Direct estimates of intra- and interspecific variation in the rate and spectrum of mitochondrial mutation in the genus Daphnia

...

# Introduction

see Ness et al., 2015 intro for review of intraspecific mutation rate variation -> unhealthy strains show elevated mutation rate. # Materials And Methods

## Data sources

(Mutation accumulation protocol details for d.magna (e.g num lines, num gen))  
In addition, sequence data for 7 MA lines in *D.pulex* was sourced from a recent MA experiment (Keith et al., 2015; PRJNA275628). Four of these lines were sourced from natural isolates from an obligate asexual genotype from Linwood, Ontario (ASEX). The other three lines were from an a cyclical parthonogenic population from Slimy Log Pond, Oregon (CYC). ASEX average gen=170, CYC average gen=85.

## Mutation Accumulation Lines

(D. magna only) The populations from which the D. magna genotypes used in this experiment are derived were initially collected along a latitudinal gradient spanning from Finland to Israel. We used three unique genotypes from each of three populations located in Finland, Germany, and Israel (9 genotypes total). The stock cultures for each genotype were maintained in 3L jars containing 2L of the medium AdAM (cite) under a constant 16L:8D photoperiod, and fed the unicellular green alga Scenedesmus obliquus ad libitum (2-3 times per week).

We established replicate clonal lineages from each of the 9 genotypes (45 lineages total) between October 2013 and April 2015. The lineages (mutation-accumulation lines [MA lines]) were established and maintained by placing a single female daphniid collected from the stock cultures in a 250 mL beaker containing 100 mL of a ADaM supplemented with S. obliquus at a concentration of 600,000 cells/mL, and maintained in a Percival environmental chamber under a 16L:8D photoperiod. The media was changed once per week and each line was fed three days after the water replacement to return the algal cell concentration in the beaker back to 600,000 cells/mL. Each lineage was propagated from generation to generation over the course of the experiment via single offspring descent by taking a single juvenile from the second clutch of the mother. A series of backups were maintained in parallel with the focal lineages in the event that the single individual intended to be used to establish the next generation died before reproduction, or was a male. In the event that the focal lineage and all backups either died before reproduction, or were all males, the lineage was declared extinct and a new replicate lineage was established from the stock cultures.

## DNA Extraction and Sequencing

*D. magna* Ten clonal individuals from each MA line and control (1 starting control and 2 existing controls per genotype) were flash frozen for DNA extractions. DNA was extracted (2 extractions per line with 5 daphnia each) using the Zymo Quick-DNA Universal Solid Tissue Prep Kit (No. D4069) following the manufacturer’s protocol (DNA was extracted from a few samples with the Qiagen DNeasy Blood and Tissue Kit, No. 69504). DNA quality was assessed by electrophoresis on 3% agarose gels and DNA concentration was determined by dsDNA HS Qubit Assay (Molecular Probes by Life Technologies, No. Q32851).

The Center for Genome Research and Biocomputing at Oregon State University generated 94 Wafergen DNA 150bp paired-end libraries using the Biosystems Apollo 324 NGS library prep system. Libraries were pooled based on qPCR concentrations across 16 lanes (2 runs) and sequenced on an Illumina Hiseq 3000. The average insert size was approximately 380bp.

## Sequence Processing and Alignment

Reads were trimmed using Skewer (V. 0.2.2; Jiang et al., 2014) as follows: \* Bases with Phred score below 30 were trimmed from 3' ends of reads. \* Illumina adapter sequences were removed (see Supplemental Material for adapter sequences). \* Reads with an overall error rate greater than 10 % of read length, or an indel error rate greater than 3 % of read length were excluded.

To account for alignment errors due to mitochondrial pseudogenes (nuMTS; Bensasson et al., 2001) and the circular structure of mtDNA we designed a two-step read mapping protocol. (Consider figure to explain this) In the first step, trimmed reads were simultaneously aligned to their corresponding nuclear (*D. magna*: Accession LRGB00000000.1; *D. pulex*: Accession ACJG00000000.1) and mitochondrial (*D. magna*: Accession NC\_026914.1; *D. pulex*: Accession NC\_000844.1) published reference sequences using Bowtie 2 with the --very-sensitive preset (V. 2.2.9; Langmead & Salzberg, 2012). The mitochondrial reference sequence was modified such that its last 1000 bp was also prepended to the start of the sequence. This ensured that reads spanning the start-end junction of the reference sequence were captured. The published *D. magna* nuclear assembly sequence contains a 14542 bp scaffold, (scaffold01847; Accession LRGB01001847.1) and a 244 bp contig (contig53563; Accession: LRGB01023672.1), which together span almost 99 % of the mitochondrial genome, with no flanking non-mitochondrial sequence. We concluded that these are mitochondrial in origin, and excluded them from the nuclear reference sequence before alignment.

Read pairs which aligned to the mitochondrial genome were then remapped both to the original mitochondrial reference sequence, and a second 'rotated' reference sequence in which the second half of the reference sequence was moved to the start (positions 7667 - 15333 for *D. pulex*; 7475 - 14948 for *D. magna*). All variant calling was confined to the middle segments of the original and rotated reference sequences. This ensured that start-end junction alignment errors could not bias variant detection.

After alignment, reads were probabilistically realigned using Lofreq's viterbi command (Wilm et al., 2012), locally realigned around indels with GATK's indelRealigner (V. 3.6.0; McKenna et al., 2010) and PCR duplicates were removed with Picard MarkDuplicates (V. 2.7.2; <http://broadinstitute.github.io/picard/>).

Default parameters were used with all tools unless otherwise noted. See supplemental materials for MultiQC (Ewels et al., 2016) reports showing sequence processing and alignment statistics.

## Sequencing error estimate

Sequencing error rates were estimated from the nuclear portion of the competitive alignments. Consensus sequences were produced for each isolate, and error rates were calculated as total #mismatches/ total #assayed sites, as previously described (Sung et al., 2012). To avoid confusing heterozygous bases with sequencing error, we ignored all positions with depth < 40 or minor allele frequency > 0.2. However, it is impossible to distinguish between heterozygosity and sequencing error with complete certainty, so our estimates are likely to be slightly upwardly biased.

## Variant Calling

Mutation calling in the mitochondrial genome requires separating low-frequency heteroplasmies from sequencing error. Additionally, pre-existing mutations must be separated from de novo mutations over the course of the MA experiment. Previous NGS mitochondrial MA studies have relied on minimum allele frequency cutoffs to distinguish low-frequency variants (Saxer et al., 2012; Sung et al., 2012; Kondra et al., 2017). This method requires the assumption that sequencing error rates are relatively constant across genomic locations, which has been shown to be untrue in practice (CITATIONS). We instead developed a novel Monte Carlo simulation variant detection method which takes advantage of the correlation in sequencing error profiles across samples. A consensus sequence was produced for each isolate which should represent the genome sequence of the original line progenitor. Then for every genome position in each isolate, the non-consensus allele with the highest allele frequency in a single sample was determined. A contingency table C can be produced for each genome position as follows: TABLE FORMAT where blah is blah and blah is blah etc.

A knitr kable.

|  |
| --- |
| Amut |
| Bmut |
| ... |
| Awt |

Mutations were detected using a novel Monte Carlo simulation based technique.

number which are protein coding. base qual > 30 above 20% of avg read depth double strand valid