Hybrid assembly of non-model lizards

UNIVERSITÄT LEIPZIG

Sree Rohit Raj Kolora

rohit@bioinf.uni-leipzig.de

18.02.2016

Background

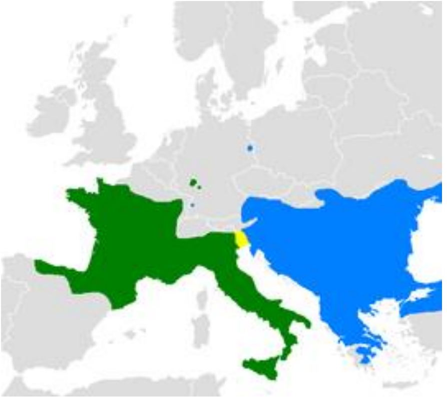
- A non-model system
- Lacertidae – 39 genera

    9 Lacertid species

- *Adaptive radiation*
- *Lacerta viridis* and *Lacerta bilineata*

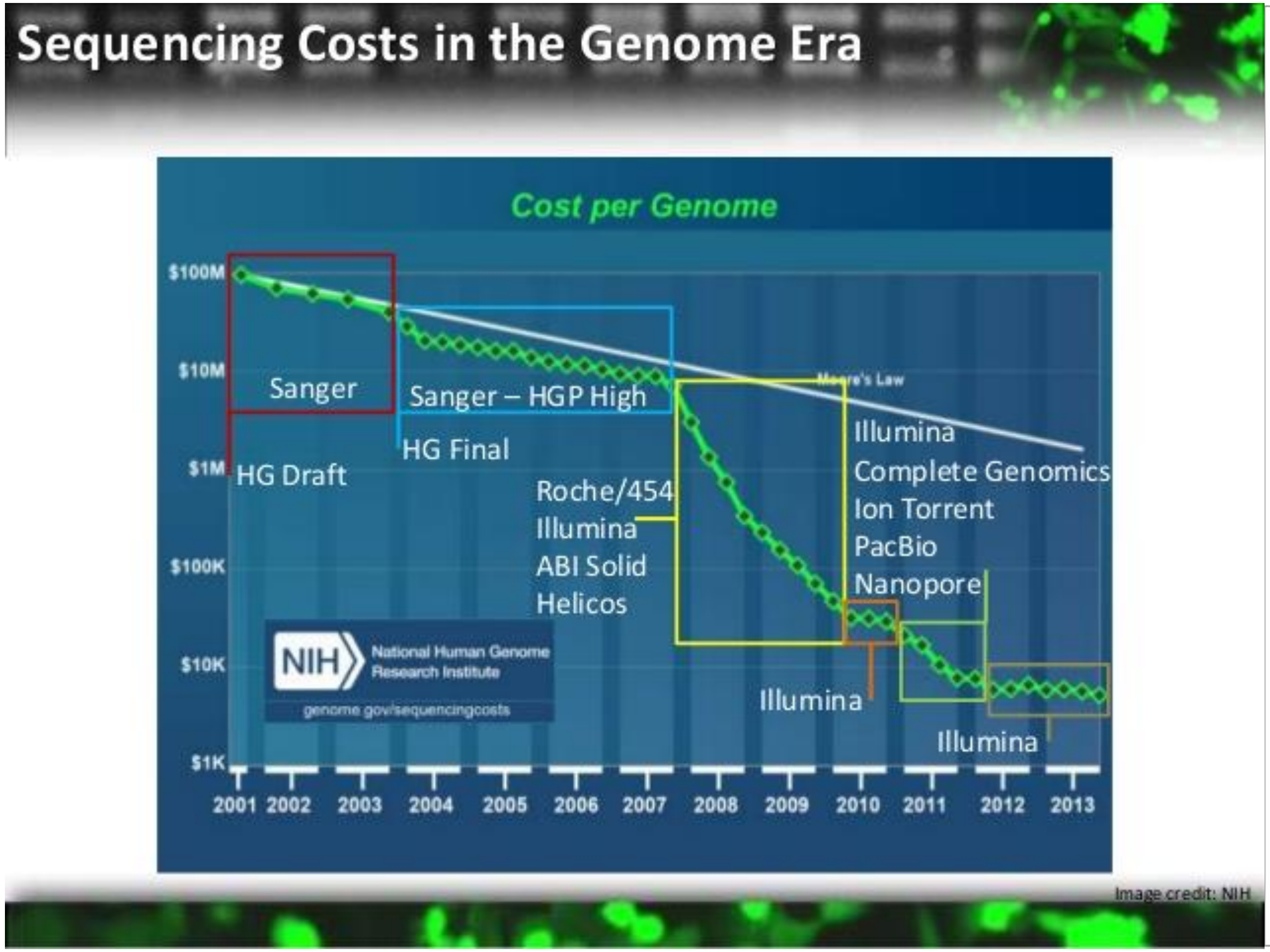
    Parapatric species – Slovenian contact zone

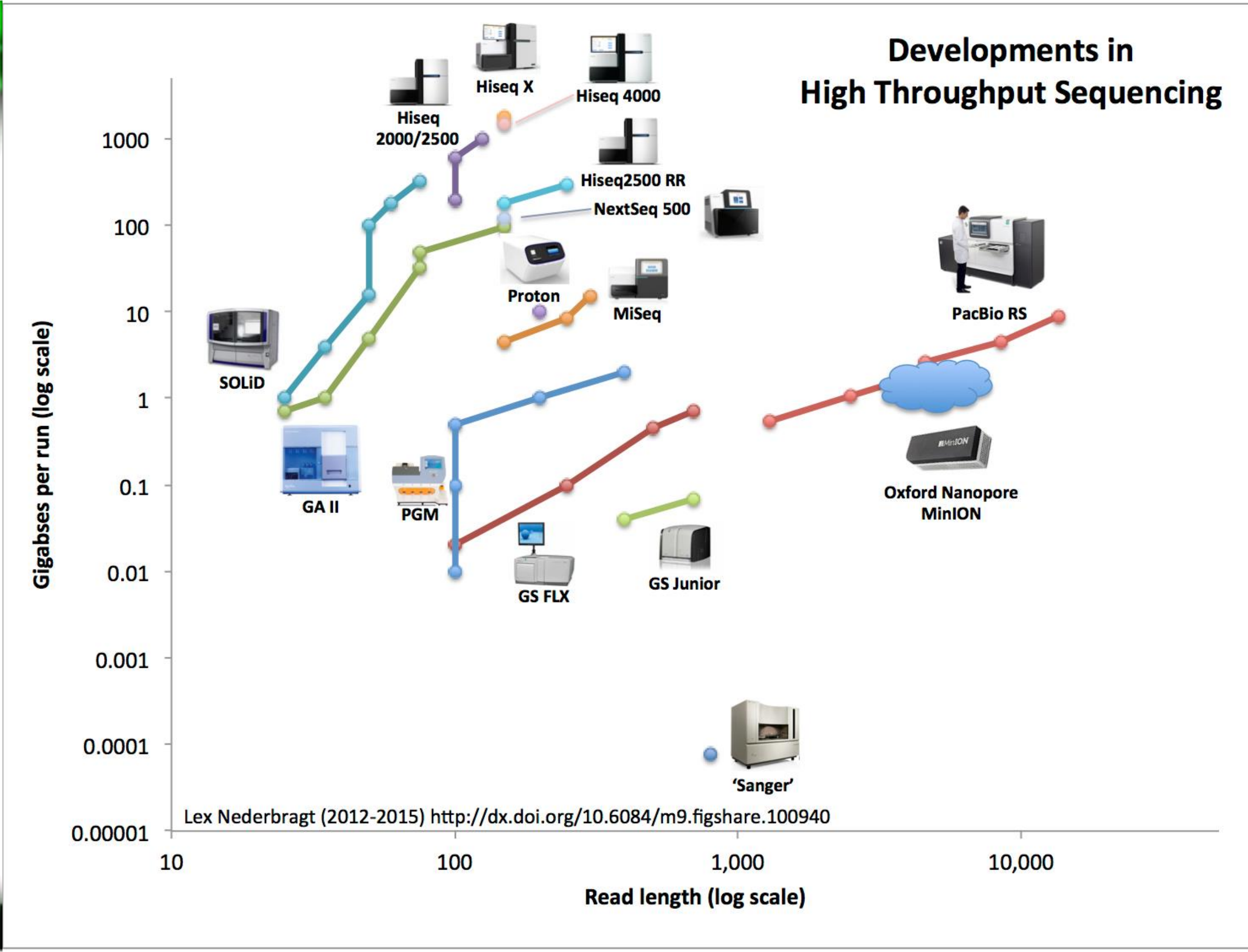
- Separated in the Pleistocene era
- Oldest lacertid fossil found 30 million years



Are Rearrangements barriers to Speciation???

CAMARGO, A., SINERVO, B., & SITES, J. (2010). Lizards as model organisms for linking phylogeographic and speciation studies. *Molecular Ecology*,*19*(16), 3250-3270. doi:10.1111/j.1365-294x.2010.04722.x





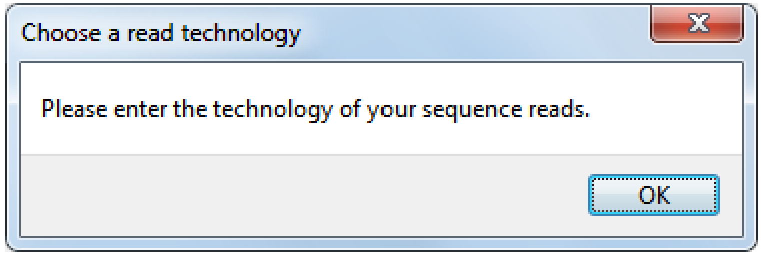
http://www.slideshare.net/Kruegsybear/high-throughput-sequencing-technologies-on-the-path-to-the-0-genome
https://flxlexblog.wordpress.com/2015/06/17/developments-in-high-throughput-sequencing-june-2015-edition/

Long Read Assembly

- Error correction with PacBio only

  30-40X coverage

- Pure PacBio assembly
- Error correction with Illumina
- Assembly of error corrected pacbio reads
- Hybrid assembly with both Illumina and PacBio with contigs and long reads



Choose a read technology

Please enter the technology of your sequence reads.

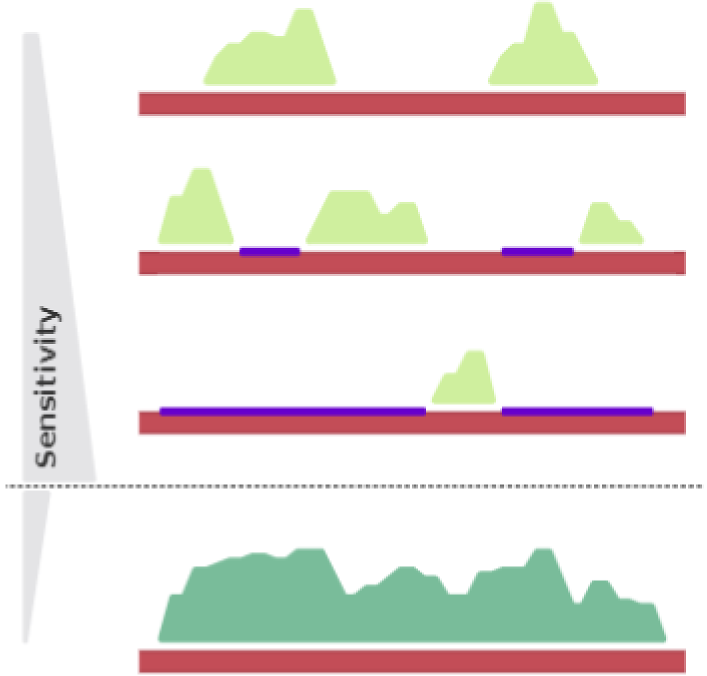
Error correction of Pacbio - **Proovread**

- Lordec, HGAP, Proovread
- Data loss in correction
- Proovread correction

  bwa mem alignment

  modified for error rates, binning, gapping

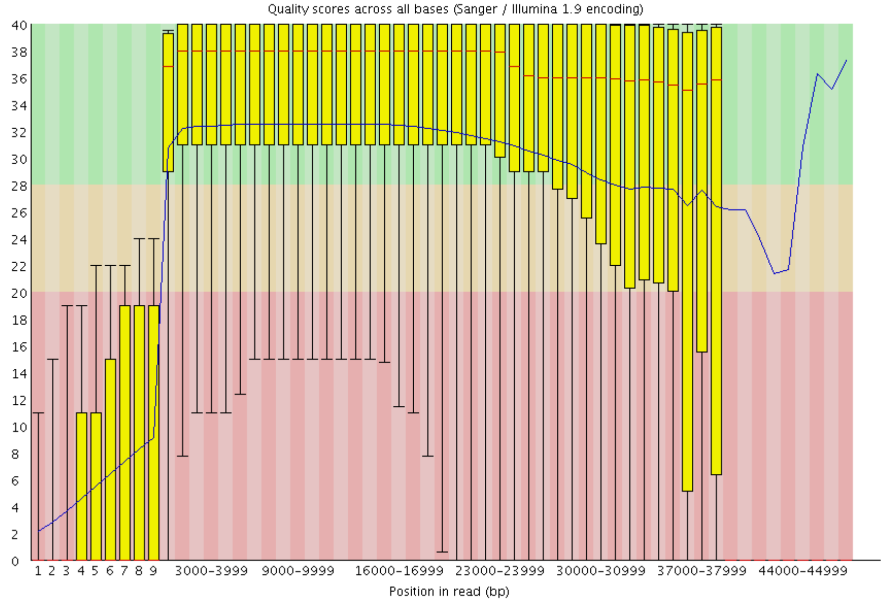
  iterations for sensitivity



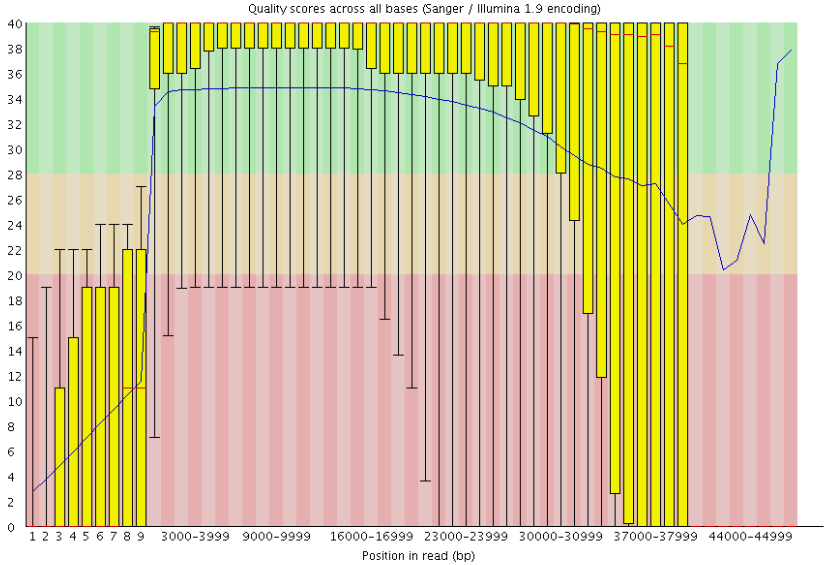
- Insertions are 2X of deletions in Pacbio i.e. 10% and 5% , substitutions are <1%
- Cons – Resources and time

PacBio Error correction - FastQC

Lacerta bilineata

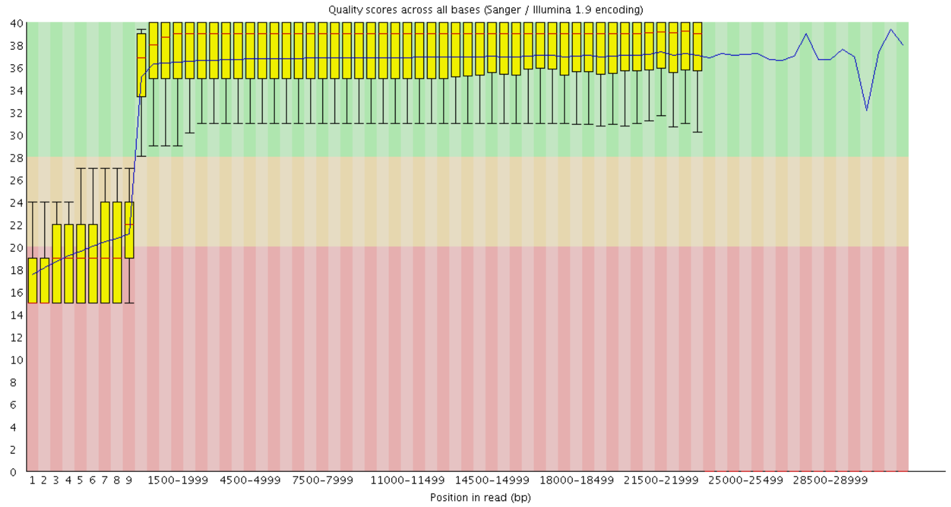


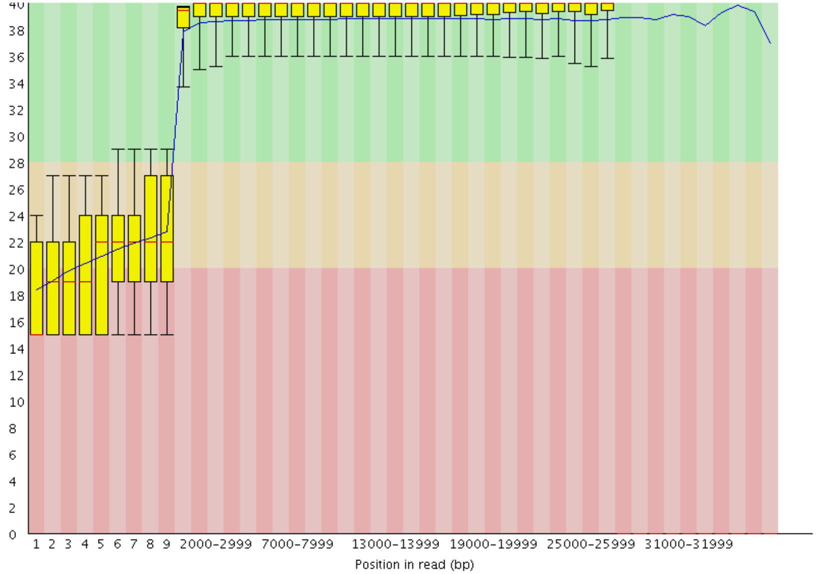
Lacerta viridis



Before

Proovread





After

PacBio hybrid assembly – **DBG2OLC**

1. Illumina contig assembly

2. Map pacbio-reads to contigs

3. Chimera removal

4. Overlap construction

5. Consensus generation

Cons – Chimeras, consensus creation - blasr

Figure 1a | Map *de Bruijn* graph contigs to the long reads. The long reads are in red, the *de Bruijn* graph contigs are in other colors. Each long read is converted into an ordered list of contigs, termed compressed reads.

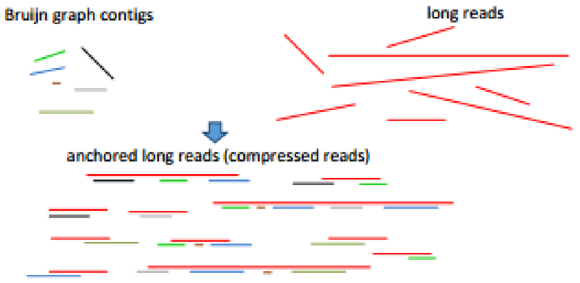


Figure 1b | Calculate overlaps between the compressed reads. The alignment is calculated using the anchors. Contained reads are removed and the reads are chained together in the best-overlap fashion.

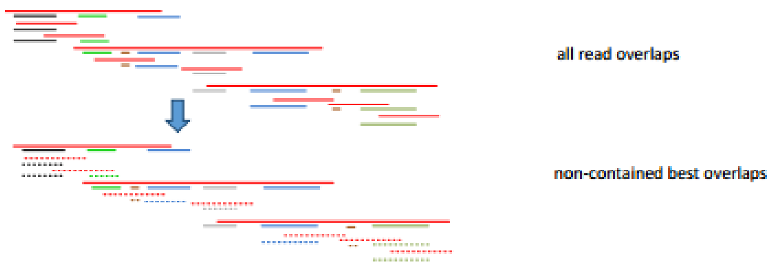
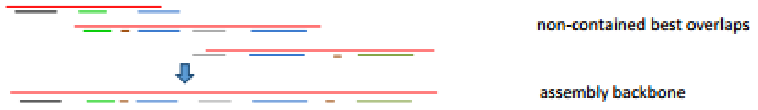


Figure 1c | Layout: construct the assembly backbone from the best overlaps.



Assembly validation

BUSCO

QUEST FOR QUALITY

"BUSCO CALIDAD"

"BUSCO QUALIDADE"

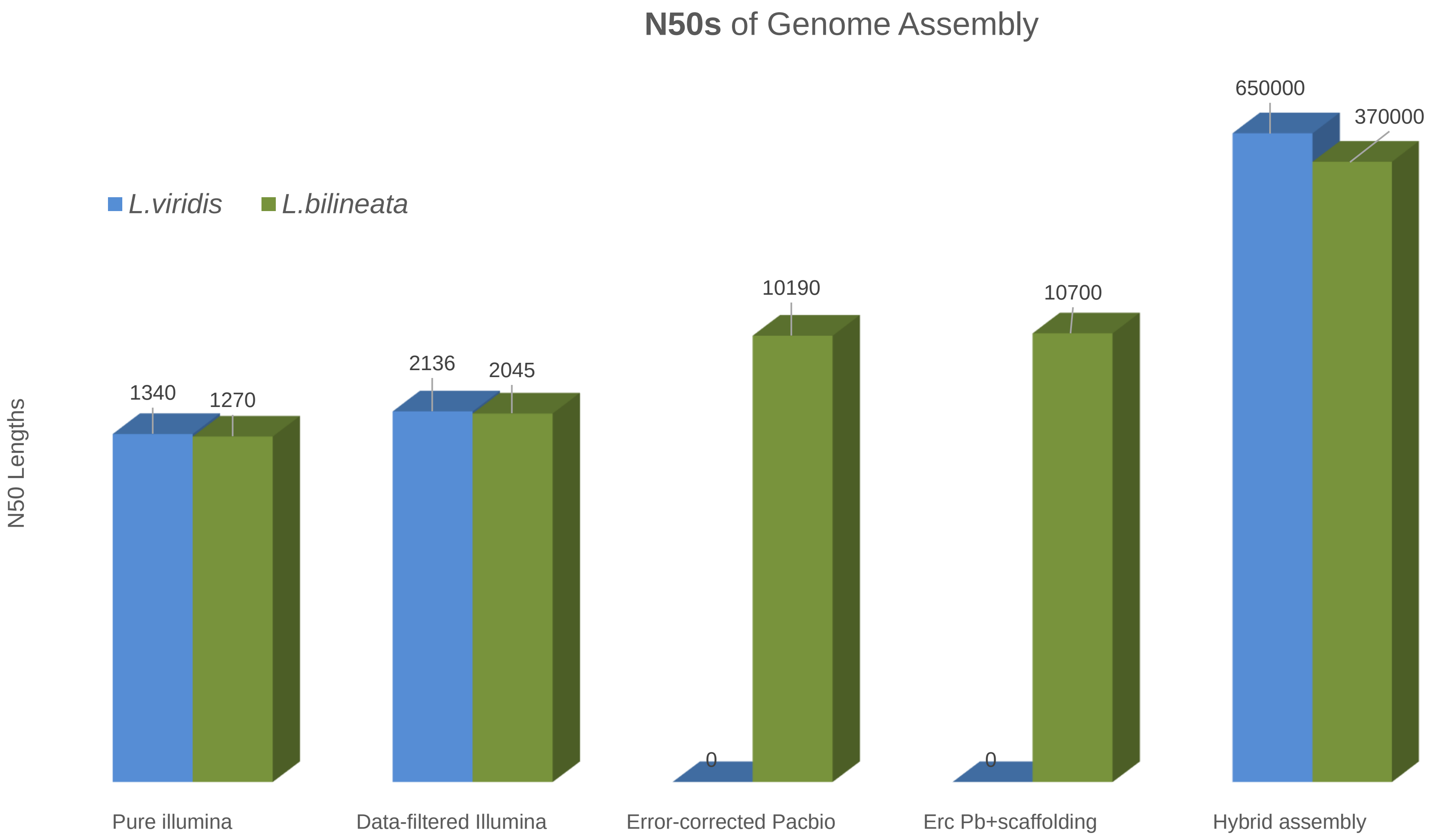
Assessing genome assembly and annotation completeness with Benchmarking Universal Single-Copy Orthologs

- Preliminary Genome annotation
- Single-copy ortholog genes

  arthropods, vertebrates, metazoans, fungi, eukaryotes, and bacteria
- Gene prediction with training set
- Completeness of assembly

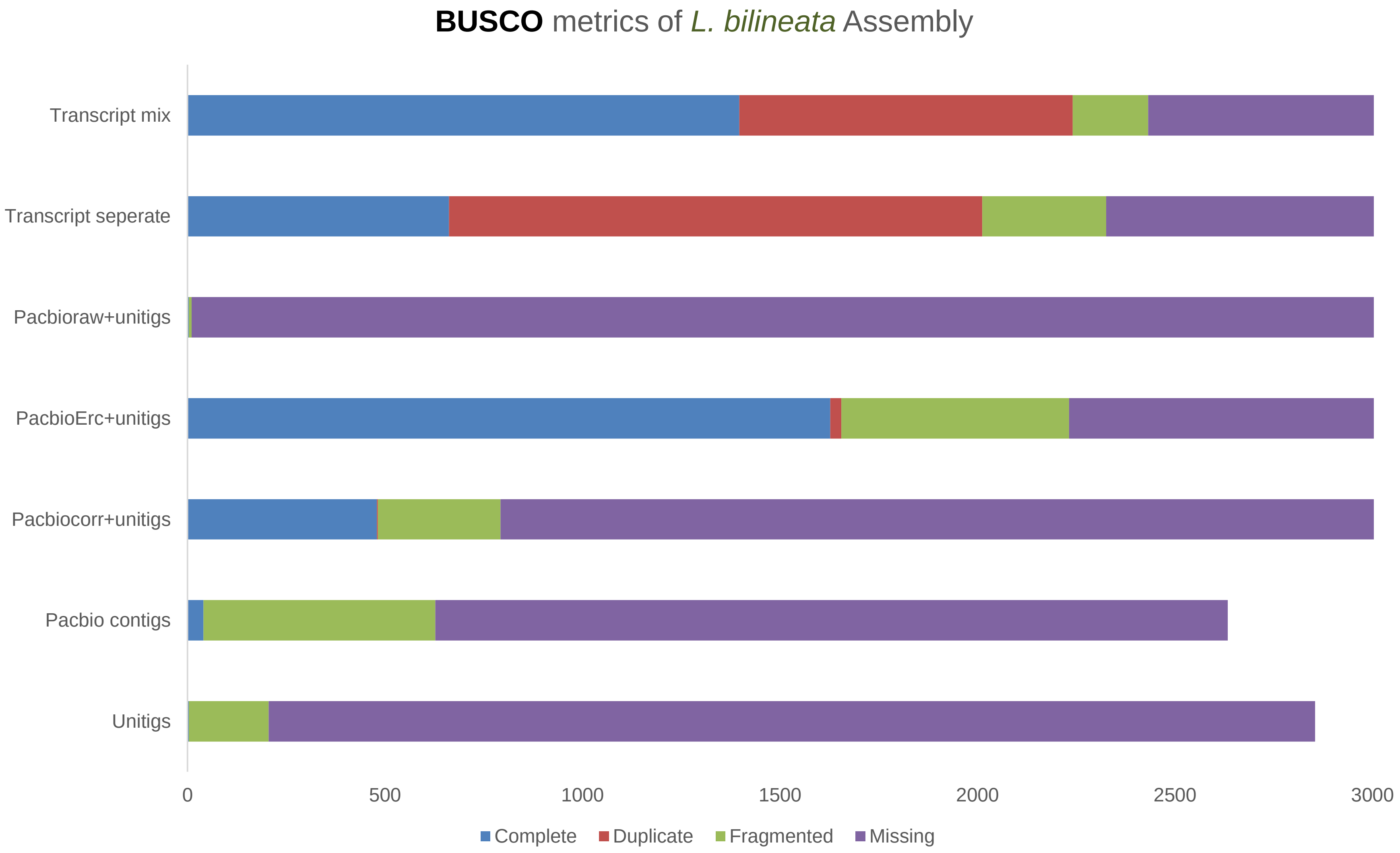
  Total, complete, duplicate, fragmented, missing
- Compare assemblies

Genome Assembly metrics



DBG2OLC: **Efficient Assembly of Large Genomes Using the Compressed Overlap Graph**

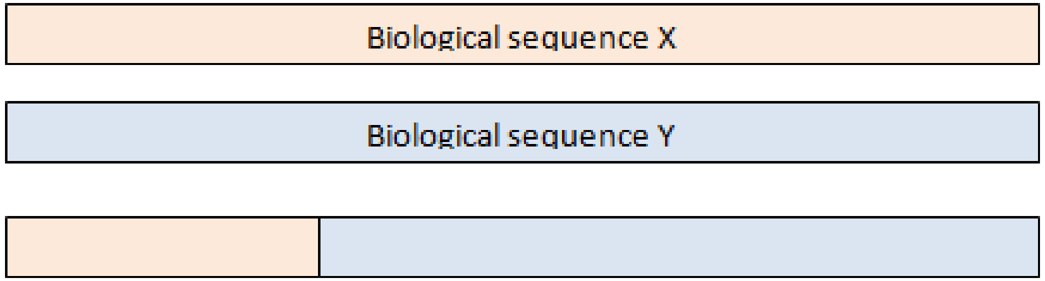
Assembly validation



Chimeric Presence

- Detection of genomic chimeras from Transcript assemblies
- Over 200 chimeras from Transcript mapping

  A case of trancript chimeras ???
- Mitogenome and NUMTS
- Variant calling and Consensus calling from the variants



Chimera formed from X and Y

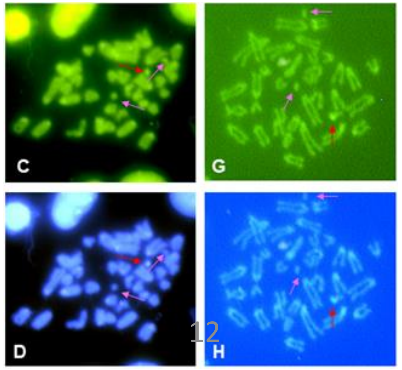
http://drive5.com/usearch/manual/chimera_formation.html

Take home messages...

- Genome assembly is far superior with **hybrid** approaches

  Intermediate corrected reads are best
- Atleast 10X PacBio and 20X illumina required, never lesser
- Error correction – a major resource consuming process
- Sanity check improvement for *de novo* genomes
- Chimeric relief from Transcript assemblies
- The futuristic NGS – Hybrid approaches with HiC

- Abundance of gene families in Lacertids
- Rearrangement detection via **syntenies** for non-reference quality genomes
- **W-chromosome** for lizards – a hotspot for divergence



iDiv

DFG Deutsche Forschungsgemeinschaft

PAC-men

- *Prof. Dr. Peter F. Stadler* (Universität Leipzig)
- *Prof. Dr. Martin Schlegel* (Universität Leipzig)
- *Dr. Katja Nowick* (Universität Leipzig)
- *Prof. Dr. Klaus Henle* (UFZ Leipzig/Halle)
- *Dr. Rui Faria* (CIBIO-University of Porto)

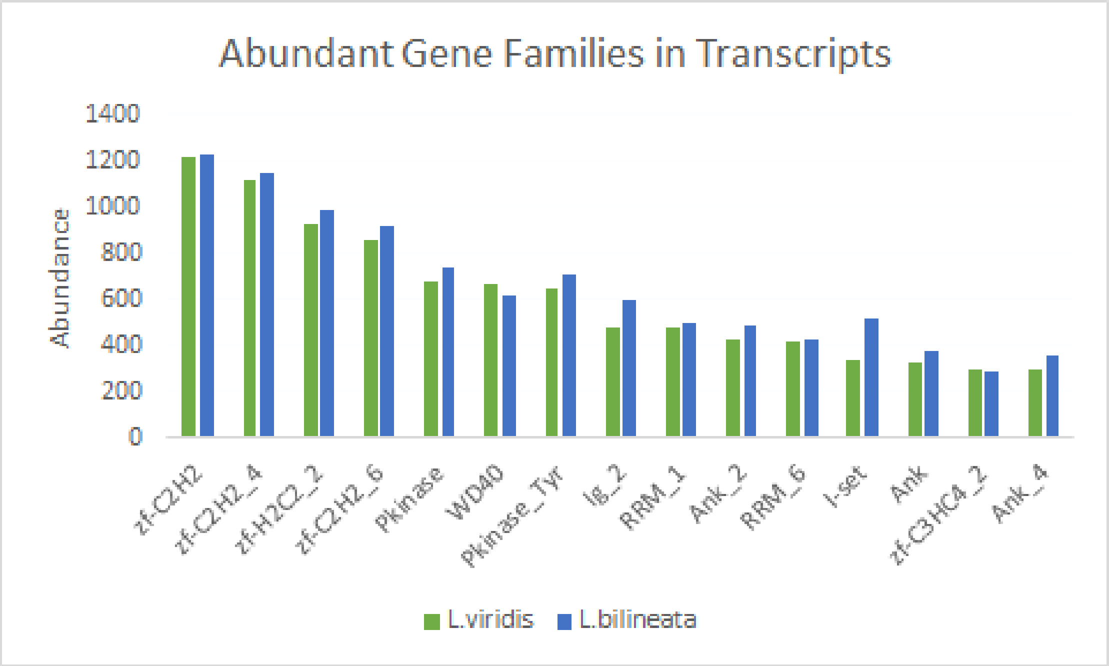
Braunschweig Sequencing group - Prof. Dr. Jorg Overmann

Experts at work
Steve, Matthias, Nowicklab, Heni, Sarah
Dr. Bleidorn, Anne, Micha, Stefan

System admins
Jens, Christian

“I kindly acknowledge the support by the German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, funded by the German Science Foundation (FZT 118).”

An overview of Gene Families



	Name	Description
	zf-C2H2	Zinc finger, C2H2 type
	zf-C2H2_4	C2H2-type zinc finger
	zf-H2C2_2	Zinc-finger double domain
	zf-C2H2_6	C2H2-type zinc finger
	Pkinase	Protein kinase domain
	WD40	WD domain, G-beta repeat
	Pkinase_Tyr	Protein tyrosine kinase
	Ig_2	Immunoglobulin domain
	RRM_1	RNA recognition motif
	Ank_2	Ankyrin repeats (3 copies)

For strictly orthologous transcripts, major differences are in

Zinc fingers – LIM, C2H2_2, F-box, H2C2_5

Repeats – Leucine rich, Ankyrin, Tetratricopeptide, Collagen triple helix

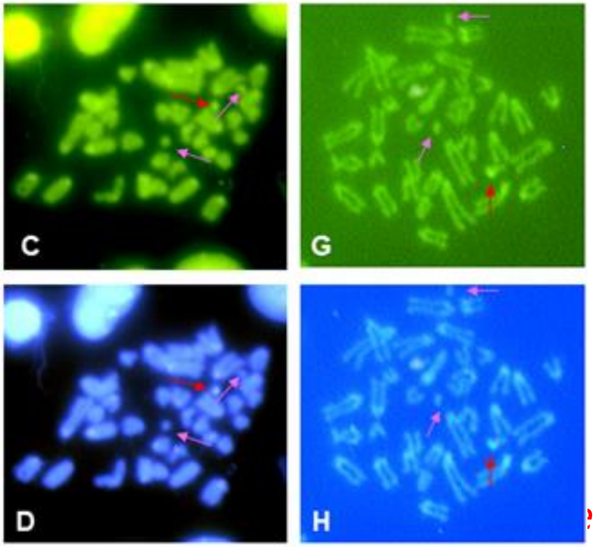
Immunoglobulins – I-set, V-set, Ig_3

Helicase – Dead box, C-terminal

EGF, transmembrane receptor (rhodopsin), SH3, SH2, kelch, SRPRB, MFS

Chromosomal structures in *bilineates*

Dr. Gaetano Odierna’s work showed marked morphological differences of W-chromosomes between L. viridis and *L. bilineata* and in-between *bilineates* (unpublished)



Pink arrows –> micro chromosome

Proovread

```
bwa mem -b 20 -l 300 -w 40 -B 11 -k 12 -T 2.5 -O 2,1 -Y -t 4 -E 4,3 -y 20 -L 30,30 -a  -r 1 -W 20 -A 5 -D 0
```

- -k 12 seed length
- -w 40 band width (gap length)
- -r 1 reseeding value (K*r), accuracy affected
- -b 20 bin size [PROOVREAD ADD]
- -B 11 mismatch penalty (high as base-errors in pacbio are not common)
- -l 300 max bp in bin [PROOVREAD ADD]
- -T 2.5 per base minimum score to ouput
- -O 2,1 gap open penalties for deletion-insertions (low as pacbio has high indel erros - more insertions than deletions)
- -Y use softclipping for supplementary alignments
- -E 4,3 gap extension penalties
- -y 20 seed occurance for 3rd round seeding [PROOVREAD ADD]
- -L 30,30 penalty for 5' and 3' clipping [PROOVREAD ADD]
- -a output all alignments for single end or unpaired (pacbio is SE)
- -W 20 discard chain if seeded bases are shorter than this (useful when seeding neighboring chains)
- -A 5 score for sequence match, this scales options 'TdBOELU' [bwa origin, proovread explained]
- -D 0 drop chains shorter than FLOAT fraction of longest overlapping chain (all chains covered by longest overlapping chain are dropped)
- Insertions are 2X of deletions in Pacbio i.e. 10% and 5% , substitutions are 1%

WHAT WE HAVE

Chimp sub-species assemblies at **6X** (testing at 30X)

20kb sequence lengths, **accuracy** to **1bp**

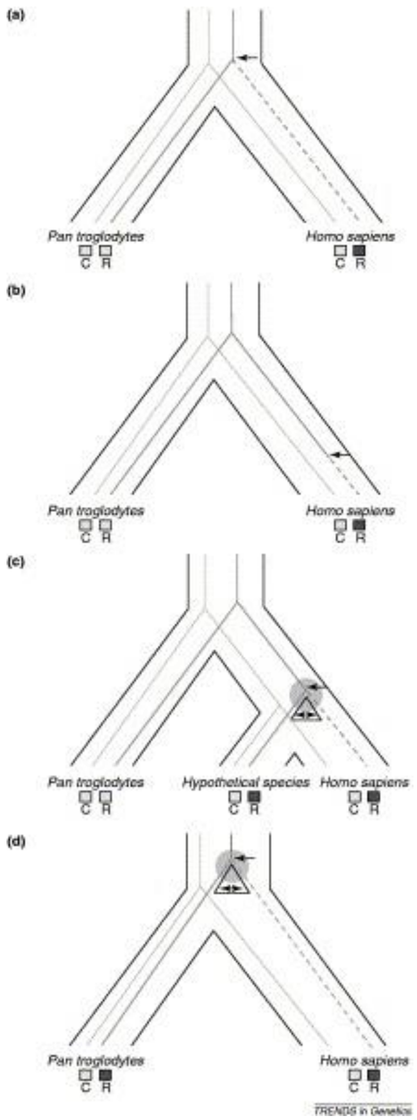
Alignment drop after **300 bp**, **split-hits** (testing)

Problem with **multiple inversions** – Scenario depiction

Convergence with read-based methods

- **Pairwise** alignment
- Uncovered regions and **missing** information
- Divergence hotspots

??? Scoring schema or accuracy points



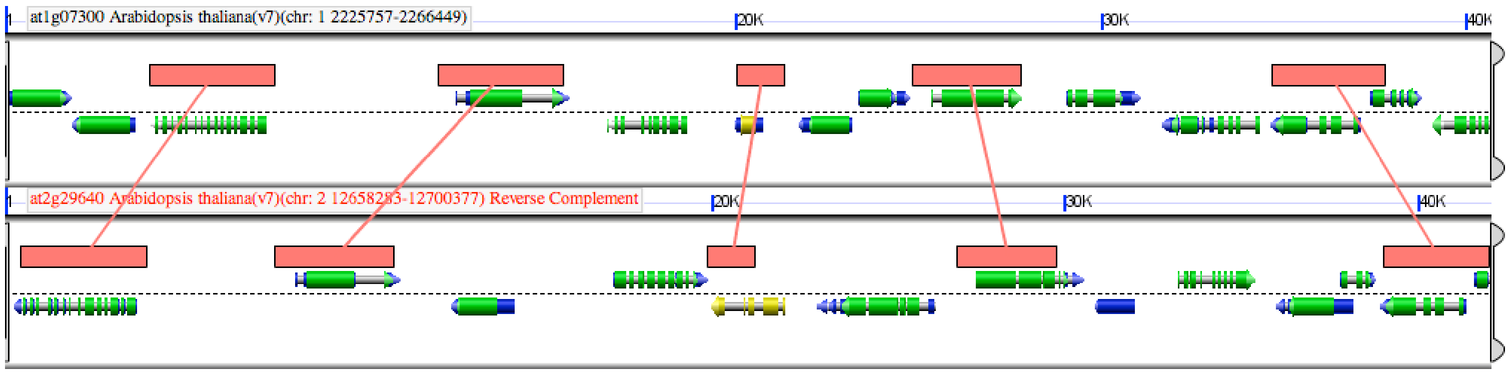
Marquès-Bonet, T., Cáceres, M., Bertranpetit, J., Preuss, T. M., Thomas, J. W. and Navarro, A.
Marquès-Bonet, T., Cáceres, M., Bertranpetit, J., Preuss, T., Thomas, J., & Navarro, A. (2004). Chromosomal rearrangements and the genomic distribution of gene-expression divergence in humans and chimpanzees.
Trends In Genetics,*20*(11), 524-529. doi:10.1016/j.tig.2004.08.009

Near-complete assemblies

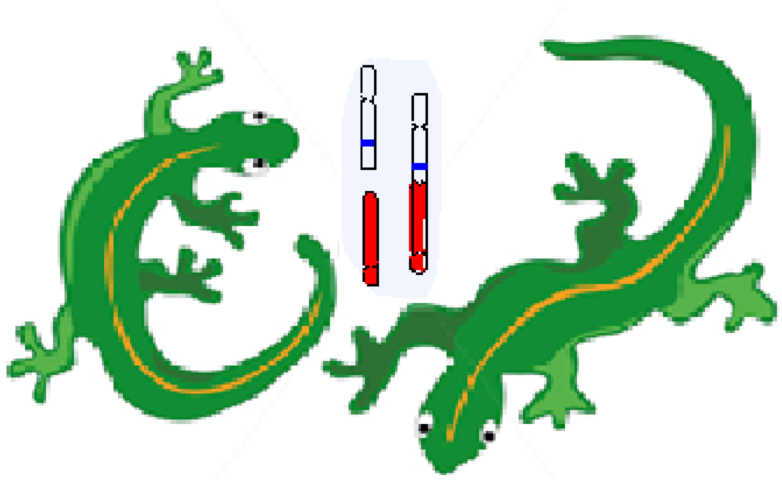
- **LAST** alignment after synteny finding

Parse PSL files

- Double genome fragmentation **unnecessary**
- **bwa mem** for alignment, **back-tracking**
- Pipeline with unaligned contigs
- Annotation
- **Graph representation** of similarity hits and weights by features


1
at1g07300 Arabidopsis thaliana(v7)(chr: 1 2225757-2266449)
20K
30K
40K
1
at2g29640 Arabidopsis thaliana(v7)(chr: 2 12658253-12700377) Reverse Complement
20K
30K
40K

Lizard project



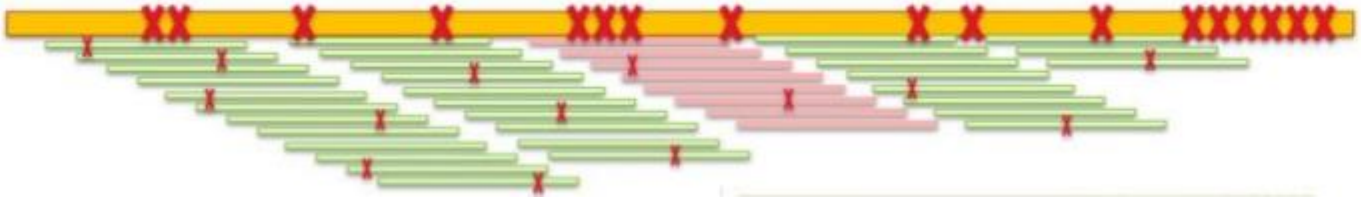
•Illumina assemblies and unitigs

•PacBio for longer ranges

•5bp indel acceptance

PacBio introduces indels sometimes after error correction

Variant calling and alternative reference generation



Koren, S., Schatz, M. C., Walenz, B. P., Martin, J., Howard, J. T., Ganapathy, G., Wang, Z., Rasko, D. A., McCombie, W. R., Jarvis, E. D. and Phillippy, A. M.

Koren, S., Schatz, M., Walenz, B., Martin, J., Howard, J., & Ganapathy, G. et al. (2012). Hybrid error correction and de novo assembly of single-molecule sequencing reads. *Nat Biotechnol*, *30*(7), 693-700. doi:10.1038/nbt.2280

Divergence and hotspots

•Divergent regions represented by regions that are completely missing from alignment blocks

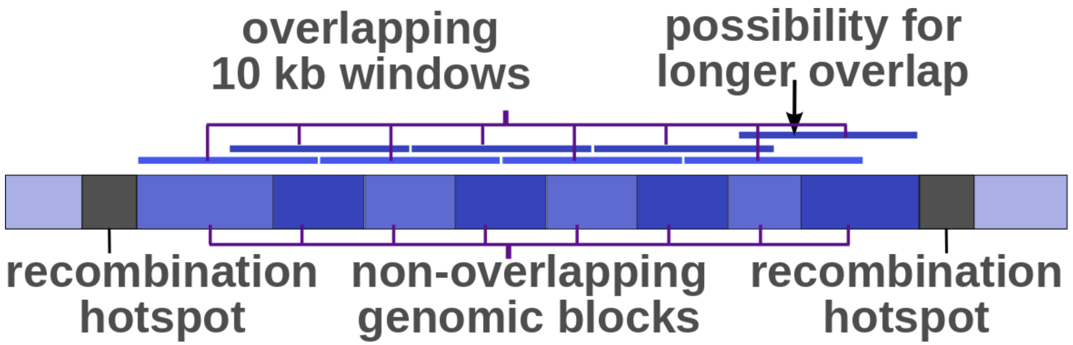
•Window based approach?

Look for mismatches in windows or the gaps observed

•Statistical support for the findings

•Splicing patterns

Transcriptome support



http://compgen.cshl.edu/INSIGHT/humanData.php