ChimPipe: a pipeline for the precise detection of chimeras from RNA-seq data

Sarah Djebali, GenPhySE, INRA Toulouse

sarah.djebali-quelen@toulouse.inra.fr
Outline

- Introduction
  - Definition
  - Mechanisms
  - Motivation
- The ChimPipe method
- Benchmarking
  - Datasets
  - Results
- ENCODE human chimeras & validation by RT-PCR
- Summary and perspectives
Chimera definition

- A chimera is a transcript encoded by several genes in the genome (Gingeras, Nature review, 2009):

Note 1: genes A & B are called the parent genes of the chimera

Note 2: this definition depends on the annotation

Note 3: there is no constraint on the relative position of genes A & B (different chromosomes, different strands are allowed)

Note 4: here we focus on transcriptional connections between exons of genes A & B
Mechanisms that can explain the formation of chimeras

- **Genomic mechanisms:**
  - Genomic rearrangements (translocation, deletion, inversion); in this case the chimera is also called a *fusion gene*.

- **Transcriptional mechanisms:**
  - *In vivo*:
    - Polymerase read-through
    - Trans-splicing
    - Polymerase slippage through Short Homologous Sequences (SHS)
  - *In vitro*:
    - Reverse transcriptase template switching
Genomic rearrangements can lead to chimeras

Here the 2 genes can be anywhere in the genome
Polymerase read-through can generate chimeras

The 2 genes are on the same chromosome, same strand and adjacent. The most common pattern is to skip the last exon of gene A and the first exon of gene B. The junction has to harbour canonical splice sites.
Trans-splicing can generate chimeras

The 2 genes can be anywhere in the genome but close in the 3D space (they are supposed to belong to the same 'transcription factory'). The chimeric junction has to harbour canonical splice sites.
Transcriptional slippage through Short Homologous Sequences (SHS) can generate chimeras

The 2 genes can be anywhere in the genome but close in the 3D space. No canonical splice sites but short homologous sequence at the junction

From Li et al, Journal of Molecular Evolution, 2009
Reverse transcriptase (RT) template switching can generate artefactual chimeras

RT-PCR sequence

Genome

Real transcript

Hyp: 2 short homologous sequences close in 3D:
→ RT directly jumps from 5' to 3' part of the sequence (internal/external RT template switching).

RT-PCR sequence

Same as for polymerase slippage but technical rather than biological artefact
Importance of chimeras

- They represent biomarkers for certain cancer types:

- They are means to create novel transcripts and proteins:
  - therefore potentially altering cells, individuals or populations' phenotype (Akiva, GR, 2006, Morgenstern et al, GR, 2012, Greger, PLoS one, 2014)

- Functionally validated chimeras are few but exist:
  - Wu et al, GR, 2014, showed that a trans-spliced transcript tRMST is responsible for maintaining cells' pluripotency
  - Babiceanu et al, NAR, 2016, knocked down 2 widely expressed chimeras in non-neoplastic cell lines, resulting in significant reduction in cell growth and motility
Computational identification of chimeras from RNA-seq

- RNA-seq is a tool of choice for surveying the transcriptome, allowing more precise transcript characterization (structure, expression) than previous microarray-based assays
- Many programs have been developed to identify chimeric transcripts from RNA-seq, and generally use a 3 step approach (Wang et al, Briefings in bioinf, 2012):
  1. Read mapping & filtering to only keep reads yielding chimera evidence
  2. Chimeric junction detection
  3. Chimera assembly and filtering
Computational identification of chimeras from RNA-seq

These programs heavily rely on a **mapper** to map the reads to the genome (and transcriptome) and make use of 2 kinds of reads:

- **Discordant paired end (PE) reads**: reads where the 2 mates map to 2 different genes; relatively easy to find but provide rough indication of chimeric junction location

- **Split-reads**: reads where one part maps to a gene and another part to another gene; more prone to mapping artefacts but provide exact junction location

Depending on whether the program uses discordant paired end reads only, split-reads only, or both, their approach is called **whole paired-end**, **direct fragmentation**, or **paired-end + fragmentation** (Beccuti, 2013)
Issues with current programs

Current programs:

- tend to output many false positives (Carrara et al, BMC bioinformatics, 2013)
- provide widely different outputs on the same input sample (Carrara et al, BMC bioinformatics, 2013)
- are designed to detect fusion genes in cancer and therefore are not always able to find:
  - read-through events
  - exact junction coordinates and several isoforms per gene pair, thus making more difficult, or even impairing, important downstream functional analyses/validation of these chimeras
ChimPipe

- A modular method
- Uses the paired-end + fragmentation approach for the complementarity of the 2 types of reads (sensitivity and exact junction detection)
- Uses a set of stringent filters (specificity)
- Can detect any kind of chimera from illumina paired-end RNA-seq from both tumor and normal samples
- Can in principle work on any eukaryote with a genome and an annotation available (human, mouse, drosophila tested)
- Can take in either sequenced reads or aligned reads (bam file)
- Provides a standard alignment bam file, therefore allowing standard downstream RNA-seq analyses
1. Read mapping

2. Chimera detection

3. Chimera filtering
1. Read mapping

RNA-seq → Healthy or Tumor cells

Paired-End (PE) reads

GEMtools rna-pipeline

Legend:
- Unmapped PE
- Concordant PE
- Consistent PE
- Inconsistent PE
- Split-read (genome)
- Split-read (transcriptome)

Genome → Mapped reads

Transcriptome → Mapped reads

De novo (1) → Unmapped reads

De novo (2) → Split-mapped reads
2. Chimera detection

- **Mapped reads**
- **Unmapped reads**
- **Split-mapped reads** (different chromosomes and strands)

**ChimSplice**

**Detection of chimeric splice junctions**
- **Gene A**
- **Gene B**
- **Chimeric junction cluster**
  - beg
  - GT
  - AG
  - end
- **Cluster split-reads into splice junctions**
- **Staggered Split-reads**
- **Consensus junction**

**ChimPE**

**Detection of supporting Paired-ends (PE)**
- **Gene A**
- **Gene B**
- **Discordant PE**
- **Evaluate discordant PE junction consistency**
- **Consistent PE**
- **Inconsistent PE**

**Chimera candidates**
3. Chimera filtering

Chimera candidates

(iv) ChimFilter

At exon boundaries
- >= 3 Total support
- >= 1 Split-reads
- >= 1 Consistent PE

Not at exon boundaries
- >= 6 Total support
- >= 3 Split-reads
- >= 3 Consistent PE

Mitochondrial Chr
Gene biotype filter
Homology filter

Highly supported chimeras

Chimera categories
- Readthrough
- Inverted
- Intrachromosomal
- Interstrand
- Interchromosomal

Supporting evidence
- 3 consistent PE
- 4 split-reads

Chimeric transcript
ChimPipe implementation

- **GitHub:**
  - https://github.com/Chimera-tools/ChimPipe

- **Documentation:**

- **Notes:**
  - ChimPipe automatically detects:
    - whether data is **directional**, and the mate configuration when it is
    - the **quality** offset encoding
  - ChimPipe associates a **class** (read-through, intrachromosomal, inverted, interstand, interchromosomal) to each chimera
  - ChimPipe provides both a complete and a final junction set, and gives the **reasons** for filtering junctions out
Benchmark data

- **Simulated** unstranded paired end RNA-seq data
  - 3 different read **lengths** (50bp, 76bp, 101bp)
  - sequencing **error** obtained from real data of the same length
  - both chimeras and **normal** transcripts included (including parent genes of the chimeras)
  - chimeras generated from **5 classes** (read-through, intrachromosomal, inverted, interstrand, interchromosomal)

- **Gold standard cancer** unstranded paired end RNA-seq data (50bp) with associated **validated chimeras**
  - leukemia/melanoma (7 cell lines, several insert sizes)
  - breast cancer (4 cell lines, several insert sizes)
Gencode v19 annotation

ChimSim
- 250 chim tr:
  - 50 read through
  - 50 intrachrom
  - 50 inverted
  - 50 interstrand
  - 50 interchrom

TranscriptSample
- 60% from coding and long non-coding gene transcripts

498 parent transcripts

101,961 transcripts

102,149 normal transcripts

102,399 transcripts

ART*
- 32.3 million 50 bp PE reads

ART*
- 21.1 million 76 bp PE reads

ART*
- 15.7 million 101 bp PE reads

* version 2.3.7
Benchmark data: gold standard cancer datasets

- **Validated fusion genes** (gene pairs + sequences) from 3 cancer types (leukemia, melanoma, breast cancer), from 3 different papers (Berger et al, GR, 2010; Edgren et al, GB, 2011; Kangaspeska et al, GB, 2011)

- We enriched these fusion genes by adding **precise junction coordinates** (DNA sequence blatted to genome + manual curation)

<table>
<thead>
<tr>
<th>Cancer dataset</th>
<th>Cell line</th>
<th>Tumor type</th>
<th>Number of validated fusion genes</th>
<th>Number of validated fusion junctions</th>
<th>Reference paper(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Berger</td>
<td>K562</td>
<td>Leukemia</td>
<td>3</td>
<td>3</td>
<td>Berger et al, Genome Research, 2010</td>
</tr>
<tr>
<td></td>
<td>501 Mel</td>
<td></td>
<td>4</td>
<td>5</td>
<td>Berger et al, Genome Research, 2010</td>
</tr>
<tr>
<td></td>
<td>M000216</td>
<td></td>
<td>1</td>
<td>1</td>
<td>Berger et al, Genome Research, 2010</td>
</tr>
<tr>
<td></td>
<td>M000921</td>
<td>Melanoma</td>
<td>2</td>
<td>3</td>
<td>Berger et al, Genome Research, 2010</td>
</tr>
<tr>
<td></td>
<td>M010403</td>
<td></td>
<td>1</td>
<td>1</td>
<td>Berger et al, Genome Research, 2010</td>
</tr>
<tr>
<td></td>
<td>M980409</td>
<td></td>
<td>1</td>
<td>1</td>
<td>Berger et al, Genome Research, 2010</td>
</tr>
<tr>
<td></td>
<td>M990802</td>
<td></td>
<td>2</td>
<td>2</td>
<td>Berger et al, Genome Research, 2010</td>
</tr>
<tr>
<td></td>
<td>All</td>
<td>All</td>
<td>14</td>
<td>16</td>
<td>Berger et al, Genome Research, 2010</td>
</tr>
<tr>
<td>Edgren</td>
<td>KPL-4</td>
<td>Breast cancer</td>
<td>3</td>
<td>3</td>
<td>Edgren et al, Genome Biology, 2011</td>
</tr>
<tr>
<td></td>
<td>MCF-7</td>
<td></td>
<td>6</td>
<td>8</td>
<td>Edgren et al, Genome Biology, 2011; Kangaspeska et al, PLOSone, 2012</td>
</tr>
<tr>
<td></td>
<td>SK-BR-3</td>
<td></td>
<td>10</td>
<td>10</td>
<td>Edgren et al, Genome Biology, 2011</td>
</tr>
<tr>
<td></td>
<td>All</td>
<td>All</td>
<td>40</td>
<td>46</td>
<td>Edgren et al, Genome Biology, 2011; Kangaspeska et al, PLOSone, 2012</td>
</tr>
</tbody>
</table>
## State of the art benchmarked programs

<table>
<thead>
<tr>
<th>Program</th>
<th>Why was it chosen for the benchmark?</th>
<th>Underlying mapper?</th>
<th>Chimera detection approach</th>
<th>What are the false positive filters used?</th>
<th>Publication</th>
</tr>
</thead>
</table>
| FusionMap       | Best according to Carrara et al paper (BMC Bioinf, 2013), and known to be good in general            | Modified GSNAP               | Direct fragmentation        | - expression  
- black gene list  
- paralogs                                                | Ge et al, Bioinformatics, 2011 (original paper)    |
| PRADA           | Used in precursor melanoma paper (Berger et al, GR, 2010)                                            | BWA                          | Paired end + fragmentation approach | - split read with mate in gene  
- similarity between genes                                 | Torres-Garcia et al, Bioinformatics, 2014 (application note) |
| Chimerascan     | Good and used in precursor paper Maher et al paper (PNAS, 2009) about RNA-seq in cancer             | Bowtie                       | Paired end + fragmentation approach | - expression  
- insert size  
- short homologous sequences                               | Iyer et al, Bioinformatics, 2011 (application note) |
| TopHatFusion    | Well known, one of the first, used extensively                                                     | Bowtie                       | Paired end + fragmentation approach | - expression  
- short homologous sequences  
- multi-copy genes  
- repeats  
- annotated gene on at least one side                      | Kim et al, Genome biology, 2011 (methods)          |
Evaluation levels and measures

Sn = \frac{TP}{TP + FN}

Pr = \frac{TP}{TP + FP}

Sn = sensitivity; Pr = precision; TP = true positive; FN = false negative; FP = false positive

A false negative is something that should be predicted and is not, a false positive the opposite
Gene pair level assessment
- ChimPipe is second after chimerascan which predicts many more cases on real data

Exact junction level assessment
- ChimPipe is the best for both kinds of datasets
## Resources needed on the simulated sets with 4 cpus

<table>
<thead>
<tr>
<th>Program</th>
<th>Max RAM used in Gb</th>
<th>Avg cumulative wallclock time in hours</th>
<th>Number of commands to launch</th>
</tr>
</thead>
<tbody>
<tr>
<td>ChimPipe</td>
<td>34</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>FusionMap</td>
<td>12</td>
<td>0.5</td>
<td>1</td>
</tr>
<tr>
<td>PRADA</td>
<td>36</td>
<td>7</td>
<td>3 (make mapping script, mapping, compute fusion)</td>
</tr>
<tr>
<td>Chimerascan</td>
<td>4.5</td>
<td>8</td>
<td>1</td>
</tr>
<tr>
<td>TophatFusion</td>
<td>8</td>
<td>4.5</td>
<td>2 (mapping + filtering)</td>
</tr>
</tbody>
</table>

FusionMap is the tool that performs best overall after ChimPipe, however its behaviour depends on the read length with 76 bp reads less well handled than 50bp and 101bp reads.
Chimeras on 108 ENCODE human RNA-seq datasets and validation by RT-PCR
108 ENCODE CSHL stranded PE 76bp long RNA-seq experiments (illumina), done in 2 bio-replicates (depth: 200 million reads):

- 3 RNA fractions (long means ≥ 200nt):
  - long polyA+
  - long polyA-
  - total long

- 6 cell compartments:
  - whole cell
  - nucleus
  - nucleolus
  - chromatin
  - nucleoplasm
  - Cytosol

- 16 cell lines (6 cancer + 10 normal)
In 2011, using Gencode v7 annotation and an ancestor of ChimPipe

400 chimeric junctions with $\geq 10$ staggered split-reads in $\geq 1$ expt

232 chimeric junctions not already annotated

116 chimeric junctions where 5' and 3' parts fall in 1 gene and this gene is protein coding

69 chimeric junctions where 5' and 3' parts fall in coding sequences

33 chimeric junctions that are read-through, intra or inter-chromosomal

6 chimeric junctions sent to RT-PCR for validation*

* chosen based on: - high expression in encode cell lines, or - expression in many encode cell lines including the ones available at CBMSO for RT-PCR
Out of 6 junctions attempted to be validated by RT-PCR, 3 were successfully validated.
RT-PCR validated junctions

- The 3 RT-PCR validated junctions were further cloned and sequenced (Sanger sequencing).
- Only 1 case (UBA-WTIP) maintained the frame of the 2 parent genes and was therefore completely sequenced.
  - 3 novel transcript structures, of which 1 was further analyzed and shows the 2 first (ThiF and UAE_Ubl) domains of the 5' UBA protein to be connected to the last 3 (LIM) domains of the 3' WTIP protein → potential novel role?
- The other 2 cases (PICALM-SYTL2 and RPL38-TTYH2) gave rise to 2 novel but incomplete transcript structures and could have a different stability than the 5' parent transcript and also affect its expression.
Protein domain analysis with SMART

1) UBA2 wild type protein (640 aa) domains

- ThiF domain = Ubiquitin-activating enzyme (E1 enzyme) (Pfam)
- UAE_UbL domain = C-term. domain of ubiquitin-activating enzyme and SUMO-activating enzyme 2 (Pfam)
- UBA2_C domain = SUMO-activating enzyme subunit 2 C-terminus (Pfam)
- Purple = Low complexity regions (SEG program)
- LIM domains = Zinc-binding domains. Some LIM domains bind protein partners via tyrosine-containing motifs (SMART)

2) WTIP wild type protein (430 aa) domains

- LIM domains

3) UBA-WTIP chimeric protein (788 aa) domains

- ThiF domain = Ubiquitin-activating enzyme (E1 enzyme) (Pfam)
- UAE_UbL domain = C-term. domain of ubiquitin-activating enzyme and SUMO-activating enzyme 2 (Pfam)
- UBA2_C domain = SUMO-activating enzyme subunit 2 C-terminus (Pfam)
- Purple = Low complexity regions (SEG program)
- LIM domains = Zinc-binding domains. Some LIM domains bind protein partners via tyrosine-containing motifs (SMART)
Summary

- ChimPipe, a method to detect any kind of chimeras from illumina paired-end RNA-seq data of eukaryotic species with a genome and a gene annotation available:
  - Exact chimeric junction detection
  - Several isoforms per gene pair detection
  - High precision and good sensitivity
- Applied to 108 encode RNA-seq datasets, it identifies 33 highly expressed chimeras of which 6 (probably read-through) were attempted to be validated by RT-PCR, and of which 3 succeeded:
  - Further cloning and sequencing revealed new transcript structures of which 3 maintain the frame of the 2 parent genes and therefore create a novel protein with the domains from the 2 parent genes
Perspectives

**Biologically:**
- Investigate the role of the novel chimeric protein found
- Apply chimpipe to many animal genomes and individuals in order to study chimera evolution and connect some of them to individuals' phenotypes
- Compare to HiC data to have a hint on mechanisms
- Use RNA FISH to confirm certain interesting cases

**Computationally:**
- Provide chimeric transcripts compatible with the junction
- Gemtools extension so as to treat internally split reads on different chromosomes or strands, and to have an internal scoring of those together with the other reads
- Implement in a pipeline language such as nextflow
Acknowledgements

**CRG/CNAG/BSC:**
- Bernardo Rodríguez Martín (CRG, BSC)
- Emilio Palumbo (CRG)
- Paolo Ribeca (CRG, CNAG)
- Santiago Marco-Sola (CNAG)
- Thasso Griebel (CNAG)
- Roderic Guigó (CRG)

**CBMSO:**
- Begoña Aguado
- Graciela Alonso
- Alberto Rastrojo
Possible mechanisms explaining the formation of chimeras

**Read-through transcription**


**Trans-splicing**


**Genomic rearrangements**


**SHS transcriptional slippage**

Benchmark data: simulated data

- **Chimeric transcripts:** 250 chimeras were generated from Gencode v19 protein coding transcripts: 50 read-throughs, 50 intra-chromosomal, 50 inverted, 50 inter-strand, 50 inter-chromosomal

- **Normal transcripts:** 60% of transcripts were sampled from the 169,935 Gencode v19 protein-coding and lncRNA genes, and added to the 498 parent transcripts of the 250 chimeras

- **Final transcripts:** 250 chimeric + 102,149 normal transcripts

- **Read simulation on final transcripts:** use ART v2.3.7 (ref) to simulate unstranded 50, 76 & 101bp paired end reads with:
  - Fragment length of 200+-20, 250+-25 and 300+-30 respect.
  - Sequencing errors obtained from real data of the corresponding lengths
  - Coverage 20
Results without read-throughs
Result for each class of chimeras (gene pair level)
Result for each class of chimeras (junction level)
Output of the 5 programs on the breast cancer dataset

The program with less unique chimeras is PRADA, then ChimPipe, FusionMap, Chimeraşcan.

There are many chimeras common to chimpipe and 2 other programs.
## Distance between predicted and true junction

### Simulation sets

<table>
<thead>
<tr>
<th>Program</th>
<th>50 bp</th>
<th></th>
<th></th>
<th>101 bp</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td># junctions</td>
<td>dist_avg+-dist_std</td>
<td># junctions</td>
<td>dist_avg+-dist_std</td>
</tr>
<tr>
<td>ChimPipe</td>
<td>158</td>
<td>0+-0</td>
<td>163</td>
<td>0+-0</td>
</tr>
<tr>
<td>FusionMap</td>
<td>57</td>
<td>0+-0</td>
<td>141</td>
<td>0.03+-0.34</td>
</tr>
<tr>
<td>PRADA</td>
<td>155</td>
<td>0+-0</td>
<td>150</td>
<td>0+-0</td>
</tr>
<tr>
<td>Chimerascan</td>
<td>193</td>
<td>2.85+-15.04</td>
<td>189</td>
<td>119.06+-1449.07</td>
</tr>
<tr>
<td>TophatFusion</td>
<td>141</td>
<td>732 227+-6 253 300</td>
<td>135</td>
<td>764 770+-6 389 800</td>
</tr>
</tbody>
</table>

### Positive sets

<table>
<thead>
<tr>
<th>Program</th>
<th>Berger</th>
<th></th>
<th>Edgren</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td># junctions</td>
<td>dist_avg+-dist_std</td>
<td># junctions</td>
</tr>
<tr>
<td>ChimPipe</td>
<td>11</td>
<td>0+-0</td>
<td>35</td>
</tr>
<tr>
<td>FusionMap</td>
<td>6</td>
<td>0+-0</td>
<td>23</td>
</tr>
<tr>
<td>PRADA</td>
<td>11</td>
<td>0+-0</td>
<td>28</td>
</tr>
<tr>
<td>Chimerascan</td>
<td>12</td>
<td>592.17+-1866.81</td>
<td>37</td>
</tr>
<tr>
<td>TophatFusion</td>
<td>7</td>
<td>2+-0</td>
<td>30</td>
</tr>
</tbody>
</table>
Output of the 5 programs on the melanoma dataset (gene pair level)
Output of the 5 programs on the breast cancer dataset (junction level)
Output of the 5 programs on the melanoma dataset (junction level)
Filtering reasons for chimpipe on the breast cancer dataset
A) With internal exons

B) Without internal exons
Highlighted in green the longest open reading frame (ORF) preserving the annotated UBA2 and WTIP CDS sequences. The ORF starts in UBA2 (NM_005499.2, RefSeq) annotated start codon and stops in WTIP (NM_001080436.1, RefSeq) stop codon, so this UBA2-WTIP chimeric transcript has the potential to encode a chimeric protein.
I read all the documents the people from Madrid sent us and I guess I already have an idea of what do they have done and how do they have done it. This is a brief summary:

1) Select 6 cases for validation from the list you sent based on their level of expression, their recurrence and the availability of cell lines.

2) Validation of chimeric junctions through RT-PCR + sanger sequencing in several cell lines. 3/6 validated

3) Verify the genes are not fused at genomic level for the 3 validated cases through PCR. No underlaying genomic rearrangement in any case, so they are transcriptional chimeras

4) Analysis of the theoretical chimeric mRNAs based on the chimeric junctions for the 3 validated cases from 2). This analysis concluded that only UBA2-WTIP has the potential to encode for a chimeric protein. The other cases are not in frame. They have a premature stop codon, so if they were translated they would lead to a truncated protein or would be degraded through non-sense mediated decay (I added this last point, it was not in their docs).
5) Amplification and sequencing of the full sequence of 3 different UBA2-WTIP chimeric mRNA isoforms. All of them are consistent with the chimeric junction reported by ChimPipe.

So, what I have done is to take the 3 sequences produced in 5 and study their protein coding potential.

- I already finish the analysis of one isoform and I confirm this chimeric transcript has the potential to encode for a chimeric protein. I send you a document with the results of the analysis. There are several details I would like to talk with you at one point. Also, please let me know if something is not clear.

- Now, I would need to do the same with the 2 remaining validated isoforms.
Chimeric junctions from Encode RNAseq experiments

Number of chimeric junctions seen by at least 10 staggered split-mappings

<table>
<thead>
<tr>
<th>Min.</th>
<th>1st Qu.</th>
<th>Median</th>
<th>Mean</th>
<th>3rd Qu</th>
<th>Max</th>
<th>Number of exp</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>3</td>
<td>12.5</td>
<td>16.98</td>
<td>27</td>
<td>74</td>
<td>108</td>
</tr>
</tbody>
</table>

- Total number of chimeric junctions seen by more than 10 staggered split-mappings, i.e. highly expressed = 400 (was 4,881 using all split-mappings).
  - a junction is seen by more than 20 experiments on average.

- On the 400 highly reliable junctions:
  - 386 are intra-chromosomal (the closer the more expressed),
  - 14 are inter-chromosomal (including two known genomic rearrangements: BCR-ABL (chr9-chr22) and ETO-AML1 (chr8-chr21)).

- On the 386 intra-chromosomal ones:
  - all are on the same strand (although not a feature of grape or gem),
  - distribution of distance is the following (1 case>100Mb on chr11):

<table>
<thead>
<tr>
<th>Min.</th>
<th>1st Qu.</th>
<th>Median</th>
<th>Mean</th>
<th>3rd Qu</th>
<th>Max</th>
<th>Number of junctions</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1,424</td>
<td>7,708</td>
<td>423,800</td>
<td>40,230</td>
<td>107,000,000</td>
<td>386</td>
</tr>
</tbody>
</table>
Classification of Encode chimeras

- On the 386 intra-chromosomal junctions, 168 connect exons of gene A and B and exons of gene A only (due to improvement of annotation from v3c to v7) → clear read-through events (usually very close, discarded),

- The remaining 218 have the following distance distribution:

<table>
<thead>
<tr>
<th>Min.</th>
<th>1st Qu.</th>
<th>Median</th>
<th>Mean</th>
<th>3rd Qu</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>231</td>
<td>6,274</td>
<td>25,020</td>
<td>736,900</td>
<td>89,150</td>
<td>107,000,000</td>
</tr>
</tbody>
</table>

- They may be investigated for mechanism and compared to chimeras found in other datasets / by other methods.

- Unexpectedly, 102/218 are not in the expected genomic order!!

- Distance distribution (bad order ones are a bit closer):

<table>
<thead>
<tr>
<th># cases / type</th>
<th>Min.</th>
<th>1st Qu.</th>
<th>Median</th>
<th>Mean</th>
<th>3rd Qu</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>102 / bad order</td>
<td>621</td>
<td>5,840</td>
<td>20,120</td>
<td>421,300</td>
<td>91,490</td>
<td>21,140,000</td>
</tr>
<tr>
<td>116 / good order</td>
<td>231</td>
<td>9,054</td>
<td>27,710</td>
<td>1,014,000</td>
<td>85,670</td>
<td>107,000,000</td>
</tr>
</tbody>
</table>
Chimeric junctions not in expected genomic order

There characteristics with respect to expected order junctions are:

- a bit less distant,
- as prevalent,
- less present in cancer cell lines,
- a bit less in polyA-,
- a bit more in nucleus.

In theory they could be due to:

- genome rearrangement,
- exon shuffling (Al-Balool et al., Genome Research, 2011),
- circular RNA (Salzman et al., PLoS ONE, 2012),

However:

- exon shuffling is not supposed to be so prevalent (I find the same proportion when considering intra-genic junctions),
- circular RNA is found more in polyA- cytosolic RNAs wrt a+ nuc.
Issues with current benchmarks

- Current benchmarks (carrara et al, nar paper, others?):
  - focus on cancer fusion genes, and therefore do not include read-through transcripts
  - do the assessment at the gene pair level, or at the junction level but allowing 20 bp difference with the true junction, and not at the exact junction level, sometimes even considering B-A as true positive when A-B must be found
  - Use simulation data that is not always very realistic, not always including the parent genes of the chimeras
  - Obtain different results for real and simulated data
How ChimPipe deals with the current issues

- **Too many false positives:**
  - Combines paired end and split reads
  - Uses several complementary filters

- **Imprecise junction coordinates:**
  - Uses gemtools and the gem rna-mapper, which are able to exhaustively split-map reads taking bases' quality and extended consensus donor/acceptor sequences into account, and with no constraint on the location of the 2 parts of the junction (rna-mapper)

- **Unordered gene pairs:**
  - Use of directionality information when data is directional and consensus donor/acceptor sequences otherwise
Confirmation par 5C de 74% des jonctions chimériques détectées par la technique de RACEarray

- 638 connexions exoniques découvertes par RACEarray sur le chromosome 21 humain (liens bleus)
- 74% d'entre elles sont confirmées par 5C (liens jaunes)