Sequencing the genome of the French wheat variety Renan
29 September 2022
— — —
Q&A session

Presenter: Frédéric Choulet (GDEC, INRAE, University Clermont Auvergne, France)

The webinar recording is available on the IWGSC YouTube channel at: https://youtu.be/JvJKIcPDhZs

---

Q: based on this experience, would you recommend PacBio or Nanopore for the initial long-read sequencing?

Timestamp: 51:52

Q: In wild genomes, still timopheevi genome sequences are n9t available on IWGSC. any idea if iwgsc is working on it or not

In 2020 Walkowiak et al. produced reads of timopheevii genome that were deposited online on SRA (PRJNA544491)

Q: Among the 160k pseudogenes did you find any matches with annotated genes in other related cultivars? I guess you mean "Low confidence genes" right? Yes, most of them are conserved between cultivars, but wheat cultivars are very close phylogenetically, so sequence conservation between cultivars is not evidence for functionality (i.e. selection pressure). So the question of the functionality of some predicted genes remains

I guess you mean "Low confidence genes" right? Yes, most of them are conserved between cultivars, but wheat cultivars are very close phylogenetically, so sequence conservation between cultivars is not evidence for functioanlity (i.e. selection pressure). So the question of the functionality of some predicted genes remains

Q: If the goal is to have more sequenced genomes than CS to get an idea about the pan-genome, is it not a problem to base assembly and annotation of Renan on CS? Meaning that you claim that woth this approach the outcome looks much better, but that is because you compare it to CS structure/annotation and you assume the same is true for other varieties.

Timestamp: 1:00:17

Q: Are renan genome sequences available on IWGSC?

Yes, they are now freely available at INRAE-URGI through a Jbrowse.

Q: The presented genome assembly has combined several approaches, like ONT, optical mapping and Illumina sequencing. Winter wheat variety Renan was reported to have three copies of vernalization gene VRN-A1, located on the chromosome 5A (Würschum et al., 2015). Using Renan sequencing data, we have found one copy only. Is it possible that regions with other two copies collapsed during the assembly? Do you think it would be possible to find the remaining VRN-A1 copies employing the raw data?

Timestamp: 1:05:22
Thanks for feedback on the Renan genome assembly. We noticed that and work on it. We will come back to you with an answer

Q: GrainGenes created a genome browser and BLAST service (nucleotide/protein) for Renan: https://wheat.pw.usda.gov/GG3/content/august-2022-released-wheat-cultivar-renan-browser-and-blast
Thanks for the link and the great job

Q: How did the error rate in reads and assembly was calculated in Nanopore sequencing data?
Nucleotide sequence conservation between wheat cultivars is close to 100%. Let say higher than 99.5% for orthologous loci. So, in the early stages of the project, we used Chinese Spring genes and TE-derived markers (uniquely mappable) to estimate the % differences in reads and contigs that gave us good estimate of error rates. At the end, when problems were solved, we estimated the global QV (Quality Value) with Merqury (https://genomebiology.biomedcentral.com/articles/10.1186/s13059-020-02134-9)

Q: Which software was used for whole genome alignment and visualization?
Timestamp: 55:15
BLAST for alignments, homemade bash scripts and R for data manipulations, and Artemis Comparison Tools for viewing

Q: Are there any publications addressing the difficulty of de novo gene annotation in wheat that you could recommend?
We mentioned these problems in the Chinese Spring RefSeq v2 paper Zhu et al. Plant Journal 2021 (https://onlinelibrary.wiley.com/doi/10.1111/tpj.15289).
Actually, we have written a book chapter specifically on wheat genome annotation which addresses these limits and difficulties. It is currently under peer-review process, and hopefully will be published soon.

Q: Would a Hybrid assembly improve the accuracy over a nanopore-only contig?
Timestamp: 54:00
Yes, and we actually used Illumina reads in order to polish the Nanopore-only contigs.

Q: Will genetic maps play a role in future assembly projects (e.g. for pseudomolecule construction)?
Timestamp: 58:26
For wheat genomes, I think no. HiC method works pretty well, is not limited by absence of recombination, and does not require production of populations, so genetic maps will probably not play a role in future wheat genome assemblies (i.e. contig/scaffold anchoring).

Q: Since there are new basecalling algorthims being developed for ONT, do you expect that this would solve more problems in wheat sequencing?
If one uses ONT for a wheat genome assembly, I think yes, basecalling improvements would be a fantastic solution to limit polishing steps, avoid producing Illumina reads, and avoid nucleotide errors in final assembly. If we imagine these improvements would allow to achieve >99% read accuracy, it would even open a scenario where one would favor ONT to Pacbio.

Q: Could you please touch on the DNA extraction optimization for obtaining long reads?
My group at GDEC Clermont was not involved in that part of the work. It has been taken in charge by collaborators at CNGRV INRAE Toulouse (A. Bellec's lab). I invite you to get in touch with the team there.

Q: Given turnover rates of TE, what are the limits of using ISPB from CS to other wheat cultivars and species? For example did the Ae. ventricosa introgressions show more errors than non-introgressed regions?

The TE turnover has erased ancestral TEs when you compare A versus B or versus D, because they have diverged long time ago. In other terms, a TE inserted on A subgenome is not conserved (because ancestrally present) in B or D. This is why an ISBP marker is specific to one single locus in the whole hexaploid genome.

Now, when you use ISBP markers to find correspondence between 2 genomes of 2 cultivars, it is completely different because what you are looking for is a conservation of markers between A and A, B and B, or D and D. For instance, chr1A of Chinese Spring and chr1A of Renan share ~90-95% of ISBP markers.

Regarding, Ae. ventricosa, it is a tetraploid with 2 subgenomes N and Dv. ISBP markers in the N subgenome are (in most cases) not conserved with any other locus in A-B-D. However, this is different for the 7D introgression because 7D from ventricosa shares TEs in common with 7D from T. aestivum.

Q: if you have 5 Mb target interval and illumina contigs at 30x coverage, what kind of ONT coverage could allow assembly of the target interval?

30x coverage is what we targeted and this is what I also recommend.

Q: A new version V2.1 of Renan assembly has been recently made available. Do you know which kind of improvements were made compared to previous V1 version ? and are you planning to transfer the annotation onto this new version ?

Changes are only minor. Yes, transfer of annotation has already been performed.

Q: Is heterozygosity an issue for future wheat assemblies, is it dependent on strain, such as level of inbreeding. Would you anticipate haplotype assemblies being needed in the future?

For the moment, I have never faced problems due to heterozygosity in bread wheat genome sequencing. I imagine that may happen with some wild Triticeae species and we may have to consider higher coverage and haplotype resolution.

Q: Which version of IWGSC RefSeq should be used to predict genes underlying QTLs or MTAs regions?

I recommend to start considering several genotypes now because Chinese Spring is not the best representative cultivar for modern varieties. But definitely CS IWGSC RefSeq v2.1 version is still the reference to use because gene/TE annotation are available, markers from many origins were placed along the chromosomes, genetic maps and QTLs were integrated, and every data was made accessible through URGI public Jbrowse.

Q: Are TE regions better assembled in these long-read assemblies compared to the CS reference (over short-read assembled regions) and does this mean that ISPB method could be improved with a long-read reference or improved CS reference based on long-reads?

I would say that TE regions are well assembled in both. The advantage with long reads is that there are fewer gaps. Although the %Ns is limited to a few %, these assemblies are still not complete (i.e., N free) and adding long reads would help achieving even higher quality. Actually, for the CS reference, we already added long reads when releasing the v2 assembly which includes bionano maps + pacbio contigs. Please see Zhu et al. Plant Journal 2021.