Discovery of sex-specific regions in a salamander genome

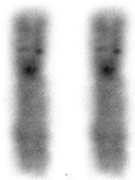
Timoshevskaya NY, Keinath MC, Smith JJ. *University of Kentucky, Department of Biology*

Biological aspects

Salamander (*Ambystoma mexicanum*) has a gigantic genome :
~32,000,000,000 bases
(10X of size of human genome):

Sex is determined by a pair of morphologically identical chromosomes:

ZZ in male
ZW in female



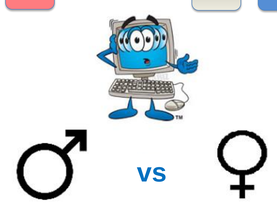
Object: Find (if there are any) genomic differences between chromosomes W and Z

Workflow:

1. Sequencing and *de novo* assembly of the reference salamander genome
2. Alignment of short sequences from male and female genomes to the reference
3. Coverage analysis

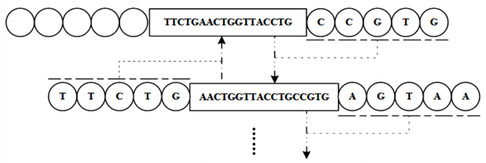
Expected relative coverage when aligning male/female reads to female genome

for male			for female		
1	1	0	1	½	½



Building *de novo* genome assembly

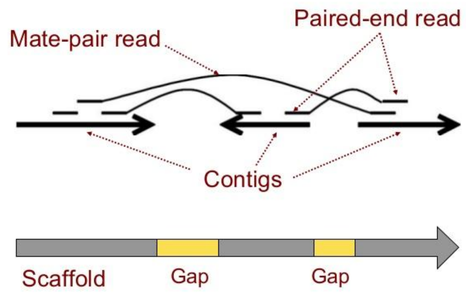
4.8Tbases of DNA sequences to build initial assembly

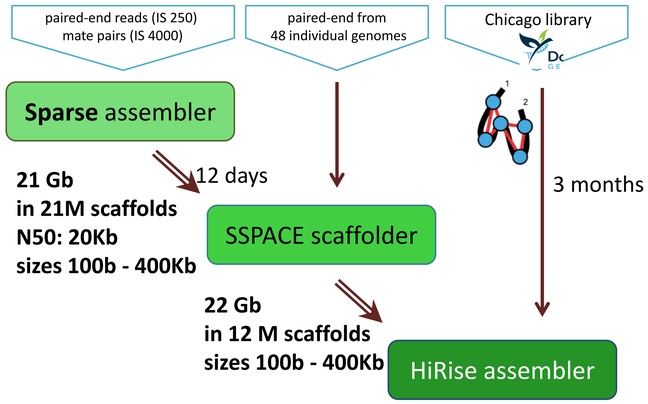


A ***SPARSE*** de Bruijn graph to surmount memory limitations

Ye C, et al. ***Exploiting sparseness in de novo genome assembly, BMC Bioinformatics*, 2012 13 (Suppl. 6):S1**

Contigs to scaffolds





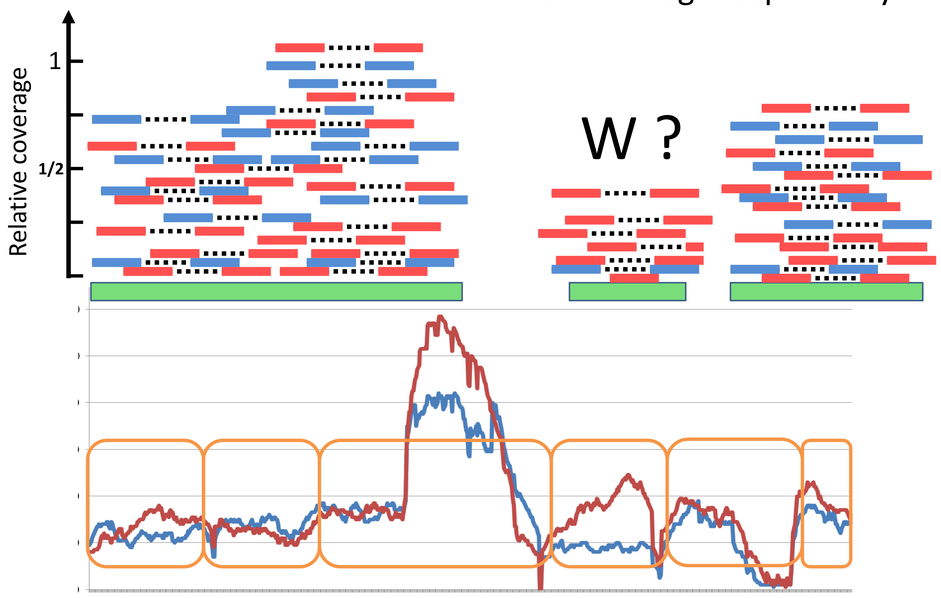
Result:
First draft of whole genome assembly published on Sal-Site (http://www.ambystoma.org)

Future task: need to improve contiguity
- gene targeted scaffolding with RNA-seq data
- gap closing

22 Gb
in 12 M scaffolds
Doubled number of scaffolds of size > 100Kb

Search of genomic regions with significantly different coverage

Alignment of paired-end reads
from sequencing of 22 females and 26 males
with 25X and 29X of combined coverage respectively



Stretched windows limited by user-defined parameters.
For each window calculate $R = \log_2(D_1/D_2)$, where D_k average coverage over *valid* bases for sample k. Search windows with $R > p$ or $R < -p$

DifCover – computational pipeline for search of genomic regions with significant difference of read coverage between two samples. The pipeline introduces ***stretched windows*** used for robust coverage analysis.

- Takes into account coverage fluctuation
- Allows exclusion of regions based on their coverage
  - under-sampled due to nucleotide content
  - repetitive sequences
- Oriented toward the analyses of large genomes and can handle very fragmented assemblies
- Inherent parallelization

Github link

Finding a needle in a haystack:
≈ 300Kb of W-specific sequences, that is
≈1/100,000th of the entire genome
with some of them validated and more being tested

1st Annual Commonwealth Computational Summit – October 17, 2017