

Copyright is owned by the Author of the thesis. Permission is given for a copy to be downloaded by an individual for the purpose of research and private study only. The thesis may not be reproduced elsewhere without the permission of the Author.

THE PATTERN AND PROCESSES OF GENOME CHANGE IN
ENDOSYMBIONTS OLD AND NEW

A thesis presented in partial fulfilment of the requirement for the degree
of
Doctor of Philosophy
in Evolutionary Biology

Submitted by
Barbara Inge Karoline Schönfeld
2012

Institute of Molecular BioSciences
Massey University
New Zealand

Abstract

Bacterial endosymbionts are an important part of eukaryote evolution as they allow their hosts to exploit bacterial abilities. Plastids, the organelles that enable plants and eukaryotic algae to photosynthesise are ancient cyanobacterial endosymbionts. Since the initial symbiosis ~1.5 billion years ago the majority of their genes has been lost or transferred to their host's nucleus. This process has carried on independently in the different lineages following the diversification of the lineage.

I have compiled a comprehensive data set of fully sequenced plastid genomes to systematically study the frequency of gene transfers from the plastid to the nucleus across the different lineages. Following the reconstruction of the Plantae phylogenetic tree from plastid encoded proteins, gene loss events were reconstructed along its branches. My calculations show that gene losses have occurred at a relative high frequency and in a lineage specific way. This challenges the original idea that gene transfers from the organelle to the nucleus are rare and chance driven events.

Bacteria and eukaryotes continue to form endosymbioses and the study of these relationships produces valuable insights into the early stages of organelle evolution, bacterial metabolic pathways and metabolic regulation. They also allow us a glimpse into the ancient history of eukaryote evolution. For this reason, diatoms that have acquired cyanobacterial endosymbionts with the capability to fix molecular nitrogen were chosen to explore the potential and limitations of high-throughput sequencing technologies for investigating this type of relationship when DNA sequences are obtained from environmental samples and in the presence of bacterial contaminants. The results of this work confirmed the suitability of this relatively new technology to sequence mixed samples but also highlighted i) difficulties in sample preparation which can bias the composition of metagenomic samples obtained, and also ii) the varying suitability of different types of samples used in high-throughput sequencing.

In Gedenken an meinen Vater Günther Schönfeld

Acknowledgements

I would like to thank my supervisors Pete Lockhart and Lesley Collins for their support and guidance. They truly care and I will be forever grateful.

Many others have helped me along the way and should know that it is deeply appreciated:

Trish McLenachan and Phil Novis volunteered their time and expertise to help me in the lab and in the field, and have taught me much. Mike Steel and Tim White worked magic with numbers and scripts to make my results 14% more likely. Bill Martin provided inspiration, support, enthusiasm and helpful criticism when it was needed most. Uwe Maier's keen interest shaped the course of this project. David Penny and the assortment of brilliant people he has gathered around him over the years who have allowed me to share in the passion, the wonder, and the drinking of malt-based beverages, that is science.

The Institute of Molecular BioSciences and the Allan Wilson Centre for Molecular Ecology and Evolution gave the financial and practical support that made my work possible.

My warmest thanks have to go to all the friends and new family who have made my time in New Zealand the best of my life:

The Piripis and Ellicots who made me feel like family. The van Hoves and McComishs for being family. The flatmates, the crafters, the gamers, the foragers, the time travellers and the stargazers, whose friendship means the world to me.

Thank you, Bennet, for being there.

Finally and most importantly I would like to acknowledge my parents, my family and my good friends back home, on whose support I can always count and whose love is always with me.

Table of Contents

Abstract	i
Acknowledgements	iii
List of Figures	ix
List of Tables	xi
1 Introduction	1
1.1 Symbiosis	1
1.1.1 Endosymbiosis	2
1.1.2 Genomic changes in vertically transmitted endosymbionts	3
1.1.3 Endosymbionts and organelles	7
1.1.4 Cyanobacteria	14
1.1.5 Symbiotic nitrogen fixation	15
1.2 Diatoms	17
1.3 Sequencing with the Illumina Genome Analyzer	19
1.4 Data analysis	22
1.5 Aims and purposes	24
Bibliography	26
2 The frequency of gene loss events during plastid evolution	35
2.1 Abstract	35
2.2 Introduction	36
2.3 Methods	42
2.3.1 Gathering and processing of data	42
2.3.2 Data verification	43
2.3.3 Protein alignments	44
2.3.4 Phylogenetic Analyses	45
2.3.5 Mapping of gene loss events	47
2.4 Results	49
2.4.1 Quality of RefSeq data base entries	49
2.4.2 The gene presence/absence matrix	50
2.4.3 The plastid phylogeny	53
2.4.4 The prevalence of plastid gene transfers to the nucleus	55
2.4.5 Lineage specific gene loss and retention	59

2.5	Discussion	61
2.5.1	Quality of RefSeq data base entries	61
2.5.2	The presence/absence matrix	63
2.5.3	The plastid phylogeny	64
2.5.4	The prevalence of plastid gene transfers to the nucleus.....	72
2.5.5	Effects of uneven lineage sampling	78
2.6	Conclusion and outlook.....	80
	Bibliography.....	82
3	Sequencing the genome of the spheroid body of <i>Rhopalodia gibba</i>	93
3.1	Abstract.....	93
3.2	Introduction.....	94
3.2.1	<i>Rhopalodia gibba</i> and its spheroid body.....	94
3.2.2	The cyanobacterium <i>Cyanothece sp.</i> ATCC 51142	95
3.2.3	The spheroid body of <i>Rhopalodia gibba</i>	96
3.3	Methods.....	98
3.3.1	Illumina sequencing.....	98
3.3.2	Assembly.....	98
3.3.3	Contig assembly and sequence comparison.....	99
3.3.4	Mapping and editing.....	100
3.3.5	Identification of bacterial contaminations.....	100
3.4	Results	101
3.4.1	Illumina sequencing.....	101
3.4.2	Assembly.....	101
3.4.3	Identification of misassemblies by mapping.....	103
3.4.4	Assembly quality	104
3.4.5	Identification of putative spheroid body sequences and contaminants	108
3.5	Discussion	111
3.5.1	Sequence assembly.....	111
3.5.2	Assembly quality and comparison to Sanger sequences	117
3.5.3	Contaminant sequences	119
3.6	Conclusion and outlook	121
	Bibliography.....	123
4	High-throughput sequencing of an environmental sample of the diatom <i>Epithemia sorex</i>.....	127

4.1	Abstract	127
4.2	Introduction	128
4.3	Methods	131
4.3.1	Study species	131
4.3.2	Sample Selection.....	132
4.3.3	DNA extraction and Illumina sequencing	136
4.3.4	Mapping.....	137
4.3.5	Assemblies.....	138
4.3.6	BLAST analyses	139
4.3.7	Analyses of eukaryotic signature proteins (ESPs).....	139
4.3.8	Sequencing of 16S and 18S rDNA genes.....	140
4.4	Results	141
4.4.1	DNA extraction and sequencing.....	141
4.4.2	Mapping.....	142
4.4.3	Assemblies.....	145
4.4.4	BLAST results.....	148
4.4.5	Eukaryotic signature proteins	150
4.4.6	16S and 18S rDNA sequencing	153
4.5	Discussion	155
4.5.1	Sample Quality.....	155
4.5.2	Sequence quality and assembly	157
4.5.3	Mapping.....	158
4.5.4	BLAST analyses with MEGAN.....	159
4.5.5	Analysis of Eukaryotic Signature Proteins	162
4.5.6	rDNA sequencing.....	164
4.6	Conclusions	168
	Bibliography	169
5	Conclusion	173
Appendix		

List of Figures

Figure 1-1	Diagram of the principles of Shotgun sequencing.....	19
Figure 1-2	Diagram of the Solexa (Illumina) sequencing process	21
Figure 1-3	Schematic of nodes and arcs of a de Bruijn graph in Velvet.....	23
Figure 2-1	Schematic of the gene presence/absence matrix.....	50
Figure 2-2	Supertree of 124 Plantae species.....	55
Figure 2-3	Gene losses in crown groups.....	58
Figure 2-4	Summary Trees for the three possible root positions.	60
Figure 3-1	<i>Rhopalodia gibba</i>	94
Figure 3-2	The spheroid body of <i>Rhopalodia gibba</i>	96
Figure 3-3	Example of a long sequence assembly for contigs from different Velvet assemblies.....	102
Figure 3-4	Reads mapped to a misassembled repeat region, visualised in Tablet	103
Figure 3-5	Example of a miss-assembly	104
Figure 3-6	Example of a coverage peak of reads derived from a subsample of fosmid inserts	105
Figure 3-7	Mapping of the indexed subset of reads to Contig 2891.....	107
Figure 3-8	MEGAN visualisation of BLAST results for CG-rich contaminant sequences	108
Figure 3-9	MEGAN visualisation of BLAST results for AT-rich contigs.....	109
Figure 3-10	Tablet screenshot of the read coverage of a 52 kb long contig that was Assembled from overlapping sequencing templates	114
Figure 3-11	Alignment of short read assemblies and Sanger sequences	116
Figure 4-1	Silicate shell of <i>Epithemia sorex</i> in valve view.....	131
Figure 4-2	<i>Epithemia sorex</i> sampling site	132
Figure 4-3	<i>Epithemia</i> cells in situ covering the stem of an aquatic plant.....	133
Figure 4-4	Environmental sample after enrichment for diatom cells.....	133
Figure 4-5	Overview of sampling sites on the south island of New Zealand.....	134
Figure 4-6	Sampling sites south of Christchurch.....	135
Figure 4-7	Sampling sites around Arthur's Pass.....	135
Figure 4-8	Environmental sample whole DNA extraction (CTAB extraction protocol)	141
Figure 4-9	Mapping of paired end reads against <i>R. gibba</i> 16S rDNA.....	144

Figure 4-10	Mapping of paired end reads against <i>E. sorex</i> 18S rDNA.....	144
Figure 4-11	Mapping of paired end reads against the sequence of <i>PsbC</i> of <i>R. contorta</i> (chloroplast).....	145
Figure 4-12	Contig size distribution of assembly results	147
Figure 4-13	Results of a BLASTn search against the nt database visualised in MEGAN.....	149
Figure 4-14	Results of a BLASTx search against all diatom sequences in GenBank visualised in MEGAN.....	150
Figure 4-15	Four examples of ESP gene trees	153
Figure 4-16	<i>Epithemia sorex</i> cells <i>in situ</i> and after incubation in RNAlater	156
Figure 4-17	Working principle of the Last Common Ancestor (LCA) algorithm used in MEGAN and the effects of poor database representation	161
Figure 4-17	The araphid, pennate diatom <i>Fragilaria ulna</i>	166

List of Tables

Table 2-1	Summary of results of Gene Loss reconstructions	56
Table 3-1	Parameters and statistics of Velvet assemblies	102
Table 3-2	Statistics of BWA mapping of indexed subset of reads to completely assembled fosmid inserts.....	106
Table 3-3	Average sequence identity between Sanger sequences and short read assemblies by alignment length.....	107
Table 4-1	Primer sequences and expected product sizes	140
Table 4-2	Bowtie2 parameters and statistics for each mapping	143
Table 4-3	Summary of the assembly statistics.....	146
Table 4-4	Average coverage of contigs assembled from untrimmed reads using a k-mer of 25.....	147
Table 4-5	ESP containing contigs and their average coverage	151
Table 4-6	Summary of BLAST results for 16S and 18S rDNA sequences	154

