Variant calling in RNA-seq data using Varscan

Elodie Girard

U900 INSERM - Mines ParisTech - Institut Curie

ITMO Roscoff – 11/19/2013
Dataset

- Public data: exome sequenced by the International HapMap Project
- Single-end reads of 100bp, Illumina Genome Analyzer IIx
- RNA-seq data of this exome available (Pickrell *et al.*, Nature, 2010)

Objectives of the workshop:

- Variant calling, filtering and annotation in exome-seq data
- Observing the potential impact of these variants by looking at the corresponding RNA-seq data

Objectif of this session:
- Variant detection using Varscan of the corresponding RNA-seq data
RNA-seq Analysis

**Difficulty: 2 types of reads**

- Exonic reads
- Splicing reads

**Goals**

- Gene expression quantification and differential expression
- Detection of Transcripts / Fusion Genes
- Variant calling: interesting (monoallelic allele, comparison with the exome-seq calling) but limited
  - Coverage linked to gene expression
  - RNA editing: post transcriptional mutations (not in DNA)
Mapping

- Standard mapping of RNA-seq reads (e.g.: with BWA)

→ A dedicated mapping strategy is needed
Mapping

- Spliced alignment of RNA-seq reads

- 2 strategies: Exon-first approach & Seed and extend approach
TopHat

Exon-First Approach

1- Exon read mapping:
   Map reads continuously onto the genome using an unspliced mapper (Bowtie)

2 - Spliced read mapping
   Unmapped reads are divided into shorter segments and mapped
Exome-seq Workflow

1. Fastq
   - Galaxy Format Conversion
     - Groomer
   - Quality Control
     - FastQC

2. Reference Genome
   - Mapping
     - BWA
   - Format conversion
     - Sam-to-Bam
   - Removing PCR duplicates
     - MarkDup

3. Preprocess GATK
   - Base Recalibration

4. Preprocess GATK
   - Indel Realignment

5. Variant Calling GATK
   - Unified Genotyper

6. VCF Filtering

7. VCF Annotation
   - Mitreup
   - Variant Calling VarScan
   - VCF Annotation
→ Mapping with TopHat is already done
→ PCR duplicates are kept in RNA-seq analysis
Files Loading

1: Go to « Shared Data » and « Published Libraries »

2: Choose this library

3: « Import History »

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>1: chr12.fa</td>
<td>Exome-seq</td>
</tr>
<tr>
<td>2: pickrell.rmdup.RG.real.recab.bam</td>
<td>RNA-seq</td>
</tr>
<tr>
<td>3: [VarScan] VarScan File</td>
<td></td>
</tr>
<tr>
<td>4: pickrell_rna_chr12.q20.bam</td>
<td></td>
</tr>
</tbody>
</table>
Visualization with IGV

1: Click on the pencil to edit the attributes of this bam

2: Edit the Database/build by choosing « Human Feb.2009 (hg19) » genome

3: Click on Save
Visualization with IGV

1: Click on the name to open the options

2: Click on « web current » to download and open IGV

3: Allow IGV

Do the same with the other BAM file (« pickrell.rmdup.RG.real.recab.bam »)

Once IGV is opened, don’t click on « web current » but on « local » to visualize an other BAM
Visualization with IGV

Enter « chr12:113,342,800-113,359,800 » and click on « Go »

RNA-seq
Tracks
Exome-seq

Ruler
Coverage
Reads
Annotation
Analysis of this session:

1. **Mpileup** of the « *pickrell_rna_chr12.q20.bam* » on the region « *chr12:112850000-113395000* » with a minimum mapping quality of **20**

2. **VarScan** of the **Mpileup output** *(with CNS: SNVs + Indels calling)*

3. **VarScan Filter** of the **VarScan output**

4. Use the « **Tag and merge multiple VarScan analysis** » to merge the outputs of **VarScan Filter for the Exome-seq and RNA-seq data**

5. Use the « **Filter SNPs on same ref position** » tool on **this merged file (Exome + RNA)**
Tools

All the tools are in the left panel:

**Mardi 19 Détection de variants**

*VarScan*

---

**Mardi 19 Detection de variants**

*VarScan*

VarScan VarScan analysis.

**POSTPROCESS TOOLS**

- **Tag and merge multiple VarScan analysis**
- **VarScan compare** Compare two varscan results files (intersect / merge / unique).
- **VarScan Filter** To filter a varscan input file.
- **Filter SNPs on same ref position**

**SAM TOOLS**

- **MPileup** SNP and indel caller
MPileup (version 0.0.1)

Choose the source for the reference list:

1 : History

BAM files

2 : Select the bam

BAM file 1

BAM file:

3 : chr12.fa

4 : Set Advanced options

Add new BAM file

Using reference file:

5 : MapQ = 20

Min. mapping quality for an alignment to be used:

5 : chr12:112850000-113395000

Only generate pileup in region:

6 : Region

chr12:112850000-113395000

Output per-sample Phred-scaled strand bias P-value:

7 : Execute

Execute
# Pileup

- **Pileup format:** describes the base-pair information at each position

<table>
<thead>
<tr>
<th>Chr</th>
<th>Position</th>
<th>Reference base</th>
<th>Read bases</th>
<th>Number of reads</th>
<th>Base qualities</th>
</tr>
</thead>
</table>
| chr12 | 112888238 | A              | ., / , = match on forward/reverse strand | 108             | =4 =??????@??@@ @=@ ?? @ ? |}
| chr12 | 112888239 | C              | $t, ,T, ,T, ,tTT, tT, tT, T, ,tT, ttttt, tTT, tT, tT, T, ,tT, ttttt, tTT, tT, tT, T, ,tT, ttttt^F,^F, | 936 78??6??6 45<875? ??? ?@6 @6????<? | =6 66= ? ???? 6=7??=???<8@7=7= ? 77?7? | 8 ??78?7????? 8 <8??88 9?8 ?0048 |
VarScan

1: Select the mpileup file

2: Pileup with Cns (calls SNVs + Indels)

3: Choose VarScan Tabulated format

4: Execute
VarScan

- Mutation caller written in **Java** (no installation required) working with **Pileup files** of Targeted, **Exome**, and Whole-Genome sequencing data

- **Multi-platforms**: Illumina, SOLiD, Life/PGM, Roche/454

- Detection of different kinds of variants (SNVs/Indels):
  - Germline variants in individual samples
  - Multi-sample variants **shared or private** in multi-sample datasets

- VarScan specificity is to be able to work with **Tumor/Normal pairs**:
  - Somatic and germline mutation, LOH events in tumor-normal pairs
  - Somatic copy number alterations (CNAs) in tumor-normal exome data
VarScan

- Most published variant callers use **Bayesian statistics** (a probabilistic framework) to detect variants and assess confidence in them (e.g.: GATK)

- VarScan uses a robust **heuristic/statistic** approach to call variants that meet desired thresholds for read depth, base quality, variant allele frequency, and statistical significance

- In Stead *et al.* (2013), they compared 3 different **somatic callers**: MuTect, Strelka, VarScan2
  - **VarScan2 performed best** overall with sequencing depths of 100x, 250x, 500x and 1000x required to accurately identify variants present at 10%, 5%, 2.5% and 1% respectively
VarScan VCF Format

2 types: **VCF** and **Tabulated** Formats - **VarScan Specific**

- VCF output doesn’t work with some GATK tools but works with other filtering or annotating tools (e.g: annovar)

- **VarScan VCF format**: classic VCF header (#) but specific variant lines

<table>
<thead>
<tr>
<th>#CHROM</th>
<th>POS</th>
<th>ID</th>
<th>REF</th>
<th>ALT</th>
<th>QUAL</th>
<th>FILTER</th>
<th>INFO</th>
<th>FORMAT</th>
<th>GENO</th>
</tr>
</thead>
<tbody>
<tr>
<td>chr12</td>
<td>250239</td>
<td>.</td>
<td>A</td>
<td>G</td>
<td>20</td>
<td>PASS</td>
<td>ADP=104;WT=0;</td>
<td></td>
<td>0/1:153:111:104:61:43:</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>HET=1;HOM=0;NC=0</td>
<td></td>
<td>41,35%:</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4,5644E-16:38:32:</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>48:13:36:7</td>
</tr>
</tbody>
</table>

- **ADP** = Average per-sample depth of bases with Phred score = 20
- **WT** = Number of samples called reference (wild-type)
- **HET** = Number of samples called heterozygous-variant
- **HOM** = Number of samples called homozygous-variant
- **NC** = Number of samples not called

Useful in **multi-sample studies**
VarScan VCF Format

- **2 types:** VCF and Tabulated Formats - **VarScan Specific**
  - VCF output doesn’t work with some GATK tools but works with other filtering or annotating tools (e.g.: annovar)

- **VarScan VCF format:** classic VCF header (#) but specific variant lines

<table>
<thead>
<tr>
<th>#CHROM</th>
<th>POS</th>
<th>ID</th>
<th>REF</th>
<th>ALT</th>
<th>QUAL</th>
<th>FILTER</th>
<th>INFO</th>
</tr>
</thead>
<tbody>
<tr>
<td>chr12</td>
<td>250239</td>
<td>.</td>
<td>A</td>
<td>G</td>
<td>20</td>
<td>PASS</td>
<td>ADP=104;WT=0; HET=1;HOM=0; NC=0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>FORMAT</th>
<th>GENO</th>
</tr>
</thead>
<tbody>
<tr>
<td>FREQ:</td>
<td>41,35%:</td>
</tr>
</tbody>
</table>

**GT=Genotype** (1/1: Homozygous ; 0/1 : Heterozygous) / **GQ= Genotype Quality**
**SDP= Raw Read Depth as reported by SAMtools**
**DP= Quality Read Depth of bases with Phred score >= 20**
**RD= Depth of reference-supporting bases**
**AD= Depth of variant-supporting bases**
**FREQ= Variant allele frequency**
VarScan VCF Format

- 2 types: **VCF** and **Tabulated** Formats - **VarScan Specific**
  - VCF output doesn’t work with some GATK tools but works with other filtering or annotating tools (e.g.: annovar)

- **VarScan VCF format**: classic VCF header (#) but specific variant lines

<table>
<thead>
<tr>
<th>#CHROM</th>
<th>POS</th>
<th>ID</th>
<th>REF</th>
<th>ALT</th>
<th>QUAL</th>
<th>FILTER</th>
<th>INFO</th>
<th>FORMAT</th>
<th>GENO</th>
</tr>
</thead>
</table>

**PVAL = P-value from Fisher's Exact Test (not computed here: default value)**

**RBQ = Average quality of reference-supporting bases**

**ABQ = Average quality of variant-supporting bases**

**RDF / RDR = Depth of reference-supporting bases on forward/reverse strand**

**ADF / ADR = Depth of variant-supporting bases on forward/reverse strand**
VarScan VCF Format

- 2 types: **VCF** and **Tabulated** Formats - **VarScan Specific**
  - VCF output doesn’t work with some GATK tools but works with other filtering or annotating tools (e.g.: annovar)

**VarScan VCF format**: classic VCF header (#) but specific variant lines

<table>
<thead>
<tr>
<th>#CHROM</th>
<th>POS</th>
<th>ID</th>
<th>REF</th>
<th>ALT</th>
<th>QUAL</th>
<th>FILTER</th>
<th>INFO</th>
<th>FORMAT</th>
<th>GENO</th>
</tr>
</thead>
</table>

PVAL = P-value from Fisher's Exact Test (**not computed here : default value**)
RBQ = Average quality of reference-supporting bases
ABQ = Average quality of variant-supporting bases
RDF / RDR = Depth of reference-supporting bases on forward/reverse strand
ADF / ADR = Depth of variant-supporting bases on forward/reverse strand
VarScan Tabulated Format

- 2 types: **VCF** and **Tabulated** Formats - **VarScan Specific**
  - Tabulated output works with other VarScan Tools
  - By default on Galaxy, VarScan outputs a line for each base covered by the selected minimum coverage **even if there is no alternative variant**
  - **Pre-process**: use « **VarScan Filter** » to keep only variants

1 : Select the VarScan Results and leave the default parameters
2 : Execute
### VarScan Tabulated Format

<table>
<thead>
<tr>
<th>Chrom</th>
<th>Position</th>
<th>Ref</th>
<th>Cons</th>
<th>Reads1</th>
<th>Reads2</th>
<th>VarFreq</th>
<th>Strands</th>
<th>Strands</th>
<th>Qual1</th>
<th>Qual2</th>
<th>Pvalue</th>
<th>Map Qual1</th>
<th>Map Qual2</th>
<th>R1 + R1 - R2 + Rs2 - Alt</th>
</tr>
</thead>
<tbody>
<tr>
<td>chr12</td>
<td>113348849</td>
<td>C</td>
<td>Y</td>
<td>31</td>
<td>30</td>
<td><strong>49.18%</strong></td>
<td>2</td>
<td>2</td>
<td>27</td>
<td>27</td>
<td>0.98</td>
<td>1</td>
<td>1</td>
<td>19  12  25  5</td>
</tr>
<tr>
<td>chr12</td>
<td>113354329</td>
<td>G</td>
<td>R</td>
<td>72</td>
<td>2</td>
<td><strong>2.70%</strong></td>
<td>2</td>
<td>2</td>
<td>31</td>
<td>26</td>
<td>0.98</td>
<td>1</td>
<td>1</td>
<td>48  24  1  1</td>
</tr>
<tr>
<td>chr12</td>
<td>113357193</td>
<td>G</td>
<td>A</td>
<td>2</td>
<td>72</td>
<td><strong>97.30%</strong></td>
<td>1</td>
<td>2</td>
<td>28</td>
<td>24</td>
<td>0.98</td>
<td>1</td>
<td>1</td>
<td>2  0  45  27</td>
</tr>
<tr>
<td>chr12</td>
<td>113357209</td>
<td>G</td>
<td>A</td>
<td>0</td>
<td>77</td>
<td><strong>100%</strong></td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>29</td>
<td>0.98</td>
<td>0</td>
<td>1</td>
<td>0  0  51  26</td>
</tr>
</tbody>
</table>

**Cons**: Consensus Genotype of Variant Called (IUPAC code):

- **M** -> A or C
- **Y** -> C or T
- **D** -> A or G or T
- **W** -> A or T
- **V** -> A or C or G
- **R** -> A or G
- **K** -> G or T
- **B** -> C or G or T
- **S** -> C or G
- **H** -> A or C or T
# Exome/RNA merging

## Tag and merge multiple VarScan analysis (version 1.0.0)

### VarScan File 1:

1. **Exome Variant File**

### Tag Analysis 1:

2. **Associate the name « Exome » to these mutations »**

### VarScan File 2:

3. **RNA Variant File**

### Tag Analysis 2:

4. **Associate the name « RNA » to these mutations »**

### Add other(s) VarScan input files

5. **Execute**
Variants Visualization

- Keeping only common variants (concatenated results)

### Filter SNPs on same ref position (version 1.0.0)

<table>
<thead>
<tr>
<th>Chr</th>
<th>Start</th>
<th>Ref Allele</th>
<th>Alt Allele</th>
<th>A/A</th>
<th>A/B</th>
<th>103/72</th>
<th>85.12%/97.3%[Total:98.8%]</th>
<th>1</th>
<th>2/2</th>
<th>29/28</th>
<th>30/24</th>
<th>0.98/0.98</th>
</tr>
</thead>
<tbody>
<tr>
<td>chr12</td>
<td>113357193</td>
<td>G</td>
<td>A</td>
<td>2</td>
<td>103/72</td>
<td>85.12%/97.3%[Total:98.8%]</td>
<td>1</td>
<td>2/2</td>
<td>29/28</td>
<td>30/24</td>
<td>0.98/0.98</td>
<td></td>
</tr>
<tr>
<td>chr12</td>
<td>112888239</td>
<td>C</td>
<td>Y</td>
<td>52</td>
<td>4/56</td>
<td>2.2%/51.85%[Total:53.5%]</td>
<td>2</td>
<td>2/2</td>
<td>31/24</td>
<td>32/28</td>
<td>0.98/0.98</td>
<td></td>
</tr>
</tbody>
</table>

- Mutation in a splice site
- Mutation only in Exome-seq data (monoallelic expression)
Mutation in a splice site

- Mutation in the 3' **splice site** of the last exon of OAS1 (chr12:113357193) → alternative isoform using a cryptic 3' splice site upstream of the mutation

(Pickrell et al., 2012)
Visualization on IGV

- Mutation splice site → alternative isoform

1. Right click: sort alignments by base
2. Right click: collapsed
3. Unzoom (-/+ buttons)
Visualization on IGV

- Monoallelic expression (Right click : sort alignments by base)

<table>
<thead>
<tr>
<th>Chr</th>
<th>Start</th>
<th>End</th>
<th>RNA Freq</th>
<th>Exome Freq</th>
<th>Total Freq</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>112888239</td>
<td>112888240</td>
<td>2.2%</td>
<td>51.85%</td>
<td>53.5%</td>
</tr>
</tbody>
</table>

**RNA-seq**

**Exome-seq**
Visualization on IGV

- RNA editing (Right click : sort alignments by base)

chr12  |  113386779  |  C  |  S  |  319  |  168  |  34.50%  |  2  |  2  |  30  |  27  |  0.98  |  1  |  1  |  205  |  114  |  111  |  57  |  G  | RNA

RNA-seq

Exome-seq
Variants visualization with IGV

1: Click on the name to open the options

2: Click on « Local » to open it on IGV (already open)

- Look at those two variants:

<table>
<thead>
<tr>
<th>Chrom</th>
<th>Position</th>
<th>Ref</th>
<th>Cons</th>
<th>Reads 1</th>
<th>Reads 2</th>
<th>VarFreq</th>
<th>Strands 1</th>
<th>Strands 2</th>
<th>Qual 1</th>
<th>Qual2</th>
<th>Pval</th>
<th>Map Qual1</th>
<th>Map Qual2</th>
<th>R1</th>
<th>R1</th>
<th>R2</th>
<th>R2</th>
<th>Rs2</th>
<th>Alt</th>
</tr>
</thead>
<tbody>
<tr>
<td>chr12</td>
<td>113357193</td>
<td>G</td>
<td>A</td>
<td>2</td>
<td>72</td>
<td>97.30%</td>
<td>1</td>
<td>2</td>
<td>28</td>
<td>24</td>
<td>0.98</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>45</td>
<td>27</td>
<td></td>
<td>A</td>
</tr>
<tr>
<td>chr12</td>
<td>112888239</td>
<td>C</td>
<td>Y</td>
<td>52</td>
<td>56</td>
<td>51.85%</td>
<td>2</td>
<td>2</td>
<td>24</td>
<td>28</td>
<td>0.98</td>
<td>1</td>
<td>1</td>
<td>20</td>
<td>32</td>
<td>24</td>
<td>32</td>
<td>T</td>
<td></td>
</tr>
</tbody>
</table>
Variants visualization with IGV
## Variants visualization with IGV

<table>
<thead>
<tr>
<th>Chrom</th>
<th>Position</th>
<th>Ref</th>
<th>Cons</th>
<th>Reads 1</th>
<th>Reads 2</th>
<th>VarFreq</th>
<th>Strands 1</th>
<th>Strands 2</th>
<th>Qual 1</th>
<th>Qual2</th>
<th>Pval</th>
<th>Map Qual1</th>
<th>Map Qual2</th>
<th>R1</th>
<th>R2</th>
<th>Rs1</th>
<th>Rs2</th>
<th>Alt</th>
</tr>
</thead>
<tbody>
<tr>
<td>chr12</td>
<td>112888239</td>
<td>C</td>
<td>Y</td>
<td>52</td>
<td>56</td>
<td>51.85%</td>
<td>2</td>
<td>2</td>
<td>24</td>
<td>28</td>
<td>0.98</td>
<td>1</td>
<td>1</td>
<td>20</td>
<td>32</td>
<td>24</td>
<td>32</td>
<td>T</td>
</tr>
</tbody>
</table>