Lawrence Berkeley National Laboratory

Title
The first complete chloroplast genome sequence of a lycophyte, Huperzia lucidula (Lycopodiaceae)

Permalink
https://escholarship.org/uc/item/3sd2g111

Authors
Wolf, Paul G.
Karol, Kenneth G.
Mandoli, Dina F.
et al.

Publication Date
2005-02-01

Peer reviewed

The first complete chloroplast genome sequence of a lycophyte,

Huperzia lucidula (Lycopodiaceae)

Keywords: organelle, evolution, gene order

Paul G. Wolf [1], Kenneth G. Karol[2], Dina F. Mandoli[2,3], Jennifer Kuehl[4], K. Arumuganathan[5], Mark W. Ellis[1], Brent D. Mishler[6,7], Dean G. Kelch[6,7], Richard G. Olmstead[2], Jeffrey L. Boore[4,6].

[1]Department of Biology, Utah State University, Logan, UT 84322-5305.

[2]Department of Biology, University of Washington, Seattle, 98195-5325.

[3]Center for Developmental Biology, University of Washington,. Seattle, 98195-5325.

[4]DOE Joint Genome Institute and Lawrence Berkeley National Laboratory, Walnut Creek, CA 94598. [5]Benaroya Research Institute at Virginia Mason, Seattle, WA 98101.

[6]Department of Integrative Biology, University of California, Berkeley, CA 94720.

[7]University Herbarium, Jepson Herbarium, University of California, Berkeley, CA 94720.

Abbreviations: LSC – large single-copy region; SSC – small single-copy region; IR – inverted repeat; RCA – rolling circle amplification; FACS - fluorescence activated cell sorting; BS – bootstrap support; Γ – gamma-distributed rates; ML – maximum likelihood; MP – maximum parsimony

Address for Correspondence: Paul G. Wolf, Department of Biology, Utah State University, 5305

Old Main Hill, Logan, UT 84322, USA. Phone (435) 797 4034; FAX (435) 797 1575; Email

wolf@biology.usu.edu.

Abstract

We used a unique combination of techniques to sequence the first complete chloroplast genome of a lycophyte, *Huperzia lucidula.* This plant belongs to a significant clade hypothesized to represent the sister group to all other vascular plants. We used fluorescence-activated cell sorting (FACS) to isolate the organelles, rolling circle amplification (RCA) to amplify the genome, and shotgun sequencing to 8x depth coverage to obtain the complete chloroplast genome sequence. The genome is 154,373 bp, containing inverted repeats of 15,314 bp each, a large single-copy region of 104,088 bp, and a small single-copy region of 19,671 bp. Gene order is more similar to those of mosses, liverworts, and hornworts than to gene order for other vascular plants. For example, the *Huperzia* chloroplast genome possesses the bryophyte gene order for a previously characterized 30 kb inversion, thus supporting the hypothesis that lycophytes are sister to all other extant vascular plants. The lycophyte chloroplast genome data also enable a better reconstruction of the basal tracheophyte genome, which is useful for inferring relationships among bryophyte lineages. Several unique characters are observed in *Huperzia*, such as movement of the gene *ndhF* from the small single copy region into the inverted repeat. We present several analyses of evolutionary relationships among land plants by using nucleotide data, amino acid sequences, and by comparing gene arrangements from chloroplast genomes. The results, while still tentative pending the large number of chloroplast genomes from other key lineages that are soon to be sequenced, are intriguing in themselves, and contribute to a growing comparative database of genomic and morphological data across the green plants.

1. Introduction

Green plants are an old group dating back about 1 billion years (Mishler, 2000). There are about half a million extant species (Mishler, 2000), including the main primary energy producers in both terrestrial and aquatic ecosystems. Reconstructing the pattern and processes of the evolution of this large and diverse group is imperative, yet challenging. Arguably, the fastest growing front in these efforts is the rapid growth in genome sequencing, which has ignited the fields of comparative and evolutionary genomics (Normile, 2001). Although large scale phylogenetic analyses of complete eukaryotic nuclear genomes are just beginning, many phylogenomic studies of the much smaller organellar genomes are complete or underway. Most of this work has been on animal mitochondrial genomes (Boore, 1999), of which over 400 species are currently represented in public databases. More recently, chloroplast genomes have been sequenced from several clades of green plants and these genomes have been found to contain considerable amounts of phylogenetically useful data (Lemieux et al., 2000).

The chloroplasts of green plants are descendents of cyanobacteria that established an endosymbiotic relationship with a primitive eukaryote. Although many proteins necessary for chloroplast functioning are imported from the cytoplasm, chloroplasts have retained their own, now diminished, genome (Stoebe et al., 1999), along with systems for expressing these genes. Across green plants, there is a high degree of consistency in chloroplast genome structure and in gene content and arrangement (Palmer and Stein, 1986). However, these features vary sufficiently among lineages to provide useful characters for phylogenetic reconstruction. Such genome-level characters have proven to be especially robust indicators of evolutionary relatedness due to their complexity and low frequency of reversal (Helfenbein and Boore, 2004).

Comparing complete chloroplast genome sequences also enables a reconstruction of events, such as gene transfers between intracellular compartments (i.e., nucleus, chloroplast, mitochondrion), and a better understanding of the evolutionary processes that account for the features of today's chloroplast genomes. Unfortunately, as of the beginning of 2004, there are still only 25 complete chloroplast genomes published and many critical clades remain unrepresented. Here we describe the first of a series of complete chloroplast genome sequences selected to fill important phylogenetic gaps, initially focusing on land plants. Currently, complete chloroplast genomes are available from each of the three main bryophyte lineages (a hornwort, a moss, and a liverwort), two ferns, two gymnosperms, and 13 angiosperms. These taxa represent the bulk of phylogenetic diversity, but no chloroplast genome sequence has been published for any lycophyte. This is somewhat surprising because the best evidence that the lycophytes are sister to remaining extant vascular plants comes from the observation of a 30 kb inversion in the chloroplast genome, detected by restriction-site mapping studies (Raubeson and Jansen, 1992). Here we describe 1, the first complete chloroplast genome sequence of a lycophyte (*Huperzia lucidula* (Michx.) Trevis.); 2, a novel method of providing chloroplast genome-enhanced material from which to obtain the sequence; and 3, the unique aspects of the genome. We also present phylogenetic analyses based on amino acid sequences and DNA sequences extracted from published land plant chloroplast genomes plus that of *H. lucidula*. Furthermore, we explore the use of genome structure to infer land plant phylogeny.

2. Materials and Methods

2.1. Preparation and DNA sequencing

Vegetative material of *H. lucidula* was collected from Balsam Gap Overlook , NC (USA). A voucher specimen (Renzaglia #3200) is deposited at the University of California Herbarium at Berkeley (UC). Purified fractions of intact chloroplasts of *H. lucidula* were collected by fluorescence-activated cell sorting (FACS). One hundred milligrams of fresh leaf tissue was placed on ice in a sterile plastic Petri dish containing 1.0 mL of an organelle isolation solution containing 0.33 M sorbitol, 50 mM HEPES at pH 7.6, 2 mM EDTA, 1 mM MgCl$_2$, 0.1% BSA, 1% PVP-40, and 5 mM β-Mercaptoethanol, and the tissue was sliced into 0.25 - 1 mm segments. Suspended organelles (chloroplasts, mitochondria, and nuclei) were withdrawn using a pipette, filtered through 30 μm nylon mesh, and stained with 2 μg/mL DAPI (Sigma-Aldrich, St. Louis, MO, USA) and 100 nM Mitotracker Green (Molecular Probes Inc., Eugene, OR, USA). The organelle suspension was incubated on ice for 15 min, then analyzed on a FACS DiVa using sterile phosphate buffered solution (Invitrogen Inc., Carlsbad, CA, USA) as sheath fluid. We used a Coherent INNOVA Enterprise Ion laser (Coherent, Inc., Santa Paula, CA, USA) emitting a 488 nm beam at 275 mW to excite chlorophyll and Mitotracker Green, and a UV beam at 30 mW to excite DAPI. Red fluorescence from chlorophyll was passed through 675±20 nm filter, held within the FL3 photomultiplier tube (PMT), and green fluorescence from Mitotracker Green was passed through a 530±30 nm filter held within the FL1 PMT. DAPI fluorescence from DNA was passed through a 424±44 nm pass filter held within the FL4 PMT. Organelles were collected into separate sterile 15 ml centrifuge tubes by flow cytometric sorting based on the respective sorting gates (Figure 1). Sorted organelles were pelleted and shipped frozen for DNA isolation and amplification.

The DNA preparation was then processed for sequencing by the Production Genomics Facility of the DOE Joint Genome Institute. Template was first amplified through rolling circle

amplification (RCA) with random hexamers (Dean et al., 2001). The DNA was then mechanically sheared into random fragments of about 3 kb by repeated passage through a narrow aperture using a Hydroshear device (Genemachines, San Carlos, CA, USA). These fragments were then enzymatically repaired to ensure blunt ends, purified by gel electrophoresis to select for a narrow distribution of fragment sizes, ligated into dephosphorylated pUC18 vector, and transformed into *E. coli* to create plasmid libraries. Automated colony pickers were used to select and transfer colonies into 384-well plates containing LB media and glycerol. After overnight incubation, a small portion was processed robotically through RCA of plasmids (Dean et al., 2001), then used as a template for DNA sequencing using Big-Dye chemistry (Applied Biosystems, Foster City, CA, USA). Sequencing reactions were cleaned using SPRI (Elkin et al., 2002), and separated electrophoretically on ABI 3730XL or Megabace 4000 automated DNA sequencing machines to produce a sequencing read from each end of each plasmid.

2.2. Assembly and annotation

Sequences were processed using Phred (Ewing and Green, 1998), trimmed for quality, screened for vector sequences, and assembled using Phrap. Quality scores were assigned automatically, and the electropherograms and assembly were viewed and verified for accuracy using Consed 12 (Gordon et al., 1998). As is typical, manual input was required to reconstruct part of one of the inverted repeat (IR) regions, since automated assembly methods cannot recognize these as different. Regions of low quality or inadequate coverage were reamplified with PCR and sequenced. The final assembly has an average depth of coverage of 8X. We assembled the sequence as a circular genome with two copies of the IR. Nucleotide numbering followed previously published chloroplast genomes by starting the genome at the beginning of

the LSC. We annotated the genome using DOGMA (Dual Organellar GenoMe Annotator), available on the web at http://phylocluster.biosci.utexas.edu/dogma/. Genes were located by using a database of previously published chloroplast genomes from which Blast searches (Altschul et al., 1997) are used to find approximate gene positions. From this initial annotation, we located hypothetical starts, stops, and intron positions based on comparisons to homologous genes in other chloroplast genomes and by considering the possibility of RNA editing, which can modify the start and stop positions.

2.3. Phylogenetic analyses – DNA and protein sequences

Seventy three protein-coding sequences were extracted from annotated chloroplast DNA genomes found in GenBank (www.ncbi.nlm.nih.gov). Because RNA editing is abundant in *Anthoceros* and *Adiantum*, cDNA sequences were used in lieu of DNA sequences. These data combined with sequences from *Huperzia* (this study) represent nineteen land plants and a single charophyte green alga (Table 1). Although additional chloroplast genome sequences are published we excluded ones that would not provide useful phylogenetic representation for a focus on land plants. Thus, we did not include two representatives of any one species, such as rice. Data sets from the 73 genes are hereafter referred to as data sets 73. Individual gene alignments were constructed using MacClade v4.0b6 (Maddison and Maddison, 2003) and assembled into a single data set. From this concatenated alignment, three data sets were generated for phylogenetic analyses: (1) nucleotide sequence data excluding unalignable regions, stop codons, and overlapping regions of *atpB*/*atpE* and *psbD*/*psbC*; (2) nucleotide sequence data described above, excluding third-codon positions; and (3) translated amino acid data excluding unalignable regions and stop codons. These data sets included 48,201 nucleotide sites, 32,135

first and second nucleotide sites, and 16,084 amino acid sites, respectively. Three additional data sets were constructed that included only genes found in all genomes (58 protein coding sequences; hereafter referred to as data sets 58). These reduced data sets included 35,571 nucleotide sites, 23,715 first and second nucleotide sites, and 11,855 amino acid sites, respectively. All data sets are available online as supplementary material (URL goes here).

Maximum likelihood (ML) and maximum parsimony (MP) analyses of the nucleotide sequence data were performed with PAUP* 4.0b10 (Swofford, 2003). Amino acid data were analyzed under MP with PAUP* and under ML with Phylip 3.6 (Felsenstein, 2004). Model selection for nucleotide data (Swofford et al., 1996) yielded the general-time-reversible model with invariable (I) sites (Hasegawa et al., 1993) and gamma-distributed (Γ) rates for variable sites as the best-fitting model. The JTT amino acid substitution model (Jones et al., 1992) was used for ML amino acid analyses along with I+Γ. Two-hundred ML bootstrap replicates and 1000 MP bootstrap replicates were performed for each data set.

Most maximum likelihood models make assumptions about equilibrium of base composition across lineages, violations of which can lead to erroneous phylogenetic inferences (Lockhart et al., 1994). We tested for compositional equilibrium using TREEPUZZLE (Strimmer and Haeseler, 1996) and we found that all taxa in our analysis failed the 5% chi-square test. Rather than proceeding with zero taxa we performed the LogDet implementation in PAUP*, which uses a transformation that is more consistent under asymmetric models of substitution (Lockhart et al., 1994). We implemented LogDet for both data sets, in each case with all codon positions and with third codons removed.

2.4. Phylogenetic analyses – Genome structure

Genomic character coding and analyses followed Kelch et al. (2004). We examined the same genomes as in the analyses of sequence data (Table 1). We used published annotations to examine gene presence and order in all selected genomes, with particular attention to regions of putative inversions. Large inverted sections of gene sequences were analyzed in reverse order to facilitate identification of additional gene rearrangements within the inverted region. Characters comprised three types: gene rearrangements representing inversions of two or more genes, gene presence/absence representing the loss or gain of a gene, and intron presence/absence representing the presence of a particular intron within chloroplast genes. Duplications of genes via inclusion in the inverted repeat (IR) were treated with gene rearrangement characters. We searched for gene order characters using basic principles of character analysis originally developed for morphological characters. Coding of inversions was binary and chosen to minimize the number of inversion characters. In addition, copies of genes or pseudogenes were coded as present or absent based on synteny. We detected 42 characters, of which 29 were potentially informative (Table 2). These characters were then coded as binary for each genome (Table 3). Phylogenetic analyses were performed using PAUP* 4.0b10 (Swofford, 2003) using MP as the optimality criterion. The matrix was analyzed using the branch-and-bound algorithm with the furthest addition sequence setting. The resulting trees were rooted using the charophyte *Chaetosphaeridium* as the outgroup. A bootstrap analysis was performed using a 1000 replicates of heuristic searches employing stepwise addition and TBR branch swapping.

3. RESULTS

Our overarching goal is to resolve the phylogeny of green plants using a wide range of data including sequences of organellar genomes (http://ucjeps.berkeley.edu/TreeofLife/). Many

of the taxa of interest are rare or have small or unicellular body plans, so traditional methods of organelle isolation, such as sucrose gradients, are not feasible because of tissue quality or quantity. However, PCR-based methods and cloning mean that even a small amount of DNA would suffice. We used a taxon, *H. lucidula*, for which tissue was abundant, to develop this methodology.

To isolate and clone the *Huperzia* chloroplast genome, we coupled FACS with RCA. As the name indicates, cells are the normal target of FACS. Organelles are at or near the size limits of state-of-the-art FACS equipment. To ensure that we were purifying the spherical chloroplasts of *H. lucidula* away from other organelles of similar size and shape, and to design sorting gates, we FACS-analyzed three organelle types (chloroplasts, nuclei, and mitochondria) from each preparation. We then simultaneously sorted putative chloroplast and putative mitochondrial fractions from each tissue preparation (Figure 1). The success of FACS and of RCA, as well as the interface between the two methodologies, was each affected by several variables, one of which was the taxon itself. We will fully describe the details and utility of FACS-RCA for several taxa in a subsequent publication. Here, the success of our marriage of FACS and RCA was clearly demonstrated by our results: a shotgun library made from the chloroplast fraction provided 2,304 clones for 4,608 sequence reads, of which 2,627 (57%) assembled into an apparent chloroplast genome.

The genome is 154,373 bp, with IRs of 15,314 bp each, an LSC region of 104,088 bp, and a small single-copy (SSC) region of 19,657 bp (Figure 2). The sequence and annotation is deposited in GenBank as Accession number **AY660566**. In addition to the fully assembled circular genome, we detected a contig of 5086 bp (GenBank accession number **AY675586**) that falsely assembled at position 111,542 and 146,920 in the IRs. This extra sequence contains

mostly repetitive DNA and we hypothesize that it is part of the nuclear genome that is adjacent

to a piece of chloroplast DNA that has recently been transferred to the nucleus. Such a transfer of

chloroplast DNA to the nuclear genome has been documented in rice (Shahmuradov et al.,

2003). Due to the repetitive nature of this putative nuclear DNA, we hypothesize that it is a false

assembly representing random scattered repetitive elements. During annotation, we located the

repertoire of genes that is typical of land plant chloroplast genomes (Figure 2). We found a few

genes with unusual features: lack of expected stop codons in *ndhJ, atpI, chlL ndhH,* and *ccsA,*

and two internal stop codons in *rps16*. We hypothesize that these are RNA editing sites but we

note that this implies considerably lower levels of RNA editing in the lycophyte chloroplast

genome than has been found in a fern (Wolf et al., 2004) or a hornwort (Kugita et al., 2003).

The overall organization of the *Huperzia* chloroplast genome is more typical of a

bryophyte than of other vascular plants. Gene order within the LSC is almost identical to that of

Anthoceros. We also detected several unique features of the genome, including placement of

ndhF into the IR. This gene actually spans the IR and SSC so that the copy in IR$_B$ is missing the

start; therefore we consider that copy a pseudogene.

3.1. Phylogenetic analyses – Sequence data

Figure 3 shows phylogenetic relationships of representative lineages of land plants

inferred using ML and MP of 73 protein-coding genes from 20 chloroplast genomes (Table 1)

without the LogDet transformation. Table 4 summarizes bootstrap values for all our analyses so

that the effects of gene inclusion, codon position, and analysis can be compared for several key

nodes. Our maximum-likelihood (ML) tree using 48,201 nucleotide sites (all codon positions) is

shown (-ln=388577.11). The ML analysis excluding 3rd codon positions yielded a similar

topology (-ln=207477.06) except that *Arabidopsis* was sister to *Oenothera* (Bootstrap

value[BS]=66%) and *Oryza* was sister to *Zea* (BS=58%). Maximum parsimony (MP) results

differed from the ML topology shown in two ways: Both MP nucleotide analyses (with and

without third codon positions) (1) placed *Huperzia* sister to the seed plants to the exclusion of the

ferns (all codon positions, BS=60%; excluding 3rd codon positions, BS=75%) %), and (2)

supported a monophyletic dicot clade with *Amborella* and *Calycanthus* sister to the remaining

dicots (all codon positions, BS=97%; excluding 3rd codon positions, BS=100%). The ML

analysis excluding 3rd codon positions yielded a similar topology (-ln=207477.05846) with the

exception of *Arabidopsis* sister to *Oenothera* (BS=66%) and *Oryza* sister to *Zea* (BS=58%). The

ML analysis using inferred amino acid data (-ln= 180033.90946) and all three MP analyses

differed in the placement of the monocots sister to a monophyletic dicot clade including

Amborella and *Calycanthus* (ML amino acid; 98%; MP all codon positions, BS=97%; MP

excluding 3rd codon positions, BS=100%; MP amino acid, BS=100%). Relationships among

angiosperms are not discussed in much further detail here because several recent papers have

addressed this area of the tree in more depth (Goremykin et al., 2004; Stefanovic et al., 2004).

Both MP nucleotide analyses placed *Huperzia* sister to the seed plants to the exclusion of the

ferns (all codon positions, BS=60%; excluding 3rd codon positions, BS=75%). In general,

bootstrap support at the base of the land plants decreased with removal of third positions and

decreased further with amino acid sequences only. This finding is similar to that of Nishiyama et

al. (2004).

The LogDet analysis resulted in similar trees to that of Figure 3 with a few notable

exceptions. All LogDet trees have monocots (represented only by three grasses) at the base of the

angiosperms, i.e., as a sister to "dicots". With third codon positions removed the three

bryophytes form a monophyletic group (BS=100 for data set 58; BS = 72 for data set 73). The

bryophytes are then sister to tracheophytes (as expected) with Huperzia sister to all other

tracheophytes. With all codon positions Physcomitrella (moss) and Marchantia (liverwort) are

sister to each other at the base of land plants with BS = 100 in both data sets. With data set 73

Huperzia and Anthoceros are sisters at the base of tracheophytes (BS=99), whereas with data set

58 the more acceptable topology has Anthoceros as sister to all tracheophytes (BS=50).

Removal of genes that were not present in all analyzed taxa had a relatively small effect

on tree topology and branch support. Table 4 compares all analyses for several critical nodes; we

focus mainly at the base of the land plants and tracheophytes. The only noticeable difference is

that the LogDet analysis with third codons removed provided higher bootstrap support for

bryophyte monophyly and for a clade of moss plus liverwort than when third codons were

included, as discussed above.

3.2. Phylogenetic analysis – Genome structure

We scored 42 characters based on differences in genome structure among the 20

chloroplast genomes examined (Table 1). Twenty nine of these characters were parsimony-

informative. Of the latter, 13 were rearrangements, 12 were gene presence/absence characters,

and 4 were intron characters. Phylogenetic analysis of these genome structure data produced 900

equally parsimonious trees (CI = 0.7, RI=0.82), of which the strict consensus is shown in Figure

4. We found strong support for a monophyletic *Pinus* (BS=98%), and grasses (BS=98%).

Huperzia was placed among the bryophyte lineages, with *Anthoceros* appearing as sister to the

euphyllophytes (= vascular plants minus lycophytes). The clade *Anthoceros* plus vascular plants

(= *Huperzia* plus euphyllophytes) was supported by the inclusion of *rps12* in the IR$_B$ region

(Figure 5, Table 2, char. 27), with a putative loss of this gene from IR_B in grasses. The clade of *Anthoceros* plus euphyllophytes was supported by an expansion of IR_B to include *rps*7 and *ndh*B (Figure 5, Table 2, char. 1). The main difference between the tree based on genome structure and that based on sequence data was that the relative positions of *Huperzia* and *Anthoceros* were reversed.

4. DISCUSSION

We first focus on our, albeit limited, phylogenetic inferences before discussing the issues associated with details of data and analysis. We present several different analyses of land plant phylogeny using both sequence data (DNA and protein) and data from genome structure. This has been made possible by the addition of a new chloroplast genome sequence from a previously unsampled clade of land plants. Any phylogenetic inference depends very much on the type of analysis used. We have performed similar analyses to those done by Nishiyama et al. (2004), using a similar (but not identical) data set, in addition to adding data from the lycophyte clade represented by *Huperzia*. Adding this taxon did not affect the recovery of a monophyletic bryophytes also detected by Nishiyama et al. (2004) in the LogDet analysis. Although the bootstrap support for this clade is high, inferring such a relationship based on only one taxon from each of the three main bryophyte clades is probably premature. Data from several critical clades within the bryophytes are needed to stabilize inferences at the base of land plants (Nishiyama et al., 2004).

The result that *Huperzia* is sister to *Anthoceros* plus euphyllophytes (Hup(Anth + Euph)) in the genome structure analysis is in conflict with our own sequence-based analysis; we did not recover this topology in any of our analyses. This topology is also inconsistent with all published

studies to date, whether based on DNA sequence data (Dombrovska and Qiu, 2004) or morphological characters (Kenrick and Crane, 1997). If the (Hup(Anth + Euph)) topology were true, it would mean that vascular plants are polyphyletic, with lycophytes and euphyllophytes having arisen separately from different bryophyte ancestors; this is an unlikely scenario. However, the (Anth + Euph) clade is supported by only one character in the genome structure analysis: the expansion of the LSC margin of the IR to include the genes *rps*7 and *ndh*B. This rather tenuous support is reflected in the low bootstrap value for this clade (Fig. 4). In addition, the documentation of several expansions and contractions (Goulding et al., 1996) suggests that the LSC margin of the IR is relatively unstable, and therefore convergence for genome structure characters in this region is likely. Such size fluctuations in the IR region could be caused by gene conversion (Goulding et al., 1996). Assuming that basal branches have retained the ancestral conditions, the trend has been generally toward an increase in size of the IR (e.g., from *Physcomitrella* to *Anthoceros* and from *Anthoceros* to *Psilotum* and *Adiantum*). However, superimposed on this trend are some subsequent gene copy losses (Kelch et al., 2004). Therefore, it seems that some genomic character types (e.g., inclusion or loss of genes from the IR of the chloroplast) may be subject to significant amounts of homoplasy, whereas other genomic characters (e.g., large inversions or gene gains and losses) may be highly resistant to homoplasy. Further sampling of relevant taxa will allow the relative homoplasy of different types of evolutionary changes to be evaluated.

The most significant aspect of gross structure of the *Huperzia* chloroplast genome may lie in its general similarity to bryophyte lineages rather than to its similarity to other vascular plants. Euphyllophytes share a large (> 20 gene, ~30 kb) inversion in the LSC relative to the bryophyte lineages and to *Huperzia*, a rearrangement detected previously with restriction site

mapping (Raubeson and Jansen, 1992). If such genomic rearrangements are relatively rare

events, then this result is in agreement with the fossil record, which indicates that lycophytes

diverged from euphyllophytes very early in vascular plant evolution (Kenrick and Crane, 1997).

Therefore, the branch leading to the common ancestor of lycophytes and euphyllophytes was

possibly short and less likely to accumulate rare genomic rearrangements than the longer branch

leading to the most recent common ancestor of the living euphyllophytes.

Currently, relationships among the three bryophyte clades and vascular plants are

unresolved; at least three alternative topologies have been proposed recently (Nishiyama et al.,

2004). We predict that this uncertainty will continue until taxon sampling improves. The

situation is somewhat different in angiosperms, where more robust phylogenetic hypotheses are

available based on several large data sets and with broad sampling among potential clades.

Amborella, along with two other dicotyledonous lineages, have been found to be sister group(s)

to all other angiosperms in studies which examine smaller data sets (but still several genes) with

larger taxon sampling (Barkman et al., 2000). Across these studies, such topologies have been

inferred using 28 different genes (Stefanovic et al., 2004). In contrast, Goremykin and coworkers

(Goremykin et al., 2004) have inferred monocots as the sister to all dicots in most (but not all) of

their analyses.   These latter studies used concatenated alignments from complete chloroplast

genome sequences (similar to our data sets), but with sparse taxon sampling especially in the

monocots, which were represented only by grasses. More recently, taxon sampling of a large

chloroplast gene data set was improved with the inclusion of the monocot *Acorus* (Stefanovic et

al., 2004). Analysis of this expanded data set retrieved the *Amborella*-basal topology. Our

sequence-based analysis found a similar instability at the base of the angiosperms; as in the

studies cited above, the topology depended much on the nature of sequences (nucleotide versus

amino acid) and the analytical approach (ML versus MP). In our examination of genome

structure we find that *Amborella* has a rather general angiosperm chloroplast organization with

no putative plesiomorphies and few autapomorphies, making it difficult to infer relationships

with confidence. However, in our analysis, only the grasses are represented at the family level by

multiple taxa (3), for which we detected five shared unique genomic characters. Therefore, it is

likely that further sampling within angiosperms will reveal genomic characters that are

phylogenetically informative at the level of family or genus.

  Clearly, robust phylogenetic analyses require both adequate taxon and character

sampling. Our coding of genome structure characters did not rely on structural rearrangements

that are unique to one taxon and therefore are uninformative phylogenetically. However, many of

these characters may become shared derived (informative) characters as we accumulate data

from more plants. In addition to such large-scale structural rearrangements, scanning genomes

for small insertions or deletions (indels) is likely to provide phylogenetically informative data

(Graham et al., 2000). Furthermore, several small inversions (~40bp - ~110 bp) have been

detected in chloroplast genomes (Kim and Lee, 2004). These rearrangements are probably too

small to be easily detected by restriction site mapping, but they could be detected by mining

genome sequence databases. Utility of such rearrangements will depend on adequate taxon

sampling. We predict that genome structure patterns will prove useful for inferring phylogeny at

several levels in the green tree of life and enable us to resolve further plant phylogeny and

improve our understanding of genome evolution.

  We explored several aspects of data analysis in this study. One of the primary issues with

concatenated data sets extracted from genome sequences is the fact that different genomes

usually contain different sets of genes. One option is to include only genes available for all taxa.

In some cases the extensive loss of data might make this option more easily attained by simply sequencing those genes from the taxa required. At the other extreme, all genes can be included, which may result in large gaps of missing data for some taxa. It appears that chloroplast genomes fall near the middle of the range, so the most effective strategy is not obvious. Our "complete" gene set had 73 genes and "no missing genes" set had 58 genes. Examination of the trees recovered and bootstrap values for several branches (Table 4) demonstrates no strong effect of missing data on our phylogenetic analyses. This is consistent with similar findings with other genome-based data sets (Stefanovic et al., 2004). However, as we emphasize below, inferences about the appropriateness of analytical methods is highly dependent on choice of taxa and the interaction among other variables. The effect of missing genes might also be different if methods of analysis were changed.

As has been noted elsewhere, taxon sampling is critical in large-scale phylogenetic analyses (Stefanovic et al., 2004). Some have argued that because taxon sampling is so important, especially for large multigene data sets, more efficient use of resources may come from using data from more taxa but fewer genes (Soltis et al., 2004). If the sole purpose of sequencing organellar genomes is to gather aligned sequence data then such arguments are probably valid. However, if genome sequence databases can be used to extract additional comparative data such as gene order, indels, small inversions, and other characteristics that may lie outside coding regions, then the genome approach may become better justified. It seems that most of the phylogenetic analyses of organellar genome data sets are using approaches that were originally developed for one or a few genes. For example, most studies use substitution models with parameters averaged across the data rather than considering different models for each gene. Furthermore, even if models are selected it is often not straightforward to test whether violations

of the assumptions are producing incorrect trees. Accounting for such factors as third codon saturation and compositional heterogeneity (Goremykin et al., 2004) can produce results that others argue are wrong (Soltis et al., 2004). Many of these debates probably emerge because different violations of assumptions may interact and are data-dependent (Ho and Jermiin, 2004).

Here, we have presented an initial attempt at analysis of data from gene order, with mixed results. Meanwhile, providing genome-scale sequence data for previously unsampled major clades will enable researchers to gain a better understanding of where future work might be fruitful.

We used a novel combination of cell and molecular methods to isolate organelles (K. Everett, A. Arumuganathan and D.F. Mandoli, in preparation), from which we determined the first complete chloroplast genome sequence of a lycophyte. This plant represents the sister group to all other vascular plants. We present exploratory phylogenetic analyses of sequence data and genome structure across land plants. We provide genome-scale sequence data for a previously unsampled major clade of plants. These data should enable researchers to gain a better understanding of where future phylogenomic work might be fruitful. Moreover, the data are likely to be useful in other, less predictable, ways. For example, most of the chloroplast genes encode products involved in the critical cellular functions of photosynthesis and chloroplast protein synthesis. Obtaining sequence data from a wider base of taxa will allow us to learn more about the evolution and variation in these genes and their products.

Literature Cited

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

Barkman, T.J., Chenery, G., McNeal, J.R., Lyons-Weiler, J., Ellisens, W.J., Moore, G., Wolfe, A.D. and dePamphilis, C.W., 2000. Independent and combined analyses of sequences from all three genomic compartments converge on the root of flowering plant phylogeny. Proc. Natl. Acad. Sci. (U.S.A.) 97, 13166-13171.

Boore, J.L., 1999. Animal mitochondrial genomes. Nucleic Acids Res. 27, 1767-1780.

Dean, F.B., Nelson, J.R., Giesler, T.L. and Lasken, R.S., 2001. Rapid amplification of plasmid and phage DNA using Phi 29 DNA polymerase and multiply-primed rolling circle. Genome Res. 11, 1095-1099.

Dombrovska, O. and Qiu, Y.L., 2004. Distribution of introns in the mitochondrial gene *nad1* in land plants: phylogenetic and molecular evolutionary implications. Molec. Phylogenet. Evol. 32, 246-263.

Elkin, C., Kapur, H., Smith, T., Humphries, D., Pollard, M., Hammon, N. and Hawkins, T., 2002. Magnetic bead purification of labeled DNA fragments for high-throughput capillary electrophoresis sequencing. Biotechniques 32, 1296-1302.

Ewing, B. and Green, P., 1998. Base-calling of automated sequencer traces using *phred*. II error probabilities. Genome Res. 8, 186-194.

Felsenstein, J., PHYLIP vs. 3.6. Computer program distributed by Department of Genetics, University of Washington, Seattle, WA, 2004.

Gordon, D., Abajian, C. and Green, P., 1998. Consed: a graphical tool for sequence finishing. Genome Res. 8, 195-202.

Goremykin, V.V., Hirsch-Ernst, K.I., Wolfl, S. and Hellwig, F.H., 2004. The chloroplast genome of *Nymphaea alba*: whole-genome analyses and the problem of identifying the most basal angiosperm. Molec. Biol. Evol. 21, 1445-1454.

Goulding, S.E., Olmstead, R.G., Morden, C.W. and Wolfe, K.H., 1996. Ebb and flow of the chloroplast inverted repeat. Molec. Gen. Genet. 252, 195-206.

Graham, S.W., Reeves, P.A., Burns, A.C.E. and Olmstead, R.G., 2000. Microstructural changes in noncoding chloroplast DNA: interpretation, evolution, and utility of indels and inversions in basal angiosperm phylogenetic inference. Int. J. Plant Sci. 161, S83-S96.

Hasegawa, M., Dirienzo, A., Kocher, T.D. and Wilson, A.C., 1993. Toward a more accurate time-scale for the human mitochondrial-DNA tree. J. Mol. Evol. 37, 347-354.

Helfenbein, K.G. and Boore, J.L., 2004. The Mitochondrial Genome of *Phoronis architecta*-- Comparisons Demonstrate that Phoronids Are Lophotrochozoan Protostomes. Molec. Biol. Evol. 21, 153-157.

Ho, S.Y.W. and Jermiin, L.S., 2004. Tracing the decay of the historical signal in biological sequence data. Syst. Biol. 53, 623-637.

Jones, D.T., Taylor, W.R. and Thornton, J.M., 1992. The rapid generation of mutation data matrices from protein sequences. Comput. Appl. Biosci. 8, 275-282.

Kelch, D.G., Driskell, A. and Mishler, B.D., Inferring phylogeny using genomic characters: a case study using land plant plastomes. In: Goffinet, B., Hollowell, V. and Magill, R. (Eds.), Monographs in Systematic Botany from the Missouri Botanical Garden, Vol. 98 Molecular Systematics of Bryophytes. MBG Press, St. Louis, MO, USA, 2004, pp. 3-11.

Kenrick, P. and Crane, P.R., The Origin and Early Diversification of Land Plants: A Cladistic Study. Smithsonian Press, Washington, DC, U.S.A., 1997.

Kim, K.-j. and Lee, H.-L., 2004. Complete chloroplast genome sequences from Korean Ginseng (*Panax schinseng* Nees) and comparative analysis of sequence evolution among 17 vascular plants. DNA Res. 11, 247-261.

Kugita, M., Yamamoto, Y., Fujikawa, T., Matsumoto, T. and Yoshinaga, K., 2003. RNA editing in hornwort chloroplasts makes more than half the genes functional. Nucleic Acids Res. 31, 2417-2423.

Lemieux, C., Otis, C. and Turmel, M., 2000. Ancestral chloroplast genome in *Mesostigma viride* reveals an early branch of green plant evolution. Nature 403, 649-652.

Lockhart, P.J., Steel, M.A., Hendy, M.D. and Penny, D., 1994. Recovering evolutionary trees under a more realistic model of sequence evolution. Molec. Biol. Evol. 11, 605-612.

Maddison, W.P. and Maddison, D.R., MacClade: Analysis of phylogeny and character evolution. Version 4.0b6. Sinauer, Sunderland, MA, 2003.

Mishler, B.D., 2000. Deep phylogenetic relationships among "plants" and their implications for classification. Taxon 49, 661-683.

Nishiyama, T., Wolf, P.G., Kugita, M., Sinclair, R.B., Sugita, M., Sugiura, C., Wakasugi, T., Yamada, K., Yoshinaga, K., Yamaguchi, K., Ueda, K. and Hasebe, M., 2004. Chloroplast phylogeny indicates that bryophytes are monophyletic. Molec. Biol. Evol. 21, 1-7.

Normile, D., 2001. Evolutionary genomics - The ups and downs of evolution. Science 294, 2281-2282.

Palmer, J.D. and Stein, D.B., 1986. Conservation of chloroplast genome structure among vascular plants. Curr. Genet. 10, 823-833.

Raubeson, L.A. and Jansen, R.K., 1992. Chloroplast DNA evidence on the ancient evolutionary split in vascular land plants. Science 255, 1697-2699.

Shahmuradov, I.A., Akbarova, Y.Y., Solovyev, V.V. and Aliyev, J.A., 2003. Abundance of plastid DNA insertions in nuclear genomes of rice and Arabidopsis. Plant Mol. Biol. 52, 923-934.

Soltis, D.E., Albert, V.A., Savolainen, V., Hilu, K., Qiu, Y.-L., Chase, M.W., Farris, J.S., Stefanovic, S., Rice, D.W., Palmer, J.D. and Soltis, P.S., 2004. Genome-scale data, angiosperm relationships, and 'ending incongruence': a cautionary tale in phylogenetics. Trends Plant Sci. 9, 477-483.

Stefanovic, S., Rice, D.W. and Palmer, J.D., 2004. Long branch attraction, taxon sampling, and the earliest angiospersm: *Amborella* or monocots? BMC Evolutionary Biology 4, 35.

Stoebe, B., Hansmann, S., Goremykin, V., Kowalik, K.V. and Martin, W., Proteins encoded in sequenced chloroplast genomes: an overview of gene content, phylogenetic information and endosymbiotic gene transfer to the nucleus. In: Hollingsworth, P.M., Batesman, R.M. and Gornall, R.J. (Eds.), Molecular Systematics and Plant Evolution. Taylor & Francis, London, 1999, pp. 327-352.

Strimmer, K. and Haeseler, A.v., 1996. Quartet puzzling: a quartet maximum likelihood method for reconstructing tree topologies. Molec. Biol. Evol. 13, 964-969.

Swofford, D.L., PAUP*: Phylogenetic analysis using parsimony (*and other methods) version 4. Sinauer Associates, Sunderland, MA, 2003.

Swofford, D.L., Olsen, G.J., Waddell, P.J. and Hillis, D.M., Phylogenetic inference. In: Hillis, D.M., Moritz, C. and Mable, B.K. (Eds.), Molecular systematics. Second edition. Sinauer, Sunderland, MA, 1996, pp. 407-514.

Wolf, P.G., Rowe, C.A. and Hasebe, M., 2004. High levels of RNA editing in a vascular plant chloroplast genome: analysis of transcripts from the fern *Adiantum capillus-veneris*. Gene In Press.

FIGURE LEGENDS

Figure 1. Sorting gates for flow cytometry on the scatter plots of red versus green fluorescence intensity are drawn around the group of events of signals from stained, putative chloroplasts and mitochondria. Note that collection as an intact putative chloroplast required only red fluorescence, but that collection as an intact putative mitochondrion required both red and green fluorescence. About twenty million chloroplasts and twenty million mitochondria were collected. Unstained and DAPI stained controls were done for each FACs run of this species (not shown).

Figure 2. Map of the chloroplast genome of *Huperzia lucidula*. Genes on the outside are transcribed clockwise and those on the inside are transcribed counter-clockwise.

Figure 3. Phylogenetic relationships of representative lineages of land plants inferred using 73 protein-coding genes from 20 chloroplast genomes. The maximum-likelihood (ML) tree using 48,201 nucleotide sites (all codon positions) is shown (-ln=388577.11007). Percent bootstrap proportions from all ML analyses (200 replicates each) are shown above branches and from all MP analyses (1000 replicates each) are shown below branches. Letters on branches: A) The ML analysis excluding 3rd codon positions yielded a similar topology (-ln=207477.05846) with the exception of *Arabidopsis* sister to *Oenothera* (BS=66%) and *Oryza* sister to *Zea* (BS=58%). B-E) The ML analysis using inferred amino acid data (-ln= 180033.90946) and all three MP analyses differed in the placement of the monocots sister to a monophyletic dicot clade including *Amborella* and *Calycanthus* (ML amino acid; 98%; MP all codon positions, BS=97%; MP excluding 3rd codon positions, BS=100%; MP amino acid, BS=100%). F-G) Both MP nucleotide analyses placed *Huperzia* sister to the seed plants to the exclusion of the ferns (all codon positions, BS=60%; excluding 3rd codon positions, BS=75%). The horizontal branch lengths are proportional to the estimated number of nucleotide substitutions per site. Common

names of higher taxonomic units are indicated on the right with square brackets. The topology is drawn with *Chaetosphaeridium* rooting the tree.

Figure 4. Strict consensus tree of 900 most parsimonious trees based on gene order analysis. Values below branches represent bootstrap values.

Fig. 5. Border of the inverted repeat A (IR$_A$) adjacent to the Large Single Copy (LSC) region, showing changes that are putative synapomorphies of the clades *Huperzia* plus *Anthoceros* plus euphyllophytes (char. 27) and *Anthoceros* plus seed plants (char. 1). One end of the IR is shown for selected taxa; genes in the inverted repeat are in bold. The complete chloroplast gene alignment is available at http://ucjeps.berkeley.edu/treeoflife/.

Table 1: GenBank accession numbers and sources of chloroplast gene maps for sampled taxa

Taxon	GenBank accession #	Taxon	GenBank accession #
Charophytes		**Angiosperms**	
Chaetosphaeridium globosum (Nordstedt) Klebahn	NC_004115	*Amborella trichopoda* Baill.	NC_005086
Liverworts		*Arabidopsis thaliana* (L.) Heynh.	NC_000932
Marchantia polymorpha L.	NC_001319	*Atropa belladonna* L.	NC_004561
Mosses		*Epifagus virginiana* L. (Bart.)	NC_001568
Physcomitrella patens (Hedw.) Bruch & W. P. Schimper	NC_005087	*Calycanthus floridus* L.	NC_004993
Hornworts		*Lotus japonicus* (Regel) K.Larsen	NC_002694
Anthoceros formosae Stephani	NC_004543	*Nicotiana tobacum* L.	NC_001879
		Oenothera elata Kunth ssp. *hookeri* (Torr. & A.Gray) W.Dietr. & W.L.Wagner	NC_002693
Lycophytes			
Huperzia lucidula (Michx.) Trevisan		*Oryza sativa* L.	NC_001320
Moniliforms		*Spinacia oleracea* L.	NC_002202
Adiantum capillis-veneris L.	NC_004766	*Triticum aestivum* L.	NC_002762
Psilotum nudum (L.) P.Beauv.	NC_003386	*Zea mays* L.	NC_001666
Conifers			
Pinus koraiensis Siebold & Zucc.	NC_004677		
Pinus thunbergiana Franco	NC_001631		

Table 2. Explanation of characters used in phylogenetic analysis.

1. Inclusion of *rps*7, *ndh*B, (and *trn*L-CAA) in the inverted repeat (IR) from the large single copy

    (LSC) margin of IR$_A$, leading to gene duplication on the LSC margin of I R$_B$.

2. Inversion of the gene order within the IRs.

3. Loss or gain of IR$_B$.

4. Large multi-gene (>20) inversion in what corresponds to the small single copy (SSC)

5. Loss or gain of *Chl*N and *Chl*L genes from IR$_A$ end of the small single copy area.

6. Inclusion of *ycf*2, *trn*H-GTG, and *psb*A within the inner edge of the IRs.

7. Loss or gain of *trn*V and *rps*12 from the LSC margin of IR$_B$ in relation to other sampled taxa.

8. Loss or gain of *rps*7 from what is IR$_B$ in other sampled taxa.

9. Large multi-gene (>25) inversion from *psb*M to *ycf*2.

10. Inclusion of *rpl*23 and *rpl*2 from IR$_B$ end of LSC into the IR region.

11. Inclusion of *trn*P-GGG, *rpl*32, and *rpl*21 into the IR region from the SSC margin of IR$_B$.

12. Large multi-gene (>50) inversion in relation to other sampled taxa.

13. Inversion of *mat*K and *trn*K-UUU in relation to other sampled taxa.

14. Multi-gene (> 10) rearrangement following *trn*G-UCC.

15. Non-alignable section in regard to other taxa.

16. Multi-gene (>20) inversion in relation to other sampled taxa.

17. Inversion of *pet*N and *psb*M.

18. Movement of multi-gene section including *psb*A, *trn*H-CAC, and *ycf*2.

19. Loss or gain of gene section between *psb*C and *trn*fM.

20. Inversion of 6 gene section (from *trn*G-GCC to *trn*T-ACC)

21. Presence/absence of *rps*12 gene between *rpl*36 and *rps*8.

22. Presence/absence of *rpl*22 between *rps*3 and *rps*19.

23. Presence/absence of *trn*H-GUG between *rps*19 and *rpl*2.

24. Presence/absence of *ycf*2 between *trnI*-CAU and *ycf*15 or *trn*L-CAA.

25. Presence/absence of *ycf*15 between *ycf*2 and *trn*L-CAA.

26. Presence/absence of *trn*L-CAA between *trn*I or *ycf*2 and *ndh*B.

27. Presence/absence of *rps*12 in IR$_B$

28. Presence/absence of rps15 at SSC margin of IR$_B$.

29. Presence/absence of *ycf*1 at SSC margin of IR$_B$.

30. Presence/absence of *rpl*21 between *ndh*F and *rpl*32.

31. Presence/absence of *trn*P-GGG between *rpl*32 and *trn*L-UAG.

32. Absence of *ycf*1 adjacent to *rps*15 (possible pseudogene present).

33. Presence/absence of *ndh*J between *trn*F-GAA and ndhK.

34. Intron missing from gene (pseudogene) of *rpl*2.

35. Lack of intron in *rps*12.

36. Lack of intron in *atp*F.

37. Lack of intron in *rpo*C1.

38. Lack of one of the introns in *ycf*3.

39. First intron missing from gene *clp*P.

40. Second intron missing from gene *clp*P.

41. Presence/absence of *rps*12 between *rpl*20 and *clp*P.

42. Presence/absence of *trn*W-CCA and *trn*P-UGG between *pet*G and *psa*J.

Table 3: Data matrix and character state assignment. Refer to Table 2 for character state explanations. Characters are binary, with ? representing unknown data.

```
                          10          20          30          40
                           .           .           .           .
Chaetosphaeridium   0001100000 0001100100 1100000000 1101101011 10

Marchantia          0001100000 0000000000 1100000000 1101111011 00

Physcomitrella      0001100000 0100000000 1100000000 ?101111111 10

Anthoceros          1001100000 0000000000 1100001000 1101111111 10

Psilotum            1001000010 1000000001 1100011000 1101?11111 10

Adiantum            1101110010 0000001001 1100001000 1101111111 00

Pinus koraiensis    ?010101010 0000000010 11000?1001 1111111?00 11

Pinus thunbergiana  1010101110 0000000010 11000?1001 1111111111 11

Oenothera           1001000011 0010000000 1101111001 0101111111 00

Oryza               1001000011 0000010000 111?011101 0001110100 00

Zea                 1001000011 0000010000 1111011101 0001110100 00

Spinacia            1001000011 0000000000 1101011001 0100111111 10

Calycanthus         1001000011 0000000000 1101011001 0101111111 10

Arabidopsis         1001000011 0000000000 0101011011 0101011011 10

Atropa              1001000011 0000000000 0101111011 0101111111 10

Nicotiana           1001000011 0000000000 0101111011 0101111111 00

Lotus               1001000011 0000000000 0001111011 0101111011 00

Epifagus            1001000011 0000000000 10010110?1 010111??11 00

Amborella           1001000011 0000000000 1101111001 0101111111 10

Triticum            1001000011 0000010000 1110011101 0001110100 00

Huperzia            0001100000 0000000000 1100000000 1001?11111 10
```

Table 4. Bootstrap values for different clades for all analyses performed for both data set 73 and data set 58. Arab…ferns denotes a branch subtending the clade including all taxa from *Arabidopsis*, in order, to the ferns, as presented in Figure 3. Hup + March denotes the clade of only the two taxa Huperzia and Marchantia, not seen on Figure 3.

	Arab… Ferns	Seed+ Huperzia No Fern	Trach	Trach + Antho	Arab… Physco	Hup+ Antho	Physco + March	Bryo mono
Data set 73:								
ML all positions	100	<50	100	100	100	<50	<50	<50
ML No 3rd	98	<50	100	100	95	<50	<50	<50
ML AA	92	<50	58	92	<50	<50	<50	<50
MP all positions	<50	60	89	99	99	<50	<50	<50
MP no 3rd	<50	75	67	75	55	<50	<50	<50
MP AA	50	<50	<50	52	<50	<50	<50	<50
LogDet all positions	90	<50	<50	100	<50	99	100	<50
LogDet no 3rd	90	<50	72	<50	<50	<50	<50	70
Data set 58:								
ML all positions	100	<50	100	100	100	<50	<50	<50
ML No 3rd	97	<50	100	100	84	<50	<50	<50
ML AA	86	<50	78	94	53	<50	<50	<50
MP all positions	<50	87	99	100	100	<50	<50	<50
MP no 3rd	<50	78	69	69	<50	<50	<50	<50
MP AA	<50	74	<50	<50	<50	<50	63	<50
LogDet all positions	50	<50	54	100	<50	<50	100	<50
LogDet no 3rd	72	<50	100	<50	<50	<50	100	99