LEIDEN UNIVERSITY MEDICAL CENTER

Copy Number Variation: available tools

Jeroen F. J. Laros

Leiden Genome Technology Center

Department of Human Genetics

Center for Human and Clinical Genetics

Introduction

A literature review of available tools.

Introduction

A literature review of available tools.

Contents:

- Coverage based.
- Breakpoint based.
- Discordant pairs based.
- De novo insertions.
- Existing pipelines.

Coverage based methods

Coverage based methods

MoGUL:

- Lee et al.
- **seunghak@cs.cmu.edu**
- Common insertions and deletions in a population.
- Clustering / Bayesian network.
- Minor Allele Frequency (MAF) for each variant.
- Especially suitable for MAF $>$ 0.06 and indels $>$ 20bp.
- Assigns confidence to variant calls.
- Proposes mrFast as initial aligner, others may work too?

Coverage based methods

FREEC:

- Boeva et al.
- **`http://bioinfo-out.curie.fr/projects/freec`**
- Window size automatically optimised.
- Works with either a control dataset...
- or uses **`GC`** content.

Coverage based methods

FREEC:

- Boeva et al.
- **`http://bioinfo-out.curie.fr/projects/freec`**
- Window size automatically optimised.
- Works with either a control dataset…
- or uses **`GC`** content.

In Boeva et al., there is a good comparison:

- CNV-seq.
- SegSeq.
- RDXplorer.

Coverage based methods

FREEC:

- Boeva et al.
- **`http://bioinfo-out.curie.fr/projects/freec`**
- Window size automatically optimised.
- Works with either a control dataset...
- or uses **`GC`** content.

In Boeva et al., there is a good comparison:

- CNV-seq.
- SegSeq.
- RDXplorer.

Coverage based methods

Name	FREEC	CNV-seq	SegSeq	RDXplorer
Refs.		(Xie and Tammi, 2009)	(Chiang et al., 2009)	(Yoon et al., 2009)
Compatibility				
Paired-ends/Mate-pairs	+	-	-	+
Input format	SAM, bowtie, eland, BED, arachne, psl, SOAP	BLAT psl, SOLiD	arachne	BAM
Genomes	Any	Any	Any	Human
Strategy				
Fixed length moving window + statistical approach	+	+	-	+
Automatic evaluation of window size	+	+	+*	-
GC-content normalization	+	-	-	+
Normalization with a control dataset	+	+	+	-
Refinement of breakpoint boundaries	-	-	+	-

Table 1: Tools for CNA/CNV prediction. (Boeva et al. 2010)

* Window size is selected separately for each region according to the number of reads in this region in the control sample.

Coverage based methods

Name	FREEC	CNV-seq	SegSeq	RDXplorer
Refs.		(Xie and Tammi, 2009)	(Chiang et al., 2009)	(Yoon et al., 2009)
Compatibility				
Minimal number of reads	Any	Any	Any	High (>1.2x)
Segmentation of profiles	+	-	+	+
Automatic annotation of gain/loss regions	+	-	-	+
Copy Number prediction**	+	-	-	+***
Output				
CNA coordinates in one file	+	-	-	+
Graphical output	+****	+	-	-
Programming language	C/C++	Perl/R	Matlab	Java/R
Computational efficiency				
Memory usage	Low	Low	Low	>2GB of RAM
Running time*****	45s w/control sample 105s w/o control	422s	250s	N/A

Table 1: Tools for CNA/CNV prediction. (Boeva et al. 2010)

** Assignment of copy number changes to gain or loss.
*** Only for diploid genomes.
**** It is possible to visualize resulted copy number profiles and CNAs with a script R (makeGraph.R) included in the package.
***** Preprocessed HCC1143 data from (Chiang et al., 2009) on a 8-core 64-bit Linux machine.

Breakpoint based methods

Breakpoint based methods

GMAP / GSNAP:

- Wu et al.
- **`http://research-pub.gene.com/gmap`**
- Index based (global search).
- Mapping RNA or ESTs.
- Can use a database, can do without.
- SNP-tolerant alignment.
    - Potentially suitable for detection of all breakpoints.

Breakpoint based methods

GMAP / GSNAP:

- Wu et al.
- **`http://research-pub.gene.com/gmap`**
- Index based (global search).
- Mapping RNA or ESTs.
- Can use a database, can do without.
- SNP-tolerant alignment.
    - Potentially suitable for detection of all breakpoints.

BWA-SW:

- Li et al.
- **`http://bio-bwa.sourceforge.net`**
- BWA for long-read alignment.
- Chimeric reads.

Discordant pairs methods

Discordant pairs methods

Breakway:

- Clark et al.
- **`http://breakway.sourceforge.net`**
- Clustering of discordantly mapped paired-end reads.
- BAM input.
    - Designed for the use in pipelines.

Discordant pairs methods

Breakway:

- Clark et al.
- **`http://breakway.sourceforge.net`**
- Clustering of discordantly mapped paired-end reads.
- BAM input.
    - Designed for the use in pipelines.

GASV:

- Sindi et al.
- **`http://code.google.com/p/gasv`**
- Deletions, Invertions.
- Translocations.
- No insertions.
- BAM input.
- Poor documentation.

Discordant pairs methods

HYDRA_SV:

- Quinlan et al.
- **`http://code.google.com/p/hydra-sv`**
- Clustering of discordant pairs.
- No classifications, but can be done with BEDTools.
- Future plans:
  - Denovo assembly of insertions.
  - Split reads.
  - Automatic annotation.
- Toolkit includes useful tools like *bamToFastq*.

Discordant pairs methods

HYDRA_SV:

- Quinlan et al.
- **`http://code.google.com/p/hydra-sv`**
- Clustering of discordant pairs.
- No classifications, but can be done with BEDTools.
- Future plans:
  - Denovo assembly of insertions.
  - Split reads.
  - Automatic annotation.
- Toolkit includes useful tools like *bamToFastq*.
  - Possibly useful when we want to realign unmapped reads.

De novo insertions

De novo insertions

TIGRA_SV:

- Chen et al.
- **`http://tigrasv.sourceforge.net`**
- Targeted local assembly.
- Iterative graph routing assembly.
- Optional breakpoint library.
- BAM input.

De novo insertions

TIGRA_SV:

- Chen et al.
- **`http://tigrasv.sourceforge.net`**
- Targeted local assembly.
- Iterative graph routing assembly.
- Optional breakpoint library.
- BAM input.

Next-generation VariationHunter:

- Hormozdiari et al.
- **`http://compbio.cs.sfu.ca/strvar.htm`**
- Transposon insertion discovery.
- Discordant pairs / conflict resolution.
- $>$ 85% discovery.
- $>$ 90% accuracy.
- DIVET input format.

De novo insertions

NovelSeq:

- Hajirasouliha et al.
- Long novel sequence insertions.
- In essence a pipeline:
  1. Alignment: mrFast.
  2. Assembler: EULER-SR / ABySS.
  3. Contamination filter: BLAST.
  4. Clustering / anchoring: mrCAR.
  5. Local assembly: mrSAAB on the output of 3 and 4.
- SAM / DIVET input.

Miscellaneous

Miscellaneous

Dindel:

- Albers et al.
- **`http://sites.google.com/site/keesalbers`**
- Localized reassembly/realignment.
- Multiple BAM files.
- Estimate haplotype frequencies.

Miscellaneous

Mosaik:

- Lee & Strömberg.
- **`https://github.com/wanpinglee/MOSAIK`**
- Smith-Waterman gapped alignment.
- Discordantly mapped paired-end reads.
- Reference-guided assemblies.
- Toolkit includes:
  - MosaikBuild.
  - MosaikAligner.
  - MosaikSort.
  - MosaikAssembler.
- Input is raw sequence data.

Existing pipelines

Existing pipelines

SVMerge:

- Wong et al.
- **`http://svmerge.sourceforge.net`**
- BreakDancer, Pindel, SE cluster, RDXplorer, RetroSeq, . . .
- Expandable.
- BAM input.

Existing pipelines

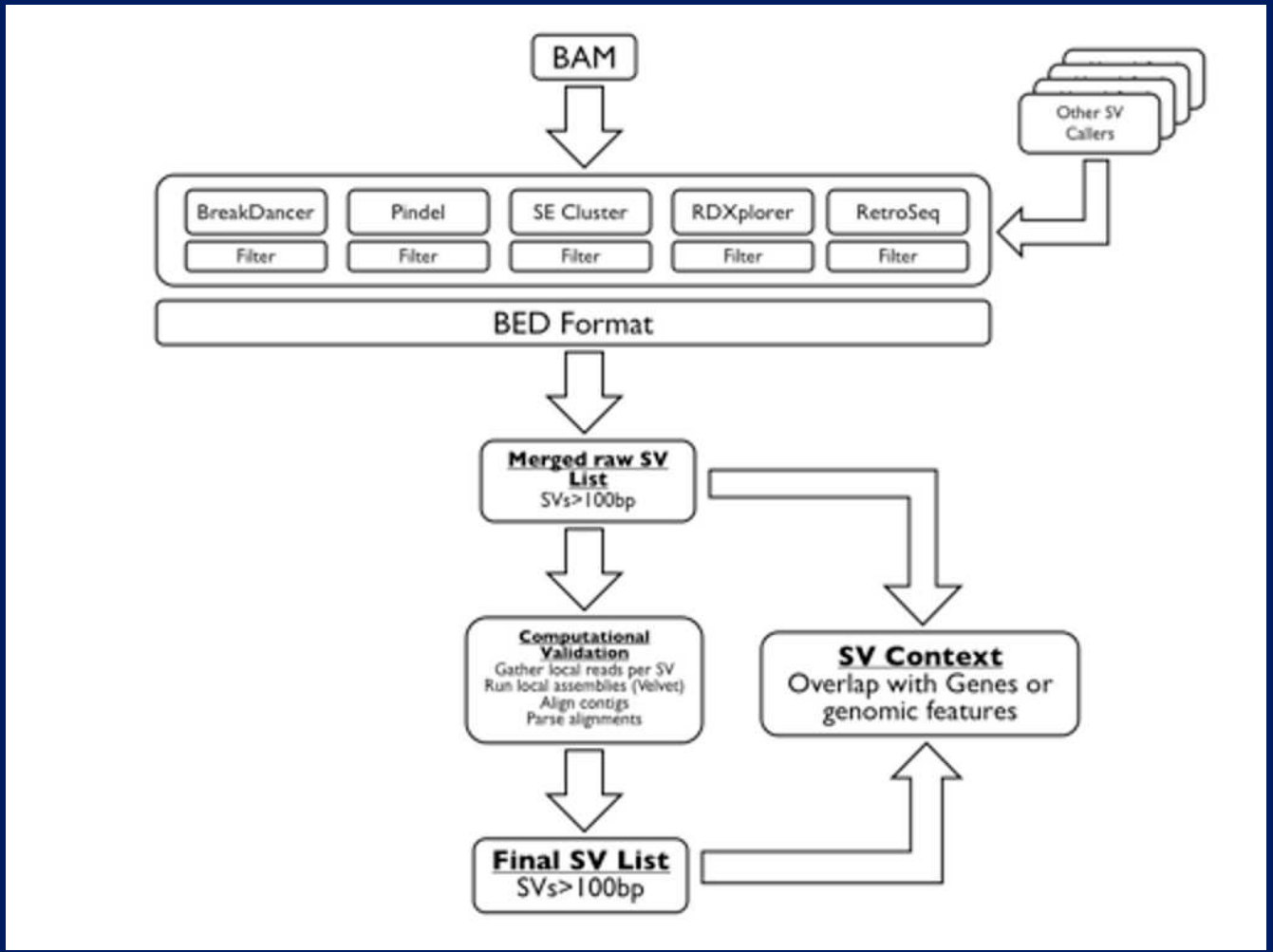

Figure 1: SVMerge pipeline.

Questions?

Acknowledgements:

Martijn Vermaat
Maarten van Iterson