CSHL Assemblathon Meeting Summary

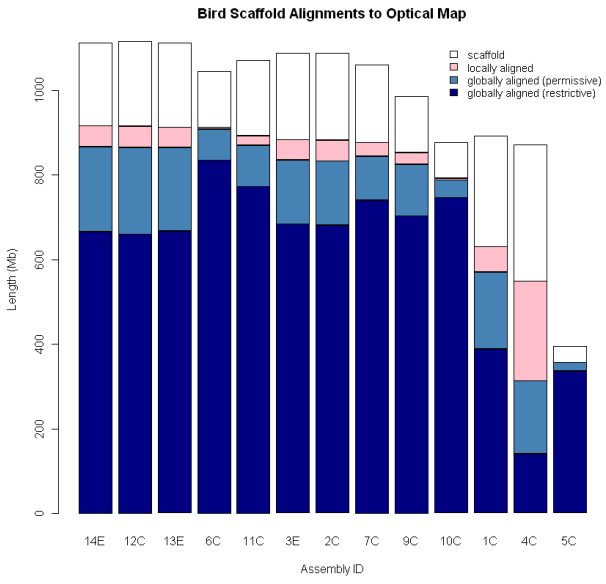
The Assemblathon meeting took place on the afternoon of November 5 2011 at Cold Spring Harbor Laboratory. The meeting was attended by 26 participants (see list at end).

The meeting began with a brief overview of the data sets, assemblies, and evaluation plans:

- Genomes
  - Snake, *Boa constrictor constrictor*, (Illumina)
  - Fish, *Maylandia zebra,* (Illumina)
  - Bird, *Melopsittacus undulatus*, (454, Illumina, PacBio)
- Assemblers
  - 21 teams
  - 43 assemblies
- Evaluation Plans
  - Optical maps
  - Random fosmid sequencing
  - Conserved proteins
  - Other sources

We then examined some of the evaluation data.

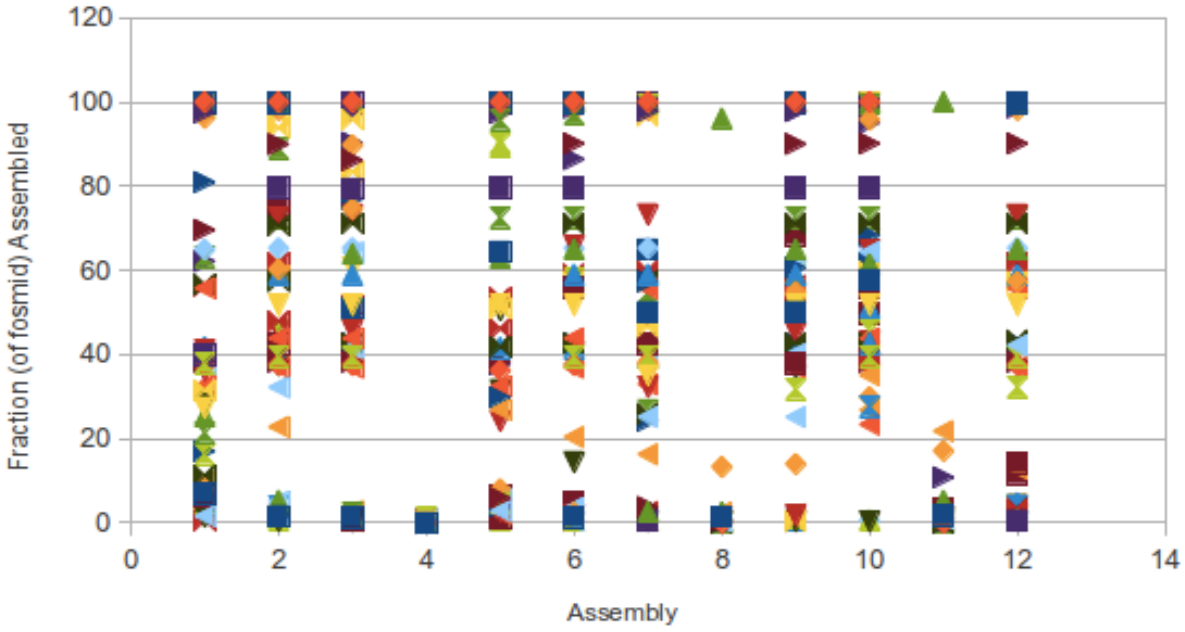
Optical Maps



Optical mapping has been completed for the bird genome by David Schwartz, Steve Goldstein and colleagues from University of Wisconsin. Technical difficulties have delayed the snake and fish analyses, but these are expected soon. Optical maps are useful for evaluating long range genome structure. Single molecule restriction maps are aligned to each other to create a genomic optical map, and genome assemblies are compared to this map. In the diagram at the right, dark blue bars show nearly perfect alignment between genome and optical map. Light blue bars correspond to lower stringency alignments. Pink is a local alignment method that is more tolerant of structural rearrangements (but still in agreement with parts of the optical map). Assemblies are coded on the X-axis (we are not disclosing which assembly is which at this time). The optical maps show a wide range of variation between different assemblies.

Fosmid Sequences

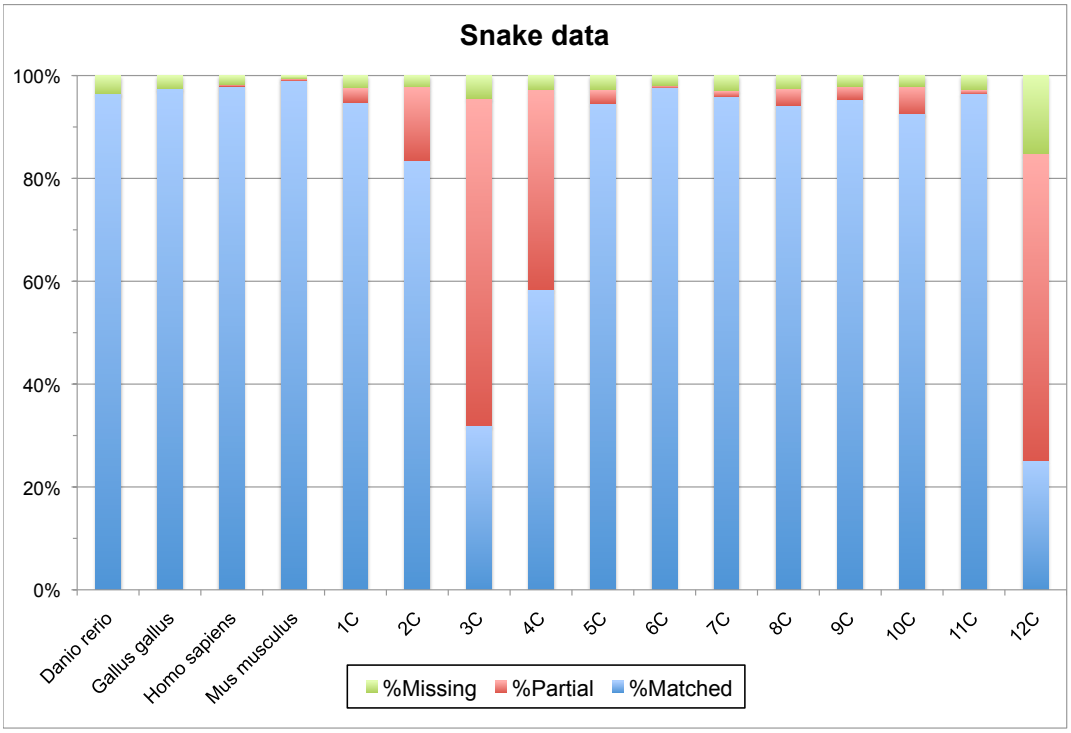
Pools of fosmids were sequenced by Jay Shendure's group at the University of Washington. These were assembled (with Velvet) and aligned (with BLAST) to the genome assemblies by the UC Davis group. Assembling complete individual fosmids is much simpler than assembling a genome, but still complicated. However, many fosmid-sized sequences emerged from the assemblies. The fact that many of these were fully aligned to the whole genome assemblies confirms that some of the pooled fosmids were assembled correctly. Curiously, the performance was highly variable. Some fosmid sequences appeared in the assemblies of one group but not others. For example, in the graph below, the blue box at the top right shows that assembly 12 was in perfect agreement with the blue-box fosmid. This fosmid sequence was not present at all in assembly 3. It is not clear why there is such variability. See below for further developments on fosmids.



Conserved Proteins

All vertebrate genomes share some highly conserved proteins in common. The UC Davis group collected 539 orthologous proteins from human, mouse, chicken, and zebrafish that share the following qualities: >=75% identity in pairwise alignment, >= 90% same length, >=70% identity in a 4-way multiple alignment, <75% identity to any known paralogs. These 539 proteins are easily aligned by BLASTX (although the shorter ones do not always align under the post-processing criteria). The graph below shows data for the snake genome. The set of conserved proteins are aligned to their original source genomes in addition to the the submitted snake assemblies. Blue bars

represent the fraction of proteins that are completely aligned; red indicates partial alignment; and green corresponds to proteins missing from the assembly. Several of the assemblers perform very well, and the resulting genome assemblies are as good as the reference assemblies, at least with respect to this evaluation metric.



Group Discussion

Our next agenda was a "Soapbox Session" where anyone with an assembly issue was invited to discuss this with the group. The plan was to use this time to kick start the following group discussion. We ended up with four groups: Genome Representation, Assemblathon 2.x, Assemblathon 3, and Combining/Merging Assemblies. Each is discussed below.

Genome Representation

One of the most important problems to solve is how to represent genome assemblies in text files. FASTA is a convenient and pervasive file format, yet it does not allow assemblers to indicate heterozygosity and/or structural ambiguity. This topic has been under discussion for some time by David Haussler, Richard Durbin, David Jaffe, Daniel Rokhsar, *et al*. It was agreed that there would be a new email list for those interested in this area (to be announced shortly via the Assemblathon website and the current

mailing list). The tentative file type is FASTG. A draft specification is expected by 2012-02-15 in time for the AGBT (Advances in Genome Biology and Technology) conference. It is likely that future Assemblathons will require assemblies in this format.

One of the problems associated with FASTG (or an equivalent) is that initially, there will be little support for existing software that works with sequence data. It will be necessary to spend some time re-engineering existing tools like BLAST and developing new tools that specifically take advantage of the new file format.

Assemblathon 2.x

The initial results of Assemblathon 2 show more variability between assembly groups than between data sources. This indicates that there are large differences in software decisions and that leveraging multiple data sources is non-trivial. It is likely that additional software development will result in better genome assemblies without generating any new data. It would therefore be possible to issue the same competition repeatedly as Assemblathon 2.1, 2.2, etc. While this would show the progress of the field and would provide better assemblies for these genomes, it creates more work for assembly teams, and it reuses old data rather than generating improved data. One very useful study that could be performed as Assemblathon 2.x, is to factor in sequencing and assembly costs as part of the competition.

Assemblathon 2.0 is not yet finished, and it is probably a good idea to focus on 2.0 before being distracted with 2.x or 3.0. The optical maps are not yet complete and the fosmids are disturbingly variable. The JGI has kindly offered to sequence ~20 fosmids with Sanger technology. This would provide an excellent assembly reference sample. Another valuable resource would be to have BAM files mapping all reads to the assemblies. Kim Worley at Baylor College of Medicine is coordinating this effort.

Assemblathon 3.0

New assembly targets the group would like to see include RNAseq, metagenome, pooled BAC/fomid, and mouse Black6 (using existing data). Additionally, plant genomes were suggested because they offer some unusual challenges due to size and ploidy.

One of the most important aspects of genome assembly is library preparation. The Assemblathon 3.0 group would like to see a competition based on differences between the various methods used for sample preparation. Certainly there is variability in this procedure that affects downstream sequencing and assembly. The costs to run such a study are probably prohibitive, so this is considered a low priority target at this time.

Merging/Combining

In the Merging/Combining subgroup, we made plans for merging the top assemblies of parrot (and later the other genomes) to create the best possible "meta-assembly". The basic strategy would be to make pairwise alignments between all assemblies and then

identify conflicting regions and gaps. For each conflicting region, a statistical test of the mate-pair separation and optical map agreement would be applied to decide which version has better evidence. After evaluating the conflicts, the best results would be combined into a composite assembly. This process could possibly be repeated to build a progressive meta-assembly of multiple draft genomes.

Once the sequence assembly is refined, we then plan to incorporate the optical map information to build optical/sequence hybrid scaffolds. These results look extremely promising suggesting that scaffold N50 lengths of over 40 Mbp may be possible (basically entire chromosomes or chromosome arms in individual scaffolds). Our timeline is to have a refined assembly available mid-Feb in time for the AGBT meeting.

Participants

Paul Baranay, Mario Caccamo, Hamidreza Chitsaz, Wen-Chi Chou, Joseph Fass, Steve Goldstein, David Jaffe, Erich Jarvis, Brian Kelly, Paul Kersey, Paul Kitts, Aaron Klammer, James Knight, Sergey Koren, Ian Korf, Zemin Ning, Adam Phillippy, Ricardo, Ramirez Gonzalez, Daniel Rokhsar, J. Graham Ruby, Surya Saha, Kristoffer Sahlin, Michael Schatz, Jared Simpson, Aarti Venkat, Kim Worley

Final Comments

Thanks to everyone who attended the meeting and to everyone on the email list. We are also grateful to those who have agreed to undertake new Assemblathon-related tasks in future. Finally, thanks to the NHGRI for providing supplemental funds for the Assemblathon 2 evaluation data.

The Assemblathon is a community effort and we all welcome your input. Sequence assembly has come a long way, but there is still a long way to go. If you are interested in assembly and are sitting quietly in the distance, please introduce yourself and become an active member.