Exploring population history and gall induction in cynipid gall wasps using genomics and transcriptomics

Jack Hearn

University of Edinburgh

Doctor of Philosophy

2013

Table of Contents

List of Figures

List of Tables

Declaration

I declare that this thesis has been written by myself and is my own work, except where indicated. In particular, Chapter 2 was carried out in collaboration with Dr K. R. Lohse (University of Edinburgh). This work has not been submitted for any other degree or professional qualification.

Jack Hearn, 2013

Abstract

Cynipid gall wasps have fascinating biology that has piqued the interest of naturalists throughout history. They induce morphologically complex, sometimes spectacular, gall structures on plants in which the larval stages develop. Gall wasps have therefore evolved an intimate association with their hosts - both metabolically, and in terms of their population histories. Gall wasps must both interact physiologically with their hosts to induce galls, and track their host plants through space and time. My thesis centres on two uses of genomic data in understanding the biology of the oak apple gall wasp *Biorhiza pallida*. I provide a comprehensive investigation into patterns of oak and gall wasp gene expression associated with gall induction, and a population genomic reconstruction of the population history of this species across the Western Palaearctic. While advances in sequencing technology and reduced costs have made these aims possible, analysis of the massive resulting datasets generated creates new challenges.

Firstly, in reconstructing the population history of *B. pallida,* I describe the use of shotgun sequencing and an informatic pipeline to generate alignments of several thousand loci for three *B. pallida* individuals sampled from putative glacial refugia across the Western Palaearctic in Iberia, the Balkans and Iran. This dataset was analysed using a new maximum likelihood method capable of estimating population splitting and admixture among refugia across very large numbers of loci. The results showed an ancient divide between Iberia and the other two refugia, followed by very recent admixture between easternmost and westernmost regions. This suggests that gall wasps have migrated westwards along the North African coast as well as through mainland Europe.

Second, I compare the gene expression profiles of gall wasp and oak tissues sampled from each of three stages of gall development, leading to new insights into potential mechanisms of gall wasp-oak interaction. A highly expressed gall wasp protein was identified that is hypothesised to stimulate somatic embryogenesis-like development of the gall through interaction with oak tissue glycoproteins. Highly expressed oak genes include those coding for nodulin-like proteins similar to those involved in legume nodule formation.

Finally, analysis of the gall wasp genome has revealed potential, but as yet unconfirmed, horizontal gene transfer events into gall wasp genomes. Genes discovered in three gall wasp genomes and expressed in three transcriptomes encode plant cell wall degrading enzymes. They are not of hymenopteran origin, and are most homologous to genes of plant pathogenic bacteria. These genes could be involved in several aspects of gall wasp biology, including feeding and developmental manipulation of host plant tissue.

Acknowledgements

I thank my supervisors Graham Stone and Mark Blaxter for mentoring me throughout this project. I am particularly grateful of the freedom I was given to design experiments around new and exciting sequencing technologies. I don't know of anywhere else I could have undertaken such a project in such a neglected, but fascinating organism as the gall wasp.

I would also like to thank the Blaxter laboratory, present and departed, for helping me become a mildly competent bioinformatician, in particular Graham Thomas, Sujai Kumar, John Davey, Georgios Koutsovolous and Ben Elsworth. From the Stone laboratory, many thanks to Frazer Sinclair, James Nicholls and especially Konrad Lohse for the experiences working together in the field and office over the last few years. The GenePool, led by Karim Gharbi was always willing to advise and help with the many special requests this thesis entailed.

Around the world I have made many helpful friends among the small but dedicated community of gall wasp researchers, including Joe Shorthouse in Canada, Warin Harrison and family in China, and Sebastien Cambier and the EVIL team in France.

Finally, I dedicate this thesis to my parents who taught me to read before I went to school. This hasn't helped with my writing, but makes everything in the world very interesting.

Chapter 1: Introduction

1.1 General aims of this thesis

This thesis uses the recent revolutionary advances in genome and transcriptome sequencing to investigate two intriguing aspects of gall wasp (family: *Cynipidae*) biology - how the gall wasp induces plant galls on hosts, and the phylogeography of gall wasps. The new technologies make a previously difficult to work with non-model organism much more accessible to study. Now it is possible to make powerful inferences by sampling natural populations of species. This thesis explores this in two ways: (1) a transcriptomic investigation into the control of gall induction by gall wasps and, (2) ascertaining the Pleistocene history of a Western Palaearctic species by genome-wide model-based phylogeography. Although they are distinct analyses, the data, DNA and RNA sequencing, is generated in the same manner. Furthermore, the dataset for one part of this thesis, for example, genome assemblies used in phylogeograpic inference, has applications in finding candidate genes for gall induction and *vice-a-versa*. Additionally, unforeseen insights into gall wasp biology can be made with genomic and transcriptomic datasets. This thesis demonstrates the versatility of high-throughput sequencing when applied to difficult to manipulate non-model organisms sampled from natural populations.

1.1.2 Overview of the introductory chapter

In this chapter I begin with an introduction of general aspects of cynipid gall wasp biology (1.2). Then the stages of gall induction are detailed and existing hypotheses for control of gall induction by gall wasps discussed (1.3). This is followed by a review of recent developments in phylogeography, particularly in non-model organisms, and the benefits of using gall wasps to study range expansions from glacial refugia (1.4). The gall wasp chosen as the basis of these investigations, *Biorhiza pallida*, is introduced in section 1.5 along with other gall wasp genetic resources used in this thesis. Finally, data sharing between the major aspects of this thesis and brief summaries of chapter objectives are given in section 1.6 and 1.7.

1.2 Cynipid gall wasp biology

1.2.1 Galling is a widespread trait

Galling is a widespread trait that has evolved repeatedly in viruses, prokaryotes, fungi, nematodes and arthropods. Galls are formed by host tissues from manipulation of host gene expression by the inducer, and therefore are an example of an extended phenotype (Dawkins, 1982). The process can result in anything from a cryptic swelling to spectacular structures that can be mistaken for fruits or inflorescences by the unaware. Arguably the most complex and beautiful galls are those induced by cynipid gall wasps (Hymenoptera, superfamily Cynipoidea) on a taxonomically diverse range of plant hosts.

Although galling has piqued the interest and comment of many naturalists from Hippocrates to Darwin (Harper et al., 2009), induction is well understood only for the bacterial crown galler *Agrobacterium tumefaciens*. During infection the bacterium transfers tumour-inducing (Ti) plasmids into host cells, and the plasmids subsequently integrate into the host genome and dictate expression. The Ti plasmid is now a powerful tool in plant genetic engineering as a gene of interest can be cloned into the plasmid and targeted to specific tissues. Much research has also focused on galls formed by plant-pathogenic nematodes and the Hessian fly *Mayetiola destructor* because of their impact on yields of economically important crops worldwide. But for Hessian flies and the cynipid gall wasps the precise mechanism of induction remains unknown, in part due to the greater complexity of the interaction between galler and host than for *A. tumafaciens*.

1.2.2 Cynipid gall wasps

There are *circa* 1400 described species of cynipid gall wasps, second only to gall midges (Diptera, family Cecidomyiidae) in diversity of galling arthropods. It is hypothesised that the phytophagous gall inducing life cycle evolved from an ancestral ectoparasitic parasitoid of larvae developing within plants such as those of wood-boring insects (Liljeblad and Ronquist, 1998). The splitting time from their entomophagous sister taxa, the *Figitidae*, is estimated to have occurred approximately 127 million years ago (Buffington et al., 2012).

The true diversity of cynipid gall wasps is much greater than currently described, as many regions, especially temperate China and tropical Southeast Asia, and potential hosts, like *Nothofagus* (family: Fagaceae, predominantly found in South America) and species of Asteraceae, have not been extensively surveyed. Consequently new species continue to be discovered and described (figure 1.1) (Pujade-Villar et al., 2010; Ide et al., 2010; Melika et al., 2011; Tang et al., 2011, Tang et al., 2012; Melika et al., 2013).

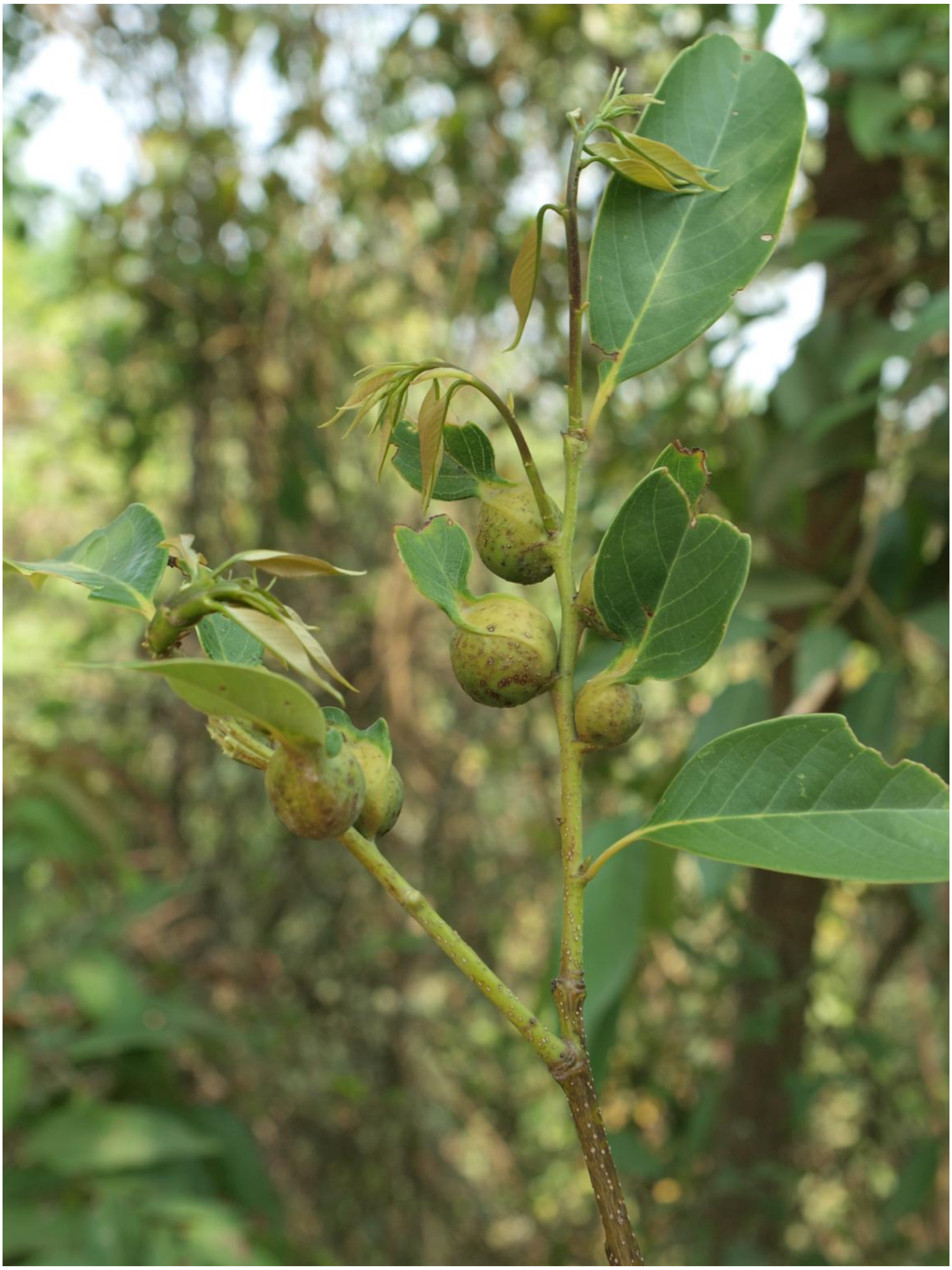

Figure 1.1. Gall of new species of gall wasp discovered by J. Hearn on a species of *Castanopsis* in Yunnan, China. Photo courtesy of Chang-ti Tang.

At approximately 1000 described species, Oak cynipids (tribe: Cynipini) have the greatest species richness of the five cynipid tribes) (Ronquist & Liljeblad, 2001). This figure however does not include gall wasps on other Fagaceae like *Castanopis* or *Lithocarpus*. The Cynipini have speciated greatly after host shifting on to trees of the Fagaceae family, the beeches and oaks. Subsequent host-shifts are very rare among gall wasps compared to the rate of host shifting for other plant-insect interactions (Ronquist & Liljeblad, 2001). When host shifts have occurred, the taxonomic distance between hosts can be great. For example, a gall wasp within the genus *Diastrophus* has shifted from a eudicot to a monocot host (Ronquist & Liljeblad, 2001). Following from this, it has been speculated that the mechanism of gall induction utilises deeply conserved plant pathways (G. Stone, personal communication). The oak cynipids induce the most morphologically diverse of cynipid galls (figure 1.2) possibly due to selection pressure from hymenopteran parasitoids (Stone & Schönrogge, 2003; Bailey et al., 2009). Not surprisingly this makes them the most popular tribe of gall wasps for study. The second most speciose gall wasp tribe, the *Rosa* galling Diplolepidini at 63 species, approaches the Cynipini in external gall complexity.

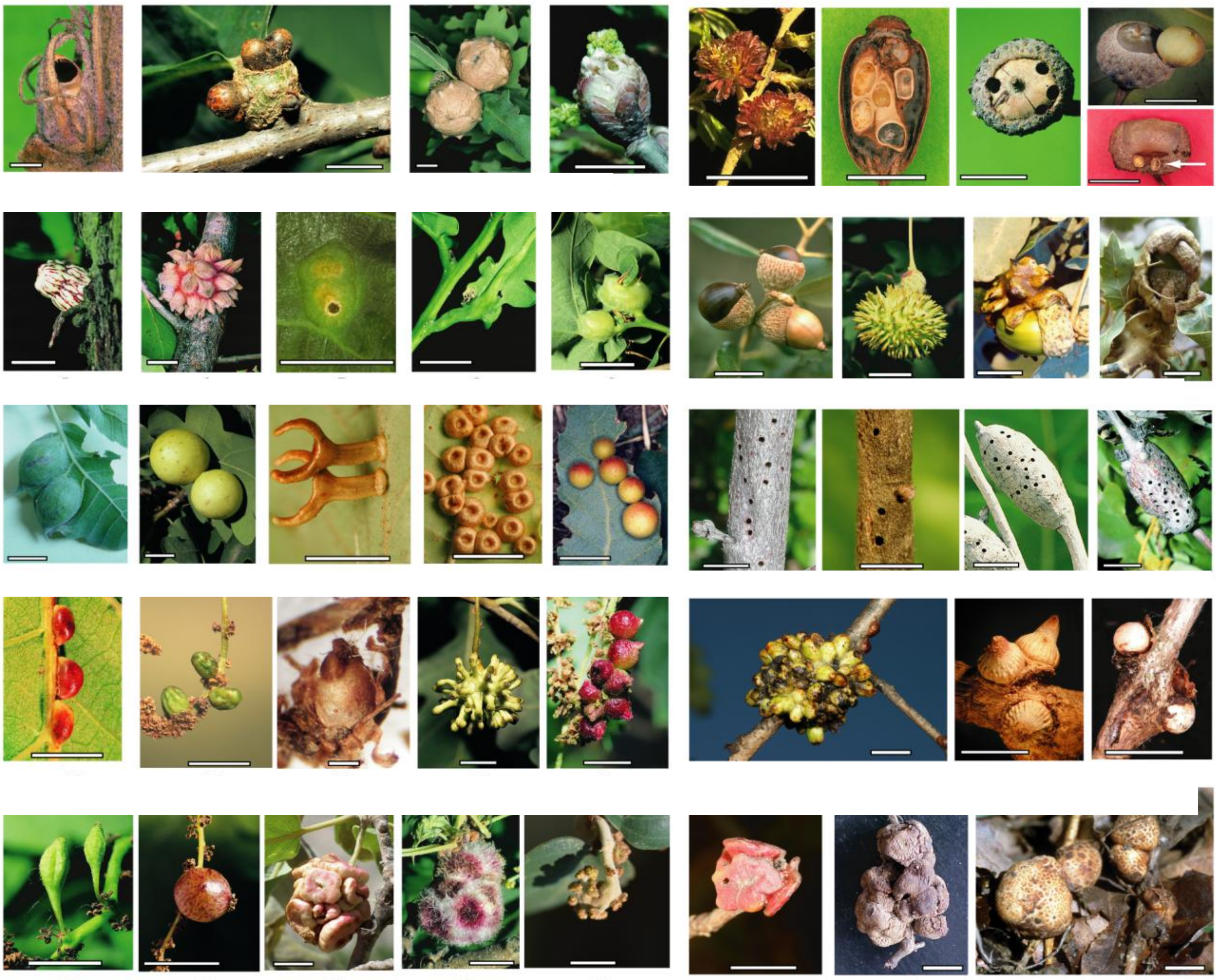

Figure 1.2. Diversity of galls induced by Cynipini gall wasps on various plant tissues (compiled by G Stone. Images by Graham Stone, György Csóka, Jose-Luis Nieves-Aldrey).

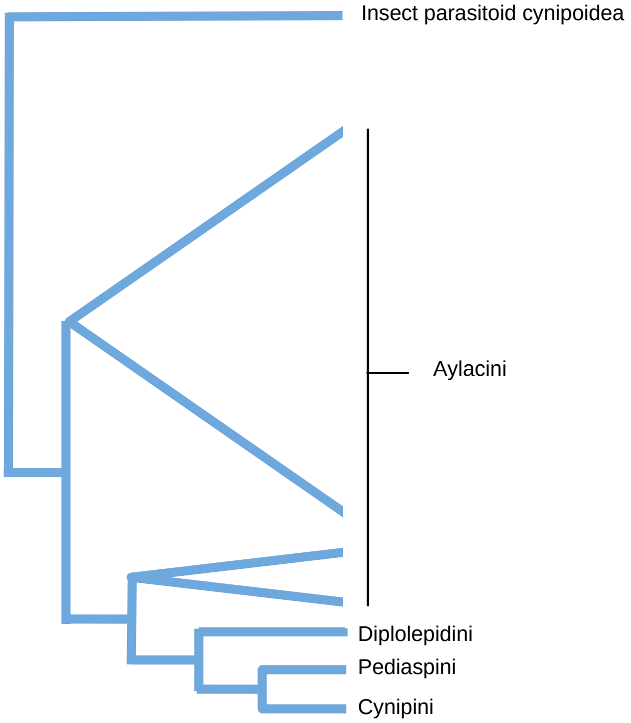

Figure 1.3. Phylogeny of gall wasp tribes demonstrating paraphyly of the Aylacini; Synergini inquilines not included, as they are unresolved. Adapted from Ronquist & Liljeblad (2001).

The herb-galling (Asteraceae) Aylacini tribe is paraphyletic but one of its lineages is probably basal within the *Cynipidae* (figure 1.3), and Ronquist & Liljeblad (2001) propose the Papaveraceae as the original host of the cynipid galls. They induce on several plant families including Asteraceae, Papaveraceae, Lamiaceae and Rosaceae. Many Aylacini have cryptic galls within stems of their hosts and as such have been less attractive as research subjects and many species and novel hosts are probably as yet unidentified. These cryptic stem galls probably represent the ancestral relationship between gall wasp and host (Ronquist & Liljeblad, 2001).

Within the *Cynipidae*, species of the tribe Synergini have evolved to attack and develop within galls of other gall wasps. They do not induce their own galls but can modify galls of other cynipids (Csóka et al., 2005). The original inhabitant of the gall may be squashed by these modifications. Species that live commensally with their hosts like this are called inquilines.

Where the Synergini sit within the gall wasp phylogeny is unknown and the tribe may be paraphyletic or polyphyletic as secondary loss of induction could have evolved many times (Nylander, 2004).

1.2.3 Many Cynipids have complex life cycles

Many rose (tribe Diplolepidini) and herb (tribe Aylacini) gall wasps have one sexually reproducing generation per year. This occurs by facultative arrhenotoky: unfertilised eggs give rise to haploid males and fertilised eggs to diploid females (Csóka et al., 2005). Parthenogenesis is also common and probably due to *Wolbachia* infection of effected species (Plantard et al., 1998; Plantard et al., 1999). In Hymenoptera the infection often induces duplication of gametes after meiosis creating homozygous diploid offspring (Plantard et al., 1998; Csóka et al., 2005). Due to the haplo-diploid sex determination of Hymenoptera this results in species consisting completely of diploid females (Csóka et al., 2005).

In contrast, the oak (tribe Cynipini) and sycamore (tribe Pediaspini) gall wasps have some of the most complex known life cycles. They undergo heterogony, meaning they alternate strictly between sexual and asexual generations (Csóka et al., 2005), which is a very rare form of life cycle among Metazoa. Both generations are completed in a year for most Cynipini, although there are exceptions where the asexual generation requires more than a year to develop (Askew, 1984; Csóka et al., 2005). The asexual generation females can be of three types: (1) androphores producing only males, (2) gynophores produce sexual females and (3) gynandrophores which give rise to both. Gall wasp gynephore asexuals contradict the complementary sex determination (CSD) found in Honeybees (*Apis mellifera*). The CSD model requires a heterozygous sex-detemining locus is required to produce females and homozygous diploids result in males. It is unknown how gynephores produce females because diploidisation would lead to homozygotes, and therefore males under the CSD model. Most gall wasp species have only androphore and gynophore females (Csóka et al., 2005). Figure 1.4 shows the lifecycle for several European species of the *Andricus, Cynips* and *Neuroterus* genera. The life cycle is further complicated in some species of *Andricus* and *Callirhytis* as they also alternate host between generations. For host-alternating *Andricus* species the asexual female oviposits on section *Cerris* oaks and the sexual female on section *Quercus* oaks (Csóka et al., 2005). The complexity and duration of gall wasp life cycles makes them challenging to study. By contrast *Drosophila melanogaster* (Diptera: Drosophilidae), an arthropod model organism, is easily reared in laboratory conditions and can complete a generation in 10 days.

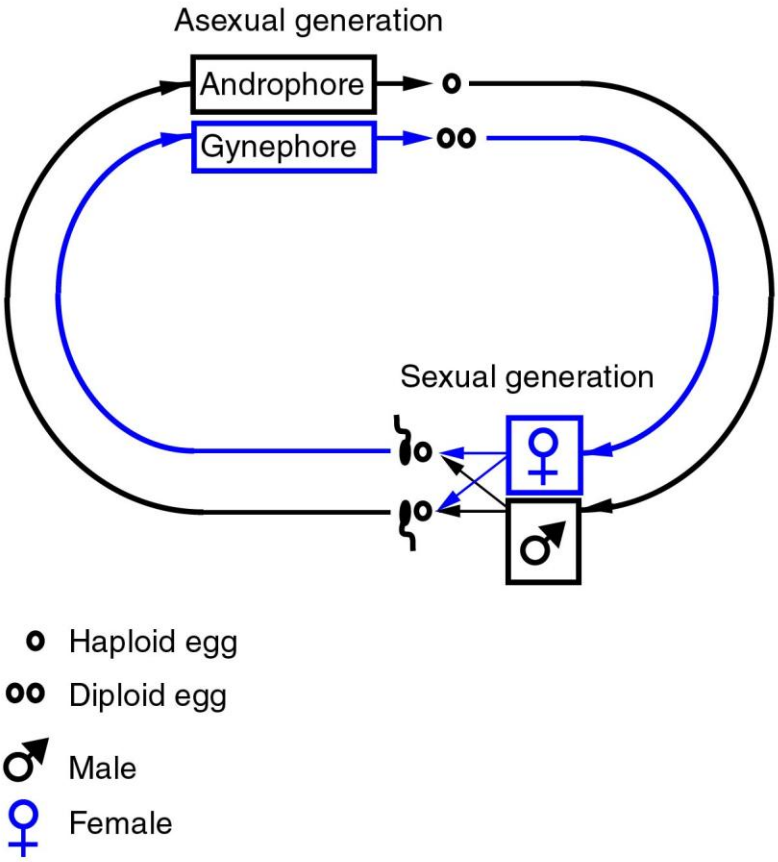

Figure 1.4. The complex bigenerational life cycle of *Andricus*, *Cynips* and *Neuroterus* oak gall wasps. Image courtesy G. Stone.

1.2.4 Oak gall wasp communities

Cynipid gall wasps are associated with a community of species, best studied in the Cynipini tribe in the Western Palaearctic. A gall can act as a useful pre-existing home for many inquilines including cynipid Synergini but also moths, midges and beetles (Stone & Schönrogge, 2003). A larva may be eaten by chalcid parasitoids oviposited directly into the larva or its chamber or by caterpillars, birds and rodents that attack from the outside. Furthermore, parasitoids may themselves be parasitised by other chalcids, an example of hyperparasitism (Sullivan & Völkl, 1999).

Complex multi-trophic interactions exist in oak gall communities. For example galls induced by species of *Andricus, Disholcaspis and Dryocosmus* produce nectar, which attracts ants that protect the gall from parasitoid attack (Stone et al., 2002). This is a tetra-trophic interaction of (1) oak, (2) gall wasp, (3) ants and (4) parasitoids. Because of the discrete structure of the galls there is an intimate association between trophic levels, particularly between plant host, gall wasp and its parasitoids.

1.2.5 Western Palaearctic phylogeography and oak gall wasps

The phylogeography of the Western Palaearctic is shaped by periods of glaciaton and the gall wasps are no exception. Species expand from refugia in Southern Europe into Northern Europe during interglacial periods and back again during glacial. The grasshopper, *Chorthippus parallelus*, is the classic insect example of refugial specific haplotype structure in Northern Europe (Hewitt, 1999). Postglacial northern European populations of this grasshopper are derived from the Balkan refugia, and the Pyrenees and Alps have acted as a natural barrier to recolonisation from Spain and Italy respectively (Hewitt, 1999). Southern European Refugia have persisted through the Pleistocene because of their mountainous nature, species could move up and down in altitude depending on environmental conditions (Hewitt, 1999).

European gall wasps migrated from Asia into Europe in the Late to Early Pleistocene, 1.3–4.2 million years ago, reaching Iberia approximately 400 000 years ago (2.5%–97.5% quantiles of 0.1-0.7 million years ago) (Stone et al., 2012). Since appearing in the Western Palaearctic changes in oak gall wasp ranges have been shaped by glacial and inter-glacial periods in a similar fashion to *C. parallelus*. Along with their hosts, and most European species of flora and fauna (Hewitt, 1999), oak gall wasps retreated to and expanded from southern refugia in Spain, Italy/the Balkans and Turkey eastwards into Iran.

Oak gall wasp host species, the *Quercus* section *Quercus* trees are a keystone taxon in Europe supporting more insect species than other forest trees (Kelly & Southwood, 1999). They have the same southern refugia as other species, and possibly another in the Caucasus. Their postglacial re-colonisation in the current interglacial period occurred from Iberia, Italy and the Balkans based on haplotypes of chloroplast DNA present in Northern Europe (Hewitt, 1999).

1.3 The stages of cynipid gall induction

There are three clearly defined phases of gall growth: initiation, growth and maturation (figure 1.5) (Harper et al., 2009). These stages are easy to differentiate visually and are natural sampling points for characterising and contrasting gene expression of gall wasp larval and host plant tissues during gall development. By the mature stage the gall has stopped growing and the larvae feed. Following this, I hypothesise that gene expression profiles of both host and inducer will change dramatically from growth to mature tissues. Additionally, although there is great variation in the outer tissues among galls of different gall wasps, the organisation of the inner tissues is highly conserved across species and host location (Stone & Schönrogge, 2003).

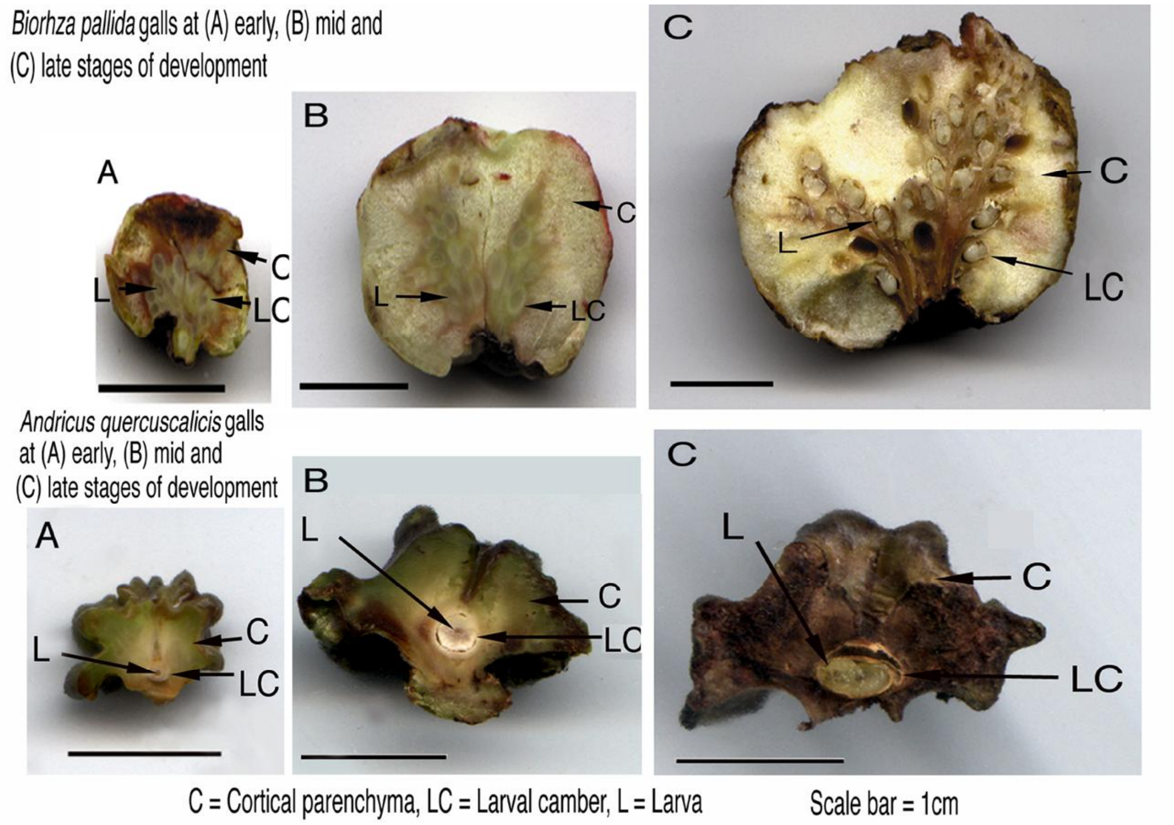
Biorhza pallida galls at (A) early, (B) mid and (C) late stages of development
A
B
C
C
L
LC
Andricus quercuscalicis galls at (A) early, (B) mid and (C) late stages of development
C = Cortical parenchyma, LC = Larval camber, L = Larva
Scale bar = 1cm

Figure 1.5. The two post-induction stages of *Biorhiza pallida* and *Andricus quercuscalicis*, multi- and single-locular gall inducers respectively. Note the progression from (A) early growth gall through to (B) mid growth to (C) mature galls. Figure courtesy of K. Schönrogge.

1.3.1 Stage 1, Induction:

A female gall wasp will oviposit one or multiple eggs in a highly specific site at the key developmental point of a specific tissue (Rohfritsch & Shorthouse, 1982). Along with the egg, maternal factors may be introduced to the host to facilitate induction. The egg must avoid or neutralise the host's immune responses after oviposition for successful gall induction. Observation of failed oviposition events and variation between trees in susceptibility to galls suggest host-immunity acts at this point in development.

Induction can occur on any plant organ including roots, flowers, fruits and buds. There is a very high specificity of the species and generation of the gall wasp to induction location. For successful induction plant tissues surrounding the egg must be capable of developing into different cell types (Harper et al., 2009). Females of some species will lay multiple eggs at one oviposition site to form a multilocular gall. In such cases, the burden of parasitism is possibly reduced by the greater size of the gall and 'sacrifice' of outer chambers to parasitoids (Stone et al., 2002).

Interactions between the galler egg and plant cell walls appear to be key to successful induction. Plant cells surrounding the egg are lysed by cellulolytic enzymes integrated into the egg wall (Shorthouse et al., 2005; Harper et al., 2009). A cavity is created into which the egg snuggly sits (figure 1.6). Genes for such enzymes are rare in metazoan genomes although they are found in some arthropods, such as termites (Davison & Blaxter, 2005; Pauchet et al., 2010). This leads to the hypothesis that gall wasps have evolved or acquired plant cell wall degrading enzymes to facilitate cell lysis. The potential presence of plant cell wall degrading enzymes raises the additional question: are these genes integrated into the gall wasp genome or encoded by a symbiont? Termites vary among species with both symbiont encoded and horizontally transferred, cellulolytic enzymes genes (Watanabe et al., 1998).

If such genes are present in the genome of the gall wasp they could result from horizontal gene transfers. Candidates are the plant host (or ancestral host) due to the intimacy of the host-galler relationship, bacteria, fungi and viruses (i.e. Virus-Like-Particles, section 1.3.5). There are many species of phytophagous plant cell wall degrading bacteria and fungi that are potential donors. Horizontal genetic transfer events from bacteria into phytophagous insects, especially among beetles, are being discovered as more non-model organism genomes and transcriptomes are sequenced (Pauchet et al., 2010; Acuña et al., 2012; Syvanen, 2012; Keeling et al., 2013).

After approximately seven days, varying by species, the egg hatches and the surrounding cells are by now de-differentiated and become wound callus-like cells (Harper et al., 2009). Whether the de-differentiation of surrounding cells is a maternal affect due to substances oviposited alongside the egg or through effectors secreted by the egg is unknown. The larva now enters a space that has formed beneath the egg, triggering rapid gall growth and tissue differentiation and transition to the growth phase. Intriguingly the larva grows very little during this phase and is assumed not to feed; it is thought that the larva 'concentrates' on induction (Harper et al., 2009).

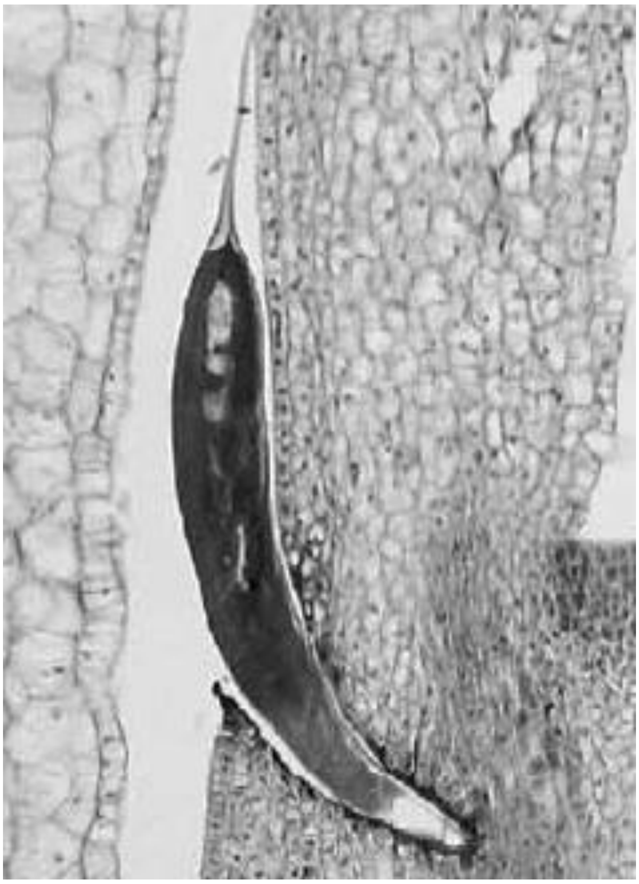

Figure 1.6. Egg of *Diplolepis spinosa* oviposited into the apical meristem of *Rosa blanda* demonstrating lysis of surrounding cells, image courtesy J. Shorthouse.

1.3.2 Stage 2, Growth:

Large nutritive cells deriving from a layer of nutritive parenchyma form around the larval chamber. Nutritive cells are the only food source available to the larva and are free from secondary plant compounds that inhibit feeding, such as tannin. The nutritive cells have high concentrations of proteins, lipids, ribosomes and nitrogenous compounds compared to non-nutritive parenchyma (Harper et al., 2009). They have also undergone many rounds of endoreduplication of the nucleus, presumably to produce the nutrients required by the gall wasp in sufficient concentrations (Harper et al., 2004).

Harper et al. (2004) isolated biotin carboxylase carrier protein (BCCP) from the nutritive cells of several gall wasps. BCCP is a protein found highly expressed in seeds of *Brassica napus* (Harper et al., 2009; Elborough et al., 1996). It is a component of the triacylglycerol lipid synthesis pathway and the resulting lipids are an energy rich food source for larvae. Many rounds of endoreduplication are necessary for nutritive cells to reach large sizes; the total number of rounds varies across galls according to inducer species (Harper et al., 2009). These processes mirror those occurring in nutritive cells of developing seeds, leading Schönrogge et al. (2000) to propose the 'galls-as-seeds' hypothesis. Under this hypothesis the inducer manipulates host seed development pathways to form nutritive tissues.

Turning host cells into nutritive factories surrounding the feeding site is also analogous to strategies of other intimately host-associated phytophagous species. Gall midges and cyst nematodes induce large high-expression cells by endoreduplication (Stuart et al., 2012) whereas root knot nematodes induce giant syncytial cells (Mitchum et al., 2012). Similarly to gall wasps, the juvenile forms of these species can feed on the nutrients within these cells. The mechanism of syncytial cell formation is unknown but it has been demonstrated that proteins secreted by the nematode localise to host-cells (Mitchum et al., 2012).

As the growth phase continues, the larval chamber enlarges, an outer layer of parenchyma develops around the nutritive cells, and gall tissue vascularises and connects to the vascular network of the host (Csóka et al., 2005). As a result the gall now acts as a resource sink for plant-fixed carbon and mineral nutrients while non-nutritive parenchyma concentrate tannins and phenolics (Csóka et al., 2005). This is easily observed by dissecting mature galls as the gall tissues oxidise rapidly on exposure to air. The outer gall cortex develops and the epidermis differentiates into species-specific structures (Harper et al., 2009). The gall itself grows rapidly during this stage but the larva(e) remain(s) small and feed(s) very little. Figure 1.5 demonstrates this in the lack of change in size of larval chambers from early (A) to growth (B) stage.

1.3.3 Stage 3, Maturity:

The gall enters the maturation phase when a layer of sclerenchyma develops within the parenchyma, splitting it into internal nutritive parenchyma and external gall parenchyma (C, figure 1.5). A mature gall ceases to be a resource sink for plant metabolites and may lignify as a defence against herbivores and parasitoids and/or detach from the host and fall to the leaf-litter (Csóka, et al., 2005). The larva feeds until the nutritive cells have been consumed and a sclerenchyma lining has been reached; to avoid fouling the food source it does not defecate (Csóka, et al., 2005). The mature stage is another candidate for plant cell wall degrading enzyme expression as the larvae needs to break down the nutritive cells to access lipids and carbohydrates. In this case the genes would be expressed by the salivary glands. Alternatively such enzymes could be expressed in the gut to break down plant cell walls into digestible products as occurs in phytophagous beetles (Pauchet et al., 2010).

Finally, the larva pupates, and may diapause according to environmental conditions, before emerging as an adult, and defecating. The adult will then mate, or not if it is an asexual generation female, and females oviposit. The gall wasp life cycle then begins again.

1.3.4 Dissecting an extended phenotype: hypotheses of gall Induction by cynipid gall wasps

The control of gall induction and growth by the eggs and larvae of cynipid gall wasps remains a mystery. Many compounds and/or mechanisms including RNA, DNA, viruses, proteins, plant hormones, oligosaccharides, arabinogalactan proteins (AGPs), NOD factors and physical action have been put forward as candidates for induction (Harper et al., 2009). These hypotheses are evaluated in this thesis, and new ones proposed based on the RNA sequencing experiment (chapter 3). It is known that killing eggs and larvae during gall formation will halt the process (Beijerinck, 1882). This observation, which confirms that the gall is an extended phenotype of the inducer, has since been replicated (Rohfritsch & Shorthouse, 1982) but not expanded upon.

The lack of understanding of cynipid galling reflects the difficulty in studying a taxon requiring a minimum of 1 year per (pair of) generation(s), a host to develop on, and the difficulty of manipulating mating and oviposition. These factors combine to make experiments manipulating gall wasp biology prior to the introduction of new sequencing technologies daunting. Research focusing on changes in the host gene expression has been more enlightening and has led to the hypothesis that galls are similar in their host plant gene expression to seeds (section 1.3.2) (Schönrogge et al., 2000; Harper et al., 2004).

1.3.5 Virus-like-particles

The problem of gall induction can be broken down into two segments: firstly, how is the gall inducing material transferred to the hosts; and secondly, how are host cells manipulated to form a gall. A hypothesis addressing the first segment is that inducing stimuli are transferred as virus-like-particles (VLPs) from galler to host, as proposed by Cornell (1983). This is a potential mechanism for the inducer to transfer the key substance(s) to the host. Cornell used an argument by analogy with endoparasitoid wasps that utilise VLPs to suppress host immune responses at oviposition (Whitfield & Asgari, 2003), although with VLP transmission controlled by the gall wasp larva and not as a maternal effect. VLPs of parasitoid wasps suppress their insect host immune responses. Bezier et al. (2009) demonstrated that braconid wasp VLP (bracoviruses) packaging proteins are of viral origin while the viral genome they carry is of wasp origin. Thus the wasps have co-opted a novel method of delivering key components for successful parasitism into the host. For cynipids VLPs could be introduced at oviposition or continuously by the larva(e) throughout gall induction and growth.

1.3.6 Secreted proteins

Another possibility is the secreted proteins observed in plant-pathogenic nematodes and gall midges like the Hessian fly (Mitchum et al., 2012; Stuart et al., 2012). In the plant-parasitic nematodes these proteins are characterised by a signal peptide and localisation to the host's extracellular matrix, apoplast, cytoplasm or nucleus cell (Mitchum et al., 2012). Their functional effects are poorly understood but some are candidate 'effector' proteins for host manipulation and suppression of immune responses (Mitchum et al., 2012). They appear to use molecular mimicry of host proteins to manipulate host expression and developmental changes (Mitchum et al., 2012). The Hessian fly is similar - more than 50% of first-instar larvae salivary glands transcripts encode a signal peptide (Stuart et al., 2012). This is the larval stage at which a compatible wheat-Hessian fly reaction occurs. Less than 5% of these transcripts have similarities to known proteins and many show evidence of positive selection (Stuart et al., 2012). The secreted proteins appear to have evolved with the galling trait in both Nematodes and gall midges. They do not have orthologs in non-galling Nematodes or midges respectively (Mitchum et al., 2012; Stuart et al., 2012).

1.3.7 NOD factors and other glycosylated molecules

Various kinds of oligosaccharide containing compounds are known to be important in plant signalling and development. The best understood gall-inducing compounds are the lipo-chitooligosaccharides, or NOD factors, of the *Rhizobium*-legume nitrogen fixing symbiosis. The lipid side chains of NOD factors are species-specific and together with the chito-oligosaccharide backbone activate host plant early nodulin genes (ENOD). ENOD genes may well represent core genes of plant development that are switched on to create the highly specialised *Rhizobium*-legume nodules. As such they represent a class of candidate genes for gall formation in cynipid galls, albeit with or without a NOD factor trigger.

Genes encoding arabinogalactan protein (AGPs) are known ENOD genes with a wide variety of plant roles compatible with gall formation. AGPs are proteoglycans consisting of less than 10% protein, the rest being predominantly arabinosyl and galactosyl residues (Schultz et al., 1998). AGPs are capable of rescuing somatic embryogenesis in arrested embryos of *Daucus carota* mutant lines. Somatic embryogenesis is the development of a plant from cells of somatic origin not normally responsible for embryogenesis (Bhojwani and Dantu, 2013).

The very first events of gall induction, the dedifferentiation of host cells surrounding the gall wasp egg and newly hatched larva, is similar to the process of somatic embryogenesis. In 2001, van Hengel et al. using endogenous carrot chitinases demonstrated that arabinogalactan proteins are capable of controlling somatic embryogenesis in carrots. It was already known that a carrot temperature-sensitive mutant, *ts11*, developmentally arrested by non-permissive temperatures at the globular, or first, stage of somatic embryogenesis is rescued by the addition of chitinase (De Jong et al., 1992; Kragh et al., 1996). Arabinogalactan proteins were candidate substrates for these chitinases as they contain cleavage sites hydrolyzable by chitinases (van Hengel et al., 2001). Van Hengel et al., (2001) compared the effect of treating carrot wild type seed protoplasts treated with either arabinogalactan proteins or arabinogalactan proteins incubated with chitinase, and control protoplasts. In controls, removal of cell walls caused a 20-fold drop in somatic embryogenesis in carrot seeds compared to normal levels. Addition of arabinogalactan proteins alone increased the rate of somatic embryogenesis to normal levels. Furthermore, addition of carrot chitinases gave a 50% increase in somatic embryogenesis over the effect of treatment with arabinogalactan.

Additional experiments showed the effect of protoplast incubation with arabinogalactan protein and chitinase to be both species- and temporally-specific. Interestingly, the active chitinases are secreted by cells that do not themselves undergo somatic embryogenesis, but act on cells that do (van Hengel et al., 1998). Modification of AGP side chains by secreted gall wasp enzymes (1.3.2), such as chitinase, is a potential mechanism of host manipulation to produce somatic embryogenesis-like results (K. Schönrogge, personal communication).

1.3.8 Manipulation of other host glycoproteins

Along with AGP, xyloglucan and pectin are found in plant cell walls and are known to transduce signals. The break down products of xyloglucan and pectin elicit defence or growth responses depending on the size of the oligomer produced. Cynipid eggs have pectinase activity when oviposited (Shorthouse et al., 2005), (figure 1.6) creating a cavity for the egg in host tissue. As pectinases are embedded in the egg surface a maternal effect is hypothesised with pectinases inserted during ovogenesis, although this has not been experimentally verified.

Additional roles for pectinase molecules are possible. Harper et al. (2009) hypothesise that cell wall loosening from the lysis of pectins, and presumably xyloglucans, could allow a large signalling molecule to permeate cell walls and induce galls.

1.4 Phylogeography: many individuals or many genes?

Phylogeography is the study of past events that result in the geographical structuring we observe in present populations of species. This is of great intrinsic interest, but is also important to other disciplines of biology both pure and applied.

Knowledge of the demographic processes shaping current populations is necessary to make unbiased ecological inferences about species. For example food webs show the relations between trophic levels of an ecological community, but the rules governing how species assemble into webs are poorly understood. Phylogeography provides a framework for testing competing hypotheses of community assembly. This is important because stable phylogeographic associations between constituent species of a food web over time predict strong coevolution and high sensitivity of food webs to species gain/loss. Whereas shuffling of species in communities by contrasting phylogeographic histories of component species predicts diffuse coevolution and greater food web resilience (Memmott, 2009).

Phylogeography also has a role in understanding the ability of invasive species to thrive in new environments, often at great economic cost. Until recently there was little economic impact of gall wasps in Europe. However, the introduction of the chestnut gall wasp, *Dryocosmus kuriphilus,* from the Far East to Southern Europe is changing this somewhat. The European chestnut *Castanea sativa* is also highly susceptible to this gall wasp. Galls on susceptible leaf tissue can reach very high. Fruit yield is reduced and the host may die (EFSA Panel on Plant Health, 2010) leading to the destruction of irreplaceable stands of ancient European chestnuts. By identifying the source population of invasive *D. kuriphilus* (assuming a single invasion event) in its native East Asian range the host chestnut populations can be found. This can potentially aid control of *D. kuriphilus* in Europe by study of native host resistance to the gall wasp.

Many phylogeographic studies generate data from few loci across large numbers of individuals to infer population histories. Such studies are shaped by the nature of Sanger sequencing. It is relatively inexpensive to design primers that amplify loci up to 1000 bases long and then sequence that region in many samples, but not to scale this approach to whole genomes. But identifying loci to amplify is a challenge in itself and the approach becomes prohibitively expensive and labour-intensive with increasing numbers of loci (Lohse et al., 2010). There is also a limit to dataset size in the tens of loci (Lohse et al., 2010). It has also been common to rely on mitochondrial sequence only, or in combination with one or two nuclear loci because of the relative ease of doing so (Rokas et al., 2001). This is potentially misleading as mitochondrial genome history may differ from that of the nuclear genome, often in insect species because of infection by cytoplasmic *Wolbachia* (Rokas et al., 2001). Secondly when only one locus is used, mitochondrial or nuclear, spurious population histories may be inferred. This is because of the stochastic nature of the coalescent (Rosenberg & Norborg, 2002). To control for coalescent variation and accurately infer population histories within a species multiple unlinked loci must be analysed. This is because unlinked genes within a genome that have experienced the same demographic events are independent replicates of the coalescent (Rosenberg & Norborg, 2002).

1.4.1 Genome-wide phylogeography in non-model organisms

Genome-wide shotgun sequencing in combination with coalescent modelling has the potential to revolutionise phylogeography. It is now possible to sequence low-coverage draft genomes or sample thousands of SNPs across the genome at (constantly decreasing) affordable cost. Genome-wide inference of a non-model organism in phylogeography was first applied to resolving the postglacial history of the pitcher plant mosquito *Wyeomia smithii*, a temperate North American species (Emerson et al., 2010). Previous, allozyme based analyses had been unable to differentiate postglacial range changes of the mosquito 19-22 000 years ago (Armbruster et al., 1998). Like gall wasps, *W. smithii's* range follows that of its host, in this case *Sarracenia purpurea*, across Canada and south to the Gulf of Mexico

(Emerson et al., 2010). From sampling across *W. smithii's* range, Emerson et al. (2010) sequenced restriction-site-associated DNA sites (RAD tags). They identified 3 741 single nucleotide polymorphisms (SNPs) across *W. smithii's* 836 megabase genome in these RAD-tags, an unprecedented dataset. A combined cytochrome oxidase 1, a mitochondrial gene, and SNP phylogenetic tree was constructed. The tree demonstrated *W. smithii's* range expansion form a southern Appalachian refugium followed the retreat of the Laurentide Ice Sheet and prevailing winds. It first spread up the Atlantic coastline and then west into Canada.

Although the method of generating the dataset was novel, the *Wyeomia smithii* analysis reflects older tree-based thinking in phylogeography (Nichols, 2001), in which the results are interpreted *post-hoc*. Newer coalescent modelling based methods, like Lohse's (2011) likelihood model (section 1.4.3) are superior, as competing models for postglacial range expansion are pre-specified and tested against one another in a likelihood framework.

Until this thesis, genome-wide methods had not been applied to phylogeography in cynipid gall wasps. Previous gall wasp phylogeography studies have sampled few markers, using little of the total information content in the genomes (Stone & Sunnucks, 1993; Rokas et al., 2001, 2003; Stone et al., 2007; Challis et al., 2007). These studies had elucidated range expansions and refugia in Western Palaearctic gall wasps similar to that observed in other European species like *C. parallelus* (Hewitt, 1999). A genome-wide study of population splitting and admixture between gall wasp refugia tests the validity of previous inferences and makes more powerful conclusions possible.

1.4.2 Estimating migration between populations using coalescent modelling

In the past decade, rigorous coalescent-based models for estimating population splitting times and admixture events between populations have been introduced that rely on multiple unlinked loci with negligible internal recombination (Hey & Nielsen, 2004; Hey, 2010; Lohse et al., 2011).

The importance of admixture in population and species histories was underlined by the discovery of segments of Neandertal (*Homo neanderthalensis)* ancestry in the genomes of non-African humans (*Homo sapiens sapiens)* (Green et al., 2010). Green et al. identified this admixture by comparing frequencies of SNPs between African and non-African human populations and the Neandertal genome. They found non-African humans were significantly enriched for shared SNPs with Neandertals than expected by chance. These regions consist of an estimated at 1-4% of non-African human genomes (Green et al., 2010). The direction of this admixture event was from the Neandertals into early-modern non-African humans.

1.4.3 The likelihood model

Lohse et al., 2011 have developed a maximum-likelihood framework to test models of divergence with gene flow between three populations using only one haploid genome from each population, in contrast to the traditional sampling approaches of phylogeography. Only one individual per refuge is sampled as the model assumes the sampling populations are panmictic, that is any one individual is completely representative of the mosaic of genealogies within that population. It assumes the sampled populations are discrete from one another as is standard in statistical phylogeography (Hickerson et al., 2010; Hey & Machado, 2003; Knowles, 2009) because such models are tractable and easy to interpret. This likelihood method is statistically optimal as it uses all available information in the data, which is key when estimating recent events, as there may be low numbers of informative mutations. This makes the method more powerful than Green et al.'s (2010) SNP only approach, but computationally difficult. A more rigorous approach to the pitcher plant mosquito discussed above is possible using this method by sampling a haploid from each of three populations of pitcher plant: in the refugium, northward along the Atlantic coast, and inland near the Great

Lakes. Then, by comparing likelihoods of different models of range expansion Emerson et al.'s (2010) inferences can be tested.

1.4.4 Gall wasps are ideal for intra-specific population studies

Gall wasps, chalcid parasitoids and most Hymenopterans are particularly well suited to population genetics as males are haploid. Single nucleotide polymorphism (SNP) calling is greatly simplified, because heterozygotes are not possible (chapter 2). Any site within an individual with more than one base present contains an error. This could be due to sequencing error or a misaligned read. Furthermore, there is no requirement to phase blocks of sequence as any SNPs within a block correspond to the same haploid chromosome.

In this thesis I have developed a pipeline to generate genome-wide alignments of across triplets of outgroup-aligned ingroup sequences and analysed it using a model-based approach. The pipeline is a standardised protocol for taking the raw data and turning it into a high quality dataset of thousands of outgroup-aligned single-copy nuclear loci. This is important because of the scale of the dataset. It is no longer possible to check by eye the quality of each final alignment as one could with small numbers of loci. Therefore each step needs to be sufficiently rigorous that the thousands of final alignments do not need to be individually checked. It would also be relatively simple to expand the pipeline to more individuals than a triplet with theoretical advances and cheaper sequencing.

The pipeline is a viable method for generating a high quality dataset of thousands of loci and megabases of sequence without needing a reference genome. It would also be relatively simple to expand the pipeline to more individuals with theoretical advances and cheaper sequencing.

Although the results contained in this thesis apply only to one trophic level the success of this initial study has led to a multi-trophic project beginning in January 2013 (chapter 5). It applies the triplet based likelihood method to multiple species of gall wasps, including a cynipid inquiline, and their parasitoids. The pipeline developed here will form the basis of the bioinformatic aspects of the project.

1.5 Selecting a model system for the study of oak gall wasp interactions and phylogeography. *Biorhiza pallida* gall wasp on *Quercus* section *Quercus* oaks

The gall wasp chosen for investigating gall induction (chapter 3) and gene flow between glacial refugia (chapter 2) is *Biorhiza pallida* (sexual generation gall, figure 1.7). It is abundant across the Western Palaearctic, easily identified and sampled, and multilocular species. A multilocular gall contains multiple developing larvae; *B. pallida* sexual generation galls may contain dozens of larvae and grow in excess of 5cm diameter. The galls of this species cannot be misidentified for other species of gall wasp. For these reasons *B. pallida* has been the focus of previous phylogeograhy (Rokas et al., 2001) and gall induction studies (Schönrogge et al., 2000; Harper et al., 2004), and as a result, morphological development of the galls is well understood. *B. pallida* galls *Quercus* section *Quercus* oaks in the Western Palaearctic. The sexual generation of *B. pallida* is well known in the United Kingdom for inducing oak apples during the spring. By contrast the asexual generation gall develops on the roots of *Quercus robur/petraea* and is seldom observed.

Western Palaearctic gall wasps show genetic structure compatible with three Pleistocene refugial areas (Iberia; Italy and the Balkans; Asia Minor and Iran) that follow those for deciduous oaks (Petit et al., 2003). Most species show patterns compatible with westwards range expansion into Europe from Asia during or before the Pleistocene (the ‘Out of Anatolia’ hypothesis, see Rokas et al., 2003; Challis et al., 2007; Stone et al., 2009), a pattern also supported by a recent meta-analysis of 19 parasitoid and 12 gall wasp species (Stone et al., 2012). The only exception to this pattern known has been *B. pallida*, for which mitochondrial and ITS nuclear sequence data show evidence of a deep east-west divide (Rokas et al., 2001). By choosing *B. pallida* for genome-wide phylogeographic inference, this anomalous pattern can be tested more rigorously and deeper insights made.

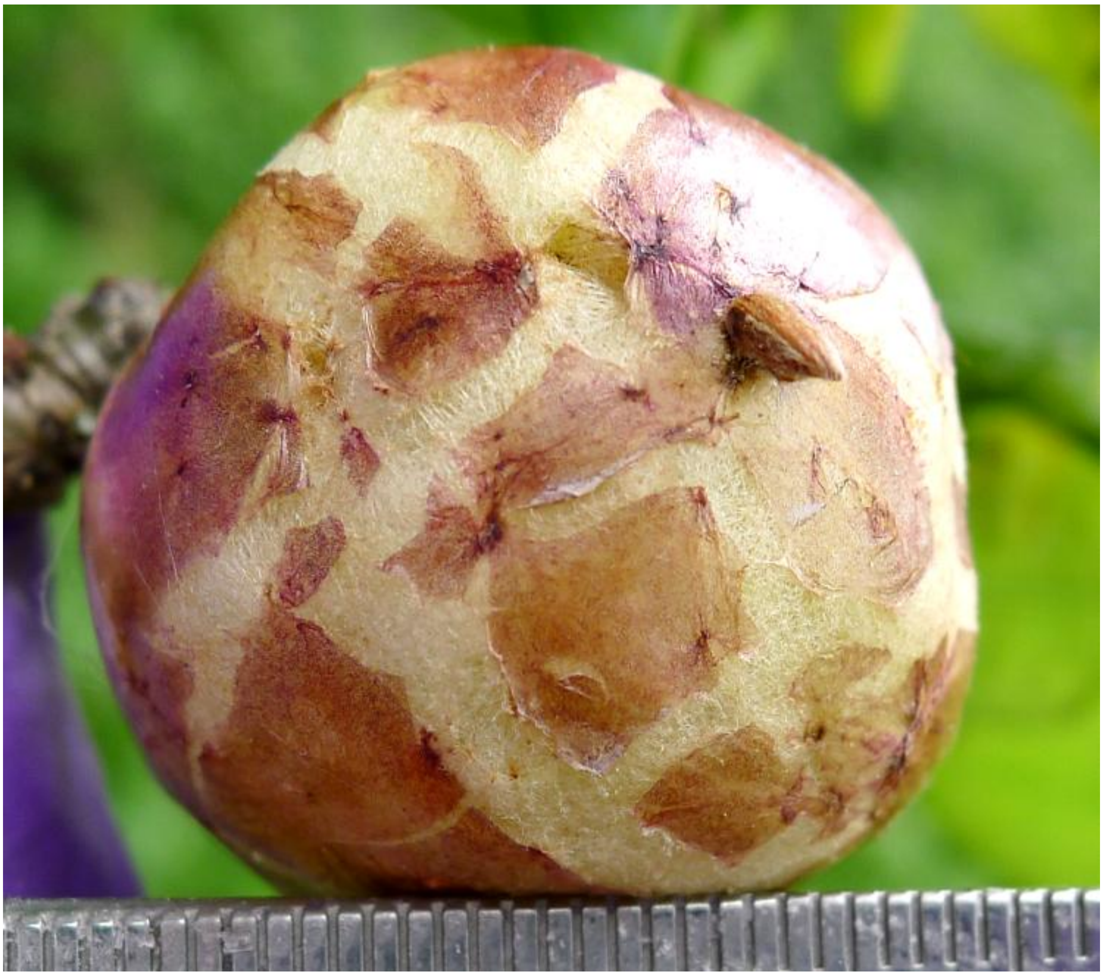

Figure 1.7. A growth stage *B. pallida* gall on *Q. robur,* photo J. Hearn.

1.5.1 *Belizinella gibbera, Diplolepis spinosa* and resources

Two other gall wasp species, *Belizinella gibbera* and *Diplolepis spinosa*, were also important to parts of this thesis. *Belizinella gibbera* was chosen as it is closely related to *B. pallida* making a suitable outgroup for genome-wide phylogeography (figure 2.2)*. B. gibbera* sequences are used to polarise SNPs within *B. pallida*. At sites with SNPs the ingroup nucleotide concordant with the outgroup nucleotide is the ancestral state. The other nucleotide is a mutation that has occurred since the population(s) that has/have it diverged from the other population(s). Unlike *B. pallida*, *B. gibbera* is unilocular and currently only asexual females have been found. Very little is known about the ecology of this species. Samples used in this thesis were collected from *Quercus dentata* (*Quercus* section *Mesobalanus*) in the Russian Far East. It is possible that an as yet unidentified sexual generation exists (Abe et al., 2007).

Diplolepis spinosa is a galler of the rose, *Rosa blanda* (Rosaceae), and forms large spiny multilocular galls on stems; it is an asexual species. It was chosen because its morphological evolution is also well understood (Shorthouse et al., 2005).

Also available for this thesis was three transcriptomes generated by the 1K Insect Transcriptome Evolution (www.1kite.org/) project. A transcriptome is the complete set of genes transcribed by the organism or tissue at the point of sampling. Two of the transcriptomes are from adult cynipids. They are the oak galler *Andricus quercuscalicis* and the sycamore galling *Pediaspis aceris*. Both are more closely related to *B. pallida* than *D. spinosa* (figure 1.3). The final transcriptome is from *Leptopilina clavipes* - a figitid parasitoid, the closest parasitoid group to the cynipids. Comparisons of transcriptome expression are less powerful than genome comparisons are as genes not present in a transcriptome may be present in the species genome but are not expressed at the sampling point.

1.6 'omics and data-sharing between different projects

This thesis aims to investigate two distinct but fascinating aspects of cynipid gall wasp biology. This is done by large scale sequencing of gall wasp genomes and transcriptomes. The research methods developed and applied here have only become possible in the past few years with the introduction of high throughput sequencing. For both chapters, the sequencing data generated primarily to answer the objectives for one chapter was applied to the other chapter (and *vice-a-versa*) and improved subsequent analyses.

1.7 Brief overview of thesis chapters

The following subsections provide an overview of the content of each chapter in this thesis.

1.7.1 Chapter 2: Genome-wide statistical phylogeography

To test models of gene flow between refugial populations of a western Palaearctic gall wasp a pipeline was developed to create single-copy nuclear-sequence alignments sampled genome-wide of a haploid *B. pallida* from each Western Palaearctic refugium plus outgroup sequence. This work was carried out in collaboration with Dr Konrad Lohse (University of Edinburgh) and others. Developing the pipeline presented many bioinformatic challenges.

The *B. pallida* transcriptome, analysed in chapter 3, was used to identify regions within the final alignments that contained expressed sequence. The proportion of expressed sequence per alignment was then used to fit heterogeneous mutational rates to the model resulting in higher likelihood scores. It was also possible to identify linked alignments using by finding those that overlapped the same transcript, a violation of the likelihood method’s assumptions.

1.7.2 Chapter 3: Identifying Candidate genes for gall induction

The transcriptomic experiment of this thesis aims to elucidate the underlying genetic control of gall induction by gall wasp larvae and the corresponding host response. The larva is essential to successful galling from initiation onward and hence is the focus of the experiment, although maternal effects at oviposition may also be important. The experimental design was of replicated *de novo* transcriptome sequencing of gall segments containing tissues of both host and galler across the three developmental stages. Draft genome assemblies of the *B. pallida* and *Q. robur* genomes were leveraged to identify the origin of reads in this mixed dataset. The quality filtered reads were aligned to both genome assemblies plus the *Q. robur/petraea* ESTs. Very simply, if a read aligned best to the B. pallida genome it was assigned to an Arthropod (i.e. gall wasp) bin. Alternatively if the best alignment was to the oak sequences then the read was assigned as of plant origin.

The expression of the mature and growth stages was used as a control to identify gall wasp and plant genes of high expression at the early stage. These genes were then annotated bioinformatically to (a) identify their function and context in genetic pathways and (b) generate more specific hypotheses for future functional studies into induction. The draft genome assemblies of *Diplolepis spinosa* and *Belizinella gibbera* were also available for querying the presence of candidate genes of interest in other gall wasp genomes.

1.7.3 Chapter 4: Horizontal gene transfer into cynipid genomes

Hypotheses of horizontal transfer events were tested using transcriptomic and genomic resources. The presence of genes of viral capsid origin (discussed in section 1.3.5) was tested across genomes. Plant genes were also candidates for horizontal genetic transfer because of the intimate relationship between host and gall wasp. Finally, plant cell-wall degrading enzymes (PCWDEs) of plant pathogens were searched for as they have been found in other phytophagous arthropods and plant-parasitic nematodes.

These tests were simple presence absence tests for such genes in the genome assemblies of *B. pallida*, *B.gibbera* and *D. spinosa*. They were also looked for in the available transcriptomes, although this was less powerful as absence could mean lack of such gene expression, not lack of these genes in the genome. A role in gall induction was considered if candidate horizontally transferred genes were differentially expressed during gall development (as established in chapter 3).

1.7.4 Chapter 5: Future proposals based on the results of this thesis

Future work is presented for genome-wide phylogeography and cynipid transcriptomics. The expansion of the methods of chapter 2 to many cynipids and their parasitoids is discussed, as this will allow testing of hypotheses of community assembly. Experimental follow up to the inferences of chapter 3 are proposed to test specific hypotheses of galler host interaction. Finally, methods of confirming horizontal gene transfer events into the cynipid genomes are discussed.

Chapter 2: Statistical phylogeography from individual *de novo* genome assemblies

2.1 Aims

The principle aim of this chapter was to develop a method for generating thousands of blocks of single-copy nuclear sequence from multiple individuals from low coverage Illumina short read data in the absence of a reference genome. The approach was tested on data from haploid individuals of the Western Palaearctic gall wasp, *Biorhiza pallida*, sampled from three European refugia. The resulting data were used to test alternative models of divergence between three refugial populations using a new maximum likelihood method (Lohse et al., 2011; Lohse et al., 2012; Lohse & Frantz 2013). The size of the dataset allowed powerful inferences of the recent, Pleistocene population history of this species, demonstrating that *de novo* genome assemblies contain detailed information about recent population parameters, such as splitting times and admixture between glacial refugia.

This chapter involved collaboration with Dr. Konrad R. Lohse (University of Edinburgh). I developed the bioinformatic pipeline and generated the dataset and K. R. Lohse developed the likelihood method. The results were analysed together.

2.2 Phylogeography, the coalescent and multiple loci

Many phylogeographic studies use data from small numbers of variable loci, such as mitochondrial DNA or microsatellites, across large numbers of individuals to infer population histories (Avise, 1987). In part, such studies are shaped by the nature of Sanger sequencing. It is relatively inexpensive to design primers that amplify loci up to 1000 bases long and then sequence that region in many samples, but not to scale this approach to whole genomes. But identifying loci to amplify is a challenge in itself and the approach becomes prohibitively expensive and labour-intensive with increasing numbers of loci (Lohse et al., 2010). There is a limit to dataset size in the tens of loci (Lohse et al., 2010). It was also common to rely on mitochondrial sequence only, or more recently in combination with one or two nuclear loci because of their variability and the relative ease of doing so (Rokas et al., 2001). However, it has long been known that many loci are preferable to one or two, but this was restricted to microsatellites and allozymes. These are difficult to generate and lack a coalescent framework for analysis; hence sequence based analyses became the norm.

For intra-specific studies the level of variability in few loci may still be insufficient for discerning recent processes. Furthermore, inferences are potentially misleading as mitochondrial genome history may differ from that of the nuclear genome, driven for example in insect species because of infection by cytoplasmic *Wolbachia* (Rokas et al., 2001). Secondly when only one locus is used, mitochondrial or nuclear, spurious population histories may be inferred because of the stochastic nature of the coalescent (Rosenberg & Norborg, 2002).

To control for coalescent variation and accurately infer population histories within a species multiple unlinked loci must be analysed in a model based framework (Nichols, 2002). This is because unlinked genes within a genome that have experienced the same demographic events are independent replicates of the coalescent process (Rosenberg & Norborg, 2002).

2.2.1 Estimating admixture between populations using coalescent modelling

In the past decade more rigorous coalescent-based models for estimating splitting times and continuous or discrete admixture (gene flow) between populations have been introduced that require multilocus data and commonly assume negligible internal recombination (Hey & Nielsen 2004, Hey 2010, Lohse et al., 2011).

The importance of admixture in population and species histories was underlined by the discovery of segments of genome with closer homology to Neandertals (*Homo neanderthalensis)* than to putative ancestral humans (*Homo sapiens*) in the genomes of modern non-African humans (Green et al., 2010). They introduced the *D-statistic* that tests for admixture by estimating enrichment for patterns of SNPs explainable by admixture. The proposed admixed regions comprise an estimated at 3-7% of non-African human genomes (Green et al., 2010). The models tested incorporated only unidirectional admixture from Neandertal to modern human. The admixture relationship is described in figure 2.1, scenario E; the direction of admixture is from the ancestral population (Neandertal) into the youngest population (non-African humans).

2.2.2 Genome-wide phylogeography in non-model organisms

Genome-wide shotgun sequencing in combination with coalescent modelling has the potential to revolutionise phylogeography. It is now possible to sequence low-coverage draft genomes at (constantly decreasing) affordable cost. However, sampling whole genomes from large numbers of individuals, as in traditional phylogeographic sampling designs, is still prohibitively expensive. Nevertheless, with further advances in sequencing and coalescent modelling this is probably the direction the field of phylogeography is heading.

The ability to analyse many homologous sequence blocks from genome sequence, as we have here, is in itself as previous genome-wide studies involved a “genomic reduction” step prior to sequencing (McCormack et al., 2013, Arnold et al., 2013). An example and the principal alternative to whole genome shotgun sequencing considered is Restriction-site Associated DNA (RAD) sequencing (Davey and Blaxter, 2010). For RAD sequencing short regions (100s of bases) around the chosen restriction site are amplified and sequenced. RAD sequencing results in allele frequency data for thousands of loci across the genome. A further advantage of this approach is that data for large numbers of individuals can be analysed if required. RAD sequencing has been used to elucidate population structure in the pitcher plant mosquito using a phylogenetic approach (Emerson et al., 2012). The disadvantages of RAD over the whole genome shotgun sequencing applied here are two-fold. Firstly, generating RAD libraries is complex and time-consuming. Secondly, the data produced are less appropriate for inferring intra-specific population histories. This is because the short sequences that are currently generated for RAD studies result in a dataset of thousands of unlinked single nucleotide polymorphisms (SNPs). More detailed information about population histories can be gained from longer blocks of sequence with multiple, linked polymorphic sites (Lohse et al., 2011). This is because linked sites allow you to generate population tree topologies (genealogies) for each sequence, including information about branch lengths. Across genealogies this distribution of branch lengths is highly informative of population history (Lohse et al., 2010), a source of information not available to RAD sequencing approaches. However, this may change if read lengths of Illumina technology continue to increase. This is because long sequence blocks linked to RAD sites will become possible, allowing genealogy based analyses as well as the currently possible SNP frequency analyses.

2.2.3 Triplet sampling and the likelihood model

An alternative strategy to "genomic reduction" is to work with whole genomes, but with the analysis restricted to a few individuals. Lohse et al. (2011) have developed and extended (Lohse et al., 2012) a maximum-likelihood framework to test models of divergence with gene flow between three populations using only one haploid genome from each population. The restriction to three individuals is not because it is superior to using multiple individuals but reflects the difficulty in expanding the model to more individuals or populations (Lohse et al., 2011). This minimal triplet sampling is uninformative about current ongoing processes within populations such as changes in effective population size (N_e), as a single haploid genome lacks the resolution needed for such parameter estimates. However, this sampling does contain much information on the historical interactions between distinct populations (Lohse et al., 2012). The Lohse method (2011) models the relationship between populations as a series of instantaneous divergence and admixture events and fits models numerically by maximizing the likelihood of parameters. This is an important advance, as most inference methods for fitting alternative models of population history do not scale up to genomic datasets or analyses take a prohibitive amount of time to complete (but see Francois et al., 2008). It assumes the sampled populations are discrete (physically separated and distinct) from one another as is standard in statistical phylogeography (Hickerson et al., 2010; Hey & Machado; 2003; Knowles, 2009) because such models are tractable and easy to interpret (Harris & Nielsen, 2013; Li & Durbin, 2011, Green et al., 2010; Lohse & Frantz 2013).

This likelihood method is statistically optimal as it uses all available information and is based on blocks of sequences, thus for every block of sequence a genealogy can be generated. This is more powerful than a RAD or *D-statistic* based analysis, because there is less information content in unlinked SNPs than for sequence blocks containing linked sites, like Lohse's (2011) method. It is therefore superior to single nucleotide polymorphism (SNP-only) based analyses. For example, it considers the distribution of polymorphisms across loci; meaning loci without any SNPs are still informative. Additionally, singleton mutations provide information on the length distribution of external branches of genealogies, whereas the D-statistic of Green et al., (2010) only measures the relative frequency of two types of shared-derived sites (which occur on internal branches of a genealogy). The extra information, and associated more powerful inferences over SNP based methods makes the informatic challenge of generating an appropriate sequence-based (rather than SNP-based) dataset worthwhile.

Other recently developed methods, like the RAD-based allele frequency spectrum approach discussed above, also consider genome-wide datasets as the field adapts to the possibilities of high-throughput sequencing. Li and Durbin (2011) have developed a hidden Markov approach for inferring past changes in effective population size from just a single diploid genome. Similarly, Harris and Nielsen (2013) use the length distribution of allelically identical (or identical by state, IBS) tracts of sequence in pairwise alignments to fit complex histories of divergence and admixture between two populations. However, both these methods are currently restricted to histories involving just one or two populations. They also rely on long sequence tracts and hence excellent genome assemblies unavailable outside of a handful of (largely model) organisms for eukaryotes.

The Lohse (2011) approach is far more applicable to non model organisms, as it requires many sequence blocks long enough to contain multiple polymorphic sites but short enough to justifiably ignore within block recombination (here block sizes of 500-2000 bases were used). This is complementary to the highly fragmented *de novo* assemblies that can be achieved with low coverage short-insert (≤ 300 base pair) paired-end Illumina data.

2.2.4 Oak gall wasp phylogeography and *Biorhiza pallida*

A suite of detailed studies have addressed phylogeographic patterns in Western Palaearctic oak gall wasp communities, both for the gall inducers (Stone & Sunnucks, 1993; Rokas et al., 2001, 2003; Stone et al., 2007; Challis et al., 2007; Stone et al., 2012) and their parasitoid enemies (Hayward & Stone, 2006; Lohse et al., 2010, 2012; Nicholls et al., 2010a, 2010b; Stone et al., 2012). European gall wasps are inferred to have migrated from Asia into Europe in the Pliocene or Pleistocene epochs (the 'Out of Anatolia' hypothesis, Rokas et al., 2003; Challis et al., 2007; Stone et al., 2009; see also Connord et al., 2012), 1.3–4.2 million years ago, reaching Iberia approximately 400 000 years ago (2.5%–97.5% quantiles of 0.1-0.7 million years ago) (Stone et al., 2012). Gall wasps have continued to enter Iberia since the initial immigration event (Stone et al., 2012). Both gall wasps and their parasitoids show genetic structure compatible with three major Pleistocene refugial areas (Iberia; Italy and the Balkans; Asia Minor and Iran) that broadly parallel those for deciduous oaks (Petit et al., 2003). The only exception to this pattern to date has been *B. pallida*, for which mitochondrial and internal transcribed spacer (ITS) nuclear sequence data show evidence of a deep east-west divide (Rokas et al., 2001). This raises the question of how general the 'Out of Anatolia' pattern is for all three trophic elements of this community (Stone et al., 2009, 2012). Here we use *B. pallida* as a case study for phylogenomic inference, and ask whether genome-level data support the apparently anomalous pattern for this species within the oak gall wasp community.

2.2.5 Gall wasps are well suited for intra-specific population studies

Gall wasps, chalcid parasitoids and most Hymenoptera are particularly well suited to for sequence-based analysis of genetic diversity as males are haploid. Because heterozygotes are not possible, single nucleotide polymorphism (SNP) calling is greatly simplified. Any site within an individual with more than one base present must contain an error. This could be due to sequencing error or a misaligned read. Furthermore, there is no requirement to phase (correctly identify alleles derived from each haploid in diploid sequence) blocks of assembled sequence as any SNPs within a block correspond to the same haploid chromosome.

Belizinella gibbera, a species closely related to *B. pallida,* was chosen as the outgroup species to polarise SNPs within *B. pallida*. At each polymorphic site mutations that are concordant between the in- and outgroup are the ancestral state. The other nucleotide is a mutation that has occurred since the population(s) that has/have it diverged from the other population(s). When two populations share the derived site the site is a parsimony informative site. The site could result from a mutation before two populations split but after splitting from the ancestral population. Alternatively it could result from a mutation in one of the populations after the populations split with subsequent admixture into the other population. A final possibility is a duplicate independent mutation (a back mutation) a violation of the infinite sites model assumed by Lohse et al.'s (2011) method. Back mutations are a danger when outgroup sequences are very divergent from ingroup sequences and individual sites reach mutational saturation (when multiple mutations at a site obscure the relationship between sequences).

2.2.6 Modelling divergence and admixture of refugial populations of *B. pallida*

Jesus et al. (2006) proposed a model of demographic changes that occur to a species undergoing alternating glacial and interglacial periods. During interglacials, like the present, the species is panmictic, but during glacial periods populations of the species fragment into small sub-populations corresponding to refugia. As this is a cyclical process, admixture between populations happens in a discrete fashion only during interglacials; thus, we modelled admixture between refugia as instantaneous and unidirectional. Population sizes of each refuge considered (Iberia - *West,* Hungary/Croatia - *Central* and Iran - *East*) were assumed to be equal. Under an infinite sites model in which any new mutations must occurs at different sites (therefore only two possible nucleotides at any one site), there are six possible branches on a triplet genealogy (figure 2.1) along which mutations can occur ($k = \{k_w, k_e, k_c, k_{we}, k_{wc}, k_{ec}\}$), where k_w is the number of singletons occurring in the western population (mutations found only in the western refugia, i.e. have occurred along a terminal branch of the genealogy, figure 2.1). A mutation on branch k_{wc} represents a shared-derived sited between the western and eastern populations; at such positions the topology is $\{E,\{W,C\}\}$. A vector of these mutational types for each alignment forms the input for the likelihood model. Different histories of population and admixture predict contrasting values in such a matrix.

For a given order of population divergence, there are six possible models (figure 2.1), each with five parameters: the time of the older split T_2; the time of the more recent split (T_1); the time of admixture, or gene flow, (T_{gf}) (all measured back in time from the present); the admixture proportion ($_f$) and the effective population size (N_e). For computational tractability, a single and constant Ne for both ancestral populations was assumed. Support for all six admixture scenarios was assessed as well as for simpler, nested models that assume no admixture and divergence between either three or two populations for each of the three possible orderings of population divergence (a total of 24 divergence and admixture models). We also quantified the support for a basal polytomy, a single panmictic population, and for distinct refugial N_e in the strict divergence models (to test whether the additional parameter substantially improved model fit without the need to invoke admixture), giving 32 models in total.

By calculating the likelihood scores of different models of population history the divergence and direction of admixture events between the three refugia resulting from past range expansions can be identified.

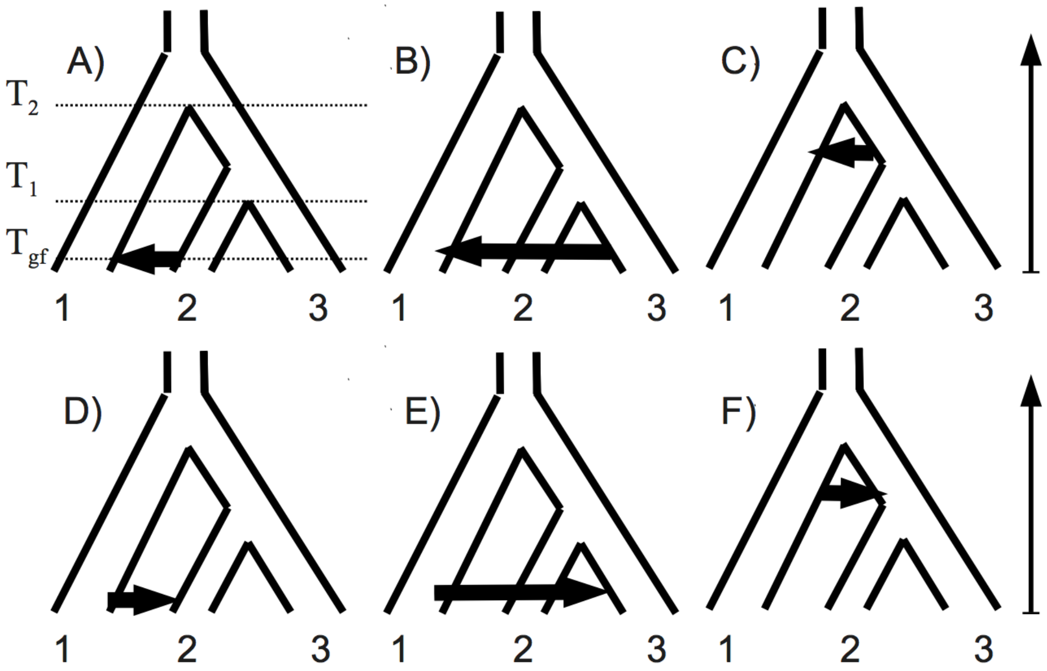

Figure 2.1. The six models of gene flow considered. T_{gf} is time of gene flow, indicated by the horizontal arrow; T_1 and T_2 are population-splitting times (figure courtesy K. R. Lohse).

2.2.7 Testing the assumption of discrete populations between refugia

The likelihood model assumes each population sampled is panmictic, and therefore that a single haploid genome can be taken as representative of the population as a whole. In other words, we assume that the haploid genome sampled in Iberia can be considered representative of the mosaic of genealogies present in Iberia. However, local genetic structure emerges as a consequence of the limited dispersal ability of individuals (Askew, 1984), and could occur within each refugial region in our analysis. The subject of this study, *B. pallida* has been observed completing multiple generations on a single oak, and there is evidence for very local, at the single tree level, adaptation in cynipids related to *B. pallida* (Egan and Ott, 2007). So any model that approximates a population occupying a large area (such as Iberia) as panmictic may well break down over recent time-scales. Therefore, a replicate haploid genome from each of the Central and Western refugia (sampled individuals: table 2.2 and sampling sites: figure 2.2) was sequenced to test the assumption of within refuge panmixis. In each refuge the replicates were both sampled approximately 400 km away from the original population sample, a distance probably well above the dispersal ability of an individual wasp (Askew, 1984). The assumption that a single haploid genome can be considered representative of a large refugium holds if the same population splitting, admixture, and parameter estimates are inferred with either refugial haploid genome. The maximum likelihood analyses were performed on all four possible combinations of West and Central individuals. Although desirable, there was no duplication of the Eastern refuge because no suitable haploid male samples were available from Iran or Asia Minor.

2.2.8 Sample selection for Genome-wide phylogeography

In total, five haploid males were selected from the sexual-generation of *B. pallida* for sequencing, two each from the Iberian and Balkan refugia and one from Iran in the east (table 2.1, figure 2.2).

As polymorphic sites within ingroup sequences had to be sorted into derived and ancestral (polarised), alignment to an outgroup was necessary. Choosing a good outgroup with optimum divergence form the ingroup was essential for a robust analysis. Firstly, Ingroup individuals should not be more divergent from each other than to outgroup sequences, this can occur because of lineage sorting (K. R. Lohse, personal communication). But they must also be close enough to avoid mutational saturation under a simple mutational model of infinite sites. Under the infinite sites assumption each new mutation occurs at a new site within genome, therefore back mutations are a violation of this assumption. Mutational saturation from sequence divergence between in- and outgroup causes back mutations that can result in errors in polarising ingroup sequences. *Belizinella gibbera* was chosen as the outgroup based on a *cytochrome B* (*cytB*), a mitochondrial gene, global phylogeny of oak gall wasps (figure 2.3, complete tree: appendix fig 2.15)

(James Nicholls, personal communication). Two *B. gibbera* females from the Russian Far East were sequenced; as the species is asexual haploid males are not available. The two sequenced females were reared from galls on the same tree.

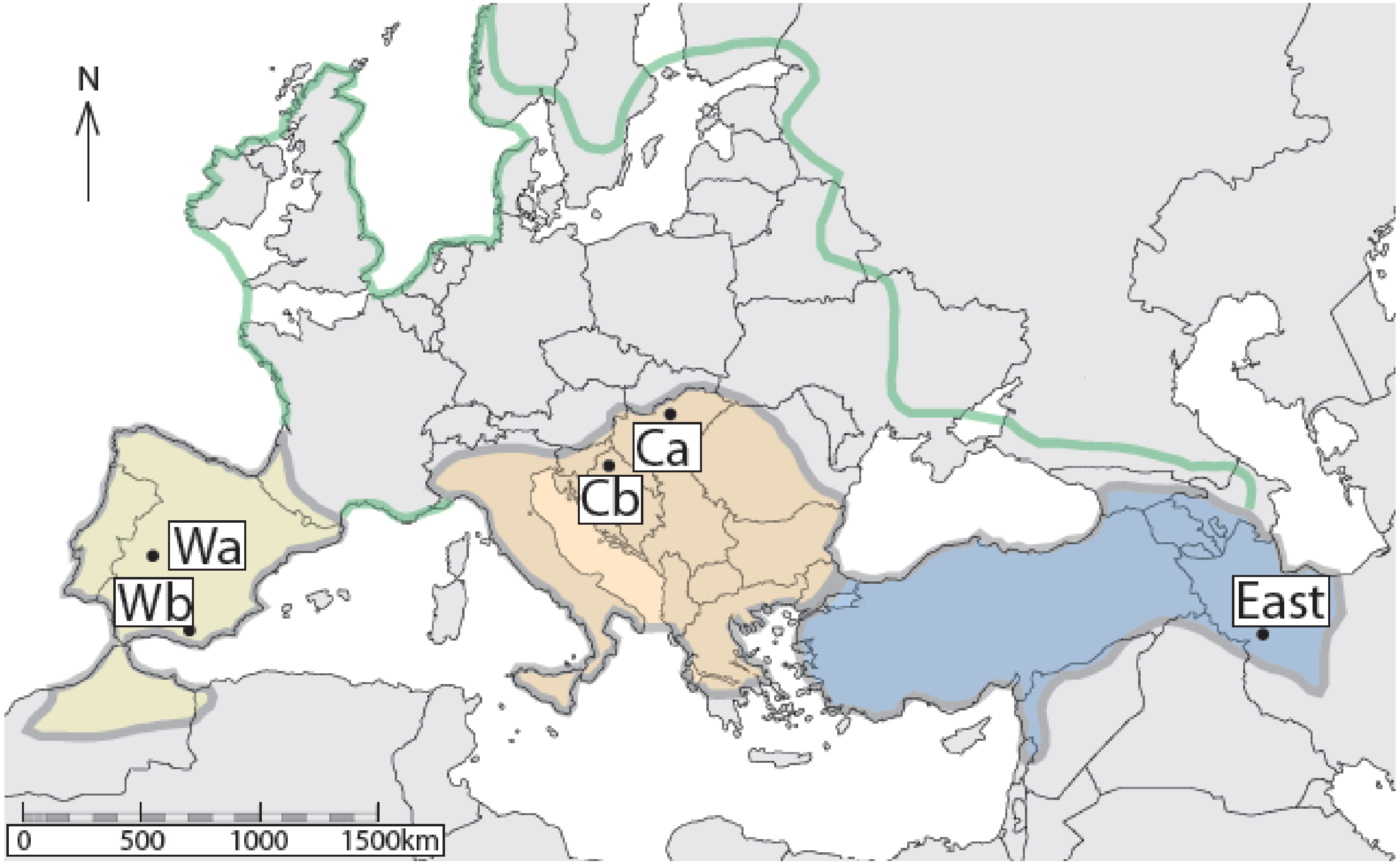

Figure 2.2. Sampling locations of the five *B. pallida* individuals (*West A, West B, Centre A, Centre B* and *East*, used for genome sequencing and population genomic analyses. Each refugium is coloured; W: green, C: orange, E: blue. The green line shows the extent of the distribution of the oak host plant. Figure courtesy of Graham Stone.

Specimen ID	Designated ID	Collection site	Region	Country	Latitude and longitude	Date	Host	Extractor
Bgib15	Outgroup	Khazan lake	Primorsky Krai	Russia	42.45 N, 130.65 E	26/09/08	*Quercus dentata*	Konrad Lohse
Bgib18	Outgroup	Khazan lake	Primorsky Krai	Russia	42.45 N, 130.65 E	26/09/08	*Q. dentata*	Jack Hearn
Bpal1398	West A	Mairena	Granada	Spain	37.37 N, 5.75 W	06/05/09	*Quercus faginea*	Jack Hearn
Bpal2	West B	Embalse de Garcia de sola	Extramadura	Spain	39.17 N, 5.22 W	12/04/05	*Q. faginea*	Konrad Lohse
Bpal1	Centre A	Szokolya	Danube-Ipoly National Park	Hungary	47.87 N, 19.02 E	15/05/98	*Quercus petraea/robur*	Konrad Lohse
Bpal1613	Centre B	Ze Medvedgrad	Zagreb	Croatia	45.86 N, 15.94 E	16/05/11	*Q. petraea*	Jack Hearn
Bpal1560	East	Bane or Merivan*	Kordestan	Iran	35.99 N, 45.90 E	01/04/11	*Q. robur*	Jack Hearn
BpalUK	BpalUK	Tarrant Valley	Dorset	UK	51.41 N, 0.64 W	20/08/09	*Q. robur*	James Nicholls

Table 2.1. Collection locations, dates, host species and extraction details for genome-wide phylogeography samples; *galls from both locations were received in Edinburgh from Iran in the same batch and not differentiated.

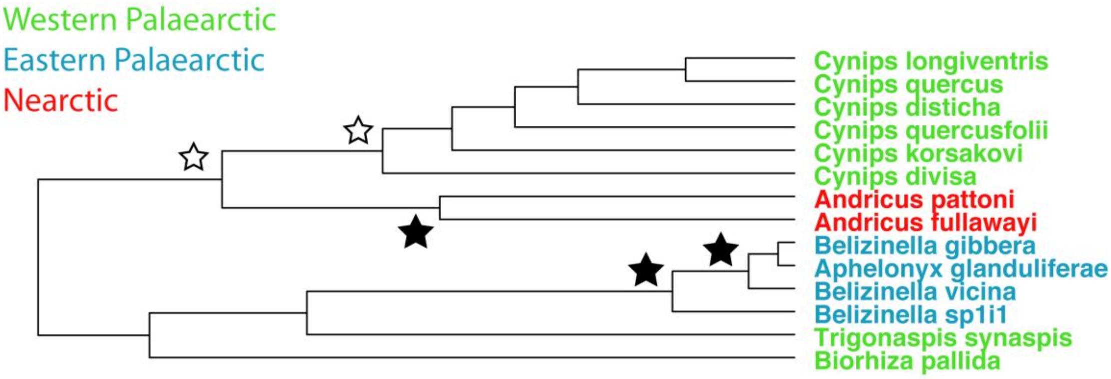

Figure 2.3. Phylogenetic relationship of *Biorhiza pallida* to *Belizinella gibbera*, courtesy of James Nicholls. Black stars represent posterior probabilities of ≥0.9, hollow stars ≥0.7.The time since the most recent common ancestor of the two species is estimated at >40 million years ago.

2.2.9 *B. pallida* and *B. gibbera* genomic DNA sequencing

Each of the seven in- and outgroup individuals were extracted using the DNEasy kit (Qiagen). Extractions with the best 260/280 ratio measured by NanoDrop spectrophotometer (Thermo Scientific) and highest mass of DNA by Qubit fluorimeter (Invitrogen) were selected from each refugium. The 260/280 ratio is a measure of extraction purity, and DNA is considered pure at a ratio of 1.8 and RNA at 2.0. Lower ratios than this can be due to contamination with protein or an extraction reagent, which may interfere with downstream processes. The DNA concentration for each sample was determined using a Qubit fluorimeter (Invitrogen) as the intercalating dye approach is more accurate than the impurity sensitive NanoDrop (Thermo Scientific).

2.2.10 Illumina Adapter and quality filtering

Raw and filtered read numbers for each individual are given in table 2.1. New genomic data was generated several times from 2009 through 2012 but filtering was standardised throughout to create a homogenous filtered dataset.

Filtering methods were adapted from Sujai Kumar's protocol (https://github.com/sujaikumar/assemblage). The raw data was first assessed using *Fastqc* (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). All reads were 3' quality trimmed using *Sickle* (https://github.com/najoshi/sickle) to a minimum quality of 20, equivalent to 99% accuracy or an error rate of one in one hundred. This Q20 filtering is commonly used and represents a trade-off between removing errors and not over-filtering the data, which can degrade assembly quality. Reads containing bases called as 'N's were removed entirely. These reads are frequently of overall low quality (S. Kumar and G. Koutsovoulos, personal communication). Adapters that had escaped basic filtering by the GenePool were removed using *Scythe* (https://github.com/vsbuffalo/scythe) and standard Illumina Paired End Adapters 1 and 2 (figure 2.4).

```
>Illumina  Paired End Adapter 1
ACACTCTTTCCCTACACGACGCTCTTCCGATCT
>Illumina  Paired End Adapter 2
CTCGGCATTCCTGCTGAACCGCTCTTCCGATCT
```

Figure 2.4. Standard Illumina paired-end adapter used to filter data.

A minimum length of 20 or 50 bases post-filtering was required for a read to be retained for 50 or 100 base-long raw reads respectively. For read pairs where one read fails quality control the other read was retained in a singles file (referred to as ‘QC singles’ in this thesis). The *Sickle* and *Scythe* commands were run as a single command as described at https://github.com/sujaikumar/assemblage. If *Fastqc* had flagged certain sequences as overrepresented in the data these were also removed using *Scythe* if they were verified as adapter sequence. Other overrepresented sequences were not removed as these can represent common sequences, or in the case of RNAseq sequences are derived from highly expressed transcripts. *Fastqc* was re-run on the filtered data to check quality control.

Sample Name	Study Name	Pairs/Singles Reads (millions)	Bases (Gb)	Filtered Pairs (millions)	Filtered Singles (millions)	Filtered Bases (Gb)
Bpal 1	CentreA	83.6	10.42	77.1	5.1	8.50
Bpal 2	WestB	41.3	6.26	35.0	1.1	5.41
BpalUK	UK	25.0	2.50	14.9	0	1.50
Bpal 1398	WestA	58.0	11.60	54.2	3.6	10.69
Bpal 1560	East	43.2	8.64	41.2	1.6	7.79
Bpal 1613	CentreB	41.0	8.20	39.3	1.7	8.24
Bgib 18	Outgroup	36.4	7.28	34.4	1.9	6.76
Bgib 15	Outgroup	93.6	11.16	58.3	33.1	9.89

Table 2.2. Combined Illumina read statistics for raw and filtered data for each individual sequenced for the genome-wide phylogeography study.

2.3 Methods and Results: creating a dataset of thousands of loci for phylogeographic inference

A flow diagram of the analysis steps in this chapter is shown in figure 2.5.

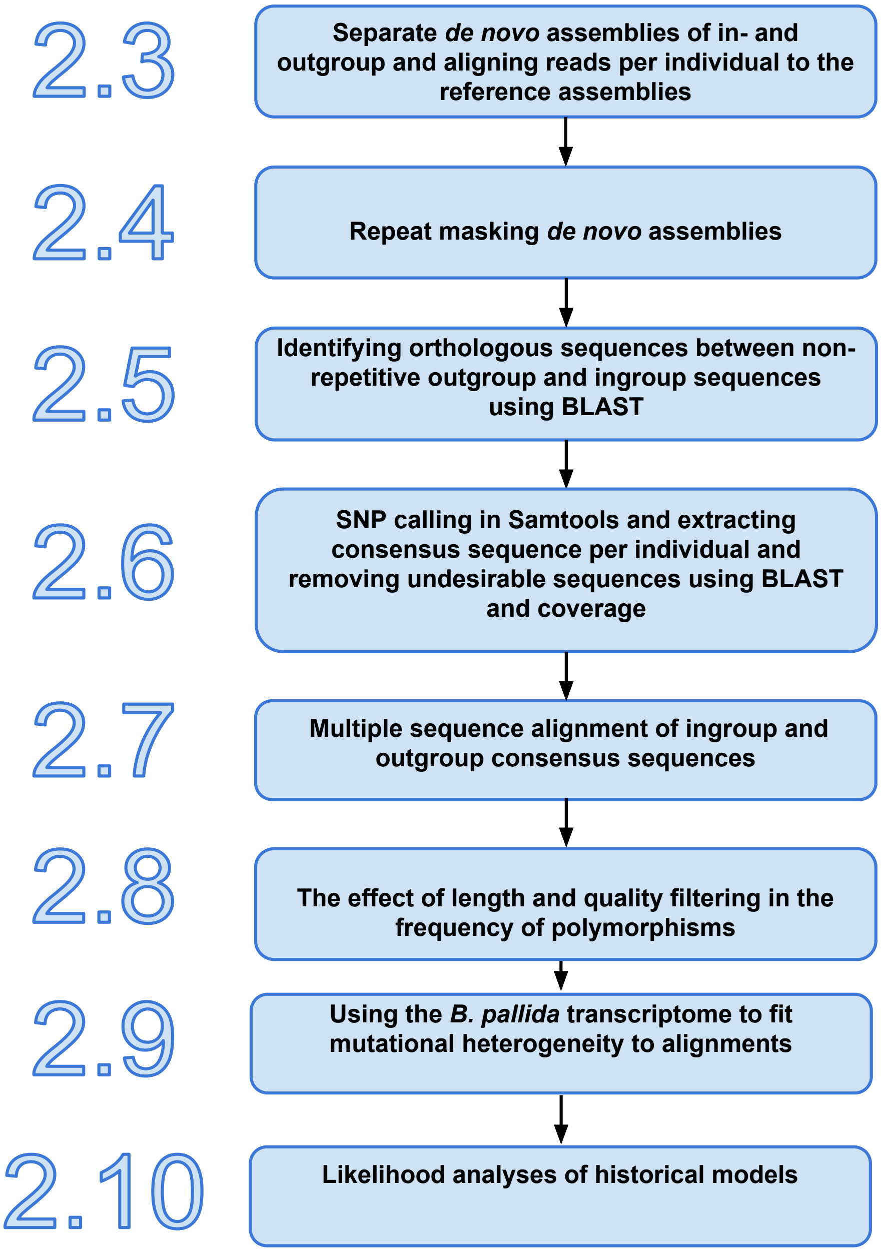
2.3
Separate *de novo* assemblies of in- and outgroup and aligning reads per individual to the reference assemblies
2.4
Repeat masking *de novo* assemblies
2.5
Identifying orthologous sequences between non-repetitive outgroup and ingroup sequences using BLAST
2.6
SNP calling in Samtools and extracting consensus sequence per individual and removing undesirable sequences using BLAST and coverage
2.7
Multiple sequence alignment of ingroup and outgroup consensus sequences
2.8
The effect of length and quality filtering in the frequency of polymorphisms
2.9
Using the *B. pallida* transcriptome to fit mutational heterogeneity to alignments
2.10
Likelihood analyses of historical models

Figure 2.5. Flow diagram of all analysis steps in this chapter. Numbering corresponds to chapter section headers.

2.3.1 De novo assemblies of *B. pallida* and *B. gibbera*

The *CLC bio de novo* assembler (http://www.clcbio.com/products/clc-assembly-cell/) was chosen for making the assemblies; *CLC bio de novo* is a de Bruijn graph assembler. *CLC bio* is proprietary software and its implementation of the de Bruijn algorithm is unpublished. *CLC bio* was chosen for its memory efficiency over other de Bruijn graph assemblers. This was important, as the genome size of oak gall wasps appears to be large for insects at 1.75Gb (± 0.286, n = 4) (Lima, 2012). The random access memory (RAM) needed to hold the graph during assembly for less memory efficient assemblers, like *Velvet* (Zerbino et al., 2008) was more than that available (512Gb RAM).

A de Bruijn graph is a directed graph of overlapping sequences of symbols (de Bruijn, 1946). By following edges through a graph complete sequences can be reconstructed. The de Bruijn graph approach first decomposes short-reads in to k-mers that become nodes on the de Bruijn graph (figure 2.6) (Schatz et al., 2011). A k-mer is a word of *k* nucleotides in length (Zerbino et al., 2008) and is a standard parameter of assemblers that can be modified to identify an optimum value. For example, *Velvet* maps multiple overlapping k-mers onto a node, and the reverse complements of the k-mers to create a bi-directed graph (Zerbino et al., 2008). A directed edge between nodes is representative of k-mers occurring consecutively in one or more reads (figure 2.6) (Schatz et al., 2011). When non-branching paths through the graph occur, like the blue k-mers of figure 2.6, unambiguous sequences of nucleotides can be strung together into contigs (Schatz et al., 2011).

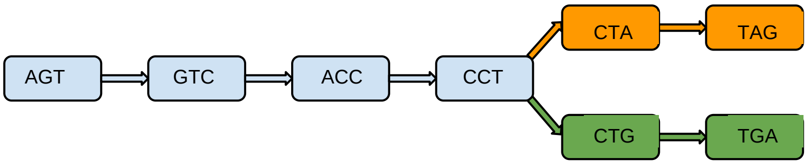

Figure 2.6. De Bruijn graph schematic of nodes and directed edges for k-mers of 3bp length. The blue boxes represent unambiguous route through the nodes. One can see that from one node to the next, one base is added and the last base is lost. The path then branches into two possible correct paths (orange and green nodes and edges). A bubble is formed if the orange and green bubbles are re-connected to form one path again.

Ideally, a de Bruijn graph resolves into a unique path through a segment of genome, but technical and biological issues can preclude this outcome. There are three common types of issues that occur with de Bruijn graph assembly. Two of these can be detected from graph topology (figure 2.6). Firstly, 'tips' are a chain of nodes disconnected from the rest of the graph at due to sequencing errors that interrupt further k-mer addition to graph. Secondly, 'bulges' or 'bubbles' occur in graphs because of discrepancies within reads. Alternatively, because the assembly attempts to collapse repeated sequences into one sequence, unique segments within repeats will cause bubbles. Algorithms that detect tips and bubbles can be implemented during assembly to clean up the graph (Zerbino et al., 2008). The third issue, incorrect connections between nodes, cannot be detected from graph topology. Instead, abrupt changes in coverage can be used to break up incorrect connections (Zerbino et al., 2008).

2.3.2 Assembly using *CLC bio de novo* assembler

The filtered reads (table 2.2) across *B. pallida* individuals were combined and assembled together using the *CLC bio de novo* assembler (version 4.0.6). This was to maximize the coverage of reads across the genome and produce the best possible assembly. Also a bias towards reads from the individual used to build the assembly is avoided in the meta-assembly approach. Data from the two outgroup individuals sequenced were combined to create a single *B. gibbera* assembly in the same way (table 2.3).

Species	Assembly N50	Number of contigs	Total bases	Average GC	Number of Ns
Biorhiza pallida	1 075	1 163 314	805 102 378	32.9	4 203 182
Belizinella gibbera	643	817 710	443 963 639	36.1	2 525 790

Table 2.3. Assembly statistics for the in- and outgroup species, *B. pallida* and *B. gibbera* respectively. Paired-end information is used to bridge unsequenced gaps in the assembly, CLC bio *de novo* places 'N's in these gaps of known length; hence number of Ns in the table.

2.3.3 Aligning reads per individual to the reference assemblies

The reads for each *B. pallida* and *B. gibbera* individual were aligned back to the respective assemblies. The *Stampy* aligner was chosen for its high sensitivity in predicting insertions or deletions (Indels), and aligning divergent reads (Lunter & Goodson 2011; Nielsen et al., 2011); it outperforms the popular *BWA* aligner (Li & Durbin, 2009). Ultimately, Indels were not incorporated into the dataset because of concerns over the accuracy of their prediction by *Stampy* (and alternatives). Neither, were SNPs within ten bases of Indels as they may have resulted from incorrect mappings.

A density plot of average contig coverage across all individuals is given for *B. pallida* and *B. gibbera* assemblies in figures 2.7-8. The dashed red lines show average coverage across all contigs demonstrating the inappropriateness of this metric for low-coverage draft genome assemblies. The few contigs of very high coverage cause a rightward skew in the coverage distribution and have a disproportionate effect on the mean. The mode is a superior descriptor of the peak of these distributions; it is less than 10 fold for both species. The numbers of reads that map for each individual are given in table 2.4. Percentages of reads mapping and pairs of reads mapping are high across all individuals. However, the percentage of properly matched pairs is low, ranging from 38-60%. This is because only pairs mapping to the same contig are reported as properly paired. There are many pairs for which one read maps to a different contig to the other read. This is expected, as genome-sequencing coverage was low, resulting in highly fragmented assemblies. The assembler did not have enough information from other read pairs to bridge the gap between effected contigs. The percentage of total reads mapping is not 100%, as PCR duplicates (separate [pairs of] reads derived from the same initial DNA molecule) were only counted once.

Individual	Total Reads mapped	% TRM	Both pairs mapping	% BPM	Properly paired mappings	% PPM
West A	110 032 493	98.15	105 275 278	97.03	41 735 608	38.47
West B	62 460 228	97.61	59 916 548	96.1	37 513 788	60.15
Centre A	151 336 967	97.4	126 679 062	95.9	77 652 270	58.81
Centre B	78 693 352	98.02	75 965 852	96.67	40 744 660	51.87
East	82 868 384	97.91	80 238 688	96.7	43 559 560	52.47
Outgroup A	68 904 531	97.43	66 085 246	96.09	31 720 514	46.12
Outgroup B	140 837 165	94.08	106 730 492	91.55	61 067 174	52.38

Table 2.4. Reads mapping to the reference assemblies for ingroup and outgroup individuals. %TRM = percentage total reads mapping; %BPM = percentage both pairs mapping; %PPM = percentage properly paired mappings.

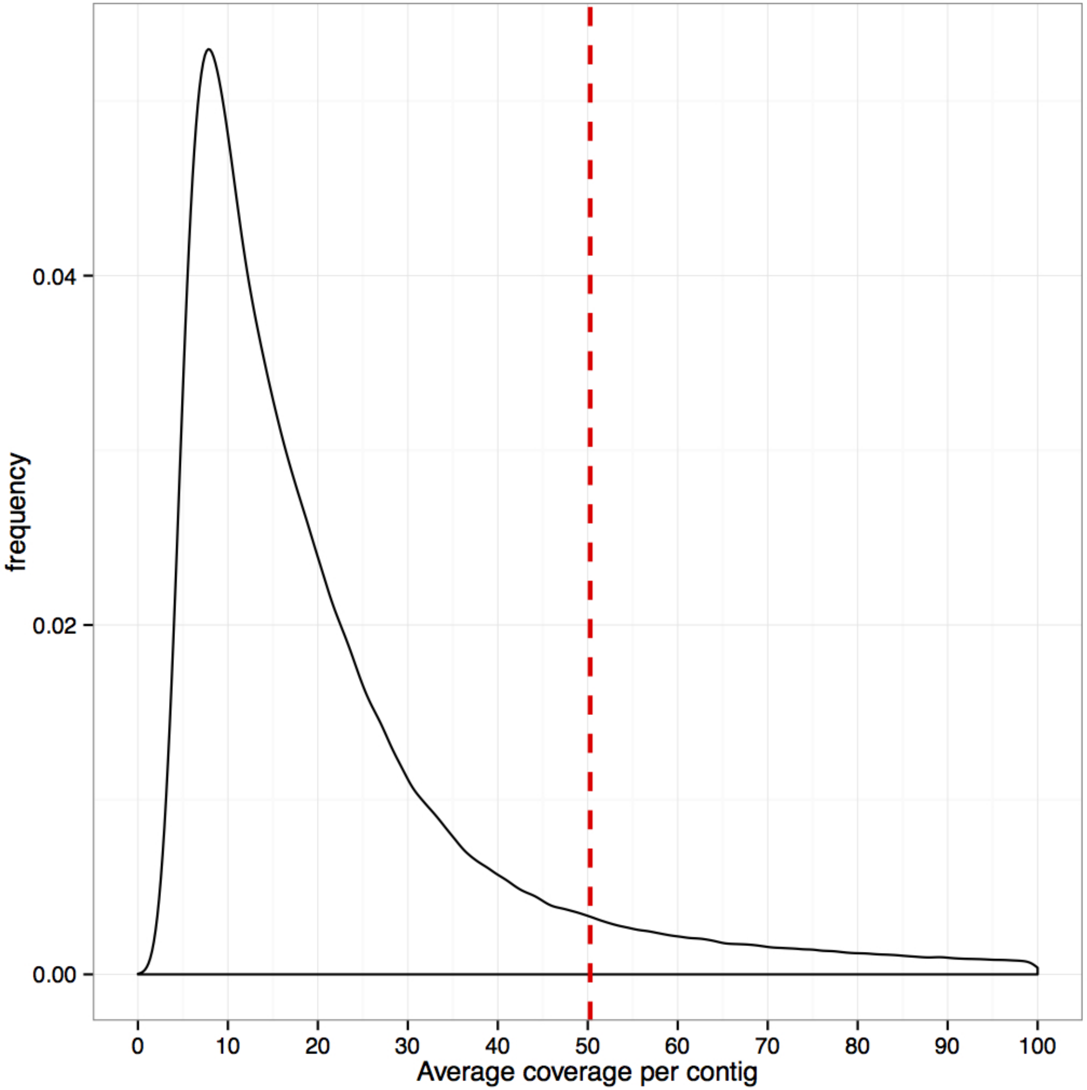

Figure 2.7. Average coverage density plot of the *B. pallida* assembly for all reads. Maximum coverage shown on this graph is 100-fold, however there are contigs with much greater coverage (>1000-fold) at low frequencies; hence the average coverage (red dashed line) is at 50-fold coverage.

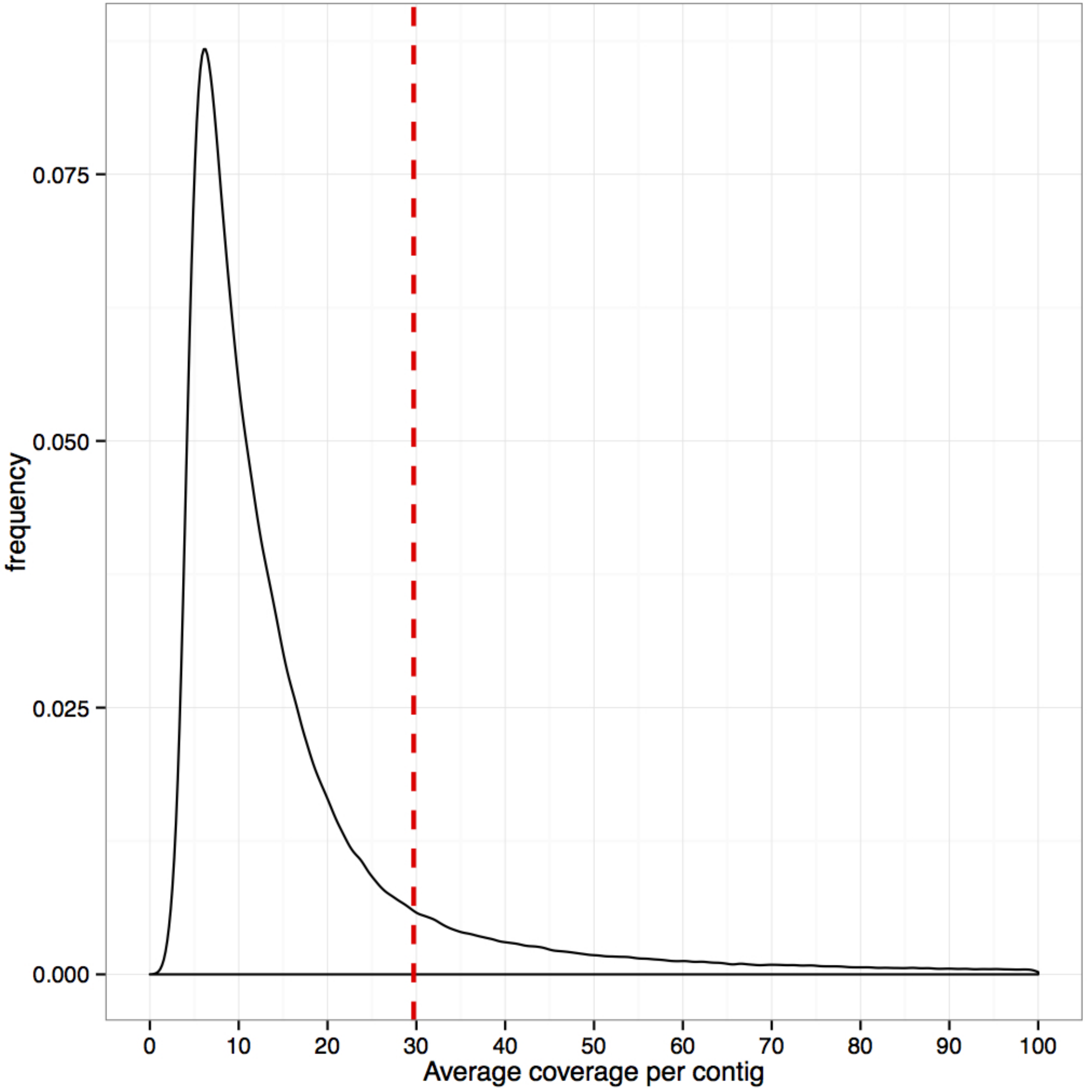

Figure 2.8. Average coverage density plot of the *B. gibbera* assembly for all reads. Maximum coverage shown on this graph is 100-fold, however there are contigs with much greater coverage (>1000-fold) at low frequencies; hence the average coverage (red dashed line) is at 30-fold coverage.

2.4.1 Repeat masking *de novo* assemblies

A repetitive sequence refers to DNA sequences that occur multiple times within a genome (Jurka et al., 2007). A large proportion of eukaryote genome can consist of various repetitive elements. For example, the human genome consists of approximately 50% repetitive elements (Treangen & Salzberg, 2012) that range in size from the tens of bases for short tandem repeats to tens of kilobases for large transposable elements.

It was important to avoid including repetitive sequences in the final dataset. This is because the ancestral relationships of duplicated sequences are hard to disentangle. This is in contrast to single copy nuclear genes that were present in their most recent common ancestor in one copy

2.4.2 *RepeatScout* and *RepeatMasker*

A combination of *RepeatMasker* (Smit et al., 2010) and *RepeatScout* (Price et al., 2005) were used to mask repetitive elements in the *B. pallida* and *B. gibbera* draft assemblies. Masking was done at this stage to greatly simplify the identification of orthologous sequences between the two assemblies (section 2.5.1).

RepeatScout (Price et al., 2005) is a *de novo* repeat finder. It predicts repeats in a genome assembly that can then be supplied to *RepeatMasker* (Smit et al., 1996-2013). *RepeatScout* works by first identifying high frequency subsequences in the input sequences of length *l*, or l-mers, as seeds (see section 2.5.1 below); l-mers are analogous to k-mers (discussed above). The most frequent l-mer is then extended to create a consensus sequence for a repeat family (Price et al., 2005). Other l-mers that belong to the same repeat family are identified and removed from the l-mer table. The process is repeated for the next most frequent l-mer until a threshold minimum l-mer frequency is reached (Price et al., 2005). *RepeatScout* was run with default settings and a fasta file of repeats was output for both assemblies for use with *RepeatMasker*.

The output fasta file containing *de novo* predicted repeats was combined with the *RepeatMasker* default repeat fasta file. This file contains repeats and low complexity sequences, such as simple tandem repeats, commonly observed in sequenced genomes (Smit et al., 2010). *RepeatMasker* was run using *RMBlast*, a modified version of *BLAST* (Altschul et al., 1997) optimized for *RepeatMasker*, to identify repeats. Masked versions of the *B. pallida* and *B. gibbera* assemblies were output with N's replacing predicted repeat sequences. *RepeatMasker* also provides detailed annotations of the masked sequences. However, most of the repeats masked were not annotated as they were identified using the *de novo RepeatScout* predicted repeats.

For *B. pallida* 51% (408 933 208 bases) and *B. gibbera* 34% (149 745 254 bases) of the assemblies were masked respectively. For comparison, using *RepeatScout* Wang et al. (2008) masked 20% of the 144Mb assembled *Drosophila melanogaster* genome and 26% of the 151Mb red flour beetle, *Tribolium castaneum,* genome.

2.5.1 Identifying orthologous sequences between outgroup and ingroup sequences

Reciprocal *discontiguous megablasts* (Altschul et al., 1990) of the masked assemblies identified orthologous regions between the two species. In explanation, for two sequences *X* and *Y* from species *x* and *y*, respectively, if sequence *X* is the best *BLAST* hit for sequence *Y* and sequence *Y* is the best *BLAST* hit for sequence *X, X* and *Y* are reciprocal best hits *(RBHs) (*explanation adapted from Salichos & Rokas, 2011*)*. The relatively simple RBH approach has been shown to compare well to more complex algorithms for ortholog identification (Altenhoff & Dessimoz, 2012). It is also appropriate for this dataset as *BLAST* works well with the short sequences representative of the assemblies. The *BLAST* algorithm can be parallelised so the reciprocal *BLASTs* were split up into hundreds of sub-jobs and run using the Edinburgh Compute and Data Facility (ECDF) (http://www.ecdf.ed.ac.uk/).

Discontiguous megablast is recommended for cross-species searches (http://www.ncbi.nlm.nih.gov/Web/Newsltr/FallWinter02/blastlab.html). The *BLAST* algorithm works by identifying short matching sequences between compared sequences. After finding shared sequences a local alignment is made, an operation called seeding. *Discontiguous megablast* differs by not using an exact contiguous word match to seed alignments. Instead, an equivalent, user-specified, number of non-contiguous positions within longer template seed alignments are used. For example, in a coding sequence, 12 bases of a template of 18 bases (or 6 amino acid codons) could be required to match exactly at codon positions 1 and 2 (equivalent to: 110110110110110110 where '1' is an exact match). The third base (represented by a zero) is allowed to wobble in accordance with third base degeneracy. This approach is more sensitive than searching for exact matches with a sequence length of 12 (http://www.ncbi.nlm.nih.gov/Web/Newsltr/FallWinter02/blastlab.html). A sequence length of 11 with a mixed coding and non-coding template (different to that above) of length of 16 was used for the *discontiguous megablasts* with an e-value cut off of 1 x 10^{-20}, and filtering of low-complexity sequences to avoid spurious matches.

The masked assemblies of *B. pallida* and *B. gibbera* were blast searched against one another. Masking of repeats (section 2.4.2) greatly reduced the number of initial blast hits, making RBH filtering simpler. Sequences around masked repeats within the same contig were kept for analysis. Multiple *BLAST* hits can occur along a contig; therefore it was possible, and common, for a contig to have several RBHs along its length to different contigs in the other species. Overlaps between best hits of 15 bases were allowed to avoid penalising good unique hits with short overlaps along one contig. A RBH was only kept if the difference in bit scores (a standardized score for the alignment comparable across separate blast searches) with the next best overlapping hit (>15 bp long) for that contig in the query species was greater than or equal to 100. This bit score filter was to avoid including false RBH orthologs because the *BLAST* algorithm chose the wrong alignment of two or more alignments with close scores. The filtering process was repeated for both reciprocal *BLASTs*: *B. pallida* versus *B.gibbera* and *B.gibbera* versus *B. pallida*. Bit score and overlap filtering were performed using a perl script on the *BLAST* output.

After RBH filtering, there were 323 693 blast hits remaining spanning 240 012 *B. pallida* and 301 282 *B. gibbera* contigs respectively. The effect of repeat masking and filtering by RBH on the distribution of contig average coverage can be seen in figure 2.9. The mean coverage is now far less skewed by high coverages as many repetitive sequences have been removed, but the peaks of the distributions are similar to those shown in figures 2.7-8.

Fewer *B. pallida* contigs were in the final RBH orthologs dataset, as the *B. pallida* assembly was superior to *B. gibbera's* (table 2.3). This has the corresponding effect that, on average, *B. pallida* contigs are longer than those for *B. gibbera*. Thus, more multiple RBHs occurred along the longer *B. pallida* contigs than for shorter *B. gibbera* contigs. A new BAM alignment file was created for each individual containing only the putatively orthologous regions. Table 2.5 shows the number of reads overlapping regions of reciprocal best hits for each individual. Table 2.6 shows the number of reads mapping to the orthologous regions. Note that the properly paired mapping percentages are much higher than for the total assembly (table 2.4).

Individual	Total Reads	Paired	Singles
West A	11 874 298	11 491 196	383 102
West B	6 694 572	6 470 325	224 247
Centre A	16 672 756	13 697 004	2 975 752
Centre B	8 490 633	8 304 868	185 765
East	9 309 041	9 123 426	185 615
Outgroup A	8 567 014	8 313 168	253 846
Outgroup B	22 536 368	16 390 864	6 145 504

Table 2.5. Reads overlapping the orthologous regions of reference assemblies for ingroup and outgroup individuals. Paired refers to paired-end read fragments and singles to single-end reads.

Individual	Total Reads mapped	% TRM	Both pairs mapping	% BPM	Properly paired mappings	% PPM
West A	10 863 854	91.49	10 356 235	90.12	8 256 622	71.85
West B	6 058 958	90.51	5 782 148	89.36	5 081 054	78.53
Centre A	14 525 611	87.12	12 069 417	88.12	10 539 095	76.94
Centre B	7 805 239	91.93	7 505 425	90.37	6 472 021	77.93
East	8 466 774	90.95	8 163 865	89.48	7 138 558	78.24
Outgroup A	7 353 309	88.45	7 165 552	86.2	6 033 392	72.58
Outgroup B	16 790 607	74.5	12 665 104	77.27	9 862 156	60.17

Table 2.6. Reads mapping to the orthologous regions of reference assemblies for ingroup and outgroup individuals. %TRM = percentage total reads mapping; %BPM = percentage both pairs mapping; %PPM = percentage properly paired mappings.

2.6.1 SNP calling and extracting consensus sequence per individual

Raw variant calling across individuals for both species was performed using *Samtools mpileup* (Li et al., 2009); indels were not called, as current aligners are not capable of accurate indel prediction. *Samtools mpileup* output was in the Variant Call Format (VCF) for storing and parsing sequence polymorphisms. Consensus sequences of minimum length 300 bases were identified for each individual per species from the VCF files by a custom perl script (all custom scripts were written by J. Hearn). For ingroup sequences; (1) the reference base was called if no variant was present or the variant did not reach a user-defined SNP quality threshold, (2) an 'N' was coded if the individual had 0 coverage at that position or was called heterozygous by *samtools* indicating a sequencing error, or more than one polymorphism was present violating the assumption of the infinite sites model (in which only one variant at any site is possible); (3) a SNP was called if it was homozygous in that individual and above the user-defined quality threshold. Ingroup consensus sequences for each individual were produced at two SNP quality thresholds by VCF file filtering: Q0, and Q20 (equivalent to an error rate of 1 in 100) for comparison of results. This was to assess the effect of quality filtering on the number of SNPs for analysis and the ratio between the six mutational types ($k = \{k_w, k_e, k_c, k_{we}, k_{wc}, k_{ec}\}$) possible.

For the two diploid outgroup individuals a genotype called '0/1' by *mpileup* could be a true heterozygote position in addition to ‘0/0’ versus ‘1/1’ homozygous variants. All polymorphic positions in the outgroup were called as 'N' to avoid including ancestral polymorphisms segregating in in- and outgroup sequences. An ancestral polymorphism occurs when the common ancestor of two alleles at a locus existed before two species became isolated (Charlesworth, 2010). This is undesirable as the divergence times between such alleles in two species are greater than that of the species divergence times (Gillespie and Langley, 1979; Charlesworth, 2010). Sites were also called as 'N's at positions with 0 read coverage in both individuals. Therefore, a single, completely homozygous outgroup sequence was created that took advantage of deeper sequencing from combining sequence data across individuals. For the two outgroup *B. gibbera* sequences 531 328 putative polymorphic sites were identified between them and masked with ‘N’s, representing 0.5% of total sites in the VCF file.

2.6.2 Removing undesirable sequences using *BLAST* and coverage

Contigs were removed from the dataset if they had nucleotide (*blastn*) or translated nucleotide (*blastx*) *BLAST* top hits to bacteria, as both species contain *Wolbachia*, mitochondrial genome sequences, and other identifiably contaminant sequence such as mouse or human DNA. Repeats that had escaped masking by *RepeatScout* and *RepeatMasker* were also identified. These are repeats that are collapsed into one contig by the de Bruijn graph during assembly, allowing them to escape prediction as repeat sequence by *RepeatScout*

To remove repeats that had been collapsed into one contig during assembly, and therefore not identified by *RepeatScout*, coverage cut-offs were used. Contigs were removed if they had coverage above an arbitrary threshold using contig average coverage distributions as a guide (figures not shown, figure 2.9 shows distributions after coverage filtering). These were set at 75 fold for *B. pallida* and 30 fold for *B. gibbera* respectively, based on inspection of the coverage distributions (figure 2.9). These coverages cut-offs are where the respective distributions approached very low frequencies in each species. The cut-offs remove the ‘long-tail’ of the frequency distributions consisting of high-coverage contigs. The final *B. pallida* dataset had a modal coverage across all individuals of ~7.5 after filtering (figure 2.9). Filtering in this way is conservative as contigs representing single-copy nuclear DNA sequences that were sequenced at greater depth than most of the genome will also be present in the ‘long tail' of the distribution After BLAST and coverage filtering, 304 027 hits remained across 232 097 *B. pallida* and 290 379 *B. gibbera* contigs respectively (table 2.7).

Species	Orthologous segment N50	Number of contigs	Total bases	Average GC
B. pallida	734	232 097	113 583 710	36.7
B. gibbera	508	290 379	111 785 775	35.9

Table 2.7. Assembly statistics for orthologous regions in the in- and outgroup after all filtering.

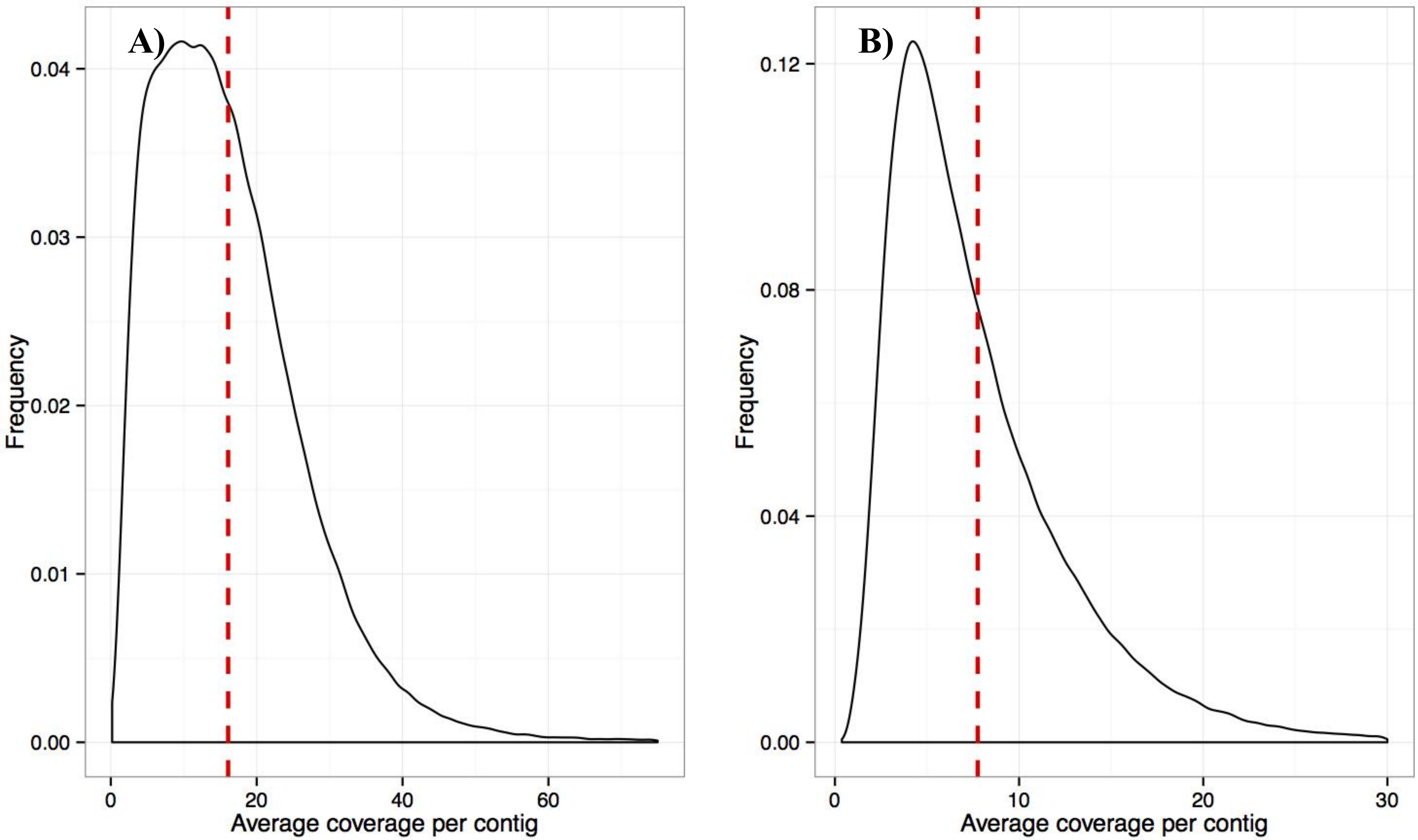

Figure 2.9. Average coverage per contig density plot of remaining orthologous regions for (A) *B. pallida* and (B) *B. gibbera* after filtering of contaminant and high-coverage sequences. The average coverage, shown by the dashed red line, is now much closer to the mode than for the plots for all of the data (figures 3.3-4) but is still effected by right skew of the data.

2.7.1 Multiple sequence alignment of ingroup and outgroup sequences

Multiple sequence alignments for each sequence block were generated for the four triplet combinations of *B. pallida* individuals and the outgroup using the aligner *MUSCLE* (Edgar, 2004). They were output as simple fasta alignments (sequence blocks) of 300 bases minimum length, an arbitrary cut-off that simplified further bioinformatic preparation of the dataset. Indels occurring between ingroup and outgroup were removed as were any sites violating the infinite sites assumption or coded as ‘N’ in the ingroup VCF filtering, using custom perl scripts. The combinations were labelled: *WaCaE WbCaE WaCbE WbCbE,* based on the three individuals combined. For example, *WaCaE* represents individuals *West A – Centre A – East*.

To avoid linkage between separate blocks in the final dataset and to increase the length of some blocks *MUSCLE* multiple sequence alignments derived from the same *B. pallida* contigs were combined linearly into single blocks. This occurs when multiple distinct *Blast* RBH alignments occur along one *B. pallida* contig to outgroup *B. gibbera* contigs. The alignment derived from the RBH closest to the start of the *B. pallida* contig was placed first, followed by the second closest and so on. However, only one alignment per outgroup contig was kept when multiple RBHs *B. gibbera* per contig occurred, meaning the other alignments along the same *B. gibbera* contig were removed from the dataset.

Finally, the raw blocks were filtered to a length of 2 000 bases (2 kb), as this length represented a good trade-off for obtaining blocks long enough to include enough polymorphic sites for inference and short enough to be unconcerned about recombination within contigs. Sub-sampling from the full set of contigs with this length cut-off gave between 2419-2889 (table 2.8) blocks (depending on the combination of W/C/E individuals), roughly 10% of the contigs meeting the initial filtering requirements.

These blocks were further sub-sampled to 1 000 and 500 bases length to assess the robustness of the likelihood method to differing block sizes. Sub-sampling the datasets using this length cut-off meant that approximately 10% (table 2.8) of bases remaining after creation of sequence blocks were used in the likelihood analysis.

Alignment	Block length	Number of contigs	Total bases (% remaining)	Filtered contigs	Filtered bases (% remaining)	Average GC
WaCbE	>300	83 383	60 277 500(100)	69 032	49 038 121(100)	38.6
WaCaE	>300	84 822	61 012 720(100)	70 286	49 630 679(100)	38.6
WbCbE	>300	77 752	54 117 641(100)	64 721	44 317 891(100)	39
WbCaE	>300	79 175	54 842 541(100)	65 931	44 902 450(100)	38.8
WaCbE	2 kb	2 889	5 778 000(9.6)	2 648	5 296 000(10.8)	39.7
WaCaE	2 kb	2 871	5 742 000 (9.4)	2 640	5 280 000(10.6)	39.5
WbCbE	2 kb	2 419	4 838 000(8.9)	2 231	4 462 000(10.1)	40.2
WbCaE	2 kb	2 419	4 838 000(8.8)	2 231	4 462 000(10.1)	40.1
WaCbE	1 kb	2 889	2 889 000(4.8)	2 648	2 648 000(5.4)	39.4
WaCaE	1 kb	2 871	2 871 000(4.7)	2 640	2 640 000(5.4)	39.3
WbCbE	1 kb	2 419	2 419 000(4.5)	2 231	2 231 000(5.0)	39.9
WbCaE	1 kb	2 419	2 419 000(4.5)	2 231	2 231 000(5.0)	39.7

Table 2.8. Alignment statistics for all four datasets for no filtering, and block lengths of 2 kb and 1 kb1 kb. Filtered columns refer to dataset after removal of linked blocks identified by transcriptomics (see section 2.9). Block length is in base pairs.

2.8.1 The effect of length and quality filtering on the frequency of polymorphisms

The full *WaCaE* datasets comprised 84 822 aligned contigs >300 bp (table 2.8) with an N50 value of 803 bases. A total of 171,694 polymorphic sites were recovered in the in-group, corresponding to an average per site diversity (as measured by Watterson's Θ_W of 0.188%) (table 2.9). Average per-site divergences between outgroup and the Eastern individual was 4%. If the ‘Out of the East’ model is true, represented by a population divergence in the order (E,(C,W)) without admixture, we expect derived sites shared by Central and Western individuals (C/W) to be more common than both derived sites shared by Central and Eastern individuals (C/E) and sites shared by

Western and Eastern individuals (W/E). Likewise, without gene flow after divergence, (C/E) and (W/E) sites, which correspond to internal branches of genealogies that are incongruent with the population history, are expected to occur at equal frequency (Hudson 1983, Tajima 1983). Analogously, under null models of a polytomic split or a single panmictic population, all three types of shared derived sites are equally likely. Contrary to these simple models, (C/E) sites were more frequent (9.6%) than (W/E) sites (5.1%), which in turn were more frequent than (W/C) sites (2.8%) (see top two rows of table 2.9 and figure 2.10 for 2 kb count distributions). This double asymmetry suggests that simple divergence models without gene flow are likely to provide a poor fit to the data. If we assume that the majority class of informative sites corresponds to the order of population divergence, then these results imply that the Western population diverged from the common ancestor of the Central and Eastern populations before these in turn diverged. Under this model, the observed excess of (W/E) sites relative to (W/C) sites could arise as a consequence of gene flow between Western and Eastern refugia after the more recent (C/E) split (Durand et al., 2011; Lohse and Frantz, 2013).

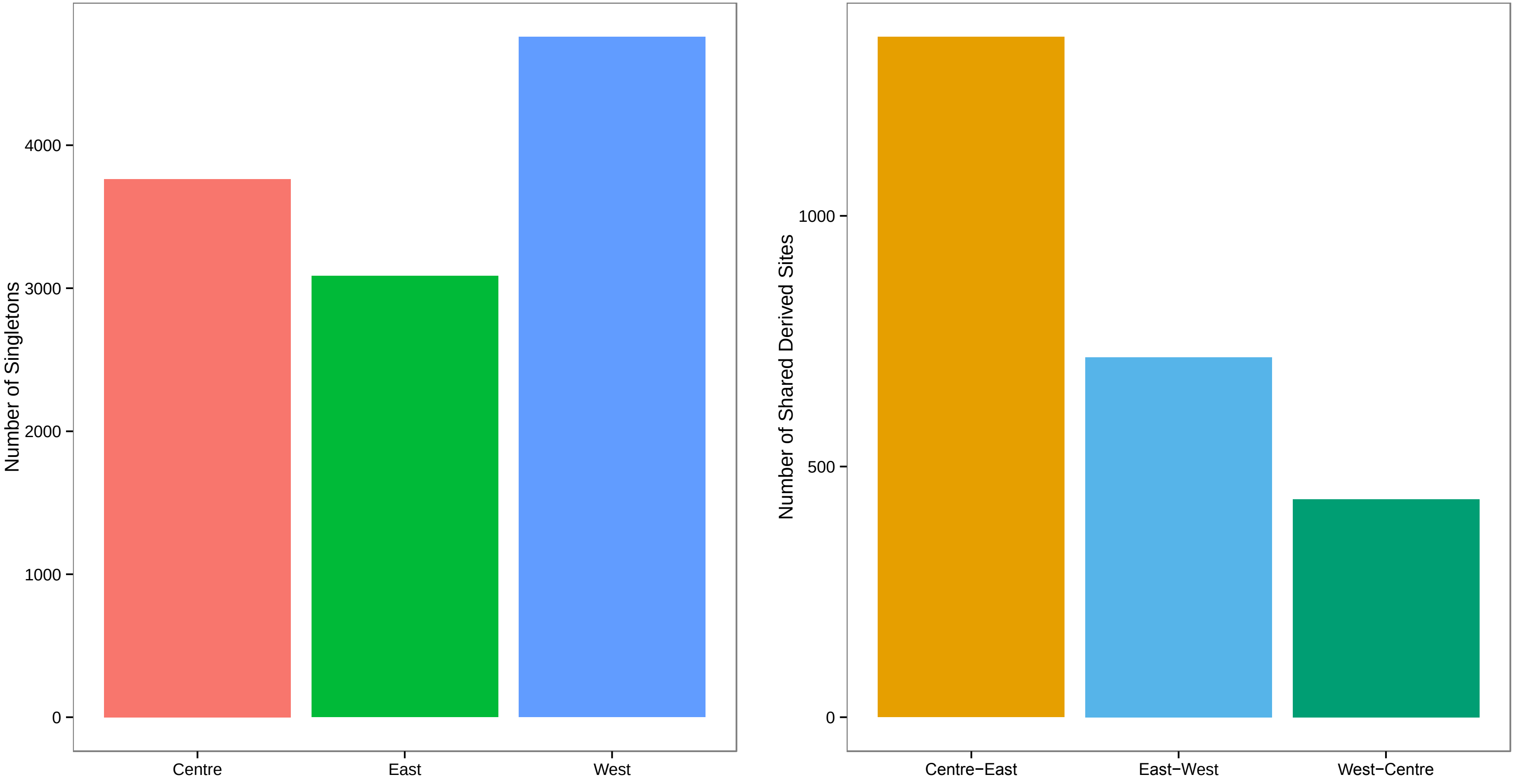

Figure 2.10. Distribution of mutational types for *WaCaE* 2 kb dataset for singletons, mutations only found in one refugia and shared-derived sites. Note the difference in scale of the Y-axis of each plot. This plot demonstrates the distributions of allele frequencies shown for the full and 2 kb datasets in table 3.7 as SNP counts.

2.8.2 Length and number of blocks

The assumption of no linkage between blocks imposes a severe limit on the number of blocks that could be analysed. Without information about the relative position of blocks in the genome, the number of blocks must be chosen such that the probability that two blocks are physically separated by less than some minimum distance by chance can be ignored. Assuming a genome size of 1.75Gb for *B. pallida* (the average measured in oak gall wasps, Lima, 2012) and sampling of blocks by chance alone, the distance between neighbouring blocks is exponentially distributed with rate n/1.75Gb (where n is the number of blocks). For example, if we classify blocks separated from their nearest neighbour by 20kb or more as being in linkage equilibrium and want to ensure that less than 5% of all blocks fall below this threshold, we could in theory sample a maximum of $-(1.75Gb \times Log[0.95])/20kb \approx 4500$ blocks. Sampling contigs longer than 2 kb from the full triplet datasets resulted in about half this theoretical maximum. For ease of comparison across different sequence blocks, we fixed the number of 2 kb blocks to the minimum of number of sequence blocks for all datasets (i.e. in each W/C/E combination of individuals, we randomly sampled that number of blocks).

Filtering contigs by length could result in various biases that might affect inference. For example, more conserved and/or structurally complex regions of the genome with lower divergence rates are expected to assemble better and align with fewer errors, and so should be represented by longer contigs. To quantify this effect, we correlated contig length against per site divergence in the *WaCaE* data. As expected, longer contigs were on average less diverged (figure 2.11) (Kendall's $\tau = -0.0419$, $p < 10-6$). Consistent with this, the average per site diversity (θ_W) in the 2 kb filtered *WaCaE* data was about half of that in the unfiltered data (table 2.9). This confirms that length filtering does enrich for conserved sequences. However, for the purpose of estimating population history, any overall bias in absolute diversity can be incorporated by a simple rescaling of the mutation rate. In contrast, to justify treating the length-filtered data as a random sample of genealogies in the genome requires that the length filtering does not affect the relative frequency of the six possible mutational types ($k = \{k_w, k_e, k_c, k_{we}, k_{wc}, k_{ec}\}$) on the genealogical branches of the three populations splitting model (2.1) (i.e. the frequency of mutational types normalized by the proportion of polymorphic sites).

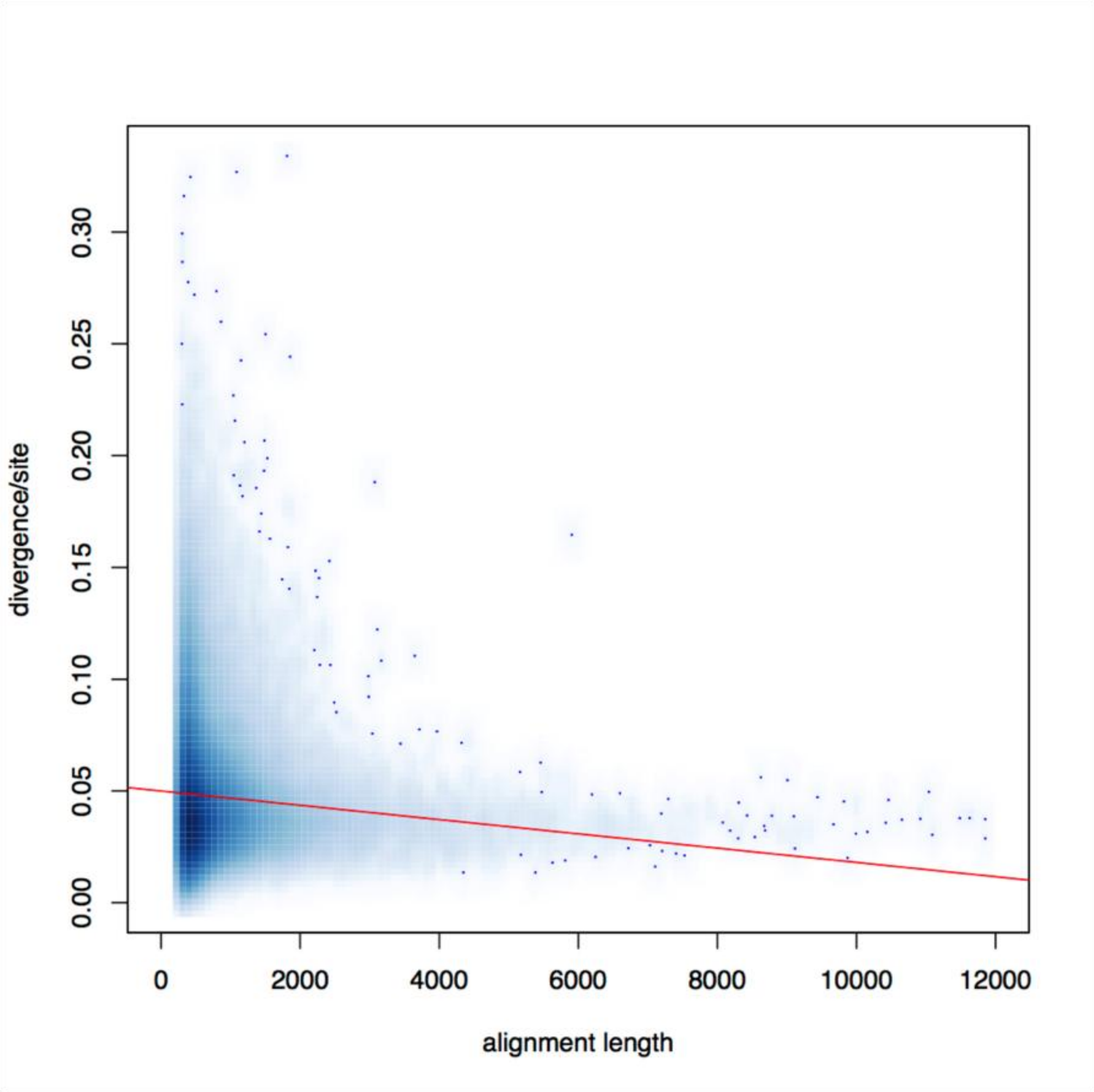

Figure 2.11. Scatter plot of divergence/site against block length indicating longer contigs are less divergent/more conserved. Individual blocks have been smeared to show densities. Red line = line of best fit. The red line is a line of best fit through the data demonstrating the negative trend.

2.8.3 Frequency of mutations between raw and length filtered datasets

To compare the frequencies of mutational types in the full and length-filtered *WaCaE* data, we obtained a random sample of unlinked SNPs in each dataset by picking one SNP at random from each sequence block. In the length- filtered data, all 2 kb blocks were included. In the full data, SNPs were drawn from a random sample of 4500 sequence blocks (the maximum estimated to be linkage free at a plausible recombination rate, as explained above) to avoid linkage effects. There was no significant difference in the relative frequencies of the three types of shared derived mutations (table 2.9) ($\chi2$ = 1.96, p = 0.38) between the filtered (length > 300bp) and unfiltered data (length 2 kb) for the *WaCaE.* However, there was a significant (but slight) excess of singleton mutations compared to shared-derived sites in the 2 kb data ($\chi2$ = 9.3, p = 0.0023) in the *WaCaE* dataset. This may be either due to assembly or alignment bias or purifying selection (which is likely to be stronger in the 2 kb filtered data as it contains a greater proportion of expressed sequence) (Fu and Li, 1993).

Dataset	**length**	ϑ_W	***W***	***C***	***E***	***W/C***	***W/E***	***C/E***
WaCaE	>300bp	0.00188	0.325	0.214	0.263	0.040	0.058	0.100
WbCbE	>300bp	0.00147	0.269	0.244	0.283	0.044	0.060	0.100
WaCaE	2 kb	0.00089	0.338	0.22	0.267	0.027	0.049	0.098
WbCbE	2 kb	0.00079	0.276	0.25	0.287	0.035	0.054	0.099

Table 2.9: Genetic diversity and relative frequencies of mutational types in *B. pallida* sequence blocks for >300 bp and 32 kb datasets.

2.8.4 The effect of quality filtering on the final datasets

The number of SNPs in the final 2 kb dataset at Q0 and Q20 filtering were compared across individuals (figure 2.12). There is very little difference between Q0 and Q20 numbers for singletons and shared-derived sites; for two comparisons the numbers are identical. The reason for the negligible effect of quality filtering is probably the strict haploid-based filtering of the SNPs on an already heavily filtered dataset.

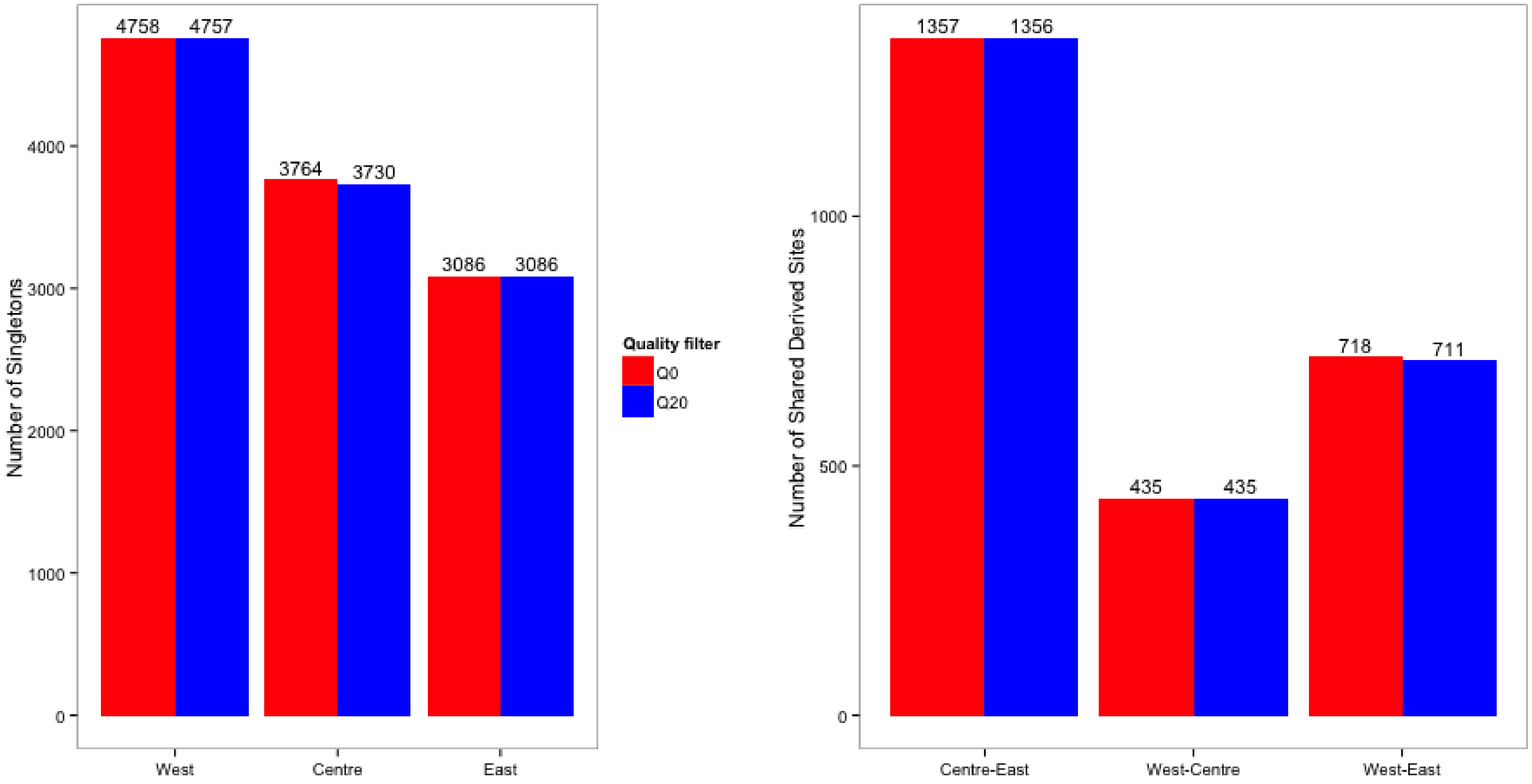

Figure 2.12. Number of singletons and shared derived sites for Q0 and Q20 filtered data for *WaCaE* 2 kb sequence blocks.

2.9.1 Leveraging the *B. pallida* transcriptome to improve the final dataset by using it to fit mutational heterogeneity to blocks

2.9.2 Proportion of expressed sequence per block

The final sequence blocks were expected to be a mix of coding and non-coding sequences. However, the term ‘expressed sequence’ is used here over coding sequence (CDS) because many of the individual transcripts of the *B. pallida* larval transcriptome contain 5’ and 3’ untranslated regions (UTRs). It is reasonable to assume that the mutation rate will differ (mutational heterogeneity) between the two types of sequence. To fit mutational heterogeneity the proportion of expressed sequence was identified for each alignment. The sequence blocks were then partitioned according to the proportion of expressed sequence. The effective neutral mutation rate was scaled to achieve a mutation rate for each of these bins. The scaling factor was the within bin divergence per site relative to the total divergence across all sites.

To identify the expressed sequences a *B. pallida* transcriptome (table 2.10, see Chapter 3) was used. The transcriptome is generated entirely from larval tissues and as such any adult specific expression is missed. Thus the transcriptome is not a complete gene set and the estimated proportion of expressed sequence is an underestimate of the true proportion of expressed sequence. The *Trinity* assembler that generated the transcriptome assembly outputs transcripts with UTRs (Grabherr et al., 2011).

***B. pallida* transcriptome assembly metrics**	
N50 (bp)	1 736
Number of transcripts	108 459
Maximum transcript size (bp)	37 465
Transcriptome length (bp)	94 447 801

Table 2.10: Basic statistics for *B. pallida* used to assign expressed sequences to sequence blocks. See chapter 3 for more in depth description.

2.9.3 Linked sequence blocks

The second application of the transcriptome was to identify linked sequence blocks. If different sequence blocks were found to match the same transcript it was assumed that they were sampled from regions adjacent to one another in the genome. In these cases only one of the linked sequence blocks was kept for analysis and the other(s) discarded (see table 2.11, effect of filtering linked blocks).

2.9.4 Assigning proportions of expressed sequence to blocks

The following method was applied to each dataset of 1 kb, 2 kb and unfiltered lengths. Firstly, the sequences for one of the individuals was removed from each alignment and placed into a fasta file. This fasta file was then searched against a nucleotide *BLAST* database (blastn) of the transcriptome (see table 2.10 for basic statistics of the transcriptome). A minimum e-value of 1 x 10^{-20} was required to accept that the sequence is expressed and a maximum of 10 hits were recorded per alignment. The blast results file (output format “6”) was then sorted to find sequence blocks with multiple hits to the transcriptome. One of these sequence blocks was kept and the others removed from the dataset, as they are probably all linked in the *B. pallida* genome (section 2.9.3). Linked sequence blocks identified in the 2 kb blast hits were used to filter the 1 kb dataset. This was to take advantage of the longer 2 kb sequence blocks and therefore chance of a blast hit to the transcriptome in the 2 kb alignment; it also meant the final 2 kb; 1 kb and 500b datasets had the same number of sequence blocks.

Then a *BED* file (a file compatible with the sequence manipulation tools of the *BEDtools* program, Quinlan & Hall, 2010) of regions within sequence blocks that matched expressed sequences was created. These regions were merged together as many transcripts overlapped the same region of the alignment. This was done using the *mergeBed* tool (Quinlan & Hall, 2010). Because the *Trinity* assembler (Grabherr et al., 2011) outputs potential isoforms of a gene such multiple mappings are expected to occur. The proportion of expressed sequence for each alignment was computed from the *mergeBed* output by dividing the region of the alignment covered by transcripts by the total length of the alignment. Proportions were then combined into one proportion per alignment using a perl script. This was necessary where transcripts matched to non-overlapping regions of the alignment, because *mergeBed* does not merge such regions. The proportions were appended to the filtered *MUSCLE* sequence blocks file for use in the maximum likelihood analyses.

Table 2.11 shows the results of filtering linked sequence blocks for the *WaCaE* dataset and table 2.8 the numbers of blocks remaining for all four comparisons. A much higher proportion of the 2 kb dataset contains expressed sequences than the full dataset at 52% to 31% compared to the unfiltered dataset. This is probably because longer blocks of sequence represent more unique regions of the genome. Such regions are easier to assemble than more repetitive regions and resulting in longer contigs assembled.

Dataset	***WaCaE* 2 kb**	***WaCaE* unfiltered**
Number of blocks	2 871	84 822
Blocks hitting transcripts	1 501	25 883
Percentage	52	31
Number blocks kept	1 92	6 501
Number removed	2 31	14 536
Remaining blocks	2 640	70 286

Table 2.11: The effect of removing linked sequence blocks on total numbers of blocks.

To be able to compare likelihoods across datasets, we fixed the number of blocks to 2 231 in all analyses, the lowest number of final sequence blocks for any combination of individuals (table 2.8, comparison *WbCbE*) after removing blocks that may be linked based on alignment of sequence blocks to the *B. pallida* transcriptome.

2.9.5 The proportion of expressed sequence in a contig correlates negatively with mutations per site

In total, 50% of all contigs in the *WaCaE* dataset had no hit to the transcriptome and are thus considered non-expressed. Across all sites the proportion of expressed sequence was 70% for those contigs that did contain expressed sequence. This together with the increased GC content in the filtered datasets (for WaCaE 38.6 – 39.5%, table 2.8 compared to table 2.3) clearly showed that our filtering strategy enriched for expressed sequence. Figure 2.13 shows the negative relationship of mutations per site versus proportion of expressed for the *WaCaE* triplet from 2 kb filtered data. The regression line indicates a strong negative relationship (Kendall's tau = -0.389, p-value = 2.22 x 10^{-16}). This correlation is much stronger than for the *WaCaE* triplet for all data unfiltered for length (Kendall's tau = -0.0419, p-value =< 2.22 x 10^{-16}), which probably reflects the greater proportion of expressed sequence in the 2 kb dataset versus the length unfiltered blocks. The plot confirms the expectation that expressed sequences are under purifying selection and many deleterious polymorphisms are removed. Much unexpressed sequence is probably under no such constraint, however purifying selection does occur in non-coding sequence (Halligan et al., 2011). Therefore, it cannot be concluded that expressed and unexpressed sequence represents a dichotomy between sequence under selection and neutral sequence.

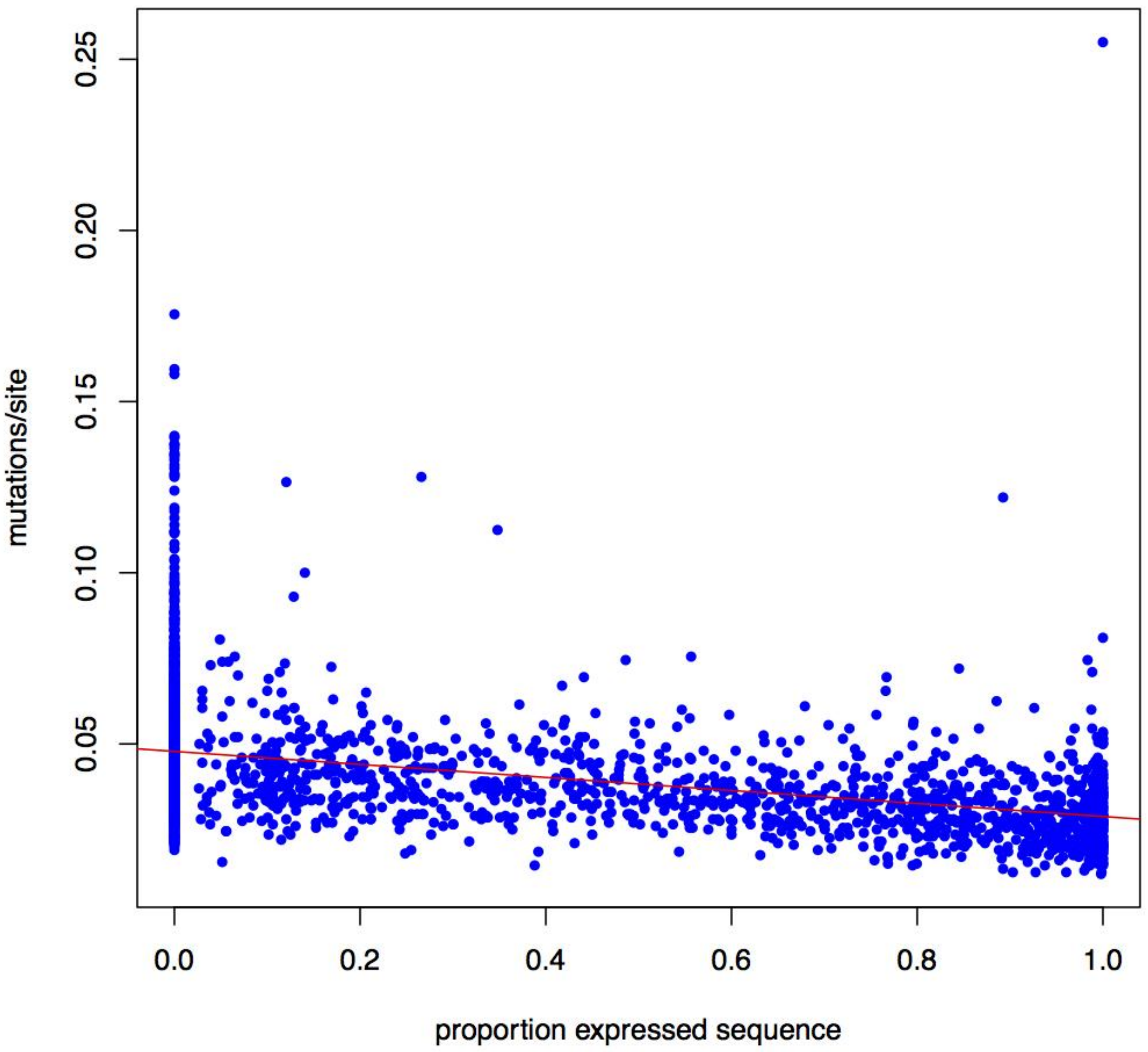

Figure 2.13. Mutations per site against proportion of expressed sequence for the *WaCaE* 2 kb dataset, the dashed red line is a line of best fit showing the negative trend in mutations/site with increasing proportion of expressed sequence.

2.10.1 Likelihood analyses of historical models

K. R. Lohse carried out the maximum likelihood analyses in *Mathematica* v8 (Wolfram Research, 2010).

The data were summarized as a vector of mutational types in each sequence block. Likelihoods of model parameters given the numbers of the six mutation types in a block were calculated numerically. The probability of observing a particular mutational configuration in a sequence block (which can be interpreted as the likelihood of the model) can be expressed in terms of the partial derivatives of a generating function (Lohse et al., 2011). Assuming that alignment blocks are unlinked and hence statistically independent, the joint logarithm of the likelihood (lnL) across blocks is the sum of individual block lnL.

To conduct a broad search of model space, we took a strict divergence model between three populations as a starting point and considered all histories that involve a single unidirectional admixture event either to or from the oldest (or first diverging) population. Models with bidirectional or multiple admixture events were not considered because the additional parameters are computationally intractable, and also because these models are biologically unexpected: expansion out of refugia is expected to be a unidirectional process. For each of the six models, we numerically computed the parameter values that maximized lnL across a large number of sequence blocks of fixed length.

2.10.2 Likelihood model results

Comparing the three possible histories of strict divergence, a population tree topology (W,(C,E)) had highest support (ΔlnL), as expected from the frequencies of shared derived sites. Allowing for different values of N_e in the two ancestral populations did improve model fit (table 2.12). However, 8–9 of the 18 models involving admixture had greater support (table 2.12). The best supported history still assumes a (W,(C,E)) population tree topology but involves substantial admixture (proportion of admixture, *f* = 0.76 − 0.83) (table 2.12) from the Eastern into the Western refuge shortly after the split between Centre and East (model B in figure 2.1).

The *WaCaE* and *WbCbE* sequence blocks yielded the same ranking of models and gave very similar parameter estimates with broadly overlapping 95% confidence intervals (tables 2.12-13 and figure 2.14). For completeness, analogous analyses for the other two possible triplet datasets (i.e. *WaCbE, WbCaE)* were also run*,* both of which again identified the same best model and gave similar admixture estimates (*WbCaE f* = 0.85; *WaCbE f* = 0.69). Interestingly however, the estimated admixture proportion *f* was slightly higher in both triplet analyses involving the individual from Southern Spain (figure 2.12) (see Discussion). Repeating the analysis for *WaCaE* at block lengths of 500 and 1000 bases resulted in the same model choice and similar parameter estimates (appendix, tables 2.15-16).

	k			
Panmixia	1	-589.3		
Polytomy	2	-88.7		
Gene flow		**$(W_1;(C_2;E_3))$**	**$(C_1;(E_2;W_3))$**	**$(E_1;(C_2;W_3))$**
A) 2➔1	5	-9.1,(T_1)	-18.8	-18.2,(f*)
B) 3➔1	5	**0**	-88.7,(T_1, T_2)	-88.7,(T_1,f*)
C) 2/3➔1	5	-4.8	-88.7,(T_{gf},f*)	-88.7,(T_{gf},T_2)
D)1➔2	5	-25.7, (f)	-18.2,(T_1)	-18.2,(f*)
E)1➔3	5	-18	-88.7,(T_1, T_2)	-88.7,(T_1,T_2)
F)1➔2/3	5	-25.7, (f*)	-79.4	-33.4,(T_{gf})
2 pop.	2	-260.8	-404	-474.5
3 pop.	2	-25.7	-88.7,(T_2)	-88.7,(T_2)
2 pop. N_e	3	-48.5	-90.1	-93.7
3 pop. N_e	4	-20.8	-88.7,(T_2)	-88.7,(T_2)

Table 2.12. Support for alternative scenarios of divergence and admixture in the oak gall wasp *B. pallida* (*WaCaE*, 1 kb data). Support (ΔlnL) relative to the best model (given a value of **0**) for alternative histories of refugial populations of *B. pallida* estimated from the *WaCaE* dataset (Model B, **$(W_1;(C_2;E_3))$** in Fig. 3.1 has highest support and is shown in bold)..The labelling of populations (1–3) and of models (A–F) corresponds to that in Fig. 3.1; all scenarios involving unidirectional admixture were assessed for each of the three possible orders of population divergence (columns 1–3). Models of strict divergence without admixture between two (2 populations i.e. $T_1 = 0$) or three (3 pop.) populations were fitted assuming either a single or two different N_e (indicated where N_e is included in the row headings) for ancestral populations Parameters for which the MLE is 0 (i.e. the model reduces to a simpler nested model) are indicated in brackets (f*) refers to complete admixture, i.e. f = 1).

	k			
Panmixia	1	-589.3		
Polytomy	2	-88.7		
Gene flow		$\mathbf{(}W_1;\mathbf{(}C_2;E_3\mathbf{))}$	$\mathbf{(}C_1;\mathbf{(}E_2;W_3\mathbf{))}$	$\mathbf{(}E_1;\mathbf{(}C_2;W_3\mathbf{))}$
A) 2➔1	5	-14.9,($\mathbf{T_1}$)	-21.1	-33.2,($\mathbf{f}$*)
B) 3➔1	5	**0**	-59.9,($\mathbf{T_1}$)	-59.4,($\mathbf{T_2}$,$\mathbf{T}_{gf}$)
C) 2/3➔1	5	-14.3	-59.9	-60.3,($\mathbf{T}_{gf}$, $\mathbf{f}$*)
D)1➔2	5	-18.0	-19.4,($\mathbf{T_1}$)	-19.4,($\mathbf{T_1}$)
E)1➔3	5	-18	-60.0,(**f**)	-60.0,($\mathbf{f}$*)
F)1➔2/3	5	-33.2,($\mathbf{f}$*)	-49.7	-14.4,($\mathbf{T}_{gf}$)
2 pop.	2	-265.3	-293.6	-386.7
3 pop.	2	-33.2	-60	-60.3,($\mathbf{T_2}$)
2 pop. N_e	3	-46.1	-60	-64.7
3 pop. N_e	4	-31.0	-60	-60.3,($\mathbf{T_2}$)

Table 2.13. Support for alternative scenarios of divergence and admixture in the oak gall wasp *B. pallida* (*WbCbE*, 1 kb data). Support (ΔlnL) relative to the best model (given a value of **0**) for alternative histories of refugial populations of *B. pallida* estimated from the *WbCbE* dataset (Model B, $\mathbf{(}W_1;\mathbf{(}C_2;E_3\mathbf{))}$ in Fig. 3.1 has highest support and is shown in bold). The labelling of populations (1–3) and of models (A–F) corresponds to that in Fig. 3.1; all scenarios involving unidirectional admixture were assessed for each of the three possible orders of population divergence (columns 1–3). Models of strict divergence without admixture between two (2 populations i.e. $T_1 = 0$) or three (3 pop.) populations were fitted assuming either a single or two different N_e (indicated where N_e is included in the row headings) for ancestral populations. Parameters for which the MLE is 0 (i.e. the model reduces to a simpler nested model) are indicated in brackets (f*) refers to complete admixture, i.e. f = 1).

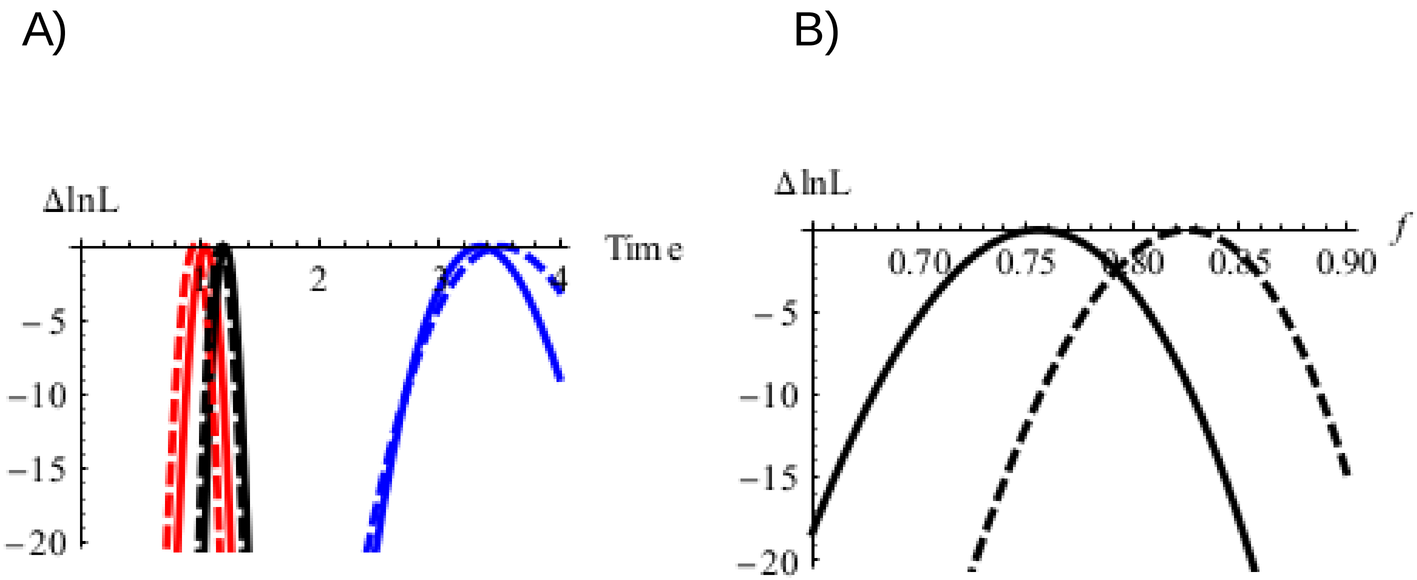

Figure 2.14. A) ΔlnL plots for the times of divergence (T_1 (black) and T_2 (blue)) and admixture T_{gf} (red). Estimates from the *WaCaE* data are shown as solid lines, those from the replicate data set *WbCbE* as dashed lines. B) *ΔlnL* for the admixture proportion *f.*

To provide an order of magnitude time calibration for the inferred history, we applied a direct, genome-wide estimate of the effective neutral mutation rate of 3.5 × 10^{-9} per site and generation measured in *Drosophila melanogaster* (Keightley, 2009). To account for the bias towards conserved sequence in our 2 kb filtered data, we scaled the *D. melanogaster* rate by the ratio of per site diversity in the filtered and unfiltered data (0.47 and 0.54 for *WaCaE* and *WbCbE* data respectively (see θ_W in table 2.9). Assuming that *B. pallida* has two generations per year (Csóka et al., 2005; Atkinson et al., 2003) this calibration gives effective population sizes between 39,000 – 52,000 (table 2.14). The time of admixture and the more recent split (t_{gf}, t_1) both date to the last glacial period (Weichselian, 12-110 thousand years ago [kya]), whereas the MLE for the oldest split (t_2) falls in the previous (Saalian, 130-200 kya) glacial period. However, because the molecular clock is from a different insect order, these absolute dates are tentative at best.

Finally, scaling the effective neutral mutation rate of each bin to account for mutational heterogeneity drastically improved model fit (table 2.14). It had no impact on the ranking of alternative models or parameter estimates under the best-supported model. However, we did find that incorporating mutational heterogeneity led to a slight reduction in both divergence time and N_e estimates (table 2.14).

Dataset	*μ het*.	*ΔlnL*	*f*	*Θ(Ne)*	*Tgf (tgf)*	*T_1 (t_1)*	*T_2 (t_2)*
***WaCaE*, 1 kb**	no	-9269.3	0.76 (0.72, 0.79)	0.69 (52 000)	1.04(54KY) (51–58KY)	1.21(63KY) (60-66KY)	3.34(173KY) (158-189KY)
***WbCbE*, 1 kb**	no	-8815.1	0.83 (0.80,0.86)	0.64 (43 000)	0.95(41KY) (38–44KY)	1.17(50KY) (51-57KY)	3.51(151KY) (135-168KY)
***WaCaE*, 1 kb**	yes	-8769.7	0.76 (0.72,0.79)	0.61 (45 900)	1.1(50KY) (47–54KY)	1.26(58KY) (55-60KY)	3.45(158KY) (143-172KY)
***WbCbE*, 1 kb**	yes	-8444	0.82 (0.79,0.85)	0.58 (39 100)	0.97(38KY) (35–40KY)	1.17(51KY) (49-54KY)	3.47(136KY) (121-151KY)

Table 2.14 Parameter estimates under the best-supported model. MLE are given for different triplet combinations and analyses with and without mutational heterogeneity. Both effective population size and divergence time parameters are scaled relative to the rate of coalescence, i.e. in 2N*e* generations. Absolute values calibrated using a direct, genome-wide mutation rate for Drosophila (Keightley et al., 2009) and assuming two generations per year are given in brackets. 95 %C.I. of scaled parameter values are given in brackets below the point estimate.

2.11 Discussion

The results show how outgroup-rooted sequence blocks of thousands of orthologous sequence blocks can be generated for multiple individuals using low-coverage genomic data and standard *de novo* assembly tools. Although the requirement for orthologous sequences in in- and outgroup and the filtering against repetitive sequences and short contigs enrich for coding and otherwise selectively constrained sequence, in the case of *B. pallida*, the frequency of mutational types is little affected. This suggests that the resulting data provide a representative sample of neutral variation in the genome that, if analysed in a multi-locus framework are highly informative about recent history

2.11.1 Admixture dominates the history of *Biorhiza pallida*

The model fit to *B. pallida* of (W,(C,E)) population divergence with strong East to West admixture differs qualitatively from previous population genomic inferences of divergence with admixture (Green et al., 2010; Lohse et al., 2013; both on gene flow between modern *Homo sapiens* and Neanderthals) in two ways. Firstly, admixture is from the more recently diverged population (E) into the older population (W), so in the opposite direction to that observed in the three-population analysis of our own Neandertal ancestry (Green et al., 2010; Durand et al., 2011). Secondly, the history of *B. pallida* is dominated by admixture rather than divergence (table 2.12, $f = 0.76 - 0.83$). Despite this, the majority class of shared derived sites is still ‘C/E’, and so concordant with the order of population divergence (W,(E,C)). This is a peculiar consequence of the direction of admixture: going backwards in time, ‘W’ lineages that trace back to the ‘E’ population via admixture only spend a short time in the ‘E’ population before they trace back to the ancestral ‘C/E’ population.

Both the order of population divergence and the direction of admixture are unexpected. First, our inference of initial divergence of the Western refuge contrasts with a previous meta-analysis of 12 oak gall wasps (including *B. pallida* and 19 associated parasitoid species (Stone et al., 2012), as well as a multi-locus study that compared the history of four oak gall parasitoid species (Lohse et al., 2012). This history is also incompatible with mitochondrial DNA gene trees and patterns in allozyme diversity in other gall wasps (Stone et al, 2001, 2007; Rokas et al., 2003; Challis et al., 2007). Both studies found a general signature of (E,(C,W)) divergence on a community scale, but had insufficient power to resolve the order of population divergence in individual species (or to fit additional admixture parameters). Interestingly, however, the deep split of the Iberian population from other refugia here inferred for (*B. pallida*) is compatible with the mitochondrial genealogy reconstructed by Rokas et al. (2001). Second, the history of (*B. pallida*) involves substantial admixture from the Middle East into Iberia bypassing the Balkans. Migration into Iberia through North Africa, possibly via a Sicilian land bridge to Tunisia, is the most plausible route of dispersal. Striking floristic links between Iberia and Asia Minor have been found across a range of plant taxa (Davis and Hedge, 1971), including oaks (Lumaret et al., 2002), and there is genetic evidence that Iberia was colonised from North Africa during the Pleistocene by some animal taxa (Griswold and Baker, 2002; Habel et al., 2008). Our finding of a higher admixture fraction from the east for the sample from Southern (*Wb*) compared to Central (*Wa*) Iberia further supports a scenario of dispersal via North Africa. Similarly, the genetic similarity of extant populations of oak gall wasps (Rokas et al., 2003) and their parasitoids (Nicholls et al., 2010) in Morocco and Spain suggests that the Strait of Gibraltar presents little or no barrier to gene flow. Given the lack of molecular calibrations for Hymenoptera in general and gall wasps in particular, our absolute time estimates are tentative at best. Nevertheless, it is clear that the divergence and admixture between refugial populations of *B. pallida* is recent, encompassing no more than two or three glacial cycles.

2.11.2 Sampling the genome and the limits of power

While in the past, most statistical analyses of phylogeographic scenarios were limited in power by the number of available loci (Carstens et al., 2009; Lohse et al., 2012), the massive replication of sequence blocks afforded by short-read sequencing overcomes this and in the case of *B. pallida* allowed us to reliably identify the best fitting history among a set of alternative divergence and admixture scenarios.

However, despite increasing the number of loci by several orders of magnitude, the difference in support we find for some alternative models (tables 2.12-13) is still relatively modest, suggesting that the power to distinguish more complex models is limited. For example, it would be hard to distinguish multiple admixture events from a single event or a model of continuous migration (Hey 2005). It is worth reiterating that the lack of linkage information for the *B. pallida* assembly imposes limits the number of blocks we were able to include in the maximum likelihood analyses, i.e. the final analysis only included 2.2Mb of sequence, a mere 0.13% of the genome. In other words, most of the assembled genome remained unused. If one had complete linkage information, i.e. if the relative position of blocks was known, one could sample blocks at fixed intervals (Lohse and Frantz, 2013), which would increase the number of blocks that can safely be taken as unlinked by an order of magnitude. However, the gain in power is limited, as increasing the number of independently segregating blocks by a factor k increases the accuracy of parameter estimates by $\sqrt{k}$ (Lohse and Frantz, 2013), although still worth exploiting. Instead, it is the recent time-scale of *the B. pallida* history that sets an inherent limit to the complexity of models that one can hope to discriminate among, using a multi-locus approach.

Given this mutational limitation, it is clear that increasing the number of individuals sampled from within each population would also only slightly improve inference: most ancestral lineages would coalesce rapidly, i.e. the vast majority of genealogical branches added by larger samples would be unresolved, and so would not give much extra information. Very large samples of a long non-recombining sequence can be informative (Kong et al., 2011), but mainly about even more recent population history than the timescale considered here. Sampling individuals a further distance apart would give extra information, but also requires more complex models, involving multiple parameters for separation times and admixture rates. In general, these considerations suggest that there will be an upper limit to the signal contained in even an extremely large number of short, unlinked sequence blocks.

In contrast, we would have far more information if we could analyse the full linear sequence and explicitly use linkage information. In *B. pallida*, a total of 3.5% of the genome would be usable after filtering for unique orthologous sequence, but allowing an arbitrary degree of linkage; ultimately, of course, we could use the whole genome in such an analysis. In lieu of this, methods that do not require outgroup alignment would increase the size of the *B. pallida* dataset further. Pairwise sequence blocks of *Centre A* and *Centre B* individuals yielded 194 231 blocks covering 230 megabases (mb) with an N50 of 898 bp and 479 590 SNPs with quality scores greater than twenty. Whereas 2 231 blocks spanning 4.5 mb of sequence for the 2 kb datasets was used in this analysis. Additionally, such a dataset is far easier to construct than the one developed here as no identification of orthologous regions between species is required. The gain in power does not come primarily from this sheer volume of data; rather, we gain extra information from the lengths of sequence blocks. For example, the length of block that shares the same genealogy within a population is inversely proportional to its coalescence time, and the length of introgressed blocks of genome decreases with the time since introgression. Thus, recombination gives an additional time-scale, beyond that provided by mutation, as used here. Barton et al. (2013) show that in a two-dimensional continuum, the distribution of block lengths shared between genomes allows inference of both dispersal rate and neighbourhood size, whereas samples of allele frequencies do not give information about dispersal rate. Li and Durbin (2011) use the distribution of heterozygous SNPs to infer ancestral population size through time, while Harris & Nielsen (2013) use this information to infer complex migration histories. However, a full statistical analysis that takes into account the linear structure of the genetic map not only remains extremely challenging analytically, but also requires much better assemblies or linkage maps than can currently be achieved for most organisms in practice.

In the meanwhile, the combination of *de novo* assembly and numerical likelihood computation we develop here provides a level of resolution far beyond that of traditional phylogeographic analyses of a few loci. The fact that our bioinformatic pipeline yielded sufficient data (and resolution to distinguish between models) in an oak gall wasp, the group with the largest known genomes in the Hymenoptera (Lima, 2012), is encouraging and suggests that analogous analyses will be feasible in a large range of organisms or even whole ecological communities (Stone et al., 2012). This is because, for species with smaller genomes adequate coverage can be achieved at lower costs; a situation further improved by future advances in Illumina (and potentially other technologies) sequencing yields. Furthermore, our sensitivity analyses suggest that such inferences based on large numbers of blocks and few individuals are robust in two fundamental ways. Firstly, and despite the fact that undetected recombination can bias multi-locus analyses (Strasburg & Rieseberg, 2009), neither model selection nor parameter estimates are much affected by the length of sequence block (table 2.16). Secondly, the fact that we recover essentially the same population history using individuals sampled many dispersal distances apart highlights that simple, discrete population models can be a useful approximation to recent, intra specific histories.

2.12 Appendix

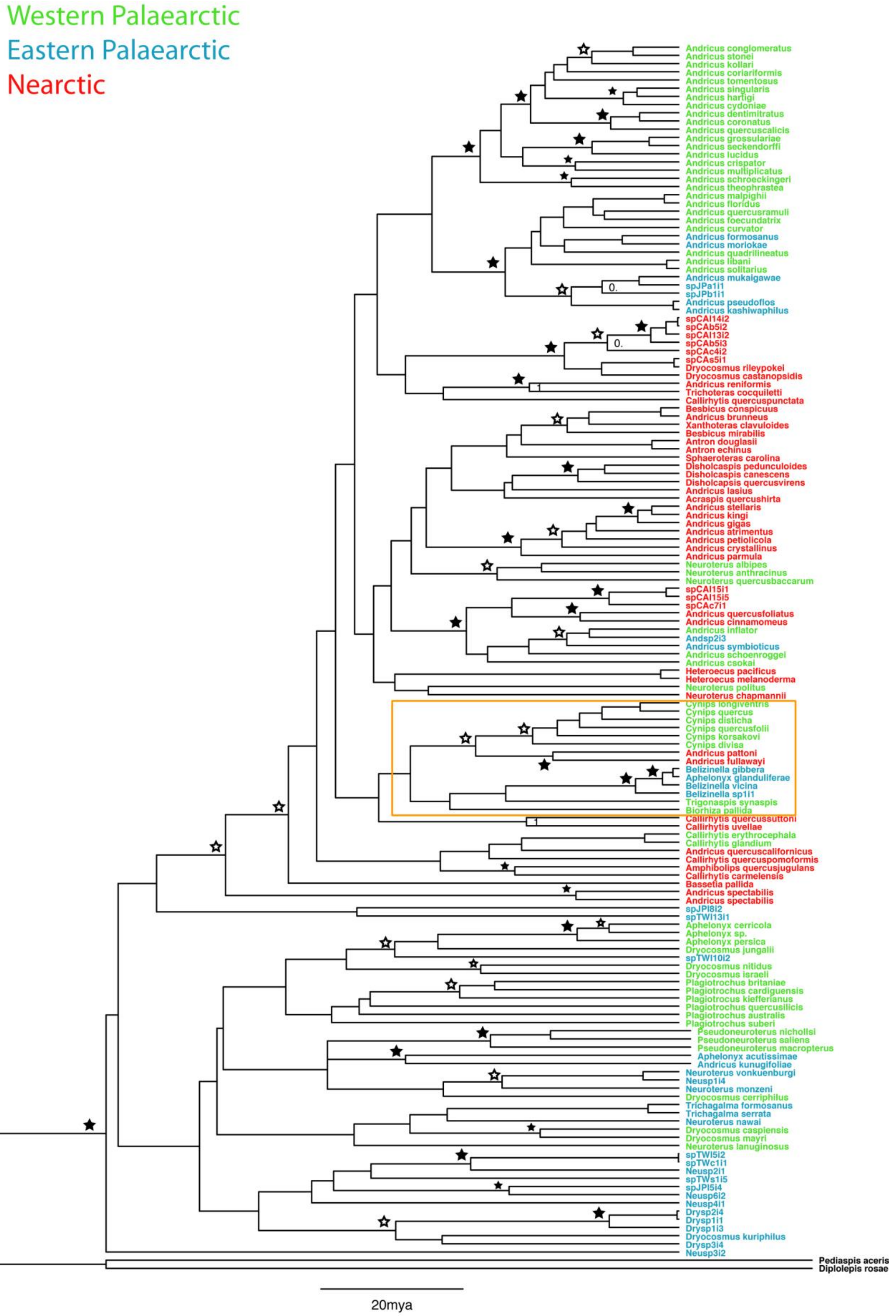

Figure 2.15. Global cytochrome B (*Cytb*) phylogenetic tree with orange box surrounding region blown-up for figure 2.2. Black stars represent posterior probabilities of ≥0.9, hollow stars ≥0.7. Phylogeny courtesy of J. Nicholls.

Model	**Admixture**	**k**		**500b**			**2 kb**	
Panmixia	No	1		-263.2			-948.7	
Polytomy	No	2		-59.3			-111.6	
Topology			***$(W_1;(E_2;C_3))$***	***$(C_1;(E_2;W_3))$***	***$(E_1;(C_2;W_3))$***	***$(W_1;(E_2;C_3))$***	***$(C_1;(E_2;W_3))$***	***$(E_1; (C_2;W_3))$***
2pop.	No	3	-109.6	-197.3	-239.8	-446.4	-601.6	-692.2
3pop.	No	4	-19.2	-59.3,(T_2)	-59.3,(T_2)	-46.2	-111.6,(T_2)	-111.6,(T_2)
A)	2→1	5	-12.8,(T_1)	-5.5	-19.2,(f*)	-16.6,(T_1)	-46.2(f*)	-28.3
B)	3→1	5	**0**	-59.3,(T_1, f*)	-59.3,(T_2, f*)	**0**	-111.6,(T_2,f)	-111.6,(T_2,f*)
C)	2/3→1	5	-12.5	-59.3,(T_{gf},f*)	-59.3,(T_{gf}, T_2),	N/A	-111.6, (T_{gf}, f*)	-111.6,(T_{gf}, f*)
D)	1→2	5	-19.2,(f)	-8.3,(T_1)	-19.2,(f*)	-30.1,(T_1)	-46.2,(f*)	-30.1
E)	1→3	5	-8.2	-59.3,(T_1,T_2)	-59.3,(T_2,f)	-46.2,(f)	-111.6, (T_1, T_2)	-111.9,(T_2,f)
F)	1→2/3	5	-19.2,(f*)	-57.5	-13.4,(T_{gf})	-45.8,	-40.4,(T_{gf})	-100.4

Table 2.15 Support (ΔlnL relative to the best model) for alternative divergence scenarios for three refugial populations of *B. pallida* without admixture or with unidirectional admixture (A–F) for alternative block lengths (500b and 2 kb). All possible scenarios (the labeling of populations (1–3) and of models (A–F) corresponds to Fig. 2) were assessed for the three possible orders of population divergence (columns 1–3). Parameters for which the MLE is 0 (i.e. the model reduces to a simpler nested model) are indicated in brackets (f* refers to complete admixture, i.e. f = 1). The model with highest support is shown in bold.

Dataset	**µ*het.***	***ΔlnL***	***f***	**θ *(Ne)***	***$T_{GF}(t_{GF})$***	***$T_1(t_1)$***	***$T_2(t_2)$***
WaCaE, 500b	no	-6560.4	0.67	0.39(59,200)	0.65(38KY)	0.93(54KY)	2.44(144KY)
WaCaE, 1 kb	no	-9269.3	0.76	0.69(52000)	1.04(54KY)	1.21(63KY)	3.34(173KY)
WaCaE, 2 kb	no	-10713.2	0.69	1.34(52900)	1.04(53KY)	1.23(62KY)	2.73(138KY)

Table 2.16 MLE of parameter estimates under the best-supported model for the *WaCaE* alignment and three different block lengths: 500b, 1 kb, and 2 kb. Both effective population size and divergence time parameters are scaled relative to the rate of coalescence, i.e. in 2Ne generations. Absolute values calibrated using the direct, genome-wide Drosophila mutation rate of Keightley et al., (2009) and assuming two generations per year are given in brackets.

Chapter 3: Dissecting an extended phenotype: candidate genes for gall induction by cynipid gall wasps

3.1 Introduction

Genes involved in gall induction and formation can be identified by comparing gene expression of the three growth stages of gall development in the gall wasp, *Biorhiza pallida*, and its oak host *Quercus robur*. It is hypothesized that early stage gall wasp larval gene expression is focused on inducing the gall, with corresponding high expression of relevant oak genes. To test this hypothesis, I compared diversity and relative levels of gene expression in three key stages of gall growth (early, growth, mature), and specifically compared early stage galls with growth and mature stages. Firstly, transcriptome sequence reads generated using RNA sequencing (RNASeq) were assigned to a species of origin as best as possible bioinformatically, and species-specific *de novo* assemblies created. Then species-assigned reads from each gall stage replicate were aligned to these assemblies and expressed genes quantified. Genes more highly expressed in the early stages were identified by differential expression analysis. Differentially expressed genes were functionally annotated and potential orthologs of larval genes identified in a phylogenetically close oak gall wasp (*Belizinella gibbera*) and a more distantly related rose gall wasp (*Diplolepis spinosa*). Finally, existing hypotheses of gall induction are evaluated in the light of my results and new, specific, hypotheses regarding gall induction are proposed.

3.1.1 The stages of gall induction are the basis of the experimental design

The stages of gall induction are discussed and referenced at length in Chapter 1 and are briefly reiterated here. There are three recognizable stages of gall induction that make excellent sampling points. The stages of gall induction I sampled can be summarised as follows:

Induction: The gall wasp egg is oviposited at a key spot in host meristem of the appropriate organ (Shorthouse et al., 2005). By the time the egg hatches, host plant cells surrounding the egg have de-differentiated to become callus-like cells. The larva lies in a space that will become its chamber surrounded by differentiating gall tissues that are undergoing rapid gall growth. Very simply, an early stage gall is small (of the order of a few mm in diameter) and so are the larvae (ca. 1mm long).

Growth: Nutritive cells form around the larval chamber from gall parenchyma. These cells are the only food source for the larva, but contain high levels of lipids and carbohydrates and have high nitrogen content. The gall tissues continue to grow, vascularize and import nutrients from the rest of the host. The gall is now much larger than the early stage but the larvae have not noticeably grown.

Mature: A layer of sclerenchyma develops between nutritive and outer parenchyma. The larvae feed on the nutritive cells and grow until the layer of sclerenchyma is reached. *B. pallida* galls lignify and take on a brown, paper-like appearance. The larvae pupate and eventually emerge as adults. However, in some species a larval or pupal diapause lasting several years may occur. The gall is the same size as at the growth stage but the larvae are much larger and their chambers enlarged as their lining of nutritive tissue is consumed.

Four replicates were sampled at each stage for sufficient statistical power for accurate identification of differentially expressed genes between stages.

3.1.2 Hypotheses for gall Induction and formation by cynipid gall wasps and their plant hosts

As laid out in chapter 1 there are very few hypotheses for how gall induction occurs. The principle hypotheses are repeated below. In the discussion, these hypotheses are evaluated, and new hypotheses are proposed based on insights from the RNAseq experiment. The hypotheses are split between those that concern the gall wasp, and those that concern the host.

3.1.3 Gall wasp based hypotheses

3.1.3.1 Virus-like-particles

Virus-like-particles (VLPs) passing from galler to host, as proposed by Cornell (1983) provide a potential mechanism for transferring the key factors of induction. Cornell used an argument by analogy with endoparasitoid wasps that utilise VLPs to suppress host immune responses at oviposition (Whitfield & Asgari, 2003). In braconid wasps it has been shown that VLP (bracovirus) packaging proteins are of viral origin while the viral genome they carry is of wasp origin (Bezier et al., 2009).

Cynipid VLP transmission is hypothesized to be under control of the gall wasp larva and not as a maternal effect (Cornell, 1983). Such VLPs would need to be produced continuously by the larva(e) at the early or growth stages to be detected by this experimental design. VLP involvement would be indicated by high expression of viral particle packaging proteins by the larvae, such as capsid proteins of viral origin. Additionally, if a distinct class of gall wasp genes is much more highly expressed than other genes this could indicate expression of VLP genome genes by host cells. This is because host expression is expected to be much higher in general than larval expression, therefore so would any gall wasp genes being expressed in the host and not the larva. The possibility of VLPs being introduced at oviposition with the egg is not addressed here, but has been explored and rejected by S. Cambier (personal communication).

3.1.3.2 Secreted proteins

A high number of secreted proteins are observed in plant-host tissues affected by plant-pathogenic nematodes and gall midges such as the Hessian fly (Mitchum et al., 2012; Stuart et al., 2012). These proteins are characterised by a signal peptide and localisation to the host's extracellular matrix, apoplast, cytoplasm or cell nucleus (Mitchum et al., 2012). The functional effects are poorly understood but some are candidate 'effector' proteins for host manipulation and suppression of immune responses (Mitchum et al., 2012). This status is probably at least in part because many of these proteins do not have orthologs in non-galling nematodes or midges respectively (Mitchum et al., 2012; Stuart et al., 2012). These secreted proteins appear to have evolved with galling in both nematodes and gall midges. In *B. pallida* proteins expressed highly in the early stage encoding a signal peptide are candidates for secretion from the larva(e) to act on host cells. As for galling nematodes and midges, a high proportion of such genes may have evolved within the *Cynipidae* and have no known orthologs.

Highly expressed gall wasp genes encoding secretory peptides are potential candidates for transmission from the larvae to the host, and by extension direct interaction between galler and host. Secreted proteins with high expression in the early stage are candidates for a role specific to induction.

3.1.3.3 Plant cell wall degrading enzymes

The potentially horizontally transferred plant cell wall degrading enzymes (PCWDEs) discovered in several gall wasp genomes including *B. pallida* have potential roles in gall induction (chapter 4). These genes include cellulases, pectin and pectate lyases and rhamnogalacturonate lyases that break down cellulose, pectin and rhamnogalacturonan of the plant cell wall. They are of probable bacterial origin as they are most homologous to plant pathogenic bacterial PCWDEs (chapter 4). But a close relative of the donor species of these genes is not identifiable.

There are two hypotheses concerning these genes in gall wasp larvae, distinguished by their contrasting predictions for gene expression profiles. Firstly, they may have a role in gall induction, for example by remodeling plant tissues. In this hypothesis, high expression in the induction phase relative to the growth and mature stages is predicted. Harper et al. (2009) hypothesise that cell wall loosening from the lysis of pectins, and presumably xyloglucans, could allow a large signalling molecule to permeate cell walls and induce galls. A second hypothesis is that these enzymes play a role in degradation of nutritive cell walls during larval feeding. This hypothesis predicts highest expression of such genes in mature galls, with very low expression early in gall development, when the larva is not feeding.

3.1.4 Plant gene expression based hypotheses

3.1.4.1 Plant hormones

Auxins and cytokinins have been implicated in cynipid gall formation. Kaldewey (1965) and Matsui and Torikata (1970) both identified an auxin, indole-3 acetic acid-like (IAA) response to an *Avena* (Poaceae) coleoptile angle bioassay using larval secretions (Harper et al., 2009). This does not mean the secretions contain IAA, but possible factors that trigger concentration of plant IAA to the secretion. The cytokinin zeatin has been isolated from cynipid larvae and hypothesized as important in gall induction (Ohkawa, 1974; Matsui and Torikata, 1970; Matsui et al., 1975; Harper et al., 2009). It is unknown whether the larva produces or concentrates zeatin from surrounding cells (Harper et al., 2009). Larval extracts have cytokinin-like effects on plant tissues, such as callus induction from stem tissues (Matsui et al., 1975). It appears that morphogens in cynipid larval extractions cause auxin- and cytokinin-like responses (Harper et al., 2009). This suggests that plant hormones are key to gall development, but whether as cause or effect remains unknown. I identify genes that show elevated expression in the early stage of gall development in the host *Q. robur* and whose annotations identify them as having probable roles in plant hormone synthesis/metabolism expression. I discuss possible roles for activity of these genes.

3.1.4.2 The galls-as-seeds hypothesis

Biotin carboxylase carrier protein (BCCP) has been isolated from the nutritive cells of several gall wasps, including *B. pallida* (Harper et al., 2004). BCCP is a protein highly expressed in seeds of *Brassica napus* (Harper et al., 2009; Elborough et al., 1996); although similar studies are lacking for oaks. It is a component of the triacylglycerol lipid synthesis pathway. These lipids are a food source for larval gall wasps, along with other nutrients present at high concentrations in the nutritive cells lining the larval chamber (Harper et al., 2009). The high nutrient content of the gall lining mirrors that found in nutritive cells of developing seeds, leading Harper et al., 2000 to propose the 'galls-as-seeds' hypothesis. Furthermore, associated with seed development are rounds of endoreduplication of nutritive cell chromosomes, and the same is also observed for gall nutritive cells (Harper et al., 2009). Under this hypothesis the inducer manipulates host seed development pathways to form nutritive tissues. To test this, I compared gene expression of BCCP and associated proteins across gall stages. If the galls-as-seeds hypothesis is correct, high expression of these genes is expected in the early and growth gall stages. Additionally, high expression of genes associated with endoreduplication in early stage galls would further support this hypothesis.

3.1.4.3 NOD factors and arabinogalactan proteins

The lipo-chitooligosaccharides, or Nod factors, of the *Rhizobium*-legume nitrogen fixing symbiosis induce nodules on host plants. They activate host plant early nodulin genes (ENOD) that form the nodules in which the symbiotic exchange of nitrogen and nutrients can occur. ENOD genes may represent core genes of plant development that are switched on to create the highly specialised *Rhizobium*-legume nodules. Because of this, ENOD genes are candidates for involvement in gall induction. Specifically, cell wall anchored arabinogalactan proteins (AGPs) are a recognized class of ENOD genes (Cassab, 1986) previously proposed for involvement in gall formation (K. Schönrogge, personal communication). They are proteoglycans consisting of less than 10% protein, the rest being predominantly arabinosyl and galactosyl monosaccharides (Schultz et al., 1998). AGPs are known to have an important role in somatic embryogenesis (van Hengel et al., 2001). They can initiate somatic embryogenesis in wild type cells of the carrot, *Daucus carota* and this ability is enhanced by addition of *D. carota* chitinases (van Hengel et al., 2001). Interestingly, *Rhizobium* Nod factors can rescue somatic embryogenesis similarly to chitinase in *D. carota* temperature sensitive mutants at non-permissive temperatures (De Jong et al., 1993).

Modification of AGP oligosaccharide side chains by secreted gall wasp enzymes is hypothesised to transduce key gall formation signals into host cells (K. Schönrogge, personal communication). AGPs are candidates for a direct interaction between gall wasp larva and host cells. Differentially expressed host genes highly expressed in the early stage will be searched for ENOD genes with a focus on arabinogalactan protein genes. Gall wasp enzymes with the potential to interact with arabinogalactan proteins will be searched for in the differentially and highly expressed genes in the early stage larvae versus the later stages of gall development.

3.2 Sampling, experimental design and challenges:

The criteria for sampling different stages of gall tissues and subsequent sequencing and quality control or reads are described in this section. The methods and results of this chapter begin at the post-sequencing stage (read statistics: table 3.1). They are split into three sections, outlined below.

A. Bioinformatic separation of reads into species. Instead of attempting to dissect larvae from gall tissue, the design called for bioinformatic separation of reads into species post-sequencing. Furthermore, key to a successful experiment was ensuring sufficient depth of *B. pallida* sequencing in the early and growth stages. Too little *B. pallida* mRNA sampling was the principal risk of the experimental design.

B. Statistical analysis of stage specific variation in gene expression. Using the counts of each gene in each replicate, differential expression analysis was performed to identify genes highly expressed in the early stage versus the later stage of gall formation. Two popular differential expression programs, *DESeq* and *EdgeR*, were compared.

C. Identifying roles of differentially expressed genes. Differentially expressed genes were annotated as well as possible using *BLAST* comparison to non-redundant nucleotide and protein sequence databases, *InterProScan* and subsequently *BLAST2GO* and GO term enrichment. These annotations were used as this basis for evaluating hypotheses and generating new ones in the discussion.

3.2.1 Sample selection for transcriptomics of gall induction

3.2.2 A Technology driven approach to experimental design

By sequencing whole galls, or segments representative of the whole, the high- throughput of Illumina technology was leveraged. RNA derived from the oak host was deliberately 'over-sequenced' to adequately sample the much lower proportion of gall wasp larval expression. In doing so, gall tissues were treated as one system containing multiple actors. This approach is well suited to exploring an extended phenotype as the host, *Q. robur,* expression became integral to the design.

The sequencing design called for four replicates at each stage giving twelve samples in total. This number of replicates represented a trade-off between enough samples for robust statistical inference and available resources. A protocol was developed for collecting gall tissues in the field using the preservative RNAlater (Ambion). It was designed to minimise changes and degradation in RNA expression due to removal of galls from the host. Sampling was carried out at several sites around the town of Blandford Forum, Dorset (50°51′43″N, 2°9′45.5″W) in Southern England. Under this design, the difficulty of dissecting larvae form gall tissue was shifted onto the bioinformatic problem of separating reads by species *post hoc.*

B. pallida is a multilocular gall meaning there are multiple larvae developing in a single gall. This causes unknown variation in the number of larvae between galls but is expected to negate the effect of parasitoids and inquilines on expression analyses. Because of the replicated design, parasitoid and inquiline expression is not expected to confound *B. pallida* expression. It also increases the proportion of gall wasp derived tissues in early stage galls compared to many unilocular gall producing gall wasps.

3.2.3 Gall collecting and dissection in the field

Gall development is a continuum but by using certain criteria galls of approximately the same stage were identified. For early galls a diameter of <0.5cm, but smaller if possible were collected (figure 3.1, early stage galls selected for sequencing). Growth stage galls were identified by their much larger size, but when dissection the larva(e) remained small. Where larvae were not visible the larval chambers could still be identified (figure 3.2, growth stage galls selected for sequencing). Growth stage tissues are moist, rapidly oxidise on exposure and vascularisation is occurring but not yet complete (figure 3.3, mature stage galls selected for sequencing). Mature galls had large growing larvae that were active and could be observed feeding using their pincer-like mandibles. At this stage larval chambers are enlarged hollows as feeding depletes nutritive tissues. *B. pallida* gall tissues also lignify and dry out becoming much harder to slice open. Externally they take on the texture and shade of a brown paper bag. Growth and mature stage galls were sliced open with a razor blade the gall in half and an internal picture taken to use the larva(e) and chambers as stage diagnostics. This was done quickly to minimise changes in gene expression and with latex gloves to avoid contamination.

Small galls were rapidly sliced into halves or quarters and immediately immersed in RNAlater (Ambion). Even when the gall was tiny it was sliced in half to allow RNAlater (Ambion) to rapidly permeate inner tissues. Larger galls were first cut in half and thin segments sliced from the centre out, akin to orange segments, and placed in RNAlater (Ambion) (figure 3.4). Each segment contained the inner and outer tissue sampled in proportions representative of the complete gall except for the epidermis.

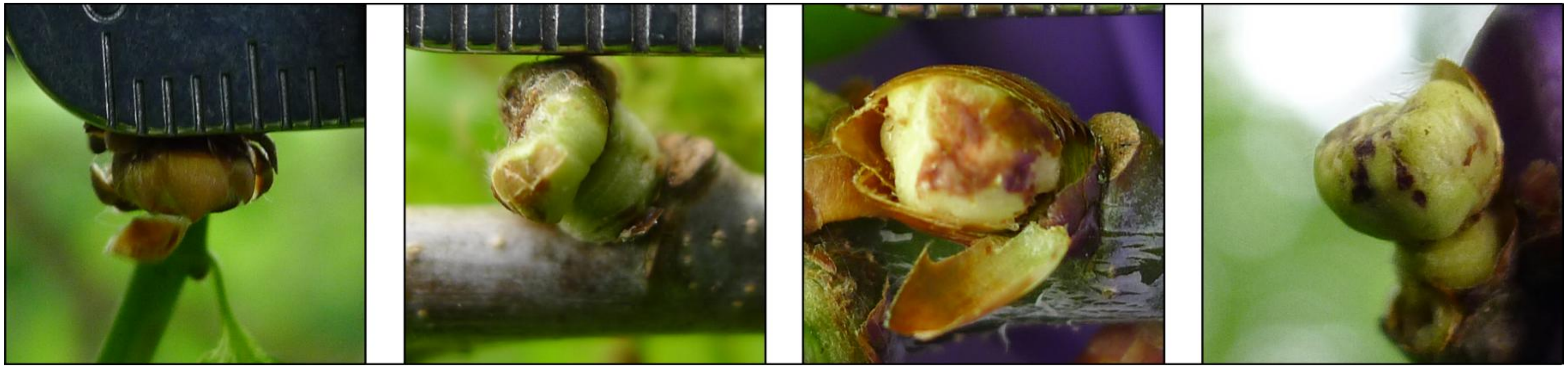

Figure 3.1. Early stage galls chosen for sequencing, collected April 2011.

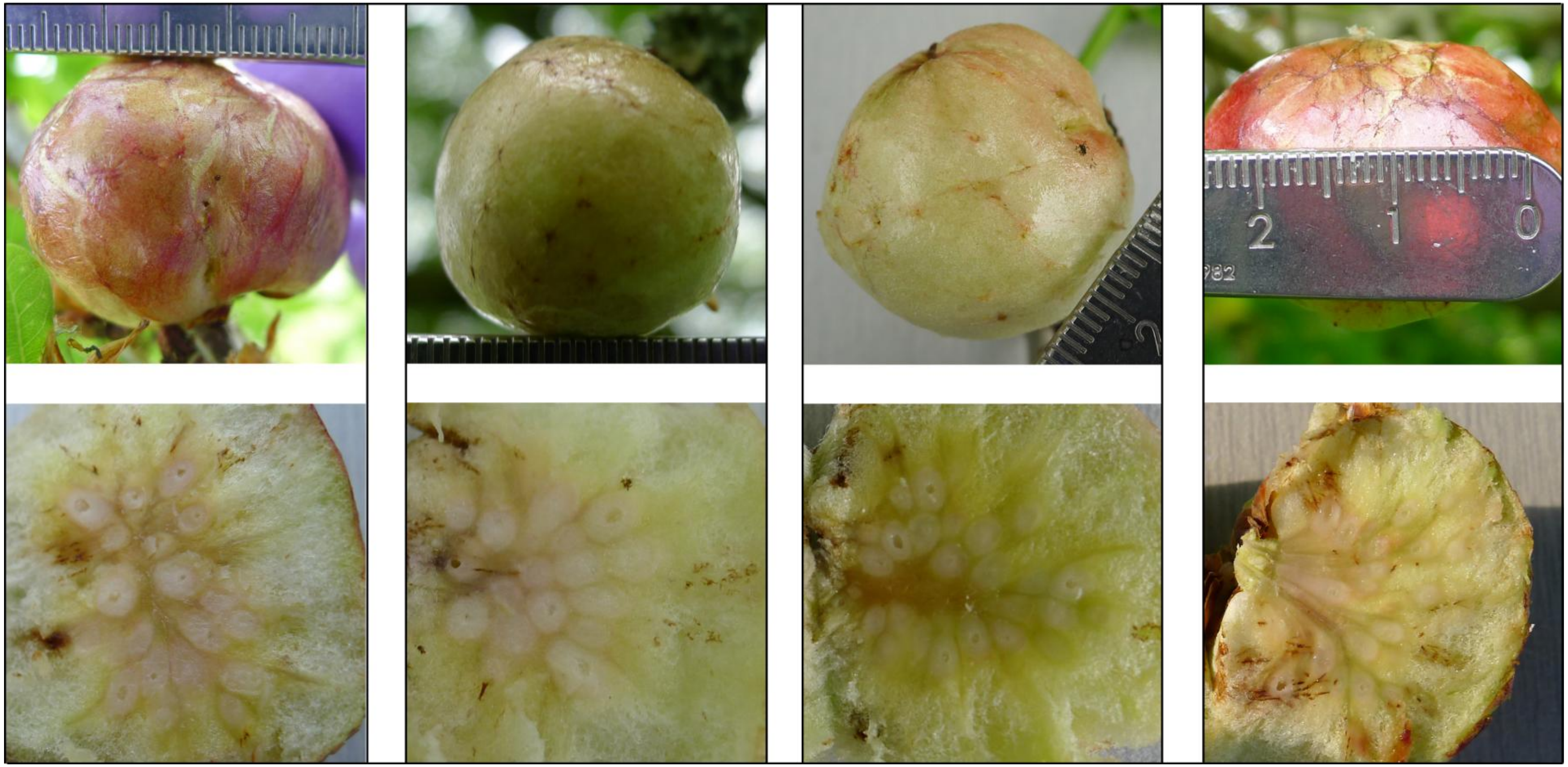

Figure 3.2. Growth stage galls chosen for sequencing, collected April 2011. The bottom row is an internal view showing small larvae with pronounced chambers and vascularisation of tissues.

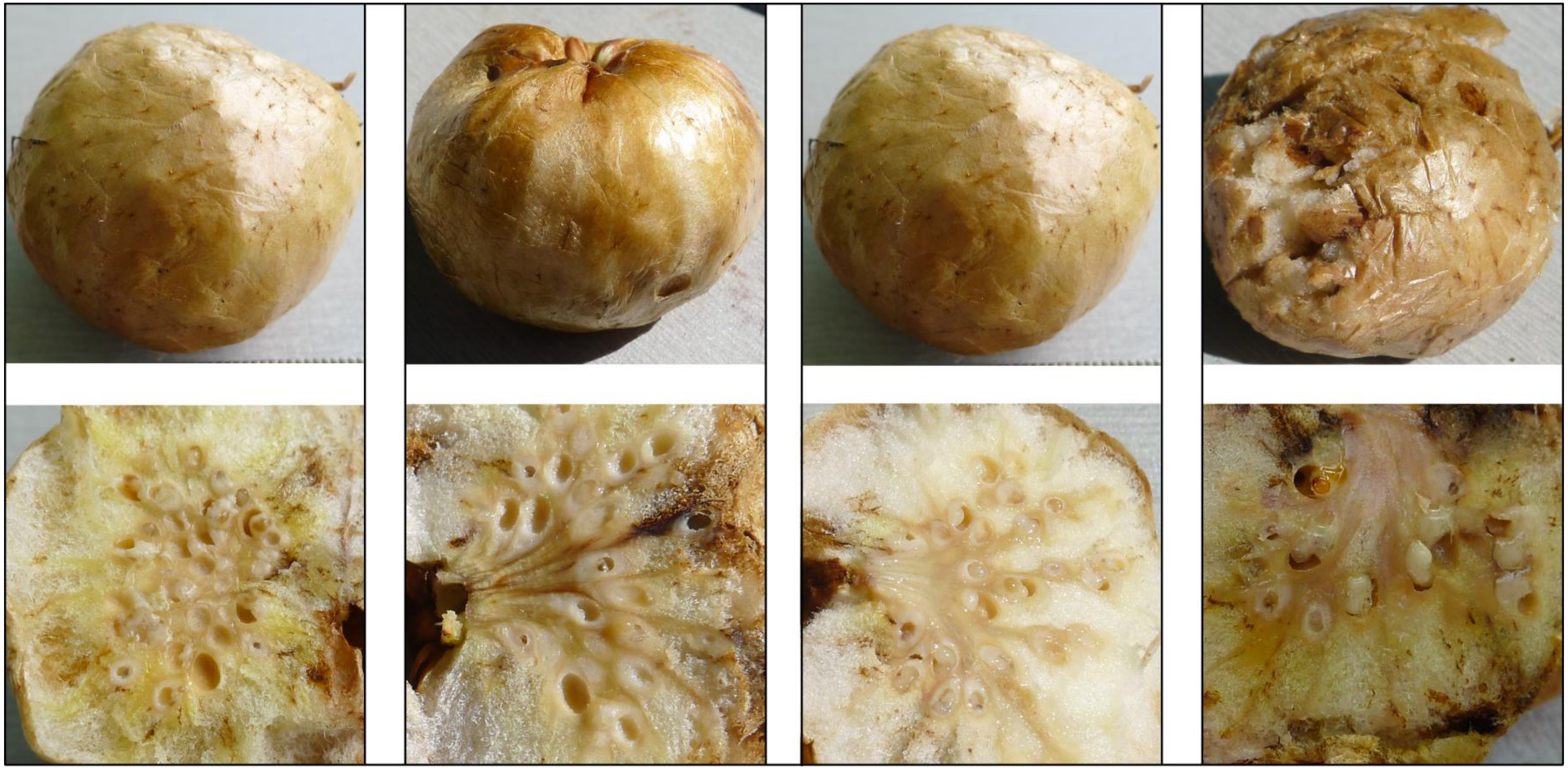

Figure 3.3. Mature stage galls chosen for sequencing, collected April 2011. The bottom row is an internal view showing large feeding larvae with pronounced hollowed chambers and lignification of tissues.

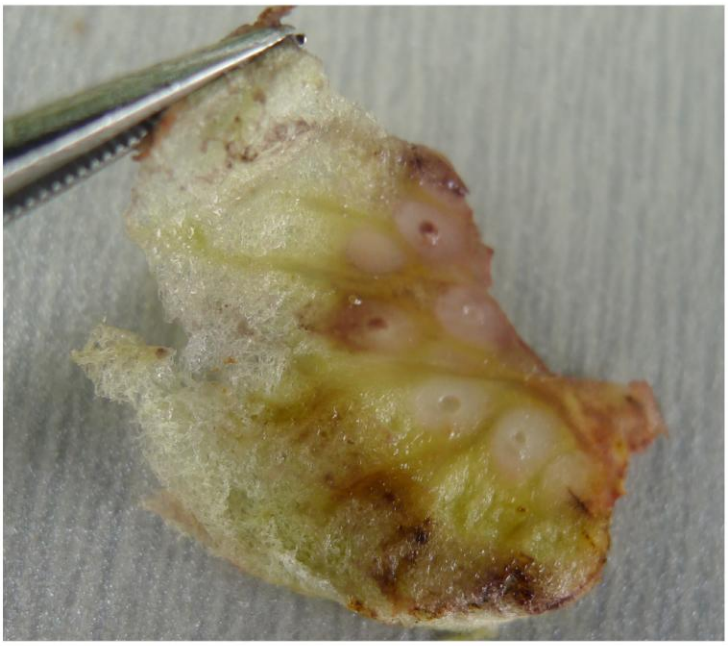

Figure 3.4. A slice of gall tissue ready to be immersed in RNAlater (Ambion) for extraction. Gall tissues oxidise rapidly on exposure to air highlighting the larval chambers.

3.2.4 Extracting RNA

Extractions for each of the twelve RNASeq experiment samples followed that of the RNEasy plant mini kit extraction protocol for plants and fungi (Qiagen) with modifications outlined here. The frozen sample tubes were weighed and then thawed at room temperature and the gall tissue placed into a mortar and pestle pre-cooled and filled with liquid nitrogen (LN). The tubes were then re-weighed without the samples for an approximate gall tissue weight. For smaller galls with diameters <0.5cm the whole gall was extracted on two extraction columns. For much larger growth and mature galls four segments were combined per gall. Multiple segments were combined to balance the effect of segments sliced poorly in the field. The sample was ground in LN until a fine powder was left and there was no resistance to the pestle. The amount of lysis buffer used depended on the number of columns used for the RNEasy (Qiagen) extractions. All extractions for a single gall/sample were combined into one tube and a small amount (15-50 micro litres) aliquoted into another Eppendorf tube for quality control. The rest was immediately frozen at -80°c until ready for sequencing.

3.2.5 Quality controlling RNA samples post extraction

Samples were assessed for RNA purity using the 260/280 and 260/230 ratios measured on a NanoDrop spectrophotometer (Thermo Scientific). The GenePool (Edinburgh) required 260/280 ratios >1.85 for RNA sequencing. The ideal 260/230 ratio is 2, however many samples from larger galls had lower ratios of approximately 1.3 possibly due to carry-over of carbohydrates and other impurities caused by column overloading during extraction. Samples with low 260/230 ratios did not cause a problem during library preparation and sequencing. One sample (270C) required additional purification; this was done using the appropriate RNEasy (Qiagen) plant mini kit protocol. All sample extractions yielded high concentrations of RNA.

To determine if the RNA had good integrity an Agilent 2100 Bioanalyzer (Agilent Technologies) total RNA nano trace was run on the samples. Assessing the ratio of the 28s to 18s ribosomal RNA peaks is a proxy for total RNA integrity (figure 3.5). All samples submitted for sequencing had no visible degradation on the Bioanalyzer traces validating the RNAlater (Ambion) based field collection protocol.

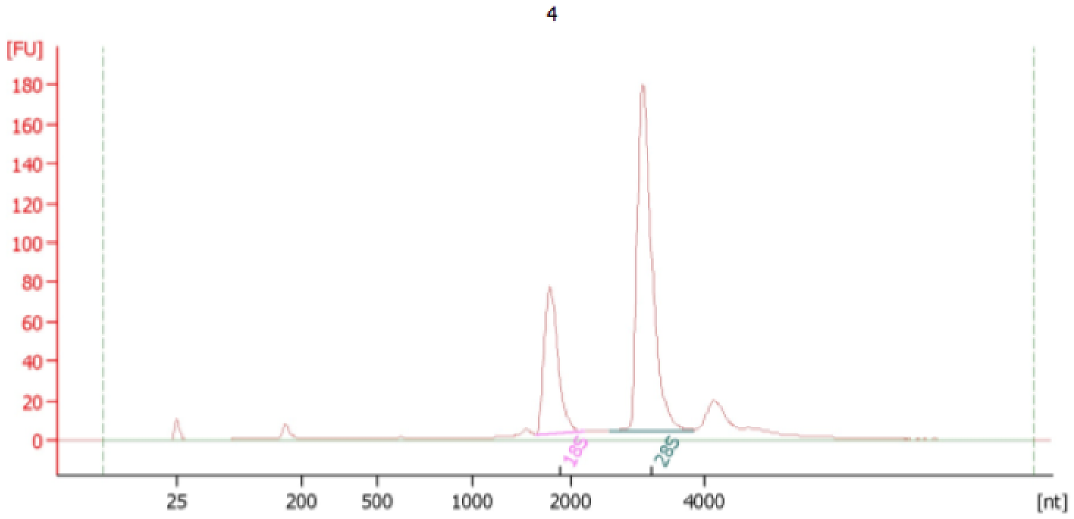

Figure 3.5. Agilent 2100 Bioanalyzer (Agilent Technologies) trace of total RNA for sample 4, x-axis is sixe of fragment in nucleotides and y-axis is fluorescence units. The lack of degradation of the 18s and 36s rRNA peak or at the baseline confirms that RNA is of sufficient quality for sequencing.

3.2.6 Reverse Transcriptase Polymerase Chain Reaction

Prior to sequencing it was not known if gall wasp RNA was at detectable levels in the sampled tissues, especially for early stage extractions. Reverse transcriptase polymerase chain reaction (RT-PCR) using a pair of exon-primed intron-crossing (EPIC) primers was used to establish if there was detectable gall wasp expression in the extraction. A positive control of *B. pallida* DNA was used to confirm cDNA was amplified and not residual genomic DNA. EPIC loci will amplify different sized bands depending on if the amplicon contains an intron or does not. The EPIC primers used were designed from the *Nasonia vitripennis* genome and ESTs and tested across a range of gall wasps and chalcid parasitoids (Lohse et al., 2010; further refined for cynipids by James Nicholls, personal communication). Primers for the genes *Receptor for Activated C Kinase 1 (RACK1)* and *Ribosomal Protein L37 Rpl37* worked best and were specific to *B. pallida.* All sequenced samples were positive for these two loci. Sanger sequencing of amplicons confirmed cDNA to be *B. pallida* by comparison to the sequence of the positive control.

3.2.7 Library preparation

All 12 samples were prepared as 100 base pair (bp) paired-end TruSeq libraries (http://www.Illumina.com/products/truseq_rna_sample_prep_kit_v2.ilmn) by The GenePool and multiplexed together for sequencing on an Illumina HiSeq 2000. Multiplexing the samples minimises inter-lane technical effects of sequencing as all samples are affected to the same extent. For a second lane of sequencing the 8 early and growth stage libraries were multiplexed together. This was to increase the number of gall wasp reads sequenced in the early stages of gall developmental when the proportion of gall wasp derived tissue is lowest.

3.2.8 Sequencing results

Quality filtering was carried as for genomic DNA in chaper 2 and before and after filtering results is given in table 3.1. One difference between transcriptome and genome data is overrepresentation of certain oligomers at the beginning of RNASeq reads. This is because of biased random hexamer priming during the reverse transcriptase stage of library preparation, (Hansen et al., 2010) and is not uncommon with RNAseq data.

Sample Name	Stage	Read Count (millions)	Bases (Gb)	Filtered Pairs (millions)	Singles (millions)	Filtered Bases (Gb)
1	Early	51.4	10.32	48.0	2.6	9.55
4	Early	39.4	7.88	37.1	1.8	7.40
8	Early	28.7	5.74	27.1	1.4	5.39
211	Early	33.9	6.78	31.6	1.8	6.29
127	Growth	45.5	9.09	41.4	3.2	8.23
148	Growth	37.9	7.58	35.1	2.2	6.98
182	Growth	17.1	3.43	45.8	0.6	3.28
224	Growth	32.0	6.40	29.5	2.0	5.85
234	Mature	11.4	2.28	10.9	0.4	2.18
252	Mature	8.7	1.75	8.3	0.3	1.68
270C	Mature	13.1	2.63	12.6	0.4	2.53
281	Mature	14.2	2.85	13.6	0.5	2.74

Table 3.1. Combined Illumina read statistics for raw and filtered data for each transcriptome sample.

3.3 Part A) Bioinformatic separation of reads into species: has enough gall wasp mRNA been sequenced?

3.3.1 Estimating the insert size of paired-end data

Estimating the insert size of paired-end data is important as overlapping reads can be combined to create super-reads for less memory demanding assemblies. It is also useful for recognizing irregularities resulting from library preparation. RNA sequencing library preparation results in an expected fragment size of 190 base pairs (bp) (GenePool, personal communication). At this fragment size an overlap of 10 bp will occur between two 100 bp paired-end reads on average; the distance the pairs overlap will increase with shorter fragment sizes. To estimate the insert size, and standard deviation of the insert size a single-ended *CLC bio de novo* assembly (v4.0.3, http://www.clcbio.com/products/clc-assembly-cell/) of the data was made and the reads mapped back to the data following the section '*How to map reads to an assembly to get insert-size and coverage information using CLC'* of the assemblage protocol (available at: https://github.com/sujaikumar/assemblage) (Kumar, 2012). The reads from each replicate were separately mapped to the single ended assembly using *CLC reference assembly* (V4.0.3). Then a script that calls *CLC assembly info* (v.4.0.3) outputs files containing coverage and insert size estimates for each pair of reads. The output files are used for the next section of the assemblage protocol '*How to make a plot of insert sizes for each library.*' The plots revealed that average insert size for all libraries was 140-145 bp with tight, very slightly right-skewed, distributions from 50-350 bp. An average fragment length of 140 bp means an average overlap of approximately 60 bp between reads of each paired-end fragment.

3.3.2 Combining overlapping reads using *FLASH*

Prior to assembly, overlapping paired-end reads were overlapped with *FLASH* (Magoč & Salzberg, 2011) to create super-reads. *FLASH* (Fast Length Adjustment of Short Reads) overlapped super-reads can be used to improve assemblies and reduce the memory requirements of the assembler (Magoč & Salzberg, 2011). Firstly, each read pair is aligned so that they overlap completely and the overlap length is calculated. A score for the overlap is given by the ratio of mismatches to overlap length. Aligning and scoring is repeated at every possible alignment length until a minimum overlap threshold is reached. The best alignment is the one with the lowest ratio and is chosen. When two overlaps have an equal score the one with lowest average quality score of mismatches decides the best alignment. Finally, the best alignment score must be lower than a mismatch threshold for an overlap to be reported (Magoč & Salzberg, 2011). *FLASH* was run on all of the paired-end reads across replicates with a minimum overlap of 10 bp. A histogram of overlapping read length output by *FLASH* corroborated the average insert size of 140-145 bp for each replicate. Table 3.2 gives the read numbers and number of bases after running *FLASH* for all replicates combined. Read pairs connected into super-reads constitute most of the data as expected from the average insert size. This dataset was assembled using *Trinity* (Grabherr et al., 2011).

The de Bruijn graph-based *Trinity* assembler was chosen for its excellent performance, user support and integration with downstream expression analyses (http://*Trinity*rnaseq.sourceforge.net/). It is a memory intensive assembler; therefore a 512Gb RAM computer was used to create assemblies. De Bruijn graph based transcriptome assembly differs from genome assembly as discussed in chapter 2 because many disconnected graphs occur, each representing a different locus. In contrast, the goal in genome assembly is to generate a minimal number of graphs corresponding to chromosomes. The program is modular, consisting of three sequential assembly steps: Inchworm, Chrysalis and Butterfly.

Library	Number of reads	Bases (Gb)
All reads pair 1	85 091 456	8.1
All reads pair 2	85 091 456	8.1
All reads overlapped	226 574 604	33.2
All reads QC singles	16 970 221	1.5
Total	413 727 737	50.9

Table 3.2. Numbers of reads remaining after FLASH overlapping of all RNAseq reads, Gb = gigabases. All reads QC singles refers to single reads in which the pair failed quality control (QC) (see Chapter 2).

3.3.3 The combined reads assembly

All reads across all replicates were combined for assembly of a reference transcriptome in *Trinity* (table 3.3). The quality of this assembly is not essential to the differential expression analysis; the transcripts just need to be contiguous enough to identify their origin using *BLAST*. The assembly has an N50 of 1468. Although N50 is not the best metric for use with a transcriptome as there is a range of expected transcript lengths, and long transcripts representing multiple isoforms of the same gene may also bias the N50 artificially upwards. The very high number of components and transcripts is likely to reflect the presence of two transcriptomes in the data and the ‘verbosity’ of the *Trinity* assembler.

N50	Number of transcripts	Number of components	Number of transcripts in N50	Maximum transcript size	Number of bases in transcripts
1468	351 215	231 436	53 017	32 024	296 033 427

Table 3.3. Assembly metrics for the all reads transcriptome assembly.

3.3.4 Assigning species to transcripts

My strategy for obtaining a higher quality *B. pallida* assembly was to separate the reads into bins representing each oak and gall wasp, remove reads derived from contaminants, and re-assemble each bin of reads separately.

To do this the transcripts were first assigned a probable taxonomic origin using several custom *BLAST* databases. A combined *BLAST* database was created using my *B. pallida* genome assembly, a recently generated *Q. robur* genome assembly (P. Fuentes, personal communication) (table 3.4) and publically available *Q. robur* ESTs (Ueno et al., 2010) (https://w3.pierroton.inra.fr/QuercusPortal/index.php?p=cgen). Additional *BLAST* databases for the Arthropoda, Plants, Fungi, Bacteria, Viruses, Mammalia, and *Castanea* chloroplasts for both protein and nucleotide sequences were also created. The combined transcriptome was then *BLAST* searched against all the custom databases and the results combined into a single output file. The Mammalia database was for identifying contamination of the dataset with mouse and human RNAs. The *Castanea mollissima* (Jansen et al., 2011) chloroplast database was created from the *Castanea mollissima* chloroplast genome to identify large contigs corresponding to chloroplast DNA. These are derived from *Q. robur* chloroplast mRNA and potentially DNA that escaped DNA digestion during extraction. *C. mollissima* is the phylogenetically closest species to *Q. robur* with an available chloroplast genome sequence. A transcript was assigned to one of several categories according to the taxonomic origin of the top-scoring transcript. A transcript was assigned to the Arthropoda category if its top hit was to either the *B. pallida* genome or the Arthropoda database, and so for all taxonomic databases. No e-value threshold was applied for the *BLAST*s against the *B. pallida* and *Q. robur* genomes, but a threshold of 1 x 10^{-5} was applied for the other databases. This was to assign as many contigs as possible to the Plants and Arthropoda categories based only on the best-hit criterion. All *BLASTs* used low complexity filtering to avoid spurious matches.

Species	Max. contig length	Number of contigs	Total bases in contigs	N50 for contigs	Contigs in N50	GC contigs	Number 'N's
Biorhiza pallida	38 791	1 163 314	805 102 378	1 075	193 792	32.9	4 203 182
Quercus robur	90 459	715 072	652 949 554	1 615	98 301	35.5	28 530 837
Q. robur ESTs	6 795	218 977	99 131 312	505	53 317	40.4	22 978

Table 3.4. Assembly metrics for draft gall wasp and oak genomes and assembled ESTs (Ueno et al., 2010) used for as references for read aligning. Number of 'N's refers to nucleotides in the database where the assembler was not able to determine the correct nucleotide but knew the position exi sts in the genome by read context.

The *B. pallida* and *Q. robur* genome assemblies are both low-coverage drafts (table 3.4), and are therefore incomplete. Because of this they may be missing genes that are expressed in the respective species transcriptomes. Therefore transcripts with best-scoring matches to the custom Arthropoda database were combined with those top-scoring to the *B. pallida* genome into one Arthropoda category. *Q. robur* genome/EST plus Plant database top hits were also combined in this way into a Plants category. Arthropoda, a broad taxonomic label, was used because parasitoid and inquiline derived sequences were implicitly assigned to this group. As expected, the majority of transcripts, 62.3%, are derived from Plants/*Q. robur* (table 3.5). The next largest category is the gall wasp proxy Arthropoda category at 31.7%. Together Plant and Arthropod categories constitute 94% of the transcripts, while the Fungi category is next most common at 4.8%. The Fungi percentage indicates an infection of plant tissues in two replicates and is controlled for during expression analysis. The viruses are predominantly single-strand RNA plant viruses. The Mammalia transcripts are 70% *Mus musculus* at very high identities and are probably laboratory or reagent contaminants.

Database	Number of transcripts	% of transcripts
Arthropoda	111 387	31.7
Plants	218 692	62.3
Fungi	16 949	4.8
Bacteria	414	0.1
Viruses	154	0.0
***Castanea chloroplast* genome**	1392	0.4
Mammals	1535	0.4
Other *BLAST* hits	447	0.1
Unassigned contigs	245	0.1
Total	351 215	100

Table 3.5. Taxonomic origins assigned to contigs from the all read transcriptome assembly.

3.3.5 Identifying and filtering non-coding and organelle derived transcripts

Concurrent with identifying the taxonomic origin of the transcripts, rRNA, mitochondrial and chloroplast gene transcripts were identified. It is prudent to remove these genes as they are often very highly expressed and may skew the normalization of replicates (see section 3.4.2) as performed by differential expression analysis programs. This is because these genes occur in multiple copies per cell in the nuclear or organelle genomes, unlike the single-copy nuclear genes that are most likely to be of interest for this experiment.

To identify the large (LSU) and small (SSU) ribosomal subunits, high quality Bacteria and Eukaryote sequences were downloaded from the *Silva* RNA database (http://www.arb-silva.de/) and searched against the combined transcriptome assembly using the *LAST* aligner (Kielbasa et al., 2011); 297 transcripts were identified as either LSU or SSU in this way. Other non-coding RNAs were identified by an adapted *BLAST* search against the Rfam database (Gardner et al., 2009). Rfam is a database of annotated RNA sequence families such as transfer RNAs (tRNAs) and micro RNAs (miRNAs). The *Rfam_scan.pl* (Gardner et al., 2009) script was used to map annotations to *BLAST* results, and a total of 1072 transcripts were identified as non-coding RNAs. Finally, 178 transcripts encoding tRNAs were identified using *tRNAscan* (Lowe and Eddy, 1997). However, after cross checking the tRNA predicted transcripts with *BLAST* results only 87 were assigned for removal as these transcripts were chloroplast or mitochondrion encoded. The retained transcripts were those apparently overlapping with nuclear genes.

3.3.6 Aligning reads to transcripts for removal

A list of transcripts for removal was created from the non-coding genes and those identified as of chloroplast, mitochondrial and mammalian origin; this amounted to 4191 transcripts. Reads were aligned to the 4191 transcripts using *bowtie2* default parameters (Langmead & Salzberg, 2012) and removed from the analysis; Remove reads totalled 68 430 077 paired-end fragments and 5 917 057 single-end reads. Replicates 127 (early), 224 (growth) and 234 (mature) had particularly high expression of the removed transcripts as evidenced by the lower remaining percentage of bases after filtering at 58.7%, 39.5%, and 55.2% (table 3.6). This may reflect overloading of the columns with gall tissue at extraction causing sub-optimal DNA degradation or polyA tail selection worked poorly for these samples.

Sample	Stage	Bases Gb	Remaining bases Gb	Percentage remaining
1	Early	9.55	9.15	95.7
4	Early	7.40	7.10	95.3
8	Early	5.39	5.18	96.0
211	Early	6.29	5.76	91.6
127	Growth	8.23	4.83	58.7
148	Growth	6.98	5.76	82.5
182	Growth	3.28	2.55	78.0
224	Growth	5.85	2.31	39.5
234	Mature	2.18	1.20	55.2
252	Mature	1.68	1.19	70.9
270C	Mature	2.53	2.13	84.2
281	Mature	2.74	1.80	65.7

Table 3.6. Percentage of bases remaining for each replicate after removal of unwanted sequences.

3.3.7 Aligning reads to the taxonomically categorized transcripts and genomic resources

The filtered reads were now ready to align to a reference of taxonomically categorized transcripts, *Q. robur* and *B. pallida* genomes, and the *Q. robur* ESTs. The reads aligned were not those that had been combined with *FLASH* for the combined assembly but the original pairs and singles. The transcripts that were filtered above were removed from their species categories. The aligner *GSNAP* (Wu and Nacu, 2010) was chosen, as it is splice-aware and therefore able to accurately align RNAseq reads to genomic contigs. *GSNAP* can align reads that bridge exons; consequently these reads need to be split when aligned to a genome (Wu and Nacu, 2010). The highest scoring mapping for each read was used to assign species of origin.

Where multiple equal highest scoring mappings occurred the read was assigned to each applicable category. This led to an inflation of the total reads assigned to each category versus true total reads. With the available resources it was not possible to further categorise these multi-mapping reads. The reads were kept so the maximum amount of reads could be used to create the best possible *B. pallida* transcriptome.

For most replicates this inflation percentage (table 3.7) was a minor percentage of the total reads. Indeed, for sample 270C there were fewer reads after *GSNAP* mapping to the references (-1.07%). However, three replicates, 127, 224 and 243 had greater mapping inflation at 4.45%, 7.14% and 3.48% respectively. These are the same replicates highlighted as having the most reads aligning to unwanted transcripts. That the same replicates are flagged as problematic suggests the same issue is affecting them.

3.3.8 Gall wasp sampling depth, numbers of reads, and dynamic range

The principal risk of the RNAseq experiment was insufficient sampling of gall wasp transcripts. A huge difference in expression proportion between gall wasp and host in favour of the host was expected. This was based on the proportion of gall tissue derived from each species and was particularly true for growth stage galls with their tiny larvae but large gall size. The presence of gall wasp RNA had been detected in each replicate by reverse transcriptase PCR, but this did not indicate whether sufficient dynamic range of gall wasp RNA would be captured for a useful differential expression analysis. The dynamic range is the ratio between maximum and minimum gene expression level, and RNAseq can detect a >9,000-fold difference (Wang et al., 2009). In terms of dynamic range, insufficient sampling of *B. pallida* RNA would drastically reduce the fold difference between minimally and maximally expressed transcripts as only the most highly expressed transcripts are captured. Additionally, a *de novo* assembly recovering many complete transcripts representative of a tissue transcriptome requires a minimum of 20 million reads (Francis et al., 2013); for whole organisms the recommendation is 30 million reads. In practice, these figures will vary by species and tissue and are only guidelines.

The total number of gall wasp paired-end and single-end reads sequenced is 43 302 057 (table 3.8). This is above the minimum recommended for transcriptome analysis (Francis et al., 2013). However, However many of these are pairs, as a result approximately 20 000 000 independent fragments have been sequenced (table 3.8). The lowest number of gall wasp reads in any replicate is 1 692 111 for early stage replicate 8, while early stage replicates 4 and 211 also have low overall counts. These are low counts and deeper sequencing would be preferable, but the highest expressed genes will be captured. As these genes are of most interest differential expression analysis is viable.

3.3.9 Approximate percentages of gall wasp, oak and other species

There is an obvious trend in the ratio of gall wasp to oak with developmental stage (table 3.9) when using Arthropoda assigned reads as a proxy for gall wasp. The early stage galls have overwhelming oak expression, and the percentage gall wasp ranges from 2.58-4.12%. Sample 211 has a Fungi proportion of 3.54%, very close to the gall wasp proportion at 4.06%. This reflects a fungal infection of the plant tissue that needs to be controlled in expression analyses. A fungal infection will result in fungal-specific response by the host and oak that may confound gall induction associated expression.

	1	4	8	211	127	148	182	224	234	252	270C	281
Number of mappings	93570526	71759215	52664528	59067477	51493017	59441197	26059225	25159422	12605455	12282898	21356161	18250667
Number of reads	93328764	71590117	52578351	58844157	49298390	58795708	25643921	23482195	12181300	12023150	21587793	18172550
Discrepancy	241762	169098	86177	223320	2194627	645489	415304	1677227	424155	259748	-231632	78117
% Mapping inflation	0.26	0.24	0.16	0.38	4.45	1.10	1.62	7.14	3.48	2.16	-1.07	0.43

Table 3.7. Discrepancy between the number of GSNAP mappings and the number of reads for individual replicates.

Origin	1	4	8	211	127	148	182	224	234	252	270C	281
Arthropod	3855332	1851710	1692111	2396419	3605113	2884112	2087645	3591806	6186109	4562420	6997659	3591621
Plant	89705901	69904305	50961803	54548292	47884940	56547111	23951601	21558828	6409592	7685409	14087999	14292587
Fungi	4001	347	8253	2089980	561	717	313	1930	806	4131	201740	2983
Virus	4742	2534	2040	17592	1642	8497	19412	5904	8469	29846	61361	361488
Bacteria	215	94	112	12573	681	500	193	866	169	1010	6298	322
Unassigned	335	225	209	2621	80	260	61	88	310	82	1104	1666
Total	93570526	71759215	52664528	59067477	51493017	59441197	26059225	25159422	12605455	12282898	21356161	18250667

Table 3.8. Total numbers of reads assigned to each taxonomic category for individual replicates. Column headers refer to sample identification.

Table 3.9. Percentage of reads assigned to each taxonomic category across replicates.

Stage	Early				Growth				Mature			
Sample	**1**	**4**	**8**	**211**	**127**	**148**	**182**	**224**	**234**	**252**	**270C**	**281**
Arthropod	4.12	2.58	3.21	4.06	7.00	4.85	8.01	14.28	49.07	37.14	32.77	19.68
Plant	95.87	97.42	96.77	92.35	92.99	95.13	91.91	85.69	50.85	62.57	65.97	78.31
Fungi	0.00	0.00	0.02	3.54	0.00	0.00	0.00	0.01	0.01	0.03	0.94	0.02
Viruses	0.01	0.00	0.00	0.03	0.00	0.01	0.07	0.02	0.07	0.24	0.29	1.98

The percentages of gall wasp reads are more variable for growth stage galls at 4.85-14.28%, but all values are higher than the early stage. This probably reflects greater variance in the precise stage of the growth gall tissues sampled. The number of larvae in the gall potentially confounds staging growth stage *B. pallida* galls, as the size may reflect maturity or the number of inhabitants. However, the percentage of gall expression in the growth stages was higher than expected. At this stage the ratio of plant to gall wasp tissue in the gall is at its largest, therefore a lower percentage of gall wasp was expected compared to the early stage.

The mature stage contrasts very strongly with both early and growth stages. Replicate 281 has the least gall wasp expression at 19.68%. The other three replicates range from 32.77-49.07%. These percentages are much greater than the other two gall stages and reflect broad changes in expression of both oak and gall wasp. Reads of replicate 270C are 0.94% fungal derived this is not large but in terms of actual read numbers is significant. Replicates 211 and 270C were combined to control for fungal specific expression in part B), differential expression analyses, below. Replicate 281 also has 1.98% viral expression, but as the focus was on identifying genes of interest in the early and growth stages this was not controlled for.

3.3.10 Creating species-specific assemblies from the separated reads

Reads from each replicate for the Arthropod and Plants data were combined to create species-specific assemblies using *Trinity*. Before assembling the Plants data it was digitally normalized because of the large number of reads.

3.3.11 Plant data normalization and combining overlapping reads

The Plant data was digitally normalized using *Trinity*’s included *in silico* normalization scripts. This was to improve assembly run times and reduce memory requirements. Normalisation refers to equalizing the coverage data around a coverage value by removing many of the reads containing very high frequency k-mers; *Trinity*'s normalization scripts were run with an average coverage of 30-fold and k-mer length of 25 for the Plants data. This read normalization was only performed for reads used in the Plants assembly. For quantification of the transcripts (section 3.4.2) the non-normalised reads were used to maintain the correct ratios of expression between transcripts. The paired and single Plants reads were normalized separately as it was not possible to perform a combined normalization. The Plants dataset was dramatically reduced by normalization; approximately 5% of the plants data was retained (table 3.10).

Plants reads	**Number of Reads**	**Number of bases (Gb)**	**Post-normalization**	**Post normalization Bases (Gb)**
Paired read 1	223 908 854	22.1	11 893 250	1.18
Paired read 2	223 908 854	22.1	11 893 250	1.18
Singles	10 186 559	0.88	4 176 744	0.37

Table 3.10. Reads remaining after normalization of plant data.

Overlapping reads were combined using *FLASH* (Magoč & Salzberg, 2011) for each species, as was done for the All data assembly. Table 3.11 shows the final numbers of overlapped reads, split paired-end reads and pre-existing single reads used for assemblies for Plants and Arthropoda. The *FLASH*-overlapped Arthropoda data were assembled with *Trinity*. For the plant data, *Trinity* assemblies of normalized Illumina only, and of Illumina plus the *Q. robur* ESTs were made. The ESTs were incorporated as an attempt to improve the Plants assembly. Default *Trinity* parameters were used. The assembly metrics (table 3.12) were compared and the most suitable Plants assembly chosen for use as a reference for quantifying transcripts.

Library	Plants number of normalized reads	Plants normalised bases (GB)	Arthropoda number of reads	Arthropoda normalised bases (GB)
Read 1	3 148 666	0.3	5 606 161	0.5
Read 2	3 148 666	0.3	5 606 161	0.5
Overlapped	8 744 584	1.3	15 522 516	2.3
QC singles	4 176 744	0.4	1 044 703	0.1

Table 3.11. *FLASH* results for both gall wasp and oak datasets.

Assembly	Arthropoda	Plants	Plants + ESTs
N50	1736	1850	1622
Number of transcripts	108459	202766	247898
Number of components	89138	118649	130151
Number of transcripts in N50	12453	34262	52801
Maximum transcript size	37465	15437	15023
Number of bases in transcripts (MB)	94.44	207.29	276.15
***CEGMA* % complete**	97.58	96.37	95.16
Average copy number	2.57	2.81	3.50
% orthology	78.51	69.46	90.25

Table 3.12. Metrics for the species-specific assemblies including *CEGMA* results and the plants + ESTs assembly.

3.3.12 The species-specific assemblies

All three assemblies appear to have assembled well, with N50s > 1kb and long maximum contigs (table 3.12). The very long maximum length Arthropoda transcript encodes an insect muscle titin, which are known to be large proteins. The two plant assemblies have more transcripts and higher numbers of transcripts per component at 1.71 (Plants, Illumina only) and 1.90 (Plants, Illumina plus ESTs) versus 1.22 for the Arthropoda assembly. This may represent underlying biology if *Q. robur* has a greater average of number of isoforms per gene than *B. pallida*.

In addition to the metrics above, *CEGMA* (Core Eukaryotic Genes Mapping Approach) (Parra et al., 2007) scores were also evaluated. Parra et al. (2007) identified a set of core eukaryotic genes (*CEG*s) present in available eukaryote genomes, and the version used (2.4) contains 248 *CEG*s. *CEGMA* combines *BLAST* (Altschul et al., 1990), *GeneWise* (Birney et al., 2004) and *geneid* (Parra et al., 2000) searches and *HMMER* (Finn et al., 2011) to identify orthologs of the *CEGMA* gene set in tested dataset. Although, *CEGMA* is intended for genomes, under the assumption that *CEGs* will be constitutively expressed because they perform essential functions it is applied here to transcriptomes. Table 3.12 provides estimates of the percentage completeness for the *CEG*s, the average number of orthologs per *CEG* and percentage of *CEG*S with more than one ortholog. The final two metrics can indicate if more than one species is present in the transcriptome, as *CEG*S are supposed to be single-copy nuclear genes. This is apparent in the *CEGMA* scores for the Arthropoda and Plants assemblies (table 3.12). This may be due to the presence of parasitoid and inquiline derived transcripts within the assembly because the filtering process has not identified them.

As a result the ortholog copy number is much greater than one and lots of *CEG*s have orthologs. Redundancy in the assembly may also have caused high copy number scores. Therefore, *CEGMA* was run with the longest transcript from each *Trinity* clustered component only. The average copy number for the Arthropoda dropped from 2.47 to 2.24, suggesting assembly redundancy does not explain a large proportion of the high *CEG* copy number. For the Plants assembly *CEG* copy number dropped from 2.81 to 2.06, but the difference was greatest for the Plants + ESTs assembly which dropped from 3.50 to 1.97. Polymorphism in the ESTs, sampled from *Q. robur* tissues collected in France versus gall tissue collected in the UK could have caused this. Highly divergent allelic polymorphism can cause copies of the same gene to be assembled separately by *Trinity*. This would also explain the greater number of transcripts, 247 898 versus 202 766, and very high percentage of CEG orthology of 90% for the Plants + ESTs dataset.

The Arthropoda and Plants transcriptomes are all close to completeness; each has greater than 95% CEGs. The missing *CEG*s may not be true core eukaryotic genes. These are genes that are present in all of the few genomes surveyed to create the current version of the *CEGMA* database but are not true universal *CEG*s. The Illumina only Plants assembly was chosen because incorporating the EST dataset did not result in an improved assembly. Additionally, the potential redundancy discussed above could complicate expression analyses.

3.3.13 Annotating the assemblies and estimating the percentage of coding sequence

The two assemblies were annotated using *BLAST* (Altschul, 1990) and *InterProScan* (Zdobnov and Apweiler, 2001). The annotations were then used to apply gene ontology (GO) terms to transcripts using *BLAST2GO* (Conesa et al., 2005). The transcriptomes were aligned against the *BLAST* non-redundant nucleotide and protein ('nt' and 'nr') databases with an e-value cut-off of 1 x 10^{-5} and complexity filters on. In total, 104 585 (45% of total transcripts) and 54 684 (52% of total transcripts) of Plants and Arthropoda transcripts were *BLAST* annotated respectively. The low percentages probably reflect the fact that no genomes for species closely related to *B. pallida* or *Q. robur* are present in the *BLAST* databases; thus, many of the expressed genes are new and orthologs are not present in the databases. It also reflects error, as not all transcripts will be derived from protein coding genes, even after filtering. For *InterProScan*, 139 693 and 87 826, Plants and Arthropoda transcripts received some form of annotation.

The *InterProScan* and *BLAST* ‘nr’ results were combined to generate GO terms with *BLAST2GO*. In total, 65 360 oak and 25 128 gall wasp transcripts were annotated with GO terms respectively. The gene ontology project is an attempt to annotate genes and their product with a structured and controlled vocabulary (Harris et al., 2008).

3.3.14 Filtering the species specific assemblies

In the species-specific assemblies certain transcripts were derived from reads that escaped filtering at the combined assembly stage. To remove non-coding RNA transcripts *Rfam scan*, *tRNAscan*, and a *LAST* search of LSU-SSU sequences were repeated on the Plants and Arthropoda assemblies. Other undesirable transcripts, mitochondrial, chloroplast and mammalian sequences were identified by *BLAST* searches against the NCBI non-redundant protein and nucleotide databases (databases downloaded January 4th 2012). In total, 867 Arthropoda and 3082 Plants transcripts were filtered by removal from the gene counts matrix created with *RSEM* below

3.3.15 Generating transcript counts using *RSEM*

To perform differential expression (part B) analysis a matrix of counts consisting of a count for each gene in each replicate is required. This is complicated for a de novo transcriptome as reads can map to multiple isoforms of the same gene. The program *RSEM* provides a method to do this and condenses the read counts to the gene level.

Correctly apportioning reads to different isoforms of the same gene is difficult. In essence, how are the reads derived from each isoform of a gene aligned to the correct isoform? If one isoform is highly expressed, other overlapping isoforms will have inflated counts from multiple mappings; hence an isoform-based analysis is not advisable. As a result, gene level counts are essential; the differential expression programs used, *DESeq* (Anders and Huber, 2010) and *EdgeR* (Robinson et al., 2010) (Part C), both require gene counts. Additionally, not knowing which transcripts are isoforms of the same gene in a *de novo* transcriptome further complicates transcript quantification.

Fortunately, *Trinity* combines sets of transcripts into components corresponding to genes. The program *RSEM* (RNA-seq by Expectation Maximization) can then estimate combined counts per component/gene while controlling for multi-mappings. *RSEM* first aligns reads to transcripts using *Bowtie* (Langmead et al., 2009), and individual reads are allowed to map to multiple locations. It then computes maximum likelihood abundance estimates using a statistical model based on the Expectation-Maximization algorithm. The contribution of multi-mapping reads to a count for an isoform is a fraction dependent on the number of mappings per read. Gene counts are given by summation of transcripts within a *Trinity* component group. *RSEM* was run in both single and paired-end mode to evaluate the best performing mode with these data. For the single ended approach read ‘1’ from each pair was used and combined with singles resulting from the initial quality filtering of the data.

Origin	Mapping method	1	4	8	211	127	148	182	224	234	252	270C	281
		Early				Growth				Mature			
Plants	Single end	95.65	95.43	95.30	94.93	93.87	94.31	94.65	94.56	94.38	94.49	93.87	94.06
	Paired end	87.92	87.37	87.35	86.53	84.86	85.51	85.88	83.54	85.40	85.79	85.73	85.51
	Difference	7.73	8.06	7.95	8.40	9.01	8.80	8.77	11.02	8.98	8.70	8.14	8.55
Arthropod	Single end	93.32	93.07	93.24	91.52	95.17	93.83	94.65	94.94	93.02	93.7	94.96	91.93
	Paired end	84.31	83.54	83.92	80.95	84.35	85.15	86.17	80.51	83.45	85.43	88.25	82.47
	Difference	9.01	9.53	9.32	10.57	10.82	8.68	8.48	14.43	9.57	8.27	6.71	9.46

Table 3.13. Percentages of reads mapping and difference between single and paired end read mapping with RSEM for Plants and Arthropod datasets.

The single end approach mapped more reads (table 3.13) for both species across all replicates. Therefore, counts derived from single end mapping were used for differential expression analysis. Both Plants and Arthropoda data have similar mapping percentages for each replicate and both types of mapping. The discrepancy between paired- and single-end occurs because *Trinity* doesn't scaffold transcripts. As a result, reads from the same paired-end fragment can map to different transcripts representing fragments of a full-length transcript. *Bowtie*, the aligner *RSEM* runs internally does not consider these mappings valid for quantification; hence, the single read provides more accurate counts. Differences between single end and paired end percentages were similar for both Plants and Arthropods. Most differences are of the order 8-10%. The counts for each replicate were combined to create the necessary count matrix for testing differential expression (an example is given in appendix table 3.25).

Part B) Statistical analysis of stage specific variation in gene expression

3.4 Identifying candidate genes for gall induction and formation: *edgeR* and *DESeq*

To identify differentially expressed genes between the early and later stages of gall development two popular differential expression programs, *edgeR* (Robinson et al., 2010) and *DESeq* (Anders and Huber, 2010), were compared and contrasted. Both programs model count data using the negative binomial distribution to estimate means and variances for each gene in the experiment.

3.4.1 Filtering genes with low counts from the analysis

Before performing expression analyses genes with little power to detect differential expression were removed from the count matrices. This increases the overall power of the experiment as each gene is tested for significance independently; hence the effect of the multiple testing adjustment is reduced by the lower number of tests overall. Both programs preferentially remove genes with low overall counts, but they do this filtering by different methods.

EdgeR recommends filtering genes with expression below a set threshold, in fewer replicates than there are replicates per condition. So, for twelve replicates, with four per stage at least four replicates need to have counts above the threshold. *EdgeR* normalizes the threshold to counts per million (CPM), meaning if three reads out of 3 000 000 are aligned to a gene in a replicate, then that gene has a CPM of one. A CPM cut off of 2 was used for *EdgeR* for both gall wasps and oak genes; and 20 940 (24%) gall wasp and 16 790 (14%) oak genes were retained for differential expression analyses.

The *DESeq* approach is more complex and involves running a full differential expression analysis (implementing a generalized linear model approach) and then ranking the unadjusted for multiple testing p-values against their expression ranking (figures 3.6, both datasets). From this plot an expression rank cut-off can be chosen and the analysis repeated. The cut-off is chosen to remove all those values with a rank expression too low to have enough power to detect differential expression. It can be seen in figure 3.6 that after ranking each gene by expression, the lowest 70% of gall wasp genes and 30% of plant genes have very few genes with unadjusted p-values less than 0.003 (or 2.5, y-axis figure 3.6). The much higher percentage for gall wasps probably reflects the shallower sequencing depth of the larval transcriptome leaving little power to detect differential expression for many genes. After filtering, 79 893 oak and 26 155 gall wasp genes remained for *DESeq* analysis. Tables 3.14-15 compares the numbers of genes called as differentially expressed when applying p-value expression based filtering and the unfiltered data, There is an increase in detection power resulting from the filtering for both Plants and Arthropod datasets.

Finally, the Plants and Arthropoda datasets were analysed in *DESeq,* but were filtered using the *edgeR* CPM method. This assessed the effect of filtering method on the final set of differentially expressed genes.

Plants/*Q. robur*	**Filtered** **Not DE**	**Filtered** **DE**	**Sum**
Unfiltered Not DE	112 519	396	112 915
Unfiltered DE	5	3 518	3 523
Sum	112 524	3 914	116 438

Table 3.14. Effect of not filtering data on differential expression (DE) when using *DESeq*. Top row are data filtered using DESeq's p-value expression ranking based filtering and the left hand-side without filtering. More genes are called differentially expressed, 396, using filtering. Only 5 genes are called DE without filtering that were not called DE after filtering.

Arthropoda/*B. pallida*	**Filtered** **Not DE**	**Filtered** **DE**	**Sum**
Unfiltered Not DE	81 365	1 568	82 933
Unfiltered DE	1	5 226	5 227
Sum	81 366	6 794	88 160

Table 3.15. Effect of not filtering data on differential expression (DE) when using *DESeq*. Top row are data filtered using DESeq's p-value expression ranking based filtering and the left hand-side without filtering. More genes are called differentially expressed, 1568, using filtering. Only 1 gene is called DE without filtering that was not called DE after filtering.

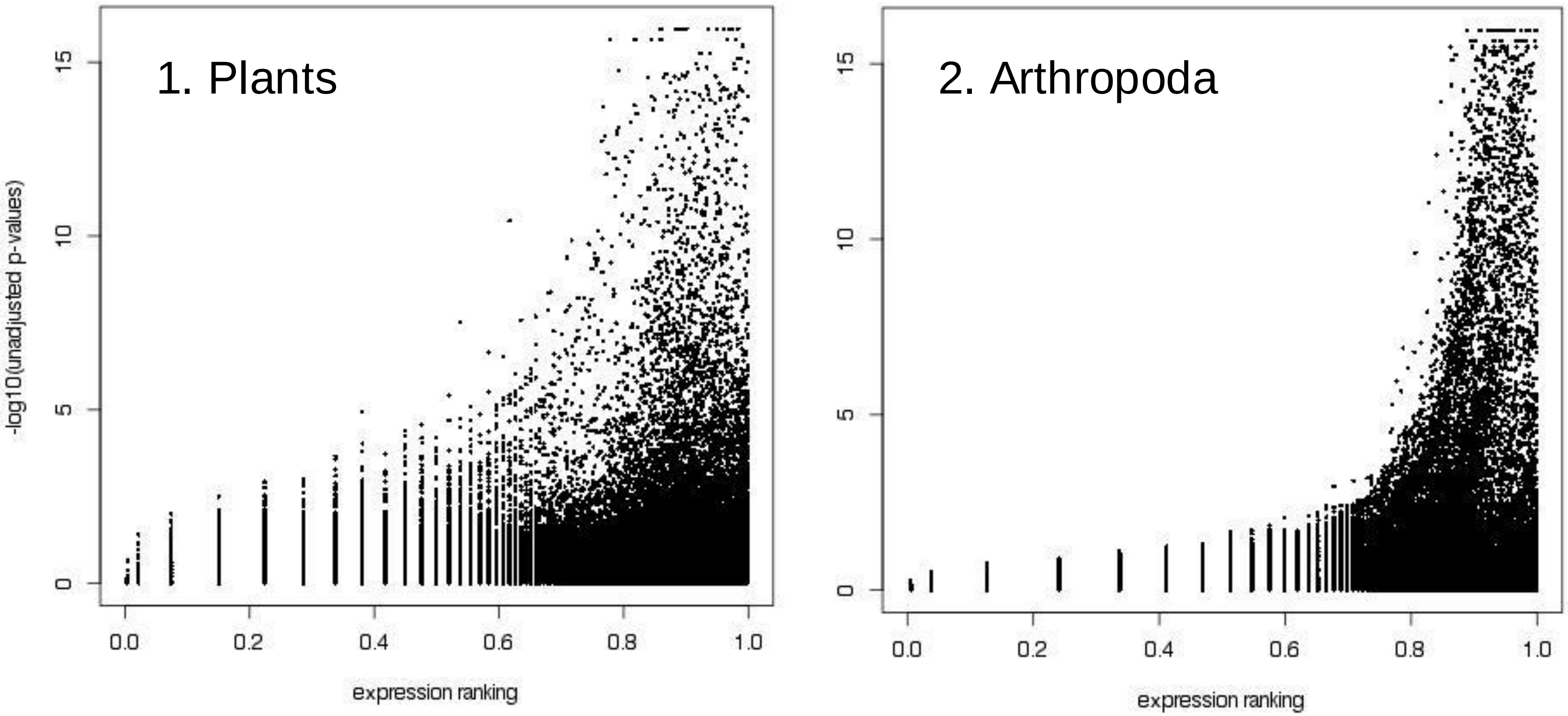

Figure 3.6. Log plot of p values versus expression demonstrating low power of lowly expressed genes. 1. Plants, 2. Arthropoda. Expression ranking is from low (0.0) to high (1.0). At high expression rankings the power to detect differential expression indicated by th e y-axis is greatly increased.

3.4.2 Normalising the datasets is essential for comparing counts between replicates

Before testing for differential expression the count datasets needed to be normalised to allow comparisons between replicates. Normalisation adjusts for differences in sequencing depth as indicated by total read counts per replicate. It also adjusts for differences in RNA composition within a replicate (Dillies et al., 2012). RNA composition adjustment is needed when a subset of genes is very highly expressed and a large proportion of reads are derived from these genes.

DESeq and *edgeR* use different normalization methods, although they both assume most genes are not differentially expressed. *DESeq* derives a scaling factor for each replicate from the ratio of the median, for each gene, of its read count divided by the geometric mean of counts for that gene across all replicates (Dillies et al., 2012). *EdgeR* uses the trimmed mean of *M*-values (TMM) normalization (Robinson and Oshlack, 2010). It uses one replicate as a reference by which the other replicates are normalized. The TMM is the weighted mean of gene-wise log expression ratios between the reference and non-reference replicates. The most expressed genes and those with the largest log ratios are trimmed from the TMM calculation (Dillies et al., 2012).

Dillies et al., (2012) compared the performance of several normalization techniques including *DESeq*, TMM, upper quartile, median, reads per kilobase per million mapped reads (RPKM) and quantile based normalization methods. They found that *DESeq* and TMM normalization were superior to the other methods and produced similar results. Robust normalization was key to this experiment as the total read counts derived from oak and gall wasp varies widely across replicates in different stages because of changes in the proportion of reads derived from oak or gall wasp among replicates (i.e. mature gall tissue has a higher proportion of total gall wasp expression than early stage tissue).

3.4.3 Clustering of replicates by global patterns of expression: are replicates from the same stage similar?

By representing the filtered count data visually, relationships between the replicates were assessed. Replicates from the same stage were expected to cluster more closely than replicates in other stages.

For clustering, *DESeq* first performs a variance-stabilising transformation (VST) of the data. This equalizes the variance across genes, allowing each gene to contribute equally to Euclidean-based clustering of the data. Otherwise the most highly expressed genes that have the most variance will dwarf any influence of other genes. VST was not applied to the actual differential expression analysis in *DESeq*. Figures 3.7-3.9 are the VST stabilized sample PCAs produced by *DESeq* for *Q. robur* and *B. pallida* respectively, for the 500 most variable genes across replicates*.* The *Q. robur* and *B. pallida* heat maps are very similar and show that the samples cluster well together, particularly in the early stage, albeit not perfectly as one growth stage replicate clusters with mature stage replicates. This was encouraging and validated the ‘wild’ sampling strategy for the gall tissues. The two fungal infected replicates, 211 of the early stage and 270C of the mature stage, sit separately from the other replicates for the *Q. robur* PCA. The *B. pallida* PCA (figure 3.8) is less well resolved than the *Q. robur* PCA. The fungus-infected replicates sit separately from other replicates in the same stage, although early stage replicate 211 sits with growth stage replicates and not as an outlier. Fungal infection has a strong effect on expression in both species, enough to form the second axis of the PCAs, and needs to be controlled for. One growth stage replicate is also an outlier in the *B. pallida* PCA. For both species the x-axis (principal component 1) differentiates stage of gall growth well.

Low sampling coverage could explain the lower resolution of *B. pallida*; alternatively, the sampling method may not fit gall wasp larval stages/global expression patterns as well as it does the oaks’. For example, a delay may occur between larval expression changes and host response.

Clustering indicated the need to control for fungal infection in replicates 211 and 270C. The effect on replicate 270C was surprising as the percentage of fungal expression in the data was low at 0.94%. Replicate 281 was not an obvious outlier in any plots despite the percentage of virus-derived reads (1.98%).

3.4.4 Controlling for fungal infection derived expression: Fitting a batch effect

A simple fungal infected (replicates 211 and 270C) versus uninfected replicates differential expression analysis, ignoring stages, was run on the Plants/*Q. robur* data in *DESeq*. In total 320 genes were differentially expressed (at adjusted alpha = 0.05). *BLAST2GO* annotations revealed that several of the genes highly expressed in 211 and 270C are annotated as chitinases that hydrolyse the cuticle of fungi to combat infections. These differentially expressed genes were removed from the gene counts matrix and a full GLM, with developmental stages fitted as factors, in *DESeq* was run. The resulting PCA plot (figure 3.9) shows that replicates 211 and 270C now cluster most closely with replicates of the same stage. The PCA plot is also very similar to the '*edgeR* filtered *DESeq'* dataset PCA plot for which no genes were removed (appendix, figure 3.15). *EdgeR* filtering probably removes many fungal-infection affected genes. This is because the *edgeR* filtering criteria requires four replicates with a counts per million greater than two. Therefore genes with expression higher than 2 CPM in two replicates (211 and 270C) only are filtered. Genes differentially expressed in response to fungal infection were not filtered from the dataset for the final analysis, as they may have additional roles during gall development.

Fungal Infection status of each replicate was fit as an additional factor to the model design for both *DESeq* and *edgeR*. By controlling for this expression, fungal-effected genes were not called as false positives in the full analyses.

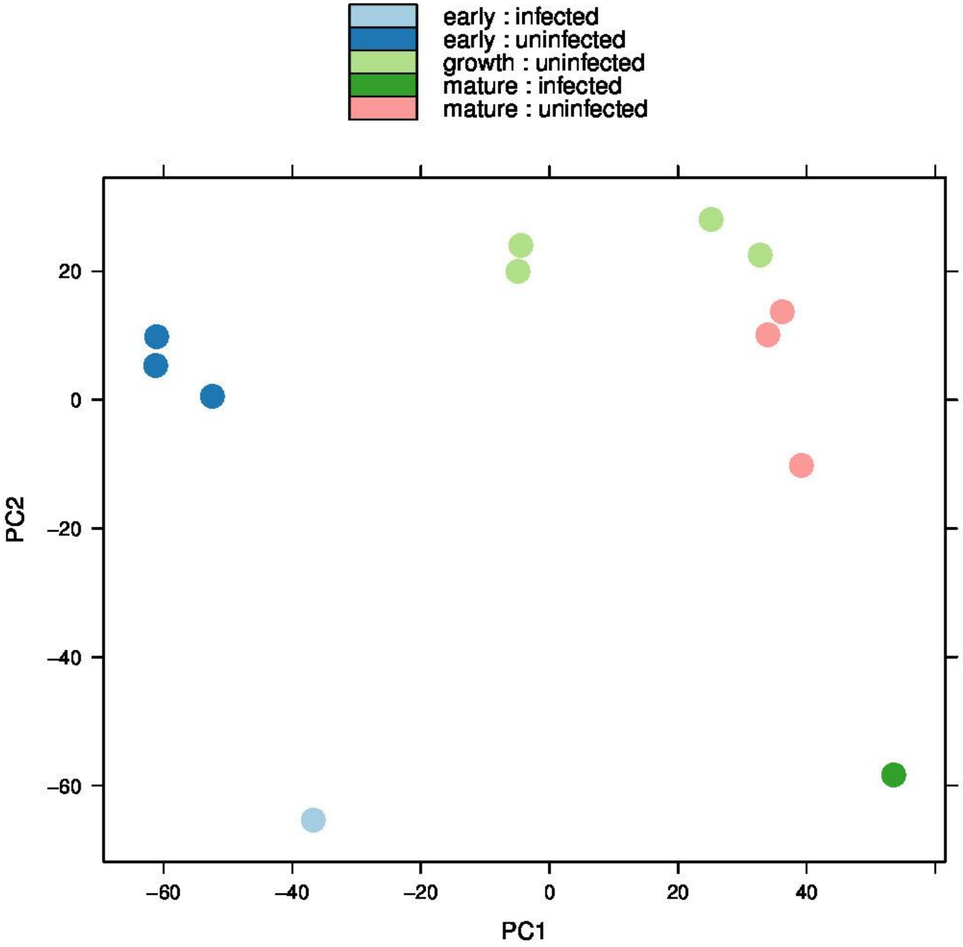

Figure 3.7. Principle components analysis of Plant (putative oak expression) replicates. The two fungal infected replicates sit separately from the others.

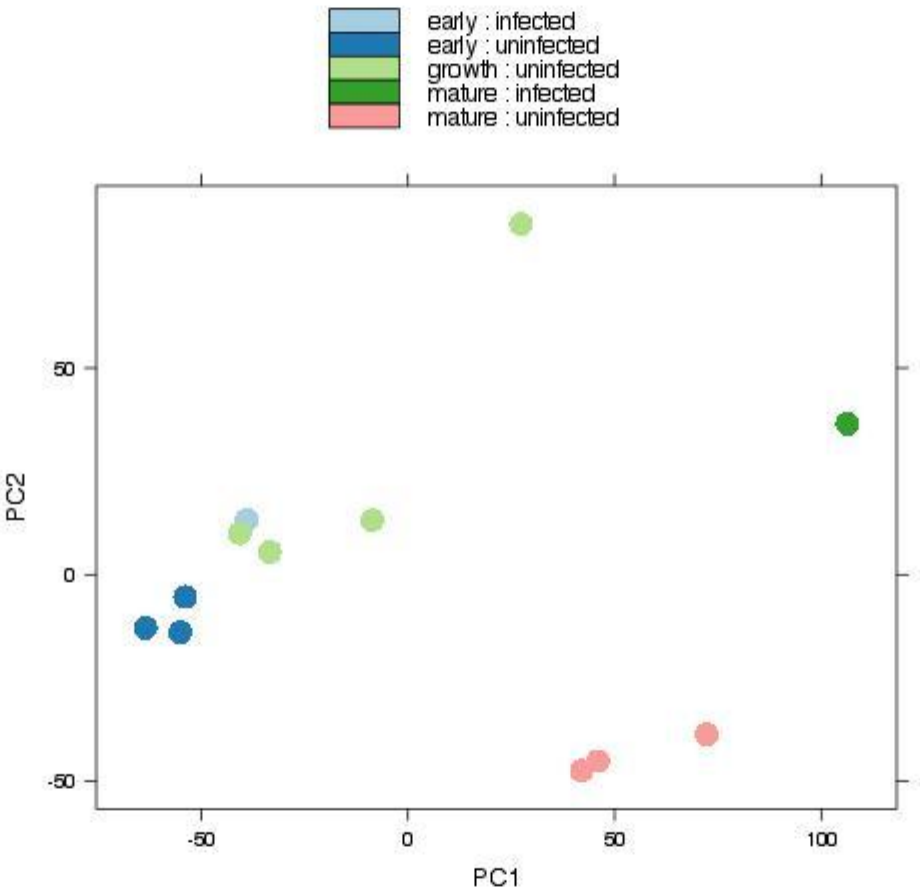

Figure 3.8. Principle components analysis of Arthropod (putative gall wasp expression) replicates.

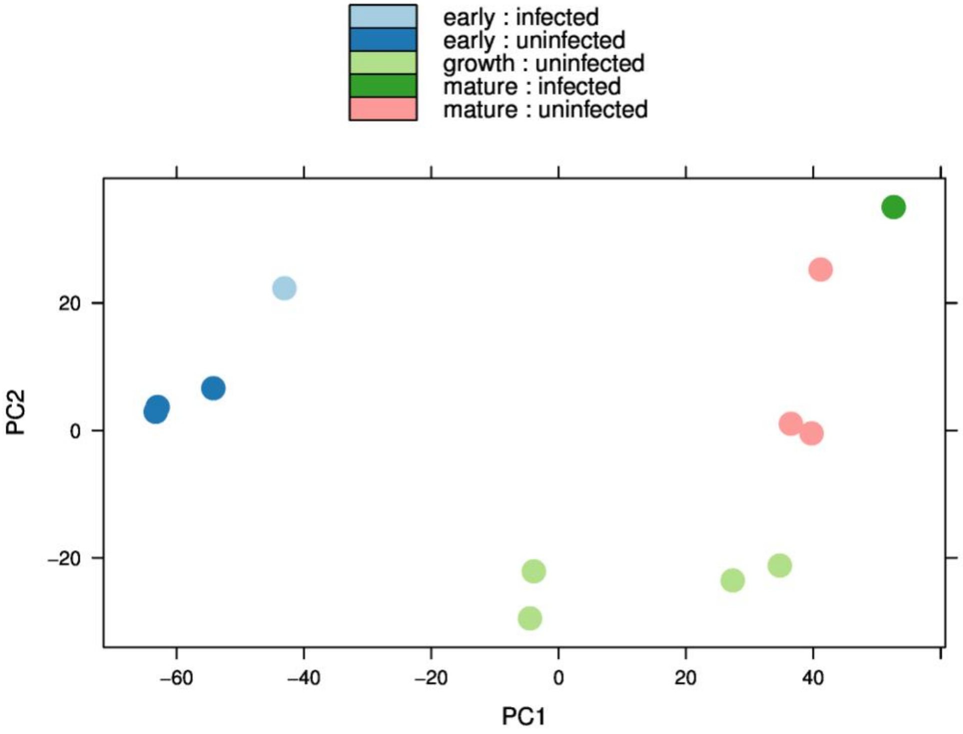

Figure 3.9. Principle components analysis of Plant replicates after removal of genes significantly differentially expressed due to fungal infection. The data clusters much better by stage compared to figure 3.7.

3.4.5 Testing for differentially expressed genes

Both programs apply a generalized linear model (GLM) based on the negative binomial distribution and then likelihood ratio tests to identify significant differential expression (adjusted α = 0.05). The null hypothesis was of no difference in expression between the stages, and the alternative 2-tailed hypothesis is of a difference in expression. Additional contrasts were made between each pair of conditions (early versus growth, early versus mature and growth versus mature) in *EdgeR*, under the same hypotheses, to produce volcano plots contrasting pairs of stages (figures 3.12-14).

P-values from the GLM likelihood-ratio tests were adjusted for multiple testing using the Benjamini-Hochberg procedure by both programs (Benjamini and Hochberg, 1995). Significant genes are given for oak and gall wasp differential expression for *DESeq*, *edgeR* and *edgeR* filtered *DESeq* dataset.

Plants/*Q. robur*	**Number DE**	**Higher in Early**	**> 2 fold change**	**Lower in Early**	**> 2 fold change**
EdgeR	5790	2748	443	2560	662
DESeq	3869	1738	598	1843	998
EdgeR* filtered *DESeq	4126	1980	405	1888	565

Arthopod/*B. pallida*	**Number DE**	**Higher in Early**	**> 2 fold change**	**Lower in Early**	**> 2 fold change**
EdgeR	9347	4554	1510	2780	1367
DESeq	6659	3214	1327	2230	1330
EdgeR* filtered *DESeq	7905	4432	1552	1881	1039

Table 3.16. Differentially expressed (DE) genes for each species for *edgeR, DESeq* and *edgeR* filtered *DESeq* analyses. Genes described as showing higher expression in the early stage, refer to those that have negative relative expression (<0) in both growth and mature stages versus the early stage; the reverse criterion is true for lower expression in the early stage. The > 2 fold change genes are DE genes with greater than 2 fold differences in expression either higher of lower in the early stage versus the later stages.

The *edgeR* contrasts (table 3.17) show for *B. pallida* a much greater change in expression between the early or growth stages versus mature, suggesting that a global change in expression patterns occurs between the growth and mature stages. In contrast, *Q. robur* has greater differential expression between the early stage versus the growth and mature stages, albeit more pronounced against the mature, suggesting a major shift in expression from early to growth stages.

EdgeR **contrasts**	**Number DE**	**Higher in Earlier stage**	**> 2 fold change**	**Lower in Earlier stage**	**> 2 fold change**
			Q. robur		
Early vs Growth	3085	1405	551	1680	823
Early vs Mature	5850	3050	1186	2800	1291
Growth vs Mature	1374	836	421	538	316
			B. pallida		
Early vs Growth	3081	1658	1406	1423	1347
Early vs Mature	8984	5867	5454	3117	2110
Growth vs Mature	6477	5133	4967	1344	931

Table 3.17. EdgeR pairwise contrast between stages used to create volcano plots, figure 3.12-14.

Figures 3.10-11 are venn diagrams of overlapping total differential expression between the three analyses. Interestingly, the *edgeR* filtered *DESeq* shares far more differential expression with the *edgeR* results than the *DESeq* only analysis for both *Q. robur* and *B. pallida*. This suggests that the initial filtering method has a strong effect on the results. Appendix figures 3.16-17 show overlapping differentially expressed genes with greater than 2 fold change between early and both later stages. They show greater proportional overlap between *edgeR* and *DESeq*, probably because these genes are highly expressed and were not filtered by either program.

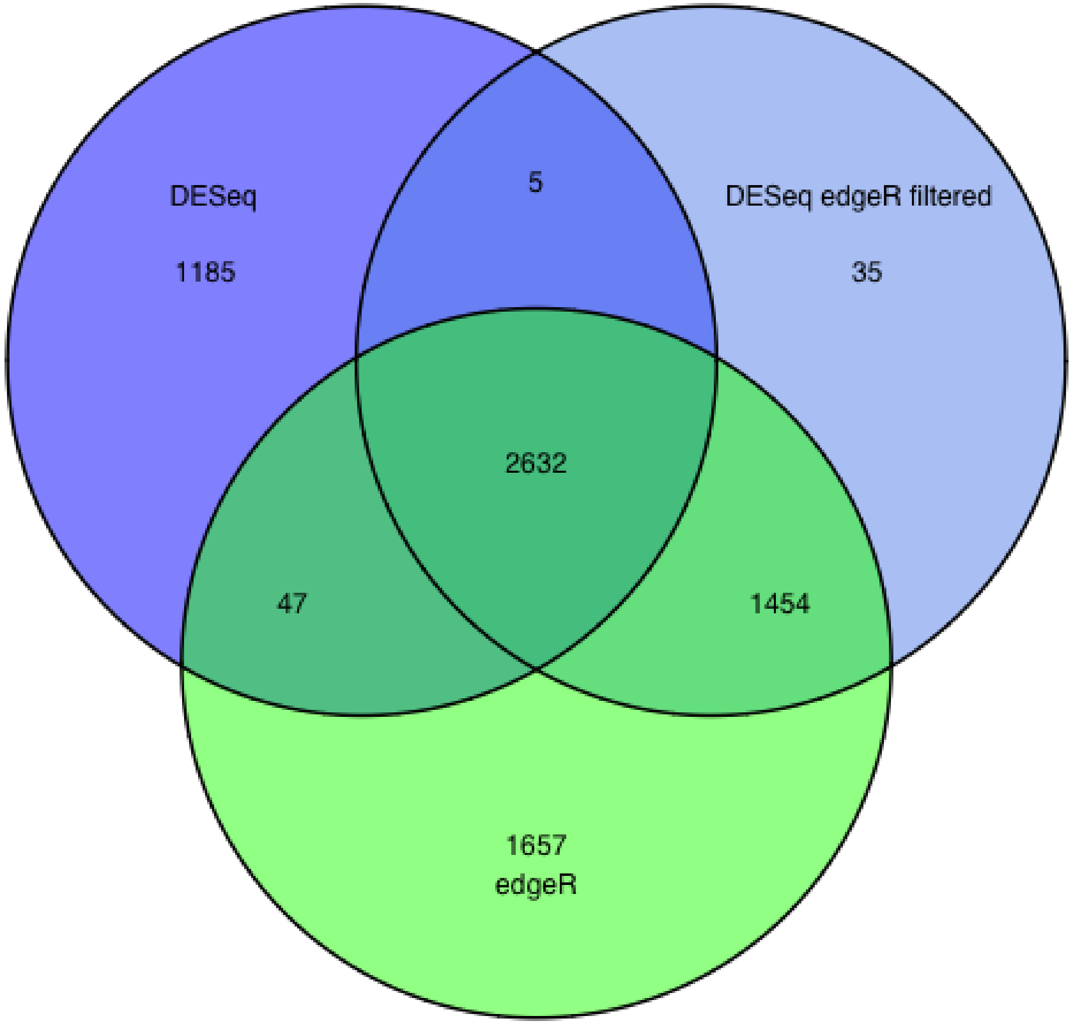

Figure 3.10. Venn diagram of overlapping genes of differential expression in oak between analyses.

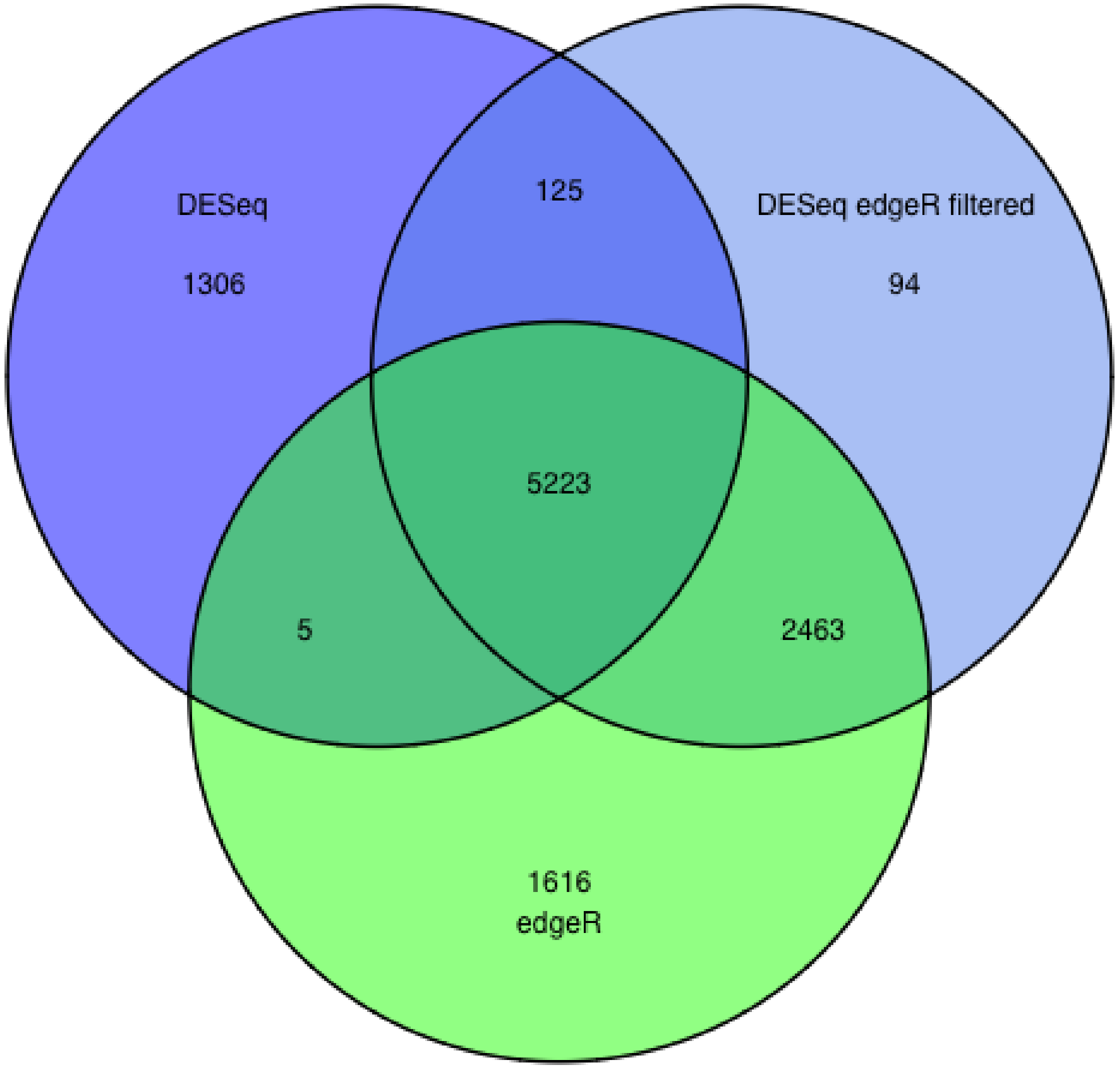

Figure 3.11. Venn diagram of overlapping genes of differential expression in gall wasp between analyses.

3.4.6 Global patterns of gene expression in gall tissues: log fold change plots of differentially expressed genes

By plotting log fold change against p-value for each gene (volcano plots), the broad changes in gene expression can be better observed, as shown for the *edgeR* contrasts between specific stages in figures 3.12-14. It was not possible to make these plots for the GLM results from *edgeR* and *DESeq* as a significantly differentially expressed gene could result from either of the early versus growth or mature comparisons, resulting in misleading plots.

The early versus growth stage reveals a group of strongly expressed plant genes in the early stage, as evidenced by the strong p-values and log fold change. There is also a right skew of gall wasp genes more highly expressed in the growth stage. On inspection there are 318 genes with a log fold change increase greater than ten in the growth stage. The raw counts indicate that these genes are affected by an unknown factor on replicate 224 only as no other growth stage replicates contain expression for these genes, and therefore are probably not important to gall development.

The early versus mature and growth versus mature plots (figure 3.12-12) are very similar (note the different y-axis scale). In both, gall wasp genes have more statistically significant genes in the earlier stage (the longer left-hand smears of the plot) while plant genes are more evenly divided. This is supported by the number of genes more highly expressed in the earlier stage versus later stage in table 3.16. Taken together, this would indicate that the mature stage represents very different global gall wasp expression to the two earlier stages, as overall 4231 genes are differentially expressed by the early and growth stage commonly versus the mature stage.

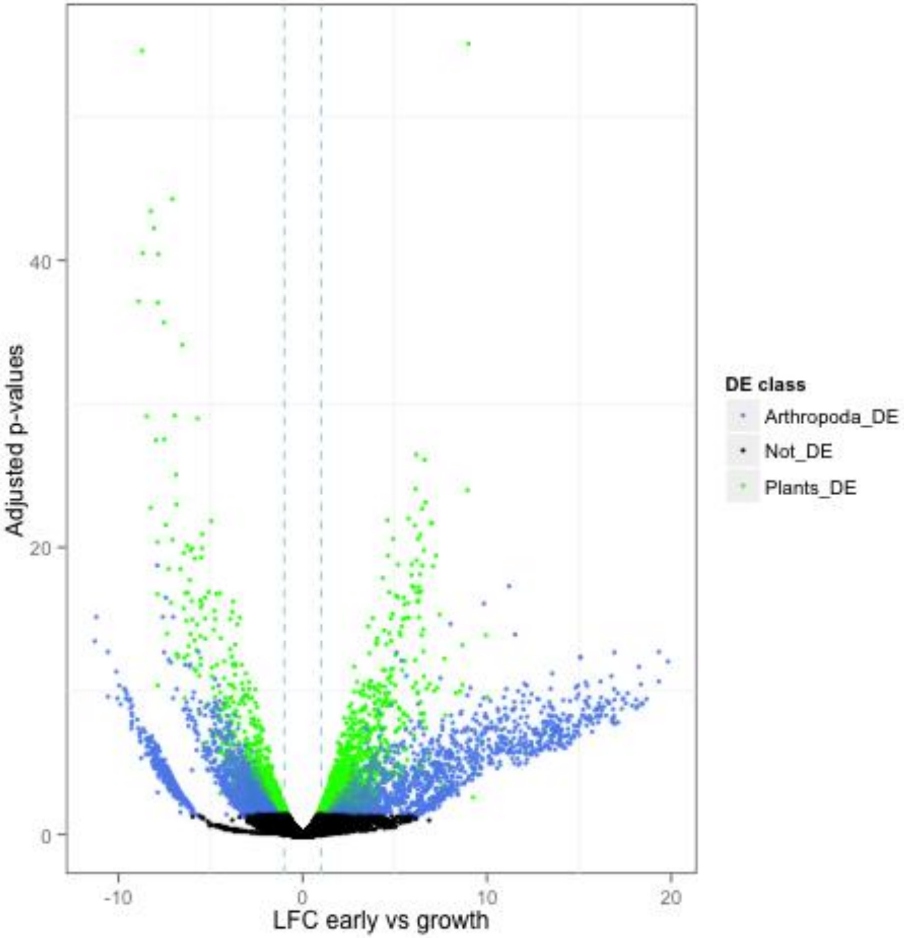

Figure 3.12. Volcano plot of early versus growth stages for both datasets. Dashed grey lines indicate a log 2 fold change in expression. LFC: log fold change; DE: differential expression.

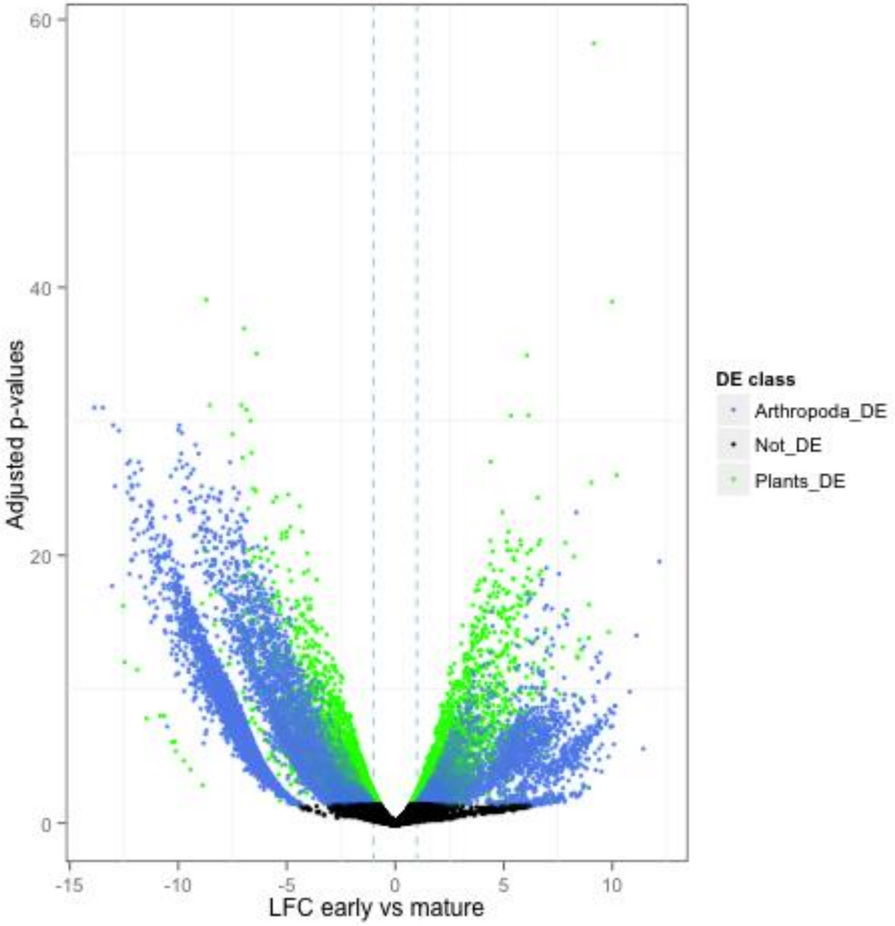

Figure 3.13. Volcano plot of early versus mature stages for both datasets. Dashed grey lines indicate a log 2 fold change in expression. LFC: log fold change; DE: differential expression.

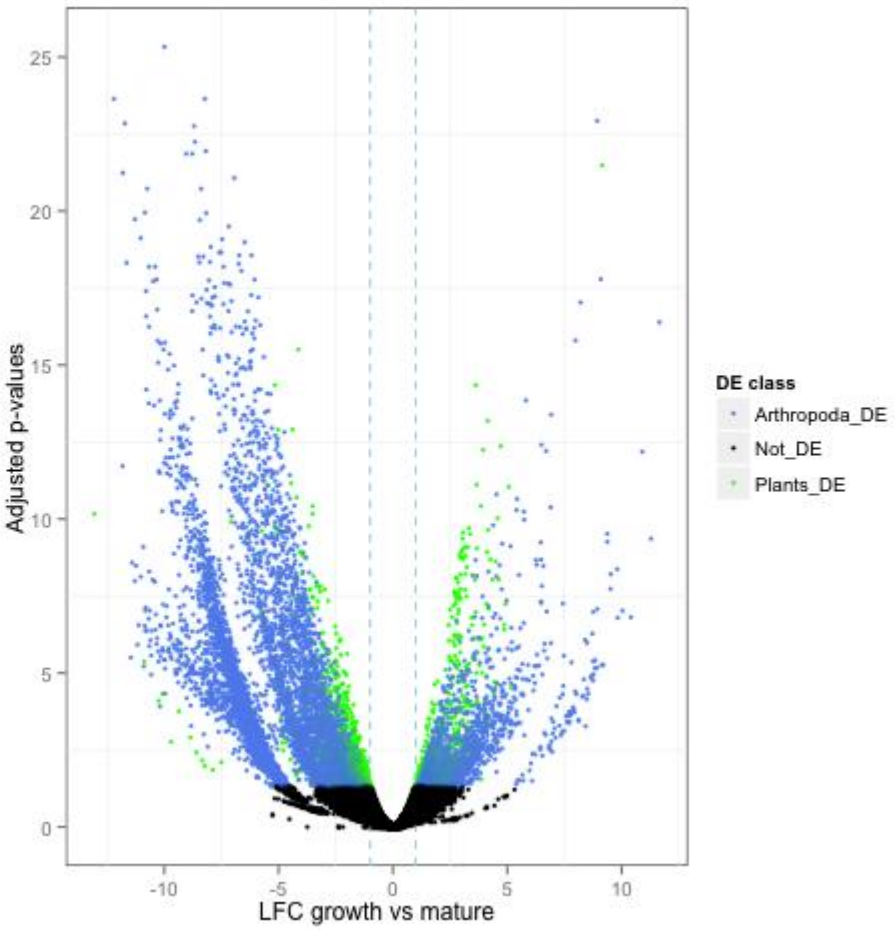

Figure 3.14. Volcano plot of growth versus mature stages for both datasets. Dashed grey lines indicate a log 2 fold change in expression. LFC: log fold change; DE: differential expression.

Part C) Identifying possible roles of differentially expressed genes

3.5 Differentially expressed genes and *BLAST* and *InterProScan* annotations

Table 3.18 outlines the numbers genes more highly and relatively less expressed in the early versus growth and mature stages, the number of genes with similarity to sequences in the non-redundant nucleotide and protein databases (*BLASTx* and *BLASTn*, e-value cut-off of 1 x 10^{-5}) and *InterProScan* annotations. Differentially expressed genes identified by both *edgeR* and *DESeq* were chosen for further analysis, and this amounted to 5228 *B. pallida* and 2679 *Q. robur* genes.

Almost all *Q. robur* differentially expressed hits have some form of annotation from *BLAST* and *InterProScan*. This is also true for those *B. pallida* genes where expression is greater in the growth and mature stages. However, it is not the case for genes more highly expressed in the early stage, as only 15% (441/2927) have *BLAST* hits. Furthermore, the percentage is even lower for genes showing more than two-fold difference in expression, at 9% (105/1138). One possible explanation is that many genes involved in gall induction may be specific to cynipid gall wasps, and hence absent from current functional databases. It is not currently possible to perform a strong test of this hypothesis, as the genome of a suitable out-group is not available. Conversely, the high number of *BLAST* hits at the later stage suggests that at this point in gall development larval gene expression is involved in universal processes, particularly feeding and growth. *InterProScan* annotated genes are high across both comparisons, allowing some inferences to be made about early differentially expressed genes without known orthologs. However, many *InterProScan* annotations are not greatly informative, such as the presence of a signal peptide or transmembrane domain.

3.4.1 *B. pallida* genes

Inspecting the list of annotated genes expressed greater than two fold higher in the early stage (table 3.18) indicates some candidates for involvement in induction and formation in both *B. pallida* and *Q. robur* respectively. Two distinct genes annotated as chitinases are differentially expressed at high absolute expression compared to other gall wasp genes (based on raw counts, table 3.19) in the gall wasp larvae versus later stages. Other differentially expressed early stage gall wasp genes with high counts include a serine protease, serine threonine protein kinase, carbonic anhydrase and glycine N-acyltransferase-like protein. Many of the genes called differentially expressed in the early stage, including proposed horizontally transferred (chapter 4) pectin and pectate lyase, have very low absolute counts (table 3.19).

Over 60% of the unannotated early differentially expressed gall wasp genes contain a signal peptide sequence indicating presence in a secretory pathway. This is not particularly noteworthy, given that 76% of the later gall wasp differentially expressed genes also contain a signal peptide.

3.5.2 *Q. robur* genes

The oak data has several excellent candidates for involvement in gall induction from *BLAST* functional annotations. Two *Q. robur* genes that are differentially, and highly, expressed in the early stage gall tissues are orthologous to early nodulin genes. In leguminous plants, early nodulin genes respond to nodulation factors (Nod) produced by nitrogen-forming bacterial symbionts to create root nodule galls. Specifically, the transcripts encode plastocyanin domain-containing arabinogalactan protein (AGP) (counts for one shown in table 3.19). A search of the *BLAST2GO* annotations for more Nod factor annotated genes revealed two more nod factor-induced genes that are differentially expressed highly in the early gall stage (although with log fold changes less than 2 versus the growth and mature stage). The other nod factor-induced gene is the highest expressed of the four and is a major facilitator superfamily (MFS) membrane transporter that transports small solutes such as sugars. There are four nod factor-induced differentially expressed (>2 fold change) genes in the growth and mature stages. In addition to the nod-induced genes, *Q. robur* genes of interest based on the rapid growth of early to growth stage galls are 13 oak cyclin and cell division kinase genes differentially expressed highly in the early stage.

Overlapping DE	**Number DE**	**Higher in Early**	**> 2 fold change**	**Lower in Early**	**> 2 fold change**
Q. robur	2679	1486	401	1070	466
B. pallida	5228	**2927**	**1138**	1360	810

***BLAST* hits**	**Number DE**	**Higher in Early**	**> 2 fold change**	**Lower in Early**	**> 2 fold change**
Q. robur	2571	1457	395	996	425
B. pallida	2032	**441**	**105**	1258	758

***Interproscan* annotations**	**Number DE**	**Higher in Early**	**> 2 fold change**	**Lower in Early**	**> 2 fold change**
Q. robur	2664	1480	398	1061	459
B. pallida	5068	**2801**	**1068**	1355	808

Table 3.18. Overlapping differentially expressed genes and *BLAST* and *InterProScan* based annotations. *B. pallida* genes that are relatively highly expressed are shown in bold to highlight their relative lack of annotation. When only the *Phobius* protein function prediction results were parsed from the *InterProScan* annotations the numbers were almost identical as full *InterProScan* annotations shown.

Stage			**Early**				**Growth**				**Mature**	
Replicate	***1***	***4***	***8***	***211***	***127***	***148***	***182***	***224***	***234***	***252***	***270C***	***281***
Chitinase	1496	825	584	757	222	149	92	15	4	0	1	1
Chitinase	2128	1553	958	1713	264	165	126	14	8	6	4	4
Pectate lyase	33	15	12	14	7	2	2	0	0	1	0	0
Pectin lyase	11	6	5	8	2	0	1	0	0	0	0	0
AGP Early nodulin 55-2 precursor	4570	2572	1718	2069	14	7	2	1	1	6	0	3
BCCP	2649	1454	1094	1203	933	1873	567	401	92	67	129	197

Table 3.19. Raw counts for several genes of interest in early, growth and late stage galls. These counts are indicative of absolute levels of gene expression, whether or not specific genes are differentially expressed across gall stages.

3.5.3 Enrichment of GO terms using a Fisher's exact test

GO term enrichment in differentially expressed genes was performed using a Fisher's exact test implemented in the *BLAST2GO* analysis suite. The test evaluates the frequency of GO terms between a test set of differentially expressed genes and a reference set (Conesa et al., 2005). The terms cover three domains: cellular component (CC), molecular function (MF) and biological process (BP). Thus a gene can be annotated with GO terms from each domain detailing its cellular position (CC), catalytic activity (MF) and involvement in a defined pathway - for example, cell division (BP).

Here the reference set consists of the other functional annotations for those genes that passed normalization in both *EdgeR* and *DESeq*. This amounted to 19 838 gall wasp and 16 790 oak genes. Genes in the differentially expressed set were removed from the reference for the tests. A 2x2 contingency table is constructed for each GO term. It consists of how many genes in the test set are assigned that GO term and how many in the test set are not, and how many genes in the reference have that term and how many in the reference set do not. The resulting P-value is corrected for multiple testing. The GO terms associated with genes in the greater than 2-fold change, up and down in both *edgeR* and *DESeq* for the early versus later stages were tested. Enrichment tests were performed with a false discovery rate cut off of 0.05. The most specific GO annotations terms of enriched GO terms were identified (table 3.20). These are the most specific GO annotations assigned to this gene, as a gene may have several enriched GO terms from very general (metabolism) to the very specific (plant cell wall degradation).

GO term enrichment	Early	Most specific GO	Later	Most specific GO
B. pallida	0	0	70	18
Q. robur	216	39	50	19

Table 3.20. GO term enrichment for greater than 2 fold differentially expressed genes for early versus growth and mature stages showing numbers all enriched GO terms and the number of most specific GO terms per gene.

Not surprisingly, as GO terms are generated from *BLAST* and *InterProScan* annotations, the early *B. pallida* genes do not have any enriched GO terms. This reinforces the idea that these genes lack orthologs in other sequenced organisms, and reflects a bias in GO terms to well-studied processes in model organisms. However, there are enriched GO terms for those genes more highly expressed in the later stages (table 3.21). The opposite effect occurs for the oak data, with more GO term enrichment identified in the early stage versus the later stages. Lists of specific GO terms for the early and later *Q. robur* stages are found in tables 3.22-23 (the complete tables containing all GO hierarchy terms, from the most general to specific, assigned to a gene: appendix tables 3.27-28)

Table 3.21. Enriched GO terms for *B. pallida* genes more highly expressed in the growth and mature stages versus the early stage. F = molecular function; C = cellular component; P = biological process. #Test = number of differentially expressed genes for this GO annotation; #Ref number of genes for this GO annotation in the reference, not including differentially expressed genes; #not in test number of differentially expressed genes not in this GO annotation; #not in Ref number of genes that do not have this GO annotation in the reference.

Table 3.22. Enriched GO terms for *Q. robur* genes more highly expressed in the early versus the growth and mature stages. F = molecular function; C = cellular component; P = biological process. #Test = number of differentially expressed genes for this GO annotation; #Ref number of genes for this GO annotation in the reference, not including differentially expressed genes; #not in test number of differentially expressed genes not in this GO annotation; #not in Ref number of genes that do not have this GO annotation in the reference.

Table 3.23. Enriched GO terms for *Q. robur* genes more highly expressed in the growth and mature stages versus the early stage. F = molecular function; C = cellular component; P = biological process. #Test = number of differentially expressed genes for this GO annotation; #Ref number of genes for this GO annotation in the reference, not including differentially expressed genes; #not in test number of differentially expressed genes not in this GO annotation; #not in Ref number of genes that do not have this GO annotation in the reference.

GO-ID	Term	Category	FDR	P-Value	#Test	#Ref	#not in Test	#not in Ref
GO:0003735	structural constituent of ribosome	F	4.06E-11	1.45E-14	63	211	447	5446
GO:0005811	lipid particle	C	1.53E-10	1.10E-13	50	145	460	5512
GO:0006412	Translation	P	1.36E-07	1.79E-10	83	419	427	5238
GO:0055114	oxidation-reduction process	P	0.001289173	2.77E-06	54	293	456	5364
GO:0005524	ATP binding	F	0.002749936	7.23E-06	73	462	437	5195
GO:0005576	extracellular region	C	0.004204039	1.41E-05	34	158	476	5499
GO:0005875	microtubule associated complex	C	0.004752315	1.77E-05	37	182	473	5475
GO:0016469	proton-transporting two-sector ATPase complex	C	0.005932773	2.62E-05	14	35	496	5622
GO:0030246	carbohydrate binding	F	0.006027306	2.74E-05	20	69	490	5588
GO:0005200	structural constituent of cytoskeleton	F	0.006729023	3.30E-05	12	26	498	5631
GO:0022627	cytosolic small ribosomal subunit	C	0.009640831	5.48E-05	11	23	499	5634
GO:0006457	protein folding	P	0.009640831	5.53E-05	22	86	488	5571
GO:0003723	RNA binding	F	0.009875232	5.79E-05	55	339	455	5318
GO:0008553	hydrogen-exporting ATPase activity, phosphorylative mechanism	F	0.014487533	9.87E-05	11	25	499	5632
GO:0015991	ATP hydrolysis coupled proton transport	P	0.018149182	1.30E-04	11	26	499	5631
GO:0030017	Sarcomere	C	0.030194911	2.27E-04	13	39	497	5618
GO:0061061	muscle structure development	P	0.046059637	3.80E-04	30	160	480	5497
GO:0007052	mitotic spindle organization	P	0.048213359	4.04E-04	20	87	490	5570

GO-ID	Term	Category	FDR	P-Value	#Test	#Ref	#not in Test	#not in Ref
GO:0051322	Anaphase	P	4.81E-29	4.36E-32	37	75	284	12628
GO:0051567	histone H3-K9 methylation	P	1.48E-25	2.11E-28	42	148	279	12555
GO:0006275	regulation of DNA replication	P	1.22E-22	2.83E-25	33	91	288	12612
GO:0016572	histone phosphorylation	P	5.92E-21	1.61E-23	25	42	296	12661
GO:0010389	regulation of G2/M transition of mitotic cell cycle	P	2.20E-18	7.97E-21	22	37	299	12666
GO:0006270	DNA-dependent DNA replication initiation	P	1.06E-17	4.25E-20	23	48	298	12655
GO:0006306	DNA methylation	P	4.13E-17	1.77E-19	33	151	288	12552
GO:0048451	petal formation	P	3.38E-16	1.58E-18	19	30	302	12673
GO:0048453	sepal formation	P	2.88E-15	1.68E-17	18	29	303	12674
GO:0051225	spindle assembly	P	2.39E-13	1.86E-15	16	27	305	12676
GO:0031048	chromatin silencing by small RNA	P	5.16E-12	5.21E-14	23	104	298	12599
GO:0006346	methylation-dependent chromatin silencing	P	6.08E-12	6.22E-14	23	105	298	12598
GO:0003777	microtubule motor activity	F	6.70E-11	7.78E-13	15	36	306	12667
GO:0005874	Microtubule	C	4.69E-08	7.41E-10	16	76	305	12627
GO:0009909	regulation of flower development	P	6.13E-08	9.76E-10	29	286	292	12417
GO:0007067	Mitosis	P	8.94E-07	1.56E-08	13	58	308	12645
GO:0007018	microtubule-based movement	P	1.62E-06	2.93E-08	12	50	309	12653
GO:0010075	regulation of meristem growth	P	6.32E-06	1.18E-07	18	146	303	12557
GO:0042023	DNA endoreduplication	P	9.92E-05	2.00E-06	12	78	309	12625
GO:0000079	regulation of cyclin-dependent protein kinase activity	P	1.48E-04	3.01E-06	7	20	314	12683
GO:0007169	transmembrane receptor protein tyrosine kinase signaling pathway	P	2.42E-04	4.98E-06	12	86	309	12617
GO:0009855	determination of bilateral symmetry	P	3.15E-04	6.61E-06	13	105	308	12598
GO:0010103	stomatal complex morphogenesis	P	4.09E-04	8.73E-06	13	108	308	12595

GO:0009524	Phragmoplast	C	0.001085499	2.33E-05	8	41	313	12662
GO:0010143	cutin biosynthetic process	P	0.003054561	6.76E-05	4	6	317	12697
GO:0004674	protein serine/threonine kinase activity	F	0.003839808	8.60E-05	34	654	287	12049
GO:0009957	epidermal cell fate specification	P	0.004547407	1.04E-04	4	7	317	12696
GO:0007000	nucleolus organization	P	0.005561425	1.30E-04	5	16	316	12687
GO:0000793	condensed chromosome	C	0.005561425	1.30E-04	5	16	316	12687
GO:0008356	asymmetric cell division	P	0.010466792	2.55E-04	5	19	316	12684
GO:0005576	extracellular region	C	0.011130584	2.72E-04	43	964	278	11739
GO:0000914	phragmoplast assembly	P	0.011410855	2.81E-04	3	3	318	12700
GO:0031225	anchored to membrane	C	0.020061757	5.08E-04	9	85	312	12618
GO:2000123	positive regulation of stomatal complex development	P	0.023420031	6.06E-04	2	0	319	12703
GO:0048443	stamen development	P	0.025120045	6.67E-04	10	108	311	12595
GO:0006323	DNA packaging	P	0.027027594	7.21E-04	7	54	314	12649
GO:0042127	regulation of cell proliferation	P	0.033869426	9.21E-04	9	93	312	12610
GO:0009955	adaxial/abaxial pattern specification	P	0.038539463	0.001057926	7	58	314	12645
GO:0000786	Nucleosome	C	0.042820307	0.001186527	5	28	316	12675

GO-ID	Term	Category	FDR	P-Value	#Test	#Ref	#not in Test	#not in Ref
GO:0010413	glucuronoxylan metabolic process	P	1.05E-05	2.18E-08	18	127	308	12571
GO:0045492	xylan biosynthetic process	P	1.11E-05	2.43E-08	18	128	308	12570
GO:2000652	regulation of secondary cell wall biogenesis	P	1.94E-04	5.02E-07	5	3	321	12695
GO:0010089	xylem development	P	5.07E-04	1.51E-06	10	48	316	12650
GO:0009809	lignin biosynthetic process	P	0.00167049	5.41E-06	8	32	318	12666
GO:0016701	oxidoreductase activity, acting on single donors with incorporation of molecular oxygen	F	0.002428961	8.81E-06	10	60	3112638	
GO:0005506	iron ion binding	F	0.003377873	1.31E-05	21	269	305	12429
GO:0006624	vacuolar protein processing	P	0.003751016	1.55E-05	3	0	323	12698
GO:0016706	oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen, 2-oxoglutarate as one donor, and incorporation of one atom each of oxygen into both donors	F	0.004214725	1.86E-05	10	66	316	12632
GO:0055114	oxidation-reduction process	P	0.008119209	3.78E-05	64	1494	262	11204
GO:0050734	hydroxycinnamoyltransferase activity	F	0.012735066	6.10E-05	3	1	323	12697
GO:0016760	cellulose synthase (UDP-forming) activity	F	0.016620734	8.39E-05	6	24	320	12674
GO:0015103	inorganic anion transmembrane transporter activity	F	0.017624606	9.13E-05	9	65	317	12633
GO:0046274	lignin catabolic process	P	0.025032974	1.39E-04	5	16	321	12682
GO:0005576	extracellular region	C	0.033647111	1.96E-04	44	963	282	11735
GO:0071702	organic substance transport	P	0.045690319	2.75E-04	23	388	303	12310
GO:0009815	1-aminocyclopropane-1-carboxylate oxidase activity	F	0.045690319	2.94E-04	3	3	323	12695
GO:0016209	antioxidant activity	F	0.045690319	2.96E-04	10	95	316	12603
GO:0015706	nitrate transport	P	0.045690319	2.96E-04	10	95	316	12603

3.5.4 GO term enrichment correlates with phenotypic observations of gall tissue

3.5.4.1 Oak GO terms over-represented in early stage galls

Enriched GO terms demonstrate a clear distinction in expression between the early and later stages in the oak tissue. The early tissues have many GO terms associated with cell division and regulation, more so in the complete GO terms table. Galls grow rapidly at this stage; hence, much cell division is expected as indicated by GO terms for cytokinesis, mitosis specifically anaphase and metaphase, regulation of DNA replication and microtubule organisation. Inspection of the differentially expressed gene list *BLAST* annotations indicates Cyclin B to be highly expressed at this stage. Cyclin B is key to progression from the G_2 to M phase of the cell cycle. Additionally, *Q. robur* DNA is being heavily reorganized at this stage, through histone and DNA modification (methylation and alkylation), chromatin silencing and condensation of chromosomes as shown by enriched GO terms (table 3.22). The gene expression profiles neatly mirror the observed phenotype of *B. pallida* galls at this stage (contrast the sizes of sampled early and growth galls in figures 3.1-2). The gall tissue also shares expression with petal and sepal formation, which is not surprising as these tissues and gall wasp galls develop from meristematic tissue.

There are also hints at the cycles of endoreduplication (table 3.22, GO: 0042023) that occur in nutritive cells forming around the larval chamber during the early stage of gall development. In contrast to this, there is enrichment for chromatin silencing by small RNA and methylation (GO: 0031048 & 0006346) associated with the repression of gene expression. This may reflect gene silencing of certain genes for which high expression is not required in the endoreduplicated nutritive cells. Additionally, in the complete GO terms table for early versus later stages (appendix table 3.27) fatty acid biosynthesis is over-represented correlating with the observed lipid-dense nutritive cell cytoplasm (Harper et al., 2004).

3.5.5 GO terms over-represented in growth and mature stage galls

3.5.5.1 *Q. robur*

Whereas early oak tissue expression appears to be dominated by cytokinesis and associated processes, the later stage tissues are enriched for processes of maturation. Four of the five lowest adjusted P-values for enriched GO terms are involved in primary and secondary cell wall growth while the other is for xylem development. Both of these observations are in line with phenotypic observations. In growth stage galls, parenchyma vascularizes so that plant nutrients and water can be imported into the gall for the larva's benefit; this explains the enrichment for a xylem development GO term. It can also be seen in internal pictures of growth stage galls. Lignification of cell walls, in part, signifies gall progression to the mature stage, as evidenced by the papery epidermis of *B. pallida* galls and hardening of vascular tissues.

The other process enriched in growth and mature galls is oxidoreductase activity (appendix table 3.28). Although there are many processes that oxidoreductases could be involved in, including cell wall biogenesis, one in line with maturing plant tissues is the production of secondary plant compounds, such as phenolics and tannins. The presence of hydroxycinnamoyltransferase activity (GO:0050734), involved in phenylpropanoid metabolism, would support this.

3.5.5.2 *B. pallida*

It is harder to draw conclusions about the gall wasp enriched GO terms. It is more telling that early stage expression has few annotations, as this indicates that previously uncharacterized genes dominate expression.

The presence of sarcomere (GO:0030017) and muscle structure development (GO:0061061) GO terms indicates the larvae are growing in the growth and mature stages, which is obvious in mature replicates but not the growth stage. Other enriched GO terms are very general, with several translation and ribosomal associated GO terms as well a lipid and RNA binding and protein folding. This is to be expected if by this stage the larvae are no longer manipulating host expression, as the gall has finished growing, and are switching to feeding on the surrounding nutritive cells.

3.5.6 Comparison of differentially expressed *B. pallida* genes with the genomes of *Diplolepis spinosa* and *Belizinella gibbera*

As many of the genes differentially expressed are new, little can be said about their function. If gall induction by cynipid gall wasps occurs by the same core processes, it follows that orthologs of the key genes are expected to be present in species across the gall-inducing *Cynipidae*. By using the genome assembly of the rose gall wasp *Diplolepis spinosa* (Diplolepidini) (see Chapter 4 for assembly) conserved orthologs in another cynipid gall wasp tribe can be identified among differentially expressed genes. Additionally, the genome of the oak gall wasp *Belizinella gibbera* can be used to check concordance in gene complement among the tribe Cynipini.

The *B. pallida* transcriptome was compared to the cynipid genome assemblies. *BLASTn* nucleotide-nucleotide and *tBLASTx* translated-nucleotide-versus-translated-nucleotide searches were run. An e-value cut-off of 1 x 10^{-5} was applied and *BLASTn* and *tBLASTx* results combined for analysis. *TBLASTx* was chosen as the divergence time of *B. pallida* and *D. spinosa* is probably tens of millions of years (Ronquist and Liljeblad, 2001) and *tBLASTx* will compare 6-frame translations of both the query and reference. Consequently coding sequence will have diverged to the extent that protein sequence alone may identify some potential orthologs.

The *D. spinosa* assembly probably contains most of this species' complement of genes, as the *CEGMA* scores (Perra et al., 2009) are good at 80% complete (Chapter 4). In contrast, the *B. gibbera* genome is far from complete with scores of 25% complete for *CEG*s. Table 3.24 gives the number of genes in each differentially expressed category with a corresponding contig in the *D. spinosa* and *B. gibbera* genomes respectively.

Shared DE	**All DE**	**Higher in Early**	**> 2 fold change**	**Lower in Early**	**> 2 fold change**
D. spinosa	2310 (44%)	662 (23%)	186 (16%)	1257 (92%)	740 (91%)
B. gibbera	5122 (98%)	2914 (100%)	1137 (100%)	1298 (95%)	769 (95%)

Table 3.24. Differentially expressed gall wasp genes with hits in a closely related (*B. gibbera*) and phylogenetically distant (*D. spinosa*) gall wasp.

The *D. spinosa* sequence similarity results share the same pattern as that for the *BLAST*s against *nt* and *nr* datasets and GO enrichment (table 3.24). A low percentage, 23%, of the differentially and highly expressed genes in the early stage has hits to the *D. spinosa* genome. This is reduced further for those with greater than 2-fold change at 16%, but this does include the chitinase and serine threonine kinase genes. In contrast, genes more highly expressed in the growth and mature stages have over 90% hits to *D. spinosa*. The closely related *B. gibbera* however contains almost all differentially expressed *B. pallida* genes across stages.

3.6 Discussion: The induction and formation of *B. pallida* galls

In the discussion each of the pre-existing hypotheses are evaluated in the same order as the introduction, except for the NOD factors and arabinogalactan hypotheses (section 3.1). New hypotheses based on the gene expression patterns identified above are proposed. Finally, the NOD factors, arabinogalactan proteins and gall wasp chitinases are discussed at length as the most exciting results and potentially fruitful directions of future research into cynipid gall induction.

3.6.1 Previously identified candidates for gall induction

3.6.1.1 Virus-like-particles

No genes were found that indicated a virus-like-particle (VLP) being involved in the gall process. Neither were any found in a *Trinity* assembly of reads assigned to the Virus group. These transcripts derived principally from plant viruses with RNA genomes (table 3.8). A VLP is very unlikely to be transferring the key affecters of gall induction into the host during gall development. Additionally, VLPs are not oviposited with the egg based on venom gland transcriptomes of *B. pallida* and *Diplolepis rosae* and electron microscopy of the venom gland (S. Cambier, personal communication).

3.6.1.2 Secreted proteins

The vast majority of early differentially expressed *B. pallida* genes are predicted to encode a secretory peptide, but so do the majority of later stage differentially expressed genes. None of the early differentially expressed genes have homology to a gall inducing nematode or gall midge sequence by *BLAST* analysis (appendix table 3.26). These genes have very little homology to other known genes as indicated by the *BLAST2GO* annotations. This contrasts with genes differentially expressed highly in the growth and mature stages that mostly have orthologs in other insects. The genes involved in gall induction are possibly unique to the *Cynipidae*, having diverged from an ancestral sequence at the same time that gall wasps diverged from non-galling ancestors. Without a close outgroup genome sequence this is not possible to state with much confidence. The low number of shared orthologs with the *D. spinosa* genome suggests there may be differences in the number of genes involved in gall induction across tribes, with a core set of conserved genes. The products of these conserved genes are candidates for secretion from the larvae of materials driving interaction with the plant host, leading to the hypothesis:

> **Highly expressed genes in the early stage with signal peptides encode proteins secreted from the larva that can interact with host factors**

Proving function requires additional experimentation, but even without knowing their function they may be informative about other aspects of gall wasp larval biology. A key question is where are these genes synthesized in the gall wasp larva? There are two hypothetical origins for routes of egress from a gall wasp larva, the larval salivary glands or Malpighian tubules (Harper et al., 2009). The Malpighian tubules of cynipids are very different to other insects and are lined by secretory cells with polytene chromosomes (Harper et al., 2009). *In situ* or immunostaining approaches can be used to identify the origin and targets of secreted larval gall wasp proteins. This is discussed further in chapter 5. Interestingly two of the few early stage genes that are highly expressed, encode a putative signal peptide, and have strong similarities to genes in other insects, are the chitinases.

3.6.1.3 Potentially horizontally transferred plant cell wall degrading enzymes

Pectin lyases most similar to bacterial sequences, which degrade cell wall pectin, were differentially expressed in the early stage gall (discussed in chapter 4). Another pectin lyase was more highly differentially expressed in the mature stage. Two distinct cellulase genes are expressed highly throughout gall developmental stages but are not significantly differentially expressed in any direction. The cellulases therefore have some role albeit one not limited to gall induction. Potentially, they could aid in larval feeding by breaking down nutritive cell walls in the larval chamber, as could the pectin lyase. Alternatively the secretion of PCWDEs could weaken cell walls allowing the passage of induction-related factors into host tissues (Harper et al., 2009)

3.6.1.4 Plant hormone related genes

In the early stage, five oak genes associated with plant hormones are highly expressed. Two are ethylene responsive transcription factors and another, 1-aminocyclopropane-1-carboxylate synthase, is an important component of the ethylene synthesis pathway. The other two genes are a gibberellin 2-oxidase, which catabolizes gibberellin (Huang et al., 2010), and an auxin transporter protein that facilitates the intercellular flow of auxin.

Ten genes are differentially expressed higher in the growth and mature versus the early stage. They include distinct (by sequence) versions of the early stage differential expressed genes, including three 1-aminocyclopropane-1-carboxylate oxidases, and a gibberellin 2-oxidase. The other genes are four auxin-induced proteins, a giberrelin receptor and abscisic insensitive 1b.

That plant hormones are involved in gall growth has been hypothesized and a high concentration of the auxin indole acetic acid has been identified in gall tissues (Harper et al., 2009). The results presented here suggest roles for, principally, ethylene and auxin during gall development but RNAseq is less a powerful tool for investigating plant hormone levels in gall tissue than specific assays. Indeed, there were no plant hormone synthesis pathways enriched GO terms in any direction.

3.6.1.5 Galls-as-seeds: Biotin carboxyl carrier protein

Biotin carboxyl carrier protein (BCCP) has been previously identified at high expression levels in the nutritive cells surrounding gall wasp larvae for several species, including *B. pallida* (Harper et al., 2004). No genes annotated as BCCP were called as differentially expressed in this experiment. However, a BCCP annotated gene was expressed at high levels throughout gall development, corroborating the assay-based work of Harper et al. (2004). I also tested acetyl-CoA carboxylase genes (35 annotated by *BLAST*) for differential expression, as BCCP is a component of this multi-subunit enzyme. No differential expression of these genes was detected, although again expression was high for many transcripts in each developmental stage, indicating that galls have a high metabolic rate.

There are other markers that suggest similarity to seed tissues, in particular the extremely high and differential expression of late embryogenic protein 14 (*lea14*). This protein belongs to the late embryogenic proteins expressed in plant seed or stressed plant tissues (Hundertmark and Hincha, 2008). Unfortunately, the functional role of these proteins is unknown, although *lea14* may help cells avoid desiccation (Singh et al., 2005). If this were the case high expression of a *lea14*-like oak protein could be explained by desiccation avoidance in early gall tissue that has not yet vascularised. Other late embryogenesis-associated proteins are also highly expressed, but not differentially nor as extremely as *lea14*.

3.6.2 New hypotheses for gall induction

3.6.2.1 New hypothesis 1: Differentially expressed genes that are conserved across gall wasp tribes are candidates for genes with key roles in gall induction.

The low proportion of *B. pallida* early differentially expressed genes with similarity to genes in the *D. spinosa* genome assembly has two explanations.

Firstly, that few gall induction genes are indeed shared by *D. spinosa* and *B. pallida;* and those that are shared represent a core set of conserved genes. These 186 genes are candidates for further investigation. For example, a chitinase gene is present in the *D. spinosa* genome and homologous to both highly expressed *B. pallida* chitinases. Is this *D. spinosa* chitinase expressed in a similar pattern to the *B. pallida* chitinases? Alternatively, there could be many orthologs between the species, but sequence divergence has resulted in little to no identity in genes performing the same functions in the two hosts. Positive selection in tandem with divergence over at least 45 million years (Ronquist and Liljeblad, 2001) between these species, could potentially result in sufficient divergence that orthology is not detectable. In this case a gradual fall in identifiable potential orthologs would be expected as one compares species across greater phylogenetic distances. Adaptation to different hosts, the *Rosaceae* and *Fagaceae*, is a potential driver for positive selection among these genes. Adaptation could also be to specific host tissues; *B. pallida* sexual generation eggs are oviposited in apical meristem tissue of bud scales (Rey, 1992) while *D. spinosa* eggs are oviposited below the apical meristem in the cortex surrounding the procambium (Shorthouse et al., 2005). Differences in host expression profiles in the differing cell types in which galls are induced (apical versus cortical meristem) could drive evolution of induction genes.

3.6.2.2 New hypothesis 2: Nutritive cell gene expression results in endocycling of chromosomes while surrounding parenchymal gene expression drives rapid cell division.

In the early stage galls phenotypic observations indicate that rapid cell division is occurring in parenchyma tissues, while nutritive cells are endocycling (Harper et al., 2004; Harper et al., 2009). A GO term for endoreduplication is enriched at the early stage for oak that could explain the observed polytene chromosomes of gall nutritive cells. However, the high expression of cyclin B genes seen in the early stage gall tissue contradicts endoreduplication, as endoreduplication in plants bypasses the M-phase of mitosis which cyclin B proteins initiate in complex with cell-division kinases (CDKs) (Breuer et al., 2010). These contradictory, cyclin GO terms may represent processes occurring in the dividing parenchyma surrounding the nutritive cells. Therefore, the contradictory expression indicated by GO term enrichment may result from distinct expression between nutritive and non-nutritive parenchymal tissues. This experiment does not discriminate between nutritive and parenchymal tissue, and thus GO terms cannot be partitioned by cell type.

To test this hypothesis future experiments could compare expression between nutritive cells and surrounding parenchyma. The expectation would be for GO terms associated with endoreduplication to occur in nutritive cells only, while cell division GO terms would be highly expressed in surrounding parenchyma.

3.6.2.3 New hypothesis 3: Gall enlargement occurs by early cell division followed by rapid cell expansion in the growth stage

The expression of many cell division GO terms and genes in the early stage gall tissue versus the later stages makes intuitive sense. But the gall grows rapidly during the growth stage so why are cell division associated genes not highly expressed in this stage as well? It may be that the size constraint on selecting growth stage galls meant that most of the cell differentiation and growth had been completed in the sample growth stage galls. An alternative explanation is that cell division and cell expansion are distinct to the early and growth stages. In this scenario, cell division occurs early in gall division; the inference would then be that observed rapid gall growth is driven predominantly by cell expansion not cell division. The GO enrichment results are very clear on cell division being an early stage gall process. Therefore observing cell expansion, by microscopy, of gall tissues, without cell division in growth stage galls would provide evidence for this hypothesis.

3.6.2.4 New Hypothesis 4: Gall wasp chitinases modify host arabinogalactans, resulting in somatic embryogenesis-like dedifferentiation and cell division in host tissues.

Differentially expressed arabinogalactan protein expression in oak tissue suggests a role for these *B. pallida* chitinases only described before for plant chitinases. Gall wasp chitinase could act on host arabinogalactan proteins (AGPs) in the extracellular matrix in a way analogous to the action of endogenous plant chitinases (van Hengel et al., 2001) by enhancing somatic embryogenesis due to cleavage of arabinogalacatan chains of AGPs. The gall wasp chitinases contain the required signal peptides for secretion into the extracellular milieu of the gall wasp larva. The high-expression of these chitinase genes compared to most other differentially expressed gall wasp genes is striking. Their possible substrate, an early nodulin associated AGP is also differentially and highly expressed by the host in early stage tissue. This AGP is multi-domain containing a phytocyanin domain as well as the AGP backbone. Poon et al. (2012) showed AGPs with this domain are capable of promoting somatic embryogenesis. The AGP identified in this experiment appears similar to that identified by Poon et al., (2012) named GhPLA1 (*Gossypium hirsute* phytocyanin-like arabinogalactan protein1) (*BLASTx* bit score of 331).

The two gall wasp larval chitinases are quite divergent from one another with only 60% amino acid identity; they both however have best *BLASTx* hits to the same *Nasonia vitripennis* chitotriosidase-1-like protein (gene ID: 345494134). The insect, family 18 glycosyl-hydrolase, chitinases normally function in turnover of extracellular chitin-containing matrices such as the insect cuticle (Arakane and Muthukrishan, 2009). Some insect chitinases are important to larval and pupal molting in *Tribolium castaneum* (Zhu et al., 2008). The two *B. pallida* chitinases have little to no expression in the mature galls when one would expect high cuticle chitinase activity as larvae prepare for pupation, although the experimental design may not be sensitive enough to detect such expression.

The gall wasp chitinase cleaving host arabinogalactans hypothesis is attractive as evolutionary predictions for the origin of this interaction make intuitive sense. Cynipid gall wasps probably evolved from parasitoids of wood-boring larvae (Ronquist and Liljeblad, 2001). In such a scenario, the ancestral gall wasp chitinase could have been used to degrade the ancestral parasitoids host-larval cuticle and would have come in to contact with host plant cells. Host cells may have then reacted to this unintended stimulus in the manner discussed above, forming somatic embryogenic calli. Such an accidental interaction could have been one of the earliest steps in the evolution of cynipid gall induction.

New hypothesis 4 makes testable predictions:

1. **Orthologs of the *B. pallida* chitinases will be present in other gall wasps and have similar expression patterns.**
2. **Incubation of oak gall wasp chitinases with host arabinogalactan proteins will increase the somatic embryogenesis potential of the AGPs.**

These hypotheses do not address how the proposed key arabinogalactan proteins came to be highly expressed in oak tissue. In short, if the hypothesis is correct, then for the chitinases to be effective the substrate arabinogalactan proteins have to be expressed. However, this observation may explain the importance of phenology to gall induction. Both van Hengel et al., (2001) and Poon et al., (2012) demonstrated that temporal expression of AGP is key to somatic embryogenesis. In cotton, the early somatic embryogenesis expressed GhPLA1 is required for somatic embryogenesis, while a different AGP expressed late in somatic embryogenesis was inhibitory to initiating somatic embryogenesis (Poon et al., 2012). Van Hengel et al., (2001) also observed this effect, although the involved AGPs were not characterized. Transient early AGP expression may represent, along with other as yet unidentified host genes, the genetic basis of the ‘window-of-opportunity’ *B. pallida* females have to successfully exploit to induce a gall on oak. If GhPLA1 is a conserved target for manipulation this statement may apply to all cynipids, leading to another new hypothesis:

> **New hypothesis 5: transient gene expression of key host genes, including orthologs of GhPLA1, in cells that go on to form a gall has driven specificity in gall wasp oviposition timing.**

It has been observed that many *B. pallida* gall wasps will preferentially oviposit on the tree they emerged from (Egan and Ott, 2007). This is consistent with a synchronizing of host phenology and cynipid oviposition timing. Cynipid gall wasps may predominately manipulate already occurring processes in meristematic host tissues as opposed to initiating them, representing a ‘simpler’ strategy of host manipulation.

3.6.2.5 New hypothesis 6: Somatic embryogenesis is induced in host apical meristem to initiate *B. pallida* gall development.

The early stages of gall induction demonstrate similarity to plant expression during somatic embryogenesis. Expression of genes known to be involved in somatic embryogenesis was investigated to find further evidence for this hypothesis. None of the enriched GO terms (full or specific) directly addressed somatic embryogenesis. Five GO terms were enriched in the early stage for post-embryonic development. These GO terms are in the same hierarchy for somatic embryogenesis. It was, however, more fruitful to look at differential expression of genes associated with somatic embryogenesis, such as somatic embryogenesis receptor kinase (*SERK*) (Karami et al., 2009).

Somatic embryogenesis receptor kinase (*SERK*) genes are expressed in early embryogenic cells of *Arabidopsis thaliana, Zea mays, Medicago trunculata, Oryza sativa, Theobroma cacao, Citrus unshiu*, and others (Karami et al., 2009). *SERK* gene overexpression increases the chance that a cell will undergo somatic embryogenesis (Karami et al., 2009). The genes encode leucine-rich repeat receptor-like kinases (*LRR-RLKS*) involved in plant signaling. One SERK gene is differentially expressed highly in the early stage of gall tissue, while several others are highly expressed throughout induction. Several other *SERK*s have high expression throughout induction but are not differentially expressed in any direction. *SERK1* has been found in complexes with brassinosteroid-insensitive 1 and its associated receptor kinase (Karami et al., 2009). This is mirrored in the oak gene early transcriptome, with one brassinosteroid-insensitive 1 associated receptor kinase differentially expressed more highly in the early stage. Again, several other brassinosteroid-insensitive 1 genes are highly expressed throughout gall development and are not differentially expressed. Binding of brassinosteroid, a plant hormone, to the *SERK*/brassinosteroid-insensitive1 receptor-kinase complexes triggers transcription of embryogenesis related genes (Karami et al., 2009). There are many other genes involved in somatic embryogenesis highly expressed during gall induction, including Glutathione s-transferases, *CLAVATA* receptors, the embryogenic cell receptor, *ECPP44*; apetala3, *WUSCHEL*, transport inhibitor proteins and other arabinogalactan proteins without phytocyanin domains (Karami et al., 2009).

This 'galls-as-somatic embryos' hypothesis is complementary to the 'galls-as-seeds' hypothesis of Harper et al., (2004). The somatic embryo hypothesis addresses the earliest stage of induction whereas the 'galls-as-seeds' hypothesis was based on studies of later tissues (Harper et al., 2004). To investigate the similarity between gall induction and somatic embryogenesis a comparison of gall tissue expression with that of oak cells undergoing somatic embryogenesis such as induced callus tissues or early acorn is recommended.

An alternative explanation to somatic embryogenesis-like expression by oak tissues is that expression common to somatic embryogenesis and maintaining the apical meristem, on which *B. pallida* sexual generation galls are initiated, is confounded. The *CLAVATA* and *WUSCHEL* genes, for example, are key regulators of cell fate in shoot apical meristem signaling in *Arabidopsis thaliana* (Schoof et al., 2000; Barton, 2010). To address this, *B.*

pallida gall gene expression should be compared to non-embryogenic meristematic tissue to assess the continuity of expression in gall tissue from the originating meristem.

3.7 Conclusions

The RNAseq experiment has identified strong candidates for control of gall development in both oak and gall wasp. It has also helped identify a potential direct interaction between oak and gall wasps in the early gall. However, many gall wasp genes identified as candidates have unknown functions as they lack identifiable homologs outside the *Cynipidae*, and potentially within the *Cynipidae*. The principal limitation of the experiment was the lower depth of gall wasp sequencing in each replicate. This did not affect the assembly, which is of good quality, or the ability to annotate *B. pallida* transcripts. The gall wasp chitinases identified are hypothesized to act directly on oak arabinogalactan proteins. The published roles of chitinases in interaction with plant arabinogalactans in initiating somatic embryogenesis prompted further investigation, as an arabinogalactan protein associate with early nodulation was highly expressed in early galls. As a result, gall induction is hypothesized to involve expression pathways commonly found during somatic embryogenesis. Many of the phenotypic observations of the initiation of gall induction are analogous to somatic embryogenesis. As previously predicted (Stone & Schönrogge, 2003) gall wasps do appear to manipulate highly conserved plant developmental pathways. Potential further experiments based on new hypotheses of gall induction are proposed in the final chapter on future research.

3.8 Appendix

	1	**4**	**8**	**211**	**127**	**148**	**182**	**224**	**234**	**252**	**270C**	**281**
comp327188_c0	1	0	0	0	0	0	2	0	0	0	1	0
comp327193_c0	2	0	0	0	0	1	0	0	0	0	0	0
comp327195_c0	1	0	0	0	1	1	0	0	0	0	0	0
comp32719_c0	124	59	35	45	41	47	72	24	141	97	121	69
comp327222_c0	2	3	0	0	0	0	0	0	0	0	1	1
comp327227_c0	1	1	0	1	0	0	0	0	0	0	0	0
comp32722_c0	37	23	13	13	13	28	25	3	51	71	88	37
comp327246_c0	1	0	0	0	0	1	0	0	0	0	2	0
comp327253_c0	2	0	0	1	0	0	0	0	0	0	0	0
comp327279_c0	1	0	0	0	0	0	0	0	0	1	0	0
comp32729_c0	86	36	20	22	14	39	31	10	42	22	30	13
comp327305_c0	1	0	0	0	0	0	0	0	1	0	0	0
comp32733_c0	82	54	38	145	44	30	54	21	195	192	241	121
comp327344_c0	1	1	0	0	0	0	0	0	0	0	0	0
comp327356_c0	3	0	1	0	0	0	0	0	0	0	0	0
comp327378_c0	1	0	0	0	0	0	0	0	0	0	2	0
comp327394_c0	2	0	1	0	0	1	0	0	0	0	0	0
comp327399_c0	3	2	0	1	0	0	0	0	0	0	0	0
comp327401_c0	1	0	1	0	0	0	0	0	0	0	2	0

Table 3.25. Example of raw counts per gene per replicate generated by RSEM.

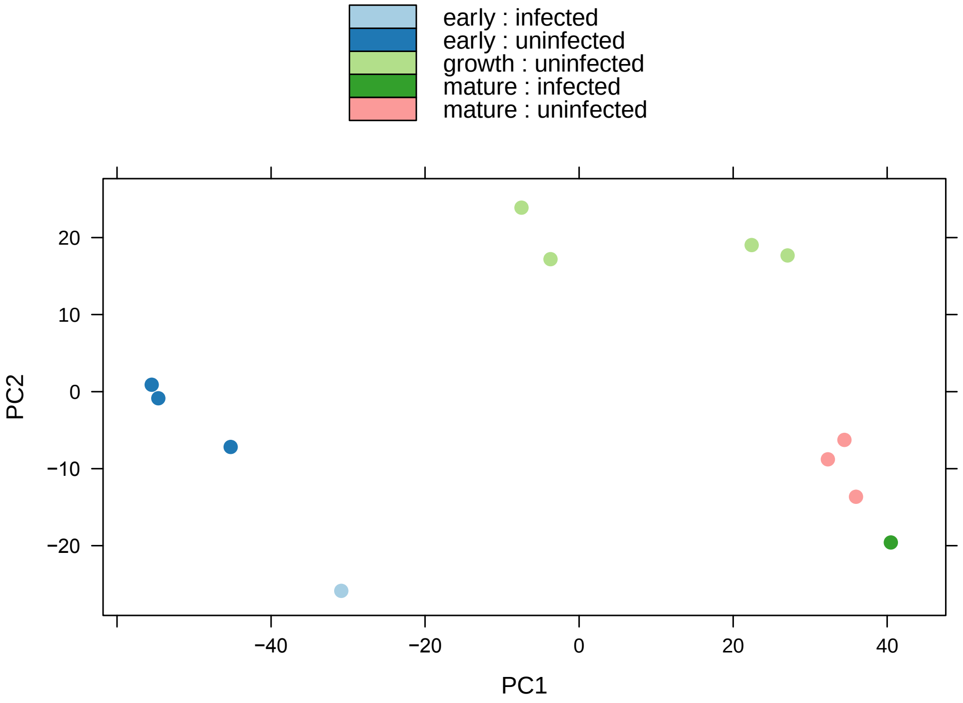

Figure 3.15. Principal components analysis of Plants data filtered by edgeR demonstrating implicit filtering of fungal infection affected gene expression.

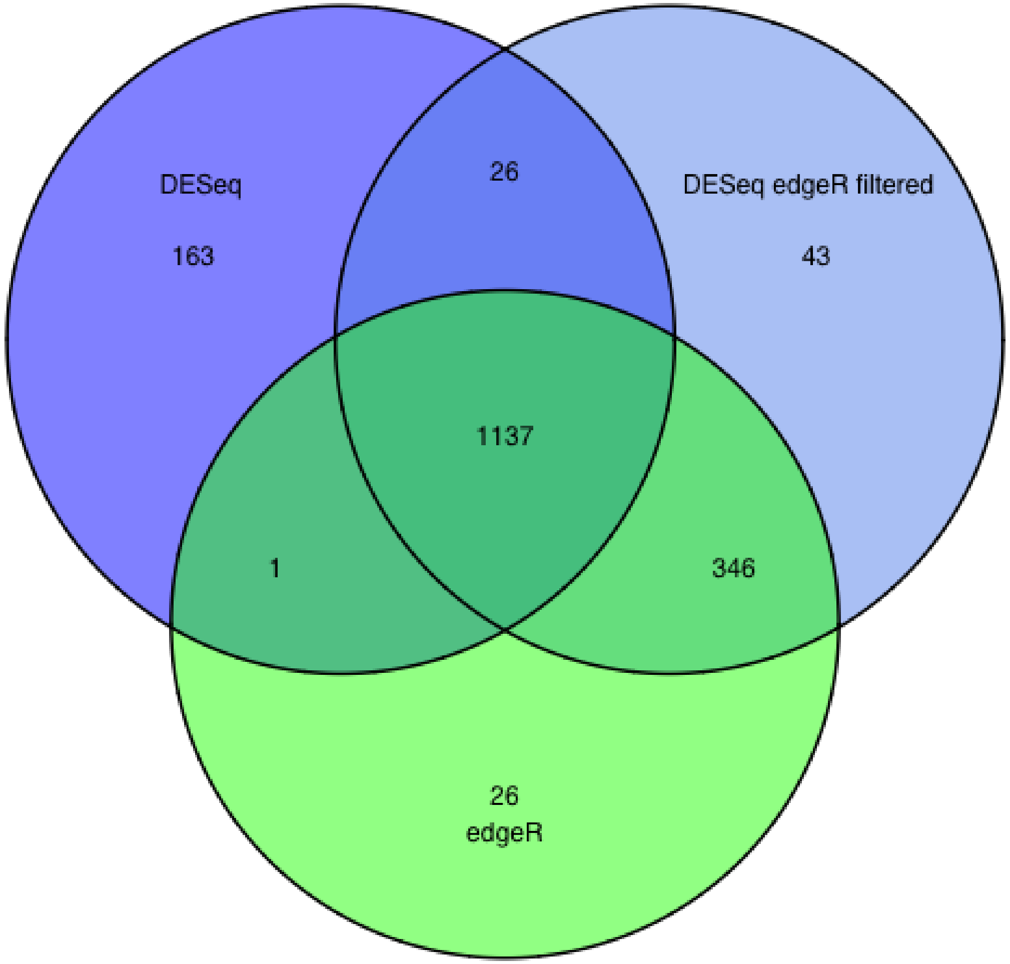

Figure 3.16. Venn diagram for Plants of differentially expressed genes exhibiting greater than 2 fold up or down changes in expression.

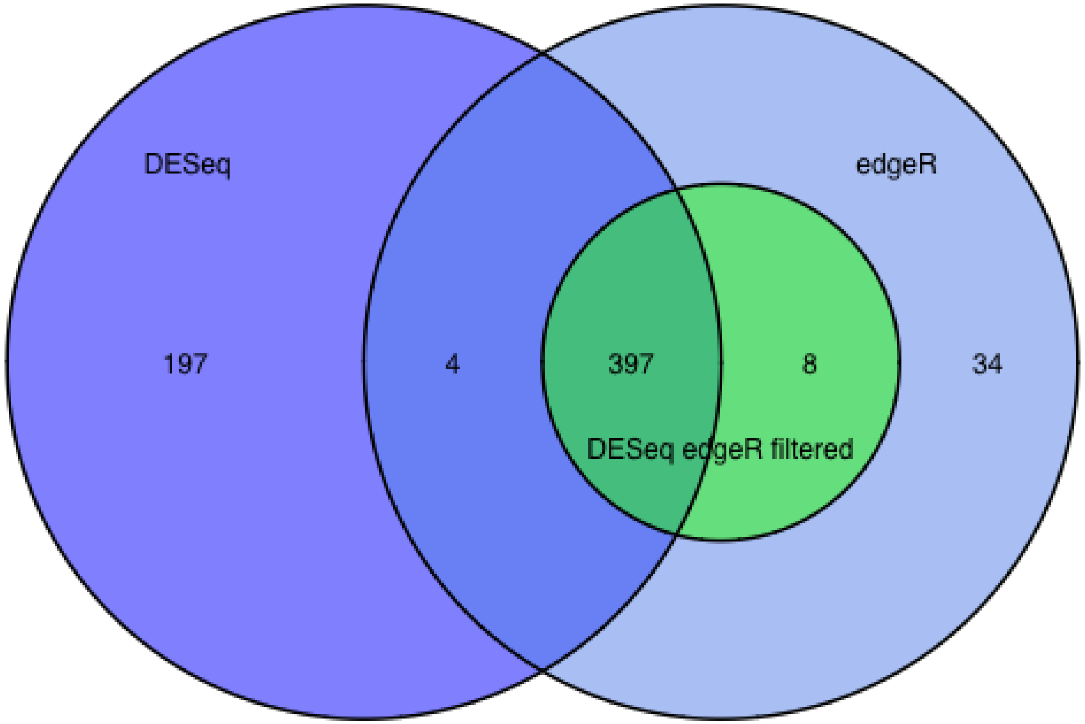

Figure 3.17. Venn diagram for Arthropoda of differentially expressed genes exhibiting greater than 2 fold up or down changes in expression.

Table 3.26. Most similar sequences from *BLAST* seqrches of *B. pallida* genes differentially highly expressed in the early stage.

Table 3.27. Complete GO terms for early stage highly expressed *Q. robur* genes. F = molecular function; C = cellular component; P = biological process. #Test = number of differentially expressed genes for this GO annotation; #Ref number of genes for this GO annotation in the reference, not including differentially expressed genes; #not in test number of differentially expressed genes not in this GO annotation; #not in Ref number of genes that do not have this GO annotation in the reference.

Table 3.28. Complete GO terms for growth and mature stage highly expressed *Q. robur* genes. F = molecular function; C = cellular component; P = biological process. #Test = number of differentially expressed genes for this GO annotation; #Ref number of genes for this GO annotation in the reference, not including differentially expressed genes; #not in test number of differentially expressed genes not in this GO annotation; #not in Ref number of genes that do not have this GO annotation in the reference.

Query	%	Length	Query start	Query end	E-value	Bit score	Target description
comp27668_c0	80.95	2320	844	3152	0	2179	Apis mellifera FoxP protein (Foxp), mRNA
comp23205_c0	75.08	1637	200	1836	0	1113	PREDICTED: Apis mellifera solute carrier family 23 member 2-like (LOC410114), mRNA
comp74505_c0	70.72	485	2261	810	0.00E+00	667	PREDICTED: hypothetical protein LOC409020 [Apis mellifera]
comp23521_c0	86.12	497	6	502	2.00E-163	585	PREDICTED: Nasonia vitripennis charged multivesicular body protein 4b-like (LOC100123786), mRNA
comp75244_c0	68.31	1537	349	1835	7.00E-148	535	PREDICTED: Bombus terrestris protein krueppel-like (LOC100642205), mRNA
comp13050_c0	50	448	42	1382	1.00E-120	438	UDP-glucuronosyltransferase 2B15 [Harpegnathos saltator]
comp95519_c0	43.19	382	519	1661	1.00E-91	342	PREDICTED: cytochrome P450 4C1 [Nasonia vitripennis]
comp17344_c0	32.44	669	2060	126	1.00E-86	326	hypothetical protein AaeL_AAEL005543 [Aedes aegypti]
comp25856_c0	44.35	345	134	1126	2.00E-80	304	PREDICTED: venom acid phosphatase Acph-1-like isoform 1 [Nasonia vitripennis]
comp13718_c0	39.86	424	428	1597	6.00E-67	260	PREDICTED: hypothetical protein LOC100643835 [Bombus terrestris]
comp20790_c0	45.66	311	134	1051	2.00E-61	241	Pectate lyase/Amb allergen [Dickeya dadantii Ech703]
comp28991_c0	47.84	255	57	800	2.00E-61	241	serine protease 4 precursor [Nasonia vitripennis]
comp27433_c0	39.78	357	1124	81	4.00E-59	234	PREDICTED: chitotriosidase-1-like [Nasonia vitripennis]
comp68430_c0	45.49	255	824	69	1.00E-57	228	PREDICTED: elongation of very long chain fatty acids protein 1-like [Apis mellifera]
comp28195_c0	41.01	356	291	1328	1.00E-54	219	PREDICTED: chitotriosidase-1-like [Nasonia vitripennis]
comp15877_c0	51.87	187	273	815	3.00E-49	201	PREDICTED: protein canopy-1-like [Bombus terrestris]
comp12097_c0	38.83	273	151	963	2.00E-44	185	PREDICTED: glycine N-acyltransferase-like protein 3-like [N. vitripennis]

comp115223_c0	40.08	252	190	945	3.00E-44	183	hypothetical protein TcasGA2_TC010509 [Tribolium castaneum]
comp104770_c0	36.23	403	7	1080	2.00E-35	155	PREDICTED: hypothetical protein LOC410107 [Apis mellifera]
comp23773_c0	72.56	266	639	904	2.00E-32	150	Drosophila mojavensis GI24266 (Dmoj\GI24266), mRNA
comp12191_c0	76.14	197	435	631	2.00E-30	143	Drosophila mojavensis GI21819 (Dmoj\GI21819), mRNA
comp26235_c0	72.95	244	159	402	8.00E-31	143	Mantispa pulchella clone Mp1 mariner transposase pseudogene, complete cds
comp28222_c0	43.69	206	855	271	4.00E-32	143	hypothetical protein SINV_08289 [Solenopsis invicta]
comp9488_c1	72.73	275	694	959	2.00E-30	143	Plasmodium knowlesi strain H chromosome 7, complete genome
comp17168_c0	72.31	260	249	507	9.00E-30	141	Herpetosiphon aurantiacus DSM 785, complete genome
comp63126_c0	30.92	304	190	1092	5.00E-31	140	Regucalcin [Camponotus floridanus]
comp23026_c1	46.1	141	5	403	3.00E-31	138	teratocyte released chitinase [Toxoneuron nigriceps]
comp30525_c0	76.47	170	280	449	1.00E-25	127	Emiliania huxleyi virus 86 isolate EhV86
comp28027_c0	32.58	267	1041	271	1.00E-26	125	carbonic anhydrase [Aedes aegypti]
comp26048_c0	68.52	324	3	324	4.00E-23	118	Nasonia vitripennis BAC NV_Bb-46A12 (Clemson University Genomics Nasonia vitripennis BAC Library) complete sequence
comp43152_c0	57	100	71	361	4.00E-24	114	PREDICTED: venom allergen 5-like [Nasonia vitripennis]
comp146762_c0	72.65	223	531	751	1.00E-20	111	Pleistodontes nigriventris clone 31.2 transposon mariner nonfunctional transposase protein gene, partial sequence
comp27060_c0	26.97	304	338	1231	5.00E-22	111	Regucalcin [Camponotus floridanus]
comp16537_c1	34.68	173	1219	1737	2.00E-21	109	reverse transcriptase, putative [Pediculus humanus corporis]
comp45923_c0	67.72	443	419	846	7.00E-20	109	Drosophila mojavensis GI22470 (Dmoj\GI22470), mRNA
comp24545_c0	15.43	350	9	1004	1.00E-20	105	late embryogenesis abundant-like protein 1 [Brachionus plicatilis]

comp65698_c0	71.57	197	367	563	1.00E-18	104	Candida tropicalis MYA-3404 predicted protein, mRNA
comp41058_c0	31.35	185	50	592	4.00E-20	103	PREDICTED: apolipoprotein D-like [Bombus impatiens]
comp115810_c0	51.96	102	1241	1546	8.00E-45	102	PREDICTED: tigger transposable element derived 6-like [Saccoglossus kowalevskii]
comp15389_c0	31.16	199	30	608	1.00E-19	101	hypothetical protein I79_025492 [Cricetulus griseus]
comp45337_c0	75	152	165	312	4.00E-17	100	PREDICTED: Acyrthosiphon pisum glutathione S-transferase-like (LOC100570856), mRNA
comp8141_c0	41.77	79	520	284	3.00E-18	97.1	hypothetical phage protein [Campylobacter phage CP220]
comp76905_c0	79.09	110	166	275	7.00E-16	95.1	Drosophila mojavensis GI19857 (Dmoj\GI19857), mRNA
comp61656_c0	26.87	268	91	873	2.00E-17	94.7	putative trypsin 2 [Phlebotomus perniciosus]
comp8386_c0	38.18	110	119	448	7.00E-17	93.6	biotin carboxylase subunit of acetyl CoA carboxylase [Plasmodium vivax Sal-1]
comp26053_c0	32.59	224	160	759	1.00E-16	91.7	APEG precursor protein [Xenopus laevis]
comp77120_c0	77.39	115	396	510	7.00E-15	91.5	Dictyostelium discoideum DrnA gene for putative RNaseIII
comp16621_c0	79.69	128	73	194	5.00E-14	89.7	PREDICTED: Nasonia vitripennis hexokinase type 2-like, transcript variant 2 (LOC100121683), mRNA
comp18400_c0	100	47	1485	1531	7.00E-13	86	Hordeum vulgare subsp. vulgare cDNA clone: FLbaf144f19, mRNA sequence
comp25915_c0	74.24	132	82	213	7.00E-13	84.2	Leishmania braziliensis MHOM/BR/75/M2904 hypothetical protein (LbrM03_V2.0720) partial mRNA
comp12159_c0	29.04	272	1710	949	1.00E-13	83.6	GK10310 [Drosophila willistoni]
comp23087_c0	25	268	476	1249	1.00E-13	82.8	PREDICTED: suprabasin-like, partial [Ornithorhynchus anatinus]
comp8146_c0	32.12	165	733	1227	3.00E-13	82.8	reverse transcriptase, putative [Pediculus humanus corporis]

comp23359_c0	29.48	173	557	69	7.00E-14	81.3	PREDICTED: similar to CG11284 CG11284-PA [Tribolium castaneum]
comp13566_c0	83.12	77	2	78	2.00E-11	80.6	Vitis vinifera contig VV78X029495.10, whole genome shotgun sequence
comp62785_c0	69.04	197	457	653	2.00E-11	80.6	Zebrafish DNA sequence DKEY-80G18 in linkage group 15,
comp25803_c0	70.55	163	489	651	8.00E-11	78.8	C.jacchus DNA sequence from clone CH259-147B13, complete sequence
comp85100_c0	27.44	277	260	1027	2.00E-11	75.5	hypothetical protein AND_12620 [Anopheles darlingi]
comp379680_c0	45.78	83	2	250	4.00E-12	75.1	Acidic mammalian chitinase [Harpegnathos saltator]
comp22909_c0	42.31	104	4	312	1.00E-11	73.2	PREDICTED: similar to chitinase 6 [Tribolium castaneum]
comp17324_c1	32.47	154	211	672	3.00E-10	71.6	PREDICTED: hypothetical protein LOC100561123 [Anolis carolinensis]
comp85680_c0	28.91	211	94	708	1.00E-10	71.6	PREDICTED: regucalcin-like [Bombus impatiens]
comp100285_c0	36.89	103	236	544	4.00E-10	71.2	hypothetical protein EAG_10027 [Camponotus floridanus]
comp59931_c0	50.82	61	3	185	6.00E-11	71.2	pectin lyase [Bacillus subtilis subsp. spizizenii ATCC 6633]
comp63509_c0	25	80	159	398	1.00E-10	71.2	conserved Plasmodium protein [Plasmodium falciparum 3D7]
comp23065_c0	33.06	124	370	8	1.00E-10	70.1	carbonic anhydrase [Aedes aegypti]
comp52503_c0	83.33	72	513	583	2.00E-08	69.8	Candidatus Carsonella ruddii PV DNA, complete genome
comp16643_c0	29.05	210	274	873	2.00E-09	69.3	trypsin precursor MDP5A [Mayetiola destructor]
comp44651_c0	26.99	226	873	256	8.00E-10	69.3	PREDICTED: hepatocyte growth factor-like isoform 1 [Equus caballus]
comp26084_c0	40.48	84	331	98	4.00E-10	68.6	chitinase [Danaus plexippus]
comp27724_c0	37.04	108	300	1	8.00E-10	67.4	carbonic anhydrase 6 precursor, putative [Pediculus humanus corporis]
comp15631_c0	84.72	72	118	189	4.00E-07	66.2	Mouse DNA sequence from clone RP23-291H20 on chromosome 2 Contains the 3' end of a novel gene, complete sequence
comp26225_c0	85.71	56	674	729	3.00E-07	66.2	Cyprinid herpesvirus 3 DNA, complete genome, strain: TUMST1

comp21444_c0	26.26	179	263	760	1.00E-08	65.9	Hypothetical protein CBG09235 [Caenorhabditis briggsae]
comp19164_c0	26.11	180	144	680	1.00E-08	65.5	PREDICTED: hypothetical protein LOC100679084 [Nasonia vitripennis]
comp22816_c0	29.66	145	811	1218	7.00E-08	64.7	hypothetical protein CPAR2_502130 [Candida parapsilosis]
comp10926_c0	86.79	53	491	543	1.00E-06	64.4	Dictyostelium discoideum snwA gene, complete cds
comp35880_c0	73.43	143	196	334	1.00E-06	64.4	Plasmodium falciparum 3D7 chromosome 9
comp46560_c0	76.11	113	51	159	7.00E-07	64.4	Zebrafish DNA sequence from clone CH211-87D5 in linkage group 7, complete sequence
comp6997_c0	72.73	110	336	445	1.00E-06	64.4	Thielavia terrestris NRRL 8126 chromosome 6, complete sequence
comp73144_c0	77.11	83	442	524	1.00E-06	64.4	Mouse DNA sequence from clone RP23-304H7 on chromosome 11 Contains a COMM domain containing 9 (Commd9) pseudogene, a novel gene and the Tcf2 gene for transcription factor 2, complete sequence
comp75670_c0	73.87	111	352	459	1.00E-06	64.4	Homo sapiens BAC clone RP11-320M2 from 2, complete sequence
comp20945_c0	62.5	56	461	294	1.00E-08	63.5	PREDICTED: hypothetical protein LOC100743197 [Bombus impatiens]
comp24597_c2	34.95	103	2	295	1.00E-08	63.5	Acidic mammalian chitinase [Camponotus floridanus]
comp11515_c0	31.78	129	2	382	3.00E-08	63.2	carbonic anhydrase II [Culex quinquefasciatus]
comp27958_c0	63.83	47	394	534	2.00E-07	63.2	hypothetical protein EAG_09607 [Camponotus floridanus]
comp107313_c0	26.67	180	546	1049	1.00E-07	62.8	hypothetical protein TcasGA2_TC004227 [Tribolium castaneum]
comp15699_c0	82.26	62	318	379	4.00E-06	62.6	PREDICTED: Strongylocentrotus purpuratus similar to 5-amp-activated protein kinase, beta subunit (LOC764925), mRNA
comp15808_c0	82.86	70	14	81	3.00E-06	62.6	Zebrafish DNA sequence from clone CH211-123B7 in linkage group 22 Contains the 5' end of the gene for a novel protein similar to vertebrate chondroitin sulfate proteoglycan family (CSPG) and three CpG islands, complete sequence
comp66187_c0	83.87	62	37	94	4.00E-06	62.6	Zebrafish sequence clone CH211-84M6 in linkage group 17, complete

comp94358_c0	36.62	71	1038	1244	4.00E-07	61.2	hypothetical protein EAG_01936 [Camponotus floridanus]
comp10820_c0	44.07	59	179	3	1.00E-07	60.5	IP20720p [Drosophila melanogaster]
comp21532_c0	34.15	123	5	346	1.00E-07	60.5	carbonic anhydrase 6 precursor, putative [Pediculus humanus corporis]
comp16052_c0	50	60	188	9	1.00E-07	60.1	predicted protein [Nematostella vectensis]
comp7526_c0	21.84	261	810	127	1.00E-06	59.7	hypothetical protein DDB_G0293586 [Dictyostelium discoideum AX4]
comp23869_c0	33.33	123	6	359	2.00E-07	59.3	Carbonic anhydrase 7 [Harpegnathos saltator]
comp7902_c0	27.22	180	976	485	1.00E-06	59.3	hypothetical protein [Paramecium tetraurelia strain d4-2]
comp25474_c0	45.79	107	2	322	2.00E-06	58.5	hypothetical protein EAG_12773 [Camponotus floridanus]
comp19740_c0	39.39	99	618	866	4.00E-06	57.4	hypothetical protein TcasGA2_TC002128 [Tribolium castaneum]
comp22015_c0	35.56	135	3	401	1.00E-06	57	hypothetical protein KGM_07092 [Danaus plexippus]
comp49700_c0	33.03	109	385	74	2.00E-06	57	conserved hypothetical protein [Streptomyces clavuligerus ATCC 27064]
comp56586_c0	26.73	202	209	754	4.00E-06	57	GF17129 [Drosophila ananassae]
comp32925_c0	52	50	258	115	2.00E-06	56.2	PREDICTED: hypothetical protein LOC100679142 [Nasonia vitripennis]
comp24290_c0	48.28	58	572	745	9.00E-06	55.8	predicted protein [Nematostella vectensis]
comp7206_c0	41.25	80	270	34	2.00E-06	55.8	carbonic anhydrase [Clonorchis sinensis]

GO-ID	Term	Category	FDR	P-Value	#Test	#Ref	#not in Test	#not int Ref
GO:0008283	cell proliferation	P	1.32E-35	1.71E-39	56	181	265	12522
GO:0051301	cell division	P	2.60E-34	6.73E-38	67	324	254	12379
GO:0000910	cytokinesis	P	9.01E-34	3.50E-37	54	183	267	12520
GO:0033205	cell cycle cytokinesis	P	2.34E-32	1.21E-35	50	158	271	12545
GO:0000911	cytokinesis by cell plate formation	P	7.76E-32	5.02E-35	49	154	272	12549
GO:0007049	cell cycle	P	1.28E-31	9.95E-35	89	725	232	11978
GO:0051322	anaphase	P	4.81E-29	4.36E-32	37	75	284	12628
GO:0007017	microtubule-based process	P	9.04E-29	9.37E-32	55	256	266	12447
GO:0022402	cell cycle process	P	6.14E-27	7.16E-30	77	625	244	12078
GO:0000226	microtubule cytoskeleton organization	P	3.33E-26	4.32E-29	48	206	273	12497
GO:0051567	histone H3-K9 methylation	P	1.48E-25	2.11E-28	42	148	279	12555
GO:0016570	histone modification	P	2.78E-25	4.33E-28	57	338	264	12365
GO:0022403	cell cycle phase	P	1.48E-24	2.49E-27	61	412	260	12291
GO:0016568	chromatin modification	P	2.46E-24	4.45E-27	61	417	260	12286
GO:0016569	covalent chromatin modification	P	2.95E-24	5.73E-27	59	388	262	12315
GO:0000279	M phase	P	3.35E-23	6.95E-26	52	304	269	12399
GO:0006325	chromatin organization	P	3.70E-23	8.14E-26	64	493	257	12210
GO:0006275	regulation of DNA replication	P	1.22E-22	2.83E-25	33	91	288	12612
GO:0034968	histone lysine methylation	P	3.07E-22	7.55E-25	44	213	277	12490

GO:0051726	regulation of cell cycle	P	1.01E-21	2.62E-24	43	208	278	12495
GO:0016572	histone phosphorylation	P	5.92E-21	1.61E-23	25	42	296	12661
GO:0051052	regulation of DNA metabolic process	P	9.51E-21	2.71E-23	37	150	284	12553
GO:0016571	histone methylation	P	1.65E-19	4.91E-22	44	256	277	12447
GO:0006260	DNA replication	P	2.38E-19	7.38E-22	44	259	277	12444
GO:0051276	chromosome organization	P	3.17E-19	1.03E-21	68	675	253	12028
GO:0006479	protein methylation	P	3.61E-19	1.26E-21	44	263	277	12440
GO:0008213	protein alkylation	P	3.61E-19	1.26E-21	44	263	277	12440
GO:0010389	regulation of G2/M transition of mitotic cell cycle	P	2.20E-18	7.97E-21	22	37	299	12666
GO:0000086	G2/M transition of mitotic cell cycle	P	3.28E-18	1.23E-20	22	38	299	12665
GO:2000602	regulation of interphase of mitotic cell cycle	P	7.35E-18	2.85E-20	22	40	299	12663
GO:0006270	DNA-dependent DNA replication initiation	P	1.06E-17	4.25E-20	23	48	298	12655
GO:0006306	DNA methylation	P	4.13E-17	1.77E-19	33	151	288	12552
GO:0006305	DNA alkylation	P	4.13E-17	1.77E-19	33	151	288	12552
GO:0006304	DNA modification	P	5.68E-17	2.50E-19	33	153	288	12550
GO:0007346	regulation of mitotic cell cycle	P	1.90E-16	8.60E-19	25	74	296	12629
GO:0048451	petal formation	P	3.38E-16	1.58E-18	19	30	302	12673
GO:0048446	petal morphogenesis	P	8.09E-16	3.88E-18	19	32	302	12671
GO:0060255	regulation of macromolecule metabolic process	P	1.47E-15	7.23E-18	101	1621	220	11082

GO:0048465	corolla development	P	1.81E-15	9.48E-18	20	41	301	12662
GO:0048441	petal development	P	1.81E-15	9.48E-18	20	41	301	12662
GO:0006261	DNA-dependent DNA replication	P	1.81E-15	9.59E-18	34	189	287	12514
GO:0007010	cytoskeleton organization	P	1.83E-15	9.95E-18	49	428	272	12275
GO:0040029	regulation of gene expression, epigenetic	P	2.48E-15	1.38E-17	47	396	274	12307
GO:0048453	sepal formation	P	2.88E-15	1.68E-17	18	29	303	12674
GO:0048447	sepal morphogenesis	P	2.88E-15	1.68E-17	18	29	303	12674
GO:0000278	mitotic cell cycle	P	2.97E-15	1.77E-17	37	236	284	12467
GO:0006259	DNA metabolic process	P	3.74E-15	2.28E-17	62	695	259	12008
GO:0048464	flower calyx development	P	6.39E-15	4.06E-17	18	31	303	12672
GO:0048442	sepal development	P	6.39E-15	4.06E-17	18	31	303	12672
GO:0009908	flower development	P	7.22E-15	4.67E-17	59	643	262	12060
GO:0000280	nuclear division	P	8.05E-15	5.31E-17	24	81	297	12622
GO:0009886	post-embryonic morphogenesis	P	1.23E-14	8.31E-17	42	329	279	12374
GO:0006342	chromatin silencing	P	1.35E-14	9.28E-17	35	220	286	12483
GO:0045814	negative regulation of gene expression, epigenetic	P	1.93E-14	1.35E-16	35	223	286	12480
GO:0009887	organ morphogenesis	P	3.29E-14	2.34E-16	43	357	278	12346
GO:0010564	regulation of cell cycle process	P	6.70E-14	4.86E-16	26	113	295	12590
GO:2000026	regulation of multicellular organismal development	P	7.86E-14	5.80E-16	53	557	268	12146

GO:0051239	regulation of multicellular organismal process	P	8.87E-14	6.66E-16	53	559	268	12144
GO:0006468	protein phosphorylation	P	1.39E-13	1.07E-15	64	803	257	11900
GO:0051225	spindle assembly	P	2.39E-13	1.86E-15	16	27	305	12676
GO:0016458	gene silencing	P	2.87E-13	2.26E-15	41	347	280	12356
GO:0048449	floral organ formation	P	3.47E-13	2.78E-15	25	111	296	12592
GO:0048645	organ formation	P	5.73E-13	4.67E-15	26	126	295	12577
GO:0048646	anatomical structure formation involved in morphogenesis	P	7.43E-13	6.16E-15	32	210	289	12493
GO:0050789	regulation of biological process	P	1.65E-12	1.39E-14	154	3498	167	9205
GO:0048285	organelle fission	P	1.84E-12	1.57E-14	24	109	297	12594
GO:0006996	organelle organization	P	1.90E-12	1.65E-14	99	1770	222	10933
GO:0019222	regulation of metabolic process	P	1.95E-12	1.72E-14	106	1974	215	10729
GO:0051329	interphase of mitotic cell cycle	P	2.13E-12	1.93E-14	27	148	294	12555
GO:0051325	interphase	P	2.13E-12	1.93E-14	27	148	294	12555
GO:0000087	M phase of mitotic cell cycle	P	2.20E-12	2.02E-14	20	67	301	12636
GO:0070925	organelle assembly	P	2.20E-12	2.05E-14	16	33	305	12670
GO:0048563	post-embryonic organ morphogenesis	P	2.25E-12	2.16E-14	25	123	296	12580
GO:0048444	floral organ morphogenesis	P	2.25E-12	2.16E-14	25	123	296	12580
GO:0007051	spindle organization	P	3.04E-12	2.95E-14	16	34	305	12669
GO:0050793	regulation of developmental process	P	3.05E-12	3.00E-14	55	661	266	12042
GO:0080090	regulation of primary metabolic process	P	5.16E-12	5.20E-14	93	1633	228	11070

GO:0031048	chromatin silencing by small RNA	P	5.16E-12	5.21E-14	23	104	298	12599
GO:0006346	methylation-dependent chromatin silencing	P	6.08E-12	6.22E-14	23	105	298	12598
GO:0048731	system development	P	1.36E-11	1.41E-13	85	1439	236	11264
GO:0043414	macromolecule methylation	P	1.52E-11	1.60E-13	45	477	276	12226
GO:0048513	organ development	P	2.34E-11	2.48E-13	84	1427	237	11276
GO:0050794	regulation of cellular process	P	2.56E-11	2.75E-13	135	2976	186	9727
GO:0016043	cellular component organization	P	3.30E-11	3.59E-13	123	2597	198	10106
GO:0065007	biological regulation	P	3.33E-11	3.67E-13	159	3807	162	8896
GO:0071842	cellular component organization at cellular level	P	3.37E-11	3.75E-13	104	2012	217	10691
GO:0019219	regulation of nucleobase-containing compound metabolic process	P	3.62E-11	4.08E-13	79	1305	242	11398
GO:0031323	regulation of cellular metabolic process	P	6.09E-11	6.94E-13	93	1709	228	10994
GO:0003777	microtubule motor activity	F	6.70E-11	7.78E-13	15	36	306	12667
GO:0048856	anatomical structure development	P	6.70E-11	7.80E-13	115	2373	206	10330
GO:0051171	regulation of nitrogen compound metabolic process	P	7.65E-11	9.01E-13	79	1326	242	11377
GO:0010556	regulation of macromolecule biosynthetic process	P	1.12E-10	1.35E-12	77	1283	244	11420
GO:2000112	regulation of cellular macromolecule biosynthetic process	P	1.12E-10	1.35E-12	77	1283	244	11420
GO:0010558	negative regulation of macromolecule biosynthetic process	P	1.13E-10	1.39E-12	37	351	284	12352

GO:2000113	negative regulation of cellular macromolecule biosynthetic process	P	1.13E-10	1.39E-12	37	351	284	12352
GO:0045892	negative regulation of transcription, DNA-dependent	P	1.16E-10	1.46E-12	36	333	285	12370
GO:0051253	negative regulation of RNA metabolic process	P	1.16E-10	1.46E-12	36	333	285	12370
GO:0031327	negative regulation of cellular biosynthetic process	P	1.74E-10	2.21E-12	37	357	284	12346
GO:0009653	anatomical structure morphogenesis	P	1.80E-10	2.31E-12	74	1217	247	11486
GO:0009890	negative regulation of biosynthetic process	P	1.84E-10	2.39E-12	37	358	284	12345
GO:0045934	negative regulation of nucleobase-containing compound metabolic process	P	2.46E-10	3.21E-12	36	343	285	12360
GO:0051172	negative regulation of nitrogen compound metabolic process	P	2.63E-10	3.47E-12	36	344	285	12359
GO:0010629	negative regulation of gene expression	P	2.94E-10	3.92E-12	42	464	279	12239
GO:0010468	regulation of gene expression	P	3.13E-10	4.22E-12	80	1396	241	11307
GO:0010605	negative regulation of macromolecule metabolic process	P	4.24E-10	5.76E-12	44	512	277	12191
GO:0032502	developmental process	P	4.87E-10	6.68E-12	125	2774	196	9929
GO:0031047	gene silencing by RNA	P	5.30E-10	7.34E-12	32	280	289	12423
GO:0031326	regulation of cellular biosynthetic process	P	5.69E-10	7.96E-12	78	1359	243	11344
GO:0009889	regulation of biosynthetic process	P	6.95E-10	9.81E-12	78	1365	243	11338
GO:0031324	negative regulation of cellular metabolic process	P	9.83E-10	1.40E-11	38	402	283	12301
GO:0007275	multicellular organismal development	P	1.17E-09	1.68E-11	118	2584	203	10119

GO:0009892	negative regulation of metabolic process	P	1.30E-09	1.88E-11	44	532	277	12171
GO:0048437	floral organ development	P	1.41E-09	2.06E-11	36	368	285	12335
GO:0003774	motor activity	F	1.97E-09	2.91E-11	15	49	306	12654
GO:0071840	cellular component organization or biogenesis	P	2.22E-09	3.31E-11	127	2906	194	9797
GO:0016310	phosphorylation	P	3.45E-09	5.18E-11	69	1167	252	11536
GO:0048580	regulation of post-embryonic development	P	4.10E-09	6.21E-11	35	364	286	12339
GO:0048438	floral whorl development	P	9.12E-09	1.39E-10	31	298	290	12405
GO:0071841	cellular component organization or biogenesis at cellular level	P	1.71E-08	2.64E-10	112	2503	209	10200
GO:0048569	post-embryonic organ development	P	2.01E-08	3.12E-10	38	451	283	12252
GO:0032501	multicellular organismal process	P	2.54E-08	3.97E-10	118	2716	203	9987
GO:0005874	microtubule	C	4.69E-08	7.41E-10	16	76	305	12627
GO:0009909	regulation of flower development	P	6.13E-08	9.76E-10	29	286	292	12417
GO:0036211	protein modification process	P	6.67E-08	1.08E-09	93	1961	228	10742
GO:0006464	cellular protein modification process	P	6.67E-08	1.08E-09	93	1961	228	10742
GO:2000241	regulation of reproductive process	P	7.82E-08	1.28E-09	30	309	291	12394
GO:0032259	methylation	P	8.76E-08	1.44E-09	45	639	276	12064
GO:2001141	regulation of RNA biosynthetic process	P	2.17E-07	3.63E-09	65	1182	256	11521
GO:0006355	regulation of transcription, DNA-dependent	P	2.17E-07	3.63E-09	65	1182	256	11521
GO:0009791	post-embryonic development	P	2.69E-07	4.53E-09	79	1592	242	11111
GO:0051252	regulation of RNA metabolic process	P	2.76E-07	4.68E-09	65	1190	256	11513

GO:0048608	reproductive structure development	P	8.64E-07	1.48E-08	67	1284	254	11419
GO:0048638	regulation of developmental growth	P	8.90E-07	1.53E-08	23	210	298	12493
GO:0048523	negative regulation of cellular process	P	8.94E-07	1.56E-08	42	620	279	12083
GO:0007067	mitosis	P	8.94E-07	1.56E-08	13	58	308	12645
GO:0015630	microtubule cytoskeleton	C	9.76E-07	1.72E-08	18	127	303	12576
GO:0090304	nucleic acid metabolic process	P	1.01E-06	1.80E-08	110	2627	211	10076
GO:0043412	macromolecule modification	P	1.16E-06	2.07E-08	99	2274	222	10429
GO:0010374	stomatal complex development	P	1.47E-06	2.64E-08	18	131	303	12572
GO:0007018	microtubule-based movement	P	1.62E-06	2.93E-08	12	50	309	12653
GO:0032774	RNA biosynthetic process	P	4.07E-06	7.42E-08	67	1341	254	11362
GO:0048519	negative regulation of biological process	P	4.86E-06	8.94E-08	52	924	269	11779
GO:0040008	regulation of growth	P	5.52E-06	1.02E-07	23	235	298	12468
GO:0010075	regulation of meristem growth	P	6.32E-06	1.18E-07	18	146	303	12557
GO:0006351	transcription, DNA-dependent	P	7.22E-06	1.36E-07	66	1334	255	11369
GO:0006796	phosphate-containing compound metabolic process	P	9.79E-06	1.86E-07	72	1525	249	11178
GO:0035266	meristem growth	P	9.79E-06	1.86E-07	18	151	303	12552
GO:0006793	phosphorus metabolic process	P	1.02E-05	1.95E-07	72	1527	249	11176
GO:0044430	cytoskeletal part	C	1.26E-05	2.43E-07	18	154	303	12549
GO:0022414	reproductive process	P	1.28E-05	2.50E-07	78	1721	243	10982
GO:0000003	reproduction	P	1.46E-05	2.86E-07	78	1727	243	10976

GO:0048509	regulation of meristem development	P	1.50E-05	2.94E-07	19	174	302	12529
GO:0001708	cell fate specification	P	3.96E-05	7.84E-07	8	24	313	12679
GO:0003006	developmental process involved in reproduction	P	4.89E-05	9.76E-07	69	1503	252	11200
GO:0005856	cytoskeleton	C	7.21E-05	1.45E-06	18	176	303	12527
GO:0042023	DNA endoreduplication	P	9.92E-05	2.00E-06	12	78	309	12625
GO:0000079	regulation of cyclin-dependent protein kinase activity	P	1.48E-04	3.01E-06	7	20	314	12683
GO:0007169	transmembrane receptor protein tyrosine kinase signaling pathway	P	2.42E-04	4.98E-06	12	86	309	12617
GO:0007167	enzyme linked receptor protein signaling pathway	P	2.42E-04	4.98E-06	12	86	309	12617
GO:0045165	cell fate commitment	P	2.83E-04	5.87E-06	8	33	313	12670
GO:0009855	determination of bilateral symmetry	P	3.15E-04	6.61E-06	13	105	308	12598
GO:0009799	specification of symmetry	P	3.15E-04	6.61E-06	13	105	308	12598
GO:0006928	cellular component movement	P	3.24E-04	6.85E-06	12	89	309	12614
GO:0010073	meristem maintenance	P	3.87E-04	8.21E-06	20	243	301	12460
GO:0010103	stomatal complex morphogenesis	P	4.09E-04	8.73E-06	13	108	308	12595
GO:0009524	phragmoplast	C	0.001085499	2.33E-05	8	41	313	12662
GO:0006139	nucleobase-containing compound metabolic process	P	0.001232245	2.66E-05	115	3222	206	9481
GO:0044427	chromosomal part	C	0.0014233	3.10E-05	12	105	309	12598
GO:0009888	tissue development	P	0.001627873	3.56E-05	40	791	281	11912

GO:0009059	macromolecule biosynthetic process	P	0.001786996	3.93E-05	99	2687	222	10016
GO:0010143	cutin biosynthetic process	P	0.003054561	6.76E-05	4	6	317	12697
GO:0034645	cellular macromolecule biosynthetic process	P	0.003419438	7.62E-05	96	2628	225	10075
GO:0004674	protein serine/threonine kinase activity	F	0.003839808	8.60E-05	34	654	287	12049
GO:0071900	regulation of protein serine/threonine kinase activity	P	0.004004757	9.05E-05	7	37	314	12666
GO:0010016	shoot morphogenesis	P	0.004004757	9.07E-05	21	315	300	12388
GO:0005694	chromosome	C	0.004542393	1.04E-04	14	160	307	12543
GO:0009957	epidermal cell fate specification	P	0.004547407	1.04E-04	4	7	317	12696
GO:0007166	cell surface receptor signaling pathway	P	0.004732981	1.09E-04	12	121	309	12582
GO:0007000	nucleolus organization	P	0.005561425	1.30E-04	5	16	316	12687
GO:0000793	condensed chromosome	C	0.005561425	1.30E-04	5	16	316	12687
GO:0048589	developmental growth	P	0.005778392	1.35E-04	28	505	293	12198
GO:0034641	cellular nitrogen compound metabolic process	P	0.006271806	1.48E-04	119	3495	202	9208
GO:0043170	macromolecule metabolic process	P	0.007124757	1.69E-04	166	5271	155	7432
GO:0044260	cellular macromolecule metabolic process	P	0.007207077	1.72E-04	157	4926	164	7777
GO:0048507	meristem development	P	0.007236161	1.73E-04	24	407	297	12296
GO:0042325	regulation of phosphorylation	P	0.009310437	2.24E-04	13	152	308	12551
GO:0007389	pattern specification process	P	0.009548186	2.31E-04	19	289	302	12414
GO:0008356	asymmetric cell division	P	0.010466792	2.55E-04	5	19	316	12684
GO:0005576	extracellular region	C	0.011130584	2.72E-04	43	964	278	11739

GO:0000914	phragmoplast assembly	P	0.011410855	2.81E-04	3	3	318	12700
GO:0071844	cellular component assembly at cellular level	P	0.013040885	3.23E-04	33	675	288	12028
GO:0006807	nitrogen compound metabolic process	P	0.015655496	3.89E-04	119	3576	202	9127
GO:0032993	protein-DNA complex	C	0.018592286	4.65E-04	6	35	315	12668
GO:0000912	assembly of actomyosin apparatus involved in cell cycle cytokinesis	P	0.01920032	4.82E-04	3	4	318	12699
GO:0031225	anchored to membrane	C	0.020061757	5.08E-04	9	85	312	12618
GO:0045859	regulation of protein kinase activity	P	0.020061757	5.12E-04	12	145	309	12558
GO:0043549	regulation of kinase activity	P	0.020061757	5.12E-04	12	145	309	12558
GO:0022607	cellular component assembly	P	0.020594265	5.28E-04	33	695	288	12008
GO:2000123	positive regulation of stomatal complex development	P	0.023420031	6.06E-04	2	0	319	12703
GO:0001932	regulation of protein phosphorylation	P	0.023420031	6.07E-04	12	148	309	12555
GO:0006633	fatty acid biosynthetic process	P	0.023590555	6.14E-04	15	216	306	12487
GO:0004672	protein kinase activity	F	0.023721585	6.20E-04	38	851	283	11852
GO:0051338	regulation of transferase activity	P	0.024393669	6.41E-04	12	149	309	12554
GO:0048466	androecium development	P	0.025120045	6.67E-04	10	108	311	12595
GO:0048443	stamen development	P	0.025120045	6.67E-04	10	108	311	12595
GO:0006323	DNA packaging	P	0.027027594	7.21E-04	7	54	314	12649
GO:0040007	growth	P	0.027838448	7.46E-04	31	651	290	12052
GO:0031032	actomyosin structure organization	P	0.028130464	7.58E-04	3	5	318	12698

GO:0016070	RNA metabolic process	P	0.028260351	7.65E-04	79	2206	242	10497
GO:0042127	regulation of cell proliferation	P	0.033869426	9.21E-04	9	93	312	12610
GO:0010467	gene expression	P	0.037562356	0.001026241	89	2579	232	10124
GO:0009955	adaxial/abaxial pattern specification	P	0.038539463	0.001057926	7	58	314	12645
GO:0071103	DNA conformation change	P	0.038794827	0.00106996	8	76	313	12627
GO:0000786	nucleosome	C	0.042820307	0.001186527	5	28	316	12675
GO:0044267	cellular protein metabolic process	P	0.04302975	0.001197902	98	2913	223	9790
GO:0003002	regionalization	P	0.049487829	0.001384096	13	187	308	12516

GO-ID	Term	Category	FDR	P-Value	#Test	#Ref	#not in Test	#not in Ref
GO:0009834	secondary cell wall biogenesis	P	8.65E-13	1.05E-16	16	39	310	20988
GO:0042546	cell wall biogenesis	P	6.75E-12	1.64E-15	33	357	293	20670
GO:0009832	plant-type cell wall biogenesis	P	3.41E-10	1.24E-13	23	184	303	20843
GO:0010383	cell wall polysaccharide metabolic process	P	8.84E-08	4.29E-11	23	251	303	20776
GO:0070592	cell wall polysaccharide biosynthetic process	P	1.01E-07	6.11E-11	20	186	306	20841
GO:0070589	cellular component macromolecule biosynthetic process	P	1.02E-07	8.65E-11	20	190	306	20837
GO:0044038	cell wall macromolecule biosynthetic process	P	1.02E-07	8.65E-11	20	190	306	20837
GO:0044036	cell wall macromolecule metabolic process	P	1.02E-07	9.85E-11	26	341	300	20686
GO:0070882	cellular cell wall organization or biogenesis	P	1.11E-07	1.21E-10	39	752	287	20275
GO:0010382	cellular cell wall macromolecule metabolic process	P	1.68E-07	2.04E-10	22	248	304	20779
GO:0045491	xylan metabolic process	P	3.95E-07	5.27E-10	19	189	307	20838
GO:0010413	glucuronoxylan metabolic process	P	5.30E-07	7.71E-10	18	171	308	20856
GO:0045492	xylan biosynthetic process	P	5.33E-07	8.40E-10	18	172	308	20855
GO:0010410	hemicellulose metabolic process	P	7.43E-07	1.26E-09	19	200	307	20827
GO:0009698	phenylpropanoid metabolic process	P	2.84E-06	5.17E-09	26	416	300	20611
GO:0071669	plant-type cell wall organization or biogenesis	P	5.59E-06	1.08E-08	27	464	299	20563
GO:0009808	lignin metabolic process	P	7.34E-06	1.51E-08	13	98	313	20929
GO:0009699	phenylpropanoid biosynthetic process	P	3.94E-05	8.61E-08	21	323	305	20704
GO:2000652	regulation of secondary cell wall biogenesis	P	4.19E-05	9.64E-08	5	4	321	21023
GO:0071554	cell wall organization or biogenesis	P	4.50E-05	1.09E-07	40	1013	286	20014
GO:0071843	cellular component biogenesis at cellular level	P	1.12E-04	2.84E-07	33	772	293	20255
GO:0010089	xylem development	P	1.13E-04	3.01E-07	10	68	316	20959
GO:0019748	secondary metabolic process	P	3.43E-04	9.55E-07	34	857	292	20170

GO:0016491	oxidoreductase activity	F	3.59E-04	1.05E-06	73	2663	253	18364
GO:0051213	dioxygenase activity	F	6.78E-04	2.05E-06	14	182	312	20845
GO:0006725	cellular aromatic compound metabolic process	P	0.001237582	3.90E-06	37	1043	289	19984
GO:0044550	secondary metabolite biosynthetic process	P	0.001694758	5.55E-06	23	496	303	20531
GO:0015698	inorganic anion transport	P	0.002244655	7.62E-06	15	235	311	20792
GO:0010087	phloem or xylem histogenesis	P	0.002437041	8.57E-06	10	102	316	20925
GO:0009809	lignin biosynthetic process	P	0.003594891	1.39E-05	8	64	318	20963
GO:0034637	cellular carbohydrate biosynthetic process	P	0.003594891	1.39E-05	28	723	298	20304
GO:0006624	vacuolar protein processing	P	0.003594891	1.39E-05	3	1	323	21026
GO:0016701	oxidoreductase activity, acting on single donors with incorporation of molecular oxygen	F	0.005636982	2.26E-05	10	115	316	20912
GO:0033692	cellular polysaccharide biosynthetic process	P	0.007898068	3.26E-05	26	677	300	20350
GO:0016706	oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen, 2-oxoglutarate as one donor, and incorporation of one atom each of oxygen into both donors	F	0.007898068	3.38E-05	10	121	316	20906
GO:0050734	hydroxycinnamoyltransferase activity	F	0.007898068	3.45E-05	3	2	323	21025
GO:0015103	inorganic anion transmembrane transporter activity	F	0.015326009	6.87E-05	9	106	317	20921
GO:0044264	cellular polysaccharide metabolic process	P	0.019856149	9.15E-05	27	765	299	20262
GO:0005506	iron ion binding	F	0.023536467	1.13E-04	21	526	305	20501
GO:0045488	pectin metabolic process	P	0.023536467	1.14E-04	7	65	319	20962
GO:0016760	cellulose synthase (UDP-forming) activity	F	0.024574769	1.22E-04	6	45	320	20982
GO:0071702	organic substance transport	P	0.027901707	1.49E-04	23	619	303	20408
GO:0016759	cellulose synthase activity	F	0.027901707	1.52E-04	6	47	320	20980
GO:0055114	oxidation-reduction process	P	0.027901707	1.52E-04	64	2606	262	18421
GO:0046274	lignin catabolic process	P	0.027901707	1.56E-04	5	29	321	20998
GO:0046271	phenylpropanoid catabolic process	P	0.027901707	1.56E-04	5	29	321	20998

GO:0005976	polysaccharide metabolic process	P	0.028377924	1.62E-04	31	971	295	20056
GO:0016705	oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen	F	0.032494219	1.89E-04	19	467	307	20560
GO:0016051	carbohydrate biosynthetic process	P	0.033619226	2.00E-04	33	1075	293	19952
GO:0005576	extracellular region	C	0.035752614	2.21E-04	44	1610	282	19417
GO:0042538	hyperosmotic salinity response	P	0.035752614	2.22E-04	9	125	317	20902
GO:0044262	cellular carbohydrate metabolic process	P	0.035752614	2.25E-04	35	1176	291	19851
GO:0003834	beta-carotene 15,15'-monooxygenase activity	F	0.03616837	2.32E-04	2	0	324	21027
GO:0019438	aromatic compound biosynthetic process	P	0.039346117	2.61E-04	26	774	300	20253
GO:0000271	polysaccharide biosynthetic process	P	0.039346117	2.76E-04	26	777	300	20250
GO:0046524	sucrose-phosphate synthase activity	F	0.039346117	2.77E-04	3	6	323	21021
GO:0009815	1-aminocyclopropane-1-carboxylate oxidase activity	F	0.039346117	2.77E-04	3	6	323	21021
GO:0015020	glucuronosyltransferase activity	F	0.039346117	2.77E-04	3	6	323	21021

Chapter 4: Cynipid gall wasp genome sequencing reveals potential horizontal gene transfer events

4.1 Introduction

Horizontal gene transfer (HGT) is the non-sexual exchange of genetic material between two organisms (Keeling et al., 2009). Widespread horizontal gene transfer across great phylogenetic distances among eukaryotes is an unexpected discovery of the recent explosion in genome sequencing (Keeling, 2009). These genes potentially have great adaptive importance to the species receiving them (Keeling, 2009). Several examples of horizontal gene transfer have been identified in the genomes of eukaryotic plant parasites (Dunning Hotopp, 2007; Mitreva et al., 2009; Sommer and Streit, 2011; Acuña et al., 2012; Kirsch et al., 2012; Pauchet and Heckel, 2013); suggesting similar processes could be important in cynipid galling.

4.1.1 Plant pathogen derived plant cell wall degrading enzymes

Many horizontal gene transfers have been identified in the plant-parasitic nematodes of the *Meloidogyne*, *Heterodera*, *Globodera*, and *Pratylenchus* genera (Dunning Hotopp, 2007; Mitreva et al., 2009; Sommer and Streit, 2011). Commonly, these genes are potential plant cell wall degrading enzymes (PCWDEs) of prokaryotic or fungal origin. (Sommer and Streit, 2011). Such genes could be crucial to successful parasitism by metabolising the components of the plant cell wall. Horizontal gene transfers into insect genomes have also been hypothesised (Keeling, 2009). Several species of plant-parasitic beetle genomes have been found to encode key enzymes of prokaryotic origin, including cellulase (Pauchet et al., 2010; Acuña et al., 2012; Pauchet and Heckel, 2013). The transfer of a prokaryotic mannanase gene into the coffee berry borer beetle, *Hypothenemus hampei*, a worldwide pest of coffee may have facilitated its adaptation to feeding on coffee beans (Acuña et al., 2012). The polysaccharide mannan occurs in high proportions in coffee beans and HGT to the beetle of mannanase allows it to metabolise this substrate (Acuña et al., 2012). The genome of the mountain pine beetle, *Dendroctonus ponderosae*, encodes the most extensive number of plant cell wall degrading enzymes known in an insect at 52 (Keeling et al., 2013). In total, Keeling et al. (2013) found six glycoside hydrolase family 48 proteins, seven polysaccharide lyase family 4 proteins, eight endo-b-1,4-glucanases, nine pectin methylesterases, and twenty-two endopolygalacturonases (cellulases).

Other candidate horizontally transferred plant cell wall degrading enzymes in animals are cellulases, hemicellulases including xylanases, pectinases, and ligninases (Calderón-Cortés et al., 2012). As more genomes are sequenced the importance of horizontal genetic transfer to eukaryotic macroevolution, particularly from prokaryotes to invertebrates, will become clearer (Dunning Hotopp, 2011). The mechanism by which these prokaryotic genes are transferred into plant parasites is unknown, but must occur into the recipient species germline (Keeling, 2009). Symbioses involving gut fungi or bacteria in their guts are common in insects, and are well understood in termites (Calderón-Cortés et al., 2012). Indeed, for some time this was the only known insect mechanism of degrading plant cell walls (Martin, 1991). The symbionts provide the necessary PCWDEs for host digestion of plant cell walls. Horizontally transferred PCWDEs of prokaryotic origin in insect genomes could be the relics of ancient symbioses. Enzymes with potential plant cell wall degrading activities were also present in the last common ancestor of bilaterian animals (Calderón-Cortés et al., 2010; Davison & Blaxter 2005; Lo et al., 2003; reviewed in Calderón-Cortés et al., 2012). Evidence for such genes in the Arthropoda was previously overlooked, as they are absent in insect model organisms such as *Drosophila melanogaster* and *Tribolium castaneum* (Pauchet et al., 2010). Thus there are three mechanisms by which insects can degrade plant cell walls: 1) ancient endogenous genes, 2) symbioses with bacteria, fungi, protists or archaea and 3) horizontal genetic transfer of PCWDEs. Discriminating between the three possible origins of PCWDEs is essential to correctly identify horizontal genetic transfers.

Eggs of the oak gall wasp *B. pallida* are known to have cellulase and pectinase activity (Bronner and Plantefol, 1973). Enzymes coating the egg lyse plant cells immediately beneath it at oviposition, creating a cavity into which the larva emerges on hatching. Other potential roles of these enzymes are in digestion of nutritive cell walls by feeding larvae, or for general re-modelling of plant tissues at the start of gall induction. By identifying genes encoding PCWDEs in gall wasp genomes and transcriptomes their origin can be ascertained. PCWDE genes in gall wasps that are most similar to beetle genes encoding endogenous PCWDEs will also reflect potential horizontal gene transfers (Pauchet et al., 2010), as beetles have acquired these genes by HGT from prokaryotes. Differentiating between symbiotic and horizontally transferred PCWDE is difficult, because the draft genome assemblies used here (table 4.2) are not contiguous enough for multiple exons per contig. Furthermore, if the PCWDE has a symbiont origin then identifiable conserved genes of the symbiont genome, like 16s rRNA for prokaryotes, are expected in the reference genome assemblies. If only the PCWDE genes are detected horizontal transfer is indicated, but not confirmed.

4.1.2 Plant genes in gall wasp genomes

The physically intimate relationship between host and galler could provide an environment in which plant genes are passed to the galler. Unlike horizontal gene transfer from prokaryotes to insects there is no evidence for plant to insect transfer as yet. It may be a more rare occurrence than gene transfer from prokaryotes to insects, with examples emerging over time. Plants do exchange genes, from host to parasitic plants for example (Mower et al., 2004; Richardson and Palmer, 2007). In the obligate parasitic plant genus *Rafflesia* 24-41% of the mitochondrial genomes are derived from horizontal genetic transfer depending on the species (Xi et al., 2013), as is 2.1% of the nuclear genome (Xi et al., 2012).

Presence of genes of plant origins in gall wasp genomes potentially gives great insights into galling. Any horizontally transferred genes could reflect the plant host of an ancestral gall wasp rather than the current host. For example, an oak gall wasp genome containing genes orthologous to poppy family genes, as a species of poppy is the proposed ancestral host of the *Cynipidae* (Ronquist and Liljeblad, 2001).

4.1.3 Virus-like-particles

Known examples of horizontal genetic transfer in the order Hymenoptera are genes encoding proteins that produce the virus-like-particles (VLPs) of braconid parasitoid wasps (Espagne et al., 2004; Bezier et al., 2009). These wasps oviposit VLPs (also known as polydnaviruses) into the host along with parasitoid eggs. The genome within the VLP encodes for immune-suppressive genes needed for successful parasitism (Espagne et al., 2004). VLP-carrying wasps have incorporated the viral coat as a means of delivering their own genes, which in turn, enhance the wasp's parasitic abilities (Bezier et al., 2009). However, genes for packaging, assembling and enveloping VLPs in the wasps *Chelonus inanitus* and *Cotesia congregata* (family: Braconidae) are derived from an ancestral virus of the Nudivirus family (Bezier et al., 2009). The original virus was integrated into an ancestral braconid genome and the VLP genomes have diversified but the nudiviral structural genes are conserved (Bezier et al., 2009). Cornell (1983) first suggested a role for VLPs in gall wasps as a mechanism for transferal of the key gall-inducing substances. However, no evidence was found for VLP related expression in larval transcriptomes (see Chapter 3). But cynipid VLPs could act as a maternal affect, analogous to braconid wasps. In this case VLP producing genes would be expressed in adult females and not the larvae and can be identified in genome assemblies. Genes with similarities to viral production and capsid genes found in braconid wasps (Bezier et al.,

2009) is evidence for cynipid VLPs.

4.2 Testing for horizontal genome transfer events in cynipid genomes

4.2.1 Gall wasp resources

Three cynipid genome assemblies were queried for potential horizontal transfer events. Two are closely related oak gall wasps (tribe: Cynipini), *Biorhiza pallida* and *Belizinella gibbera*, and the other a rose gall wasp *Diplolepis spinosa* (section 4.2.2) (tribe*:* Diplolepidini). Transcriptomes from two gall wasp species, one on oak, *Andricus quercuscalicis* and the other on *Acer, Pediaspis aceris* (tribe: Pediaspini) were provided by the 1KITE project (1K Transcriptome Evolution: www.**1kite**.org/*)*. The Pediaspini are a sister tribe to the Cynipini, and the Diplolepidini are the next most closely related tribe forming a monophyletic clade within the *Cynipidae* (figure 4.1) (Ronquist and Liljeblad, 2001). A third transcriptome of a figitid parasitoid, *Leptopilina clavipes,* also from the 1KITE project, serves as an outgroup for analyses of the cynipid datasets, albeit an imperfect one. This is because the transcriptomes, created from adult wasps will not be expressing their full gene sets. Chapter 3 demonstrated that gall wasps have very different expression profiles across larval stages. Adult gall wasps probably have just as distinct expression from larval stages. The genomes of the transcriptome-sequenced species may therefore contain horizontally transferred genes that are not detected because they are not expressed. Thus it is not possible to say with certainty if candidate horizontally transferred genes are exclusive to the *Cynipidae*, as the outgroup figitid data are transcriptomic and not genomic. Statistics for all assemblies used including N50s and Core Eukaryotic Genes Mapping Approach (*CEGMA*) (Parra et al., 2007) completeness scores are shown in tables 4.2-3.

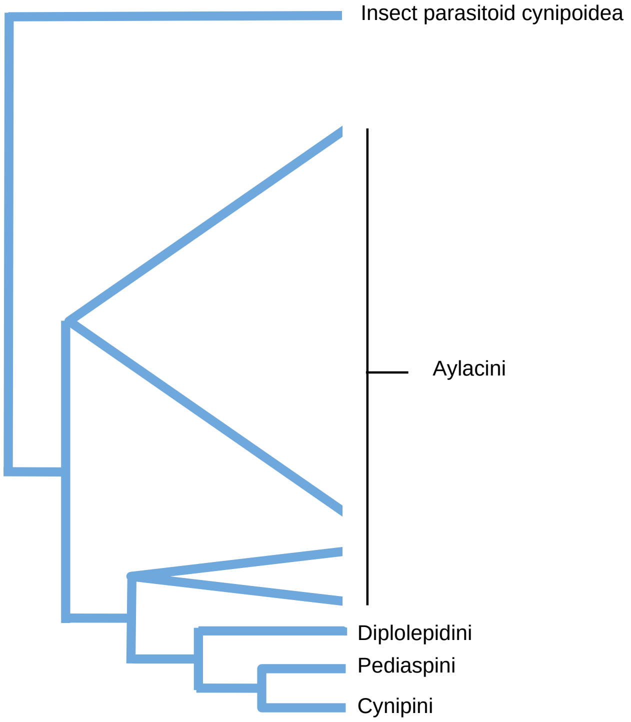

Figure 4.1. Phylogeny of gall wasp tribes demonstrating paraphyly of the Aylacini; Synergini inquilines not included, as they are unresolved. Adapted from Ronquist & Liljeblad (2001).

4.2.2 Sequencing and Assembling the *D. spinosa* genome and other genomic resources

D. spinosa is asexual so diploid females were sequenced. Four individuals were extracted and combined for 454 sequencing and three extracted for Illumina sequencing. The specimens were collected in Ontario, Canada by Dr Joe Shorthouse (Laurentian University, Ontario Canada). DNEasy (Qiagen) extractions were made of the samples and quality controlled by 260/280 ratio and DNA concentration. Three paired-end libraries were prepared for Illumina sequencing by the GenePool (University of Edinburgh) at 50, 75 and 100 base pairs (bp). The 50bp library was sequenced over two lanes of the Illumina GAIIx and the 75bp and 100bp sequenced on one lane each. After *Fastqc* inspection, only Q20 filtering using an in house perl script (courtesy S. Kumar) was run on the data; singles were not retained (table 4.1)

Lane	Read Length	Pairs (millions)	Bases (Gb)	Filtered Pairs (millions)	Filtered Bases (Gb)
1	50	14.9	1.49	12.1	1.16
2	50	18.8	1.88	18.0	1.82
3	75	23.3	3.50	22.3	3.28
4	100	38.1	7.63	37.2	7.27
Combined	mixed	95.1	14.5	89.6	13.53

Table 4.1. Illumina read statistics for each lane of *D. spinosa* genome sequencing. Combined data is in the bottom row.

Pre-existing *B. pallida* and *B. gibbera* assemblies (chapter 3) were used in this chapter (table 4.2). The *D. spinosa* genome was assembled using CLC bio (version 3.3.0, http://www.clcbio.com/products/clc-assembly-cell/) *de novo* de Bruijn graph based assembler with a paired-end insert size range of 0-400. Illumina reads were quality trimmed to Q20 (data used in assembly, table 4.1) and combined with 587 132 raw 454 reads to create a hybrid assembly.

Tables 4.2 and 4.3 show several metrics used to assess the assemblies. In addition to the N50 and other simple metrics, Core Eukaryotic Genes Mapping Approach (Parra et al., 2007) scores were also evaluated. Parra et al. (2007) identified a set of core eukaryotic genes (*CEGs*) present in all available eukaryote genomes, and the version used here (2.4) contains 248 of these *CEGs*. *CEGs* are supposed to represent single-copy nuclear genes. *CEGMA* combines *BLAST* (Altschul et al., 1990), *GeneWise* (Birney et al., 2004) and *geneid* (Parra et al., 2000) searches and *HMMER* (Finn et al., 2011) to identify orthologs of the *CEGMA* gene set in the tested dataset. Although, *CEGMA* is intended for genomes, under the assumption that core genes will be constitutively expressed it is applied here to transcriptomes. Tables 4.2-4.3 provide estimates of the percentage completeness for the *CEGs* (the percentage of complete *CEGs* in the dataset), the average copy number of orthologs per *CEG* and percentage of *CEGS* with more than one ortholog. The final two metrics indicate an excess of orthologs in the dataset. Haploid assembly of the data can explain this, when two copies of a gene assembled for a diploid genome instead of one due to sequence divergence. Alternatively, these metrics are explained by sequences of more than one species present in the dataset inflating the scores.

Genome assemblies	*B. pallida*	*B. gibbera*	*D. spinosa*
N50	1 075	643	1 729
Number of contigs	1 163 314	817 710	302 575
Assembly length	805 102 378	443 963 639	329 859 230
Average GC	32.9	35.84	32.8
Number of N's	4 203 182	2 525 790	2 286 759
CEGMA % completeness	37.9	25.0	79.8
Average copy number	1.19	1.23	1.15
% orthology	17.02	19.35	11.62

Transcriptome assemblies	*A. quercuscalicis*	*P. aceris*	*L. clavipes*
N50	2495	2115	1819
Number of contigs	22651	31282	21313
Assembly length	30 260 505	36 365 349	22 931 017
Average GC	39.6	38.6	36.8
Number of N's	3 607	3 718	1 281
CEGMA % completeness	95.56	97.18	96.77
Average copy number	1.94	1.85	1.81
% orthology	50.21	46.06	50.42

Table 4.2 & 4.3 Assembly metrics for cynipid genome assemblies and transcriptomes used in this chapter. Metrics generated using a perl script (courtesy S. Kumar) and CEGMA (version 2.4) (Parra et al., 2007).

The *D. spinosa* genome assembly is superior to the Cynipini (oak gall wasp) assemblies. *D. spinosa* has a higher N50 indicating greater contiguity. Far more *CEGs* are complete at 80% than *B. pallida* (38%) and *B. gibbera* (19%) and these *CEGs* have lower copy number and % orthology. This is despite the greater number of Illumina reads used for the cynipid genomes. The final *D. spinosa* assembly is also substantially smaller than the Cynipini assemblies, and its *CEGMA* scores better. A lower genome size in *D. spinosa* genome size compared to *B. gibbera* and *B. pallida* explains these observations. The genome size of *D. spinosa* is estimated at 0.63 gigabases (Lima, 2012) while the average oak gall wasp genome size is much larger at 1.71Gb (± 0.286, n = 4) (Lima, 2012). The discrepancy between the *D. spinosa* assembly length (number of bases in the assembly) and measure genome size may be explained by insufficient sequencing.

The transcriptome assemblies are all very close to *CEGMA* completeness as all are >95% complete. They have higher average *CEG* copy number and % orthology than the genome assemblies; this could result from isoform expression of the *CEGs*. If a *CEG* has more than one isoform expressed in the transcriptome this will inflate both copy number and percentage orthology.

4.2.3 *BLAST* searches for candidate horizontally transferred genes

The analysis is a simple presence or absence test for genes with a possible origin in a different Kingdom of life. Potential horizontally transferred genes were identified from *BLAST* (version 2.2.25) (Altschul et al., 1990) outputs against the NCBI non-redundant nucleotide (nt) and protein (nr) databases (databases downloaded January 4th 2012) using an e-value cut-off of 1e-5 and low sequence complexity filtering. All searches were performed using the Edinburgh Compute and Data Facility (ECDF), University of Edinburgh. Contigs or transcripts most similar to fungi, plants, bacterial or viral genomes were selected for further analysis. Bit scores were chosen because they can be compared across separate *BLAST* analyses, unlike e-values. *InterProScan* (Zdobnov and Apweiler, 2001) was used for further annotation of candidate genes. The insect symbiont *Wolbachia* (order: Rickettsiales) is present in all three species' genomes (table 4.4).

Equivalent Potential horizontal transfers identified in gall wasp resources were tested for in the *Nasonia vitripennis* genome as a non-galling control using the EvidentialGene (http://arthropods.eugenes.org/EvidentialGene/) annotator predictions (which incorporates BLAST predictions) (downloaded from http://arthropods.eugenes.org/EvidentialGene/nasonia/genes/nvit2_evigenes_pub11u.attr.simple.txt).

Species	Number of contigs	Combined length of contigs (megabases)	Most closely related *Wolbachia*	Genome size (Mb)	*Wolbachia* supergroup
B. pallida	561	1.53	*Drosophila melanogaster*	1.27[1]	A
B. gibbera	858	1.40	*D. melanogaster*	1.27[1]	A
D. spinosa	471	1.66	*Culex quinquefasciatus*	1.48[2]	B

Table 4.4. Wolbachia statistics for contigs in each genome assembly, including host of the most closely related *Wolbachia* genome, its genome size and supergroup. 1. Wu et al., 2004; 2. Klasson et al., 2008).

Wolbachia nuclear insertions into insect genomes are known to occur (Klasson et al., 2009). However, within the *Cynipidae, Wolbachia* is not present in all species (Stone, personal communication). Indeed, the supergroup of *Wolbachia* differs, for *D. spinosa* it is 'B' and for *B. pallida* and *B. gibbera* it is 'A'. Suggesting *Wolbachia* entered the *Cynipidae* more than once and possibly after evolution of the *Cynipidae*. Therefore *Wolbachia* were most probably not essential to the evolution of galling. For this reason, potential *Wolbachia* horizontal genetic transfers were not searched for in the available datasets.

4.2.4 Viral packaging proteins and horizontally transferred genes of plant origin

No viral packaging proteins or genes of plant origin were discovered that are shared across the genomes or transcriptomes. In *B. pallida* there are 27 contigs of putative *Quercus* and *Castanea* (the chestnuts) origin. These contigs are, on average of percentage *BLAST* identities normalized by alignment length, >97% identical to their most similar *Quercus* and *Castanea* sequences. They mainly encode for chloroplast and ribosomal associated genes. These contigs are not found in the *B. gibbera* or *D. spinosa* assemblies. They are probably remnants of *Q. robur* tissue on the body surface, or from gut of *B. pallida* individuals sequenced.

Neither plant horizontal transfer events nor a VLP-like system are detectable in any of the datasets tested.

4.2.5 Plant cell wall degrading enzymes of bacterial origin

BLAST searches revealed several genes encoding plant cell wall degrading enzymes (PCWDEs) of prokaryotic origin in all cynipid species tested. As expected the genomes have a greater number of unique matches than the transcriptomes. The outgroup figitid transcriptome had no corresponding expression. The *B. pallida* and *B. gibbera* genomes have the most potential PCWDEs, at 35 and 37 contigs respectively (*BLAST* best matches organized by PCWDE gene type, table 4.5); *D. spinosa* has less with 13 contigs (table 4.5).

The *BLAST* results also indicated many genes encoding polysaccharide lyase family 4 genes in several cynipid species. These genes are most similar to polysaccharide lyases found in the mountain pine beetle, *Dendroctonus ponderosae*. In *D. ponderosae,* based on phylogenetic analysis these polysaccharide lyases are hypothesized to result from horizontal gene transfer from plant pathogenic bacteria (Pauchet et al., 2010; Pauchet and Heckel, 2013). Therefore, these genes were included in table 4.5 and further analyses.

The *N. vitripennis* control genome only contains one potential PCWDE gene, an endoglucanase E-4-like, a glycosyl hydrolase 9 family cellulase (GH9). This gene only has protein BLAST homologs to other metazoans. Best matches are to the hymenoptera species *Bombus impatiens* (bit score 765) and *Apis mellifera* (bit score 764). Additionally, the three gall wasp genomes all contain contigs with endoglucanase E-4-like matching best to either of the bumblebees *Bombus terrestris* and *B. impatiens*. Davison & Blaxter (2005) provided phylogenetic evidence for an ancient eukaryotic origin for GH9 family cellulases a eukaryotic cellulase of ancient origin. Endoglucanase E-4-like cellulase was discounted from the candidate HGT set, and therefore no PCWDE genes of potential prokaryotic origin were found in the control *N. vitripennis* genome.

Table 4.5. *BLAST* results for each cynipid species sorted by PCWDE type. It includes Accession, 1) % identity, 2) alignment length to target, 3) mismatches, 4) gaps opened, Q. start = query start, Q. end = query end, R. start = reference start, R. end = reference end, e-value, bit score and reference (target) description. Contaminant sequences (section 5.2.6) have been removed from this list.

Contig	Accession	1	2	3	4	Q. start	Q. end	R. Start	R. End	E-value	Bit score	Target description
							Belizinella gibbera					
Cellulases												
contig_122111	gi|192360233|ref|YP_001983438.1|	65.02	303	99	5	567	1466	36	334	6.00E-107	393	endo-1,4-beta glucanase [Cellvibrio japonicus Ueda107]
contig_229402	gi|90022879|ref|YP_528706.1|	60.8	301	113	3	368	1264	37	334	1.00E-104	385	cellulase [Saccharophagus degradans 2-40]
contig_118239	gi|90022881|ref|YP_528708.1|	65	260	90	1	1	780	80	338	9.00E-101	372	cellulase [Saccharophagus degradans 2-40]
contig_63378	gi|269965254|dbj|BAI50016.1|	58.72	298	122	1	918	25	36	332	4.00E-100	370	endoglucanase [Saccharophagus sp. JAM-R001]
contig_133195	gi|90022879|ref|YP_528706.1|	67.07	249	82	0	1	747	84	332	1.00E-97	361	cellulase [Saccharophagus degradans 2-40]
contig_133196	gi|192360233|ref|YP_001983438.1|	66.54	257	85	1	83	853	40	295	1.00E-96	358	endo-1,4-beta glucanase [Cellvibrio japonicus Ueda107]
contig_131985	gi|269965254|dbj|BAI50016.1|	61.07	298	114	2	6	896	37	333	2.00E-95	353	endoglucanase [Saccharophagus sp. JAM-R001]
contig_302705	gi|90022879|ref|YP_528706.1|	62.14	206	78	0	137	754	49	254	2.00E-70	271	cellulase [Saccharophagus degradans 2-40]
contig_266172	gi|90022881|ref|YP_528708.1|	56.45	186	78	2	715	158	121	303	4.00E-50	203	cellulase [Saccharophagus degradans 2-40]
contig_270499	gi|192360233|ref|YP_001983438.1|	40.21	97	51	3	153	422	225	321	7.00E-09	64.3	endo-1,4-beta glucanase [Cellvibrio japonicus Ueda107]
Pectinases												
contig_372460	gi|308068153|ref|YP_003869758.1|	50	320	146	4	1073	147	29	345	2.00E-75	288	pectin lyase [Paenibacillus polymyxa E681]
contig_240234	gi|357201957|gb|AET59854.1|	46.46	226	106	4	303	944	29	251	7.00E-50	202	pectin lyase [Paenibacillus terrae HPL-003]
contig_326134	gi|307129345|ref|YP_003881361.1|	47.25	218	112	2	647	3	3	220	1.00E-49	202	pectate lyase [Dickeya dadantii 3937]
contig_120058	gi|357201957|gb|AET59854.1|	46.19	210	98	4	308	901	35	241	2.00E-44	184	pectin lyase [Paenibacillus terrae HPL-003]
contig_167362	gi|307129345|ref|YP_003881361.1|	49.21	189	93	2	560	3	3	191	1.00E-41	174	pectate lyase [Dickeya dadantii 3937]
contig_312349	gi|343096079|emb|CCC84288.1|	41.92	229	114	4	654	1	17	237	2.00E-38	164	pectate lyase [Paenibacillus polymyxa M1]
contig_330331	gi|310640947|ref|YP_003945705.1|	47.62	189	85	5	1	543	126	308	1.00E-34	152	pectate lyase, polysaccharide lyase family 1 [Paenibacillus polymyxa SC2]
contig_203617	gi|357201957|gb|AET59854.1|	45.33	150	71	2	4	426	26	173	1.00E-28	129	pectin lyase [Paenibacillus terrae HPL-003]
contig_27409	gi|16078925|ref|NP_389746.1|	48.03	127	65	1	10	390	219	344	4.00E-27	129	pectin lyase [Bacillus subtilis subsp. subtilis str. 168]
contig_745810	gi|350266199|ref|YP_004877506.1|	48.97	145	69	2	3	431	204	345	4.00E-27	124	pectin lyase [Bacillus subtilis subsp. spizizenii TU-B-10]

contig_744443	gi|357201957|gb|AET59854.1|	56	100	41	1	3	302	194	290	5.00E-26	121	pectin lyase [Paenibacillus terrae HPL-003]
contig_209160	gi|357201957|gb|AET59854.1|	41.77	158	66	3	92	493	10	165	6.00E-24	114	pectin lyase [Paenibacillus terrae HPL-003]
contig_644178	gi|1230540|gb|AAA92512.1|	44.53	128	71	0	37	420	3	130	1.00E-23	113	pectate lyase [Pseudomonas marginalis]
contig_301755	gi|229589782|ref|YP_002871901.1|	45.74	129	70	0	389	3	2	130	7.00E-23	110	pectin lyase [Pseudomonas fluorescens SBW25]
contig_430430	gi|310640947|ref|YP_003945705.1|	47.47	99	52	0	439	143	210	308	5.00E-20	101	pectate lyase, polysaccharide lyase family 1 [Paenibacillus polymyxa SC2]
contig_619773	gi|253687763|ref|YP_003016953.1|	43.52	108	57	3	344	33	70	177	5.00E-16	88.2	pectate lyase/Amb allergen [Pectobacterium carotovorum subsp. carotovorum PC1]
contig_105368	gi|52081406|ref|YP_080197.1|	58.57	70	29	0	215	6	347	416	6.00E-16	87.8	pectate lyase, polysaccharide lyase family 1 [Bacillus licheniformis ATCC 14580]
contig_415386	gi|349594864|gb|AEP91051.1|	51.39	72	35	0	391	176	99	170	4.00E-15	85.1	pectin lyase [Bacillus subtilis subsp. subtilis RO-NN-1]
contig_325752	gi|307129345|ref|YP_003881361.1|	49.45	91	43	1	265	2	70	160	2.00E-14	82.4	pectate lyase [Dickeya dadantii 3937]
contig_461437	gi|52081406|ref|YP_080197.1|	47.56	82	42	1	248	3	191	271	9.00E-14	80.5	pectate lyase, polysaccharide lyase family 1 [Bacillus licheniformis ATCC 14580]
contig_455965	gi|16078925|ref|NP_389746.1|	48	75	39	0	376	152	269	343	1.00E-12	76.6	pectin lyase [Bacillus subtilis subsp. subtilis str. 168]
contig_671218	gi|308068153|ref|YP_003869758.1|	48.1	79	32	2	76	285	37	115	2.00E-10	69.7	pectin lyase [Paenibacillus polymyxa E681]
contig_591792	gi|308068153|ref|YP_003869758.1|	45.68	81	35	1	216	1	25	105	5.00E-10	68.2	pectin lyase [Paenibacillus polymyxa E681]
contig_437115	gi|308068153|ref|YP_003869758.1|	39.05	105	54	2	293	6	7	110	8.00E-10	67.4	pectin lyase [Paenibacillus polymyxa E681]
Rhamnogalacturonate lyases												
contig_140448	gi|261820229|ref|YP_003258335.1|	35.05	194	118	2	730	170	42	234	2.00E-22	110	Rhamnogalacturonate lyase [Pectobacterium wasabiae WPP163]
contig_646631	gi|307129727|ref|YP_003881743.1|	42.86	63	36	0	11	199	94	156	6.00E-08	61.2	Rhamnogalacturonate lyase [Dickeya dadantii 3937]
contig_555256	gi|307129727|ref|YP_003881743.1|	39.68	63	38	0	11	199	94	156	8.00E-08	60.8	Rhamnogalacturonate lyase [Dickeya dadantii 3937]
Polysaccharide lyases												
contig_36809	gi|315570656|gb|ADU33332.1|	46.62	547	271	9	311	1924	30	564	2.00E-145	521	polysaccharide lyase family protein 4 [Dendroctonus ponderosae]
contig_56209	gi|315570656|gb|ADU33332.1|	45.08	539	268	7	1644	64	39	561	1.00E-131	475	polysaccharide lyase family protein 4 [Dendroctonus ponderosae]
contig_238739	gi|315570650|gb|ADU33329.1|	42.8	542	306	3	43	1659	9	549	2.00E-129	468	polysaccharide lyase family protein 4 [Dendroctonus ponderosae]
contig_272997	gi|315570656|gb|ADU33332.1|	48.42	349	173	3	1942	905	217	561	1.00E-95	356	polysaccharide lyase family protein 4 [Dendroctonus ponderosae]

contig_184148	gi|315570656|gb|ADU33332.1|	47.66	363	182	3	1227	154	101	460	2.00E-92	345	polysaccharide lyase family protein 4 [Dendroctonus ponderosae]
contig_166411	gi|315570650|gb|ADU33329.1|	43.22	354	181	4	1061	3	8	342	2.00E-76	291	polysaccharide lyase family protein 4 [Dendroctonus ponderosae]
contig_100272	gi|315570648|gb|ADU33328.1|	52.63	247	116	1	11	748	78	324	8.00E-71	271	polysaccharide lyase family protein 4 [Dendroctonus ponderosae]
contig_229203	gi|315570648|gb|ADU33328.1|	46.35	233	125	0	3	701	111	343	3.00E-53	213	polysaccharide lyase family protein 4 [Dendroctonus ponderosae]
contig_166410	gi|315570656|gb|ADU33332.1|	47.22	216	108	3	639	1	345	557	5.00E-51	205	polysaccharide lyase family protein 4 [Dendroctonus ponderosae]
contig_137836	gi|315570648|gb|ADU33328.1|	60	140	56	0	422	3	135	274	4.00E-45	184	polysaccharide lyase family protein 4 [Dendroctonus ponderosae]
contig_338719	gi|315570650|gb|ADU33329.1|	53.73	134	58	1	52	453	392	521	2.00E-35	152	polysaccharide lyase family protein 4 [Dendroctonus ponderosae]
contig_338718	gi|315570656|gb|ADU33332.1|	52.14	117	53	2	1	342	368	484	4.00E-33	125	polysaccharide lyase family protein 4 [Dendroctonus ponderosae]
contig_140446	gi|315570656|gb|ADU33332.1|	38.13	139	84	1	411	1	289	427	8.00E-21	103	polysaccharide lyase family protein 4 [Dendroctonus ponderosae]
contig_533748	gi|315570656|gb|ADU33332.1|	60.76	79	29	2	37	267	395	473	1.00E-19	99.8	polysaccharide lyase family protein 4 [Dendroctonus ponderosae]
contig_563084	gi|315570656|gb|ADU33332.1|	45.05	111	57	2	598	269	457	564	1.00E-17	94	polysaccharide lyase family protein 4 [Dendroctonus ponderosae]
contig_347641	gi|315570650|gb|ADU33329.1|	49.25	67	30	1	207	7	459	521	1.00E-11	73.2	polysaccharide lyase family protein 4 [Dendroctonus ponderosae]
contig_353796	gi|315570656|gb|ADU33332.1|	49.15	59	30	0	203	27	322	380	4.00E-11	71.6	polysaccharide lyase family protein 4 [Dendroctonus ponderosae]
contig_140447	gi|315570656|gb|ADU33332.1|	35.06	77	50	0	231	1	289	365	8.00E-08	60.8	polysaccharide lyase family protein 4 [Dendroctonus ponderosae]
						Biorhiza pallida						
Cellulases												
contig_167705	gi|90022881|ref|YP_528708.1|	64.55	299	105	1	1505	609	41	338	4.00E-115	421	cellulase [Saccharophagus degradans 2-40]
contig_249861	gi|90022881|ref|YP_528708.1|	60.2	304	118	2	1217	306	38	338	2.00E-105	388	cellulase [Saccharophagus degradans 2-40]
contig_69918	gi|192360233|ref|YP_001983438.1|	62.99	308	107	5	1257	2171	36	339	2.00E-104	385	endo-1,4-beta glucanase [Cellvibrio japonicus Ueda107]
contig_366356	gi|90022879|ref|YP_528706.1|	60	295	113	3	359	1237	34	325	4.00E-101	373	cellulase [Saccharophagus degradans 2-40]
contig_31255	gi|269965254|dbj|BAI50016.1|	54.68	331	145	2	1402	2382	3	332	5.00E-98	365	endoglucanase [Saccharophagus sp. JAM-R001]

contig_17336	gi|192360233|ref|YP_001983438.1|	66.27	255	85	1	1	765	85	338	8.00E-93	346	endo-1,4-beta glucanase [Cellvibrio japonicus Ueda107]
contig_17337	gi|90022879|ref|YP_528706.1|	66.67	195	65	0	3	587	37	231	1.00E-72	276	cellulase [Saccharophagus degradans 2-40]
contig_331473	gi|269965254|dbj|BAI50016.1|	60.44	182	71	1	547	2	41	221	5.00E-49	197	endoglucanase [Saccharophagus sp. JAM-R001]
contig_216706	gi|90022879|ref|YP_528706.1|	66.42	137	46	0	554	144	63	199	4.00E-57	192	cellulase [Saccharophagus degradans 2-40]
contig_173681	gi|192360233|ref|YP_001983438.1|	66.67	93	30	1	280	2	132	223	6.00E-29	130	endo-1,4-beta glucanase [Cellvibrio japonicus Ueda107]
contig_399218	gi|269965254|dbj|BAI50016.1|	65.52	87	30	0	350	610	37	123	7.00E-28	128	endoglucanase [Saccharophagus sp. JAM-R001]
contig_540965	gi|269965254|dbj|BAI50016.1|	62.5	88	33	0	379	642	37	124	2.00E-27	126	endoglucanase [Saccharophagus sp. JAM-R001]
contig_501105	gi|806574|emb|CAA60493.1|	32.47	194	99	4	1093	602	143	334	3.00E-16	91.7	endo-1,4-beta-glucanase [Cellvibrio japonicus]
contig_226353	gi|192360233|ref|YP_001983438.1|	68.66	67	20	1	241	44	218	284	2.00E-13	79.7	endo-1,4-beta glucanase [Cellvibrio japonicus Ueda107]
contig_402741	gi|90022881|ref|YP_528708.1|	53.23	62	29	0	1336	1151	276	337	6.00E-12	77.8	cellulase [Saccharophagus degradans 2-40]
Pectinases												
contig_441974	gi|350266199|ref|YP_004877506.1|	47.2	339	158	7	773	1753	15	344	3.00E-79	301	pectin lyase [Bacillus subtilis subsp. spizizenii TU-B-10]
contig_10790	gi|357201957|gb|AET59854.1|	46.53	346	162	7	2828	1833	8	344	4.00E-76	292	pectin lyase [Paenibacillus terrae HPL-003]
contig_165961	gi|308068153|ref|YP_003869758.1|	50.31	320	145	4	1671	745	29	345	1.00E-74	286	pectin lyase [Paenibacillus polymyxa E681]
contig_7692	gi|349594864|gb|AEP91051.1|	43.2	331	173	4	4080	3121	18	344	6.00E-73	281	pectin lyase [Bacillus subtilis subsp. subtilis RO-NN-1]
contig_115387	gi|307129345|ref|YP_003881361.1|	49.41	255	125	3	755	3	11	265	2.00E-62	245	pectate lyase [Dickeya dadantii 3937]
contig_11862	gi|242240781|ref|YP_002988962.1|	45.66	311	163	4	3270	4187	3	312	2.00E-60	241	Pectate lyase/Amb allergen [Dickeya dadantii Ech703]
contig_34168	gi|357201957|gb|AET59854.1|	44.56	285	152	3	916	71	63	344	3.00E-60	239	pectin lyase [Paenibacillus terrae HPL-003]
contig_481385	gi|357201957|gb|AET59854.1|	44.12	238	127	3	1026	322	110	344	1.00E-48	198	pectin lyase [Paenibacillus terrae HPL-003]
contig_668987	gi|310640947|ref|YP_003945705.1|	38.85	296	152	4	86	904	19	308	4.00E-46	191	pectate lyase, polysaccharide lyase family 1 [Paenibacillus polymyxa SC2]
contig_441114	gi|308068153|ref|YP_003869758.1|	45.3	234	105	7	845	1501	7	232	2.00E-41	175	pectin lyase [Paenibacillus polymyxa E681]
contig_611491	gi|357201957|gb|AET59854.1|	53.42	161	62	5	2	463	136	290	8.00E-39	163	pectin lyase [Paenibacillus terrae HPL-003]
contig_188178	gi|52081406|ref|YP_080197.1|	34.98	283	146	5	2485	1745	205	485	3.00E-37	162	pectate lyase, polysaccharide lyase family 1 [Bacillus licheniformis ATCC 14580]

contig_302522	gi|310640947|ref|YP_003945705.1|	47.17	159	81	1	620	144	153	308	2.00E-37	159	pectate lyase, polysaccharide lyase family 1 [Paenibacillus polymyxa SC2]
contig_763479	gi|310640947|ref|YP_003945705.1|	50.93	108	51	1	4	321	201	308	3.00E-25	118	pectate lyase, polysaccharide lyase family 1 [Paenibacillus polymyxa SC2]
contig_349598	gi|307129345|ref|YP_003881361.1|	41.43	70	37	1	2002	1793	151	216	6.00E-07	62.4	pectate lyase [Dickeya dadantii 3937]
Rhamnogalacturonate lyases												
contig_270176	gi|251790809|ref|YP_003005530.1|	41.92	396	222	2	1608	445	27	422	4.00E-95	354	Rhamnogalacturonate lyase [Dickeya zeae Ech1591]
contig_310874	gi|307129727|ref|YP_003881743.1|	45.71	326	167	4	955	2	28	351	8.00E-83	313	Rhamnogalacturonate lyase [Dickeya dadantii 3937]
contig_380131	gi|261820229|ref|YP_003258335.1|	57.89	76	32	0	228	1	243	318	1.00E-19	100	Rhamnogalacturonate lyase [Pectobacterium wasabiae WPP163]
contig_319025	gi|307129727|ref|YP_003881743.1|	31.86	113	69	2	879	1196	29	140	4.00E-11	74.3	Rhamnogalacturonate lyase [Dickeya dadantii 3937]
contig_632138	gi|251790809|ref|YP_003005530.1|	42.53	87	48	1	468	208	485	569	1.00E-09	67	Rhamnogalacturonate lyase [Dickeya zeae Ech1591]
Polysaccharide lyases												
contig_311700	gi|315570656|gb|ADU33332.1|	46.25	547	273	9	1725	112	30	564	2.00E-143	515	polysaccharide lyase family protein 4 [Dendroctonus ponderosae]
contig_211523	gi|315570656|gb|ADU33332.1|	44.42	547	276	8	1036	2655	39	564	1.00E-134	486	polysaccharide lyase family protein 4 [Dendroctonus ponderosae]
contig_319529	gi|315570650|gb|ADU33329.1|	42.41	547	309	4	27	1658	6	549	1.00E-128	466	polysaccharide lyase family protein 4 [Dendroctonus ponderosae]
contig_211524	gi|315570656|gb|ADU33332.1|	45.21	511	250	8	891	2399	39	527	2.00E-128	465	polysaccharide lyase family protein 4 [Dendroctonus ponderosae]
contig_24334	gi|315570656|gb|ADU33332.1|	49.05	367	180	3	1877	786	199	561	2.00E-100	371	polysaccharide lyase family protein 4 [Dendroctonus ponderosae]
contig_95640	gi|315570648|gb|ADU33328.1|	43.94	396	215	3	2710	1532	109	500	3.00E-93	349	polysaccharide lyase family protein 4 [Dendroctonus ponderosae]
contig_391986	gi|315570656|gb|ADU33332.1|	41.79	347	173	5	47	1078	241	561	3.00E-70	271	polysaccharide lyase family protein 4 [Dendroctonus ponderosae]
contig_40263	gi|315570656|gb|ADU33332.1|	46.98	215	111	2	3	638	276	490	3.00E-52	209	polysaccharide lyase family protein 4 [Dendroctonus ponderosae]
contig_261197	gi|315570656|gb|ADU33332.1|	42.11	247	130	3	1239	508	328	564	6.00E-47	193	polysaccharide lyase family protein 4 [Dendroctonus ponderosae]
contig_399225	gi|315570650|gb|ADU33329.1|	42.98	242	119	3	723	1	8	231	4.00E-46	191	polysaccharide lyase family protein 4 [Dendroctonus ponderosae]
contig_366060	gi|315570656|gb|ADU33332.1|	54.47	123	56	0	373	5	221	343	1.00E-37	161	polysaccharide lyase family protein 4 [Dendroctonus ponderosae]

contig_327626	gi|315570656|gb|ADU33332.1|	52	150	67	2	1242	796	419	564	7.00E-36	156	polysaccharide lyase family protein 4 [Dendroctonus ponderosae]
contig_321778	gi|315570648|gb|ADU33328.1|	53.54	127	54	2	2699	2328	179	303	3.00E-33	144	polysaccharide lyase family protein 4 [Dendroctonus ponderosae]
contig_741455	gi|315570650|gb|ADU33329.1|	48.97	145	70	1	812	378	410	550	2.00E-32	144	polysaccharide lyase family protein 4 [Dendroctonus ponderosae]
contig_270175	gi|315570656|gb|ADU33332.1|	42.98	114	63	1	338	3	308	421	6.00E-22	108	polysaccharide lyase family protein 4 [Dendroctonus ponderosae]
contig_273713	gi|315570656|gb|ADU33332.1|	64.94	77	26	1	2	229	415	491	5.00E-22	107	polysaccharide lyase family protein 4 [Dendroctonus ponderosae]
contig_273714	gi|315570656|gb|ADU33332.1|	64.38	73	25	1	3	218	419	491	5.00E-20	101	polysaccharide lyase family protein 4 [Dendroctonus ponderosae]
contig_572545	gi|315570656|gb|ADU33332.1|	63.01	73	26	1	3	218	419	491	4.00E-19	98.2	polysaccharide lyase family protein 4 [Dendroctonus ponderosae]
contig_327649	gi|315570656|gb|ADU33332.1|	49.43	87	42	1	259	5	338	424	1.00E-15	87	polysaccharide lyase family protein 4 [Dendroctonus ponderosae]
contig_331094	gi|315570650|gb|ADU33329.1|	38.33	120	69	3	491	138	15	131	8.00E-14	82.4	polysaccharide lyase family protein 4 [Dendroctonus ponderosae]
contig_379595	gi|315570648|gb|ADU33328.1|	65.38	52	18	0	453	608	175	226	1.00E-13	81.3	polysaccharide lyase family protein 4 [Dendroctonus ponderosae]
contig_759421	gi|315570656|gb|ADU33332.1|	45.12	82	42	1	1	246	483	561	9.00E-12	73.9	polysaccharide lyase family protein 4 [Dendroctonus ponderosae]
contig_438304	gi|315570656|gb|ADU33332.1|	49.23	65	33	0	208	14	497	561	1.00E-10	70.1	polysaccharide lyase family protein 4 [Dendroctonus ponderosae]
							Diplolepis spinosa					
Cellulases												
contig_59010	gi|90022881|ref|YP_528708.1|	61.33	300	115	1	2413	3312	39	337	2.00E-106	392	cellulase [Saccharophagus degradans 2-40]
contig_37918	gi|269965254|dbj|BAI50016.1|	54.43	327	146	2	1567	587	29	352	5.00E-100	369	endoglucanase [Saccharophagus sp. JAM-R001]
Pectinases												
contig_169528	gi|351470365|gb|EHA30503.1|	50	348	154	5	491	1483	9	353	3.00E-93	347	pectate lyase [Bacillus subtilis subsp. subtilis str. SC-8]
contig_101079	gi|242240781|ref|YP_002988962.1|	44.59	314	172	1	705	1640	1	314	1.00E-73	281	Pectate lyase/Amb allergen [Dickeya dadantii Ech703]
contig_128531	gi|242240781|ref|YP_002988962.1|	45.19	312	169	1	1530	601	3	314	4.00E-72	277	Pectate lyase/Amb allergen [Dickeya dadantii Ech703]
contig_58231	gi|343096079|emb|CCC84288.1|	44.51	319	163	3	1174	2097	28	343	8.00E-71	273	pectate lyase [Paenibacillus polymyxa M1]

contig_5773	gi|308068153|ref|YP_003869758.1|	51.25	160	76	2	1	480	188	345	2.00E-40	171	pectin lyase [Paenibacillus polymyxa E681]
contig_116836	gi|307129345|ref|YP_003881361.1|	45.22	157	84	1	468	4	3	159	2.00E-33	148	pectate lyase [Dickeya dadantii 3937]
contig_145587	gi|307129345|ref|YP_003881361.1|	43.95	157	86	1	127	591	3	159	1.00E-31	139	pectate lyase [Dickeya dadantii 3937]
contig_48663	gi|308068153|ref|YP_003869758.1|	48.84	129	64	1	279	659	88	216	2.00E-36	119	pectin lyase [Paenibacillus polymyxa E681]
contig_65337	gi|310640947|ref|YP_003945705.1|	43.14	102	56	1	2	307	210	309	1.00E-18	96.7	pectate lyase, polysaccharide lyase family 1 [Paenibacillus polymyxa SC2]
Rhamnogalacturonate lyases												
contig_167541	gi|261820229|ref|YP_003258335.1|	45.9	549	286	4	2724	1111	22	570	3.00E-141	507	Rhamnogalacturonate lyase [Pectobacterium wasabiae WPP163]
contig_58856	gi|227326316|ref|ZP_03830340.1|	43.32	554	304	4	1804	173	10	563	5.00E-131	472	rhamnogalacturonate lyase [Pectobacterium carotovorum subsp. carotovorum WPP14]
Polysaccharide lyases												
contig_84907	gi|315570650|gb|ADU33329.1|	44.55	541	290	5	263	1879	21	553	7.00E-137	493	polysaccharide lyase family protein 4 [Dendroctonus ponderosae]
contig_106711	gi|315570650|gb|ADU33329.1|	46.55	537	278	5	466	2073	22	550	8.00E-133	480	polysaccharide lyase family protein 4 [Dendroctonus ponderosae]
Andricus quercuscalicis												
Cellulases												
C77523	gi|269965254|dbj|BAI50016.1|	59.42	313	123	2	74	1012	24	332	2.00E-103	380	endoglucanase [Saccharophagus sp. JAM-R001]
C43076	gi|192360233|ref|YP_001983438.1|	69.62	79	23	1	239	3	158	235	2.00E-25	119	endo-1,4-beta glucanase [Cellvibrio japonicus Ueda107]
Pectinases												
scaffold5068	gi|357201957|gb|AET59854.1|	45.26	285	150	3	205	1050	63	344	1.00E-60	239	pectin lyase [Paenibacillus terrae HPL-003]
C39706	gi|52081406|ref|YP_080197.1|	47.22	72	38	0	2	217	190	261	2.00E-12	76.3	pectate lyase, polysaccharide lyase family 1 [Bacillus licheniformis ATCC 14580]
Pediaspis aceris												
Pectinases												
C96711	gi|308068153|ref|YP_003869758.1|	49.06	267	130	3	3	797	82	344	4.00E-64	249	pectin lyase [Paenibacillus polymyxa E681]
scaffold1651	gi|242240781|ref|YP_002988962.1|	49.61	256	127	1	94	855	3	258	6.00E-60	235	Pectate lyase/Amb allergen [Dickeya dadantii Ech703]
scaffold1653	gi|357201957|gb|AET59854.1|	50.65	231	109	2	163	849	63	290	6.00E-58	229	pectin lyase [Paenibacillus terrae HPL-003]

scaffold1652	gi|242240781|ref|YP_002988962.1|	50.34	149	72	1	13	453	113	261	4.00E-38	161	Pectate lyase/Amb allergen [Dickeya dadantii Ech703]
Rhamnogalacturonate lyases												
scaffold2513	gi|307129727|ref|YP_003881743.1|	43.37	475	261	1	32	1432	29	503	2.00E-119	434	Rhamnogalacturonate lyase [Dickeya dadantii 3937]
C48768	gi|271500632|ref|YP_003333657.1|	42.67	75	40	1	222	7	466	540	3.00E-10	68.9	Rhamnogalacturonate lyase [Dickeya dadantii Ech586]
Polysaccharide lyases												
C106083	gi|315570656|gb|ADU33332.1|	43.65	520	281	7	1547	9	34	548	6.00E-129	466	polysaccharide lyase family protein 4 [Dendroctonus ponderosae]
scaffold2514	gi|315570656|gb|ADU33332.1|	48.68	152	73	2	5	457	393	540	6.00E-37	157	polysaccharide lyase family protein 4 [Dendroctonus ponderosae]

4.2.6 Probable contaminant sequences

Some top hits to plant cell wall degrading enzymes are due to contamination of the sequenced individuals with other organisms. In total there are 1004 contigs in the *B. gibbera* assembly most similar to *Cordyceps* and the closely related *Beauvaria* (Rehner and Buckley, 2005). *Cordyceps* and *Beauvaria* species are entomopathogenic fungi. This includes the 28s and 18s rRNAs of *Beauvaria bassiana*. Furthermore, *Cordyceps* and *Beauvaria* sequences are not found in the other genomes or transcriptomes. The contigs indicate the presence of a complete, if partially sequenced, *Cordyceps* or *Beauvaria*-like fungal genome in the *B. gibbera* assembly and not a HGT event. Fungal spores carried by one of the *B. gibbera* individuals sequenced can explain the presence of this genome in the *B. gibbera* assembly.

A bacterial contaminant most closely related to, *Limnobacter species MED105* (http://www.ncbi.nlm.nih.gov/genome/13680?project_id=54689), was identified in the *B. pallida* assembly by its cellulase. In total 2896 *B. pallida* contigs had closest match to this *Limnobacter* species (Accession: NZ_ABCT00000000.1) including 23s and 16s rRNA genes. As was the case for *Cordyceps* in *B. gibbera*, *Limnobacter*-like sequences are not present in the other datasets. Again the genome of this species was probably sequenced alongside *B. pallida* in one or more of the individuals sequenced.

4.2.7 Shared top hits across cynipid resources

Four plant cell wall degrading functions were identified across the tested species, they were: 1) cellulase/endoglucanase, 2) rhamnogalacturonate lyase, 3) polysaccharide lyases and 4) pectin lyases. A lyase is an enzyme that breaks down a substrate through a mechanism other than hydrolysis or oxidation. These enzymes catalyse the break down of plant cell wall components cellulose, rhamnogalacturonans, and cell wall pectins. Polysaccharide lyases also break down rhamnogalacturonans, which are complex pectic polysaccharides of the cell wall.

The pectinases are the most common contigs occurring in all cynipid assemblies. They are most frequent in *B. gibbera* and *B. pallida* at 27 and 20 contigs respectively, with *D. spinosa* slightly less at 9 contigs. Rhamnogalacturonate lyases are found in all three genomes and another is expressed in *P. aceris*. The most similar species to these cynipid PCWDE genes are the plant pathogens *Dickeya dadantii* or *Pectobacterium wasabiae* of the family Enterobacteriaceae. *Dickeya* species are closely related to *Pectobacterium* species, and they were previously assigned to the genus *Pectobacterium* (Samson et al., 2005).

Cellulases are present in all gall wasp species except *P. aceris* (though might be present in the genome), most similar to either *Cellvibrio* or *Saccharophagus* species; they are particularly numerous in *B. pallida* and *B. gibbera*. In total, only seven genera of bacteria were responsible for best hits, five of which are class Gammaproteobacteria (*Cellvibrio, Dickeya, Pectobacterium, Pseudomonas*, and *Saccharophagus*), and two, most similar to the pectinases, are Bacilli (*Bacillus* and *Paenibacillus*).

4.2.8 Relationships among HGT candidates: phylogenies of PCWDE enzymes

A phylogenetic approach can tell us if cellulase genes from the same species are more similar to one another than to cellulases in another species or *vice-a-versa*. If such genes share orthologs within species then they have duplicated since that species lineage split from other species in the analysis. Alternatively, if a gene encoding a cellulase is more similar to a gene in another species than it is to cellulases in its own genome assembly, gene duplication more ancient than the split between those species is inferred. Whether this has happened in the gall wasp genome or an unidentified symbiont is not indicated by this analysis.

Coding sequences overlapping the *BLAST* hit to PCWDE genes were extracted from genomic and transcriptomic contigs for each species using *EMBOSS getorf* (Longden and Bleasby, 2000) and translated into amino acid sequence. The resulting proteins were kept for further analysis if they were at least 100 amino acids long. Amino acids were aligned using *MUSCLE* (Edgar, 2004) with default settings. Maximum likelihood phylogenies were created in *phyML* (version 3.0) (Guindon et al., 2009) using the Whelan and Goldman (WAG) substitution model hosted at www.phylogeny.fr (Dereeper et al., 2008). Branch support was provided by the *Shimodaira-Hasegawa-like* approximate likelihood ratio test for branches (Anisimova and Gascuel, 2006), branches with less than 50% support were collapsed. Trees were annotated using *Figtree* (version 1.4) (http://tree.bio.ed.ac.uk/software/figtree/) and *Inkscape* (0.48.4) (www.inkscape.org).

Outgroup sequences were included for cellulase, polysaccharide lyase and pectinase trees. For the polysaccharide lyase tree, three *Dendroctonus ponderosae* polysaccharide lyase sequences (Pauchet et al., 2010) (Accessions: 315570650, 315570648 and 315570656) were used to root the tree. These are the three *D. ponderosae* polysaccharide lyase family 4 genes most similar to gall wasp contigs and transcripts in table 4.5. For the pectinase tree, one pectin lyase sequence was used from *Bacillus subtilis* (Accession 16078925), as all the pectin lyases outgroup sequences in table 4.5 are outgroup to cynipid sequences when tested (appendix figure 4.11). For the cellulases, five outgroup sequences were used, two from *Cellvibrio japonicus* (Accessions: 192360233, 806574) and three from *Saccharophagus spp.* (Accessions: 269965254, 90022879, 90022881). These proteins were trimmed to 340 amino acids reflecting the maximum alignment position in table 4.5 for cynipid to bacterial cellulase sequences. The rhamnogalacturonate lyase was not rooted with outgroup sequences (appendix figure 4.10).

One can immediately see that broader phylogenetic relationships among cynipid tribes are maintained across the four genes (figure 4.1 compared with figure 4.2-4 and appendix figures 4.11-12), sequences from each species are labelled with different colours. *B. pallida* (coloured red) and *B. gibbera* (blue) intermingle across phylogenies, in many cases genes there are pairs of PCWDEs in the phylogenies for these two species. *A. quercuscalicis* (yellow) nestles with the other Cynipini in the pectinase phylogeny (figure 4.3), but is unresolved in the cellulase tree (figure 4.2). *D. spinosa* (green) and *P. aceris* (fuchsia) sequences are consistently separate to the other species (figure 4.2-4). The Cynipini and Pediaspini are proposed as sister tribes (figure 4.1, and Ronquist & Liljeblad, 2001) within the *Cynipidae*. However the polysaccharide lyase phylogeny (figure 4.4) shows *D. spinosa* and *P. aceris* forming a monophyletic clade in the tree that splits basally from the Cynipini sequences.

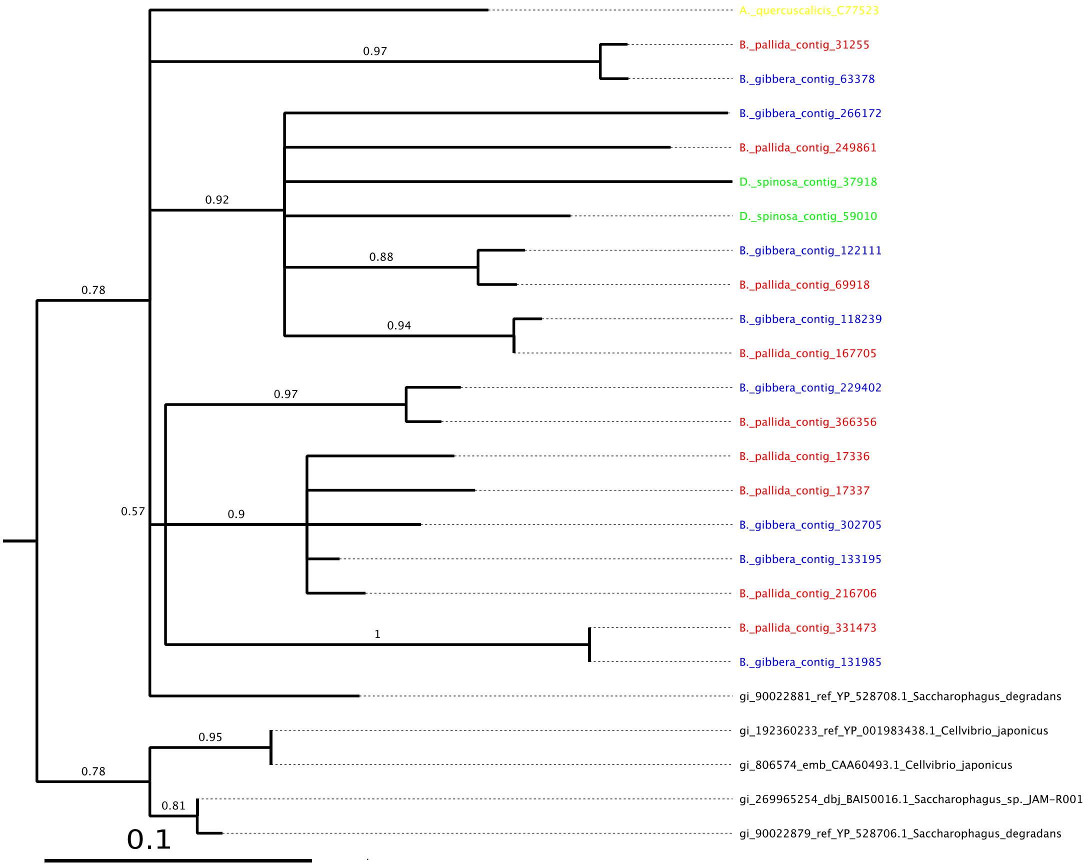

Figure 4.2. Phylogeny of cellulase genes identified and passing alignment criteria *B. gibbera* = red, *B. pallida* = blue, *D. spinosa* = green, *A. quercuscalicis* = yellow, *P. aceris* = fuchsia and outgroup sequences = black. Scale bar is substitutions per site and branches are labeled with approximate likelihood ratio test support values.

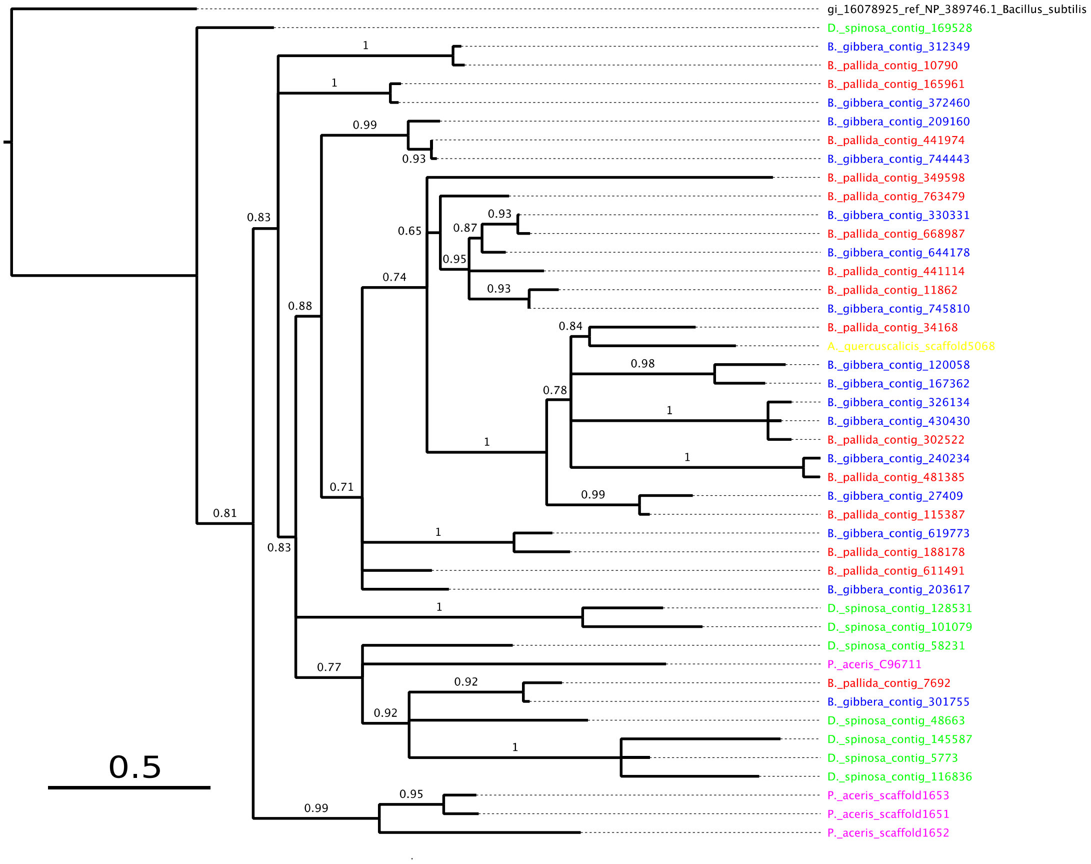

Figure 4.3. Phylogeny of pectinase genes identified and passing alignment criteria *B. gibbera* = red, *B. pallida* = blue, *D. spinosa* = green, *A. quercuscalicis* = yellow, *P. aceris* = fuchsia and outgroup sequences = black. Scale bar is substitutions per site and branches are labeled with approximate likelihood ratio test support values.

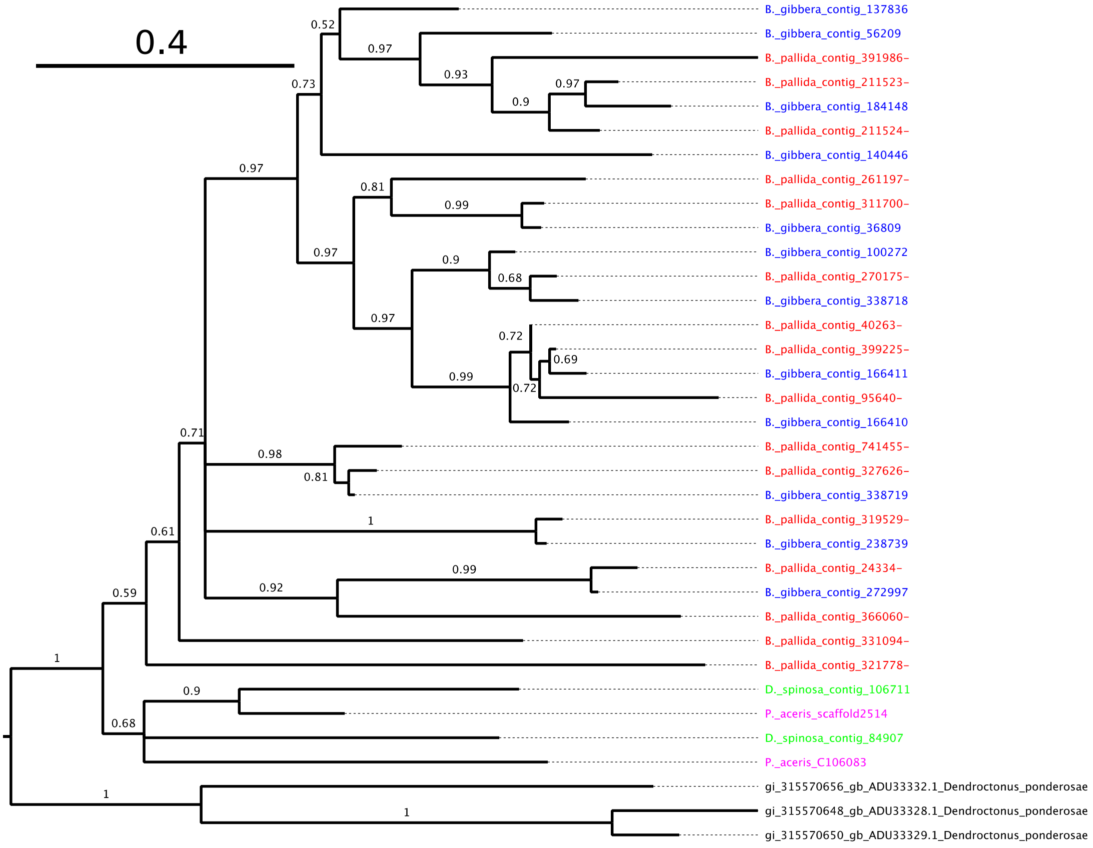

Figure 4.4. Phylogeny of polysaccharide lyase genes identified and passing alignment criteria *B. gibbera* = red, *B. pallida* = blue, *D. spinosa* = green, *A. quercuscalicis* = yellow, *P. aceris* = fuchsia and outgroup sequences = black. Scale bar is substitutions per site and branches are labeled with approximate likelihood ratio test support values.

4.2.9 Expression of PCWDE genes in *B. pallida* larvae

Plant cell wall degrading enzyme like-genes expressed in the *B. pallida* larval transcriptome in the RNA sequencing experiment (Chapter 3) were identified from *BLAST* annotations of transcripts (Chapter 3, part C). For expressed PCWDE candidate genes differentially and highly expressed in early larvae a role in gall induction is possible. Alternatively, expression at later stages could indicate a role in digestion of host cell walls during feeding by the mature larvae.

A total of thirty-three *B. pallida* larval transcripts were annotated as PCWDE of non-insect origin, of which twenty-four are detectable in the *B. pallida* genome assembly using *BLAST* (without an e-value threshold) and nine are not (tables 4.7-8). Two transcripts encoding cellulase, with divergent sequences, are highly expressed through gall development (expression counts across replicates, appendix table 4.12), but were not differentially expressed in either *edgeR* or *DESeq* analyses (Chapter 3, part B) across larval developmental stages. Because they are not differentially expressed, these cellulases may have a continuous role in gall development.

Eighteen of the thirty-three transcripts are most similar to PCWDE genes in the mountain pine beetle *D. ponderosae* (table 4.7-8). These *D. ponderosae* genes are hypothesized to result from ancient horizontal gene transfers from bacteria (Pauchet et al., 2010). The *D. ponderosae*-like transcripts encode nine polysaccharide lyases, glycoside hydrolase family protein 48 (cellulase), endo-beta-1,4-glucanase (xylanase), endopolygalacturonase and pectin methylesterase. However, only half (nine) of the transcripts most similar to *D. ponderosae* PCWDE genes have corresponding contigs in the genome assembly, and all of these are polysaccharide lyases family 4 proteins (table 4.7, *B. pallida* larval PCWDE transcripts without corresponding *B. pallida* genomic regions).

The thirty-three expressed *B. pallida* larval PCWDE encoding transcripts were also *BLAST* searched against the *B. gibbera* assembly and the same twenty-four contigs were detectable as against the *B. pallida* assembly (appendix table 4.13). Therefore, the same nine transcripts without corresponding contigs in *B. pallida* (table 4.8) have no corresponding contigs in the *B. gibbera* assembly; they may not derive from *B. pallida*. This potential expression from other species is possible in the gall wasp transcriptome, as the transcriptome was not filtered for sequences from closely related species (Chapter 3, part A). For example, the three *D. ponderosae* glycoside hydrolase family protein 48 (cellulases) transcripts are only expressed in replicate 270C (appendix table 4.12). Possible alternative sources of this PCWDE expression are parasitoids, cynipid inquilines, and other gall inhabiting insects. A parasitoid that eats plant tissue prior to feeding on gall wasp larvae could also have acquired PCWDEs, an example of a koinobiont lifestyle. The cynipid chalcid parasitoid *Eurytoma brunniventris* is known to eat gall tissue before galler larvae in this manner (Askew, 1984).

None of the *B. pallida* larval transcripts have high expression other than the two cellulases, but three polysaccharide genes and four pectin lyases are differentially expressed (*DESeq* results, table 4.6 and expression counts across replicates appendix table 4.12).

Gene	**Annotation**	**Early vs. Growth**	**Early vs. Mature**	**P-value**	**Adjusted P-value**
comp14239_c0	polysaccharide lyase family protein 4	-0.08	1.52	3.46E-05	0.000290757
comp18312_c0	pectin lyase	0.12	-5.53	0.002830762	0.01425849
comp20769_c0	pectin lyase	-1.89	2.53	0.00162793	0.00892196
comp20790_c0	pectate lyase	-2.20	-6.90	7.43E-06	7.25E-05
comp26878_c0	polysaccharide lyase family protein 4	-0.84	-3.99	1.22E-07	1.61E-06
comp26878_c1	polysaccharide lyase family protein 4	-1.92	-6.08	0.002447804	0.012616273
comp59931_c0	pectin lyase	-2.55	-32.89	0.002100293	0.011078433

Table 4.6. Significantly differentially expressed PCWDE genes in the B. pallida transcriptome (see Chapter 3). Early versus Growth and Early vs. Mature indicate fold change in expression. A negative value means the gene is more highly expressed in the early stage and *vice-a-versa* for positive values. Adjusted p-value is the p-value for this gene after correction for multiple testing of all the genes analysed in the RNASeq experiment in chapter 4.

Table 4.7 *BLAST* best hits for larvally expressed PCWDE-encoding *B. pallida* transcripts in the *B. pallida* genome assembly.

Transcript	Genome tophit	% identity	Alignment length	Mismatches	Gaps	Query start	Query end	Ref. Start	Ref. End	E-value	Bit score	Transcript PCWDE
comp103141_c0	contig_10790	99.07	429	4	0	1	429	1773	2201	0	756	pectin lyase
comp132692_c0	contig_668987	87.77	319	2	1	33	351	38	319	6.00E-119	430	pectin lyase
comp13836_c0	contig_167705	100	990	0	0	175	1164	1505	516	0	1786	cellulase
comp14276_c0	contig_165961	98.93	560	6	0	1	560	1163	1722	0	984	pectin lyase
comp14276_c1	contig_165961	99.8	496	1	0	1	496	690	1185	0	890	pectin lyase
comp18312_c0	contig_10790	99.66	589	2	0	268	856	2793	2205	0	1054	pectin lyase
comp20769_c0	contig_441974	99.91	1056	1	0	7	1062	1848	793	0	1900	pectin lyase
comp20790_c0	contig_11862	99.9	979	1	0	100	1078	3236	4214	0	1761	pectin lyase
comp258830_c0	contig_763479	99.64	279	1	0	1	279	370	92	1.00E-139	499	pectin lyase
comp27915_c0	contig_331473	95.62	548	24	0	215	762	548	1	0	881	cellulase
comp282070_c0	contig_11862	85.57	201	29	0	1	201	3772	3972	2.00E-59	232	pectin lyase
comp326228_c0	contig_366356	100	239	0	0	1	239	473	711	1.00E-119	432	cellulase
comp57254_c0	contig_69918	99.9	985	1	0	88	1072	1269	2253	0	1772	cellulase
comp59931_c0	contig_188178	99.15	468	3	1	15	481	1916	1449	0	823	pectin lyase
comp98838_c0	contig_331094	99.45	361	0	1	202	560	138	498	0	639	Rhamnogalacturonate lyase
comp27190_c0	contig_211523	99.84	1893	3	0	102	1994	2872	980	0	3400	polysaccharide lyase family protein4
comp14239_c0	contig_319529	100	2208	0	0	77	2284	60	2267	0	3983	polysaccharide lyase family protein4
comp26878_c0	contig_310874	100	753	0	0	126	878	973	221	0	1359	polysaccharide lyase family protein4
comp24519_c0	contig_40263	95.61	638	28	0	12	649	1	638	0	1025	polysaccharide lyase family protein4
comp100388_c0	contig_24334	99.94	1609	1	0	17	1625	1	1609	0	2897	polysaccharide lyase family protein4
comp26878_c1	contig_327649	100	248	0	0	219	466	261	14	3.00E-124	448	polysaccharide lyase family protein4
comp198252_c0	contig_95640	100	415	0	0	1	415	2356	2770	0	749	polysaccharide lyase family protein4
comp53817_c0	contig_327626	90.02	571	49	5	1	570	1092	529	0	758	polysaccharide lyase family protein4
comp24519_c1	contig_759421	100	370	0	0	265	634	370	1	0	668	polysaccharide lyase family protein4

Transcript	Accession	% identity	length	Mismatch	Gap	Query start	Query end	Ref. Start	Ref. End	E-value	Bit score	Annotation
comp93909_c0	gi|315570583|gb|HM175791.1|	77.24	769	167	4	1	765	834	70	4.00E-164	587	glycoside hydrolase family protein 48
comp150589_c0	gi|315570583|gb|HM175791.1|	76.3	422	94	4	8	426	1545	1127	2.00E-77	298	glycoside hydrolase family protein 48
comp14831_c0	gi|315570583|gb|HM175791.1|	74.51	455	104	3	4	458	1969	1527	2.00E-76	295	glycoside hydrolase family protein 48
comp136735_c0	gi|315570603|gb|HM175801.1|	70.45	538	153	4	1	535	1036	502	5.00E-60	241	endopolygalacturonase (GH28Pect-5)
comp151243_c0	gi|315570568|gb|ADU33288.1|	76.56	128	30	0	385	2	23	150	3.00E-47	191	endo-beta-1,4-glucanase
comp350819_c0	gi|315570565|gb|HM175782.1|	72.08	308	83	1	18	322	37	344	5.00E-38	167	endo-beta-1,4-glucanase
comp258611_c0	gi|315570574|gb|ADU33291.1|	78.48	79	17	0	4	240	59	137	1.00E-31	139	endo-beta-1,4-glucanase
comp213754_c0	gi|315570632|gb|ADU33320.1|	55.24	143	57	3	13	438	1	137	3.00E-29	132	endopolygalacturonase
comp397147_c0	gi|315570640|gb|ADU33324.1|	61.97	71	27	0	213	1	143	213	1.00E-19	100	pectin methylesterase

Table 4.8 *BLAST* results for transcripts without corresponding regions in the *B. pallida* genome assembly. The same nine transcripts also lack corresponding regions in the *B. gibbera* genome assembly (table 4.13).

4.3 Evidence for a prokaryotic or eukaryotic origin of the PCWDE genes

4.3.1 Introns are present in some PCWDE genes

Introns are characteristic of eukaryotic genomes, intronless genes do exist but are very rare (Sakharkar et al., 2002; Sakharkar and Kangueane, 2004) Introns were searched for in *B. pallida* genomic contigs corresponding to the expressed *B. pallida* larval PCWDEs.

This was restricted to the twenty-four PCWDEs expressed in the *B. pallida* larval transcriptome with corresponding contigs in the *B. pallida* genome assembly (table 4.7). The expression of these genes is assumed to indicate that they are functional. These genes were nucleotide *BLAST* (*BLASTn*) searched against the *B. pallida* genome without an e-value thresholds or maximum number of hits. The purpose was to identify the entire length of the expressed transcripts across one or more genomic contigs. Figure 4.5 explains the logic used to identify introns occurring in the genome using PCWDE transcripts. Firstly 1), when different exons of the transcript fall in different contigs of the genome assembly. This is because of the short average length of contigs in the draft assembly (table 4.3). The second scenario, 2) is preferable as the two exons fall along the same contig and the intron length can be estimated.

Only transcripts with alignments to contigs that did not overlap the start or end of genomic contigs were considered. This is because a transcript alignment overlapping a contig end indicates an incompletely assembled genomic fragment in which the start or end of an exon cannot be predicted.

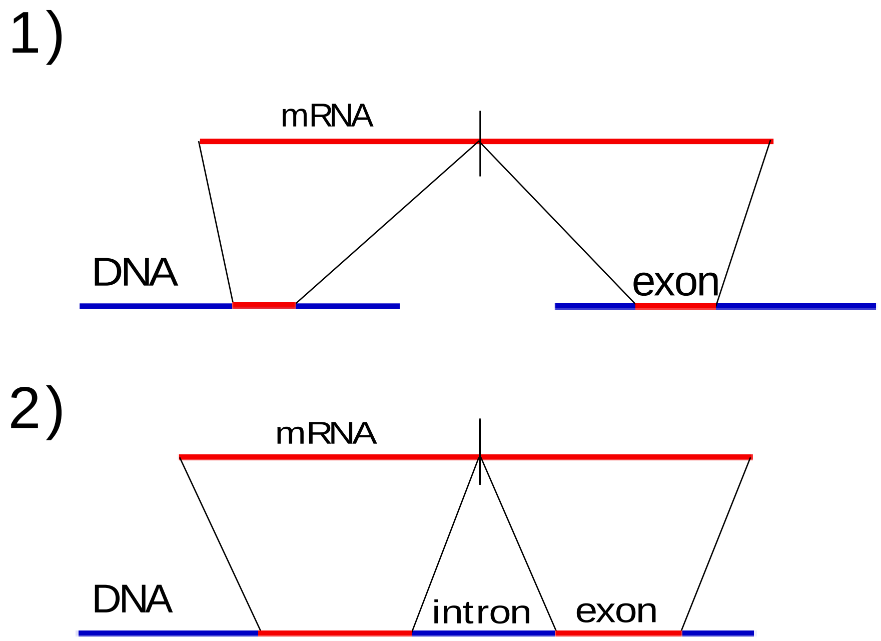

Figure 4.5. The two scenarios possible when *BLAST* searching for introns by comparison of expressed transcripts (mRNA) against a draft genome (DNA). Scenario 1) shows the expectation when the transcript maps to different genomic contigs, meaning the length of the intron cannot be known but intron donor and acceptor sites are identifiable.

Five *B. pallida* transcripts had evidence for introns in corresponding genomic contigs, four of which conform to scenario 1) above and one to scenario 2) (table 4.9). Table 4.9 shows that *B. pallida Contig_69918-*, encoding a cellulase (table 4.5) contains an intron 543 bp long, flanked by exons for the transcript *comp57254_c0*. To confirm this, 5' and 3' intron splicing consensus sequences were identified. At the intron start position (base 726, end of first exon) there is a 5' intron splicing site consensus sequence (donor) 'TGGTAAGT' and at the end of the intron (base 1269, start of second exon) the 3’ intron splicing site consensus sequence (acceptor), ‘CAG’. These splice sites were also predicted by the splice site predictor *SplicePort* (Dogan et al., 2007). *SplicePort* was also used to predict splice sites at the boundaries of *BLAST* alignments occurring for the other candidate intron containing contigs (table 4.10). *SplicePort* classifies intron splices sites as donor (5’) or acceptor (3’) and provides a score for predicted sites. This score is derived from the number of splice site criteria a site has, the higher the score the more confident *SplicePort* is of correct splice site identification (Dogan et al., 2007). For the genomic contigs with scenario 1-like potential introns (figure 4.6) the intron acceptor and donor splice sites are on different contigs so it is not possible to know intron length. In total seven of fourteen potential splice sites had *SplicePort* detectable splice sites.

Transcript	Genomic contig containing exon	Scenario	Alignment length	Transcript length	Transcript start	Transcript end	Exon start	Exon end	Contig length	Bit score
comp57254_c0_	Biorhiza_pallida_contig_69918-	2	88	1076	1	88	639	726	2888	159
comp57254_c0	Biorhiza_pallida_contig_69918-	2	985	1076	88	1072	1269	2253	2888	1772
comp27190_c0	Biorhiza_pallida_contig_560948-	1	88	1909	1	88	110	197	388	150
comp27190_c0	Biorhiza_pallida_contig_211523-	1	1729	1909	94	1822	2708	980	2872	3104
comp26878_c0	Biorhiza_pallida_contig_864867-	1	68	1150	1	68	187	120	267	123
comp26878_c0	Biorhiza_pallida_contig_311700-	1	73	1150	70	142	1935	1863	2846	87.8
comp26878_c0	Biorhiza_pallida_contig_95640-	1	995	1150	146	1140	2958	2001	3626	890
comp20769_c0	Biorhiza_pallida_contig_441974-	1	1056	1175	7	1062	1848	793	2356	1900
comp20769_c0	Biorhiza_pallida_contig_381950-	1	114	1175	1062	1175	2204	2091	2618	206
comp18312_c0	Biorhiza_pallida_contig_236172-	1	269	856	1	269	1636	1904	2131	486
comp18312_c0	Biorhiza_pallida_contig_10790-	1	589	856	268	856	2793	2205	4631	1054
comp132692_c0	Biorhiza_pallida_contig_375511-	1	37	351	1	37	1798	1834	2679	62.6
comp132692_c0	Biorhiza_pallida_contig_668987-	1	319	351	33	351	38	319	1447	430

Table 4.9 *BLAST* results for transcripts with potential introns corresponding regions of the *B. pallida* genome assembly. Scenario is described by figure 4.5. There is only one example of scenario 2, in which the intron is on the same genomic contig as the flanking exons. Exon start and Exon end columns refer to the regions within a genomic contig mapping to different regions of an expressed PCWDE transcript.

Transcript	**Contig**	**Contig position**	**Present**	**Donor/ Acceptor**	**Splice site**	**Score**
comp57254_c0_	Biorhiza_pallida_contig_69918-	726	Yes	D	tgctggtaagta	1.65797
comp57254_c0	Biorhiza_pallida_contig_69918-	1269	Yes	A	ttaacagatgtg	0.813637
comp27190_c0	Biorhiza_pallida_contig_560948-	197	No	D	N/A	N/A
comp27190_c0	Biorhiza_pallida_contig_211523-	2708	No	A	N/A	N/A
comp26878_c0	Biorhiza_pallida_contig_864867-	2001	Yes	D	cttcagtaagtc	0.595475
comp26878_c0	Biorhiza_pallida_contig_311700-	1935	Yes	A	ttatcagttgaa	1.88363
comp26878_c0	Biorhiza_pallida_contig_311700-	1863	Yes	D	atggtgtaagtc	0.991692
comp26878_c0	Biorhiza_pallida_contig_95640-	2958	No	A	N/A	N/A
comp20769_c0	Biorhiza_pallida_contig_441974-	793	No	D	N/A	N/A
comp20769_c0	Biorhiza_pallida_contig_381950-	2204	No	A	N/A	N/A
comp18312_c0	Biorhiza_pallida_contig_236172-	1904	No	D	N/A	N/A
comp18312_c0	Biorhiza_pallida_contig_10790-	2793	Yes	A	tatgcaggcgct	1.72909
comp132692_c0	Biorhiza_pallida_contig_375511-	1834	Yes	D	ttttggtgagat	0.568529
comp132692_c0	Biorhiza_pallida_contig_668987-	38	No	A	N/A	N/A

Table 4.10 *SplicePort* predictions from potential intron donor and acceptor splice sites derived from table 4.9.

For comparison to cynipid PCWDE genes, the PCWDE polysaccharide lyases of the beetle *D. ponderosae* (Genbank accessions: 315570650, 315570648 and 315570656) similar to cynipid contigs and transcripts were checked for introns. *D. ponderosae* expressed sequence tags (ESTs) of the polysaccharide lyases (Pauchet et al., 2010) were *BLAST* searched against the draft *D. ponderosae* genome (Keeling et al., 2013) (http://www.ncbi.nlm.nih.gov/genome/11242). An EST to genome *BLAST* revealed one of the polysaccharide lyases (Accession: 31557064), 1 778 bp long, contained an obvious intron between positions 363 and 416. This was confirmed by a *GENSCAN* (http://genes.mit.edu/GENSCAN.html) (Burge and Karlin, 1997) search for introns in the corresponding genomic contig (Accession: 315570648).

4.3.2 Hymenopteran genes present on genomic contigs encoding PCWDEs

Another way of demonstrating a HGT event is when genes of host origin surround the candidate HGT genes; i.e. genes shared by other hymenopterans and insects for cynipids. This is easiest to demonstrate with finished genomes, for example the horizontally transferred PCWDE genes of the plant parasitic nematode *Meloidogyne incognita* (Abad et al., 2008). Additionally, in *M. incognita* the PCWDE genes have evolved introns since their acquisition by the nematode (Abad et al., 2008).

With the draft assemblies available here this is much more difficult, PCWDEs of putative prokaryote origin have been detected in contigs of ~6000 bp maximum length. The relative shortness of these contigs reduces the chance of identifying multiple open reading frames per contig. Open reading frames were extracted from genomic contigs using *getorf* from the EMBOSS package (Longden and Bleasby, 2000). These ORFs were then *BLAST* searched against the non-redundant (nr) *BLAST* database.

One polysaccharide lyase encoding contig of *B. pallida* (contig_321778-, 3 183 bp length) also had an ORF most similar to a hypothetical gene of the hymenopteran *Nasonia vitripennis* (e-value 1 x 10^{-6}; LOC100114674; Accession: 156540059). The region of the hypothetical gene overlapping the *B. pallida* contig encodes a non-LTR RNAse H class I domain of reverse transcriptase. Additionally, *B. pallida* contig_11862 (length 5652 bp) encoding a pectate lyase, also contains a transposase most similar to a DDE superfamily endonuclease *Caenorhabditis briggsae* (e-value 2 x 10^{-23}). The best hits to the transposase-encoding region of contig_11862 from the *B. pallida* transcriptome (comp157814_c0_seq1, e-value 4 x 10^{-90}), *B. gibbera*, and *D. spinosa* assemblies are also *Caenorhabditis spp.* transposases of the same DDE superfamily.

4.3.3 Codon usage bias of plant cell wall degrading enzyme genes

Codon usage refers to the frequency of each synonymous codon for a particular amino acid; if one codon is preferred over the others there is a bias in codon usage for that amino acid. Pauchet et al. (2010) used codon usage bias as evidence for a coleopteran bias in PCWDEs of hypothesized prokaryotic origin identified in beetle ESTs. They demonstrated that codon usage frequencies were similar to that observed in other insects and contrasted with that of *Wolbachia* as an example of a prokaryote.

A similar analysis was repeated here for cynipid PCWDEs using the *B. pallida* larval transcriptome (chapter 3) and jewel wasp (*Nasonia vitripennis*) official gene set version 1.2 (Munoz-Torres et al., 2011) as hymenopteran references. For potential donor bacteria species, gene sets of species close most similar to the identified PCWDEs were chosen (downloaded from http://www.ncbi.nlm.nih.gov/genome). Bacterial species chosen were *Bacillus subtilis, Cellvibrio japonicus, Dickeya dadantii, Paenibacillus subtilis,* and *Saccharophagus degradans*. The program *General Codon Usage Analysis* (*GCUA)* (McInerney, 1998) analysed codon usage for each dataset. The relative synonymous codon usage (RSCU) value was used to compare codons among datasets. An RSCU value is the number of times a codon is observed divided by the expected number of observations in the absence of codon usage bias (McInerney, 1998). Thus a value greater than one indicates bias for a particular codon, and a number less than one a bias against.

The results (table 4.11) are equivocal. There is no clear pattern to PCWDE codon usage bias with respect to the eukaryote versus prokaryote references; some codons are more similar to hymenopteran sequences, others to the prokaryotic sequences. However, a principal components analysis (PCA) of the RSCU shows clear associations (figure 4.6). The first two components of the PCA explain 40 and 32% of variation in the data respectively, the third 11%. The PCWDEs are highly correlated with the *B. pallida* transcriptome and the *N. vitripennis* gene set. The bacteria cluster together but are not correlated with the hymenopteran sequences and PCWDEs.

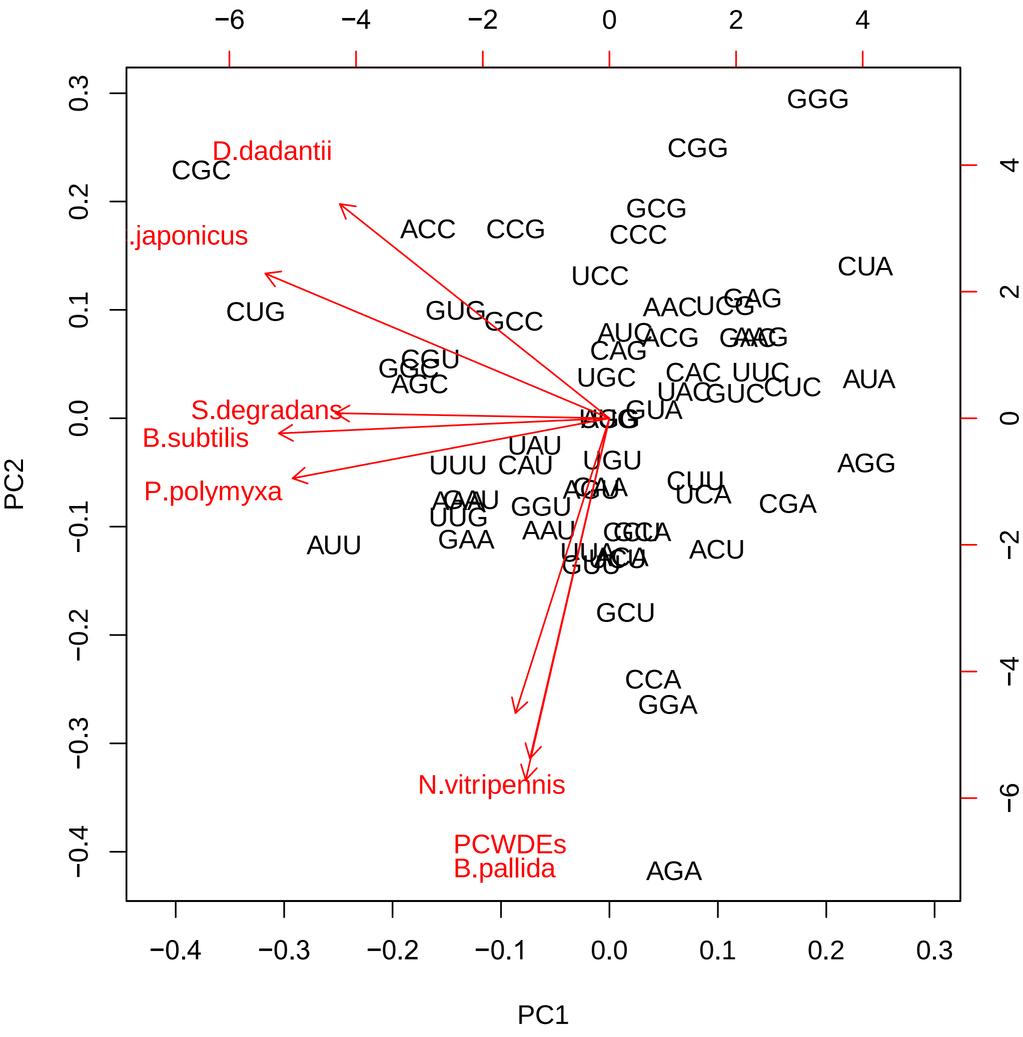

Figure 4.6. Principal components analysis biplot of amino acid (black) usage and species included in the RSCU analysis (red). Plant cell wall degrading enzymes were collated together across species and are labelled PCWDEs.

Table 4.11. Codon usage table of RSCU values for each amino acid across the gene sets tested, numbers in red are the highest RSCU value for that codon for that species. All datasets were of coding sequence beginning at position 1 forward strand.

Amino Acid	**Codon**	**Cynipid PCWDEs**	***B. pallida* transcriptome**	***N. vitripennis***	***B. subtilis***	***P. polymxa***	***C. japonicus***	***D. dadantii***	***S. degradans***
Phe	UUU	**1.35**	0.98	**1.04**	**1.42**	**1.33**	**1.28**	**0.93**	**1.49**
	UUC	0.65	**1.02**	0.96	0.58	0.67	0.72	1.07	0.51
Leu	UUA	**1.77**	0.95	1.10	**1.28**	0.81	0.64	0.33	**1.65**
	UUG	1.17	**1.24**	1.12	**1.28**	1.47	1.41	0.90	1.22
	CUU	1.11	1.17	0.86	0.95	0.88	0.5	0.32	0.99
	CUC	0.44	1.02	1.03	0.71	0.55	0.71	0.48	0.40
	CUA	0.75	0.54	0.65	0.53	0.50	0.22	0.18	0.97
	CUG	0.77	1.09	**1.24**	1.25	**1.79**	**2.52**	**3.80**	0.78
Tyr	UAU	**1.45**	0.94	0.92	**1.24**	**1.34**	**1.13**	0.98	0.90
	UAC	0.55	**1.06**	**1.08**	0.76	0.66	0.87	**1.02**	**1.10**
His	CAU	**1.27**	**1.05**	0.96	**1.30**	**1.44**	**1.04**	0.96	0.82
	CAC	0.73	0.95	**1.04**	0.70	0.56	0.96	**1.04**	**1.18**
Gln	CAA	**1.18**	**1.08**	**1.02**	**1.14**	0.98	0.84	0.50	**1.28**
	CAG	0.82	0.92	0.98	0.86	**1.02**	**1.16**	**1.50**	0.72
Ile	AUU	**1.51**	**1.26**	**1.07**	**1.71**	**1.66**	**1.43**	1.13	**1.68**
	AUC	0.62	0.99	0.98	0.99	0.90	1.22	**1.65**	0.45
	AUA	0.88	0.75	0.95	0.30	0.44	0.36	0.22	0.86
Met	AUG	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**
Asn	AAU	**1.38**	**1.16**	**1.10**	**1.19**	**1.22**	**1.06**	0.78	0.94
	AAC	0.62	0.84	0.90	0.81	0.78	0.94	**1.22**	**1.06**
Lys	AAA	**1.35**	**1.15**	**1.13**	**1.41**	**1.16**	**1.25**	**1.32**	**1.41**
	AAG	0.65	0.85	0.87	0.59	0.84	0.75	0.68	0.59
Val	GUU	**1.48**	**1.27**	**1.07**	1.06	0.98	0.94	0.55	1.18

	GUC	0.58	1.02	1.06	0.85	0.74	0.78	0.94	0.22
	GUA	1.15	0.83	0.90	0.71	1.12	0.77	0.47	1.40
	GUG	0.79	0.89	0.97	**1.39**	**1.17**	**1.50**	**2.03**	**1.20**
Asp	GAU	**1.37**	**1.13**	**1.06**	**1.28**	**1.38**	**1.33**	0.99	**1.27**
	GAC	0.63	0.87	0.94	0.72	0.62	0.67	**1.01**	0.73
Glu	GAA	**1.55**	**1.23**	**1.10**	**1.42**	**1.18**	**1.18**	**1.26**	**1.27**
	GAG	0.45	0.77	0.90	0.58	0.82	0.82	0.74	0.73
Ser	UCU	**1.63**	**1.24**	0.92	1.06	1.03	0.63	0.48	1.18
	UCC	0.63	0.79	0.77	1.20	**1.19**	1.19	1.20	0.53
	UCA	1.38	1.04	0.98	0.93	0.82	0.68	0.48	0.61
	UCG	0.34	0.85	1.02	0.52	0.8	0.76	1.31	0.87
	AGU	1.21	1.08	1.09	1.07	0.99	1.15	0.60	1.06
	AGC	0.81	0.98	**1.22**	**1.23**	1.16	**1.59**	**1.93**	**1.75**
Cys	UGU	**1.38**	0.98	0.92	**1.10**	**1.08**	0.900	0.67	0.91
	UGC	0.62	**1.02**	**1.08**	0.90	0.92	**1.10**	**1.33**	**1.09**
Trp	UGG	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**
Pro	CCU	1.23	1.27	0.96	0.92	1.17	0.72	0.41	1.01
	CCC	0.59	0.63	0.76	**1.30**	0.62	**1.38**	0.68	1.00
	CCA	**1.70**	**1.39**	**1.29**	0.66	0.93	0.63	0.31	**1.27**
	CCG	0.48	0.71	0.99	1.11	**1.28**	1.27	**2.60**	0.72
Arg	CGU	1.30	0.91	0.73	1.17	**1.80**	1.42	1.53	1.33
	CGC	0.57	0.83	0.90	**1.38**	1.23	**3.17**	**2.59**	**2.86**
	CGA	1.28	1.04	1.00	0.60	0.75	0.32	0.32	0.61
	CGG	0.45	0.43	0.53	1.32	0.98	0.61	1.15	0.35
	AGA	**1.69**	**1.73**	**1.75**	0.98	0.83	0.17	0.22	0.52

	AGG	0.71	1.06	1.10	0.55	0.40	0.30	0.19	0.33
Thr	ACU	**1.45**	**1.19**	1.02	0.81	0.73	0.56	0.33	0.87
	ACC	0.67	0.83	0.88	**1.30**	0.94	**2.10**	**2.17**	**1.58**
	ACA	1.38	1.13	**1.11**	1.06	1.15	0.71	0.30	0.93
	ACG	0.50	0.84	0.99	0.84	**1.18**	0.63	1.20	0.62
Ala	GCU	1.46	**1.31**	**1.16**	0.96	**1.15**	0.65	0.39	0.89
	GCC	0.54	0.97	1.03	**1.35**	0.88	**1.61**	**1.53**	1.10
	GCA	**1.48**	1.09	1.03	0.80	1.09	0.82	0.36	**1.11**
	GCG	0.51	0.64	0.77	0.89	0.88	0.91	1.72	0.91
Gly	GGU	1.42	1.16	1.05	0.92	1.12	1.36	0.80	1.42
	GGC	0.62	1.05	1.24	**1.30**	1.03	**1.61**	**2.16**	**1.84**
	GGA	**1.62**	**1.45**	**1.31**	0.97	**1.17**	0.45	0.31	0.23
	GGG	0.35	0.34	0.41	0.82	0.68	0.58	0.73	0.51

4.3.4 Identifying potential Shine-Dalgarno sequences in the 5' untranslated region of expressed PCWDEs

Within the 5' untranslated region (UTR) of most prokaryotic mRNAs is the Shine-Dalgarno sequence (AGGAGG) (Shine and Dalgarno, 1975). The sequence acts as a ribosomal binding site for 16s ribosomal RNA and occurs 6-7 bases upstream of the start codon. The 5' untranslated regions (UTR) of the genomic contigs corresponding to expressed *B. pallida* PCWDE transcripts were tested for the Shine-Dalgarno (SD) sequence and subsequences (GAGG, AGGA, and GGAG) to account for variation in the SD sequence. The presence of a SD sequence would confirm the contigs as of prokaryote origin. However, the lack of an SD sequence would not confirm the transcripts are eukaryotic as some prokaryotes lack the SD sequence (Lim et al., 2012).

The UTR regions of PCWDE transcripts were extracted, and reverse complemented where necessary and the twenty bases adjacent to the start codon tested. Twenty-three contigs with UTRs greater than 5 bp were tested. No transcripts contained the full SD consensus sequence or any subsequences within 20 bases of the start codon.

4.3.5 *InterProScan* predicted structures of cynipid PCWDE genes

Figures 4.7-10 show the *InterProScan* predicted (Zdobnov and Apweiler, 2001) structure of the PCWDEs for each enzyme type discovered. The cellulase and pectin/pectate lyases have the simplest structure encoding only the enzyme domain and a eukaryotic signal peptide. The polysaccharide lyases and rhamnogalacturonate lyases have similar structures. They both contain a rhamnogalacturonate lyase, a galactose mutarotase-like domain, carboxypeptidase and a galactose-binding domain. The polysaccharide lyases from genomic contigs also encode eukaryotic signal peptides while the rhamnogalacturonate lyase of a *P. aceris* transcript does not. This *P. aceris* rhamnogalacturonate lyase is the only complete open reading frame for this enzyme across genetic resources tested. Incomplete rhamnogalacturonate lyases, with the 5' end present in the other species genome and transcriptome assemblies also do not have signal peptides. This may reflect variation in signal peptide presence between cynipid PCWDEs genes.

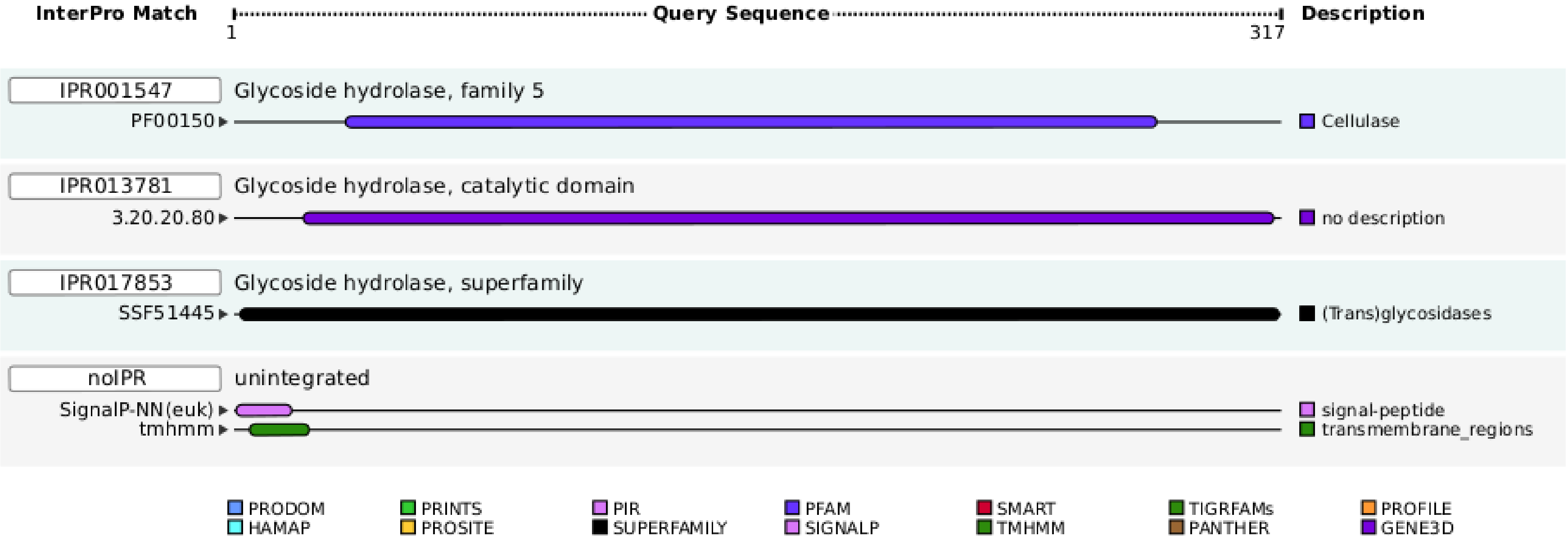

Figure 4.7. Cellulase protein encoded by comp_13836_c0 and contig_167705 of *B. pallida* transcriptome and genome assemblies respectively This diagram was generated using *InterProScan* http://www.ebi.ac.uk/Tools/pfa/iprscan/ web search using all available databases. Length of query sequences is in amino acids above annotations. Each row indicates a different annotation to the protein from a different database included in the *InterProScan* search, the colour of the annotation matches the colour of the database from which the annotation is derived.

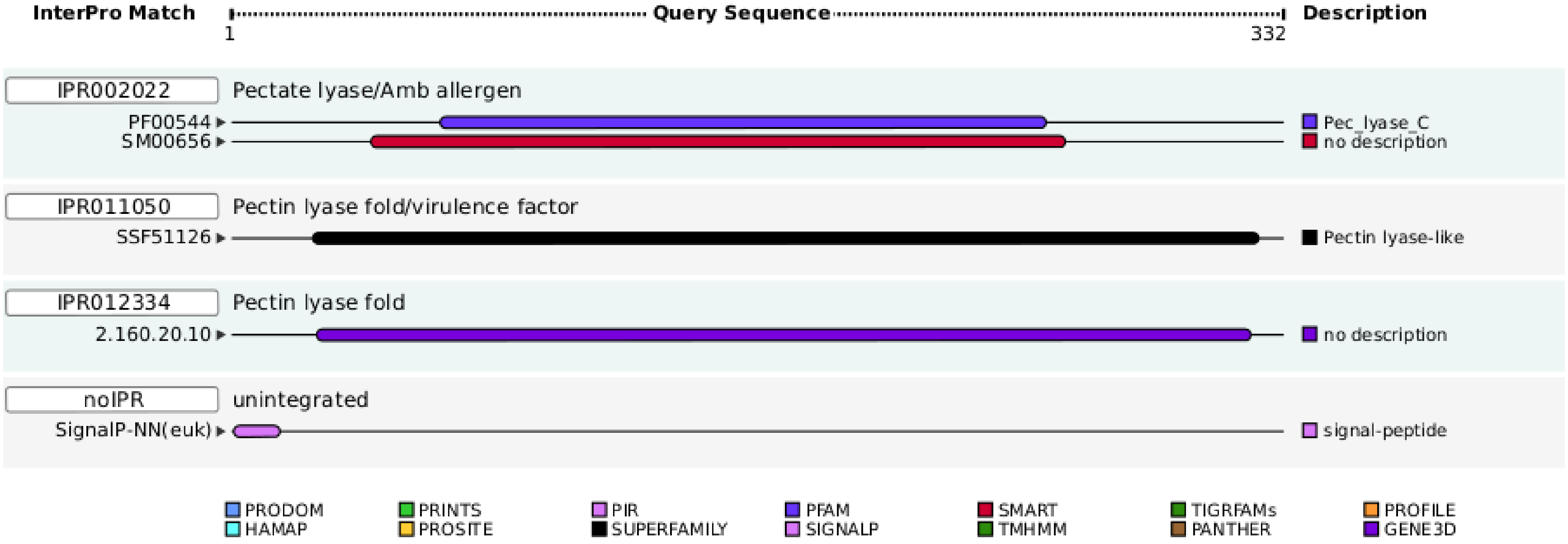

Figure 4.8. Pectin lyase protein encoded by comp_20769_c0 and contig_441974 of *B. pallida* transcriptome and genome assemblies respectively. This diagram was generated using *InterProScan* http://www.ebi.ac.uk/Tools/pfa/iprscan/ web search using all available databases. Length of query sequences is in amino acids above annotations. Each row indicates a different annotation to the protein from a different database included in the *InterProScan* search, the colour of the annotation matches the colour of the database from which the annotation is derived.

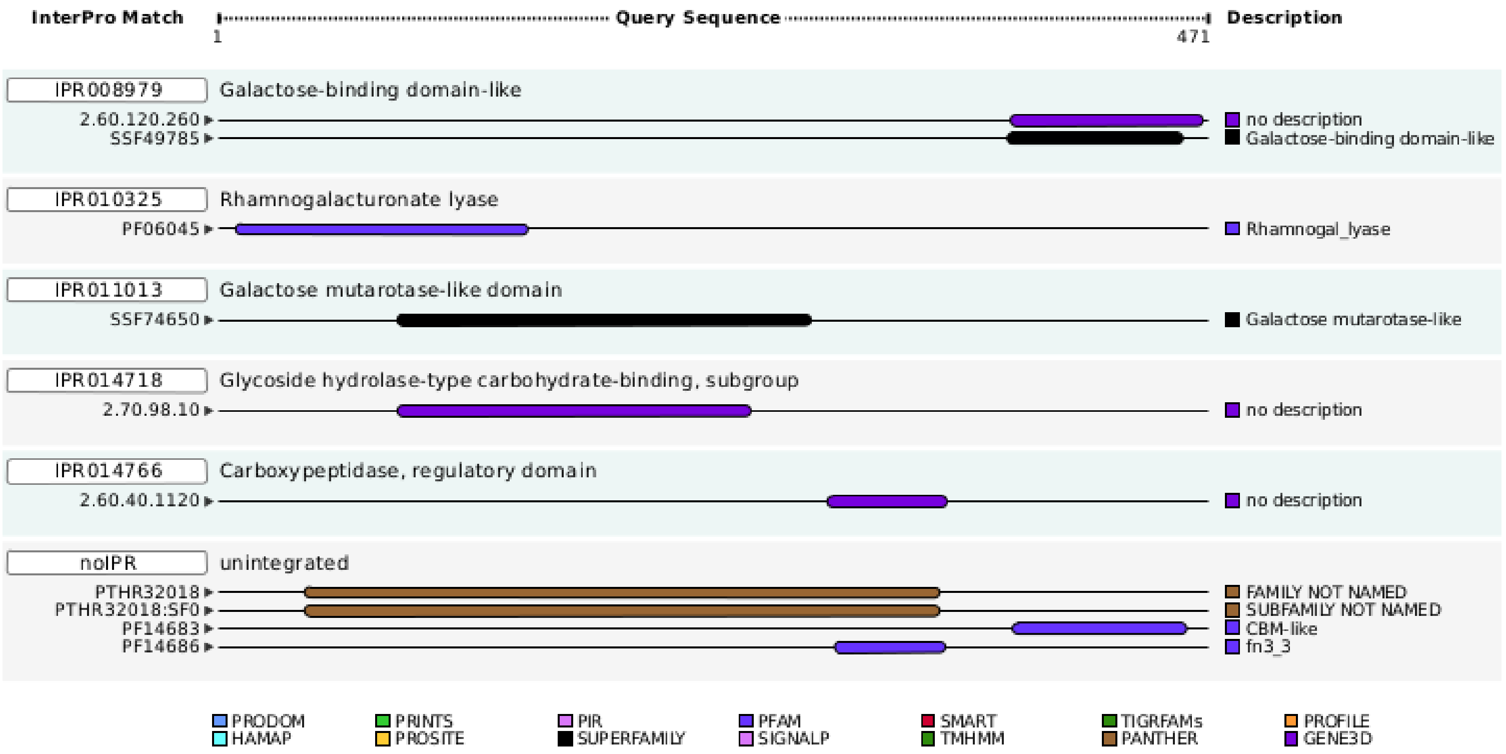

Figure 4.9. Rhamnogalacturonate lyase protein encoded by scaffold2513 of the *P. aceris* transcriptome. This diagram was generated using *InterProScan* http://www.ebi.ac.uk/Tools/pfa/iprscan/ web search using all available databases. Length of query sequences is in amino acids above annotations. Each row indicates a different annotation to the protein from a different database included in the *InterProScan* search, the colour of the annotation matches the colour of the database from which the annotation is derived.

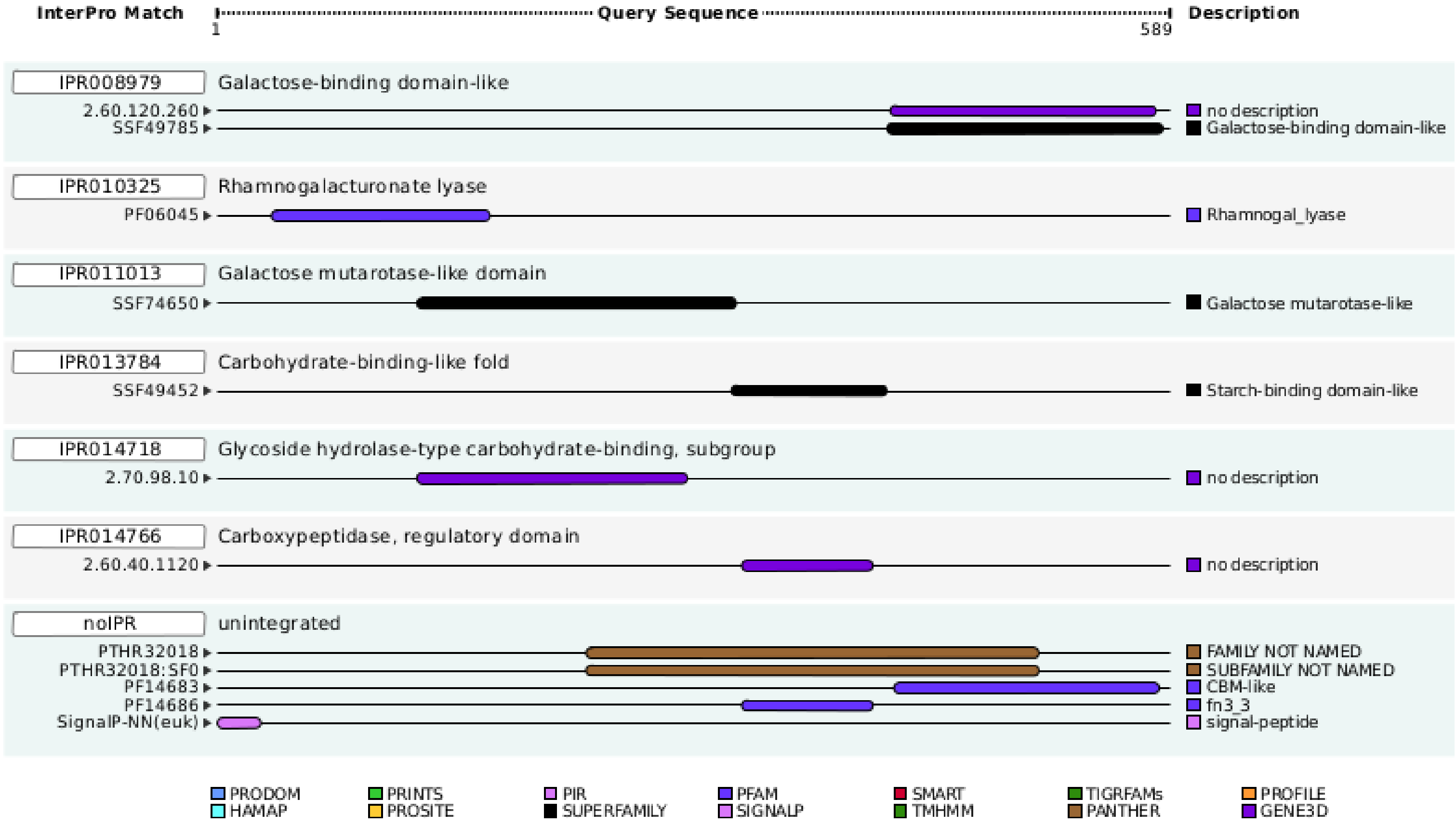

Figure 4.10. Polysaccharide lyase protein encoded by comp_27190_c0 and contig_211523 of *B. pallida* transcriptome and genome assemblies respectively. This diagram was generated using *InterProScan* http://www.ebi.ac.uk/Tools/pfa/iprscan/ web search using all available databases. Length of query sequences is in amino acids above annotations. Each row indicates a different annotation to the protein from a different database included in the *InterProScan* search, the colour of the annotation matches the colour of the database from which the annotation is derived.

4.4 Discussion

4.4.1 The presence of PCWDE genes in cynipid genomes by horizontal gene transfer is strongly indicated but remains unconfirmed

Plant cell wall degrading enzymes are present in genomic assemblies of the cynipid gall wasps *D. spinosa* (Diplolepidini), *B. gibbera* and *B. pallida* (Cynipini). These genes are not expressed in the transcriptome of an outgroup with a parasitoid life history, *L. clavipes*. Several PCWDEs genes are expressed during larval development in *B. pallida* (tables 4.7-8), and in the venom gland of adult cynipids (S. Cambier, personal communication). Phylogenies (figures 4.2-4) show these genes to be probable orthologs, and the relationships between species are broadly concordant with cynipid tribes (figure 4.1). Although, the relationship between tribes is contradicted for the polysaccharide lyase family 4 phylogeny (figure 4.4), as the Pediaspini are more closely related to the Diplolepidini sequences than to those of the Cynipini.

The prokaryote sequences cynipid PCWDEs are most similar to are derived from phylogenetically disparate bacteria in different classes. Separate horizontal transfer events from different donors could explain this. A hypothesis that can be answered by more in depth phylogenetics, in which cynipid PCWDE enzyme genes are placed into a broader bacterial context. However, there is very little positive evidence for these genes belonging to an unidentified prokaryote symbiont of the gall wasps. Only the PCWDE genes are present in the assemblies/transcriptome and not the full complement of genes one would expect if a symbiont were present. Highly conserved genes, such as ribosomal RNAs should be detectable if a symbiont is present. Additionally, the 5’ UTRs of genomic contigs corresponding to PCWDE genes expressed in the *B. pallida* larval transcriptome do not contain Shine-Dalgarno sequences expected of bacterial mRNAs.

Furthermore, several expressed PCWDE genes contain potential introns, and one has a confirmed intron. The presence of which is positive evidence that these genes are eukaryotic, and therefore encoded in cynipid nuclear genomes. The intron containing contig (*Contig_69918-*) is present in the cellulase phylogeny (figure 4.2) and is paired with a *B. gibbera* contig (*Contig_122111*). However, more introns confirmed in PCWDE genes are required to conclude that PCWDEs are a class of nuclear encoded genes; this is better explored with superior genome assemblies (see Chapter 5). Finding introns in horizontally transferred genes is concordant with the process of intron insertion observed for horizontally transferred PCWDEs of nematodes and a fungal-derived carotenoid in aphids (Blaxter 2007; Moran & Jarvik 2010; Mayer et al., 2011). Except in one case where a *N. vitripennis* transposase is present on the same contig as a PCWDE gene, the genomic contigs are too short to contain exons of genes of unambiguously hymenopteran origin up- or downstream of PCWDEs. Further, positive evidence for being encoded in the cynipid genomes are the *InterProScan* predicted eukaryotic signal peptides of cellulase, pectin lyase and polysaccharide lyase, although the rhamnogalacturonate lyases do not have this domain (figures 4.7-10).

Finally, the codon usage analysis indicates a very strong eukaryotic codon bias for the PCWDE enzymes in comparison to those bacterial genomes containing genes with the greatest homology to cynipid PCWDEs. This differentiation between eukaryote and prokaryote codon usage is commonly observed (Gustaffson et al., 2004). In particular, arginine appears to have a strong effect on the PCA. The arginine codon ‘AGA’ clusters with eukaryote (Gustaffson et al., 2004) sequences, while an alternative arginine codon ‘CGC’ is close to *Dickeya* and *Cellvibrio* sequences.

4.4.2 Gene expression indicates successful co-option of PCWDE genes

Two cellulase genes are expressed highly throughout gall induction, but are not differentially expressed. Four pectin lyase genes are differentially expressed although at much lower overall levels than the cellulases. For the larval stage, further experiments are required before functions can be assigned to these enzymes. Two possible functions are feeding or general re-modelling of host cells. Both could be true for different genes in the oak gall wasps if sub-functionalization of the PCWDEs has occurred; for example, some cellulases re-model host cells during induction, while others digest cellulose in the larval gut. These genes do appear to have integrated into the biology of the cynipids, one of Blaxter's (2007) criteria for a successful horizontal gene transfer. The second criterion, longevity, is indicated by presence of PCWDEs in three cynipid tribes that last shared a common ancestor approximately 54 millions of years ago (Buffington et al., 2012). To confirm this requires phylogenetic analysis of the PCWDEs with wider sampling of the *Cynipidae* (Chapter 5, further work). Such an analysis would also indicate whether PCWDEs are shared across all cynipid tribes. A hypothesis explaining widespread PCWDE genes in the *Cynipidae* is that their acquisition occurred during the evolution of gall induction in the *Cynipidae*. The phylogenetic differentiation of PCWDE genes into Cynipini, Pediaspini and Diplolepidini genomes does support an ancient presence of these genes in the *Cynipidae* before the splitting of these lineages from their most recent common ancestor.

4.5 Appendix

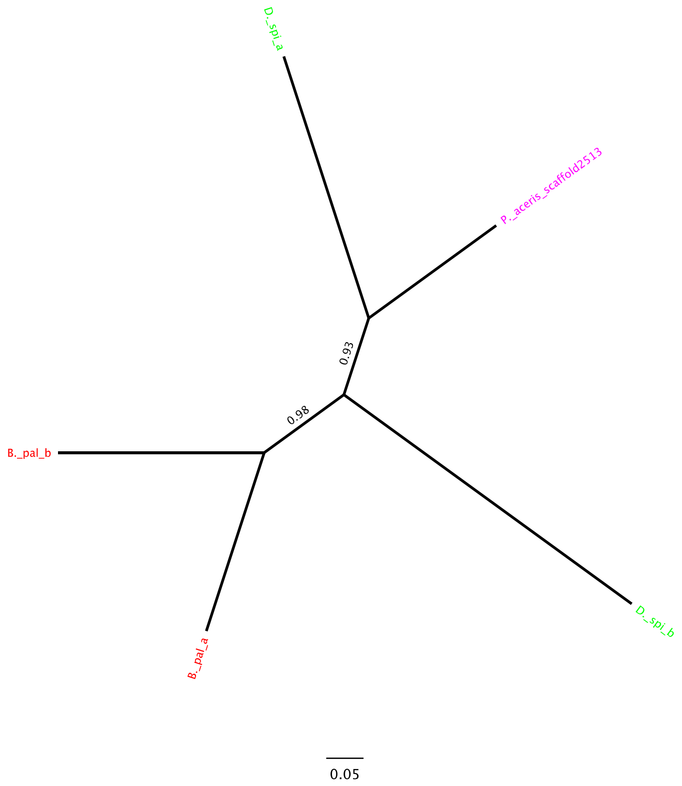

Figure 4.11. Phylogeny of rhamnogalacturonate lyase genes identified and passing alignment criteria *B. gibbera* = red, *B. pallida* = blue, *D. spinosa* = green, *A. quercuscalicis* = yellow, and *P. aceris* = fuchsia. Scale bar is substitutions per site and branches are labeled with approximate likelihood ratio test support values. B_pal_a = *B. pallida* contig_310874-, B_pal_b = *B. pallida* contig_270176-, D_spi_A = *D. spinosa* contig_58856 and, D_spi_B = *D. spinosa* contig_167541.

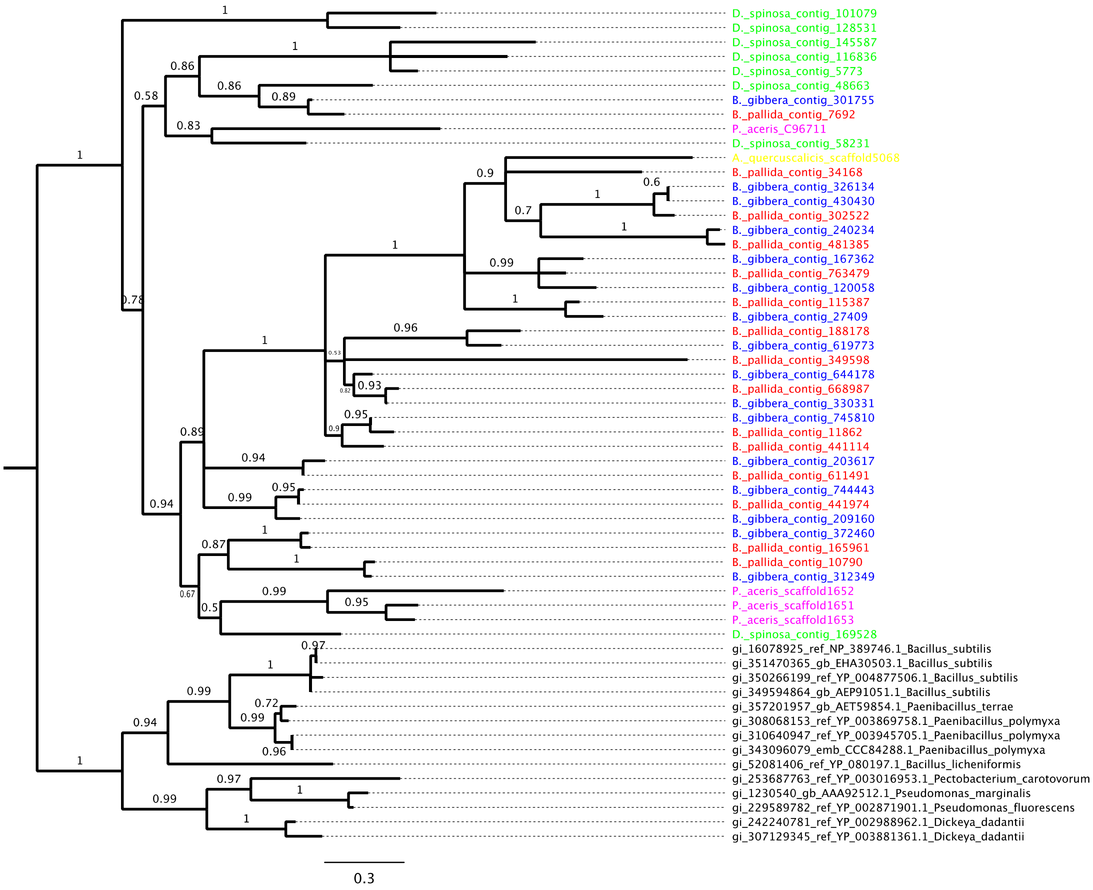

Figure 4.12. Phylogeny of pectinase genes including all possible outgroup sequences identified and passing alignment criteria *B. gibbera* = red, *B. pallida* = blue, *D. spinosa* = green, *A. quercuscalicis* = yellow, *P. aceris* = fuchsia and outgroup sequences = black. Scale bar is substitutions per site and branches are labeled with approximate likelihood ratio test support values.

Table 4.12. RSEM generated counts for each expressed *B. pallida* larval transcriptome PCWDE gene for each replicate. The two more highly expressed cellulases are in bold.

Table 4.13. *BLAST* genome best hits for larvally expressed *B. pallida* PCWDE genes in the *B. gibbera* assembly.

Transcript	Annotation	1	4	8	211	127	148	182	224	234	252	270C	281
comp103141_c0	pectin lyase	4	3	2	6	0	0	0	0	0	0	0	0
comp132692_c0	pectin lyase	2	1	1	4	0	1	0	0	0	0	0	0
comp13836_c0	**cellulase**	**207**	**60**	**70**	**613**	**462**	**439**	**1297**	**372**	**1650**	**302**	**562**	**2226**
comp14276_c0	pectin lyase	8	7	3	7	2	3	1	0	1	0	1	4
comp14276_c1	pectin lyase	8	1	2	6	0	0	1	0	1	0	0	1
comp18312_c0	pectin lyase	11	2	4	15	6	5	3	2	0	1	0	0
comp20769_c0	pectin lyase	29	17	25	21	2	2	10	5	626	388	82	54
comp20790_c0	pectin lyase	33	15	12	14	7	2	2	0	0	1	0	0
comp258830_c0	pectin lyase	0	1	0	2	0	0	0	0	0	0	0	0
comp27915_c0	**cellulase**	**1133**	**857**	**671**	**1263**	**222**	**198**	**913**	**245**	**2450**	**527**	**33**	**2288**
comp282070_c0	pectin lyase	2	0	1	0	0	0	0	0	0	0	0	0
comp326228_c0	cellulase	0	1	0	2	0	0	0	0	0	0	0	0
comp57254_c0	cellulase	2	3	2	6	2	0	6	1	28	5	5	13
comp59931_c0	pectin lyase	11	6	5	8	2	0	1	0	0	0	0	0
comp98838_c0	Rhamnogalacturonate lyase	3	8	5	0	2	3	2	0	1	0	0	0
comp100388_c0	pectin lyase	13	8	5	6	2	4	8	2	2	0	0	2
comp136735_c0	endopolygalacturonase (GH28Pect-5)	0	0	0	0	0	0	0	0	0	0	14	0
comp14239_c0	polysaccharide lyase family protein4	45	30	30	36	27	30	46	19	381	212	19	212
comp14831_c0	glycoside hydrolase family protein48	0	0	0	0	0	0	0	0	0	0	21	0
comp150589_c0	glycoside hydrolase family protein48	0	0	0	0	0	0	0	0	0	0	10	0
comp151243_c0	comp151243_c0	0	0	0	0	0	0	0	0	0	2	8	0
comp198252_c0	pectin lyase	1	1	0	2	1	0	1	0	0	0	0	0
comp213754_c0	endopolygalacturonase	0	0	0	0	0	0	0	0	0	0	7	0
comp24519_c0	pectin lyase	30	5	10	6	8	4	10	0	1	1	0	4
comp24519_c1	pectin lyase	6	2	6	4	7	2	7	0	1	1	0	1

comp258611_c0	comp151243_c0	0	0	0	0	0	0	0	0	0	0	3	0
comp26878_c0	polysaccharide lyase family protein4	51	38	32	37	20	18	22	7	1	3	2	9
comp26878_c1	polysaccharide lyase family protein4	10	8	8	11	0	7	0	1	1	0	0	0
comp27190_c0	pectin lyase	34	15	14	13	41	116	263	56	143	18	3	331
comp350819_c0	comp151243_c0	0	0	0	0	0	0	0	0	0	0	4	0
comp397147_c0	pectin methylesterase	0	0	0	0	0	0	0	0	0	1	2	0
comp53817_c0	pectin lyase	1	11	3	2	0	3	0	1	2	0	1	0
comp93909_c0	glycoside hydrolase family protein48	0	0	0	0	0	0	0	0	0	0	21	0

Transcript	Genome tophit	% identity	length	Misma -tches	Gaps	Query start	Query end	Ref. Start	Ref. End	E-value	Bit score	Transcript top hit
comp103141_c0	contig_455965	98.25	285	5	0	1	285	92	376	2.00E-137	491	pectin lyase
comp132692_c0	contig_644178	95.3	319	15	0	33	351	3	321	2.00E-142	508	pectin lyase
comp13836_c0	contig_118239	95.76	873	37	0	292	1164	1	873	0	1407	cellulase
comp14276_c0	contig_372460	96.61	560	19	0	1	560	565	1124	0	924	pectin lyase
comp14276_c1	contig_372460	94.76	496	25	1	1	496	93	587	0	774	pectin lyase
comp18312_c0	contig_312349	97.45	589	15	0	268	856	643	55	0	994	pectin lyase
comp20769_c0	contig_203617	96.77	433	14	0	340	772	434	2	0	719	pectin lyase
comp20790_c0	contig_330331	96.65	626	21	0	189	814	1	626	0	1034	pectin lyase
comp258830_c0	contig_330331	82.01	278	49	1	2	279	590	314	5.00E-72	273	pectin lyase
comp27915_c0	contig_131985	95.91	1051	43	0	204	1254	6	1056	0	1701	pectin lyase
comp282070_c0	contig_105368	99	200	2	0	2	201	216	17	4.00E-96	352	cellulase
comp326228_c0	contig_229402	96.23	239	9	0	1	239	473	711	2.00E-107	390	cellulase
comp57254_c0	contig_122111	96.82	942	29	1	88	1028	579	1520	0	1561	cellulase
comp59931_c0	contig_330331	75.6	209	51	0	1	209	359	567	8.00E-34	147	pectin lyase
comp98838_c0	contig_272997	91.67	540	19	2	23	560	1954	2469	0	787	Rhamnogalacturonate lyase
comp100388_c0	contig_272997	93.28	1638	90	7	1	1625	98	1728	0	2444	polysaccharide lyase family protein 4
comp14239_c0	contig_238739	96.08	1991	73	2	77	2067	61	2046	0	3234	polysaccharide lyase family protein 4
comp198252_c0	contig_229203	92.02	351	28	0	1	351	351	1	2.00E-142	508	polysaccharide lyase family protein 4
comp24519_c0	contig_36809	83.99	893	143	0	1	893	1033	1925	0	966	polysaccharide lyase family protein 4
comp24519_c1	contig_166410	95.73	234	10	0	401	634	1	234	4.00E-103	378	polysaccharide lyase family protein 4
comp26878_c0	contig_166411	96.52	804	28	0	179	982	1025	222	0	1324	polysaccharide lyase family protein 4
comp26878_c1	contig_36809	82.07	463	83	0	4	466	1015	1477	3.00E-128	461	polysaccharide lyase family protein 4
comp27190_c0	contig_184148	92.4	1302	53	1	507	1808	154	1409	0	1941	polysaccharide lyase family protein 4
comp53817_c0	contig_454089	93.83	243	15	0	182	424	271	29	5.00E-101	370	polysaccharide lyase family protein 4

Chapter 5: Planned and proposed future work based on the results of this thesis

In this short final chapter I propose future experiments to expand on the conclusions of this thesis. It is split into three sections concerning (5.1) genome wide multi-trophic phylogeography; (5.2) experiments to increase understanding of gall induction by cynipid gall wasps; and (5.3) horizontal gene transfer in cynipid genomes.

5.1 Genome wide multi-trophic phylogeography of gall wasps, fig wasps and their parasitoids

In chapter 2, I demonstrated that it is now possible to generate genome-level datasets for three individuals in non-model species to address population-level questions. The dataset was capable of disentangling population splitting from geneflow using maximum likelihood methods over Pleistocene timescales.

The bioinformatic and population genetic methods developed in this thesis, by Dr K. R. Lohse (University of Edinburgh) and I are being applied to a much larger project on two distinct hymenopteran gall systems, gall wasps and fig wasps. For this project I will build and help analyse the datasets. From each system several galler species and their parasitoid natural enemies will be sampled from different glacial refugia. For gall wasps and their parasitoids the Western Palaearctic will be sampled from the same eastern, central, and western refugia as for chapter 2. For fig wasps, which are also gall inducers, and their parasitoids, samples are to be collected in a series of separate latitudinal refugia of the Great Dividing Range, Australia. The results of this expanded analysis will be used to infer how species assemble temporally into the communities we observe.

There are several questions that will be addressed by this project. Do species that interact today show concordant phylogeographic histories, implying sustained interactions? For example, do the parasitoids of *B. pallida* share the East to West via North Africa migration event discovered in chapter 2? Or have communities been reshuffled by contrasting phylogeographic histories among component species? Do particular guilds or trophic levels experience contrasting levels of migration during divergence of populations in different glacial refugia? The answers to these questions are crucial for predicting the strength and direction of coevolution between gallers and their parasitoids. They have practical importance in ecology, as stable associations predict strong coevolution and high sensitivity of food webs to species gain/loss, while instability predicts diffuse coevolution and greater food web resilience (Memmott, 2009).

Illumina Hi-Seq sequencing of haploid males from a total of 20 species will provide the data for this project. In each system, groups of closely related species that can act as reciprocal out groups in analyses were selected. In the gall wasp system, five gall wasps - *Andricus coriarius, A. kollari*, *A. quercustozae, Cynips quercus, C. disticha* - and six of their most important parasitoids -*Megastigmus dorsalis, M. stigmatizans* (family Torymidae), *Eurytoma brunniventris, Sycophila biguttata* (family: Eurytomidae), *Mesopolobus amaenus* and *Mesopolobus tibialis* (family: Pteromalidae). For the fig wasps, fig-specific *Pleistodontes* pollinating fig wasps and six parasitoids (a pair of *Sycoscapter* and *Philotrypesis* species associated with each fig) have been selected. Because of the possibility of cryptic taxa and sample misidentification identifications for gall parasitoids and all fig-associates will be confirmed using multilocus DNA barcodes. Each individual will be sequenced to a depth of ~6x coverage. Two individuals per refugium will be sequenced to check robustness of inferences as was done for *B. pallida* in chapter 2. The methodology of chapter 2 will be adapted to assemble the datasets. The bioinformatics criterion remains the same: triplet alignments of ingroup individuals plus the outgroup sequence. However, new ways of assembling the dataset will be explored to (a) produce more robust ortholog groupings and (b) increase parallelization to cope with the much greater data throughput. Such as exploring alternatives to the *discontiguous* *megablast* (Altschul et al., 1997) reciprocal best hits (RBH) approach. Although RBH alternatives will need to work at the level of DNA as much of the datasets will consist of non-coding sequence. Automation of dataset generation stages through better optimization of scripting will aid parallelization of the bioinformatics.

5.2 Proposed experiments to confirm and test candidate gene involvement in gall induction

Before attempting further experiments validation of highly differentially expressed genes using reverse transcriptase quantitative PCR (rt-qPCR) in biological replicates will confirm the RNASeq results. Primers will be designed within exonic regions of transcripts tested and a set of control housekeeping genes for both plant and gall wasp genes. Rt-qPCR will be performed on four biological replicates to mirror the RNAseq experiment. RNAseq and rt-qPCR log fold changes can be plotted (with error bars) against one another and a correlation coefficient determined. A high correlation value and low log fold change values for housekeeping genes will validate the results. This analysis can be broadened to other gall wasp species for validated genes to test differential expression of candidate genes across the *Cynipidae*.

The RNAseq experiment also lacks controls in the form of plant and insect tissues at equivalent stages of development. Controls would have helped identify gall-specific expression and resulted in less (currently unknowable) errors in specifying new hypotheses. Although such experiments were planned a lack of resources prevented sequencing of figitid larvae for comparison of the gall wasp transcriptome and plant tissue to gall tissues. This will be partially rectified for plant tissue by our collaborators experiment discussed below. A figitid comparison would confirm if the chitinase gene expression pattern is unique to gall wasps. If figitids do have similar chitinase expression patterns an alternative explanation is breakdown of a chitin containing egg by larvae at hatching. Currently gall wasp eggs are not known to contain chitin, however the eggs of the mosquito *Aedes aegypti* do (Moreira et al., 2007). They identifiec chitinase activity by *A. aegypti* eggs and newly hatched larvae, although the chitinase itself was not identified (Moreira et al., 2007). Eggshell degrading chitinase activity would be consistent with the expression pattern and therefore needs to be addressed before further the experiments outlined below are attempted. In addition to an expression experiment on a non-galling outgroup, gall wasp eggs ovaries can be tested for chitin.

A similar RNA sequencing experiment to that of chapter 3 is underway by collaborators of the Stone laboratory at the Chinese Academy of Sciences research station at Xishuangbanna Tropical Botanical Gardens (XTBG), China. They are sequencing early and growth stage gall tissue and control leaf bud tissue sampled from *Dryocosmus cannoni* (tribe: Cynipini*)* on tropical chestnuts (*Castanopis spp.)*. The bud tissues will indicate normal developmental processes occurring in ungalled tissues. By comparing the results of their differential expression analysis with the results of chapter Chapter 3 orthologous candidates for gene expression can be identified. Addtionally, unique gall expression can be identified reducing the gene set on which to base further experiments. I hypothesise that gall wasp genes differentially expressed in the early stages in this experiment on *B. pallida* will be shared with *D. cannoni*. If this is the case previously unknown genes, with little functional annotations, dominate Cynipini tribe gall induction. Confirmation of differential gall wasp chitinase expression is also very important on the bases of the hypotheses specified in chapter 3.

The two sections below describe separate experiments that can indicate if the genes differentially expressed in early stage *B. pallida* are key actors in gall induction. Section 5.2.1 describes an *in situ* experiment and 5.2.2 an *in vivo* experiment.

5.2.1 RNA *in situ* hybridization of candidate genes and immunolocalisation of their proteins

Many *B. pallida* genes with expression patterns that potentially indicate a role in induction have no identifiable homologs. This makes inferences about their function impossible without further experiment. One route is to find where these proteins are expressed, and identify the sites at which they act, potentially in or on host cells. This is a test of gall wasp secreted proteins as a mechanism of manipulating the host during gall induction. If the hypothesis is true, the site of origin of these secreted proteins in the gall wasp is identifiable. There are two hypotheses for this: (1) the salivary glands or (2) the Malpighian tubules, found in the digestive tract that have endoreduplicated secretory cells in cynipid larvae (Harper et al., 2009).

An experiment, to discern the origin of secretory proteins is possible with a combination of RNA *in situ* hybridization and immunodetection of proteins. Fluorescence *in situ* hybridization has already been performed on cynipid gall tissues to identify BCCP in nutritive cells surrounding the larval chamber (Harper et al., 2004). However, unlike Harper et al. (2004) staining would be of larval RNA and not host DNA. RNA *in situ* hybridization can show where candidate genes are expressed in the larvae using confocal microscopy of cross sections of gall tissues as in Harper et al. (2004). Subsequently, immunodetection of candidate proteins using antibodies raised against them can then show if they act directly on the host after secretion from the larvae. This technique has been applied successfully in root knot nematodes to identify nematode proteins aggregating in the apoplasm (the diffusional space outside the plasma membrane of plant cells) and host nuclei (Vieira et al., 2010; Jaouannet et al., 2012). Good candidate genes for this experiment are the differentially expressed gall wasp chitinases. If they do interact with host cell wall-bound arabinogalactan proteins as discussed in chapter 3, chitinases are expected to localize to host cell walls or extracellular space (Poon et al., 2012). Antibody labelling of the candidate arabinogalactan protein (AGP) makes for an even more sensitive test of interaction. In this case, labelled chitinases could be observed interacting with labelled AGPs. Confirming an interaction between gall wasp chitinases and AGPs is a good pre-condition for the *in vitro* experiment described below. For those highly differentially expressed genes without functional annotations, this experiment will also show where they may act on host tissues.

5.2.2 An *in vitro* experiment to test the effect of AGP and chitinase on somatic embryogenesis in *Quercus robur*

In chapter 3, I proposed that Gall wasp chitinases modify host arabinogalactan proteins (AGP) resulting in somatic embryogenesis-like dedifferentiation and cell division; this dedifferentiation of host cells is a key step in successful development of an early gall. Poon et al., (2012) demonstrated that a cotton arabinogalactan protein (AGP) with a phytocyanin domain, similar to the *Q. robur* AGP early nodulation factor identified in chapter 3, promotes somatic embryogenesis. Chitinases also increase somatic embryogenesis in plant tissue by cleaving AGP (van Hengel et al., 2001; van Hengel et al., 2002). The high expression of an arabinogalactan in oak tissue and chitinase by gall wasp larvae in early stage galls led to the hypothesis that they interact to promote somatic embryogenesis-like processes in gall tissues. Using the bioassays developed by Poon et al. (2012) as a basis, I propose an experiment to test this.

Q. robur hypocotyl explants (tissues isolated from the stem of a germinating seedling and) (Cuenca et al., 1999) can be grown on various culture medias and the rates of embryogenic calli (parenchyma arising from cultured explants) development compared. The effect of different culture medias on the rate of somatic embryogenesis can be tested. A possible set of culture medias are AGPs extracted from early gall tissue, AGPs extracted from late gall tissue, developing buds, embryogenic calli, and non-embryogenic tissue. Replicates of each comparison with AGP pre-treated with gall wasp chitinases and subsequently re-precipitated, would assess the effect of gall wasp chitinases on inducing somatic embryogenesis-like processes (van Hengel et al., 2001). The hypothesis predicts that growth on media containing embryonic AGPs (isolated from early gall, developing bud or embryogenic calli) will be significantly enhanced relative to controls and non-embryogenic tissue. This effect should be further enhanced by treatment with early stage cynipid larval chitinases.

Each bioassay would need to be run enough times (n = 10, Poon et al., 2012) to allow adequate statistical power for the null hypotheses to be rejected. The results are analysable using odds ratios: the odds of an explant line developing embryogenic calli on each tested medium over the odds of explant line development on control medium (Poon et al., 2012).

Further *in vitro* experiments could involve RNA interference (RNAi) expression suppression of gall wasp candidate mRNAs. This however is technically challenging in the gall system. Rose gallers of *Diplolepis* are good candidates as galls have been induced in controlled conditions on roses (Harper et al., 2009). RNAi transformed rose callus tissue would act as a substrate for induction. Comparing the rate of induction between RNAi transformed and untransformed calli would indicate if the candidate gene is essential to gall induction.

5.3 Further analysis of plant cell wall degrading enzymes (PCWDEs) and their origin

To confirm the presence of PCWDE genes in the gall wasp genomes requires very long contiguous sequences (>20 000bp) to identify exons of hymenopteran origin, and rule out other possible origins of PCWDE genes. This can be achieved by deeper shotgun sequencing of cynipid genomes and better genome assemblies, as is occurring for the cynipids chosen for multi-trophic genome wide phylogeography (see section 5.1). Alternatively, a genomic library, bacterial artificial chromosome (BAC), of gall wasp genomes approach can be used. Clones containing PCWDE genes can be identified by library screening and sequenced by high throughput technology. For a bacterial artificial chromosome (BAC) the clone insert size is up to 350 kilobases. PCWDE gene containing BACS can be sequenced, assembled and any hymenopteran genes present identified. This approach would also confirm if a cryptic prokaryote was responsible for the PCWDEs in gall wasp genomes and transcriptomes. If PCWDE genes are as present within the gall was genome, and hence horizontally transferred into it, the BAC sequencing approach can be replicated in another gall wasp species. This tests for synteny among orthologous PCWDEs. Synteny predicts that orthologous genes are found up- and downstream of PCWDE orthologs in the genomes of the compared species.

Advances in sequencing and related technologies may also aid confirmation by presence of unambiguous hymenopteran genes and PCWDE genes on contiguous sequence. Pacific Biosciences technology can produce reads up to 10 kilobases (kb) with high error rates, or better quality but shorter reads of approximately 2kb (Shin et al., 2013). With further advances in length and quality of this technology, identifying multiple exons along a single read without the need for assembly should be possible. Optical mapping is another alternative for confirmation of PCWDE genes in gall wasp genomes. For optical mapping of a genome, a single molecule of DNA is stretched onto a slide and digested with restriction enzymes (Dimalanta et al., 2004). Each piece of DNA along the slide is fluorescently labelled and its size determined using the intensity of fluorescence (Dimalanta et al., 2004). Repeating this across thousands of molecules, allows a genome wide consensus optical map to be created. Contigs can be mapped to this consensus optical map; PCWDE containing contigs that do so are confirmed as present in the genome.

Further cynipid transcriptomes are being sequenced as part of the 1KITE project (http://www.1kite.org/) to sequence a thousand insect transcriptomes, as is a partial replication of the RNA sequencing experiment (chapter 3) in *Dryocosmus cannoni* (tribe: *Cynipini)* on tropical chestnuts (*Castanopis spp*). Additionally the deeper sequencing of cynipid genomes (section 5.1) will result in better assemblies, and therefore PCWDE gene-containing contigs of greater average length. This will demonstrate if the *Cynipini* do have higher copy number of particular PCWDEs than found in other gall wasp tribes. It will also provide a greater number of sequences for deeper phylogenetic analyses of gall wasp PCWDEs. Although, sampling of gall wasp genomes from the herb galling Aylacini tribe is needed for better representation of cynipid diversity.

Phylogenetic analysis of a broader sampling of PCWDE genes can confirm if PCWDEs in gall wasp genomes are ancient, for the second criterion of Blaxter's (2007) requirements for identifying a successful horizontal gene transfer. Hypotheses about the evolution of PCWDEs within the *Cynipidae* can be tested. For example, have the PCWDEs evolved into gene families? Have genes evolved for specific roles in gall wasps? More specifically, are larvally expressed cellulases of *B. pallida* and *D. cannoni* (assuming cellulases are present) orthologs? If true, larval cellulases should cluster together across species in a phylogeny and adult (venom gland) cellulases will form distinct clusters. Furthermore, if the new Cynipini genome assemblies also have lots of PCWDEs and large genome assembly sizes genomic duplication(s) in the Cynipini should be investigated. To test this, the number of genes in gene families of hymenopteran origin should be compared to the equivalent number in the *D. spinosa* rose gall wasp assembly.

Finally, several expressed PCWDEs were identified in chapter 4 (table 4.7) that were not identified in either the *B. pallida* or *B. gibbera* genome assemblies. Several of these PCWDEs are distinct from those confirmed as present in the *B. pallida* assembly and other genetic resources. They encode glycoside hydrolase family 48 proteins, endopolygalacturonase, and pectin methylesterase. One explanation is that the two assemblies are incomplete, however the same transcripts are missing in both species. This is explainable if the genes encoding these transcripts are located in difficult to sequence regions for Illumina technology. An alternative hypothesis is that these transcripts derive from other inhabitants of the gall, which are most commonly parasitoids or cynipid inquilines. These candidates for expressing cryptic PCWDEs may also have evolved, or acquired by horizontal gene transfer, plant cell wall degrading enzymes, as the true extent of HGT in the Arthropoda remains unknown. This is testable by searching the parasitoid genomes sequenced in section 5.1 above for PCWDE genes. For the inquilines, a tribe Synergini species, *Synergus umbraculus*, is also being genome sequenced (G. Stone, personal communication). The PCWDE complement of a cynipid inquiline can therefore be compared to gall inducers for losses or gains of particular PCWDEs.

Bibliography

Abad, P., Gouzy, J., Aury, J.-M., Castagnone-Sereno, P., Danchin, E. G. J., Deleury, E., … Blok, V. C. (2008). Genome sequence of the metazoan plant-parasitic nematode *Meloidogyne incognita*. *Nature biotechnology*, *26*(8), 909–15. doi:10.1038/nbt.1482

Abe, Y., Melika, G., & Stone, G. N. (2007). The diversity and phylogeography of cynipid gallwasps (Hymenoptera : Cynipidae) of the oriental and eastern Palearctic regions, and their associated communities. *Oriental Insects*, *41*, 169–212. Retrieved from http://elibrary.ru/item.asp?id=14252867

Acuña, R., Padilla, B. E., Flórez-Ramos, C. P., Rubio, J. D., Herrera, J. C., Benavides, P., … Rose, J. K. C. (2012). Adaptive horizontal transfer of a bacterial gene to an invasive insect pest of coffee. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(11), 4197–202. doi:10.1073/pnas.1121190109

Altenhoff, A. M., & Dessimoz, C. (2009). Phylogenetic and functional assessment of orthologs inference projects and methods. *PLoS computational biology*, *5*(1), e1000262. doi:10.1371/journal.pcbi.1000262

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*. *215*, 403-410.

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research*, *25*(17), 3389-3402.

Anders, S., & Huber, W. (2010). Differential expression analysis for sequence count data. *Genome biology*, *11*(10), R106. doi:10.1186/gb-2010-11-10-r106

Anisimova, M., & Gascuel, O. (2006). Approximate likelihood-ratio test for branches: A fast, accurate, and powerful alternative. *Systematic biology*, *55*(4), 539-552.

Arakane, Y., & Muthukrishnan, S. (2010). Insect chitinase and chitinase-like proteins. *Cellular and molecular life sciences*, *67*(2), 201-216.

Armbruster, P., Bradshaw, W. E., & Holzapfel, C. M. (1998). Effects of postglacial range expansion on allozyme and quantitative genetic variation of the pitcher-plant mosquito, *Wyeomyia smithii*. *Evolution*, 1697-1704.

Arnold, B., Corbett-Detig, R. B., Hartl, D., & Bomblies, K. (2013). RADseq underestimates diversity and introduces genealogical biases due to nonrandom haplotype sampling. *Molecular Ecology*, *22*(11), 3179–3190. doi:10.1111/mec.12276

Askew, R. R. (1984). The biology of gall wasps. *Biology of gall insects (TN Anantakrishnan, ed.). Edward Arnold, London*, 223-271.

Atkinson, R. J., Brown, G. S., & Stone, G. N. (2003). Skewed sex ratios and multiple founding in galls of the oak apple gall wasp *Biorhiza pallida*. *Ecological Entomology*, *28*(1), 14–24.

Avise, J. (1987). Intraspecific phylogeography: the mitochondrial DNA bridge between population genetics and systematics. *Annual Review of Ecology and Systematics*, *18*, 489–522.

Bailey, R., Schönrogge, K., Cook, J. M., Melika, G., Csóka, G., Thúroczy, C., & Stone, G. N. (2009). Host niches and defensive extended phenotypes structure parasitoid wasp communities. *PLoS Biology*, *7*(8), e1000179.

Barton, M. K. (2010). Twenty years on: the inner workings of the shoot apical meristem, a developmental dynamo. *Developmental biology*, *341*(1), 95–113. doi:10.1016/j.ydbio.2009.11.029

Barton, N. H., Kelleher, J. H., & Etheridge, A. M. (2010). A new model for extinction and recolonisation in two dimensions: Quantifying phylogeography. *Evolution*, *64*(9), 2701–2715.

Barton, N.H., Etheridge, A.M., Kelleher, J. & Véber, A. (2013). Inference for the spatial lambda-Fleming-Viot process. *Theoretical Population Biology, page in press*.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 289-300.

Beijerinck, M.W. (1882) *Beobachtungen u□ber die ersten Entwicklungsphasen einger Cynipidengallen.* Mu□ller, Amsterdam. Pp. 198-250.

Bézier, A., Annaheim, M., Herbinière, J., Wetterwald, C., Gyapay, G., Bernard-Samain, S., … Drezen, J.-M. (2009). Polydnaviruses of Braconid Wasps Derive from an Ancestral Nudivirus. *Science*, *323* (5916), 926–930. doi:10.1126/science.1166788

Bigot, Y., Samain, S., Augé-Gouillou, C., & Federici, B. a. (2008). Molecular evidence for the evolution of ichnoviruses from ascoviruses by symbiogenesis. *BMC evolutionary biology*, *8*, 253. doi:10.1186/1471-2148-8-253

Birney, E., Clamp, M., & Durbin, R. (2004). GeneWise and genomewise. *Genome research*, *14*(5), 988-995.

Bhojwani, S. S., & Dantu, P. K. (2013). Plant Tissue Culture: An Introductory Text, 75–92. doi:10.1007/978-81-322-1026-9

Blaxter, M. (2007). Symbiont genes in host genomes: fragments with a future? *Cell host & microbe*, *2*(4), 211–3. doi:10.1016/j.chom.2007.09.008

Breuer, C., Ishida, T., & Sugimoto, K. (2010). Developmental control of endocycles and cell growth in plants. *Current opinion in plant biology*, *13*(6), 654–60. doi:10.1016/j.pbi.2010.10.006

Bronner, R. & Plantefol, L. (1973) Propriétés lytiques des oeufs de Biorhiza pallida Ol. *Comptes Rendus de L'Academie des Sciences Paris, Série D*. 276, 189-192.

Buffington, M. L., Brady, S. G., Morita, S. I., & Van Noort, S. (2012). Divergence estimates and early evolutionary history of Figitidae (Hymenoptera: Cynipoidea). *Systematic Entomology*, *37*(2), 287-304.

Burge, C., & Karlin, S. (1997). Prediction of complete gene structures in human genomic DNA. *Journal of molecular biology*, *268*(1), 78-94.

Calderón-Cortés, N., Quesada, M., Watanabe, H., Cano-Camacho, H., & Oyama, K. (2012). Endogenous Plant Cell Wall Digestion: A Key Mechanism in Insect Evolution. *Annual Review of Ecology, Evolution, and Systematics*, *43*(1), 45–71. doi:10.1146/annurev-ecolsys-110411-160312

Carstens, B. C., Stoute, H. N., & Reid, N. M. (2009). An information-theoretical approach to phylogeography. *Molecular Ecology*, *18*(20), 4270–4282.

Cassab, G. I. (1986). Arabinogalactan proteins during the development of soybean root nodules. *Planta*, *168*(4), 441-446.

Challis, R. J., Mutun, S., Nieves-Aldrey, J.-L., Preuss, S., Rokas, A., Aebi, A., … Stone, G. N. (2007). Longitudinal range expansion and cryptic eastern species in the western Palaearctic oak gallwasp *Andricus coriarius*. *Molecular Ecology*, *16*(10), 2003–2014

Charlesworth, D. (2010). Don't forget the ancestral polymorphisms. *Heredity*, *105*(6), 509–510. Retrieved from http://dx.doi.org/10.1038/hdy.2010.14

Cock, P. J. A., Fields, C. J., Goto, N., Heuer, M. L., & Rice, P. M. (2010). The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Research* , *38* (6), 1767–1771. doi:10.1093/nar/gkp1137

Conesa, A., Götz, S., García-Gómez, J. M., Terol, J., Talón, M., & Robles, M. (2005). Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, *21*(18), 3674-3676.

Connord, C., Gurevitch, J., & Fady, B. (2012). Large-scale longitudinal gradients of genetic diversity: a meta-analysis across six phyla in the Mediterranean basin. *Ecology and Evolution*, *2*, 2600–2614.

Cornell, H. V. (1983). The Secondary Chemistry and Complex Morphology of Galls Formed by the Cynipinae (Hymenoptera): Why and How? *American Midland Naturalist*, *110*(2), 225–234. doi:10.2307/2425263

Csóka, G., Stone, G. N., & Melika, G. (2005). The biology, ecology and evolution of Gall-inducing Cynipidae. In C. Raman, W. Schaefer, & T. M. Withers (Eds.), *Biology, ecology and evolution of gall inducing insects* (pp. 573–642). Enfield, New Hampshire: Science Publisher.

Cuenca, B., San-José, M. C., Martínez, M. T., Ballester, A., & Vieitez, A. M. (1999). Somatic embryogenesis from stem and leaf explants of *Quercus robur* L. *Plant Cell Reports*, *18*(7-8), 538–543. doi:10.1007/s002990050618

Davey, J. W., & Blaxter, M. L. (2010). RADSeq: next-generation population genetics. *Briefings in Functional Genomics*, *9*(5-6), 416-423.

Davis, P. H., & Hedge, I. C. (1971). Floristic links between NW Africa and SW Asia. *Annales Naturhistorisches Museum Wien*, *75*(1), 43–57.

Davison, A., & Blaxter, M. (2005). Ancient Origin of Glycosyl Hydrolase Family 9 Cellulase Genes. *Mol Biol Evol*, *22*(5), 1273–1284. doi:10.1093/molbev/msi107

Dawkins, R. (1983). *The extended phenotype: The long reach of the gene*. Oxford; New York: Oxford University Press.

De Bruijn, N. G. (1946). A Combinatorial Problem. Koninklijke Nederlandse Akademie v. *Wetenschappen* 49: 758–764.

De Jong, a J., Cordewener, J., Lo Schiavo, F., Terzi, M., Vandekerckhove, J., Van Kammen, a, & De Vries, S. C. (1992). A carrot somatic embryo mutant is rescued by chitinase. *The Plant cell*, *4*(4), 425–33. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=160142&tool=pmcentrez&r

endertype=abstract

De Jong, A. J., Heidstra, R., Spaink, H. P., Hartog, M. V., Meijer, E. A., Hendriks, T., ... & De Vries, S. C. (1993). Rhizobium lipooligosaccharides rescue a carrot somatic embryo mutant. *The Plant Cell Online*, *5*(6), 615-620.

Dereeper, A., Guignon, V., Blanc, G., Audic, S., Buffet, S., Chevenet, F., ... & Gascuel, O. (2008). Phylogeny. fr: robust phylogenetic analysis for the non-specialist. *Nucleic acids research*, *36*(suppl 2), W465-W469.

Dimalanta, E. T., Lim, A., Runnheim, R., Lamers, C., Churas, C., Forrest, D. K., ... & Schwartz, D. C. (2004). A microfluidic system for large DNA molecule arrays. *Analytical chemistry*, *76*(18), 5293-5301.

Dillies, M. A., Rau, A., Aubert, J., Hennequet-Antier, C., Jeanmougin, M., Servant, N., ... & Jaffrézic, F. (2012). A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Briefings in Bioinformatics*.

Dogan, R. I., Getoor, L., Wilbur, W. J., & Mount, S. M. (2007). SplicePort—an interactive splice-site analysis tool. *Nucleic acids research*, *35*(suppl 2), W285-W291.

Hotopp, J. C. D., Clark, M. E., Oliveira, D. C., Foster, J. M., Fischer, P., Torres, M. C. M., ... & Werren, J. H. (2007). Widespread lateral gene transfer from intracellular bacteria to multicellular eukaryotes. *Science*, *317*(5845), 1753-1756.

Dunning Hotopp, J. C. (2011). Horizontal gene transfer between bacteria and animals. *Trends in genetics : TIG*, *27*(4), 157–63. doi:10.1016/j.tig.2011.01.005

Durand, E. Y., Patterson, N., Reich, D., Slatkin, M., & M., S. (2011). Testing for ancient admixture between closely related populations. *Mol Biol Ecol*, *28*(8), 2239–2252.

Edgar, R. C. (2004). *MUSCLE*: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, *32*(5), 1792–1797.

EFSA Panel on Plant Health (PLH), 2010. Risk assessment of the oriental chestnut gall wasp, *Dryocosmus kuriphilus* for the EU territory and identification and evaluation of risk management options. *EFSA Journa*l. *8* (6): 1619. doi: 10.2903/j.efsa.2010.1619

Egan, S. P., & Ott, J. R. (2007). Host plant quality and local adaptation determine the distribution of a gall-forming herbivore. *Ecology*, *88*(11), 2868-2879.

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., … Bettman, B. (2009). Real-Time DNA Sequencing from Single Polymerase Molecules. *Science* , *323* (5910), 133–138. doi:10.1126/science.1162986

Elborough, K., Winz, R., Deka, R., Markham, J., White, A., Rawsthorne, S., & Slabas, A. (1996). Biotin carboxyl carrier protein and carboxyltransferase subunits of the multi-subunit form of acetyl-CoA carboxylase from Brassica napus: cloning and analysis of expression during oilseed rape embryogenesis. *Biochem. J*, *315*, 103-112.

Emerson, K. J., Merz, C. R., Catchen, J. M., Hohenlohe, P. A., Cresko, W. A., Bradshaw, W. E., & Holzapfel, C. M. (2010). Resolving postglacial phylogeography using high-throughput sequencing. *Proceedings of the National Academy of Sciences*, *107*(37), 16196–16200. doi:10.1073/pnas.1006538107

Espagne, E., Dupuy, C., Huguet, E., Cattolico, L., Provost, B., Martins, N., … Drezen, J. M. (2004). Genome Sequence of a Polydnavirus: Insights into Symbiotic Virus Evolution.

Science, *306* (5694), 286–289. doi:10.1126/science.1103066

Field, L. M., James, A. A., Plantard, O., Rasplus, J. Y., Mondor, G., 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.

Finn, R. D., Clements, J., & Eddy, S. R. (2011). HMMER web server: interactive sequence similarity searching. *Nucleic acids research*, *39*(suppl 2), W29-W37.

Francis, W. R., Christianson, L. M., Kiko, R., Powers, M. L., Shaner, N. C., & Haddock, S. H. (2013). A comparison across non-model animals suggests an optimal sequencing depth for de novo transcriptome assembly. *BMC genomics*, *14*(1), 167.

François, O., Blum, M. G. B., Jakobsson, M., & Rosenberg, N. A. (2008). Demographic History of European Populations of *Arabidopsis thaliana*. *PLoS Genet*, *4*(5), e1000075. doi:10.1371/journal.pgen.1000075

Fu, Y. X., & Li, W. H. (1993). Statistical Tests of Neutrality of Mutations. *Genetics*, *133*(3), 693–709. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1208208&tool=pmcentrez&rendertype=abstract

Gardner, P. P., Daub, J., Tate, J. G., Nawrocki, E. P., Kolbe, D. L., Lindgreen, S., ... & Bateman, A. (2009). Rfam: updates to the RNA families database. *Nucleic acids research*, *37*(suppl 1), D136-D140.

Gillespie, J., & Langley, C. (1979). Are evolutionary rates really variable? *Journal of Molecular Evolution*, *13*(1), 27–34. doi:10.1007/BF01732751

Goecks, J., Mortimer, N. T., Mobley, J. a, Bowersock, G. J., Taylor, J., & Schlenke, T. a. (2013). Integrative approach reveals composition of endoparasitoid wasp venoms. *PloS one*, *8*(5), e64125. doi:10.1371/journal.pone.0064125

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., … Regev, A. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotech*, *29*(7), 644–652. Retrieved from http://dx.doi.org/10.1038/nbt.1883

Green, R. E., et. al, Krause, J., Briggs, A. W., Maricic, T., Stenzel, U., … Zhai, W. (2010). A draft sequence of the Neanderthal genome. *Science*, *328*(5979), 710–722. doi:10.1126/science.1188021

Griswold, C. K., & Baker, A. J. (2002). Time to the most recent common ancestor and divergence times of populations of common chaffinches *Fringilla coelebs* in Europe and North Africa: Insights into Pleistocene refugia and curent levels of migration. *Evolution*, *56*(1), 143–153.

Guindon, S., Delsuc, F., Dufayard, J. F., & Gascuel, O. (2009). Estimating maximum likelihood phylogenies with *PhyML*. In *Bioinformatics for DNA Sequence Analysis* (pp. 113-137). Humana Press.

Gustafsson, C., Govindarajan, S., & Minshull, J. (2004). Codon bias and heterologous protein expression. *Trends in Biotechnology*, *22*(7), 346–353. doi:http://dx.doi.org/10.1016/j.tibtech.2004.04.006

Habel, J. C., Meyer, M., El Mousadik, A., & Schmitt, T. (2008). Africa goes Europe: The complete phylogeography of the marbled white butterfly species complex *Melanargia* *galathea/M. lachesis* (Lepidoptera: Satyridae). *Organisms Diversity & Evolution*, *8*(2), 121–129.

Halligan, D. L., & Keightley, P. D. (2006). Ubiquitous selective constraints in the *Drosophila* genome revealed by a genome-wide interspecies comparison. *Genome Research*, *16*(7), 875–884.

Hansen, K. D., Brenner, S. E., & Dudoit, S. (2010). Biases in Illumina transcriptome sequencing caused by random hexamer priming. *Nucleic acids research*, *38*(12), e131. doi:10.1093/nar/gkq224

Harper, L., Schönrogge, K., & Lichtenstein, C. (2009). Chapter 4: Mechanisms of gall induction 1, *19*(c), 1–36. In Stone, Melika and Csóka (editors). *Oak gallwasps of the western palaearctic: ecology and evolution*. The Ray Society, London. In preparation.

Harper, L. J., Schönrogge, K., Lim, K. Y., Francis, P., & Lichtenstein, C. P. (2004). Cynipid galls : insect-induced modifications of plant development create novel plant organs, 327–335.

Harris, M. A., Deegan, J. I., Lomax, J., Ashburner, M., Tweedie, S., Carbon, S., ... & Huntley, R. (2008). The gene ontology project in 2008. *Nucleic Acids Res*, *36*, D440-D444.

Harris, K., & Nielsen, R. (2013). Inferring demographic history from a spectrum of shared haplotype lengths. *PLoS Genetics*, *in press*.

Hayward, A., & Stone, G. N. (2006). Comparative phylogeography across two trophic levels: the oak gall wasp *Andricus kollari* and its chalcid parasitoid *Megastigmus stigmatizans*. *Molecular Ecology*, *15*(2), 479–489.

Hewitt, G. M. (1999). Post-glacial re-colonization of European biota. *Biological Journal of the Linnean Society*, *68*(1-2), 87–112.

Hey, J. (2005). On the number of New World founders: a population genetics portrait of the peopling of the Americas. *PLoS Biology*, *3*(6), e193.

Hey, J. (2010). Isolation with migration models for more than two populations. *Molecular Biology and Evolution*, *27*, 905–920.

Hey, J., & Machado, C. A. (2003). The study of structured populations - new hope for a difficult and divided science. *Nature Reviews Genetics*, *4*(7), 535–543.

Hey, J., & Nielsen, R. (2004). Multilocus Methods for Estimating Population Sizes, Migration Rates and Divergence Time, With Applications to the Divergence *Drosophila pseudoobscura* and *D. persimilis*. *Genetics*, *167*(2), 747–760.

Hickerson, M. J., Carstens, B. C., Cavender-Bares, J., Crandall, K. A., Graham, C. H., Johnson, J. B., … Yoder, A. D. (2010). Phylogeography’s past, present, and future: 10 years after Avise 2000. *Molecular Phylogenetics and Evolution*, *54*(1), 291–301.

Huang, J., Tang, D., Shen, Y., Qin, B., Hong, L., You, A., … Cheng, Z. (2010). Activation of gibberellin 2-oxidase 6 decreases active gibberellin levels and creates a dominant semi-dwarf phenotype in rice (Oryza sativa L.). *Journal of genetics and genomics = Yi chuan xue bao*, *37*(1), 23–36. doi:10.1016/S1673-8527(09)60022-9

Hudson, R. R. (1983). Testing the constant-rate neutral allele model with protein sequence data. *Evolution*, *37*, 203–217.

Hundertmark, M., & Hincha, D. K. (2008). LEA (late embryogenesis abundant) proteins and their encoding genes in Arabidopsis thaliana. *BMC genomics*, *9*, 118. doi:10.1186/1471-2164-9-118

Ide, T., Wachi, N., & Abe, Y. (2010). Discovery of a new Plagiotrochus species (Hymenoptera: Cynipidae) inducing galls on the evergreen oak in Japan. *Annals of the Entomological Society of America*, *103*(6), 838-843.

Jansen, R. K., Saski, C., Lee, S. B., Hansen, A. K., & Daniell, H. (2011). Complete plastid genome sequences of three rosids (Castanea, Prunus, Theobroma): evidence for at least two independent transfers of rpl22 to the nucleus. *Molecular biology and evolution*, *28*(1), 835-847.

Jaouannet, M., Perfus-Barbeoch, L., Deleury, E., Magliano, M., Engler, G., Vieira, P., ... & Rosso, M. N. (2012). A root-knot nematode-secreted protein is injected into giant cells and targeted to the nuclei. *New Phytologist*, *194*(4), 924-931.

Jurka, J., Kapitonov, V. V., Kohany, O., & Jurka, M. V. (2007). Repetitive sequences in complex genomes: structure and evolution. *Annual Reviews of Genomics and Human Genetics.*, *8*, 241-259.

Kaldewey, H. (1965) Wachstumsregulatoren aus Pflanzengallen und Larven der Gallenbewohner. *Berichte der Deutschen Botanischen Gesellschaft. 78*, 73-84.

Karami, O., Aghavaisi, B., & Mahmoudi Pour, A. (2009). Molecular aspects of somatic-to-embryogenic transition in plants. *Journal of chemical biology*, *2*(4), 177–90. doi:10.1007/s12154-009-0028-4

Keeling, P. J. (2009). Functional and ecological impacts of horizontal gene transfer in eukaryotes. *Current opinion in genetics & development*, *19*(6), 613–9. doi:10.1016/j.gde.2009.10.001

Keeling, C. I., Yuen, M. M., Liao, N. Y., Docking, T. R., Chan, S. K., Taylor, G. a, … W Huber, D. P. (2013). Draft genome of the mountain pine beetle, Dendroctonus ponderosae Hopkins, a major forest pest. *Genome biology*, *14*(3), R27. doi:10.1186/gb-2013-14-3-r27

Kelly, C. K., & Southwood, T. R. E. (1999). Species richness and resource availability: A phylogenetic analysis of insects associated with trees. *Proc. Nat. Acad. Sci. USA*, *96*, 8013–8016.

Keightley, P. D., Trivedi, U., Thomson, M., Oliver, F., Kumar, S., & Blaxter, M. L. (2009). Analysis of the genome sequences of three Drosophila melanogaster spontaneous mutation accumulation lines. *Genome Research*, *19*(7), 1195–1201.

Kiełbasa, S. M., Wan, R., Sato, K., Horton, P., & Frith, M. C. (2011). Adaptive seeds tame genomic sequence comparison. *Genome research*, *21*(3), 487-493.

Kirsch, R., Wielsch, N., Vogel, H., Svatoš, A., Heckel, D. G., & Pauchet, Y. (2012). Combining proteomics and transcriptome sequencing to identify active plant-cell-wall-degrading enzymes in a leaf beetle. *BMC genomics*, *13*(1), 587. doi:10.1186/1471-2164-13-587

Knowles, L. L. (2002). Statistical phylogeography. *Annual Review of Ecology, Evolution, and*

Systematics, *11*(1), 593–612.

Kong, Q. P., Sun, C., Wang, H. W., Zhaio, M., Wang, W. Z., Zhong, L., … Zhang, Y. P. (2011). Large-scale mtDNA screening reveals a surprising matrilineal complexity in East Asia and its implications to the peopling of the region. *MBE*, *28*, 513–522.

Kragh, K. M., Hendriks, T., de Jong, a J., Lo Schiavo, F., Bucherna, N., Højrup, P., … de Vries, S. C. (1996). Characterization of chitinases able to rescue somatic embryos of the temperature-sensitive carrot variant ts 11. *Plant molecular biology*, *31*(3), 631–45. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/8790295

Kumar, S. 2013. *Next-generation nematode genomes*. University of Edinburgh.

Langmead, B., Trapnell, C., Pop, M., & Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*, *10*(3), R25.

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature methods*, *9*(4), 357-359.

Li, H., Wysoker A Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., & Durbin, R. (2009). The Sequence Alignment/Map SAM Format. *Bioinformatics*, 2078–2079.

Li, H., & Durbin, R. (2011). Inference of human population history from individual whole-genome sequences. *Nature*, *475*(7357), 493–496.

Liljeblad, J., & Ronquist, F. (1998). A phylogenetic analysis of higher-level gall wasp relationships (Hymenoptera: Cynipidae). *Systematic Entomology*, *23*(3), 229-252.

Lim, K., Furuta, Y., & Kobayashi, I. (2012). Large variations in bacterial ribosomal RNA genes. *Molecular biology and evolution*, *29*(10), 2937–48. doi:10.1093/molbev/mss101

Lima, J. (2012). *Species Richness and Genome Size Diversity in Hymenoptera with Different Developmental Strategies: A DNA Barcoding Enabled Study*. University of Guelph

Lohse, K., Sharanowski, B., & Stone, G. N. (2010). Quantifying the population history of the oak gall parasitoid *C. fungosa*. *Evolution*, *58*(4), 439–442.

Lohse, K., Harrison, R. J., & Barton, N. H. (2011). A general method for calculating likelihoods under the coalescent process. *Genetics*, *189*(3), 977–87. doi:10.1534/genetics.111.129569

Lohse, K., Barton, N. H., Melika, G., & Stone, G. N. (2012). A likelihood-based comparison of population histories in a parasitoid guild. *Molecular ecology*, *21*(18), 4605–17. doi:10.1111/j.1365-294X.2012.05700.x

Lohse, K., & Frantz, L. (2013). Maximum likelihood evidence for Neandertal admixture in Eurasian populations from three genomes. *Molecular Biology & Evolution*, *in review*.

Lowe, T. M., & Eddy, S. R. (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic acids research*, *25*(5), 0955-964.

Lumaret, R., Mir, C., Michaud, H., & Raynal, V. (2002). Phylogeographical variation of chloroplast DNA in holm oak *Quercus ilex*. *Molecular Ecology*, *11*(11), 2327–2336.

Lunter, G., & Goodson, M. (2011). Stampy: A statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Research*, *21*(6), 936–939. doi:10.1101/gr.111120.110

Klasson, L., Walker, T., Sebaihia, M., Sanders, M. J., Quail, M. A., Lord, A., … Parkhill, J. (2008). Genome Evolution of Wolbachia Strain wPip from the Culex pipiens Group. *Molecular Biology and Evolution* , *25* (9), 1877–1887. doi:10.1093/molbev/msn133

Klasson, L., Kambris, Z., Cook, P. E., Walker, T., & Sinkins, S. P. (2009). Horizontal gene transfer between Wolbachia and the mosquito Aedes aegypti. *Bmc Genomics*, *10*(1), 33.

Magoč, T., & Salzberg, S. L. (2011). FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics*, *27*(21), 2957-2963.

Martin, M. M., Jones, C. G., & Bernays, E. A. (1991). The Evolution of Cellulose Digestion in Insects [and Discussion]. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* , *333* (1267), 281–288. doi:10.1098/rstb.1991.0078

Matsui, S., & Torikata, H. (1970) Studies on the resistance of chestnut trees to chestnut gall wasps. III.Plant growth regulators contained in chestnut gall wasps and host gall tissue. *Journal of the Japanese Society of Horticultural Science. 39*, 115-123.

Matsui, S., Torikata, H. & Munakata, K. (1975) Studies on the resistance of chestnut trees to chestnut gall wasps. V. Cytokinin activity in leaves of gall wasps and callus formation of chestnut stem sections by larval extracts. *Journal of the Japanese Society of Horticultural Science*. *43*, 415-422.

Mayer, W. E., Schuster, L. N., Bartelmes, G., Dieterich, C., & Sommer, R. J. (2011). Horizontal gene transfer of microbial cellulases into nematode genomes is associated with functional assimilation and gene turnover. *BMC evolutionary biology*, *11*(1), 13. doi:10.1186/1471-2148-11-13

McCarthy, D. J., Chen, Y., & Smyth, G. K. (2012). Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic acids research*, *40*(10), 4288-4297.

McInerney, J. O. (1998). GCUA: general codon usage analysis. *Bioinformatics* , *14* (4), 372–373. doi:10.1093/bioinformatics/14.4.372

McCormack, J. E., Hird, S. M., Zellmer, A. J., Carstens, B. C., & Brumfield, R. T. (2013). Applications of next-generation sequencing to phylogeography and phylogenetics. *Molecular Phylogenetics and Evolution*, *66*, 526–538.

Melika, G., Tang, C. T., Nicholls, J. A., Yang, M. M., & Stone, G. N. (2011). Four New Species of Dryocosmus gallwasps from Taiwan (Hymenoptera: Cynipidae: Cynipini). *ISRN Zoology*, *2011*.

Melika, G. (2013). A new genus of oak gallwasp, Cyclocynips Melika, Tang & Sinclair (Hymenoptera: Cynipidae: Cynipini), with descriptions of two new species from Taiwan. *Zootaxa*, *3630*(3), 534-548.

Memmott, J. (2009). Food webs: a ladder for picking strawberries or a practical tool for practical problems? *Phil Trans Roy Soc*, *364*, 1693–1699.

Mitchum, M. G., Wang, X., Wang, J., & Davis, E. L. (2012). Role of nematode peptides and other small molecules in plant parasitism. *Annual review of phytopathology*, *50*, 175–95. doi:10.1146/annurev-phyto-081211-173008

Mitreva, M., Smant, G., & Helder, J. (2009). Role of horizontal gene transfer in the evolution of plant parasitism among nematodes. In *Horizontal Gene Transfer* (pp. 517-535). Humana Press.

Moran, N. a, & Jarvik, T. (2010). Lateral transfer of genes from fungi underlies carotenoid production in aphids. *Science (New York, N.Y.)*, *328*(5978), 624–7. doi:10.1126/science.1187113

Moreira, M. F., Dos Santos, A. S., Marotta, H. R., Mansur, J. F., Ramos, I. B., Machado, E. a, … Vasconcellos, A. M. H. (2007). A chitin-like component in *Aedes aegypti* eggshells, eggs and ovaries. *Insect biochemistry and molecular biology*, *37*(12), 1249–61. doi:10.1016/j.ibmb.2007.07.017

Mortimer, N. T., Goecks, J., Kacsoh, B. Z., Mobley, J. a, Bowersock, G. J., Taylor, J., & Schlenke, T. a. (2013). Parasitoid wasp venom SERCA regulates Drosophila calcium levels and inhibits cellular immunity. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(23), 9427–32. doi:10.1073/pnas.1222351110

Mower, J. P., Stefanovic, S., Young, G. J., & Palmer, J. D. (2004). Plant genetics: Gene transfer from parasitic to host plants. *Nature*, *432*(7014), 165–166. Retrieved from http://dx.doi.org/10.1038/432165b

Munoz-Torres, M. C., Reese, J. T., Childers, C. P., Bennett, A. K., Sundaram, J. P., Childs, K. L., … Elsik, C. G. (2011). Hymenoptera Genome Database: integrated community resources for insect species of the order Hymenoptera. *Nucleic acids research*, *39*(Database issue), D658–62. doi:10.1093/nar/gkq1145

Nicholls, J., Fuentes-Utrilla, P., Hayward, A., Melika, G., Csoka, G., Nieves-Aldrey, J.-L., … Stone, G. (2010). Community impacts of anthropogenic disturbance: natural enemies exploit multiple routes in pursuit of invading herbivore hosts. *BMC Evolutionary Biology*, *10*(1), 322.

Nicholls, J. A., Preuss, S., Hayward, A., Melika, G., Csóka, G., Nieves-Aldrey, J.-L., … Stone, G. N. (2010). Concordant phylogeography and cryptic speciation in two Western Palaearctic oak gall parasitoid species complexes. *Molecular Ecology*, *19*, 592–609.

Nichols, R. (2001). Gene trees and species trees are not the same. *Trends in Ecology & Evolution*, *16*(7), 358–364

Nielsen, R., Paul, J. S., Albrechtsen, A., & Song, Y. S. (2011). Genotype and SNP calling from next-generation sequencing data. *Nature reviews. Genetics*, *12*(6), 443–51. doi:10.1038/nrg2986

Nylander, J. A. (2004). *Bayesian Phylogenetics and the Evolution of Gall Wasps BY*. Upsaliensis, Acta Universitatis.

Ohkawa, M. (1974) Isolation of zeatin from larvae of *Dryocosmus kuriphilus* Yasumatsu. *Hortscience*. *9*, 458-459.

Parra, G., Bradnam, K., & Korf, I. (2007). CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics (Oxford, England)*, *23*(9), 1061–7. doi:10.1093/bioinformatics/btm071

Pauchet, Y., Wilkinson, P., Chauhan, R., & Ffrench-Constant, R. H. (2010). Diversity of beetle genes encoding novel plant cell wall degrading enzymes. *PloS one*, *5*(12), e15635. doi:10.1371/journal.pone.0015635

Pauchet, Y., & Heckel, D. G. (2013). The genome of the mustard leaf beetle encodes two active xylanases originally acquired from bacteria through horizontal gene transfer. *Proceedings. Biological sciences / The Royal Society*, *280*(1763), 20131021. doi:10.1098/rspb.2013.1021

Petit, R. J., Csaikl, U. M., Bordacs, S., & Burg K. Coart, E. et al. (2003). Chloroplast DNA variation in European white oaks phylogeography and patterns of diversity based on data from over 2600 populations. *Forest Ecology and Management*, *176*, 595–599.

Plantard, O., Rasplus, J. Y., Mondor, G., Le Clainche, I., & Solignac, M. (1998). Wolbachia–induced thelytoky in the rose gallwasp 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.

Poon, S., Heath, R. L., & Clarke, A. E. (2012). A chimeric arabinogalactan protein promotes somatic embryogenesis in cotton cell culture. *Plant physiology*, *160*(2), 684–95. doi:10.1104/pp.112.203075

Price, A. L., Jones, N. C., & Pevzner, P. A. (2005). De novo identification of repeat families in large genomes. *Bioinformatics*, *21*(suppl 1), i351–i358. doi:10.1093/bioinformatics/bti1018

Pujade-Villar, J., Romero-Rangel, S., Chagoyán-García, C., Equihua-Martínez, A., Estrada-Venegas, E. G., & Melika, G. (2010). A new genus of oak gallwasps, Kinseyella Pujade-Villar & Melika, with a description of a new species from Mexico (Hymenoptera: Cynipidae: Cynipini). *Zootaxa*, *2335*, 16-28.

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, *26*(6), 841–842. doi:10.1093/bioinformatics/btq033

Rehner, S. A., & Buckley, E. (2005). A Beauveria phylogeny inferred from nuclear ITS and EF1-α sequences: evidence for cryptic diversification and links to Cordyceps teleomorphs. *Mycologia* , *97* (1), 84–98. doi:10.3852/mycologia.97.1.84

Rey, L.A. (1992) Developmental morphology of two types of hymenopterous galls. Pp. 118-140 in: *Biology of insect-induced galls* (eds. J.D.Shorthouse & O. Rohfritsch). Oxford University Press, New York.

Rice, P., Longden, I., & Bleasby, A. (2000). EMBOSS: The European Molecular Biology Open Software Suite. *Trends in Genetics*, *16*(6), 276–277. doi:http://dx.doi.org/10.1016/S0168-9525(00)02024-2

Richardson, A. O., & Palmer, J. D. (2007). Horizontal gene transfer in plants. *Journal of experimental botany*, *58*(1), 1–9. doi:10.1093/jxb/erl148

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics (Oxford, England)*, *26*(1), 139–40. doi:10.1093/bioinformatics/btp616

Robinson, M. D., & Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol*, *11*(3), R25.

Rohfritsch, O. & Shorthouse, J.D. (1982) Insect galls. Pp. 131-152 in: Molecular biology of plant tumours (eds. O. Rohfritsch & J.D. Shorthouse). Academic Press. London.

Rokas, A., Atkinson, R., Brown, G., West, S. A., & Stone, G. N. (2001). Understanding patterns of genetic diversity in the oak gallwasp *Biorhiza pallida*: demographic history or a *Wolbachia* selective sweep? *Heredity*, *87*, 294–305.

Rokas, A., Atkinson, R. J., Webster, L., Csóka, G., & Stone, G. N. (2003). Out of Anatolia: longitudinal gradients in genetic diversity support an eastern origin for a circum-Mediterranean oak gallwasp *Andricus quercustozae*. *Molecular Ecology*, *12*(8), 2153–2174.

Ronquist, F., and J. Liljeblad. 2001. Evolution of the gall wasp-host plant association. *Evolution 55*, 2503-2522.

Rosenberg, N. a, & Nordborg, M. (2002). Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. *Nature reviews. Genetics*, *3*(5), 380–90. doi:10.1038/nrg795

Sakharkar, M. K., Kangueane, P., & Dmitri, A. (2002). SEGE: A database on 'intron less/single exonic' genes from eukaryotes. *Bioinformatics: Applications Note*. *18*(9), 1266–1267.

Sakharkar, M. K., & Kangueane, P. (2004). Genome SEGE: a database for "intronless" genes in eukaryotic genomes. *BMC bioinformatics*, *5*, 67. doi:10.1186/1471-2105-5-67

Salichos, L., & Rokas, A. (2011). Evaluating ortholog prediction algorithms in a yeast model clade. *PloS one*, *6*(4), e18755. doi:10.1371/journal.pone.0018755

Samson, R., Legendre, J. B., Christen, R., Fischer-Le Saux, M., Achouak, W., & Gardan, L. (2005). Transfer of *Pectobacterium chrysanthemi* (Burkholder et al. 1953) Brenner et al. 1973 and *Brenneria paradisiaca* to the genus *Dickeya* gen. nov. as *Dickeya chrysanthemi* comb. nov. and *Dickeya paradisiaca* comb. nov. and delineation of four novel species, Dick. *International journal of systematic and evolutionary microbiology*, *55*(Pt 4), 1415–27. doi:10.1099/ijs.0.02791-0

Schatz, M. C., Delcher, A. L., & Salzberg, S. L. (2010). Assembly of large genomes using second-generation sequencing. *Genome research*, *20*(9), 1165–73. doi:10.1101/gr.101360.109

Schönrogge, K., Harper, L. J., & Lichtenstein, C. P. (2000). The protein content of tissues in cynipid galls (Hymenoptera : Cynipidae): Similarities between cynipid, 215–222.

Schoof, H., Lenhard, M., Haecker, A., Mayer, K. F., Jürgens, G., & Laux, T. (2000). The Stem Cell Population of *Arabidopsis* Shoot Meristems Is Maintained by a Regulatory Loop between the *CLAVATA* and *WUSCHEL* Genes. *Cell*, *100*(6), 635-644.

Schultz, C., Gilson, P., Oxley, D., Youl, J., & Bacic, A. (1998). GPI-anchors on arabinogalactan-proteins: implications for signalling in plants. *Trends in Plant Science*, *3*(11), 426–431. doi:http://dx.doi.org/10.1016/S1360-1385(98)01328-4

Schulz, M. H., Zerbino, D. R., Vingron, M., & Birney, E. (2012). Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics* , *28* (8), 1086–1092. doi:10.1093/bioinformatics/bts094

Shine, J., & Dalgarno, L. (1975). Determinant of cistron specificity in bacterial ribosomes. *Nature*, *254*(5495), 34-38.

Singh, S., Cornilescu, C. C., Tyler, R. C., Cornilescu, G., Tonelli, M., & Lee, M. I. N. S. (2005). Solution structure of a late embryogenesis abundant protein (LEA14) from Arabidopsis thaliana , a cellular stress-related protein, 2601–2609. doi:10.1110/ps.051579205.different

Smit AFA, H. R. & G. P. (n.d.). RepeatMasker Open-3.0. 1996-2010.

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. doi:10.1016/j.baae.2005.07.006

Sommer, R. J., & Streit, A. (2011). Comparative genetics and genomics of nematodes: genome structure, development, and lifestyle. *Annual review of genetics*, *45*, 1–20. doi:10.1146/annurev-genet-110410-132417

Stone, G. N., & Sunnucks, P. (1993). Genetic consequences of an invasion through a patchy environment - the cynipid gallwasp *Andricus quercuscalicis* (Hymenoptera: Cynipidae). *Molecular Ecology*, *2*(4), 251–268.

Stone, G.N., K. Schönrogge, R.J. Atkinson, D. Bellido, and J. Pujade-Villar. 2002. The population biology of oak gall wasps (Hymenoptera: Cynipidae). Annual Review of Entomology, *47*:633–668.

Stone, G. N., & Schönrogge, K. (2003). The adaptive significance of insect gall morphology. *Trends in Ecology & Evolution*, *18*(10), 512-522.

Stone, G. N., Challis, R. J., Atkinson, R. J., Csóka, G., Hayward, A., Melika, G., … Schönrogge, K. (2007). The phylogeographical clade trade: tracing the impact of human-mediated dispersal on the colonization of northern Europe by the oak gallwasp *Andricus kollari*. *Molecular Ecology*, *16*, 2768–2781.

Stone, G. N., Hernandez-Lopez, A., Nicholls, J. A., di Pierro, E., Pujade-Villar, J., Melika, G., … Abbot, P. (2009). Extreme host plant conservatism during at least 20 million years of host plant pursuit by oak gallwasps. *Evolution*, *63*(4), 854–869.

Stone, G. N., Lohse, K., Nicholls, J. A., Fuentes-Utrilla, P., Sinclair, F., Schönrogge, K., … Hickerson, M. J. (2012). Reconstructing community assembly in time and space reveals enemy escape in a western palaearctic insect community. *Current Biology*, *22*(6), 531–537.

Strand, M. R., & Burke, G. R. (2012). Polydnaviruses as symbionts and gene delivery systems. *PLoS pathogens*, *8*(7), e1002757. doi:10.1371/journal.ppat.1002757

Strasburg, J. L., & Rieseberg, L. H. (2009). How robust are isolation with migration analysies to violations of the IM model? A simulation study. *Molecular Biology and Evolution*, *27*(2), 297–310.

Stuart, J. J., Chen, M.-S., Shukle, R., & Harris, M. O. (2012). Gall midges (Hessian flies) as plant pathogens. *Annual review of phytopathology*, *50*, 339–57. doi:10.1146/annurev-phyto-072910-095255

Sullivan, D. J., & Volkl, W. (1999). Hyperparasitism: Multitrophic ecology and behavior. *Annual Review of Entomology*, *44*, 291–315.

Syvanen, M. (2012). Evolutionary implications of horizontal gene transfer. *Annual review of genetics*, *46*, 341–58. doi:10.1146/annurev-genet-110711-155529

Tajima, F. (1983). Evolutionary relationships of DNA sequences in finite populations. *Genetics*, *105*(2), 437–460.

Tang, C. (2011). A new genus of oak gallwasps, *Cycloneuroterus* Melika & Tang, with the description of five new species from Taiwan (Hymenoptera: Cynipidae: Cynipini), *62*, 5326.

Tang, C. T., Sinclair, F., Yang, M. M., & Melika, G. (2012). A new *Andricus* Hartig oak gallwasp species from China (Hymenoptera: Cynipidae: Cynipini). *Journal of Asia-Pacific Entomology*, *15*(4), 601-605.

Treangen, T. J., & Salzberg, S. L. (2011). Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nature Reviews Genetics*, *13*(1), 36-46.

Ueno, S., Le Provost, G., Leger, V., Klopp, C., Noirot, C., Frigerio, J.-M., … Plomion, C. (2010). Bioinformatic analysis of ESTs collected by Sanger and pyrosequencing methods for a keystone forest tree species: oak. *BMC Genomics*, *11*(1), 650. doi:10.1186/1471-2164-11-650

Van Bakel, H., Stout, J., Cote, A., Tallon, C., Sharpe, A., Hughes, T., & Page, J. (2011). The draft genome and transcriptome of Cannabis sativa. *Genome Biology*, *12*(10), R102. doi:10.1186/gb-2011-12-10-r102

Van Hengel AJ, Guzzo, F., van Kammen A, & de Vries SC. (1998). Expression pattern of the carrot EP3 endochitinase genes in suspension cultures and in developing seeds. *Plant physiology*, *117*(1), 43–53. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=35020&tool=pmcentrez&rendertype=abstract

Van Hengel, a J., Tadesse, Z., Immerzeel, P., Schols, H., van Kammen, a, & de Vries, S. C. (2001). N-acetylglucosamine and glucosamine-containing arabinogalactan proteins control somatic embryogenesis. *Plant physiology*, *125*(4), 1880–90. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=88843&tool=pmcentrez&rendertype=abstract

Van Hengel, A. J., Van Kammen, A., & De Vries, S. C. (2002). A relationship between seed development, Arabinogalactan-proteins (AGPs) and the AGP mediated promotion of somatic embryogenesis. *Physiologia plantarum*, *114*(4), 637–644. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11975739

Vieira, P., Danchin, E. G., Neveu, C., Crozat, C., Jaubert, S., Hussey, R. S., ... & Rosso, M. N. (2011). The plant apoplasm is an important recipient compartment for nematode secreted proteins. *Journal of experimental botany*, *62*(3), 1241-1253.

Volkoff, A.-N., Jouan, V., Urbach, S., Samain, S., Bergoin, M., Wincker, P., … Drezen, J.-M. (2010). Analysis of virion structural components reveals vestiges of the ancestral ichnovirus genome. *PLoS pathogens*, *6*(5), e1000923. doi:10.1371/journal.ppat.1000923

Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, *10*(1), 57-63.

Wang, S., Lorenzen, M. D., Beeman, R. W., & Brown, S. J. (2008). Analysis of repetitive DNA distribution patterns in the Tribolium castaneum genome. *Genome biology*, *9*(3), R61. doi:10.1186/gb-2008-9-3-r61

Watanabe, H., Noda, H., Tokuda, G., & Lo, N. (1998). A cellulase gene of termite origin. *Nature*, *394*(6691), 330-331.

Whelan, S., & Goldman, N. (2001). A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Molecular biology and evolution*, *18*(5), 691-699.

Whitfield, J. B., & Asgari, S. (2003). Virus or not? Phylogenetics of polydnaviruses and their wasp carriers. *Journal of Insect Physiology*, *49*(5), 397-405.

Whittemore, a T., & Schaal, B. a. (1991). Interspecific gene flow in sympatric oaks. *Proceedings of the National Academy of Sciences of the United States of America*, *88*(6), 2540–4. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=51268&tool=pmcentrez&rendertype=abstract

Wolfram Research, I. (2010). *Mathematica, Version 8.0*. Champaign, Illinois: Wolfram Research, Inc.

Wu, M., Sun, L. V, Vamathevan, J., Riegler, M., Deboy, R., Brownlie, J. C., … Eisen, J. A. (2004). Phylogenomics of the Reproductive Parasite *Wolbachia pipientis wMel*: A Streamlined Genome Overrun by Mobile Genetic Elements. *PLoS Biol*, *2*(3), e69. Retrieved from http://dx.doi.org/10.1371%2Fjournal.pbio.0020069

Wu, T. D., & Nacu, S. (2010). Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics*, *26*(7), 873-881.

Zdobnov, E. M., & Apweiler, R. (2001). InterProScan–an integration platform for the signature-recognition methods in InterPro. *Bioinformatics*, *17*(9), 847-848.

Zerbino, D. R., & Birney, E. (2008). Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome research*, *18*(5), 821–9. doi:10.1101/gr.074492.107

Zhu, Q., Arakane, Y., Beeman, R. W., Kramer, K. J., & Muthukrishnan, S. (2008). Functional specialization among insect chitinase family genes revealed by RNA interference. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(18), 6650–5. doi:10.1073/pnas.0800739105

Xi, Z., Bradley, R. K., Wurdack, K. J., Wong, K., Sugumaran, M., Bomblies, K., … Davis, C. C. (2012). Horizontal transfer of expressed genes in a parasitic flowering plant. *BMC genomics*, *13*(1), 227. doi:10.1186/1471-2164-13-227

Xi, Z., Wang, Y., Bradley, R. K., Sugumaran, M., Marx, C. J., Rest, J. S., & Davis, C. C. (2013). Massive mitochondrial gene transfer in a parasitic flowering plant clade. *PLoS genetics*, *9*(2), e1003265. doi:10.1371/journal.pgen.1003265