nature genetics

SUPPLEMENTARY INFORMATION

https://doi.org/10.1038/s41588-019-0405-z

In the format provided by the authors and unedited.

The genome sequence of segmental allotetraploid peanut *Arachis hypogaea*

David J. Bertioli[1,2,3,30*], Jerry Jenkins[4,30], Josh Clevenger[1,2,3,30], Olga Dudchenko[5], Dongying Gao[1], Guillermo Seijo[6,7], Soraya C. M. Leal-Bertioli[1,2,8], Longhui Ren[9], Andrew D. Farmer[10], Manish K. Pandey[11], Sergio S. Samoluk[6,7], Brian Abernathy[1], Gaurav Agarwal[8], Carolina Ballén-Taborda[2], Connor Cameron[10], Jacqueline Campbell[12], Carolina Chavarro[1,2], Annapurna Chitikineni[11], Ye Chu[13], Sudhansu Dash[10], Moaine El Baidouri[14,15], Baozhu Guo[16], Wei Huang[12], Kyung Do Kim[1,17], Walid Korani[1], Sophie Lanciano[15,18,19], Christopher G. Lui[5], Marie Mirouze[15,18,19], Márcio C. Moretzsohn[20], Melanie Pham[5], Jin Hee Shin[1,17], Kenta Shirasawa[21], Senjuti Sinharoy[22], Avinash Sreedasyam[4], Nathan T. Weeks[23], Xinyou Zhang[24,25], Zheng Zheng[24,25], Ziqi Sun[24,25], Lutz Froenicke[26], Erez L. Aiden[5], Richard Michelmore[26], Rajeev K. Varshney[11], C. Corley Holbrook[27], Ethalinda K. S. Cannon[12], Brian E. Scheffler[28], Jane Grimwood[4], Peggy Ozias-Akins[2,13], Steven B. Cannon[23,31], Scott A. Jackson[1,2,3,31*] and Jeremy Schmutz[4,29,31*]

[1]Center for Applied Genetic Technologies, University of Georgia, Athens, GA, USA. [2]Institute of Plant Breeding, Genetics and Genomics, University of Georgia, Athens, GA, USA. [3]Department of Crop and Soil Science, University of Georgia, Athens, GA, USA. [4]HudsonAlpha Institute of Biotechnology, Huntsville, AL, USA. [5]The Center for Genome Architecture, Baylor College of Medicine, Houston, TX, USA. [6]Instituto de Botánica del Nordeste (CONICET-UNNE), Corrientes, Argentina. [7]FACENA, Universidad Nacional del Nordeste, Corrientes, Argentina. [8]Department of Plant Pathology, University of Georgia, Tifton, GA, USA. [9]Interdepartmental Genetics Graduate Program, Iowa State University, Ames, IA, USA. [10]National Center for Genome Resources, Santa Fe, NM, USA. [11]Center of Excellence in Genomics & Systems Biology, International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), Hyderabad, India. [12]Department of Computer Science, Iowa State University, Ames, IA, USA. [13]Department of Horticulture, University of Georgia, Tifton, GA, USA. [14]UMR5096, Laboratoire Génome et Développement des Plantes, CNRS, Perpignan, France. [15]UMR5096, Laboratoire Génome et Développement des Plantes, Université de Perpignan, Perpignan, France. [16]Crop Protection and Management Research Unit, US Department of Agriculture, Agricultural Research Service, Tifton, GA, USA. [17]Corporate R&D, LG Chem, Seoul, Republic of Korea. [18]UMR232, Diversité, Adaptation et Développement des Plantes, IRD, Montpellier, France. [19]UMR232, Diversité, Adaptation et Développement des Plantes, Université de Montpellier, Montpellier, France. [20]Embrapa Genetic Resources and Biotechnology, Brasília, Brazil. [21]Department of Frontier Research and Development, Kazusa DNA Research Institute, Kisarazu, Japan. [22]National Institute of Plant Genome Research, New Delhi, India. [23]Corn Insects and Crop Genetics Research Unit, US Department of Agriculture Agricultural Research Service, Ames, IA, USA. [24]Henan Provincial Key Laboratory for Genetic Improvement of Oil Crops, Industrial Crops Research Institute, Henan Academy of Agricultural Sciences, Zhengzhou, China. [25]Key Laboratory of Oil Crops in Huanghuaihai Plains, Ministry of Agriculture and Rural Affairs, Zhengzhou, China. [26]Genome Center, University of California, Davis, Davis, CA, USA. [27]Crop Genetics and Breeding Research Unit, US Department of Agriculture Agricultural Research Service, Tifton, GA, USA. [28]Genomics and Bioinformatics Research Unit, US Department of Agriculture Agricultural Research Service, Stoneville, MS, USA. [29]Department of Energy, Joint Genome Institute, Walnut Creek, CA, USA. [30]These authors contributed equally: David J. Bertioli, Jerry Jenkins, Josh Clevenger. [31]These authors jointly supervised this work: Steven B. Cannon, Scott A. Jackson, Jeremy Schmutz. *e-mail: bertioli@uga.edu; sjackson@uga.edu; jschmutz@hudsonalpha.org

Supplementary Note 1

For**:** The genome sequence of segmental allotetraploid peanut *Arachis hypogaea*

Evidence of the use and movement of wild *Arachis* species by ancient inhabitants of South America

All *Arachis* species produce soft nutritious seeds, an attractive food that have long been used by humans. Archaeological remains and remnant populations of *Arachis* species far from their natural distributions testify to the human use and cultivation of *Arachis* species since prehistoric times. In this Supplemental Note we present the most compelling evidence with which we are familiar. In the annotated map (**Supplementary Note Fig. 1-1;** see below), inferred places of origin and movement are indicated by stars and dashed-line arrows respectively. Solid circles and distributions outlined by dashed lines indicate anthropogenic populations. Dotted circles indicate archeological remains of *Arachis* fruits.

It is important to emphasize that the topography of the regions that separate the areas of origin and derived locations present great difficulties for natural dispersion of *Arachis* seeds. This is because of the unusual reproductive biology of the genus; whilst the flowers develop above ground, a special 'peg' structure pushes the young pod underground, where development is completed (Smith 1950). This limits the usual dispersion of seeds to within an area of roughly 1 m in diameter covered by the mother plant. Therefore, populations are quite static over long periods of time: over a thousand years, they can usually move only about 1 km. Rarely, water-driven soil erosion will disperse seeds downhill. This pattern of dispersal has led to the distribution of species being heavily influenced by hydrographic basins (Krapovickas and Gregory 2007).

Red: Fragments of a peanut hull (morphologically compatible with fruits of wild species) radiocarbon dated to ~8,500 years before present were recovered from a buried house in the Ñanchoc valley, pacific coast of Peru (Dillehay et al. 2007). The sites at which they were recovered are far removed from the known range of wild *Arachis*. The closest wild populations (*A. williamsii* Krapov. & W.C. Gregory*, A. trinitensis* Krapov. & W.C. Gregory) are known from the Mamoré River in Bolivia, on the other side of the Andean Cordillera whose passes are above 4500 m in elevation, 1500 km distant from the Pacific coast (Krapovickas and Gregory 2007). Since the Valley is not a domestication center, the

adoption of peanut and other crops suggests that these plants must have been cultivated elsewhere earlier than this date, after which groups of local traders, mobile horticulturalists or others brought them into the valley. With the Andean barrier, the transportation by humans is very evident in this case.

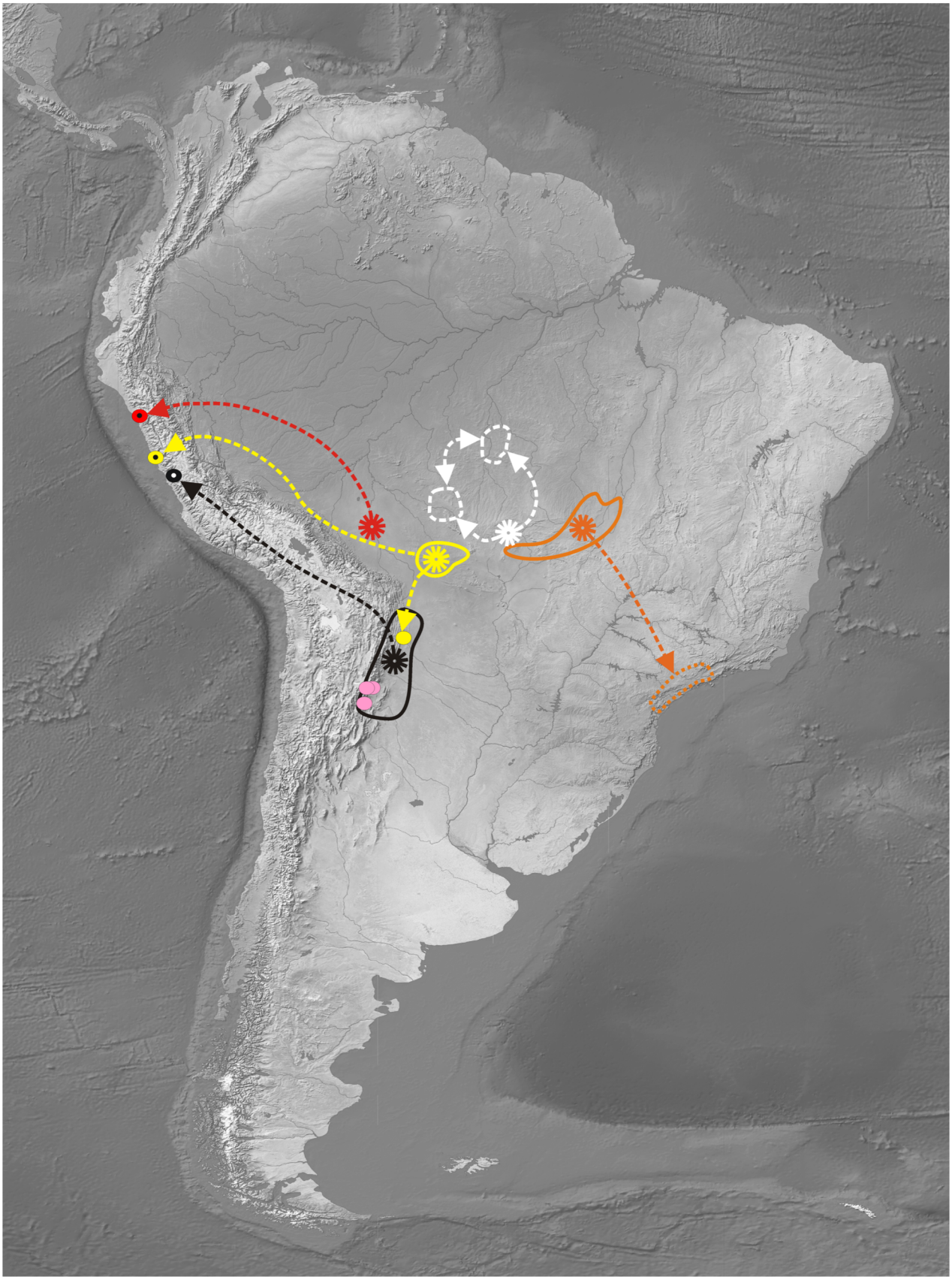
White: *Arachis villosulicarpa* Hoehne is characterized by its large (10-18mm) S-shaped seed and pegs much stronger than can be found in wild species. It is only known as cultivated by indigenous groups of west-central Mato Grosso, like the Nambiquara people. With the exception of this species, none of the other species of the botanical section Extranervosae has ever been collected west of the Paraguay River, which from its source, constitutes the western limit for the section. The closest related wild species is *A. pietrarelli* Krapov. & W.C. Gregory that lives in a small region to the SE of the localities in which *A. villosulicarpa* were found. In all the populations collected, the natives spoken to had no knowledge of this species in the wild state, each tribe maintaining their own seed (Krapovickas and Gregory 2007). All evidence supports this species as an independent domestication of a diploid *Arachis* species.

Orange: *Arachis stenosperma* Krapov. & W.C. Gregory is characterized by the long, narrow, cylindrical fruit and seeds. It occurs in the southeast part of the state of Mato Grosso, within the area where other species of the botanical section *Arachis* naturally occur. It also grows on the Atlantic coast far from any other species of the section, where it is found in soils of almost pure sand, from Rio de Janeiro to Paranaguá in the state of Paraná (Brazil). This disjunct range was interpreted as an obvious case of cultivation and transportation by humans from the Mato Grosso region to the Atlantic coast of Brazil. It is interesting to note that on the Atlantic coast it was always found in ruderal habitats, while in Mato Grosso, some populations grew in a relatively undisturbed environment. It was hypothesized that seed transportation may have occurred through a pre-Hispanic road called “Peabirá”, that connected the Paraná River with the Atlantic coast (JFM Valls pers. commun. in Krapovickas and Gregory 2007).

Yellow. *Arachis ipaënsis* Krapov. & W.C. Gregory, the B genome donor of peanut (Bertioli et al. 2016), is known from only one population located at ~600 km distant from any other species of the B genome (Krapovickas and Gregory 2007). No ancient nor present river exists to explain the movement of this species in a NE-NW direction by hydrochory. Niche modelling of the B species could not explain the location of *A. ipaënsis* by natural dispersion at the site in which it was collected (Seijo et al. unpublished). Archeological peanut shells which closely resemble those of *A. magna* Krapov., W.C. Gregory and C.E. Simpson, *A. ipaënsis*, and/or *A. monticola* Krapov. and Rigoni were excavated in the Casma and Bermejo Valleys, Peru, from a layer where there was no indication of the presence of corn. These shells were dated at 1800 to 1500 B.C. (Simpson et al. 2001). These findings strongly indicate movement and use of the B genome wild species by humans.

Black: In a dig near Casma Valley, shells were found that closely resemble *A. duranensis* Krapov. and W.C. Gregory dated at about 1800 to 1500 B.C. (D.J. Banks, pers. commun. in Simpson et al. 2001). Movement of *A. duranensis* seeds by humans were also proposed to account for the geographical distributions of polymorphisms in chloroplast (Grabiele et al. 2012) and genomic DNA (this work) for this species.

Pink: *Arachis monticola* Krapov. and Rigoni has been found at only three different sites in two unconnected valleys on the sources of the Bermejo (Río Grande de Jujuy) and Salado (Rio Juramento) Rivers in NW Argentina. This wild form of peanut was only found in places that were occupied and cultivated by ancient natives at altitudes higher than either of the diploid progenitor species occur.



Supplementary Note Fig. 1-1. Annotated map of evidence of cultivation of *Arachis* in prehistory, inferred places of origin and movement are indicated by stars and dashed-line arrows respectively. Solid circles and distributions outlined by dashed lines indicate anthropogenic populations. Dotted circles indicate archeological remains of *Arachis* fruits. Please refer to main text for color keys and explanations. The figure was generated using Natural Earth.

References

Bertioli, D.J. et al The genome sequences of *Arachis duranensis* and *Arachis ipaensis*, the diploid ancestors of cultivated peanut. *Nature Genetics*, **47**, 438. (2015).

Dillehay, T.D., Rossen, J., Andres, T.C. and Williams, D.E., Preceramic adoption of peanut, squash, and cotton in northern Peru. *Science*, **316**, 1890-1893 (2007).

Grabiele, M., Chalup, L., Robledo, G. & Seijo, G. Genetic and geographic origin of domesticated peanut as evidenced by 5S rDNA and chloroplast DNA sequences. *Plant Syst. Evol.* **298**, 1151–1165 (2012).

Krapovickas, A. & Gregory, W.C. Taxonomy of the genus *Arachis* (Leguminosae). *Bonplandia* **16** (Supl.), 1–205 (2007) [transl.].

Simpson, C.E., Krapovickas, A. & Valls, J.F.M. History of *Arachis* including evidence of *A. hypogaea* L. progenitors. *Peanut Sci.* **28**, 78–80 (2001).

Smith, B. *Arachis hypogaea*, aerial flower and subterranean fruit. *Am. J. Bot.* **37**, 802–850 (1950).

Supplementary Note 2

For: The genome sequence of segmental allotetraploid peanut *Arachis hypogaea*

Repetitive DNA

Identifying repetitive DNA in the Tifrunner genome. Mobile elements were identified using a number of homology and *de novo* structural pattern finding algorithms and manual curation.

LTR retrotransposons were identified in the assembled genome sequence. Both LTR_FINDER (Xu and Wang 2007) and LTRharvest (Ellinghaus et al. 2008) were used with default parameters except with minimum LTR length and minimum distance between LTRs of 50 bp and minimum LTR size of 50 bp with LTR_Finder for identifying small LTR retrotransposons. Sequences were extracted and grouped using BLASTN (Altschul et al. 1990) and Perl scripts. Retrotransposon sequences were manually classified and annotated with the aid of the output from LTR_FINDER and hmmsearch (Eddy 2011). False positive annotations including tandem repeats, fragmental elements and other sequences were discarded.

Long interspersed nuclear elements (LINEs) were identified using tblastn homology searches with reference LINE reverse transcriptase domains as queries (Wicker et al. 2007). All significant hits (E-value < 10^{-6}) and the flanking sequences (5 kb for each side) were extracted and inspected. Complete LINEs were determined by the terminal motifs including poly-A (L1 group) and short repeat (RTE group) and target site duplication (TSD). To identify potential short interspersed nuclear elements (SINEs), regions flanking polyA sequences were extracted using a custom Perl script. Candidates were also identified by the SINE-Finder software (Wenke et al. 2011) Sequences were manually inspected for poly-A tails and TSDs to identify *bona fide* elements. Elements were grouped using BLASTN, and representatives selected.

DNA transposons were detected by using the conserved domains of different DNA transposon superfamilies as queries in tblastn homology searches. Similarities with e-value < 10^{-6} and their 20 Kb flanking sequences (10 Kb each side) were extracted and aligned to define the boundaries. Additionally, MITE-Hunter (Han and Wessler 2010) was used to identify the small DNA transposons that encode no DNA transposases. Endogenous plant pararetroviruses (EPRVs) were identified using homology searches and exemplars extracted for characterization using Perl scripts. Identified transposons were compared with our previous

transposon sequences in the diploid wild species (Bertioli et al. 2016). The names of homologous transposons were used except adding 'Ah' before the names.

To determine the transposon distributions, all transposons were combined together and used as a library to screen the genome using RepeatMasker (http://www.repeatmasker.org/) with default parameters except with -nolow and -norna to not mask low-complexity sequences and rDNA, respectively. The output files were summarized using a custom Perl script, and regions masked by more than one sequence in the repeat library were recognized and counted only once. Base-pair counts excluded gaps.

Identifying active transposable elements though their Circular DNAs. Transposable elements generate extrachromosomal circular DNAs when active (Lanciano et al. 2017). Extrachromosomal circular DNAs were isolated from genomic DNA using the PlasmidSafe DNase (Epicentre, USA) according to the manufacturer's instructions except that the 37°C incubation was performed for 17 h. Circular DNAs were then amplified by random rolling circle amplification using the Illustra TempliPhi kit (GE Healthcare, USA) according to the manufacturer's instructions except that the incubation was performed for 65 h at 28°C. Amplified DNA was used to prepare libraries and sequenced using the MiSeq platform (Illumina, USA) using 250 bp paired-end sequencing.

After quality filtering, to remove any read originating from organelle circular genomes, sequence data was mapped against the mitochondria and chloroplast genomes using Bowtie2 version 2.2.2 71 with sensitive local mapping. Unmapped reads were then mapped against the appropriate reference genomes using parameters: –sensitive local, -k 1. Finally, the bam alignment files were normalized and visualized with the Integrative Genomics Viewer (IGV) software (Broad Institute, https://www.broadinstitute.org/igv/). Sequence reads were also assembled *de novo* using the A5-miseq pipeline (Coil et al. 2014). For each library, fasta and bam files were obtained and the idxstats module of SAMtools was used to determine the read number corresponding to each assembled scaffold. The coverage data was normalized by the total number of reads used for the *de novo* assembly. Filtered scaffolds were annotated using a BLAST analysis (-p -m 8) against organelle genomes and the databases of repetitive DNAs allowing for one hit per scaffold (-b 1 -v 1 options) and for an e-value < 10^{-2}.

References

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).

Bertioli, D.J. *et al* The genome sequences of *Arachis duranensis* and *Arachis ipaensis*, the diploid ancestors of cultivated peanut. *Nature Genetics*, **48**, 438. (2016).

Coil, D., Jospin, G. & Darling, A.E. A5-miseq: an updated pipeline to assemble microbial genomes from Illumina MiSeq data. *Bioinformatics* **31**, 587-589. (2014).

Eddy, S.R. Accelerated Profile HMM Searches. PLoS Computation Biology **7**, e1002195 (2011).

Ellinghaus, D., *et al.* LTRharvest, an efficient and flexible software for de novo detection LTR retrotransposons. *BMC Bioinformatics* **9**,18 (2008).

Han, Y. & Wessler, S.R. MITE-Hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences. *Nucleic Acids Res*. **38**, e199 (2010).

Lanciano, S. *et al.* Sequencing the extrachromosomal circular mobilome reveals retrotransposon activity in plants. *PLoS Genet.* **13**, 1–20 (2017).

Wenke, T. *et al.* Targeted identification of short interspersed nuclear element families shows their widespread existence and extreme heterogeneity in plant genomes. *Plant Cell* **23**, 3117-2128 (2011).

Wicker, T. *et al.* A unified classification system for eukaryotic transposable elements. *Nat. Rev. Genet*. **8**, 973-982 (2007).

Supplementary Note 3

For: The genome sequence of segmental allotetraploid peanut *Arachis hypogaea*

Statistical analysis.

The comparison of expression of homeologous gene pairs was done using the DESeq2 (Love et al. 2014)) package in R based on the negative binomial distribution. Only genes with $\log_2$ fold change >=1, Benjamini-Hochberg adjusted P <0.05 were retained. The comparison of highly expressed homeologous gene pairs between subgenomes in different tissues was carried out using binomial test with the odds of a subgenome being more highly expressed at 0.5 probability, P <0.05 were considered significant. Gene Ontology (GO) enrichment analysis was carried out using topGO (Alexa et al. 2006 ; Alexa et al 2016) an R Bioconductor package with Fisher's exact test; only GO terms with a P <0.05 (FDR < 0.05) were considered significant. Exact P values are listed in **Dataset 2** (hosted at https://doi.org/10.25739/hb5x-wx74, Cyverse and Peanutbase). The statistical analysis of genome methylation was carried out using the Wilcoxon rank-sum test, P values are listed on **Supplementary Fig. 18**.

References

Love, M.I., Huber, W. & Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**, 550 (2014).

Alexa, A., Rahnenführer, J., Lengauer, T. Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics* **22**, 1600–7 (2006).

Alexa, A., & Rahnenfuhrer, J. topGO: Enrichment Analysis for Gene Ontology. *R Package version* **2**, 2240. (2016)

Supplementary Tables for

The genome sequence of segmental allotetraploid peanut *Arachis hypogaea*

Supplementary Table 1
Summary assembly statistics for chromosome scale assembly

Scaffold total	384
Contig total	4,037
Scaffold length total	2,556.3 Mb
Chromosome Sequence	2,534.1 Mb
Contig sequence total	2,552.5 Mb (0.1% gap)
Scaffold N/L50	9 / 134.9 Mb
Contig N/L50	461 / 1.5 Mb

Supplementary Table 2
Chromosomal pseudomolecule and scaffold metrics for genome assembly

Scaffold name	No. Contigs	Size (bp)	%GC
Arahy.15	150	160,879,708	36.59%
Arahy.19	233	158,625,764	36.59%
Arahy.16	222	154,808,347	36.61%
Arahy.11	163	149,299,306	36.90%
Arahy.13	174	146,725,006	36.24%
Arahy.20	183	143,980,330	36.75%
Arahy.03	180	143,813,506	35.94%
Arahy.14	152	143,237,272	36.49%
Arahy.18	206	135,150,084	36.65%
Arahy.17	163	134,922,436	36.72%
Arahy.04	238	128,801,742	36.54%
Arahy.12	136	120,579,088	36.10%
Arahy.09	130	120,519,698	36.32%
Arahy.10	236	117,088,237	36.15%
Arahy.05	173	115,930,344	36.06%
Arahy.06	213	115,504,342	36.48%
Arahy.01	157	112,420,854	35.90%
Arahy.02	195	102,981,163	35.91%
Arahy.07	113	81,119,488	35.79%
Arahy.08	45	51,897,010	33.47%
21	35	1,982,220	24.95%
22	24	1,569,740	44.68%
23	35	1,519,527	38.42%
25	23	983,129	31.34%
26	17	868,297	24.55%
27	19	745,707	41.34%
28	3	672,389	41.71%
29	13	553,903	30.76%
30	13	419,763	38.68%
31	3	322,620	27.05%
32	10	267,858	51.22%
33	5	202,291	31.99%
34	6	195,354	35.09%
35	1	181,829	32.24%
36	1	169,023	36.91%
37	1	168,196	39.82%
38	4	160,818	22.76%
39	4	160,336	45.81%
40	2	160,171	26.80%
42	3	128,001	28.68%
43	1	107,484	33.73%
44	3	102,637	40.57%

Supplementary Table 3
The summary of repeat annotation

Transposable element	Copy number ($X10^3$)	Coverage (bp)	Content (%)
Class I	1,827.41	1,644,831,268	64.43
LTR retrotransposon	1,694.04	1,578,596,840	61.83
Ty1/copia	159.55	120,565,582	4.72
Ty3/gypsy	707.58	942,750,004	36.93
TRIM	16.42	5,225,833	0.20
other	810.48	595,215,524	23.32
Non-LTR retrotransposon	133.37	72,064,474	2.82
LINEs	116.53	69,834,293	2.74
SINEs	16.84	2,256,947	0.09
Class II	592.33	282,447,447	11.06
CACTA	220.61	141,956,302	5.56
Harbinger/PIF	2.63	1,703,651	0.07
hAT	67.24	25,712,050	1.01
Helitron	65.50	22,345,038	0.88
mutator	236.34	92,387,871	3.62
Pararetrovirus	2.02	6,575,859	0.26
Tandem repeats		36,911,444	1.45
Total TE coverage	2,421.75	1,890,047,658	74.03

Note: Overlapping regions were counted only once for the Total TE coverage.
Total TE coverage excludes tandem repeats.

Supplementary Table 4
Summary of repeat annotation for each Chromosome

Chromosome	RNA transposon			DNA transposon			Total transposon		
	Copy number ($X10^3$)	Coverage (bp)	Content (%)	Copy number ($X10^3$)	Coverage (bp)	Content (%)	Copy number ($X10^3$)	Coverage (bp)	Content (%)
Arahy.01	79.07	70,434,548	62.74	26.2	13,076,060	11.64	105.47	83,510,608	74.39
Arahy.02	70.21	63,373,940	61.68	26.65	12,370,699	12.04	95.86	75,744,639	73.72
Arahy.03	99.73	87,591,932	61.00	34.84	16,914,242	11.78	134.57	104,506,174	72.78
Arahy.04	91.19	85,380,662	66.42	29.51	14,862,594	11.56	120.70	100,243,256	77.98
Arahy.05	79.25	70,836,832	61.19	29.14	13,593,359	11.74	108.39	84,430,191	72.94
Arahy.06	79.41	73,565,858	63.81	27.29	13,377,300	11.60	106.71	86,943,158	75.42
Arahy.07	56.853	48,981,019	60.46	21.00	9,881,427	12.20	77.85	58,862,446	72.66
Arahy.08	29.57	19,463,094	37.54	18.87	6,334,441	12.22	48.43	25,797,535	49.76
Arahy.09	86.36	77,996,439	64.79	28.09	14,209,957	11.80	114.45	92,206,396	76.59
Arahy.10	80.97	75,960,376	65.00	26.57	13,353,587	11.43	107.54	89,313,963	76.43
Arahy.1-10	752.60	673,584,700	61.89	267.35	127,973,666	11.76	1,019.95	801,558,366	73.65
Arahy.11	113.00	104,453,996	70.04	31.68	15,799,513	10.59	144.68	120,253,509	80.63
Arahy.12	86.91	77,157,318	64.06	29.19	13,380,777	11.11	116.11	90,538,095	75.17
Arahy.13	101.63	89,753,035	61.25	36.01	15,854,493	10.82	137.64	105,607,528	72.07
Arahy.14	106.32	96,636,566	67.54	32.66	15,690,775	10.97	138.97	112,327,341	78.50
Arahy.15	121.82	109,950,628	68.42	34.31	15,967,351	9.94	156.13	125,917,979	78.36
Arahy.16	113.14	103,230,754	66.79	35.37	16,864,452	10.91	148.51	120,095,206	77.70
Arahy.17	99.71	91,015,261	67.54	28.88	13,419,852	9.96	128.60	104,435,113	77.50
Arahy.18	99.34	90,469,021	67.05	29.57	14,066,832	10.43	128.91	104,535,853	77.48
Arahy.19	118.32	108,210,961	68.32	34.47	17,528,860	11.07	152.80	125,739,821	79.39
Arahy.20	107.56	99,325,152	69.07	31.50	15,213,929	10.58	139.06	114,539,081	79.65
Arahy.11-20	1,067.75	970,202,692	67.08	323.65	153,786,834	10.63	1,391.40	1,123,989,526	77.71

Supplementary Table 5
Summary of genome methylation

	Number of methylated Cytosines	Number of unmethylated Cytosines	Percentage methylated
A-subgenome			
CG	42027680	13269654	76.00
CHG	40507958	25135604	61.71
CHH	23827157	439972180	5.14
B-subgenome			
CG	57173235	13855996	80.49
CHG	53749002	28764160	65.14
CHH	31790492	544836904	5.51
Both subgenomes			
CG	99200915	27125650	78.53
CHG	94256960	53899764	63.62
CHH	55617649	984809084	5.35

Supplementary Table 6. Single nucleotide polymorphisms assignable to ancestral A/B genomes in Tifrunner chromosomes

Whole Chromosomes - assignable single nucleotide polymorphisms (SNPs)

	A SNPs	B SNPs	Proportion B		A SNPs	**B SNPs**	Proportion A
Arahy.01	**572,704**	13,090	0.0223	**Arahy.11**	4,664	**660,427**	0.0070
Arahy.02	**455,861**	13,819	0.0294	**Arahy.12**	15,711	**526,486**	0.0290
Arahy.03	**696,004**	33,160	0.0455	**Arahy.13**	6,926	**732,512**	0.0094
Arahy.04	**546,306**	30,670	0.0532	**Arahy.14**	3,192	**666,930**	0.0048
Arahy.05	**561,488**	17,020	0.0294	**Arahy.15**	40,182	**672,022**	0.0564
Arahy.06	**546,081**	19,016	0.0337	**Arahy.16**	19,975	**618,253**	0.0313
Arahy.07	**439,632**	12,546	0.0277	**Arahy.17**	4,219	**597,357**	0.0070
Arahy.08	**393,684**	10,424	0.0258	**Arahy.18**	7,449	**578,008**	0.0127
Arahy.09	**581,794**	19,326	0.0321	**Arahy.19**	4,260	**713,842**	0.0059
Arahy.10	**561,733**	15,037	0.0261	**Arahy.20**	3,289	**681,023**	0.0048
A sub-gen. sum	**5,355,287**	184,108	0.0332	**B subgen. sum**	109,867	**6,446,860**	0.0168

Supplementary Table 7
Assignable single nucleotide polymorphisms in chromosome segments derived from ancestral homeologs

Only chromosomes with segments from homeologous subgenome shown

	A SNPs	B SNPs	chrom. region bp
Arahy.04			**13,873**
Arahy.04	545,846	17,863	**126,021,859**
Arahy.04			**126,022,314**
Arahy.04	9	3,865	**126,861,072**
Arahy.04			**126,871,298**
Arahy.04	426	54	**126,999,780**
Arahy.04			**127,000,094**
Arahy.04	25	8,888	**128,750,405**

	A SNPs	B SNPs	chrom. region bp
Arahy.06			**2,098**
Arahy.06	536,244	16,448	**112,992,505**
Arahy.06			**112,992,729**
Arahy.06	9	2,224	**113,529,472**
Arahy.06			**113,530,170**
Arahy.06	9,828	344	**115,496,864**

B SNPs within segments in A chromosomes
14,977

	A SNPs	B SNPs	chrom. region bp
Arahy.12			**27,779**
Arahy.12	2,351	525,917	**118,257,157**
Arahy.12			**118,257,663**
Arahy.12	13,360	569	**120,571,742**

	A SNPs	B SNPs	chrom. region bp
Arahy.13			**5,399**
Arahy.13	3,088	732,489	**146,346,380**
Arahy.13			**146,346,556**
Arahy.13	3,837	23	**146,718,123**

	A SNPs	B SNPs	chrom. region bp
Arahy.14			**2,538**
Arahy.14	2,773	657,818	**141,299,785**
Arahy.14			**141,299,804**
Arahy.14	403	96	**141,435,136**
Arahy.14			**141,435,994**
Arahy.14	16	9,016	**143,225,797**

	A SNPs	B SNPs	chrom. region bp
Arahy.15			**33**
Arahy.15	37,006	1,671	**6,308,081**
Arahy.15			**6,308,629**
Arahy.15	3,176	670,352	**160,872,434**

	A SNPs	B SNPs	chrom. region bp
Arahy.16			**6,153**
Arahy.16	3,725	581,125	**143,754,943**
Arahy.16			**143,755,256**
Arahy.16	7,350	519	**146,257,546**
Arahy.16			**146,258,032**
Arahy.16	134	36,196	**152,833,466**
Arahy.16			**152,834,963**
Arahy.16	8,764	412	**154,798,381**

	A SNPs	B SNPs	chrom. region bp
Arahy.17			**22,563**
Arahy.17	13	3,111	**628,749**
Arahy.17			**629,946**
Arahy.17	1,113	86	**1,438,126**
Arahy.17			**1,441,816**
Arahy.17	0	594,159	**134,918,226**

	A SNPs	B SNPs	chrom. region
Arahy.18			**2,956**
Arahy.18	1,247	34	**366,912**
Arahy.18			**374,792**
Arahy.18	6,194	577,956	**135,142,461**

	A SNPs	B SNPs	chrom. region
Arahy.19			3,733
Arahy.19	353	57	191,811
Arahy.19			204,118
Arahy.19	3,907	713,785	158,598,766

A SNPs within segments in B chromosomes
73,433

Supplementary Table 8

Estimated DNA identities by chromosome of different *A. duranensis* accessions to the A subgenome of *A. hypogaea* cv. Tifrunner*

(using averages of Illumina whole genome sequences)

Accession	Voucher	Location	Arahy.01	Arahy.02	Arahy.03	Arahy.04	Arahy.05	Arahy.06	Arahy.07	Arahy.08	Arahy.09	Arahy.10
PI-468202	KGBSPSc 30065	Rio Seco	99.57	99.31	99.25	99.45	99.45	99.45	99.39	99.20	99.35	99.53
PI-468201	KGBSPSc 30067	Rio Seco	99.57	99.31	99.25	99.45	99.44	99.45	99.38	99.19	99.37	99.53
S2741	S2741	Rio Seco	99.51	99.17	99.08	99.32	99.34	99.34	99.26	99.08	99.24	99.46
V14167	V14167	Salta city	99.39	99.24	98.90	99.09	99.21	99.12	98.93	98.88	99.27	99.42
ICG8138	?	Rio Arenales, Salta	99.35	99.11	98.68	98.98	99.64	99.05	98.76	98.88	99.18	99.39
PI-468323	KGBSPSc 30077	Chuquisaca, Bolivia	98.89	99.19	98.98	99.19	99.28	99.12	99.01	98.79	99.06	99.19
PI-475845	KGBSPSc 30070	Tarija, Bolivia	98.14	98.15	98.04	98.27	98.29	98.22	98.19	97.91	98.06	98.25

Supplementary Table 9

Exome similarity metrics of *A. duranensis* accessions to A subgenome of *A. hypogaea* cv. Tifrunner per chromosome*

Accession	Voucher	Arahy.01	Arahy.02	Arahy.03	Arahy.04	Arahy.05	Arahy.06	Arahy.07	Arahy.08	Arahy.09	Arahy.10
PI-468201	KGBSPSc 30065	0.9227	0.8885	0.8867	0.9068	0.9052	0.9070	0.9056	0.9069	0.8954	0.9146
PI-468202	KGBSPSc 30067	0.9164	0.8899	0.8880	0.9064	0.9061	0.9060	0.9038	0.9057	0.8953	0.9131
Grif-14269	WiSVg 1296	0.9163	0.8992	0.8844	0.8968	0.8971	0.9050	0.9078	0.9042	0.8892	0.9062
PI-497269	KSSc 38905	0.9139	0.9001	0.8827	0.8955	0.8996	0.9038	0.9045	0.9038	0.8906	0.9070
PI-468200	KGBSPSc 30064	0.9131	0.8962	0.8739	0.8855	0.8941	0.8839	0.8801	0.8974	0.9003	0.9144
Grif-14265	WiSVg 1272	0.9121	0.8978	0.8736	0.8797	0.8873	0.8824	0.8763	0.8956	0.8998	0.9128
PI-475886	KSBScC 36006	0.9110	0.9011	0.8746	0.8806	0.9208	0.8827	0.8780	0.8949	0.9019	0.9121
	V14167	0.9110	0.8991	0.8759	0.8834	0.8945	0.8846	0.8797	0.8964	0.9024	0.9101
PI-497263	ScV 21764	0.9109	0.9015	0.8759	0.8826	0.8929	0.8843	0.8798	0.8969	0.9026	0.9135
PI-468198	KGBSPSc 30061	0.9109	0.8941	0.8695	0.8802	0.8883	0.8802	0.8739	0.8937	0.8988	0.9115
PI-475882	KSBScC 36002	0.9105	0.8964	0.8793	0.8799	0.9026	0.8877	0.8876	0.8985	0.8928	0.9093
PI-468197	KGBSPSc 30060	0.9101	0.8937	0.8710	0.8796	0.8883	0.8810	0.8748	0.8941	0.8992	0.9114
PI-497262	ScV 21763	0.9099	0.8993	0.8736	0.8839	0.8890	0.8835	0.8780	0.8932	0.9015	0.9117
	Se2741	0.9088	0.8918	0.8815	0.8989	0.9025	0.9002	0.8986	0.9006	0.8903	0.9055
PI-497264	ScV 21766	0.9082	0.8983	0.8722	0.8805	0.8905	0.8812	0.8763	0.8944	0.8994	0.9132
PI-475884	KSBScC 36004	0.9068	0.8929	0.8708	0.8782	0.8886	0.8826	0.8746	0.8936	0.8963	0.9056
PI-262133	GKP 10038	0.9055	0.8928	0.8678	0.8742	0.9387	0.8756	0.8703	0.8888	0.8953	0.9053
	S2822	0.9049	0.8962	0.8690	0.8821	0.8982	0.8833	0.8737	0.8920	0.8882	0.9054
PI-475883	KSBScC 36003	0.9038	0.8914	0.8712	0.8712	0.8947	0.8790	0.8786	0.8921	0.8880	0.9039
	S2856	0.9021	0.8987	0.8687	0.8825	0.8927	0.8834	0.8781	0.8942	0.8973	0.9092
PI-497265	ScV 21767	0.9006	0.8961	0.8693	0.8792	0.8888	0.8807	0.8743	0.8937	0.8978	0.9097
PI-475844	KGBSPSc 30069	0.8956	0.8909	0.8702	0.8732	0.8974	0.8803	0.8801	0.8923	0.8854	0.9044
PI-497270	KSSc 38906	0.8935	0.8910	0.8695	0.8687	0.8930	0.8814	0.8729	0.8887	0.8873	0.8982
	S2848	0.8922	0.8845	0.8685	0.8694	0.8931	0.8654	0.8718	0.8834	0.8787	0.8955
PI-468324	KGBSPSc 30078	0.8906	0.8821	0.8671	0.8738	0.9321	0.8744	0.8781	0.8897	0.8950	0.9019
PI-475847	KGBSPSc 30072	0.8671	0.8813	0.8705	0.8868	0.8983	0.8886	0.8819	0.8945	0.8710	0.8976
	K7988	0.8658	0.8861	0.8598	0.8821	0.8873	0.8738	0.8803	0.8843	0.8715	0.8936
PI-475846	KGBSPSc 30071	0.8648	0.8807	0.8699	0.8823	0.9007	0.8806	0.8859	0.8967	0.8745	0.9045
PI-497268	KSSc 38904	0.8646	0.8870	0.8760	0.8783	0.9044	0.8849	0.8845	0.8836	0.8750	0.9011
PI-497267	KSSc 38903	0.8640	0.8798	0.8649	0.8856	0.8968	0.8855	0.8819	0.8858	0.8738	0.8981
PI-219823	No data	0.8640	0.8868	0.8640	0.8845	0.8940	0.8774	0.8856	0.8851	0.8745	0.8948
PI-497266	KSSc 38900	0.8630	0.8937	0.8675	0.8786	0.8986	0.8827	0.8755	0.8915	0.8842	0.9027
	Se3350	0.8622	0.8768	0.8661	0.8817	0.8933	0.8856	0.8806	0.8869	0.8655	0.8938
PI-475845	KGBSPSc 30070	0.8534	0.8753	0.8651	0.8837	0.8925	0.8808	0.8782	0.8834	0.8652	0.8974
	Se3146	0.8506	0.8832	0.8616	0.8612	0.8882	0.8658	0.8691	0.8764	0.8622	0.8836
PI-497484	KSSc 38902	0.8504	0.8827	0.8603	0.8735	0.8874	0.8804	0.8727	0.8825	0.8651	0.8849
PI-468319	KGBSPSc 30073	0.8493	0.8742	0.8620	0.8814	0.8921	0.8867	0.8733	0.8833	0.8619	0.8875
PI-468203	GKBSPSc 30064	0.8491	0.8823	0.8671	0.8811	0.8956	0.8853	0.8731	0.8812	0.8645	0.8872
Grif-15039	WiSVgJsQ 1507	0.8465	0.8638	0.8622	0.8606	0.8715	0.8645	0.8591	0.8760	0.8626	0.8668
Grif-14264	WiSVg 1275	0.8461	0.8631	0.8590	0.8619	0.8702	0.8675	0.8633	0.8724	0.8615	0.8641
	Se3139	0.8458	0.8766	0.8547	0.8789	0.8860	0.8704	0.8652	0.8749	0.8623	0.8748
PI-475885	KSBScC 36005	0.8450	0.8649	0.8420	0.8375	0.8563	0.8563	0.8521	0.8642	0.8346	0.8426
PI-468323	KGBSPSc 30077	0.8447	0.8856	0.8593	0.8796	0.8968	0.8686	0.8701	0.8785	0.8626	0.8751
PI-468320	KGBSPSc 30074	0.8441	0.8803	0.8600	0.8799	0.8935	0.8813	0.8700	0.8773	0.8589	0.8839
PI-468321	KGBSPSc 30075	0.8426	0.8785	0.8529	0.8666	0.8799	0.8729	0.8690	0.8719	0.8626	0.8730
PI-497483	KSSc 38901	0.8423	0.8768	0.8538	0.8677	0.8778	0.8726	0.8662	0.8746	0.8637	0.8774
Grif-14263	WiSVg 1274	0.8389	0.8529	0.8554	0.8465	0.8601	0.8570	0.8486	0.8664	0.8536	0.8545
Grif-15036	PI 666084/WiSVg 1510-B	0.8389	0.8563	0.8565	0.8516	0.8630	0.8590	0.8525	0.8675	0.8575	0.8579
	K30078	0.8372	0.8456	0.8469	0.8432	0.8688	0.8306	0.8414	0.8539	0.8260	0.8497
Grif-15035	WiSVgJsQ 1510-A	0.8353	0.8562	0.8526	0.8453	0.8565	0.8581	0.8493	0.8642	0.8531	0.8583
Grif-15038	WiSVgJsQ 1506-W	0.8325	0.8507	0.8502	0.8512	0.8575	0.8517	0.8472	0.8639	0.8514	0.8531
Grif-14261	WiSVg 1268	0.8315	0.8450	0.8509	0.8427	0.8520	0.8494	0.8437	0.8607	0.8490	0.8481
Grif-14262	WiSVg 1270	0.8307	0.8494	0.8518	0.8423	0.8509	0.8500	0.8427	0.8614	0.8486	0.8496
Grif-15037	PI 666085/WiSVgJsQ 1506-E	0.8286	0.8475	0.8495	0.8456	0.8544	0.8500	0.8464	0.8629	0.8504	0.8506
	K30077	0.8283	0.8535	0.8422	0.8452	0.8382	0.8539	0.8372	0.8587	0.8319	0.8281

*Note exome similarity metrics are not DNA identity. Values have been red-yellow-green heat mapped

Supplementary Table 10
Tetraploid / diploid chromosomal pseudomolecule sizes and ratios for whole chromosomes and and segments inverted relative

Chroms.	Aradu	A subgenome	Tet/Dip
01	107035537	112420854	1.05
02	93869048	102981163	1.1
03	135057546	143813506	1.06
04	123556382	128801742	1.04
05	110037037	115930344	1.05
06	112752717	115504342	1.02
07	79126724	81119488	1.03
08	49462234	51897010	1.05
09	120672674	120519698	1.00
10	109463236	117088237	1.07

Choms.	Araip	B subgenome	Tet/Dip
01/11	137414913	149299306	1.09
02/12	108997779	120579088	1.11
03/13	136109863	146725006	1.08
04/14	133615181	143237272	1.07
05/15	149900536	160879708	1.07
06/16	137147148	154808347	1.13
07/17	126351151	134922436	1.07
08/18	129606920	135150084	1.04
09/19	147089397	158625764	1.08
10/20	136175642	143980330	1.06

"New"* inverted chromosome segments

	Aradu	A sub genome	Tet/Dip
05	32093060	28515710	0.89
07	8496956	8930392	1.05

"New" inverted chromosome segments

	Araip	B sub genome	Tet/Dip
01	9862242	10619870	1.08

Almost all chromosomal pseudomolecules in the tetraploid assembly are slightly larger than the corresponding chromosomal pseudomolecules in the diploids. Only the inversion on Arahy.05 is significantly smaller. *Evidence from genetic mapping supports the inversion in Arahy.07 being present in the diploid ancestor.

Supplementary Table 11
Collection localities and overall exome similarity metrics of *A. duranensis* accessions (and control species) to A subgenome of *A. hypogaea* cv. Tifrunner

Color code	Accession	Voucher	Similarity Metric to Tifrunner A subgenome. Proportion Potential Polymorphic Sites matching (Not base similarity)	Country	Dept/Province	Locality	Latitude				Longitude				altitude	
	PI-468202	KGBSPSc 30067	0.907677	Argentina	Salta	Seco River	23	2	0	S	63	56	0	W	330	
	PI-468201	KGBSPSc 30065	0.905946	Argentina	Salta	Seco River	23	4	0	S	63	54	0	W	329	
		Se, Sn, Ho, Ch 2741	0.905423	Argentina	Salta	Seco River	23	1	47	S	63	51	11	W	355	
	Grif-14269	WiSVg 1296	0.905107	Bolivia	Cordillera	San José de Chiquitos	18	32	8	S	62	38	36	W	310	
	PI-497269	KSSc 38905	0.904747	Argentina	Jujuy	Fraile Pintado	23	58	23	S	67	40	47	W	465	
	PI-262133	GKP 10038	0.896765	Argentina	Salta	Arenales River, El Prado	24	50	17	S	65	30	19	W	1283	
	PI-475886	KSBScC 36006	0.896633	Argentina	Salta	Salta City, Grad Bourg	24	44	59	S	65	28	50	W	1325	
	PI-468324	KGBSPSc 30078	0.896571	Bolivia	Chuquisaca	Carandaity	20	45	0	S	62	45	0	W	500	
	PI-475882	KSBScC 36002	0.895606	Argentina	Salta	14 Km NW El Tunal	25	16	0	S	64	32	0	W	600	
	PI-497263	ScV 21764	0.893808	Argentina	Salta	Arenales River	24	49	51	S	65	29	48	W	1280	
	Grif-14265	WiSVg 1272	0.893212	Bolivia	Cordillera	Izozog	19	20	35	S	62	25	54	W	500	
	PI-475884	KSBScC 36004	0.892900	Argentina	Jujuy	Palpala, Jujuy.	24	16	0	S	65	12	0	W	1100	
	PI-497262	ScV 21763	0.892863	Argentina	Salta	Salta City, Sporting Club	24	46	32	S	65	23	41	W	1180	
	PI-468200	KGBSPSc 30064	0.892066	Argentina	Jujuy	Jujuy City, Airport	24	16	0	S	65	12	0	W	1100	
	PI-497264	ScV 21766	0.891603	Argentina	Jujuy	Jujuy City, airport	24	16	0	S	65	12	0	W	1100	
	PI-468197	KGBSPSc 30060	0.891095	Argentina	Jujuy	Perico River bridge	24	21	37	S	65	6	16	W	940	
	PI-468198	KGBSPSc 30061	0.890329	Argentina	Jujuy	Palpalá	24	16	0	S	65	12	0	W	1100	
	PI-475844	KGBSPSc 30069	0.890222	Bolivia	Tarija	30 Km N Yacuiba, Caiza	21	45	0	S	63	32	0	W	601	
		S. 2856	0.889464	Argentina	Jujuy	El Carmen, Airport	24	22	23	S	65	5	35	W	918	
	PI-497265	ScV 21767	0.889236	Argentina	Jujuy	Palpalá	24	15	21	S	65	13	25	W	1150	
	PI-475883	KSBScC 36003	0.888894	Argentina	Salta	El Tunal	25	16	0	S	64	32	0	W	470	
		Se2822	0.888198	Argentina	Salta	El Tunal	25	14	16	S	64	28	28	W	470	
	PI-497270	KSSc 38906	0.888096	Argentina	Salta	Anta	24	40	0	S	64	15	0	W	650	
		Se SN 2848	0.887701	Argentina	Salta	Anta, El Ceibalito	24	55	42	S	64	25	13	W	639	
	V14167	V14167	0.886098	Argentina	Salta	Salta city	24	45	0	S	65	26	0	W	1130	
	PI-475846	KGBSPSc 30071	0.883291	Bolivia	Tarija	2 W Saladillo	21	36	0	S	63	45	0	W	1000	
	PI-497266	KSSc 38900	0.882793	Argentina	Salta	Las Lajitas	24	40	0	S	64	15	0	W	465	
	PI-219823	No data	0.880659	Argentina	Salta	Campo Durán	22	11	59	S	63	29	39	W	480	
	PI-475847	KGBSPSc 30072	0.879638	Bolivia	Tarija	14 Km N of Carapari	21	34	0	S	63	45	0	W	870	
	PI-497268	KSSc 38904	0.879594	Argentina	Salta	5 Km W of Dragones	23	14	58	S	63	20	47	W	241	
	PI-497267	KSSc 38903	0.87919	Argentina	Salta	Campo Duran	22	11	59	S	63	29	39	W	480	
	PI-468203	GKBSPSc 30064	0.877073	Argentina	Salta	14 Km SW of A. Quijarro	22	21	0	S	63	55	59	W	350	
	PI-475845	KGBSPSc 30070	0.876100	Bolivia	Tarija	18 Km N Yacuiba	21	50	0	S	63	37	0	W	600	
	PI-468319	KGBSPSc 30073	0.876061	Bolivia	Tarija	Fork to Sanandita	21	43	0	S	63	30	0	W	540	
	PI-468320	KGBSPSc 30074	0.873617	Bolivia	Tarija	2 Km of Palmar Grande	21	37	0	S	63	30	0	W	400	
		Se 3350	0.872834	Bolivia	Tarija	6.3 Km E de Caiza	21	49	8	S	63	31	59	W	559	
		K 7988	0.870045	Argentina	Salta	Campo Durán	22	11	59	S	63	29	39	W	480	
	Grif-15036	I 666084/WiSVg 1510-I	0.868686	Paraguay	Garay	Cano Martinez Primo	29	9	24	S	61	46	17	W	285	
	PI-468323	KGBSPSc 30077	0.868586	Bolivia	Chuquisaca	l Salvador Experimental Statio	20	45	0	S	63	13	0	W	500	
	PI-497484	KSSc 38902	0.868188	Bolivia	Tarija	n S of Pilcomayo River, Villa M	21	26	35	S	63	29	56	W	480	
	PI-468321	KGBSPSc 30075	0.866205	Bolivia	Tarija	5 Km S of Villa Montes	21	18	0	S	63	30	0	W	450	
		Se 3146	0.864574	Bolivia	Chuquisaca	Carandayti	20	40	55	S	63	6	38	W	678	
	PI-497483	KSSc 38901	0.864425	Bolivia	Tarija	Pilcomayo River, 4Km N of Vil	21	12	18	S	63	25	54	W	550	
		Se 3139	0.861219	Bolivia	Tarija	Ipa community.	21	3	15	S	63	24	53	W	558	
	Grif-14264	WiSVg 1275	0.860364	Bolivia	Cordillera	el Izozog,19 Km NW of Com. Sa	19	15	27	S	62	38	36	W	500	
	Grif-15035	WiSVgJsQ 1510-A	0.852400	Paraguay	Garay	Martinez Primo, 5 Km S of Parque Cue									285	
	Grif-15039	WiSVgJsQ 1507	0.851378	Paraguay	Garay	3 Km N of Gral. E.A. Garay	20	29	38	S	62	8	11	W	370	
	Grif-14263	WiSVg 1274	0.85104	Bolivia	Cordillera	ozog, 2.3 Km west of Comunida	19	21	2	S	62	36	7	W	500	
	Grif-14261	WiSVg 1268	0.847946	Bolivia	Cordillera	del Izozog, 5 Km west of Com.	19	32	52	S	62	35	42	W	500	
	Grif-14262	WiSVg 1270	0.847611	Bolivia	Cordillera	l Izozog, 1 Km west of Comunic	19	32	44	S	62	34	24	W	520	
	Grif-15038	WiSVgJsQ 1506-W	0.846135	Paraguay	Garay	29 Km NW of Nueva Asunción	20	32	13	S	62	7	26	W	350	ver coordenad
	Grif-15037	666085/WiSVgJsQ 150	0.843754	Paraguay	Garay	29 Km NW of Nueva Asunción									350	
		KGBSPSc 30078	0.840561	Bolivia	Chuquisaca	7 Km W Carandayti	20	45	0	S	63	9	0	W	500	
	PI-475885	KSBScC 36005	0.835900	Argentina	Salta	Salta city, Airport	24	50	20	S	65	28	40	W	1247	
		KGBSPSc 30077	0.827864	Bolivia	Chuquisaca	l Salvador Experimental Statio	20	45	0	S	63	13	0	W	500	
		A. cardenasii	0.819587	Bolivia												A. cardenasii
	PI-468325	GKBSPSc 30079	0.756208	Bolivia	Cordillara	28km of Camiri	20	16	59	S	63	28	0	W	900	A. batizocoi
	PI-331189	?	0.722979													
		GKBSPSc 30076	0.698245	Bolivia	Tarija	Ipa, 30 km N of Villa Montes	21	0	0	S	63	24	0	W	500	A. ipaënsis

Notes:

1) Germplasm accessions with "PI" or "Grif" denominations are generally available from the USDA Germplasm Resources Information Network (https://www.ars-grin.gov)

2) The top ranked two accessions KGBSPSc 30067 and KGBSPSc 30065 were both collected during an expedition in 1977 between 15th March and 19th May by A. Krapovickas, W.C. Gregory, D.J. Banks, J.R. Pietrarelli, A. Schinini & C.E. Simpson. Remarkably, this was the same expedition where *A. ipaensis* KGBSPSc 30076, the probable present day representative of the B subgenome donor of *Arachis hypogaea* (Bertioli et al 2016), was collected (Krapovickas et al 2007; "KGBSPSc" is mostly referred to in the abbreviated form "K"). We are greatly indebted to these collectors with their ingenuity and adventurous spirits, the funding agencies and Institutions that supported them, and the regulatory environment that allowed the collection and distribution of these collections to seed banks world-wide. (Sadly, since these collections, regulations following from the Convention on Biological Diversity, implemented in 1993, the Nagoya Protocol and the Andean Pact have seriously undercut the ability to collect and share germplasm.)

Supplementary Table 12

Genomic libraries included in the *Arachis hypogaea* genome assembly and their respective assembled sequence coverage levels in the final release.

PACBIO reads were used for the assembly, Illumina reads for polishing homozygous SNPs and indels, HiC was used for scaffolding

*Average read length of PACBIO reads

Library	Sequencing Platform	Average Insert Size	Read Size (bp)	Read Number	Assembled Sequence Coverage (x)
	PACBIO		**9,784***	**17,747,748**	**76.74**
ICIH	Illumina-Frag	800	2x250	294,724,162	29.59
ICID	Illumina-Frag	800	2x250	333,638,898	33.50
Fragment Total				**628,363,060**	**63.09**
AAAA	Illumina-HiC		2x85	33,208,834	1.13
AAAB	Illumina-HiC		2x85	28,934,386	0.99
AAAC	Illumina-HiC		2x85	30,030,102	1.03
AAAD	Illumina-HiC		2x85	26,565,192	0.91
AAAE	Illumina-HiC		2x150	211,752,480	12.75
AAAF	Illumina-HiC		2x150	12,366,688	0.74
AAAG	Illumina-HiC		2x150	415,779,674	25.04
AAAH	Illumina-HiC		2x150	21,117,210	1.27
Hi-C Total				**779,754,566**	**43.86**

Supplementary Table 13

PACBIO library statistics for single pass yield of the 51 chips included in the *Arachis hypogaea* genome assembly and their respective assembled sequence coverage levels

Cutoff	Number of Reads	Basepairs	Average Read Length	Coverage
0	17,747,784	207,196,434,170	9,784	76.74x
1000	16,984,792	206,776,022,418	10,262	76.58x
2000	16,068,940	205,392,960,654	10,854	76.07x
3000	15,083,004	202,928,022,234	11,523	75.16x
4000	14,105,835	199,510,501,164	12,217	73.89x
5000	13,153,838	195,229,606,670	12,928	72.31x
6000	12,219,149	190,089,983,704	13,657	70.40x
7000	11,296,495	184,094,072,448	14,407	68.18x
8000	10,394,841	177,334,261,665	15,169	65.68x
9000	9,524,492	169,939,908,810	15,950	62.94x
10000	8,699,996	162,111,554,928	16,738	60.04x
11000	7,925,503	153,984,133,737	17,525	57.03x
12000	7,203,695	145,687,893,797	18,309	53.96x
13000	6,530,054	137,271,980,668	19,096	50.84x
14000	5,896,752	128,725,684,685	19,896	47.68x
15000	5,295,774	120,014,218,754	20,716	44.45x
16000	4,735,569	111,335,189,066	21,553	41.24x
17000	4,219,000	102,815,474,357	22,401	38.08x
18000	3,741,661	94,465,942,929	23,266	34.99x
19000	25,272,751	808,466,536,231	24,147	32.00x

Supplementary Table 14

Summary statistics of the raw output of the MECAT whole genome shotgun assembly. The table shows total contigs and total assembled basepairs for each set of scaffolds greater than the size listed in the left hand column

Minimum Scaffold Length	Number of Scaffolds	Number of Contigs	Scaffold Size	Basepairs	% Non-gap Basepairs
5 Mb	23	23	150,942,560	150,942,560	100.00%
2.5 Mb	97	97	404,936,510	404,936,510	100.00%
1 Mb	501	502	1,001,909,947	1,001,909,795	100.00%
500 Kb	1,247	1,249	1,512,793,957	1,512,793,073	100.00%
250 Kb	2,591	2,595	1,988,505,988	1,988,499,990	100.00%
100 Kb	4,778	4,783	2,349,231,029	2,349,223,361	100.00%
50 Kb	6,252	6,258	2,456,140,461	2,456,132,755	100.00%
25 Kb	7,303	7,309	2,495,638,348	2,495,630,642	100.00%
10 Kb	7,662	7,668	2,502,304,643	2,502,296,937	100.00%
5 Kb	7,692	7,698	2,502,553,973	2,502,546,267	100.00%
2.5 Kb	7,692	7,698	2,502,553,973	2,502,546,267	100.00%
1 Kb	7,692	7,698	2,502,553,973	2,502,546,267	100.00%
0 bp	7,692	7,698	2,502,553,973	2,502,546,267	100.00%

Supplementary Table 15

Summary of the tetrasomic regions identified and duplicated in the final release assembly

Homeologous pair	Size of Region	Duplicated From	Added to
01/11	N/A		
02/12	2,322,927	bottom of Arahy.12	bottom of Arahy.02
03/13	342,613	bottom of Arahy.03	bottom of Arahy.13
04/14	2,767,012	bottom of Arahy.14	bottom of Arahy.04
05/15	6,264,747	top of Arahy.05	top of Arahy.15
06/16	2,501,280	bottom of Arahy.06	bottom of Arahy.16
07/18	579,447	bottom of Arahy.07	top of Arahy.18
09/19	193,833	top of Arahy.09	top of Arahy.19
10/20	N/A		

Supplementary Table 16
Summary of sequences used for transcript assembly

	Tifrunner	Florida-07	Gregory	NC 3033	C76-16	A72	Reference	# Cleaned reads
Vegetative and reproductive development	X						Clevenger et al. 2016. Front. Plant Sci.	1.3E+09
Preharvest aflatoxin	X	X		X	X	X	Clevenger et al. 2016. Toxins	6.9E+08
Postharvest aflatoxin		X					Korani et al. 2018. Genetics	8.6E+08
Nematode			X				Clevenger et al. 2017. Sci. Rep.	6.5E+07
Late leaf spot (0,2,4,5,8,20d post-inoculation)		X					unpublished	2.9E+08
Nodulation (0,6,10,15,21d post-inoculation)	X						unpublished	7.7E+08
Seed and pericarp development (R3,R4-5,R6,R7 stages)	X			X			unpublished	1.9E+09
								TOTAL 5.9E+09

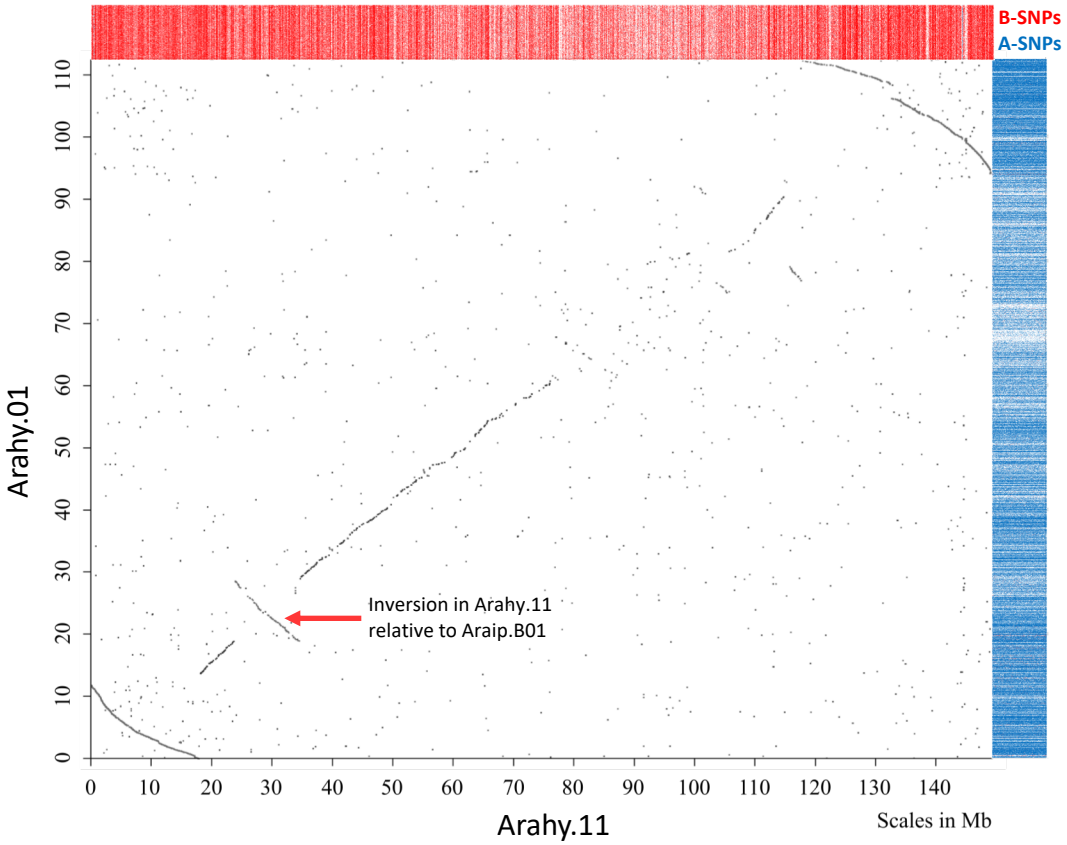
Supplementary Figures for

The genome sequence of segmental allotetraploid peanut *Arachis hypogaea*

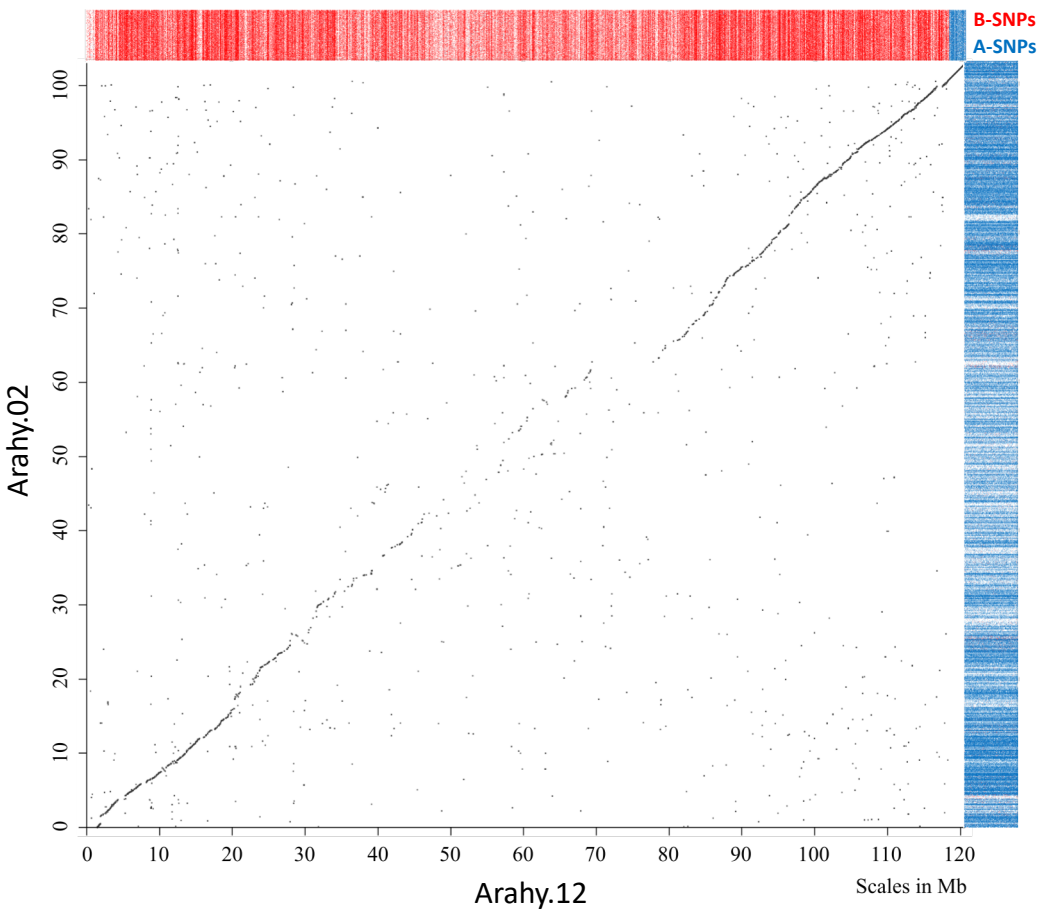
Overview legend for Supplementary Figures 1-12

Figures show a dot plot for each homeologous pair of chromosomes from the A and B subgenomes of *Arachis hypogaea* cv. Tifrunner. Colored bars along the axes indicate the genome compositions. Bars are colored with dots indicating assignable ancestral single nucleotide polymorphisms (SNPs), red for ancestral B genome and blue for ancestral A genome.

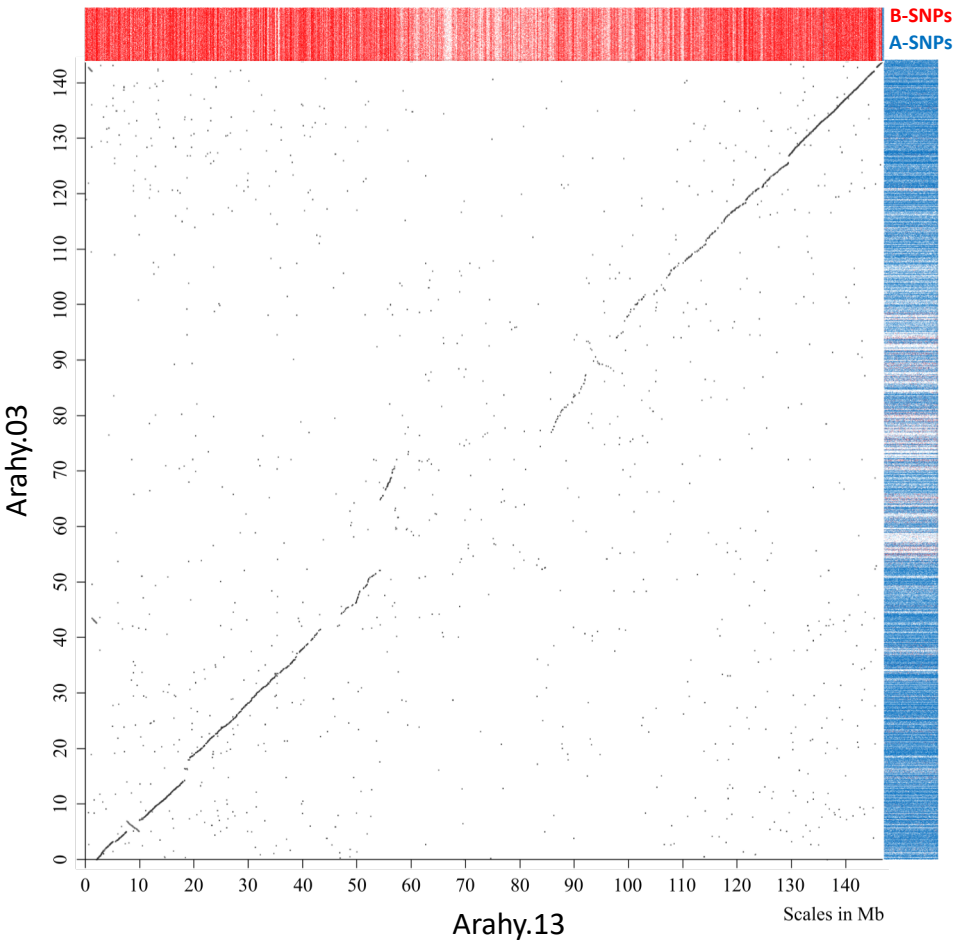
Supplementary Figure 1
Comparison and genome compositions of Arahy.01 *vs.* Arahy.11



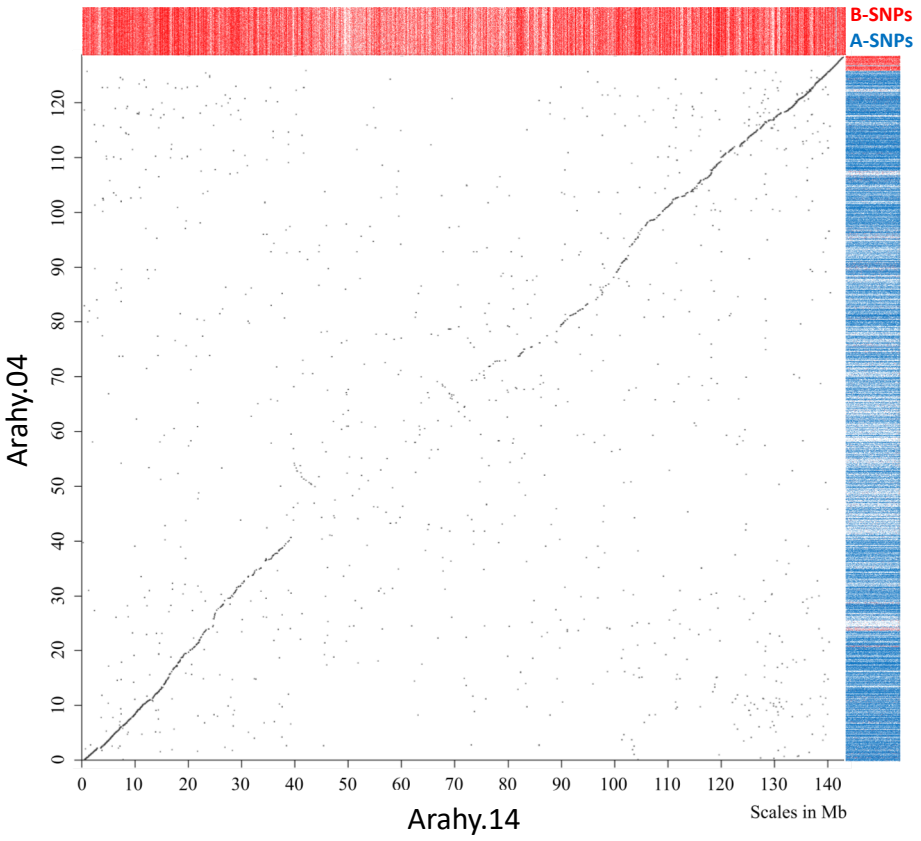
Supplementary Figure 2
Comparison and genome compositions of Arahy.02 *vs.* Arahy.12



Supplementary Figure 3
Comparison and genome compositions of Arahy.03 *vs.* Arahy.13

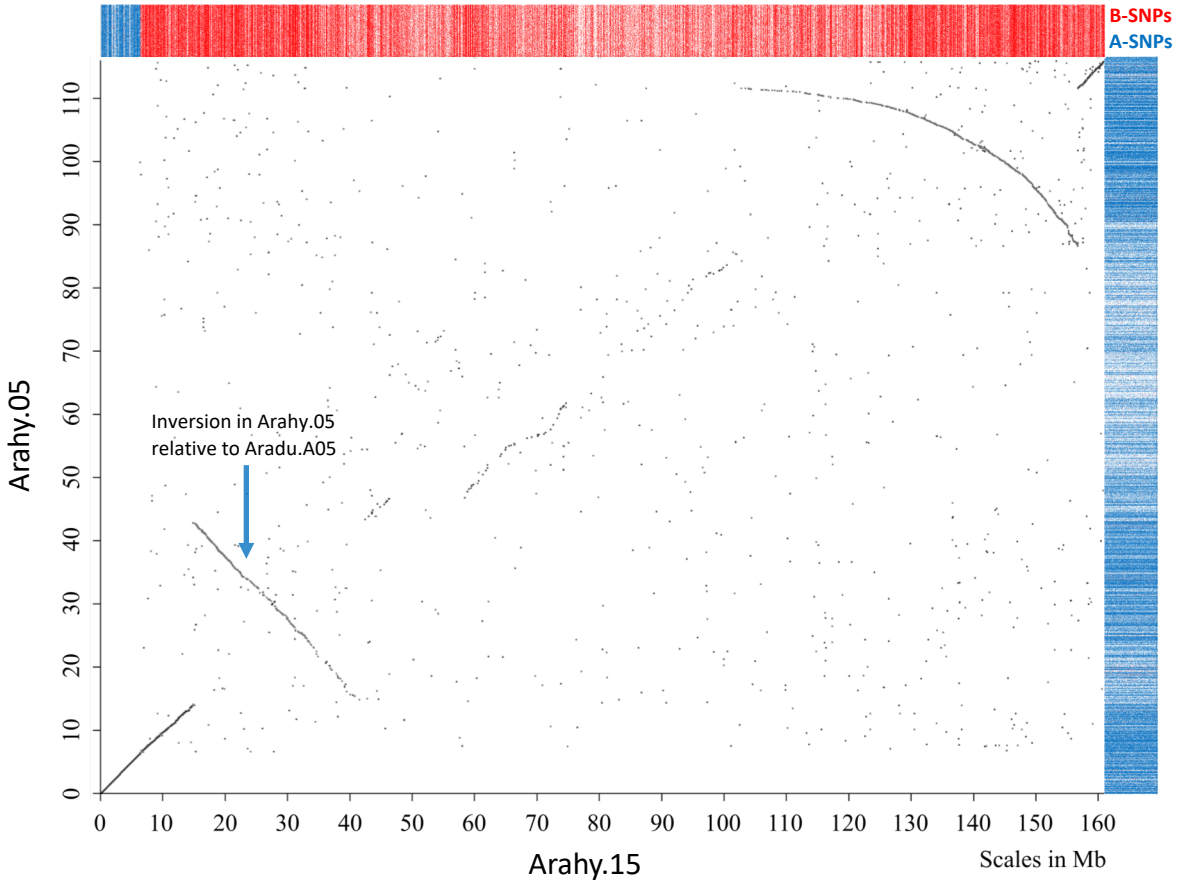


Supplementary Figure 4
Comparison and genome compositions of Arahy.04 *vs.* Arahy.14



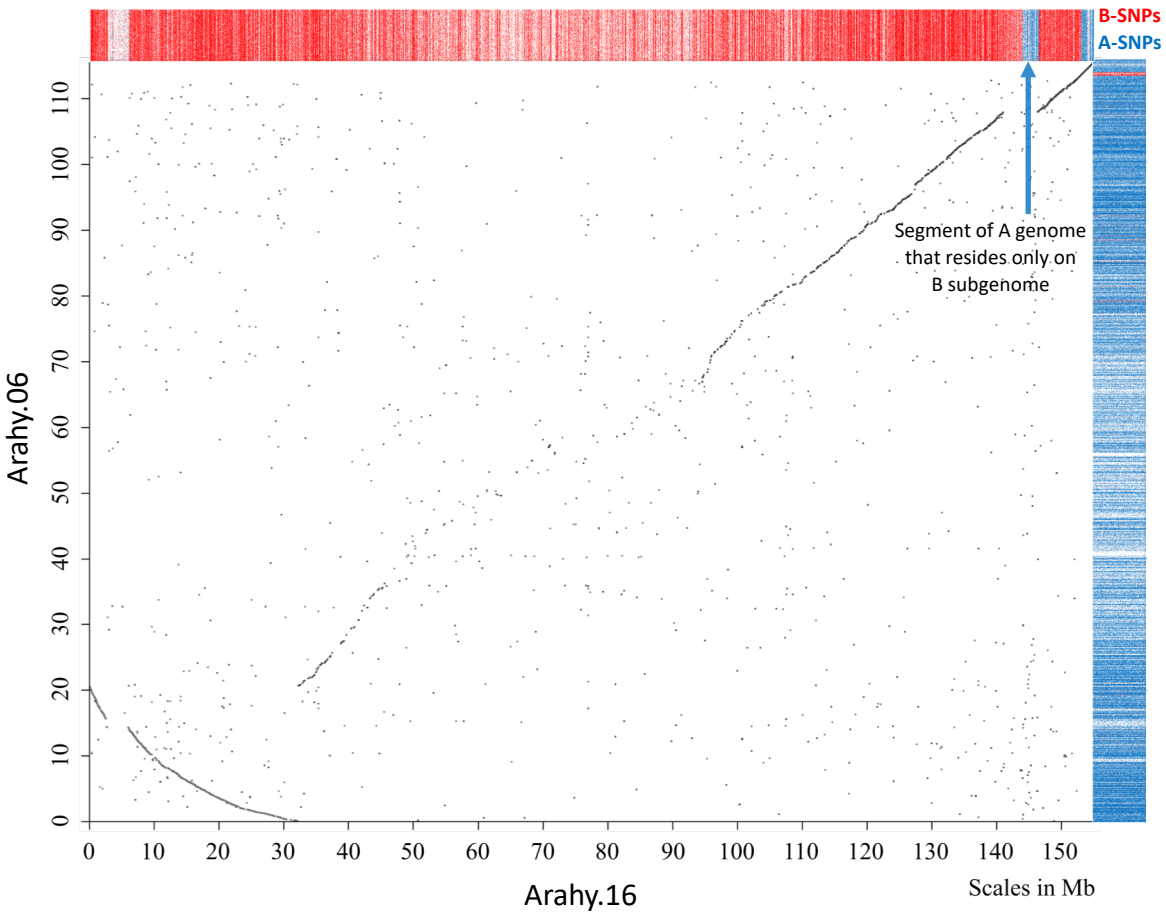
Supplementary Figure 5

Comparison and genome compositions of Arahy.05 *vs.* Arahy.15

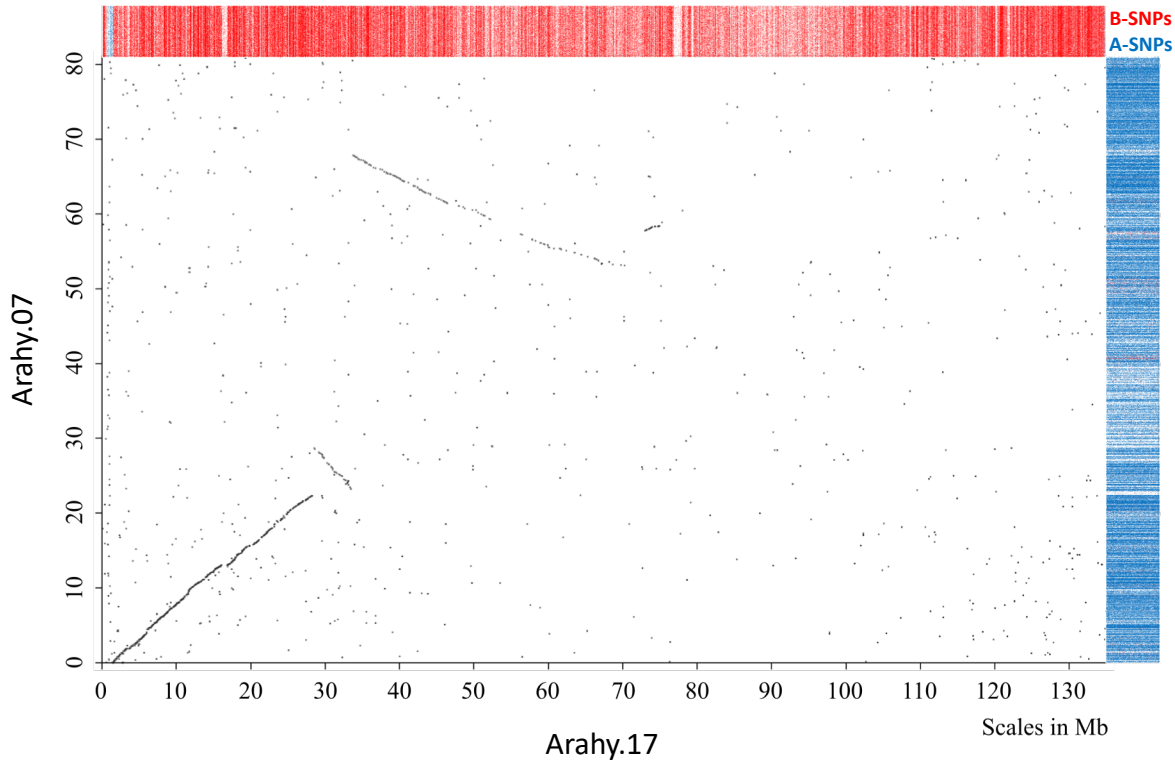


Supplementary Figure 6

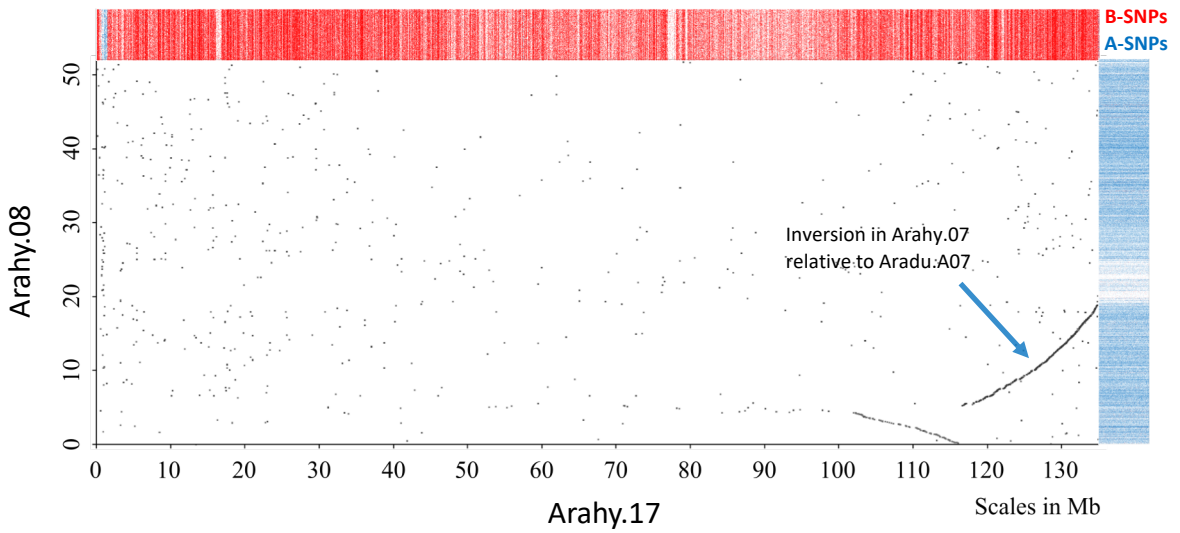
Comparison and genome compositions of Arahy.06 *vs.* Arahy.16



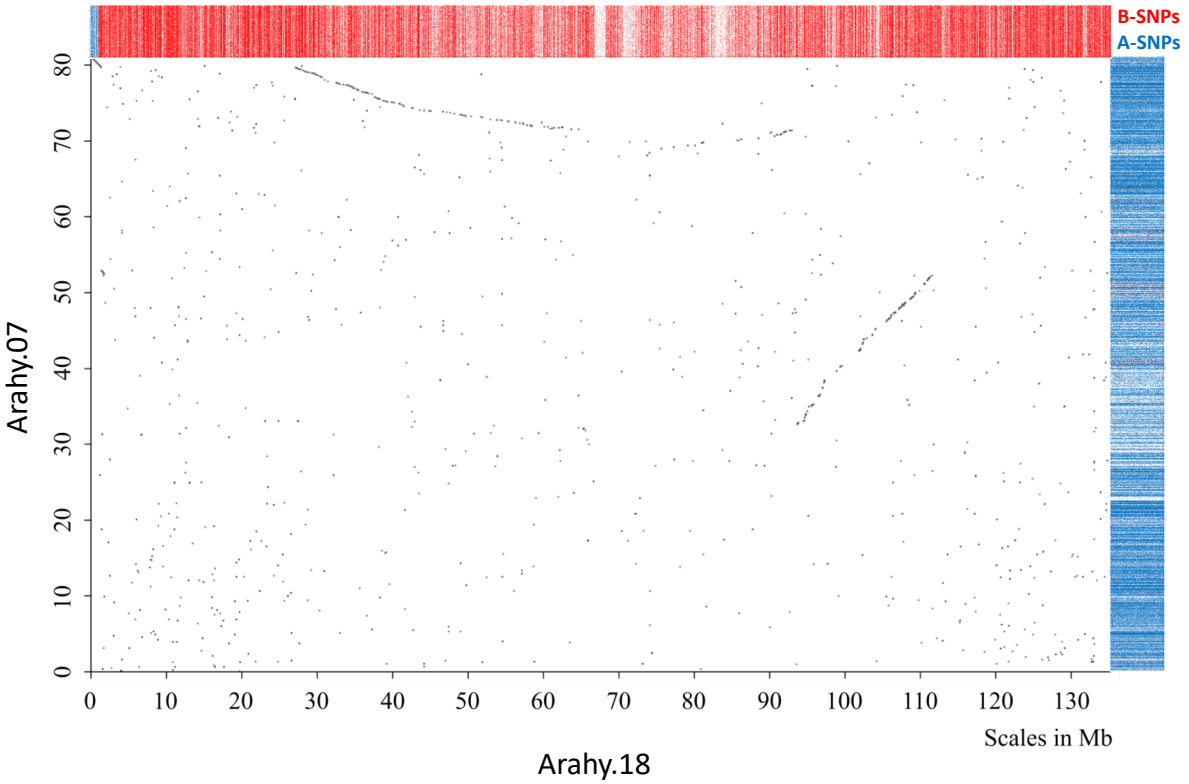
Supplementary Figure 7
Comparison and genome compositions of Arahy.07 *vs.* Arahy.17



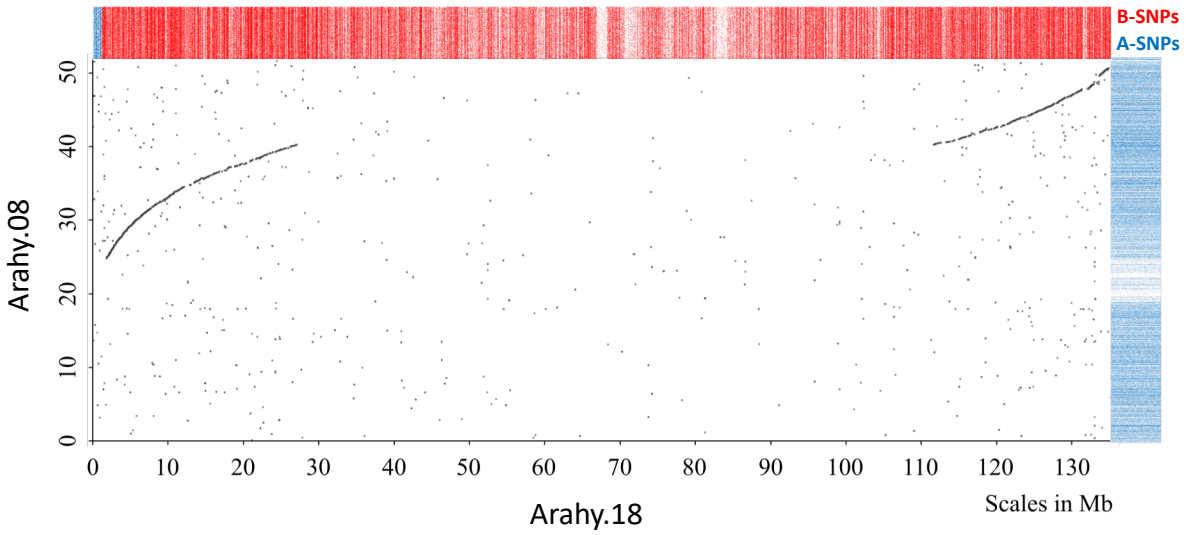
Supplementary Figure 8
Comparison and genome compositions of Arahy.08 *vs.* Arahy.17



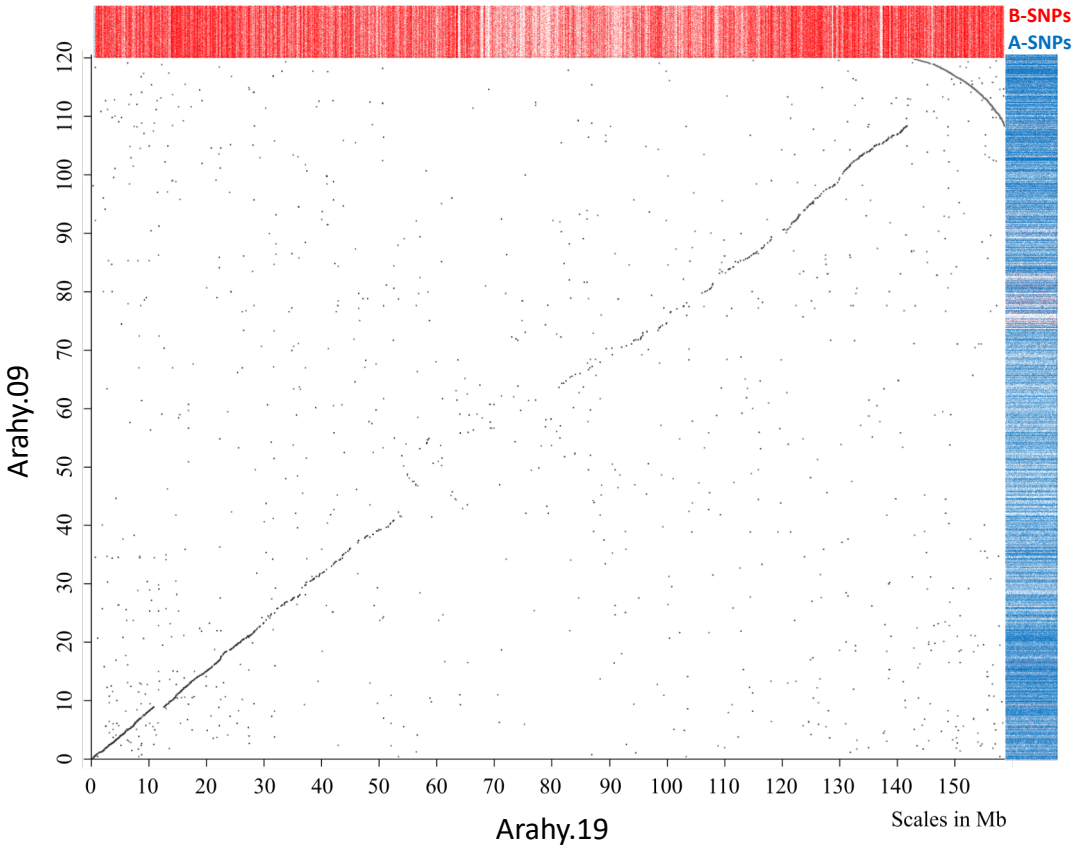
Supplementary Figure 9
Comparison and genome compositions of Arahy.07 *vs.* Arahy.18



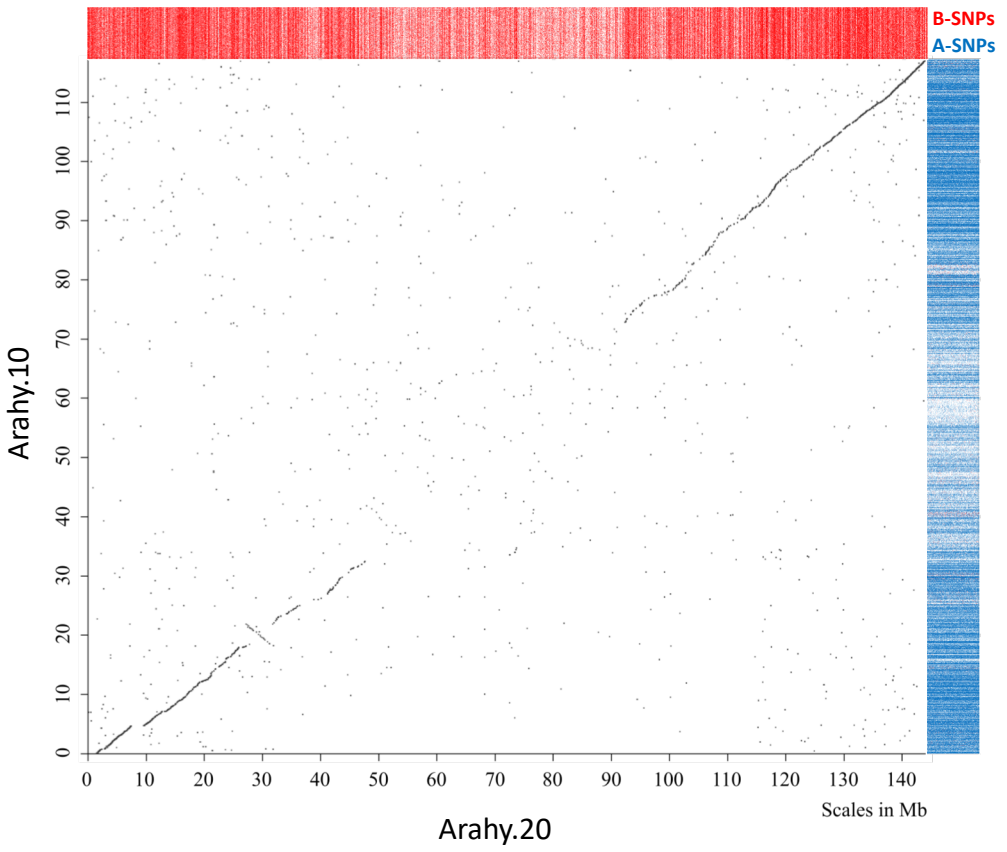
Supplementary Figure 10
Comparison and genome compositions of Arahy.08 *vs.* Arahy.18



Supplementary Figure 11
Comparison and genome compositions of Arahy.09 *vs.* Arahy.19



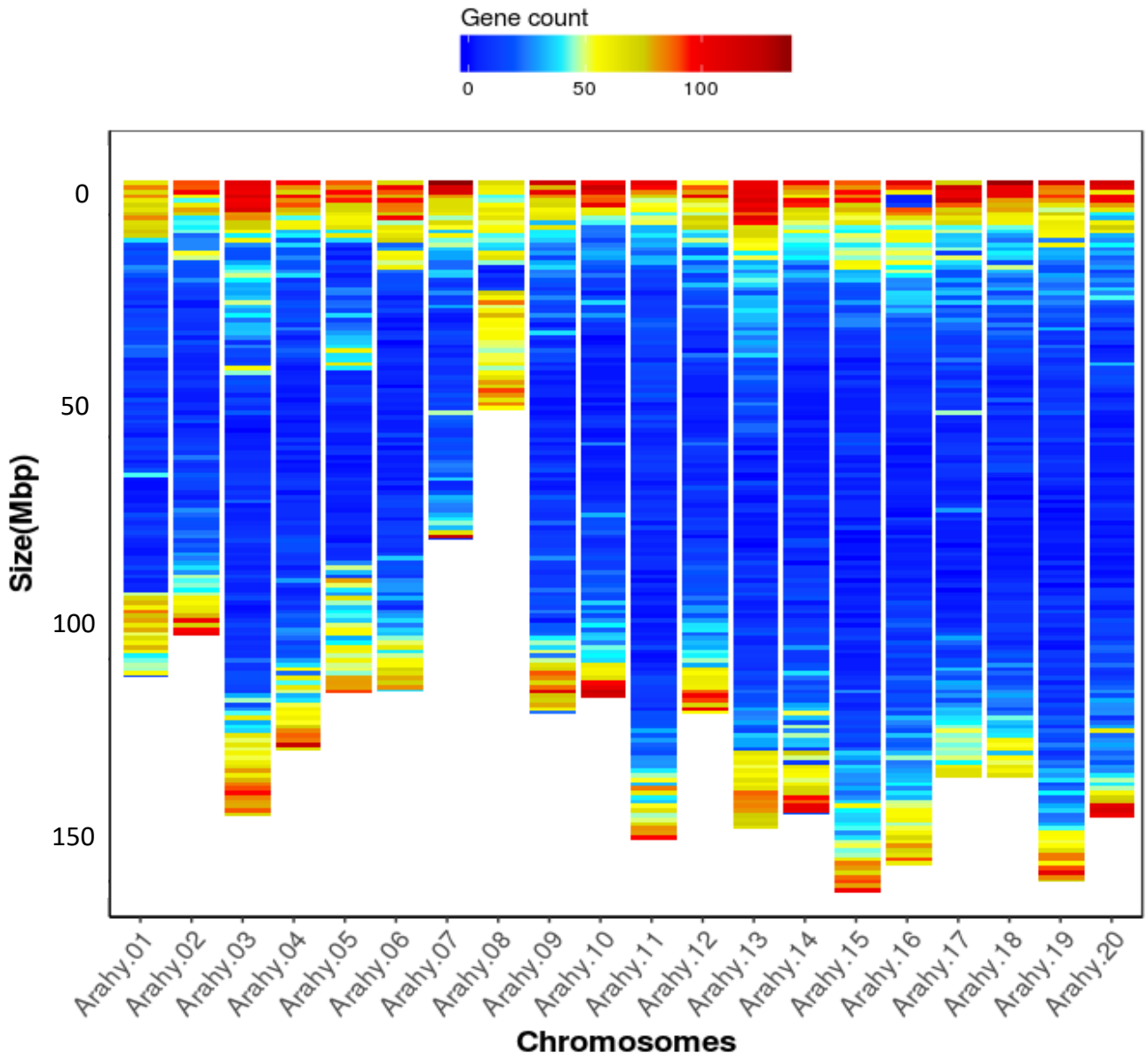
Supplementary Figure 12
Comparison and genome compositions of Arahy.10 *vs.* Arahy.20



Supplemental Figure 13

Heatmap of gene density along the 20 *Arachis hypogaea* cv. Tifrunner chromosomal pseudomolecules.

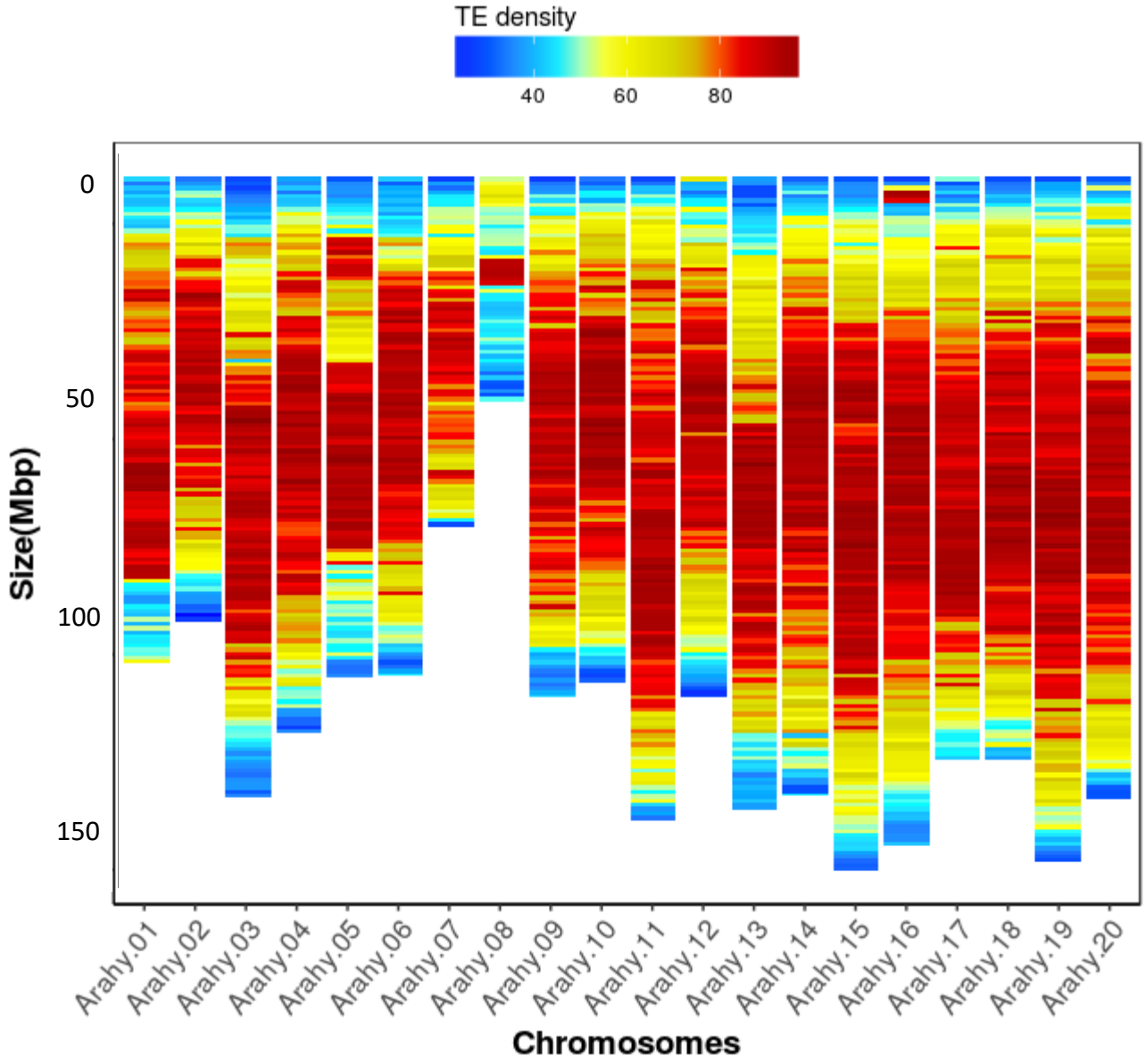
The gradient color corresponds to the number of genes within 1 Mb windows



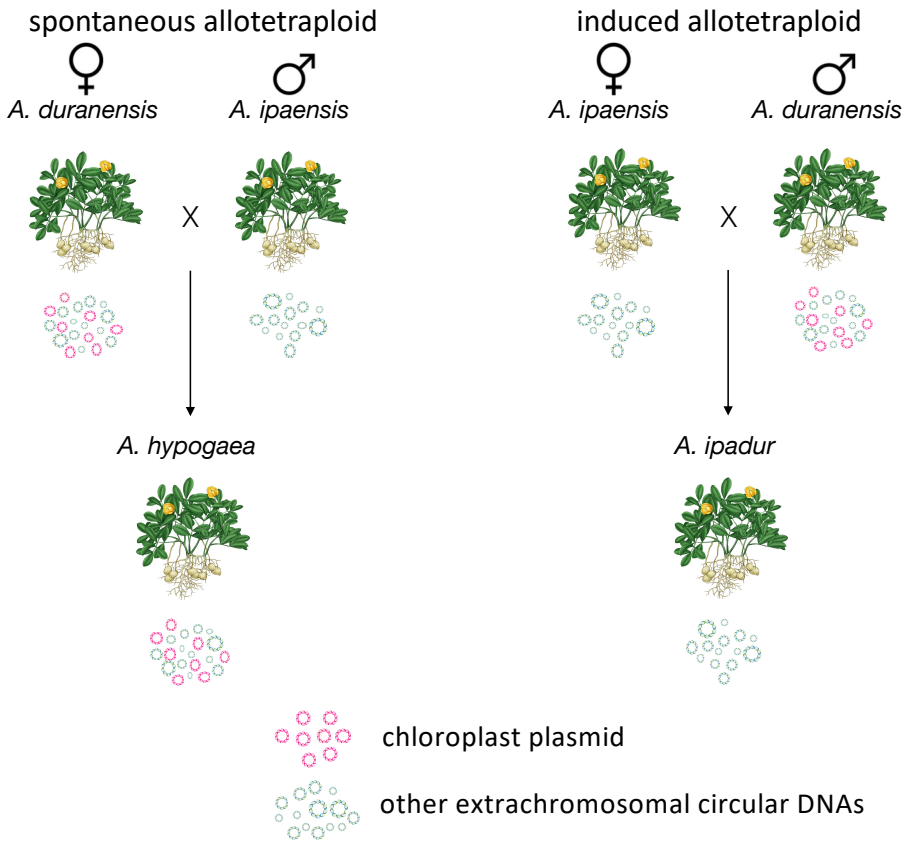
Supplemental Figure 14

Heatmap of TE density along the 20 *Arachis hypogaea* cv. Tifrunner chromosomal pseudomolecules.

The gradient color corresponds to the percentage coverage of TEs within 200 kbp windows

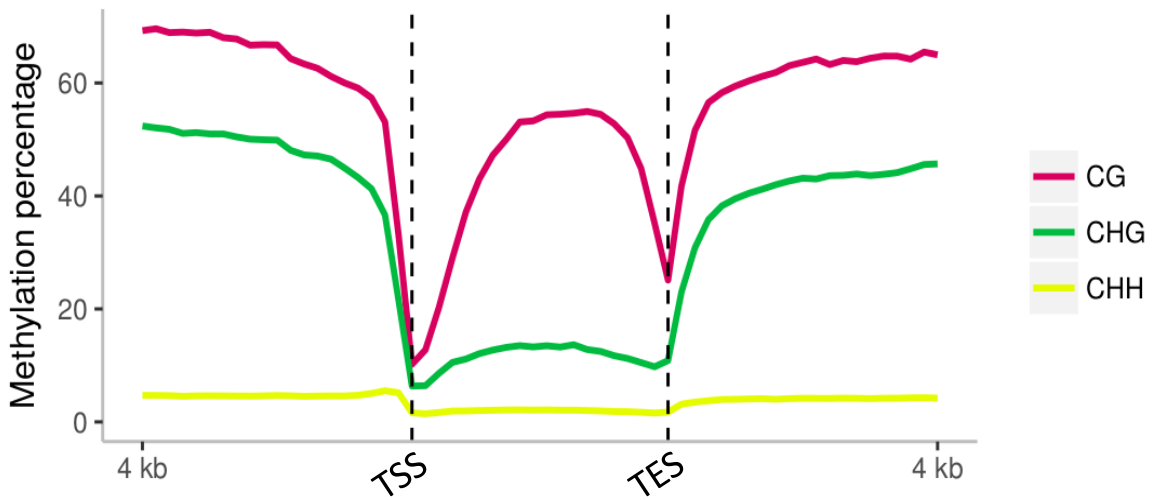


Supplementary Figure 15
inheritance of chloroplast plasmid



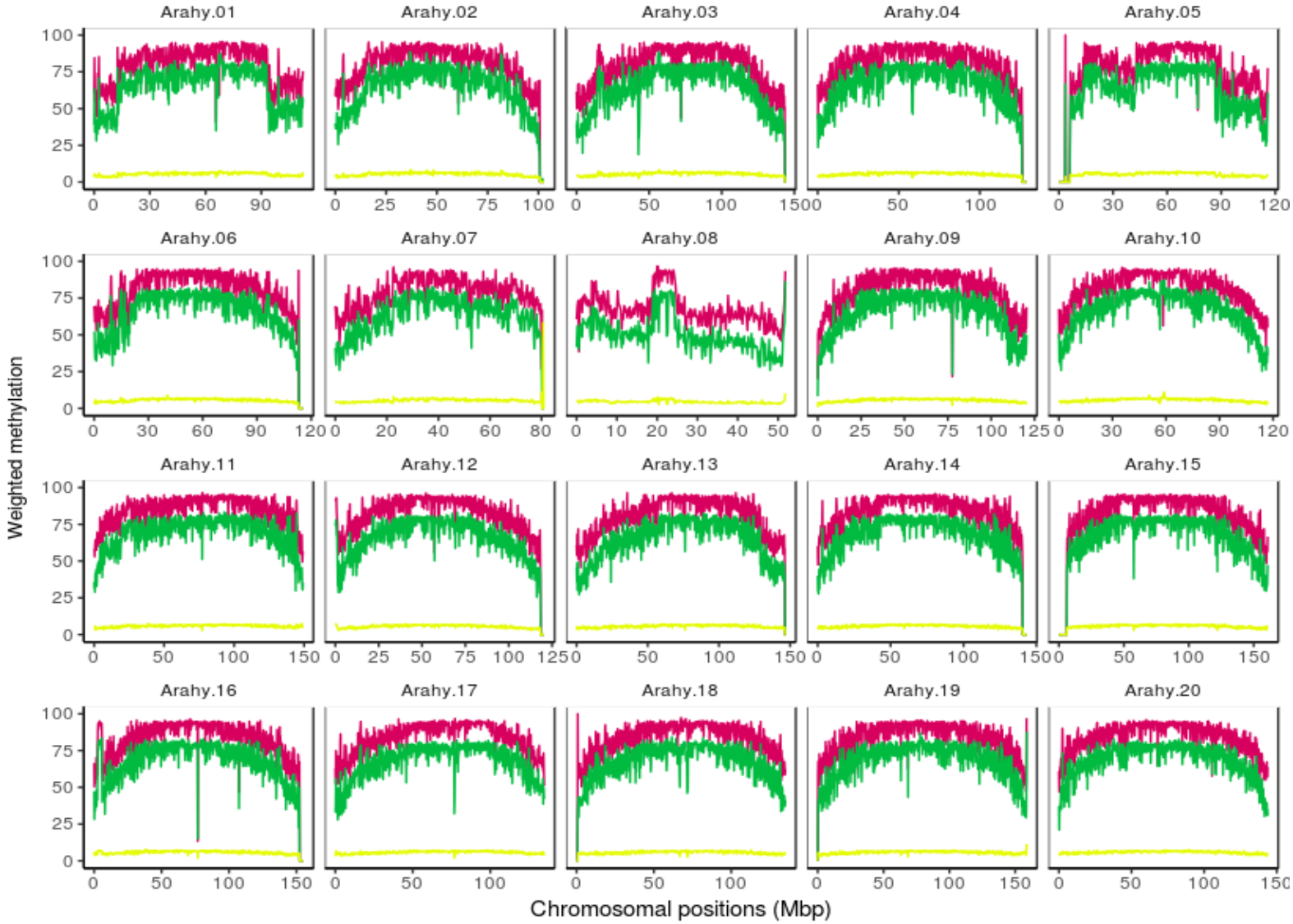
Supplementary Figure 16
Methylation in and near genes

Metaplots showing methylation for ~15,000 genes with annotated 5' and 3' untranslated regions (UTRs) and flanking regions at CG (pink), CHG (green) and CHH (yellow) sites (where H is a non G base). Scale on x-axis is in bp. Genes show typical patterns of methylation observed in other species: mCG, mCHG, and mCHH methylation levels are higher upstream and downstream of gene bodies. mCG methylation decreases considerably around transcription start sites (TSS), increases across the gene-bodies, and decreases at transcription termination sites (TES), whereas mCHG and mCHH are low across entire gene-bodies

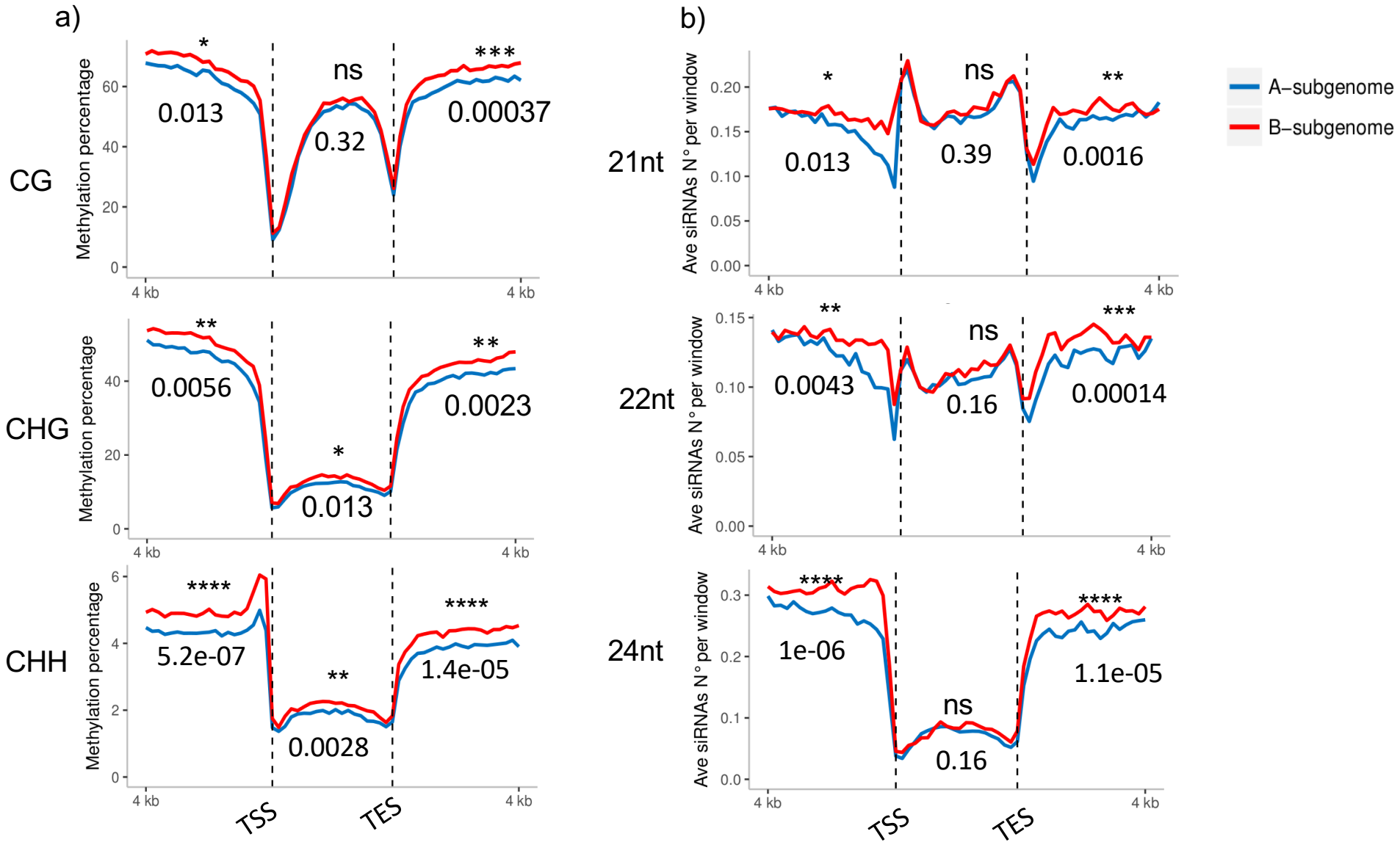


Supplementary Figure 17

Methylation percentage in chromosomal pseudomolecules of *A. hypogaea* at CG (in fuchsia), CHG (in green) and CHH (in yellow) sites. Values where calculated in 200 kb windows.



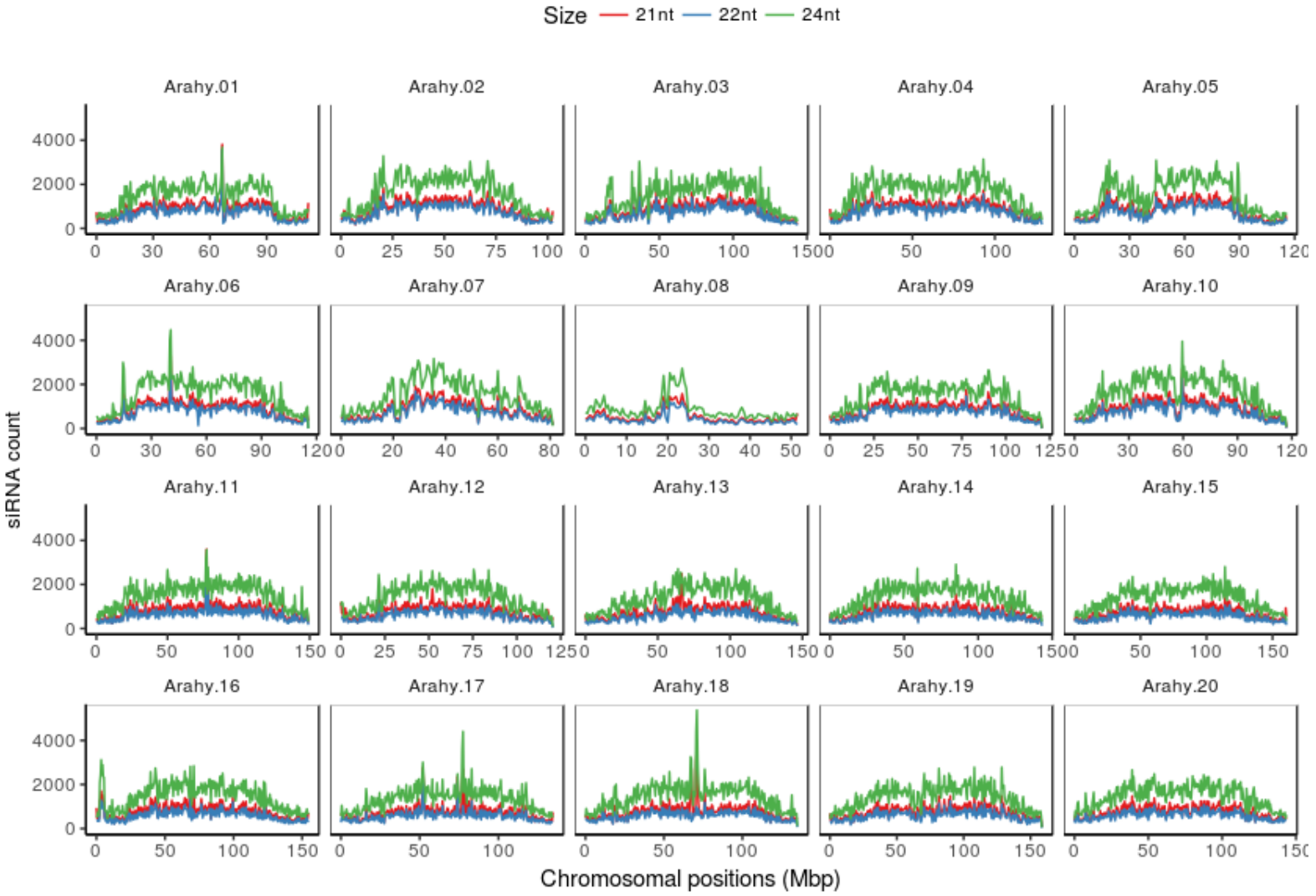
Supplemental Figure 18: Metaplots of gene bodies and flanking regions in A and B-subgenomes comparing (a) methylation percentage (CG, CHG and CHH) and (b) siRNAs count (21, 22 and 24 nt). A homeologs (blue), B homeologs (red). TSS : transcriptional start site, TES : transcriptional end site. the p-values for significance of difference correspond to Wilcoxon rank-sum statistical test. ns: not significant, * : P ≤ 0.05, ** : P ≤ 0.01, *** P : ≤ 0.001, **** : P ≤ 0.0001.



Supplemental Figure 19: densities of DNA sequences corresponding to sRNAs in chromosomal pseudomolecules of *A. hypogaea* cv. Tifrunner

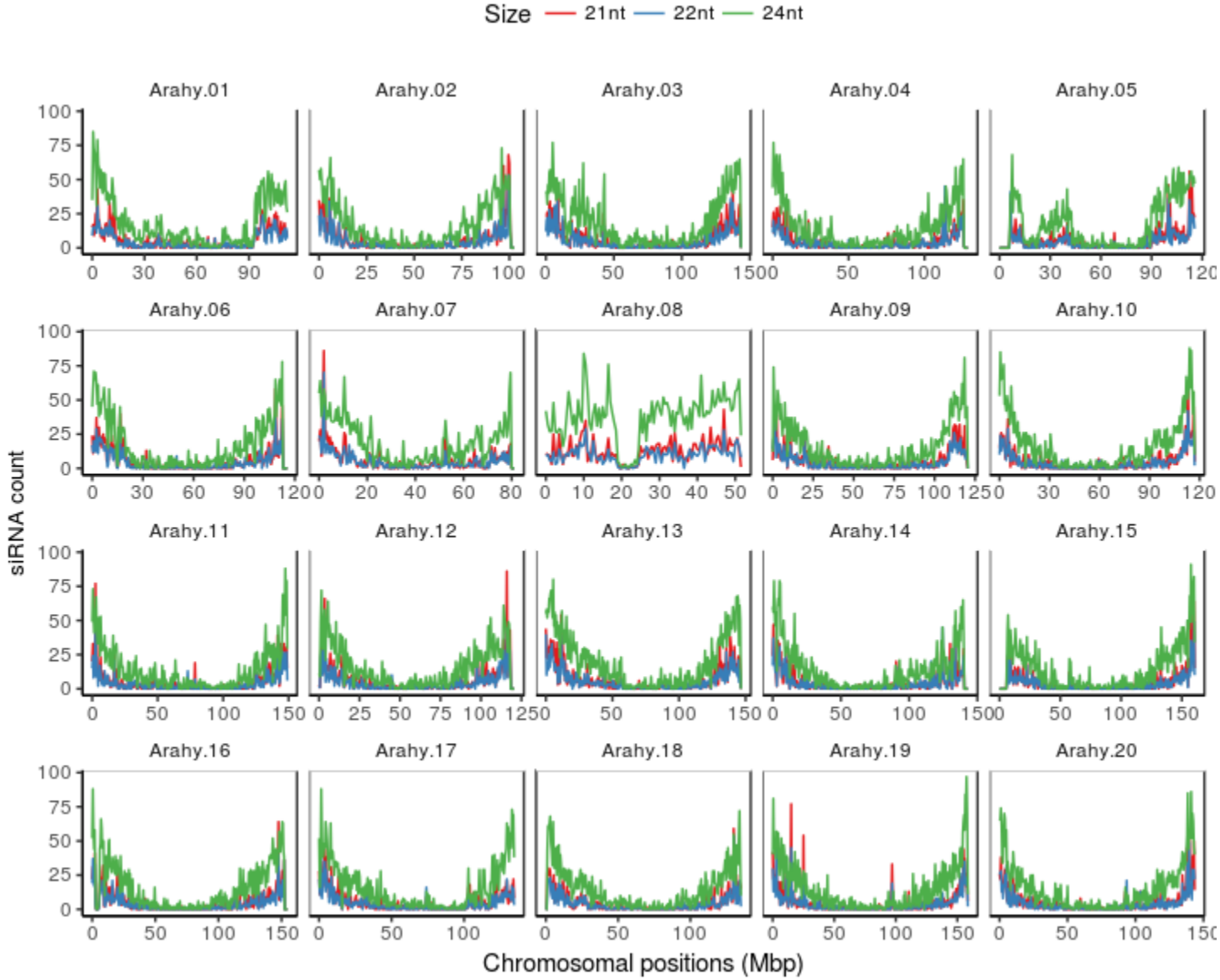
21nt (in red), 22nt (in green) and 24nt (in blue) length**.**

Values where calculated in 200 kb windows.



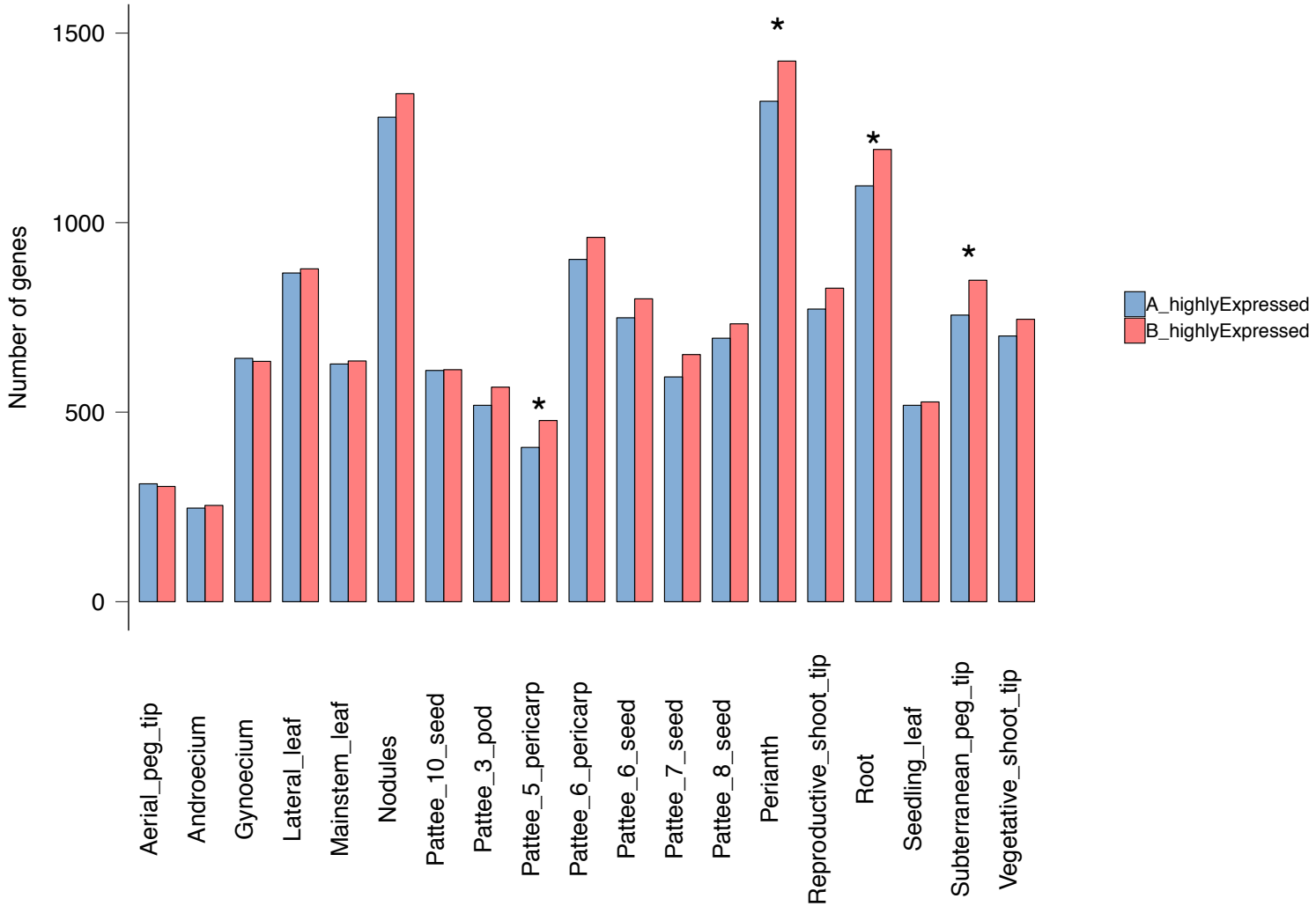
Supplemental Figure 20: densities of DNA sequences corresponding to uniquely mapping sRNAs in chromosomal pseudomolecules of *A. hypogaea* cv. Tifrunner.

21nt (in red), 22nt (in green) and 24nt (in blue) length. Values where calculated in 200 kb windows.



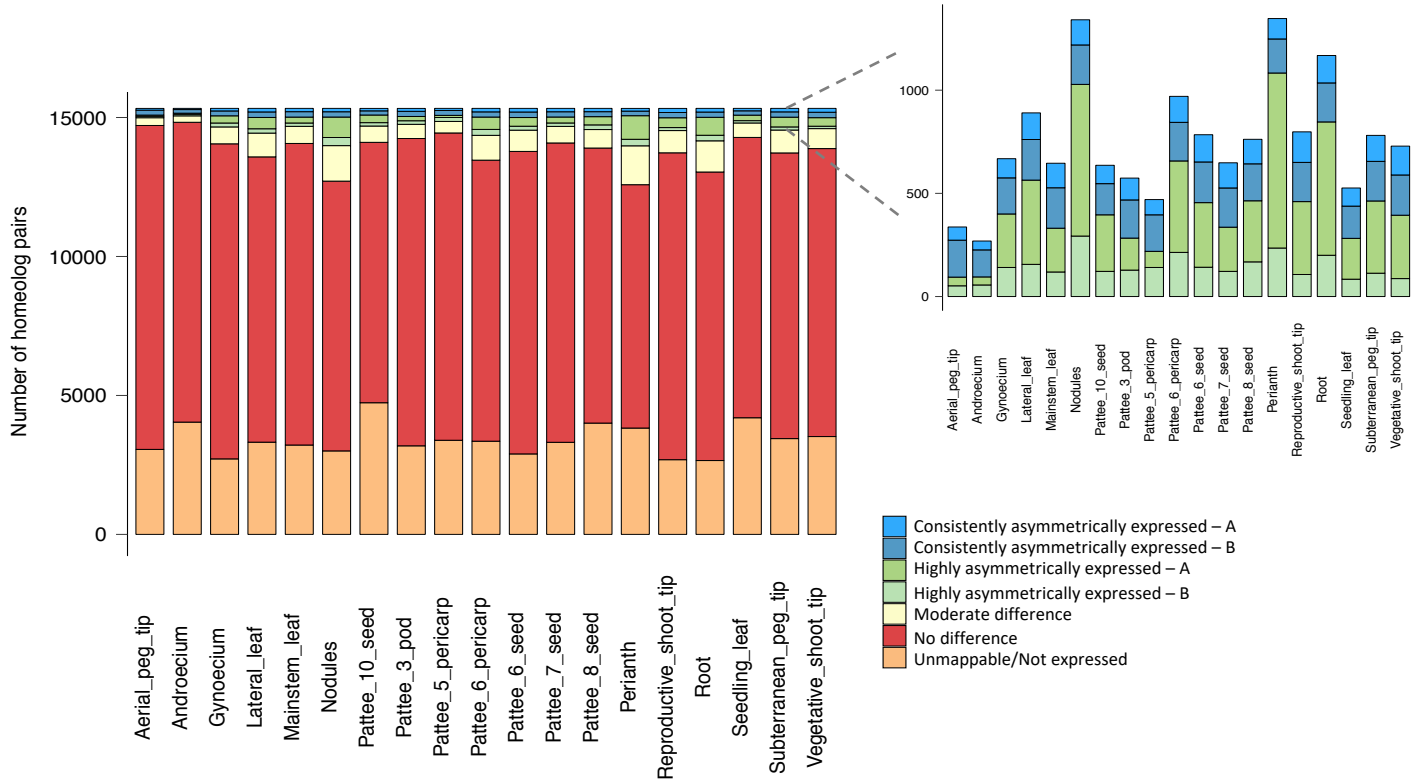
Supplementary Figure 21

Homeolog expression differences in *A. hypogaea* cv. Tifrunner. Homeologs were compared to check for differences in gene expression levels in different tissues and developmental stages of peanut pods. The number of homeologous genes more highly expressed ($\log_2$ fold change ≥ 1, Benjamini-Hochberg adjusted $P < 0.05$; Wald test) in each subgenome is represented. *P*-value correspond to binomial test with the odds of A genes being more highly expressed at 0.5 probability. * : $P < 0.05$, others : not significant.



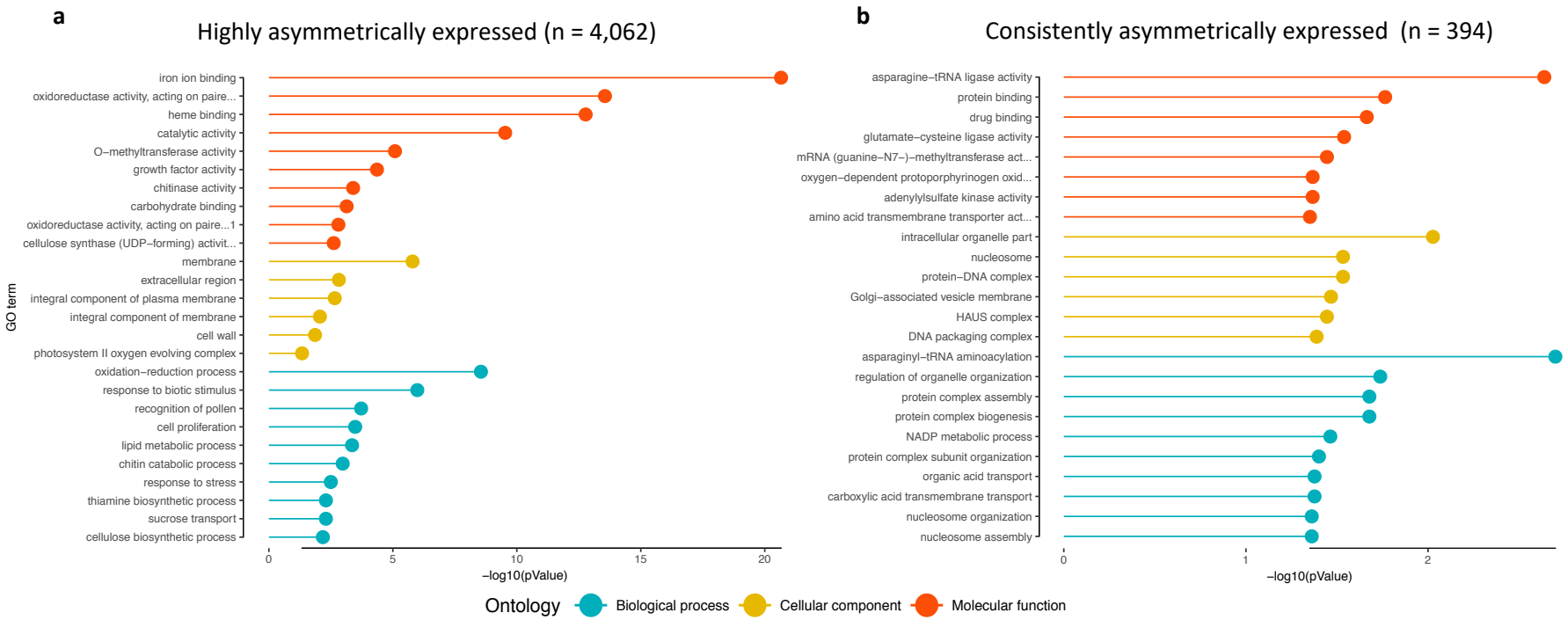
Supplementary Figure 22

Classification of homeologous pairs by expression patterns. Differentially expressed pairs (n = 15,328) in tissues and developmental stages of pods in *A. hypogaea* were classified into five categories: i) highly asymmetrically expressed, $\log_2$ expression ratios > 3; ii) consistently asymmetrically expressed, same expression pattern in at least half of the tissues; iii) moderate difference $\log_2$expression ratios ≥ 1 and < 3; iv) no difference; and v) unmappable/not expressed. The inset plot shows the distribution of highly and consistently asymmetrically expressed pairs in each subgenome.



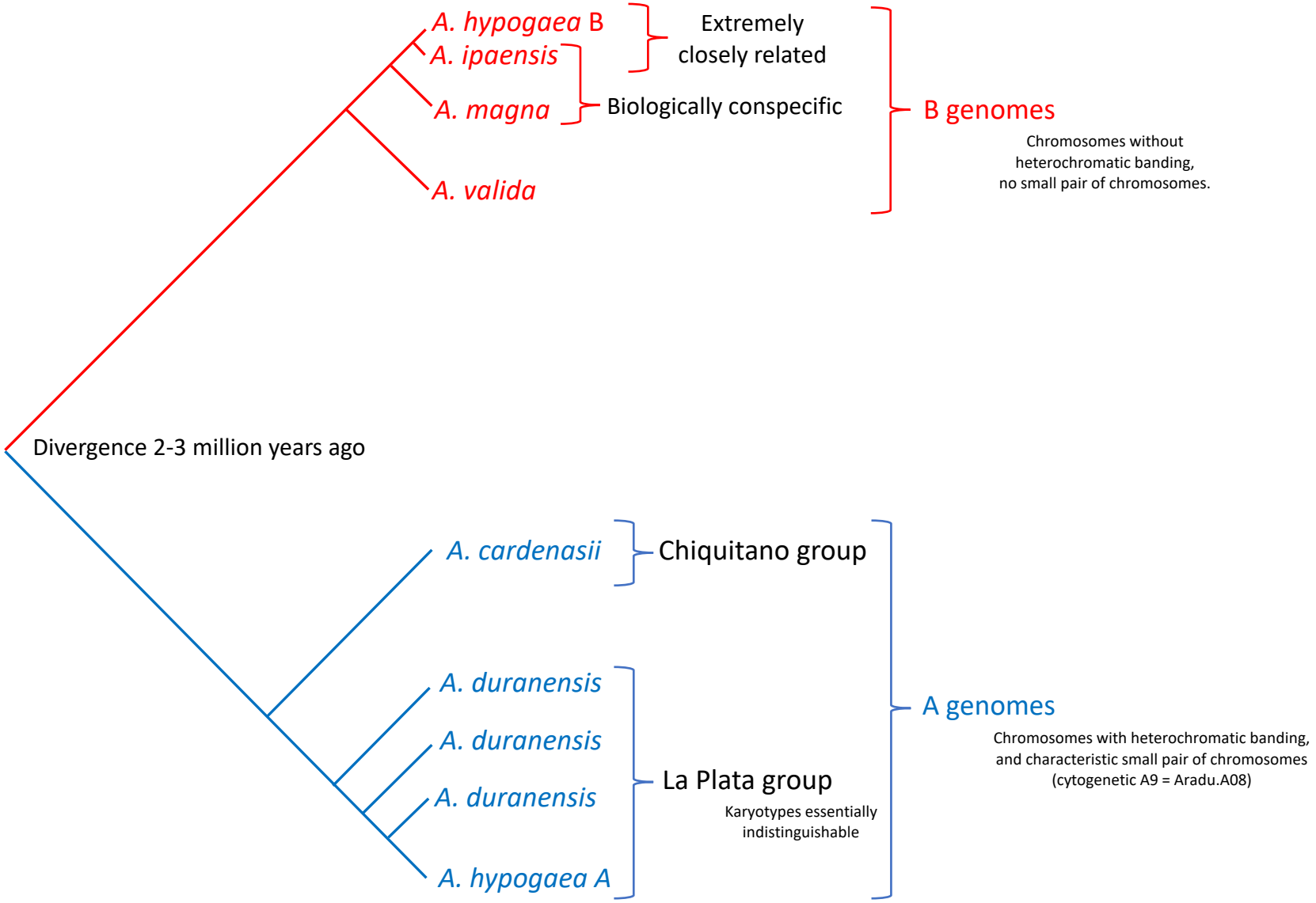
Supplementary Figure 23

Top 10 Gene Ontology terms (in each of three categories - biological process, cellular component and molecular function) enriched ($P < 0.05$, Fisher's exact test) among a) highly asymmetrically expressed and b) consistently asymmetrically expressed homeolog pairs.



Supplementary Figure 24

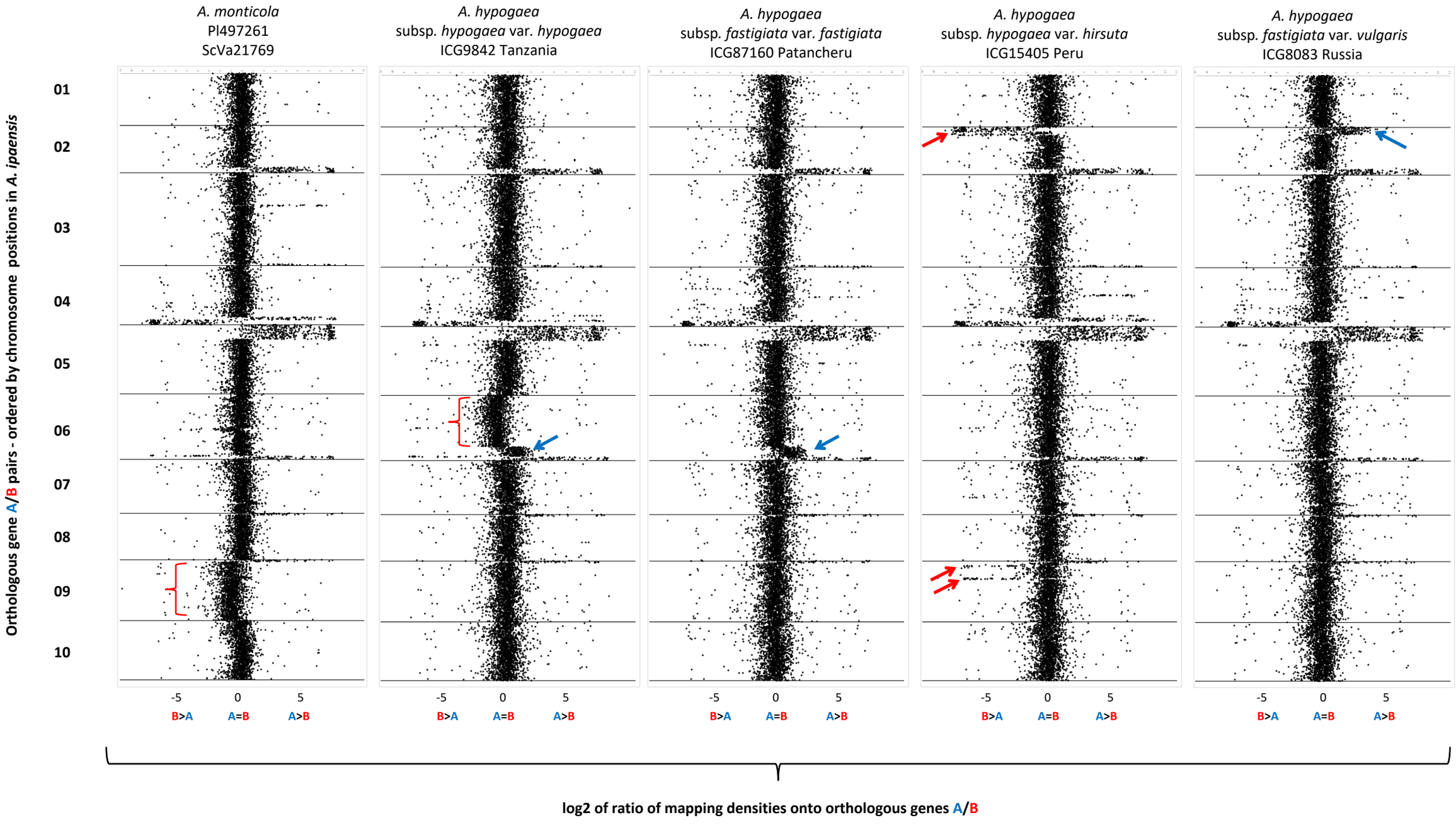
Schematic tree of relationships of species used in identifying single nucleotide polymorphisms characteristic of ancestral A and B genomes. We consider the topography of relationships extremely strongly supported: by biogeography, cross-compatibilities of species and accessions (Krapovickas et al 2007; Moretzsohn et al 2009), karyotypes (Robledo et al 2009; Robledo et al 2010) and molecular phylogenies (Moretzsohn et al 2013).



Tree not to scale.

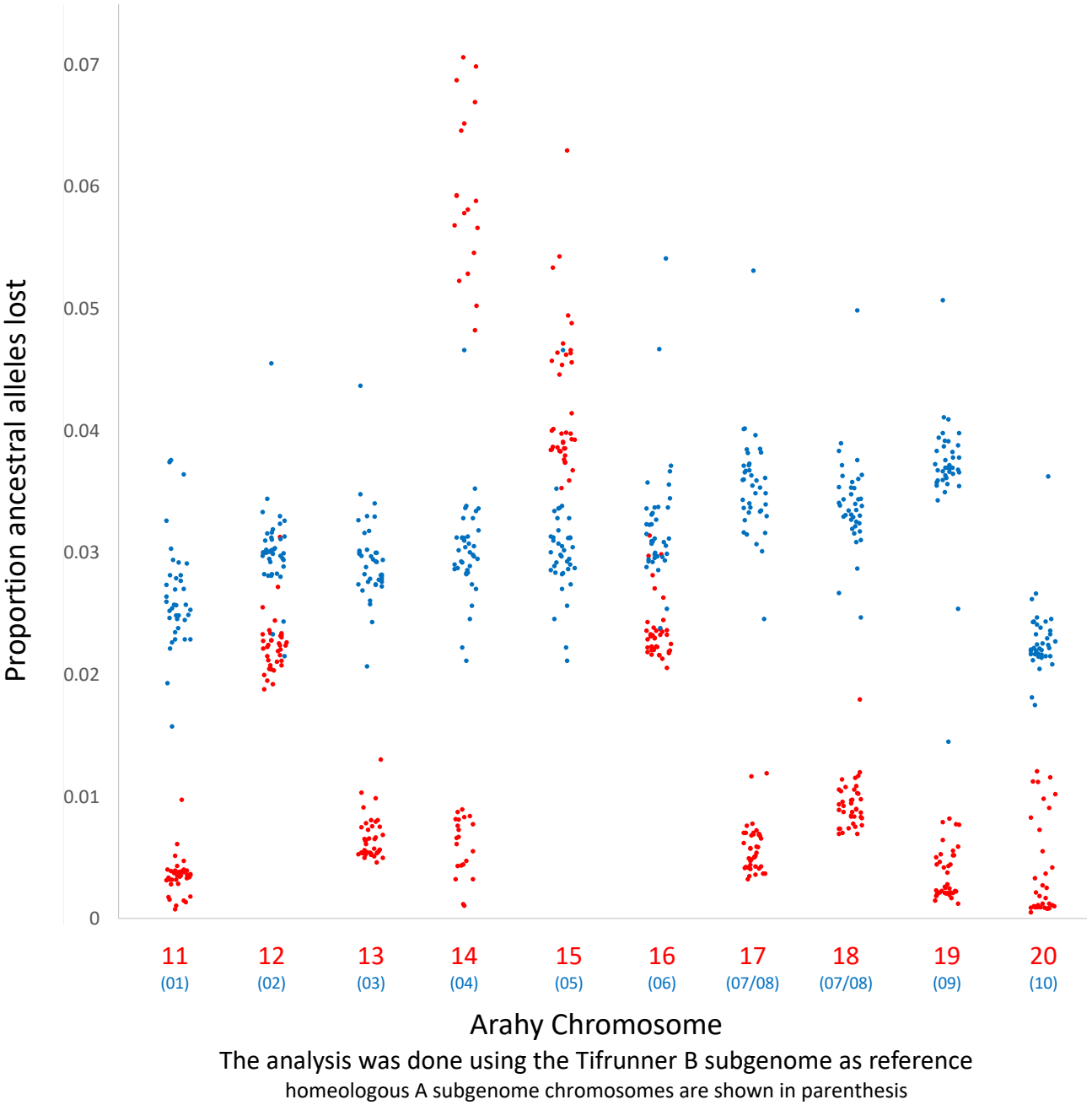
Supplementary Figure 25

This Figure is a supplement for **Fig. 2** in main manuscript. Overview of A/B genome compositions in selected *A. hypogaea* and *A. monticola* that show variations in genome structures visible on this scale and using these methods. Unusual *v*ariations are highlighted with with colored arrows and brackets. (note how the deviations in structure in the two right hand plots, at the top of A02/B02 (02/12) are in opposite directions)



Supplementary Figure 26
Ancestral allele loss in a selection of *A. monticola* and *A. hypogaea*

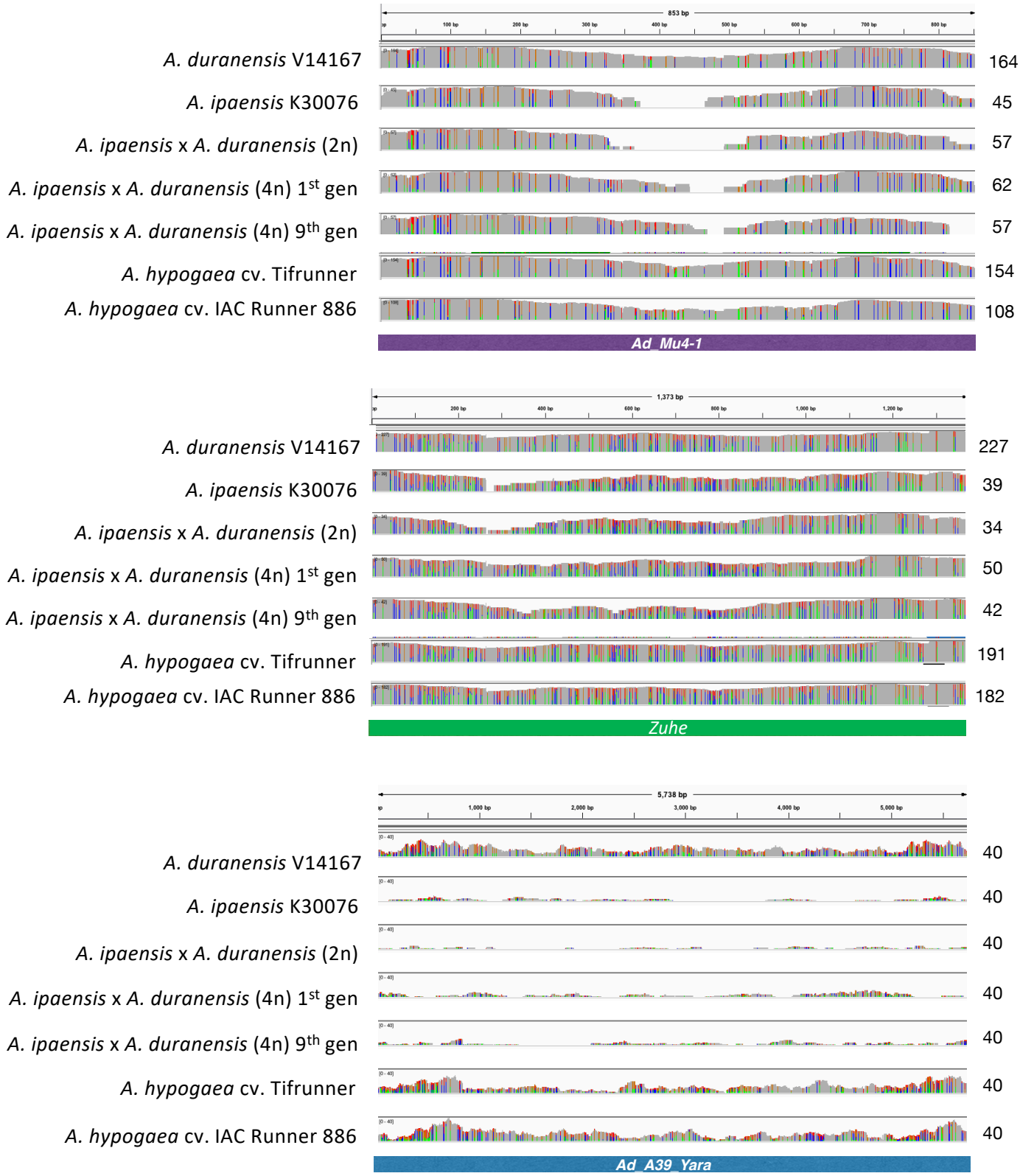
Estimated proportion of ancestral alleles lost by homeologous recombination and deletions, in one *A. monticola* and 38 *A. hypogaea* accessions. Genotypes are Tifrunner, NC3303, and the "China landraces and modern cultivars" genotypes (**Dataset 2**). Each genotype is represented by 20 dots. Illumina whole genome sequences were mapped to the B subgenome of Tifrunner; 825,960 ancestral A and B alleles were inferred from their differentiation of representatives of A and B genome diploid species (**Supplementary Fig. 24**). The raw counts that generated this graph are in **Dataset 5**, file Data-AB-SNP-counts-high-coverage.txt.



The analysis was done using the Tifrunner B subgenome as reference
homeologous A subgenome chromosomes are shown in parenthesis

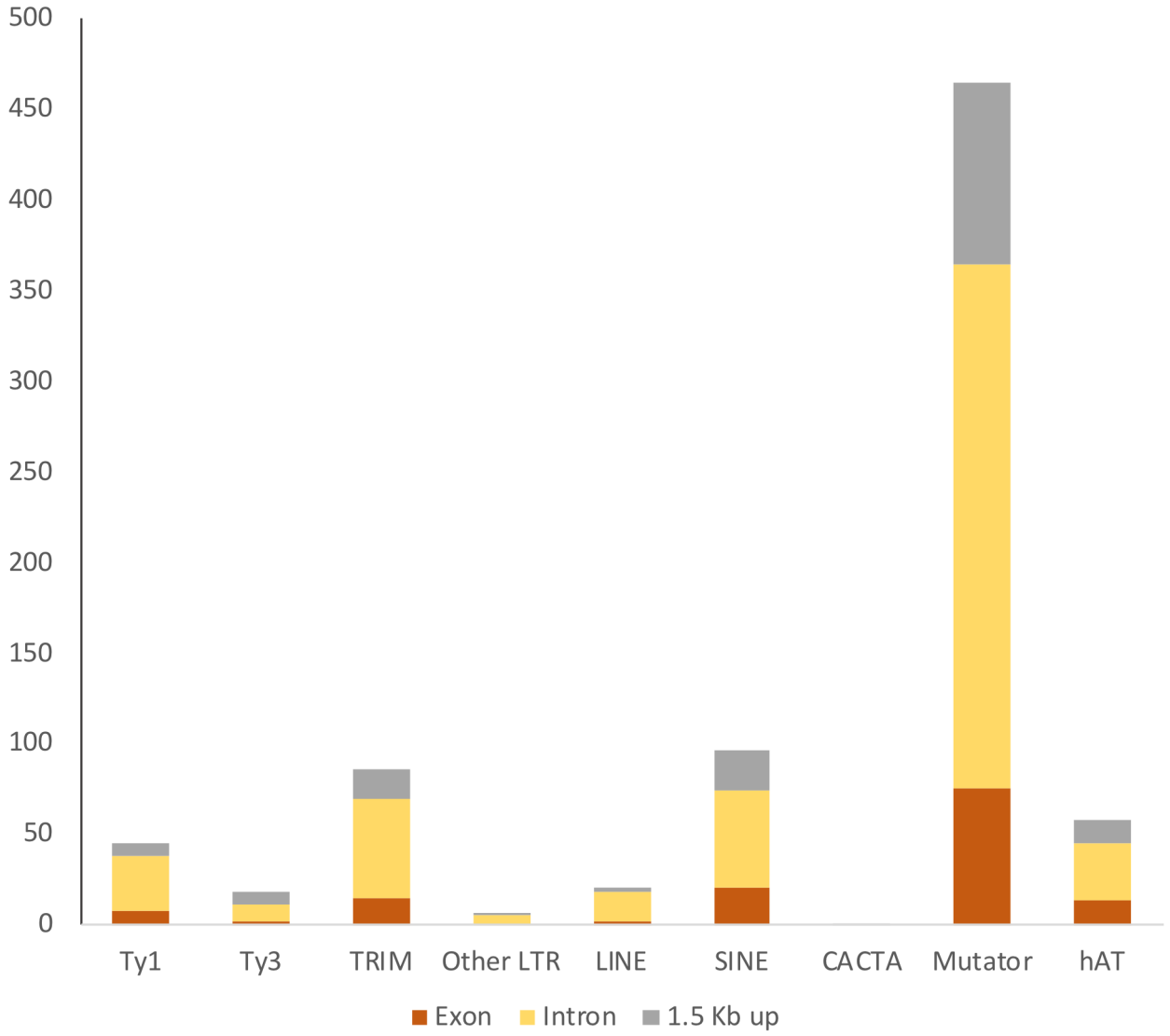
Supplementary Figure 27
Extrachromosomal circular DNAs detected in diploid and tetraploid *Arachis*

Depths of coverage of mapped reads from extrachromosomal circular DNA onto the abundant transposable elements detected in the libraries. Genotypes on left. Colors indicate the presence of SNPs, maximum coverage is indicated on the right. *Mu4-1* is a DNA MULE transposon, *Zuhe* is a Ty3-*gypsy* LTR transposon, *Yara* is a Ty1-*copia* LTR transposon. No abundant circular DNAs were detected in hybrids or *A. hypogaea* that were not detected in one or both of the ancestral diploids.



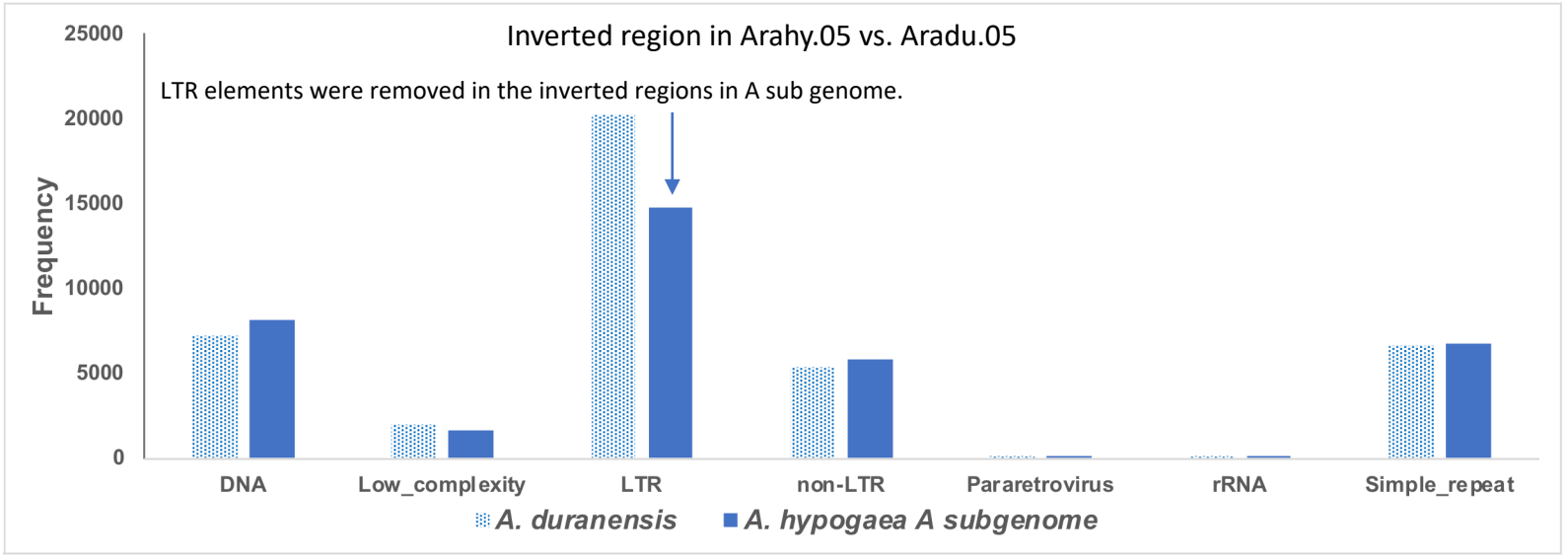
Supplementary Figure 28
Mobile elements recently inserted near Tifrunner genes

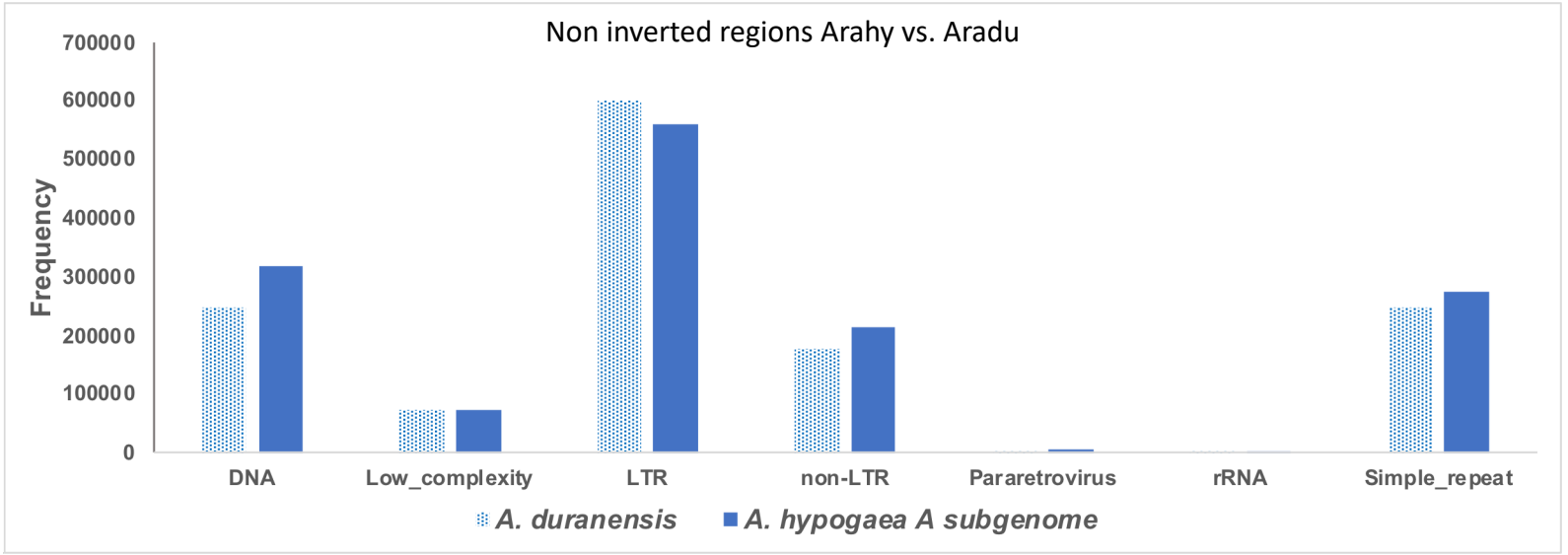
Recently inserted transposons located in or near annotated genes in the Tifrunner genome. Mutator-like elements (MULEs) dominate detected transposon activity near genes.



Supplementary Fig 29

Repeat counts in inverted and non-inverted regions of the tetraploid chromosomes of *A. hypogaea* cv. Tifrunner





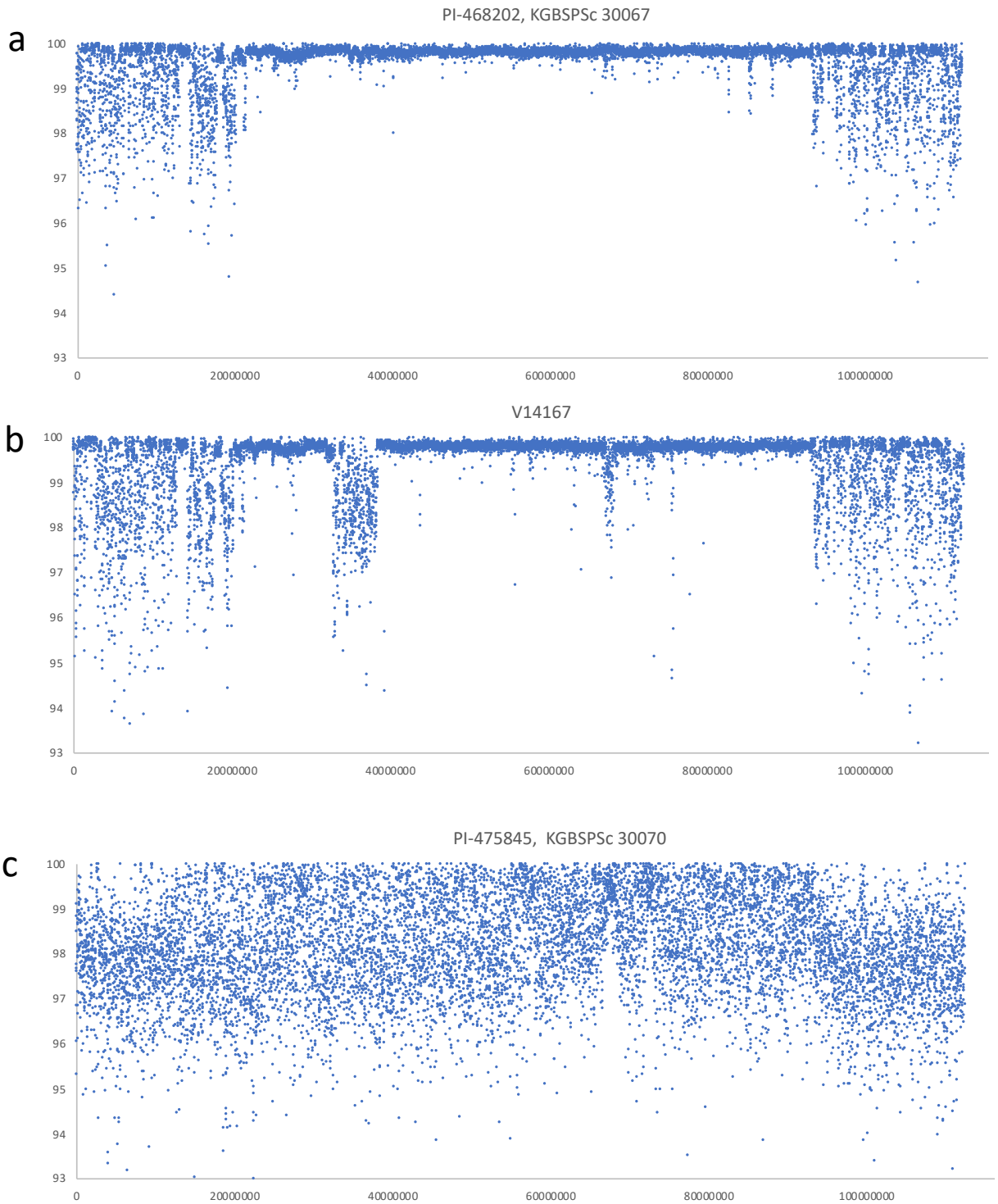
Supplementary Figure 30

Percentage identities (y axes) in 10 kb windows of three different *A. duranensis* accessions vs. distance along *A. hypogaea* cv. Tifrunner Arahy.01 (x axis). Estimated by alignment of Illumina whole genome sequences to Arahy.01

(a) PI-468202, from Rio Seco, Argentina, one of the two accessions with highest DNA similarity to the A subgenome of *A. hypogaea;*

(b) V14167 from Salta, Argentina, with a sequenced genome (Bertioli et al 2015);

(c) PI-475845 from Tarija, Bolivia, with a partially assembled genome (Chen et al 2016)



References

Bertioli et al 2016 Nature Genetics, 47, 438.

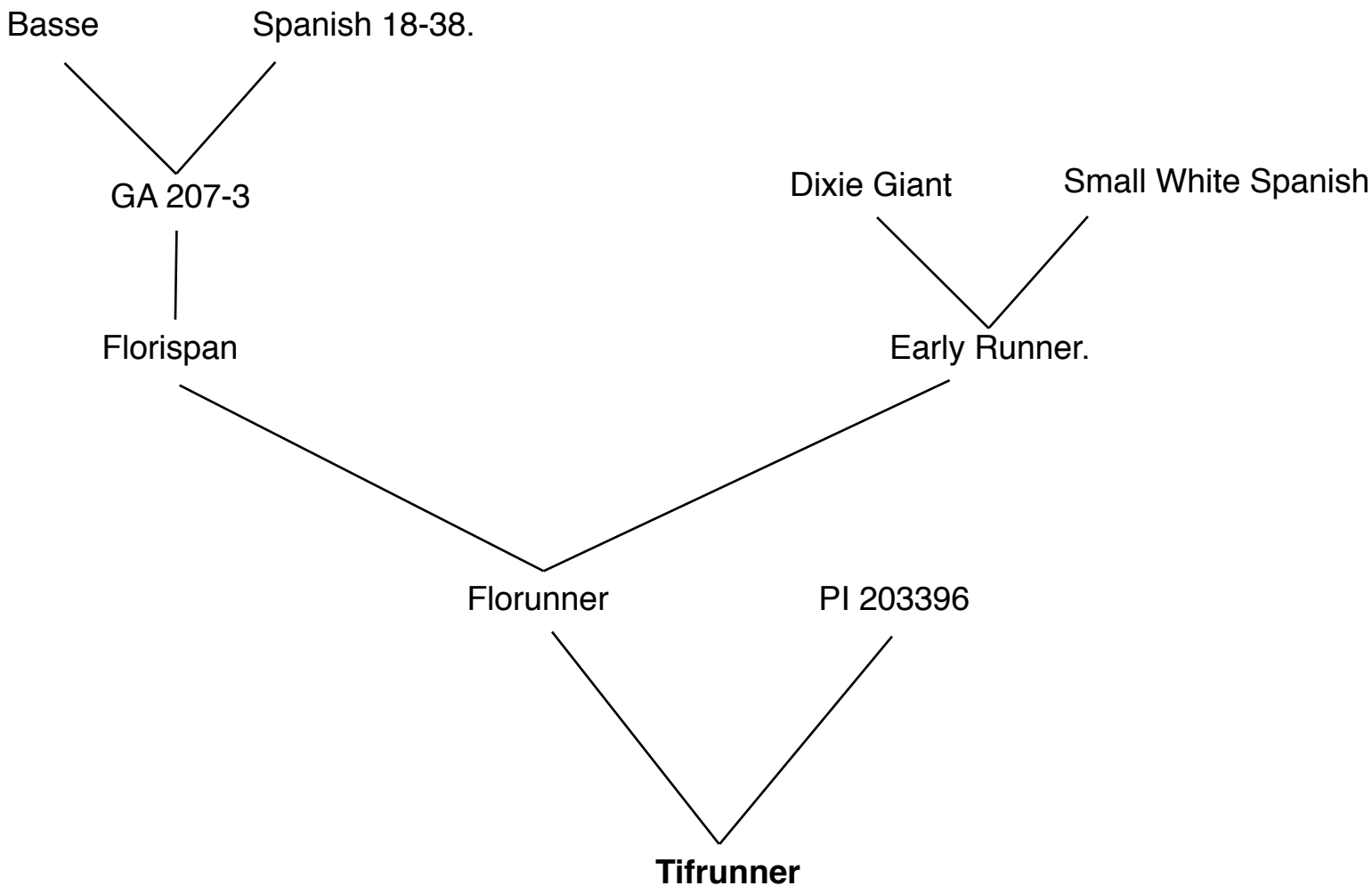
Chen et al 2016 Proc. Natl. Acad. Sci. U.S.A. 113: 6785-6790.

Supplementary Figure 31

Recorded pedigree of *A. hypogaea* cv. Tifrunner

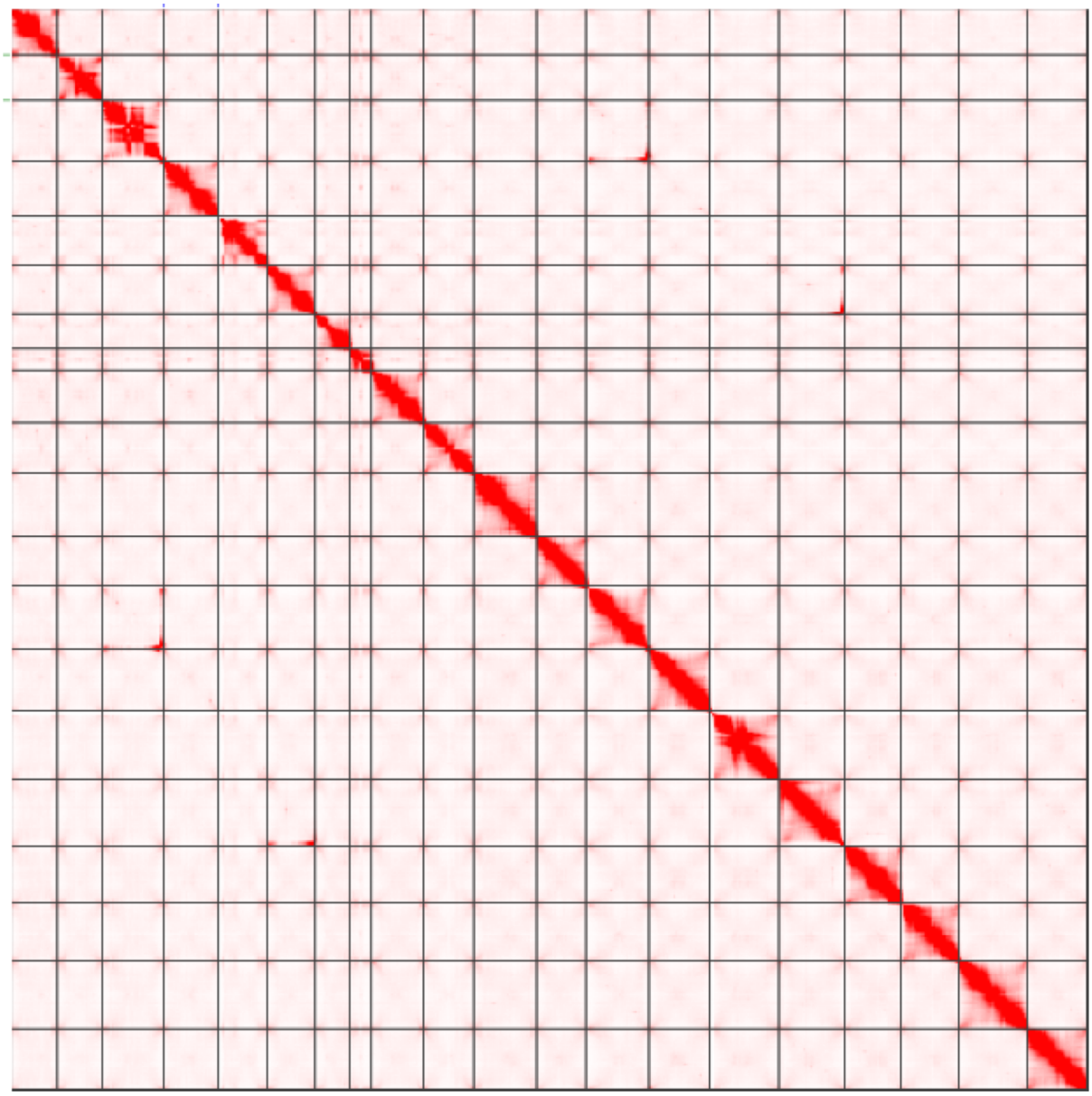
Basse, Dixie Giant and PI 203396 are thought to be *A. hypogaea* subsp. *hypogaea* var. *hypogaea*

Spanish 18-38 and Small White Spanish may be *A. hypogaea* subsp. *fastigiata* var. *vulgaris* (but not verified in this work)



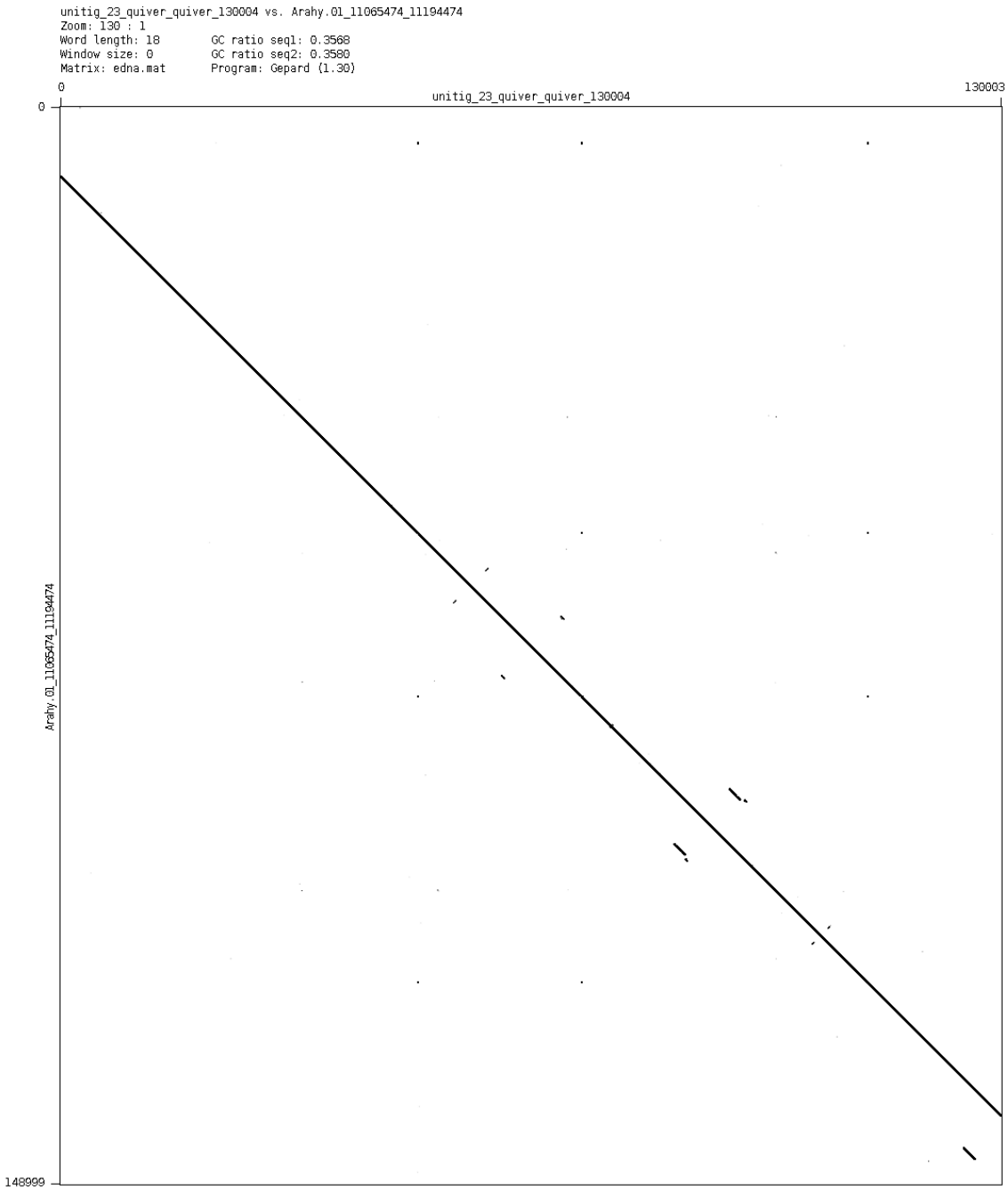
Supplementary Figure 32

Contact map visualization produced by the software Juicebox for the 20 chromosomal pseudomolecules of *A. hypogaea* cv. Tifrunner



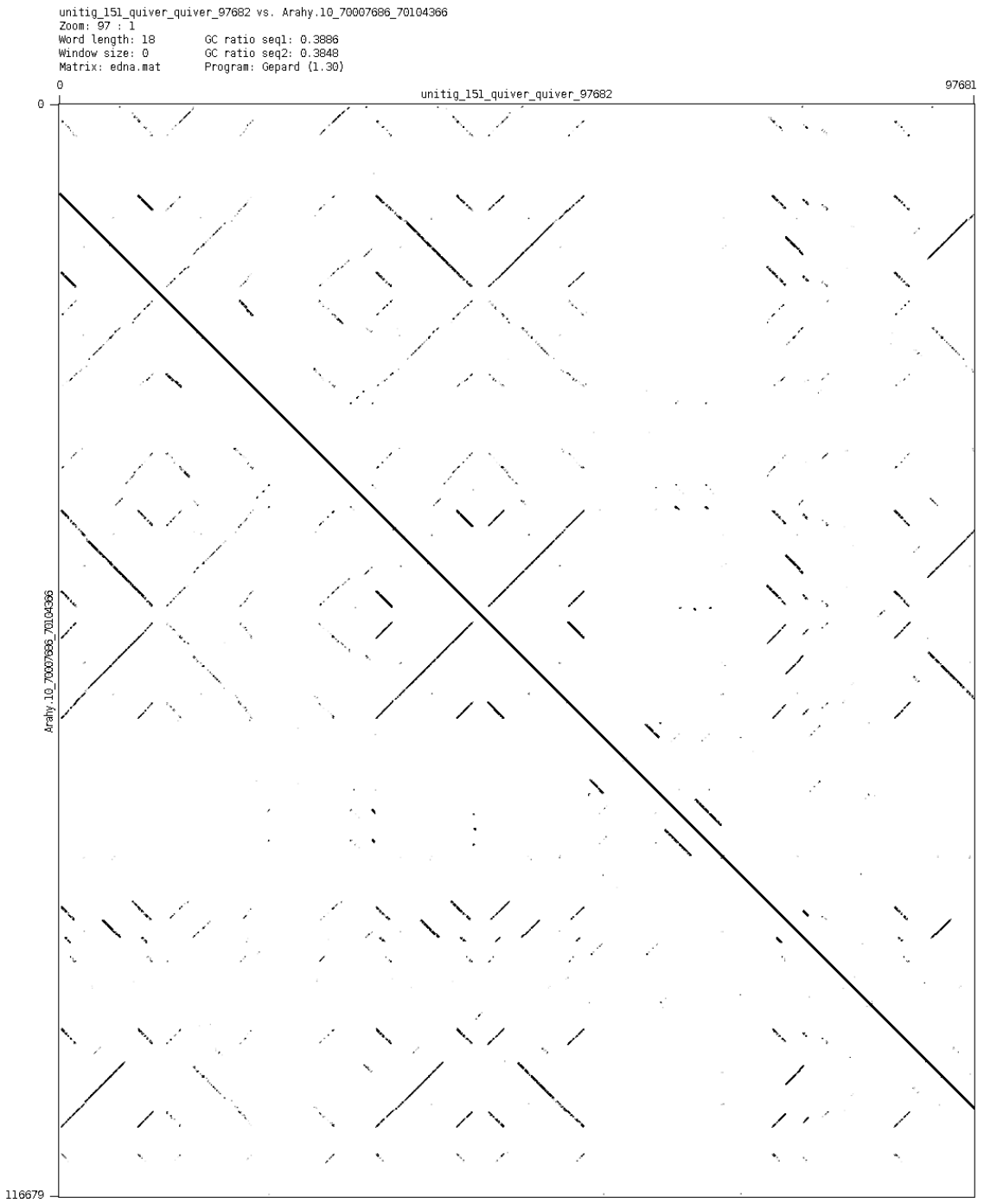
Supplementary Figure 33

Dot plot of BAC contig 159899 on a region of Arahy.01. This alignment is representative of the high quality alignments in 79 of the 175 available BAC contigs.



Supplementary Figure 34

Dot plot of BAC contig 97682 on a region of Arahy.10, which is representative of the 86 contigs from repetitive regions of the genome.



Supplementary Figure 35

Dot plot of BAC contig 140600 on a region of Arahy.17, which is representative of the 10 that align to an area where the adjacent genome contigs indicate that they overlap.