Comparative transcriptome profiling in wild species uncovering gene expression signatures of mating systems

Ockendon, Nina

Award date:
2015

Awarding institution:
University of Bath

Comparative transcriptome profiling in wild species: uncovering gene expression signatures of mating systems

Nina Frances Ockendon

A thesis submitted for the degree of Doctor of Philosophy

University of Bath

Department of Biology and Biochemistry

November 2014

Supervisors

Dr. Araxi Urrutia, Dept. of Biology and Biochemistry, University of Bath, UK

Prof. Tamás Székely, Dept.of Biology and Biochemistry, University of Bath, UK

Declaration

This thesis is entirely the work of Nina Frances Ockendon with advice, guidance and support from her supervisors, except those instances stated below. It is understood that these instances have not been submitted for other degrees, except where stated.

- Chapter 2:
  - All bacterial culturing, phenotype microarrays, and RNA-seq data acquisition had been performed previously in the Waterfield research group, University of Bath. RNA-seq data quality control and alignments had previously been performed by Mr. Paul Wilkinson, University of Bath.
- Chapter 3:
  - Dr. Lauren O'Connell, Harvard University, performed all assembly construction;
  - Dr. Stephen Bush, University of Bath, performed all assembly homology searching;
  - Ms. Holly Barnes, University of Bath, performed all primate direct genome mapping.
- Chapter 4:
  - Dr. Ákos Pogány, Eötvös Loránd University, led the fieldwork trip, determining locations for searching for birds and capturing all individuals;
  - Dr. Gergely Zachar, Semmelweis University, sacrificed all songbirds used in this study and performed all brain and some other tissue dissections;
  - Dr. Alexander Ball, University of Bath, performed sperm morphology analysis (understood to have been submitted as part of his thesis, 2014) and some tissue dissections;
  - Songbird morphological measurements were performed by Dr.'s Ball, Pogány, and Zachar.

Acknowledgements

Firstly, I am tremendously grateful to my supervisors Araxi and Tamas for their unswerving support throughout my PhD: they have made this challenging process interesting, exciting, and inspiring. I will always value their opinions and advice – I have learned so much from them both. I also acknowledge the support from all of my lab mates and students in the Urrutia research group for the laughs, jokes and general support and advice.

Thank you to my family and friends for all their love and support – they know what achieving my PhD means to me. Particular thanks go to my Mother Mary and my Fiancé Jonny for your love and support – mentally, spiritually, and, on occasions, financially.

I would like to acknowledge my collaborators, particularly Prof. Hans Hofmann, without whom our interesting songbird project would not have materialised, and I would not have had the opportunity to spend time in his lab in Texas – a brilliant experience that I will always remember. Thanks also to Dr. Lauren O'Connell – an inspirational collaborator whose expertise and vision has spurred me in this very exciting field. Thanks to Prof. Andras Csillag for his kind hospitality while in Budapest – it was highly valued. Extreme thanks to Dr.'s Pogány and Zachar for allotting the time to be part of the fieldtrip – they completed a great team and it was a great trip. Thanks also to Prof. Nick Waterfield for the opportunity to work on a very interesting area that I had not expected to work on: having the opportunity to work across multiple 'omics data streams has been a tremendously useful experience. Thanks go to the BBSRC for the Systems Biology Studentship I was awarded to fund this work, and additionally to the Company of Biologists and the Korner Foundation for the Travelling Fellowships they respectively awarded me in support of the work undertaken herein.

Table of Contents

Table of figures and supplementary figures

Chapter 4:

Table of tables and supplementary tables

1 General Introduction

1.1 Abstract

Understanding the molecular processes underlying adaptation of complex phenotypes presents major challenges in evolutionary biology. An important question currently is how to accurately use the plethora of 'omics data to better understand ecological variation. Using RNA-seq transcriptome data from many lineages, I demonstrate the power of this data type when studying the molecular basis of complex phenotypes. My work has produced three major results. Firstly, I have integrated bacterial RNA-seq data with high throughput phenotype microarrays, providing the first indication of functional pathways implicated at genomic and phenotypic levels in trait evolution related to host switching and proliferation in *Photorhabdus* species. Secondly, since genome sequence data are currently unavailable for most species, I present an optimised methodology for RNA-seq transcriptome annotation for species with no sequenced genome. This shows that direct mapping of RNA-seq short reads to a reference genome – from the same species or a closely-related species – is the most effective, accurate and least functionally biased strategy for annotating transcriptomes compared to currently popular transcriptome assembly methods. Thirdly, I have contributed genomic resources to the scientific community by obtaining brain transcriptomes from two non-sequenced songbird species that represent interesting ecological models of mating behaviour. Applying my direct genome mapping annotation strategy to the novel data, I have described the transcriptomes via gene expression profiling and functional characterisation, amongst others methods. I have provided a first indication of genes differentially regulated during the breeding seasons of typically monogamous and polygamous songbirds. Overall, I have provided insight into the performance of state-of-the-art high throughput genomic and phenotypic analyses, identifying genes and functional pathways potentially important in the evolution and development of specific complex phenotypes across a variety of taxa. Thus, my work provides an excellent basis for further studies to disentangle how these phenotypes evolved and dissect the mechanisms by which they operate.

1.2 Exploring the genomic basis of complex trait evolution in non-model species

How have complex traits, such as animal social behaviour, evolved? Understanding this remains one of the biggest questions for biologists today. Combining neuroscience (encompassing neurobiology and neuroethology), evolutionary biology, and the study of developmental process, the field of "neuro-evo-devo" has made important advances in our appreciation of how morphological variation, molecular patterning, such as receptor expression in various brain regions, and genetic and epigenetic factors have contributed to the evolution of conserved brain circuitry that integrates fundamental aspects of animal social traits (O'Connell & Hofmann, 2011; Robinson et al., 2008; Toth & Robinson, 2007). With the advancement of genomic sequencing technology, and other 'omic technologies, in recent years data capture of epic proportions has been possible which is allowing researchers to begin describing the molecular genomic and cellular landscapes of the brain during various behavioural states. This builds upon existing knowledge of the genes that underlie particular behavioural traits, such as pair bonding (oxytocin and vasopressin receptor genes, *OTR* and *V1aR*; Gimpl & Fahrenholz, 2001; Ophir et al., 2012; Ross et al., 2009; Walum et al., 2008; Young et al., 1999; Young & Wang, 2004), aggression (monoamine oxidase A gene, *Maoa*, and the serotonin transporter gene, *5HTT*; Cases et al., 1995; Holmes et al., 2002; Trainor et al., 2009), vocal learning and recognition (early growth response 1 gene, *Egr1;* Mello et al., 1992), amongst others (Robinson et al., 2005). Fig. 1 illustrates the complex interactions between brain, genome and the social environment, and highlights some further examples of genes that have been implicated in specific traits (Robinson *et al.* 2008). However, despite our existing knowledge and given that animal behaviour is a highly complex entity, many major questions still remain. For instance, how do these genetic and cellular pathways and networks operate in time and space in response to external stimuli and facilitate the internal changes required to cause appropriate reactions (Robinson *et al.* 2008)? How do genomic and epigenomic architecture and activity integrate with sensory experience and learning to impinge on structural plasticity within the brain (Caroni et al., 2012)? Commitment to understanding these issues has been established under initiatives such as the National Institute of Health's BRAIN (brain research through advancing innovative neurotechnologies) initiative (www.nih.gov/science/brain/). Some progress has already been made in this area: the field of optogenetics, cell-specific loss or gain of function via combined genetic and optical methods, has developed rapidly since the light-activated, membrane potential-altering "opsin" genes were brought into neurogenetic studies (Deisseroth 2010, 2011), allowing scientists to control well-defined, single-neuron events in space and time. While this technology is beginning to generate insight into the function of specific pathways in the brain in regulating aspects of behaviour and indeed delve deeper into the basis of certain disease states (Zalocusky & Deisseroth 2013), many questions still remain.

Where model species offer salient context-dependent proxies for a wide range of processes and traits (Bolker 2014), non-model species offer additional flexibility and the crucial ecological relevance for exploring the molecular basis of particular traits in greater detail (Parsons & Albertson 2013). Limitations of using non-model species present themselves in the relative lack of genomic and other resources, but advancing technology, lowering costs and more robust computational tools make these sorts of organisms ever-increasingly easy to work with (Bräutigam et al., 2008; Grabherr et al., 2011; Wheat, 2010). Many recent comparative transcriptomic studies using RNA-seq have made use of the technology's independence of reference sequences to shed the first light on the molecular basis of traits relevant to those species (Kawahara-Miki et al., 2011; Moghadam et al., 2013; Schunter et al., 2014; Shi et al., 2011). These studies, while insightful, lack a thorough characterisation of the error and bias associated with annotating the transcriptome of a particular species using reference sequences from a different species. Additionally, while the emergence of many novel and useful computation tools for the assembly, annotation and analysis of transcript sequences from next generation sequencing technology has propelled the field forward, the relative merits, errors and bias associated with each, particularly when used with species lacking an annotated genome, have yet to be thoroughly explored.

Comparative transcriptomics allows the exploration of interesting changes in gene expression and transcript complexity such that modules and networks of genes influencing phenotypes can be identified. Prior to the common usage of RNA-seq, microarray studies contributed, and indeed continue to contribute, insights into the genomic influence and regulation of trait evolution and development (Aubin-Horth et al., 2007; Brunberg et al., 2013; Czibere et al., 2011; Renn et al., 2008). It is often the case that the most interesting ecological models of a given trait are not adequately presented by those species typically considered as 'model' for which the greatest quantity and quality of genomic resources are available. Although genomic resources have expanded hugely, and continue to do so, it remains costly in both time and money to sequence and annotate an entire genome. Using microarrays, heterologous hybridisation of probes and transcript sequences between different species has been shown to be useful in identifying species-specific gene expression, although this is impacted by sequence divergence between the sequences used (Machado et al, 2009; Renn et al., 2004). Harnessing the benefits of RNA-seq, comparative transcriptomics can be far more effectively applied to non-sequenced as well as sequenced species (Wang *et al.* 2009a; Ozsolak & Milos 2011), and is also being developed for simultaneous host-pathogen sequencing (reviewed by Westermann et al., 2012). However, despite the increasing number of studies using RNA-seq in comparative transcriptome analyses of non-sequenced species (Collins et al., 2008; Crawford et al., 2010; Kawahara-Miki et al., 2011; Künstner et al., 2010; Garg et al., 2011; Dassanayake et al., 2009), it remains unclear exactly how sequence divergence between the transcriptome and reference species impacts on the efficacy and accuracy of transcriptome recovery. Additionally, given that there is still uncertainty over the most appropriate

method for analysis when annotating the transcriptome of a species even when using its own genome (Garber *et al.* 2011), choosing the best strategy to annotate the transcriptome of a species with no genomic resources available presents a continual problem to researchers. To our knowledge, a comprehensive analysis of the most effective, accurate and least functionally biased transcriptome annotation strategy when using a reference sequence from an alternative species has yet to be conducted.

1.3 Sexual selection, social behaviour and mating systems

1.3.1 Sexual selection

The term sexual selection was coined by Darwin (1871) to explain the evolution of characteristics that do not confer advantages via natural selection. The influence of sexual selection on brain gene expression to modulate aspects of behaviour, such as mating, currently remains poorly understood. Sexual selection arises from the competition between individuals for access to reproductive resources resulting from variation in the number and quality of those reproductive resources (Emlen & Oring 1977). Intersexual competition represents the variable ability of individuals to be selected for mating by members of the opposite sex (see (Petrie 1983; Arnqvist 1992; Székely *et al.* 2010) and often leads to the evolution of sexual dimorphism and ornamentation, postulated to reflect individual genetic quality (von Schantz et al., 1999). Intrasexual competition drives variance in the ability of members of one sex to exclude other members of that sex from reproductive opportunities, leading to the evolution of traits such as weaponry (Emlen, 2008). Sperm competition, occurring between sperm from different males delivered into the female reproductive tract and impacting on their to fertilise the ova, is also one such trait and represents a key influencer of male reproductive success (Møller & Ninni 1998).

Intersexual competition most commonly represents female mate choice, whereas intrasexual competition occurs most often between males. This imbalance results from the typical physiological limits on reproductive success per sex: females invest heavily in producing offspring and hence there is a low maximum capacity relative to that of males, where reproductive success is positively correlated with the quantity of mating events. Intrasexual competition in females is known to occur and may manifest as the competition between females for reproductive access to males, or specifically to high quality males where mate quality impinges on reproductive success (Clutton-Brock, 2007; Rosvall, 2011). The shift between which sex competes for access to reproductive resources depends upon the ratio of each sex that are ready for mating within the population (adult operational sex ratio); the level of polygamy, which defines the mating system (see below), depends on the degree of imbalance in the adult operational sex ratio (Emlen & Oring, 1977).

Songbirds provide an ideal model within which to study sexual selection by virtue of the extraordinary diversity observable in sexually selected traits, such as sexual dimorphism, ornamentation and song, the enormous diversity of species, particularly within the Passeriformes, and the wealth of ecological and behavioural data available. Revised nomenclature of the songbird brain (Reiner *et al.* 2004; Jarvis *et al.* 2005) and the recent sequencing of several bird genomes, including the songbird *Taeniopygia guttata* (zebra finch, Warren et al., 2010) have, respectively, demonstrated surprising homology with mammalian brain regions and provided useful resources for comparative genomics studies using bird species. This has facilitated the emergence of songbirds as excellent candidate systems for the exploration of genotype-phenotype interactions particularly related to brain and behaviour (Clayton et al., 2009). As such, this study utilises two species of songbird to explore the molecular basis of differences in mating system, described below.

1.3.2 Evolution and neurocircuitry of social behaviour: pair bonding

Animals, from ant to elephant, display profound variation in social behaviour. Social traits, such as group living, cooperation, affiliation, aggression, communication, and parental care, have been the subject of intense study for many years. Recent work with many species has revealed that complex social traits, including vocal learning, social dominance and pair bonding, have strong genetic underpinnings (Aubin-Horth et al., 2007; Garfield et al., 2011; Mello et al., 1992; Young & Wang, 2004). These studies have begun to reveal how genes and neural substrates lead to the diverse social behaviour that has puzzled evolutionary biology ever since Charles Darwin (1871).

Neuroethological studies have identified particular regions of the brain that are conserved across many taxa and have prominent roles in facilitating these behaviours. In particular, Sarah Newman's (1999) synthesis of the animal social behaviour network, comprising specific nodes of the brain's limbic system that are involved in reproductive, parental and aggressive behaviours in both sexes via sensitivity to hormones and neurochemicals (Newman 1999), has laid the groundwork for further identification of integrated brain regions and gene networks with conserved functions related to social behaviour. In particular, O'Connell and Hofmann's findings that regions of the brain expressing neurochemical genes implicated in social behaviour and decision-making are conserved across major vertebrate lineages (reptiles, birds, mammals, amphibians, and teleosts) have led to the description of the vertebrate social decision making network (SDMN), which incorporates Newman's social behaviour network and the mesolimbic reward system, which are functionally interconnected (O'Connell & Hofmann 2012a) – see Fig. 2. A key finding of theirs was the inherent variation in the spatial expression profiles of ligands but not of receptors,

indicating that these brain circuits exhibit a conserved 'hard-wired' signalling infrastructure and a flexible ligand signalling system. The exact nature of this varies according to region, perhaps reflecting the effect of different selection pressures acting on regions with more basic physiological roles versus those with more receptive, cognitive functions.

Animal social decision-making encompasses the evolution of patterns that operate at the individual and the population level, manifesting as approach or avoidance mechanisms of behaviour in response to challenges or opportunities (O'Connell & Hofmann, 2011). Tendencies for affiliation and subsequent attachment between individuals, termed pair bonding, are social traits where selective and preferential associations occur between individuals – parents and offspring, or between adults – representing one of the most basic functions of the social brain. The demands of pair bond formation are perhaps the most important aspect of intra-specific behaviour for the evolution of the social brain (Dunbar & Shultz 2007). Pair bonds, as key influencers of behaviour, underpin many core aspects that define animal and indeed human societies, such as group living, the spread of cultural information, extending to, in humans, the arts and politics (Massey 2002). In humans, the emergence of pair bonding behaviour most likely coincided with the development of increased cranial capacity and the laterality of the brain in *Homo erectus* (Massey 2002).

Having a strong neurobiological basis, various neuropeptides, neurotransmitters, and their respective genes have been implicated in modulating pair bond formation (Young & Wang, 2004). Detailed studies using *Microtus* voles have implicated the conserved neuropeptides oxytocin and arginine vasopressin (AVP) in impacting upon mating systems (polygamy versus monogamy) by affecting the formation of pair bonds (Ahern & Young, 2009; Cho et al., 1999; McGraw & Young, 2010; Ophir et al., 2012) – homologous effects have since been demonstrated across different vertebrate lineages (Bielsky et al., 2004; Clipperton-Allen et al., 2012; Oldfield & Hofmann, 2010; Sala et al., 2011). Significantly, the same neuropeptides have been linked to human behavioural disorders such as autism spectrum disorder (Kim *et al.* 2002; Jacob *et al.* 2007). The signalling pathways of other neuropeptide hormones and neurotransmitters, such as dopamine (Aragona et al., 2006; Goodson et al., 2009a; Shahrokh et al., 2010) and serotonin (Cases *et al.* 1995; Holmes *et al.* 2002), and gonadotrophin-releasing hormone (GnRH, Maruska et al., 2011; White et al., 2002), have been implicated in the modulation of various behavioural states that impact on pair bond formation, such as affiliation, aggression and reward. The expression distribution in the brain of receptors for oxytocin and AVP, along with other receptors for ligands important in modulating various social traits, occurring within regions of the SDMN is conserved across many lineages (O'Connell & Hofmann 2012a), suggesting that the key molecular components mediating pair bonding are evolutionarily ancient. Together, these observations indicate that while the pathways that modulate pair bonding may be centred on oxytocin and AVP receptor signalling, they are

integrated within circuits that modulate other, related social traits and hence, there may be some degree of crosstalk and/or redundancy by which internal responses to a wide variety of external cues impact differentially on pair bond behaviour.

Pair bonding in a mating context represents reproductive opportunities that lead to learned and remembered preferences for affiliation ("approach") via social and sensory experience, reward pathways, hormonal activity, and neural plasticity (Goodson et al., 2009a; O'Connell & Hofmann, 2011). As such, the neural processes, and indeed the effects of those processes, occurring within the brains of typically monogamous individuals versus typically polygamous species are likely to be very different. While oxytocin and AVP signalling may underpin the formation of partner preferences, given that choosing a mate impacts on subsequent trait expression (mating, reproduction, aggression, parental behaviour) and that oxytocin and AVP receptor signalling in the brain is integrated with pathways that impact on other traits, it is likely that there exists complex inter-regulation between the causes and effects of all these traits. As sexual behaviour can be linked to dominance (Clutton-Brock et al., 2006; Maruska et al., 2011b), and social stress levels (often measured via the activity of the hypothalamic-pituitary-adrenal [HPA] axis) are also linked to dominance (Kotrschal et al., 1998; Stefanski & Engler, 1999), it is likely that sexual behaviour in at least males is related to stress levels. Given that activity of the HPA axis can be modulated by and impact upon neurochemical signalling pathways such as AVP, serotonin and dopamine pathways (reviewed by Blanchard et al., 2001), it is therefore likely that the occurrence and impacts of stress in the brain from social cues and responses is related to pair bond formation. Disentangling the genomic drivers and regulators of these behaviours and their effects remains a key challenge for molecular genomic studies.

1.3.3 Mating systems

The evolved tendencies of members of a species to form pair bonds between unrelated individuals, categorised as mating systems, can vary between the extremes of monogamy, where two individuals preferentially mate with each other, and polygamy (including polygyny, polyandry, and polygynandry) where individuals mate with multiple individuals. Mating systems can be characterised on their social and genetic bases, for instance individuals who preferentially affiliate and mate with each other (socially monogamous) may also engage in extra pair copulations (EPCs) leading to extra pair young (EPY, genetically promiscuous, Westneat et al., 1990). The extent of successful mating outside a social pairing, measured as extra-pair paternity (EPP), the proportion of offspring sired by alternative males, can be used as a measure of genetic monogamy and indicates strong pair bonding (Griffith et al., 2002), although this can be impacted by other factors (Cohas & Allainé 2009). The rate of EPC is thought to be influenced by ecological (population-specific)

factors, whereas the probability of EPY occurring subsequently is likely modulated by processes that are consistent above the species level (Brommer *et al.* 2010). Given that pair bonding tendencies vary considerably across many taxa, using proxies for pair bond strength, such as EPP, mating systems can be used in comparative studies to investigate the ecological and molecular components that influence and underlie pair bond formation.

Mating systems shape and are shaped by other factors such as adult operational sex ratio, parental care requirements, food resource abundance (Emlen & Oring, 1977), philopatry (Greenwood 1980) and social structure/cooperation (Clutton-Brock et al., 2006). The mating system (particularly monogamy) is associated with impacts on sexually dimorphic gene expression (Hollis *et al.* 2014), which can translate into effects on phenotypic sexual dimorphism in birds (Pointer *et al.* 2013). The type of mating system operating within a population exemplifies the extent of sexual selection acting on individuals as it modulates the competition for access to reproductive resources: polygamous species experience stronger sexual selection as more members of those species are competing for mates compared to monogamous species. Increased sexual selection in polygamous species across a wide range of taxa is associated with increased relative testes size (gonadosomatic index, GSI, see Calhim & Birkhead, 2006). GSI correlates with EPP and with circulating testosterone levels, a key male sex hormone contributing to the development of sexual characteristics and behaviour (Garamszegi et al., 2005). GSI has also been shown to correlate with number and activity (Fos production) of dopaminergic neurones in the ventral tegmental area (VTA) region of the zebra finch brain, known to regulate reward pathways, in male subjects exposed to females (Goodson et al., 2009b). As such, mating systems provide useful associated proxies for the tendencies towards pair bonding and the levels of sexual selection acting on a given species, from which the differential impacts on gene expression and the brain can be assessed through comparative study. However, the molecular basis of mating system evolution remains poorly understood, particularly in songbird species.

1.4 Using next generation sequencing to understand complex traits

The dramatic advancement of sequencing technologies has led to a vast increase in the amount of sequence data that can be generated to investigate biological questions, particularly how molecular factors interact to produce complex traits. A complex trait is, by definition, "any phenotype that does not exhibit classic Mendelian recessive or dominant inheritance attributable to a single gene locus" (Lander & Schork 1994). Complex traits, such as social behaviour, are often impacted by multiple cues of differing context – both external, from the physical and social environment, and internal, by way of epistatic and epigenomic modulation of functionally-related gene expression (Székely et al., 2010). As such, many gene loci often underlie the development and expression of a

given complex trait and exert relatively weak influences such that identifying the key genomic loci that underpin the trait is exceedingly difficult. Much has been learned from genome-wide association studies concerning the identification of genes important in certain phenotypes and also the relative contribution of multiple genomic loci to quantitative traits (Mackay et al., 2009). However, the amount of phenotypic variation that these explain is often very small: generating sufficient data from enough biological samples to adequately identify and map variants associated with quantitative traits has posed problems in terms of both technological availability, and the knowledge of how to accurately analyse the resultant data (Houle 2010). Gene identification following quantitative trait locus (QTL) analysis requires further fine mapping using techniques such as chromosome dissection and positional cloning (Flint & Mott 2001). Thankfully, recent advances in genomic sequencing technology have led to the development of a variety of next generation sequencing platforms which have revolutionised our approaches to molecular explorations of complex traits by providing rapid, plentiful, accurate, and relatively inexpensive data generation (Ozsolak & Milos, 2011; Wang et al., 2009). For example, Colgan et al. (2011) used Roche 454 sequencing of the bumble bee *Bombus terrestris* transcriptomes from different developmental stages, sexes and castes to explore the differential gene expression associated with biological processes linked to polyphenism, identifying candidate phenotypic influencers (Colgan *et al.* 2011). Shi et al. (2011) used the Illumina platform to deep sequence the transcriptome of the tea plant *Camellia sinensis*, generating insight into pathways influencing metabolism and the production of compounds important to the quality of tea (Shi *et al.* 2011).

Next generation transcriptome sequencing, termed RNA-seq, provides genome-wide deep sequencing of RNA transcripts within a sample to single base resolution, from which gene expression profiles can be elucidated, providing snapshots of genome activity, transcriptome composition and complexity (Wang et al., 2009), see Fig. 3. Where Sanger sequencing used dye-labelled dideoxynucleotides in a chain termination reaction to extend primer sequences on PCR-ed template sequences, next generation technologies have developed around several principles, including liquid-phase emulsion PCR and solid-phase amplification (reviewed by Metzker, 2010). Prior to RNA-seq, microarrays formed the main method of exploring gene expression profiles. Here, gene sequence probes are immobilised on a tiling array, to which cDNA (derived from RNA samples) is applied and expression is detected via fluorescence (see Benes & Muckenthaler, 2003). Although microarrays are still a popular method to use, especially where a known group of well characterised sequences are being probed for, RNA-seq holds several advantages over microarrays: it requires a relatively low amount of RNA, gene sequences and polymorphisms can be obtained *de novo* rather than needing to be specified *a priori*; computational annotation of RNA-seq data does not restrict the quantity of genes that can be detected, as is the case when probing microarrays; RNA-seq generates negligible background signal; genes with a high dynamic range of expression can be detected with RNA-seq (Wang *et al.* 2009a). However, RNA-seq is not without its flaws:

due to the fragmentation process involved, RNA-seq can generate bias according to transcript length, where longer transcripts can be over-represented in the data set – a feature that is not exhibited by microarrays (Oshlack & Wakefield 2009).

The double-edged sword of RNA-seq, and 'omics technologies overall, is manifested in the deluge of data that can be, and has been, generated. Where challenges have been posed for computing storage capacities and processing power – the requirements for which were predicted to double every 18 months but in fact have increased by five times that amount every year since 2002 – further challenges have become apparent in the need to identify subtle indicators among multiple data streams and appraise these in the context of existing findings (Berger et al., 2013). In response, the scientific community has produced and benchmarked many new computation tools for the alignment, assembly and annotation of short read sequences plus transcript expression analysis (Garber et al., 2011), and functional pathway and network exploration (Huang et al., 2009; Langfelder & Horvath, 2008), amongst others.

1.5 Integrating 'omics

Alongside the rapid expansion of genomic sequencing capacity for genomics and transcriptomics, so too have other high throughput 'omics technologies advanced with the goal of easing the characterisation of 'genotype-phenotype' maps. These include protein sequencing and analysis (proteomics; Roepstorff, 2012) and identification of metabolite profiles (metabolomics; Fuhrer & Zamboni, 2015). The most challenging of these, due to its multi-dimensional, temporally and spatially variable nature, remains the complete repertoire of phenotypes displayed by an individual or species, the phenome (Chen et al., 2014), and how genomic and environmental influences contribute to this dizzying diversity (Houle et al., 2010). Integrating these various 'omics technologies presents an additional layer of analytical complexity over and above each individual data stream. However, in recent years, this has become a priority for the international community, with the designation of funds and effort dedicated to advancing this field (please refer to the NERC-funded Environmental 'Omics Synthesis centre, environmentalomics.org). In response, these techniques have been discussed (Pathak & Davé, 2014; Yang et al., 2011) and advanced over the last few years across a wide range of life science fields such that insight has been generated into the functional linkages between levels of molecular complexity (Ahn *et al.* 2011; Durban *et al.* 2013; Urich *et al.* 2013). As a field in its infancy, and given the continuous increasing availability of open source tools for analysing 'omics data and reducing data generation costs, it provides exciting strategies to explore and shed light into complex biological questions to a much deeper degree than has hitherto been generally possible. For example, prominent challenges currently facing human health, such as understanding how pathogen genomes allow them to functionally

adapt to and become stable in new hosts, evading host immune systems, provide exciting areas to which these techniques can be applied. Certain *Photorhabdus* species of bacteria have recently been found to infect mammals as well as their typical insect hosts – a host switch that may have resulted from adaptation of metabolic pathways allowing survival at higher mammalian temperatures.

1.6 Thesis objectives

The overarching aim of this PhD thesis was to apply state-of-the-art systems biology techniques to questions concerning the genomic evolution of ecological variation in complex phenotypes, using next generation transcriptome sequencing complemented by other methods. To achieve this aim, the specific thesis objectives were to learn the techniques required to process, analyse, and evaluate next generation transcriptome sequencing methods, how these could be integrated with other 'omics data streams, and then to apply these methods to novel transcriptome data to explore the genomic basis of differences in social traits related to mating systems. As such, in Chapter 2, I worked with existing transcriptome and phenotype data in collaboration with the Waterfield research group at the University of Bath, analysing and integrating bacterial RNA-seq and phenotype microarray data. I integrated comparative functional genomics with high throughput phenotype microarray data sets in the *Photorhabdus* system to explore the molecular basis of host switching. Phenotype microarrays in this context involve culturing bacteria on various substrates immobilised on multiwall plates, measuring respiration via a reporter dye, allowing the quantitative measurement of many cellular phenotypes at one time. I have integrated comparative RNA-seq and phenotype microarray data sets from insect-restricted species and those derived from mammalian clinical isolates when cultured under various growth conditions. I have found that varying substrates elicits the greatest changes in gene expression, compared to temperature and growth phase, and adaptations to differing host environments may be centred around specific metabolic pathways.

To explore the molecular underpinnings of mating system evolution, we chose to work with non-model songbird species: by virtue of their huge phenotypic diversity within closely related clades and the large volume of documented ecological traits including paternity data, they provide a wealth of resources. We were able to identify a pair of species with opposing mating systems, the water pipit and the dunnock, that were closely related to each other and the closest species with an available reference genome sequence, the zebra finch. However, given that we decided to use non-model species with no genomic resources available and hence had to use the reference genome from a different species to annotate the transcriptomes, a key problem to solve before we embarked on this task was to identify the most appropriate transcriptome annotation method to use with non-

model species, and explore how sequence divergence between the transcriptome and reference sequences impacted on gene detection efficacy and accuracy. Therefore, in Chapter 3, I used published RNA-seq and genome data from the *Drosophila* family to conduct a comprehensive power analysis of transcriptome annotation methods and varying degrees of divergence between transcriptome and reference genome sequences on accurate and unbiased transcriptome recovery. To understand the impacts of sequence divergence between transcriptome and reference sequences when derived from different species, and to identify a preferred strategy for annotating the transcriptome of a species with no available genome sequence, I compared widely used methods for the annotation of transcriptome data with a novel strategy of short read-to-genome mapping. By sequentially mapping RNA-seq data from *Drosophila melanogaster* to its own genome and those of the 11 other sequenced *Drosophila* species, I have characterised the efficacy, accuracy, and functional bias associated with two commonly used transcriptome annotation strategies (*de novo*, and genome-guided assembly of RNA-seq data, followed by homology searching) compared to a novel, simpler approach whereby short RNA-seq reads are aligned directly to a reference genome and assigned to genes based on coordinates. I have found that this latter technique, termed direct genome mapping (DGM), outperformed both of the assembly-based methods in all tests performed, indicating that it is the most appropriate method for recovering an accurate and representative profile of expressed genes both when a reference genome is available and when a closely-related alternative must be used.

With knowledge generated in Chapter 3, I could then embark on Chapter 4: the sequencing, annotation, functional characterisation, and comparison of the water pipit and dunnock brain transcriptomes – the first exploration of the genomic differences between wild monogamous and polygamous songbirds. We aimed to use both males and females, to explore sexual dimorphism in gene expression related to mating systems. We chose to use free-living animals as we wished to capture the natural, rather than laboratory influenced, context of genome-wide differences. In order to gain insights into the potential molecular basis of mating behaviour, I obtained, annotated, and analysed novel brain transcriptome data from two wild-caught songbird species, the water pipit and dunnock, which hitherto did not have any genomic resources available. The water pipit is highly monogamous, whereas the dunnock is highly polygamous. These species are closely related to each other and their closest common reference species, which is also classified within the family Passeridae, the zebra finch – the genome of which has recently been published (Warren *et al.* 2010). I obtained brain samples from wild-caught individuals during their breeding season, which were sequenced using RNA-seq, and subsequently analysed using DGM, as this was identified to be the most appropriate technique in Chapter 3. By characterising and comparing the gene expression profiles from these species, I have provided the first insight into the genome-wide differences in gene expression that may underlie the causes and/or effects of behavioural differences around pair bonding preferences in these species – factors that may underpin

differential mating system evolution in songbirds. These findings provide a proof of principle for this type of analysis in wild species with no available reference sequence and gene identities to guide further molecular explorations of the pathways that underlie mating system evolution. This work was conducted as part of a wider study using species pairs from a wide variety of lineages to identify the key functional molecular components and pathways that may underlie the evolution of pair bond formation.

1.7 Thesis impact

This thesis provides a synthesis of 'omics data that has several overarching impacts. Integrating bacterial RNA-seq and phenoarray data, I identify possible functional pathways that have evolved differentially to facilitate host switching that has led to the occurrence of clinical pathologies. As such, this work paves the way for further studies into the molecular mechanisms that permit certain *Photorhabdus* species to infect mammals, including humans, which could lead to more effective therapies. Having presented a thorough assessment of transcriptome annotation techniques and identified the most appropriate for use with species lacking a genome sequence – which has been submitted for publication in a peer reviewed journal – this provides the international scientific community with clear guidelines of how to more effectively design experiments using species that lack genomic resources for comparative genomic studies. By enhancing the efficacy, accuracy, and robustness of data generated from such studies, the advice provided herein has the potential to improve the range of conclusions that can be drawn across all areas of molecular ecology, enhancing the output of the community at large. Having applied my transcriptome annotation method to data derived from a pair of wild-caught songbird species, I provide not only a proof of principle for this type of analysis but I also implicate specific genes and pathways within the brain as being involved in the causes and/or effects of behavioural choices related to monogamy versus polygamy in songbirds. This provides an excellent basis for developing this avenue of inquiry into a greater programme of research, which has the potential, when combined with findings from other lineages and further molecular biological studies, to elucidate the impacts of sexual selection on the genome and its activity in the brain, and how these impacts translate into behavioural programmes.

1.8 Figure legends

Fig. 1. **Taken from Robinson et al.'s review entitled *Genes and social behavior* (2008).** The central diagram illustrates the complex interconnections between the brain, genome and the social environment. The authors eloquently describe how "these relationships operate over three time scales: (i) physiological time via effects on brain activity (solid lines), (ii) developmental time via slower effects on brain development and genome modification (dotted lines), and (iii) evolutionary time via the processes of natural selection (dashed line)" (Robinson *et al.* 2008). Vector 1 refers to the directional effects of social information toward altered brain and behaviour via neural transduction leading to genome responses and modification. Vector 2 indicates how genetic variability impacts on social behaviour via the action of RNA and protein expression and activity impacting on brain cells and systems. The surrounding images present a selection of the animals, social traits, and genes discussed in their review.

Fig. 2. **Schematic representation of two neural circuits implicated in modulating social behaviour in the mammalian brain: the mesolimbic reward circuit (MRC, top), and the social behaviour network (SBN, bottom), taken from O'Connell & Hofmann, 2011.** The specific brain regions involved in each circuit are labelled with colour – blue for the MRC, yellow for the SBN. Shared regions are shown in green. Directionality of functional connections is indicated with arrows. Abbreviations: AH: anterior hypothalamus; blAMY: basolateral amygdala; BNST: bed nucleus of the stria terminalis; HIP: hippocampus; LS: lateral septum; meAMY: medial amygdala; NAcc: nucleus accumbens; PAG/CG: periaquaductal gray/central gray; POA: preoptic area; STR: Striatum; VMH: ventromedial hypothalamus; VP: ventral pallidum; VTA: ventral tegmental area.

Fig. 3. **Schematic overview of the principles of RNA-seq data generation and annotation (adapted from Park, 2009; Wang et al., 2009).** The coding population of mRNAs are separated from total RNA, fragmented, and reverse transcribed, adding adaptor sequences. These are then used as templates for high throughput sequencing. The resulting sequences can then be aligned to a reference genome, permitting classification and annotation of the mRNA short reads. ORF: open reading frame.

Fig. 1

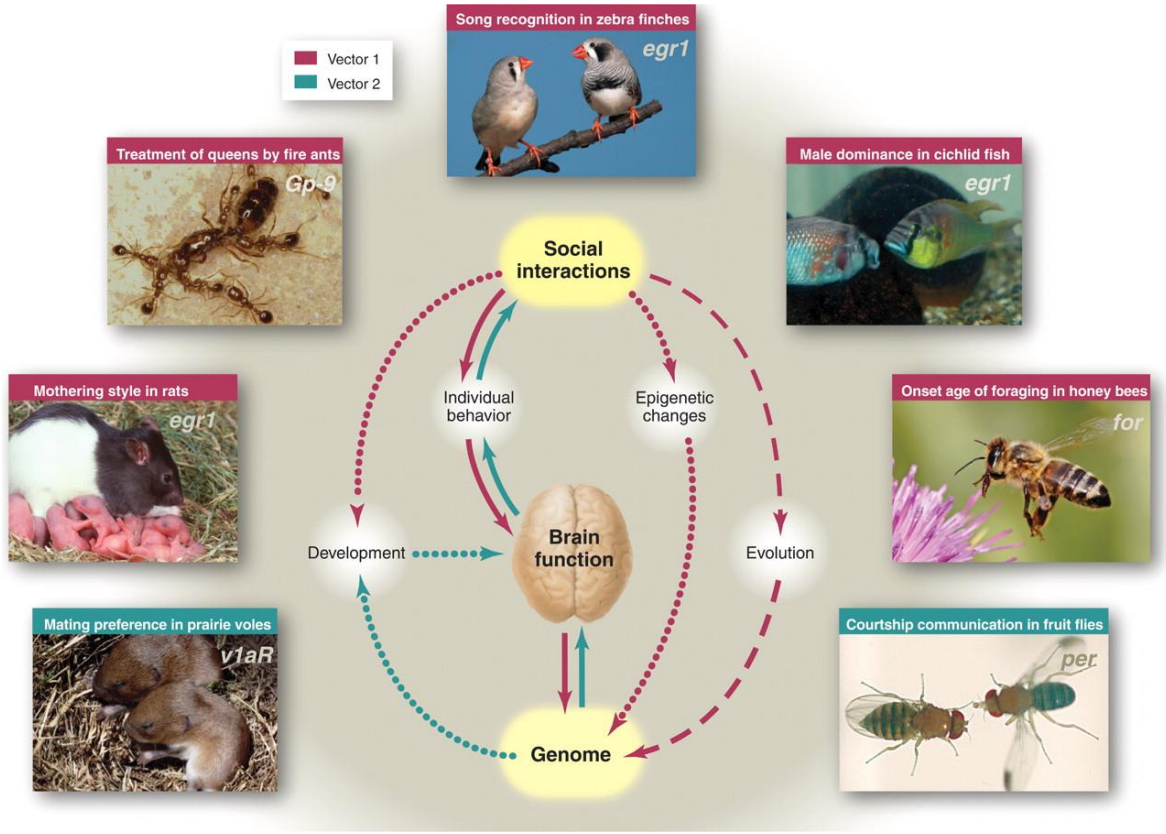

Vector 1
Vector 2
Song recognition in zebra finches
egr1
Treatment of queens by fire ants
Gp-9
Male dominance in cichlid fish
egr1
Social interactions
Individual behavior
Epigenetic changes
Development
Brain function
Evolution
Genome
Mothering style in rats
egr1
Onset age of foraging in honey bees
for
Mating preference in prairie voles
v1aR
Courtship communication in fruit flies
per

Fig. 2

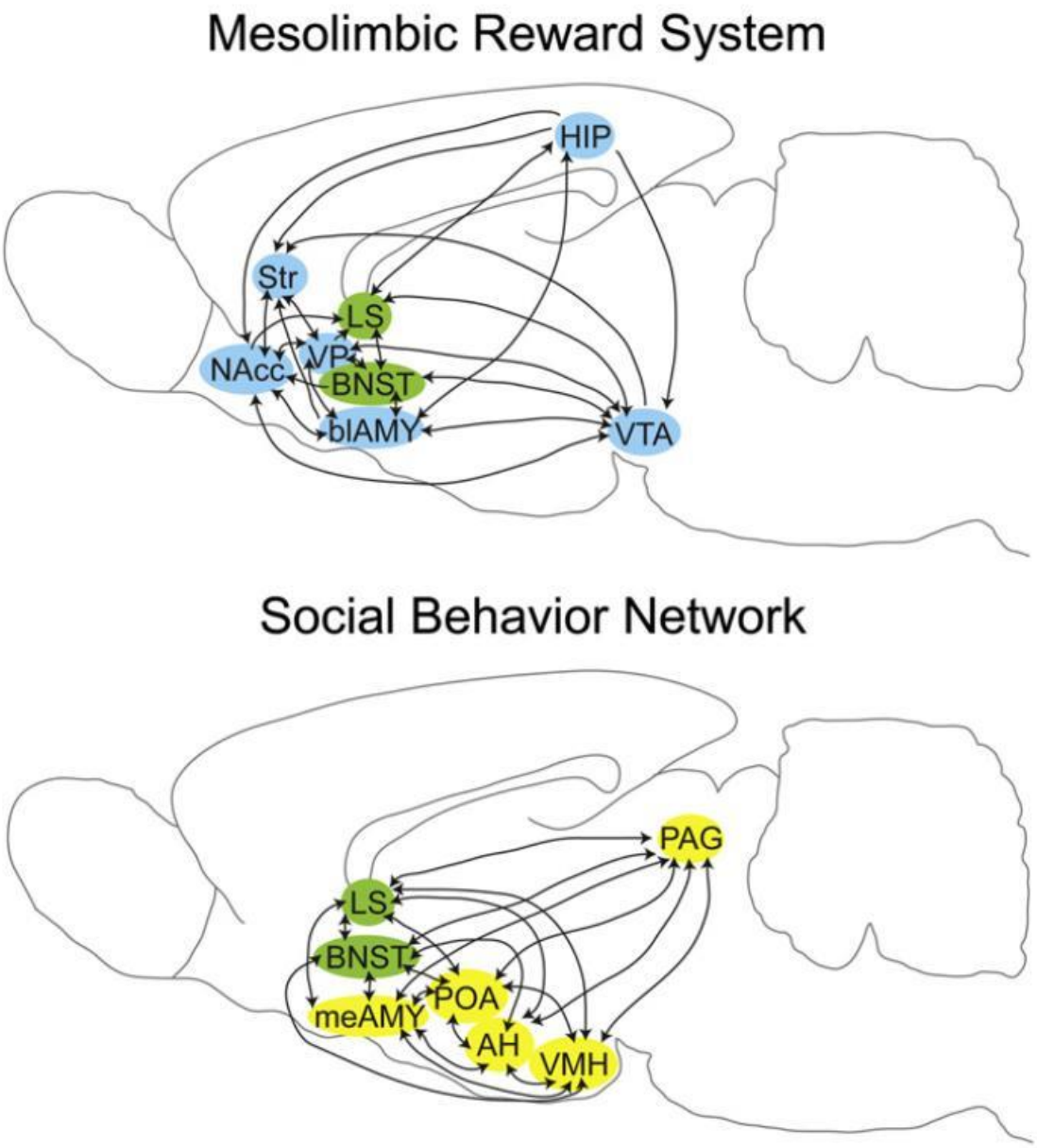
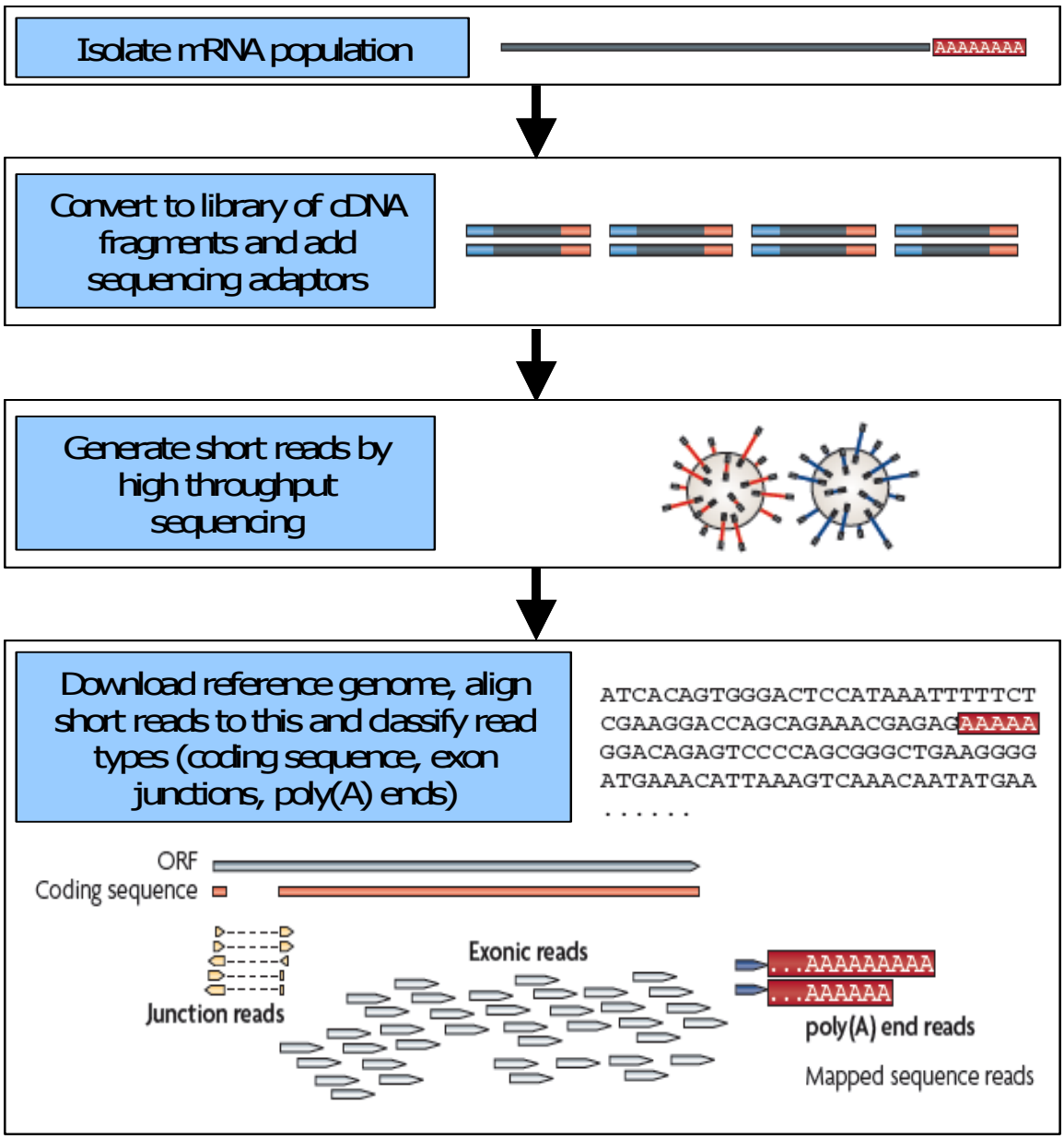

Mesolimbic Reward System
HIP
Str
LS
NAcc
VP
BNST
blAMY
VTA
Social Behavior Network
PAG
LS
BNST
meAMY
POA
AH
VMH

Fig. 3


Isolate mRNA population
AAAAAAAA
Convert to library of cDNA fragments and add sequencing adaptors
Generate short reads by high throughput sequencing
Download reference genome, align short reads to this and classify read types (coding sequence, exon junctions, poly(A) ends)
ATCACAGTGGGACTCCATAAATTTTTCT
CGAAGGACCAGCAGAAACGAGAGAAAAA
GGACAGAGTCCCCAGCGGGCTGAAGGGG
ATGAAACATTAAAGTCAAACAATATGAA
. . . . . .
ORF
Coding sequence
Exonic reads
Junction reads
...AAAAAAAAA
...AAAAAA
poly(A) end reads
Mapped sequence reads

2 Exploring the transcriptomic and metabolic basis of differential host targeting in *Photorhabdus*

2.1 Abstract

Photorhabdus, bacterial symbionts of the entomopathogenic *Heterorhabditis* nematode worm, have recently been found to infect human as well as insect hosts, causing serious disease states. The molecular basis permitting certain *Photorhabdus* species to survive and proliferate in mammalian systems remains unclear. Taking an integrative approach, RNA-seq and phenotype microarray (phenoarray) data were obtained from insect-restricted (*P. luminescens* TT01, Pl^{TT01}), and human clinical isolates (*P. asymbiotica* ATCC43949, $Pa^{ATCC43949}$) of *Photorhabdus* species grown under various conditions, varying temperature, media and growth phase. We found that growth medium elicited the greatest difference in gene expression, leading to changes in specific gene functional pathways. In particular, the glycine, serine and threonine metabolic pathway appears to integrate species-specific differences in gene regulation between Pl^{TT01} and $Pa^{ATCC43949}$. RNA-seq data available from a second strain of *P. asymbiotica*, Kingscliff ($Pa^{Kingscliff}$), when analysed using the limited orthology data available for these species showed, that one gene - encoding a putative phage tail fibre protein - is up-regulated in both *P. asymbiotica* strains when grown in human serum-supplemented medium compared to basic medium, suggesting that acquisition of phage-derived elements may have occurred in tandem with adaptation to mammalian tissues. The phenoarray data showed that *P. asymbiotica* respiration at 37^{o}C was overall significantly lower than at 28^{o}C. This indicates that although this species can survive and proliferate at higher temperatures, there are associated metabolic changes. These findings open the way for further disentanglement of the molecular adaptations of *Photorhabdus* species to mammalian systems, aiding the development of therapeutic interventions.

2.2 Introduction

Here we investigate the molecular differences between species of *Photorhabdus* that infect different host lineages (insects versus mammals) to better understand genomic and metabolic changes that allow bacteria to survive in and infect new hosts, causing disease. Many human pathogens first arose from host switching events where a bacteria or virus which would normally infect other organisms acquires the ability to infect, reproduce and achieve effective transmission to other individuals in the human host. Indeed, it is now believed that some of the most devastating human epidemics and pandemics, such as bubonic plague (*Yersinia pestis*) and influenza, resulted from zoonotic infections which subsequently became stable and transmissible between humans (Keeling & Gilligan 2000; Taubenberger & Kash 2010). The human immunodeficiency virus (HIV) is known to have adapted to humans from simian immunodeficiency viruses (SIVs) in great ape populations (Sharp & Hahn 2011) and *Staphylococcus* species have been documented to transfer from dogs to humans (*S. intermedius*, Tanner et al., 2000), and from humans to poultry (*S. aureus*, Lowder et al., 2009). In many cases, novel human pathogens have serious health implications for the infected individuals. In general, most events of host switching tend to occur among more related species than between those more distant. This is likely to result from the fact that the molecular targets for the pathogen to achieve infection will be more similar if the pathogen finds itself in a related host. Other factors such as the host body temperature might also play a role in determining the chance of success in any event of host switching. For example, in human fungal pathogens thermotolerance is universal but is achieved through a variety of mechanisms (reviewed by Cooney & Klein, 2008).

Photorhabdus species are gram negative bacterial symbionts of entomopathogenic *Heterorhabditis* nematode worm (EPN), which infect, kill, and reproduce inside insects. Recently, cases have been reported of humans becoming infected by these mutualistic species pairs, representing an evolutionary shift in target host (Farmer et al., 1989; Gerrard et al., 2006; Gerrard et al., 2003; Peel et al., 1999; Plichta et al., 2009). There are three currently known *Photorhabdus* species: *P. luminescens*, *P. asymbiotica*, and *P. temperata* (Fischer-Le Saux et al., 1999). *P. asymbiotica* is the only species known to currently infect mammals (including humans) as well as insects, and can be classified into at least two subclades, prevalent in the USA and Australia respectively, plus two distinct European strains which may represent a third subspecies, see Fig. 1. It is currently unclear how *P. asymbiotica* acquired the ability to survive in mammalian systems to cause infection, and how *P. asymbiotica* differs metabolically from *P. luminescens*. The *Photorhabdus* life cycle is inextricably linked to that of the nematode, together forming a formidable symbiosis: bacteria inhabiting the host infective juvenile (IJ) nematode intestines are regurgitated upon entry into a prey organism and set up lethal septicaemia. Within the prey cadaver, the bacteria then biotransform the tissue into further bacteria which the nematode feed upon. A proportion of bacteria remain within the nematode as an intestinal biofilm and when the nematode then begins to

reproduce, these remaining bacteria are transmitted to the new IJ worms, which develop, cause matricide and disperse in search of further prey. *P. asymbiotica* infections in humans have been reported sporadically around the globe and its exact incidence remains unclear. Clinical presentations are typically characterised by invasive or disseminated soft tissue infections, where additional sites of soft tissue infection may develop, indicating systemic spread. Treatment responds to antibiotics although relapses may occur. See Waterfield et al., 2009 for an overview.

The pathogenic switch that has permitted *P. asymbiotica* to infect mammals seems likely to be at least in part mediated by an ability to grow at 37-42^{o}C rather than being restricted to 28-34^{o}C but may also stem from adaptation to alternative substrates present in mammalian tissues (Line et al., 2010). The molecular and genetic basis of the phenotypic changes to facilitate this switch remain poorly understood, although preliminary (unpublished) data from the Waterfield research group at the University of Bath (now University of Warwick) indicate that key pathways differing between insect-restricted strains and those derived from clinical isolates include asparagine and pyruvate metabolism (Prof. N. Waterfield, personal communication).

Given the existing detailed knowledge of the physiological, cellular and genetic mechanisms underlying the mutualistic relationships and typical prey attack of photorhabdus, and because of the apparent relatively recent host shift, the case of *Photorhabdus* represents an ideal system to explore the molecular adaptations involved in host shift events and how these are encoded in the genome. Recent advances in high throughout technologies have led to 'omics level data that can be integrated across multiple levels of complexity to yield far deeper insights into biological mechanisms underlying shifts in ecology than has so far been possible, being particularly useful in microbial systems (Urich et al., 2013; Yang et al., 2011). Next generation transcriptome sequencing (RNA-seq) has in very recent years, proved itself to be an extremely useful tool for exploring the genomic basis of phenotypic differences in both model and non-model organisms (Collins et al., 2008; Pinto et al., 2011; Wang et al., 2009), as discussed in Chapters 3 and 4, using techniques such as differential gene expression analysis with tools such as DESeq (Anders & Huber 2010). This technique is currently being used in bacterial systems to explore the molecular factors underlying virulence and infection (Engelmann *et al.* 2011; Mandlik *et al.* 2011).

Bacterial systems, in contrast to more complex animal systems, are extremely well suited to high throughput approaches that provide in-depth exploration of important phenotypes such as growth: by virtue of their ease of colonising diverse, overlapping and often extreme ecological niches they can be successfully cultured in small volumes in a vast array of nutrients and subject to varying environmental challenges (such as chemical stress, aerobic versus anaerobic conditions, Borglin et al., 2012). This permits real time measurements of whole system effects to be documented. One

such system is the phenotype microarray, or phenoarray (such as the OmniLog system from Biolog, Inc): bacteria are cultured in various 96 well plates that provide a wide range of controlled environmental conditions. Respiration levels are recorded by colorimetric changes effected by reduction of a redox-sensitive dye (Bochner et al., 2001; Bochner & Savageau, 1977). This system allows exploration of genetic differences between samples affecting nutrient usage (Bochner et al., 2001).

Assessing significant effects is especially important for the phenoarray when, in contrast to most other high throughput 'omics technologies, it has the longitudinal element of respiration over time to consider. There are various characteristics of a bacterial respiration curve that provide valuable information describing the behaviour of the colony within its respective growth conditions: the length of the lag phase (λ), the increase in respiration rate (the curve slope, μ), the maximum cell respiration achieved (the maximum value recorded for that curve, A), and derived from these, the area under the curve (AUC). The phenoarray manufacture's tools for analysing curve parameters have been found to lose much of the detail of the results and do not provide robust methods for the accurate analysis and comparison of bacterial respiration curves to distinguish statistically significant differences (Vaas et al., 2012). However, recently published statistical analysis packages, such as *grofit* (Kahm et al., 2010), used with the statistical programming language R (R Core Development Team, 2010), can be manipulated to address such challenges, as tested by Vaas et al. (Vaas et al., 2012). The authors rigorously benchmarked measures for quality control and compared model-fitting with model-free curve analysis to ensure reproducibility and reliability when comparing multiple data sets. A key factor in the processing of phenoarray data is in the generation of confidence intervals for the parameters of respiration curves in test wells: when comparing curves, overlapping confidence intervals of curve parameters indicate that the respiration occurring in those wells is statistically similar to that level of confidence. Confidence intervals that do not overlap identify those cases that are most significantly likely to be different from each other (Vaas et al., 2012). A key element of the authors' suggestions was that the thresholds and parameters utilised in any given study to define statistically significant differences between respiration curves should be flexible rather than prescriptive for the user to accommodate the nature of the microbial systems used.

Given that insect-restricted growth is the ancestral state for *P. luminescens* and *P. asymbiotica*, and that divergence from that occurred relatively recently for *P. asymbiotica*, it is likely that *P. asymbiotica* has recently acquired the ability to survive in mammalian systems: to cope with increased temperature and/or different tissue substrates. If the ability to survive in such a different environment is recently acquired, it is likely that survival pathways remain suboptimal and as such, we expected that *P. asymbiotica* growth at higher temperatures and in mammalian-like media would be more erratic than at lower insect temperatures. Considering recent observations in

Campylobacter species that higher temperatures result in changes in specific carbon source usage (Line *et al.* 2010), we hypothesised that there would be significant differences to respiration patterns and gene expression when *P. asymbiotica* is cultured at higher temperatures and in human serum, and that these differences may be restricted to specific metabolic pathways (null: there are no significant differences in growth patterns or gene expression). This would indicate that particular metabolic pathways have adapted in *P. asymbiotica* to mammalian systems to permit survival and growth. We predicted that differential gene expression would occur between different temperatures, growth media, and growth phase. If *P. asymbiotica* had adapted to higher temperatures rather than mammalian substrate usage, recruiting different molecular pathways, growth at higher temperatures would result in a comparatively greater proportion of differentially expressed genes. However, if this was not the case, and *P. asymbiotica* had adapted to mammalian tissue substrates rather than temperature, we would expect to see a greater proportion of differentially expressed genes when grown in different media than at different temperatures.

To explore the molecular and genomic differences between the insect-restricted *P. luminescens* species and the insect and human pathogen *P. asymbiotica*, a detailed integrative analysis of RNA-seq data with phenoarray data from both species grown in different conditions was performed. Analytical approaches for phenoarrays incorporated the best practice guidelines suggested by Vaas et al. (2012) for deriving respiration curve parameters and comparisons, using the *lattice* (Sarkar 2008) and *grofit* R packages, with common sense approaches that accommodated the highly variable nature of *Photorhabdus in vitro* growth (Prof. N. Waterfield, personal communication).

2.3 Materials and Methods

2.3.1 RNA-seq

RNA-seq data (Illumina HiSeq 2000) were available from the Waterfield group for *Pl*^{TT01}, *Pa*^{ATCC43949}, the *Pa*^{ATCC43949} plasmid pPAU1, and *Pa*^{Kingscliff}. Each species/strain were cultured at different growth phases (stationary versus exponential), at 28^{o}C and, for *Pa*^{ATCC43949} and *Pa*^{Kingscliff}, also at 37^{o}C. They were cultured in different media: all in lysogeny broth (LB); *Pl*^{TT01} also in LB supplemented with insect haemolymph (LBHm); *Pa*^{ATCC43949} and *Pa*^{Kingscliff} also in LB supplemented with normal human serum (LBNHS). There was only one biological replicate available for each species/strain in each growth condition. Short transcriptome read alignments had previously been performed using SSAHA (Ning et al., 2001) by Mr. Paul Wilkinson, University of Bristol: *Pl*^{TT01} reads were mapped to the published *P. luminescens* genome (Duchaud *et al.* 2003) and the *Pa*^{ATCC43949} and *Pa*^{Kingscliff} reads were mapped to the published *P. asymbiotica* genome (Wilkinson *et al.* 2009), downloadable from the NCBI Nucleotide database (reference sequences NC_005126.1 and NC_012962.1, respectively). Mapped reads were then assigned to gene model coordinates using custom Python scripts (gene coordinates files kindly supplied by Dr. G. Mulley, University of Reading). To integrate the phenoarray data with detailed gene-level data to provide gene level insight into possible pathways that had divergent function between *P. asymbiotica* and *P. luminescens*, gene differential expression analysis was performed using the R package DESeq (Anders & Huber 2010) on each combination of condition per species. A non-generalised linear model fitting option was used for consistency as not all the data would fit to a generalised linear model. Genes were called as significantly differentially expressed if the adjusted p value was under 0.05. A significant caveat on the interpretation of the differential expression analysis is that no replicate RNA-seq data sets were available and therefore this analysis should be considered as a preliminary exploration of possible gene-level and functional effects and a proof of principle. The differential expression tool used, DESeq, does not recommend using it without biological replicates as, without having several expression values for each gene from replicated experiments, it is not possible to accurately estimate natural variation in the expression levels of individual genes. However, DESeq does provide an option for using it with single or partial replicates by estimating variation, not from a range of expression values for the same gene, but from the range of expression across all genes in the list. The differential expression analysis allowed identification of the most statistically significantly differentially expressed genes between the growth conditions per species. To establish broad functional changes in gene profiles, genes were assigned, using custom Python scripts, to KEGG pathway annotations (Kanehisa & Goto, 2000; Kanehisa et al., 2012) downloaded from the KEGG database and results were collated for the number of differentially expressed genes and the number of annotated KEGG pathways identified per comparison.

2.3.2 Phenotype microarrays (phenoarrays)

Phenoarray data were available from the Waterfield group for *Pl*^{TT01}, *Pa*^{ATCC43949}, and *Pa*^{Kingscliff}: the ability of these strains to utilise different carbon and nitrogen sources for respiration and to tolerate a range of pH and osmolyte-stress environments was assessed at 28°C and, for the *Pa* strains, 37°C also, using the Biolog Phenotype Microarray system. Specifically, carbon (PM01, PM02), nitrogen (PM3B) and peptide (PM6, 7 and 8) plates were used, with biological duplicates for each. It should be noted that in order to use the Biolog Phenotype Microarray system it was necessary to make certain adaptations to the standard protocols. It was found that IF0A media was toxic to *Photorhabdus* and did not support respiration and therefore this was replaced with M9 salts in plates PM01, PM02, PM3B, PM06, PM07 and PM08). The carbon plates (PM01 and PM02) were supplemented with Casamino Acids (0.05% w/v) and 20 mM D-mannose was provided as a carbon source in the nitrogen (PM3B) and peptide plates (PM06, PM07, PM08). It was necessary to supplement the carbon, nitrogen and peptide plates with 1x RPMI vitamin mix to support respiration at 37^{o}C.

The phenoarray data comprised a series of 96 well plates that assessed respiration levels on key metabolite families (see Fig. S1). Individual replicates of the respiration data and means of those replicates per well of each plate were visualised in a series of plots using the *lattice* package in R (see Fig. S2). From these, it was possible to obtain an overall, visual assessment of firstly the success of obtaining suitable control well respiration replicates (as without this, further analysis of the experimental wells is not possible) and secondly, the general trends of respiration behaviour on the different substrates in each well of the plates. However, from these it is not possible to robustly assess whether (a) the control well replicates are significantly similar enough to represent a true control, (b) the experimental well replicates were similar enough to be included in the comparison, and (c) which experimental wells are significantly different enough from each other to represent a biological effect. Therefore, we developed a strategy to, as far as possible and practical, compare replicates and subsequently compare means of the replicates to determine significant differences between respiration curves. This comprised several steps: the R package *grofit* was used to calculate and compare curve parameters, and this was combined with a suite of custom Python scripts and excel spreadsheets to extract, evaluate and analyse information from the phenoarrays. The overall pipeline for this is outlined in Fig. 2.

Grofit calculates respiration curve parameters, including the lag phase (λ), the respiration rate (slope, μ), the maximum respiration level (A), and the area under the curve (AUC) – see Fig. 3. These parameters are used by *grofit* in a bootstrapping process to calculate confidence intervals for each parameter. Respiration curves can be compared to each other using the curve parameters and their respective confidence intervals (Vaas *et al.* 2012). Vaas *et al*. (2012) determined that the curve parameters A (maximum height) and AUC (area under the curve) are more robust than λ and

μ in determining differences between respiration curves with diverse shapes. Statistically significant differences between curves can be identified by comparing the overlap of confidence intervals for the different parameters of those curves: overlapping confidence intervals indicate statistical similarity between curve parameters (Vaas *et al.* 2012). The level of statistical significance of such differences depends upon where the confidence intervals are set.

Photorhabdus growth *in vitro* is sensitive and highly variable (Prof. N. Waterfield, personal communication). Given the inherent variability of the system, i.e. the erratic growth patterns that were observed with *Photorhabdus*, the level to where the confidence intervals were set needed adjustment as this determined the overall sensitivity of the results. Setting the confidence intervals too stringently may have excluded many of the data points and some meaningful biological insights might have been missed. Thus, with the *Photorhabdus* data presented here, a common sense approach was required to prevent exclusion of the majority of data, as suggested by Vaas *et al*. (2012). To set the confidence intervals for the A and AUC curve parameters to a level that included a large proportion of the data but excluded those data sets that were extremely different, we decided to use an approach that combined automated assessments of curve similarity, using *grofit*, with the smallest amount of manual assessment as possible. Firstly, each person within the project team (Prof. Nick Waterfield, Dr. Araxi Urrutia and Miss Nina Ockendon) independently marked lattice *plots* of the two replicates for each species in a given well as being similar or different to each other. Those replicates that all three people agreed were different were noted for all plates. It was assumed that this processes provided a relatively consistent level of comparison across all data sets as a starting point for setting common sense statistical thresholds.

Grofit was run on the raw data, which used bootstrapping to calculate the 95% confidence intervals. An excel spreadsheet was used to process this data from all wells of a given plate alongside the sets of replicates determined by the project team to be dissimilar. The spreadsheet used the data output from *grofit* to calculate wider confidence intervals by increasing the number of standard deviations, which included more of the data points. By varying the , confidence boundaries, it was possible to determine the optimum confidence intervals that excluded at a minimum all those replicates determined by the project team to be dissimilar for that plate. This provided thresholds for each plate that allowed curves from different wells on that plate to be compared to each other and determine differences. As this approach retained the greatest proportion of all data, it was hence termed the ‘maximum confidence interval’ method. Inevitably, more wells were discounted than were manually selected by the project team, highlighting the inadequacies of relying solely on visual assessments but also indicating that the manual assessments were relatively conservative.

The plate-specific thresholds for the A and AUC parameters of curves were subsequently inputted into the analysis pipeline: they were fed into *grofit* to determine whether curves were significantly different to each other. Comparisons were conducted on a per-plate basis (plates were not compared to each other) for both replicate and mean replicate data. Firstly, all replicate comparisons were conducted, and secondly the means of those replicates were compared between species per well. Thus, for each plate all differences between replicates and mean experimental data between species were to the same level of confidence. This process was performed for (a) 95% confidence intervals, the default value calculated by *grofit*, (b) maximum confidence intervals that excluded at a minimum all of the group-defined wells, and (c) medium confidence intervals that excluded approximately 20% of all wells. Comparing the results of these, it was deemed best to use the maximum confidence intervals in order to not exclude large quantities of the data (data not shown). One trade-off in using this approach against the benefit that all the wells can be compared at the same level of confidence is that by increasing the confidence interval boundaries to include a greater proportion of the data, the level of statistical significance in any findings is concurrently reduced. As different boundaries were calculated per plate, the statistical confidence across the set of plates analysed was variable and must be factored into any conclusions drawn from the complete data set.

The maximum confidence interval method was tested against an alternative approach whereby the effect size for each respiration curve parameter (A and AUC) of each well was utilised (data not shown). The effect size is an index of the magnitude of the mean differences between the curves being compared, highlighted by Vaas et al. as being helpful for enabling the user to identify biologically meaningful results (Vaas *et al.* 2012). For those wells on a given plate deemed to be notably different by the group effort, the minimum effect size that excluded all such wells was set as a threshold against which all other wells were assessed – if the effect size exceeded this threshold, the wells were excluded from the analyses. However, this method was found not to be as effective and consistent at removing all those wells selected by the project team and hence the maximum confidence interval method was used instead.

Custom scripts within the pipeline also calculate the type of growth occurring in each well: negative, minimal, positive, or optimal versus sub-optimal, depending on the plate used. This was used to provide further insight into the respiratory impacts of each growth substrate – see Results.

2.3.3 'Omics synthesis

The RNA-seq and phenoarray data streams were integrated to derive biological insights into the evolutionary differences of *Photorhabdus* concerning host growth conditions at a higher level of

functional complexity than can be gained using either data stream alone. Integration was achieved by condensing and refining the outputs of both analyses to human-readable levels: simple tables that were cross-referenced, either manually or using custom Python scripts. Phenoarray data for those growth conditions where the most differential gene expression occurred and the most KEGG pathways were modulated were cross-referenced to the KEGG pathways to see if particular substrates implicated by the KEGG pathway did indeed cause statistically significantly altered respiration. Orthologous genes differentially expressed within these pathways were selected to see if patterns of up-, or down-regulation were consistent with phenotypic observations.

2.4 Results

2.4.1 RNA-seq analysis implicates specific functional pathways in mediating differences between species with different host adaptability

Gene expression profiling revealed that more than 90% of all gene features probed against are detected for all *Photorhabdus* strains tested under the various conditions (Table 1). Differential gene expression analysis on transcriptome read counts per gene indicate that, in general, more genes are differentially expressed as a result of changes to the growth medium than changes to temperature or growth phase, when using a significance threshold of less than 0.05 for the adjusted p value (Fig. 3 and Table 2; gene expression levels not shown). For instance, Pl^{TT01} cultured in LB versus LBHm at 28^{o}C during stationary and exponential growth caused significant differential expression in 18 genes and 61 genes , respectively. 12 more genes were expressed were expressed when Pl^{TT01} was cultured in LBHm compared to LB at 28^{o}C during stationary growth. 57 fewer genes were expressed when Pl^{TT01} was cultured in LBHm compared to LB at 28^{o}C during exponential growth. Conversely, no genes were differentially expressed between stationary and exponential growth of Pl^{TT01} in either LB or LBHm. Given that only one replicate was available for each growth condition, the statistical confidence in any findings is low, and as such further biological replicates (at least two) are required to add robustness. The results presented here provide an indication of the genes and pathways that may be involved in functional differences in response to environmental factors.

In $Pa^{ATCC43949}$, 30 genes were differentially expressed between LB and LBHm at 28^{o}C during exponential growth (with 19 fewer genes detected in LBHm compared to LB); between LB and LBNHS at 37^{o}C during stationary, 54 genes were differentially expressed (33 more genes detected in LBNHS compared to LB); between LB and LBNHS at 37^{o}C during exponential growth, 32 genes were differentially expressed (14 more genes detected in LBNHS compared to LB). When LBHm at 28^{o}C was compared to LBNHS at 37^{o}C during exponential growth, only 4 genes were differentially expressed (18 more genes were detected in LBNHS compared to LBHm). $Pa^{Kingscliff}$ displayed notable differences to $Pa^{ATCC43949}$: although a similar quantity of genes were differentially expressed between LB and LBHm at 28^{o}C during exponential growth (19 genes), 170 fewer genes were detected. Comparing LB to LBNHS at 37^{o}C during stationary growth, only 7 genes were differentially expressed, and 155 fewer genes were detected. Again, LB at 28^{o}C versus LB at 37^{o}C during exponential growth results in few differentially expressed genes (6 genes), however 72 fewer genes were detected. The occurrence of so many genes in $Pa^{Kingscliff}$ being switched off compared to $Pa^{ATCC43949}$ indicates significant differences in gene regulation which may be in response to environmental cues.

Genes on the *P. asymbiotica* plasmid pPAU1 do not appear to be involved in adaptation to human tissues and temperatures as no genes from this were differentially expressed. Limited orthology data for *P. luminescens* and *P. asymbiotica* is currently available with which to be able to draw comparisons of common genes modulated under the differing growth conditions. However, KEGG functional annotations are available for *Pl*^{TT01}, *Pa*^{ATCC434949} and *Pa*^{Kingscliff} so that differentially expressed genes could be mapped to the corresponding functional pathways, allowing identification of common areas of differential regulation – see Tables 3-5. The condition where notably high numbers of differentially expressed genes occurred in both *Pl*^{TT01} and *Pa*^{ATCC43949}, LB versus LBHm at 28^{o}C during exponential growth, also exhibited some of the same KEGG pathways to which similar numbers of genes were assigned. Glycine, serine and threonine metabolism showed two genes up-regulated for *Pl*^{TT01} in LBHm and two genes down-regulated for *Pa*^{ATCC43949} in LBHm. Glycolysis/Gluconeogenesis showed three genes, two up-, and one down-regulated for *Pl*^{TT01} in LBHm whereas in *Pa*^{ATCC43949}, two genes were up-regulated in LBHm. For pyruvate metabolism, two genes were up-regulated for both *Pl*^{TT01} and *Pa*^{ATCC43949} in LBHm. *Pl*^{TT01} also demonstrated modulation of tyrosine metabolic pathways, and *Pa*^{ATCC43949} additionally demonstrated modulation of histidine metabolism, both with two genes down-regulated. These findings indicate that the differentially expressed genes impact on some similar functional pathways, and where they are increased or decreased similarly, they may represent orthologous genes – something that could be confirmed as further orthology data becomes available. As these instances of functional pathway regulation appear common to both *P. luminescens* and *P. asymbiotica*, they may represent evolutionarily ancient adaptation to their common host, insects. However, observing the pathways that are modulated when *Pa*^{ATCC43949} is cultured in comparable conditions for its recent human hosts, LB versus LBNHS at 37^{o}C during stationary and exponential growth, the greatest number of genes are differentially expressed for this species and a different set of pathways are identified. During exponential growth, four genes are down-regulated for the citrate (tricarboxylic acid) cycle and five genes are down-regulated for porphyrin/ chlorophyll metabolism. Interestingly, in *Pa*^{Kingscliff} this comparison causes the most genes to be differentially expressed and modulates the largest number of functional pathways. However, the pathways are generally different from those altered in *Pa*^{ATCC43949}: the greatest number of genes are assigned to metabolic pathways, including cofactors and vitamins (three genes) and porphyrin and chlorophyll (three genes) amongst others. However, one orthologous gene, which unfortunately does not have a KEGG pathway assignment, is up-regulated in both *Pa*^{ATCC43949} and *Pa*^{Kingscliff} grown in LBNHS at 37^{o}C compared to LB: PAU_01648/PAK_1624, a putative tail fibre protein. Tail fibre genes originate from the integration of bacteriophage genomic material into the host bacterial genome, and these can be associated with acquisition of virulence (see Boyd & Brüssow, 2002). These findings may indicate that in these two strains that were isolated from disparate geographical locations, different gene sets have been acquired and/or selected for, permitting either distinct substrates to be utilised for respiration at 28^{o}C in insect hosts, subsequently co-opted for survival in mammalian systems, or permitting alternative survival pathways to be implemented at this higher temperature. However, it is possible to speculate that similarities in genetic responses, such as the

common up-regulation of a putative phage tail fibre protein, may represent ancestral acquired virulence factors maintained prior to sub-speciation.

2.4.2 Phenoarray analysis identifies specific substrates that mediate differences in Photorhabdus species respiration

Using *grofit's* default calculations, approximately 6% of wells were kept from the whole data set whereas when using the maximum confidence interval method (see Materials and Methods) approximately 81% of all data was kept. The *grofit* results summaries of statistically significant differences between mean replicate data for Pl^{TT01} and $Pa^{ATCC43949}$ per plate are summarised in Tables S1-6 and further condensed in Table 6. More substrates were found to give rise to significantly different respiration when $Pa^{ATCC43949}$ cultured at 28°C was compared to itself cultured at 37°C (19.6% of all wells) than when either of these were compared to Pl^{TT01} at 28°C (Table 6). These were found to result in overall lower growth (minimal, negative or sub-optimal) compared to both species at 28°C for all plates tested (see Tables S1-6). Of the categories of substrates included in the set of Biolog plates used, osmolytes, pH, and particular groups of peptide nitrogen sources elicited the greatest number of wells where significantly different growth was observed (Table 6). The results also allow identification of those substrates that promote $Pa^{ATCC43949}$ respiration at 37°C compared to itself or Pl^{TT01} at 28°C (positive versus negative respiration). There were only two instances of this observed: the di-peptides His-Trp and Gly-Asn, although positive respiration of $Pa^{ATCC43949}$ at 37°C was also observed on L-tyrosine and Thr-Glu. This gives weight to the idea that $Pa^{ATCC43949}$ adaptation to 37°C has been mediated by selection on specific and limited functional pathways and, in general, other pathways that normally facilitate growth at 28°C are negatively impacted at 37°C.

Comparing Pl^{TT01} (at 28°C only) to $Pa^{ATCC43949}$ at both 28°C and 37°C, there are no instances of Pl^{TT01} exhibiting positive growth where $Pa^{ATCC43949}$ exhibits negative growth, indicating that on these substrates tested, $Pa^{ATCC43949}$ is more prolific than Pl^{TT01}, particularly at 28°C. This may mean that key substrates for positive Pl^{TT01} growth have been omitted from the study, that Pl^{TT01} respiration is restricted to relatively few substrates, or that Pl^{TT01} respiration is markedly lower in general compared to $Pa^{ATCC43949}$. In either case, this indicates that $Pa^{ATCC43949}$ respires at a higher level on a wider range of common substrates than Pl^{TT01}, a characteristic that may have contributed to its ability to adapt to a new host environment. $Pa^{ATCC43949}$ at 28°C exhibits positive respiration where Pl^{TT01} is negative for N-acetyl-D-glucosamine adenosine, and the di-peptides Ala-Glu, Ala-Gln, and Leu-His. Functional pathways that metabolise these substrates may represent key differences in *P. luminescens* versus *P. asymbiotica* respiration leading to greater adaptive capabilities. One caveat of this data is that *Photorhabdus* growth is known to be variable and sensitive to growth conditions (Prof. N. Waterfield, personal communication) and the lattice plots

allow us to see that the negative and positive growth controls (well A01 in plates 3-8, and well A02 in plates 6-8, respectively) do not necessarily behave as desired, even with two replicates. The low number of replicates weakens the confidence that can be placed in this study's findings: three replicates should be used as a minimum – more are preferable given the erratic respiration patterns of these species. As such, these results are to be taken as indicative of possible substrates involved in respiration differences between *P. asymbiotica* and *P. luminescens*. Future expansion of this work should ideally include further replicates of these phenoarrays to ensure the highest possible level of consistency is attained.

2.4.3 A metabolic switch in glycine, serine and threonine metabolic pathways may underlie Photorhabdus adaptation to different host species

The phenoarrays show that, at 28°C, Pl^{TT01} exhibits minimal growth on L-serine whereas $Pa^{ATCC43949}$ exhibits positive growth (Table S1). Additionally, on the di-peptide Ala-Ser, the same pattern is observed (Table S2). This indicates a functional switch in metabolic pathways between these two species where serine-derived substrates promote $Pa^{ATCC43949}$ but not Pl^{TT01} respiration at 28°C. Exploring the KEGG pathway diagram for glycine, serine and threonine metabolism, it appears that this network may indeed represent the set of pathways where an adaptive switch has occurred between Pl^{TT01} and $Pa^{ATCC43949}$. At 28°C in LBHm compared to LB, the KEGG annotated RNA-seq results show that genes responsible for glycolysis/gluconeogenesis and pyruvate metabolism are up-regulated in both Pl^{TT01} and $Pa^{ATCC43949}$ (illustrated in Fig. 4). In contrast, the genes detected for glycine, serine and threonine metabolism are up-regulated in Pl^{TT01} and down-regulated in $Pa^{ATCC43949}$, indicating a phenotypic switch. This is accompanied in Pl^{TT01} by an up-regulation of valine, leucine and isoleucine biosynthesis, down-regulation of arginine and proline metabolism and glyoxylate metabolism, and in $Pa^{ATCC43949}$, an up-regulation of methane and sulphur metabolism. For $Pa^{ATCC43949}$ at 37°C versus 28°C, notable changes also occur within this network: the genes involved in the citrate cycle are down-regulated, and porphyrin and chlorophyll metabolism is also down-regulated (porphyrin biosynthesis is included in the glycine, serine and threonine metabolism network).

The KEGG annotations of the differentially expressed gene data also indicate a difference in amino acid usage between Pl^{TT01} and $Pa^{ATCC43949}$ at 28°C in LBHm compared to LB: Pl^{TT01} showed two genes to up-regulated for tyrosine metabolism when in LBHm whereas $Pa^{ATCC43949}$ showed two genes up-regulated for histidine metabolism in LBHm. The phenoarray data gave moderate support for this also: there were five instances where Pl^{TT01} respiration on several histidine-containing wells was negative whereas it was minimal or positive for $Pa^{ATCC43949}$ – there were only two occurrences where this was reversed, and one occurrence where respiration level was the same. There was one

instance where *Pl*^{TT01} respiration on tyrosine-containing wells was minimal or positive and *Pa*^{ATCC43949} respiration was negative, and two instances where the respiration level was the same.

2.5 Discussion

This study has utilised the integrative state-of-the-art high throughput techniques of RNA-seq and phenoarray to characterise the genomic and phenotypic effects of *Photorhabdus* species growth under various conditions. This has permitted exploration of the environmental factors that may underlie cellular changes facilitating adaptation of *P. asymbiotica* strains, such as $Pa^{ATCC43949}$, to human hosts as well as insects, compared to insect-restricted strains such as Pl^{TT01}. We found that bacterial growth medium impacted on gene expression to a greater extent than temperature, and that growth of $Pa^{ATCC43949}$ at human temperature (37^{o}C) resulted in a notable drop in growth on a wide range of substrates. These findings indicate that adaptation to human hosts is focused on functional metabolic pathway alterations in response to environmental substrates rather than an enhanced ability of typical pathways to operate at higher temperatures, and that growth at higher temperatures in fact results in sub-optimal growth, most likely due to de-regulation of key survival pathways.

'Omics technologies have greatly expanded and advanced in recent years, mainly by virtue of leaps in the technology available for molecular sequencing, the digital capture of data, and the development of robust computational tools for analysing the vastly increasing volume of data that these technologies generate (Berger et al., 2013). The techniques used herein, RNA-seq and phenoarray, generate accurate, high throughout data (Bochner et al., 2001; Wang et al., 2009) that can be integrated to enhance the granularity and confidence in the results from each data stream, overcoming their respective inherent limitations (Ge et al., 2003). Indeed, several recent studies have done just that, illustrating the power of combining phenoarray with sequence data (Schuller *et al.* 2004; Pietiäinen *et al.* 2009; Yan *et al.* 2013; Minato *et al.* 2014). Here, we are able to show that components of specific functional genomic pathways are differentially expressed in response to temperature and growth medium conditions and that there are phenotypic effects relating to specific substrates consistent with differential modulation of functional metabolic networks.

An initial hypothesis of this study was that *P. asymbiotica* had adapted to survival at the higher temperatures of mammalian systems, allowing it to colonise human soft tissue when introduced via its nematode symbiont. However, our results do not necessarily support this as it was observed that growth medium impacted the most on gene expression, rather than temperature. The combined findings from the RNA-seq and phenoarray data indicate that functional metabolic differences between *P. asymbiotica* and *P. luminescens* may have manifested within the pathway network related to glycine, serine and threonine metabolism: $Pa^{ATCC43949}$ and Pl^{TT01} exhibited striking similarities and differences when grown on LBHm compared to LB at 28^{o}C. In both species, genes involved in glycolysis/ gluconeogenesis and pyruvate metabolism were up-regulated, whereas there appeared to be a striking switch in the activity of glycine, serine and threonine metabolism. Two

genes were found to be significantly differentially expressed in both species, both genes being up-regulated in *Pl*^{TT01} and both down-regulated in *Pa*^{ATCC43949}. Unfortunately, these genes were not included in the gene orthology data currently available so we are not able to speculate whether these represent orthologs. *Pa*^{ATCC43949} additionally shows four genes down-regulated in the citrate cycle. This indicates that the glycine, serine and threonine metabolism pathway may integrate differential amino acid usage by each species via glycolysis/ gluconeogenesis. Concurrently, the phenoarray data demonstrated a significant difference in respiration on L-serine: *Pl*^{TT01} respiration was minimal whereas *Pa*^{ATCC43949} respiration was positive. Although no significant effects were observed in the differential expression analysis for genes within these pathways when *Pa*^{ATCC43949} was cultivated at 37^{o}C, the phenoarray data indicates that the function of these pathways are adequately operational at this higher temperature: respiration was positive for *Pa*^{ATCC43949} at both 28^{o}C and 37^{o}C on the dipeptides Thr-Glu and Gly-Asn. *Pl*^{TT01} respiration was significantly lower on Gly-Asn at 28^{o}C than *Pa*^{ATCC43949} at 37^{o}C: given that there were relatively few occurrences of positive versus negative respiration, the differential utilisation of these substrates as nitrogen sources may underlie core differences in metabolic preferences of these bacteria, manifesting in their respective host restrictions.

Recently, it has been found that insect-restricted *Pl*^{TT01} cells can rapidly be selected to survive at mammalian temperatures and that approximately 1x10^{-6} cells within a normal population will have this ability (other cells die at around 34^{o}C, Prof. N. Waterfield, personal communication). This indicates that random mutation or a shift in gene regulation conferring the ability to survive at higher temperatures is a regular naturally occurring phenomenon in *Photorhabdus*, perhaps aiding survival in variable environmental conditions, as has been noted for other environmentally-derived pathogens (Cooney & Klein 2008). Fixation of this ability in the *P. asymbiotica* communities from clinical isolates is likely to have arisen following the acquisition of resistance to mammalian immunity. It has recently been found in *P. asymbiotica* that flagella biosynthesis transcription is switched on when bacteria are grown at 37^{o}C in LB medium compared to 28^{o}C, but that this returns to levels seen during growth at 28^{o}C when human serum is added. Flagellin proteins strongly stimulate host immunity. Our findings here partially support this observation: in *Pa*^{ATCC43949}, we observed that a flagella assembly gene is down-regulated in LBNHS at 37^{o}C compared to LB at the same temperature, during exponential growth. Conversely, we found that *Pa*^{Kingscliff} exhibited significant up-regulation of a gene involved in flagella assembly (different to the gene detected in *Pa*^{ATCC43949}) during stationary growth at 37^{o}C in LBNHS compared to growth at 37^{o}C in LB. Given the immunogenicity of flagellin proteins, this differential activity may contribute to differences in the pattern of pathogenesis between these two strains that may be exploited during clinical assessment and treatment of mammalian infections. Further recent RNA-seq analyses of *Pa*^{ATCC43949} and *Pa*^{Kingscliff} indicate that these strains differ significantly in their genomic response to growth at 37^{o}C (Prof. N. Waterfield, personal communication), presenting interesting avenues for further exploration of the comparative genomic differences within the *P. asymbiotica* species that

may underlie differences in the acquisition of virulence in mammalian hosts in response to environment-dependent selection pressures.

A limitation to inferences drawn from these results regarding the molecular basis of the phenotypic switch that has occurred to allow *P. asymbiotica* species to infect humans as well as insects is the omission of tests of both *P. luminescens* and *P. asymbiotica* growth in exponential and stationary phases in LBHm at 37^{o}C. Although this is not necessarily a biological situation that would be encountered in nature, it would allow disentanglement of the genes that are specifically involved in growth at higher temperatures, rather than at high temperatures and in human serum, as tested here. One of the main challenges with high throughput data is in the application of rigorous statistical methods to disentangle the significant effects. For example, when conducting differential gene expression, independent biological replicates are required to robustly assess natural variation in gene expression prior to classing pairs of genes as differentially expressed (Anders 2012). Given that no replicates of the RNA-seq data were performed, all findings from these data sets may only be considered as preliminary and further replicates should be performed if these findings are to be confirmed. It is our understanding that these replicates are currently in process. Additionally, quantitative reverse transcription polymerase chain reaction (qRT-PCR) should ideally be used to verify key gene expression differences.

2.6 Conclusions

RNA-seq integrated with phenoarray provides a good methodology for exploring the phenotypic effects concurrent with gene expression changes, from which to move forward into more direct molecular biological approaches to dissect the mechanistic basis of these differences. *Photorhabdus* provide a highly interesting model of pathogenic diversity within a complex system of symbiosis. Their highly variable growth tendencies, rapid evolvability, and clear pathogenic differences make them a challenging subject, but the use of high throughput methods has allowed us to shed light on the key differences between insect-restricted *P. luminescens* and the insect and human pathogen *P. asymbiotica*. We show that there are clear demonstrable differences in the genes being expressed and substrates being utilised for respiration in various host-representative conditions: these differences appear to mostly be stimulated by differing substrates rather than temperature, as first thought. Within-species differences observed between *Pa*$^{\text{ATCC43949}}$ and *Pa*$^{\text{Kingscliff}}$, isolated from the USA and Australia, respectively, indicate that distinct substrate utilisation can occur. The impact this may have on the clinical presentation of infection is unknown and may represent an area worth exploring to aid effective treatment.

Moving forward, further independent replicates of RNA-seq and phenoarray data would be obtained to add greater certainty to the statistically significant differences observable under these conditions. This would additionally permit the construction of gene co-expression networks which would allow a deeper understanding of the gene modules recruited or suppressed in response to environmental cues. A further condition testing growth of *P. luminescens* and *P. asymbiotica* in LBHm at 37^{o}C would be conducted to identify those genes modulated in response to temperature versus medium, and the substrates which can be utilised in each. Following this, qRT-PCR would be used to confirm gene expression differences. Knock-down and knock-in experiments could then be performed to explore the specific activities of particular genes and gene networks: by silencing genes identified in the differential expression analyses as up-regulated in response to a particular condition, the dependence of survival on that gene, or its modulation of a given phenotype, could be assessed. By rescuing the expression of knocked out genes, the dependence of the phenotype on that gene could be confirmed.

2.7 Figure legends

Fig. 1. Phylogeny of *Photorhabdus* sp. with additional details: clinical isolate location, typical host and temperatures at which they are known to survive (adapted from a figure kindly provided by Prof. N. Waterfield, not to scale). The branch point at which a phenotypic switch may have occurred to enable survival within mammalian tissues is indicated (red circle). Strains employed in this study are highlighted in light blue.

Fig. 2. Analysis pipeline for phenoarray data, using custom Python scripts, spreadsheets, incorporating the R packages, *lattice* and *grofit*.

Fig. 3. Graphical representation of bacterial respiration curve phases and parameters, as estimated by the R package *grofit*. Bacterial respiration curves can be subdivided into a series of phases, including: the lag phase, λ, before respiration begins; the respiration rate phase, μ, which corresponds to the slope of the curve; the maximum respiration, A, which corresponds to the maximum value recorded within the curve, and a derivative of all these being the area under the curve, AUC.

Fig. 4. Barplots of numbers of genes significantly differentially expressed per comparison per strain: (A) *Pl*^{TT01}, (B) *Pa*^{ATCC43949}, (C) *Pa*^{Kingscliff}. These plots demonstrate that changes in growth medium generally result in the greatest impact on gene expression, rather than changes in temperature or growth phase. LB28ex/st: LB broth at 28^{o}C during exponential/stationary growth phase; LBHm28ex/st: LB broth supplemented with insect haemolymph at 28^{o}C during exponential/stationary growth phase; LB37ex/st: LB broth at 37^{o}C during exponential/stationary growth phase; LBNHS37ex/st: LB broth supplemented with normal human serum at 37^{o}C during exponential/stationary growth phase.

Fig. 5. The glycine, serine and threonine metabolic pathway appears to integrate specific similarities and differences between *Pa*^{ATCC43949} (red) and *Pl*^{TT01} (blue) when growth at 28^{o}C in LBHm is compared to growth in LB. Differences between *Pa*^{ATCC43949} grown at 28^{o}C and 37^{o}C also manifest within this network and are shown in green. The presence of boxes alone indicates a path is up-regulated, whereas boxes plus arrows indicate down-regulation. Adapted from the KEGG database.

Fig. 1

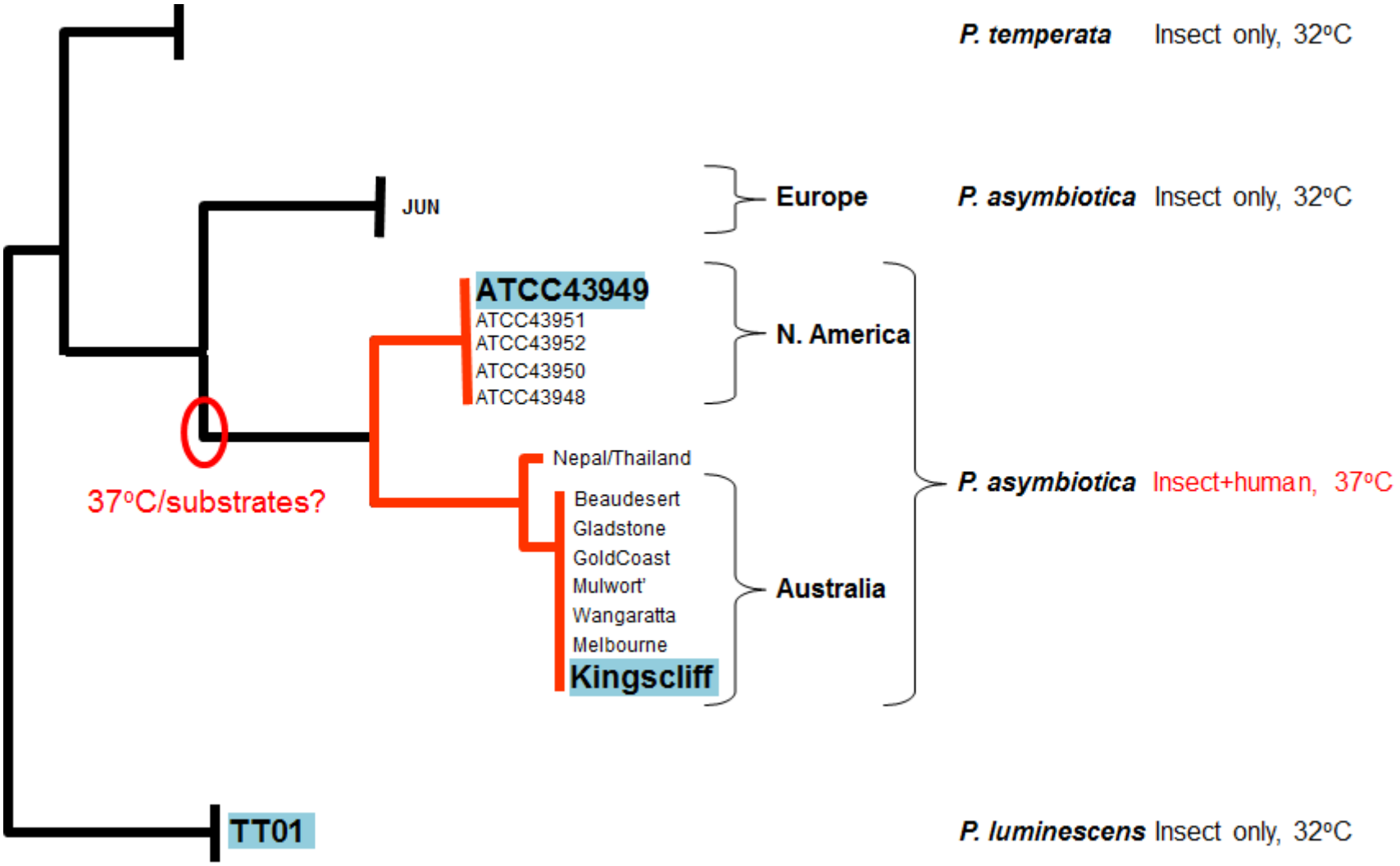

P. temperata Insect only, 32°C
JUN
Europe
P. asymbiotica Insect only, 32°C
ATCC43949
ATCC43951
ATCC43952
ATCC43950
ATCC43948
N. America
37°C/substrates?
Nepal/Thailand
Beaudesert
Gladstone
GoldCoast
Mulwort'
Wangaratta
Melbourne
Kingscliff
Australia
P. asymbiotica Insect+human, 37°C
TT01
P. luminescens Insect only, 32°C

Fig. 2

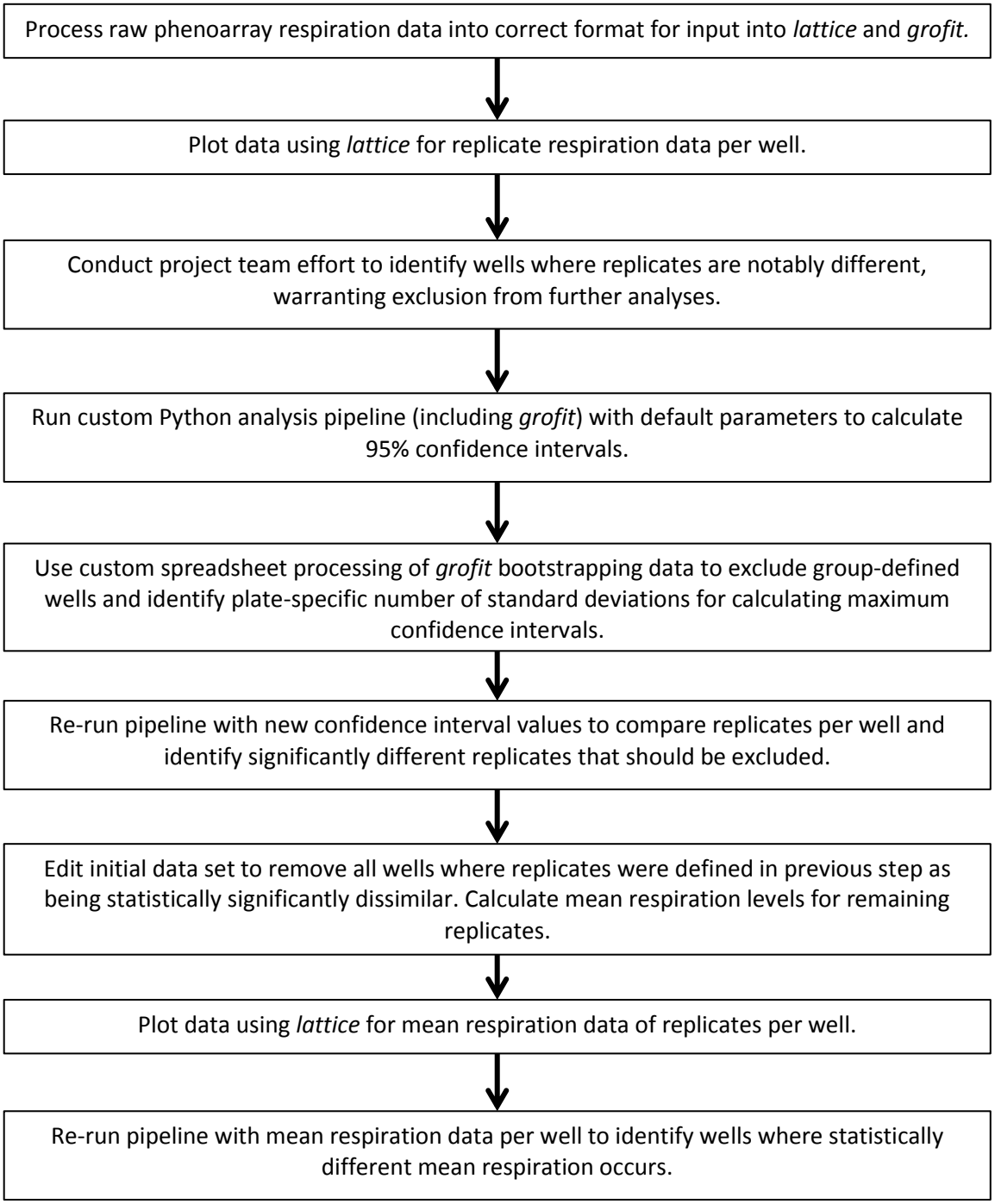

Process raw phenoarray respiration data into correct format for input into *lattice* and *grofit.*
Plot data using *lattice* for replicate respiration data per well.
Conduct project team effort to identify wells where replicates are notably different, warranting exclusion from further analyses.
Run custom Python analysis pipeline (including *grofit*) with default parameters to calculate 95% confidence intervals.
Use custom spreadsheet processing of *grofit* bootstrapping data to exclude group-defined wells and identify plate-specific number of standard deviations for calculating maximum confidence intervals.
Re-run pipeline with new confidence interval values to compare replicates per well and identify significantly different replicates that should be excluded.
Edit initial data set to remove all wells where replicates were defined in previous step as being statistically significantly dissimilar. Calculate mean respiration levels for remaining replicates.
Plot data using *lattice* for mean respiration data of replicates per well.
Re-run pipeline with mean respiration data per well to identify wells where statistically different mean respiration occurs.

Fig. 3

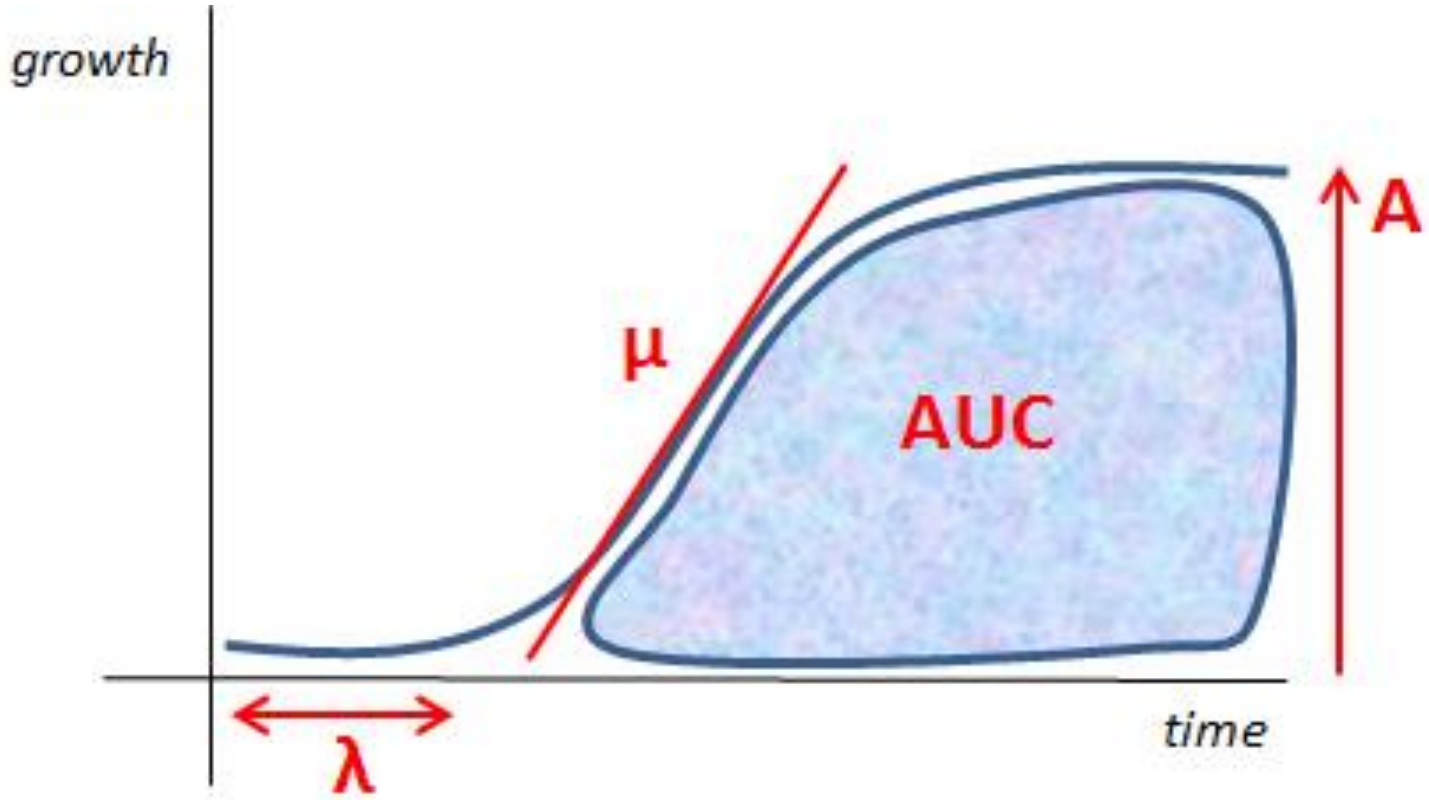


Fig. 4

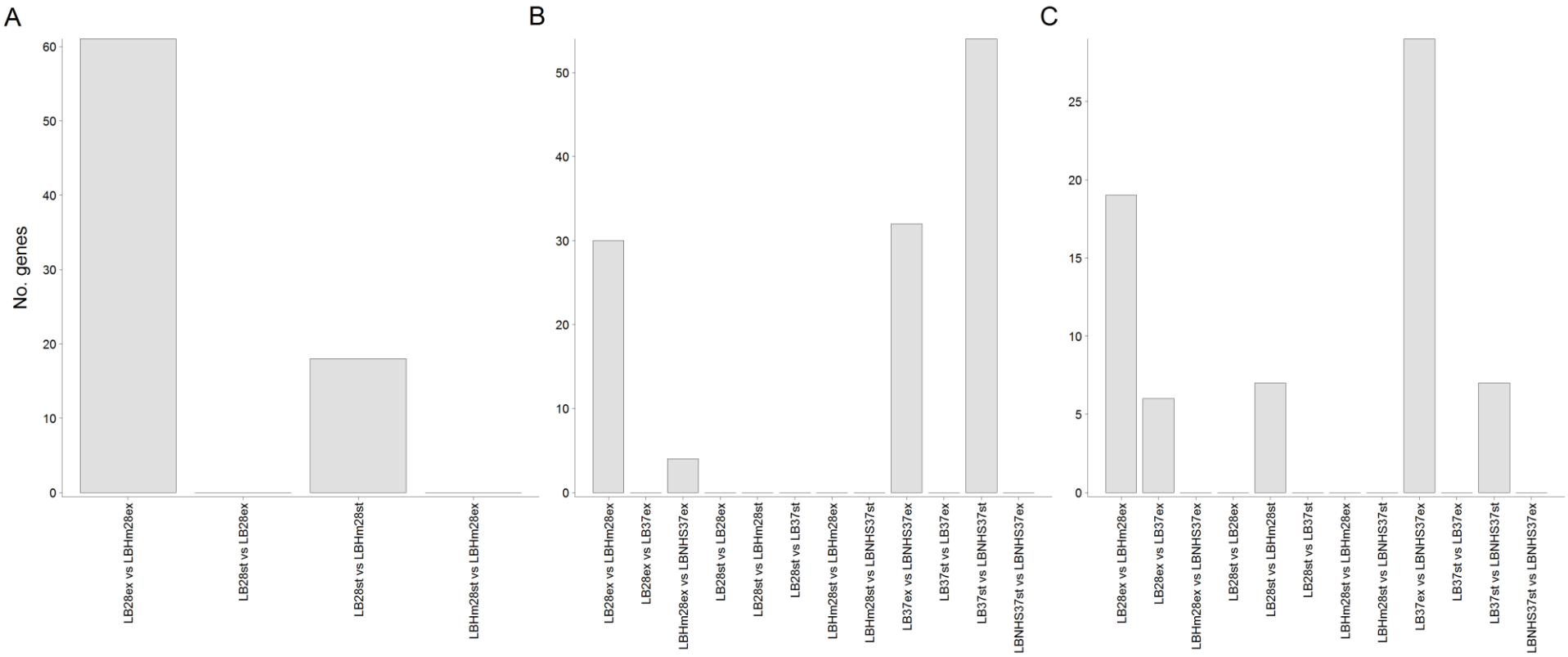

A
B
C
No. genes
LB28ex vs LBHm28ex
LB28st vs LB28ex
LB28st vs LBHm28st
LBHm28st vs LBHm28ex
LB28ex vs LB37ex
LBHm28ex vs LBNHS37ex
LB28st vs LB28ex
LB28st vs LBHm28st
LB28st vs LB37st
LBHm28st vs LBHm28ex
LBHm28st vs LBNHS37st
LB37ex vs LBNHS37ex
LB37st vs LB37ex
LB37st vs LBNHS37st
LBNHS37st vs LBNHS37ex

Fig. 5

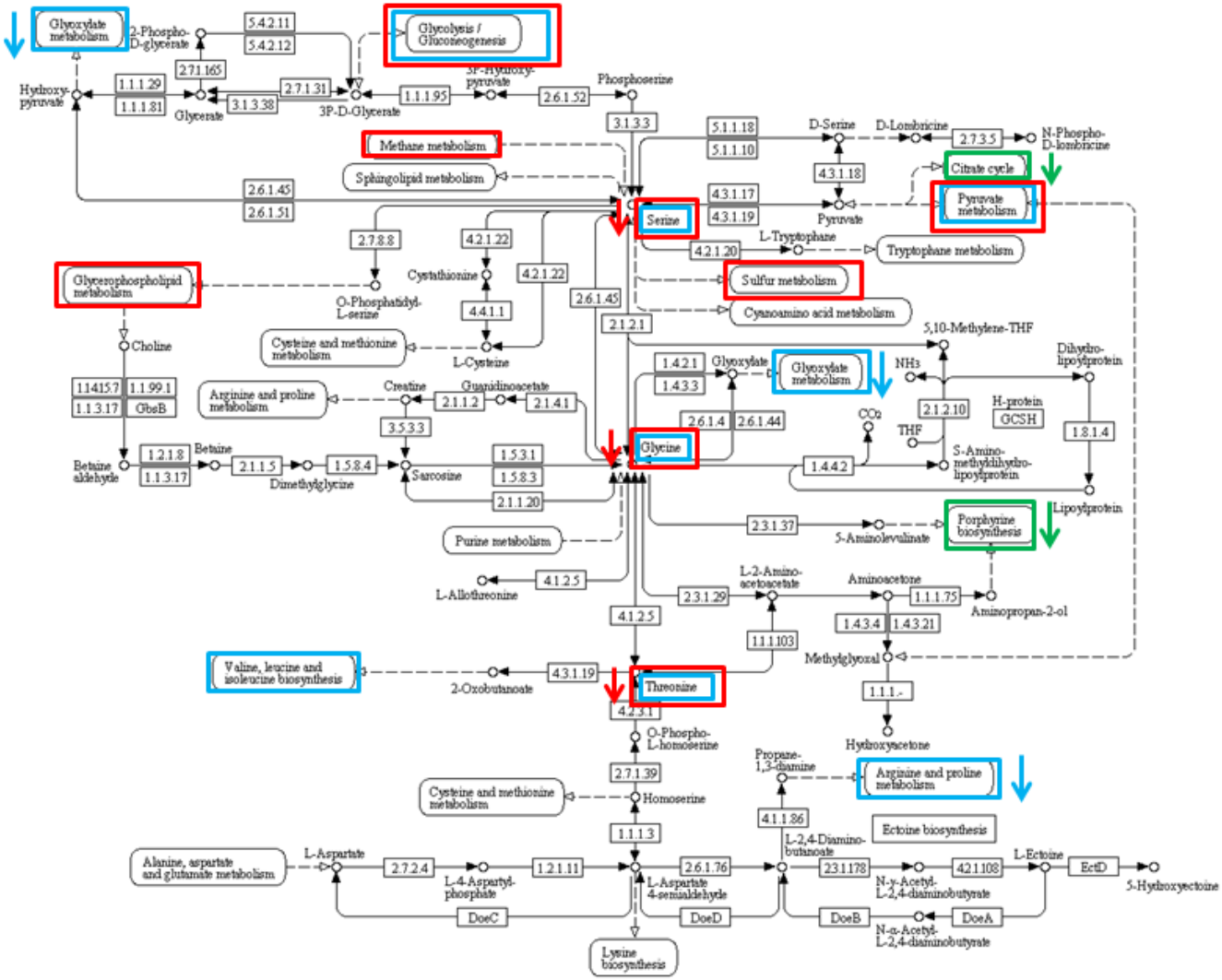

Glyoxylate metabolism
2-Phospho-D-glycerate
5.4.2.11
5.4.2.12
Glycolysis / Gluconeogenesis
2.7.1.165
Hydroxy-pyruvate
1.1.1.29
1.1.1.81
Glycerate
3.1.3.38
2.7.1.31
3P-D-Glycerate
1.1.1.95
3P-Hydroxy-pyruvate
2.6.1.52
Phosphoserine
3.1.3.3
Methane metabolism
Sphingolipid metabolism
2.6.1.45
2.6.1.51
Serine
5.1.1.18
5.1.1.10
D-Serine
D-Lombricine
2.7.3.5
N-Phospho-D-lombricine
4.3.1.18
Citrate cycle
4.3.1.17
4.3.1.19
Pyruvate
Pyruvate metabolism
L-Tryptophane
4.2.1.20
Tryptophane metabolism
2.7.8.8
4.2.1.22
Cystathionine
Glycerophospholipid metabolism
O-Phosphatidyl-L-serine
4.4.1.1
Sulfur metabolism
Cyanoamino acid metabolism
2.1.2.1
5,10-Methylene-THF
Choline
Cysteine and methionine metabolism
L-Cysteine
1.4.2.1
1.4.3.3
Glyoxylate
Glyoxylate metabolism
NH3
Dihydro-lipoylprotein
1.1415.7
1.1.99.1
1.1.3.17
GbeB
Arginine and proline metabolism
Creatine
Guanidinoacetate
2.1.1.2
2.1.4.1
3.5.3.3
2.6.1.4
2.6.1.44
CO2
2.1.2.10
H-protein
GCSH
THF
1.8.1.4
Betaine
Betaine aldehyde
1.2.1.8
1.1.3.17
2.1.1.5
1.5.8.4
Dimethylglycine
Sarcosine
1.5.3.1
1.5.8.3
2.1.1.20
Glycine
1.4.4.2
S-Amino-methyldihydro-lipoylprotein
Porphyrine biosynthesis
Lipoylprotein
2.3.1.37
5-Aminolevulinate
Purine metabolism
L-Allothreonine
4.1.2.5
L-2-Amino-acetoacetate
2.3.1.29
Aminoacetone
1.1.1.75
Aminopropan-2-ol
1.4.3.4
1.4.3.21
11.1.103
Methylglyoxal
Valine, leucine and isoleucine biosynthesis
2-Oxobutanoate
4.3.1.19
Threonine
4.2.3.1
1.1.1.-
Hydroxyacetone
O-Phospho-L-homoserine
2.7.1.39
Propane-1,3-diamine
Arginine and proline metabolism
Cysteine and methionine metabolism
Homoserine
1.1.1.3
4.1.1.86
Ectoine biosynthesis
L-2,4-Diamino-butanoate
Alanine, aspartate and glutamate metabolism
L-Aspartate
2.7.2.4
L-4-Aspartyl-phosphate
1.2.1.11
2.6.1.76
L-Aspartate 4-semialdehyde
2.3.1.178
N-γ-Acetyl-L-2,4-diaminobutyrate
4.2.1.108
L-Ectoine
EctD
5-Hydroxyectoine
DoeC
DoeD
DoeB
N-α-Acetyl-L-2,4-diaminobutyrate
DoeA
Lysine biosynthesis

Table 1: Numbers of genes detected by each growth condition for each species.

Strain	Growth media	Temperature (^{o}C)	Growth phase	Number of genes detected (proportion of total features)
Pl^{TT01}	LB	28	Exponential	4796 (92.5%)
Pl^{TT01}	LBHm	28	Exponential	4739 (91.4%)
Pl^{TT01}	LB	28	Stationary	4767 (91.9%)
Pl^{TT01}	LBHm	28	Stationary	4779 (92.1%)
Pa^{ATCC43949}	LB	28	Exponential	4422 (94.9%)
Pa^{ATCC43949}	LBHm	28	Exponential	4403 (94.4%)
Pa^{ATCC43949}	LB	28	Stationary	4420 (94.8%)
Pa^{ATCC43949}	LBHm	28	Stationary	4440 (95.2%)
Pa^{ATCC43949}	LB	37	Exponential	4407 (94.5%)
Pa^{ATCC43949}	LBNHS	37	Exponential	4421 (94.8%)
Pa^{ATCC43949}	LB	37	Stationary	4395 (94.3%)
Pa^{ATCC43949}	LBNHS	37	Stationary	4428 (95.0%)
Pa^{ATCC43949} pPAU1	LB	28	Exponential	28 (96.6%)
Pa^{ATCC43949} pPAU1	LBHm	28	Exponential	28 (96.6%)
Pa^{ATCC43949} pPAU1	LB	28	Stationary	28 (96.6%)
Pa^{ATCC43949} pPAU1	LBHm	28	Stationary	28 (96.6%)
Pa^{ATCC43949} pPAU1	LB	37	Exponential	28 (96.6%)
Pa^{ATCC43949} pPAU1	LBNHS	37	Exponential	28 (96.6%)
Pa^{ATCC43949} pPAU1	LB	37	Stationary	28 (96.6%)
Pa^{ATCC43949} pPAU1	LBNHS	37	Stationary	28 (96.6%)
Pa^{Kingscliff}	LB	28	Exponential	4839 (94.3%)
Pa^{Kingscliff}	LBHm	28	Exponential	4669 (91.0%)
Pa^{Kingscliff}	LB	28	Stationary	4821 (93.9%)
Pa^{Kingscliff}	LBHm	28	Stationary	4809 (93.7%)
Pa^{Kingscliff}	LB	37	Exponential	4767 (92.9%)
Pa^{Kingscliff}	LBNHS	37	Exponential	4818 (93.9%)
Pa^{Kingscliff}	LB	37	Stationary	4867 (94.8%)
Pa^{Kingscliff}	LBNHS	37	Stationary	4712 (91.8%)

Table 2: Differential gene expression results.

Source	Comparison	No. significant DE genes (p<=0.05)
Pl^{TT01}	LB 28°C Expo vs. LBHm 28°C Expo	61
	LB 28°C Stat vs. LB 28°C Exp	0
	LB 28°C Stat vs. LBHm 28°C Stat	18
	LBHm 28°C Stat vs. LBHm 28°C Expo	0
Pa^{ATCC43949}	LB 28°C Expo vs. LBHm 28°C Expo	30
	LB 28°C Expo vs. LB 37°C Expo	0
	LBHm 28°C Expo vs. LBNHS 37°C Expo	4
	LB 28°C Stat vs. LB 28°C Expo	0
	LB 28°C Stat vs. LBHm 28°C Stat	0
	LB 28°C Stat vs. LB 37°C Stat	0
	LBHm 28°C Stat vs. LBHm 28°C Expo	0
	LBHm 28°C Stat vs. LBNHS 37°C Stat	0
	LB 37°C Expo vs. LBNHS 37°C Expo	32
	LB 37°C Stat vs. LB 37°C Expo	0
	LB 37°C Stat vs. LBNHS 37°C Stat	54
	LBNHS 37°C Stat vs. LBNHS 37°C Expo	0
Pa^{ATCC43949} plasmid	LB 28°C Expo vs. LBHm 28°C Expo	0
	LB 28°C Expo vs. LB 37°C Expo	0
	LBHm 28°C Expo vs. LBNHS 37°C Expo	0
	LB 28°C Stat vs. LB 28°C Expo	0
	LB 28°C Stat vs. LBHm 28°C Stat	0
	LB 28°C Stat vs. LB 37°C Stat	0
	LBHm 28°C Stat vs. LBHm 28°C Expo	0
	LBHm 28°C Stat vs. LBNHS 37°C Stat	0
	LB 37°C Expo vs. LBNHS 37°C Expo	0
	LB 37°C Stat vs. LB 37°C Expo	0
	LB 37°C Stat vs. LBNHS 37°C Stat	0

	LBNHS 37^{o}C Stat vs. LBNHS 37^{o}C Expo	0
Pa^{Kingscliff}	LB 28^{o}C Expo vs. LBHm 28^{o}C Expo	19
	LB 28^{o}C Expo vs. LB 37^{o}C Expo	6
	LBHm 28^{o}C Expo vs. LBNHS 37^{o}C Expo	0
	LB 28^{o}C Stat vs. LB 28^{o}C Expo	0
	LB 28^{o}C Stat vs. LBHm 28^{o}C Stat	7
	LB 28^{o}C Stat vs. LB 37^{o}C Stat	0
	LBHm 28^{o}C Stat vs. LBHm 28^{o}C Expo	0
	LBHm 28^{o}C Stat vs. LBNHS 37^{o}C Stat	0
	LB 37^{o}C Expo vs. LBNHS 37^{o}C Expo	29
	LB 37^{o}C Stat vs. LB 37^{o}C Expo	0
	LB 37^{o}C Stat vs. LBNHS 37^{o}C Stat	7
	LBNHS 37^{o}C Stat vs. LBNHS 37^{o}C Expo	0

Table 3: Genes differentially expressed per KEGG pathway – Pl^{TT01}.

KEGG pathway	LB 28°C Expo vs. LBHm 28°C Expo	LB 28°C Stat vs. LB 28°C Exp	LB 28°C Stat vs. LBHm 28°C Stat	LBHm 28°C Stat vs. LBHm 28°C Expo
ABC transporters	1			
Arginine and proline metabolism	1			
Bacterial secretion system	1			
Butanoate metabolism				
Fatty acid biosynthesis	1		1	
Fructose and mannose metabolism			1	
Glycine serine and threonine metabolism	2			
Glycolysis / Gluconeogenesis	3			
Glyoxylate and dicarboxylate metabolism	1			
Inositol phosphate metabolism	1			
Oxidative phosphorylation	1			
Phenylalanine metabolism				

Pyruvate metabolism	2			
Two-component system	1			
Tyrosine metabolism	2			
Valine leucine and isoleucine biosynthesis	1			
Total genes	61	0	18	0

Table 4: Genes differentially expressed per KEGG pathway – $Pa^{ATCC43949}$.

KEGG pathway	LB 28°C Expo vs. LBHm 28°C Expo	LB 28°C Expo vs. LB 37°C Expo	LBHm 28°C Expo vs. LBNHS 37°C Expo	LB 28°C Stat vs. LB 28°C Expo
ABC transporters	1			
Bacterial secretion system				
Benzoate degradation	1			
Biotin metabolism				
Citrate cycle (TCA cycle)				
Fatty acid metabolism				
Flagellar assembly				
Glycerophospholipid metabolism	1			
Glycine serine and threonine metabolism	2			
Glycolysis / Gluconeogenesis	2			
Histidine metabolism	2			
Methane metabolism	1			
Porphyrin and chlorophyll metabolism				
Propanoate metabolism	1			

Purine metabolism				
Pyruvate metabolism	2			
Sulfur metabolism	1			
Two-component system				
Ribosome				
Total genes	30	0	4	0

Table 4: cont.

KEGG pathway	LB 28°C Stat vs. LBHm 28°C Stat	LB 28°C Stat vs. LB 37°C Stat	LBHm 28°C Stat vs. LBHm 28°C Expo	LBHm 28°C Stat vs. LBNHS 37°C Stat
ABC transporters				
Bacterial secretion system				
Benzoate degradation				
Biotin metabolism				
Citrate cycle (TCA cycle)				
Fatty acid metabolism				
Flagellar assembly				
Glycerophospholipid metabolism				
Glycine serine and threonine metabolism				
Glycolysis / Gluconeogenesis				
Histidine metabolism				
Methane metabolism				
Porphyrin and chlorophyll metabolism				
Propanoate metabolism				

Purine metabolism				
Pyruvate metabolism				
Sulfur metabolism				
Two-component system				
Ribosome				
Total genes	0	0	0	0

Table 4: cont.

KEGG pathway	LB 37°C Expo vs. LBNHS 37°C Expo	LB 37°C Stat vs. LB 37°C Expo	LB 37°C Stat vs. LBNHS 37°C Stat	LBNHS 37°C Stat vs. LBNHS 37°C Expo
ABC transporters			1	
Bacterial secretion system			1	
Benzoate degradation				
Biotin metabolism	1			
Citrate cycle (TCA cycle)	4			
Fatty acid metabolism				
Flagellar assembly	1			
Glycerophospholipid metabolism				
Glycine serine and threonine metabolism				
Glycolysis / Gluconeogenesis				
Histidine metabolism				
Methane metabolism				
Porphyrin and chlorophyll metabolism	5			
Propanoate metabolism				

Purine metabolism				
Pyruvate metabolism				
Sulfur metabolism				
Two-component system				
Ribosome			2	
Total genes	32	0	54	0

Table 5: Genes differentially expressed per KEGG pathway – *Pa*^{Kingscliff}.

KEGG pathway	LB 28°C Expo vs. LBHm 28°C Expo	LB 28°C Expo vs. LB 37°C Expo	LBHm 28°C Expo vs. LBNHS 37°C Expo	LB 28°C Stat vs. LB 28°C Expo
ABC transporters	1			
Fatty acid metabolism				
Flagellar assembly				
Glycerophospholipid metabolism	1			
Glycine serine and threonine metabolism	1			
Methane metabolism	1			
Nitrogen metabolism				
Polyketide sugar unit biosynthesis	1			
Porphyrin and chlorophyll metabolism				
Purine metabolism				
Pyruvate metabolism				
Two-component system	2			
Total genes	19	6	0	0

Table 5: cont.

KEGG pathway	LB 28°C Stat vs. LBHm 28°C Stat	LB 28°C Stat vs. LB 37°C Stat	LBHm 28°C Stat vs. LBHm 28°C Expo	LBHm 28°C Stat vs. LBNHS 37°C Stat
ABC transporters				
Fatty acid metabolism				
Flagellar assembly				
Glycerophospholipid metabolism				
Glycine serine and threonine metabolism				
Methane metabolism				
Nitrogen metabolism				
Polyketide sugar unit biosynthesis				
Porphyrin and chlorophyll metabolism				
Purine metabolism				
Pyruvate metabolism				
Two-component system				
Total genes	7	0	0	0

Table 5: cont.

KEGG pathway	LB 37°C Expo vs. LBNHS 37°C Expo	LB 37°C Stat vs. LB 37°C Expo	LB 37°C Stat vs. LBNHS 37°C Stat	LBNHS 37°C Stat vs. LBNHS 37°C Expo
ABC transporters	3			
Fatty acid metabolism	1			
Flagellar assembly			1	
Glycerophospholipid metabolism				
Glycine serine and threonine metabolism				
Methane metabolism				
Nitrogen metabolism	1			
Polyketide sugar unit biosynthesis				
Porphyrin and chlorophyll metabolism	3			
Purine metabolism	1			
Pyruvate metabolism	1			
Two-component system	2			
Total genes	29	0	7	0

Table 6: Condensed phenoarray results of mean replicate data: numbers of wells found to be significantly different per comparison, using the maximum confidence intervals method.

Plate substrates	*Pa*^{ATCC43949} 28°C vs. *Pa*^{ATCC43949} 37°C	*Pa*^{ATCC43949} 28°C vs. *Pl*^{TT01} 28°C	*Pa*^{ATCC43949} 37°C vs. *Pl*^{TT01} 28°C	**Total per plate**
Nitrogen sources	7	11	4	**22**
Peptide nitrogen sources	21	18	4	**43**
Peptide nitrogen sources (further)	11	8	4	**23**
Peptide nitrogen sources (further)	4	4	1	**9**
Osmolytes	30	14	29	**73**
pH	40	14	34	**88**
Total per comparison	**113 (19.6%)**	**69 (12.0%)**	**76 (13.2%)**	

2.8 Supplementary information

2.8.1 Supplementary figure legends

Fig. S1. Biolog Phenotype Microarray (phenoarray) plate information. The phenoarray plates are set out as 96 well plates with different substrates adhered to the bottom of the wells. The plates shown are those used in this study. This information is also available at: http://www.biolog.com/products/?product=Phenotype%20MicroArrays%20for%20Microbial%20Cells&view=Product%20Literature.

Fig. S2. Lattice plots of mean replicate data for Pl^{TT01} and *Pa*^{ATCC43949} respiration on the phenoarray plates shown in Fig. S1. The x axis is time and the y axis is respiration level.

Fig. S1.

PM1 MicroPlate™ Carbon Sources

A1 Negative Control	A2 L-Arabinose	A3 N-Acetyl-D-Glucosamine	A4 D-Saccharic Acid	A5 Succinic Acid	A6 D-Galactose	A7 L-Aspartic Acid	A8 L-Proline	A9 D-Alanine	A10 D-Trehalose	A11 D-Mannose	A12 Dulcitol
B1 D-Serine	B2 D-Sorbitol	B3 Glycerol	B4 L-Fucose	B5 D-Glucuronic Acid	B6 D-Gluconic Acid	B7 D,L-α-Glycerol-Phosphate	B8 D-Xylose	B9 L-Lactic Acid	B10 Formic Acid	B11 D-Mannitol	B12 L-Glutamic Acid
C1 D-Glucose-6-Phosphate	C2 D-Galactonic Acid-γ-Lactone	C3 D,L-Malic Acid	C4 D-Ribose	C5 Tween 20	C6 L-Rhamnose	C7 D-Fructose	C8 Acetic Acid	C9 α-D-Glucose	C10 Maltose	C11 D-Melibiose	C12 Thymidine
D-1 L-Asparagine	D2 D-Aspartic Acid	D3 D-Glucosaminic Acid	D4 1,2-Propanediol	D5 Tween 40	D6 α-Keto-Glutaric Acid	D7 α-Keto-Butyric Acid	D8 α-Methyl-D-Galactoside	D9 α-D-Lactose	D10 Lactulose	D11 Sucrose	D12 Uridine
E1 L-Glutamine	E2 M-Tartaric Acid	E3 D-Glucose-1-Phosphate	E4 D-Fructose-6-Phosphate	E5 Tween 80	E6 α-Hydroxy Glutaric Acid-γ-Lactone	E7 α-Hydroxy Butyric Acid	E8 β-Methyl-D-Glucoside	E9 Adonitol	E10 Maltotriose	E11 2-Deoxy Adenosine	E12 Adenosine
F1 Glycyl-L-Aspartic Acid	F2 Citric Acid	F3 M-Inositol	F4 D-Threonine	F5 Fumaric Acid	F6 Bromo Succinic Acid	F7 Propionic Acid	F8 Mucic Acid	F9 Glycolic Acid	F10 Glyoxylic Acid	F11 D-Cellobiose	F12 Inosine
G1 Glycyl-L-Glutamic Acid	G2 Tricarballylic Acid	G3 L-Serine	G4 L-Threonine	G5 L-Alanine	G6 L-Alanyl-Glycine	G7 Acetoacetic Acid	G8 N-Acetyl-β-D-Mannosamine	G9 Mono Methyl Succinate	G10 Methyl Pyruvate	G11 D-Malic Acid	G12 L-Malic Acid
H1 Glycyl-L-Proline	H2 p-Hydroxy Phenyl Acetic Acid	H3 m-Hydroxy Phenyl Acetic Acid	H4 Tyramine	H5 D-Psicose	H6 L-Lyxose	H7 Glucuronamide	H8 Pyruvic Acid	H9 L-Galactonic Acid-γ-Lactone	H10 D-Galacturonic Acid	H11 Phenylethyl-amine	H12 2-Aminoethanol

PM3B MicroPlate™ Nitrogen Sources

A1 Negative Control	A2 Ammonia	A3 Nitrite	A4 Nitrate	A5 Urea	A6 Biuret	A7 L-Alanine	A8 L-Arginine	A9 L-Asparagine	A10 L-Aspartic Acid	A11 L-Cysteine	A12 L-Glutamic Acid
B1 L-Glutamine	B2 Glycine	B3 L-Histidine	B4 L-Isoleucine	B5 L-Leucine	B6 L-Lysine	B7 L-Methionine	B8 L-Phenylalanine	B9 L-Proline	B10 L-Serine	B11 L-Threonine	B12 L-Tryptophan
C1 L-Tyrosine	C2 L-Valine	C3 D-Alanine	C4 D-Asparagine	C5 D-Aspartic Acid	C6 D-Glutamic Acid	C7 D-Lysine	C8 D-Serine	C9 D-Valine	C10 L-Citrulline	C11 L-Homoserine	C12 L-Ornithine
D-1 N-Acetyl-D,L-Glutamic Acid	D2 N-Phthaloyl-L-Glutamic Acid	D3 L-Pyroglutamic Acid	D4 Hydroxylamine	D5 Methylamine	D6 N-Amylamine	D7 N-Butylamine	D8 Ethylamine	D9 Ethanolamine	D10 Ethylenediamine	D11 Putrescine	D12 Agmatine
E1 Histamine	E2 β-Phenylethyl-amine	E3 Tyramine	E4 Acetamide	E5 Formamide	E6 Glucuronamide	E7 D,L-Lactamide	E8 D-Glucosamine	E9 D-Galactosamine	E10 D-Mannosamine	E11 N-Acetyl-D-Glucosamine	E12 N-Acetyl-D-Galactosamine
F1 N-Acetyl-D-Mannosamine	F2 Adenine	F3 Adenosine	F4 Cytidine	F5 Cytosine	F6 Guanine	F7 Guanosine	F8 Thymine	F9 Thymidine	F10 Uracil	F11 Uridine	F12 Inosine
G1 Xanthine	G2 Xanthosine	G3 Uric Acid	G4 Alloxan	G5 Allantoin	G6 Parabanic Acid	G7 D,L-α-Amino-N-Butyric Acid	G8 γ-Amino-N-Butyric Acid	G9 ε-Amino-N-Caproic Acid	G10 D,L-α-Amino-Caprylic Acid	G11 δ-Amino-N-Valeric Acid	G12 α-Amino-N-Valeric Acid
H1 Ala-Asp	H2 Ala-Gln	H3 Ala-Glu	H4 Ala-Gly	H5 Ala-His	H6 Ala-Leu	H7 Ala-Thr	H8 Gly-Asn	H9 Gly-Gln	H10 Gly-Glu	H11 Gly-Met	H12 Met-Ala

PM6 MicroPlate™ Peptide Nitrogen Sources

A1 Negative Control	A2 Positive Control: L-Glutamine	A3 Ala-Ala	A4 Ala-Arg	A5 Ala-Asn	A6 Ala-Glu	A7 Ala-Gly	A8 Ala-His	A9 Ala-Leu	A10 Ala-Lys	A11 Ala-Phe	A12 Ala-Pro
B1 Ala-Ser	B2 Ala-Thr	B3 Ala-Trp	B4 Ala-Tyr	B5 Arg-Ala	B6 Arg-Arg	B7 Arg-Asp	B8 Arg-Gln	B9 Arg-Glu	B10 Arg-Ile	B11 Arg-Leu	B12 Arg-Lys
C1 Arg-Met	C2 Arg-Phe	C3 Arg-Ser	C4 Arg-Trp	C5 Arg-Tyr	C6 Arg-Val	C7 Asn-Glu	C8 Asn-Val	C9 Asp-Asp	C10 Asp-Glu	C11 Asp-Leu	C12 Asp-Lys
D1 Asp-Phe	D2 Asp-Trp	D3 Asp-Val	D4 Cys-Gly	D5 Gln-Gln	D6 Gln-Gly	D7 Glu-Asp	D8 Glu-Glu	D9 Glu-Gly	D10 Glu-Ser	D11 Glu-Trp	D12 Glu-Tyr
E1 Glu-Val	E2 Gly-Ala	E3 Gly-Arg	E4 Gly-Cys	E5 Gly-Gly	E6 Gly-His	E7 Gly-Leu	E8 Gly-Lys	E9 Gly-Met	E10 Gly-Phe	E11 Gly-Pro	E12 Gly-Ser
F1 Gly-Thr	F2 Gly-Trp	F3 Gly-Tyr	F4 Gly-Val	F5 His-Asp	F6 His-Gly	F7 His-Leu	F8 His-Lys	F9 His-Met	F10 His-Pro	F11 His-Ser	F12 His-Trp
G1 His-Tyr	G2 His-Val	G3 Ile-Ala	G4 Ile-Arg	G5 Ile-Gln	G6 Ile-Gly	G7 Ile-His	G8 Ile-Ile	G9 Ile-Met	G10 Ile-Phe	G11 Ile-Pro	G12 Ile-Ser
H1 Ile-Trp	H2 Ile-Tyr	H3 Ile-Val	H4 Leu-Ala	H5 Leu-Arg	H6 Leu-Asp	H7 Leu-Glu	H8 Leu-Gly	H9 Leu-Ile	H10 Leu-Leu	H11 Leu-Met	H12 Leu-Phe

PM7 MicroPlate™ Peptide Nitrogen Sources

A1 Negative Control	A2 Positive Control: L-Glutamine	A3 Leu-Ser	A4 Leu-Trp	A5 Leu-Val	A6 Lys-Ala	A7 Lys-Arg	A8 Lys-Glu	A9 Lys-Ile	A10 Lys-Leu	A11 Lys-Lys	A12 Lys-Phe
B1 Lys-Pro	B2 Lys-Ser	B3 Lys-Thr	B4 Lys-Trp	B5 Lys-Tyr	B6 Lys-Val	B7 Met-Arg	B8 Met-Asp	B9 Met-Gln	B10 Met-Glu	B11 Met-Gly	B12 Met-His
C1 Met-Ile	C2 Met-Leu	C3 Met-Lys	C4 Met-Met	C5 Met-Phe	C6 Met-Pro	C7 Met-Trp	C8 Met-Val	C9 Phe-Ala	C10 Phe-Gly	C11 Phe-Ile	C12 Phe-Phe
D1 Phe-Pro	D2 Phe-Ser	D3 Phe-Trp	D4 Pro-Ala	D5 Pro-Asp	D6 Pro-Gln	D7 Pro-Gly	D8 Pro-Hyp	D9 Pro-Leu	D10 Pro-Phe	D11 Pro-Pro	D12 Pro-Tyr
E1 Ser-Ala	E2 Ser-Gly	E3 Ser-His	E4 Ser-Leu	E5 Ser-Met	E6 Ser-Phe	E7 Ser-Pro	E8 Ser-Ser	E9 Ser-Tyr	E10 Ser-Val	E11 Thr-Ala	E12 Thr-Arg
F1 Thr-Glu	F2 Thr-Gly	F3 Thr-Leu	F4 Thr-Met	F5 Thr-Pro	F6 Trp-Ala	F7 Trp-Arg	F8 Trp-Asp	F9 Trp-Glu	F10 Trp-Gly	F11 Trp-Leu	F12 Trp-Lys
G1 Trp-Phe	G2 Trp-Ser	G3 Trp-Trp	G4 Trp-Tyr	G5 Tyr-Ala	G6 Tyr-Gln	G7 Tyr-Glu	G8 Tyr-Gly	G9 Tyr-His	G10 Tyr-Leu	G11 Tyr-Lys	G12 Tyr-Phe
H1 Tyr-Trp	H2 Tyr-Tyr	H3 Val-Arg	H4 Val-Asn	H5 Val-Asp	H6 Val-Gly	H7 Val-His	H8 Val-Ile	H9 Val-Leu	H10 Val-Tyr	H11 Val-Val	H12 Y-Glu-Gly

PM8 MicroPlate™ Peptide Nitrogen Sources

A1 Negative Control	A2 Positive Control: L-Glutamine	A3 Ala-Asp	A4 Ala-Gln	A5 Ala-Ile	A6 Ala-Met	A7 Ala-Val	A8 Asp-Ala	A9 Asp-Gln	A10 Asp-Gly	A11 Glu-Ala	A12 Gly-Asn
B1 Gly-Asp	B2 Gly-Ile	B3 His-Ala	B4 His-Glu	B5 His-His	B6 Ile-Asn	B7 Ile-Leu	B8 Leu-Asn	B9 Leu-His	B10 Leu-Pro	B11 Leu-Tyr	B12 Lys-Asp
C1 Lys-Gly	C2 Lys-Met	C3 Met-Thr	C4 Met-Tyr	C5 Phe-Asp	C6 Phe-Glu	C7 Gln-Glu	C8 Phe-Met	C9 Phe-Tyr	C10 Phe-Val	C11 Pro-Arg	C12 Pro-Asn
D1 Pro-Glu	D2 Pro-Ile	D3 Pro-Lys	D4 Pro-Ser	D5 Pro-Trp	D6 Pro-Val	D7 Ser-Asn	D8 Ser-Asp	D9 Ser-Gln	D10 Ser-Glu	D11 Thr-Asp	D12 Thr-Gln
E1 Thr-Phe	E2 Thr-Ser	E3 Trp-Val	E4 Tyr-Ile	E5 Tyr-Val	E6 Val-Ala	E7 Val-Gln	E8 Val-Glu	E9 Val-Lys	E10 Val-Met	E11 Val-Phe	E12 Val-Pro
F1 Val-Ser	F2 β-Ala-Ala	F3 β-Ala-Gly	F4 β-Ala-His	F5 Met-β-Ala	F6 β-Ala-Phe	F7 D-Ala-D-Ala	F8 D-Ala-Gly	F9 D-Ala-Leu	F10 D-Leu-D-Leu	F11 D-Leu-Gly	F12 D-Leu-Tyr
G1 Ɣ-Glu-Gly	G2 Ɣ-D-Glu-Gly	G3 Gly-D-Ala	G4 Gly-D-Asp	G5 Gly-D-Ser	G6 Gly-D-Thr	G7 Gly-D-Val	G8 Leu-β-Ala	G9 Leu-D-Leu	G10 Phe-β-Ala	G11 Ala-Ala-Ala	G12 D-Ala-Gly-Gly
H1 Gly-Gly-Ala	H2 Gly-Gly-D-Leu	H3 Gly-Gly-Gly	H4 Gly-Gly-Ile	H5 Gly-Gly-Leu	H6 Gly-Gly-Phe	H7 Val-Tyr-Val	H8 Gly-Phe-Phe	H9 Leu-Gly-Gly	H10 Leu-Leu-Leu	H11 Phe-Gly-Gly	H12 Tyr-Gly-Gly

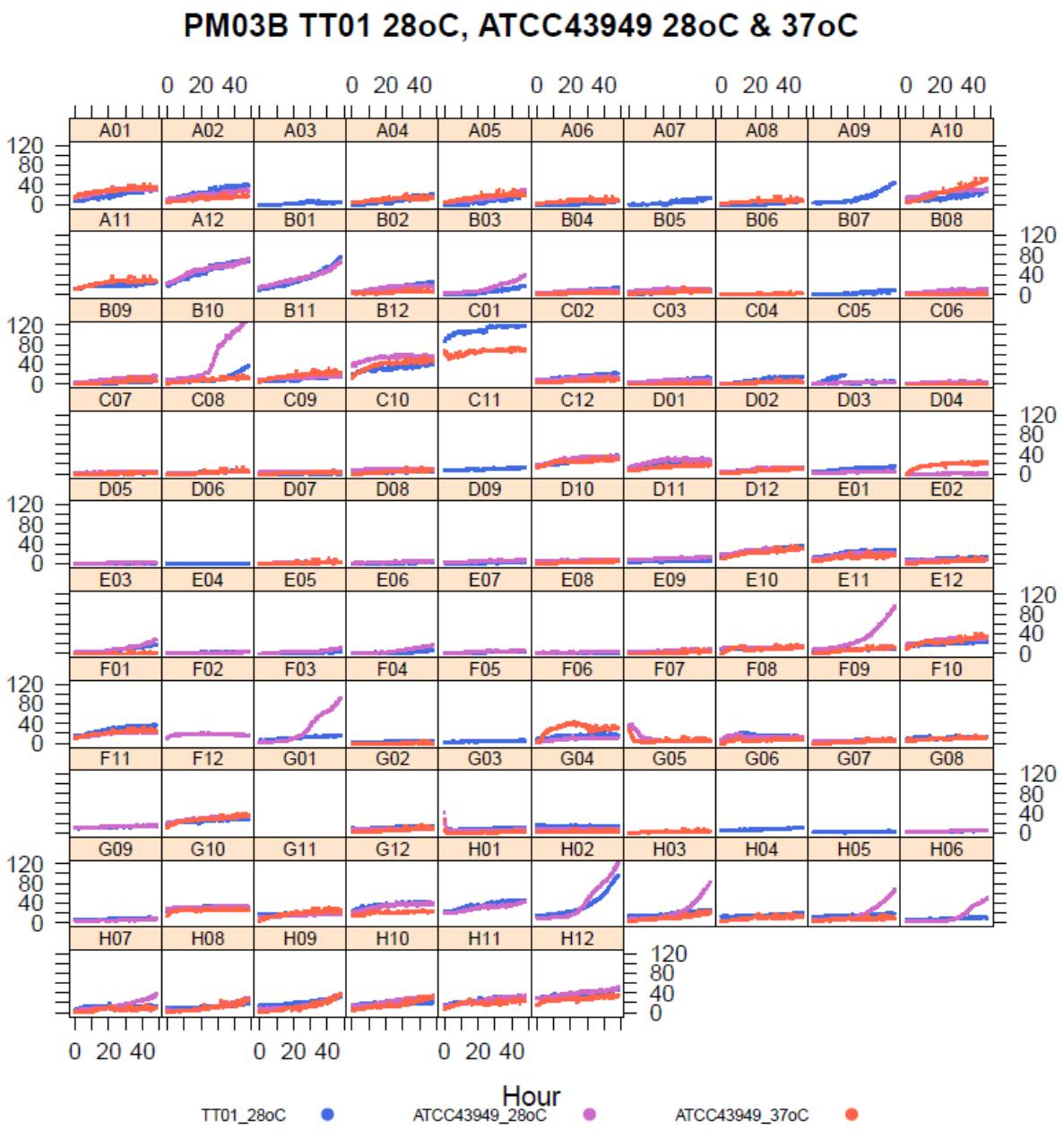
PM9 MicroPlate™ Osmolytes

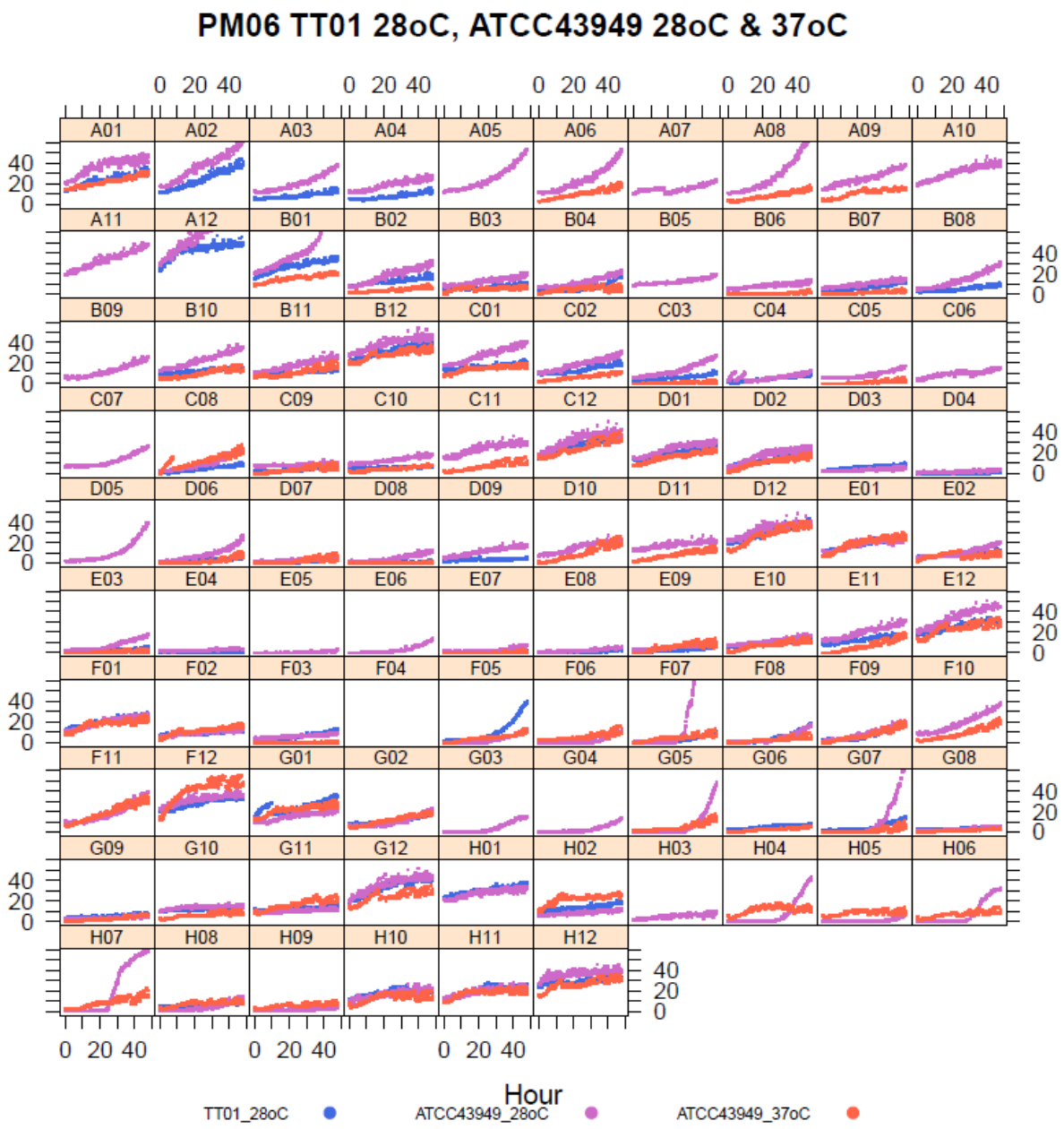
A1 NaCl 1%	A2 NaCl 2%	A3 NaCl 3%	A4 NaCl 4%	A5 NaCl 5%	A6 NaCl 5.5%	A7 NaCl 6%	A8 NaCl 6.5%	A9 NaCl 7%	A10 NaCl 8%	A11 NaCl 9%	A12 NaCl 10%
B1 NaCl 6%	B2 NaCl 6% + Betaine	B3 NaCl 6% + N-N Dimethyl glycine	B4 NaCl 6% + Sarcosine	B5 NaCl 6% + Dimethyl sulphonyl propionate	B6 NaCl 6% + MOPS	B7 NaCl 6% + Ectoine	B8 NaCl 6% + Choline	B9 NaCl 6% + Phosphoryl choline	B10 NaCl 6% + Creatine	B11 NaCl 6% + Creatinine	B12 NaCl 6% + L- Carnitine
C1 NaCl 6% + KCl	C2 NaCl 6% + L-proline	C3 NaCl 6% + N-Acethyl L-glutamine	C4 NaCl 6% + β-Glutamic acid	C5 NaCl 6% + γ–Amino -n-butyric acid	C6 NaCl 6% + Glutathione	C7 NaCl 6% + Glycerol	C8 NaCl 6% + Trehalose	C9 NaCl 6% + Trimethylamine -N-oxide	C10 NaCl 6% + Trimethylamine	C11 NaCl 6% + Octopine	C12 NaCl 6% + Trigonelline
D-1 Potassium chloride 3%	D2 Potassium chloride 4%	D3 Potassium chloride 5%	D4 Potassium chloride 6%	D5 Sodium sulfate 2%	D6 Sodium sulfate 3%	D7 Sodium sulfate 4%	D8 Sodium sulfate 5%	D9 Ethylene glycol 5%	D10 Ethylene glycol 10%	D11 Ethylene glycol 15%	D12 Ethylene glycol 20%
E1 Sodium formate 1%	E2 Sodium formate 2%	E3 Sodium formate 3%	E4 Sodium formate 4%	E5 Sodium formate 5%	E6 Sodium formate 6%	E7 Urea 2%	E8 Urea 3%	E9 Urea 4%	E10 Urea 5%	E11 Urea 6%	E12 Urea 7%
F1 Sodium Lactate 1%	F2 Sodium Lactate 2%	F3 Sodium Lactate 3%	F4 Sodium Lactate 4%	F5 Sodium Lactate 5%	F6 Sodium Lactate 6%	F7 Sodium Lactate 7%	F8 Sodium Lactate 8%	F9 Sodium Lactate 9%	F10 Sodium Lactate 10%	F11 Sodium Lactate 11%	F12 Sodium Lactate 12%
G1 Sodium Phosphate pH 7 20mM	G2 Sodium Phosphate pH 7 50mM	G3 Sodium Phosphate pH 7 100mM	G4 Sodium Phosphate pH 7 200mM	G5 Sodium Benzoate pH 5.2 20mM	G6 Sodium Benzoate pH 5.2 50mM	G7 Sodium Benzoate pH5.2 100mM	G8 Sodium Benzoate pH 5.2 200mM	G9 Ammonium sulfate pH6 10mM	G10 Ammonium sulfate pH 6 20mM	G11 Ammonium sulfate pH 6 50mM	G12 Ammonium sulfate pH6 100mM
H1 Sodium Nitrate 10mM	H2 Sodium Nitrate 20mM	H3 Sodium Nitrate 40mM	H4 Sodium Nitrate 60mM	H5 Sodium Nitrate 80mM	H6 Sodium Nitrate 100mM	H7 Sodium Nitrite 10mM	H8 Sodium Nitrite 20mM	H9 Sodium Nitrite 40mM	H10 Sodium Nitrite 60mM	H11 Sodium Nitrite 80mM	H12 Sodium Nitrite 100mM

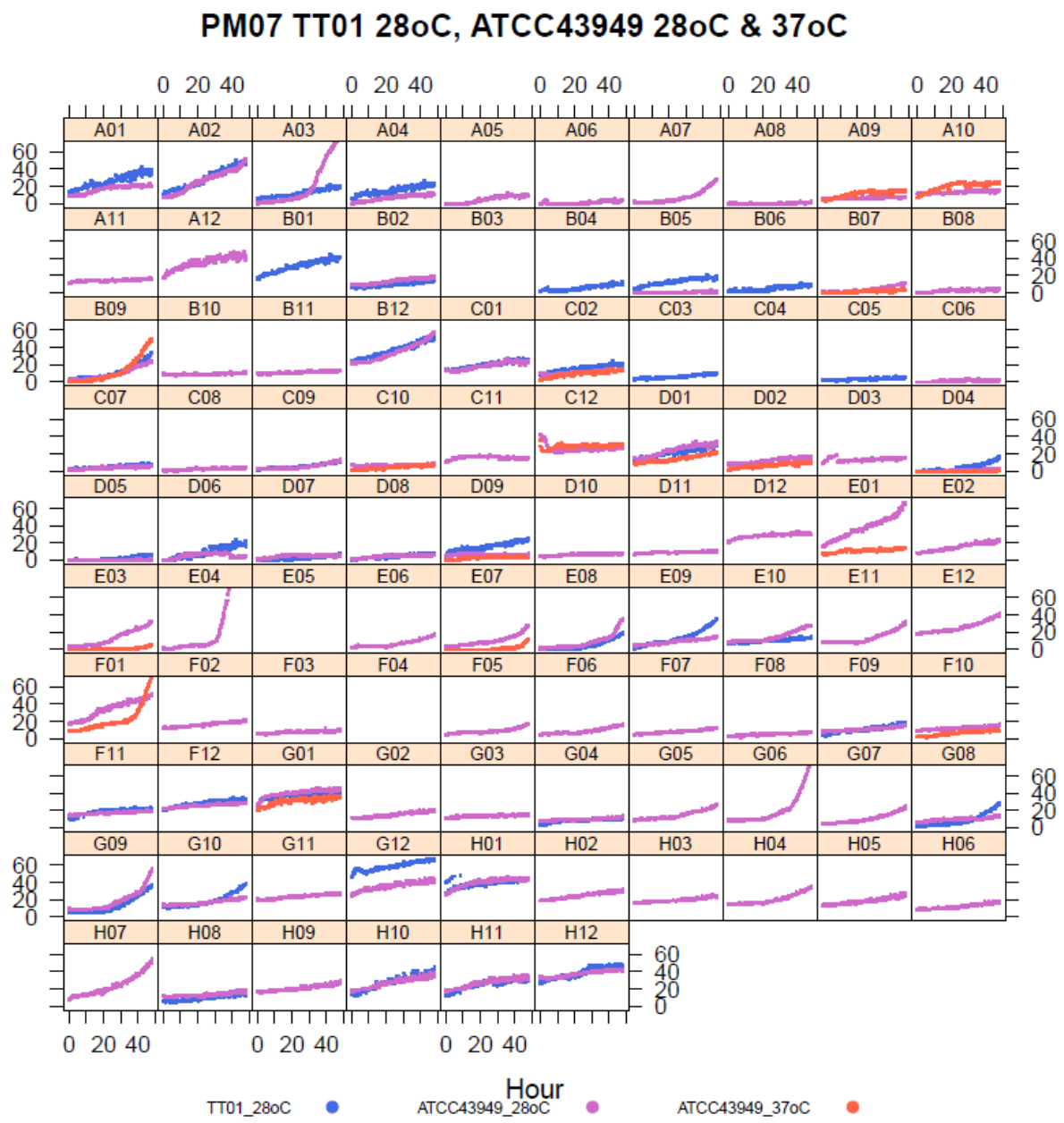
PM10 MicroPlate™ pH

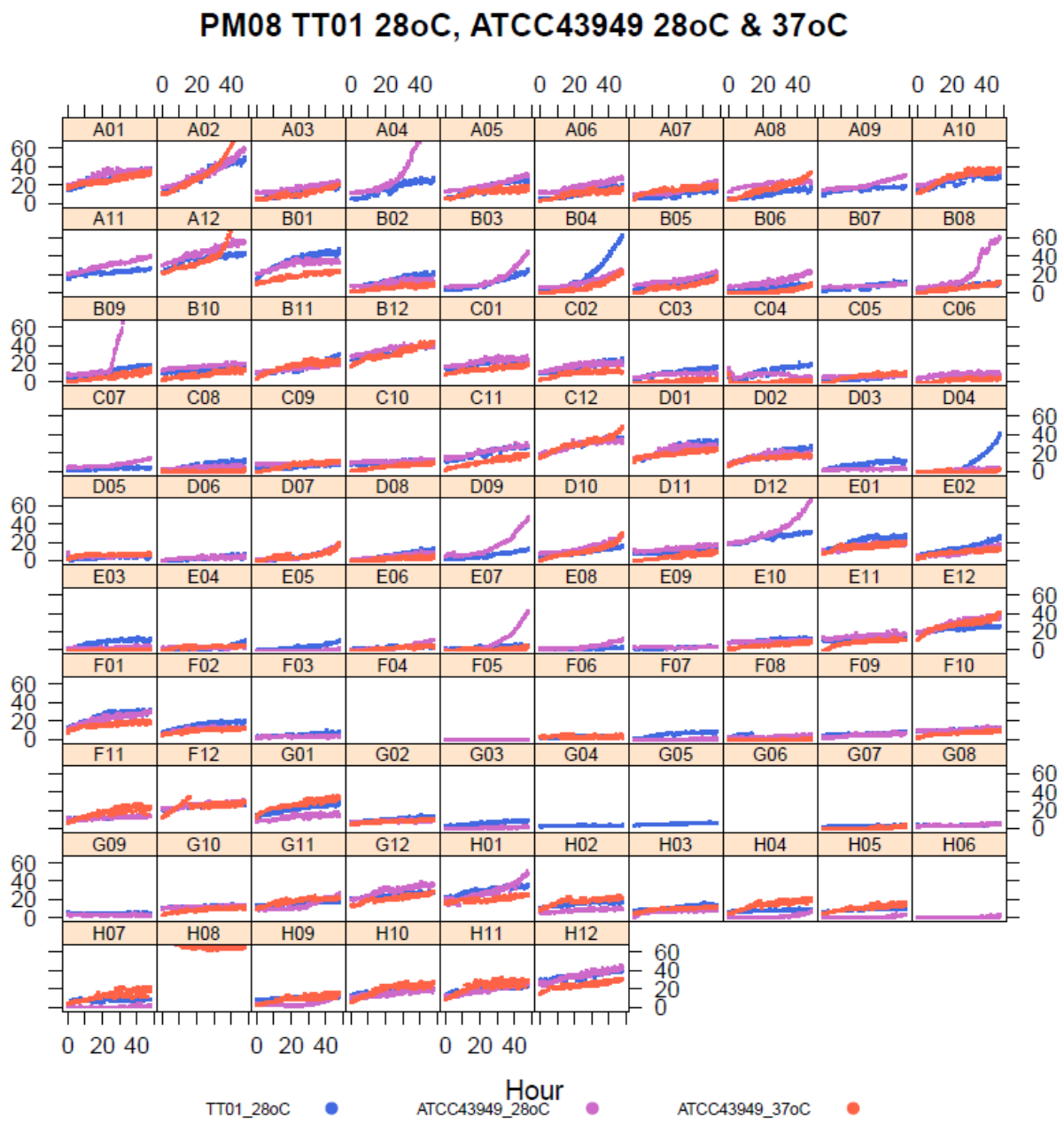
A1 pH 3.5	A2 pH 4	A3 pH 4.5	A4 pH 5	A5 pH 5.5	A6 pH 6	A7 pH 7	A8 pH 8	A9 pH 8.5	A10 pH 9	A11 pH 9.5	A12 pH 10
B1 pH 4.5	B2 pH 4.5 + L-Alanine	B3 pH 4.5 + L-Arginine	B4 pH 4.5 + L-Asparagine	B5 pH 4.5 + L-Aspartic Acid	B6 pH 4.5 + L-Glutamic Acid	B7 pH 4.5 + L-Glutamine	B8 pH 4.5 + Glycine	B9 pH 4.5 + L-Histidine	B10 pH 4.5 + L-Isoleucine	B11 pH 4.5 + L-Leucine	B12 pH 4.5 + L-Lysine
C1 pH 4.5 + L-Methionine	C2 pH 4.5 + L- Phenylalanine	C3 pH 4.5 + L-Proline	C4 pH 4.5 + L-Serine	C5 pH 4.5 + L-Threonine	C6 pH 4.5 + L-Tryptophan	C7 pH 4.5 + L-Tyrosine	C8 pH 4.5 + L-Valine	C9 pH 4.5 + Hydroxy- L-Proline	C10 pH 4.5 + L-Ornithine	C11 pH 4.5 + L-Homoarginine	C12 pH 4.5 + L-Homoserine
D-1 pH 4.5 + Anthranilic acid	D2 pH 4.5 + L-Norleucine	D3 pH 4.5 + L-Norvaline	D4 pH 4.5 + α- Amino-N- butyric acid	D5 pH 4.5 + p- Aminobenzoate	D6 pH 4.5 + L-Cystelc acid	D7 pH 4.5 + D-Lysine	D8 pH 4.5 + 5-Hydroxy Lysine	D9 pH 4.5 + 5-Hydroxy Tryptophan	D10 pH 4.5 + D,L-Diamino pimelic acid	D11 pH 4.5 + Trimethyl amine-N-oxide	D12 pH 4.5 + Urea
E1 pH 9.5	E2 pH 9.5 + L-Alanine	E3 pH 9.5 + L-Arginine	E4 pH 9.5 + L-Asparagine	E5 pH 9.5 + L-Aspartic Acid	E6 pH 9.5 + L-Glutamic Acid	E7 pH 9.5 + L-Glutamine	E8 pH 9.5 + Glycine	E9 pH 9.5 + L-Histidine	E10 pH 9.5 + L-Isoleucine	E11 pH 9.5 + L-Leucine	E12 pH 9.5 + L-Lysine
F1 pH 9.5 + L-Methionine	F2 pH 9.5 + L- Phenylalanine	F3 pH 9.5 + L-Proline	F4 pH 9.5 + L-Serine	F5 pH 9.5 + L-Threonine	F6 pH 9.5 + L-Tryptophan	F7 pH 9.5 + L-Tyrosine	F8 pH 9.5 + L-Valine	F9 pH 9.5 + Hydroxy- L-Proline	F10 pH 9.5 + L-Ornithine	F11 pH 9.5 + L-Homoarginine	F12 pH 9.5 + L-Homoserine
G1 pH 9.5 + Anthranilic acid	G2 pH 9.5 + L-Norleucine	G3 pH 9.5 + L-Norvaline	G4 pH 9.5 + Agmatine	G5 pH 9.5 + Cadaverine	G6 pH 9.5 + Putrescine	G7 pH 9.5 + Histamine	G8 pH 9.5 + Phenylethylamin e	G9 pH 9.5 + Tyramine	G10 pH 9.5 + Creatine	G11 pH 9.5 + Trimethyl amine-N-oxide	G12 pH 9.5 + Urea
H1 X-Caprylate	H2 X-α-D- Glucoside	H3 X-β-D- Glucoside	H4 X-α-D- Galactoside	H5 X-β-D- Galactoside	H6 X-α- D- Glucuronide	H7 X-β- D- Glucuronide	H8 X-β-D- Glucosaminide	H9 X-β-D- Galactosaminid e	H10 X-α-D- Mannoside	H11 X-PO4	H12 X-SO4

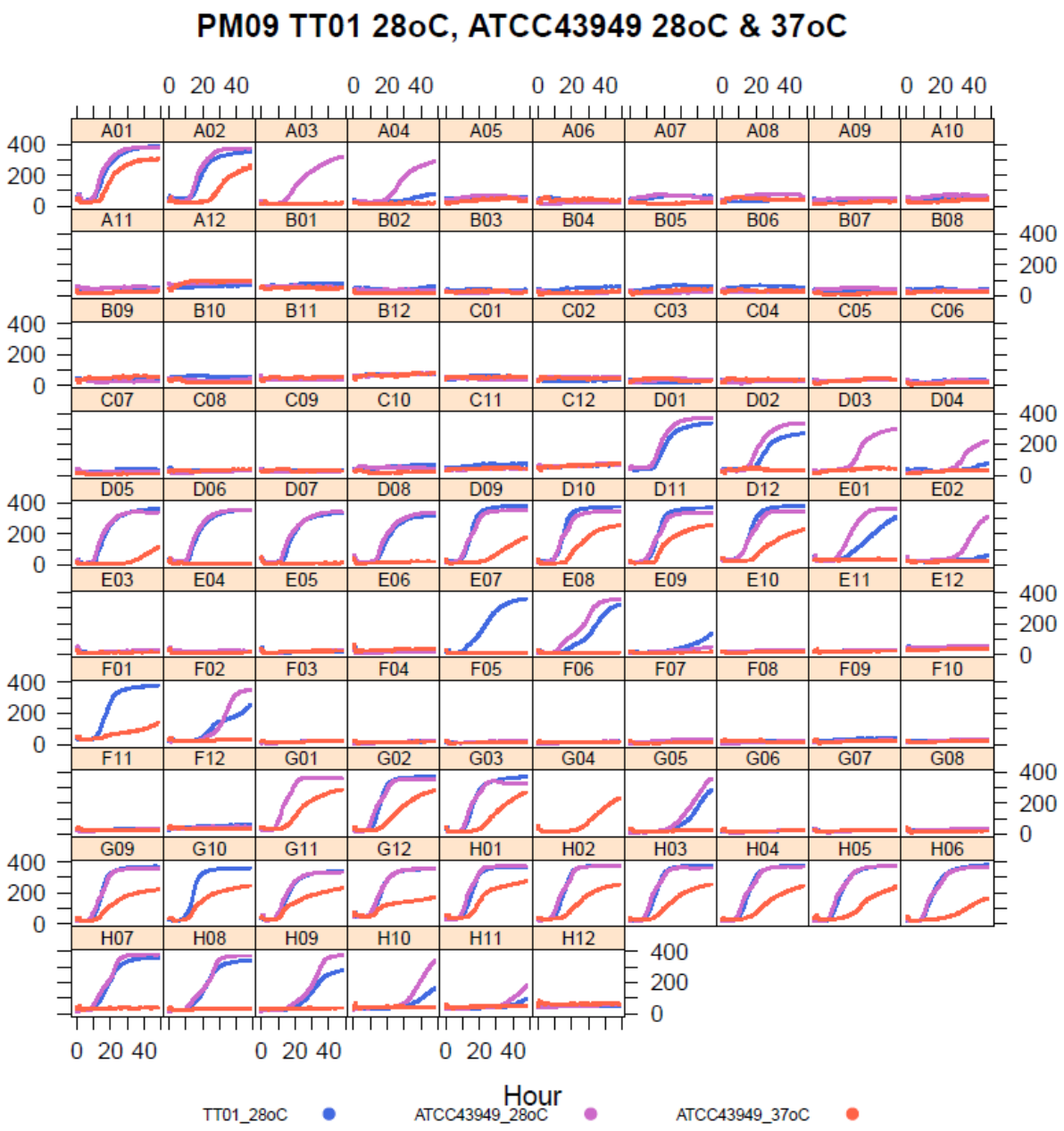
Fig. S2.


PM03B TT01 28oC, ATCC43949 28oC & 37oC
A01 A02 A03 A04 A05 A06 A07 A08 A09 A10
A11 A12 B01 B02 B03 B04 B05 B06 B07 B08
B09 B10 B11 B12 C01 C02 C03 C04 C05 C06
C07 C08 C09 C10 C11 C12 D01 D02 D03 D04
D05 D06 D07 D08 D09 D10 D11 D12 E01 E02
E03 E04 E05 E06 E07 E08 E09 E10 E11 E12
F01 F02 F03 F04 F05 F06 F07 F08 F09 F10
F11 F12 G01 G02 G03 G04 G05 G06 G07 G08
G09 G10 G11 G12 H01 H02 H03 H04 H05 H06
H07 H08 H09 H10 H11 H12
0 20 40
0 80 40 120
Hour
TT01_28oC
ATCC43949_28oC
ATCC43949_37oC


PM06 TT01 28oC, ATCC43949 28oC & 37oC
A01 A02 A03 A04 A05 A06 A07 A08 A09 A10
A11 A12 B01 B02 B03 B04 B05 B06 B07 B08
B09 B10 B11 B12 C01 C02 C03 C04 C05 C06
C07 C08 C09 C10 C11 C12 D01 D02 D03 D04
D05 D06 D07 D08 D09 D10 D11 D12 E01 E02
E03 E04 E05 E06 E07 E08 E09 E10 E11 E12
F01 F02 F03 F04 F05 F06 F07 F08 F09 F10
F11 F12 G01 G02 G03 G04 G05 G06 G07 G08
G09 G10 G11 G12 H01 H02 H03 H04 H05 H06
H07 H08 H09 H10 H11 H12
0 20 40
Hour
TT01_28oC
ATCC43949_28oC
ATCC43949_37oC


PM07 TT01 28oC, ATCC43949 28oC & 37oC
A01 A02 A03 A04 A05 A06 A07 A08 A09 A10
A11 A12 B01 B02 B03 B04 B05 B06 B07 B08
B09 B10 B11 B12 C01 C02 C03 C04 C05 C06
C07 C08 C09 C10 C11 C12 D01 D02 D03 D04
D05 D06 D07 D08 D09 D10 D11 D12 E01 E02
E03 E04 E05 E06 E07 E08 E09 E10 E11 E12
F01 F02 F03 F04 F05 F06 F07 F08 F09 F10
F11 F12 G01 G02 G03 G04 G05 G06 G07 G08
G09 G10 G11 G12 H01 H02 H03 H04 H05 H06
H07 H08 H09 H10 H11 H12
0 20 40
0 20 40
0 20 40
0 20 40
0 20 40
0 20 40
0 20 40
0 20 40
0 20 40
0 20 40
0 20 40
0 20 40
Hour
TT01_28oC
ATCC43949_28oC
ATCC43949_37oC


PM08 TT01 28oC, ATCC43949 28oC & 37oC
A01 A02 A03 A04 A05 A06 A07 A08 A09 A10
A11 A12 B01 B02 B03 B04 B05 B06 B07 B08
B09 B10 B11 B12 C01 C02 C03 C04 C05 C06
C07 C08 C09 C10 C11 C12 D01 D02 D03 D04
D05 D06 D07 D08 D09 D10 D11 D12 E01 E02
E03 E04 E05 E06 E07 E08 E09 E10 E11 E12
F01 F02 F03 F04 F05 F06 F07 F08 F09 F10
F11 F12 G01 G02 G03 G04 G05 G06 G07 G08
G09 G10 G11 G12 H01 H02 H03 H04 H05 H06
H07 H08 H09 H10 H11 H12
0 20 40
0 20 40
0 20 40
0 20 40
0 20 40
0 20 40
0 20 40
0 20 40
0 20 40
0 20 40
0 20 40
0 20 40
Hour
TT01_28oC
ATCC43949_28oC
ATCC43949_37oC


PM09 TT01 28oC, ATCC43949 28oC & 37oC
A01 A02 A03 A04 A05 A06 A07 A08 A09 A10
A11 A12 B01 B02 B03 B04 B05 B06 B07 B08
B09 B10 B11 B12 C01 C02 C03 C04 C05 C06
C07 C08 C09 C10 C11 C12 D01 D02 D03 D04
D05 D06 D07 D08 D09 D10 D11 D12 E01 E02
E03 E04 E05 E06 E07 E08 E09 E10 E11 E12
F01 F02 F03 F04 F05 F06 F07 F08 F09 F10
F11 F12 G01 G02 G03 G04 G05 G06 G07 G08
G09 G10 G11 G12 H01 H02 H03 H04 H05 H06
H07 H08 H09 H10 H11 H12
0 20 40
0
200
400
Hour
TT01_28oC
ATCC43949_28oC
ATCC43949_37oC

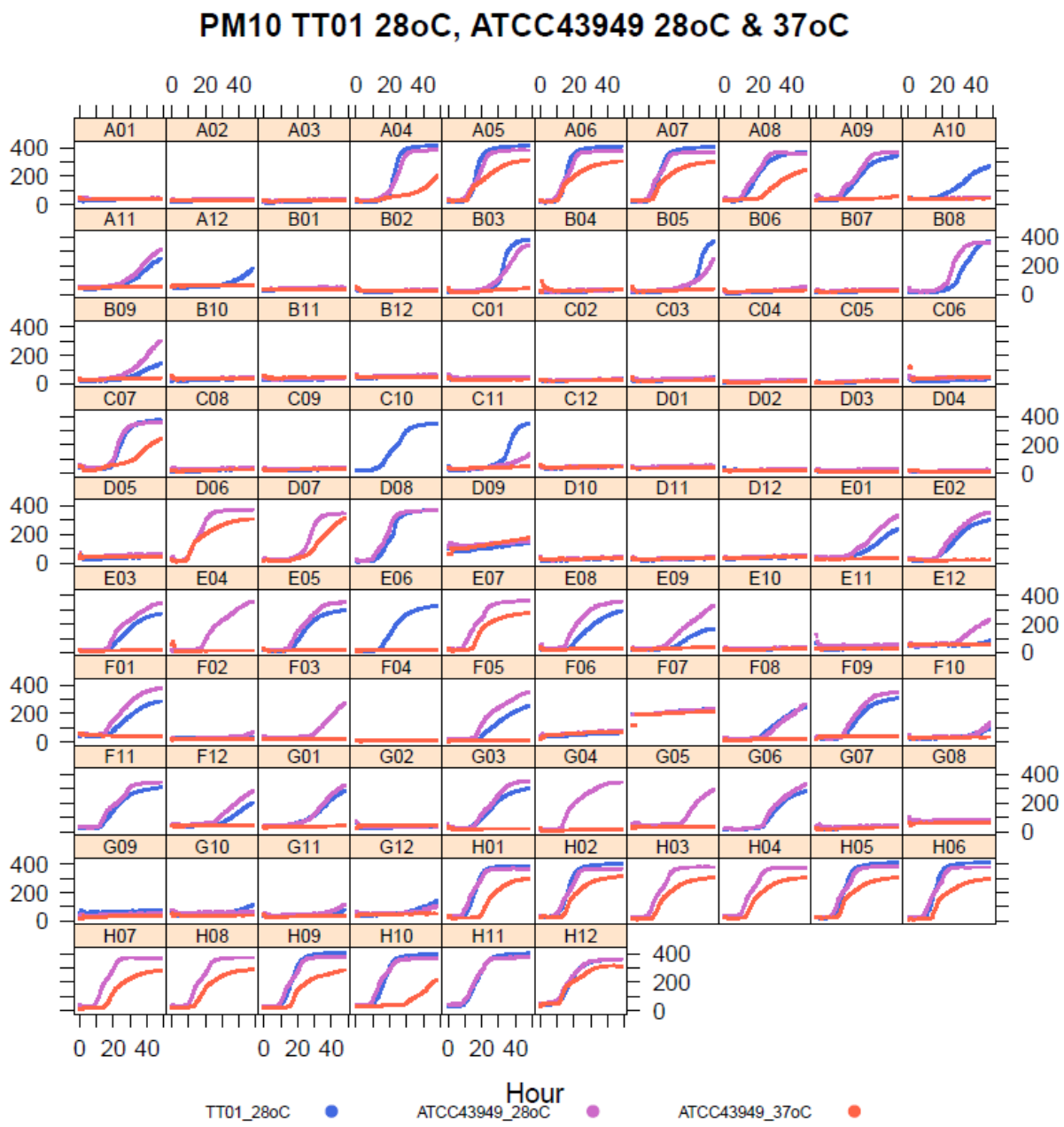

PM10 TT01 28oC, ATCC43949 28oC & 37oC
A01 A02 A03 A04 A05 A06 A07 A08 A09 A10
A11 A12 B01 B02 B03 B04 B05 B06 B07 B08
B09 B10 B11 B12 C01 C02 C03 C04 C05 C06
C07 C08 C09 C10 C11 C12 D01 D02 D03 D04
D05 D06 D07 D08 D09 D10 D11 D12 E01 E02
E03 E04 E05 E06 E07 E08 E09 E10 E11 E12
F01 F02 F03 F04 F05 F06 F07 F08 F09 F10
F11 F12 G01 G02 G03 G04 G05 G06 G07 G08
G09 G10 G11 G12 H01 H02 H03 H04 H05 H06
H07 H08 H09 H10 H11 H12
0 20 40
400
200
0
Hour
TT01_28oC
ATCC43949_28oC
ATCC43949_37oC

Table S1: Phenoarray – PM3B nitrogen sources

Species 1	Temperature 1	Respiration type 1	Species 2	Temperature 2	Respiration type 2	Substrate
Pa^{ATCC43949}	28	Negative	*Pa*^{ATCC43949}	37	Negative	L-Phenylalanine
Pa^{ATCC43949}	28	Positive	*Pa*^{ATCC43949}	37	Negative	L-Serine
Pa^{ATCC43949}	28	Negative	*Pa*^{ATCC43949}	37	Negative	Hydroxylamine
Pa^{ATCC43949}	28	Minimal	*Pa*^{ATCC43949}	37	Negative	Tyramine
Pa^{ATCC43949}	28	Positive	*Pa*^{ATCC43949}	37	Negative	N-Acetyl-D-Glucosamine
Pa^{ATCC43949}	28	Negative	*Pa*^{ATCC43949}	37	Minimal	Guanine
Pa^{ATCC43949}	28	Positive	*Pa*^{ATCC43949}	37	Negative	Ala-Glu
Pa^{ATCC43949}	28	Minimal	*Pl*^{TT01}	28	Negative	L-Histidine
Pa^{ATCC43949}	28	Negative	*Pl*^{TT01}	28	Negative	L-Proline
Pa^{ATCC43949}	28	Positive	*Pl*^{TT01}	28	Minimal	L-Serine
Pa^{ATCC43949}	28	Negative	*Pl*^{TT01}	28	Negative	D-Valine
Pa^{ATCC43949}	28	Minimal	*Pl*^{TT01}	28	Negative	Tyramine
Pa^{ATCC43949}	28	Positive	*Pl*^{TT01}	28	Negative	N-Acetyl-D-Glucosamine
Pa^{ATCC43949}	28	Positive	*Pl*^{TT01}	28	Negative	Adenosine
Pa^{ATCC43949}	28	Negative	*Pl*^{TT01}	28	Negative	Alloxan
Pa^{ATCC43949}	28	Positive	*Pl*^{TT01}	28	Negative	Ala-Glu
Pa^{ATCC43949}	28	Minimal	*Pl*^{TT01}	28	Negative	Ala-His
Pa^{ATCC43949}	28	Minimal	*Pl*^{TT01}	28	Negative	Ala-Leu
Pa^{ATCC43949}	37	Minimal	*Pl*^{TT01}	28	Negative	L-Aspartic Acid
Pa^{ATCC43949}	37	Positive	*Pl*^{TT01}	28	Positive	L-Tyrosine
Pa^{ATCC43949}	37	Negative	*Pl*^{TT01}	28	Negative	Thymine
Pa^{ATCC43949}	37	Negative	*Pl*^{TT01}	28	Negative	Alloxan

Table S2: Phenoarray – PM6 peptide nitrogen sources

Species 1	Temperature 1	Respiration type 1	Species 2	Temperature 2	Respiration type 2	Substrate
Pa^{ATCC43949}	28	Minimal	*Pa*^{ATCC43949}	37	Negative	Ala-Glu
Pa^{ATCC43949}	28	Minimal	*Pa*^{ATCC43949}	37	Negative	Ala-His
Pa^{ATCC43949}	28	Negative	*Pa*^{ATCC43949}	37	Negative	Ala-Leu
Pa^{ATCC43949}	28	Positive	*Pa*^{ATCC43949}	37	Negative	Ala-Ser
Pa^{ATCC43949}	28	Negative	*Pa*^{ATCC43949}	37	Negative	Ala-Thr
Pa^{ATCC43949}	28	Negative	*Pa*^{ATCC43949}	37	Negative	Arg-Asp
Pa^{ATCC43949}	28	Negative	*Pa*^{ATCC43949}	37	Negative	Arg-Ile
Pa^{ATCC43949}	28	Negative	*Pa*^{ATCC43949}	37	Negative	Arg-Met
Pa^{ATCC43949}	28	Negative	*Pa*^{ATCC43949}	37	Negative	Arg-Phe
Pa^{ATCC43949}	28	Negative	*Pa*^{ATCC43949}	37	Negative	Arg-Ser
Pa^{ATCC43949}	28	Negative	*Pa*^{ATCC43949}	37	Negative	Arg-Tyr
Pa^{ATCC43949}	28	Negative	*Pa*^{ATCC43949}	37	Negative	Glu-Glu
Pa^{ATCC43949}	28	Negative	*Pa*^{ATCC43949}	37	Negative	Gly-Arg
Pa^{ATCC43949}	28	Negative	*Pa*^{ATCC43949}	37	Negative	Gly-Tyr
Pa^{ATCC43949}	28	Minimal	*Pa*^{ATCC43949}	37	Negative	His-Leu
Pa^{ATCC43949}	28	Minimal	*Pa*^{ATCC43949}	37	Negative	Ile-Gln
Pa^{ATCC43949}	28	Minimal	*Pa*^{ATCC43949}	37	Negative	Ile-His
Pa^{ATCC43949}	28	Negative	*Pa*^{ATCC43949}	37	Minimal	Ile-Tyr
Pa^{ATCC43949}	28	Minimal	*Pa*^{ATCC43949}	37	Negative	Leu-Ala
Pa^{ATCC43949}	28	Negative	*Pa*^{ATCC43949}	37	Negative	Leu-Asp
Pa^{ATCC43949}	28	Minimal	*Pa*^{ATCC43949}	37	Negative	Leu-Glu
Pa^{ATCC43949}	28	Negative	*Pl*^{TT01}	28	Negative	Ala-Ala
Pa^{ATCC43949}	28	Positive	*Pl*^{TT01}	28	Minimal	Ala-Pro
Pa^{ATCC43949}	28	Positive	*Pl*^{TT01}	28	Minimal	Ala-Ser
Pa^{ATCC43949}	28	Negative	*Pl*^{TT01}	28	Negative	Arg-Gln

Pa^{ATCC43949}	28	Negative	*Pl*^{TT01}	28	Negative	Arg-Ile
Pa^{ATCC43949}	28	Negative	*Pl*^{TT01}	28	Minimal	Arg-Met
Pa^{ATCC43949}	28	Negative	*Pl*^{TT01}	28	Negative	Arg-Ser
Pa^{ATCC43949}	28	Negative	*Pl*^{TT01}	28	Negative	Asn-Val
Pa^{ATCC43949}	28	Negative	*Pl*^{TT01}	28	Negative	Asp-Glu
Pa^{ATCC43949}	28	Negative	*Pl*^{TT01}	28	Negative	Glu-Glu
Pa^{ATCC43949}	28	Negative	*Pl*^{TT01}	28	Negative	Glu-Gly
Pa^{ATCC43949}	28	Negative	*Pl*^{TT01}	28	Negative	Gly-Ala
Pa^{ATCC43949}	28	Negative	*Pl*^{TT01}	28	Negative	Gly-Arg
Pa^{ATCC43949}	28	Negative	*Pl*^{TT01}	28	Negative	Gly-Pro
Pa^{ATCC43949}	28	Negative	*Pl*^{TT01}	28	Minimal	His-Asp
Pa^{ATCC43949}	28	Minimal	*Pl*^{TT01}	28	Negative	His-Leu
Pa^{ATCC43949}	28	Negative	*Pl*^{TT01}	28	Minimal	His-Tyr
Pa^{ATCC43949}	28	Minimal	*Pl*^{TT01}	28	Negative	Ile-His
Pa^{ATCC43949}	37	Negative	*Pl*^{TT01}	28	Minimal	Ala-Ser
Pa^{ATCC43949}	37	Negative	*Pl*^{TT01}	28	Negative	Gly-Tyr
Pa^{ATCC43949}	37	Negative	*Pl*^{TT01}	28	Minimal	His-Asp
Pa^{ATCC43949}	37	Positive	*Pl*^{TT01}	28	Minimal	His-Trp

Table S3: Phenoarray – PM7 peptide nitrogen sources

Species 1	Temperature 1	Respiration type 1	Species 2	Temperature 2	Respiration type 2	Substrate
Pa^{ATCC43949}	28	Negative	*Pa*^{ATCC43949}	37	Negative	Lys-Ile
Pa^{ATCC43949}	28	Negative	*Pa*^{ATCC43949}	37	Minimal	Lys-Leu
Pa^{ATCC43949}	28	Negative	*Pa*^{ATCC43949}	37	Negative	Phe-Gly
Pa^{ATCC43949}	28	Positive	*Pa*^{ATCC43949}	37	Minimal	Phe-Pro
Pa^{ATCC43949}	28	Negative	*Pa*^{ATCC43949}	37	Negative	Phe-Ser
Pa^{ATCC43949}	28	Negative	*Pa*^{ATCC43949}	37	Negative	Pro-Leu
Pa^{ATCC43949}	28	Positive	*Pa*^{ATCC43949}	37	Negative	Ser-Ala
Pa^{ATCC43949}	28	Minimal	*Pa*^{ATCC43949}	37	Negative	Ser-His
Pa^{ATCC43949}	28	Minimal	*Pa*^{ATCC43949}	37	Negative	Ser-Pro
Pa^{ATCC43949}	28	Positive	*Pa*^{ATCC43949}	37	Positive	Thr-Glu
Pa^{ATCC43949}	28	Negative	*Pa*^{ATCC43949}	37	Negative	Trp-Gly
Pa^{ATCC43949}	28	Minimal	*Pl*^{TT01}	28	Minimal	Negative Control
Pa^{ATCC43949}	28	Negative	*Pl*^{TT01}	28	Negative	Leu-Trp
Pa^{ATCC43949}	28	Negative	*Pl*^{TT01}	28	Negative	Lys-Tyr
Pa^{ATCC43949}	28	Negative	*Pl*^{TT01}	28	Negative	Pro-Ala
Pa^{ATCC43949}	28	Negative	*Pl*^{TT01}	28	Negative	Pro-Asp
Pa^{ATCC43949}	28	Negative	*Pl*^{TT01}	28	Negative	Pro-Gly
Pa^{ATCC43949}	28	Negative	*Pl*^{TT01}	28	Negative	Pro-Leu
Pa^{ATCC43949}	28	Positive	*Pl*^{TT01}	28	Positive	Tyr-Phe
Pa^{ATCC43949}	37	Negative	*Pl*^{TT01}	28	Negative	Met-Leu
Pa^{ATCC43949}	37	Minimal	*Pl*^{TT01}	28	Negative	Phe-Pro
Pa^{ATCC43949}	37	Negative	*Pl*^{TT01}	28	Negative	Pro-Ala
Pa^{ATCC43949}	37	Negative	*Pl*^{TT01}	28	Negative	Pro-Leu

Table S4: Phenoarray – PM8 peptide nitrogen sources

Species 1	Temperature 1	Respiration type 1	Species 2	Temperature 2	Respiration type 2	Substrate
Pa^{ATCC43949}	28	Positive	*Pa*^{ATCC43949}	37	Positive	Gly-Asn
Pa^{ATCC43949}	28	Minimal	*Pa*^{ATCC43949}	37	Negative	Leu-Asn
Pa^{ATCC43949}	28	Positive	*Pa*^{ATCC43949}	37	Negative	Leu-His
Pa^{ATCC43949}	28	Minimal	*Pa*^{ATCC43949}	37	Negative	Val-Gln
Pa^{ATCC43949}	28	Positive	*Pl*^{TT01}	28	Negative	Ala-Gln
Pa^{ATCC43949}	28	Minimal	*Pl*^{TT01}	28	Negative	Leu-Asn
Pa^{ATCC43949}	28	Positive	*Pl*^{TT01}	28	Negative	Leu-His
Pa^{ATCC43949}	28	Minimal	*Pl*^{TT01}	28	Negative	Val-Gln
Pa^{ATCC43949}	37	Positive	*Pl*^{TT01}	28	Minimal	Gly-Asn

Table S5: Phenoarray – PM9 osmolytes

Species 1	Temperature 1	Respiration type 1	Species 2	Temperature 2	Respiration type 2	Substrate
Pa^{ATCC43949}	28	Sub-optimal	*Pa*^{ATCC43949}	37	Sub-optimal	NaCl
Pa^{ATCC43949}	28	Sub-optimal	*Pa*^{ATCC43949}	37	Sub-optimal	NaCl
Pa^{ATCC43949}	28	Sub-optimal	*Pa*^{ATCC43949}	37	Sub-optimal	Potassium chloride 4%
Pa^{ATCC43949}	28	Sub-optimal	*Pa*^{ATCC43949}	37	Sub-optimal	Potassium chloride 5%
Pa^{ATCC43949}	28	Sub-optimal	*Pa*^{ATCC43949}	37	Sub-optimal	Potassium chloride 6%
Pa^{ATCC43949}	28	Sub-optimal	*Pa*^{ATCC43949}	37	Sub-optimal	Sodium sulfate 2%
Pa^{ATCC43949}	28	Sub-optimal	*Pa*^{ATCC43949}	37	Sub-optimal	Sodium sulfate 3%
Pa^{ATCC43949}	28	Sub-optimal	*Pa*^{ATCC43949}	37	Sub-optimal	Ethylene glycol 5%
Pa^{ATCC43949}	28	Sub-optimal	*Pa*^{ATCC43949}	37	Sub-optimal	Ethylene glycol 10%
Pa^{ATCC43949}	28	Sub-optimal	*Pa*^{ATCC43949}	37	Sub-optimal	Ethylene glycol 15%
Pa^{ATCC43949}	28	Sub-optimal	*Pa*^{ATCC43949}	37	Sub-optimal	Ethylene glycol 20%
Pa^{ATCC43949}	28	Sub-optimal	*Pa*^{ATCC43949}	37	Sub-optimal	Sodium formate 1%
Pa^{ATCC43949}	28	Sub-optimal	*Pa*^{ATCC43949}	37	Sub-optimal	Sodium formate 2%
Pa^{ATCC43949}	28	Sub-optimal	*Pa*^{ATCC43949}	37	Sub-optimal	Urea 3%
Pa^{ATCC43949}	28	Sub-optimal	*Pa*^{ATCC43949}	37	Sub-optimal	Sodium Lactate 2%
Pa^{ATCC43949}	28	Optimal	*Pa*^{ATCC43949}	37	Sub-optimal	Sodium Phosphate pH 7 20mM
Pa^{ATCC43949}	28	Sub-optimal	*Pa*^{ATCC43949}	37	Sub-optimal	Sodium Phosphate pH 7 50mM
Pa^{ATCC43949}	28	Sub-optimal	*Pa*^{ATCC43949}	37	Sub-optimal	Sodium Phosphate pH 7 100mM
Pa^{ATCC43949}	28	Sub-optimal	*Pa*^{ATCC43949}	37	Sub-optimal	Sodium Benzoate pH 5.2 20mM
Pa^{ATCC43949}	28	Optimal	*Pa*^{ATCC43949}	37	Sub-optimal	Ammonium sulfate pH8 10mM
Pa^{ATCC43949}	28	Sub-optimal	*Pa*^{ATCC43949}	37	Sub-optimal	Ammonium sulfate pH 8 50mM
Pa^{ATCC43949}	28	Sub-optimal	*Pa*^{ATCC43949}	37	Sub-optimal	Ammonium sulfate pH8 100mM
Pa^{ATCC43949}	28	Optimal	*Pa*^{ATCC43949}	37	Sub-optimal	Sodium Nitrate 10mM
Pa^{ATCC43949}	28	Optimal	*Pa*^{ATCC43949}	37	Sub-optimal	Sodium Nitrate 20mM
Pa^{ATCC43949}	28	Optimal	*Pa*^{ATCC43949}	37	Sub-optimal	Sodium Nitrate 40mM

Pa^{ATCC43949}	28	Optimal	*Pa*^{ATCC43949}	37	Sub-optimal	Sodium Nitrate 60mM
Pa^{ATCC43949}	28	Sub-optimal	*Pa*^{ATCC43949}	37	Sub-optimal	Sodium Nitrate 100mM
Pa^{ATCC43949}	28	Optimal	*Pa*^{ATCC43949}	37	Sub-optimal	Sodium Nitrite 10mM
Pa^{ATCC43949}	28	Sub-optimal	*Pa*^{ATCC43949}	37	Sub-optimal	Sodium Nitrite 20mM
Pa^{ATCC43949}	28	Optimal	*Pa*^{ATCC43949}	37	Sub-optimal	Sodium Nitrite 40mM
Pa^{ATCC43949}	28	Sub-optimal	*Pl*^{TT01}	28	Sub-optimal	NaCl
Pa^{ATCC43949}	28	Sub-optimal	*Pl*^{TT01}	28	Sub-optimal	NaCl
Pa^{ATCC43949}	28	Sub-optimal	*Pl*^{TT01}	28	Sub-optimal	Potassium chloride 3%
Pa^{ATCC43949}	28	Sub-optimal	*Pl*^{TT01}	28	Sub-optimal	Potassium chloride 4%
Pa^{ATCC43949}	28	Sub-optimal	*Pl*^{TT01}	28	Sub-optimal	Potassium chloride 6%
Pa^{ATCC43949}	28	Sub-optimal	*Pl*^{TT01}	28	Optimal	Ethylene glycol 5%
Pa^{ATCC43949}	28	Sub-optimal	*Pl*^{TT01}	28	Optimal	Ethylene glycol 10%
Pa^{ATCC43949}	28	Sub-optimal	*Pl*^{TT01}	28	Optimal	Ethylene glycol 15%
Pa^{ATCC43949}	28	Sub-optimal	*Pl*^{TT01}	28	Optimal	Ethylene glycol 20%
Pa^{ATCC43949}	28	Sub-optimal	*Pl*^{TT01}	28	Sub-optimal	Sodium formate 2%
Pa^{ATCC43949}	28	Sub-optimal	*Pl*^{TT01}	28	Optimal	Sodium Phosphate pH 7 100mM
Pa^{ATCC43949}	28	Sub-optimal	*Pl*^{TT01}	28	Sub-optimal	Sodium Nitrite 20mM
Pa^{ATCC43949}	28	Optimal	*Pl*^{TT01}	28	Sub-optimal	Sodium Nitrite 40mM
Pa^{ATCC43949}	28	Sub-optimal	*Pl*^{TT01}	28	Sub-optimal	Sodium Nitrite 60mM
Pa^{ATCC43949}	37	Optimal	*Pl*^{TT01}	28	Optimal	NaCl
Pa^{ATCC43949}	37	Sub-optimal	*Pl*^{TT01}	28	Sub-optimal	Potassium chloride 4%
Pa^{ATCC43949}	37	Sub-optimal	*Pl*^{TT01}	28	Sub-optimal	Potassium chloride 6%
Pa^{ATCC43949}	37	Sub-optimal	*Pl*^{TT01}	28	Sub-optimal	Sodium sulfate 2%
Pa^{ATCC43949}	37	Sub-optimal	*Pl*^{TT01}	28	Sub-optimal	Sodium sulfate 3%
Pa^{ATCC43949}	37	Sub-optimal	*Pl*^{TT01}	28	Optimal	Ethylene glycol 5%
Pa^{ATCC43949}	37	Sub-optimal	*Pl*^{TT01}	28	Optimal	Ethylene glycol 10%
Pa^{ATCC43949}	37	Sub-optimal	*Pl*^{TT01}	28	Optimal	Ethylene glycol 15%

Pa^{ATCC43949}	37	Sub-optimal	*Pl*^{TT01}	28	Optimal	Ethylene glycol 20%
Pa^{ATCC43949}	37	Sub-optimal	*Pl*^{TT01}	28	Sub-optimal	Sodium formate 1%
Pa^{ATCC43949}	37	Sub-optimal	*Pl*^{TT01}	28	Sub-optimal	Urea 2%
Pa^{ATCC43949}	37	Sub-optimal	*Pl*^{TT01}	28	Sub-optimal	Urea 3%
Pa^{ATCC43949}	37	Sub-optimal	*Pl*^{TT01}	28	Optimal	Sodium Lactate 1%
Pa^{ATCC43949}	37	Sub-optimal	*Pl*^{TT01}	28	Sub-optimal	Sodium Lactate 2%
Pa^{ATCC43949}	37	Sub-optimal	*Pl*^{TT01}	28	Optimal	Sodium Phosphate pH 7 50mM
Pa^{ATCC43949}	37	Sub-optimal	*Pl*^{TT01}	28	Optimal	Sodium Phosphate pH 7 100mM
Pa^{ATCC43949}	37	Sub-optimal	*Pl*^{TT01}	28	Sub-optimal	Sodium Benzoate pH 5.2 20mM
Pa^{ATCC43949}	37	Sub-optimal	*Pl*^{TT01}	28	Optimal	Ammonium sulfate pH8 10mM
Pa^{ATCC43949}	37	Sub-optimal	*Pl*^{TT01}	28	Optimal	Ammonium sulfate pH 8 20mM
Pa^{ATCC43949}	37	Sub-optimal	*Pl*^{TT01}	28	Sub-optimal	Ammonium sulfate pH 8 50mM
Pa^{ATCC43949}	37	Sub-optimal	*Pl*^{TT01}	28	Optimal	Ammonium sulfate pH8 100mM
Pa^{ATCC43949}	37	Sub-optimal	*Pl*^{TT01}	28	Optimal	Sodium Nitrate 10mM
Pa^{ATCC43949}	37	Sub-optimal	*Pl*^{TT01}	28	Optimal	Sodium Nitrate 20mM
Pa^{ATCC43949}	37	Sub-optimal	*Pl*^{TT01}	28	Optimal	Sodium Nitrate 40mM
Pa^{ATCC43949}	37	Sub-optimal	*Pl*^{TT01}	28	Optimal	Sodium Nitrate 60mM
Pa^{ATCC43949}	37	Sub-optimal	*Pl*^{TT01}	28	Optimal	Sodium Nitrate 100mM
Pa^{ATCC43949}	37	Sub-optimal	*Pl*^{TT01}	28	Sub-optimal	Sodium Nitrite 10mM
Pa^{ATCC43949}	37	Sub-optimal	*Pl*^{TT01}	28	Sub-optimal	Sodium Nitrite 20mM
Pa^{ATCC43949}	37	Sub-optimal	*Pl*^{TT01}	28	Sub-optimal	Sodium Nitrite 40mM

Table S6: Phenoarray – PM10 pH

Species 1	Temperature 1	Respiration type 1	Species 2	Temperature 2	Respiration type 2	Substrate
Pa^{ATCC43949}	28	Optimal	*Pa*^{ATCC43949}	37	Optimal	pH 5
Pa^{ATCC43949}	28	Optimal	*Pa*^{ATCC43949}	37	Optimal	pH 5.5
Pa^{ATCC43949}	28	Optimal	*Pa*^{ATCC43949}	37	Optimal	pH 6
Pa^{ATCC43949}	28	Optimal	*Pa*^{ATCC43949}	37	Optimal	pH 7
Pa^{ATCC43949}	28	Optimal	*Pa*^{ATCC43949}	37	Optimal	pH 8
Pa^{ATCC43949}	28	Optimal	*Pa*^{ATCC43949}	37	Optimal	pH 8.5
Pa^{ATCC43949}	28	Optimal	*Pa*^{ATCC43949}	37	Optimal	pH 9.5
Pa^{ATCC43949}	28	Optimal	*Pa*^{ATCC43949}	37	Sub-optimal	pH 4.5 + L-Arginine
Pa^{ATCC43949}	28	Optimal	*Pa*^{ATCC43949}	37	Sub-optimal	pH 4.5 + L-Histidine
Pa^{ATCC43949}	28	Optimal	*Pa*^{ATCC43949}	37	Optimal	pH 4.5 + L-Tyrosine
Pa^{ATCC43949}	28	Optimal	*Pa*^{ATCC43949}	37	Optimal	pH 4.5 + L-Cysteic acid
Pa^{ATCC43949}	28	Optimal	*Pa*^{ATCC43949}	37	Sub-optimal	pH 9.5
Pa^{ATCC43949}	28	Optimal	*Pa*^{ATCC43949}	37	Sub-optimal	pH 9.5 + L-Alanine
Pa^{ATCC43949}	28	Optimal	*Pa*^{ATCC43949}	37	Sub-optimal	pH 9.5 + L-Arginine
Pa^{ATCC43949}	28	Optimal	*Pa*^{ATCC43949}	37	Optimal	pH 9.5 + L-Asparagine
Pa^{ATCC43949}	28	Optimal	*Pa*^{ATCC43949}	37	Sub-optimal	pH 9.5 + L-Aspartic Acid
Pa^{ATCC43949}	28	Optimal	*Pa*^{ATCC43949}	37	Optimal	pH 9.5 + L-Glutamine
Pa^{ATCC43949}	28	Optimal	*Pa*^{ATCC43949}	37	Sub-optimal	pH 9.5 + Glycine
Pa^{ATCC43949}	28	Optimal	*Pa*^{ATCC43949}	37	Sub-optimal	pH 9.5 + L-Histidine
Pa^{ATCC43949}	28	Optimal	*Pa*^{ATCC43949}	37	Optimal	pH 9.5 + L-Lysine
Pa^{ATCC43949}	28	Optimal	*Pa*^{ATCC43949}	37	Optimal	pH 9.5 + L-Methionine
Pa^{ATCC43949}	28	Optimal	*Pa*^{ATCC43949}	37	Sub-optimal	pH 9.5 + L-Proline
Pa^{ATCC43949}	28	Optimal	*Pa*^{ATCC43949}	37	Sub-optimal	pH 9.5 + L-Threonine
Pa^{ATCC43949}	28	Optimal	*Pa*^{ATCC43949}	37	Sub-optimal	pH 9.5 + L-Valine
Pa^{ATCC43949}	28	Optimal	*Pa*^{ATCC43949}	37	Sub-optimal	pH 9.5 + Hydroxy-L-Proline

Pa^{ATCC43949}	28	Optimal	*Pa*^{ATCC43949}	37	Optimal	pH 9.5 + L-Homoserine
Pa^{ATCC43949}	28	Optimal	*Pa*^{ATCC43949}	37	Sub-optimal	pH 9.5 + Anthranilic acid
Pa^{ATCC43949}	28	Optimal	*Pa*^{ATCC43949}	37	Sub-optimal	pH 9.5 + L-Norvaline
Pa^{ATCC43949}	28	Optimal	*Pa*^{ATCC43949}	37	Sub-optimal	pH 9.5 + Agmatine
Pa^{ATCC43949}	28	Optimal	*Pa*^{ATCC43949}	37	Sub-optimal	pH 9.5 + Cadaverine
Pa^{ATCC43949}	28	Optimal	*Pa*^{ATCC43949}	37	Optimal	X-Caprylate
Pa^{ATCC43949}	28	Optimal	*Pa*^{ATCC43949}	37	Optimal	X–α-D-Glucoside
Pa^{ATCC43949}	28	Optimal	*Pa*^{ATCC43949}	37	Optimal	X-ß-D-Glucoside
Pa^{ATCC43949}	28	Optimal	*Pa*^{ATCC43949}	37	Optimal	X-α-D-Galactoside
Pa^{ATCC43949}	28	Optimal	*Pa*^{ATCC43949}	37	Optimal	X-ß-D-Galactoside
Pa^{ATCC43949}	28	Optimal	*Pa*^{ATCC43949}	37	Optimal	X-α-D-Glucuronide
Pa^{ATCC43949}	28	Optimal	*Pa*^{ATCC43949}	37	Optimal	X-ß-D-Glucuronide
Pa^{ATCC43949}	28	Optimal	*Pa*^{ATCC43949}	37	Optimal	X-ß-D-Glucosaminide
Pa^{ATCC43949}	28	Optimal	*Pa*^{ATCC43949}	37	Optimal	X-ß-D-Galactosaminide
Pa^{ATCC43949}	28	Optimal	*Pa*^{ATCC43949}	37	Optimal	X-α-D-Mannoside
Pa^{ATCC43949}	28	Optimal	*Pl*^{TT01}	28	Optimal	pH 5
Pa^{ATCC43949}	28	Optimal	*Pl*^{TT01}	28	Optimal	pH 5.5
Pa^{ATCC43949}	28	Optimal	*Pl*^{TT01}	28	Optimal	pH 6
Pa^{ATCC43949}	28	Optimal	*Pl*^{TT01}	28	Optimal	pH 7
Pa^{ATCC43949}	28	Optimal	*Pl*^{TT01}	28	Optimal	pH 9
Pa^{ATCC43949}	28	Optimal	*Pl*^{TT01}	28	Optimal	pH 4.5 + L-Arginine
Pa^{ATCC43949}	28	Optimal	*Pl*^{TT01}	28	Optimal	pH 4.5 + L-Homoarginine
Pa^{ATCC43949}	28	Optimal	*Pl*^{TT01}	28	Optimal	pH 9.5 + L-Arginine
Pa^{ATCC43949}	28	Optimal	*Pl*^{TT01}	28	Optimal	pH 9.5 + L-Histidine
Pa^{ATCC43949}	28	Optimal	*Pl*^{TT01}	28	Optimal	pH 9.5 + L-Methionine
Pa^{ATCC43949}	28	Optimal	*Pl*^{TT01}	28	Optimal	pH 9.5 + Hydroxy-L-Proline
Pa^{ATCC43949}	28	Optimal	*Pl*^{TT01}	28	Optimal	X–α-D-Glucoside

Pa^{ATCC43949}	28	Optimal	*Pl*^{TT01}	28	Optimal	X-α-D-Glucuronide
Pa^{ATCC43949}	28	Optimal	*Pl*^{TT01}	28	Optimal	X-ß-D-Galactosaminide
Pa^{ATCC43949}	37	Optimal	*Pl*^{TT01}	28	Optimal	pH 5
Pa^{ATCC43949}	37	Optimal	*Pl*^{TT01}	28	Optimal	pH 5.5
Pa^{ATCC43949}	37	Optimal	*Pl*^{TT01}	28	Optimal	pH 6
Pa^{ATCC43949}	37	Optimal	*Pl*^{TT01}	28	Optimal	pH 7
Pa^{ATCC43949}	37	Optimal	*Pl*^{TT01}	28	Optimal	pH 8
Pa^{ATCC43949}	37	Optimal	*Pl*^{TT01}	28	Optimal	pH 8.5
Pa^{ATCC43949}	37	Optimal	*Pl*^{TT01}	28	Optimal	pH 9
Pa^{ATCC43949}	37	Optimal	*Pl*^{TT01}	28	Optimal	pH 9.5
Pa^{ATCC43949}	37	Sub-optimal	*Pl*^{TT01}	28	Optimal	pH 4.5 + L-Arginine
Pa^{ATCC43949}	37	Sub-optimal	*Pl*^{TT01}	28	Optimal	pH 4.5 + L-Histidine
Pa^{ATCC43949}	37	Optimal	*Pl*^{TT01}	28	Optimal	pH 4.5 + L-Tyrosine
Pa^{ATCC43949}	37	Optimal	*Pl*^{TT01}	28	Optimal	pH 4.5 + L-Homoarginine
Pa^{ATCC43949}	37	Sub-optimal	*Pl*^{TT01}	28	Optimal	pH 9.5
Pa^{ATCC43949}	37	Sub-optimal	*Pl*^{TT01}	28	Optimal	pH 9.5 + L-Alanine
Pa^{ATCC43949}	37	Sub-optimal	*Pl*^{TT01}	28	Optimal	pH 9.5 + L-Arginine
Pa^{ATCC43949}	37	Sub-optimal	*Pl*^{TT01}	28	Optimal	pH 9.5 + L-Aspartic Acid
Pa^{ATCC43949}	37	Sub-optimal	*Pl*^{TT01}	28	Optimal	pH 9.5 + L-Glutamic Acid
Pa^{ATCC43949}	37	Sub-optimal	*Pl*^{TT01}	28	Optimal	pH 9.5 + Glycine
Pa^{ATCC43949}	37	Sub-optimal	*Pl*^{TT01}	28	Optimal	pH 9.5 + L-Histidine
Pa^{ATCC43949}	37	Optimal	*Pl*^{TT01}	28	Optimal	pH 9.5 + L-Methionine
Pa^{ATCC43949}	37	Sub-optimal	*Pl*^{TT01}	28	Optimal	pH 9.5 + L-Threonine
Pa^{ATCC43949}	37	Optimal	*Pl*^{TT01}	28	Optimal	pH 9.5 + L-Tyrosine
Pa^{ATCC43949}	37	Sub-optimal	*Pl*^{TT01}	28	Optimal	pH 9.5 + L-Valine
Pa^{ATCC43949}	37	Sub-optimal	*Pl*^{TT01}	28	Optimal	pH 9.5 + Hydroxy-L-Proline
Pa^{ATCC43949}	37	Optimal	*Pl*^{TT01}	28	Optimal	pH 9.5 + L-Homoserine

Pa^{ATCC43949}	37	Sub-optimal	*Pl*^{TT01}	28	Optimal	pH 9.5 + Anthranilic acid
Pa^{ATCC43949}	37	Sub-optimal	*Pl*^{TT01}	28	Optimal	pH 9.5 + L-Norvaline
Pa^{ATCC43949}	37	Sub-optimal	*Pl*^{TT01}	28	Optimal	pH 9.5 + Tyramine
Pa^{ATCC43949}	37	Optimal	*Pl*^{TT01}	28	Optimal	X-Caprylate
Pa^{ATCC43949}	37	Optimal	*Pl*^{TT01}	28	Optimal	X–α-D-Glucoside
Pa^{ATCC43949}	37	Optimal	*Pl*^{TT01}	28	Optimal	X-ß-D-Galactoside
Pa^{ATCC43949}	37	Optimal	*Pl*^{TT01}	28	Optimal	X-α-D-Glucuronide
Pa^{ATCC43949}	37	Optimal	*Pl*^{TT01}	28	Optimal	X-ß-D-Galactosaminide
Pa^{ATCC43949}	37	Optimal	*Pl*^{TT01}	28	Optimal	X-α-D-Mannoside

3 Optimisation of next generation sequencing transcriptome annotation for species that lack sequenced genomes

3.1 Abstract

Next generation sequencing methods, such as RNA-seq, have permitted the exploration of gene expression in a range of organisms which have been studied in ecological contexts but without a sequenced genome. However, the efficacy and accuracy of RNA-seq annotation methods using reference genomes from related species have yet to be robustly characterised. Here we conduct a comprehensive power analysis employing RNA-seq data from *Drosophila melanogaster* in conjunction with 11 additional genomes from related *Drosophila* species to compare annotation methods and quantify the impact of evolutionary divergence between transcriptome and reference genome. Our analyses demonstrate that, regardless of the level of sequence divergence, direct genome mapping, where transcript short reads are aligned directly to the reference genome, significantly outperforms the widely used *de novo* and genome-guided assembly-based methods in both the quantity and accuracy of gene detection. Our analysis also reveals that direct genome mapping significantly reduces biases in the distribution of genes across Gene Ontology functional categories, which are often used to interpret emergent patterns in genome-wide expression analyses. Lastly, analysis of available primate RNA-seq data demonstrates the applicability of our observations across evolutionarily diverse taxa. Our quantification of annotation accuracy and transcriptome recovery decay associated with sequence divergence thus provide empirically derived guidelines for the design of future gene expression studies in species without sequenced genomes.

3.2 Introduction

Next generation transcriptome sequencing (RNA-seq) has transformed global analyses of gene expression by overcoming the limitations of microarray platforms, including most importantly transcriptional characterisation in species yet to have sequenced genomes (Wilhelm & Landry, 2009; Wang et al, 2009). These species often represent interesting ecological or behavioural model systems, where transcriptome profiling can provide valuable insights into the molecular and physiological underpinnings of complex phenotypic traits. As RNA-seq is not dependent on a predefined set of probes corresponding to a particular set of genes, as is the case with microarrays, it has been used in transcriptome profiling of species lacking sequenced genomes where transcriptome annotation is performed using the genome of a related species as a reference (Colgan et al., 2011; Esteve-Codina et al., 2011; Collins et al., 2008; Crawford et al., 2010; Kawahara-Miki et al., 2011; Künstner et al., 2010; Wolf et al., 2010; Garg et al., 2011; Dassanayake et al., 2009; Toth et al., 2007). However, when using reference genomes from a different species in transcriptome annotation, accuracy is highly sensitive to sequence divergence (Renn et al., 2004; Machado et al., 2009). For example, array-based comparative genomic hybridisation (aCGH) analyses of various drosophilid transcriptomes using *Drosophila melanogaster* as a reference results in a diminishing number of orthologous genes detected with increasing sequence divergence and loses its utility at <92% sequence identity, even if correction procedures are applied (Renn et al., 2010). A systematic characterisation of the accuracy of using reference genomes in transcriptomic studies is therefore required.

Annotation of RNA-seq transcriptome data for species that lack sequenced genomes has been carried out using approaches already employed in the annotation of transcriptomes from species with sequenced genomes. These strategies generate a reference transcriptome by assembling raw short transcript reads which are then annotated by homology searching against annotated sequences. Reference transcriptomes are assembled primarily in two ways: (a) using genome or transcriptome sequences from a closely-related species as a guide (‘reference sequence-guided’ transcriptome assembly), or (b) carrying out a reference sequence-independent assembly (‘*de novo*’ transcriptome assembly) (Garber et al., 2011). Both of these strategies suffer from a significant reduction in the proportion of the transcript sequences that have homology to the reference genome of related species as sequence divergence increases (Shi et al., 2011; Colgan et al., 2011; Kawahara-Miki et al., 2011; Balakrishnan et al., 2013; Moghadam, et al., 2013). Furthermore, it is not clear whether contigs produced from the assembly process are accurately assigned to the corresponding orthologous sequence to which a contig would have been assigned if the genome of the same species had been available. Given the relevance of comparative gene expression analyses in exploring the molecular basis and evolution of biological traits, it is of paramount importance to maximise gene detection rates while minimising gene identification errors inherent to transcriptome characterisation before analysing and interpreting expression profiles.

The accuracy of transcript-to-gene assignment obtained with different transcriptome assembly and annotation strategies when using divergent reference genomes has only recently been addressed. Hornett & Wheat (2012) explored the efficacy of gene recovery using *de novo* assembly construction approaches applied to longer 454 reads and shorter Illumina reads with increasingly divergent reference genomes. As expected, increasing divergence resulted in an increased rate of error and the extent of functional bias in the recovered transcriptome (Hornett & Wheat, 2012). However, they concluded that the use of reference genomes of up to 100mya divergence were still suitable for transcriptome annotation in species with no sequenced genome. Vijay et al. (2013) explored transcriptome annotation in non-model species by comparing the performance of *de novo* assemblies constructed from simulated transcriptome reads against a 'mapping assembly' approach where transcript consensus sequences were obtained from simulated reads aligned to reference genomes with a range of divergence levels (Vijay, et al., 2013). They found that, when considering the proportion of the reference transcriptome recovered in the assemblies, mapping assemblies performed better than *de novo* assemblies with up to 15% sequence divergence, including a minimal reduction in accuracy. Furthermore, when assigning gene IDs to assembled contigs, all assembly types exhibited increasing error with increasing sequence divergence and the use of a subset of tissue-specific genes resulted in misassignment even in the absence of divergence. Lu et al. (2013) compared *de novo* and genome-guided assembly methods and demonstrated substantial variability in the performance of different tools. For example, they find that these methods are comparable in terms of the completeness of assembled transcripts, but genome-guided assemblies perform better regarding contiguity (proportion of known transcripts covered by a transcribed sequence fragment), while *de novo* assemblers perform better in terms of generating fewer chimeric transcripts and in variant resolution. However, these studies did not assess the performance of transcriptome annotation using a simpler method of directly mapping reads to the reference genome, bypassing assembly of reads into contigs.

These previous studies indicate that assembly quality can be highly variable and that simpler mapping-based methods can be highly effective for gene detection in non-model species. A direct approach, where transcript short reads are mapped to the genome sequence of the closest available reference species, might be superior in detecting genes relevant to a particular trait of interest. Such an approach aims to provide the fullest possible complement of genomic information of relevance to a closely related species, including regions no longer expressed in the reference species that would hence be absent from its transcriptome, while avoiding potential biases that may stem from the inherently variable nature of transcriptome content (Sims et al., 2014). It is noteworthy that this approach is currently used by 'splicing aware' alignment tools such as TopHat (Trapnell et al., 2009) to accurately locate splice junctions via gapped alignment prior to transcript assembly. However, a comparison of *de novo* and guided transcriptome assembly methods, using reference

genomes of varying levels of divergence, remains to be performed. Furthermore, given that a substantial portion of transcriptome sequences may map to multiple genomic locations, there is uncertainty in the accuracy of their annotation. For gene expression studies, it has been noted that multi-match reads should be included to provide more representative expression profiles (Mortazavi *et al.* 2008). Some annotation tools have been developed to help deal with these problematic sequences, such as ERANGE (Mortazavi et al, 2008), BM-MAP (Ji *et al.* 2011), RSEM (Li *et al.* 2010), and SeqEm (Paşaniuc *et al.* 2011). However, error rates associated with sequences mapping to single versus multiple locations, and how these rates are impacted by sequence divergence and annotation method, have yet to be determined. Despite this, multi-match sequences are often incorporated into transcriptome analyses to increase the quantity of annotated transcripts and genes detected despite the uncertainty about the bias this may introduce (Mortazavi et al., 2008; Brawand et al., 2011).

In the present study we quantitatively assess the impact of sequence divergence between transcriptome and reference species on the performance of a range of next generation transcriptome annotation strategies. Using published RNA-seq data from *Drosophila melanogaster* and genome sequences for 12 *Drosophila* species, the efficacy of two widely used transcript annotation strategies, reference genome-guided assembly and *de novo* assembly, and a direct genome mapping method which bypasses transcriptome assembly are compared for the first time. The accuracy of gene detection using transcript sequences aligned to single versus multiple locations and biases in gene function categories associated with each annotation methodology are assessed. Lastly, RNA-seq data from four primate species are used to confirm the generality of these findings. Our results clearly demonstrate in multiple taxa that the power to accurately recover genes detected as expressed from RNA-seq data is significantly impacted by the level of divergence between transcriptome and reference species and, more importantly, the annotation method used. We find that, regardless of the level of sequence divergence, direct genome mapping significantly outperforms *de novo* and genome-guided assembly-based strategies in both the quantity and accuracy of gene detection. As such, these results present guidelines for the design of future studies in species without sequenced genomes.

3.3 Materials and Methods

3.3.1 Drosophila genome sequences and orthology annotations

Genome releases for *Drosophila melanogaster* (Adams 2000) and 11 additional *Drosophila* species (Richards et al., 2005; Clark et al., 2007) and orthology relationships were obtained from Flybase (www.flybase.org). See Table S1 for genome releases. Nested and/or overlapping genes were removed from all analyses.

3.3.2 RNA-seq data download and pre-processing

D. melanogaster transcriptome Illumina derived short reads were downloaded from the modENCODE database (www.modencode.org, data set 2027: The modENCODE Consortium et al., 2011). Short reads (n = 9,663,442) were pre-processed in the Penn State Galaxy server (http://galaxyproject.org; Goecks et al., 2010; Giardine et al., 2005). Reads were groomed into fastqsanger format, sequencing artefacts were removed, and the remaining read set was quality filtered using the following criteria: each base was required to satisfy a minimum PHRED quality score of 20, equating to approximately a 1% error rate, allowing less than 10% of the read length (3 bases of 36 base reads) with quality scores below this (Crawford et al., 2010;Cloonan et al., 2008). This left 6,863,396 reads remaining (71.02% of the original read set).

3.3.3 Transcriptome annotation through assembly-based methods

A reference genome-guided assembly and a *de novo* assembly were generated using software packages Velvet Columbus (Zerbino & Birney, 2008; Zerbino, 2010) and Velvet Oases (Schulz et al., 2012), respectively. An additional *de novo* assembly was produced using Trinity (Grabherr *et al.* 2011). All assemblies were performed by Dr. Lauren O'Connell, Harvard University. All necessary alignments between pre-processed *D. melanogaster* reads and the 12 annotated *Drosophila* genomes were performed using the gapped short read alignment mapping program, SHRiMP (Rumble et al., 2009; David et al., 2011) with default parameters, outputting all unaligned reads to the alignment file. For both Velvet assemblies, a multiple k-mer approach (k= 23, 25, 27, 29, 31, 33) was used and the *mergeAssembly* function was used to merge the multiple kmers. Then, CD-HIT-EST (Li & Godzik, 2006) was used to remove contig redundancy that can occur by merging multiple assemblies. Given that redundant contigs can represent alternative spice variants, polymorphisms among the pooled individuals, or sequencing errors, a conservative threshold of 98% sequence similarity was used. All contigs below 100bp were removed as likely artefacts of the merging and clustering process.

All assemblies were subjected to homology searching using Blast v2.2.26+ (Altschup et al., 1990), with threshold value E = $1e^{-10}$, and local alignment against chromosomal databases, as these performed better than coding sequence (CDS) or exon data bases (data not shown). Significant hits were then verified using the Smith–Waterman algorithm (fasta36.3.5d with parameters –a –A) (Pearson 2000). Homology searching and verification were performed by Dr. Stephen Bush (Urrutia group, University of Bath).

3.3.4 Direct genome mapping (DGM) transcript annotation

Processed *D. melanogaster* reads were sequentially aligned against each of the 12 *Drosophila* genomes using SHRiMP. Alignments were generated using default parameters, and reads were subsequently assigned to genes based on alignment coordinates using custom Python scripts. Alignments were not filtered by mapping quality.

3.3.5 Assessment of annotation accuracy

Transcript sequences were segregated into those that mapped to single locations within the genome (unassembled reads) or single genes (assembled contigs), termed single-match sequences (often referred to as uniquely mapped sequences), and those that mapped to multiple locations/genes, termed multi-match sequences. Multi-match sequences were filtered to only contain those matches where, for a given transcript sequence, there was a demonstrably higher scoring alignment ('top hit'). All other multi-matching transcripts were excluded from further analyses as they could not be unambiguously assigned to a specific location. Error rates in the annotation of transcript sequences mapping to genes were assessed separately for the two groups, using increasingly divergent reference genomes with the *D. melanogaster* transcript annotation as benchmark. Each transcript sequence was then classed as correctly assigned if it was assigned to the corresponding orthologue detected by that sequence in *D. melanogaster*, or it was classed as incorrectly assigned if it did not map to the same orthologue detected by *D. melanogaster*. Gene detection accuracy was benchmarked using orthologous gene sets detected using alternative genomes compared to those identified using the *D. melanogaster* genome. Gene detection accuracy was calculated as the proportion of orthologous genes correctly identified using each alternative genome, out of the total number identified using that alternative genome.

3.3.6 Gene functional classification

Genes detected by single-match reads in *D. melanogaster* were assigned to GO slim categories (gene associations [CVS revision 1.220, GOC validation date 24/01/2012] and generic GO slim terms [CVS revision 1.864, dated 15/08/2011] obtained from the Gene Ontology Consortium: The

Gene Ontology Consortium, 2000). Species were grouped according to their varying levels of divergence from *D. melanogaster* (Clark et al., 2007; Tamura et al., 2004): *D. sechellia* and *D. simulans*, *D. erecta* and *D. yakuba*, *D. pseudoobscura* and *D. persimilis*. Orthologous genes detected by single-match reads for each group were combined into a single list and then assigned to GO slim categories. The observed distribution of these orthologous genes in the combined list to GO slim categories were compared to the expected distribution and the fold change from the expected values was calculated. Expected values were calculated using the proportion of *D. melanogaster* genes detected overall, assuming that the reduction in this would be reflected in the same proportional decrease in the number of genes assigned to each GO slim category.

3.3.7 Primate RNA-seq

Genome sequences and gene annotations for human and four additional primate species (chimpanzee, gorilla, orangutan and macaque) were downloaded from Ensembl (www.ensembl.org; Flicek et al., 2014). Orthology annotations were obtained from Brawand et al. (Brawand *et al.* 2011). Publicly available single-end human RNA-seq data was downloaded from NCBI Sequence Read Archive (www.ncbi.nlm.nih.gov/sra, sample ERS045944). The reads were filtered for sequencing artefacts and subsequently quality filtered to the same stringency as the *D. melanogaster* short reads (minimum score of 20 with 10% of the read length [5 bases of 50 base reads] allowed below this), reducing the total number of reads from 29,849,485 to 5,025,987. These were then sequentially directly mapped to each genome. Single-match mapped reads were extracted and annotated as above. Miss Holly Barnes (Urrutia group, University of Bath) performed all primate DGM. Accuracy was calculated using single-match human reads assigned to the human genome as benchmark.

3.3.8 Statistical analysis

All statistical tests were performed in R (R: A Language and Environment for Statistical Computing, 2010). Shapiro-Wilk tests were used to test for normal data distributions. When all data sets within a given comparison were normally distributed, t tests and F tests were used to test for statistically significant differences in data means and variances, respectively. Where data sets were not consistently normally distributed, comparisons were performed using Mann-Whitney (two-sample Kruskal-Wallis) tests.

3.4 Results

3.4.1 Differential impact of sequence divergence on transcript mapping

When using either the *D. melanogaster* genome or increasingly divergent genomes from other species, transcript assembly-based methods give very similar results to each other with up to 90% of contigs matching to annotated genes while only up to 35% of reads matched to annotated genes using DGM (Fig 1 and Fig. S1). Furthermore, all methods display a significant reduction in the proportion of mapped sequences with increasing levels of divergence (Fig. 1 and Fig. S1). When considering single-match and multi-match sequences, substantial differences in the relative proportions of sequences that are assigned to single versus multiple targets were identified (Fig. S1). DGM provides a significantly lower proportion of single-match sequences (DGM versus each assembly method, $1.733e^{-07} \leq p \leq 0.007$), and a significantly higher proportion of multi-match sequences than any of the assembly methods (DGM versus each assembly, $7.878e^{-06} \leq p \leq 0.007$). This is expected as assembled contigs are significantly longer (data not shown) and thus have greater specificity. In all cases, the vast majority (95-98%) of sequences correctly assigned to orthologous genes are those that map to a single genomic location.

3.4.2 DGM identifies more genes than alternative strategies

DGM recovered over twice as many genes than any of the assembly-based methods (Fig. 2A and Table S2): 11,173 genes were detected using DGM (10,914 using single-matches), whereas 3,722 genes were detected by both Velvet genome-guided and *de novo* assembly (3,544 and 3,578 using single-matches, respectively), and 2,360 genes were detected by Trinity *de novo* assembly (2,285 using single-matches). Additionally, DGM recovered a significantly higher proportion of orthologous genes than any of the assembly methods for each of the 11 genomes analysed ($1.308e^{-07} \leq p \leq 3.041e^{-06}$; Fig. 2B). In contrast, the assembly strategies displayed poorer performance, detecting only 20-35% of orthologous genes in low divergence genomes (*D. ananassae* and more closely related species) and as low as 10% in the more divergent species (Fig. 2B). These results indicate that DGM (a) identifies more genes when using the same genome as a reference and (b) displays superior performance across increasingly divergent genomes, despite the reduction in reads that are mapped in comparison to assembly-based approaches.

3.4.3 Increased accuracy of DGM in gene detection

As expected, all annotation strategies were associated with higher error rates with increasing sequence divergence and multi-match transcripts were associated with markedly higher error rates compared to single match transcripts (Fig. 3 and Figs. S2 and S3). Nonetheless, DGM provided the lowest error rate (6-10%) in gene detection across all species tested, and for both single-match (Fig.

3A; $4.76x10^{-4} \leq p \leq 0.012$) and multi-match sequences (Fig. 3B; $4.05x10^{-4} \leq p \leq 0.0004763$). Error rates were substantially higher using both *de novo* assemblies (11-16%) or the guided-genome assemblies (~30%). It is noteworthy that filtering alignment scores and read counts per gene resulted in marginal improvements in DGM accuracy although this vastly compromised the number of genes detected (Fig. S4), hence using all alignment results is recommended for optimal gene detection. These findings indicate that DGM is significantly more accurate than genome-guided or *de novo* assembly when a corresponding reference genome sequence is unavailable and that this effect is particularly enhanced for multi-match sequences.

3.4.4 DGM is associated with minimal functional biases in resulting transcriptome annotations

Due to the non-uniformity of evolutionary rates, transcriptome annotation accuracy using diverged genomes is expected to suffer for rapidly evolving genes (Le Quéré *et al.* 2006) and have a pronounced effect on related gene ontology analyses. For example, housekeeping genes tend to evolve more slowly (Duret & Mouchiroud, 2000; Lercher et al., 2004; Zhang & Li, 2004) whereas immune and reproductive genes evolve at a faster rate (see Dorus et al., 2010). All annotation methods were associated with a functional profile bias but DGM resulted in substantially less bias than the other annotation methods, particularly at higher levels of divergence (Fig. 4A) while detecting more GO slim terms (Fig. 4B). Importantly, analysis of the mean error scores for genes detected per GO slim term revealed substantial variance across GO slim terms (Fig. S5). As expected, GO terms representing highly conserved functions, such as translation, chromosome organisation, and response to stress, display consistently low error levels (Table S3 gives terms with zero error). Similarly, several GO terms associated with rapidly evolving processes, such as reproduction (Dorus *et al.* 2010) and mRNA processing (Marz et al., 2008), exhibit consistently high error rates in both *Drosophila* and primates (see below) (Table S4).

3.4.5 Corroborating DGM performance in alternative taxa

. Compared to *Drosophila*, the proportion of single-match reads mapping to orthologous genes is slightly lower in primates (50% versus 79% at the lowest level of divergence for primates and *Drosophila*, respectively) as is the proportion of orthologous genes detected (Fig. 5). This is likely to be accounted for in part by the greater amounts of repetitive sequences in the primate genomes (Liu et al., 2003) and the higher proportion of genes within large, highly homologous gene families. However, the proportions of single-match reads and genes detected across increasingly divergent genomes are well maintained. This is perhaps expected given the low range of divergence amongst the primate genomes analysed but is nonetheless informative in choosing an alternative reference genome for transcriptome analysis by DGM. Error rates for gene detection are also lower for primates compared to *Drosophila*, a finding which we attribute to the longer read

length in primates (primate = 50nt; *Drosophila* = 36nt). Lastly, functional bias in gene detection, as was the case with *Drosophila*, was determined to (a) have a positive relationship with divergence, and (b) to vary across functional terms (Fig. S6). Consistent with low divergence in these primate genomes, GO slim terms detected in primates exhibited very low error, especially within those with highly conserved functions (Fig. S7).

3.5 Discussion

Recent studies have begun to explore the impact of sequence divergence between study species and reference species on next generation transcriptome annotation (Lu et al., 2013; Vijay et al., 2013; Hornett & Wheat, 2012). However, none have compared transcriptome assembly with a simple process where reads are mapped directly to the reference genome. Also, it is unclear whether, when using reference sequences from a different species, whether using RNA-seq reads that map uniquely versus to multiple locations is informative or indeed accurate. We have conducted a systematic performance comparison of two assembly-based methods and direct read-to-genome mapping (DGM) when applied to the annotation of transcriptome data for species without sequenced genomes. *D. melanogaster* Illumina reads were annotated using 11 other *Drosophila* genomes to measure efficiency and accuracy as a function of nucleotide divergence, with key findings validated using primate species. We found that DGM is substantially more effective and efficient in gene recovery both when transcriptome and reference sequence are the same or divergent. Specifically, DGM detects over twice the number of genes, and concurrently far more functional gene categories, than the best assembly-based methods in the absence of divergence, its superior performance increasing with divergence. Importantly, we were able to benchmark gene annotation accuracy and assess bias in the detection of gene functional categories: DGM displayed the highest accuracy in gene detection and the lowest functional bias across wide ranges of divergence. This indicates that DGM is more robust at detecting the functional complexity of transcriptome profiles when there is divergence between transcriptome and reference species, and demonstrates that studies aiming to characterise novel transcriptomes should benefit from this powerful and relatively error-free technique compared to assembly-based methods. To help inform the design of future comparative functional genomics studies aiming to use multiple transcriptome/reference species with differing divergence levels between them, we related functional gene category detection error to divergence. Error differs according to the term with many showing stable error across the divergence levels tested. Similar trends are observed with primate data and comparing the two lineages, categories with consistently high (reproduction, biosynthetic process, and mRNA processing) or low (translation, chromosome organisation, and response to stress) error can be observed. Using a reference species with the lowest possible nucleotide divergence from the transcriptome species, and only utilising single-match reads from DGM is therefore recommended for gene detection studies.

As expected, decreased gene detection, increased gene detection error and functional bias with increasing divergence of the reference genome was observed with all annotation methods tested. This is consistent with similar previous studies using assembly-based transcriptome annotation methods (Lu et al., 2013; Vijay et al., 2013; Hornett & Wheat, 2012). Indeed, the trends of gene detection and transcript assignment error with increasing divergence of our primate results recapitulates Hornet and Wheat's (2012) findings using assembled primate sequences. It is worth

noting, however, that these studies did not assess the performance of transcriptome annotation using direct read-to-genome mapping which bypasses assembly of reads into contigs. Lu et al. (2013) advocated an approach integrating aspects of genome-guided and *de novo* assembly methods when there is no sequence divergence between transcriptome and reference species. Our findings do not support this: when comparing *de novo* with guided assembly methods, though these approaches performed comparably for the quantity of genes detected, both our *de novo* assemblies performed significantly better than the genome-guided assemblies regarding accuracy of gene detection and transcript assignment across large evolutionary distances. Interestingly, despite detecting different numbers of genes when transcriptome and reference species are the same (Velvet/Oases: 3,992; Trinity: 2,886), the two *de novo* assemblies showed similar levels of accuracy with increasing divergence. This indicates that the method of *de novo* assembly is more appropriate for annotating transcriptomes using alternative reference sequences by generating contigs that, when combined, are more representative of expressed transcripts. This is likely to be due to species-specific variation in transcriptome structure and complexity: assembly methods that reply on reference sequences, such as genome-guided assembly, may not cope as well with species-specific differences in transcriptome structure as reference sequence-independent methods, such as *de novo* assembly. Although many transcripts, or large portions of transcripts, may be conserved between closely related species, it may be that genome-guided transcriptome assembly tools underestimate the differences in transcript structure and diversity across small evolutionary distances, overly relying on reference sequence structure, which is avoided in *de novo* assembly. Indeed, the poor performance of the assemblies overall compared to DGM may be related to the introduction of computational steps in generating longer transcript sequences. Where DGM utilises the unadulterated sequences produced directly from the sequencing platform, assemblies reply on artificial construction of transcripts. Most recent assemblers use de Bruijn graphs to construct transcripts, a method which was developed for Illumina/Solexa sequences (compared to the overlap-layout-consensus approach used by Roche 454 sequence assemblers, see Martin & Wang 2011; Ren *et al.* 2012). Interestingly, Trinity was recently found to outperform other de Bruijn assemblers and was also found to perform well on 454 data (Ren *et al.* 2012). In that study, the authors report that there can be significant variations in the performance of each algorithm on the accuracy and efficacy of transcript construction. This indicates that assembly parameters exert a large effect on performance and thus warrant careful consideration. Importantly, the authors report that using a multiple k-mer approach as used in this study, which improves assembly sensitivity, also results in reduced specificity by producing overlapping, redundant transcripts. This may explain some of the discrepancies we observe between the performance of DGM and the *de novo* assemblies: if many overlapping contigs were indeed generated, which mapped to orthologous genes, these may have been considered as multi-match contigs and hence removed from the analyses. This would have reduced the proportion of genes detected and may have impacted on the reported accuracy (Ren *et al.* 2012).

The relative number of transcripts produced by each assembly tool (Velvet/Oases: 14,555; Trinity: 5,026) suggests that Trinity may be more efficient at constructing valid, longer-length transcripts.

We further show that transcript sequences mapping to a single location or gene are far more accurate than the top-scoring hits of multi-match sequences. Our results demonstrate that the inclusion of multi-match reads, or indeed contigs that map to multiple genes, in any transcriptome study of a species lacking a sequenced genome would introduce high levels of error in both transcript sequence assignment and gene detection and hence should be avoided, especially if the divergence between transcriptome and reference species is high. One way to incorporate these approaches, potentially bolstering the gene expression profile to a more representative degree without compromising excessively on accuracy, may be to obtain the list of genes identified by single-match reads and subsequently incorporate only those multi-match reads that aligned to genes in that list. As sequencing technologies improve, increasing read length will aid reduction in multi-match reads, thereby reducing ambiguity and enabling transcriptomic analysis of a wider repertoire of organisms, with potentially greater evolutionary divergence between themselves and the closest available annotated reference species.

Previous microarray studies using multiple transcriptome/reference species pairs from various taxa have highlighted a key issue: when comparing the transcriptomes of species lacking sequenced genomes that have been annotated using a related genome sequence, the gene lists identified and subsequently compared need to be standardised (Renn et al., 2010; Machado et al., 2009). Our results in both *Drosophila* and primate species not only reiterate this issue, highlighting how the choice of annotation strategy influences the degree of function bias, but also demonstrate that common functional categories suffer similarly from gene detection error induced by divergence. This may reflect certain gene categories being associated with similar rates of sequence divergence across metazoan lineages. Particular terms are observed to consistently present relatively high error rates in both *Drosophila* and primate species, such as reproduction, biosynthetic process and mRNA processing. This may be explained by a number of factors, including comparatively high rates of evolution (Dorus *et al.* 2010), gene duplication and rapid synteny changes (Marz et al., 2008) operating on such types of genes, but also lineage-specific changes in exon usage via differentially regulated alternative splicing (Blekhman et al., 2010). However, this may also be contributed to by inconsistent gene ontology annotations across the range of species used (Khatri & Drăghici 2005), particularly where the gene ontology annotations include multiple terms.

We expect that our findings regarding gene detection capabilities and error, functional bias, and the manner with which these are exacerbated with increasing nucleotide divergence, will aid the

interpretation of transcriptome annotation using species lacking sequenced genomes in comparative analyses, particularly where multiple species pairs are to be compared. The linear relationships of both gene detection and gene detection error with sequence divergence enables us to illustrate thresholds of divergence between transcriptome and reference species below which gene detection is maximised while gene detection error remains low (Fig. 6). We observe a similar pattern with the primate data where the crossover between detection and error is shifted to a lower region of sequence divergence. This is most likely due to the low sequence divergence between the species used and relatively few data points (data not shown). Together, this illustrates how factors such as divergence between transcriptome and reference species, and genomic features such as complexity, repetitive sequence, and gene length, can impact on the power to accurately recover genes. Additionally, as high levels of divergence can lead to the enrichment of slow evolving, highly conserved genes over-powering depleted fast-evolving genes, using multiple transcriptome/reference species pairs with varying degrees of divergence between them may lead to non-comparable results. As such, the selection of species with no available genome sequence should be based on the availability of the closest possible reference sequence and consider the knowledge base surrounding that reference sequence. For those wishing to extract further information than the genes detectable from transcriptome sequences, one suggestion might be to establish gene lists using single-match reads from DGM, and subsequently restrict the gene models used to *de novo* assemble transcripts to those within the DGM list. This would deliver the gene detection accuracy and minimal functional bias of DGM with *de novo* transcript assembly specificity. Using only those assembled transcripts that map to single genes would provide further confidence in their accuracy.

3.6 Conclusions

Our results demonstrate that, compared to the conventionally used assembly based methods *de novo* and genome-guided assembly, DGM has superior performance when applied to RNA-seq short read transcriptome data annotation from a species lacking a sequenced genome using the annotated reference sequence from a closely-related species. Importantly, DGM is also associated with the greatest accuracy over large evolutionary distances and recovers a more representative functional profile (as assessed by GO slim categories) of genes than the other strategies.

Compared to the assembly-based methods, DGM is a very simple process to employ: it requires few steps and a small amount of ‘hands-on’ time to implement – it requires no optimisation, except for establishing the user’s preferred levels of short read pre-processing and alignment parameters, and no subsequent homology searching. Together, our findings pave the way for the utilisation of a wide variety of non-model species in transcriptome studies where the closest available reference species is not necessarily a very close relative with minimal loss of gene detection and error rates.

3.7 Figure legends

Fig. 1. Negative relationship between orthologous sequence mapping and divergence. The proportions of pre-processed *D. melanogaster* RNA-seq sequences (unassembled reads or assembled contigs) assigned to orthologous genes when mapped to each of the alternative *Drosophila* genomes were plotted against sequence divergence for the following annotation strategies: DGM (stars), genome-guided assemblies using Velvet Columbus (open diamonds), *de novo* assembly using Velvet Oases (inverted open triangles), and *de novo* assembly using Trinity (filled black circles). "% sample" indicates the percentage of the total number of sequences for each annotation strategy thatcwere assigned to orthologues. Data points for the Velvet-based genome-guided and *de novo* assemblies overlap to a high degree.

Fig. 2. DGM detects more genes than alternative assembly methods. The efficacy of each transcriptome annotation strategy at recovering genes was assessed using the same reference species and different reference sequences at increasing levels of sequence divergence. (A) Total numbers of genes detected by each strategy (whole bars) when *D. melanogaster* RNA-seq sequences are annotated using its own genome: DGM: direct genome mapping; GGV: genome-guided assemblies using Velvet Columbus; DNV: *de novo* assembly using Velvet Oases; DNT: *de novo* assembly using Trinity. Genes detected by single-match sequences are indicated by wide striped sections. (B) The proportion of orthologous genes that are detected at increasing levels of sequence divergence by direct genome mapping (stars), genome-guided assemblies (diamonds), and *de novo* assembly using Velvet Oases (inverted triangles) or Trinity (filled circles).

Fig. 3. Direct genome mapping results in lower gene detection error than alternative assembly methods. (A) The proportion of orthologous genes detected incorrectly by single-match sequences (unassembled reads or assembled contigs) is the lowest for direct genome mapping, compared to the assembly methods. (B) The proportion of orthologous genes detected incorrectly by multi-match sequences is the lowest for direct genome mapping, compared to the assembly methods. Single-match sequences display significantly lower gene detection error compared to multi-match sequences. Results for direct genome mapping (stars), genome-guided assemblies (diamonds), *de novo* assembly using Velvet Oases (inverted triangles), and *de novo* assembly using Trinity (filled circles) are indicated.

Fig. 4. Direct genome mapping introduces less functional gene category bias and detects more functional gene categories than assembly-based methods. Bias in GO slim term detection was assessed by calculating the log(2) fold change between observed and expected values for numbers of genes assigned to each GO slim term. Expected values for the numbers of genes assigned to each GO slim term were generated assuming a linear loss of genes per term with increasing divergence.

Mean bias was plotted for *D. erecta* and *D. yakuba* using DGM (solid line), genome-guided assemblies (dashed line), *de novo* assembly using Velvet Oases (dotted line) and Trinity (dot-dashed line). (B) The number of GO slim terms detected using the *D. melanogaster* genome (light hatching) and the four most divergent species (*D. willistoni*, *D. mojavensis*, *D. virilis*, and *D. grimshawi*, dense hatching) by DGM, genome-guided assembly (GGV), *de novo* assembly using Velvet (DNV), and *de novo* assembly using Trinity (DNT) are shown. DGM detected more GO slim terms and demonstrated less of a drop in term detection with high levels of divergence than other assembly strategies.

Fig. 5. DGM using primate sequences generates similar trends to *Drosophila* sequences. (A) The proportion of human orthologues detected correctly by single-match reads when human RNA-seq reads were mapped to the alternative non-human primate genomes (solid triangles) was marginally less than the proportion of genes detected overall by single-match sequences (open triangles), displaying a slower reduction with increasing divergence. (B) Read assignment error shows a similar trend to the *Drosophila* alignments but is lower overall.

Fig. 6. Considering gene detection capabilities and read assignment error, limits of sequence divergence between transcriptome and reference species can be inferred. Trend lines for gene detection and read assignment error as a function of sequence divergence intersect at approximately 2.0 substitutions per site, indicating that below this level gene detection significantly outweighs read assignment error, hence ensuring confidence in gene recovery.

Fig. 1

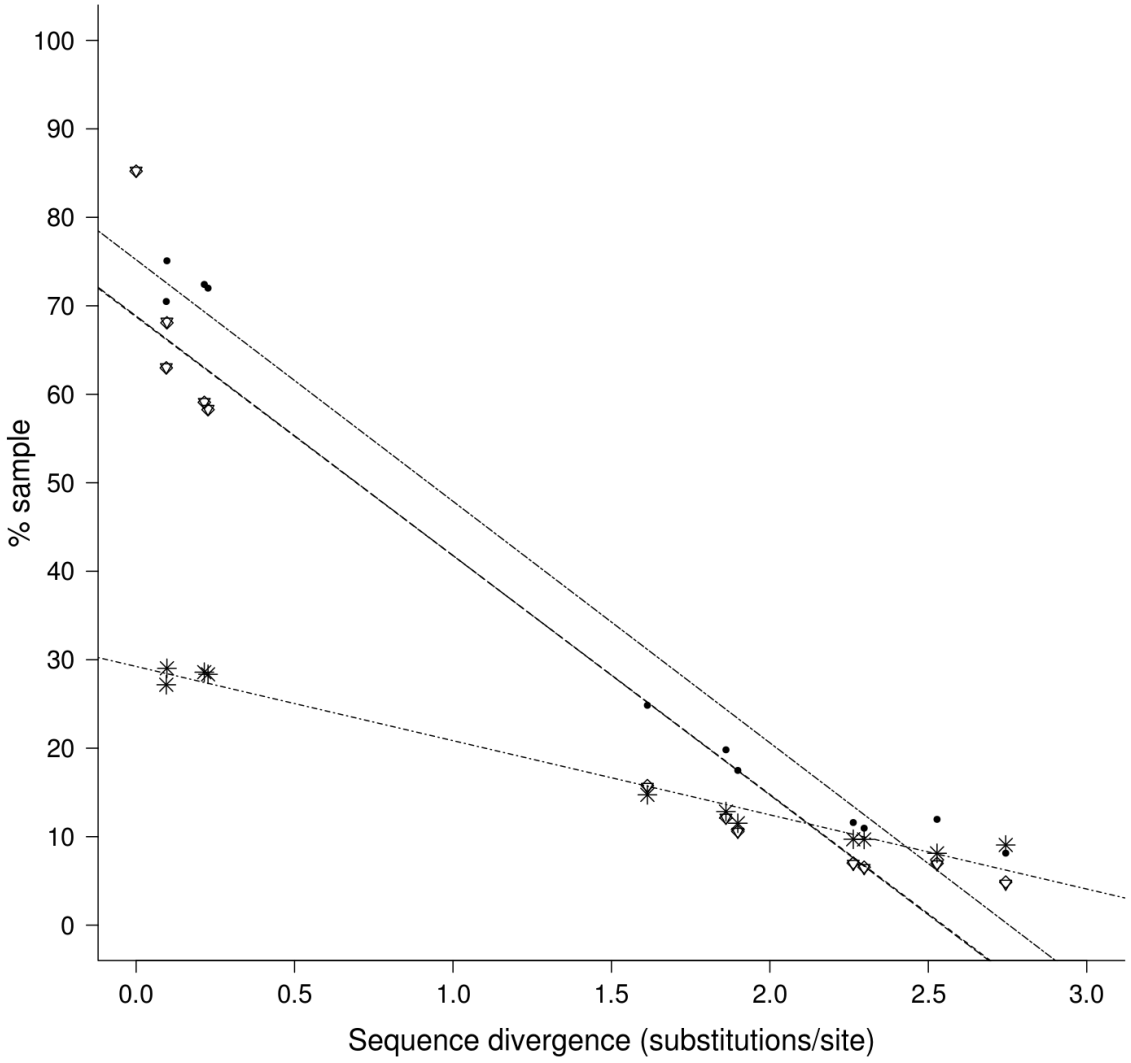

100
90
80
70
60
50
40
30
20
10
0
% sample
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Sequence divergence (substitutions/site)

Fig. 2

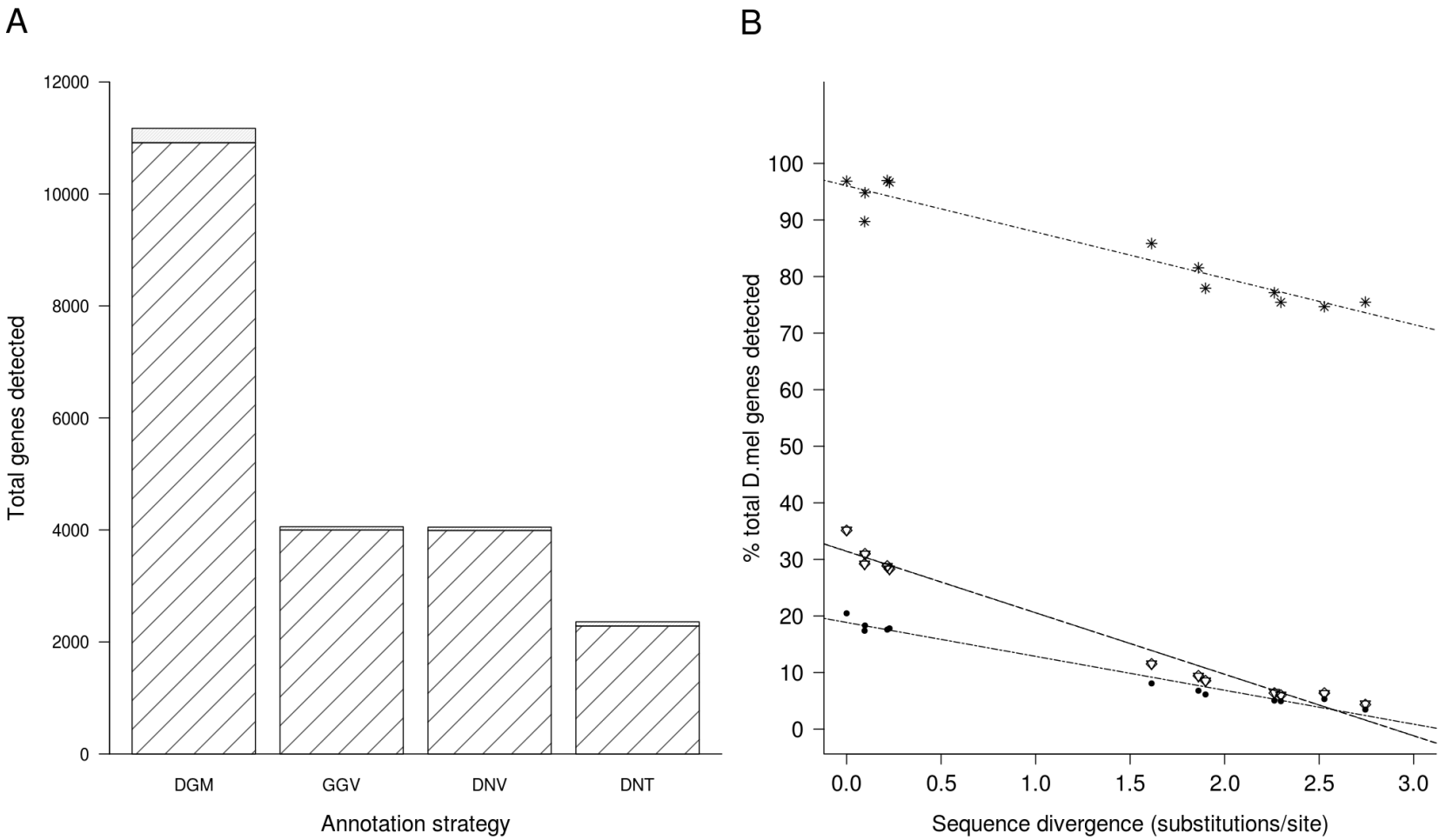

A
12000
10000
8000
6000
4000
2000
0
Total genes detected
DGM
GGV
DNV
DNT
Annotation strategy
B
100
90
80
70
60
50
40
30
20
10
0
% total D.mel genes detected
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Sequence divergence (substitutions/site)

Fig. 3

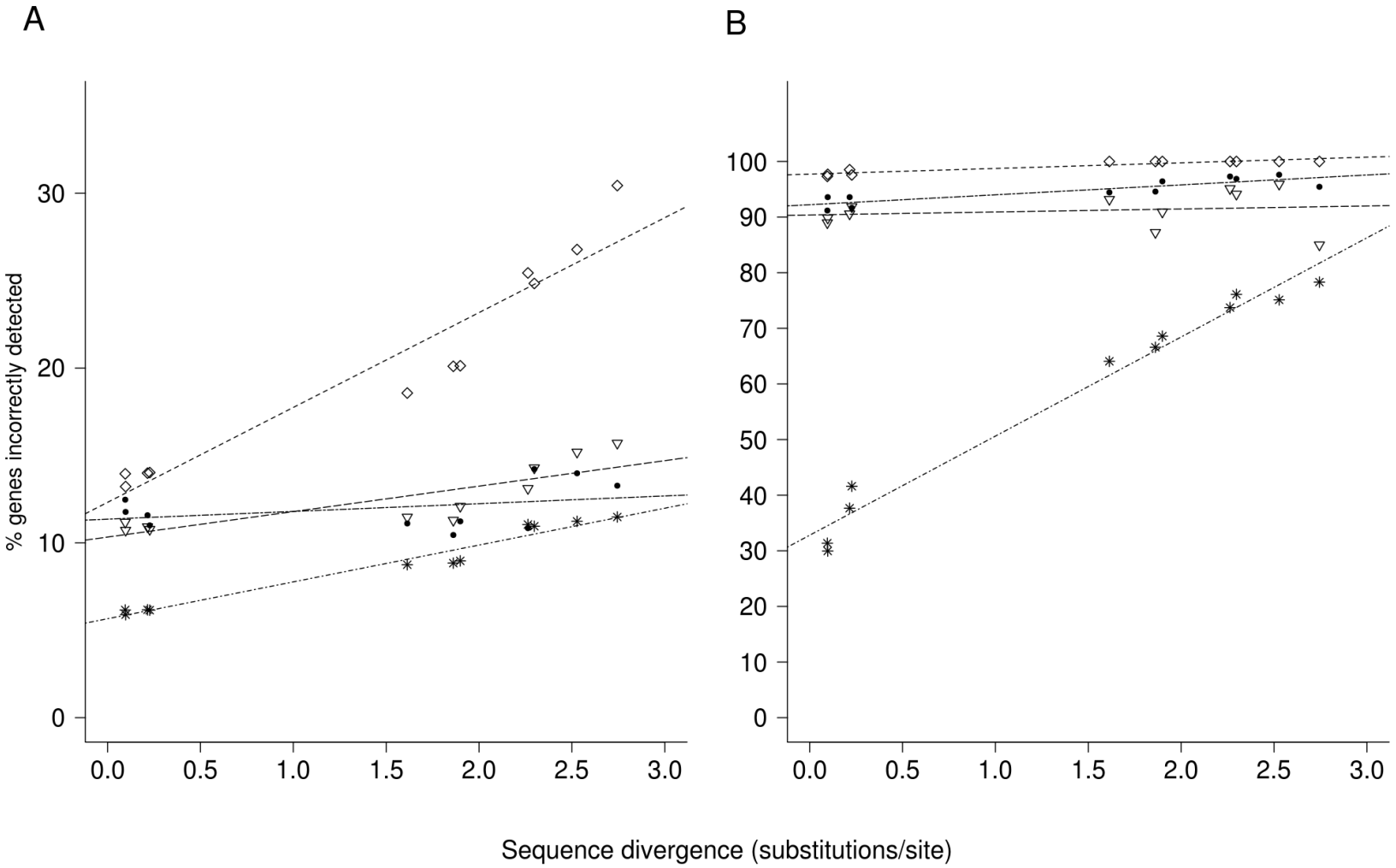

A
B
% genes incorrectly detected
0
10
20
30
0
10
20
30
40
50
60
70
80
90
100
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Sequence divergence (substitutions/site)

Fig. 4

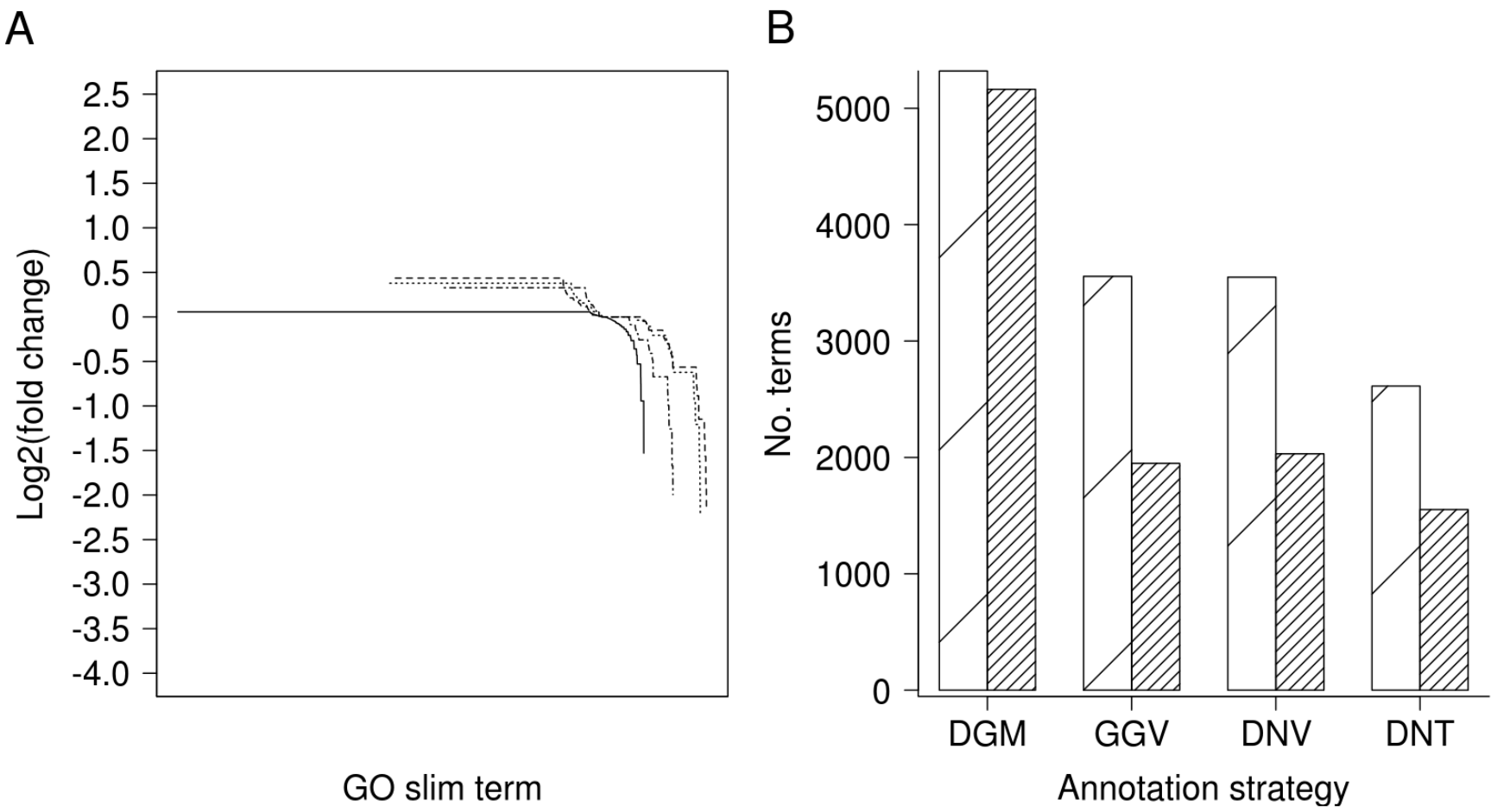

A
2.5
2.0
1.5
1.0
0.5
0
-0.5
-1.0
-1.5
-2.0
-2.5
-3.0
-3.5
-4.0
Log2(fold change)
GO slim term
B
5000
4000
3000
2000
1000
0
No. terms
DGM
GGV
DNV
DNT
Annotation strategy

Fig. 5

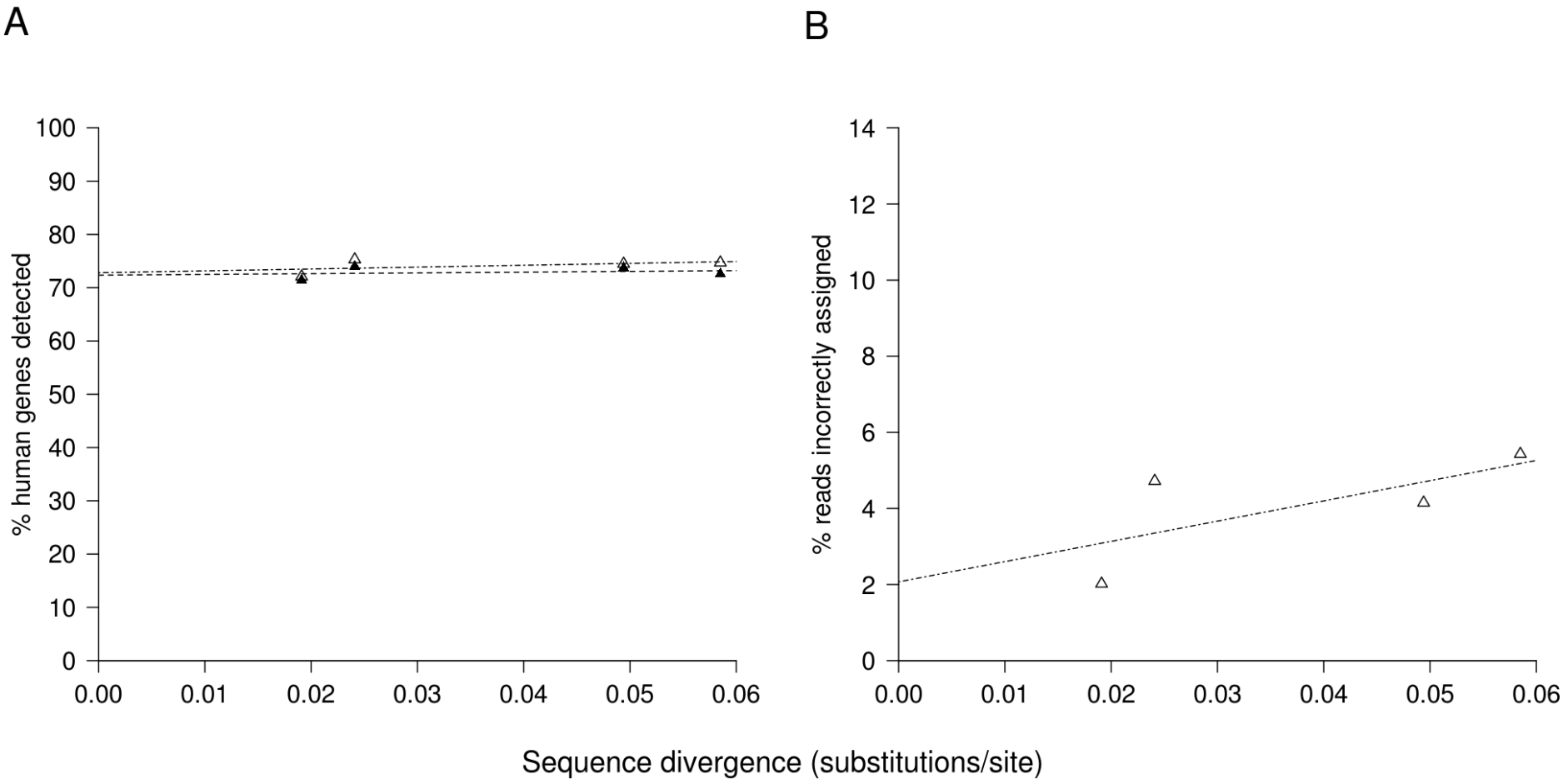
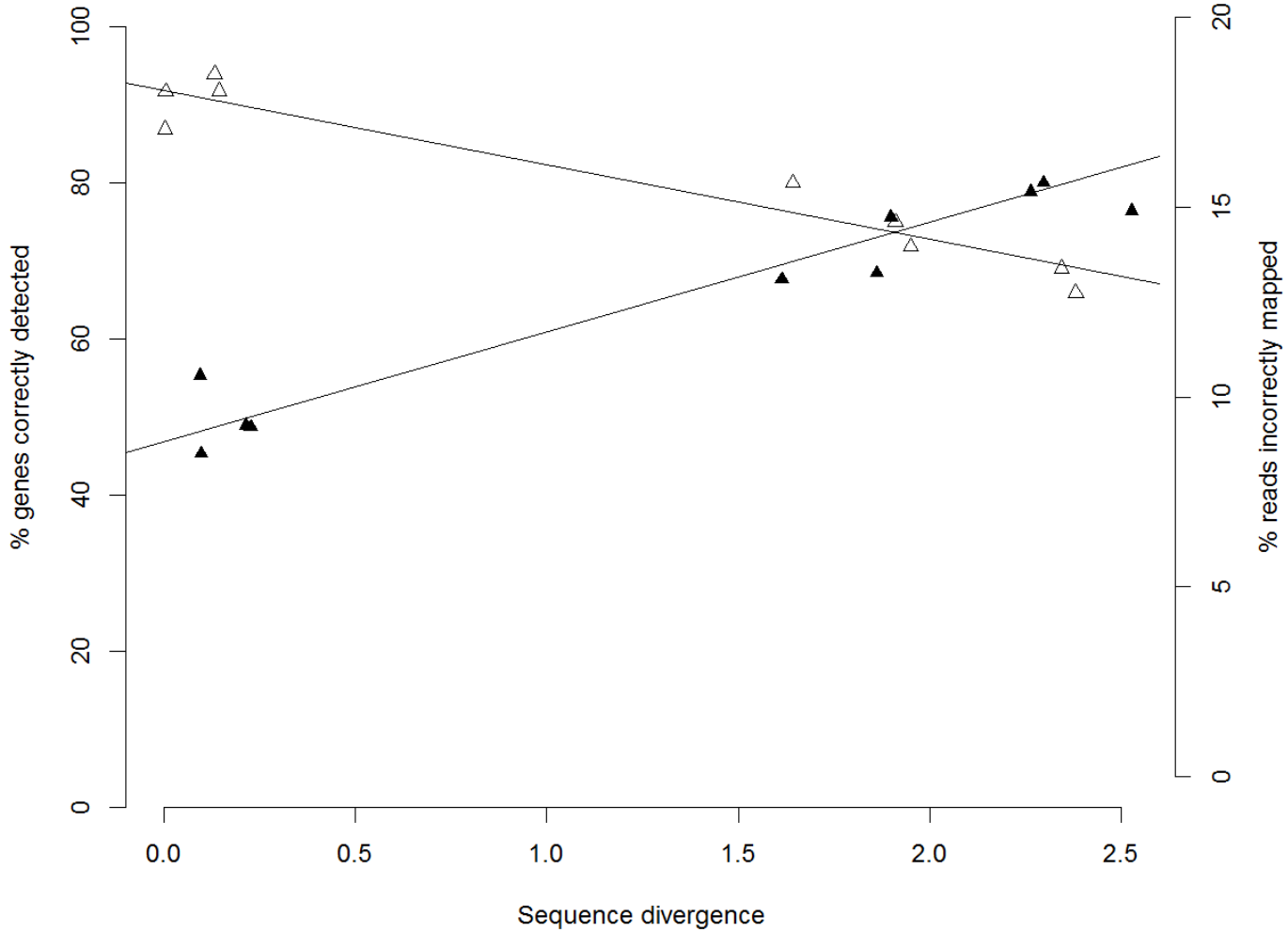

A
B
% human genes detected
% reads incorrectly assigned
Sequence divergence (substitutions/site)
0
10
20
30
40
50
60
70
80
90
100
0
2
4
6
8
10
12
14
0.00
0.01
0.02
0.03
0.04
0.05
0.06

Fig. 6


% genes correctly detected
% reads incorrectly mapped
Sequence divergence
0
20
40
60
80
100
0
5
10
15
20
0.0
0.5
1.0
1.5
2.0
2.5

3.8 Supplementary information

3.8.1 Supplementary figure legends

Fig. S1. Negative relationship between orthologous sequence mapping and divergence – trend recapitulated when results are plotted against divergence in MYA. The proportions of pre-processed *D. melanogaster* sequences (unassembled reads or assembled contigs) assigned to orthologous genes when mapped to each alternative *Drosophila* genome were plotted against *Drosophila* species divergence in MYA for (A) DGM; B) genome-guided assemblies using Velvet/Columbus; C) *de novo* assembly using Velvet/Oases; D) *de novo* assembly using Trinity. Total unassembled read or assembled contig matches (open squares) and single-match sequences (open triangles) are shown.

Fig. S2. Direct genome mapping displays lower gene detection error than alternative assembly methods - trend recapitulated when results are plotted against divergence in MYA. (A) The proportion of orthologous genes detected incorrectly by single-match sequences (unassembled reads or assembled contigs) was the lowest for direct genome mapping, compared to the assembly methods. (B) The proportion of orthologous genes detected incorrectly by multi-match sequences was the lowest for direct genome mapping, compared to the assembly methods. Single-match sequences displayed significantly lower gene detection error compared to multi-match sequences. Results for direct genome mapping (stars), genome-guided assemblies (diamonds), *de novo* assembly using Velvet/Oases (inverted triangles), and *de novo* assembly using Trinity (filled circles) are indicated.

Fig. S3. Single-match sequences show far lower error in assignment than multi-match sequences. (A) The proportion of single-match sequences (unassembled reads or assembled contigs) incorrectly assigned to orthologous genes was similar between direct genome mapping and the *de novo* assembly methods – all of which were lower than the error for the genome-guided assembly. (B) The proportion of multi-match sequences incorrectly assigned was the lowest for direct genome mapping, compared to the assembly methods. Results for direct genome mapping (stars), genome-guided assemblies (diamonds), *de novo* assembly using Velvet/Oases (inverted triangles), and *de novo* assembly using Trinity (filled circles) are indicated.

Fig. S4. Increased DGM annotation accuracy using reads filtered for low alignment scores and higher read counts per gene. Given that DGM performed the best for gene detection, gene detection accuracy was explored in greater depth. Allocating reads to bins according to score, reads with the lowest score range (< 199) were significantly different from the others in terms of the proportion of reads in that bin that were correctly assigned (ANOVA: $p < 2.2e^{-16}$, Tukey HSD test: $p < 2.2e^{-16}$ for '<199 score' bin compared to all others). Similarly, when allocating genes to bins according to read count, genes with less than 5 reads assigned were significantly different from the others in terms of the proportion of genes in that bin that were correctly detected (ANOVA: $p < 2.2e^{-16}$, Tukey HSD test: $p < 2.2e^{-16}$ for '<5 reads per gene' bin compared to all others). Hence, data for genes detected by single-match reads were filtered to, firstly, remove reads with an alignment score of less than 199, and, secondly, remove genes with fewer than 5 reads assigned. This caused a moderate drop in the proportion of orthologous *D. melanogaster* genes that can be detected (A) and improves DGM accuracy by a small amount (B), particularly at high levels of divergence.

Fig. S5. Gene detection error varies with functional gene category. Mean error scores of gene detection using DGM per GO slim term for (A) *D. sechellia* and *D. simulans*, (B) *D. erecta* and *D. yakuba*, and (C) *D. pseudoobscura* and *D. persimilis* were plotted (employing a minimum threshold of 20 *D. melanogaster* genes per GO slim term). Particular GO slim terms show heightened mean error scores across all levels of divergence, such as lysosome, whereas other terms maintain low levels of error, such as translation.

Fig. S6. Positive relationship between functional gene category detection bias and sequence divergence in primate species. Bias in GO slim term detection using DGM was assessed by calculating the log(2) fold change between observed and expected values for numbers of genes assigned to each GO slim term. Expected values for the numbers of genes assigned to each GO slim term were generated assuming a linear loss of genes per term with increasing divergence. These values were lower for DGM than the assembly methods. Bias was plotted at levels of increasing divergence for chimpanzee (solid line), gorilla (dashed line), orang-utan (dotted line), and macaque (dot-dashed line).

Fig. S7. Gene detection error varies with functional gene category in primate species. Mean error scores of gene detection using DGM per GO slim term for (A) chimpanzee, (B) gorilla, (C) orang-utan, and (D) macaque were plotted employing a minimum threshold of 20 human genes per GO slim term and selecting the 50 terms with highest error per species. Particular GO slim terms show heightened mean error scores across all levels of divergence, such as external encapsulating structure.

Fig. S1

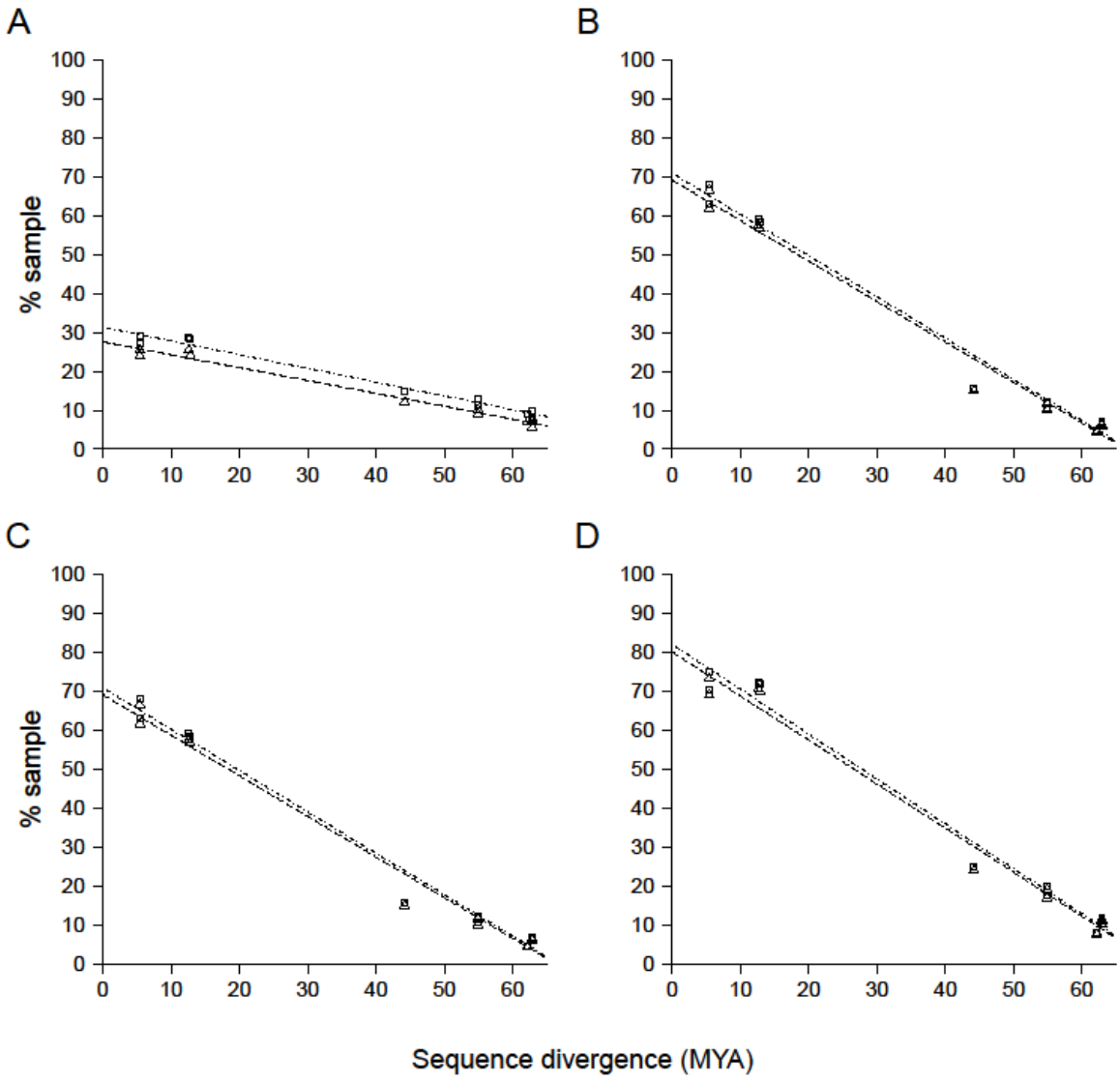

A
B
C
D
% sample
% sample
Sequence divergence (MYA)
0
10
20
30
40
50
60
70
80
90
100

Fig. S2

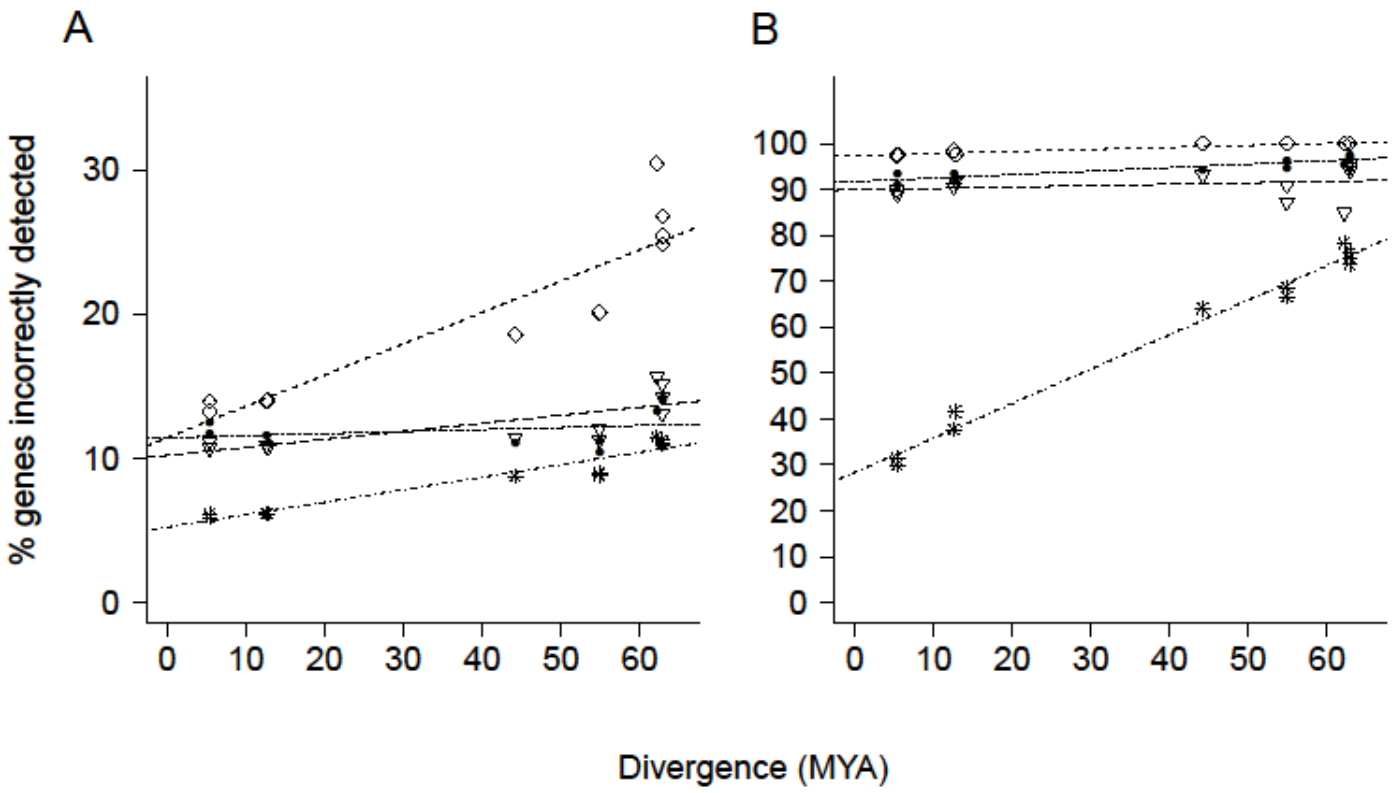

A
B
% genes incorrectly detected
0
10
20
30
40
50
60
70
80
90
100
Divergence (MYA)

Fig. S3

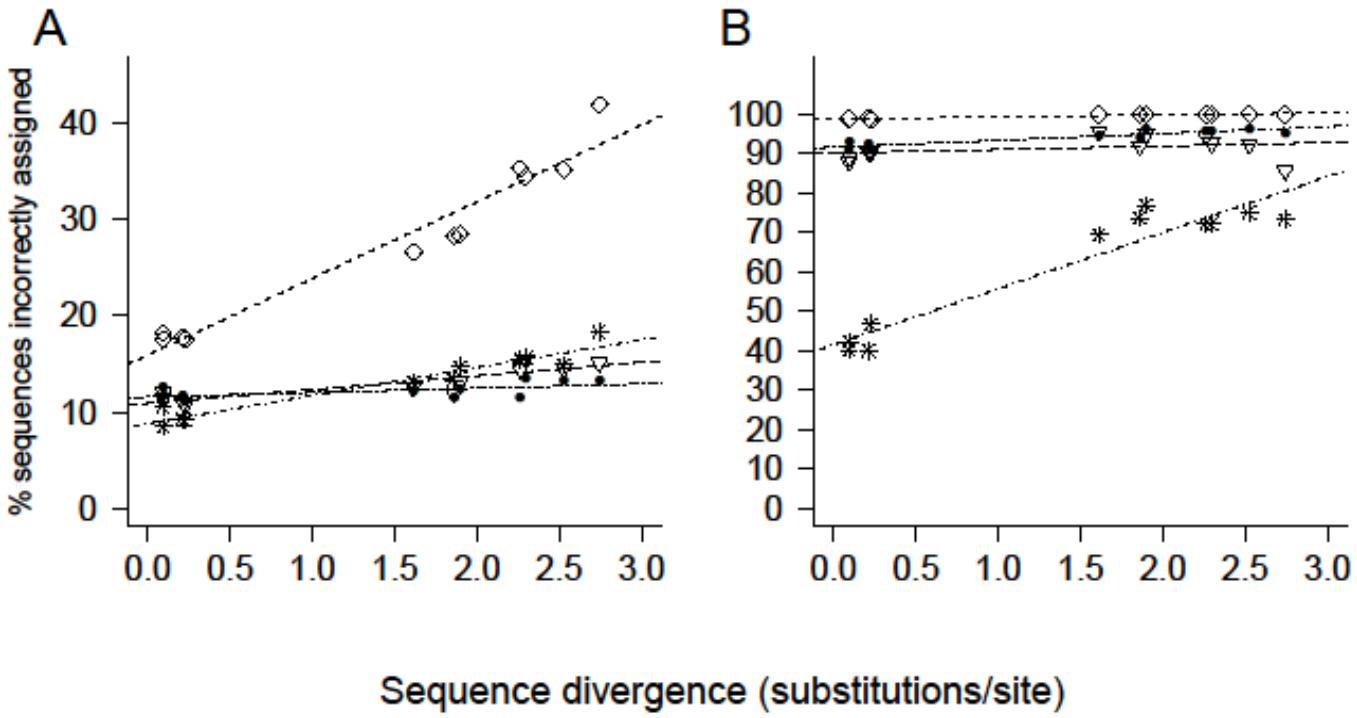

A
B
% sequences incorrectly assigned
0
10
20
30
40
50
60
70
80
90
100
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Sequence divergence (substitutions/site)

Fig. S4

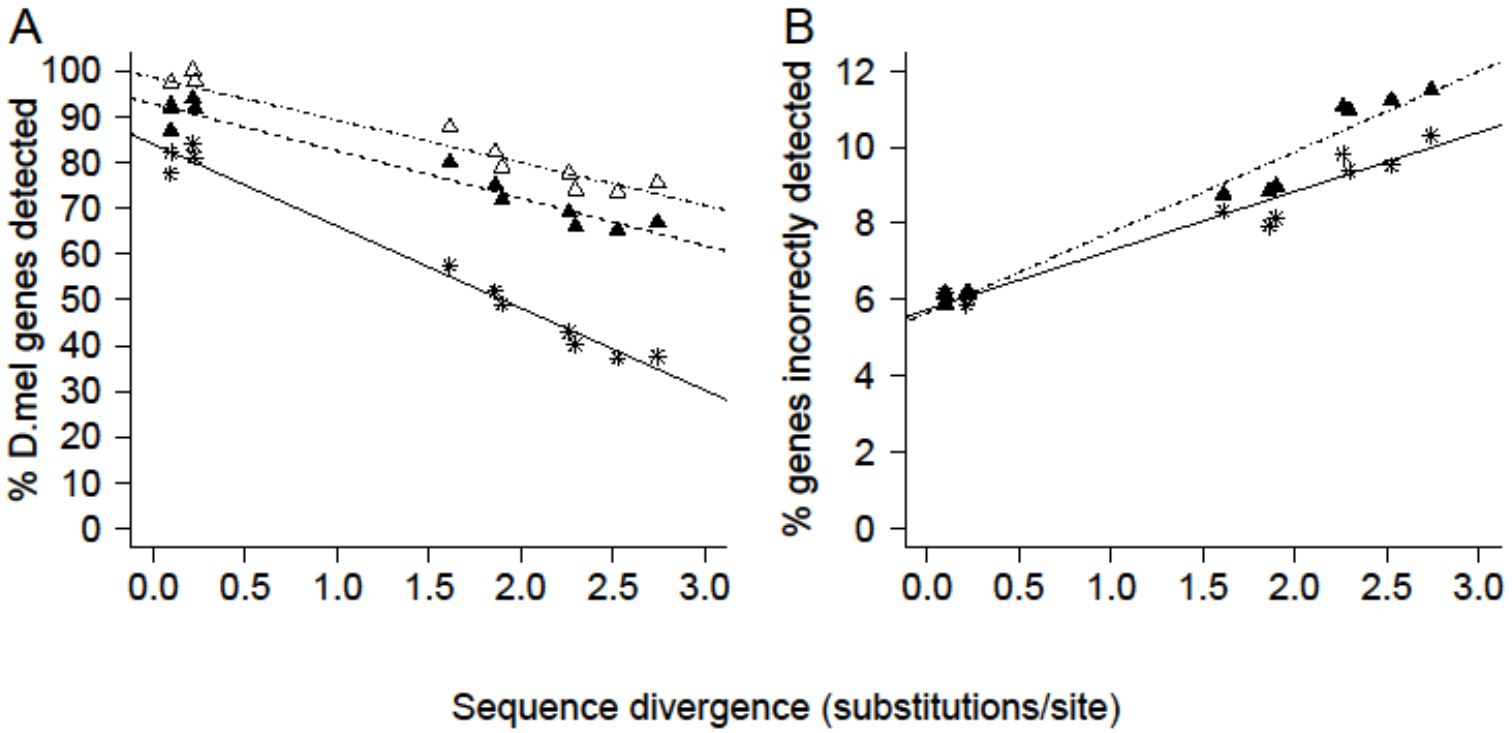

A
% D.mel genes detected
100
90
80
70
60
50
40
30
20
10
0
0.0
0.5
1.0
1.5
2.0
2.5
3.0
B
% genes incorrectly detected
12
10
8
6
4
2
0
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Sequence divergence (substitutions/site)

Fig. S5

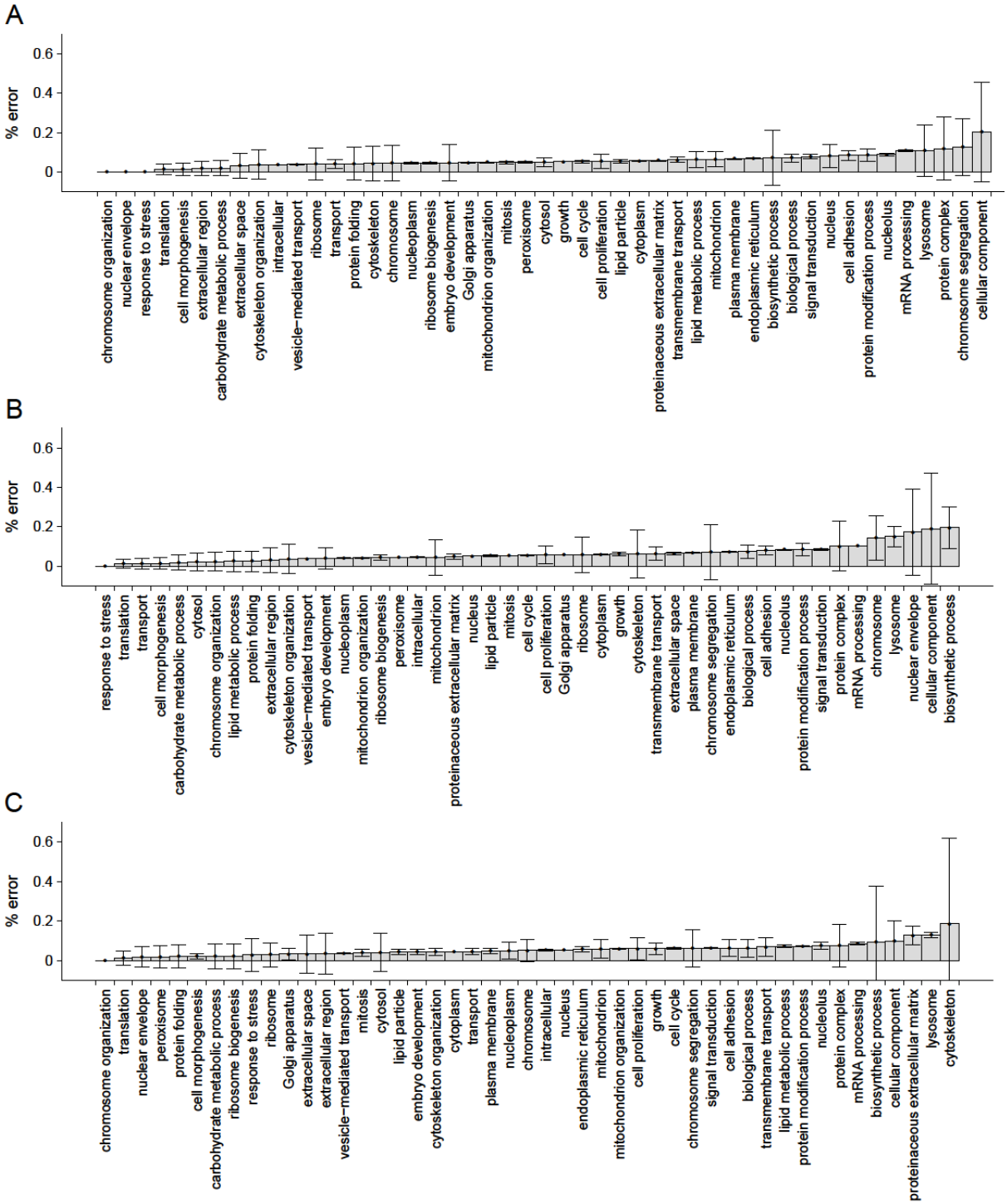

A
% error
0.6
0.4
0.2
0
chromosome organization
nuclear envelope
response to stress
translation
cell morphogenesis
extracellular region
carbohydrate metabolic process
extracellular space
cytoskeleton organization
intracellular
vesicle−mediated transport
ribosome
transport
protein folding
cytoskeleton
chromosome
nucleoplasm
ribosome biogenesis
embryo development
Golgi apparatus
mitochondrion organization
mitosis
peroxisome
cytosol
growth
cell cycle
cell proliferation
lipid particle
cytoplasm
proteinaceous extracellular matrix
transmembrane transport
lipid metabolic process
mitochondrion
plasma membrane
endoplasmic reticulum
biosynthetic process
biological process
signal transduction
nucleus
cell adhesion
protein modification process
nucleolus
mRNA processing
lysosome
protein complex
chromosome segregation
cellular component
B
% error
0.6
0.4
0.2
0
response to stress
translation
transport
cell morphogenesis
carbohydrate metabolic process
cytosol
chromosome organization
lipid metabolic process
protein folding
extracellular region
cytoskeleton organization
vesicle−mediated transport
embryo development
nucleoplasm
mitochondrion organization
ribosome biogenesis
peroxisome
intracellular
mitochondrion
proteinaceous extracellular matrix
nucleus
lipid particle
mitosis
cell cycle
cell proliferation
Golgi apparatus
ribosome
cytoplasm
growth
cytoskeleton
transmembrane transport
extracellular space
plasma membrane
chromosome segregation
endoplasmic reticulum
biological process
cell adhesion
nucleolus
protein modification process
signal transduction
protein complex
mRNA processing
chromosome
lysosome
nuclear envelope
cellular component
biosynthetic process
C
% error
0.6
0.4
0.2
0
chromosome organization
translation
nuclear envelope
peroxisome
protein folding
cell morphogenesis
carbohydrate metabolic process
ribosome biogenesis
response to stress
ribosome
Golgi apparatus
extracellular space
extracellular region
vesicle−mediated transport
mitosis
cytosol
lipid particle
embryo development
cytoskeleton organization
cytoplasm
transport
plasma membrane
nucleoplasm
chromosome
intracellular
nucleus
endoplasmic reticulum
mitochondrion
mitochondrion organization
cell proliferation
growth
cell cycle
chromosome segregation
signal transduction
cell adhesion
biological process
transmembrane transport
lipid metabolic process
protein modification process
nucleolus
protein complex
mRNA processing
biosynthetic process
cellular component
proteinaceous extracellular matrix
lysosome
cytoskeleton

Fig. S6

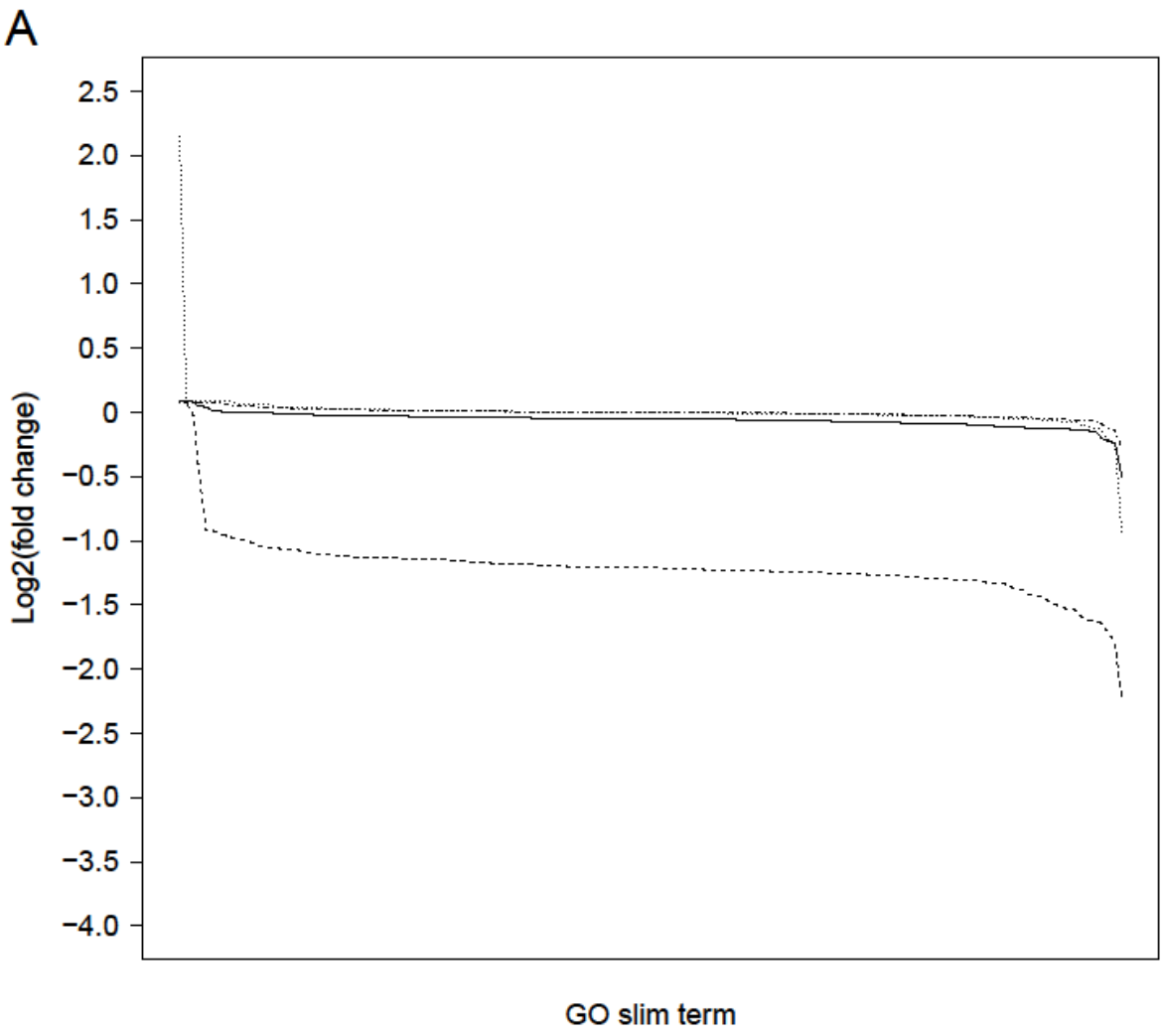

A
2.5
2.0
1.5
1.0
0.5
0
−0.5
−1.0
−1.5
−2.0
−2.5
−3.0
−3.5
−4.0
Log2(fold change)
GO slim term

Fig. S7

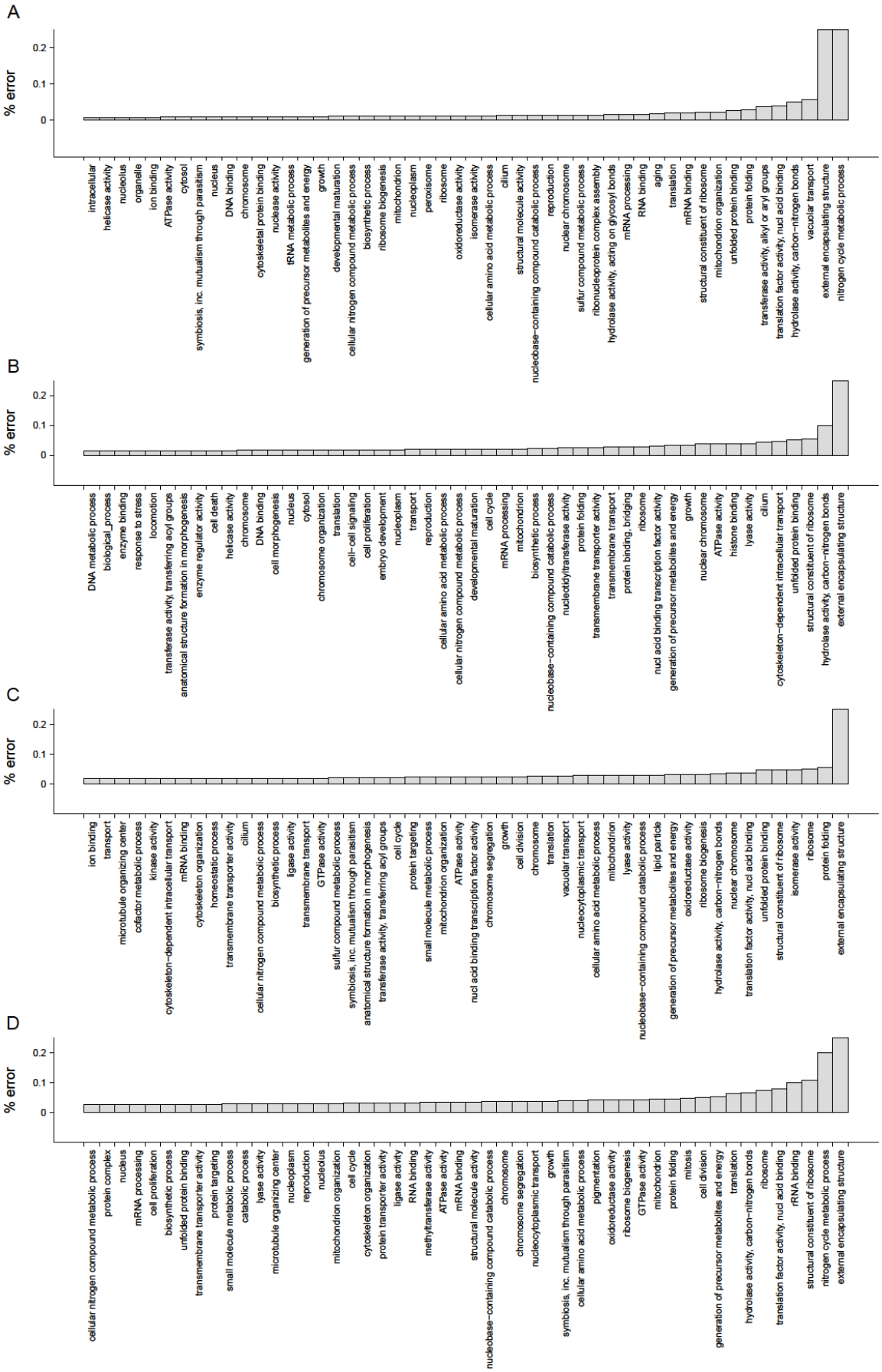

A
% error
0.2
0.1
0
intracellular
helicase activity
nucleolus
organelle
ion binding
ATPase activity
cytosol
symbiosis, inc. mutualism through parasitism
nucleus
DNA binding
chromosome
cytoskeletal protein binding
nuclease activity
tRNA metabolic process
generation of precursor metabolites and energy
growth
developmental maturation
cellular nitrogen compound metabolic process
biosynthetic process
ribosome biogenesis
mitochondrion
nucleoplasm
peroxisome
ribosome
oxidoreductase activity
isomerase activity
cellular amino acid metabolic process
cilium
structural molecule activity
nucleobase−containing compound catabolic process
reproduction
nuclear chromosome
sulfur compound metabolic process
ribonucleoprotein complex assembly
hydrolase activity, acting on glycosyl bonds
mRNA processing
RNA binding
aging
translation
mRNA binding
structural constituent of ribosome
mitochondrion organization
unfolded protein binding
protein folding
transferase activity, alkyl or aryl groups
translation factor activity, nucl acid binding
hydrolase activity, carbon−nitrogen bonds
vacuolar transport
external encapsulating structure
nitrogen cycle metabolic process
B
% error
0.2
0.1
0
DNA metabolic process
biological_process
enzyme binding
response to stress
locomotion
transferase activity, transferring acyl groups
anatomical structure formation in morphogenesis
enzyme regulator activity
cell death
helicase activity
chromosome
DNA binding
cell morphogenesis
nucleus
cytosol
chromosome organization
translation
cell−cell signaling
cell proliferation
embryo development
nucleoplasm
transport
reproduction
cellular amino acid metabolic process
cellular nitrogen compound metabolic process
developmental maturation
cell cycle
mRNA processing
mitochondrion
biosynthetic process
nucleobase−containing compound catabolic process
nucleotidyltransferase activity
protein folding
transmembrane transporter activity
transmembrane transport
protein binding, bridging
ribosome
nucl acid binding transcription factor activity
generation of precursor metabolites and energy
growth
nuclear chromosome
ATPase activity
histone binding
lyase activity
cilium
cytoskeleton−dependent intracellular transport
unfolded protein binding
structural constituent of ribosome
hydrolase activity, carbon−nitrogen bonds
external encapsulating structure
C
% error
0.2
0.1
0
ion binding
transport
microtubule organizing center
cofactor metabolic process
kinase activity
cytoskeleton−dependent intracellular transport
mRNA binding
cytoskeleton organization
homeostatic process
transmembrane transporter activity
cilium
cellular nitrogen compound metabolic process
biosynthetic process
ligase activity
transmembrane transport
GTPase activity
sulfur compound metabolic process
symbiosis, inc. mutualism through parasitism
anatomical structure formation in morphogenesis
transferase activity, transferring acyl groups
cell cycle
protein targeting
small molecule metabolic process
mitochondrion organization
ATPase activity
nucl acid binding transcription factor activity
chromosome segregation
growth
cell division
chromosome
translation
vacuolar transport
nucleocytoplasmic transport
cellular amino acid metabolic process
mitochondrion
lyase activity
nucleobase−containing compound catabolic process
lipid particle
generation of precursor metabolites and energy
oxidoreductase activity
ribosome biogenesis
hydrolase activity, carbon−nitrogen bonds
nuclear chromosome
translation factor activity, nucl acid binding
unfolded protein binding
structural constituent of ribosome
isomerase activity
ribosome
protein folding
external encapsulating structure
D
% error
0.2
0.1
0
cellular nitrogen compound metabolic process
protein complex
nucleus
mRNA processing
cell proliferation
biosynthetic process
unfolded protein binding
transmembrane transporter activity
protein targeting
small molecule metabolic process
catabolic process
lyase activity
microtubule organizing center
nucleoplasm
reproduction
nucleolus
mitochondrion organization
cell cycle
cytoskeleton organization
protein transporter activity
ligase activity
RNA binding
methyltransferase activity
ATPase activity
mRNA binding
structural molecule activity
nucleobase−containing compound catabolic process
chromosome
chromosome segregation
nucleocytoplasmic transport
growth
symbiosis, inc. mutualism through parasitism
cellular amino acid metabolic process
pigmentation
oxidoreductase activity
ribosome biogenesis
GTPase activity
mitochondrion
protein folding
mitosis
cell division
generation of precursor metabolites and energy
translation
hydrolase activity, carbon−nitrogen bonds
ribosome
translation factor activity, nucl acid binding
rRNA binding
structural constituent of ribosome
nitrogen cycle metabolic process
external encapsulating structure

Table S1. Genome sequence versions.

Genome sequences	**Release details***
Drosophila melanogaster	5.41
Drosophila ananassae	1.3
Drosophila erecta	1.3
Drosophila grimshawi	1.3
Drosophila mojavensis	1.3
Drosophila persimilis	1.3
Drosophila pseudoobscura	2.24
Drosophila sechellia	1.3
Drosophila simulans	1.3
Drosophila virilis	1.2
Drosophila willistoni	1.3
Drosophila yakuba	1.3
Homo sapiens	68
Pan troglodytes	68
Gorilla gorilla	69
Pongo abelii	68
Macaca mulatta	68
Taeniopygia guttata	66

* Fly genome sequences downloaded from Flybase (www.flybase.org) and primate genomes downloaded from Ensembl (www.ensembl.org; Flicek et al., 2014).

Table S2. Orthologous gene detection using alternative annotation strategies for total sequences (total) and single-match sequences (SM). Divergence given as both sequence divergence (total substitutions, 'subst') and million years ago ('MYA').

Species	DGM	Genome-guided assembly (Velvet Columbus)	*de novo* assembly (Velvet Oases)	*de novo* assembly (Trinity)
D. melanogaster (0sd, 0MYA)	11173 (total), 10914 (SM)	3722 (total), 3574 (SM)	3722 (total), 3578 (SM)	2360 (total), 2285 (SM)
D. sechellia (0.097subst, 5.4MYA)	10272 (total), 9998 (SM)	3121 (total), 3073 (SM)	3271 (total), 3183 (SM)	2112 (total), 2004 (SM)
D. simulans (0.095subst, 5.4MYA)	9698 (total), 9470 (SM)	2934 (total), 2888 (SM)	3074 (total), 2990 (SM)	2005 (total), 1908 (SM)
D. yakuba (0.227subst, 12.8MYA)	10398 (total), 10010 (SM)	2825 (total), 2791 (SM)	2992 (total), 2897 (SM)	2053 (total), 1926 (SM)
D. erecta (0.215subst, 12.6MYA)	10429 (total), 10246 (SM)	2863 (total), 2817 (SM)	3040 (total), 2932 (SM)	2028 (total), 1925 (SM)
D. ananassae (1.613subst, 44.2MYA)	8930 (total), 8721 (SM)	1120 (total), 1104 (SM)	1239 (total), 1194 (SM)	931 (total), 854 (SM)
D. pseudoobscura (1.861subst, 54.9MYA)	8450 (total), 8174 (SM)	894 (total), 875 (SM)	1001 (total), 957 (SM)	782 (total), 707 (SM)
D. persimilis (1.899subst, 54.9MYA)	8080 (total), 7835 (SM)	813 (total), 798 (SM)	914 (total), 876 (SM)	706 (total), 649 (SM)
D. willistoni (2.744subst, 62.2MYA)	7552 (total), 7292 (SM)	365 (total), 356 (SM)	443 (total), 410 (SM)	397 (total), 323 (SM)
D. mojavensis (2.528subst, 62.9MYA)	7377 (total), 7115 (SM)	515 (total), 503 (SM)	612 (total), 574 (SM)	612 (total), 464 (SM)
D. virilis (2.263subst, 62.9MYA)	7790 (total), 7527 (SM)	635 (total), 592 (SM)	534 (total), 520 (SM)	581 (total), 442 (SM)
D. grimshawi (2.297subst, 62.9MYA)	7585 (total), 7183 (SM)	595 (total), 556 (SM)	512 (total), 501 (SM)	568 (total), 421 (SM)

Table S3. GO slim terms with zero gene detection error for *Drosophila*. SimSec: From combined gene detections lists for D. simulans and D. sechellia. YakEre: From combined gene detections lists for D. yakuba and D. erecta. PsePer: From combined gene detections lists for D. pseudoobscura and D. persimilis. WilMojVirGri: From combined gene detections lists for D. willistoni, D. mojavensis, D. virilis, and D. grimshawi.

SimSec	YakEre	PsePer	WilMojVirGri
aging	aging	catabolic process	catabolic process
catabolic process	catabolic process	cell division	cytoplasmic membrane-bounded vesicle
cell death	cell-cell signaling	cellular amino acid metabolic process	
cell-cell signaling	cellular amino acid metabolic process	chromosome organization	
cellular amino acid metabolic process	cytoplasmic membrane-bounded vesicle	cytoskeleton-dependent intracellular transport	
chromosome organization	extracellular matrix organization	extracellular matrix organization	
extracellular matrix organization	nuclear chromosome	nuclear chromosome	
nuclear chromosome	nucleocytoplasmic transport	nucleocytoplasmic transport	
nuclear envelope	protein complex assembly	ribonucleoprotein complex assembly	
nucleocytoplasmic transport	protein targeting	vacuolar transport	
protein complex assembly	response to stress		
response to stress	ribonucleoprotein complex assembly		
ribonucleoprotein complex assembly	vacuolar transport		
tRNA metabolic process			
vacuolar transport			

Table S4. Top 20% of GO slim terms ranked by gene detection error for *Drosophila* and primate species. The terms reproduction, biosynthetic process, and mRNA processing are highlighted (bold italic) as they exhibit consistently high error in all species tested, both *Drosophila* and primate.

(A)

Drosophila	GO slim term	Mean error
	sulfur compound metabolic process	0.4938
	reproduction	0.3635
	cell division	0.2100
	protein targeting	0.2083
	cellular_component	0.1509
	lysosome	0.1470
	microtubule organizing center	0.1305
	biosynthetic process	0.1201
	protein complex	0.1097
	proteinaceous extracellular matrix	0.1055
	mRNA processing	0.1025
	cell differentiation	0.0992
	nucleolus	0.0944
	cell adhesion	0.0918
	protein modification process	0.0898
	biological_process	0.0823
	cell wall	0.7500
Primates	external encapsulating structure	0.2500
	hydrolase activity, acting on carbon-nitrogen (but not peptide) bonds	0.0627
	structural constituent of ribosome	0.0580
	ribosome	0.0414
	protein folding	0.0390
	translation factor activity, nucleic acid binding	0.0389
	unfolded protein binding	0.0381
	generation of precursor metabolites and energy	0.0319
	translation	0.0313
	nuclear chromosome	0.0276
	Mitochondrion	0.0267
	Growth	0.0266
	ATPase activity	0.0265

	nucleobase-containing compound catabolic process	0.0256
	cellular amino acid metabolic process	0.0252
	oxidoreductase activity	0.0249
	cilium	0.0231
	chromosome	0.0219
	nucleic acid binding transcription factor activity	0.0214
	nucleocytoplasmic transport	0.0213
	symbiosis, encompassing mutualism through parasitism	0.0203
	transmembrane transporter activity	0.0202
	biosynthetic process	0.0199
	reproduction	0.0197
	mRNA processing	0.0196
	cellular nitrogen compound metabolic process	0.0189
	nucleoplasm	0.0186

4 Brain transcriptomes of two non-sequenced wild, free-living songbird species, the dunnock and the water pipit: exploring the genomic basis of differences in behavioural ecology

4.1 Abstract

Phenotypic evolution and development in an ecological context can be meaningfully explored using non-traditional species in next generation transcriptome sequencing (RNA-seq) studies. Songbirds present diverse phenotypic variation, particularly regarding social behaviour and communication, and are increasingly important in behavioural and molecular ecology. We obtained Illumina RNA-seq data from brain samples from males of two songbird species that currently lack genome sequences, the water pipit, *Anthus spinoletta*, and the dunnock, *Prunella modularis*. This pair of species represents a comparative model of mating system evolution: the water pipit is monogamous whereas the dunnock is highly polygamous. Sperm morphology confirms differing levels of sexual selection operating in these two species. The transcriptome of each species was assembled using the zebra finch's genome as reference. Additionally, we used a direct read-to-genome mapping technique for transcriptome annotation which we have previously shown to be more effective for annotating transcriptome data of species lacking sequenced genomes. We detected expression of over 15,000 genes in each species, representing over 90% of annotated zebra finch genes. In contrast, assembly based methods allowed the detection of only around 46% of zebra finch genes. We conducted differential gene expression analysis to explore candidate genes that may underlie species-specific differences related to phenotypic variation, identifying 62 genes to be significantly differentially expressed (adjusted p-value <0.05). These genes have been associated with defence against stress, energy balance and neurogenesis and may underlie some of the observed differences in mating behaviour between the two species. This study provides the first indication of the differences in brain gene expression profiles associated with monogamous and polygamous mating behaviour in songbirds.

4.2 Introduction

Birds are excellent model systems for the study of the ecology and evolution of sexual behaviour as they exhibit great diversity in mating systems and parental care. Traits including sexual dimorphism, ornamentation, sperm competition, and social behaviours have been described in detail for many species (Garamszegi et al., 2005; Griffith et al., 2002; Møller & Briskie, 1995; Sol et al., 2007; Székely et al., 2007; Székely et al., 2004; van Dijk et al., 2010) making birds an increasingly attractive target for neuroethological and genomic studies over traditional rodent models. The availability of many avian genomes, from chicken, *Gallus gallus* (International Chicken Genome Sequencing Consortium 2004) and zebra finch, *Taeniopygia guttata* (Warren et al., 2010) to the recent release of near 50 avian genomes (Zhang *et al.* 2014;, see http://phybirds.genomics.org.cn) enables comparative genomic studies using bird species, and, in particular, has placed the diverse and well-described oscine Passerine species at the centre of the field of avian behavioural genomics (Clayton et al., 2009). Next generation transcriptome sequencing (RNA-seq, Wang et al., 2009; Wilhelm & Landry, 2009) has allowed genome-wide exploration of factors involved in complex trait development and evolution (Shi *et al.* 2011) and has further allowed the study of transcriptome profiles for species with no current available genome sequence, due to its independence of the need for species-specific probe sequences, such as with array-based methods (Barakat et al., 2009; Colgan et al., 2011; Collins et al., 2008; Crawford et al., 2010; Esteve-Codina et al., 2011; Kawahara-Miki et al., 2011; Künstner et al., 2010; Moghadam et al., 2013; Wolf et al., 2010).

Recent studies involving genome wide transcriptome profiling provide interesting insights into the molecular basis of complex traits. Balakrishnan et al. (2013) characterised the brain transcriptome of the violet-eared waxbill, *Uraeginthus granatina*, via Roche 454 sequencing using the zebra finch's genome as reference for transcript annotation. This study compared differences in gene sets and patterns of polymorphisms in the violet-eared waxbill and the zebra finch, and identified genomic differences that may underpin some of the marked differences in social behaviour, including group living and territoriality, between the two species (Balakrishnan *et al.* 2013). The use of 454 sequencing, however, limited the study's ability to characterise transcript abundance patterns. Moghadam et al. (2013) explored sex-biased gene expression in the Kentish plover, *Charadrius alexandrinus*, neurotranscriptome using Illumina RNA-seq technology and a *de novo* assembly method for transcriptome annotation, identifying categories of gene function that are significantly different between females and males (Moghadam *et al.* 2013). Both of these studies chose to present brain transcriptomes, highlighting the potential for avian species to shed important insight into the neurogenomics of social trait evolution.

Here we compare the brain transcriptomes and investigate brain gene expression patterns in two wide-spread Eurasian species, the water pipit, *Anthus spinoletta*, and the dunnock, *Prunella modularis*. These songbirds are members of the family *Passeridae* and are almost equally diverged from each other as they are from their common reference species, the zebra finch (Fjeldså et al, 2010). Water pipits are general larger than dunnocks, with males being generally larger than females in both species. Similar in habitat and feeding preferences, these two species seem to differ primarily around mating system and song (Table 1). The water pipit is a typically monogamous species, displaying typically very low amounts of extra pair paternity (EPP) per breeding season (5.2% EPP, Griffith et al., 2002; Reyer et al., 1997), and exhibiting simple song patterns where occurrence of a particular buzz (the 'snarr') predicts mating success (Rehsteiner et al., 1998). The dunnock exhibits a highly variable socially and sexually polygamous mating system including polyandry and polygynandry that may involve several males and several females (EPP up to 44.1%, Burke et al., 1989; Davies, 1992; Griffith et al., 2002). The dunnock has a highly complex song repertoire, which is variable depending on the social context: it may be used territorially over large distances, or between individuals, such as during courtship (Snow & Snow 1983). As such, given their relatively similar morphology and ecology, the water pipit and the dunnock present an excellent species pair for the exploration of species-specific differences in gene expression that might underlie differences in behaviour related to differential mating system evolution. Neither of these species have sequenced genomes or any form of genome-wide expressed sequences publically available; nor, to our knowledge, do any other members of their respective families (water pipit: Motacillidae; dunnock: Prunellidae).

To provide an initial assessment of the utility of the water pipit and dunnock in comparative genomics studies and a preliminary exploration of the impacts of differing levels of sexual selection on brain gene expression in songbirds, we present a characterisation and comparative analysis of the transcriptomes of the water pipit and the dunnock. Using RNA-seq, we have sequenced and analysed the pooled brain transcriptomes of a number of wild, free-living males from both species, generating approximately 100 million reads per species. These were assembled using the closest available reference sequence, the genome of the zebra finch, *Taeniopygia guttata*. As the annotation of transcriptome sequences from species lacking an available genome relies upon homology detection with annotated regions of the closest reference species, this presents challenges and limitations inherent to the evolutionary divergence between the species used (Renn et al., 2004; Machado et al., 2009). However, we recently compared the efficacy and accuracy of various currently popular transcriptome annotation techniques, and concluded that direct genome mapping (DGM) detects by far the greatest number of genes with the lowest error and functional bias (Ockendon et al. submitted). We predicted that, given the close evolutionary relationships between the water pipit, dunnock and zebra finch, the transcriptomes of the two study species would be relatively similar in terms of the number of genes that could be detected and their overall patterns

of genetic variation and sequence evolution compared to the zebra finch. However, given that the dunnock is documented to be monophyletic with the zebra finch and not the water pipit, we anticipated that the relative levels would be lower for the dunnock relative to the zebra finch compared to the water pipit relative to the zebra finch. We hypothesised that a proportion of transcripts would exhibit enhanced rates of evolution indicating positive selection between the dunnock and the water pipit, which may represent selective differences due to their natural histories (null: no transcripts would exhibit enhanced rates of evoltuon). Additionally we hypothesised that there would be significant differences in brain gene expression patterns between the water pipit and the dunnock, which may represent variation in functional pathways related to behavioural differences (null: there would be no significant differences in brain gene expression patterns between the water pipit and the dunnock).

Since sperm morphology variance is a proxy for EPP in birds (Lifjeld et al., 2010) reflecting the level of sexual selection operating (Møller & Ninni, 1998) we have additionally analysed sperm length from individuals of both species collected in the field to identify whether indeed they are subject to differential levels of sexual selection as predicted by previous behavioural and paternity studies of these species. More intense sexual selection – a process that involves directional selection – would be indicated by low variance in sperm morphology (Lifjeld *et al.* 2010). We hypothesised that increased sexual selection in the dunnock specimens used in the study would be reflected in low sperm morphology variance compared to the water pipit (null hypothesis: we would observe no significant differences in sperm morphology variance when comparing the two species).

Using direct genome mapping of transcriptome short reads, we detected the greatest number of genes compared to the assemblies: over 90% of zebra finch protein coding genes (with an estimated error rate of approximately 10%; Ockendon et al. submitted). We detected over 14,000 single nucleotide polymorphisms (SNPs) and approximately 100,000 insertions/deletions (indels) in each species, between 20% and 30% of which mapped to gene regions. Analysing molecular rates of evolution, we found that approximately 7% of all transcripts in both species may be experiencing adaptive evolution. Differential gene expression analysis indicated that 62 genes were significantly differentially expressed, implicating pathways involved in defence against stress, energy balance and neurogenesis. Given that individuals used in the study could not be sequenced separately, it was not possible to robustly determine natural variation in gene expression and hence these results are caveated accordingly: they should be considered as preliminary findings worthy of further study. Combined, these results demonstrate how useful non-model species with no available genomes can be in comparative transcriptomics studies, and provide the first insight into the genomic landscapes and comparative functional genomic features of these two interesting songbird

species. We identify genes that could be considered as candidates underlying differences in mating behaviour: further songbird species pairs should be used to enhance these assertions.

4.3 Materials and Methods

4.3.1 Fieldwork and wild songbird brain samples

Fieldwork was conducted in the Harghita region of Transylvania, Romania, to collect tissue from wild populations of water pipit and dunnock during their breeding season in May-June 2011 (under permit: Ministerial Order no. 1470/2011). Using song playback, four water pipit and five dunnock adult males were lured into mist nests. We were granted permission to obtain both females and males, although since only one female was trapped, she was released and no samples were obtained from her. Morphometric data were collected and collated for each bird. Birds were sacrificed by decapitation within four minutes of capture to prevent stress-induced changes to circulating testosterone levels and gene expression (Deviche et al., 2010; Van Hout et al., 2010). Whole brains were dissected out, hindbrains were removed and the remaining material was finely chopped and placed in Eppendorf tubes free from DNA, DNase and RNase, and flooded with RNAlater to remove any air bubbles. Testes dimensions were obtained using sterile callipers. Samples were stored on ice for between 8 and 12 hours, to allow the RNAlater to permeate the whole tissue (Applied Biosciences protocol, Ambion), before being stored at approximately –17^{o}C for up to 10 days before being frozen to –80^{o}C.

4.3.2 Sperm morphology analysis

Sperm samples were taken from each individual bird that was used in this study and measured (performed by Dr. Alexander Ball, Szekely lab, University of Bath). 10µl sperm samples in formalin solution were air-dried on microwell dishes prior to addition of 10µl 0.6µmol DAPI solution. Sperm were visualised using a Zeiss LSM510Meta confocal laser-scanning microscope. Suitable individual sperm were located using a 20x Phase 2 air objective microscope under white light, then viewed using a 488nm argon laser and digitally photographed. Morphometric measurements were derived using the image processing program, ImageJ (Abràmoff et al., 2004). Ten sperm per male were used: each was measured three times and mean values were used in further analyses.

4.3.3 RNA-seq

RNA was extracted and tested for integrity using the Genome Analyser. The three best quality samples per species (Figs. S1 and S2) were then pooled for paired-end sequencing using the Illumina HiSeq 2000 platform (Illumina, Inc). Two lanes were sequenced per species. Total transcriptome short read samples were pre-processed using the FASTX toolkit (http://hannonlab.cshl.edu/fastx_toolkit/index.html). Sequencing artefacts were removed, adaptor and barcode sequences were clipped, and the remaining reads were quality filtered to meet a minimum PHRED score of 20 per base with 10% of the read length allowed below this (Crawford

et al. 2010). Table 2 provides quantities and mean quality scores for reads pre-, and post-processing.

4.3.4 Transcriptome assemblies

Genome-guided assemblies were generated using software packages Velvet Columbus (Zerbino & Birney, 2008; Zerbino, 2010), performed by Dr. Lauren O'Connell, Harvard University. All necessary alignments (single-end and paired-end) between the pre-processed reads and the zebra finch genome were performed using the gapped short read alignment mapping programme, SHRiMP (Rumble et al., 2009; David et al., 2011) with default parameters while outputting all unaligned reads to the alignment file. A multiple k-mer approach (k= 23, 25, 27, 29, 31, 33) was used and the *mergeAssembly* function was used to merge the multiple kmers. Then, CD-HIT-EST (Li & Godzik, 2006) was used to remove contig redundancy that can occur by merging multiple assemblies. Given that redundant contigs can represent alternative spice variants, polymorphisms among the pooled individuals, or sequencing errors, a conservative threshold of 98% sequence similarity was used. Assemblies were subjected to a custom Perl script that performed homology searching and local alignment against zebra finch CDS sequences (performed by Dr. Stephen Bush, Urrutia lab, University of Bath). Homology searches were conducted using Blast v2.2.26+ (Altschup et al., 1990) with threshold value E = 1e-10. Significant hits were then verified using the Smith–Waterman algorithm (fasta36.3.5d with parameters –a –A, Pearson, 2000). Custom Python scripts were then used to select and annotate only those contigs that matched to a single gene (single-match contigs).

4.3.5 Annotation using direct genome mapping (DGM)

Pre-processed transcriptome short reads were aligned (single-end and paired-end) against the closest available reference sequence, the zebra finch genome (Warren *et al.* 2010), using the gapped short read alignment mapping programme, SHRiMP (Rumble et al., 2009; David et al., 2011). The zebra finch genome sequence and corresponding gene annotations were downloaded from Ensembl (http://www.ensembl.org; Flicek et al., 2014). Alignments were generated in SAM output format (Li et al., 2009) using default parameters with the correct quality value offsets (+33) and were analysed using a pipeline constructed of custom Python scripts and Python-based tools. The custom scripts selected only those reads that map to a single location (single match reads) as these have been shown to be the most accurate population of aligned reads across large evolutionary distances between transcriptome and reference species (Ockendon et al. submitted). Alignments were not filtered by mapping quality. The Python-based tool HTSeq (Anders et al., 2014) was used to generate read counts per gene. Detected genes were subsequently assigned to GO slim categories (downloaded from Ensembl).

4.3.6 Gene ontology annotation

Genes were assigned to GO slim terms, downloaded from Ensembl Biomart (www.ensembl.org; Flicek et al., 2014), using custom Python scripts and subject to hypergeometric tests for over-, and under-representation, performed in R (The R Development Core Team 2010).

4.3.7 Sequence variation detection and analysis

Zebra finch chromosome information was obtained from NCBI Genome database (www.ncbi.nlm.nih.gov/genome). Single nucleotide polymorphisms, SNPs, were identified from SAM format alignments using SAMtools mpileup (Li et al., 2009) and segregated into those shared with the zebra finch, those not shared with ZF and indels. These features were then mapped to genes regions. All SNP and indel lists were assigned to chromosomes. SNPs mapping to genes were explored for functional enrichment using DAVID (Huang et al., 2009; Jiao et al., 2012).

4.3.8 Molecular rate analysis

Reads aligned using SHRiMP were processed using custom Perl scripts (performed by Dr. Stephen Bush, Urrutia lab, University of Bath) to extract transcript sequences from the water pipit and the dunnock. Extracted sequences were aligned to homologous zebra finch sequences and inputted into PAML (Yang 1997), generating dN and dS data from which dN/dS, the metric typically used to assess rates of molecular evolution (Yang & Bielawski 2000), was calculated. Molecular rates were filtered to remove items where $dS < 0.02$, $dS > 2$, or $dN > 2$ (Löytynoja & Goldman 2008). Transcripts were ranked according to dN/dS and assigned to GO slim terms. Gene functional enrichment/depletion were assessed using hypergeometric tests, as before.

4.3.9 Differential expression

Raw read counts per gene, or gene counts per GO slim term, were inputted into the differential expression package, DESeq (Anders & Huber 2010) implemented within the statistical language, R (The R Development Core Team 2010). Differential gene expression was performed with default parameters, whereas differential GO slim term expression was implemented using the local fit parameter as default parameters (parametric fit) failed.

4.4 Results

4.4.1 Sperm and body morphological variance

As expected, the within-male variance in mean sperm length is lower for the dunnock than for the water pipit (Fig. 1, Lifjeld et al., 2010), indicating higher levels of sexual selection operating in dunnock than in water pipit (Lifjeld et al., 2010; Møller & Ninni, 1998), hence verifying our choice of these species in experiencing opposing levels of sexual selection. The dunnock presents larger cloacal protuberances and greater testes volume, as expected (Fig. 2, A and B, respectively, Birkhead et al., 1993; Schut et al., 2012; Wolfson, 1952). However, the dunnock displays substantially more variance than the water pipit for cloacal protuberance volume (Fig. 2, A), indicating that there is lower constraint on this feature. Testes volume variance appears slightly lower for the dunnock compared to the water pipit (Fig. 2, B), reminiscent of sperm length variance, suggesting that testes volume is impacted by sexual selection. Incongruence between cloacal protuberance and testes volumes is not unexpected given that cloacal protuberance volume is impacted by not only sperm length but also number of sperm and may be additionally be affected by the typical mating rates for these species – low for water pipit and high for dunnock (Birkhead *et al.* 1993).

4.4.2 Transcriptome sequencing and annotation

It is possible to align reads in single-end, or paired-end modes and as there are no current guidelines as to which method is most appropriate when using a species with no sequenced genome, we performed both for the assemblies and the DGM to assess which approach returned the greatest gene detection. From the assemblies, more genes were detected in the water pipit than the dunnock by the single-end alignments (8,188 and 8,112, respectively; Table 3), whereas with the paired-end alignments, more genes were detected in the dunnock than the water pipit (8,627 and7,496, respectively; Table 3). DGM identified far more genes than the assembly method, both when using single-, and paired-end mapping, and single-end mapping detected the greatest number of genes using DGM (Table 3). Overall, the greatest number of genes was detected using DGM in single-end mode: 15,837 were detected in the water pipit and 15,740 were detected in the dunnock, representing 90.6% and 90.0% of annotated zebra finch genes, respectively (Table 3). Just over 400 fewer genes were detected for the dunnock compared to the water pipit. Given that there were 70% as many raw reads generated for the dunnock (65 million paired reads) compared to the water pipit (92 million paired reads), this is not surprising and also shows that over a certain level, acquiring more reads does not necessarily result in a proportionally greater power to detect genes. The different number of reads generated is most likely due to comparatively better RNA integrity for the water pipit compared to the dunnock (Figs. S1 and S2). As DGM in single-end alignment mode detected the highest number of genes, all further analyses were conducted using these results.

4.4.3 Gene functional characterisation

Over- and under-represented zebra finch GO slim terms, detected for the water pipit and the dunnock are shown in Figs. 3-6. 138 of the 139 zebra finch GO slim terms were detected in both the water pipit and the dunnock, respectively: the only term not detected was extracellular matrix organisation, which has 3 zebra finch genes in this term. As there are 27 (nearly 20% of) terms with fewer than 10 genes assigned in the zebra finch, the failure to detect extracellular matrix organisation may represent specific differences between the water pipit/dunnock transcriptomes and the zebra finch. We identified the same terms as over-, and under-represented in both the water pipit and dunnock, indicating that the relative divergence of each species from the zebra finch is equal enough to return comparable functional transcriptome profiles. As expected, terms for cellular components and housekeeping processes, such as terms for ribosome, cytoplasm, endoplasmic reticulum, Golgi apparatus, and translation, were over-represented. Also, those terms relevant to high energy demands and processes typical of brain tissue were enriched, such as mitochondrion, ATPase activity, and vesicle-mediated transport.

4.4.4 Distribution of genetic variation within the brain transcriptomes

As reads were obtained from pooled samples from three individuals per species we were able to detect SNPs, as well as insertions and deletions (indels) when compared to the zebra finch genome. Table 4 outlines the quantities of features identified per species overall and those that mapped to gene models. Far more indels were detected than SNPs, and distributions were similar between the water pipit and the dunnock, usually scaling with chromosome length (Figs. 7-9). Interestingly, the distributions of SNPs where one of the possible nucleotide variants is shared with the zebra finch was much more diffuse (Fig. 9) than those that are not shared with the zebra finch (Fig. 8), and chromosome 13 appeared as an extreme outlier of both songbird species (Fig. 9). These findings are similar to those by Balakrishnan et al. (2013) who found that SNPs on the Z chromosome and chromosome 4A in the violet-eared waxbill transcriptome did not scale with chromosome size, as did the other chromosomes (Balakrishnan *et al.* 2013). Given the strong synteny in avian genomes (Völker *et al.* 2010), this indicates that chromosome 13 may be of interest for further exploration of the species-specific differences between the water pipit, dunnock and zebra finch. Between approximately three-, and five-fold fewer features mapped to annotated gene regions (Table 4), although highly similar distributions were observed (Figs. S7-S9. This indicates that the majority of features, mapping outside genes, represent water pipit-, or dunnock-specific transcripts. SNPs mapping to genes were explored for functional enrichment using DAVID (Huang *et al.* 2009; Jiao *et al.* 2012), but none were found to be statistically significant (data not shown).

4.4.5 Patterns of sequence evolution

Estimates of synonymous substitutions, dS, can represent the underlying mutation rates, assuming that selection operating on these sites is neutral (see Yang & Bielawski, 2000). Mean dS estimates were lower for the water pipit relative to the dunnock (0.123, SD: 0.131) compared to either of these species relative to the zebra finch (water pipit: 0.136, SD: 0.156; dunnock: 0.133, SD: 0.157). Recent published phylogenetic information for these songbirds indicate that the zebra finch is monophyletic with the dunnock but not the water pipit (Fjeldså et al., 2010; Garamszegi & Møller, 2004). The results presented here indicate that mutation rates between the water pipit and dunnock are generally lower than between either of these and the zebra finch, suggesting that they may in fact be more closely related to each other than to the zebra finch as the phylogenies suggest. Mean dN/dS estimates were again lower for the water pipit relative to the dunnock (0.190, SD: 0.370), than for either species relative to the zebra finch (water pipit: 0.228, SD: 0.413; dunnock: 0.224, SD: 0.457). Of the 13,698 water pipit and 13,484 dunnock transcripts constructed, 960 (7.00%) and 926 (6.87%) transcripts displayed dN/dS>1 when aligned to the zebra finch, respectively, indicative of adaptive evolution (Yang & Bielawski 2000). When the water pipit and dunnock were aligned to each other, 808 transcripts displayed dN/dS>1, suggesting that fewer transcripts were under positive selection between these two species than either of them compared to the zebra finch.

Pairwise dN/dS estimates collated per chromosome (Fig. 10) for the three combinations of species showed that rates were lower when the water pipit was compared to the dunnock than when either species is compared to the zebra finch. In all combinations the Z chromosome displayed the greatest dN/dS, as was found by Balakrishnan et al. when comparing the violet-eared waxbill to the zebra finch (Balakrishnan *et al.* 2013). This indicates that selection on Z chromosome genes may be elevated in songbirds. More specifically, our results indicate that Z chromosome selection is higher when the water pipit is compared to the zebra finch (dN/dS approximately 0.29), than when the dunnock is compared to the zebra finch (dN/dS approximately 0.275). Consistent with the dN/dS levels for the other chromosomes investigated, selection on the Z chromosome appears lower between the water pipit and the dunnock (dN/dS approximately 0.25), than either compared to the zebra finch. Chromosome 9 displayed the lowest mean dN/dS, significantly lower than most other chromosomes, when the water pipit was compared to both the zebra finch and the dunnock, but not when the dunnock was aligned to the zebra finch. Balakrishnan et al. did not test chromosome 9 but instead found that chromosome 4A is the lowest (Balakrishnan *et al.* 2013). We find that mean dN/dS on chromosome 4A is comparable to chromosome 9 when the water pipit is compared to the zebra finch, but not in any other comparison (Fig. 10).

Tables 5-7 outline the significant enrichment or depletion of genes with dN/dS>1 assigned to GO slim terms. Terms such as mitochondrion, ribosome, translation and cytoplasm appear consistently

over-represented and contain a relatively large number of genes, indicating that selective pressures may be operating differentially between these species on these types of genes. However, these terms were also found to be over-represented in the transcriptomes generally (see Figs. 3 and 4). The only term found not to be over-represented in the whole transcriptome yet over-represented in the lists of genes with dN/dS>1 when both species were compared to the zebra finch was organelle, indicating that genes of this function are indeed evolving more rapidly than those in the zebra finch. The terms aging, cell death, and transferase activity transferring alkyl or aryl (other than methyl) groups were not enriched in the water pipit transcriptome overall but were in genes with dN/dS>1; in the dunnock, this was true of reproduction and nucleolus. The term sulfur compound metabolic process was enriched in the water pipit genes with dN/dS>1 compared to both the dunnock and zebra finch. Terms that were extremely under-represented, i.e. those where dN/dS was consistently less than 1 such that they were not present in these lists, such as external encapsulating structure, cell wall organization or biogenesis, and extracellular matrix organization, may indicate genes with function that is either under stabilising selection or exhibits the same degree of genetic drift across the species tested. It should be noted, however, that these categories generally possess relatively few genes, and hence they may not necessarily represent a consistent pattern. Also, extracellular matrix organization is found to be under-represented in both the water pipit and the dunnock transcriptomes generally, which may be biasing this result. In contrast, external encapsulating structure was over-represented in the whole water pipit transcriptome but not detected in the dunnock, so the finding that this category was depleted in genes with dN/dS over 1 in the water pipit, indicates that water pipit genes of this category are under similar rates of evolution as such genes in the zebra finch. Terms with relatively many genes typically under-represented where dN/dS>1 yet not depleted in the transcriptome overall are more accurate indicators of specific selective effects. In both the water pipit and dunnock, these included protein modification process and kinase activity. When the water pipit was compared to the zebra finch and the dunnock, ribonucleoprotein complex assembly was depleted. There were no categories that were specifically depleted in the dunnock compared to the other species.

4.4.6 Differential gene expression

The following results should be considered alongside the following caveat: the RNA samples were pooled for sequencing and hence natural variation in gene expression level could not be calculated. The expression variation estimated by the differential expression tool used, DESeq, when using single replicates is derived from the overall variation in expression level for all genes in the lists being compared. To increase statistical confidence in these results, further replicates are required. 62 genes were found to be statistically significantly differentially expressed (DE) by DESeq (Table 8) – 39 where expression was highest in the polygamous dunnock, and 23 where expression was highest in the monogamous water pipit. The most highly significant DE gene, where expression was high in the dunnock and low in the water pipit, was CYP2D6 which is involved in

biotransformation pathway for defence against oxidative stress (Meyer 1996). Additionally, another key component of this pathway was found to be significantly DE in the same pattern, glutathione S-transferase (Meyer 1996), indicating that the biotransformation pathway may be differentially modulated between these two species. The genes with the greatest magnitude of differential expression overall, being expressed in the dunnock but not in the water pipit, were two uncharacterised proteins and NECAB1, a neuronal calcium ion-binding protein (Sugita et al., 2002; Wu et al., 2007). The gene most highly differentially expressed that was expressed in both species was MLF1IP, related to centrosome function (Minoshima *et al.* 2005; Suzuki *et al.* 2007), which was again more highly expressed in the dunnock compared to the water pipit.

The most highly significant DE genes where expression was greater in the water pipit than the dunnock included several uncharacterised proteins, PTCD2, and LRRC34. PTCD2 is a member of the pentatricopeptide repeat-containing protein genes, involved in regulating mitochondrial gene expression in mammals (Lightowlers & Chrzanowska-Lightowlers, 2008; Rackham & Filipovska, 2011; Rackham et al., 2011; Xu et al., 2012). LRRC34, a leucine rich repeat containing protein, is putatively involved in ribosomal biogenesis, particularly in pluripotent stem cells (Lührig *et al.* 2014) but has also been linked to centrosomal structures (Firat-Karalar et al., 2014).

In terms of overall gene function, explored using gene functional categories, which may be differentially regulated between the water pipit and the dunnock, only one category was significantly differentially expressed: neurological system process. This was more highly expressed in the water pipit than the dunnock, which is slightly surprising given that the majority of significantly expressed genes were more highly expressed in the dunnock. One gene from the list of differentially expressed genes fell within this term: ENSTGUG00000010887, an uncharacterised protein that also was allocated to other terms for ligase activity, cell-cell signalling, cytoplasmic membrane-bounded vesicle, and transport, among others.

4.5 Discussion

Presented here are the brain transcriptomes of two non-model songbird species with no previously available genomic resources. Songbirds present excellent subjects for comparative phenotypic evolutionary studies, given their rapid radiation: there are many closely related species with well documented behavioural differences. Additionally, avian genomic resources have expanded dramatically in recent years, with the sequencing and annotation of the chicken (International Chicken Polymorphism Map Consortium 2004) and zebra finch (Warren *et al.* 2010) genomes, with many more on the way (Zhang *et al.* 2014). The water pipit and dunnock present an interesting ecological comparison of mating system evolution as the water pipit is primarily monogamous (Griffith et al., 2002; Reyer et al., 1997) whereas the dunnock is polygamous (Burke *et al.* 1989; Griffith *et al.* 2002). Morphological and sperm characterisation obtained from the individuals sampled is consistent with differing levels of sexual selection in the two species as expected from the reported differences in mating systems.

Using Illumina RNA-seq combined with the DGM transcriptome annotation strategy allowed the detection of expression of around 90% of zebra finch orthologs in the two species profiled. This is a markedly higher number of annotated genes than those obtained in comparable recent studies based on transcriptome assembly methods: Balakrishnan et al. (2013) identified 11,084 zebra finch genes using their assembled contigs and singletons of the violet-eared waxbill transcriptome (Balakrishnan *et al.* 2013), and Moghadam et al. (2013) detected expression of 8,963 chicken and 9,247 zebra finch 1:1 orthologues in the Kentish plover, respectively (Moghadam *et al.* 2013). This highlights the effectiveness of the DGM technique employed. We have shown previously that DGM performs significantly better than *de novo* or genome-guided assemblies in detecting genes in an accurate and unbiased fashion (Ockendon et al. submitted). Given the evolutionary divergence between the zebra finch and each of our study species (38.2 MYA for the water pipit, and 36.2 MYA for the dunnock; Fjeldså et al., 2010), and based on our previous characterisation of the effect of sequence divergence between the species profiled and that used as reference for the annotation we conservatively estimate a gene detection error rate of approximately 10% to these gene detection values.

Compared to a similar assessment of the functional bias of the violet-eared waxbill, as performed by Balakrishnan et al. (2013), we detected fewer terms as over-, or under-represented, indicating that our recovered transcriptome profiles were more similar to the expected form despite the higher level of divergence between the water pipit/dunnock and the zebra finch, compared to the waxbill and zebra finch (Balakrishnan *et al.* 2013). This may indicate reduced functional bias in our transcriptomes compared to theirs, or that the brain transcriptomes of our songbird species are more similar to that of the zebra finch than that of the waxbill is to the zebra finch. However, we used

hypergeometric tests whereas Balakrishnan et al. used Fisher's exact tests and, although they can give similar results (Rivals et al., 2007), depending on the sample size they approximate to different distributions. Indeed, using our data we observed more terms as enriched/depleted using the Fisher's exact test compared to the hypergeometric test (data not shown). As hypergeometric tests are known to be appropriate for sampling from small and large numbers of genes (Hong et al., 2014), approximating to the same class of distribution used by the tool we used for the differential expression analysis, we deemed this approach most suitable for these tests. Interestingly, many (approximately half) of the terms over-represented in the water pipit and dunnock transcriptomes were also identified as enriched in the waxbill transcriptome, suggesting that perhaps in songbirds genes with these sorts of functions are relatively slow evolving (data not shown). However, far fewer terms that were depleted were the same between the water pipit/dunnock transcriptomes and the waxbill, indicating that these terms may be more highly evolving between these species. Terms that were similarly depleted in the water pipit/dunnock and waxbill transcriptomes were signal transducer activity, signal transduction, cytoskeleton, and extracellular region.

Our detailed characterisation of the transcriptomes highlights evolutionary differences, such as SNP and indel occurrence, and rates of evolution (dN/dS), between these species and their closest available reference species, the zebra finch, and each other. The distributions of SNPs and indels follow similar trends as reported recently for a different songbird transcriptome (Balakrishnan *et al.* 2013), generally scaling with chromosome size. Enrichment and depletion of expressed functional terms may reflect either comparative gene expression differences of the water pipit and dunnock transcriptomes to that of the zebra finch, or the effect of sequence divergence between the water pipit/dunnock transcriptomes and the zebra finch reference genome in recovering a representative expressed gene list. Balakrishnan et al. (2013), in their exploration of the violet-eared waxbill transcriptome, do not appear to consider the latter effect, appearing to assume that bias was due only to failure to express genes of certain classes rather than a failure to detect them, although they do acknowledge their inability to "attain 'complete' transcriptome coverage" (Balakrishnan *et al.* 2013). They detect many more terms as over-, or under-represented than we do here, which, given the lower reported divergence between the zebra finch and the waxbill compared to that between the zebra finch and the water pipit/dunnock, indicates that their transcriptome profile is indeed biased in the detection of gene function.

We have also explored rates of evolution, highlighting the quantities of transcripts and gene functional categories that may be subject to adaptive evolution, showing how, at the genome-wide scale, these two non-model species may be under similar selective effects. We report similar results to Balakrishnan et al. (2013) in terms of average rates of molecular evolution (dN/dS) per chromosome: the Z chromosome displays consistently the highest level and chromosome 4A

displays among the lowest (Balakrishnan *et al.* 2013). High levels of Z chromosome synteny and conservation have been reported across many bird species and selection on the Z chromosome may be related to the evolution of male sexually selected traits, particularly in species with a ZW sex chromosome system where the trend is towards male-biased expression (Kirkpatrick & Hall 2004). Our results indicate that there is a greater difference in dN/dS between the Z chromosome and the other macro chromosomes in the water pipit than in the dunnock. As evolution under different mating systems is known to impact on sex biased gene expression, and the Z chromosome is important in dosage compensation, although at the level of the gene in avian species (Mank & Ellegren 2009), these findings may reflect the differential impacts of mating system on Z chromosome evolution. We find that overall dN/dS levels on the Z chromosome are higher between the water pipit and the zebra finch, compared to those between the dunnock and the zebra finch, or between the water pipit and the dunnock. Given that the water pipit is taken to be the outgroup species, this is not unexpected. Although, if the Z chromosome contains regions that are under sexual selection in songbirds, and assuming that the water pipits used in this study have been subject to lower sexual selection (as suggested by the sperm morphology data), this may represent lower evolutionary constraint increasing genetic drift compared to the dunnock. The fact that substitution rates overall are lower between the water pipit and the dunnock than between either species and the zebra finch indicates that the water pipit and dunnock are more similar overall at the level of the sequences generated in this study. This possibly indicates a closer evolutionary relationship than is currently documented. However, the fact that the zebra finch sequences were generated in a separate study and probably sequenced to greater depth should be considered as this may have resulted in artefactual differences between the data sets.

Considering those genes that appear to be adaptively evolving relative to the zebra finch, those expressed by the water pipit appear to be related to aging and cell death, whereas those enriched in the dunnock are related to reproduction and the nucleolus. It is known that mating system impacts upon longevity (Liker & Szkely 2005; Clutton-Brock & Isvaran 2007), and of course upon reproduction, therefore these findings may highlight categories of genes worthy of further exploration in conjunction with the differential expression results (see below).

These two species represent interesting comparative models of behaviour: opposing mating systems and song complexity. As such, differential expression analysis has permitted identification of genetic factors that may underlie these differences. This highlights the importance and usefulness of novel species in sequencing projects where genomic resources are not necessarily readily available but where interesting ecological traits are present. The limitations of our findings stem from the pooling of RNA samples prior to sequencing, negating the ability to robustly estimate natural variation in gene expression. Additionally, given that we were only able to collect

samples from one species pair, the differences in gene expression we report may not necessarily be related to behavioural differences but instead general physiological or behavioural differences. Hence to increase confidence that the genes we report are indeed related to mating systems in songbirds, further songbird species pairs with opposing mating systems should be used. We find 62 genes to be significantly differentially expressed between the two songbird species: these may relate to brain differences that either modulate, or are impacted by their respective behaviour. Of those genes more highly expressed in the polygamous dunnock than the monogamous water pipit, there are two genes of the biotransformation pathway, involved in defence against oxidative stress: CYP2D6 and glutathione S-transferase (see Meyer, 1996). This indicates that oxidative stress genes may either be involved with pathways facilitating behavioural differences around mating and song, or that the increased sexual selection experienced by the dunnock due to its mating preferences compared to the water pipit may have manifested in differential regulation of these pathways in these species, perhaps as a protective mechanism. Indeed, members of the Ritchie lab, University of St. Andrews, have found that this pathway, and indeed one of the same genes that we identify as differentially expressed (glutathione S-transferase) may be involved in differences between monogamous and polygamous *Drosophila* species (PopGroup 2013 presentation by Dr. Paris Veltsos and personal communication). If so, this may represent an evolutionarily conserved path of either how sexual selection manifests within the brain, or how differences in mating behaviour are mediated alongside neuroprotective mechanisms. Sexual selection is known to act on the development of sexually dimorphic ornamentation commonly seen in birds, such as elaborate plumage and wattles often with carotendoid-based colours. As these features have been found to predict sensitivity to oxidative stress (Mougeot et al., 2010), this therefore provides a clear way for females to be able to visually detect the relative fitness of potential mates (von Schantz et al., 1999). Males engaged in intermale competition have been found to exhibit high circulating stress hormone (glucocorticoid) levels (Orchinik et al., 1988; Reedy et al., 2014), indicating that in species where intermale competition is high, i.e. polygamous species, increased stress hormone levels may be an important difference to species where intermale competition is low, i.e. monogamous species. Considering that glucocorticoids promote oxidative stress, and that brain tissue is highly susceptible to this (Costantini et al., 2011), it is tempting to postulate that, compared to males of monogamous species, males of polygamous species typically experience higher levels of circulating stress hormones which promotes oxidative stress within the brain and that coordinated expression of biotransformation pathway genes has hence evolved alongside polygamous tendencies to mitigate the impacts of heightened oxidative damage. If so, it may be that regulation of these genes crosstalks with pathways in the brain regulating polygamous behaviour. Indeed CYP2D6 has been shown to exert modest effects on the formation of specific oestrogen metabolites, as do many of the other cytochrome P450 isoforms (Zhu & Lee 2005), which may represent a route to modulation of behaviourally-relevant substrates. Oestrogens are documented to exert neuroprotective effects, resisting the effects of oxidative stress (Behl *et al.* 1997). These effects have recently been linked to mitochondrial function modulating cell survival

(Simpkins et al., 2010) and interestingly, two mitochondrial genes are significantly more highly expressed in the dunnock compared to the water pipit: NADH-dehydrogenase subunit 6 (ND6), a key part of the complex 1 of the electron transport chain, and G elongation factor mitochondrial 2 (GFM2), a mitochondrial gene expression facilitator. IL-18 was also significantly up-regulated in the polygamous dunnock compared to the monogamous water pipit. IL-18 is a proinflammatory cytokine documented to increase in production during periods of oxidative stress following hypoxia (Ikonomidou & Kaindl 2011) and following brain injury (Felderhoff-Mueser et al., 2005), and can be neuroprotective against infection (Kawakami *et al.* 1997). This lends additional weight to the possibility that brain gene expression in these polygamous songbirds is impacted by increased oxidative stress, potentially deriving from intermale competition, to a greater extent than the monogamous species. High oxidative stress has been shown to impair mitochondrial function, disrupting the healthy energy balance within the brain which facilitates neurotransmission and plasticity (Ikonomidou & Kaindl 2011; Picard & McEwen 2014). As such, and given that these findings suggest increased biotransformation pathway activity in the dunnock brain which may lower or balance enhanced oxidative stress levels, it seems that during the breeding season, gene expression in the dunnock brain is focused on managing overall brain health compared to the water pipit. CEP89, a centrosomal protein (Jakobsen et al., 2011; Kumar et al., 2013) is more highly expressed in the dunnock than the water pipit, and has been found to be involved in mitochondrial and neuronal function (van Bon *et al.* 2013). The function of centrosomes during neurogenesis remains unclear although it is apparent that the expression, localisation and function of these structures are highly important in establishing cytoskeletal polarisation, impacting key processes of neurogenesis: proliferation, migration, and differentiation (see Higginbotham & Gleeson, 2007). Another gene significantly more highly expressed in the dunnock was MLF1IP, which is linked to centromere function (Minoshima *et al.* 2005), indicating that pathways regulating mitosis are up-regulated in the dunnock compared to the water pipit.

Genes more highly expressed in the water pipit than the dunnock included PTCD2, LRRC34, a putative ribosomal protein, LOC100223017, and LIPA. PTCD2 and LRRC34 both appear to be involved in modulation of RNA transcription, be that either gene expression or ribosomal RNA expression (Lightowlers & Chrzanowska-Lightowlers, 2008; Lührig et al., 2014; Rackham & Filipovska, 2011; Rackham et al., 2011; Xu et al., 2012). PTCD2 is linked to mitochondrial enzyme complex function (Lightowlers & Chrzanowska-Lightowlers 2008; Xu *et al.* 2012) and specifically has been shown to decrease levels of the mitochondrial long non-coding (lnc) RNAs including lncND6, which is located within the region complementary to the ND6 gene (Rackham *et al.* 2011) – which we found to be more highly expressed in the dunnock compared to the water pipit. LncND6, along with other mitochondrial lncRNAs can form double-stranded intermolecular complexes, possibly presenting a mechanism by which they can regulate the availability of their complementary coding counterparts (Rackham *et al.* 2011). If so, this may suggest that ND6

expression is differentially modulated between the dunnock and the water pipit, possible via the activity of PTCD2 in the water pipit at least. LRRC34 has been shown to be a marker of pluripotent stem cells and may be involved in regulation of pluripotent cell ribosomal biogenesis (Lührig et al., 2014), but has also been linked to centrosome function (Firat-Karalar et al., 2014), indicating that processes related to neurogenesis and mitosis specifically may be also occurring in the water pipit but perhaps according to different patterns or modes compared to the dunnock. LOC100223017 is putative ribosomal protein L35a, part of the large ribosomal subunit important in protein synthesis (Herzog et al., 1990), which has been implicated in inhibiting cell death (Lopez et al., 2002) and identified as commonly over-expressed in malignant brain tumours (Kroes *et al.* 2000). LIPA, lysosomal acid lipase A, cholesteryl ester hydrolase, has been recently found to correlate strongly with brain phospholipid levels and its expression increased in brain tissue of humans who suffered violet death suicides (Freemantle et al., 2013), indicating a role in behavioural modulation.

Overall, these results suggest that there were some key differences between molecular pathways operating within the dunnock and water pipit brains, such as defence against oxidative stress in the dunnock and ribosomal function in the water pipit. Additionally, there appear to be some similar functions occurring but modulated by different genetic components, such as mitochondrial function and neurogenesis. The latter process is particularly relevant in songbirds during their breeding season where the higher vocal centre expands significantly, involving the generation and recruitment of new neurons (Louissaint et al., 2002; Tramontin & Brenowitz, 2000). These differences may reflect the differing neural priorities and capabilities for brains responding to low and high levels of sexual selection, respectively, and/or the different behavioural scenarios and the internal and external responses that those scenarios necessitate, in terms of both gene expression, neurogenesis and energy balance.

4.6 Conclusions

Here we demonstrate the usefulness of non-model species without sequenced genomes in comparative genomic studies aiming to explore the molecular basis of phenotypic differences. The preferred transcriptome annotation method DGM detects many more genes than similar recent studies have done using more common assembly-based methods. Performing functional investigations of genes detected and of sequences with interesting evolutionary characteristics, we have shown that water pipit and dunnock brain transcriptomes are functionally similar to each other, with respect to their common closest reference species, the zebra finch. Key differences have been noted, indicating points of potential functional variation in gene regulation that may underlie phenotypic differences between these species, and statistically significantly differentially expressed genes and functional categories have been identified which provide a first port of call for further exploring the molecular basis of mating system evolution in songbirds. Future work should aim to obtain further biological replicates to increase statistical confidence in the genes and pathways detected as differentially expressed. In particular, obtaining female brain samples would enable comparison of sexually dimorphic gene expression occurring in the context of different mating systems (Pointer *et al.* 2013; Hollis *et al.* 2014), which may enable identification of sex-role-specific versus core pathways underlying broad traits in mating system. Additionally, it would be interesting to explore the expression profiles of specific regions of the songbird brain, particularly within the nodes of the animal social decision making network (O'Connell & Hofmann 2012b) to observe whether the highly complex mating dynamics of songbirds are integrated within this, or whether additional regions are recruited.

4.7 Figure legends

Fig. 1. **A: Within-male sperm variance measures in dunnock (n=5) and water pipit ('Wpipit', n=4). B: Between-male variance in sperm length versus EPY in passerine birds.** Data from Lifjeld et al. 2010, with our dunnock and water pipit data added (red dots, left and right, respectively). Light micrographs of C: water pipit sperm, and D: dunnock sperm. Performed by Dr. Alexander Ball, University of Bath, Szekely lab.

Fig. 2. **Variance in morphological characteristics for the dunnock and water pipit.** A: cloacal protuberance volume; B: testes volume.

Fig 3. **Over-represented zebra finch GO slim terms detected in the water pipit transcriptome (hypergeometric test).** A: The blue section illustrates the proportion of zebra finch genes detected in the water pipit transcriptome, whereas the red portion reflects that not detected. B: The table provides the number of zebra finch genes it is possible to detect in each GO slim category.

Fig. 4. **Over-represented zebra finch GO slim terms detected in the dunnock transcriptome (hypergeometric test).** A: The blue section illustrates the proportion of zebra finch genes detected in the water pipit transcriptome, whereas the red portion reflects that not detected. B: The table provides the number of zebra finch genes it is possible to detect in each GO slim category.

Fig. 5. **Under-represented zebra finch GO slim terms detected in the water pipit transcriptome (hypergeometric test).** A: The blue section illustrates the proportion of zebra finch genes detected in the water pipit transcriptome, whereas the red portion reflects that not detected. B: The table provides the number of zebra finch genes it is possible to detect in each GO slim category.

Fig. 6. **Under-represented zebra finch GO slim terms detected in the dunnock transcriptome (hypergeometric test).** A: The blue section illustrates the proportion of zebra finch genes detected in the water pipit transcriptome, whereas the red portion reflects that not detected. B: The table provides the number of zebra finch genes it is possible to detect in each GO slim category.

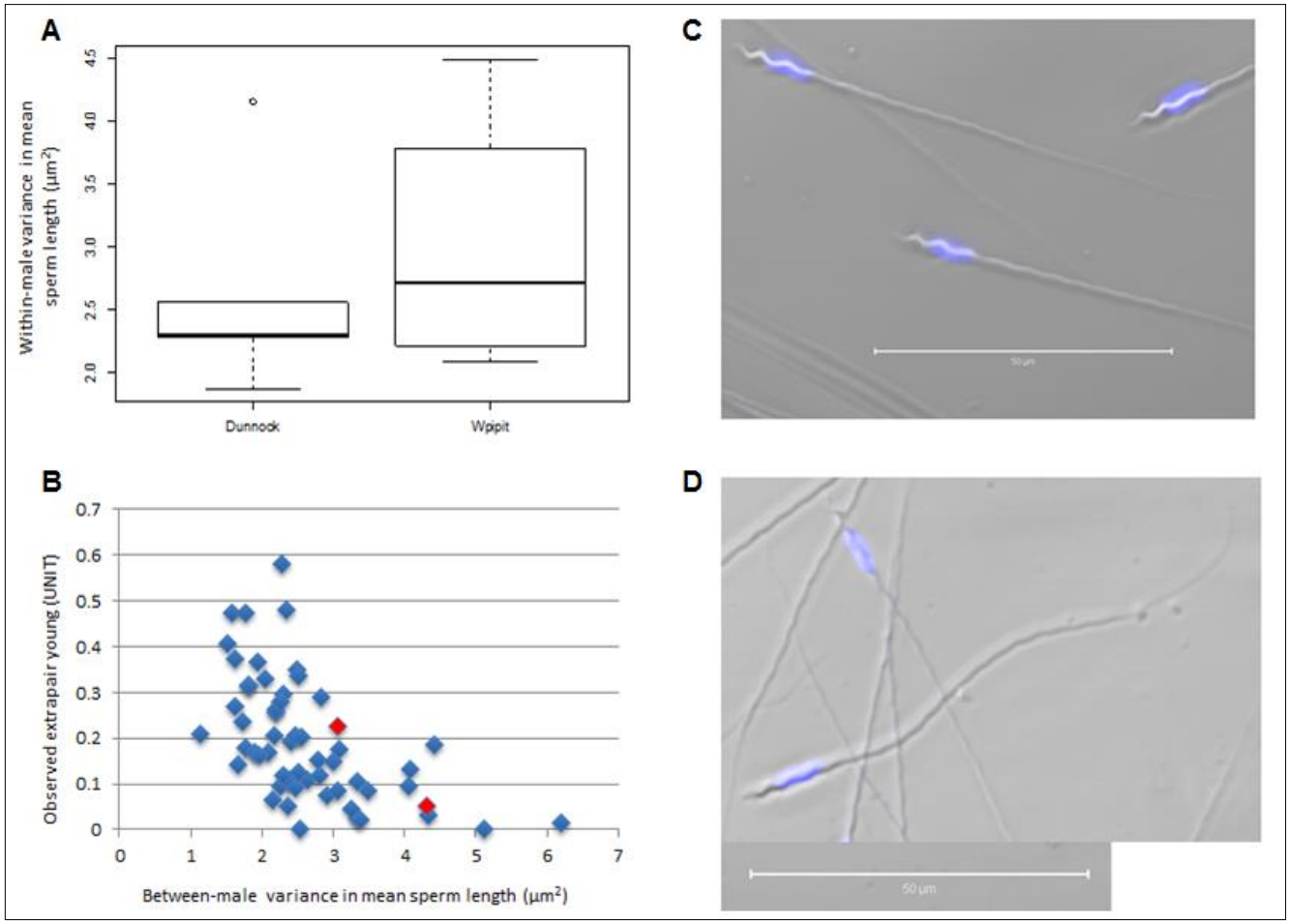
Fig. 7. **Indels identified using the zebra finch genome against chromosome size identified in the (A) water pipit, and (B) dunnock.** Indel frequency scales well with chromosome size.

Fig. 8. **SNPs identified using the zebra finch genome against chromosome size identified in the (A) water pipit, and (B) dunnock.** Again, SNP frequency scales well with chromosome size.

Fig. 9. **SNPs shared with the zebra finch, identified using the zebra finch genome against chromosome size identified in the (A) water pipit, and (B) dunnock.** Shared SNP frequency does not scale as well with chromosome size as indels or unshared SNPs, plus chromosome 13 is a clear outlier.

Fig. 10. **Rates of molecular evolution (dN/dS) collated per macrochromosome.** The Z chromosome displays consistently the greatest mean dN/dS values, as has been found previously in a songbird transcriptome (Balakrishnan *et al.* 2013). A: water pipit:zebra finch; B: dunnock:zebra finch; C: water pipit:dunnock.

Fig. 1.


A
Within-male variance in mean sperm length (µm²)
2.0
2.5
3.0
3.5
4.0
4.5
Dunnock
Wpipit
B
Observed extrapair young (UNIT)
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0
1
2
3
4
5
6
7
Between-male variance in mean sperm length (µm²)
C
50 µm
D
50 µm

140

Fig. 2.

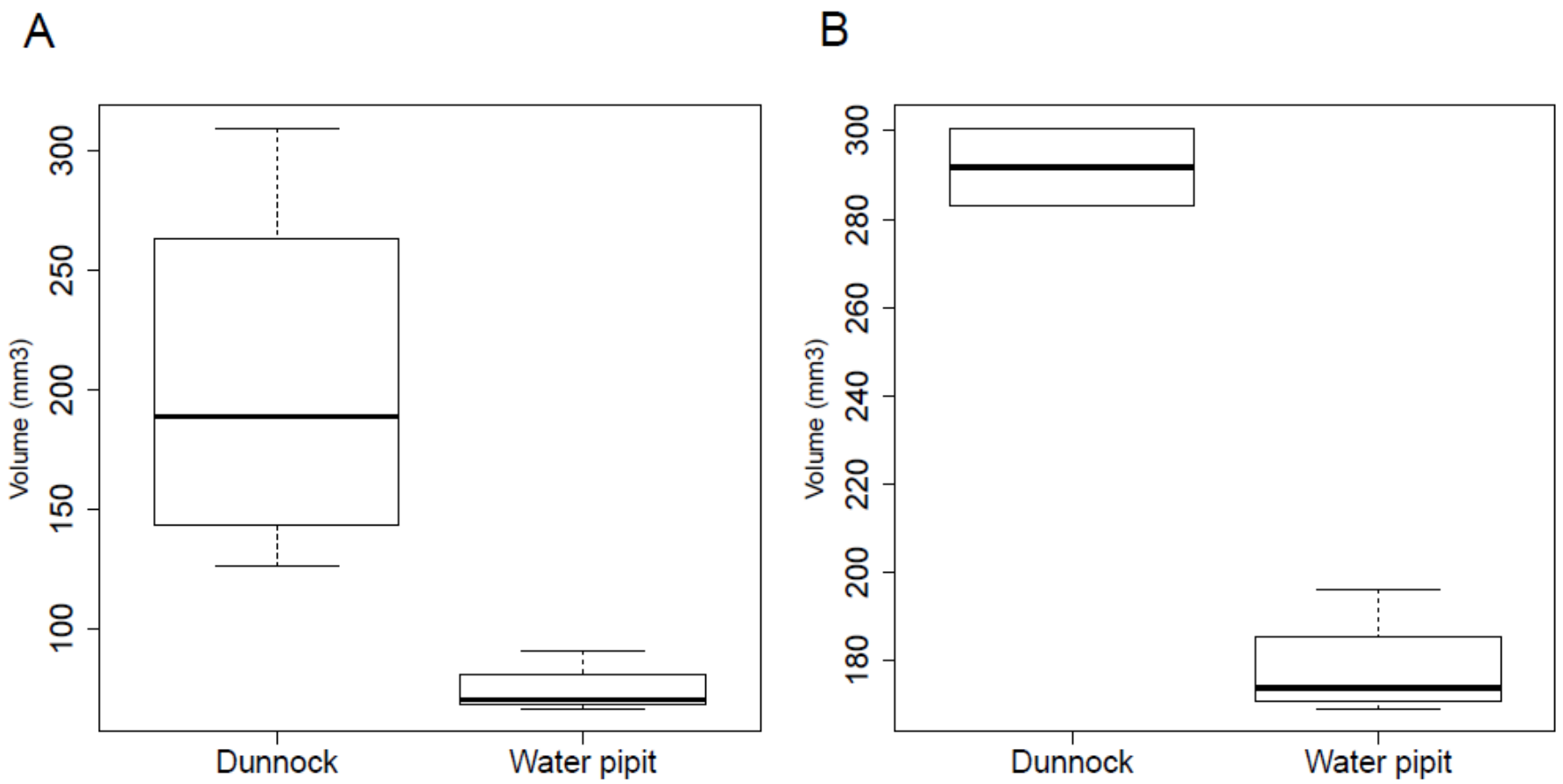

A
Volume (mm3)
100
150
200
250
300
Dunnock
Water pipit
B
Volume (mm3)
180
200
220
240
260
280
300
Dunnock
Water pipit

Fig. 3.

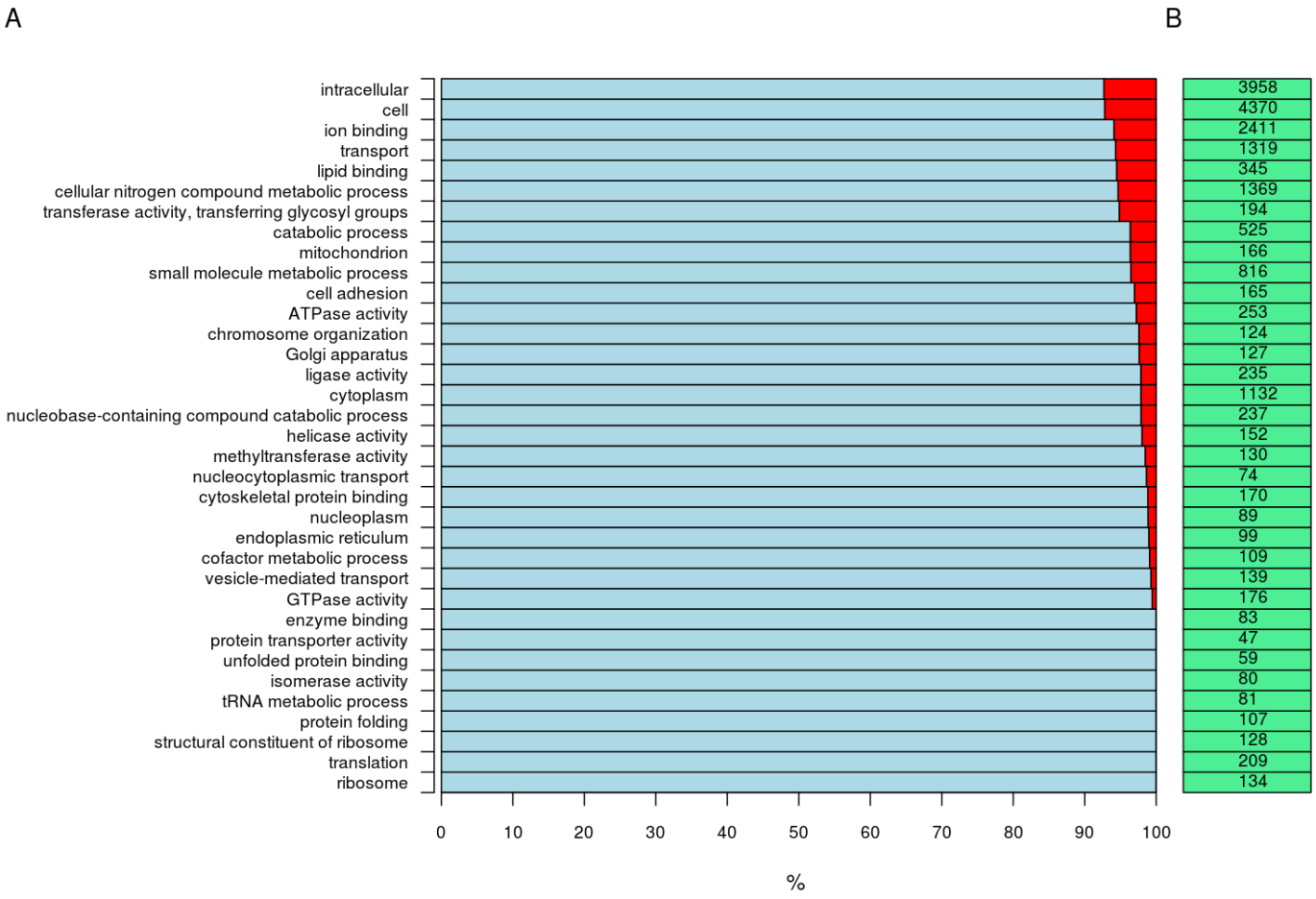

A
B
intracellular
cell
ion binding
transport
lipid binding
cellular nitrogen compound metabolic process
transferase activity, transferring glycosyl groups
catabolic process
mitochondrion
small molecule metabolic process
cell adhesion
ATPase activity
chromosome organization
Golgi apparatus
ligase activity
cytoplasm
nucleobase-containing compound catabolic process
helicase activity
methyltransferase activity
nucleocytoplasmic transport
cytoskeletal protein binding
nucleoplasm
endoplasmic reticulum
cofactor metabolic process
vesicle-mediated transport
GTPase activity
enzyme binding
protein transporter activity
unfolded protein binding
isomerase activity
tRNA metabolic process
protein folding
structural constituent of ribosome
translation
ribosome
3958
4370
2411
1319
345
1369
194
525
166
816
165
253
124
127
235
1132
237
152
130
74
170
89
99
109
139
176
83
47
59
80
81
107
128
209
134
0
10
20
30
40
50
60
70
80
90
100
%

Fig. 4.

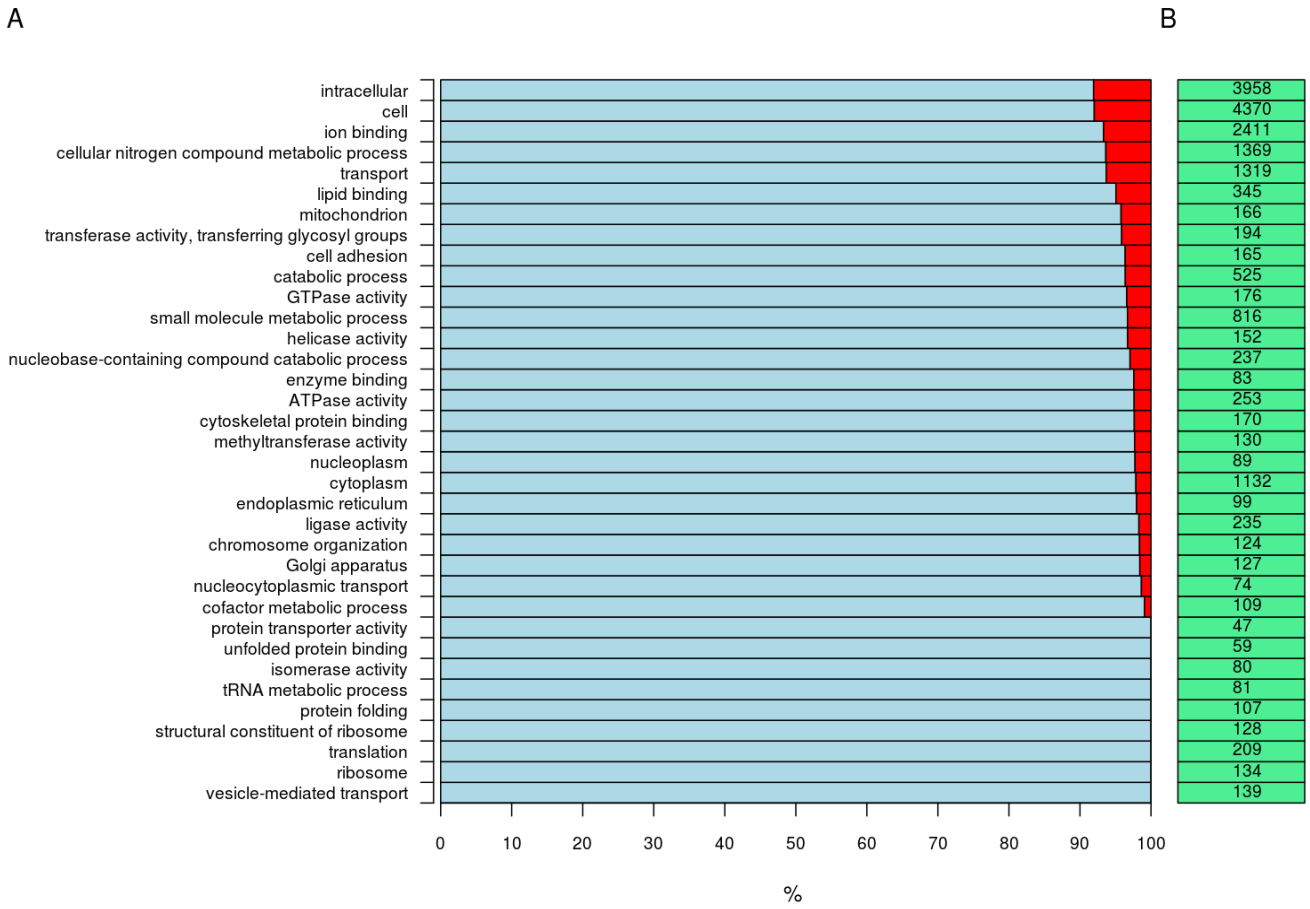

A
B
intracellular
cell
ion binding
cellular nitrogen compound metabolic process
transport
lipid binding
mitochondrion
transferase activity, transferring glycosyl groups
cell adhesion
catabolic process
GTPase activity
small molecule metabolic process
helicase activity
nucleobase-containing compound catabolic process
enzyme binding
ATPase activity
cytoskeletal protein binding
methyltransferase activity
nucleoplasm
cytoplasm
endoplasmic reticulum
ligase activity
chromosome organization
Golgi apparatus
nucleocytoplasmic transport
cofactor metabolic process
protein transporter activity
unfolded protein binding
isomerase activity
tRNA metabolic process
protein folding
structural constituent of ribosome
translation
ribosome
vesicle-mediated transport
3958
4370
2411
1369
1319
345
166
194
165
525
176
816
152
237
83
253
170
130
89
1132
99
235
124
127
74
109
47
59
80
81
107
128
209
134
139
0
10
20
30
40
50
60
70
80
90
100
%

Fig. 5.

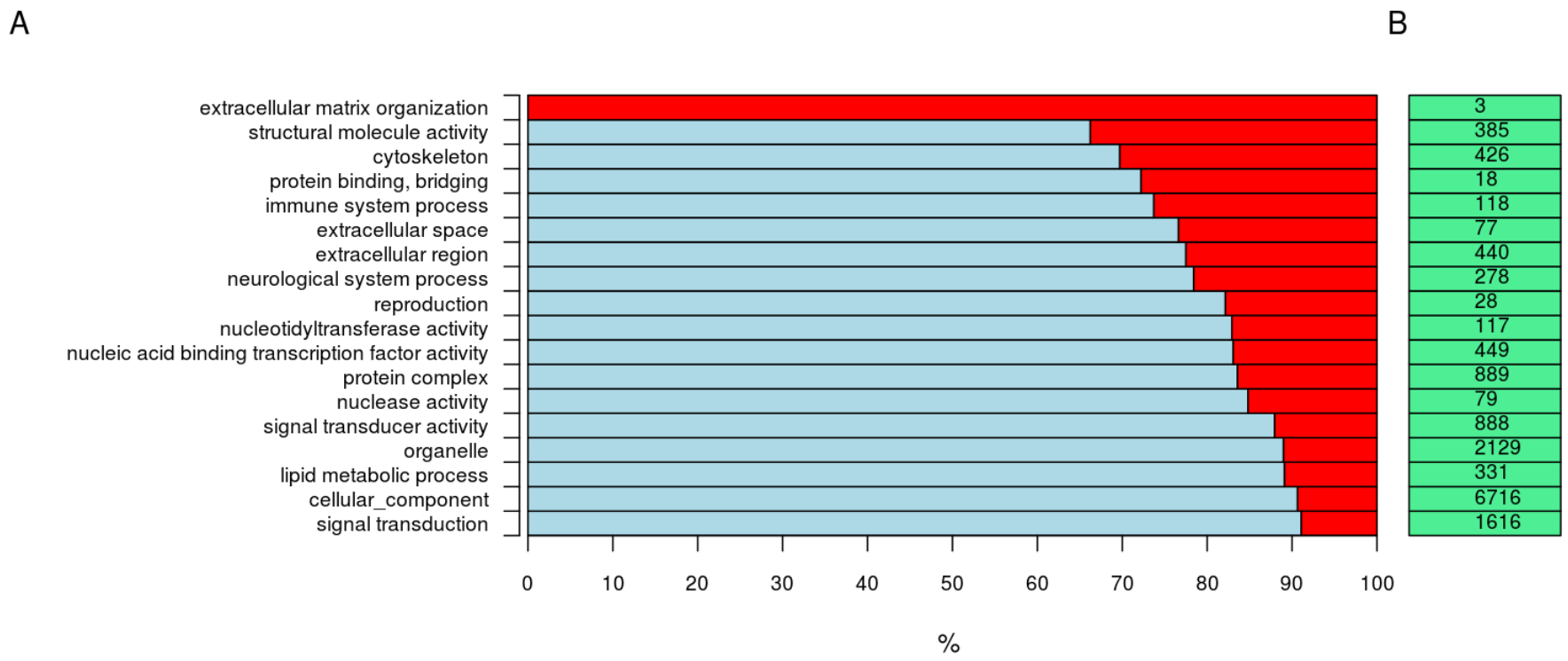

A
B
extracellular matrix organization
structural molecule activity
cytoskeleton
protein binding, bridging
immune system process
extracellular space
extracellular region
neurological system process
reproduction
nucleotidyltransferase activity
nucleic acid binding transcription factor activity
protein complex
nuclease activity
signal transducer activity
organelle
lipid metabolic process
cellular_component
signal transduction
3
385
426
18
118
77
440
278
28
117
449
889
79
888
2129
331
6716
1616
0
10
20
30
40
50
60
70
80
90
100
%

Fig. 6

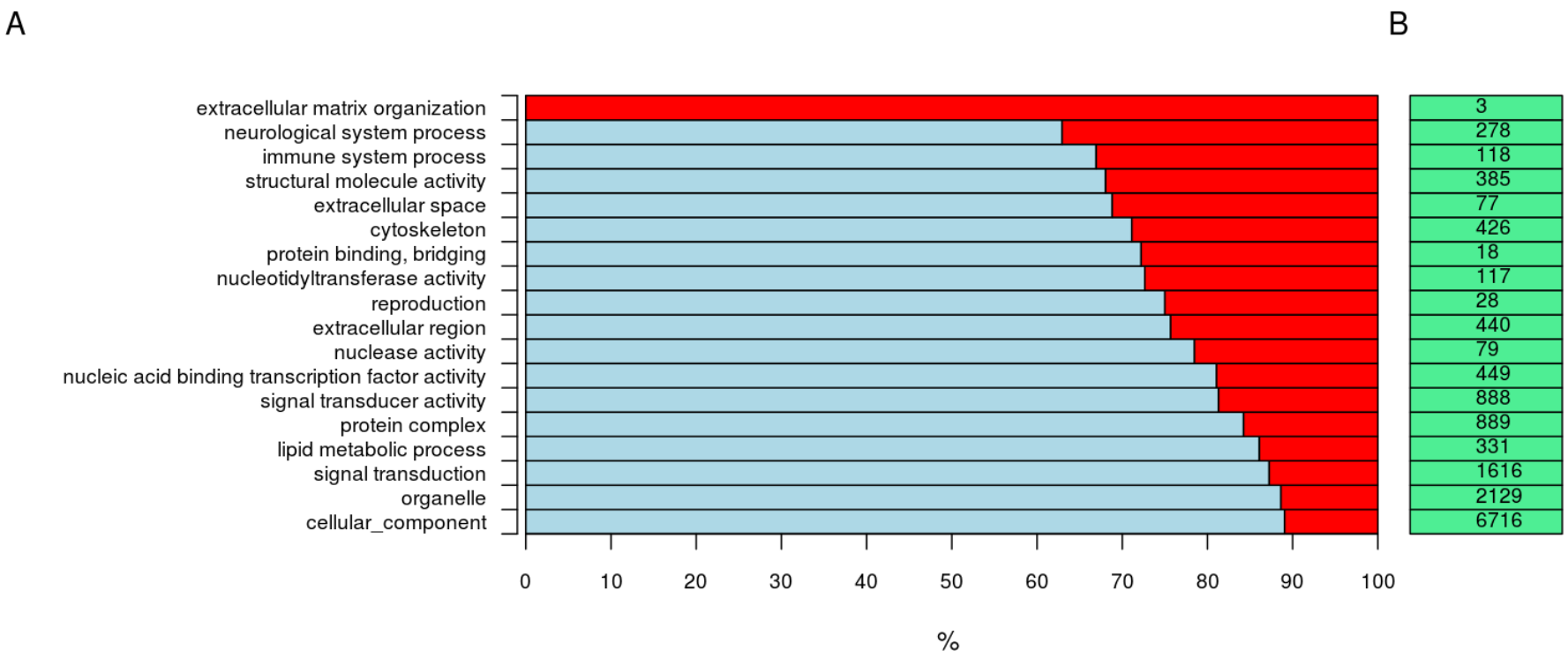

A
B
extracellular matrix organization
neurological system process
immune system process
structural molecule activity
extracellular space
cytoskeleton
protein binding, bridging
nucleotidyltransferase activity
reproduction
extracellular region
nuclease activity
nucleic acid binding transcription factor activity
signal transducer activity
protein complex
lipid metabolic process
signal transduction
organelle
cellular_component
3
278
118
385
77
426
18
117
28
440
79
449
888
889
331
1616
2129
6716
0
10
20
30
40
50
60
70
80
90
100
%

Fig. 7.

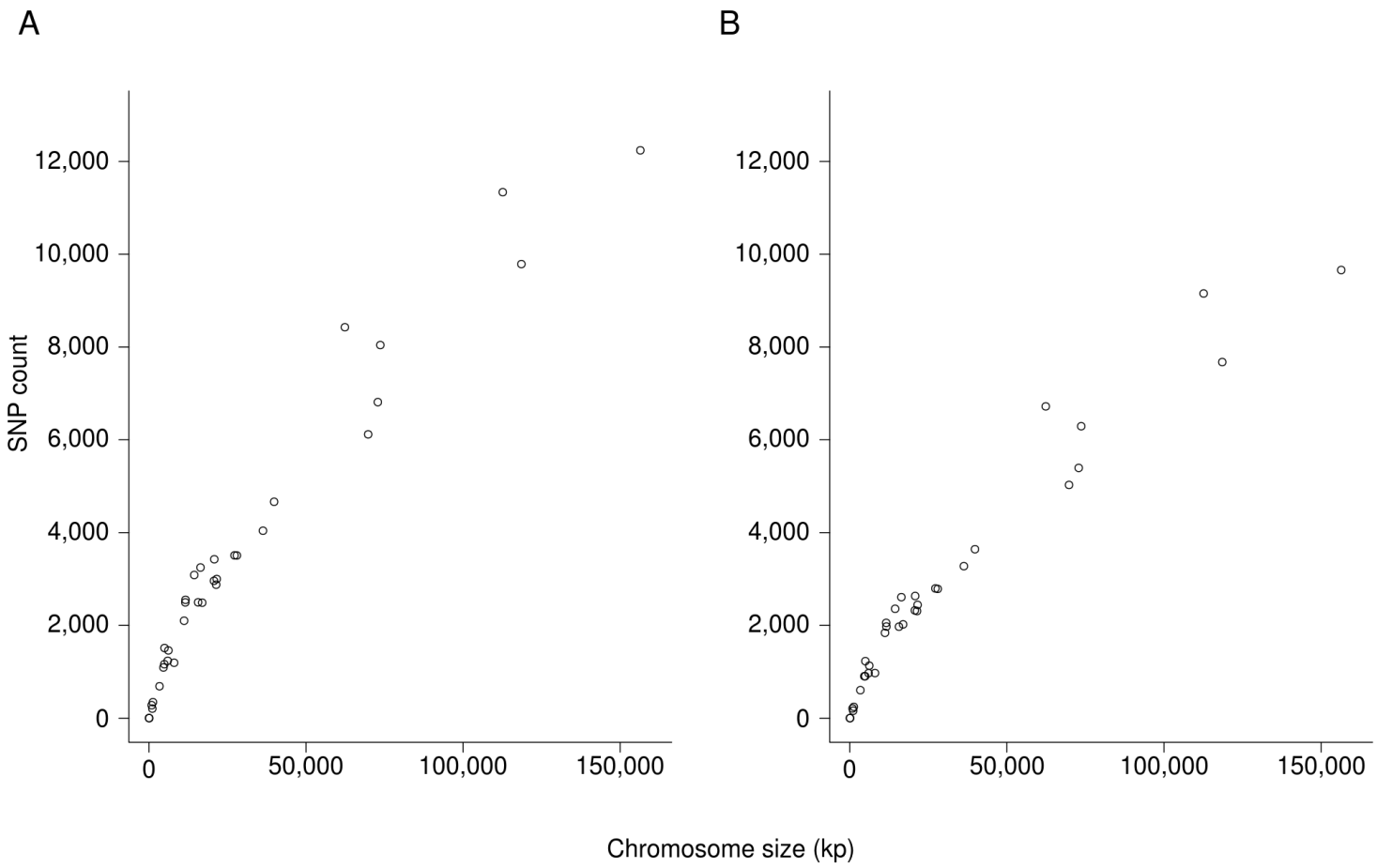

A
B
SNP count
12,000
10,000
8,000
6,000
4,000
2,000
0
0
50,000
100,000
150,000
Chromosome size (kp)

Fig. 8.

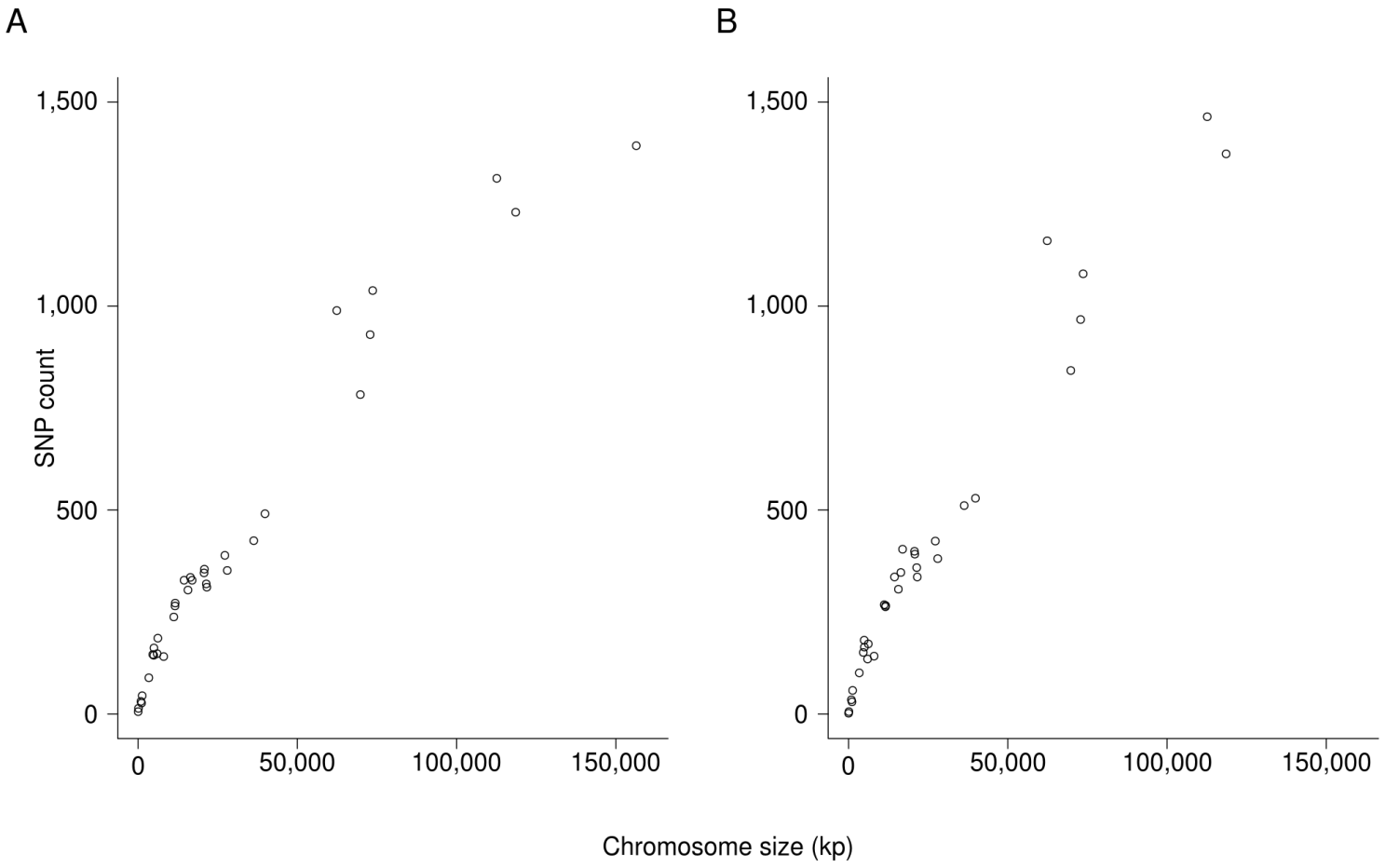

A
B
1,500
1,000
500
0
SNP count
0
50,000
100,000
150,000
Chromosome size (kp)

Fig. 9.

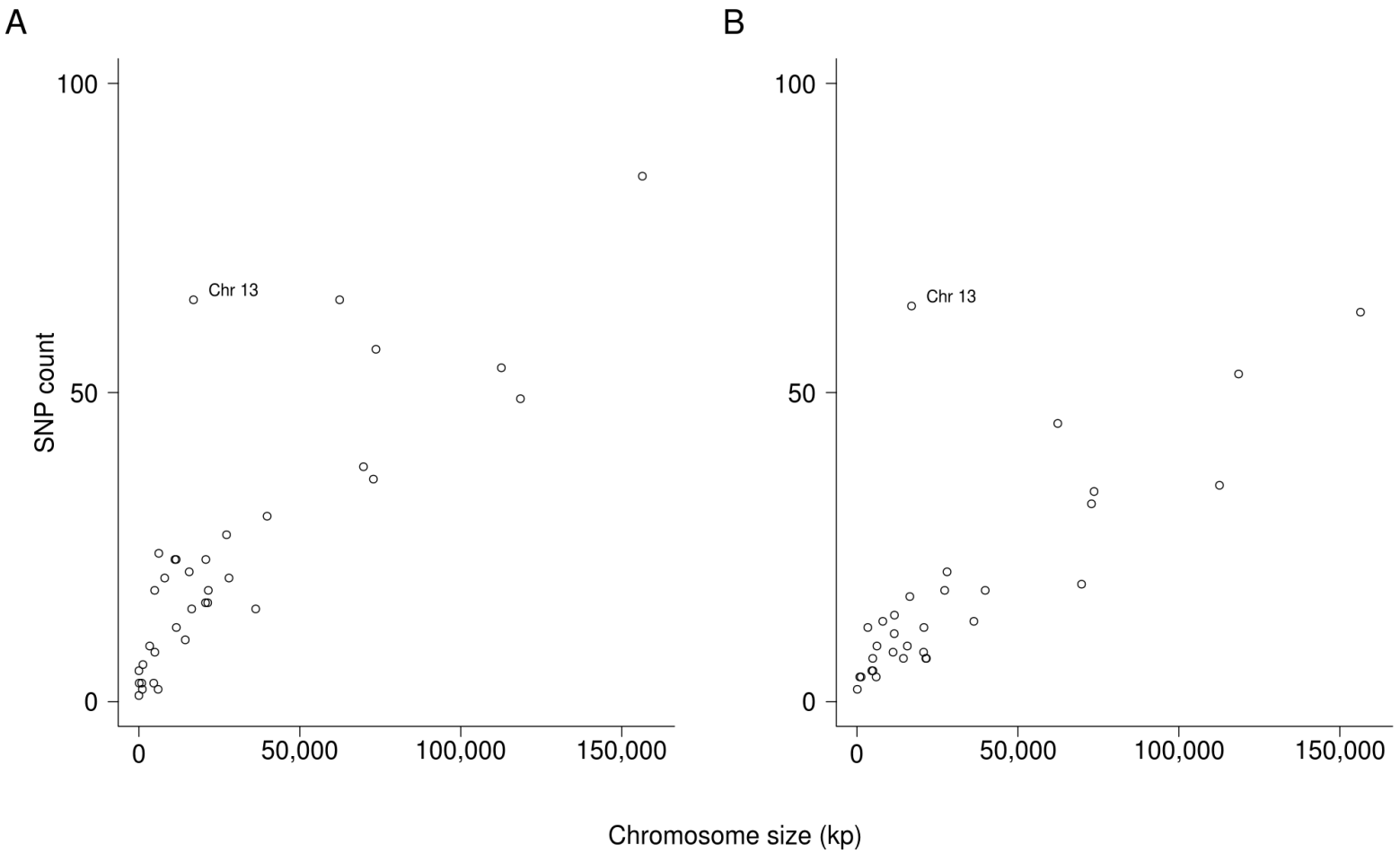

A
B
100
50
0
SNP count
Chr 13
Chr 13
0
50,000
100,000
150,000
Chromosome size (kp)

Fig. 10.

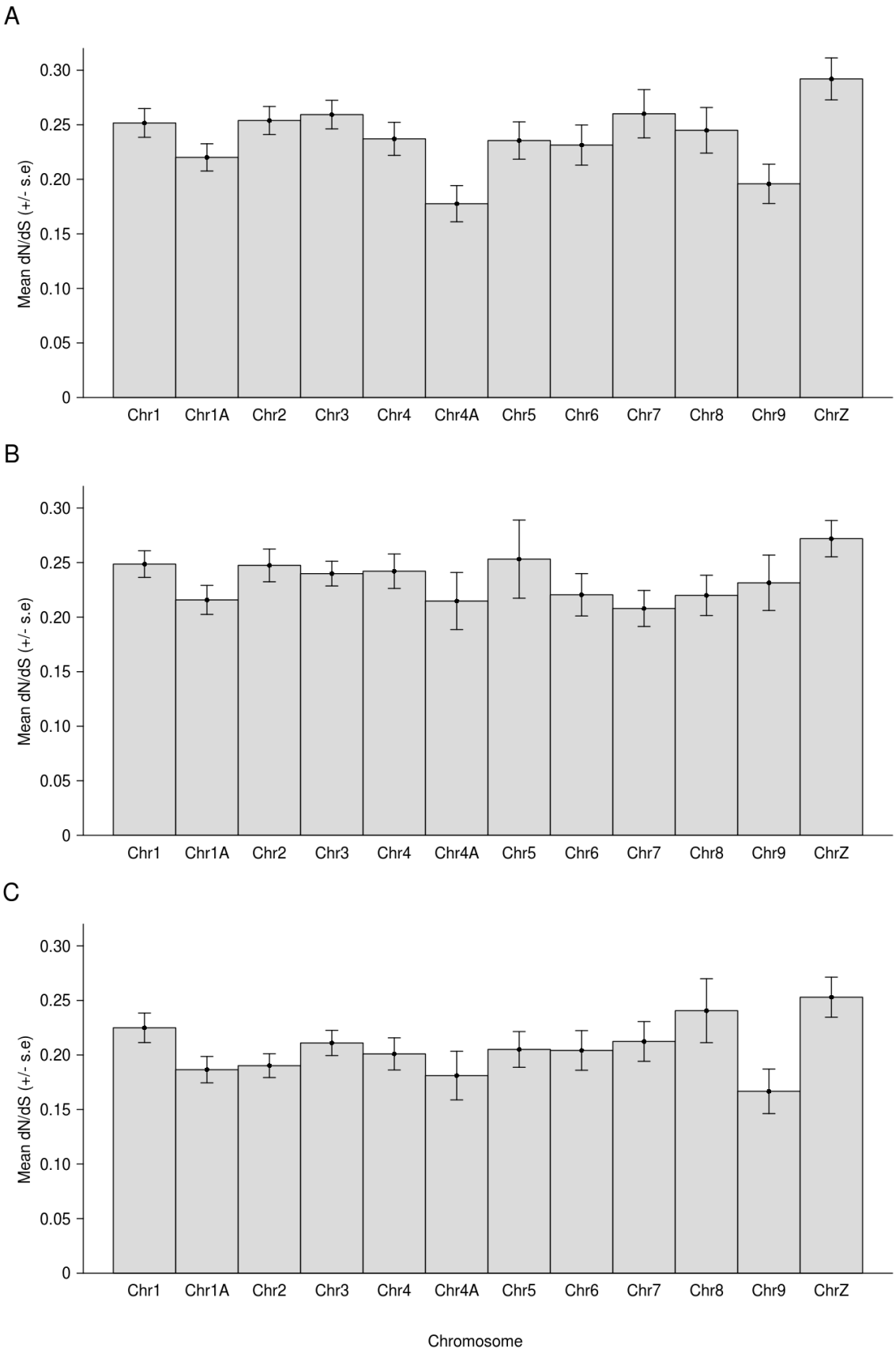

A
B
C
Mean dN/dS (+/- s.e)
0
0.05
0.10
0.15
0.20
0.25
0.30
Chr1
Chr1A
Chr2
Chr3
Chr4
Chr4A
Chr5
Chr6
Chr7
Chr8
Chr9
ChrZ
Chromosome

Table 1: Descriptive comparison of traits and ecology of the water pipit, *Anthus spinoletta*, and the dunnock, *Prunella modularis*.

Trait	Water pipit	Dunnock
Field characters	17-17.5cm, wing-span 24-29cm.	14.5cm, wing-span 19-21cm. Ground-creeping passerine.
Wing length	Male 88-96mm, female 82-90mm.	Male 68-74mm, female 65-72mm.
Weight	Both sexes mostly 19-27g	Mostly male 17-25g, female 16-24g.
Habitat	Montane, breeds in western Palearctic, middle and lower latitudes at considerable elevations (in Switzerland, infrequently below 1400-1800m and up to +2600m). Prefers area where stunted trees and sparse ground cover/ moist meadows, often near glaciers, on steep bare crags, even above snow line. Descends in Winter to lower ground or banks of mountain streams, Spring: boggy lowland with shrubs, sandy lowlands and arable land. In Western Europe, descends to flooded lowlands or damp meadows, watercress beds, estuaries and seashores including mudflats.	Upper and middle latitudes, mainly temperate but marginally subarctic, boreal, Mediterranean, between July isotherms 13-26^{o}C. Apparently evolved in scrub and stunted coniferous arctic-alpine and wooded tundra habitats, which it still occupies in south and north-east of range. In southern areas, mainly near tree line in mountains. In north, mainly spruce but also mixed and broad-leaf woodland, esp. along rivers/ streams.

Food	Mainly invertebrates, some plant material. Feeds mainly on the ground, sometimes catches insects in flight. In cold spells at high altitudes sometimes feeds around burrow entrances of marmots.	Mainly insects but with large proportion of small seeds in winter. Predominantly a ground feeder, under bushes, hedges, young conifers, among piles of twigs, roots, leaf litter. Steady hop, ceaseless pecking, never retraces steps, will peck seeds from vegetation.
Social pattern & behaviour	Solitary and gregarious outside breeding season. Reports that birds use the same area for feeding over several weeks. Flocks occur for roosting, and, usually loose-knit, for feeding: often 2-5, sometimes 20-60 or up to +200. Territorial in breeding season, usually monogamous.	Essentially solitary outside breeding season, occupying individual home ranges but can form local feeding aggregations. Male and female home ranges are independent and residents are dominant over intruders.
Breeding	Two broods. Site: steep bank or hollow, well concealed with overhanging vegetation, sometimes at end of short tunnel. Nest: cup of grass stems, leaves, moss, with slight lining of finer leaves and some hairs. Eggs: sub-elliptical, smooth, glossy. Grey-white, heavily mottled brown and grey, sometimes with dark zone or cap at broad end. Clutch: 4-6(-7). Incubation: 14-15 days. Fledging: 14-15 days. Presence of extra pair young in the nest is characterised by	Two, occasionally three broods. Site: bush, hedge, low tree, bank side, normally well concealed. Nest: substantial cup of twigs, leaves, stems, roots and other plant material, lined with wool, hair, moss, sometimes feathers. Eggs: sub-elliptical, smooth, glossy, bright blue and rarely also with some reddish spots. Clutch: 4-6 (3-7). Incubation: 12-13 days. Fledging: 11-12 days. Variable mating system including monogamy, polyandry

	asynchronous clutch initiation. Intraspecific brood parasitism, resulting from egg dumping, is characterised by a greater overlap with neighbouring territories and closer proximity to communal feeding sites, suggesting that EPP occurs more as a chance event related to ecological factors rather than female search for genetic/ phenotypic benefits. Additionally: occasional polygyny, slightly male-biased sex ratio, annual mortality rate of 44% (see Rehsteiner et al., 1998). Thus, some 'bachelors' per season, and expect strong selection on traits that improve chances in male-male competition and female attraction. Reproductive failure generally due to either snow or adder predation.	(usually one female and two males), and polygynandry (usually two females and two males). Where more than one member of one sex is present at the nest, there will be a dominant and a subordinate. Dominant males have greater access to females and both types of males will provide parental care.
Song	Song element, the 'snarr' has been identified as key for mating success: high 'snarr' males were mated more (irrespective of offspring survival or no. offspring) and their territories overlapped less with those of neighbours. Frequency of snarr correlated with body condition (weight) but not male age, territory size, quality of territory (food) and paternal performance. Therefore, high snarr scores likely	Male dunnocks have highly complex song. Their territories often overlap, sometimes completely, where a dominance hierarchy results although both males will sing (Birkhead 1981, Snow and Snow 1982). Male song repertoire includes a number of song types, each of which contain passages that are highly similar to neighbouring males' songs. As such, the song types within a male's repertoire are more different to each other than

	to represent greater social dominance rather than females preferentially choosing high snarr males. Male song duration approx. 15s, repertoire size of 3-4 elements, and used 2-3 elements per song. Sequence of the elements and other specific features of a males' song was determined during the first year of life and did not change thereafter (Rehsteiner *et al.* 1998)	they are to those of a neighbouring male. A repertoire varies to a small extent year on year: song types may be modified or lost all together (Snow and Snow 2009).

Table 2. RNA-seq data details: sequence quantities and quality scoring.

Species sequencing run	Raw reads		Pre-processed reads		
	Number of paired reads	Mean base quality	Number of reads	Mean base quality	Number of paired reads
Water pipit, lane 1	45,420,488	33.6 forward, 34.0 reverse	31,852,271 forward, 34,343,307 reverse	35.9 forward, 36.3 reverse	25,624,746
Water pipit, lane 2	46,180,373	33.4 forward, 33.8 reverse	32,172,205 forward, 34,672,942 reverse	35.8 forward, 36.2 reverse	25,820,618
Dunnock, lane 1	28,421,675	32.9 forward, 33.2 reverse	19,649,969 forward, 20,327,899 reverse	35.3 forward, 35.8 reverse	15,189,363
Dunnock, lane 2	36,391,353	33.5 forward, 33.6 reverse	25,245,144 forward, 26,523,295 reverse	35.8 forward, 36.1 reverse	19,757,879

Table 3. Gene detection for different short read treatments. All gene detection was performed using single-match transcriptome sequences.

Read treatment	Water pipit: genes detected	Dunnock: genes detected
Single-end aligned reads	15,837	15,740
Paired-end aligned reads	15,272	14,867
Genome-guided assembly: single-end alignments	8,188	8,112
Genome-guided assembly: paired-end alignments	7,496	8,627

Table 4. Variant detection. Features that mapped to genes are given in parenthesis.

Feature	Water pipit	Dunnock
SNPs not shared with ZF	13,873 (3,841)	15,245 (4,444)
SNPs shared with ZF	822 (191)	580 (150)
Indels	118,417 (27,999)	94,313 (18,415)

Table 5. Functional enrichment and depletion of genes indicated to exhibit adaptive evolution: water pipit versus zebra finch.

GOslim term	Detected (%)	Total ZF genes in term
Over-represented		
aging	75.00	4
cytosol	24.14	58
mitochondrion	23.49	166
transferase activity, transferring alkyl or aryl (other than methyl) groups	23.08	26
cell death	21.05	38
ribosome	17.91	134
homeostatic process	16.85	89
structural constituent of ribosome	15.63	128
generation of precursor metabolites and energy	13.89	72
translation	12.44	209
cytoplasm	10.69	1132
RNA binding	9.63	270
oxidoreductase activity	8.21	560
organelle	7.09	2129
Under-represented		
biological process	4.68	7884
molecular function	4.57	11734
cellular component	4.54	6716

ion binding	3.98	2411
protein modification process	3.68	1034
kinase activity	3.07	750
transmembrane transport	3.03	462
neurological system process	2.52	278
signal transduction	2.35	1616
nucleic acid binding transcription factor activity	2.23	449
signal transducer activity	1.91	888
external encapsulating structure	0.00	16
photosynthesis	0.00	6
ribonucleoprotein complex assembly	0.00	6
cell wall organization or biogenesis	0.00	6
mRNA binding	0.00	5
vacuolar transport	0.00	5
cilium	0.00	3
extracellular matrix organization	0.00	3
small conjugating protein binding	0.00	3
endosome	0.00	2
secondary metabolic process	0.00	2
developmental maturation	0.00	2
cell junction organization	0.00	1

Table 6. Functional enrichment and depletion of genes indicated to exhibit adaptive evolution: dunnock versus zebra finch.

GOslim term	Detected (%)	Total ZF genes in term
Over-represented		
anatomical structure formation involved in morphogenesis	50.00	6
cytosol	29.31	58
nucleolus	25.00	24
mitochondrion	22.29	166
reproduction	21.43	28
ribosome	18.66	134
structural constituent of ribosome	16.41	128
homeostatic process	14.61	89
anatomical structure development	13.89	72
translation	13.88	209
cytoplasm	10.60	1132
RNA binding	10.00	270
organelle	7.00	2129
Under-represented		
biological process	4.46	7884
molecular function	4.45	11734
cellular component	4.36	6716
ion binding	3.86	2411
protein modification process	3.19	1034

peptidase activity	2.95	441
plasma membrane	2.73	440
signal transduction	2.48	1616
nucleic acid binding transcription factor activity	2.45	449
kinase activity	2.13	750
signal transducer activity	1.91	888
transferase activity, transferring glycosyl groups	1.55	194
external encapsulating structure	0.00	16
photosynthesis	0.00	6
cell wall organization or biogenesis	0.00	6
mRNA binding	0.00	5
vacuolar transport	0.00	5
cilium	0.00	3
cell proliferation	0.00	3
extracellular matrix organization	0.00	3
small conjugating protein binding	0.00	3
endosome	0.00	2
secondary metabolic process	0.00	2
developmental maturation	0.00	2

Table 7. Functional enrichment and depletion of genes indicated to exhibit adaptive evolution: water pipit versus dunnock.

GOslim term	Detected (%)	Total ZF genes in term
Over-represented		
aging	75.00	4
nucleolus	25.00	24
sulfur compound metabolic process	20.83	24
cytosol	20.69	58
mitochondrion	19.28	166
homeostatic process	14.61	89
ribosome	10.45	134
translation	9.09	209
cytoplasm	8.92	1132
oxidoreductase activity	7.68	560
Under-represented		
cellular component	3.89	6716
molecular function	3.89	11734
ion binding	3.53	2411
protein modification process	3.09	1034
kinase activity	2.27	750
signal transduction	2.23	1616
transferase activity, transferring glycosyl groups	1.55	194
signal transducer activity	1.24	888

nucleic acid binding transcription factor activity	1.11	449
growth	0.00	26
external encapsulating structure	0.00	16
transcription factor binding	0.00	12
lysosome	0.00	11
photosynthesis	0.00	6
rRNA binding	0.00	6
ribonucleoprotein complex assembly	0.00	6
cell wall organization or biogenesis	0.00	6
vacuolar transport	0.00	5
histone binding	0.00	5
symbiosis, encompassing mutualism through parasitism	0.00	5
cilium	0.00	3
extracellular matrix organization	0.00	3
small conjugating protein binding	0.00	3
endosome	0.00	2
secondary metabolic process	0.00	2
developmental maturation	0.00	2

Table 8: Differential gene expression results (using DESeq).

Gene	Name/ Description	Log2 fold change	Adjusted p value	Associated GO slim terms
Dunnock (polygamous) > water pipit (monogamous)				
ENSTGUG00000009674	CYP2D6	5.73	4.92E-15	Ion binding, oxidoreductase activity, molecular function, biological process
ENSTGUG00000009671		5.63	2.08E-14	Molecular function
ENSTGUG00000006821		3.84	7.08E-13	Ribosome, structural molecule activity, structural constituent of ribosome, translation, organelle, biosynthetic process, molecular function, cytoplasm, cellular component, intracellular, cell, biological process
ENSTGUG00000006711	MLF1IP	6.87	3.36E-11	None
ENSTGUG00000009751	TCTE3	4.93	7.29E-09	None
ENSTGUG00000017463		Inf	1.17E-06	None
ENSTGUG00000000297	IL18	6.04	3.89E-06	Extracellular space, extracellular region, molecular function, cellular component
ENSTGUG00000002115	SCLT1	2.83	2.13E-05	None

ENSTGUG00000009375		4.60	3.02E-05	Signal transduction, molecular function, cellular component, intracellular, cell, biological process
ENSTGUG00000017554	LOC100190559	2.30	8.60E-05	Ribosome, structural molecule activity, structural constituent of ribosome, translation, organelle, biosynthetic process, molecular function, cytoplasm, cellular component, intracellular, cell, biological process
ENSTGUG00000014971		4.51	0.00049	None
ENSTGUG00000013602		Inf	0.00172	None
ENSTGUG00000018767	ND6	1.84	0.00436	Mitochondrion, organelle, oxidoreductase activity, molecular function, cytoplasm, cellular component, intracellular, cell, biological process
ENSTGUG00000004972		5.85	0.00503	None
ENSTGUG00000017385		4.75	0.00597	Peptidase activity, proteinaceous extracellular matrix, extracellular region, ion binding, molecular function, cellular component
ENSTGUG00000009556		4.35	0.00597	None
ENSTGUG00000010266	CTH	2.66	0.00597	Cellular amino acid metabolic process, biosynthetic process, small molecule metabolic process, molecular function, cellular nitrogen compound metabolic process, biological process
ENSTGUG00000012881		2.24	0.00623	Transferase activity transferring alkyl or aryl (other than methyl) groups, molecular function,

				biological process
ENSTGUG00000000766	NMRK1	2.93	0.00674	None
ENSTGUG00000011832	NECAB1	Inf	0.00890	Ion binding, molecular function
ENSTGUG00000008993	NCOA3	2.26	0.00959	Transferase activity, transferring acyl groups, protein binding transcription factor activity, chromosome organization, protein modification process, signal transduction, signal transducer activity, organelle, biosynthetic process, nucleus, molecular function, cellular component, intracellular, cell, cellular nitrogen compound metabolic process, biological process
ENSTGUG00000016716		2.11	0.01045	Ion binding, molecular function, biological process
ENSTGUG00000005985	GFM2	2.13	0.01241	Nucleobase-containing compound catabolic process, GTPase activity, catabolic process, small molecule metabolic process, molecular function, cellular nitrogen compound metabolic process, biological process
ENSTGUG00000001486	TMOD1	2.35	0.01243	Cytoskeletal protein binding, cytoskeleton, organelle, molecular function, cellular component, intracellular, cell
ENSTGUG00000013143	MATN3	3.10	0.01313	Ion binding, molecular function
ENSTGUG00000001016	ATCAY	1.84	0.01609	None

ENSTGUG00000007199		1.93	0.01661	Ribosome, structural molecule activity, structural constituent of ribosome, translation, organelle, biosynthetic process, molecular function, cytoplasm, cellular component, intracellular, cell, biological process
ENSTGUG00000013381		3.69	0.01661	None
ENSTGUG00000010887		1.97	0.01957	Ligase activity, cell-cell signalling, neurological system process, cytoplasmic membrane-bounded vesicle, organelle, transport, molecular function, cytoplasm, cellular component, intracellular, cell, biological process
ENSTGUG00000008939		1.76	0.01957	Molecular function
ENSTGUG00000010624	RGN	2.61	0.02023	Enzyme regulator activity, ion binding, molecular function, cytoplasm, cellular component, intracellular, cell, biological process
ENSTGUG00000018494		2.32	0.02120	None
ENSTGUG00000009207	CEP89	2.27	0.02125	None
ENSTGUG00000011814	RFC3	1.89	0.02530	DNA binding, nucleotidyltransferase activity, biosynthetic process, molecular function, protein complex, cellular component, intracellular, cell, cellular nitrogen compound metabolic process, biological process, DNA metabolic process

ENSTGUG00000007486	DNAH5	2.12	0.02901	Nucleobase-containing compound catabolic process, ATPase activity, cytoskeleton, organelle, catabolic process, small molecule metabolic process, molecular function, protein complex, cellular component, intracellular, cell, cellular nitrogen compound metabolic process, biological process
ENSTGUG00000002713		1.81	0.03163	Ion binding, molecular function, biological process
ENSTGUG00000017338		2.21	0.04240	None
ENSTGUG00000001572	ALB	6.78	0.04450	Extracellular space, extracellular region, transport, cellular component, biological process
ENSTGUG00000018513		1.97	0.04838	None
Water pipit (monogamous) > dunnock (polygamous)				
ENSTGUG00000015535		4.39	7.40E-11	Organelle, biosynthetic process, nucleus, cell cycle, cellular component, intracellular, cell, response to stress, cellular nitrogen compound metabolic process, biological process, DNA metabolic process
ENSTGUG00000002791		4.28	1.04E-10	Signal transduction, signal transducer activity, molecular function, cellular component, biological process
ENSTGUG00000005705	PTCD2	5.72	6.32E-09	None
ENSTGUG00000015745		4.17	1.15E-07	Signal transduction, cellular component, biological process

ENSTGUG00000015647		5.38	6.70E-07	Ion binding, molecular function, cellular component, intracellular, cell
ENSTGUG00000014425		Inf	3.00E-06	None
ENSTGUG00000015724		Inf	1.06E-05	None
ENSTGUG00000010997	LRRC34	3.64	8.46E-05	Molecular function
ENSTGUG00000014429		4.63	8.73E-05	None
ENSTGUG00000002657		3.38	0.00012	Nucleobase-containing compound catabolic process, transmembrane transporter activity, ATPase activity, transmembrane transport, catabolic process, transport, small molecule metabolic process, molecular function, cellular component, cellular nitrogen compound metabolic process, biological process
ENSTGUG00000018337		3.71	0.00014	Signal transduction, signal transducer activity, molecular function, cellular component, biological process
ENSTGUG00000007597		2.84	0.00067	None
ENSTGUG00000009679	LOC100223017	2.43	0.00089	Ribosome, structural molecule activity, structural constituent of ribosome, translation, organelle, biosynthetic process, ribosome biogenesis, molecular function, cytoplasm, cytosol, cellular component, intracellular, cell, cellular nitrogen compound metabolic process, biological process

ENSTGUG00000011753		6.19	0.00125	
ENSTGUG00000017188		3.12	0.00125	Hydrolase activity, acting on carbon-nitrogen (but not peptide) bonds, molecular function, cytoplasm, cellular component, intracellular, cell
ENSTGUG00000002332		3.05	0.00128	Molecular function
ENSTGUG00000008318	LIPA	2.32	0.00302	Lipid metabolic process, molecular function, biological process
ENSTGUG00000009573	PACRGL	2.50	0.01300	Molecular function
ENSTGUG00000011834	COCH	2.32	0.01661	Molecular function
ENSTGUG00000010084	FAM211B	2.44	0.02120	
ENSTGUG00000012005	HRSP12	1.99	0.02755	
ENSTGUG00000002687	ERCC8	2.06	0.03188	Molecular function
ENSTGUG00000003586	PBLD	3.69	0.03821	Biosynthetic process, molecular function, biological process

4.8 Supplementary information

4.8.1 Supplementary figure legends

Fig. S1. Electropheragrams of dunnock (*Prunella modularis*) RNA extracted from brain tissue samples from each bird. It can be seen that birds 1, 6 and 8 have the best quality RNA due to the presence of more distinct 18S and 28S peaks. Hence, RNA from these birds has been pooled for high throughput sequencing.

Fig. S2. Electropheragrams of water pipit (*Anthus spinoletta*) RNA extracted from brain tissue samples from each bird. It can be seen that birds 4, 5 and 10 have the best quality RNA due to the presence of more distinct 18S and 28S peaks. Hence, RNA from these birds has been pooled for high throughput sequencing.

Fig. S3. Quality score boxplots of raw read samples for (A) AsL1R1, (B) AsL1R2, (C) AsL2R1, (D) AsL2R2.

Fig. S4. Quality score boxplots of processed reads for (A) AsL1R1, (B) AsL1R2, (C) AsL2R1, (D) AsL2R2.

Fig. S5. Quality score boxplots of raw read samples for (A) DpL4R1, (B) DpL4R2, (C) DpL5R1, (D) DpL5R2.

Fig. S6. Quality score boxplots of processed reads for (A) DpL4R1, (B) DpL4R2, (C) DpL5R1, (D) DpL5R2.

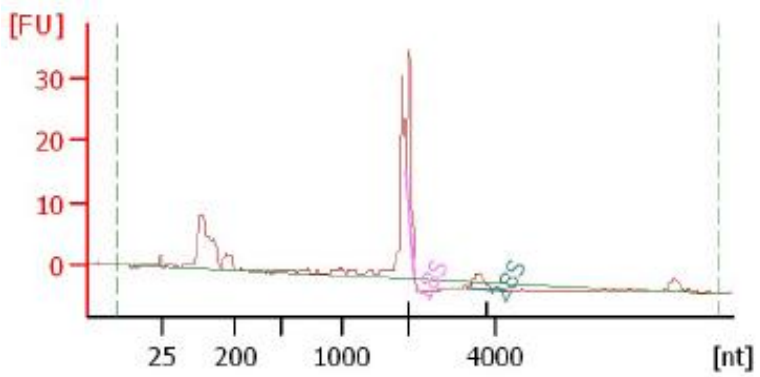
Fig. S7. Indels that map to genes identified using the zebra finch genome against chromosome size identified in the (A) water pipit, and (B) dunnock. Indel frequency scales well with chromosome size.

Fig. S8. SNPs that map to genes identified using the zebra finch genome against chromosome size identified in the (A) water pipit, and (B) dunnock. Again, SNP frequency scales well with chromosome size.

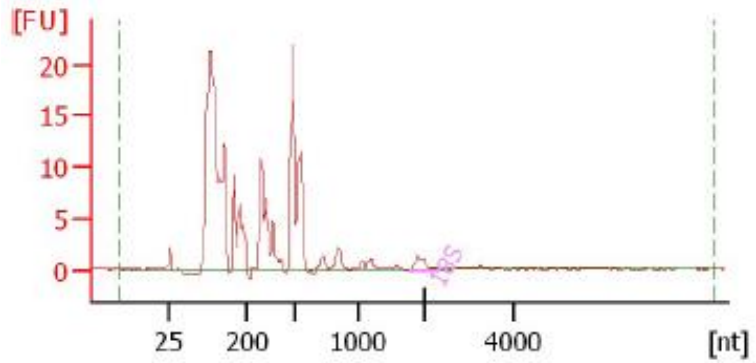
Fig. S9. SNPs shared with the zebra finch that map to genes, identified using the zebra finch genome against chromosome size identified in the (A) water pipit, and (B) dunnock. Shared SNP frequency does not scale as well with chromosome size as indels or unshared SNPs.

Fig. S1.

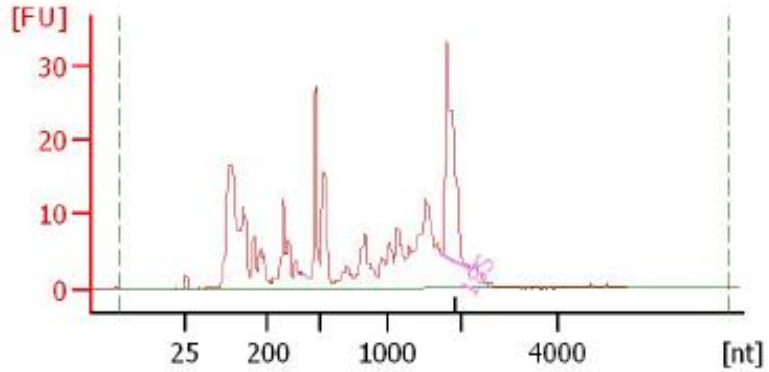
Bird ID 1



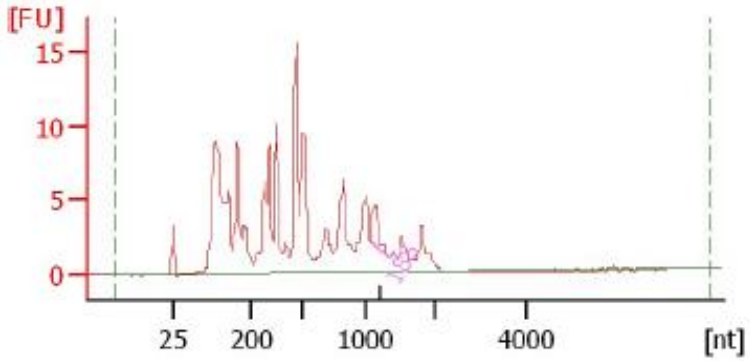
Bird ID 2



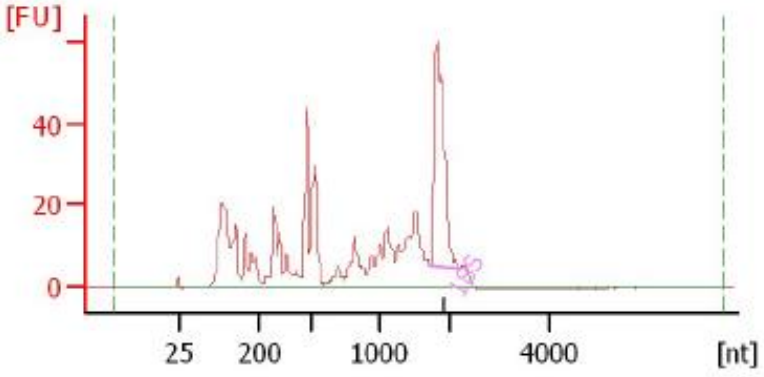
Bird ID 6



Bird ID 7



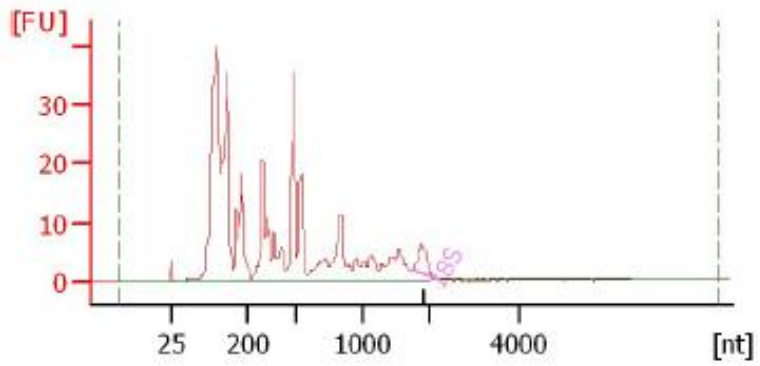
Bird ID 8



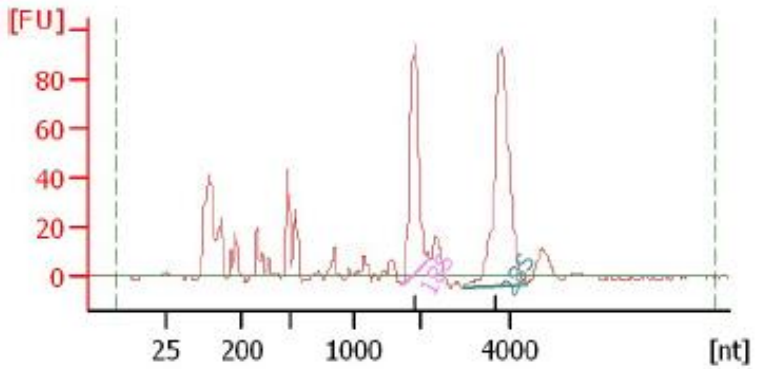
172

Fig. S2.

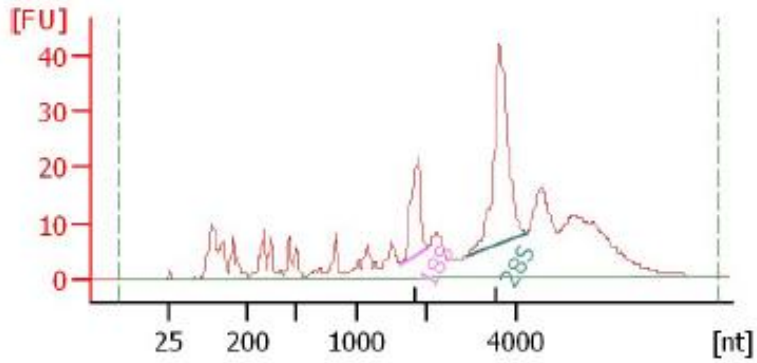
Bird ID 3



Bird ID 4



Bird ID 5



Bird ID 10

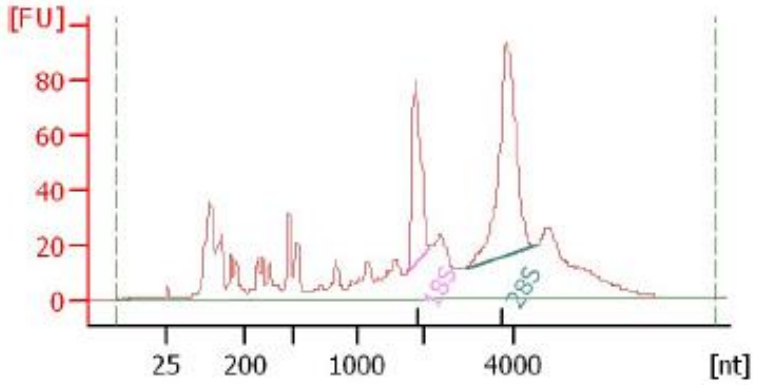
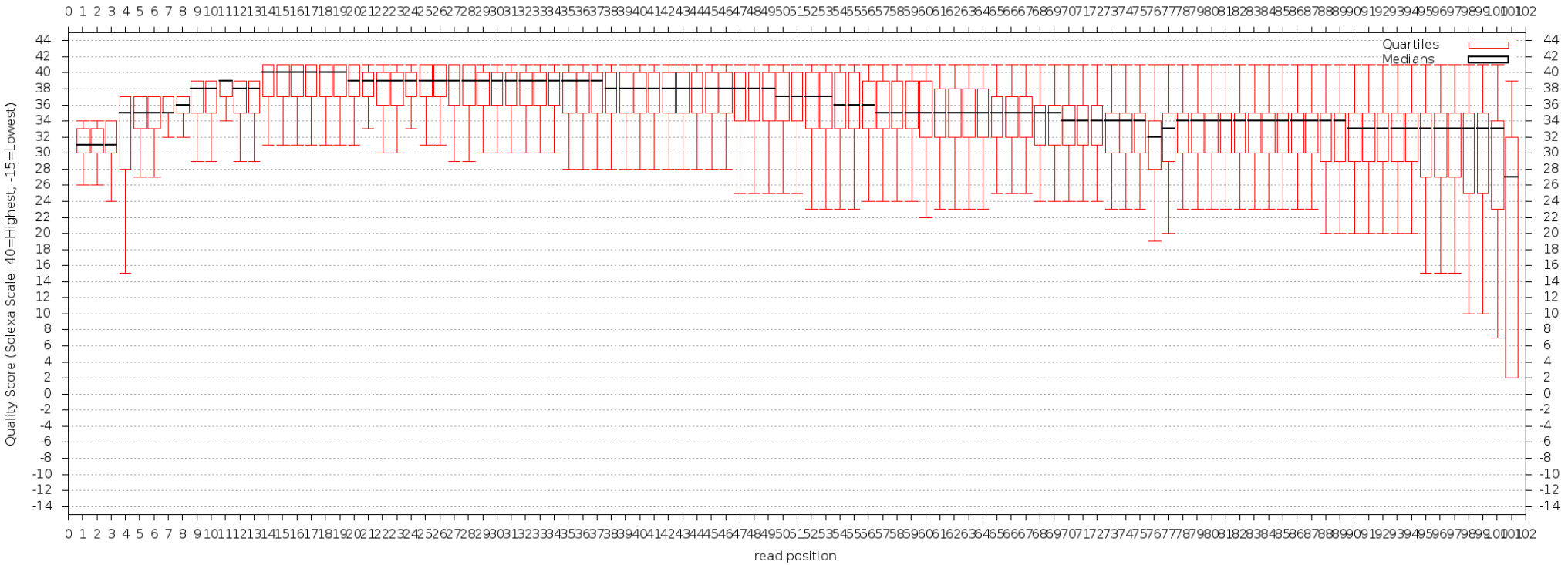
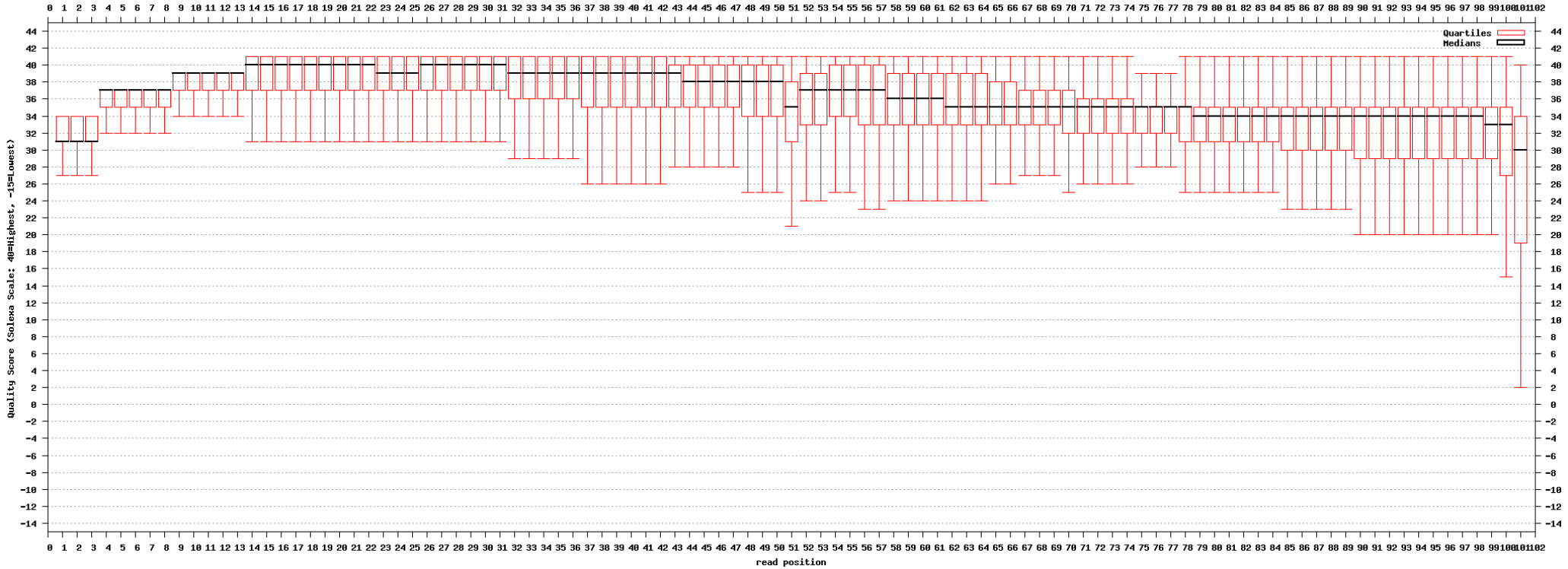


Fig. S3.

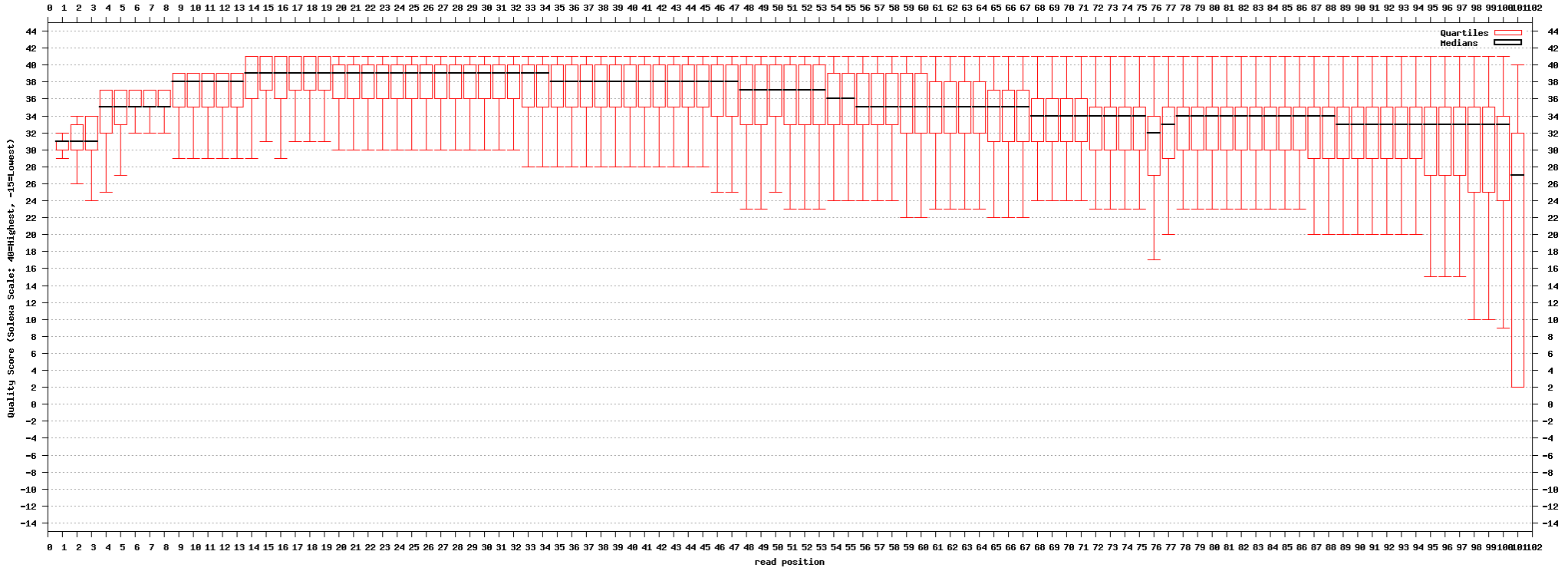
(A)


Quartiles
Medians
Quality Score (Solexa Scale: 40=Highest, -15=Lowest)
read position

(B)


Quartiles
Medians
Quality Score (Solexa Scale: 40=Highest, -15=Lowest)
read position

(C)


Quartiles
Medians
Quality Score (Solexa Scale; 40=Highest, -15=Lowest)
read position

(D)

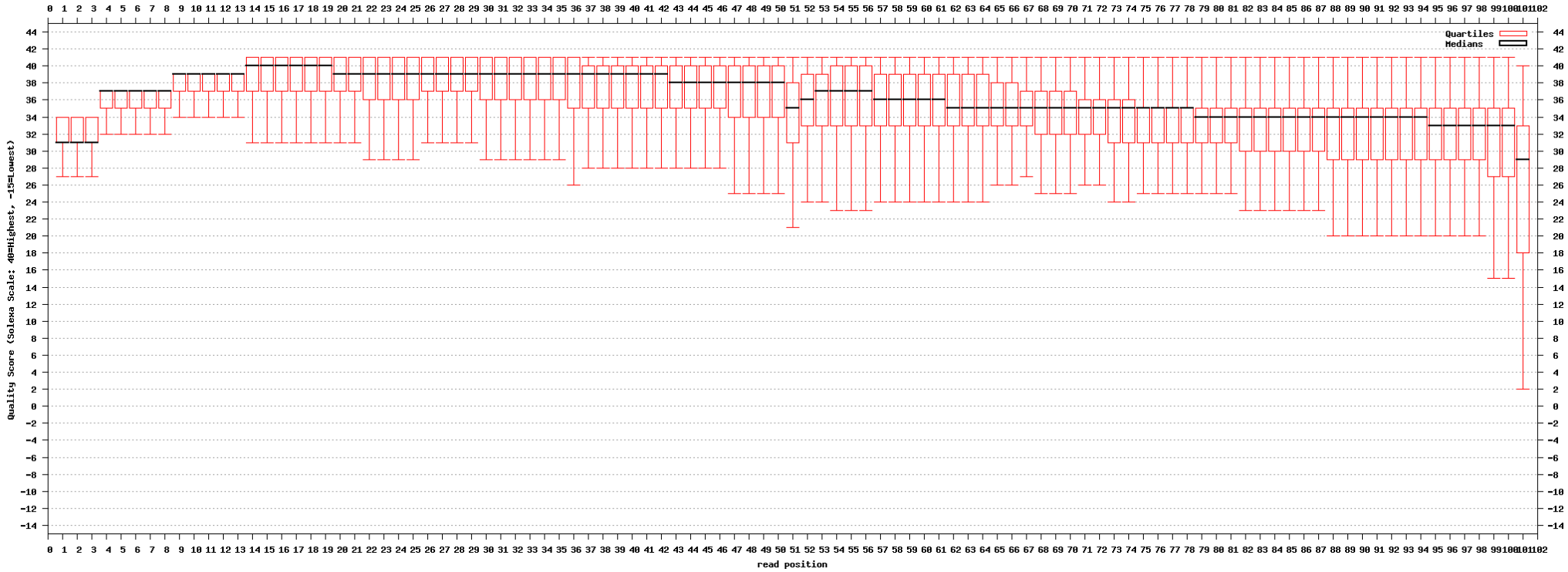
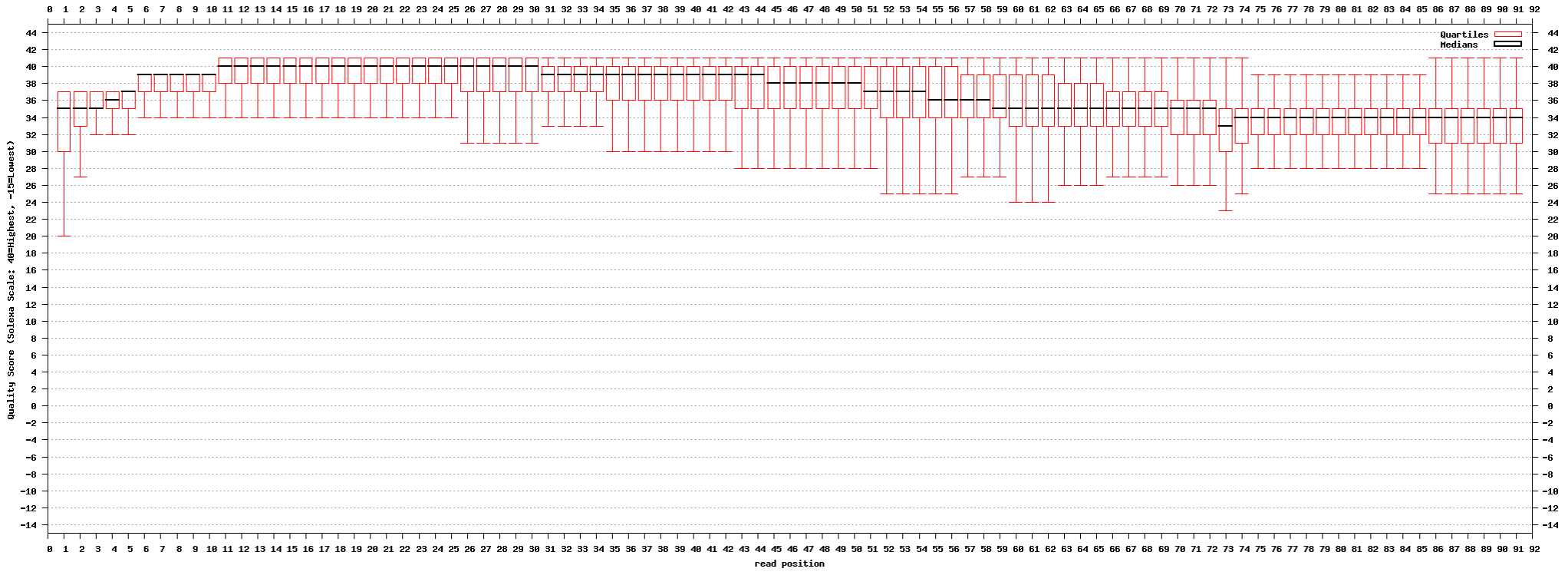
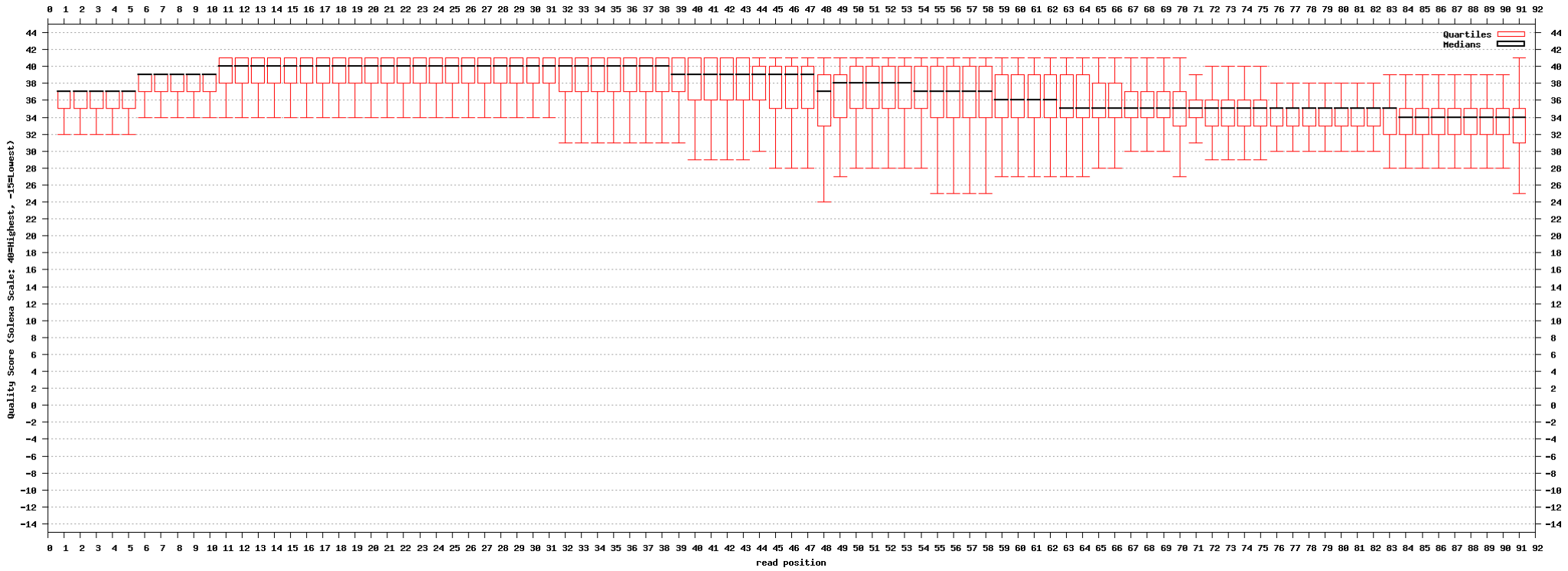

Quartiles
Medians
Quality Score (Solexa Scale; 40=Highest, -15=Lowest)
read position

Fig. S4.

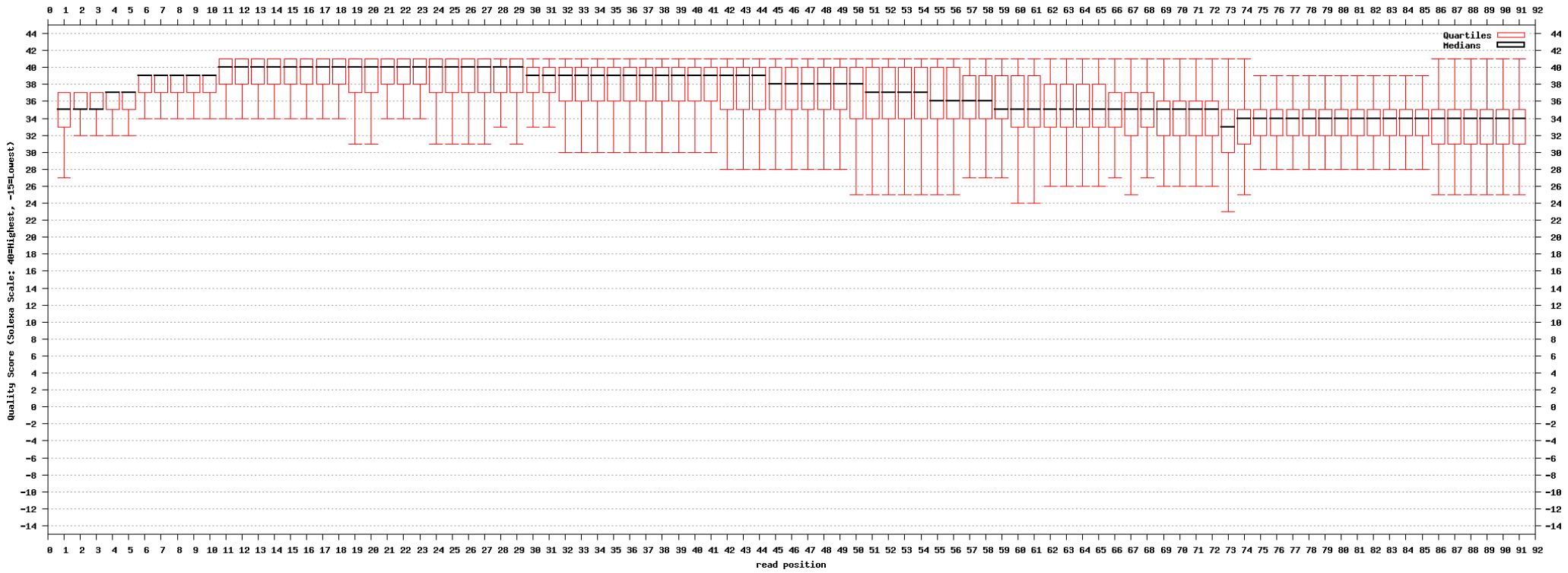
(A)


Quartiles
Medians
Quality Score (Solexa Scale; 40=Highest, -15=Lowest)
read position

(B)


Quartiles
Medians
Quality Score (Solexa Scale: 40=Highest, -15=Lowest)
read position

(C)


Quartiles
Medians
Quality Score (Solexa Scale; 40=Highest, -15=Lowest)
read position

(D)

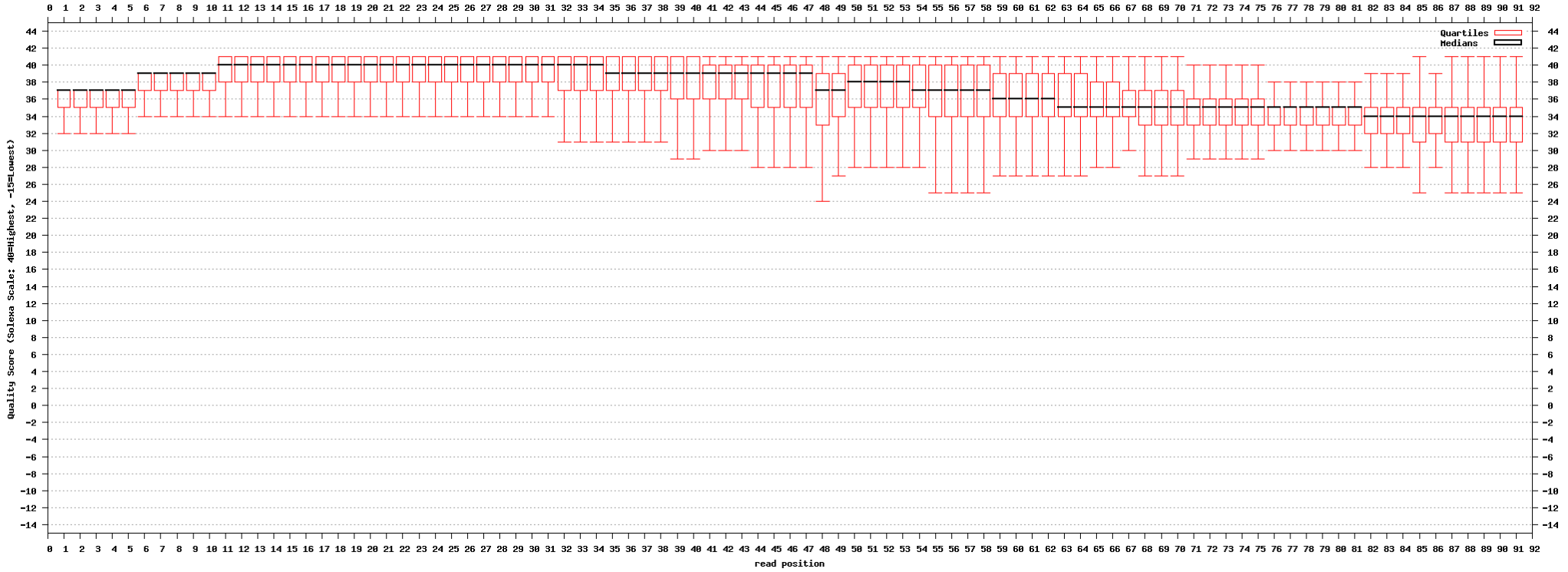
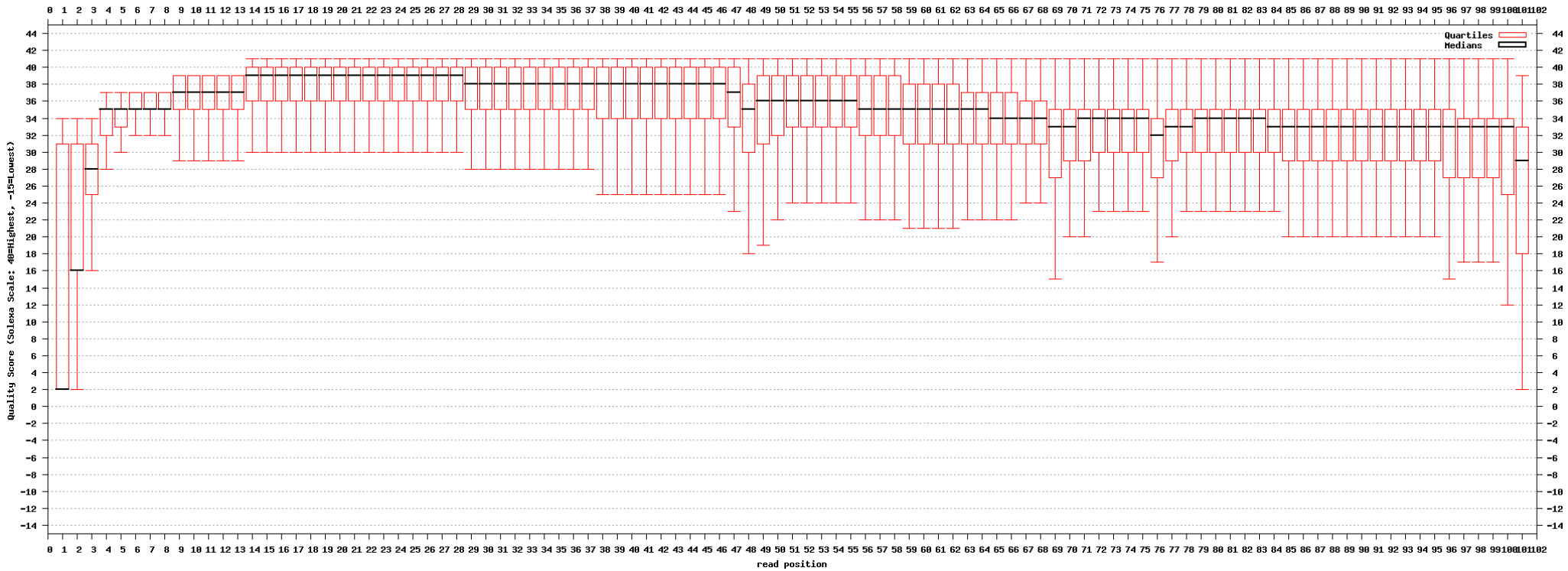
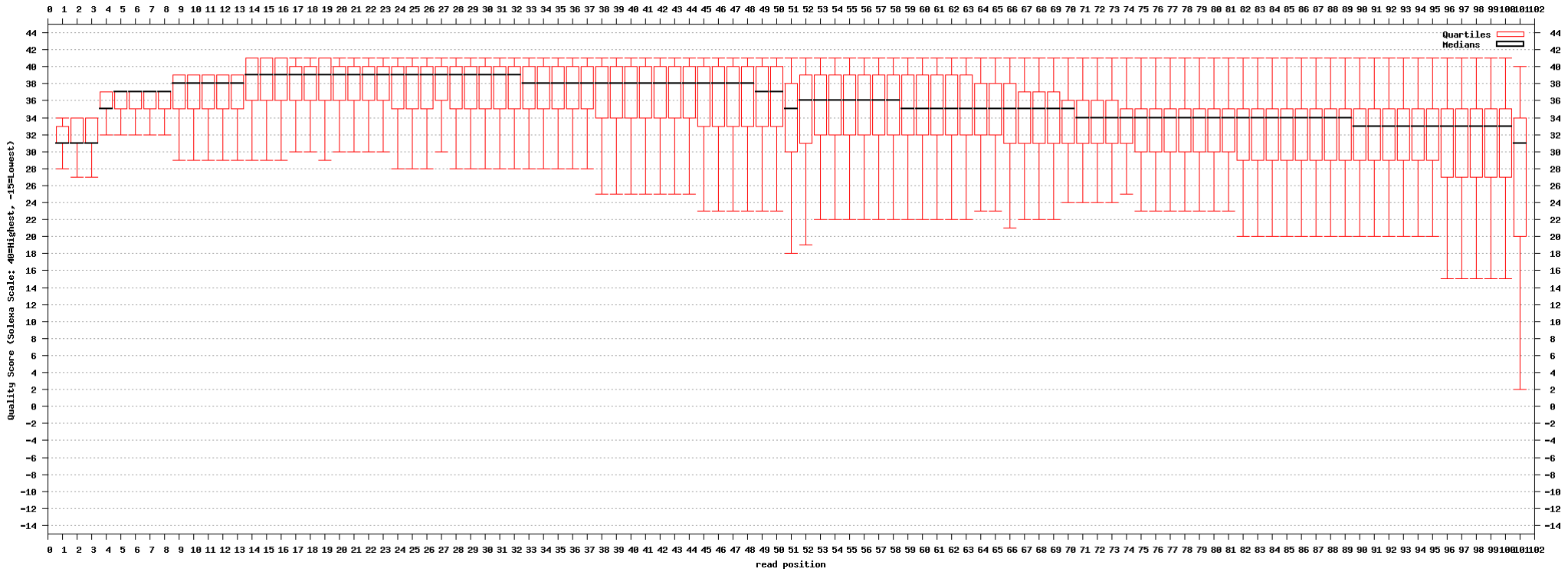

Quartiles
Medians
Quality Score (Solexa Scale; 40=Highest, -15=Lowest)
read position

Fig. S5.

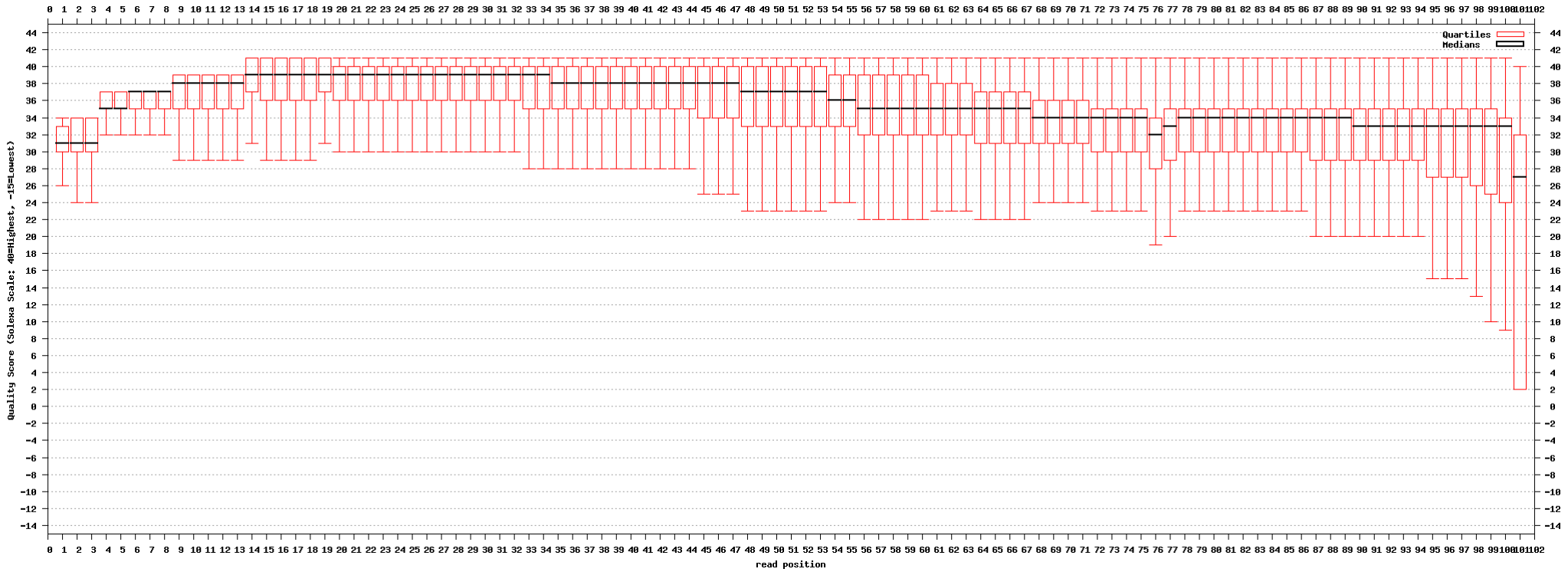
(A)


Quartiles
Medians
Quality Score (Solexa Scale: 40=Highest, -15=Lowest)
read position

(B)


Quartiles
Medians
Quality Score (Solexa Scale: 40=Highest, -15=Lowest)
read position

(C)


Quartiles
Medians
Quality Score (Solexa Scale: 40=Highest, -15=Lowest)
read position

(D)

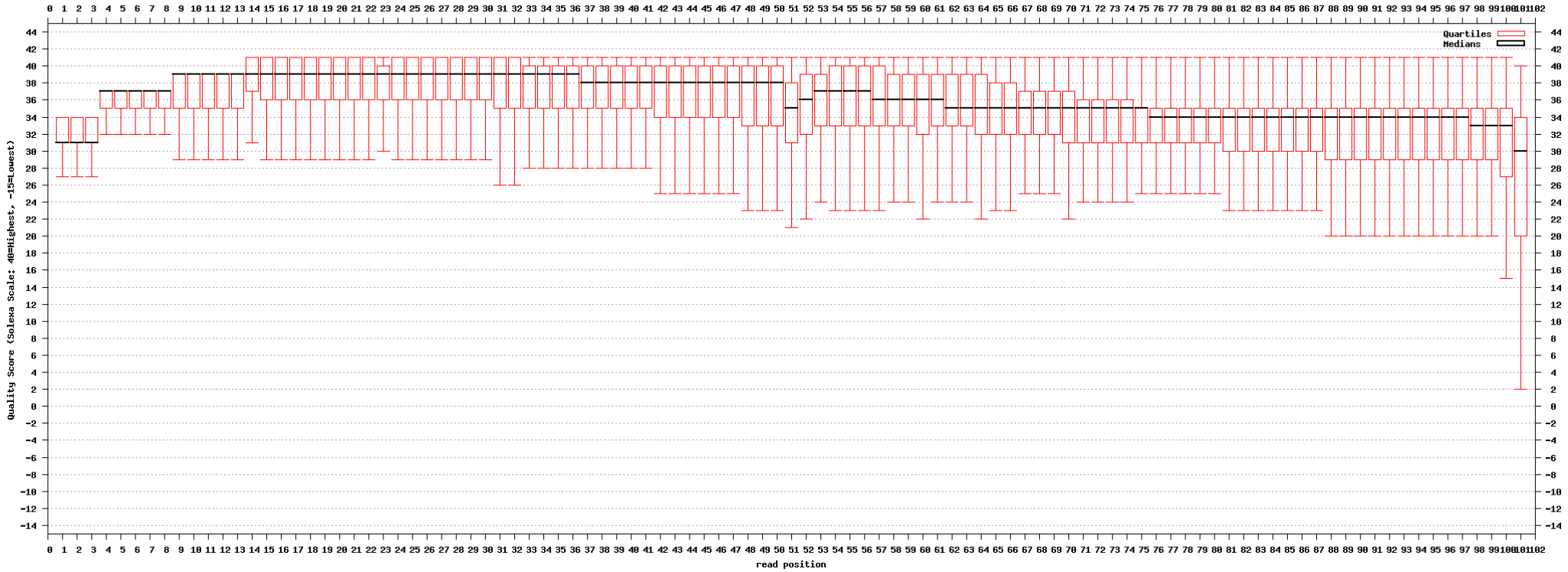
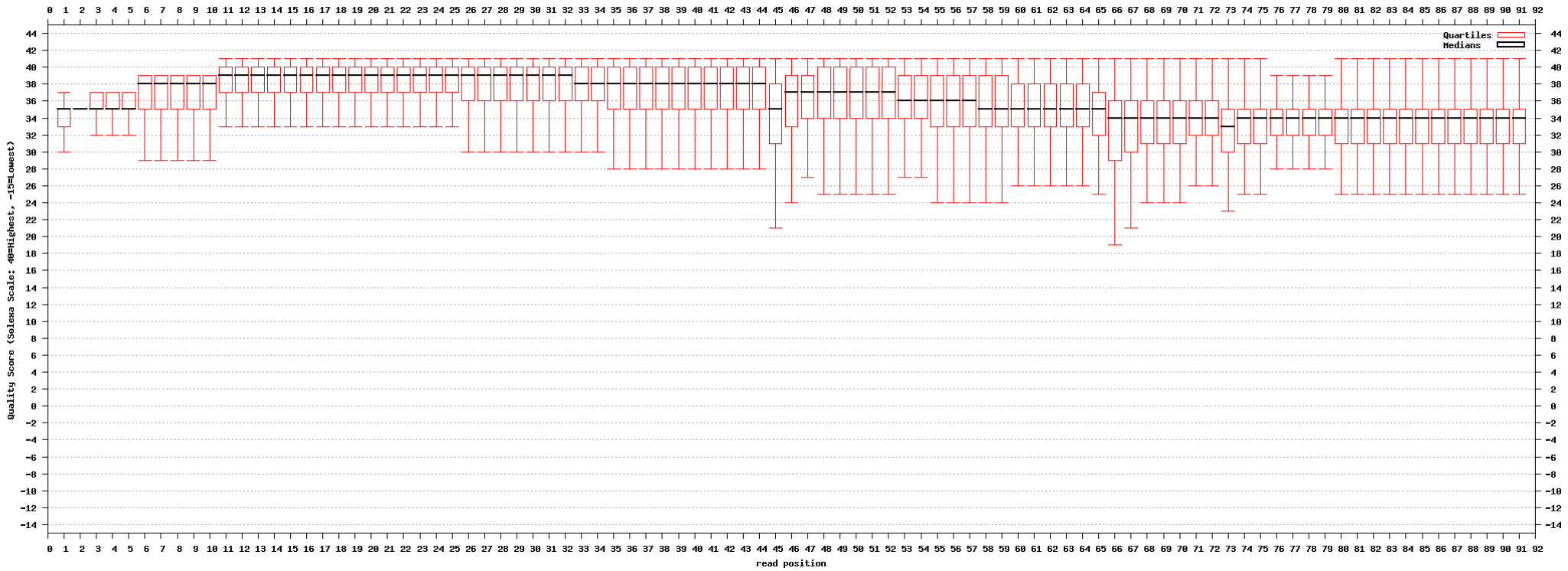
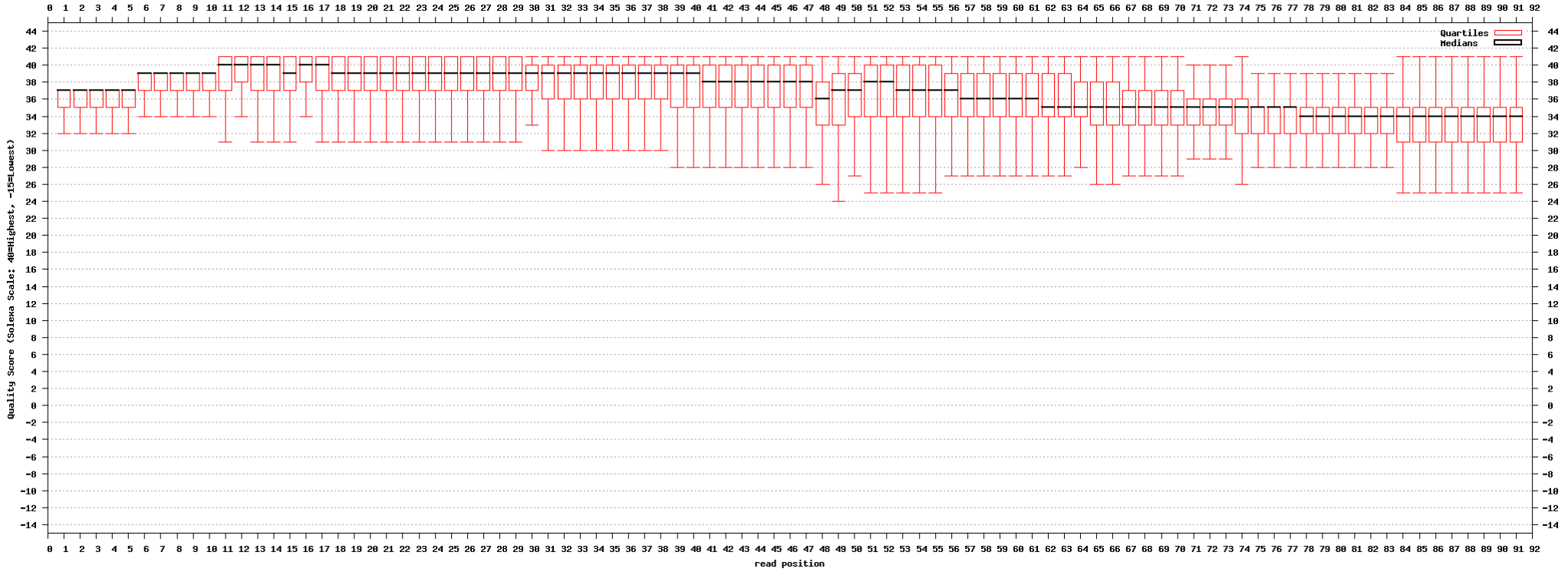

Quartiles
Medians
Quality Score (Solexa Scale: 40=Highest, -15=Lowest)
read position

Fig. S6.

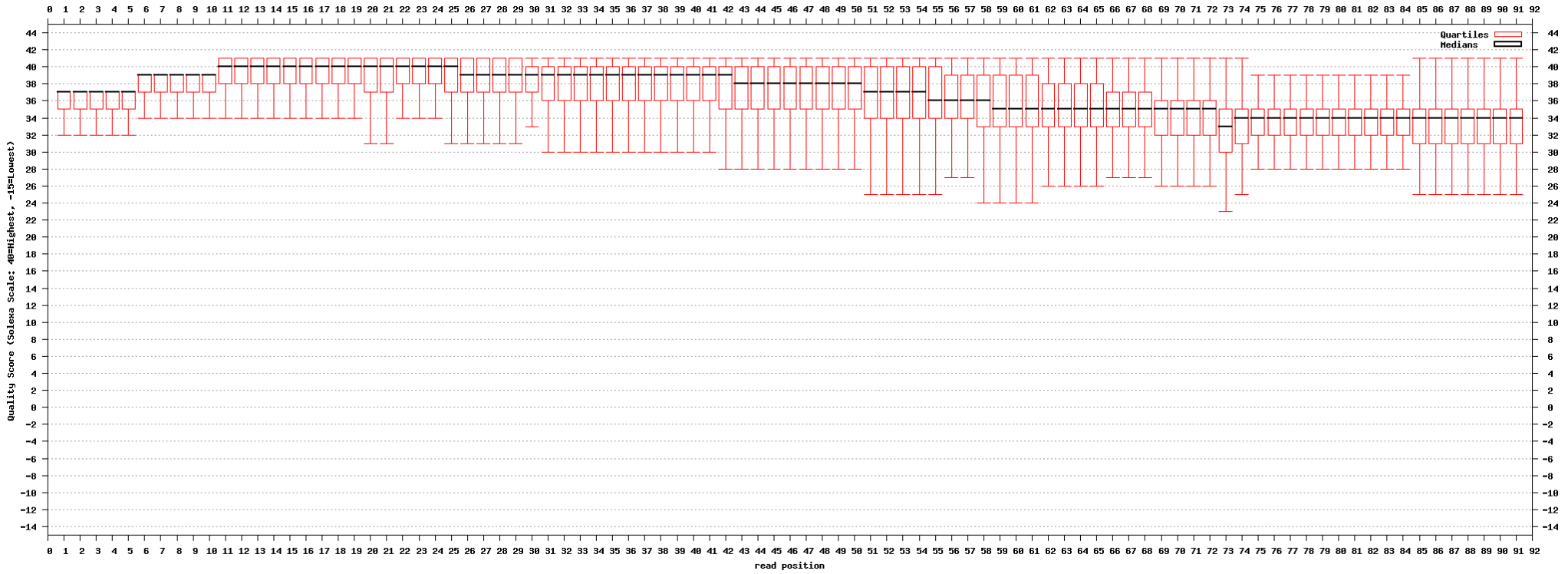
(A)


Quartiles
Medians
Quality Score (Solexa Scale: 40=Highest, -15=Lowest)
read position

(B)


Quartiles
Medians
Quality Score (Solexa Scale; 40=Highest, -15=Lowest)
read position

(C)


Quartiles
Medians
Quality Score (Solexa Scale: 40=Highest, -15=Lowest)
read position

(D)

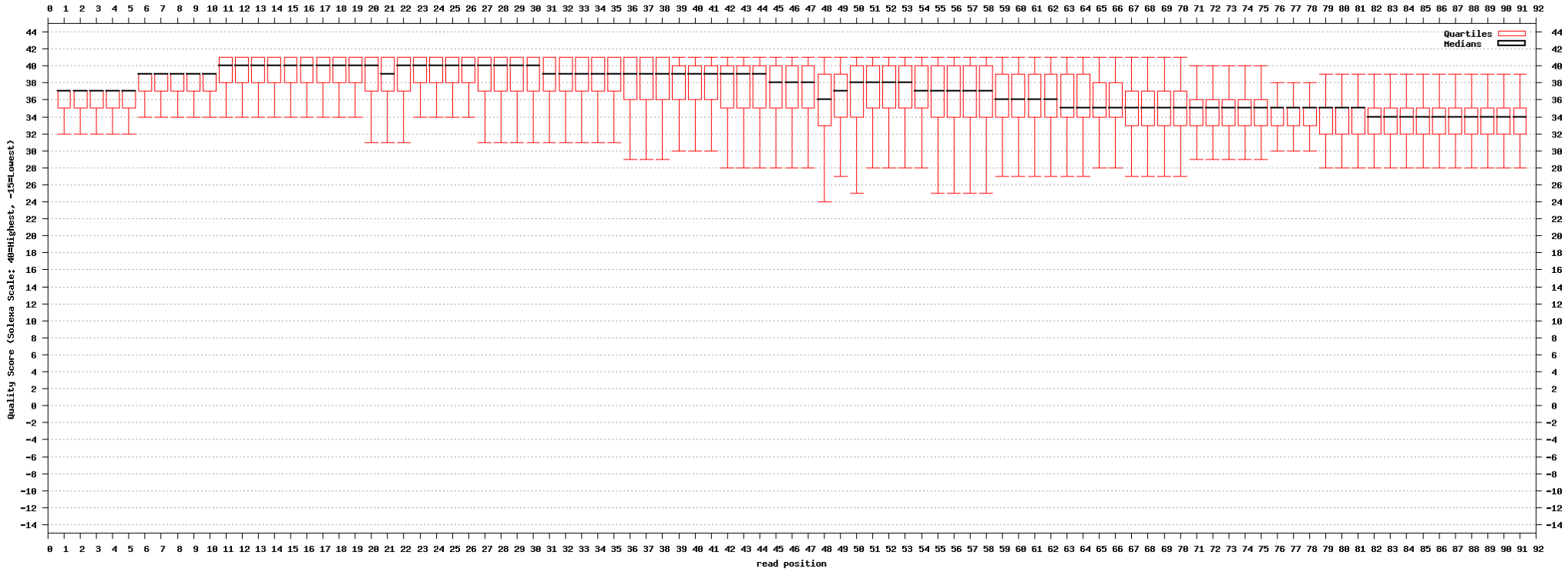

Quartiles
Medians
Quality Score (Solexa Scale: 40=Highest, -15=Lowest)
read position

Fig. S7.

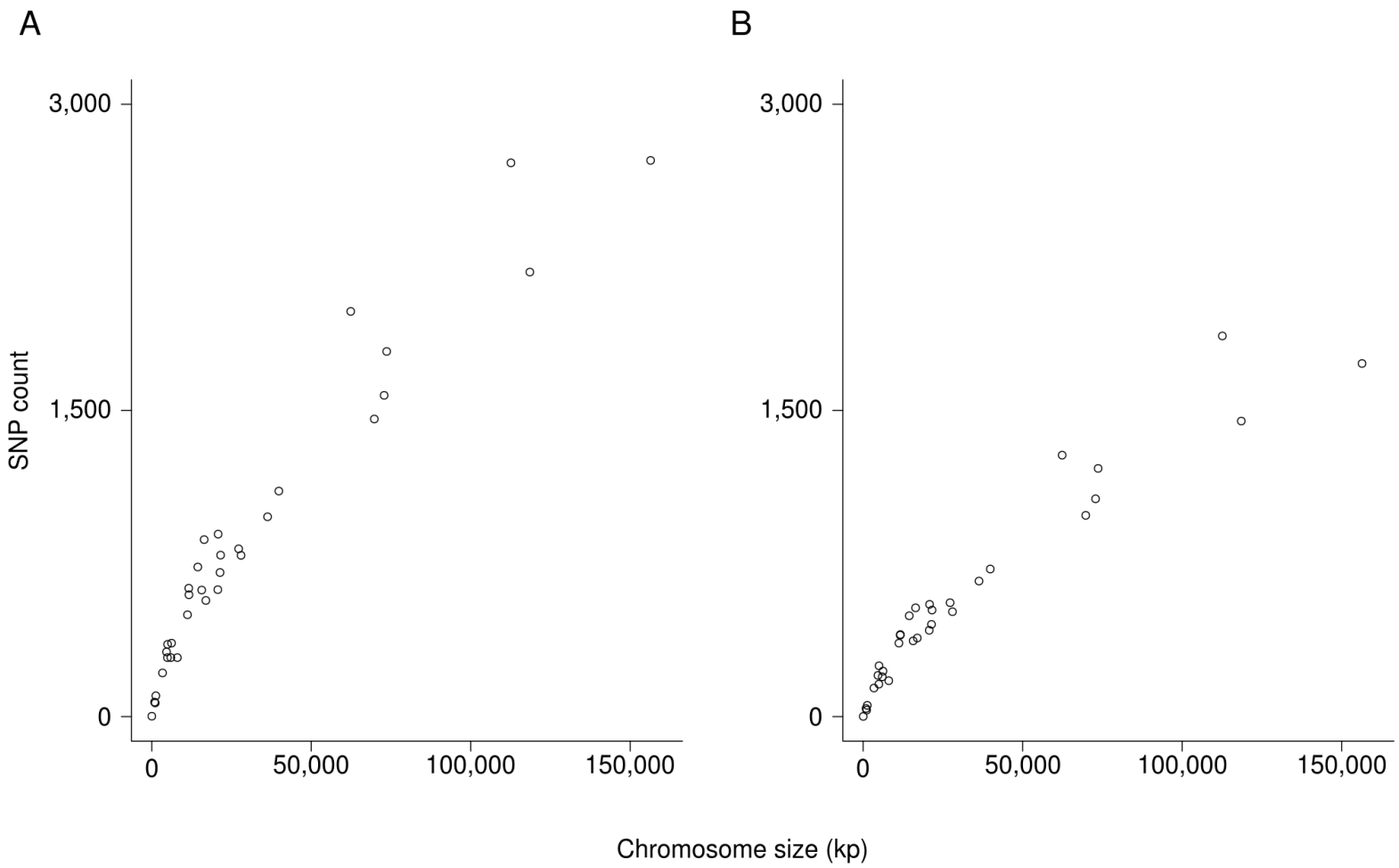

A
B
3,000
1,500
0
SNP count
0
50,000
100,000
150,000
Chromosome size (kp)

Fig. S8.

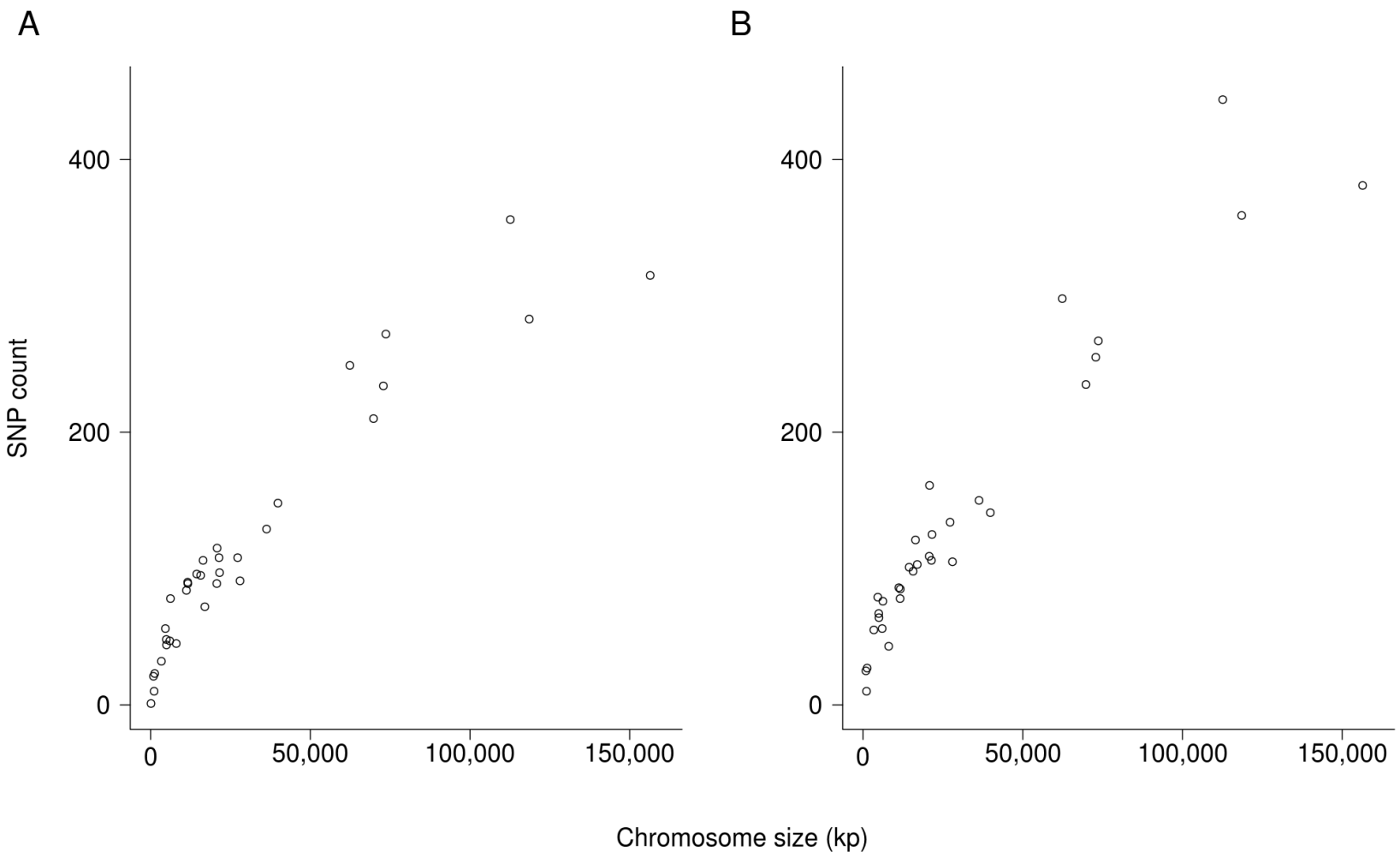
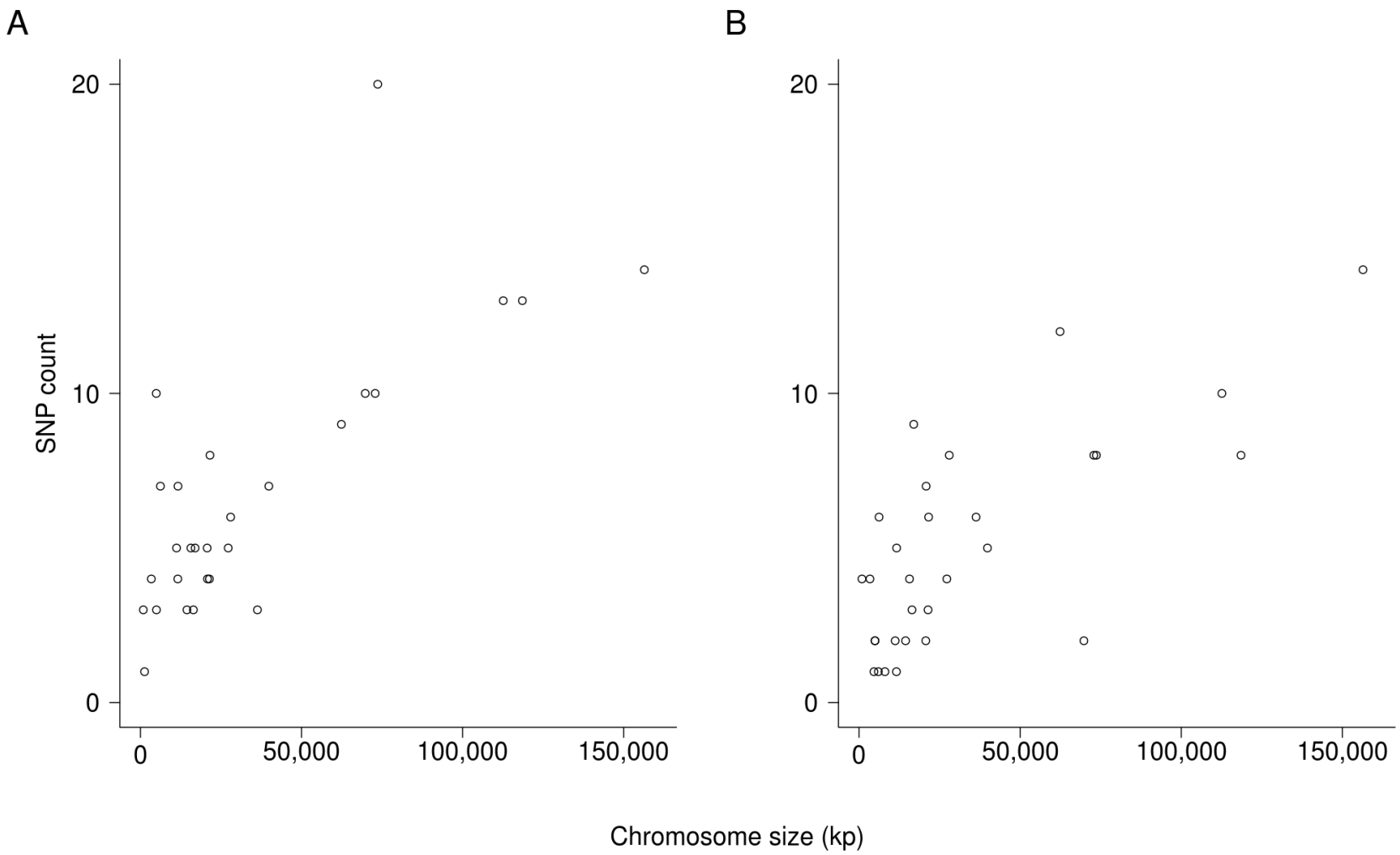

A
B
SNP count
400
200
0
0
50,000
100,000
150,000
Chromosome size (kp)

Fig. S9.



5 Overall conclusions and future directions

The molecular basis of ecological variation is one of the fundamental research areas in the post genomic era. While the molecular basis of phenotypes controlled by single genes is now relatively easy to elucidate, the identity of and mechanisms by which genes control more complex phenotypes remains poorly understood. Integrating 'omics data streams has the power to generate more comprehensive understanding of the layers of complexity that operate and interact in the evolution and development of complex traits. This relies upon the utilisation of accurate and unbiased data capture and analysis tools (Berger et al., 2013).

The overarching aim of my PhD thesis was to apply state-of-the-art systems biology techniques to questions concerning the genomic evolution of variation in complex phenotypes, focusing on comparative transcriptome profiling of wild species that exhibit different mating systems as different ecological strategies. I exploit the results of natural experiments that have produced different mating systems, pair-bonding and parental care between different populations (Balshine, 2012; McGraw et al., 2010). Whilst the behavioural ecology of mating systems and parental care is well-established although a rapidly developing field itself, my work has focused on a new aspect of mating system evolution: brain gene expression.

Using RNA-seq and phenotypic data, I sought to identify transcriptomic signatures of different phenotypes and apply rigorous statistical methods to reduce false positives and ensure confidence in my findings. My thesis presents several key outputs: (a) I have demonstrated how to apply and integrate 'omics technologies, shedding insight into genomic variation underlying complex traits; (b) I have identified a preferred annotation technique to apply to transcriptome data derived from species lacking sequenced genomes, and; (c) I have identified candidate genes expressed in the brain that may underlie differences in mating behaviour between monogamous and polygamous songbird species during their breeding seasons. In achieving these outputs, I have analysed a combination of previously obtained data from different 'omics data streams, and data that I collected myself by conducting fieldwork.

5.1 Integrating 'omics technologies to explore the genomic basis of complex trait evolution: functional genomics and phenotypic consequences of host switching in Photorhabdus species

In Chapter 2, I analysed and integrated RNA-seq and phenotype microarray (phenoarray) data to explore the molecular basis of phenotypic differences that may underlie host switching in *Photorhabdus* species of bacteria. *P. luminescens* (Pl^{TT01}) is restricted to insect hosts, where *P. asymbiotica* ($Pa^{ATCC43949}$ and $Pa^{Kingscliff}$) were isolated from clinical samples. By comparing gene expression and respiration patterns for each species/strain under various environmental conditions (temperature and substrate), I have identified specific metabolic pathways that may represent functional differences underlying the different host specificity in these species. The data available only provided one biological replicate of RNA-seq data and two replicates of phenoarray data, thus the statistical power to draw robust inferences in this study was limited and, as such, the findings presented here should be considered as preliminary indications of possible molecular differences.

Previous studies in bacterial systems have implicated changes to the function of various metabolic pathways in facilitating adaptation and reaction to different host conditions (Gray et al., 2006; Line et al., 2010; Stevenson et al., 1995). Here we present evidence that this may indeed be the case with *Photorhabdus* species. It appears that gene expression in both the insect-restricted (Pl^{TT01}), and the insect and mammalian pathogens ($Pa^{ATCC43949}$ and $Pa^{Kingscliff}$) responds most significantly to changes in growth medium than to changes to temperature or growth phase, indicating direct effects of environmental sensing on metabolism. The results indicate that $Pa^{ATCC43949}$ and $Pa^{Kingscliff}$ may be more responsive to growth conditions in mammalian systems than insect systems as more genes are switched on under mammalian-type conditions than in insect-type conditions, which is interesting given that survival in insect hosts is presumed to be the ancestral state. However, an alternative scenario should also be considered: increased transcriptional activation may represent response to the stress of mammalian-type conditions if these are suboptimal conditions for growth, as has been demonstrated in other species (Goh et al., 2002; Mostertz et al., 2004). Previous studies have highlighted high levels of variation in bacterial transcriptomic, and specifically metabolic, activity in response to environmental cues (Buescher *et al.* 2012; Nicolas *et al.* 2012), suggesting that the bacterial transcriptome machinery may be highly adaptable to enable survival. Given that *Photorhabdus* species have not maintained a large proportion of Enterobacterial ancestral genes (Baumler et al., 2013) and have a relatively high propensity toward gene duplication which may confer environmental adaptation (Bratlie *et al.* 2010), it may be that the genus possesses inherently high levels of transcriptome activity variation, which enables rapid adaptation to host conditions.

KEGG pathway analysis of RNA-seq data point to species-specific differences centred around glycine, serine and threonine metabolism: this pathway appears to be a switch where two genes were significantly up-regulated in *P. luminescens* and two were down-regulated in *P. asymbiotica*. Phenoarray data lends support to these observations: Pl^{TT01} respiration was significantly lower than $Pa^{ATCC43949}$ on L-serine and the dipeptide glycine-asparagine (Gly-Asn). Phenoarray data from *Campylobacter jejuni* indicates that growth temperature variation prompts differential carbon usage, with L-serine utilisation specifically up-regulated at higher temperatures (Line *et al.* 2010). It is known that glycine betaine, the trimethylated derivative of glycine, provides tolerance to osmotic stress in some *Enterobacteriaceae* when accumulated intracellularly (Le Rudulier & Bouillard 1983). Thus, these findings may represent pathways that can be found in a range of bacteria conferring the ability to survive at higher temperatures. The availability of phenotypic data to add insight into the functional mechanisms that are different between these two species provides substantiation of the most salient differential gene expression results, reinforcing the choices of the best candidate genes and pathways to investigate further.

5.2 Transcriptome annotation in species lacking a sequenced genome: the impact of sequence divergence and annotation strategy on efficacy, accuracy and functional bias

Using publically available RNA-seq data from *Drosophila melanogaster* and genome sequences from the 12 sequenced *Drosophila* species, I have characterised the impact of sequence divergence and strategy on the efficacy, accuracy and functional bias of transcriptome annotation to highlight the most appropriate technique to use with species lacking a sequenced genome, attaching to these approximate error values (Chapter 3). Observations with the *Drosophila* data were verified using human RNA-seq data in conjunction with primate genome sequences. I demonstrate that direct genome mapping (DGM) outperforms the assembly methods we assess (genome-guided and two types of *de novo* assembly) in terms of gene detection, accuracy, and functional bias of detected genes. With all transcriptome annotation methods tested, there is variation in both the detection of gene functional categories. Also, there is substantial variation in the error of gene detection per functional category which can be exploited for comparative molecular ecology studies to select the functional categories that are most likely to contain accurately detected genes.

Matching transcriptome sequences of one species to genomic DNA sequences from another species, either *in vitro* (heterologous hybridisation) or *in silico* (computational sequence alignment) is a useful technique for identifying orthologous sequences (Renn et al., 2004; Renn et al., 2010; Schunter et al., 2014; Shi et al., 2011). Others have previously shown that, as expected, increasing sequence divergence between transcriptome and genome sequences has a negative relationship with the proportion of sequences that have matches, and hence the proportion of orthologous genes

that can be detected (Hornett & Wheat, 2012; Renn et al., 2010). Transcriptome assembly methods aim to re-construct transcript sequences, either *de novo* or using a reference as a guide, which can then be matched against reference sequences for annotation. However, the processes by which the transcripts are assembled can be prone to error (Jain et al., 2013; Martin & Wang, 2011). Hence we sought to explore the impact of sequence divergence on the ability to map transcriptome sequences from three different annotation strategies: *de novo* and genome guided assembly, plus direct genome mapping (DGM). We used a gapped aligner to help overcome the effects of sequence divergence to a degree, maximising the proportion of sequences that will have matches. Some recent studies have assessed the accuracy of various assembly tools in recovering transcripts at the base level (Lu et al., 2013; Vijay et al., 2013) but there has been a lack of characterisation of the error associated with complete sequences being assigned to correct genomic locations and the error and bias in detecting correct orthologous genes, which is relevant for gene profiling studies seeking to explore functional differences between ecologically interesting scenarios.

Consistent with previous findings (Hornett & Wheat, 2012; Renn et al., 2010), sequence divergence between transcriptome and reference species has a negative relationship with the proportion of transcriptome sequences that are assigned to orthologous genes, and that error in transcript sequence assignment and gene detection increases with increasing sequence divergence. My work, however, goes beyond the aforementioned studies by showing that the differences in error between assembly-based annotation strategies and DGM are significant: DGM recovers more genes, both when a genome sequence is available for the transcriptome species and when it is not, and DGM is more accurate than the assembly-based methods. These findings indicate that there can be significant errors within the typical methods of transcriptome assembly (Ren *et al.* 2012) which can be avoided by using a simpler and more direct annotation technique like DGM. Where transcriptome assembly can be useful when assembling transcripts from a species with a sequenced genome (Martin & Wang 2011), my findings indicate that this does not necessarily hold true when assembling the transcriptomes of species that lack a sequenced genome. It would be expected that in the latter circumstance, transcript fragments generated directly by the sequencing platform are likely to be the most accurate at representing expressed sequences from that species, due to the low error rates of current next generation sequencing technologies. As such, the improvement of sequenced read lengths is likely to promote more accurate transcript detection over computationally combining sequences. To calculate gene detection accuracy of each annotation strategy we have used data sets of orthologous genes across the *Drosophila* and primate species. This provides a way of assessing whether orthologous genes to those detected when transcriptome and reference species are the same are also detected when using an alternative reference genome. Where the primate data set included only 1-to-1 orthologues, the *Drosophila* set contained 1-to-many. Where using all genes detected could have led to enhanced gene detection values, our gene detection method involved isolating only those reads that mapped to a single location and mapped to a gene. As such,

for any read that mapped to a gene in *D. melanogaster* and mapped to more than one orthologous gene in an alternative species, this would have been removed, essentially enforcing detection of only 1-to-1 orthologues.

When endeavouring to recover a representative transcriptome profile of a species lacking a sequenced genome, sequence divergence may lead to skew in the function of genes detected: fast evolving genes are likely to exhibit enhanced sequence divergence with respect to their orthologues compared to their more conserved counterparts. This issue has been raised previously by others in the context of microarray studies (Le Quéré et al., 2006; Renn et al., 2010) but had not, until now, been fully explored in the context of next generation sequencing. Hornet and Wheat (2012) reported proportions of gene families that were biased with increasing divergence when using assembly-based methods but did not demonstrate how the degree of skew varied per GO slim term with increasing divergence, nor how gene detection error varied per term (Hornett & Wheat 2012). By exploring these aspects, I demonstrate here that there is considerable variation in functional term detection, by way of the proportion of genes detected per term compared to the expected values, and that error levels vary per term. These findings indicate that sequence divergence has a significant impact on the overall functional distribution of the transcriptome, which must be considered when conducting comparative studies in species lacking sequenced genomes. However, these results also clearly show that DGM considerably outperforms the assembly-based methods in the degree of functional bias induced. As such, comparative studies can dramatically minimise functional bias by choosing DGM over assembly-based methods. Some functional terms consistently exhibit low or zero gene detection error and hence these terms are good candidates for core gene expression comparisons between species.

Given the superior efficacy, accuracy, and relatively unbiased nature of DGM over assembly-based methods, this has quite profound implications for previous studies that have not only identified genes based on assembly methods but also drawn functional inferences from those gene lists. It is anticipated that re-annotation of previously published data using DGM could yield larger and more reliable annotated transcriptome data sets, which could help generate more in depth and robust insight into the evolution and development of complex traits.

5.3 Uncovering the brain gene expression signatures of mating system evolution: novel sequencing, annotation and functional comparison of the water pipit and dunnock brain transcriptomes

Songbirds have been demonstrated to be an excellent model for exploring the molecular basis of social behaviour related to mating (Clayton et al., 2009; Goodson et al., 2009). Where some studies have begun to uncover the genetic basis of pair bonding behaviour (Ahern & Young, 2009; Cho et al., 1999; McGraw & Young, 2010; Ophir et al., 2012), there has, until now, been no genome-wide exploration of brain gene expression underlying differences between monogamous and polygamous species of bird. To investigate the brain gene expression profiles underlying differences in mating system evolution in songbirds, we obtained, sequenced and analysed brain transcriptomes from wild-caught songbird species that have opposing mating systems but, as yet, no genomic resources available (Chapter 4). Water pipits are typically monogamous whereas dunnocks are variable and can be highly polygamous (Bollmann & Reyer, 1999; Burke et al., 1989; Griffith et al., 2002). By mapping the transcriptomes of these species to the genome of the closest available reference, the zebra finch, using the highly effective, accurate and efficient DGM technique, I have characterised the functional gene expression profiles of both species, providing the first genomic resources for these species, and conducted the first comparison of brain gene expression differences between a monogamous and a polygamous bird species. As the differential brain gene expression comparison was for only one pair of species, the results indicate either general species-specific differences that may or may not be related to mating behaviour, or simply neutral divergence of no phenotypic effect. Also, given that RNA samples for each species could not be sequenced individually, natural variation in gene expression could not be calculated and as such these findings are preliminary indications of the gene expression differences between these species.

Other recent studies have presented transcriptome characterisation of songbirds that lack an available genome sequence (Balakrishnan et al., 2013; Moghadam et al., 2013), using assembly-based methods. Having demonstrated that DGM is the most appropriate method for accurately annotating the transcriptomes of species that lack sequenced genomes, I verify that observation using these novel transcriptomes, generating larger expressed gene lists than assemblies could achieve. The RNA-seq data sets detected over 90% of annotated zebra finch genes. Functional analysis returned similar types of genes as enriched/depleted within each data set, indicating that the annotated transcriptomes were functionally similar and therefore comparable for the purposes of this study. I find 62 genes as significantly differentially expressed which indicate specific pathways and functions as different between the male water pipit and dunnock brains during their breeding season. However, as we were unable to sequence individual transcriptomes separately, these results provide a proof of principle and a preliminary indication of the genes that may represent species-specific differences that contribute to differences in mating behaviour.

It appears that key functional molecular differences between the water pipit and the dunnock are related to neuroprotection from oxidative stress/inflammation, metabolic control and neurogenesis within the brain. However, both species appear to express genes from similar functional pathways, albeit different genes, which may indicate that similar functional programmes are mediated by different genomic factors. These differences may reflect the differing impacts of neurological activity related to mating choices and their afferent and efferent signals. The dunnock and water pipit differ around expression of genes involved in detoxification, defence against oxidative stress, and mitochondrial function (higher in the dunnock), and genes that engender greater flexibility in gene expression, which may crosstalk with mitochondrial processes and may be related to neurogenesis (higher in the water pipit). Steroid hormones are known to impact upon the metabolic functions of neurons and environmental cues feed in via receptor-mediated signals, converging on mitochondria function. Energy demand is critical within the central nervous system (CNS) for maintaining membrane ionic gradients, requiring high ATP metabolism. During the breeding season of birds, significant changes take place within the brain, impacting on cell number, location and activity (Tramontin & Brenowitz 2000). Our findings suggest that although similar processes may be occurring within the brains of each species, around maintaining a good energy balance and neurogenesis, these may comprise slightly different pathways, the operation of which may reflect the overall requirements of the tissue, such as defence against oxidative/inflammatory stress in the polygamous species. If this is indeed the case, that polygamous mating choices go hand-in-hand with the need to reduce the impact of neural stressors not similarly experienced in monogamous species, this represents an interesting possibility. It may be that the evolution of mating behaviour differences between these species has been shaped by the stressful internal physiological impacts of responses to external environmental opportunities and challenges related to the availability of reproductive resources and that, as a result, behavioural pathways are integrated within those required to maintain the overall health of the brain.

5.4 Future directions

To functionally validate my observations in *Photorhabdus,* genetic knock-down experiments (using methods that have been demonstrated in bacterial systems such as using RNAi [Blau & McManus, 2013; Szaszák et al., 2013] or the CRISPR-Cas system, [Sander & Joung, 2014]) could be performed on the genes that appear to form the functional switch within the glycine, serine and threonine pathway and phenotypic effects could be observed by repeating the phenoarray assays. To gain further insight into temperature-dependent metabolic differences between *P. luminescens* and *P. asymbiotica*, a greater range of temperatures could be used for both comparative gene expression profiling and phenoarray with many biological replicates to ensure that the high level of natural variation in *Photorhabdus* growth is adequately accounted for. Specifically, it would be

important to include data points for *P. luminescens* and *P. asymbiotica* cultured in lysogeny broth supplemented with insect haemoplymph (LBHm) at 37^{o}C.

To take my work on optimising RNA-seq annotation strategies further, it would be beneficial to incorporate RNA-seq data sets from more advanced platforms that are able to generate significantly longer reads than those used here with comparable sequencing accuracy. The average length of RNA-seq reads has increased from 25bp from Solexa's first platform to the 150bp of the recent Illumina HiSeq instrument (Mardis 2013). As mentioned earlier, this is likely to have a beneficial impact on the accuracy of gene detection when directly mapping reads to the genome. Additionally, as computational tools are constantly developing and new ones are appearing on a regular basis, it would be useful to monitor the relative performance of DGM against these. Given that the cost of sequencing has dramatically decreased since next generation sequencing first appeared, it is entirely likely in a matter of a few years that the cost of sequencing the genome of the species of interest, if not immediately available, will negate the need to use a reference sequence from a related species. The findings presented here will continue to be of use in this situation with regards to the accuracy of single versus multimatch sequences and overall gene detection capabilities of DGM compared to more complex methods.

The major limitation of my songbird brain transcriptome study is the lack of individually sequenced RNA-seq samples: this precluded us from establishing natural variation in the levels of gene expression and therefore reduced the confidence that could be vested in the differential expression results. It also meant that allelic expression could not be detected. Given that species categorised as having a particular mating system exhibit variation in the extent of extra pair paternity (Brommer *et al.* 2010), it may be that certain alleles of key genomic loci are strong influencers of mating behaviour and the relative abundance of such alleles within a population therefore exerts a strong influence over the overall levels of extra pair paternity. To explore this aspect and determine the genomic loci of greatest importance for influencing mating choices and partner preference formation, quantitative-trait loci (QTL) mapping could be performed. By increasing the number of individual samples sequenced per species, it would be possible to construct gene co-expression networks which would allow the identification of putative gene regulatory modules, differences in which between species may identify possible differences in gene expression regulation that impacts upon and/or is impacted by differences in mating system. This technique has been recently used to identify gene modules involved in phenotypic expression (Ficklin et al., 2010; Filteau et al., 2013). Alternatively, the variation in individual partner preferences may result from incomplete penetrance of contributing genetic factors or phenocopy. Additionally, epigenomic factors may exert variable influence over resultant behaviour. To tackle this, bisulfite sequencing, sequencing of non-coding RNAs (ncRNAs), and CHiP-seq (chromatin

immunoprecipitation sequencing) could be used to identify possible regions and perhaps even genes that are epigenetically silenced, which in conjunction with expanded RNA-seq data could help generate more holistic insight into differentially regulated genes pathways between the water pipit and the dunnock.

At the very inception of this project, it was aimed to obtain not only several songbird species pairs but females as well as males. More pairs of species would have aided the detection of any conserved differentially expressed genes and pathways in the evolution of mating systems in songbirds. Having only one species pair we can currently only postulate that the genes and pathways we identify are linked to mating system, where they may instead be linked simply to species-specific differences, or indeed to neutral divergence of no phenotypic effect. Using females as well as males would have allowed us to not only strengthen the statistical power of our inferences in determining sex-independent gene expression patterns, but would have also allowed us to determine sexually dimorphic expression. This may have shed some light onto the sex-specific impacts of sexual selection on the brain and paved the way for further investigation of male versus female-directed mate choice.

5.5 Concluding remarks

Gene expression studies are immensely useful tools for uncovering genes and functional pathways underlying complex traits such as social behaviour. In spite of immense advances in recent years, significant challenges remain in this area. One issue, particularly within the brain, is the spatial restriction of gene expression: only a small population of cells may express the genes pivotal of influence. Optogenetics is making headway into this area (Deisseroth 2010, 2011) but there are many more hurdles to overcome. Given that genes act in concert, within complex positive and negative feedback modules, as do cells within neural circuits, integrative functional network approaches at different levels of complexity will be useful for identifying and disentangling the critical paths and components for different traits. Additionally, when considering the biologically meaningful context of living in nature, gene expression and cellular activity may vary throughout the day/season, responding to internal physiological and external environmental influences on the individual. As such, it may be necessary to computationally model the molecular responses of key cell types and neural circuits in a laboratory setting, and develop predictors of this activity, such as circulating hormone levels, or dynamic epigenomic modification, for use in (semi-)natural populations. The development of non-invasive genomic and cellular predictors would enable the real-time modelling of the internal changes related to behavioural fluctuation, particularly where related and unrelated groups of individuals are used.

Recent years have seen the rapid development and expansion of genomic sequencing technologies, alongside other 'omics technologies such as proteomics, metabolomics, and even phenomics. The amount and diversity of this data that these technologies accumulates threatens to outstrip our computation power to analyse it, although integrating 'omics has developed into a field of inquiry in its own right (Berger et al., 2013). Within genomic sequencing alone, there is heterogeneity in the data generated from different sequencing platforms, presenting additional computational challenges particularly where comparative studies seek to integrate data from a large number of previously published data (Berger et al., 2013). Therefore, the need to optimise methods for integrating and analysing data both within and between 'omics streams is paramount in order to better understand relevant processes in living organisms. Indeed, the former must occur before the latter can proceed. This thesis stimulates a questioning of some of the currently widely accepted methods for transcriptome annotation of species without available genome sequences, by illustrating the error and bias inherent to these methods and demonstrating that a far simpler method proves superior. With the continual advancement of data generation technologies and the development of novel computational tools for data processing, there is a risk that the identification and optimisation of the most appropriate analytical techniques for the biological question at hand may be overshadowed by the real and perceived benefits provided by new approaches. This thesis has highlighted the importance and relevance of continual critical assessment of the best available analytical options to ensure that knowledge is maximised from all data generated, particularly when animals have been used.

Looking forward, some of the work started here is being advanced by further students. For example, I used the DGM technique to re-annotate previously published brain transcriptomes from a variety of bird species and this data is now being used to explore the relationship between brain gene expression and various phenotypes among birds. Additional brain transcriptomes are planned to be obtained from wild bird species with interesting ecological models of mating and parental behaviour, which will enable further insight to be generated regarding the molecular genomic basis of aspects of avian social behaviour, advancing my findings. We are additionally looking to explore comparative brain gene co-expression networks in birds and mammals to highlight areas of conservation and divergence. Taking a broad phylogenetic viewpoint will hopefully allow us to highlight genes, perhaps even gene modules, of major importance in the evolution of social traits in many species, which, when expanded in the context of integrative, whole system approaches, will help us unravel the secrets of the animal social brain.

Abràmoff MD, Magalhães PJ, Ram SJ (2004) Image Processing with ImageJ. *Biophotonics International*, **11**, 36–42.

Adams MD (2000) The Genome Sequence of *Drosophila melanogaster*. *Science*, **287**, 2185–95.

Ahern TH, Young LJ (2009) The impact of early life family structure on adult social attachment, alloparental behavior, and the neuropeptide systems regulating affiliative behaviors in the monogamous prairie vole (*Microtus ochrogaster*). *Frontiers in behavioral neuroscience*, **3**, 1–19.

Ahn S-Y, Jamshidi N, Mo ML et al. (2011) Linkage of organic anion transporter-1 to metabolic pathways through integrated "omics"-driven network and functional analysis. *Journal of biological chemistry*, **286**, 31522–31.

Alagna F, D'Agostino N, Torchia L et al. (2009) Comparative 454 pyrosequencing of transcripts from two olive genotypes during fruit development. *BMC Genomics*, **10**, 399.

Altschup SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic Local Alignment Search Tool. *Journal of molecular biology*, **215**, 403–10.

Anders S (2012) Analysing RNA-Seq data with the DESeq package. *European Molecular Biology Laboratory*, 1–28.

Anders S, Huber W (2010) Differential expression analysis for sequence count data. *Genome biology*, **11**, R106.

Anders S, Pyl PT, Huber W (2014) HTSeq – A Python framework to work with high-throughput sequencing data. *Bioinformatics*, 1–4.

Aragona BJ, Liu Y, Yu YJ et al. (2006) Nucleus accumbens dopamine differentially mediates the formation and maintenance of monogamous pair bonds. *Nature neuroscience*, **9**, 133–9.

Arnqvist G (1992) Pre-copulatory fighting in a water strider: inter-sexual conflict or mate assessment? *Animal Behaviour*, **43**, 559–67.

Aubin-Horth N, Desjardins JK, Martei YM, Balshine S, Hofmann HA (2007) Masculinized dominant females in a cooperatively breeding species. *Molecular ecology*, **16**, 1349–58.

Balakrishnan CN, Chapus C, Brewer MS, Clayton DF (2013) Brain transcriptome of the violet-eared waxbill *Uraeginthus granatina* and recent evolution in the songbird genome. *Open biology*, **3**, 130063.

Balshine S (2012) Patterns of parental care in vertebrates. In: *The evolution of parental care*. (eds Royle N, Smiseth P, Molliker M), pp. 60–80. Oxford University Press.

Barakat A, DiLoreto DS, Zhang Y et al. (2009) Comparison of the transcriptomes of American chestnut (*Castanea dentata*) and Chinese chestnut (*Castanea mollissima*) in response to the chestnut blight infection. *BMC plant biology*, **9**, 51.

Baumler DJ, Ma B, Reed JL, Perna NT (2013) Inferring ancient metabolism using ancestral core metabolic models of enterobacteria. *BMC systems biology*, **7**, 46-63.

Behl C, Skutella T, Lezoualc'h F et al. (1997) Neuroprotection against Oxidative Stress by Estrogens : Structure-Activity Relationship. *Molecular Pharmacology*, **51**, 535–41.

Benes V, Muckenthaler M (2003) Standardization of protocols in cDNA microarray analysis. *Trends in biochemical sciences*, **28**, 244–9.

Berger B, Peng J, Singh M (2013) Computational solutions for omics data. Nature reviews. *Genetics*, **14**, 333–46.

Bielsky IF, Hu S-B, Szegda KL, Westphal H, Young LJ (2004) Profound impairment in social recognition and reduction in anxiety-like behavior in vasopressin V1a receptor knockout mice. *Neuropsychopharmacology*, **29**, 483–93.

Birkhead TR, Briskie J V, Möller AP (1993) Male sperm reserves and copulation frequency in birds. *Behavioral Ecology and Sociobiology*, **32**, 85–93.

Blanchard RJ, McKittrick CR, Blanchard DC (2001) Animal models of social stress: effects on behavior and brain neurochemical systems. *Physiology & behavior*, **73**, 261–71.

Blau JA, McManus MT (2013) Renewable RNAi. *Nature biotechnology*, **31**, 319–20.

Blekhman R, Marioni JC, Zumbo P, Stephens M, Gilad Y (2010) Sex-specific and lineage-specific alternative splicing in primates. *Genome research*, **20**, 180–9.

Bochner BR, Gadzinski P, Panomitros E (2001) Phenotype microarrays for high-throughput phenotypic testing and assay of gene function. *Genome research*, **11**, 1246–55.

Bochner BR, Savageau MA (1977) Generalized indicator plate for genetic, metabolic, and taxonomic studies with microorganisms. *Applied and Environmental Microbiology*, **33**, 434–44.

Bolker JA (2014) Model species in evo-devo: a philosophical perspective. *Evolution & development*, **16**, 49–56.

Bollmann K, Reyer H (1999) Why does monogamy prevail in the Alpine Water Pipit *Anthus spinoletta*? *Proceedings of the International Ornithological Congress*, **22**, 2666–88.

Van Bon BWM, Oortveld MAW, Nijtmans LG et al. (2013) CEP89 is required for mitochondrial metabolism and neuronal function in man and fly. *Human molecular genetics*, **22**, 3138–51.

Borglin S, Joyner D, DeAngelis KM et al. (2012) Application of phenotypic microarrays to environmental microbiology. *Current opinion in biotechnology*, **23**, 41–8.

Boyd EF, Brüssow H (2002) Common themes among bacteriophage-encoded virulence factors and diversity among the bacteriophages involved. *Trends in microbiology*, **10**, 521–9.

Bratlie MS, Johansen J, Sherman BT et al. (2010) Gene duplications in prokaryotes can be associated with environmental adaptation. *BMC genomics*, **11**, 588-605.

Bräutigam A, Shrestha RP, Whitten D et al. (2008) Low-coverage massively parallel pyrosequencing of cDNAs enables proteomics in non-model species: comparison of a species-specific database generated by pyrosequencing with databases from related

species for proteome analysis of pea chloroplast envelopes. *Journal of biotechnology*, **136**, 44–53.

Brawand D, Soumillon M, Necsulea A et al. (2011) The evolution of gene expression levels in mammalian organs. *Nature*, **478**, 343–8.

Brommer JE, Alho JS, Biard C et al. (2010) Passerine extrapair mating dynamics: a bayesian modeling approach comparing four species. *The American naturalist*, **176**, 178–87.

Brunberg E, Jensen P, Isaksson A, Keeling LJ (2013) Brain gene expression differences are associated with abnormal tail biting behavior in pigs. *Genes, brain, and behavior*, **12**, 275–81.

Buescher JM, Liebermeister W, Jules M et al. (2012) Global network reorganization during dynamic adaptations of *Bacillus subtilis* metabolism. *Science*, **335**, 1099–103.

Burke T, Davies N, Bruford M, Hatchwell B (1989) Parental care and mating behaviour of polyandrous dunnocks *Prunella modularis* related to paternity by DNA fingerprinting. *Nature*, **338**, 249–51.

Calhim S, Birkhead TR (2006) Testes size in birds: quality versus quantity--assumptions, errors, and estimates. *Behavioral Ecology*, **18**, 271–5.

Caroni P, Donato F, Muller D (2012) Structural plasticity upon learning: regulation and functions. *Nature reviews Neuroscience*, **13**, 478–90.

Cases O, Seif I, Curie I et al. (1995) Aggressive behavior and altered amounts of brain serotonin and norepinephrine in mice lacking MAOA. *Science*, **268**, 1763–6.

Chen D, Chen M, Altmann T, Klukas C (2014) *Approaches in Integrative Bioinformatics* (M Chen, R Hofestädt, Eds,). Springer Berlin Heidelberg, Berlin, Heidelberg.

Cho MM, DeVries AC, Williams JR, Carter CS (1999) The effects of oxytocin and vasopressin on partner preferences in male and female prairie voles (*Microtus ochrogaster*). *Behavioral neuroscience*, **113**, 1071–9.

Clark AG, Eisen MB, Smith DR et al. (2007) Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature*, **450**, 203–18.

Clayton DF, Balakrishnan CN, London SE (2009) Integrating genomes, brain and behavior in the study of songbirds. *Current biology*, **19**, R865–73.

Clipperton-Allen AE, Lee AW, Reyes A et al. (2012) Oxytocin, vasopressin and estrogen receptor gene expression in relation to social recognition in female mice. *Physiology & behavior*, **105**, 915–24.

Cloonan N, Forrest ARR, Kolle G et al. (2008) Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nature Methods*, **5**, 613–9.

Clutton-Brock T (2007) Sexual selection in males and females. *Science*, **318**, 1882–5.

Clutton-Brock TH, Hodge SJ, Spong G et al. (2006) Intrasexual competition and sexual selection in cooperative mammals. *Nature*, **444**, 1065–8.

Clutton-Brock TH, Isvaran K (2007) Sex differences in ageing in natural populations of vertebrates. *Proceedings of The Royal Society Biological sciences*, **274**, 3097–104.

Cohas A, Allainé D (2009) Social structure influences extra-pair paternity in socially monogamous mammals. *Biology letters*, **5**, 313–6.

Colgan TJ, Carolan JC, Bridgett SJ et al. (2011) Polyphenism in social insects: insights from a transcriptome-wide analysis of gene expression in the life stages of the key pollinator, *Bombus terrestris. BMC genomics*, **12**, 623.

Collins LJ, Voelckel C, Biggs PJ, Joly S (2008) An approach to transcriptome analysis of non-model organisms using short-read sequences. *Genome Informatics*, **21**, 3–14.

Consortium T modENCODE, Roy S, Ernst J et al. (2011) Identification of functional elements and regulatory circuits by *Drosophila* modENCODE. *Science*, **330**, 1787–97.

Cooney NM, Klein BS (2008) Fungal adaptation to the mammalian host: it's a new world, after all. *Current opinion in microbiology*, **11**, 511–6.

Costantini D, Marasco V, Møller AP (2011) A meta-analysis of glucocorticoids as modulators of oxidative stress in vertebrates. *Journal of comparative physiology*, *B*, **181**, 447–56.

Crawford JE, Guelbeogo WM, Sanou A et al. (2010) *De novo* transcriptome sequencing in *Anopheles funestus* using Illumina RNA-seq technology. *PloS one*, **5**, e14202.

Czibere L, Baur LA, Wittmann A et al. (2011) Profiling trait anxiety: transcriptome analysis reveals cathepsin B (Ctsb) as a novel candidate gene for emotionality in mice. *PloS one*, **6**, e23604.

Darwin C (1871) The descent of man, and Selection in relation to sex, Vol 2. John Murray, London, England.

Dassanayake M, Haas JS, Bohnert HJ, Cheeseman JM (2009) Shedding light on an extremophile lifestyle through transcriptomics. *The New phytologist*, **183**, 764–75.

David M, Dzamba M, Lister D, Ilie L, Brudno M (2011) SHRiMP2: sensitive yet practical SHort Read Mapping. *Bioinformatics*, **27**, 1011–2.

Davies NB (1992) Dunnock behaviour and social evolution. Oxford University Press.

Deisseroth K (2010) Controlling the brain with light. *Scientific American*, 48–55.

Deisseroth K (2011) Optogenetics. *Nature methods*, **8**, 26–29.

Deviche PJ, Hurley LL, Fokidis HB et al. (2010) Acute stress rapidly decreases plasma testosterone in a free-ranging male songbird: potential site of action and mechanism. *General and comparative endocrinology*, **169**, 82–90.

Van Dijk RE, Mészáros LA, van der Velde M et al. (2010) Nest desertion is not predicted by cuckoldry in the Eurasian penduline tit. *Behavioral ecology and sociobiology*, **64**, 1425–35.

Dorus S, Wasbrough ER, Busby J, Wilkin EC, Karr TL (2010) Sperm proteomics reveals intensified selection on mouse sperm membrane and acrosome genes. *Molecular biology and evolution*, **27**, 1235–46.

Duchaud E, Rusniok C, Frangeul L et al. (2003) The genome sequence of the entomopathogenic bacterium Photorhabdus luminescens. *Nature biotechnology*, **21**, 1307–13.

Durban J, Pérez A, Sanz L et al. (2013) Integrated “omics” profiling indicates that miRNAs are modulators of the ontogenetic venom composition shift in the Central American rattlesnake, *Crotalus simus simus*. *BMC genomics*, **14**, 234.

Duret L, Mouchiroud D (2000) Determinants of substitution rates in mammalian genes: expression pattern affects selection intensity but not mutation rate. *Molecular biology and evolution*, **17**, 68–74.

Emlen DJ (2008) The Evolution of Animal Weapons. *Annual Review of Ecology, Evolution, and Systematics*, **39**, 387–413.

Emlen ST, Oring LW (1977) Ecology, sexual selection, and the evolution of mating systems. *Science*, **197**, 215–23.

Engelmann I, Griffon A, Tichit L et al. (2011) A comprehensive analysis of gene expression changes provoked by bacterial and fungal infection in *C. elegans*. *PloS one*, **6**, e19055.

Esteve-Codina A, Kofler R, Palmieri N et al. (2011) Exploring the gonad transcriptome of two extreme male pigs with RNA-seq. *BMC genomics*, **12**, 552.

Farmer JJ, Jorgensen JH, Grimont PA et al. (1989) *Xenorhabdus luminescens* (DNA hybridization group 5) from human clinical specimens. *Journal of clinical microbiology*, **27**, 1594–600.

Felderhoff-Mueser U, Schmidt OI, Oberholzer A, Bührer C, Stahel PF (2005) IL-18: a key player in neuroinflammation and neurodegeneration? *Trends in neurosciences*, **28**, 487–93.

Ficklin SP, Luo F, Feltus FA (2010) The association of multiple interacting genes with specific phenotypes in rice using gene coexpression networks. *Plant physiology*, **154**, 13–24.

Filteau M, Pavey SA, St-Cyr J, Bernatchez L (2013) Gene coexpression networks reveal key drivers of phenotypic divergence in lake whitefish. *Molecular biology and evolution*, **30**, 1384–96.

Firat-Karalar EN, Sante J, Elliott S, Stearns T (2014) Proteomic analysis of mammalian sperm cells identifies new components of the centrosome. *Journal of cell science*, **127**, 4128–33.

Fischer-Le Saux M, Viallardt V, Brunelt B, Normand P, Boemarel NE (1999) Polyphasic classification of the genus *Photorhabdus* and proposal of new taxa*: P. luminescens* subsp. *luminescens* subsp. nov., *P. luminescens* subsp . *akhurstii* subsp . nov ., *P. luminescens* subsp . *laumondii* subsp . nov ., *P. temperata* sp. nov., *P. tempera*. *International Journal of Systematic Bacteriology*, **49**, 1645–56.

Fjeldså J, Irestedt M, Ericson PGP, Zuccon D (2010) The Cinnamon Ibon *Hypocryptadius cinnamomeus* is a forest canopy sparrow. *Ibis*, **152**, 747–60.

Flicek P, Amode MR, Barrell D et al. (2014) Ensembl 2014. *Nucleic acids research*, **42**, D749–55.

Flint J, Mott R (2001) Finding the molecular basis of quantitative traits: successes and pitfalls. *Nature reviews Genetics*, **2**, 437–45.

Franssen SU, Shrestha RP, Bräutigam A, Bornberg-Bauer E, Weber APM (2011) Comprehensive transcriptome analysis of the highly complex *Pisum sativum* genome using next generation sequencing. *BMC genomics*, **12**, 227-43.

Freemantle E, Mechawar N, Turecki G (2013) Cholesterol and phospholipids in frontal cortex and synaptosomes of suicide completers: relationship with endosomal lipid trafficking genes. *Journal of psychiatric research*, **47**, 272–9.

Fuhrer T, Zamboni N (2015) High-throughput discovery metabolomics. *Current Opinion in Biotechnology*, **31**, 73–8.

Garamszegi LZ, Eens M, Hurtrez-Boussès S, Møller AP (2005) Testosterone, testes size, and mating success in birds: a comparative study. *Hormones and behavior*, **47**, 389–409.

Garamszegi LZ, Møller AP (2004) Extrapair paternity and the evolution of bird song. *Behavioral Ecology*, **15**, 508–19.

Garber M, Grabherr MG, Guttman M, Trapnell C (2011) Computational methods for transcriptome annotation and quantification using RNA-seq. *Nature methods*, **8**, 469–77.

Garfield AS, Cowley M, Smith FM et al. (2011) Distinct physiological and behavioural functions for parental alleles of imprinted Grb10. *Nature*, **469**, 534–8.

Garg R, Patel RK, Tyagi AK, Jain M (2011) *De novo* assembly of chickpea transcriptome using short reads for gene discovery and marker identification. *DNA research*, **18**, 53–63.

Ge H, Walhout AJM, Vidal M (2003) Integrating "omic" information: a bridge between genomics and systems biology. *Trends in genetics*, **19**, 551–60.

Gerrard JG, Joyce SA, Clarke DJ et al. (2006) Nematode Symbiont for *Photorhabdus asymbiotica*. *Emerging infectious diseases*, **12**, 1562–4.

Gerrard JG, McNevin S, Alfredson D, Forgan-Smith R, Fraser N (2003) *Photorhabdus* species: bioluminescent bacteria as emerging human pathogens? *Emerging infectious diseases*, **9**, 251–4.

Giardine B, Riemer C, Hardison RC et al. (2005) Galaxy : A platform for interactive large-scale genome analysis. *Genome research*, **15**, 1451–5.

Gimpl G, Fahrenholz F (2001) The oxytocin receptor system: structure, function, and regulation. *Physiological reviews*, **81**, 629–83.

Goecks J, Nekrutenko A, Taylor J (2010) Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome biology*, **11**, R86.

Goh E, Yim G, Tsui W et al. (2002) Transcriptional modulation of bacterial gene expression by subinhibitory concentrations of antibiotics. *Proceedings of the National Academy of Sciences*, **99**, 17025–30.

Goodson JL, Kabelik D, Kelly AM, Rinaldi J, Klatt JD (2009a) Midbrain dopamine neurons reflect affiliation phenotypes in finches and are tightly coupled to courtship.

Proceedings of the National Academy of Sciences of the United States of America, **106**, 8737–42.

Goodson JL, Rinaldi J, Kelly AM (2009b) Vasotocin neurons in the bed nucleus of the stria terminalis preferentially process social information and exhibit properties that dichotomize courting and non-courting phenotypes. *Hormones and behavior*, **55**, 197–202.

Grabherr MG, Haas BJ, Yassour M et al. (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature biotechnology*, **29**, 644–52.

Gray MJ, Freitag NE, Boor KJ (2006) How the bacterial pathogen *Listeria monocytogenes* mediates the switch from environmental Dr . Jekyll to pathogenic Mr . Hyde. *Infection and immunity*, **74**, 2505–12.

Greenwood PJ (1980) Mating systems, philopatry and dispersal in birds and mammals. *Animal*, **28**, 1140–62.

Griffith SC, Owens IPF, Thuman KA (2002) Extra pair paternity in birds: a review of interspecific variation and adaptive function. *Molecular ecology*, **11**, 2195–212.

Hao DC, Ge G, Xiao P, Zhang Y, Yang L (2011) The first insight into the tissue specific taxus transcriptome via Illumina second generation sequencing. *PloS one*, **6**, e21220.

Herzog H, Höfferer L, Schneider R, Schweiger M (1990) cDNA encoding the human homologue of rat ribosomal protein L35a. *Nucleic acids research*, **18**, 4600.

Higginbotham HR, Gleeson JG (2007) The centrosome in neuronal development. *Trends in neurosciences*, **30**, 276–83.

Hollis B, Houle D, Yan Z, Kawecki TJ, Keller L (2014) Evolution under monogamy feminizes gene expression in *Drosophila melanogaster*. *Nature communications*, **5**, 3482-7.

Holmes A, Murphy DL, Crawley JN (2002) Reduced aggression in mice lacking the serotonin transporter. *Psychopharmacology*, **161**, 160–7.

Hong G, Zhang W, Li H, Shen X, Guo Z (2014) Separate enrichment analysis of pathways for up- and downregulated genes. *Journal of the Royal Society Interface*, **11**, 20130950.

Hornett EA, Wheat CW (2012) Quantitative RNA-Seq analysis in non-model species: assessing transcriptome assemblies as a scaffold and the utility of evolutionary divergent genomic reference species. *BMC genomics*, **13**, 361-77.

Houle D (2010) Numbering the hairs on our heads: the shared challenge and promise of phenomics. *Proceedings of the National Academy of Sciences of the United States of America*, **107 Suppl** , 1793–9.

Houle D, Govindaraju DR, Omholt S (2010) Phenomics: the next challenge. *Nature reviews Genetics*, **11**, 855–66.

Van Hout AJ-M, Eens M, Darras VM, Pinxten R (2010) Acute stress induces a rapid increase of testosterone in a songbird: implications for plasma testosterone sampling. *General and comparative endocrinology*, **168**, 505–10.

Huang DW, Sherman BT, Lempicki RA (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols*, **4**, 44–57.

Ikonomidou C, Kaindl AM (2011) Neuronal death and oxidative stress in the developing brain. *Antioxidants & Redox Signaling*, **14**, 1535–50.

International Chicken Genome Sequencing Consortium (2004) Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature*, **432**, 695–716.

International Chicken Polymorphism Map Consortium (2004) A genetic variation map for chicken with 2.8 million single-nucleotide polymorphisms. *Nature*, **432**, 717–722.

Iorizzo M, Senalik DA, Grzebelus D et al. (2011) *De novo* assembly and characterization of the carrot transcriptome reveals novel genes, new markers, and genetic diversity. *BMC genomics*, **12**, 389-404.

Jacob S, Brune CW, Carter CS et al. (2007) Association of the oxytocin receptor gene (OXTR) in Caucasian children and adolescents with autism. *Neuroscience letters*, **417**, 6–9.

Jain P, Krishnan NM, Panda B (2013) Augmenting transcriptome assembly by combining *de novo* and genome-guided tools. *PeerJ*, **1**, e133.

Jakobsen L, Vanselow K, Skogs M et al. (2011) Novel asymmetrically localizing components of human centrosomes identified by complementary proteomics methods. *EMBO Journal*, **30**, 1520–35.

Jarvis ED, Güntürkün O, Bruce L et al. (2005) Avian brains and a new understanding of vertebrate brain evolution. Nature reviews. *Neuroscience*, **6**, 151–9.

Ji Y, Xu Y, Zhang Q et al. (2011) BM-Map: Bayesian mapping of multireads for next-generation sequencing data. *Biometrics*, **67**,1215-24.

Jiao X, Sherman BT, Huang DW et al. (2012) DAVID-WS: a stateful web service to facilitate gene/protein list analysis. *Bioinformatics*, **28**, 1805–6.

Kahm M, Hasenbrink G, Lichtenberg-Frate H, Ludwig J, Kschischo M (2010) grofit : Fitting biological growth curves with R. *Journal of statistical software*, **33**, 1-21.

Kanehisa M, Goto S (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, **28**, 27–30.

Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M (2012) KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic acids research*, **40**, D109–14.

Kawahara-Miki R, Wada K, Azuma N, Chiba S (2011) Expression profiling without genome sequence information in a non-model species, Pandalid shrimp (*Pandalus latirostris*), by next-generation sequencing. *PloS one*, **6**, e26043.

Kawakami K, Qureshi MH, Zhang T et al. (1997) IL-18 protects mice against pulmonary and disseminated infection with *Cryptococcus neoformans* by inducing IFN-gamma production. *Journal of Immunology*, **159**, 5528–34.

Keeling MJ, Gilligan CA (2000) Bubonic plague: a metapopulation model of a zoonosis. *Proceedings of The Royal Society Biological sciences*, **267**, 2219–30.

Khatri P, Drăghici S (2005) Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics*, **21**, 3587–95.

Kim S-J, Young LJ, Gonen D et al. (2002) Transmission disequilibrium testing of arginine vasopressin receptor 1A (AVPR1A) polymorphisms in autism. *Molecular psychiatry*, **7**, 503–7.

Kirkpatrick M, Hall DW (2004) Male-biased mutation sex linkage and the rate of adaptive evolution. *Evolution*, **58**, 437–40.

Kotrschal K, Hirschenhauser K, Mostl E (1998) The relationship between social stress and dominance is seasonal in greylag geese. *Animal behaviour*, **55**, 171–6.

Kroes RA, Jastrow A, Mclone MG et al. (2000) The identification of novel therapeutic targets for the treatment of malignant brain tumors. *Cancer letters*, **156**, 191–8.

Kumar A, Rajendran V, Sethumadhavan R, Purohit R (2013) CEP proteins: the knights of centrosome dynasty. *Protoplasma*, **250**, 965–83.

Künstner A, Wolf JBW, Backström N et al. (2010) Comparative genomics based on massive parallel transcriptome sequencing reveals patterns of substitution and selection across 10 bird species. *Molecular ecology*, **19 Suppl 1**, 266–76.

Lander ES, Schork NJ (1994) Genetic dissection of complex traits. *Science*, **265**, 2037–48.

Langfelder P, Horvath S (2008) WGCNA: an R package for weighted correlation network analysis. *BMC bioinformatics*, **9**, 559.

Lercher MJ, Chamary J-V, Hurst LD (2004) Genomic regionality in rates of evolution is not explained by clustering of genes of comparable expression profile. *Genome research*, **14**, 1002–13.

Li W, Godzik A (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**, 1658–9.

Li H, Handsaker B, Wysoker A et al. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–9.

Li B, Ruotti V, Stewart RM, Thomson JA, Dewey CN (2010) RNA-Seq gene expression estimation with read mapping uncertainty. Bioinformatics, **26**, 493–500.

Lifjeld JT, Laskemoen T, Kleven O, Albrecht T, Robertson RJ (2010) Sperm length variation as a predictor of extrapair paternity in passerine birds (T Tregenza, Ed,). *PLoS One*, **5**, e13456.

Lightowlers RN, Chrzanowska-Lightowlers ZMA (2008) PPR (pentatricopeptide repeat) proteins in mammals: important aids to mitochondrial gene expression. *The Biochemical journal*, **416**, e5–6.

Liker A, Szkely T (2005) Mortality costs of sexual selection and parental care in natural populations of birds. *Evolution*, **59**, 890–897.

Lin X, Zhang J, Li Y et al. (2011) Functional genomics of a living fossil tree, *Ginkgo*, based on next-generation sequencing technology. *Physiologia plantarum*, **143**, 207–18.

Line JE, Hiett KL, Guard-Bouldin J, Seal BS (2010) Differential carbon source utilization by *Campylobacter jejuni* 11168 in response to growth temperature variation. *Journal of microbiological methods*, **80**, 198–202.

Lopez CD, Martinovsky G, Naumovski L (2002) Inhibition of cell death by ribosomal protein L35a. *Cancer letters*, **180**, 195–202.

Louissaint A, Rao S, Leventhal C, Goldman SA (2002) Coordinated interaction of neurogenesis and angiogenesis in the adult songbird brain. *Neuron*, **34**, 945–60.

Lowder B V, Guinane CM, Ben Zakour NL et al. (2009) Recent human-to-poultry host jump, adaptation, and pandemic spread of *Staphylococcus aureus*. *Proceedings of the National Academy of Sciences of the United States of America*, **106**, 19545–50.

Löytynoja A, Goldman N (2008) Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. *Science*, **320**, 1632–5.

Lu B, Zeng Z, Shi T (2013) Comparative study of *de novo* assembly and genome-guided assembly strategies for transcriptome reconstruction based on RNA-Seq. *Science China Life sciences*, **56**, 143–55.

Lührig S, Siamishi I, Tesmer-Wolf M et al. (2014) Lrrc34, a novel nucleolar protein, interacts with Npm1 and Ncl and has an impact on pluripotent stem cells. *Stem cells and development*, **00**, 1–13.

Machado HE, Pollen AA, Hofmann HA, Renn SCP (2009) Interspecific profiling of gene expression informed by comparative genomic hybridization: A review and a novel approach in African cichlid fishes. *Integrative and comparative biology*, **49**, 644–59.

Mackay TFC, Stone EA, Ayroles JF (2009) The genetics of quantitative traits: challenges and prospects. *Nature reviews Genetics*, **10**, 565–77.

Mandlik A, Livny J, Robins WP et al. (2011) RNA-Seq-based monitoring of infection-linked changes in *Vibrio cholerae* gene expression. *Cell host & microbe*, **10**, 165–74.

Mank JE, Ellegren H (2009) All dosage compensation is local: gene-by-gene regulation of sex-biased expression on the chicken Z chromosome. *Heredity*, **102**, 312–20.

Mardis ER (2013) Next-generation sequencing platforms. *Annual review of analytical chemistry*, **6**, 287–303.

Martin JA, Wang Z (2011) Next-generation transcriptome assembly. *Nature reviews Genetics*, **12**, 671–82.

Maruska KP, Levavi-Sivan B, Biran J, Fernald RD (2011) Plasticity of the reproductive axis caused by social status change in an african cichlid fish: I. Pituitary gonadotropins. *Endocrinology*, **152**, 281–90.

Marz M, Kirsten T, Stadler PF (2008) Evolution of spliceosomal snRNA genes in metazoan animals. *Journal of molecular evolution*, **67**, 594–607.

Massey DS (2002) A brief history of human society: the origin and role of emotion in social life. *American Sociological Review*, **67**, 1–29.

McGraw L, Szekely T, Young LJ (2010) Pair bonds and parental behaviour. In: *Social behaviour: genes, ecology and evolution*. (eds Szekely T, Moore A, Komdeur J), pp. 271–301. Cambridge University Press.

McGraw LA, Young LJ (2010) The prairie vole: an emerging model organism for understanding the social brain. *Trends in neurosciences*, **33**, 103–9.

Mello C V, Vicario DS, Clayton DF (1992) Song presentation induces gene expression in the songbird forebrain. *Proceedings of the National Academy of Sciences of the United States of America*, **89**, 6818–22.

Metzker ML (2010) Sequencing technologies - the next generation. *Nature reviews Genetics*, **11**, 31–46.

Meyer UA (1996) Overview of enzymes of drug metabolism. *Journal of Pharmacokinetics and Biopharmaceutics*, **24**, 449–59.

Minato Y, Fassio SR, Kirkwood JS et al. (2014) Roles of the sodium-translocating NADH:quinone oxidoreductase (Na+-NQR) on vibrio cholerae metabolism, motility and osmotic stress resistance. *PloS one*, **9**, e97083.

Minoshima Y, Hori T, Okada M et al. (2005) The constitutive centromere component CENP-50 is required for recovery from spindle damage. *Molecular and cellular biology*, **25**, 10315–28.

Moghadam HK, Harrison PW, Zachar G, Székely T, Mank JE (2013) The plover neurotranscriptome assembly: transcriptomic analysis in an ecological model species without a reference genome. *Molecular ecology resources*, **13**, 696–705.

Møller AP, Briskie J V. (1995) Extra-pair paternity, sperm competition and the evolution of testis size in birds. *Behavioral Ecology and Sociobiology*, **36**, 357–65.

Møller AP, Ninni MP (1998) Sperm competition and sexual selection : a meta-analysis of paternity studies in birds. *Behvaioural Ecology and Sociobiology*, **43**, 345–358.

Mortazavi A, Williams BA, Mccue K, Schaeffer L, Wold B (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*, **5**, 621–8.

Mostertz J, Scharf C, Hecker M, Homuth G (2004) Transcriptome and proteome analysis of *Bacillus subtilis* gene expression in response to superoxide and peroxide stress. *Microbiology*, **150**, 497–512.

Mougeot F, Martínez-Padilla J, Blount JD et al. (2010) Oxidative stress and the effect of parasites on a carotenoid-based ornament. *Journal of experimental biology*, **213**, 400–7.

Newman SW (1999) The medial extended amygdala in male reproductive behavior. A node in the mammalian social behavior network. *Annals of the New York Academy of Sciences*, **877**, 242–57.

Nicolas P, Mäder U, Dervyn E et al. (2012) Condition-dependent transcriptome reveals high-level regulatory architecture in *Bacillus subtilis*. *Science*, **335**, 1103–6.

Ning Z, Cox AJ, Mullikin JC (2001) SSAHA: a fast search method for large DNA databases. *Genome research*, **11**, 1725–9.

O'Connell LA, Hofmann HA (2011) Genes, hormones, and circuits: An integrative approach to study the evolution of social behavior. *Frontiers in neuroendocrinology*, **32**, 320–35.

O'Connell LA, Hofmann HA (2012a) Evolution of a vertebrate social decision-making network. *Science*, **336**, 1154–7.

O'Connell LA, Hofmann HA (2012b) SUPPL Evolution of a vertebrate social decision-making network. *Science*, **336**, 1154–7.

Oldfield RG, Hofmann HA (2010) Neuropeptide regulation of social behavior in a monogamous cichlid fish. *Physiology & behavior*, **102**, 296–303.

Ophir AG, Gessel A, Zheng D-J, Phelps SM (2012) Oxytocin receptor density is associated with male mating tactics and social monogamy. *Hormones and behavior*, **61**, 445–53.

Orchinik M, Licht P, Crews D (1988) Plasma steroid concentrations change in response to sexual behavior in *Bufo marinus*. *Hormones and behavior*, **22**, 338–50.

Oshlack A, Wakefield MJ (2009) Transcript length bias in RNA-seq data confounds systems biology. *Biology direct*, **4**, 14-24.

Ozsolak F, Milos PM (2011) RNA sequencing: advances, challenges and opportunities. *Nature reviews Genetics*, **12**, 87–98.

Park PJ (2009) ChIP-seq: advantages and challenges of a maturing technology. *Nature reviews Genetics*, **10**, 669–80.

Parsons KJ, Albertson RC (2013) Unifying and generalizing the two strands of evo-devo. *Trends in ecology & evolution*, **28**, 584–91.

Paşaniuc B, Zaitlen N, Halperin E (2011) Accurate estimation of expression levels of homologous genes in RNA-seq experiments. *Journal of computational biology*, **18**, 459–68.

Pathak RR, Davé V (2014) Integrating omics technologies to study pulmonary physiology and pathology at the systems level. *Cellular physiology and biochemistry*, **33**, 1239–60.

Pearson W (2000) Flexible sequence similarity searching with the FASTA3 program package. In: *Methods in molecular biology* (Clifton, NJ), pp. 132:185–219.

Peel MM, Alfredson DA, Gerrard JG et al. (1999) Isolation , identification , and molecular characterization of strains of *Photorhabdus luminescens* from infected humans in Australia. *Journal of clinical microbiology*, **37**, 3647-53.

Petrie M (1983) Female moorhens compete for small fat males. *Science*, **220**, 413–15.

Picard M, McEwen BS (2014) Mitochondria impact brain function and cognition. *Proceedings of the National Academy of Sciences of the United States of America*, **111**, 7–8.

Pietiäinen M, François P, Hyyryläinen H-L et al. (2009) Transcriptome analysis of the responses of *Staphylococcus aureus* to antimicrobial peptides and characterization of the roles of vraDE and vraSR in antimicrobial resistance. *BMC genomics*, **10**, 429.

Pinto AC, Melo-Barbosa HP, Miyoshi A, Silva A, Azevedo V (2011) Application of RNA-seq to reveal the transcript profile in bacteria. *Genetics and molecular research*, **10**, 1707–18.

Plichta KL, Joyce SA, Clarke D, Waterfield N, Stock SP (2009) *Heterorhabditis gerrardi* n. sp. (*Nematoda: Heterorhabditidae*): the hidden host of *Photorhabdus asymbiotica* (*Enterobacteriaceae*: *gamma-Proteobacteria*). *Journal of helminthology*, **83**, 309–20.

Pointer MA, Harrison PW, Wright AE, Mank JE (2013) Masculinization of gene expression is associated with exaggeration of male sexual dimorphism. *PLoS Genetics*, **9**, 1-9.

Le Quéré A, Eriksen KA, Rajashekar B et al. (2006) Screening for rapidly evolving genes in the ectomycorrhizal fungus *Paxillus involutus* using cDNA microarrays. *Molecular ecology*, **15**, 535–50.

Rackham O, Filipovska A (2011) The role of mammalian PPR domain proteins in the regulation of mitochondrial gene expression. *Biochimica et biophysica acta*, **1819**, 1008–16.

Rackham O, Shearwood A-MJ, Mercer TR et al. (2011) Long noncoding RNAs are generated from the mitochondrial genome and regulated by nuclear-encoded proteins. *RNA*, **17**, 2085–93.

Reedy AM, Edwards A, Pendlebury C et al. (2014) An acute increase in the stress hormone corticosterone is associated with mating behavior in both male and female red-spotted newts, *Notophthalmus viridescens*. *General and comparative endocrinology*, **208**, 57–63.

Rehsteiner U, Geisser H, Reyer H (1998) Singing and mating success in water pipits: one specific song element makes all the difference. *Animal behaviour*, **55**, 1471–81.

Reiner A, Perkel DJ, Bruce LL et al. (2004) Revised nomenclature for avian telencephalon and some related brainstem nuclei. *Journal of comparative neurology*, **473**, 377–414.

Ren X, Liu T, Dong J et al. (2012) Evaluating de Bruijn Graph assemblers on 454 transcriptomic data. *PLoS ONE*, **7**, e51188.

Renn SCP, Aubin-Horth N, Hofmann HA (2004) Biologically meaningful expression profiling across species using heterologous hybridization to a cDNA microarray. *BMC genomics*, **5**, 42-55.

Renn SCP, Aubin-Horth N, Hofmann HA (2008) Fish and chips: functional genomics of social plasticity in an African cichlid fish. *Journal of experimental biology*, **211**, 3041–56.

Renn SCP, Machado HE, Jones A et al. (2010) Using comparative genomic hybridization to survey genomic sequence divergence across species: a proof-of-concept from *Drosophila*. *BMC Genomics*, **11**, 271-83.

Reyer H-U, Bollmann K, Schläpfer AR, Schymainda A, Klecack G (1997) Ecological determinants of extrapair fertilizations and egg dumping in Alpine water pipits (*Anthus spinoletta*). *Behavioral Ecology*, **8**, 534–43.

Richards S, Liu Y, Bettencourt BR et al. (2005) Comparative genome sequencing of *Drosophila pseudoobscura*: chromosomal, gene, and cis-element evolution. *Genome research*, **15**, 1–18.

Rismani-Yazdi H, Haznedaroglu BZ, Bibby K, Peccia J (2011) Transcriptome sequencing and annotation of the microalgae *Dunaliella tertiolecta*: pathway description and gene discovery for production of next-generation biofuels. *BMC genomics*, **12**, 148-65.

Rivals I, Personnaz L, Taing L, Potier M-C (2007) Enrichment or depletion of a GO category within a class of genes: which test? *Bioinformatics*, **23**, 401–7.

Robinson GE, Fernald RD, Clayton DF (2008) Genes and social behavior. *Science*, **322**, 896–900.

Robinson GE, Grozinger CM, Whitfield CW (2005) Sociogenomics: social life in molecular terms. *Nature reviews Genetics*, **6**, 257–70.

Roepstorff P (2012) Mass spectrometry based proteomics, background, status and future needs. *Protein & cell*, **3**, 641–7.

Ross HE, Freeman SM, Spiegel LL et al. (2009) Variation in oxytocin receptor density in the nucleus accumbens has differential effects on affiliative behaviors in monogamous and polygamous voles. *Journal of Neuroscience*, **29**, 1312–8.

Rosvall KA (2011) Intrasexual competition in females: evidence for sexual selection? *Behavioral ecology*, **22**, 1131–40.

Le Rudulier D, Bouillard L (1983) Glycine betaine , an osmotic effector in *Klebsiella pneumoniae* and other members of the *Enterobacteriaceae*. *Applied and Environmental Microbiology*, **46**, 152–9.

Rumble SM, Lacroute P, Dalca A V et al. (2009) SHRiMP: accurate mapping of short color-space reads. *PLoS computational biology*, **5**, e1000386.

Sala M, Braida D, Lentini D et al. (2011) Pharmacologic rescue of impaired cognitive flexibility, social deficits, increased aggression, and seizure susceptibility in oxytocin receptor null mice: a neurobehavioral model of autism. *Biological psychiatry*, **69**, 875–82.

Sander JD, Joung JK (2014) CRISPR-Cas systems for editing, regulating and targeting genomes. *Nature biotechnology*, **32**, 347–55.

Sarkar D (2008) Lattice: Multivariate Data Visualization with R. Springer, New York.

Von Schantz T, Bensch S, Grahn M, Hasselquist D, Wittzell H (1999) Good genes, oxidative stress and condition-dependent sexual signals. *Proceedings of The Royal Society Biological sciences*, **266**, 1–12.

Schuller C, Mamnun YM, Mollapour M et al. (2004) Global phenotypic analysis and transcriptional profiling defines the weak acid stress response regulon in *Saccharomyces cerevisiae*. *Molecular Biology of the Cell*, **15**, 706–20.

Schulz MH, Zerbino DR, Vingron M, Birney E (2012) Oases: robust *de novo* RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics*, **28**, 1086–92.

Schunter C, Vollmer S V, Macpherson E, Pascual M (2014) Transcriptome analyses and differential gene expression in a non-model fish species with alternative mating tactics. *BMC genomics*, **15**, 167-80.

Schut E, Magrath MJL, Oers K Van, Komdeur J (2012) Volume of the cloacal protuberance as an indication of reproductive state in male blue tits *Cyanistes caeruleus*. *Ardea*, **100**, 202–5.

Shahrokh DK, Zhang T-Y, Diorio J, Gratton A, Meaney MJ (2010) Oxytocin-dopamine interactions mediate variations in maternal behavior in the rat. *Endocrinology*, **151**, 2276–86.

Sharp PM, Hahn BH (2011) Origins of HIV and the AIDS pandemic. *Cold Spring Harbor perspectives in medicine*, **1**, a006841.

Shi C-Y, Yang H, Wei C-L et al. (2011) Deep sequencing of the *Camellia sinensis* transcriptome revealed candidate genes for major metabolic pathways of tea-specific compounds. *BMC genomics*, **12**, 131-50.

Simpkins JW, Yi KD, Yang S-H, Dykens JA (2010) Mitochondrial mechanisms of estrogen neuroprotection. *Biochimica et biophysica acta*, **1800**, 1113–20.

Sims D, Sudbery I, Ilott NE, Heger A, Ponting CP (2014) Sequencing depth and coverage: key considerations in genomic analyses. *Nature reviews Genetics*, **15**, 121–32.

Snow DW, Snow BK (1983) Territorial song of the Dunnock *Prunella modularis*. *Bird Study*, **30**, 51–6.

Sol D, Székely T, Liker A, Lefebvre L (2007) Big-brained birds survive better in nature. *Proceedings. Of The Royal Society Biological sciences*, **274**, 763–9.

Stefanski V, Engler H (1999) Social stress , dominance and blood cellular immunity. *Journal of neuroimmunology*, **94**, 144–52.

Stevenson B, Schwan TG, Rosa PA (1995) Temperature-related differential expression of antigens in the Lyme disease spirochete, *Borrelia burgdorferi*. *Infection and Immunity*, **63**, 4535-39.

Sugita S, Ho A, Südhof TC (2002) NECABs: a family of neuronal Ca(2+)-binding proteins with an unusual domain structure and a restricted expression pattern. *Neuroscience*, **112**, 51–63.

Suzuki H, Arakawa Y, Ito M et al. (2007) MLF1-interacting protein is mainly localized in nucleolus through N-terminal bipartite nuclear localization signal. *Anticancer research*, **27**, 1423–30.

Szaszák M, Shima K, Käding N et al. (2013) Host metabolism promotes growth of *Chlamydia pneumoniae* in a low oxygen environment. *International journal of medical microbiology*, **303**, 239–46.

Székely T, Freckleton RP, Reynolds JD (2004) Sexual selection explains Rensch's rule of size dimorphism in shorebirds. *Proceedings of the National Academy of Sciences of the United States of America*, **101**, 12224–7.

Székely T, Lislevand T, Figuerola J (2007) Sexual size dimorphism in birds. In: *Sex, size, and gender roles: evolutionary studies of sexual size dimorphism* (eds Fairbairn D, Blanckenhorn W, Székely T), pp. 27–37. Oxford University Press.

Székely T, Moore AJ, Komdeur J (2010) Social Behaviour: Genes, Ecology and Evolution. Cambridge University Press.

Tamura K, Subramanian S, Kumar S (2004) Temporal patterns of fruit fly (*Drosophila*) evolution revealed by mutation clocks. *Molecular biology and evolution*, **21**, 36–44.

Tanner MA, Everett CL, Youvan DC (2000) Molecular phylogenetic evidence for noninvasive zoonotic transmission of *Staphylococcus intermedius* from a canine pet to a human. *Journal of clinical microbiology*, **38**, 1628–31.

Taubenberger JK, Kash JC (2010) Influenza virus evolution, host adaptation, and pandemic formation. *Cell host & microbe*, **7**, 440–51.

The Gene Ontology Consortium (2000) Gene Ontology : tool for the unification of biology. *Nature genetics*, **25**, 25–9.

The R Development Core Team (2010) R : A Language and Environment for Statistical Computing.

Toth AL, Robinson GE (2007) Evo-devo and the evolution of social behavior. *Trends in genetics*, **23**, 334–41.

Toth AL, Varala K, Newman TC et al. (2007) Wasp gene expression supports an evolutionary link between maternal behavior and eusociality. *Science*, **318**, 441–4.

Trainor BC, Sisk CL, Nelson RJ (2009) Hormones and the development and expression of aggressive behavior. In: *Hormones, Brain and Behavior*, 2nd Edition (Editors: Pfaff, D. W; Arnold, A. P; Etgen, A. M; Farhbach, S. E.; Rubin, R. T.), pp. 167–203.

Tramontin AD, Brenowitz EA (2000) Seasonal plasticity in the adult brain. *Trends in neurosciences*, **23**, 251–8.

Trapnell C, Pachter L, Salzberg SL (2009) TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, **25**, 1105–11.

Urich T, Lanzén A, Stokke R et al. (2013) Microbial community structure and functioning in marine sediments associated with diffuse hydrothermal venting assessed by integrated meta-omics. *Environmental microbiology*, **16**, 2699–710.

Vaas LAI, Sikorski J, Michael V, Göker M, Klenk H-P (2012) Visualization and curve-parameter estimation strategies for efficient exploration of phenotype microarray kinetics. *PloS one*, **7**, e34846.

Vijay N, Poelstra JW, Künstner A, Wolf JBW (2013) Challenges and strategies in transcriptome assembly and differential gene expression quantification. A comprehensive in silico assessment of RNA-seq experiments. *Molecular ecology*, **22**, 620–34.

Völker M, Backström N, Skinner BM et al. (2010) Copy number variation, chromosome rearrangement, and their association with recombination during avian evolution. *Genome research*, **20**, 503–11.

Walum H, Westberg L, Henningsson S et al. (2008) Genetic variation in the vasopressin receptor 1a gene (AVPR1A) associates with pair-bonding behavior in humans.

Proceedings of the National Academy of Sciences of the United States of America, **105**, 14153–6.

Wang Z, Gerstein M, Snyder M (2009a) RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews Genetics*, **10**, 57–63.

Wang W, Wang Y, Zhang Q, Qi Y, Guo D (2009b) Global characterization of *Artemisia* annua glandular trichome transcriptome using 454 pyrosequencing. *BMC genomics*, **10**, 465–75.

Warren WC, Clayton DF, Ellegren H et al. (2010) The genome of a songbird. *Nature*, **464**, 757–62.

Waterfield NR, Ciche T, Clarke D (2009) *Photorhabdus* and a host of hosts. *Annual review of microbiology*, **63**, 557–74.

Westermann AJ, Gorski SA, Vogel J (2012) Dual RNA-seq of pathogen and host. *Nature reviews Microbiology*, **10**, 618–30.

Westneat DF, Sherman PW, Morton ML (1990) The ecology and evolution of extra-pair copulations in birds. *Current ornithology*, **7**, 331–69.

Wheat CW (2010) Rapidly developing functional genomics in ecological model systems via 454 transcriptome sequencing. *Genetica*, **138**, 433–51.

White SA, Nguyen T, Fernald RD (2002) Social regulation of gonadotropin-releasing hormone. *Journal of experimental biology*, **205**, 2567–81.

Wilhelm BT, Landry J-R (2009) RNA-Seq-quantitative measurement of expression through massively parallel RNA-sequencing. *Methods*, **48**, 249–57.

Wilkinson P, Waterfield NR, Crossman L et al. (2009) Comparative genomics of the emerging human pathogen *Photorhabdus asymbiotica* with the insect pathogen *Photorhabdus luminescens*. *BMC genomics*, **10**, 302–324.

Wolf JBW, Bayer T, Haubold B et al. (2010) Nucleotide divergence vs. gene expression differentiation: comparative transcriptome sequencing in natural isolates from the carrion crow and its hybrid zone with the hooded crow. *Molecular ecology*, **19 Suppl 1**, 162–75.

Wolfson A (1952) The Cloacal Protuberance : A means for determining breeding condition in live male passerines. *Bird-banding*, **23**, 159–65.

Wu H, Li D, Shan Y et al. (2007) EFCBP1/NECAB1, a brain-specifically expressed gene with highest abundance in temporal lobe, encodes a protein containing EF-hand and antibiotic biosynthesis monooxygenase domains. *DNA sequence*, **18**, 73–9.

Xu F, Addis JBL, Cameron JM, Robinson BH (2012) LRPPRC mutation suppresses cytochrome oxidase activity by altering mitochondrial RNA transcript stability in a mouse model. *Biochemical Journal*, **441**, 275–83.

Yan Q, Power KA, Cooney S et al. (2013) Complete genome sequence and phenotype microarray analysis of *Cronobacter sakazakii* SP291: a persistent isolate cultured from a powdered infant formula production facility. *Frontiers in microbiology*, **4**, 1-20.

Yang Z (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. *Bioinformatics*, **13**, 555–6.

Yang Z, Bielawski JP (2000) Statistical methods for detecting molecular adaptation. *Trends in Ecology & Evolution*, **15**, 496–503.

Yang JY, Karr JR, Watrous JD, Dorrestein PC (2011) Integrating "-omics" and natural product discovery platforms to investigate metabolic exchange in microbiomes. *Current opinion in chemical biology*, **15**, 79–87.

Young LJ, Nilsen R, Waymire KG, MacGregor GR, Insel TR (1999) Increased affiliative response to vasopressin in mice expressing the V1a receptor from a monogamous vole. *Nature*, **400**, 766–8.

Young LJ, Wang Z (2004) The neurobiology of pair bonding. *Nature neuroscience*, **7**, 1048–54.

Zalocusky K, Deisseroth K (2013) Optogenetics in the behaving rat: integration of diverse new technologies in a vital animal model. *Optogenetics*, **1**, 1–17.

Zerbino D (2010) Using the Columbus extension to Velvet. 1–5.

Zerbino DR, Birney E (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome research*, **18**, 821–9.

Zhang G, Jarvis ED, Gilbert MTP (2014) A flock of genomes. *Science*, **346**, 1308–9.

Zhang L, Li W-H (2004) Mammalian housekeeping genes evolve more slowly than tissue-specific genes. *Molecular biology and evolution*, **21**, 236–9.

Zhu BT, Lee AJ (2005) NADPH-dependent metabolism of 17beta-estradiol and estrone to polar and nonpolar metabolites by human tissues and cytochrome P450 isoforms. *Steroids*, **70**, 225–44.