Genomic resources and genetic studies of parasitic flukes, with an emphasis on *Clonorchis sinensis*

Daxi Wang

ORCID ID: 0000-0002-0979-1924

Submitted in fulfilment of the requirements of the degree of Doctor of Philosophy

March 2019

The Faculty of Veterinary and Agricultural Sciences

The University of Melbourne

SUMMARY

Clonorchiasis is a complex hepatobiliary disease caused by the foodborne parasite, *Clonorchis sinensis* (family Opisthorchiidae). This disease can induce cholangiocarcinoma (CCA), a malignant cancer of the bile ducts, and has a major socioeconomic impact on ~ 35 million people predominantly in East Asia. Currently, no vaccine is available to prevent clonorchiasis, and repeated use of the only recommended drug, praziquantel (PZQ) increases the risk of developing drug resistance. Further understanding of the disease epidemiology relies on the knowledge of genetic variation of *C. sinensis* in endemic areas. Moreover, evidence of karyotypic variation within *C. sinensis* highlights the importance of comparing the genomes of geographically distinct isolates of this parasite.

The two predominant research aims of this thesis were to decode the mitochondrial (mt) and nuclear genomes of a Korean isolate of *C. sinensis* and assess genetic variation, using high-throughput sequencing technologies and advanced bioinformatic workflows. The mt and nuclear genomes for a *C. sinensis* isolate from Korea (*Cs*-k2) were sequenced, assembled, characterised and compared with one or more isolates. In addition, a refined bioinformatic workflow was designed to infer high quality syntenic blocks between the nuclear genome and a previously published draft genome of another isolate from China. The results not only reveal a high level of nucleotide similarity within the syntenic regions, but also pinpoint variable genes that might be central to infection and/or adaptive process.

The mt and nuclear genomes and the syntenic blocks now serve as a solid foundation for a future genetic analysis of *C. sinensis*. The mt genome, on one hand, confirmed the specific identity of the specimen, on the other hand, highlighted potential challenges with using mtDNA markers for genetic analyses. In contrast to the mt genome, the syntenic blocks of the nuclear genome exhibit major potential for future genetic studies due to a vast extent of nucleotide differences in coding regions. These blocks also contain a substantial number of genetic loci that might enhance knowledge of host-parasite relationships in an evolutionary context. Compared with coding regions, the genetic variation in the intronic regions showed an improved phylogenetic signal at both the whole genome and individual gene levels.

In future, improved quality of the assembly and annotation of nuclear genomes should be achieved using long read data, allowing a broader range of genetic and structural variation to be identified using whole genomic data of representative individuals. Furthermore, a systematic bioinformatic framework is required to discover individual variants, infer population structure and identify adaptive selection, with the consideration of parasitic life cycle and demographic history. Although the present thesis focused predominantly on *C. sinensis*, the work extended logically to another trematode. A third research aim was addressed to explore the population genetic structure of a related trematode parasite, *Schistosoma japonicum* in China and to identify genes under positive selection in particular geographical locations. Clearly, the findings of this thesis and the approaches established should have important and broad implications for studies of a range of flatworm parasites.

DECLARATION

The work described in the thesis was performed in the Faculty of Veterinary and Agricultural Sciences of The University of Melbourne between February 2016 and March 2019. The scientific work was performed by the author, with the exception of the assistance which has been specifically acknowledged. The thesis is less than 100,000 words in length, exclusive of tables, figures, references and appendices. No part of this thesis has been submitted for any other degree or diploma.

……………………

Daxi Wang

15 March 2019

ACKNOWLEDGEMENTS

I wish to express my deep sense of gratitude to my principal supervisor, Dr Neil D. Young, for his great patience, insightful suggestions and sustained support during my research project. Neil spent a lot of time providing revisions and comments during the drafting of my thesis and guiding me to develop scientific thinking and method. I also wish to express great appreciation to my co-supervisor, Professor Robin B. Gasser, for his valuable suggestions and insightful discussion during the research. Robin has impacted me a lot by his great passion and rigorous attitude to parasitology and research. In addition, I wish to sincerely thank my co-supervisor, Dr Pasi K. Korhonen, for his guidance and great help with bioinformatics. By working with Pasi, I gained experience and skills in data analysis. Without the help and guidance from my supervisors, it would not have been possible for my thesis to be completed.

I wish to thank the other members of the parasitology laboratory, particularly Dr Anson V. Kohler, for his guidance and help in my research, and Mr Ross S. Hall for computational support. Thanks also to my friends, including Yaqing Jiao, Guangxu Ma and Yan Zhang, for their support. I also wish to acknowledge the help provided by Dr Woon-Mok Sohn (Department of Parasitology and Tropical Medicine, and Institute of Health Sciences, Gyeongsang National University, School of Medicine, Jinju, Korea) who provided the parasite samples of *Clonorchis sinensis* (Chapters 2 and 3). I am also grateful for the Melbourne Research Scholarships (MRS) from the University of Melbourne. Finally, I sincerely thank my wife Xiaoli Yang for her unconditional love and support during the time that I worked on this thesis.

PUBLICATIONS

Wang, D.*, Young, N.D.*, Korhonen, P.K., Gasser, R.B., 2018. *Clonorchis sinensis* and clonorchiasis: the relevance of exploring genetic variation. Adv. Parasitol. 100, 155-208 (Chapter 1). * First authors.

Wang, D., Young, N.D., Koehler, A.V., Tan, P., Sohn, W.M., Korhonen, P.K., Gasser, R.B., 2017. Mitochondrial genomic comparison of *Clonorchis sinensis* from South Korea with other isolates of this species. Infect. Genet. Evol. 51, 160-166 (Chapter 2).

Wang, D., Korhonen, P.K., Gasser, R.B., Young, N.D., 2018. Improved genomic resources and new bioinformatic workflow for the carcinogenic parasite *Clonorchis sinensis*: Biotechnological implications. Biotechnol. Adv. 36, 894-904. (Chapter 3).

Wang, D.*, Zhao, Q.P.*, Emery, A.M., Nie, P., Rollinson, D., Li, Y.W., Allan, F., Gasser, R.B., Webster, B.L., Korhonen, P.K., Dong, H.F., Young, N.D., 2019. *Schistosoma japonicum*: improving our understanding of population genetic variation in China using an expanded genomic data set. (Chapter 4). * First authors.

ORAL PRESENTATIONS

Wang, D., 2018. Exploring genetic variation within *Clonorchis sinensis*. *Faculty of Veterinary and Agricultural Sciences, The University of Melbourne*, 11 April 2018 (PhD Confirmation Seminar).

Wang, D., Korhonen, P.K., Gasser, R.B., Young, N.D., 2018. Improved genomic resources for the carcinogenic parasite *Clonorchis sinensis* achieved using a new bioinformatic workflow. Proceedings of 2018 Australian Society for Parasitology Conference, Melbourne, Australia, 25 September 2018.

Wang, D., 2018. Genomes of *Clonorchis sinensis* – resources, challenges and implications for population genetic studies of flukes. *Faculty of Veterinary and Agricultural Sciences, The University of Melbourne*, 07 March 2019 (PhD Completion Seminar).

TABLE OF CONTENTS

Chapter 1 - Literature review

1.1. Introduction

Clonorchis sinensis, *Opisthorchis felineus* and *Opisthorchis viverrini* (phylum Platyhelminthes; class Trematoda; family Opisthorchiidae) are important foodborne liver flukes that infect ~ 24 million people worldwide and cause a global burden of more than 349,737 disability-adjusted life years (DALYs) (Furst et al., 2012). As a widespread liver fluke, *C. sinensis* infects at least 15 million people predominantly in China, Vietnam, Korea and the Russian Far East (Furst et al., 2012; Qian et al., 2016).

Clonorchis sinensis infection causes clonorchiasis, which is a neglected tropical disease (NTDs) (Qian et al., 2016). Infection with *C. sinensis* often leads to chronic hepatobiliary diseases, such as hepatic fibrosis, and can induce cholangiocarcinoma (CCA), a malignant cancer of the biliary system. Hence, *C. sinensis* has been classified as a Class I carcinogen by the International Agency for Research on Cancer (IARC) (Choi et al., 2004, 2011; Grosse et al., 2009;).

To date, no vaccine is available to prevent clonorchiasis, and humans have no resistance to reinfection (Qian et al., 2016). The repeated use of the only recommended drug, praziquantel (PZQ), increases the risk of the fluke developing drug resistance (WHO, 2013). To better control clonorchiasis, research has been conducted on the biology and epidemiology of the parasite, and focused on diagnosis and treatment. A deep understanding of the molecular biology of *C. sinensis* is now possible, underpinned by the use of new transcriptomic and genomic resources produced using high-throughput sequencing technologies (Young et al., 2010; Wang et al., 2011; Yoo et al., 2011; Huang et al., 2013). For example, the genome of *C. sinensis* from China provides a basis to explore molecular pathways and processes in this liver fluke. For example, the genome of *C. sinensis* encodes a full complement of genes required to metabolise host lipids but does not contain genes linked to canonical fatty acid biosynthesis (Wang et al., 2011; Huang et al., 2013). In addition, this genome encodes a large number of excretory-secretory proteins (ESPs), many of which are inferred to be involved in host-parasite interactions. However, there is no detailed information on the functions of biologically relevant genes/gene products or levels of genetic variation among different geographical isolates of *C. sinensis*.

It has been assumed that a single nuclear genome of *C. sinensis* is sufficient to represent all geographically distinct isolates of this species (Wang et al., 2011; Huang et al., 2013). However, evidence of karyotypic variation within *C. sinensis* and cryptic species within the related opisthorchiid fluke, *O. viverrini*, highlight the importance of comparing the genomes of *C. sinensis* from disparate geographical regions. Previous studies of selected genetic loci suggest that variation in this parasite has been influenced by multiple factors, such as life cycle, climatic change and environmental adaptation (Tatonova et al., 2012; Tatonova et al., 2013; Chelomina et al., 2014). For example, asexual reproduction enables genetically identical or similar organisms to aggregate in the same host, which can increase inbreeding and decrease genetic diversity (Glémin and Galtier, 2012; Vilas et al., 2012). Furthermore, it appears that low nucleotide and high haplotypic variation in the mitochondrial *cox*1 gene within *C. sinensis* relates to a rapid population expansion after the ice age (Tatonova et al., 2012; Chelomina et al., 2014). However, such a conclusion requires in depth population genetic investigations.

It is now feasible to undertake such investigations on a whole-genome level using high-throughput sequencing technologies and advanced bioinformatics. Such research would enable explorations of the genetic structures and substructures of *C. sinensis* populations, and of variation in genes/gene products that are involved in infection and/or adaptive processes. The purpose of this chapter is to provide a background on the biology, epidemiology, pathogenesis, diagnosis and control of *C. sinensis*/clonorchiasis; to critically review current knowledge of the genetics, molecular biology and genomics of *C. sinensis*; and highlight new sequencing technologies as well as genomic tools and resources that can now be employed for future genetic studies of this and other important trematodes.

1.2. Biology

Clonorchis sinensis is a leaf-shaped fluke that is about 10-25 mm long and 3 mm wide. It has a complex life cycle (Fig. 1-1) involving asexual reproduction in an intermediate host (freshwater snail), encystment within a secondary intermediate host (freshwater fish) and sexual reproduction in a definitive, piscivorous mammal (e.g., dog, cat and human) (Kaewkes, 2003). The ovoid eggs of *C. sinensis* are yellowish, 26-30 µm long and 15-17 µm wide (Lun et al., 2005). Each egg, which is operculated at one end, contains a well-formed miracidium (Lun et al., 2005). After being released from adult flukes in the definitive host, the eggs, which can be infectious for 2 or 3 weeks (Echaubard et al., 2016), are eaten by an intermediate freshwater snail host. In the digestive tract of the snail, the eggs release miracidia, which then penetrate the intestinal wall and enter the snail's haemocoel and digestive gland (Bogitsh et al., 2013). Within ~ 4 h, miracidia transform into sporocysts that begin developing to rediae within germ balls *via* asexual reproduction (Lun et al., 2005). After ~ 17 days, each sporocyst produces 20-50 rediae. Within germ balls in the brood chambers, 5-50 cercariae will differentiate within 21 to 30 days (Lun et al., 2005). These cercariae leave the snail host, enter the water and actively swim off to seek a cyprinid fish, the second intermediate host. The cercariae are able to stay active in the water for 24 to 48 h (Bogitsh et al., 2013). After finding and attaching to the fish, the cercariae invade the dermis and the musculature beneath it, lose their eyespots and tails, and encyst there as metacercariae. These metacercariae mature within 5-6 weeks (Bogitsh et al., 2013) and can remain viable in tissues for more than a year (Cheng et al., 2007). When humans or other fish-eating mammals eat uncooked infected fish, encysted metacercariae excyst in the small intestine and then migrate *via* the ampulla of Vater to the biliary ducts. Within the biliary tree, juvenile flukes continue to develop, and mature into hermaphroditic adults within the intrahepatic biliary tree within ~ 4 weeks (Lun et al., 2005). Adult flukes have been reported to survive within the biliary tree for decades (Attwood and Chou, 1978). Fertilised eggs pass through the bile duct to the small intestine and are released in the faeces, to start a new life cycle (Lun et al., 2005). Adult flukes can lay 1,000 to 4,000 eggs per day (Lun et al., 2005).

First intermediate hosts

The geographical range and distribution of the snail intermediate host(s) determine the distribution of *C. sinensis*. Compared with *Opisthorchis* spp., which infect only bithyniid snails, *C. sinensis* infects at least 10 distinct snail species of four families (Bithyniidae, Semisulcospiridae, Assimineidae and Thiaridae) (Table 1-1; Fig. 1-2; cf. Petney et al., 2013). *Parafossarulus manchouricus* is the most widespread intermediate snail host, inhabiting all known endemic countries, including Russia, Japan, Korea, China and Vietnam (Tang et al., 2016). Besides *P. manchouricus*, the distribution of other snail hosts is variable, which is likely linked to climatic variation. For example, *P. spiridonovi* is only endemic in Russia Far East, whereas *Semisulcospira scancellata* is in Russia, North China, Central and South China (Fattakhov et al., 2012; Chelomina et al., 2014). In addition, *Melanoides tuberculata* is endemic only in South China, Vietnam and Thailand (Doanh and Nawa, 2016; Tang et al., 2016).

Paleontologic evidence suggests that *P. manchouricus* has inhabited China for more than 2.6 million years (Li et al., 2015), whereas the earliest fossils of *P. manchouricus* from Japan are about 781 thousand years old (Isaji and Ugai, 2006). The distribution of *P. manchouricus* is closely linked with climatic factors, especially temperature (⩾ 12° C) and rainfall (Chung et al., 1980). *P. manchouricus* can live for up to 16 months (Li et al., 1979) and is active in slow-flowing water, such as ponds and rice fields (Li et al., 1979; Chung et al., 1980). In winter, the snail can hibernate (Li et al., 1979).

Currently, there are numerous species of snail hosts in Central and South China. Although the prevalence of infected snail hosts is usually low, prevalence can vary considerably (Lun et al., 2005), particularly in South China. For example, the prevalence of infection in *B. longicornis* can be as high as 27% in Guangdong, and 15% in Guangxi (Lun et al., 2005). This might be explained by the climatic differences between North and South China. For example, the longer summer in South China extends the time in which the snails stay active (i.e. extends the feeding time) and increases their likelihood of ingesting *C. sinensis* eggs, which would relate to an increase in the prevalence of infection (Yang et al., 1994).

Second intermediate hosts

Clonorchis sinensis has a large range of second intermediate hosts (Lun et al., 2005; Tang et al., 2016) (Table 1-2; Fig. 1-2). In China, more than 30 fish genera belonging to 10 families and four shrimp species have been shown to be hosts of *C. sinensis* (reviewed by Tang et al., 2016). In Korea, for example, 31 fish genera representing 6 families can host metacercariae of *C. sinensis* (Sohn, 2009). In Russian Far East, three species of freshwater fish have been found to harbour this parasite (Fattakhov et al., 2012). Across these geographical regions, 71 species of cyprinid fish (family Cyprinidae) have been reported to harbour metacercariae of *C. sinensis* (see Lun et al., 2005).

Among them, *Pseudorasbora parva*, *Ctenopharyngodon idellus, Carassius auratus* and *Hypophthalmichthys nobilis* are the commonest fish hosts of *C. sinensis*, and have a high prevalence of infection in endemic areas (Lun et al., 2005; Tang et al., 2016). Due to an exponential asexual reproduction of cercariae in snails, the prevalence of *C. sinensis* in freshwater fish can be significantly higher than observed in the snail hosts (Cheng et al., 2007). In North East China, *C. sinensis* infects, on average, ~ 20% of the freshwater fishes, including *P. parva* (43%), *Hemicculter leuciclus* (23%), *C. auratus* (20%) and *Rhodeus ocellatus* (11%) (Zhang et al., 2014). In the Pearl River Delta (South China), 37% of fish and 3% of shrimps (*Caridina* spp.) examined were infected with *C. sinensis* (see Chen et al., 2010). The prevalence of *C. sinensis* recorded in *C. idellus*, *H. nobilis*, *C. auratus* and *P. parva* was 52%, 22%, 14% and 94%, respectively (Chen et al., 2010).

Of freshwater fish hosts, small fishes, such as *P. parva*, *R. ocellatus* and *H. leuciclus*, are more likely to be infected with *C. sinensis,* and have a higher infection intensity compared with large fish (Lun et al., 2005). For example, the prevalence in *P. parva*, whose body weight is generally less than 1.5 g, can be up to 93% in Guangdong, with an average of ~117 encysted metacercariae per infected fish (Chen et al., 2010). The high prevalence of *C. sinensis* in *P. parva* can be explained by the ecology of this fish and its first intermediate host(s). For example, *P. parva* and snails both prefer the habitats with slow-flowing or stagnant water, such as in rice fields (Li et al., 1979; Katano et al., 2003). Another factor contributing to the high prevalence of *C. sinensis* infection in *P. parva* is a relatively low innate resistance (Bui et al., 2016). For instance, infection experiments show that *C. sinensis* cercariae could readily invade the muscles of *P. parva* fry, whereas the cercariae attached to *C. carpio* and *C. auratus* all died on the surface of the fish (Chun, 1964). This finding was further supported by the difference in the wormicidal effect of the epidermal mucus of different fishes. For example, epidermal mucus from *C. scarassius* can kill the cercariae of *C. sinensis* within 14 min, whereas it takes 180 min in the case of *P. parva* (see Rhee et al., 1980).

1.3. Epidemiology

Clonorchis sinensis appeared in China at least 2,300 years ago, as evidenced by the discovery of *C. sinensis* eggs in the faeces and bile ducts of a human corpse from the

Western Han Dynasty (buried in 167 BC) in Jiangling, Hubei province, China (Wu et al., 1980). However, this parasite was not identified until the 19th century. In 1875, James McConnell first discovered *C. sinensis* in the bile ducts of a 20-year-old Chinese male in India (Qian et al., 2016). Later, the first clonorchiasis case in China was reported in 1908 (Pan et al., 2000). Currently*, C. sinensis* infects more than 15 million people in East Asia, including China, Vietnam, Korea, Thailand and the Russian Far East (Furst et al., 2012). In China, ~ 13 million people are infected (Furst et al., 2012; Qian et al., 2012), and the prevalence is reported to be particularly high in Guangdong (16.4%), Guangxi (9.8%), Heilongjiang (4.7%), Jilin (2.9%) and Hunan (1.3%) (Fang et al., 2008). Most infections (78.9%) are of a relatively low intensity (Fang et al., 2008). *Clonorchis sinensis* is also reported to infect one million people in South Korea (Kim et al., 2009a), the Russian Far East (3,000) (Abdussalam et al., 1995) and North Vietnam (1 million) (Qian et al., 2016). However, because *O. viverrini* is also widespread in Vietnam and the eggs of *O. viverrini* and *C. sinensis* are morphologically similar, the exact number of people infected with *C. sinensis* in Vietnam remains uncertain (Qian et al., 2016). *Clonorchis sinensis* may also co-exist with *O. viverrini* in central Thailand (Traub et al., 2009).

Aspects of relevance to epidemiology

Climate. In Guangzhou, China, temperature and rainfall correlate positively with the prevalence of clonorchiasis in humans (Li et al., 2014a). For example, in 2008, the relatively low prevalence of *C. sinensis* infection in fish in Foshan, China, related to a period of below-average winter temperatures in this geographical region (Chen et al., 2010; Zhou et al., 2011). High temperatures can significantly enhance snail activity (Ye et al., 1997; Laoprom et al., 2016) and can increase the production rate of cercariae (Poulin, 2006). Given climatic changes due to industrialisation and deforestation (Miettinen et al., 2011), there is a possibility of clonorchiasis outbreaks in the future (cf. Poulin, 2006).

Aspects relating to aquaculture. Currently, aquaculture in China is responsible for >60% of the world's fish production, with an estimated tripling in production of capture and farmed fish over the past two decades (Cao et al. 2015). However, common domesticated fishes, particularly cyprinids, are frequently infected with *C. sinensis* (see

Lun et al., 2005). In southern China, the prevalence can be up to 37% in domesticated freshwater fishes, with an average number of 11 metacercariae per fish (Chen et al., 2010). Given the high prevalence of *C. sinensis* in domesticated fish, the rapid development of aquaculture also contributes to the widespread distribution of *C. sinensis* in parts of East Asia. The total area used for fish farming has more than doubled since 1979 (Li et al., 2011). Humans in villages that are close to aquaculture systems have been reported to have a high prevalence of *C. sinensis* infection (see Keiser and Utzinger, 2005), which is further supported by a recent network analysis (Vinh et al., 2017). In these villages, toilets are usually built adjacent to or over a fish-pond; excrement containing fluke eggs is released directly into the pond water (Lun et al., 2005), after which snail hosts consume these eggs. Furthermore, infected fish can be transported to other regions. Even when frozen at -20°C, metacercariae of *C. sinensis* can remain viable and infective (Fan, 1998), which can be a public health concern in non-endemic areas.

Reservoir hosts. Clonorchis sinensis infects a wide range of piscivorous mammals, including cats, dogs and foxes, and it can also experimentally infect rabbits, rats and mice (Lun et al., 2005). Dogs and cats appear to be the commonest reservoir hosts in China, particularly in the Guangdong, Guangxi, Hubei, Hunan and Heilongjiang provinces of China (Lun et al., 2005). For instance, in Guangdong province, 20.5% of dogs and 41.8% of cats are infected with *C. sinensis* (see Lin et al., 2011). Compared with dogs, cats are reported to have higher infection intensities in this region of China (Lin et al., 2011). Especially, stray cats usually move freely on streets and, in villages, easily get access to uncooked fish from household waste.

The custom of eating raw fish. Humans become infected with *C. sinensis* by eating raw fish, particularly males and aged individuals (Qian et al., 2016). For the last thirty years, the estimated infected population has increased from ~ 7 million in the 1990s to at least 15 million in the 2000s (Qian et al., 2012). The increased prevalence of *C. sinensis* infection has been attributed to: (1) an increased income (per capita), offering more people opportunity to afford and regularly consume raw fish at restaurants (Qian et al., 2013a); (2) the common misconception that drinking liquor will kill *C. sinensis* metacercariae in raw fish (Qian et al., 2016); and (3) improved trade and traffic networks that facilitate the transport of fish to non-endemic provinces, enabling the geographic spread of *C. sinensis*. Despite efforts to stop the custom of humans eating raw fish, the cycle of *C. sinensis* continues.

1.4. Pathogenesis, host susceptibility and immunity

Following ingestion of infected fish by the human or other piscivorous host (e.g., dog or cat), the metacercariae hatch and juveniles migrate and develop to adult flukes in the biliary system (Sripa et al., 2010). The adult flukes use suckers to attach to the epithelium cells and, *via* mechanical irritation, cause damage to the bile ducts (Kim et al., 2009b). Adult flukes also frequently obstruct the bile duct and increase bile duct pressure, which causes bile stagnation, pigment deposition and the formation of bile duct stones (chololithiasis) (Qian et al., 2016). Moreover, fluke eggs are readily trapped in tissues and can induce granulomatous lesions (Sripa et al., 2007). In addition to mechanical irritation, the adult flukes can alter surrounding host cells and excretory-secretory products (ESPs) can stimulate the proliferation of epithelium cells, regulating cancer-related genes, and inducing the production of host inflammatory cytokines (Kim et al., 2010; Zhang et al., 2013; Pak et al., 2014). A number of ESPs have been shown to have a proliferative effect on host cells (Kim et al., 2008; Zhang et al., 2013; Chen et al., 2015; Wang et al., 2017a). For instance, *C. sinensis* granulin (*Cs*GRN) is a growth factor that enhances migration and invasion of host cells (Wang et al., 2017a). In addition, *C. sinensis* secretes phospholipase A2 (*Cs*sPLA2) that can regulate the synthesis of host collagens, and contributes to the development of liver fibrosis (Hariprasad et al., 2012; Zhang et al., 2013; Wu et al., 2017).

Clonorchis sinensis infection can induce cholangiocarcinoma (CCA) (Won et al., 2014). *C. sinensis* ESPs have been shown to enhance the expression of adhesion proteins and metalloproteinases (MMPs) in CCA cells *in vitro* (Won et al., 2014). Adhesion proteins facilitate cell aggregation, and MMPs can promote the invasion of CCA cells into neighbouring extracellular matrices (Won et al., 2014). Furthermore, ESPs can alter the expression of cancer-associated miRNAs, which are usually involved in inflammatory processes, oncogene regulation and cell proliferation (Pak et al., 2014). For instance, the down-regulation of let-7i, a miRNA that functions in tumour suppression, is suggested to be associated with the immune response to the *C. sinensis* infection (Pak et al., 2014).

During the host immune reaction, Toll-like receptors play key roles in recognising *C. sinensis* antigens (Yan et al., 2015). The expression of two types of Toll-like receptors (TLR2 and TLR4) have been found to be induced by *C. sinensis* ESPs (Yan et al., 2015). Interestingly, Yan et al. (2015) showed that the elevated TLR4 promotes the expression of the nuclear-factor κB (NF-κB) and tumour necrosis factor-α (TNF-α) in ESP-treated biliary epithelial cells. In addition, free radicals, which can be triggered by the *C. sinensis* ferritin heavy chain protein (*Cs*FHC), have been shown to promote the expression of pro-inflammatory cytokines (IL-1β and IL-6) *via* the NF-κB signalling pathway (Nam et al., 2012; Mao et al., 2015). NF-κB also triggers the production of cyclooxygenase-2 (COX-2) and inducible nitric oxide synthase (iNOS) (Zheng et al., 2017). High-level expression of COX-2 can have a proliferative effect on tumour cells (Kim et al., 2009c), and the accumulation of iNOS generates endogenous nitric oxide (NO), which ultimately causes DNA damage and inhibits cell apoptosis (Zheng et al., 2017). NO can also lead to the formation of N-nitroso compounds, such as nitrosamine (Sripa et al., 2007). Being a primary carcinogen, nitrosamine can induce mutagenesis, and thus leads to a malignant transformation of epithelial cells (Sripa et al., 2007). Currently, the exact mechanism of how *C. sinensis* regulates host immune responses is still unclear, and requires further investigation.

The immune response to clonorchiasis varies depending on definitive host species (Sohn et al., 2006). For example, some mouse strains have low susceptibility to *C. sinensis* infection (Uddin et al., 2012) - both type 1 T helper (Th1) and type 2 T helper (Th2) responses have been observed, supporting a significant up-regulation of immunoglobulin E (IgE), interferon gamma (IFN-ϒ) and interleukin 13 (IL-13) (Uddin et al., 2012). Compared with mice, rats are more susceptible and have an exclusive Th2 response against clonorchiasis (Wang et al., 2009). Nonetheless, previously exposed rats are resistant to the re-infection with *C. sinensis*. Such rats exhibit high levels of serum IgE and bile immunoglobulin A (IgA) (Zhang et al., 2008a). In contrast to mice and rats, humans become readily re-infected (Sohn et al., 2006). In the bile and serum of infected humans, the levels of immunoglobulin G (IgG) and IgG4 correlate with infection intensity, whereas IgA and IgE are usually only moderately up-regulated (Yen et al., 1992). In the future, the cause of such differences in immune responses should be studied, which might provide opportunities for developing new ways of preventing infection or disease.

1.5. Diagnosis, treatment and control

Diagnosis

Clonorchiasis can be diagnosed using various methods, including the microscopic detection of eggs in faeces, PCR-based approaches, immunological techniques and/or ultrasound imaging (reviewed in Qian et al., 2016). Among these, egg detection methods are most widely used (Hong et al., 2003; Qian et al., 2013b). However, the eggs of *C. sinensis* are difficult to distinguish from those of other liver flukes, particularly those of *O. viverrini* (see Traub et al., 2009).

Molecular methods are usually more sensitive and specific for diagnosis than traditional coproscopic or serological methods (reviewed by Qian et al., 2016). For example, a multiplex PCR technique that uses 2 pairs of primers from mitochondrial sequences has been applied to effectively detect *C. sinensis* and discriminate *C. sinensis* from *O. viverrini* (see Le et al., 2006). In addition to mitochondrial sequences, internal transcribed spacer (ITS) sequences have also been applied for molecular detection (Traub et al., 2009). Real-time fluorescence PCR technique has been combined with melting curve analysis to specifically detect *C. sinensis* and *O. viverrini* by amplifying mitochondrial *nad*2 or *cox*1 sequences from these two species (Sanpool et al., 2012; Cai et al., 2014). This technique has a high sensitivity, and the detection limit can be less than 1 pg of genomic DNA (Cai et al., 2014). Another rapid DNA detection method, called loop-mediated isothermal amplification (LAMP), seems quite promising for use in a field situation due to its high diagnostic specificity (100%) and sensitivity (~97%) (Cai et al., 2010; Chen et al., 2013; Rahman et al., 2017). This method can effectively detect *C. sinensis* in intermediate hosts and human faecal samples. In the future, advanced molecular diagnostic tools should be developed for detailed studies of biology and epidemiology, and for the monitoring of infection prevalence.

Treatment and control

No vaccine is yet available to prevent clonorchiasis in humans or any other definitive hosts (reviewed by Qian et al., 2016). The control of clonorchiasis relies on mass chemotherapy, health education, community awareness of the impact of clonorchiasis and CCA.

There is a long tradition of eating raw cyprinid fish in many parts of South China and Southeast Asia (Broglia et al., 2011; Vinh et al., 2017). In these areas, some people believe that eating raw fish improves their health and vitality (Hung et al., 2013), while others may know that the consumption of raw cyprinid fish leads to clonorchiasis (Vinh et al., 2017). However, social customs still encourage people in endemic regions to regard raw fish (infected with *C. sinensis* or not) as a delicacy (Qian et al., 2013a). This habit might be changed through health education, and can be applied successfully in combination with chemotherapy to decrease the prevalence of infection and the re-infection rate in endemic areas (Oh et al., 2014).

Mass chemotherapy with the only recommended drug, praziquantel (PZQ), is applied in areas endemic for clonorchiasis (Choi et al., 2010). Three doses of 25 mg/kg per day for 1 or 2 days are usually recommended (Hong and Fang, 2012; Qian et al., 2013c). Using this regimen, the cure rate can reach 83%, and the faecal egg count reduction can reach 99% (Rim, 1986). However, 40% of people treated with the recommended dosage have adverse effects, including dizziness, sleepiness, headaches and/or even anaphylactic reactions (Shen et al., 2007). Another approved broad-spectrum drug, tribendimidine, was tested *in vitro* and *in vivo* (Keiser and Vargas, 2010; Soukhathammavong et al., 2011; Xiao et al., 2011) and is reported to have an efficacy similar to PZQ for treating *C. sinensis* infection. A dose of 400 mg tribendimidine was found to have the best-tolerated effect, with a cure rate of 44-50% (Qian et al., 2013c; Xu et al., 2014). In addition, 400 mg tribendimidine daily for 3 days can achieve a cure rate of 58%.

Other methods that rely on breaking the life cycle could be applied to control this parasite, but might have some challenges. For instance, the snail hosts of *C. sinensis* might be controlled using molluscicides (Wang et al., 2007). However, as molluscicides can be harmful to fishes and plants, this method has not been widely adopted. Eliminating the release of *C. sinensis* eggs into aquaculture environments by removing toilets near or over fish-ponds or waterways is a more practical and effective approach (Qian et al., 2016). With the rapid expansion of aquaculture industry in southern China, for example, fish are being increasingly cultured in existing rivers and lakes, making control much harder (Qian et al., 2016). Given the strengths and weakness of each of the above methods, an integrated management strategy (including education campaigns, breaking the lifecycle of the parasite and strategic PZQ treatment) is recommended to achieve effective control of *C. sinensis* and clonorchiasis.

Potential for drug resistance development

Although the use of PZQ can achieve a substantial reduction in faecal egg counts in people (~ 99%) (Rim, 1986), cure rates vary in different geographic regions and are usually less than 85% (Choi et al., 2010). For instance, in Qiyang, China, with 75 mg/kg given in 4 doses, the cure rate of PZQ was 57% (Xu et al., 2014). Treatment failures often go unnoticed in patients with asymptomatic infection (Echaubard et al., 2016). Although there is no unequivocal evidence of drug resistance in *C. sinensis* (see Hong et al., 2012; Qian et al., 2016; Tang et al., 2016), resistance could develop, given reported cases of limited effectiveness of PZQ against *C. sinensis* in North Vietnam (Tinga et al., 1999) and the human blood fluke, *Schistosoma mansoni*, in Egypt (Ismail et al., 1999). In addition, drug resistant schistosome strains can be induced artificially within only five generations (Mwangi et al., 2014). Hence, it is reasonable to propose that drug-resistant strains of *C. sinensis* can develop. An understanding of the genetics and genomics of *C. sinensis* should facilitate studies of resistance, and also could assist in finding new interventions against clonorchiasis.

1.6. Aspects of the genetics of *C. sinensis*

Studies of *C. sinensis* from China and Russian Far East have reported that the karyotype of this species is 2n = 14 (Gao et al., 1987; Li, 1989; Gao et al., 1993; Zadesenets et al., 2012), which is partly supported by the similar chromosome number of *O. felineus* (2n = 14) (Polyakov et al., 2010; Zadesenets et al., 2012) and *O. viverrini* (2n = 12) (Kaewkong et al., 2012; Zadesenets et al., 2012), whereas other research from China and Korea (Park et al., 2000) has shown a distinct number of chromosomes (2n = 56). An explanation might be that the latter karyotype (Park et al., 2000) might represent octoploid cells of normal *C. sinensis* (Zadesenets et al., 2012). In the gonads of opisthorchids, germ cell precursors, such as clusters containing eight cells or even octoploid cells were frequently observed (Gao et al., 1987; Gao et al., 1993; Zadesenets et al., 2012). Currently, the mechanism for maintaining diploidy in this fluke is unclear, although some evidence from planarians suggests that the number of chromosomes can be restored by expelling sets of chromosomes from polyploid cells (Lentati, 1970). Another explanation is that these octoploid cells are formed by the fusion of two tetraploid gametes. Evidence of unreduced (tetraploid) eggs has been reported for *Fasciola hepatica* and *Paragonimus westermani* (see Terasaki et al., 1996; Fletcher et al., 2004). In addition, polyploidy is also often associated with parthenogenesis, which can lead to a novel lineage (Terasaki et al., 1996; Fletcher et al., 2004).

1.7. Genetic variation within *C. sinensis*

An approach to identifying cryptic species is to explore genetic variation among geographically distinct isolates. Numerous nuclear and mitochondrial loci have been sequenced, mostly for *C. sinensis* samples from China (Table 1-3). When considering the number of specimens sequenced and countries sampled, ITS-1 and *cox1* are the most commonly used loci. By investigating this range of markers, most genetic variation studies suggest a low divergence of *C. sinensis* across endemic areas (Park and Yong, 2001; Lee and Huh, 2004; Le et al., 2006; Cai et al., 2012; Liu et al., 2012; Tatonova et al., 2012; Tatonova et al., 2013; Chelomina et al., 2014). Low genetic divergence within flukes is frequently reported and attributed, to some extent, to their complex life cycle, involving different life stages and up to three hosts (cf. Criscione et al., 2005; Prugnolle et al., 2005a; Prugnolle et al., 2005b; Criscione et al., 2011; Auld and Tinsley, 2015). Because *C. sinensis* is a hermaphrodite with a typical digenean life cycle, it can self-fertilise or might mate with other genetically identical clones within the same definitive host (cf. Lun et al., 2005). Selfing or inbred organisms are expected to be less genetic diverse than random-mating organisms due to an increased genome-wide homozygosity and a reduced population size (Glemin and Galtier, 2012), as seen in the hermaphroditic tapeworm, *Oochoristica javaensis* (see Detwiler and Criscione, 2017). Currently, the fertilisation process is poorly studied in *C. sinensis*. In other opisthorchiid liver flukes, such as *O. viverrini* and *O. felineus*, self-fertilisation (selfing) is thought to prevail, as heterozygote deficiency is commonly observed (Laoprom et al., 2012; Zhigileva et al., 2014; Pitaksakulrat et al., 2017). In this case, a factor predicted to play a determining role in the mode of fertilisation is parasite density within the definitive host (Detwiler et al., 2017). In contrast, a high level of genetic diversity has been reported in the common liver fluke, *F. hepatica* (see Walker et al., 2011; Beesley, 2016; Beesley et al., 2017). In *F. hepatica* populations in the UK, out-crossing was predicted to be more frequent than selfing, particularly in sheep (Beesley, 2016; Beesley et al., 2017). It remains to be determined whether out-crossing plays a significant role in *C. sinensis* reproduction; this aspect warrants detailed investigation.

Genetically identical *C. sinensis* cercariae derived from the asexual reproduction in snails might attach to the same freshwater fish and aggregate in the same definitive host. This bias may result in incorrect estimates of key population parameters (Criscione et al., 2005; Prugnolle et al., 2005b). For example, aggregated clones may mate with each other, increasing the degree of inbreeding and lowering the effective population size (Vilas et al., 2012). At the same time, clones are produced within the sedentary snail hosts, which contributes to local genetic homogeneity and prevents genetic exchange (Zhigileva et al., 2014). Furthermore, gene flow caused by host migration offsets the genetic diversification and local adaptation of the flukes (Gandon and Michalakis, 2002). High levels of gene flow have been found among different *C. sinensis* isolates from geographically distinct areas (Tatonova et al., 2013; Chelomina et al., 2014), which appears to be introduced by hosts with high mobility (i.e. humans) (Criscione et al., 2005; Prugnolle et al., 2005a; Prugnolle et al., 2005b). In contrast to *C. sinensis*, which cycles in freshwater and terrestrial hosts, trematodes that only cycle in freshwater hosts, such as *Deropegus aspina* and *Plagioporus shawi*, have been proven to have well-structured populations and low levels of gene flow due to a relatively limited mobility of first intermediate hosts in freshwater (Criscione and Blouin, 2004).

Previous research has revealed that genetic diversity patterns are linked to the geographical origins of *C. sinensis* (see Liu et al., 2012; Tatonova et al., 2012; Tatonova et al., 2013; Chelomina et al., 2014). Based on the variation in the ITS-1 sequence, Tatonova et al. (2012) divided *C. sinensis* into "northern" and "southern" groups. In other research, data for four mitochondrial loci (*cox*1, *cox*2, *nad*1 and *nad*2) suggested that a *C. sinensis* isolate from Heilongjiang (North China) clustered with a Russia isolate (Liu et al., 2012). These findings are supported by the proposed rapid expansion of *C. sinensis* from central China at the end of the Last Glacial Maximum (LGM) (15,000 yr BP) (Chelomina et al., 2014). During the glacial period, many areas in central and south China were predicted to have a relatively warm and stable climate and, thus, might have served as glacial refuges for different species (cf. You et al., 2010; Qiu et al., 2011) including the intermediate hosts of liver flukes. In Europe, previous research found that most *Bythinella* spp. survived the Pleistocene in restricted refuges, whereas other lineages in permafrost regions perished (Benke et al., 2009).

With the increased temperature following the ice age, many species migrated to the previously iced territory from their original refuges (Hewitt, 1999). This hypothesis is also consistent with the star-like haplotype network of Chinese liver fluke from endemic areas (Tatonova et al., 2013). In particular, a single *cox*1 haplotype lineage was detected in samples collected from geographically distinct areas (China, Russia and Vietnam) (Tatonova et al., 2013), suggesting that those individuals might have descended from a small population. Moreover, *cox*1 and ITS-1 sequences for *C. sinensis* individuals from China show a higher genetic diversity than those from Russia and Korea (Tatonova et al., 2012; Tatonova et al., 2013), which is consistent with a low similarity coefficient detected in the Guangdong group (South China) compared with the Heilongjiang group (North China) using PCR analysis of mobile genetic elements (MGE-PCR) and random amplification of polymorphic DNA (RAPD) (Lai et al., 2008), although the latter molecular methods have their limitations (De Wolf et al., 2004). Besides China, the Korean peninsula is also likely to have contained glacial refuges (Qiu et al., 2011). However, there is no clear evidence that *C. sinensis* has been subjected to long-term isolation, although isozyme profiles of adult flukes from China and Korea showed differences between the two populations (Park and Yong, 2001).

1.8. Current genetic markers for *C. sinensis*

Most genetic studies of *C. sinensis* have used a limited number of mitochondrial and/or nuclear loci (Park and Yong, 2001; Lee and Huh, 2004; Le et al., 2006; Cai et al., 2012; Liu et al., 2012; Tatonova et al., 2012; Sun et al., 2013; Tatonova et al., 2013; Chelomina et al., 2014) (Table 1-3). These loci display different levels of variability (Sun et al., 2013). For example, the ITS-2 sequence was reported to be conserved in multiple studies (Liu et al., 2007; Tatonova et al., 2012; Shin et al., 2013), whereas the *cox*1 sequence varied by 2.17% (Sun et al., 2013). Compared with ITS-2 sequences, the *cox*1 sequence is employed for assessing levels of intraspecific variation (Vilas et al., 2005; Vanhove et al., 2013; Blasco-Costa et al., 2016). The level of intraspecific divergence in mitochondrial DNA (mtDNA) usually varies considerably in platyhelminths (Vilas et al., 2005). However, which mtDNA marker is most suitable for genetic variation research in trematodes is still contentious (Blasco-Costa et al., 2016). In future, this problem might be addressed by comparing the variability of each mt protein-encoding gene among geographically distinct isolates of *C. sinensis*. With the newly published mt genomes from China, Russia and Korea, such an analysis is feasible and might provide an opportunity to define "variable" markers in *C. sinensis* (see Shekhovtsov et al., 2010; Cai et al., 2012; Wang et al., 2017b). It would also be useful to explore genetic variation using a large number of nuclear genomic markers. Microsatellite markers were developed for *C. sinensis* (Nguyen et al. 2015), but have not yet been used for population genetic studies of this species. The advent of next-generation sequencing (NGS) now enables high-throughput sequencing of whole-genomes, which provides an opportunity to better define relationships within and among geographically distinct populations, and to investigate loci involved in the infection process and parasitism.

Methods for studying genetic variation in genomes of C. sinensis and other trematodes

Currently, most genome-wide variation studies of trematodes have identified nucleotide-level variants by mapping read data to a single reference sequences, assuming that this sequence is sufficient to represent all the isolates (see Clement et al., 2013; Cwiklinski et al., 2015; Young et al., 2015; Crellen et al., 2016). This approach is usually suited for the comparison of isolates representing the same species of the same karyotype. For instance, *Schistosoma japonicum*, another important trematode endemic in East Asia, has the same karyotype as other schistosome species (seven pairs of autosomal chromosomes and one pair of sex chromosomes) (Rollinson et al., 1997; Lawton et al., 2011), which enables detailed phylogeographic investigations using a read-mapping approach (cf. Young et al., 2015). However, mapping reads to a single reference genome can create false-positive read alignments or may miss strain-specific sequences when genomes with large-scale rearrangements are compared, especially for the taxa representing a species complex, which might be the case for *C. sinensis*, given the observed evidence of karyotypic variation among geographically distinct isolates (Park and Yong, 2001; Zadesenets et al., 2012).

An alternative method would be to de novo-assemble genomes from geographically distinct isolates and then to evaluate genetic variation within syntenic blocks among the assembled genomes. Although the comparison of genomes might require sophisticated bioinformatic analyses and substantial computational resources, this approach is particularly applicable only a few geographical distinct isolates within a suspected species complex are compared. Given the evidence of cryptic species and/or species complexes within opisthorchiid flukes and other trematodes (e.g., Saijuntha et al., 2007; Leung et al., 2009; Criscione et al., 2011), selecting a suitable approach for genomic/genetic comparisons is critical for future genetic studies.

1.9. Molecular investigations using genomic, transcriptomic and/or proteomic tools

Molecular research has been benefited enormously from the rapid development of sequencing and bioinformatic techniques (Korhonen et al., 2016). In 2011, the first genome assembly for *C. sinensis* was published (Wang et al., 2011). Two years later, a refined genome, with a re-annotation and corresponding transcriptomic analysis, was published (Huang et al., 2013). The latest assembly is 547 Mb in size, and the N50 size is 417 kb; 13,634 genes were predicted for the genome, 79.6% of which is annotated (Huang et al., 2013). This genome provided a foundation for the study of a wide range of biological processes and mechanisms.

Explorations of key biological processes in C. sinensis

After being consumed and digested by the host, metacercariae migrate and excyst in upper small intestine (Sripa et al., 2010). During this process, the excystment is not only modulated by the host digestive enzymes, but also facilitated by the proteases secreted by the metacercariae (Li et al., 2004). These proteases can cleave the disulfide bonds in proteins of the cyst wall (Li et al., 2004). Families of *C. sinensis* proteases, including serine and cysteine proteases, have been annotated, and are highly transcribed in metacercariae (Nithikathkul et al., 2007; Yoo et al., 2011; Huang et al., 2013).

Following excystment, a juvenile liver fluke relies mainly on the energy from lipoproteins, which are abundant in different forms in bile ducts (Manzato et al., 1976). Studies have shown that *C. sinensis* has evolved a series of molecular mechanisms needed to absorb and digest nutrients from the host. For example, pathways in the tricarboxylic acid (TCA) cycle, fatty acid metabolism and glycolysis were identified (Wang et al., 2011). However, the *de novo* fatty acid biosynthesis pathway is absent (Huang et al., 2013), which means that the acquisition of fatty acid from the host animal is crucial to *C. sinensis*. In addition, multiple Niemann-Pick C2 (NPC2) genes, which function in lipid transportation and chemical homeostasis, have been annotated and are highly transcribed in *C. sinensis* (see Huang et al., 2013). Transporter genes for fatty acids, glucose and other nutrients are also highly transcribed (Huang et al., 2013), and free amino acids can also be transformed into part of TCA cycle. Moreover, cathepsin F genes, a key component of the lysosomal proteolytic system (Santamaria et al., 1999), are highly transcribed/expressed in adult *C. sinensis* (see Kang et al. 2010).

During the energy metabolism of the flukes, substantial oxygen is required to oxidise FADH2 and NADH, to generate ATP (Tielens et al., 1984). This requires a relatively high utilisation of oxygen in the micro-aerobic environment of the biliary system (Rashid et al., 1997). Three globin genes have been identified in *C. sinensis* (see Huang et al., 2013), of which myoglobin was found to be highly transcribed (Huang et al., 2013). Myoglobins are known to have a high oxygen affinity in trematodes, suggesting an ability of *C. sinensis* to efficiently consume oxygen (Rashid et al., 1997).

Excretory-secretory (ES) proteins play critical roles in migration, excystment, detoxification and feeding (Zheng et al., 2011). To date, 297 ES protein genes have been identified in the *C. sinensis* genome, which mainly function in lipid-binding, transportation, peptide hydrolysis and/or peptidase inhibition (Huang et al., 2013). Among them, six secretory cysteine protease genes are highly transcribed in *C. sinensis*, reflecting their roles in the degradation of host proteins during parasite invasion (Na et al., 2006; Huang et al., 2013). Moreover, nine *C. sinensis* ES protein genes have been proposed to function in cell proliferation. Among them, a granulin gene is highly transcribed in the sucker of the fluke. In humans, granulins have been shown to play roles in mediating cell growth during normal development and in tumorigenesis (Bateman and Bennett, 2009). The high expression of granulin in *C. sinensis* suggests that this gene might be involved in the genesis of cholangiocarcinoma (Huang et al., 2013).

1.10. Conclusions

The appraisal of the literature shows that *C. sinensis* is a highly significant carcinogenic fluke of humans and other fish-eating mammals. Although aspects of the biology, epidemiology, pathogenesis and control of *C. sinensis* have been investigated, the genetics and genomics of this parasite has only recently been tackled. The published genome for a Chinese isolate of *C. sinensis* will assist in improving the understanding of this parasite and the diseases that it causes. However, to date, no study has yet focused on assessing genetic variation at the whole mitochondrial or nuclear genome level. Furthermore, there is a paucity of information about genetic variation in protein-encoding genes of known or inferred biological relevance.

This thesis focuses on addressing three research aims:

(1) To assess variation in the mitochondrial genome between *C. sinensis* from China, Korea and Russia (Chapter 2);

(2) To sequence, assemble and annotate a nuclear genome of a Korean isolate and compare it with that of the published genome of a Chinese isolate. (Chapter 3); and

(3) To explore the population structure of *Schistosoma japonicum* in China and detect genes under positive selection within geographically distinct locations (Chapter 4).

1.11. References

Abdussalam, M., Kaferstein, F.K., Mott, K.E., 1995. Food safety measures for the control of foodborne trematode infections. Food Control 6, 71-79.

Attwood, H.D., Chou, S.T., 1978. The longevity of *Clonorchis sinensis*. Pathology 10, 153-156.

Auld, S.K., Tinsley, M.C., 2015. The evolutionary ecology of complex lifecycle parasites: linking phenomena with mechanisms. Heredity (Edinb.) 114, 125-132.

Bateman, A., Bennett, H.P., 2009. The granulin gene family: from cancer to dementia. Bioessays 31, 1245-1254.

Beesley, N.J., 2016. Population genetic structure of *Fasciola hepatica* in Great Britain. Doctoral dissertation, University of Liverpool.

Beesley, N.J., Williams, D.J., Paterson, S., Hodgkinson, J., 2017. *Fasciola hepatica* demonstrates high levels of genetic diversity, a lack of population structure and high gene flow: possible implications for drug resistance. Int. J. Parasitol. 47, 11-20.

Benke, M., Brandle, M., Albrecht, C., Wilke, T., 2009. Pleistocene phylogeography and phylogenetic concordance in cold-adapted spring snails (*Bythinella* spp.). Mol. Ecol. 18, 890-903.

Blasco-Costa, I., Cutmore, S.C., Miller, T.L., Nolan, M.J., 2016. Molecular approaches to trematode systematics: 'best practice' and implications for future study. Syst. Parasitol. 93, 295-306.

Blum, T., Briesemeister, S., Kohlbacher, O., 2009. MultiLoc2: integrating phylogeny and Gene Ontology terms improves subcellular protein localization prediction. BMC Bioinf. 10, 274.

Bogitsh, B.J., Carter, C.E., Oeltmann, T.N., 2013. Human Parasitology, fourth ed. Academic Press.

Broglia, A., Kapel, C., 2011. Changing dietary habits in a changing world: emerging drivers for the transmission of foodborne parasitic zoonoses. Vet. Parasitol. 182, 2-13.

Bui, T.N., Pham, T.T., Nguyen, N.T., Nguyen, H.V., Murrell, D., Phan, V.T., 2016. The importance of wild fish in the epidemiology of *Clonorchis sinensis* in Vietnam. Parasitol. Res. 115, 3401-3408.

Cai, X.Q., Liu, G.H., Song, H.Q., Wu, C.Y., Zou, F.C., Yan, H.K., Yuan, Z.G., Lin, R.Q., Zhu, X.Q., 2012. Sequences and gene organization of the mitochondrial genomes of the liver flukes *Opisthorchis viverrini* and *Clonorchis sinensis* (Trematoda). Parasitol. Res. 110, 235-243.

Cai, X.Q., Xu, M.J., Wang, Y.H., Qiu, D.Y., Liu, G.X., Lin, A., Tang, J.D., Zhang, R.L., Zhu, X.Q., 2010. Sensitive and rapid detection of *Clonorchis sinensis* infection in fish by loop-mediated isothermal amplification (LAMP). Parasitol. Res. 106, 1379-1383.

Cai, X.Q., Yu, H.Q., Li, R., Yue, Q.Y., Liu, G.H., Bai, J.S., Deng, Y., Qiu, D.Y., Zhu, X.Q., 2014. Rapid detection and differentiation of *Clonorchis sinensis* and *Opisthorchis viverrini* using real-time PCR and high resolution melting analysis. Scientific World Journal 2014, 893981.

Cao, L., Naylor, R., Henriksson, P., Leadbitter, D., Metian, M., Troell, M., Zhang, W., 2015. Global food supply. China's aquaculture and the world's wild fisheries. Science 347, 133-135.

Chelomina, G.N., Tatonova, Y.V., Hung, N.M., Ngo, H.D., 2014. Genetic diversity of the Chinese liver fluke *Clonorchis sinensis* from Russia and Vietnam. Int. J. Parasitol. 44, 795-810.

Chen, D., Chen, J., Huang, J., Chen, X., Feng, D., Liang, B., Che, Y., Liu, X., Zhu, C., Li, X., Shen, H., 2010. Epidemiological investigation of *Clonorchis sinensis* infection in freshwater fishes in the Pearl River Delta. Parasitol. Res. 107, 835-839.

Chen, W., Ning, D., Wang, X., Chen, T., Lv, X., Sun, J., Wu, D., Huang, Y., Xu, J., Yu, X., 2015. Identification and characterization of *Clonorchis sinensis* cathepsin B proteases in the pathogenesis of clonorchiasis. Parasit. Vectors 8, 647.

Chen, Y., Wen, T., Lai, D.H., Wen, Y.Z., Wu, Z.D., Yang, T.B., Yu, X.B., Hide, G., Lun, Z.R., 2013. Development and evaluation of loop-mediated isothermal amplification (LAMP) for rapid detection of *Clonorchis sinensis* from its first intermediate hosts, freshwater snails. Parasitology 140, 1377-1383.

Cheng, R., Cao, Y., Yu, S., Yang, Y., Qin, X., Chen, T., 2007. Clonorchiasis transmission dynamics among hosts in China. Parasitoses Infect. Dis. 5, 203–205 (in Chinese).

Choi, B.I., Han, J.K., Hong, S.T., Lee, K.H., 2004. Clonorchiasis and cholangiocarcinoma: etiologic relationship and imaging diagnosis. Clin. Microbiol. Rev. 17, 540-552.

Choi, D.W., 1984. *Clonorchis sinensis*: life cycle, intermediate hosts, transmission to man and geographical distribution in Korea. Arzneimittelforschung 34, 1145-1151.

Choi, M.H., Chang, Y.S., Lim, M.K., Bae, Y.M., Hong, S.T., Oh, J.K., Yun, E.H., Bae, M.J., Kwon, H.S., Lee, S.M., Park, H.W., Min, K.U., Kim, Y.Y., Cho, S.H., 2011. *Clonorchis sinensis* infection is positively associated with atopy in endemic area. Clin. Exp. Allergy 41, 697-705.

Choi, M.H., Park, S.K., Li, Z.M., Ji, Z., Yu, G., Feng, Z., Xu, L.Q., Cho, S.Y., Rim, H.J., Lee, S.H., Hong, S.T., 2010. Effect of control strategies on prevalence, incidence and re-infection of clonorchiasis in endemic areas of China. PLoS Negl. Trop. Dis. 4, e601.

Chun, S.K., 1964. Studies on the experimental mode of infections of *Clonorchis sinensis*: III. Studies on the wormicidal effect of external mucous substance of some fresh water fish on the larva of *Clonorchis sinensis*. Kisaengchunghak Chapchi 2, 148-158 (in Korean).

Chung, B.J., Joo, C.Y., Choi, D.W., 1980. Seasonal variation of snail population of *Parafossarulus manchouricus* and larval trematode infection in river Kumho, Kyungpook province, Korea. Kisaengchunghak Chapchi 18, 54-64 (in Korean).

Clement, J.A., Toulza, E., Gautier, M., Parrinello, H., Roquis, D., Boissier, J., Rognon, A., Mone, H., Mouahid, G., Buard, J., Mitta, G., Grunau, C., 2013. Private selective sweeps identified from next-generation pool-sequencing reveal convergent pathways under selection in two inbred *Schistosoma mansoni* strains. PLoS Negl. Trop. Dis. 7, e2591.

Crellen, T., Allan, F., David, S., Durrant, C., Huckvale, T., Holroyd, N., Emery, A.M., Rollinson, D., Aanensen, D.M., Berriman, M., Webster, J.P., Cotton, J.A., 2016. Whole genome resequencing of the human parasite *Schistosoma mansoni* reveals population history and effects of selection. Sci. Rep. 6, 20954.

Criscione, C.D., Blouin, M.S., 2004. Life cycles shape parasite evolution: comparative population genetics of salmon trematodes. Evolution 58, 198-202.

Criscione, C.D., Poulin, R., Blouin, M.S., 2005. Molecular ecology of parasites: elucidating ecological and microevolutionary processes. Mol. Ecol. 14, 2247-2257.

Criscione, C.D., Vilas, R., Paniagua, E., Blouin, M.S., 2011. More than meets the eye: detecting cryptic microgeographic population structure in a parasite with a complex life cycle. Mol. Ecol. 20, 2510-2524.

Cwiklinski, K., Dalton, J.P., Dufresne, P.J., La Course, J., Williams, D.J.L., Hodgkinson, J., et al., 2015. The *Fasciola hepatica* genome: gene duplication and polymorphism reveals adaptation to the host environment and the capacity for rapid evolution. Genome Biol. 16, 71.

De Wolf, H., Blust, R., Backeljau, T., 2004. The use of RAPD in ecotoxicology. Mutat. Res. 566, 249-262.

Detwiler, J.T., Caballero, I.C., Criscione, C.D., 2017. Role of parasite transmission in promoting inbreeding: I. Infection intensities drive individual parasite selfing rates. Mol. Ecol. 26, 4391-4404.

Detwiler, J.T., Criscione, C.D., 2017. Role of parasite transmission in promoting inbreeding: II. Pedigree reconstruction reveals sib-transmission and consequent kin-mating. Mol. Ecol. 26, 4405-4417.

Doanh, P.N., Nawa, Y., 2016. *Clonorchis sinensis* and *Opisthorchis* spp. in Vietnam: current status and prospects. Trans. R. Soc. Trop. Med. Hyg. 110, 13-20.

Echaubard, P., Sripa, B., Mallory, F.F., Wilcox, B.A., 2016. The role of evolutionary biology in research and control of liver flukes in Southeast Asia. Infect. Genet. Evol. 43, 381-397.

Fan, P.C., 1998. Viability of metacercariae of *Clonorchis sinensis* in frozen or salted freshwater fish. Int. J. Parasitol. 28, 603-605.

Fang, Y.Y., Chen, Y.D., Li, X.M., Wu, J., Zhang, Q.M., Ruan, C.W., 2008. Current prevalence of *Clonorchis sinensis* infection in endemic areas of China. Chin. J. Parasitol. Parasit. Dis. 26 (99–103), 109 (in Chinese).

Fattakhov, R.G., Ushakov, A.V., Stepanova, T.F., Ianovich, V.A., Kopylov, P.V., 2012. Epizootiological characteristics of clonorchiasis foci in the Amur River ecosystem in the Jewish autonomic region. Med. Parazitol. (Mosk) 4, 15–18.

Fletcher, H.L., Hoey, E.M., Orr, N., Trudgett, A., Fairweather, I., Robinson, M.W., 2004. The occurrence and significance of triploidy in the liver fluke, *Fasciola hepatica*. Parasitology 128, 69-72.

Furst, T., Keiser, J., Utzinger, J., 2012. Global burden of human food-borne trematodiasis: a systematic review and meta-analysis. Lancet Infect. Dis. 12, 210-221.

Gandon, S., Michalakis, Y., 2002. Local adaptation, evolutionary potential and host–parasite coevolution: interactions between migration, mutation, population size and generation time. J. Evol. Biol. 15, 451-462.

Gao, L., You, S., Chen, S., Wu, M., 1993. Study on meiosis of *Clonorchis sinensis*. Chin. J. Schistosomiasis Control 5, 230–233 (in Chinese).

Gao, L., You, S., Chen, S., Wu, M., Li, G., Li, W., You, T., 1987. Primary analysis of karyotypes of *Clonorchis sinensis*. J. Hengyang Med. Coll. 15, 108–112 (in Chinese).

Glémin, S., Galtier, N., 2012. Genome evolution in outcrossing versus selfing versus asexual species. Methods. Mol. Biol. 855, 311-335.

Grosse, Y., Baan, R., Straif, K., Secretan, B., Ghissassi, F., Bouvard, V., Benbrahim-Tallaa, L., Guha, N., Galichet, L., Cogliano, V., 2009. WHO International Agency for Research on Cancer Monograph Working Group. A review of human carcinogens — Part A: pharmaceuticals. Lancet Oncol. 10 (1), 13–14.

Hane, J.K., Anderson, J.P., Williams, A.H., Sperschneider, J., Singh, K.B., 2014. Genome sequencing and comparative genomics of the broad host-range pathogen *Rhizoctonia solani* AG8. PLoS Genet. 10, e1004281.

Hariprasad, G., Kaur, P., Srinivasan, A., Singh, T.P., Kumar, M., 2012. Structural analysis of secretory phospholipase A2 from *Clonorchis sinensis*: therapeutic implications for hepatic fibrosis. J. Mol. Model 18, 3139-3145.

Hewitt, G.M., 1999. Post-glacial re-colonization of European biota. Biol. J. Linn. Soc. 68, 87-112.

Hong, S.T., Choi, M.H., Kim, C.H., Chung, B.S., Ji, Z., 2003. The Kato-Katz method is reliable for diagnosis of *Clonorchis sinensis* infection. Diagn. Microbiol. Infect. Dis. 47, 345-347.

Hong, S.T., Fang, Y., 2012. *Clonorchis sinensis* and clonorchiasis, an update. Parasitol. Int. 61, 17-24.

Huang, Y., Chen, W., Wang, X., Liu, H., Chen, Y., Guo, L., Luo, F., Sun, J., Mao, Q., Liang, P., Xie, Z., Zhou, C., Tian, Y., Lv, X., Huang, L., Zhou, J., Hu, Y., Li, R., Zhang, F., Lei, H., Li, W., Hu, X., Liang, C., Xu, J., Li, X., Yu, X., 2013a. The carcinogenic liver fluke, *Clonorchis sinensis*: new assembly, reannotation and analysis of the genome and characterization of tissue transcriptomes. PLoS One 8, e54732.

Hung, N.M., Madsen, H., Fried, B., 2013. Global status of fish-borne zoonotic trematodiasis in humans. Acta. Parasitol. 58, 231-258.

Isaji, S., Ugai, H., 2006. Freshwater molluscan fossil assemblages of the Middle Pleistocene Kiyokawa formation, Shimosa group, with special reference to the pH conditions of their habitat. Quat. Res. (Tokyo) 45, 169.

Ismail, M., Botros, S., Metwally, A., William, S., Farghally, A., Tao, L.F., Day, T.A., Bennett, J.L., 1999. Resistance to praziquantel: direct evidence from *Schistosoma mansoni* isolated from Egyptian villagers. Am. J. Trop. Med. Hyg. 60, 932-935.

Kaewkes, S., 2003. Taxonomy and biology of liver flukes. Acta Trop. 88, 177-186.

Kaewkong, W., Choochote, W., Kanla, P., Maleewong, W., Intapan, P.M., Wongkham, S., Wongkham, C., 2012. Chromosomes and karyotype analysis of a liver fluke, *Opisthorchis viverrini*, by scanning electron microscopy. Parasitol. Int. 61, 504-507.

Kang, S., Sultana, T., Loktev, V.B., Wongratanacheewin, S., Sohn, W.M., Eom, K.S., Park, J.K., 2008. Molecular identification and phylogenetic analysis of nuclear rDNA sequences among three opisthorchid liver fluke species (Opisthorchiidae: Trematoda). Parasitol. Int. 57, 191-197.

Kang, J.M., Bahk, Y.Y., Cho, P.Y., Hong, S.J., Kim, T.S., Sohn, W.M., Na, B.K., 2010. A family of cathepsin F cysteine proteases of *Clonorchis sinensis* is the major secreted proteins that are expressed in the intestine of the parasite. Mol. Biochem. Parasitol. 170, 7-16.

Kašnŷ, M., Mikeš, L., Hampl, V., Dvořák, J., Caffrey, C.R., Dalton, J.P., Horák, P., 2009. Chapter 4. Peptidases of trematodes. Adv. Parasitol. 69, 205-297.

Katano, O., Hosoya, K., Iguchi, K.i., Yamaguchi, M., Aonuma, Y., Kitano, S., 2003. Species diversity and abundance of freshwater fishes in irrigation ditches around rice fields. Environ. Biol. Fishes 66, 107-121.

Keiser, J., Utzinger, J., 2005. Emerging foodborne trematodiasis. Emerg. Infect. Dis. 11, 1507-1514.

Keiser, J., Vargas, M., 2010. Effect of artemether, artesunate, OZ78, praziquantel, and tribendimidine alone or in combination chemotherapy on the tegument of *Clonorchis sinensis*. Parasitol. Int. 59, 472-476.

Kim, D.W., Kim, J.Y., Moon, J.H., Kim, K.B., Kim, T.S., Hong, S.J., Cheon, Y.P., Pak, J.H., Seo, S.B., 2010. Transcriptional induction of minichromosome maintenance protein 7 (Mcm7) in human cholangiocarcinoma cells treated with *Clonorchis sinensis* excretory-secretory products. Mol. Biochem. Parasitol. 173, 10-16.

Kim, E.M., Kim, J.S., Choi, M.H., Hong, S.T., Bae, Y.M., 2008. Effects of excretory/secretory products from *Clonorchis sinensis* and the carcinogen dimethylnitrosamine on the proliferation and cell cycle modulation of human epithelial HEK293T cells. Korean J. Parasitol. 46, 127-132.

Kim, H.G., Han, J., Kim, M.H., Cho, K.H., Shin, I.H., Kim, G.H., Kim, J.S., Kim, J.B., Kim, T.N., Kim, T.H., Kim, T.H., Kim, J.W., Ryu, J.K., Moon, Y.S., Moon, J.H., Park, S.J., Park, C.G., Bang, S.J., Yang, C.H., Yoo, K.S., Yoo, B.M., Lee, K.T., Lee, D.K., Lee, B.S., Lee, S.S., Lee, S.O., Lee, W.J., Cho, C.M., Joo, Y.E., Cheon, G.J., Choi, Y.W., Chung, J.B., Yoon, Y.B., 2009b. Prevalence of clonorchiasis in patients with gastrointestinal disease: a Korean nationwide multicenter survey. World J. Gastroenterol. 15, 86-94.

Kim, T.I., Na, B.K., Hong, S.J., 2009a. Functional genes and proteins of *Clonorchis sinensis*. Korean J. Parasitol. 47, S59-S68.

Kim, Y.J., Choi, M.H., Hong, S.T., Bae, Y.M., 2009c. Resistance of cholangiocarcinoma cells to parthenolide-induced apoptosis by the excretory-secretory products of *Clonorchis sinensis*. Parasitol. Res. 104, 1011-1016.

Köhler, F., 2016. Rampant taxonomic incongruence in a mitochondrial phylogeny of *Semisulcospira* freshwater snails from Japan (Cerithioidea: Semisulcospiridae). J. Moll. Stud. 82, 268–281.

Korhonen, P.K., Young, N.D., Gasser, R.B., 2016. Making sense of genomes of parasitic worms: Tackling bioinformatic challenges. Biotechnol. Adv. 34, 663-686.

Lai, D.H., Wang, Q.P., Chen, W., Cai, L.S., Wu, Z.D., Zhu, X.Q., Lun, Z.R., 2008. Molecular genetic profiles among individual *Clonorchis sinensis* adults collected from cats in two geographic regions of China revealed by RAPD and MGE-PCR methods. Acta Trop. 107, 213-216.

Laoprom, N., Kiatsopit, N., Sithithaworn, P., Kopolrat, K., Namsanor, J., Andrews, R.H., Petney, T.N., 2016. Cercarial emergence patterns for *Opisthorchis viverrini* sensu lato infecting *Bithynia siamensis goniomphalos* from Sakon Nakhon Province, Thailand. Parasitol. Res. 115, 3313-3321.

Laoprom, N., Sithithaworn, P., Andrews, R.H., Ando, K., Laha, T., Klinbunga, S., Webster, J.P., Petney, T.N., 2012. Population genetic structuring in *Opisthorchis viverrini* over various spatial scales in Thailand and Lao PDR. PLoS Negl. Trop. Dis. 6, e1906.

Le, T.H., Van De, N., Blair, D., Sithithaworn, P., McManus, D.P., 2006. *Clonorchis sinensis* and *Opisthorchis viverrini:* development of a mitochondrial-based multiplex PCR for their identification and discrimination. Exp. Parasitol. 112, 109-114.

Lee, S.U., Huh, S., 2004. Variation of nuclear and mitochondrial DNAs in Korean and Chinese isolates of *Clonorchis sinensis*. Korean J. Parasitol. 42, 145-148.

Leung, T.L., Keeney, D.B., Poulin, R., 2009. Cryptic species complexes in manipulative echinostomatid trematodes: when two become six. Parasitology 136, 241-252.

Li, B., Wang, C., Li, D., Liu, T., Deng, L., Liu, Y., 1979. Ecology research in *Parafossarulus striatulus,* the first intermediate host of *Clonorchis sinensis* in Liaoning Province. J. China Med. Univ. 4, 3–6 (in Chinese).

Li, J., 1989. Initial report of Karyotypes of *Clonorchis sinensis*. Chin. J. Zoonoses 5, 57-58 (in Chinese).

Li, S., Chung, Y.B., Chung, B.S., Choi, M.H., Yu, J.R., Hong, S.T., 2004. The involvement of the cysteine proteases of *Clonorchis sinensis* metacercariae in excystment. Parasitol. Res. 93, 36-40.

Li, T., Yang, Z., Wang, M., 2014a. Correlation between clonorchiasis incidences and climatic factors in Guangzhou, China. Parasit. Vectors 7, 29.

Li, X., Li, J., Wang, Y., Fu, L., Fu, Y., Li, B., Jiao, B., 2011. Aquaculture industry in China: current state, challenges, and outlook. Rev. Fish. Sci. 19, 187-200.

Li, Z., Li, Y., Li, W., Li, Y., Ma, K., Gao, L., 2015. A new discovery of gastropod opercula from Youhe fomation, Weihe area. Quaternary Sciences 35, 642-649 (in Chinese).

Lin, R.Q., Tang, J.D., Zhou, D.H., Song, H.Q., Huang, S.Y., Chen, J.X., Chen, M.X., Zhang, H., Zhu, X.Q., Zhou, X.N., 2011. Prevalence of *Clonorchis sinensis* infection in dogs and cats in subtropical southern China. Parasit. Vectors 4, 180.

Liu, G.H., Li, B., Li, J.Y., Song, H.Q., Lin, R.Q., Cai, X.Q., Zou, F.C., Yan, H.K., Yuan, Z.G., Zhou, D.H., Zhu, X.Q., 2012a. Genetic variation among *Clonorchis sinensis* isolates from different geographic regions in China revealed by sequence analyses of four mitochondrial genes. J. Helminthol. 86, 479-484.

Liu, W.Q., Liu, J., Zhang, J.H., Long, X.C., Lei, J.H., Li, Y.L., 2007. Comparison of ancient and modern *Clonorchis sinensis* based on ITS1 and ITS2 sequences. Acta Trop. 101, 91-94.

Lun, Z.R., Gasser, R.B., Lai, D.H., Li, A.X., Zhu, X.Q., Yu, X.B., Fang, Y.-Y., 2005. Clonorchiasis: a key foodborne zoonosis in China. Lancet Infect. Dis. 5, 31–41.

Manzato, E., Fellin, R., Baggio, G., Walch, S., Neubeck, W., Seidel, D., 1976. Formation of lipoprotein-X. Its relationship to bile compounds. J. Clin. Invest. 57, 1248-1260.

Mao, Q., Xie, Z.Z., Wang, X.Y., Chen, W.J., Ren, M.Y., Shang, M., Lei, H.L., Tian, Y.L., Li, S., Liang, P., Chen, T.J., Liang, C., Xu, J., Li, X.R., Huang, Y., Yu, X.B., 2015. *Clonorchis sinensis* ferritin heavy chain triggers free radicals and mediates inflammation signaling in human hepatic stellate cells. Parasitol. Res. 114, 659-670.

McDonald, J.H., Kreitman, M., 1991. Adaptive protein evolution at the Adh locus in *Drosophila*. Nature 351, 652.

Miettinen, J., Shi, C.H., Liew, S.C., 2011. Deforestation rates in insular Southeast Asia between 2000 and 2010. Glob. Chang. Biol. 17, 2261-2270.

Mwangi, I.N., Sanchez, M.C., Mkoji, G.M., Agola, L.E., Runo, S.M., Cupit, P.M., Cunningham, C., 2014. Praziquantel sensitivity of Kenyan *Schistosoma mansoni* isolates and the generation of a laboratory strain with reduced susceptibility to the drug. Int. J. Parasitol. Drugs Drug Resist. 4, 296-300.

Na, B.K., Kim, S.H., Lee, E.G., Kim, T.S., Bae, Y.A., Kang, I., Yu, J.R., Sohn, W.M., Cho, S.Y., Kong, Y., 2006. Critical roles for excretory-secretory cysteine proteases during tissue invasion of *Paragonimus westermani* newly excysted metacercariae. Cell Microbiol. 8, 1034-1046.

Nagarajan, N., Pop, M., 2013. Sequence assembly demystified. Nat. Rev. Genet. 14, 157-167.

Nam, J.H., Moon, J.H., Kim, I.K., Lee, M.R., Hong, S.J., Ahn, J.H., Chung, J.W., Pak, J.H., 2012. Free radicals enzymatically triggered by *Clonorchis sinensis* excretory-secretory products cause NF-kappa B-mediated inflammation in human cholangiocarcinoma cells. Int. J. Parasitol. 42, 103-113.

Nguyen, T.T., Arimatsu, Y., Hong, S.J., Brindley, P.J., Blair, D., Laha, T., Sripa, B., 2015. Genome-wide characterization of microsatellites and marker development in the carcinogenic liver fluke *Clonorchis sinensis*. Parasitol. Res. 114, 2263-2272.

Nithikathkul, C., Tesana, S., Sithithaworn, P., Balakanich, S., 2007. Early stage biliary and intrahepatic migration of *Opisthorchis viverrini* in the golden hamster. J. Helminthol. 81, 39-41.

Oh, J.K., Lim, M.K., Yun, E.H., Cho, H., Park, E.Y., Choi, M.H., Shin, H.R., Hong, S.T., 2014. Control of clonorchiasis in Korea: effectiveness of health education for community leaders and individuals in an endemic area. Trop. Med. Int. Health 19, 1096-1104.

Ooi H.K., Chen, C.I., Lin, S.C., Tung, K.C., Wang, J.S., Kamiya, M., 1997. Metacercariae in fishes of Sun Moon lake which is an endemic area for *Clonorchis sinensis* in Taiwan. Southeast Asian J. Trop. Med. Public Health 28 (Suppl. 1), 222-223.

Pak, J.H., Kim, I.K., Kim, S.M., Maeng, S., Song, K.J., Na, B.K., Kim, T.S., 2014. Induction of cancer-related microRNA expression profiling using excretory-secretory products of *Clonorchis sinensis*. Parasitol. Res. 113, 4447-4455.

Pan, B., Fan, Y., Yang, W., 2000. Current situation and control strategy of parasitic diseases in Guangdong province. Ann. Bull. Soc. Parasitol. Guangdong 22, 5 (in Chinese).

Park, G.M., Im, K., Huh, S., Yong, T.S., 2000. Chromosomes of the liver fluke, *Clonorchis sinensis*. Korean J. Parasitol. 38, 201-206.

Park, G.M., Yong, T.S., 2001. Geographical variation of the liver fluke, *Clonorchis sinensis*, from Korea and China based on the karyotypes, zymodeme and DNA sequences. Southeast Asian J. Trop. Med. Public Health 32 (Suppl. 2), 12-16.

Petney, T.N., Andrews, R.H., Saijuntha, W., Wenz-Mucke, A., Sithithaworn, P., 2013. The zoonotic, fish-borne liver flukes *Clonorchis sinensis*, *Opisthorchis felineus* and *Opisthorchis viverrini*. Int. J. Parasitol. 43, 1031-1046.

Pitaksakulrat, O., Kiatsopit, N., Laoprom, N., Webster, B.L., Webster, J.P., Lamberton, P.H., Laha, T., Andrews, R.H., Petney, T.N., Blair, D., Carlton, E.J., Spear, R.C., Sithithaworn, P., 2017. Preliminary genetic evidence of two different populations of *Opisthorchis viverrini* in Lao PDR. Parasitol. Res. 116, 1247-1256.

Polyakov, A.V., Katokhin, A.V., Bocharova, T.A., Romanov, K.V., L'vova, M.N., Bonina, O.M., Yurlova, N.I., Mordvinov, V.A., 2010. Comparative Analysis of Karyotypes of *Opisthorchis felineus* from West Siberia. Contemp. Probl. Ecol. 3, 1-3.

Poulin, R., 2006. Global warming and temperature-mediated increases in cercarial emergence in trematode parasites. Parasitology 132, 143-151.

Prugnolle, F., Liu, H., de Meeus, T., Balloux, F., 2005a. Population genetics of complex life-cycle parasites: an illustration with trematodes. Int. J. Parasitol. 35, 255-263.

Prugnolle, F., Roze, D., Theron, A., T, D.E.M., 2005b. F-statistics under alternation of sexual and asexual reproduction: a model and data from schistosomes (platyhelminth parasites). Mol. Ecol. 14, 1355-1365.

Qian, M.B., Chen, Y.D., Fang, Y.Y., Tan, T., Zhu, T.J., Zhou, C.H., Wang, G.F., Xu, L.Q., Zhou, X.N., 2013b. Epidemiological profile of *Clonorchis sinensis* infection in one community, Guangdong, People's Republic of China. Parasit. Vectors 6, 194.

Qian, M.B., Chen, Y.D., Liang, S., Yang, G.J., Zhou, X.N., 2012. The global epidemiology of clonorchiasis and its relation with cholangiocarcinoma. Infect. Dis. Poverty. 1, 4.

Qian, M.B., Utzinger, J., Keiser, J., Zhou, X.N., 2016. Clonorchiasis. The Lancet 387, 800-810.

Qian, M.B., Yap, P., Yang, Y.C., Liang, H., Jiang, Z.H., Li, W., Tan, Y.G., Zhou, H., Utzinger, J., Zhou, X.N., Keiser, J., 2013a. Efficacy and safety of tribendimidine against *Clonorchis sinensis*. Clin. Infect. Dis. 56, E76-E82.

Qian, M.B., Yap, P., Yang, Y.C., Liang, H., Jiang, Z.H., Li, W., Utzinger, J., Zhou, X.N., Keiser, J., 2013c. Accuracy of the Kato-Katz method and formalin-ether concentration technique for the diagnosis of *Clonorchis sinensis*, and implication for assessing drug efficacy. Parasit. Vectors 6, 314.

Qiu, Y.X., Fu, C.X., Comes, H.P., 2011. Plant molecular phylogeography in China and adjacent regions: Tracing the genetic imprints of Quaternary climate and environmental change in the world's most diverse temperate flora. Mol. Phylogenet. Evol. 59, 225-244.

Rahman, S.M.M., Song, H.B., Jin, Y., Oh, J.K., Lim, M.K., Hong, S.T., Choi, M.H., 2017. Application of a loop-mediated isothermal amplification (LAMP) assay targeting cox1 gene for the detection of *Clonorchis sinensis* in human fecal samples. PLoS Negl. Trop. Dis. 11, e0005995.

Rashid, A.K., Van Hauwaert, M.L., Haque, M., Siddiqi, A.H., Lasters, I., De Maeyer, M., Griffon, N., Marden, M.C., Dewilde, S., Clauwaert, J., Vinogradov, S.N., Moens, L., 1997. Trematode myoglobins, functional molecules with a distal tyrosine. J. Biol. Chem. 272, 2992-2999.

Rhee, J.K., Baek, B.K., Ahn, B.Z., Park, Y.J., 1980. The wormicidal substances of fresh water fishes on *Clonorchis sinensis*: II. Preliminary research on the wormicidal substances from mucous substances of various fresh water fishes. Kisaengchunghak Chapchi 18, 98-104 (in Korean).

Rim, H.J., 1986. The current pathobiology and chemotherapy of clonorchiasis. Kisaengchunghak Chapchi 24 (Suppl.), 1-141 (in Korean).

Russell, S.J., Norvig, P., 2003. Artificial intelligence : a modern approach, second ed. Prentice Hall, Upper Saddle River, NJ/Great Britain.

Saijuntha, W., Sithithaworn, P., Wongkham, S., Laha, T., Pipitgool, V., Tesana, S., Chilton, N.B., Petney, T.N., Andrews, R.H., 2007. Evidence of a species complex within the food-borne trematode *Opisthorchis viverrini* and possible co-evolution with their first intermediate hosts. Int. J. Parasitol. 37, 695-703.

Sala-Bozano, M., Ketmaier, V., Mariani, S., 2009. Contrasting signals from multiple markers illuminate population connectivity in a marine fish. Mol. Ecol. 18, 4811-4826.

Sanpool, O., Intapan, P.M., Thanchomnang, T., Janwan, P., Lulitanond, V., Doanh, P.N., Van Hien, H., Dung, D.T., Maleewong, W., Nawa, Y., 2012. Rapid detection and differentiation of *Clonorchis sinensis* and *Opisthorchis viverrini* eggs in human fecal samples using a duplex real-time fluorescence resonance energy transfer PCR and melting curve analysis. Parasitol. Res. 111, 89-96.

Santamaria, I., Velasco, G., Pendas, A.M., Paz, A., Lopez-Otin, C., 1999. Molecular cloning and structural and functional characterization of human cathepsin F, a new cysteine proteinase of the papain family with a long propeptide domain. J. Biol. Chem. 274, 13800-13809.

Shekhovtsov, S.V., Katokhin, A.V., Kolchanov, N.A., Mordvinov, V.A., 2010. The complete mitochondrial genomes of the liver flukes *Opisthorchis felineus* and *Clonorchis sinensis* (Trematoda). Parasitol. Int. 59, 100-103.

Shen, C., Choi, M.H., Bae, Y.M., Yu, G., Wang, S., Hong, S.T., 2007. A case of anaphylactic reaction to praziquantel treatment. Am. J. Trop. Med. Hyg. 76, 603-605.

Shin, D.H., Oh, C.S., Lee, H.J., Chai, J.Y., Lee, S.J., Hong, D.W., Lee, S.D., Seo, M., 2013. Ancient DNA analysis on *Clonorchis sinensis* eggs remained in samples from medieval Korean mummy. J. Archaeol. Sci. 40, 211-216.

Sohn, W.M., 2009. Fish-borne zoonotic trematode metacercariae in the Republic of Korea. Korean J. Parasitol. 47 (Suppl.), S103-113.

Sohn, W.M., Zhang, H., Choi, M.H., Hong, S.T., 2006. Susceptibility of experimental animals to reinfection with *Clonorchis sinensis*. Korean J. Parasitol. 44, 163-166.

Soukhathammavong, P., Odermatt, P., Sayasone, S., Vonghachack, Y., Vounatsou, P., Hatz, C., Akkhavong, K., Keiser, J., 2011. Efficacy and safety of mefloquine, artesunate, mefloquine-artesunate, tribendimidine, and praziquantel in patients with *Opisthorchis viverrini*: a randomised, exploratory, open-label, phase 2 trial. Lancet Infect. Dis. 11, 110-118.

Sripa, B., Kaewkes, S., Intapan, P.M., Maleewong, W., Brindley, P.J., 2010. Food-borne trematodiases in Southeast Asia epidemiology, pathology, clinical manifestation and control. Adv. Parasitol. 72, 305-350.

Sripa, B., Kaewkes, S., Sithithaworn, P., Mairiang, E., Laha, T., Smout, M., Pairojkul, C., Bhudhisawasdi, V., Tesana, S., Thinkamrop, B., Bethony, J.M., Loukas, A., Brindley, P.J., 2007. Liver fluke induces cholangiocarcinoma. PLoS Med. 4, e201.

Sun, J., Xu, J., Liang, P., Mao, Q., Huang, Y., Lv, X., Deng, C., Liang, C., de Hoog, G.S., Yu, X., 2011. Molecular identification of *Clonorchis sinensis* and discrimination with other opisthorchid liver fluke species using multiple ligation-depended probe amplification (MLPA). Parasit. Vectors 4, 98.

Sun, J., Huang, Y., Huang, H., Liang, P., Wang, X., Mao, Q., Men, J., Chen, W., Deng, C., Zhou, C., Lv, X., Zhou, J., Zhang, F., Li, R., Tian, Y., Lei, H., Liang, C., Hu, X., Xu, J., Li, X., Xinbingyu, 2013. Low divergence of *Clonorchis sinensis* in China based on multilocus analysis. PLoS One 8, e67006.

Tang, Z.L., Huang, Y., Yu, X.B., 2016. Current status and perspectives of *Clonorchis sinensis* and clonorchiasis: epidemiology, pathogenesis, omics, prevention and control. Infect. Dis. Poverty 5, 71.

Tatonova, Y.V., Chelomina, G.N., Besprosvannykh, V.V., 2012. Genetic diversity of nuclear ITS1-5.8S-ITS2 rDNA sequence in *Clonorchis sinensis* Cobbold, 1875 (Trematoda: Opisthorchidae) from the Russian Far East. Parasitol. Int. 61, 664-674.

Tatonova, Y.V., Chelomina, G.N., Besprozvannykh, V.V., 2013. Genetic diversity of *Clonorchis sinensis* (Trematoda: Opisthorchiidae) in the Russian southern Far East based on mtDNA *cox*1 sequence variation. Folia Parasitol. (Praha) 60, 155-162.

Tatonova, Y.V., Chelomina, G.N., Nguyen, H.M., 2017. Inter-individual and intragenomic variations in the ITS region of *Clonorchis sinensis* (Trematoda: Opisthorchiidae) from Russia and Vietnam. Infect. Genet. Evol. 55, 350-357.

Terasaki, K., Shibahara, T., Noda, Y., Kayano, H., 1996. The oocyte of triploid fluke receiving intrusion of sperm from a diploid fluke--evidence for the origin of tetraploids in *Paragonimus westermani*. J. Parasitol. 82, 947-950.

Tielens, A.G., van den Heuvel, J.M., van den Bergh, S.G., 1984. The energy metabolism of *Fasciola hepatica* during its development in the final host. Mol. Biochem. Parasitol. 13, 301-307.

Tinga, N., De, N., Vien, H.V., Chau, L., Toan, N.D., Kager, P.A., Vries, P.J., 1999. Little effect of praziquantel or artemisinin on clonorchiasis in Northern Vietnam. A pilot study. Trop. Med. Int. Health 4, 814-818.

Traub, R.J., Macaranas, J., Mungthin, M., Leelayoova, S., Cribb, T., Murrell, K.D., Thompson, R.C., 2009. A new PCR-based approach indicates the range of *Clonorchis sinensis* now extends to Central Thailand. PLoS Negl. Trop. Dis. 3, e367.

Uddin, M.H., Li, S., Bae, Y.M., Choi, M.H., Hong, S.T., 2012. Strain variation in the susceptibility and immune response to *Clonorchis sinensis* infection in mice. Parasitol. Int. 61, 118-123.

Van De, N., Le, T.H., Murrell, K.D., 2012. Prevalence and intensity of fish-borne zoonotic trematodes in cultured freshwater fish from rural and urban areas of northern Vietnam. J. Parasitol. 98, 1023-1025

Vanhove, M.P., Tessens, B., Schoelinck, C., Jondelius, U., Littlewood, D.T., Artois, T., Huyse, T., 2013. Problematic barcoding in flatworms: A case-study on monogeneans and rhabdocoels (Platyhelminthes). Zookeys, 355-379.

Vilas, R., Criscione, C.D., Blouin, M.S., 2005. A comparison between mitochondrial DNA and the ribosomal internal transcribed regions in prospecting for cryptic species of platyhelminth parasites. Parasitology 131, 839-846.

Vilas, R., Vazquez-Prieto, S., Paniagua, E., 2012. Contrasting patterns of population genetic structure of *Fasciola hepatica* from cattle and sheep: implications for the evolution of anthelmintic resistance. Infect. Genet. Evol. 12, 45-52.

Vinh, H.Q., Phimpraphai, W., Tangkawattana, S., Smith, J.F., Kaewkes, S., Dung, D.T., Duong, T.T., Sripa, B., 2017. Risk factors for *Clonorchis sinensis* infection transmission in humans in northern Vietnam: A descriptive and social network analysis study. Parasitol. Int. 66, 74-82.

Walker, S.M., Johnston, C., Hoey, E.M., Fairweather, I., Borgsteede, F., Gaasenbeek, C., Prodohl, P.A., Trudgett, A., 2011. Population dynamics of the liver fluke, *Fasciola hepatica*: the effect of time and spatial separation on the genetic diversity of fluke populations in the Netherlands. Parasitology 138, 215-223.

Wang, C., Lei, H., Tian, Y., Shang, M., Wu, Y., Li, Y., Zhao, L., Shi, M., Tang, X., Chen, T., Lv, Z., Huang, Y., Tang, X., Yu, X., Li, X., 2017a. *Clonorchis sinensis* granulin: identification, immunolocalization, and function in promoting the metastasis of cholangiocarcinoma and hepatocellular carcinoma. Parasit. Vectors 10, 262.

Wang, D., Young, N.D., Koehler, A.V., Tan, P., Sohn, W.M., Korhonen, P.K., Gasser, R.B., 2017b. Mitochondrial genomic comparison of *Clonorchis sinensis* from South Korea with other isolates of this species. Infect. Genet. Evol. 51, 160-166.

Wang, D., Korhonen, P.K., Gasser, R.B., Young, N.D., 2018. Improved genomic resources and new bioinformatic workflow for the carcinogenic parasite *Clonorchis sinensis*: Biotechnological implications. Biotechnol. Adv. 36, 894-904.

Wang, Q.P., Chen, X.G., Lun, Z.R., 2007. Invasive freshwater snail, China. Emerg. Infect. Dis. 13, 1119-1120.

Wang, X., Chen, ., Huang, Y., Sun, J., Men, J., Liu, H., Luo, F., Guo, L., Lv, X., Deng, C., Zhou, C., Fan, Y., Li, X., Huang, L., Hu, Y., Liang, C., Hu, X., Xu, J., Yu, X., 2011. The draft genome of the carcinogenic human liver fluke *Clonorchis sinensis*. Genome Biol. 12, R107.

Wang, X.Y., Liang, C., Chen, W.J., Fan, Y.X., Hu, X.C., Xu, J., Yu, X.B., 2009. Experimental model in rats for study on transmission dynamics and evaluation of *Clonorchis sinensis* infection immunologically, morphologically, and pathologically. Parasitol. Res. 106, 15-21.

WHO, 2013. Sustaining the Drive to Overcome the Global Impact of Neglected Tropical Diseases: Second WHO Report on Neglected Tropical Diseases. WHO.

Won, J., Ju, J.W., Kim, S.M., Shin, Y., Chung, S., Pak, J.H., 2014. *Clonorchis sinensis* infestation promotes three-dimensional aggregation and invasion of cholangiocarcinoma cells. PLoS One 9, e110705.

Wu, Y., Li, Y., Shang, M., Jian, Y., Wang, C., Bardeesi, A.S., Li, Z., Chen, T., Zhao, L., Zhou, L., He, A., Huang, Y., Lv, Z., Yu, X., Li, X., 2017. Secreted phospholipase A2 of *Clonorchis sinensis* activates hepatic stellate cells through a pathway involving JNK signalling. Parasit. Vectors 10, 147.

Wu, Z.B., Tian, H.S., Zeng, Y.N., 1980. Study of the ancient corpse of the Western Han Dynasty unearthed from tomb no. 168 on Phoenix Hill at Jiangling county. Wuhan Yi Xue Yuan Xue Bao 1, 1-10 (in Chinese).

Xiao, J.Y., Gao, J.F., Cai, L.S., Dai, Y., Yang, C.J., Luo, L., Agatsuma, T., Wang, C.R., 2013. Genetic variation among *Clonorchis sinensis* isolates from different hosts and geographical locations revealed by sequence analysis of mitochondrial and ribosomal DNA regions. Mitochondrial DNA 24, 559-564.

Xiao, S.H., Xue, J.A., Xu, L.L., Zhang, Y.N., Qiang, H.Q., 2011. Comparative effect of mebendazole, albendazole, tribendimidine, and praziquantel in treatment of rats infected with *Clonorchis sinensis*. Parasitol. Res. 108, 723-730.

Xu, L.L., Jiang, B., Duan, J.H., Zhuang, S.F., Liu, Y.C., Zhu, S.Q., Zhang, L.P., Zhang, H.B., Xiao, S.H., Zhou, X.N., 2014. Efficacy and safety of praziquantel, tribendimidine and mebendazole in patients with co-infection of *Clonorchis sinensis* and other helminths. PLoS Negl. Trop. Dis. 8, e3046.

Yan, C., Li, X.Y., Li, B., Zhang, B.B., Xu, J.T., Hua, H., Yu, Q., Liu, Z.Z., Fu, L.L., Tang, R.X., Zheng, K.Y., 2015. Expression of Toll-like receptor (TLR) 2 and TLR4 in the livers of mice infected by *Clonorchis sinensis*. J. Infect. Dev. Ctries. 9, 1147-1155.

Yang, L., Gui, A., Zuo, S., Song, X., 1994. The continuous observation of freshwater snails infected by cercaria of *Clonorchis sinensis* in Hubei province. Chin. J. Schistosomiasis Control 6, 150–151 (in Chinese).

Ye, X.P., Fu, Y.L., Wu, Z.X., Anderson, R.M., Agnew, A., 1997. The effects of temperature, light and water upon the hatching of the ova of *Schistosoma japonicum*. Southeast Asian J. Trop. Med. Public Health 28, 575-580.

Yen, C.M., Chen, E.R., Hou, M.F., Chang, J.H., 1992. Antibodies of different Immunoglobulin isotypes in serum and bile of patients with clonorchiasis. Ann. Trop. Med. Parasitol. 86, 263-269.

Yoo, W.G., Kim, D.W., Ju, J.W., Cho, P.Y., Kim, T.I., Cho, S.H., Choi, S.H., Park, H.S., Kim, T.S., Hong, S.J., 2011. Developmental transcriptomic features of the carcinogenic liver fluke, *Clonorchis sinensis*. PLoS Negl. Trop. Dis. 5, e1208.

Yoshida, Y., 2012. Clonorchiasis--a historical review of contributions of Japanese parasitologists. Parasitol. Int. 61, 5-9.

You, Y., Sun, K., Xu, L., Wang, L., Jiang, T., Liu, S., Lu, G., Berquist, S.W., Feng, J., 2010. Pleistocene glacial cycle effects on the phylogeography of the Chinese endemic bat species, *Myotis davidii*. BMC Evol. Biol. 10, 208.

Young, N.D., Campbell, B.E., Hall, R.S., Jex, A.R., Cantacessi, C., Laha, T., Sohn, W.M., Sripa, B., Loukas, A., Brindley, P.J., Gasser, R.B., 2010. Unlocking the transcriptomes of two carcinogenic parasites, *Clonorchis sinensis* and *Opisthorchis viverrini*. PLoS Negl. Trop. Dis. 4, e719.

Young, N.D., Chan, K.G., Korhonen, P.K., Min Chong, T., Ee, R., Mohandas, N., Koehler, A.V., Lim, Y.L., Hofmann, A., Jex, A.R., Qian, B., Chilton, N.B., Gobert, G.N., McManus, D.P., Tan, P., Webster, B.L., Rollinson, D., Gasser, R.B., 2015. Exploring molecular variation in *Schistosoma japonicum* in China. Sci. Rep. 5, 17345.

Zadesenets, K.S., Katokhin, A.V., Mordvinov, V.A., Rubtsov, N.B., 2012. Comparative cytogenetics of opisthorchid species (Trematoda, Opisthorchiidae). Parasitol. Int. 61, 87-89.

Zhang, F., Liang, P., Chen, W., Wang, X., Hu, Y., Liang, C., Sun, J., Huang, Y., Li, R., Li, X., Xu, J., Yu, X., 2013. Stage-specific expression, immunolocalization of *Clonorchis sinensis* lysophospholipase and its potential role in hepatic fibrosis. Parasitol. Res. 112, 737-749.

Zhang, X., Jin, Z., Da, R., Dong, Y., Song, W., Chen, X., Huang, Q., Ling, H., Che, Y., Li, Y., Zhang, F., 2008a. Fas/FasL-dependent apoptosis of hepatocytes induced in rat and patients with *Clonorchis sinensis* infection. Parasitol. Res. 103, 393-399.

Zhang, Y., Chang, Q.C., Zhang, Y., Na, L., Wang, W.T., Xu, W.W., Gao, D.Z., Liu, Z.X., Wang, C.R., Zhu, X.Q., 2014. Prevalence of *Clonorchis sinensis* infection in freshwater fishes in northeastern China. Vet. Parasitol. 204, 209-213.

Zheng, M., Hu, K., Liu, W., Hu, X., Hu, F., Huang, L., Wang, P., Hu, Y., Huang, Y., Li, W., Liang, C., Yin, X., He, Q., Yu, X., 2011. Proteomic analysis of excretory secretory products from *Clonorchis sinensis* adult worms: molecular characterization and serological reactivity of a excretory-secretory antigen-fructose-1,6-bisphosphatase. Parasitol. Res. 109, 737-744.

Zheng, S.H., Zhu, Y., Zhao, Z.J., Wu, Z.D., Okanurak, K., Lv, Z.Y., 2017. Liver fluke infection and cholangiocarcinoma: a review. Parasitol. Res. 116, 11-19.

Zhigileva, O.N., Ozhirelev, V.V., Stepanova, T.F., Moiseenko, T.I., 2014. Population structure of *Opisthorchis felineus* (Trematoda) and its second intermediate hosts - cyprinid fishes in the Ob-Irtysh focus of opisthorchiasis, based on allozyme data. Helminthologia 51, 309-317.

Zhou, B., Gu, L., Ding, Y., Shao, L., Wu, Z., Yang, X., Li, C., Li, Z., Wang, X., Cao, Y., Zeng, B., 2011. The great 2008 Chinese ice storm: its socioeconomic–ecological impact and sustainability lessons learned. Bull. Am. Meteorol. Soc. 92, 47-60.

Table 1-1 Species of freshwater snails presently recognised as first intermediate hosts of *Clonorchis sinensis*

Country	Region/province	Family	Species	References
Russia	Far East	Bithyniidae	*Parafossarulus manchouricus* (=*P. striatulus*), *P. spiridonovi*	Fattakhov et al. (2012); Chelomina et al. (2014); Köhler (2016)
		Semisulcospiridae	*Semisulcospira* (=*Koreoleptoxis*) *amurensis*[*]	
Japan		Bithyniidae	*P. manchouricus*	Yoshida (2012)
Republic of Korea		Bithyniidae	*P. manchouricus*	Choi (1984)
China	North-East (Heilongjiang, Liaoning)	Bithyniidae	*P. manchouricus*， *P. anomalospiralis*	Lun et al. (2005); Tang et al. (2016)
		Semisulcospiridae	*S. cancellata, S. amurensis*	
	Central China (Anhui, Henan, Hubei, Hunan, Jiangxi)	Bithyniidae	*Bithynia longicornis* (=*Alocinma longicornis*), *B. fuchsianus, B. misella, P. manchouricus, P. sinensis, P. anomalospiralis*	Lun et al. (2005); Tang et al. (2016)
		Semisulcospiridae	*S. cancellata*	
	South China (Guangdong, Guangxi)	Bithyniidae	*B. longicornis, B. fuchsianus, B. misella, P. manchouricus, P. sinensis*	Lun et al. (2005); Zhang et al. (2007); Tang et al. (2016)
		Semisulcospiridae	*S. cancellata*	
		Assimineidae	*Assiminea lutea*	
		Thiaridae	*Melanoides tuberculata*	
Vietnam	North	Bithyniidae	*P. manchouricus, B. siamensis, B. longicornis*[*], *B. fuchsiana*[*]	Chelomina et al. (2014); Doanh and Nawa (2016)
		Thiaridae	*M. tuberculata*	
Thailand		Thiaridae	*M. tuberculata*	Traub et al. (2009)

[*] Potential host only

Table 1-2 Families and genera of freshwater fishes presently recognised as secondary intermediate hosts of *Clonorchis sinensis*

Country/ region	Family	Genus	References
Russia	Bagridae; Cyprinidae	*Pseudobagrus; Tachysyrus; Liocassis*	Fattakhov et al. (2012)
Republic of Korea	Bagridae; Cyprinidae; Osmeridae; Osphronemidae; Percichthyidae; Pristigasteridae	*Abbottina; Acanthorhodeus; Acheilognathus; Aphyocypris; Carassius; Coreobagrus; Coreoleuciscus; Coreoperca; Culter; Cyprinus; Erythroculter; Gnathopogon; Hemiculter; Hemibarbus; Hypomesus; Ilisha; Macropodus; Microphysogobio; Opsariichthys; Phoxinus; Pseudogobio; Pseudorasbora; Puntungia; Rhodeus; Saurogobio; Sarcocheilichthys; Siniperca; Squaliobarbus; Squalidus; Tribolodon; Zacco;*	Sohn (2009)
Taiwan	Cyprinidae	*Hemiculter*	Ooi et al (1997)
China	Bagridae; Cichlidae; Channidae; Cobitidae; Cyprinidae; Gobiidae; Odontobutidae; Poeciliidae; Sebastidae; Siluridae	*Abbottina; Acrossocheilus; Carassius; Channa; Cirrhinus; Ctenogobius; Ctenopharyngodon; Cyprinus; Distoechodon; Erythroculter; Gambusia; Gobio; Hemiculter; Hemiculterella; Heros; Hypophthalmichthys; Misgurnus; Megalobrama; Mylopharyngodon; Opsariichthys; Oreochromis; Parabramis; Parasilurus; Pelteobagrus; Perccottus; Phoxinus; Pseudorasbora; Rhodeus; Saurogobio; Sebastiscus*	Lun et al. (2005); Tang et al. (2016)
Vietnam*	Cichlidae; Cyprinidae	*Anabas; Carassius; Cirrhina; Ctenopharyngodon; Cyprinus; Hypophthalmichthys; Mylopharyngodon; Tilapia*	Van De et al. (2012); Doanh and Nawa (2016)

* Reported as infected with *C. sinensis* - but may have been infected with metacercariae of multiple fish-borne trematodes

Table 1-3 Nuclear and mitochondrial genome loci used to characterise genetic variation in *Clonorchis sinensis*

Locus	Individuals sequenced for each country					References
	Russia	Japan	Republic of Korea	China	Vietnam	
In nuclear genome						
18S rRNA	-	-	15	3	-	Park and Yong (2001); Lee et al. (2004); Kang et al. (2008)
ITS-1	65	-	20	384	26	Lee et al. (2004); Kang et al. (2008); Sun et al. (2011); Tatonova et al. (2012); Shin et al. (2013); Sun et al. (2013); Tatonova et al. (2017)
5.8S	39	-	1	2	26	Tatonova et al. (2012); Tatonova et al. (2017)
ITS-2	39	1	2	3	26	Park and Yong (2001); Lee et al. (2004); Tatonova et al. (2012); Shin et al. (2013); Tatonova et al. (2017)
act	-	-	-	256	-	Sun et al. (2013)
tub	-	-	-	256	-	Sun et al. (2013)
ef-1*a*	-	-	-	256	-	Sun et al. (2013)
Microsatellite markers	-	-	5	5	5	Nguyen et al. (2015)
MGE and RAPD*	-	-	-	44	-	Lai et al. (2008)
Whole nuclear genome	-	-	1 pooled	1	-	Wang et al. (2011); Huang et al. (2013); Wang et al. (2018)
In mitochondrial genome						
*cox*1	40	1	2	318	28	Park and Yong (2001); Lee et al (2004); Liu et al. (2012); Shin et al. (2013); Sun et al. (2013); Tatonova et al. (2013); Xiao et al. (2013); Chelomina et al. (2014)
*cox*2	-	-	-	31	-	Liu et al. (2012)
*cox*3	-	-	-	257	-	Park and Yong (2001); Sun et al. (2013)
*nad*1	-	-	-	31	-	Liu et al. (2012)
*nad*2	-	-	-	59	-	Liu et al. (2012); Xiao et al. (2013)
*nad*4	-	-	-	256	-	Sun et al. (2013)
*nad*5	-	-	-	284	-	Sun et al. (2013); Xiao et al. (2013)
Whole mitochondrial genome	1 pooled	-	2 pooled	1 pooled	-	Shekhovtsov et al. (2010); Cai et al. (2012); Wang et al. (2017b)

*Mobile genetic elements (MGEs); random amplified polymorphic DNA (RAPD)

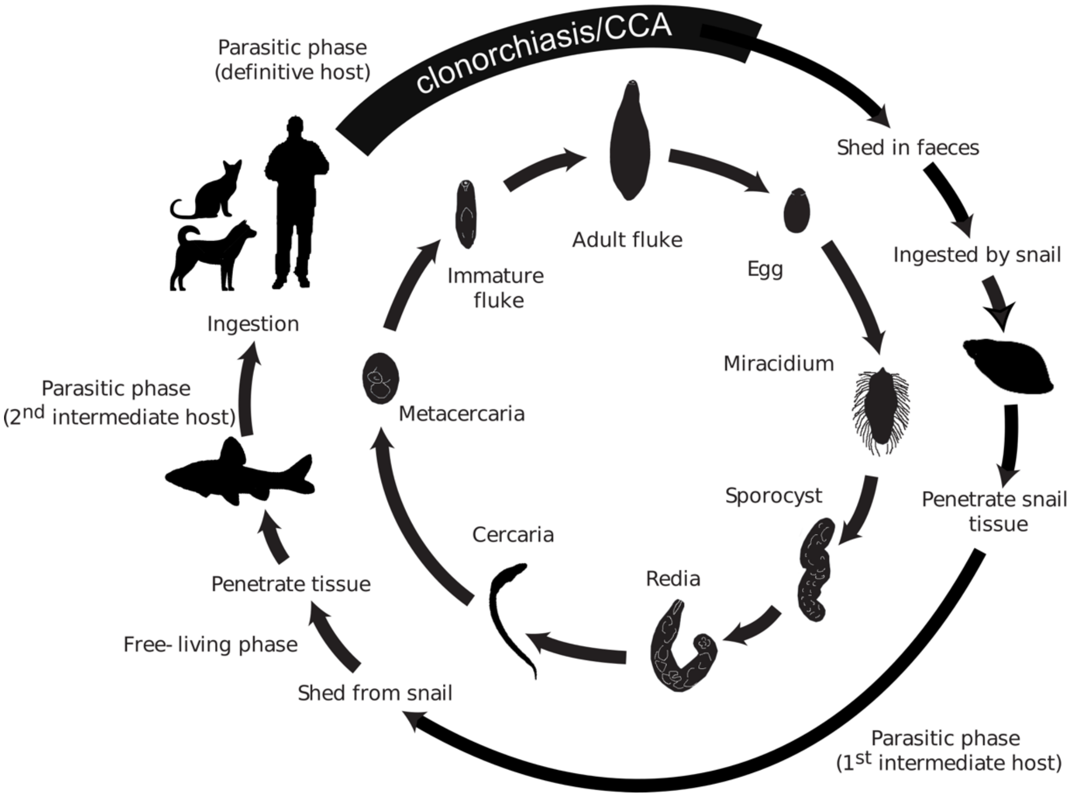
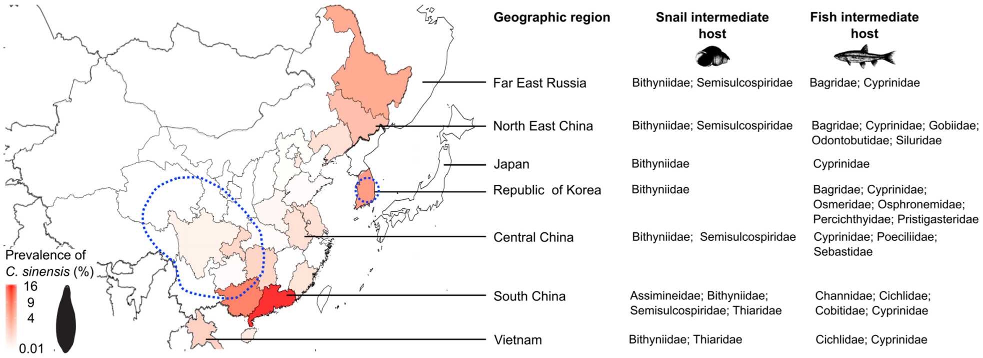


Fig. 1-1 Life cycle of *Clonorchis sinensis.* Eggs from adult flukes in the definitive host are released into the environment and then eaten by a first intermediate snail host (e.g., *Parafossarulus manchouricus*) in freshwater; following asexual replication, cercariae are shed from the snail into water, undergo host searching and then invade a freshwater fish (e.g., cyprinid fish); metacercariae in infected fish are ingested by a definitive host (human, cat or dogs), excyst and then develop to juveniles which mature into reproductively active hermaphroditic adults. CCA = cholangiocarcinoma.



Geographic region	Snail intermediate host	Fish intermediate host
Far East Russia	Bithyniidae; Semisulcospiridae	Bagridae; Cyprinidae
North East China	Bithyniidae; Semisulcospiridae	Bagridae; Cyprinidae; Gobiidae; Odontobutidae; Siluridae
Japan	Bithyniidae	Cyprinidae
Republic of Korea	Bithyniidae	Bagridae; Cyprinidae; Osmeridae; Osphronemidae; Percichthyidae; Pristigasteridae
Central China	Bithyniidae; Semisulcospiridae	Cyprinidae; Poeciliidae; Sebastidae
South China	Assimineidae; Bithyniidae; Semisulcospiridae; Thiaridae	Channidae; Cichlidae; Cobitidae; Cyprinidae
Vietnam	Bithyniidae; Thiaridae	Cichlidae; Cyprinidae

Fig. 1-2 The geographical distributions of intermediate hosts of *Clonorchis sinensis* in East Asia. This map indicates the distribution of the families of snail and fish intermediate hosts in relation to estimated prevalences (colour grading) of *Clonorchis sinensis* and regions in China, Vietnam and the Republic of Korea. Geographic boundaries are indicated by black lines; proposed refugia following glacial events are circumscribed by dotted blue lines (cf. Qiu et al., 2011).

Chapter 2 - Mitochondrial genomic comparison of *Clonorchis sinensis* from South Korea with other isolates of this species

Abstract

Clonorchiasis is a neglected tropical disease that affects more than 35 million people in China, Vietnam, South Korea, Thailand and parts of Russia. The disease-causing agent, *Clonorchis sinensis*, is a liver fluke of humans and other piscivorous animals, and has a complex aquatic life cycle involving snails and fish intermediate hosts. Chronic infection in humans causes liver disease and associated complications including malignant bile duct cancer. Central to control and to understanding the epidemiology of this disease is knowledge of the specific identity of the causative agent as well as genetic variation within and among populations of this parasite. Although most published molecular studies seem to suggest that *C. sinensis* represents a single species and that genetic variation within the species is limited, karyotypic variation within. *C. sinensis* among China, Korea (2n = 56) and Russian Far East (2n = 14) suggests that this taxon might contain sibling species. Here, we assessed and applied a deep sequencing-bioinformatic approach to sequence and define a reference mitochondrial (mt) genome for a particular isolate of *C. sinensis* from Korea (*Cs*-k2), to confirm its specific identity, and compared this mt genome with homologous data sets available for this species. Comparative analyses revealed consistency in the number and structure of genes as well as in the lengths of protein-encoding genes, and limited genetic variation among isolates of *C. sinensis.* Phylogenetic analyses of amino acid sequences predicted from mt genes showed that representatives of *C. sinensis* clustered together, with absolute nodal support, to the exclusion of other liver fluke representatives, but sub-structuring within *C. sinensis* was not well supported. The plan now is to proceed with the sequencing, assembly and annotation of a high-quality draft nuclear genome of this defined isolate (*Cs*-k2) as a basis for a detailed investigation of molecular variation within *C. sinensis* from disparate geographical locations in parts of Asia and to prospect for cryptic species.

2.1. Introduction

After the Director General of the World Health Organization (WHO), Dr Margaret Chan, called for urgent action to eliminate human suffering from neglected tropical diseases (NTDs) (WHO, 2013) and the London Declaration in 2012 (http://www.globalnetwork.org/london-declaration), there has been a major resolve to understand and tackle problems associated with NTDs, which are some of the most insidious and chronic illnesses of mankind, affecting billions of people worldwide (Molyneux et al., 2016). Amongst these diseases are opisthorchiasis and clonorchiasis, which are caused by liver flukes of the family Opisthorchiidae, with tens of millions of people affected, and ~600 million people at risk of infection (Keiser and Utzinger, 2005; Sripa et al., 2007; Young et al., 2010).

Clonorchiasis is caused by a chronic infection by *Clonorchis sinensis* and, alone, is estimated to affect at least 35 million people predominantly in China, Vietnam, Korea, Thailand and some parts of Russia (Lun et al., 2005; Qian et al., 2016). This disease is linked to cholangitis and associated problems in the liver, most importantly, malignant bile duct cancer (cholangiocarcinoma) (Lun et al., 2005). Humans act as definitive hosts of *C. sinensis* (as do canids and felids), and become infected by eating raw cyprinoid fish containing a larval (metacercarial) stage of the parasite (cf. Young et al., 2010). Although this disease could be prevented if people simply consumed cooked instead of raw fish, it persists in endemic regions due to people's cultural/eating habits (Qian et al., 2016). There is no anti-*C. sinensis* vaccine, and the treatment of infected people relies on the use of the anthelmintic praziquantel. In spite of having insights into some aspects of clonorchiasis (Lun et al., 2005; Qian et al., 2016), little is known about the epidemiology of this disease complex.

Central to understanding the epidemiology of this disease is knowledge of the specific identity of the causative agent (*C. sinensis*), being able to track the transmission of this pathogen through different hosts, and genetic variation within and among populations of this parasite. However, it is not possible to unequivocally identify the species by microscopy, because some developmental stages, such as eggs and metacercariae, cannot be differentiated reliably from those of some other fluke species (e.g., *Opisthorchis viverinni*) (cf. Le et al., 2006). Currently, most published molecular evidence seems to suggest that *C. sinensis* represents a single species and that genetic variation within the species is relatively low (Park and Yong, 2001; Lee and Huh, 2004; Petney et al., 2013). However, to date, most studies have used a relatively small number of genetic loci for specific identification (e.g., internal transcribed spacers of nuclear ribosomal DNA) and/or population genetic studies (e.g., *act*, *tub*, *ef-1a*, *cox*1, *cox*3, *nad*4 and *nad*5) (Park and Yong, 2001; Lee and Huh, 2004; Liu et al., 2007; Park, 2007; Katokhin et al., 2008; Lai et al., 2008; Shekhovtsov et al., 2009; Liu et al., 2012b; Shin et al., 2013; Sun et al., 2013; Tatonova et al., 2013; Xiao et al., 2013; Chelomina et al., 2014), and it is not yet known whether "cryptic" (= morphologically similar, but genetically distinct) species exist within *C. sinensis*. Nonetheless, karyotypic differences in *C. sinensis* observed among China, Korea (2n = 56) and the Russian Far East (2n = 14) (Park and Yong, 2001; Zadesenets et al., 2012) do suggest that this might be the case.

Relatively recently, three mitochondrial (mt) genomes were sequenced from single isolates of *C. sinensis* from Russia (Shekhovtsov et al., 2010), China and Korea (Cai et al., 2012) as a possible source of genetic markers to undertake future systematic and/or population genetic investigations. In the latter study, Cai et al. (2012) utilised their data sets (concatenated amino acid sequence data for 12 mt protein-coding genes) to infer the phylogenetic relationships among selected parasitic trematodes. In this published analysis, *C. sinensis* samples from China and Korea were more closely related to one another than either was to *C. sinensis* from Russia. Although the extent of variability within *C. sinensis* appears to be low, there have been no comprehensive studies to date to explore population genetic variation among relatively large numbers of isolates from disparate geographic locations using complete or near complete mt and/or nuclear genomic data sets. Nonetheless, such studies have been conducted on other flatworm parasites, such as the blood flukes *Schistosoma japonicum* and *S. mansoni*, using advanced deep sequencing and bioinformatics (Young et al., 2015; Crellen et al., 2016). In the present study, our goal was to assess and apply a deep sequencing-bioinformatic approach to sequence and define a reference mt genome for a particular isolate of *C. sinensis* from Korea, in order to confirm the specific identity of the parasite, and to compare this reference mt genome to all mt genomic data sets currently available for this and selected species of trematodes, as a basis for a nuclear genomic sequencing project and a future global analysis of genetic variation within and among isolates from parts of Southeast Asia.

2.2. Materials and methods

Isolation and procurement of the parasite

Metacercariae of *C. sinensis* were isolated from tissues from naturally infected cyprinoid fish, *Pseudorasbora parva*, in the Jinju-si, South Gyeongsang Province, South Korea. This isolate was designated *Cs*-k2 and procured using established methods (Sohn et al., 2006). In brief, the fish were ground and digested in hydrochloric acid (0.1 M HCl, pH 1.5) containing 0.01% pepsin (Sigma) for 2 h at 37°C, and the metacercariae were isolated by sieving (0.5 mm aperture), washing and sedimentation in physiological saline. Although the specific identity of metacercariae could not be unequivocally established by light microscopy (40-times magnification), their size and shape were consistent with those of *C. sinensis* (Kobayashi, 1917). Helminth-free, inbred Syrian golden hamsters (*Mesocricetus auratus*) were infected with metacercariae (n ~ 50) as described previously (Chung and Choi, 1988; Sohn et al., 2006), in accordance with protocols approved by the animal ethics committee of Gyeongsang National University. Eight weeks after infection, adult worms were collected from the bile ducts from hamster livers and then briefly cultured *in vitro* to allow the worms to regurgitate caecal contents using an established technique (Young et al., 2010). Then, worms were washed extensively in phosphate-buffered saline (pH 7.4) and frozen at -80ºC.

Sequencing of total genomic DNA, and assembly and annotation of the mt genome

High molecular weight genomic DNA was isolated from a pool of 95 adults of *C. sinensis* using an established protocol (Sambrook, 1989). The total DNA amount was determined using a Qubit fluorometer dsDNA HS kit (Invitrogen), according to the manufacturer's instructions. Genomic DNA integrity was verified by agarose gel electrophoresis and using a BioAnalyzer (2100, Agilent). A paired-end genomic library (500 bp insert size) was built, and assessed for both size distribution and quality also employing the BioAnalyzer. All sequencing was carried out on the HiSeq 2000 sequencing platform (Illumina). The sequence data generated from the library were verified, and low-quality sequences, base-calling duplicates and adapters removed (Li et al., 2010). In total, 95 million reads were generated and exported to FASTQ (Cock et al., 2010). Several steps were taken to enforce read quality. Custom Perl scripts were used to trim the last nucleotide from each read; nucleotides with a quality score of < 3, and Ns were removed. Quality-trimmed reads were retained if they were ⩾ 90 nt in length. Each set of quality-filtered, and paired-end reads were mapped to a reference mt genome (GeneBank accession no. JF729304.1; Cai et al., 2012) using the program Bowtie2 (Langmead and Salzberg, 2012), using the default parameters for a gap penalty, with a maximum edit distance of 0.1 (allowing for 10% mismatch). Custom scripts were used to extract individual read pairs, which were then merged to form a paired-end interleaved FASTQ file. Reads were then assembled independently using the program Velvet (Zerbino and Birney, 2008), employing *k*-mers of 17-79 bp for de Bruijn graph construction. All paired-end reads were then aligned to the contigs to achieve scaffolding, and any remaining gaps were closed.

Sequences with nucleotide homology to a previously published mt genome from *C. sinensis* from Korean (*Cs*-k1; accession no. JF729304.1; Cai et al., 2012) were identified by BLASTn (default settings, E-value: 10^{-5}). Scaffolds containing parts of the mt genome were assessed for completeness and then trimmed to ensure that only a single copy of the mt genome remained. Then, mt protein-encoding genes, large and small subunits of the mt ribosomal (r) RNA genes (*rrn*S and *rrn*L, respectively) and the transfer (t) RNA genes were identified by pairwise alignment with the annotated mt genome of *Cs*-k1 using the program MUSCLE v.3.7 (Edgar, 2004). Annotated sequence data were imported using the program SEQUIN (available *via* http://www.ncbi.nlm.nih.gov/Sequin/) for the final verification of the mt genome organization/annotation prior to submission to the GenBank database.

Sliding window analysis

This analysis was performed on the aligned mt genome sequences of the four isolates of *C. sinensis* (Table 2-1) using an R package, “PopGenome” (Pfeifer et al., 2014). The sequences were first aligned using MUSCLE v.3.7 (Edgar, 2004); keeping the nucleotides in frame, there were no ambiguously aligned regions. A sliding window of 300 bp (steps of 10 bp) was used to estimate nucleotide diversity (π) (Nei and Li, 1979) among four members (pairwise) of the *C. sinensis*. Nucleotide diversity for the alignments was plotted against midpoint positions of each window, and gene boundaries were defined. Separating the analyses in this way allowed a pairwise comparison of general patterns within *C. sinensis*, in order to highlight conserved regions and other areas with potential for the definition of additional mt genetic markers with low, medium or high variability among representative isolates.

Phylogenetic analysis

The amino acid sequences conceptually translated from the mt genome of *Cs*-k2 were aligned with those predicted from other, publicly available mt genomes representing *C. sinensis* from Korea (*Cs*-k1), China (*Cs*-c1) and Russia (*Cs*-r1) (accession nos. JF729304.1, JF729303.1 and FJ381664.2, respectively; Cai et al., 2012; Shekhovtsov et al., 2010) and other selected trematodes (liver flukes; Table 2-1) using MUSCLE v3.7 (Edgar, 2004). Finally, all aligned sequence blocks were concatenated, and alignments assessed by eye. The optimal protein evolutionary model for each sequence was then assessed using the program ProtTest3.4.2 (Darriba et al., 2011) employing default settings. The concatenated, aligned sequences were then subjected to phylogenetic analysis using Bayesian inference (BI) and maximum likelihood (ML) methods. BI analysis was conducted employing Monte Carlo Markov Chain analysis in the program MrBayes v.3.2.2 (Ronquist et al., 2012); the optimal model for each partition of the concatenated sequences was applied in the inference, involving four chains and 2,000,000 iterations, sampling every 100th iteration; the first 500,000 iterations were removed from the analysis as burn-in. ML analysis was carried out using the RAxML program (Stamatakis, 2014); the optimal amino acid substitution model inferred using the program ProtTest3.4.2 for each data partition was used in the 'PROTGAMMAGTR' option for 1000 bootstrap replicates. The unrooted trees were viewed and drawn using the programs FigTree (http://tree.bio.ed.ac.uk/software/figtree/) and modified utilising the program PowerPoint (https://products.office.com/en-au/powerpoint).

2.3. Results and discussion

The circular mt genome assembled for *C. sinensis* from South Korea (*Cs*-k2) was 13,877 bp in size. By comparison with another representative sequence for *C. sinensis* from Korea (*Cs*-k1; accession no. JF729304), we identified 36 genes (Table 2-2), including 12 protein-coding genes (*cox*1-3, *nad*1-6, *nad*4L, *atp*6 and *cyt*b), 22 tRNA genes, two rRNA genes and two non-coding regions. As expected, this mt genome lacks an *atp*8 gene, consistent with the other three previously published mt genomes representing *C. sinensis* from Korea, China and Russia (Cai et al., 2012; Shekhovtsov et al., 2010). All protein-coding genes have an ATG or GTG as an initiation codon (encoding for methionine and valine) and a TAG or TAA as a termination codon. The mt gene order is consistent with that of the Opisthorchiidae, Fasciolidae and Paragonimidae (cf. Le et al., 2000; Le et al., 2002; Cai et al., 2012). The nucleotide content of the mt genomic sequence is biased toward A+T (60%), with T (43%) being the most favoured nucleotide, and C (12.5%) the least favoured, in accord with mt genomes of some other trematodes sequenced to date (Le et al., 2002).

Having established the features of the present mt genome of *C. sinensis* (*Cs*-k2), we established levels of nucleotide variation (%) along the whole mt genome between *C. sinensis* from Korea (accession no. KY564177) and other distinct isolates from Korea (*Cs*-k1), China (*Cs*-c1) and Russia (*Cs*-r1) (accession nos. JF729304, JF729303 and FJ381664, respectively; Cai et al., 2012; Shekhovtsov et al., 2010). Sliding window analyses (Fig. 2-1) showed that the present isolate (*Cs*-k2) of *C. sinensis* from Korea is genetically most distinct from the sample (*Cs*-c1) from China (mean π = 0.0051), followed by the isolate (*Cs*-r1) from Russia (FJ381664) (mean π = 0.0027), followed by another sample (*Cs*-k1) from Korea (JF729304) (mean π = 0.0023). In total, there were 96 variable sites, 74 of which were located in protein-encoding genes, 24 of which were non-synonymous alterations (Table 2-3). In the 12 protein coding genes, *nad*5 had most non-synonymous sites (*n* = 4), while *atp*6*, nad*1 and *nad*3 had only one non-synonymous site each.

Using available mt genomic data sets (see Table 2-1), we were able to assess the genetic relationships of the four distinct isolates of *C. sinensis*, including in the analyses sequences of eight selected species of trematodes (liver flukes) for comparative purposes (Fig. 2-2). First, we assessed phylogenetic 'informativeness' at individual positions of the aligned, concatenated amino acid sequences. At the amino acid level, 2,026 of 3,441 alignment positions were phylogenetically informative and 1,264 (36.7%) were invariable (Table 2-4). Phylogenetic trees constructed using amino acid sequence data employing the BI and ML methods showed that all four *C. sinensis* isolates (*Cs*-k2, *Cs*-k1, *Cs*-c1 and *Cs*-r1) clustered together, with absolute nodal support, to the exclusion of other liver fluke representatives. Nonetheless, the clustering within *C. sinensis* was not well supported (nodal support: <0.5 or <50%) in the analyses utilising each of the two tree-building methods, precluding further interpretation (Fig. 2-2).

Taken together, the comparison of mt genomes of *C. sinensis* in the present study showed consistency in the number and structure of genes as well as the lengths of protein-encoding genes, as reported previously for this trematode (cf. Shekhovtsov et al., 2010; Cai et al., 2012). Comparative analyses of the protein-coding genes showed limited or no sequence variation (mean: 0-0.21%) in *nad*4L and *nad*4, moderate variability (0.32-0.58%) in *atp*6, *cox*1, *cox*3, *cyt*b, *nad*1, *nad*2, *nad*3 and *nad*5, and most variability (0.66-0.76%) in *cox*2 and *nad*6 (Table 2-4). This information suggests that most genes (used together or individually) should be suited for diagnostic applications (i.e. specific identification and/or differentiation), but the levels of nucleotide variation and phylogenetic signal established for the latter two most variable genes alone are inadequate to establish genetic relationships with high statistical support. Therefore, it is still unclear whether these two genes will find some continued utility for molecular epidemiological and/or genetic studies of *C. sinensis.* These findings suggest that future work should focus on nuclear genomic markers for such investigations.

From a technical perspective, the deep sequencing-assembly-annotation approach taken in this study allowed us to rapidly characterise the mt genome of a *C. sinensis* isolate, whose specific identity required molecular confirmation. Considering labour, time as well as laboratory consumable- and salary costs, in our experience, this procedure is more cost effective to carry out than using PCR and/or cloning methods (cf. Shekhovtsov et al., 2010; Cai et al., 2012;). In addition, deep sequencing has the advantage that it achieves much higher genome coverage (>100 times) than the Sanger sequencing of amplicons, and allows the ready detection of nucleotide and amino acid sequence variability within and among samples. Our aim was to produce a (consensus) mt reference sequence for *C. sinensis* from Korea (*Cs*-k2), in which dominant nucleotides were recorded at individual sequence positions. Although some nucleotide variation was detected, it was subtle, consistent with limited sequence diversity recorded in mt DNA within most species of trematodes studied to date (Le et al., 2002; Li et al., 2010). Minor within-species nucleotide variability might relate to distinct substitution rates in mt genomes of distinct mitochondrion populations in various tissues of the worms or mt sequence differences among individual specimens within the pool of worms used here to prepare DNA for sequencing. However, establishing nucleotide variability within isolates (often containing tens to hundreds of individuals) is considered less critical than defining positions of unambiguous nucleotide difference by comparison to sequences from other isolates. However, for any future population genetic or molecular epidemiological studies of *C. sinensis*, careful consideration must be given to "background" nucleotide variability that might occur within isolates containing multiple worms, such that no incorrect conclusions are made regarding nucleotide substitution rates in mt DNA.

In conclusion, given the limited extent of variability in mt genomic sequences between/among isolates of *C. sinensis* studied thus far, it will be important to explore levels of variation in nuclear DNA on a global scale. Although a draft nuclear genome of *C. sinensis* from China is publicly accessible (Wang et al., 2011; Huang et al., 2013), there is a need to produce a higher quality nuclear genome reference sequence, to enable direct genome-genome comparisons and future applications of available genomic data sets. Therefore, now that we have verified the specific identity of the present isolate of *C. sinensis* from Korea (*Cs*-k2), we are now in a good position to proceed with the sequencing, assembly and annotation of a high-quality draft nuclear genome of this isolate, as a basis for a profound investigation of molecular variation within *C. sinensis* from disparate geographical locations in one or more parts of Asia or to prospect for cryptic species.

2.4. References

Cai, X.Q., Liu, G.H., Song, H.Q., Wu, C.Y., Zou, F.C., Yan, H.K., Yuan, Z.G., Lin, R.Q., Zhu, X.Q., 2012. Sequences and gene organization of the mitochondrial genomes of the liver flukes *Opisthorchis viverrini* and *Clonorchis sinensis* (Trematoda). Parasitol. Res. 110, 235-243.

Chelomina, G.N., Tatonova, Y.V., Hung, N.M., Ngo, H.D., 2014. Genetic diversity of the Chinese liver fluke *Clonorchis sinensis* from Russia and Vietnam. Int. J. Parasitol. 44, 795-810.

Chung, D.I., Choi, D.W., 1988. Intensity of infection and development of adult *Clonorchis sinensis* in hamsters. Kisaengchunghak Chapchi 26, 9-14.

Cock, P.J.A., Fields, C.J., Goto, N., Heuer, M.L., Rice, P.M., 2010. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Res. 38, 1767-1771.

Crellen, T., Allan, F., David, S., Durrant, C., Huckvale, T., Holroyd, N., Emery, A.M., Rollinson, D., Aanensen, D.M., Berriman, M., Webster, J.P., Cotton, J.A., 2016. Whole genome resequencing of the human parasite *Schistosoma mansoni* reveals population history and effects of selection. Sci. Rep. 6, 20954.

Darriba, D., Taboada, G.L., Doallo, R., Posada, D., 2011. ProtTest 3: fast selection of best-fit models of protein evolution. Bioinformatics 27, 1164-1165.

Edgar, R.C., 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32, 1792-1797.

Huang, Y., Chen, W., Wang, X., Liu, H., Chen, Y., Guo, L., Luo, F., Sun, J., Mao, Q., Liang, P., Xie, Z., Zhou, C., Tian, Y., Lv, X., Huang, L., Zhou, J., Hu, Y., Li, R., Zhang, F., Lei, H., Li, W., Hu, X., Liang, C., Xu, J., Li, X., Yu, X., 2013. The carcinogenic liver fluke, *Clonorchis sinensis*: new assembly, reannotation and analysis of the genome and characterization of tissue transcriptomes. PLoS One 8, e54732.

Katokhin, A.V., Shekhovtsov, S.V., Konkow, S., Yurlova, N.I., Serbina, E.A., Vodianitskaia, S.N., Fedorov, K.P., Loktev, V.B., Muratov, I.V., Ohyama, F., Makhneva, T.V., Pel'tek, S.E., Mordvinov, V.A., 2008. Assessment of the genetic distinctions of *Opisthorchis felineus* from *O. viverrini* and *Clonorchis sinensis* by ITS2 and CO1 sequences. Dokl. Biochem. Biophys. 421, 214-217.

Keiser, J., Utzinger, J., 2005. Emerging foodborne trematodiasis. Emerg. Infect. Dis. 11, 1507-1514.

Kobayashi, H., 1917. On the life history and morphology of the liver distoma (*Clonorchis sinensis*). Mitt. Med. Hochsch. Keijo 1, 251-289.

Lai, D.H., Wang, Q.P., Chen, W., Cai, L.S., Wu, Z.D., Zhu, X.Q., Lun, Z.R., 2008. Molecular genetic profiles among individual *Clonorchis sinensis* adults collected from cats in two geographic regions of China revealed by RAPD and MGE-PCR methods. Acta Trop. 107, 213-216.

Langmead, B., Salzberg, S.L., 2012. Fast gapped-read alignment with Bowtie 2. Nat. Methods 9, 357-359.

Le, T.H., Blair, D., McManus, D.P., 2000. Mitochondrial genomes of human helminths and their use as markers in population genetics and phylogeny. Acta Trop. 77, 243-256.

Le, T.H., Blair, D., McManus, D.P., 2001. Complete DNA sequence and gene organization of the mitochondrial genome of the liverfluke, *Fasciola hepatica* L. (Platyhelminthes; Trematoda). Parasitology 123, 609-621.

Le, T.H., Blair, D., McManus, D.P., 2002. Mitochondrial genomes of parasitic flatworms. Trends Parasitol. 18, 206-213.

Le, T.H., Van De, N., Blair, D., Sithithaworn, P., McManus, D.P., 2006. *Clonorchis sinensis* and *Opisthorchis viverrini:* development of a mitochondrial-based multiplex PCR for their identification and discrimination. Exp. Parasitol. 112, 109-114.

Lee, S.U., Huh, S., 2004. Variation of nuclear and mitochondrial DNAs in Korean and Chinese isolates of *Clonorchis sinensis*. Korean J. Parasitol. 42, 145-148.

Li, J., Zhao, G.H., Zou, F.C., Mo, X.H., Yuan, Z.G., Ai, L., Li, H.L., Weng, Y.B., Lin, R.Q., Zhu, X.Q., 2010a. Combined mitochondrial 16S and 12S rDNA sequences: an effective genetic marker for inter-species phylogenetic analysis of zoonotic trematodes. Parasitol. Res. 107, 561-569.

Li, R., Fan, W., Tian, G., Zhu, H., He, L., Cai, J., Huang, Q., Cai, Q., Li, B., Bai, Y., Zhang, Z., Zhang, Y., Wang, W., Li, J., Wei, F., Li, H., Jian, M., Li, J., Zhang, Z., Nielsen, R., Li, D., Gu, W., Yang, Z., Xuan, Z., Ryder, O.A., Leung, F.C., Zhou, Y., Cao, J., Sun, X., Fu, Y., Fang, X., Guo, X., Wang, B., Hou, R., Shen, F., Mu, B., Ni, P., Lin, R., Qian, W., Wang, G., Yu, C., Nie, W., Wang, J., Wu, Z., Liang, H., Min, J., Wu, Q., Cheng, S., Ruan, J., Wang, M., Shi, Z., Wen, M., Liu, B., Ren, X., Zheng, H., Dong, D., Cook, K., Shan, G., Zhang, H., Kosiol, C., Xie, X., Lu, Z., Zheng, H., Li, Y., Steiner, C.C., Lam, T.T., Lin, S., Zhang, Q., Li, G., Tian, J., Gong, T., Liu, H., Zhang, D., Fang, L., Ye, C., Zhang, J., Hu, W., Xu, A., Ren, Y., Zhang, G., Bruford, M.W., Li, Q., Ma, L., Guo, Y., An, N., Hu, Y., Zheng, Y., Shi, Y., Li, Z., Liu, Q., Chen, Y., Zhao, J., Qu, N., Zhao, S., Tian, F., Wang, X., Wang, H., Xu, L., Liu, X., Vinar, T., Wang, Y., Lam, T.W., Yiu, S.M., Liu, S., Zhang, H., Li, D., Huang, Y., Wang, X., Yang, G., Jiang, Z., Wang, J., Qin, N., Li, L., Li, J., Bolund, L., Kristiansen, K., Wong, G.K., Olson, M., Zhang, X., Li, S., Yang, H., Wang, J., Wang, J., 2010b. The sequence and *de novo* assembly of the giant panda genome. Nature 463, 311-317.

Liu, G.H., Gasser, R.B., Young, N.D., Song, H.Q., Ai, L., Zhu, X.Q., 2014a. Complete mitochondrial genomes of the 'intermediate form' of *Fasciola* and *Fasciola gigantica*, and their comparison with *F. hepatica*. Parasit. Vectors 7, 150.

Liu, G.H., Li, B., Li, J.Y., Song, H.Q., Lin, R.Q., Cai, X.Q., Zou, F.C., Yan, H.K., Yuan, Z.G., Zhou, D.H., Zhu, X.Q., 2012. Genetic variation among *Clonorchis sinensis* isolates from different geographic regions in China revealed by sequence analyses of four mitochondrial genes. J. Helminthol. 86, 479-484.

Liu, G.H., Yan, H.B., Otranto, D., Wang, X.Y., Zhao, G.H., Jia, W.Z., Zhu, X.Q., 2014b. *Dicrocoelium chinensis* and *Dicrocoelium dendriticum* (Trematoda: Digenea) are distinct lancet fluke species based on mitochondrial and nuclear ribosomal DNA sequences. Mol. Phylogenet. Evol. 79, 325-331.

Liu, W.Q., Liu, J., Zhang, J.H., Long, X.C., Lei, J.H., Li, Y.L., 2007. Comparison of ancient and modern *Clonorchis sinensis* based on ITS1 and ITS2 sequences. Acta Trop. 101, 91-94.

Lun, Z.R., Gasser, R.B., Lai, D.H., Li, A.-X., Zhu, X.Q., Yu, X.B., Fang, Y.Y., 2005. Clonorchiasis: a key foodborne zoonosis in China. Lancet Infect. Dis. 5, 31-41.

Ma, J., He, J.J., Liu, G.H., Leontovyc, R., Kasny, M., Zhu, X.Q., 2016. Complete mitochondrial genome of the giant liver fluke *Fascioloides magna* (Digenea: Fasciolidae) and its comparison with selected trematodes. Parasit. Vectors 9, 429.

Molyneux, D.H., Savioli, L., Engels, D., 2016. Neglected tropical diseases: progress towards addressing the chronic pandemic. Lancet. 389, 312-315.

Nei, M., Li, W.H., 1979. Mathematical-model for studying genetic-variation in terms of restriction endonucleases. Proc. Natl. Acad. Sci. U. S. A. 76, 5269-5273.

Park, G.M., 2007. Genetic comparison of liver flukes, *Clonorchis sinensis* and *Opisthorchis viverrini*, based on rDNA and mtDNA gene sequences. Parasitol. Res. 100, 351-357.

Park, G.M., Yong, T.S., 2001. Geographical variation of the liver fluke, *Clonorchis sinensis*, from Korea and China based on the karyotypes, zymodeme and DNA sequences. Southeast Asian J. Trop. Med. Public Health 32 (Suppl. 2), 12-16.

Petney, T.N., Andrews, R.H., Saijuntha, W., Wenz-Mucke, A., Sithithaworn, P., 2013. The zoonotic, fish-borne liver flukes *Clonorchis sinensis*, *Opisthorchis felineus* and *Opisthorchis viverrini*. Int. J. Parasitol. 43, 1031-1046.

Pfeifer, B., Wittelsburger, U., Ramos-Onsins, S.E., Lercher, M.J., 2014. PopGenome: an efficient Swiss army knife for population genomic analyses in R. Mol. Biol. Evol. 31, 1929-1936.

Qian, M.B., Utzinger, J., Keiser, J., Zhou, X.N., 2016. Clonorchiasis. The Lancet 387, 800-810.

Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D.L., Darling, A., Hohna, S., Larget, B., Liu, L., Suchard, M.A., Huelsenbeck, J.P., 2012. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Syst. Biol. 61, 539-542.

Sambrook, J., Fritch, E.F., Maniatis, T., 1989. Molecular Cloning: A Laboratory Manual, second ed. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Shekhovtsov, S.V., Katokhin, A.V., Kolchanov, N.A., Mordvinov, V.A., 2010. The complete mitochondrial genomes of the liver flukes *Opisthorchis felineus* and *Clonorchis sinensis* (Trematoda). Parasitol. Int. 59, 100-103.

Shekhovtsov, S.V., Katokhin, A.V., Romanov, K.V., Besprozvannykh, V.V., Fedorov, K.P., Yurlova, N.I., Serbina, E.A., Sithithaworn, P., Kolchanov, N.A., Mordvinov, V.A., 2009. A novel nuclear marker, Pm-int9, for phylogenetic studies of *Opisthorchis felineus*, *Opisthorchis viverrini*, and *Clonorchis sinensis* (Opisthorchiidae, Trematoda). Parasitol. Res. 106, 293-297.

Shin, D.H., Oh, C.S., Lee, H.J., Chai, J.Y., Lee, S.J., Hong, D.W., Lee, S.D., Seo, M., 2013. Ancient DNA analysis on *Clonorchis sinensis* eggs remained in samples from medieval Korean mummy. J. Archaeol. Sci. 40, 211-216.

Sohn, W.M., Zhang, H., Choi, M.H., Hong, S.T., 2006. Susceptibility of experimental animals to reinfection with *Clonorchis sinensis*. Korean J. Parasitol. 44, 163-166.

Sripa, B., Kaewkes, S., Sithithaworn, P., Mairiang, E., Laha, T., Smout, M., Pairojkul, C., Bhudhisawasdi, V., Tesana, S., Thinkamrop, B., Bethony, J.M., Loukas, A., Brindley, P.J., 2007. Liver fluke induces cholangiocarcinoma. PLoS Med. 4, e201.

Stamatakis, A., 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30, 1312-1313.

Sun, J., Huang, Y., Huang, H., Liang, P., Wang, X., Mao, Q., Men, J., Chen, W., Deng, C., Zhou, C., Lv, X., Zhou, J., Zhang, F., Li, R., Tian, Y., Lei, H., Liang, C., Hu, X., Xu, J., Li, X., Yu, X., 2013. Low divergence of *Clonorchis sinensis* in China based on multilocus analysis. PLoS One 8, e67006.

Tatonova, Y.V., Chelomina, G.N., Besprozvannykh, V.V., 2013. Genetic diversity of *Clonorchis sinensis* (Trematoda: Opisthorchiidae) in the Russian southern Far East based on mtDNA cox1 sequence variation. Folia Parasitol. (Praha) 60, 155-162.

Uniting to Combat NTDs, 2016. London Declaration on Neglected Tropical Diseases. http://unitingtocombatntds.org/sites/default/files/document/london_declaration_on_ntds.pdf (accessed Dec 7, 2016).

Wang, X., Chen, W., Huang, Y., Sun, J., Men, J., Liu, H., Luo, F., Guo, L., Lv, X., Deng, C., Zhou, C., Fan, Y., Li, X., Huang, L., Hu, Y., Liang, C., Hu, X., Xu, J., Yu, X., 2011. The draft genome of the carcinogenic human liver fluke *Clonorchis sinensis*. Genome Biol. 12, R107.

World Health Organization, 2012. Accelerating work to overcome the global impact of neglected tropical diseases: a roadmap for implementation. Geneva: World Health Organization; 2012 WHO/HTM/NTD/2012.1. http://www.who.int/neglected_diseases/NTD_RoadMap_2012_Fullversion.pdf (accessed Dec 7, 2016)

Xiao, J.Y., Gao, J.F., Cai, L.S., Dai, Y., Yang, C.J., Luo, L., Agatsuma, T., Wang, C.R., 2013. Genetic variation among *Clonorchis sinensis* isolates from different hosts and geographical locations revealed by sequence analysis of mitochondrial and ribosomal DNA regions. Mitochondrial DNA 24, 559-564.

Young, N.D., Campbell, B.E., Hall, R.S., Jex, A.R., Cantacessi, C., Laha, T., Sohn, W.M., Sripa, B., Loukas, A., Brindley, P.J., Gasser, R.B., 2010. Unlocking the transcriptomes of two carcinogenic parasites, *Clonorchis sinensis* and *Opisthorchis viverrini*. PLoS Negl. Trop. Dis. 4, e719.

Young, N.D., Chan, K.G., Korhonen, P.K., Min Chong, T., Ee, R., Mohandas, N., Koehler, A.V., Lim, Y.L., Hofmann, A., Jex, A.R., Qian, B., Chilton, N.B., Gobert, G.N., McManus, D.P., Tan, P., Webster, B.L., Rollinson, D., Gasser, R.B., 2015. Exploring molecular variation in *Schistosoma japonicum* in China. Sci. Rep. 5, 17345.

Zadesenets, K.S., Katokhin, A.V., Mordvinov, V.A., Rubtsov, N.B., 2012. Comparative cytogenetics of opisthorchid species (*Trematoda*, *Opisthorchiidae*). Parasitol. Int. 61, 87-89.

Zerbino, D.R., Birney, E., 2008. Velvet: Algorithms for *de novo* short read assembly using de Bruijn graphs. Genome Res. 18, 821-829.

Table 2-1 Mitochondrial genomic sequences for *Clonorchis sinensis* and other trematodes from various geographical origins used in the present study, with accession numbers and references listed.

Species (code)	Geographical origin	GenBank accession no.	References
Clonorchis sinensis (*Cs*-k2)	Jinju-si, Gyeongsangnam-do Province, South Korea	KY564177	**Present study**
C. sinensis (*Cs*-k1)	Korea	JF729304.1	(Cai et al., 2012)
C. sinensis (*Cs*-c1)	China	JF729303.1	(Cai et al., 2012)
C. sinensis (*Cs*-r1)	Khabarovsk, Khabarovskiy Krai, Russia	FJ381664.2	(Shekhovtsov et al., 2010)
Dicrocoelium chinensis	Gansu, China	NC_025279.1	(Liu et al., 2014b)
D. dendriticum	Gansu, China	NC_025280.1	(Liu et al., 2014b)
Fasciola gigantica	Guangxi, China	NC_024025.1	(Liu et al., 2014a)
F. hepatica	Victoria, Australia	NC_002546.1	(Le et al., 2001)
Fasciola sp. ('intermediate form'	Heilongjiang, China	KF543343.1	(Liu et al., 2014a)
Fascioloides magna	Czech Republic	NC_029481.1	(Ma et al., 2016)
Opisthorchis felineus	Novosibirsk Oblast, Russia	NC_011127.2	(Shekhovtsov et al., 2010)
O. viverrini	Laos	JF739555.1	(Cai et al., 2012)

Table 2-2 Comparisons of positions and nucleotide sequence lengths of genes as well as initiation and termination codons for protein-coding genes for mitochondrial genomes of *Clonorchis sinensis* currently accessible in public databases (cf. Table 2-1). *Clonorchis sinensis* from South Korea (*Cs*-k1 and *Cs*-k2), China (*Cs*-c1) and Russia (*Cs*-r1).

Gene/region	Positions and nucleotide sequence lengths (bp)				Initiation/termination codons			
	Cs-k2	*Cs*-k1	*Cs*-c1	*Cs*-r1	*Cs*-k2	*Cs*-k1	*Cs*-c1	*Cs*-r1
*cox*3	1-642 (642)	1-642 (642)	1-642 (642)	1-642 (642)	ATG/TAG	ATG/TAG	ATG/TAG	ATG/TAG
tRNA-His (H)	673-739 (67)	673-739 (67)	673-739 (67)	673-739 (67)				
*cyt*b	748-1860 (1113)	748-1860 (1113)	748-1860 (1113)	748-1860 (1113)	ATG/TAG	ATG/TAG	ATG/TAG	ATG/TAG
*nad*4L	1869-2132 (264)	1869-2132 (264)	1869-2132 (264)	1869-2132 (264)	ATG/TAG	ATG/TAG	ATG/TAG	ATG/TAG
*nad*4	2093-3370 (1278)	2093-3370 (1278)	2093-3370 (1278)	2093-3370 (1278)	GTG/TAG	GTG/TAG	GTG/TAG	GTG/TAG
tRNA-Gln (Q)	3383-3445 (63)	3383-3445 (63)	3383-3445 (63)	3383-3445 (63)				
tRNA-Phe (F)	3463-3528 (66)	3463-3528 (66)	3463-3528 (66)	3463-3527 (65)				
tRNA-Met (M)	3530-3597 (68)	3530-3597 (68)	3530-3597 (68)	3529-3596 (68)				
*atp*6	3598-4113 (516)	3598-4113 (516)	3598-4113 (516)	3597-4112 (516)	ATG/TAG	ATG/TAG	ATG/TAG	ATG/TAG
*nad*2	4149-5021 (873)	4149-5021 (873)	4150-5022 (873)	4148-5020 (873)	ATG/TAG	ATG/TAG	GTG/TAG	GTG/TAG
tRNA-Val (V)	5030-5094 (65)	5030-5094 (65)	5031-5095 (65)	5029-5093 (65)				
tRNA-Ala (A)	5112-5174 (63)	5112-5174 (63)	5113-5175 (63)	5111-5173 (63)				
tRNA-Asp (D)	5180-5248 (69)	5180-5248 (69)	5181-5249 (69)	5179-5247 (69)				
*nad*1	5252-6154 (903)	5252-6154 (903)	5253-6155 (903)	5251-6153 (903)	GTG/TAG	GTG/TAG	GTG/TAG	GTG/TAG
tRNA-Asn (N)	6154-6222 (69)	6154-6222 (69)	6155-6223 (69)	6153-6221 (69)				
tRNA-Pro (P)	6231-6298 (68)	6231-6298 (68)	6232-6299 (68)	6230-6297 (68)				
tRNA-Ile (I)	6299-6360 (62)	6299-6360 (62)	6300-6361 (62)	6298-6359 (62)				
tRNA-Lys (K)	6383-6450 (68)	6383-6450 (68)	6384-6451 (68)	6382-6449 (68)				
*nad*3	6454-6810 (357)	6454-6810 (357)	6455-6811 (357)	6453-6809 (357)	GTG/TAG	GTG/TAG	GTG/TAG	GTG/TAG
tRNA-SerAGN (S1)	6822-6881 (60)	6822-6881 (60)	6823-6882 (60)	6820-6880 (61)				
tRNA-Trp (W)	6899-6967 (69)	6899-6967 (69)	6900-6968 (69)	6898-6966 (69)				
*cox*1	6971-8530 (1560)	6971-8530 (1560)	6972-8531 (1560)	6970-8529 (1560)	GTG/TAA	GTG/TAA	GTG/TAA	GTG/TAA
tRNA-Thr (T)	8544-8607 (64)	8544-8607 (64)	8545-8608 (64)	8543-8606 (64)				
16S rRNA	8608-9605 (998)	8608-9605 (998)	8609-9607 (999)	8607-9604 (998)				
tRNA-Cys (C)	9606-9664 (59)	9606-9664 (59)	9608-9666 (59)	9605-9663 (59)				
18S rRNA	9665-10,443 (779)	9665-10,443 (779)	9667-10,445 (779)	9664-10,442 (779)				
*cox*2	10,444-11,079 (636)	10,444-11,079 (636)	10,446-11,081 (636)	10,443-11,078 (636)	ATG/TAG	ATG/TAG	ATG/TAG	ATG/TAA
*nad*6	11,099-11,560 (462)	11,099-11,560 (462)	11,101-11,562 (462)	11,098-11,559 (462)	GTG/TAA	GTG/TAA	GTG/TAA	GTG/TAA
tRNA-Tyr (Y)	11,567-11,628 (62)	11,567-11,628 (62)	11,569-11,630 (62)	11,566-11,627 (62)				
tRNA-LeuCUN (L1)	11,629-11,694 (66)	11,629-11,694 (66)	11,631-11,696 (66)	11,628-11,693 (66)				
tRNA-SerUCN (S2)	11,692-11,760 (69)	11,692-11,760 (69)	11,694-11,762 (69)	11,691-11,759 (69)				
tRNA-LeuUUR (L2)	11,765-11,829 (65)	11,765-11,829 (65)	11,767-11,831 (65)	11,764-11,828 (65)				
tRNA-Arg (R)	11,842-11,906 (65)	11,842-11,906 (65)	11,844-11,908 (65)	11,841-11,905 (65)				
*nad*5	11,908-13,512 (1605)	11,908-13,512 (1605)	11,910-13,514 (1605)	11,907-13,511 (1605)	GTG/TAA	GTG/TAA	GTG/TAA	GTG/TAA
tRNA-Glu (E)	13,523-13,589 (67)	13,523-13,589 (67)	13,525-13,591 (67)	13,522-13,588 (67)				
Non-coding region (NL)	13,590-13,743 (154)	13,590-13,742 (153)	13,592-13,744 (153)	-				
tRNA-Gly (G)	13,744-13,810 (67)	13,743-13,809 (67)	13,745-13,811 (67)	13,748 – 13,808 (67)				
Non-coding region (NS)	13,811-13,877 (67)	13,810-13,877 (68)	13,812-13,879 (68)	-				

Table 2-3 Numbers of synonomous (S) and non-synonomous (NS) nucleotide alterations in the mitochondrial genes among all available mitochondrial genome sequences available for *Clonorchis sinensis* from Korea (*Cs*-k1 and *Cs*-k2), China (*Cs*-c1) and Russia (*Cs*-r1).

Gene	*Cs*-k2		*Cs*-c1		*Cs*-r1	
	S	NS	S	NS	S	NS
*atp*6	1	0	2	1	1	1
*cox*1	6	3	0	0	6	3
*cox*2	4	1	0	0	4	2
*cox*3	2	2	4	2	3	2
*cyt*b	0	2	6	1	3	1
*nad*1	1	0	4	1	3	0
*nad*2	1	0	4	3	3	2
*nad*3	0	0	2	0	1	1
*nad*4L	0	0	0	0	0	0
*nad*4	1	1	1	3	0	2
*nad*5	2	1	7	4	4	1
*nad*6	0	0	4	2	3	0
Totals	18	10	34	17	31	15

Table 2-4 Summary of the features of the alignments of amino acid sequences predicted for individual protein-coding genes in the mitochondrial genomes of *Clonorchis sinensis* (n = 4) and eight other liver flukes (see Table 2-1) to establish alignment quality as well as the numbers of informative and invariable positions.

Amino acid sequence	Average length	Alignment length	Standard deviation	Shortest (bp)	Longest (bp)	Informative positions	Invariable positions
ATP6	171	174	2.92	170	172	120	48
COX1	515	522	7.57	510	520	220	288
COX2	207	225	18.76	200	214	125	72
COX3	213	216	2.50	213	216	157	55
CYTB	370	372	1.91	369	371	161	208
NAD1	299	301	1.22	299	300	153	145
NAD2	289	293	3.84	288	290	221	66
NAD3	117	119	1.53	116	118	83	32
NAD4L	88	90	2.45	87	90	47	40
NAD4	420	433	16.04	399	426	270	124
NAD5	527	543	17.29	518	534	362	143
NAD6	151	153	1.78	150	153	107	43

Table 2-5 Pairwise comparison of levels of sequence variation (%) in the mitochondrial gene or inferred amino acid sequences derived from mitochondrial genomes representing *Clonorchis sinensis* from Korea (*Cs*-k1 and *Cs*-k2), China (*Cs*-c1) and Russia (*Cs*-r1) (cf. Table 2-1).

Gene (length (nt))	Nucleotide (nt) variation (%)						Predicted protein (length (aa))	Amino acid (aa) variation (%)					
	Cs-k2 vs. *Cs*-k1	*Cs*-k2 vs. *Cs*-c1	*Cs*-k2 vs. *Cs*-r1	*Cs*-k1 vs. *Cs*-c1	*Cs*-k1 vs. *Cs*-r1	*Cs*-c1 vs. *Cs*-r1		*Cs*-k2 vs. *Cs*-k1	*Cs*-k2 vs. *Cs*-c1	*Cs*-k2 vs. *Cs*-r1	*Cs*-k1 vs. *Cs*-c1	*Cs*-k1 vs. *Cs*-r1	*Cs*-c1 vs *Cs*-r1
*atp*6 (516)	0.19	0.39	0.19	0.58	0.39	0.19	ATP6 (171)	0	0.58	0.58	0.58	0.58	0
*cox*1 (1560)	0.58	0.58	0.13	0	0.58	0.58	COX1 (519)	0.58	0.58	0	0	0.58	0.58
*cox*2 (636)	0.79	0.79	0.47	0	0.94	0.94	COX2 (211)	0.47	0.47	0.47	0	0.95	0.95
*cox*3 (642)	0.78	0.31	0.16	1.09	0.93	0.16	COX3 (213)	0.47	0	0	0.47	0.47	0
*cyt*b (1113)	0.18	0.63	0.36	0.63	0.36	0.27	CYTB (370)	0.54	0.27	0.27	0.27	0.27	0
*nad*1 (903)	0.11	0.66	0.44	0.55	0.33	0.22	NAD1 (300)	0	0.33	0	0.33	0	0.33
*nad*2 (873)	0.23	0.69	0.46	0.92	0.69	0.46	NAD2 (290)	0	1.03	0.69	1.03	0.69	0.34
*nad*3 (357)	0	0.56	0.56	0.56	0.56	0.56	NAD3 (118)	0	0	0.85	0	0.85	0.85
*nad*4 (1278)	0.16	0.31	0.16	0.31	0.16	0.16	NAD4 (425)	0.24	0.47	0.24	0.71	0.47	0.24
*nad*4L (264)	0	0	0	0	0	0	NAD4L (87)	0	0	0	0	0	0
*nad*5 (1605)	0.19	0.62	0.37	0.69	0.31	0.37	NAD5 (534)	0.19	0.56	0	0.75	0.19	0.56
*nad*6 (462)	0	1.3	0.65	1.3	0.65	0.65	NAD6 (153)	0	1.31	0	1.31	0	1.31

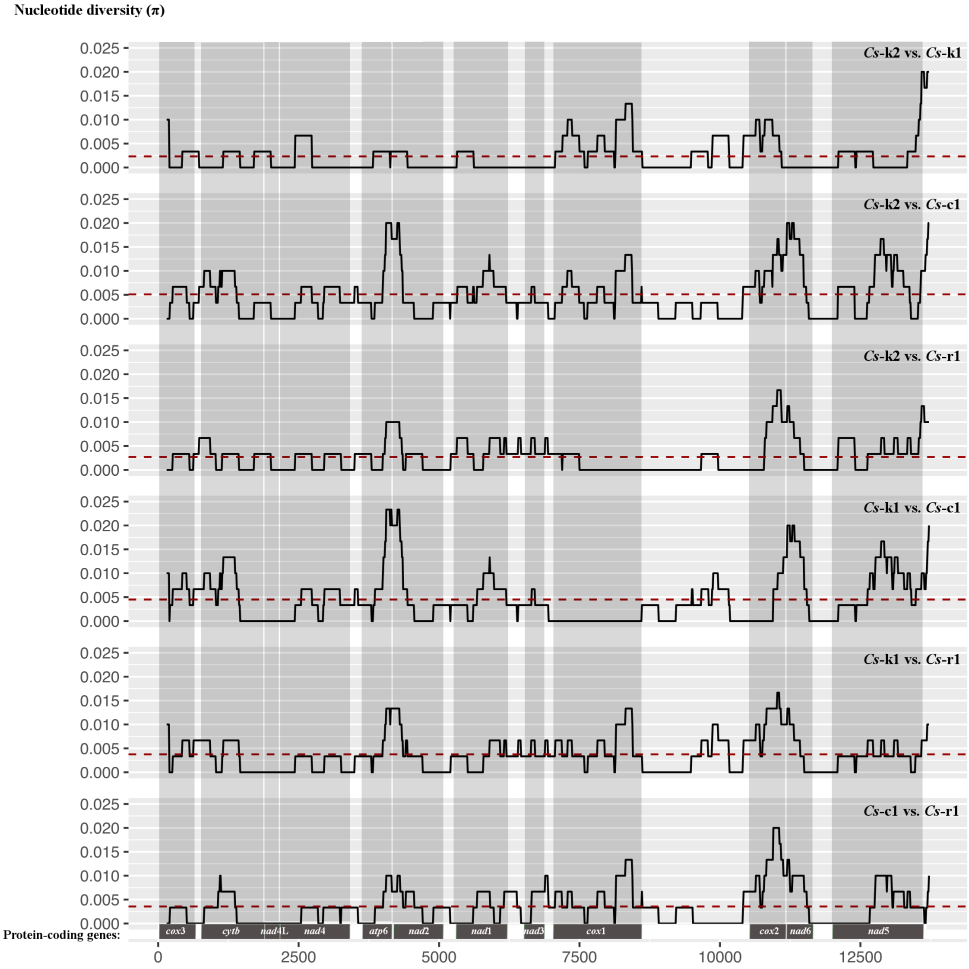


Fig. 2-1 Sliding window analyses between mitochondrial genomes of *Clonorchis sinensis* from Korea (*Cs*-k2 and *Cs*-k1), China (*Cs*-c1) and Russia (*Cs*-r1) (cf. Table 2-1) upon pairwise comparison (*Cs*-k2 vs. *Cs*-k1, *Cs*-k2 vs. *Cs*-c1, *Cs*-k2 vs. *Cs*-r1, *Cs*-k1 vs. *Cs*-c1, *Cs*-k1 vs. *Cs*-r1 and *Cs*-c1 vs. *Cs*-r1). Nucleotide diversity (π) was measured iteratively with a 10 bp-step using a 300 bp-window. Average diversity indicated by a dashed line.

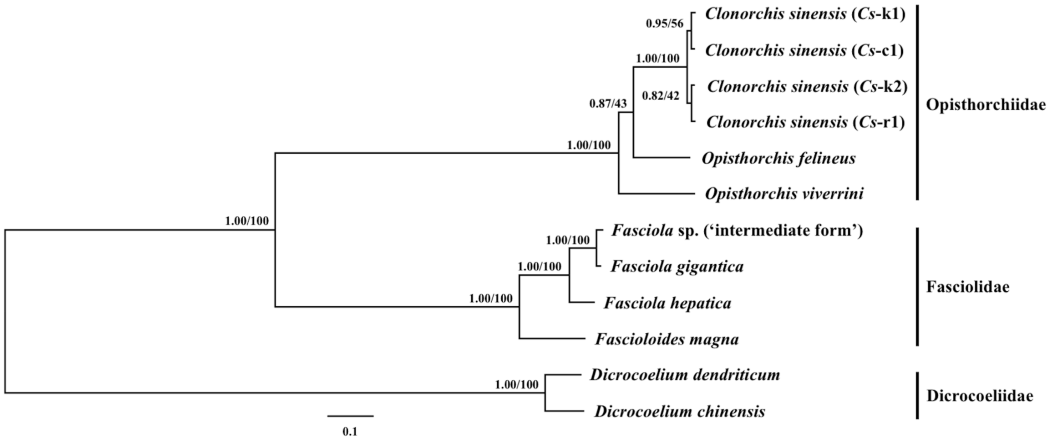


Fig. 2-2 Phylogenetic relationship of *Clonorchis sinensis* (isolate *Cs*-k2) with three other representatives of this species as well as other (selected) species of trematodes (cf. Table 2-1) inferred based on analyses of aligned concatenated amino acid sequences derived from mitochondrial genomes using two distinct tree-building algorithms (Bayesian inference [BI] and maximum likelihood [ML]). The presented branch length is inferred using the program RAxML (Stamatakis, 2014). The scale bar represents 0.1 substitution per site. Posterior probability (pp) support and bootstrap (bs) support values for BI and ML, respectively, are indicated at each node of the tree (in this order). *Dicrocoelium* species used as outgroups.

Chapter 3 - Improved genomic resources and new bioinformatic workflow for the carcinogenic parasite *Clonorchis sinensis*: Biotechnological implications

Abstract

Clonorchis sinensis (family Opisthorchiidae) is an important foodborne parasite that has a major socioeconomic impact on ~ 35 million people predominantly in China, Vietnam, Korea and the Russian Far East. In humans, infection with *C. sinensis* causes clonorchiasis, a complex hepatobiliary disease that can induce cholangiocarcinoma (CCA), a malignant cancer of the bile ducts. Central to understanding the epidemiology of this disease is knowledge of genetic variation within and among populations of this parasite. Although most published molecular studies seem to suggest that *C. sinensis* represents a single species, evidence of karyotypic variation within *C. sinensis* and cryptic species within a related opisthorchiid fluke (*Opisthorchis viverrini*) emphasise the importance of studying and comparing the genes and genomes of geographically distinct isolates of *C. sinensis*. Recently, we sequenced, assembled and characterised a draft nuclear genome of a *C. sinensis* isolate from Korea and compared it with a published draft genome of a Chinese isolate of this species using a bioinformatic workflow established for comparing draft genome assemblies and their gene annotations. We identified that 50.6% and 51.3% of the Korean and Chinese *C. sinensis* genomic scaffolds were syntenic, respectively. Within aligned syntenic blocks, the genomes had a high level of nucleotide identity (99.1%) and encoded 15 variable proteins likely to be involved in diverse biological processes. Here, we review current technical challenges of using draft genome assemblies to undertake comparative genomic analyses to quantify genetic variation between isolates of the same species. Using a workflow that overcomes these challenges, we report on a high-quality draft genome for *C. sinensis* from Korea and comparative genomic analyses, as a basis for future investigations of the genetic structures of *C. sinensis* populations, and discuss the biotechnological implications of these explorations.

3.1. Introduction

Parasitic flatworms (phylum Platyhelminthes; class Trematoda) include socioeconomically important foodborne pathogens that are responsible for some neglected tropical diseases that affect humans and other vertebrates, particularly in Asian countries (Furst et al., 2012; Torgerson, et al., 2015). For example, *Clonorchis sinensis* causes clonorchiasis, which impacts ~ 35 million people predominantly in China, Vietnam, Korea and the Russian Far East (Lun et al., 2005; Qian et al., 2012). More than 13.5 million people are infected with *C. sinensis* in China and the Republic of Korea alone (Qian et al., 2012). Clonorchiasis is a chronic hepatobiliary disease that can induce cholangiocarcinoma (CCA), a fatal cancer of the biliary system. Despite *C. sinensis* being classified as a Class I carcinogen by the International Agency for Research on Cancer (Choi et al., 2004, 2011), the complex biology of *C. sinensis* and the persistent cultural practice of eating raw fish have impeded efforts to control clonorchiasis and associated CCA in many regions of Asia.

The life cycle of *C. sinensis* commences when its eggs are ingested by a freshwater snail (e.g., bithynid) (Kaewkes 2003). Within the snail, a miracidium emerge from each egg, transforms into a sporocyst, which then undergoes asexual reproduction to eventually produce cercariae. Cercariae leave the snail, penetrate the skin of a freshwater cyprinid fish and encyst as metacercariae in tissues (Kaewkes 2003). When the definitive host ingests an infected fish, encysted metacercariae pass through the gastrointestinal tract, excyst in the small intestine and the juvenile flukes pass through the ampulla of Vater to establish in the biliary and/or pancreatic ducts. Within the intrahepatic biliary tree, the worms rapidly mature into hermaphroditic adults, which produce fertilised eggs that are released via faeces into the environment.

Chronic *C. sinensis* infection causes cholangitis, fibrosis, cholecystitis and associated cancer (CCA) (Qian et al., 2016). Despite the impact of clonorchiasis, no vaccines are available to prevent infection, and integrated control programs rely on health education, community awareness of the disease and chemotherapy with praziquantel - the only drug in current use to treat this disease (Oh et al., 2014). Although these strategies can be effective at controlling clonorchiasis, the prevalence of infection remains high in endemic areas due to the cultural practice of eating raw freshwater fish and because *C. sinensis* does not induce a protective immune response in humans (Oh et al., 2014; Qian et al., 2016). The repeated and prolonged use of a single drug can increase the risk of *C. sinensis* developing resistance to praziquantel (Tinga et al., 1999) and highlights the importance of developing alternative methods to prevent infection, including novel approaches to perturb cellular processes in *C. sinensis* within the mammalian host (Qian et al., 2016). To underpin these alternative approaches, a detailed understanding of the molecular biology of this parasite is required.

In 2010, the first transcriptome of adult *C. sinensis* was assembled *de novo* (Young et al., 2010), and in 2011 the first draft genomes of a Chinese isolate (designated here as *Cs*-c2) was published (Huang et al., 2013a; Wang et al., 2011). Profiling of the transcriptomes of eggs, metacercariae and adults (Yoo et al., 2011) and selected tissues of adult worms (Huang et al., 2013a) have provided transcriptional evidence of gene regulation throughout the life cycle of *C. sinensis*. This new genomic resource provided first insights into the molecular biology of this worm. For instance, the *C. sinensis* genome encodes genes required for fatty acid metabolism but lacks genes essential for fatty acid biosynthesis which are present in other eukaryotes, suggesting that *C. sinensis* has evolved to depend on host lipids (Huang et al., 2013a; Wang et al., 2011). In addition, predicted excretory-secretory products (ESPs) and tegumental proteins have been characterised and reported to include known flatworm antigens and other immunomodulators (Bian et al., 2014; Chen et al., 2014; Huang et al. 2012a, 2013b; Liang et al., 2014). This knowledge has led to a new understanding of complex host-pathogen interactions (Hu et al., 2014; Li et al., 2014a; Liang et al., 2014; Chen et al., 2015; Wang et al., 2014, 2017a) and has the potential to guide vaccine development (Huang et al., 2012a; Chen et al., 2014).

Thus far, it has been assumed that a reference genome of a single *C. sinensis* isolate is sufficient to represent all geographically distinct isolates of this species (Wang et al., 2011; Huang et al., 2013a). However, some studies have shown that *C. sinensis* from China and Korea have a karyotype (2n = 56) (Park and Yong 2001) that is distinct from that (2n = 14) of the Russian Far East (Zadesenets et al., 2012), suggesting the existence of cryptic species. Reports of significant genetic differences between geographically distinct isolates of *Opisthorchis viverrini,* a related opisithorchiid fluke, provides evidence that cryptic species occur (Saijuntha et al., 2007; Laoprom et al., 2009; Kiatsopit et al., 2011). Studies exploring genetic variation within *C. sinensis* have predicted low divergence among parasite populations using a small number of genetic loci (Lee and Huh, 2004; Le, et al. 2006; Cai et al., 2012; Liu, et al. 2012; Tatonova, et al. 2012, 2013; Sun et al. 2013; Chelomina et al., 2014). Recently, we compared mitochondrial genomes of *C. sinensis* isolates from Korea (designated *Cs*-k1 and *Cs*-k2; Cai et al., 2012; Wang et al., 2017b), China (*Cs*-c1; Cai et al., 2012) and Russia (designated *Cs*-r1; Shekhovtsov et al., 2010) and also observed limited genetic variation within and among these isolates (Wang et al., 2017b). However, no study has yet assessed genetic variation in *C. sinensis* at the whole nuclear genomic level. In other flatworm species, a genome sequence survey approach was used to characterise variation within *Fasciola hepatica* (see Cwiklinski et al., 2015), *Schistosoma mansoni* (see Clement et al., 2013; Crellen et al., 2016) and *Schistosoma japonicum* (see Young et al., 2015; Yin et al., 2016). These studies assumed that an existing genome was a suitable reference template for the mapping of sequence reads, to identify nucleotide-level variants between or among isolates. An alternative approach would be to independently assemble genomes from geographically distinct isolates and then to establish levels of genome synteny and genetic variation within syntenic regions of individual genomes. This alternative approach is more applicable when it is unclear that a single reference genome is suitable for comparative analysis of isolates within a suspected species complex, such as *C. sinensis*.

Here, we review a validated bioinformatic workflow system, designed for the comparison of draft genome assemblies and derived gene annotations. Using this workflow, we compare a draft genome of a Korean *C. sinensis* isolate (*Cs*-k2) with that of Chinese *C. sinensis* isolate (*Cs*-c2; Huang et al., 2013a) and reveal a high level of nucleotide similarity within the syntenic regions of the two genomes and identified two variable genes that encode proteins likely to participate in lipid metabolism and linked to the life of *C. sinensis* in the host's biliary system. By comparing aligned syntenic blocks between the draft genomes, we estimate nucleotide variability between the isolates, and assess synonymous and nonsynonymous mutations in protein-coding genes involved in diverse biological processes. The present article emphasises the challenges associated with genomic comparisons and encourages future investigations of the involvement of variable genes in host-parasite relationships, and their biological and biotechnological relevance.

3.2. Material and methods

Metacercariae of *C. sinensis* were isolated from naturally infected cyprinoid fish, *Pseudorasbora parva*, in the Jinju-si, Gyeongsangnam-do province, South Korea using established approaches (Sohn et al., 2006). In brief, fish were ground and digested in pepsin-hydrochloric acid (HCl) (0.01% pepsin (Sigma) in 0.1 M HCL) for 2 h at 37°C; then, metacercariae were isolated by sieving (0.5 mm aperture), washing and sedimentation in physiological saline. The identity of metacercariae was confirmed by light microscopy (40-times magnification) using established methods (Sohn 2009). Subsequently, helminth-free, inbred Syrian golden hamsters (*Mesocricetus auratus*) were infected with metacercariae (n = 50) as described previously (Sohn et al., 2006), in accordance with protocols approved by the Gyeongsang National University animal ethics committee. Juvenile (2 week-old) or adult (8 week-old) worms were collected from the bile ducts of euthanased hamsters, and worms were cultured *in vitro* to allow them to regurgitate caecal contents (Young et al., 2010). Subsequently, all developmental stages of *C. sinensis* were washed separately and extensively in physiological saline, snap-frozen in liquid nitrogen and then stored at -80 ºC.

High molecular weight genomic DNA (> 15 µg) was isolated from 95 adult (8 week-old) worms using an established protocol (Brindley et al., 1989). The DNA amount was determined using a Qubit fluorometer employing the dsDNA HS kit (Invitrogen), according to the manufacturer's instructions. DNA integrity was verified by agarose gel electrophoresis. Whole genome amplification (WGA) using the REPLI-g Midi Kit (Qiagen) and 200 ng of genomic template was used to produce the required amount of DNA for the construction of the 10 kb library. Short-insert (170 bp and 500 bp) and mate-pair (800 bp, 2 kb, 5 kb and 10 kb) genomic DNA libraries were constructed and paired-end sequenced using TruSeq sequencing chemistry employing the HiSeq 2000 sequencing platform (Illumina) (Supplementary File 3-1). The sequence data produced from each genomic DNA library were verified, and low quality sequences, base-calling duplicates and adapters removed using established methods (Li et al., 2010). Briefly, all sequences produced from short-insert libraries were corrected via majority voting over aligned *k*-mers (n = 17) (Li et al., 2010). In addition, reads were filtered if: (1) > 10% of the read contained continuous adenosine mono-phosphates (poly-A) or ambiguous (designated as 'N') bases; (2) > 65% of all bases in small insert-size libraries (i.e. < 800 bp) had a Phred quality of < 8; (3) > 80% of all bases in large insert libraries (2 kb, 5 kb and 10 kb) had a Phred quality of < 8; (4) reads were PCR duplicates (i.e. identical); (5) reads contained matches to Illumina adaptor sequence; and (6) paired-reads from 500 bp and 800 bp libraries overlapped by > 10 bp (< 10% mismatch). The final quality of each library was verified using the program FASTQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/).

Paired-end sequence data from all genomic DNA libraries were used to assemble the draft genome using the assembler SOAPdenovo2 r240 (Li et al., 2010). Initially, short-insert, paired-end reads were used to assemble contigs employing a *k*-mer value of 35 bp. Using all paired-end reads (from short-insert and mate-pair libraries), contigs were joined iteratively into scaffolds using a step-wise process, with > 3 read pairs required to form a connection. HaploMerger v2.0.0 (Huang et al., 2012b) was used to remove scaffold redundancy and re-scaffold using large-insert (2 kb, 5 kb and 10 kb) libraries and SSPACE v3.0 (Boetzer et al., 2011). Using data from small-insert (170 bp, 500 bp) libraries, gaps were closed using the program GapCloser v1.12 (http://soap.genomics.org.cn/about.html).

A full poly(A)-selected transcriptomic sequencing approach (RNA-seq) was applied to metacercarial, juvenile and adult stages of *Cs*-k2. In brief, total RNA was isolated from metacercaria (n = 200), juvenile (n = 40) or adult (n = 10) stages using the TriPure reagent (Roche), according to manufacturer's protocol, and then treated with *DN*ase I (DNase-Free, Ambion). Total RNA amounts and integrity were verified by using a 2100 BioAnalyzer (Agilent). Polyadenylated (polyA+) RNA was purified from 5-20 μg of total RNA using Sera-mag oligo(dT) beads, fragmented to a length of 160-240 base pairs (bp), reverse-transcribed using random hexamers, end-repaired and adaptor-ligated according to the manufacturer' s protocol (Illumina). Ligated products of ~168 bp were excised from agarose and then PCR-amplified (15 cycles). Products were purified using a MinElute column (Qiagen) and subjected to paired-end RNA-seq using HiSeq 2000 (Illumina) and assessed for quality and adaptor sequence.

Repetitive element libraries, which were then used for repeat masking by the program RepeatMasker (4.0.5) (Tarailo-Graovac and Chen 2009), were *de novo* predicted using RepeatModeler (1.0.8) (Tarailo-Graovac and Chen 2009) and LTR_FINDER (Xu and Wang 2007). The annotated and unknown repetitive elements identified by both programs were merged into separate 'annotated repeat' and 'unknown repeat' databases, respectively. Initially, the repetitive regions of the genome were annotated and hard-masked using the program RepeatMasker (Tarailo-Graovac and Chen 2009) with the ‘annotated repeat’ database. Then, the hard-masked genome was further annotated for repetitive elements and hard-masked using RepeatMasker with the ‘unknown repeat’ database. Other repetitive elements, such as simple repeats, satellites and low complexity repeats, were also annotated using RepeatMasker. Then, a consensus repeat annotation library for the draft genome was created by merging all the RepeatMasker output files (extension '.out') derived from the above steps using the program ProcessRepeats within the RepeatMasker package (Tarailo-Graovac and Chen 2009). In addition, Transfer RNA (tRNA) genes were predicted using the program tRNAscan-SE (1.3.1) (Lowe and Eddy 1997) with default settings. By searching against the Rfam database (12.1) (Griffiths-Jones et al., 2003) using the program INFERNAL (1.1.1) (Nawrocki et al., 2009), non-coding RNAs (e.g., ribosomal RNA) were identified in the *C. sinensis* draft genome.

3.2.1. Gene prediction, assessment and curation

3.2.1.1. Independent gene prediction

An independent *Cs*-k2 gene set was predicted using available genomic and transcriptomic data and the MAKER2 software framework v2.3.8 (Holt and Yandell, 2011).

First, transcriptomic support for each gene element was generated. This included: the identification of genomic regions encoding transcripts, identified by mapping RNAseq reads of both Korean (*Cs*-k1) and Chinese (*Cs*-c1 and *Cs*-c2) isolates (BioProject ID: PRJDA72781) to the assembled *Cs*-k2 (Korea) genome using TopHat2 v2.1.0 (Trapnell et al., 2014), identifying transcript locations using Cufflinks v2.2.1 (Trapnell et al., 2014), and assembling a transcriptome via both genome-guided and *de novo* methods using the program Trinity v2.2.0 (Haas et al., 2013) with the same RNAseq reads. The program Transdecoder 2.1.0 (http://transdecoder.sf.net) was used to select full-length transcripts from the assembly. These high-quality transcripts were used as a training set for *ab initio* gene prediction tools.

Second, we conducted *ab initio* gene prediction using the programs AUGUSTUS v3.1 (Stanke et al., 2008), SNAP v6.7 (Korf 2004) and GENEMARK v4.2.9 (Lukashin and Borodovsky 1998). The resultant *ab initio* gene predictions were then combined in the MAKER2 framework, together with the Cufflinks transcript models, genome-guided and *de novo*-assembled transcripts and proteomes from representative trematodes with published genomes, including *F. hepatica* (see Cwiklinski et al., 2015), *O. viverrini* (see Young et al., 2014) and *S. mansoni* (see Berriman et al., 2009). Further curation of the MAKER2 gene set was undertaken using the program Evidence Modeler (EVM) (Haas et al., 2003). In brief, the PASA pipeline v2.0.2 (Haas et al., 2003) was used to infer evidence based gene model candidates using high-quality Trinity transcripts. To infer an independent predicted gene set, these candidates were then combined with all putative gene models and the mappings of transcripts and proteins derived from the results of MAKER2, and given to the program EVM (Haas et al., 2003).

3.2.1.2. Gene transfer

Due to the differences between the gene prediction methodologies of *Cs*-c2 (Huang et al., 2013a) and *Cs*-k2, genes predicted in one genome may not be predicted in the other. Therefore, we downloaded the FASTA file (Accession number: GCA_000236345.1) of *Cs*-c2 genome and the corresponding GFF annotation file from WormBase (Howe et al., 2016) and used the program RATT (Otto et al., 2011) to transfer the protein coding gene models from the *Cs*-c2 genome to the *Cs*-k2 genome using the evidence of co-linear blocks and sequence homology. The coordinates of the transferred genes were compared to those of the independently predicted gene models. The transferred *Cs*-c2 genes that coincided to the intergenic areas and opposite strands of the predicted *Cs*-k2 gene models were identified and added to the EVM gene set.

3.2.1.3. Gene set annotation and curation

Amino acid sequences of the combined gene set were subjected to BLASTp search (default settings, E-value: 10^{-8}) (Altschul et al., 1997) against protein databases of *O. viverrini* (see Young et al., 2014), *F. hepatica* (see Cwiklinski et al., 2015) and *S. mansoni* (see Berriman et al., 2009) and the NCBI non-redundant protein database (Pruitt et al., 2007), Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa and Goto, 2000) and UniProtKB database (The UniProt Consortium, 2017). The BLASTp results were used to annotate and characterise protein functions. Based on similarity to KEGG orthologous gene terms, protein classes and biological pathways were assigned using KEGG PATHWAY (Kanehisa et al., 2007) and KEGG BRITE (Kanehisa et al., 2007) hierarchy information. In addition, we searched for conserved domains in the predicted proteins using the program InterProScan (Quevillon et al., 2005) using default settings. The results were then used to assign GO categories

(Ashburner et al., 2000) to the predicted genes. For the prediction of secreted proteins, SignalP v4.0 (Nielsen et al., 1997) and TMHMM v2.0 (Krogh et al., 2001) were used to identify signal peptides and transmembrane domains in predicted proteins, respectively. We defined genes encoding signal peptides (but not transmembrane domains) as putative ES proteins, whose cellular location was predicted using the program MultiLoc2 v26-10-2009 (Blum et al., 2009). ES proteins predicted to be secreted to lysosome and extracellular environment represented the final ES molecular set.

The functional annotation for each protein was used to select the final gene set. Repetitive elements within the predicted gene set were identified and removed if the description of their orthologous groups in the KEGG, Swiss-Prot and InterPro databases contained repeat-related keywords (e.g., mobile, transposon, transcriptase, transposase, transposable) and RepeatMasker identified an overlapping repetitive region (> 80%) within the protein-encoding sequence. Putative multi-copy gene families were identified by grouping predicted coding domains against themselves using the program OrthoMCL v.2.0.4 (Li et al., 2003). Inferred multi-copy gene family members without protein functional description were removed from the gene set if single exon genes overlapped in repetitive regions by more than 20% or multi-exon genes overlapped in repetitive regions by > 80%. Finally, the remaining genes were analysed using the program Transdecoder v.2.1.0 (http://transdecoder.github.io) to resolve coding regions that present complete open reading frames (ORF) (i.e. contain start and stop codons). The genes (with complete ORFs) containing > 80% repetitive regions and < 20% transcriptomic data-covered regions and the un-resolved genes (without complete ORFs) containing > 80% repetitive regions or < 20% transcriptomic data-covered regions or with only one exon were discarded. The remaining genes with < 30% of repetitive regions and the genes associated with KEGG, Swiss-Prot or InterPro databases were retained in the final gene set. Following these filtering steps, the final gene set was subjected to the program BUSCO v1.1 (Benchmarking Universal Single-Copy Orthologs) to assess completeness (Simao et al., 2015).

The predicted protein-encoding genes were classified and labelled based on the presence of InterProScan conserved domains and their matches to the items in the Swiss-Prot database (E-value: 10^{-8}). Genes without matches to either database were labelled as hypothetical proteins. All genes were renamed with the locus tag “CSKR”. After the curation of gene descriptors and the removal of overlapped genes, the final gene set of *Cs*-k2 was deposited in NCBI database under Bioproject PRJNA386618.

3.2.2. Orthologous protein groups

Orthologous protein groups of *C. sinensis* (*Cs*-k2 and *Cs*-c2) and of other representative trematodes whose genomes are published - *O. viverrini* (see Young et al., 2014), *F. hepatica* (see Cwiklinski et al., 2015) and *S. mansoni* (see Berriman et al., 2009) - were inferred using the program OrthoMCL (Li et al., 2003) using default parameters. Results were presented as an Edwards-Venn diagram (Edwards 2004) (modified from https://commons.wikimedia.org/wiki/File:Edwards-Venn-five.svg), created using the list of orthologous groups and a custom PERL script.

3.2.3. Comparison of gene sets

Based on nucleotide similarity, coding sequences of the two *C. sinensis* draft genomes (*Cs*-k2 and *Cs*-c2) were grouped using OrthoMCL (Li et al., 2003) employing default settings. Groups that contained only one *Cs*-k2 gene and one *Cs*-c2 gene were defined as single-copy gene pairs (SCGPs). The coding sequences of defined SCGPs were also subjected to the BLAST-Like Alignment Tool (BLAT) for comparison of aligned regions. The coverage and identity of aligned blocks of coding and gene regions of each SCGP were calculated using the BLAT alignment file and a custom PERL script that calculated the proportion of total coding domain or gene region length that was aligned.

3.2.4. Genome sequence alignment and synteny analysis

To detect homologous sequences between the draft genomes of *Cs*-k2 and *Cs*-c2, aligned genome scaffolds and regions of synteny were identified. First, to estimate overall nucleotide identity, a map of aligned nucleotide blocks between the masked and unmasked *Cs*-k2 and *Cs*-c2 genomes was created using the program nucmer employing the default settings and retaining matches with a minimum nucleotide identity of 90% (Kurtz et al., 2004). Only alignment blocks of > 200 nucleotides in length within paired scaffolds with > 10,000 aligned nucleotides were retained. Next, paired SCGPs, the genome scaffold coordinates of each SCGP and OrthoCluster (Vergara and Chen, 2010) were used to identify syntenic blocks that contained two or more SCGPs with a conserved gene order. To determine the nucleotide identity and assess variant effects within syntenic blocks, the unmasked nucleotide sequence of each *Cs*-c2 syntenic block (including 3'- and 5'- 10,000 nucleotide extensions) was aligned to the corresponding *Cs*-k2 scaffold using the nucmer and dnadiff packages within the program MUMmer3 and using a minimum alignment length of 200 nucleotides and a minimum identity of 95%. The coverage of the alignments was calculated using the GFF files, BEDTools v2.21.0 (Quinlan and Hall 2010) and a custom PERL script that used the intersection results from each BED file to calculate the total alignment length and the proportion of total scaffold length aligned.

Images to visually compare regions of nucleotide identity and synteny between the two draft genomes were created using Circos (v0.69) (Krzywinski et al., 2009). For each Circos plot, the order of chromosomes was optimised using the tool orderchr employing default settings (Krzywinski et al., 2009). To represent genome synteny, *Cs*-k2 scaffolds containing syntenic blocks with ≥ 10 SCGPs were displayed with their corresponding *Cs*-c2 scaffolds with ≥ 5 SCGPs. The largest *Cs*-k2 scaffold and nucleotide aligned *Cs*-c2 scaffolds were selected to represent the relationship between aligned nucleotide regions (from MUMmer3) and regions of synteny.

3.2.5. Identification of variable and invariable genes

Inferred nucleotide differences within the syntenic blocks were identified using the nucmer alignment output and using the show-snps package in MUMmer3 employing default settings (Kurtz et al., 2004). The location of alternative and reference nucleotides was summarised in a Variant Call Format (VCF) file and used, with the *Cs*-k2 gene annotation GFF file, to annotate variant effects with SnpEff (Cingolani et al., 2012) using default settings. Nucleotide differences in coding regions of *Cs*-k2 gene models were reported if such differences and insertion/deletion events did not change the function and/or position of splice sites, or start or stop codons. The final number of synonymous and non-synonymous mutations within coding regions was summarised using custom PERL scripts. Nucleotide differences in *Cs*-k2 coding regions were used to calculate the pairwise nucleotide identity of SCGPs conserved in the *Cs*-c2 and *Cs*-k2 genomes, and SCGPs with a pairwise nucleotide identity of < 98% were reported as variable. SCGPs with a pairwise nucleotide identity of > 99.8% were reported to be conserved. Functional annotation of variable SCGPs was inferred from the description of associated orthologous groups in the KEGG, Swiss-Prot and InterPro databases (Bairoch and Apweiler, 1999; Kanehisa and Goto, 2000; Quevillon et al., 2005).

3.3. Results

3.3.1. High quality nuclear genome for C. sinensis from Korea (Cs-k2)

From 54.8 Gb of high quality sequence reads representing ~90-fold genome coverage, we initially produced an assembly of 601 Mb in size (longest scaffold = 2.2 Mb, N50 = 314 kb, N content = 3.9%). Following the removal of redundancy, this assembly was re-scaffolded and gaps were closed (Table 3-1; Supplementary File 3-1), achieving a predicted genome size of 562,768,885 bp (BioProject: PRJNA386618; Accession no. NIRI00000000; 562.77 Mb) in 2,776 scaffolds with a GC-content of 43.95%. These steps increased the N50 value from 314 kb to 1.6 Mb. The size and GC-content of the draft genome of *Cs*-k2 were similar to the *Cs*-c2 draft genome (Accession no. GCA_000236345.1; 547.29 Mb; 4,348 scaffolds; 44.05% GC-content) (Huang et al., 2013a) (Table 3-1). The mean length of the assembled *Cs*-k2 scaffolds (longest scaffold = 8.86 Mb, N50 = 1,629 kb) was greater than that of the assembled *Cs*-c2 scaffolds (longest scaffold = 2.05 Mb, N50 = 417 kb), with more gaps supported by mate-pair physical coverage than *Cs*-c2 scaffolds (total = 3.16% and 0.03% ambiguous nucleotide (N) content in *Cs*-k2 and *Cs*-c2 scaffolds, respectively). The *Cs*-k2 draft genome was independently annotated and then used for comparative analyses of genomes and gene sets (Tables 3-1 and 3-2).

3.3.2. Repetitive elements

Approximately one third (33.32%) of the genome assembly for *Cs*-k2 encodes repetitive elements, and most (29.4%) were interspersed repeat elements (Supplementary File 3-2). A similar proportion of repeat elements was recorded in the *Cs*-c2 draft genome (32%). For the interspersed repeat elements, the percentages of LTR retrotransposons, long-interspersed nuclear element (LINE)-like and short-interspersed nuclear element (SINE)-like elements in *Cs*-k2 (LTR: 2.25%; LINES: 19.54%; SINES: 0.27%) were also consistent with those identified in the *Cs*-c2 genome (LTR: 1.97%; LINES: 15.07%;

SINES: 0.13%). Fewer annotated repetitive regions in *Cs*-k2 (5.08%) were unclassified compared with the *Cs*-c2 genome (14.32%).

3.3.3. Gene prediction and annotation

We predicted and curated 14,538 protein-encoding genes from the repeat-masked draft genome of *Cs*-k2; 13,643 of these genes were supported by transcriptomic data (Table 3-2; Supplementary File 3-3). Average gene length (15,556 bp), coding sequence length (1,437 bp) and exon length (234 bp) in the *Cs*-k2 gene set were similar to the published 13,634 gene set of *Cs*-c2 (average gene length: 17,761 bp; average coding sequence length: 1,591 bp; average exon length: 232 bp) (Huang et al., 2013a) (Table 3-2). More complete benchmarking eukaryotic single-copy orthologs (BUSCO) were identified in the *Cs*-k2 gene set (n = 555; 65%) than in that of *Cs*-c2 (n = 493; 58%), and fewer BUSCO groups were predicted to be fragmented or missing in *Cs*-k2 (79/209) than in *Cs*-c2 (116/234) (Table 3-2).

From the annotation of the 14,538 *Cs*-k2 protein encoding genes (Supplementary Files 3-3 and 3-4), we inferred 8,696 (59.82%) proteins with one or more Pfam (n = 6,288; 43.25%), PANTHER (n = 7,772; 53.46%), PRINTS (n = 1,234; 8.49%) and PIRSF (n = 251; 1.73%) conserved amino acid domains, 9,737(66.7%) proteins homologous (BLASTp, E-value ,$\leqslant 10^{-8}$) to proteins submitted to Swiss-Prot (n = 7,090; 48.77%) and/or Kyoto Encyclopaedia of Genes and Genomes (KEGG) (n = 9,367; 64.43%) databases. Proteins annotated with KEGG orthologous gene terms (n = 3,591) were mapped to 245 KEGG pathways and 43 KEGG BRITE protein families. Using KEGG BRITE (Supplementary File 3-5), we identified 276 peptidases, 332 kinases, 313 phosphatases, 91 GTP-binding proteins, 151 ion channels, 311 transporters and 161 receptors. Cysteine (n = 90, 32.5%) and metallo- (n = 75, 27.0%) peptidases formed the most abundant peptidase families among the aspartic, cysteine, metallo-, serine and threonine peptidases identified within the *Cs*-k2 gene set. Of the 313 phosphatases classified, most were serine/threonine (256) or tyrosine phosphatases (34). We also identified 91 GTPase, 71 small, monomeric G-proteins, 20 heterotrimeric G-proteins and 150 ion channel proteins including 55 voltage-gated cation channels. Classified receptors (n = 161) included 70 GPCRs (54 Class A rhodopsin family-proteins). The *Cs*-k2 draft genome encoded 207 predicted ES (excretory-secretory) proteins (Supplementary File 3-6), including 101 proteins with KEGG (n = 97), Pfam (n = 59), PANTHER (n = 63) and SWISSPROT (n = 61) annotation. Based on KEGG BRITE annotation, ES proteins included 10 peptidases, 4 collagen proteins, 3 Niemann-Pick C2 (NPC2) proteins, 1 hydrolase and 1 glycosyltransferase.

3.3.4. Trematode orthologs

Proteins predicted for *Cs*-k2 were compared with those of *Cs*-c2 and three other species of trematode, *O. viverrini*, *F. hepatica* and *S. mansoni* (Fig. 3-1 and Supplementary File 3-7). Of the 13,259 orthologous groups identified in all species, 8,394 groups contained proteins common to *Cs*-k2 and one or more trematode species and 8,398 groups were common to *Cs*-c2 proteins and one or more trematode species (Supplementary File 3-7). The predicted proteomes of *C. sinensis* clustered mostly into common orthologous groups with *O. viverrini* (7,887 for *Cs*-k2, and 7,812 for *Cs*-c2), followed by *F. hepatica* (6,295 for *Cs*-k2, and 6,427 for *Cs*-c2) and *S. mansoni* (5,798 for *Cs*-k2, and 5,937 for *Cs*-c2) (Supplementary File 3-7). A total of 791 orthologous groups were common to *Cs*-k2 and one or more other trematode species, but were not detected in the *Cs*-c2 gene set (Supplementary File 3-7). In contrast, 795 groups were common to *Cs*-c2 and one or more other trematode species, but were not found in the *Cs*-k2 gene set (Supplementary File 3-7). Combining orthologous groups in *Cs*-k2 and *Cs*-c2, 4,796 orthologous groups were common to all flatworms; 157 or 261 supplementary groups were common to *Cs*-k2 or *Cs*-c2 and the other flatworm taxa assessed here (Fig. 3-1). Differences between the *Cs*-k2 and *Cs*-c2 gene sets were further investigated by the identification and pairwise alignment of orthologous single copy *C. sinensis* gene pairs (SCGPs).

3.3.5. Genomic comparisons

In total, 7,886 SCGPs (54.24% of the *Cs*-k2 gene get) were identified in *Cs*-k2 and *Cs*-c2 gene sets, with 6,733 (85.38%) and 3,421 (43.38%) aligning across 50% and 90% of their total sequence lengths, respectively (Table 3-2). A total of 2,464 (31.25%) SCGPs shared conserved splice site and start/stop codon positions. The gene order of inferred SCGPs (n = 7,886) was used to identify and characterise conserved syntenic blocks between the *Cs*-k2 and *Cs*-c2 draft genomes.

To compare nucleotide identity across all scaffolds, repeat masked and unmasked *Cs*-k2 and *Cs*-c2 genomes were aligned (Table 3-3; Supplementary File 3-8). Using masked genomes, ~ 319 Mb (56.73%) of 610 *Cs*-k2 scaffolds were aligned (average length = 1,323 bp) to *Cs*-c2 genome scaffolds for an alignment length of > 200 nucleotides. Using unmasked genomes, ~ 512 Mb (90.17%) of 640 *Cs*-k2 genome scaffolds were aligned to *Cs*-c2 genome scaffolds for an alignment length of > 200 nucleotides (average length = 9,601 bp) (Table 3-3; Supplementary File 3-8). Scaffolds with one or more aligned nucleotide blocks represent ~99% and ~98% of the size of the *Cs*-k2 and *Cs*-c2 genomes, respectively (Table 3-3). Based on the location of 6,421 SCGPs with conserved gene order (81.4% of total SCGPs) in *Cs*-k2 and *Cs*-c2 genome scaffolds, 1,827 syntenic blocks (total length = 285 Mb; 50.6% and 51.3% of the *Cs*-k2 and *Cs*-c2 genomes, respectively) were identified (Fig. 3-2; Supplementary File 3-8). Syntenic blocks included a minimum of two and a maximum of 18 genes, and were contained within 484 *Cs*-k2 scaffolds (total length = 532 Mb) and 1,262 *Cs*-c2 scaffolds (total length = 455 Mb). Scaffolds with one or more syntenic blocks represents 96% and ~83% of *Cs*-k2 and *Cs*-c2 genome size, respectively (Table 3-3). Within aligned syntenic blocks, *Cs*-k2 and *Cs*-c2 genome scaffolds shared 99.1% nucleotide identity with an average length of 153,712 nucleotides (Fig. 3-3; Table 3-3). For comparative purposes, aligned nucleotide and syntenic blocks in *Cs*-k2 genome scaffold sc00000020 are presented with corresponding *Cs*-c2 scaffolds (Fig. 3-3). In this example, most aligned nucleotide blocks were contained within syntenic blocks (with conserved gene order); however, several aligned nucleotide blocks were also observed in scaffolds with predicted gene and repetitive sequence annotation, but were not predicted as syntenic (Fig. 3-3).

3.3.6. Variable and invariable genes

Nucleotide conservation and variation between *Cs*-c2 and *Cs*-k2 were then determined within the coding domains of 5,379 SCGP *Cs*-k2 gene models in aligned syntenic blocks (Supplementary File 3-9). Nucleotide differences were observed in intronic (398,727 nucleotide differences; 4.32 per kb) and exonic (15,096 nucleotide differences; 1.72 per kb) regions, and included 6,863 non-synonymous and 8,233 synonymous mutations within coding domains. Most SCGPs were relatively well conserved, with 68.20% (3,669) aligning with a pairwise nucleotide identity of > 99.8% (Supplementary File 3-

9). Fifteen variable SCGPs (0.28% of total) with synonymous and nonsynonymous mutations had an observed pairwise nucleotide identity of ≤ 98% (Table 3-4, Supplementary File 3-10). These variable genes included six genes with functional annotation inferred from KEGG and/or InterProScan databases (Table 3-4, Supplementary File 3-10), cathepsin-D peptidase (CSKR_13438s; 96.17% pairwise nucleotide identity), two Niemann-Pick C2 (NPC2) proteins (CSKR_9510s and CSKR_11331s; 96.99 and 97.38%, respectively), a calponin-like protein (CSKR_709s; 97.40%), histone 2A (CSKR_482s; 97.62%) and a homologue of protein FAM161A (CSKR_2578s; 97.86%).

3.4. Dicussion

Previous studies using a small panel of genetic loci or whole mitochondrial genomic data suggested genetic variation among geographically distinct *C. sinensis* isolates is limited (Lee and Huh 2004; Le et al., 2006; Cai et al., 2012; Chelomina et al., 2014; Liu et al., 2012; Tatonova et al., 2012, 2013; Sun et al., 2013; Wang et al., 2017b), a finding that is not consistent with karyotypic variation recorded among *C. sinensis* isolates (Park and Yong, 2001; Zadesenets et al., 2012). Here, we explored the genome of a *C. sinensis* isolate from Korea (*Cs*-k2), and characterised genome-wide nucleotide variation between *Cs*-k2 and a published genome for an isolate from China (*Cs*-c2). Unlike most studies that have relied on mapping short sequence reads to an existing reference genome (Jirimutu et al., 2012; Clement et al., 2013; Huang et al., 2014a; Crellen et al., 2016; Hane et al., 2017), here we utilised an independently assembled genome to unambiguously explore syntenic blocks and identify regions of sequence variation between the two isolates.

A comparison of the two genomes (*Cs*-k2 and *Cs*-c2) and corresponding gene sets revealed the presence of unique orthologous protein groups for each isolate and identified syntenic blocks covering ~ 50% of the two genomes. Limited synteny between these genomes suggests a lack of assembly contiguity and/or genetic distinctiveness in structure and composition; thus, at this time-point, a single draft reference genome is not representative of *C. sinensis* isolates from distinct geographical areas. Similar results have also been reported in previous genome comparisons (Li et al., 2014b; Zhou et al., 2017). Based on these findings, we suggest that genetic investigations by genome sequence survey (GSS) should quantify read-mapping rates to both genomes before selecting an annotated reference. Identification and curation of syntenic blocks containing unambiguously aligned sequences between the two genomes enabled the quantitation of nucleotide variation in coding domains of SCGPs. The syntenic blocks and evidence of genetic variation identified here are expected to be useful for ensuing population genetic studies of *C. sinensis* isolates from geographically distinct areas, which should further validate this conclusion. In addition, the methodology applied here provides a cost-effective way to define variable genetic markers by a direct comparison of draft genomes, which could be applied more broadly, without needing to achieve chromosomal contiguity in assembly.

Defining synteny between *Cs*-k2 and *Cs*-c2 revealed varying levels of contiguity between the two assemblies. For example, one *Cs*-k2 scaffold often aligned to multiple *Cs*-c2 scaffolds, and *vice versa*. In total, fewer *Cs*-k2 scaffolds (n = 484, 532 Mb) were aligned to more *Cs*-c2 scaffolds (n = 1,262, 455 Mb), and the aligned *Cs*-k2 scaffolds contained a larger proportion of the total genomic content than in *Cs*-c2. Statistical indicators support the quality of the genomic assemblies (Table 3-1). Interestingly, the N50 value of the *Cs*-k2 assembly was three times higher than that of *Cs*-c2, although the total number of *Cs*-k2 scaffolds was less than that of *Cs*-c2. Multiple factors could contribute to such a difference in contiguity. First, *Cs*-k2 DNA was isolated from 95 adult worms, which might have contained more euchromatin and, hence, improved sequence coverage of the genomic DNA template and a more complete genome assembly. Second, the DNA libraries constructed for *Cs*-k2 had insert sizes of 170 bp to 10 kb, while those of *Cs*-c2 ranged from 300 bp to 5 kb (Huang et al., 2013a). Paired-end reads with larger insert sizes, spanning longer genomic regions, can lead to improved assembly quality (van Heesch et al., 2013). Moreover, a difference in the assembly and post-assembly workflows for the *Cs*-k2 and *Cs*-c2 genomic data sets (cf. Myers et al., 2000; Li et al., 2010; Huang et al., 2013a) would have led to the distinctiveness in quality between the *Cs*-k2 and *Cs*-c2 draft genomes. Using our workflow steps (section 2), the N50 value of the *Cs*-k2 assembly increased from 314 kb to 1.6 Mb. Similar improvements have been reported also in other studies using a similar approach (Hane et al., 2014; Huang et al., 2014b). Compared with the relatively high cost of PacBio-sequencing genomes of organisms with genomes of more than 100 Mb in size, the present approach established here provides a cost-effective alternative for sequencing and assembly.

The high-quality *Cs*-k2 assembly provided a solid foundation for the prediction of the gene set for *C. sinensis*, which was assessed for quality using the benchmark BUSCO groups (Simao et al., 2015). The results (Table 3-2) showed that the *Cs*-k2 gene set is more complete than that of *Cs*-c2. Improvements to the gene-prediction workflow often lead to a more complete gene set and the incorporation of multiple data sets, including the proteomes of related taxa and gene transcription evidence, usually reduce the probability of predicting invalid gene models (Elsik et al., 2014). In our workflow system, proteomes of related trematode species were included, and transcriptomic data derived from the two *C. sinensis* isolates were compared with *ab initio* gene predictions conducted using the gene set merger programs Maker2 and EVM, in order to infer the final gene set. The *Cs*-c2 gene models that aligned to the *Cs*-k2 genome in regions where no existing gene model was predicted were also included to increase the completeness of the *Cs*-k2 gene set. Gene transfer has been used frequently to annotate new or refined genome assemblies with existing gene models of the same species (e.g., Protasio et al., 2012; Otto et al., 2014). In addition to differences in the gene set completeness predicted using the BUSCO groups (Simao et al., 2015), the inference of orthologous groups using data from other trematodes revealed differences between the two *C. sinensis* gene sets. Orthologous groups shared by flatworms but not present in either the *Cs*-k2 or *Cs*-c2 gene set further support the differences observed in the contiguity of assembly and/or the distinctiveness in genomic structure and composition. This observation may also reflect the difference in the assembly and annotation methodologies applied to the two genomes. For instance, within syntenic blocks, the predicted gene structures (linked to start and stop codons, and splice sites) often varied between *Cs*-k2 and *Cs*-c2, which was supported by the distinctiveness in alignment coverage between SCGPs (i.e. single-copy gene pairs). To unambiguously assess genetic variation between the two isolates, we focused on establishing genetic variation in syntenic blocks.

The high degree of shared nucleotide identity between aligned syntenic blocks suggests limited divergence between *Cs*-k2 and *Cs*-c2, which is consistent with previous genetic studies of *C. sinensis* (see Lee and Huh 2004; Le et al., 2006; Wang et al., 2011; Cai et al., 2012; Liu et al., 2012; Sun et al., 2013; Chelomina et al., 2014; Tatonova et al., 2012, 2013). Nonetheless, we did detect a large number of nucleotide differences in intronic and exonic regions. The lower density of nucleotide differences in the exonic areas (1.725 per kb) compared with intronic areas (4.32 per kb) is consistent with the selective pressure to preserve codons in mature messenger RNAs (Betts et al., 2001). However, fewer observed non-synonymous nucleotide differences (6,863) than synonymous nucleotide differences (8,233) in the coding domains of SCGPs provided weak support for selective pressure within the coding domains of the *C. sinensis* SCGPs. A possible explanation is that molecular changes in some SCGPs are adaptive or neutral, which requires more work to determine the selective pressure on individual genes in the future. In general, single-copy genes tend to be more conserved (Han et al., 2014), the number of nucleotide differences detected per kb in exonic regions of the single copy genes in *C. sinensis* (1.72 per kb) is markedly lower than that in *S. japonicum* (average of 6.4 per kb) (Young et al., 2015), and 68.20% (3,669 of 6,069) of the SCGPs were found to be highly conserved (> 99.8% nucleotide identity) in the present study. Only 15 SCGPs had a nucleotide identity of less than 98%. Collectively, these findings might be explained by hermaphroditism (Saijuntha et al., 2008). In *O. viverrini*, self-fertilization has been suggested to cause heterozygote deficiency due to a relative low parasite burden within the mammalian host (Kiatsopit et al., 2014), although similar observations have not been reported for *C. sinensis* by other researchers. Detailed genomic comparisons of large numbers of individuals of *C. sinensis* from a broad geographical range is now required to confirm the low genetic variation and variable gene set we predicted in this study.

Surprisingly, among the translated products of the 15 variable SCGPs, we identified two NPC2 proteins containing a conserved MD-2 related lipid-recognition domain. The NPC2 protein is a lipid-binding protein involved in lipid transportation and chemical homeostasis (Xu et al., 2007). Four NPC2 genes have been shown to be highly transcribed (FPKM of > 100) in the oral sucker of *C. sinensis* (see Huang et al., 2013a). We also identified a family of 40 proteins with NPC2-like lipid-binding domains in *Cs*-k2 compared with 25 in *O. viverrini* (see Young et al., 2014), suggesting an expansion of this gene family in some opisthorchiid flukes with weaker purifying selection and rapid evolution of the group (Francino, 2005). Interestingly, of the 11 nucleotide differences identified in one of the variable NPC2 genes (CSKR_11331s), nine were non-synonymous, which is consistent with a rapid evolution of this gene. Future studies should explore NPC2 genes and proteins in *C. sinensis* and related liver flukes to assess variation among geographically distinct isolates and to determine the functional relevance of this gene family.

Here, we focused on single nucleotide differences within syntenic blocks. Without a high level of contiguity, direct comparison between draft genomes of the same species remains a challenging task. Most of the previous intraspecific genome comparison studies have been conducted between draft assemblies and a complete or near-complete reference genome (Hester et al., 2013; Huang et al., 2014a; Hane et al., 2017; Zhou et al., 2017). In these studies, the detection of structural variation has relied on the relative location of mapped sequences to the reference. In the absence of an assembly with chromosomal contiguity, it is not possible to reliably identify structural variation and the genetic variation located at the ambiguous (gap) or un-assembled areas. With the decreasing cost of third-generation sequencing techniques (cf. Korhonen et al., 2016), we hope that it will be possible to produce a complete *C. sinensis* genome in the near future to establish a broader range of genetic and structural variation. For now, the new, high-quality draft genome for *C. sinensis* from Korea provides a sound basis for future investigations of the genetic structures of *C. sinensis* populations in Asia.

3.5. Conclusions

This article highlights the technical challenges of comparing draft genome assemblies, to estimate the magnitude of genetic variation between isolates of a particular species. We report on a practical workflow that we have developed to overcome these challenges. Using this workflow, we revealed a high level of nucleotide similarity within the syntenic regions of the two genomes and identified two variable genes that encode proteins likely to participate in lipid metabolism and linked to the life of *C. sinensis* in the host's biliary system. Further studies should explore the involvement of these variable genes in host-parasite relationships in a biotechnological context.

3.6. References

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389-3402.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., Sherlock, G., 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat. Genet. 25, 25-29.

Bairoch, A., Apweiler, R., 1999. The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1999. Nucleic Acids Res. 27, 49-54.

Berriman, M., Haas, B.J., LoVerde, P.T., Wilson, R.A., Dillon, G.P., Cerqueira, G.C., Mashiyama, S.T., Al-Lazikani, B., Andrade, L.F., Ashton, P.D., Aslett, M.A., Bartholomeu, D.C., Blandin, G., Caffrey, C.R., Coghlan, A., Coulson, R., Day, T.A., Delcher, A., DeMarco, R., Djikeng, A., Eyre, T., Gamble, J.A., Ghedin, E., Gu, Y., Hertz-Fowler, C., Hirai, H., Hirai, Y., Houston, R., Ivens, A., Johnston, D.A., Lacerda, D., Macedo, C.D., McVeigh, P., Ning, Z., Oliveira, G., Overington, J.P., Parkhill, J., Pertea, M., Pierce, R.J., Protasio, A.V., Quail, M.A., Rajandream, M.A., Rogers, J., Sajid, M., Salzberg, S.L., Stanke, M., Tivey, A.R., White, O., Williams, D.L., Wortman, J., Wu, W., Zamanian, M., Zerlotini, A., Fraser-Liggett, C.M., Barrell, B.G., El-Sayed, N.M., 2009. The genome of the blood fluke *Schistosoma mansoni*. Nature 460, 352-358.

Betts, M.J., Guigo, R., Agarwal, P., Russell, R.B., 2001. Exon structure conservation despite low sequence similarity: a relic of dramatic events in evolution? EMBO J. 20, 5354-5360.

Bian, M., Li, S., Wang, X., Xu, Y., Chen, W., Zhou, C., Chen, X., He, L., Xu, J., Liang, C., Wu, Z., Huang, Y., Li, X., Yu, X., 2014. Identification, immunolocalization, and immunological characterization of nitric oxide synthase-interacting protein from *Clonorchis sinensis*. Parasitol. Res. 113, 1749-1757.

Blum, T., Briesemeister, S., Kohlbacher, O., 2009. MultiLoc2: integrating phylogeny and Gene Ontology terms improves subcellular protein localization prediction. BMC Bioinf. 10, 274.

Boetzer, M., Henkel, C.V., Jansen, H.J., Butler, D., Pirovano, W., 2011. Scaffolding pre-assembled contigs using SSPACE. Bioinformatics 27, 578-579.

Brindley, P.J., Lewis, F.A., McCutchan, T.F., Bueding, E., Sher, A., 1989. A genomic change associated with the development of resistance to hycanthone in *Schistosoma mansoni*. Mol. Biochem. Parasitol. 36, 243-252.

Cai, X.Q., Liu, G.H., Song, H.Q., Wu, C.Y., Zou, F.C., Yan, H.K., Yuan, Z.G., Lin, R.Q., Zhu, X.Q., 2012. Sequences and gene organization of the mitochondrial genomes of the liver flukes *Opisthorchis viverrini* and *Clonorchis sinensis* (Trematoda). Parasitol. Res. 110, 235-243.

Chelomina, G.N., Tatonova, Y.V., Hung, N.M., Ngo, H.D., 2014. Genetic diversity of the Chinese liver fluke *Clonorchis sinensis* from Russia and Vietnam. Int. J. Parasitol. 44, 795-810.

Chen, W., Ning, D., Wang, X., Chen, T., Lv, X., Sun, J., Wu, D., Huang, Y., Xu, J., Yu, X., 2015. Identification and characterization of *Clonorchis sinensis* cathepsin B proteases in the pathogenesis of clonorchiasis. Parasit. Vectors 8, 647.

Chen, W., Wang, X., Lv, X., Tian, Y., Xu, Y., Mao, Q., Shang, M., Li, X., Huang, Y., Yu, X., 2014. Characterization of the secreted cathepsin B cysteine proteases family of the carcinogenic liver fluke *Clonorchis sinensis*. Parasitol. Res. 113, 3409-3418.

Choi, B.I., Han, J.K., Hong, S.T., Lee, K.H., 2004. Clonorchiasis and cholangiocarcinoma: etiologic relationship and imaging diagnosis. Clin. Microbiol. Rev. 17, 540-552.

Choi, M.H., Chang, Y.S., Lim, M.K., Bae, Y.M., Hong, S.T., Oh, J.K., Yun, E.H., Bae, M.J., Kwon, H.S., Lee, S.M., Park, H.W., Min, K.U., Kim, Y.Y., Cho, S.H., 2011. *Clonorchis sinensis* infection is positively associated with atopy in endemic area. Clin. Exp. Allergy 41, 697-705.

Cingolani, P., Platts, A., Wang, L.L., Coon, M., Nguyen, T., Wang, L., Land, S.J., Lu, X.Y., Ruden, D.M., 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w(1118); iso-2; iso-3. Fly 6, 80-92.

Clement, J.A., Toulza, E., Gautier, M., Parrinello, H., Roquis, D., Boissier, J., Rognon, A., Mone, H., Mouahid, G., Buard, J., Mitta, G., Grunau, C., 2013. Private selective sweeps identified from next-generation pool-sequencing reveal convergent pathways under selection in two inbred *Schistosoma mansoni* strains. PLoS Negl. Trop. Dis. 7, e2591.

Crellen, T., Allan, F., David, S., Durrant, C., Huckvale, T., Holroyd, N., Emery, A.M., Rollinson, D., Aanensen, D.M., Berriman, M., Webster, J.P., Cotton, J.A., 2016. Whole genome resequencing of the human parasite *Schistosoma mansoni* reveals population history and effects of selection. Sci. Rep. 6, 20954.

Cwiklinski, K., Dalton, J.P., Dufresne, P.J., La Course, J., Williams, D.J.L., Hodgkinson, J., et al., 2015. The *Fasciola hepatica* genome: gene duplication and polymorphism reveals adaptation to the host environment and the capacity for rapid evolution. Genome Biol. 16, 71.

Edwards, A.W.F., 2004. Cogwheels of the mind : the story of Venn diagrams. Johns Hopkins University Press, Baltimore.

Elsik, C.G., Worley, K.C., Bennett, A.K., Beye, M., Camara, F., Childers, C.P., de Graaf, D.C., Debyser, G., Deng, J., Devreese, B., Elhaik, E., Evans, J.D., Foster, L.J., Graur, D., Guigo, R., teams, H.p., Hoff, K.J., Holder, M.E., Hudson, M.E., Hunt, G.J., Jiang, H., Joshi, V., Khetani, R.S., Kosarev, P., Kovar, C.L., Ma, J., Maleszka, R., Moritz, R.F., Munoz-Torres, M.C., Murphy, T.D., Muzny, D.M., Newsham, I.F., Reese, J.T., Robertson, H.M., Robinson, G.E., Rueppell, O., Solovyev, V., Stanke, M., Stolle, E., Tsuruda, J.M., Vaerenbergh, M.V., Waterhouse, R.M., Weaver, D.B., Whitfield, C.W., Wu, Y., Zdobnov, E.M., Zhang, L., Zhu, D., Gibbs, R.A., Honey Bee Genome Sequencing, C., 2014. Finding the missing honey bee genes: lessons learned from a genome upgrade. BMC Genomics 15, 86.

Francino, M.P., 2005. An adaptive radiation model for the origin of new gene functions. Nat. Genet. 37, 573-577.

Furst, T., Keiser, J., Utzinger, J., 2012. Global burden of human food-borne trematodiasis: a systematic review and meta-analysis. Lancet Infect. Dis. 12, 210-221.

Griffiths-Jones, S., Bateman, A., Marshall, M., Khanna, A., Eddy, S.R., 2003. Rfam: an RNA family database. Nucleic Acids Res. 31, 439-441.

Haas, B.J., Delcher, A.L., Mount, S.M., Wortman, J.R., Smith, R.K., Jr., Hannick, L.I., Maiti, R., Ronning, C.M., Rusch, D.B., Town, C.D., Salzberg, S.L., White, O., 2003. Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. Nucleic Acids Res. 31, 5654-5666.

Haas, B.J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P.D., Bowden, J., Couger, M.B., Eccles, D., Li, B., Lieber, M., Macmanes, M.D., Ott, M., Orvis, J., Pochet, N., Strozzi, F., Weeks, N., Westerman, R., William, T., Dewey, C.N., Henschel, R., Leduc, R.D., Friedman, N., Regev, A., 2013. *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat. Protoc. 8, 1494-1512.

Han, F.M., Peng, Y., Xu, L.J., Xiao, P.G., 2014. Identification, characterization, and utilization of single copy genes in 29 angiosperm genomes. BMC Genomics 15, 504.

Hane, J.K., Anderson, J.P., Williams, A.H., Sperschneider, J., Singh, K.B., 2014. Genome sequencing and comparative genomics of the broad host-range pathogen *Rhizoctonia solani* AG8. PLoS Genet. 10, e1004281.

Hane, J.K., Ming, Y., Kamphuis, L.G., Nelson, M.N., Garg, G., Atkins, C.A., Bayer, P.E., Bravo, A., Bringans, S., Cannon, S., Edwards, D., Foley, R., Gao, L.L., Harrison, M.J., Huang, W., Hurgobin, B., Li, S., Liu, C.W., McGrath, A., Morahan, G., Murray, J., Weller, J., Jian, J.B., Singh, K.B., 2017. A comprehensive draft genome sequence for lupin (*Lupinus angustifolius*), an emerging health food: insights into plant-microbe interactions and legume evolution. Plant Biotechnol. J. 15, 318-330.

van Heesch, S., Kloosterman, W.P., Lansu, N., Ruzius, F.P., Levandowsky, E., Lee, C.C., Zhou, S.G., Goldstein, S., Schwartz, D.C., Harkins, T.T., Guryev, V., Cuppen, E., 2013. Improving mammalian genome scaffolding using large insert mate-pair next-generation sequencing. BMC Genomics 14, 257.

Hester, J., Chan, E.R., Menard, D., Mercereau-Puijalon, O., Barnwell, J., Zimmerman, P.A., Serre, D., 2013. *De novo* assembly of a field isolate genome reveals novel *Plasmodium vivax* erythrocyte invasion genes. PLoS Negl. Trop. Dis. 7, e2569.

Holt, C., Yandell, M., 2011. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinf. 12, 491.

Howe, K.L., Bolt, B.J., Cain, S., Chan, J., Chen, W.J., Davis, P., Done, J., Down, T., Gao, S., Grove, C., Harris, T.W., Kishore, R., Lee, R., Lomax, J., Li, Y., Muller, H.M., Nakamura, C., Nuin, P., Paulini, M., Raciti, D., Schindelman, G., Stanley, E., Tuli, M.A., Van Auken, K., Wang, D., Wang, X., Williams, G., Wright, A., Yook, K., Berriman, M., Kersey, P., Schedl, T., Stein, L., Sternberg, P.W., 2016. WormBase 2016: expanding to enable helminth genomic research. Nucleic Acids Res. 44, D774-780.

Hu, Y., Zhang, E., Huang, L., Li, W., Liang, P., Wang, X., Xu, J., Huang, Y., Yu, X., 2014. Expression profiles of glyceraldehyde-3-phosphate dehydrogenase from

Clonorchis sinensis: a glycolytic enzyme with plasminogen binding capacity. Parasitol. Res. 113, 4543-4553.

Huang, J., Zhao, Y., Shiraigol, W., Li, B., Bai, D., Ye, W., Daidiikhuu, D., Yang, L., Jin, B., Zhao, Q., Gao, Y., Wu, J., Bao, W., Li, A., Zhang, Y., Han, H., Bai, H., Bao, Y., Zhao, L., Zhai, Z., Zhao, W., Sun, Z., Zhang, Y., Meng, H., Dugarjaviin, M., 2014a. Analysis of horse genomes provides insight into the diversification and adaptive evolution of karyotype. Sci. Rep. 4, 4958.

Huang, S., Chen, Z., Huang, G., Yu, T., Yang, P., Li, J., Fu, Y., Yuan, S., Chen, S., Xu, A., 2012b. HaploMerger: reconstructing allelic relationships for polymorphic diploid genome assemblies. Genome Res. 22, 1581-1588.

Huang, S., Chen, Z., Yan, X., Yu, T., Huang, G., Yan, Q., Pontarotti, P.A., Zhao, H., Li, J., Yang, P., Wang, R., Li, R., Tao, X., Deng, T., Wang, Y., Li, G., Zhang, Q., Zhou, S., You, L., Yuan, S., Fu, Y., Wu, F., Dong, M., Chen, S., Xu, A., 2014b. Decelerated genome evolution in modern vertebrates revealed by analysis of multiple lancelet genomes. Nat. Commun. 5, 5896.

Huang, Y., Chen, W., Wang, X., Liu, H., Chen, Y., Guo, L., Luo, F., Sun, J., Mao, Q., Liang, P., Xie, Z., Zhou, C., Tian, Y., Lv, X., Huang, L., Zhou, J., Hu, Y., Li, R., Zhang, F., Lei, H., Li, W., Hu, X., Liang, C., Xu, J., Li, X., Yu, X., 2013a. The carcinogenic liver fluke, *Clonorchis sinensis*: new assembly, reannotation and analysis of the genome and characterization of tissue transcriptomes. PLoS One 8, e54732.

Huang, Y., Li, W., Huang, L., Hu, Y., Chen, W., Wang, X., Sun, J., Liang, C., Wu, Z., Li, X., Xu, J., Yu, X., 2012a. Identification and characterization of myophilin-like protein: a life stage and tissue-specific antigen of *Clonorchis sinensis*. Parasitol. Res. 111, 1143-1150.

Huang, Y., Liao, H., Li, W., Hu, Y., Huang, L., Wang, X., Sun, J., Chen, W., Deng, C., Liang, C., Wu, Z., Li, X., Xu, J., Yu, X., 2013b. Identification, sequence analysis and characterization of *Clonorchis sinensis* ubiquitin. Exp. Parasitol. 133, 62-69.

Jirimutu, Wang, Z., Ding, G., Chen, G., Sun, Y., Sun, Z., Zhang, H., Wang, L., Hasi, S., Zhang, Y., Li, J., Shi, Y., Xu, Z., He, C., Yu, S., Li, S., Zhang, W., Batmunkh, M., Ts, B., Narenbatu, Unierhu, Bat-Ireedui, S., Gao, H., Baysgalan, B., Li, Q., Jia, Z., Turigenbayila, Subudenggerile, Narenmanduhu, Wang, Z., Wang, J., Pan, L., Chen, Y., Ganerdene, Y., Dabxilt, Erdemt, Altansha, Altansukh, Liu, T., Cao, M., Aruuntsever, Bayart, Hosblig, He, F., Zha-ti, A., Zheng, G., Qiu, F., Sun, Z., Zhao, L., Zhao, W., Liu, B., Li, C., Chen, Y., Tang, X., Guo, C., Liu, W., Ming, L., Temuulen, Cui, A., Li, Y., Gao, J., Li, J., Wurentaodi, Niu, S., Sun, T., Zhai, Z., Zhang, M., Chen, C., Baldan, T., Bayaer, T., Li, Y., Meng, H., 2012. Genome sequences of wild and domestic bactrian camels. Nat. Commun. 3, 1202.

Kaewkes, S., 2003. Taxonomy and biology of liver flukes. Acta Trop. 88, 177-186.

Kanehisa, M., Araki, M., Goto, S., Hattori, M., Hirakawa, M., Itoh, M., Katayama, T., Kawashima, S., Okuda, S., Tokimatsu, T., 2007. KEGG for linking genomes to life and the environment. Nucleic Acids Res. 36, D480-D484.

Kanehisa, M., Goto, S., 2000. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 28, 27-30.

Kiatsopit, N., Sithithaworn, P., Saijuntha, W., Pitaksakulrat, O., Petney, T.N., Webster, J.P., Andrews, R.H., 2014. Analysis of the population genetics of *Opisthorchis viverrini* sensu lato in the Nam Ngum River wetland, Lao PDR, by multilocus enzyme electrophoresis. Parasitol. Res. 113, 2973-2981.

Kiatsopit, N., Sithithaworn, P., Sithithaworn, J., Boonmars, T., Tesana, S., Pitaksakulrat, O., Saijuntha, W., Petney, T.N., Andrews, R.H., 2011. Genetic relationships within the *Opisthorchis viverrini* species complex with specific analysis of *O. viverrini* from Savannakhet, Lao PDR by multilocus enzyme electrophoresis. Parasitol. Res. 108, 211-217.

Korf, I., 2004. Gene finding in novel genomes. BMC Bioinf. 5, 59.

Korhonen, P.K., Young, N.D., Gasser, R.B., 2016. Making sense of genomes of parasitic worms: Tackling bioinformatic challenges. Biotechnol. Adv. 34, 663-686.

Krogh, A., Larsson, B., von Heijne, G., Sonnhammer, E.L., 2001. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J. Mol. Biol. 305, 567-580.

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S.J., Marra, M.A., 2009. Circos: an information aesthetic for comparative genomics. Genome Res. 19, 1639-1645.

Kurtz, S., Phillippy, A., Delcher, A.L., Smoot, M., Shumway, M., Antonescu, C., Salzberg, S.L., 2004. Versatile and open software for comparing large genomes. Genome Biol. 5.

Laoprom, N., Saijuntha, W., Sithithaworn, P., Wongkham, S., Laha, T., Ando, K., Andrews, R.H., Petney, T.N., 2009. Biological variation within *Opisthorchis viverrini* sensu lato in Thailand and Lao PDR. J. Parasitol. 95, 1307-1313.

Le, T.H., Van De, N., Blair, D., Sithithaworn, P., McManus, D.P., 2006. *Clonorchis sinensis* and *Opisthorchis viverrini:* development of a mitochondrial-based multiplex PCR for their identification and discrimination. Exp. Parasitol. 112, 109-114.

Lee, S.U., Huh, S., 2004. Variation of nuclear and mitochondrial DNAs in Korean and Chinese isolates of *Clonorchis sinensis*. Korean J. Parasitol. 42, 145-148.

Li, L., Stoeckert, C.J., Jr., Roos, D.S., 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. 13, 2178-2189.

Li, R., Fan, W., Tian, G., Zhu, H., He, L., Cai, J., Huang, Q., Cai, Q., Li, B., Bai, Y., Zhang, Z., Zhang, Y., Wang, W., Li, J., Wei, F., Li, H., Jian, M., Li, J., Zhang, Z., Nielsen, R., Li, D., Gu, W., Yang, Z., Xuan, Z., Ryder, O.A., Leung, F.C., Zhou, Y., Cao, J., Sun, X., Fu, Y., Fang, X., Guo, X., Wang, B., Hou, R., Shen, F., Mu, B., Ni, P., Lin, R., Qian, W., Wang, G., Yu, C., Nie, W., Wang, J., Wu, Z., Liang, H., Min, J., Wu, Q., Cheng, S., Ruan, J., Wang, M., Shi, Z., Wen, M., Liu, B., Ren, X., Zheng, H., Dong, D., Cook, K., Shan, G., Zhang, H., Kosiol, C., Xie, X., Lu, Z., Zheng, H., Li, Y., Steiner, C.C., Lam, T.T., Lin, S., Zhang, Q., Li, G., Tian, J., Gong, T., Liu, H., Zhang, D., Fang, L., Ye, C., Zhang, J., Hu, W., Xu, A., Ren, Y., Zhang, G., Bruford, M.W., Li, Q., Ma, L., Guo, Y., An, N., Hu, Y., Zheng, Y., Shi, Y., Li, Z., Liu, Q., Chen, Y., Zhao, J., Qu, N., Zhao, S., Tian, F., Wang, X., Wang, H., Xu, L., Liu, X., Vinar, T., Wang, Y., Lam, T.W., Yiu, S.M., Liu, S., Zhang, H., Li, D., Huang, Y., Wang, X., Yang, G., Jiang, Z., Wang, J., Qin, N., Li, L., Li, J., Bolund, L., Kristiansen, K., Wong, G.K., Olson, M., Zhang, X., Li, S., Yang, H., Wang, J., Wang, J., 2010. The sequence and *de novo* assembly of the giant panda genome. Nature 463, 311-317.

Li, S., Bian, M., Wang, X.Y., Chen, X.Q., Xie, Z.Z., Sun, H.C., Jia, F.F., Liang, P., Zhou, C.H., He, L., Mao, Q., Huang, B., Liang, C., Wu, Z.D., Li, X.R., Xu, J., Huang, Y., Yu, X.B., 2014a. Molecular and biochemical characterizations of three fructose-1,6-bisphosphate aldolases from *Clonorchis sinensis*. Mol. Biochem. Parasitol. 194, 36-43.

Li, Y.H., Zhou, G.Y., Ma, J.X., Jiang, W.K., Jin, L.G., Zhang, Z.H., Guo, Y., Zhang, J.B., Sui, Y., Zheng, L.T., Zhang, S.S., Zuo, Q.Y., Shi, X.H., Li, Y.F., Zhang, W.K., Hu, Y.Y., Kong, G.Y., Hong, H.L., Tan, B., Song, J., Liu, Z.X., Wang, Y.S., Ruan, H., Yeung, C.K.L., Liu, J., Wang, H.L., Zhang, L.J., Guan, R.X., Wang, K.J., Li, W.B., Chen, S.Y., Chang, R.Z., Jiang, Z., Jackson, S.A., Li, R.Q., Qiu, L.J., 2014b.

De novo assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits. Nat. Biotechnol. 32, 1045-1052.

Liang, P., He, L., Xu, Y.Q., Chen, X.Q., Huang, Y., Ren, M.Y., Liang, C., Li, X.R., Xu, J., Lu, G., Yu, X.B., 2014. Identification, immunolocalization, and characterization analyses of an exopeptidase of papain superfamily, (cathepsin C) from *Clonorchis sinensis*. Parasitol. Res. 113, 3621-3629.

Liu, G.H., Li, B., Li, J.Y., Song, H.Q., Lin, R.Q., Cai, X.Q., Zou, F.C., Yan, H.K., Yuan, Z.G., Zhou, D.H., Zhu, X.Q., 2012. Genetic variation among *Clonorchis sinensis* isolates from different geographic regions in China revealed by sequence analyses of four mitochondrial genes. J. Helminthol. 86, 479-484.

Lowe, T.M., Eddy, S.R., 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. 25, 955-964.

Lukashin, A.V., Borodovsky, M., 1998. GeneMark.hmm: new solutions for gene finding. Nucleic Acids Res. 26, 1107-1115.

Lun, Z.R., Gasser, R.B., Lai, D.H., Li, A.X., Zhu, X.Q., Yu, X.B., Fang, Y.Y., 2005. Clonorchiasis: a key foodborne zoonosis in China. Lancet Infect. Dis. 5, 31-41.

Myers, E.W., Sutton, G.G., Delcher, A.L., Dew, I.M., Fasulo, D.P., Flanigan, M.J., Kravitz, S.A., Mobarry, C.M., Reinert, K.H., Remington, K.A., Anson, E.L., Bolanos, R.A., Chou, H.H., Jordan, C.M., Halpern, A.L., Lonardi, S., Beasley, E.M., Brandon, R.C., Chen, L., Dunn, P.J., Lai, Z., Liang, Y., Nusskern, D.R., Zhan, M., Zhang, Q., Zheng, X., Rubin, G.M., Adams, M.D., Venter, J.C., 2000. A whole-genome assembly of *Drosophila*. Science 287, 2196-2204.

Nawrocki, E.P., Kolbe, D.L., Eddy, S.R., 2009. Infernal 1.0: inference of RNA alignments. Bioinformatics 25, 1335-1337.

Nielsen, H., Engelbrecht, J., Brunak, S., von Heijne, G., 1997. Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. Protein Eng. 10, 1-6.

Oh, J.K., Lim, M.K., Yun, E.H., Cho, H., Park, E.Y., Choi, M.H., Shin, H.R., Hong, S.T., 2014. Control of clonorchiasis in Korea: effectiveness of health education for community leaders and individuals in an endemic area. Trop. Med. Int. Health 19, 1096-1104.

Otto, T.D., Dillon, G.P., Degrave, W.S., Berriman, M., 2011. RATT: Rapid Annotation Transfer Tool. Nucleic Acids Res. 39, e57.

Otto, T.D., Rayner, J.C., Bohme, U., Pain, A., Spottiswoode, N., Sanders, M., Quail, M., Ollomo, B., Renaud, F., Thomas, A.W., Prugnolle, F., Conway, D.J., Newbold, C., Berriman, M., 2014. Genome sequencing of chimpanzee malaria parasites reveals possible pathways of adaptation to human hosts. Nat. Commun. 5, 4754.

Park, G.M., Yong, T.S., 2001. Geographical variation of the liver fluke, *Clonorchis sinensis*, from Korea and China based on the karyotypes, zymodeme and DNA sequences. Southeast Asian J. Trop. Med. Public Health 32 (Suppl. 2), 12-16.

Protasio, A.V., Tsai, I.J., Babbage, A., Nichol, S., Hunt, M., Aslett, M.A., De Silva, N., Velarde, G.S., Anderson, T.J., Clark, R.C., Davidson, C., Dillon, G.P., Holroyd, N.E., LoVerde, P.T., Lloyd, C., McQuillan, J., Oliveira, G., Otto, T.D., Parker-Manuel, S.J., Quail, M.A., Wilson, R.A., Zerlotini, A., Dunne, D.W., Berriman, M., 2012. A systematically improved high quality genome and transcriptome of the human blood fluke *Schistosoma mansoni*. PLoS Negl. Trop. Dis. 6, e1455.

Pruitt, K.D., Tatusova, T., Maglott, D.R., 2007. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res. 35, D61-65.

Qian, M.B., Chen, Y.D., Liang, S., Yang, G.J., Zhou, X.N., 2012. The global epidemiology of clonorchiasis and its relation with cholangiocarcinoma. Infect. Dis. Poverty. 1, 4.

Qian, M.B., Utzinger, J., Keiser, J., Zhou, X.N., 2016. Clonorchiasis. The Lancet 387, 800-810.

Quevillon, E., Silventoinen, V., Pillai, S., Harte, N., Mulder, N., Apweiler, R., Lopez, R., 2005. InterProScan: protein domains identifier. Nucleic Acids Res. 33, W116-W120.

Quinlan, A.R., Hall, I.M., 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26, 841-842.

Saijuntha, W., Sithithaworn, P., Wongkham, S., Laha, T., Pipitgool, V., Tesana, S., Chilton, N.B., Petney, T.N., Andrews, R.H., 2007. Evidence of a species complex within the food-borne trematode *Opisthorchis viverrini* and possible co-evolution with their first intermediate hosts. Int. J. Parasitol. 37, 695-703.

Saijuntha, W., Sithithaworn, P., Wongkham, S., Laha, T., Satrawaha, R., Chilton, N.B., Petney, T.N., Andrews, R.H., 2008. Genetic variation at three enzyme loci within a Thailand population of *Opisthorchis viverrini*. Parasitol. Res. 103, 1283-1287.

Shekhovtsov, S.V., Katokhin, A.V., Kolchanov, N.A., Mordvinov, V.A., 2010. The complete mitochondrial genomes of the liver flukes *Opisthorchis felineus* and *Clonorchis sinensis* (Trematoda). Parasitol. Int. 59, 100-103.

Simao, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V., Zdobnov, E.M., 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31, 3210-3212.

Sohn, W.M., 2009. Fish-borne zoonotic trematode metacercariae in the Republic of Korea. Korean J. Parasitol. 47 (Suppl.), S103-113.

Sohn, W.M., Zhang, H., Choi, M.H., Hong, S.T., 2006. Susceptibility of experimental animals to reinfection with *Clonorchis sinensis*. Korean J. Parasitol. 44, 163-166.

Stanke, M., Diekhans, M., Baertsch, R., Haussler, D., 2008. Using native and syntenically mapped cDNA alignments to improve *de novo* gene finding. Bioinformatics 24, 637-644.

Sun, J., Huang, Y., Huang, H., Liang, P., Wang, X., Mao, Q., Men, J., Chen, W., Deng, C., Zhou, C., Lv, X., Zhou, J., Zhang, F., Li, R., Tian, Y., Lei, H., Liang, C., Hu, X., Xu, J., Li, X., Xinbingyu, 2013. Low divergence of *Clonorchis sinensis* in China based on multilocus analysis. PLoS One 8, e67006.

Tarailo-Graovac, M., Chen, N., 2009. Using RepeatMasker to identify repetitive elements in genomic sequences. Curr. Protoc. Bioinformatics Chapter 4, Unit 4. 10.

Tatonova, Y.V., Chelomina, G.N., Besprosvannykh, V.V., 2012. Genetic diversity of nuclear ITS1-5.8S-ITS2 rDNA sequence in *Clonorchis sinensis* Cobbold, 1875 (Trematoda: Opisthorchidae) from the Russian Far East. Parasitol. Int. 61, 664-674.

Tatonova, Y.V., Chelomina, G.N., Besprozvannykh, V.V., 2013. Genetic diversity of *Clonorchis sinensis* (Trematoda: Opisthorchiidae) in the Russian southern Far East based on mtDNA *cox1* sequence variation. Folia Parasitol. (Praha) 60, 155-162.

The UniProt Consortium, 2017. UniProt: the universal protein knowledgebase. Nucleic Acids Res. 45, D158-D169.

Tinga, N., De, N., Vien, H.V., Chau, L., Toan, N.D., Kager, P.A., Vries, P.J., 1999. Little effect of praziquantel or artemisinin on clonorchiasis in Northern Vietnam. A pilot study. Trop. Med. Int. Health 4, 814-818.

Torgerson, P.R., Devleesschauwer, B., Praet, N., Speybroeck, N., Willingham, A.L., Kasuga, F., Rokni, M.B., Zhou, X.N., Fevre, E.M., Sripa, B., Gargouri, N., Furst, T., Budke, C.M., Carabin, H., Kirk, M.D., Angulo, F.J., Havelaar, A., de Silva, N., 2015.

World Health Organization Estimates of the global and regional disease burden of 11 foodborne parasitic diseases, 2010: a data synthesis. PLoS Med. 12, e1001920.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D.R., Pimentel, H., Salzberg, S.L., Rinn, J.L., Pachter, L., 2014. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat. Protoc. 9, 2513.

Vergara, I.A., Chen, N.S., 2010. Large synteny blocks revealed between *Caenorhabditis elegans* and *Caenorhabditis briggsae* genomes using OrthoCluster. BMC Genomics 11, 516.

Wang, C., Lei, H., Tian, Y., Shang, M., Wu, Y., Li, Y., Zhao, L., Shi, M., Tang, X., Chen, T., Lv, Z., Huang, Y., Tang, X., Yu, X., Li, X., 2017a. *Clonorchis sinensis* granulin: identification, immunolocalization, and function in promoting the metastasis of cholangiocarcinoma and hepatocellular carcinoma. Parasit. Vectors 10, 262.

Wang, D., Young, N.D., Koehler, A.V., Tan, P., Sohn, W.M., Korhonen, P.K., Gasser, R.B., 2017b. Mitochondrial genomic comparison of *Clonorchis sinensis* from South Korea with other isolates of this species. Infect. Genet. Evol. 51, 160-166.

Wang, X., Chen, W., Huang, Y., Sun, J., Men, J., Liu, H., Luo, F., Guo, L., Lv, X., Deng, C., Zhou, C., Fan, Y., Li, X., Huang, L., Hu, Y., Liang, C., Hu, X., Xu, J., Yu, X., 2011. The draft genome of the carcinogenic human liver fluke *Clonorchis sinensis*. Genome Biol. 12, R107.

Wang, X.Y., Hu, F.Y., Hu, X.C., Chen, W.J., Huang, Y., Yu, X.B., 2014. Proteomic identification of potential *Clonorchis sinensis* excretory/secretory products capable of binding and activating human hepatic stellate cells. Parasitol. Res. 113, 3063-3071.

Xu, S., Benoff, B., Liou, H.L., Lobel, P., Stock, A.M., 2007. Structural basis of sterol binding by NPC2, a lysosomal protein deficient in Niemann-Pick type C2 disease. J. Biol. Chem. 282, 23525-23531.

Xu, Z., Wang, H., 2007. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. Nucleic Acids Res. 35, W265-W268.

Yin, M., Liu, X., Xu, B., Huang, J., Zheng, Q., Yang, Z., et al., 2016. Genetic variation between *Schistosoma japonicum* lineages from lake and mountainous regions in China revealed by resequencing whole genomes. Acta Trop. 161, 79-85.

Yoo, W.G., Kim, D.W., Ju, J.W., Cho, P.Y., Kim, T.I., Cho, S.H., Choi, S.H., Park, H.S., Kim, T.S., Hong, S.J., 2011. Developmental transcriptomic features of the carcinogenic liver fluke, *Clonorchis sinensis*. PLoS Negl. Trop. Dis. 5, e1208.

Young, N.D., Campbell, B.E., Hall, R.S., Jex, A.R., Cantacessi, C., Laha, T., Sohn, W.M., Sripa, B., Loukas, A., Brindley, P.J., Gasser, R.B., 2010. Unlocking the transcriptomes of two carcinogenic parasites, *Clonorchis sinensis* and *Opisthorchis viverrini*. PLoS Negl. Trop. Dis. 4, e719.

Young, N.D., Chan, K.G., Korhonen, P.K., Min Chong, T., Ee, R., Mohandas, N., Koehler, A.V., Lim, Y.L., Hofmann, A., Jex, A.R., Qian, B., Chilton, N.B., Gobert, G.N., McManus, D.P., Tan, P., Webster, B.L., Rollinson, D., Gasser, R.B., 2015. Exploring molecular variation in *Schistosoma japonicum* in China. Sci. Rep. 5, 17345.

Young, N.D., Nagarajan, N., Lin, S.J., Korhonen, P.K., Jex, A.R., Hall, R.S., Safavi-Hemami, H., Kaewkong, W., Bertrand, D., Gao, S., Seet, Q., Wongkham, S., Teh, B.T., Wongkham, C., Intapan, P.M., Maleewong, W., Yang, X., Hu, M., Wang, Z., Hofmann, A., Sternberg, P.W., Tan, P., Wang, J., Gasser, R.B., 2014. The *Opisthorchis viverrini* genome provides insights into life in the bile duct. Nat. Commun. 5, 4378.

Zadesenets, K.S., Katokhin, A.V., Mordvinov, V.A., Rubtsov, N.B., 2012. Comparative cytogenetics of opisthorchid species (Trematoda, Opisthorchiidae). Parasitol. Int. 61, 87-89.

Zhou, P., Silverstein, K.A., Ramaraj, T., Guhlin, J., Denny, R., Liu, J., Farmer, A.D., Steele, K.P., Stupar, R.M., Miller, J.R., Tiffin, P., Mudge, J., Young, N.D., 2017. Exploring structural variation and gene family architecture with *de novo* assemblies of 15 Medicago genomes. BMC Genomics 18, 261.

Table 3-1 Characteristics of the *Clonorchis sinensis* draft genomes.

Characteristics of genome assembly	Korean isolate (*Cs*-k2)	Chinese isolate (*Cs*-c2)
Total size of scaffolds (before post-assembly*) (bp)	562,768,885 (601,230,684)	547,288,241
Number of scaffolds (before post-assembly)	2,776 (218,683)	4,348
Longest scaffold (before post-assembly) (bp)	8,861,937 (2,207,854)	2,050,842
Shortest scaffold (before post-assembly) (bp)	501(100)	300
Number of scaffolds > 1kb; > 100kb; > 1Mb	1564; 542; 204 (7113; 1598; 37)	4,016; 1,411; 44
Mean/median scaffold size (bp)	202,727/1,175 (2,749/135)	125,871/ 23,185
N50 scaffold length (before post-assembly) (bp)	1,628,761 (313,682)	417,486
Genomic DNA GC content (excluding Ns)	43.95% (43.84%)	44.05%
N content	3.16% (3.92%)	0.03%

*Statistics of initial assembly

Table 3-2 Features of the *Clonorchis sinensis* gene sets.

Gene set features	Korean isolate (*Cs*-k2)[a]	Chinese isolate (*Cs*-c2)[b]
Gene number (transferred genes)	14,538 (804)	13,634
Genes supported by transcriptomic data	13,688	-
Gene length (average bp ± s.d.[c])	15,556 ± 17369	17,761 ± 19,228
Coding domain length (average bp ± s.d.)	1,437 ± 1464	1,591 ± 1,567
Exon length (average bp ± s.d.)	234 ± 323	232 ± 297
Exon number per gene (average ± s.d.)	6.14 ± 5.62	6.8 ± 6.40
Complete BUSCOs (%)	555 (65%)	493 (58%)
Complete duplicated BUSCOs (%)	64 (7.5%)	62 (7.3%)
Fragmented BUSCOs (%)	79 (9.3%)	116 (13%)
Missing BUSCOs (%)	209 (24%)	234 (27%)
Total BUSCO groups searched (%)	843 (100%)	843 (100%)
SCGPs[d]	7,886	7,886
SCGPs with > 50% alignment coverage	6,733	6,733
SCGPs with > 90% alignment coverage	3,421	3,421
SCGPs with 100% alignment coverage	2,464	2,464

[a] Filtered and curated gene set of a Korean isolate, *Cs*-k2

[b] Published gene set of a Chinese isolate, *Cs*-c2

[c] Standard deviation

[d] Single copy gene pairs identified in the two gene sets

Table 3-3 Pairwise nucleotide identity and synteny between *Clonorchis sinensis* draft genomes.

Block features	Korean isolate (*Cs*-k2)	Chinese isolate (*Cs*-c2)
Length of each aligned block (masked; unmasked)[a]	1,323 ± 1,319; 9,601 ± 11,807	1,322 ± 1,317; 9,591 ± 11,799
Total aligned nucleotides (masked; unmasked) [b]	319,305,900 (56.73%); 512,007,250 (90.17%)	318,806,533 (58.25%); 511,940,867 (93.54%)
Aligned scaffolds (masked; unmasked)[c]	610; 640	2142; 2374
Total size of scaffolds with aligned nucleotides[b] (masked; unmasked)	557,353,501 (99.04%); 558,116,334 (99.17%)	535,103,559 (97.77%); 539,840,071 (98.64%)
Number of syntenic blocks	1,827	1,827
Total number of SCGPs in syntenic blocks	6,421	6,421
Blocks per scaffold (min; max; average)	2; 18; 3	2; 18; 3
Block length	155,915 ± 130,281	153,712 ± 128,164
Total length of syntenic blocks[b]	284,856,993 (50.6%)	280,831,525 (51.3%)
Number of syntenic scaffolds	484	1262
Total length of syntenic scaffolds[b]	540,285,816 (96.0%)	454,624,656 (83.1%)
Average identity of blocks	99.10%	99.10%

[a] Average nucleotide base pairs (bp) ± standard deviation
[b] Length (bp) and percentage of the total size of scaffolds
[c] Scaffolds containing one or more syntenic blocks

Table 3-4 Summary of nucleotide differences (NDs) in annotated, variable single copy gene pairs with ⩽ 98% pairwise nucleotide identity in coding domains.

GeneID	Non-synonymous NDs (%)	Synonymous NDs (%)	Identities (%)	Length of coding domains (bp)	Gene product
CSKR_13438s	31 (2.82%)	11 (1.00%)	96.17%	1,098	Cathepsin D
CSKR_9510s	5 (1.37%)	6 (1.64%)	96.99%	366	Niemann-Pick C2 protein
CSKR_11331s	9 (2.14%)	2 (0.48%)	97.38%	420	Niemann-Pick C2 protein
CSKR_709s	1 (0.22%)	11 (2.38%)	97.40%	462	Calponin putative
CSKR_482s	3 (0.79%)	6 (1.59%)	97.62%	378	Histone H2A
CSKR_2578s	23 (1.12%)	21 (1.02%)	97.86%	2,058	Protein FAM161A

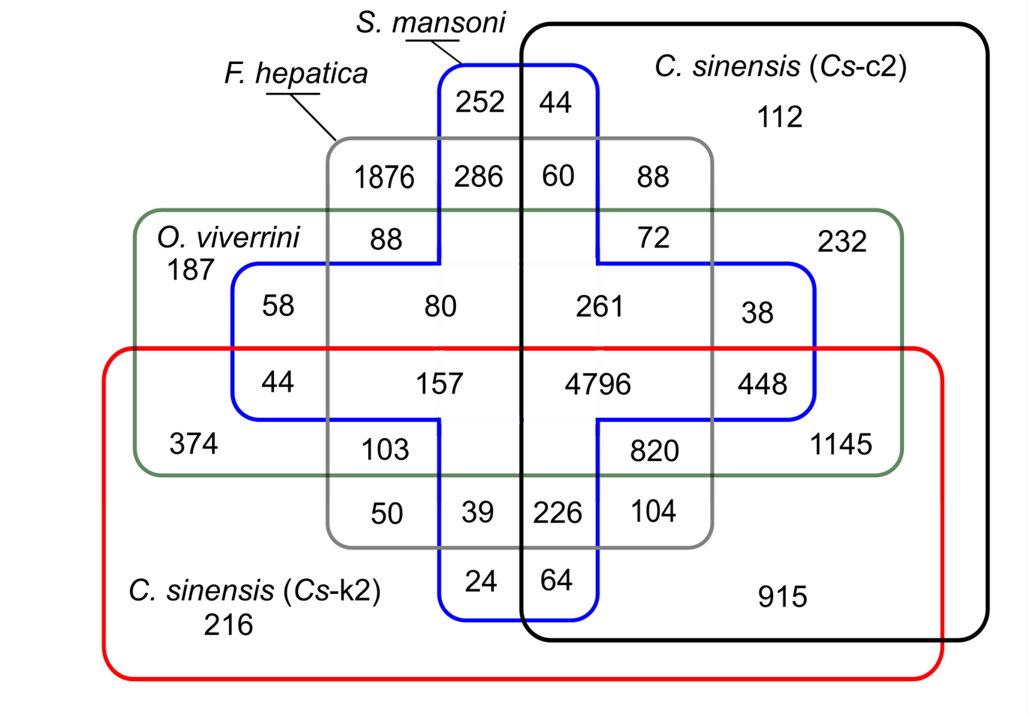


Fig. 3-1 Comparison of the predicted proteomes of Korean and Chinese *Clonorchis sinensis* isolates with respect to other trematode species.

Venn diagram of orthologous protein groups between Korean (*Cs*-k2) and Chinese (*Cs*-c2) *C. sinensis* isolates and *Opisthorchis viverrini, Fasciola hepatica and Schistosoma mansoni.* The numbers represent the number of orthologous groups shared between these species/isolates.

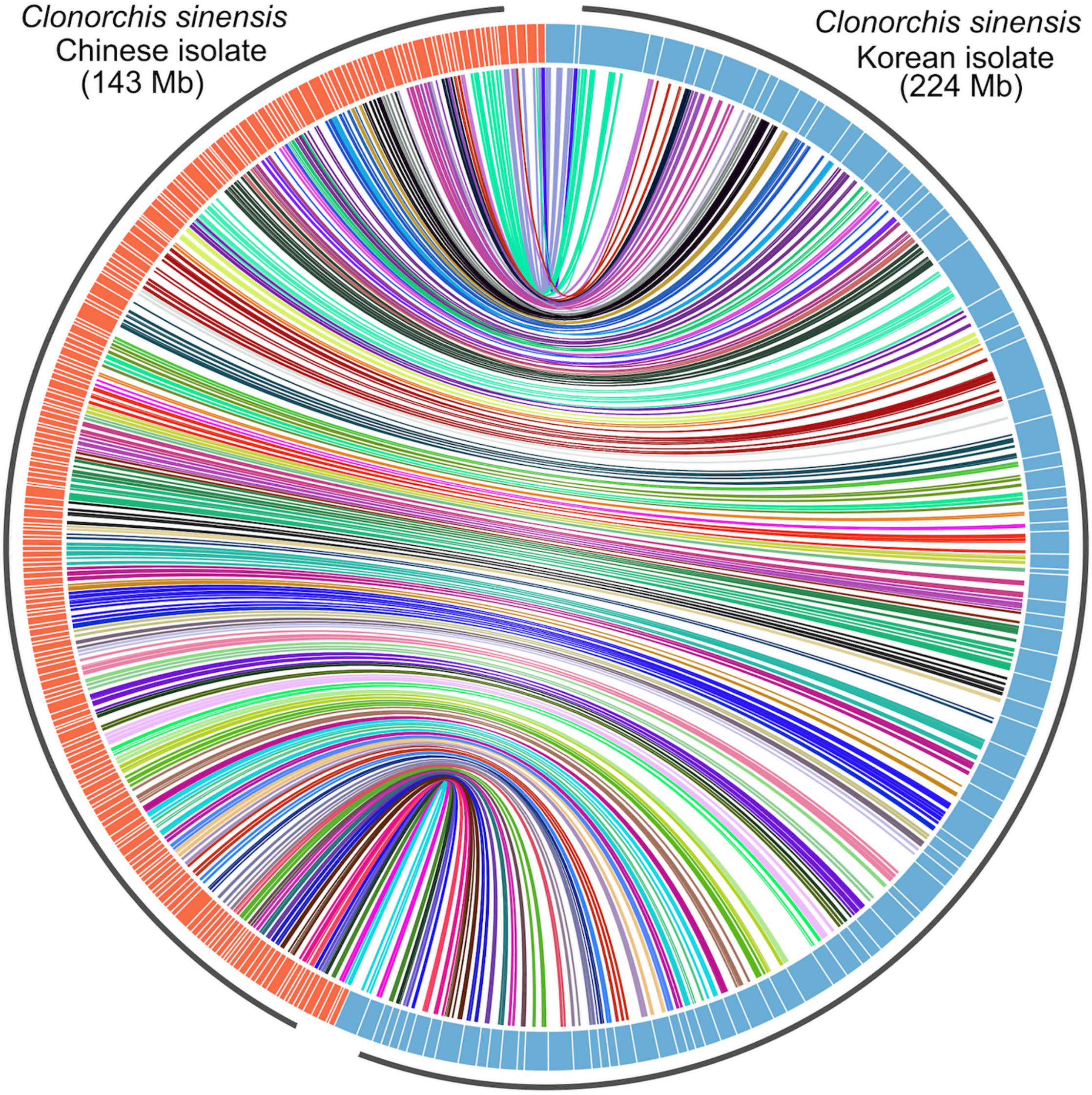


Fig. 3-2 Circos plot of syntenic blocks aligned between draft genomes of Korean and Chinese isolates of *Clonorchis sinensis.*

Each blue segment is a *Clonorchis sinensis* Korea isolate (*Cs*-k2) scaffold with at least 10 single-copy gene pairs (SCGPs) conserved within aligned *C. sinensis* China isolate (*Cs*-c2) scaffolds. Each red segment is a *Cs*-c2 scaffold with at least 5 SCGPs aligned within *Cs*-k2 scaffolds. Each line connecting *Cs*-k2 and *Cs*-c2 scaffolds represents a SCGP between the two genomes with a line colour unique to each *Cs*-k2 scaffold.

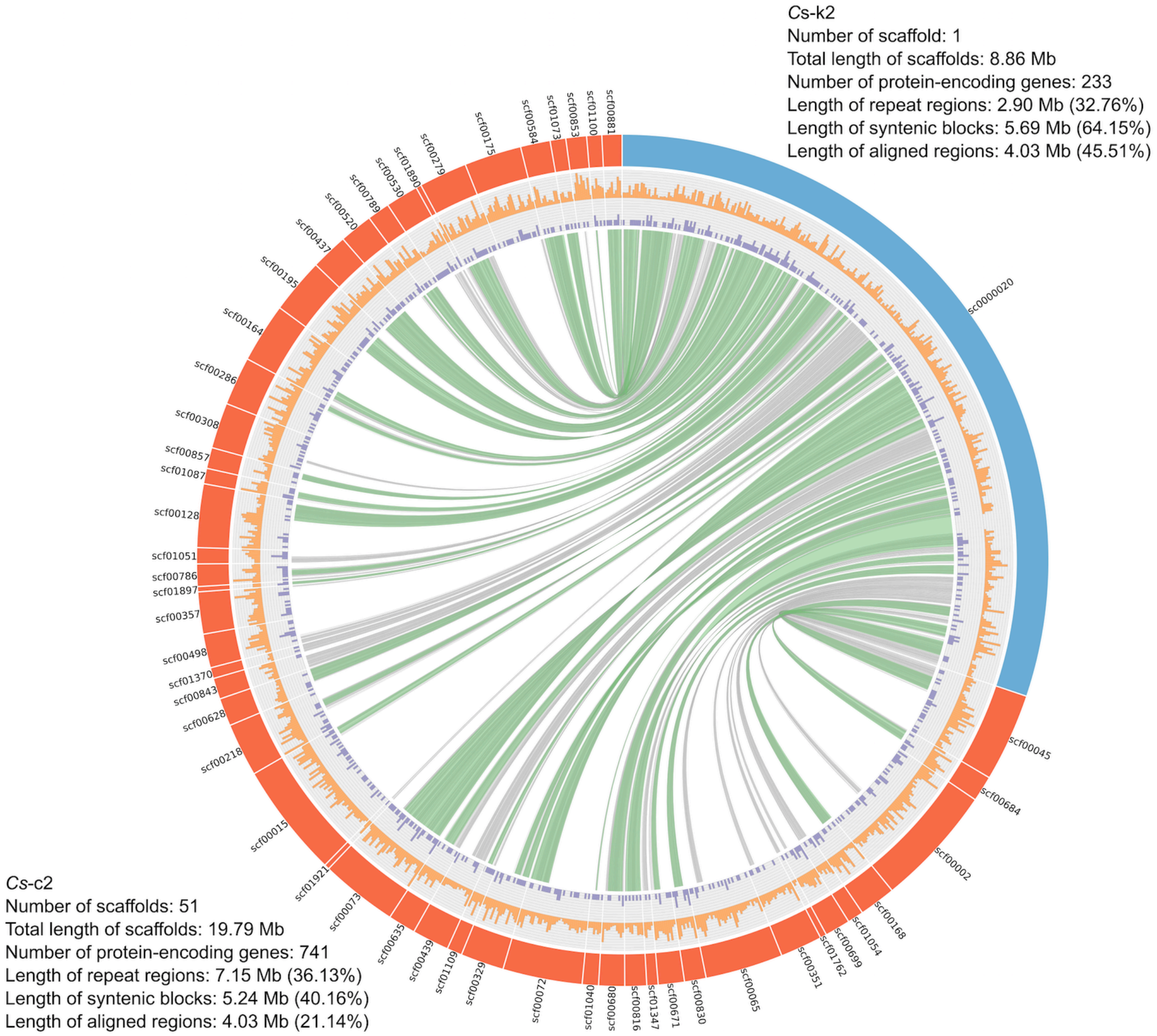


Fig. 3-3 Representative Circos plot of nuclear and syntenic block alignment between draft genomes of Korean and Chinese isolates of *Clonorchis sinensis*.
Summary of the *Clonorchis sinensis* China isolate (*Cs*-c2; red scaffolds) scaffolds > 10 kbp aligned to the largest *C. sinensis* Korea isolate (*Cs*-k2; sc0000020; blue scaffold) scaffold. Light grey ribbons connect *Cs*-k2 and *Cs*-c2 scaffolds within aligned nucleotide blocks (length > 200bp). Light green ribbons connect *Cs*-k2 and *Cs*-c2 scaffolds aligned in syntenic blocks with conserved gene order. The outer histogram track (orange bar chart) represents the percentage coverage of repetitive regions in non-overlapping 25 kb sliding windows. The inner histogram track (purple bar chart) represents the total number of genes in non-overlapping 25 kb sliding windows.

Chapter 4 - *Schistosoma japonicum*: improving our understanding of population genetic variation in China using an expanded genomic data set

Abstract

Schistosoma japonicum is an important blood fluke that infects humans and more than 40 animal species in East Asia. The implementation of schistosomiasis control over the past decades has significantly reduced the prevalence of schistosomiasis japonica in humans and bovines in China. The changing distribution of intermediate hosts of *S. japonicum* caused by the construction of the Three Gorges Dam (TGD) and floods highlights a need to re-examine the population structuring of this parasite. Here, we inferred the genetic relationships between two independent nuclear genomic data sets for *S. japonicum*, and estimated selective pressures on *S. japonicum* populations. We confirmed significant genetic differentiation between isolates from Southwest region (SW) and central and lower reaches of the Yangtze River (CL). We observed inconsistent topologies between concatenated and coalesced phylogenetic trees constructed, suggesting incomplete lineage sorting (ILS) and/or frequent introgression of this parasite. We also identified genes that appear to relate to local adaptations of the life cycle of *S. japonicum*. The present study highlights the technical challenges associated with combining independent data sets and estimating genomic variation within and among *S. japonicum* populations represented by isolates (of pooled adult worms). Future population genetic studies should focus on genomic sequencing from individual worms.

4.1. Introduction

Schistosomiasis is a neglected tropical disease that affects more than 250 million people globally (Colley et al., 2014; WHO, 2015). Of six species of *Schistosoma* which can cause disease in humans, *Schistosoma japonicum* is the only schistosome for which zoonotic transmission is considered important in East Asia. After asexual production in a snail intermediate host, the cercariae of this parasite can infect humans as well as a range of species of wild and domesticated animals as definitive hosts to complete its complex lifecycle. *S. japonicum* is endemic to regions along the central and lower reaches of the Yangtze River (CL) and in the mountainous regions of Southwest China (SW). It is also present in parts of the Philippines and parts of Indonesia (Colley et al., 2014; Soares Magalhaes et al., 2014).

In China, with over 60 years of nationwide schistosomiasis control and environmental modification (Wang et al., 2009; Collins et al., 2012; Cao et al., 2016; Xu et al., 2016), the transmission source of this parasite has been constantly changing. For example, in humans, the number of infected patients in China has reduced from 12 million in 1950s to 77,200 in 2014 (Ross et al., 2001; Sun et al., 2017; Wang et al., 2017), and domestic animals have become the most significant reservoirs for zoonotic transmission (Rudge et al., 2013). In China, more than 75% of current transmission is attributable to bovines in the CL regions (Guo et al., 2006; Gray et al., 2009). The distribution of intermediate hosts of *S. japonicum*, *Oncomelania hupensis*, also determines the current distribution of this parasite. Since the 1950s, the geographical range of *O. hupensis* has reduced by 25% (Shi et al., 2016), particularly following the completion of the Three Gorges Dam (TGD) in 2006. A reduction in water flow along the Yangtze River and its tributaries correlated with a reduction in the prevalence and density of snails, likely as a result of marshlands receiving less water each summer (Zhou et al., 2016). These changes have been predicted to accelerate the reduction of *S. japonicum* populations (Zhou et al., 2016). However, previous studies provide conflicting evidence for (Ding et al., 2017) and against (Yin et al., 2016a) recent bottlenecks in *S. japonicum* populations in the CL regions. To better understand transmission and population dynamics of *S. japonicum* in these regions, in-depth investigations of the population genetic structure(s) of this parasite are required (Steinauer et al., 2010).

Studies have shown significant genetic differentiation between populations of *S. japonicum* from SW and CL regions of China (Shrivastava et al., 2005; Zhao et al., 2009; Zhao et al., 2012; Yin et al., 2015a; Young et al., 2015). Based on the phylogeny of 119 mitochondrial genomes of isolates from major endemic areas, Yin *et al.* (2015) concluded that SW and other southeast Asian lineages originated from one of the CL populations, coinciding with human migration and the spread of rice planting. This hypothesis has been further supported by the relative lower genetic diversity found in SW regions compared with that of CL populations (Zhao et al., 2012; Yin et al., 2015a; Yin et al., 2015b), which could be a signal of founder effects, if the novel population was established by a small number of individuals from the original population (Ellegren and Galtier, 2016). Presently, the genetic structure of *S. japonicum* within CL regions remains unresolved (Yin et al., 2015a). For instance, many studies speculate that most CL regions actually comprise multiple, "migrated" *S. japonicum* lineages (Shrivastava et al., 2005; Rudge et al., 2009; Lu et al., 2010; Zhao et al., 2012). This hypothesis is consistent with a diversification of the intermediate snail host of *S. japonicum* in the CL regions (Zhao et al., 2010), suggesting to be the result of the frequent flooding of the Yangtze River and "driving" the proposed widespread dispersion and complex migration of *S. japonicum* (see Zhao et al., 2012; Attwood et al., 2015). Due to the evidence of genetic hybridisation in some schistosomes (Webster et al., 2013; Leger and Webster, 2017), the dispersion hypothesis and potential hybridisation between sibling lineages of *S. japonicum* warrants testing.

Hypotheses regarding dispersal of *S. japonicum* have been tested mostly using microsatellite and/or mitochondrial sequence data sets (Zhao et al., 2012; Attwood et al., 2015; Bian et al., 2015; Yin et al., 2015a; Yin et al., 2015b; Yin et al., 2016a; Ding et al., 2017). Mitochondrial gene sequences do not provide phylogenetic data that is sufficiently informative to resolve genetic relationships of population variants of *S. japonicum*, particularly in the CL reaches (Zhao et al., 2012; Young et al., 2015). Furthermore, conclusions derived from the analysis of microsatellite markers have been inconsistent with those of mitochondrial markers (Yin et al., 2015a). To address such limitations, genetic variation has been explored in protein-coding genes of the nuclear genome, including genes essential for adapting to different habitat types and/or a novel subspecies of the intermediate host (Li et al., 2017). For example, a single gene encoding a tegumental protein, *SjT*22.6, was reported to be under positive selection and readily differentiated SW and CL populations of *S. japonicum* (Li et al., 2017). This finding was consistent with a whole-nuclear genome investigation which explored variation in functional genes in a genome-wide manner, and further assisted in establishing the evolutionary relationships of geographically distinct *S. japonicum* isolates using a high throughput sequencing-based approach (Young et al., 2015; Yin et al., 2016b). However, inconsistencies in predicted relationships were still observed among independent studies. For example, although two lake/marshland isolates (AHGC and SJ2) were collected from the same county (Guichi, Anhui Province), AHGC clustered with the SW mountainous populations in one study (Yin et al., 2016b), whereas SJ2 clustered with the CL lake/marshland populations in another (Young et al., 2015). These inconsistencies might have resulted from biological factors (micro-habitat of isolates or well-mixed populations), but they may also relate to technical issues pertaining to sample size or genome sequence coverage. Clearly, these aspects warrant further exploration.

In the present study, we re-examined the structuring and substructuring of *S. japonicum* population in the CL and SW regions of China through analyses of two independent nuclear genomic data sets. These data were used to construct concatenated and coalesced phylogenetic trees using protein-coding or intronic regions, and explore intra- and inter-species selective pressures on genes of *S. japonicum*. Through these efforts, we revealed a clear genetic differentiation between SW and CL populations, and a predominant selective pressure on the *S. japonicum* genome. The present study highlights some salient technical challenges associated with selecting a reliable approach for phylogenetic analyses and population genomic studies of *S. japonicum*, considering the potential hybridisation of mixed lineages and dynamic changes in the rate of evolution (e.g., local adaptation and relaxed selection) in particular contexts.

4.2. Material and methods

Schistosoma japonicum **samples**

Sixteen isolates of *S. japonicum* adults were used in this study. Nine of these isolates (Group 1; labelled with a prefix “N”) represented eight populations from regions along the central and lower reaches of the Yangtze River (CL) and one mountainous population from SW (Fig. 4-1A; Table 4-1; Supplementary File 4-1). These isolates were collected as part of a previous study (Zhao et al., 2012) aimed at continuously sampling population diversity of *S. japonicum* and *O. hupensis*. Briefly, adult worms were raised in mice, with each mouse infected with 40 cercariae shed from multiple snails representing each population/location (Supplementary File 4-1). Six weeks following the infection, adult worms were retrieved by perfusion from mesenteric veins using 0.9% NaCl, pooled, washed extensively in saline and stored in 95% ethanol at 4°C. Group 2 included seven isolates of adult worms (SJ1-SJ7; Table 4-1) that were previously whole genome sequenced (Bioproject PRJNA286685; Young et al., 2015), including five populations (SJ1-SJ5) from CL regions and two mountainous populations (SJ6 and SJ7) in the SW (Fig. 4-1A). Adult worms for each location were produced in rabbits infected with 1000 cercariae from 10 snails (3 snails for SJ6) in 1999 (Chilton et al., 1999). SJ1 had been passaged in the laboratory through mice and snails for 20 years (Chilton et al., 1999).

Genomic DNA library construction and sequencing of Group1 adult worms

Group 1 worms representing each population were pooled for DNA isolation (Supplementary File 4-1) using established method (Zhao et al., 2012). Total DNA was quantified by Nanodrop (ThermoFisher), and DNA integrity was assessed by agarose gel electrophoresis. High-quality genomic DNA was used to construct short-insert genomic DNA libraries using a TruSeq DNA library construction kit (Illumina) and paired-end sequenced as 100 nucleotide (nt) reads using the HiSeq-2500 platform (Illumina).

Sequence read mapping, library normalisation, and identification and curation of nucleotide polymorphisms

First, low quality bases (Phred quality: < 25), adapters and reads of < 50 nt in length were removed using Trimmomatic v.0.32 (Bolger et al., 2014) and sequence quality was confirmed using FastQC v.0.11.2 (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Next, high-quality reads were mapped to scaffolds of the published reference genome of *S. japonicum* (SjRef; Bioproject PRJEA34885) (Zhou et al., 2009) using Bowtie2 v.2.2.5 (Langmead and Salzberg, 2012), and read alignments were stored in the BAM format. Each BAM file was sorted, duplicates were removed, insertion-deletion events curated and the quality of read alignments established using PICARD tools v.1.123

(http://broadinstitute.github.io/picard) according to best practice (i.e. GATK guidelines) (McKenna et al., 2010). To adjust for variability in library sizes between the datasets derived from Groups 1 and 2, BAM files were down-sampled, so that each library contained 21 million random pairs of reads that were mapped to the *S. japoncium* reference genome using SAMtools v.1.6 (Li et al., 2009). An MPILEUP format file was created from each down-sampled BAM file using SAMtools, and the frequencies of SNPs were estimated using the program VarScan v.2.3.7 (Koboldt et al., 2009) with the following settings: --min-coverage: 5 -- min-reads: 5 -- p-value: 0.95, and stored in VCF format. In each VCF file, SNPs were removed if the reference allele had an ambiguous nucleotide (N) or if SNPs were common to all Group 1 or all Group 2 datasets. Reported SNP variants were annotated based on their genomic locations and predicted coding effects using SnpEff v.4.0e (Cingolani et al., 2012) and a GFF annotation file available for the reference genome SjRef.

Selection of consensus SNPs and construction of *S. japonicum* gene groups

Nucleotide positions in the reference genome with a SNP at greater than 50% allelic frequency in at least one library (consensus SNP) were selected for further processing. When a consensus SNP was called by VarScan v.2.3.7 (Koboldt et al., 2009) in one or more library, alternative alleles in any other library with greater than 50% allelic frequency were also recorded as a consensus SNP. To predict the coding domains for each *S. japonicum* library, consensus SNPs were transferred to the reference genome sequence using VCFtools v.0.1.12b, vcf-consensus (Danecek et al., 2011) and coding domains and amino acid sequences were extracted from each genome using GAG v.2.0.1 (http://genomeannotation.github.io/GAG) and the annotated SjRef genome (Zhou et al., 2009). Intronic sequences were extracted from each genome using the program gffread (Trapnell et al., 2010) and the SjRef genome. For both CRs and IRs, the intersections between shared and unique consensus SNPs among the 16 isolates were summarised and visualised using the R package UpSetR v.1.3.3 (Lex et al., 2014). Coding and intronic consensus SNPs were also subjected to principal component analysis (PCA) using the R package SNPRelate v 1.8.0 (Zheng et al., 2012).

Single-copy gene groups (SCGGs) in coding and intronic regions were selected from the *S. japonicum* gene sets if: (1) < 80% nucleotide identity to other *S. japonicum* coding domain; (2) < 10% ambiguous nucleotides (Ns) within coding or intronic domains; (3)

coverage of ⩾ 5 mapped reads to more than 80% of the coding or intronic domains in all libraries; and (4) no internal stop codons within the SjRef reference coding domain.

Assessing genetic variation and construction of phylogenetic trees

To compare geographic and genetic distances among isolates, the pairwise geographic distance between isolates was calculated from the longitude and latitude of the sampling locations and using the R package "geosphere" (https://github.com/cran/geosphere). Pairwise genetic distance between isolates was estimated from the concatenated coding and intronic sequences of SCCGs and using the "dist.dna" function within the R package "ape" v.5.1 (Paradis et al., 2004) and applying the with TN93 nucleotide substitution model. Spearman's rank correlation coefficient between geographic and genetic distance among all isolate pairs, isolate pairs within CL or SW and between CL and SW were calculated in R v.3.3.2 (Team, 2013) and plotted using the R package "ggplot2" (Wickham and Chang, 2008).

Coding (CR) and intronic (IR) nucleotide regions of SCGGs were subjected to phylogenomic analysis using Bayesian inference (BI), maximum likelihood (ML) and coalescent-based summary methods. For BI, the concatenated CRs or IRs were subjected to BI analysis using MrBayes v.3.2.2 (Ronquist et al., 2012), with the general time reversible (GTR) and invariant and discrete gamma model. Trees were constructed, employing the Monte Carlo Markov chain method (n chains = 4) over 500,000 generations, with every 200th tree being saved; 25% of the first saved trees were discarded to ensure a convergence of the nodal split frequencies. Consensus (50% majority rule) trees were constructed from all remaining trees, with nodal support expressed as a posterior probability (pp). For ML, the same concatenated sequences were subjected to the phylogenetics program RAxML v.8.2.9 (Stamatakis, 2014) using GTR model, and setting four discrete rate categories. Nodal support values were inferred from 100 bootstrap replicates. Consensus trees (50% majority rule) were created using sumtrees.py in the DendroPy v.4.1.0 (Sukumaran and Holder, 2010). Trees were displayed and re-labelled using the program Figtree v.1.4 (http://tree.bio.ed.ac.uk/software/figtree/).

A coalescence approach to tree construction was also applied, following established methods (Jarvis et al., 2014; Mirarab et al., 2014). First, individual majority rule gene trees were constructed from the CRs or IRs of each gene using RAxML and using the

GTR substitution model. Then, CR or IR gene trees were subjected to the program ASTRAL v.3 (Mirarab et al., 2014) to construct a coalesced "summary" tree. To construct such a tree, the topologies of all possible quartet trees around internal branches were compared, and concordant branches among genes were used to construct the final tree. Posterior probability (PP) support for an internal branch was estimated from the frequency of quartet trees sharing compatible topologies (Sayyari and Mirarab, 2016). Quartet frequencies were used to derive estimates of internal branch lengths in coalescent units (Degnan and Rosenberg, 2006).

Calculation of interspecific divergence and intraspecific polymorphism

Single-copy orthologous (SCO) coding sequences shared by *S. japonicum* and *S. mansoni* (Berriman et al., 2009) were identified using the program OrthoMCL (v.2.0.4) (Li et al., 2003) using default settings. The resultant groups that contained only one *S. japonicum* gene and one *S. mansoni* gene were selected as SCOs. Codons of *S. japonicum* and *S. mansoni* orthologs were aligned using PRANK v.170427 using default settings (Loytynoja and Goldman, 2010). Gaps that were longer than 1/3 of an alignment length were removed. The average number of nonsynonymous and synonymous substitutions per site between *S. japonicum* and *S. mansoni* (d_N, d_S) and the average number of nonsynonymous and synonymous polymorphism per site within *S. japonicum* (π_N, π_S) were calculated employing an established method (Chen et al., 2009). Nonsynonymous or synonymous sites were calculated using nucleotide differences and an established protocol (Ina, 1995). Nucleotide diversity (π) of each coding sequence was estimated by calculating the average pairwise nucleotide differences per site. Spearman correlation coefficient and Fisher's exact test were performed in R v.3.3.2 (Team, 2013).

The McDonald-Kreitman test was used to compare interspecies nonsynonymous (Dn) and synonymous (Ds) divergent bases with intraspecies nonsynonymous (Pn) and synonymous (Ps) polymorphic bases (McDonald and Kreitman, 1991). The nonsynonymous and synonymous nucleotide differences in retained codon-based *S. japonicum* and *S. mansoni* orthologue alignments were designated as Dn and Ds, respectively. Pn and Ps were calculated based on the consensus SNPs that were combined and merged from all *S. japonicum* data sets. For each gene, Dn, Ds, Pn, Ps values constitute a 2×2 contingency, the significance of which was determined using Fisher's exact test (p < 0.05). All the significant SCGGs were subjected to the test of direction of selective forces: DoS = Dn/(Dn+Ds)-Pn/(Pn+Ps) (Stoletzki and Eyre-Walker, 2011). Positive and negative DoS values indicate adaptive and balancing/relaxed selection, respectively. Significantly enriched protein families containing representatives with homology (BLASTP, E-value, 1e-05) to proteins submitted to the Kyoto Encyclopedia of Genes and Genomes (KEGG) Brite hierarchy database (Kanehisa et al., 2011; Xie et al., 2011) were determined using the Fisher's exact test ($p < 0.05$).

We compared the frequency of the consensus SNPs between SW and CL isolates, to identify fixed SNPs in each region. SNPs present in all SW isolates, but absent from most CL isolates (n ≥ 10 of 13 isolates) were inferred to be fixed SNPs in SW population. SNPs present in most CL isolates (n ≥ 10 of 13 isolates), but were missing from the SW isolates were inferred to be fixed SNPs in CL populations. SCGGs with more than 3 non-synonymous fixed SNPs were recorded to relate to high amino-acid variability. Functional annotation of the these SCGGs was inferred from a previous annotation of *S. japonicum* (Young et al., 2015).

4.3. Results

Illumina sequencing data and pre-processing of reads

Nine Group 1 libraries were constructed and whole-genome sequenced. Each Group 1 library contained 8 to 12 Gb of clean sequence data (NCBI BioProject accession number: PRJNA354903), comprising 71.7-109.8 million high quality reads (Supplementary File 4-1). Overall, 80.4% to 89.2% of these reads mapped to the reference genome, with 80.4% to 89.2% of them mapping as pairs (Supplementary File 4-1). Whole genome sequence data for each Group 2 library contained 16 to 20 Gb of high quality genomic sequence data, comprising 159.2 to 203.6 million high quality reads (NCBI BioProject accession number: PRJNA286685; (Young et al., 2015)), and 82.6% to 88.0% of these reads mapped to the reference genome, and ~95% mapped as pairs. Because of the observed differences in sequence mapping depth between Group 1 (16.94-fold to 26.42-fold) and Group 2 (34.73-fold to 41.99-fold) libraries

(Supplementary File 4-2), all 16 libraries were normalised by down-sampling to 21 million paired reads (42 million reads in total).

SNP calling, consensus SNPs filtration, and coding regions extraction

Normalised data sets were used to infer 1,643,746 ± 262,616.3 final consensus SNPs (ranging from 1,267,537 to 2,092,259 SNPs) with high quality (> 34) and uniform read coverage (Table 4-1). Of these, 71.7% (n = 911,779 to 1,492,913), 26.6% (n = 333,505 to 561,682) and 1.8% (n = 22,253 to 37,664) of SNPs were within intergenic, intronic and protein-coding regions, respectively (Table 4-1).

After filtering for predicted paralogs and excluding gene regions with low read mapping coverage, we curated SNPs located in 7,986,150 bp of mapped coding region from 6,978 SCGGs, and 29,033,261 bp of mapped intronic region from 4,638 of these SCGGs, which collectively represent 9.2% of the *S. japonicum* genome assembly. These SCGGs contained 84,304 to 145,760 consensus SNPs within curated gene regions of all 16 isolates, of which 12% (n = 9,946 to 17,180) and 88% (n = 74,358 to 128,580) were located in coding and intronic regions, respectively. Within the coding regions, we identified 4,579 to 8,143 non-synonymous and 5,367 to 9,037 synonymous SNPs in each isolate (Table 4-2). For all the SCGGs, the average number of SNPs per kb of the coding regions ranged from 1.25 to 2.15 and 2.56 to 4.43 for intronic regions.

Subsequently, we compared the number of consensus SNPs unique to, or shared by, the 16 isolates (Fig. 4-2). The 30 largest unique or intersecting SNP sets were comprised of between 3,822 to 259 and 25,870 to 2,114 SNPs in coding and intronic regions, respectively. The 8 largest SNP sets in coding and intronic regions were unique to the isolates with the most called consensus SNPs when compared to the reference genome (Fig. 4-2). For example, isolates SJ7, SJ6, N1, N8 and N10 contained the most total and unique SNPs in both coding and intronic regions, whereas isolates N24, N31 and SJ3 contained the least SNPs in both regions (Fig. 4-2, Table 4-2). In the intersections of two or more isolates, SW isolates shared the most common SNPs in both coding and intronic regions. For example, SJ7 and SJ6 share 697 and 5,576 common SNPs within coding and intronic regions, respectively, most of which were also common to the third SW isolate (N10). In comparison, very few SNPs (< 250 within coding regions and < 2000 within introns) were common among all 13 CL isolates.

PCA of consensus SNPs in coding and intronic regions showed relatedness among the 16 isolates (Fig. 4-2C,D). The first and second principal components explained

25.19 % of variation in coding regions (Fig. 4-2C) and 23.82 % of variation in intronic regions (Fig. 4-2D). Using both coding and intronic SNPs, we observed clear division of SW and CL isolates, and little apparent population structure among the CL isolates, apart from the divergence of N1 and N8 isolates.

Correlation between geographic and genetic distance

Pairwise comparisons among all 16 isolates suggest a modest but significant correlation (n = 120 comparisons; $r_s = 0.58$; p-value $< 1\times10^{-11}$) between genetic and geographic distances (Fig. 4-3). However, no significant correlation was observed for the pairwise comparison among CL (n = 78 comparisons; $r_s = 0.16$; p-value = 0.16) and SW (n = 3 comparisons; $r_s = 1$; p-value = 0.33) isolates alone or pairwise comparison between the SW and CL isolates (n = 39 comparisons; $r_s = 0.26$; p-value = 0.11).

Phylogenomic relationships of 16 *S. japonicum* study populations

Concatenated coding and intronic data sets. Concatenated sequences of: (1) coding regions (CRs) and (2) intronic regions (IRs) of consensus SCGGs were subjected to Bayesian inference (BI) and maximum likelihood (ML) analyses. Maximum likelihood trees inferred from coding regions (CR-ML) and intronic regions (IR-ML) had variable tree topologies (Fig. 4-4A,B). For both ML analyses, SW isolates formed a well-supported group, to the exclusion of all CL isolates (bootstrap values > 90%), although the relationships between the three SW isolates in this group were not entirely consistent (Fig. 4-4A,B). For the CL isolates, N8 and N1, N37 and N21 and SJ4 and SJ5 grouped with high nodal support (> 50 %). In the IR-ML tree, isolate N22 was positioned basal to other CL isolates, with strong support, and for Group 2 isolates, SJ1 to SJ5 consistently clustered with high nodal support (BS >90%). The trees build from coding region data by BI did not converge, and thus, no consensus tree was produced. The tree constructed by BI analysis of intronic data converged to produce a tree that was consistent with that built by ML (Fig. 4-4B). In the IR-BI tree grouped N1, N8, N21, N37, N9 and N31 with high nodal support (posterior probability = 0.997). Here, N9 and N31 grouped with strong nodal support (posterior probability = 1). Hence, the reliability and accuracy of the inferred trees varied markedly. For example, the CR-ML tree only had 5 of 13 possible internal nodes with BS > 90%, whereas the IR-ML tree had 9 of 13 possible internal nodes with bootstrap (BS) > 90% (Table 4-3). In addition, the IR trees had longer internal branch lengths (IR 0.00083; CR 0.00062) and lower negative log likelihood values (IR -45,666,864 ± 5290.55; CR -11,779,789 ± 5088.11) than the CR trees.

Coalesced gene tree. To investigate how the phylogenetic signals of each gene affect the topology of ML and BI trees, we constructed a coalesced tree for each data set (CR, IR) using trees constructed from individual SCGGs. Average bootstrap support for nodes in individual gene trees was low (CR 2.3%; IR 19.2%; Table 4-3). An unrooted, coalesced summary tree (CS) tree was then inferred for CR and IR data sets with the ML trees for each bin (Figure 4-4C,D). Agreement between gene trees and their corresponding CS tree was reported as the proportion of quartet tree topologies congruent with the input gene trees and summarised as a normalised branch quartet score which was used to derive localised posterior probability support for each node (Fig 4-4C,D and Table 4-3). Generally, the nodal support for CR-CS (Fig. 4-4C) and IR-CS (Fig. 4-4D) trees was low. Although some groups were consistent among CS, ML and/or BI trees, the nodes within the CS trees frequently had lower PP support (from 0.37 to 1 PP). In addition, both CS trees had fewer supported branches than the CR and IR trees, with 4 of 13 and 5 of 13 branches with PP > 0.95 in the CR-CS and IR-CS tree, respectively. Approximately 40% of the quartet trees (normalised branch quartet score ~ 0.4) were represented in the CS trees for both data sets (Fig 4-4C,D and Table 4-3). Consistent with the trees constructed from the concatenated coding and intronic data sets, both CS trees supported a clear separation of SW and CL isolates and the grouping of isolates N37 and N21 and SJ4 and SJ5. In addition, the IR-CS tree supports the grouping of isolates SJ1, SJ4 and SJ5 as well as N1 and N8.

Genetic divergence and polymorphism within and between schistosome species

Genetic variation within the protein-coding region of each *S. japonicum* SCCG was characterised by quantifying and comparing average nucleotide diversity of synonymous (π_S) and nonsynonymous polymorphisms (π_N) within species isolates (Table 4-4). Of the curated 6,978 *S. japonicum* SCGGs, 1,564 contained no synonymous polymorphisms ($\pi_S = 0$) and were excluded from further analysis. The remaining 5,414 SCGGs had an average nucleotide diversity (π) of 0.0021 ± 0.0020 (average ± standard deviation) and a π_N/π_S ratio of 0.2995 ± 0.4372 (average ± standard deviation) (Table 4-4). Only a weak but significant correlation ($r_s = 0.28$; *p*-value <

2.2e-16) was observed when π and π_N/π_S were compared for each *S. japonicum* gene (Table 4-4). A significant correlation between π and π_N/π_S was also observed when considering distinct geographic regions (CL isolates, $r_s = 0.27$; *p*-value = 2.2e-16 and SW isolates, $r_s = 0.37$; *p*-value = 2.2e-16).

Subsequently, we characterised genetic variation between schistosome species, using the *S. mansoni* reference genome (Bioproject PRJEA36577) (Berriman et al., 2009; Protasio et al., 2012) as a representative of this species. First, we identified 3,317 predicted *S. japonicum* and *S. mansoni* orthologues from the 6,978 *S. japonicum* SCGGs (Table 4-4) and removed 567 genes containing no interspecies synonymous polymorphisms (d_S). We then quantified the nucleotide diversity between species at both nonsynonymous and synonymous positions in the remaining 2,750 orthologues. Between the two schistosomes, the average number of nonsynonymous (d_N) and synonymous (d_S) substitutions per site was 0.1493 ± 0.0913 and 0.7368 ± 0.2511 (average ± standard deviation), respectively. The ratio of nonsynonymous and synonymous nucleotide polymorphisms between species (d_N/d_S) was 0.2186 ± 0.1387 (average ± standard deviation) (Table 4-4) and correlated with nucleotide polymorphisms within all *S. japonicum* isolates (π_N/π_S) ($r_s = 0.45$; *p*-value = 2.2e-16) and within distinct geographic regions (CL isolates , $r_s = 0.45$; *p*-value = 2.2e-16 and SW isolates, $r_s = 0.43$; *p*-value = 2.2e-16) (Fig. 4-5). Differences in interspecific nonsynonymous (Dn) and synonymous (Ds) divergence and intraspecific nonsynonymous (Pn) and synonymous (Ps) polymorphisms were then used to predict that 217 genes were under significant selection pressure (Fisher's exact test; *p*-value < 0.05). Based on estimates of direction of selective force (DoS) for *S. japonicum* and *S. mansoni* orthologues (McDonald-Kreitman Test; Supplementary File 4-4), 193 and 24 of these genes were under adaptive and balancing/relaxed selection, respectively (Fig. 4-5 and Supplementary File 4-4). Only 51.8% (n=100) of these genes encode proteins homologous (BLASTP; E-value < 1e-05) to curated KEGG proteins with protein family and/or pathway annotation. Among the 82 annotated genes under adaptive selection (DoS > 0), no pathways were significantly enriched, and only the transcription factors (ko0003; 14 of 142 transcription factors within orthologues; Supplementary File 4-5) were significantly enriched (Fisher's exact test; *p*-value < 0). Among those transcription factors, three genes contained Cys2-His2 zinc finger (C2H2-ZF) domains, and four contained helix-turn-helix domains.

We also compared the frequency of the consensus SNPs between SW and CL isolates to identify fixed SNPs for each geographic location. We identified 1,060 SNPs in all the SW isolates (SW-SNPs) that were absent from at least ten CL isolates (n ≥ 10 isolates) and 502 SNPs (CL-SNPs) present in at least ten CL isolates (n ≥ 10) but absent from all the SW isolates. For SW isolates, the 1,060 SW-SNPs were within the gene regions of 741 SCGGs, 14 of which had amino-acid variability linked to > 3 non-synonymous SNPs, including a methionine sulfoxide reductase, a G protein-coupled receptor and a calcium-activated potassium channel protein (Table 4-5 and Supplementary File 4-6). For CL isolates, the 502 CL-SNPs were within the gene regions of 434 SCGGs, 3 of which had amino-acid variability, including a glucose-6-phophate 1-dehydrogenase.

4.4. Discussion

Here, we combined two whole genome data sets and re-analysed the population genetic structure of *S. japonicum*. To do this, we constructed concatenated and coalesced phylogenetic trees using nucleotide sequences representing coding and intronic regions. Despite the low phylogenetic resolution of CR trees, the inferred relationships among Group 2 isolates were generally compatible with the previous publication (Young et al., 2015), being consistent in geographic distribution. In contrast, geographic associations were not clear for the Group 1 isolates. Most Group 1 isolates did not group with geographically adjacent isolates, except for N21 and N37, suggesting a complex geographic distribution of *S. japonicum*. Coincidently, a severe flooding event occurred along the Yangtze River in 1998, followed by the completion of the Three Gorges Dam in 2006. Given that the two groups of isolates were sampled at very different time points, marked temporal changes of the distribution of *S. japonicum* might have occurred in the last few decades, leading to a complex dispersal pattern for this parasite in the CL regions of China.

Nonetheless, all phylogenetic trees showed clear genetic differentiation between SW and CL isolates, supporting the hypothesis that SW population originated from one of the CL populations about 5,000 years ago, following human migration and the spread of rice planting (Yin et al., 2015a). A significant correlation was found between genetic and geographic distance among the 16 isolates, which is likely to relate to geographic isolation due to the mountain ranges and river gorges separating the SW and CL regions. The possible effect of these geographical barriers appears to be observed also in the Three Gorges Area located between SW and CL regions, where schistosomiasis is not endemic (Zhou et al., 2010). Another possible explanation for this genetic differentiation would be a local adaptation to distinct subspecies of *O. hupensis* in the mountainous regions of Southwest China (*O. hupensis robertsoni*) *versus* the lake/marshland CL regions (*O. hupensis hupensi*) (Zhao et al., 2010; Attwood et al., 2015). Given that geographically distinct lineages of *O. hupensis* differ in their susceptibility to *S. japonicum* infection (Cross et al., 1984; He et al., 1990), SW and CL isolates might have gradually adapted to and proposed in respective local environments. During such adaptation, distinct gene alleles might have been favoured and thus led to further genetic distance between SW and CL isolates.

Surprisingly, although the IR and CR trees separated SW and CL isolates, there were clear differences in the phylogenetic support in these trees. Compared with CRs, IR consensus trees had longer internal branch lengths, lower negative log-likelihood values, and more internal nodes with a high nodal support (BS > 90%). IR trees are usually constructed using significantly more phylogenetically informative sites (Jarvis et al., 2014), which was attributed here to a longer alignment length (IR: 29 Mb vs CR: 8 Mb) and a greater SNP density (IR: 2.56 to 4.43 per kb; CR: 1.25 to 2.15 per kb). In complex cases, limited informative sites can create error-prone phylogenetic signals (Xi et al., 2015) and, thus, reduce the accuracy of the tree. Herein, the lower resolution of CR trees may be the result of too few informative sites, but other possibilities, such as convergent evolution, need to be considered (cf. Poulin and Randhawa, 2015), given the reported convergence of genes linked to proteolysis in *S. mansoni*, a related schistosome (Clément et al., 2013). In contrast, the lack of strong natural selection in IR regions (Saeb, 2015) relates to more polymorphic sites, which lends greater power to overcome inconsistent phylogenetic signals and, hence, provides better support for a more reliable phylogeny.

For comparative purposes, consensus SNPs were also used to construct individual gene trees that were summarised as coalesced "summary" (CS) trees for coding and intronic regions. Overall, CS trees were usually congruent with concatenated sequence trees; however, CS trees only represented 40% of the gene trees. Discrepancies between individual gene trees and CS trees are quite common (Kutschera et al., 2014; Shipham et al., 2015) and might reflect a more accurate evolutionary scenario compared with trees based on concatenated sequences. Given the proposed complex dispersal of this parasite in CL regions (Yin et al., 2015a), the low proportion of compatible gene trees could be caused by incomplete lineage sorting (ILS) and/or introgression (Long and Kubatko, 2018). ILS occurs when ancestral polymorphisms retained in descendant lineages do not completely match the evolutionary relationships among lineages, which occur more frequently in recently diverged lineages or lineages derived from a larger ancestral population (Rogers and Gibbs, 2014). Based on the hypothesised evolutionary history of *S. japonicum* (Yin et al., 2015b), large-scale dispersal of this parasite in China was not likely to occur until 22,000 years ago (Yin et al., 2015a), which might have led to a genome-wide ILS in populations or subpopulations.

Another explanation for the phylogenetic discrepancies is introgression. Frequent introgression can result in decreasing posterior probabilities, underestimates of branch length and may alter the topology of a CS tree (Leaché et al., 2013; Kutschera et al., 2014). Both ILS and introgression can strongly confound the inference of a phylogeny. However, disentangling the effect of ILS and introgression is challenging due to their similar "footprints" of genetic variation (Leaché et al., 2013). Nonetheless, our results indicate that phylogenetic relationships that have been inferred using a few genetic markers should be interpreted with caution. With the decreasing cost high throughput sequencing, genome-wide information should be obtained and utilised to achieve robust phylogenies.

Another factor that deeply affects genetic variation is the nature and extent of natural selection (Akey et al., 2004). The relatively low d_N/d_S of *S. japonicum* and *S. mansoni* orthologs shows that most nonsynonymous mutations in these genes were deleterious and, hence, have been removed by negative selection. Surprisingly, we observed a positive correlation between π_N/π_S and d_N/d_S, suggesting a consistent natural selection on both deep and recent timescales. Although most common S. *japonicum*/*S. mansoni* orthologs were inferred to be under purifying selection, 193 orthologs were predicted to be under a long-term, positive selection. Genes encoding transcription factors were significantly enriched, including three containing Cys2-His2 zinc finger (C2H2-ZF) domains. In humans, most genes encoding C2H2-ZF domains are also adaptive and regulate a diverse range of genes (Najafabadi et al., 2015). Given the rapid lineage-specific expansions of the C2H2-ZF family in many species (Emerson and Thomas, 2009), the adaptive C2H2-ZF proteins identified might also have evolved rapidly to regulate existing genes in response to environmental changes. Interestingly, another adaptive gene encoding a methionine sulfoxide reductase (Msr) contains seven nonsynonymous SNPs that were fixed in the SW isolates; an ortholog of this gene has been reported to be highly transcribed in the eggs of *S. mansoni* and predicted to play a role in repairing oxidised methionine residues (Oke et al., 2009). More broadly, this enzyme has been reported to reduce oxidative stress and aid the survival of the schistosomes in the intermediate host (Oke et al., 2009). Given that *S. japonicum* infects two distinct subspecies of *O. hupensis* in SW and CL regions, respectively, the nonsynonymous fixed SNPs in the Msr gene may reflect a recent adaptation to distinct intermediate host taxa (Zhao et al., 2010). Further studies are required to explore the role(s) of the Msr gene in schistosomes and other flatworms and their evolution.

In the present research, we explored genetic variation among *S. japonicum* populations based on the sequencing data of pooled individual worms (Pool-Seq). In such studies, it is possible to infer population statistics for individual pools (isolates), including nucleotide diversity and allelic frequency (Nolte et al., 2013; Guo et al., 2015; Kapun et al., 2016; Dal Grande et al., 2017); however, this approach requires deep sequencing of many (n > 50) pooled, possibly unrelated individuals (cf. Schlötterer et al., 2014). Meeting Pool-Seq requirements for flatworms can be challenging due to their large genome size and clonal replication in the snail intermediate host. Distinct clones of cercariae in individual snails can ultimately lead to significant variation in the genetic composition of adult worms among different definitive host individuals (Prugnolle et al., 2005). Another issue is that population statistics may be unreliable for pools of worms in temporal studies of *S. japonicum* conducted, for example, along the Yangtze River, particularly given that flooding, at distinct time points, might have led to a mixing of distinct *S. japonicum* lineages, complicating interpretation. Hence, instead of characterising population statistics for each pool, we elected to focus on constructing a phylogeny based on consensus SNPs. With the decreasing cost of the high-throughput sequencing, we expect that whole-genome sequencing of large-scale individual worms should be feasible in the near future and enable the inference of genetic relationships among individuals prior to the estimation of population statistics.

4.5. Conclusions

By constructing the phylogeny of *S. japonicum*, we confirmed the genetic differentiation between SW and CL isolates, and revealed evidence of dramatic temporal distributional changes of this parasite in CL regions. Applying multiple phylogenetic techniques, we improved the accuracy of the phylogenetic construction, and identified the importance of intronic regions for future phylogenetic inferences using whole genome sequence data. Within the genome, high proportions of inconsistent gene trees suggest a genome wide ILS and/or a relatively frequent introgression, which might be associated with a complex dispersal history of *S. japonicum*. Interestingly, we identified genes that are under positive selection and found evidence for the adaptation of the parasite to local intermediate hosts. The present study identifies and emphasises the technical challenges associated with studying genetic variation in *S. japonicum* and other flatworms using pooled worms. With a decreasing cost of the high-throughput sequencing, the focus should be on using whole genome data sets from large numbers of individual worms for detailed population genetic investigations.

4.6. References

Akey, J.M., Eberle, M.A., Rieder, M.J., Carlson, C.S., Shriver, M.D., Nickerson, D.A., Kruglyak, L., 2004. Population history and natural selection shape patterns of genetic variation in 132 genes. PLoS Biol. 2, e286.

Attwood, S.W., Ibaraki, M., Saitoh, Y., Nihei, N., Janies, D.A., 2015. Comparative Phylogenetic Studies on *Schistosoma japonicum* and Its Snail Intermediate Host *Oncomelania hupensis*: Origins, Dispersal and Coevolution. PLoS Negl. Trop. Dis. 9, e0003935.

Berriman, M., Haas, B.J., LoVerde, P.T., Wilson, R.A., Dillon, G.P., Cerqueira, G.C., Mashiyama, S.T., Al-Lazikani, B., Andrade, L.F., Ashton, P.D., Aslett, M.A., Bartholomeu, D.C., Blandin, G., Caffrey, C.R., Coghlan, A., Coulson, R., Day, T.A., Delcher, A., DeMarco, R., Djikeng, A., Eyre, T., Gamble, J.A., Ghedin, E., Gu, Y., Hertz-Fowler, C., Hirai, H., Hirai, Y., Houston, R., Ivens, A., Johnston, D.A., Lacerda, D., Macedo, C.D., McVeigh, P., Ning, Z., Oliveira, G., Overington, J.P., Parkhill, J., Pertea, M., Pierce, R.J., Protasio, A.V., Quail, M.A., Rajandream, M.A., Rogers, J., Sajid, M., Salzberg, S.L., Stanke, M., Tivey, A.R., White, O., Williams, D.L., Wortman, J., Wu, W., Zamanian, M., Zerlotini, A., Fraser-Liggett, C.M., Barrell, B.G., El-Sayed, N.M., 2009. The genome of the blood fluke *Schistosoma mansoni*. Nature 460, 352-358.

Bian, C.R., Gao, Y.M., Lamberton, P.H., Lu, D.B., 2015. Comparison of genetic diversity and population structure between two *Schistosoma japonicum* isolates--the field and the laboratory. Parasitol. Res. 114, 2357-2362.

Bolger, A.M., Lohse, M., Usadel, B., 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30, 2114-2120.

Cao, Z.G., Zhao, Y.E., Lee Willingham, A., Wang, T.P., 2016. Towards the elimination of schistosomiasis japonica through control of the disease in domestic animals in the People's Republic of China: a tale of over 60 years. Adv. Parasitol. 92, 269-306.

Chen, J.Q., Wu, Y., Yang, H., Bergelson, J., Kreitman, M., Tian, D., 2009. Variation in the ratio of nucleotide substitution and indel rates across genomes in mammals and bacteria. Mol. Biol. Evol. 26, 1523-1531.

Chilton, N.B., Bao-Zhen, Q., Bogh, H.O., Nansen, P., 1999. An electrophoretic comparison of *Schistosoma japonicum* (Trematoda) from different provinces in the

People's Republic of China suggests the existence of cryptic species. Parasitology 119 (Pt 4), 375-383.

Cingolani, P., Platts, A., Wang, L.L., Coon, M., Nguyen, T., Wang, L., Land, S.J., Lu, X.Y., Ruden, D.M., 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w(1118); iso-2; iso-3. Fly 6, 80-92.

Clément, J.A., Toulza, E., Gautier, M., Parrinello, H., Roquis, D., Boissier, J., Rognon, A., Moné, H., Mouahid, G., Buard, J., 2013. Private selective sweeps identified from next-generation pool-sequencing reveal convergent pathways under selection in two inbred *Schistosoma mansoni* strains. PLoS Negl. Trop. Dis. 7, e2591.

Colley, D.G., Bustinduy, A.L., Secor, W.E., King, C.H., 2014. Human schistosomiasis. The Lancet 383, 2253-2264.

Collins, C., Xu, J., Tang, S., 2012. Schistosomiasis control and the health system in P.R. China. Infect. Dis. Poverty 1, 8.

Cross, J.H., Zaraspe, G., Lu, S., Chiu, K., Hung, H., 1984. Susceptibility of *Oncomelania hupensis* subspecies to infection with geographic strains of *Schistosoma japonicum*. Southeast Asian J. Trop. Med. Public Health 15, 155-160.

Dal Grande, F., Sharma, R., Meiser, A., Rolshausen, G., Budel, B., Mishra, B., Thines, M., Otte, J., Pfenninger, M., Schmitt, I., 2017. Adaptive differentiation coincides with local bioclimatic conditions along an elevational cline in populations of a lichen-forming fungus. BMC Evol. Biol. 17, 93.

Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., 2011. The variant call format and VCFtools. Bioinformatics 27, 2156-2158.

Degnan, J.H., Rosenberg, N.A., 2006. Discordance of species trees with their most likely gene trees. PLoS Genet. 2, e68.

Ding, H., Lu, D.B., Gao, Y.M., Deng, Y., Li, Y., 2017. Genetic diversity and structure of *Schistosoma japonicum* within two marshland villages of Anhui, China, prior to schistosome transmission control and elimination. Parasitol. Res. 116, 569-576.

Ellegren, H., Galtier, N., 2016. Determinants of genetic diversity. Nat. Rev. Genet. 17, 422-433.

Emerson, R.O., Thomas, J.H., 2009. Adaptive evolution in zinc finger transcription factors. PLoS Genet. 5, e1000325.

Gray, D.J., Williams, G.M., Li, Y., Chen, H., Forsyth, S.J., Li, R.S., Barnett, A.G., Guo, J., Ross, A.G., Feng, Z., McManus, D.P., 2009. A cluster-randomised intervention trial against *Schistosoma japonicum* in the Peoples' Republic of China: bovine and human transmission. PLoS One 4, e5900.

Guo, B., DeFaveri, J., Sotelo, G., Nair, A., Merila, J., 2015. Population genomic evidence for adaptive differentiation in Baltic Sea three-spined sticklebacks. BMC Biol. 13, 19.

Guo, J., Li, Y., Gray, D., Ning, A., Hu, G., Chen, H., Davis, G.M., Sleigh, A.C., Feng, Z., McManus, D.P., Williams, G.M., 2006. A drug-based intervention study on the importance of buffaloes for human *Schistosoma japonicum* infection around Poyang Lake, People's Republic of China. Am. J. Trop. Med. Hyg. 74, 335-341.

He, Y., Guo, Y., Ni, C., Xia, F., Liu, H., Yu, Q., Hu, Y., 1990. Studies on the strain differences of *Schistosoma japonicum* in the mainland of China. I. Compatibility between schistosomes and their snail hosts. Chin. J. Parasitol. Parasit. Dis. 8, 92-95 (in Chinese)..

Ina, Y., 1995. New methods for estimating the numbers of synonymous and nonsynonymous substitutions. J. Mol. Evol. 40, 190-226.

Jarvis, E.D., Mirarab, S., Aberer, A.J., Li, B., Houde, P., Li, C., Ho, S.Y., Faircloth, B.C., Nabholz, B., Howard, J.T., 2014. Whole-genome analyses resolve early branches in the tree of life of modern birds. Science 346, 1320-1331.

Kanehisa, M., Goto, S., Sato, Y., Furumichi, M., Tanabe, M., 2011. KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Res. 40, D109-D114.

Kapun, M., Fabian, D.K., Goudet, J., Flatt, T., 2016. Genomic Evidence for Adaptive Inversion Clines in *Drosophila melanogaster*. Mol. Biol. Evol. 33, 1317-1336.

Koboldt, D.C., Chen, K., Wylie, T., Larson, D.E., McLellan, M.D., Mardis, E.R., Weinstock, G.M., Wilson, R.K., Ding, L., 2009. VarScan: variant detection in massively parallel sequencing of individual and pooled samples. Bioinformatics 25, 2283-2285.

Kutschera, V.E., Bidon, T., Hailer, F., Rodi, J.L., Fain, S.R., Janke, A., 2014. Bears in a forest of gene trees: phylogenetic inference is complicated by incomplete lineage sorting and gene flow. Mol. Biol. Evol. 31, 2004-2017.

Langmead, B., Salzberg, S.L., 2012. Fast gapped-read alignment with Bowtie 2. Nat. Methods 9, 357-359.

Leaché, A.D., Harris, R.B., Rannala, B., Yang, Z., 2013. The influence of gene flow on species tree estimation: a simulation study. Syst. Biol. 63, 17-30.

Leger, E., Webster, J.P., 2017. Hybridizations within the Genus *Schistosoma*: implications for evolution, epidemiology and control. Parasitology 144, 65-80.

Lex, A., Gehlenborg, N., Strobelt, H., Vuillemot, R., Pfister, H., 2014. UpSet: Visualization of Intersecting Sets. IEEE Trans. Vis. Comput. Graph. 20, 1983-1992.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., Genome Project Data Processing, S., 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078-2079.

Li, L., Stoeckert, C.J., Jr., Roos, D.S., 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. 13, 2178-2189.

Li, Y., Yin, M., Wu, Q., McManus, D.P., Blair, D., Li, H., Xu, B., Mo, X., Feng, Z., Hu, W., 2017. Genetic diversity and selection of three nuclear genes in *Schistosoma japonicum* populations. Parasit. Vectors 10, 87.

Long, C., Kubatko, L., 2018. The effect of gene flow on coalescent-based species-tree inference. Syst. Biol. 67, 770-785.

Loytynoja, A., Goldman, N., 2010. webPRANK: a phylogeny-aware multiple sequence aligner with interactive alignment browser. BMC Bioinf. 11, 579.

Lu, D.B., Rudge, J.W., Wang, T.P., Donnelly, C.A., Fang, G.R., Webster, J.P., 2010. Transmission of *Schistosoma japonicum* in Marshland and Hilly Regions of China: Parasite Population Genetic and Sibship Structure. PLoS Negl. Trop. Dis. 4, e781.

McDonald, J.H., Kreitman, M., 1991. Adaptive protein evolution at the Adh locus in *Drosophila*. Nature 351, 652.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20, 1297-1303.

Mirarab, S., Reaz, R., Bayzid, M.S., Zimmermann, T., Swenson, M.S., Warnow, T., 2014. ASTRAL: genome-scale coalescent-based species tree estimation. Bioinformatics 30, i541-i548.

Najafabadi, H.S., Mnaimneh, S., Schmitges, F.W., Garton, M., Lam, K.N., Yang, A., Albu, M., Weirauch, M.T., Radovani, E., Kim, P.M., 2015. C2H2 zinc finger proteins greatly expand the human regulatory lexicon. Nat. Biotechnol. 33, 555-562.

Nolte, V., Pandey, R.V., Kofler, R., Schlötterer, C., 2013. Genome-wide patterns of natural variation reveal strong selective sweeps and ongoing genomic conflict in *Drosophila mauritiana*. Genome Res. 23, 99-110.

Oke, T.T., Moskovitz, J., Williams, D.L., 2009. Characterization of the methionine sulfoxide reductases of *Schistosoma mansoni*. J. Parasitol. 95, 1421-1428.

WHO., 2015. Schistosomiasis: number of people treated worldwide in 2013. Wkly. Epidemiol. Rec. 90, 25-32.

Paradis, E., Claude, J., Strimmer, K., 2004. APE: Analyses of Phylogenetics and Evolution in R language. Bioinformatics 20, 289-290.

Poulin, R., Randhawa, H.S., 2015. Evolution of parasitism along convergent lines: from ecology to genomics. Parasitology 142, S6-S15.

Protasio, A.V., Tsai, I.J., Babbage, A., Nichol, S., Hunt, M., Aslett, M.A., De Silva, N., Velarde, G.S., Anderson, T.J., Clark, R.C., Davidson, C., Dillon, G.P., Holroyd, N.E., LoVerde, P.T., Lloyd, C., McQuillan, J., Oliveira, G., Otto, T.D., Parker-Manuel, S.J., Quail, M.A., Wilson, R.A., Zerlotini, A., Dunne, D.W., Berriman, M., 2012. A systematically improved high quality genome and transcriptome of the human blood fluke *Schistosoma mansoni*. PLoS Negl. Trop. Dis. 6, e1455.

Prugnolle, F., Liu, H., de Meeus, T., Balloux, F., 2005. Population genetics of complex life-cycle parasites: an illustration with trematodes. Int. J. Parasitol. 35, 255-263.

Rogers, J., Gibbs, R.A., 2014. Comparative primate genomics: emerging patterns of genome content and dynamics. Nat. Rev. Genet. 15, 347.

Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D.L., Darling, A., Hohna, S., Larget, B., Liu, L., Suchard, M.A., Huelsenbeck, J.P., 2012. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Syst. Biol. 61, 539-542.

Ross, A.G., Sleigh, A.C., Li, Y., Davis, G.M., Williams, G.M., Jiang, Z., Feng, Z., McManus, D.P., 2001. Schistosomiasis in the People's Republic of China: prospects and challenges for the 21st century. Clin. Microbiol. Rev. 14, 270-295.

Rudge, J.W., Lu, D.B., Fang, G.R., Wang, T.P., Basanez, M.G., Webster, J.P., 2009. Parasite genetic differentiation by habitat type and host species: molecular epidemiology of *Schistosoma japonicum* in hilly and marshland areas of Anhui Province, China. Mol. Ecol. 18, 2134-2147.

Rudge, J.W., Webster, J.P., Lu, D.B., Wang, T.P., Fang, G.R., Basanez, M.G., 2013. Identifying host species driving transmission of schistosomiasis japonica, a multihost parasite system, in China. Proc. Natl. Acad. Sci. U. S. A. 110, 11457-11462.

Saeb, A.T., 2015. Heat shock protein Hsp70 multigene family as a new genetic target for the differentiation and identification of entomopathogenic nematodes (Rhabditida: Heterorhabditidae). Advances in Life Sciences and Health 2, 16-30.

Sayyari, E., Mirarab, S., 2016. Fast coalescent-based computation of local branch support from quartet frequencies. Mol. Biol. Evol. 33, 1654-1668.

Schlötterer, C., Tobler, R., Kofler, R., Nolte, V., 2014. Sequencing pools of individuals - mining genome-wide polymorphism data without big funding. Nat. Rev. Genet. 15, 749-763.

Shi, L., Li, W., Wu, F., Zhang, J.F., Yang, K., Zhou, X.N., 2016. Epidemiological Features and Control Progress of Schistosomiasis in Waterway-Network Region in The People's Republic of China. Adv. Parasitol. 92, 97-116.

Shipham, A., Schmidt, D.J., Joseph, L., Hughes, J.M., 2015. Phylogenetic analysis of the Australian rosella parrots (Platycercus) reveals discordance among molecules and plumage. Mol. Phylogenet. Evol. 91, 150-159.

Shrivastava, J., Qian, B.Z., McVean, G., Webster, J.P., 2005. An insight into the genetic variation of *Schistosoma japonicum* in mainland China using DNA microsatellite markers. Mol. Ecol. 14, 839-849.

Soares Magalhaes, R.J., Salamat, M.S., Leonardo, L., Gray, D.J., Carabin, H., Halton, K., McManus, D.P., Williams, G.M., Rivera, P., Saniel, O., Hernandez, L., Yakob, L., McGarvey, S., Clements, A., 2014. Geographical distribution of human *Schistosoma japonicum* infection in The Philippines: tools to support disease control and further elimination. Int. J. Parasitol. 44, 977-984.

Stamatakis, A., 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30, 1312-1313.

Steinauer, M.L., Blouin, M.S., Criscione, C.D., 2010. Applying evolutionary genetics to schistosome epidemiology. Infect. Genet. Evol. 10, 433-443.

Stoletzki, N., Eyre-Walker, A., 2011. Estimation of the neutrality index. Mol. Biol. Evol. 28, 63-70.

Sukumaran, J., Holder, M.T., 2010. DendroPy: a Python library for phylogenetic computing. Bioinformatics 26, 1569-1571.

Sun, L.P., Wang, W., Hong, Q.B., Li, S.Z., Liang, Y.S., Yang, H.T., Zhou, X.N., 2017. Approaches being used in the national schistosomiasis elimination programme in China: a review. Infect. Dis. Poverty 6, 55.

Team, R.C., 2013. R: A language and environment for statistical computing.

Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J., Pachter, L., 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat. Biotechnol. 28, 511-515.

Wang, L.D., Chen, H.G., Guo, J.G., Zeng, X.J., Hong, X.L., Xiong, J.J., Wu, X.H., Wang, X.H., Wang, L.Y., Xia, G., Hao, Y., Chin, D.P., Zhou, X.N., 2009. A strategy to control transmission of *Schistosoma japonicum* in China. N. Engl. J. Med. 360, 121-128.

Wang, X., Wang, W., Wang, P., 2017. Long-term effectiveness of the integrated schistosomiasis control strategy with emphasis on infectious source control in China: a 10-year evaluation from 2005 to 2014. Parasitol. Res. 116, 521-528.

Webster, B.L., Diaw, O.T., Seye, M.M., Webster, J.P., Rollinson, D., 2013. Introgressive hybridization of *Schistosoma haematobium* group species in Senegal: species barrier break down between ruminant and human schistosomes. PLoS Negl. Trop. Dis. 7, e2110.

Wickham, H., Chang, W., 2008. ggplot2: An implementation of the Grammar of Graphics. URL: http://CRAN.R-project.org/package=ggplot2.

Xi, Z., Liu, L., Davis, C.C., 2015. Genes with minimal phylogenetic information are problematic for coalescent analyses when gene tree estimation is biased. Mol. Phylogenet. Evol. 92, 63-71.

Xie, C., Mao, X., Huang, J., Ding, Y., Wu, J., Dong, S., Kong, L., Gao, G., Li, C.-Y., Wei, L., 2011. KOBAS 2.0: a web server for annotation and identification of enriched pathways and diseases. Nucleic Acids Res. 39, W316-W322.

Xu, J., Steinman, P., Maybe, D., Zhou, X.N., Lv, S., Li, S.Z., Peeling, R., 2016. Evolution of the National Schistosomiasis Control Programmes in The People's Republic of China. Adv. Parasitol. 92, 1-38.

Yin, M., Li, H., Blair, D., Xu, B., Feng, Z., Hu, W., 2016a. Temporal genetic diversity of *Schistosoma japonicum* in two endemic sites in China revealed by microsatellite markers. Parasit. Vectors 9, 36.

Yin, M., Li, H., McManus, D.P., Blair, D., Su, J., Yang, Z., Xu, B., Feng, Z., Hu, W., 2015a. Geographical genetic structure of *Schistosoma japonicum* revealed by analysis of mitochondrial DNA and microsatellite markers. Parasit. Vectors 8, 150.

Yin, M., Liu, X., Xu, B., Huang, J., Zheng, Q., Yang, Z., Feng, Z., Han, Z.G., Hu, W., 2016b. Genetic variation between *Schistosoma japonicum* lineages from lake and mountainous regions in China revealed by resequencing whole genomes. Acta Trop. 161, 79-85.

Yin, M., Zheng, H.X., Su, J., Feng, Z., McManus, D.P., Zhou, X.N., Jin, L., Hu, W., 2015b. Co-dispersal of the blood fluke *Schistosoma japonicum* and Homo sapiens in the Neolithic Age. Sci. Rep. 5, 18058.

Young, N.D., Chan, K.G., Korhonen, P.K., Min Chong, T., Ee, R., Mohandas, N., Koehler, A.V., Lim, Y.L., Hofmann, A., Jex, A.R., Qian, B., Chilton, N.B., Gobert, G.N., McManus, D.P., Tan, P., Webster, B.L., Rollinson, D., Gasser, R.B., 2015. Exploring molecular variation in *Schistosoma japonicum* in China. Sci. Rep. 5, 17345.

Zhao, G.H., Mo, X.H., Zou, F.C., Li, J., Weng, Y.B., Lin, R.Q., Xia, C.M., Zhu, X.Q., 2009. Genetic variability among *Schistosoma japonicum* isolates from different endemic regions in China revealed by sequences of three mitochondrial DNA genes. Vet. Parasitol. 162, 67-74.

Zhao, Q.P., Jiang, M.S., Dong, H.F., Nie, P., 2012. Diversification of *Schistosoma japonicum* in Mainland China revealed by mitochondrial DNA. PLoS Negl. Trop. Dis. 6, e1503.

Zhao, Q.P., Jiang, M.S., Littlewood, D.T., Nie, P., 2010. Distinct genetic diversity of *Oncomelania hupensis*, intermediate host of *Schistosoma japonicum* in mainland China as revealed by ITS sequences. PLoS Negl. Trop. Dis. 4, e611.

Zheng, X., Levine, D., Shen, J., Gogarten, S.M., Laurie, C., Weir, B.S., 2012. A high-performance computing toolset for relatedness and principal component analysis of SNP data. Bioinformatics 28, 3326-3328.

Zhou, X.N., Bergquist, R., Leonardo, L., Yang, G.J., Yang, K., Sudomo, M., Olveda, R., 2010. Schistosomiasis japonica: control and research needs. Adv. Parasitol. 72, 145-178.

Zhou, Y.B., Liang, S., Chen, Y., Jiang, Q.W., 2016. The Three Gorges Dam: Does it accelerate or delay the progress towards eliminating transmission of schistosomiasis in China? Infect. Dis. Poverty 5, 63.

Zhou, Y., Zheng, H., Chen, X., Zhang, L., Wang, K., Guo, J., Huang, Z., Zhang, B., Huang, W., Jin, K., 2009. The *Schistosoma japonicum* genome reveals features of host-parasite interplay. Nature 460, 345.

Table 4-1 Summary of the final consensus SNPs in each population after down-sampling and filtration.

Isolate code	Sample location (County, Province)	No. of SNP called	Average allelic frequency (%)*	Mapping depth*	Genotype call quality*	Intergenic SNPs	Coding SNPs	Intronic SNPs	Sequencing depth within coding domains*
N1	Zongyang, Anhui	1,936,137	83.45 ± 18.89	12.31 ± 6.77	34.34 ± 6.14	1,379,308 (71.24%)	36,181 (1.87%)	520,648 (26.89%)	14.23 ± 6.62
N8	Honghu, Hubei	1,863,304	82.11 ± 17.47	12.33 ± 7.37	34.32 ± 6.03	1,327,769 (71.26%)	34,417 (1.85%)	501,118 (26.89%)	14.46 ± 14.61
N9	Xinjian, Jiangxi	1,364,921	74.47 ± 15.38	11.94 ± 6.69	34.50 ± 5.35	980,372 (71.83%)	24,022 (1.76%)	360,527 (26.41%)	13.95 ± 12.07
N10	Xichang, Sichuan	1,844,939	81.15 ± 16.65	12.43 ± 7.48	34.16 ± 5.90	1,317,972 (71.44%)	34,227 (1.86%)	492,740 (26.71%)	14.64 ± 7.86
N21	Jianglin, Hubei	1,638,431	78.30 ± 16.61	12.13 ± 6.52	34.42 ± 5.64	1,174,111 (71.66%)	29,671 (1.81%)	434,649 (26.53%)	14.32 ± 8.93
N22	Jinzhou, Hubei	1,447,217	74.85 ± 15.71	12.23 ± 7.01	34.60 ± 5.31	1,039,775 (71.85%)	25,669 (1.77%)	381,773 (26.38%)	14.46 ± 7.05
N24	Yangxin, Hubei	1,267,537	72.28 ± 14.60	11.95 ± 6.55	34.69 ± 5.10	911,779 (71.93%)	22,253 (1.76%)	333,505 (26.31%)	14.01 ± 10.73
N31	Dongtinghu, Hunan	1,329,657	73.93 ± 15.44	11.75 ± 5.99	34.58 ± 5.27	956,050 (71.90%)	23,241 (1.75%)	350,366 (26.35%)	13.97 ± 10.77
N37	Jianglin, Hubei	1,633,765	77.69 ± 16.82	12.13 ± 6.74	34.44 ± 5.55	1,170,244 (71.63%)	29,016 (1.78%)	434,505 (26.60%)	14.27 ± 8.68
SJ1	Jiashan, Zhejiang	1,751,097	77.58 ± 16.15	11.80 ± 5.75	34.41 ± 5.31	1,252,396 (71.52%)	30,216 (1.73%)	468,485 (26.75%)	13.67 ± 8.49
SJ2	Guichi, Anhui	1,443,885	73.50 ± 15.10	11.84 ± 5.47	34.61 ± 4.99	1,037,180 (71.83%)	24,573 (1.70%)	382,132 (26.47%)	13.76 ± 7.72
SJ3	Yongxia, Jiangxi	1,368,776	72.76 ± 14.74	11.70 ± 5.48	34.62 ± 4.95	986,931 (72.10%)	22.993 (1.68%)	358,852 (26.22%)	13.51 ± 7.41
SJ4	Wuhan, Hubei	1,650,198	76.29 ± 15.92	11.83 ± 5.55	34.47 ± 5.22	1,184,008 (71.75%)	28,025 (1.70%)	438,165 (26.55%)	13.65 ± 6.60
SJ5	Yueyang, Hunan	1,606,912	75.54 ± 15.74	11.75 ± 5.53	34.47 ± 5.16	1,153,590 (71.79%)	27,065 (1.68%)	426,257 (26.53%)	13.56 ± 6.81
SJ6	Tianquan, Sichuan	2,060,895	82.51 ± 16.42	12.01 ± 6.22	34.17 ± 5.68	1,472,301 (71.44%)	36,684 (1.78%)	551,910 (26.78%)	13.91 ± 6.38
SJ7	Dali, Yunnan	2,092,259	81.69 ± 16.92	12.37 ± 8.06	34.12 ± 5.60	1,492,913 (71.35%)	37,664 (1.80%)	561,682 (26.85%)	14.21 ± 6.76

*mean ± standard deviation

Table 4-2 Summary of SNPs in coding and intronic regions of 6,978 single-copy gene groups in each isolate.

Isolate code	**Total SNPs**	**Retained intronic SNPs**	**Intronic SNPs per 1 kb***	**Retained coding SNPs**	**Coding SNPs per 1 kb***	**Non-Synonymous SNPs**	**Synonymous SNPs**
N1	137,205	120,441 (87.78%)	4.15	16,764 (12.22%)	2.10	7,840	8,924
N8	131,985	116,254 (88.08%)	4.00	15,731 (11.92%)	1.97	7,278	8,453
N9	91,411	80,789 (88.38%)	2.78	10,622 (11.62%)	1.33	4,775	5,847
N10	129,954	114,212 (87.89%)	3.93	15,742 (12.11%)	1.97	7,399	8,343
N21	112,007	98,572 (88.01%)	3.40	13,435 (11.99%)	1.68	6,178	7,257
N22	97,498	85,973 (88.18%)	2.96	11,525 (11.82%)	1.44	5,383	6,142
N24	84,304	74,358 (88.20%)	2.56	9,946 (11.80%)	1.25	4,579	5,367
N31	88,449	78,216 (88.43%)	2.69	10,233 (11.57%)	1.28	4,776	5,457
N37	110,961	97,837 (88.17%)	3.37	13,124 (11.83%)	1.64	6,069	7,055
SJ1	118,386	104,991 (88.69%)	3.62	13,395 (11.31%)	1.68	6,146	7,249
SJ2	95,234	84,354 (88.58%)	2.91	10,880 (11.42%)	1.36	5,029	5,851
SJ3	89,553	79,365 (88.62%)	2.73	10,188 (11.38%)	1.28	4,695	5,493
SJ4	110,153	97,753 (88.74%)	3.37	12,400 (11.26%)	1.55	5,665	6,735
SJ5	107,388	95,354 (88.79%)	3.28	12,034 (11.21%)	1.51	5,534	6,500
SJ6	142,442	125,816 (88.33%)	4.33	16,626 (11.67%)	2.08	7,711	8,915
SJ7	145,760	128,580 (88.21%)	4.43	17,180 (11.79%)	2.15	8,143	9,037

Table 4-3 Summary of phylogenetic analyses of the *Schistosoma japonicum* coding and intronic region data sets.

Features	Coding regions	Introns
Concatenated sequence approach		
Bootstrap support (BS) of the 13 branches (branches >90%; >50%; Mean %)	4;5;50	9;9;79
ML[a] Internal branch length (BS >50%)	0.00062	0.00083
ML Bootstrap Likelihood[b]	-11,779,789 ± 5088.11	-45,666,864 ± 5290.55
BI[c] likelihood (run1; run2)	-	-44,291,865.54; -44,292,093.00
Summary tree approach		
Normalised quartet score for coalesced summary tree	0.405	0.422

[a] Maximum likelihood
[b] Average ± standard deviation
[c] Bayesian inference

Table 4-4 Characteristics of nucleotide diversity, polymorphism and divergence among schistosomes and among/between *Schistosoma japonicum* isolates in the central and lower reaches of the Yangtze River (CL) and in the mountainous regions of Southwest China (SW).

Type of analysis	**Total**	**CL**	**SW**	***S. japonicum* and *S. mansoni***
d_N[a]				0.1493 ± 0.0913
d_S[a]				0.7368 ± 0.2511
d_N/d_S[a]				0.2186 ± 0.1387
π[a]	0.0021 ± 0.0020	0.0020 ± 0.0019	0.0028 ± 0.0026	
π_N[a]	0.0012 ± 0.0015	0.0011 ± 0.0015	0.0015 ± 0.0020	
π_S[a]	0.0057 ± 0.0054	0.0054 ± 0.0052	0.0078 ± 0.0074	
π_N/π_S*[a]	0.2995 ± 0.4372	0.2950 ± 0.4436	0.2432 ± 0.3173	
π_N/π_S vs. π[b]	0.28; *p*-value < 2.2e-16	0.27; < 2.2e-16	0.37; *p*-value <2.2e-16	
d_N/d_S vs. π_N/π_S[b]	0.45; *p*-value < 2.2e-16	0.45; < 2.2e-16	0.43; *p*-value<2.2e-16	

[a] Average (± standard deviation) nucleotide differences per site

[b] Spearman correlation coefficient (r_s); *p*-value

Table 4-5 Genes with most fixed SNPs in *Schistosoma japonicum* genes between central and lower reaches of the Yangtze River (CL) and in the mountainous regions of Southwest China (SW).

Regions	Genes	Polymorphic sites	Annotation
SW			
	Sjp_0059310	Pn:7/29; Ps:3/14	Methionine sulfoxide reductase, MK adaptive genes
	Sjp_0071550	Pn:4/14; Ps:2/14	5-hydroxytryptamine receptor 7 G protein-coupled receptors
	Sjp_0067610	Pn:4/17; Ps:4/18	Calcium-activated potassium channel
CL			
	Sjp_0026090	Pn:4/9; Ps:2/17	glucose-6-phosphate 1-dehydrogenase

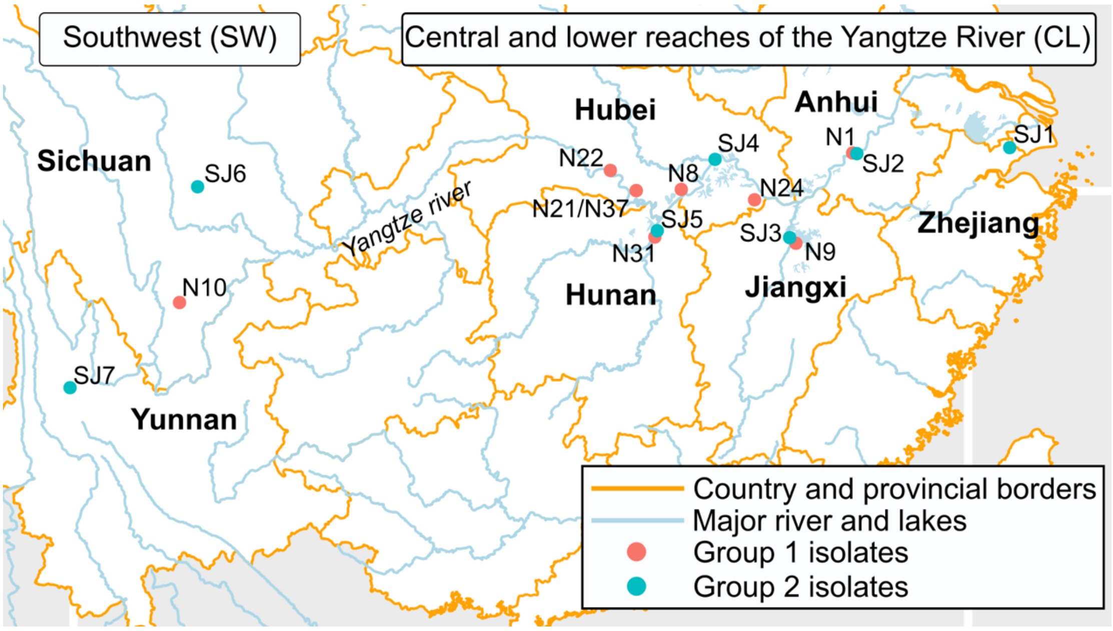


Fig. 4-1 Map of locations of the 16 isolates. This map includes 9 geographical populations (isolate with N prefix) in Group 1 (red circles) and 7 geographical populations (isolates with SJ prefix) in Group 2 (blue circles). The Yangtze River and other major rivers and lakes are shown in light blue colour. Boundaries of countries and provinces are shown as orange lines. The map was generated and modified based on the resources of www.naturalearthdata.com.

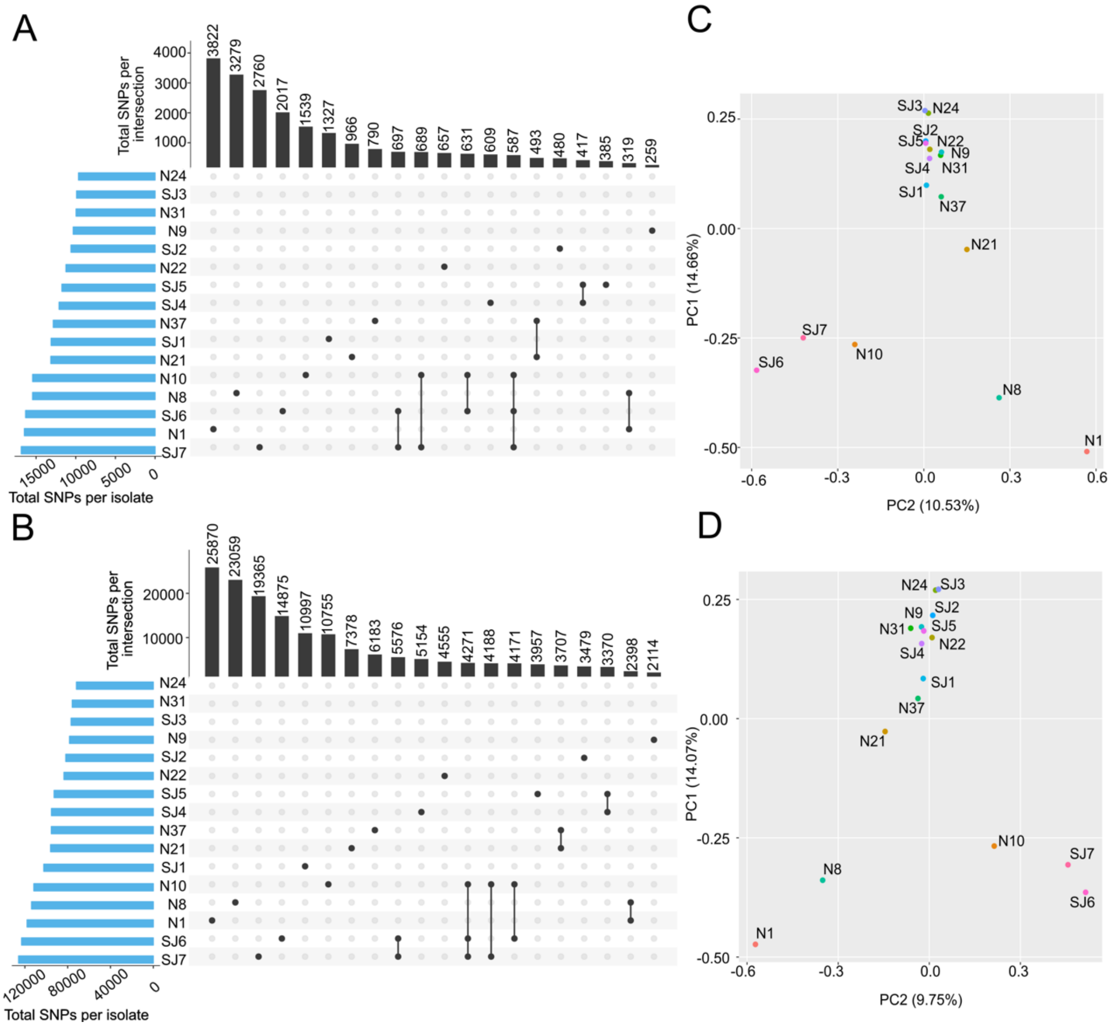


Fig. 4-2 Consensus SNPs distribution and principal component analysis of *Schistosoma japonicum* from 16 locations in China. UpSet plot showing the distribution of consensus coding SNPs (A) and intronic SNPs (B) of the 16 isolates. The black vertical bars and the numbers on top represent frequency of unique SNPs in each intersection of isolate group(s). Black dot(s) below each vertical bar showing the isolates that present the SNPs. The lined dots show that the intersection is the common unique SNPs shared between two or more isolates. The number of consensus SNPs in each isolate is shown by horizontal blue bars on the bottom left of the figure. Principal component analysis (PCA) with consensus SNPs located in coding regions (C), introns (D) are shown as 2-dimension plots. The two axes represent two most crucial components. The percentage of variance for each component is shown within brackets. Each isolate is shown as circle with corresponding colour.

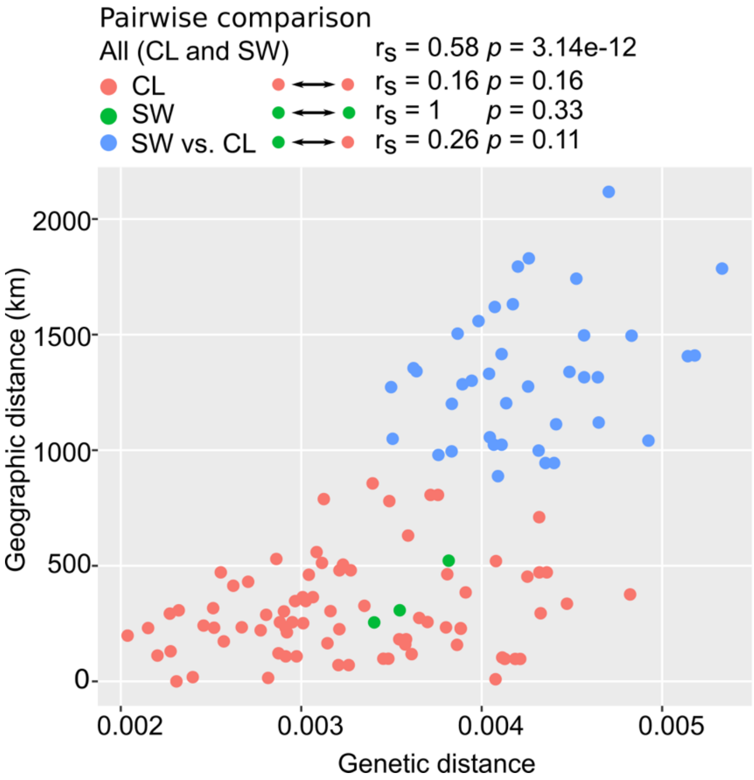


Fig. 4-3 Correlation between genetic distance and geographical location. Red dots represent isolate pairs from central and lower reaches of Yangtze River (CL). Green dots represent isolate pairs from Southwest (SW) region. Blue dots represent pairs comprised of one SW isolate and one CL isolate (SW-vs-CL). Spearman's rank correlation coefficient (r_s) and the corresponding *p*-value (p) for each group and all the pairs are listed at the right.

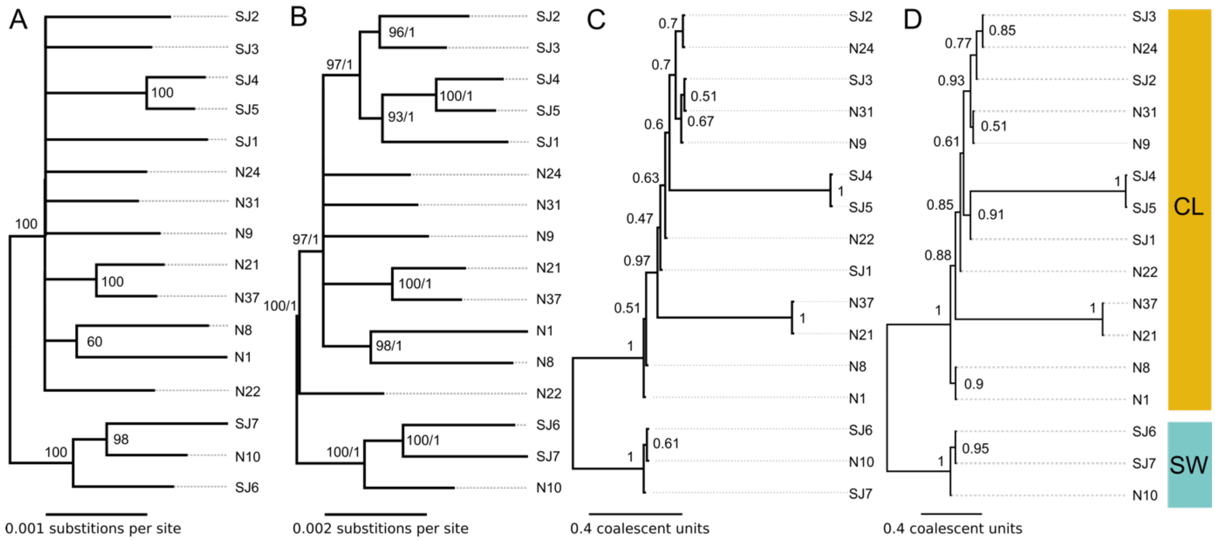


Fig. 4-4 Phylogenetic relationships of *Schistosoma japonicum* from 16 locations in China. **(**A) Majority rule maximum likelihood (ML) tree constructed using concatenated nucleotide sequences of coding regions single copy gene groups (SCGGs). (B) Majority rule ML tree constructed using concatenated nucleotide sequences of intronic regions of SCGGs. Nodal bootstrap or posterior probability values on ML trees are indicated in the following order: ML/BI. (C) Coalesced summary tree constructed using nucleotide sequences of coding regions of SCGGs. (D) Coalesced summary (CS) tree constructed using nucleotide sequences of intronic regions of SCGGs. Nodal posterior probability values are shown on each CS tree.

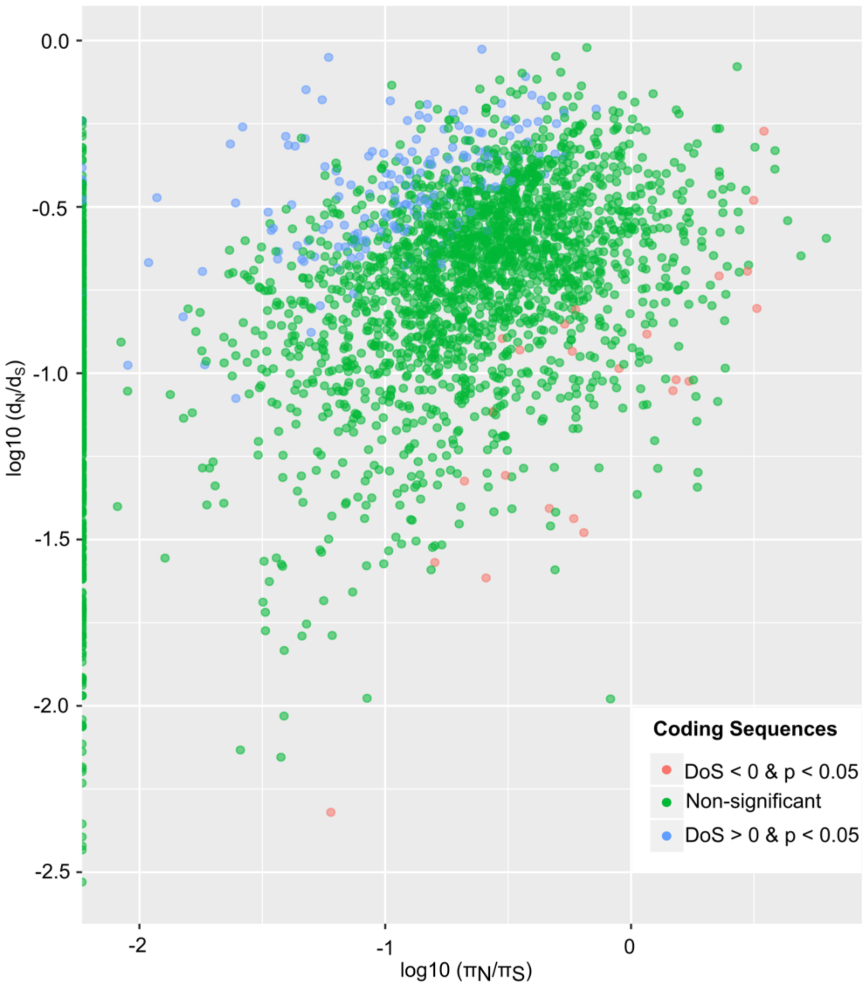


Fig. 4-5 Estimation of neutrality index of coding sequence. The two axes showing logarithmic values (base: 10) of the ratio of interspecific nonsynonymous (d_N) and synonymous (d_S) divergence per site and the ratio of intraspecific nonsynonymous (π_N) and synonymous (π_S) polymorphisms per site for each single copy gene shared between *Schistosoma japonicum* and *S. mansoni*. Genes with a significant McDonald-Kreitman result (Fisher's exact test, $p < 0.05$) are shown in red or blue. Within these genes, the genes with a positive direction of selection force (DoS) value are indicated as blue dots. And the genes with a negative DoS value are indicated as red dots. Genes with a non-significant Fisher's exact test value are shown in green dots.

Chapter 5 - General discussion

The aims of this thesis were to: (1) assess sequence variation in the mitochondrial (mt) genomes between multiple *Clonorchis sinensis* isolates; (2) sequence, assemble and annotate a nuclear genome of a Korean isolate and compare it with that of the published genome of a Chinese isolate; and (3) explore the population structure of *Schistosoma japonicum* in China and detect genes under positive selection in geographically distinct locations. All of these aims were achieved.

The purpose of this chapter is to summarise the main research achievements, to discuss aspects relating to the methods developed in this thesis, the genetics of *C. sinensis*, *S. japonicum* and parasitic helminths, more generally, and to identify some key topics for the future investigations.

5.1. Mitochondrial genomes of *C. sinensis*

In Chapter 2, a deep sequencing-bioinformatics approach was used to sequence and define a reference mitochondrial (mt) genome for an isolate of *C. sinensis* from Korea (*Cs*-k2). Compared with traditional polymerase chain reaction (PCR) and/or cloning methods (e.g., Le et al., 2001; Notsu et al., 2002; Gibson et al., 2007), the deep sequencing approach allows the assembly of high-quality reference mt genomes in a time- and cost-effective manner (Goldman and Domschke, 2014). By using this approach, the specific identity of *Cs*-k2 was unequivocally confirmed based on mt genomic sequence and size, genome composition and gene order. In addition, the whole mt genome sequences from isolates from Korea (*Cs*-k2 and *Cs*-k1), China (*Cs*-c1) and Russia (*Cs*-r1) (Shekhovtsov et al., 2010; Cai et al., 2012) were aligned and compared, and the amino acid sequence data were used to infer phylogenetic relationships among the four isolates and other representative trematodes. The relatively small number (24) of variable amino acids led to an unresolved relationship of the four isolates, whereas the pairwise nucleotide diversity in the whole mt genome overcame this problem, which might be partly contributed by the variable sites in the non-coding regions and, hence, warrants future consideration and investigation.

In mammalian mtDNA, a strong selective force is responsible for more synonymous than non-synonymous mutations (Stewart et al., 2008). In mice, nonsynonymous mtDNA mutations in protein-encoding genes are rapidly eliminated from mtDNA during the reproductive stage, suggesting a negative selection against amino acid changes (Stewart et al., 2008). Currently, genetic loci under selective forces have been reported in a number of mtDNA studies (e.g., Nachman et al., 1996; Hung and Zink, 2014), which would have impacts on demographic studies using mtDNA markers. For example, selective forces could result in patterns similar to demographic changes (Bazin et al., 2006; Ramirez-Soriano et al., 2008) and might increase the difficulty in interpreting key population estimators, such as effective population size (N_e) and Tajima's D (Tajima, 1989).

In addition, we observed gene-specific mutations in the mt genomes that were not associated with the geographical origins of the samples (Chapter 2). For example, for nine of the 12 mt genes, the lowest level of nucleotide variation was between the two Korean isolates. However, for two (*cox*1 and *cox*2) of the other three genes, the lowest level of nucleotide variation was between *Cs*-k1 and *Cs*-c1. A similar observation was made in a previous study that reports a recombination event of mtDNA within an Australian lizard (Ujvari et al., 2007). Traditionally, based on the assumptions of homoplasy and strict maternal inheritance, mtDNA has long been believed not to recombine (White et al., 2008). However, evidence to the contrary has been reported. For example, mitochondrial heteroplasmy and biparental inheritance were both observed in animals (Jannotti-Passos et al., 2001; Sutovsky et al., 2004; Breton et al., 2007). These findings suggest that mitochondria originating from both parents are able to exist within the same cell of offspring. Furthermore, the hypothesis of recombination in mt genomes is also supported by the presence of mitochondrial fusion mechanisms (Yaffe, 1999) and enzymes that function in DNA exchange and repairing in mitochondria (Lakshmipathy and Campbell, 1999; Santel and Fuller, 2001). In addition, from an evolutionary perspective, recombination could eliminate some deleterious mutations from a population (Tsaousis et al., 2005). These studies might be able to explain, at least partly, the increasing number of mtDNA recombination events reported in plants, fungi and animals (Piganeau et al., 2004; Tsaousis et al., 2005; Gibson et al., 2007; Ujvari et al., 2007; Wang et al., 2017c), although these recombination signals might also be introduced due to bioinformatic or experimental artefacts (Piganeau et al., 2004). Given the evidence of biparental inheritance of mtDNA in *Schistosoma mansoni* (Jannotti-Passos et al., 2001) and the observed pattern genetic variation in *C. sinensis*, the possibility that recombination occurs occasionally in trematodes should not be excluded. Clearly, caution is required when interpreting intraspecific variation in mtDNA.

Chapter 2 identified several potential issues when using mtDNA markers in population genetic studies. Such issues have been reviewed previously (see Sunnucks, 2000; Zhang and Hewitt, 2003; Vilas et al., 2005; Galtier et al., 2009). For example, smaller N_e of mtDNA (one half for hermaphrodites) increases the impact of genetic drift effect, leading to a rapid allele extinction rate and oversimplification of the real population structure (Zhang and Hewitt, 2003). In addition, due to a relatively high gene density of the mt genome, allele frequencies are altered by natural selection on the nearby locus that is in linkage disequilibrium (Awadalla et al., 1999), and can create conflicts in lineage relationships (Galtier et al., 2009). These factors might contribute to the reported discordance between mt and nuclear genomic data sets (Toews and Brelsford, 2012), which might be the signals of hybridisation and/or adaptive selection on mtDNA (Detwiler and Criscione, 2010; Toews and Brelsford, 2012). Hence, a number of studies have employed both mtDNA and nuclear DNA markers to ensure an accurate estimation of genetic structure (reviewed by Toews and Brelsford, 2012). In contrast to mtDNA, the nuclear genome provides a vast amount of relatively neutral (e.g., intronic areas) and non-neutral genetic loci (e.g., coding domains that are under natural selection) (Kirk and Freeland, 2011). The use of a large number of genetic loci can increase the robustness of evolutionary inference (Allendorf et al., 2010), and nuclear genomic sequences can assist in this endeavour.

5.2. Nuclear genome of *C. sinensis*

In Chapter 3, the nuclear genome of a *C. sinensis* isolate from Korea was sequenced, assembled and characterised. In addition, a refined bioinformatic workflow was designed and applied to compare this assembly with a previously published draft genome of another isolate from China. Using this workflow, nucleotide differences were identified and evaluated by aligning nucleotide sequences within the syntenic blocks with conserved gene orders (Chapter 3). Given the reported karyotypic variation among *C. sinensis* isolates (Park et al., 2000; Park and Yong, 2001; Zadesenets et al., 2012), the workflow developed here could also be used to study genetic variation within *C. sinensis* from Russian Far East in the future.

Currently, evaluating genetic variation between nuclear genomes of geographically distinct isolates can be a major challenge. Most studies have detected genetic variation by mapping reads to a reference genome (Jirimutu et al., 2012; Clement et al., 2013; Huang et al., 2014a; Crellen et al., 2016; Hane et al., 2017). However, the quality of "variation calling" is affected by ambiguous read alignments in repetitive regions (Treangen and Salzberg, 2011). In addition, genome-wide comparisons that map to a single reference genome may miss strain-specific sequences and structural variations (SVs), particularly for geographically distinct isolates. Given the observed karyotypic variation observed among *C. sinensis* isolates (Gao et al., 1987; Gao et al., 1993; Park et al., 2000; Park and Yong, 2001; Zadesenets et al., 2012), it is likely that one reference genome does not represent all intraspecific populations. Besides a read-mapping approach, previous studies have aligned and compared assemblies to a reference genome to identify SVs and changes in the sizes of gene families (Xue et al., 2012; Hester et al., 2013; Menard et al., 2013; Sun et al., 2015; Ansari et al., 2016; Zhou et al., 2017) - based on the assumption that variation can be detected directly from the comparison of assemblies and predicted gene sets. This approach is more likely to detect evolutionary changes in genomic content, but inferences might be affected by sequencing artefacts and/or mis-assemblies, particularly when draft genomes are compared (Denton et al., 2014).

Draft genomes often contain misassembled regions. For instance, based on a linkage map, over 43% of the assembly of *S. mansoni* was found to be incorrect (Criscione et al., 2009). Similar findings have also been reported for the genome of the mollusc *Crassostrea gigas*, for which 38.5% of the genomic scaffolds contain mis-assemblies (Hedgecock et al., 2015). Factors contributing to these mis-assemblies can relate to aspects including contamination during DNA extraction (Hashimoto et al., 2016), genomic structure, and systematic errors of sequencing platforms and assemblers, resulting in a false estimation of the number of predicted genes (Denton et al., 2014).

DNA contamination can confound evolutionary conclusions. For instance, 223 bacterial homologs in the initial human genome assembly were reported as the result of horizontal genetic transfer (Lander et al., 2001). Subsequent research, however, revealed that most of these homologs were caused by bacterial contaminants (Salzberg et al., 2001). In addition to contamination, the presence of heterozygosity and repetitive elements also imposes challenges in achieving a complete and accurate genome (Vinson et al., 2005). Heterozygous regions are often assembled into separate loci and, hence, can increase the size of draft assemblies and the number of predicted genes (Holt et al., 2002; Jones et al., 2004; Hahn et al., 2014).

Although the level of heterozygosity in animals might be reduced through inbreeding, generating inbred lines can be expensive and time consuming (Kajitani et al., 2014). Another approach is to allow more mismatches in the alignment of sequences to be merged into the same locus during assembly (Hahn et al., 2014). However, such an approach would also increase the possibility of tandem repetitive elements "collapsing" into one sequence. Presence of repetitive elements is a common reason for unsuccessful assemblies. For example, in human genome projects, repetitive elements can lead to a loss of ~16% of the reference genome in draft assemblies (Alkan et al., 2011). In contrast to the repeat-rich human genome, the chicken genome was predicted to have a lower repeat content (10%), and 5% of the available reference genome was missing from the draft assemblies (Ye et al., 2011). Similar to mammals, *C. sinensis* is a diploid organism and has a relatively high repeat content (~ 33%), which creates challenges in addressing these issues (mis-assemblies caused by the presence of heterozygosity and repetitive elements) using short-read NGS data alone.

Another feature of draft assemblies is fragmentation, which increases the number of predicted genes within a genome. In humans, the number of predicted genes decreased from 30,000 to 40,000 (Lander et al., 2001) to 20,687 (Harrow et al., 2012), with an increasing level of contiguity of different assembly versions and a better understanding of the non-coding and protein coding genetic elements. In addition, 30% of the human reference genes were present in more than one scaffold of draft genome assemblies (Alkan et al., 2011). Portions of several gene sequences were identified in up to 200 scaffolds (Alkan et al., 2011), which further emphasises the importance of correct scaffolding and assembly.

Currently, the detection of mis-assemblies in draft genomes remains a challenging task (Muggli et al., 2015). One approach is to compare draft genomes against an "error-free" reference genome (e.g., QUAST Gurevich et al., 2013). However, for species without complete/near-complete genomes, a high-quality reference genome is not achievable. An alternative is to detect mis-assembled regions by mapping read data to the draft genomes. However, such tools either lack sensitivity or accuracy due to ambiguous mapping of short read data. For example, although REAPR (Hunt et al.,

2013) has a low false-positive rate (FDR) in detecting misassembled regions, this tool can only identify a small proportion of such regions (Muggli et al., 2015). Another program, misSEQuel, is able to identify most misassembled regions, whereas it has a high FDR (Muggli et al., 2015) when using only NGS data. Hence, at the current stage, it is not possible to confidently quantify misassembled regions in draft genomes without experimental (e.g., physical or linkage map) or long sequence-read data.

Therefore, for *C. sinensis*, the risk of incorrectly recording a genetic variant (i.e. a "false-positive") was avoided by exclusively comparing assembled genomic regions within syntenic blocks with conserved gene orders (Chapter 3). This method greatly simplifies the complex syntenic relationship resulting from the presence of repetitive elements (Bourque et al., 2004), which is evidenced by the larger average size of syntenic blocks compared with aligned regions (of both masked and unmasked genomes) based on nucleotide similarities (Chapter 3). Moreover, because the single-copy genes within the syntenic blocks were predicted independently with a consistent order and orientation on two different assemblies, the inferred gene models should be more reliable than those outside of the syntenic blocks. However, using this approach, the size of syntenic blocks is still limited by fragmentation. Although several measures were implemented to increase the contiguity and completeness of the assembly, such as employing DNA from pooled worms, linking scaffolds with large insert size read pairs and employing post-assembly program HaploMerger2 (Huang et al., 2012a) to remove redundant sequences and improve assembly quality, the sequence contiguity does still not represent a chromosome-level assembly. Substantial gaps remained in the assembly. One reason for this could be unsuccessful gap filling, which might be related to false scaffolding or the presence of repetitive elements (Hunt et al., 2014). Although the contiguity of assembly could be further improved through the construction of a linkage map (Fierst, 2015) and/or a physical map (Mayer et al., 2012), these approaches are time consuming and costly to perform. Another way might be to employ long sequence-read data (cf. Korhonen et al., 2016, 2019). With the decreasing cost of third-generation sequencing, it is expected that assembly contiguity will increase significantly in the coming years. In synteny analysis, the improved assembly should effectively extend the average size of the syntenic blocks and reduce the number of predicted breakpoints relating to mis-assemblies and fragmentation.

Another consideration is that the accuracy of the method based on single-copy genes is highly dependent on the quality of a gene model. For instance, the orthologous genes that are used to define syntenic blocks need to be correctly predicted in both genomes being compared; otherwise, the sensitivity for the prediction of syntenic blocks will be low. To increase the completeness of the *Cs*-k2 gene set, evidence from multiple sources, including proteomes of relatively close-related species and transcriptomic data of the two *C. sinensis* isolates, was used to predict candidate gene models. These candidate gene models were then compared and merged with *ab initio* gene predictions into a consensus gene set using two gene set merger programs (Maker2 and EVM). By applying these steps, the *Cs*-k2 gene set displayed a higher level of completeness compared with *Cs*-c2 (Chapter 3). The orthologous grouping of the proteins, however, showed that each isolate still had a substantial number of unique genes, which could not be explained by nucleotide differences alone. The number of such unique genes is also affected by the quality and fragmentation of the assemblies. For example, one gene might be assembled into multiple copies due to heterozygosity (Denton et al., 2014). In addition, two very similar genes could be assembled into a single gene due to an oversimplification of repetitive regions. Other reasons might be a failed transfer of some gene models from *Cs*-c2 to *Cs*-k2 or a lack of sufficient transcriptomic evidence for all developmental stages to support and refine gene models. Currently, due to the requirement of sequence alignment between two genomes (Otto et al., 2011), the effectiveness of gene transfer is still limited by sequence fragmentation, which further highlights the importance of assembly quality.

Another reason leading to the incompleteness of gene prediction might be caused by a systematic error in the gene filtering procedure. In Chapter 3, a decision tree was designed and applied to filter gene candidates. Although a number of factors, such as exon number, transcriptomic support, repeat coverage and functional annotation, were taken into account, protein-encoding genes might still be filtered out during this step. Currently, it is a challenge to capture all the signatures of a functional gene, particularly for gene models without close homologs in public databases. Machine learning algorithms are quite promising in this area. These algorithms are able to automatically divide inputs into different groups through supervised or unsupervised learning of defined features (Libbrecht et al., 2015). In this way, a more systematic gene filtering approach could be developed in the future to infer accurate gene sets.

5.3. Associating patterns of genetic variation of *C. sinensis* with demographic and local adaptation events

Using the workflow designed in Chapter 3, a high degree of nucleotide identity was shown in the aligned syntenic blocks, suggesting a relatively low divergence between Korean and Chinese isolates. In contrast, the mt genome results (Chapter 2) revealed a pairwise nucleotide diversity pattern linked to geographical location. Based on pairwise nucleotide diversities in mt DNA, *Cs*-k2 is close genetically to *Cs*-k1, followed by *Cs*-r1 and *Cs*-c1. This result is consistent with the previous genetic studies in *C. sinensis* (Park and Yong, 2001; Lee and Huh, 2004; Le et al., 2006; Cai et al., 2012; Liu et al., 2012a; Tatonova et al., 2012; Tatonova et al., 2013; Chelomina et al., 2014).

An hypothesis to explain these findings is that this parasite experienced a rapid population expansion after the last glacial maximum (LGM) (You et al., 2010; Liu and Jiang, 2016). This hypothesis can be derived from the dynamic changes of the living conditions of the first intermediate hosts of the flukes. For example, the main first intermediate host of *C. sinensis*, a freshwater snail *Parafossarulus manchouricus*, is usually active when the water temperature is above 12C° —the snail requires at least 4 months for the process from mating to the sexual maturity of offspring (Li et al., 1979; Chung et al., 1980). Currently, the most northern habitat of the snail hosts is in the Amur river basin, near the border between China and Russia. In the glacial period, the annual average temperature in North China was predicted to be at least 5C° colder than presently is the case (Liu and Jiang, 2016). Such climate changes would not allow most snail hosts of *C. sinensis* to survive, and, hence, would have limited population growth in this geographical region.

In contrast to North China, several regions in South and Central China had a relatively warm and stable climate during LGM and, hence, would have served as glacial refugia for different species (You et al., 2010; Qiu et al., 2011). Following the LGM, the population of snail hosts and the worms would have rapidly expanded as the climate outside of the refugia began to warm up and, hence, became favourable for snails to reproduce again (Hewitt, 1999; You et al., 2010). Patterns of genetic variation caused by post-glacial expansion have been widely investigated (Hewitt, 1999; Hoarau et al., 2007; Provan and Bennett, 2008; Allcock and Strugnell, 2012). For example, during the expansion, a low frequency of random mutations would have generated large numbers of haplotypes that are highly similar to the ancestral haplotype, and, hence, led to low nucleotide diversity and high haplotype diversity. Such patterns were reported in recent studies of *C. sinensis* (Tatonova et al., 2013; Chelomina et al., 2014). Other evidence of post-glacial expansion has also been reported, such as low genetic diversity in Northeast China and relatively high diversity in Central and South China (Liu et al., 2012a; Tatonova et al., 2012). This pattern is considered to be a signal of post-glacial recolonisation (Hewitt, 1999). Due to the smaller size of newly established populations, genetic drift causes lower genetic variation at the frontier of the recolonised area (Allcock and Strugnell, 2012). Therefore, the population structure in an original glacial refugium is usually more diverse compared with that in a previous iced territory (Hewitt, 1999).

Currently, the number of geographically distinct glacial refugia for *C. sinensis* is unclear. However, such information might be gleaned from the genetic changes due to isolation for long periods of time. For example, despite the low nucleotide divergence within the syntenic blocks, half of the two assemblies could not be aligned in syntenic blocks. Besides mis-assembly and genome assembly fragmentation, the lack of synteny could be caused by genome rearrangements, including duplication, insertion and/or translocation, occurring in geographic isolation. For instance, a long-term geographic isolation might have occurred in a glacial refugium on the Korean Peninsula during the ice age, as reported for several other animal and plant species (Zhang et al., 2008a; Bai et al., 2010; Qiu et al., 2011; Zhang et al., 2016). Furthermore, geographical barriers, such as Yalu, the Tumen River and the Changbai Mountains, may have prevented gene flow between the Korean Peninsula and the Chinese mainland. Gene flow introduced by human migration may also have been limited by the Great Wall (since 220 BC) between central and Northeast China.

In addition to demographic events caused by climate change and human activities, environmental differences between China and Korea are also likely to exert selective forces and might “drive” the evolution and adaptation of this parasite. Interestingly, this study (Chapter 3) showed that most (31/42) nucleotide differences (NDs) in one variable gene encoding cathepsin D (CSKR_13438s) are nonsynonymous, a genetic pattern that is consistent with adaptive evolution.

Cathepsins mainly function in protein degradation and might be critical in miracidial penetration of the snail host (Yoshino et al., 1993). Hence, beneficial mutations in cathepsin genes may increase the fitness of *C. sinensis* to adapt to a novel intermediate host. Such an ability could also be reflected by the large range of snail hosts (representing four families) of *C. sinensis* (Lun et al., 2005). Compared with *P. manchouricus*, other snail host species have relatively narrow distributions in East Asia. However, all of these regions overlap in the habitats of *P. manchouricus*, a phenomenon that is not likely to be a random event. In contrast, this might also be the result of natural selection due to long-term host-parasite interaction (Peters and Lively, 1999; Decaestecker et al., 2007; King et al., 2011). For example, the increasing density of non-host snails living in a particular area could impose strong selective pressure on the parasite by reducing its ability to infect susceptible host snails (King et al., 2011; Civitello et al., 2015). Therefore, the lineage of *C. sinensis* that can infect more diverse snail hosts is likely to produce more offspring. Another trematode, *S. mansoni*, was also recorded to have an increased fitness in snail host after only 3-4 generations of artificial selection (Webster et al., 2007). Hence, it seems reasonable to propose that the Chinese liver fluke might also have the capacity to rapidly adapt to new host snails. However, no studies have yet been conducted to investigate the compatibility and evolutionary relationship between *C. sinensis* and its snail hosts. This is an area that warrants investigation.

5.4. Selecting genetic markers for *C. sinensis*

In Chapter 3, nucleotide differences (NDs) in the aligned nuclear genomes of *C. sinensis* isolates were curated. Syntenic blocks, which cover ~ 50% of each genome, provided a template for high quality read alignment and variation detection. These syntenic blocks (with conserved gene order) serve as a solid foundation for linkage disequilibrium and selective sweep analyses. Within the coding regions of orthologous genes, 15,096 NDs were identified (Chapter 3). Among them, the number of non-synonymous NDs and synonymous NDs is more even in nuclear genomes than in mt genomes, reflecting a lower level of overall purifying selection on the nuclear genomes.

Demographic events and/or natural selection might have shaped genetic diversity within *C. sinensis*. However, for the two pooled isolates analysed herein, there is insufficient whole genomic data to differentiate genetic variation patterns linked to demographic events and natural selection. Most of the demographic inference methods are based on neutral theories and assume that intraspecific variation has little impact on the fitness of an organism (Holderegger et al., 2006). Without considering natural selection, the maintenance of genetic variation represents a dynamic balance between random mutations and a constant loss of alleles caused by random sampling from a given population ("genetic drift") (Leffler et al., 2012). Within this balance, the rate of random mutation and genetic drift correlates positively and negatively with N_e, respectively (Hartl, 1980). Based on these assumptions, one could estimate a series of population parameters, such as N_e and population differentiation, using observed data (e.g., allele frequencies) as well as, for example, mutation and recombination rates (Leffler et al., 2012). However, these population parameters need to be interpreted with caution for liver flukes due to their complex life cycles (Criscione et al., 2005). For example, an increased selfing rate and the variance of clonal reproduction in snail hosts could significantly decrease N_e for the parasite (cf. Prugnolle et al., 2005a). Furthermore, selfing organisms are expected to have a lower recombination rate, reflecting an increased linkage disequilibrium. Therefore, compared with outcrossing organisms, natural selection in selfing organisms is expected to act on longer linkage disequilibrium blocks and likely to affect larger genomic regions (Thomas et al., 2015).

To date, the understanding of *C. sinensis* evolution has been enhanced by studies of genetic variation in mtDNA (Park and Yong, 2001; Huang et al., 2012b; Tatonova et al., 2013; Chelomina et al., 2014). However, as indicated, there are some limitations with using mt genome markers, such as an oversimplification of population structure and genetic hitchhiking caused by natural selection and linkage disequilibrium (Zhang and Hewitt, 2003; Galtier et al., 2009). Recently, genome-wide microsatellite DNA markers have been characterised in *C. sinensis* (Nguyen et al., 2015). Microsatellite markers have been utilised in population analyses due to their neutrality and high levels of polymorphism (see Li et al., 2002). However, microsatellite markers usually contain more than two alleles at the same locus and have complex evolutionary patterns (Zhang and Hewitt, 2003). Furthermore, the mutational mechanisms in microsatellite markers are often unclear (Selkoe and Toonen, 2006), and the mutation rates can vary substantially among different species (Xu et al., 2000). Due to these limitations, microsatellites are unsuitable genetic markers for inferring ancestral states.

In contrast to neutral loci, non-neutral (or adaptive) loci are impacted by both demographic events and selective pressure. The use of non-neutral loci for demographic inference, without considering natural selection, might lead to a false conclusion. For instance, 5% of non-neutral sites are able to alter the predicted fixation index (F_{ST})

value by 30% to 50%, and, hence, would alter the predicted population structure (Allendorf et al., 2010). To date, it is still a challenge to effectively detect selection signatures (Kirk and Freeland, 2011). Such detection, however, is more frequently conducted in humans (Oleksyk et al., 2010), and might be applicable to similar analyses in parasites. In addition, these nuclear DNA loci are often found to relate to infection and adaptation (Kelley and Swanson, 2008; Oleksyk et al., 2010). For example, the NDs identified in Chapter 3 provide valuable information for identifying such loci, and might ultimately contribute to the identification of drug and vaccine targets. Besides cathepsin D, an adaptive evolutionary pattern was identified for the Niemann-Pick C2 (NPC2) protein. This protein is predicted to function in lipid transportation and chemical homeostasis (Xu et al., 2007). The adaptive evolution of this protein and its gene might facilitate *C. sinensis* to recognise and transport a variety of lipid molecules and increase its fitness in the definitive host. It might also have potential as a vaccine candidate, as supported by research findings showing that *C. sinensis* DNA encodes a lipid-binding protein that triggers protective immune response in rats (Lee et al., 2006).

5.5. Genetic variation within *S. japonicum*

Although the direct comparison of genome assemblies can be used to establish levels of genetic variation, this approach is computationally intensive, requires a robust and effective bioinformatic workflow system and can be challenging, particularly when large genomes representing many geographical isolates are compared. For species with no evidence of genomic rearrangement linked to a conserved karyotype, mapping short read DNA sequence data to a well-assembled reference genome is an efficient, alternative approach for estimating population genetic variation. Since schistosome species usually have eight pairs of chromosomes (Rollinson et al., 1997; Lawton et al., 2011), the latter approach was employed in Chapter 4. Here, read data for pooled adult worms of *S. japonicum* representing distinct isolates were individually mapped to a reference genome for consensus base-calling to select representative SNPs in non-repetitive gene regions. Using these SNP data, there was a clear genetic differentiation between SW and CL isolates using both concatenated and coalesced phylogenies. In contrast to the approach used in Chapter 4, a number of previous population studies (using sequence data for isolates of pooled worms) estimated allelic frequency based on statistics for individual pools, and used them to identify adaptive selection and to infer inter-population relationships (Nolte et al., 2013; Guo et al., 2015; Kapun et al., 2016; Dal Grande et al., 2017), with the assumption that DNA libraries for sequencing evenly represent genetic variation across all individuals in a pool (Chen et al., 2012; Hivert et al., 2018). Applying such libraries and analyses can be challenging in flatworms, particularly when genetic variation within and among libraries is high, given the potential aggregation of clonal adult worms due to their asexual reproductive stage (Prugnolle et al., 2005).

The complexity of "pooled sequencing" analysis is enhanced when population structure is unclear. For example, sampling individuals from a population that contains cryptic subdivisions can lead to an underestimate of heterozygosity (Wahlund effect) (Wahlund, 1928). This phenomenon has been frequently reported in flatworms (Vilas et al., 2003; Saijuntha et al., 2009; Criscione et al., 2011), which might be associated with the clonal reproduction within their life cycle (Halkett et al., 2005). Given the temporal changes and the high proportion of contradictory gene trees (Chapter 4), such cases may be common in *S. japonicum*. A skewed proportion of individual DNA also impedes a comprehensive detection of adaptive loci due to the biased estimation of allelic frequency (Chen et al., 2012). To overcome the influences caused by uncertain sampling, adaptive loci were identified based on the comparison of different categories of genetic variation (McDonald–Kreitman test; McDonald and Kreitman, 1991), assuming that both categories are equally affected by the same sampling and demographic changes. However, due to the limited information of intrapopulation polymorphism and linkage disequilibrium, how those adaptive genes interact with flanking regions remains to be explored. In contrast to "pooled sequencing" techniques, "individual sequencing" is able to identify genetically identical individuals, which enables a reliable estimation of intra-population polymorphism, genome-wide haplotype diversity and linkage disequilibrium of target species. With the decreasing cost of high throughput sequencing, further exploration should be performed to investigate the population structure of *S. japonicum* and functional relevance of the identified adaptive genes in the context of specified inbreeding and demographic models.

5.6. Future work

The high-quality nuclear genome of *C. sinensis* isolate from Korea, and genetic markers identified in syntenic blocks between Korea and China (Chapter 3) provide a sound basis for population genetic investigations of *C. sinensis* from geographically distinct areas. However, given the complex biology of flukes, such investigations require careful experimental design and a systematic bioinformatic framework. To achieve this, individuals should be randomly sampled from all the endemic areas. Following DNA sequencing of individuals, reads should be mapped to the most appropriate reference genome for analysis (i.e. *Cs*-k2 or *Cs*-c2). After selecting the reference genome and calling variants, genetic relationships among individuals should be examined to exclude clonal individuals and define population structures preceding further analyses. In addition, neutral loci, such as synonymous polymorphic sites and intronic regions, should be selected and applied to demographic analyses. Based on the patterns of variation and heterozygosity of neutral loci, estimates, such as self-fertilisation rate, genetic diversity, recombination rate and historical population size, should be calculated (Cutter, 2006; Li and Durbin, 2010; Thomas et al., 2015). These estimates are critical to understand the interaction between demographic events, reproduction and lineage divergence, and can be used to simulate selective and neutral loci with the same demographic history. The comparison between simulated and observed data sets would predict patterns of natural selection and demographic events. Based on this information, genetic loci related to local adaptive process might be reliably identified. Studying such loci will be particularly important when working toward defining new drug and vaccine targets.

5.7. Concluding remarks

In this thesis, both mt and nuclear genomes of *C. sinensis* Korean isolate were assembled, annotated and compared with other isolates. The comparison of mt genomes (Chapter 2) not only confirmed the specific identity of our specimen, but also highlighted potential issues with using mtDNA markers for genetic analyses. In contrast to mt genomes, exploring genetic variation in nuclear genomes (Chapter 3) showed major promise due to the nature and extent of nucleotide divergence. Although focused mainly on *C. sinensis*, some of this work extended to an investigation of the population genetic structure of a related worm, *S. japonicum*, in China (Chapter 4) and of genes that under positive selection in particular geographical regions.

To overcome the technical challenges of comparing draft genomes and investigating the application potential of nuclear genomes in population genetic studies, a refined bioinformatic workflow was established, and successfully used to define high-quality syntenic blocks between two genomes. The nuclear genome and the high quality syntenic blocks defined in this thesis now serve as a foundation for a future population genetic analysis of *C. sinensis*. The coding regions of these blocks contain a substantial number of genetic loci that might serve well for the discovery of new fluke-specific intervention targets.

Compared with coding regions, the genetic variation in the intronic regions showed an improved phylogenetic signal at both the whole genome and individual gene levels (Chapter 4). The variance of phylogenetic composition among individual genes reflect the complexity of pool sequencing, historical events and local adaptation. In future, population genomic research of parasitic flatworms should be achieved using whole genomic data of representative individuals. Furthermore, a systematic bioinformatic framework is required to robustly discover individual variants, infer population structure and identify adaptive selection, with the consideration of parasitic life cycle and demographic history. Although the present thesis focused on *C. sinensis* and *S. japonicum*, the findings and the approaches established here have broader implications for studies of other flatworm parasites.

5.8. References

Alkan, C., Sajjadian, S., Eichler, E.E., 2011. Limitations of next-generation genome sequence assembly. Nat. Methods 8, 61-65.

Allcock, A.L., Strugnell, J.M., 2012. Southern Ocean diversity: new paradigms from molecular ecology. Trends Ecol. Evol. 27, 520-528.

Allendorf, F.W., Hohenlohe, P.A., Luikart, G., 2010. Genomics and the future of conservation genetics. Nat. Rev. Genet. 11, 697-709.

Ansari, H.R., Templeton, T.J., Subudhi, A.K., Ramaprasad, A., Tang, J., Lu, F., Naeem, R., Hashish, Y., Oguike, M.C., Benavente, E.D., Clark, T.G., Sutherland, C.J., Barnwell, J.W., Culleton, R., Cao, J., Pain, A., 2016. Genome-scale comparison of expanded gene families in *Plasmodium ovale wallikeri* and *Plasmodium ovale curtisi* with *Plasmodium malariae* and with other *Plasmodium* species. Int. J. Parasitol. 46, 685-696.

Awadalla, P., Eyre-Walker, A., Smith, J.M., 1999. Linkage disequilibrium and recombination in hominid mitochondrial DNA. Science 286, 2524-2525.

Bai, W.N., Liao, W.J., Zhang, D.Y., 2010. Nuclear and chloroplast DNA phylogeography reveal two refuge areas with asymmetrical gene flow in a temperate walnut tree from East Asia. New Phytol. 188, 892-901.

Bazin, E., Glémin, S., Galtier, N., 2006. Population size does not influence mitochondrial genetic diversity in animals. Science 312, 570-572.

Bourque, G., Pevzner, P.A., Tesler, G., 2004. Reconstructing the genomic architecture of ancestral mammals: lessons from human, mouse, and rat genomes. Genome Res. 14, 507-516.

Breton, S., Beaupre, H.D., Stewart, D.T., Hoeh, W.R., Blier, P.U., 2007. The unusual system of doubly uniparental inheritance of mtDNA: isn't one enough? Trends Genet. 23, 465-474.

Cai, X.Q., Liu, G.H., Song, H.Q., Wu, C.Y., Zou, F.C., Yan, H.K., Yuan, Z.G., Lin, R.Q., Zhu, X.Q., 2012. Sequences and gene organization of the mitochondrial genomes of the liver flukes *Opisthorchis viverrini* and *Clonorchis sinensis* (Trematoda). Parasitol. Res. 110, 235-243.

Chelomina, G.N., Tatonova, Y.V., Hung, N.M., Ngo, H.D., 2014. Genetic diversity of the Chinese liver fluke *Clonorchis sinensis* from Russia and Vietnam. Int. J. Parasitol. 44, 795-810.

Chen, X., Listman, J.B., Slack, F.J., Gelernter, J., Zhao, H., 2012. Biases and errors on allele frequency estimation and disease association tests of next-generation sequencing of pooled samples. Genet. Epidemiol. 36, 549-560.

Chung, B.J., Joo, C.Y., Choi, D.W., 1980. Seasonal variation of snail population of *Parafossarulus manchouricus* and larval trematode infection in river Kumho, Kyungpook province, Korea. Kisaengchunghak Chapchi 18, 54-64 (in Korean).

Civitello, D.J., Cohen, J., Fatima, H., Halstead, N.T., Liriano, J., McMahon, T.A., Ortega, C.N., Sauer, E.L., Sehgal, T., Young, S., Rohr, J.R., 2015. Biodiversity inhibits parasites: Broad evidence for the dilution effect. Proc. Natl. Acad. Sci. U. S. A. 112, 8667-8671.

Clement, J.A., Toulza, E., Gautier, M., Parrinello, H., Roquis, D., Boissier, J., Rognon, A., Mone, H., Mouahid, G., Buard, J., Mitta, G., Grunau, C., 2013. Private selective sweeps identified from next-generation pool-sequencing reveal convergent pathways under selection in two inbred *Schistosoma mansoni* strains. PLoS Negl. Trop. Dis. 7, e2591.

Crellen, T., Allan, F., David, S., Durrant, C., Huckvale, T., Holroyd, N., Emery, A.M., Rollinson, D., Aanensen, D.M., Berriman, M., Webster, J.P., Cotton, J.A., 2016. Whole genome resequencing of the human parasite *Schistosoma mansoni* reveals population history and effects of selection. Sci. Rep. 6, 20954.

Criscione, C.D., Poulin, R., Blouin, M.S., 2005. Molecular ecology of parasites: elucidating ecological and microevolutionary processes. Mol. Ecol. 14, 2247-2257.

Criscione, C.D., Valentim, C.L., Hirai, H., LoVerde, P.T., Anderson, T.J., 2009. Genomic linkage map of the human blood fluke *Schistosoma mansoni*. Genome Biol. 10, R71.

Criscione, C.D., Vilas, R., Paniagua, E., Blouin, M.S., 2011. More than meets the eye: detecting cryptic microgeographic population structure in a parasite with a complex life cycle. Mol. Ecol. 20, 2510-2524.

Cutter, A.D., 2006. Nucleotide polymorphism and linkage disequilibrium in wild populations of the partial selfer *Caenorhabditis elegans*. Genetics 172, 171-184.

Dal Grande, F., Sharma, R., Meiser, A., Rolshausen, G., Budel, B., Mishra, B., Thines, M., Otte, J., Pfenninger, M., Schmitt, I., 2017. Adaptive differentiation coincides with local bioclimatic conditions along an elevational cline in populations of a lichen-forming fungus. BMC Evol. Biol. 17, 93.

Decaestecker, E., Gaba, S., Raeymaekers, J.A., Stoks, R., Van Kerckhoven, L., Ebert, D., De Meester, L., 2007. Host-parasite 'Red Queen' dynamics archived in pond sediment. Nature 450, 870-873.

Denton, J.F., Lugo-Martinez, J., Tucker, A.E., Schrider, D.R., Warren, W.C., Hahn, M.W., 2014. Extensive error in the number of genes inferred from draft genome assemblies. PLoS Comput. Biol. 10, e1003998.

Detwiler, J.T., Criscione, C.D., 2010. An infectious topic in reticulate evolution: introgression and hybridization in animal parasites. Genes (Basel) 1, 102-123.

Fierst, J.L., 2015. Using linkage maps to correct and scaffold *de novo* genome assemblies: methods, challenges, and computational tools. Front. Genet. 6, 220.

Galtier, N., Nabholz, B., Glémin, S., Hurst, G.D., 2009. Mitochondrial DNA as a marker of molecular diversity: a reappraisal. Mol. Ecol. 18, 4541-4550.

Gao, L., You, S., Chen, S., Wu, M., 1993. Study on meiosis of *Clonorchis sinensis*. Chin. J. Schistosomiasis Control 5, 230–233 (in Chinese).

Gao, L., You, S., Chen, S., Wu, M., Li, G., Li, W., You, T., 1987. Primary analysis of karyotypes of *Clonorchis sinensis*. J. Hengyang Med. Coll. 15, 108-112 (in Chinese).

Gibson, T., Blok, V.C., Phillips, M.S., Hong, G., Kumarasinghe, D., Riley, I.T., Dowton, M., 2007. The mitochondrial subgenomes of the nematode *Globodera pallida* are mosaics: evidence of recombination in an animal mitochondrial genome. J. Mol. Evol. 64, 463-471.

Goldman, D., Domschke, K., 2014. Making sense of deep sequencing. Int. J. Neuropsychopharmacol. 17, 1717-1725.

Guo, B., DeFaveri, J., Sotelo, G., Nair, A., Merila, J., 2015. Population genomic evidence for adaptive differentiation in Baltic Sea three-spined sticklebacks. BMC Biol. 13, 19.

Gurevich, A., Saveliev, V., Vyahhi, N., Tesler, G., 2013. QUAST: quality assessment tool for genome assemblies. Bioinformatics 29, 1072-1075.

Hahn, M.W., Zhang, S.V., Moyle, L.C., 2014. Sequencing, assembling, and correcting draft genomes using recombinant populations. G3 (Bethesda) 4, 669-679.

Halkett, F., Simon, J.-C., Balloux, F., 2005. Tackling the population genetics of clonal and partially clonal organisms. Trends Ecol. Evol. 20, 194-201.

Hane, J.K., Ming, Y., Kamphuis, L.G., Nelson, M.N., Garg, G., Atkins, C.A., Bayer, P.E., Bravo, A., Bringans, S., Cannon, S., Edwards, D., Foley, R., Gao, L.L., Harrison, M.J., Huang, W., Hurgobin, B., Li, S., Liu, C.W., McGrath, A., Morahan, G., Murray, J., Weller, J., Jian, J.B., Singh, K.B., 2017. A comprehensive draft genome sequence for lupin (*Lupinus angustifolius*), an emerging health food: insights into plant-microbe interactions and legume evolution. Plant Biotechnol. J. 15, 318-330.

Harrow, J., Frankish, A., Gonzalez, J.M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B.L., Barrell, D., Zadissa, A., Searle, S., Barnes, I., Bignell, A., Boychenko, V., Hunt, T., Kay, M., Mukherjee, G., Rajan, J., Despacio-Reyes, G., Saunders, G., Steward, C., Harte, R., Lin, M., Howald, C., Tanzer, A., Derrien, T., Chrast, J., Walters, N., Balasubramanian, S., Pei, B., Tress, M., Rodriguez, J.M., Ezkurdia, I., van Baren, J., Brent, M., Haussler, D., Kellis, M., Valencia, A., Reymond, A., Gerstein, M., Guigo, R., Hubbard, T.J., 2012. GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res. 22, 1760-1774.

Hartl, D.L., 1980. Principles of population genetics. Sinauer Associates, Sunderland, Mass.

Hashimoto, T., Horikawa, D.D., Saito, Y., Kuwahara, H., Kozuka-Hata, H., Shin, I.T., Minakuchi, Y., Ohishi, K., Motoyama, A., Aizu, T., Enomoto, A., Kondo, K., Tanaka, S., Hara, Y., Koshikawa, S., Sagara, H., Miura, T., Yokobori, S., Miyagawa, K., Suzuki, Y., Kubo, T., Oyama, M., Kohara, Y., Fujiyama, A., Arakawa, K., Katayama, T., Toyoda, A., Kunieda, T., 2016. Extremotolerant tardigrade genome and improved radiotolerance of human cultured cells by tardigrade-unique protein. Nat. Commun. 7, 12808.

Hedgecock, D., Shin, G., Gracey, A.Y., Den Berg, D.V., Samanta, M.P., 2015. Second-generation linkage maps for the pacific oyster *Crassostrea gigas* reveal errors in assembly of genome scaffolds. G3 (Bethesda) 5, 2007-2019.

Hester, J., Chan, E.R., Menard, D., Mercereau-Puijalon, O., Barnwell, J., Zimmerman, P.A., Serre, D., 2013. *De novo* assembly of a field isolate genome reveals novel *Plasmodium vivax* erythrocyte invasion genes. PLoS Negl. Trop. Dis. 7, e2569.

Hewitt, G.M., 1999. Post-glacial re-colonization of European biota. Biol. J. Linn. Soc. 68, 87-112.

Hivert, V., Leblois, R., Petit, E.J., Gautier, M., Vitalis, R., 2018. Measuring genetic differentiation from Pool-seq data. Genetics 210, 315-330.

Hoarau, G., Coyer, J.A., Veldsink, J.H., Stam, W.T., Olsen, J.L., 2007. Glacial refugia and recolonization pathways in the brown seaweed *Fucus serratus*. Mol. Ecol. 16, 3606-3616.

Holderegger, R., Kamm, U., Gugerli, F., 2006. Adaptive vs. neutral genetic diversity: implications for landscape genetics. Landsc. Ecol. 21, 797-807.

Holt, R.A., Subramanian, G.M., Halpern, A., Sutton, G.G., Charlab, R., Nusskern, D.R., Wincker, P., Clark, A.G., Ribeiro, J.M., Wides, R., Salzberg, S.L., Loftus, B., Yandell, M., Majoros, W.H., Rusch, D.B., Lai, Z., Kraft, C.L., Abril, J.F., Anthouard, V., Arensburger, P., Atkinson, P.W., Baden, H., de Berardinis, V., Baldwin, D., Benes, V., Biedler, J., Blass, C., Bolanos, R., Boscus, D., Barnstead, M., Cai, S., Center, A., Chaturverdi, K., Christophides, G.K., Chrystal, M.A., Clamp, M., Cravchik, A., Curwen, V., Dana, A., Delcher, A., Dew, I., Evans, C.A., Flanigan, M., Grundschober-Freimoser, A., Friedli, L., Gu, Z., Guan, P., Guigo, R., Hillenmeyer, M.E., Hladun, S.L., Hogan, J.R., Hong, Y.S., Hoover, J., Jaillon, O., Ke, Z., Kodira, C., Kokoza, E., Koutsos, A., Letunic, I., Levitsky, A., Liang, Y., Lin, J.J., Lobo, N.F., Lopez, J.R., Malek, J.A., McIntosh, T.C., Meister, S., Miller, J., Mobarry, C., Mongin, E., Murphy, S.D., O'Brochta, D.A., Pfannkoch, C., Qi, R., Regier, M.A., Remington, K., Shao, H., Sharakhova, M.V., Sitter, C.D., Shetty, J., Smith, T.J., Strong, R., Sun, J., Thomasova, D., Ton, L.Q., Topalis, P., Tu, Z., Unger, M.F., Walenz, B., Wang, A., Wang, J., Wang, M., Wang, X., Woodford, K.J., Wortman, J.R., Wu, M., Yao, A., Zdobnov, E.M., Zhang, H., Zhao, Q., Zhao, S., Zhu, S.C., Zhimulev, I., Coluzzi, M., della Torre, A., Roth, C.W., Louis, C., Kalush, F., Mural, R.J., Myers, E.W., Adams, M.D., Smith, H.O., Broder, S., Gardner, M.J., Fraser, C.M., Birney, E., Bork, P., Brey, P.T., Venter, J.C., Weissenbach, J., Kafatos, F.C., Collins, F.H., Hoffman, S.L., 2002. The genome sequence of the malaria mosquito *Anopheles gambiae*. Science 298, 129-149.

Huang, J., Zhao, Y., Shiraigol, W., Li, B., Bai, D., Ye, W., Daidiikhuu, D., Yang, L., Jin, B., Zhao, Q., Gao, Y., Wu, J., Bao, W., Li, A., Zhang, Y., Han, H., Bai, H., Bao, Y., Zhao, L., Zhai, Z., Zhao, W., Sun, Z., Zhang, Y., Meng, H., Dugarjaviin, M., 2014. Analysis of horse genomes provides insight into the diversification and adaptive evolution of karyotype. Sci. Rep. 4, 4958.

Huang, S., Chen, Z., Huang, G., Yu, T., Yang, P., Li, J., Fu, Y., Yuan, S., Chen, S., Xu, A., 2012a. HaploMerger: reconstructing allelic relationships for polymorphic diploid genome assemblies. Genome Res. 22, 1581-1588.

Huang, S.Y., Zhao, G.H., Fu, B.Q., Xu, M.J., Wang, C.R., Wu, S.M., Zou, F.C., Zhu, X.Q., 2012b. Genomics and molecular genetics of *Clonorchis sinensis*: current status and perspectives. Parasitol. Int. 61, 71-76.

Hung, C.M., Zink, R.M., 2014. Distinguishing the effects of selection from demographic history in the genetic variation of two sister passerines based on mitochondrial-nuclear comparison. Heredity (Edinb.) 113, 42-51.

Hunt, M., Kikuchi, T., Sanders, M., Newbold, C., Berriman, M., Otto, T.D., 2013. REAPR: a universal tool for genome assembly evaluation. Genome Biol. 14, R47.

Hunt, M., Newbold, C., Berriman, M., Otto, T.D., 2014. A comprehensive evaluation of assembly scaffolding tools. Genome Biol. 15, R42.

Jannotti-Passos, L.K., Souza, C.P., Parra, J.C., Simpson, A.J.G., 2001. Biparental mitochondrial DNA inheritance in the parasitic trematode *Schistosoma mansoni*. J. Parasitol. 87, 79-82.

Jirimutu, Wang, Z., Ding, G., Chen, G., Sun, Y., Sun, Z., Zhang, H., Wang, L., Hasi, S., Zhang, Y., Li, J., Shi, Y., Xu, Z., He, C., Yu, S., Li, S., Zhang, W., Batmunkh, M., Ts, B., Narenbatu, Unierhu, Bat-Ireedui, S., Gao, H., Baysgalan, B., Li, Q., Jia, Z., Turigenbayila, Subudenggerile, Narenmanduhu, Wang, Z., Wang, J., Pan, L., Chen, Y., Ganerdene, Y., Dabxilt, Erdemt, Altansha, Altansukh, Liu, T., Cao, M., Aruuntsever, Bayart, Hosblig, He, F., Zha-ti, A., Zheng, G., Qiu, F., Sun, Z., Zhao, L., Zhao, W., Liu, B., Li, C., Chen, Y., Tang, X., Guo, C., Liu, W., Ming, L., Temuulen, Cui, A., Li, Y., Gao, J., Li, J., Wurentaodi, Niu, S., Sun, T., Zhai, Z., Zhang, M., Chen, C., Baldan, T., Bayaer, T., Li, Y., Meng, H., 2012. Genome sequences of wild and domestic bactrian camels. Nat. Commun. 3, 1202.

Jones, T., Federspiel, N.A., Chibana, H., Dungan, J., Kalman, S., Magee, B.B., Newport, G., Thorstenson, Y.R., Agabian, N., Magee, P.T., Davis, R.W., Scherer, S., 2004. The diploid genome sequence of *Candida albicans*. Proc. Natl. Acad. Sci. U. S. A. 101, 7329-7334.

Kajitani, R., Toshimoto, K., Noguchi, H., Toyoda, A., Ogura, Y., Okuno, M., Yabana, M., Harada, M., Nagayasu, E., Maruyama, H., Kohara, Y., Fujiyama, A., Hayashi, T., Itoh, T., 2014. Efficient *de novo* assembly of highly heterozygous genomes from whole-genome shotgun short reads. Genome Res. 24, 1384-1395.

Kapun, M., Fabian, D.K., Goudet, J., Flatt, T., 2016. Genomic Evidence for Adaptive Inversion Clines in *Drosophila melanogaster*. Mol. Biol. Evol. 33, 1317-1336.

Kelley, J.L., Swanson, W.J., 2008. Positive selection in the human genome: from genome scans to biological significance. Annu. Rev. Genomics Hum. Genet. 9, 143-160.

King, K.C., Jokela, J., Lively, C.M., 2011. Trematode parasites infect or die in snail hosts. Biol. Lett. 7, 265-268.

Kirk, H., Freeland, J.R., 2011. Applications and implications of neutral versus non-neutral markers in molecular ecology. Int. J. Mol. Sci. 12, 3966-3988.

Korhonen, P.K., Hall, R.S., Young, N.D., Gasser, R.B., 2019. Common Workflow Language (CWL)-based software pipeline for *de novo* genome assembly from long- and short-read data. GigaScience. (in press)

Korhonen, P.K., Young, N.D., Gasser, R.B., 2016. Making sense of genomes of parasitic worms: Tackling bioinformatic challenges. Biotechnol. Adv. 34, 663-686.

Lakshmipathy, U., Campbell, C., 1999. Double strand break rejoining by mammalian mitochondrial extracts. Nucleic Acids Res. 27, 1198-1204.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J.P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J.C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R.H., Wilson, R.K., Hillier, L.W., McPherson, J.D., Marra, M.A., Mardis, E.R., Fulton, L.A., Chinwalla, A.T., Pepin, K.H., Gish, W.R., Chissoe, S.L., Wendl, M.C., Delehaunty, K.D., Miner, T.L., Delehaunty, A., Kramer, J.B., Cook, L.L., Fulton, R.S., Johnson, D.L., Minx, P.J., Clifton, S.W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J.F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R.A., Muzny, D.M., Scherer, S.E., Bouck, J.B., Sodergren, E.J., Worley, K.C., Rives, C.M., Gorrell, J.H., Metzker, M.L., Naylor, S.L., Kucherlapati, R.S., Nelson, D.L., Weinstock, G.M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Yang, H.M., Yu, J., Wang, J., Huang, G.Y., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S.Z., Davis, R.W., Federspiel, N.A., Abola, A.P., Proctor, M.J., Myers, R.M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D.R., Olson, M.V., Kaul, R., Raymond, C., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G.A., Athanasiou, M., Schultz, R., Roe, B.A., Chen, F., Pan, H.Q., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W.R., de la Bastide, M., Dedhia, N., Blocker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J.A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D.G., Burge, C.B., Cerutti, L., Chen, H.C., Church, D., Clamp, M., Copley, R.R., Doerks, T., Eddy, S.R., Eichler, E.E., Furey, T.S., Galagan, J., Gilbert, J.G.R., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W.H., Johnson, L.S., Jones, T.A., Kasif, S., Kaspryzk, A., Kennedy, S., Kent, W.J., Kitts, P., Koonin, E.V., Korf, I., Kulp, D., Lancet, D., Lowe, T.M., McLysaght, A., Mikkelsen, T., Moran, J.V., Mulder, N., Pollara, V.J., Ponting, C.P., Schuler, G., Schultz, J.R., Slater, G., Smit, A.F.A., Stupka, E., Szustakowki, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y.I., Wolfe, K.H., Yang, S.P., Yeh, R.F., Collins, F., Guyer, M.S., Peterson, J., Felsenfeld, A., Wetterstrand, K.A., Patrinos, A., Morgan, M.J., Conso, I.H.G.S., 2001. Initial sequencing and analysis of the human genome. Nature 409, 860-921.

Lawton, S.P., Hirai, H., Ironside, J.E., Johnston, D.A. and Rollinson, D., 2011. Genomes and geography: genomic insights into the evolution and phylogeography of the genus Schistosoma. Parasit. Vectors 4, 131.

Le, T.H., Blair, D., McManus, D.P., 2001. Complete DNA sequence and gene organization of the mitochondrial genome of the liverfluke, *Fasciola hepatica* L. (Platyhelminthes; Trematoda). Parasitology 123, 609-621.

Le, T.H., Van De, N., Blair, D., Sithithaworn, P., McManus, D.P., 2006. *Clonorchis sinensis* and *Opisthorchis viverrini:* development of a mitochondrial-based multiplex PCR for their identification and discrimination. Exp. Parasitol. 112, 109-114.

Lee, J.S., Kim, I.S., Sohn, W.M., Lee, J., Yong, T.S., 2006. A DNA vaccine encoding a fatty acid-binding protein of *Clonorchis sinensis* induces protective immune response in Sprague-Dawley rats. Scand. J. Immunol. 63, 169-176.

Lee, S.U., Huh, S., 2004. Variation of nuclear and mitochondrial DNAs in Korean and Chinese isolates of *Clonorchis sinensis*. Korean J. Parasitol. 42, 145-148.

Leffler, E.M., Bullaughey, K., Matute, D.R., Meyer, W.K., Segurel, L., Venkat, A., Andolfatto, P., Przeworski, M., 2012. Revisiting an old riddle: what determines genetic diversity levels within species? PLoS Biol. 10, e1001388.

Li, B., Wang, C., Li, D., Liu, T., Deng, L., Liu, Y., 1979. Ecology research in *Parafossarulus striatulus,* the first intermediate host of *Clonorchis sinensis* in Liaoning Province. J. China Med. Univ. 4, 3-6 (in Chinese).

Li, H., Durbin, R., 2010. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 26, 589-595.

Li, Y.C., Korol, A.B., Fahima, T., Beiles, A., Nevo, E., 2002. Microsatellites: genomic distribution, putative functions and mutational mechanisms: a review. Mol. Ecol. 11, 2453-2465.

Libbrecht, M.W., Noble, W.S., 2015. Machine learning applications in genetics and genomics. Nat. Rev. Genet. 16, 321.

Liu, G.H., Li, B., Li, J.Y., Song, H.Q., Lin, R.Q., Cai, X.Q., Zou, F.C., Yan, H.K., Yuan, Z.G., Zhou, D.H., Zhu, X.Q., 2012. Genetic variation among *Clonorchis sinensis* isolates from different geographic regions in China revealed by sequence analyses of four mitochondrial genes. J. Helminthol. 86, 479-484.

Liu, Y., Jiang, D., 2016. Last glacial maximum permafrost in China from CMIP5 simulations. Palaeogeogr. Palaeoclimatol. Palaeoecol. 447, 12-21.

Lun, Z.R., Gasser, R.B., Lai, D.H., Li, A.X., Zhu, X.Q., Yu, X.B., Fang, Y.Y., 2005. Clonorchiasis: a key foodborne zoonosis in China. Lancet Infect. Dis. 5, 31–41.

Mayer, K.F., Waugh, R., Brown, J.W., Schulman, A., Langridge, P., Platzer, M., Fincher, G.B., Muehlbauer, G.J., Sato, K., Close, T.J., Wise, R.P., Stein, N., 2012. A physical, genetic and functional sequence assembly of the barley genome. Nature 491, 711-716.

Menard, D., Chan, E.R., Benedet, C., Ratsimbasoa, A., Kim, S., Chim, P., Do, C., Witkowski, B., Durand, R., Thellier, M., Severini, C., Legrand, E., Musset, L., Nour, B.Y.M., Mercereau-Puijalon, O., Serre, D., Zimmerman, P.A., 2013. Whole Genome sequencing of field isolates reveals a common duplication of the duffy binding protein gene in Malagasy *Plasmodium vivax* Strains. PLoS Negl. Trop. Dis. 7, e2489.

Muggli, M.D., Puglisi, S.J., Ronen, R., Boucher, C., 2015. Misassembly detection using paired-end sequence reads and optical mapping data. Bioinformatics 31, i80-88.

Nachman, M.W., Brown, W.M., Stoneking, M., Aquadro, C.F., 1996. Nonneutral mitochondrial DNA variation in humans and chimpanzees. Genetics 142, 953-963.

Nguyen, T.T., Arimatsu, Y., Hong, S.J., Brindley, P.J., Blair, D., Laha, T., Sripa, B., 2015. Genome-wide characterization of microsatellites and marker development in the carcinogenic liver fluke *Clonorchis sinensis*. Parasitol. Res. 114, 2263-2272.

Nolte, V., Pandey, R.V., Kofler, R., Schlötterer, C., 2013. Genome-wide patterns of natural variation reveal strong selective sweeps and ongoing genomic conflict in *Drosophila mauritiana*. Genome Res. 23, 99-110.

Notsu, Y., Masood, S., Nishikawa, T., Kubo, N., Akiduki, G., Nakazono, M., Hirai, A., Kadowaki, K., 2002. The complete sequence of the rice (*Oryza sativa L.*) mitochondrial genome: frequent DNA sequence acquisition and loss during the evolution of flowering plants. Mol. Genet. Genomics 268, 434-445.

Oleksyk, T.K., Smith, M.W., O'Brien, S.J., 2010. Genome-wide scans for footprints of natural selection. Philos. Trans. R. Soc. Lond. B. Biol. Sci. 365, 185-205.

Otto, T.D., Dillon, G.P., Degrave, W.S., Berriman, M., 2011. RATT: Rapid Annotation Transfer Tool. Nucleic Acids Res. 39, e57.

Park, G.M., Im, K., Huh, S., Yong, T.S., 2000. Chromosomes of the liver fluke, *Clonorchis sinensis*. Korean J. Parasitol. 38, 201-206.

Park, G.M., Yong, T.S., 2001. Geographical variation of the liver fluke, *Clonorchis sinensis*, from Korea and China based on the karyotypes, zymodeme and DNA sequences. Southeast Asian J. Trop. Med. Public Health 32 (Suppl. 2), 12-16.

Peters, A.D., Lively, C.M., 1999. The Red Queen and Fluctuating Epistasis: A population genetic analysis of antagonistic coevolution. Am. Nat. 154, 393-405.

Piganeau, G., Gardner, M., Eyre-Walker, A., 2004. A broad survey of recombination in animal mitochondria. Mol. Biol. Evol. 21, 2319-2325.

Provan, J., Bennett, K.D., 2008. Phylogeographic insights into cryptic glacial refugia. Trends Ecol. Evol. 23, 564-571.

Prugnolle, F., Liu, H., de Meeus, T., Balloux, F., 2005. Population genetics of complex life-cycle parasites: an illustration with trematodes. Int. J. Parasitol. 35, 255-263.

Qiu, Y.X., Fu, C.X., Comes, H.P., 2011. Plant molecular phylogeography in China and adjacent regions: Tracing the genetic imprints of Quaternary climate and environmental change in the world's most diverse temperate flora. Mol. Phylogenet. Evol. 59, 225-244.

Ramirez-Soriano, A., Ramos-Onsins, S.E., Rozas, J., Calafell, F., Navarro, A., 2008. Statistical power analysis of neutrality tests under demographic expansions, contractions and bottlenecks with recombination. Genetics 179, 555-567.

Rollinson D., Kaukas A., Johnston D.A., Simpson A.J., Tanaka M., 1997. Some molecular insights into schistosome evolution. Int. J. Parasitol. 27, 11-28.

Saijuntha, W., Sithithaworn, P., Chilton, N.B., Petney, T.N., Klinbunga, S., Satrawaha, R., Webster, J.P., Andrews, R.H., 2009. Impact of temporal changes and host factors on the genetic structure of a population of *Opisthorchis viverrini* sensu lato in Khon Kaen Province (Thailand). Parasitology 136, 1057-1063.

Salzberg, S.L., White, O., Peterson, J., Eisen, J.A., 2001. Microbial genes in the human genome: lateral transfer or gene loss? Science 292, 1903-1906.

Santel, A., Fuller, M.T., 2001. Control of mitochondrial morphology by a human mitofusin. J. Cell Sci. 114, 867-874.

Selkoe, K.A., Toonen, R.J., 2006. Microsatellites for ecologists: a practical guide to using and evaluating microsatellite markers. Ecol. Lett. 9, 615-629.

Shekhovtsov, S.V., Katokhin, A.V., Kolchanov, N.A., Mordvinov, V.A., 2010. The complete mitochondrial genomes of the liver flukes *Opisthorchis felineus* and *Clonorchis sinensis* (Trematoda). Parasitol. Int. 59, 100-103.

Stewart, J.B., Freyer, C., Elson, J.L., Wredenberg, A., Cansu, Z., Trifunovic, A., Larsson, N.G., 2008. Strong purifying selection in transmission of mammalian mitochondrial DNA. PLoS Biol. 6, e10.

Sun, Y.B., Xiong, Z.J., Xiang, X.Y., Liu, S.P., Zhou, W.W., Tu, X.L., Zhong, L., Wang, L., Wu, D.D., Zhang, B.L., Zhu, C.L., Yang, M.M., Chen, H.M., Li, F., Zhou, L., Feng, S.H., Huang, C., Zhang, G.J., Irwin, D., Hillis, D.M., Murphy, R.W., Yang, H.M., Che, J., Wang, J., Zhang, Y.P., 2015. Whole-genome sequence of the Tibetan frog *Nanorana parkeri* and the comparative evolution of tetrapod genomes. Proc. Natl. Acad. Sci. U. S. A. 112, E1257-E1262.

Sunnucks, P., 2000. Efficient genetic markers for population biology. Trends Ecol. Evol. 15, 199-203.

Sutovsky, P., Van Leyen, K., McCauley, T., Day, B.N., Sutovsky, M., 2004. Degradation of paternal mitochondria after fertilization: implications for heteroplasmy, assisted reproductive technologies and mtDNA inheritance. Reproductive Biomedicine Online 8, 24-33.

Tajima, F., 1989. Statistical-method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123, 585-595.

Tatonova, Y.V., Chelomina, G.N., Besprosvannykh, V.V., 2012. Genetic diversity of nuclear ITS1-5.8S-ITS2 rDNA sequence in *Clonorchis sinensis* Cobbold, 1875 (Trematoda: Opisthorchidae) from the Russian Far East. Parasitol. Int. 61, 664-674.

Tatonova, Y.V., Chelomina, G.N., Besprozvannykh, V.V., 2013. Genetic diversity of *Clonorchis sinensis* (Trematoda: Opisthorchiidae) in the Russian southern Far East based on mtDNA *cox*1 sequence variation. Folia Parasitol. (Praha) 60, 155-162.

Thomas, C.G., Wang, W., Jovelin, R., Ghosh, R., Lomasko, T., Trinh, Q., Kruglyak, L., Stein, L.D., Cutter, A.D., 2015. Full-genome evolutionary histories of selfing, splitting, and selection in *Caenorhabditis*. Genome Res. 25, 667-678.

Toews, D.P., Brelsford, A., 2012. The biogeography of mitochondrial and nuclear discordance in animals. Mol. Ecol. 21, 3907-3930.

Treangen, T.J., Salzberg, S.L., 2011. Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nat. Rev. Genet. 13, 36-46.

Tsaousis, A.D., Martin, D.P., Ladoukakis, E.D., Posada, D., Zouros, E., 2005. Widespread recombination in published animal mtDNA sequences. Mol. Biol. Evol. 22, 925-933.

Ujvari, B., Dowton, M., Madsen, T., 2007. Mitochondrial DNA recombination in a free-ranging Australian lizard. Biol. Lett. 3, 189-192.

Vilas, R., Criscione, C.D., Blouin, M.S., 2005. A comparison between mitochondrial DNA and the ribosomal internal transcribed regions in prospecting for cryptic species of platyhelminth parasites. Parasitology 131, 839-846.

Vilas, R., Paniagua, E., Sanmartin, M.L., 2003. Genetic variation within and among infrapopulations of the marine digenetic trematode *Lecithochirium fusiforme*. Parasitology 126, 465-472.

Vinson, J.P., Jaffe, D.B., O'Neill, K., Karlsson, E.K., Stange-Thomann, N., Anderson, S., Mesirov, J.P., Satoh, N., Satou, Y., Nusbaum, C., Birren, B., Galagan, J.E., Lander, E.S., 2005. Assembly of polymorphic genomes: algorithms and application to *Ciona savignyi*. Genome Res. 15, 1127-1135.

Wahlund, S., 1928. Zusammensetzung von Populationen und Korrelationserscheinungen vom Standpunkt der Vererbungslehre aus betrachtet. Hereditas 11, 65-106.

Wang, P., Sha, T., Zhang, Y., Cao, Y., Mi, F., Liu, C., Yang, D., Tang, X., He, X., Dong, J., Wu, J., Yoell, S., Yoell, L., Zhang, K.Q., Zhang, Y., Xu, J., 2017. Frequent heteroplasmy and recombination in the mitochondrial genomes of the basidiomycete mushroom *Thelephora ganbajun*. Sci. Rep. 7, 1626.

Webster, J.P., Shrivastava, J., Johnson, P.J., Blair, L., 2007. Is host-schistosome coevolution going anywhere? BMC Evol. Biol. 7, 91.

White, D.J., Wolff, J.N., Pierson, M., Gemmell, N.J., 2008. Revealing the hidden complexities of mtDNA inheritance. Mol. Ecol. 17, 4925-4942.

Xu, S., Benoff, B., Liou, H.L., Lobel, P., Stock, A.M., 2007. Structural basis of sterol binding by NPC2, a lysosomal protein deficient in Niemann-Pick type C2 disease. J. Biol. Chem. 282, 23525-23531.

Xu, X., Peng, M., Fang, Z., Xu, X., 2000. The direction of microsatellite mutations is dependent upon allele length. Nat. Genet. 24, 396-399.

Xue, M.F., Yang, J., Li, Z.G., Hu, S.N.A., Yao, N., Dean, R.A., Zhao, W.S., Shen, M., Zhang, H.W., Li, C., Liu, L.Y., Cao, L., Xu, X.W., Xing, Y.F., Hsiang, T., Zhang, Z.D., Xu, J.R., Peng, Y.L., 2012. Comparative Analysis of the Genomes of Two Field Isolates of the Rice Blast Fungus *Magnaporthe oryzae*. PLoS Genet. 8, e1002869.

Yaffe, M.P., 1999. The machinery of mitochondrial inheritance and behavior. Science 283, 1493-1497.

Ye, L., Hillier, L.W., Minx, P., Thane, N., Locke, D.P., Martin, J.C., Chen, L., Mitreva, M., Miller, J.R., Haub, K.V., Dooling, D.J., Mardis, E.R., Wilson, R.K., Weinstock, G.M., Warren, W.C., 2011. A vertebrate case study of the quality of assemblies derived from next-generation sequences. Genome Biol. 12, R31.

Yoshino, T.P., Lodes, M.J., Rege, A.A., Chappell, C.L., 1993. Proteinase activity in miracidia, transformation excretory-secretory products, and primary sporocysts of *Schistosoma mansoni*. J. Parasitol. 79, 23-31.

You, Y., Sun, K., Xu, L., Wang, L., Jiang, T., Liu, S., Lu, G., Berquist, S.W., Feng, J., 2010. Pleistocene glacial cycle effects on the phylogeography of the Chinese endemic bat species, *Myotis davidii*. BMC Evol. Biol. 10, 208.

Zadesenets, K.S., Katokhin, A.V., Mordvinov, V.A., Rubtsov, N.B., 2012. Comparative cytogenetics of opisthorchid species (Trematoda, Opisthorchiidae). Parasitol. Int. 61, 87-89.

Zhang, D., Ye, Z., Yamada, K., Zhen, Y., Zheng, C., Bu, W., 2016. Pleistocene sea level fluctuation and host plant habitat requirement influenced the historical phylogeography of the invasive species *Amphiareus obscuriceps* (Hemiptera: Anthocoridae) in its native range. BMC Evol. Biol. 16, 174.

Zhang, D.X., Hewitt, G.M., 2003. Nuclear DNA analyses in genetic studies of populations: practice, problems and prospects. Mol. Ecol. 12, 563-584.

Zhang, H., Yan, J., Zhang, G., Zhou, K., 2008. Phylogeography and demographic history of Chinese black-spotted frog populations (*Pelophylax nigromaculata*): evidence for independent refugia expansion and secondary contact. BMC Evol. Biol. 8, 21.

Zhou, P., Silverstein, K.A., Ramaraj, T., Guhlin, J., Denny, R., Liu, J., Farmer, A.D., Steele, K.P., Stupar, R.M., Miller, J.R., Tiffin, P., Mudge, J., Young, N.D., 2017. Exploring structural variation and gene family architecture with *de novo* assemblies of 15 Medicago genomes. BMC Genomics 18, 261.

LIST OF SUPPLEMENTARY FILES

CHAPTER 3

Supplementary File 3-1 Libraries used in the present study.

Supplementary File 3-2 Transposons predicted in the Korean *Clonorchis sinensis* genome.

Supplementary File 3-3 Annotation summary for genes of the Korean *Clonorchis sinensis* isolate.

Supplementary File 3-4 Korean *Clonorchis sinensis* genes linked to KEGG pathway.

Supplementary File 3-5 Classification of the Korean *Clonorchis sinensis* genes.

Supplementary File 3-6 The secretome of the Korean *Clonorchis sinensis* isolate.

Supplementary File 3-7 Summary of predicted orthologous groups.

Supplementary File 3-8 Synteny results of Korean and Chinese *Clonorchis sinensis* isolates.

Supplementary File 3-9 Summary of nucleotide differences between Korean and Chinese *Clonorchis sinensis* isolates.

Supplementary File 3-10 Functional annotation for selected variable *Clonorchis sinensis* genes.

CHAPTER 4

Supplementary File 4-1 Summary of sampling and read data of 9 *Schistosoma japonicum* isolates in group 1.

Supplementary File 4-2 Summary of SNPs called in each isolate before downsampling.

Supplementary File 4-3 Annotations of 6,978 single-copy gene groups (SCGGs) shared by 16 *Schistosoma japonicum* isolates.

Supplementary File 4-4 Summary of 2,750 SCGGs shared between *Schistosoma japonicum* and *S. mansoni*.

Supplementary File 4-5 Summary of transcription factors under adaptive selection.

Supplementary File 4-6 Summary and annotation of genes with fixed SNPs in SW and CL population of *Schistosoma japonicum*.

Author/s:
Wang, Daxi

Title:
Genomic resources and genetic studies of parasitic flukes, with an emphasis on Clonorchis sinensis

Date:
2019

Persistent Link:
http://hdl.handle.net/11343/225654

File Description:
Thesis