4/26/05

# 4th Genome build for Xenopus tropicalis.

The new sequence data, beyond what was included in the last build (xentr3), are BAC and fosmid ends, and finished clones sequences. For this reason, and because localized re-assembly tests and parameter tuning have not revealed obvious ways to improve contig formation, this assembly is essentially a re-build of the scaffolds from the same contigs from the previous build according to the protocol described below.

## Dataset summary

| | Reads | Seq Depth (X) | Clone Depth (X) |
|---|---|---|---|
| 3 Kb | 9.4 M | 3.3 | 6.9 |
| 8 Kb | 11.0 M | 3.75 | 19.6 |
| 40 Kb | 1.9 M | 0.55 | 14.5 |
| BAC ends | 158 K | 0.05 | 3.9 |

## Placement of fosmid end, BAC end and EST clusters

The long insert clone end reads (fosmid and BACs ends) were aligned to the repeat-masked xentr3 scaffolds using BLAT.
Reads which either have a unique best placement on the scaffolds, or have two placements of roughly equal quality, and one of them is to a small scaffold (<10kb), are accepted as placed.
In order to link scaffolds together based on split gene sequences, a set of assembled EST sequences (JGI EST cluster build 20050214) were aligned to the scaffolds and processed as follows for inclusion as links by the assembler:
For each EST sequence aligned with BLAT, the top three hits are examined (the three hits with the most matching bases). If the third best hit has no more than half as many hits as the second best hit, and if the sum of the numbers of matches in the first two hits is less than the length of the EST, and if the length of the overlap of the footprints of the two alignments on the EST is less than 25 bases, then the EST is considered split between the first two hits. A virtual read pair is created, one read being placed in each of the two best-hit positions. These pairs are fed into the proceeding assembly steps as linking information in the form of a virtual shotgun library with nominal insert size of 10kb +/- 10kb.

## Initial scaffold breaking

All xentr3 scaffold sequences were virtually broken at every internal gap of length greater than 8kb. There were a total of 3830 breaks. This produces as set of broken scaffold sequences and a coordinate mapping for transferring placements on the xentr3 scaffolds to the broken scaffolds.

## Assembly

The broken scaffolds are ordered and oriented using the JAZZ layout engine based on the placement of clone end and EST sequences in a trial assembly.

## Additional scaffold breaks

Based on the results of the trial assembly, 310 potential mis-joins in xentr3 scaffolds were detected. These breaks were added to the original set of large-gap breaks, and the read placements were re-mapped to the updated set of broken scaffolds.

## Re-assembly

The broken scaffolds were ordered and oriented with JAZZ, and new scaffold sequences constructed.

## Incorporation of finished clone sequence

The new scaffold sequences were aligned with BLAT to a set of 128 finished clone sequences from the CH216 and ISB BAC libraries. See the file "finished_clone_sequences.fasta" for the sequences used and their GenBank accession information. The resulting alignments were tested for consistency with the *fortify* program (Ho, Skrainka, Putnam, unpublished) and the finished sequence inserted into the scaffold for accepted matches.

## Assembly statistics

| | |
|---|---|
| total length of scaffolds | 1513925492 |
| total length of contigs | 1359411890 |
| total number of scaffolds | 19759 |
| total number of contigs | 191480 |
| N50 scaffold number | 272 |
| N50 scaffold size (L50) | 1564123 |
| N50 net scaffold number | 240 |
| N50 net scaffold size (L50) | 1600435 |
| N50 contig number | 22314 |
| N50 contig size (L50) | 17021 |

## Current Resources

The resources being released at the time are located in the assembly release directory:

- xentr4.fasta : The scaffold sequences for the new assembly. Where the sequence has been taken directly from a segment of the xentr3 assembly, repeat masking has been preserved.
- contiguity_fig.ps : A figure comparing the contiguity of the current release to that of xentr3.
- finished_clone_sequences.fasta : the sequences of the finished clones which were available for incorporation into the assembly.
- xentr3_xentr4_coordmap.txt : coordinate mapping file which shows how xentr3 scaffolds and fragments of scaffolds are joined into xentr4 scaffolds.
- blast index files
- xentr4.fasta.*N50 : N50.pl output contiguity summary files.

## Resources still to be prepared

- Layout of the reads in the final coordinate system of the new scaffolds.
- Fasta file with the sequences of reads which were not assembled.
- Quality scores for the final scaffolds
- NSP inventory for the current assembly