Example of a NGS workflow in KNIME and connecting KNIME with BioXM

Sebastian Kopetzky
Biomax Informatics AG
Overview

1  BioXM
2  KNIME NGS workflow
3  Interaction BioXM ↔ KNIME
Semantic Networks
Overview — Agile Solution Building

Query the knowledge network, explore the graph and report query results

**Step 1**
Specification
Designing the data model

**Step 2**
Implementation
Importing information

**Step 3**
Use Query building and information retrieval

Define the domain-specific data model

Instantiate the knowledge network with data and information from external resources
Access to the semantic model

Find this patient by the human readable query above - BioXM speaks your language
Overview

1. BioXM
2. KNIME NGS workflow
3. Interaction BioXM ↔ KNIME
- **Input**: Raw reads (fastq)
- **Output**: Annotated SNPs/InDels (vcf)
- **FastQC**: Read quality statistics/visualization
  
  modified by Jonathan Hoser, Helmholtz Center Munich

- **RawReadManipulator**: Read filtering
  
  modularized, extensible and multicore-capable read-filtering tool by Jonathan Hoser
**- BWA: Map reads to reference genome**

- **BAMSAMConverter**: convert BWA output to BAM
- **SNPcall**: call SNPs using mpileup, bcftools, vcfutils

- **SnpEffGetDB**: download reference database
- **SnpEff**: annotate SNPs
- **SnpSift**: filter SNPs (e.g. quality)

Overview

1 BioXM

2 KNIME NGS workflow

3 Interaction BioXM ↔ KNIME
BioXM Reader node
BioXM Importer node

RunFastQC → FastQC → FastQC → RawReadManipulator → Node 6

BWA → BAMSAMConverter → SNPcall → Node 7 → Node 8

FastQC → Node 3

SnpEffGetDB

SnpEff → SnpSift → VCFtoTable → Node 10 → Node 2

SnpEffGetDB

Node 9

Node 13

Node 14

Node 15

Node 2

Dialog - 2.2 - BioXM Importer

Connection  Tab Import  Flow Variables  Memory Policy

Scripts:

Ok, Redo!  Script name: vcf_import

Settings:

Import Policy: ALLOW_COMPLETE_IMPORT_ONLY

Include Header

Detailed Log

Allowed Errors: 0

OK  Apply  Cancel
BioXM Importer node

### General Elements

**View:** SNP Info

#### Object

<table>
<thead>
<tr>
<th>Object</th>
<th>Reference</th>
<th>Alternative</th>
<th>Quality</th>
<th>Read depth</th>
<th>Zygoity</th>
<th>Chromosome</th>
<th>Pos...</th>
<th>Effect(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NC_010473A</td>
<td>T</td>
<td>C</td>
<td>40.0</td>
<td>46</td>
<td>0/1</td>
<td>NC_010473</td>
<td>1398034</td>
<td>SYNTHOMATIC_CODING</td>
</tr>
<tr>
<td>NC_010473A</td>
<td>G</td>
<td>A</td>
<td>220.0</td>
<td>51</td>
<td>1/1</td>
<td>NC_010473</td>
<td>1390190</td>
<td>SYNTHOMATIC_CODING</td>
</tr>
<tr>
<td>NC_010473A</td>
<td>G</td>
<td>S</td>
<td>45.0</td>
<td>83</td>
<td>0/1</td>
<td>NC_010473</td>
<td>1396610</td>
<td>DOWNSTREAM</td>
</tr>
<tr>
<td>NC_010473A</td>
<td>T</td>
<td>T</td>
<td>27.0</td>
<td>44</td>
<td>0/1</td>
<td>NC_010473</td>
<td>1409592</td>
<td>STOP_GAINED [HIGH</td>
</tr>
<tr>
<td>NC_010473A</td>
<td>G</td>
<td>A</td>
<td>30.0</td>
<td>26</td>
<td>0/1</td>
<td>NC_010473</td>
<td>1417132</td>
<td>NON_SYNTHOMATIC_CODING</td>
</tr>
<tr>
<td>NC_010473A</td>
<td>G</td>
<td>G</td>
<td>220.0</td>
<td>32</td>
<td>0/1</td>
<td>NC_010473</td>
<td>1423504</td>
<td>DOWNSTREAM</td>
</tr>
<tr>
<td>NC_010473A</td>
<td>A</td>
<td>A</td>
<td>47.0</td>
<td>32</td>
<td>0/1</td>
<td>NC_010473</td>
<td>1422026</td>
<td>NON_SYNTHOMATIC_CODING</td>
</tr>
<tr>
<td>NC_010473A</td>
<td>G</td>
<td>G</td>
<td>53.0</td>
<td>34</td>
<td>0/1</td>
<td>NC_010473</td>
<td>1427423</td>
<td>DOWNSTREAM</td>
</tr>
<tr>
<td>NC_010473A</td>
<td>T</td>
<td>T</td>
<td>87.0</td>
<td>46</td>
<td>0/1</td>
<td>NC_010473</td>
<td>1426296</td>
<td>DOWNSTREAM</td>
</tr>
<tr>
<td>NC_010473A</td>
<td>G</td>
<td>G</td>
<td>37.0</td>
<td>80</td>
<td>0/1</td>
<td>NC_010473</td>
<td>1431090</td>
<td>DOWNSTREAM</td>
</tr>
<tr>
<td>NC_010473A</td>
<td>G</td>
<td>G</td>
<td>21.0</td>
<td>19</td>
<td>0/1</td>
<td>NC_010473</td>
<td>1444086</td>
<td>NON_SYNTHOMATIC_CODING</td>
</tr>
<tr>
<td>NC_010473A</td>
<td>T</td>
<td>T</td>
<td>222.0</td>
<td>42</td>
<td>1/1</td>
<td>NC_010473</td>
<td>1446379</td>
<td>NON_SYNTHOMATIC_CODING</td>
</tr>
<tr>
<td>NC_010473A</td>
<td>C</td>
<td>C</td>
<td>34.0</td>
<td>32</td>
<td>0/1</td>
<td>NC_010473</td>
<td>14477207</td>
<td>SYNTHOMATIC_CODING</td>
</tr>
<tr>
<td>NC_010473A</td>
<td>G</td>
<td>G</td>
<td>74.0</td>
<td>26</td>
<td>0/1</td>
<td>NC_010473</td>
<td>1447907</td>
<td>NON_SYNTHOMATIC_CODING</td>
</tr>
</tbody>
</table>

#### Related objects

- **Organism:** E. coli
- **Description:** YidB, antibiotic ABC transporter
- **Gene:** yidB

### Export

- [Export](#)
- [Graph](#)
- [List](#)
- [Folder](#)
- [Annotate](#)

### 932 objects
Executing KNIME from BioXM
Executing KNIME from BioXM
## Executing KNIME from BioXM

### SNP Query Variables

<table>
<thead>
<tr>
<th>Object</th>
<th>SNP Position</th>
<th>SNP Reference</th>
<th>SNP Alternative</th>
<th>SNP Quality</th>
<th>SNP Read-depth</th>
<th>SNP Zygosity</th>
<th>SNP Affected Genes</th>
<th>SNP Effect(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NC_01047</td>
<td>14644995.0</td>
<td>T</td>
<td>G</td>
<td>10.0</td>
<td>40.0</td>
<td>0/1</td>
<td>NON_SYNONYMOUS_CODING</td>
<td>MODERATE</td>
</tr>
<tr>
<td>NC_01047</td>
<td>14654955.0</td>
<td>A</td>
<td>G</td>
<td>58.0</td>
<td>37.0</td>
<td>0/1</td>
<td>NON_SYNONYMOUS_CODING</td>
<td>MODERATE</td>
</tr>
<tr>
<td>NC_01047</td>
<td>14652520.0</td>
<td>T</td>
<td>G</td>
<td>38.0</td>
<td>29.0</td>
<td>0/1</td>
<td>NON_SYNONYMOUS_CODING</td>
<td>MODERATE</td>
</tr>
<tr>
<td>NC_01047</td>
<td>14691344.0</td>
<td>T</td>
<td>C</td>
<td>35.0</td>
<td>42.0</td>
<td>0/1</td>
<td>NON_SYNONYMOUS_CODING</td>
<td>MODERATE</td>
</tr>
<tr>
<td>NC_01047</td>
<td>14627426.0</td>
<td>A</td>
<td>C</td>
<td>37.0</td>
<td>34.0</td>
<td>0/1</td>
<td>NON_SYNONYMOUS_CODING</td>
<td>MODERATE</td>
</tr>
<tr>
<td>NC_01047</td>
<td>14615952.0</td>
<td>T</td>
<td>G</td>
<td>33.0</td>
<td>36.0</td>
<td>0/1</td>
<td>NON_SYNONYMOUS_CODING</td>
<td>MODERATE</td>
</tr>
<tr>
<td>NC_01047</td>
<td>14613448.0</td>
<td>T</td>
<td>C</td>
<td>25.0</td>
<td>21.0</td>
<td>0/1</td>
<td>NON_SYNONYMOUS_CODING</td>
<td>MODERATE</td>
</tr>
<tr>
<td>NC_01047</td>
<td>14626596.0</td>
<td>G</td>
<td>T</td>
<td>42.0</td>
<td>29.0</td>
<td>0/1</td>
<td>NON_SYNONYMOUS_CODING</td>
<td>MODERATE</td>
</tr>
<tr>
<td>NC_01047</td>
<td>14652336.0</td>
<td>T</td>
<td>C</td>
<td>56.0</td>
<td>32.0</td>
<td>0/1</td>
<td>NON_SYNONYMOUS_CODING</td>
<td>MODERATE</td>
</tr>
<tr>
<td>NC_01047</td>
<td>14656430.0</td>
<td>A</td>
<td>C</td>
<td>35.0</td>
<td>35.0</td>
<td>0/1</td>
<td>NON_SYNONYMOUS_CODING</td>
<td>MODERATE</td>
</tr>
<tr>
<td>NC_01047</td>
<td>14643896.0</td>
<td>T</td>
<td>G</td>
<td>78.0</td>
<td>34.0</td>
<td>0/1</td>
<td>NON_SYNONYMOUS_CODING</td>
<td>MODERATE</td>
</tr>
<tr>
<td>NC_01047</td>
<td>14641914.0</td>
<td>T</td>
<td>C</td>
<td>72.0</td>
<td>37.0</td>
<td>0/1</td>
<td>NON_SYNONYMOUS_CODING</td>
<td>MODERATE</td>
</tr>
<tr>
<td>NC_01047</td>
<td>14627026.0</td>
<td>T</td>
<td>G</td>
<td>119.0</td>
<td>33.0</td>
<td>0/1</td>
<td>NON_SYNONYMOUS_CODING</td>
<td>MODERATE</td>
</tr>
<tr>
<td>NC_01047</td>
<td>14527023.0</td>
<td>A</td>
<td>C</td>
<td>38.0</td>
<td>36.0</td>
<td>0/1</td>
<td>NON_SYNONYMOUS_CODING</td>
<td>MODERATE</td>
</tr>
<tr>
<td>NC_01047</td>
<td>14626943.0</td>
<td>C</td>
<td>A</td>
<td>25.0</td>
<td>37.0</td>
<td>0/1</td>
<td>NON_SYNONYMOUS_CODING</td>
<td>MODERATE</td>
</tr>
<tr>
<td>NC_01047</td>
<td>14646875.0</td>
<td>T</td>
<td>G</td>
<td>21.0</td>
<td>35.0</td>
<td>0/1</td>
<td>NON_SYNONYMOUS_CODING</td>
<td>MODERATE</td>
</tr>
<tr>
<td>NC_01047</td>
<td>14646115.0</td>
<td>T</td>
<td>C</td>
<td>60.0</td>
<td>29.0</td>
<td>0/1</td>
<td>NON_SYNONYMOUS_CODING</td>
<td>MODERATE</td>
</tr>
<tr>
<td>NC_01047</td>
<td>14638626.0</td>
<td>A</td>
<td>G</td>
<td>88.0</td>
<td>34.0</td>
<td>0/1</td>
<td>NON_SYNONYMOUS_CODING</td>
<td>MODERATE</td>
</tr>
<tr>
<td>NC_01047</td>
<td>14648446.0</td>
<td>T</td>
<td>C</td>
<td>22.0</td>
<td>32.0</td>
<td>0/1</td>
<td>NON_SYNONYMOUS_CODING</td>
<td>MODERATE</td>
</tr>
<tr>
<td>NC_01047</td>
<td>14636346.0</td>
<td>A</td>
<td>C</td>
<td>55.0</td>
<td>44.0</td>
<td>0/1</td>
<td>NON_SYNONYMOUS_CODING</td>
<td>MODERATE</td>
</tr>
<tr>
<td>NC_01047</td>
<td>14634816.0</td>
<td>T</td>
<td>G</td>
<td>65.0</td>
<td>44.0</td>
<td>0/1</td>
<td>NON_SYNONYMOUS_CODING</td>
<td>MODERATE</td>
</tr>
<tr>
<td>NC_01047</td>
<td>146322496.0</td>
<td>A</td>
<td>C</td>
<td>35.0</td>
<td>27.0</td>
<td>0/1</td>
<td>NON_SYNONYMOUS_CODING</td>
<td>MODERATE</td>
</tr>
</tbody>
</table>

322 object(s)
Goals:

- Facilitate interaction between BioXM and KNIME

- Access arbitrary workflows from BioXM
  ⇒ combine the potentials of both systems
Thank you for your attention!

Co-developers of the NGS pipeline:

Jan-Dominik Quell, Helmholtz Center Munich
Maximilian Hastreiter, Helmholtz Center Munich