Reference-guided de novo assembly: improved genome of a non-model plant species

URPP Evolution in Action

University of Zurich[UZH]

Heidi E.L. Lischer[1,2,3], Damianos Melidis[2], Masaomi Hatakeyama[2,4], Kentaro Shimizu[1,2]

[1]URPP Evolution in Action; [2]Institute of Evolutionary Biology and Environmental Studies (IEU), University of Zurich, Switzerland; [3]Swiss Institute of Bioinformatics (SIB), Lausanne, Switzerland; [4]Functional Genomics Center Zurich, ETH Zurich, Switzerland

Swiss Institute of Bioinformatics

Problem

- De novo assembly of genomes is still challenging:
  - Short read length
  - Repetitive regions
  - Uneven coverage, even missing data
  - Polymorphisms and sequencing errors
  - High computational resources required

Solution

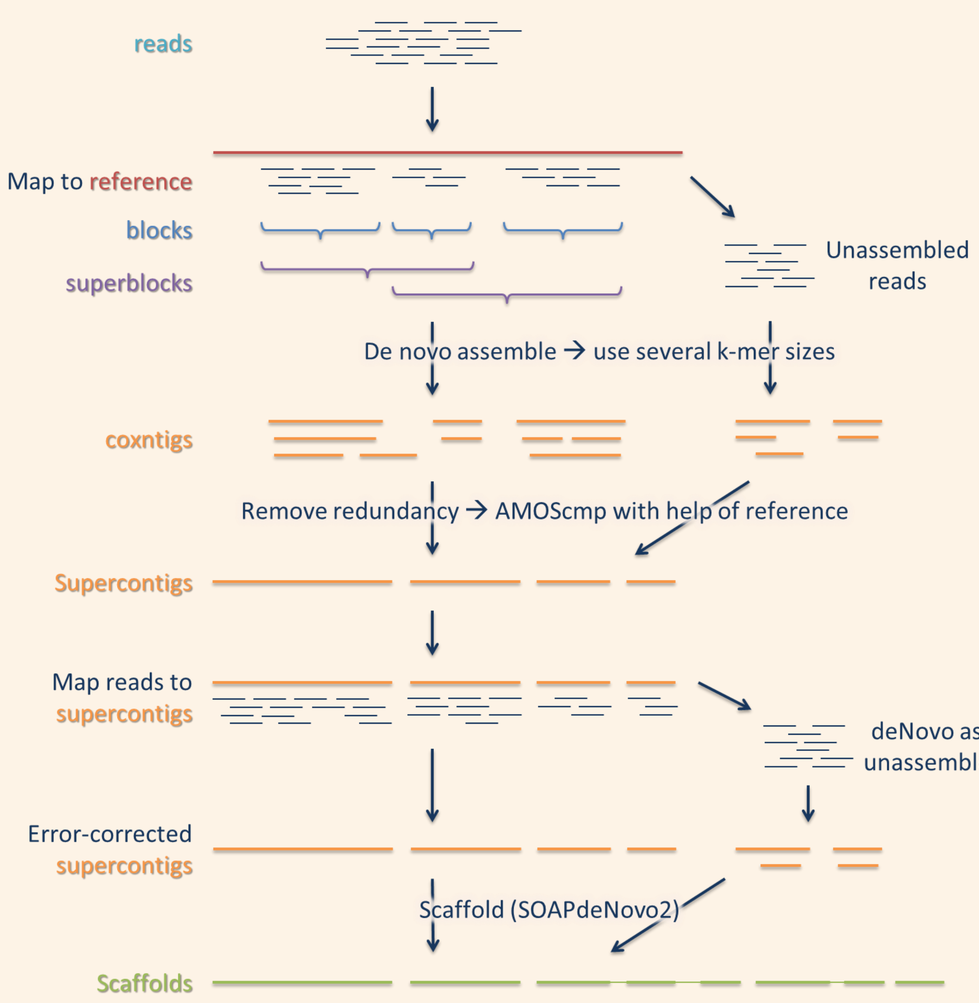
- Reference-guided assembly approach: use an available reference sequence from a close relative to improve genome assembly
  (adapted and extended from Schneeberger *et al.* 2011 [1])

Conclusions

- Reference-guided de novo assembly approach
  - Reduces K-mer problem
  - Less memory resources required, but longer runtime
  - Improved genome assembly, especially in genic regions
  - → The closer related the reference genome the better result is expected
- Idba:
  - Best genome assembly results, but requires 2-3 times more memory resources compared to reference-guided approach with ALLPATHS-LG
- *Arabidopsis halleri*:
  - Improved genome with reference-guided de novo assembly approach
  - Idba de novo assembly leads to less good results in real data set

Methods

Reference-guided de novo assembly pipeline:



Evaluation:

- Target genome: Simulate reads from *Arabidopsis lyrata* (204 Mb)
  - 1% heterozygosity
  - 330 million paired-end reads (150, 200, 500 bp insertion length)
  - 610 million mate-pair reads (3, 5, 7, 11, 15 kb insertion length)
- Reference genome: *Arabidopsis thaliana* (120 Mb)
  - >80 % overall sequence identity with *A. lyrata*, but >50% missing [2]

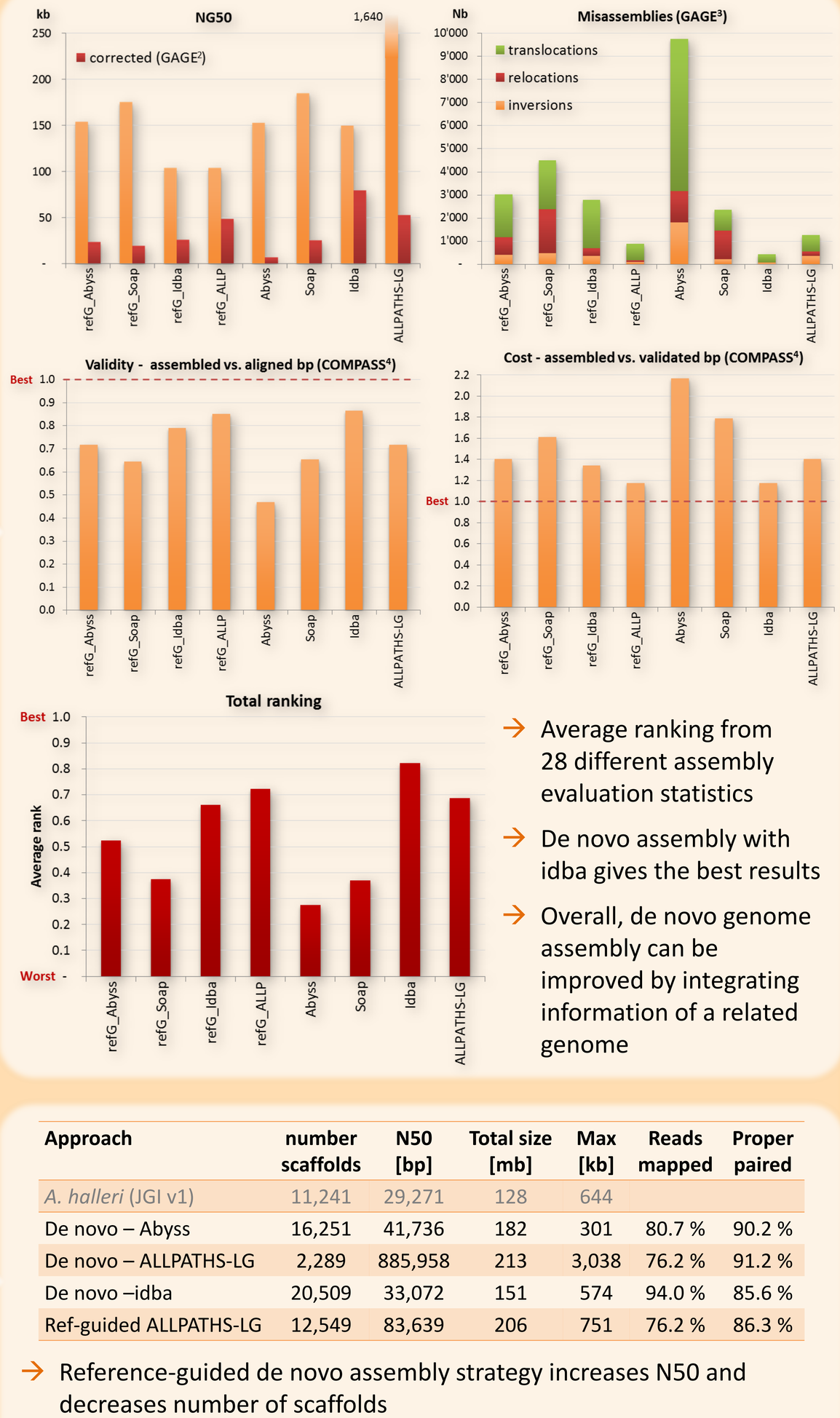
Application to real data set - *Arabidopsis halleri*

- Important model species to study environmental adaptation and polyploid speciation
- *A. lyrata* genome available, but *A. halleri* is missing

A. lyrata (LL) x *A. halleri* (HH)
↓
A. kamchatica (LLHH)

- Resources:
  - 300 million paired-end reads (150, 300, 750 bp insertion length)
  - 850 million mate-pair reads (3, 5, 7, 11, 15, 22 kb insertion length)
  - Reference genome: *Arabidopsis lyrata*

Results



- → Average ranking from 28 different assembly evaluation statistics
- → De novo assembly with idba gives the best results
- → Overall, de novo genome assembly can be improved by integrating information of a related genome

Approach	number scaffolds	N50 [bp]	Total size [mb]	Max [kb]	Reads mapped	Proper paired
A. halleri (JGI v1)	11,241	29,271	128	644		
De novo – Abyss	16,251	41,736	182	301	80.7 %	90.2 %
De novo – ALLPATHS-LG	2,289	885,958	213	3,038	76.2 %	91.2 %
De novo –idba	20,509	33,072	151	574	94.0 %	85.6 %
Ref-guided ALLPATHS-LG	12,549	83,639	206	751	76.2 %	86.3 %

- → Reference-guided de novo assembly strategy increases N50 and decreases number of scaffolds
- → Improved genome will facilitate the study of polyploid speciation

References

[1] Schneeberger K, et al. 2011. Reference-guided assembly of four diverse *Arabidopsis thaliana* genomes. PNAS 108: 10249-10254
[2] Hu TT, et al. 2011. The *Arabidopsis lyrata* genome sequence and the basis of rapid genome size change. NatGenet 43: 476-481
[3] Salzberg SL, et al. 2012. GAGE: A critical evaluation of genome assemblies and assembly algorithms. GenomeRes 22: 557-567
[4] Bradnam KR, et al. 2013. Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. GigaScience: 2: 10

Contact: heidi.lischer@ieu.uzh.ch

Funding: URPP Evolution in Action

Link to poster: