Internship Presentation

Halina Krzystek
MPS Candidate in Biomedical and Health Informatics
Part I: An Evaluation of Copy Number Variant Calling Algorithms for a Clinical Genomics Pipeline Using Exome Sequencing
The NCGENES2 project aims to generate evidence for the use of exome sequencing as a first-line diagnostic tool. Seeking to expand the diagnostic yield of its exomes sequencing by identifying additional variants (CNVs) which may contribute to a clinical phenotype.
My Project Goals

(1) Explore the CNV calling tools available for exome sequencing in a literature review

(2) Evaluate their appropriateness for the NCGENES2 pipeline, and then

(3) Compare the tools’ performance on data from the 1000 Genomes Project
Copy Number Variants (CNVs)

• Variations from the normal copy number of 2 for a diploid organism

Rice and Lysaght, 2017
Pros of Exome Sequencing for CNV detection

- Decreasing cost of sequencing
- Previous microarray-based methods: typically had a resolution of 400kb
- ES could replace current methods that employ microarrays + sequencing
- ES could have a finer granularity in CNV detection, identifying new CNVs as small as one or two exons
- Shortcomings: low sensitivity and high number of false positives
The Read-Depth Method

Fromer and Purcell, 2014
The Read-Depth Method

- Normalization
  - Of raw read count
  - To remove bias such as GC content, mappability, and capture

- Segmentation
  - Segment the CNV calls into chromosomal regions that may span several exons

- CNV calling

Fromer and Purcell, 2014
Methods

- ExomeDepth, CoNIFER, XHMM, CN.Mops and CODEX were appropriate for detecting CNVs in germline samples.

- However, XHMM and CoNIFER require a large number of samples for their normalization steps.

- Thus, ExomeDepth, CN.Mops, and CODEX were selected for evaluation.

- All tools were run using default values for comparison.

- Ran through the framework Ximmer.
• ExomeDepth:
  • Normalization: Robust Beta-Binomial distribution
  • Segmentation: Hidden Markov Model on log ratio of read counts

• CODEX
  • Normalization: Poisson latent model
  • Segmentation: Exon-level threshold to a Circular Binary Segmentation (CBS) algorithm

• CN.Mops:
  • Normalization: Mixture of Poissons models
  • Segmentation: Based on models’ I/NI calls, joining segments with similar I/NI calls
Results

(1) CNV Size and Distribution

(2) Concordance between 3 Callers

(3) Concordance with 1000 Genomes GS Call Set
1. CNV Size and Distribution

- CN.Mops could not identify CNVs less than 1kb in length
- ExomeDepth had the greatest range in CNV size
- ExomeDepth identified CNVs as small as one exon and CODEX identified CNVs as small as three exons
- CN.Mops and CODEX identified larger CNVs
2. Concordance Between Callers

- ExomeDepth (9374)
- CN.Mops (32095)
- CODEX (14887)

Intersection counts:
- 2760
- 3022
- 2277
- 1315
- 8127
- 3168
- 18669
3. Concordance with GS call set

- Only 180 of the 1213 (15%) target exons were identified by all three callers.

- ExomeDepth achieved the highest sensitivity at 40.73%, precision at 5.3%

- CODEX had a sensitivity of 30.67%, precision of 2.4%

- CN.Mops had a sensitivity of 29.60%, precision of 1.1%
All 3 callers: 15% sensitivity, 8% precision, 2097 FP
Two or more callers: 37.5% sensitivity, 14288 FP
ExomeDepth + 1 or more callers: 35.0% sensitivity, 6189 FP
<table>
<thead>
<tr>
<th>Callers</th>
<th>Sensitivity</th>
<th>Precision</th>
<th>False Positives</th>
</tr>
</thead>
<tbody>
<tr>
<td>ExomeDepth alone</td>
<td>40.7%</td>
<td>5.3%</td>
<td>8,880</td>
</tr>
<tr>
<td>CODEX alone</td>
<td>30.7%</td>
<td>2.4%</td>
<td>14,528</td>
</tr>
<tr>
<td>CN.Mops alone</td>
<td>29.6%</td>
<td>1.1%</td>
<td>31,733</td>
</tr>
<tr>
<td>All 3 Callers</td>
<td>14.8%</td>
<td>8%</td>
<td>2,097</td>
</tr>
<tr>
<td>2 or More Callers</td>
<td>37.5%</td>
<td>3.1%</td>
<td>14,288</td>
</tr>
<tr>
<td>ExomeDepth + 1 or more callers</td>
<td>35.0%</td>
<td>6.4%</td>
<td>6,189</td>
</tr>
</tbody>
</table>
631 missed exons are largely on the X chromosome
Corresponding 127 CNVs’ Size

Deletions

Duplications

CN > 4
Takeaways and Future Directions

1. The best rule is EXOMEDEPTH + one or more callers

2. Can machine learning be applied to CNV calling?
Part II: The Application of Machine Learning Clustering on MicroRNAs as a Quality Analysis and Control Tool for Large Cancer Genomics Projects
microRNAs

- miRNAs are small non-coding RNAs that can regulate genes
- Identified as significant in multiple cancers: breast, ovarian, glioblastoma, leukemia
Clustering is a type of unsupervised machine learning—one that requires no training set.

Groups together samples by similarity in traits.

In our case, samples are grouped by similarity in microRNA expression.

Two widely used methods: $k$-means and hierarchical clustering.
Results of Hierarchical Clustering
Lung Adenocarcinoma Clustering - Pilot
Lung Squamous Cell Carcinoma - Pilot
Normal Adjacents- Pilot
Amyloid Leukemia Clustering
Glioblastoma Clustering
Observations

- Amyloid Leukemia and Glioblastomas cluster well in retrospective data
- Lung Adenocarcinoma and Lung Squamous Cell Carcinoma can cluster separately
- Normal Adjacent samples cluster separately from tumors
Lessons Learned from Internship

FORENSIC BIOINFORMATICS

POWER AND LIMITATIONS OF SEQUENCING TECHNOLOGY IN HEALTHCARE

HOW TECH CAN BE LEVERAGED TO IMPROVE HEALTH OUTCOMES

• Rice, A., McLysaght, A. Dosage sensitivity is a major determinant of human copy number variant pathogenicity. *Nat Commun* **8**, 14366 (2017). [https://doi.org/10.1038/ncomms14366](https://doi.org/10.1038/ncomms14366)