BMC Biology

RESEARCH ARTICLE

Open Access

Hybridisation has shaped a recent radiation of grass-feeding aphids

Thomas C. Mathers[1,2*], Roland H. M. Wouters[1], Sam T. Mugford[1], Roberto Biello[1], Cock van Oosterhout[3] and Saskia A. Hogenhout[1*]

Abstract

Background Aphids are common crop pests. These insects reproduce by facultative parthenogenesis involving several rounds of clonal reproduction interspersed with an occasional sexual cycle. Furthermore, clonal aphids give birth to live young that are already pregnant. These qualities enable rapid population growth and have facilitated the colonisation of crops globally. In several cases, so-called "super clones" have come to dominate agricultural systems. However, the extent to which the sexual stage of the aphid life cycle has shaped global pest populations has remained unclear, as have the origins of successful lineages. Here, we used chromosome-scale genome assemblies to disentangle the evolution of two global pests of cereals—the English (*Sitobion avenae*) and Indian (*Sitobion miscanthi*) grain aphids.

Results Genome-wide divergence between *S. avenae* and *S. miscanthi* is low. Moreover, comparison of haplotype-resolved assemblies revealed that the *S. miscanthi* isolate used for genome sequencing is likely a hybrid, with one of its diploid genome copies closely related to *S. avenae* (~ 0.5% divergence) and the other substantially more divergent (> 1%). Population genomics analyses of UK and China grain aphids showed that *S. avenae* and *S. miscanthi* are part of a cryptic species complex with many highly differentiated lineages that predate the origins of agriculture. The complex consists of hybrid lineages that display a tangled history of hybridisation and genetic introgression.

Conclusions Our analyses reveal that hybridisation has substantially contributed to grain aphid diversity, and hence, to the evolutionary potential of this important pest species. Furthermore, we propose that aphids are particularly well placed to exploit hybridisation events via the rapid propagation of live-born "frozen hybrids" via asexual reproduction, increasing the likelihood of hybrid lineage formation.

Keywords Insect crop pest, Introgression, Comparative genomics, Population genomics, Genome assembly, *Sitobion avenae*, *Sitobion miscanthi*, *Metopolophium dirhodum*

Background

Crop pests and pathogens have evolved from species colonising wild plants to take advantage of new niches created by agriculture [1, 2]. Often, these pests and pathogens can evolve rapidly to overcome pesticides, the introduction of resistant crop varieties or to subvert other control measures [3]. Furthermore, pests and pathogens may periodically undergo host jumps causing disease outbreaks [4–6]. The rapid evolution of pests and pathogens may occur via selection acting on standing genetic variation [7–9], novel innovations derived

*Correspondence:
Thomas C. Mathers
thomas.mathers@jic.ac.uk; tm18@sanger.ac.uk
Saskia A. Hogenhout
saskia.hogenhout@jic.ac.uk
[1] Department of Crop Genetics, John Innes Centre, Norwich Research Park, Norwich, UK
[2] Present Address: Tree of Life, Welcome Sanger Institute, Hinxton, Cambridge, UK
[3] School of Environmental Sciences, University of East Anglia, Norwich, UK

from de novo mutation [10, 11] or by acquiring genetic innovations from other species or populations through hybridisation and introgression [12–15]. Understanding the origins, diversity and evolutionary potential of pests and pathogens is therefore of fundamental importance.

Among insect crop pests, aphids—a diverse group of sap-sucking insects from order Hemiptera—are particularly important due to their role as vectors of plant disease agents [16–18]. A key aspect of aphid success as crop pests is their "best of both worlds" approach to reproduction that involves multiple rounds of asexual reproduction alternated with occasional sexual reproduction [19–21]. This reproductive mode, known as cyclical parthenogenesis, involves specialised reproductive morphs and maximises population expansion via clonal propagation whilst providing opportunity for mixing of genetic variation during the sexual cycle. In spring and summer, asexual females rapidly produce live-born, genetically identical (with the exception of de novo mutation and gene conversion) offspring via apomictic parthenogenesis [22–24]. Population expansions during this phase are further accelerated by telescoping of generations, whereby aphid females give birth to multiple live young (viviparity) that already have daughters developing inside them. Furthermore, winged asexually reproducing morphs may be generated that facilitate dispersal, potentially over long distances [25]. In the autumn and winter, differentiated male and female morphs are induced that reproduce sexually, producing overwintering eggs that hatch as asexual females the following spring, restarting the cycle. As such, unlike strictly sexual species, high fitness aphid genotypes fortuitously produced by sexual reproduction can be rapidly amplified during the asexual stage. In many cases, this has led to the proliferation of so-called "super clones" which dominate aphid populations and can spread globally [26–29]. However, the origins of highly successful aphid pest lineages are often unknown.

Cereal aphids present an ideal opportunity to investigate the evolutionary genomics of crop pest emergence, particularly given the global cultivation of wheat and other cereal crops. Specialisation on cereals and other grasses has occurred multiple times during aphid evolution and several grass-specialists have become important pests of crops [30, 31]. Among aphid cereal-specialised lineages, the English and Indian grain aphids, *Sitobion avenae* and *Sitobion miscanthi*, are particularly destructive due to their global distribution and role as vectors of barley yellow dwarf virus [32]. Along with two other grain aphid species, *S. fragariae* and *S. akebiae*, the English and Indian grain aphids form a closely related complex [31]. Across Europe and in China, population genetic studies have revealed high genetic diversity of *S. avenae* and *S. miscanthi* [33–35]. However, the genome-wide diversity and diversification of *S. avenae* and *S. miscanthi*, and their evolutionary origins, is currently unknown as studies have either focused on each lineage in isolation, or only made use of a small number of microsatellite markers.

Here, we investigate the evolution of grain aphids from the *Sitobion* genus using chromosome-scale genome assemblies and population genomics. We generate chromosome-scale genome assemblies of *S. avenae* and a divergent grass-feeding aphid, *Rhopalosiphum padi* (bird cherry-oat aphid) and reassemble a recently published chromosome-scale genome sequence of *S. miscanthi* [36]. We also generated a high-quality draft genome assembly of *Metopolophium dirhodum* (rose-grain aphid), another important crop pest of grains, from a sister genus to *Sitobion* that serves as an outgroup in our analyses. On finding that *S. avenae* and *S. miscanthi* have low genome-wide sequence divergence and that the strain used to assemble the *S. miscanthi* genome is likely of hybrid origin, we reanalysed published population genomic data for *S. miscanthi* and *S. avenae* from the UK and China [35]. Using these data, we revealed that *S. avenae* and *S. miscanthi* are part of a larger cryptic species complex shaped by hybridisation.

Results

Chromosome-scale genome assemblies of Sitobion miscanthi and Sitobion avenae and a short-read assembly of Metopolophium dirhodum

We first assessed the recently published chromosome-scale genome assembly of *S. miscanthi* (Simis_v1) that derives from a Chinese lab colony dubbed Langfang-1 [36]. Simis_v1 was assembled with a combination of PacBio long reads (85 × coverage), Illumina short reads (105 × coverage) and in vivo Hi-C data (76 × coverage) for long-range scaffolding (Additional File 1: Table S1). The total length of the assembly is 398 Mb, it has a contig N50 of 1.6 Mb and a scaffold N50 of 36.3 Mb, with the nine longest scaffolds in the assembly accounting for 95% of the assembled genome content (Table 1). These nine super scaffolds are assumed to correspond to the nine chromosomes of *S. miscanthi* [36].

Alignment of Simis_v1 with chromosomes of the closely related model aphid *Acyrthosiphon pisum* [37] reveals fragmentation of the X chromosome into three chunks as well as substantial rearrangement of the autosomes (Additional File 2: Figure S1). This is unexpected as we, and others, have recently shown long-term conservation of aphid X chromosome structure across divergent lineages [37, 38]. Fragmentation of the *S. miscanthi* X chromosome may represent genuine chromosome fission events or be the result of genome assembly

Table 1 Genome assembly and annotation statistics

Species	*S. miscanthi*	*S. miscanthi*	*S. avenae*	*M. dirhodum*
Assembly	Simis_v1	Simis_v2	Siave_v2.1	Siave_v1.1
Sequencing approach[a]	IL + PB + HiC	IL + PB + HiC	IL + HiC	IL
Base pairs (Mb)	397.9	403.1	366.0	387.0
% Ns	0.01	0.13	0.80	0.08
Number of contigs[b]	1,148	1,889	23,024	28,240
Contig N50 (Mb)[b]	1.61	1.92	0.04	0.04
Number of scaffolds	655	833	14,626	24,973
Scaffold N50 (Mb)	32.70	37.52	29.84	0.05
% of asembly in chromosome length scaffolds	94.8	96.2	74.0	NA
Protein coding genes	16,006	21,798	19,919	22,349
Transcripts	16,006	23,875	23,368	24,826
Reference	Jiang et al. 2019	This study	This study	This study

[a] *IL* Illumina short reads, *Hi-C* High-throughput chromatin conformation capture, *PB* PacBio long reads

[b] Scaffolds split on runs of 10 or more Ns

error. To assess the quality of Simis_v1 chromosome-scale genomic scaffolds, we generated a genome-wide Hi-C contact map using data from the original genome assembly (Fig. 1a). Visual inspection of the Simis_v1 Hi-C contact map shows multiple off diagonal Hi-C contacts that are indicative of large-scale assembly error within scaffolds [39, 40]. Furthermore, several scaffolds show regions that have very low contact frequencies with adjacent sequence in the scaffold, potentially indicating incorrect assignment of scaffold start and end points. Taken together, these analyses suggest that the assembled chromosomes of Simis_v1 are likely to be inaccurate and that there may be only a single X chromosome in *S. miscanthi*.

Given the apparent scaffolding errors in Simis_v1, we reassembled the *S. miscanthi* genome using the original sequence data to create Simis_v2. In total, Simis_v2 spans 403 Mb and 96% of the assembly is contained in nine chromosome-scale super scaffolds that have consistent Hi-C contact frequencies along their full length (Table 1; Fig. 1b). Compared to Simis_v1, the contig N50 size of Simis_v2 is modestly improved (1.9 Mb vs 1.6 Mb; Table 1). Furthermore, based on the representation of arthropod Benchmarking sets of Universal Single-Copy Orthologs (BUSCOs; $n = 1066$), we substantially reduced the amount of missing (2.6% vs 5.3%) and duplicated (3.1% vs 4.7%) assembly content (Additional File 2: Figure S2). The improved sequence content of Simis_v2 compared to Simis_v1 is also supported by K-mer analysis of the raw Illumina reads with either genome assembly version (Additional File 2: Figure S3) and a taxon-annotated GC content-coverage plot indicates that Simis_v2 is free from obvious contamination (Additional File 2: Figure S4).

To further validate our new assembly of *S. miscanthi* and generate additional genomic resources for the *Sitobion* genus, we also generated a chromosome-scale assembly of the English grain aphid (*S. avenae*), using a combination of PCR-free Illumina short-read sequencing (75 × coverage) and in vivo Hi-C (Additional File 1: Table S1). As we are currently assembling a diverse range of aphid species [42], including several that are maintained at the John Innes Centre (JIC) Insectary, we experimented with using a mixed species sample to reduce Hi-C library preparation costs. We pooled *S. avenae* individuals with another aphid species—the bird cherry-oat aphid (*Rhopalosiphum padi*)—and sent the resulting pooled sample to Dovetail Genomics (Santa Cruz, CA) for Hi-C library preparation. As nuclei are cross-linked in vivo during Hi-C library preparation, species-specific chromatin conformation information is maintained, allowing a single library to be used to scaffold multiple species [43]. Furthermore, high sequence divergence between *S. avenae* and *R. padi* (Aphidini vs Macrosyphini synonymous site divergence = ~ 34% [37]) minimises the chance of Hi-C reads mismapping between the two species. To support this effort, we also generated a new draft genome assembly from the *R. padi* clonal lineage maintained at JIC using 10 × genomics linked reads.

In total, we generated 20.7 Gb of multi-species Hi-C data giving ~ 10 × coverage of the *S. avenae* genome and ~ 30 × coverage of the *R. padi* genome. The final assembly of *S. avenae* (Siave_v2.1) spans 369 Mb with a contig N50 size of 40 kb (Table 1). BUSCO and K-mer analysis shows that the assembly is highly complete with

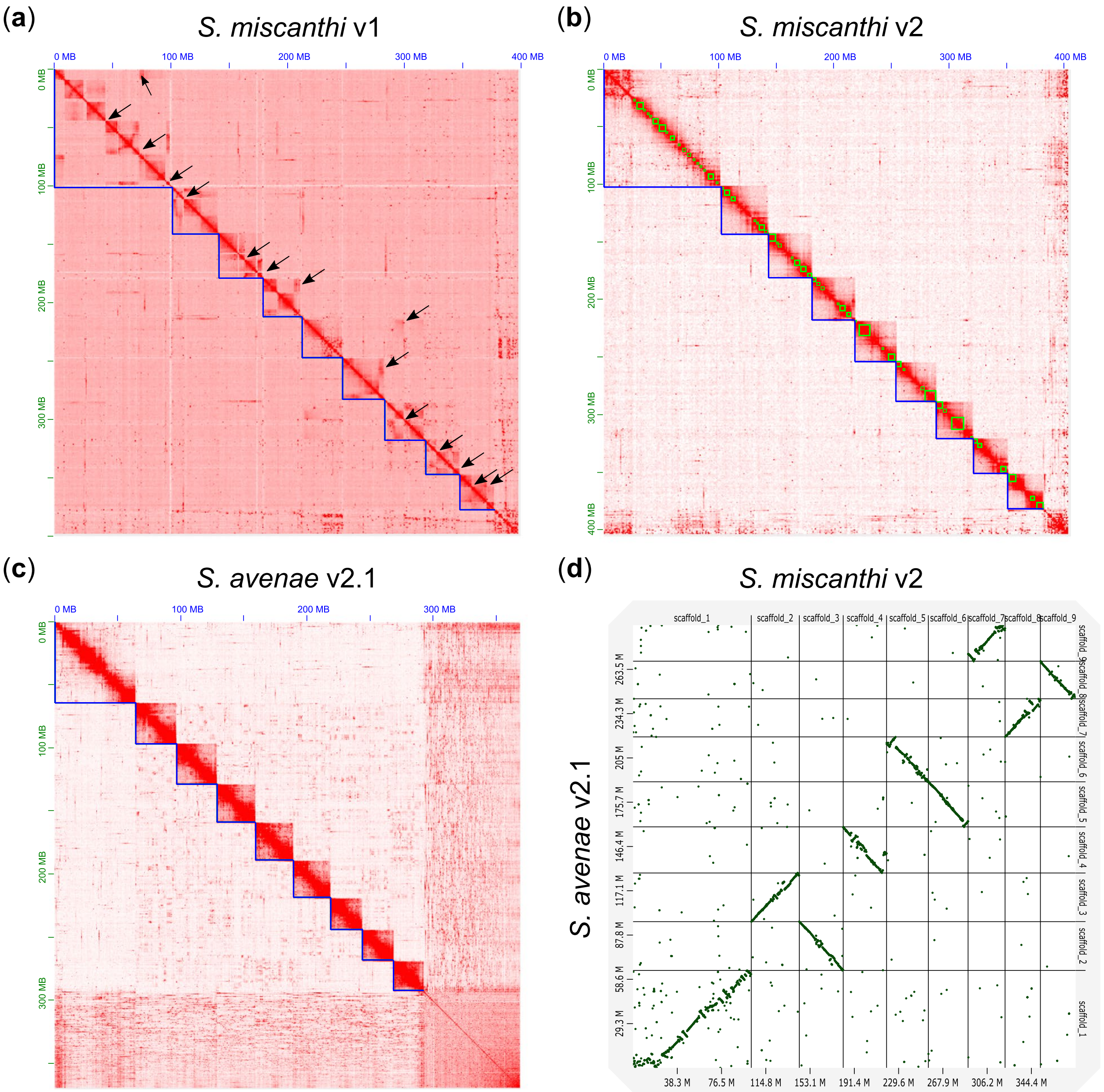

Fig. 1 New chromosome-scale assemblies of *S. miscanthi* and *S. avenae* are high quality and show conserved synteny. **a** Hi-C contact map for the *S. miscanthi* v1 genome assembly. Blue lines show chromosome-scale super scaffolds. Arrows indicate likely assembly errors. Genomic scaffolds are ordered from longest to shortest with the *x*- and *y*-axis showing cumulative length in millions of base pairs (Mb). **b** Hi-C contact map for the *S. miscanthi* v2 genome assembly. Green lines show contigs. **c** Hi-C contact map for the *S. avenae* v2.1 genome assembly. **d** *Dot plot* showing a MashMap [41] whole genome alignment between the *S. miscanthi* v2 and *S. avenae* v2.1 genome assemblies. For clarity, only chromosome-scale scaffolds are included. Scaffolds in each assembly are ordered from longest to shortest. The *x*- and *y*-axis show cumulative scaffold length in Mb

little missing or duplicated content (Additional File 2: Figures S2 and S5) and a taxon-annotated GC content-coverage plot indicates that Siave_v2.1 is free from obvious contamination (Additional File 2: Figure S6). The shorter total assembly size of Siave_v2.1 compared to Simis_v2 (369 Mb vs 403 Mb) is likely the result of missing and collapsed repeat content due to the use of short Illumina reads for de novo assembly. Despite high fragmentation, scaffolding with Hi-C enabled the placement of 74% of the *S. avenae* assembly onto nine chromosome-scale super scaffolds (Fig. 1c). Whole genome alignment of Siave_v2.1 and Simis_v2 chromosome-scale scaffolds reveals broad structural agreement between the two independent assemblies, supporting the accuracy of our multi-species Hi-C scaffolding approach (Fig. 1d). We also scaffolded our draft assembly of *R. padi* using the same multi-species Hi-C library, placing 95% of the draft assembly onto four chromosome-scale super scaffolds (Additional File 2: Figure S7) corresponding to the expected *R. padi* karyotype (2*n* = 8 [44]). We make our new *R. padi* assembly available here, but it will be described in more detail elsewhere.

Finally, to aid comparative genome analysis, we generated a short-read genome assembly of the rose-grain aphid, *Metopolophium dirhodum*, which is thought to be closely related to *Sitobion* [45]. Following the low-cost genome assembly approach set out in Mathers et al. [46], we generated 23.9 Gb (62 × coverage) of PCR-free Illumina genome sequence data and 14.8 Gb of strand-specific RNA-seq data for genome assembly scaffolding and genome annotation (Additional File 1: Table S1). We assembled these data into 24,973 scaffolds spanning 387 Mb (Medir_v1.1; Table 1). Although fragmented (scaffold N50 = 49 kb), this short-read assembly is highly complete at the gene level (BUSCO complete = 97.1%; Additional File 2: Figure S2) and K-mer analysis reveals the absence of excessive missing or duplicated genome content (Additional File 2: Figure S8). The assembly is also free from obvious contamination based on a taxon-annotated GC content-coverage plot (Additional File 2: Figure S9).

All three new grain aphid genome assemblies (Simis_v2, Siave_v2.1 and Medir_v1.1) were annotated using the same gene annotation pipeline that we previously applied to the model aphid species *A. pisum* and *Myzus persicae* [37], incorporating evidence from RNA-seq data (Additional File 1: Table 1). In total, we annotated 21,798 genes (23,875 transcripts) in *S. miscanthi*, 19,919 genes (22,368 transcripts) in *S. avenae* and 22,349 genes (24,826 transcripts) in *M. dirhodum* (Table 1). We also annotated our new chromosome-scale assembly of *R. padi* using the same procedure, identifying 16,977 genes (19,137 transcripts). Compared to Simis_v1, our annotation of Simis_v2 identifies an additional 5792 genes. This large increase in gene count is likely due to a combination of improved genome assembly completeness in Simis_v2 and the use of different gene annotation pipelines. Indeed, BUSCO analysis of the Simis_v1 and Simis_v2 gene sets reveals that Simis_v2 is substantially more complete than Simis_v1 (Additional File 2: Figure S10), with completely missing BUSCO genes reduced by 55% in Simis_v2 (*n* = 68 vs *n* = 28). Furthermore, RNA-seq pseudoalignment rates of the mixed whole-body sample of *S. miscanthi* from Jiang et al. [36] to the annotated gene models increased from 78% in Simis_v1 to 83% in Simis_v2 (Additional File 3: Table S2). Taken together, our highly complete genome assemblies and annotations of *S. miscanthi*, *S. avenae* and *M. dirhodum* provide a solid foundation to study grain aphid biology and complement two other contig-level long-read genome assemblies of *S. avenae* (clone SA3 [47] and SaG1 [48]) and a chromosome-scale assembly of *M. dirhodum* [49] that were published during the completion of this study.

Genome evolution in dactynotine aphids

Increasing numbers of sequenced aphid genomes are allowing finer scale analysis of aphid genome evolution (e.g. Julca et al. [50]). To place our new assemblies of *S. miscanthi*, *S. avenae* and *M. dirhodum* in a phylogenetic context, we clustered their proteomes with eight other aphid species from the aphid tribes Macrosiphini and Aphidini (Additional File 4: Table S3). In total, we clustered 270,894 proteins into 23,712 orthogroups (gene families) and 19,653 singleton genes (Additional File 5: Table S4). Maximum likelihood phylogenetic analysis based on a concatenated alignment of 5091 conserved single-copy genes produced a fully resolved species tree that places *Sitobion* and *Metopolophium* in a monophyletic group that is closely related to *A. pisum* (Fig. 2a; Additional File 2: Figure S11). These findings are in agreement with previous studies using small numbers of loci and larger sets of taxa [45, 51].

Assembly of three genomes closely related to the model aphid *A. pisum*—all belonging to the dactynotine subtribe [51, 55]—allows additional insights into gene family dynamics in aphids. *A. pisum* was the first aphid species to have its genome sequenced and this effort revealed substantial gene family expansion and, at the time, the largest gene count of any sequenced insect species [56]. Subsequent reassembly of *A. pisum* has led to refinement of the estimated gene count for this species, supporting the large number of genes [37, 57]. Here, we find lower gene counts (19,919–22,349 vs 30,784) and smaller genome sizes (366–403 Mb vs 525 Mb) in *Sitobion* and *Metopolophium* compared to *A. pisum*, indicating that genome expansion is limited to the *A. pisum* lineage and occurred after divergence of the dactynotine common ancestor (Fig. 2a). These results are consistent with flow cytometry-based estimates of genome size for *S. avenae*, *M. dirhodum* and *A. pisum* [58] and a long-read genome assembly of *M. dirhodum* (assembly size = 446 Mb) that was published during the completion of this study [49].

In addition to dynamic gene family evolution, we and others have identified high rates of autosomal chromosome rearrangement in aphids [37, 38]. To investigate chromosome evolution in *Sitobion* and their dactynotine relatives, we identified syntenic genome regions between Simis_v2 and the chromosome-scale assemblies of *A. pisum* and *M. persicae* using MCscanX [59]. This analysis confirms homology and conservation of the aphid X

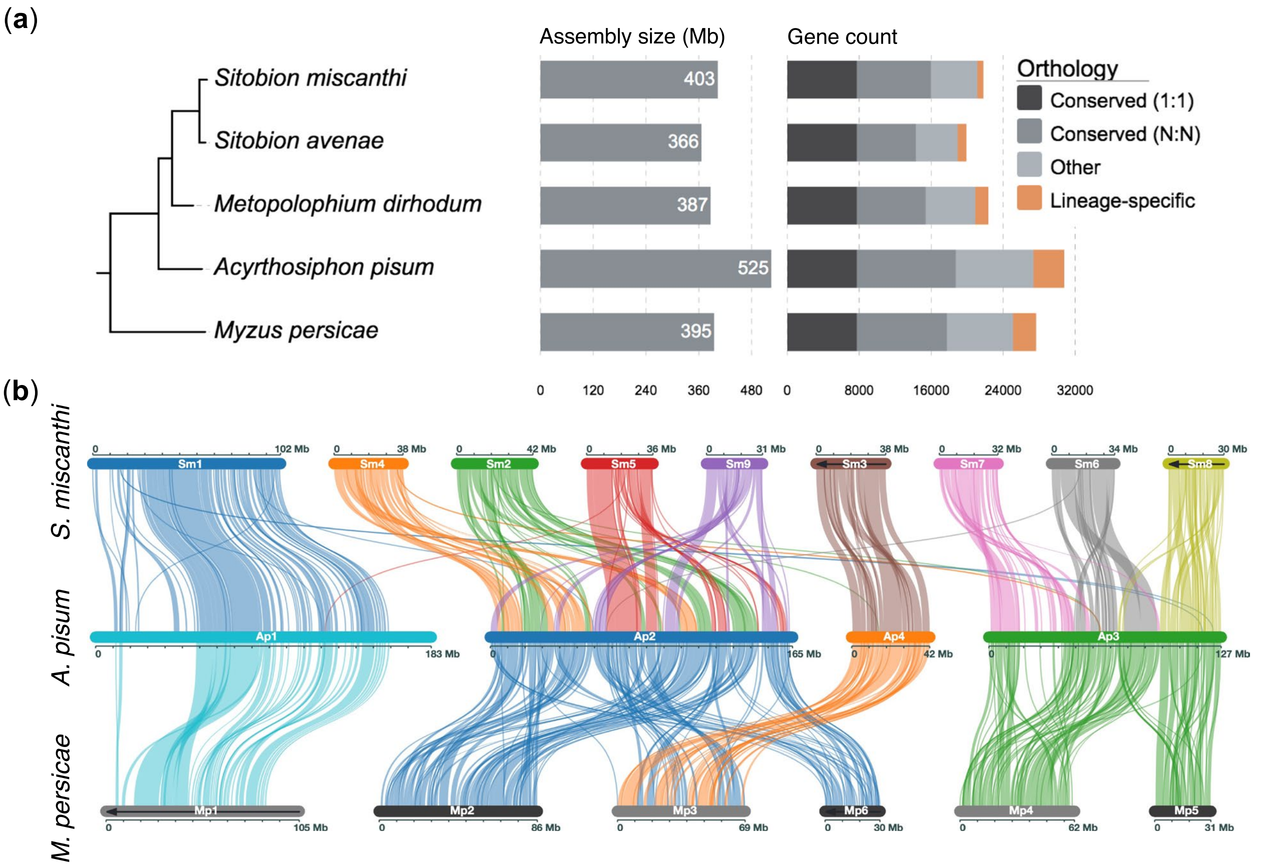

Fig. 2 Comparative genomics of the dactynotine sub-tribe. **a** Maximum likelihood phylogeny of dactynotine aphids based on a concatenated alignment of 5091 conserved single-copy protein coding genes. We included seven outgroup aphid species from Aphidinae but for simplicity only show *M. persicae*. The full phylogeny is shown in Additional File 2: Figure S11. All branches received maximal support according to the Shimodaira-Hasegawa test [52] implemented in FastTree [53, 54] with 1000 resamples. Branch lengths are in amino acid substitutions per site. The tree is annotated with genome assembly length and gene counts coloured by orthology relationships across the full phylogeny. **b** Chromosome evolution in the dactynotine sub-tribe. Plot shows blocks of syntenic genes identified between *S. miscanthi* (top), *A. pisum* (middle) and *M. persicae* (bottom) chromosomes. Chromosomes containing black arrows are visualised as the reverse compliment to aid clarity. Ap1 and Mp1 in *A. pisum* and *M. persicae*, respectively, have previously been identified as the X (sex) chromosome and are homologous to Sm1 (scaffold_1) in our new *S. miscanthi* (Simis_v2) genome assembly

chromosome as a single linkage group within Macrosiphini and reveals substantial autosomal genome reorganisation over the course of dactynotine aphid diversification (Fig. 2b).

So far, the high rate of autosome rearrangement and small number of chromosome-scale genome assemblies have hampered the inference of specific chromosome rearrangement events that have led to extant aphid karyotypes [37]. However, we previously hypothesised that *A. pisum* chromosome 3 (Ap3) was formed by a fusion event involving homologues of *M. persicae* chromosomes 4 (Mp4) and 5 (Mp5) [37]. This scenario is confirmed by alignment of our new *S. miscanthi* assembly with *A. pisum* and *M. persicae*, with *S. miscanthi* chromosome 8 (Sm8) and *M. persicae* chromosome 5 (Mp5) sharing synteny with the final third (orientation as per the *A. pisum* JIC1 assembly) of Ap3 (Fig. 2b). Additionally, the alignment of Mp4 and the first two thirds of Ap3 to *S. miscanthi* chromosome 7 (Sm7) and chromosome 6 (Sm6) reveals that Sm7 and Sm6 were formed by a chromosome fission event in the *Sitobion* lineage. As such, we can infer at least one unambiguous chromosome fusion event in the *A. pisum* lineage and one unambiguous chromosome fission event in the *Sitobion* lineage. Interestingly, we also find that the two large autosomes found in pea aphid (Ap2 and Ap3) exhibit distinct rearrangement patterns. Ap2 is homologous to four small (42–31 Mb) chromosomes in *S. miscanthi*, the order of which is shuffled in *A. pisum*. In contrast, Ap3 is homologous to Sm6, Sm7 and Sm8, which each align to distinct territories along Ap3. In the future, additional chromosome-scale assemblies of *S. miscanthi* and *A. pisum* close relatives will further illuminate the complex history of aphid chromosome evolution.

Low genome-wide divergence between *S. miscanthi* and *S. avenae*

Phylogenetic analysis revealed a short branch length between *S. miscanthi* and *S. avenae* indicating recent divergence from a common ancestor (Fig. 2a). To further investigate genome-wide patterns of sequence divergence among *Sitobion* aphids and their relatives, we generated a reference-free, four-way, whole genome alignment between *S. miscanthi*, *S. avenae*, *M. dirhodum* and *A. pisum* using Progressive Cactus [60]. Alignment coverage of our highly complete long-read-based *S. miscanthi* genome assembly ranged from 72 to 93% (Fig. 3a panel 1) whereas alignment coverage of *S. avenae* was slightly higher (75 to 97%; Fig. 3a panel 2), likely due to lower representation of hard to align repetitive regions in this short-read assembly. Using these alignments, we estimated pairwise sequence divergence in 1 Mb fixed windows along *S. miscanthi* chromosome-scale scaffolds (Fig. 3b). We used alignments anchored to autosomal chromosomes to estimate genome-wide sequence divergence as aphid X chromosomes are known to exhibit elevated substitution rates [61, 62]. Using these autosomal alignments, we estimate median sequence divergence between *S. miscanthi* and *S. avenae* to be only 0.78% (Fig. 3c). In contrast, *S. miscanthi* divergence from *M. dirhodum* and *A. pisum* is 5.63 and 8.86%, respectively.

Phased genome assemblies reveal hybrid origins of the *S. miscanthi* Langfang-1 lab population

Intriguingly, we noticed that the genome-wide divergence between *S. miscanthi* and *S. avenae* (~0.78%) is lower

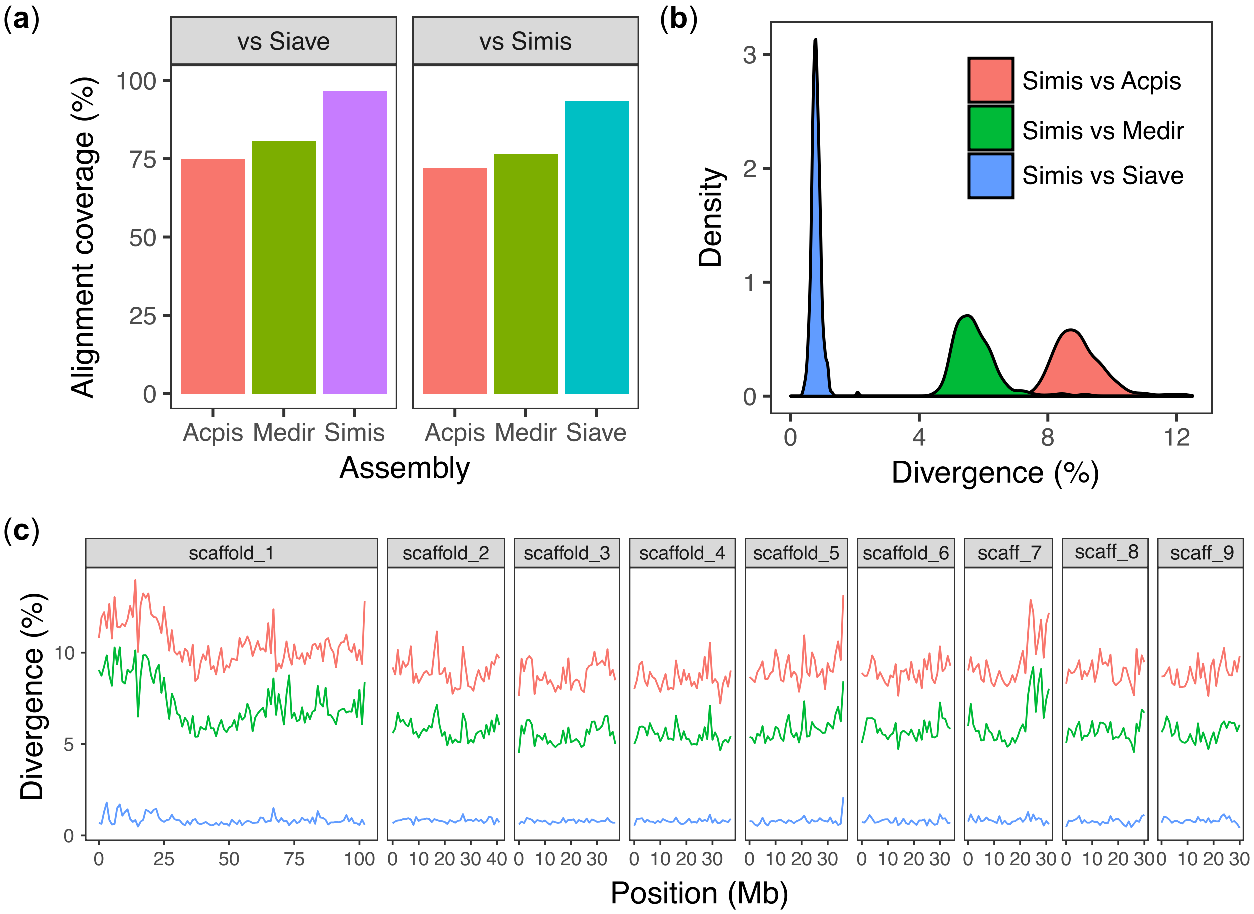

Fig. 3 Reference-free whole genome alignment reveals low sequence divergence between *S. miscanthi* and *S. avenae*. **a** Alignment coverage on either *S. avenae* (Siave) or *S. miscanthi* (Simis) of the *A. pisum* (Acpis), *M. dirhodum* (Medir) and either *S. miscanthi* (Simis) or *S. avenae* (Siave) genome assemblies. **b** Pairwise sequence divergence between *S. miscanthi* and either *A. pisum* (red line), *M. dirhodum* (green line) or *S. avenae* (blue line) in 1 Mb fixed windows along the *S. miscanthi* genome assembly. scaffold_1 = the X (sex) chromosome. **c** *Density plot* showing the distribution of pairwise sequence divergence (as in **b**) for *S. miscanthi* autosomes vs those of *A. pisum* (median divergence = 8.86%), *M. dirhodum* (median divergence = 5.63%) and *S. avenae* (median divergence = 0.78%)

than the predicted diversity (i.e. heterozygosity) found within the previously sequenced *S. miscanthi* clonal lineage dubbed Langfang-1 (LF1) (0.98% based on k-mer analysis of short reads [36]). To further investigate intra- and inter-individual patterns of sequence divergence in *S. miscanthi* and *S. avenae*, we reconstructed independent phased haplotypes for each assembly and generated a four-way whole genome alignment with sibeliaZ [63]. In total, we phased 3,043,224 (>99.99%) and 1,033,184 (97.68%) heterozygous single-nucleotide polymorphisms (SNPs) and small indel variants on *S. miscanthi* (LF1) and *S. avenae* (JIC1) chromosomes, respectively. Due to the inclusion of long-range phase information from our Hi-C data, a single-phase block covered ≥ 99.96% of each chromosome in both JIC1 and LF1 (Additional File 6: Table S5).

Using our whole genome haplotype alignment, we estimated pairwise divergence in 100 kb fixed windows along the *S. avenae* reference genome. Genome-wide within-individual haplotype divergence (heterozygosity) is significantly lower in JIC1 than in LF1 (0.32% ± 0.0039% [mean ± *SE*] vs 0.83% ± 0.016%; Wilcoxon signed-rank test, $p < 2.2 \times 10^{-16}$; Fig. 4a). This is possibly caused by an extreme founder event and/or inbreeding in the UK *S. avenae* population [35]. Consistent with this, we find mega base-scale stretches of near-zero haplotype divergence (i.e. long runs of homozygosity) on several *S. avenae* JIC1 chromosomes (Fig. 4b; Additional File 2: Figure S12). Surprisingly, the two haplotypes found within LF1 substantially differ in their divergence from both JIC1 haplotypes, with one haplotype diverged by ~ 0.5% and the other by ~ 1.0% (Fig. 4a). This unusual pattern of haplotype divergence is maintained across most chromosomes without any "switching" of haplotypes that would be expected if recombination had taken place (Fig. 4b; Additional File 2: Figure S13). As such, we hypothesise that the LF1 clonal lineage is a "frozen hybrid", in particular, a first-generation (F1) clonal descendant of a cross between two lineages that differ in their divergence from *S. avenae*.

S. miscanthi and *S. avenae* are part of a cryptic species complex

To investigate the origins of the LF1 clone and gain a greater understanding of population-level divergence and diversity in *S. avenae* and *S. miscanthi*, we reanalysed genotyping by sequencing (GBS) data for 100 *S. miscanthi* individuals from China and 119 *S. avenae* individuals from the UK from a recent study by Morales-Hojas et al. [35] (Additional File 7: Table S6). Previously, these data were analysed separately using different GBS protocols. However, given our finding that sequence divergence between *S. avenae* and *S. miscanthi* is very low, we reanalysed these data together and searched for overlapping SNP markers between the two sets of samples. In total, including variants from the JIC1 and LF1 whole genome samples, we identified 3,246,566 biallelic SNPs (min. depth ≥ 2 and site quality ≥ 30). Of these sites, we retained 4359 that were covered (and called) in at least 75% of samples. We further refined the dataset by removing 60 samples that had more that 30% missing data. The final dataset contains markers spread across all nine *Sitobion* chromosomes ($n = 332–978$ per chromosome; Additional File 8: Table S7) and includes 149 samples, 52 from the UK and 97 from China, allowing us to investigate diversity and differentiation within and between *S. avenae* and *S. miscanthi* populations.

Previously, Morales-Hojas et al. [35] identified six highly differentiated *S. miscanthi* populations in China (genome-wide FST = 0.13–0.79). These results are recapitulated in our analysis using the shared SNP set, with highly similar groupings based largely on geographic location (Fig. 5a). Furthermore, the LF1 and JIC1 whole genome samples cluster within their expected populations based on geography, i.e. LF1 groups with the Langfang Chinese *S. miscanthi* samples and JIC1 groups with the UK *S. avenae* samples. Surprisingly, by combing data for *S. avenae* and *S. miscanthi*, we find that the UK *S. avenae* population is more closely related to one of the Chinese *S. miscanthi* populations (TG_YC). This suggests that *S. avenae* sensu stricto may be part of a larger cryptic species complex that includes multiple diverged *S. miscanthi* lineages.

To gain more insight into the putative hybrid origin of the *S. miscanthi* LF1 sample used for de novo genome assembly (Fig. 4), we phased the shared SNP set using beagle [64] and estimated a SplitsTree [65] network of the resulting haplotypes using data from the longest *S. avenae* autosome, chromosome 2 (Fig. 5b). This analysis reveals that one LF1 haplotype clusters with the Langfang *S. miscanthi* population and that the second falls out on its own between the monophyletic groups containing the UK *S. avenae*-like samples and the *S. miscanthi* KM population. Phased haplotype networks generated from SNP's on the X chromosome and autosomes 3, 5, 8 and 9 also show a similar pattern, with one LF1 haplotype clustering with the Langfang population and the second falling out in the middle of the network (Additional File 2: Figures S13, S14, S15, S17, S20 and S21). However, haplotype networks generated from SNP's on autosomes 4, 6 and 7 show clustering of both LF1 haplotypes with the Langfang population (Additional File 2: Figures S16, S18 and S19)). The similar LF1 haplotypes observed on autosomes 4, 6 and 7 when averaging across entire chromosomes are likely the result of one or a few generations of

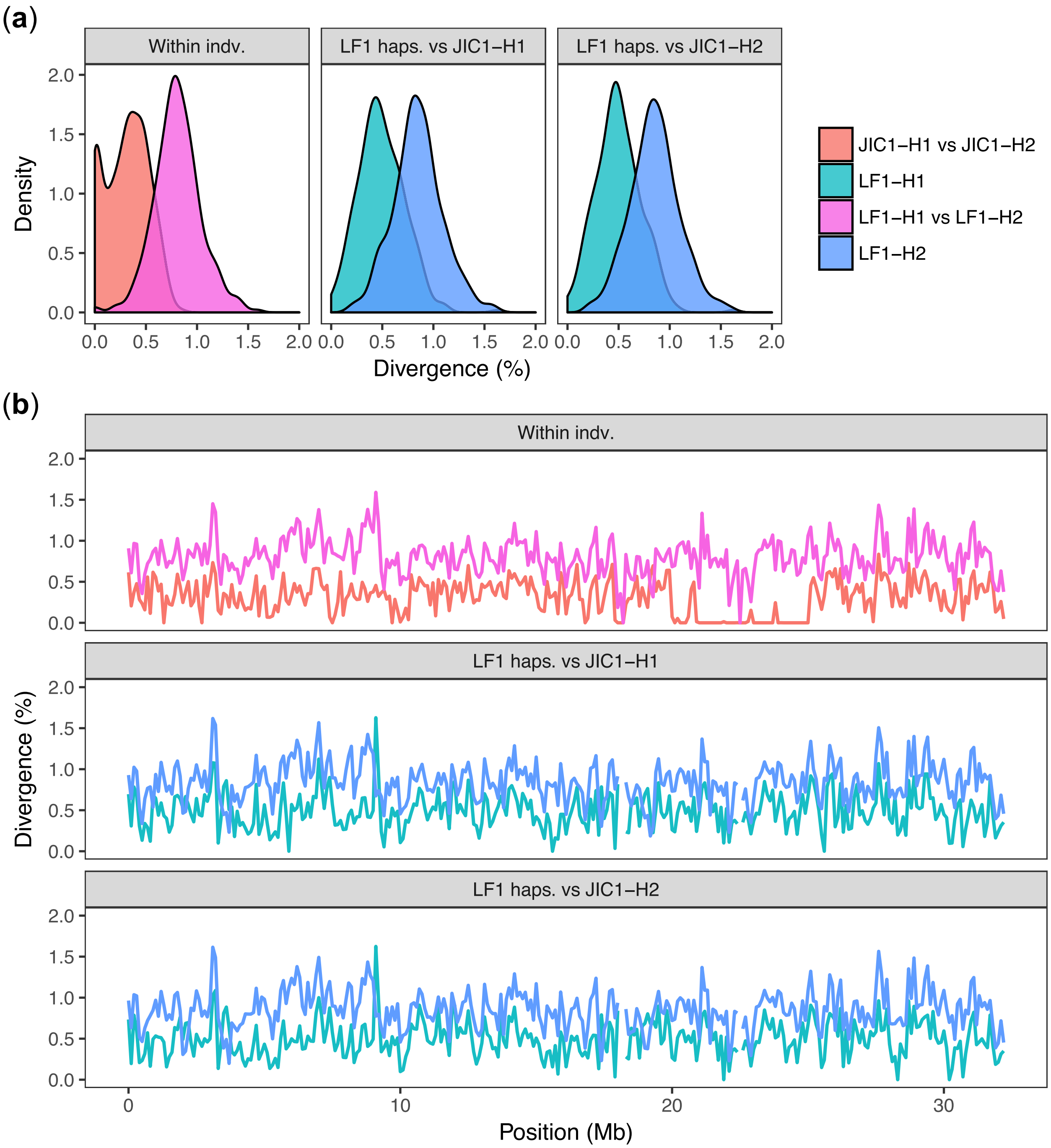

Fig. 4 Highly differentiated haplotypes within the *S. miscanthi* LF1 clonal lineage differ in their in their divergence from *S. avenae*. **a** Sequence divergence distribution within and between phased haplotypes of JIC1 and LF1 for *S. avenae* chromosome (scaffold) 2. Comparing the LF1 haplotypes (LF1-H1 and LF1-H2) to either JIC1 haplotype reveals distinct patterns of divergence, with LF1-H1 diverged ~ 0.5% from either JIC1 haplotype, and LF1-H2 diverged by ~ 1%. **b** Haplotype divergence within and between JIC1 and LF1 in 100-kb fixed windows along *S. avenae* chromosome (scaffold) 2. The unusual pattern of haplotype divergence observed in LF1 is maintained across the full length of chromosome 2 indicating an absence of recombination. Divergence patterns for all chromosomes are shown in Additional File 2: Figure S12

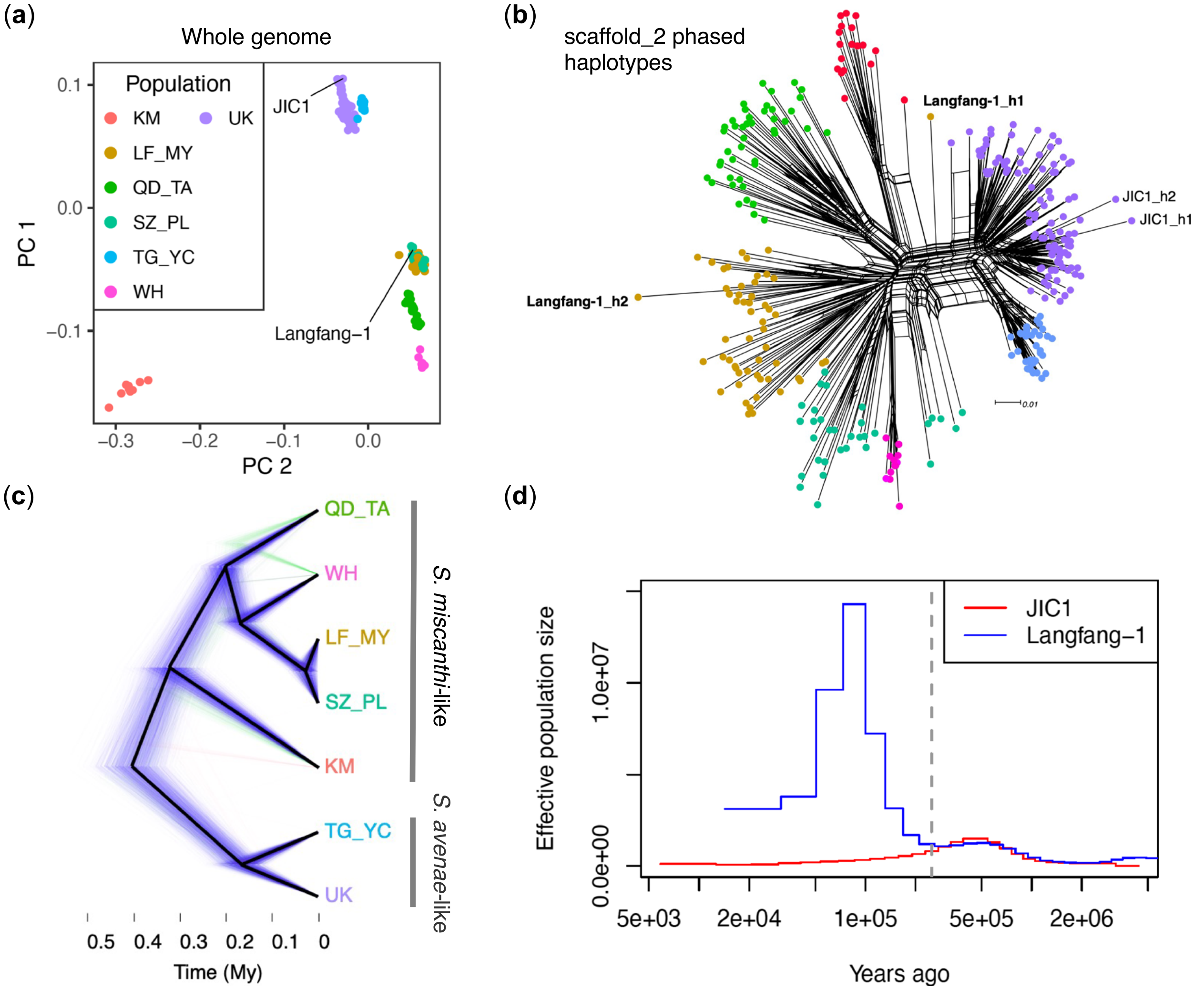

Fig. 5 Population genomics of *Sitobion avenae* and *Sitobion miscanthi* from the UK and China. **a** Principal component analysis based on the thinned (max 1 SNP per 25 kb) shared SNP set (*n* = 1772) reveals previously described population structure in China [35] and the close relationship of the UK *S. avenae* (purple dots) to a Chinese *S. miscanthi* lineage (TG_YC; light blue dots). Samples used for genome assembly of *S. avenae* (JIC1) and *S. miscanthi* (LF1) are highlighted. Populations are named following Morales-Hojas et al. (2020) according to geographic location: KM = Kunming, LF_MY = Langfang and Mianyang, QD_TA = Qingdao and Tai'an, SZ_PL = Suzhou and Pingliang, TG_YC = Taigu and Yinchuan, WH = Wuhan, UK = United Kingdom. **b** SplitsTree network of samples shown in **a** based on phased haplotypes for SNPs on *S. aveane* chromosome 2. The two haplotypes from the *S. avenae* (JIC1) and *S. miscanthi* (Langfang-1) samples used for genome assembly are highlighted. **c** "Cloudogram" showing a time calibrated phylogeny of *Sitobion* lineages rooted with *M. dirhodum* (not shown). A posterior sample of 1801 trees are drawn with the first, second and third most common topologies coloured blue, green and red, respectively. Black lines show the maximum clade credibility tree. Tip labels (showing populations) are coloured according to **a**. **d** Demographic history of the JIC1 (derived from the UK population) and LF1 (derived from the LF_MY population) whole genome sequence isolates estimated with MSMC2. The dashed vertical line indicates the approximate time of divergence between the two samples 250,000 years ago (Kya) when the population histories become clearly separated

backcrossing with the LF_MY parent population. Taken together our results suggest that the LF1 clone is likely the product of a hybridisation event between the Langfang population and another, as yet unsampled, *Sitobion* lineage.

Next, we set out to date the radiation of the *Sitobion avenae* / *miscanthi* complex. The spontaneous mutation rate of the closely related aphid *A. pisum* has recently been inferred, enabling the estimation of species and population divergence times from genome sequence data [66]. However, our (primarily) GBS-based SNP dataset does not include invariant sites and so likely suffers from ascertainment bias, making direct dating based on a known mutation rate challenging [67]. To avoid applying an unrealistic mutation rate to our SNP data, we first estimated the divergence time of the sister genera

Sitobion and *Metopolophium* to provide a calibration point for estimating divergence times within *Sitobion*. Using 8462 autosomal phylogenetically inferred single-copy orthologs between *S. miscanthi* and our new genome sequence of the rose-grain aphid (*M. dirhodum*), we estimated the most recent common ancestor (MRCA) of *Sitobion* and *Metopolophium* to have accrued 7.27 million years ago (Mya) based on median autosomal synonymous site divergence of 5.9% (Additional File 2: Figure S22) and the *A. pisum* spontaneous mutation rate of 2.7×10^{-10} per haploid genome per generation and an average of 15 generations (14 asexual and 1 sexual) per year [68]. We then mapped genomic reads for *M. dirhodum* to the *S. avenae* reference genome and called SNPs alongside the *Sitobion* GBS and whole genome sequence samples, identifying 3043 shared biallelic SNPs.

Using the SNP dataset and calibration point inferred above, we jointly estimated a population tree and divergence times under the multi-species coalescent (MSC) with SNAPP [69] following recommendations by Stange et al. [67]. This analysis recovers a well-supported tree (Fig. 5b; Additional File 2: Figure S23) that places the UK *S. avenae*-like population in a monophyletic group with the Chinese TG_YC population (Bayesian posterior probability (BPP) = 1) and the remaining *S. miscanthi*-like Chinese populations in second monophyletic group (BPP = 0.99). The deep split between the *S. miscanthi*-like and *S. avenae*-like groups is inferred to have occurred around 404 thousand years ago (Kya) (95% highest posterior density (HPD) = 328 – 464 Kya). Within the *S. miscanthi*-like group, the KM population forms a third highly differentiated lineage that diverged around 323 Kya (95% HPD = 274 – 383 Kya), although there is some support for an alternative topology which places this group as an outgroup to all other included *Sitobion* lineages pushing back the inferred split time. Within the *S. avenae*-like lineage, we estimate that the UK lineage diverged from the Chinese TG_YC lineage around 165 Kya (95% HPD = 124–213 Kya).

To provide an independent estimate of the primary split time between the *S. avenae*-like lineage and the *S. miscanthi*-like lineage based on whole genome sequences (rather than GBS samples), we estimated the past demographic history of the JIC1 and LF1 clonal lineages using MSMC2 [70] which implements the multiple sequentially Markovian coalescent (MSMC) model. This analysis indicates that JIC1 and LF1 demographic histories begin to diverge around 500 Kya ago and are fully separated by approximately 250 Kya (Fig. 5d), providing a slightly earlier estimate of divergence between the *S. avenae*-like and *S. miscanthi*-like lineages than the SNPAPP analysis (Fig. 5c). Furthermore, the LF1 sample shows a very sharp rise in effective population size around 100 Kya in Fig. 5d. Typically, in MSMC models of hybrids, the effective population size goes to infinity at the time when gene flow stopped between the parental lineages [71, 72]. The sharp rise in effective population size may therefore represent the divergence time of the LF_MY lineage from the second unidentified lineage that contributed to the LF1 hybrid. The failure to reach infinity may reflect a complex history of admixture between the parental lineages, or backcrossing following the initial hybridisation event. This could retain a signal of coalescence in small parts of the genome, stopping MSMC from estimating an infinite effective population size. Additional sequencing of Chinese *Sitobion* populations will likely shed further light on these processes. Nonetheless, our analyses reveal multiple highly differentiated lineages within the *S. miscanthi* / *avenae* complex that substantially predate the origins of agriculture.

Hybridisation has shaped the Sitobion radiation

Finally, given the putative hybrid origins of the *S. miscanthi* LF1 lab population, we asked whether hybridisation has occurred more widely between lineages in the *S. avenae* / *miscanthi* complex. Using *M. dirhodum* as an outgroup, the SNAPP species/population tree, and excluding the previously identified Langfang-1 hybrid isolate, we summarised admixture across the *S. avenae* / *miscanthi* complex using Patterson's *D* (ABBA–BABA test) [73] and the *f*-branch (f_b) statistic [74], both of which use patterns of allele sharing to infer gene flow. First, to gain a course and conservative overview of admixture history across the complex, we calculated the minimum *D* statistic (D_{min}) irrespective of phylogeny for all trios of ingroup lineages ($n = 35$) [74]. In total, 34% of trios have significant D_{min} (Bonferroni corrected $p < 0.05$; Additional File 9: Table S8), indicating moderate levels of admixture among members of the *S. avenae* / *miscanthi* complex.

Significant D_{min} may be caused by current admixture (i.e. hybridisation) between extant taxa, or by historical admixture between ancestral lineages. Historic admixture is expected to inflate the number of trios with significant D_{min} due to non-independence. To map admixture to specific lineages of the *S. avenae* / *miscanthi* complex, we calculated the f_b statistic which summarises all possible f_4 admixture ratios for a given phylogeny and reports the admixture proportion between all compatible pairs of branches and taxa in a phylogeny [74]. We find the largest admixture proportions between WH and the QD_TA ($f_b = 0.57$) and KM ($f_b = 0.28$) lineages (Fig. 6; Additional File 9: Table S8). There are also strong bidirectional signatures of admixture between the *S. avenae*-like group and the *S. miscanthi*-like group. In particular, admixture is detected between the TG_YC *S. avenae*-like

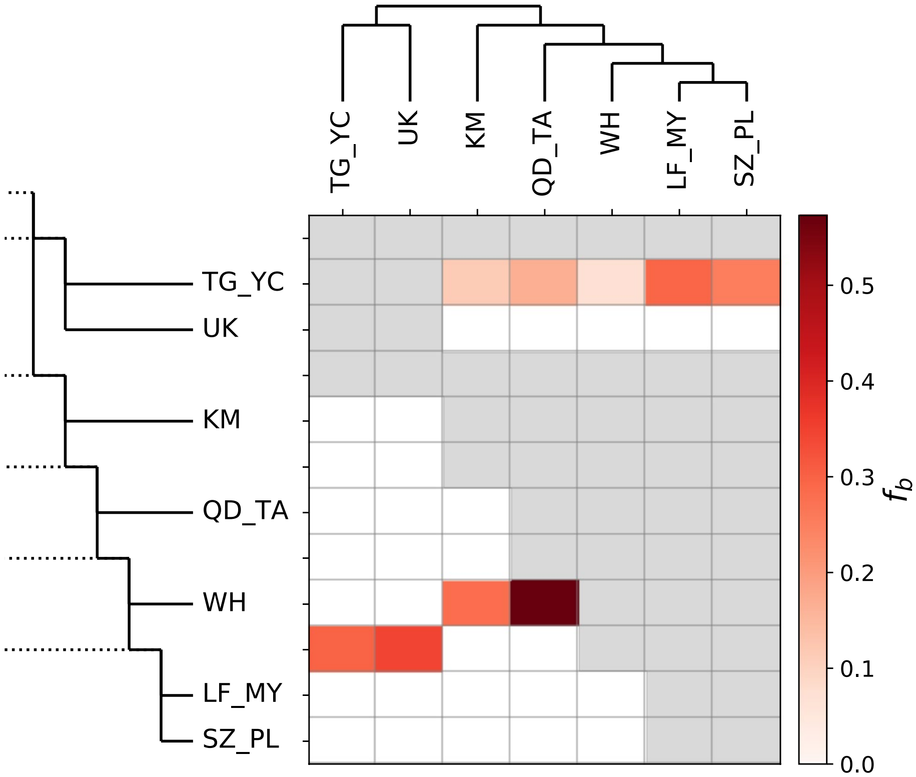

Fig. 6 Excess allele sharing between *Sitobion* lineages. The *heatmap* shows the magnitude of the f_b ratio [74] between each branch on the *y*-axis and the sample on the *x*-axis. Grey squares indicate comparisons that cannot be made. Comparisons where corresponding *D* statistics (Additional File 9: Table S8) are non-significant ($p > 0.01$) are set to zero. Tip names correspond to populations/lineages identified in Fig. 5a

lineage in China and among all members of the *S. miscanthi*-like group ($f_b = 0.07–0.29$; LF_MY strongest), and also between the common ancestor of *S. avenae*-like LF_MY + SZ_PL. This latter signal may reflect admixture event(s) prior to the divergence of the two sampled *S. avenae*-like lineages. However, better sampling and an improved understanding of global *S. avenae / S. miscanthi* diversity and phylogeography will be required to better understand specific gene flow events within the complex.

Next, to investigate admixture across the *S. avenae / miscanthi* complex at greater resolution, we generated a second SNP dataset containing only the Chinese GBS samples and the two whole genome sequences of Langfang-1 and JIC1 ($n = 98$). As all the Chinese samples derive from the same GBS experiment, we were able to recover a much larger set of SNPs enabling fine-scale introgression analysis across the genome. In total, we identified 73,903 SNPs that are present in at least 90% of samples. A phylogenetic network based on this larger SNP set reveals large reticulations indicating substantial admixture (or hybridisation) between lineages of the *Sitobion miscanthi / avenae* complex (Fig. 7a), consistent with our previous analysis based on the smaller SNP set (Fig. 6).

Using the large SNP set, we investigated genome-wide patterns of introgression focussing on the WH, QD_TA and KM lineages, which are estimated to have the highest rates of hybridization based on the f_b statistic (Fig. 6). We estimated phylogenetic trees in 50 SNP windows for all samples and summarised the distribution of all possible topologies among the three focal lineages and the UK *S. avenae* lineage using topology weighting by iterative sampling of subtrees (Twisst; [75]). As expected, Twisst recovers the SNAPP species topology as the highest weighted topology genome-wide (topo2: Fig. 7b)—this tree places the KM lineage as a sister to the WH and QD_TA lineages. However, the second most common topology (topo3) also receives high weighting and groups the KM and WH lineages together with QD_TA as an outgroup (Fig. 7b). This pattern is consistent with admixture between the WH and KM lineages (or their ancestors).

Surprisingly, the distribution of the three topologies across the genome is non-random with large regions of each chromosome predominantly having either the species tree (topo2) or the admixture tree (topo3) (Fig. 6c). A striking example is found on chromosome 2 where a clear

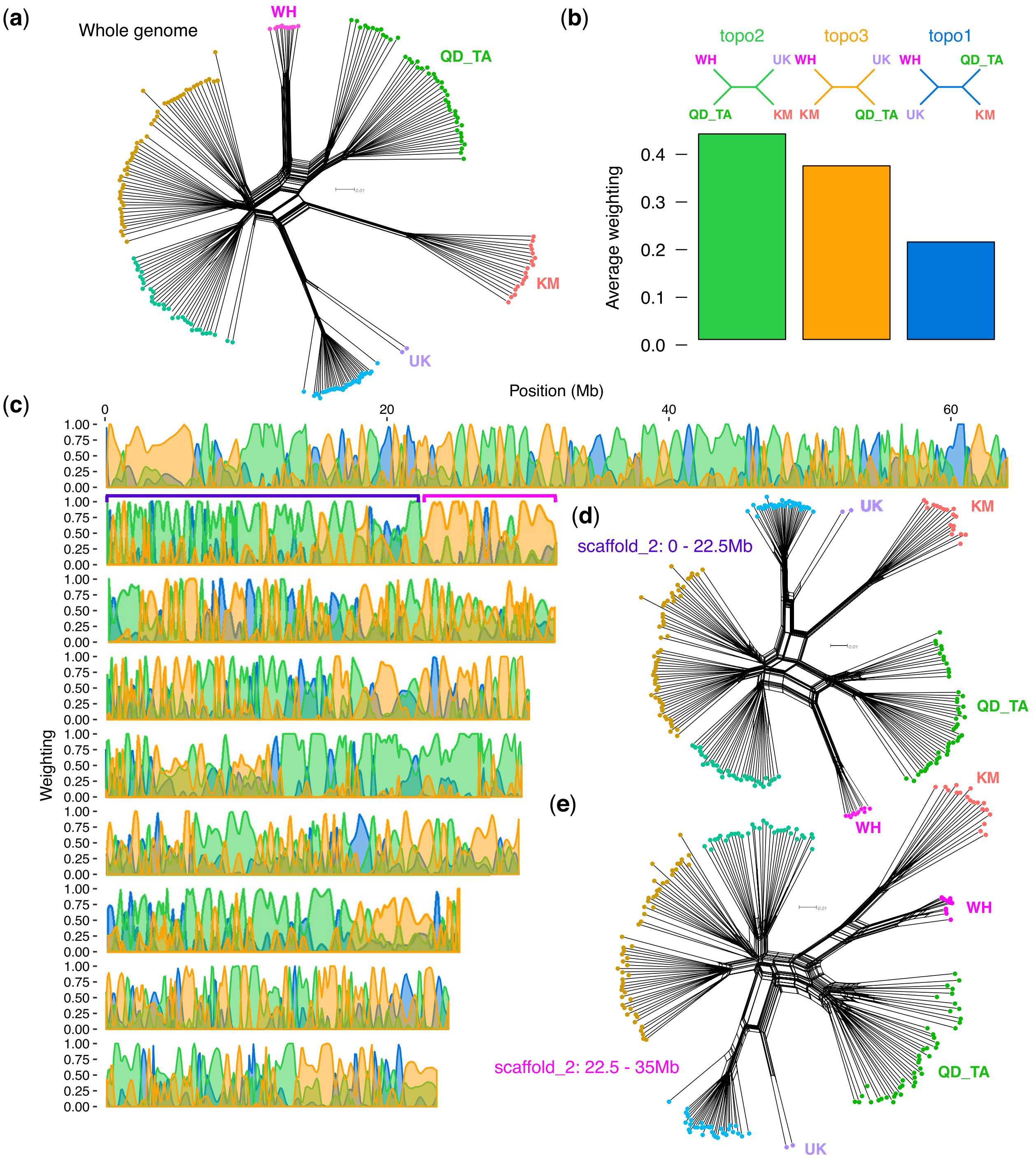

Fig. 7 Topology weighting reveals massive introgressed blocks. **a** SplitsTree [65] network of phased haplotypes based on 73,903 genome-wide SNPs for Chinese GBS samples (Additional File 7: Table S6) plus the whole genome sequences of *S. avenae* JIC1 from the UK and *S. miscanthi* LF1 from the LF_MY population. Tips are coloured by population/lineage as per Fig. 5. Three focal *S. miscanthi*-like lineages (WH, QD_TA and KM) and one outgroup *S. avenae* lineage (UK) are highlighted. **b** Phylogenetic trees were estimated in 50 SNP windows across all *S. avenae* chromosomes for all samples included in **a**. The *bar chart* summarises the average genome-wide weighting (determined by Twisst [75], see *main text*) of the three possible topologies between WH, KM, QD_TA and UK focal lineages. The tree with the highest weighting (topo2) corresponds to the species tree estimated with SNAPP (Fig. 5c). Alternative topologies are also highly weighted. **c** The distribution of topology weightings across all nine *S. avenae* chromosomes reveals large blocks with a different evolutionary history to the species tree. Example regions on chromosome (scaffold) 2 are highlighted with purple (predominantly weighted towards the species tree (topo2)) and pink (predominantly weighted towards topo3). **d**, **e** show SplitsTree networks of phased haplotypes for the example regions of chromosome 2 highlighted in **c**. For each network, the focal lineages are indicated

switch point can be seen at ~ 22.5 Mb, with the beginning of the chromosome (*region 1*) strongly weighted towards the species tree, and the final 12.5 Mb (*region 2*) strongly weighted towards the admixture tree. Network analysis based on SNPs from either of these two regions shows that the KM individuals group closely to the *S. avenae*-like group (UK + TG_YC) in *region 1* (Fig. 7d), but that they group more closely to the WH individuals in *region 2* (Fig. 7e). Given the very large blocks of alternative ancestry found in across all chromosomes and that all sampled members of the KM population share the admixed regions, we speculate that the KM lineage may have been formed by a hybridisation event that was followed by very low levels of backcrossing with its parent populations.

Discussion

It has been debated whether *S. avenae* and *S. miscanthi* should be considered separate species (reviewed in Choe et al. [31]). The high-quality chromosome-scale genome assemblies generated here, and our analysis of sequence divergence patterns, revealed that *S. avenae* and *S. miscanthi* have less than 1% genome-wide sequence divergence (Fig. 3), and hence, that they are closely related. Surprisingly, we found that the *S. miscanthi* isolate used for genome assembly is likely an F1 or recent hybrid, with one of its haplotypes being more closely related to *S. avenae* than the other (Fig. 4). Interestingly, the haplotype most closely related to *S. avenae* appears to be from an unsampled *S. miscanthi*-like "ghost lineage" [76–78], hinting at the presence of significant as yet undescribed *S. miscanthi* diversity in China. Indeed, rather than being a two species system, *S. avenae* and *S. miscanthi* appear to belong to a species complex with multiple highly differentiated lineages that predate modern agriculture, having diverged between 404 and 27 thousand years ago based on coalescent analysis (Fig. 5). The lineages in this species complex have a highly reticulated evolutionary history (Fig. 6) with evidence of hybridization, particularly among three Chinese lineages where we find large blocks of chromosomes with alternative ancestry shared by all members of each lineage (Fig. 7). We hypothesise that hybrid speciation is responsible for the evolution of new lineages in the *Sitobion* genus.

The unusual reproductive mode employed by aphids—cyclical parthenogenesis—may enable occasional high fitness combinations of parental lineages to rapidly expand through clonal propagation, effectively freezing a mosaic genome architecture as observed in Chinese *S. miscanthi*-like populations in this study. By avoiding or minimising sexual recombination after the hybridisation event, the descendants of hybrid aphids do not suffer from hybrid-vigour breakdown. Such breakdown leads to a fitness loss in sexual descendants, which is caused by segregation of the initially heterozygous loci and the breakdown of positive epistatic interactions [79, 80]. Therefore, aphids may capitalise on the initial advantage of hybridisation (increasing variation and masking the genetic load [81]), whilst their ability to proliferate asexually after the event reduces future fitness costs. These factors may increase the likelihood of hybrid lineage formation in aphids. Indeed, the clonal proliferation of "frozen hybrid" lineages may be pervasive across aphid evolution as signatures of ancient hybridisation have been recently detected in a phylogenomic analysis of Aphididae [82].

Finally, the genetic exchange between previously isolated lineages may play an important role in biological invasions of pests in our increasingly globalised world [15, 83, 84]. The benefits that hybridisation offers is preserved in clonal lineages, which gives clonal reproduction in facultative sexuals two important advantages. First, clonal reproduction increases colonising ability of pest species by not being reliant on conspecifics for reproduction. Second, hybridisation between diverged lineages generates novel genotypic variation that is important in adaptive evolution. Both advantages are currently capitalised on by many pest species because they can take advantage of human-mediated transport to colonise new habitats and host species. In addition, given that the hosts in agriculture tend to have little genetic diversity, a single successful genotype of a parasite or pest could infect an entire crop [1, 4, 85]. However, our current study shows that this is not an evolutionary scenario unique to modern times, but that aphids may have exploited these evolutionary advantages of high mobility and cyclical parthenogenesis long before the advent of agriculture.

Conclusions

We have used comparative genomics and population genetics to dissect the evolutionary history of the English and Indian grain aphids, *Sitobion avenae* and *Sitobion miscanthi*—two important pests of cereal crops. Our analyses reveal that both species are closely related and belong to a diverse species complex that predates modern agriculture. The grain aphid complex has a highly reticulated evolutionary history, and hybridisation appears to have driven the emergence of new lineages. We propose that aphids are particularly well placed to exploit hybridisation events via the rapid propagation of live-born "frozen hybrids" via asexual reproduction, increasing the likelihood hybrid lineage formation. As such, hybridisation has likely contributed to the success of aphids in the past and may pose future threats to agriculture.

Methods

Genome assembly approach and quality control

For this study, we generated de novo genome assemblies of *S. miscanthi*, *S. avenae*, *M. dirhodum* and *R. padi* using a variety of sequencing approaches detailed in the sections below. Regardless of the method used, we aimed to generate high-quality haploid genome assemblies maximising assembly completeness and minimising the inclusion of erroneously duplicated content (i.e. haplotigs). Genome assemblies were assessed by generating K-mer spectra, a procedure that involved comparing K-mer content of the raw sequencing reads to the K-mer content of the genome assembly with the K-mer analysis toolkit (KAT [86];). We also assessed assembly completeness and duplication levels by searching for arthropod Benchmarking sets of Universal Single-Copy Orthologs (BUSCOs; $n = 1066$) using BUSCO v3 [87, 88]. BUSCO and K-mer spectra analyses were used throughout the assembly process and to assess the final frozen genome assembly of each species. Each de novo assembly was checked for contamination and the presence of symbiont genomes by generating taxon-annotated GC content-coverage plots (known as a “BlobPlots”) with BlobTools v1.0.1 [89, 90]). Where symbiont genomes were co-assembled with their aphid host, we have included them as separate assembly files as part of the data release for this study. However, symbiont genomes have not been subjected to further curation or quality control.

S. miscanthi v1 genome evaluation

We assessed the quality of the previously published *S. miscanthi* genome (Simis_v1 [36];) using BUSCO and by generating a K-mer spectra comparing the published Illumina short reads (NCBI accession number: SRX5767526) to the genome assembly. To check synteny with the closely related species *A. pisum*, we aligned Simis_v1 chromosome-scale scaffolds to chromosome-scale scaffolds from the *A. pisum* JIC1 v1 assembly [37] using the D-GENIES server [91]. To assess scaffolding quality, we visualised Hi-C contacts across the published genome assembly using Juicebox Assembly Tools (JBAT [40];). Hi-C reads (SRX5767527) from Jiang et al. [36] were mapped to Simis_v1 using Juicer [92] with default settings, and the resulting merged_nodups.txt file was processed using the run-assembly-visualizer.sh script from the 3dDNA assembly pipeline [93].

Reassembly of *S. miscanthi*

We reassembled the *S. miscanthi* genome using sequence data from Jiang et al. [36]. These data include PacBio long reads (SRX5767529; 85 × coverage), Illumina short reads (SRX5767526; 105 × coverage) and in vivo Hi-C data (SRX5767527; 76 × coverage). De novo assemblies of the PacBio long reads were generated with Flye v2.8.1 [94] using default PacBio parameters (“–pacbio-raw”) and wtdgb2 v2.3 [95] with default parameters. The wtdgb2 assembly was subjected to a single round of long-read polishing with wtpoa-cns using minimap v2.14 [96]. PacBio read alignments with the parameter “-ax map-pb”. The Flye and wtdgb2 assemblies were merged using quickmerge v0.3 [97] with the parameters “-l 1,256,119 -ml 10,000 [Flye_assembly_fasta] [wtdgb2_assembly_fasta]”. The “-l” flag was conservatively set to the N50 of wtdbg2 assembly as the Flye assembly had an N50 below 1 Mb (scaffold N50 = 583 kb) and low values of “ − l” may lead to increased misjoins. The Flye assembly was used as the “query” sequence because preliminary analysis showed it to be more complete than the wtdgb2 assembly and quickmerge assemblies predominantly contain sequence content from the “query” assembly. The quickmerge assembly was subjected to a single round of long-read polishing using the Flye polisher and followed by three rounds of short-read polishing with Pilon v1.22 [98]. Redundant haplotigs were removed from the polished assembly using purge_dups [99]. For purge_dups, scaffold coverage was estimated by mapping the PacBio long reads with mininmap v2.16 with the parameter “-x map-pb” and assembly self-alignment was carried out with minimap v2.16 with the parameter “-xasm5 -DP”. Scaffold coverage cutoffs for purge_dups were estimated automatically using the calcuts script.

We scaffolded the draft assembly into chromosome-scale super scaffolds using the published Hi-C data [36]. The Juicer pipeline was used to identify Hi-C contacts and the 3D-DNA assembly pipeline was used for assembly scaffolding (with default parameters), followed by manual curation with JBAT. We found that the Hi-C library had low resolution, resulting in sub-optimal scaffolding performance by 3D-DNA. However, 3D-DNA first orders the input assembly into a single super scaffold before breaking the assembly into putative chromosome-scale fragments. Inspection of the initial round of scaffolding revealed sufficient signal to manually assemble the *S. miscanthi* chromosomes in JBAT (Additional File 2: Figure S 24). The scaffolded assembly was screened for contamination based on manual inspection of “BlobPlots”. Briefly, short reads were aligned to the assembly with BWA mem v0.7.7 [100] and used to estimate average coverage per scaffold. Additionally, each scaffold in the assembly was compared to the NCBI nucleotide database (nt; downloaded 13th October 2017) with BLASTN v2.2.31 [101] with the parameters “-task megablast -culling_limit 5 -evalue 1e-25 -outfmt ’6 qseqid staxids bitscore std sscinames sskingdoms stitle’”. Read mappings and blast results were passed to BlobTools v1.0.1 which was used to generate “BlobPlots” annotated at the order and genus level (Additional File 2: Figure S25 and S26). We removed two scaffolds belonging to the obligate *Buchnera* endosymbiont and additional scaffolds that had low coverage (< 30 × Illumina short-read converge). The remaining scaffolds were ordered by size and assigned a numbered scaffold ID with SeqKit v0.9.1 [102] to create a frozen release for downstream analysis (Simis_v2).

Sequencing and de novo assembly of *S. avenae JIC1* and *M. dirhodum*

S. avenae and *M. dirhodum* individuals were sampled from clonal lineages maintained at the JIC insectary on *Avena sativa* (oats). The *S. avenae* colony (dubbed JIC1) was originally obtained from the University of Newcastle in 2012. The original plant host is unknown. The *M. dirhodum* colony (dubbed UK035) was originally collected from a rose bush in Norwich in 2015.

We followed the procedures described in Mathers et al. [46] to generate low-cost short-read de novo genome assemblies of *S. avenae* and *M. dirhodum*. Briefly, DNA was extracted from a single individual and sent to Novogene (China) where a PCR-free Illumina sequencing library was prepared with a target insert size of 500–1000 bp and sequenced on an Illumina HiSeq 2500 instrument with 250 bp paired-end chemistry. We also extracted total RNA from bulked adult unwinged asexual female individuals from each species and sent it to Novogene for strand-specific library preparation and sequencing on an Illumina platform with 150 bp paired-end chemistry. Genomic reads were processed with trim_galore (http://www.bioinformatics.babraham.ac.uk/projects/trim_galore) to remove adapters with the parameters "–quality –paired –length 150" and then assembled using Discovar de novo (https://software.broadinstitute.org/software/discovar/blog/) with default parameters. Erroneously duplicated content (i.e. haplotigs) in the initial draft assemblies was identified and removed using the K-mer-based deduplication pipeline described in Mathers et al. [46]. Following deduplication, the assemblies were screened for contamination based on manual inspection of "BlobPlots" generated as described above for *S. miscanthi* (*S. avenae*: Additional File 2: Figure S27 and S28; *M. dirhodum*: Additional File 2: Figure S29 and S30). For *S. aveane*, we identified and removed three circular scaffolds corresponding to the chromosome (636 kb long) and two plasmids of the obligate endosymbiont *Buchnera aphidicola*. We also removed additional scaffolds that had low coverage (< 15 × Illumina short-read converge). For *M. dirhodum*, we identified and removed a circular scaffold corresponding to the *Buchnera* chromosome (642 kb long) and three scaffolds (two circular) corresponding to *Buchnera* plasmids. We also identified and removed 121 scaffolds corresponding to the secondary symbiont *Regiella insecticola*. The *R. insecticola* scaffolds spanned 2.8 Mb which is similar to the reported genome size of an *R. insecticola* isolate from pea aphid [103], indicating we have likely assembled the complete genome of this bacterium. We also removed additional scaffolds that had low coverage (< 10 × Illumina short-read converge).

We further improved the contiguity of the *S. avenae* and *M. dirhodum* assemblies using RNA-seq scaffolding with P_RNA_scaffolder [104] as described in Mathers et al. [46]. For *S. avenae*, the RNA-seq scaffolded assembly was carried forward for further scaffolding with Hi-C data (see section below). For *M. dirhodum*, an additional round of assembly deduplication was carried out using purge_dups with assembly self-alignment carried out as for *S. miscanthi* and scaffold coverage estimated by mapping the PCR-free Illumina library to the draft assembly with BWA mem v0.7.7 with default parameters. Coverage cutoffs for purge_dups were set manually with the calcuts script with the parameters "-l 5 -m 25 -u 90". Finally, the *M. dirhodum* assembly was ordered by size and assigned a numbered scaffold ID with SeqKit v0.9.1 to create a frozen release for downstream analysis (Medir_v1.1).

Sequencing and de novo assembly of *R. padi*

R. padi individuals were sampled from a clonal lineage (dubbed JIC1) maintained at the JIC insectary on *A. sativa* (oats). The colony was originally sampled in 2005. The original plant host and sampling location is unknown.

We followed procedures described in Biello et al. [105] to extract high molecular weight DNA from a single *R. padi* individual and sent this to Novogene for 10 × genomics link-read sequencing [106]. We generated an initial de novo assembly with Supernova v2.1.1 [107] with the parameter "–maxreads = 143,797,524" set to give approximately 56 × coverage. To increase contiguity of the Supernova assembly, we carried out two rounds of linked-read scaffolding with scaff10x (https://github.com/wtsi-hpag/Scaff10X) with the parameters "-longread 0 -edge 45,000 -block 45,000" followed by a single round of misjoin detection and scaffolding with Tigmint v1.1.2 [108] using default parameters. The resulting draft assembly was carried forward for further scaffolding with Hi-C data (see section below).

S. avenae and *R. padi* multi-species Hi-C library preparation and genome scaffolding

Whole bodies of *R. padi* and *S. avenae* individuals from clonally reproducing colonies maintained at the JIC Insectary were snap frozen in liquid nitrogen, pooled in approximately equal numbers (~ 100 aphids in total) and sent to Dovetail Genomics (Santa Cruz, CA) for

Hi-C library preparation and sequencing on an Illumina HiSeq X instrument with 150 bp paired-end chemistry. Hi-C library preparation was carried out using the DpnII restriction enzyme following a similar protocol to Lieberman-Aiden et al. [109]. We scaffolded both assemblies using the same multi-species Hi-C library with the 3dDNA assembly pipeline (with default settings) followed by manual curation with JBAT. Pre-curation Hi-C contact maps are shown in Additional File 2: Figure S31 and S32 for *S. avaenae* and *R. padi*, respectively.

For *R. padi*, we screened the resulting chromosome-scale assembly for contamination based on manual inspection of "BlobPlots" generated as described above for *S. miscanthi* (Additional File 2: Figure S33 and S34) and identified a fragmented assembly (159 short scaffolds) of the obligate endosymbiont *Buchnera* which was removed from the final assembly. We also removed additional low coverage scaffolds (< 20 × linked-read coverage). For *S. avenae*, following manual curation of the 3dDNA assembly, we removed additional duplicated content (haplotigs) from the assembly with purge_dups with scaffold coverage estimated from mapping the PCR-free Illumina library with BWA mem v0.7.7 and assembly self-alignment with minimap v2.16 with the parameters "-xasm5 -DP". Coverage cutoffs for purge_dups were estimated automatically using the calcuts script. Finally, the *S. avenae* and *R. padi* genome assemblies were ordered by size and assigned a numbered scaffold ID with SeqKit v0.9.1 to create a frozen release for downstream analysis (*S. avenae*: Siave_v2.1; *R. padi*: Rhpad_v1).

Genome annotation

Our new assemblies of *S. miscanthi*, *S. avenae*, *M. dirhodum* and *R. padi* were annotated following Mathers et al. [37] incorporating evidence from RNA-seq data. Each genome was soft-masked with RepeatMasker v4.0.7 [110, 111] using known Insecta repeats from Repbase [112] with the parameters "-e ncbi -species insecta -a -xsmall -gff". RNA-seq reads were mapped to the genomes with HISAT2 v2.0.5 [113] with the parameters "–max-intronlen 25,000 –dta-cufflinks" followed by sorting and indexing with SAMtools v1.3 [114]. Where strand-specific RNA-seq reads were available, we included the parameter "–rna-strandness RF". We then ran BRAKER2 [115, 116] with UTR training and prediction enabled with the parameters "–softmasking –gff3 –UTR = on". Strand-specific RNA-seq alignments were split by forward and reverse strands and passed to BRAKER2 as separate BAM files to improve the accuracy of UTR models as recommended in the BRAKER2 documentation. For *S. miscanthi*, we used unstranded RNA-seq data from Jiang et al. [36]. For *S. avenae* and *M. dirohdum*, we used strand-specific RNA-seq generated for this study derived from pools of unwinged adult asexual females. For *R. padi*, we used unstranded RNA-seq data from Thorpe et al. [117]. Full details of RNA-seq libraries used for genome annotation are given in Additional File 1: Table S1. Following gene prediction, genes were removed that contained in frame stop codons using the BRAKER2 script getAnnoFastaFromJoingenes.py and the completeness of each gene set was checked with BUSCO v3 with the Arthropoda gene set ($n = 1066$), using the longest transcript of each gene as the representative transcript. For *S. miscanthi*, we compared RNA-seq pseudo alignment rates between the published v1 annotation from Jiang et al. [36] and our new annotation based on the Simis_v2 assembly. The *S. miscanthi* RNA-seq library used for both annotations was pseudo aligned to the Simis_v1 and Simis_v2 transcript sets with Kallisto v0.44.0 [118] with 100 bootstrap replicates (all other parameters were default) and alignment rates were extracted from the Kallisto run reports.

Phylogenomic analysis

Protein sequences from our new genome assemblies of *S. miscanthi*, *S. avenae* and *M. dirhodum* and eight previously published Aphidinae genomes [37, 46, 119–122] were clustered into orthogroups with OrthoFinder version 2.3.8 [123, 124]. Genome assembly and annotation versions are summarised in Additional File 4: Table S3. Where genes had multiple annotated transcripts, we used the longest transcript to represent the gene model. OrthoFinder was run in multiple sequence alignment mode ("-M msa -S diamond -T fasttree") with DIAMOND version 0.9.14 [125], Multiple Alignment using Fast Fourier Transform (MAFFT) version 7.305 [126] and FastTree version 2.1.7 [53, 54] used for protein similarity searches, multiple sequence alignment and gene and species tree estimation, respectively. The OrthoFinder species tree was automatically rooted based on informative gene duplications with Species Tree Root Inference from Gene Duplication Events (STRIDE [127]). For visualisation, the species tree was pruned to only include dactynotine aphids (*S. miscanthi*, *S. avenae*, *M. dirhodum* and *A. pisum*) and *M. persicae* (outgroup) using ape v5.1 [128]. Genome size, gene counts and orthology information were visualised on the phylogeny with Evolview v3 [129].

Synteny analysis

We identified syntenic blocks of genes between *S. avenae* (Simis_v2.1) and *S. miscanthi* (Simis_v2), and the published chromosome-scale genome assemblies of *A. pisum* (JIC1 v1) and *M. persicae* (clone O v2) [37] using MCScanX v1.1 [59]. For each comparison, we carried out an all versus all BLAST search of annotated protein sequences using BLASTALL v2.2.22 [130] with the parameters "-p BlastP -e 1e-10 -b 5 -v 5 -m 8" and ran MCScanX with the parameters "-s 5 -b 2," requiring synteny blocks to contain at least five consecutive genes and to have a gap of no more than 20 genes. MCScanX results were visualised with SynVisio [131].

Whole genome alignment and estimation of sequence divergence.

We used Progressive Cactus v1.0.0 [60] to align *S. avenae* v2.1 (Simis), *S. miscanthi* v2 (Simis), *A. pisum* JIC1 v1 (Acpis) and *M. persicae* clone O v2 (Myper) genomes given the phylogeny (((Siave,Simis),Medir),Acpis) with default parameters. Tools from the hal [132] and PHAST v1.5 [133] packages were used to manipulate the alignment and calculate divergence statistics. Alignment coverage statistics relative to *S. avenae* and *S. miscanthi* were calculated using halStats with the "-coverage" option. To carry out window-based pairwise sequence divergence analysis relative to the *S. miscanthi* v2 reference genome, we specified fixed 1-Mb windows along *S. miscanthi* chromosome-scale scaffolds using makewindows from bedtools v2.28.0 [134] and extracted alignments for each window in maf format using hal2maf with the parameters "–refGenome Simis –noAncestors –onlyOrthologs –refTargets [window bed file]". The maf files were post processed with maf_stream merge_dups (https://github.com/joelarmstrong/maf_stream) in "consensus" mode to resolve alignments to multiple genomic copies as described Feng et al. [135]. The maf_stream processed alignment files were converted to fasta format with msa_view with the parameter "–soft-masked". To generate pairwise divergence estimates for each genomic window, we reduced the fasta formatted alignment files to contain sequences from *S. miscanthi* and one other target species (either *S. avenae*, *M. dirhodum* or *A. pisum*) with SeqKit v0.9.1 (seqkit grep) and passed these files to phlyoFit which was run with default settings to estimate divergence in substitutions per site under the REV model.

JIC1 and Langfang-1 haplotype divergence

To investigate intra- and inter-individual patterns of sequence divergence in *S. miscanthi* and *S. avenae*, we reconstructed independent phased haplotypes for each assembly using HapCUT2 v1.1 [136] and generated a four-way whole genome alignment with sibeliaZ [63]. Our approach took advantage of the availability of in vivo Hi-C data for both isolates, which contains accurate long-range phasing information [136], and the unique biology of aphids which means lab-reared colonies can be maintained as clonal lineages in the absence of recombination (aphid parthenogenesis is apomictic [22–24]). As such, although sequence data for each isolate is derived from pools of individuals (except from PCR-free Illumina sequence data for *S. avenae*), all individuals sequenced for a given isolate are expected to contain the same two haplotypes and so sequence data can be combined to reconstruct fully phased haplotypes for each isolate.

We followed the HapCUT2 pipeline to assemble chromosome-scale haplotypes for *S. aveane* JIC1 and *S. miscanthi* Langfang-1. Short-read data for each isolate was mapped to its respective reference genome with BWA mem v0.7.17 and the resulting alignments were sorted and indexed with SAMtools v1.7 followed by PCR duplicate marking with picard MarkDuplicates v2.1.1 (https://broadinstitute.github.io/picard/). Using these data, we called single-nucleotide polymorphisms (SNPs) and short structural variants (indels) with Freebayes v1.3.1 [137] with default parameters. The initial variant sets were filtered with BCFtools v1.8 [138] and VcfFilter (https://github.com/biopet/vcffilter) to retain biallelic sites and remove low-quality sites (QUAL < 30) and sites with low sequence coverage (DP < 5). The filtered variant file (in vcf format) was then split by chromosome using BCFtools view for processing with HapCUT2. Next, we extracted haplotype informative information from our read sets for each chromosome using extractHAIRS from HapCUT2. For *S. miscanthi*, we used phase information from PacBio long reads and in vivo Hi-C data. PacBio reads were aligned to Simis_v2 using minimap v2.14 with the parameter "-ax map-pb" and the resulting alignments sorted with SAMtools v1.9 and passed to extractHAIRS with the parameters "–pacbio 1 –new_format 1 –indels 1". Hi-C reads were aligned separately for read 1 and read 2 using BWA mem v0.7.12 and the resulting alignments sorted by read name with sambamba [139] and passed to the HapCUT2 script HiC_repair.py to generate a merged alignment file. The repaired Hi-C mapping file was sorted by read name with sambamba, processed with SAMtools fixmate, sorted again by coordinate and PCR duplicates marked with picard MarkDuplicates v2.1.1. The processed Hi-C alignments were passed to extractHAIRS with the parameters "–HiC 1 –new_format 1 –indels 1". For *S. avenae*, in the absence of long-read data, we used phase information from our PCR-free reads and our in vivo Hi-C data. Preliminary analysis showed that including phase information from the PCR-free Illumina reads increased the proportion of phased variants on scaffold_1 (the longest chromosome in the assembly) from 79% (137,885 / 172,422) to 97% (166,986 / 172,422) compared to just using phase information from the Hi-C reads. *S. avenae* Hi-C reads were aligned to Siave_v2.1 and processed following the procedure described for *S. miscanthi*. For the *S. avenae* PCR-free Illumina reads, we used the alignment file generated for variant calling and passed it to extractHAIRS with the parameters "–new_fromat 1 –indels 1". Using the variant calls and phase information generated by extractHAIRS, we ran HapCUT2 separately for each chromosome of *S. miscanthi* and *S. aveane* with the parameters "–HiC 1 –ea 1 –nf 1 –outvcf 1". Phasing statistics were extracted from the resulting vcf files with WhatsHap stats v0.17 [140], and we made fasta files of each haplotype (per chromosome) using BCFtools consensus v1.8 with haplotypes specified with either "-H 1" or "-H 2" and concatenated them by haplotype and sample (either *S. avaenae* JIC1 or *S. miscanthi* Langfang-1). Overall, this pipeline generated four independent haplotype assemblies (incorporating SNPs and indels) with ≥ 99.96% of each chromosome contained in a single-phase block (Additional File 6: Table S5). Although each chromosome is nearly fully phased for each sample, the assignment of H1 or H2 haplotype IDs is arbitrary between chromosomes.

To estimate divergence between the assembled haplotypes of *S. avenae* JIC1 and *S. miscanthi* Langfang-1, we generated a four-way whole genome alignment with sibeliaZ using default settings and processed the alignment with MafFilter v1.3.1 [141] and mafTools v0.2 [142]. MafFilter subset with the parameters "species = (JIC1_H1,JIC1_H2,LF1_H1,LF1_H2),strict = yes,keep = no,remove_duplicates = yes)" was used to retain alignment blocks that are covered by all haplotypes and remove blocks containing paralogs. The filtered alignment was ordered using JIC1 haplotype 1 (JIC1_H1) as the reference with mafRowOrderer with the parameter "–order JIC1_H1,JIC1_H2,LF1_H1,LF1_H2" then processed with mafStrander and mafSorter, both with the parameter "–seq = JIC1_H1". We specified 100 kb fixed genomic windows relative to the JIC1_H1 assembly and estimated pairwise sequence divergence with phyloFit as for the divergence estimates generated for the progressive cactus alignment described in the section above, with the exception that mafExtractor was used to generate window-specific alignment files from the pre-processed sibeliaZ maf file.

To confirm that the divergent haplotypes observed in the Langfang-1 isolate were from a single clonal lineage and were not the result of sampling a mixed population of aphids, we compared the read depth of our assembled haplotypes using the subsampled Langfang-1 Illumina short reads and the Langfang-1 PacBio long reads. If the two haplotypes are derived from an isogenic isolate of a single asexually reproducing diploid female (as described in the original study [36]), we would expect equal coverage of both haplotypes, whereas if the haplotypes are derived from a mixed sample, we would expect uneven coverage of the haplotypes. We generated a merged fasta file of the Langfang-1 H1 and H2 chromosomes assembled with HapCut2. Langfang-1 short and long reads were mapped to the merged haplotype assembly as described above for the HapCut2 pipeline. For each readset, we calculated the average sequencing depth per haplotype, per chromosome in 100 kb fixed windows using Sambamba [139]. Both the short- and long-read datasets had equal coverage of each haplotype for all chromosomes, as expected for an isogenic isolate (Additional File 2: Figure S35 and S36).

Samples, read mapping and genotyping

We obtained genotyping be sequencing (GBS) data for *S. avenae* UK populations (119 samples) and *S. miscanthi* Chinese populations (100 samples) from Morales-Hojas et al. (2020). Sample information is provided Additional File 7: Table S6. Reads were trimmed for adapters and low-quality bases with trim_galore v0.4.5 with the parameters "–paired –length 100". We also included Illumina short reads from the isolates of *S. avaenae* (JIC1) and *S. miscanthi* (Langfang-1) used for genome assembly. These data were subsampled to give approximately 25 × coverage using SeqKit sample v0.9.1 with the parameters "-p 0.25 -s 1234" for *S. miscanthi* Langfang-1 and "-p 0.4 -s 1234" for of *S. avaenae* JIC1. All read sets were mapped to the *S. avenae* v2.1 genome assembly using BWA mem v0.7.17 with default parameters and the alignments were sorted and indexed with SAMtools v1.7 followed by PCR duplicate marking with picard MarkDuplicates v2.1.1. Mapping statistics were gathered with QualiMap v2.2.1 [143] and we omitted all samples with less than 1,000,0000 aligned reads ($n = 12$) from downstream analyses. Variant calling was carried out with Freebayes v1.3.1 with default parameters. Variants were filtered with BCFtools v1.8 as follows: we retained only biallelic SNP sites located on one of the nine chromosome-scale *S. avenae* v2 scaffolds, we removed sites with low-quality (QUAL < 30) and individual genotype calls with fewer than two supporting reads (FORMAT/DP < 2). We then further filtered the variant set to remove sites with more than 25% missing data using VCFtools v0.1.15 [144] with the parameter "–max-missing 0.75". Following these steps, we identified and removed 60 samples that had more than 30% missing data (across all retained sites) using VCFtools v0.1.15 [144].

Principal component analysis (PCA)

We investigated relationships among the Chinese and UK samples using principal component analysis with SNPrelate [145]. To minimise the effects of linkage disequilibrium (LD), SNPs were thinned to one SNP every 25 kb with VCFtools v0.1.15, reducing the filtered variant set to 1772 sites. Plotting of principal components 1 and 2 revealed clustering of the Chinese samples in accordance with the populations identified by Morales-Hojas et al.

[35] which cluster largely based on geographic location. We therefore assigned each sample to either one of six Chinese populations identified by Morales-Hojas et al. [35] or to the UK population Additional File 7: Table S6. Seven samples (out of 149) clustered with a different population to that expected by their geographic origin — these samples were assigned to populations based their genetic identity (ascertained by PCA).

Phylogenetic network analysis

To further visualise population structure in our data and to investigate the origin of the Langfang-1 individual used for *S. miscanthi* genome assembly, we generated a distance-based split network using the neighbour-net algorithm with SplitsTree v4.14.6 [65]. To generate the network, we phased the filtered variant set with BEAGLE v5.1 [64] using default settings and thinned the phased SNP set to one SNP every 25 kb with VCFtools v0.1.15. Haplotypes from *S. avenae* v2.1 scaffold_2 (the longest autosome) were extracted in fasta format using PGDspider v2.1.1.5 [146] with the parameters "FASTA_WRITER_HAPLOID_QUESTION = false VCF_PARSER_EXC_MISSING_LOCI_QUESTION = true VCF_PARSER_MONOMORPHIC_QUESTION = false VCF_PARSER_PLOIDY_QUESTION = DIPLOID".

Divergence time analysis

We used the pea aphid spontaneous mutation rate (2.7×10^{-10} per haploid genome per generation [66]) to estimate the divergence time between *Sitobion* and *Metopolophium*. From our OrthoFinder phylogenomic analysis of aphids (see section above), we identified 11,702 phylogenetically inferred 1-to-1 orthologs between *S. miscanthi* and *M. dirhodum*. For each pair of orthologous genes, we extracted coding sequences, generated a codon alignment with PRANK v150803 [147] with the parameter "-codon" and estimated synonymous site (third codon positions) divergence using paml v4.9 [148] with YN00 [149]. Genes were categorised based on their location (X chromosome or autosome) in the Simis_v2 genome assembly and genes on unplaced scaffolds were excluded. We then estimated the divergence time between *Sitobion* and *Metopolophium* (in number of generations) using the formula $T = d_S / 2\mu$ [150], where d_S is the median sequence divergence at autosomal synonymous sites and μ is the pea aphid spontaneous mutation rate in substitutions per haploid genome per generation. To convert our estimate from number of generations to years, we divided T by 15 which corresponds to the estimated number of aphid generations per year assuming 14 asexual generations and one sexual generation [68].

To date the divergence of ingroup *Sitobion* lineages, we jointly called variants among the *Sitobion* GBS samples (filtered set, see above), *S. miscanthi* Langfang-1 whole genome sample, *S. avaenae* JIC1 whole genome sample and the *M. dirhodum* whole genome sample and estimated a species / population tree under the MSC with SNAPP v1.5.1 [69]. To generate the variant set used for coalescent analysis, we mapped *M. dirhodium* PCR-free Illumina short reads to the *S. avaene* v2.1 genome assembly using BWA mem v0.7.17 with default parameters, sorted and indexed the alignments with SAMtools v1.7 and marked duplicates with picard MarkDuplicates v2.1.1. Variants among the *M. dirhodium* sample and the filtered set of *Sitobion* samples were called using Freebayes v1.3.1 with default parameters. Variants were filtered as for the *Sitobion*-only analysis with the exception that we removed sites with more than 10% missing data using VCFtools v0.1.15 with the parameter "–max-missing 0.9". For the SNAPP analysis, we selected the two highest coverage samples from each population (Additional File 7: Table S6). Where populations contained samples from two locations, we selected the highest coverage sample from each location. The Langfang-1 sample was excluded due to putative hybrid origin. SNAPP xml files were prepared following Stange et al. [67] using the script snapp_prep.rb. We set a starting tree that specified a split between *Sitobion* and *Metopolophium* with all other relationships unresolved ("(MedirOG:7.27,(UK:0.6,KM:0.6,LF_MY:0.6,SZ_PL:0.6,QD_TA:0.6,TG_YC:0.6,WH:0.6):3);") and applied a normally distributed prior centred at 7.27 Mya (SD = 0.5 Mya) on the split between *Sitobion* and *Metopolophium* to calibrate the molecular clock. Ingroup *Sitobion* lineages were constrained to be monophyletic with respect to *M. dirhodum*. We carried out two independent SNAPP runs with BEAST v2.6.3 [151], running each for 1 million MCMC iterations and taking samples every 500 iterations. We checked stationarity and convergence of the runs with Tracer v1.7.1 (effective sample size > 100 for all parameters) and generated a maximum clade credibility tree using TreeAnnotator v2.6.3, discarding the first 10% of samples as burn in.

Demographic history

We reconstructed historical changes in effective population size for *S. miscanthi* and *S. avaenae* using MSMC2 v2.0 [70], which implements the multiple sequentially Markovian coalescent (MSMC) model. We used read mappings (against the *S. avenae* JIC1 v2.1 reference genome) from the population genomic analysis described above for the Langfang-1 and JIC1 whole genome samples and called variants in each sample with SAMtools mpileup v1.3 (parameters: "-q 20 -Q 20 -C 50 -u ") and BCFtools call v1.3.1 (parameters: "-c -V indels "). Variant calls from BCFtools were passed to the bamCaller.

py script from the msmc-tools repository (github.com/stschiff/msmc-tools) to generate vcf and mask files which were in turn passed to the generate_multihetsep.py script (also from the msmc-tools repository) to generate the required input files for MSMC2. For bamCaller.py, we provided the average sequencing depth for the Langfang-1 (23x) and JIC1 (26x) samples as calculated from the output of SAMtools depth using only chromosome-scale scaffolds. MSMC2 was run with default setting and the output was scaled for plotting using the pea aphid spontaneous mutation rate (2.7×10^{-10} per haploid genome per generation [66]) and 15 generations per year [68].

D statistics

We summarised admixture across the *S. avenae* / *miscanthi* complex using Patterson's *D* [73] and the *f*-branch (f_b) statistic with Dsuite v0.4r38 [152]. We used Dtrios from Dsuite with the SNP set generated for the SNAPP phylogenomic analysis (described above) and the SNAPP phylogeny to calculate D_{min}, the minimum amount of allele sharing regardless of any assumptions made about the tree topology, for all trios of ingroup lineages ($n=35$). We also summarised rates of introgression using the f_b statistic with Fbranch from Dsuite using the ".tree" file generated by Dtrios and the SNAPP phylogeny. f_b statsitics were plotted on the SNAPP phylogeny using dtools.py, specifying *M. dirhodum* as the outgroup. We note that the Langfang-1 whole genome sample was excluded from this analysis due to putative hybrid origin.

Topology weighting

To investigate genome-wide patterns of introgression and hybridisartion among *Sitobion* lineages at greater resolution, we re-filtered the raw variant calling results generated for the SNAPP phylogenomic analysis, excluding all the UK *S. avenae* GBS samples and the *M. dirhodum* sample. We expected the resulting set of filtered variants to contain more sites because we had previously implemented strict criteria requiring called sites to be shared by at least 90% of samples and the UK GBS samples had been prepared using a different restriction enzyme to the Chinese samples, meaning only a small number of sites overlapped in the two sets of GBS samples due chance proximity of restriction enzyme cut sites. We note that, in this reduced dataset, the UK *S. avenae* population is still represented by the JIC1 whole genome sample. After removing the UK GBS samples from the raw variant file, we applied the following filtering criteria: we retained only biallelic SNP sites located on one of the nine chromosome-scale *S. avenae* v2 scaffolds, we removed sites with low-quality (QUAL < 30) and individual genotype calls with fewer than two supporting reads (FORMAT/DP < 2), we removed sites with a called genotype in less than 90% of the samples. The filtered vcf file was then phased with BEAGLE v5.1 using default settings. We refer to this set of variants as the "large SNP set" in the section below.

We used topology weighting by iterative sampling of subtrees (Twisst [75]) to explore phylogenetic relationships across the genome, focussing on three focal lineages (WH, KM and QD_TA) inferred to have high levels of introgression based on the f_b analysis. We processed the phased "large SNP set" with scripts from the genomics_general repository (https://github.com/simonhmartin/genomics_general) to create phylogenetic trees (containing two haplotypes per sample) in 50 SNP windows across the *S. avenae* JIC1 v2.1 reference genome using PhyML v3.3 [153] with the GTR substitution model. We then ran Twisst to calculate topology weightings for the three possible topologies describing relationships between the KM, WH, QD_TA and UK (included as an outgroup) lineages. Samples not belonging to the four lineages of interest were ignored. For visualisation of topology weightings along *S. avenae* JIC1 v2.1 chromosomes, a smoothing parameter was applied with a loess span of 1,000,000 bp, with a 100,000 bp spacing. SplitsTree networks were generated for the whole genome and for regions of interest as described above for the smaller SNP set.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12915-023-01649-4.

Additional file 1: Table S1. Summary of sequence data used for genome assembly and annotation.

Additional file 2: Figures S1 - S36. Figure S1. Whole genome alignment of the S. miscanthi v1 and A. pisum JIC1 v1 genome assemblies. **Figure S2.** BUSCO completeness plot. **Figures S3, S5 and S8.** KAT k-mer spectra plots. **Figures S4, S6, S9, S25 - S30, S33 and S44.** Taxon-annotated GC content-coverage plots. **Figures S7, S24, S31 and S32.** Hi-C contact maps. **Figure S10.** Gene set BUSCO completeness plot. **Figure S11.** Maximum likelihood phylogeny of 11 aphid species. **Figure S12.** JIC1 and LF1 within individual haplotype divergence. **Figure S13.** JIC1 and LF1 between individual haplotype divergence. **Figures S14 - S21.** Per chromosome SplitsTree networks of phased haplotypes for S. miscanthi and S. avaenae GBS samples and the Langfang-1 and JIC1 whole genome samples. **Figure S22.** Synonymous site divergence between *S. miscanthi* and *M. dirhodum* one-to-one orthologs. **Figure S23.** SNAPP maximum-clade-credibility time calibrated phylogeny of Sitobion lineages. **Figures S35 and S36.** Per haplotype, per chromosome sequencing depth for the Hapcut2 phased assembly of *S. miscanthi* Langfang-1 based on Illuminaand PacBioread mapping.

Additional file 3: Table S2. RNA-seq pseudoalignment stats.

Additional file 4: Table S3. Genomes used for phylogenomic analysis.

Additional file 5: Table S4. Gene family clustering summary statistics.

Additional file 6: Table S5. HapCUT2 phasing statistics.

Additional file 7: Table S6. Sample information and mapping statistics for population genomic analysis of *S. miscanthi* and *S. avenae.*

Additional file 8: Table S7. Per chromosome shared SNP counts among UK and Chinese GBS Sitobion samples.

Additional file 9: Table S8. Dsuite summary statistics for all combinations of UK and Chinese Sitobion lineages.

Acknowledgements
This research was supported by JIC Technology Platforms, particularly the Entomology and Insectary facilities and staff. We thank the NBI Computing Infrastructure for Science Group, which provides technical support and maintenance to the John Innes Centre's high-performance computing cluster and storage systems.

Authors' contributions

T.C.M. and S.H. conceived the study. S.H. obtained funding for the project. R.H.M.W., S.T.M. and R.B. extracted DNA and RNA. T.C.M. carried out bioinformatic analysis. T.C.M, S.H. and C.V.O. wrote the manuscript. All authors read and approved the final manuscript.

Funding
The described work was supported by a Biotechnology and Biological Sciences Research Council (BBSRC) grant (BB/R009481/1) and a CEPAMS grant (17.03.2) awarded to S.A.H. and a BBSRC Future Leader Fellowship (BB/R01227X/1) awarded to T.C.M.. R.H.M.W. was funded from the BBSRC Norwich Research Park Biosciences Doctoral Training Partnership Award (BB/M011216/1). Additional support was provided by the BBSRC Institute Strategy Programs (BBS/E/J/000PR9797 and BBS/E/J/000PR9798) awarded to the John Innes Centre. The JIC is grant-aided by the John Innes Foundation.

Availability of data and materials

Raw sequence data generated for this study are available at the NCBI short-read archive under BioProject PRJNA880698 [154]. Supporting data including OrthoFinder gene family clustering results, whole genome alignments, variant calls and SNAPP configuration files are available from Zenodo (https://doi.org/10.5281/zenodo.7108778) [155]. Genome assemblies and annotations generated in this study are available from the Aphidinae comparative genomics resource (https://doi.org/10.5281/zenodo.5908005) [42].

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Received: 19 December 2022 Accepted: 13 June 2023
Published online: 13 July 2023

References

1. Stukenbrock EH, McDonald BA. The origins of plant pathogens in agro-ecosystems. Annu Rev Phytopathol. 2015;2008(46):75–100.
2. Bernal JS, Medina RF. Agriculture sows pests: how crop domestication, host shifts, and agricultural intensification can create insect pests from herbivores. Curr Opin Insect Sci. 2018;26:76–81.
3. McDonald BA, Stukenbrock EH. Rapid emergence of pathogens in agro-ecosystems: Global threats to agricultural sustainability and food security. Philos Trans R Soc B Biol Sci. 2016;371(1709):20160026.
4. Inoue Y, Vy TTP, Yoshida K, Asano H, Mitsuoka C, Asuke S, et al. Evolution of the wheat blast fungus through functional losses in a host specificity determinant. Science (80-). 2017;357:80–3.
5. Couch BC, Fudal I, Lebrun MH, Tharreau D, Valent B, Van Kim P, et al. Origins of host-specific populations of the blast pathogen *Magnaporthe oryzae* in crop domestication with subsequent expansion of pandemic clones on rice and weeds of rice. Genetics. 2005;170:613–30.
6. Grünwald NJ, Flier WG. The biology of *Phytophthora infestans* at its center of origin. Annu Rev Phytopathol. 2005;43:171–90.
7. Taylor KL, Hamby KA, DeYonke AM, Gould F, Fritz ML. Genome evolution in an agricultural pest following adoption of transgenic crops. Proc Natl Acad Sci U S A. 2021;118:1–10.
8. Panini M, Chiesa O, Troczka BJ, Mallott M, Manicardi GC, Cassanelli S, et al. Transposon-mediated insertional mutagenesis unmasks recessive insecticide resistance in the aphid *Myzus persicae*. Proc Natl Acad Sci U S A. 2021;118:1–10.
9. Pélissié B, Chen YH, Cohen ZP, Crossley MS, Hawthorne DJ, Izzo V, et al. Genome resequencing reveals rapid, repeated evolution in the Colorado potato beetle. Mol Biol Evol. 2022;39(2):msac016.
10. Hartmann FE, Vonlanthen T, Singh NK, McDonald MC, Milgate A, Croll D. The complex genomic basis of rapid convergent adaptation to pesticides across continents in a fungal plant pathogen. Mol Ecol. 2021;30:5390–405.
11. Dong S, Stam R, Cano LM, Song J, Sklenar J, Yoshida K, et al. Effector specialization in a lineage of the Irish potato famine pathogen. Science. 2014;343:552–5.
12. Menardo F, Praz CR, Wyder S, Ben-David R, Bourras S, Matsumae H, et al. Hybridization of powdery mildew strains gives rise to pathogens on novel agricultural crop species. Nat Genet. 2016;48:201–5.
13. McMullan M, Gardiner A, Bailey K, Kemen E, Ward BJ, Cevik V, et al. Evidence for suppression of immunity as a driver for genomic introgressions and host range expansion in races of *Albugo candida*, a generalist parasite. Elife. 2015;2015:1–24.
14. Valencia-Montoya WA, Elfekih S, North HL, Meier JI, Warren IA, Tay WT, et al. Adaptive introgression across semipermeable species boundaries between local *Helicoverpa zea* and invasive *Helicoverpa armigera* moths. Mol Biol Evol. 2020;37:2568–83.
15. Rogério F, Van Oosterhout C, Ciampi-Guillardi M, Correr FH, Hosaka GK, Cros-Arteil S, et al. Means, motive and opportunity for biological invasions: Genetic introgression in a fungal pathogen. Mol Ecol. 2022;January:1–15.
16. Hogenhout SA, Ammar ED, Whitfield AE, Redinbaugh MG. Insect vector interactions with persistently transmitted viruses. Annu Rev Phytopathol. 2008;46:327–59.
17. Whitfield AE, Falk BW, Rotenberg D. Insect vector-mediated transmission of plant viruses. Virology. 2015;479–480:278–89.
18. Ng JCK, Falk BW. Virus-vector interactions mediating nonpersistent and semipersistent transmission of plant viruses. Annu Rev Phytopathol. 2006;44:183–212.
19. Dixon AFG. Aphid ecology: life cycles, polymorphism, and population regulation. Annu Rev Ecol Syst. 1977;8:329–53.
20. Simon JC, Stoeckel S, Tagu D. Evolutionary and functional insights into reproductive strategies of aphids. Comptes Rendus - Biol. 2010;333:488–96.
21. Moran NA. The evolution of aphid life cycles. Annu Rev Entomol. 1992;37:321–48.
22. Hales FD, Wilson ACC, Sloane M a, Simon J-C, le Gallic J-F, Sunnucks P. Lack of detectable genetic recombination on the X chromosome during the parthenogenetic production of female and male aphids. Genet Res. 2002;79:203–9.
23. Blackman RL. Stability and variation in aphid clonal lineages. Biol J Linn Soc. 1979;11:259–77.
24. Tomiuk J, Wöhrmann K. Comments on the genetic stability of aphid clones. Experientia. 1982;38:320–1.
25. Irwin ME, Thresh JM, Harrison BD. Long-range aerial dispersal of cereal aphids as virus vectors in North America. Philos Trans R Soc London B, Biol Sci. 1988;321:421–46.
26. Margaritopoulos JT, Kasprowicz L, Malloch GL, Fenton B. Tracking the global dispersal of a cosmopolitan insect pest, the peach potato aphid. BMC Ecol. 2009;9:1–13.

27. Haack L, Simon JC, Gauthier JP, Plantegenest M, Dedryver CA. Evidence for predominant clones in a cyclically parthenogenetic organism provided by combined demographic and genetic analyses. Mol Ecol. 2000;9:2055–66.
28. Figueroa CC, Simon JC, Le Gallic JF, Prunier-Leterme N, Briones LM, Dedryver CA, et al. Genetic structure and clonal diversity of an introduced pest in Chile, the cereal aphid *Sitobion avenae*. Heredity. 2005;95:24–33.
29. Peccoud J, Figueroa CC, Silva AX, Ramirez CC, Mieuzet L, Bonhomme J, et al. Host range expansion of an introduced insect pest through multiple colonizations of specialized clones. Mol Ecol. 2008;17:4608–18.
30. Blackman RL, Eastop VF. Aphids on the world's crops: an identification and information guide. Chichester: Wiley; 1984.
31. Choe HJ, Lee SH, Lee S. Morphological and genetic indiscrimination of the grain aphids, *Sitobion avenae* complex (Hemiptera: Aphididae). Appl Entomol Zool. 2006;41:63–71.
32. Vickerman GP, Wratten SD. The biology and pest status of cereal aphids (Hemiptera: Aphididae) in Europe: A review. Bull Entomol Res. 1979;69:1–32.
33. Papura D, Simon JC, Halkett F, Delmotte F, Le Gallic JF, Dedryver CA. Predominance of sexual reproduction in Romanian populations of the aphid *Sitobion avenae* inferred from phenotypic and genetic structure. Heredity (Edinb). 2003;90:397–404.
34. Simon JC, Baumann S, Sunnucks P, Hebert PDN, Pierre JS, Gallic JFLE, et al. Reproductive mode and population genetic structure of the cereal aphid *Sitobion avenae* studied using phenotypic and microsatellite markers. Mol Ecol. 1999;8:531–45.
35. Morales-Hojas R, Sun J, Alvira Iraizoz F, Tan X, Chen J. Contrasting population structure and demographic history of cereal aphids in different environmental and agricultural landscapes. Ecol Evol. 2020;10:9647–62.
36. Jiang X, Zhang Q, Qin Y, Yin H, Zhang S, Li Q, et al. A chromosome-level draft genome of the grain aphid *Sitobion miscanthi*. Gigascience. 2019;8:giz101.
37. Mathers TC, Wouters RHM, Mugford ST, Swarbreck D, van Oosterhout C, Hogenhout SA, et al. Chromosome-scale genome assemblies of aphids reveal extensively rearranged autosomes and long-term conservation of the X chromosome. Mol Biol Evol. 2021;38:856–75.
38. Li Y, Zhang B, Moran NA. The aphid X chromosome is a dangerous place for functionally important genes: diverse evolution of hemipteran genomes based on chromosome-level assemblies. Mol Biol Evol. 2020;37:2357–68.
39. Howe K, Chow W, Collins J, Pelan S, Pointon DL, Sims Y, et al. Significantly improving the quality of genome assemblies through curation. Gigascience. 2021;10:1–9.
40. Dudchenko O, Shamim MS, Batra S, Durand NC, Musial NT, Mostofa R, et al. The Juicebox Assembly Tools module facilitates de novo assembly of mammalian genomes with chromosome-length scaffolds for under $1000. bioRxiv. 2018;:254797.
41. Jain C, Koren S, Dilthey A, Phillippy AM, Aluru S. A fast adaptive algorithm for computing whole-genome homology maps. Bioinformatics. 2018;34:i748–56.
42. Mathers TC, Mugford ST, Wouters RHM, Heavens D, Botha A-M, Swarbreck D, et al. Aphidinae comparative genomics resource (Version v1) [Data set]. Zenodo.2022. https://doi.org/10.5281/ZENODO.5908005.
43. Marbouty M, Cournac A, Flot JF, Marie-Nelly H, Mozziconacci J, Koszul R. Metagenomic chromosome conformation capture (meta3C) unveils the diversity of chromosome organization in microorganisms. Elife. 2014;3: e03318.
44. Monti V, Manicardi GC, Mandrioli M. Distribution and molecular composition of heterochromatin in the holocentric chromosomes of the aphid *Rhopalosiphum padi* (Hemiptera: Aphididae). Genetica. 2010;138:1077–84.
45. Choi H, Shin S, Jung S, Clarke DJ, Lee S. Molecular phylogeny of Macrosiphini (Hemiptera: Aphididae): an evolutionary hypothesis for the Pterocomma-group habitat adaptation. Mol Phylogenet Evol. 2018. https://doi.org/10.1016/j.ympev.2017.12.021.
46. Mathers TC, Mugford ST, Hogenhout SA, Tripathi L. Genome sequence of the banana aphid, *Pentalonia nigronervosa* Coquerel (Hemiptera: Aphididae) and its symbionts. G3 Genes, Genomes, Genet. 2020;10:4315–21.
47. Byrne S, Schughart M, Carolan JC, Gaffney M, Thorpe P, Malloch G, et al. Genome sequence of the English grain aphid, *Sitobion avenae* and its endosymbiont *Buchnera aphidicola*. G3 Genes, Genomes, Genet. 2022;12(3):jkab418.
48. Villarroel CA, González-González A, Alvarez-Baca JK, Villarreal P, Ballesteros GI, Figueroa CC, et al. Genome sequencing of a predominant clonal lineage of the grain aphid *Sitobion avenae*. Insect Biochem Mol Biol. 2022;143 February:103742.
49. Zhu B, Wei R, Hua W, Li L, Zhang W, Liang P, et al. A high-quality chromosome-level assembly genome provides insights into wing dimorphism and xenobiotic detoxification in *Metopolophium dirhodum* (Walker). Res Sq. 2022;1–24.
50. Julca I, Marcet-houben M, Cruz F, Vargas-chavez C, Spencer J, Frias L, et al. Phylogenomics identifies an ancestral burst of gene duplications predating the diversification of Aphidomorpha. Mol Biol Evol. 2020;37:730–56.
51. von Dohlen CD, Rowe C a, Heie OE. A test of morphological hypotheses for tribal and subtribal relationships of Aphidinae (Insecta: Hemiptera: Aphididae) using DNA sequences. Mol Phylogenet Evol. 2006;38:316–29.
52. Shimodaira H, Hasegawa M. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. Mol Biol Evol. 1999;16:1114–6.
53. Price MN, Dehal PS, Arkin AP. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. Mol Biol Evol. 2009;26:1641–50.
54. Price MN, Dehal PS, Arkin AP. FastTree 2 - approximately maximum-likelihood trees for large alignments. PLoS ONE. 2010;5:e9490.
55. Börner C, Heinze K. Aphidina-Aphidoidea. In: Sorauer P, editor. Handbuch der Pflanzenkrankheiten. Berlin: Paul Parey; 1957. p. 1–402.
56. IAGC. Genome sequence of the pea aphid *Acyrthosiphon pisum*. PLoS Biol. 2010;8:e1000313.
57. Li Y, Park H, Smith TE, Moran NA. Gene family evolution in the pea aphid based on chromosome-level genome assembly. Mol Biol Evol. 2019;36:2143–56.
58. Wenger JA, Cassone BJ, Legeai F, Johnston JS, Bansal R, Yates AD, et al. Whole genome sequence of the soybean aphid, *Aphis glycines*. Insect Biochem Mol Biol. 2017. https://doi.org/10.1016/j.ibmb.2017.01.005.
59. Wang Y, Tang H, Debarry JD, Tan X, Li J, Wang X, et al. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. Nucleic Acids Res. 2012;40:e49.
60. Armstrong J, Hickey G, Diekhans M, Fiddes IT, Novak AM, Deran A, et al. Progressive Cactus is a multiple-genome aligner for the thousand-genome era. Nature. 2020;587:246–51.
61. Jaquiéry J, Stoeckel S, Rispe C, Mieuzet L, Legeai F, Simon JC. Accelerated evolution of sex chromosomes in aphids, an X0 system. Mol Biol Evol. 2012;29:837–47.
62. Jaquiéry J, Peccoud J, Ouisse T, Legeai F, Prunier-Leterme N, Gouin A, et al. Disentangling the causes for faster-X evolution in aphids. Genome Biol Evol. 2018;10:507–20.
63. Minkin I, Medvedev P. Scalable multiple whole-genome alignment and locally collinear block construction with SibeliaZ. Nat Commun. 2020;11:6327.
64. Browning SR, Browning BL. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am J Hum Genet. 2007;81:1084–97.
65. Huson DH, Bryant D. Application of phylogenetic networks in evolutionary studies. Mol Biol Evol. 2006;23:254–67.
66. Fazalova V, Nevado B, True J. Low spontaneous mutation rate and pleistocene radiation of pea aphids. Mol Biol Evol. 2020;37:2045–51.
67. Stange M, Sánchez-Villagra MR, Salzburger W, Matschiner M. Bayesian divergence-time estimation with genome-wide single-nucleotide polymorphism data of sea catfishes (Ariidae) supports miocene closure of the Panamanian Isthmus. Syst Biol. 2018;67:681–99.
68. Loxdale HD, Balog A. Aphid specialism as an example of ecological–evolutionary divergence. Biol Rev. 2018;93:642–57.
69. Bryant D, Bouckaert R, Felsenstein J, Rosenberg NA, Roychoudhury A. Inferring species trees directly from biallelic genetic markers: bypassing gene trees in a full coalescent analysis. Mol Biol Evol. 2012;29:1917–32.
70. Malaspinas AS, Westaway MC, Muller C, Sousa VC, Lao O, Alves I, et al. A genomic history of Aboriginal Australia. Nature. 2016;538:207–14.

71. Li H, Durbin R. Inference of human population history from individual whole-genome sequences. Nature. 2011;475:493–6.
72. Cahill JA, Soares AER, Green RE, Shapiro B. Inferring species divergence times using pairwise sequential markovian coalescent modelling and low-coverage genomic data. Philos Trans R Soc B Biol Sci. 2016;371(1699):20150138.
73. Patterson N, Moorjani P, Luo Y, Mallick S, Rohland N, Zhan Y, et al. Ancient admixture in human history. Genetics. 2012;192:1065–93.
74. Malinsky M, Svardal H, Tyers AM, Miska EA, Genner MJ, Turner GF, et al. Whole-genome sequences of Malawi cichlids reveal multiple radiations interconnected by gene flow. Nat Ecol Evol. 2018;2:1940–55.
75. Martin SH, Van Belleghem SM. Exploring evolutionary relationships across the genome using topology weighting. Genetics. 2017;206 May:429–38.
76. Beerli P. Effect of unsampled populations on the estimation of population sizes and migration rates between sampled populations. Mol Ecol. 2004;13:827–36.
77. Tricou T, Tannier E, De Vienne DM. Ghost lineages highly influence the interpretation of introgression tests. Syst Biol. 2022;71:1147–58.
78. Pang X-X, Zhang D-Y. Impact of ghost introgression on coalescent-based species tree inference and estimation of divergence time. Syst Biol. 2022;0 July:1–15.
79. Burke JM, Arnold ML. Genetics and the fitness of hybrids. Annu Rev Genet. 2001;35:31–52.
80. Moulia C, Le Brun N, Loubes C, Marin R, Renaud F. Hybrid vigour against parasites in interspecific crosses between two mice species. Heredity (Edinb). 1995;74:48–52.
81. Bertorelle G, Raffini F, Bosse M, Bortoluzzi C, Iannucci A, Trucchi E, et al. Genetic load: genomic estimates and applications in non-model animals. Nat Rev Genet. 2022;23 August:492–503.
82. Owen CL, Miller GL. Phylogenomics of the Aphididae: deep relationships between subfamilies clouded by gene tree discordance, introgression and the gene tree anomaly zone. Syst Entomol. 2021;2022:1–17.
83. Stukenbrock EH. The role of hybridization in the evolution and emergence of new fungal plant pathogens. Phytopathology. 2016;106:104–12.
84. Sotiropoulos AG, Arango-Isaza E, Ban T, Barbieri C, Bourras S, Cowger C, et al. Global genomic analyses of wheat powdery mildew reveal association of pathogen spread with historical human migration and trade. Nat Commun. 2022;13:1–14.
85. Latorre SM, Reyes-Avila CS, Malmgren A, Win J, Kamoun S, Burbano HA. Differential loss of effector genes in three recently expanded pandemic clonal lineages of the rice blast fungus. BMC Biol. 2020;18:1–15.
86. Mapleson D, Accinelli GG, Kettleborough G, Wright J, Clavijo BJ. KAT: A K-mer analysis toolkit to quality control NGS datasets and genome assemblies. Bioinformatics. 2017;33:574–6.
87. Waterhouse RM, Seppey M, Simao FA, Manni M, Ioannidis P, Klioutchnikov G, et al. BUSCO applications from quality assessments to gene prediction and phylogenomics. Mol Biol Evol. 2018;35:543–8.
88. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015;31:3210–2.
89. Laetsch DR, Blaxter ML. BlobTools: Interrogation of genome assemblies. F1000Research. 2017;6:1287.
90. Kumar S, Jones M, Koutsovoulos G, Clarke M, Blaxter M. Blobology: exploring raw genome data for contaminants, symbionts, and parasites using taxon-annotated GC-coverage plots. Front Genet. 2013;4:1–12.
91. Cabanettes F, Klopp C. D-GENIES: Dot plot large genomes in an interactive, efficient and simple way. PeerJ. 2018;6:e4958.
92. Durand NC, Shamim MS, Machol I, Rao SSP, Huntley MH, Lander ES, et al. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. Cell Syst. 2016;3:95–8.
93. Dudchenko O, Batra SS, Omer AD, Nyquist SK, Hoeger M, Durand NC, et al. De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. Science. 2017;356:92–5.
94. Kolmogorov M, Yuan J, Lin Y, Pevzner PA. Assembly of long, error-prone reads using repeat graphs. Nat Biotechnol. 2019;37:540–6.
95. Ruan J, Li H. Fast and accurate long-read assembly with wtdbg2. Nat Methods. 2020;17:55–158.
96. Li H. Sequence analysis Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34 May:3094–100.
97. Chakraborty M, Baldwin-Brown JG, Long AD, Emerson JJ. Contiguous and accurate de novo assembly of metazoan genomes with modest long read coverage. Nucleic Acids Res. 2016;44:1–12.
98. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, et al. Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS ONE. 2014;9:e112963.
99. Guan D, Guan D, McCarthy SA, Wood J, Howe K, Wang Y, et al. Identifying and removing haplotypic duplication in primary genome assemblies. Bioinformatics. 2020;36:2896–8.
100. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv. 2013;:arXiv:1303.3997v2.
101. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinformatics. 2009;10:421.
102. Shen W, Le S, Li Y, Hu F. SeqKit: A cross-platform and ultrafast toolkit for FASTA/Q file manipulation. PLoS ONE. 2016;11:e0163962.
103. Nikoh N, Tsuchida T, Koga R, Oshima K, Hattori M, Fukatsu T. Genome analysis of "Candidatus Regiella insecticola" strain TUt, facultative bacterial symbiont of the pea aphid *Acyrthosiphon pisum*. Microbiol Resour Announc. 2020;9:e00598–620.
104. Zhu BH, Xiao J, Xue W, Xu GC, Sun MY, Li JT. P_RNA_scaffolder: a fast and accurate genome scaffolder using paired-end RNA-sequencing reads. BMC Genomics. 2018;19:1–13.
105. Biello R, Singh A, Godfrey CJ, Fernández FF, Mugford ST, Powell G, et al. A chromosome-level genome assembly of the woolly apple aphid, *Eriosoma lanigerum* Hausmann (Hemiptera: Aphididae). Mol Ecol Resour. 2021;21:316–26.
106. Zheng GXY, Lau BT, Schnall-Levin M, Jarosz M, Bell JM, Hindson CM, et al. Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. Nat Biotechnol. 2016;34:303–11.
107. Weisenfeld NI, Kumar V, Shah P, Church DM, Jaffe DB. Direct determination of diploid genome sequences. Genome Res. 2017;27:757–67.
108. Jackman SD, Coombe L, Chu J, Warren RL, Vandervalk BP, Yeo S, et al. Tigmint: correcting assembly errors using linked reads from large molecules. BMC Bioinformatics. 2018;19:1–10.
109. Lieberman-Aiden E, Van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science (80-). 2009;33292 October:289–94.
110. Tarailo-Graovac M, Chen N. Using RepeatMasker to identify repetitive elements in genomic sequences. Curr Protoc Bioinforma. 2009;SUPPL. 25:1–14.
111. Smit AFA, Hubley R, Green P. RepeatMasker Open-4.0. 2013. https://bmcbiol.biomedcentral.com/articles/10.1186/s12915-021-01158-2.
112. Bao W, Kojima KK, Kohany O. Repbase Update, a database of repetitive elements in eukaryotic genomes. Mob DNA. 2015;6:4–9.
113. Kim D, Langmead B, Salzberg SL. HISAT: A fast spliced aligner with low memory requirements. Nat Methods. 2015;12:357–60.
114. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25:2078–9.
115. Hoff KJ, Lange S, Lomsadze A, Borodovsky M, Stanke M. BRAKER1: unsupervised RNA-Seq-based genome annotation with GeneMark-ET and AUGUSTUS. Bioinformatics. 2015;32:767–9.
116. Hoff KJ, Lomsadze A, Borodovsky M, Stanke M. Whole-genome annotation with BRAKER. In: Kollmar M, editor. Gene Prediction: Methods and Protocols. Springer New York: New York, NY; 2019. p. 65–95.
117. Thorpe P, Cock PJA, Bos J. Comparative transcriptomics and proteomics of three different aphid species identifies core and diverse effector sets. BMC Genomics. 2016;17:172.
118. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol. 2016;34:525–7.
119. Mathers TC. Improved genome assembly and annotation of the soybean aphid (*Aphis glycines* Matsumura). G3 Genes, Genomes, Genet. 2020;10:899–906.
120. Nicholson SJ, Nickerson ML, Dean M, Song Y, Hoyt PR, Rhee H, et al. The genome of *Diuraphis noxia*, a global aphid pest of small grains. BMC Genomics. 2015;16:429.

121. Thorpe P, Escudero-Martinez CM, Cock PJAA, Eves-Van Den Akker S, Bos JIBB, Akker SE Den, et al. Shared transcriptional control and disparate gain and loss of aphid parasitism genes. Genome Biol Evol. 2018;10:2716–33.
122. Chen W, Shakir S, Bigham M, Richter A, Fei Z, Jander G. Genome sequence of the corn leaf aphid (Rhopalosiphum maidis Fitch). Gigascience. 2019;8:1–12.
123. Emms DM, Kelly S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol. 2015;16:157.
124. Emms DM, Kelly S. OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol. 2019;20:1–14.
125. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nat Methods. 2014;12:59–60.
126. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013;30:772–80.
127. Emms DM, Kelly S. STRIDE: species tree root inference from gene duplication events. Mol Biol Evol. 2017;34:3267–78.
128. Paradis E, Schliep K. Ape 5.0: An environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics. 2019;35:526–8.
129. Subramanian B, Gao S, Lercher MJ, Hu S, Chen WH. Evolview v3: a webserver for visualization, annotation, and management of phylogenetic trees. Nucleic Acids Res. 2019;47:W270–5.
130. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215:403–10.
131. Bandi V, Gutwin C. Interactive exploration of genomic conservation. Proc - Graph Interface. 2020;2020-May.
132. Hickey G, Paten B, Earl D, Zerbino D, Haussler D. HAL: A hierarchical format for storing and analyzing multiple genome alignments. Bioinformatics. 2013;29:1341–2.
133. Hubisz MJ, Pollard KS, Siepel A. Phastand Rphast: Phylogenetic analysis with space/time models. Brief Bioinform. 2011;12:41–51.
134. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26:841–2.
135. Feng S, Stiller J, Deng Y, Armstrong J, Fang Q, Reeve AH, et al. Dense sampling of bird diversity increases power of comparative genomics. Nature. 2020;587:252–7.
136. Edge P, Bafna V, Bansal V. HapCUT2: Robust and accurate haplotype assembly for diverse sequencing technologies. Genome Res. 2017;27:801–12.
137. Garrison E, Marth G. Haplotype-based variant detection from short-read sequencing. arXiv. 2012;:1207.3907.
138. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. Gigascience. 2021;10:1–4.
139. Tarasov A, Vilella AJ, Cuppen E, Nijman IJ, Prins P. Sambamba: fast processing of NGS alignment formats. Bioinformatics. 2015;31:2032–4.
140. Martin M, Patterson M, Garg S, Fischer SO, Pisanti N, Gunnar W, et al. WhatsHap : fast and accurate read-based phasing. bioRxiv. 2016;:085050.
141. Dutheil JY, Gaillard S, Stukenbrock EH. MafFilter: a highly flexible and extensible multiple genome alignment files processor. BMC Genomics. 2014;15:53.
142. Earl D, Nguyen N, Hickey G, Harris RS, Fitzgerald S, Beal K, et al. Alignathon: a competitive assessment of whole-genome alignment methods. Genome Res. 2014;24:2077–89.
143. Okonechnikov K, Conesa A, García-Alcalde F. Qualimap 2: Advanced multi-sample quality control for high-throughput sequencing data. Bioinformatics. 2016;32:292–4.
144. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. 2011;27:2156–8.
145. Zheng X, Levine D, Shen J, Gogarten SM, Laurie C, Weir BS. A high-performance computing toolset for relatedness and principal component analysis of SNP data. Bioinformatics. 2012;28:3326–8.
146. Lischer HEL, Excoffier L. PGDSpider: an automated data conversion tool for connecting population genetics and genomics programs. Bioinformatics. 2012;28:298–9.
147. Löytynoja A. Phylogeny-aware alignment with PRANK. In: Multiple sequence alignment methods. New York, Heidelberg, Dordrecht: Humana Press, Springer; 2014. p. 155–70.
148. Yang Z. PAML 4: Phylogenetic analysis by maximum likelihood. Mol Biol Evol. 2007;24:1586–91.
149. Yang Z, Nielsen R. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. Mol Biol Evol. 2000;17:32–43.
150. Cutter AD. Divergence times in *Caenorhabditis* and *Drosophila* inferred from direct estimates of the neutral mutation rate. Mol Biol Evol. 2008;25:778–86.
151. Bouckaert R, Vaughan TG, Barido-Sottani J, Duchêne S, Fourment M, Gavryushkina A, et al. BEAST 2.5: an advanced software platform for Bayesian evolutionary analysis. PLoS Comput Biol. 2019;15:1–28.
152. Malinsky M, Matschiner M, Svardal H. Dsuite - Fast D-statistics and related admixture evidence from VCF files. Mol Ecol Resour. 2021;21:584–95.
153. Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol. 2010;59:307–21.
154. Mathers TC, Wouters RHM, Mugford ST, Biello R, Van Oosterhout C, Hogenhout SA. Hybridisation has shaped a recent radiation of grass-feeding aphids. GenBank. 2022. https://www.ncbi.nlm.nih.gov/bioproject/880698.
155. Mathers TC, Wouters RHM, Mugford ST, Biello R, Van Oosterhout C, Hogenhout SA. Supplementary data for: Hybridisation has shaped a recent radiation of grass-feeding aphids. 2022. Zenodo. https://doi.org/10.5281/zenodo.7108778.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.