

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.

Plastid genes across the Great Divide

A thesis presented in partial fulfilment
of the requirements for the degree of

Doctor of Philosophy
in Evolutionary Genetics/Bioinformatics

Institute of Fundamental Sciences
Massey University, Manawatu
New Zealand

Simon J.L. Cox

2014

Abstract

Nearly all life that is visible to the naked eye is arguably a direct consequence of one or both endosymbiotic events that took place early in evolution and eventually resulted in the mitochondrion and the chloroplast. The timing of the mitochondrial endosymbiotic event weighs argument around the nature of LUCA (Last Universal Common Ancestor) being complex or simple and challenge the commonly taught view of bacteria being the first kingdom to emerge from the primordial state.

The ancient metabolic pathways of amino acid and vitamin biosynthesis are examined and Ancestral Sequences constructed in order to discover the endosymbiotic signature within the nucleus of eukaryotes. Cyanobacterial and plant enzymes from these pathways are tracked as they cross from a prokaryotic coding environment to a eukaryotic one. If the eukaryote that took up the chloroplast ancestor was heterotrophic then it probably got some of its co-factors (vitamins) and essential amino acids from its diet. However, in order to become autotrophic it would have to be able to synthesise these amino acids and co-factors directly. The most likely source of these elements would have been the cyanobacterium; therefore cyanobacterial homologs should be found in the nuclear genome of plants.

Ancestral Sequence Reconstruction (ASR) had a negligible effect on uncovering deeper endosymbiotic homologs. However ASR did confirm ancestral convergence between chloroplast and cyanobacterial homologs and between eukaryote nuclear genes and their cyanobacterial counterparts for vitamin and amino acid biosynthetic pathways. The results, all significant, show that the convergence is much stronger between organisms from the same coding environment (prokaryote [chloroplast] versus prokaryote [cyanobacteria]) than from different coding environments (eukaryote [nuclear] versus prokaryote [cyanobacteria]).

Contents

Abstract.....	2
Contents.....	3
Table of Figures.....	6
Tables.....	7
Photographs.....	9
Chapter 1. Literature Reviews	10
Introduction	10
Literature Review - Ancestral Sequence Reconstruction (ASR) Methodology	14
Introduction	14
Construction of Ancestral Sequences	18
Summary	19
Literature Review- Last Eukaryotic Common Ancestor (LECA)	20
The Problem of LECA.....	20
Reconstruction of Ancient Relationships.....	21
Theories on the origin of LECA.....	22
Eocyte theory	23
2 Kingdoms versus 3 Kingdoms.....	24
Another possibility	28
The tree of life and phylogenetic markers from the bacterial perspective.....	32
Viruses - a fourth super kingdom?.....	33
The “When” of LECA	33
Vitamins, Essential Amino Acids and Metabolic pathways.....	34
Summary	37
Chapter 2. Nucleotide Approach	39
Introduction	39
Nucleotide approach.....	39
Archaeplastida	40
Pathways.....	41
KEGG (Kyoto Encyclopaedia of Genes and Genomes)	41
BLAST.....	43
PFAM	43
ASR (Ancestral Sequence Reconstruction).....	44
Method	45

Results.....	46
Discussion.....	47
Enzyme Variants.....	47
KEGG vs. Blastn	49
Markov Models	51
Rationale for the next experiment.....	52
Summary	53
Chapter 3. Amino Acid Approach	55
Background	55
Method	56
Results.....	57
Analysis of Protein Families – 2 samples.	57
Problems, Comments and Potential for further study	63
Possible reasons on why ASR isn't working as efficiently as hoped.	71
Summary	72
Chapter 4. Further Analysis	73
Rationale	73
Methods.....	74
Results.....	80
Discussion.....	83
Chloroplast Transit Peptides.....	83
KEGG Efficacy	83
Truncated Sequences.....	85
AS Nucleotide and Protein Comparison.....	85
Results.....	87
Summary	88
Chapter 5. Ancestral Sequence Reconstruction	89
Introduction.....	93
Results and Discussion	96
Materials and Methods.....	99
Literature Cited	102
Chapter 6. Summary, Conclusions and Future Directions	117
Summary	117
Where next	118

Percentage of bacterial and archaeal genes in eukaryotes	118
The Scientific Process.....	119
Location of Cyanobacterial homologs in biosynthetic pathways	120
Rising enzyme lengths and multiple enzymes	122
References	124
Appendix 1	138
ASR Perl Script-.....	138
Control File-.....	142
Table 1.....	144
Table 2.....	148
Table 3.....	152
Table 4.....	153

Table of Figures

Figure 1-1-illustrating the 3 domains of life as we now know it. While this model (Gogarten et al., 1989) is readily accepted there is dispute about when the archaea and eukaryotes diverged, and about the nature of the out-group. This is discussed at length, along with associated implications later in this thesis.	11
Figure 1-2- Illustrates the depth of AS used to find homologs from Collins <i>et al.</i> , 2003. The letters correspond to reconstructions used to aid in the discovery of more distant homologs for the underlined taxa; branch lengths do not correspond to phylogenetic distance.	16
Figure 1-3- Models for the Origin of Eukaryotes.	24
Figure 1-4- Relationships proposed between eukarya, archaea and bacteria under 2D and 3D scenarios.	25
Figure 1-5- Illustration of eukarya properties being basal. The dashed box illustrates uncertainty about the split of bacteria and archaea from the eukaryotic lineage.	30
Figure 1-6- The unrooted tree (1A) for the three groups' Archaea, Bacteria and Caryotes (eukaryotes).	31
Figure 1-7- The problems of a complex out-group.	32
Figure 1-8- Summarizes key features of LECA problem.	35
Figure 2-1-Overview of methodology.	40
Figure 2-2-Thiamine metabolic pathway; green boxes indicate enzymes that are found in <i>A.thaliana</i> , numbers indicate the EC number of that reaction.	42
Figure 2-3 Illustrating exon rearrangements leading to differing gene lengths. The differing dash-arrangement surrounding the boxes represents different exons in different orders.	49
Figure 2-4 - Lists of archaeplastida cyanobacterial homologs from KEGG for E.C. 3.5.4.25 from the Riboflavin pathway.	50
Figure 2-5-Reliability of Markov modelling with addition of other models.	52
Figure 2-6-From nucleotide to amino acid to domain to motif.	53
Figure 3-1: PFAM result for Maize EC 2.4.2.17.	65
Figure 3-2: PFAM result for Castor Bean EC 2.4.2.17.	65
Figure 3-3: PFAM result for <i>A. thaliana</i> EC 2.4.2.17.	66
Figure 3-4: PFAM result for <i>Cyanidioschyzon merolae</i> EC 2.4.2.17.	66
Figure 3-5: Diagram of protein domain location on <i>A. thaliana</i> EC 2.8.1.7, Pyridoxal motif.	67
Figure 3-6 - Diagram of protein domain location on <i>A. thaliana</i> EC 2.8.1.7, Cys motif.	68
Figure 3-7 - Diagram of protein domain location on <i>Z.mays</i> EC 2.8.1.7, Cys motif.	68
Figure 3-8 - Diagram of protein domain location on <i>Z.mays</i> EC 2.8.1.7, Pyridoxal motif.	69

Figure 3-9 - Diagram of protein domain location on Synechococcus JA-3-3Ab EC 2.8.1.7, Cys motif	69
Figure 3-10 - Diagram of protein domain location on Synechococcus JA-3-3Ab EC 2.8.1.7, Pyridoxal motif	70
Figure 3-11 - Diagram of protein domain location on <i>A. variabilis</i> EC 2.8.1.7, Cys motif.....	70
Figure 3-12 - Diagram of protein domain location on <i>A.variabilis</i> EC 2.8.1.7, Pyridoxal motif	71
Figure 4-1- MSA of 12 variants for 3.1.4.4 AA. AthAT4G11840, circled in red, was chosen for its region of conserved sequence shown as a solid black line. AthAT4G11830 could also have been chosen.....	76
Figure 4-2- Signal output for <i>Medicago truncatula</i> E.C.2.7.8.1 from the ether lipid pathway. The predicted cleavage site is between position 30 and 31.	77
Figure 4-3-Pre-alignment of EC 3.1.1.4 amino acids (AA) (set of 50). The third sequence “bpg” was removed for its excess length.....	77
Figure 4-4- Unedited alignment of 3.1.1.4 AA, set of 50. Note the large gaps.....	78
Figure 4-5- Edited alignment of 3.1.1.4 AA (set of 50).	78
Figure 4-6- the consensus sequence from 3.1.1.4 amino acid (set of 50) alignment. Green indicates 100% agreement, brown between 100%- 30% and red below 30%. The brown area between 60 and 600 indicates the more conserved portion of this enzyme.....	78
Figure 4-7- N-J tree for 3.1.1.4 AA (set of 10).	79
Figure 4-8: Graphic Blastp result of E.C. 1.1.1.3	86
Figure 5-1- The trees for two subgroups X and Y are illustrated along with the ancestral sequences ax and ay, from (White et al., 2013). Although the subgroups X and Y are based initially on ‘prior knowledge’, they are tested objectively on the data as used.....	91
Figure 6-1- Illustrating the decrease in the probability of finding a correct result at deeper times when using HMM to model evolutionary mutation rates. The "senility zone” represents time periods when HMM no longer works effectively.	120

Tables

Table 1-1-List of attributes thought to be properties of the Last Eukaryotic Common Ancestor [LECA] from (Poole and Neumann, 2011).	11
Table 1-2- List of taxa and associated evolutionary reconstructions from Wolf <i>et al.</i> , 2013.....	17
Table 1-3- Summary of seven recent large scale phylogenetic analyses, based on Gribaldo <i>et al</i> , 2010. (ML = Maximum Likelihood, MP = Maximum Parsimony).	26

Table 1-4- Summarises the problems with theories about Eukaryotic origins and the number of Kingdoms; adapted from Poole & Penny (2007) and Gribaldo <i>et al.</i> , (2010).....	28
Table 2-1-List of archaeplastida used in this study.....	41
Table 2-2-The 18 amino acid and vitamin metabolic pathways used in this study.....	41
Table 2-3-Tables listing both the number of EC 1.2.1.3 variants per organism (left-hand side) and the variety of pathways EC 1.2.1.3 is involved in (right-hand side).	48
Table 2-4-List of AS archaeplastida homologs from BLAST; no cyanobacterial homologs were found in the first 30 bacterial homologs.....	51
Table 3-1-List of cyanobacterial species	56
Table 3-2: Analysis of Protein Family E-scores in the Isoleucine and Niacin Pathways compared to the Ancestral Sequence E-scores	58
Table 3-3- Comparison of BLAST results to Phyla for the Ancestral Sequence and <i>A. thaliana</i>	59
Table 3-4: Result for Enzyme LeuB- a Simple Similarity Table constructed from a MUSCLE Alignment and expressed as a percentage. A score of 100% would indicate that the two sequences were identical.	60
Table 3-5: Result for Enzyme 3.5.4.26- a Simple Similarity Table constructed from a MUSCLE Alignment and expressed as a percentage.....	61
Table 3-6: Summary of Simple Similarity results for all enzymes.....	62
Table 3-7: EC 4.2.1.35- multiple isozymes, variants and PFAM protein domains that occur in 13 archaeplastida-enzyme lengths are expressed in the main body of the Table. For example <i>A.thaliana</i> has four different Aconitase enzymes that range in length from 222-509 amino acids (aa).....	64
Table 3-8- Summary of different lengths and identified protein families for organisms coding for EC 2.4.2.17. The average length for the 13 archaeplastida is 380 aa; these four organisms are the exception that may be stymieing the multiple sequence alignment.	67
Table 3-9- <i>A.thaliana</i> EC 2.8.1.7 Pyridoxal Motif (NFS1) detailing location and E-score of protein domains.	67
Table 3-10- <i>A.thaliana</i> EC 2.8.1.7 Cys Motif (CPNIFS) detailing location and E-score of protein domains	68
Table 3-11- <i>Zea mays</i> EC 2.8.1.7 Cys Motif (CPNIFS) detailing location and E-score of protein domains	68
Table 3-12- <i>Zea mays</i> EC 2.8.1.7 Pyridoxal Motif (NFS1) detailing location and E-score of protein domains	69

Table 3-13- <i>Synechococcus JA-3-3Ab</i> EC 2.8.1.7 Cys Motif (CPNIFS) detailing location and E-score of protein domains.....	69
Table 3-14- <i>Synechococcus JA-3-3Ab</i> EC 2.8.1.7 Pyridoxal Motif (NFS1) detailing location and E-score of protein domains.....	70
Table 3-15- <i>A. variabilis</i> EC 2.8.1.7 Cys Motif (CPNIFS) detailing location and E-score of protein domains	70
Table 3-16- <i>A.variabilis</i> EC 2.8.1.7 Pyridoxal Motif (NFS1) detailing location and E-score of protein domains	71
Table 4-1-Organisms used in this study from the KEGG database.	74
Table 4-2-the 10 enzymes from the Lysine and Ether Lipid pathways	75
Table 4-3- Comparison of Ancestral Sequences (expressed as nucleotides) and <i>A. thaliana</i> blasted against bacteria.	80
Table 4-4-Comparison of Ancestral Sequences (expressed as amino acids) and <i>A. thaliana</i> blasted against bacteria.	81
Table 4-5: cTP results from the Ether Lipid enzyme 2.7.8.1	82
Table 4-6: Comparison of cTP-trimmed AS with untrimmed sequences.....	82
Table 4-7: Comparison of KEGG and UniProt databases for E.C. 1.1.1.3 found in the Lysine biosynthetic pathway.	84
Table 4-8: Comparison of KEGG and UniProt databases for E.C. 4.3.3.7 found in the Lysine biosynthetic pathway	85
Table 4-9: Comparison of nucleotide and protein ancestral sequences from the Lysine pathway.	86
Table 5-1-List of all core chloroplast genes present in essentially all chloroplast genomes from photosynthetic organisms. Adapted from Barbrook et al, (2010).	91
Table 6-1- Position of cyanobacterial homologs in the biosynthetic pathways analysed for this thesis	122

Photographs

Photo 1-1- Illustrating <i>Gemmata obscuriglobus</i> and the three layers surrounding its nucleoid; from Fuerst, 2005.	29
-----------------------------------------------------------------------------------------------------------------------------	----