Systematic study

Wittall J.B. et al. (2010): Finding a (pine) needle in a haystack: chloroplast genome sequence divergence in rare and widespread pines. *Molecular Ecology* 19, 100-114.

Reasons for the study

- compare level of whole chloroplast differentiation in pines with narrow and broad distribution
- chloroplast differentiation between two subspecies of *P. torreyana*
- compare differentiation with other species pairs
- test NGS for reliable SNP detection
- divergence dating

Chloroplast genome

- predominant uniparental inheritance – paternal in conifers – tracks pollen dispersal
- conservative mutation rate – 100x lower than animal mitochondria
- primarilly microsatellites studied – highly variable but high degree of homoplasy
- A/T rich (~ 62%) – can cause biased sequencing errors – problem when surveying for rare polymorphism

Study species

Pinus torreyana – 2 populations in California

- mainland – *P. torreyana* subsp. *torreyana* (81)
- island – *P. torreyana* subsp. *insularis* (86)

P. monticola S – N

P. lambertiana S – N

P. lambertiana N – *P. albicaulis*

P. ayacahuite – P. flexilis (~2200 km distant)

P. cembra – P. sibirica (~4800 km distant)

Methods

- 35 separate PCR reactions – to amplify whole chloroplast (Cronn et al. 2008)
- quantification, equimolar pooling, barcoded Illumina libraries
- pooling – 4 libraries (full chloroplast) or 16 (partial)
- de novo assembly (VELVET, EDENA) – minimum depth 5x, minimum contig length 100 bp
- alignment of de novo conting to a reference chloroplast (*P. ponderosa*, *P. koraiensis*) – CODONCODE
- consensus sequence (BioEdit) + reference -> 'chimeric pseudoreference'
- microread mapped onto pseudoreference (RGA) – minimum depth 2x, 70% majority minimum for SNP
- alignment of genomes – MAFFT
- annotation (DOGMA)

Methods

- *P. torreyana* – SNP validation by Sanger sequencing (regions flanking putative SNPs)
  - identification of false-positives and false-negatives
- pairwise comparison of genomes (MEGA)
  - minimum depth 25x, 85% majority base call
  - uncorected pairwise distances
  - silent sites (d*S* – synonymous)
  - non-synonymous sites (d*N*)
- AMOVA – hierarchical structure in *P. monticola*
- *P. torreyana* – SNP genotyping using dCAP assay (*derived cleaved amplified fragment length polymorphism*)
- divergence dating – calibrated with chloroplast-specific mutation rate estimated for *Pinus*

Results

- 1 336 085 microreads (33-37 bp) on average per genome
- de novo asseblies consistently interrupted at priming sites
- *P. torreyana* – 32 putative SNPs (Table 2, Fig. 2, Fig. 3), bi-allelic
  - 5 validated by Sanger sequencing
  - false positives (not confirmed) – low sequencing depth
  - 7 false negatives (consistently present in Sanger sequences) – no novel SNPs
- uneven distribution of variable sites across genome
- differences between genomes (Table 3)
  - no – *P. sibirica* vs. *P. cembra*
  - 382 – within *P. lambertiana*
  - divergence dates
- spatial differentiation
  - *P. torreyana* – 5 validated SNPs fixed between populations
  - 10 *P. monticola* individuals – 9 distinct haplotypes – no geographic pattern (in contrast to nuclear differentiation)

Discussion

- chloroplast genome-wide sequence variation is very low in pine species – all comparison fewer than 18 SNPs
  - even for geographically widespread species
  - low variation in *P. torreyana* is not due to its rarity but it is a norm for *Pinus*
  - > full chloroplast genomes are required for robust resolution
- uneven distribution of variation
  - no best highly variable region region
  - > again plastome scale approach necessary
- chloroplast introgression of *P. albicaulis* to northern population of *P. lambertiana*
- future prospects – comparison of microsatellite and NGS analysis – longer reads necessary for direct comparison