Strategies for dealing with large scale Zebrafish repeats

Alan Tracey on behalf of the Wellcome Trust Sanger Institute.

Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, UK

As part of ongoing involvement with large genomes the Wellcome Trust Sanger Institute (WTSI) uses a clone based strategy to finish clones to phase 3. The majority of clones are successfully resolved with our normal strategies including autoprefinishing, custom oligos and alternative chemistries. Some of the clones which typically evade successful resolution by the normal methods include Zebrafish Large Tandem repeats (ZLTs). These are defined as tandem repeats >10kb which can be very difficult to construct using automated assembly programs such as Phrap (Green) in isolation and may even be resistant to preliminary manual finishing efforts. Such clones may warrant specialist attention in order to resolve the assembly problems. The WTSI uses the specialist skills of the most experienced finishers to tackle these clones and assess the likelihood of resolution before committing a significant amount of resources.
ZLTs occur in 3.3% of 9815 clones seen in finishing to date equating to approximately 330 clones, and of this only around 5% of ZLTs have been found to contain gene objects. The prevalence of non-coding ZLTs has led us to seek alternative strategies to resolve some of these regions. Pre-analysis (defined later**) is undertaken to select those repeats which warrant additional effort. Coding repeats are often tackled using transposon libraries generated from large insert subclones; a highly effective method for tackling large repeats. Non-coding repeats are scaffolded to agree with the repeat pattern and restriction digest data. This strategy gives a better representation of the clone than force joining which could remove valuable data or even make clones appear deleted.

Gene objects found ← **Pre-analysis** → **No gene objects found**

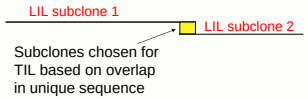
Transposon Insert Library (TIL) derived from a Large Insert Library (LIL) used to finish a gene object repeat cluster.

**pre-analysis is a computational method of screening unfinished sequence for disrupted or missing genes within a repeat region. Selected clone sequence is entered into the analysis pipeline which includes blast searches against nucleotide and protein databases alongside *ab-initio* gene predictions. This data can then be viewed manually through the otter annotation system.

Alternative strategy for tackling non-coding ZLTs

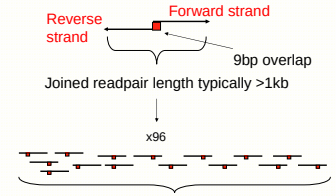
The LIL-TIL strategy

The LIL-TIL strategy involves the generation of a large insert library. Large insert subclones are then chosen for TIL based on overlap in sequence believed to occur only once in the clone. 2 overlapping LIL subclones can generate up to 24kb of reliable sequence to help build the assembly.



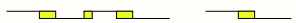
The aim is to build each LIL subclone using a TIL. Once joined, the end product is a length of dependable sequence which aids repeat construction.

TILs are generated using a kit based approach. Transposons randomly insert into the subclone and are sequenced outwards from the transposon insertion site. TIL readpairs can then be joined up based on the 9bp duplication site to give contiguous TIL readpairs in excess of 1kb.



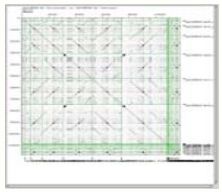
The result is a contiguous LIL subclone up to 12kb in length.

The diagram below illustrates how, by choosing a number of subclones that overlap in unique sequence, a minimum tile path can be built to correctly assemble large sections of repeat.

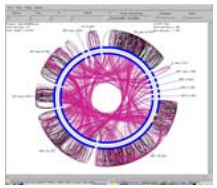


The combination of the LIL-TIL strategy, existing readpair information and restriction digest data is a very powerful means of building a correct assembly. The strategy is particularly useful where the repeat unit is larger than the subclone. In this case, individual TILs can be phrap'd with no chance of misassembly. This strategy is less useful where multiple repeat copies occur within the length of a subclone.

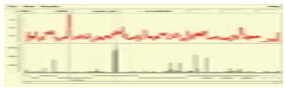
Initial problematic assembly



•Dotter (Sonnhammer) shows numerous contigs containing repetitive sequence.

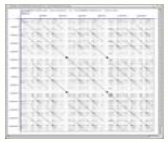


•Orchid (Flowers) shows a high proportion of bad readpairs.

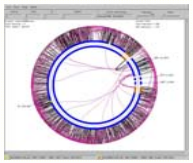


•The Gap4 (Staden) reading coverage histogram (red) shows variable coverage and the diploid graph (black) shows discrepancies.

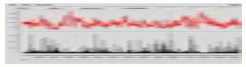
LILTILs enable resolution of repetitive clone



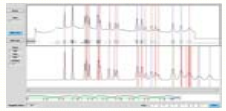
•Dotter shows the repetitive nature of the clone which caused the assembly problems.



•Orchid shows a high level of good readpairs and no groups of bad readpairs.



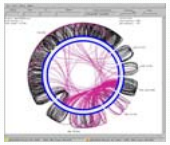
•The reading coverage histogram (red) shows a fairly uniform level of coverage and the diploid graph (black) shows no significant discrepancies.



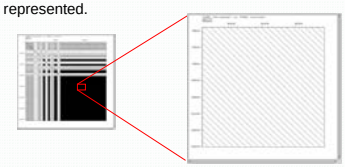
•In-house software Confirm (Attwood) shows no assembly problems

BX005078 (in the illustration above) is an example of a clone identified by pre-analysis as containing several copies of a small gene cluster. The clone was successfully constructed using the LILTIL strategy.

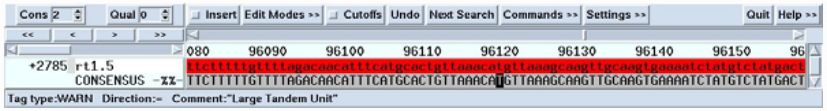
Where collapsed repeats are shown to be non-coding, they are scaffolded to match existing copies of the repeat found in the clone. The repeat is built to match the size given by restriction digest data. By scaffolding existing repeat sequence to match the digests, the finished product will closely match the original clone sequence. Ambiguous sequence is removed from the unsure region and replaced with scaffold sequence taken from known repeat so that only true repeat sequence is represented.



An orchid display of CR381705 shows a problematic repeat.

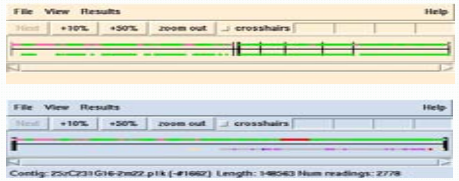


A zoomed in dotter shows that the repeat unit is very small (<200bp) whereas the total repeat size is in excess of 80kb. The repeat is also highly conserved making this clone very resistant to normal finishing strategies including the LILTIL strategy.

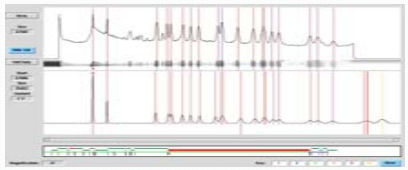


The screenshot above shows the repeat scaffold that has been generated by reproducing known repeat within the clone and adding quality values to it. An annotation tag of at least 1kb is added across sequence that has been inserted in this manner and this is visible in the EMBL/Genbank header:

```
CC   Any regions longer than 1kb tagged as misc-feature "unsure" are part of
CC   a tandem repeat of more than 10kb in length where it has not been possible
CC   to anchor the base differences between repeat copies. The region has been
CC   built up based on the repeat element to match the total size of repeat
CC   indicated by restriction digest, but repeat copies may not be in the
CC   correct order and the usual finishing criteria may not apply.
```



The two gap4 contig selector windows shown above illustrate that contig number is very quickly brought down from several to one using scaffold sequence (tagged red).



The restriction digest data show that the repeat is now the correct size. Restriction digest data shown using Confirm.

The advantage of this strategy compared with force joining with large amounts of data missing is that the finished assembly gives a true reflection of the clone size which is useful for genome assembly.

Where sequence features such as direct repeats or unique islands are present these are placed within the ZLT as accurately as possible using restriction digest data. Where unique sequence is confidently placed it is not included within the unsure tag and normal finishing rules apply. Where several unique islands exist within ZLTs, placement can be problematic despite the use of restriction digest data. Where placement is unsure, annotation tags are used to indicate this.

Acknowledgements

Mapping Core
Sean Humphray, Claire Riddle, David Elliott, Glen Thredgold

Subcloning
Matt Jones, David Willey

Shotgun
John Burton, Karen Oliver, Chris Clee, Sarah Sims

Auto-prefinishing
Stuart McLaren

Finishing
Darren Grafham, Karen McLaren, Lucy Matthews, Sophie Palmer

HAVANNA Group
Jennifer Harrow

Zebrafish Analysis Group
Kerstin Howe, Gavin Laird

Head of Genome Biology
Jane Rogers