CNV Consensus Track

Introduction

Currently there are a relatively large number of CNV data control sets available. These sets tend to be comprised of large numbers of copy number variable regions derived from a number of different scientific studies. The accurate combination of control sets can be complicated by a number of factors, which include differences in array designs, discovery sensitivity and study size.

The CNV consensus set is comprised of a number of carefully selected, high-resolution controls sets where frequency information is available. As the understanding of CNV increases it is important to remain aware of the frequency spectrum of a given variant. Observing a variant only once (singleton) across a large control set is very different to having observed the variant in every sample (100%).

Included Data Sets

- 42 Million study - raw call lists.
- 42 Million study - genotyped regions.
- WTCCC study - merged Affy6 data set.
- 1000 Genomes pilot - merged deletions.
- 1000 Genomes pilot - tandem duplications.
- DDD study - national blood service controls.
- DDD study - generation Scotland controls.

General Approach

First features that obey certain rules are merged within each individual set separately. “Loss” and “gain” sets are created from each starting set and “unknown” (or “loss/gain”) features are included within both sets. These sets are then analysed in the following way:

Feature merging

If a pairwise comparison between two features within a set shows a reciprocal overlap greater than 50%. Feature clusters are built for all features showing greater than 50% overlap. All features within the resulting cluster are merged, creating new features for each, these (identical) features display new breakpoints of 90% of the inner to outer breakpoint distances.

This analysis continues in an iterative manner and closure is achieved once no feature shows greater than 50% reciprocal overlap with any other feature within the set.

Frequency calculation

If frequency information is present in a set that information is carried though the analysis and combined if features are merged i.e. if two feature containing frequency information are merged the number of observations are summed across the two (or more) features.

If frequency information was not available it is calculated with each feature initially being assigned only a single observation.

Combining sets and frequency data

Sets displaying the same type i.e. loss or gain can be combined.

First the number of samples within each set is summed to yield the overall number of samples in the combined set.

Then feature merging between sets is applied using the same principal as above.

If any resulting feature clusters are to be merged:

- Breakpoint locations are adjusted as above.
- Frequency information is combined as above.

The overall frequency of each feature is then adjusted given the new overall number of samples within the combined set.

Method Implementation

All methods describe above are implemented in a single C binary. The different parameters for all the individual steps can be easily adjusted. The feature comparison and clustering methods are highly efficient and can be applied to many millions of features with remarkable performance characteristics.

Comments

We are always interested in receiving feedback on how useful this resource is to Decipher users. Please contact us at decipher@sanger.ac.uk with any queries.