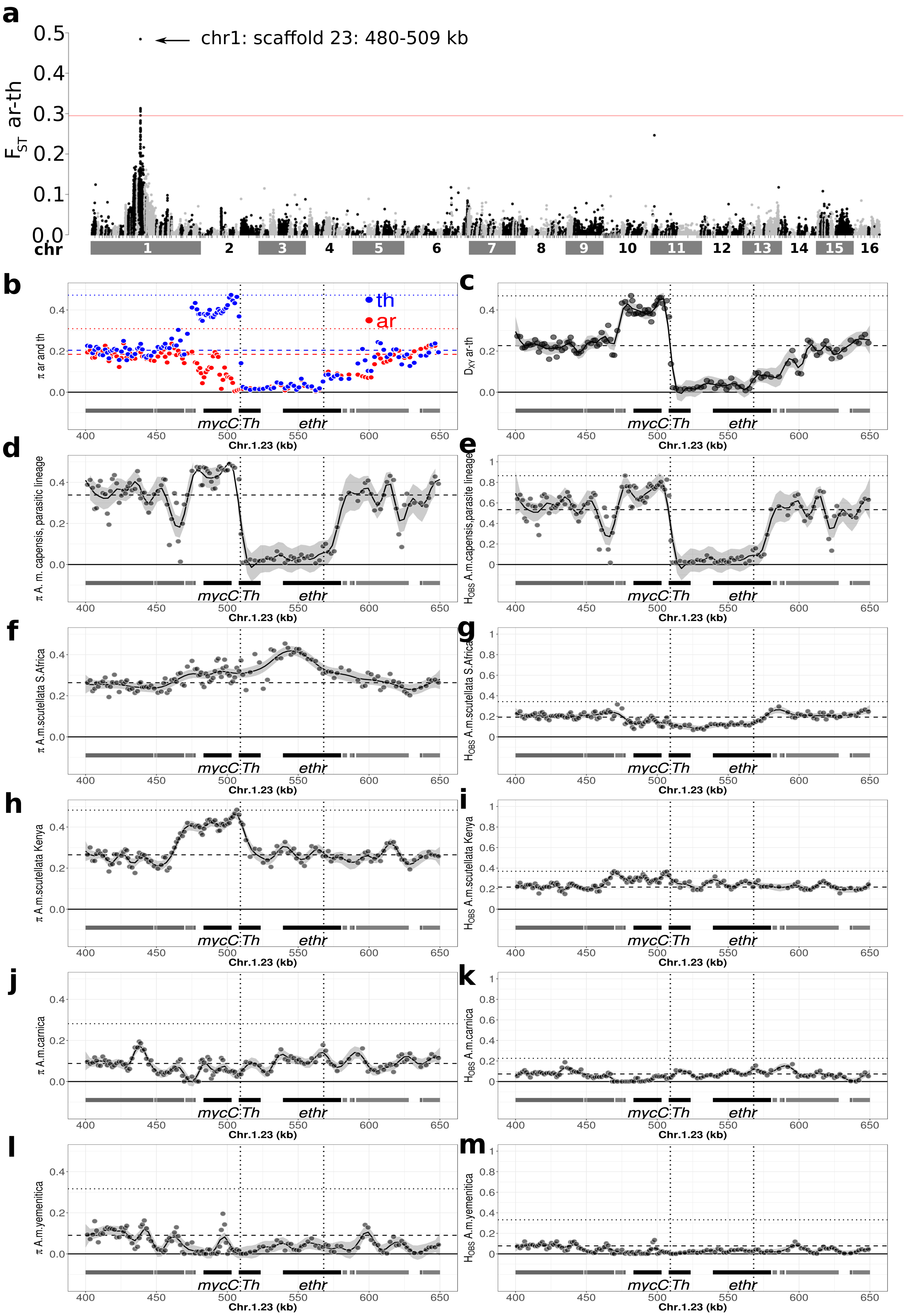
Supplementary Information

A single SNP turns a social honey bee (*Apis mellifera*) worker into a selfish parasite

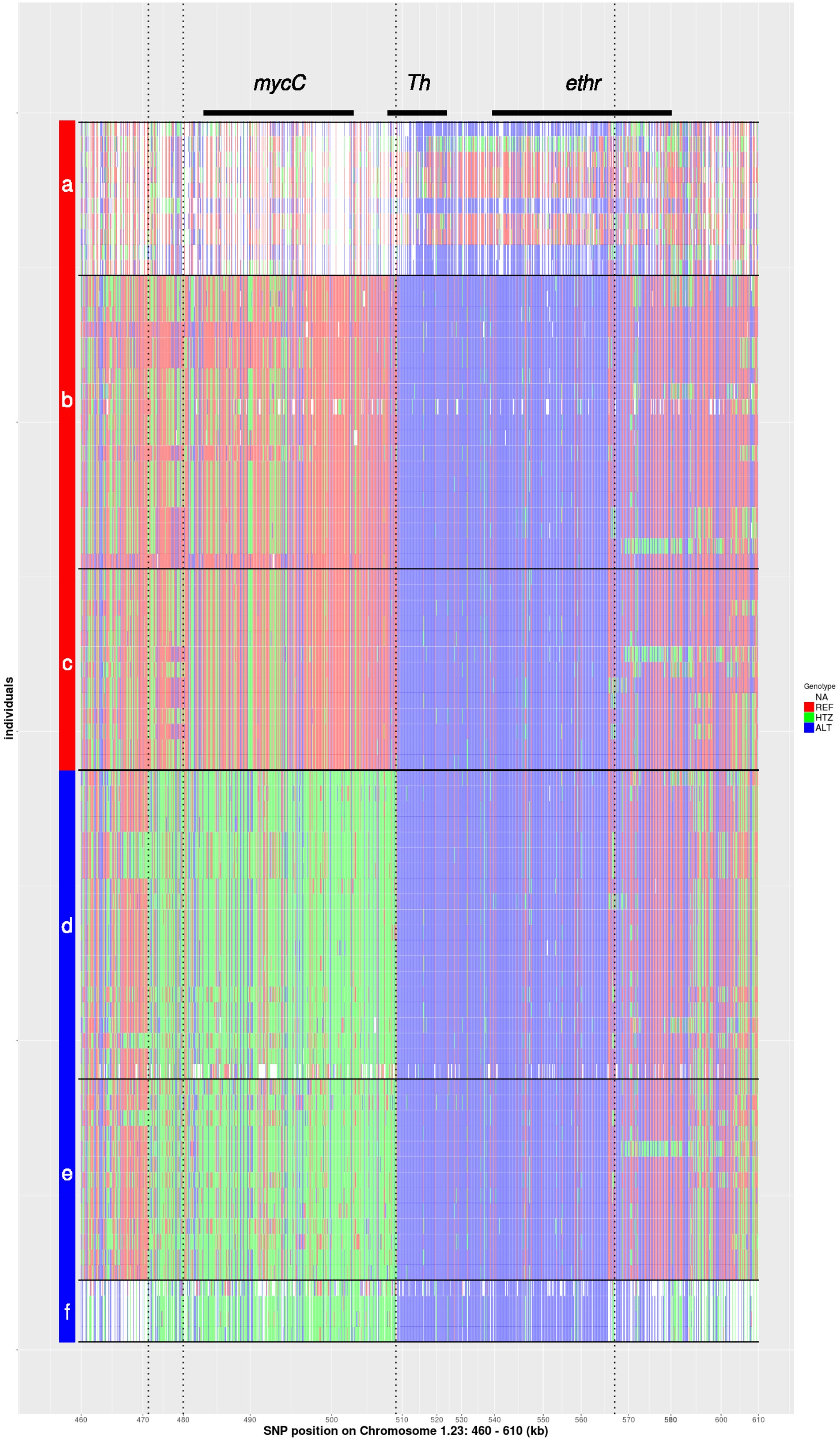
Aumer et al.

Contents

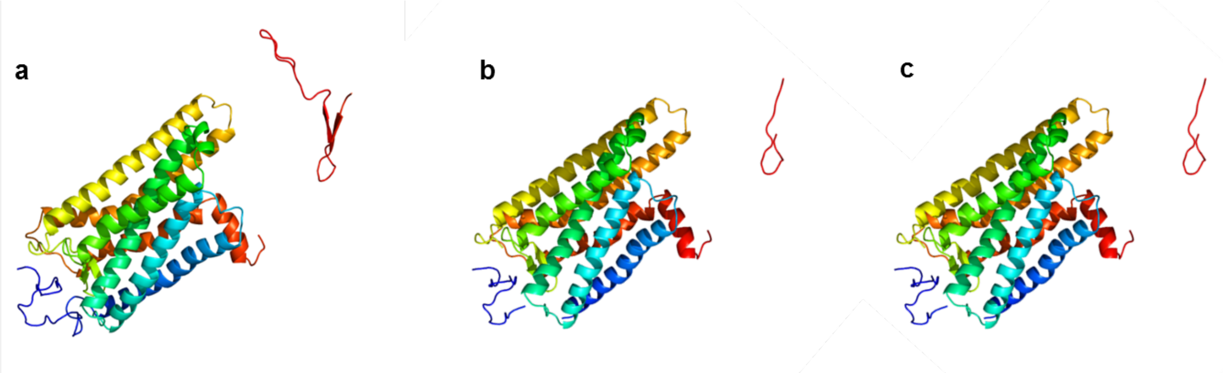


a
0.5
0.4
0.3
0.2
0.1
0.0
F_{ST} ar-th
chr1: scaffold 23: 480-509 kb
chr
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
b
π ar and th
● th
● ar
c
D_{XY} ar-th
d
π A. m. capensis, parasitic lineage
e
H_{OBS} A.m.capensis parasite lineage
f
π A.m.scutellata S.Africa
g
H_{OBS} A.m.scutellata S.Africa
h
π A.m.scutellata Kenya
i
H_{OBS} A.m.scutellata Kenya
j
π A.m.carnica
k
H_{OBS} A.m.carnica
l
π A.m.yemenitica
m
H_{OBS} A.m.yemenitica
0.0
0.2
0.4
0.6
0.8
1
mycC
Th
ethr
400
450
500
550
600
650
Chr.1.23 (kb)

Suplementary FIG. 1: Genomic summary data for the thelytoky region. The first panel (**a**) shows F_{ST} values (mean per 100 SNP window, sliding in 50 SNP steps) between thelytokous pseudoqueens from the parasitic clonal lineage (n=4) and the core set of arrhenotokous females from the mapping population (n=13). The horizontal line denotes the 99.99th percentile. The second panel (**b, c**) shows nucleotide diversity (π, **b**) in thelytokous (blue) and arrhenotokous individuals (red) of the core set of the mapping population (n=13 each), and the mean of pairwise differences between them (D_{XY}, **c**). The remainder panels show nucleotide diversity (π, **left**) and observed heterozygosity (H_{OBS}, **right**) per individual as mean in 100 SNP windows (sliding in 50 SNP steps) in thelytokous pseudoqueens from the parasitic clonal lineage (n=4) (**d,e**), South African *A. m. scutellata* (n=10) (**f,g**), Kenyan *A. m. scutellata* (n=24) (**h,i**), A. m. carnica (n=9) (**j,k**) and A. m. yemenitica (n=10) (**l,m**). In each graph (**b-m**), the genome-wide average is shown as dashed horizontal line, the 99.99th percentile is shown as dotted horizontal line, the position of the non-synonymous SNP and the border of the selective sweep are shown as the two vertical dotted lines.


mycC
Th
ethr
a
b
c
d
e
f
individuals
Genotype
NA
🟥 REF
🟩 HTZ
🟦 ALT
460
470
480
490
500
510
520
530
540
550
560
570
580
590
600
610
SNP position on Chromosome 1.23: 460 - 610 (kb)

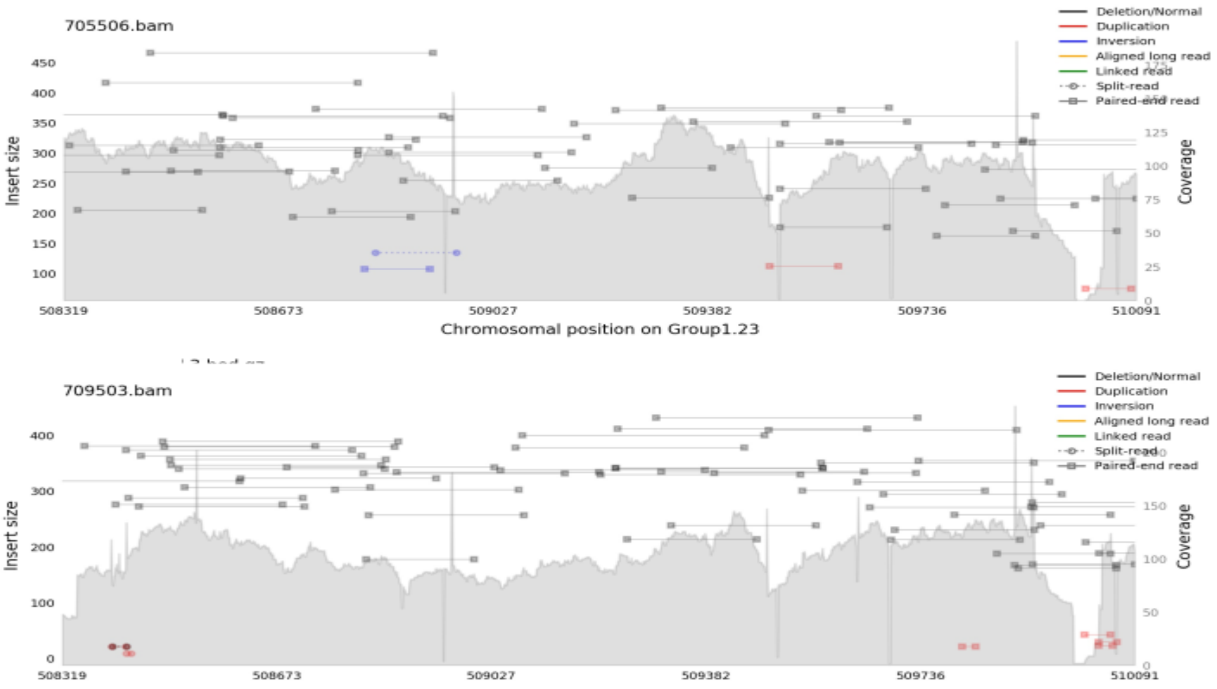
Supplementary FIG. 2: Heatmap showing genotypes of biallelic SNPs in the region of the thelytoky locus (red = homozygous for the reference genome sequence allele, blue = homozygous for the alternative allele, green = heterozygous) per individual (1 row per individual). The dotted vertical lines denote the locations where heterozygosity or nucleotide diversity change, i.e. borders of the selective sweep and the heterozygous region. The coloured vertical bars at the left denote arrhenotokous (red) and thelytokous (blue) individuals. Groups of individuals are separated by black lines: **a)** arrhenotokous *A. m. scutellata* (South Africa) (n=10), **b)** other arrhenotokous workers from the *A. m. capensis* mapping population (n=20), **c)** core arrhenotokous workers from the mapping population (n=13), **d)** other thelytokous workers from the mapping population (n=19), **e)** core thelytokous workers from the mapping population (n=13), **f)** thelytokous pseudoqueens from the *A. m. capensis* parasitic clonal lineage (n=4).



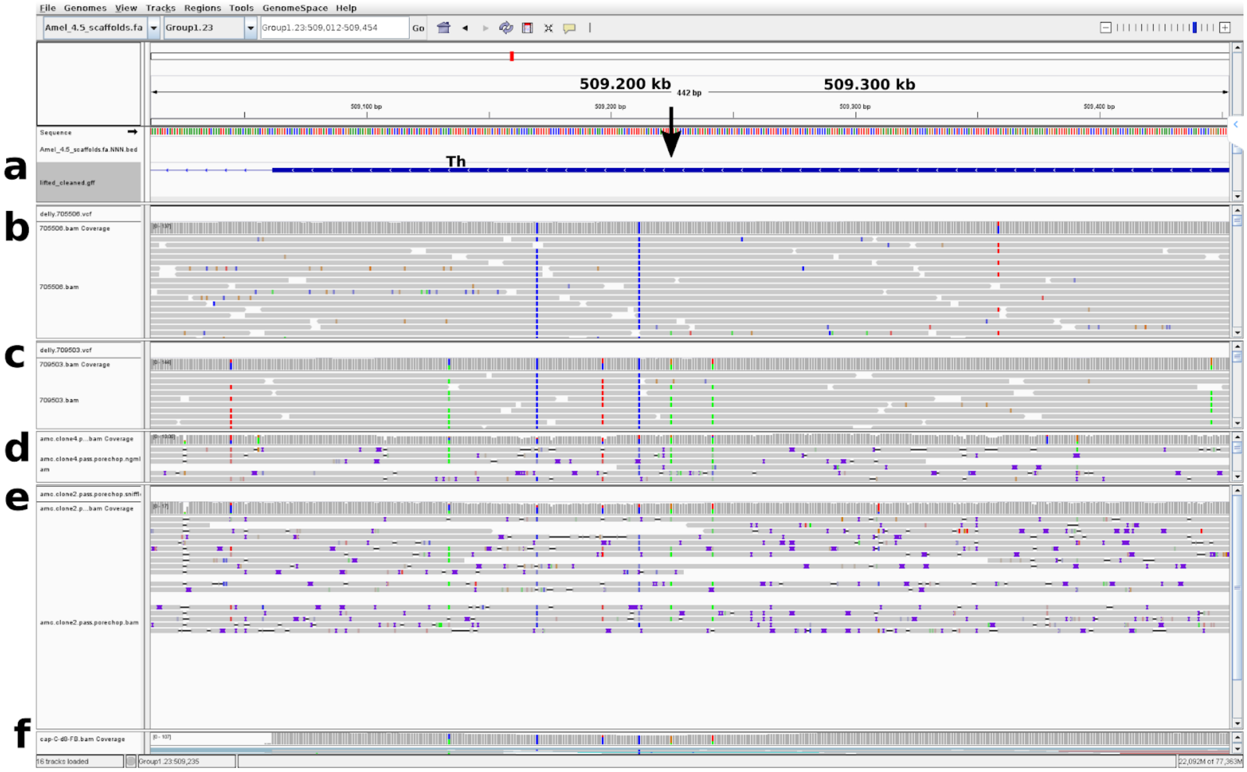
Supplementary FIG. 3: Modelled tertiary structures of *Ethr* isoforms. a) The reference sequence obtained from NCBI (XP_006570145.1). 365 residues (84% of sequence) have been modelled with 100% confidence. **b)** The first functional variant detected in *A. m. capensis* (with Asp at AA197 in the amino acid sequence). 329 residues (78% sequence) have been modelled with 100% confidence. **c)** the second functional variant detected in *A. m. capensis* (with Asn at AA197 in the amino acid sequence). 329 residues (78% sequence) have been modelled with 100% confidence. The models are shown in rainbow colours from the N-terminus (blue) to the C-terminus (red) (Kelley et al. 2015).

Supplementary figures 4 to 10

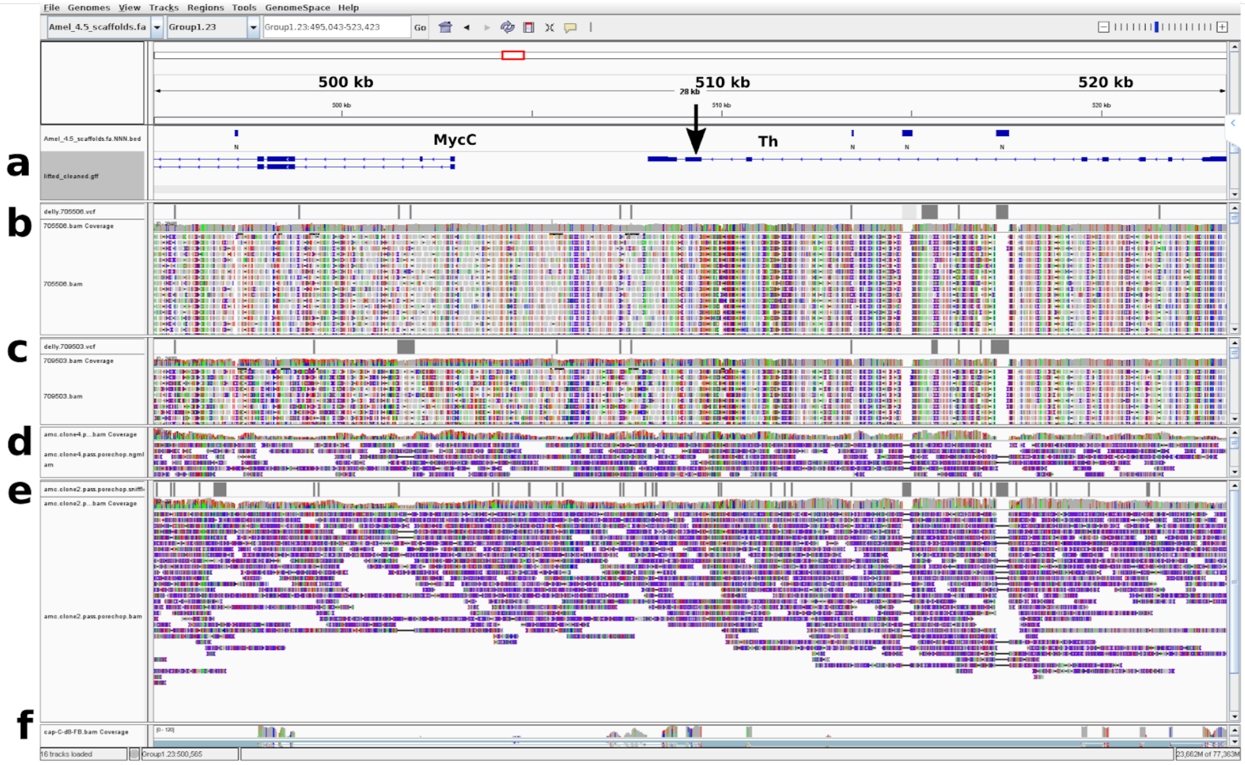
The following supplementary figures show screenshots of genomic data visualized in samplot or IGV (integrated genome viewer, Thorvaldsdóttir et al. 2013). The first set of figures (Supplementary fig. 4-7) shows different-sized regions based on Amel_4.5 as a reference genome sequence (identical to the reference used for the mapping and heterozygosity analyses in the main text). The second set (Supplementary fig. 8-10) shows the same data but mapped to the recently published genome sequence assembly of the honeybee (Amel_HAv3.1, Wallberg et al. 2018). SV callers did not detect inversion in or around the thelytoky locus, neither in short reads (delly) nor long reads (sniffles), these visualisations provide further evidence for an absence of an inversion at the thelytoky locus.



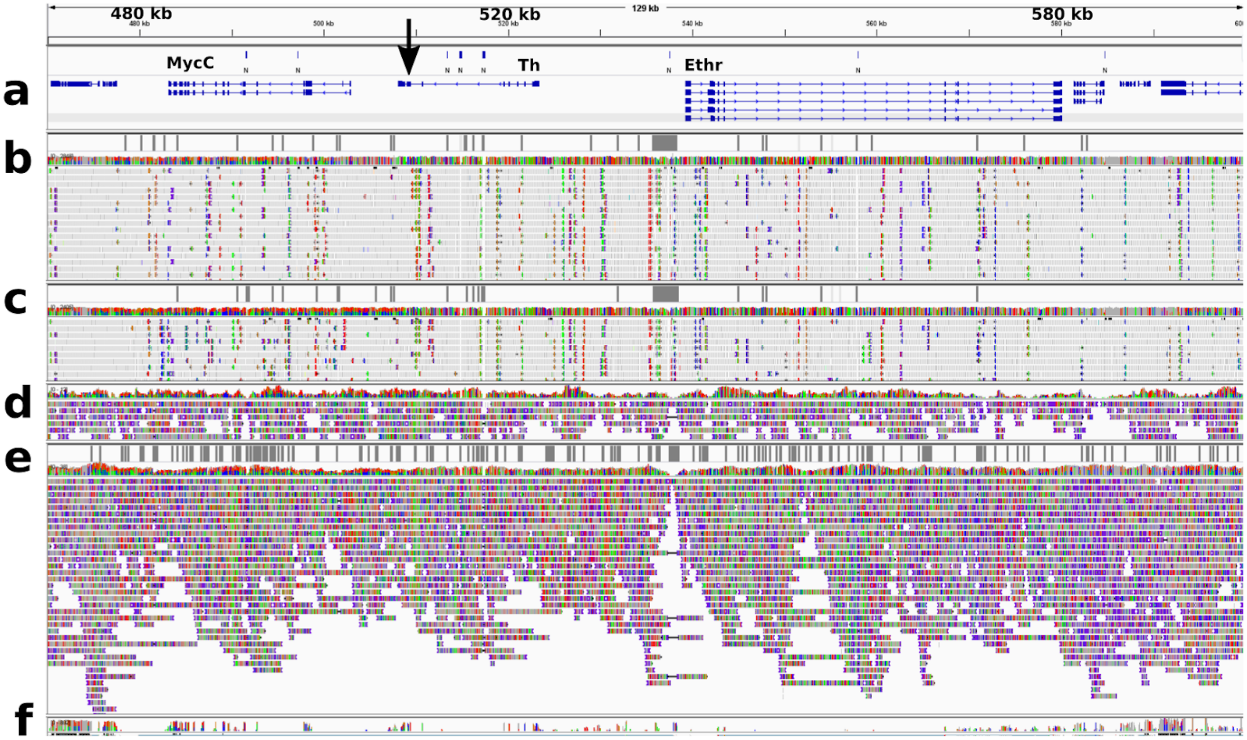
Supplementary FIG. 4: Samplot visualisation of read mapping positions and read splittings in the highest coverage arrhenotokous individual (barcode 705506, top) and the highest coverage thelytokous individual (barcode 709503, bottom).



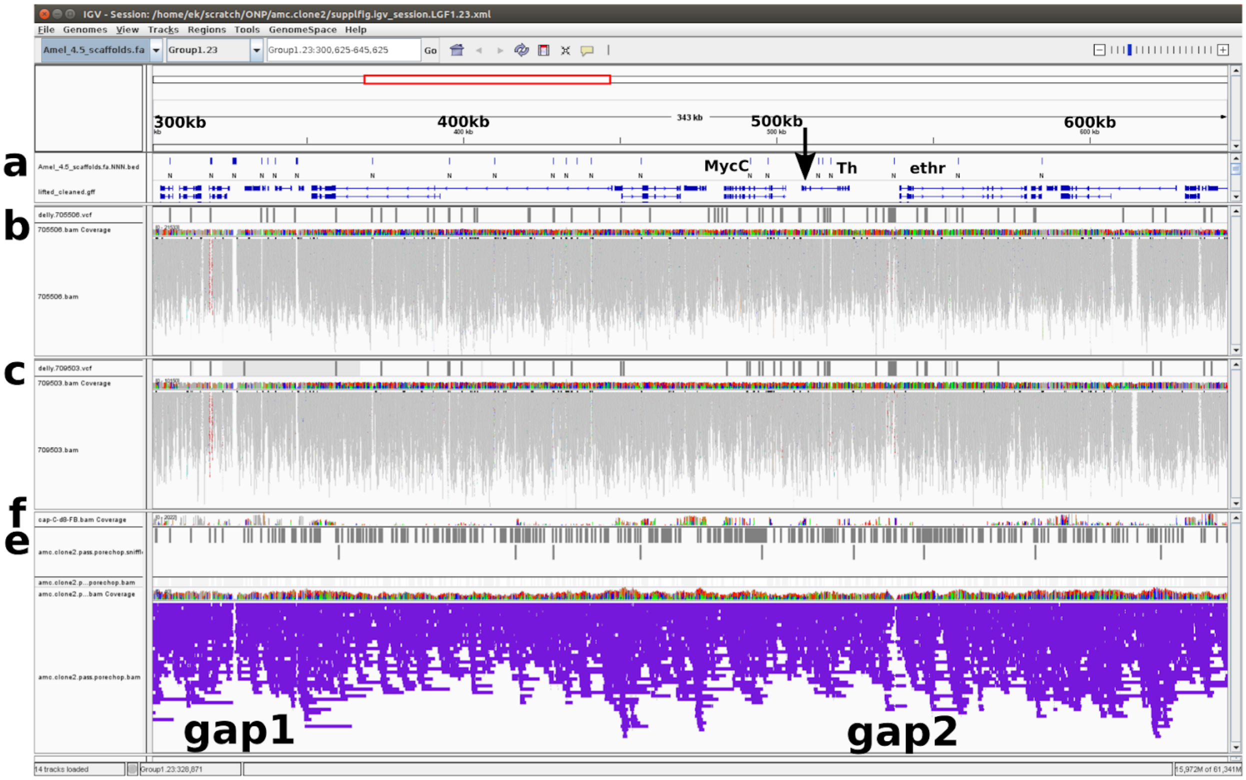
Supplementary FIG. 5: IGV visualisation of the thelytoky locus (arrow denoting the non-synonymous, heterozygous SNP in the second last exon of *Th*). Data tracks show **a)** annotation, **b)** highest coverage arrhenotokous individual (Illumina short reads, incl. coverage track in log-scale), **c)** highest coverage thelytokous individual (Illumina short reads, incl. coverage track in log-scale). **d)** individual 1 (~5x genomic coverage) of the four pseudoqueens from the thelytokous parasitic lineage (Oxford Nanopore long reads, incl. coverage track in log-scale), **e)** individual 2 (~25x genomic coverage) of the four pseudoqueens from the thelytokous parasitic lineage (Oxford Nanopore long reads, incl. coverage track in log-scale), **f)** fatbody RNAseq reads from a pseudoqueen from the thelytokous parasitic lineage (only as coverage track in log-scale).


File Genomes View Tracks Regions Tools GenomeSpace Help
Amel_4.5_scaffolds.fa
Group1.23
Group1.23:495,043-523,423
Go
500 kb
510 kb
520 kb
28 kb
MycC
Th
N
a
Amel_4.5_scaffolds.fa.NNN.bed
lifted_cleaned.gff
b
delly.705506.vcf
705506.bam Coverage
705506.bam
c
delly.709503.vcf
709503.bam Coverage
709503.bam
d
amc.clone4.p...bam Coverage
amc.clone4.pass.porechop.ngmlam
e
amc.clone2.pass.porechop.sniffl
amc.clone2.p...bam Coverage
amc.clone2.pass.porechop.bam
f
cap-C-d8-FB.bam Coverage
[0 - 120]
16 tracks loaded
Group1.23:500,565

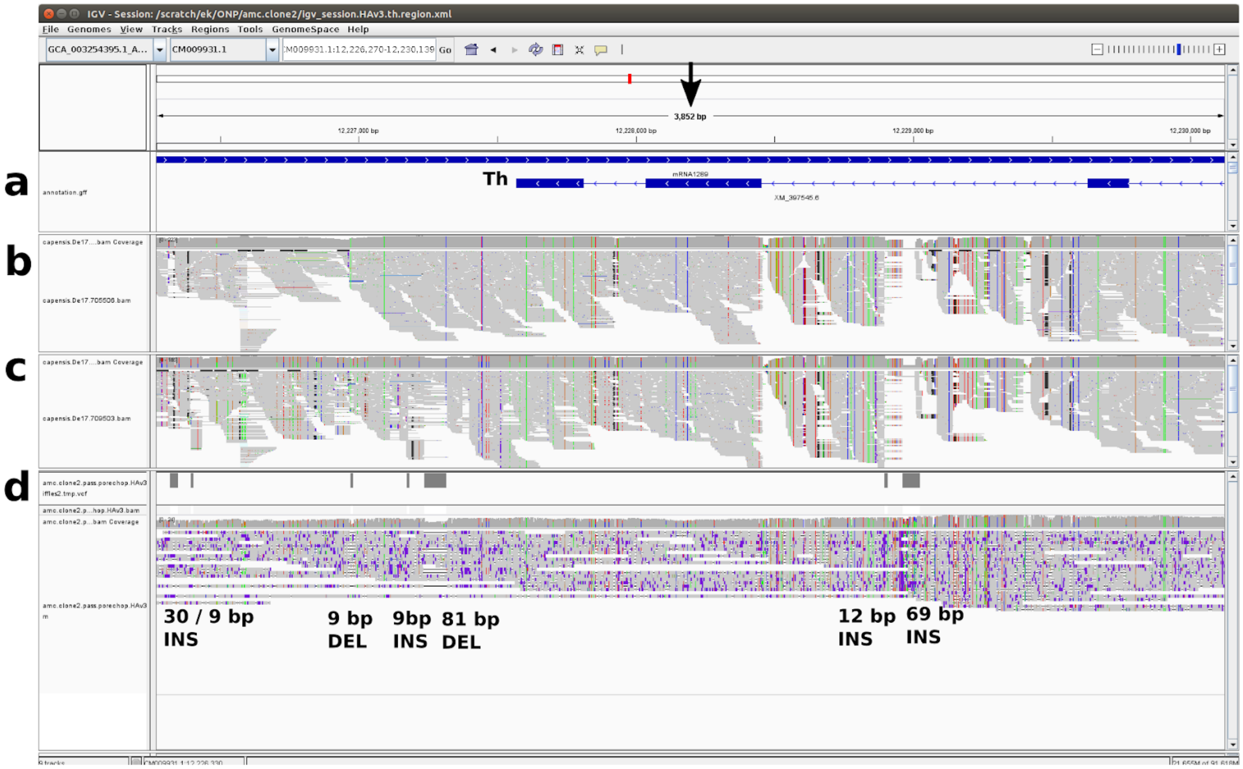
Supplementary FIG. 6: IGV visualisation of the thelytoky locus (arrow denoting the non-synonymous, heterozygous SNP in the second last exon of *Th*). Data tracks show **a)** annotation, **b)** highest coverage arrhenotokous individual (Illumina short reads, incl. coverage track in log-scale and SV calls by delly), **c)** highest coverage thelytokous individual (Illumina short reads, incl. coverage track in log-scale and SV calls by delly). **d)** individual 1 (~5x genomic coverage) of the four pseudoqueens from the thelytokous parasitic lineage (Oxford Nanopore long reads, incl. coverage track in log-scale), **e)** individual 2 (~25x genomic coverage) of the four pseudoqueens from the thelytokous parasitic lineage (Oxford Nanopore long reads, incl. coverage track in log-scale and SV calls by sniffles), **f)** fatbody RNAseq reads from a pseudoqueen from the thelytokous parasitic lineage (only as coverage track in log-scale).



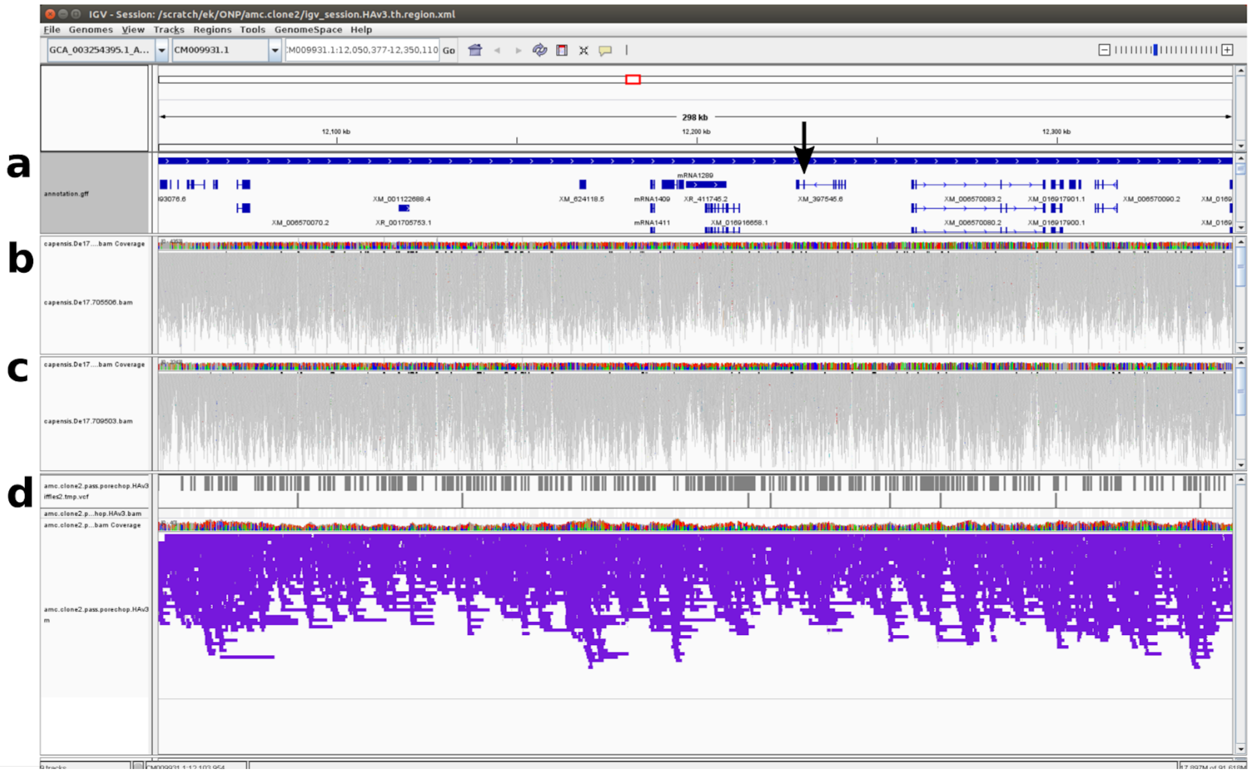
Supplementary FIG. 7: IGV visualisation of the thelytoky locus (arrow denoting the non-synonymous, heterozygous SNP in the second last exon of *Th*). Data tracks show **a)** annotation, **b)** highest coverage arrhenotokous individual (Illumina short reads, incl. coverage track in log-scale and SV calls by delly), **c)** highest coverage thelytokous individual (Illumina short reads, incl. coverage track in log-scale and SV calls by delly). **d)** individual 1 (~5x genomic coverage) of the four pseudoqueens from the thelytokous parasitic lineage (Oxford Nanopore long reads, incl. coverage track in log-scale), **e)** individual 2 (~25x genomic coverage) of the four pseudoqueens from the thelytokous parasitic lineage (Oxford Nanopore long reads, incl. coverage track in log-scale and SV calls by sniffles), **f)** fatbody RNAseq reads from a pseudoqueen from the thelytokous parasitic lineage (only as coverage track in log-scale).


IGV - Session: /home/ek/scratch/ONP/amc.clone2/supplfig.igv_session.LGF1.23.xml
File Genomes View Tracks Regions Tools GenomeSpace Help
Amel_4.5_scaffolds.fa
Group1.23
Group1.23:300,625-645,625
Go
300kb
400kb
343 kb
500kb
600kb
a
Amel_4.5_scaffolds.fa.NNN.bed
lifted_cleaned.gff
MycC
Th
ethr
b
delly.705506.vcf
705506.bam Coverage
705506.bam
c
delly.709503.vcf
709503.bam Coverage
709503.bam
f
e
cap-C-d8-FB.bam Coverage
amc.clone2.pass.porechop.sniffl
amc.clone2.p...porechop.bam
amc.clone2.p...bam Coverage
amc.clone2.pass.porechop.bam
gap1
gap2
14 tracks loaded
Group1.23:328,871

Supplementary FIG. 8: IGV visualisation of the thelytoky locus based on the improved and recently published assembly Amel_HAv3.1 (Wallberg et al. 2018) (arrow denoting the non-synonymous, heterozygous SNP in the second last exon of *Th*). Data tracks show **a)** annotation, **b)** highest coverage arrhenotokous individual (Illumina short reads, incl. coverage track in log-scale and SV calls by delly), **c)** highest coverage thelytokous individual (Illumina short reads, incl. coverage track in log-scale and SV calls by delly). **d)** individual 1 (~5x genomic coverage) of the four pseudoqueens from the thelytokous parasitic lineage (Oxford Nanopore long reads, incl. coverage track in log-scale), **e)** individual 2 (~25x genomic coverage) of the four pseudoqueens from the thelytokous parasitic lineage (Oxford Nanopore long reads, incl. coverage track in log-scale and SV calls by sniffles), **f)** fatbody RNAseq reads from a pseudoqueen from the thelytokous parasitic lineage (only as coverage track in log-scale). Gap1 and gap2 denote coverage gaps in both short reads and long reads datasets caused by stretches of “N” (undefined bases, i.e. assembly gaps). Both gaps and coverage reduction are not present in an improved genome assembly (see next graphs).


IGV - Session: /scratch/ek/ONP/amc.clone2/igv_session.HAv3.th.region.xml
File Genomes View Tracks Regions Tools GenomeSpace Help
GCA_003254395.1_A...
CM009931.1
CM009931.1:12,226,270-12,230,139
Go
3,852 bp
12,227,000 bp
12,228,000 bp
12,229,000 bp
12,230,000 bp
a
annotation.gff
Th
mRNA1289
XM_397545.6
b
capensis.De17....bam Coverage
capensis.De17.705506.bam
c
capensis.De17....bam Coverage
capensis.De17.709503.bam
d
amc.clone2.pass.porechop.HAv3 iffles2.tmp.vcf
amc.clone2.p...hop.HAv3.bam
amc.clone2.p...bam Coverage
amc.clone2.pass.porechop.HAv3 m
30 / 9 bp INS
9 bp DEL
9bp INS
81 bp DEL
12 bp INS
69 bp INS
9 tracks

Supplementary FIG. 9: IGV visualisation of the thelytoky locus based on the improved and recently published assembly Amel_HAv3.1 (Wallberg et al. 2018) (arrow denoting the non-synonymous, heterozygous SNP in the second last exon of *Th*). Data tracks show **a)** annotation, **b)** highest coverage arrhenotokous individual (Illumina short reads, incl. coverage track in log-scale), **c)** highest coverage thelytokous individual (Illumina short reads, incl. coverage track in log-scale), **d)** individual 2 (~25x genomic coverage) of the four pseudoqueens from the thelytokous parasitic lineage (Oxford Nanopore long reads, incl. coverage track in log-scale and SV calls by sniffles). Note that sniffles did not call other small SVs. Due to a length threshold of 8 bp.



Supplementary FIG. 10: IGV visualisation of the thelytoky locus based on the improved and recently published assembly Amel_HAv3.1 (Wallberg et al. 2018) (arrow denoting the non-synonymous, heterozygous SNP in the second last exon of *Th*). Data tracks show **a)** annotation, **b)** highest coverage arrhenotokous individual (Illumina short reads, incl. coverage track in log-scale), **c)** highest coverage thelytokous individual (Illumina short reads, incl. coverage track in log-scale), **d)** individual 2 (~25x genomic coverage) of the four pseudoqueens from the thelytokous parasitic lineage (Oxford Nanopore long reads, incl. coverage track in log-scale and SV calls by sniffles).

Citations

Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJE. 2015. The Phyre2 web portal for protein modeling, prediciton and analysis, Nat. Protoc. 10:845-858.

Thorvaldsdóttir H, Robinson JT, Mesirov JP. 2013. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief. Bioinform14: 178-192.

Wallberg A, Bunikis I, Pettersson OV, Mosbech M-B, Childers AK, Evans JD, Mikheyev AS, Robertson HM, Robinson GE, Webster MT. 2018. A hybrid *de novo* genome assembly of the honeybee, *Apis mellifera*, with chromosome-length scaffolds. bioRxiv:doi: https://doi.org/10.1101/361469.