Science Advances

AAAS

advances.sciencemag.org/cgi/content/full/4/4/eaaq0392/DC1

Supplementary Materials for

Improved de novo genomic assembly for the domestic donkey

Gabriel Renaud, Bent Petersen, Andaine Seguin-Orlando, Mads Frost Bertelsen, Andrew Waller, Richard Newton, Romain Paillot, Neil Bryant, Mark Vaudin, Pablo Librado, Ludovic Orlando

Published 4 April 2018, *Sci. Adv.* **4**, eaaq0392 (2018)
DOI: 10.1126/sciadv.aaq0392

This PDF file includes:

- table S3. Human phenotypes, human diseases, and pathways associated with genes enriched in detected ROHs.
- table S4. Horse sequences used for the detection of donkey scaffolds pertaining to the Y chromosome.
- table S5. Heterozygosity rates for various species of asses and zebras computed when aligning to the donkey reference described in this study and recomputed on the basis of the data reported by Jónsson *et al.* (*9*), which were aligned to the horse reference.
- table S6. Listing missing proteins in complete and partially complete Eukaryotic Orthologous Groups from the Core Eukaryotic Genes Mapping Approach.
- table S7. Repeat elements and low-complexity DNA sequences masked in the donkey genome using RepeatMasker.
- table S8. Repeat elements and low-complexity DNA sequences masked in the donkey genome using the second of the RepeatMasker using the model generated from RepeatModeler as custom library input on the previously masked genome.
- table S9. Statistics of the completeness of the different versions of the donkey genome based on 248 Core Eukaryotic Genes.
- References (*44–62*)

Supplementary Materials and Methods

section S1.

Chicago library sequencing

Two Chicago libraries were prepared as described previously (*14*). Briefly, for each library, 500 ng of high-molecular-weight genomic DNA (~50 kb mean fragment size) was reconstituted into chromatin *in vitro* and fixed with formaldehyde. Fixed chromatin was then digested with the MboI enzyme, the 5' overhangs were filled in with biotinylated nucleotides and then free blunt ends were ligated. After ligation, crosslinks were reversed and the DNA purified from protein. Purified DNA was treated to remove biotin that was not internal to ligated fragments. The DNA was then sheared to ~350 bp mean fragment size and sequencing libraries were generated using NEBNext Ultra enzymes and Illumina-compatible adapters. Biotin-containing fragments were then isolated using streptavidin beads before PCR enrichment of the library. The libraries were sequenced on the Illumina (*44*) HiSeq 2500 instrument at the Danish National DNA Sequencing Center to produce 365M 2X150 bp read pairs, providing 101x physical coverage (1-50 kb pairs). Physical coverage measures the average number of times that a read-pair of 1-100 kb span a given nucleotide in the genome.

Scaffolding the draft genome with HiRiSE and quality metrics

A draft genome assembly previously reported (*2*) , representing 2,320 Mb with a scaffold N50 of 434 kb, Illumina shotgun sequence data, and Chicago library read pairs in FASTQ format were used as input data for HiRiSE, a software pipeline designed specifically for using Chicago data to assemble genomes (*14*). Shotgun and Chicago library sequences were aligned to the draft input assembly using a modified SNAP read mapper (http://snap.cs.berkeley.edu). The separation of Chicago read pairs mapped within draft scaffolds were analyzed by HiRise to produce a likelihood model for genomic distance between read pairs, and the model was used to identify putative misjoins and score prospective joins. After scaffolding, shotgun sequences were used to close gaps between contigs. Quality metrics for this assembly were obtained with Quast (*45*) with default parameters.

Repeat masking

Repeats and low complexity DNA sequences were masked in the genome prior to gene annotation using RepeatMasker version 4.0.5 (*46*) using the species repeat database 'mammals' with RepBase Update 20150807, RM database version 20150807 **(**table S7**).**

Remaining donkey specific repetitive elements were predicted *de novo* using RepeatModeler version open-1.0.8 (*46*) on the masked genome. Subsequently, a second round of RepeatMasker was run with the model generated from RepeatModeler as custom library input on the previously masked genome **(**tables S8-S9**).**

Gene Annotation

Genome annotation was performed using the genome annotation pipeline Maker2 version 2.31.8 (*47*) with ab-initio and homology-based gene predictions. Protein sequences from *homo sapiens, Equus caballus* and *Mus musculus* was used for homology-based gene prediction. As no training gene models were available for Equids, we used CEGMA (*48,49*) to train the ab-initio gene predictor SNAP (*50*)**,** rather than using the *de-novo* gene predictor Augustus (*51*) . Maker2 was run with "model_org=simple, softmask=1, augustus_species=human" and the "snaphmm" parameter was set to the HMM generated in the manual training of SNAP. Missing proteins in complete and partial complete KOGs can be found in table S6.

Orthologs in the horse genome were obtained using OrthoFinder (*52*) with default parameters. The horse protein-coding genes were obtained from Ensembl Genes (EquCab2.0, version 86). The parsing of the output obtained in OrthoFinder was done in-house using custom scripts. The gene symbols for the horse proteins were obtained using Biomart (*53*).

Heterozygosity and estimates of effective population size

Mapping of shotgun data from different *Equus* species to the donkey reference was performed using BWA v. 0.5.9 (*54*) with default parameters. Heterozygosity rates, both globally and locally, were computed using ANGSD v. 0.915-26 (*55*). Confidence bounds for the rate of

heterozygosity were obtained using a standard error interval for a binomial distribution. The local estimates of heterozygosity were performed using a window size of 50 kb with a step of 10 kb. The effective population size over time for the different species aligned to this donkey reference was performed using PSMC v. 0.6.5-r67 (*56*) using a base quality filter of 35, and parameters "-N25 -r5 -p 4+25*2+4+6". The results were plotted using mutation rate of $7.242x10^{-9}$ mutations per generation and site, and assuming a generation time of eight years. To minimize biases due to sex chromosomes, only donkey scaffolds aligning to horse autosomal chromosomes were considered.

Runs of homozygosity were identified using overlapping windows of 50 kb with a heterozygosity rate consistently less than the overall average of 0.068% with a total combined length greater than 500 kb. The analysis for pathway enrichment was performed using WebGestalt (*57*), using the total set of annotated genes for the donkey genomes as background reference set. Again, only genes within scaffolds that were aligned to horse autosomal chromosomes were considered. The ROHs plots were generated using the qqman package (*58*)

Genome-wide alignments

The alignment of the different scaffolds to their corresponding genome chromosomes was performed using the nucmer program part of mummer package (*59*). This correspondence was established using the same 101-mers alignments that were used for the synteny plot.

As the orientation of the scaffold was not known *a priori*, we oriented the scaffolds in order to minimize the number of rearrangements. Furthermore, by leveraging on the chromosome map between horse and donkey genomes described in (*60*), we manually reverse complemented certain scaffolds to make sure that the orientation of the scaffolds was consistent with their map. However, in the main genome-wide plot presented in Fig. 4, the orientation of the scaffolds was selected using an automated procedure which maximizes similar chromosomal strands rather than minimizing the number of rearrangements.

The donkey divergence to the horse genome was computed using Nei's standard genetic distance (D) (*61*) using windows of 30 kb. Scaffolds were assigned as potentially coming from the Y chromosome by aligning donkey scaffolds to a set of 19 contigs from the horse Y chromosome (see table S4) using BLAT v.35 (*62*) with default parameters. To avoid spurious alignments to the X chromosome, the gene annotation for each scaffold potentially originating from the Y chromosome were aligned using NCBI Blast against the non-redundant protein database 'nr'. If a scaffold contained genes mapping to the X chromosome in horses, this scaffold was flagged as being potentially from the X chromosome rather than the Y one.

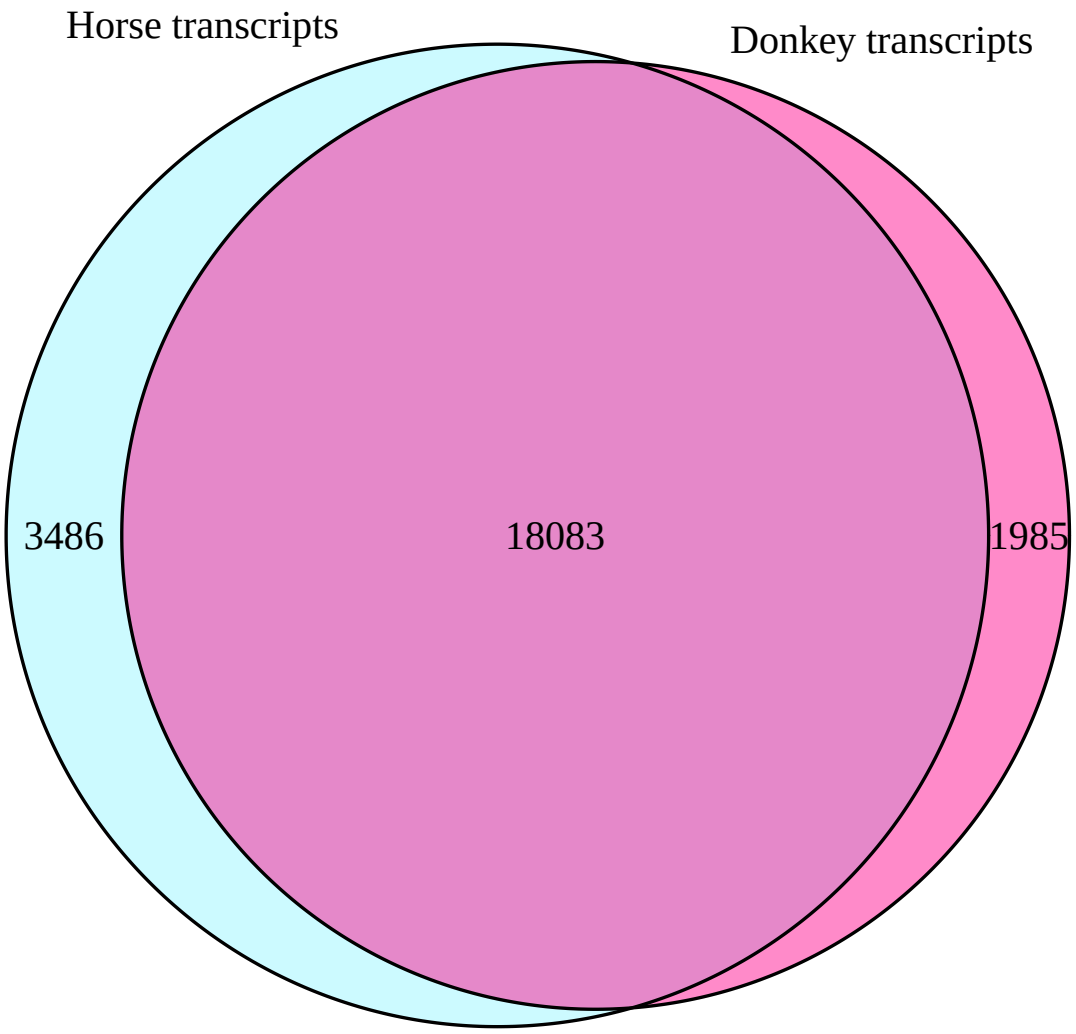


fig. S1. Venn diagram of the protein-coding genes that were annotated in the donkey assembly versus the protein-coding gene annotation for the horse. The reference for the horse (*Equus caballus)* genome was EquCab2 and Ensembl Genes (version 86) were used. The comparison to the horse annotation was performed using a single transcript per predicted protein-coding gene.

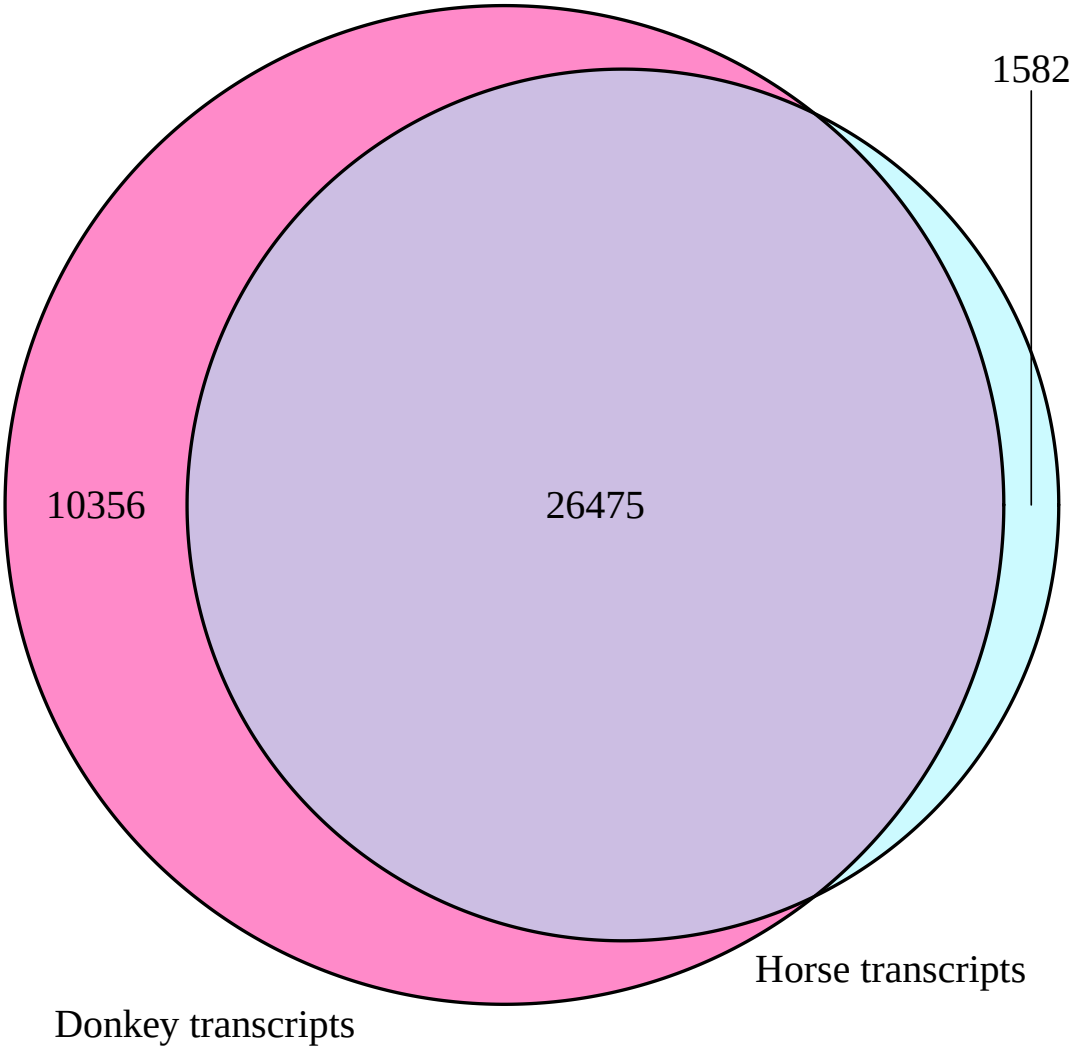


fig. S2. Venn diagram of the protein-coding genes that were annotated in the donkey assembly published by Huang *et al.* (*15*) versus the protein-coding gene annotation for the *E. caballus* genome (version EquCab2.0) using Ensembl genes (version 86).

Fig. S3

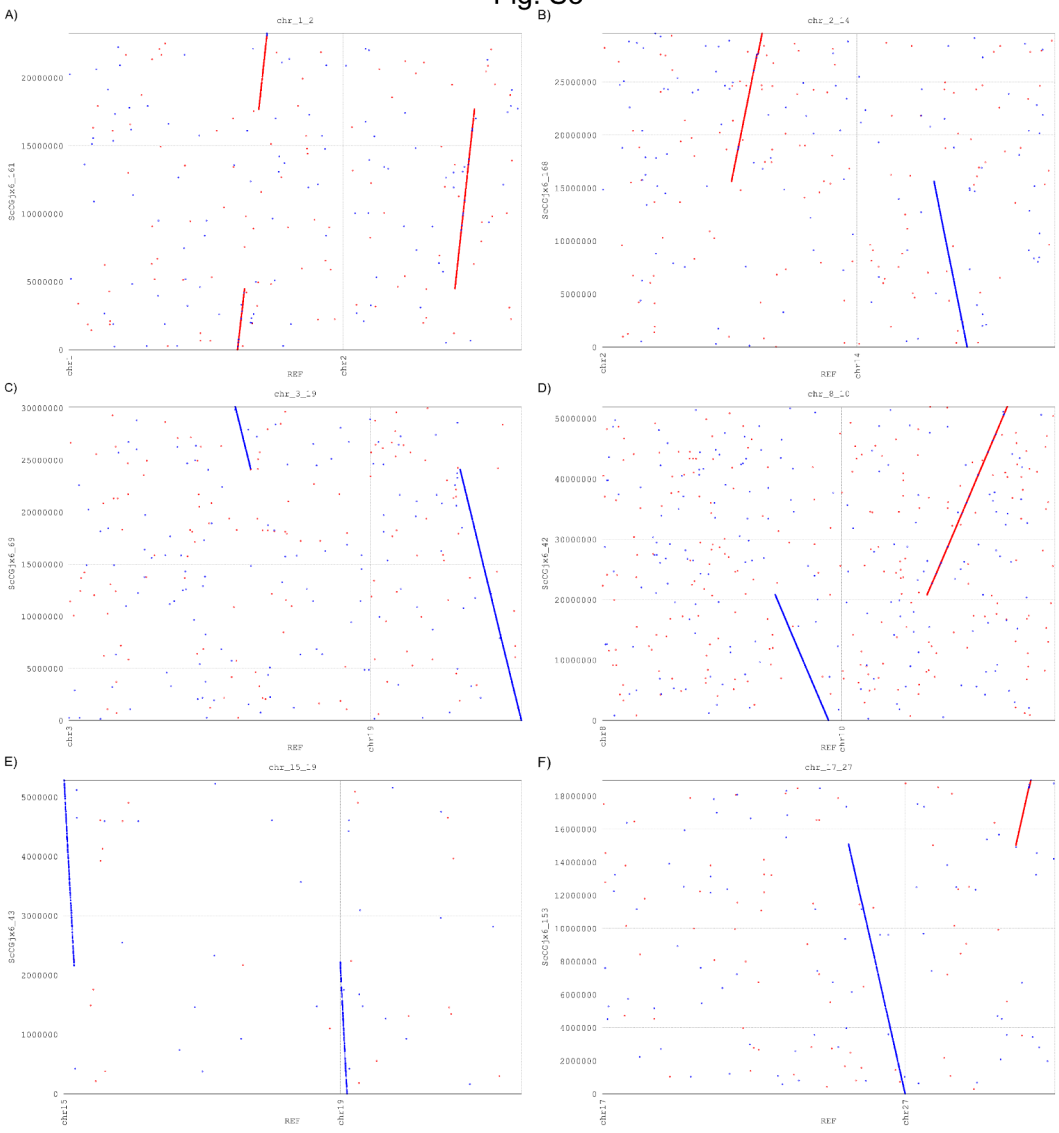


fig. S3. Alignment of horse chromosomes to six donkey scaffolds with putative signs of translocations. These alignments were performed with MUMmer v3.23.

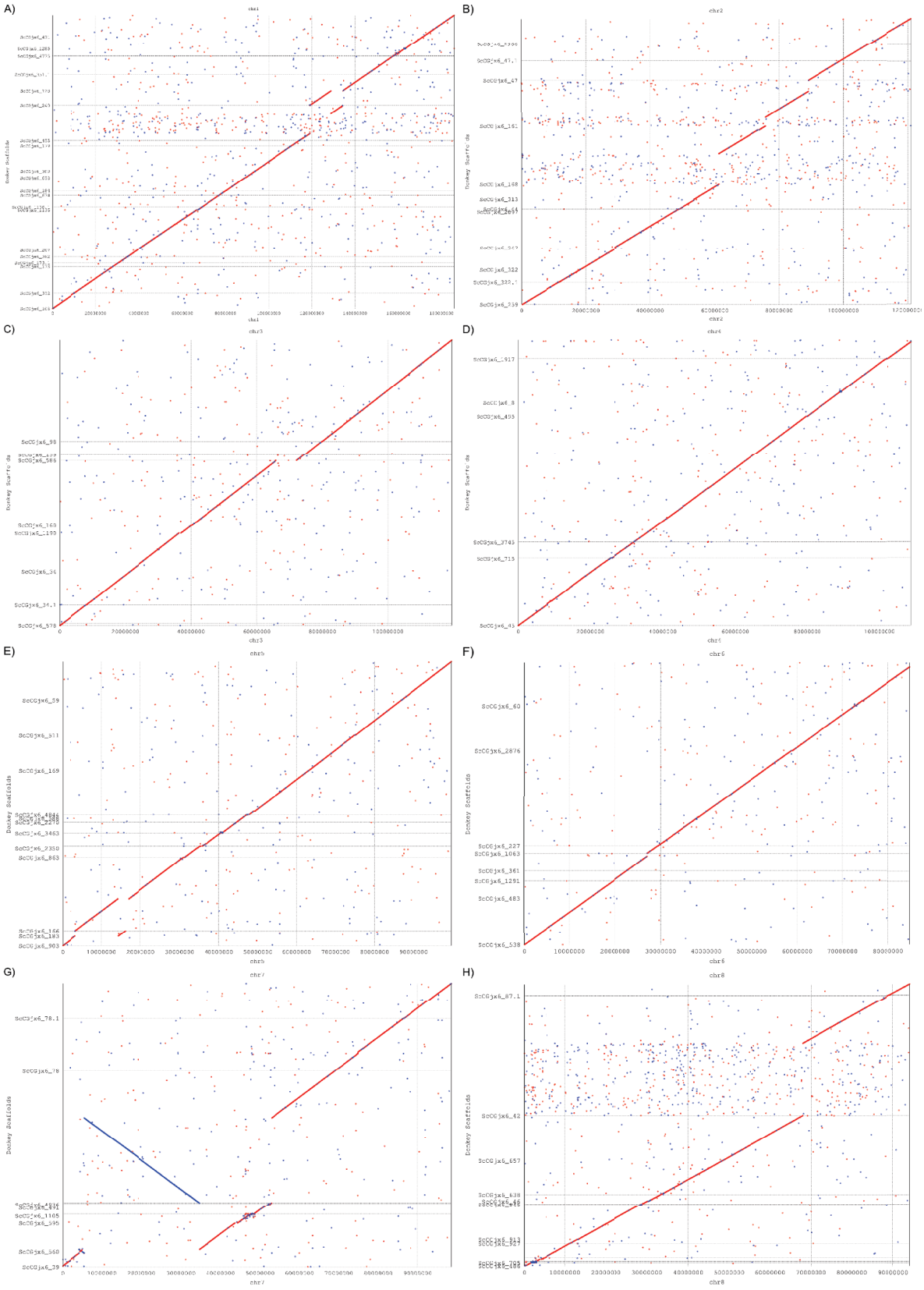


fig. S4. Alignment of donkey scaffolds to corresponding horse chromosomes. These alignments were performed with MUMmer v3.23.

Fig. S4 2 of 4

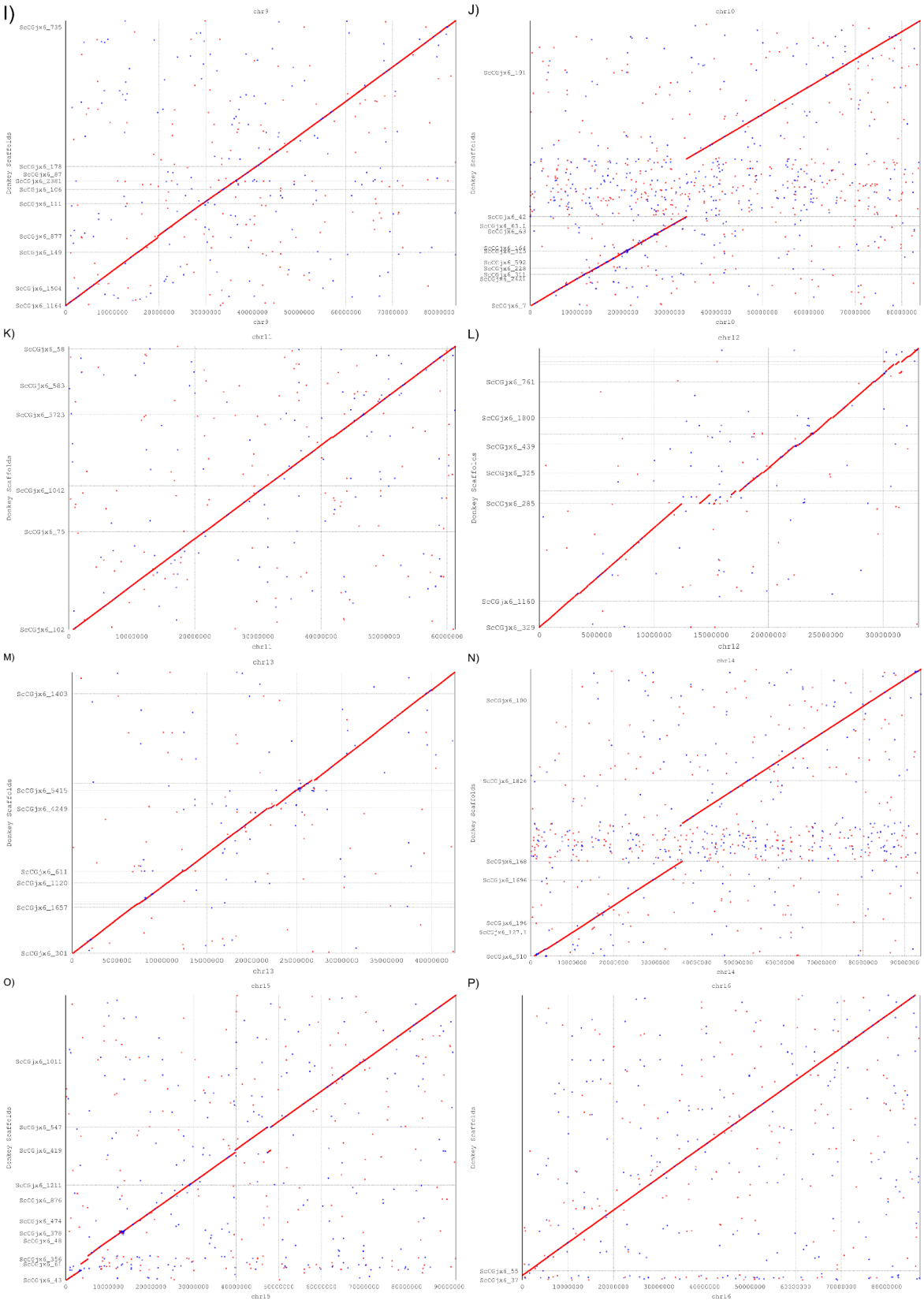

I)
chr9
Donkey Scaffolds
chr9
J)
chr10
Donkey Scaffolds
chr10
K)
chr11
Donkey Scaffolds
chr11
L)
chr12
Donkey Scaffolds
chr12
M)
chr13
Donkey Scaffolds
chr13
N)
chr14
Donkey Scaffolds
chr14
O)
chr15
Donkey Scaffolds
chr15
P)
chr16
Donkey Scaffolds
chr16

Fig. S4 3 of 4

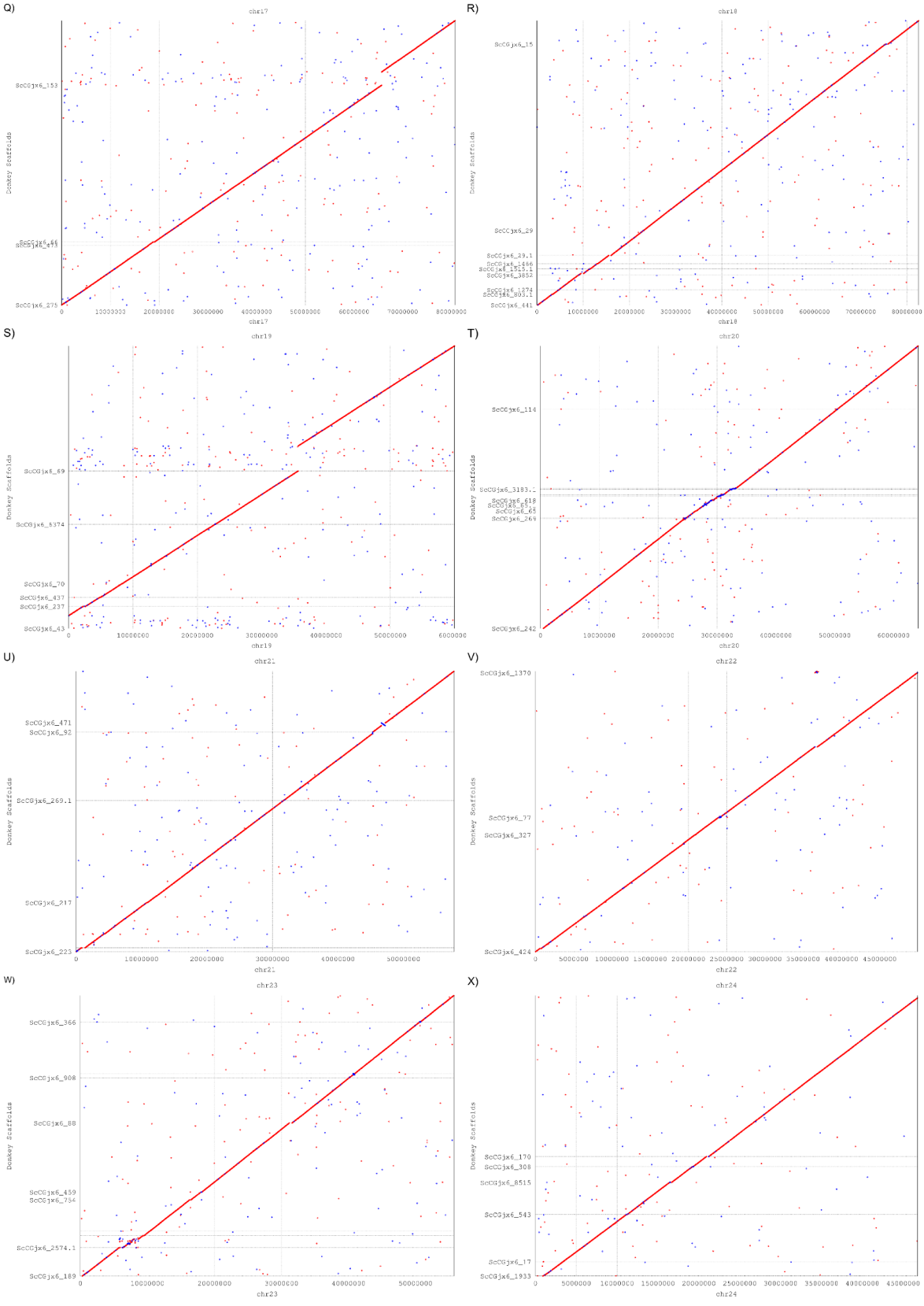
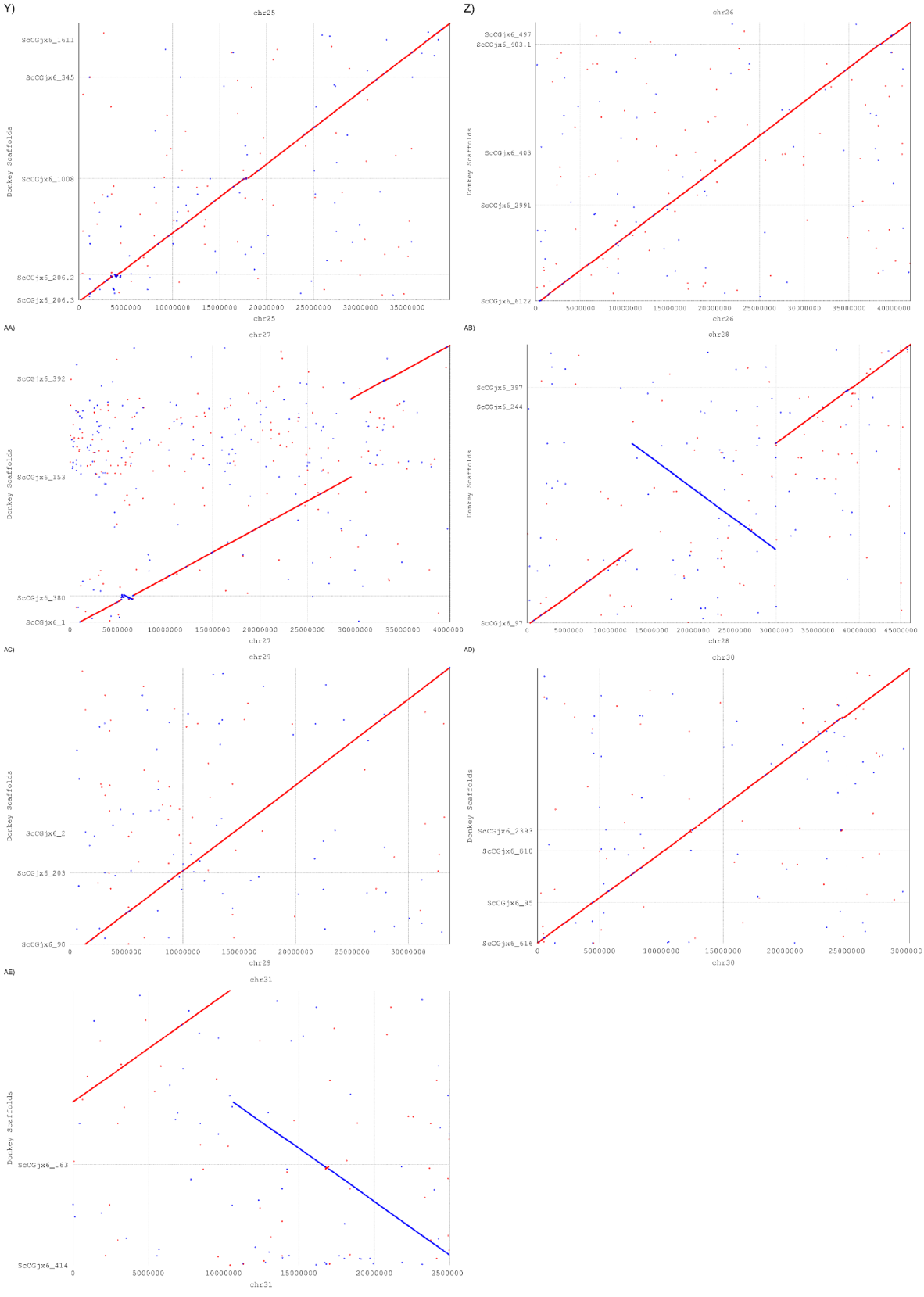

Q)
chr17
Donkey Scaffolds
ScCGjx6_153
ScCGjx6_275
chr17
R)
chr18
Donkey Scaffolds
ScCGjx6_15
ScCGjx6_29
ScCGjx6_29.1
ScCGjx6_441
chr18
S)
chr19
Donkey Scaffolds
ScCGjx6_69
ScCGjx6_5374
ScCGjx6_70
ScCGjx6_437
ScCGjx6_237
ScCGjx6_43
chr19
T)
chr20
Donkey Scaffolds
ScCGjx6_114
ScCGjx6_242
chr20
U)
chr21
Donkey Scaffolds
ScCGjx6_471
ScCGjx6_92
ScCGjx6_269.1
ScCGjx6_217
ScCGjx6_223
chr21
V)
chr22
Donkey Scaffolds
ScCGjx6_1370
ScCGjx6_77
ScCGjx6_327
ScCGjx6_424
chr22
W)
chr23
Donkey Scaffolds
ScCGjx6_366
ScCGjx6_908
ScCGjx6_88
ScCGjx6_459
ScCGjx6_734
ScCGjx6_2574.1
ScCGjx6_189
chr23
X)
chr24
Donkey Scaffolds
ScCGjx6_170
ScCGjx6_308
ScCGjx6_8515
ScCGjx6_543
ScCGjx6_17
ScCGjx6_1933
chr24

Fig. S4 4 of 4


Y)
chr25
ScCGjx6_1611
ScCGjx6_345
ScCGjx6_1008
ScCGjx6_206.2
ScCGjx6_206.3
Donkey Scaffolds
0
5000000
10000000
15000000
20000000
25000000
30000000
35000000
chr25
Z)
chr26
ScCGjx6_497
ScCGjx6_403.1
ScCGjx6_403
ScCGjx6_2991
ScCGjx6_6122
Donkey Scaffolds
0
5000000
10000000
15000000
20000000
25000000
30000000
35000000
40000000
chr26
AA)
chr27
ScCGjx6_392
ScCGjx6_153
ScCGjx6_380
ScCGjx6_1
Donkey Scaffolds
0
5000000
10000000
15000000
20000000
25000000
30000000
35000000
40000000
chr27
AB)
chr28
ScCGjx6_397
ScCGjx6_244
ScCGjx6_97
Donkey Scaffolds
0
5000000
10000000
15000000
20000000
25000000
30000000
35000000
40000000
45000000
chr28
AC)
chr29
ScCGjx6_2
ScCGjx6_203
ScCGjx6_90
Donkey Scaffolds
0
5000000
10000000
15000000
20000000
25000000
30000000
chr29
AD)
chr30
ScCGjx6_2393
ScCGjx6_810
ScCGjx6_95
ScCGjx6_616
Donkey Scaffolds
0
5000000
10000000
15000000
20000000
25000000
3000000
chr30
AE)
chr31
ScCGjx6_:63
ScCGjx6_414
Donkey Scaffolds
0
5000000
10000000
15000000
20000000
2500000
chr31

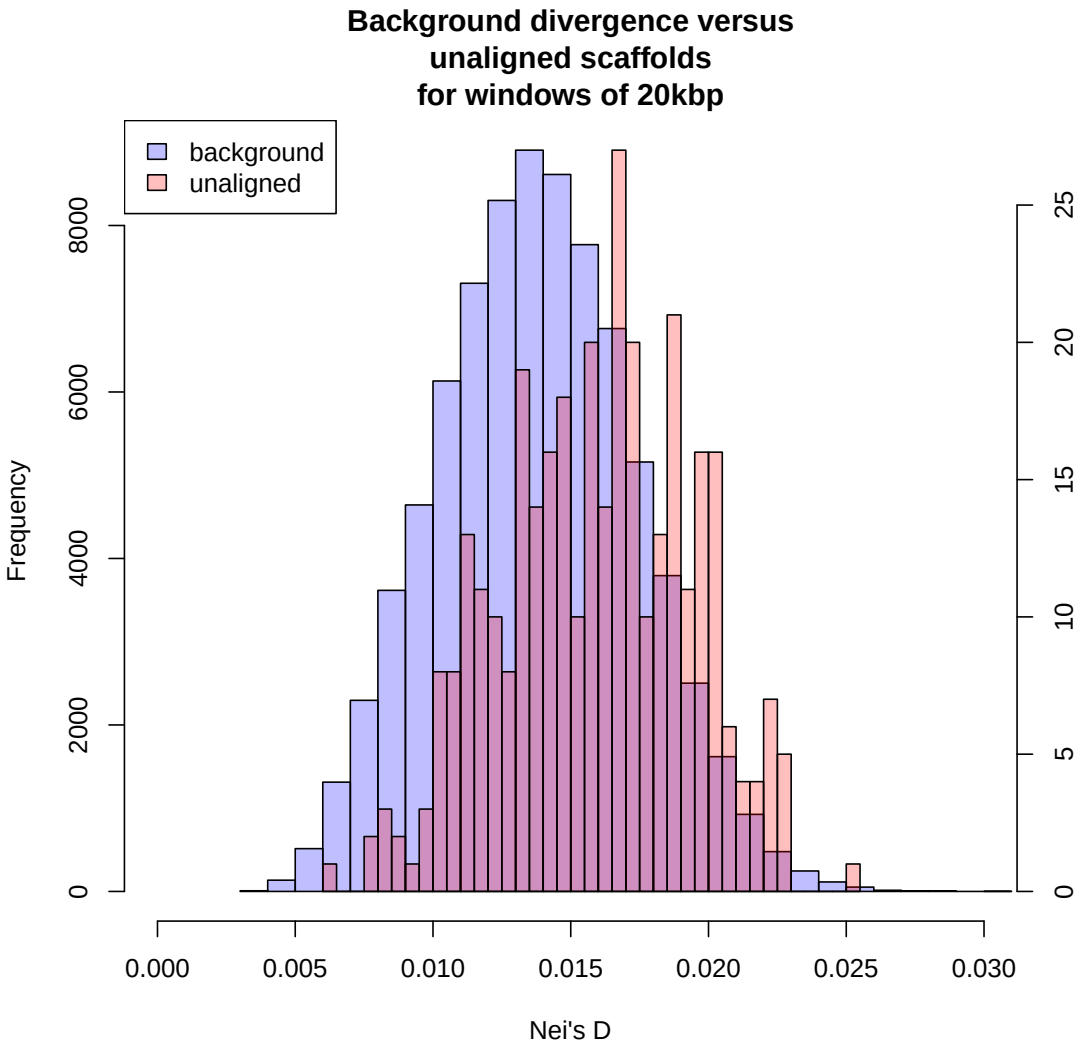


fig. S5. Genetic distance between scaffolds spanning the gap on ECA12 versus the background.

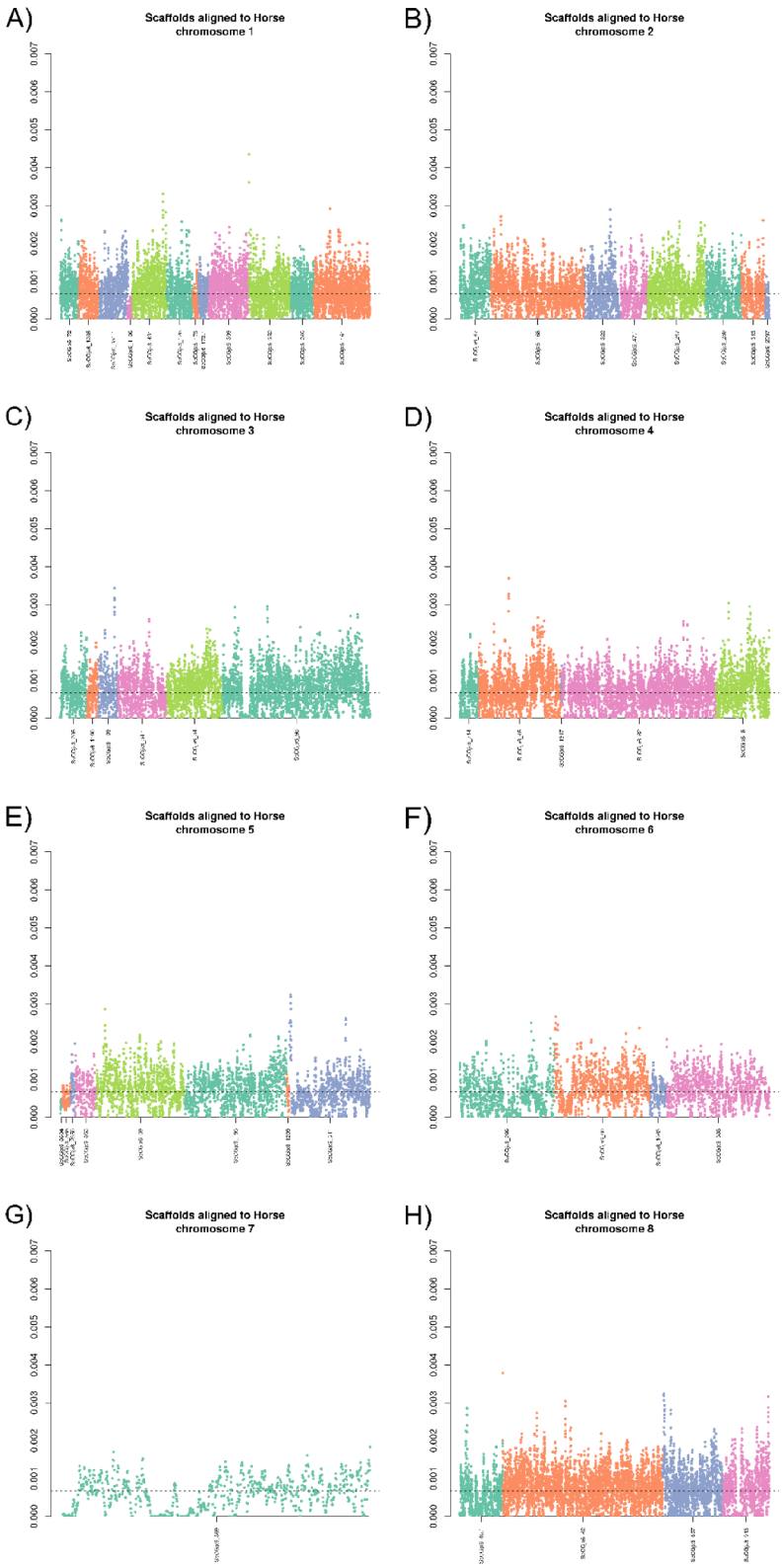


fig. S6. Measured heterozygosity rates for the donkey scaffolds aligned to the various horse chromosomes.

Fig. S6 2 of 4

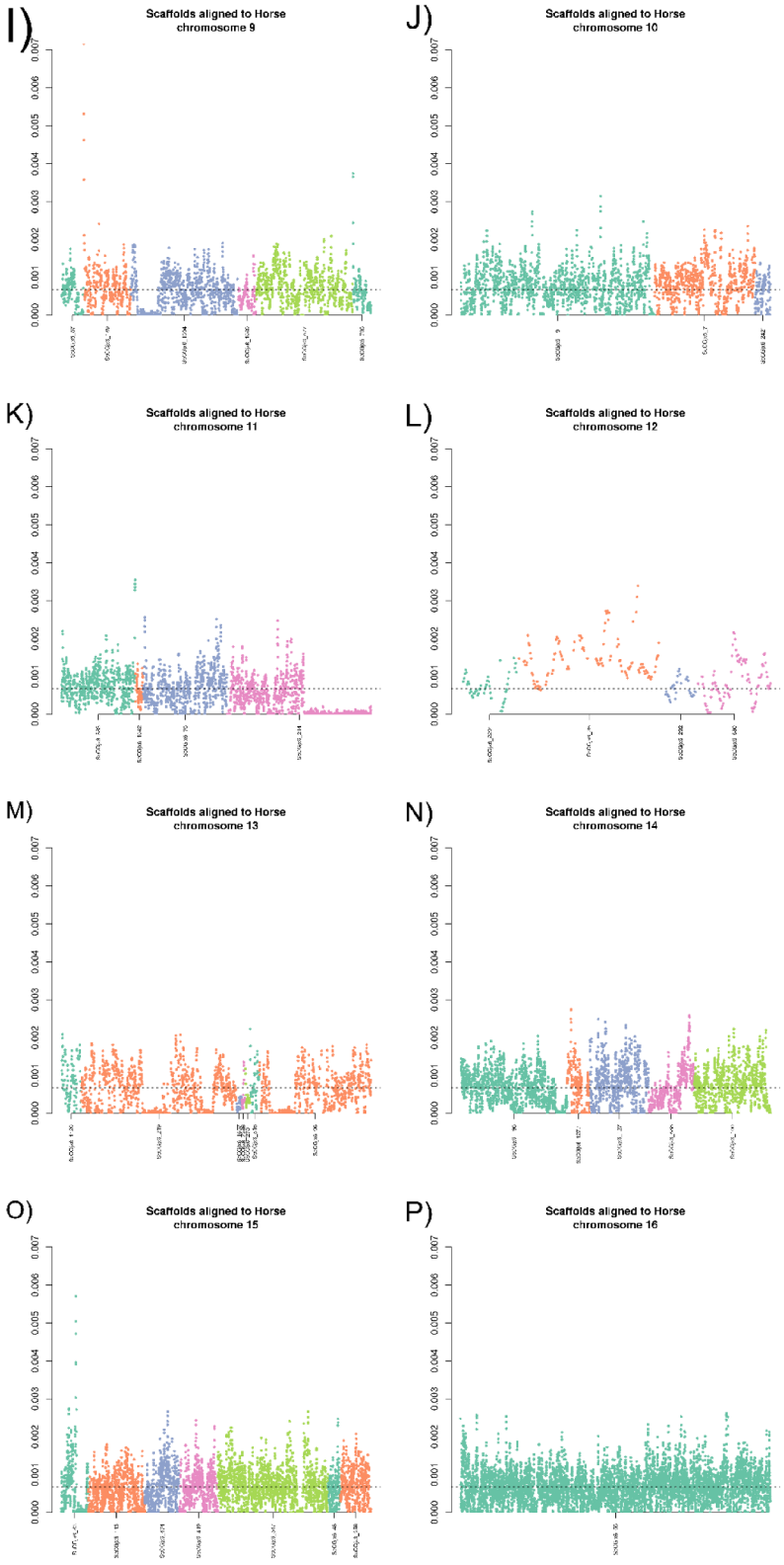

I)
Scaffolds aligned to Horse
chromosome 9
J)
Scaffolds aligned to Horse
chromosome 10
K)
Scaffolds aligned to Horse
chromosome 11
L)
Scaffolds aligned to Horse
chromosome 12
M)
Scaffolds aligned to Horse
chromosome 13
N)
Scaffolds aligned to Horse
chromosome 14
O)
Scaffolds aligned to Horse
chromosome 15
P)
Scaffolds aligned to Horse
chromosome 16

Fig. S6 3 of 4

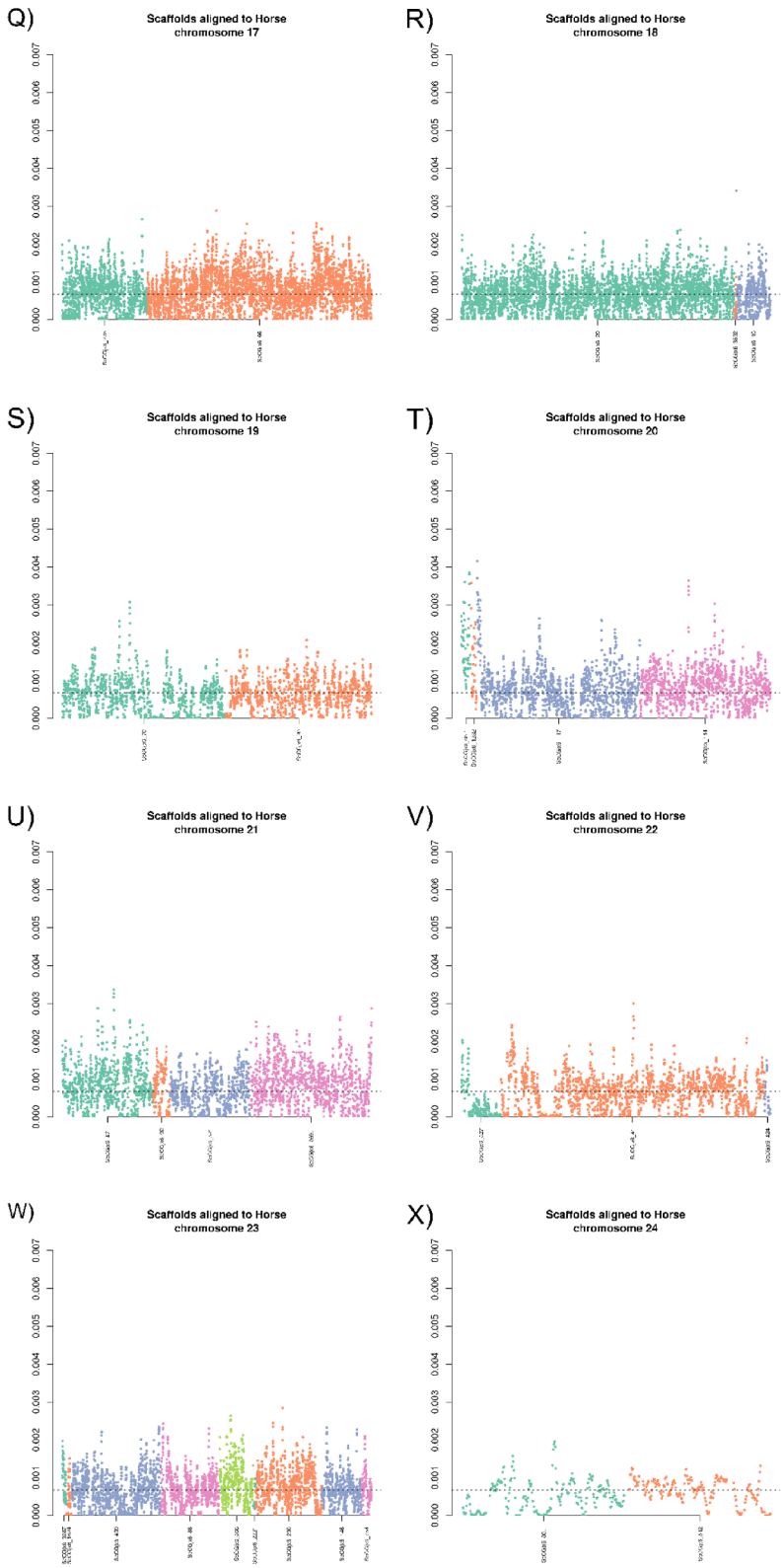

Q)
Scaffolds aligned to Horse
chromosome 17
R)
Scaffolds aligned to Horse
chromosome 18
S)
Scaffolds aligned to Horse
chromosome 19
T)
Scaffolds aligned to Horse
chromosome 20
U)
Scaffolds aligned to Horse
chromosome 21
V)
Scaffolds aligned to Horse
chromosome 22
W)
Scaffolds aligned to Horse
chromosome 23
X)
Scaffolds aligned to Horse
chromosome 24

Fig. S6 4 of 4

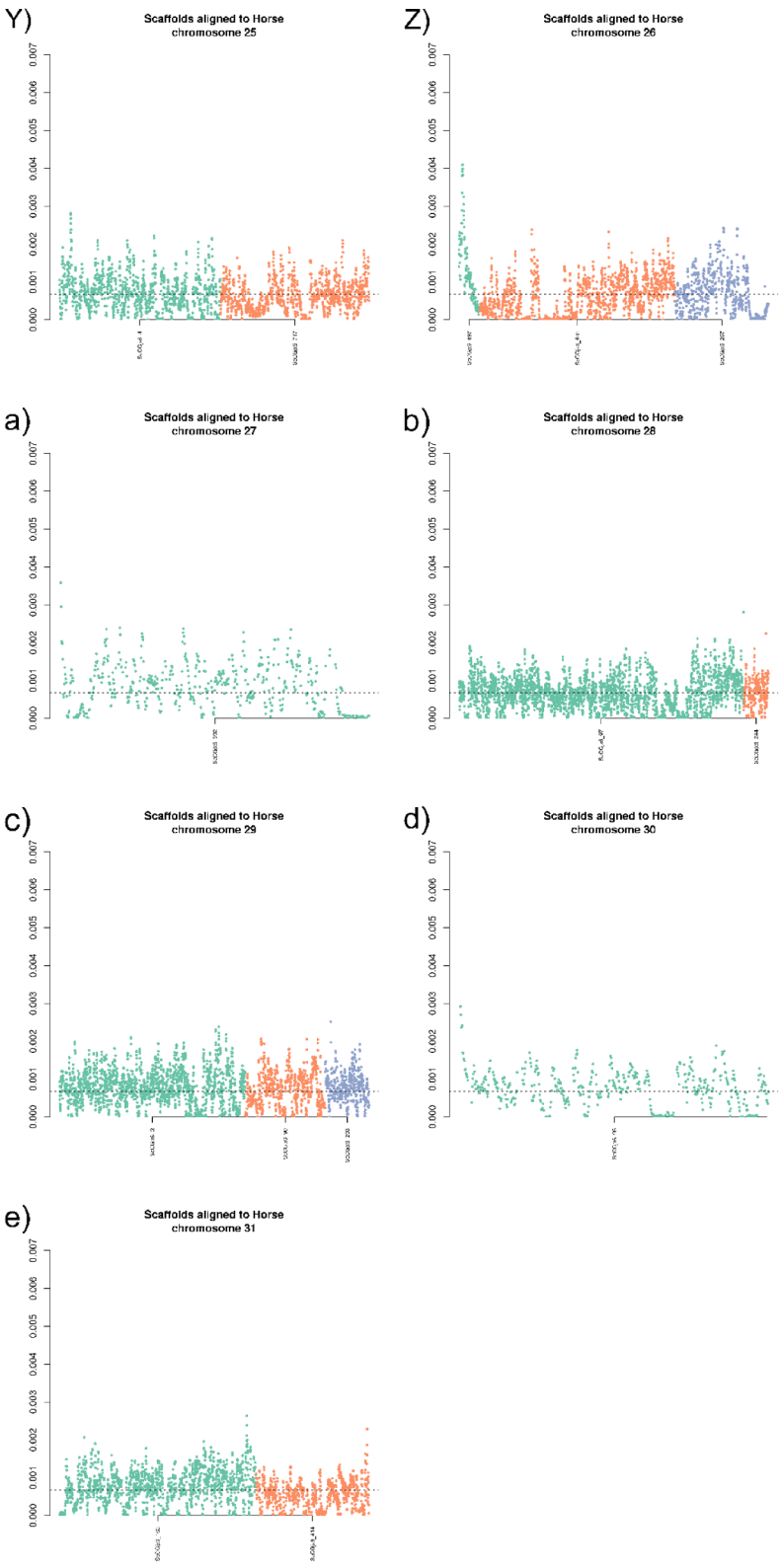

Y)
Scaffolds aligned to Horse
chromosome 25
Z)
Scaffolds aligned to Horse
chromosome 26
a)
Scaffolds aligned to Horse
chromosome 27
b)
Scaffolds aligned to Horse
chromosome 28
c)
Scaffolds aligned to Horse
chromosome 29
d)
Scaffolds aligned to Horse
chromosome 30
e)
Scaffolds aligned to Horse
chromosome 31

Fig. S7

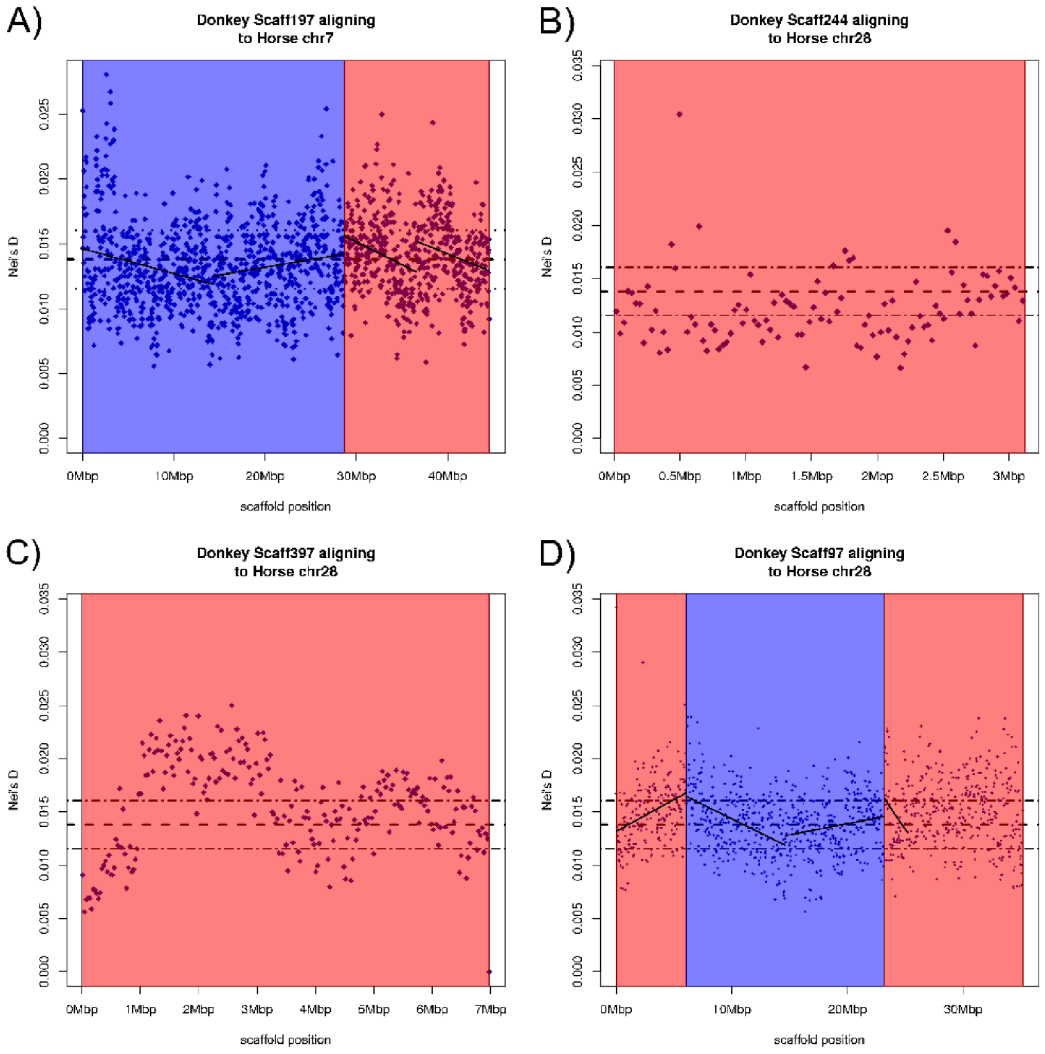


fig. S7. Nei's genetic distance by windows of 30 kb between donkey and horse chromosomes for scaffolds with signs of inversions. The areas in blue are potentially on different strands and in red, on the same strand.

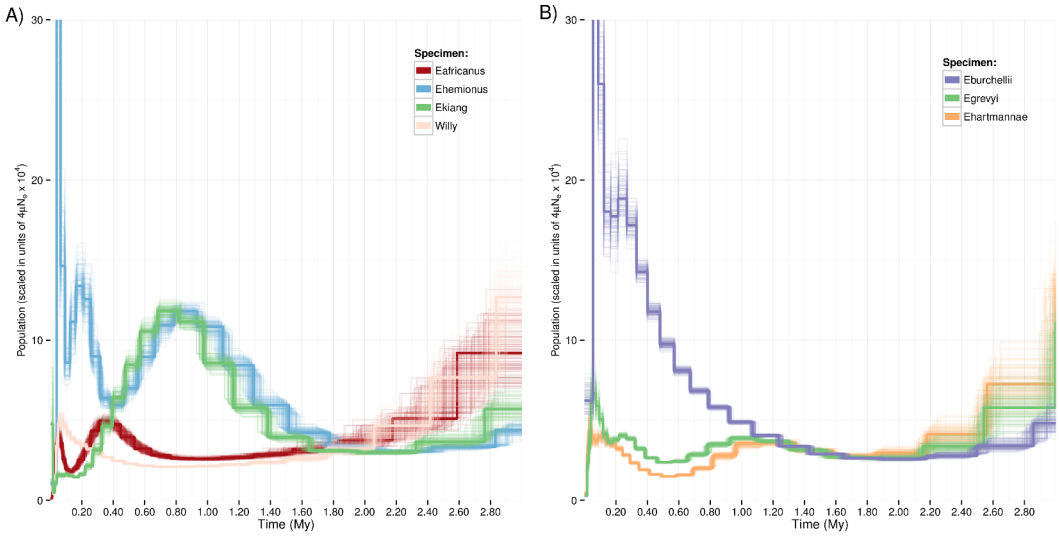


fig. S8. Effective population size over time by aligning to the horse reference. PSMC reconstruction of the effective population size over time using the data from (*9*) which had been aligned to the horse genome, for different *ass* species (**A**) and zebra species (**B**). For both, the effective population over time are estimated to be lower when the new donkey reference is used which is likely due to the greater phylogenetic proximity of this new reference.

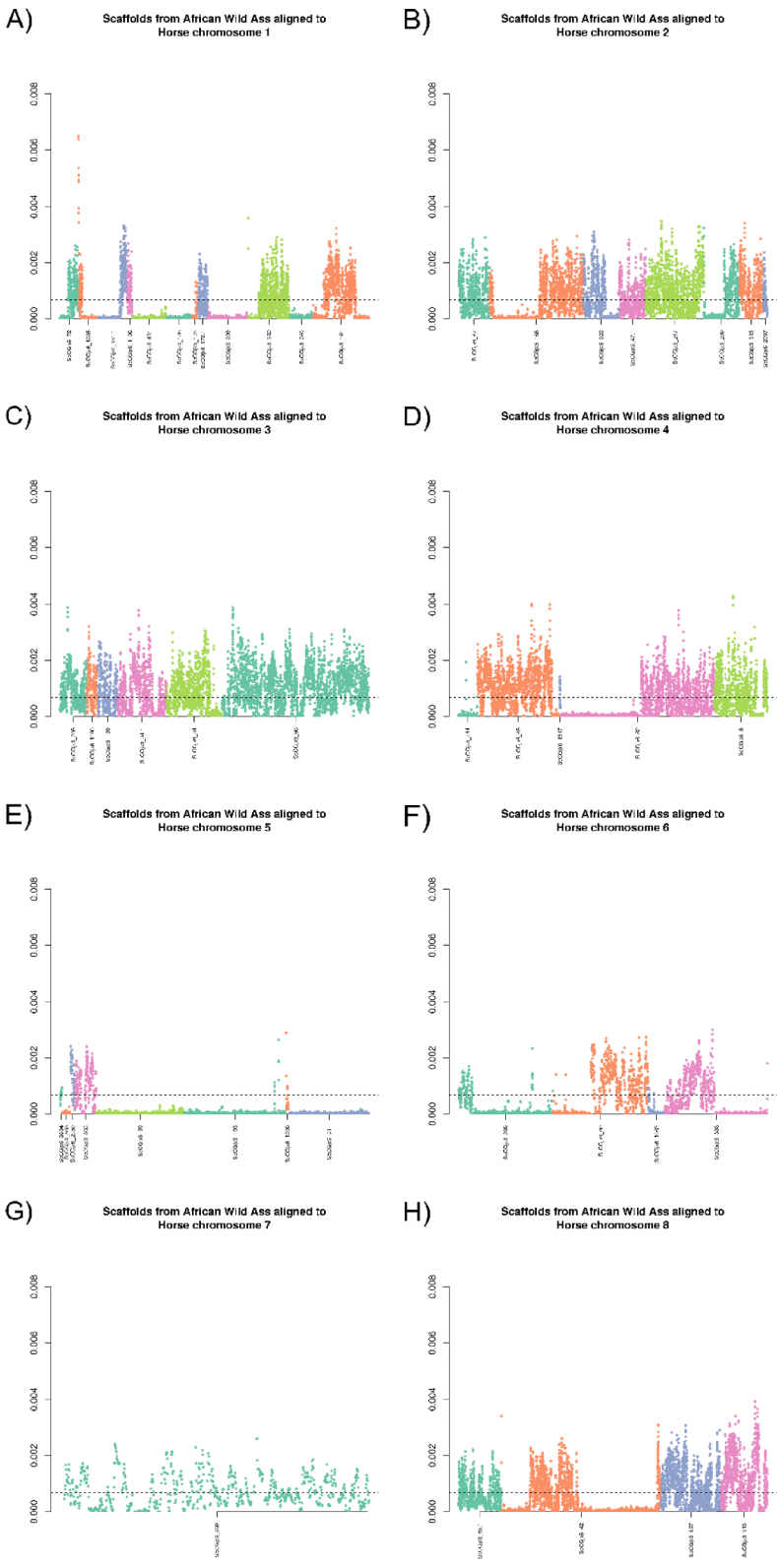


fig. S9. Measured heterozygosity rates for the African wild ass using the donkey scaffolds aligned to the horse chromosomes.

Fig. S9 2 of 4

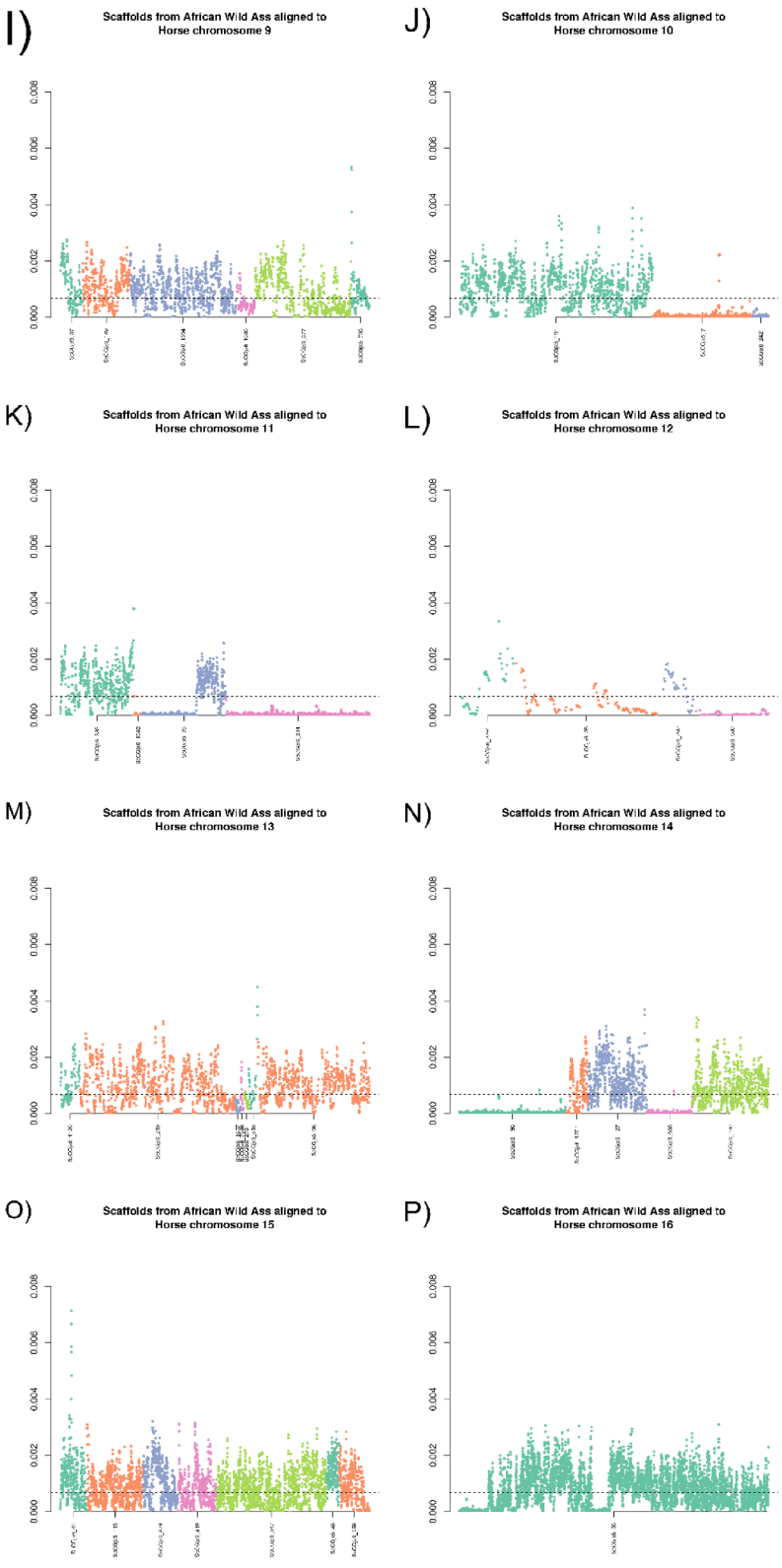

I)
Scaffolds from African Wild Ass aligned to Horse chromosome 9
J)
Scaffolds from African Wild Ass aligned to Horse chromosome 10
K)
Scaffolds from African Wild Ass aligned to Horse chromosome 11
L)
Scaffolds from African Wild Ass aligned to Horse chromosome 12
M)
Scaffolds from African Wild Ass aligned to Horse chromosome 13
N)
Scaffolds from African Wild Ass aligned to Horse chromosome 14
O)
Scaffolds from African Wild Ass aligned to Horse chromosome 15
P)
Scaffolds from African Wild Ass aligned to Horse chromosome 16
0.000
0.002
0.004
0.006
0.008

Fig. S9 3 of 4

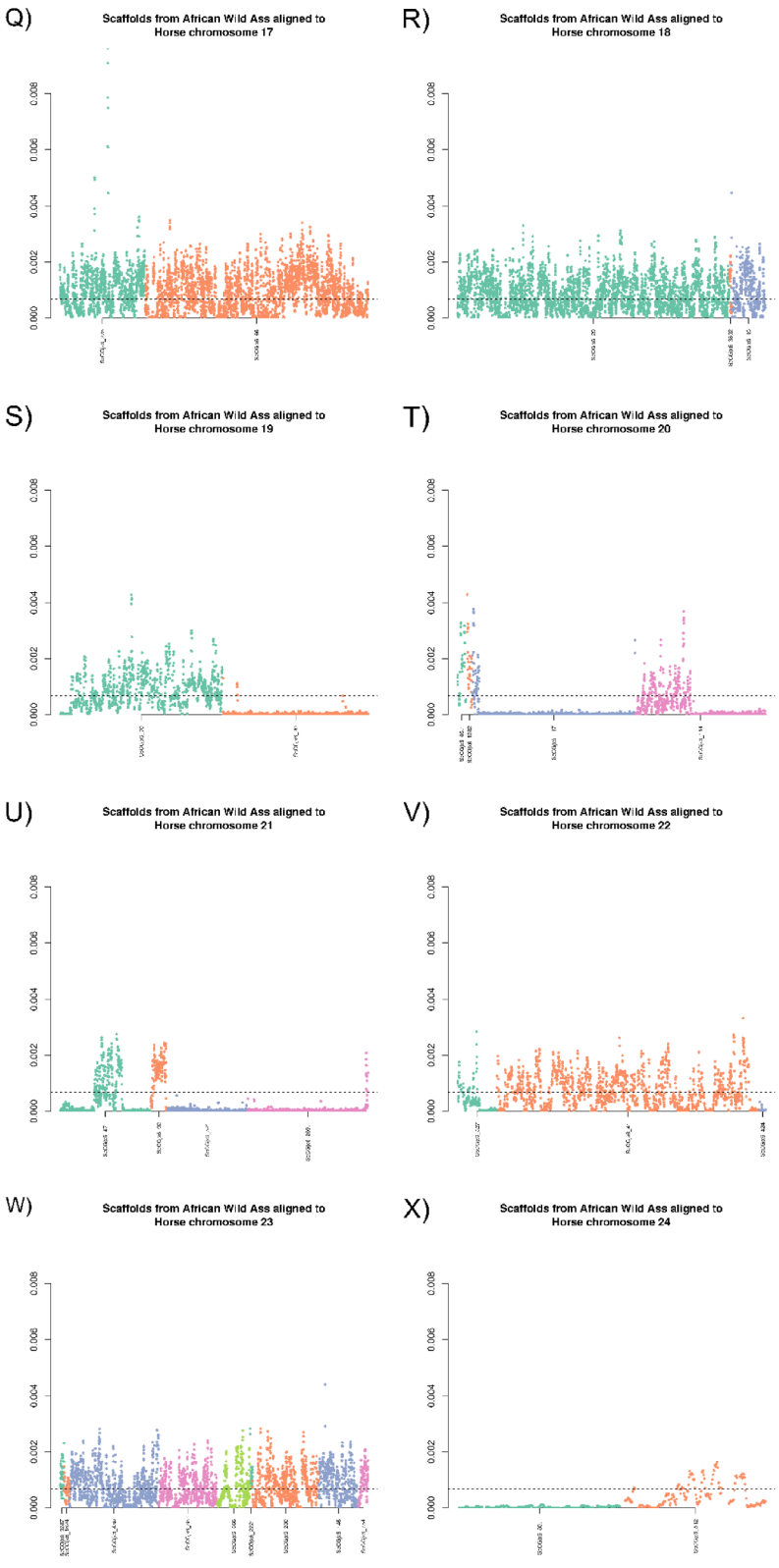

Q)
Scaffolds from African Wild Ass aligned to Horse chromosome 17
R)
Scaffolds from African Wild Ass aligned to Horse chromosome 18
S)
Scaffolds from African Wild Ass aligned to Horse chromosome 19
T)
Scaffolds from African Wild Ass aligned to Horse chromosome 20
U)
Scaffolds from African Wild Ass aligned to Horse chromosome 21
V)
Scaffolds from African Wild Ass aligned to Horse chromosome 22
W)
Scaffolds from African Wild Ass aligned to Horse chromosome 23
X)
Scaffolds from African Wild Ass aligned to Horse chromosome 24

Fig. S9 4 of 4

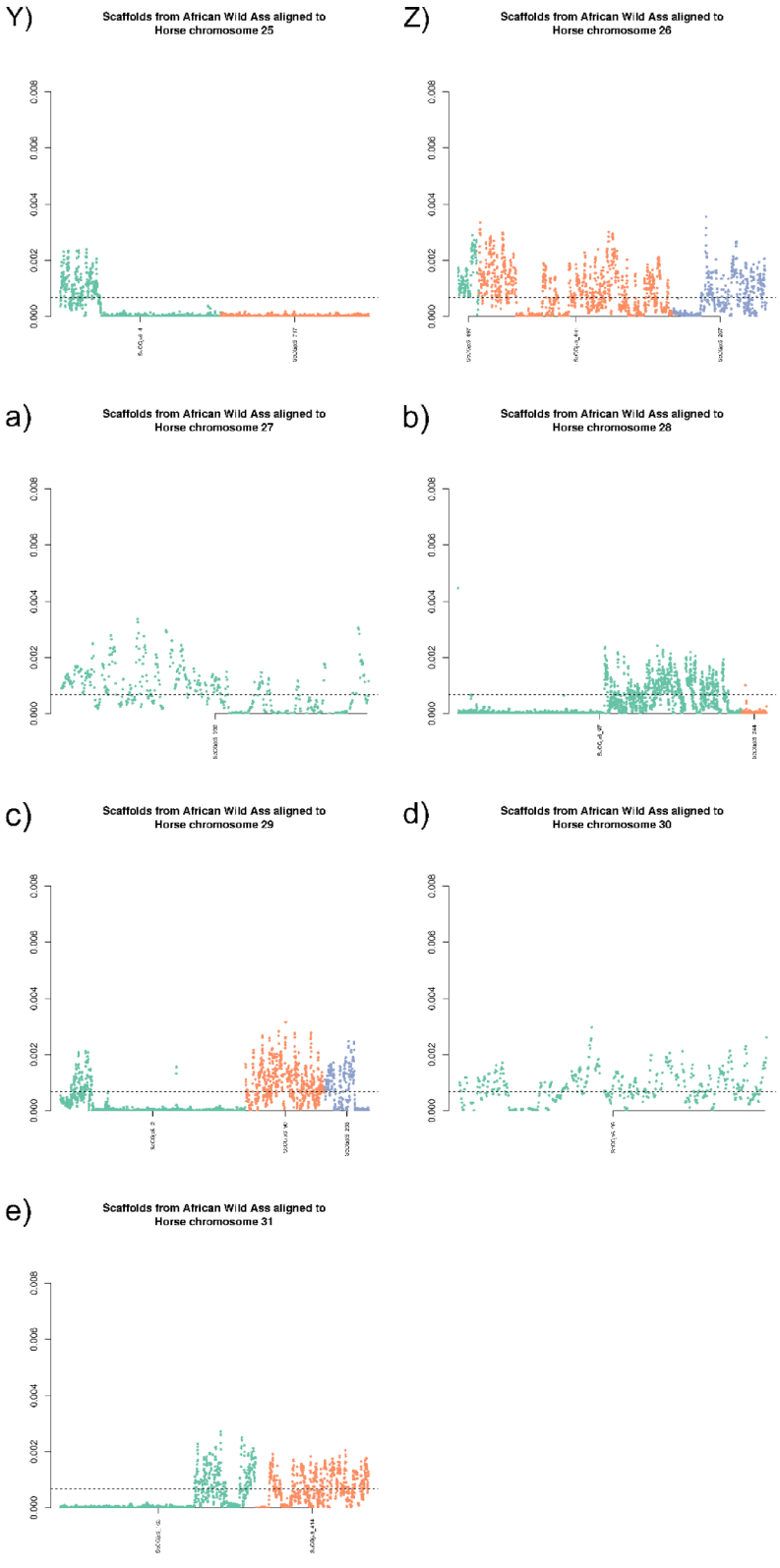

Y)
Scaffolds from African Wild Ass aligned to Horse chromosome 25
Z)
Scaffolds from African Wild Ass aligned to Horse chromosome 26
a)
Scaffolds from African Wild Ass aligned to Horse chromosome 27
b)
Scaffolds from African Wild Ass aligned to Horse chromosome 28
c)
Scaffolds from African Wild Ass aligned to Horse chromosome 29
d)
Scaffolds from African Wild Ass aligned to Horse chromosome 30
e)
Scaffolds from African Wild Ass aligned to Horse chromosome 31
0.000
0.002
0.004
0.006
0.008

table S1. Translocations found between the donkey and horse scaffolds. The first three represent translocations for donkey scaffolds aligning to the same horse chromosome. The last six translocations are for donkey scaffolds that align to two different horse chromosomes.

	Donkey scaffolds	Horse coordinates	Description	Figure
Translocation 1	ScCGjx6_113:1,497-10,261,736	chr15:29,315,938-39,718,792	the last ~742 kb of ScCGjx6_113 map 8Mb away from the rest of the aligned scaffold, ScCGjx6_419 is found in between	fig. S4 O)
	ScCGjx6_419:1,3994-7,541,780	chr15:39,767,787-47,354,651		
	ScCGjx6_113:10,278,753-11,020,782	chr15:47,359,895-48,089,680		
Translocation 2	ScCGjx6_161:13,460-4,479,819	chr1:114,297,765-119,020,414	The alignment of ScCGjx6_161 is interrupted in the middle by an alignment of 9,7Mb to ScCGjx6_240, the unaligned portion of ScCGjx6_161 is found on ECA2	fig. S4 A)
	ScCGjx6_240:10,024-9,454,019	chr1:119,040,857-128,716,420		
	ScCGjx6_161:17,705,081-23,104,118	chr1:128,737,306-134,312,170		
Translocation 3	ScCGjx6_77:20,221-12,409,593	chr22:24,084,620-36,522,341	An alignment of 245 kb to ScCGjx6_1370 is found in the middle of scaffold ScCGjx6_77 (25Mb)	fig. S4 V)
	ScCGjx6_1370:4395-253254	chr22:36,548,355-36,794,213		
	ScCGjx6_77:12,414,766-25,397,100	chr22:36,851,421-49,910,038		
Translocation 4	ScCGjx6_168:15,609,437-9,407	chr14:36,550,698-52,230,055	ScCGjx6_168 spans the middle of ECA2 and 14	fig. S3 B)
	ScCGjx6_168:15,636,073-25,881,750	chr2:61,288,241-75,810,515		
Translocation 5	ScCGjx6_161:13,460-4,479,819	chr1:114,297,765-119,020,414	ScCGjx6_161 starts on ECA1,	fig. S3 A)

	ScCGjx6_161:4,498,007-17,696,940	chr2:75,816,382-89,056,597	moves to ECA2 and moves back to ECA1	
	ScCGjx6_161:17,705,081-23,122,153	chr1:128,737,306-134,312,170		
Translocation 6	ScCGjx6_69: 24,084,324-3,843	chr19:35,686,863-59,971,246	ScCGjx6_69 joins the middle of ECA3 and the end of ECA19	fig. S3 C)
	ScCGjx6_69:24,089,115-30,125,315	chr3:65,994,971-72,154,088		
Translocation 7	ScCGjx6_42:20,842,070-2,313	chr8:67,985,150-88,979,166	ScCGjx6_42 joins the middle of ECA10 and chromosome 8 (towards the end)	fig. S3 D)
	ScCGjx6_42:20,842,070-51,953,946	chr10:33,700,746-65,205,297		
Translocation 8	ScCGjx6_43:2,190,086-2,233	chr19:29,071-2,259,915	ScCGjx6_43 spans the beginning of ECA15 and ECA19	fig. S3 E)
	ScCGjx6_43:2,190,086-5,271,059	chr15:203,277-3,453,145		
Translocation 9	ScCGjx6_153:9,695-15,088,019	chr17:65,701,400-80,747,902	ScCGjx6_153 joins the end of ECA17 and the middle of ECA27	fig. S3 F)
	ScCGjx6_153:15,088,019-18,932,489	chr27:29,582,427-33,520,420		

table S2. Gene ontologies of biological processes and enriched Reactome pathways associated with genes found in donkey scaffolds with signs of inversions when compared to the horse genome.

Gene ontology	P-value	FDR	Genes
Organonitrogen compound catabolic process	7.00E-05	5.33E-02	PDE10A;KERA;AMDHD1;ALDH1L2;DCN;HAL;LTA4H;LUM;PAH;STAB2;TDG
Regulation of DNA metabolic process	3.32E-04	1.26E-01	APAF1;IGF1;MAP3K4;KITLG;PARPBP;PPP2R1A;RFC5;TCP1;UBE2N;CDK1
Aromatic compound catabolic process	8.22E-04	1.92E-01	PDE10A;AMDHD1;PNLDC1;ALDH1L2;HAL;APAF1;PAH;PPP2R1A;RPL11;TDG
Organic cyclic compound catabolic process	1.01E-03	1.92E-01	PDE10A;AMDHD1;PNLDC1;ALDH1L2;HAL;APAF1;PAH;PPP2R1A;RPL11;TDG
Carbohydrate derivative catabolic process	1.64E-03	2.50E-01	PDE10A;KERA;DCN;LUM;STAB2;TDG
Heterocycle catabolic process	2.13E-03	2.64E-01	PDE10A;AMDHD1;PNLDC1;ALDH1L2;HAL;APAF1;PPP2R1A;RPL11;TDG
Cellular nitrogen compound catabolic process	2.42E-03	2.64E-01	PDE10A;AMDHD1;PNLDC1;ALDH1L2;HAL;APAF1;PPP2R1A;RPL11;TDG
Regulation of embryonic development	5.04E-03	4.80E-01	AMOT;DUSP6;IGF1;CDK1
Regulation of protein stability	6.10E-03	5.16E-01	GAPDH;IGF1;USP28;TCP1;USP4;USP2
Dicarboxylic acid metabolic process	7.49E-03	5.70E-01	AMDHD1;ALDH1L2;HAL;NR1H4
Pathway	**P-value**	**FDR**	**Genes**
Cilium Assembly	2.43E-04	1.06E-01	FGFR1OP;NEDD1;CEP83;PPP2R1A;TCP1;TCTE3;CEP290;CDK1
Anchoring of the basal body to the plasma membrane	3.34E-04	1.06E-01	FGFR1OP;NEDD1;CEP83;PPP2R1A;CEP290;CDK1
Loss of Nlp from mitotic centrosomes	4.94E-04	1.06E-01	FGFR1OP;NEDD1;PPP2R1A;CEP290;CDK1

Loss of proteins required for interphase microtubule organization from the centrosome	4.94E-04	1.06E-01	FGFR1OP;NEDD1;PPP2R1A;CEP290;CDK1
Organelle biogenesis and maintenance	5.48E-04	1.06E-01	FGFR1OP;NEDD1;MRPL42;MRPL18;CEP83;PPP2R1A;TCP1;TCTE3;CEP290;CDK1
AURKA Activation by TPX2	6.12E-04	1.06E-01	FGFR1OP;NEDD1;PPP2R1A;CEP290;CDK1
Recruitment of mitotic centrosome proteins and complexes	6.79E-04	1.06E-01	FGFR1OP;NEDD1;PPP2R1A;CEP290;CDK1
Centrosome maturation	6.79E-04	1.06E-01	FGFR1OP;NEDD1;PPP2R1A;CEP290;CDK1
Regulation of PLK1 Activity at G2/M Transition	1.09E-03	1.53E-01	FGFR1OP;NEDD1;PPP2R1A;CEP290;CDK1
Diseases associated with glycosaminoglycan metabolism	1.46E-03	1.84E-01	KERA;DCN;LUM

table S3. Human phenotypes, human diseases, and pathways associated with genes enriched in detected ROHs.

Human phenotype	P-value	FDR	Genes
ROHs >1Mb			
Deep palmar crease	2.66E-04	9.85E-01	ASXL1;PAFAH1B1;YWHAE
Recurrent aspiration pneumonia	1.79E-03	1.00E+00	PAFAH1B1;YWHAE
Thick upper lip vermilion	2.66E-03	1.00E+00	PAFAH1B1;YWHAE
Aspiration pneumonia	2.66E-03	1.00E+00	PAFAH1B1;YWHAE
ROH 500 kb-1Mb			
Polydipsia	6.55E-04	8.98E-01	KCNJ1;AVP;CEP290
Abnormal drinking behavior	6.55E-04	8.98E-01	KCNJ1;AVP;CEP290
Hypokalemia	8.74E-04	8.98E-01	KCNJ1;KCNJ5;AVP
Polyuria	1.00E-03	8.98E-01	KCNJ1;AVP;CEP290
ROHs 100 kb-500 kb			
Stomatocytosis	6.41E-04	1.00E+00	RHAG;ABCG5;ABCG8
Biliary tract abnormality	2.54E-03	1.00E+00	EHHADH;IFT172;IL12RB1;DCTN4;PEX12;PHKG2;PIK3CA;PKHD1;PTPN3;PEX2;BBS10;CLDN1
CNS demyelination	3.02E-03	1.00E+00	LRPPRC;PEX12;FOXRED1;NDUFA12;PEX2;SDHA;EIF2B4;EIF2B5
Abnormality of the biliary system	4.12E-03	1.00E+00	EHHADH;ANKS6;IFT172;IL12RB1;MPV17;DCTN4;PEX12;PHKG2;PIK3CA;PKHD1;PTPN3;PEX2;RHAG;ABCG8;BBS10;CLDN1
Human disease	**P-value**	**FDR**	**Genes**
ROHs >1Mb			
Muscle Weakness	3.82E-04	1.62E-01	HUNK;MIS18A;SCAF4;SOD1;TIAM1;URB1
Down Syndrome	5.21E-04	1.62E-01	HUNK;MIS18A;SCAF4;SOD1;TIAM1;URB1
Epithelial ovarian cancer	7.29E-04	1.62E-01	TIPARP;ATAD5;CHMP4C
Turner Syndrome	1.97E-03	2.43E-01	NOS2;SOD1

ROH 500 kb-1Mb			
Down Syndrome	0.00E+00	0.00E+00	USP16;CCT8;CYYR1;GART;N6AMT1;DONSON;IFNAR1;IFNGR2;MAP3K7CL;BACH1;ITSN1;SON;TMEM50B;ADAMTS1;CRYZL1
Muscle Weakness	0.00E+00	0.00E+00	USP16;CCT8;CYYR1;GART;N6AMT1;DONSON;IFNAR1;IFNGR2;MAP3K7CL;BACH1;ITSN1;SON;TMEM50B;ADAMTS1;CRYZL1
Small cell carcinoma of lung	1.43E-02	1.00E+00	AVP;SOX2
Myocardial Ischemia	1.69E-02	1.00E+00	APLP2;JAK2;KITLG;VEGFA
ROHs 100 kb-500kb			
Muscle Weakness	2.75E-06	1.71E-03	MORC3;CLDN14;GABPA;BRWD1;MRPL39;JAM2;SH3BGR;SIM2;CHAF1B;PSMG1;SYNJ1;PAXBP1
Down Syndrome	5.13E-06	1.71E-03	MORC3;CLDN14;GABPA;BRWD1;MRPL39;JAM2;SH3BGR;SIM2;CHAF1B;PSMG1;SYNJ1;PAXBP1
Neuroectodermal Tumor, Primitive	1.29E-03	2.87E-01	EWSR1;HEY1;HES1
Weight Gain	3.96E-03	5.96E-01	MAPRE1;PRKD3;IPO11;PIK3CA;SPARC;ADIPOQ;FEZ2
Pathways	**P-value**	**FDR**	**Genes**
ROHs >1Mb			
Hormone-sensitive lipase (HSL)-mediated triacylglycerol hydrolysis	3.42E-05	4.80E-02	FABP4;FABP5;FABP9;FABP12
Antimicrobial peptides	5.70E-03	1.00E+00	BPIFB6;BPIFB4;BPIFB2
HSF1 activation	7.40E-03	1.00E+00	RPA1;YWHAE
AURKA Activation by TPX2	1.27E-02	1.00E+00	MAPRE1;TPX2;PAFAH1B1;YWHAE
ROH 500 kb-1Mb			
Inwardly rectifying K+ channels	4.63E-03	1.00E+00	KCNJ1;KCNJ5
Regulation of IFNG signaling	5.44E-03	1.00E+00	IFNGR2;JAK2
Platelet degranulation	6.92E-03	1.00E+00	APLP2;GTPBP2;CDC37L1;VEGFA

Response to elevated platelet cytosolic Ca2+	7.20E-03	1.00E+00	APLP2;GTPBP2;CDC37L1;VEGFA
ROHs 100 kb-500kb			
Hyaluronan metabolism	4.99E-03	1.00E+00	ABCC5;GUSB;STAB2
O-linked glycosylation of mucins	6.84E-03	1.00E+00	MUC20;GALNT10;WBSCR17;ST6GAL1;GALNT12
Metabolism of carbohydrates	8.55E-03	1.00E+00	ABCC5;KERA;GAPDH;GLB1;GUSB;NDST1;SLC35B2;PHKG1;PHKG2;STAB2;XYLT2;SLC2A2;SLC5A2;NUP37
RNA Polymerase III Transcription Initiation From Type 2 Promoter	9.38E-03	1.00E+00	CRCP;GTF3C2;POLR2H;POLR3B

table S4. Horse sequences used for the detection of donkey scaffolds pertaining to the Y chromosome.

NCBI nucleotide ID	GenBank accession	Length (bp)
406356568	JX647038.1	27711
406356560	JX647030.1	34694
406356544	JX647022.1	11087
406356536	JX647014.1	10486
406356528	JX647006.1	5528
406356520	JX646998.1	7619
406356512	JX646990.1	8830
406356504	JX646982.1	9393
406356496	JX646974.1	8508
406356488	JX646966.1	10880
406356480	JX646958.1	17810
406356472	JX646950.1	18678
406356464	JX646942.1	14899
20373117	G72338.1	400
20373114	G72335.1	528
20373118	G72339.1	255
29126040	AB091794.1	5591
42525419	AY532879.1	452
20373115	G72336.1	508

table S5. Heterozygosity rates for various species of asses and zebras computed when aligning to the donkey reference described in this study and recomputed on the basis of the data reported by Jónsson *et al.* (*9*), which were aligned to the horse reference.

Species name	Binomial nomenclature	Depth-of-coverage (reads)	Heterozygosity rate (aligned to donkey)	Heterozygosity rate (aligned to horse)
Somali wild ass	*Equus africanus somaliensis*	26.9	0.05747±0.00265%	0.081221%
Grévy's zebra	*Equus grevyi*	20.8	0.09410±0.00266%	0.113645%
Onager	*Equus hemionus*	22.0	0.18138±0.00266%	0.210421%
Kiang	*Equus kiang*	12.1	0.10419±0.00266%	0.128137%
Burchell's zebra	*Equus quagga burchellii*	26.4	0.21491±0.00266%	0.265465%
Hartmann's mountain zebra	*Equus zebra hartmannae*	21.5	0.08159±0.00266%	0.0999795%
Domestic donkey	*Equus africanus asinus*	61.2	0.06814±0.00264%	0.113711%

table S6. Listing missing proteins in complete and partially complete Eukaryotic Orthologous Groups from the Core Eukaryotic Genes Mapping Approach.

Complete KOGs							
KOG0018	KOG0025	KOG0176	KOG0179	KOG0182	KOG0184	KOG0188	KOG0209
KOG0261	KOG0276	KOG0291	KOG0292	KOG0357	KOG0361	KOG0364	KOG0365
KOG0376	KOG0400	KOG0424	KOG0434	KOG0462	KOG0477	KOG0481	KOG0556
KOG0559	KOG0741	KOG0780	KOG0862	KOG0871	KOG0964	KOG0969	KOG0985
KOG0991	KOG1058	KOG1099	KOG1112	KOG1123	KOG1137	KOG1145	KOG1185
KOG1211	KOG1241	KOG1299	KOG1335	KOG1349	KOG1355	KOG1358	KOG1373
KOG1458	KOG1463	KOG1498	KOG1532	KOG1540	KOG1549	KOG1555	KOG1647
KOG1746	KOG1795	KOG1816	KOG1872	KOG1889	KOG1942	KOG2004	KOG2036
KOG2044	KOG2303	KOG2311	KOG2415	KOG2446	KOG2451	KOG2472	KOG2481
KOG2531	KOG2535	KOG2537	KOG2572	KOG2575	KOG2613	KOG2623	KOG2680
KOG2719	KOG2775	KOG2807	KOG2909	KOG2916	KOG2930	KOG2967	KOG3013
KOG3049	KOG3157	KOG3174	KOG3180	KOG3189	KOG3239	KOG3297	KOG3313
KOG3404	KOG3855	KOG3954					
Partial complete KOGs							
KOG0365	KOG0741	KOG1358	KOG1816	KOG1872	KOG1889	KOG1942	KOG2451
KOG2472	KOG2531	KOG2575	KOG2613	KOG2719	KOG3954		

table S7. Repeat elements and low-complexity DNA sequences masked in the donkey genome using RepeatMasker. The table is showing the number of elements for the different types of SINEs, LINEs, LTR elements, DNA elements and small rNA and satellites masked in the donkey genome. Furthermore, the total length in bp and the percentage of masked sequence for each category is listed.

bases masked: 920,991,241 bp (39.68 %)				
		Number of elements*	Length occupied	Percentage of sequence
SINEs:		**981,574**	**167,801,528 bp**	**7.23 %**
	Alu/B1	9	707 bp	0.00 %
	MIRs	589,657	82,854,837 bp	3.57 %
LINEs:		**1,004,122**	**445,910,954 bp**	**19.21 %**
	LINE1	518,939	315,463,444 bp	13.59 %
	LINE2	417,636	113,787,977 bp	4.90 %
	L3/CR1	53,387	12,250,891 bp	0.53 %
	RTE	13,118	4,224,920 bp	0.18 %
LTR elements:		**428,117**	**171,420,149 bp**	**7.39 %**
	ERVL	108,139	52,444,566 bp	2.26 %
	ERVL-MaLRs	175,510	65,261,362 bp	2.81 %
	ERV_classI	79,265	41,584,018 bp	1.79 %
	ERV_classII	33,143	2,942,704 bp	0.13 %
DNA elements:		**417,275**	**91,394,731 bp**	**3.94 %**
	hAT-Charlie	229,200	46,401,915 bp	2.00 %
	TcMar-Tigger	73,698	21,448,450 bp	0.92 %
Unclassified:		**7,850**	**1,541,532 bp**	**0.07 %**
Total interspersed repeats:			**878,068,894 bp**	**37.83 %**
Small RNA:		372,003	80,855,800 bp	3.48 %

Satellites:		84,842	42,804,139 bp	1.84 %
Simple repeats:		0	0	0.00 %
Low complexity:		0	0	0.00 %

* most repeats fragmented by insertions or deletions have been counted as one element

table S8. Repeat elements and low-complexity DNA sequences masked in the donkey genome using the second of the RepeatMasker using the model generated from RepeatModeler as custom library input on the previously masked genome. The table is showing the number of elements for the different types of SINEs, LINEs, LTR elements, DNA elements and small RNA and satellites masked in the donkey genome. Furthermore the total length in bp and the percentage of masked sequence for each category is listed.

Bases masked: 7657033 bp (0.33 %)				
		Number of elements*	Length occupied	Percentage of sequence
SINEs:		**0**	**0 bp**	**0.00 %**
	Alu/B1	0	0 bp	0.00 %
	MIRs	0	0 bp	0.00 %
LINEs:		**6,264**	**2,169,302 bp**	**0.09 %**
	LINE1	4,010	1,025,059 bp	0.04 %
	LINE2	598	39,096 bp	0.00 %
	L3/CR1	0	0 bp	0.00 %
LTR elements:		**3,577**	**637,692 bp**	**0.03 %**
	ERVL	903	228,918 bp	0.01 %
	ERVL-MaLRs	1,936	265,014 bp	0.01 %
	ERV_classI	738	143,760 bp	0.01 %
	ERV_classII	0	0 bp	0.00 %
DNA elements:		**12,599**	**1,584,888 bp**	**0.07 %**
	hAT-Charlie	0	0 bp	0.00 %

	TcMar-Tigger	733	264,345 bp	0.01 %
Unclassified:		**5,591**	**1,761,449 bp**	**0.08 %**
Total interspersed repeats:			**6,153,331 bp**	**0.27 %**
Small RNA:		0	0 bp	0.00 %
Satellites:		616	425,075 bp	0.02 %
Simple repeats:		6,922	1,080,429 bp	0.05 %
Low complexity:		0	0	0.00 %

* most repeats fragmented by insertions or deletions have been counted as one element

table S9. Statistics of the completeness of the different versions of the donkey genome based on 248 Core Eukaryotic Genes. 'Complete (%)' refers to the predicted proteins that could be aligned to the HMMs of a KOG for a given protein family from the CEGMA dataset consisting of 248 CEGs. 'Partial complete (%)' refers to incomplete proteins. Complete genomes will also be included in the 'Partial' set.

Genome sequence	**Complete (%)**	**Partial complete (%)**
Current Donkey assembly repeatmasked	60.08	94.35
Current Donkey assembly non-repeatmasked	60.48	94.35
Huang et al. 2015	57.26	92.34
Orlando et al 2013	41.13	86.69