Applications of a generic model of genomic variations functional analysis

Sarah N. Mapelli, Uberto Pozzoli
The tools developer point of view: a general analysis flow

Step 1: obtain information about a given genomic element

Definitions:

Genomic element (GE): A region identified by an interval along a chromosome strand
GE annotations: A set of functionally meaningful positions and intervals (along a GE).
GE sequence: The genomic sequence corresponding to a GE

Resources involved:

Annotation databases (UCSC, ENSEMBL,...)
Complex File formats (GFF, BED, ...)
Genome assemblies (NCBI36, GRCh37, ...)

Development Issues:

- Data structures need to be defined to describe annotations.
- Annotations must be retrieved (different formats, storage systems, strandness)
- Sequences are also needed (different assemblies)
The tools developer point of view: a general analysis flow

Step 2: map relevant data and/or calculate features along the GE

Definitions:
Calculated feature: A quantitative feature that can be calculated by algorithms
Retrieved feature: A quantitative feature obtained by experimental data

Resources involved:
Annotation databases (UCSC, ENSEMBL,...)
Experiments results (many different formats)
Algorithms for feature calculation

Development Issues
• Conversion to/from external algorithms and data formats.
• Different coordinates systems (reference/element)
• Features are often tightly connected with element annotation
The tools developer point of view: a general analysis flow

Step 3: map genomic variations on a GE to obtain a “varied” GE (vGE)

Definitions:
Genomic variation (GV): Any genomic variation with respect to a reference sequence
Varied GE (vGE): A Genomic Element whose interval, annotations, sequence and features are modified according to one or more GV.

Resources involved:
Variation databases (HapMap, 1000 Genomes Project, ...)
Resequencing results (vcf files)

Development Issues
- Variations are usually stored in huge databases/files that needs specific tools to be queried (i.e. vcf tools)
- vGE annotations must reflect positional changes induced by Indels, CNV and SV.
- Feature recalculation can be time consuming.
- Difficult identification of intervals that needs feature recalculation due to the presence of variations.
- Varied sequence must be “constructed” to feed “sequence driven” algorithms.
The tools developer point of view: a general analysis flow

Step 4: compare a GE with the corresponding vGE and report differences.

This is the final step of our work flow we have everything we need to detect modifications induced by variations.

It usually involves the comparison of some genomic feature or annotation between the two elements.

The details are strongly dependent on what we are looking for:

Still, we have common...

Development Issues

- It is not trivial to identify matching positions along the two elements (i.e. GE and vGE) due to the changes possibly induced by variations.
- Annotations can be deleted or newly inserted.
The tools developer point of view: implementation Issues

- Data structures need to be defined to describe annotations.
- Annotations must be retrieved (different formats, storage systems, strandness)
- Sequences are also needed (different assemblies)
- Conversion to/from external algorithms and data formats.
- Different coordinates systems (reference/element)
- Features are often tightly connected with element annotation
- Variations are usually stored in huge databases/files that needs specific tools to be queried (i.e. vcf tools)
- vGE annotations must reflect positional changes induced by Indels, CNV and SV.
- Feature recalculation can be time consuming.
- Difficult identification of intervals that needs feature recalculation due to the presence of variations.
- Varied sequence must be “constructed” to feed “sequence driven” algorithms.
- It is not trivial to identify matching positions along the two elements (i.e. GE and vGE) due to the changes possibly induced by variations.
- Annotations can be deleted or newly inserted
The tools developer point of view: a possible solution

The definition of GE and GE Annotations allows us:

- To define a data structure able to contain GE interval and annotations
- To link this data structure to a set of methods to set / get Annotation information
- To link this data structure to a set of methods to convert between reference and element coordinate systems

We call this structure and its methods “gElment”: functions can be developed to instantiate gElements from different annotation formats.

The gElements instantiated will be independent from the data source.
The tools developer point of view: a possible solution

We define `gSequenceRetriever` as an abstract method to retrieve sequence information within given a `gElement` interval.

We also define `gFeatureRetriever` as an abstract method to retrieve data or to calculate features in a given interval of a given `gElement`.

At this point we can add to `gElements` methods that use `gSequenceRetriever`s and `gFeatureRetriever`s to retrieve/calculate sequence and features respectively.

Due to their abstract nature sequence and feature retrievers can be developed for specific data formats, databases or algorithms.
The tools developer point of view: a possible solution

We define gVariations as a data structure that can contain information about genomic variations. A set of methods, associated to this data structure, allow us to get/set variations information.

gVariations can be instantiated by using method written to get information from different sources. Once instantiated they will be independent from the data source.
The tools developer point of view: a possible solution

We can obtain a new instance of a gElement by adding a gVariation. We can add methods to the gElement in order to modify annotations, sequences and features according to the variations.

In particular the methods to modify features will identify the regions that needs recalculation and call the feature retrievers only in those regions.

Positions on the originally instantiated gElement can be mapped to the modified one thanks to the mechanism of coordinate conversion.
The tools developer point of view: a possible solution

- gSequenceRetriever instance
- gFeatureRetriever instance

Comparison features, annotations

- gElement instance (GE)
- gVariations instantiation function

- varied gElement instance (vGE)

Object oriented approach
Language: C++
Library: GeCo++
The tools developer point of view: implementation issues

- Data structures need to be defined to describe annotations.
- Annotations must be retrieved (different formats, storage systems, strandness).
- Sequences are also needed (different assemblies).
- Conversion to/from external algorithms and data formats.
- Different coordinate systems (reference/element).
- Features are often tightly connected with element annotation.
- Variations are usually stored in huge databases/files that needs specific tools to be queried (i.e. vcftools).
- vGE annotations must reflect positional changes induced by Indels, CNV and SV.
- Feature recalculation can be time-consuming.
- Difficult identification of intervals that needs feature recalculation due to the presence of variations.
- Varied sequence must be “constructed” to feed “sequence-driven” algorithms.
- It is not trivial to identify matching positions along the two elements (i.e. GE and vGE) due to the changes possibly induced by variations.
- Annotations can be deleted or newly inserted.
```bash
$ time ./whatsup --method=simpleGenes --genes=DMD --subjects=P01_337_07
response
variantTable:

<table>
<thead>
<tr>
<th>gene</th>
<th>transcript</th>
<th>subject</th>
<th>chrom</th>
<th>position</th>
<th>rs_id</th>
<th>ref</th>
<th>p</th>
<th>m</th>
<th>f</th>
<th>mend</th>
<th>type</th>
</tr>
</thead>
<tbody>
<tr>
<td>DMD</td>
<td>uc004dxc.2</td>
<td>P01_337_07</td>
<td>chrX</td>
<td>31196945</td>
<td>rs72466541</td>
<td>T</td>
<td>T/T</td>
<td>T/T</td>
<td>C/C</td>
<td>1.000000</td>
<td>intron</td>
</tr>
<tr>
<td>DMD</td>
<td>uc004dxc.2</td>
<td>P01_337_07</td>
<td>chrX</td>
<td>31200831</td>
<td>rs72466545</td>
<td>A</td>
<td>A/A</td>
<td>A/A</td>
<td>AT/AT</td>
<td>1.000000</td>
<td>intron</td>
</tr>
<tr>
<td>DMD</td>
<td>uc004dcy.1</td>
<td>P01_337_07</td>
<td>chrX</td>
<td>31224683</td>
<td>rs2293668</td>
<td>A</td>
<td>G/G</td>
<td>G/G</td>
<td>G/G</td>
<td>0.000000</td>
<td>intron</td>
</tr>
<tr>
<td>DMD</td>
<td>uc004dxc.2</td>
<td>P01_337_07</td>
<td>chrX</td>
<td>31496349</td>
<td>rs1800280</td>
<td>C</td>
<td>T/T</td>
<td>C/T</td>
<td>T/T</td>
<td>0.000000</td>
<td>missense</td>
</tr>
<tr>
<td>DMD</td>
<td>uc004dcy.1</td>
<td>P01_337_07</td>
<td>chrX</td>
<td>31854781</td>
<td>rs1379871</td>
<td>G</td>
<td>G/G</td>
<td>G/G</td>
<td>C/C</td>
<td>1.000000</td>
<td>intron</td>
</tr>
<tr>
<td>DMD</td>
<td>uc004dcr.1</td>
<td>P01_337_07</td>
<td>chrX</td>
<td>31893306</td>
<td>rs1800275</td>
<td>T</td>
<td>T/T</td>
<td>T/T</td>
<td>G/G</td>
<td>1.000000</td>
<td>cds</td>
</tr>
<tr>
<td>DMD</td>
<td>uc004dcw.2</td>
<td>P01_337_07</td>
<td>chrX</td>
<td>32503193</td>
<td>rs2200266</td>
<td>T</td>
<td>T/T</td>
<td>T/C</td>
<td>C/C</td>
<td>1.000000</td>
<td>cds</td>
</tr>
<tr>
<td>DMD</td>
<td>uc010ngr.1</td>
<td>P01_337_07</td>
<td>chrX</td>
<td>32663134</td>
<td>rs1800266</td>
<td>T</td>
<td>T/T</td>
<td>T/G</td>
<td>T/T</td>
<td>0.000000</td>
<td>cds</td>
</tr>
<tr>
<td>DMD</td>
<td>uc010ngr.1</td>
<td>P01_337_07</td>
<td>chrX</td>
<td>32867944</td>
<td></td>
<td>T</td>
<td>T/T</td>
<td>T/TA</td>
<td>T/T</td>
<td>0.000000</td>
<td>intron</td>
</tr>
</tbody>
</table>
```

real    0m0.692s
user    0m0.566s
sys     0m0.050s
Example
<table>
<thead>
<tr>
<th>gene</th>
<th>transcript</th>
<th>subject</th>
<th>class</th>
<th>position</th>
<th>rs_id</th>
<th>ref</th>
<th>p</th>
<th>a</th>
<th>f</th>
<th>strand</th>
<th>type</th>
</tr>
</thead>
<tbody>
<tr>
<td>00D</td>
<td>ac0844cx.2</td>
<td>P01_337_07</td>
<td>chrX</td>
<td>31,196,945</td>
<td>rs7246541</td>
<td>T</td>
<td>T/T</td>
<td>T/T</td>
<td>C/C</td>
<td>1</td>
<td>Intron</td>
</tr>
<tr>
<td>00D</td>
<td>ac0844cx.2</td>
<td>P01_337_07</td>
<td>chrX</td>
<td>31,209,833</td>
<td>rs7246545</td>
<td>A</td>
<td>A/A</td>
<td>A/A</td>
<td>A/T</td>
<td>1</td>
<td>Intron</td>
</tr>
<tr>
<td>00D</td>
<td>ac0844cx.1</td>
<td>P01_337_07</td>
<td>chrX</td>
<td>31,224,682</td>
<td>rs2293684</td>
<td>A</td>
<td>G/G</td>
<td>G/G</td>
<td>G/G</td>
<td>0</td>
<td>Intron</td>
</tr>
<tr>
<td>00D</td>
<td>ac0844cx.2</td>
<td>P01_337_07</td>
<td>chrX</td>
<td>31,496,249</td>
<td>rs1808266</td>
<td>C</td>
<td>T/T</td>
<td>C/T</td>
<td>T/T</td>
<td>0</td>
<td>nonsense</td>
</tr>
<tr>
<td>00D</td>
<td>ac0844cv.1</td>
<td>P01_337_07</td>
<td>chrX</td>
<td>31,854,782</td>
<td>rs1379871</td>
<td>G</td>
<td>G/G</td>
<td>G/G</td>
<td>G/G</td>
<td>1</td>
<td>Intron</td>
</tr>
<tr>
<td>00D</td>
<td>ac0844cv.2</td>
<td>P01_337_07</td>
<td>chrX</td>
<td>31,991,306</td>
<td>rs1808275</td>
<td>T</td>
<td>T/T</td>
<td>T/T</td>
<td>G/G</td>
<td>1</td>
<td>cds</td>
</tr>
<tr>
<td>00D</td>
<td>ac0844cv.2</td>
<td>P01_337_07</td>
<td>chrX</td>
<td>32,503,193</td>
<td>rs228406</td>
<td>T</td>
<td>T/T</td>
<td>T/C</td>
<td>C/C</td>
<td>1</td>
<td>cds</td>
</tr>
<tr>
<td>00D</td>
<td>ac0310mgr.1</td>
<td>P01_337_07</td>
<td>chrX</td>
<td>32,663,134</td>
<td>rs1808266</td>
<td>T</td>
<td>T/T</td>
<td>T/6</td>
<td>T/T</td>
<td>0</td>
<td>cds</td>
</tr>
<tr>
<td>00D</td>
<td>ac0310mgr.1</td>
<td>P01_337_07</td>
<td>chrX</td>
<td>32,867,941</td>
<td></td>
<td>T</td>
<td>T/T</td>
<td>T/16</td>
<td>T/T</td>
<td>0</td>
<td>Intron</td>
</tr>
</tbody>
</table>