Genetic basis of gall formation and modes of reproduction in gall wasps (Hymenoptera, Cynipidae)

Ksenia Mozhaitseva

THÈSE DE DOCTORAT

NNT : 2023UPASB081

université PARIS-SACLAY

Genetic basis of gall formation and modes of reproduction in gall wasps (Hymenoptera, Cynipidae)

Bases génétiques de la formation des galles et des modes de reproduction chez les guêpes à galles (Hymenoptera, Cynipidae)

Thèse de doctorat de l'université Paris-Saclay

École doctorale n° 567 : sciences du végétal : du gène à l'écosystème (SEVE)
Spécialité de doctorat : Écologie
Graduate School : BioSphERA. Référent : Faculté des Sciences d'Orsay

Thèse préparée dans l'unité de recherche **EGCE** (Université Paris-Saclay, CNRS, IRD), sous la direction d'**Antoine BRANCA**, Maître de conférences

Thèse soutenue à Paris-Saclay, le 12 décembre 2023, par

Ksenia MOZHAITSEVA

Composition du Jury

Membres du jury avec voix délibérative

Emmanuelle BAUDRY Professeure Université Paris-Saclay, ESE IDEEV	Présidente
Christoph HAAG Directeur de Recherche CNRS, CEFE Montpellier	Rapporteur & Examinateur
Jean-Christophe SIMON Directeur de Recherche INRAE, IGEPP Le Rheu	Rapporteur & Examinateur
Héloïse BASTIDE Maîtresse de conférences Université Paris-Saclay, EGCE IDEEV	Examinatrice
Olivier PLANTARD Directeur de Recherche INRAE, BIOEPAR Nantes	Examinateur

ÉCOLE DOCTORALE
Sciences du végétal : du gène à l'écosystème (SEVE)

Titre : Bases génétiques de la formation des galles et modes de reproduction chez les guêpes à galles (Hymenoptera : Cynipidae)

Mots clés : guêpe à galles, Cynipidae, génomique des populations, transcriptomique, *thélytoquie*, parthénogenèse cyclique

Résumé : La sélection naturelle agit sur le phénotype, qui est principalement le produit du génotype. Il reste compliqué de lier un trait observable à sa base moléculaire car leurs interactions peuvent être complexes et inclure des interactions entre différents gènes et entre les gènes et l'environnement. Néanmoins, il existe diverses méthodologies pour associer le phénotype et le génotype.

La thèse visait à identifier les gènes candidats impliqués dans la formation des galles chez les guêpes à galles (Hymenoptera : Cynipidae). Une galle est un tissu ou un organe végétal nouveau et anormal, dont la formation représente une réponse spécifique provoquée par d'autres organismes. Les guêpes à galles sont un groupe d'insectes qui induisent de galles notamment sur les chênes et les églantiers. Les morphologies des galles des Cynipidae varient de légères modifications des tissus de la plante à des structures complexes. La galle est un trait adaptatif qui sert de ressource nutritionnelle et d'abri pour les guêpes à galles. Le processus de formation de la galle passe par trois étapes : l'initiation, la croissance et la maturation. Jusqu'à aujourd'hui, les facteurs provoquant l'initiation de la galle restent méconnus chez les Cynipidae. On suppose que des molécules effectrices produites par les œufs, les glandes salivaires des larves, les glandes à venin des femelles adultes et/ou leurs microorganismes symbiotiques peuvent induire la formation de la galle. Dans notre étude, nous avons recherché des traces de sélection dans les génomes des Cynipidae en utilisant la génomique des populations. Ensuite, nous les avons liées à l'expression des gènes lors de la formation des galles grâce à la transcriptomique. Nous avons étudié deux organismes modèles : *Diplolepis rosae* et *Cynips quercusfolii*. *D. rosae* est une guêpe à galles qui parasite les églantiers sauvages. Elle se reproduit principalement par la reproduction asexuée. *Cynips quercusfolii* est une guêpe à galles qui parasite les chênes. Cette espèce a deux générations alternantes par an : une génération sexuelle sur les bourgeons de printemps et une génération asexuée sur les feuilles. La génération asexuée est composée de deux types de femelles : celles qui ne produisent que des mâles et celles qui ne produisent que des femelles. D'abord, nous avons analysé des génomes et la structure des populations des deux espèces. Nous avons identifié deux lignées de *D. rosae* qui étaient fortement différenciées sur le territoire français. L'une des lignées présentait un niveau de recombinaison plus élevé et une hétérozygotie plus élevée par rapport à une autre lignée. Nous avons découvert que les gènes enrichis en fonctions liées aux traits mâles étaient sous sélection négative dans la lignée à plus forte fréquence de recombinaison, tandis que les mêmes gènes étaient sous sélection balancée ou relâchée dans la deuxième lignée. Chez *C. quercusfolii*, nous avons effectué une analyse préliminaire du génome et de la structure de la population. Une différence dans les types de femelles asexuées pourrait être dans la structure du génome et le niveau d'hétérozygotie dans des régions particulières du génome. Cependant, en appliquant la méthode de la génomique des populations, nous n'avons pas trouvé de gènes potentiellement impliqués dans l'induction de la galle chez *D. rosae* ni *C. quercusfolii.*

Ensuite, nous avons réalisé l'analyse le transcriptome de *D. rosae* à partir de galles collectées en conditions naturelles. Nous avons identifié 11916 gènes surexprimés au cours de la formation des galles. Nous avons démontré une surexpression des gènes codant des enzymes de dégradation de la paroi cellulaire végétale ce qui pourrait être lié à la formation des galles. De plus, ces gènes ont déjà été démontrés chez d'autres Cynipidae comme *Biorhiza pallida*. Nous avons aussi démontré une surexpression des gènes ayant les mêmes annotations fonctionnelles que ceux exprimés dans les venins de divers hyménoptères parasitoïdes. Ces gènes pourraient être impliqués dans la réponse immunitaire des guêpes à galles contre les parasitoïdes, le microbiome végétal et/ou le système de défense des plantes hôtes. Notre étude contribue au domaine de la recherche sur les Cynipidae. Nous avons démontré la structure génomique et la structure de population de *D. rosae* et *C. quercusfolii*, et nous les avons reliées à leurs modes de reproduction. Puis, nous avons constaté que la sélection peut agir sur différents traits en fonction du mode de reproduction chez les cynipidés sexués et asexués. Enfin, nous avons révélé un ensemble de gènes qui pourraient être impliqués dans la formation de la galle et dans la défense contre les ennemis naturels des Cynipidae.

Title : Genetic basis of gall formation and modes of reproduction in gall wasps (Hymenoptera: Cynipidae)

Keywords : gall wasp, Cynipidae, populations genomics, transcriptomics, thelytoky, cyclical parthenogenesis

Abstract : Natural selection acts on an phenotype that is mostly a product of the genotype. It is challenging to link an observable feature and its molecular basis because their relationships can be complex and include interactions between different genes and interactions between genes and the environment. Despite this complexity, there are various methodologies used to associate the phenotype and the genotype.

The thesis aimed to list candidate genes involved in gall formation in gall wasps (Hymenoptera: Cynipidae). Gall is an abnormal outgrowing novel plant tissue or organ whose formation represents a specific response caused by other organisms. Gall wasps are a group of gall-inducing insects parasitizing mostly oaks and wild roses. Cynipid gall morphologies vary from little plant tissue modifications to complex multi-chamber structures. The gall is an adaptive trait that serves as a nutritional resource and a shelter for gall wasps. The process of gall formation undergoes three stages: initiation, growth, and maturation. Until today, the initial triggers produced by gall wasps that provoke gall initiation remain unknown. Effector molecules produced by eggs, salivary glands of larvae, venom glands of female adults, and/or their symbiotic microorganisms are assumed to act in gall induction. In our study, we searched for traces of selection in cynipid genomes using population genomics and related them to gene expression during gall formation using transcriptomics. We used two model organisms: *Diplolepis rosae* and *Cynips quercusfolii*. *D. rosae* is a gall wasp that parasitizes wild roses. It exhibits a highly female-biased sex ratio, suggesting predominantly asexual reproduction. *Cynips quesrcusfolii* is a gall wasp that parasitizes oaks. This species has two alternating generations per year: an inconspicuous sexual generation on dormant buds and a well visible asexual generation on leaves. The asexual generation is presented by two types females: those that produce only males and those that produce only females.

Firstly, we performed population genomic analysis in both species. In *D. rosae*, we identified two highly differentiated peripatric lineages in France. One of the lineages showed a higher recombination rate and higher per-individual heterozygosity than the other lineage. We found that genes enriched in functions related to male traits were under negative selection in the more recombining lineage, whereas in the less recombining lineage, the same genes were under balancing or relaxed selection. In *C. quercusfolii*, we performed a preliminary analysis of the genome and population structure. A difference in types of asexual females could be in genome structure and level of heterozygosity in particular regions of the genome. However, by applying population genomics, we did not find candidate genes potentially involved in gall induction in *D. rosae* nor *C. quercusfolii*.

Secondly, we performed the transcriptome analysis of *D. rosae* from galls collected in natural conditions. We revealed 11,916 genes overexpressed during the observable gall formation period. Firstly, in the young developing larvae, there was an overexpression of genes encoding plant cell wall degrading enzymes which can be associated with gall formation. These genes have been already found in other Cynipidae like *Biorhiza pallida*. Secondly, we detected genes showing the same functional annotations as those expressed in venoms of various parasitic Hymenoptera. These genes could be involved in the gall wasp immune response against the parasitoids, plant microbiome, and/or host plant defense system.

While our study did not reveal obvious candidate genes acting at the initial stages of gall formation, it contributes to the broader field of cynipid research. Firstly, we demonstrated the genomic and population structure of two gall wasp species, *D. rosae* and *C. quercusfolii*, and related it with their modes of reproduction. Secondly, we found that selection can act upon different traits depending on the mode of reproduction in sexual-asexual cynipids. Lastly, we revealed a set of genes that could play a role in the defense against natural enemies of Cynipidae and gall formation.

«Делай не благодаря, а вопреки»

Эта диссертация полностью посвящена моему отчиму, человеку, благодаря которому я сейчас здесь. Спасибо Вам за то, что на протяжении пятнадцати лет Вы вкладывали в меня свое время, знания и опыт.

Cette thèse est entièrement dédiée à mon beau-père, grâce à qui je suis là aujourd'hui. Merci d'avoir investi quinze ans de votre temps, de vos connaissances et de votre expérience en moi.

REMERCIEMENTS

Tout d'abord, je remercie **Antoine Branca** de m'avoir choisie pour cette thèse et de m'offrir une opportunité d'obtenir une très belle expérience professionnelle et d'apprendre le sujet si challenging. Surtout, merci pour votre patience quand je disais ou écrivais des bêtises et de m'avoir expliqué les choses plusieurs fois. Merci également du fait que vous étiez toujours là au cas où, surtout dans le contexte actuel. C'est majoritairement grâce à vous que je pourrai poursuivre ma carrière aujourd'hui.

Bien sûr que je remercie **Zoé Tourrain** avec qui j'ai passé la première année. Merci pour ton soutien, ton énergie et ton humour. Merci pour ta très belle personnalité. Merci de ton attention à moi jusqu'à aujourd'hui et que tu me sortais de ma zone de confort.

Je remercie les membres de mon comité de thèse, **Florence Mougel**, **Jacqui Shykoff**, **Amir Yassin**, **Thierry Robert**, et surtout **Olivier Plantard** qui est également membre du jury. Merci d'avoir donné vos idées et partagé vos expériences pour qu'on puisse améliorer le travail.

Je remercie également les membres du jury de thèse, **Christoph Haag**, **Jean-Christophe Simon**, **Héloïse Bastide** et **Emmanuelle Baudry** d'avoir pris votre temps pour lire et évaluer mon travail.

Je remercie toute l'équipe ESE de m'avoir accueillie pour la première année de thèse, et surtout **Jeanne Ropars** et **Olivier Chauveau** pour le parrainage et de m'avoir guidée quand j'ai passé un moment difficile. Merci également à **Xavier Aubriot** pour les small talks.

Merci à toute l'équipe EGCE de m'avoir accueillie pour ma deuxième et troisième année et particulier à **Taiadjana Fortuna** et Mme **Laure Kaiser-Arnauld** pour votre gentillesse et petites conversations qui me faisaient du bien.

Merci à mon école doctorale d'avoir assuré le bon déroulement de ma thèse et en particulier à Mme **Sophie Nadot** de votre bienveillance et d'avoir toujours répondu aux questions si besoin.

Et merci aux stagiaires, **Anaïs Pourtoy** et **Xavier Vincent**, pour m'avoir offert l'opportunité de guider votre expérience et pour vos contributions que j'ai pu apporter à mon manuscrit.

Enfin, je remercie **Laurent**, mon cher ami parisien, qui a facilité ma vie à l'étranger et qui m'a beaucoup aidé dans la vie quotidienne. Merci d'avoir été là pour m'aider en quoi que ce soit.

TABLE OF CONTENTS

RESUME SUBSTANTIEL EN FRANCAIS

La sélection naturelle agit sur le phénotype, qui est principalement le produit du génotype. Il reste compliqué de lier un trait observable à sa base moléculaire car leurs interactions peuvent être complexes et inclure des interactions entre différents gènes et entre les gènes et l'environnement. Néanmoins, il existe diverses méthodologies pour associer le phénotype et le génotype.

La thèse visait à identifier les gènes candidats impliqués dans la formation des galles chez les guêpes à galles (Hymenoptera : Cynipidae).

Une galle est un tissu ou un organe végétal nouveau et anormal, dont la formation représente une réponse spécifique provoquée par d'autres organismes. Les guêpes à galles sont un groupe d'insectes qui induisent de galles notamment sur les chênes et les églantiers. Les morphologies des galles des Cynipidae varient de légères modifications des tissus de la plante à des structures complexes. La galle est un trait adaptatif qui sert de ressource nutritionnelle et d'abri pour les guêpes à galles. Le processus de formation de la galle passe par trois étapes : l'initiation, la croissance et la maturation. Jusqu'à aujourd'hui, les facteurs provoquant l'initiation de la galle restent méconnus chez les Cynipidae. On suppose que des molécules effectrices produites par les œufs, les glandes salivaires des larves, les glandes à venin des femelles adultes et/ou leurs microorganismes symbiotiques peuvent induire la formation de la galle.

Dans notre étude, nous avons recherché des traces de sélection dans les génomes des Cynipidae en utilisant la génomique des populations. Ensuite, nous les avons liées à l'expression des gènes lors de la formation des galles grâce à la transcriptomique.

Nous avons étudié deux organismes modèles : *Diplolepis rosae* et *Cynips quercusfolii*. *D. rosae* est une guêpe à galles qui parasite les églantiers sauvages. Elle se reproduit principalement par la reproduction asexuée. *Cynips quercusfolii* est une guêpe à galles qui parasite les chênes. Cette espèce a deux générations alternantes par an : une génération sexuelle sur les bourgeons de printemps et une génération asexuée sur les feuilles. La génération asexuée est composée de deux types de femelles : celles qui ne produisent que des mâles et celles qui ne produisent que des femelles.

D'abord, nous avons analysé des génomes et la structure des populations des deux espèces. Nous avons identifié deux lignées de *D. rosae* qui étaient fortement différenciées sur le territoire français. L'une des lignées présentait un niveau de recombinaison plus élevé et une hétérozygotie plus élevée par rapport à une autre lignée. Nous avons découvert que les gènes enrichis en fonctions liées aux traits mâles étaient sous sélection négative dans la lignée à plus forte fréquence de recombinaison, tandis que les mêmes gènes étaient sous sélection balancée ou relâchée dans la deuxième lignée. Chez *C. quercusfolii*, nous avons effectué une analyse préliminaire du génome et de la structure de la population. Une différence dans les types de femelles asexuées pourrait être dans

la structure du génome et le niveau d'hétérozygotie dans des régions particulières du génome. Cependant, en appliquant la méthode de la génomique des populations, nous n'avons pas trouvé de gènes potentiellement impliqués dans l'induction de la galle chez *D. rosae* ni *C. quercusfolii*.

Ensuite, nous avons réalisé l'analyse le transcriptome de *D. rosae* à partir de galles collectées en conditions naturelles. Nous avons identifié 11916 gènes surexprimés au cours de la formation des galles. Nous avons démontré une surexpression des gènes codant des enzymes de dégradation de la paroi cellulaire végétale ce qui pourrait être lié à la formation des galles. De plus, ces gènes ont déjà été démontrés chez d'autres Cynipidae comme *Biorhiza pallida*. Nous avons aussi démontré une surexpression des gènes ayant les mêmes annotations fonctionnelles que ceux exprimés dans les venins de divers hyménoptères parasitoïdes. Ces gènes pourraient être impliqués dans la réponse immunitaire des guêpes à galles contre les parasitoïdes, le microbiome végétal et/ou le système de défense des plantes hôtes.

Notre étude contribue au domaine de la recherche sur les Cynipidae. Nous avons démontré la structure génomique et la structure de population de *D. rosae* et *C. quercusfolii*, et nous les avons reliées à leurs modes de reproduction. Puis, nous avons constaté que la sélection peut agir sur différents traits en fonction du mode de reproduction chez les cynipidés sexués et asexués. Enfin, nous avons révélé un ensemble de gènes qui pourraient être impliqués dans la formation de la galle et dans la défense contre les ennemis naturels des Cynipidae.

GENERAL INTRODUCTION

Being in an ongoing arms race with natural enemies, species must run, that is to evolve, to keep in the same place, that is to survive (Van Valen 1973). This phenomenon is particularly evident in parasite-host interactions. Parasites must continually adapt to the selective pressures exerted by the defense systems of their hosts. Concurrently, hosts must undergo continuous adaptations to counter the manipulating and mimicking strategies applied by their parasites. Being under biotic selective pressures, species also face to a changing physical environment, such as climate shifts, tectonic events, and various random perturbations (Barnosky 1999). The adaptation to such a changing environment is manifested through a phenotype, known as all observable traits of an organism, that is encoded in its genes, known as a genotype. However, the relationships between the phenotype and the genotype are not simple: one observable feature can be encoded by several genes, and one gene can be involved in multiple phenotypic traits. For example, the trait such as a plant's resistance to a pathogen is usually encoded in multiple genes known as R genes; a mutation in one gene encoding the enzyme phenylalanine hydroxylase provokes phenylketonuria, a human disease showing multiple phenotypes including mental disorders, skin rashes, and pigment defects. Furthermore, the phenotype can be impacted by environmental conditions. For instance, in crocodiles, male and female phenotypes are determined by temperature: eggs incubated at low temperatures produce one sex, and eggs incubated at higher temperatures produce another sex.

To associate observable traits with genes encoding them in natural populations is often challenging but, despite this complexity, there are various methodologies employed to link the phenotype and the genotype using genomic data from natural populations whose lifecycle is difficult, or impossible, to control for experiments. One approach is to reconstruct the gains and the losses of genes along the phylogeny. Indeed, the emergence of certain key traits in a specific group of organisms, such as mammalian adaptations to the marine environment (Chikina et al. 2016) and the flower in flowering plants (Panchy et al. 2016), is usually correlated with major genomic changes like gene losses and duplications. The reverse ecology approach uses population genomics to identify different molecular signatures of selection acting upon genes of interest through the calculation of various metrics like genetic diversity within a pool of individuals (Li et al. 2008). The methods of functional genomics are focused in measuring gene expression at the DNA, RNA, protein, or metabolite level and associating it with the studied phenotype.

Inference from evolutionary genomic analyses can give candidates to reveal gene function using, for instance, gene editing technologies. These methods are designed to

introduce, modify, or delete the sequence encoding a candidate gene and understand how such modifications affect the phenotype.

Cynipid gall: an adaptive extended phenotype in gall wasps

Forty years ago, Richard Dawkins coined a new concept on the genotype-phenotype relationship: the extended phenotype (Dawkins 1982). He defined the phenotype as "all the effects of a gene on the world" (Dawkins 1982, 2004). In other words, the phenotype is an individual's observable traits together with all adaptive environmental modifications it makes. One of the most frequently cited extended phenotypes is physical features like caddis houses, beaver dams, and bird nests (Dawkins 1982). Besides, all types of intra- and interspecific interactions between organisms can be studied within this concept (Hunter 2009; Bailey 2012). Hence, the idea of the extended phenotype was a great stimulus for further discussions and investigations.

In our study, we focused on the spectacular instance of the extended phenotype, the product of a host-parasite interaction: a **gall**. The gall is an abnormal outgrowing plant tissue or organ induced by another organism for its own benefit (Rohfritsch and Shorthouse 1982; Meyer and Maresquelle 1983). Galls can be induced by different microorganisms, nematodes, and arthropods such as mites and insects. There is a wide range of attacked plants, from mosses to woody angiosperms, and diverse target organs, e.g. leaves, buds, flowers, and fruits (**fig. 1**).

One of the groups of gall-inducing insects is gall wasps (Hymenoptera: Cynipidae). Cynipidae includes at least 1400 species being the second largest group of gall-inducing insects after gall midges (Diptera: Cecidomyiidae), and occur on all continents, except the Antarctic (Ronquist et al. 2015). Apart from gall-inducing species, Cynipidae (s. lat.) includes inquilines and parasitoids (Hearn et al. 2023). Inquilines are cynipids that cannot induce gall formation *de novo* but occupy galls caused by gall inducers. Parasitoids are organisms that are parasites at an immature stage and free-living organisms at a mature stage; they feed on another organism, eventually killing it. In Cynipidae (s. lat.), parasitoids are presented by the family Figitidae and principally include parasitoids of Diptera and some parasitoids of cynipid gall inducers (Paretas-Martínez et al. 2011). The most recent phylogenetic studies (Ronquist et al. 2015; Blaimer et al. 2020; Hearn et al. 2023) demonstrated multiple transitions between the lifestyles within Cynipidae (s. lat.). The first scenario suggested parasitoid lifestyle to be an ancestral state giving rise to gall inducers that gave rise to inquilines. Subsequently, multiple transitions occurred between galler and inquiline lifestyles (Ronquist et al. 2015; Hearn et al. 2023). The second scenario suggested that gall inducers were the ancestral lifestyle. Inquilines were evolved from gall inducers, followed by multiple transitions between the two strategies, and parasitoid forms originated from both inquilines and gall inducers (Hearn et al. 2023).

In gall-inducing cynipids, each species typically attacks a single host plant species or genus. Most of the host plants belong to Fagaceae, Rosaceae, Asteraceae, Lamiaceae, and Papaveraceae. For instance, *Diplolepis* spp. induce galls in wild roses (*Rosa* spp.), the members of the Cynipini tribe induce galls in oaks (*Quercus* spp.), *Diastrophus* spp. attack *Rubus* spp., and *Pediaspis aceris* attacks *Acer* spp. In addition, each species causes a particular form of gall in a particular plant organ. Gall forms vary from a slight modification of plant tissue to complex multi-chamber structures (Stone and Schönrogge 2003) (**fig. 2**). Even so, the correlation between gall structure and the target organ, as well as the correlation between gall structure and the gall wasp taxon, is often unclear. For example, little flat *Neuroterus* spp. galls and spheric *Cynips* spp. galls can be found together on oak leaves; the galls having the same spheric structure are induced by *Diplolepis eglanteriae* (Diplolepidini) and *Diplolepis nervosa* (Diplolepidini) on wild roses (**fig. 3**), by *Belonocnema kinseyi* (Cynipini) on oaks, or by *Pediaspis aceris* (Pediaspidini) on sycamores. Finally, a great variety of gall morphologies can be found within one tribe like Cynipini attacking oak leaves, stems, and fruits and producing gall structures varying from simple plates to very complex morphologies such as 'artichoke' galls, 'hairy' galls, or such extraordinary forms as that of *A. dentimitratus*.

For gall wasps, the adaptive significance of diverse gall morphologies can be explained by their functions: **nutrition** resource, **microenvironment**, and **protection** against natural enemies (Stone and Schönrogge 2003). Firstly, gall tissue surrounding a chamber with a developing gall inducer contains an increased level of sugars, proteins, amino acids, lipids, and minerals and is physiologically similar to those of seeds (Schönrogge 2000; Giron et al. 2016). Therefore, the inner surface of this tissue can be enhanced, for example, by creating internal folds or multiple connections instead of one round chamber. Secondly, galls can decrease permeability by developing protective tissue containing more waxes. Furthermore, gall-inducing organisms are exposed to natural enemies such as predators and parasitoids. Various complex morphologies, such as 'hairy' galls, and thicker gall walls could be developed to reduce accessibility for the natural enemies of Cynipidae (Stone and Schönrogge 2003).

In spite of the diverse morphologies of cynipid galls, their inner structure and gall formation process are principally the same (**fig. 4**). The first stage of gall formation (galling) begins with female venom injection and oviposition into a particular plant organ and tissue, usually meristematic. It provokes gall **initiation**. After that, plant cells around the egg lyse and create a small chamber for a future larva. At the next step, **growing**, the larva continues to develop with simultaneous chamber enlargement and differentiation of meristematic plant cells becoming nutritive tissue. The nutritive layer is surrounded by vacuolate parenchyma also destined to be nutritive tissue. It is covered by the inner parenchyma. The outer layers are presented by sclerenchyma serving as protective tissue. The final step of gall formation is **maturation**: plant cell differentiation stops, and the

larva feeds on surrounding nutriments until entering diapause. At the same time, gall tissue becomes more and more sclerotized allowing the gall wasp to overwinter in it (Stone et al. 2002; Stone and Schönrogge 2003). The entire process of gall formation can last from the emergence of adults in early spring to overwintering nymphs or imagos in late autumn. Thus, gall formation represents a significant part of the cynipid life span.

Diverse life cycles of cynipid wasps

Cynipidae exhibit various types of life cycle that can include one sexual generation (**1)**, one asexual generation (**2**), or alternating sexual and asexual generations (bivoltine life cycle) (**3**) (Hood et al. 2018). Although some exceptions in the bivoltine life cycle, it commonly spans a single year. Each type of life cycle reflects a specific mode of reproduction of gall wasps. Similar to most other Hymenoptera, sexual reproduction (**1**) in Cynipidae is arrhenotokous parthenogenesis, where diploid females develop from fertilized eggs, while haploid males develop from unfertilized eggs. Cynipids can also undergo asexual reproduction (**2**), where virgin females produce females (Rabeling and Kronauer 2013) (**CHAPTER I**). Asexual reproduction is prevalent in herb gall wasps such as 'Aylacini' and rose gall wasps such as Diplolepidini. It is hypothesized (Stone et al. 2002) that the loss of sex in Cynipidae is probably due to *Wolbachia*, the most widespread intracellular bacterium in insect infecting from 25% to 70% of all species (Kozek and Rao 2007). In Hymenoptera, *Wolbachia* can provoke thelytoky, where females produce females without mating (Stouthamer and Kazmer 1994). In Cynipidae, several experiments (Plantard et al. 1998, 1999) linked the absence of males in studied populations with the presence of *Wolbachia* in females. The last type of life cycle is more complex and includes alternating sexual and asexual generations (**fig. 5**) (**CHAPTER III**). Bivoltine life cycle (**3**) is common in oak (Cynipini tribe) and sycamore gall wasps (Pediaspini tribe). In general, both generations are morphologically and ecologically distinguishable and induce the galls with different morphologies. Sexual generation typically develops in spring and is often inconspicuous or even unknown in certain species (Pujade-Villar et al. 2001; Hood et al. 2018). The galls are usually invisible and often develop in hidden or small host plant organs, such as roots (Zhang et al. 2021) or catkins (Brandão-Dias et al. 2022). Conversely, asexual generation is widely described. Besides, many oak Cynipidae are known only from asexually reproducing stage (Pujade-Villar et al. 2001). It typically develops from summer to winter. These galls are clearly visible and primarily form on host plant tissue such as leaves and buds.

In the bivoltine life cycle, an intriguing observation is that it is not lined to *Wolbachia* infection. Indeed, even when infected, oak Cynipidae still produce males, thereby maintaining sexual reproduction. However, the sexual generation may be lost resulting in a complete asexual reproduction (Pujade-Villar et al. 2001). This could appear in certain environmental conditions. For instance, obligatory asexual reproduction was

demonstrated in an invasive North American form of *Plagiotrochus suberi*, an originally European species reproducing by cyclical parthenogenesis (Bailey and Stange 1966). In *Andricus mukaigawae*, only the asexual form was present in warmer climate conditions, whereas both alternating generations were present in colder conditions (Abe 1986). The capacity of oak cynipids to reproduce only asexually or sexually raised the question of why such complex bivoltine life cycle can be maintained. From an ecological point of view, a species with alternating generations can experience various environmental conditions. Firstly, it can expand its ecological niche in space and time, thereby reducing intraspecific competition (Wolda 1988). Secondly, the alternation of generations might reduce the selective pressure from natural enemies such as predators and parasites (Pujade-Villar et al. 2001). We can also consider the persistence of such life cycle in the context of the benefits of both modes of reproduction. Asexual reproduction is advantageous in stable environment, while sexual reproduction provides genetic diversity through recombination when exposed to unfavorable environments and engaged in evolutionary arm races with natural enemies (Williams 1966; Van Valen 1973; Maynard-Smith 1978; Hamilton 1980).

Cynipid gall is not an enemy-free space

The gall is not only a product of interaction between a gall wasp and its host plant but a whole ecosystem that represents various complex interactions between the host plant, the gall wasp, its inquilines, parasitoids, hyperparasitoids ('parasitoids of parasitoids'), opportunistic species, as well as symbiotic microorganisms of all these organisms (Stone at al. 2002). For instance, in the *Diplolepis rosae* gall complex, Williams (2013) listed one inquiline species (*Pericistus brandtii*), and sixteen hymenopteran parasitoid species (**fig. 6**), and in asexual galls of *Belonocnema treatae*, Forbes et al. (2016) revealed 25 species of its natural enemies from Hymenoptera, Coleoptera, Diptera, and Lepidoptera orders.

Although the gall is a heterogeneous community, its morphological, anatomical, and biochemical features aim primarily to protect and feed the gall wasp and, therefore, reduce selective pressure from parasitoids, pathogens, and the plant defense system. In order to reduce selective pressure imposed by parasitoid attacks, various traits such as gall size, gall wall thickness, and number of chambers may undergo natural selection (László and Tóthmérész 2013). For example, in *Dryocosmus kuriphilus* parasitizing the chestnut and *Diplolepis rosae* parasitizing wild roses, the parasitism rate reduces with an increase in gall size, thereby suggesting that selection favors larger galls (László and Tóthmérész 2013; Gil-Tapetado et al. 2021). Gall wasps may also suffer from endophytic microorganisms. Thus, traits such as gall wall thickness, concentration of antimicrobial substances composing gall tissue, defense molecules produced by the immune system of gall wasps, and other biological features aimed at inhibiting pathogen activity may be under selection. For instance, in the *D. kuriphilus* gall community, the gall wasp seems to

be affected by various fungi that compose the host plant microbiome: larger galls and galls with thicker sclerenchyma layers contained fewer fungal lesions (Cooper and Rieske, 2010). Another study (Martinson et al. 2022) revealed that genes involved in the synthesis of antimicrobial molecules (flavonoids) were overexpressed in the external layers of gall tissue in red oaks parasitized by *Dryocosmus quercuspalustris*. The authors hypothesized that hyperproduction of flavonoids could inhibit pathogenic activity of the host plant microbiome. Furthermore, the authors demonstrated the downregulation of genes associated with the host plant defense system in the internal gall layers that directly surrounded the larva, thereby showing that gall wasps have developed tools to inhibit host plant immunity.

Study question: how gall formation affects the genome of gall wasps in regards of their mode of reproduction?

Unlike gall-inducing microorganisms, whose activity leads to the formation of unstructured tumor tissue in plants (Gätjens-Boniche 2019), and herbivore animals that simply damage plant tissues, cynipid wasps employ mechanisms that not only inhibit plant immunity but also cause significant alterations in host plant metabolism and reprogram cell differentiation, ultimately resulting in the development of a new structured organ. It is fascinating how such a complex phenotype as the cynipid gall can be initiated by molecular triggers originating from the insect beyond its body. Today, the genetic basis, i.e. the genes encoding the molecular triggers that provoke gall formation in Cynipidae, remains a challenging question.

Multiple hypotheses, that do not mutually exclude one another, propose potential molecular triggers of gall formation (**CHAPTER II**). Firstly, as gall wasps manipulate plant cell differentiation and metabolic pathways, they are suggested to produce various molecules that mimic phytohormones (Yamaguchi et al., 2012). Secondly, since gall wasp larvae reside within the gall chamber, their salivary gland secretions may contribute to gall formation by damaging plant tissues (Hearn et al., 2019). Thirdly, gall initiation is believed to occur at the moment of oviposition, suggesting that components detected in female venom glands and egg secretions could serve as potential triggers (Cambier et al., 2019; Gobbo et al., 2020). Finally, we can also consider the role of symbiotic microorganisms or their genes acquired through horizontal gene transfer, which may also play a part in gall initiation (Bartlett and Connor, 2014).

When considering coevolution between the gall wasps and the host plants, gall formation represents an adaptive trait that implies signatures of positive and balancing selection in cynipid genomes. For instance, in the gall-inducing species *Synergus itoensis* reproducing mostly by arrhenotokous parthenogenesis, certain genes being associated with insect-plant interaction were found to be under positive selection (Gobbo et al. 2020), as elaborated in detail in **CHAPTER II**. The signatures of positive selection can be

identified through various methods, which is demonstrated in the example of selective sweeps. The selective sweep refers to a scenario where a new, advantageous mutation arises in a population and increases in frequency leading to the fixation. This results in a decrease in genetic diversity within the surrounding genomic region, where neutral mutations are linked to the advantageous allele. Consequently, this localized signal can be identified as a decline in genetic diversity in this region when compared to the genomic background.

In our analyses, it is crucial to take into account the predominantly asexual mode of reproduction of the studied cynipids (**CHAPTER I**). Indeed, the asexual mode of reproduction influences genome structure in terms of homozygosity and linkage disequilibrium, which may result in a challenge to properly detect positive selection. For instance, *Diplolepis rosae* primarily reproduces through thelytokous parthenogenesis via gamete duplication (**CHAPTER I**). This process involves restoring diploidy by duplicating initially haploid genetic material. Consequently, the female offspring will be entirely homozygous and will reproduce clonally, resulting in complete linkage of all alleles in the genome. Other cynipids, such as *D. eglanteriae* and *D. nervosa*, exhibit apomixis (Sanderson, 1988), where virgin females produce daughter offspring through mitosis. This maintains the ancestral homozygosity in the genome and all alleles in complete linkage. Thus, the absence of recombination, which breaks links between alleles, leads to the fixation of the beneficial mutation along with the entire genomic background. As a result, detecting positive selection in non-recombining lineages becomes challenging due to the difficulty to distinguish localized signals, such as selective sweeps, from the genomic background.

Furthermore, asexual reproduction leads to a decrease in the efficiency of selection. Due to high linkage disequilibrium, the advantageous mutation linked to a deleterious allele may be quickly purged from the population. In addition, in the absence of recombination, the advantageous mutation may take more time to be fixed. For instance, in a sexually reproducing lineage, consider two beneficial mutations that occur independently. Thanks to recombination, the genotype containing both alleles will be predominant, which can be detected. However, in the non-recombining lineage or one where recombination is highly reduced, we will observe several competing sub-populations exhibiting separately beneficial mutations. This will dilute the signal of selection.

In addition, cynipid populations are supposed to have a reduced effective population size (Zayed, 2004; **CHAPTER I**). In such lineages, allele frequencies are influenced by random fluctuations rather than selection. Thus, due to genetic drift, the advantageous mutation can either be fixed rapidly or be lost by chance.

However, thelytokous hymenopteran populations still occasionally produce rare males (Rabeling and Kronauer 2013), indicating the presence of sexual reproduction and,

consequently, recombination that breaks linked alleles. That is why assessing the efficiency of recombination within a predominantly thelytokous species was necessary to adapt our approaches to search for signatures of selection.

Identifying candidate genes responsible for gall formation: initial strategy

In our study, we aimed to contribute to the research focused on detecting potential triggers involved in gall initiation in cynipid wasps. Our strategy employed two distinct approaches: **population genomics** and **transcriptomics**.

We employed population genomics to detect signatures for selection within a gall wasp genome. Given the ongoing evolutionary arms race (Van Valen 1973) between gall wasps and their host plants, we hypothesized that gall wasps must generate new genetic variants to evade the plant immune system and successfully manipulate the host plant metabolism. Therefore, we sought to identify evidence of balancing selection maintaining elevated genetic diversity, and positive selection favoring novel beneficial mutations. To do so, we calculated various metrics, such as nucleotide diversity and the relative frequency of polymorphisms, across the genome. Additionally, considering that closely related species like *Diplolepis* spp. often induce galls with entirely different morphologies (**Figure**), we expected to detect signatures of divergent selection, notably an elevated relative frequency of divergent sites.

Through transcriptomics, we aimed to conduct a time-series greenhouse experiment in order to measure gene expression levels in various gall wasp tissues at different steps of gall formation (**fig. 7**). Our aim was to identify genes that are upregulated during the early stages (0 – 72 hours, 1 week, and 1 month) compared to the later stages.

Subsequently, we expected to refine the list of candidate genes by comparing two sets of candidates identified through population genomics and transcriptomics. On the one hand, we sought to exclude genes that might be under selection but not be expressed during the initial stages of gall formation. On the other hand, the set of genes under positive and balancing selection could help to exclude conserved genes overexpressed during gall initiation but likely responsible for other processes, such as insect development.

We used two model organisms: ***Diplolepis rosae*** and ***Cynips quercusfolii***. *D. rosae* is a univoltine gall wasp species that parasitizes wild roses (*Rosa spp*.). It exhibits a highly female-biased sex ratio, suggesting predominantly asexual reproduction. *C. quercusfolii* is a bivoltine gall wasp species that parasitizes oaks (*Quercus spp*.). In this project, we focused on studying the females of the asexual generation. Biology of both species is described in the respective chapters.

FIGURES

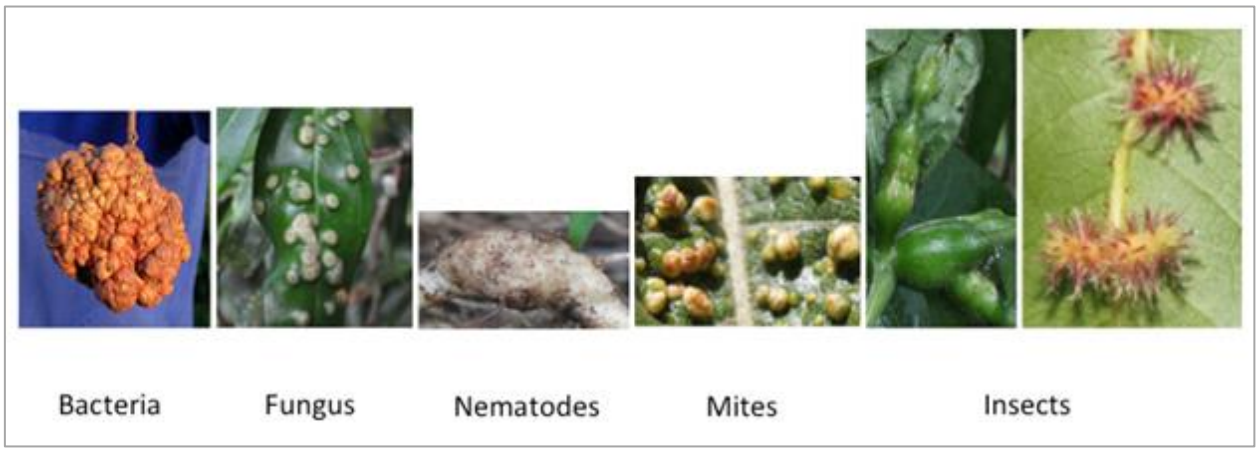


Fig. 1. Examples of plant galls induced by different taxons of gall-inducing organisms. Adapted from (Gätjens-Boniche 2019).

Fig. 2. Morphological diversity of oak galls induced by gall wasps from Cynipini tribe on different oak species (*Quercus* spp.) (Pascual-Alvarado et al. 2017). 1: Gall on *Quercus conzati* (petiole). 2: *Q. sideroxyla* (branch). 3 and 4: *Q. magnoliifolia* (leaves). 5: *Q. resinosa* (catkin). 6: *Q. crassifolia* (bud). 7: *Q. ocoteifolia* (yema). 8: *Q. uxoris* (leaf). 9: *Q. sp.* (branch). 10: *Q. castanea* (branch). 11: *Q. segoviensis* (catkin). 12: *Q. polymorpha* (leaf). 13: *Q. arizonica* (leaf). 14. *Q. segoviensis* (leaf). 15: *Q. cupreata* (root). 16: *Q. obtusata* (leaf). 17: *Q. microphylla* (acorn). 18: *Q. laurina* (branch). 19: *Q. gregii* (leaf). 20: *Q. laeta* (bud). 21: *Q. rugosa* (bud). 22: *Q. conspersa* (branch). 23: *Q. frutex* (leaf). 24: *Q. viminea* (bud). 25: *Q. deserticola* (leaf).

Fig. 3. Galls induced by *Diplolepis* spp. A: *D. rosae*. B: *D. mayri*. C: *D. spinosissimae*. D: *D. eglanteriae/nervosa*. Adapted from (Sardon-Gutierrez et al. 2021).

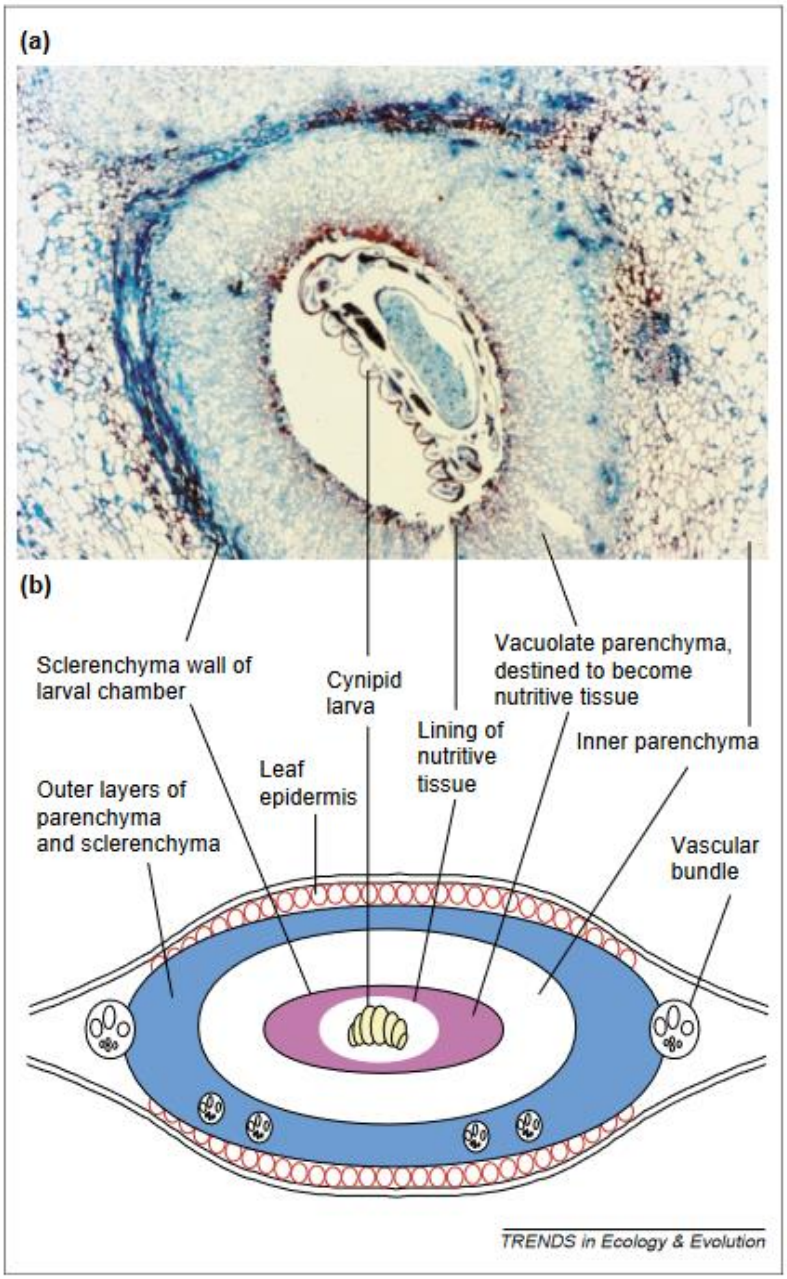


Fig. 4. Anatomical structure of cynipid gall (Stone and Schönrogge 2003).

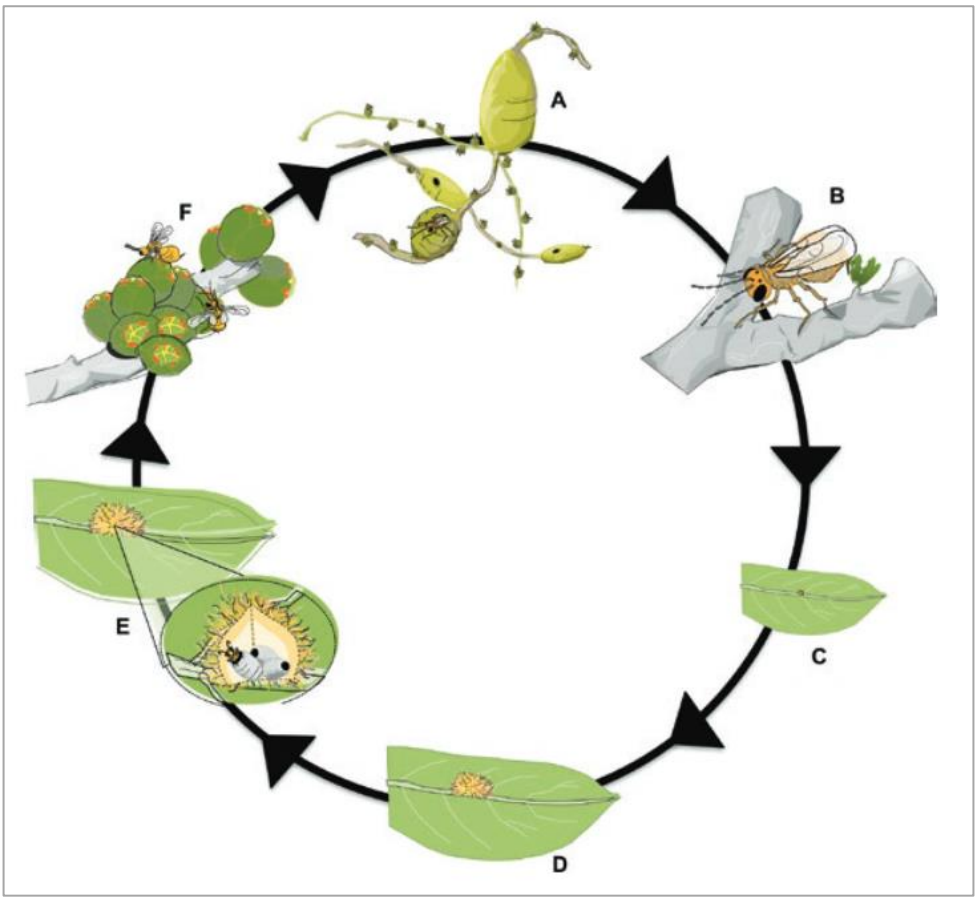


Fig. 5. Bivoltine life cycle of *Andricus quercuslanigera* (Hood et al. 2018). A: sexual generation emerging from oak catkin galls in mid-March to early April. **B**: Oviposition by the female of sexual generation into leaves. **C**: asexual gall in mid-summer. **D**: asexual gall in late summer. **E**: Developing female of asexual generation. **F**: asexual female emerging from September to late February.

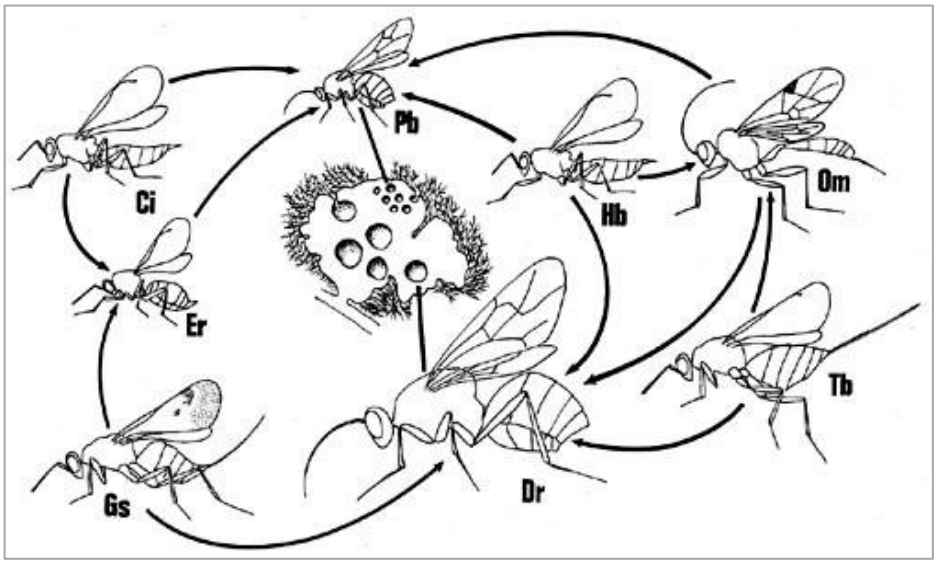


Fig. 6. Example of parasitic complex of *Diplolepis rosae* bedeguar gall *(*Schilthuizen and Stouthamer 1998*)*. Dr: *Diplolepis rosae*. **Pb**: *Periclistus brandtii*. **Om***: Orthopelma mediator*. **Hb***: Habrocytus bedeguaris*. **Tb**: *Torymus bedeguaris*. **Gs**: *Glyphomerus stigma*. **Er**: *Eurytoma rosae*. **Ci**: *Caenacis inflexa*.

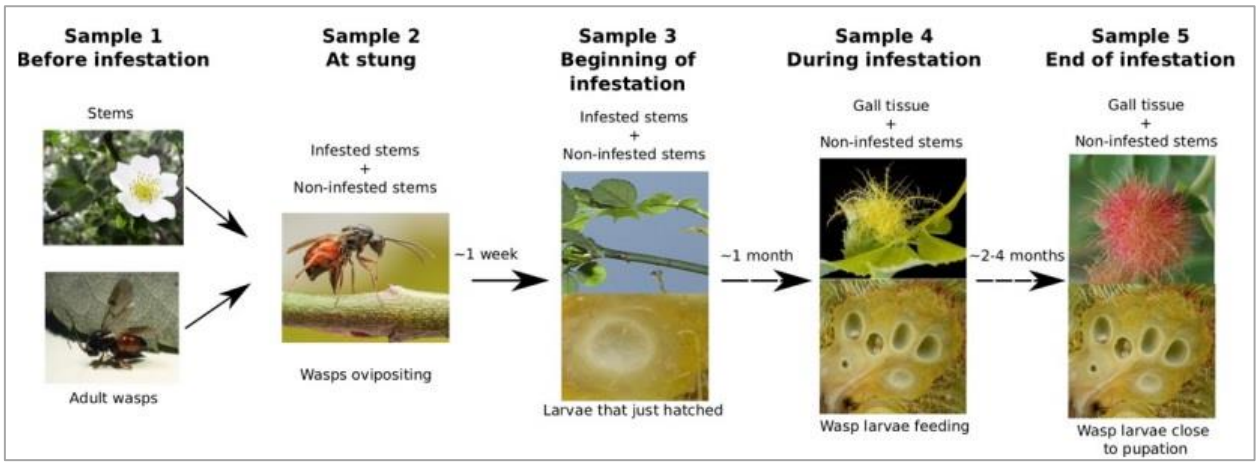


Fig. 7. Initial transcriptomics sampling scheme (A. Branca, ANR-19-CE02-0008 project proposal).

AIM AND SCOPE OF THE STUDY

The thesis aimed to reveal candidate genes involved in gall formation in gall wasps (Hymenoptera: Cynipidae) in the context of their modes of reproduction. We studied the genome structure of two cynipid genomes and searched for traces of selection using population genomics and related them to gene expression during gall formation using transcriptomics.

The thesis consists of three chapters:

CHAPTER I: population genomics of *Diplolepis rosae*. In this part, we studied the genome and population structure of *D. rosae* and searched the traces of selection by population genomic approach. We demonstrated how genome structure of *D. rosae* was influenced by mostly asexual mode of reproduction. Using a high-quality reference genome, we identified specific patterns of differentiation, genetic diversity, and homozygosity. The results have been published in ***Genome Biology and Evolution***.

CHAPTER II: transcriptome analysis of *D. rosae*. We identified genes overexpressed at the early stages of gall formation and performed a test for selection for the genes of interest. The results have been submitted to ***Insect Molecular Ecology***.

CHAPTER III: preliminary results of the genome and population structure of *Cynips quercusfolii*. We searched on highly heterozygous and highly homozygous genome regions and structural variations in seven asexual *C. quercusfolii* females. We hypothesized that a particular genome structure underlined the type of asexual female and the maintenance of the life cycle in heterogonic Cynipidae.

Each chapter consists of an Introduction specialized on the given question, Materials and Methods, Results, Discussion, Conclusion, Tables, and Figures. Tables and Figures of each chapter have their own numbering. The thesis ends with a General Discussion and Conclusion.

CHAPTER I. Population genomics of the mostly thelytokous *Diplolepis rosae* reveals population-specific selection for sex

ARTICLE SUMMARY

CHAPTER I is dedicated to population genomics of *Diplolepis rosae* and presented in the form of scientific article published in ***Genome Biology and Evolution***. In this section, we described the population structure of *D. rosae* in France and searched for specific patterns of differentiation, recombination, and homozygosity in its genome in the context of its mode of reproduction. *Diplolepis rosae* is a gall wasp inducing bedeguars on the dog roses (**fig. 1**) and reproducing mostly by thelytokous parthenogenesis (thelytoky) when virgin females produce only females (**fig. 2**). In order to assess the effect of thelytoky on the genome and population structure of *D. rosae,* we assembled a high-quality reference genome using Oxford Nanopore long-read technology and sequenced 17 samples collected in France with high-coverage Illumina reads. We revealed two highly differentiated *D. rosae* peripatric lineages that differed in the level of recombination and homozygosity. One of the *D. rosae* lineages showed a higher recombination rate and higher per-individual heterozygosity. Conversely, another *D. rosae* lineage showed a higher number of widespread (several Mbp) runs of homozygosity, i.e. contiguous regions of the genome where an individual is homozygous across all sites (Ceballos et al. 2018). In several regions of the *D. rosae* genome, we noticed a decrease in Fst and a simultaneous change in nucleotide diversity (π). Therefore, we developed a composite score (CS) metric that summarized Fst and π, which enabled us to detect the regions showing a drop in Fst with a simultaneous increase (positive CS) or decrease (negative CS) in π. A negative outlier CS value reflected negative selection, and a positive outlier CS value could be the result of balancing or relaxed selection. We observed that several traits are under contrasted selective regimes in both *D. rosae* lineages. In the more recombining lineage, the genes enriched in the 'sperm competition', 'insemination', and 'copulation' gene ontology terms, the functions related to male traits, showed the signatures of purifying selection. In the less recombining lineage, the same genes were under balancing or relaxed selection. We hypothesized that the genes involved in male traits and important for efficient recombination through the production of males. Sexual reproduction could generate genetic diversity via allele combinations. Higher genetic diversity would then give an advantage in survival in the face of a highly prevalent and diverse community of parasitoids attacking *D. rosae* (Stille 1984; Rizzo and Massa 2006; Todorov et al. 2012). Thus, the more recombining ('sexual') *D. rosae* lineage creates genetic diversity for adaptation to changing environments, and the less recombining ('asexual') lineage provides a rapid, less costly reproduction (Williams 1966; Maynard-Smith 1978; Hamilton 1980).

REMARK

The results described in **ARTICLE** were supported by additional studies of the M1 internship students Anaïs Pourtoy and Xavier Vincent (Université Paris-Sud, Biodiversité Ecologie Evolution). The aim of their internships was to assess the differences between the two *D. rosae* lineages in terms of the gall hymenopteran community composition, parasite rate, and gall characteristics. The principal result was that in the more recombining *D. rosae* lineage, the proportion of parasitized gall chambers in a gall was 30%, whereas in the less recombining lineage, the percentage of parasitized cells was 80% (**fig. 3**). Therefore, this provides an additional argument for the hypothesis advanced in **ARTICLE**: the maintenance of sexual reproduction in the mostly thelytokous *D. rosae* provides genetic diversity against its natural enemies.

FIGURES

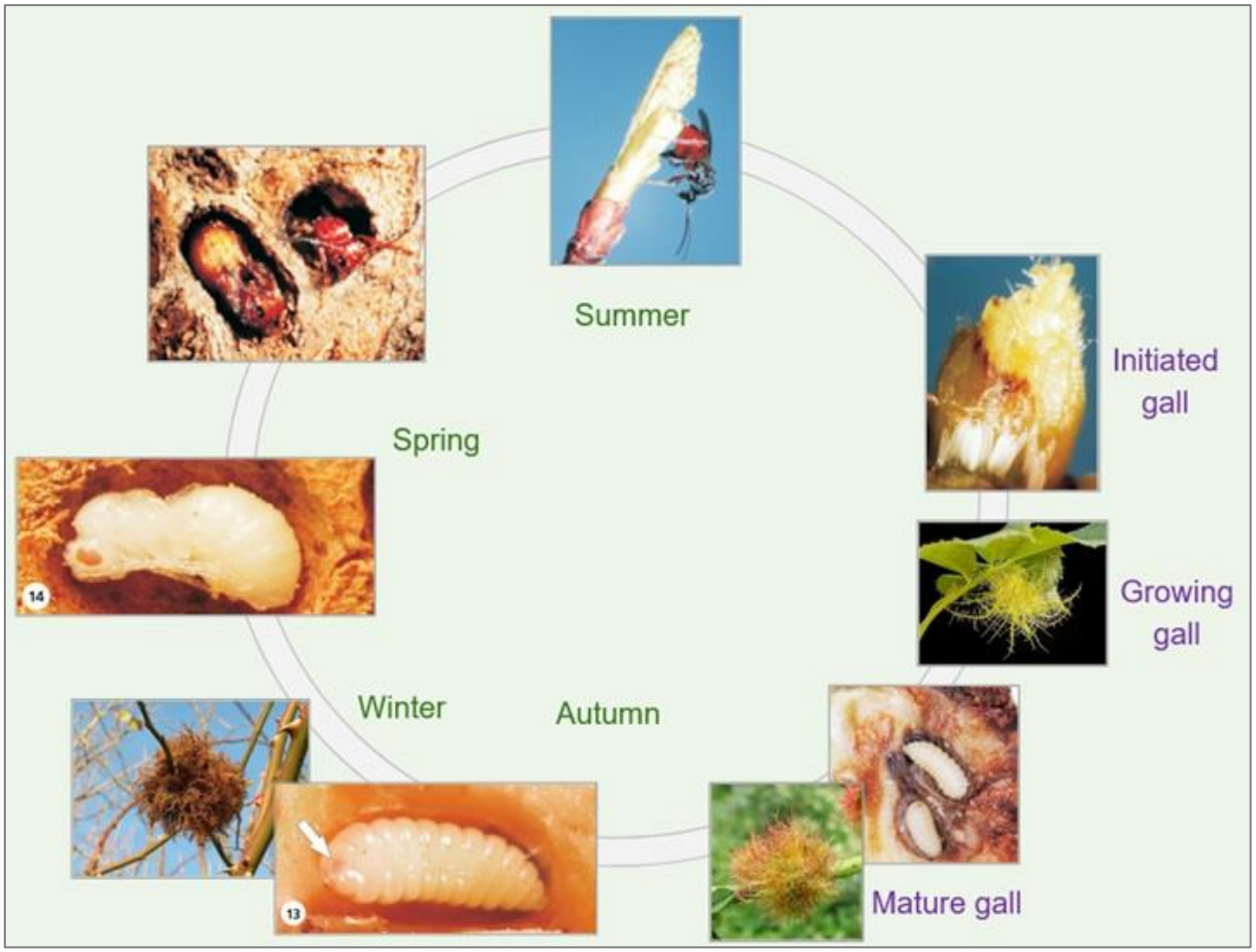


Fig. 1. *Diplolepis rosae* life cycle (adapted from Shorthouse and Floate 2010). *D. rosae* (Linnaeus, 1758) (Cynipidae: Diplolepidini) is a holarctic gall wasp species that induces galls called bedeguars in wild roses *Rosa* spp. sect. Caninae (Rosaceae). It has one generation per year (Shorthouse and Floate 2010) and reproduces mostly by thelytokous parthenogenesis. Adults emerge from May to June and are synchronized with the development of suitable host plant tissues for gall induction (Shorthouse and Floate 2010). After emergence, the females immediately oviposit into epidermal plant cells located between the developing leaflets of an expanding bud (Bronner 1985). Once eggs are laid, gall tissue begins to develop. One gall can contain up to one hundred larvae (Rizzo and Massa 2006; laboratory observations). The feeding larvae are surrounded by gall cells and spend at pre-nymph stage before the next spring (Shorthouse and Floate 2010).

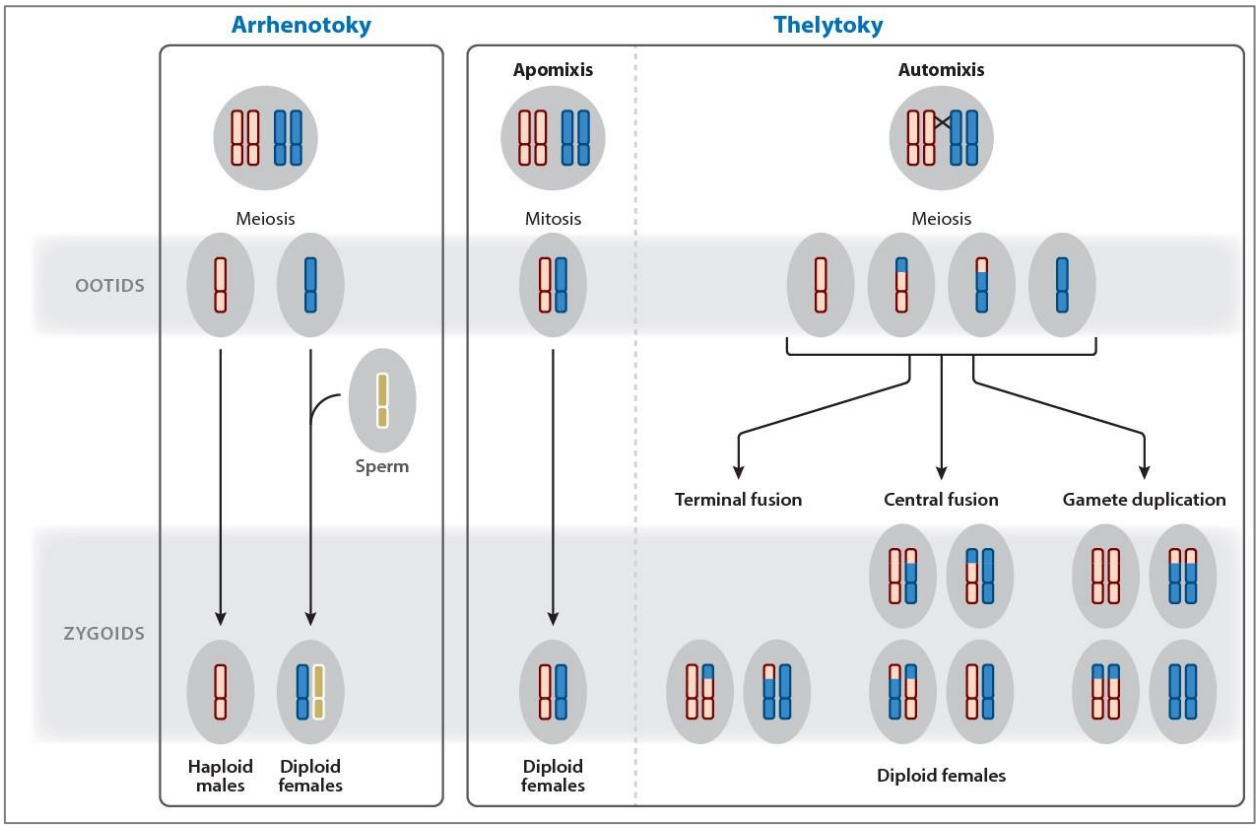


Fig. 2. Reproductive modes in Hymenoptera (Rabeling and Kronauer 2013). Arrhenotoky: development of haploid males and diploid females from unfertilized and fertilized eggs, respectively. Thelytoky: production of diploid females by a virgin female via apomixis (mitotic division of ootids) or automixis (fusion or duplication of meiotic products).

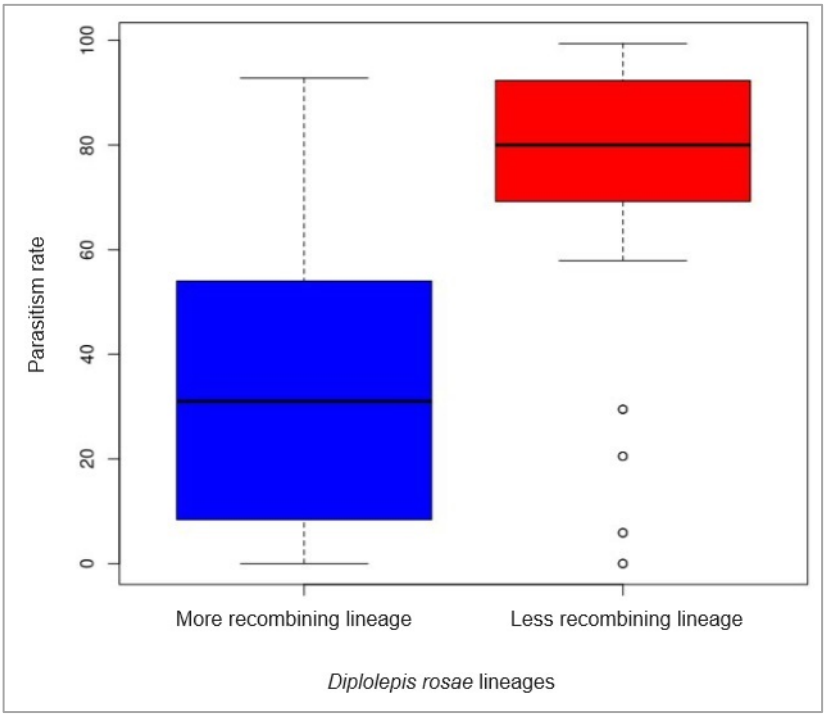


Fig. 3. Parasitism rate in the more recombining and the less recombining *Diplolepis rosae* lineages. Binomial generalized linear model (ANOVA chi2, $p < 0.001$). Adapted from A. Pourtoy.

ARTICLE

Population genomics of the mostly thelytokous Diplolepis rosae (Linnaeus, 1758) (Hymenoptera: Cynipidae) reveals population-specific selection for sex

Ksenia Mozhaitseva[1]*, Zoé Tourrain[1], and Antoine Branca[1]

[1] Laboratoire Evolution, Génomes, Comportement, Ecologie, l'Institut Diversité, Ecologie et Evolution du Vivant, Université Paris-Saclay, Gif-sur-Yvette, France

*Author for correspondence: Ksenia Mozhaitseva, Laboratoire Evolution, Génomes, Comportement, Ecologie, l'Institut Diversité, Ecologie et Evolution du Vivant, Université Paris-Saclay, Gif-sur-Yvette, France, ksenia.mozhaitseva@universite-paris-saclay.fr

Genome Biology and Evolution, evad185, https://doi.org/10.1093/gbe/evad185

Published:

13 October 2023

Abstract

In Hymenoptera, arrhenotokous parthenogenesis (arrhenotoky) is a common reproductive mode. Thelytokous parthenogenesis (thelytoky), when virgin females produce only females, is less common and is found in several taxa. In our study, we assessed the efficacy of recombination and the effect of thelytoky on the genome structure of *Diplolepis rosae*, a gall wasp producing bedeguars in dog roses. We assembled a high-quality reference genome using Oxford Nanopore long-read technology and sequenced 17 samples collected in France with high-coverage Illumina reads. We found two *D. rosae* peripatric lineages that differed in the level of recombination and homozygosity. One of the *D. rosae* lineages showed a recombination rate that was 13.2 times higher and per-individual heterozygosity that was 1.6 times higher. In the more recombining lineage, the genes enriched in functions related to male traits ('sperm competition', 'insemination', and 'copulation' gene ontology terms) showed signals of purifying selection, whereas in the less recombining lineage, the same genes showed traces pointing towards balancing or relaxed selection. Thus, although *D. rosae* reproduces mainly by thelytoky, selection may act to maintain sexual reproduction.

Significance

Many organisms can alternate between sexual and asexual reproduction in different ways. Sexual reproduction is essential to create genetic diversity for adaptation to changing environments, whereas asexual reproduction is important in the short term and in stable environments. Using genomic data, we demonstrated the existence of two lineages in the rose bedeguar wasp *Diplolepis rosae* previously shown to reproduce mainly by thelytokous parthenogenesis, giving almost only females. One of the lineages showed higher recombination, higher heterozygosity, and genes involved in male traits under purifying selection. This could be linked to the expected advantages of maintaining sexual reproduction in natural populations.

Keywords: *Diplolepis rosae*, thelytoky, recombination, runs of homozygosity, selection, population genomics.

Introduction

It is commonly asserted that sexual reproduction, i.e. recombination, is advantageous because it provides the opportunity for organisms to generate genetic diversity, enabling them to overcome environmental perturbations. However, recombination can also break linked advantageous alleles and broaden the variance around a fitness optimum, leading to the counterselection of sexual lineages (Crow and Kimura 1965). Therefore, the coexistence of sexually and asexually reproducing forms, which exhibit variability in the frequency of recombination and heterozygosity, is related to the cost of sex from both short- and long-term perspectives (Williams 1966; Maynard-Smith 1978; Hamilton 1980). Species that possess both sexual and asexual lineages are especially interesting for studying the conditions that determine their selection for recombination.

In Hymenoptera, some species reproduce both sexually and asexually. Arrhenotokous parthenogenesis (arrhenotoky) is the ancestral sexual reproductive mode (Heimpel and De Boer 2008; Rabeling and Kronauer 2013). Diploid females develop from fertilised eggs, whereas haploid males develop from unfertilised eggs. Another less frequent mode of reproduction is thelytokous parthenogenesis (thelytoky), which relates to asexuality, in which virgin females produce only females (Heimpel and De Boer 2008). Thelytoky can be encoded in the genomes of hymenopterans (Wenseleers and Billen 2000; Belshaw and Quicke 2003; Engelstädter et al. 2011; Foray et al. 2013; Capdevielle-Dulac et al. 2022) or induced by endosymbionts (Stouthamer et al. 1990; Stouthamer and Kazmer 1994). Genetically based thelytoky exists in the form of automixis, i.e. gamete fusion or gamete duplication after meiosis, and in the form of apomixis, i.e. mitotic division of ootids with no meiosis (Heimpel and De Boer 2008; Schön et al. 2009; Queffelec et al. 2021). Notably, automictic Hymenopteran females can still produce rare males (Stenberg and Saura 2009). The presence of males in thelytokous populations can also be explained by a failure of the mechanism inducing thelytoky or by occasional gene flow between the arrhenotokous and thelytokous populations (Stouthamer and Kazmer 1994; Engelstädter et al. 2011). Thelytoky leads to a decrease in recombination and individual genetic diversity compared to arrhenotoky; therefore, each mode of reproduction will leave a contrasting pattern of genetic diversity across the genome (Tvedte et al. 2019).

Endosymbiont-induced thelytoky is caused by endocytoplasmic bacteria, such as *Rickettsia*, *Cardinium*, and *Wolbachia* (Werren et al. 2008; Giorgini et al. 2009; Adachi-Hagimori et al. 2011). *Wolbachia* is the most widespread intracellular parasite infecting 25 to 70% of all insect species (Kozek and Rao 2007). *Wolbachia*-induced thelytoky has been extensively studied in *Trichogramma* spp. (Stouthamer et al. 1990; Stouthamer and Kazmer 1994), where thelytokous females exposed to antibiotic treatment or high temperatures produced males and arrhenotokous females (Stouthamer et al. 1990). Thus far, *Wolbachia*-mediated thelytoky has only been shown to be induced via gamete

duplication, i.e. the failure of chromosome segregation in unfertilised eggs during anaphase I and subsequent duplication of terminal meiotic products. This mechanism leads to completely homozygous females (Stouthamer and Kazmer 1994). Hence, endosymbiont-induced thelytoky may result in the rapid spread of several locally adapted parthenogenetic lineages, each originating from a single female, which might lead to speciation (Werren 1998; Schilder et al. 1999; Bordenstein 2003; Adachi-Hagimori et al. 2011).

Diplolepis rosae (Hymenoptera: Cynipidae) is a cynipid wasp that parasitises wild dog roses (*Rosa* spp. section *caninae*) and causes specific plant outgrowths called rose bedeguar galls (Shorthouse and Floate 2010; Giron et al. 2016). One cytological study of meiosis in *D. rosae* showed that it reproduces by thelytokous parthenogenesis via gamete duplication; after anaphase II, one of the haploid ootids enters mitotic division. Thereafter, the fusion of the two daughter products results in the restoration of diploidy, thereby leading to completely homozygous females (Stille and Dävring 1980). Indeed, *D. rosae* individuals collected from several European locations consisted of almost only females, with the proportion of males varying from 1 to 4% (Nordlander 1973). A further study based on the electrophoresis of 27 isozymes showed that *D. rosae* females sampled in Sweden, Germany, and Greece were completely homozygous (Stille 1985). Additionally, females of *D. rosae* have been shown to be infected with *Wolbachia*, which was supposed to induce gamete duplication thelytoky in this species (Van Meer et al. 1995; Schilthuizen and Stouthamer 1998; Plantard et al. 1999).

In summary, the previously published *D. rosae* studies demonstrated that populations were strongly female-biased. Based on several genetic markers, females were mostly homozygous. The prevalence of females and homozygosity were associated with *Wolbachia* infection. Therefore, we expected that *Wolbachia* infection in *D. rosae* might lead to gamete duplication thelytoky and a fine-scale population structure coupled with high homozygosity across the genome. However, any selection for recombination and genetic diversity would act to maintain sexual reproduction and leave different signatures on the genome.

Thus, the objective of our study was to assess the effect of thelytoky on the genome of *D. rosae*. First, using a high-quality reference genome of *D. rosae*, we investigated whether thelytoky led to a fine-scale population structure of *D. rosae* in France. Second, we analysed the patterns of diversity, homozygosity, and recombination to show the consequences of thelytoky on the genome of *D. rosae*. Third, we assessed whether selection favoured recombination over thelytoky by searching for regions that showed specific patterns of differentiation, recombination, and homozygosity.

Results

Genome structure. The total length of the *D. rosae* reference genome assembly was estimated at 760.7 Mbp, with a total number of sequences of 757, an N50 of 7,663,408 bp, the largest scaffold of 33,454,033 bp, and an L50 of 25 (**supplementary table S1**). Repetitive sequences represented 69.28% of the genome assembly: 48.50% unclassified repeats, 11.40% retroelements, 8.66% DNA transposons, and 0.72% other repeats (rolling circles, small RNA, satellites, simple repeats, and small complexity repeats) (**supplementary table S2**). Using the BUSCO hymenoptera_odb10 dataset (Manni et al. 2021), we found a genome completeness of 91.8% of complete and single-copy genes, 0.5% of complete and duplicated genes, 1.6% of fragmented genes, and 6.1% of missing genes. The total number of genes predicted by BRAKER2 (Hoff et al. 2019) *in silico* was 20,301, of which 14,559 were partially or fully annotated. The assembled genome of *D. rosae* is available at the NCBI platform (https://www.ncbi.nlm.nih.gov/): a DDBJ/ENA/GenBank accession is JAPYXD000000000 (version JAPYXD000000000.1), a BioProject accession is PRJNA914909, and a BioSample identifier is SAMN32363506.

Population structure. The most likely population structure for *D. rosae* in France was two lineages, lineage 1 and lineage 2. This was based on the number of examined lineages (K), which was equal to 2 (Raj et al. 2014) (**fig. 1**). One individual, *D. rosae*-652, showed admixed ancestry, with 86.2% of its genome assigned to lineage 2 and 13.8% to lineage 1. The negative value of F3-admixture statistics (Patterson et al. 2012; Peter 2016) ($F3 = -0.0797 \pm 0.0150$, $z = -5.29$, $p < 0.0001$) suggests that this individual is admixed between the *D. rosae* lineages related to lineage 1 and lineage 2. The additional genotyping confirmed two lineages distributed in France (**fig. 2, supplementary fig. S1**).

Demographic scenario. According to the inference method (*dadi*) based on the joint site frequency spectrum of genetic variants (Gutenkunst et al. 2009), the most probable demographic scenario for *D. rosae* was a bottleneck of an ancestral population, followed by exponential growth, then a split into two populations with no gene flow between them (lowest AIC = 74419.2, **table S3**, **supplementary fig. S2**). The bottleneck time and split time were estimated at 0.22*2Neff generations ago and 0.20*2Neff generations ago, respectively, where Neff is the effective population size (diploid individuals). The likelihood ratio test showed no significant difference between the models with and without the migration parameter (Dadj = 0.37, p = 0.28). The migration rate (m) was estimated at 0.81/(2*Neff) migrants per generation. Additional examination by approximate Bayesian computation (ABC) (Csilléry et al. 2012) confirmed the demographic model inferred by *dadi* to provide a good fit to the data (1 million simulations, Goodness of Fit, dist. = 3365.0, p = 0.21). The prediction errors for all model

parameters (effective population size, bottleneck time, and split time) were close to 1 (**supplementary table S4**). Posteriors had the same or broader distributions as the priors, except for effective population sizes (**supplementary fig. S3**). The population sizes for both *D. rosae* lineages were estimated to be 500 diploid individuals.

Population genomic statistics. The fixation index Fst and the absolute divergence Dxy values between the two lineages were 0.81 ± 0.25 and 0.0033 ± 0.0024, respectively. Lineage 1 was characterised by a population-scaled recombination rate (ρ) that was 13.2 times higher (U = 511,350, z = 61.8, p = 0.0001) and a heterozygosity (H) that was 1.6 times higher (U = 9, z = 2.55, p = 0.0108) than lineage 2 (**fig. 3**). In lineage 1, the median Tajima's D value did not differ from zero (one-sample Wilcoxon test, comparison with the median u = 0: W = 9.5e+08, p = 0.22) (**fig. 3**). In lineage 2, the distribution of Tajima's D showed a slight excess of negative values (Tajima's D median = −0.84, comparison with the median u = 0: W = 5.5e+08, p < 0.0001). There was no correlation between nucleotide diversity and recombination rate in lineage 1 (Pearson's correlation r = 0.0626, t = 0.259, p = 0.799) or lineage 2 (r = 0.0287, t = 0.119, p = 0.907). There was no correlation between the protein-coding sequence density and nucleotide diversity in lineage 1 (r = −0.0142, t = −0.218, p = 0.827) or lineage 2 (r = −0.0237, t = −0.365, p = 0.715) (**supplementary fig. S4**).

Runs of homozygosity. Lineage 1 showed more frequent short (0.01–0.1 Mbp) runs of homozygosity (ROHs) than lineage 2 (Mann–Whitney U test: U = 9, z = 2.55, p = 0.0108) (**Fig. S5**). In contrast, lineage 2 showed more widespread and larger ROHs (0.1–0.5 Mbp ROHs: U = 9, z = 2.454, p = 0.0141; 0.5–1.0 Mbp ROHs: U = 9, z = 2.550, p = 0.0108; >1 Mbp ROHs: U = 9, z = 2.550, p = 0.0108). In lineage 1, the ROHs covered 74 to 82% of the genome assembly length, whereas in lineage 2, the coverage varied from 83 to 91%. In the admixed individual, *D. rosae*-652 (**fig. 1**), the ROHs covered 52% of the genome.

Detection of homozygous genomic regions with low differentiation. In several regions of the *D. rosae* genome, longer ROHs overlapped with a decrease in Fst and a change in genetic diversity (π) (**fig. 4**). Therefore, we developed a composite score (CS) that summarises Fst and π. It enabled us to detect the regions showing a drop in Fst with a simultaneous increase (positive CS) or decrease (negative CS) in π. We detected a positive CS (**table 1**) associated with an increase in recombination rate in scaffold 204 (26.0 and 28.0 Mbp regions) (**fig. 5**), scaffold 325 (5.1–7.5 Mbp) (**fig. 6**), and scaffold 414 (5.5–9.0 Mbp) (**fig. 6**) in lineage 1: some alleles segregated in both *D. rosae* lineages but recombination broke the linkage between alleles in lineage 1. In scaffold 313 (6.1–6.2 Mbp) (**fig. 5**), scaffold 325 (2.5–3.0 Mbp) (**fig. 6**), and scaffold 762 (12.5–20.0 Mbp) (**fig. 7**), the positive CS values (**table 1**) in both lineages showed an increase in diversity π but a decrease in recombination rate ρ, indicating several haplotypes segregating in both

lineages with complete linkage. Furthermore, in lineage 1, scaffold 325 (2.5–3.0 Mbp Mbp) (**fig. 6**) and scaffold 414 (5.8–6.0 Mbp) (**fig. 6**) demonstrated a higher π_N/π_S ratio, indicating an increase in non-synonymous site diversity relative to synonymous site diversity. In scaffold 204 (30.0 Mbp) (**fig. 5**), scaffold 313 (4.1 Mbp) (**fig. 5**), and scaffold 414 (3.0 Mbp) (**fig. 6**), we detected regions with negative CS outliers (**table 1**) associated with a lower recombination rate ρ: some linked alleles segregated in both *D. rosae* lineages. In scaffold 204 (30.1–33.0 Mbp) (**fig. 5**), scaffold 313 (0.0–4.0 Mbp and 4.2–5.9 Mbp), and scaffold 523 (0.0–2.0 Mbp) (**fig. 7**), we observed the opposite trend with the CS values (**table 1**): positive outliers in lineage 1 but negative outliers in lineage 2. Gene set enrichment analysis (GSEA) of the genome regions showing outlier composite score values revealed several significantly enriched Gene Ontology (GO) terms (**supplementary table S5**). In 'Biological Process' GO, the term 'commissural neuron axon guidance' corresponded to a negative score outlier in lineage 2 and a positive score outlier in lineage 1 (**supplementary table S5**, **supplementary fig. S6**). The terms 'sperm competition', 'insemination', and 'copulation' corresponded to a negative score outlier in lineage 1 (**supplementary table S5**, **supplementary fig. S7**) and a positive score outlier in lineage 2. In 'Molecular Function' GO, the terms 'metalloendopeptidase activity' and 'metallopeptidase activity' corresponded to negative outliers in lineage 1 (**supplementary table S5**, **supplementary fig. S8**) and positive outliers in lineage 2. The term 'commissural neuron axon guidance' annotated a group of genes found in scaffold 313 (1.7 – 2.7 Mbp region) (**fig. 5**). The terms 'sperm competition', 'insemination', 'copulation', 'metalloendopeptidase activity', and 'metallopeptidase activity' matched genes found in scaffold 204 (28.7–28.9 Mbp region) (**fig. 5**).

***Wolbachia* identification.** *Wolbachia* contigs belonging to supergroups A and B were identified in 5 and 14 *D. rosae* individuals, respectively (**supplementary table S6**). In *D. rosae*-078 and *D. rosae*-117, both supergroups were identified in the same bin in the metagenome assembly. The population structure of *Wolbachia* did not follow that of *D. rosae* (**supplementary fig. S9–S13**). The average number of *Wolbachia* supergroup A copies per cell (coverage) varied from 0 to 2.4. The coverage of *Wolbachia* supergroup B varied from 0 to 14.0. There was no significant difference between the *D. rosae* lineages in terms of *Wolbachia* coverage (supergroup A: Mann–Whitney U test, $U = 25.5$, $z = 0.276$, $p = 0.782$; supergroup B: $U = 20$, $z = 0.868$, $p = 0.385$). The '*Wolbachia* supergroup A/B' ratio was the same in both *D. rosae* lineages ($U = 23$, $z = 0.621$, $p = 0.535$).

Discussion

Our results showed that two peripatric, highly differentiated lineages of *D. rosae* (hereafter 'lineage 1' and 'lineage 2') occurred in France. The two lineages differed in homozygosity level and recombination rate. In lineage 1, the recombination rate was 13.2 times higher, and per-individual heterozygosity was 1.6 times higher compared to lineage 2. Thus, *D. rosae* is another illustrative example of hymenopteran species exhibiting population variability in heterozygosity and the frequency of recombination. Another example of such organisms is *Asobara japonica* showing both *Wolbachia*-free arrhenotokous lineages and *Wolbachia*-infected thelytokous lineages, with *Wolbachia* infection becoming obligatory for asexual reproduction (Kremer et al. 2009). Within Cynipidae s. lat., *Diplolepis spinosissimae*, a closely related species to *D. rosae*, exhibited geographically distant *Wolbachia*-infected and *Wolbachia*-free lineages in France (Plantard et al. 1998). In the *Wolbachia*-infected coastal lineages, males represented 1.3% of individuals, and no heterozygous females were found. Conversely, in the *Wolbachia*-free continental lineages, males represented 21–29% of individuals, and 78–96% of females were heterozygous. In *Leptopilina clavipes*, a figitid parasitoid member of Cynipidae s. lat., studied across several European countries, Pannebakker et al. (2004a) revealed several northern clonal *Wolbachia*-infected lineages and southern *Wolbachia*-free lineages. *Wolbachia*-infected *L. clavipes* was shown to reproduce by gamete duplication thelytoky (Pannebakker et al. 2004). Interestingly, *Wolbachia*-free and *Wolbachia*-infected lineages were highly differentiated but showed the same level of genetic variation (Pannebakker et al. 2004a). Furthermore, the authors demonstrated that several different thelytokous lineages were infected with the same *Wolbachia* strain. Based on these results, the authors concluded that various *L. clavipes* thelytokous lineages originated from an arrhenotokous lineage through the horizontal transmission of *Wolbachia* from infected to uninfected lineages. In another parasitic wasp, *Ventura canescens*, Schneider et al. (2002) also demonstrated the coexistence of arrhenotokous and thelytokous lineages. However, contrary to the other species, thelytokous *V. canescens* lineages were not associated with *Wolbachia* infection: thelytoky was shown to be genetically based and occur via central fusion (Beukeboom and Pijnacker 2000). An intriguing remark made by Schneider et al. (2002) was that the thelytokous *V. canescens* lineages were predominant in man-made habitats, such as grain stores infested by lepidopteran pests, whereas the arrhenotokous lineages prevailed in outdoor habitats. This suggests that clonal reproduction could be advantageous in stable environments, whereas sexual reproduction could be favoured in unstable environments.

In our study, both *D. rosae* lineages were recovered across France, with no clear geographic distribution. Nonetheless, lineage 1 seemed to be more frequent in the north of France, notably in the Île-de-France region, whereas lineage 2 seemed to be more

frequent in the southeast of France. The first structuring factor that could reflect the distribution of *D. rosae* is the host plant genotype, but this is unlikely to be the case. Indeed, different *D. rosae* genotypes have been found to parasitise the same *R. canina* genotype, and the same *D. rosae* genotype has also been observed in different *Rosa* spp. (Stille 1985; Kohnen et al. 2011). Furthermore, in our study, in three instances among 61 localities, individuals from different lineages were found on the same branch of the same host plant specimen (**fig. 2**). Nevertheless, it is possible that each population exploits different microhabitats in *Rosa* shrubs (for example, different elevations of parasitised leaf buds or the orientation of the galls towards the sun). The second structuring factor could be an infection with different *Wolbachia* strains. However, we found that the *Wolbachia* infection did not explain the population structure of *D. rosae* as it could be for the other hymenopterans (Plantard et al. 1998; Pannebakker et al. 2004a; Kremer et al. 2009). Both lineages were infected with the same *Wolbachia* strains belonging to supergroups A and B in varying quantities. They could acquire *Wolbachia* infection independently by horizontal transmission either via parasitoids (Werren et al. 1995) or via the same host plants through the phloem (Schilthuizen and Stouthamer 1998). Lastly, each *D. rosae* lineage could be under top-down control and attacked by a specific parasite community that controls population density. Local variation in the presence of parasite species would then determine the prevalence of each lineage. Indeed, gall wasps are exposed to intense attacks by natural enemies, notably hymenopteran parasitoids and cynipid inquilines (Stille 1984; Rizzo and Massa 2006; Todorov et al. 2012; Laszlo et al. 2014). For instance, Rizzo and Massa (2006) showed that, on average, only 5% of the *D. rosae* progeny per gall survived in each generation. Therefore, parasitism avoidance is under strong selection (Stone and Schönrogge 2003) and might act as a top-down structuring factor.

According to the best supported demographic scenario, the *D. rosae* lineages originated from a bottleneck of an ancestral population that split into two populations with no migration since the split. However, we found the same demographic scenario, with the migration parameter being equally probable as the simpler model. Indeed, we observed one admixed individual, *D. rosae*-652 (**fig. 1**), belonging to lineage 2 but showing 13.8% of the genome assigned to lineage 1. This could be due to a shared ancestral polymorphism or recent gene flow. Gene flow between the two lineages is more likely because this individual showed the highest heterozygosity (**fig. 3**) and the lowest number of widespread ROHs (**supplementary fig. S5**). Therefore, we concluded that the two *D. rosae* lineages were well-differentiated lineages with rare gene flow.

Background selection is expected to be a major force acting on polymorphism in thelytokous organism, such as *D. rosae*, because populations consist of almost only females (Nordlander 1973; Stille 1985; Plantard et al. 1999; laboratory observations) and show low levels of heterozygosity (an average of 3 and 2 heterozygous sites per 10 kbp in lineage 1 and lineage 2, respectively). A negative correlation between the density of

protein-coding sequences and nucleotide diversity (π) usually indicates selection at linked sites, such as background selection, because deleterious alleles are more likely to occur in genes (Payseur and Nachman 2002). However, we did not observe such a pattern, which is probably the result of an intense purge of deleterious alleles (**supplementary fig. S4**). Indeed, there are two reasons to expect the genetic load to be low in *D. rosae*. First, in Hymenoptera, deleterious mutations are usually purged in haploid males. Second, in thelytokous organisms, deleterious mutations are also purged in highly homozygous females (Pearcy et al. 2006).

Thelytoky also led to a reduced recombination rate and widespread ROHs, which we observed in both lineages (**fig. 4–7**). This is concordant with previous studies suggesting that *D. rosae* reproduces mostly by thelytoky (Nordlander 1973; Stille and Dävring 1980; Stille 1985; Plantard et al. 1999). As previously stated by Stille and Dävring (1980) and discussed in other works (Van Meer et al. 1995; Schilthuizen and Stouthamer 1998; Plantard et al. 1999), we suggest that gamete duplication is the main mechanism of thelytoky in *D. rosae*. Indeed, gamete duplication leads to complete homozygosity across the genome. Although we did not observe complete homozygosity, we detected ROHs reaching several Mbp and making up to 90% of the genome, which is concordant with frequent gamete duplication thelytoky. Thus, other mechanisms of thelytoky retaining local heterozygosity (Rabeling and Kronauer 2013), e.g. via central or terminal fusion or apomixis, are less likely to occur or less frequent. Another clue that *D. rosae* reproduces by gamete duplication thelytoky is the presence of *Wolbachia*. *Wolbachia* is known to induce thelytoky via gamete duplication in several Hymenoptera (Stouthamer et al. 1990; Stouthamer and Kazmer 1994; Gottlieb et al. 2002). However, we acknowledge that we have only indirect proof of gamete duplication thelytoky. Only laboratory experiments should be able to provide clear-cut evidence (for example, Stouthamer et al. 1990).

Both thelytokous *D. rosae* lineages differed in the intensity of recombination and heterozygosity. Lineage 1 showed higher heterozygosity and higher recombination rates than lineage 2. This could reflect sexual reproduction or the use of an alternative mechanism of thelytoky. To reveal the factors that could explain the difference between the two lineages, we scanned the *D. rosae* genome for regions deviating from the background. Using the composite score (CS) summarising Fst and π values, we unravelled regions with low differentiation and low or high nucleotide diversity. Indeed, the decrease in Fst indicates low differentiation between lineages, which is maintained by local gene flow or by selection. If it is associated with low diversity, it could indicate negative selection on linked sites that have maintained low differentiation between lineages since their split. If it is associated with high diversity, it could reflect balancing selection when the same diversity is maintained since the population split. It could also be a relaxed negative selection that allows a new allele to emerge that would have been

previously deleterious. We acknowledge that this approach has some limitations in detecting selection. First, it is challenging to distinguish between balancing selection/relaxed constraint and the effect of recombination in the regions showing a local increase in nucleotide diversity. Therefore, we estimated the recombination rate across the genome to contrast the diversity level and recombination. Therefore, when we observed a high composite score in the non-recombining regions, putative balancing selection or relaxed constraint was likely at play. Second, the method is limited in terms of negative and balancing selection. In the case of positive selection or negative selection acting on different haplotypes in the different lineages, we observed high Fst, low π, and CS close to 0. However, the Fst between the two lineages was already close to 1 in most of the genome (**fig. 4**). Therefore, to properly detect positive selection, one should perform other analyses (for example, the McDonald–Kreitman test that needs an outgroup) that are not affected by the mode of reproduction or demography (McDonald and Kreitman 1991; Eyre-Walker 2002; Parsch et al. 2009).

Despite these limitations, we revealed that the two lineages demonstrated opposite composite score outliers in several ROHs, which could be due to contrasted selective regimes. In the more recombining lineage 1, the genes associated with 'sperm competition', 'insemination', 'copulation', 'metallopeptidase activity', and 'metalloendopeptidase activity' terms showed patterns typical of negative selection, whereas in the less recombining lineage 2, they would be putatively under balancing or relaxed selection. Genes associated with the term 'commissural neuron axon guidance' showed the opposite trend with genomic signatures of balancing/relaxed selection in lineage 1 and negative selection in lineage 2. The changes in the selective regime of those specific genes could be related to different selective processes acting in the *D. rosae* lineages and could explain the maintenance of the two lineages despite a shared habitat. In lineage 1, genes involved in sexual reproduction are important for efficient recombination through the production of males, thereby generating genetic diversity in terms of allele combination. Higher genetic diversity would then give an advantage in survival in the face of a highly prevalent and diverse community of parasites (Stille 1984; Rizzo and Massa 2006; Todorov et al. 2012; Laszlo et al. 2014). Our suggestion could be supported by the study of Rizzo and Massa (2006), that showed a possible association between average parasitism and the percentage of *D. rosae* males that emerged. Parasitism rates were 30.5% and 34.3% in two *D. rosae* populations consisting of 21% and 15.6% males, respectively. In other populations, there were no males, and the average parasitism rate reached 57.6% (varied from 12.5% to 100%) (Rizzo and Massa 2006). In lineage 2, showing more widespread ROHs and lower recombination, the candidate genes from the terms 'sperm competition', 'insemination', and 'copulation' showed signatures of either balancing selection or relaxed selection. Regarding the extreme level of homozygosity in lineage 2, male-related genes could become

unnecessary and be under relaxed selection. However, balancing selection could also take place to maintain male alleles through frequency-dependent selection, as alleles important for male function would increase in frequency when selection for recombination occurs. In contrast to male-related function, the genes involved in 'commissural neuron axon guidance' showed a signature typical of purifying selection in lineage 2 but balancing selection or relaxed constraint in lineage 1. We hypothesise that a less recombining lineage shows a conservative host-searching behavioural pattern. Indeed, Ramirez-Romero et al. (2012) demonstrated differences in host-searching behaviour between thelytokous and arrhenotokous populations of *Odontosema anastrephae* (Hymenoptera: Cynipidae s. lat.), the figitid parasitoid of the fruit fly. In this study, thelytokous females showed a basic behavioural sequence when exploring the odour source, whereas arrhenotokous females demonstrated more complex behaviour (Ramirez-Romero et al. 2012).

Conclusion

We demonstrated the existence of two highly differentiated peripatric lineages of *D. rosae* that differ in the level of recombination and homozygosity across the genome. The maintenance of these lineages might be due to selection acting upon different traits. Further research could explore the natural history of the two lineages by examining the following questions. Does each lineage really differ in the frequency of male production? Are there differences in the parasite communities attacking them that would explain top-down control? Do they show the same phenology? These are key questions to answer to determine which selective pressure leads to the maintenance of recombination in *D. rosae*, despite mostly thelytokous reproduction.

Materials and Methods

Sampling. Eighteen dog rose (*Rosa canina*) bud galls with *D. rosae* were collected from April 2020 to December 2020 in France for sequencing (**supplementary table S7**). After sampling, the galls were kept in plastic bags at room temperature. The insect material was removed by dissection of the gall tissue for further DNA extraction. A one-tube sample contained emerged adults or larvae from the same gall. Further DNA extractions were performed from larvae or adult females.

DNA extraction for Illumina sequencing. Prior to DNA extraction, the insects were frozen at −80°C overnight. The initial mass of the material provided for DNA extraction varied from 13.7 to 56.0 mg. The insect material was homogenised using a TissueLyser (Qiagen, Haan, Germany) with a metallic bead. Insect DNA was extracted using a DNeasy Blood and Tissue Kit (Qiagen, Germany). After extraction, the purity of the samples was evaluated by estimating the 260/280 and 260/230 ratios measured using a Thermo Scientific NanoDrop 2000 Spectrophotometer. The 260/280 ratio varied from 1.69 to 2.08; the 260/230 ratio varied from 0.79 to 2.08. The DNA quantity was estimated using a Qubit dsDNA BR Assay Kit (Invitrogen) and the Qubit Fluorometer and varied from 0.98 to 6.9 µg. Illumina sequencing was performed by Genotul, Toulouse, France (https://get.genotoul.fr). Illumina *D. rosae* reads are available at the NCBI (https://www.ncbi.nlm.nih.gov/) platform: SRA accessions are SRS16596603−SRS16596619, and BioSample accessions are SAMN32903234−SAMN32903250.

DNA extraction for Nanopore sequencing. Prior to DNA extraction, one *D. rosae* larva (sample ESE-709, **supplementary table S7**) was ground by hand with a sterile pestle. The DNA was extracted using the NucleoBond Buffer Set IV kit (Macherey-Nagel, Germany) and NucleoBond AXG 20 columns (Macherey-Nagel, Germany). Long-read sequencing was performed using a Flow Cell Wall Wash Kit (EXP-WSH004) (Oxford Nanopore Technologies, UK) and an Oxford Nanopore MinION Flow Cell R10 (Oxford Nanopore Technologies, UK).

Genome assembly. Oxford Nanopore Technology (ONT) raw reads were base called using Guppy basecaller v. 6.1.2+e0556ff (Oxford Nanopore Technologies, UK) with accurate mode and dna_r9.4.1_450bps_sup.cfg config. The resulting *.fastq* reads were assembled with a Flye assembler (Kolmogorov et al. 2019) using a --nano-hq flag, an estimated coverage of 18x, and an error rate of 0.1. The input genome size was 580 Mbp, based on a k-mer estimate using GenomeScope (Vurture et al. 2017). Because of the high error rate of ONT reads, the resulting assembly was polished using NextPolish (Hu et al. 2020) with high-coverage Illumina reads (119x) using a sample from the same population (sample ESE-219, **supplementary table S7**). Repeats were discovered in the

de novo assembly using RepeatModeler v. 2.0.3 (Flynn et al. 2020) using RMBlast v. 2.11.0. *De novo*-detected repeats were assessed for potential wrong assignment to repeats of highly duplicated genes by blasting each consensus repeat sequence onto the NCBI nr database. Repeats matching known protein-coding genes not related to known repeats were removed, and filtered repeats were then used to build the repeat database. *De novo* assembly was then masked using the newly built repeat database. The frequency of repeats was calculated in sliding windows and presented as the ratio between the repeat length and the length of the given genome region. The assembly quality was evaluated by computing different metrics using QUAST v. 5.1.0rc1 (Mikheenko et al. 2018). Protein-coding genes were predicted using BRAKER2 with ab-initio mode (Hoff et al. 2019) using protein homology from OrthoDB Metazoa, Fungi, Plants, and Bacteria for the ProtHint (Bruna et al. 2020) step. The completeness of annotated genes in terms of their expected gene content was evaluated using BUSCO v. 5.3.2 (Manni et al. 2021). Predicted genes were functionally annotated using the eggNOG-mapper v. 2.1.7 web server (Huerta-Cepas et al. 2019; Cantalapiedra et al. 2021).

SNP calling. All reads of *D. rosae* were first aligned to the reference using Bowtie 2 v. 2.3.5.1 (Langmead and Salzberg 2012). The alignments were processed for further manipulation using samtools v. 1.7 (view, sort, index, depth, and stat options) (Danecek et al. 2021). The quality of each alignment was estimated by calculating the average percentage of the mapped reads (57.3–98.1%), the average quality (35.4–35.7), the average coverage (25–119×), and the error rate (0.0078–0.015). Subsequently, the identification of polymorphisms across the genome was performed using a pipeline described in GenomeAnalysisToolkit (GATK) v. 4.0 (AddOrReplaceReadGroups, HaplotypeCaller, CombineGVCFs, and GenotypeGVCFs tools) (https://gatk.broadinstitute.org/hc/en-us/articles/360035890411-Calling-variants-on-cohorts-of-samples-using-the-HaplotypeCaller-in-GVCF-mode). Indels and non-biallelic sites were removed from the *.vcf* file using bcftools v. 1.7 (filter command) (Danecek et al. 2021). The rare (--maf 0.05), low-quality (--minQ 30), low-depth (cutoff below 2.5th percentile), and high-depth (cutoff above 97.5th percentile) sites were removed using vcftools 0.1.17 (Danecek et al. 2021). The *.vcf* file was phased by beagle v. 5.4 (Browning et al. 2021). The total number of examined polymorphisms was 2,903,839.

Population structure. The population structure of *D. rosae* was inferred according to the fastStructure algorithm (Raj et al. 2014) based on the calculation of the allele frequency spectra from SNP data. Prior to fastStructure, indels and site linkage disequilibrium ($r^2 > 0.2$) were filtered out using bcftools v. 1.7 (filer and prune commands) (Danecek et al. 2021) from the initial *.vcf* file that was then converted to a *.bed* file by

plink v. 1.9 (Purcell et al. 2007). After filtering, the total number of examined SNPs was 12,309. The population structure was assessed with the number of populations, K, varying from 1 to 4. To show individuals admixed between the lineages, F3 statistics representing the covariance of allele frequency differences between populations were calculated using ADMIXTOOLS 2.0.0 (Patterson et al. 2012; Peter 2016).

DNA extraction for genotyping. Prior to DNA extraction, the insects were frozen at −20°C overnight. The DNA was extracted using either a DNeasy Blood and Tissue Kit (Qiagen, Germany) or Chelex 100 Resin (Bio-Rad).

Genotyping. A genetic marker helping to distinguish the lineages of *D. rosae* was searched along the genome by choosing a 15 kbp window showing the Fst value closest to 1 and substantial polymorphism. A 706-bp sequence containing 9 SNPs was selected, and the primers were designed using Primer3web 4.1.0 (Untergasser et al. 2012). The marker was then amplified using polymerase chain reaction (PCR) (**supplementary protocol S1**). The presence of the PCR product was verified by performing electrophoresis in 3.5% agarose gel and sequencing using Eurofins Genomics. The sequences were assessed visually on the trace file using SnapGene Viewer 6.0.2 ("SnapGene software" www.snapgene.com) and quality trimmed. Subsequently, the sequences were used for cladogram construction (maximum likelihood statistical method, Tamura-Nei substitution model (default)) by MEGA 11 (Tamura et al. 2021). A total of 123 *D. rosae* individuals from 61 locations were genotyped for the marker. Samples were collected from different habitats in France, from different host plant individuals from the same habitat, and from different galls from the same host plant.

Demographic scenarios. Scenarios describing the demographic history of *D. rosae* were examined using *dadi* software based on the diffusion (continuous approximation) approach (Gutenkunst et al. 2009). The following standard two-population models were examined: bottleneck followed by exponential growth, then split without (i) and with migration (ii), split into two populations of specified size (iii), isolation with exponential population growth (iv), isolation with exponential population growth, and a size change prior to splitting (v) (**supplementary code S1**). A set of joint allele frequency spectra (AFS) was generated to compare the model with the data. Each model was examined by varying population sizes, time of split/isolation, and migration parameter (0, symmetric, or asymmetric). The models were ranked by calculating the Akaike Information Criterion (AIC). A likelihood ratio test was performed to compare the best models according to AIC.

Estimation of model parameters. To confirm the demographic scenario found by *dadi* and estimate parameters describing the demography of *D. rosae* (**supplementary code S2**), the approximate Bayesian computation (ABC) was applied. One million simulations

of the population parameters (effective population size Ne, mutation rate μ, bottleneck time, split time) were performed, and the resulting summary statistics were calculated by the msprime simulator v. 1.1.1 (Baumdicker et al. 2022). Subsequently, the goodness-of-fit and validation of the model were performed using the *abc* R package v. 2.2.1 (Csilléry et al. 2012).

Population genomic statistics. Absolute divergence D_{xy} (Nei 1987) and the fixation index Fst (Weir and Cockerham 1984) were calculated in the genome 10-kbp windows using pixy v. 1.2.6.beta1 (Korunes and Samuk 2021). Tajima's D and nucleotide diversity π (in 10-kbp windows) and per-individual heterozygosity H were calculated using vcftools 0.1.17 (Danecek et al. 2021). The Watterson estimator θ*w* was calculated as the number of segregating sites (provided in the output table using vcftools 0.1.17 when calculating π) in the genome 10-kbp windows divided by the sum of the (n−1) first harmonic means, where n is the number of haplotypes (Watterson 1975). Nucleotide diversity at nonsynonymous π_N and synonymous π_S sites was calculated in protein coding sequences (CDS) using the PolydNdS programme (Thornton 2003). Each CDS was extracted from the genome using the vcf2fasta.py script (https://github.com/santiagosnchez/vcf2fasta). The effect of background selection on the genome was assessed by measuring the correlation between gene density (the proportion of nucleotides assigned to a protein-coding sequence) and π within each scaffold. The population-scaled recombination rate ρ was estimated by ReLERNN (Adrion et al. 2020) using the ReLERNN_SIMULATE - ReLERNN_TRAIN - ReLERNN_PREDICT - ReLERNN_BSCORRECT pipeline. In ReLERNN_TRAIN, --nEpochs (time) was estimated at 500, corresponding to a minimal convergence of loss (mean squared error) between the training set and the validation set.

Runs of homozygosity. ROHs were detected across the *D. rosae* genome using bcftools v. 1.7 (roh command) with the option -G, the phred-scaled genotype likelihoods, set to 30 (Danecek et al. 2021). The frequency of 0.01–0.1, 0.1–0.5, 0.5–1.0, and >1 Mbp runs was estimated as the total number of ROHs from each category divided by the total number of ROHs.

Detection of genome regions under selection. To detect regions under putative selection in the *D. rosae* genome, we calculated Fst and π in 10-kbp windows using pixy v. 1.2.6.beta1 (Korunes and Samuk 2021). To distinguish between genomic regions with low differentiation/high nucleotide diversity and low differentiation/low diversity, we created a composite score that summarised Fst and π. The score was equal to (1 − Fst)*2(F(π) − 0.5), where F(π) is the cumulative distribution function in each lineage. Composite score outliers below −0.5 reflected genomic regions with low differentiation and low genetic diversity, and composite score outliers above 0.5 indicated regions with

low differentiation and high diversity (**fig. S14, fig. S15**). Subsequently, annotated gene sets (eggNOG-mapper) found in these regions were used in the Gene set enrichment analysis (GSEA). GSEA was performed using the *topGO* R package v. 2.48.0 by applying the Fisher exact test (Alexa and Rahnenfuhrer 2022). The obtained raw p-values were adjusted using the *p.adjust* function (R Core Team 2022). In total, 8721 genes were used. The examined number of genes showing a composite score below −0.5 was 75 for lineage 1 and 333 for lineage 2, respectively. The examined number of genes showing a composite score above 0.5 was 177 and 80 for lineage 1 and lineage 2, respectively.

Identification of *Wolbachia*. The assembly of the *D. rosae* genome using Nanopore data shows only one completely assembled genome of *Wolbachia* belonging to supergroup B (Wang et al. 2020). To distinguish between *Wolbachia* genomes from supergroup B and supergroup A, contaminant reads were removed from the *D. rosae* reference (Bowtie 2 --un-conc-gz option) and assembled using the MEGAHIT metagenome assembler v. 1.2.9 (Li et al. 2015). The obtained contigs were used to reconstruct genomes with MetaBAT v. 2.12.1 (Kang et al. 2015). After binning, *Wolbachia* supergroups were identified according to the multilocus sequence typing (MLST) system based on five genes (*coxA*, *gatB*, *hcpA*, *ftsZ*, and *fbpA*) (Wang et al. 2020). The coverage of *Wolbachia* contigs was given by MetaBAT and normalised by the coverage of the corresponding *D. rosae* individual.

Statistics. All statistical analyses were performed using R v 4.2.2 (R Core Team 2022). The significance level was set to 0.05. The figures were produced using R v. 4.2.2 and Microsoft Excel 2010.

Acknowledgements

We would like to thank David Ogereau (CNRS, EGCE) for providing DNA for Nanopore sequencing. We also thank Jacqui Shykoff (Paris-Saclay University, ESE), Olivier Plantard (INRAE), Amir Yassin (CNRS, EGCE), Florence Mougel (CNRS, EGCE), and Thierry Robert (Paris-Saclay University, ESE). The study received a grant from the National Agency for Research (France) (Project ANR-19-CE02-0008 "Tracing back the history of an adaptive trait: genetic basis of plant host manipulation by gall wasps - BETAGALL").

Author Contributions

KM performed DNA extraction (genotyping), simulations, and data analysis, and wrote the manuscript. AB conceived and designed the study, wrote the manuscript, performed genome assembly, described the *D. rosae* genome, and performed GSEA. AB, KM, and ZT collected gall samples. ZT kept the gall samples and performed DNA extraction (Illumina). All authors discussed the results and contributed to the final manuscript.

Data Availability

The assembled genome of *D. rosae* is available at the NCBI platform (https://www.ncbi.nlm.nih.gov/): a DDBJ/ENA/GenBank accession is JAPYXD000000000 (version JAPYXD000000000.1), a BioProject accession is PRJNA914909, and a BioSample identifier is SAMN32363506. Illumina *D. rosae* reads are available at the NCBI platform: SRA accessions are SRS16596603–SRS16596619, and BioSample accessions are SAMN32903234–SAMN32903250.

Conflict of Interest

The authors declare no conflict of interest.

Tables

Table 1. Composite score outlier values detected in the *Diplolepis rosae* genome regions with low differentiation.

Composite score (CS) outliers in *D. rosae* lineages	Scaffold number	Region, Mbp	Other signal(s)
Negative (CS < −0.5) lineage 1	204	30.0	Decrease in ρ
	313 (**fig. 5**)	4.1	Decrease in ρ
	414 (**fig. 6**)	3.0	Decrease in ρ
Negative (CS < −0.5) lineage 2	204	30.0	Decrease in ρ
	204	30.1–33.0	Decrease in ρ
	313	0.0–5.9	Decrease in ρ
	414	3.0	Decrease in ρ
	523 (**fig. 7**)	0.0–2.0	–
Positive (CS > 0.5) lineage 1	204	26.0	Increase in ρ
	204	28.0	Increase in ρ
	204	30.1–33.0	Decrease in ρ
	313	0.0–4.0	–
	313	4.2–5.9	–
	313	6.1–6.2	–
	325 (**fig. 6**)	2.5–3.0	Increase in π, decrease in ρ, and increase in πN/πS
	414	5.5–9.0	Increase in ρ
	414	5.8–6.0	Increase in πN/πS
	523	0.0–2.0	–
	762 (**fig. 7**)	12.5–17.5	Increase in π and decrease in ρ
Positive (CS > 0.5) lineage 2	204	26.0	–
	204	28.0	–
	313	6.1–6.2	Increase in π
	325	2.5–3.0	Increase in π and decrease in ρ
	414	5.5–9.0	–
	762	12.5–17.5	Increase in π and decrease in ρ

Figures

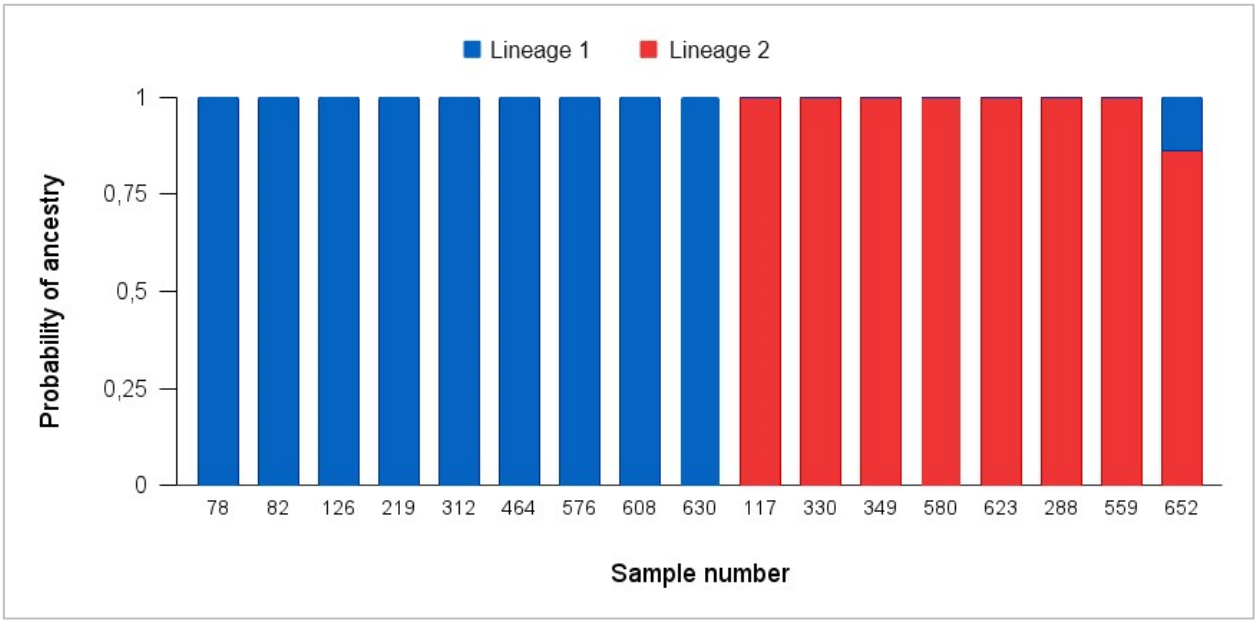


Fig. 1. Probability of ancestry of each *Diplolepis rosae* sample using ancestry proportions inferred by the fastStructure algorithm for K = 2.

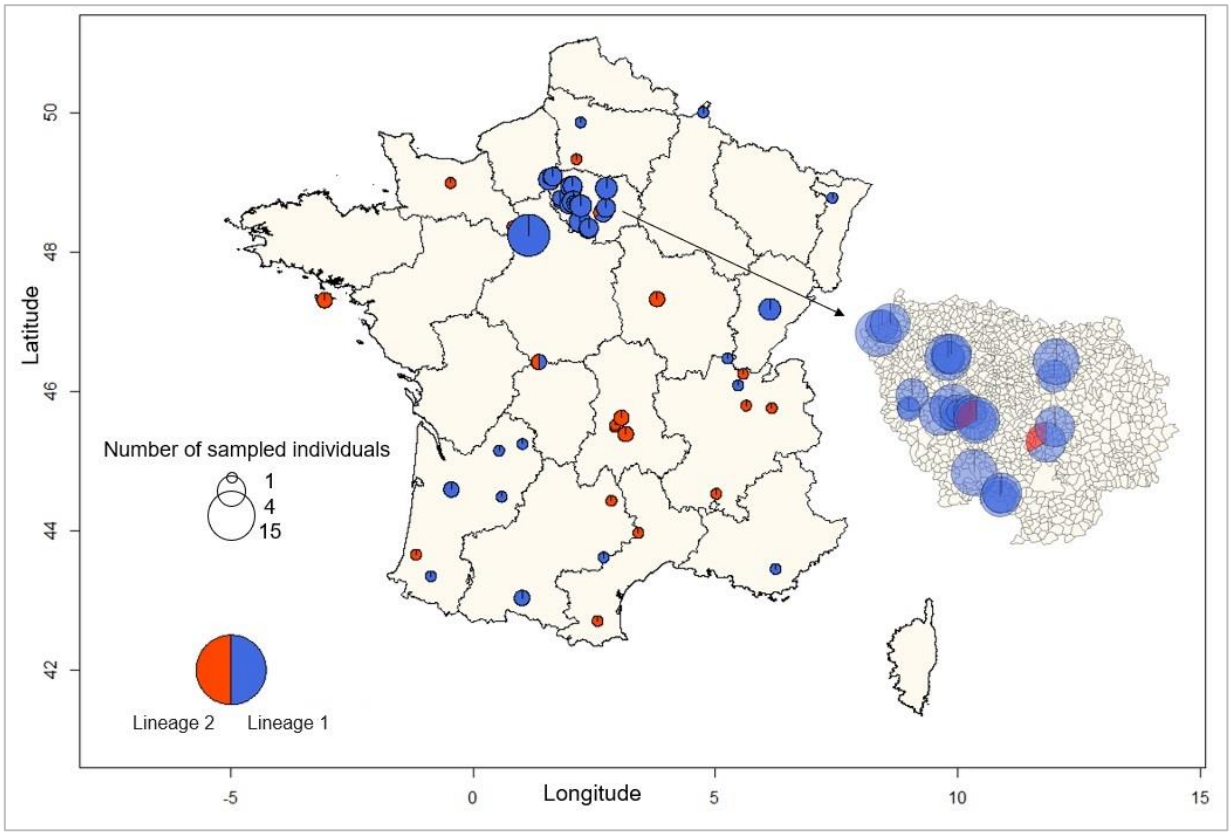


Fig. 2. Geographical distribution of *Diplolepis rosae* lineages in France according to one population-specific genetic marker (see supplementary protocol S1).

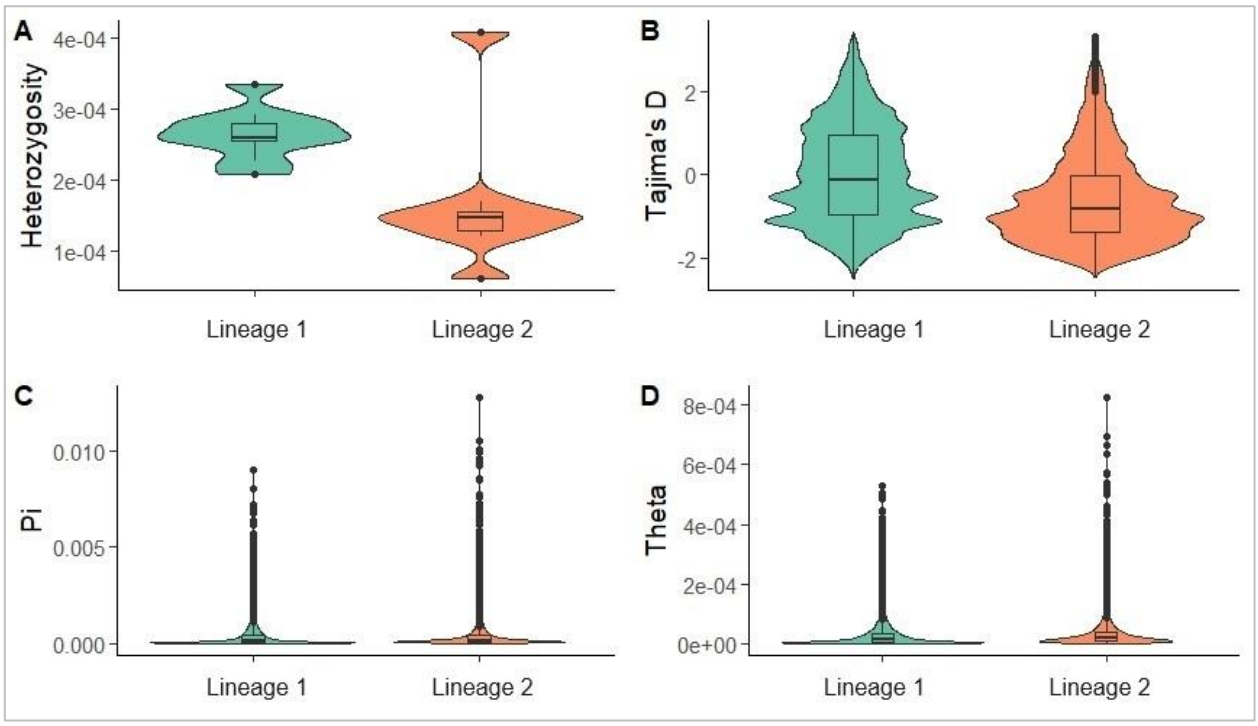


Fig. 3. Population genomic statistics for *Diplolepis rosae*. A: Per-individual heterozygosity (proportion of heterozygous sites in the total assembly length); in lineage 2, the highest value corresponds to the admixed individual *D. rosae-652* (**fig. 1**). **B**: Tajima's D. **C**: Nucleotide diversity, Pi. **D**: Watterson estimator, Theta.

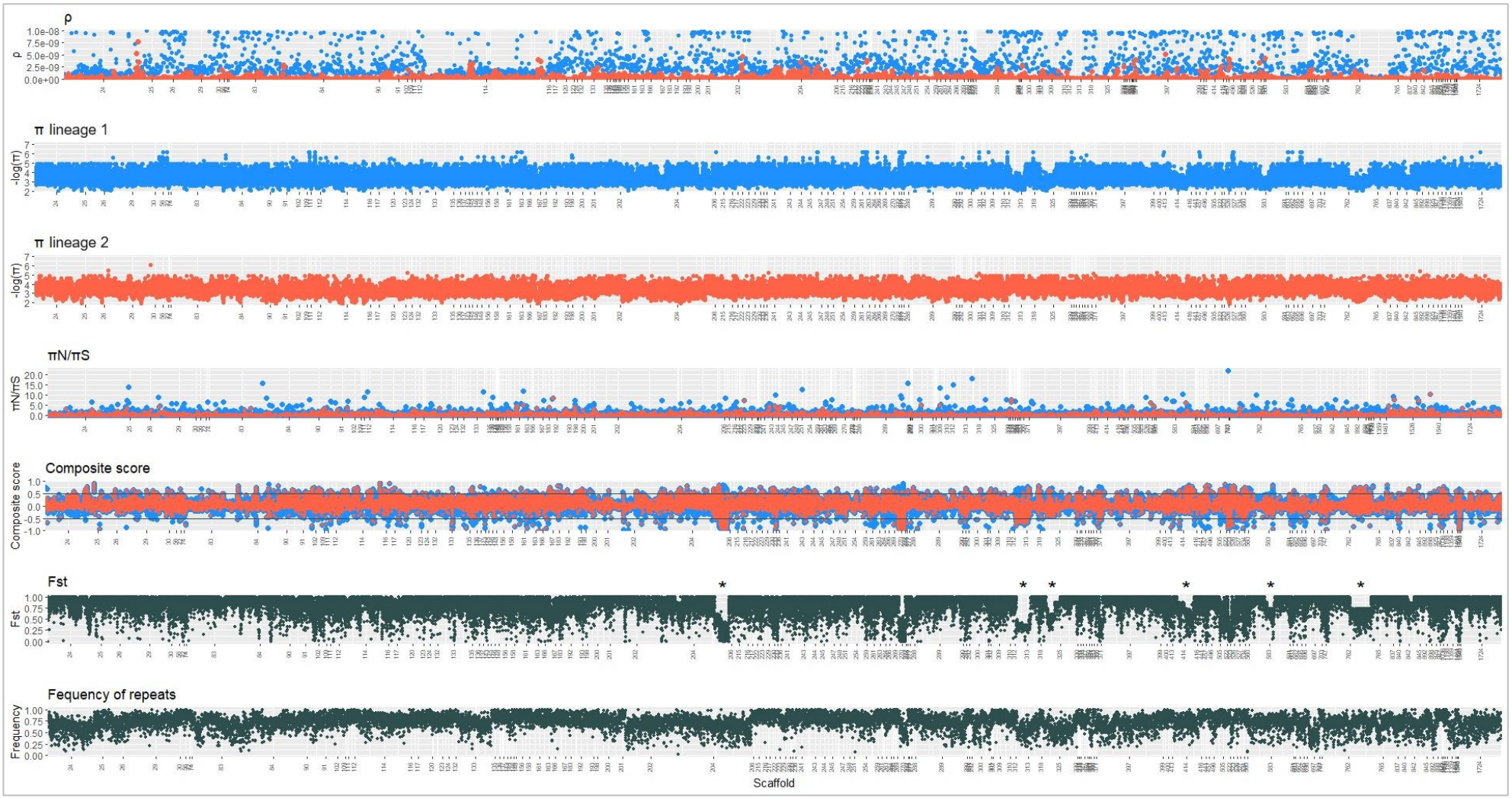

ρ
1.0e-08
7.5e-09
5.0e-09
2.5e-09
0.0e+00
ρ
π lineage 1
-log(π)
π lineage 2
-log(π)
πN/πS
πN/πS
Composite score
Composite score
Fst
Fst
Fequency of repeats
Frequency
Scaffold

Fig. 4. Genome-wide scan of population scaled per-bp recombination rate ρ, nucleotide diversity π, the ratio of nonsynonymous and synonymous mutations π_N/π_S, composite score, fixation index Fst, and frequency of repetitive sequences for *Diplolepis rosae*. Recombination rate ρ is presented in units of 4*Neff generations, where Neff is the effective population size (haploid individuals). The frequency of repetitive sequences in non-recombining regions does not significantly differ from that in regions with detected recombination in both *D. rosae* lineages (Mann–Whitney U test: U = 294, z = 1.195, p = 0.232). The composite score is equal to (1 − Fst)*2(F(π) − 0.5), where F(π) is the cumulative distribution function. Dark grey lines show thresholds corresponding to the outlier values of −0.5 and 0.5. The x-axis represents the scaffold numbers. In the Fst panel, the asterisks indicate scaffolds 204, 313, 325, 414, 523, and 762, showing an extended decrease in Fst and Runs of Homozygosity (see in detail **fig. 5, fig. 6,** and **fig. 7**). Blue points: lineage 1. Red points: lineage 2.

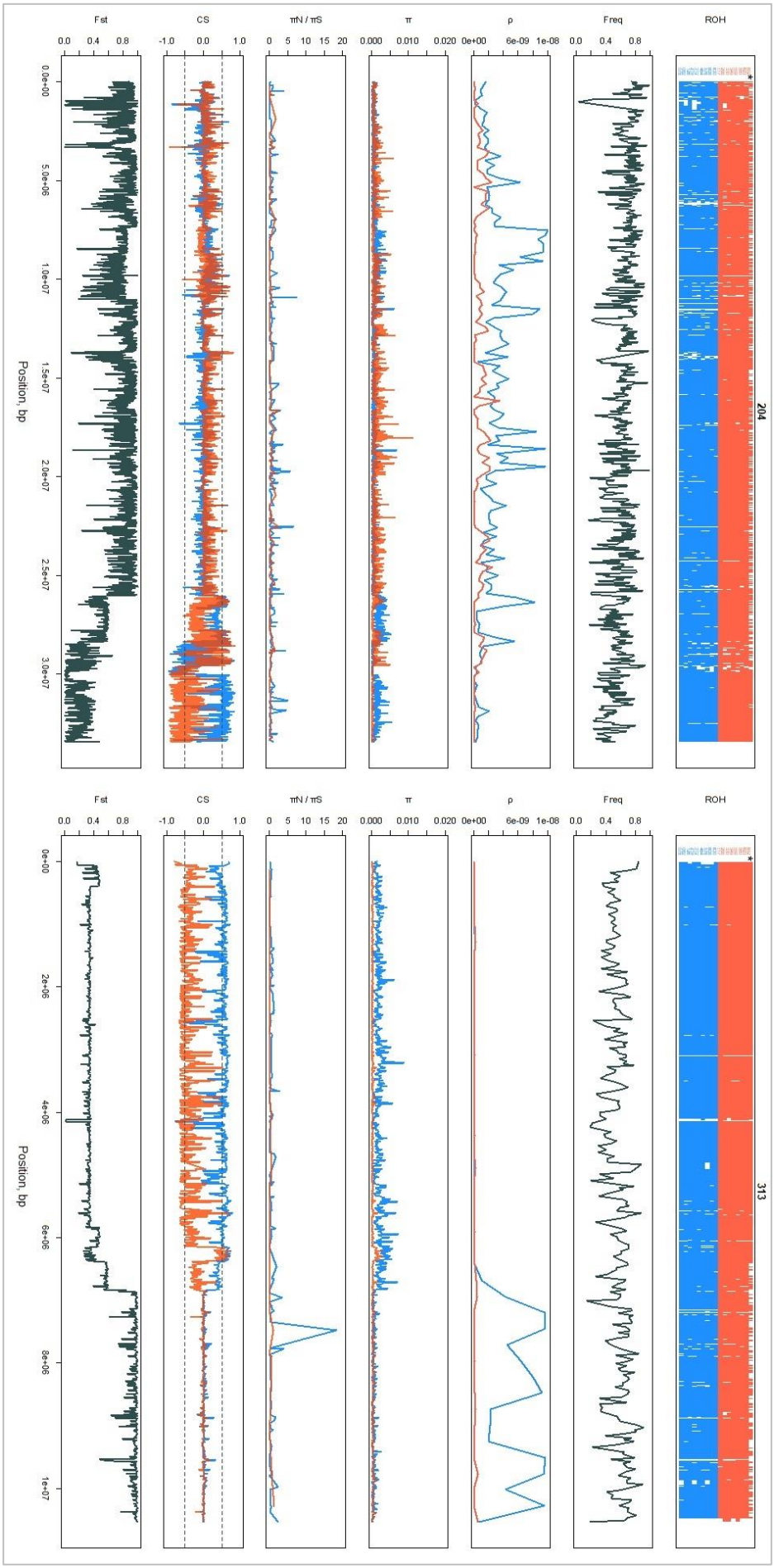

Fst
CS
πN / πS
π
ρ
Freq
ROH
Position, bp
204
313

Fig. 5. Runs of homozygosity (ROHs), frequency (Freq) of repetitive sequences, scan of population scaled per-bp recombination rate ρ, nucleotide diversity π, the ratio of nonsynonymous and synonymous mutations π_N/π_S, composite score (CS), and the fixation index Fst for *Diplolepis rosae* lineage 1 (blue) and lineage 2 (red) in scaffolds 204 and 313. In the ROH panel, the numbers (y-axis) indicate *D. rosae* individual names, and the asterisk (*) denotes the admixed individual *D. rosae*-652 (**fig. 1**). Recombination rate ρ is presented in units of 4*Neff generations, where Neff is the effective population size (haploid individuals). The composite score is equal to (1 − Fst)*2(F(π) − 0.5), where F(π) is the cumulative distribution function. In the CS panel, the dark grey dashed lines represent thresholds corresponding to the outlier values of −0.5 and 0.5. In scaffold 204 (26.0–33.0 Mbp), ROHs overlap with the decrease in Fst and the simultaneous change in π in both lineages, resulting in outlier CS. In the 26.0 Mbp and 28.0 Mbp regions, positive CS is associated with an increase in ρ. In the 30.0 Mb region, both *D. rosae* lineages showed negative CS. In the 30.1–33.0 Mbp region, lineage 1 and lineage 2 demonstrate opposite CS and a decrease in ρ. In scaffold 313 (0.0–6.5 Mbp), ROHs overlap with the zero recombination rate ρ and the decrease in Fst. In the 4.1 Mbp region, negative CS is associated with a decrease in π in both lineages. In the 6.1–6.2 Mbp region, both lineages show positive CS.

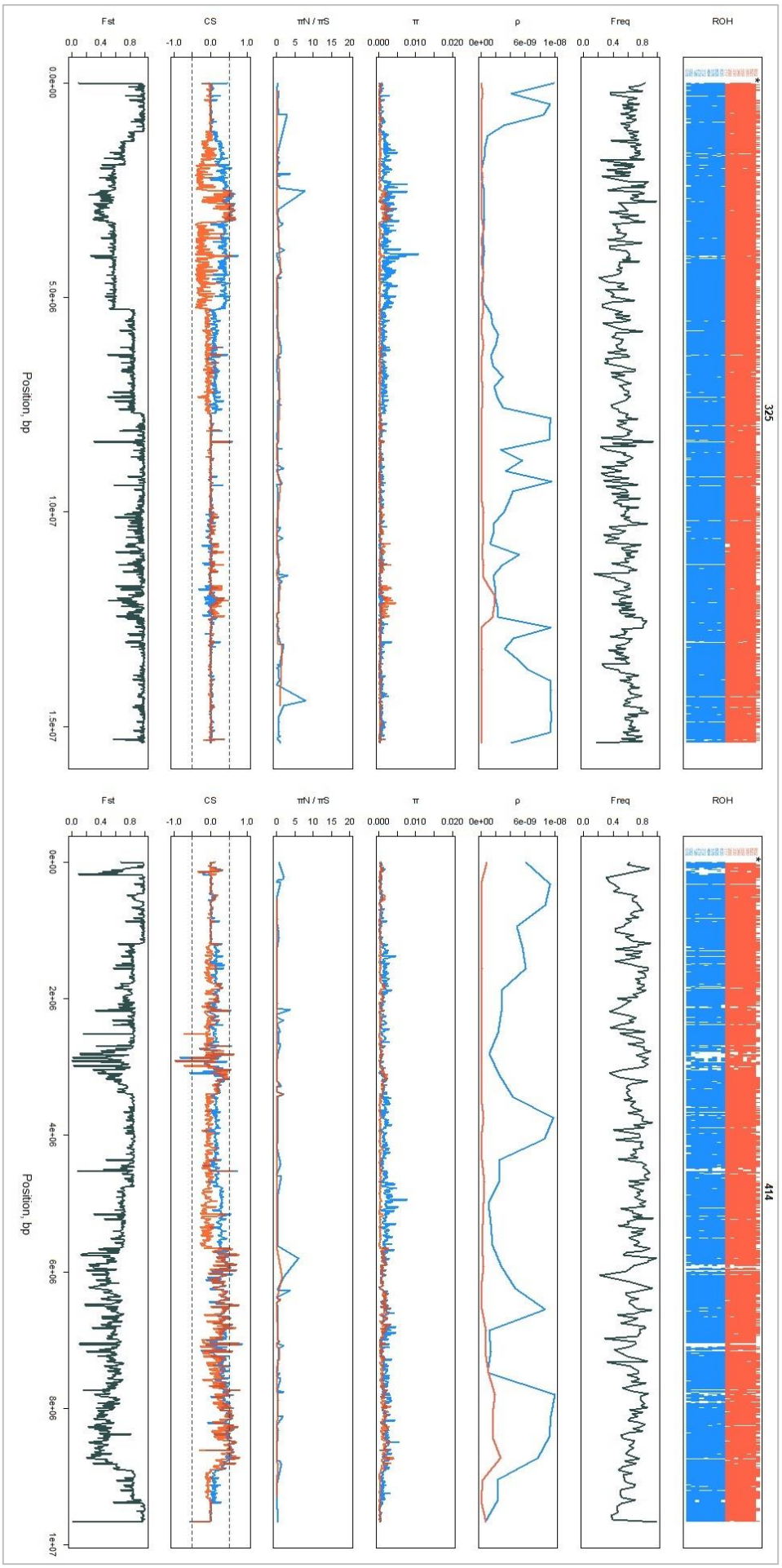

Fst
CS
πN / πS
π
ρ
Freq
ROH
Position, bp
325
414

Fig. 6. Runs of homozygosity (ROHs), frequency (Freq) of repetitive sequences, scan of population scaled per-bp recombination rate ρ, nucleotide diversity π, the ratio of nonsynonymous and synonymous mutations π_N/π_S, composite score (CS), and the fixation index Fst for *Diplolepis rosae* lineage 1 (blue) and lineage 2 (red) in scaffolds 325 and 414. In the ROH panel, the numbers (y-axis) indicate *D. rosae* individual names, and the asterisk (*) denotes the admixed individual *D. rosae*-652 (**fig. 1**). Recombination rate ρ is presented in units of 4*Neff generations, where Neff is the effective population size (haploid individuals). The composite score is equal to (1 − Fst)*2(F(π) − 0.5), where F(π) is the cumulative distribution function. In the CS panel, the dark grey dashed lines represent thresholds corresponding to the outlier values of −0.5 and 0.5. In scaffold 325, the 2.5–3.0 Mbp region shows zero recombination rate ρ, the increase in π, and the decrease in Fst, resulting in positive outlier CS in both *D. rosae* lineages. In scaffold 414, negative CS in 3.0 Mbp is associated with a decrease in ρ in both lineages. In the 5.5–9.0 Mbp region, the increase in CS is associated with the increase in ρ.

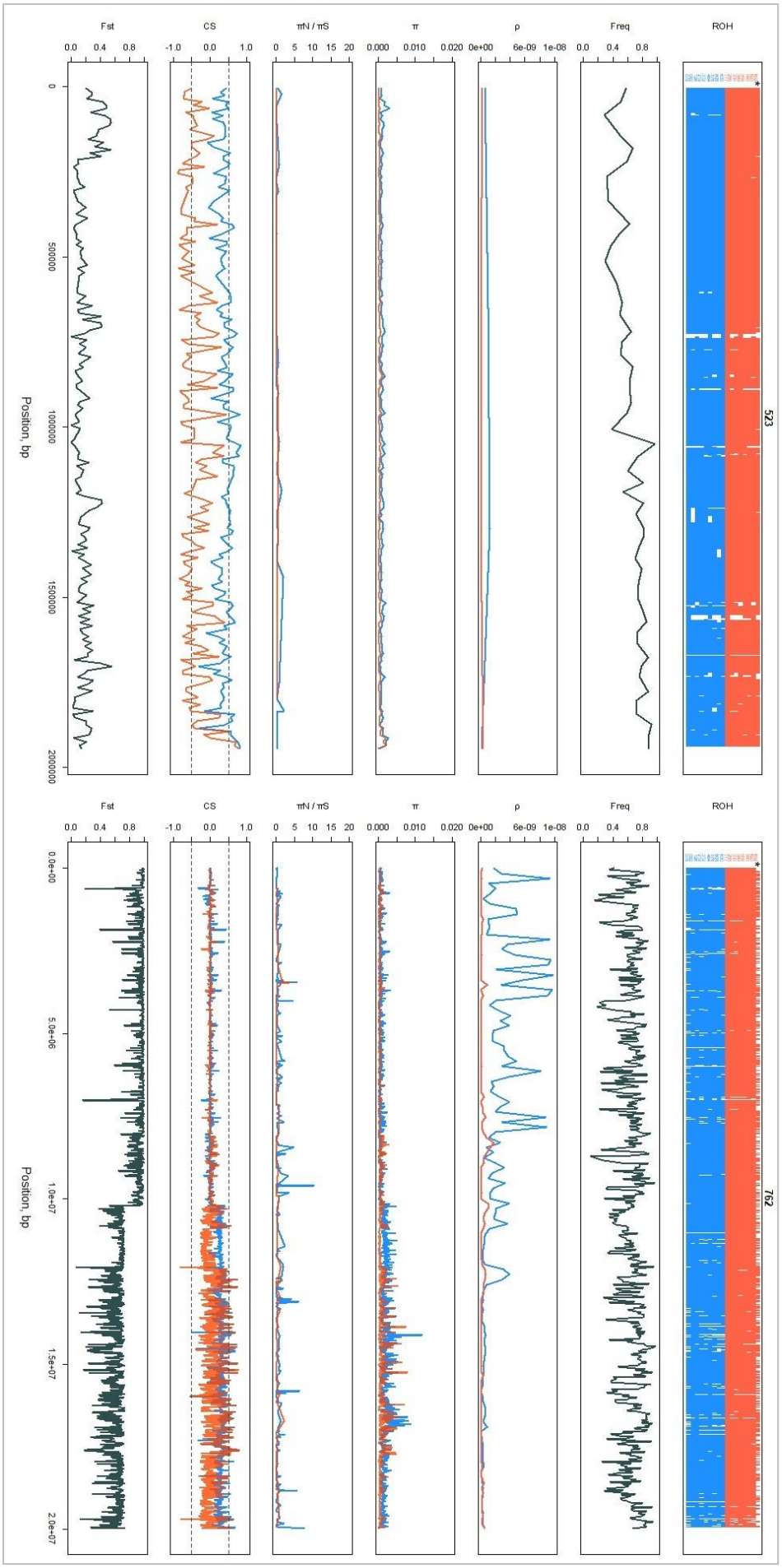

Fst
CS
πN / πS
π
ρ
Freq
ROH
Position, bp
523
762

Fig. 7. Runs of homozygosity (ROHs), frequency (Freq) of repetitive sequences, scan of population scaled per-bp recombination rate ρ, nucleotide diversity π, the ratio of nonsynonymous and synonymous mutations π_N/π_S, composite score (CS), and the fixation index Fst for *Diplolepis rosae* lineage 1 (blue) and lineage 2 (red) in scaffolds 523 and 762. In the ROH panel, the numbers (y-axis) indicate *D. rosae* individual names, and the asterisk (*) denotes the admixed individual *D. rosae*-652 (**fig. 1**). Recombination rate ρ is presented in units of 4*Neff generations, where Neff is the effective population size (haploid individuals). The composite score equals (1 − Fst)*2(F(π) − 0.5), where F(π) is the cumulative distribution function. In the CS panel, the dark grey dashed lines represent thresholds corresponding to the outlier values of −0.5 and 0.5. In scaffold 523, lineage 1 and lineage 2 demonstrate the opposite CS values associated with the zero recombination rate. In scaffold 762, the 12.5–20.0 Mbp region shows positive CS associated with a decrease in ρ in both *D. rosae* lineages.

CHAPTER II. Transcriptome analysis of *Diplolepis rosae*: revealing overexpression of genes potentially associated with insect immune response and gall formation at early larval stages

The results presented in this Chapter have been submitted to ***Insect Molecular Biology***.

Abstract

Insect parasites can provoke drastic changes in host plant physiology by affecting cell differentiation pathways and various metabolic processes. An intriguing example of such interaction is a gall, a novel outgrowing plant organ, induced by another organism for its own benefit. Cynipidae is a family of gall-inducing hymenopterans that induce galls with a complex anatomical structure. Gall formation involves three stages: initiation, growth, and maturation. Until today, the mechanism of gall initiation remains unknown. In this study, we aimed to reveal candidate genes involved in gall induction in *Diplolepis rosae*, a gall wasp inducing bedeguars in wild roses. We performed differential expression analysis of the gall wasp larva transcriptome. We observed that genes encoding plant cell wall degrading enzymes are upregulated during gall developoment. These enzymes may contribute to the formation of a chamber for a developing larva by lysing plant tissues. We also demonstrated the overexpression of genes encoding podocan, vasorin-like protein, toll-like receptor 7, tetraspanin, lipase, peroxidase, phospholipase A2, and venom acid phosphatase. These genes may be involved in insect development and the immune response against parasitoids, host plant microbiome, and host plant defense systems. Additionally, we performed a test for selection to detect *D. rosae* genes under positive selection. However, we detected only one gene encoding a transposable element. The mostly asexual reproductive mode of this species leads probably to a decrease power to detect signatures of positive selection in the genome. Our study contributes to understanding the processes occurring in cynipid wasps during gall formation and creates opportunities for further investigations of other candidate genes.

Key words: gall formation, transcriptomics, *Diplolepis rosae*, insect immune response

Introduction

In host-plant - parasite interactions, parasites can manipulate their host plants through cell reprogramming, metabolic alterations, and immune system suppression. A fascinating example of such an interaction is the formation of galls, abnormal plant outgrowths induced by foreign organisms for their own benefit (Gätjens-Boniche 2019). Gall-inducing organisms use galls as a nutritional resource and protection against unfavorable biotic and abiotic conditions (Stone and Schönrogge 2003).

Gall formation can be caused by both micro- and macroorganisms from different taxa (Gätjens-Boniche 2019). However, the fine molecular mechanism underlying gall formation is well understood only for a handful of microorganisms like *Sinorhizobium* spp. inducing root nodules in Fabaceae, *Agrobacterium tumefaciens* provoking crown galls, and *Ustilago maydis*, the smut fungus of maize (Le Fevre et al. 2015; Hearn et al. 2019). In these microorganisms, the mechanism of gall formation includes (1) detection of released host plant molecules by the infectious agent, (2) release of gall inducer effector molecules interacting with specific plant receptors, and (3) further induction of plant growth response resulting in the formation of gall tissue (Le Fevre et al. 2015).

Unlike microorganisms, gall formation by animals is more complex and less understood, except for a few economically important invertebrates such as root-knot and cyst nematodes (Mejias et al. 2019), several aphids (Nabity et al. 2013; Korgaonkar et al. 2021), and the Hessian fly (*Mayetiola destructor*) (Stuart et al. 2012). Invertebrate gallers are shown to induce changes in sugar and nitrogen metabolism, disruption of the defense system, and cellular modifications in host plants (Giron et al. 2016). However, little is known about triggers encoded in the genomes of gall-inducing invertebrates, such as insects, that could be responsible for gall initiation. Unlike gall-inducing microorganisms, whose activity leads to the formation of unstructured tumor tissue in plants (Gätjens-Boniche 2019), and herbivore animals that simply damage plant tissues, insect gall inducers employ mechanisms that cause significant alterations in host plant metabolism and reprogram cell differentiation, ultimately resulting in the development of a structured new organ. It is fascinating how such a complex phenotype as the insect gall can be initiated by molecular triggers originating from the insect beyond its body. The genetic basis, i.e. the genes encoding the molecular triggers that provoke gall formation in insect gallers, remains a relevant question.

Today, multiple hypotheses, that do not mutually exclude one another, propose potential molecular triggers for gall formation. Firstly, gall initiation is believed to occur at the moment of oviposition, suggesting that components detected in female venom glands and egg secretions could serve as potential triggers (Cambier et al., 2019; Gobbo et al., 2020). Secondly, salivary gland secretions of developing larvae may contribute to gall formation (Zhao et al. 2015; Hearn et al., 2019; Korgaonkar et al. 2021). Thirdly, as

gallers manipulate plant cell differentiation and metabolic pathways, they are suggested to produce various molecules that mimic phytohormones (Yamaguchi et al., 2012). Lastly, we can also consider the role of mutualistic or pathogenic microorganisms or microbial genes acquired through horizontal gene transfer, which may also play a part in gall initiation (Bartlett and Connor, 2014; Hearn et al. 2019).

The model that has been the most study in relation to genes potentially involved in galling is *M. destructor* (Diptera: Cecidomyiidae), one of the most destructive crop pests. More than 7% of the Hessian fly genome has been estimated to encode putative effector proteins. This group of proteins includes secreted salivary gland protein (SSGP)-71. This protein contains leucine-rich repeats (LRR) that mediate protein-protein interactions (Ho et al. 2006). In plants, proteins containing LRRs play a role in plant development and immunity (Zhao et al. 2015). Interestingly, the whole structure of (SSGP)-71 resembles ubiquitin E3 ligases in plants and E3-ligase-mimicking effectors in plant pathogenic bacteria. SSGP-71 protein has been shown to interact with wheat Skp signal protein *in vivo*. Mutations in the gene encoding SSGP-71 have been shown to avoid the effector-triggered immunity in the host plant. According to these results, the authors supposed this protein to be a potential trigger of gall formation (Zhao et al. 2015).

In aphids, the salivary glands of *Hormaphis cornu* making galls on the witch-hazel (*Hamamelis virginiana*) produce a specific determinant of gall color (DGC) protein. The production of this protein is associated with the regulation of anthocyanin synthesis in the host plant (Korgaonkar et al. 2021). Anthocyanins are pigments responsible for the formation of red, purple, and blue colors in plants. In the study of Korgaonkar et al. (2021), the hyperproduction of red pigment in forming galls was correlated with the upregulation of seven genes coding enzymes acting in anthocyanin synthesis in the plant and simultaneous hyperproduction of the DGC protein in the insect. The authors supposed the triggering of red gall to be due to the injection of this potential effector protein from salivary glands into a plant tissue. Furthermore, the gene encoding DGC displays a high dN/dS ratio indicating positive selection. Thus, it shows a potential role of this gene in the galling in the context of the evolutionary arms race of aphids and their host plants.

The following studies were dedicated to another group of gall inducing organisms, gall wasps (Hymenoptera: Cynipidae). Cynipid wasps include at least 1400 species being the second largest group of gall-inducing insects after gall midges (Diptera: Cecidomyiidae), and occur on all continents, except the Antarctic (Ronquist et al. 2015). Each gall wasp species typically attacks a single host plant species or genus. In addition, each species causes a particular form of gall in a particular plant organ. Gall forms vary from slight tissue modification to complex multi-chamber structures (Stone and Schönrogge 2003).

In Cynipidae a study (Gobbo et al. 2020) demonstrated positive selection for the genes associated with gall formation in *Synergus itoensis* (Cynipidae: Synergini), the only

one known gall inducer species from the genus *Synergus*. The authors calculated a pairwise dN/dS ratio between *S. itoensis* and three inquiline *Synergus* species and performed gene set enrichment analysis of the genes showing a higher dN/dS. Cynipid inquiline is a form that had lost the ability to induce galls *de novo* but occupies the galls induced by a gall inducer species (Ronquist 1994). The gene set of *S. itoensis* showing signature of positive selection, unlike those of the inquiline species, was enriched in the 'ovarian follicle cell development', 'heart development', 'axonogenesis', and 'axon development' terms (Gobbo et al. 2020). The genes enriched in these terms were supposed to reflect the ability to induce galls. The authors hypothesized that the secretions coating the egg surface are known to induce plant immunity (Dobens and Raftery 2000; Hilker and Fatouros 2015). The plant immune response was supposed to accompany the initial steps of gall induction just after oviposition (Gobbo et al. 2020).

Other candidate genes acting on gall formation were found in the venom glands of *Diplolepis rosae* and *Biorhiza pallida*. Transcriptome analysis revealed the overexpression of genes encoding serine proteases, phospholipases, lipases, esterases, and peroxidases (Cambier et al. 2019). These enzymes have no evident role in regulation of plant immunity or plant development. Nonetheless, in the *B. pallida* venom the authors also detected cellulases of bacterial origin, which was supposed to contribute to the lysis of a plant cell wall. Furthermore, another transcriptome study (Hearn et al. 2019) showed the overexpression of genes encoding different plant cell wall degrading enzymes (PCWDEs) like pectate lyases, rhamnogalacturonan lyases, and cellulases in *B. pallida* larvae. PCWDEs were both encoded in the insect genome and most certainly acquired via horizontal gene transfer from bacteria.

In this study, we aimed to reveal candidate genes that could be responsible for the initial stages of gall formation, i.e. gall initiation in *Diplolepis rosae*, a holarctic gall wasp causing bedeguars in wild roses *Rosa* spp. sect. Caninae (Rosaceae). *D. rosae* is a univoltine species reproducing mostly by thelytokous parthenogenesis, where virgin females produce females (Nordlander 1973; Stille and Dävring 1980; Heimpel and De Boer 2008). Adults emerge from May to June and are synchronized with the development of suitable host plant tissues for gall induction (Shorthouse and Floate 2010). After emergence, the females immediately oviposit into epidermal plant cells located between the developing leaflets of an expanding bud (Bronner 1985). Once eggs are laid, gall tissue begins to develop. One gall can contain up to one hundred larvae (Rizzo and Massa 2006). The feeding larvae are surrounded by gall cells and spend at the pre-nymph stage (from early November) before the next spring (Shorthouse and Floate 2010).

Firstly, we hypothesized that gall wasps must generate new genetic variants to evade plant immune system and successfully manipulate host plant metabolism (Van Valen 1973). Therefore, we sought to identify evidence of positive selection in the genome of *D. rosae* by performing a test for selection as in the study of Korgaonkar et al. 2021.

Secondly, we employed transcriptomics and measured gene expression levels in various *D. rosae* tissues at different steps of gall formation. Our aim was to identify genes that are overexpressed during the early stages compared to the later stages. Subsequently, we expected to refine the list of candidate genes by comparing two sets of candidates identified by both approaches. On the one hand, we sought to exclude genes that might be under selection but not be expressed during the initial stages of gall formation. On the other hand, the set of genes under positive selection could help to exclude conserved genes overexpressed during gall initiation but likely responsible for other processes, such as insect development.

Materials and Methods

Sampling. Seventeen *D. rosae* (**supplementary table S7**) and two *D. eglanteriae* (**supplementary table S8**) galls were collected in France from September 2019 to December 2020. *D. rosae* individuals were collected from the same host plant specimen and came from one female due to a clonal mode of reproduction (Mozhaitseva et al. 2023). After sampling, the galls were kept in plastic bags at room temperature. Insect material was removed by dissection of gall tissue for further DNA extraction.

DNA extraction for Illumina sequencing. The larvae were homogenized in 2-mL plastic tubes by a TissueLyser (Qiagen) with adding a metallic bead. The initial mass of larvae varied 1.8 to 56.0 mg. DNA was extracted using a DNeasy Blood and Tissue Kit (QIAGEN, Germany) (**supplementary protocol S2**). 260/280 and 260/230 ratios varied from 1.35 to 2.08 and from 0.40 to 2.08, respectively. DNA quantity ranged from 0.42 to 6.9 µg (**supplementary protocol S3**). Before the sequencing, *D. eglanteriae* was verified using the PCR of the gene encoding cytochrome oxidase c subunit I (**supplementary protocol S4-S5**) to distinguish between the gall wasp larvae and their parasitoids. The sequence of a PCR product was verified in the Nucleotide collection (nt) database (the NCBI platform: https://blast.ncbi.nlm.nih.gov) . After that, the Illumina sequencing was performed by Genotoul, Toulouse, France (https://get.genotoul.fr). *D. rosae* Illumina reads are available in the SRA database (the NCBI platform) under the accessions SRX19185216-SRX19185235.

SNP calling (supplementary code S3). All reads of *Diplolepis spp.* were firstly aligned to the *D. rosae* reference (accession JAPYXD000000000 in DDBJ/ENA/GenBank) using Bowtie 2 v. 2.3.5.1 (Langmead and Salzberg 2012) with --end-to-end (default, *D. rosae* reads) or --local (*D. eglanteriae* reads) mode. The alignments were processed for further manipulations using Samtools 1.7 (view, sort, index, depth, and stat options) (Danecek et al. 2021). The quality of each alignment was estimated by calculating the average percentage of the mapped reads (varied between 56.6% and 98.1%), the average quality (35.4 - 35.7), the average coverage (22x - 119x), and the error rate (0.0079 - 0.015). After that, the identification of polymorphisms across the genome was performed by using a pipeline proposed by GenomeAnalysisToolkit (GATK) v. 4.0 (AddOrReplaceReadGroups, HaplotypeCaller, CombineGVCFs, and GenotypeGVCFs tools) (https://gatk.broadinstitute.org/hc/en-us/articles/360035890411-Calling-variants-on-cohorts-of-samples-using-the-HaplotypeCaller-in-GVCF-mode). Indels and non-biallelic sites were removed from the *.vcf* file using bcftools v. 1.7 (filter command) (Danecek et al. 2021). Phasing of the *.vcf* file was performed by beagle v. 5.4 (Browning et al. 2021) to show two haplotypes corresponding to each individual.

McDonald–Kreitman test (supplementary code S4). The *.fasta* alignments were extracted from the phased *.vcf* file using the vcf2fasta.py program (https://github.com/santiagosnchez/vcf2fasta). Gene coordinates were shown in a *.gff* annotation file and the reference *.fasta* file. Each alignment corresponded to concatenated coding sequences (CDS) belonging to the same transcript. For each *D. rosae* alignment, the number of polymorphic synonymous/non-synonymous and divergent synonymous/non-synonymous mutations was inferred by the MKtest program from libsequence (Thornton 2003) using *D. eglanteriae* as an outgroup. The mean number of synonymous and replacement sites per gene was calculated by the polydNdS program from libsequence (Thornton 2003). Obtained data were used to estimate the direction of selection, the mean value of selection coefficient *s* and lower and upper *s* bounds using SnIPRE, a McDonald-Kreitman type analysis based on a generalized linear mixed model (Eilertson et al. 2012).

RNA extraction. *D. rosae* galls developing on the same dog rose were sampled from May 2022 to November 2022 in Bures-sur-Yvette (48°42′12″N, 2°9′35″E). *D. eglanteriae* dog rose galls were sampled in mid-July 2022. The insect material was removed by gall dissection. Before RNA extraction, the insects were kept in a 2-ml Eppendorf tube containing 1 ml of RNAlater Solution (Invitrogen, Lithuania) at 4 °C (**supplementary protocol S6**). Insect tissue was homogenized by TissueLyser (Qiagen) using a metallic bead. RNA was extracted using a RNeasy Mini Kit (QIAGEN, Germany). After extraction, the purity of the samples was estimated using 260/280 (varied between 1.90 and 2.25) and 260/230 (varied between 0.31 and 2.25) ratios measured by a Thermo Scientific NanoDrop 2000 Spectrophotometer. RNA quantity was estimated using a Qubit RNA BR Assay Kit (Molecular Probes, USA) and a Qubit Fluorometer and varied 0.82 to 5.3 µg (**supplementary protocol S7**).

Sampling. Since we did not master the life cycle of *D. rosae* for a laboratory experiment, we collected the galls in natural conditions. We obtained the following samples:

- mid-July larva (first visible gall);
- early September larva (growing gall);
- early October larva salivary glands (mature gall);
- early November larva/pre-nymph salivary glands (wintering gall);
- head from the female adult emerging from the gall kept in the laboratory (control sample);
- egg removed from the emerged female adult (control sample);
- additionally, mid-July *D. eglanteriae* larva to exclude species-specific genes.

We expected to detect candidate genes specifically expressed:

- at each life cycle stage or tissue: for instance, to show up-regulated genes in the mid-July *D. rosae* larva, we searched for an overlap between the up-regulated genes by comparing the pairs:
    - mid-July larva - egg,
    - mid-July larva - early September larva,
    - mid-July larva - early October larva salivary glands,
    - mid-July larva - early November larva salivary glands,
    - mid-July larva - female adult head;
- during the active gall growth ('growth' vs no 'growth'): comparison between the sample set pair 'mid-July larva + early September larva + mid-July *D. eglanteriae* larva' and 'egg + early October larva salivary glands + early November larva salivary glands + female adult head',
- during the whole gall formation ('gall' vs 'no gall'): comparison between the sample set pair 'mid-July larva + early September larva + mid-July *D. eglanteriae* larva + early October larva salivary glands + early November larva salivary glands' and 'egg + female adult head'.

Relative differential gene expression data analysis (supplementary code S5). cDNA sequencing and library preparation (Illumina NovaSeq 50 M 150-bp reads, PolyA enrichment, non-stranded)) was performed by Novogene Europe (Cambridge, UK). Quality control of raw reads was performed using the fastQC program v. 0.12.1 (Andrews 2010) (**supplementary table S9**). Then, the reads were aligned to the genome reference by STAR v. 2.7.10b (Dobin et al. 2013). Genome coordinates were provided in a *.gff3* file generated by BRAKER v. 3.0.3 (Stanke et al. 2006, 2008; Hoff et al. 2016, 2019; Bruna et al. 2021) from the reference genome and .bam alignments obtained at the previous step. The count of aligned reads to annotated genes was performed by featureCounts v. 2.0.6 (Liao et al. 2014) using gene coordinates given in the *.gff3* gene prediction file. The counting was performed at gene level. The summarizing gene count matrix was then used in the relative differential expression analysis performed by DESeq2 v. 1.36.0 (Love et al. 2014). Up- and down-regulated genes specifically expressed at each stage or tissue were assessed by pairwise comparison of the following samples: egg, mid-July larva, early September larva, October salivary glands, November salivary glands, and female adult head. Data quality was assessed by calculating pairwise Euclidean distances (**supplementary fig. S16-S18**), performing the principal component analysis (**supplementary fig. S19-S21**), and plotting dispersion estimates (**supplementary fig. S22-S24**).

Gene annotation and alignment. *D. rosae* genes were functionally annotated by eggNOG-mapper v. 2.1.7 (Cantalapiedra et al. 2021; Huerta-Cepas et al. 2021) and using the Non-redundant protein sequences (nr) database (the NCBI platform).

Firstly, we examined whether the genes specific to the *D. rosae* venom gland (Cambier et al. 2019) were upregulated during the early stages of gall formation. The reads from one available adult female venom gland sample (run SRR8501630) and one available adult female ovary sample (run SRR8501629) obtained from the SRA database (the NCBI platform) were aligned to the reference genome (**CHAPTER I**) by STAR v. 2.7.10b (Dobin et al. 2013). The count of reads aligned to the annotated genes was performed by featureCounts v. 2.0.6 (Liao et al. 2014). The number of read counts was then normalized by the total number of reads in the respective library. According to Cambier et al. (2019), genes showing at least 20-times higher number of read counts in the *D. rosae* venom gland compared to the ovary were considered up-regulated. Finally, the presence of these genes was assessed in the list of those overexpressed in the mid-July and the early September *D. rosae* larvae by aligning the gene sets using blastn of BLAST v. 2.14.0 (Camacho et al. 2009).

Secondly, we examined whether *D. rosae* orthologous genes encoding venom components (Cambier et al. 2019) and plant cell wall degrading enzymes (Hearn et al. 2019) in *B. pallida* were overexpressed during gall formation. Candidate orthologous genes overexpressed in the mid-July *D. rosae* larva, the early September larva, October salivary glands, and November salivary glands were identified by applying the bidirectional best hit strategy (Smith and Waterman 1981). First, the *D. rosae* genes (protein sequence, -query tag) were aligned to the *B. pallida* transcriptome (-db) (Hearn et al. 2023) by tblastn of BLAST v. 2.14.0 (Camacho et al. 2009). Next, the *B. pallida* sequences (-query) showing the highest bit score were aligned to the *D. rosae* genes (-db) by blastx of BLAST v. 2.14.0. (Camacho et al. 2009). The -db *D. rosae* genes showing the highest bit score were then compared with the -query *D. rosae* genes from the tblastn output: if the same *D. rosae* gene matched the same *B. pallida* sequence, they were considered orthologs.

Statistics. All statistical analyses were performed by R v 4.2.2 (R Core Team 2022 https://www.R-project.org/). The significance level was set to 0.05.

Results

Relative differential gene expression data analysis. BRAKER prediction (*.gff3* file) (Stanke et al. 2006, 2008; Hoff et al. 2016, 2019; Bruna et al. 2021) of the *D. rosae* revealed 125,626 genes, 135,510 mRNA transcripts, 346,652 exons and protein-coding sequences (CDS), and 211,236 introns. In the RNAseq analysis, the total number of reads aligned to the *gff3*. annotation varied between 98.61 million and 151.2 million depending on the library. The percentage of reads overlapping with the *D. rosae* genes varied between 38.72 % and 69.45 % (**supplementary table S10**). Pairwise differential expression analysis (**supplementary table S11**) revealed genes specifically expressed in the *D. rosae* egg (1028 genes), mid-July larva (2390), early September larva (455), October salivary gland (112), November salivary gland (2516), and female adult head (855) (**fig. 1**). The number of genes upregulated during gall formation (combined sample 'mid-July *D. rosae* larva + mid-July *D. eglanteriae* larva + early September *D. rosae* larva + October *D. rosae* salivary gland + November *D. rosae* salivary gland' vs combined sample 'head + egg') was 11,916. Genes encoding proteins containing leucine-rich repeat, plant cell wall degrading enzymes, and venom-like enzymes were found to be up-regulated during the whole process of galling.

Gene annotations and dynamics of gene expression. The total number of annotated genes by eggNOG (Cantalapiedra et al. 2021; Huerta-Cepas et al. 2021) was 62,690 including 47,871 genes annotated in Cluster of Orthologous Groups (COG) database and 12,461 genes in the Gene Ontology (GO). Among the 11,916 genes up-regulated at least at one stage during the whole process of gall formation (combined sample 'mid-July *D. rosae* larva + mid-July *D. eglanteriae* larva + early September *D. rosae* larva + October *D. rosae* salivary gland + November *D. rosae* salivary gland' vs combined sample 'female adult head + egg'), 5,152 genes were annotated at the protein families level (Finn et al. 2016), 4,088 genes were annotated with COG, and 2,208 genes were annotated with GO. The number of upregulated genes encoding proteins containing leucine-rich repeats (LRR), plant cell wall degrading enzymes (PCWDE) and venom-like enzymes was 15, 6, and 30, respectively (**supplementary table S12**). Five genes encoding LRR proteins and 13 genes encoding venom-like enzymes showed greater expression during July and September (active gall growth) compared to October and November (**fig. 2-4**). Among PCWDEs, the gene g94279 encoding cellulase was highly expressed from July to November (**fig. 5**). The other genes encoding PCWDEs showed higher expression from July to October, followed by a decline in November.

McDonald–Kreitman test. The initial number of extracted alignments (concatenated protein-coding sequences, CDS) was 135,472. The number of alignments showing at least

one polymorphic or divergent site was 120,442. The number of alignments showing at least one divergent site was 266. Genome-wide selection coefficient *s* was estimated at -2.1. Among genes specifically expressed in the mid-July larva, the early September larva, during active gall growing, and during whole gall formation, one gene was under positive selection (mean *s* estimation provided by SniPRE (Eilertson et al. 2012) was 0.74), 348 genes were under neutral selection (mean *s*: [-1.49; 0.38]), and 11,567 genes were under negative selection (mean *s*: [-4.32; -0.60]) (**fig. 6**). The gene (g50314) under positive selection encoded a transposable element.

Discussion

In this study, we performed the transcriptome analysis of *D. rosae* to identify candidate genes involved in gall formation. We focused on the genes that were upregulated during the earliest observable *D. rosae* larval stages in nature, i.e. mid-July and early September. We gave a particular focus on the candidate genes that were similarly annotated as those previously reported in other studies of gall-inducing insects (Zhao et al. 2015; Cambier et al. 2019; Hearn et al. 2019) for two reasons. The first reason is the absence of a significant number of gene annotations. Among the genes that were upregulated at least at one stage from July to November, only 43% of genes had available protein family annotations. Orthologous Group and Gene Ontology annotations were available for 34% and 19% of the up-regulated genes, respectively. Hence, performing an enrichment analysis solely on genes with available functional annotations could lead to inaccurate results. The second reason is that approximately half of the available functional annotations were attributed to transposable elements, whose role in galling is challenging to evaluate; the simultaneous overexpression of genes encoding transposable elements can be explained by an active transposition occurring during insect development. Hence many transposable are active in *D. rosae* genome. Thus, we investigated the *D. rosae* genes (1) encoding any proteins with the leucine-rich domain, as detected in the Hessian fly salivary gland protein (Zhao et al. 2015), (2) similar to those overexpressed in the *D. rosae* venom gland (Cambier el al. 2019), and (3) orthologous to those overexpressed in the *B. pallida* larva and venom gland (Cambier el al. 2019; Hearn et al. 2019).

The first group of up-regulated genes in the mid-July and early September *D. rosae* larvae encodes leucine-rich repeat (LRR) proteins: slit homolog, neuronal protein 2, podocan, vasorin-like protein, and Toll-like receptor 7. Slit homolog is involved in neural development and controls axon crossing in *Drosophila* (Brose et al. 1999; Kidd et al. 1999). The peptide matching the LRR domain of neuronal protein 2 is likely to be involved in synapse functioning (Linhoff et al. 2009). Thus, the expression of these genes can be related to insect development rather than being involved in galling. Other genes encoding LRR proteins can be associated with insect immune response. For example, podocan homolog was shown to be overexpressed in *Cnaphalocrocis medinalis* (Lepidoptera: Crambidae) in response to a baculovirus infection (Han et al. 2021). Vasorin-like protein and Toll-7 receptor belong to Toll-like receptors, a group of transmembrane proteins containing extracellular LRR motifs that play an essential role in invertebrate immunity and contribute to embryonic development in insects. These receptors recognize specific molecular patterns of various pathogenic microorganisms and initiate immune response (Medzhitov 2001; Leulier and Lemaitre 2008). Notably, Fjøsne et al. 2015 showed the vasorin-like protein to be overexpressed in *Dentrobaena veneta* (Annelida, Lambricidae) in response to a bacterial infection. Another study (Park et al. 2019) demonstrated that

the expression of Toll-like receptor 7 was induced by both bacterial and fungal infections in *Tenebrio molitor* (Coleoptera: Tenebrionidae).

The second group of examined genes was those encoding proteins previously detected in the *D. rosae* venom gland (Cambier et al. 2019): tetraspanin and putative transposase. Additionally, we examined the genes annotated as 'venom acid phosphatase' and orthologous to those up-regulated in the *B. pallida* venom gland (Cambier et al. 2019). Tetraspanins are a group of proteins implicated in multiple biological processes including insect development, reproductive processes, extracellular matrix organization, vesicle formation, and host-pathogen interactions (Todres et al. 2000; Hemler 2003). In host-pathogen interactions, tetraspanin serves as surface marker of immune cells and involved in signal transduction when generating immune response (Zhuang et al. 2007). For instance, Mei et al. (2023) showed overexpression of tetraspanin when *Bombyx mori* (Lepidoptera: Bombycidae) was exposed to a viral infection, and Mahadav et al. (2008) demonstrated that the parasitism by the wasp *Eretmocerus mundus* (Hymenoptera: Aphelinidae) induced the expression of tetraspanin in *Bemisia tabaci* (Hemiptera: Aleyrodidae). Other genes overexpressed in both *D. rosae* venom glands (Cambier et al. 2019) and the young larvae encode venom acid phosphatases, phospholipases A2, lipases, and peroxidase. Notably, some of these enzymes were detected in venoms of various parasitic Hymenoptera (Colinet et al. 2013; Poirie et al. 2014). The main role of parasitoid venom is to inhibit the insect host immune response. Common examples of immunomodulating parasitoid venom components are serine proteases (Asgari et al. 2003) and serine protease inhibitors (Colinet et al. 2009; Qian et al. 2015) that interrupt the formation of the melanin protective capsule in the insect host. However, the function of the other hydrolytic enzymes in regulation of the host immunity is unclear or has not been confirmed (Dani et al. 2005; Colinet et al. 2013; Dorémus et al. 2013; Poirie et al. 2014). Hence, the uncertain role of these proteins in parasitoid venoms presents a challenge when hypothesizing their role in gall wasps. Nonetheless, we can suppose that the hyperproduction of such proteins is due to insect development and immune challenges in the young *D. rosae* larvae. Firstly, lipases and phospholipases A2 are likely to be implicated in fatty acid metabolism during the gall wasp development. This could find support in the study of Akpinar et al. 2017 that demonstrated changes in fatty acid composition during the whole life cycle of *Diplolepis fructuum*. Secondly, these enzymes are involved in various processes including lipid metabolism and lipid signaling in insect development, reproduction, neurotransmission, as well as stress and immune responses (Stanley 2006; Shrestha et al. 2010; Hossen et al. 2016). Another enzyme showing diverse functions in insects is acid phosphatase playing a role in different biosynthetic and signaling pathways (Xia et al. 2000; Hossen et al. 2016; Lehmann 2021). An increase in peroxidase activity might be necessary to deactivate toxic molecules and radicals generated during the immune response that could take place during the gall wasp

development. Indeed, *D. rosae* larvae are exposed to high parasitic pressure: the percentage of parasitoid individuals in the *D. rosae* gall community can reach more than 70% (Rizzo and Massa 2006; Todorov et al. 2012). Furthermore, one of the most common parasitoids of *D. rosae*, *Orthopelma mediator* (Hymenoptera: Ichneumonidae), attacks *D. rosae* even before the gall begins to develop (Stille 1984). This can trigger an immune response in the gall wasp larvae at the early stages of gall formation. In addition, *D. rosae* can be also exposed to endophytic microorganisms. For instance, the microbiome of *Rosa* spp. consists of bacteria, such as *Bacillus* and *Staphylococcus* (Xia et al. 2020), as well as fungi (Zhao et al. 2018). Furthermore, hyperproduction of peroxidase in the young *D. rosae* larvae may be interpreted as a response to the defensive mechanisms of the host plant. When attacked by phyotphagous organisms, plants initiate a defensive cascade, releasing signaling molecules such as ATP, Ca^{2+}, and H_2O_2. These molecules trigger the expression of defense genes and, subsequently, plant defense responses (Guiguet et al. 2016). Hence, peroxidase of *D. rosae* could neutralize H_2O_2 produced by the host plant, thereby allowing it to evade plant immune response. Furthermore, the hyperproduction of enzymes like peroxiredoxin, which are also involved in the breakdown of H_2O_2, was detected in salivary glands of various insect herbivores (Guiguet et al. 2016). Lastly, in summer *D. rosae* larvae, we detected the overexpression of a gene encoding testicular haploid expressed repeat-like motif that has unclear role in invertebrates.

The third group comprises *D. rosae* genes that are orthologous to those encoding plant cell wall degrading enzymes (PCWDE) in *B. pallida* (Hearn et al. 2019): cellulase, pectate lyase, and rhamnogalacturonate lyase. Similar to other insect herbivores (Wybouw et al. 2016), the *D. rosae* PCWDE genes were acquired from bacteria via horizontal gene transfer . The PCWDE genes are found in cynipid gallers and inquilines, which is associated with their role in the formation of gall chambers for developing larvae. Furthermore, the PCWDEs genes have been lost in the members from the nested parasitoid family Figitidae, or are presented as fragments of functional domains or pseudogenes (Hearn et al. 2019, 2023). In *D. rosae*, we observed that the PCWDE genes showed the same expression level during the whole examined period of galling, i.e. July-November. However, we could suppose that PCWDEs may play a role not only in the formation of a gall chamber during larval nutrition but also in the earlier steps of gall formation, i.e. initiation. Indeed, the functioning of these enzymes leads to the release of degradation products serving as plant signaling molecules (Vallarino and Osorio 2012; De Lorenzo et al. 2019; Hearn et al. 2019). These metabolites could potentially modulate the differentiation of plant cells and provoke the development of gall tissue. Furthermore, the enzyme such as cellulase was detected in the *B. pallida* venom gland (Cambier et al. 2019). It could contribute to the degradation of plant cell wall during oviposition and also lead to release of metabolites, thereby initiating gall formation. Regrettably, due to the limitations in data quality (Cambier et al. 2019), we do not dispose of complete information about the venom

composition of *D. rosae* that could provide an additional argument. Nevertheless, *D. rosae* remains another example of Cynipidae highlighting the role of the genes encoding PCWDEs in gall formation, particularly during the growth stage.

As we have observed, it is challenging to link the studied genes with gall formation in *D. rosae*. Most of the examined genes can be predominantly involved in insect development and nutrition rather than host plant manipulation. Nonetheless, we observed that various upregulated genes were associated with the immune response of insects. The last observation highlights that almost all genes were under negative selection, and only one gene encoding a transposable element was under positive selection. It can be explained by the lack of power to detect signature of positive selection in the asexual mode of reproduction of *D. rosae*, i.e. thelytokous parthenogenesis via gamete duplication (Nordlander 1973; Stille and Dävring 1980; Heimpel and De Boer 2008). This mode of reproduction results in complete homozygosity where all alleles are in complete linkage. Consequently, in highly homozygous thelytokous *D. rosae* females, disadvantageous alleles and all linked alleles are purged quickly (Pearcy et al. 2006). However, positive selection may occur but remain undetectable due to the extremely low recombination rate in *D. rosae* (Mozhaitseva et al., 2023). Males constitute only 0-4% of *D. rosae* populations (Nordlander, 1973; laboratory observations) and have minimal contribution to each generation. This leads to the phenomenon of clonal interference (Muller 1932). In asexually reproducing organisms, a beneficial mutation takes a longer time to be fixed or can be lost in the absence of recombination and the inability to spread in a population. Thus, we were unable to detect the signatures of positive selection such as elevated relative frequency of a non-synonymous polymorphism in the protein-coding sequences of *D. rosae*.

Conclusion

Transcriptome analysis revealed the overexpression of 11,916 *D. rosae* genes during the observable stages of gall formation (mid-July to early November), with 7,153 genes being overexpressed specifically during the active gall growth phase (mid-July to early September). Among the examined genes, those encoding plant cell wall degrading enzymes can be associated with galling. The other upregulated genes are likely to be implicated in insect development and the immune response to parasitoids and the host plant microbiota. Almost all genes have been found to be under negative selection. It could be explained by the mostly asexual mode of reproduction in *D. rosae*, which results in a rapid purge of deleterious and all linked alleles. Besides, positive selection could be undetectable because of reduced recombination in this species and the effect of clonal interference.

Further investigations could be focused on designing experiments that allow to examine the *D. rosae* transcriptome sampled during gall initiation. One should master the life cycle of *D. rosae* in greenhouse conditions. It could help to pinpoint the moment and location of oviposition during the experiment and measure gene expression levels in *D. rosae* eggs and young larvae after the first hours and days. Additionally, other groups of genes, such as those involved in the disrupting danger signal molecules in plants, could be examined using the same methodology as in this study. Improving gene annotations and conducting enrichment analyses would help to identify gene groups with specific functions. Lastly, we could analyze the evolution of candidate genes associated with galling within Cynipidae s. lat. and infer whether any genes present in gall inducers and absent in their parasitic members.

Figures

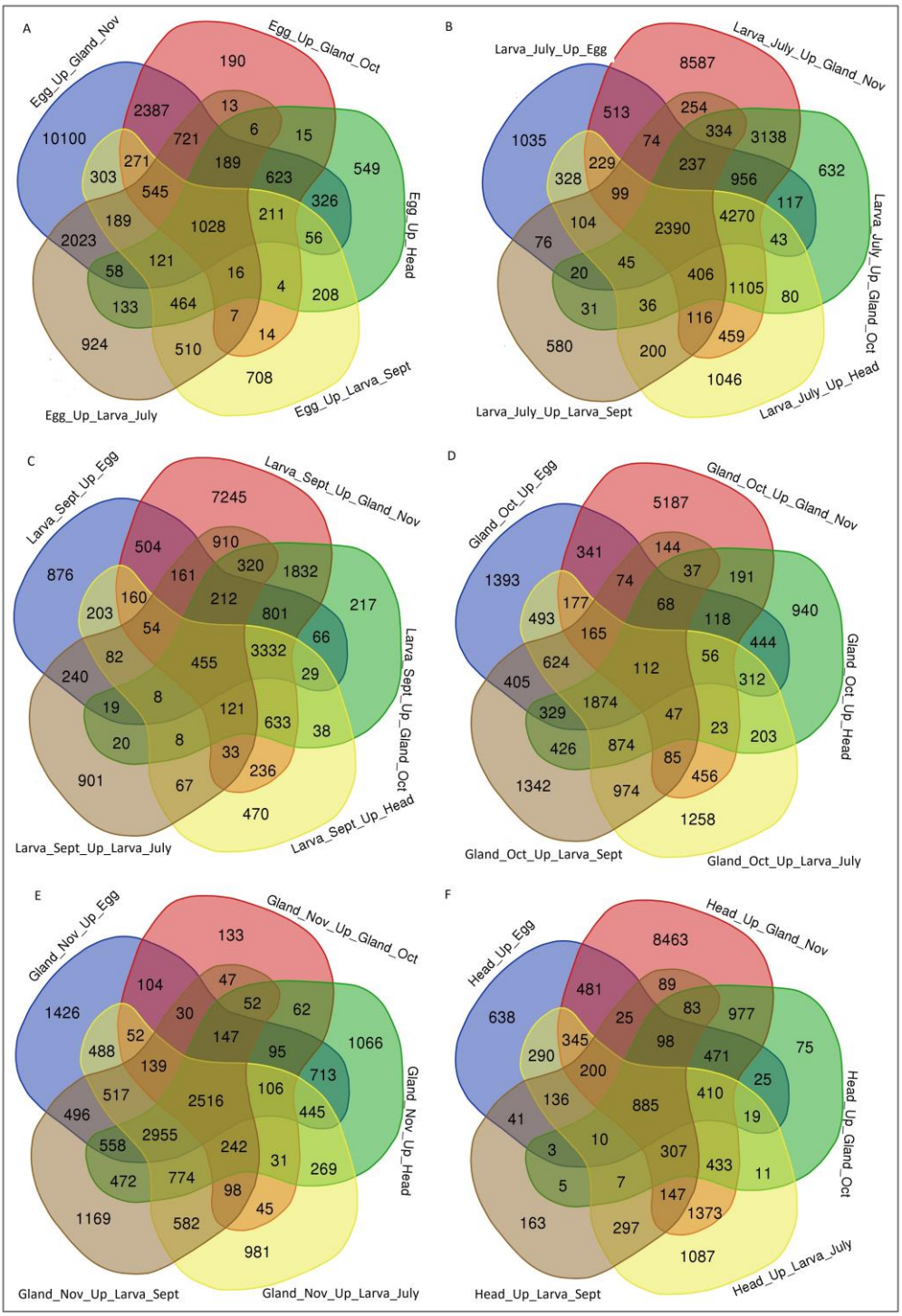


Fig. 1. Number of genes specifically expressed in the *Diplolepis rosae* (A) eggs removed from a female adult (Egg_Up), (B) mid-July larva (Larva_July_Up), (C) early September larva (Larva_Sept_Up), (D) October larva salivary gland (Gland_Oct_Up), (E) November larva salivary gland (Gland_Nov_Up), and (F) female adult head (Head_Up) found by relative differential gene expression analysis. Up: relatively over-expressed genes found in the first sample compared to the second one. Venn diagrams were drawn by the tool available at https://bioinformatics.psb.ugent.be/webtools/Venn/.

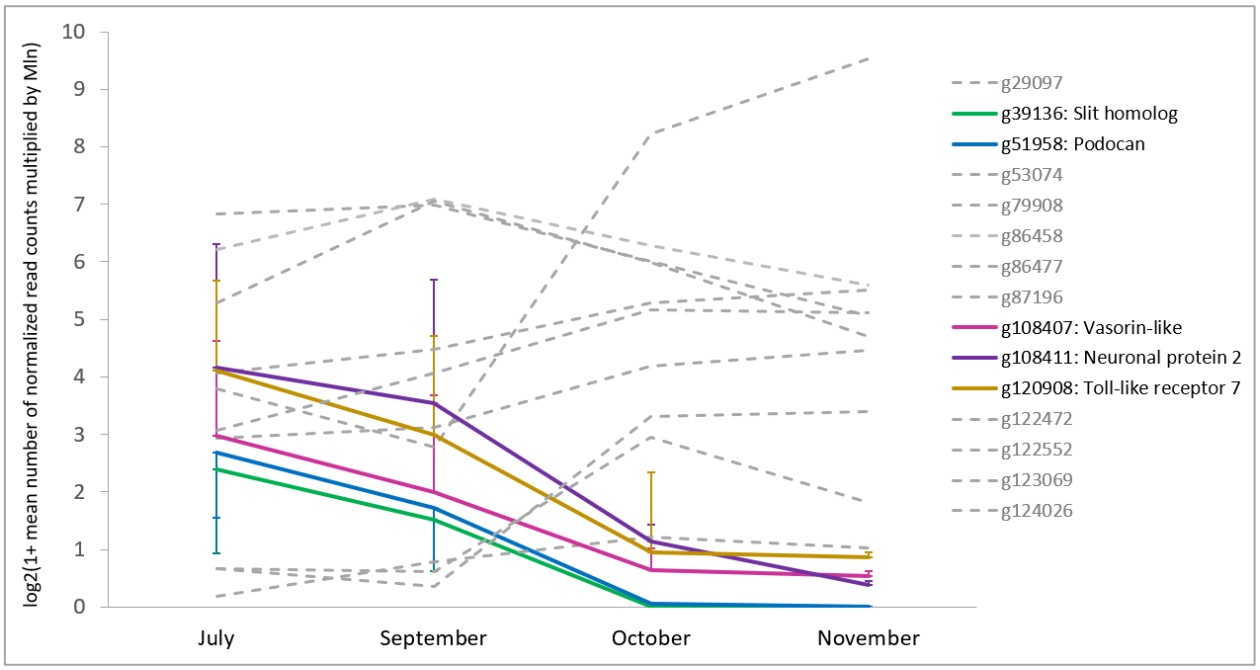


Fig. 2. Gene expression dynamics of proteins containing leucine-rich repeats in *Diplolepis rosae*. Results are presented as mean ± SEM. Genes g108407, g108411, and g120908 show only the upper SEM bound. Genes g39136 and g51958 show only the lower SEM bound. Colored lines represent genes with significantly higher expression in July and September (combined sample 'mid-July *D. rosae* larva + early September *D. rosae* larva + mid-July *D. eglanteriae* larva') compared to October and November ('October *D. rosae* larva salivary glands + November *D. rosae* larva salivary glands'). Dashed grey lines represent genes that did not show significant overexpression in July and September compared to October and November. g29097: slit homolog; g53074: chaoptin-like protein; g79908: U2 small ribonucleoprotein A'; g86458: peroxidasin-like protein; g86477: Protein phosphatase 1 regulatory subunit 42-like; g87196: chaoptin; g122472: slit-like protein; g122552: insulin-like growth factor-binding protein; g123069: fibronectin type III domain containing protein; g124026: follicle-stimulating hormone receptor.

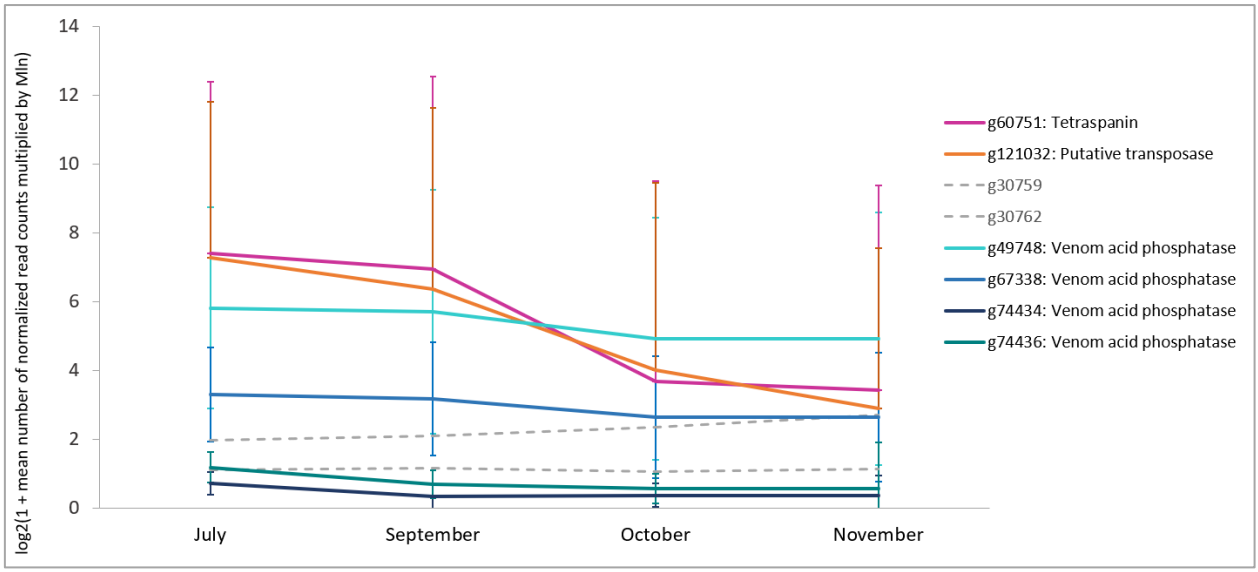


Fig. 3. Expression of genes up-regulated in the venom gland and the larvae of *Diplolepis rosae*. Results are presented as mean ± SEM. Genes g60751 and g121032 show only the upper SEM bound. Colored lines represent genes with significantly higher expression in July and September (combined sample 'mid-July *D. rosae* larva + early September *D. rosae* larva + mid-July *D. eglanteriae* larva') compared to October and November ('October *D. rosae* larva salivary glands + November *D. rosae* larva salivary glands'). Dashed grey lines represent genes that did not show significant overexpression in July and September compared to October and November. g30759 and g30762: Bi-VSP-like venom serine protease.

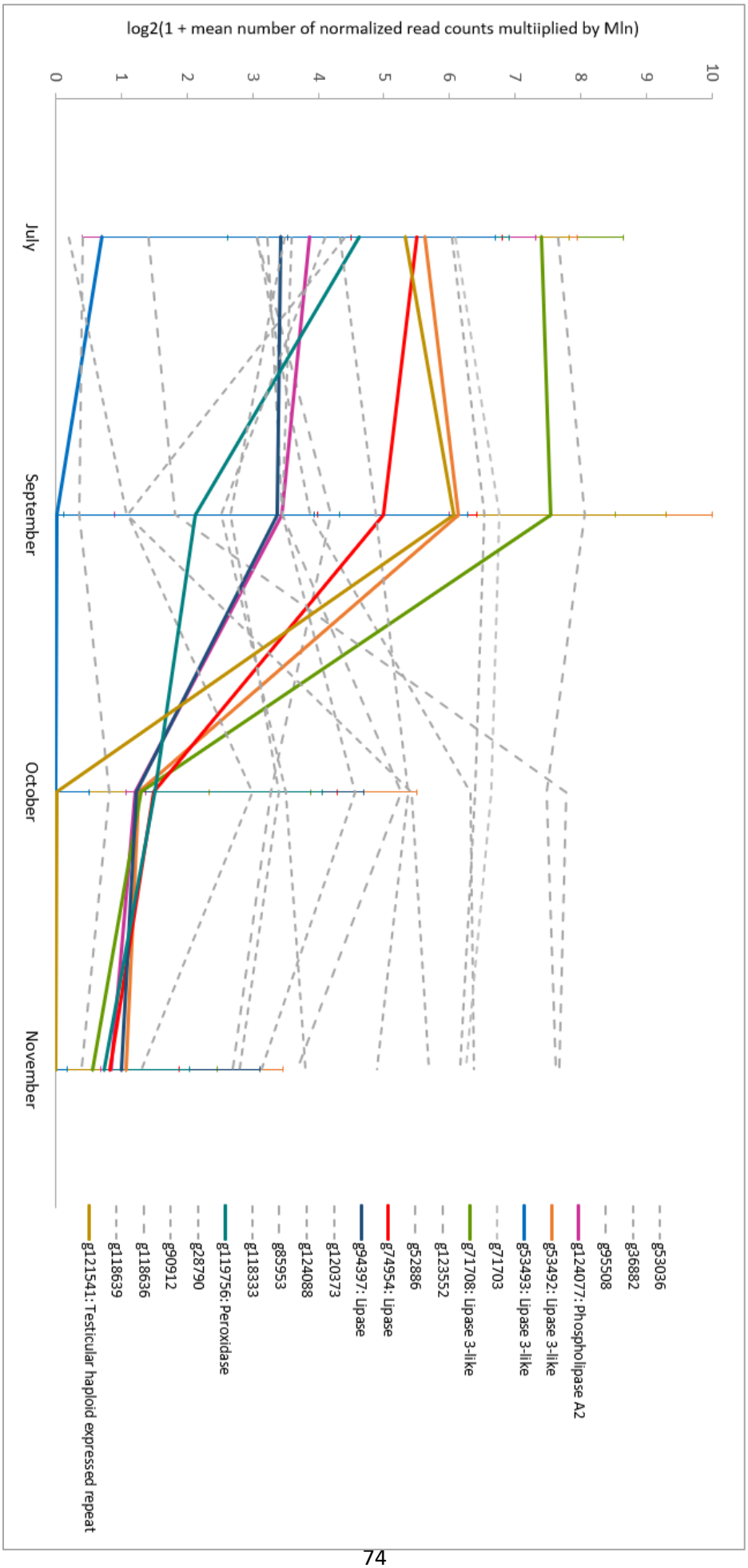

log2(1 + mean number of normalized read counts multiiplied by Mln)
0
1
2
3
4
5
6
7
8
9
10
July
September
October
November
g53036
g36882
g95508
g124077: Phospholipase A2
g53492: Lipase 3-like
g53493: Lipase 3-like
g71703
g71708: Lipase 3-like
g123552
g52886
g74954: Lipase
g94397: Lipase
g120373
g124088
g85953
g118333
g119756: Peroxidase
g28790
g90912
g118636
g118639
g121541: Testicular haploid expressed repeat

Fig. 4. Expression of *Diplolepis rosae* genes orthologous to genes up-regulated in the venom gland of *Biorhiza pallida*. Results are presented as mean ± SEM (upper bounds). Colored lines represent genes with significantly higher expression in July and September (combined sample 'mid-July *D. rosae* larva + early September *D. rosae* larva + mid-July *D. eglanteriae* larva') compared to October and November ('October *D. rosae* larva salivary glands + November *D. rosae* larva salivary glands'). Dashed grey lines represent genes that did not show significant overexpression in July and September compared to October and November. g53036: chitooligosaccharidolytic beta-N-acetylglucosaminidase; g36882: group XII secretory phospholipase A2; g95508: phospolipase A2-like; g71703: lipase-3-like enzyme; g123552: lipase family; g52886: lipase; g120373: pancreatic triacylglycerol lipase; g124088: peroxidase-like enzyme; g85953: peroxidase; g118333: peroxidasin homolog; g28790: protein D2-like; g90912: apyrase; g118636: inosine-uridine preferring nucleoside hydrolase; g118639: lysozyme C1-like.

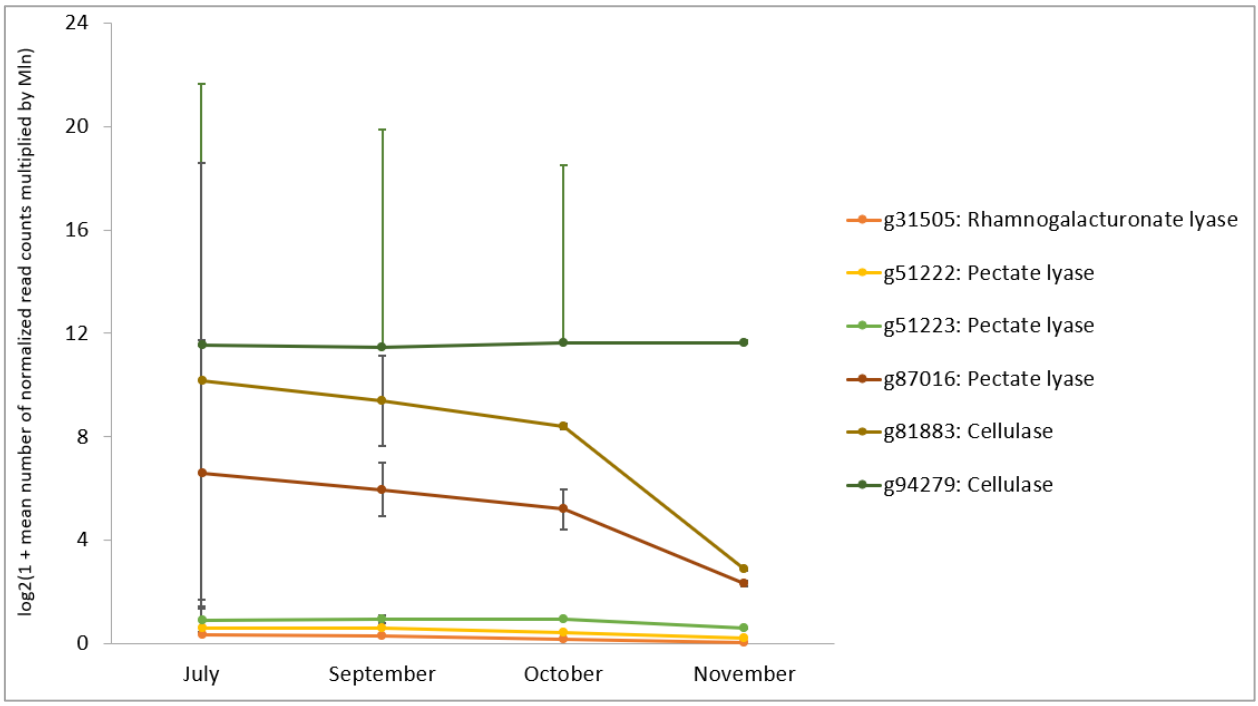


Fig. 5. Gene expression dynamics of plant cell wall degrading enzymes in *Diplolepis rosae*. Results are presented as mean ± SEM. Genes g81883 and g94279 shows only the upper SEM bound. When comparing gene expression between summer (combined sample 'mid-July *D. rosae* larva + early September *D. rosae* larva + mid-July *D. eglanteriae* larva') and autumn ('October *D. rosae* larva salivary glands + November *D. rosae* larva salivary glands'), no significant results were found.

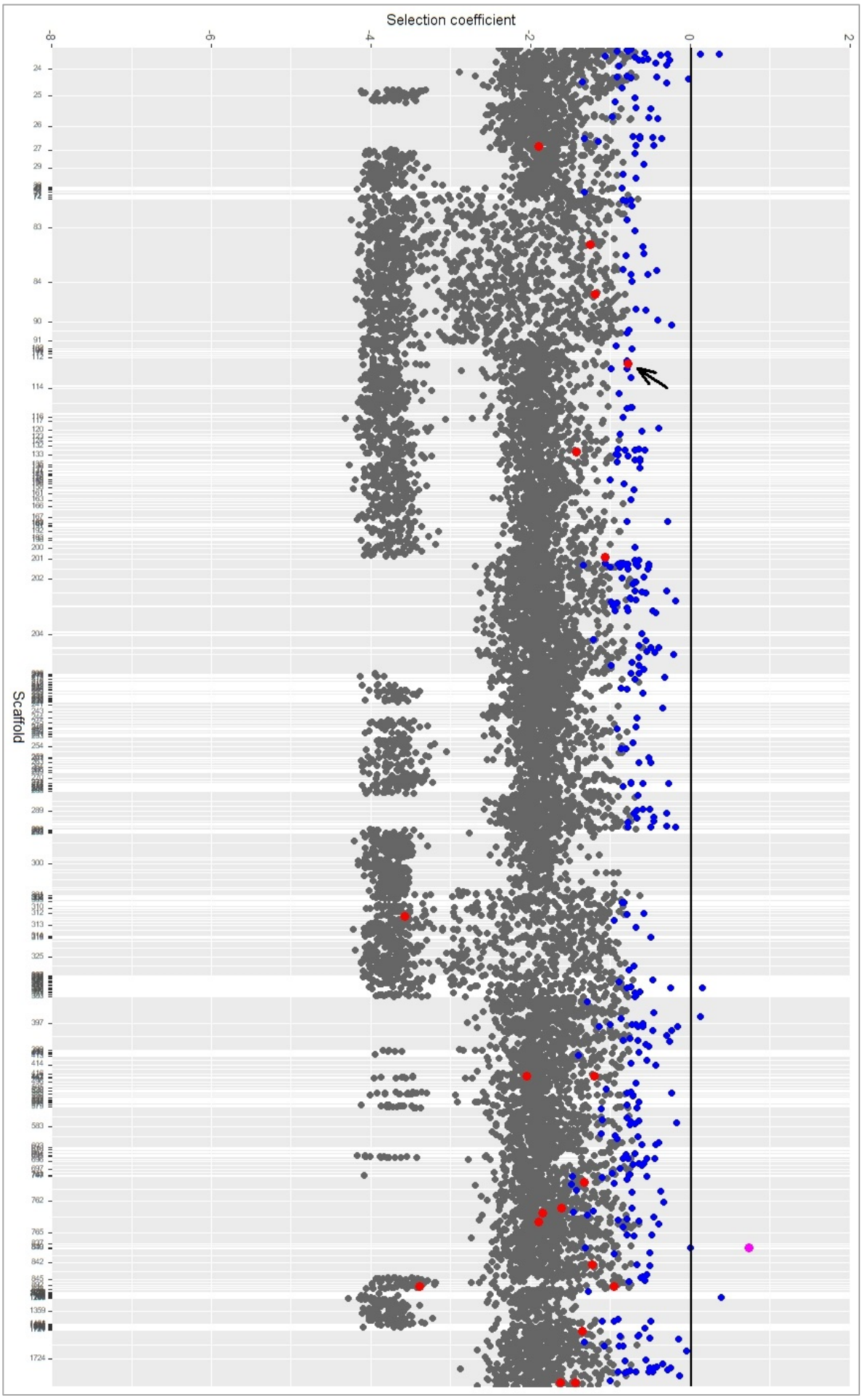

Selection coefficient
-8
-6
-4
-2
0
2
Scaffold
24
25
26
27
29
83
84
90
91
114
120
133
202
204
254
289
300
313
325
397
583
762
785
842
1359
1724

Fig. 6. Genomic scan of the selection coefficient (s) estimated by SnIPRE (Eilertson et al. 2012) in the protein-coding sequences of the *Diplolepis rosae* genome. The numbers indicate scaffolds in the genome assembly. The black line indicates the selection coefficient equal to zero, which reflects the absence of selection. Dark grey points: mean s estimations corresponding to genes under negative selection. Blue points: mean s estimations corresponding to genes under neutral selection. Rose point: mean s estimation corresponding to a gene under positive selection. Red points: genes that were found to be up-regulated in summer *D. rosae* larvae compared to autumn *D. rosae* larvae (combined sample 'mid-July *D. rosae* larva + early September *D. rosae* larva + mid-July *D. rosae* larva' vs combined sample 'October *D. rosae* salivary gland + November *D. rosae* salivary gland') (**Fig. 2-4**). All these genes are under negative selection except for that encoding tetraspanin (g60751, marked by flash), that is under neutral selection.

CHAPTER III. Do structural variations in the genome of *Cynips quercusfolii* reflect the types of asexual females?

Introduction

Cyclical parthenogenesis, or heterogony, is a reproductive mode when organisms obligately alternate between asexually and sexually reproducing forms. Heterogony is present in the trematodes, the rotifers, the crustaceans, and several insect orders (Hemiptera, Diptera, Coleoptera, and Hymenoptera) (Bell 1982). In heterogonic arthropods, such as *Daphnia* and aphids, the number of generations of clonally reproducing females varies and depends on environmental conditions: the switch to sexual reproduction is triggered, for instance, by colder temperatures and shorter daylight (Simon et al. 2002; Decaestecker et al. 2009).

In Cynipidae, the two tribes, Cynipini and Pediaspidini, reproduce by cyclical parthenogenesis (Stone et al. 2002). The origin of heterogony and the genetic mechanism explaining the alternation between both generations in heterogonic cynipids remains little studied (Pujade-Villar et al. 2001; Stone et al. 2002).

Contrary to crustaceans and aphids, a general life cycle of bivoltine cynipids involves the **obligate alternation of two generations: one asexual generation and one sexual generation**. The asexual generation is presented by two types of females: gynephores and androphores (**fig. 1A**). Gynephores produce diploid sexual females by thelytoky, and androphores produce haploid males through unfertilized eggs. There were two major schemes explaining how sexual generation could produce either gynephores or androphores: 1) gynephores produce two types of sexual females that, after mating, give either gynephores or androphores (**fig. 1B**), and 2) androphores produce two types of males that mate with identical sexual females, resulting in two types of asexual females (**fig. 1C**). Atkinson (2000) and Folliot (1964) examined both hypotheses and supported the second one: gynephores produce genetically identical sexual females by clonal apomixis (Atkinson 2000) or via gamete duplication (automixis) (Folliot 1964) and androphores produce haploid males.

The described life cycle scheme (Fig 1) is the least complex and mostmost common in heterogonic Cynipidae. Nevertheless, there are several deviations. First, in some species like *Biorhiza pallida* and some *Andricus* spp., there is only one type of asexual females (termed gynandrophores) producing both females and males (Folliot 1964; Atkinson 2000). Second, in rare cases (*B. pallida* and *Pediaspis* sp.), sexual females are capable to produce a small number of asexual females without mating (Folliot 1964). Third, the deviations concern the duration of the life cycle. In general, there are two generations per year. However, depending on environmental conditions the asexual generation can

facultatively or obligatory take more than one year to develop. In other cases, each generation takes one year (Pujade-Villar et al. 2001; Stone et al. 2002).

Cynips quercusfolii (Linnaeus, 1758) (Cynipidae: Cynipini) is a gall wasp species common in Europe (Dinç 2017) that causes cherry galls on leaves in oaks *Quercus* spp. (Fagaceae). In *C. quercusfolii*, the females of the asexual generation emerge in late autumn and oviposit into dormant oak lateral leaf buds. In spring, small red sexual galls develop from the buds. The females and males of the sexual generation emerge from these galls in early summer and mate. The females oviposit into oak leaf veins inducing gall formation (Giertych et al. 2013). Asexual females develop before next autumn (**fig. 2**).

In this chapter, we present the initial findings about the genome and population structure of *C. quercusfolii*, taking into account its mode of reproduction. Considering the obligate alternation of sexual and asexual reproductive modes, we suggest that such switch from one generation to another should be encoded in the *C. quercusfolii* genome. Firstly, we hypothesized that the distribution of highly homozygous and heterozygous regions could determine the type of asexual female (gynephore or androphore) in *C. quercusfolii*. Secondly, we examined whether the asexual females differed in terms of structural variations, such as inversions, deletions, and duplications. We supposed that chromosomal rearrangements could occur in genes involved in meiotic recombination. Indeed, inversions can play an important role in blocking of recombination, thereby leading to co-segregating alleles at the same locus, used to be called a 'supergene' (Kelly 2000). In *C. quercusfolii*, we can suppose a scheme involving multiple mutations that led to each the type of asexual female (**fig. 3**). Our scheme fits the Folliot's (1964) model involving automictic parthenogenesis and production of two male genotypes. In addition, we can consider another type of thelytokous parthenogenesis, i.e. automictic thelytoky via central or terminal fusion or gamete duplication thelytoky. Atkinson (2000) supposed clonal apomixis in gynephores that produce sexual female offspring. However, in apomixis, diploid ootids undergo only mitotic division, which leads to clonal female offspring (Rabeling and Kronauer 2013). In this case, it is unclear why this female offspring being absolutely identical to its asexual gynephore mother would mate to complete the cycle.

Materials and Methods

Sampling. Seven *C. quercusfolii* asexual galls were collected from *Quercus spp*. from August 2019 to October 2021 in France (**supplementary table S13**). The galls were kept in plastic bags at room temperature. Wasp females were removed by dissection of gall tissue and homogenized in 2-ml plastic tubes by a TissueLyser (Qiagen) with adding a metallic bead. DNA was extracted from one individual per sample using a DNeasy Blood and Tissue Kit (QIAGEN, Germany). Before the sequencing, *C. quercusfolii* was confirmed by performing the PCR of the gene encoding cytochrome oxidase c subunit I. The sequence of a PCR product was examined in the Nucleotide collection (nt) database (the NCBI platform: https://www.ncbi.nlm.nih.gov/). After that, the Illumina sequencing was performed by Genotoul Sequencing Platform, Toulouse, France (https://get.genotoul.fr).

Reference genome. The reconstruction of the *C. quercusfolii* genome was performed by A. Branca. Short reads were assembled using the GATB pipeline (https://github.com/GATB/gatb-minia-pipeline) that use the minia assembler and BESST scaffolding (Sahlin et al. 2014). Then scaffolding of the GATB assembly was performed with RagTag using *B. kinseyi* as backbone (Cynipidae: Cynipini) (RefSeq assembly accession GCF_010883055.1) (Alonge et al. 2022). A total length of the assembly was 1.57 Gb, a total number of contigs was 424,389, an N50 was 82.1 Mbp, an L50 was 9, and the largest scaffold was 108.6 Mp long, the percentage of detected BUSCOs (Manni et al. 2021) for hymenopteran (N=5991) was 86%, among which 85% in complete single copy). In this study, only the longest scaffolds of 68.8 – 108.6 Mbp long were examined, which represented 55.4% of the assembly.

SNP calling. *C. quercusfolii* reads of the different individuals were aligned to the reference using Bowtie 2 v. 2.3.5.1 (Langmead and Salzberg 2012). The alignments were processed for further manipulations using Samtools 1.7 (view, sort, index, depth, and stat options) (Danecek et al. 2021). The quality of each alignment was estimated by calculating the average percentage of the mapped reads (varied between 31.0% and 97.3%), the average quality (35.6 - 35.8), the average coverage (12x - 42x), and the error rate (0.015 - 0.024). Thereafter, the identification of polymorphisms across the genome was performed using a pipeline proposed by GenomeAnalysisToolkit (GATK) v. 4.0 (AddOrReplaceReadGroups, HaplotypeCaller, CombineGVCFs, and GenotypeGVCFs tools) (https://gatk.broadinstitute.org/hc/en-us/articles/360035890411-Calling-variants-on-cohorts-of-samples-using-the-HaplotypeCaller-in-GVCF-mode). A final multi-sampled *.vcf* file was created by bcftools v. 1.7 merge (Danecek et al. 2021). Indels and non-biallelic sites were removed from the *.vcf* file using bcftools v. 1.7 (filter command) (Danecek et al. 2021). The rare (--maf 0.05), low-quality (--minQ 30), low-depth (cutoff below 2.5th

percentile), and high-depth (cutoff above 97.5th percentile) sites were removed using vcftools 0.1.17 (Danecek et al. 2021). Phasing of the *.vcf* file was performed by beagle v. 5.4 (Browning et al. 2021). The total number of examined SNPs was 11,016,204.

Population genomic statistics and population structure (supplementary code S6). The pairwise absolute divergence Dxy (Nei 1987) was calculated within each chromosome using pixy v. 1.2.6.beta1 and normalized by the total number of polymorphisms (Korunes and Samuk 2021). Per-individual heterozygosity H was calculated by vcftools 0.1.17 (Danecek et al. 2021) and presented as a number of heterozygous SNPs normalized by the assembly length. Population structure was studied using principal component analysis (PCA). The PCA was performed by plink v. 1.9 (Purcell et al. 2007). The output *.eugenvec* and *.eugenval* files were used to plot the PCA by the *tidiverse* R package v. 1.3.0 (Wickham et al. 2019).

Detection of runs of homozygosity and runs of heterozygosity (supplementary code S7). Detection of runs of homozygosity (ROH) and runs of heterozygosity (ROHet) was performed using the detectRUNS R package v. 0.9.6 (slidingRuns mode, window size 10, min number of SNP 10, and min run length 10 kbp) (Biscarini et al. 2018; Purcell et al. 2007). Necessary *.map* and *.ped* files were created by plink v. 1.9 (Purcell et al. 2007). The data were visualized in R v 4.2.2 (R Core Team 2022).

Detection of structural variations (supplementary code S8). Detection of chromosomal rearrangements was performed by Delly v. 1.1.7.0 (Rausch et al. 2012) in scaffold 57 and scaffold 62 of the *C. quercusfolii* assembly. Prior to generation of a *.vcf* file showing structural variations, duplicated aligned reads (sorted *.bam* alignments) were located by MarkDuplicates (Picard v. 2.1.18) (https://gatk.broadinstitute.org/hc/en-us/articles/360037052812-MarkDuplicates-Picard-). Visualization of structural variations was performed by the *intansv* R package v. 1.40.0 (Yao 2023). The minimum and the maximum size of variations to visualize was 10 kbp and 0.1 Gb, respectively.

***Wolbachia* identification.** *Wolbachia* sequences we searched in the *C. quercusfolii* genome by the method described in **CHAPTER I** ('Materials and Methods: *Wolbachia* identification' section).

McDonald–Kreitman test. The MK test was performed as described in **CHAPTER II** ('McDonald–Kreitman test' in the 'Materials and methods' section). *Belonocnema kinseyi* (RefSeq assembly accession GCF_010883055.1) was used as an outgroup. Among a total of 20,163 genes detected in the studied assembly, 5,757 monoexonic genes were examined in the study.

Statistics. All statistical analyses were performed by R v 4.2.2 (R Core Team 2022 https://www.R-project.org/). The significance level was set to 0.05.

Results

Population structure. Principal component analysis of the population structure showed the first components explaining 40.00% and 23.36% of the variance, respectively (**fig. 4-5**). There was no correlation between the geographical distance between samples and pairwise absolute divergence Dxy (Spearman correlation test: ρ = -0.323, p = 0,153) (**fig. 6**, **table 1**). Per-individual heterozygosity varied between 0.00088 and 0.0031 (**table 2**). The sample 359c, which is the most differentiated, is also the one showing the lowest heterozygosity.

Runs of homozygosity and runs of heterozygosity. A total number of homozygous blocks per individual varied between 0 and 3197 (**table 2, fig. 7, supplementary fig. S27-S36**). The largest contiguous homozygous region reached 418,650 bp long and was detected in *Cynips quercusfolii*-490. A total number of contiguous heterozygous blocks varied 43 to 993 (**table 2, fig. 7, supplementary fig. S27-S36**). The largest heterozygous region was detected in *Cynips quercusfolii*-705.5 and reached 74,908 bp. The sample 60a presented by far the highest number of runs of homozygosity and runs of heterozygosity.

Structural variations. In scaffold 57 (**fig. 8, supplementary table S14**), the individuals *C. quercusfolii*-60a and 638b showed the same 28.7-Mbp deletion; *C.quercusfolii*-305a and 359c showed the same 2.06-Mbp deletion. In scaffold 62 (**fig. 8, supplementary table S14**), no individuals showed any shared deletion. In scaffold 57 **(fig. 9, supplementary table S14**), the individuals 18c, 638b, and 705.5 had the same 43.1-Mbp duplication; the individuals 305a and 359c showed the same 9.30-Mbp duplication. In scaffold 62 (**fig. 9, supplementary table S14**), the individuals 18c, 60a, 638b, and 705.5 had the same 25.8-Mbp duplication. In scaffold 57 (**fig. 10, supplementary table S14**), the individuals 18c, 638b, and 705.5 had the same 46.7-Mbp inversion that containing 1109 genes. In scaffold 62 (**fig. 10, supplementary table S14**), the individuals 18c, 305a, 359c, and 490 showed the same 52.2-Mbp inversion, in which no genes were detected; the 305a, 359c, and 490 showed the same 8.16-Mbp inversion containing 262 genes. All discussed inversions were in homozygous state.

***Wolbachia* identification.** No *Wolbachia* sequences were identified in the *C. quercusfolii* genome.

McDonald–Kreitman test. The initial number of extracted alignments was 5757. The number of alignments showing at least one polymorphic or divergent site was 4730. Genome-wide selection coefficient *s* was estimated at -1.5. Ten genes were under positive selection (mean *s* estimation varied 0.58 to 1.7), 2734 genes were under neutral selection

(mean *s*: [-2.4; 0.81]), and 1986 genes were under negative selection (mean *s*; [-6.6; -0.74]). Among genes under positive selection, 8 genes exhibited no functional annotation, and 2 genes were annotated as transposable elements. Two genes under positive selection encoded uncharacterized proteins on chromosome 57 (chr_57) (**Table 3**) were located in the inversion (scaffold 57) shared among *C. quercusfolii*-60a, 638b, 705.5, and 18c (**Fig. 10**).

Discussion and Conclusion

In this part, we performed the first steps to reveal the genome and population structure of *C. quercusfolii*. We acknowledge that our observations were restricted to only seven diploid individuals and half of the examined genome assembly. Nonetheless, we draw the following preliminary discussion.

C. quercusfolii exhibits no particular population structure with two individuals, *C. quercusfolii*-359c (PC1 in **fig. 5**) and *C. quercusfolli*-60a (PC2 in **fig. 5**), being distinct from the others. *C.quercusfolii*-359c shows the lowest per-individual heterozygosity, whereas *C. quercusfolli*-60a shows the highest number of repetitive homozygous and heterozygous regions across the genome.

In the studied chromosomes we found several chromosomal rearrangements. Interestingly, in scaffold 57, the females 60a, 638b, 705.5 shared the same inversion, whereas in scaffold 62, the females 305a, 359c, and 490 shared another inversion. Furthermore, the last female, *C. quercusfolii*-18c, presented both inversions (**fig. 10**). The size of these inversions reaches several tens of Mbp, that contains several hundreds of genes. This could block the recombination in the inverted genome regions and, therefore, lead to balancing selection maintaining different allele combinations. The two alleles resulting from the inversion could encode two male genotypes resulting in the two types of asexual females in heterogonic Cynipidae after mating. However, all examined females could belong to only one type of asexual females. The observed structural variations could reflect various *C. quercusfolii* lineages adapted to local conditions. The detected inversions could lead to a combination of linked alleles advantageous in the given environment. For instance, an inversion in a quantitative trait locus was associated with two ecotypes of the rough periwinkle (*Littorina saxatilis*): a larger and thicker ecotype occurred in boulder fields and a smaller wave ecotype occurred in rocky shores (Koch et al. 2021). In the case of *C. quercusfolii*, the observed inversions might be linked to locally adapted genotypes, such as those specific to the host plant genotype or the parasitic complex.

The additional test for selection revealed that the majority of genes in *C. quercusfolii* are under either negative or neutral selection. Genes under positive selection either lack functional annotations and or annotated as transposable elements. This presents a challenge to discuss their role in *C. quercusfolii* and other cynipids.

To confirm our observations, we require more evidence. Firstly, we need a high-quality genome assembly and more individuals to confirm our findings. To collect more asexual females is quite easy, but to assemble the genome of Cynipidae is challenging. The main reason is the high number of repetitive sequences (up to 35%). It leads to errors during the assembly and a worst quality of the final data, that cannot be taken into account during the examination of the genome structure. Secondly, it would be valuable to collect and sequence the sexual generation of *C. quercusfolii*. The genomes of sexual

individuals could demonstrate whether the inversion polymorphism is shared within males or females. Thirdly, one should annotate genes found in the inverted genome regions of *C. quercusfolii* and evaluate whether their function can be associated with the maintenance of the heterogonic life cycle.

Tables

Table 1. Geographical distance (km) (yellow) and pairwise absolute divergence (green) Dxy between *Cynips quercusfolii* individuals.

Distance, km / Dxy	18c	305a	359c	490	60a	638b	705.5
18c	-	0.211	0.144	0.196	0.298	0.199	0.217
305a	60.2	-	0.157	0.208	0.310	0.238	0.229
359c	378	340	-	0.142	0.254	0.173	0.165
490	3.54	56.8	372	-	0.296	0.223	0.213
60a	46.2	21	344	43.5	-	0.282	0.310
638b	375	366	103	374	342	-	0.244
705.5	125	184	460	122	167	457	-

Each number (18c – 705.5) represent one *C. quercusfolli* individual.

Table 2. Total number of runs of homozygosity (ROHs) and runs of heterozygosity (ROHets) and per-individual heterozygosity of *Cynips quercusfolii*.

Individual / Characteristic	Total number of ROHs	Total number of ROHets	Per-individual heterozygosity
Cynips quercusfolii-18c	31	285	0.0017
Cynips quercusfolii-305a	23	553	0.0017
Cynips quercusfolii-359c	0	43	0.00088
Cynips quercusfolii-490	15	330	0.0015
Cynips quercusfolii-60a	3197	993	0.0023
Cynips quercusfolii-638b	1	78	0.0031
Cynips quercusfolii-705.5	42	915	0.0021

Heterozygosity is presented as a number of heterozygous polymorphisms normalized by the total number of SNPs.

Table 3. **Functional annotation of the genes under positive selection detected in the *Cynips quercusfolii* genome.**

Gene id	L, aa	blastp sp align	Blastp match	% cov	% id	E value	tot sc
chr_61_1.0	54	*Phymasttichus coffea* (Chalcidoidea)	GVQW3 motif containing protein-like	100	75.93	3e-20	89.4
chr_61_1.0	54	*Acromyrmex insinuator* (Formicoidea)	MOS1T transposase	98	79.25	4e-20	87.8
chr_57_20.5	531	*Belonocnema sp.*	Uncharacterized protein	100	79.28	0.0	861
chr_57_26.6	202	*Belonocnema sp.*	Uncharacterized protein	100	46.34	3e-37	139
chr_60_4.1	106	*Belonocnema sp.*	Tigger transposon	100	69.81	6e-41	145
chr_64_4.7	490	*Belonocnema sp.*	Uncharacterized protein	98	33.88	2e-88	293
chr_60_86.9	102	*Belonocnema sp.*	Uncharacterized protein	100	57.84	9e-32	121
chr_65_8.8	70	*Belonocnema sp.*	Uncharacterized protein	100	52.11	5e-06	52.0
chr_58_90.0	129	*Belonocnema sp.*	Uncharacterized protein	98	36.84	7e-13	73.6
chr_59_9.0	114	*Belonocnema sp.*	Uncharacterized protein	100	71.05	1e-43	149
chr_58_97.0	111	*Belonocnema sp.*	Uncharacterized protein	96	39.64	3e-11	68.6

Gene id: gene identifier, chromosome (scaffold) number followed by a start position of a protein-coding sequence, in Mbp. L, aa: protein length, amino acids. blastp sp align sp: taxonomic name of the species showing protein sequences matching to the query *Cynips quercusfolii* sequence. % cov: query cover, proportion of the query *C. quercusfolii* sequence that is aligned with the database sequence. % id: percent of identity between the query *C. quercusfolii* sequence and the database sequence. E value: expectation value of the alignment between the query *C. quercusfolii* sequence and the database sequence. tot sc: total score reflecting the strength of the match between the query *C. quercusfolii* sequence and the database sequence.

Figures

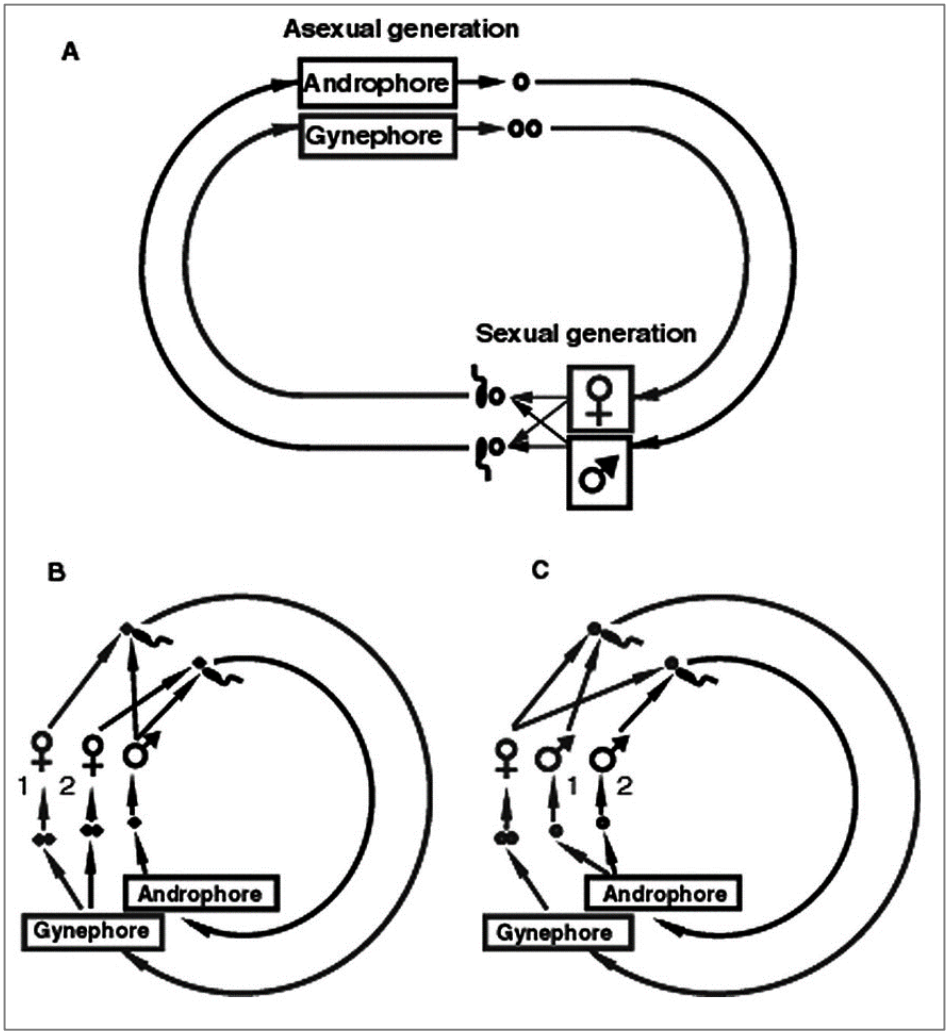


Fig. 1. Schematic presentation of cyclical parthenogenesis in heterogonic Cynipidae (A) and hypothetic mechanisms explaining the maintenance of heterogony based either on two different types of sexual generation females (B) or two different types of males (C) (Stone et al. 2002).

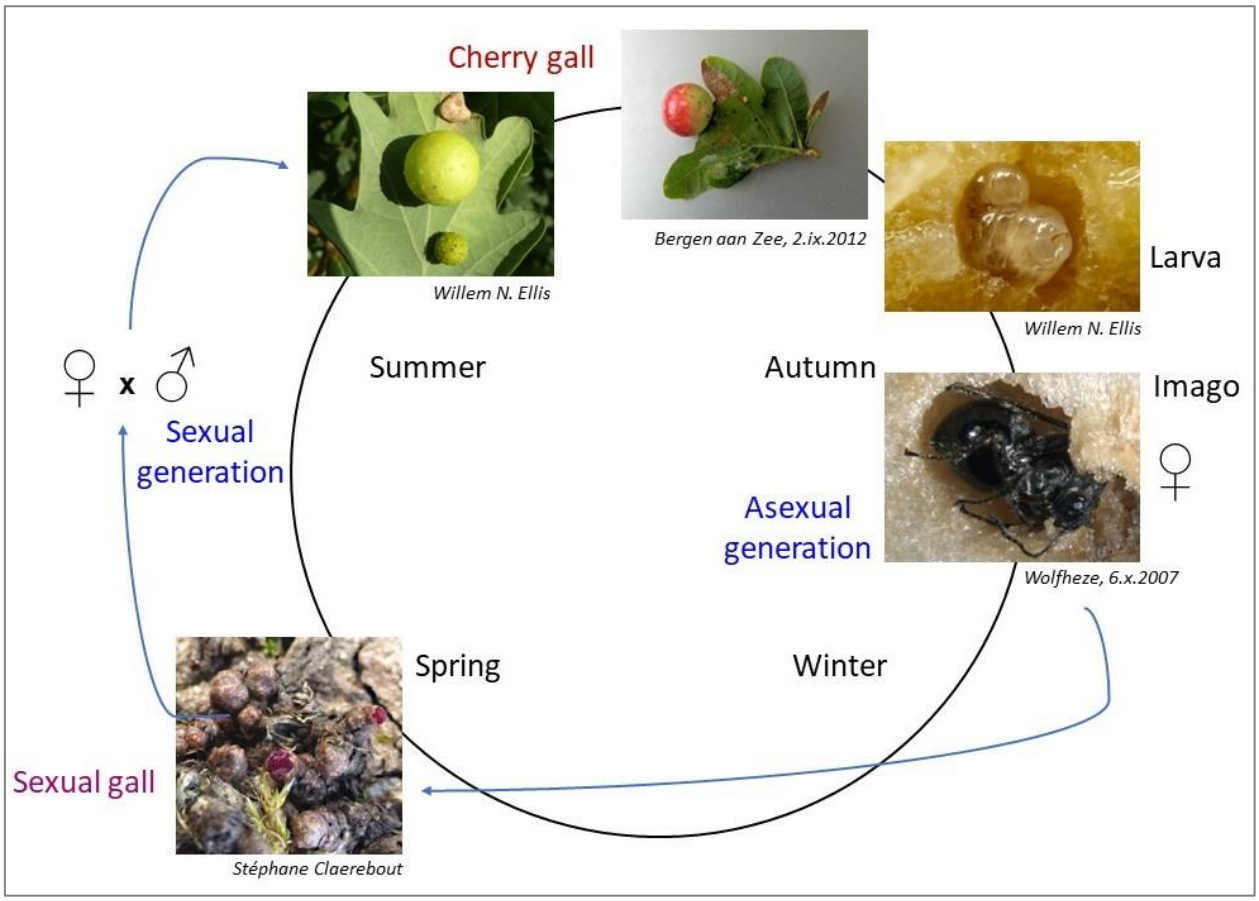


Fig. 2. *Cynips quercusfolii* life cycle (adapted from Ellis 2007).

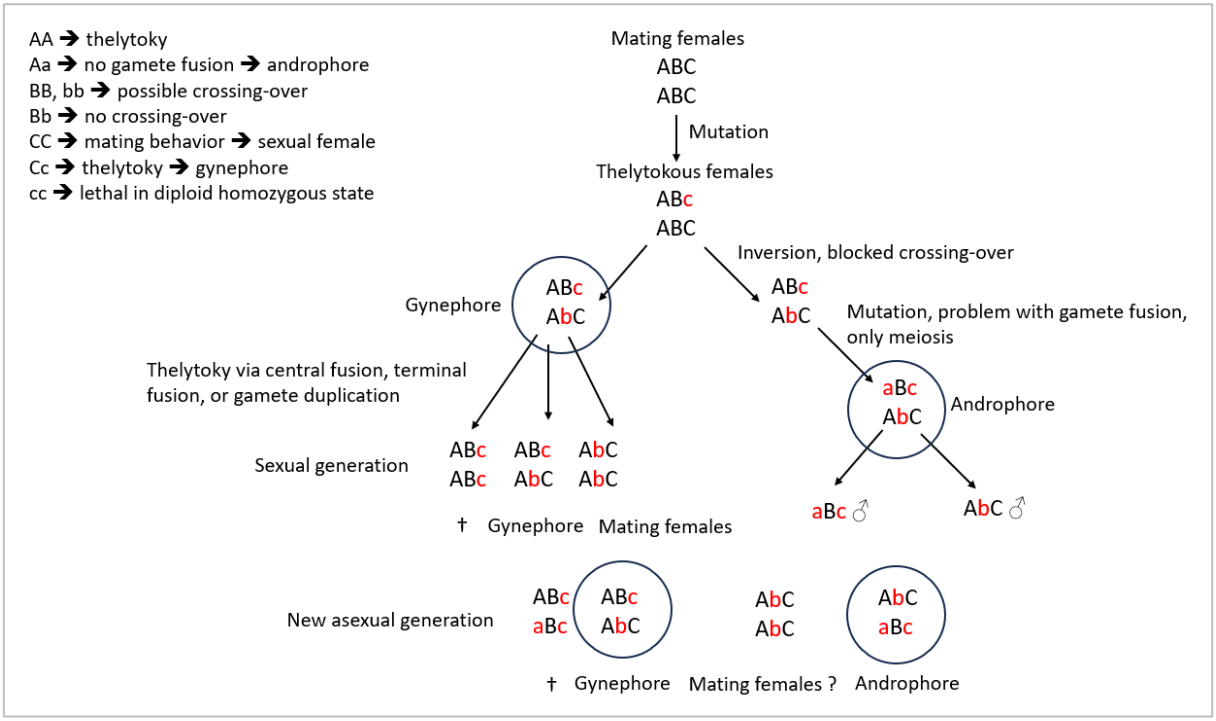


Fig. 3. Theoretical scenario proposing the origin of cyclical parthenogenesis in Cynipini. First, in an arrhenotokous population with the AABBCC female genotype, a mutation ***c*** occurred and led to the thelytokous female genotype AABBCc. The thelytokous genotype spread in the population due to asexual reproduction and the possibility of thelytokous females to mate. In summer, thelytokous genotype prevailed over the arrhenotokous genotype. Second, in a part of thelytokous females a novel mutation ***b***, i.e. an inversion blocking crossing-over in a chromosome region, occurred and spread in the population giving thelytokous females AABbCc (gynephores). Any of the types of automixis could lead to the following possible genotypes: AABBcc (supposed to be lethal), AAABbCb (gynephores), and AAbbCC (arrhenotokous females). In spring, the arrhenotokous genotype prevailed over the thelytokous genotype. Third, in a part of gynephores, a new mutation ***c*** led to a failure of gamete fusion, which led to a haploid chromosome number and the development of males aBc and AbC. Thus, the AaBbCc females became androphores. The aBc and AbC males mate with predominant arrhenotokous AAbbCC females giving the next generation of androphores and gynephores (and sexual females (sign '?' in the scheme) that could be less present compared to predominant asexually reproducing females). The inversion (Bb genotype) allowed to keep gynophore and androphore genotypes.

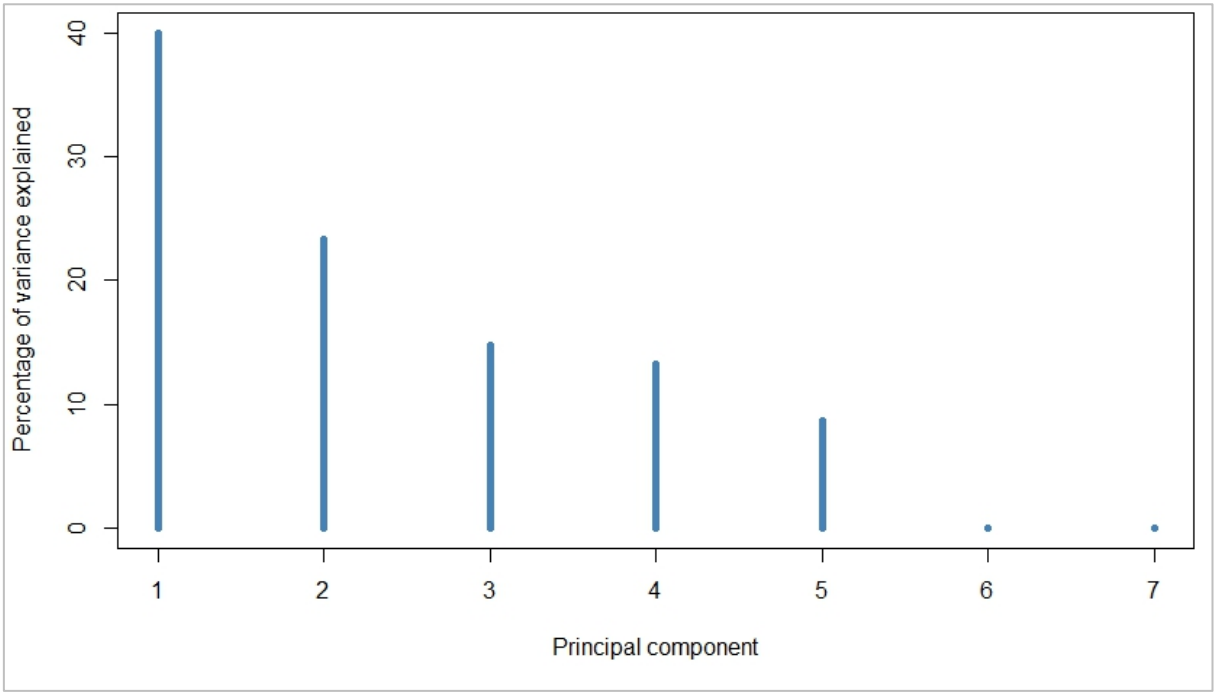


Fig. 4. Percentage of explained variance of each principal components using whole genome genotype data of *Cynips quercusfolii* .

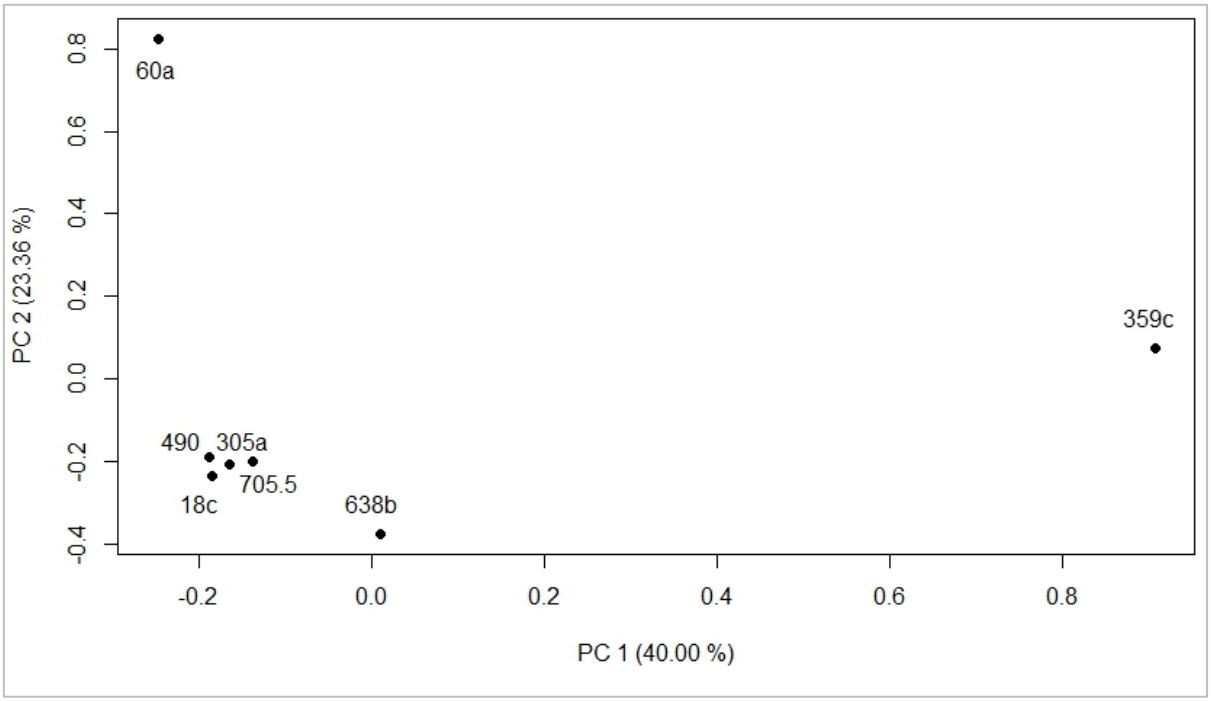


Fig. 5. Clusters of samples after performing principal component analysis of the population structure of *Cynips quercusfolii*. PC1 and PC2 represent the first two components explaining 40.00% and 23.36%, respectively. The numbers corresponding to each point represent the individual identifier.

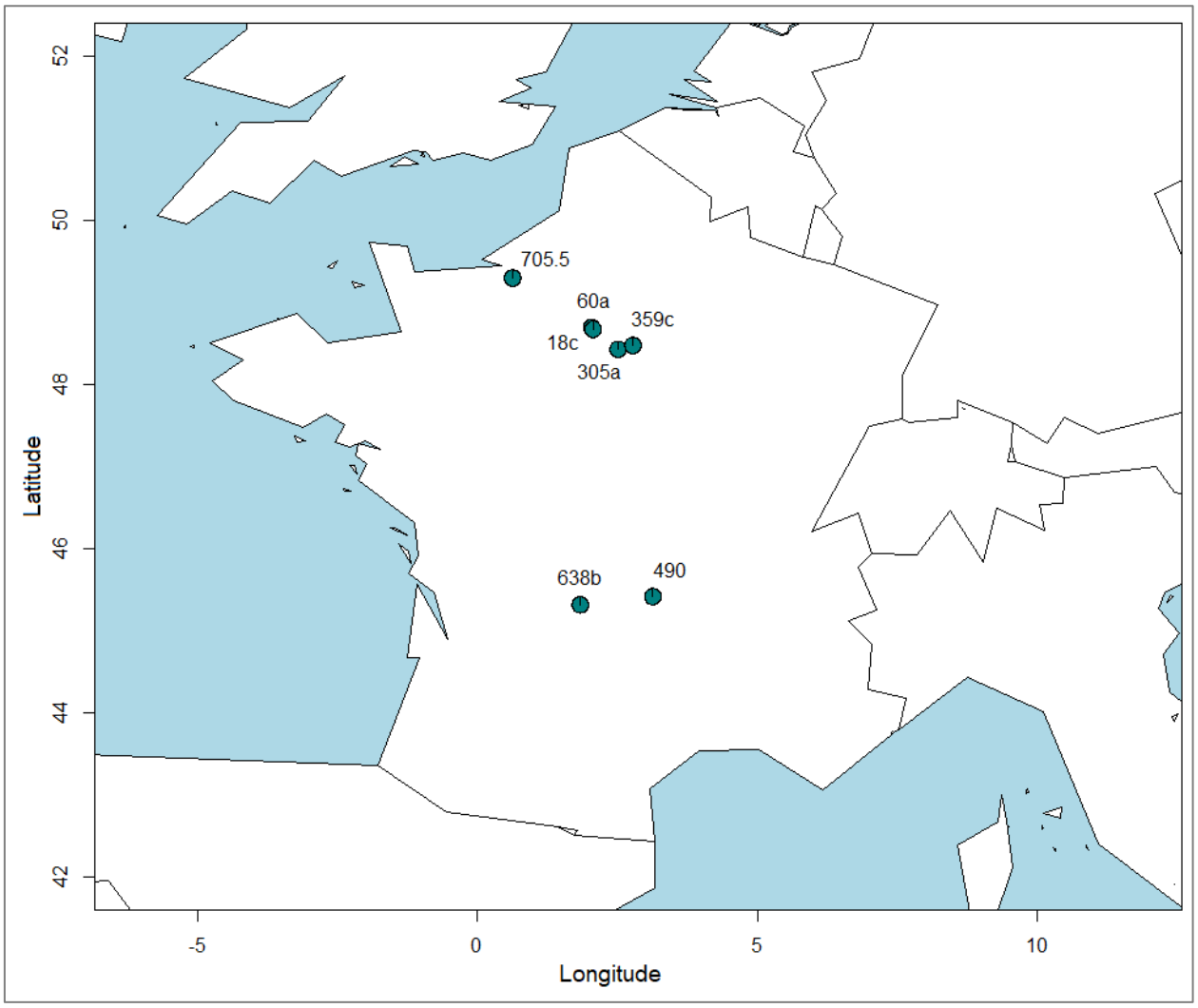


Fig. 6. **Geographical distribution of *Cynips quercusfolii* in France.** The numbers represent the individual identifier.

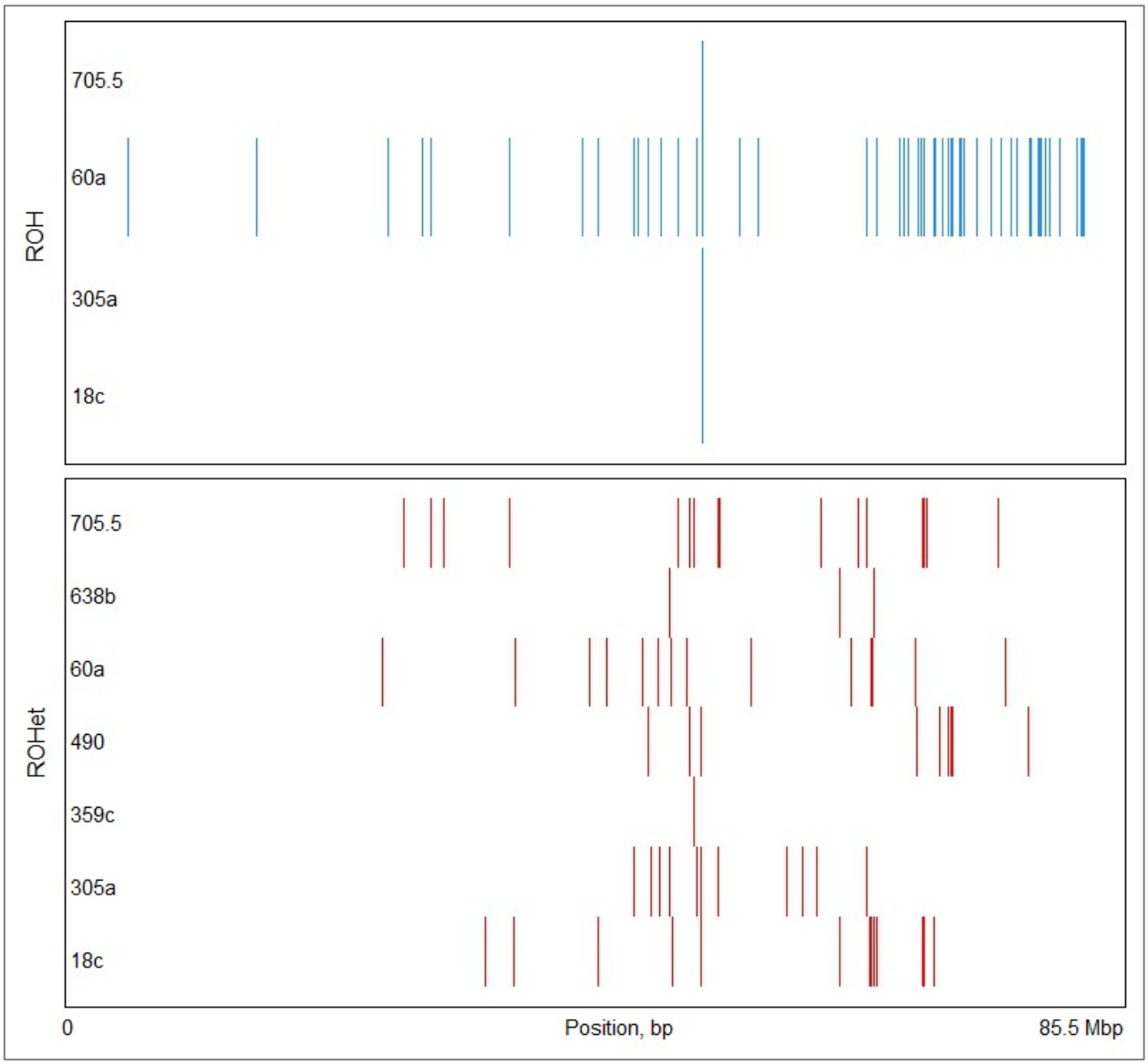
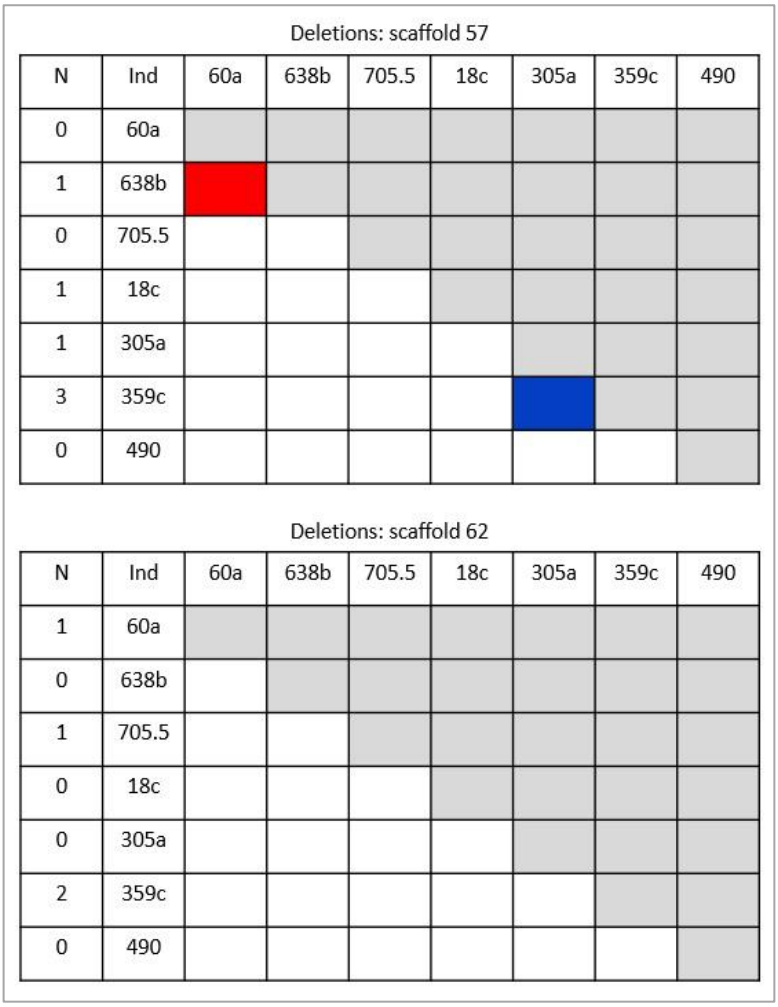


Fig. 7. Runs of homozygosity (ROH) and runs of heterozygosity (ROHet) detected in scaffold 63 (chr_63) of *Cynips quercusfolii*.



Deletions: scaffold 57

N	Ind	60a	638b	705.5	18c	305a	359c	490
0	60a							
1	638b							
0	705.5							
1	18c							
1	305a							
3	359c							
0	490							

Deletions: scaffold 62

N	Ind	60a	638b	705.5	18c	305a	359c	490
1	60a							
0	638b							
1	705.5							
0	18c							
0	305a							
2	359c							
0	490							

Fig. 8. 0.01-100 Mbp deletions detected in scaffold 57 and scaffold 62 of the assembled genome of *Cynips quercusfolii*. Ind: *C. quercusfolii* individual identifier. N: number of detected deletions in each individual. In scaffold 57, red color represents the individuals *C. quercusfolii*-638b and 60a exhibiting the same deletion, and blue color represents the individuals *C. quercusfolii*-305a and 359c showing the same deletion.

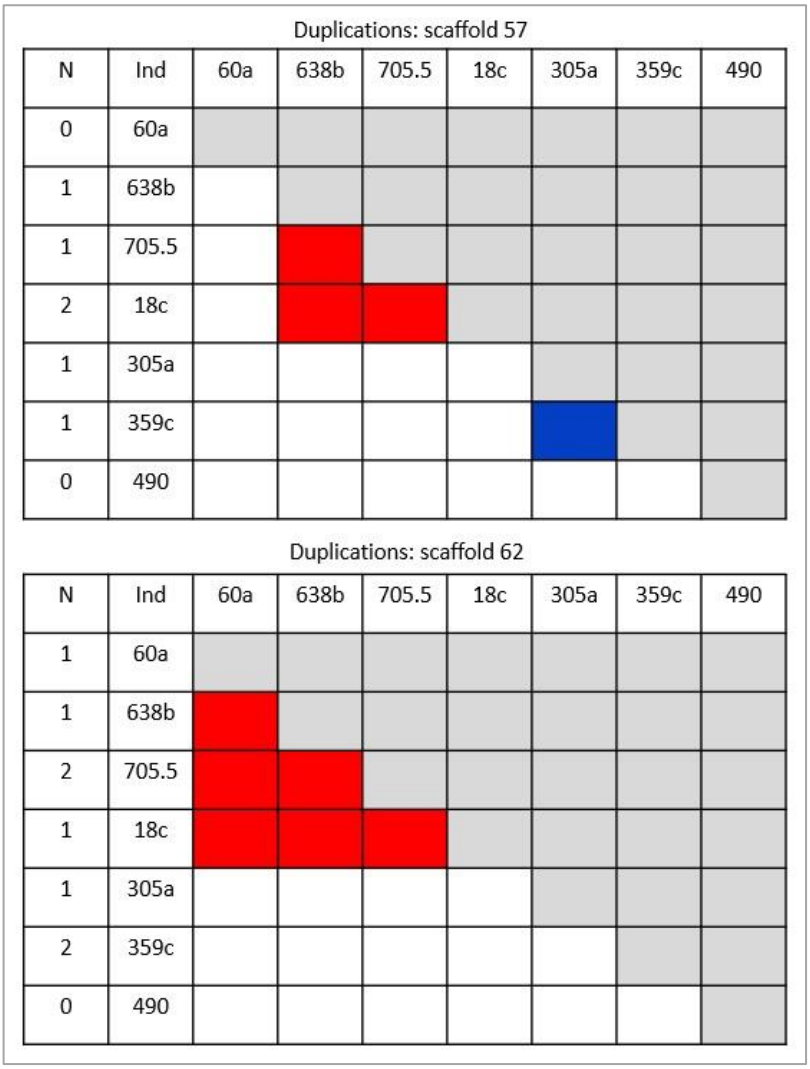


Fig. 9. 0.01-100 Mbp duplications detected in scaffold 57 and scaffold 62 of the assembled genome of *Cynips quercusfolii*. Ind: *C. quercusfolii* individual identifier. N: number of detected duplications in each individual. In scaffold 57, red color represents the individuals *C. quercusfolii*-638b, 705.5, and 18c showing the same duplication, and blue color represents the individuals *C. quercusfolii*-305a and 359c sharing the same duplication. In scaffold 62, red color represents the individuals *C. quercusfolii*-60a, 638b, 705.5, and 18c sharing the same duplication.

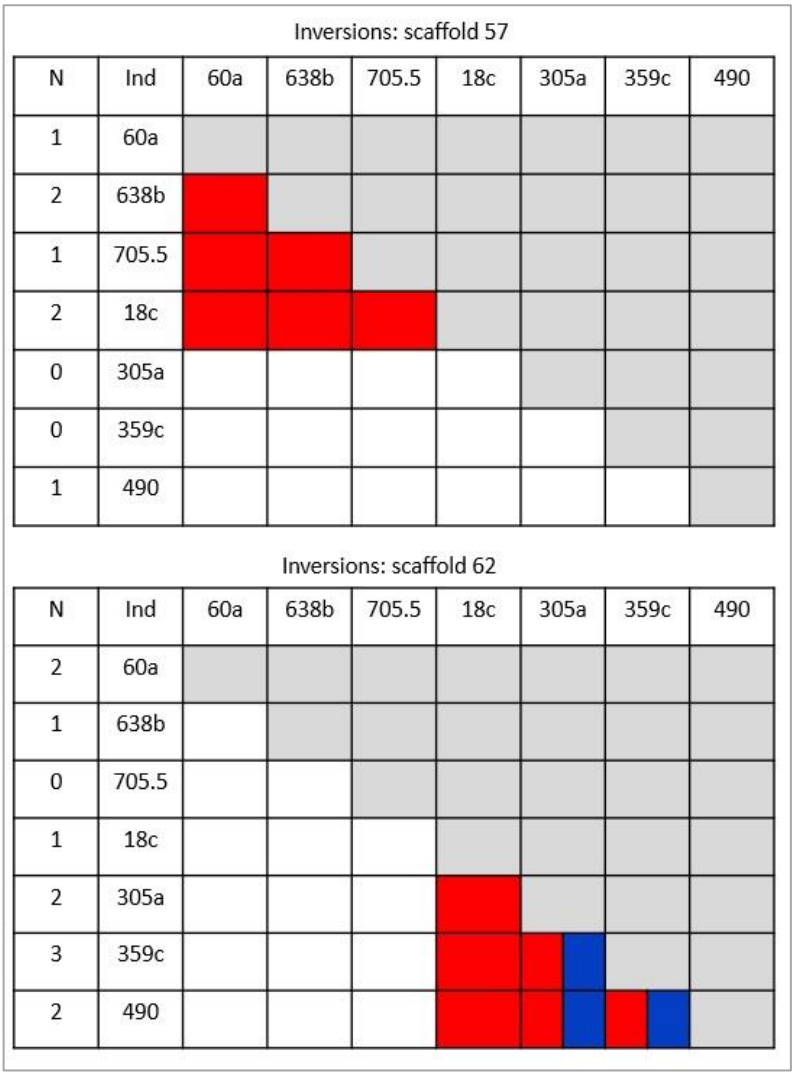


Fig. 10. 0.01-100 Mbp inversions detected in scaffold 57 and scaffold 62 of the assembled genome of *Cynips quercusfolii*. Ind: *C. quercusfolii* individual identifier. N: number of detected inversions in each individual. In scaffold 57, red color represents the individuals *C. quercusfolii*-60a, 638b, 705.5, and 18c sharing the same inversion. In scaffold 62, red color represents the individuals *C. quercusfolii*-18c, 305a, 359c, and 490 exhibiting the same inversion, and blue color indicates the individuals *C. quercusfolii*-305a, 359c, and 490 sharing the same inversion.

GENERAL DISCUSSION

Main observations

The thesis aimed to list candidate genes involved in galling in gall wasps (Hymenoptera: Cynipidae) in the context of their reproductive modes. Our strategy was to detect traces of selection in two cynipid genomes, *Diplolepis rosae* and *Cynips quercusfolii*, using population genomics and relate them to gene expression during gall formation using transcriptomics.

The most crucial step of our study was to demonstrate how reproductive modes used by Cynipidae influenced their genomes, particularly the distribution of widespread homozygous regions and the frequency of polymorphisms. We assembled genomes of *D. rosae* and *C. quercusfolli*, revealed the population structure of both species and applied various methodologies to detect signatures of selection. The mode of reproduction of both species posed a significant challenge that required us to adapt to the original objective of the study. Since asexual reproduction is the predominant reproductive mode in both cynipids, this leads to an almost complete linkage of alleles in the genome. Thus, we could not apply standard population genomics tools developed for sexually reproducing organisms to search for signatures for positive selection like selective sweeps or elevated relative frequency of divergent or polymorphic sites. Therefore, we developed the composite score approach to detect signatures for purifying and balancing/frequency-dependent selection in the genome of *D. rosae* (**CHAPTER I**). To detect traces of positive selection, we performed the McDonald–Kreitman test that is supposed not to depend on demography and mode of reproduction (McDonald and Kreitman 1991; Eyre-Walker 2002; Parsch et al. 2009) (**CHAPTER II**). Positive selection in a group of genes associated with galling was detected in the aphid *Hormaphis cornu* (Korgaonkar et al. 2021) exhibiting cyclical parthenogenesis and the cynipid *Synergus itoensis* (Gobbo et al. 2020) reproducing by arrhenotoky (Abe et al. 2011). However, in *D. rosae*, we detected only one gene encoding a transposable element, and in *C. quercusfolii,* we revealed several genes with unknown functions. In these species, detecting positive selection was challenging, probably due to a high purge of deleterious alleles along with all linked sites in highly homozygous thelytokous females. Additionally, the phenomenon of clonal interference occurring in mostly asexually reproducing populations could also to be at play (**CHAPTER II**).

That is why we identified population structure of *D. rosae* in France and performed the genome scan of several metrics, such as Fst, π, and ρ (**CHAPTER I**). We detected several regions with low differentiation and change in nucleotide diversity. Based on Fst and π, we developed the composite score (CS) summarizing both metrics: negative CS outliers (low Fst and low π) reflected negative selection, and positive CS outliers (low Fst, high π, and zero recombination) reflected balancing/frequency-dependent selection or

relaxed constraint. We revealed two highly differentiated peripatric *D. rosae* lineages that differ in the intensity of recombination and heterozygosity. Both lineages reproduce mostly by thelytokous parthenogenesis, but one the lineages exhibits a higher recombination rate and heterozygosity compared to another lineage. The most intriguing observation was that in the more recombining lineage, the group of genes related to male characteristics was under negative selection. In contrast, in the less recombining lineage, these genes were under balancing selection or relaxed constraint. Such an opposite trend in selection observed in the *D. rosae* lineages could serve as an additional illustrative example when discussing the advantages of both sexual and asexual reproduction. In the less recombining lineage ('thelytokous'), these genes were found to be under relaxed negative selection or under balancing selection. We suppose that in the thelytokous females, asexual reproduction allows the accumulation of deleterious alleles or the maintenance previously existing allelic diversity in genes related to male function. Asexual reproduction of this 'thelytokous' *D. rosae* lineage could be advantageous in local environments. To examine our hypothesis, we could assess the fine-scale *D. rosae* population structure by using several genetic markers and demonstrate multiple local 'thelytokous' lineages. In the more recombining lineage ('arrhenotokous'), the genes related to male traits demonstrated the signatures of purifying selection, which could preserve the potential for sexual reproduction. This could enable recombination and generation of genetic diversity in unfavorable biotic and abiotic conditions.

Our findings could be supported by the existence of several hymenopteran species exhibiting the same arrhenotokous-thelytokous population structure, such as *Diplolepis spinosissimae* (Plantard et al. 1998), *Asobara japonica* (Kremer et al. 2009), *Leptopilina clavipes* (Pannebakker et al. 2004a), and *Venturia canescens* (Schneider et al. 2002) (**CHAPTER I**). Among these species, *V. canescens* is the most fascinating example of how arrhenotokous and thelytokous lineages could be ecologically distinguishable. *V. canescens* is a parasitoid wasp parasitizing several lepdopteran pests, such as *Ephestia kuehniella* and *Plodia interpunctella*, that often infest bakery and granary stores. *V. canescens* strains collected in such locations representing a relatively stable environment were known to reproduce only by thelytoky (Whiting 1928; Diamond 1930; Cline et at al. 1983; Schneider et al. 2002). Simultaneously, in field conditions both arrhenotokous and thelytokous lineages were identified, with a prevalence of the sexualy reproducing form (Schneider et al. 2002). Consequently, in this species, the prevalence of one lineage over another could be associated with the advantages of one or another reproductive mode in a particular environment (Schneider et al., 2002). In *D. rosae*, among the samples collected throughout France, the more recombining ('arrhenotokous') form prevailed over the less recombining ('thelytokous'), more localized, form (**fig. 2, CHAPTER I**). Thus, taking into account the case of *V. canescens*, could we, for instance, link specific habitat

characteristics where we found only 'thelytokous' *D. rosae* lineage and differentiate them from those where the 'arrhenotokous' lineage was predominant?

The hypothesis of the importance of preserving the potential for sexual reproduction in the mostly asexually reproducing *D. rosae* could be also supported by the analysis of transcriptomic data (**CHAPTER II**). We demonstrated the overexpression of genes that could be associated with immune response in the young *D. rosae* larvae developing in the one-month (mid-July) gall. These genes encode proteins like Toll-like receptor 7, tetraspanin, venom acid phosphatase, phospholipase A2, and peroxidase. *D. rosae* could generate the immune response against parasitoids and host plant microbiome. This leads to the importance of maintaining sexual reproduction because it is a source of genetic diversity. This could provide the production of various protein isoforms in different combinations when the gall wasp is in an arms race with its natural enemies. Indeed, *D. rosae* faces a high parasitic pressure. The *D. rosae* parasitic complex comprises at least sixteen parasitoid species (Williams 2013), and the mortality rate can reach 75–100% (Stille 1984; Rizzo and Massa 2006). Furthermore, *D. rosae* is under continuous parasitoid attack throughout the summer while the gall is actively growing. One of the most widespread parasitoids of *D. rosae*, *Orthopelma mediator*, lays its eggs into the *D. rosae* larvae when the gall begins to grow. After that, another common parasitoid, *Torymus bedeguaris*, attacks *D. rosae* in later summer by laying its eggs on the surface of the host larvae (László and Tóthmérész 2013). Consequently, it could be essential for *D. rosae* to activate their defense system even at the early steps of gall formation.

Finally, the observations made by two internship students Anaïs Pourtoy and Xavier Vincent (Université Paris-Saclay, Master Biodiversité Ecologie Evolution) could also confirm our hypothesis about the maintenance of sexual reproduction in the mostly asexual *D. rosae* (**CHAPTER I, REMARK**). They showed that in the more recombining *D. rosae* lineage ('arrhenotokous'), the parasitism rate was twice less than in the less recombining ('thelytokous') lineage. Previously, Rizzo and Massa (2006) demonstrated that in the *D. rosae* populations where males comprised 15.6–21%, the parasitism rate was 34.4–35%, whereas in the populations with no males, the parasitism rate was at an average of 57.6%. Another example supporting the argument for the maintenance of sexual reproduction is the New Zealand mud snail (*Potamopyrgus antipodarum*) that inhabits freshwater and faces parasitism from trematodes like *Microphallus* sp. The snail populations consist of coexisting sexual lineages and asexually reproducing females. In such mixed populations, Jokela et al. (2009) examined the population dynamics of clonal and sexual genotypes and related it to the parasitism rate. They observed that certain clonal genotypes initially exhibited lower susceptibility to parasitism, but over time, they underwent a drastic decrease in frequency and could be replaced by other clonal

genotypes. In contrast, sexual lineages, initially were sensible to parasitism but remain relatively constant in frequency and could persist to parasitic pressure over the years.

In summary, sexual-asexual populations expand ecological niches and reduce the pressure from their natural enemies through the benefits of both reproductive strategies. Another intriguing way that helps to profit advantages of both reproductive modes in space and time is cyclical parthenogenesis that we observed in the second studied cynipid, *C. quercusfolii*. Cyclical parthenogenesis (heterogony) is a mode of reproduction when organisms obligately alternate between asexually and sexually reproducing forms (Bell 1982). The ecological benefits of this life cycle can be demonstrated in heterogonic animals, such as *Daphnia* spp. and aphids. These organisms reproduce asexually during spring and summer (favorable conditions) and produce several asexual generations. In autumn (unfavorable conditions, i.e. colder temperatures, and a shorter photoperiod), they switch to sexual reproduction and produce cold-resistant eggs that enter diapause before the next spring (Simon et al. 2002; Decaestecker et al. 2009). In aphids, the switch from one generation to another was found to be associated with a change in juvenile hormone concentration. For instance, Ishikawa et al. (2012) demonstrated that in the pea aphid (*Acyrthosiphon pisum*) females starting to produce sexual morphs in shorter photoperiods, there was a lower concentration of juvenile hormone and an upregulation of genes responsible for its degradation compared to the asexually reproducing females reared in longer photoperiods.

Similar to heterogonic crustaceans and aphids, *C. quercusfolii* and other heterogonic cynipids exhibit the well-present asexual generation, and the less present sexual generation. While cyclical parthenogenesis in Cynipini and Pediaspidini can be explained from the ecological perspective, the genetic basis underlying such strict alternation of the two generations remains completely unresolved. Because Cynipidae are partially parasitic organisms, their life cycle is synchronized with that of their host plants and relies on the physiology of parasitized host plant organs: Cynipidae are capable of inducing the galls only ovipositing in meristematic tissues. Furthermore, gall wasps have their own developmental rate, which could result in limited opportunities to generate multiple generations during the summer period. We can also consider this life cycle as an extreme case of ecological specialization of previously co-existing sexual-asexual populations (Pujade-Villar et al. 2001). Ecological specialization has been demonstrated for arrhenotokous-thelytokous populations, such as *V. canescens* (Schneider et al. 2002). Additionally, sexual and asexual generations of Cynipini often occupy different host plant tissues: sexual generation usually develops in tissues such as oak catkins and buds in spring, and asexual forms develop in leaves in summer. Consequently, we can hypothesize that cyclical parthenogenesis in Cynipini may have evolved from previously co-existing arrhenotokous ('sexual') and thelytokous ('asexual') lineages, with one or the other being seasonally predominant.

Nevertheless, while we may give ecological explanations, the genetic basis maintaining cyclical parthenogenesis in Cynipini remains unknown. We have performed the preliminary observations of the genome structure of *C. quercusfolii* and have demonstrated various structural variations in the asexual females (**CHAPTER III**). The hypothesis suggesting that chromosomal rearrangements may suppress recombination and maintain a group of specific alleles associated with the asexual female type requires further exploration. This is essential for making conclusions about the persistence of cyclical parthenogenesis in Cynipini.

What about galling genes?

In our study, we applied population genomics, transcriptomics, and the test for selection to reveal the genes under selection in *D. rosae*and *C. quercusfolii*. Population genomics showed the opposite trends in selection in genes related to sexual reproduction and neural function in the 'sexual-asexual' population of *D. rosae* (**CHAPTER I**), and different chromosomal rearrangements in asexual females within one *C. quercusfolii* population (**CHAPTER III**). Transcriptomics revealed the overexpression of genes associated with insect development and the immune response in the young *D. rosae*larvae (**CHAPTER II**). The test for selection revealed the genes encoding transposable elements and proteins with unknown functions in both *D. rosae* (**CHAPTER II**) and *C. quercusfolii* (**CHAPTER III**). However, apart from plant cell wall degrading enzymes that have been already known in Cynipidae (Hearn et al. 2019), we did not detect candidate genes that we could associate with gall induction in *D. rosae* nor in *C. quercusfolii*.

The first reason why we did not detect putative triggers of gall formation is the lack of power of the used tools and methods. Population genomics tools to investigate predominantly asexually reproducing organisms are often limited, which rendered the study of such organisms challenging. For instance, we could not use the tools to detect the patterns of positive selection (e.g. selective sweeps) even in the more recombining *D. rosae* lineage because of limited recombination and widespread linkage of alleles. We tried to apply a method based on polymorphisms to detect selective sweeps in the *D. rosae* genome. We expected positive selection in the regions where polymorphism, represented as the site frequency spectrum, was significantly different from the genome background. However, detected signals were false positives: they reflected either balancing selection or a point of recombination rather than positive selection. In addition, many genes were not functionally annotated, which also led to the lack of information. Indeed, the majority of genes had no gene ontology annotations nor cluster of orthologous group annotations. That is why we could not perform the enrichment analyses and reveal the functional groups of genes. Finally, we met the challenges when performing a greenhouse experiment for transcriptome analysis. Even if the gall wasp

stung the plant and we could catch this moment, the galls did not develop. Thus, we did not analyze the gall wasp transcriptome at the earliest moment of gall induction.

The second reason is that positive selection within the gall wasp-host plant system may have occurred in the past. A new beneficial allele of a gene involved in gall induction had been fixed in a population, and positive selection turned into purifying selection that we observe today (Bazykin and Kondrashov 2011). It is also possible that there is no intense coevolution between gall wasp species and their host plants. For instance, Stille (1985) and Kohnen et al. (2011) demonstrated that there was no correlation between the *D. rosae* genotype and the host plant genotype: one *D. rosae* genotype could be sampled in several *Rosa* spp. genotypes and one host plant genotype could be parasitized by different gall wasp genotypes. In addition, gall wasps typically have a limited impact on the population dynamics of their host plants. Occasional damage was usually detected when studying introduced or invasive gall wasp species, such as the Asian chestnut gall wasp (*Dryocosmus kuriphilus*) (Giertych et al. 2013; Avtzis et al. 2019). However, in the native gall wasp-host plant system, when analyzing 15 oak trees, Giertych et al. (2013) found that only 0.05–1.56% of leaves were parasitized by *C. quercusfolii*. In our study, we successfully sampled *D. rosae* galls from wild roses growing near roads or in urban areas, but it was challenging to find galls in forests. We concluded that gall wasps could successfully parasitize weaker shrubs that are under pressure from anthropogenic factors, while shrubs growing in relatively wild habitats could resist attacks by gall wasps. Nonetheless, gall wasps are likely under intense coevolution with their natural enemies. As discussed above, can be attacked by tens of parasitoid species (Williams 2013; Forbes et al. 2016), and the mortality rate can reach 75–100% (Stille 1984; Rizzo and Massa 2006). Thus, gall characteristics and/or gall wasp behavior could be optimized in response to selective pressure from parasitoids and predators. Parasitism rate and predation rate have been demonstrated to be the factors determining the optimal gall size in *D. rosae* (Laszlo et al. 2014). The authors showed a negative correlation between gall size and parasitism rate: smaller galls were parasitized more than larger galls. Conversely, they revealed a positive correlation between gall size and predation rate: vertebrate predators preferred larger, more visible galls. The authors concluded that the optimal gall size could be determined by the opposite trends of directional selection from both factors.

As a result, we may assume that selection acts on the location of galls and the moment of oviposition. A more favorable location could help to avoid parasitism and predation and produce more nutritious galls for developing larvae. Giertych et al. (2013) demonstrated that *C. quercusfolii* preferred to oviposit on leaves with a large surface area, suggesting that larger leaves might provide more nutrients for developing galls. The placement of galls (e.g. primary/secondary branch, position on the leaves, choice of leaf size, and distance from the soil) is likely determined by gall wasp behavior. Consequently, these traits are likely associated with genes involved in neural function. Genes linked to

behavioral traits could be under positive or balancing selection, while genes involved in gall induction could be conserved within one clade. This hypothesis was developed thanks to the findings of Shorthouse et al. (2005): several *Diplolepis* spp. oviposited in different tissues of rose leaf buds. *D. polita* and *D. bicolor* stung leaflets within leaf buds; *D. nodulosa*, *D. triforma*, and *D. spinosa* oviposited between leaf primordia; *D. fusiformans* attacked the surface of stems. Hence, gall morphology may depend on the location and type of meristematic plant tissue, rather than solely on the initial trigger produced by gall wasps. Egg or larval secretions damage plant cells, and diverse metabolites resulting from plant tissue lysis may trigger different signal pathways. Consequently, these various signals can initiate multiple cell differentiation pathways within different plant tissues, potentially leading to such diverse gall morphologies.

Lastly, we could also examine the potential role of transposable elements (TE) in gall induction. Indeed, gall wasp genomes are rich in repetitive sequences, including various TEs. We estimated that 69% of the *D. rosae* genome was presented by repetitive elements (**CHAPTER I**), making it one of the insect species with the highest proportion of TEs in its genome (Petersen et al. 2019). We can hypothesize that the functioning and the expression of genes involved in galling may change due to TE transpositions notably in regulatory regions. As a result, this could lead to the production of different signal proteins, the generation of different signal pathways, and, consequently, diverse gall morphologies. Additionally, it would be also interesting to examine whether the TE content correlates with the various lifestyles (galler, inquiline, and parasitoid) within Cynipidae s. lat. and study the role of specific TE groups.

Deeper understanding of cynipid natural history: additional steps we could take

The results presented in the thesis introduce new questions that we could explore to advance our understanding of cynipid natural history. The following tasks concern more detailed analyses of the population structure of the studied cynipids and investigation of other gene groups potentially associated with galling.

Firstly, we could examine whether anatomical characteristics of galls (e.g. gall size and gall wall thickness) and parasitic complexes correlate with the reproductive strategies within the *D. rosae* populations. We could expect that, for instance, the more recombining D. rosae lineage is affected by a smaller number of parasitoid species and/or demonstrates thicker gall walls. In contrast, the less recombining lineage induces larger galls with more gall cells and/or produces more eggs. It could demonstrate the advantages of maintaining the 'sexual-asexual' population. The more recombining lineage would tend to relax parasitic pressure and the less recombining lineage would tend to produce more offspring without a two-fold cost of sex. Furthermore, we could also determine the fine-scale population structure of *D. rosae* using several genetic

markers. It could precise the geographic distribution of the *D. rosae* lineages and demonstrate it there are 'sexual-asexual' strains adapted to different local conditions.

Secondly, it would be valuable examine repetitive sequences and genes acquired through horizontal gene transfer, such as those encoding plant cell wall-degrading enzymes, to determine if there is a correlation with cynipid lifestyles. We could expect that, for example, genomes of gall-inducing cynipids contain specific genes or transposable elements (TE) absent in both inquilines and parasitoids. This could show new candidate genes involved in galling. In the case of specific TEs detected in gall-inducing cynipids, we could search for their positions in the genome and suppose what genes they could affect.

Besides that, it would be interesting to analyze the expression dynamics of genes encoding calcium-binding proteins, ATPases, and peroxidases. These molecules could play a role in the initial steps of gall formation by damaging danger signal molecules produced by the host plant due to herbivore attacks (Giron et al. 2016).

Furthermore, one should find a way to improve gene annotations and perform a robust gene enrichment analysis. Accurately identified functional gene groups could be associated with galling and show other means that gall wasps may use to manipulate host plant metabolism. For instance, Martinson et al. (2022) performed the enrichment analysis of the genes expressed in the oak parasitized by *Dryocosmus quercuspalustris* and revealed gene ontology terms related to gene silencing, RNA methylation, RNA modification, and mRNA transport in the inner gall tissue. These findings led to the hypothesis that gall induction may involve RNA interference, the process wherein small regulatory RNAs from gall wasps may alter the expression of plant genes by modifying transcription and translation.

In gall wasp – host plant interaction, it is essential to consider how parasitism impacts plant cell differentiation. We could hypothesize that gall morphology is primarily determined by the type and location of the affected meristematic plant tissue rather than a specific effector produced by the gall wasp. For instance, gall tissues are known to have similar anatomy and nutritive content to fruits and seeds (Schönrogge et al. 2000). Consequently, we can search candidate genes or even entire signaling pathways in the existing literature that are responsible for cell differentiation leading to fruit or seed formation. By investigating whether these genes are up-regulated during the development of galls, we can gain insights into whether gall wasps employ or manipulate regulatory molecules similar to those involved in plant cell differentiation processes.

We have to identify the factors that did not allow the greenhouse experiment organized for analyzing of the *D. rosae* transcriptome immediately after oviposition (female venom gland and egg) and within several hours and days (first larval instar). We need to establish a laboratory rearing of *D. rosae*, i.e. create conditions in a greenhouse where *D. rosae* females can induce galls on *R. canina* shrubs.

Finally, it is essential to analyze a complete and less fragmented genome assembly of *C. quercusfolii* and improve alignment quality. Sampling and sequencing of additional asexual females could confirm whether the identified structural variations are indicators of population structure rather than alignment artifacts. Cytological analysis of asexual females could reveal the type of thelytokous parthenogenesis and demonstrate if the female type (gynephore or androphore) is associated with the phenomena like aneuploidy (abnormal number of chromosomes like 2n+1 or 2n-1) or B chromosomes (additional, but non-essential for a species' survival chromosome). It is also important to sequence the sexual generation of *C. quercusfolii*. It would show whether balancing selection in genes found in the inversions acts within males or females. Therefore, we could conclude whether the type of asexual female (androphore or gynephore) depends on male or female genotype of sexual generation. Thirdly, one should annotate genes found in the inverted genome regions to conclude whether their function can be associated with the maintenance of the life cycle in heterogonic Cynipidae.

CONCLUSION

Gall wasps are fascinating organisms that use unknown molecular tools to manipulate host plant metabolism and generate a great variety of gall morphologies. While our study may not have unveiled clear candidate genes encoding potential molecular triggers of gall formation, it makes a contribution to the broader field of cynipid research. We have demonstrated the genome and population structure of the two cynipid wasp species exhibiting different life cycles. Furthermore, our findings regarding the contrasting selective processes acting in sexual-asexual cynipid populations can be valuable within the context of one of the fundamental questions in evolutionary biology: what are the evolutionary advantages of sexual reproduction that offset the two-fold cost? Finally, we have addressed the question of the genetic basis underlying thelytokous reproduction and the maintenance of a bivoltine life cycle in Cynipidae.

While significant progress has been made in gall wasp research, particularly thanks to the ongoing advancements in molecular biology and bioinformatics tools, there is still substantial work remaining for a deeper understanding of cynipid natural history. The questions of the molecular mechanisms behind gall induction, diversity of gall morphologies, selective forces driving gall wasp evolution, and factors underlying the diversity of cynipid life cycles remain relevant.

To conclude my thesis, I would like to reference the words of H. Hagen from *The Canadian Entomologist* (1878), which are still applicable to contemporary cynipid studies:

"*The natural history of the interesting gall insects is still somewhat mysterious. A large number of observations have been made here and in Europe by prominent Entomologists; nevertheless, a careful study of the most detailed papers always gives the impression that something is still wanting to explain the various facts related by the authors*"

REFERENCES

Abe, Y. (1986) Taxonomic status of the *Andricus mukaigawae* complex and its speciation with geographic parthenogenesis (Hymenoptera: Cynipidae). *Applied entomology and Zoology*, 21(3), 436–447. Available from: https://doi.org/10.1303/aez.21.436.

Abe, Y., Ide, T., & Wachi, N. (2011). Discovery of a new gall-inducing species in the inquiline tribe Synergini (Hymenoptera: Cynipidae): inconsistent implications from biology and morphology. *Annals of the entomological society of America*, 104(2), 115-120. Available from: https://doi.org/10.1603/AN10149.

Akpınar, M. A., Görgün, S., Gençer, L., & Aktümsek, A. (2017) Fatty acid composition of *Diplolepis fructuum* (Rübsaamen, 1895) (Hymenoptera: Cynipidae) during its developmental stages. *Journal of the Entomological Research Society*, 19(3), 109–118. Available from: https://entomol.org/journal/index.php/JERS/article/view/1372.

Alexa, A., Rahnenfuhrer, J. (2022) *topGO: Enrichment Analysis for Gene Ontology. R package version 2.50.0*. Available at: https://bioconductor.org/packages/release/bioc/html/topGO.html [Accessed 23rd September 2023].

Alonge, M., Lebeigle, L., Kirsche, M., Jenike, K., Ou, S., Aganezov, S., ... & Soyk, S. (2022) Automated assembly scaffolding using RagTag elevates a new tomato system for high-throughput genome editing. *Genome biology*, 23(1), 1–19. Available from: https://doi.org/10.1186/s13059-022-02823-7.

Andrews, S. (2010) *FastQC: a quality control tool for high throughput sequence data*. Available at: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/ [Accessed 23rd September 2023].

Asgari, S., Zhang, G., Zareie, R., & Schmidt, O. (2003) A serine proteinase homolog venom protein from an endoparasitoid wasp inhibits melanization of the host hemolymph. *Insect Biochemistry and Molecular Biology*, 33(10), 1017–1024. Available from: https://doi.org/10.1016/S0965-1748(03)00116-4.

Atkinson, R. J. (2000) *The genetic analysis of natural history, reproductive strategy, and population structure in European oak gall wasps (Hymenoptera: Cynipidae) (Doctoral dissertation)*. University of Oxford.

Avtzis, D. N., Melika, G., Matošević, D., & Coyle, D. R. (2019). The Asian chestnut gall wasp *Dryocosmus kuriphilus*: a global invader and a successful case of classical biological control. *Journal of pest science*, 92, 107–115. Available from: https://doi.org/10.1007/s10340-018-1046-1.

Bazykin, G. A., & Kondrashov, A. S. (2011). Detecting past positive selection through ongoing negative selection. *Genome Biology and Evolution*, 3, 1006–1013. Available from: https://doi.org/10.1093/gbe/evr086.

Bailey, N. W. (2012) Evolutionary models of extended phenotypes. *Trends in ecology & evolution*, 27(10), 561–569. Available from: https://doi.org/10.1016/j.tree.2012.05.011.

Bailey, S. F., & Stange, L. A. (1966) The twig wasp of Cork Oak—Its biology and control. *Journal of Economic Entomology*, 59(3), 663–668. Available from: https://doi.org/10.1093/jee/59.3.663.

Barnosky, A. D. (1999) Does evolution dance to the Red Queen or the Court Jester. *Journal of Vertebrate Paleontology*, 19, 31A. Available from: https://doi.org/10.1080/02724634.1999.10011202.

Bartlett, L., & Connor, E. F. (2014) Exogenous phytohormones and the induction of plant galls by insects. *Arthropod-Plant Interactions*, 8, 339–348. Available from: https://doi.org/10.1007/s11829-014-9309-0.

Bell, G. (1982) *The paradox of sexuality. The masterpiece of nature: The evolution and genetics of sexuality*, 160–331. Berkeley: University of California Press.

Beukeboom, L. W., & Pijnacker, L. P. (2000). Automictic parthenogenesis in the parasitoid *Venturia canescens* (Hymenoptera: Ichneumonidae) revisited. *Genome*, *43*(6), 939–944. Available from: https://doi.org/10.1139/g00-06.

Biscarini, F, Cozzi, P, Gaspa, G, and Marras, G. (2018) Detectruns: An R Package to Detect Runs of Homozygosity and Heterozygosity in Diploid Genomes. CRAN (The Comprehensive R Archive Network). Available at: https://orca.cardiff.ac.uk/108906/ [Accessed 23rd September 2023].

Blaimer, B. B., Gotzek, D., Brady, S. G., & Buffington, M. L. (2020). Comprehensive phylogenomic analyses re-write the evolution of parasitism within cynipoid wasps. *BMC Evolutionary Biology*, 20, 1-22. Available from: https://doi.org/10.1186/s12862-020-01716-2.

Brandão-Dias, P. F., Zhang, Y. M., Pirro, S., Vinson, C. C., Weinersmith, K. L., Ward, A. K., ... & Egan, S. P. (2022) Describing biodiversity in the genomics era: A new species of Nearctic Cynipidae gall wasp and its genome. *Systematic Entomology*, 47(1), 94–112. Available from: https://doi.org/10.1111/syen.12521.

Bronner, R. (1985) Anatomy of the ovipositor and oviposition behavior of the gall wasp *Diplolepis rosae* (Hymenoptera: Cynipidae). *The Canadian Entomologist*, 117(7), 849–858. Available from: https://doi.org/10.4039/Ent117849-7.

Brose, K., Bland, K. S., Wang, K. H., Arnott, D., Henzel, W., Goodman, C. S., ... & Kidd, T. (1999) Slit proteins bind Robo receptors and have an evolutionarily conserved role in repulsive axon guidance. *Cell*, 96(6), 795–806. Available from: https://doi.org/10.1016/S0092-8674(00)80590-5.

Browning, B. L., Tian, X., Zhou, Y., & Browning, S. R. (2021) Fast two-stage phasing of large-scale sequence data. *The American Journal of Human Genetics*, 108(10), 1880–1890. Available from: https://doi.org/10.1016/j.ajhg.2021.08.005.

Bruna, T., Hoff, K.J., Lomsadze, A., Stanke, M., & Borodovsky, M. (2021) BRAKER2: Automatic Eukaryotic Genome Annotation with GeneMark-EP+ and AUGUSTUS Supported by a Protein Database. *NAR Genomics and Bioinformatics,* 3(1):lqaa108. Available from: https://doi.org/10.1093/nargab/lqaa108.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009) BLAST+: architecture and applications. *BMC bioinformatics*, 10, 1–9. Available from: https://doi.org/10.1186/1471-2105-10-421.

Cambier, S., Ginis, O., Moreau, S. J., Gayral, P., Hearn, J., Stone, G. N., ... & Drezen, J. M. (2019) Gall wasp transcriptomes unravel potential effectors involved in molecular dialogues with oak and rose. *Frontiers in physiology*, 10, 926. Available from: https://doi.org/10.3389/fphys.2019.00926.

Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P., & Huerta-Cepas, J. (2021) eggNOG-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale. *Molecular biology and evolution*, 38(12), 5825–5829. Available from: https://doi.org/10.1093/molbev.

Ceballos, F. C., Hazelhurst, S., & Ramsay, M. (2018) Assessing runs of Homozygosity: a comparison of SNP Array and whole genome sequence low coverage data. *BMC genomics*, 19, 1–12. Available from: . https://doi.org/10.1186/s12864-018-4489-0.

Chikina, M., Robinson, J. D., & Clark, N. L. (2016) Hundreds of genes experienced convergent shifts in selective pressure in marine mammals. *Molecular biology and evolution*, 33(9), 2182–2192. Available from: https://doi.org/10.1093/molbev/msw112.

Cline, L. D., Flaherty, B. R., & Press, J. W. (1983). Response of selected parasitoids and predators of stored-product insects to whitelight or blacklight traps. *Journal of Economic Entomology*, 76(2), 298–301. Available from: https://doi.org/10.1093/jee/76.2.298.

Colinet, D., Dubuffet, A., Cazes, D., Moreau, S., Drezen, J. M., & Poirié, M. (2009) A serpin from the parasitoid wasp *Leptopilina boulardi* targets the *Drosophila* phenoloxidase cascade. *Developmental & Comparative Immunology*, 33(5), 681–689. Available from: https://doi.org/10.1016/j.dci.2008.11.013.

Colinet, D., Mathe-Hubert, H., Allemand, R., Gatti, J. L., & Poirié, M. (2013) Variability of venom components in immune suppressive parasitoid wasps: from a phylogenetic to a population approach. *Journal of Insect Physiology*, 59(2), 205–212. Available from: https://doi.org/10.1016/j.jinsphys.2012.10.013.

Cooper, W. R., & Rieske, L. K. (2010) Gall structure affects ecological associations of *Dryocosmus kuriphilus* (Hymenoptera: Cynipidae). *Environmental Entomology*, 39(3), 787–797. Available from: https://doi.org/10.1603/EN09382.

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., ... & 1000 Genomes Project Analysis Group. (2011) The variant call format and VCFtools. *Bioinformatics*, 27(15), 2156–2158. Available from: https://doi.org/10.1093/bioinformatics/btr330.

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., ... & Li, H. (2021) Twelve years of SAMtools and BCFtools. *Gigascience*, 10(2), giab008. Available from: https://doi.org/10.1093/gigascience/giab008.

Dani, M. P., Edwards, J. P., & Richards, E. H. (2005) Hydrolase activity in the venom of the pupal endoparasitic wasp, *Pimpla hypochondriaca*. *Comparative Biochemistry and Physiology Part B: Biochemistry and Molecular Biology*, 141(3), 373–381. Available from: https://doi.org/10.1016/j.cbpc.2005.04.010.

Dawkins, R. (1982) *The extended phenotype: The gene as the unit of selection*. Oxford, UK: Oxford University Press.

Dawkins, R. (2004) Extended phenotype–but not too extended. A reply to Laland, Turner and Jablonka. *Biology and Philosophy*, 19, 377–396. Available from: https://doi.org/10.1023/B:BIPH.0000036180.14904.96.

De Lorenzo, G., Ferrari, S., Giovannoni, M., Mattei, B., & Cervone, F. (2019) Cell wall traits that influence plant development, immunity, and bioconversion. *The Plant Journal*, 97(1), 134–147. Available from: https://doi.org/10.1111/tpj.14196.

Decaestecker, E., De Meester, L., Mergeay, J. (2009). Cyclical Parthenogenesis in *Daphnia*: Sexual Versus Asexual Reproduction. In: Schön, I., Martens, K., Dijk, P. (eds) Lost Sex. Springer, Dordrecht. pp. 295–316. Available from: https://doi.org/10.1007/978-90-481-2770-2_15.

Diamond, V. R. (1930). The Biology of *Nemeritis canescens*, a Parasite of the Mediterranean Flour Moth. *60th Annual Report of the Entomological Society of Ontario 1929.* pp. 84–89.

Dinç, S. (2017) *Phylogenetic analysis and phylogeographic structure of Cynips quercusfolii (Hymenoptera: Cynipidae) as inferred from mtDNA and nDNA sequences (Doctoral dissertation)*. Abant Izzet Baysal University: Graduate School of Natural and Applied Sciences, Department of Biology.

Dobens, L. L., & Raftery, L. A. (2000) Integration of epithelial patterning and morphogenesis in *Drosophila* ovarian follicle cells. *Developmental dynamics: an official publication of the American Association of Anatomists*, 218(1), 80–93. Available from: https://doi.org/10.1002/(SICI)1097-0177(200005)218:1<80::AID-DVDY7>3.0.CO;2-8.

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., ... & Gingeras, T. R. (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1), 15–21. Available from: https://doi.org/10.1093/bioinformatics/bts635.

Dorémus, T., Urbach, S., Jouan, V., Cousserans, F., Ravallec, M., Demettre, E., ... & Volkoff, A. N. (2013) Venom gland extract is not required for successful parasitism in the polydnavirus-associated endoparasitoid *Hyposoter didymator* (Hym. Ichneumonidae) despite the presence of numerous novel and conserved venom proteins. *Insect Biochemistry and Molecular Biology*, 43(3), 292–307. Available from: https://doi.org/10.1016/j.ibmb.2012.12.010.

Eilertson, K. E., Booth, J. G., & Bustamante, C. D. (2012) SnIPRE: selection inference using a Poisson random effects model. *PLoS computational biology*, 8(12), e1002806. Available from: https://doi.org/10.1371/journal.pcbi.1002806.

Ellis, W. N. (2007) *Leafminers and plant galls of Europe*. *Cynips quercusfolii.* Available at: https://bladmineerders.nl/ [Accessed 23rd September 2023].

Eyre-Walker, A. (2002). Changing effective population size and the McDonald-Kreitman test. *Genetics*, 162(4), 2017-2024. Availbale from: https://doi.org/10.1093/genetics/162.4.2017.

Finn, R. D., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Mistry, J., Mitchell, A. L., ... & Bateman, A. (2016) The Pfam protein families database: towards a more sustainable future. *Nucleic acids research*, 44(D1), D279–D285. Available from: https://doi.org/10.1093/nar/gkr1065.

Fjøsne, T. F., Stenseth, E. B., Myromslien, F., & Rudi, K. (2015) Gene expression of TLR homologues identified by genome-wide screening of the earthworm Dendrobaena veneta. *Innate Immunity*, 21(2), 161–166. Available from: https://doi.org/10.1177/1753425914523056.

Folliot, R. (1964) *Contribution à l'étude de la biologie des Cynipides gallicoles (Hyménoptères, Cynipoidea) (Doctoral dissertation)*. Université de Rennes: Faculté des Sciences.

Forbes, A. A., Hall, M. C., Lund, J., Hood, G. R., Izen, R., Egan, S. P., & Ott, J. R. (2016) Parasitoids, hyperparasitoids, and inquilines associated with the sexual and asexual generations of the gall former, *Belonocnema treatae* (Hymenoptera: Cynipidae). *Annals of the Entomological Society of America*, 109(1), 49–63. Available from: https://doi.org/10.1093/aesa/sav112.

Gätjens-Boniche, O. (2019) The mechanism of plant gall induction by insects: Revealing clues, facts, and consequences in a cross-kingdom complex interaction. *Revista de Biología Tropical*, 67(6), 1359–1382. Available from: http://dx.doi.org/10.15517/rbt.v67i6.33984.

Giertych, M. J., Jagodzinski, A. M., & Karolewski, P. (2013) Spatial distribution of *Cynips quercusfolii* (Hymenoptera: Cynipidae) galls on leaves and within the crowns of oak trees. European *Journal of Entomology*, 110(4), 657–661. Available from: http://www.eje.cz/pdfs/110/4/657.

Gil-Tapetado, D., Castedo-Dorado, F., Nieves-Aldrey, J. L., & Lombardero, M. J. (2021) Gall size of *Dryocosmus kuriphilus* limits down-regulation by native parasitoids. *Biological Invasions*, 23(4), 1157–1174. Available from: https://doi.org/10.1007/s10530-020-02427-x.

Giron, D., Huguet, E., Stone, G. N., & Body, M. (2016) Insect-induced effects on plants and possible effectors used by galling and leaf-mining insects to manipulate their host-plant. *Journal of Insect Physiology*, 84, 70–89. Available from: https://doi.org/10.1016/j.jinsphys.2015.12.009.

Gobbo, E. (2022) *Gall induction in gall wasps (Cynipidae s. lat.): Insights from comparative genomics (Doctoral dissertation)*. Stockholm University: Department of Zoology.

Gobbo, E., Lartillot, N., Hearn, J., Stone, G. N., Abe, Y., Wheat, C. W., ... & Ronquist, F. (2020) From inquilines to gall inducers: Genomic signature of a lifestyle transition in Synergus gall wasps. *Genome biology and evolution*, 12(11), 2060–2073. Available from: https://doi.org/10.1093/gbe/evaa204.

Guiguet, A., Dubreuil, G., Harris, M. O., Appel, H. M., Schultz, J. C., Pereira, M. H., & Giron, D. (2016) Shared weapons of blood-and plant-feeding insects: surprising commonalities for manipulating hosts. *Journal of Insect Physiology*, 84, 4–21. Available from: https://doi.org/10.1016/j.jinsphys.2015.12.006.

Hagen, H. (1878). On the natural history of gall insects. *The Canadian Entomologist*, 10(5), 85–94. Available from: https://doi.org/10.4039/Ent1085-5.

Hamilton, W. D. (1980) Sex versus non-sex versus parasite. *Oikos*, 282–290. Available from: https://doi.org/10.2307/3544435.

Han, G., Liu, Q., Li, C., Xu, B., & Xu, J. (2021) Transcriptome sequencing reveals *Cnaphalocrocis medinalis* against baculovirus infection by oxidative stress. *Molecular Immunology*, 129, 63–69. Available from: https://doi.org/10.1016/j.molimm.2020.10.020.

Hearn, J., Blaxter, M., Schönrogge, K., Nieves-Aldrey, J. L., Pujade-Villar, J., Huguet, E., ... & Stone, G. N. (2019) Genomic dissection of an extended phenotype: Oak galling by a cynipid gall wasp. *PLoS genetics*, 15(11), e1008398. Available from: https://doi.org/10.1371/journal.pgen.1008398.

Hearn, J., Gobbo, E., Nieves-Aldrey, J. L., Branca, A., Nicholls, J. A., Koutsovoulos, G., ... & Ronquist, F. (2023). Phylogenomic analysis of protein-coding genes resolves complex gall wasp relationships. *Systematic Entomology*. Available from: https://doi.org/10.1111/syen.12611

Heimpel, G. E., & De Boer, J. G. (2008) Sex determination in the Hymenoptera. *Annual Review of Entomology*, 53, 209–230. Available from: https://doi.org/10.1146/annurev.ento.53.103106.093441.

Hemler, M. E. (2003) Tetraspanin proteins mediate cellular penetration, invasion, and fusion events and define a novel type of membrane microdomain. *Annual review of cell and developmental biology*, 19(1), 397–422. Available from: https://doi.org/10.1146/annurev.cellbio.19.111301.153609.

Hilker, M., & Fatouros, N. E. (2015) Plant responses to insect egg deposition. *Annual Review of Entomology*, 60, 493–515. Available from: https://doi.org/10.1146/annurev-ento-010814-020620.

Ho, M. S., Tsai, P. I., & Chien, C. T. (2006) F-box proteins: the key to protein degradation. *Journal of biomedical science*, 13, 181–191. Available from: https://doi.org/10.1007/s11373-005-9058-2.

Hoff, K.J., Lange, S., Lomsadze, A., Borodovsky, M. and Stanke, M. (2016) BRAKER1: unsupervised RNA-Seq-based genome annotation with GeneMark-ET and AUGUSTUS. *Bioinformatics,* 32(5), 767–769. Available from: https://doi.org/10.1093/bioinformatics/btv661.

Hoff, K.J., Lomsadze, A., Borodovsky, M. and Stanke, M. (2019) Whole-Genome Annotation with BRAKER. *Methods in Molecular Biology,* 1962, 65–95. Available from: https://doi.org/10.1007/978-1-4939-9173-0_5.

Hood, G. R., Zhang, L., Topper, L., Brandão-Dias, P. F., Del Pino, G. A., Comerford, M. S., & Egan, S. P. (2018) 'Closing the life cycle'of *Andricus quercuslanigera* (Hymenoptera: Cynipidae). *Annals of the Entomological Society of America*, 111(3), 103–113. Available from: https://doi.org/10.1093/aesa/say005.

Hossen, M. S., Shapla, U. M., Gan, S. H., & Khalil, M. I. (2016) Impact of bee venom enzymes on diseases and immune responses. *Molecules*, 22(1), 25. Available from: https://doi.org/10.3390/molecules22010025.

Huerta-Cepas, J., Szklarczyk, D., Heller, D., Hernández-Plaza, A., Forslund, S. K., Cook, H., ... & Bork, P. (2019) eggNOG 5.0: a hierarchical, functionally, and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic acids research*, 47(D1), D309–D314. Available from: https://doi.org/10.1093/nar/gky1085.

Hunter, P. (2009) Extended phenotype redux: How far can the reach of genes extend in manipulating the environment of an organism?. *EMBO reports*, 10(3), 212–215. Available from: https://doi.org/10.1038/embor.2009.18.

Ishikawa, A., Ogawa, K., Gotoh, H., Walsh, T. K., Tagu, D., Brisson, J. A., ... & Miura, T. (2012). Juvenile hormone titre and related gene expression during the change of reproductive modes in the pea aphid. *Insect molecular biology*, 21(1), 49-60. Available from: https://doi.org/10.1111/j.1365-2583.2011.01111.x.

Jokela, J., Dybdahl, M. F., & Lively, C. M. (2009). The maintenance of sex, clonal dynamics, and host-parasite coevolution in a mixed population of sexual and asexual snails. *the american naturalist*, 174(S1), S43-S53. Available from: https://doi.org/10.1086/599080.

Kidd, T., Bland, K. S., & Goodman, C. S. (1999) Slit is the midline repellent for the robo receptor in *Drosophila*. *Cell*, 96(6), 785–794. Available from: https://doi.org/10.1016/S0092-8674(00)80589-9.

Koch, E. L., Morales, H. E., Larsson, J., Westram, A. M., Faria, R., Lemmon, A. R., ... & Butlin, R. K. (2021). Genetic variation for adaptive traits is associated with polymorphic inversions in *Littorina saxatilis*. *Evolution letters*, 5(3), 196–213. Available from: https://doi.org/10.1002/evl3.227.

Kohnen, A., Wissemann, V., & Brandl, R. (2011) No host-associated differentiation in the gall wasp *Diplolepis rosae* (Hymenoptera: Cynipidae) on three dog rose species. *Biological Journal of the Linnean Society*, 102(2), 369–377. Available from: https://doi.org/10.1111/j.1095-8312.2010.01582.x.

Korgaonkar, A., Han, C., Lemire, A. L., Siwanowicz, I., Bennouna, D., Kopec, R. E., ... & Stern, D. L. (2021) A novel family of secreted insect proteins linked to plant gall development. *Current Biology*, 31, 1836–1849. Available from: https://doi.org/10.1016/j.cub.2021.01.104.

Korunes, K. L., & Samuk, K. (2021) pixy: Unbiased estimation of nucleotide diversity and divergence in the presence of missing data. *Molecular ecology resources*, 21(4), 1359–1368. Available from: https://doi.org/10.1111/1755-0998.13326.

Kozek, W. J., & Rao, R. U. (2007) The discovery of *Wolbachia* in arthropods and nematodes – A historical perspective. *Wolbachia: a bug's life in another bug*, 5, 1–14. Available from: https://doi.org/10.1159/000104228.

Kremer, N., Charif, D., Henri, H., Bataille, M., Prevost, G., Kraaijeveld, K., & Vavre, F. (2009). A new case of *Wolbachia* dependence in the genus *Asobara*: evidence for parthenogenesis induction in *Asobara japonica*. *Heredity*, 103(3), 248–256. Available from: https://doi.org/10.1038/hdy.2009.63.

Langmead, B., & Salzberg, S. L. (2012) Fast gapped-read alignment with Bowtie 2. *Nature methods*, 9(4), 357–359. Available from: https://doi.org/10.1038/nmeth.1923.

Laszlo, Z., Solyom, K., Prazsmari, H., Barta, Z., & Tothmeresz, B. (2014). Predation on rose galls: parasitoids and predators determine gall size through directional selection. *PLoS One*, 9(6), e99806. Available from: https://doi.org/10.1371/journal.pone.0099806.

László, Z., & Tóthmérész, B. (2013). Landscape and local effects on multiparasitoid coexistence. *Insect Conservation and Diversity*, 6(3), 354-364. Available from: https://doi.org/10.1111/j.1752-4598.2012.00225.x.

László, Z., & Tóthmérész, B. (2013) The enemy hypothesis: correlates of gall morphology with parasitoid attack rates in two closely related rose cynipid galls. *Bulletin of entomological research*, 103(3), 326–335. Available from: https://doi.org/10.1017/S0007485312000764.

Le Fevre, R., Evangelisti, E., Rey, T., & Schornack, S. (2015) Modulation of host cell biology by plant pathogenic microbes. *Annual review of cell and developmental biology*, 31, 201–229. Available from: https://doi.org/10.1146/annurev-cellbio-102314-112502.

Lehmann, M. (2021) Diverse roles of phosphatidate phosphatases in insect development and metabolism. *Insect biochemistry and molecular biology*, 133, 103469. Available from: https://doi.org/10.1016/j.ibmb.2020.103469.

Leulier, F., & Lemaitre, B. (2008) Toll-like receptors — taking an evolutionary approach. *Nature Reviews Genetics*, 9(3), 165–178. Available from: https://doi.org/10.1038/nrg2303.

Li, Y. F., Costello, J. C., Holloway, A. K., & Hahn, M. W. (2008) "Reverse ecology" and the power of population genomics. *Evolution*, 62(12), 2984–2994. Available from: https://doi.org/10.1111/j.1558-5646.2008.00486.x.

Liao, Y., Smyth, G. K., & Shi, W. (2014) featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, 30(7), 923–930. Available from: https://doi.org/10.1093/bioinformatics/btt656.

Linhoff, M. W., Laurén, J., Cassidy, R. M., Dobie, F. A., Takahashi, H., Nygaard, H. B., ... & Craig, A. M. (2009) An unbiased expression screen for synaptogenic proteins identifies the LRRTM protein family as synaptic organizers. *Neuron*, 61(5), 734–749. Available from: https://doi.org/10.1016/j.neuron.2009.01.017.

Love, M. I., Huber, W., & Anders, S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology*, 15(12), 1–21. Available from: https://doi.org/10.1186/s13059-014-0550-8.

Mahadav, A., Gerling, D., Gottlieb, Y., Czosnek, H., & Ghanim, M. (2008) Parasitization by the wasp *Eretmocerus mundus* induces transcription of genes related to immune response and symbiotic bacteria proliferation in the whitefly *Bemisia tabaci*. *BMC genomics*, 9, 1–11. Available from: https://doi.org/10.1186/1471-2164-9-342.

Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A., & Zdobnov, E. M. (2021) BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Molecular biology and evolution*, 38(10), 4647–4654. Available from: https://doi.org/10.1093/molbev/msab199.

Martinson, E. O., Werren, J. H., & Egan, S. P. (2022) Tissue-specific gene expression shows a cynipid wasp repurposes oak host gene networks to create a complex and novel parasite-specific organ. *Molecular Ecology*, 31(11), 3228–3240. Available from: https://doi.org/10.1111/mec.16159.

Maynard-Smith, J. (1978) *The evolution of sex*. Cambridge: Cambridge University Press.

McDonald, J. H., & Kreitman, M. (1991). Adaptive protein evolution at the Adh locus in Drosophila. *Nature*, 351(6328), 652–654. Available from: https://doi.org/10.1038/351652a0.

Medzhitov, R. (2001) Toll-like receptors and innate immunity. *Nature Reviews Immunology*, 1(2), 135–145. Available from: https://doi.org/10.1038/35100529.

Mei, X., Qiao, P., Ma, H., Qin, S., Song, X., Zhao, Q., & Shen, D. (2023) *Bombyx mori* Tetraspanin A (BmTsp. A) is a facilitator in BmNPV invasion by regulating apoptosis. *Developmental & Comparative Immunology*, 146, 104736. Available from: https://doi.org/10.1016/j.dci.2023.104736.

Mejias, J., Truong, N. M., Abad, P., Favery, B., & Quentin, M. (2019) Plant proteins and processes targeted by parasitic nematode effectors. *Frontiers in plant science*, 10, 970. Available from: https://doi.org/10.3389/fpls.2019.00970.

Meyer, J., & Maresquelle, H. J. (1983) Anatomie des galles. Berlin: Gebrüder Borntraeger.

Mozhaitseva, K., Tourrain, Z., Branca, A. (2023) Population genomics of the mostly thelytokous *Diplolepis rosae* (Linnaeus, 1758) (Hymenoptera: Cynipidae) reveals population-specific selection for sex. *Genome Biology and Evolution*, evad185. Available from: https://doi.org/10.1093/gbe/evad185.

Muller, H. J. (1932) Some genetic aspects of sex. *The American Naturalist*, 66(703), 118–138. Available from: https://doi.org/10.1086/280418.

Nabity, P. D., Haus, M. J., Berenbaum, M. R., & DeLucia, E. H. (2013) Leaf-galling phylloxera on grapes reprograms host metabolism and morphology. *Proceedings of the National Academy of Sciences*, 110(41), 16663–16668. Available from: https://doi.org/10.1073/pnas.1220219110.

National Center for Biotechnology Information (NCBI) (1988) *Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information.* Available at: https://www.ncbi.nlm.nih.gov/ [Accessed 25th July 2023].

Nei, M. (1987) *Molecular evolutionary genetics*. Columbia university press.

Nordlander G. (1973) Parasitsteklar i galler av *Diplolepis rosae* (L.) och *D. mayri* Schlechtd. (Hym. Cynipidae) (Hym. Ichneumonoidea, Chalcidoidea, Cynipoidea). *Entomologisk Tidskrift*. 94(3–4), 148–176.

Oates, C. N., Denby, K. J., Myburg, A. A., Slippers, B., & Naidoo, S. (2016) Insect gallers and their plant hosts: from omics data to systems biology. *International journal of molecular sciences*, 17(11), 1891. Available from: https://doi.org/10.3390/ijms17111891.

Panchy, N., Lehti-Shiu, M., & Shiu, S. H. (2016) Evolution of gene duplication in plants. *Plant Physiology*, 171(4), 2294–2316. Available from: https://doi.org/10.1104/pp.16.00523.

Pannebakker, B. A., Pijnacker, L. P., Zwaan, B. J., & Beukeboom, L. W. (2004). Cytology of *Wolbachia*-induced parthenogenesis in *Leptopilina clavipes* (Hymenoptera: Figitidae). *Genome*, 47(2), 299–303. Available from: https://doi.org/10.1139/g03-137.

Pannebakker, B. A., Zwaan, B. J., Beukeboom, L. W., & Van Alphen, J. J. (2004a). Genetic diversity and *Wolbachia* infection of the *Drosophila* parasitoid *Leptopilina clavipes* in western Europe. *Molecular Ecology*, *13*(5), 1119-1128. Available from: https://doi.org/10.1111/j.1365-294X.2004.02147.x.

Paretas-Martínez, J., Restrepo-Ortiz, C., Buffington, M., & Pujade-Villar, J. (2011) Systematics of *Australian Thrasorinae* (Hymenoptera, Cynipoidea, Figitidae) with descriptions of Mikeiinae, new subfamily, two new genera, and three new species. *ZooKeys*, 108, 21–48. Available at: https://doi.org/10.3897/zookeys.108.829.

Park, S., Jo, Y. H., Park, K. B., Ko, H. J., Kim, C. E., Bae, Y. M., ... & Han, Y. S. (2019) TmToll-7 plays a crucial role in innate immune responses against Gram-negative bacteria by regulating 5 AMP genes in *Tenebrio molitor*. *Frontiers in Immunology*, 10, 310. Available from: https://doi.org/10.3389/fimmu.2019.00310.

Parsch, J., Zhang, Z., & Baines, J. F. (2009). The influence of demography and weak selection on the McDonald–Kreitman test: an empirical study in Drosophila. *Molecular Biology and Evolution*, 26(3), 691-698. Available from: https://doi.org/10.1093/molbev/msn297.

Pascual-Alvarado, E., Nieves-Aldrey, J. L., Castillejos-Lemus, D. E., Cuevas-Reyes, P., & Oyama, K. (2017) Diversity of galls induced by wasps (Hymenoptera: Cynipidae, Cynipini) associated with oaks (Fagaceae: Quercus) in Mexico. *Botanical Sciences*, 95(3), 461–472. Available from: https://doi.org/10.17129/botsci.1215.

Pearcy, M., Hardy, O., & Aron, S. (2006) Thelytokous parthenogenesis and its consequences on inbreeding in an ant. *Heredity*, 96(5), 377–382. Available from: https://doi.org/10.1038/sj.hdy.6800813.

Pearson, W. R. (2013) An introduction to sequence similarity ("homology") searching. *Current protocols in bioinformatics*, 42(1), 3–1. Available from: https://doi.org/10.1002/0471250953.bi0301s42.

Petersen, M., Armisén, D., Gibbs, R. A., Hering, L., Khila, A., Mayer, G., ... & Misof, B. (2019). Diversity and evolution of the transposable element repertoire in arthropods with particular reference to insects. *BMC Ecology and Evolution*, 19(1), 1–15. Available from: https://doi.org/10.1186/s12862-018-1324-9.

Plantard O, Rasplus J-Y, Mondor G, Le Clainche I, Solignac M. (1998) *Wolbachia*-induced thelytoky in the rose gall wasp *Diplolepis spinosissimae* (Giraud) (Hymenoptera: Cynipidae), and its consequences on the genetic structure of its host. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 265(1401), 1075–1080. Available from: https://doi.org/10.1098/rspb.1998.0401.

Plantard, O., Field, L. M., James, A. A., Rasplus, J. Y., Mondor, G., Le Clainche, I., & Solignac, M. (1999) Distribution and phylogeny of *Wolbachia* inducing thelytoky in Rhoditini and 'Aylacini'(Hymenoptera: Cynipidae). *Insect Molecular Biology*, 8(2), 185–191. Available from: https://doi.org/10.1046/j.1365-2583.1999.820185.x.

Poirié, M., Colinet, D., & Gatti, J. L. (2014) Insights into function and evolution of parasitoid wasp venoms. *Current Opinion in Insect Science*, 6, 52–60. Available from: https://doi.org/10.1016/j.cois.2014.10.004.

Pujade-Villar, J., Bellido, D., Segu, G., & Melika, G. (2001) Current state of knowledge of heterogony in Cynipidae (Hymenoptera, Cynipoidea). *Sessió Conjunta d'Entomologia*, 11, 87–107. Available from: https://raco.cat/index.php/SessioEnto/article/view/192160.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., ... & Sham, P. C. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American journal of human genetics*, 81(3), 559–575. doi: 10.1086/519795.

Qian, C., Fang, Q., Wang, L., & Ye, G. Y. (2015) Molecular cloning and functional studies of two Kazal-type serine protease inhibitors specifically expressed by *Nasonia vitripennis* venom apparatus. *Toxins*, 7(8), 2888–2905. Available from: https://doi.org/10.3390/toxins7082888.

R Core Team (2022) *R: A language and environment for statistical computing*. Available at: https://www.R-project.org/ [Accessed 24 September 2023].

Rabeling, C., & Kronauer, D. J. (2013) Thelytokous parthenogenesis in eusocial Hymenoptera. *Annual review of entomology*, 58, 273–292. Available from: https://doi.org/10.1146/annurev-ento-120811-153710.

Rausch, T., Zichner, T., Schlattl, A., Stütz, A. M., Benes, V., & Korbel, J. O. (2012). DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics*, *28*(18), i333–i339. Available from: https://doi.org/10.1093/bioinformatics/bts378.

Rizzo, M. C., & Massa, B. (2006) Parasitism and sex ratio of the Bedeguar gall wasp *Diplolepis rosae* (L.) (Hymenoptera: Cynipidae) in Sicily (Italy). *Journal of Hymenoptera Research*, 15(2), 2779285. Available from: http://biostor.org/reference/372.

Rohfritsch, O., & Shorthouse, J. D. (1982) Insect galls. *Molecular biology of plant tumors*, 131–152. Available from: https://doi.org/10.1016/B978-0-12-394380-4.50011-6.

Ronquist, F. (1994) Evolution of parasitism among closely related species: phylogenetic relationships and the origin of inquilinism in gall wasps (Hymenoptera, Cynipidae). *Evolution*, 48(2), 241–266. Available from: https://doi.org/10.1111/j.1558-5646.1994.tb01310.x.

Ronquist, F., Nieves-Aldrey, J. L., Buffington, M. L., Liu, Z., Liljeblad, J., & Nylander, J. A. (2015) Phylogeny, evolution, and classification of gall wasps: the plot thickens. *PLoS One*, 10(5), e0123301. Available from: https://doi.org/10.1371/journal.pone.0123301.

Sahlin, K., Vezzi, F., Nystedt, B., Lundeberg, J., & Arvestad, L. (2014) BESST-efficient scaffolding of large fragmented assemblies. *BMC bioinformatics*, 15, 1–11. Available from: https://doi.org/10.1186/1471-2105-15-281.

Sanderson, A. R. (1988) Cytological investigations of parthenogenesis in gall wasps (Cynipidae, Hymenoptera). *Genetica*, 77(3), 189–216. Available from: https://doi.org/10.1007/BF00122389.

Sardon-Gutierrez, S., Gil, D., Gómez, J. F., & Nieves-Aldrey, J. L. (2021) Ecological niche modelling of species of the rose gall wasp *Diplolepis* (Hymenoptera: Cynipidae) on the Iberian Peninsula. *European Journal of Entomology* 118, 31–45. doi: 10.14411/eje.2021.004

Simon, J. C., Rispe, C., & Sunnucks, P. (2002). Ecology and evolution of sex in aphids. *Trends in Ecology & Evolution*, 17(1), 34–39. Available from: https://doi.org/10.1016/S0169-5347(01)02331-X.

Schilthuizen, M., & Stouthamer, R. (1998) Distribution of *Wolbachia* among the guild associated with the parthenogenetic gall wasp *Diplolepis rosae*. *Heredity*, 81(3), 270–274. Available from: https://doi.org/10.1046/j.1365-2540.1998.00385.x.

Schneider, M. V., Beukeboom, L. W., Driessen, G., Lapchin, L., Bernstein, C., & Van Alphen, J. J. (2002) Geographical distribution and genetic relatedness of sympatrical thelytokous and arrhenotokous populations of the parasitoid *Venturia canescens* (Hymenoptera). *Journal of Evolutionary Biology*, *15*(2), 191–200. Available from: https://doi.org/10.1046/j.1420-9101.2002.00394.x

Schönrogge, K., Harper, L. J., & Lichtenstein, C. P. (2000) The protein content of tissues in cynipid galls (Hymenoptera: Cynipidae): similarities between cynipid galls and seeds. *Plant, Cell & Environment*, 23(2), 215–222. Available from: https://doi.org/10.1046/j.1365-3040.2000.00543.x.

Shorthouse, J. D., & Floate, K. D. (2010) Galls induced by cynipid wasps of the genus *Diplolepis* (Hymenoptera: Cynipidae) on the roses of Canada's grasslands. *Arthropods of Canadian Grasslands. Biological Survey of Canada*, 251–279. Available from: https://doi.org/10.3752/9780968932148.CH12.

Shorthouse, J. D., Leggo, J. J., Sliva, M. D., & Lalonde, R. G. (2005). Has egg location influenced the radiation of *Diplolepis* (Hymenoptera: Cynipidae) gall wasps on wild roses?. *Basic and Applied Ecology*, 6(5), 423–434. Available from: https://doi.org/10.1016/j.baae.2005.07.006.

Shrestha, S., Park, Y., Stanley, D., & Kim, Y. (2010) Genes encoding phospholipases A2 mediate insect nodulation reactions to bacterial challenge. *Journal of insect physiology*, 56(3), 324–332. Available from: https://doi.org/10.1016/j.jinsphys.2009.11.008.

Smith, T. F., & Waterman, M. S. (1981) Identification of common molecular subsequences. *Journal of molecular biology*, 147(1), 195–197. doi: 10.1016/0022-2836(81)90087-5.

Stanke, M., Diekhans, M., Baertsch, R., & Haussler, D. (2008) Using native and synthetically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics*, 24(5), 637–644. Available from: https://doi.org/10.1093/bioinformatics/btn013.

Stanke. M., Schöffmann, O., Morgenstern, B. & Waack, S. (2006) Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. *BMC Bioinformatics*, 7, 62. Available from: https://doi.org/10.1186/1471-2105-7-62.

Stanley, D. (2006) The non-venom insect phospholipases A2. *Biochimica et Biophysica Acta (BBA)-Molecular and Cell Biology of Lipids*, 1761(11), 1383–1390. Available from: https://doi.org/10.1016/j.bbalip.2006.05.011.

Stille, B. (1984) The effect of hosptlant and parasitoids on the reproductive success of the parthenogenetic gall wasp *Diplolepis rosae* (Hymenoptera, Cynipidae). *Oecologia*, 63, 364–369. Available from: https://doi.org/10.1007/BF00390666.

Stille, B. O., & Dävring, (1980) Meiosis and reproductive strategy in the parthenogenetic gall wasp *Diplolepis rosae* (L.) (Hymenoptera, Cynipidae). *Hereditas*, 92(2), 353–362. Available from: https://doi.org/10.1111/j.1601-5223.1980.tb01720.x.

Stone, G. N., & Schönrogge, K. (2003) The adaptive significance of insect gall morphology. *Trends in Ecology & Evolution*, 18(10), 512–522. Available from:https://doi.org/10.1016/S0169-5347(03)00247-7.

Stone, G. N., Schönrogge, K., Atkinson, R. J., Bellido, D., & Pujade-Villar, J. (2002) The population biology of oak gall wasps (Hymenoptera: Cynipidae). *Annual review of entomology*, 47(1), 633–668. Available from: https://doi.org/10.1146/annurev.ento.47.091201.145247.

Stuart, J. J., Chen, M. S., Shukle, R., & Harris, M. O. (2012) Gall midges (Hessian flies) as plant pathogens. *Annual review of phytopathology*, 50, 339–357. Available from: https://doi.org/10.1146/annurev-phyto-072910-095255.

Stouthamer, R., & Kazmer, D. J. (1994) Cytogenetics of microbe-associated parthenogenesis and its consequences for gene flow in *Trichogramma* wasps. *Heredity*, 73(3), 317–327. Available from: https://doi.org/10.1038/hdy.1994.139.

Thornton, K. (2003) Libsequence: a C++ class library for evolutionary genetic analysis. *Bioinformatics*, 19(17), 2325–2327. Available from: https://doi.org/10.1093/bioinformatics/btg316.

Todorov, I., Stojanova, A., Parvanov, D., & Boyadzhiev, P. (2012) Studies on the gall community of *Diplolepis rosae* (Hymenoptera: Cynipidae) in Vitosha Mountain, Bulgaria. *Acta Zoologica Bulgarica*, 4, 27–37.

Todres, E., Nardi, J. B., & Robertson, H. M. (2000) The tetraspanin superfamily in insects. *Insect molecular biology*, 9(6), 581–590. Available from: https://doi.org/10.1046/j.1365-2583.2000.00222.x.

Vallarino, J. G., & Osorio, S. (2012) Signaling role of oligogalacturonides derived during cell wall degradation. *Plant signaling & behavior*, 7(11), 1447–1449. Available from: https://doi.org/10.4161/psb.21779.

Van Valen, L. (1973) A new evolutionary law. Evolutionary Theory, 1, 1–30.

Wade M.J. (eds) *Epistasis and the Evolutionary Process*. Oxford University Press: New York. pp. 146–157.

Werren, J. H., Zhang, W., & Guo, L. R. (1995). Evolution and phylogeny of *Wolbachia*: reproductive parasites of arthropods. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 261(1360), 55–63. Available from: https://doi.org/10.1098/rspb.1995.0117.

Whiting, P. W. (1928). Biological notes on *Nemeritis canescens* (Grav.) (Ichneumonidae). *Psyche: A Journal of Entomology*, 35(2), 125–125.

Williams, G. C. (1966) Natural selection, the costs of reproduction, and a refinement of Lack's principle. *The American Naturalist*, 100(916), 687–690. Available from: https://doi.org/10.1086/282461.

Williams, R. (2013) The Natural History of the Rose Bedeguar Gall: Field Work and Keys to the Insect Inhabitants. *The British Plant Gall Society*. Haverhill, UK: Red Side Up Limited.

Wolda, H. (1988) Insect seasonality: why?. *Annual review of ecology and systematics*, 19(1), 1–18. Available from: https://doi.org/10.1146/annurev.es.19.110188.000245.

Wybouw, N., Pauchet, Y., Heckel, D. G., & Van Leeuwen, T. (2016) Horizontal gene transfer contributes to the evolution of arthropod herbivory. *Genome biology and evolution*, 8(6), 1785–1801. Available from: https://doi.org/10.1093/gbe/evw119.

Xia, Y., Dean, P., Judge, A. J., Gillespie, J. P., Clarkson, J. M., & Charnley, A. K. (2000) Acid phosphatases in the haemolymph of the desert locust, *Schistocerca gregaria*, infected with the entomopathogenic fungus *Metarhizium anisopliae*. *Journal of Insect Physiology*, 46(9), 1249–1257. Available from: https://doi.org/10.1016/S0022-1910(00)00045-7.

Xia, A. N., Liu, J., Kang, D. C., Zhang, H. G., Zhang, R. H., & Liu, Y. G. (2020) Assessment of endophytic bacterial diversity in rose by high-throughput sequencing analysis. *PloS one*, *15*(4), e0230924. Available from: https://doi.org/10.1371/journal.pone.0230924.

Yamaguchi, H., Tanaka, H., Hasegawa, M., Tokuda, M., Asami, T., & Suzuki, Y. (2012) Phytohormones and willow gall induction by a gall-inducing sawfly. *New Phytologist*, 196(2), 586–595. Available from: https://doi.org/10.1111/j.1469-8137.2012.04264.x.

Yao W. (2023). *intansv: Integrative analysis of structural variations*. *R package version 1.40.0*. Available at: https://www.bioconductor.org/packages/release/bioc/html/intansv.html [Accessed 3rd October 2023].

Zayed, A. (2004) Effective population size in Hymenoptera with complementary sex determination. *Heredity*, 93(6), 627–630. Available from: https://doi.org/10.1038/sj.hdy.6800588.

Zhang, Y. M., Egan, S. P., Driscoe, A. L., & Ott, J. R. (2021) One hundred and sixty years of taxonomic confusion resolved: *Belonocnema* (Hymenoptera: Cynipidae: Cynipini) gall wasps associated with live oaks in the USA. *Zoological Journal of the Linnean Society*, 193(4), 1234–1255. Available from: https://doi.org/10.1093/zoolinnean/zlab001.

Zhao, C., Escalante, L. N., Chen, H., Benatti, T. R., Qu, J., Chellapilla, S., ... & Richards, S. (2015) A massive expansion of effector genes underlies gall-formation in the wheat pest *Mayetiola destructor*. *Current Biology*, 25(5), 613–620. Available from: https://doi.org/10.1016/j.cub.2014.12.057.

Zhao, Y., Xiong, Z., Wu, G., Bai, W., Zhu, Z., Gao, Y., ... & Li, H. (2018) Fungal endophytic communities of two wild *Rosa* varieties with different powdery mildew susceptibilities. *Frontiers in Microbiology*, 9, 2462. Available from: https://doi.org/10.3389/fmicb.2018.02462.

Zhuang, S., Kelo, L., Nardi, J. B., & Kanost, M. R. (2007) An integrin-tetraspanin interaction required for cellular innate immune responses of an insect, *Manduca sexta*. *Journal of Biological Chemistry*, 282(31), 22563–22572. Available from: https://doi.org/10.1074/jbc.M700341200.

SUPPLEMENTARY MATERIAL

Tables

Table S1. **Quality of the *Diplolepis rosae* genome assembly.**

Parameter	Value
# contigs (>= 0 bp)	757
# contigs (>= 1000 bp)	729
# contigs (>= 5000 bp)	569
# contigs (>= 10000 bp)	429
# contigs (>= 25000 bp)	326
# contigs (>= 50000 bp)	285
Total length (>= 0 bp)	760724616
Total length (>= 1000 bp)	760705830
Total length (>= 5000 bp)	760203332
Total length (>= 10000 bp)	759172802
Total length (>= 25000 bp)	757619231
Total length (>= 50000 bp)	756165228
Largest contig, bp	33459371
Total length, bp	760724616
% GC bases	33.43
N50	7663408
N90	1376597
L50	25
L90	117
# N's per 100 kbp	0.00

Table S2. Classification and percentage of repetitive sequences in the genome of *Diplolepis rosae*.

Class of repetitive sequence	Number of elements †	Length occupied, bp	% of genome assembly
Retroelements:	125770	86726618	11.40
SINEs:	10149	1528961	0.20
Penelope	39915	16183944	2.13
LINEs:	66655	38394411	5.05
CRE/SLACS	0	0	0
L2/CR1/Rex	0	0	0
R1/LOA/Jockey	21111	19220324	2.53
R2/R4/NeSL	872	334972	0.04
RTE/Bov-B	2151	558997	0.07
L1/CIN4	0	0	0
LTR elements:	48966	46803246	6.15
BEL/Pao	2918	2652590	0.35
Ty1/Copia	2178	1798818	0.24
Gypsy/DIRS1	43259	41907222	5.51
Retroviral	0	0	0
DNA transposons:	176423	65850608	8.66
hobo-Activator	13608	8583491	1.13
Tc1-IS630-Pogo	101005	38660213	5.08
En-Spm	0	0	0
MuDR-IS905	0	0	0
PiggyBac	20152	4871256	0.64
Tourist/Harbinger	1306	206185	0.03
Other (Mirage, P-element, Transib)	6541	2519567	0.33
Rolling-circles	1558	288437	0.04
Unclassified	1198129	368923706	48.50
Small RNA	7582	822276	0.11
Satellites	378	84139	0.01
Simple repeats	78424	3534451	0.46
Low complexity	15988	798635	0.10

† Most repeats fragmented by insertions or deletions have been counted as one element.

Table S3. Akaike Information Criterion (AIC) values for examined demographic scenarios for *Diplolepis rosae*.

Model	AIC
bottlegrowth_split_no_migration	74419.2
bottlegrowth_split_migration	79152.6
split_no_migration	144643.8
split_asymmetric_migration	145062.8
split_symmetric_migration	233881.6
isolation_no_migration	154600.6
isol_asymmetric_migration	146823.6
isol_symmetric_migration	159047.4
exp_pop_growth_size_change_split_no_migration	372553.8
exp_pop_growth_size_change_split_asymmetric_migration	371429.2
exp_pop_growth_size_change_split_symmetric_migration	371833.6

Table S4. Parameter estimations provided by approximate Bayesian computation for the demographic scenario 'bottleneck of an ancestral population ➔ split to two populations' examined for *Diplolepis rosae*.

Parameter	N1	N2	proportion	Tbot	Tsplit	μ
Prediction error	0.96	1.1	0.76	1.0	0.96	0.88
Estimation, median	511	494	0.62	1780	849	5.5e-7
Units	Number of diploid individuals		-	Generations ago		per bp per generation

Estimation is considered correct if its prediction value is less than 1. N1, N2: effective population sizes of lineage 1 and lineage 2, respectively; proportion: ratio between a number of individuals of *D. rosae* from a population after a bottleneck to an ancestral one; Tbot, Tsplit: bottleneck and split time, respectively; μ: mutation rate.

Table S5. Gene ontology terms showing outlier composite score values in the two Diplolepis rosae populations detected by Gene Set Enrichment Analysis.

Composite score (CS) outliers in *D. rosae* lineages	Significant GO annotations
Negative (CS < -0.5) population 1	GO:0046692: sperm competition (BP): 8/29 (p < 0.0001) GO:0007320: insemination (BP): 8/32 (p < 0.0001) GO:0007620: copulation (BP): 8/41 (p < 0.0001) GO:0008237: metallopeptidase activity (MF): 11/123 (p = 0.000253) GO:0004222: metalloendopeptidase activity (MF): 9/75 (p = 0.000253)
Negative (CS < -0.5) population 2	GO:0071105: response to interleukin-1 (BP): 6/8 (p = 0.000278) GO:0071348: cellular response to interleukin-11 (BP): 6/8 (p = 0.000278) GO:1903544: response to butyrate (BP): 6/8 (p = 0.000278) GO:1903545: cellular response to butyrate (BP): 6/8 (p = 0.000278) GO:2001028: positive regulation of endothelial cell (BP): 7/13 (p = 0.000448) GO:0099641: anterograde axonal protein transport (BP): 6/9 (p = 0.000462) GO:1905907: negative regulation of amyloid fibril (BP): 6/9 (p = 0.000462) GO:0032463: negative regulation of protein homooligomerization (BP): 7/14 (p = 0.000481) GO:2001026: regulation of endothelial cell chemotaxis (BP): 7/14 (p = 0.000481) GO:0038033: positive regulation of endothelial cell (BP): 6/10 (p = 0.000711) GO:0042595: behavioral response to starvation (BP): 6/10 (p = 0.000711) GO:1905906: regulation of amyloid fibril formation (BP): 7/16 (p = 0.00113) GO:0032460: negative regulation of protein oligomer (BP): 7/17 (p = 0.00139) GO:0032462: regulation of protein homooligomerization (BP): 7/17 (p = 0.00139) GO:0035767: endothelial cell chemotaxis (BP): 7/17 (p = 0.00139) GO:0042026: protein refolding (BP): 7/17 (p = 0.00139) GO:0038089: positive regulation of cell migration (BP): 6/12 (p = 0.00190) GO:0032459: regulation of protein oligomerization (BP): 8/25 (p = 0.00202) GO:1905383: protein localization to presynapse (BP): 6/13 (p = 0.00305) GO:1990000: amyloid fibril formation (BP): 7/20 (p = 0.00401) GO:0007021: tubulin complex assembly (BP): 6/14 (p = 0.00468) GO:0032731: positive regulation of interleukin-1 (BP): 6/15 (p = 0.00718)
	GO:1902176: negative regulation of oxidative stress (BP): 7/22 (p = 0.00718) GO:0051016: barbed-end actin filament capping (BP): 6/16 (p = 0.00909) GO:0098840: protein transport along microtubule (BP): 6/16 (p = 0.00909) GO:0099118: microtubule-based protein transport (BP): 6/16 (p = 0.00909) GO:0099640: axo-dendritic protein transport (BP): 6/16 (p = 0.00909) GO:1990776: response to angiotensin (BP): 8/32 (p = 0.0100) GO:0009631: cold acclimation (BP): 6/17 (p = 0.0122) GO:0032732: positive regulation of interleukin-1 (BP): 6/17 (p = 0.0122)

	GO:0071679: commissural neuron axon guidance (BP): 6/18 (p = 0.0169) GO:0032651: regulation of interleukin-1 beta product (BP): 7/26 (p = 0.0169) GO:1902175: regulation of oxidative stress-induced (BP): 7/26 (p = 0.0169) GO:0032611: interleukin-1 beta production (BP): 7/28 (p = 0.0277) GO:0002931: response to ischemia (BP): 7/29 (p = 0.0333) GO:0032652: regulation of interleukin-1 production (BP): 7/29 (p = 0.0333) GO:0035994: response to muscle stretch (BP): 6/22 (p = 0.0450) GO:0051693: actin filament capping (BP): 6/22 (p = 0.0450) GO:0008089: anterograde axonal transport (BP): 8/41 (p = 0.0450) GO:1900408: negative regulation of cellular response (BP): 8/41 (p = 0.0450) GO:1903202: negative regulation of oxidative stress (BP): 8/41 (p = 0.0450) GO:0032612: interleukin-1 production (BP): 7/31 (p = 0.0450) GO:0008426: protein kinase C inhibitor activity (MF): 7/10 (p = 0.000159) GO:0005212: structural constituent of eye lens (MF): 6/11 (p = 0.00776) GO:0097512: cardiac myofibril (CC): 6/10 (p = 0.00527)
Positive (CS > 0.5) population 1	GO:0071679: commissural neuron axon guidance (BP): 6/18 (p = 0.0139)
Positive (CS > 0.5) population 2	GO:0046692: sperm competition (BP): 9/29 (p < 0.0001) GO:0007320: insemination (BP): 9/32 (p < 0.0001) GO:0007620: copulation (BP): 9/41 (p < 0.0001) GO:0007617: mating behavior (BP): 10/167 (p = 0.00940) GO:0044706: multi-multicellular process (BP): 9/152 (p = 0.0271) GO:0007618: mating (BP): 9/160 (p = 0.0341) GO:0008237: metalloendopeptidase activity (MF): 75/10 (p < 0.0001) GO:0004222: metallopeptidase activity (MF): 11/123 (p < 0.0001) GO:0004175: endopeptidase activity (MF): 11/203 (p = 0.00780) GO:0004175: peptidase activity (MF): 13/340 (p = 0.0300)

Table S6. Coverage *Wolbachia* supergroup A and B bins (contigs built after metagenome assembly and binning) recovered from *Diplolepis rosae* Illumina reads.

Population	*D. rosae* individual	bin	Coverage	Identified *Wolbachia* Supergroup (A or B)	*D. rosae* genome coverage	Normalized *Wolbachia* coverage
1	078	bin.1.fa	9.82822		44.8897	
1	078	bin.2.fa	36.2733			
1	078	bin.3.fa	17.3943			
1	078	bin.4.fa	5.21198			
1	078	bin.5.fa	13.0497	A and B		0.29
1	082	bin.1.fa	9.17111		58.1444	
1	082	bin.2.fa	14.354	B		0.25
1	082	bin.3.fa	9.34931			
1	082	bin.4.fa	7.32544			
1	082	bin.5.fa	13.8626			
1	082	bin.6.fa	14.8484			
1	082	bin.7.fa	28.1723			
1	082	bin.8.fa	9.44117			
2	117	bin.1.fa	9.9021		25.9941	
2	117	bin.2.fa	23.8607	A and B		0.92
2	117	bin.3.fa	18.5399			
2	117	bin.4.fa	43.5673			
1	126	bin.1.fa	24.5021		27.1813	
1	126	bin.2.fa	4.5765			
1	126	bin.3.fa	13.4358			
1	126	bin.4.fa	6.61558			
1	126	bin.5.fa	28.1247			
1	126	bin.6.fa	13.6376	B		0.50
1	126	bin.7.fa	26.043	A		0.96
1	219	bin.1.fa	46.5367		119.241	
1	219	bin.2.fa	23.4625			
1	219	bin.3.fa	10.7127			

1	219	bin.4.fa	42.6195			
1	219	bin.5.fa	107.281	B		0.90
1	219	bin.6.fa	17.9632			
2	288	bin.1.fa	8.33563		38.8936	
2	288	bin.2.fa	8.5904			
2	288	bin.3.fa	93.9574	A		2.4
2	288	bin.4.fa	122.614	B		3.2
2	288	bin.5.fa	148.543	B		3.8
1	312	bin.1.fa	251.813	B	34.1124	7.4
1	312	bin.2.fa	8.66442			
2	330	bin.1.fa	7.54022		36.7328	
2	330	bin.2.fa	7.33231			
2	330	bin.3.fa	8.48644			
1	464	bin.1.fa	5.54094		29.4562	
1	464	bin.2.fa	348.081			
1	464	bin.3.fa	14.0629			
1	464	bin.4.fa	308.676	B		10.5
1	464	bin.5.fa	12.6503			
1	464	bin.6.fa	13.2359			
1	576	bin.10.fa	84.098	B	38.3787	2.2
1	576	bin.11.fa	16.5527	B		0.43
1	576	bin.12.fa	11.6308			
1	576	bin.13.fa	11.2718			
1	576	bin.1.fa	12.8816			
1	576	bin.2.fa	11.2316			
1	576	bin.3.fa	11.9606			
1	576	bin.4.fa	11.3751			
1	576	bin.5.fa	11.9592			
1	576	bin.6.fa	7.63389			
1	576	bin.7.fa	7.69986			

1	576	bin.8.fa	11.5755			
1	576	bin.9.fa	11.18			
2	580	bin.1.fa	9.88938		39.3057	
2	580	bin.2.fa	13.0791			
2	580	bin.3.fa	10.38			
2	580	bin.4.fa	8.82744			
2	580	bin.5.fa	10.5102			
2	580	bin.6.fa	7.80195			
2	580	bin.7.fa	12.2507			
2	580	bin.8.fa	152.488	B		3.9
1	608	bin.1.fa	89.8094	B	36.2919	2.5
1	608	bin.2.fa	20.7932			
1	608	bin.3.fa	8.68584			
1	608	bin.4.fa	190.495	A		5.2
1	608	bin.5.fa	8.2291			
2	623	bin.1.fa	474.038	B	33.7402	14.0
2	623	bin.2.fa	554.346			
2	623	bin.3.fa	6.25831			
1	630	bin.1.fa	19.6523		32.8526	
1	630	bin.2.fa	16.5952			
1	630	bin.3.fa	106.839	B		0.60
1	630	bin.4.fa	33.1374			
1	630	bin.5.fa	7.15868			
1	630	bin.6.fa	7.00814			
2	652	bin.1.fa	146.607	B	46.3449	3.2
2	652	bin.2.fa	8.9036			
2	652	bin.3.fa	8.3211			

Table S7. Characteristics of *Diplolepis rosae* samples used in Illumina sequencing.

Species	Sample name	Stage	Number of individuals per sample	Tissue	Sex	Host plant	Isolation source	Collection date	Latitude	Longitude
Diplolepis rosae	ESE-078	Imago	6	Whole individual	Female	*Rosa canina*	Gall	2020-05-12	48.70199	2.173458
Diplolepis rosae	ESE-082	Imago	7	Whole individual	Female	*Rosa canina*	Gall	2020-05-13	48.70371	2.083596
Diplolepis rosae	ESE-117	Imago	8	Whole individual	Female	*Rosa canina*	Gall	2020-04-04	45.75876	6.159383
Diplolepis rosae	ESE-126	Imago	8	Whole individual	Female	*Rosa canina*	Gall	2020-06-14	47.24930	6.073626
Diplolepis rosae	ESE-219	Imago	30	Whole individual	Female	*Rosa canina*	Gall	2020-08-10	46.42217	1.353230
Diplolepis rosae	ESE-288	Larva	7	Whole individual	Female	*Rosa canina*	Gall	2020-08-26	47.32478	3.792692
Diplolepis rosae	ESE-312	Larva	10	Whole individual	Female	*Rosa canina*	Gall	2020-09-02	50.00647	4.743896
Diplolepis rosae	ESE-330	Larva	20	Whole individual	Female	*Rosa canina*	Gall	2020-09-05	44.52756	5.016034
Diplolepis rosae	ESE-349	Larva	10	Whole individual	Female	*Rosa canina*	Gall	2020-09-08	45.62013	3.057751
Diplolepis rosae	ESE-464	Larva	10	Whole individual	Female	*Rosa canina*	Gall	2020-09-23	43.44791	6.239238
Diplolepis rosae	ESE-559	Larva	12	Whole individual	Female	*Rosa canina*	Gall	2020-09-10	43.96794	3.405437
Diplolepis rosae	ESE-576	Larva	10	Whole individual	Female	*Rosa canina*	Gall	2020-10-12	43.03309	1.003795
Diplolepis rosae	ESE-580	Larva	10	Whole individual	Female	*Rosa canina*	Gall	2020-10-09	42.70232	2.564203
Diplolepis rosae	ESE-608	Larva	8	Whole individual	Female	*Rosa canina*	Gall	2020-10-20	44.58980	-0.454849

Diplolepis rosae	ESE-623	Larva	7	Whole individual	Female	*Rosa canina*	Gall	2020-10-20	43.65517	-1.180453
Diplolepis rosae	ESE-630	Larva	10	Whole individual	Female	*Rosa canina*	Gall	2020-10-20	45.15035	0.529976
Diplolepis rosae	ESE-652	Larva	5	Whole individual	Female	*Rosa canina*	Gall	2020-12-27	47.30910	-3.067183

Table S8. **Characteristics of *Diplolepis eglanteriae* samples used in Illumina sequencing.**

Sample name	Stage	Num Ind	Tissue	Sex	Host plant	Isolation source	Collect date	Lat	Long
54	-	1	Whole Ind	Female	*Rosa sp.*	Gall	2019-10-09	48.421111	0.923611
281.1	Imago	1	Whole Ind	Female	*Rosa canina*	Gall	2020-08-27	47.286478	4.7273

Num ind: number of individuals per sample. Whole Ind: whole individual. Collect date: collection date. Lat: latitude. Long: longitude.

Table S9. Quality control of *Diplolepis rosae* RNAseq libraries used in RNAseq analysis.

Lib	Num pairs, Mln	Num bases, Gb	Effect, %	Error rate, %	Q20, %	Q30, %	GC cont, %	Per base GC cont	Per base seq cont	Per base N cont	Seq dup levels	Over seq	Adapt
Head_1	106.0	15.9	98.82	0.03	97.69	93.43	34.19	Fail	Fail	Pass	Fail	Warn	Pass
Head_2	106.0	15.9	98.33	0.03	97.52	93.00	34.26	Fail	Fail	Pass	Fail	Warn	Pass
Egg_1	99.66	14.9	97.64	0.03	97.34	92.79	35.52	Fail	Fail	Pass	Fail	Warn	Pass
Egg_3	91.81	13.8	98.21	0.03	97.73	93.47	35.37	Pass	Fail	Pass	Fail	Warn	Pass
Egg_3	11.74	1.76	98.12	0.03	96.59	90.82	35.15	Pass	Fail	Pass	Fail	Warn	Pass
Egg_4	97.91	14.7	98.76	0.03	97.45	93.00	35.99	Pass	Fail	Pass	Fail	Warn	Pass
Larva_ July_1	7.72	11.6	98.76	0.03	96.17	90.06	40.74	Fail	Fail	Pass	Fail	Warn	Pass
Larva_ July_1	94.44	14.2	98.75	0.03	97.45	92.93	40.96	Fail	Fail	Pass	Fail	Warn	Pass
Larva_ July_2	99.76	15.0	98.88	0.03	97.78	93.71	40.42	Warn	Fail	Pass	Fail	Warn	Pass
Larva_ July_3	81.15	12.2	98.92	0.03	97.65	93.40	39.06	Pass	Fail	Pass	Fail	Warn	Pass
Larva_ July_3	24.50	3.68	98.93	0.03	96.47	90.69	38.86	Pass	Fail	Pass	Fail	Warn	Pass
Larva_ July_4	28.50	4.28	98.79	0.03	97.37	92.89	42.47	Fail	Fail	Pass	Fail	Warn	Pass
Larva_ July_4	79.72	12.0	98.81	0.03	97.79	93.72	42.56	Fail	Fail	Pass	Fail	Warn	Pass
Larva_Sept_1	18.61	2.79	98.92	0.03	97.47	93.05	36.95	Fail	Fail	Pass	Fail	Warn	Pass
Larva_Sept_1	89.26	13.4	98.90	0.03	97.59	93.24	37.00	Fail	Fail	Pass	Fail	Warn	Pass

Larva_Sept_2	98.36	14.8	98.82	0.03	97.37	92.80	36.77	Fail	Fail	Pass	Fail	Warn	Pass
Larva_Sept_3	103.8	15.6	98.61	0.03	97.68	93.36	37.31	Warn	Fail	Pass	Fail	Warn	Pass
Larva_Sept_4	16.66	2.50	98.74	0.03	97.44	93.01	37.67	Warn	Fail	Pass	Fail	Warn	Pass
Larva_Sept_4	91.55	13.7	98.72	0.03	97.57	93.20	37.72	Warn	Fail	Pass	Fail	Warn	Pass
Gland_Oct_1	6.79	10.2	99.00	0.03	96.55	90.81	36.89	Warn	Fail	Pass	Fail	Warn	Pass
Gland_Oct_1	96.56	14.5	99.00	0.03	97.70	93.44	37.14	Warn	Fail	Pass	Fail	Warn	Pass
Gland_Oct_2	81.95	12.3	98.54	0.03	97.58	93.25	35.34	Fail	Fail	Pass	Fail	Warn	Pass
Gland_Oct_2	24.36	3.65	98.52	0.03	96.57	90.90	35.05	Fail	Fail	Pass	Fail	Warn	Pass
Gland_Oct_3	115.18	17.3	99.27	0.03	97.82	93.68	39.18	Warn	Fail	Pass	Fail	Warn	Pass
Gland_Oct_4	82.06	12.3	98.71	0.03	96.63	91.13	38.17	Warn	Fail	Pass	Fail	Warn	Pass
Gland_Oct_4	28.17	4.23	98.80	0.03	96.96	91.79	38.30	Warn	Fail	Pass	Fail	Warn	Pass
Gland_Nov_1	122.9	18.4	99.29	0.03	97.64	93.25	38.11	Warn	Fail	Pass	Fail	Pass	Pass
Gland_Nov_2	134.7	20.2	99.18	0.03	97.62	93.20	37.96	Pass	Fail	Pass	Fail	Pass	Pass
Gland_Nov_3	64.96	9.74	98.67	0.03	97.54	93.13	38.49	Pass	Fail	Pass	Fail	Pass	Pass
Gland_Nov_3	82.02	12.3	98.74	0.03	96.85	91.68	38.43	Pass	Fail	Pass	Fail	Pass	Pass
Gland_Nov_4	83.21	12.5	99.02	0.03	97.63	93.34	38.20	Warn	Fail	Pass	Fail	Pass	Pass

Gland_Nov_4	29.86	4.48	99.06	0.03	96.38	90.50	37.99	Warn	Fail	Pass	Fail	Pass	Pass
D_eglant_2	83.89	10.6	98.43	0.03	97.07	92.12	38.22	Warn	Fail	Pass	Fail	Warn	Pass
D_eglant_2	67.21	10.1	98.32	0.03	97.68	93.42	38.27	Warn	Fail	Pass	Fail	Warn	Pass
D_eglant_4	60.33	9.05	99.07	0.03	97.77	93.67	37.13	Warn	Fail	Pass	Fail	Warn	Pass
D_eglant_4	97.28	14.6	99.10	0.03	97.14	92.31	37.08	Warn	Fail	Pass	Fail	Warn	Pass

Head: *Diplolepis rosae* female adult head. Egg: *D. rosae* eggs removed from the female adults. Larva_July: *D. rosae* early larva removed from the gall in mid-July. Larva_Sept: *D. rosae* larva removed from the gall in early September. Gland_Oct: salivary glands from the *D. rosae* larva collected in October. Gland_Nov: salivary glands from the *D. rosae* larva collected in November. D_eglant: *D. eglanteriae*. Lib: RNAseq library. Num pairs: number of pairs of 150-bp pair-end reads. Num bases: total number of bases in the library. Effect: effective rate (100%*(clean reads / raw reads)). Error rate: base error rate. Q20: 100%*(base count of Qphred > 20 / total base count). Q30: 100%*(base count of Qphred > 30 / total base count). Q20 shows an error probability < 1/100. Q30 shows an error probability < 1/1000. GC cont: 100%*((G and C base count) / total base count). Per base GC cont: presence of contaminant sequences (Pass: GC distribution is near to theoretical normal distribution of an organism; Fail: presence of contaminant or overexpressed sequences). Per base seq cont: per base sequence content (always Fail in the RNA sequencing because of using random hexamer primers). Per base N cont: number of unidentified bases (Pass: the number of Ns is negligible). Seq dup levels: presence of duplicated sequences (Pass: < 20%; Warn (Warning): > 20%; Fail: > 50%) being overrepresented in the case of too little initial RNA quantity or too many PCR cycles. Over seq: sequences (at least 20 bp) occurring in more than 0.1% of the total number of sequences (overexpressed gene or adapter). Adapt: adapter sequences (Pass: removed by Novogene).

Table S10. Percentage of different classes of reads aligned to the annotated genes of *Diplolepis rosae*.

Sample	% Assigned	% Unassigned_ NoFeatures	% Unassigned_ Ambiguity	Total number of reads
Egg_1	46.57	52.40	1.03	105,692,304
Egg_3	50.60	48.40	1.00	109,586,716
Egg_4	53.18	45.81	1.01	99,616,178
Gland_Nov_1	66.81	31.92	1.27	128,332,522
Gland_Nov_2	66.20	32.46	1.34	141,362,650
Gland_Nov_3	69.45	29.06	1.49	151,175,994
Gland_Nov_4	68.03	30.54	1.43	117,059,536
Gland_Oct_1	53.06	45.92	1.02	115,042,208
Gland_Oct_2	38.72	60.47	0.81	137,425,368
Gland_Oct_3	64.30	34.44	1.26	119,692,118
Gland_Oct_4	59.03	39.80	1.17	118,450,874
Head_1	41.52	57.64	0.84	122,053,964
Head_2	43.12	56.02	0.86	120,933,152
Larva_July_1	61.56	36.68	1.76	98,612,530
Larva_July_2	56.60	41.80	1.60	101,374,936
Larva_July_3	63.17	35.61	1.22	105,756,354
Larva_July_4	65.36	32.47	2.17	103,313,482
Larva_Sept_1	50.27	48.57	1.16	134,714,808
Larva_Sept_2	49.08	49.78	1.14	122,851,446
Larva_Sept_3	49.95	48.60	1.45	115,339,990
Larva_Sept_4	51.67	47.02	1.31	127,945,160
D_eglant_2	56.10	24.16	1.25	151,090,542
D_eglant_4	50.84	26.23	1.08	157,615,770

Egg: *Diplolepis rosae* eggs removed from the female adults. Gland_Nov: salivary glands from the *D. rosae* larva collected in November. Gland_Oct: salivary glands from the *D. rosae* larva collected in October. Head: *D. rosae* female adult head. Larva_July: *D. rosae* early larva removed from the gall in mid-July. Larva_Sept: *D. rosae* larva removed from the gall in early September. D_eglant: *D. eglanteriae*. Assigned: reads successfully assigned to one or another feature (gene). Only this class of reads was taken into account in the differential expression analysis. Unassigned_NoFeatures: alignments that do not overlap any feature (gene). Unassigned_Ambiguity: alignments that overlap more than one feature (gene). In the categories 'Unassigned_Unmapped' (unmapped reads cannot be assigned), 'Unassigned_Read_Type' (reads showing unexpected, i.e. no-pair-end type), 'Unassigned_Singleton' (read pairs showing only one read mapped), 'Unassigned_MappingQuality' (reads showing the mapping quality score less than the threshold of 0 by default), 'Unassigned_Chimera' (two ends in a paired-end alignment are located on different chromosomes), 'Unassigned_FragmentLength' (alignments showing the length out of the min and max thresholds, i.e. 50 bp and 600 bp by default), 'Unassigned_Duplicate' (duplicated reads), 'Unassigned_MultiMapping' (reads mapping different genome positions), 'Unassigned_Secondary' (secondary alignments, i.e. that can match another genome location but with a less probability than the first one), 'Unassigned_NonSplit' (useful if the count of exon-spanning reads is required), 'Unassigned_Overlapping_Length' (minimum number of overlapping bases in the alignment required to be assigned, 1 by default), the number of alignments is 0 in all samples (Liao et al. 2014).

Table S11. Number of differentially expressed genes after performing pairwise comparisons between different stages and tissues of *Diplolepis rosae*.

Pair of samples / Counts	Total non-zero	Up	Down	Outliers	Low counts (mean count <)
Larva_July vs Egg	101,399	10,536	6,947	21	40,765 (< 3)
Larva_Sept vs Egg	91,907	7,202	4,655	723	36,697 (< 3)
Gland_Oct vs Egg	98,161	6,985	6,240	839	39,377 (< 3)
Gland_Nov vs Egg	95,520	10,787	19,151	174	32,782 (< 2)
Head vs Egg	90,432	4,077	4,007	2	41,215 (< 4)
Larva_Sept vs Larva_July	100,904	3,611	5,002	1,234	44,415 (< 3)
Larva_July vs Gland_Oct	105,466	13,840	7,733	1,293	38,441 (< 2)
Larva_July vs Gland_Nov	103,097	23,167	10,240	166	33,589 (< 2)
Larva_July vs Head	101,299	10,956	5,957	33	46,539 (< 3)
Larva_Sept vs Gland_Oct	93,784	8,111	7,580	2,256	35,688 (< 3)
Larva_Sept vs Gland_Nov	90,040	17,009	10,794	1486	27,337 (< 2)
Larva_Sept vs Head	84,961	5,929	2,496	243	40,143 (< 5)
Gland_Oct vs Gland_Nov	92,754	7,281	3,899	2,097	38,824 (< 3)
Head vs Gland_Oct	92,287	3,819	6,054	1,089	42,125 (< 4)
Head vs Gland_Nov	88,876	14,787	10,503	10	33712 (< 2)
Growth vs no_Growth	114,798	11,050	5,293	0	42745 (< 1)
Gall vs No_gall	114,625	4,827	2,039	1,403	48757 (< 2)

Egg: *Diplolepis rosae* eggs removed from the female adults. Gland_Nov: salivary glands from the *D. rosae* larva collected in November. Gland_Oct: salivary glands from the *D. rosae* larva collected in October. Head: *D. rosae* female adult head. Larva_July: *D. rosae* larva removed from the gall in mid-July. Larva_Sept: *D. rosae* larva removed from the gall in early September. No_growth: sample set which does not correspond to the active gall growth stage (Egg + Head + Gland_Oct + Gland_Nov). Growth: sample set corresponding to the active gall growth stage (Larva_July + Larva_Sept + D_eglant, i.e. mid-July *D. eglanteriae* larva). No_gall: sample set which does not correspond to the whole gall stage (Egg + Head): Gall: sample set corresponding to the whole gall stage (Larva_July + Larva_Sept + Gland_Oct + Gland_Nov + D_eglant, mid-July *D. eglanteriae* larva). Total non-zero: number of genes that reads mapped at least one time. Up: number of up-regulated genes, i.e. whose expression is more in the first *D. rosae* tissue/stage compared to the second one in the first *D. rosae* tissue/stage compared to the second one. The number of up-regulated genes was defined by positive log2(read counts in the first tissue/read counts of the second tissue) (p adj < 0.05) (**supplementary fig. S25-S27**). Down: number of down-regulated genes, i.e. whose expression is less in the first *D. rosae* tissue/stage compared to the second one. The number of down-regulated genes was defined by negative log2(read counts in the first tissue/read counts of the second tissue) (p adj < 0.05) (**supplementary fig. S25-S27**). Outliers: number of genes where reads show outlier counts. Low counts (mean count <): number of genes where a number of mapped reads is less than the given threshold. The total number of examined genes was 125,626.

Table S12. Functional annotation and selection coefficients of the genes encoding proteins with leucine-rich repeats, plants cell wall degrading enzymes, and venom-like proteins that were overexpressed during gall formation from mid-July to early November in *Diplolepis rosae*.

Gene id	Annot lev	Func annot eggNOG	MK (*s*)	N tr	L, aa	blastp sp align	blastp match	% cov	% id	E value	tot sc
g29097	Hym	Leucine-rich repeat	-3,34 (neg)	2	776	*Belonocnema* sp.	Slit homolog	97	77.92	0	1068
g39136*	Hym	Leucine-rich repeat	-1.27 (neg)	2	498	*Belonocnema* sp.	Slit homolog	84	47.31	2e-142	446
g51958*	Hym	Leucine-rich repeat	-1.19 (neg)	1	723	*Belonocnema* sp.	Podocan	97	69.08	0	911
g53074	Hym	Leucine-rich repeat	-0.80 (neg)	2	1234	*Belonocnema* sp.	Chaoptin-Like	100	85.33	0	2100
g79908	Hym	Leucine-rich repeat	-0.92 (neg)	2	247	*Belonocnema* sp.	U2 small ribonucleo-protein A'	100	90.69	3e-162	459
g86458	Hym	Leucine-rich repeat	-1.29 (neg)	2	298	*Belonocnema* sp.	Peroxidasin-like	95	69.18	4e-129	380
g86477	Hym	Leucine-rich repeat	-1.55 (neg)	2	351	*Belonocnema* sp.	Protein phosphatase 1 regulatory subunit 42-like	100	72.75	0	533
g87196	Hym	Leucine-rich repeat	-1.09 (neg)	1	1351	*Leptopilina* sp.	Chaoptin	99	80.43	0	2149
g108407*	Hym	Leucine-rich repeat	-2.05 (neg)	1	699	*Leptopilina* sp.	Vasorin -like	34	66.39	2e-101	330
g108407*	Hym	Leucine-rich repeat	-2.05 (neg)	1	699	*Leptopilina* sp.	Neuronal protein 1-like	34	66.39	6e-94	311
g108411*	Hym	Leucine-rich repeat	-1.20 (neg)	1	133	*Harpegnathos* sp.	Neuronal protein 2	79	67.92	9e-46	165
g120908*	Hym	Leucine-rich repeat	-1.69 (neg)	2	568	*Belonocnema* sp.	Carboxypeptidase N subunit 2-like	95	65.81	0	701
g120908*	Hym	Leucine-rich repeat	-1.69 (neg)	2	568	*Nomia* sp.	Toll-like receptor 7	97	40.04	3e-125	389

g122472	Insect	Leucine-rich repeat	-1.34 (neg)	1	383	*Lasius* sp.	Slit-like	37	61.74	1e-50	186
g122552	Hym	Leucine-rich repeat	-1.34 (neg)	2	611	*Leptopilina* sp.	Insulin-like growth factor-binding protein	100	78.57	0	993
g122552	Hym	Leucine-rich repeat	-1.45 (neg)	2	611	*Belonocnema* sp.	Carboxy-peptidase N subunit 2	100	84.90	0	1059
g123069	Hym	Leucine-rich repeat	-0.75 (neg)	2	902	*Belonocnema* sp.	Fibronectin type-III domain containing protein	99	89.64	0	1583
g123069	Hym	Leucine-rich repeat	-0.75 (neg)	2	902	*Leptopilina* sp.	Vasorin	98	86.49	0	1553
g124026	Hym	Leucine-rich repeat	-0.74 (neg)	1	927	*Leptopilina* sp.	Follicle-stimulating hormone receptor	98	84.95	0	1612
g31505	GPA	Rhamno-galacturonate lyase	-1.33 (neg)	1	595	*Belonocnema* sp.	Rhamno-galacturonate lyase-like	97	54.02	0	630
g51222	GPA	Pectate lyase	-1.17 (neg)	1	337	*Belonocnema* sp.	Pectin lyase-like	89	48.68	6e-99	305
g51223	GPA	Pectate lyase	-1.22 (neg)	1	264	*Belonocnema* sp.	Pectin lyase-like	99	50.00	4e-47	254
g81883	GPA	Cellulase	-1.43 (neg)	3	373	*Belonocnema* sp.	Endo glucanase Z-like	80	65.22	1e-139	410
g87016	GPA	Pectate lyase	-1.60 (neg)	1	330	*Belonocnema* sp.	Pectin lyase-like	90	63.67	9e-134	393
g94279	GPA	Cellulase	-1.21 (neg)	1	512	*Belonocnema* sp.	Endo glucanase Z-like	60	71.61	2e-163	476
g30759	Hym	Venom protease-like / Trypsin	-3.69 (neg)	1	187	*Leptopilina* sp.	Venom serine protease Bi-VSP-like	61	55.65	2e-32	130
g30762	Hym	Venom protease-like / Trypsin	-1.84 (neg)	1	245	*Belonocnema* sp.	Venom serine protease Bi-VSP-like	40	72.28	2e-42	155
g30762	Hym	Venom protease-like / Trypsin	-1.84 (neg)	1	245	*Leptopilina* sp.	Venom serine protease Bi-VSP-like	46	62.07	2e-40	153
g49748 *	Hym	Venom acid phosphatase / Histidine phosphatase 2	-1.26 (neg)	1	302	*Belonocnema* sp.	Venom acid phosphatase Acph-1-like	99	72.09	2e-162	467
g49748 *	Hym	Venom acid phosphatase / Histidine phosphatase 2	-1.26 (neg)	1	302	*Leptopilina* sp.	Venom acid phosphatase Acph-1-like	99	67.00	5e-151	438

g67338 *	Hym	Venom acid phosphatase / Histidine phosphatase 2	-1.43 (neg)	1	375	*Belonocnema* sp.	Venom acid phosphatase Acph-1-like	100	61.27	8e-170	489
g67338 *	Hym	Venom acid phosphatase / Histidine phosphatase 2	-1.43 (neg)	1	375	*Leptopilina* sp.	Venom acid phosphatase Acph-1-like	99	56.23	3e-149	437
g74434 *	Hym	Venom acid phosphatase / Histidine phosphatase 2	-1.44 (neg)	1	192	*Leptopilina* sp.	Venom acid phosphatase Acph-1-like	81	48.72	4e-45	163
g74436 *	Hym	Venom acid phosphatase / Histidine phosphatase 2	-1.63 (neg)	1	214	*Belonocnema* sp.	Venom acid phosphatase Acph-1-like	57	56.10	9e-39	147
g74436 *	Hym	Venom acid phosphatase / Histidine phosphatase 2	-1.63 (neg)	1	214	*Leptopilina* sp.	Venom acid phosphatase Acph-1-like	53	54.78	5e-37	147
g60751 *	Hym	Tetraspanin	-0.78 (neut)	2	290	*Leptopilina* sp.	CD63 antigen-like	100	82.07	3e-168	478
g121032 *	Meta	Domain of unknown function	-1.85 (neg)	1	119	*Trichonephila clavipes*	Putative transposase	85	57.84	5e-36	130
g28790	Hym	Phosphatidylethanolamine-binding protein	-1.40 (neg)	3	208	*Belonocnema* sp.	Protein D2-like	100	86.12	4e-134	385
g53036	Hym	Chitooligosaccharidolytic beta-N-acetylglucosaminidase	-3.36 (neg)	2	599	*Leptopilina* sp.	Chitooligosaccharidolytic beta-N-acetylglucosaminidase	100	68.50	0	896
g53036	Hym	Chitooligosaccharidolytic beta-N-acetylglucosaminidase	-3.36 (neg)	2	599	*Nasonia vitripennis*	Chitooligosaccharidolytic beta-N-acetylglucosaminidase	100	61.53	0	808
g53036	Hym	Chitooligosaccharidolytic beta-N-acetylglucosaminidase	-3.36 (neg)	2	599	*Venturia canescens*	Chitooligosaccharidolytic beta-N-acetylglucosaminidase	96	62.76	0	806
g90912	Hym	Apyrase	-1.41 (neg)	1	131	*Belonocnema* sp.	Soluble calcium-activated nucleotidase 1	90	75.00	8e-58	194
g90912	Hym	Apyrase	-1.41 (neg)	1	131	*Leptopilina* sp.	Soluble calcium-activated nucleotidase 1	90	67.50	1e-51	178
g90912	Hym	Apyrase	-1.41 (neg)	1	131	*Leptopilina* sp.	Apyrase	90	69.17	4e-51	177
g118636	Hym	Inosine-uridine preferring nucleoside hydrolase	-1.89 (neg)	2	342	*Belonocnema* sp.	Probable uridine nucleosidase 2	97	63.06	2e-149	434
g118639	Hym	Lysozyme C	-1.56 (neg)	1	144	*Belonocnema* sp.	Lysozyme C1-like	96	74.82	7e-75	230
g121541 *	Insect	Testicular haploid expressed repeat	-1.90 (neg)	1	130	*Leptopilina* sp.	Testicular haploid expressed protein-like	79	69.90	5e-36	136

g36882	Hym	Group XII secretory phospho-lipase A2 precursor	-0.80 (neg)	1	264	*Leptopilina* sp.	Group XII secretory phospho-lipase A2	77	82.24	1e-126	369
g36882	Hym	Group XII secretory phospho-lipase A2 precursor	-0.80 (neg)	1	264	*Belonocnema* sp.	Group XII secretory phospho-lipase A2	77	83.33	2e-124	363
g36882	Hym	Group XII secretory phospho-lipase A2 precursor	-0.80 (neg)	1	264	*Nasonia vitripennis*	Group XII secretory phospho-lipase A2	79	78.95	9e-121	353
g95508	Hym	Phospho-lipase A2-like	-1.29 (neg)	2	189	*Belonocnema* sp.	Phospho-lipase A2-like	95	62.43	7e-83	255
g95508	Hym	Phospho-lipase A2-like	-1.29 (neg)	2	189	*Leptopilina* sp.	Phospho-lipase A2-like	97	59.46	4e-79	245
g124077*	Hym	Phospho-lipase A2	-1.23 (neg)	1	230	*Belonocnema* sp.	Phospho-lipase A2	86	76.50	4e-111	328
g124077*	Hym	Phospho-lipase A2	-1.23 (neg)	1	230	*Leptopilina* sp.	Phospho-lipase A2-like	84	73.71	6e-102	305
g124077*	Hym	Phospho-lipase A2	-1.23 (neg)	1	230	*Cyphomyrmex costatus*	Predicted: phospho-lipase A2 A2-actitoxin-Usc2a	83	64.62	9e-91	277
g124077*	Hym	Phospho-lipase A2	-1.23 (neg)	1	230	*Colletes gigas*	Phospho-lipase A2 A2-actitoxin-Usc2a	85	62.50	1e-90	276
g53492*	Hym	AB hydrolase superfamily: lipase family	-1.02 (neg)	2	612	*Leptopilina* sp.	Lipase-3-like	63	69.82	0	606
g53493*	Hym	AB hydrolase superfamily: lipase family	-0.68 (neut)	2	429	*Belonocnema* sp.	Lipase-3-like	85	83.11	0	950
g53493*	Hym	AB hydrolase superfamily: lipase family	-0.68 (neut)	2	429	*Leptopilina* sp.	Lipase-3-like	99	71.96	0	659
g71703	Hym	AB hydrolase superfamily: lipase family	-1.46 (neg)	2	448	*Leptopilina* sp.	Lipase-3-like	93	67.38	0	613
g71703	Hym	AB hydrolase superfamily: lipase family	-1.46 (neg)	2	448	*Nasonia vitripennis*	Lipase-3-like	85	55.93	4e-151	447
g71708*	Hym	AB hydrolase superfamily: lipase family	-1.35 (neg)	1	402	*Belonocnema* sp.	Lipase-3-like	96	54.64	1e-177	513
g123552	Hym	AB hydrolase superfamily: lipase family	-1.45 (neg)	1	286	*Leptopilina* sp.	Lipase-3-like	100	64.34	2e-139	410
g52886	Hym	Lipase	-1.40 (neg)	1	433	*Leptopilina* sp.	Pancreatic lipase-related protein-1-like	86	59.36	2e-169	491

g52886	Hym	Lipase	-1.40 (neg)	1	433	*Orussus abietinus*	Pancreatic triacyl glycerol lipase	84	54.59	6e-150	441
g74954 *	Hym	Lipase	-1.07 (neg)	3	313	*Belonocnema* sp.	Lipase member H	100	92.97	0	617
g74954 *	Hym	Lipase	-1.07 (neg)	3	313	*Leptopilina* sp.	Phospho-lipase A1 memeber A	100	88.33	0	588
g74954 *	Hym	Lipase	-1.07 (neg)	3	313	*Odontomachus brunneus*	Pancreatic lipase-related protein-2-like	100	84.98	0	573
g94397 *	Hym	Lipase	-1.46 (neg)	1	355	*Belonocnema* sp.	Pancreatic triacyl glycerol lipase	93	67.98	4e-175	500
g94397 *	Hym	Lipase	-1.90 (neg)	1	355	*Leptopilina* sp.	Pancreatic lipase-related protein-1-like	89	45.99	4e-93	293
g120373	Hym	Lipase	-1.46 (neg)	1	388	*Belonocnema* sp.	Pancreatic triacyl glycerol lipase-like	96	78.46	0	630
g120373	Hym	Lipase	-1.46 (neg)	1	388	*Leptopilina* sp.	Pancreatic triacyl glycerol lipase-like	98	72.92	0	600
g124088	Hym	Animal haem peroxidase	-1.39 (neg)	1	698	*Leptopilina* sp.	Peroxidase-like	100	68.62	0	998
g85953	Hym	Animal haem peroxidase	-1.61 (neg)	1	1312	*Belonocnema* sp.	Peroxidase	51	89.36	0	1321
g118333	Hym	Animal haem peroxidase	-1.22 (neg)	1	693	*Leptopilina* sp.	Peroxidasin homolog	99	83.21	0	1238
g119756 *	Hym	Animal haem peroxidase	-1.33 (neg)	1	1280	*Belonocnema* sp.	Peroxidasin	100	81.74	0	2259

Annot lev: annotation level. Hym: Hymenoptera. GPA: Gammaproteobacteria. Insect: Insecta. Met: Metazoa. Func annot eggNOG: functional annotation provided by eggnog-mapper v. 2.1.7 (Cantalapiedra et al. 2021; Huerta-Cepas et al. 2021). MK (*s*): selection coefficient estimated by SnIPRE, McDonald-Kreitman type analysis (Eilertson et al. 2012); neg: negative selection; neut: neutral selection. N tr: number of expressed gene transcripts. L, aa: protein length, amino acids. Blastp sp align: taxonomic name of the species showing protein sequences matching to the query *Diplolepis rosae* sequence. % cov: query cover, proportion of the query *D. rosae* sequence that is aligned with the database sequence. % id: percent of identity between the query *D. rosae* sequence and the database sequence. E val: expectation value of the alignment between the query *D. rosae* sequence and the database sequence. Tot sc: total score reflecting the strength of the match between the query *D. rosae* sequence and the database sequence. The query *D. rosae* protein and the database protein showing E value < 10e-06 and total score < 50 were considered homologous (Pearson 2013). Blue columns indicate genes encoding proteins with leucine rich repeat (Zhao et al. 2015). Yellow columns indicate genes encoding plant cell wall degrading enzymes (Hearn et al. 2019). Red columns indicate genes encoding proteins found in the *D. rosae* venom glands (Cambier et al. 2019). Green columns indicate the *D. rosae* genes orthologous to the genes overexpressed in the *B. pallida* venom glands (Cambier et al. 2019). The asterisks (*) denotes genes highly expressed in the mid-July and early September *D. rosae* larvae.

Table S13. Characteristics of *Cynips quercusfolii* samples used in Illumina sequencing.

Sample name	Stage	Num Ind	Tissue	Sex	Host plant	Isol source	Collect date	Lat	Long
18c	-	1	Whole Ind	Female	*Quercus robur*	Gall	2019-08-28	48.695	2.037
60a	-	1	Whole Ind	Female	*Quercus robur*	Gall	2019-10-15	48.428333	2.512222
305a	-	1	Whole Ind	Female	*Quercus petraea*	Gall	2020-08-30	48.474959	2.777347
359c	Imago	1	Whole Ind	Female	*Quercus pubescens*	Gall	2020-09-08	45.41458	3.131403
490	Imago	1	Whole Ind	Female	*Quercus sp.*	Gall	2020-09-26	48.671951	2.063188
638b	-	1	Whole Ind	Female	*Quercus pubescens*	Gall	2020-10-20	45.314087	1.834971
705.5	-	1	Whole Ind	Female	*Quercus sp.*	Gall	2021-10-16	49.297035	0.626649

Num ind: number of individuals per sample. Whole Ind: whole individual. Isol source: isolation source. Collect date: collection date. Lat: latitude. Long: longitude.

Table S14. Structural variations identified in scaffold 57 (NC_046657.1) and scaffold 62 (NC_046662.1) of the *C. quercusfolii* assembled genome.

Deletions

60a

	chromosome	pos1	pos2	size
1	NC_046662.1	5698245	5740661	42416
2	NC_046662.1	110383318	117385992	7002674

638b

	chromosome	pos1	pos2	size
1	NC_046657.1	668622	29437416	28768794
2	NC_046662.1	110383315	117385971	7002656

705.5

	chromosome	pos1	pos2	size
1	NC_046662.1	79369564	79467827	98263
2	NC_046662.1	110383355	117386020	7002665

18c

	chromosome	pos1	pos2	size
1	NC_046657.1	668622	29437413	28768791
2	NC_046662.1	110383361	117385978	7002617

305a

	chromosome	pos1	pos2	size
1	NC_046657.1	112305912	114370091	2064179
2	NC_046662.1	110383355	117385965	7002610

359c

	chromosome	pos1	pos2	size
1	NC_046657.1	76564617	85365605	8800988
2	NC_046657.1	112305926	114370089	2064163
3	NC_046657.1	162659659	166436375	3776716
4	NC_046662.1	21869455	21885762	16307
5	NC_046662.1	71284786	75619888	4335102
6	NC_046662.1	110383376	117386006	7002630

490

	chromosome	pos1	pos2	size
1	NC_046662.1	110383313	117386005	7002692

Duplications

60a

	chromosome	pos1	pos2	size
1	NC_046662.1	111910195	137726172	25815977

638b

	chromosome	pos1	pos2	size
1	NC_046657.1	126428286	169568683	43140397
2	NC_046662.1	111910193	137726173	25815980

705.5

	chromosome	pos1	pos2	size
1	NC_046657.1	126428269	169568683	43140414
2	NC_046662.1	23218392	23241109	22717
3	NC_046662.1	111910192	137726172	25815980

18c

	chromosome	pos1	pos2	size
1	NC_046657.1	19973721	19986641	12920
2	NC_046657.1	126428281	169568682	43140401
3	NC_046662.1	111910195	137726173	25815978

305a

chromosome pos1 pos2 size 1 NC_046657.1 167006191 176308381 9302190 2 NC_046662.1 132088919 137917122 5828203 **359c** chromosome pos1 pos2 size 1 NC_046657.1 167006181 176308366 9302185 2 NC_046662.1 80620599 80634191 13592 3 NC_046662.1 110384577 117386794 7002217 **490** NULL
Inversions
60a chromosome pos1 pos2 size 1 NC_046657.1 5940792 52689078 46748286 2 NC_046662.1 18972863 55404158 36431295 3 NC_046662.1 137725975 142949282 5223307 **638b** chromosome pos1 pos2 size 1 NC_046657.1 5940739 52689070 46748331 2 NC_046657.1 113030510 152739218 39708708 3 NC_046662.1 74819707 130256120 55436413 **705.5** chromosome pos1 pos2 size 1 NC_046657.1 5940792 52689080 46748288 **18c** chromosome pos1 pos2 size 1 NC_046657.1 5940792 52689080 46748288 2 NC_046657.1 158849119 169568615 10719496 3 NC_046662.1 137725979 142949280 5223301 **305a** chromosome pos1 pos2 size 1 NC_046662.1 81987894 90150503 8162609 2 NC_046662.1 137725979 142949282 5223303 **359c** chromosome pos1 pos2 size 1 NC_046662.1 7888604 7903135 14531 2 NC_046662.1 81987902 90150509 8162607 3 NC_046662.1 137725973 142949282 5223309 **490** chromosome pos1 pos2 size 1 NC_046657.1 169568618 176310189 6741571 2 NC_046662.1 81987894 90150503 8162609 3 NC_046662.1 137725979 142949282 5223303

Chromosome: scaffold. Pos 1 and pos 2: start and end positions of the variation, bp. Size: variation length, bp.

Fig.s

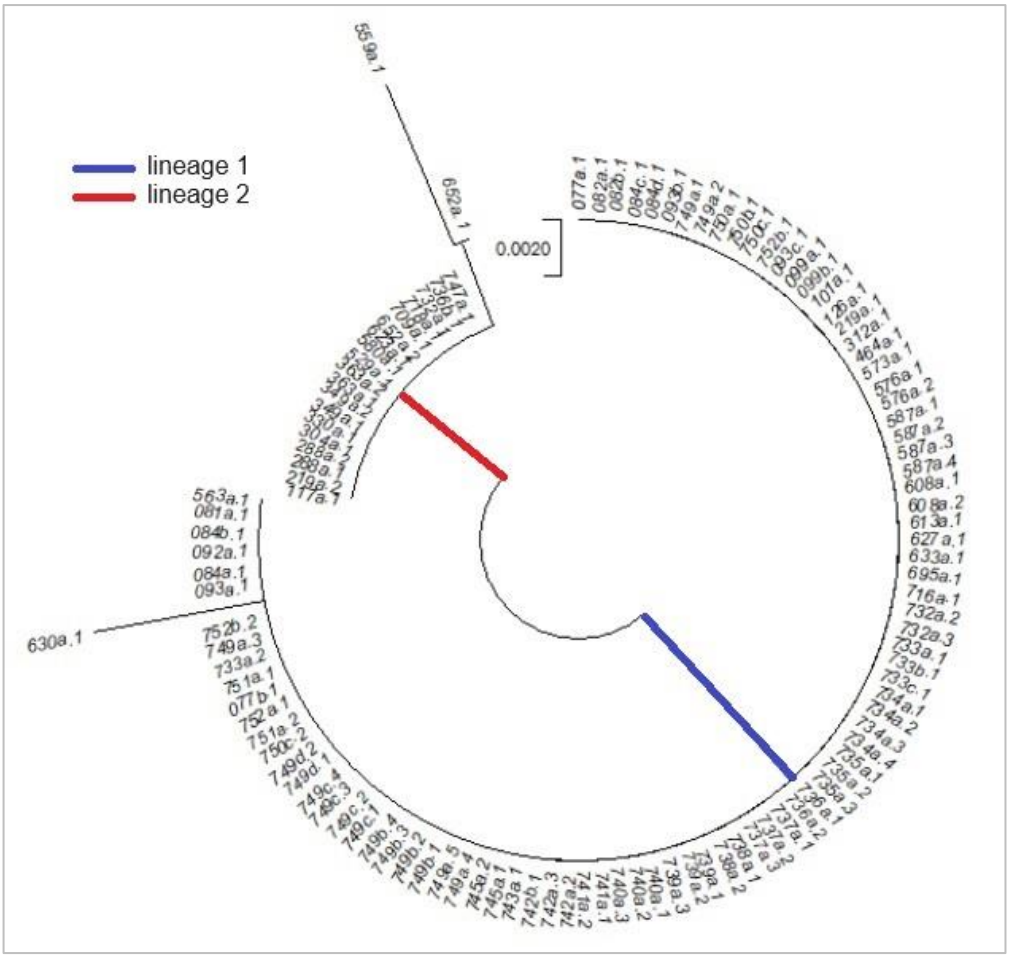


Fig. S1. Phylogram of *Diplolepis rosae* genotypes collected in France based on one marker. Two branches correspond to the two *D. rosae* lineages.

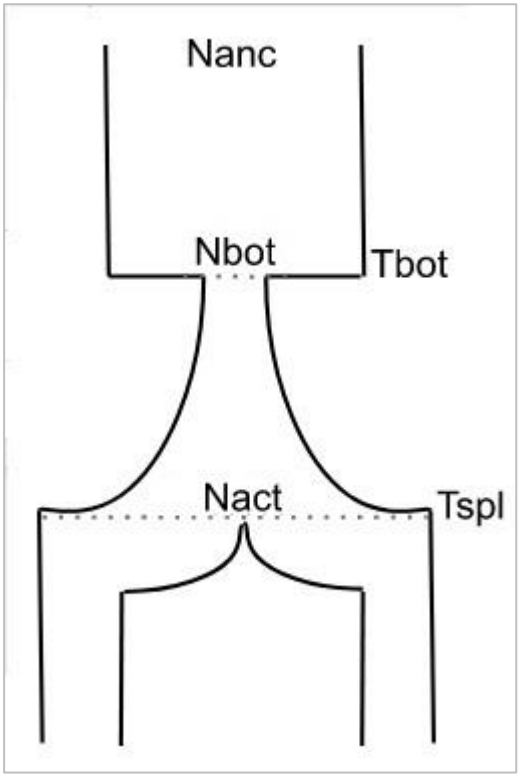


Fig. S2. Schema describing the best-fitted model for the demographic scenario of *Diplolepis rosae*: a bottleneck of an ancestral population followed by exponential growth then split into two populations with no gene flow between them. Nanc, Nbot, Nact: population sizes (ancestral, after the bottleneck, just before the split, respectively); Tbot, Tspl: times (bottleneck, split, respectively). Line sizes are scaled relatively to *dadi* estimates: Nbot/Nanc = 0.13; Nact/Nanc = 2.4; bottleneck time = 0.22; split time = 0.20; time unit is 2*Neff generations ago; Neff: effective population size (diploid individuals).

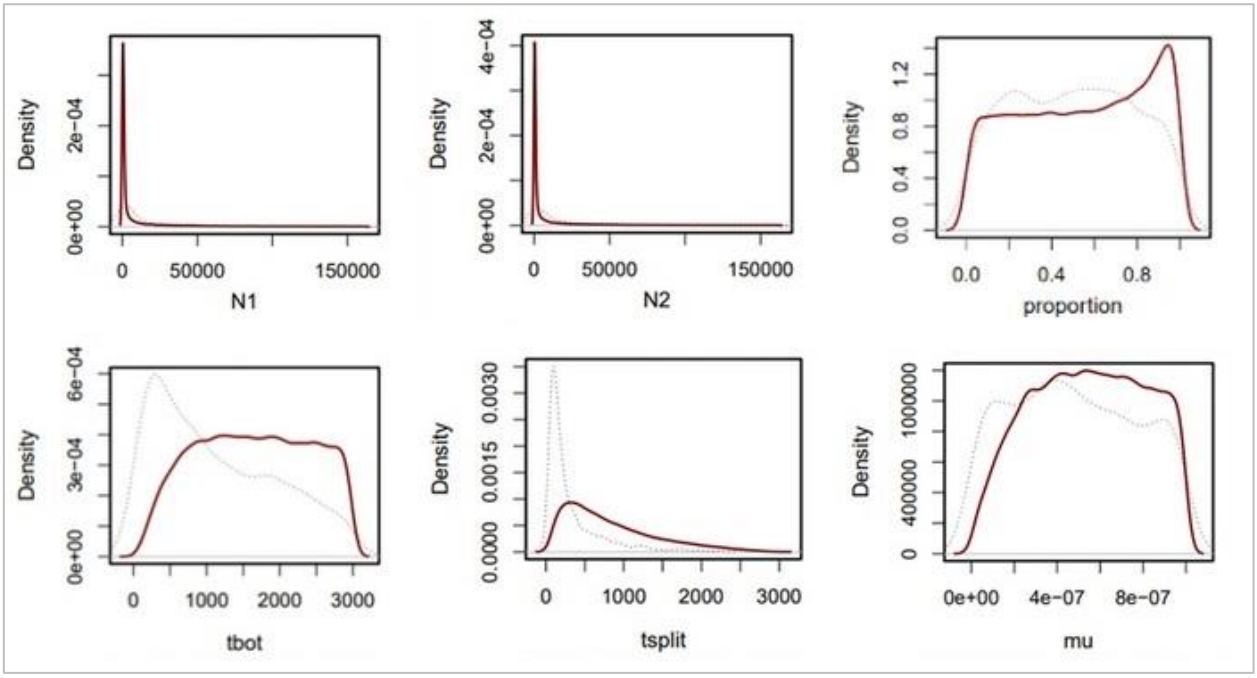


Fig. S3. Parameter inference of simulated demographic model "Bottleneck_growth_split" in *Diplolepis rosae* using approximate Bayesian computation (ABC). Grey line: given distribution of simulated model parameters (priors). Red line: distribution accepted model parameters (posteriors). N1, N2: effective population sizes of lineage 1 and lineage 2, respectively; proportion: ratio between a number of individuals of *D. rosae* from a population after a bottleneck to an ancestral one; tbot, tsplit: bottleneck and split time, respectively; mu: mutation rate.

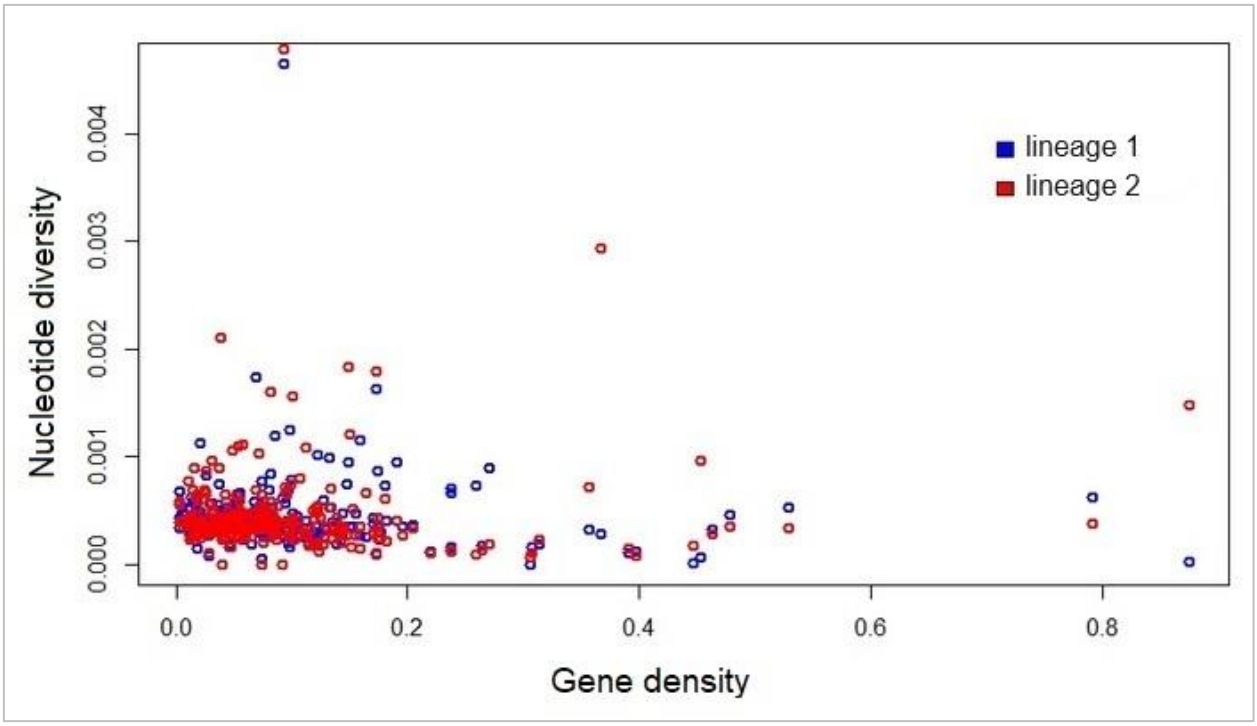


Fig. S4. Relation between gene density (the proportion of nucleotides assigned to a protein coding sequence in a genome fragment) and nucleotide diversity (π) in the two lineages of *Diplolepis rosae*.

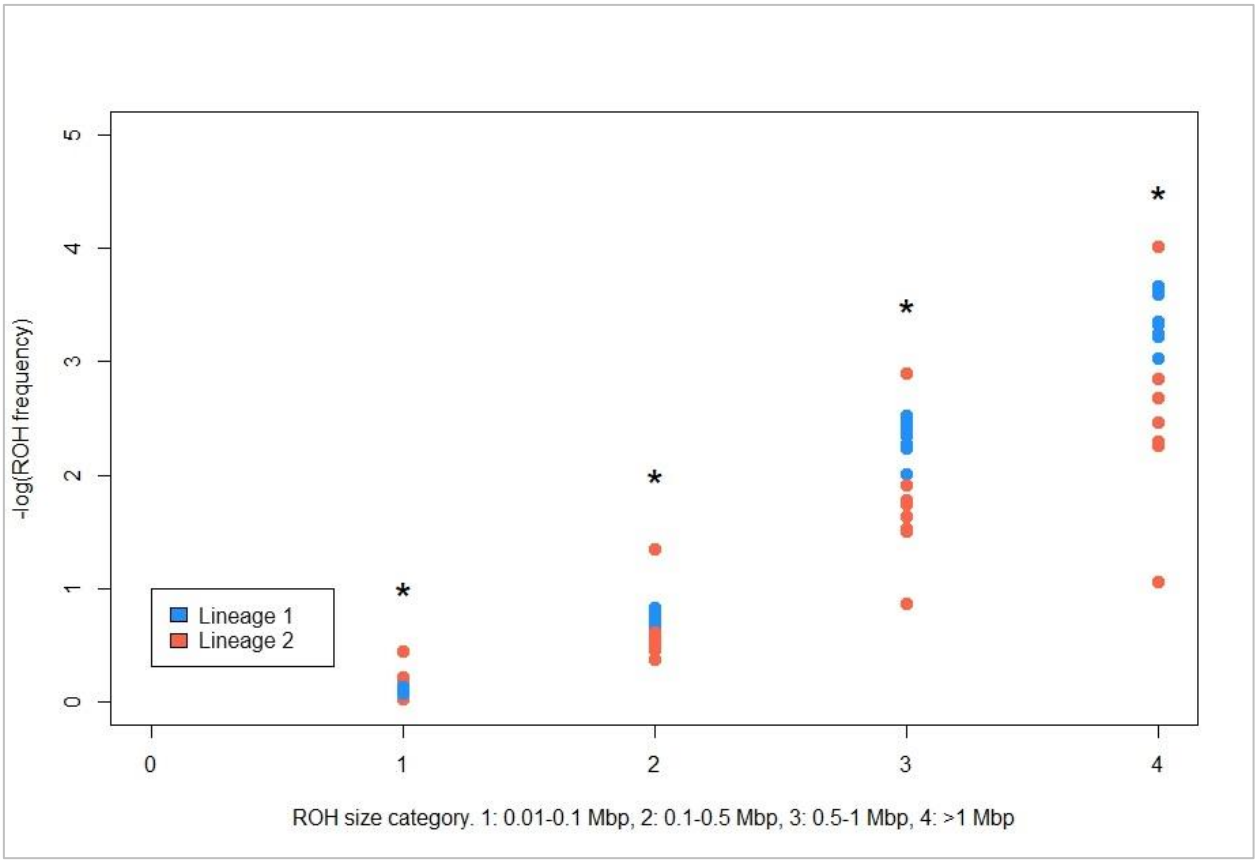


Fig. S5. Frequency of runs of homozygosity (ROH) of different size categories in the *Diplolepis rosae* genome. Each point represents one *D. rosae* individual. The ROH frequency is presented as the ratio between the number of ROHs from each size category and the total number of ROHs on a negative logarithmic scale. The asters indicate a significant difference in ROH frequencies between lineage 1 and lineage 2 (Mann-Whitney U test, $p < 0.05$).

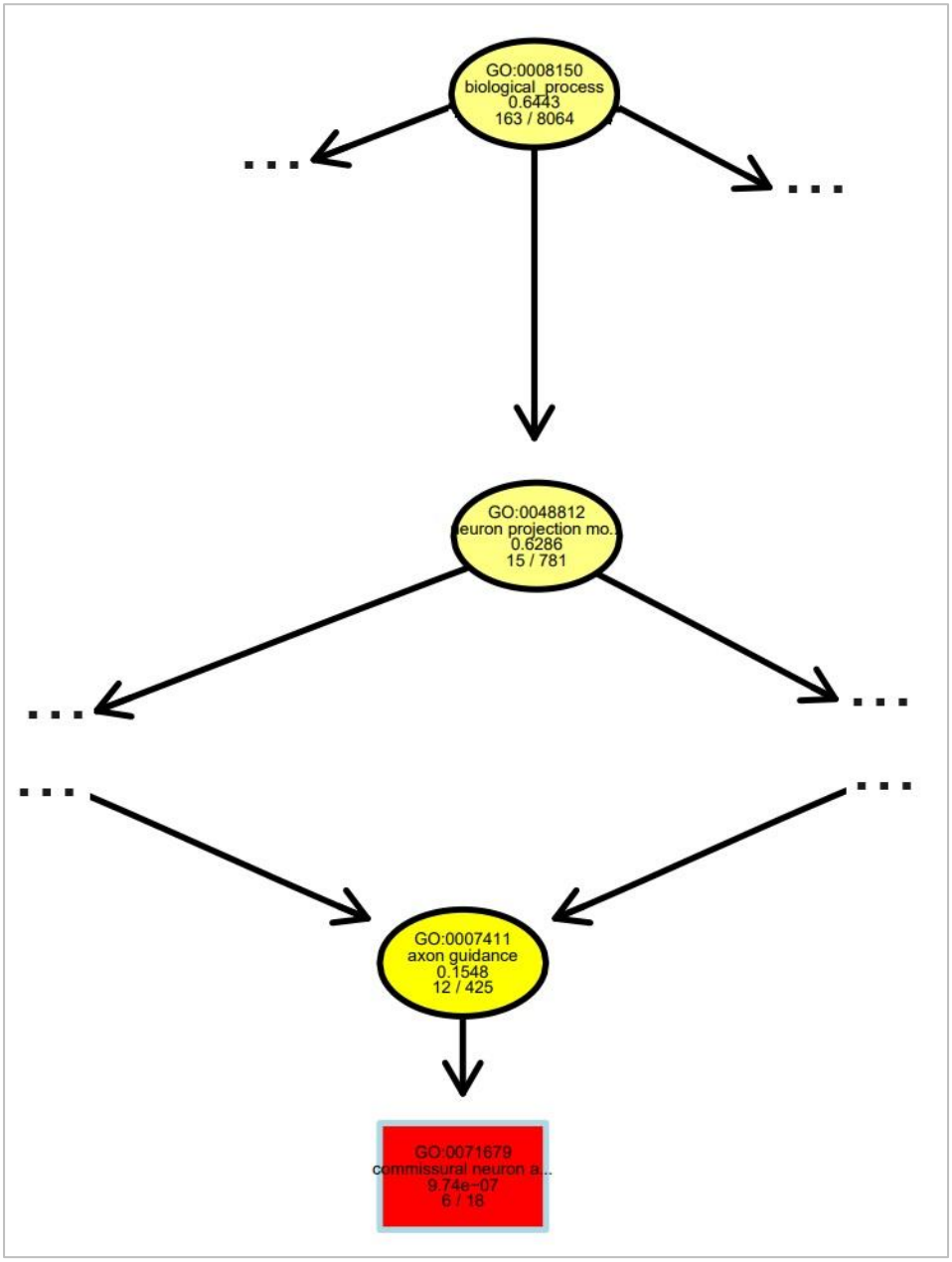


Fig. S6. Partial relation in the 'Biological process' ontology between gene terms showing a positive outlier composite score in *Diplolepis rosae* lineage 1. Results are presented as Gene ontology (GO) term, annotation, raw p-value (Fisher's exact test), number of detected significant genes in *D. rosae* / total number of genes in the full gene set.

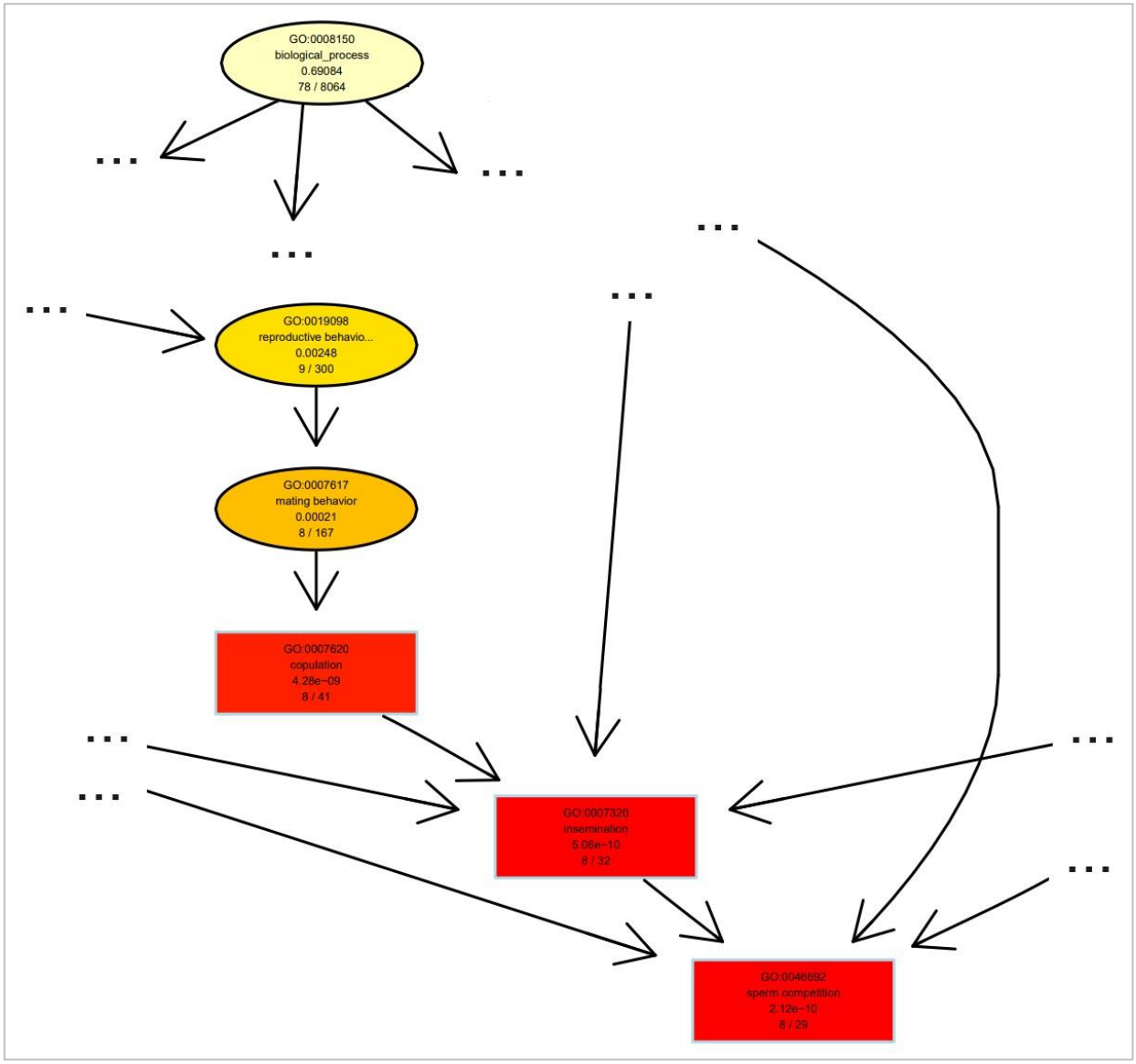


Fig. S7. Relation in the 'Biological process' ontology between gene terms showing a negative outlier composite score in *Diplolepis rosae* lineage 1. Results are presented as Gene ontology (GO) term, annotation, raw p-value (Fisher's exact test), number of detected genes in *D. rosae* / total number of genes in the full gene set.

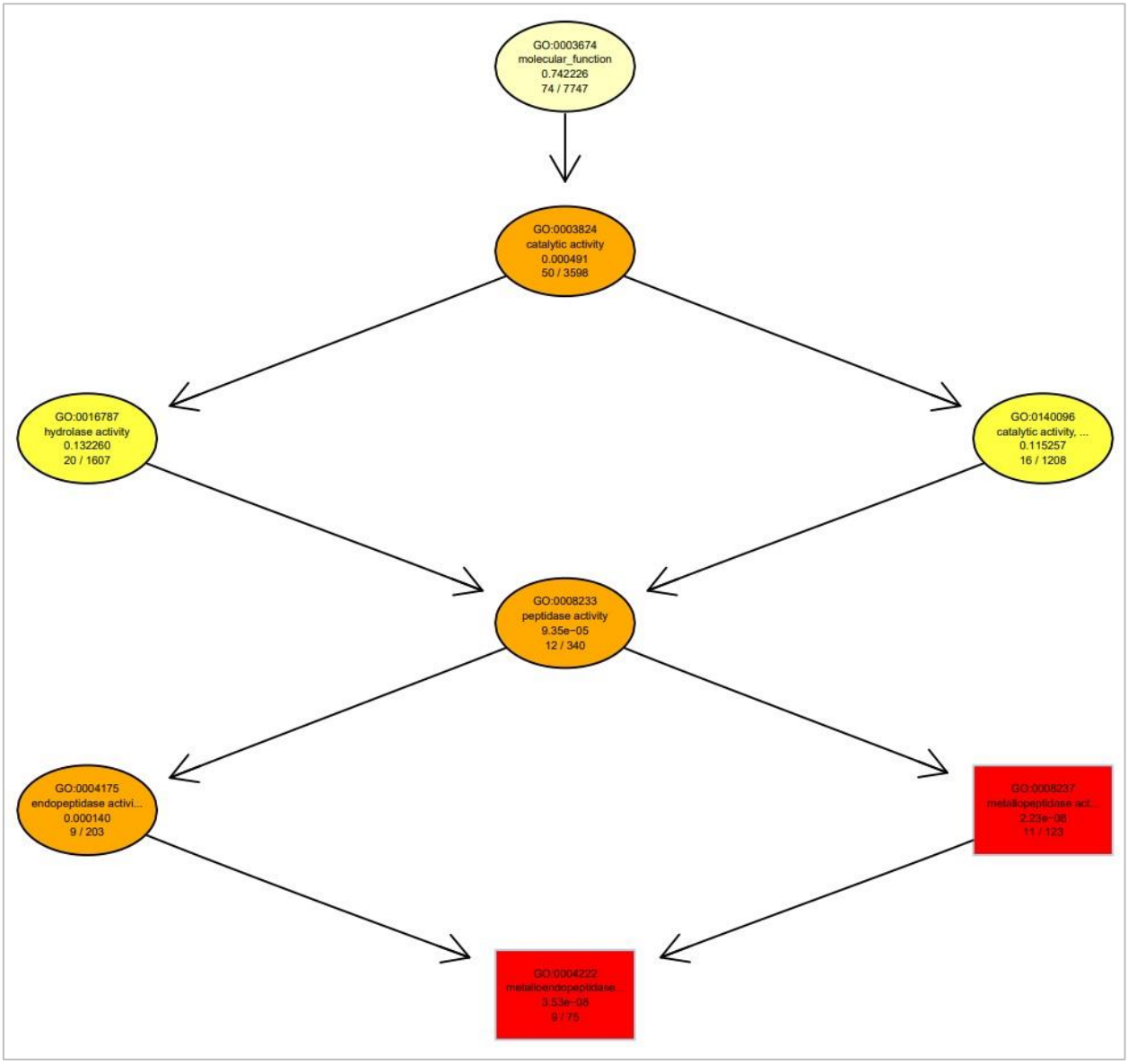


Fig. S8. Relation in the 'Molecular Function' ontology between gene terms showing a negative outlier composite score in *Diplolepis rosae* lineage 1. Results are presented as Gene ontology (GO) term, annotation, raw p-value (Fisher's exact test), number of detected genes in *D. rosae* / total number of genes in the full gene set.

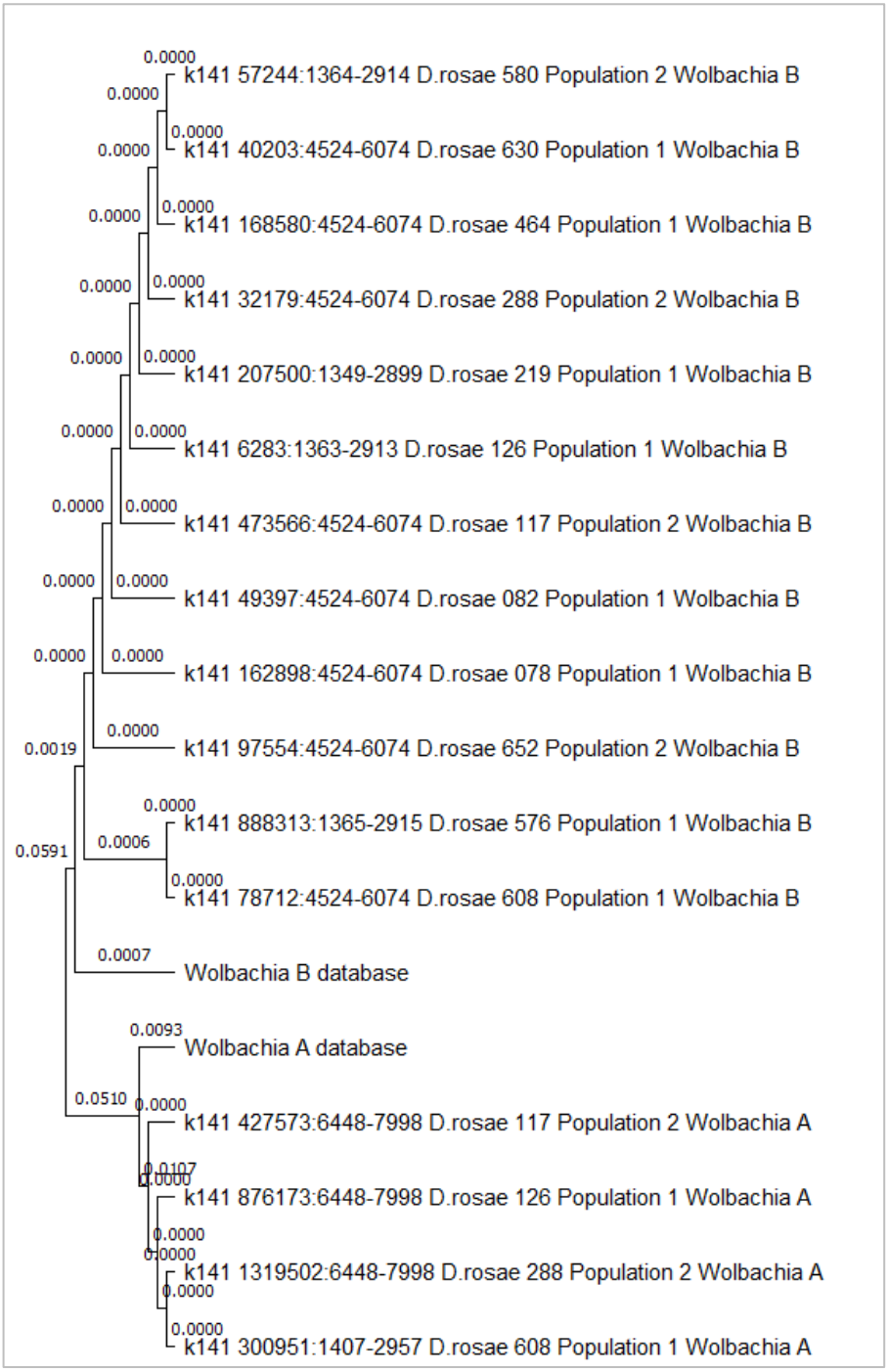


Fig. S9. **Phylogram of *Wolbachia* genotypes (*coxA* gene, see details in Wang et al. 2020) recovered from the *Diplolepis rosae* Illumina sequences.** Each sample corresponds to *Wolbachia* contig identifier: start – end position of the gene_corresponding *D.rosae* sample name_*D. rosae* lineage_*Wolbachia* supergroup. The phylogram was constructed using the Maximum likelihood statistical method (Tamura-Nei substitution model (default)) by MEGA 11 (Tamura et al. 2021). Population 1 and population 2 refer to the *D. rosae* lineage 1 and lineage 2, respectively.

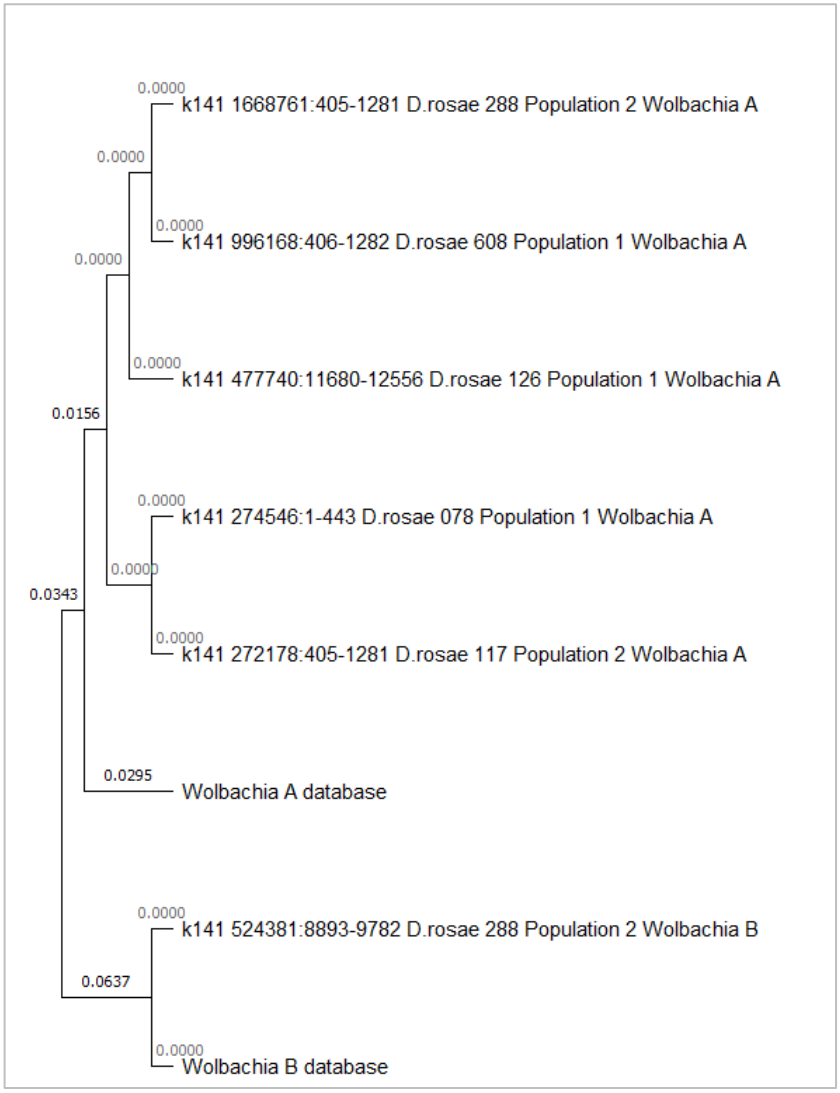


Fig. S10. Phylogram of *Wolbachia* genotypes (*fbpA* gene, see details in Wang et al. 2020) recovered from *Diplolepis rosae* Illumina sequences. Each sample corresponds to *Wolbachia* contig identifier: start – end position of the gene_corresponding *D.rosae* sample name_*D. rosae* lineage_*Wolbachia* supergroup. The phylogram was constructed using the Maximum likelihood statistical method (Tamura-Nei substitution model (default)) by MEGA 11 (Tamura et al. 2021). Population 1 and population 2 refer to the *D. rosae* lineage 1 and lineage 2, respectively.

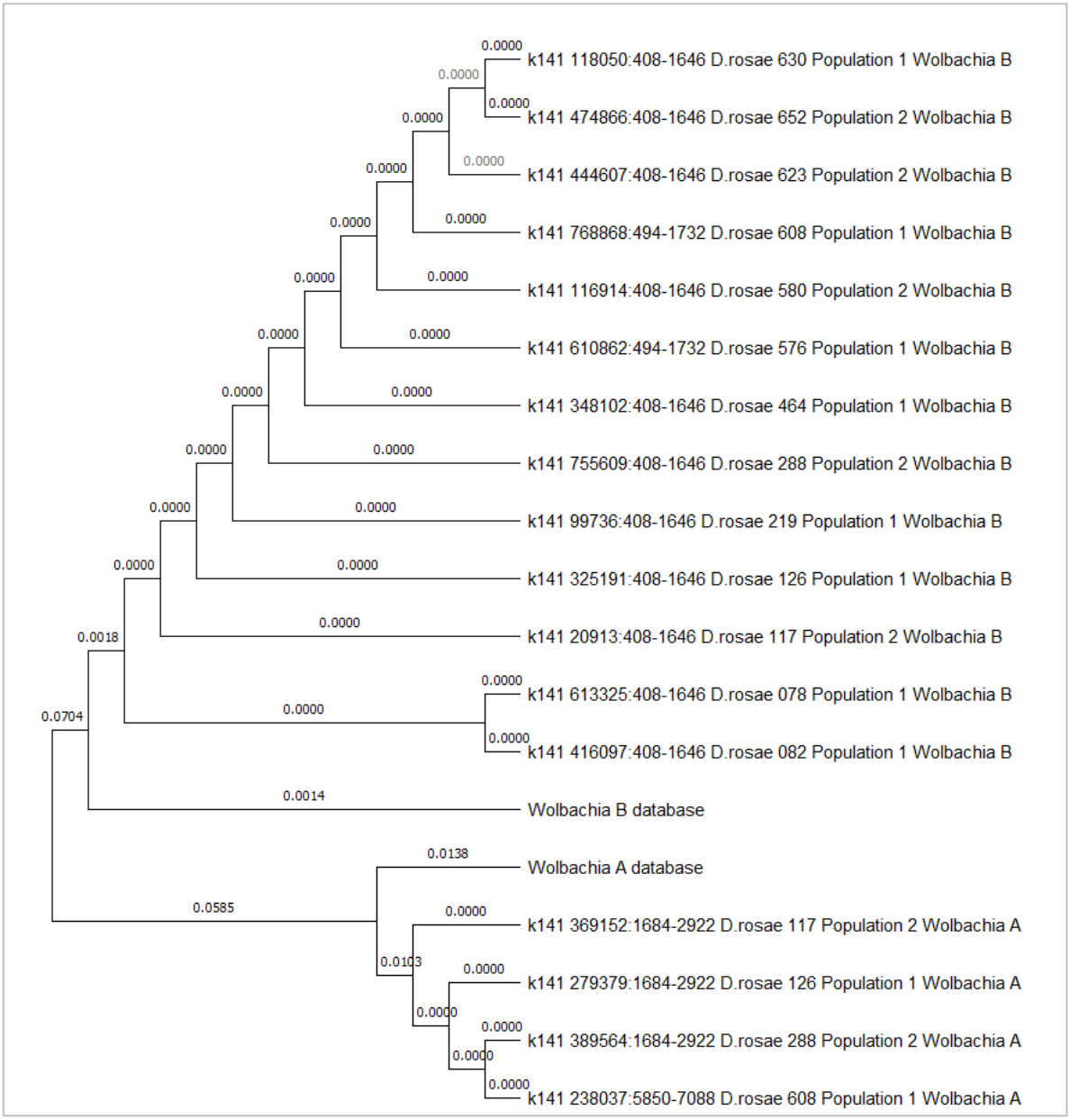


Fig. S11. **Phylogram of *Wolbachia* genotypes (*ftsZ* gene, see details in Wang et al. 2020) recovered from *Diplolepis rosae* Illumina sequences**. Each sample corresponds to *Wolbachia* contig identifier: start – end position of the gene_corresponding *D.rosae* sample name_*D. rosae* lineage_*Wolbachia* supergroup. The phylogram was constructed using the Maximum likelihood statistical method (Tamura-Nei substitution model (default)) by MEGA 11 (Tamura et al. 2021). Population 1 and population 2 refer to the *D. rosae* lineage 1 and lineage 2, respectively.

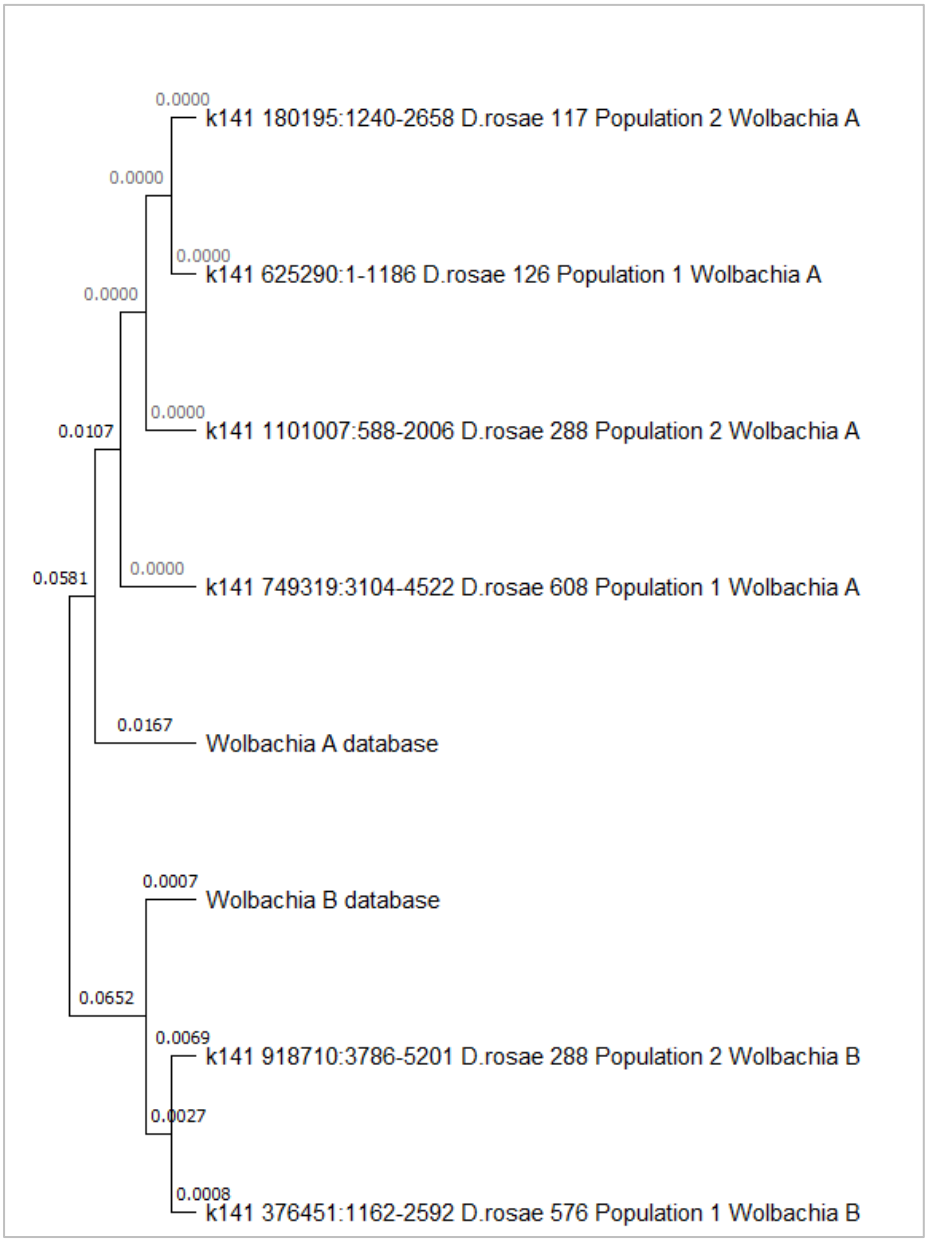


Fig. S12. **Phylogram of *Wolbachia* genotypes (*gatB* gene, see details in Wang et al. 2020) recovered from *Diplolepis rosae* Illumina sequences.** Each sample corresponds to *Wolbachia* contig identifier: start – end position of the gene_corresponding *D.rosae* sample name_*D. rosae* lineage_*Wolbachia* supergroup. The phylogram was constructed using the Maximum likelihood statistical method (Tamura-Nei substitution model (default)) by MEGA 11 (Tamura et al. 2021). Population 1 and population 2 refer to the *D. rosae* lineage 1 and lineage 2, respectively.

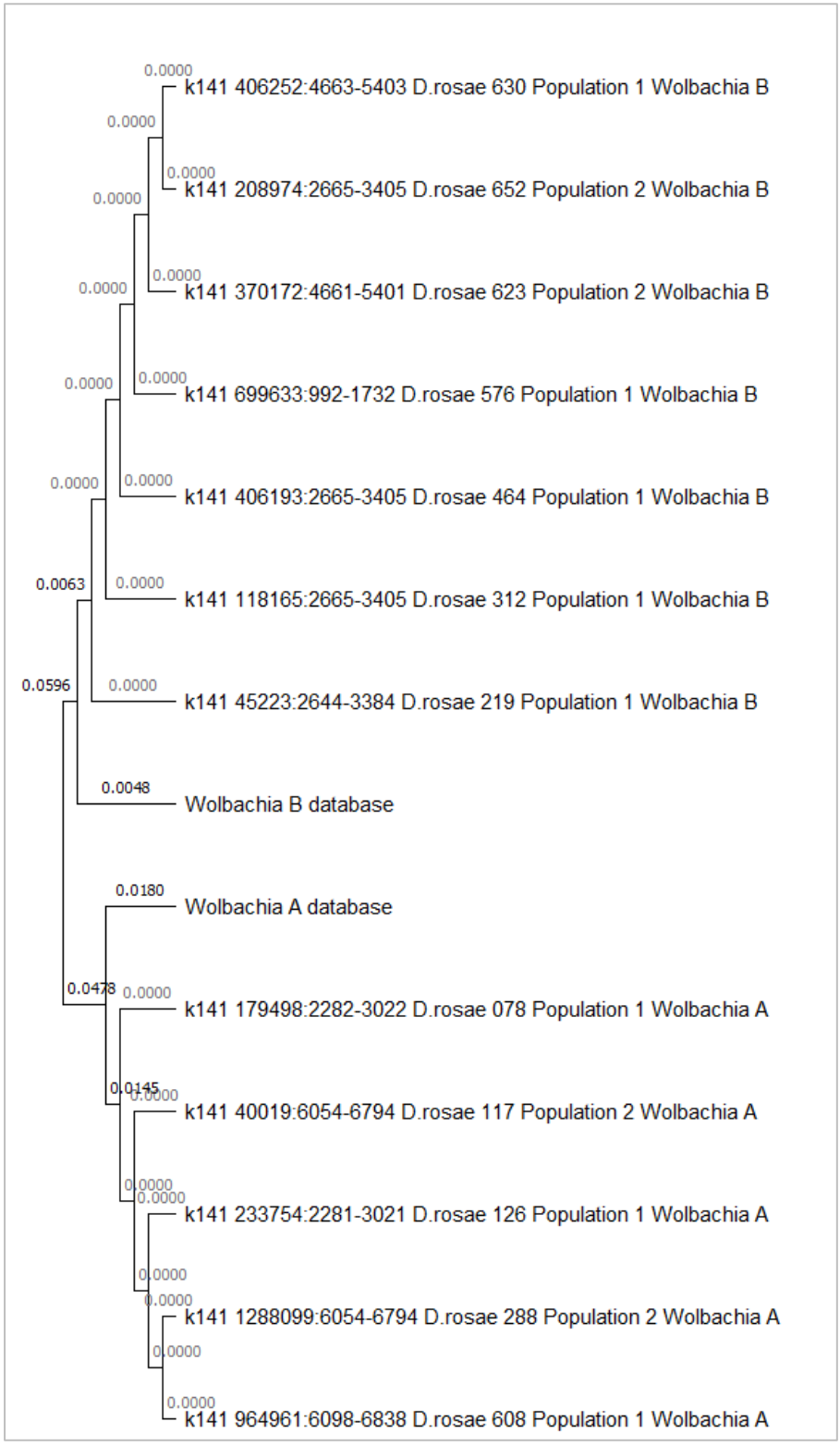


Fig. S13. **Phylogram of *Wolbachia* genotypes (*hcpA* gene, see details in Wang et al. 2020) recovered from *Diplolepis rosae* Illumina sequences.** Each sample corresponds to *Wolbachia* contig identifier: start – end position of the gene_corresponding *D.rosae* sample name_*D. rosae* lineage_*Wolbachia* supergroup. The phylogram was constructed using the Maximum likelihood statistical method (Tamura-Nei substitution model (default)) by MEGA 11 (Tamura et al. 2021). Population 1 and population 2 refer to the *D. rosae* lineage 1 and lineage 2, respectively.

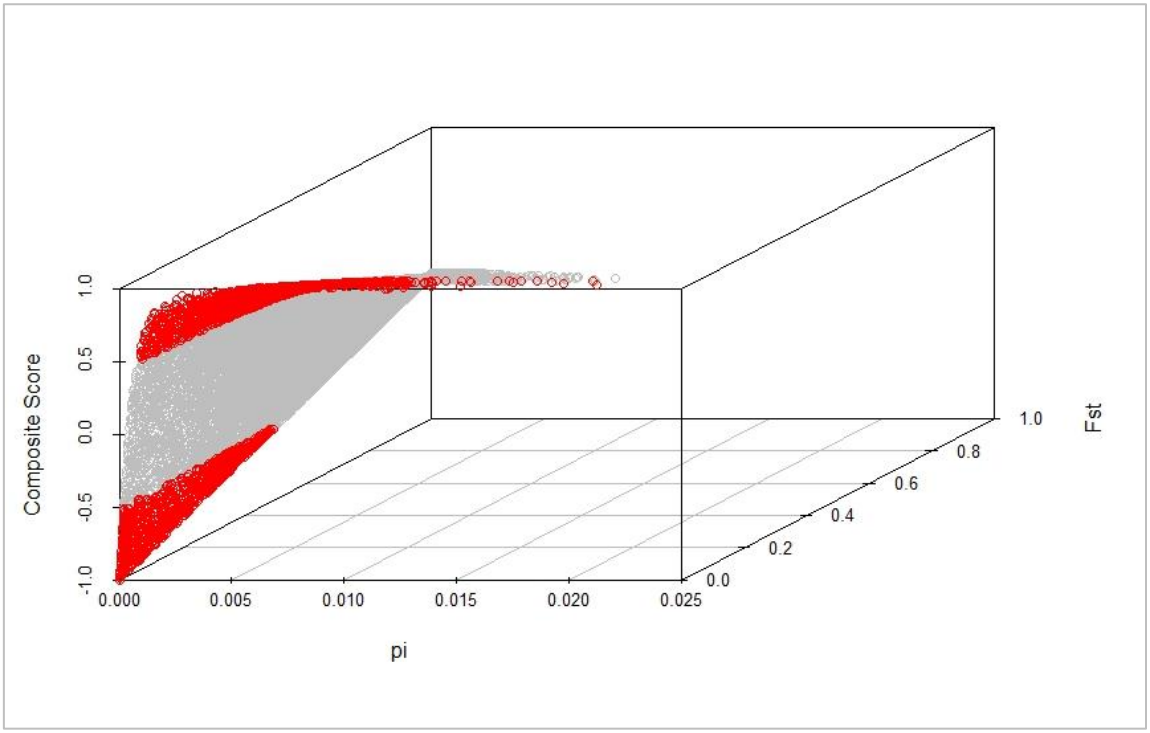


Fig. S14. Relation between nucleotide diversity (pi), the fixation index Fst, and a composite score (CS) equal to (1 - Fst)*2(F(pi) - 0.5), where F(pi) is cumulative distribution function. Red points represent CS outliers below -0.5 and above 0.5 corresponding to the genome regions of *Diplolepis rosae* examined in gene set enrichment analysis.

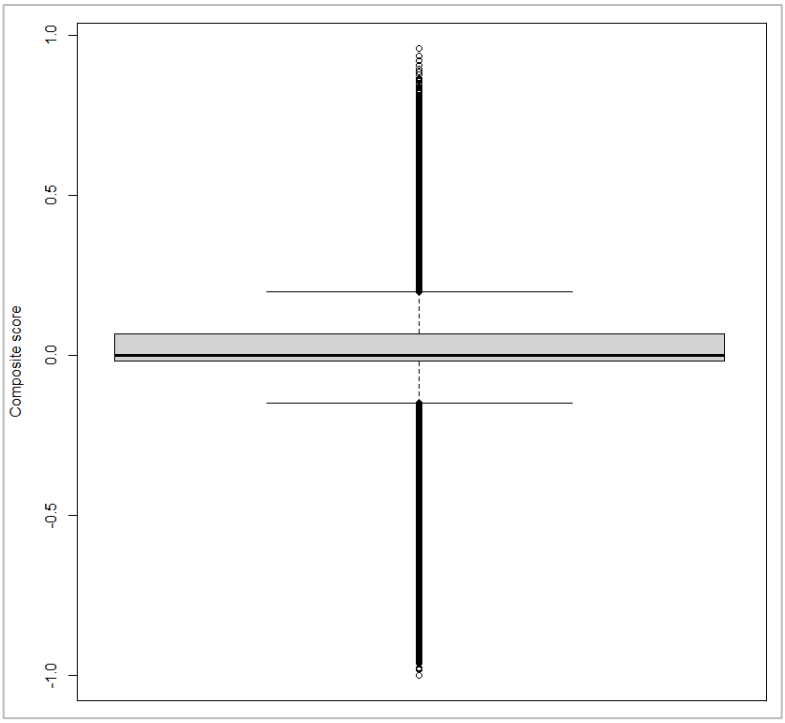


Fig. S15. **Distribution of composite score (CS) values calculated for 10-kbp windows in the *Diplolepis rosae* genome.** CS equals (1 - Fst)*2(F(π) - 0.5), where Fst is the fixation index calculated between the two *D. rosae* populations, π is nucleotide diversity calculated for each *D. rosae* lineage, and F(π) is cumulative distribution function.

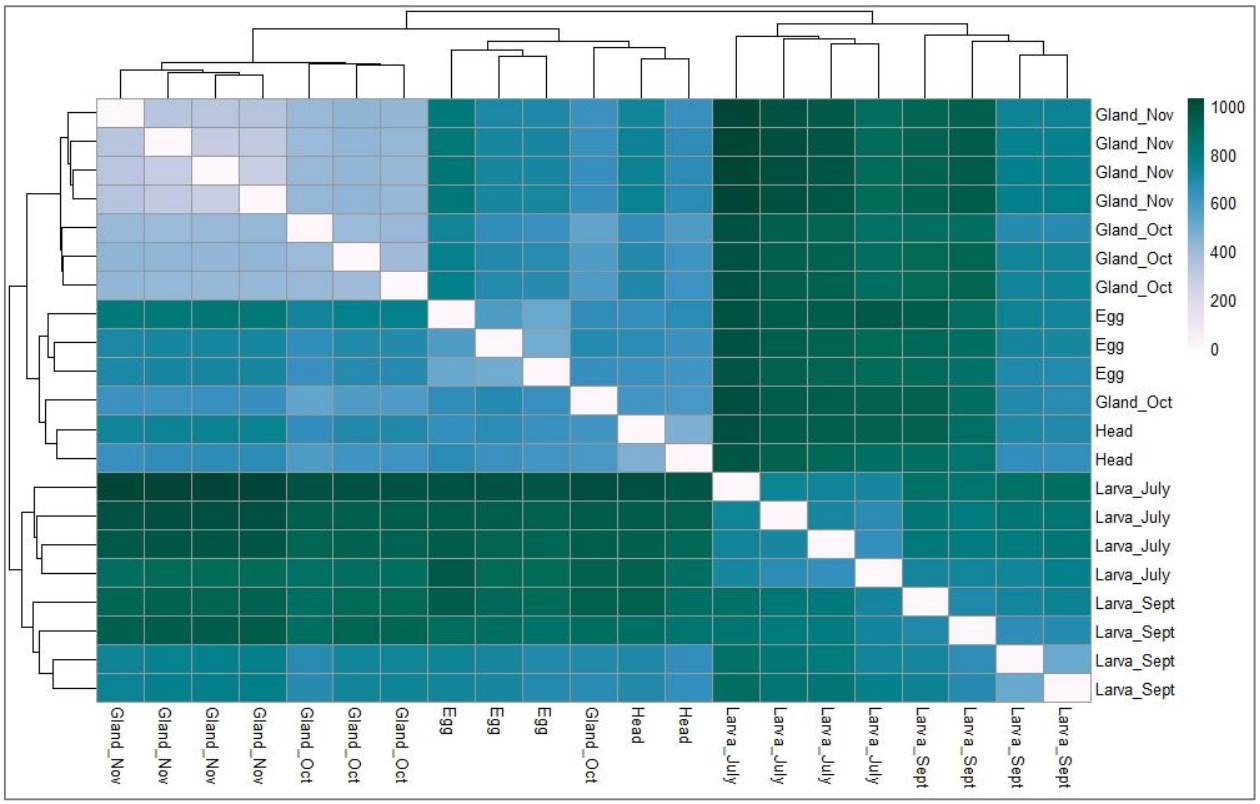


Fig. S16. **Pairwise Euclidean distances between the *Diplolepis rosae* samples used in the relative differential gene expression analysis.** Egg: egg removed from a female adult. Larva_July: mid-July larva. Larva_Sept: early September larva. Gland_Oct: October larva salivary glands. Gland_Nov: November larva salivary glands. Head: female adult head. The clustering is based on the distances between the rows/columns of the distance matrix.

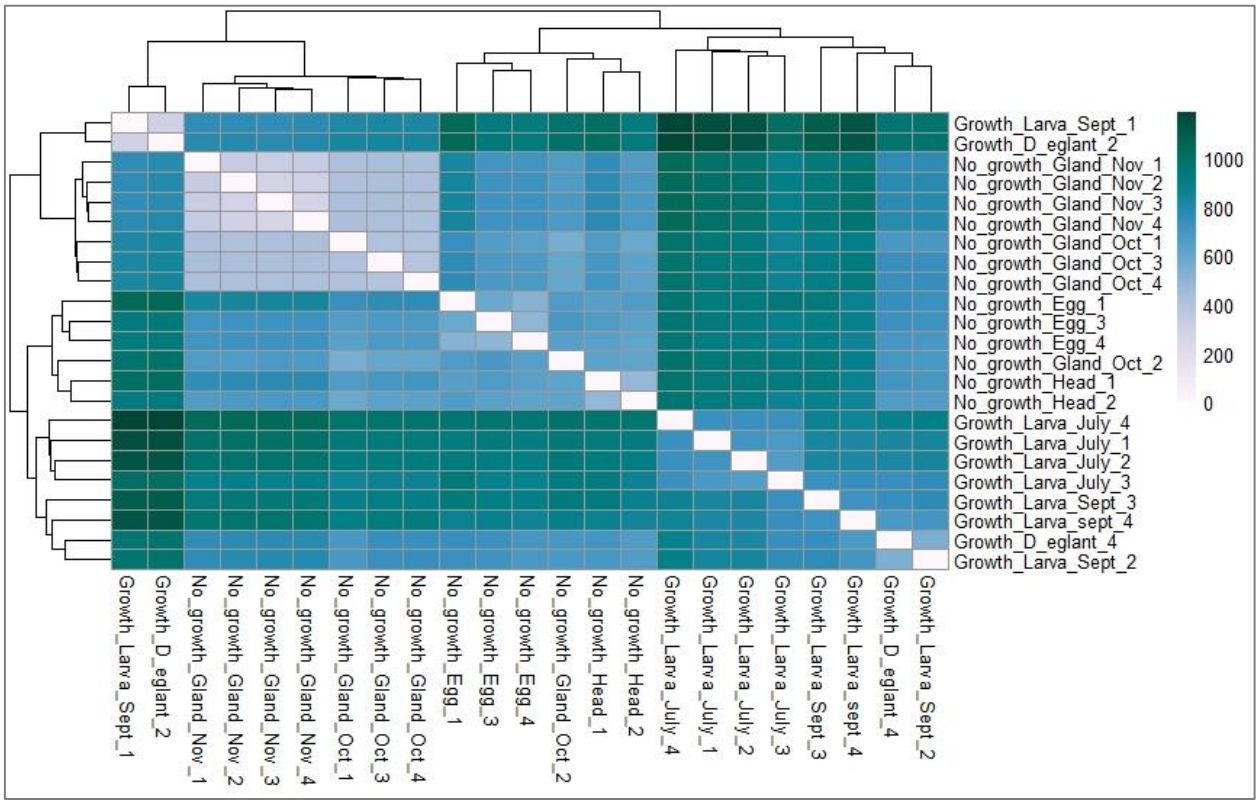


Fig. S17. **Pairwise Euclidean distances between the No_growth' and 'Growth' *Diplolepis rosae* samples used in the relative differential gene expression analysis.** No_growth: sample set that does not correspond to the active gall growth stage (egg removed from a female adult + female adult head + October larva salivary gland + November larva salivary gland). Growth: sample set corresponding to the active gall growth stage (mid-July larva + early September larva + mid-July *Diplolepis eglanteriae* larva). The clustering is based on the distances between the rows/columns of the distance matrix.

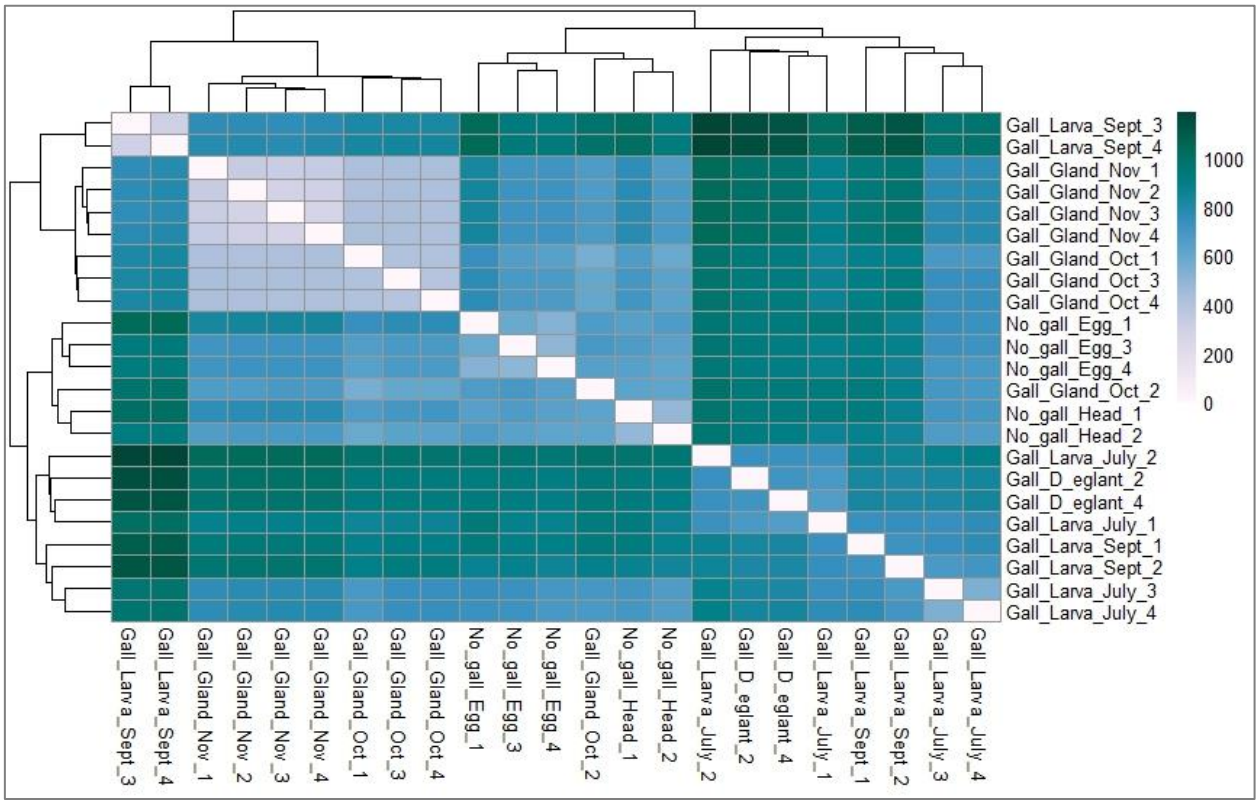


Fig. S18. **Pairwise Euclidean distances between the No_gall' and 'Gall' *Diplolepis rosae* samples used in the relative differential gene expression analysis**. No_gall: sample set which does not correspond to the whole gall stage (egg removed from a female adult + female adult head): Gall: sample set corresponding to the whole gall stage (mid-July larva + early September larva + mid-July *D. eglanteriae* larva + October salivary gland + November salivary gland). The clustering is based on the distances between the rows/columns of the distance matrix.

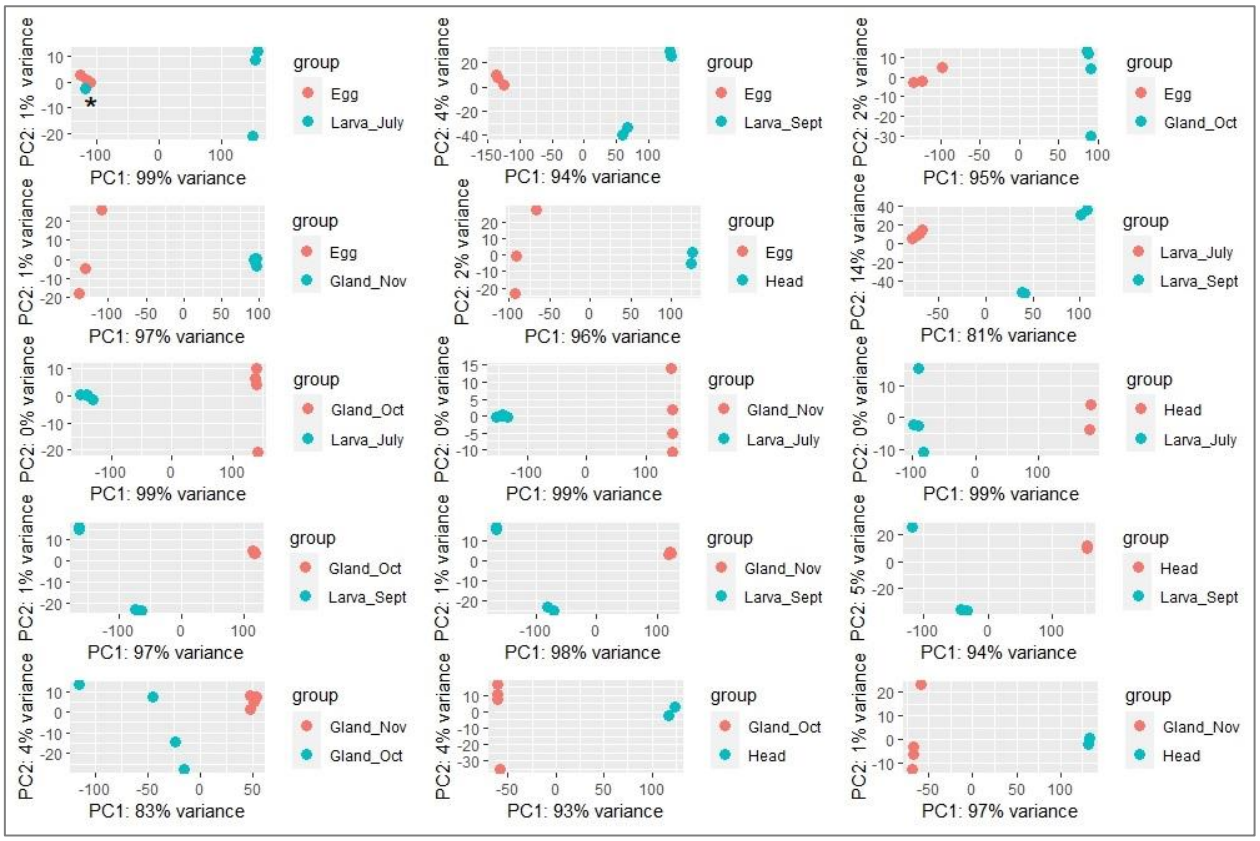


Fig. S19. Percentage of dispersion explained by the first (PC1) and the second (PC2) components after performing principal component analysis of the *Diplolepis rosae* samples used in the relative differential gene expression analysis. Egg: egg removed from a female adult. Larva_July: mid-July larva. Larva_Sept: early September larva. Gland_Oct: October larva salivary glands. Gland_Nov: November larva salivary glands. Head: female adult head. *: sample removed from the analysis.

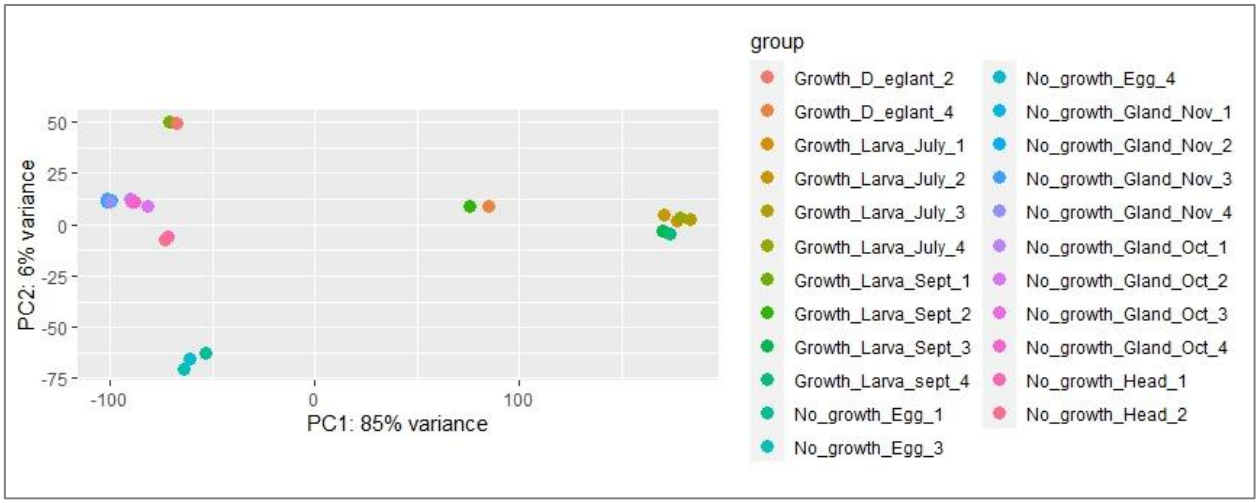


Fig. S20. Percentage of dispersion explained by the first (PC1) and the second (PC2) components after performing principal component analysis of the 'No_growth' and 'Growth' *D. rosae* samples used in the relative differential gene expression analysis. No_growth: sample set that does not correspond to the active gall growth stage (egg removed from a female adult + female adult head + October larva salivary gland + November larva salivary gland). Growth: sample set corresponding to the active gall growth stage (mid-July larva + early September larva + mid-July *Diplolepis eglanteriae* larva).

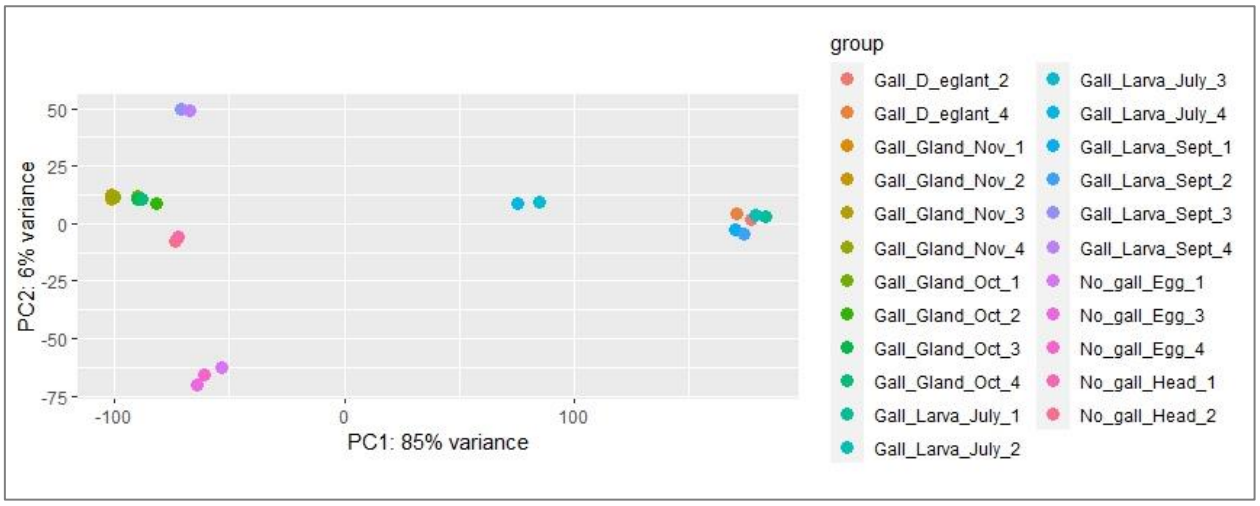


Fig. S21. Percentage of dispersion explained by the first (PC1) and the second (PC2) components after performing principal component analysis of the 'No_gall' and 'Gall' *Diplolepis rosae* samples used in the relative differential gene expression analysis. No_gall: sample set which does not correspond to the whole gall stage (egg removed from a female adult + female adult head): Gall: sample set corresponding to the whole gall stage (mid-July larva + early September larva + mid-July *Diplolepis eglanteriae* larva + October salivary gland + November salivary gland).

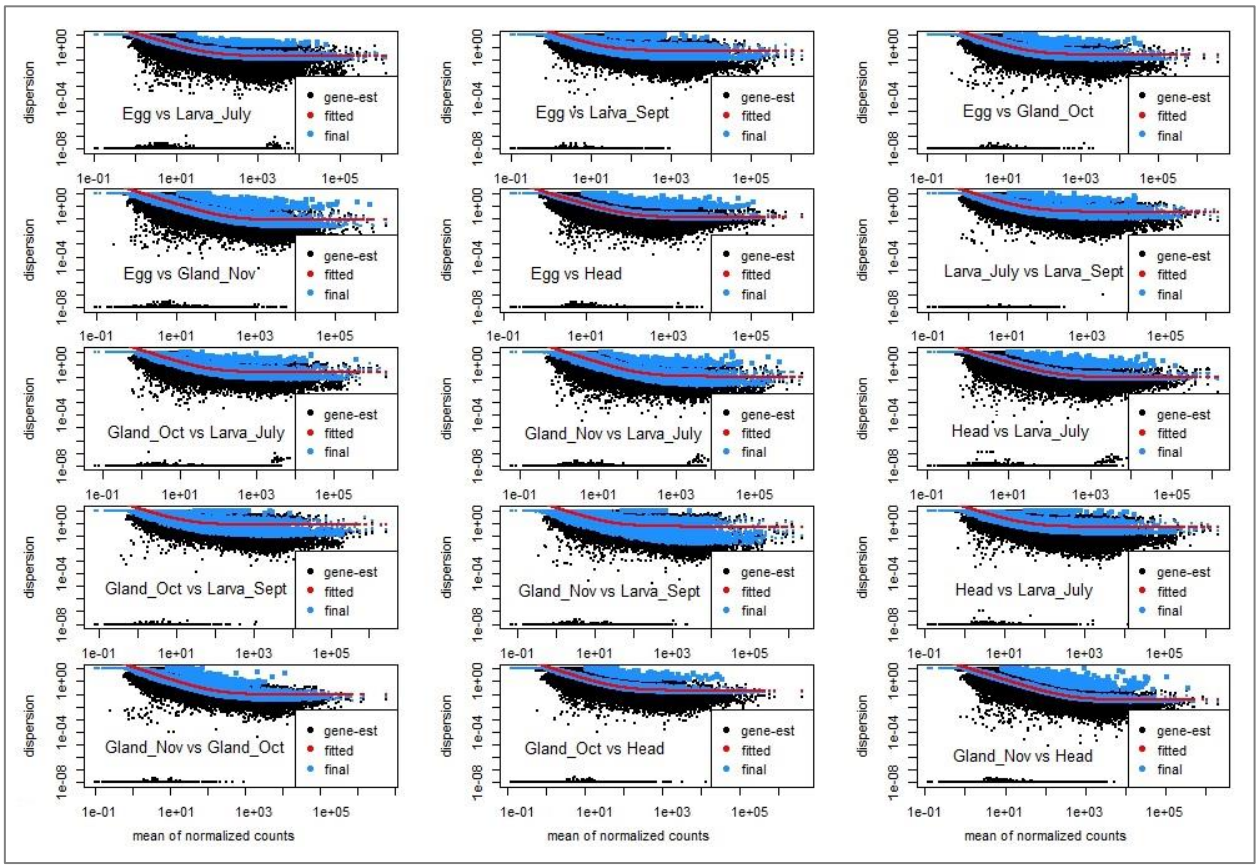


Fig. S22. **Dispersion estimates provided by the pairwise relative differential gene expression analysis of the *Diplolepis rosae* samples**. gene-est (black points): gene estimation; dispersion values calculated for normalized mean counts of reads mapped to each gene. fitted (red line): fitting curve constructed according to a generalized linear model used in the analysis. final (blue points): shrank gene-wise dispersion estimates towards values predicted by the fitting curve. Egg: egg removed from a female adult. Larva_July: mid-July larva. Larva_Sept: early September larva. Gland_Oct: October larva salivary glands. Gland_Nov: November larva salivary glands. Head: female adult head.

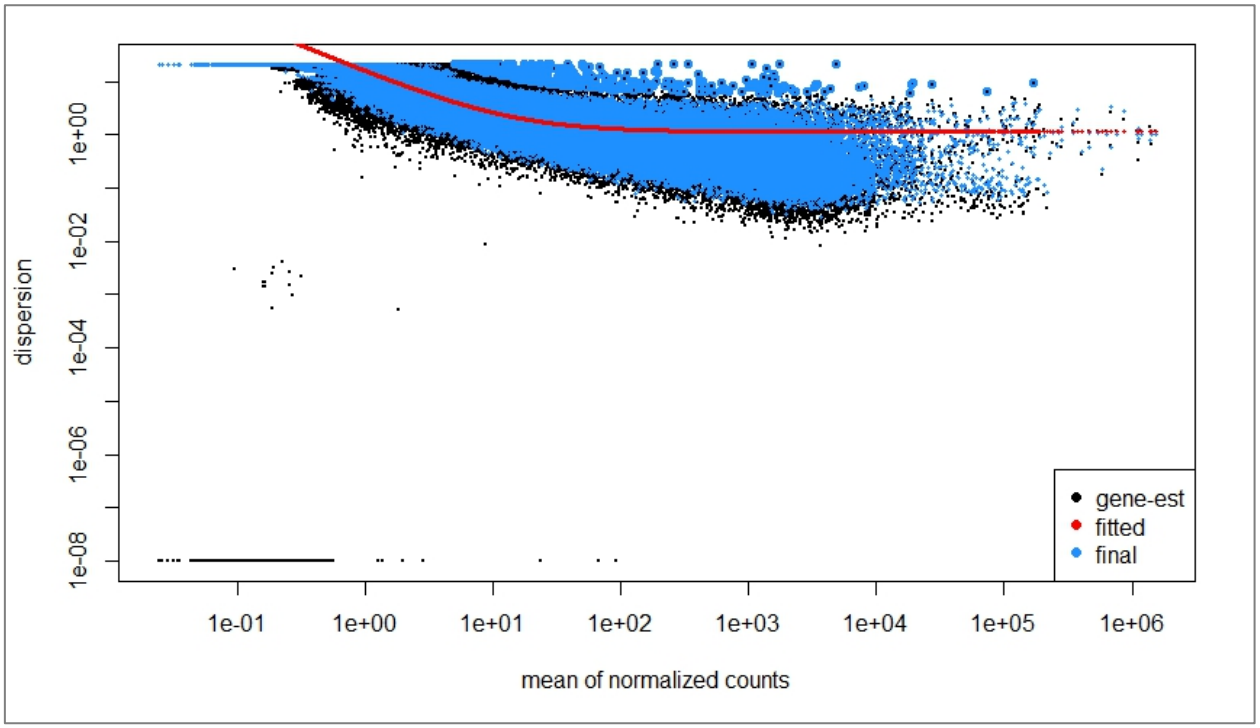


Fig. S23. **Dispersion estimates provided by the relative differential gene expression analysis of the sample pair 'No_growth *Diplolepis rosae* - Growth *Diplolepis rosae*'**. No_growth: sample set that does not correspond to the active gall growth stage (egg removed from a female adult + female adult head + October larva salivary gland + November larva salivary gland). Growth: sample set corresponding to the active gall growth stage (mid-July larva + early September larva + mid-July *Diplolepis eglanteriae* larva). gene-est (black points): gene estimation; dispersion values calculated for normalized mean counts of reads mapped to each gene. fitted (red line): fitting curve constructed according to a generalized linear model used in the analysis. final (blue points): shrank gene-wise dispersion estimates towards values predicted by the fitting curve.

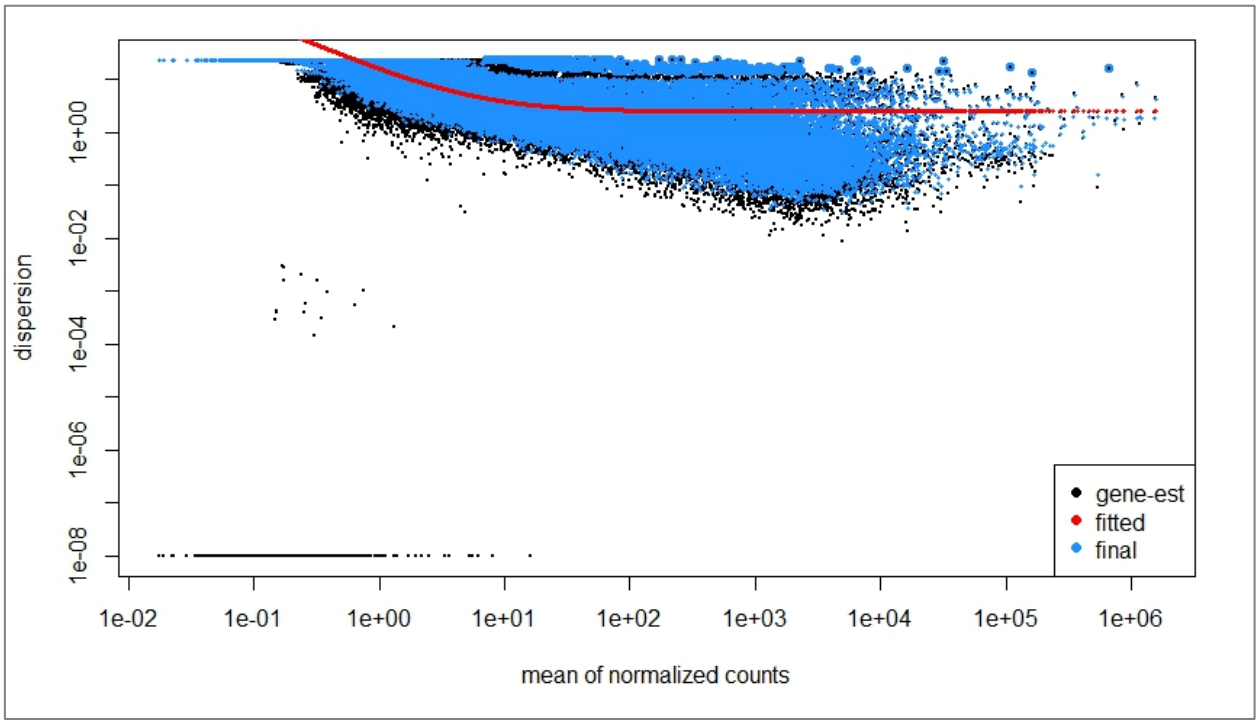


Fig. S24. Dispersion estimates provided by the relative differential gene expression analysis of the sample pair 'No_gall *Diplolepis rosae* - Gall *Diplolepis. rosae*'. No_gall: sample set which does not correspond to the whole gall stage (egg removed from a female adult + female adult head): Gall: sample set corresponding to the whole gall stage (mid-July larva + early September larva + mid-July *Diplolepis eglanteriae* larva + October salivary gland + November salivary gland). gene-est (black points): gene estimation; dispersion values calculated for normalized mean counts of reads mapped to each gene. fitted (red line): fitting curve constructed according to a generalized linear model used in the analysis. final (blue points): shrank gene-wise dispersion estimates towards values predicted by the fitting curve.

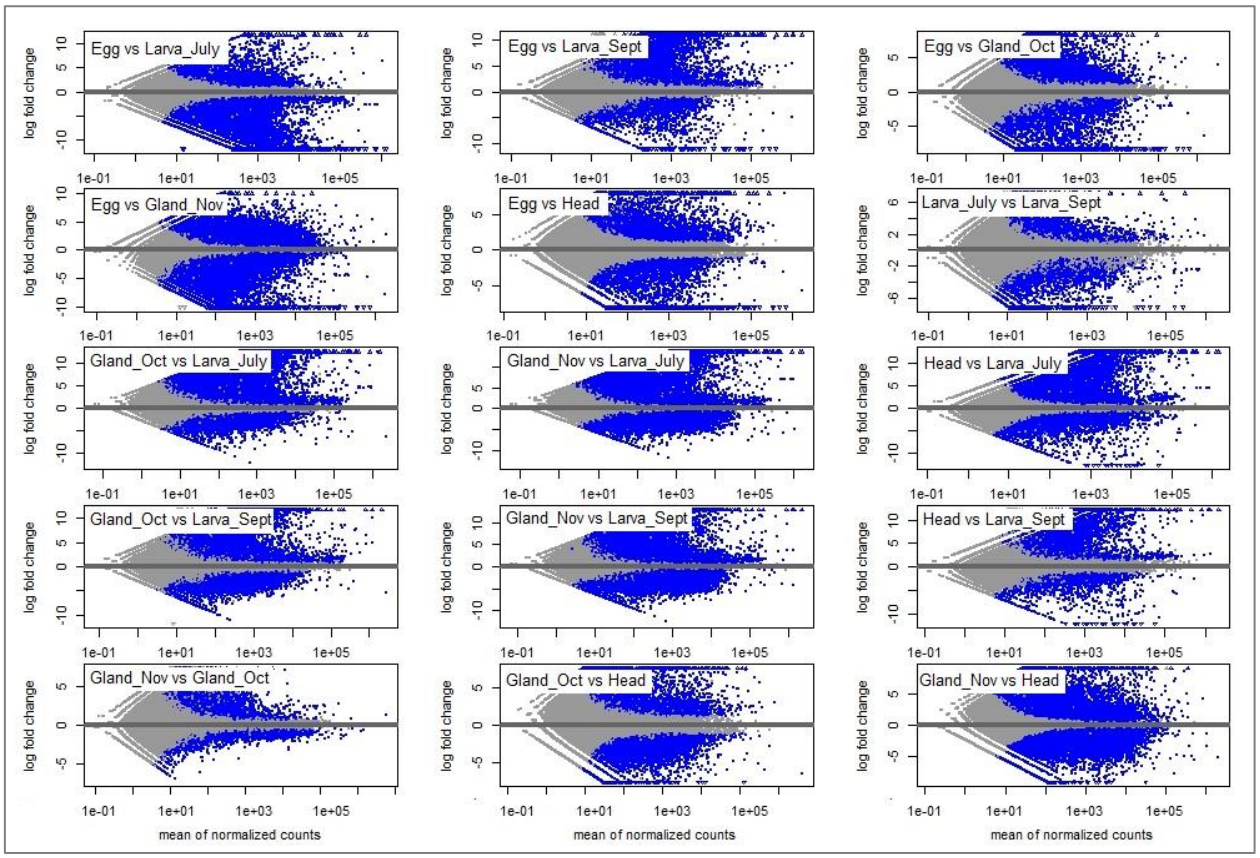


Fig. S25. **Log2 fold changes of the gene expression levels calculated for each gene over the mean of normalized counts for all the samples used in the pairwise relative differential expression analysis**. Blue points show the values with the adjusted p-value < 0.1. Log2 fold change is calculated as log2 (normalized mean counts of reads mapped to a gene of sample 1 / normalized mean counts of reads mapped to a gene of sample 2). Egg: egg removed from a female adult. Larva_July: mid-July larva. Larva_Sept: early September larva. Gland_Oct: October larva salivary glands. Gland_Nov: November larva salivary glands. Head: female adult head.

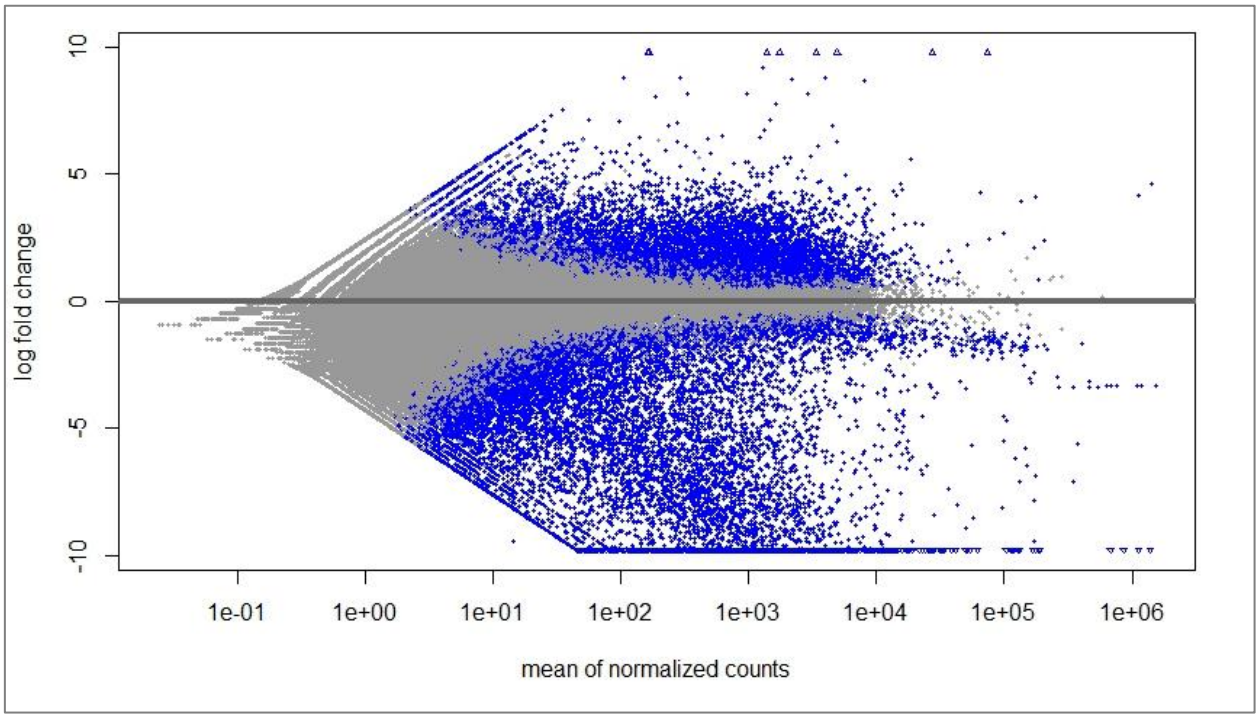


Fig. S26. **Log2 fold changes of the gene expression levels calculated for each gene over the mean of normalized counts for all the 'No_growth' and 'Growth' *Diplolepis rosae* samples used in the relative differential expression analysis**. Blue points show the values with the adjusted p-value < 0.1. Log2 fold change is calculated as log2 (normalized mean counts of reads mapped to a gene of the 'No_growth' sample set / normalized mean counts of reads mapped to a gene of the 'Growth' sample set). No_growth: sample set which does not correspond to the active gall growth stage (egg removed from a female adult + female adult head + October larva salivary gland + November larva salivary gland). Growth: sample set corresponding to the active gall growth stage (mid-July larva + early September larva + mid-July *Diplolepis eglanteriae* larva).

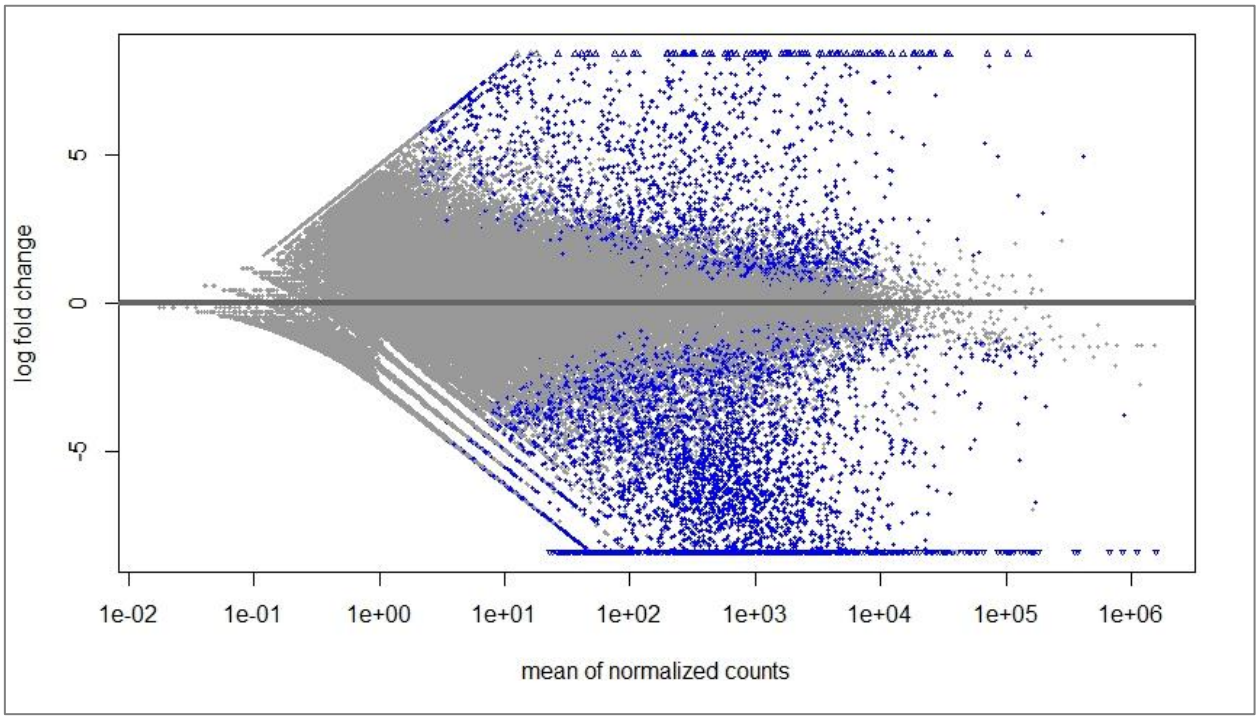


Fig. S27. **Log2 fold changes of the gene expression levels calculated for each gene over the mean of normalized counts for all the 'No_gall' and 'Gall' *Diplolepis rosae* samples used in the relative differential expression analysis**. Blue points show the values with the adjusted p-value < 0.1. Log2 fold change is calculated as log2 (normalized mean counts of reads mapped to a gene of the 'No_gall' sample set / normalized mean counts of reads mapped to a gene of the 'Gall' sample set). No_gall: sample set which does not correspond to the whole gall stage (egg removed from a female adult + female adult head): Gall: sample set corresponding to the whole gall stage (mid-July larva + early September larva + mid-July *Diplolepis eglanteriae* larva + October salivary gland + November salivary gland).

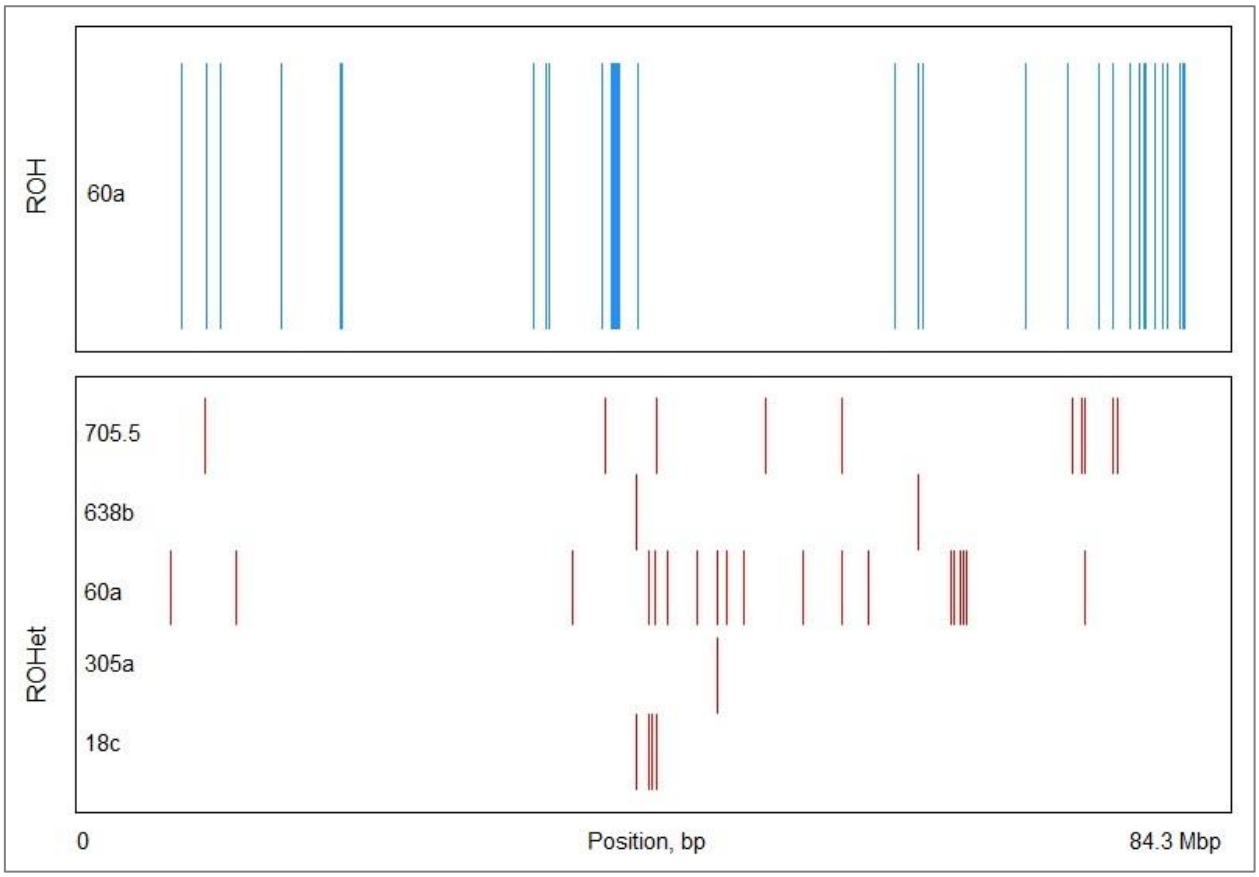


Fig. S28. Runs of homozygosity (ROH) and runs of heterozygosity (ROHet) detected in scaffold 57 (chr_57) of *Cynips quercusfolii*.

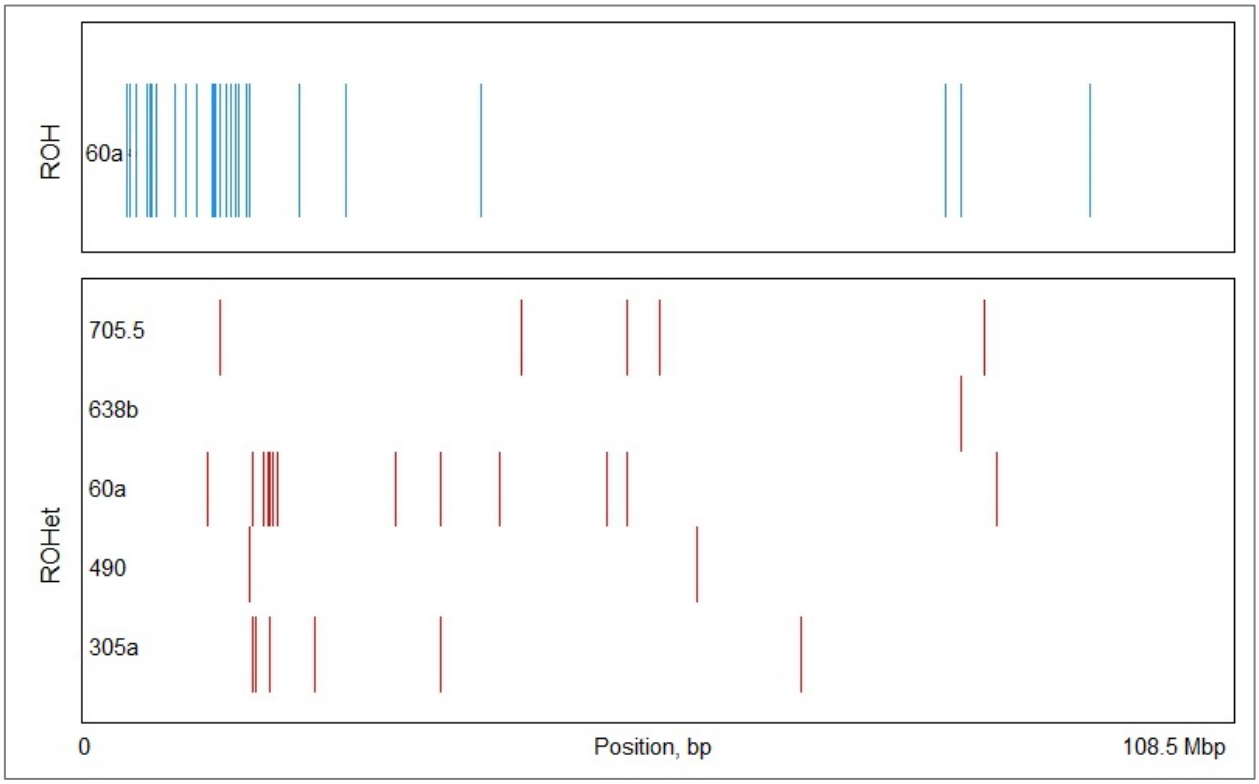


Fig. S29. Runs of homozygosity (ROH) and runs of heterozygosity (ROHet) detected in scaffold 58 (chr_58) of *Cynips quercusfolii*.

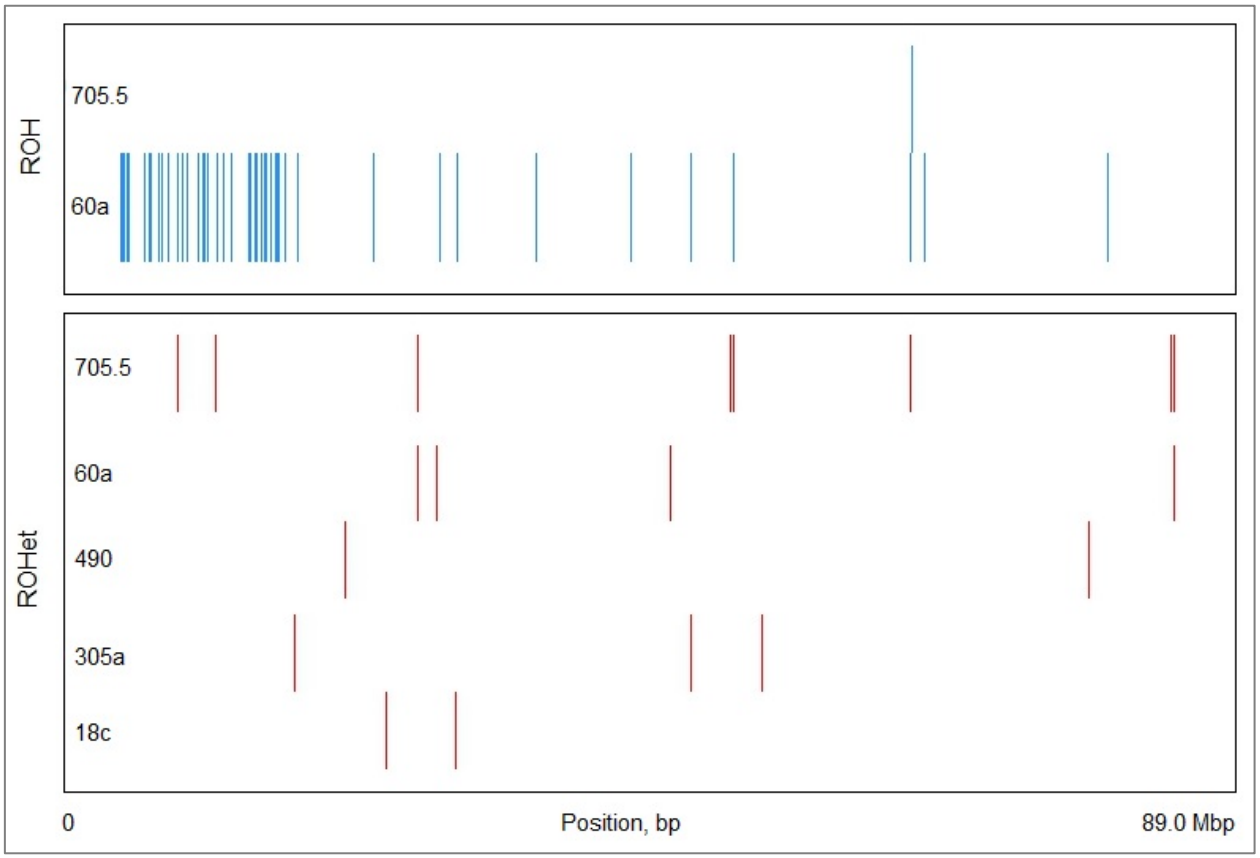


Fig. S30. Runs of homozygosity (ROH) and runs of heterozygosity (ROHet) detected in scaffold 59 (chr_59) of *Cynips quercusfolii*.

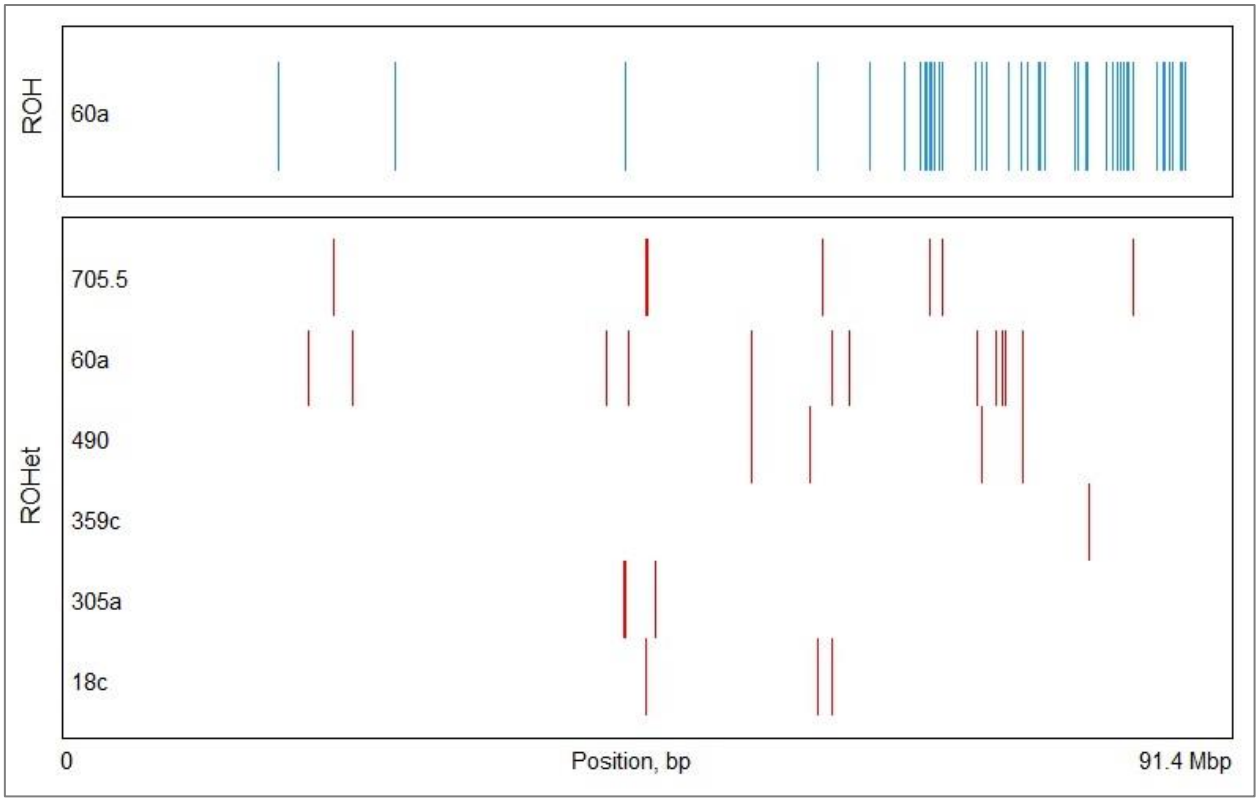


Fig. S31. Runs of homozygosity (ROH) and runs of heterozygosity (ROHet) detected in scaffold 60 (chr_60) of *Cynips quercusfolii*.

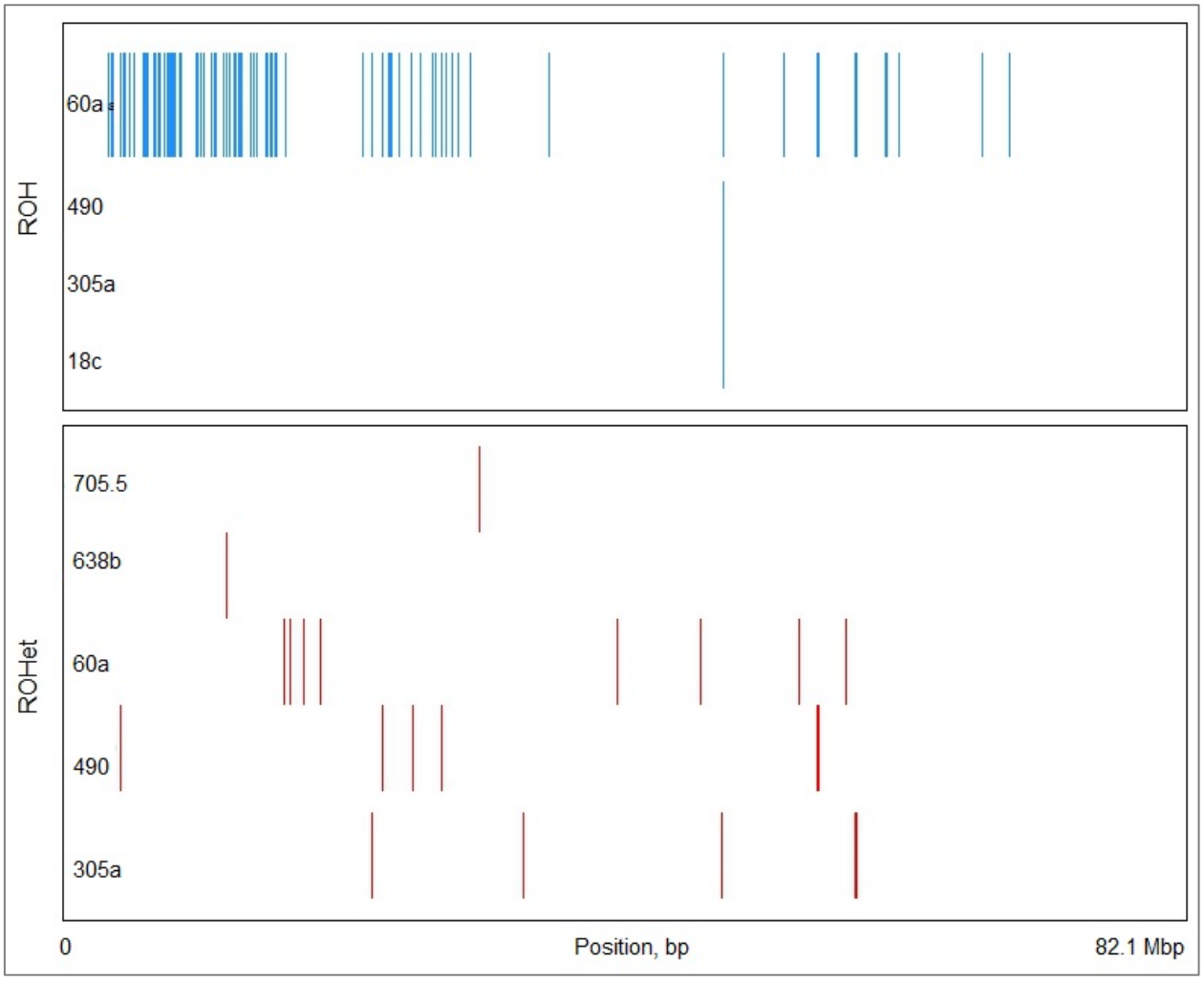


Fig. S32. Runs of homozygosity (ROH) and runs of heterozygosity (ROHet) detected in scaffold 61 (chr_61) of *Cynips quercusfolii*.

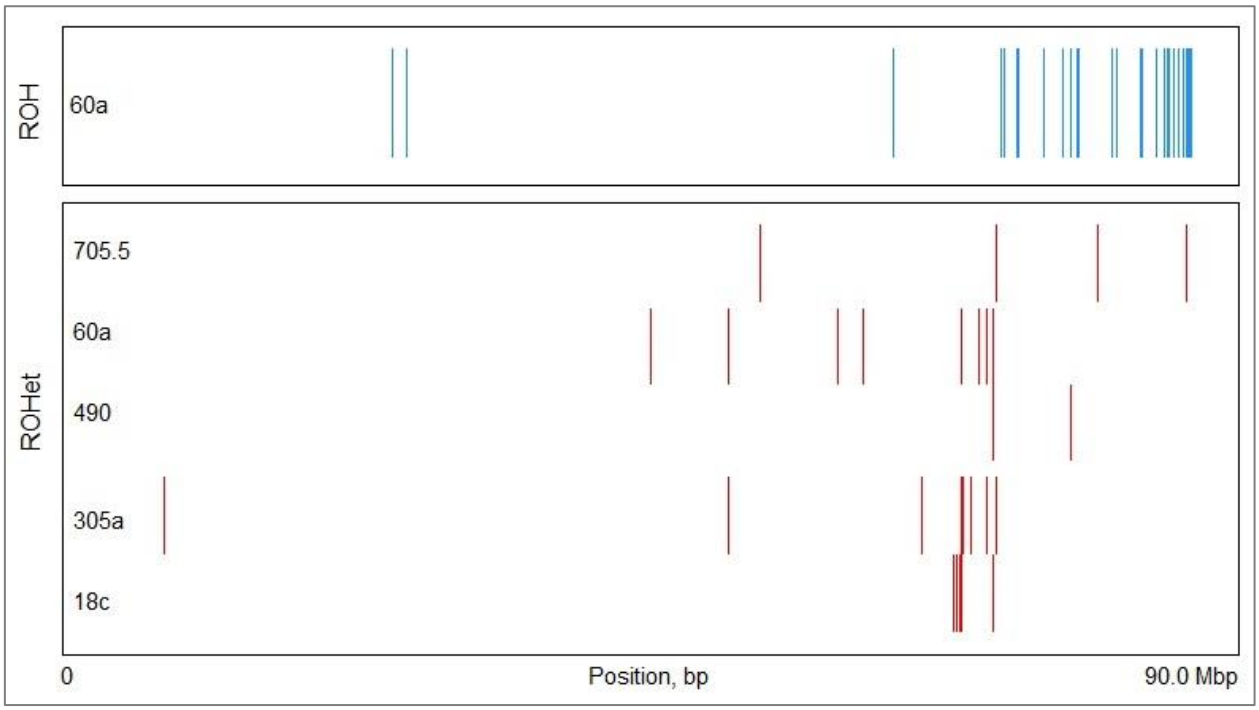


Fig. S33. Runs of homozygosity (ROH) and runs of heterozygosity (ROHet) detected in scaffold 62 (chr_62) of *Cynips quercusfolii*.

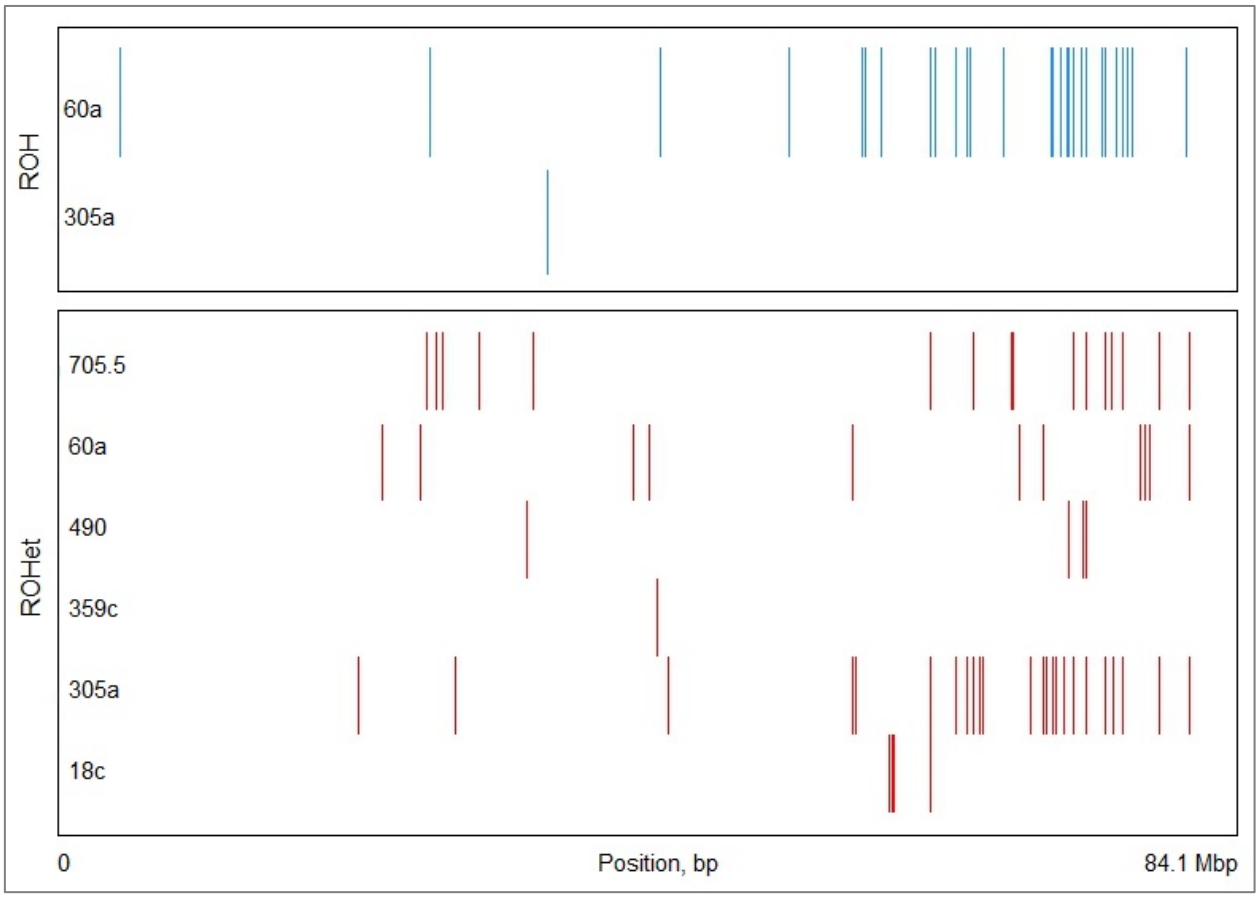


Fig. S34. Runs of homozygosity (ROH) and runs of heterozygosity (ROHet) detected in scaffold 64 (chr_64) of *Cynips quercusfolii*.

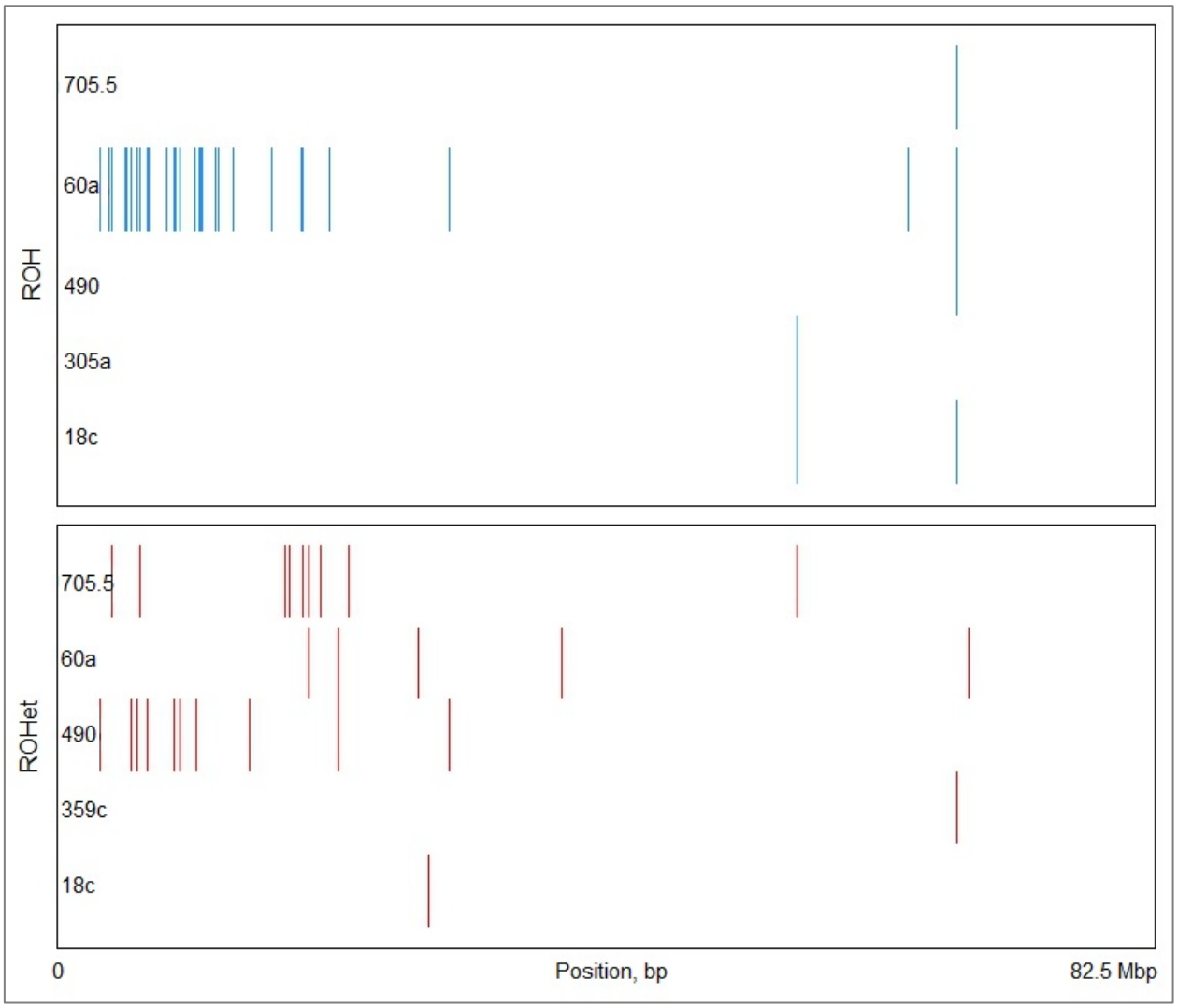


Fig. S35. Runs of homozygosity (ROH) and runs of heterozygosity (ROHet) detected in scaffold 65 (chr_65) of *Cynips quercusfolii*.

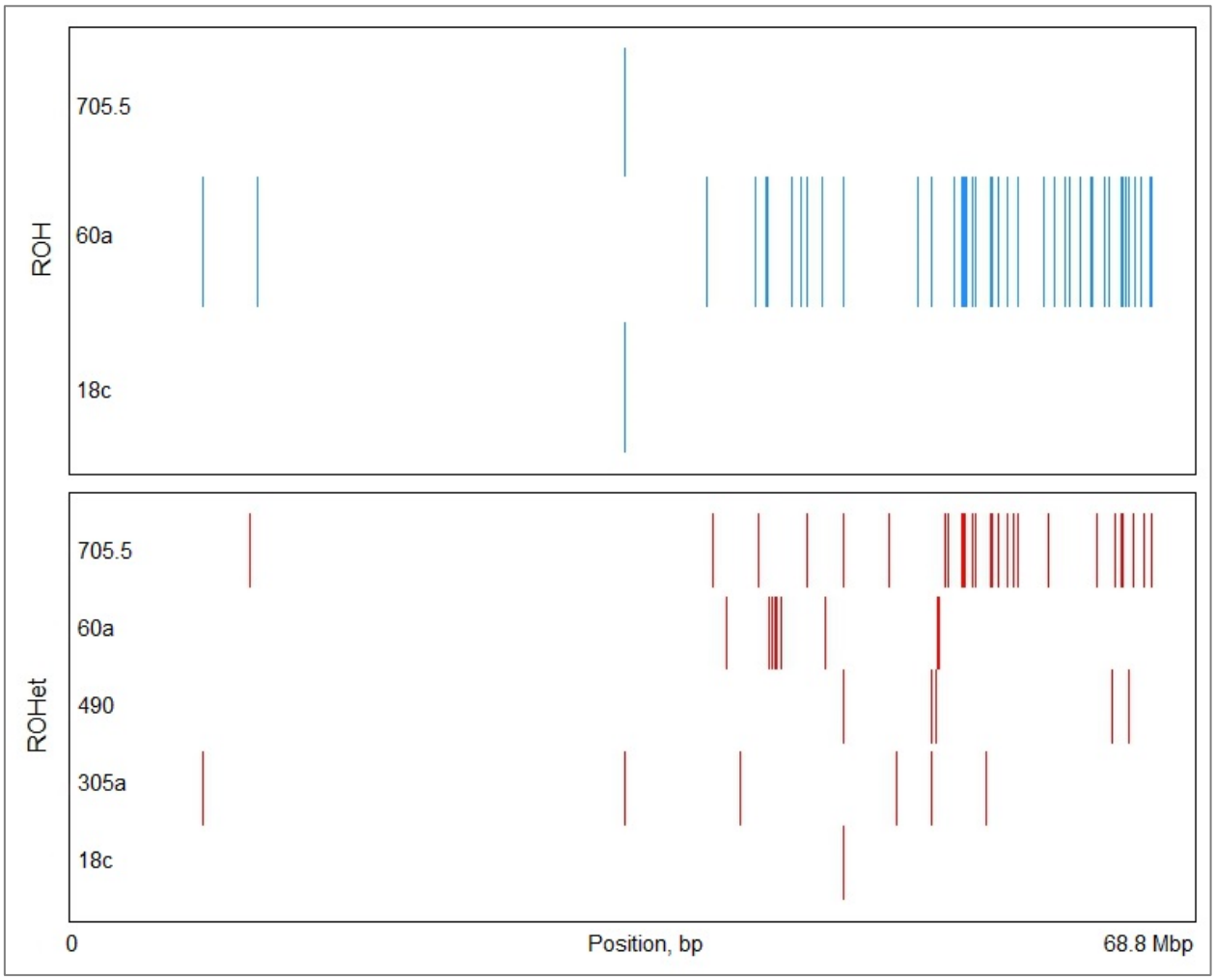


Fig. S36. Runs of homozygosity (ROH) and runs of heterozygosity (ROHet) detected in scaffold 66 (chr_66) of *Cynips quercusfolii*.

Protocols

Protocol S1. Polymerase chain reaction for *Diplolepis rosae* genetic marker.

1. For one sample in 200-µL plastic tubes prepare 25 µL of a solution containing:
    - 5 µL of 5X Taq Buffer with MgCl2 (Promega)
    - 0.5 µL of 10 mM dNTPs (Promega)
    - one unit of Taq DNA polymerase (0.2 µL) (Promega)
    - 0.7 µL of 10 mM forward primer †
    - 0.7 µL of 10 mM reverse primer †
    - 16.9 µL of nuclease-free water (Promega)
    - 1 µL of extracted DNA
2. PCR thermocycling profile:
    - 1 cycle of 2 min at 95 °C
    - 45 cycles of 45 sec at 95 °C, 45 sec at 65.5 °C, and 60 sec at 72 °C
    - final extension of 5 minutes at 72 °C
    - store at 4 °C until necessary

† *Information about the marker:*

Sequence:

GCATGAGATTGAGAAGCGGGAGTGGTGGCTTTGACGACTGCCACCCTCGAAGACTAGAAGA
AGGGCTTATTCCGCCGTCGTTTCAACCGGCGATGATCATTAAAAGTTTAAGCGTCCCCTATTC
AAGTACAGATTCAACGGCCGATTCAGATTCTACTATGGACGGCGCTATTTACTTTCTGCTCTCG
TCAAACTTGCTAAGAGAAAAGAGAGAGAGCAATTATTGGTAATTCGGTTTCATAAATTTCGGT
AAAATGAATTTCAGGAAATATGGAATTTGGGGAATATTCATGTGCCCCTCCGAATAACTATGC
GAAGTGTGTTCAGAAAGTAATGATAAAATCTTCAGCTGTATTCATTGTTTACTTTGTCCAGGCA
CGGCCAAGTCTTATAGGAGGAATTGGCAAAATTTAGAACTTGGACGAAGGGATGAAGGGGG
AAGTGCGGGGTGTTGAGTAGTTATCTGGTCCCTGGTAGGTGGCTCTGGATATGATGACTAGAT
GTATCTGCGGGGAAAGGGGATTTGAATAATGCAAACACGAGACAGTCTCAATTTTGGTTAGT
TGCTGCCAGTTATTTATGGAAAATGGATCGTAATAAATGCATTGTGCCAGCTTGTAATTCATTA
CGGAAGAATGAAAAAGTTCCAATACATCGACTTCCAAGAAATAGTGTCACTGGACGTGAGTG
GTTAGAGTTGTGCGGA

Primer design (Untergasser et al. 2012):

OLIGO	start	len	tm	gc%	any_th	3'_th	hairpin	seq
LEFT PRIMER	7319	20	59.06	55.00	0.00	0.00	0.00	
5′ GCATGAGATTGAGAAGCGGG 3′								
RIGHT PRIMER	8024	20	58.95	50.00	0.00	0.00	0.00	
5′ TCCGCACAACTCTAACCACT 3′								

SEQUENCE SIZE: 8323
INCLUDED REGION SIZE: 8323
PRODUCT SIZE: 706, PAIR ANY_TH COMPL: 5.58, PAIR 3'_TH COMPL: 8.79

Protocol S2. DNA extraction (Qiagen DNA insect kit)

1. Place the sample in a 2-mL Eppendorf tube
2. Add 180 µL of ATL, add two metallic beads, and homogenize the sample by using the tissueLyser (2x 30 Hertz /1 min)
3. Centrifuge at 13000 rpm for 2 min, and take the supernatant
4. Add 20 µL of proteinase K and 200 µL of AL
5. Vortex, incubate at 56 C for 3 hours
6. Vortex, add 200 µL of ethanol (96-100%), vortex
7. Pipet the mixture (including any precipitate) into the DNeasy Mini spin column placed in a 2 ml collection tube (provided).
8. Centrifuge at ≥6000 x g (8000 rpm) for 1 min. Discard flow-through and collection tube
9. Place the DNeasy Mini spin column in a new 2 ml collection tube (provided), add 500 µl of AW1, and centrifuge for 1 min at ≥6000 x g (8000 rpm). Discard flow-through and collection tube
10. Place the DNeasy Mini spin column in a new 2 ml collection tube (provided), add 500 µl of AW2, and centrifuge for 3 min at 20,000 x g (14,000 rpm) to dry the DNeasy membrane. Discard flow-through and collection tube
11. Place the DNeasy Mini spin column in a clean 2 ml microcentrifuge tube (not provided), and pipet 10-50 µl of AE directly onto the DNeasy membrane
12. Incubate at room temperature for 1 min, and then centrifuge for 1 min at ≥6000 x g (8000 rpm) to elute
13. Store at 4 °C

Protocol S3. Estimation of the quantity of dsDNA (Qubit dsDNA BR Assay Kit)

1. Prepare a Solution containing 199 µL of a Buffer and 1 µL of a 200X Reagent per sample
2. Prepare a Standart 1 in 200-µL plastic tubes (provided): 10 µL of the Standart 1 (0 ng/µL in TE buffer) + 190 µL of Solution
3. Prepare a Standart 2: 10 µL of the Standart 2 (100 ng/µL in TE buffer) + 190 µL of Solution
4. Prepare the samples: 4 µL of the extracted DNA + 196 µL of Solution
5. Measure the concentration of DNA using Qubit Fluorometer ('Broad range' option)

Protocol S4. Polymerase chain reaction for the gene encoding cytochrome c oxidase I (710 bp)

1. For one sample in 200-µL plastic tubes prepare 25 µL of a solution containing:
    - 5 µL of 5X Taq Buffer with MgCl2 (Promega)
    - 0.5 µL of 10 mM dNTPs (Promega)
    - one unit of Taq DNA polymerase (0.2 µL) (Promega)
    - 0.7 µL of 10 mM forward primer (LCO1490 (5' GGTCAACAAATCATAAAGATATTGG 3'))
    - 0.7 µL of 10 mM reverse primer (HCO2198 (5' TAAACTTCAGGGTGACCAAAAAAT 3'))
    - 16.9 µL of nuclease-free water (Promega)

- 1 µL of extracted DNA

2. PCR thermocycling profile:
    - 1 cycle of 2 min at 95 °C
    - 30 cycles of 45 sec at 95 °C, 45 sec at 59 °C, and 60 sec at 72 °C
    - final extension of 5 minutes at 72 °C
    - store at 4 °C until necessary

Protocol S5. Electrophoresis of PCR products

1. In a 500-ml chemical flask, prepare a solution containing 3.5 g of agarose in 100 ml 1X Tris-Acetate-EDTA buffer, and heat to dissolve the agarose (800 WM for 2 min in a microwave)
2. Add 10 µL of 10000X SYBR Safe DNA gel stain (Invitrogen)
3. Put the solution into a plastic electrophoresis cuve, and wait for 20 min
4. Add 2.5-5 µL of a PCR product and 2.5- 5 µL of SmartLadder - 200 - 10000 bp (Eurogentec)
5. Electrophoresis parameters (Mupid One electrophoresis system): 100 V for 40 min, in 1X Tris-Acetate-EDTA buffer
6. Visualize the PCR Product in a UV Transilluminator

Protocol S6. **RNA extraction (RNeasy Mini Kit)**

1. Prepare a 1.5-mL Eppendorf tube, add two metallic beads, 100 µL of the RNAlater Solution, and 600 µL of RLT
2. Place the insect sample into solution
3. Homogenize the sample using the tissueLyser (2x 30 Hertz /2 min)
4. Centrifuge at 13000 rpm for 3 min
5. Transfer the supernatant into a new 1.5-mL Eppendorf tube
6. Add 700 µL of 70% ethanol and mix by pipetting
7. Transfer the mixture into the RNeasy Mini spin column placed in a 2 ml collection tube (provided).
8. Centrifuge at ≥ 10000 rpm for 1 min. Discard flow-through and collection tube
9. Place the RNeasy Mini spin column in a new 2 ml collection tube (provided), add 700 µl of RW1, and centrifuge at ≥ 10000 rpm for 1 min. Discard flow-through and collection tube
10. Place the RNeasy Mini spin column in a new 2 ml collection tube (provided), add 500 µl of RPE, and centrifuge at ≥ 10000 rpm for 1 min. Discard flow-through and collection tube
11. Place the RNeasy Mini spin column in a new 2 ml collection tube (provided), add 500 µl of RPE, and centrifuge at ≥ 10000 rpm for 2 min. Discard flow-through and collection tube
12. Place the RNeasy Mini spin column in a new 2 ml collection tube (provided) and centrifuge at 13000 rpm for 1 min to eliminate any possible carryover of RPE

13. Place the RNeasy Mini spin column in a new 1.5 ml collection tube (provided), add 30 µl of RNAse-free water, and centrifuge at ≥ 10000 rpm for 1 min to elute the RNA
14. Store at 4 °C

Protocol S7. Estimation of the quantity of RNA (Qubit RNA BR Assay Kit)

6. Prepare a Solution containing 199 µL of a Buffer and 1 µL of a 200X Reagent per sample
7. Prepare a Standart 1 in 200-µL plastic tubes (provided): 10 µL of the Standart 1 (0 ng/µL in TE buffer) + 190 µL of Solution
8. Prepare a Standart 2: 10 µL of the Standart 2 (100 ng/µL in TE buffer) + 190 µL of Solution
9. Prepare the samples: 4 µL of the extracted DNA + 196 µL of Solution
10. Measure the concentration of RNA using Qubit Fluorometer ('Broad range' option)

Codes

Code S1. Scripts used to find the best scenario describing the demography of *Diplolepis rosae* by continuous approximation) approach (Gutenkunst et al. 2009). The bricks are available from https://dadi.readthedocs.io/en/latest/api/dadi/.

```
python3
import numpy as np
import dadi

#General remarks
#Time T is given in units of 2*Neff generations
#Migration rate m12 is given in units of 2*Neff*mig12, mig12 is a
#fraction of individuals each generation in pop 1 that are new migrants #from
pop 2 rates. No migration m12=m21=0; simmetric migration m12=m21=m
#pts_0 = [20, 40, 60] is the number of grid points used in the calculation

#Importing data
dd = dadi.Misc.make_data_dict_vcf("varian_call_Drosae_ancestral_allele.vcf",
"pops.txt")
fs = dadi.Spectrum.from_data_dict(dd, ['pop1', 'pop2'], projections = [18, 16],
polarized=True)

#Joint polarized AFS
import pylab
pylab.figure(figsize=(10,10))
dadi.Plotting.plot_single_2d_sfs(fs, vmin = 1)
pylab.savefig('AFS', dpi=250)

#Single-population statistics
thetaW = fs.Watterson_theta()
pi = fs.pi()
D = fs.Tajima_D()

#Multi-population statistics
S = fs.S()
Fst = fs.Fst()

#split model: split into two populations of a specified size with migration
#nu1, nu2: population sizes after split
#T: time in the past of split
#m12, m21: migration

def split_mig(params, ns, pts):
      nu1,nu2,T,m12,m21 = params
      xx = dadi.Numerics.default_grid(pts
      phi = dadi.PhiManip.phi_1D(xx)
      phi = dadi.PhiManip.phi_1D_to_2D(xx, phi)
      phi = dadi.Integration.two_pops(phi, xx, T, nu1, nu2, m12=12, m21=m21)
      fs = dadi.Spectrum.from_phi(phi, ns, (xx,xx))
      return fs

#optimisation of parameters
my_extrap_func = dadi.Numerics.make_extrap_func(split_mig)
pts_0 = [20, 40, 60]
upper_bound = [2000, 2000, 0.5, 0.15, 0.15]
lower_bound = [500, 500, 0.1, 0.05, 0.05]
#set of parameters from which start optimisation
p0 = [1250, 1250, 0.25, 0.1, 0.1]
p0    =    dadi.Misc.perturb_params(p0,    fold=1,    upper_bound=upper_bound,
lower_bound=lower_bound)
```

```
print('Beginning optimisation')
popt    =    dadi.Inference.optimize(p0,    fs,    my_extrap_func,    pts_0,
lower_bound=lower_bound, upper_bound=upper_bound, verbose=len(p0))
print('Finished optimisation')
print('Best-fit parameters: {0}'.format(popt))

#Best-fit parameters after optimisation
#When assymetric migration:
#popt   =   [7.51053057e+02   1.98023428e+03   1.69649627e-01   5.28043606e-02
#1.48500000e-01]
#When symmetric migration m12=m21=m:
#popt = [1.02722156e+03 1.56986352e+03 2.17817079e-01 1.48500000e-01]
#When no migration:
#popt = [1.16747258e+03 1.99816894e+03 1.74841169e-01]


#isolation model: isolation with exponential pop growth
#s: proportion of individuals of pop 1 after split (pop 2 has size 1-s)
#nu1, nu2: final population sizes (after exponential growth)
#T: time in the past of split
#m12, m21: migration

def isol_mig(params, ns, pts):
    s,nu1,nu2,T,m12,m21 = params
    xx = dadi.Numerics.default_grid(pts)
    phi = dadi.PhiManip.phi_1D(xx)
    phi = dadi.PhiManip.phi_1D_to_2D(xx, phi)
    nu1_func = lambda t: s * (nu1/s)**(t/T)
    nu2_func = lambda t: (1-s) * (nu2/(1-s))**(t/T)
    phi = dadi.Integration.two_pops(phi, xx, T, nu1_func, nu2_func, m12=12,
m21=21)
    fs = dadi.Spectrum.from_phi(phi, ns, (xx,xx))
    return fs

#optimisation of parameters
my_extrap_func = dadi.Numerics.make_extrap_func(isol_mig)
pts_0 = [20, 40, 60]
upper_bound = [0.9, 2000, 2000, 1.5, 1.5, 1.5]
lower_bound = [0.1, 500, 500, 0.05, 0.05, 0.05]
p0 = [0.5, 1000, 1000, 0.1, 1, 01]
p0    =    dadi.Misc.perturb_params(p0,    fold=1,    upper_bound=upper_bound,
lower_bound=lower_bound)
print('Beginning optimisation')
popt    =    dadi.Inference.optimize(p0,    fs,    my_extrap_func,    pts_0,
lower_bound=lower_bound, upper_bound=upper_bound, verbose=len(p0))
print('Finished optimisation')
print('Best-fit parameters: {0}'.format(popt))

#Best-fit parameters after optimisation
#When assymetric migration:
#popt   =   [8.21020680e-01   1.55906681e+03   1.20456650e+03   9.10803571e-01
#1.35621133e+00 7.18192319e-0]
#When symmetric migration m12=m21=m:
#popt   =   [1.00004257e-01   5.26620953e+02   8.98966318e+02   1.46967456e-01
#7.69940616e-01]
#When no migration:
#popt = [8.95917278e-01 6.99188738e+02 1.66546204e+03 4.53489214e-01]


#Isolation-with-migration model with exponential pop growth and a size change
prior to split
#nuPre: size after first size change
#TPre: time before split of first size change
```

```
#s: fraction of nuPre that goes to pop 1 => pop 2 has size nuPre*(1-s).)
#nu1, nu2: final population sizes
#T: Time in the past of split
#m12, m21: migration
#n1,n2: sample sizes of resulting spectrum

def isol_mig_pre(params, ns, pts):
      nuPre,TPre,s,nu1,nu2,T,m12,m21 = params
      xx = Numerics.default_grid(pts)
      phi = PhiManip.phi_1D(xx)
      phi = Integration.one_pop(phi, xx, TPre, nu=nuPre)
      phi = PhiManip.phi_1D_to_2D(xx, phi)
      nu1_0 = nuPre*s
      nu2_0 = nuPre*(1-s)
      nu1_func = lambda t: nu1_0 * (nu1/nu1_0)**(t/T)
      nu2_func = lambda t: nu2_0 * (nu2/nu2_0)**(t/T)
      phi = Integration.two_pops(phi, xx, T, nu1_func, nu2_func, m12=m12,
m21=m21)
      fs = Spectrum.from_phi(phi, ns, (xx,xx))
      return fs

#optimisation of parameters
my_extrap_func = dadi.Numerics.make_extrap_func(isol_mig_pre)
pts_0 = [20, 40, 60]
upper_bound = [2000, 6, 0.9, 1000, 1000, 1.5, 1.5, 1.5]
lower_bound = [500, 5, 0.1, 800, 800, 0.5, 0.5, 0.5]
p0 = [1000, 5.5, 0.5, 900, 900, 1, 1, 1]
p0 = dadi.Misc.perturb_params(p0, fold=1, upper_bound=upper_bound,
lower_bound=lower_bound)
print('Beginning optimisation')
popt = dadi.Inference.optimize(p0, fs, my_extrap_func, pts_0,
lower_bound=lower_bound, upper_bound=upper_bound, verbose=len(p0))
print('Finished optimisation')
print('Best-fit parameters: {0}'.format(popt))

#Best-fit parameters after optimisation
#When assymetric migration:
#popt = [7.64925622e+02 5.01183360e+00 7.53743684e-01 9.89999994e+02
#8.07999931e+02 8.54984055e-01 8.16222033e-01 9.00362495e-01]
#When symmetric migration m12=m21=m:
#popt = [9.88797106e+02 5.05000000e+00 8.73576179e-01 9.90000000e+02
#8.53851232e+02 7.73547199e-01 9.16075629e-01]
#When no migration:
#popt = [5.15031419e+02 5.05000000e+00 7.07834470e-01 9.90000000e+02
#8.08000000e+02 1.48500000e+00]


#bottlegrowth_split model: instantanous size change followed by exponential
growth then split
#nuB: ratio of population size after instantanous change to ancient #population
size
#nuF: ratio of contempoary to ancient population size
#T: time in the past at which instantaneous change happened and growth #began
#Ts: time in the past at which the two populations split

def bottlegrowth_split(params, ns, pts):
    nuB,nuF,T,Ts = params
    return bottlegrowth_split_mig((nuB,nuF,0,T,Ts), ns, pts)

#optimisation of parameters
my_extrap_func = dadi.Numerics.make_extrap_func(bottlegrowth_split)
pts_0 = [20, 40, 60]
upper_bound = [0.2, 3, 0.5, 0.5]
```

```
lower_bound = [0.05, 1, 0.05, 0.05]
p0 = [0.1, 2, 0.1, 0.1]
p0    =    dadi.Misc.perturb_params(p0,    fold=1,    upper_bound=upper_bound,
lower_bound=lower_bound)
print('Beginning optimisation')
popt    =    dadi.Inference.optimize(p0,    fs,    my_extrap_func,    pts_0,
lower_bound=lower_bound, upper_bound=upper_bound, verbose=len(p0))
print('Finished optimisation')
print('Best-fit parameters: {0}'.format(popt))

#Best-fit parameters after optimisation
#popt = [0.13041597 2.37120407 0.21729583 0.20394267]
```

#bottlegrowth_split_migration model: instantanous size change followed by exponential growth then split with migration

```
#nuB: ratio of population size after instantanous change to ancient #population
size
#nuF: Ratio of contempoary to ancient population size
#m: migration
#T: time in the past at which instantaneous change happened and growth began
#Ts: Time in the past at which the two populations split

def bottlegrowth_split_mig(params, ns, pts):
      nuB,nuF,m,T,Ts = params
      xx = dadi.Numerics.default_grid(pts)
      phi = dadi.PhiManip.phi_1D(xx)
      if T >= Ts:
            nu_func = lambda t: nuB*numpy.exp(numpy.log(nuF/nuB) * t/T)
            phi = dadi.Integration.one_pop(phi, xx, T-Ts, nu_func)
            phi = dadi.PhiManip.phi_1D_to_2D(xx, phi)
            nu0 = nu_func(T-Ts)
            nu_func = lambda t: nu0*numpy.exp(numpy.log(nuF/nu0) * t/Ts)
            phi  =  dadi.Integration.two_pops(phi,  xx,  Ts,  nu_func,  nu_func,
m12=m, m21=m)
      else:
            phi = dadi.PhiManip.phi_1D_to_2D(xx, phi)
            phi = dadi.Integration.two_pops(phi, xx, Ts-T, 1, 1, m12=m, m21=m)
            nu_func = lambda t: nuB*numpy.exp(numpy.log(nuF/nuB) * t/T)
            phi  =  dadi.Integration.two_pops(phi,  xx,  T,  nu_func,  nu_func,
m12=m, m21=m)

       fs = dadi.Spectrum.from_phi(phi, ns, (xx,xx))
       return fs

#optimisation of parameters
my_extrap_func = dadi.Numerics.make_extrap_func(bottlegrowth_split_mig)
pts_0 = [20, 40, 60]
upper_bound = [0.9, 5, 1, 5, 5]
lower_bound = [0.01, 0.1, 0.1, 0.1, 0.1]
p0 = [0.1, 1, 0.5, 1, 1]
p0    =    dadi.Misc.perturb_params(p0,    fold=1,    upper_bound=upper_bound,
lower_bound=lower_bound)
print('Beginning optimisation')
popt    =    dadi.Inference.optimize(p0,    fs,    my_extrap_func,    pts_1,
lower_bound=lower_bound, upper_bound=upper_bound, verbose=len(p0))
print('Finished optimisation')
print('Best-fit parameters: {0}'.format(popt))

#Best-fit parameters after optimisation
#popt = [0.05008461 1.93543999 0.18266133 0.12533342 0.81468721]
```

```
#likelihood
model = my_extrap_func(popt, ns, pts_0)
ll = dadi.Inference.ll_multinom(model, fs)
print(ll)

#likelihood ratio test
#my_extrap_func_1: extrapolation of more complex model to an infinitely #fine
grid
#ptc_1: number of grid points used for more complex model
#p1: optimised parameters for more complex model
#ll_0, ll_1: composite likelihood of a simplier and more complex model,
respectively
#bootstrapping for more complex model
chunks = dadi.Misc.fragment_data_dict(dd, chunk_size=15000)
boots = dadi.Misc.bootstraps_from_dd_chunks(chunks, Nboot=10, pop_ids=['pop0',
'pop1'], projections = [18, 16], mask_corners=True, polarized=True)
#Multiplicative adjustment to the likelihood ratio test statistic
adj = dadi.Godambe.LRT_adjust(my_extrap_func_1, ptc_1, boots, p1, fs,
nested_indices=[5], multinom = True, eps=0.01)
#adjusted D-statistics
D=adj*2*(ll_1 - ll_0)
print (D)
p=dadi.Godambe.sum_chi2_ppf(D, weights=(0, 1))
print(p)
```

Code S2. Script used to estimate model parameters of *Diplolepis rosae* demography (approximate Bayesian computation approach).

Simulations (Baumdicker et al. 2022)

```
import numpy as np

import msprime

import tskit

import csv


header = ['Nanc', 'N1', 'N2', 'tbot', 'proportion', 'tsplit', 'mu', 'mean_pi1',
'var_pi1',   'mean_pi2',   'var_pi2',   'mean_Tajimas_D1',   'var_Tajimas_D1',
'mean_Tajimas_D2', 'var_Tajimas_D2', 'mean_d', 'var_d', 'mean_Fst', 'var_Fst']

f = open("simulations_Drosae.csv", 'w', encoding='UTF8', newline='')

writer = csv.writer(f, delimiter="\t")

writer.writerow(header)


#In a log uniform distribution, the log transformed random variable is #assumed
to be uniformly distributed

#Thus logU(a, b) ~ exp(U(log(a), log(b))

#Thus, we could create a log-uniform distribution using numpy

#mu: mutation rate (per sequence length per generation)

#N1, N2: diploid size of ancestral and two populations of D. rosae,
#respectively

#tbot: bottleneck time of ancestral population, in generations

#proportion: proportion of individuals after bottleneck compared to the #size
of ancestral population

#tsplit: split time of the ancestral population to two populations of D. #rosae,
in generations


i = 1

while i <= 1000000:

#Random parameters, simulations

    mu = np.random.uniform(1e-10, 9.99e-7)

    N1 = np.exp(np.random.uniform(5, 12, size=None))

    N2 = np.exp(np.random.uniform(5, 12, size=None))

    tbot = np.exp(np.random.uniform(5, 9, size=None))

    proportion = np.random.uniform(0, 1)

    tsplit = np.exp(np.random.uniform(5, 9, size=None))

    if tbot> tsplit:
```

```
#Demography

            demography = msprime.Demography()

            demography.add_population(name="Pop1", initial_size=N1)

            demography.add_population(name="Pop2", initial_size=N2)

            demography.add_population(name="Anc", initial_size=N1+N2)

            demography.add_population_split(time=tsplit,      derived=["Pop1",
      "Pop2"], ancestral="Anc")

            demography.add_simple_bottleneck(time=tbot, proportion=proportion,
population="Anc")

#Tree sequence

#Sequence length: D. rosae assembly length 497872038 bp => 10000 times shorter


            sequence_length = 49787

            ts   =   msprime.sim_ancestry(samples={"Pop1":   9,   "Pop2":   8},
demography=demography, ploidy=2, random_seed=np.random.randint(1, 5000 + 1),
sequence_length=sequence_length)

#Mutation model

            model=msprime.BinaryMutationModel()

#Mutations

            mutations       =       msprime.sim_mutations(ts,       rate=mu,
random_seed=np.random.randint(1, 5000 + 1), model=model)

#Statistics in each 1000 bp region

            num_windows = 50

            pi1                                                              =
            mutations.diversity(sample_sets=mutations.samples(population=0),
      windows=np.linspace(0, sequence_length, num_windows +1))

            pi2                                                              =
      mutations.diversity(sample_sets=mutations.samples(population=1),
      windows=np.linspace(0, sequence_length, num_windows +1))

            Tajimas_D1                                                       =
      mutations.Tajimas_D(sample_sets=mutations.samples(population=0),
      windows=np.linspace(0, sequence_length, num_windows +1))

            Tajimas_D2                                                       =
      mutations.Tajimas_D(sample_sets=mutations.samples(population=1),
      windows=np.linspace(0, sequence_length, num_windows +1))

            d                                                                =
      mutations.divergence(sample_sets=[mutations.samples(population=0),
      mutations.samples(population=1)],                windows=np.linspace(0,
      sequence_length, num_windows +1))

            Fst  =  mutations.Fst(sample_sets=[mutations.samples(population=0),
      mutations.samples(population=1)],                windows=np.linspace(0,
      sequence_length, num_windows +1))

            data  =  [N1,  N2,  tbot,  proportion,  tsplit,  mu,  np.mean(pi),
      np.var(pi),   np.mean(pi1),   np.var(pi1),   np.mean(pi2),   np.var(pi2),
```

```
      np.mean(Tajimas_D),       np.var(Tajimas_D),       np.mean(Tajimas_D1),
      np.var(Tajimas_D1), np.mean(Tajimas_D2), np.var(Tajimas_D2), np.mean(d),
      np.var(d), np.mean(Fst), np.var(Fst)]
            writer.writerow(data)
            i += 1
      else:
            pass
```

Approximate Bayesian Computation (abc package in R) (Csilléry et al. 2012)

A) Calculation of summary statistics (observed data, obs.txt)

B) Parameter file (params.txt): simulation of N1, N2, mu, proportion, tbot, tsplit

C) Sim_sumstas file (sim_sumstats.txt): simulations of pi, Tajima's D, dxy, and Fst

```
#Import the data
obs <- read.table("obs.txt", header=TRUE)
params <- read.table("params.txt", header=TRUE)
sim_sumstats <- read.table("sim_sumstats.txt", header=TRUE)
#Goodness-of-fit
fit <- gfit(target=obs, sumstat=sim_sumstats, nb.replicate=100)
#Cross-validation
cross <- cv4abc(param=params, sumstat=sim_sumstats, abc.out = NULL, nval=100,
tols=c(0.005, 0.01, 0.05), method="neuralnet")
#Parameter inference
rej   <-   abc(target=obs,   param=params,   sumstat=sim_sumstats,   tol=0.01,
method="neuralnet")
#Diagnostic plot
plot <- gfitpca(target=obs, sumstat=sim_sumstats, index="D", cprob=0.05)
```

Code S3. Variant calling

```
bowtie2-build reference_genome.fasta reference_genome

bowtie2 -x reference_genome -1 clean_reads_1.fastq.gz -2 clean_reads_2.fastq.gz
-S alignment.sam

samtools view -S -b alignment.sam > alignment.bam

samtools sort alignment.bam -o alignment_sorted.bam

samtools stats alignment_sorted.bam ⇒ make after alignment

samtools depth  alignment_sorted.bam  |  awk '{sum+=$3} END { print "Average =
",sum/NR}' ⇒ make after alignment

samtools index -b alignment_sorted.bam
```

```
gatk AddOrReplaceReadGroups -I alignment_sorted.bam -O alignment_sorted_rg.bam
-LB lib -PL ILLUMINA -PU unit -SM sample_name

samtools index -b alignment_sorted_rg.bam

samtools faidx reference_genome.fasta

samtools dict reference_genome.fasta

gatk HaplotypeCaller -I alignment_sorted_rg.bam -O genome_variant.gvcf -R
reference_genome.fasta --emit-ref-confidence GVCF
```

Diplolepis rosae

```
gatk CombineGVCFs -R reference_genome.fasta -V genome_variant_1.gvcf -V
genome_variant_2.gvcf … -V genome_variant_n.gvcf -O
combined_genome_variant.gvcf

gatk GenotypeGVCFs -R reference_genome.fasta -V combined_genome_variant.gvcf -
O variant_call_file.vcf
```

Cynips quercusfolii

```
gatk GenotypeGVCFs -R reference_genome.fasta -V genome_variant.gvcf -O
variant_call_file.vcf

bcftools merge variant_call_file_1.vcf.gz variant_call_file_2.vcf.gz …
variant_call_file_7.vcf.gz -Ov > merged_variant_call_file.vcf


java -Xmx100g -jar beagle.22Jul22.46e.jar gt=merged_variant_call_file.vcf
out=phased_variant_call_file ne=500 window=50

bcftools filter -i 'TYPE="snp"' merged_variant_call_file.vcf -Ov >
vcf_snp.vcf

bcftools view --max-alleles 2 --exclude-types indels vcf_snp.vcf -Ov >
vcf_snp_2all.vcf

bgzip vcf_snp_2all.vcf

tabix vcf_snp_2all.vcf.gz
```

Code S4. Identification of genes under selection (McDonald–Kreitman test)

```
java -Xmx100g -jar beagle.22Jul22.46e.jar gt=variant_call_file.vcf
out=phased_variant_call_file ne=500 window=50

bcftools filter -i 'TYPE="snp"' variant_call_file.vcf -Ov > vcf_snp.vcf

bcftools view --max-alleles 2 --exclude-types indels vcf_snp.vcf -Ov >
vcf_snp_2all.vcf

samtools faidx reference.fasta

bgzip phased_variant_call_file.vcf

tabix phased_variant_call_file.vcf.gz
```

```
vcf2fasta.py -f reference.fasta -v phased_variant_call_file.vcf.gz -g
annotaed_masked_from_transposons_file.gff3 -b -e CDS

MKtest -i alignment.fasta -I 1 1 (infoseq alignment.fasta | tail -1 | awk
'{print $6}') -n N
```

-n: number of haplotypes (N = 34 in D. rosae)

```
polydNdS -i alignment.fasta -I 1 1 (infoseq $file | tail -1 | awk '{print $6}')
-O N
```

-O: outgroup individual number (N = 35 corresponding to Diplolepis eglanteriae)

Data to SnIPRE (Eilertson et al. 2012) script (R):

	geneID	FR	PR	FS	PS	Trepl	Tsil	nout	npop
1	g1000	0	0	0	0	409.167	91.8333	10	34
2	g1001	0	0	0	0	163.500	40.5000	10	34
3	g1002	0	7	0	3	1365.280	332.7220	10	34
4	g1003	0	2	0	2	1295.720	294.2780	10	34
5	g1004	0	4	0	1	820.648	193.3520	10	34
6	g1005	0	2	0	1	2015.020	477.9810	10	34

…

```
## Part (1)  Empirical Bayes Implementation  (lme4 package, SnIPRE_source.R)
## Part (2)  Bayesian Implementation (R2WinBUGS package, B_SnIPRE_source.R, and
WinBUGS or OpenBUGS) necessary
setwd("~/Dropbox/SnIPRE_code_JAGS")

#####################################################################
## Part (1)  Empirical Bayes Implementation  (lme4 package)
#####################################################################

source("SnIPRE_source.R")
source("my.jags2.R")
library(lme4)
library(R2jags)
library(arm)


data <- read.table("SnIPRE_table.txt", header = TRUE)  # sample data set

#SnIPRE <-function(mydata)
# mydata: name of data set;
# mydata must have a header with the following columns: PS, PR, FS, FR, npop,
nout, Tsil, Trepl (no particular order)
# outputs 2 objects:  new.dataset & model
# new.dataset contains the estimates of the selection effect, and selection
coefficient (gamma); as well as the estimates of constraint effect (Rest) and
constraint (f).
eb.res = SnIPRE(data)

res = eb.res$new.dataset
model = eb.res$model

write.table(res, file = "eb_results.csv", sep  = ",", row.names = FALSE)
```

```
#####################################################################
## Part (2)  Bayesian Implementation (R2WinBUGS package,
##          B_SnIPRE_source.R, and WinBUGS or OpenBUGS) necessary
#####################################################################

source("B_SnIPRE_source.R")
source("my.jags2.R")
library(lme4)
library(R2jags)
library(arm)


test <- read.table("Snipre_Drosae_Dspinosa_Bayes_check_positive.txt", header =
TRUE)  # sample data set
#BSnIPRE.run <- function(mydata, path = ".", burnin = 500, thin = 5, iter =
2500){
  # path will be where the chains are stored, and must also be where the ".bug"
model is located
  # burnin, thin, and iter (number iterations after burnin) are for MCMC samples
BSnIPRE.run(test, burnin = 10000, thin = 4, iter = 15000)

# check to make sure it finished correctly:
# if a "sample" file is in your working directory (getwd()), or the path you
specified)
# is empty or not there, there is a problem


load("samples")

res.mcmc <- samples

#BSnIPRE <- function(data.mcmc,mydata){
# outputs 2 objects:  new.dataset & effects
# new.dataset contains the estimates of the selection effect, and selection
coefficient (gamma); as well as the estimates of constraint effect (Rest) and
constraint (f).
# the "effects" may be useful if you are interested in estimation
# of population parameters (gamma, constraint) with other assumptions than the
PRF

b.res <- BSnIPRE(res.mcmc, test)

bres = b.res$new.dataset

write.table(bres, file = "bayesian_results.csv", sep  = ",", row.names = FALSE)
```

Code S5. **RNAseq analysis**

Quality control:

```
fastqc file.fastq.gz
```

Alignment:

```
STAR  --runThreadN  20  --runMode  genomeGenerate  --genomeDir  Alignments  --
genomeFastaFiles           /home/ksenia/assembly.fasta           --sjdbGTFfile
/home/ksenia/gene_prediction.gtf --sjdbOverhang 149 --genomeSAindexNbases 13
```

```
STAR --runThreadN 10 --genomeDir Alignments --readFilesCommand zcat --readFilesIn R1.fq.gz R2.fq.gz --chimOutType withinBAM --outSAMtype BAM SortedByCoordinate --outFileNamePrefix Out
```

Genome annotation:

```
braker.pl --genome=assembly.fasta --bam=RNAseq.out.bam --gff3 --useexisting
```

Feature counts:

```
featureCounts RNAseq.out.bam -a gene_prediction.gff3 -T 5 -M -p -t gene -o gene_counts
```

Differential expression analysis (Love et al. 2014):
The pipeline is available at: http://bioconductor.org/packages/devel/bioc/vignettes/DESeq2/inst/doc/DESeq2.html

```
#DESeq2 R package
library(DESeq2)
library(pheatmap)
library(vsn)
library(RColorBrewer)
library(remotes)
library(ggpubr)
library(ggplot2)

#Upload the count matrix
cts <- as.matrix(read.table("C:/cygwin64/home/Ksenia/DESeq2_gene_matrix.txt", header=T))
cts0 <- cts[,c("Egg_1", "Egg_3", "Egg_4", "Larva_July_1", "Larva_July_2", "Larva_July_3", "Larva_July_4", "Larva_Sept_1", "Larva_Sept_2", "Larva_Sept_3", "Larva_Sept_4", "Gland_Oct_1", "Gland_Oct_2", "Gland_Oct_3", "Gland_Oct_4", "Gland_Nov_1", "Gland_Nov_2", "Gland_Nov_3", "Gland_Nov_4", "Head_1", "Head_2")]

#Upload sample type (factor) matrix
coldata <- read.table("C:/cygwin64/home/Ksenia/DESeq2_coldata.txt", header=T, row.names = 1)
coldata0 <- coldata[c("Egg_1", "Egg_3", "Egg_4", "Larva_July_1", "Larva_July_2", "Larva_July_3", "Larva_July_4", "Larva_Sept_1", "Larva_Sept_2", "Larva_Sept_3", "Larva_Sept_4", "Gland_Oct_1", "Gland_Oct_2", "Gland_Oct_3", "Gland_Oct_4", "Gland_Nov_1", "Gland_Nov_2", "Gland_Nov_3", "Gland_Nov_4", "Head_1", "Head_2"),]
coldata0 <- data.frame(coldata0)
coldata0$coldata0 <- factor(coldata0$coldata0)

#Sample distance matrix and visualisation
dds0 <- DESeqDataSetFromMatrix(countData = cts0, colData = coldata0, design = ~ coldata0)
ntd0 <- normTransform(dds0)
sampleDists <- dist(t(assay(ntd0)))
sampleDistMatrix <- as.matrix(sampleDists)
rownames(sampleDistMatrix) <- paste(ntd0$coldata0, sep="-")
colnames(sampleDistMatrix) <- paste(ntd0$coldata0, sep="-")
colors <- colorRampPalette(brewer.pal(9, "PuBuGn"))(255)
pheatmap(sampleDistMatrix,
         clustering_distance_rows=sampleDists,
         clustering_distance_cols=sampleDists,
         col=colors)
```

```
#Count matrices for pairwise comparisons
cts1 <- cts[,c("Larva_July_1", "Larva_July_2", "Larva_July_3", "Larva_July_4",
"Egg_1", "Egg_3", "Egg_4")]

#Factor matrices for pairwise comparisons
coldata1    <-    coldata[c("Egg_1",    "Egg_3",    "Egg_4",    "Larva_July_1",
"Larva_July_2", "Larva_July_3", "Larva_July_4"),]
coldata1 <- data.frame(coldata1)
coldata1$coldata1 <- factor(coldata1$coldata1)

#Create the object dds from the matrices for further differential expression
analysis
dds1 <- DESeqDataSetFromMatrix(countData = cts1, colData = coldata1, design =
~ coldata1)
#log2(n + 1) transformation of the counts for the principal component analysis
ntd1 <- normTransform(dds1)

#Principal Component Analysis
pca1 <- plotPCA(ntd1,intgroup = "coldata1")
pca1

#Differential expression analysis. The function DESeq comprises (1) estimation
of size factors, (2) estimation of dispersion, and negative binomial generalized
linear model fitting for the log2 fold changes in gene counts (see in detail
Love et al. 2014)
dds1 <- DESeq(dds1)

#Plotting dispersion and assessing model fitting
plotDispEsts(dds1)

#Show significant (adjusted p-value < 0.05) log2 fold changes (<0 for down-
regulated genes and  >0 for up-regulated genes)
res1 <- results(dds1)
summary(res1, alpha=0.05)

#Visualise log2 fold changes
plotMA(res1, alpha=0.05)
```

Bidirectional best hits:

```
tblastn        -query        putative_larva_venom_genes_Drosae.aa        -db
Biorhiza_transcriptome.fa -outfmt 6
```

Choose query_seq_ids and subject_seq_ids with the highest bit score from tblastn output (List 1)

```
makeblastdb -in putative_larva_venom_genes_Drosae.aa -dbtype prot -parse_seqids
blastx            -query            Biorhiza_best_hit_seq.fa            -db
putative_larva_venom_genes_Drosae.aa -outfmt 6
```

Choose query_seq_ids and subject_seq_ids with the highest bit score from blastx output (List 2).

Compare List 1 and List 2 and choose the genes showing the same alignments.

Code S6. Population structure of *Cynips quercusfolii*

Pairwise divergence Dxy:

```
pixy --vcf vcf_file.vcf.gz --stats dxy --window_size 10000 --populations
Cynips_sample_1_Cynips_sample_2.txt
```

#Cynips_sample_1_Cynips_sample_2.txt file:

Cynips_sample_1 pop1

Cynips_sample_2 pop2

Per-individual heterozygosity:

```
vcftools --het --vcf vcf_file.vcf
```

Population structure:

```
bcftools +prune -l 0.05 -w 1000 vcf_file.vcf -Ov > vcf_file_filter_1.vcf | sed
's/ID=[[:digit:]]/ID=contig_/g'       vcf_file_filter_2.vcf       |       sed
's/^[[:digit:]]/contig_/g'  > vcf_file_filter_3.vcf

./plink --pca --vcf vcf_file_filter_3.vcf --make-bed --double-id --allow-extra-
chr --out variant_bed
```

Output files used to perform the PCA in R (tidyverse package) (Wickham et al. 2019):

variant_bed.eigenval: the eigenvalues from our analysis
variant_bed.eigenvec: the eigenvectors from our analysis

The pipeline is available at: https://speciationgenomics.github.io/pca/

```
library(tidyverse)
# read the data
pca <- read_table2("variant_bed.eigenvec", col_names = FALSE)
eigenval <- scan("variant_bed.eigenval")
# sort out the pca data
# remove nuisance column
pca <- pca[,-1]
# set names
names(pca)[1] <- "ind"
names(pca)[2:ncol(pca)] <- paste0("PC", 1:(ncol(pca)-1))
# make plot
a <- ggplot(pve, aes(PC, pve)) + geom_bar(stat = "identity")
a + ylab("Percentage variance explained") + theme_light()
# first convert to percentage variance explained
pve <- data.frame(PC = 1:7, pve = eigenval/sum(eigenval)*100)
# calculate the cumulative sum of the percentage variance explained
cumsum(pve$pve)
#plot pca
plot(x=pca$PC1, y=pca$PC2)
```

Code S7. Detection of Runs of Heterzygosity (ROHet) and Runs of Homozygosity (ROH)

Runs of Homozygosity (Danecek et al. 2021)

```
bcftools roh -G 30 -O -r snp.vcf.gz > ROH_table.txt
```

Visualising in R:

```
Ind<-read.table("ROH_table.txt", header=T)
```

```
Ind<-Ind[Ind$V3=="Scaffold_name",]
NumSampleInd<-length(unique(Ind[,2]))
Chrom<-unique(Ind$V3)
Sampleind<-unique(Ind$V2)
SizeScaf<-read.table("Scaffold_sizes.txt",row.names=1)
for(chr in Chrom){
  if(SizeScaf[chr,]>100000){
    plot(c(1,SizeScaf[chr,]),c(0,1),col="white",,main="Scaffold_name",yaxt="n
",xaxt="n",xlab="",ylab="ROH")
    Pos=0
    for(ind in SampleInd){
      NumROH<-length(Ind[Ind[,2]==ind & Ind[,3]==chr,3])
      if(NumROH>0){
        for(j in 1:NumROH){
          rect(Ind[Ind[,2]==ind  &  Ind[,3]==chr,4][j],Pos,Ind[Ind[,2]==ind  &
Ind[,3]==chr,5][j],Pos+1/(NumSampleInd),col="#0072B2",border = NA)
        }

      }
      text(-
0.015*SizeScaf[chr,],Pos+1/(2*(NumSampleInd)),labels=substr(ind,16,nchar(ind)
),col="black",cex=0.6)
      Pos=Pos+1/(NumSampleInd)
    }

    }
  }
```

Runs of Heterozygosity (Biscarini et al. 2018; Purcell et al. 2007):

```
./plink --recode --double-id --allow-extra-chr --vcf vcf_file.vcf --out ROHet
```

Use ROHet.map and ROHet.ped output files in R script:
The script is available at: https://cran.r-project.org/web/packages/detectRUNS/vignettes/detectRUNS.vignette.html

```
#First part ⇒ detecting runs
library(detectRUNS)
slidingRuns <- slidingRUNS.run(
  genotypeFile = "C:/cygwin64/home/Ksenia/Cynips_ROHet.ped",
  mapFile = "C:/cygwin64/home/Ksenia/Cynips_ROHet.map",
  windowSize = 10,
  threshold = 0.05,
  minSNP = 10,
  ROHet = FALSE / TRUE,
  maxOppWindow = 1,
  maxMissWindow = 1,
  maxGap = 10^6,
  minLengthBps = 10000,
  minDensity = 1/10^3, # SNP/kbps
  maxOppRun = NULL,
  maxMissRun = NULL
)


#Second part => visualizing data
Ind<-slidingRuns_het[slidingRuns_het$chrom =="Chromosome_name",]
NumSampleInd<-length(unique(Ind[,2]))
Chrom<-unique(Ind$chrom)
SampleInd<-unique(Ind$id)
```

```
SizeScaf<-read.table("Scaffold_sizes.txt",row.names=1)
for(chr in Chrom){
  if(SizeScaf[chr,]>100000){
  plot(c(1,SizeScaf[chr,]),c(0,1),col="white",,main="scaffold_NC_046657",yaxt
="n",xaxt="n",xlab="",ylab="ROHet")
    Pos=0
    for(ind in SampleInd){
      NumROH<-length(Ind[Ind[,2]==ind & Ind[,3]==chr,3])
      if(NumROH>0){        for(j in 1:NumROH){
          rect(Ind[Ind[,2]==ind  &  Ind[,3]==chr,5][j],Pos,Ind[Ind[,2]==ind  &
Ind[,3]==chr,6][j],Pos+1/(NumSampleInd),col="red",border = NA)
        }

      }
      text(-
0.015*SizeScaf[chr,],Pos+1/(2*(NumSampleInd)),labels=substr(ind,16,nchar(ind)
),col="black",cex=0.6)
      Pos=Pos+1/(NumSampleInd)
    }

  }
}
```

Code S8. Detection of structural variations

```
java      -jar      picard.jar      MarkDuplicates      I=sorted_alignments.bam
O=sorted_alignments_with_tag_duplicated_reads.bam
M=sorted_alignments_with_tag_duplicated_reads_metrics.txt

delly               call               -g               genome_reference.fasta
sorted_alignments_with_tag_duplicated_reads.bam > structural_variations.vcf
```

Intansv R package:

```
result   <-   read.Delly(structural_variations.vcf,   regSizeLowerCutoff=10000,
regSizeUpperCutoff=100000000, readsSupport=5, method=”DELLY”)

result
```