Chromosome-level genomes provide insights into genome evolution, organization and size in *Epichloë* fungi

Artemis D. Treindl a, *, Jessica Stapley a, David J. Winter b, Murray P. Cox b, Adrian Leuchtmann a

a Plant Ecological Genetics Group, Institute of Integrative Biology, ETH Zurich, 8092 Zurich, Switzerland
b Statistics and Bioinformatics Group, School of Fundamental Sciences, Massey University, Palmerston North 4442, New Zealand

**ARTICLE INFO**

**Keywords:**
- Genome compartmentalization
- Two-speed genomes
- Repeat-induced point mutation
- Fungal plant pathogens
- Transposable elements
- *Epichloë* endophytes

**ABSTRACT**

*Epichloë* fungi are endophytes of cool season grasses, both wild species and commercial cultivars, where they may exhibit mutational or pathogenic lifestyles. The *Epichloë*-grass symbiosis is of great interest to agricultural research for the fungal bioprotective properties conferred to host grasses but also serves as an ideal system to study the evolution of fungal plant-pathogens in natural environments. Here, we assembled and annotated gapless chromosome-level genomes of two pathogenic *Epichloë* sibling species. Both genomes have a bipartite genome organization, with blocks of highly syntenic gene-rich regions separated by blocks of AT-rich DNA. The AT-rich regions show an extensive signature of RIP (repeat-induced point mutation) and the expansion of this compartment accounts for the large difference in genome size between the two species. This study reveals how the rapid evolution of repeat structure can drive divergence between closely related taxa and highlights the evolutionary role of dynamic compartments in fungal genomes.

1. Introduction

Despite their small size, fungal genomes can provide big insights into eukaryote genome organization and evolution. Plant-colonizing fungi and oomycetes, in particular, frequently have genomes with a bipartite organization: a conserved "core genome" containing the majority of genes and a highly variable compartment characterized by a high repeat-content and few genes [1]. In some cases these repeat-rich regions are targeted by repeat-induced point mutation (RIP), a defense mechanism against transposable element proliferation, which leads to locally high mutation rates, specifically C-to-T mutations, reduced G/C content and increased genetic diversity [2,3]. Regional variation in the levels of structural and sequence polymorphism gives rise to different rates of evolutionary change across the genome, referred to as the "two-speed genome" model [4]. Regions with higher rates of evolution are enriched for repetitive elements and effector genes that are involved in plant-pathogen interactions, and these regions are often important in adaptation [4–7]. Despite recent progress, we still lack an understanding of how widespread this bipartite genome organization is, how it varies across taxa and how it relates to evolutionary processes [8].

To date, the most evidence for bipartite genomes comes from pathogens of agricultural crops with large economic impacts. Studies of these agriculturally important pathogens have shown that the distribution of repeat rich compartments can vary. They can cover whole accessory or lineage specific chromosomes, for example in *Fusarium oxysporum* [9] and *Zymoseptoria tritici* [10]. In other species, this bipartite architecture is organized in a mosaic-like structure with alternating blocks of genomic compartments distributed along chromosomes; for example, chromosomes of a dothideomycete pathogen (*Leptosphaeria maculans*) contain structural isochores, scattered sequence blocks of distinct base composition and are enriched in transposable elements [11,12]. These so-called AT-isochores are largely devoid of genes, similar to the genome of the oomycete *Phytophthora infestans*, which also consists of alternating repeat-poor/gene-rich and repeat-rich/gene-poor regions [13]. The genome organization of fungal plant pathogens in natural ecosystems are less well studied, in part because fewer high quality, whole genome sequences are available for these systems. One exception is the facultative pathogen *Epichloë festucae*. A recent study by Winter et al. [14] found that the genome of *E. festucae* also displays a two-speed organization, with AT and GC-rich regions alternating to create a patchwork genome structure and this structure affects the three-dimensional configuration of chromosomes creating compartments of distinct gene expression. This suggests a functional role of genome organization in *Epichloë*, although its evolutionary importance remains largely obscure.

---

* Corresponding author at: Swiss Federal Research Institute WSL, Biodiversity and Conservation Biology, Zürcherstrasse 111, CH-8903 Birmensdorf, Switzerland.
E-mail address: artemis.treindl@env.ethz.ch (A.D. Treindl).

https://doi.org/10.1016/j.jgeno.2021.11.009
Received 26 February 2021; Received in revised form 21 October 2021; Accepted 8 November 2021
Available online 12 November 2021
0888-7543/© 2021 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).
To better understand bipartite genome evolution in natural fungal plant pathogens, we set out to produce high quality reference genome for two pathogenic *Epichloë* species. The genus *Epichloë* (Ascomycota, Clavicipitaceae) consists of obligate biotrophic fungi that form life-long systemic infections of cool-season grasses (Poaceae, tribe Pooidae which includes most species of wild and pasture grasses in temperate areas) [15]. The interaction between *Epichloë* and their host grasses serves as a unique model in the study of fungal adaptation because they span the whole symbiotic continuum from antagonistic to mutualistic depending on the mode of reproduction [16]. Sexually reproducing *Epichloë* are haploid outcrossers and usually a single genotype infects the whole above-ground tissue of the host grass [17]. They grow asymptotically to the point of host flowering and then cause choke disease, whereby the undeveloped host inflorescences are enclosed in the fungal fruiting structure [18]. This pathogenic stage can sterilize the host plant and enables the fungus to horizontally transmit meiotic spores and spread to other plants. Purely asexual *Epichloë*, on the other hand, form infections that do not cause visible symptoms in the host plant and transmit vertically by dissemination via seeds. These mutualistic symbionts can confer various benefits to the host plant including increased drought resistance [19], and protection against herbivorous mammals and invertebrates mediated by the production of bioactive alkaloid compounds [20]. *Epichloë* endophytes are therefore widely employed in commercial agriculture [21–23]. Interestingly, many of these mutualistic symbionts are heteroploid (aneuploid, diploid, or polyploid) interspecific hybrids of two or more ancestral sexual species [24]. The two species studied here are both sexually reproducing pathogens and possess haploid genomes.

The objective of this study was to investigate how the structural organization of fungal genomes may be linked to the divergence of species and to gain an understanding of the genomic mechanisms driving such evolutionary processes. We therefore sequenced DNA and RNA from two closely related taxa in the *Epichloë typhina* species complex (*E. typhina* (Pers.) Brockm. and *E. clarkii* J.F. White) that infect different host grasses, to create finished reference genomes. We combined Illumina and Pacific Biosciences (PacBio) DNA sequencing to generate complete chromosome-level genome assemblies, and used Illumina RNAseq data to improve gene annotation. Using these new genomic resources, we compared (1) patterns of genome organization, (2) gene and repeat content, and (3) synteny between the two species.

2. Methods

2.1. Taxonomic status and nomenclature

This study focuses on two members of the *E. typhina* species complex: *E. typhina* infecting the grass species *Dactylis glomerata* L. and *E. clarkii* infecting *Holcus lanatus* L. The two sibling taxa are currently assigned to the geometric rank of subspecies based on their sexual compatibility and ability to hybridize in experimental crosses [15,25]. Evidence from previous studies suggests, however, that *E. typhina* and *E. clarkii* are reproductively isolated biological entities at an advanced stage of divergence [26–28], and our ongoing work further supports this, detecting no evidence of gene flow among natural sympatric populations of the two taxa (unpublished data). For the purpose of this and future studies we therefore consider *E. typhina* and *E. clarkii* as closely related yet distinct species. Further, as now suggested by the International Code of Nomenclature [29], we refer to *Epichloë* without its historical diarhesis, aiding to make the data, metadata and analysis open, reusable and reproducible.

2.2. Fungal isolates

*Epichloë typhina* strain Ety_1756 and *Epichloë clarkii* strain Ecl_1605_22 were isolated from infected plant tissue collected from naturally infected grass accessions in Switzerland (Supplementary Table S1 and Fig. S1). Note that in subsequent figures we use the abbreviations Ety and Ecl, respectively. Pure fungal cultures were obtained from surface sterilized leaf-sheaths of host plants *D. glomerata* and *H. lanatus* following the procedure described by Leuchtmann and Clay [30]. Isolates were grown in liquid V8 medium [31] for ten days, strained, washed with sterile deionized water to remove residue medium and subsequently freeze dried.

2.3. DNA preparation and sequencing

DNA was extracted according to the simple method for extraction of fungal genomic DNA of Al-Samarrai and Schmid [32]. In short, the protocol involves lysis of 30 mg of freeze dried ground mycelium in fresh sodium dodecyl sulphate buffer, detachment of DNA from polysaccharides by mild shearing, NaCl precipitation of polysaccharides and protein, chloroform extraction and ethanol precipitation. DNA quality and quantity was assessed by automated electrophoresis (2200 TapeStation, Agilent, Santa Clara, CA, USA), spectrophotometric ratios (A260/280 > 1.8 and A260/230 > 2, Nanodrop spectrophotometer, Thermo Fisher Scientific, Waltham, MA, USA) and fluorometric analysis (Quantum™ Fluorometer, Promega Corporation, Wisconsin, USA). For each isolate we obtained around 20 μg of high molecular weight, high purity genomic DNA. PacBio SMRTbell libraries were prepared at the Functional Genomics Center of Zurich (FGCZ), using 15 μg of DNA. Sequencing was performed on a PacBio RSII instrument at the FGCZ (Sequencing and Genomic Technologies Shared Resource, Duke University, NC, USA). For assembly polishing, we generated Illumina paired-end libraries (insert size of 330 bp) using NEBNext® Ultra™ II DNA Library Prep Kits and sequenced these on a HiSeq4000 Illumina sequencer, at a final coverage of approximately 50× for *E. typhina* and 25× for *E. clarkii*.

2.4. RNA extraction and sequencing

RNA was isolated from reference *Epichloë* strains Ety_1756 and Ecl_1605_22 grown as axenic culture, as well as in association with their host grasses *D. glomerata* and *H. lanatus*, respectively. For genes expressed in culture, RNA was extracted from strains grown on supplemented malt-extract agar (containing 1% malt extract, 1% glucose, 0.25% bacto peptone, and 0.25% yeast extract) for 10 days at room temperature. Three technical replicates per strain were used. For *in planta* RNA extraction, endophyte-free seedlings grown from commercial cultivar seeds of *D. glomerata* (cv. ‘Prato’), *Agroscop Reckenholz-Tänikon*) and *H. lanatus* (3624/K01/0.1, Otto Hauenstein seeds) were inoculated with the respective isolates Ety (*D. glomerata*) and Ecl (*H. lanatus*) following the procedure of Latch and Christiansen (1985) [33]. Inoculated plants were kept for 3–5 days in the dark and then grown in multipot trays in commercial soil (Ökohum Bio Universalerde) at 22 °C with a photoperiod of 16 h of light. After 10 weeks, plants were screened for infection by examining aniline blue stained leaf-sheath tissue of three tillers per plant using a light microscope at 400× magnification. Five plants of each species with infection in all three tillers were re-potted into larger pots and moved to the greenhouse. Pseudostems of three plants (biological replicates) with two technical replicates per plant were collected and frozen in liquid nitrogen resulting in a total of six samples per strain. We extracted total RNA using the RNeasy Plant Mini Kit (QiAGEN, Hilden, Germany) and removed residual genomic DNA by DNase treatment (DNase I RNaSe-Free, New England Biolabs, MA, USA). We assessed RNA concentration by fluorometric analysis (Quantum™ Fluorometer, Promega Corporation, Wisconsin, USA), and quality by automated electrophoresis (2200 TapeStation, Agilent, Santa Clara, CA, USA). Paired-end cDNA libraries were prepared with the Illumina TruSeq RNA kit and sequencing was performed at the FGCZ on an Illumina NovaSeq 6000 (Illumina, San Diego, CA, USA).
2.5. Genome assembly

Initial assemblies were produced using three different programs: Canu v1.8 [34], FALCON v1.2.2 [35] and MECAT2 v20190314 [36]. By comparing assemblies using D-GENIES v1.2.0 [37], we found that discontinuities present in one assembly were generally well-resolved in one or both of the other assemblies. As FALCON produced the most contiguous assembly for each of our strains, we used the assemblies produced by this program as the base of our final reference sequence and used overlapping segments present in other assemblies to resolve the remaining gaps. Base-level errors in the final assembly were polished using Pilon v1.23 [38], using Illumina DNA reads as input.

2.6. Gene annotation

The genome was annotated using v1.4 of the funannotate pipeline [39]. This pipeline makes use of Augustus, SNAP, glimmerHMM and CodingQuarry for gene-calling. Step-by-step details and example code are available at https://github.com/adtreindl/Epichloe_genomes/blob/master/RNASeq_Data_Analysis_predict.md. Briefly, we mapped RNA-seq data using STAR v2.5.3 [40], and identified transcripts using stringtie v1.3.3 [41]. We then used the funannotate predict functions to predict genes, incorporating the transcript information. Gene/protein function was predicted using InterProScan v5.17–56.0 [42] and we identified putative effectors using signalP v4.1 [43], and effectorP v2.0 [44]. We identified orthologous genes shared by both genomes using proteinortho v6.0 beta [45]. Additionally, the same analysis was performed to identify orthologs shared between E. typhina, E. clarkii and the outgroup species E. festucae.

2.7. Transposable element annotation

To identify transposable elements (TEs) we produced de novo TE libraries for each genome using the RepeatModeler v2.0.1 pipeline [46], which uses RepeatScout to identify multicopy regions of the genome and RECON [47] to discover known TE motifs. We then identified specific elements from each family in our de novo libraries using RepeatMasker v4.0.6 [48].

2.8. Identification of genome compartmentalization

We scanned genomes for signatures of RIP using The RIPPer (access date 23 October 2020), with default settings (1000 bp window size with 500 bp step size, with substrate index values ≤0.75 and product index values ≥1.1.6) and identified C-to-T mutations resulting in a reduced GC content. We also identified AT-rich regions using OccultCut v1.1 [7], which detects bimodal patterns in GC distribution across genome sequences and then segments them into regions of differing nucleotide-content using the Jensen–Shannon divergence (DJS). The RIPPer accounts for the overall AT-content of a genome and implements a quantitative measure of RIP activity using the RIP index of Margolin [3], whereas OccultCut is a qualitative measure assigning genomic regions to defined categories. Given that both methods rely on nucleotide proportions, we expect results from these analyses to be strongly correlated.

2.9. Synteny

We assessed synteny, the shared physical orientation of genomic sequence between species, at two levels. For a comparison at the nucleotide level, we performed pairwise alignments of three genomes, the two new genomes and the E. festucae F1 reference genome, with minmap2 v2.12 using the preset option for cross-species full genome alignment (asm10 for sequence divergence around 2% and asm20 for sequence divergence up to 10%, giving very similar results). Secondly, synteny was assessed at the gene level by identifying orthologous genes shared by both genomes using proteinortho v6.0 beta [45]. We used coordinates of single copy orthologous genes to investigate gene order and co-linearity.

2.10. Assembly completeness

The completeness of each genome assembly was estimated using BUSCO v3.0.2 [50], by searching for conserved genes in the Sordariomycete dataset (library sordariomycete_odb9).

3. Results

3.1. Assembly and annotation of two complete chromosome-level genomes

We combined PacBio long-read and Illumina short-read DNA sequencing to generate high quality contiguous reference genomes of E. typhina and E. clarkii. PacBio sequencing yielded 810,432 and 807,685 subreads with an N50 of 14,350 and 17,455 base pairs, respectively. In total we obtained 6.8 and 6.6 Gb representing a mean coverage of 200× for the E. typhina genome and 144× for the E. clarkii genome (Table 1). We used a set of different long-read assemblers to generate separate assemblies from these reads. As FALCON produced the most contiguous assembly for each of our strains, we used the assemblies produced by this program as the base of our final assemblies, covering any breaks in the FALCON assembly with near chromosome-length overlapping segments present in other assemblies. All break points in one assembly that were completely resolved in another assembly were confirmed by manual inspection, giving us strong confidence in our final assemblies (see section data accessibility).

The final genome assemblies consist of seven nuclear chromosomes and the complete mitochondrial genome for each species. Nuclear genome size differs between the two sibling species (Table 2), with the E. typhina genome spanning 33.83 Mb whereas the E. clarkii genome is more than a third larger at 45.62 Mb. Genome sequences were deposited on GenBank (BioProject ID PRJNA533212 for E. typhina Ety_1756 and PRJNA533212 for E. clarkii Ecl_1605_22). We identified 8165 and 8736 protein coding genes in the genomes of E. typhina and E. clarkii, respectively. We assigned functional annotations to most of the genes identified; 81.3% and 77.1% of E. typhina and E. clarkii protein coding genes, respectively. We also identified putative signal peptides in both species (584 in E. typhina and 587 in E. clarkii), and of these, we identified 138 and 144 putative effectors in E. typhina and E. clarkii, respectively (see supplementary files here: https://github.com/adtreindl/Epichloe_genomes). We identified 6923 one-to-one orthologs shared by both species. Duplicated genes were rare with only three two-to-two orthologs identified representing genes duplicated in both species (paralogs). Of all gene pairs, 5715 (86.1%) formed an orthology group with a single gene from the sordariomycete dataset (library sordariomyceta_odb9).

3.2. Bipartite genome structure

Both the E. typhina and E. clarkii genomes have been profoundly
affected by RIP, generating the distinctive bipartite genome structure seen in some fungal pathogens. In line with the previously sequenced *E. festucae*, both new *Epichloe* genomes exhibited a clear bimodal distribution of GC-content (Fig. 1) corresponding to two genome compartments with distinct GC-content ranges (*E. typhina*: 0–37.6% and 37.6–100% GC-content; *E. clarkii*: 0–39.9% and 39.9–100% GC-content). Chromosomes consisted of blocks of sequence with high AT nucleotide content (AT-rich regions with 24.4% and 25.7% GC-content on average in *E. typhina* and *E. clarkii*, respectively), and alternating blocks of sequence with approximately equal nucleotide content (52.7% GC-content on average, Table 2 and Fig. 2). We found concordance between quantitative signatures of RIP in genomic windows (identified by The RIPper), and qualitative assignment of AT-rich regions identified by OcculterCut (Figs. S2 and S3), which was expected given that both methods principally rely on nucleotide distribution along the genome.

The genome-wide extent of RIP estimated by both methods was very similar: 31.7% and 30.7% in *E. typhina* were affected by RIP and 48.6% and 45.6% in *E. clarkii* - the former percentages refer to AT-rich compartment size identified by OcculterCut (see Table 2), and the latter correspond to the percentage of bases assigned to Large RIP Affected Regions (LRARs) identified by The RIPper. These genomes thus have a substantially larger proportion of the genome affected by RIP compared to other Sordariomycetes [51], and similarly large RIPped contents were also detected among members of the genus *Claviceps*, the sister group to *Epichloe* [52]. The extent and variation in the proportion of the genome affected by RIP in two closely related lineages clearly demonstrates the impact of this process in the recent history of this taxonomic group.

### 3.3. Distribution of genes

For subsequent analyses, we used the dichotomous categorization into AT-rich and GC-rich regions made by OcculterCut and use the term “compartment” to refer to the entire genomic proportion made up by either sequence category.

The GC-rich compartment contained the vast majority of annotated genes (99.88% and 99.39% in *E. typhina* and *E. clarkii*, respectively), and is henceforth referred to as the gene-rich compartment. In contrast, the AT-rich compartment of these two *Epichloe* genomes were largely devoid of genes (10 (0.12%) and 53 (0.61%) genes in *E. typhina* and *E. clarkii* respectively, and one and two of these genes were putative effectors), in line with other fungal genomes investigated [7]. Consistent with this, we found that of 6923 orthologs shared by both species, most were located in AT-rich regions, whereas only one gene in *E. typhina* (Ety_007361) and two in *E. clarkii* (Ecl_003362 and Ecl_005753) were located in AT-rich regions (Fig. 3, yellow lines).

While RIP specifically targets repeats, RIP-induced mutations have also been shown to leak beyond duplicated regions and occur in single copy sequences located in close proximity to RIP-affected repeats in other fungi (*Neurospora crassa* [53] and *Leptosphaeria maculans* [54]). We therefore tested if putative effector genes were more likely to be located either in or adjacent to AT-rich regions (1 kb up- and down-stream), a genomic location in which they may experience faster rates of sequence evolution which may foster rapid adaptive innovation [see 5]. We used permutation tests implemented using the ‘overlapPermTest’ function in the R package regionR with 1000 iterations [55]. We found that effector genes were significantly less likely to be located in AT-rich regions than expected by chance, in line with the overall low number of annotated genes there. However, effector genes were enriched in the regions directly bordering AT-rich regions (Fig. S16), supporting the idea that this genomic environment may help maintain a dynamic effector repertoire in *Epichloe* and provide an adaptive advantage.

### 3.4. Variation in genome size and repeat content

The genome of *E. clarkii* was a third larger than *E. typhina* and this is largely explained by the repeat content and an expansion of the AT-rich compartment in *E. clarkii*, whereas the gene-rich compartments of the genomes were approximately equal in size (23.1 Mb and 23.4 Mb). The
AT-rich compartment made up nearly half (48.6% or 22.2 Mb) of the genome sequence in *E. clarkii* in contrast to 31.7% (10.7 Mb) in *E. typhina* (Table 2), and *E. clarkii* contained a larger number of AT-rich regions (648 vs 382) and had on average slightly longer AT-rich regions (mean 29.5 kb *E. typhina* vs. 34.2 kb *E. clarkii*) (Fig. 2). This difference in compartment size is congruent with the difference in repeat content between genomes: 18.9 Mb (41.4%) of the *E. clarkii* genome was assigned to a TE family compared to 9.1 Mb (27.0%) TE sequence in *E. typhina*. In both genomes, 95% of all bases annotated as being part of a TE fell into AT-rich regions demonstrating that AT-rich compartments of *E. typhina* and *E. clarkii* are largely comprised of repetitive DNA (Supplementary Table S3 and Fig. S4). The composition of TEs in each genome was similar, with the majority of TE sequences (82% *E. typhina* and 86% in *E. clarkii*) being full or partial LTR retrotransposons and the remainder comprised of DNA transposons, LINE elements and unclassified elements. Almost all LTRs were located in AT-rich regions (99% in *E. typhina* and 97% in *E. clarkii*) whereas a relatively high proportion of unclassified elements and DNA transposons occurred in the gene-rich regions of the genome. Unclassified elements likely include short non-autonomous elements such as MITES, which may function as regulators of gene expression in *Epichloe* [14].

### 3.5. Conserved synteny between sibling taxa

We found complete conservation of synteny between *E. typhina* and *E. clarkii* at the chromosome level, with no evidence of major rearrangements between non-homologous chromosomes (Fig. 3). The chromosome numbers do not correspond because *Epichloe* chromosomes are numbered by size with the suspected short arm of the chromosome first; as a result, chromosome 1 in *E. typhina* corresponds to the inverted chromosome 3 in *E. clarkii*. The synteny of these two sibling taxa were compared to the synteny with the more distantly related species *E. festucae* using the F11 reference genome, and, as expected, whole genome alignments with *E. festucae* showed substantially more rearrangements among the seven chromosomes (Figs. S5–S7) compared to the alignments between *E. typhina* and *E. clarkii*. In gene-rich regions of homologous chromosomes, both macrosynteny (the overall chromosomal gene content) and microsynteny (the order of genes) were highly conserved (Fig. 4). The mapping of positions of 6923 single copy orthologous genes showed high gene co-linearity (Figs. S8–S10B), and in pairwise alignments of homologous chromosomes between *E. typhina* and *E. clarkii*, we found only one large rearrangement affecting more than 47 genes in a gene-rich region (Fig. S10).

### 4. Discussion

The genome size difference between *E. typhina* and *E. clarkii* is much larger than might be expected given they are closely related members of the same species complex. The difference between these sibling species is almost as large as the largest differences seen among all *Epichloe* species where genome size has been estimated from genome sequence data [56]. Although no clear association between genome size and phylogeny was detected, different strains of the same species tended to have more similar genome sizes [57]. Caution is needed when comparing genome size estimates using different sequencing technologies (short-read based genome assemblies may have misassembled or collapsed large repetitive regions), and yet it appears that these sibling species have diverged considerably in genome size compared to other...
sibling taxa. Despite the difference in overall size, gene content was similar in the two genomes; we found a similar number of genes most of which had high sequence homology between species and were identified as orthologous pairs (84.8% and 79.2% of genes in *E. typhina* and *E. clarkii*, respectively).

### 4.1. What leads to genome expansion?

RIP is a genomic defense mechanism that counteracts TE activity in fungal genomes [2]. RIP activity is linked to sexual reproduction (meiosis), and targets duplicated sequences such as TEs specifically inducing frequent C-to-T mutations which renders the TE inactive and causes a strong depletion in GC-content over time [58]. This generates AT-rich sequences that are confined to the repeat-rich regions of fungal genomes and contain a high frequency of TpA di-nucleotides, the primary product of RIP [7,59]. We found clear evidence of RIP in the form of AT-rich regions across both studied *Epichloë* genomes and the genome-wide extent of RIP was among the highest reported to date for sordariomycete fungi [see 7,51]. Interestingly, the AT-rich compartment was much larger in *E. clarkii* compared to *E. typhina*. This appears to result from both novel AT-rich regions that exist in one species but not the other, as well as expansions of shared AT-rich regions. Genome compartmentalization in eukaryotes is thought to result from uneven rates of TE insertions along the genome. TE insertions in gene-rich regions...
regions are more likely to be deleterious by disrupting a coding sequence or regulatory region of a gene and will be removed from the population by purifying selection. However, in largely non-coding sequence such as gene-sparse, repeat-rich regions, insertions may have little or no effect on fitness and thus they can persist and proliferate [8]. RIP can inactivate the TEs in these non-coding regions, and this generates blocks of AT-rich sequence. We show that the AT-rich compartments of Epichloe genomes are largely comprised of repetitive DNA. This is in line with the previously published E. festucae F1 reference genome, where preexisting repeats were interrupted by new TE-insertions generating nested long repeats, all of which have been heavily targeted by RIP [14].

The size difference in the AT-rich compartment, and coincidently genome size, between E. typhina and E. clarkii suggests that these genomes likely experienced independent surges of TE activity and subsequent inactivation by RIP. Expansion of the AT-rich compartment does not appear to simply be the result of elongated AT-rich regions, which could be caused by recombination processes, such as replication slippage increasing the size of existing AT-rich regions. Instead, the lengths of AT-rich regions are similar, and each genome contains species-specific AT-rich regions supporting new TE invasions. The divergent remodeling of the AT-rich compartment could be explained by differences in the ecological and evolutionary histories of these species, altering the relative strength of selection and genetic drift acting on genomes, via changes in effective population size between species or populations [60]. A reduction in effective population size, for example following a bottleneck, could lead to a stronger effect of genetic drift and a temporally weaker effect of purifying selection. This in turn can result in TE invasion of previously “protected” parts of the genome, giving rise to new repetitive regions that are targeted by RIP and results in the emergence of novel AT-rich regions [61]. Similarly, phases of asexual reproduction and the absence of recombination may reduce the efficacy of purifying selection and lead to TE proliferation [62]. Simultaneously, asexual reproduction renders meiotic genome defenses such as RIP inactive or ineffective and may thereby further promote the spread of TEs. Thus, fungal species that rarely undergo meiosis and have reduced RIP activity should exhibit more TE expansion compared to species with frequent sexual cycles and this is indeed supported by a growing number of studies (see [63]). The bottleneck associated with a host jump, invasion and range expansion is expected to drastically reduce effective population size, and could potentially be followed by a prolonged phase of clonal spread through the new host population, as has been suggested in other fungal pathogens [12,64–66], for example, when only one mating type is present in a founder population. In such a scenario, neither natural selection nor RIP could prevent the invasion of the genome by TEs, and the consequent genomic expansion until sexual reproduction (and RIP) becomes prevalent again with pressure to adapt and diversify. Given the differentially affected genomic composition in Epichloe genomes, we hypothesize that the divergence of E. typhina and E. clarkii may have been accompanied by a prolonged phase of asexual reproduction, and small effective population size in E. clarkii, leading to the observed genome expansion while TE proliferation remained more limited in E. typhina.

The similarity of the two Epichloe genomes in terms of their chromosome-level synteny in the gene-rich compartment (core genome) indicates recent divergence of the sibling species E. typhina and E. clarkii. In artificial laboratory settings, strains of these species remain interfertile, supporting only recent divergence [28]. Despite this, the AT-rich compartments have diverged at a much higher rate compared to the core genome, consistent with the two-speed genome model. TE-dynamics and the activity of RIP have played key roles in the genetic divergence of these species, and the genomic environment of AT-rich regions may create hotspots of genetic diversity that serve as cradles for rapid adaptive evolution contributing to local adaptation and divergence at the population level. Here we discuss the mechanisms that may have differentially shaped fungal genomes in their evolutionary past and outline how a bottleneck and relaxed selection could explain the observed patterns. We propose that AT-rich regions may be important for evolution at both the within- and between-species level, however, testing this hypothesis will require genomic analyses of more than a single isolate. Future studies that combine comparative and population genomic approaches in these sibling species, will provide important insight into the role of genome compartmentalization in shaping genomic diversity, and understanding how this influences adaptation and speciation.

5. Conclusions

We have assembled and annotated chromosome-level genome sequences for two Epichloe sibling species, E. typhina and E. clarkii. Both genomes display a remarkable bipartite organization, with a gene-rich core genome that is highly conserved between species, interspersed with blocks of divergent AT-rich sequence. AT-rich regions are the result
of repeat-induced point mutation (RIP) countering TE proliferation in fungal genomes. Variation of repeat content and consequently the AT-rich compartment between ecologically similar and evolutionarily close species of these fungi accounts for the large increase in genome size in *E. clarkii* compared to *E. typhina*. These results demonstrate the important role that TE-activity and genome defense mechanisms play in the genetic divergence of closely related taxa and highlights the role of such dynamic genomic compartments in shaping genome evolution. The genomic resource in the form of complete, high-quality reference genomes of two closely related *Epichloë* species, combined with population data, will serve as an ideal study system for the comparison of genomic architecture and patterns of sequence evolution both within and between species.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ygeno.2021.11.009.

**Data deposition**

Genome assemblies have been uploaded to GenBank (Accession numbers PRJN533210 for *E. typhina* Etv.1756 and PRJN533212 for *E. clarkii* Ecj.1605.22). PacBio and Illumina raw DNA reads and RNAseq data have been deposited in the Sequence Read Archive under the same respective accession numbers. A repository providing the scripts and files associated with this manuscript is available at https://github.com/adtreindl/Epichloe_genomes.

**Funding**

This research was supported by the Swiss National Science Foundation (grant number 31003A_169269), the New Zealand Tertiary Education Commission via a Bio-Protection Research Centre grant, and the Royal Society Te Apārangi Marsden Fund (grant 18-MAU-02). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Author statement**

We confirm that the revised manuscript has been read and approved by all named authors.

**Declaration of Competing Interest**

All authors declare that they have no competing interests.

**Acknowledgements**

We would like to kindly thank Claudia Michel and Beatrice Arnold for laboratory assistance, Daniel Berry for advice concerning RNA extractions, and Stefan Zoller and Niklaus Zemp for bioinformatic support. We would like to kindly thank Claudia Michel and Beatrice Arnold for laboratory assistance, Daniel Berry for advice concerning RNA extractions, and Stefan Zoller and Niklaus Zemp for bioinformatic support.