RESEARCH ARTICLE

Figures and figure supplements

Global genome diversity of the *Leishmania donovani* complex

Susanne U Franssen *et al*

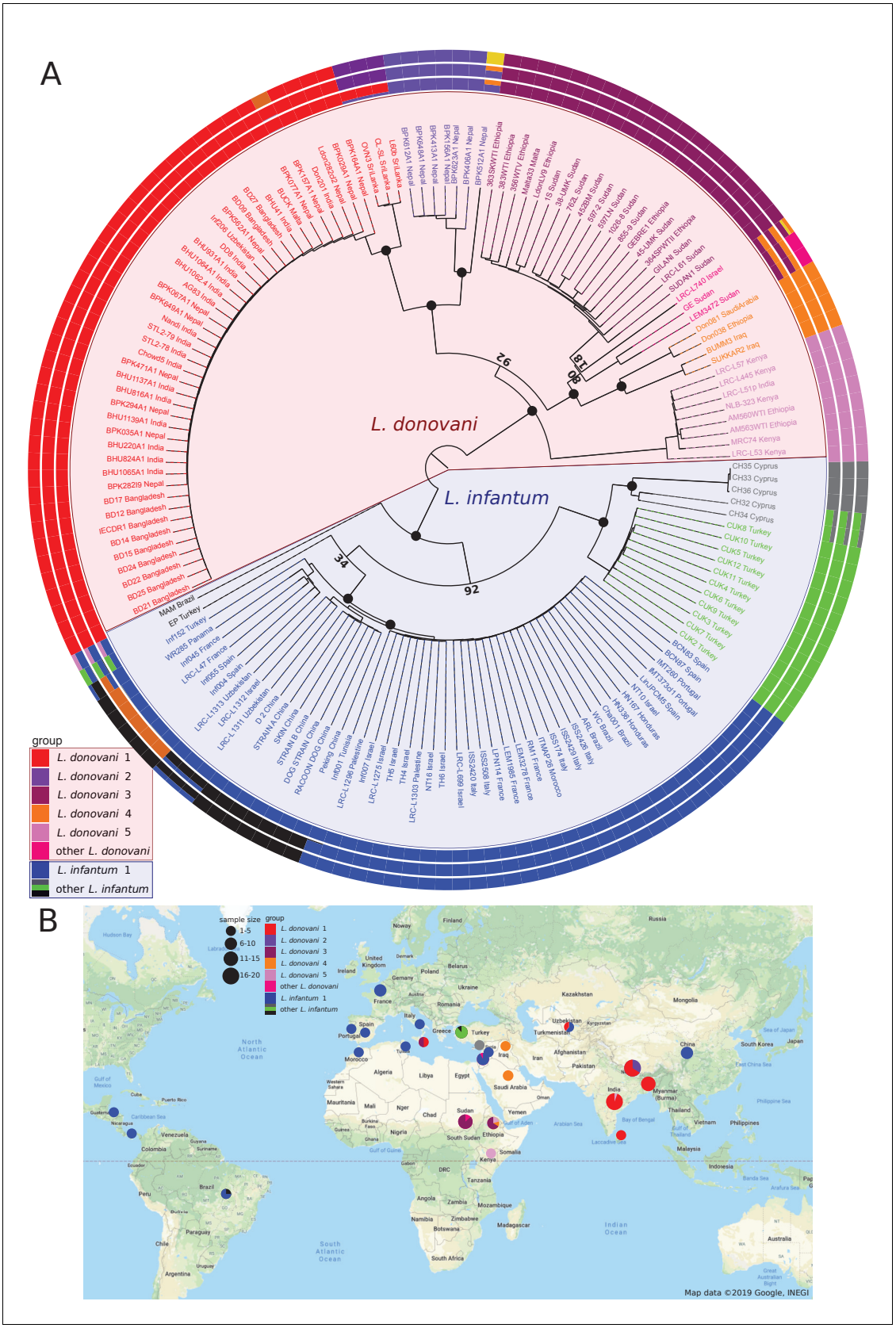

Figure 1. Sample phylogeny and distribution. (**A**) Phylogeny of all 151 samples of the *L. donovani* complex. The phylogeny was calculated with neighbour joining based on Nei's distances using genome-wide SNPs and rooted based on the inclusion of isolates of *L. mexicana* (U1103.v1), *L.*

Figure 1 continued on next page

Figure 1 continued

tropica (P283) and *L. major* (LmjFried) (outgroups not shown in the phylogeny). Bootstrap support is shown for prominent nodes in the phylogeny as black circles for values of 100% and otherwise the respective support value in % based on 1000 replicates. The groupings shown in the outer circles were calculated by admixture with $K = 8$, $K = 11$ and $K = 13$ (see Materials and methods). Groups labelled with different colours were defined based on the phylogeny and include monophyletic groups as well as groups that are polyphyletic and/or largely influenced by hybridisation (indicated by 'other'). (**B**) Map of the sampling locations. Groups are indicated by the different colours. Sample sizes by country of origin are visualised by the sizes of the circles.

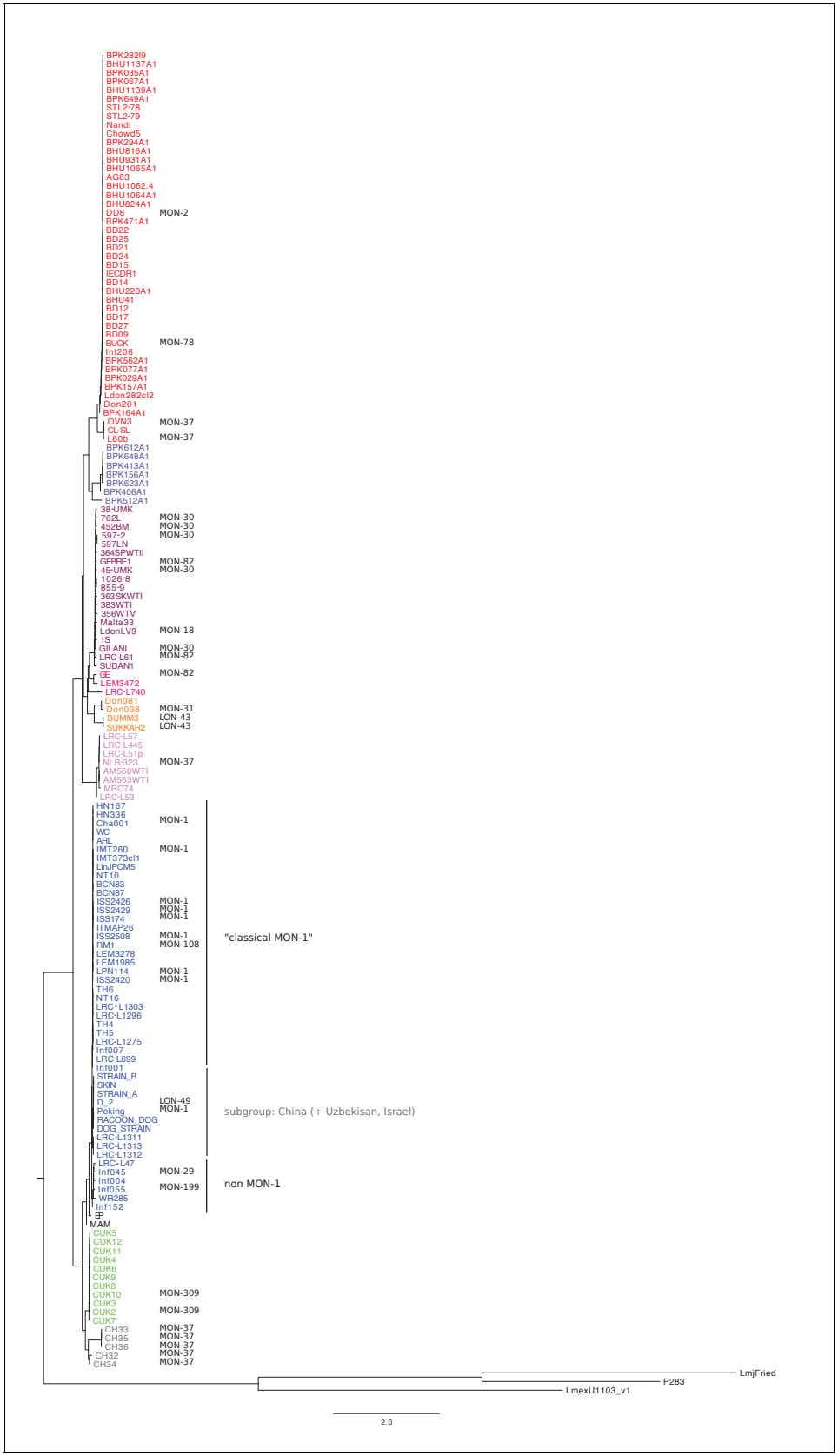

Figure 1—figure supplement 1. Phylogenetic reconstruction of all 151 samples of the *L. donovani* complex. The phylogeny was calculated with neighbour joining based on Nei's distances including isolates of *L. mexicana* (U1103.v1), *L. tropica* (P283) and *L. major* (LmjFried) as outgroups. Group

Figure 1—figure supplement 1 continued on next page

Figure 1—figure supplement 1 continued

colours are identical to those used throughout the manuscript. Phylogenetic grouping as suggested by Multilocus Enzyme Electrophoresis are indicated in black (see ***Supplementary file 1*** for references).

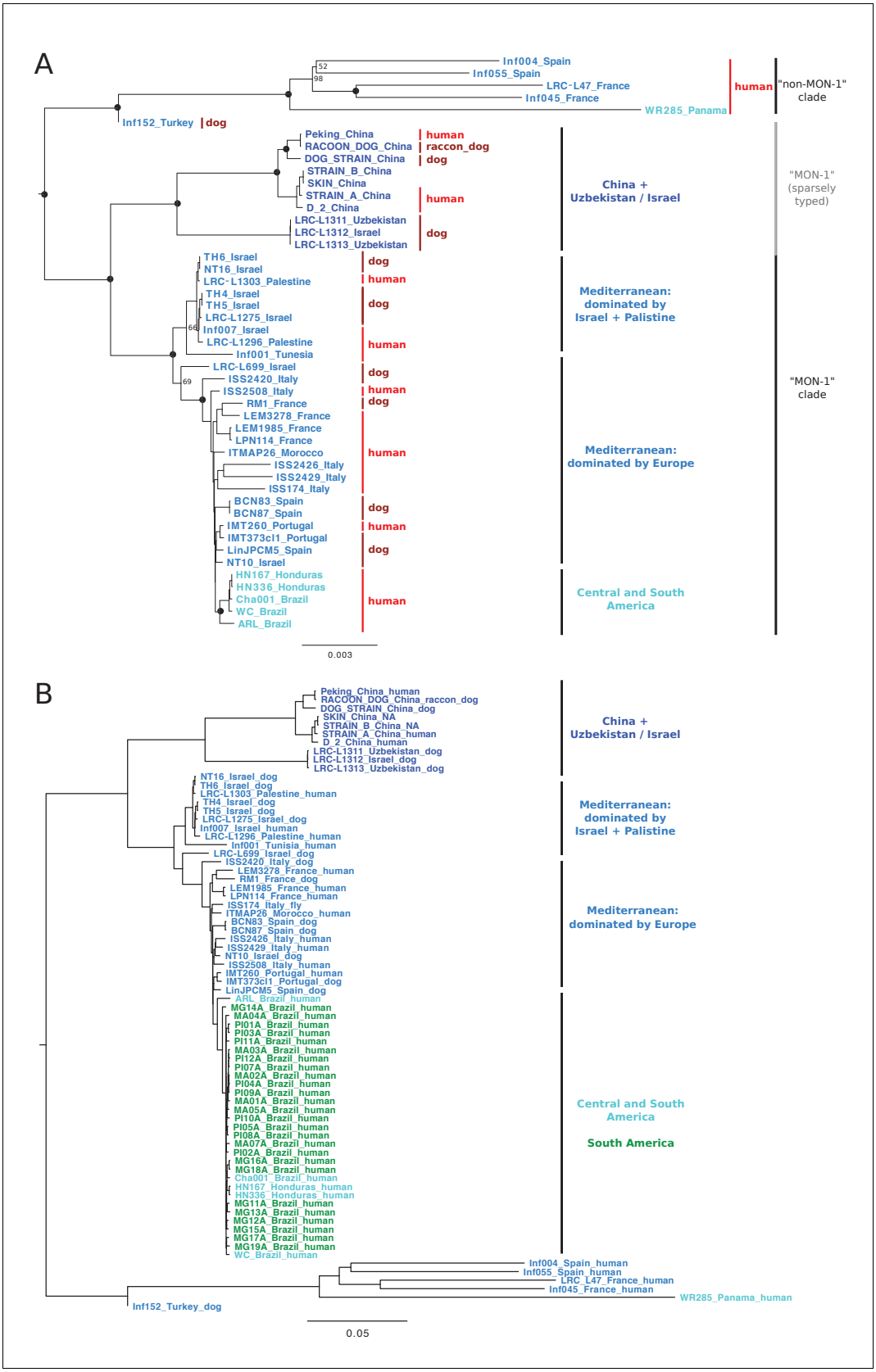

Figure 1—figure supplement 2. Sample phylogeny of the Linf1 group. (**A**) The phylogeny displays a sub-tree of the phylogeny in ***Figure 1***. Bootstrap values are shown for prominent nodes in the phylogeny as black circles for values of 100 and otherwise the respective support value. The two main

Figure 1—figure supplement 2 continued on next page

Figure 1—figure supplement 2 continued

subclades differentiate MON-1 from non-MON-1 samples (annotation based on ***Figure 1—figure supplement 1***) and sample colours indicate the geographic origin, that is Asia (dark blue), Mediterranean (pale blue) and South American (turquois). Sample names are annotated by their country of origin. The host the parasite was isolated from is indicated in bright and dark red, respectively. (**B**) Phylogeny of Linf1 samples including additional strains isolated from human infections from three different states in Brazil: Maranho (MA), Minas Gerais (MG) and Piau (PI) (***Carnielli et al., 2018***) (coloured in green).

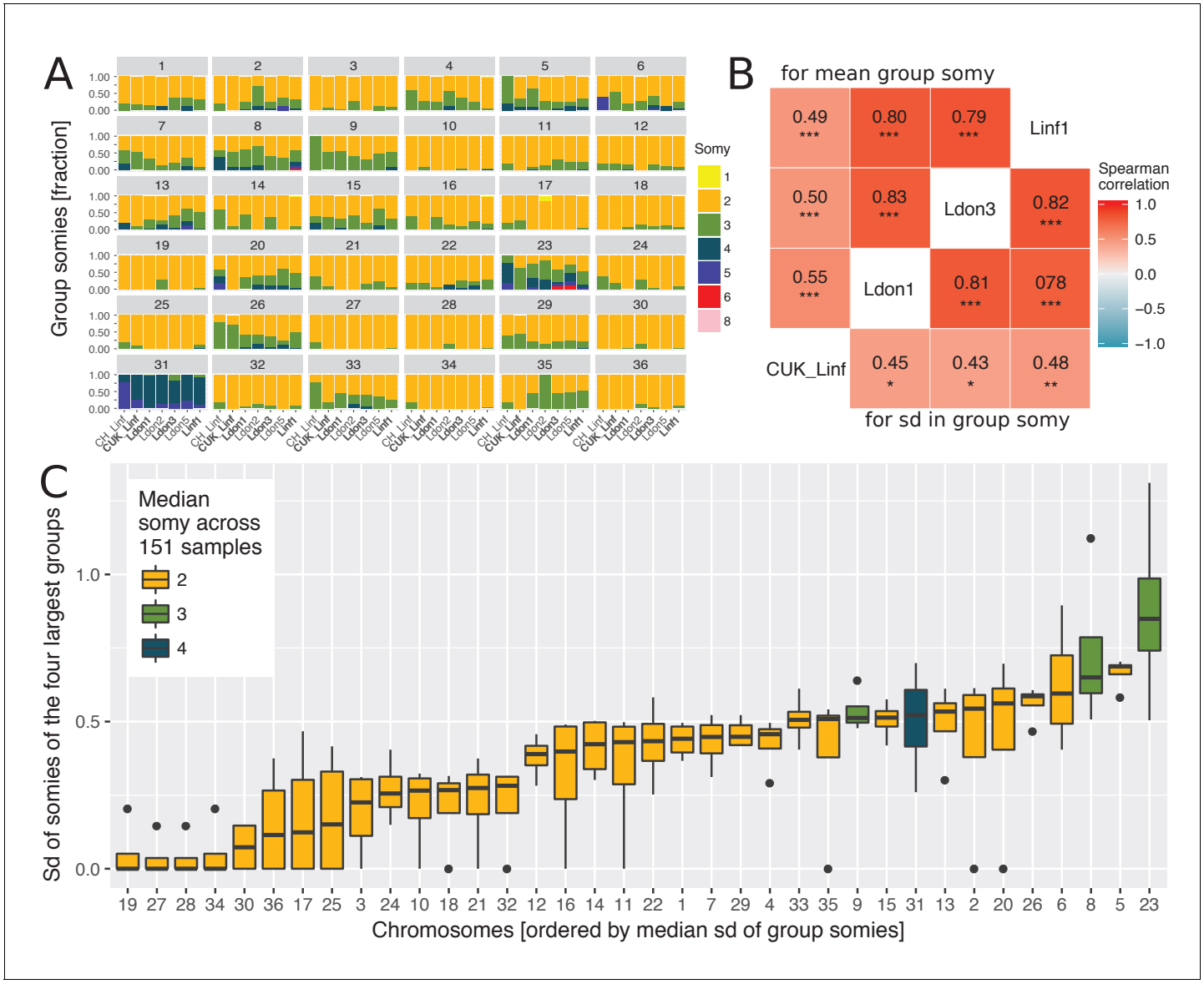

Figure 2. Chromosome-specific somy variability. (**A**) Somy variability is displayed for the 7 largest groups (≥ 5 isolates) for each chromosome as fractions of isolates with the respective somies. The four largest groups (≥ 9 samples per group) are indicated in bold. (**B**) The heatmap shows the Spearman correlations of chromosome-specific somy statistics between the four largest groups, measured as the mean group somies (upper triangle) and the standard deviation (sd) of chromosome somies (lower triangle), respectively. False discovery rates (FDR) of each correlation are indicated by asterisks (*: < 0.05, **: < 0.01, ***: < 0.001). (**C**) Boxplots show the distribution of variability in chromosome-specific somy across the four largest groups used as independent replicates across the species range. Medians estimate the chromosome-specific variation in somy.

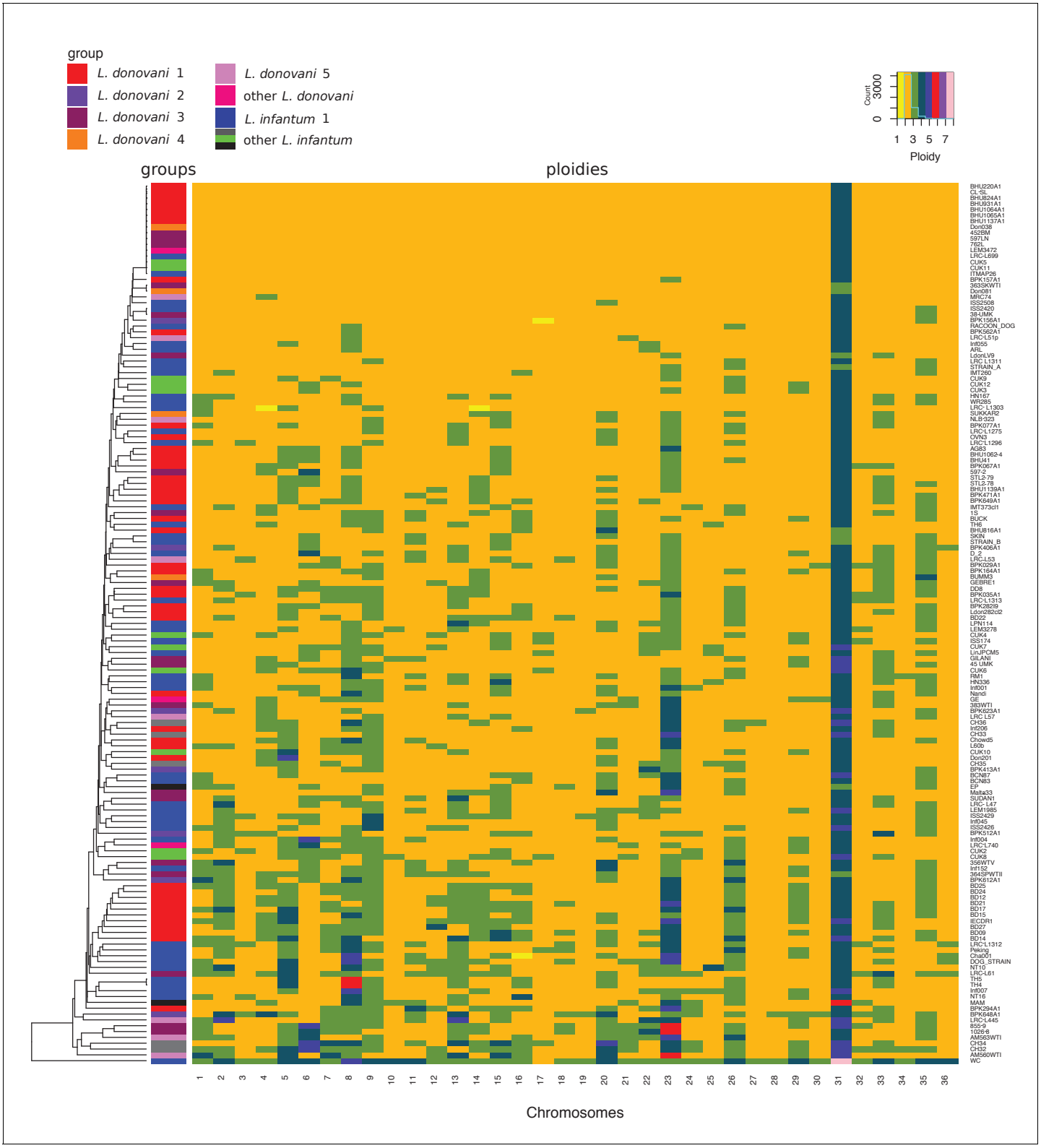

Figure 2—figure supplement 1. Aneuploidy patterns across all 151 samples. The heatmap displays the ploidies of the individual chromosomes and samples. Samples, displayed in different rows, are ordered by average linkage clustering and chromosomes are shown in different columns. Different ploidies are indicated by the colours in the heatmap (legend: upper right corner). The column on the left indicates the different phylogenetic groups (legend: upper left corner).

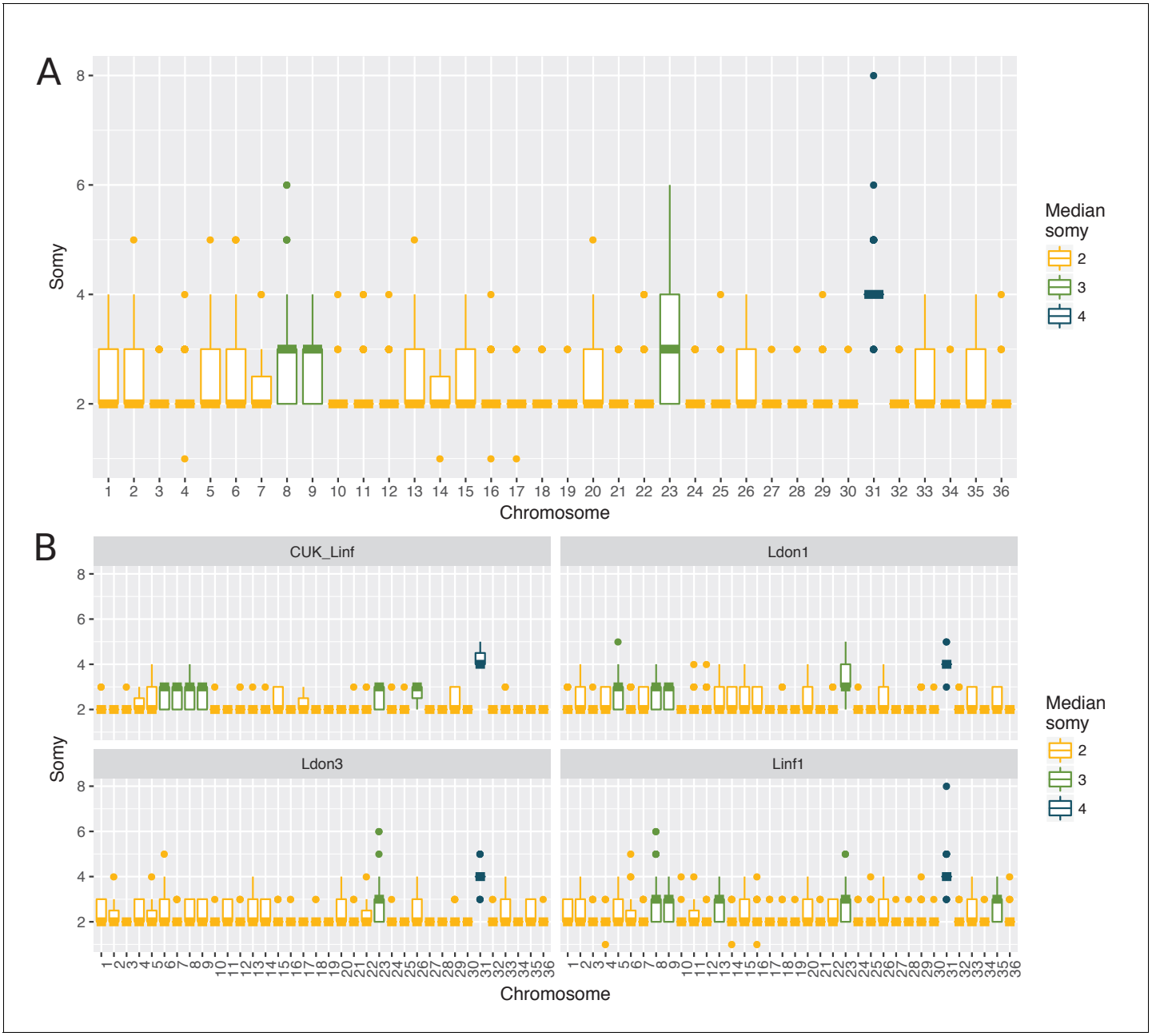

Figure 2—figure supplement 2. Aneuploidy distributions for the different chromosomes. (**A**) The boxplots show the distributions of all observed ploidies across all 151 samples for each chromosome. (**B**) Ploidy distributions are shows across the four largest (sub-)groups (>=9 samples per group) identified in the data set.

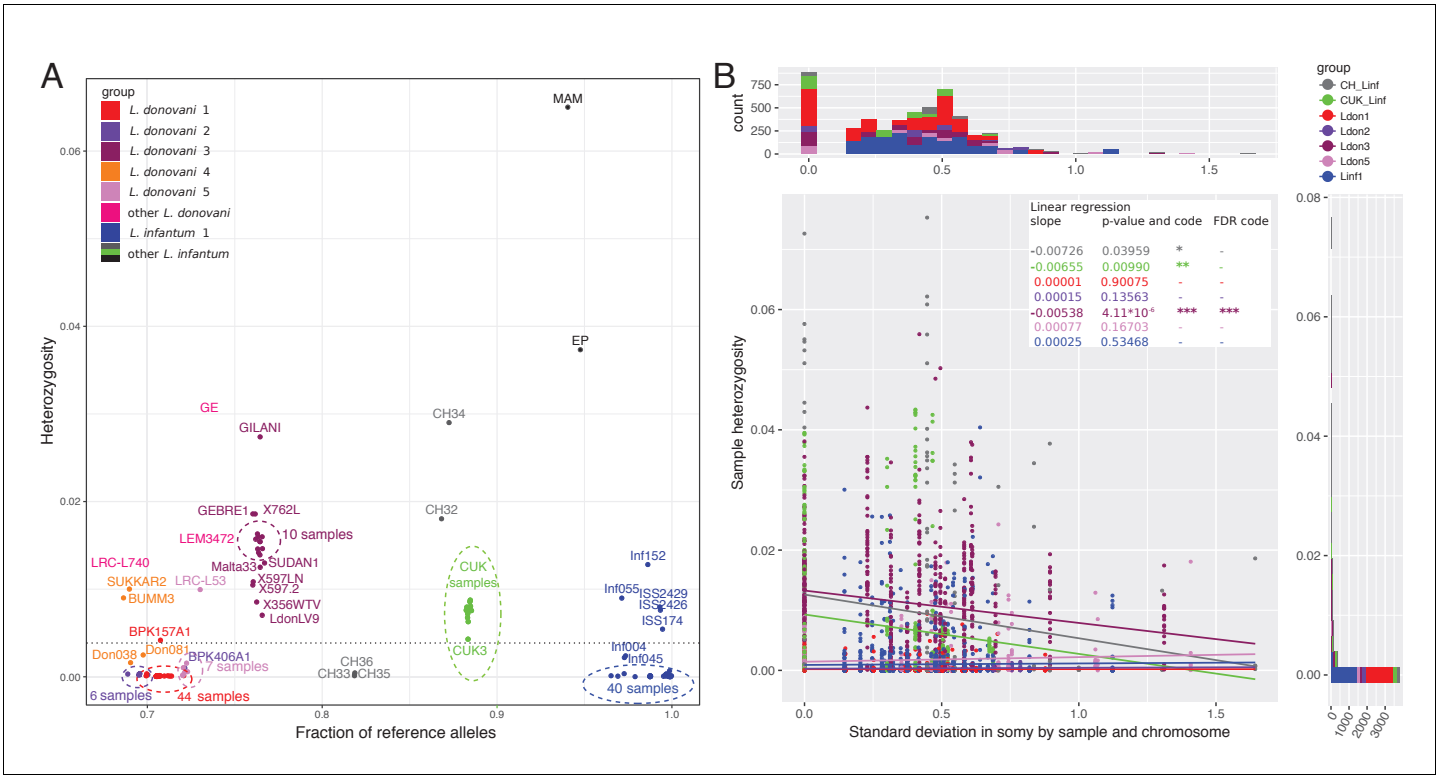

Figure 3. Whole genome sample heterozygosities. (**A**) Whole genome heterozygosities versus fraction of reference alleles. The fraction of reference alleles is calculated across all 395,602 SNP loci in the data set. Isolate names are written unless they are present in dense clusters indicated by dashed-line circles. Groups are indicated by colour as defined in figure 1. The dashed horizontal line at a genome-wide heterozygosity of 0.004 was chosen to separate samples with putative recent between-strain hybridisation history. (**B**) Relationship between chromosome-specific somy variability and sample heterozygosity. The scatterplot describes the relationship between the standard deviation in chromosome-specific somy by group (groups with $\geq$ 5 samples) against the chromosome-specific sample heterozygosity. Linear regressions were performed for each group. Asterisks indicate statistical significance of the estimated regression slope with *: < 0.05, **: < 0.01, ***: < 0.001 or '-' for not significant. Marginal histograms on the top and on the right correspond to the x-values and the y-values of the scatterplot, respectively. Groups are indicated by the different colours.

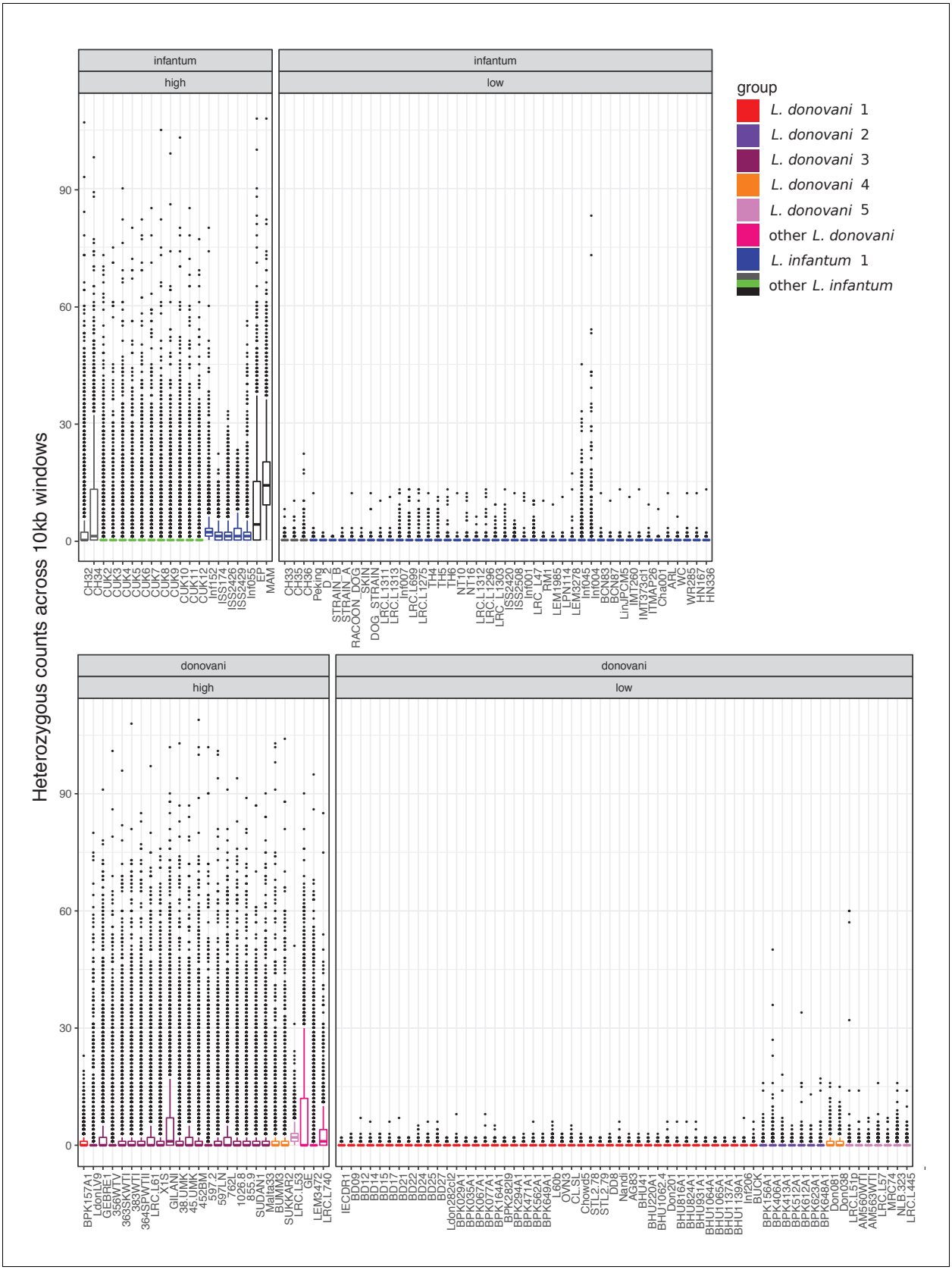

Figure 3—figure supplement 1. Distribution of heterozygous sites across the genome. Boxplots show the distribution of heterozygous counts for 10 kb windows across the genome for each sample. Upper panels show sample categorisation by species and genome-wide 'high' versus 'low'

Figure 3—figure supplement 1 continued on next page

Figure 3—figure supplement 1 continued

heterozygosity samples (i.e. above or below a genome-wide heterozygosity of 0.004). Groups are indicated by the different colours used throughout this study.

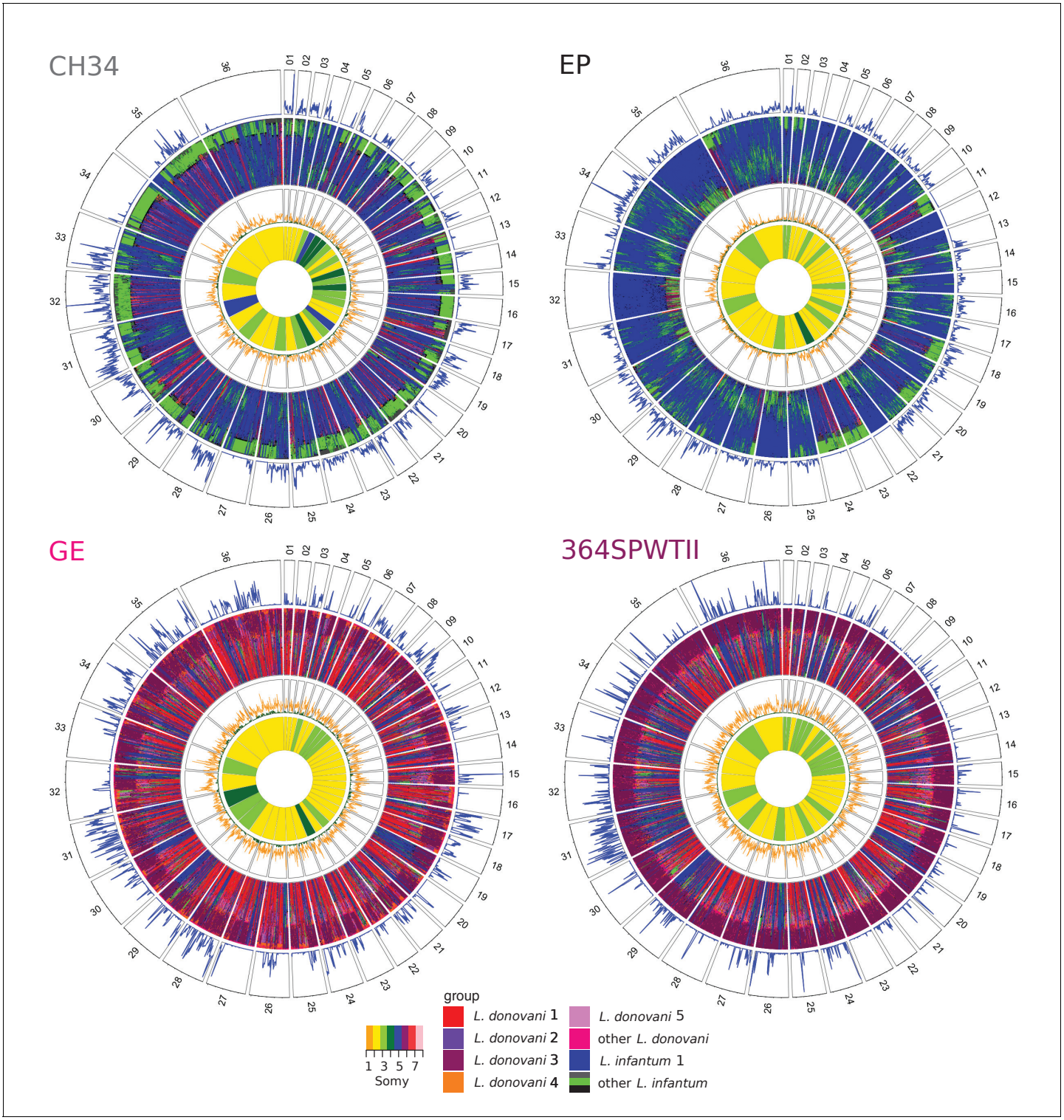

Figure 4. Window-based analysis of relatedness. Each circos plot shows four different genomic features of the isolate named in each top left corner. In the four different rings, pies correspond to the different chromosomes labelled by chromosome number. The three outer rings show a window-based analysis for a window size of 10 kb. Starting from the outer ring, they show: 1. Heterozygosity with the number of heterozygous sites ranging from 0 to 98, 146, 90 and 85 sites per window for CH34, EP, GE and 364SPWTII, respectively, 2. A heatmap coloured by groups of the 60 genetically closest isolates based on Nei's D and starting with the closest sample at the outer margin and the 60th furthest isolate at the inner margin, 3. Nei's D to the closest (green) and the 60th closest isolate (orange) scaled from 0 to 1. The innermost circle shows the colour-coded somy.

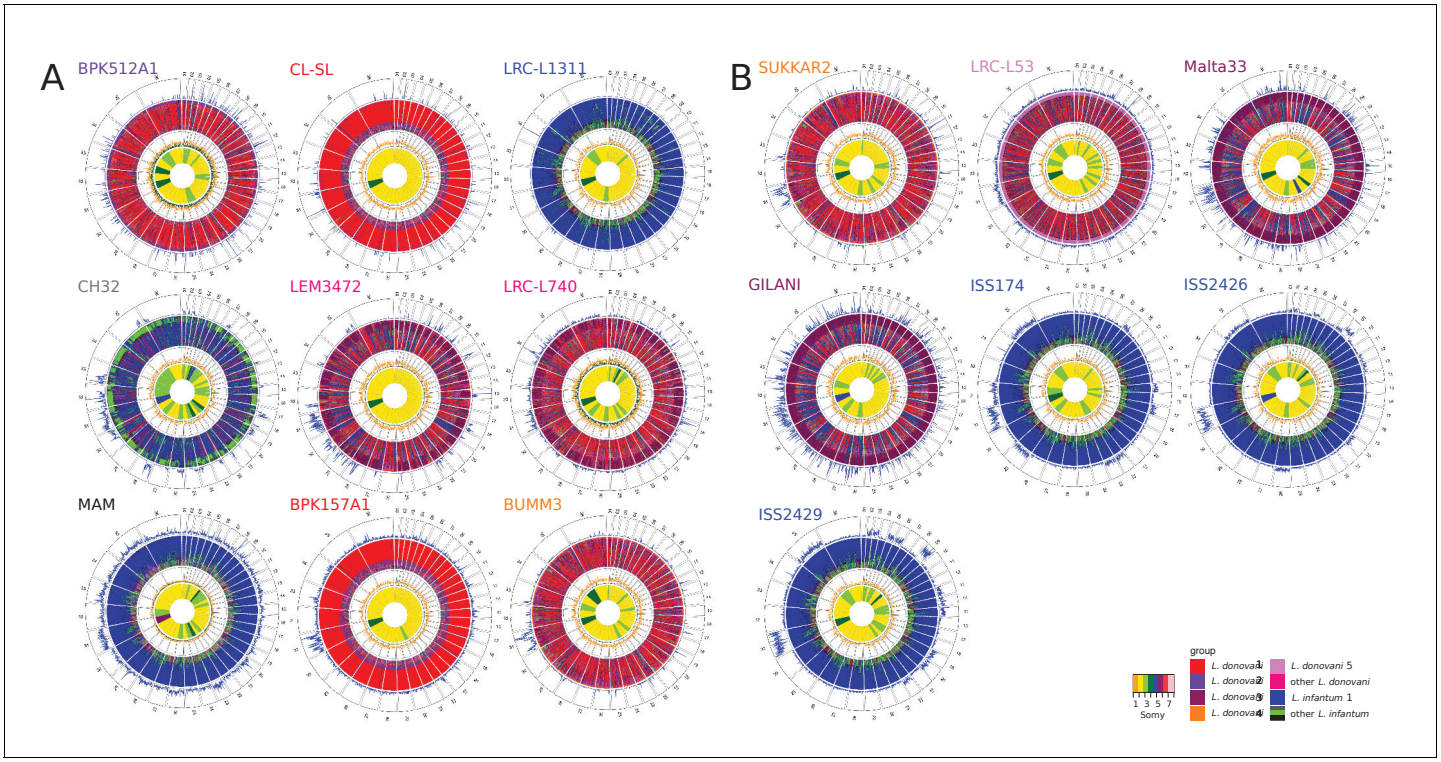

Figure 4—figure supplement 1. Window based analysis of relatedness for a subset of samples. Each circos plot shows four different genomic features of the isolate named in each top left corner. In the four different rings, pies correspond to the different chromosomes labelled by the chromosome number. The three outer rings show a window-based analysis for a window size of 10 kb. Starting from the outer ring, they show: 1. Heterozygosity (number of heterozygous sites; range from **A**) 0 to 5 (BPK512A1), 5 (CL-SL), 10 (LRC-L1311), 107 (CH32), 95 (LEM3472), 85 (LRC-L740), 108 (MAM), 23 (BPK157A1) and 103 (BUMM3) and **B**) from 0 to 104 (SUKKAR2), 51 (LRC-L53), 90 (Malta33), 102 (GILANI), 22 (ISS174), 33 (ISS2426) and 23 (ISS2429) heterozygous sites per 10 kb, respectively), 2. A heatmap coloured by groups of the 60 genetically closest isolates (based on Nei's D), starting with the closest sample at the outer margin and the 60th furthest isolate at the inner margin, 3. Nei's D to the closest isolate (green) and the 60th closest sample (orange), range from 0 to 1. The innermost circle shows the colour-coded somy.

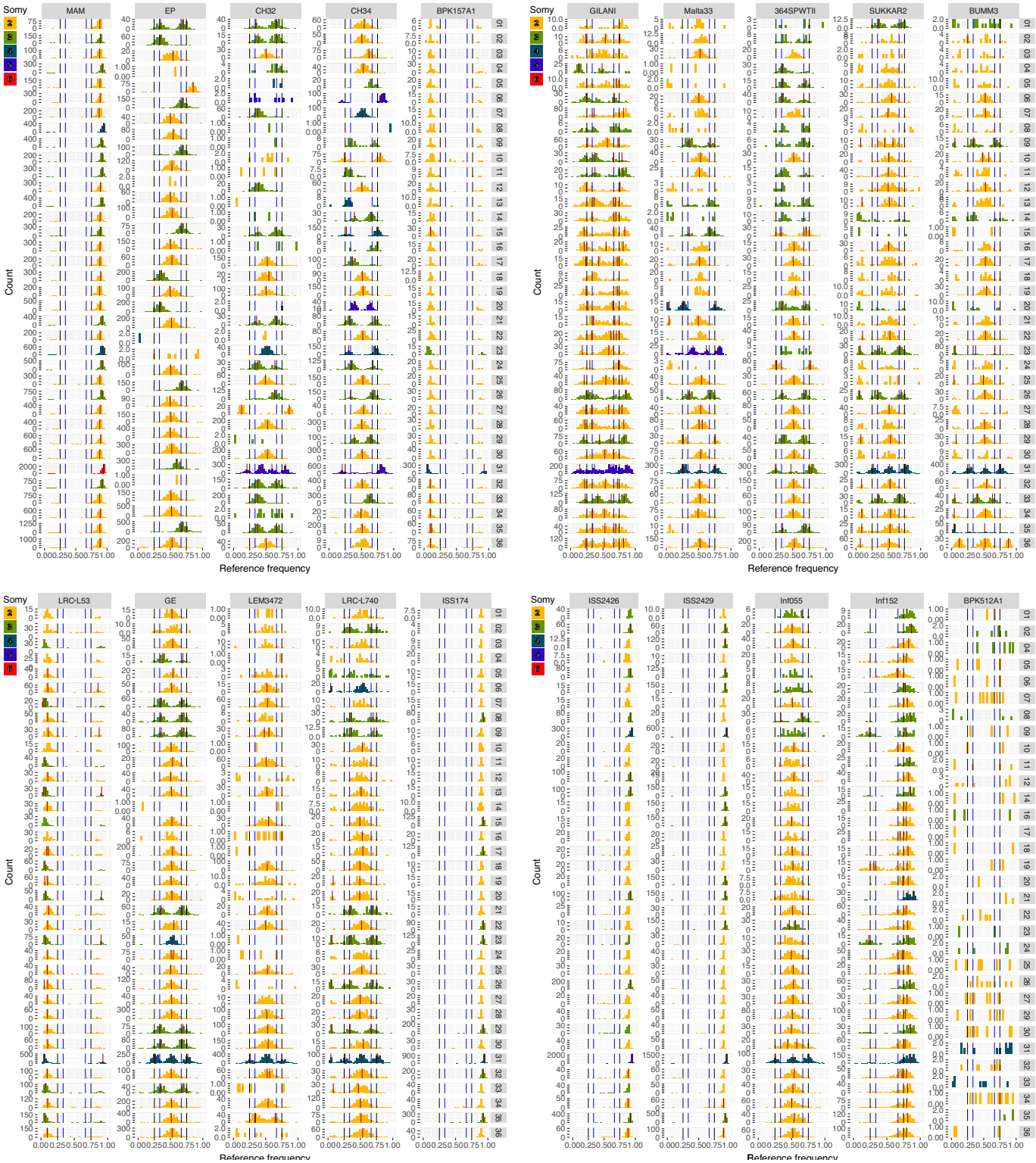

Figure 4—figure supplement 2. Allele frequency distributions by isolate. Histograms of allele frequency distributions are shown per chromosome for different isolates indicated at the top of each plot. Allele frequencies of $\frac{1}{3}\&\frac{2}{3}$ and $\frac{1}{4}\&\frac{3}{4}$ are visualised by blue and black vertical lines, respectively. Red

Figure 4—figure supplement 2 continued on next page

Figure 4—figure supplement 2 continued

vertical lines indicate estimated peaks of the distribution, which are only shown for distributions with at least 100 SNPs (see Materials and methods). Colours of the distributions indicate the estimated somy based on chromosomal coverage.

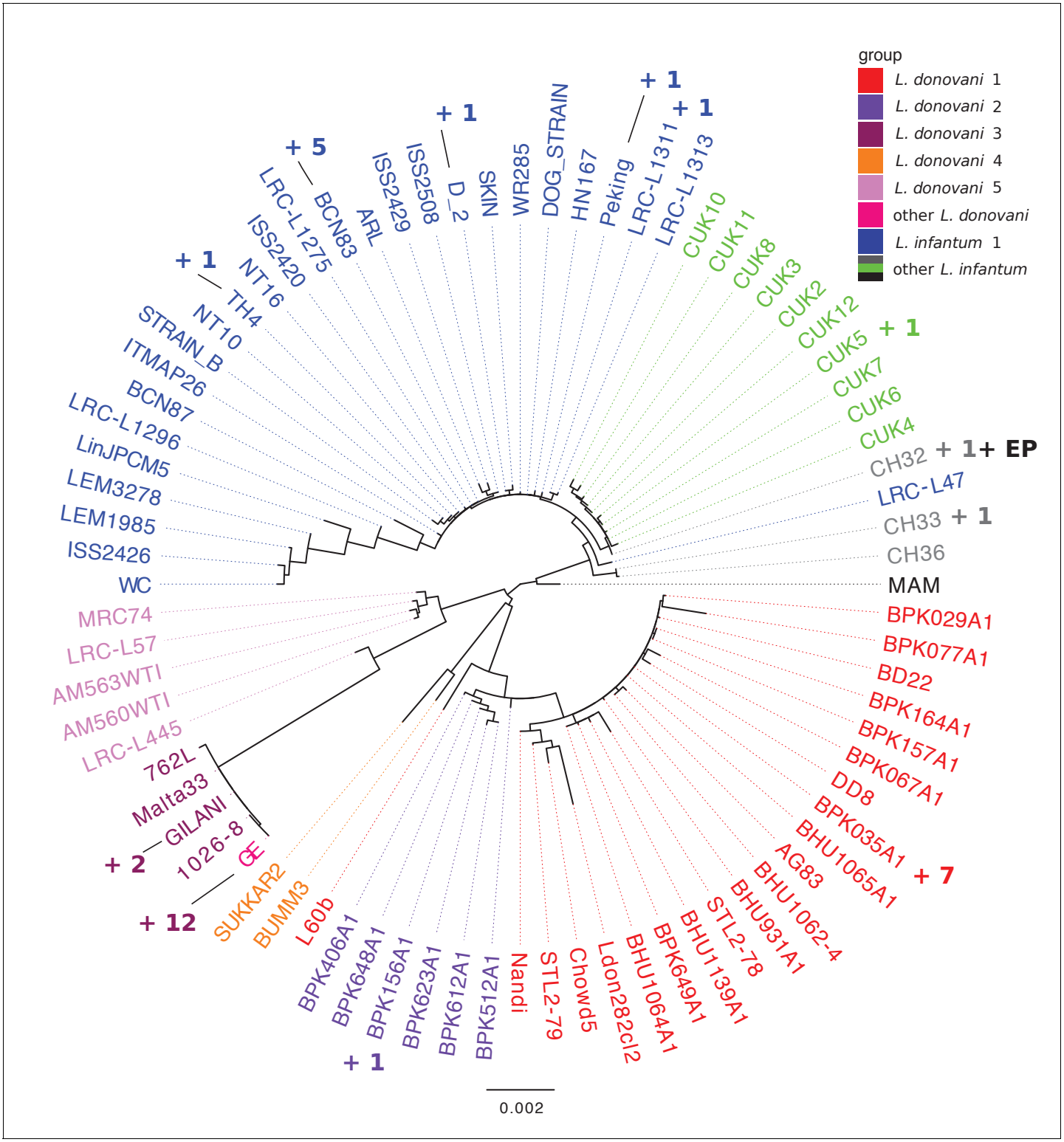

Figure 4—figure supplement 3. Phylogenetic tree based on the maxicircle DNA. Unrooted tree constructed using RaxML. Only a part of the samples is shown in the tree as the library preparation of some samples removed the kDNA and some samples have identical sequences to the ones shown and were therefore removed for tree estimation. The samples with identical sequences are listed in ***Supplementary file 3*** and the counts of removed samples are indicated near the respective tree leaves. Group identity of each isolate is indicated by the different colours. Phylogenetic reconstruction was repeated with four outgroup isolates of *L. mexicana* (U1103.v1), *L. tropica* (P283) and *L. amazonensis* (two samples). All of those four outgroup-samples were clearly different to the samples of our study (data not shown).

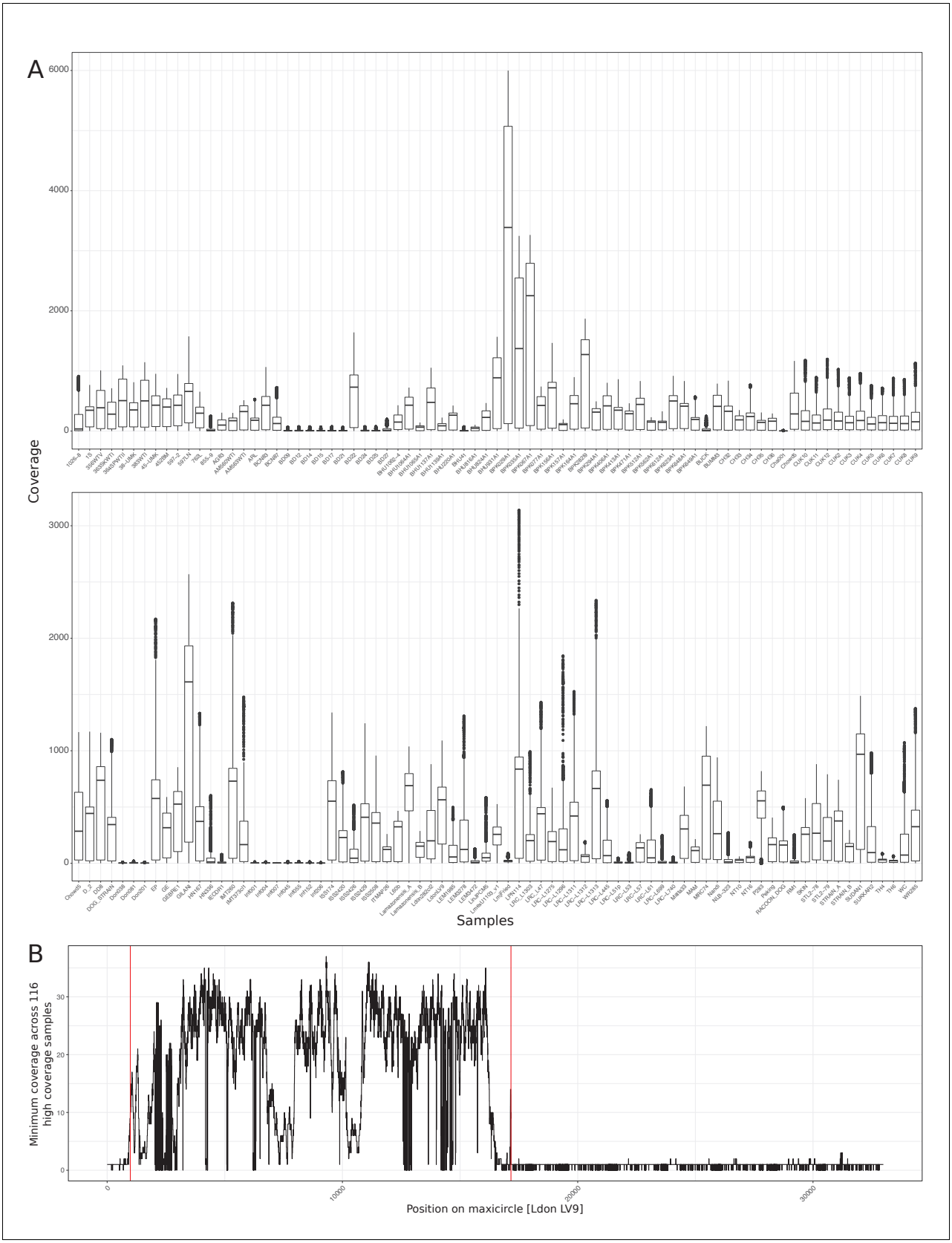

Figure 4—figure supplement 4. Coverage of maxicircle DNA. (**A**) Coverage of maxicircle DNA across isolates. Only isolates 116 of the 151 isolates had a median coverage >= 20 and were used for phylogenetic reconstruction of the maxicircle DNA. (**B**) Region of high confidence mapping to the

Figure 4—figure supplement 4 continued on next page

Figure 4—figure supplement 4 continued

maxicircle DNA across isolates. The minimum coverage across all 116 isolates with a good maxicircle coverage (median coverage >= 20, see **A**) is shown along the maxicircle DNA. The region within the red lines was chosen for phylogenetic reconstruction.

Figure 4—figure supplement 5. Somy evaluation based on allele frequency profiles. Somies for the different isolates and chromosomes were calculated based on relative chromosome wide coverages within a sample. For heterozygous samples and chromosomes with at least 100 SNPs, these estimates were evaluated based on the expected frequency distributions. Errors above 0.2 suggest deviances in somy estimates larger than expected by sampling error. In a few cases, the frequency profiles clearly suggest another somy than estimated by chromosomal coverage.

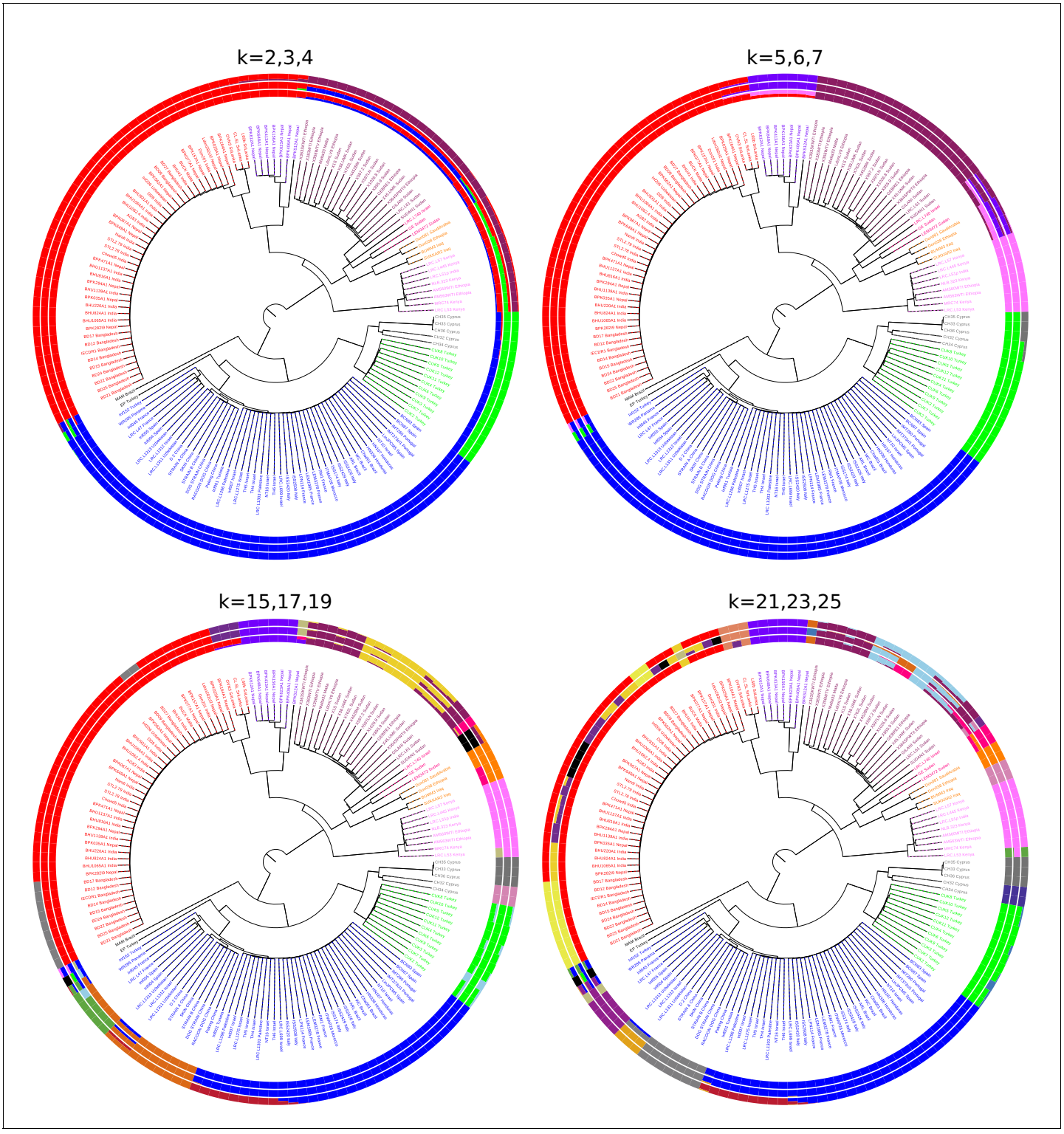

Figure 4—figure supplement 6. Sample phylogeny and admixture analysis across a range of *K* values. The phylogenetic reconstruction is identical to ***Figure 1*** but admixture results shown are for a range of different *K* values. Group colours in the phylogeny are identical to the ones used throughout this study and are matched for admixture results where possible.

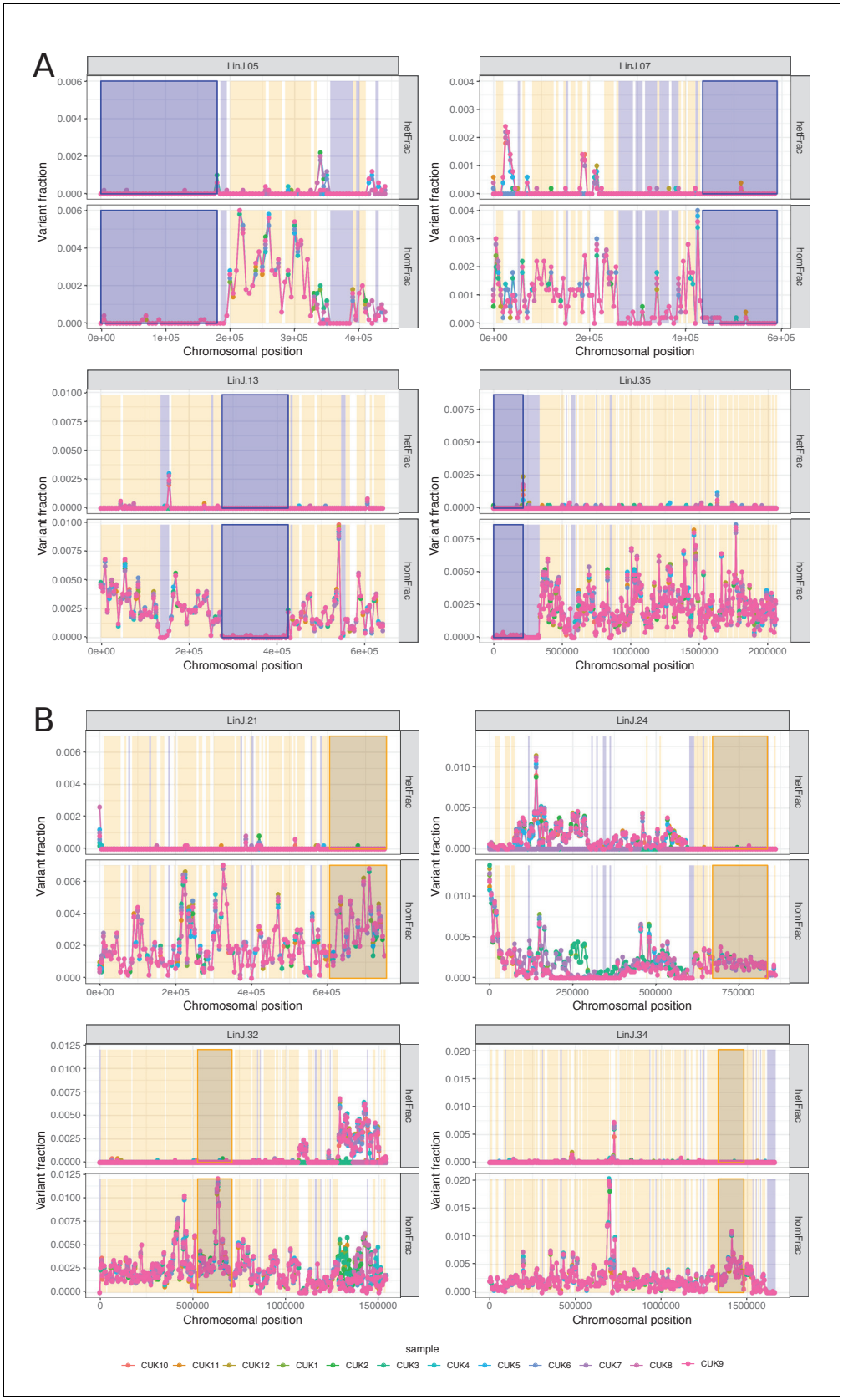

Figure 4—figure supplement 7. Genomic regions used for haplotype-based parent identification. Figures show the four largest identified genomic regions specific to either parent used for parent characterisation: (**A**) JPCM5-like parent and (**B**) other unknown parent. Each figure shows the fractions

Figure 4—figure supplement 7 continued on next page

Figure 4—figure supplement 7 continued

of heterozygote SNPs or fixed homozygous SNPs with respect to the JPCM5 reference for each of the 12 isolates from the Cukurova region, Turkey (CUK, ***Rogers et al., 2014***). Conservatively called regions originating from a specific parent are coloured in blue (JPCM5-like) and orange (other). Genomic regions used for phylogenetic reconstruction are framed.

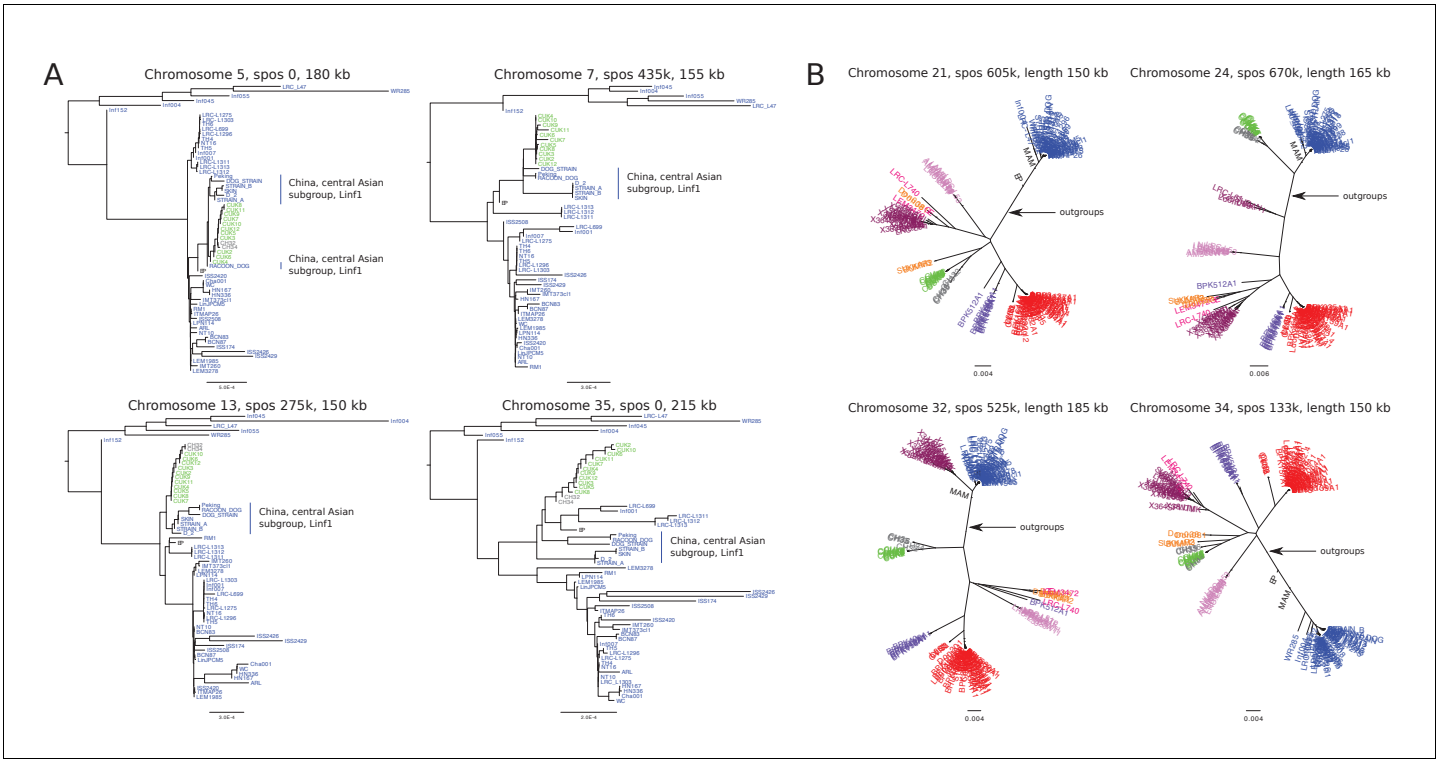

Figure 4—figure supplement 8. Putative parents of CUK samples. Phylogenetic trees were reconstructed using Nei's D and neighbour joining including all 151 samples and three outgroup samples, *L. mexicana* (U1103.v1), *L. tropica* (P283) (see Materials and methods). Trees were done for the four largest homozygous genome regions across the CUK genomes either almost devoid of differences A) or with increased fixed differences B) to the JPCM5 reference. (**A**) Phylogenetic trees using genomic regions putatively from a JPCM5-like parent in the CUK samples. For a better resolution, only the subtree for the Linf1 clade is shown, respectively. (**B**) Phylogenetic trees using genomic regions with increased numbers of fixed differences to the JPCM5 reference in the CUK samples. For a better resolution the position of the out-group branch is only indicated.

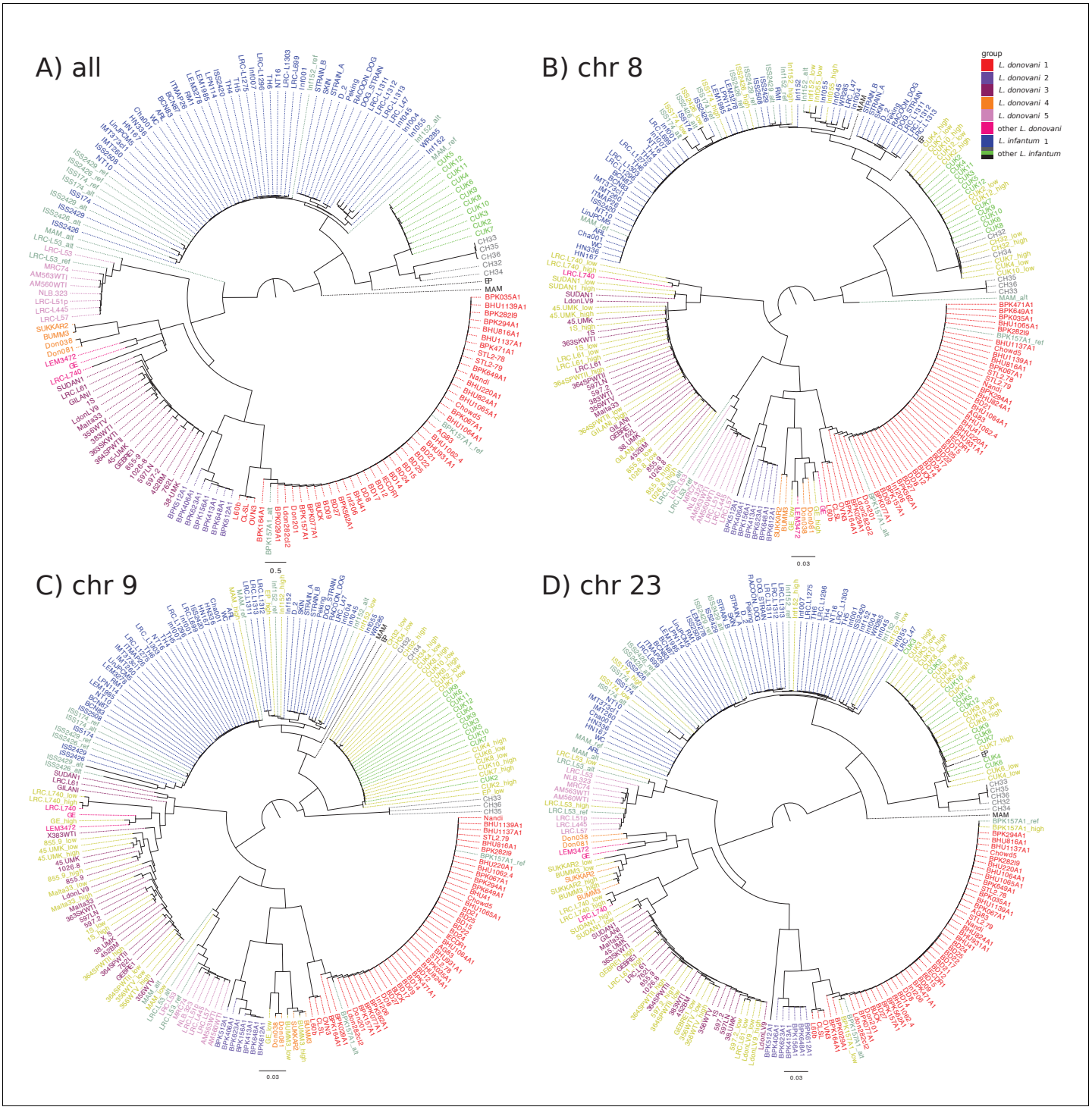

Figure 4—figure supplement 9. Sample phylogeny based on genomic SNP variation including phased samples with skewed allele frequency spectra. Phylogenetic reconstruction was equivalent to the reconstruction shown in ***Figure 1***. (**A–D**) Samples with strongly skewed allele frequency spectra across all chromosomes were phased based on high versus low allele frequency variants and also included in the tree. Resulting haplotypes are coloured in greenish grey and labelled with ref (reference allele; reference *L. infantum* JPCM5) and alt (alternate allele) depending on the polarisation of the majority if SNPs in the respective haplotype. (**B–D**) Individual chromosomes that were triploid were phased based on the allele frequency spectra. Resulting haplotypes are coloured in yellowish grey and labelled with "high", where the JPCM5 reference allele is at high $(\sim \frac{2}{3})$ and the non-reference allele is at low $(\sim \frac{1}{3})$ sample frequency. The haplotype resulting from the opposite scenario is labelled with 'low' The remaining samples are coloured in their typical group colours.

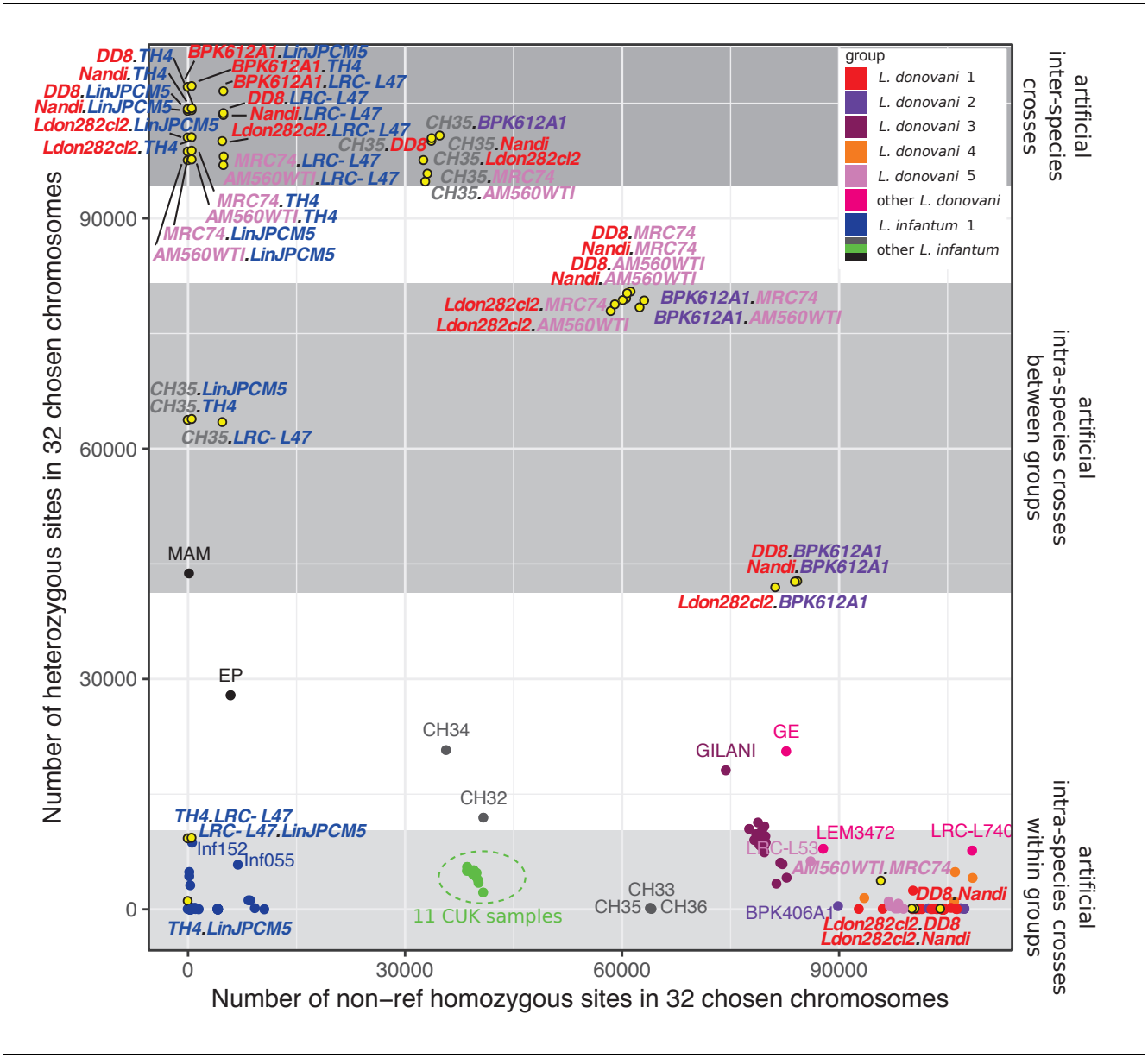

Figure 4—figure supplement 10. Heterozygosity of artificial F1 hybrids. Artificial F1 hybrids were generated by matching ten almost entirely homozygous samples across the phylogeny. Chromosomes 2, 8, 22 and 31 were excluded since they did not satisfy these criteria. The number of non-reference homozygous sites is plotted against the number of heterozygous sites. The plot includes real samples as well as artificial hybrids. Groups are indicated by colours. Artificial hybrids are displayed by yellow dots surrounded by black circles and their names are written in bold italics indicating the names of the combined samples in their respective group colours separated by a black dot. Grey shaded areas indicate heterozygosity levels that are spanned by different levels of artificial hybrids including inter-species and intra-species hybrids between and within identified groups.

Figure 4—figure supplement 11. Verification of skewed allele frequency spectra in a subset of isolated strains. For all 11 samples, where sample allele frequency distributions indicated mixed infections, heterozygous SNPs were filtered for highest quality SNPs (SNP calling quality of 99 and presence of the alternate allele in at least five other samples as homozygous call). Frequency distributions of the high vs. lower quality heterozygous SNPs are shown for the different samples. Signals for the suggested highly homozygous sub-clones BPK157A1, Inf152, ISS174, ISS2426, ISS2429, LRC-L53 and MAM, and the sub-clones BUMM3, LRC-L740, Malta33 and SUKKAR2 with the high-frequency one being heterozygous are mainly supported by highest quality SNPs.

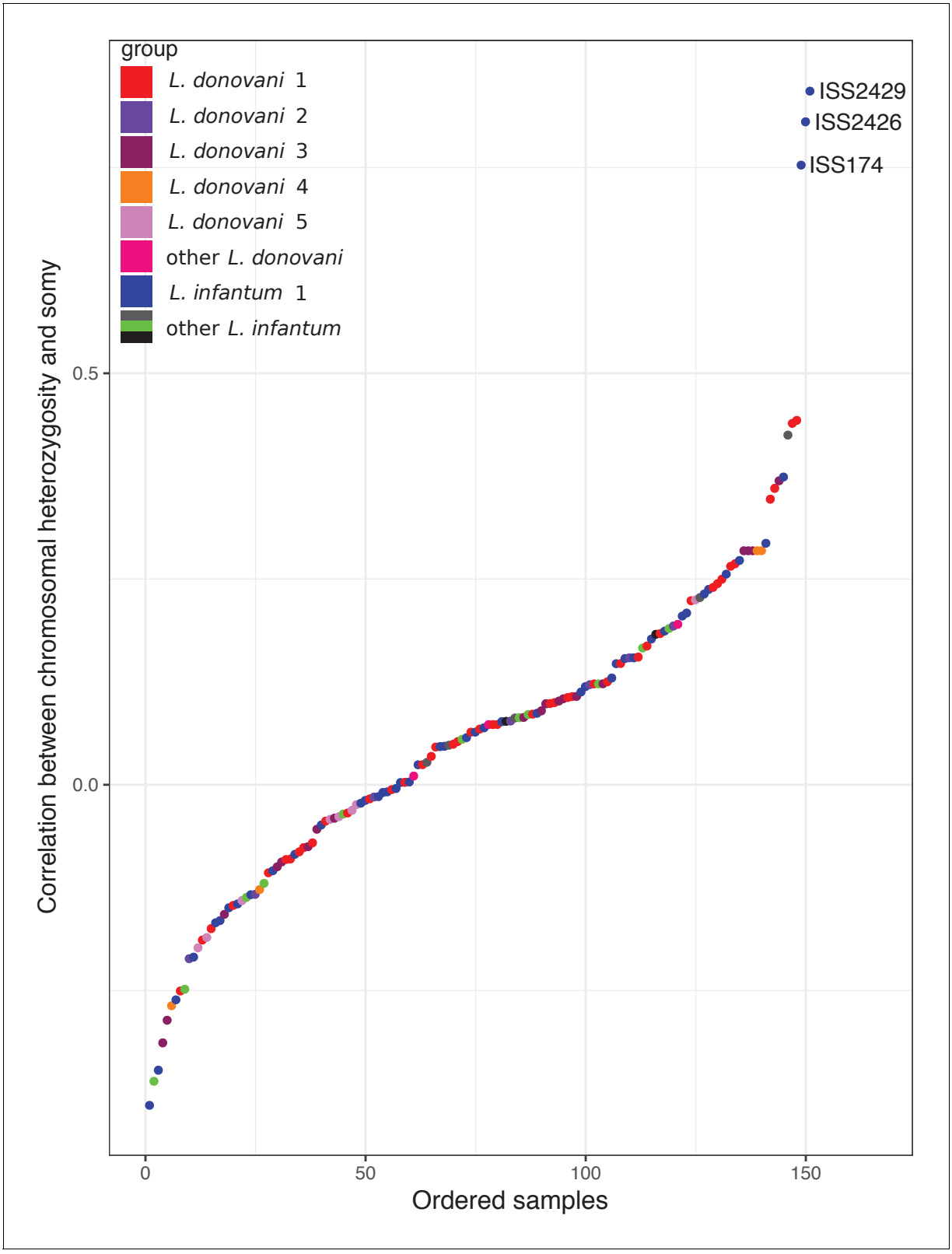

Figure 4—figure supplement 12. Correlation between somies and heterozygosities across chromosomes. For all 151 isolates aneuploidy profiles were correlated with respective chromosome-specific heterozygosity values. Names of isolates with significant correlations (Spearman, FDR <= 0.001) are printed next to the respective correlation value.

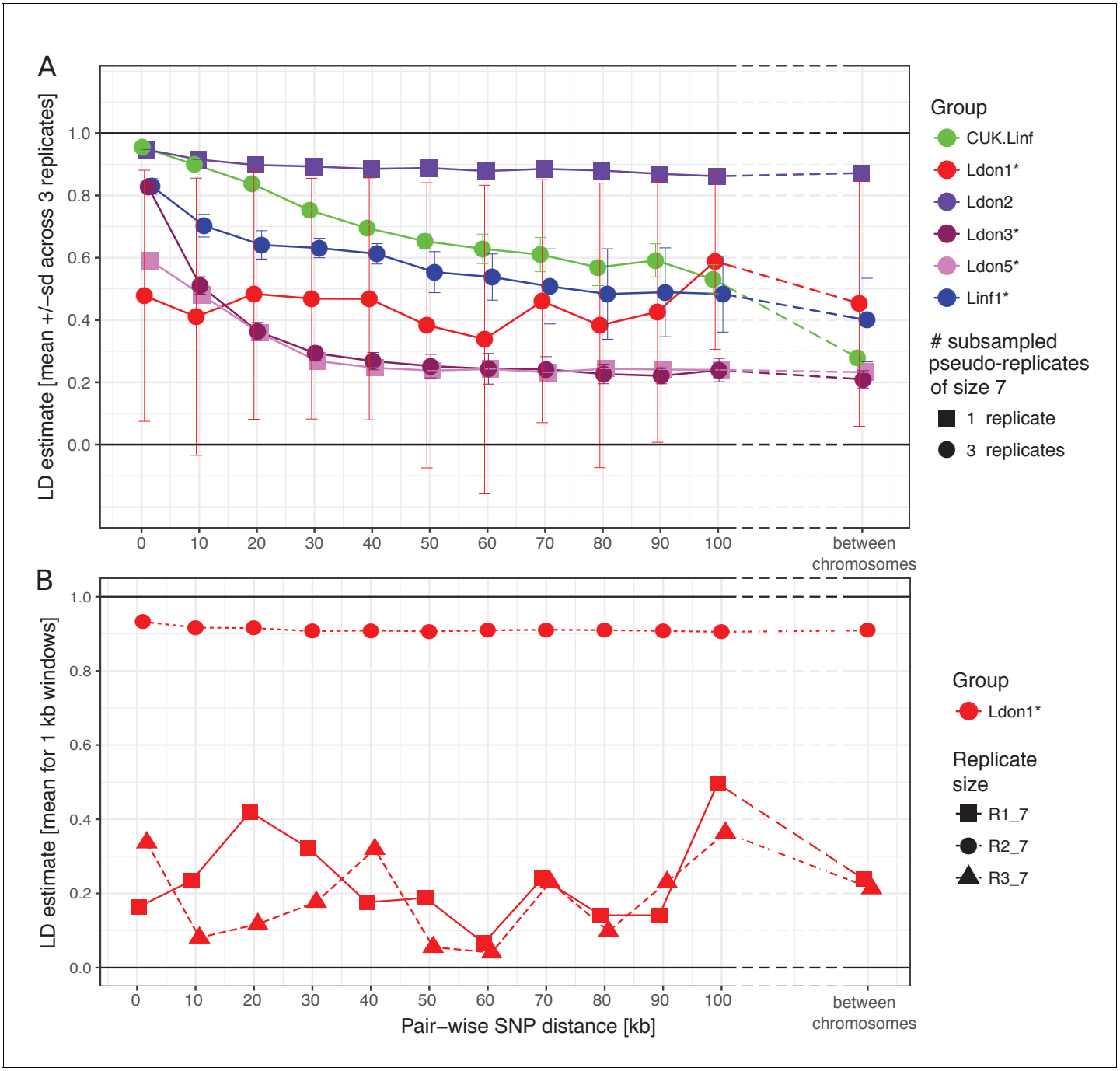

Figure 5. LD decay with genomic distance. (**A**) LD decay was measured for the six largest groups removing isolates that were identified as putative strain mixtures (indicated by *; see Materials and methods). Groups with more than seven isolates per group were sub-sampled to three pseudo-replicates of seven isolates (round symbols) to make LD estimates comparable between groups. Mean and standard deviation across the three pseudo-replicates are shown where applicable. Groups with only seven isolates were not sub-sampled and are indicated by squared symbols. (**B**) LD decay with distance is shown for the three pseudo-replicates for the Ldon1 group. (**A and B**) Data for individual replicates was calculated as means of 1 kb windows for SNP pairs of the stated genomic distance. For LD estimates between chromosomes, 100 SNPs were randomly sampled per chromosome and means across all pair-wise combinations between chromosomes are shown. This procedure was done twice independently but as differences between both such replicates were negligible, only the results of one replicate are shown.

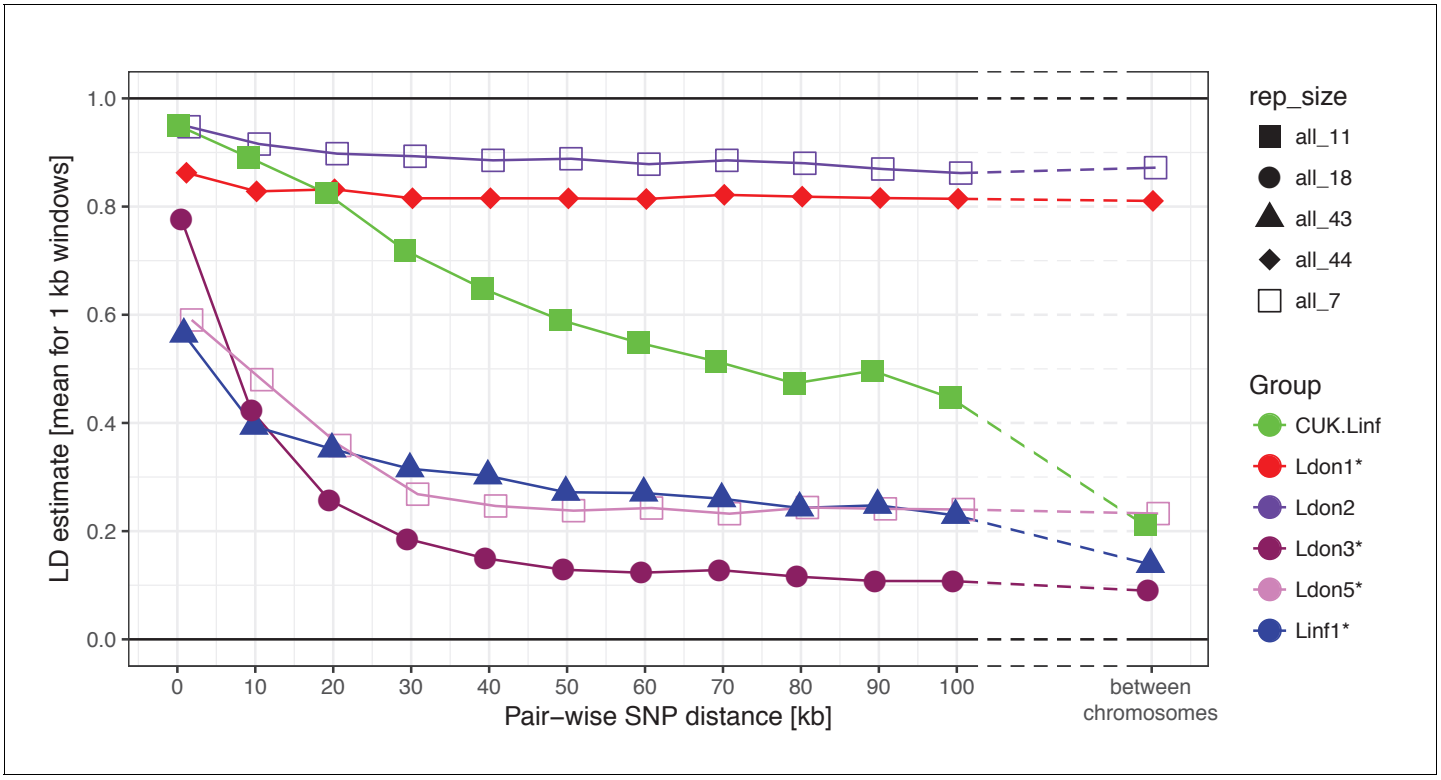

Figure 5—figure supplement 1. LD decay with genomic distance. LD decay was measured for the six largest groups removing isolates that were identified as putative strain mixtures (indicated by *; see Materials and methods). Shown are means of 1 kb wide windows for SNP pairs of the stated genomic distance. For LD estimates between chromosomes, 100 SNPs were randomly sampled per chromosome and means across all pair-wise combinations between chromosome are shown. This procedure was done twice independently but as differences between both such replicates were negligible only the results of one replicate are shown. Samples sizes used for LD estimates vary between groups depending on the group size.

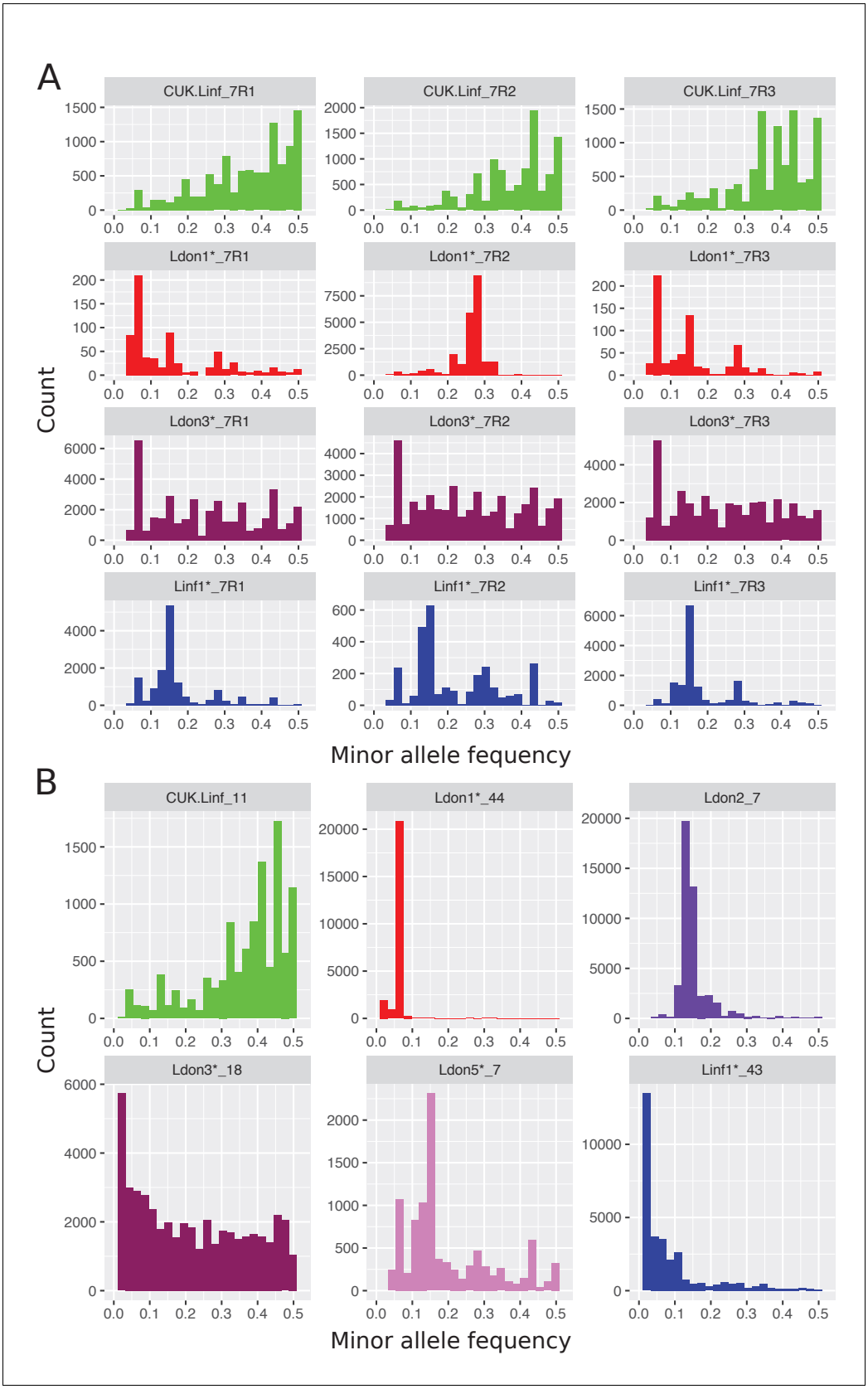

Figure 5—figure supplement 2. Folded site frequency spectra of the six largest groups. (**A**) The six largest groups were randomly sub-sampled to seven samples into three pseudo-replicates and respective SFS are shown. (**B**) SFS are shown for the six largest groups including all available samples.

Figure 5—figure supplement 2 continued on next page

Figure 5—figure supplement 2 continued

Group names are indicated at the top of each plot. Asterisks indicate that samples that were identified as clone mixtures were removed (see Materials and methods). Sample size and replicate number if applicable are written next to the group names.

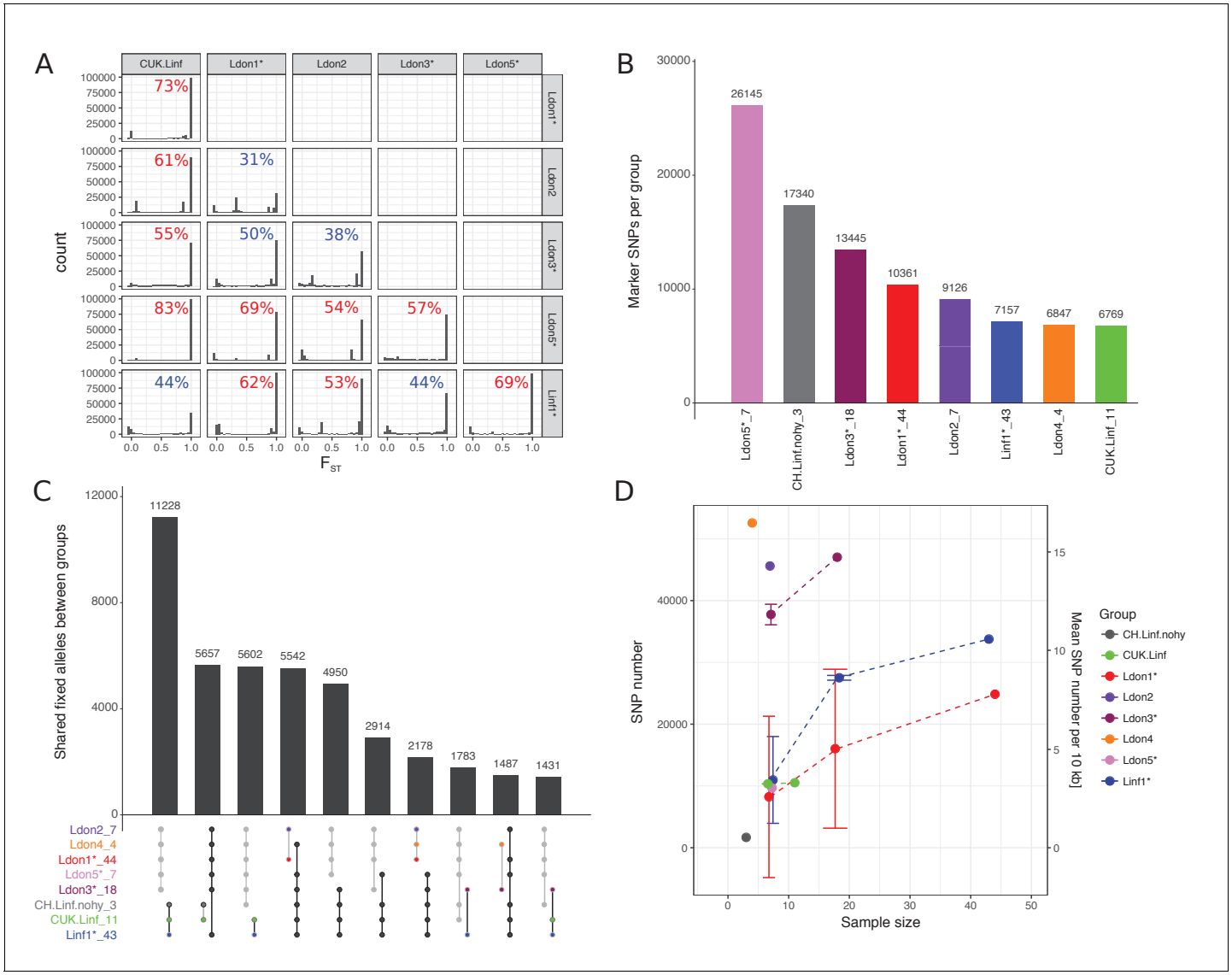

Figure 6. Differentiated and segregating SNPs between and within groups. For this analysis isolates that were shown to be mixtures of clones or hybrids between groups were removed (indicated by '*', see also Materials and methods). Groups sizes after removal of those isolates are specified in panels A and C. (**A**) F_{ST} values between pairwise group comparisons. The fraction of differentially fixed SNPs (F_{ST} = 1) for each pairwise group comparison is indicated at the top right corner of each plot. Percentages larger than 50% are coloured in red, otherwise blue. (**B**) The number of marker SNPs for each group, that is SNPs that are differentially fixed in one group versus all others. (**C**) Number of SNPs that are differentially fixed between sets of groups. Groups fixed for the same allele are indicated in the bottom panel through connecting points corresponding to the specific groups. Grey and black lines connect sets of groups monomorphic for the alternate and reference allele, respectively. (**D**) Number and density of SNPs segregating in the respective groups. As sample sizes of the different groups vary, figures are also shown for three random sub-samples of the larger groups. Results of sub-sampling are displayed as mean and sd.

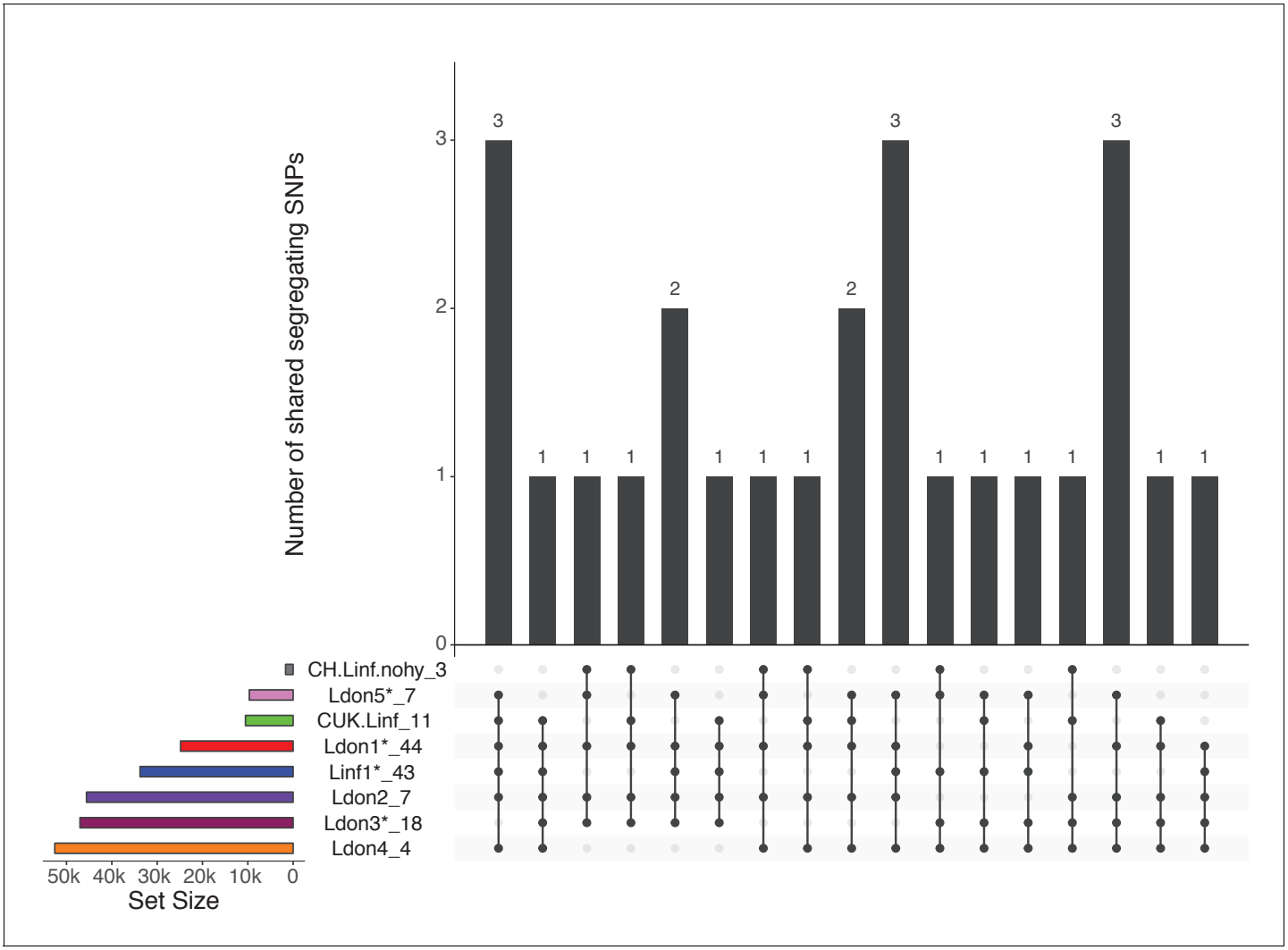

Figure 6—figure supplement 1. Polymorphism sharing between groups. The histogram lists the number of SNPs that are segregating in multiple groups with the respective groups indicated by a black dot in the panel below. Polymorphism sharing between groups is only shown for sites that are shared by at least five of the eight groups. The histogram on the left shows the number of segregating polymorphisms in each group individually.

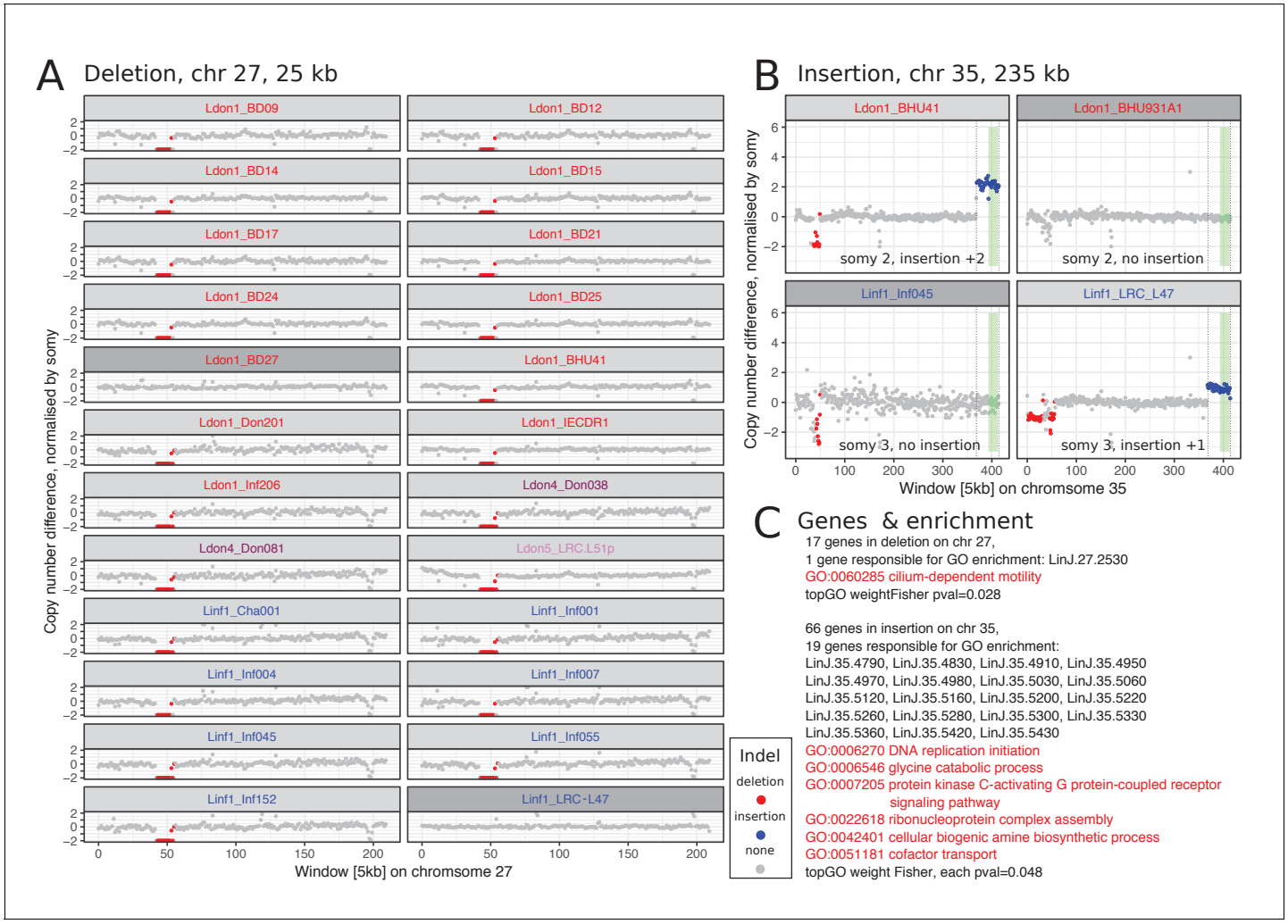

Figure 7. Two large CNVs that are shared between both species. (**A**) Chromosome 27 has a 25 kb long deletion that is present in 15% of all samples and four different groups. All chromosomes 27 that have this deletion in our dataset are diploid and the deletion results in a loss of this allele in the respective sample. (**B**) The duplication on chromosome 35 is 235 kb long and present in one isolate of group Ldon1 and Linf1, respectively. The insertion is once present on a disomic background with a 2-fold increase and once on a trisomic background with a 1-fold increase. The green rectangle marks the CD1/LD1 locus sequences for *L. infantum* described in ***Sunkin et al. (2001)*** (***Supplementary file 8***). For A) and B) a few closely related samples not harbouring the respective CNV are also displayed and highlighted in dark grey. Group identities are indicated by colours of the isolate name. (**C**) Genes present in the respective CNV along with GO enrichment results using topGO (***Alexa et al., 2006***). Details on both CNVs can be found in ***Supplementary file 7***: unique CNVs with ids 150 and 215, respectively. The CNV characterisation of the corresponding isolates can be found in ***Supplementary file 6***.

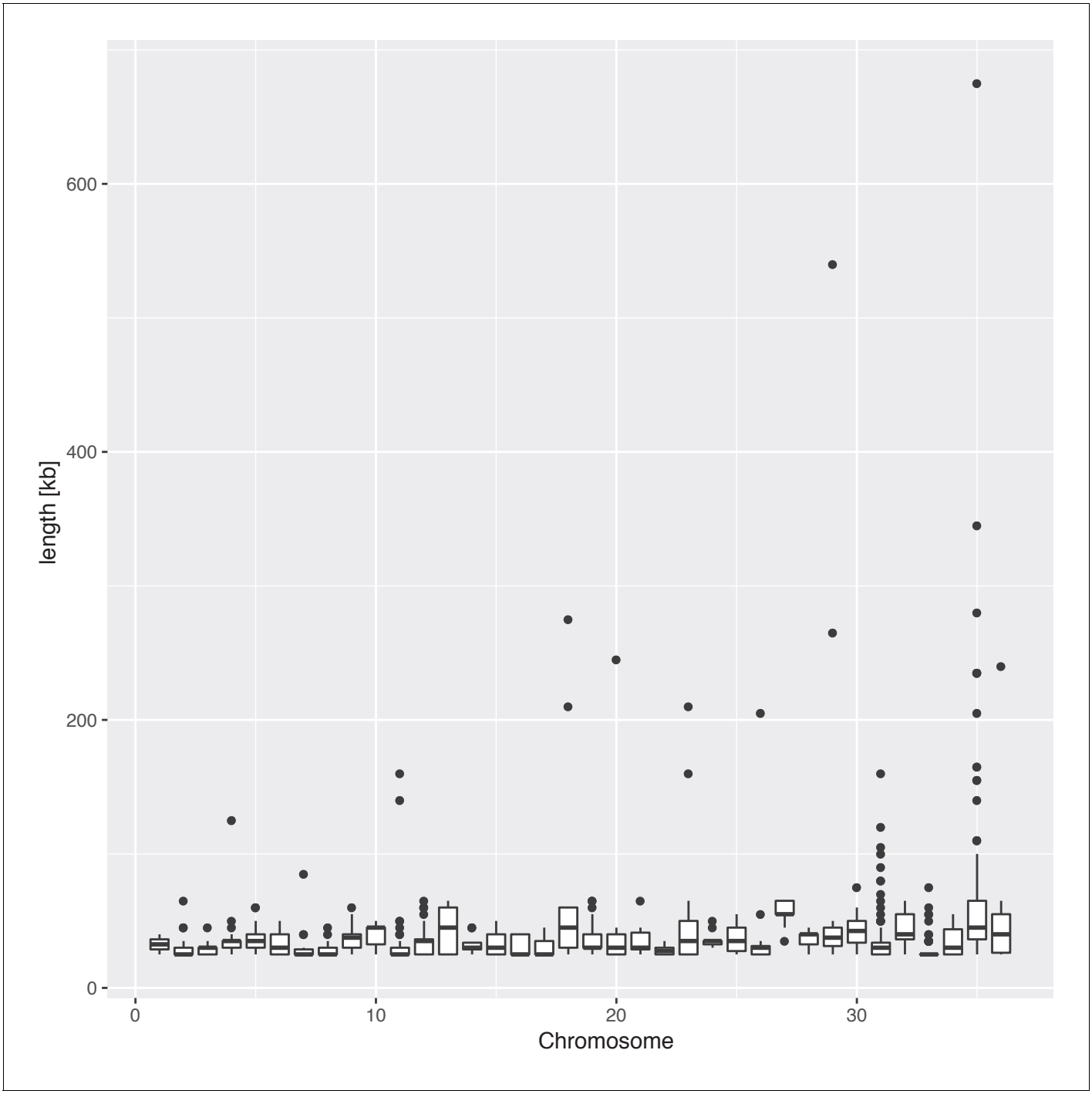

Figure 7—figure supplement 1. Length distribution of large CNVs by chromosome. Large CNVs were called using a minimum length threshold of 25 kb (see Materials and methods).

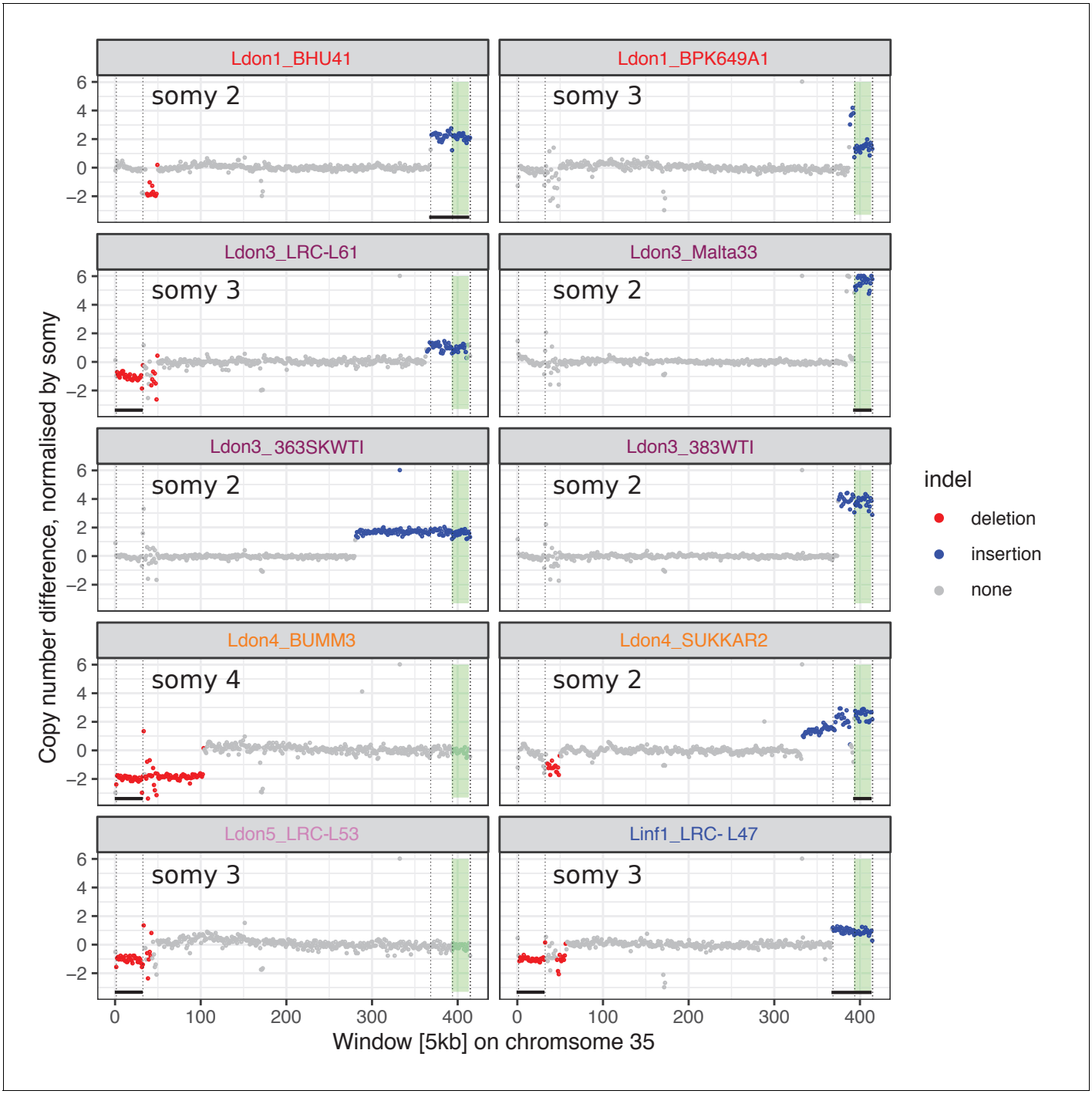

Figure 7—figure supplement 2. Most chromosome scale CNVs are located on chromosome 35. Shown are genome coverages of chromosome 35 for all samples that harbour at least one chromosome-scale CNV (>100 kb). Genome coverage for 5 kb windows was normalised by the sample and chromosome specific somy and coloured in red and blue for duplications and deletions, respectively. The respective chromosome-specific somy is indicated in each plot. Vertical lines mark indel boundaries and horizontal black bars below indicate indels with shared identical boundaries between samples. The green rectangles mark the CD1/LD1 locus sequences for *L. infantum* described in ***Sunkin et al. (2001)***. Group origin of the different samples is indicated by the group colours used throughout this study.

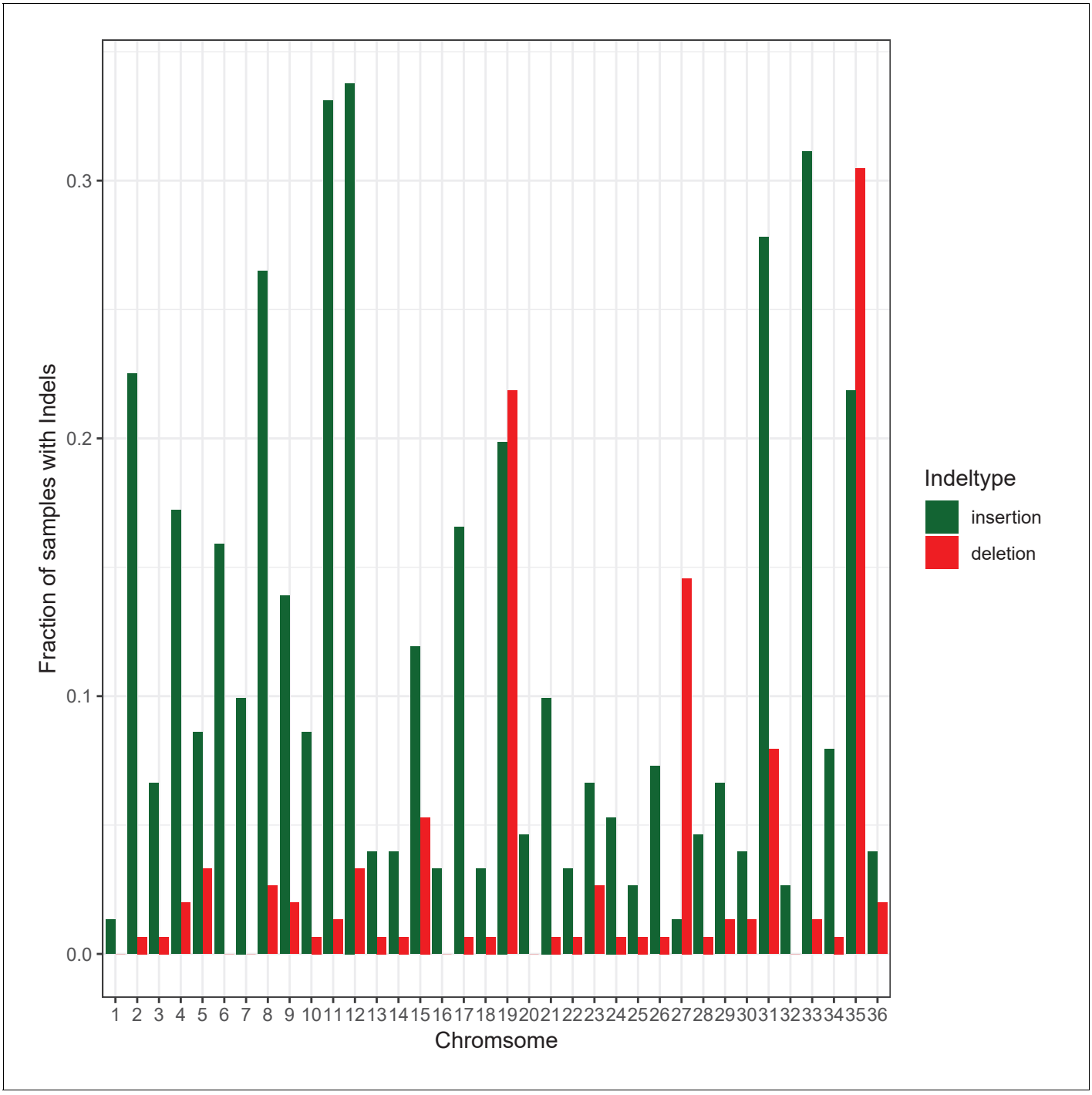

Figure 7—figure supplement 3. Fraction of large CNVs across chromosomes. Shown is the fraction of all 151 samples that contain at least one large copy number variant (>=25 kb; see Materials and methods) of the respective type for each chromosome.

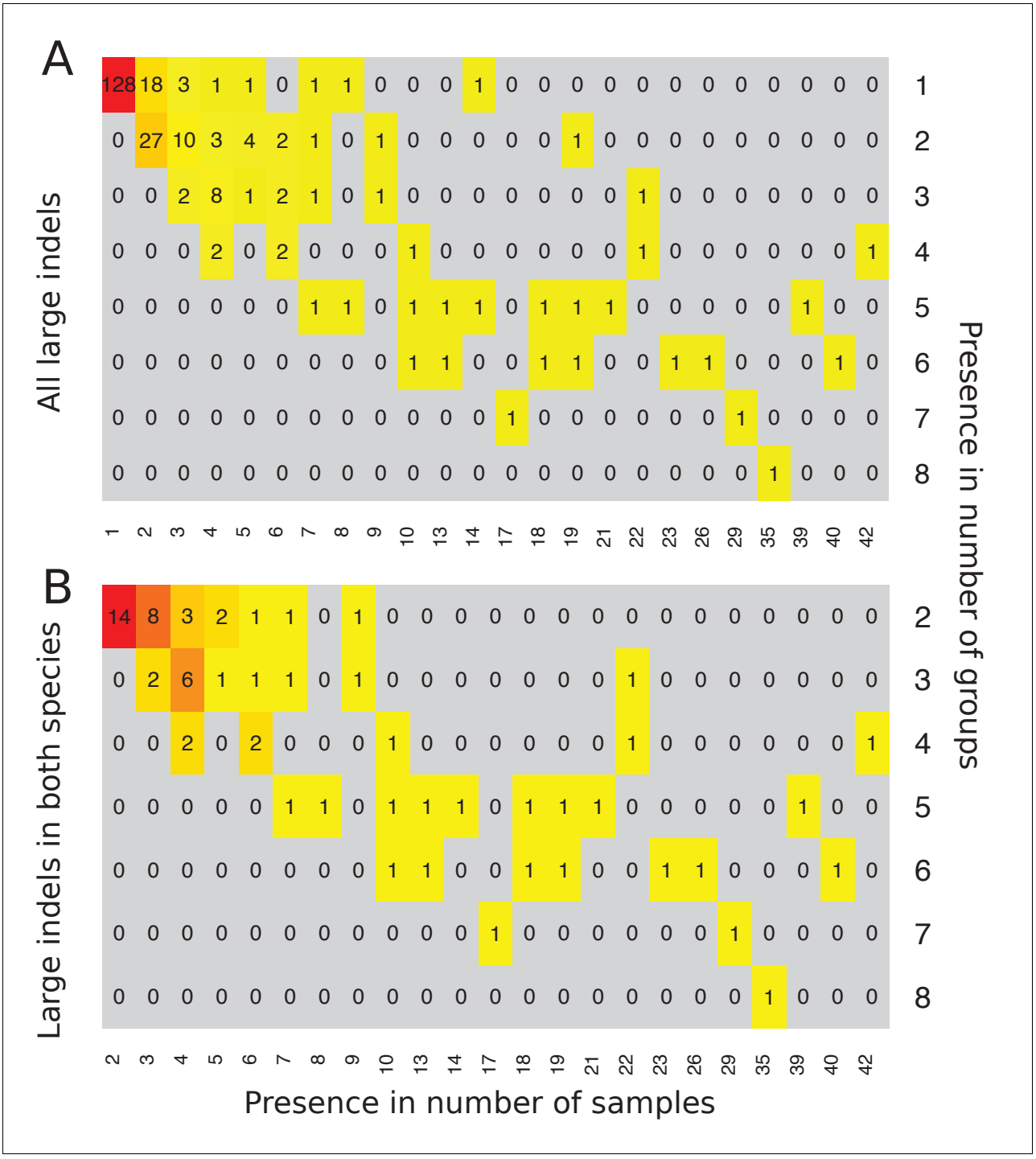

Figure 7—figure supplement 4. Large CNVs shared across samples and groups. Sharing of large CNVs (>=25 kb) is shown between samples and groups. (**A**) All large CNVs identified across all 151 isolates. (**B**) All large CNVs that have been found in both species, *L. donovani* and *L. infantum*.

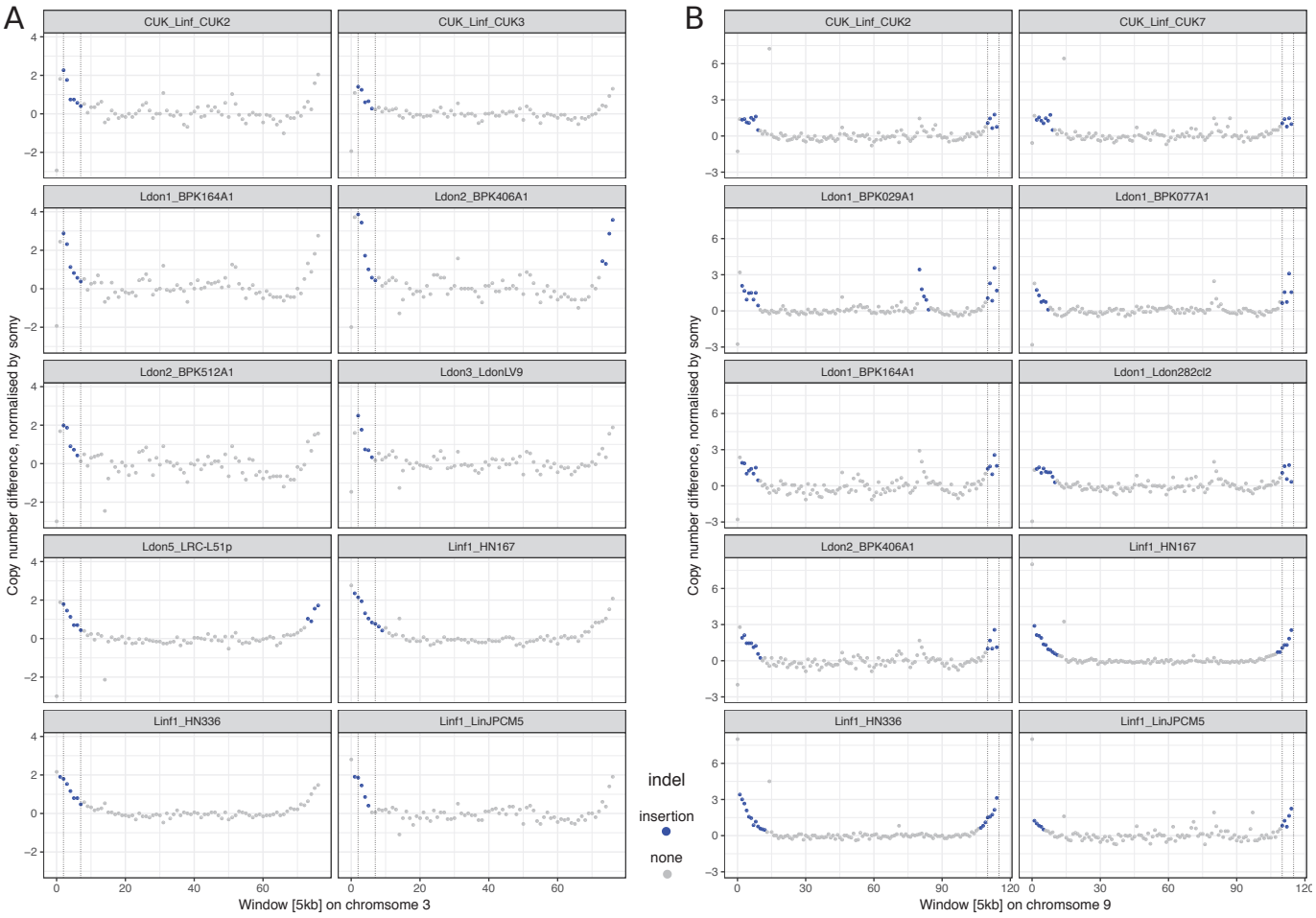

Figure 7—figure supplement 5. Increased coverage of samples towards chromosome ends. Samples are shown with called duplications at chromosome ends that show a gradual coverage increase. Plots show median window coverage across 5 kb windows. Called duplications are indicated by blue dots and boundaries are indicated by vertical bars. (**A**) Examples for chromosome 3. (**B**) Examples for chromosome 9.

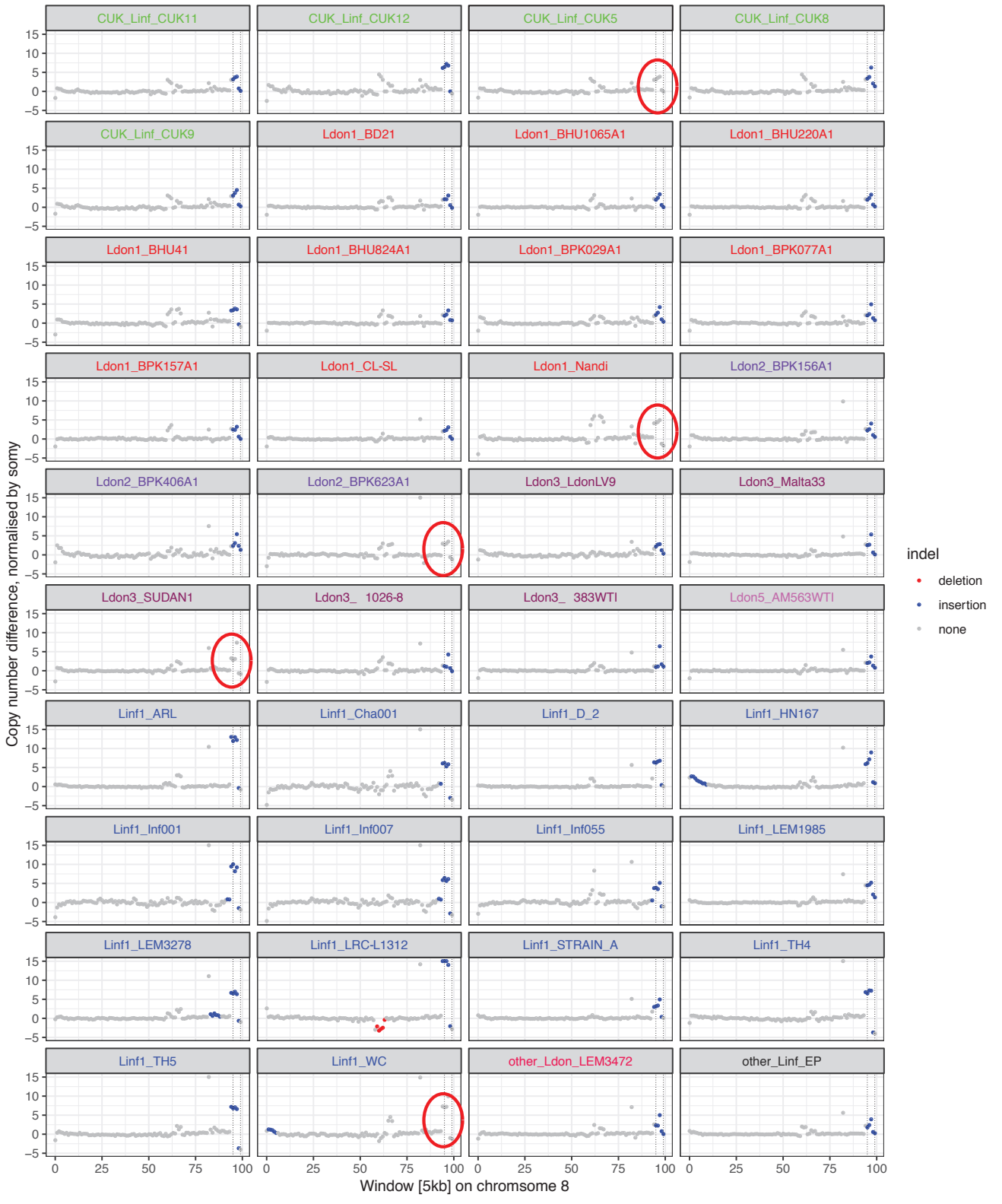

Figure 7—figure supplement 6. Indication of a putative assembly error in the reference genome. A common duplication of 25 kb on chromosome 8, position 470–495 kb, was found in 35 samples across eight different groups (indicated by vertical bars and highlighted in blue). When inspecting

Figure 7—figure supplement 6 continued on next page

Figure 7—figure supplement 6 continued

remaining samples, however, a copy number increase was also present in all other 116 remaining samples, which failed to meet the CNV calling threshold. Five of these 116 samples are also shown and the non-called copy number increase is indicated by a red circle. As the copy number increase varies between samples, these regions may still be copy number variable between isolates.

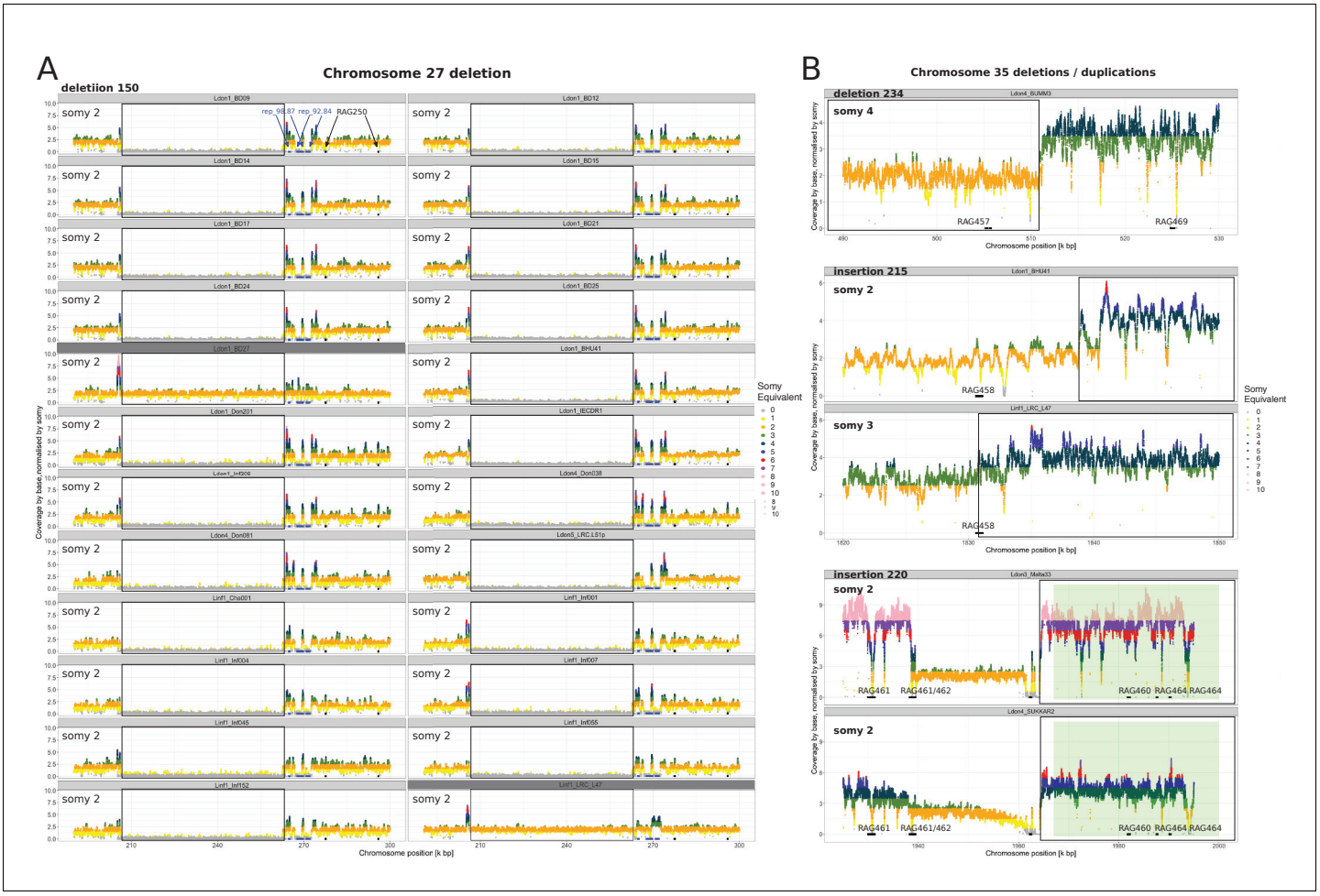

Figure 7—figure supplement 7. CNV association with repeat sequences in the genome. The per base coverage is shown for break-point regions of insertions on chromosomes 27 A) and 35 B) in relevant strains (***Figure 7*** and ***Figure 7—figure supplement 2***). The sequencing coverage is normalised by the haploid sequencing coverage estimated across all chromosomes for the respective strain. The somy of the respective strains and chromosome are indicated in the left top corner of each subplot and the local 'somy equivalent' is indicated by the respective colour. Repeated sequences described in ***Ubeda et al. (2014)*** are indicated by black bars and annotated with their repeat alignment group (RAG). Newly identified repeated sequences that were not present in the reference genome version used by ***Ubeda et al. (2014)*** are indicated by blue bars and are annotated with identity between two repeated sequences (A, see ***Figure 7—figure supplement 8***). The copy number variant type, that is insertion/deletion is indicated in the top left corner of each subplot along with its id as stated in ***Supplementary file 7*** and the region of the respective variant is indicated by a black frame. For deletion 150 in chromosome 27 the coverage is additionally shown for two samples that do not harbour the deletion as a control (indicated by a dark grey header, (**A**). The first half of CD1/LD1 locus sequences described in ***Sunkin et al. (2001)*** is present in the breakpoint region of insertion 220 and is indicated by the green rectangle (**B**).

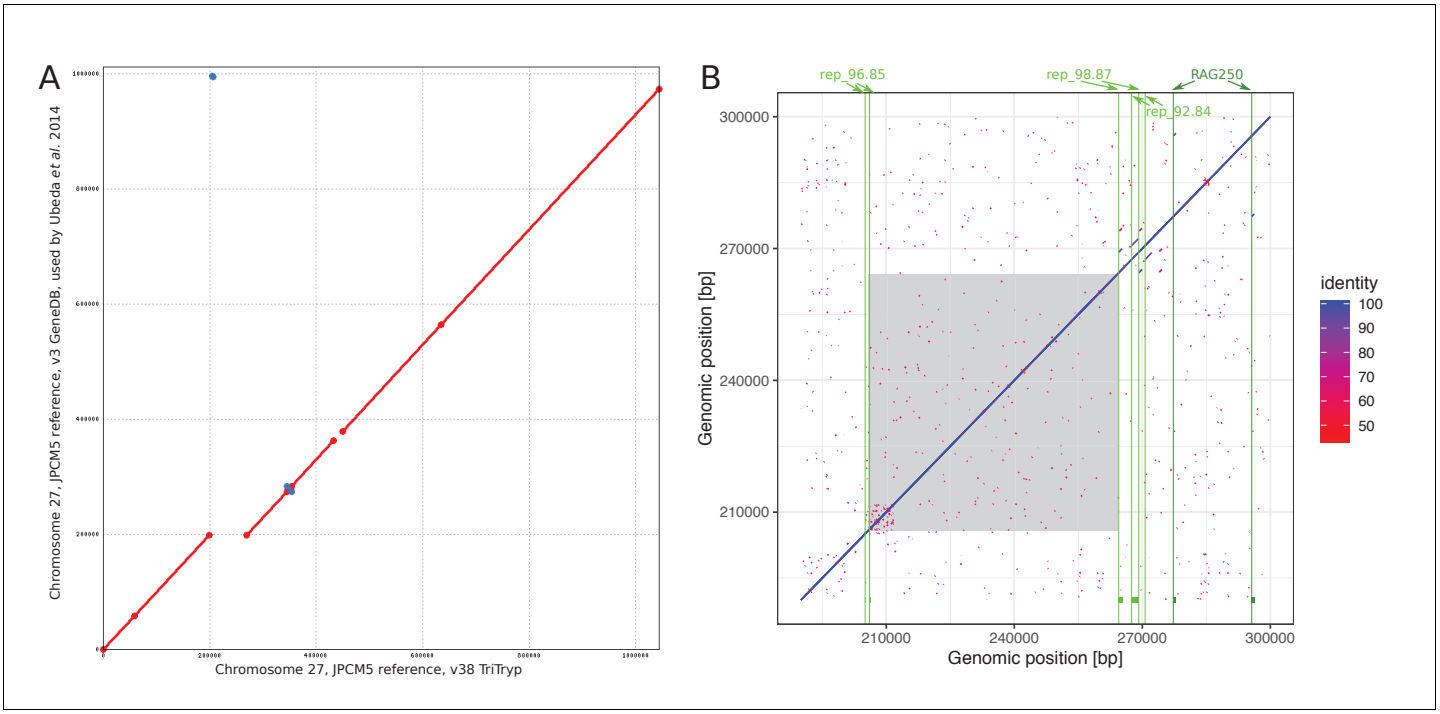

Figure 7—figure supplement 8. Identification of novel repeated sequences on chromosome 27. (**A**) Positions 199,468–269,164 in the *L. infantum* JPCM5 reference genome (TriTrypDB, (**v38**) were not present in the reference assembly used by ***Ubeda et al. (2014)***. (**B**) Repeated sequences within chromosome 27, positions 190,000–300,000 in the reference genome, JPCM5 (TriTrypDB v38). The dot plot shows the comparison of the sequence region against itself. Similar sequences are coloured by their % identity. The grey shaded area indicates the common deletion found in a subset of all our strains (***Figure 7A***, ***Figure 7—figure supplement 7A***). Green bars at the bottom indicate the location of repeat regions with a vertical line indicating their start position. Dark green indicated repeats originally described in ***Ubeda et al. (2014)*** and light indicates newly identified repeats. Coordinates of the newly identified repeats are summarised in ***Supplementary file 13***.

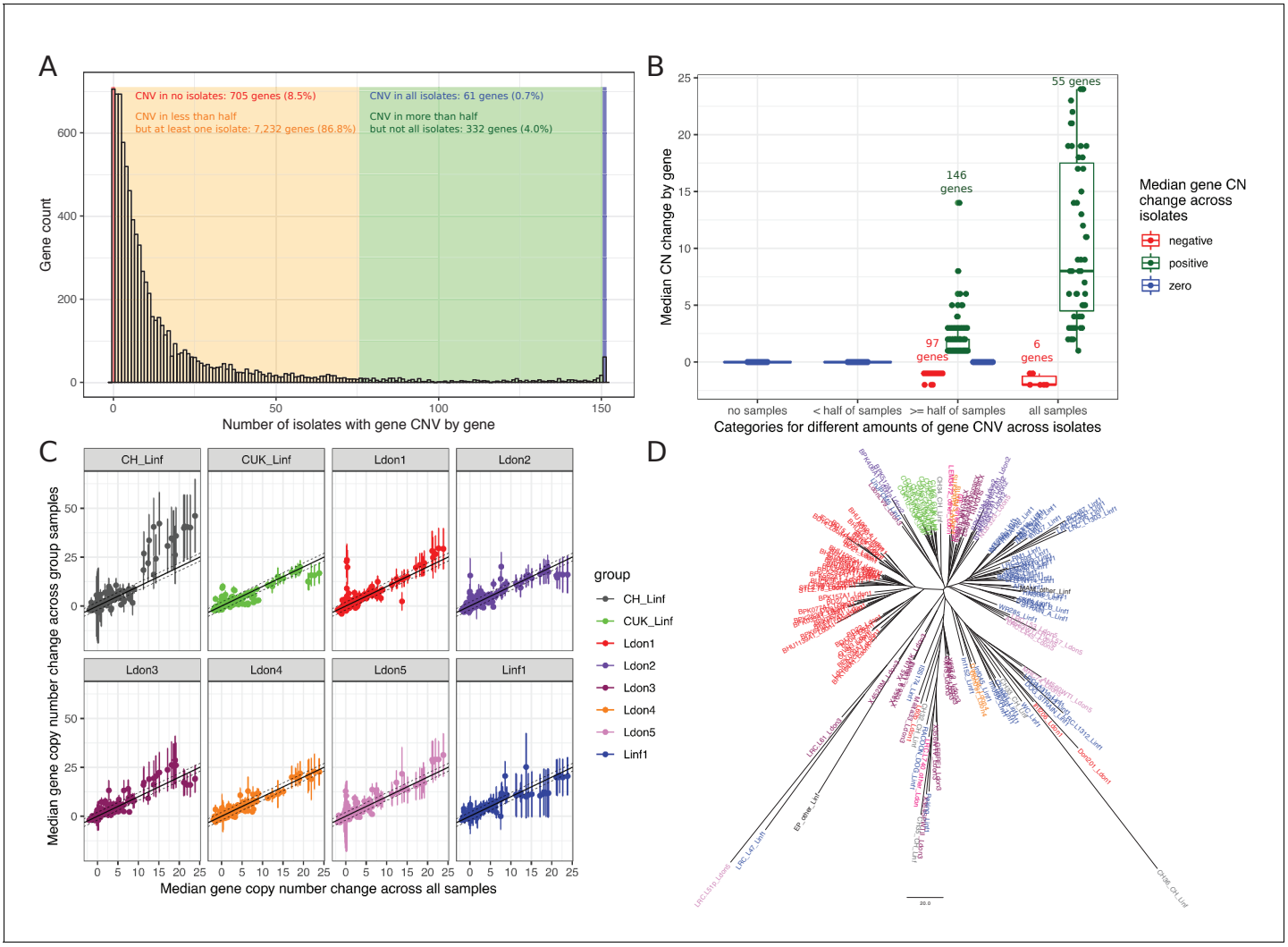

Figure 8. Gene copy number variation across groups. (**A**) CN abundances by gene across all 151 isolates. Genes are grouped in four categories (identified by different colours) depending on how many isolates are affected by CN variation in the respective gene. (**B**) Median copy number changes for each gene are shown (individual dots) and summarised for the four different categories also used in sub-figure A including the direction of effect sizes using boxplots. (**C**) Correlations of the median gene copy number across all samples and each respective phylogenetic group. (**D**) Neighbour joining tree using gene CN profiles for each sample.

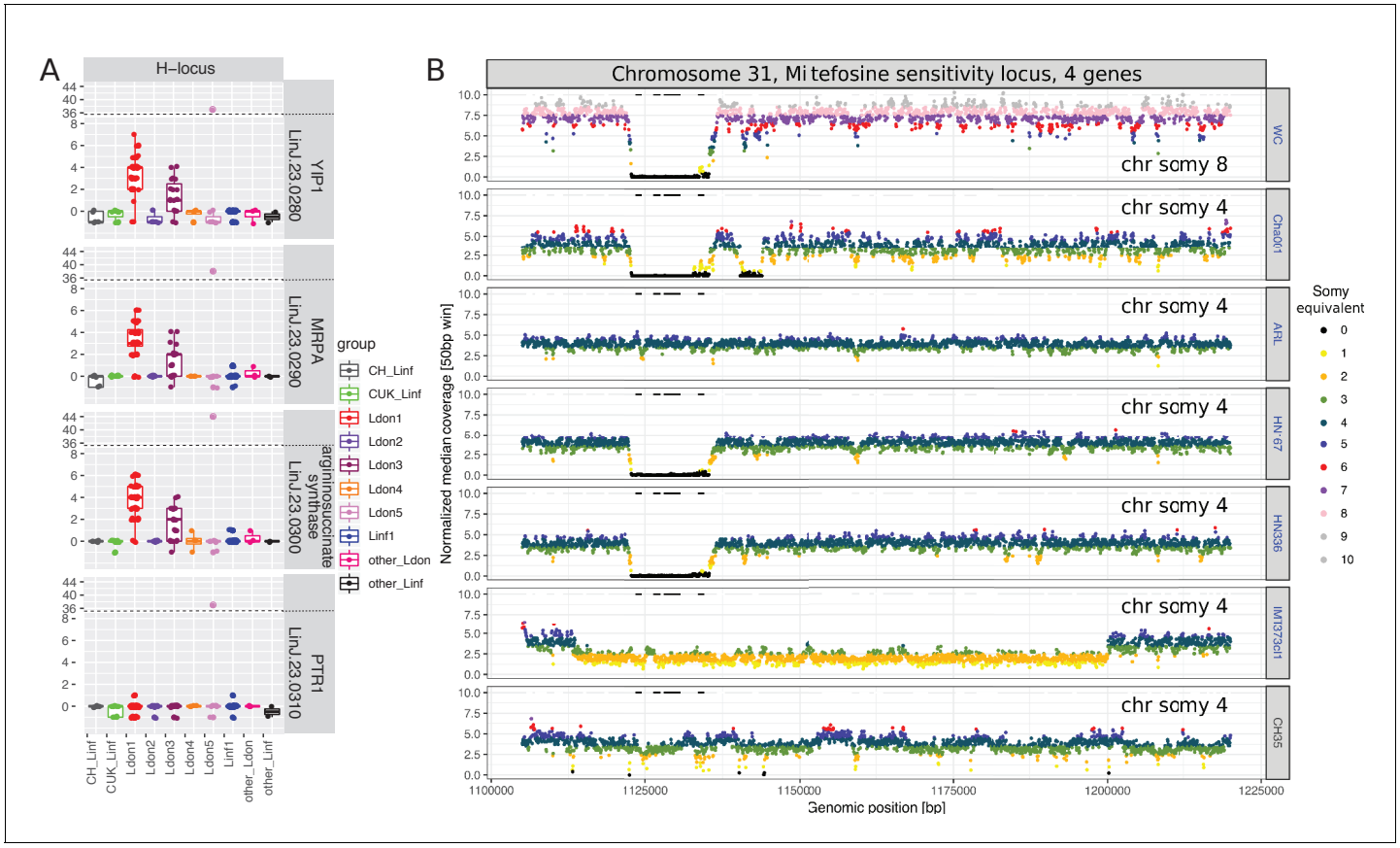

Figure 9. Copy number variation of putative drug resistance genes. (**A**) Copy numbers (CNs) for all four genes on the H-locus are shown for all 151 samples across all 10 different (sub-)groups. (**B**) Genome coverage in the genomic regions surrounding the MSL in all six samples showing a deletion and one sample with no CN reduction. Genome coverage for 50 bp windows is normalised by the haploid chromosome coverage and colours indicate the somy equivalent coverage of the respective window. The genes, LinJ.31.2370, LinJ.31.2380, LinJ.31.2390 and LinJ.31.2400, are marked as black horizontal lines. Colours of the sample names indicate group colours used throughout this study.

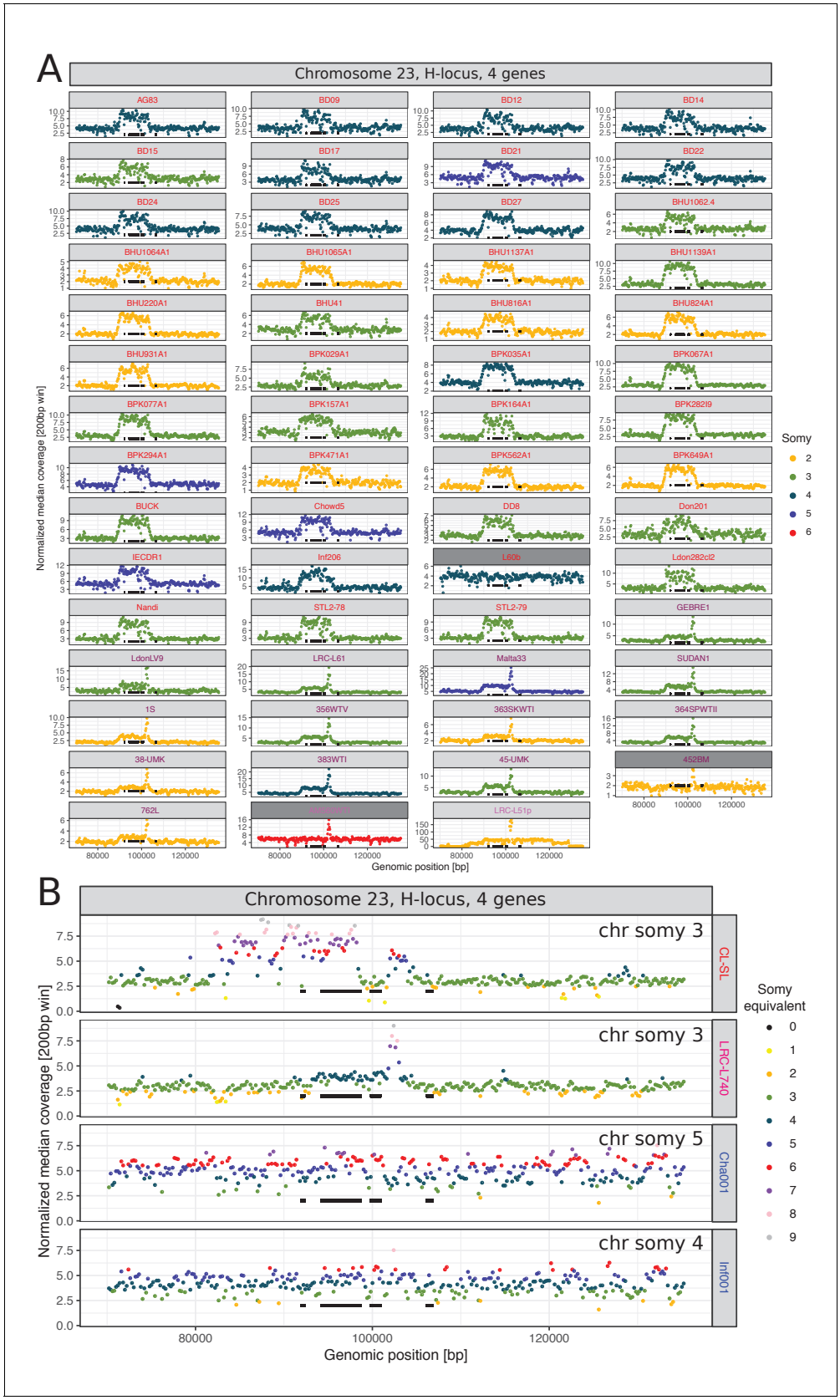

Figure 9—figure supplement 1. Copy number increase at the H-locus. Shown are coverage plots across isolates for the H-locus on chromosome 23 highlighting the four genes (YIP1, MRPA, argininosuccinate synthase, PTR1) present at this locus by black, horizontal bars. Isolate names are coloured

Figure 9—figure supplement 1 continued on next page

Figure 9—figure supplement 1 continued

by the group colours used throughout this study. (**A**) Copy number variation is shown for all 37% of isolates that have a copy number increase of at least three genes at the H-locus. Coverages of each isolate are coloured by the somy of chromosome 23 in the respective isolate. For comparison for each group, the coverage of one isolate with no copy number increase is also plotted (dark grey headers). (**B**) Copy number variation of all isolates that show a copy number increase in only two of the associated genes. Window-specific coverages are coloured by its somy equivalent. Somies of chromosome 23 for each isolate are indicated in the respective row.

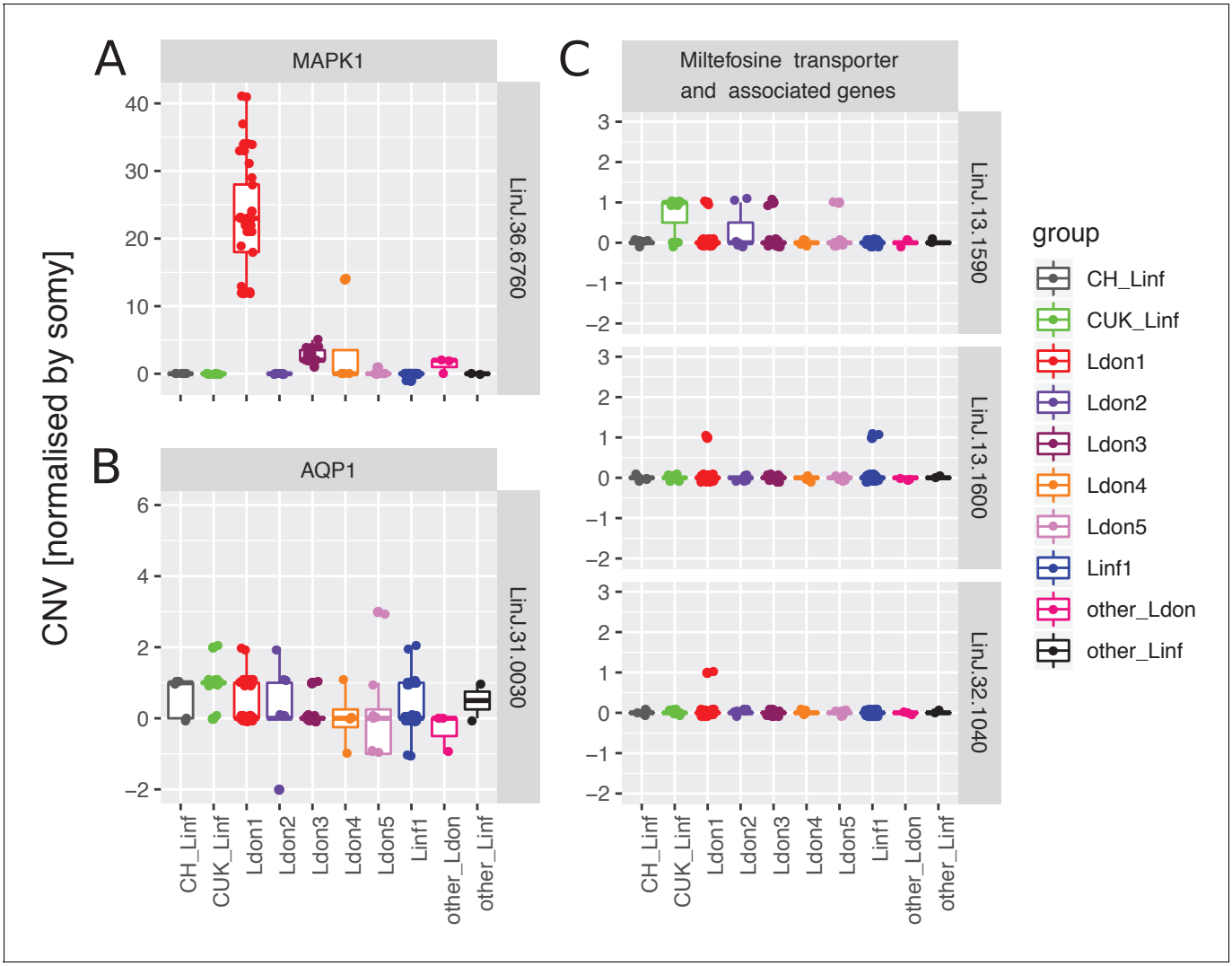

Figure 9—figure supplement 2. Copy number variation of putative drug resistance genes. Gene copy number for all four genes and all 151 samples is shown across 10 different groups for three different loci putatively involved in drug resistance. (**A**) MAPK1 (LinJ.36.6760) and (B) AQP1 (LinJ.31.0030) genes are putatively involved antimony drug resistance. (**C**) The Miltefosine transporter (LinJ.13.1590), the adjacent gene (LinJ.13.1600) and the Ros3 (LinJ.32.1040) gene are putatively involved in Miltefosine resistance.

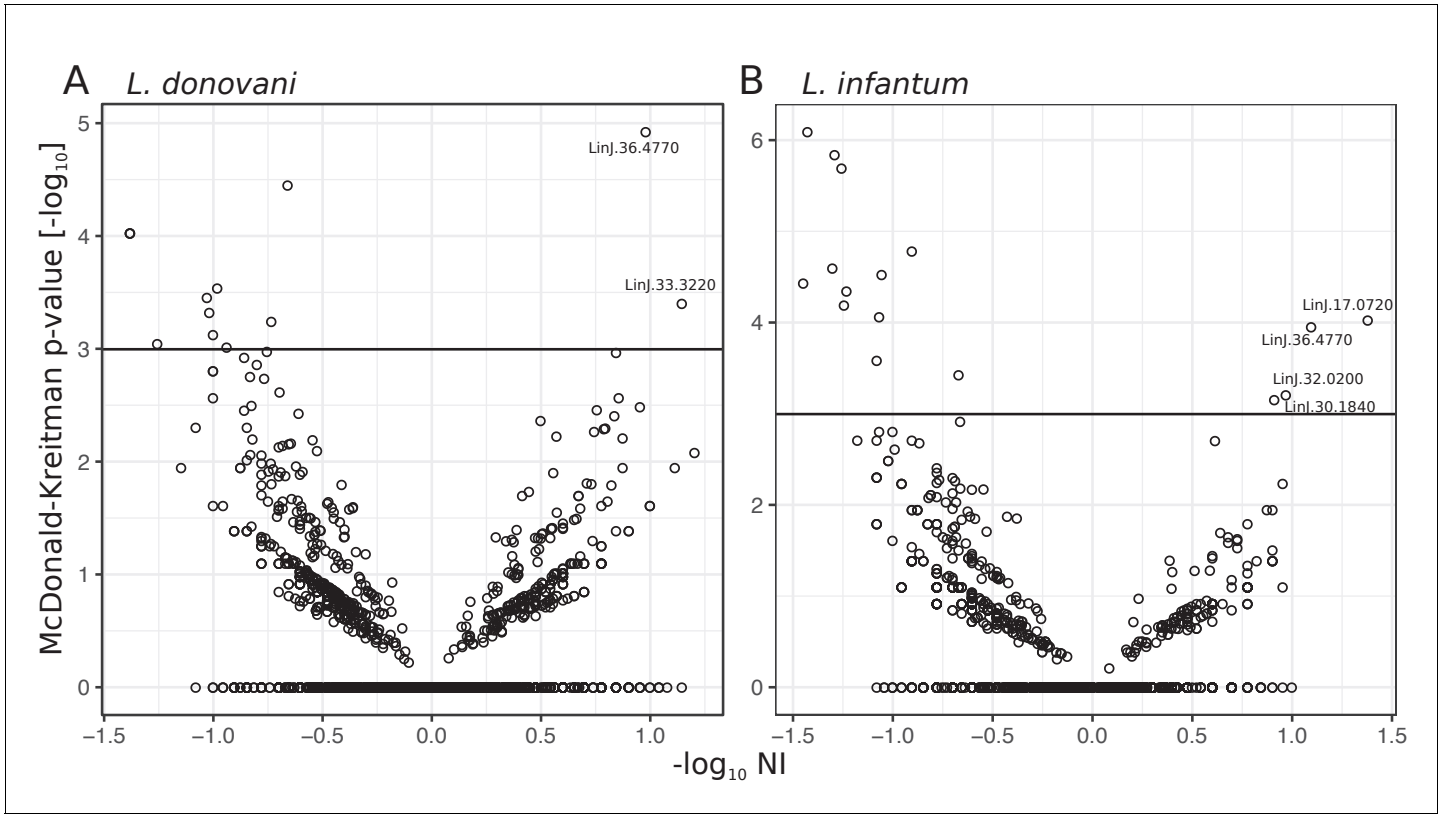

Figure 9—figure supplement 3. Measures of adaptive evolution. Species species-specific evolution is measured for (**A**) *L. donovani* and (**B**) *L. infantum*. Each point represents the neutrality index (NI) and the associated p-value of the McDonald-Kreitman test for each of 8234 genes. The horizontal line represents the. -$\log_{10}$ value equivalent to a p-value of 0.05. None of the shown values passes multiple testing correction.

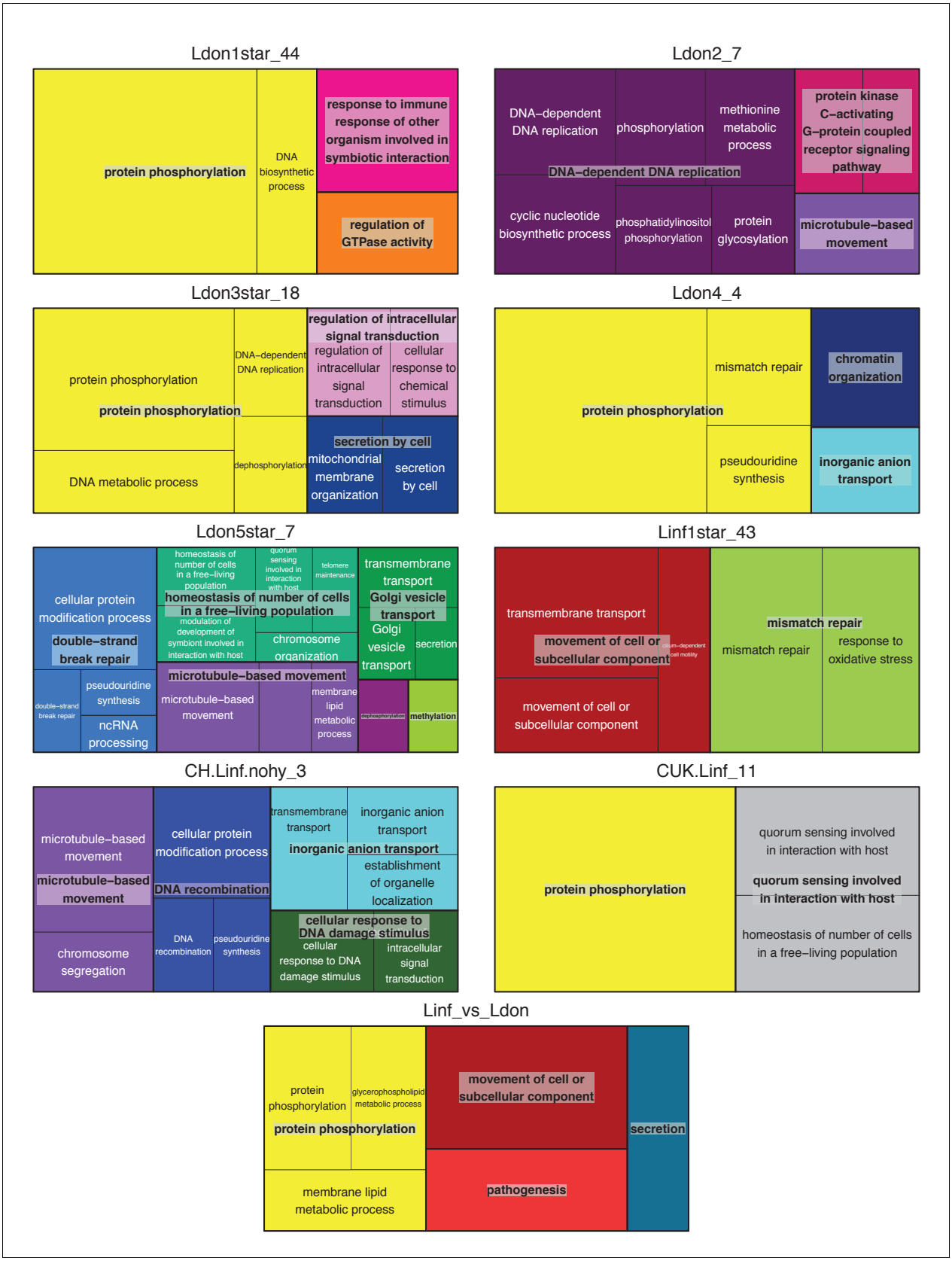

Figure 9—figure supplement 4. Gene ontology enrichment of marker genes with putative biological impact. For each group our species-specific GO enrichment, biological process results are shown for genes with at least one moderate or high effect variant according to SNPeff annotation

Figure 9—figure supplement 4 continued on next page

Figure 9—figure supplement 4 continued

(***Supplementary file 3***). Plots show enrichment for the lenient cut-off of a p-value<0.05 using the weighted Fisher test statistic (weightFish, topGO, ***Alexa et al., 2006***) using Revigo (***Supek et al., 2011***). Sizes of rectangles are normalised by absolute $\log_{10}$ p-value. Plot titles indicate the group marker sets or the species comparison, respectively. Stars in the group names indicate that samples that have previously been identified as mixtures of clones have been removed (***Table 1*** B3 and B4). Additionally, hybrids between the major groups (***Table 1*** B2) were removed from this analysis. Group sample sizes are indicated at the end of each name.