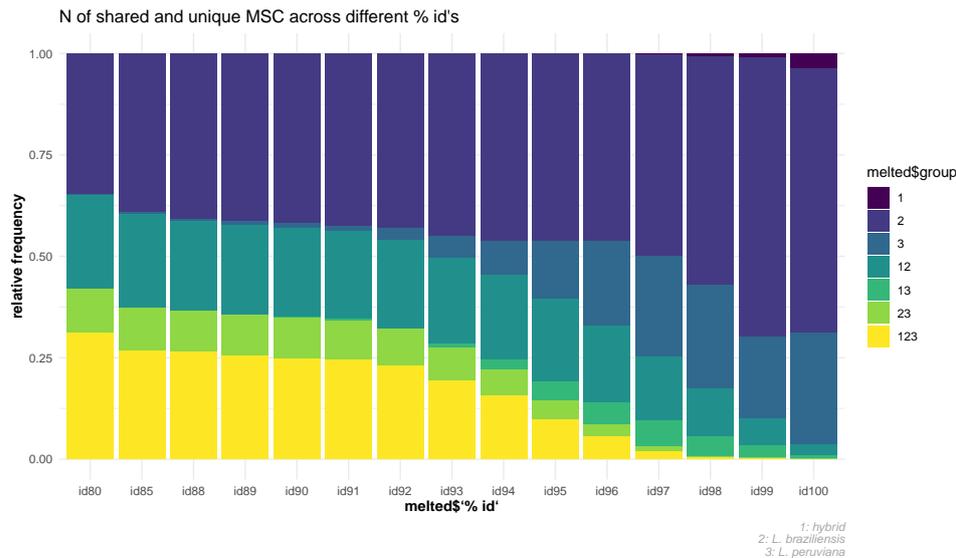

Figure 9: Barplots show the proportion of minicircle sequence classes that are unique or shared between *L. braziliensis*, *L. peruviana* and their hybrids, for each % identity threshold used during the clustering analyses.

```
c(sim$relfreq$id97["2"]*100,
  sim$relfreq$id97["3"]*100)
```

```
       2        3
49.89107 24.72767
```

11 PCA

Principal Component Analysis based on minicircle sequence similarity (i.e. MSC presence/absence per isolate) separated *L. braziliensis* from *L. peruviana* on the first axis and three *L. peruviana* populations on the second axis (Figure 10). The three *L. peruviana* populations correspond to the Porculla lineage that circulates in the tropical deciduous forests of Peru, and the two Surco lineages that circulate in desert shrubland on the Pacific Coast (Surco North/Central and Surco Central/South). Hybrids did not cluster with either parental species, in contrast to what was observed for the uniparentally inherited kinetoplast, but instead occupied an intermediate position between *L. braziliensis* and the *L. peruviana* Surco Central/South lineage (Figure 10), again consistent with mixing of the parental minicircle populations.

```
res.pca <- lapply(matrices, function(x) msc.pca(x, samples = exData$samples
    ,
                   groups = exData$species, n=30, labels=FALSE, title=NULL))
res.pca$id97$plot
```
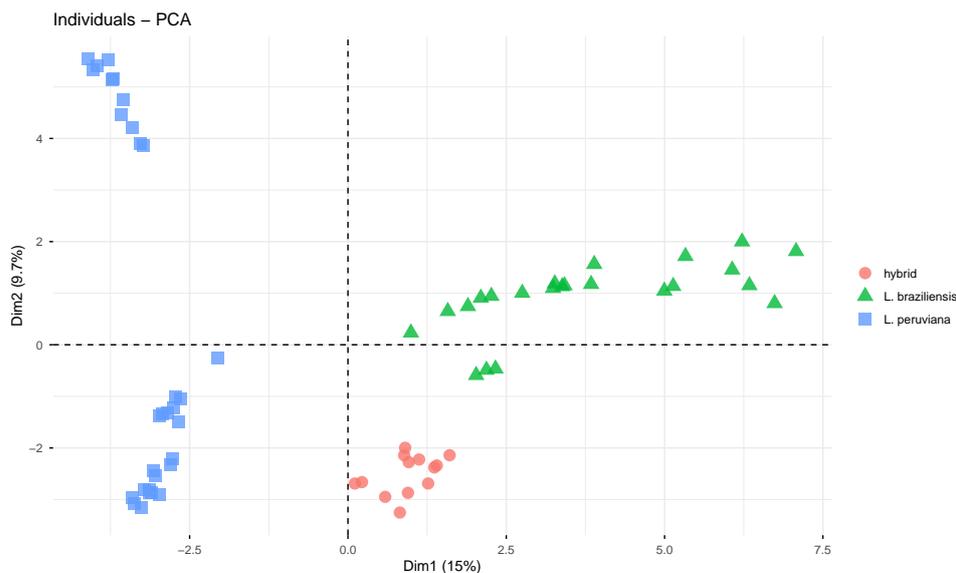

Figure 10: Principal Component Analysis based on sequence similarity between MSCs at 97% identity.