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

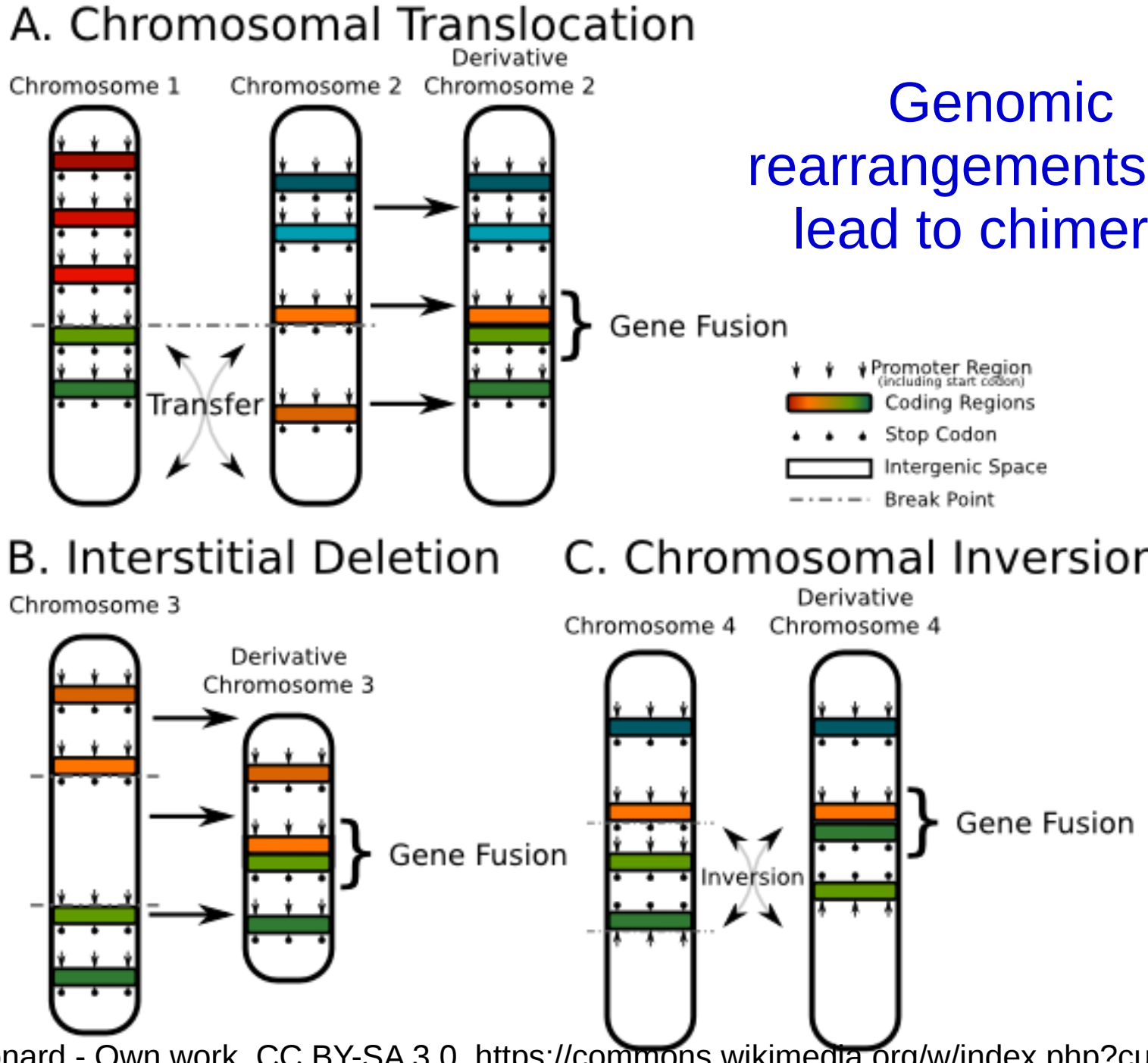
Chimera definition

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

- Note1: genes A & B are called the parent genes of the chimera
- Note2: this definition depends on the annotation
- Note3: there is no constraint on the relative position of genes A & B (different chromosomes, different strands are allowed)
- Note4: here we focus on transcriptional connections between **exons** of genes A & B

Mechanisms that can explain the formation of chimeras

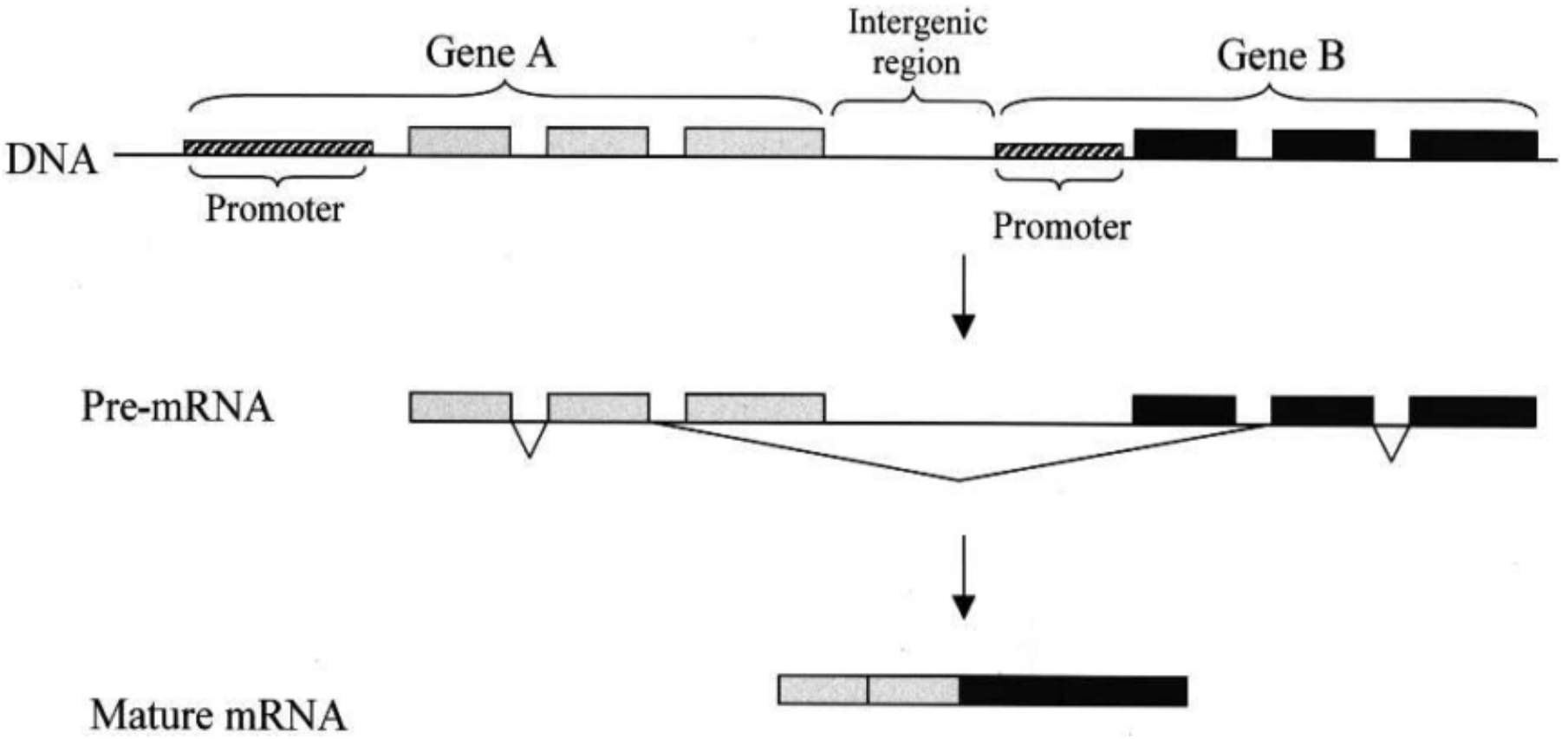
- Genomic mechanisms:
  - Genomic rearrangements (translocation, deletion, inversion); in this case the chimera is also called a fusion gene
- Transcriptional mechanisms:
  - *In vivo:*
    - Polymerase read-through
    - Trans-splicing
    - Polymerase slippage through Short Homologous Sequences (SHS)
  - *In vitro:*
    - Reverse transcriptase template switching



Genomic rearrangements can lead to chimeras

Here the 2 genes can be anywhere in the genome

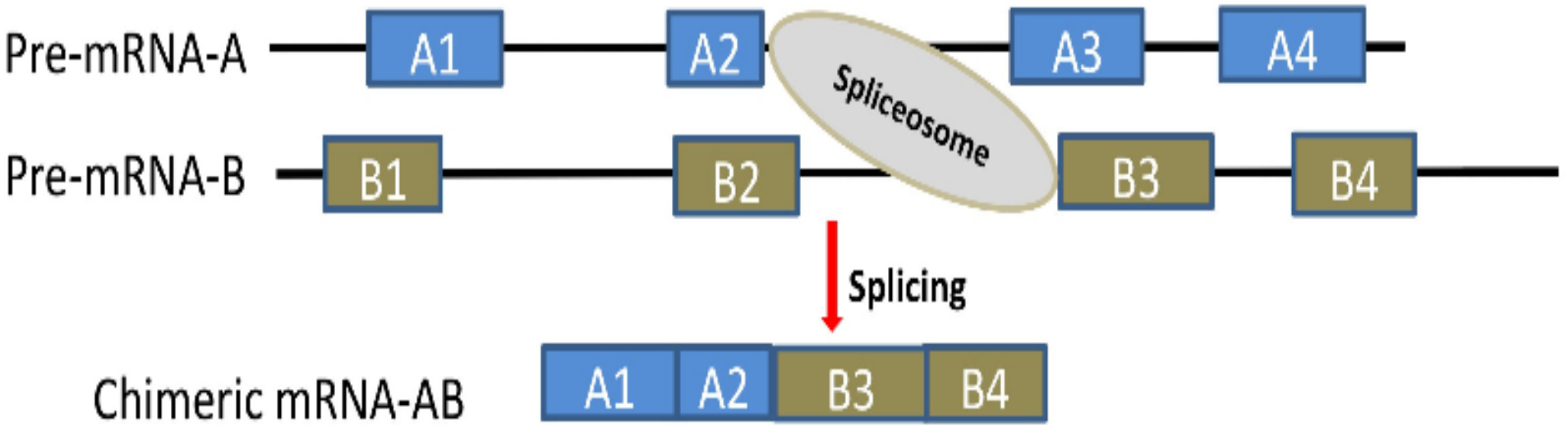
Polymerase read-through can generate chimeras



From Akiva et al, Genome research, 2006

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.

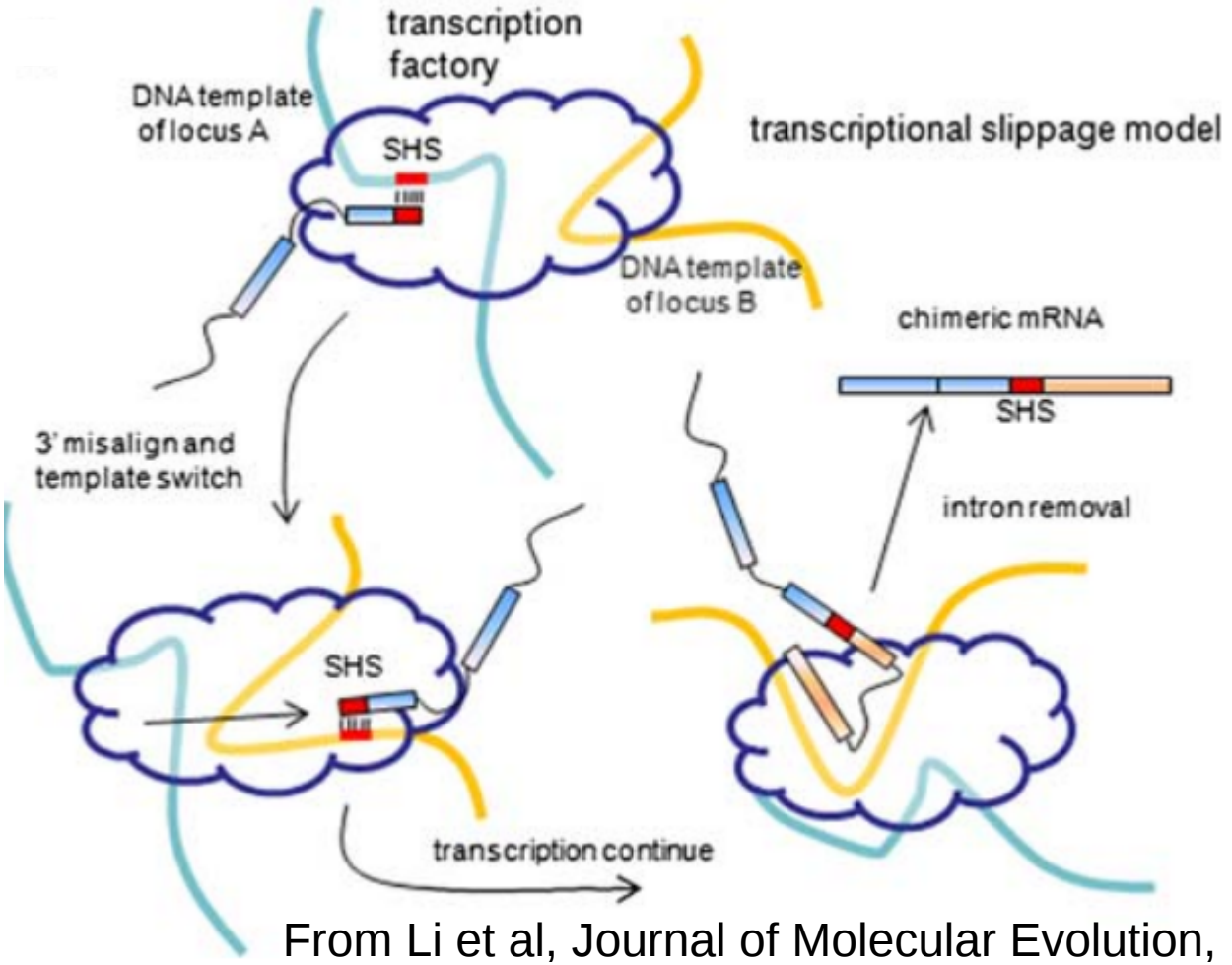
Trans-splicing can generate chimeras



From Zhou et al, BMB reports, 2012

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

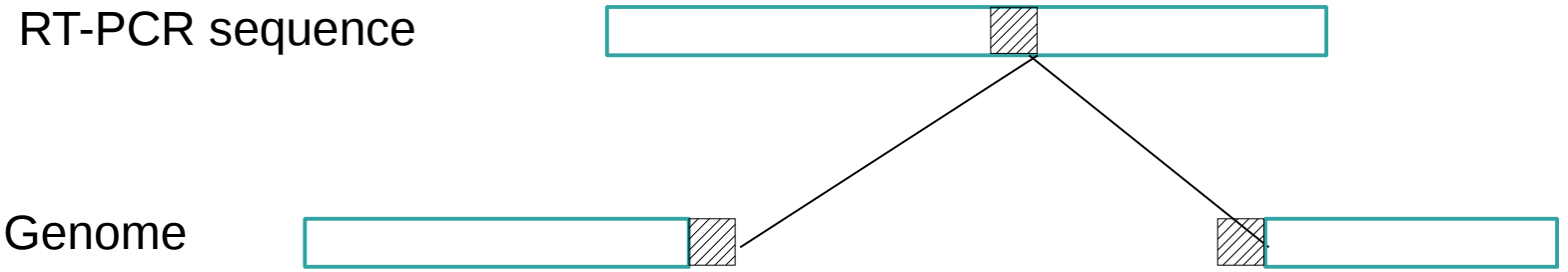
Transcriptional slippage through Short Homologous Sequences (SHS) can generate chimeras

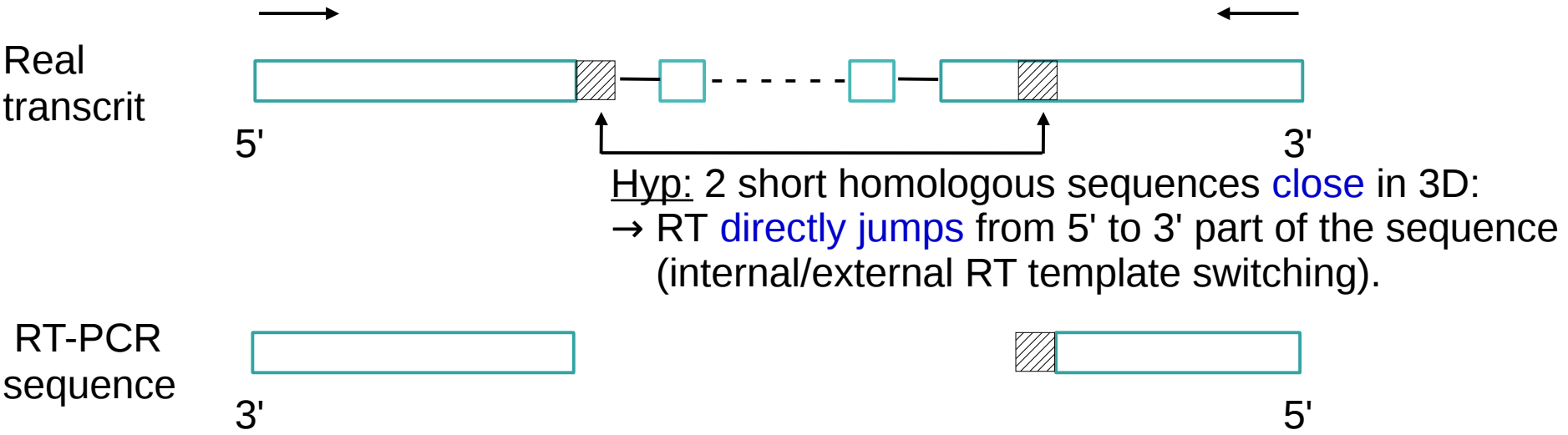


From Li et al, Journal of Molecular Evolution, 2009

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

Reverse transcriptase (RT) template switching can generate artefactual chimeras





Same as for polymerase slippage but technical rather than biological artefact

Importance of chimeras

- They represent biomarkers for certain cancer types:
  - BCR-ABL1 in chronic myeloid leukemia (CML) (Mitelman et al, Nature Review Cancer, 2007)
  - TMPRSS2-ERG2 in prostate cancer (urine test) (Thomlins et al, Science, 2005, Thomlins et al, Nature, 2007)
- 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

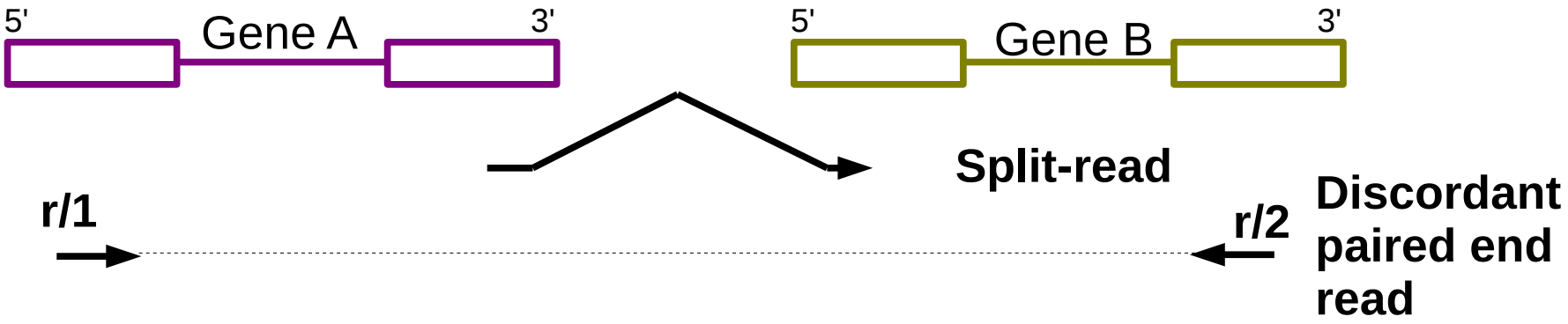
Computational identification of chimeras from RNA-seq

- 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

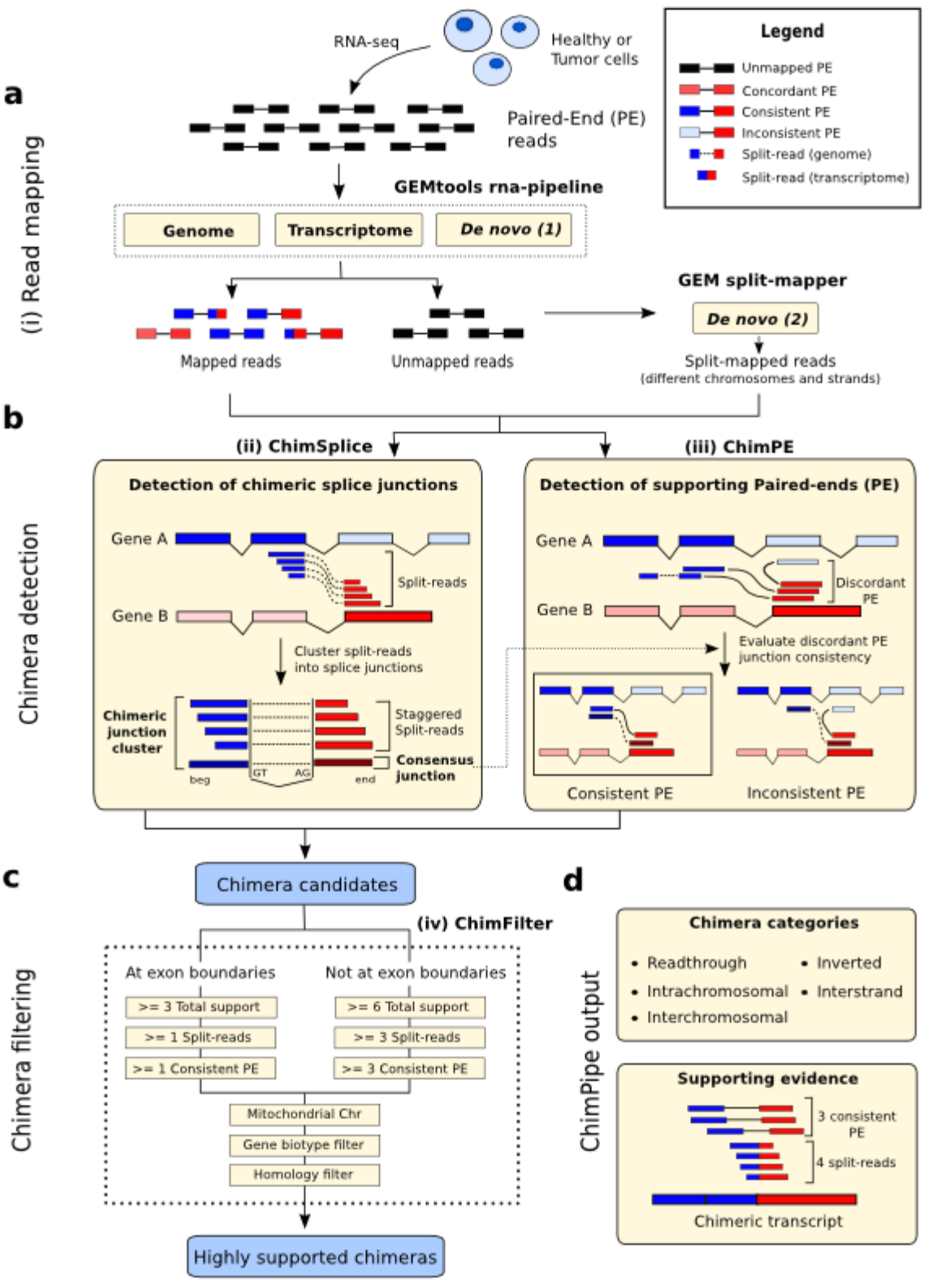
ChimPipe

- 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

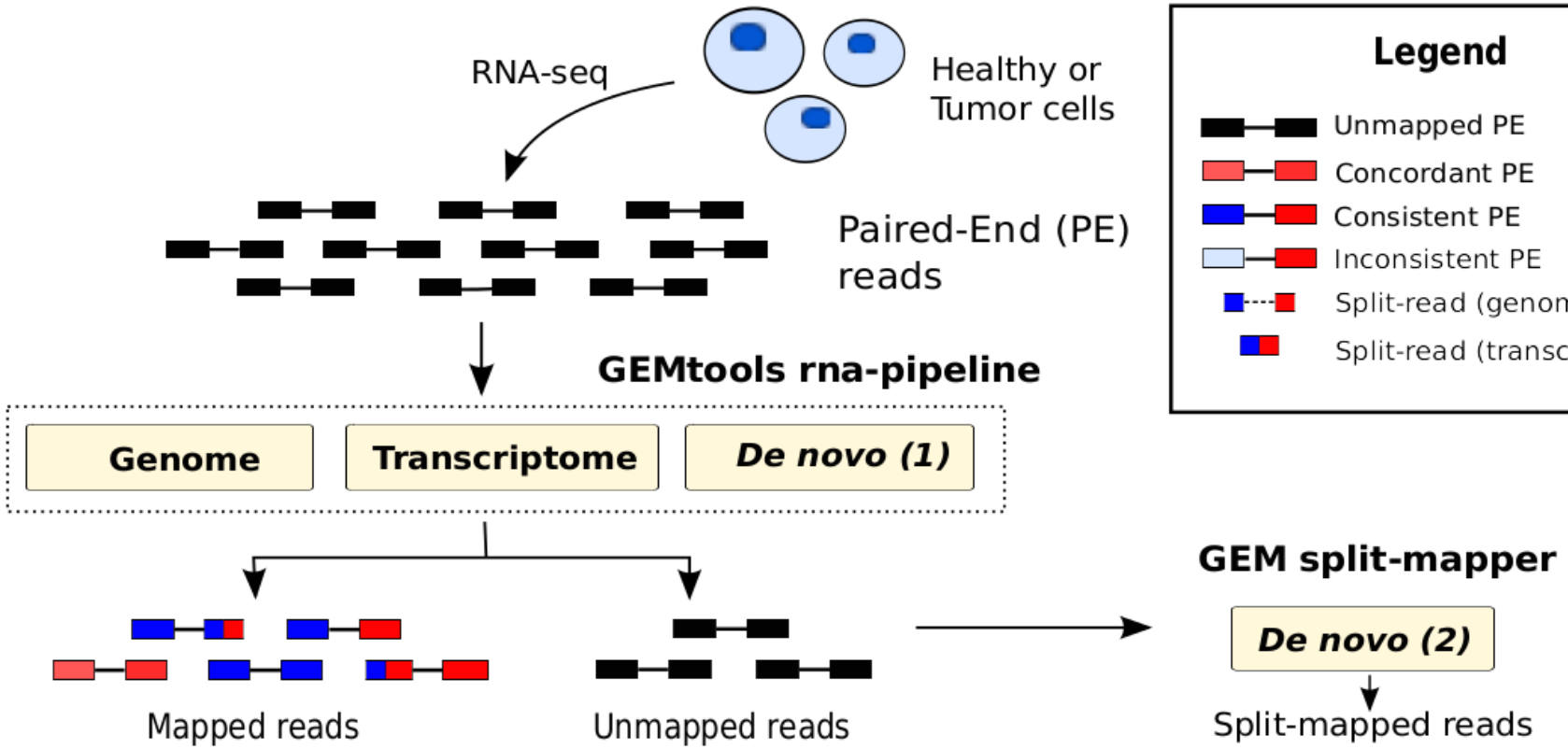
2. Chimera detection

3. Chimera filtering

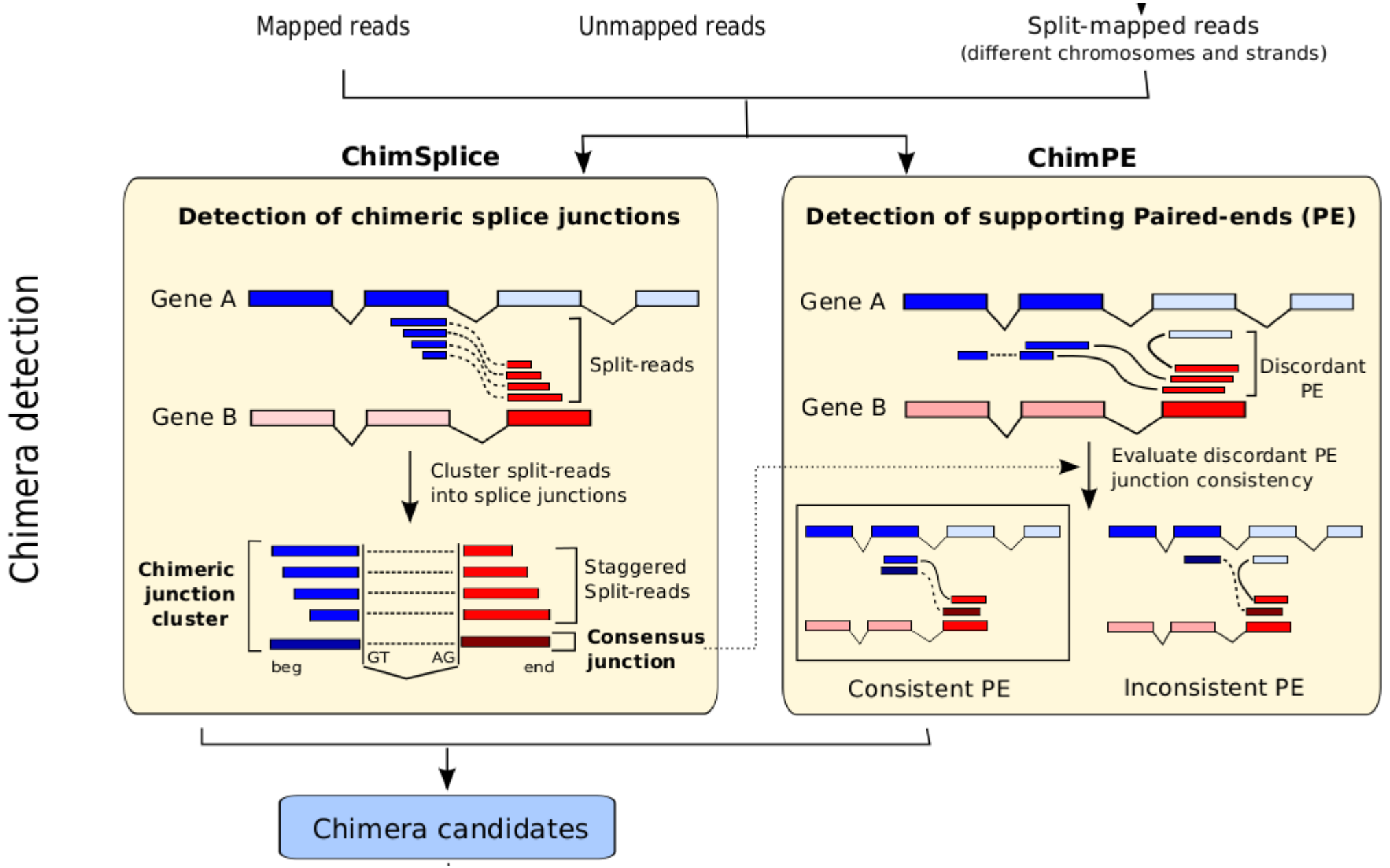


1. Read mapping

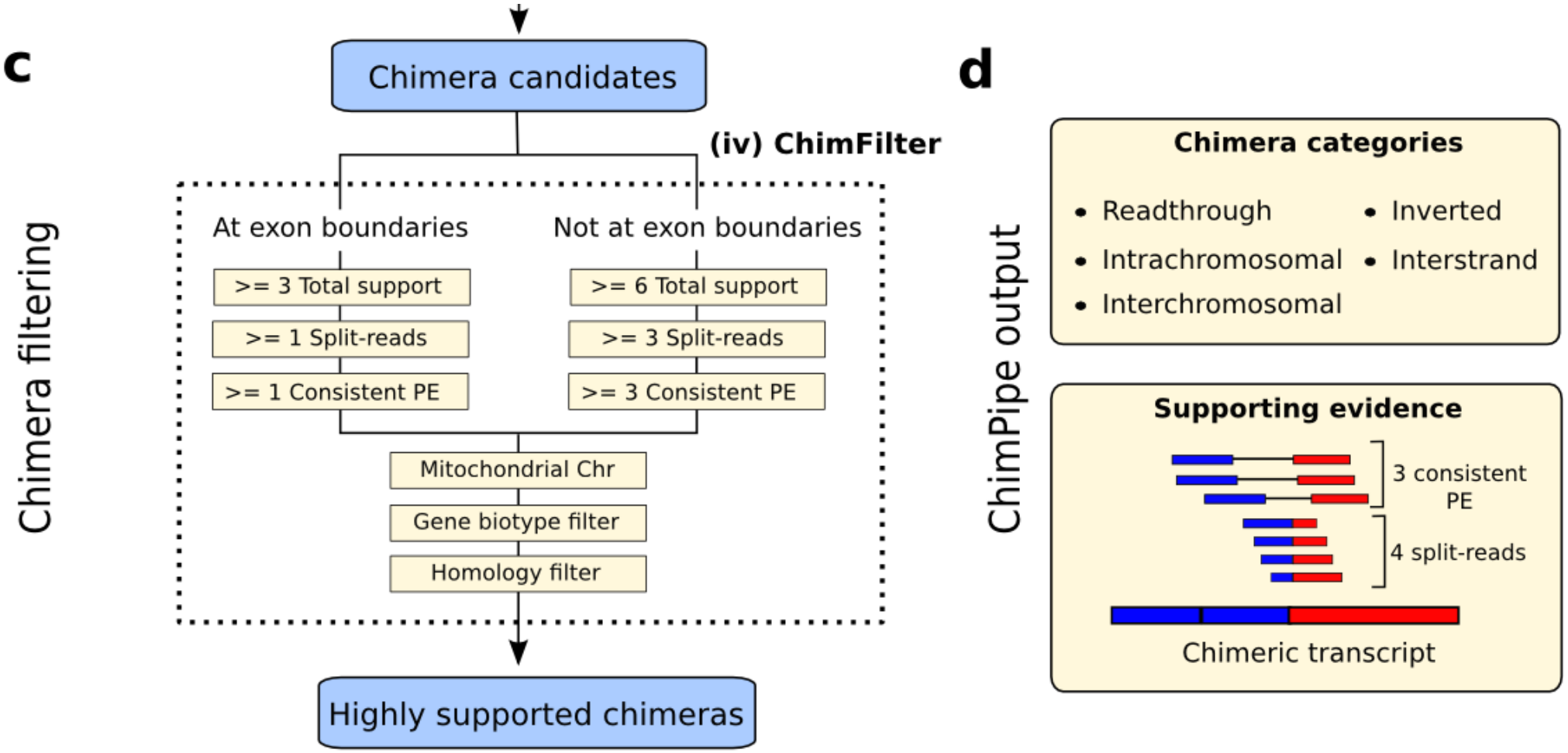
Read mapping



2. Chimera detection



3. Chimera filtering

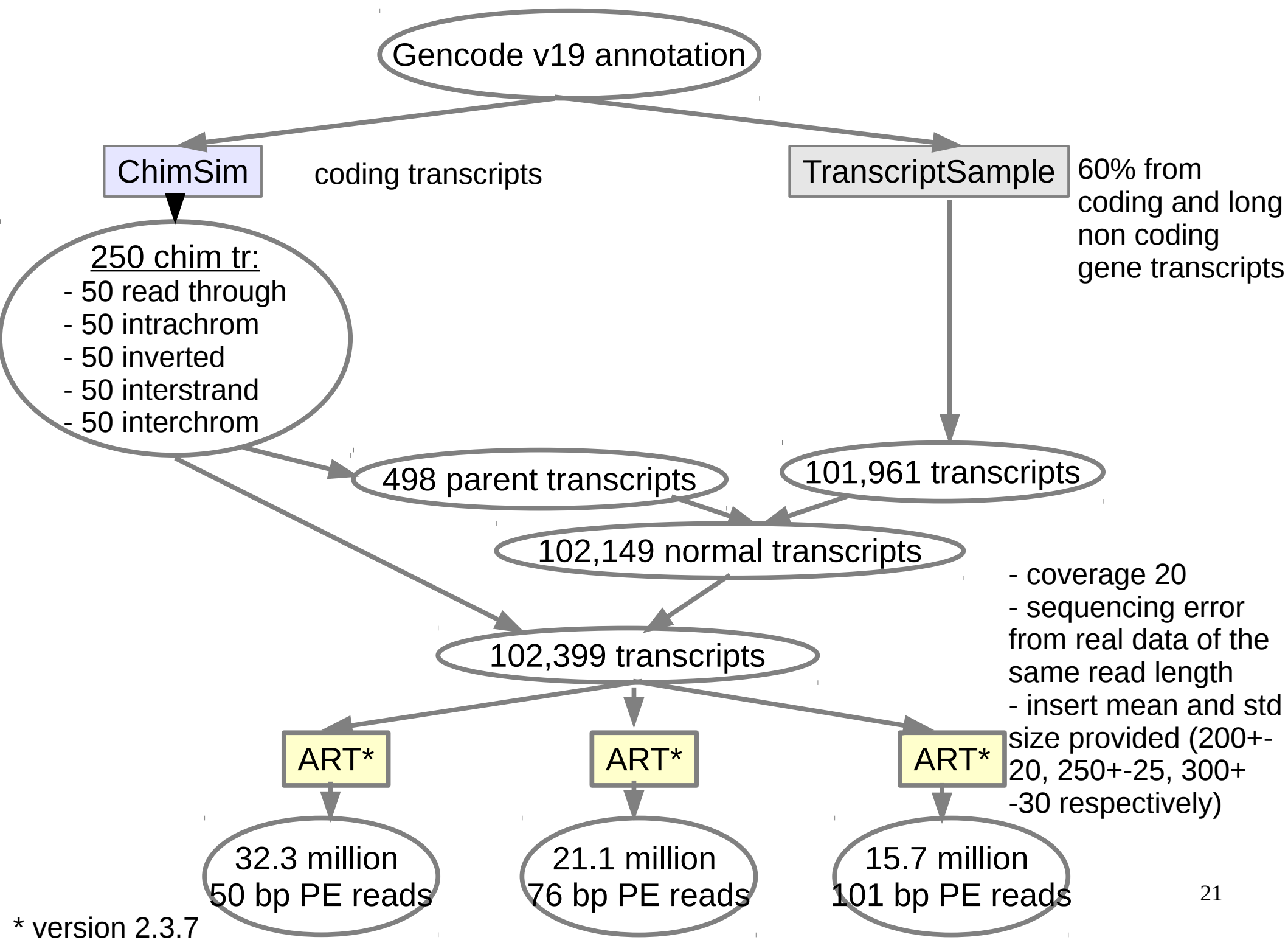


ChimPipe implementation

- GitHub:
  - https://github.com/Chimera-tools/ChimPipe
- Documentation:
  - https://chimpipe.readthedocs.org/en/latest/
- 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
coding transcripts
TranscriptSample
60% from coding and long non coding gene transcripts
250 chim tr:
- 50 read through
- 50 intrachrom
- 50 inverted
- 50 interstrand
- 50 interchrom
498 parent transcripts
101,961 transcripts
102,149 normal transcripts
102,399 transcripts
- coverage 20
- sequencing error from real data of the same read length
- insert mean and std size provided (200+-20, 250+-25, 300+-30 respectively)
ART*
ART*
ART*
32.3 million 50 bp PE reads
21.1 million 76 bp PE reads
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)

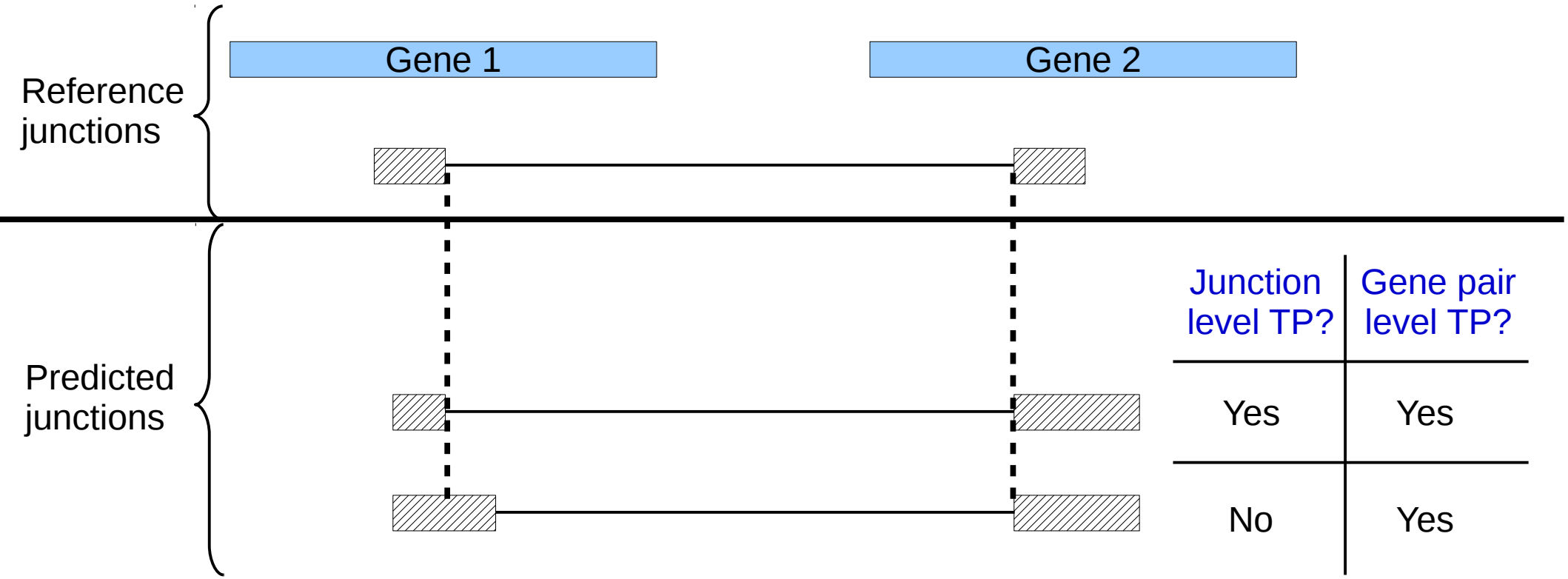
Cancer dataset	Cell line	Tumor type	Number of validated fusion genes	Number of validated fusion junctions	Reference paper(s)
Berger	K562	Leukemia	3	3	Berger et al, Genome Research, 2010
	501 Mel	Melanoma	4	5	Berger et al, Genome Research, 2010
	M000216		1	1	Berger et al, Genome Research, 2010
	M000921		2	3	Berger et al, Genome Research, 2010
	M010403		1	1	Berger et al, Genome Research, 2010
	M980409		1	1	Berger et al, Genome Research, 2010
	M990802		2	2	Berger et al, Genome Research, 2010
	All	All	**14**	**16**	Berger et al, Genome Research, 2010
Edgren	KPL-4	Breast cancer	3	3	Edgren et al, Genome Biology, 2011
	MCF-7		6	8	Edgren el al, Genome Biology, 2011; Kangaspeska et al, PLOSone, 2012
	BT-474		21	25	Edgren el al, Genome Biology, 2011; Kangaspeska et al, PLOSone, 2012
	SK-BR-3		10	10	Edgren et al, Genome Biology, 2011
	All		**40**	**46**	Edgren el al, Genome Biology, 2011; Kangaspeska et al, PLOSone, 2012

State of the art benchmarked programs

Program	Why was it chosen for the benchmark?	Underlying mapper?	Chimera detection approach	What are the false positive filters used?	Publication
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

$$Sn = \frac{TP}{TP + FN} \qquad Pr = \frac{TP}{TP + FP}$$

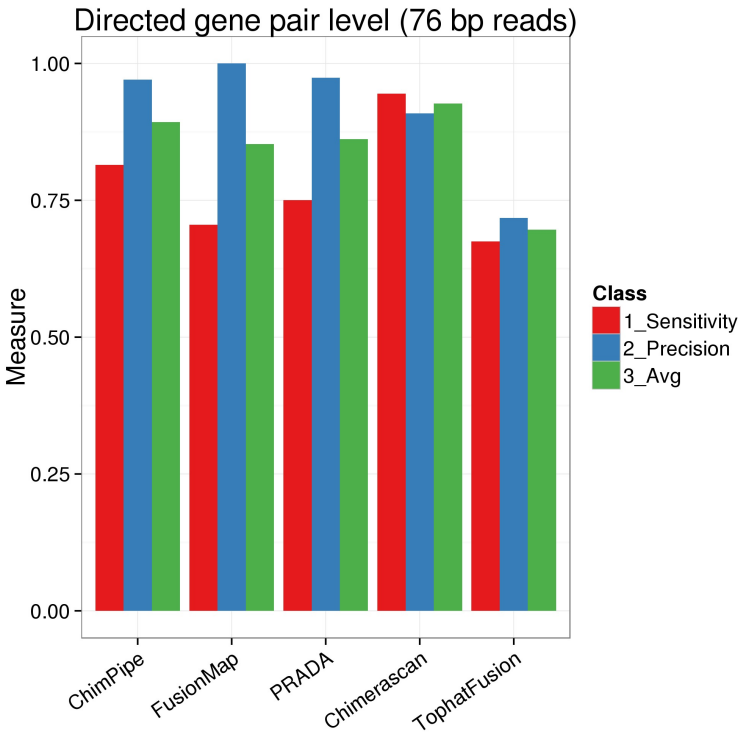


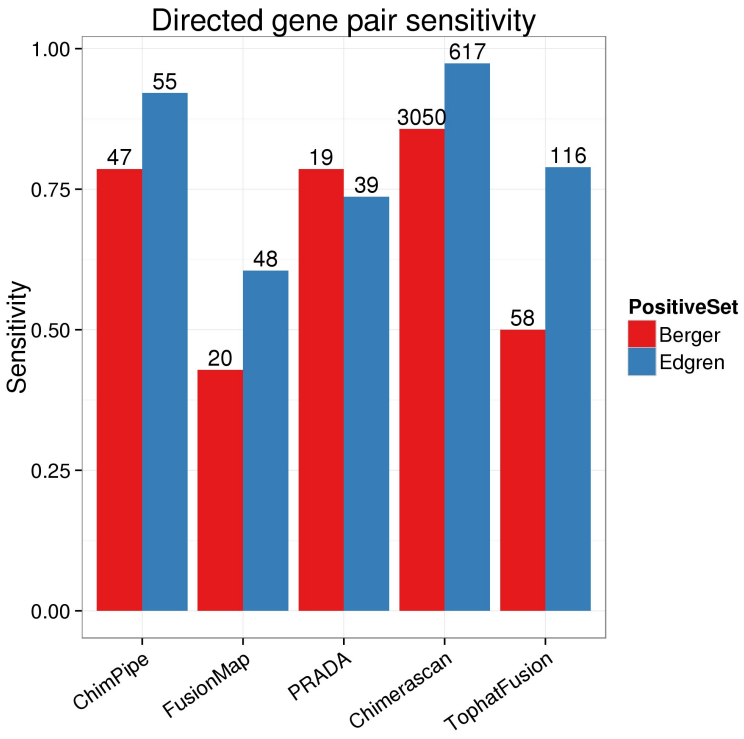
Junction level TP?	Gene pair level TP?
Yes	Yes
No	Yes

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

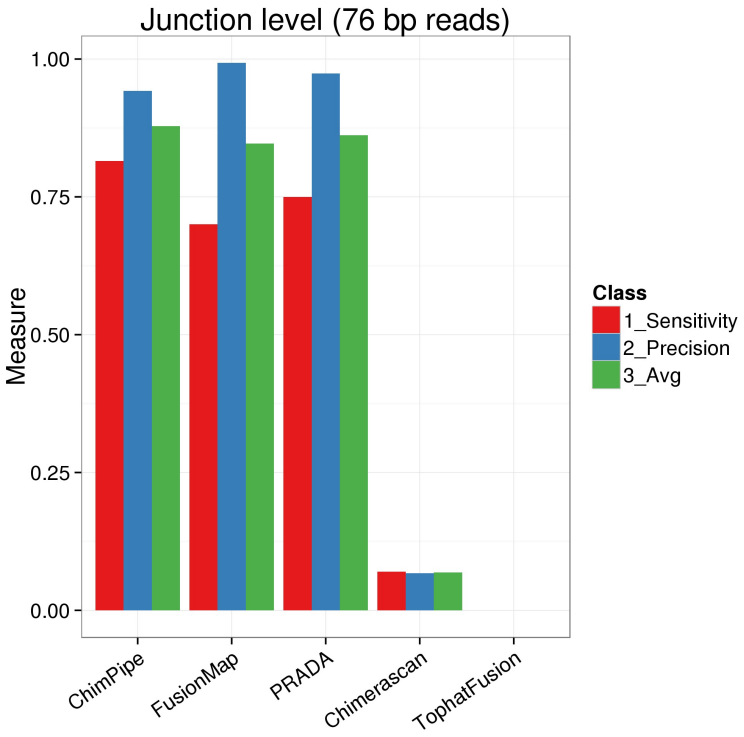
- ChimPipe is second after chimerascan which predicts many more cases on real data

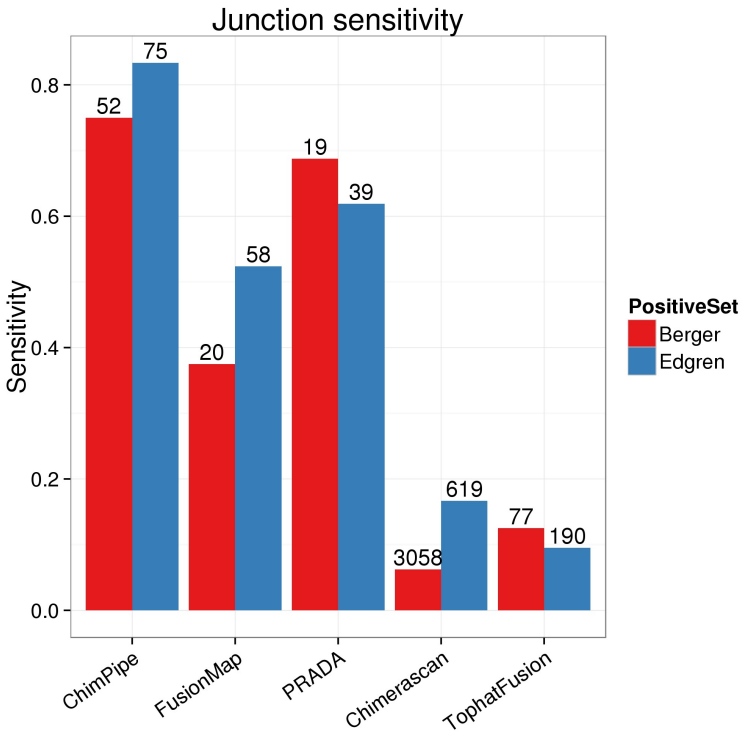




Exact junction level assessment

- ChimPipe is the best for both kinds of datasets





Resources needed on the simulated sets with 4 cpus

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

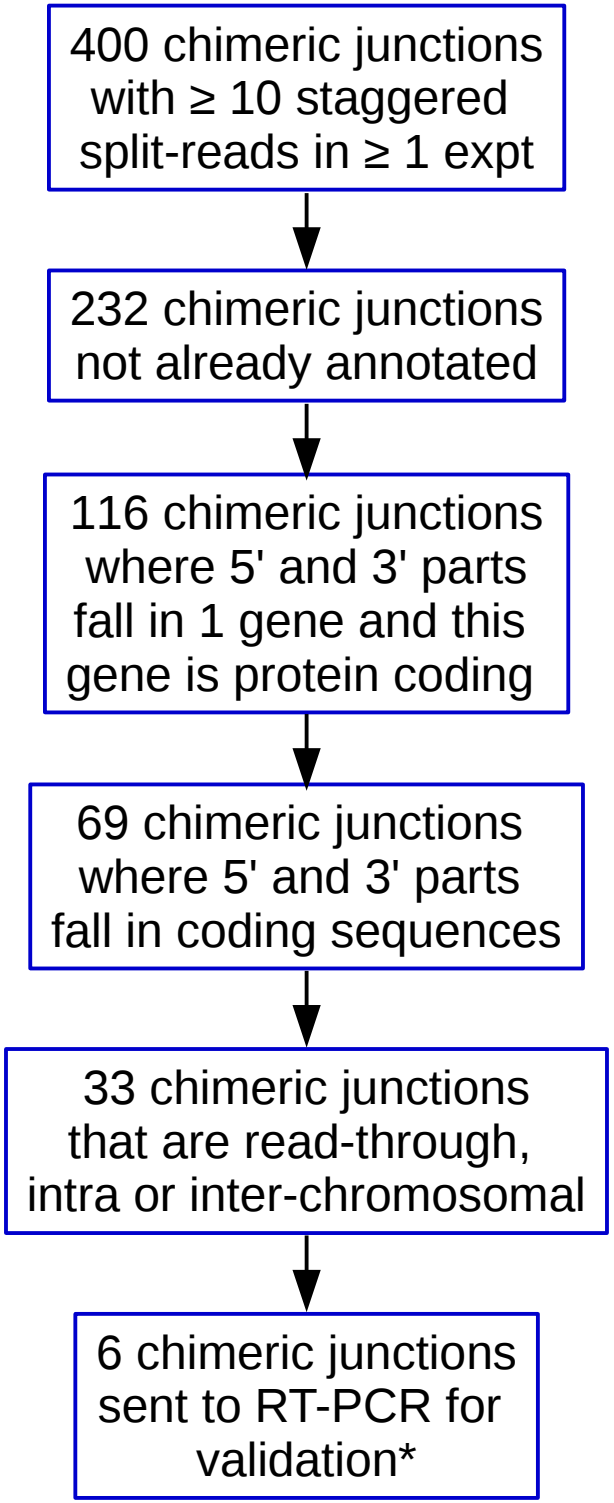
FusionMap is the tool that performs best overall after ChimPipe, however its behaviour depends on the read length with 76 bp reads less well handled than 50bp and 101bp reads

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)

RNA fraction	Cell line	Cell compartment	RNA fraction	Cell line	Cell compartment
LONGNONPOLYA	A549	CELL	LONGPOLYA	A549	CELL
LONGNONPOLYA	A549	CELL	LONGPOLYA	A549	CELL
LONGNONPOLYA	AG04450	CELL	LONGPOLYA	AG04450	CELL
LONGNONPOLYA	AG04450	CELL	LONGPOLYA	AG04450	CELL
LONGNONPOLYA	BJ	CELL	LONGPOLYA	GM12878	CELL
LONGNONPOLYA	BJ	CELL	LONGPOLYA	GM12878	CELL
LONGNONPOLYA	GM12878	CELL	LONGPOLYA	GM12878	CYTOSOL
LONGNONPOLYA	GM12878	CELL	LONGPOLYA	GM12878	CYTOSOL
LONGNONPOLYA	GM12878	CYTOSOL	LONGPOLYA	GM12878	NUCLEUS
LONGNONPOLYA	GM12878	CYTOSOL	LONGPOLYA	GM12878	NUCLEUS
LONGNONPOLYA	GM12878	NUCLEUS	LONGPOLYA	H1HESC	CELL
LONGNONPOLYA	GM12878	NUCLEUS	LONGPOLYA	H1HESC	CELL
LONGNONPOLYA	H1HESC	CELL	LONGPOLYA	H1HESC	CYTOSOL
LONGNONPOLYA	H1HESC	CELL	LONGPOLYA	H1HESC	NUCLEUS
LONGNONPOLYA	H1HESC	CYTOSOL	LONGPOLYA	HELAS3	CELL
LONGNONPOLYA	H1HESC	NUCLEUS	LONGPOLYA	HELAS3	CELL
LONGNONPOLYA	HELAS3	CELL	LONGPOLYA	HELAS3	CYTOSOL
LONGNONPOLYA	HELAS3	CELL	LONGPOLYA	HELAS3	CYTOSOL
LONGNONPOLYA	HELAS3	CYTOSOL	LONGPOLYA	HELAS3	NUCLEUS
LONGNONPOLYA	HELAS3	CYTOSOL	LONGPOLYA	HELAS3	NUCLEUS
LONGNONPOLYA	HELAS3	NUCLEUS	LONGPOLYA	HEPG2	CELL
LONGNONPOLYA	HELAS3	NUCLEUS	LONGPOLYA	HEPG2	CELL
LONGNONPOLYA	HEPG2	CELL	LONGPOLYA	HEPG2	CYTOSOL
LONGNONPOLYA	HEPG2	CELL	LONGPOLYA	HEPG2	CYTOSOL
LONGNONPOLYA	HEPG2	CYTOSOL	LONGPOLYA	HEPG2	NUCLEUS
LONGNONPOLYA	HEPG2	CYTOSOL	LONGPOLYA	HEPG2	NUCLEUS
LONGNONPOLYA	HEPG2	NUCLEUS	LONGPOLYA	HSMM	CELL
LONGNONPOLYA	HEPG2	NUCLEUS	LONGPOLYA	HSMM	CELL
LONGNONPOLYA	HMEC	CELL	LONGPOLYA	HUVEC	CELL
LONGNONPOLYA	HSMM	CELL	LONGPOLYA	HUVEC	CELL
LONGNONPOLYA	HSMM	CELL	LONGPOLYA	HUVEC	CYTOSOL
LONGNONPOLYA	HUVEC	CELL	LONGPOLYA	HUVEC	CYTOSOL
LONGNONPOLYA	HUVEC	CELL	LONGPOLYA	HUVEC	NUCLEUS
LONGNONPOLYA	HUVEC	CYTOSOL	LONGPOLYA	HUVEC	NUCLEUS
LONGNONPOLYA	HUVEC	NUCLEUS	LONGPOLYA	K562	CELL
LONGNONPOLYA	HUVEC	NUCLEUS	LONGPOLYA	K562	CELL
LONGNONPOLYA	K562	CELL	LONGPOLYA	K562	CYTOSOL
LONGNONPOLYA	K562	CYTOSOL	LONGPOLYA	K562	CYTOSOL
LONGNONPOLYA	K562	NUCLEUS	LONGPOLYA	K562	NUCLEUS
LONGNONPOLYA	K562	NUCLEUS	LONGPOLYA	K562	NUCLEUS
LONGNONPOLYA	MCF7	CELL	LONGPOLYA	MCF7	CELL
LONGNONPOLYA	MCF7	CELL	LONGPOLYA	MCF7	CELL
LONGNONPOLYA	NHEK	CELL	LONGPOLYA	NHEK	CELL
LONGNONPOLYA	NHEK	CELL	LONGPOLYA	NHEK	CELL
LONGNONPOLYA	NHEK	CYTOSOL	LONGPOLYA	NHEK	NUCLEUS
LONGNONPOLYA	NHEK	CYTOSOL	LONGPOLYA	NHEK_BROAD	CELL
LONGNONPOLYA	NHEK	NUCLEUS	LONGPOLYA	NHLF	CELL
LONGNONPOLYA	NHEK	NUCLEUS	LONGPOLYA	NHLF	CELL
LONGNONPOLYA	NHEK_BROAD	CELL	LONGPOLYA	SKNSHRA	CELL
LONGNONPOLYA	NHLF	CELL	TOTAL	K562	CHROMATIN
LONGNONPOLYA	NHLF	CELL	TOTAL	K562	CHROMATIN
LONGNONPOLYA	SKNSHRA	CELL	TOTAL	K562	NUCLEOLUS

In 2011, using Gencode v7 annotation and an ancestor of ChimPipe



* 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

Chimeric junction number	Chimeric junction identifier	start	end	cis?	Gene1 (biological name)	Gene2 (biological name)	max local % nt similarity	Average number of staggered reads supporting the junction per experiment in a cell line																lines in which is it expressed	List of cell lines	OnlyCancer OnlyNotCancer | Mix
								K562:16	GM12878:12	HELAS3:12	HEPG2:12	HUVEC:11	NHEK:9	H1HESC:8	A549:4	AG04450:4	HSMM:4	MCF7:4	NHLF:4	SKNSHRA:3	BJ:2	NHEK_BROAD_2	HMEC:1			
1	chr11_85685855_-:chr11_85468668_-	85,685,796	85,468,727	yes	PICALM	SYTL2	0	0	0	0	0	0	0	0	0	0	0	15.75	0	0	0	0	0	1	MCF7,	OnlyCancer
2	chr12_52911947_-:chr12_52885306_-	52,911,906	52,885,355	yes	KRT5	KRT6A	100	0	0	0	0	0	5.66667	0	0	0	0	0	0	0	0	6.5	0	2	NHEK,NHEK_BROAD,	OnlyNotCancer
3	chr12_56113007_+:chr12_56115473_+	56,112,949	56,115,531	yes	BLOC1S1	RDH5	0	0.6875	2.41667	3.58333	0.5	0.54546	0.22222	2.375	0	0	0.5	0.5	1	0.66667	0	0	0	11	G2,HUVEC,NHEK,H1HESC,HSMM,MCF7,NHLF,SKNSHRA	Mix
4	chr14_39722375_+:chr14_39746138_+	39,722,315	39,746,195	yes	MIA2	CTAGE5	0	0	0	0.25	4.58333	0	0	0	4	0	0	0	0	0	0	0	0	3	HELAS3,HEPG2,A549,	OnlyCancer
5	chr16_19603196_+:chr16_19867809_+	19,603,143	19,867,866	yes	C16orf62	IQCK	0	0	0	0	0	0	0	0	0	0	0	15	0	0	0	0	0	1	MCF7,	OnlyCancer
6	chr17_72200329_+:chr17_72218624_+	72,200,271	72,218,684	yes	RPL38	TTYH2	0	0.125	1.91667	0.16667	0.5	0.54546	0.11111	0.25	0	0	0.25	0.5	0.25	0.33333	0	0	0	11	G2,HUVEC,NHEK,H1HESC,HSMM,MCF7,NHLF,SKNSHRA	Mix
7	chr17_7474797_+:chr17_7477578_+	7,474,738	7,477,627	yes	SENP3	EIF4A1	0	1.375	2.33333	2.66667	1.66667	2.45455	1.11111	3.75	3.75	2	4	1.5	3.25	2.33333	1.5	2.5	1	16	G2,HUVEC,NHEK,H1HESC,A549,AG04450,HSMM,MCF7,NHLF,SKNSHRA,BJ,NHEK_BR	Mix
8	chr17_76160445_+:chr17_76166898_+	76,160,386	76,166,956	yes	C17orf99	SYNGR2	0	4.125	0.33333	0.08333	0.75	0	0.33333	0	0	0	0	0	0	0	0	0	0	5	K562,GM12878,HELAS3,HEPG2,NHEK,	Mix
9	chr19_33878987_-:chr19_33450806_-	33,878,930	33,450,868	yes	PEPD	CCDC123	0	0	0	4.75	0.75	0.27273	0.22222	0	0	0	0	0	0	0	0	0	0	4	HELAS3,HEPG2,HUVEC,NHEK,	Mix
10	chr19_34957919_+:chr19_34981281_+	34,957,865	34,981,333	yes	UBA2	WTIP	0	1.4375	0.91667	2.75	0.75	0.54546	0.66667	0.75	4.75	1.25	0.5	0.75	0.75	1	0	0.5	0	14	G2,HUVEC,NHEK,H1HESC,A549,AG04450,HSMM,MCF7,NHLF,SKNSHRA,BJ,NHEK_BR	Mix
11	chr20_35207369_+:chr20_35236118_+	35,207,311	35,236,175	yes	TGIF2	C20orf24	0	0.25	0.91667	0.33333	1.16667	0.45455	0.55556	1.25	0.25	1	1.5	4	1	1.33333	0	0.5	0	14	G2,HUVEC,NHEK,H1HESC,A549,AG04450,HSMM,MCF7,NHLF,SKNSHRA,NHEK_BROA	Mix
12	chr2_71148415_+:chr2_71170788_+	71,148,356	71,170,835	yes	VAX2	ATP6V1B1	0	0	0	0	2.5	0.63636	0.11111	0	0.75	0	0	0.25	1	1.33333	0.5	0	0	8	HEPG2,HUVEC,NHEK,A549,MCF7,NHLF,SKNSHRA,BJ,	Mix
13	chr2_85806290_+:chr2_85818848_+	85,806,232	85,818,901	yes	VAMP8	VAMP5	0	0.125	0.41667	0.25	0.25	0.45455	1.88889	0.25	0.25	0	0.75	0.75	0.25	0	0	1	0	12	G2,HUVEC,NHEK,H1HESC,A549,HSMM,MCF7,NHLF,NHE	Mix
14	chr5_60053472_-:chr5_59934577_-	60,053,412	59,934,635	yes	ELOVL7	DEPDC1B	0	0	0	0	0	0	0	0	0	0	0	13.75	0	0	0	0	0	1	MCF7,	OnlyCancer
15	chr6_147830523_+:chr6_148711270_+	147,830,464	148,711,328	yes	SAMD5	SASH1	0	0	0	0	0	0.18182	0	0	0.5	0.5	5.25	0	1	0	0	0.5	0	6	HUVEC,A549,AG04450,HSMM,NHLF,NHEK_BROAD,	Mix
16	chr6_26020927_+:chr6_26045849_+	26,020,883	26,045,903	yes	HIST1H3A	HIST1H3C	100	0	0.16667	1	0.58333	0	0	0	0	0	0	0	0	0	0	0	0	3	GM12878,HELAS3,HEPG2,	Mix
17	chr6_26020963_+:chr6_26045885_+	26,020,924	26,045,941	yes	HIST1H3A	HIST1H3C	100	0	0	1.16667	0.41667	0.36364	0	0.25	0	0	0	0	0	0	0.5	0	0	5	HELAS3,HEPG2,HUVEC,H1HESC,BJ,	Mix
18	chr6_26021013_+:chr6_26045933_+	26,020,958	26,045,981	yes	HIST1H3A	HIST1H3C	100	0	0	0.91667	0.25	0	0	0	0	0	0	0	0	0	0	0	0	2	HELAS3,HEPG2,	OnlyCancer
19	chr6_26045884_+:chr6_27778098_+	26,045,830	27,778,143	yes	HIST1H3C	HIST1H3H	100	0	0.33333	1.41667	0.66667	0	0	0	0	0	0	0	0	0	0	0	0	3	GM12878,HELAS3,HEPG2,	Mix
20	chr6_26124559_+:chr6_26217302_+	26,124,501	26,217,355	yes	HIST1H2AC	HIST1H2AE	100	0	0	0.83333	0	0	0	0	0	0	0	0	0	0	0	0	0	1	HELAS3,	OnlyCancer
21	chr6_26197151_-:chr6_26031962_-	26,197,102	26,032,021	yes	HIST1H3D	HIST1H3B	87.85	0.0625	0.16667	2	0.91667	0.09091	0	0.375	0	0	0	0	0	0	0	0	0	6	K562,GM12878,HELAS3,HEPG2,HUVEC,H1HESC,	Mix
22	chr6_26216645_-:chr6_26123902_-	26,216,590	26,123,962	yes	HIST1H2BG	HIST1H2BC	95.45	0.4375	1.5	0.25	0.33333	0	0	1	0	0	0	0	0	0	0	0	0	5	K562,GM12878,HELAS3,HEPG2,H1HESC,	Mix
23	chr6_26250506_-:chr6_26031962_-	26,250,464	26,032,022	yes	HIST1H3F	HIST1H3B	100	0.125	0.58333	1.75	0.41667	0.09091	0	0.125	0	0	0	0.5	0.25	0	0	0	0	8	K562,GM12878,HELAS3,HEPG2,HUVEC,H1HESC,MCF7,NHLF	Mix
24	chr6_31833561_-:chr6_31619433_-	31,833,501	31,619,493	yes	SLC44A4	BAG6	0	19	0	0	0	0.09091	0	0	0	0	0	0	0	0	0	0	0	2	K562,HUVEC,	Mix
25	chr6_32632691_-:chr6_32489799_-	32,632,661	32,489,852	yes	HLA-DQB1	HLA-DRB5	100	0	2.75	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	GM12878,	OnlyNotCancer
26	chr7_111409733_-:chr7_111127294_-	111,409,675	111,127,354	yes	DOCK4	IMMP2L	0	5.875	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	K562,	OnlyCancer
27	chr9_139701124_+:chr9_139717977_+	139,701,065	139,718,036	yes	KIAA1984	C9orf86	100	0.0625	0.08333	0	2.08333	0	0	0	0	0	0	0	0	0	0	0	0	3	K562,GM12878,HEPG2,	Mix
28	chr6_26124529_+:chr1_149858594_+	26,124,490	149,858,650	No	HIST1H2AC	HIST2H2AC	95.24	0.0625	0	1	0.33333	0	0	0	0	0	0	0	0	0	0	0	0	3	K562,HELAS3,HEPG2,	OnlyCancer
29	chr22_23632600_+:chr9_133729451_+	23,632,539	133,729,510	No	BCR	ABL1	0	16.4375	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	K562,	OnlyCancer
30	chr20_49411710_+:chr17_59445688_+	49,411,650	59,445,743	No	BCAS4	BCAS3	0	0	0	0	0	0	0	0	0	0	0	19.75	0	0	0	0	0	1	MCF7,	OnlyCancer
31	chr10_104019873_+:chr20_14665489_+	104,019,813	14,665,547	No	GBF1	MACROD2	0	0	0	0	0	0	0	0	0	0	0	8.25	0	0	0	0	0	1	MCF7,	OnlyCancer
32	chr17_56801461_+:chr3_63965591_+	56,801,402	63,965,648	No	RAD51C	ATXN7	0	0	0	0	0	0	0	0	0	0	0	5.75	0	0	0	0	0	1	MCF7,	OnlyCancer
detected by	chr17_79477859_-:chr7_5567635_-	79,477,815	5,567,698	No	ACTG1	ACTB	96.23	0.0625	1.41667	0.16667	0.16667	3	1.11111	2.25	1.5	4	6.5	7.5	14	0.66667	1	3.5	2	16	G2,HUVEC,NHEK,H1HESC,A549,AG04450,HSMM,MCF7,NHLF,SKNSHRA,BJ,NHEK_BR	Mix

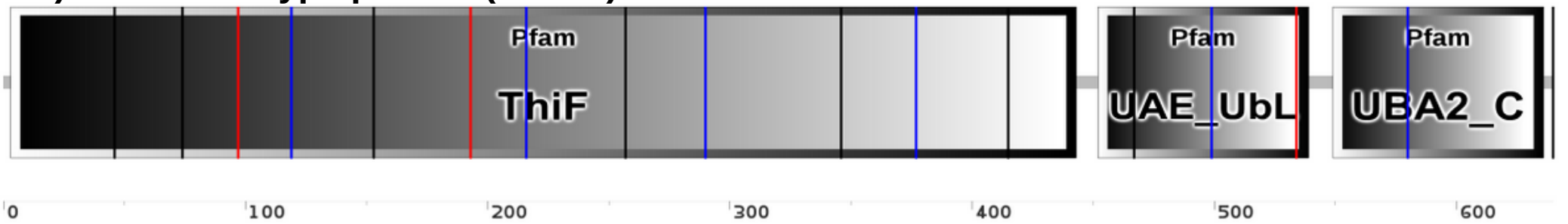
Out of 6 junctions attempted to be validated by RT-PCR, 3 were successfully validated

RT-PCR validated junctions

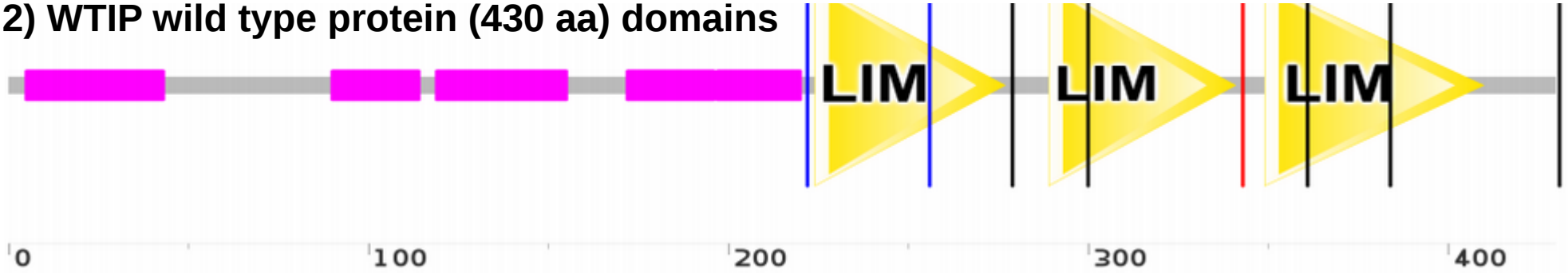
- The 3 RT-PCR validated junctions were further cloned and sequenced (Sanger sequencing)
- Only 1 case (UBA-WTIP) maintained the frame of the 2 parent genes and was therefore completely sequenced
  - 3 novel transcript structures, of which 1 was further analyzed and shows the 2 first (ThiF and UAE_Ubl) 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

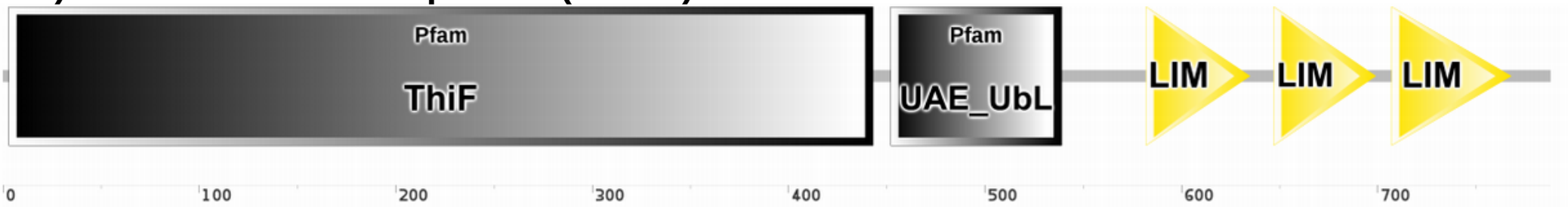
1) UBA2 wild type protein (640 aa) domains



2) WTIP wild type protein (430 aa) 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

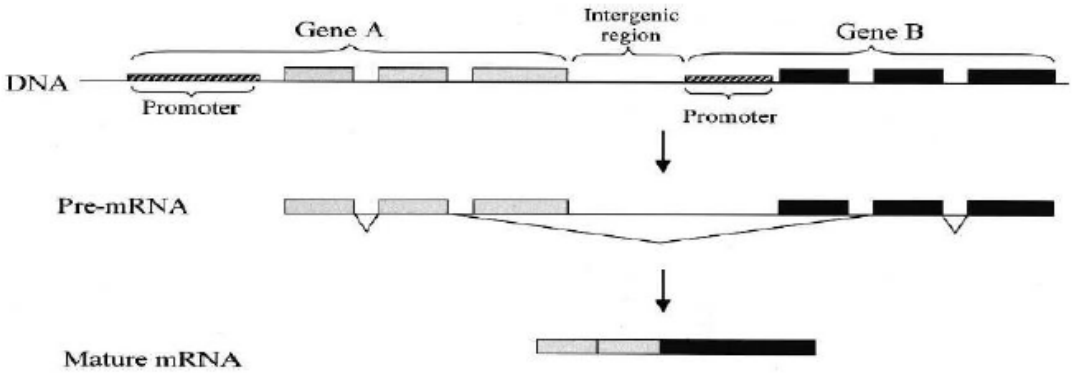
- Biologically:
  - Investigate the role of the novel chimeric protein found
  - Apply chimpipe to many animal genomes and individuals in order to study chimera evolution and connect some of them to individuals' phenotypes
  - Compare to HiC data to have a hint on mechanisms
  - Use RNA FISH to confirm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

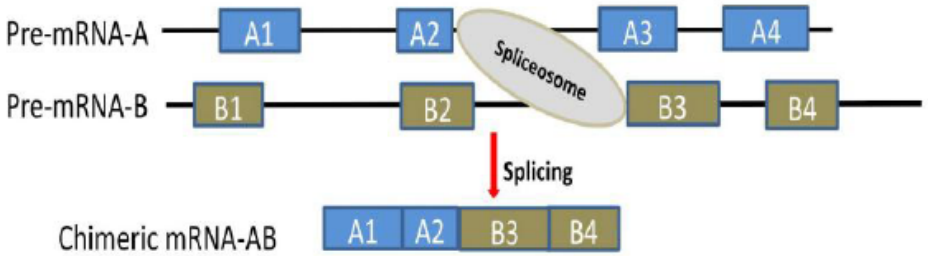
Possible mechanisms explaining the formation of chimeras

Read-through transcription



Akiva P. *et al.* (2006) Genome Res.

Trans-splicing



Zhou J. *et al.* (2012) BMB Rep.

Genomic rearrangements

Mitelman, Johansson, Mertens (2007)
Nature Review Cancer

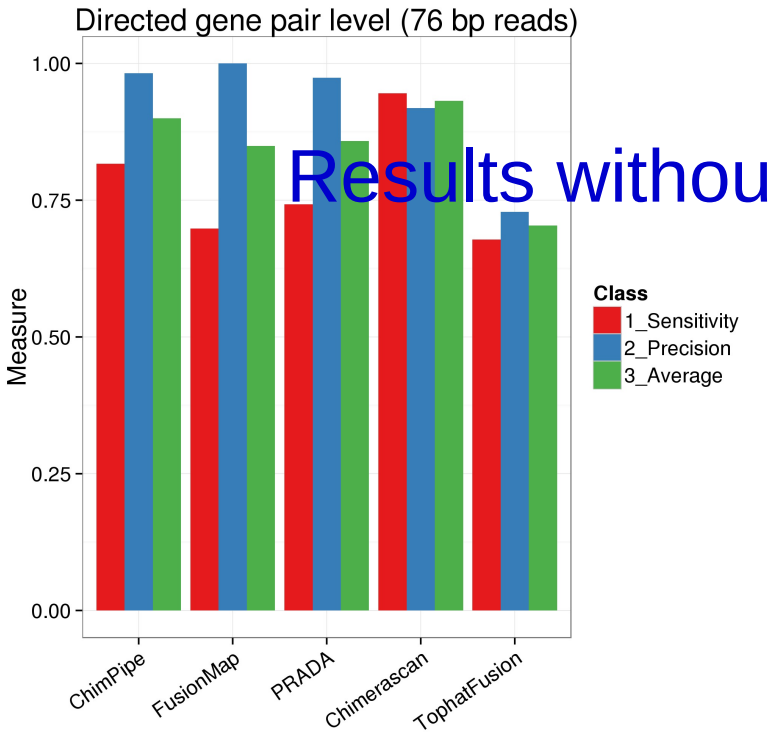
SHS transcriptional slippage

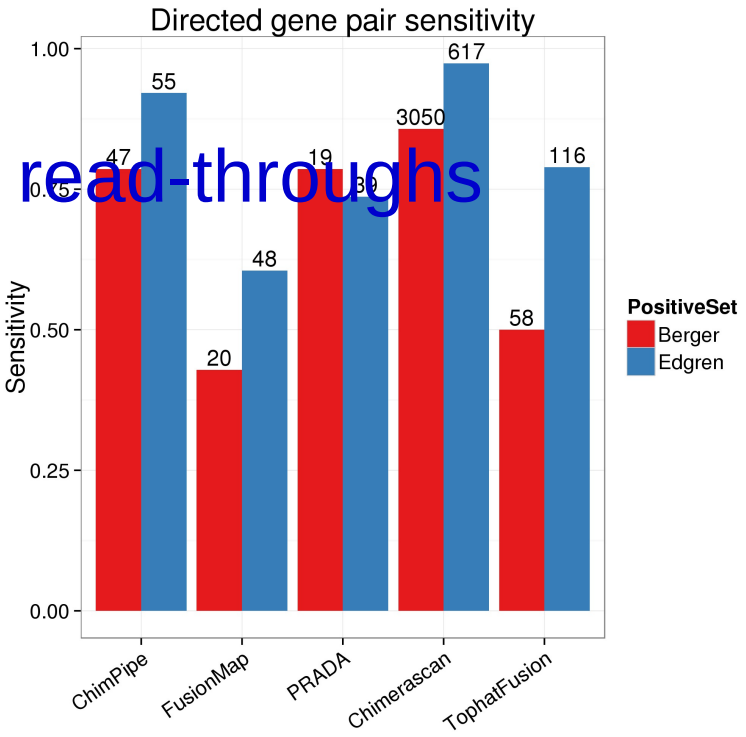
Lin X. *et al.* (2008) J Mol Evol

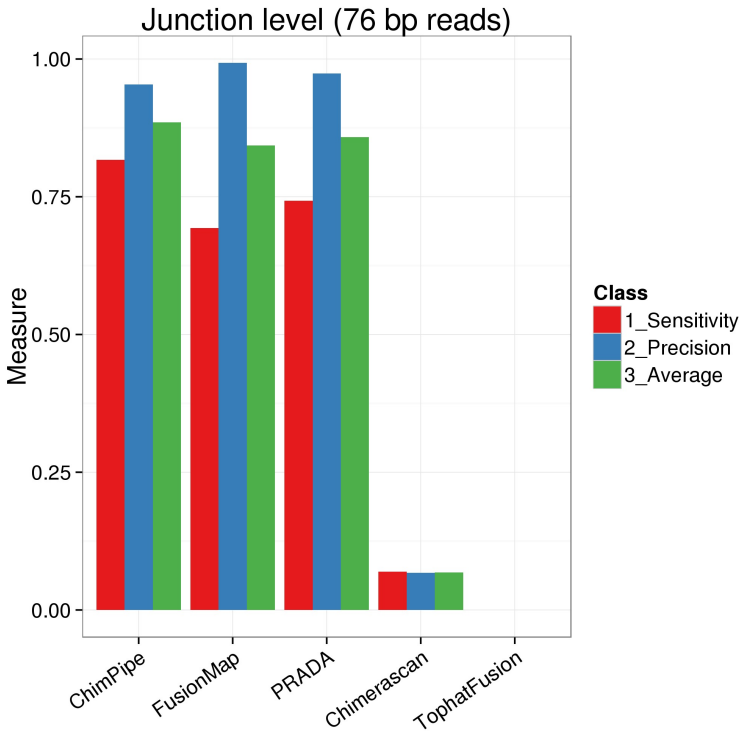
Benchmark data: simulated data

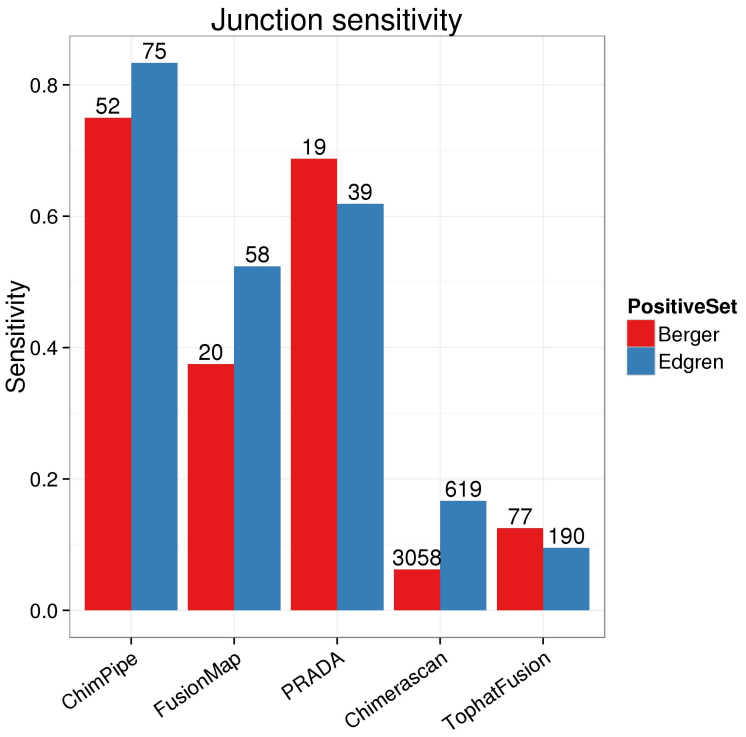
- Chimeric transcripts: 250 chimeras were generarted from Gencode v19 protein coding transcripts: 50 read-throughs, 50 intra-chromosomal, 50 inverted, 50 inter-strand, 50 inter-chromosomal
- Normal transcripts: 60% of transcripts were sampled from the 169,935 Gencode v19 protein-coding and lncRNA genes, 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

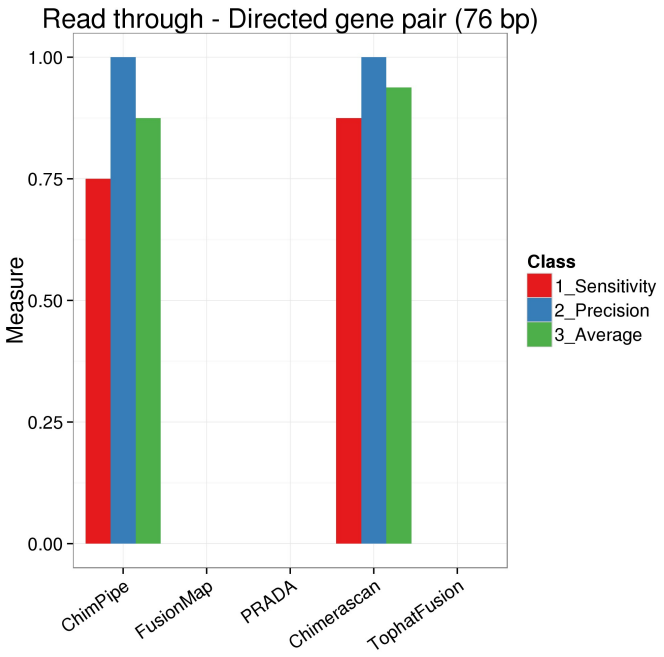


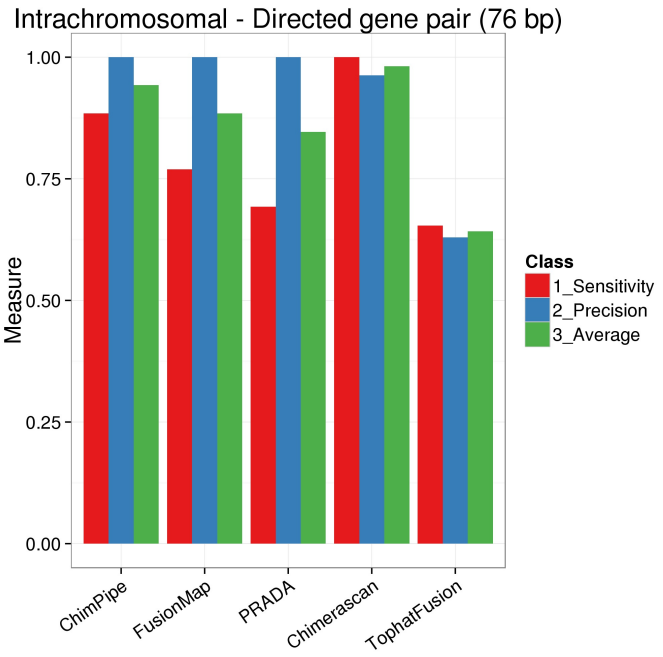


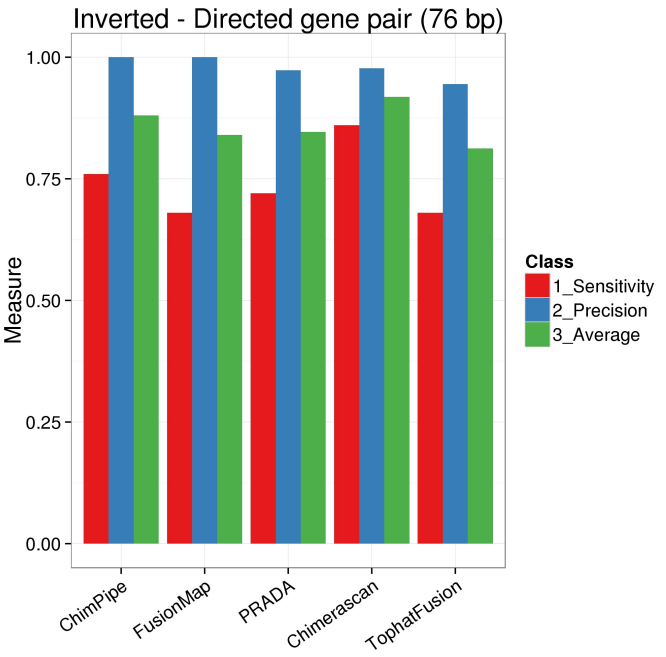


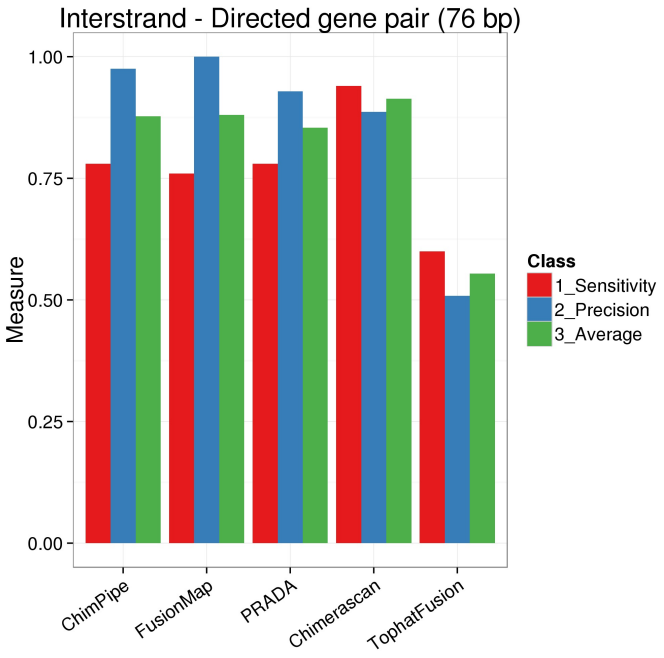


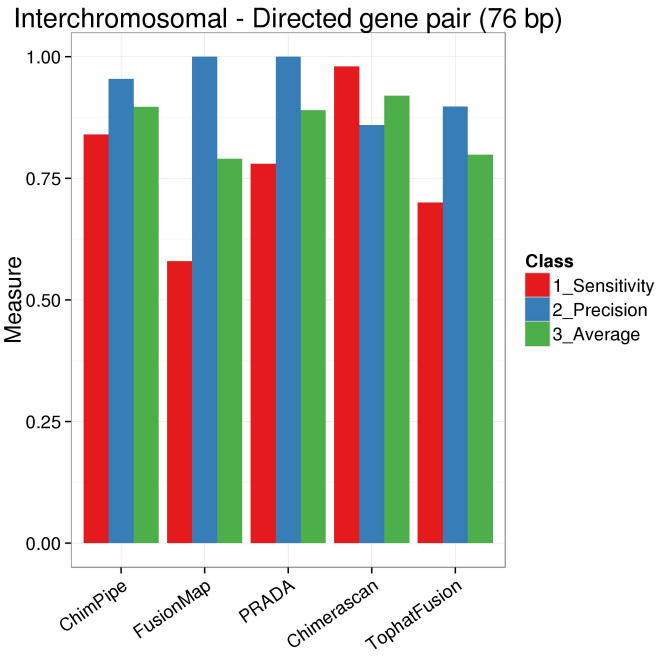
Result for each class of chimeras (gene pair level)

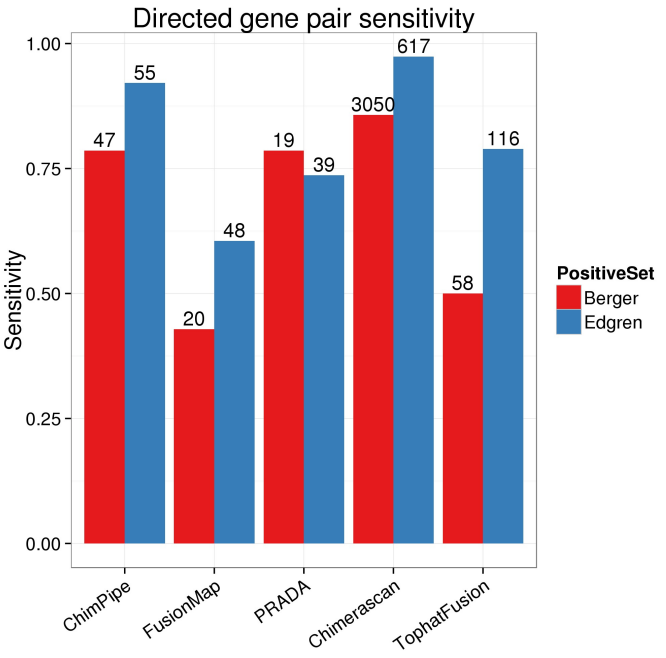




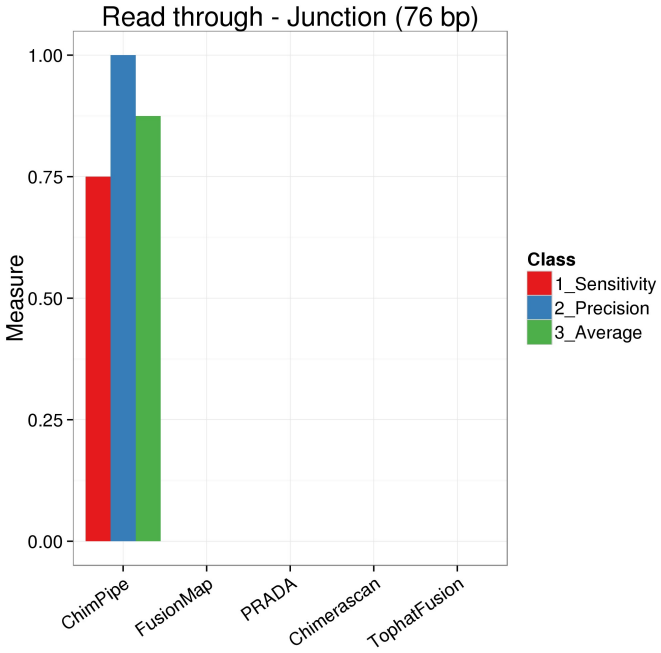


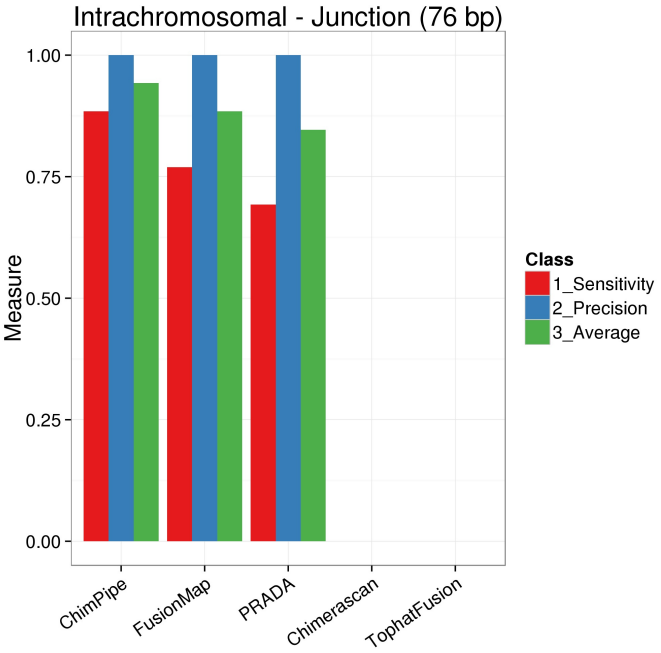


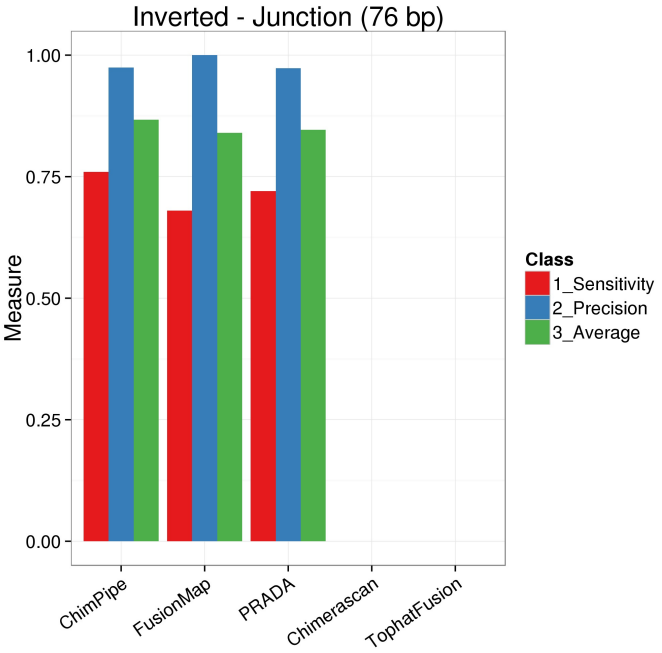


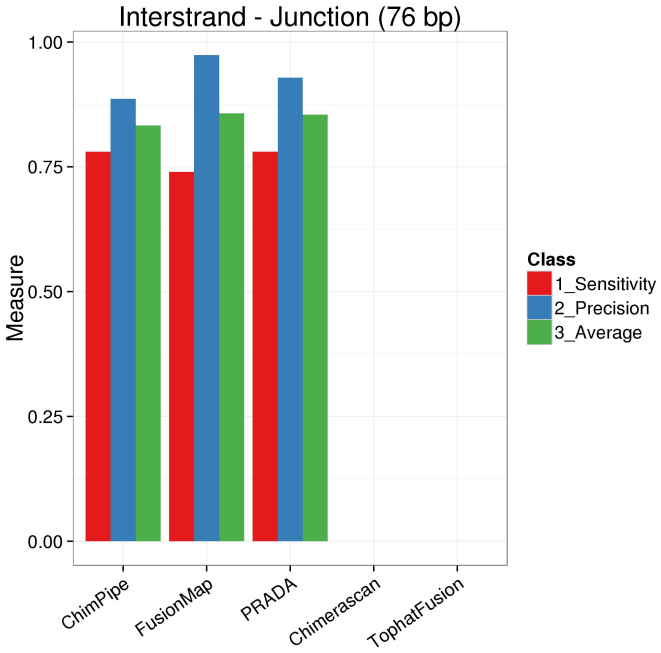


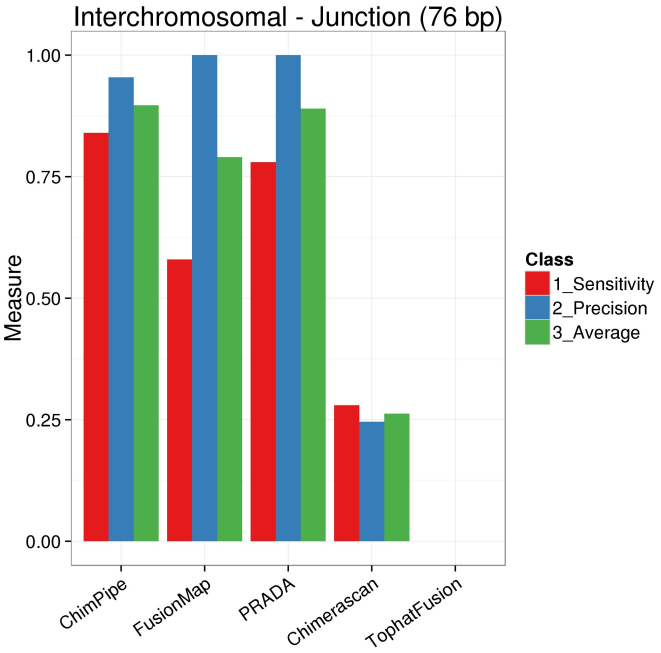
Result for each class of chimeras (junction level)

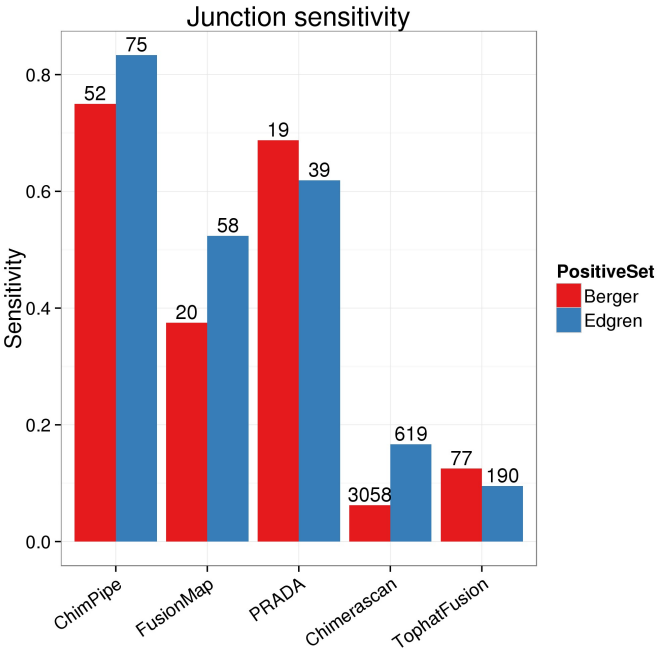


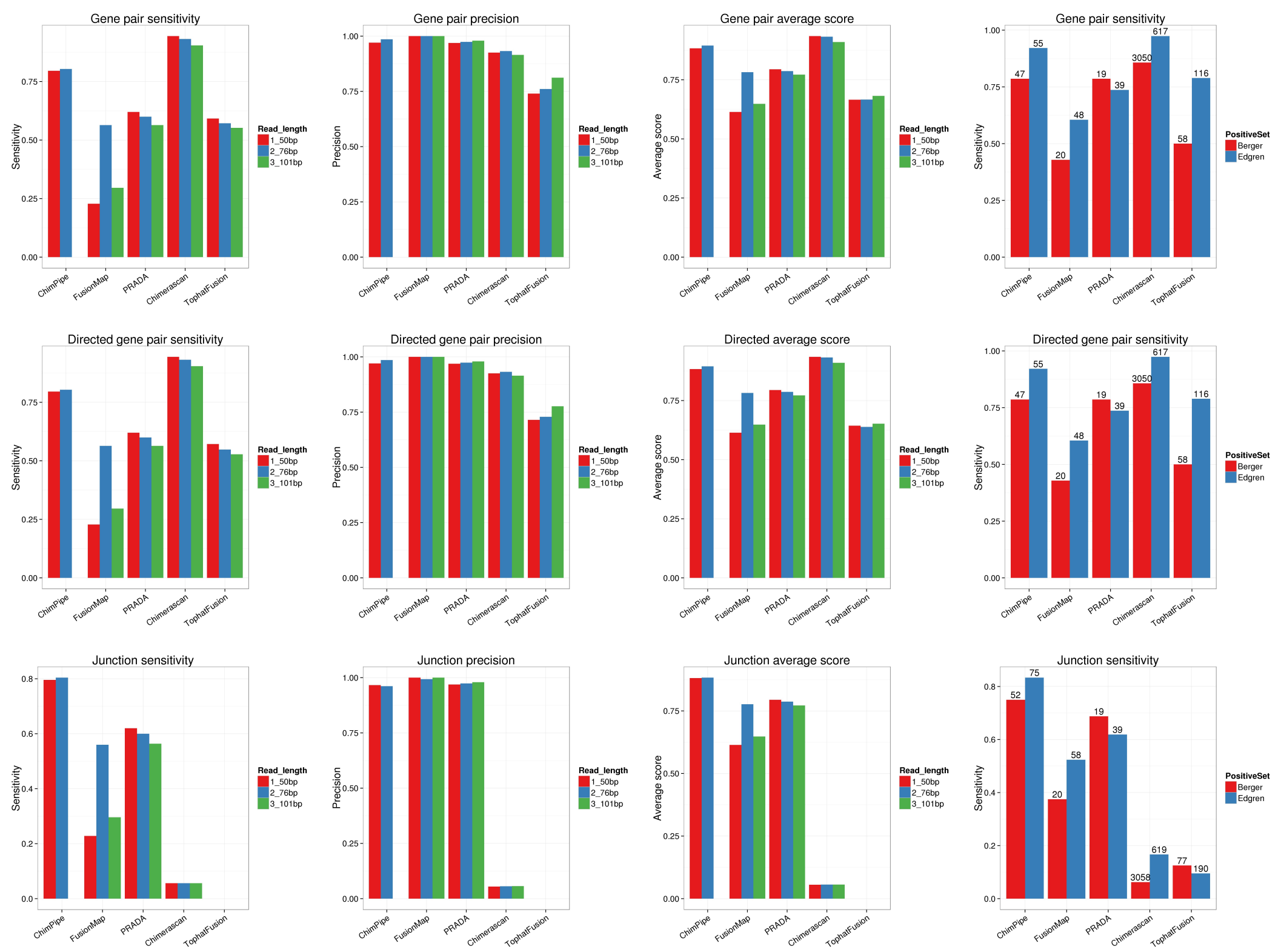




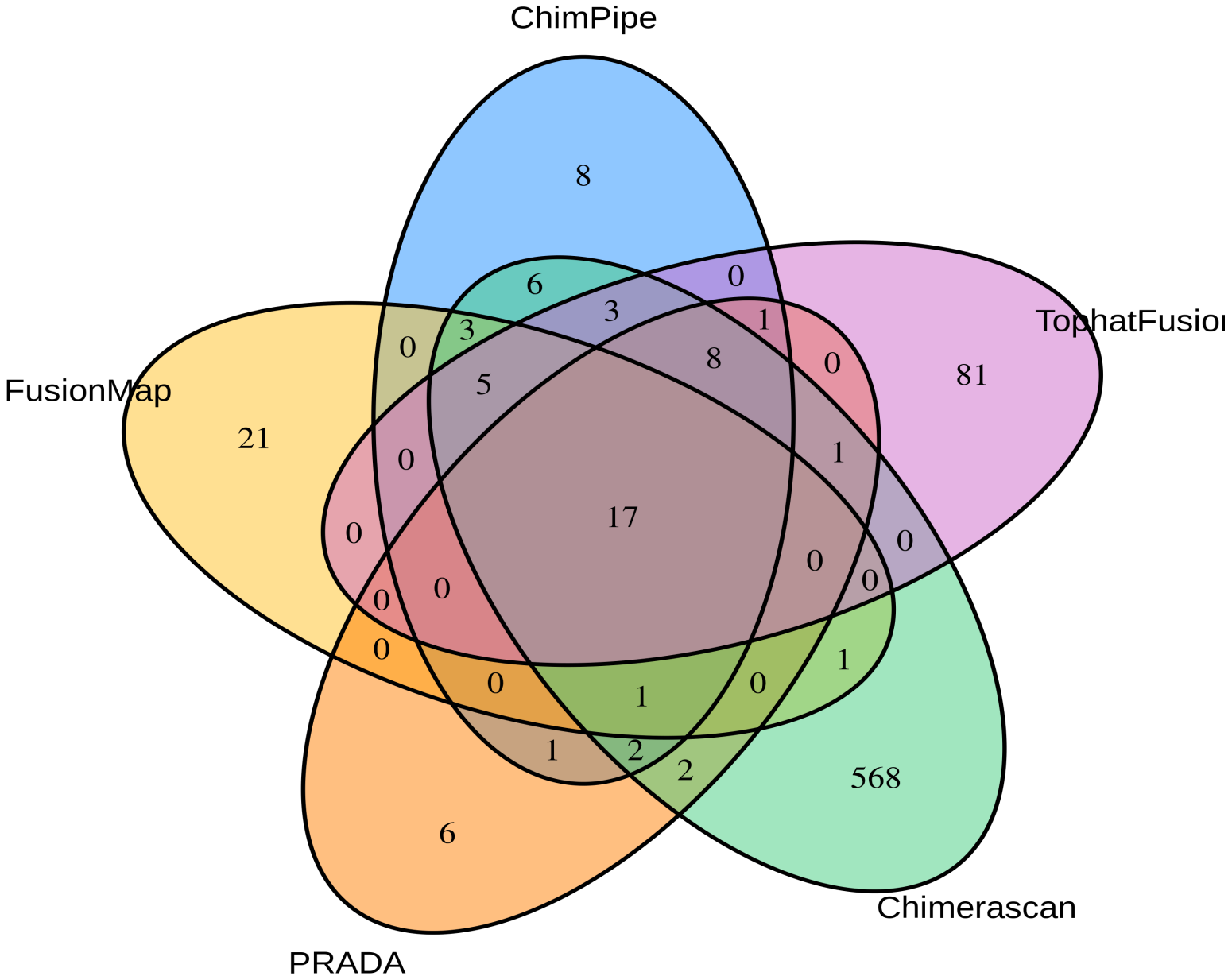







Gene pair sensitivity
Sensitivity
0.00
0.25
0.50
0.75
ChimPipe
FusionMap
PRADA
Chimerascan
TophatFusion
Read_length
1_50bp
2_76bp
3_101bp
Gene pair precision
Precision
0.00
0.25
0.50
0.75
1.00
Gene pair average score
Average score
Gene pair sensitivity
47
55
20
48
19
39
3050
617
58
116
PositiveSet
Berger
Edgren
Directed gene pair sensitivity
Directed gene pair precision
Directed average score
Directed gene pair sensitivity
Junction sensitivity
0.0
0.2
0.4
0.6
0.8
Junction precision
Junction average score
Junction sensitivity
52
75
20
58
19
39
3058
619
77
190

Output of the 5 programs on the breast cancer dataset



The program with less unique chimeras is PRADA, then ChimPipe, FusionMap, Chimerascan
There are many chimeras common to chimpipe and 2 other programs

Distance between predicted and true junction

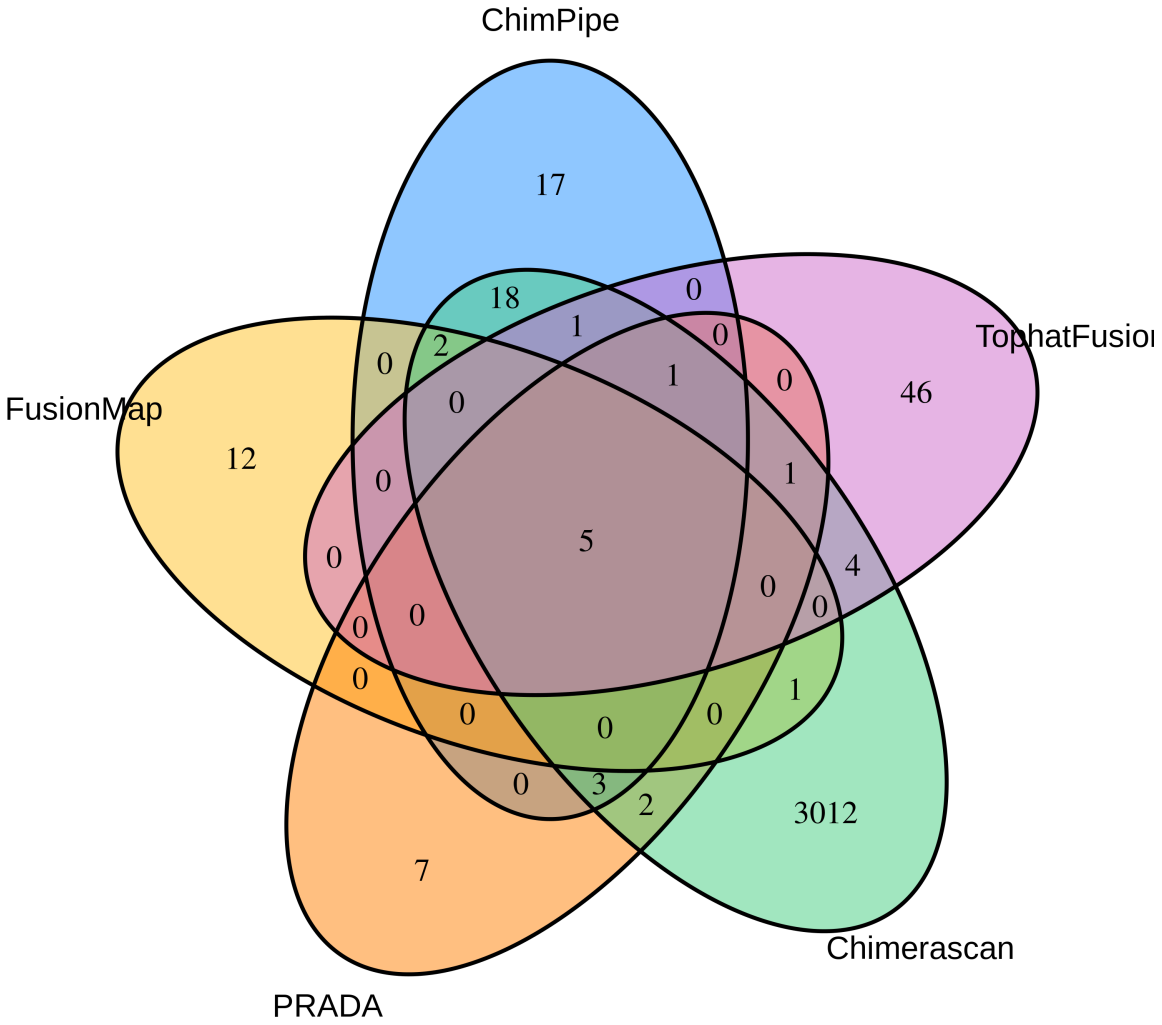
Simulation sets

Program	50 bp		76 bp		101 bp	
	# junctions	dist_avg+-dist_std	# junctions	dist_avg+-dist_std	# junctions	dist_avg+-dist_std
ChimPipe	158	0+-0	163	0+-0	NA	0+-0
FusionMap	57	0+-0	141	0.03+-0.34	73	0+-0
PRADA	155	0+-0	150	0+-0	141	0+-0
Chimerascan	193	2.85+-15.04	189	119.06+-1449.07	183	157.98+-1187.20
TophatFusion	141	732 227+-6 253 300	135	764 770+-6 389 800	130	1 706 020+-14 986 400

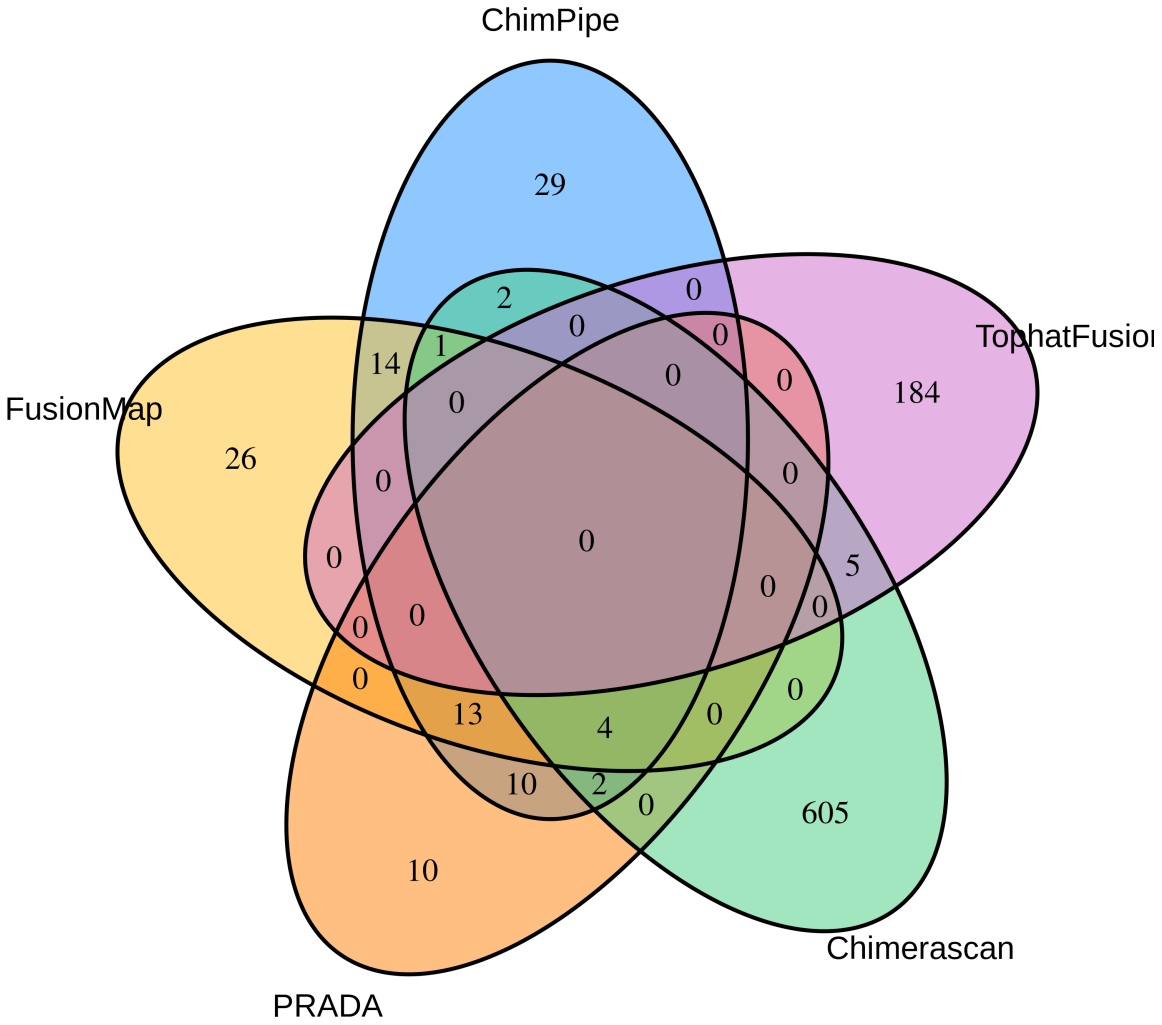
Positive sets

Program	Berger		Edgren	
	# junctions	dist_avg+-dist_std	# junctions	dist_avg+-dist_std
ChimPipe	11	0+-0	35	247.60+-1426.69
FusionMap	6	0+-0	23	27.57+-95.72
PRADA	11	0+-0	28	274.36+-1408.88
Chimerascan	12	592.17+-1866.81	37	402.97+-1639.35
TophatFusion	7	2+-0	30	1 015 780+-5 563 540

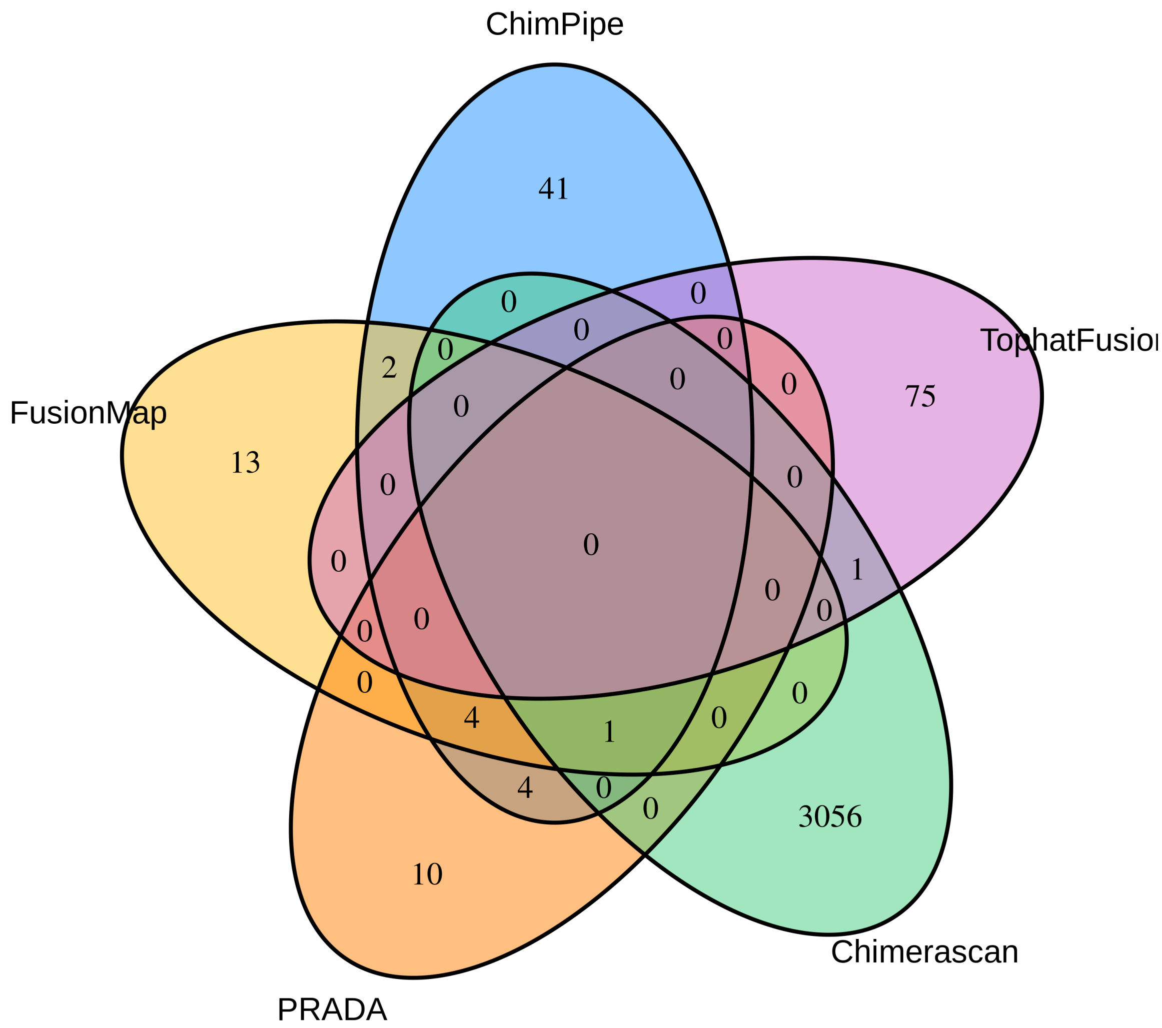
Output of the 5 programs on the melanoma dataset (gene pair level)



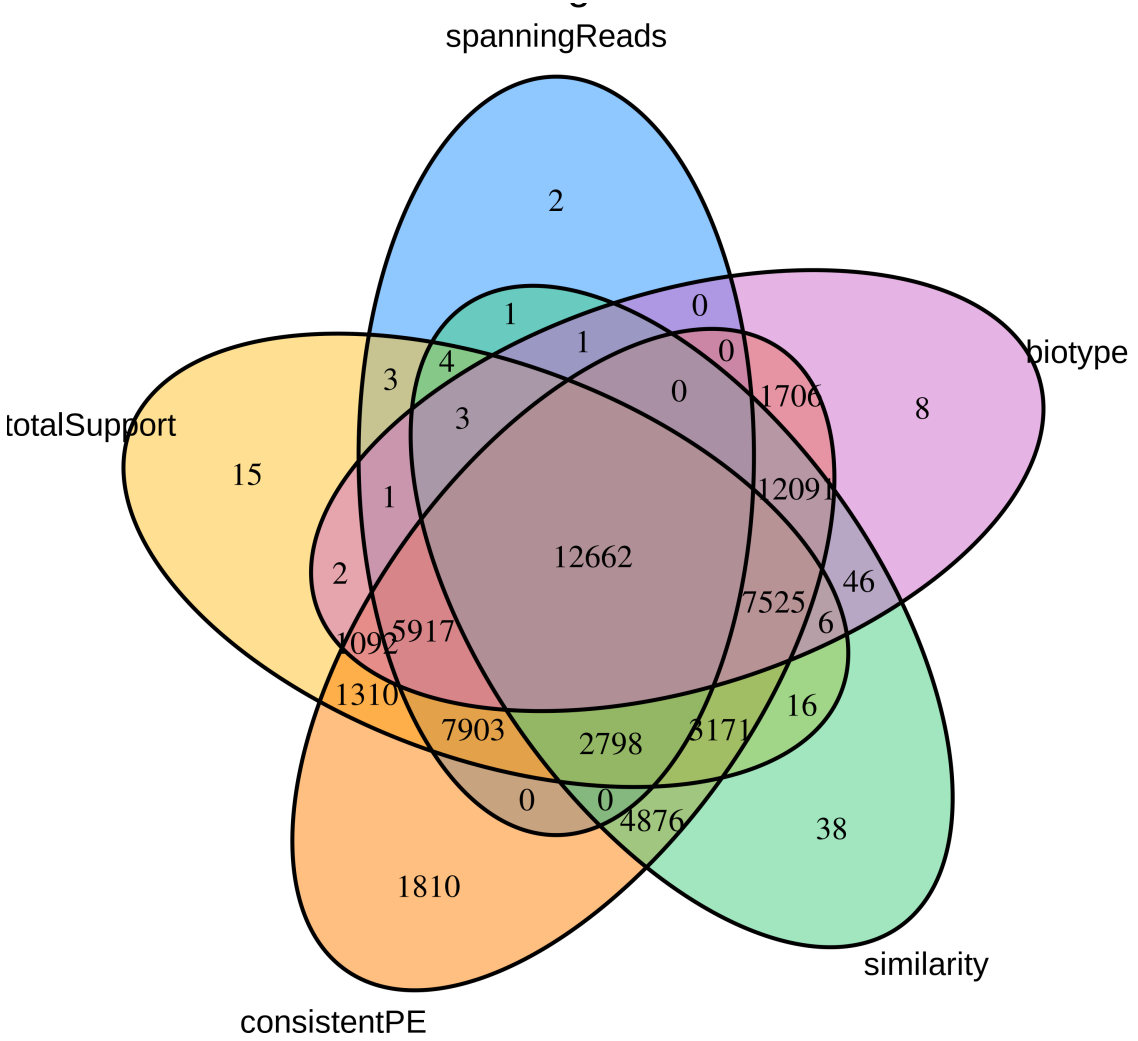
Output of the 5 programs on the breast cancer dataset (junction level)



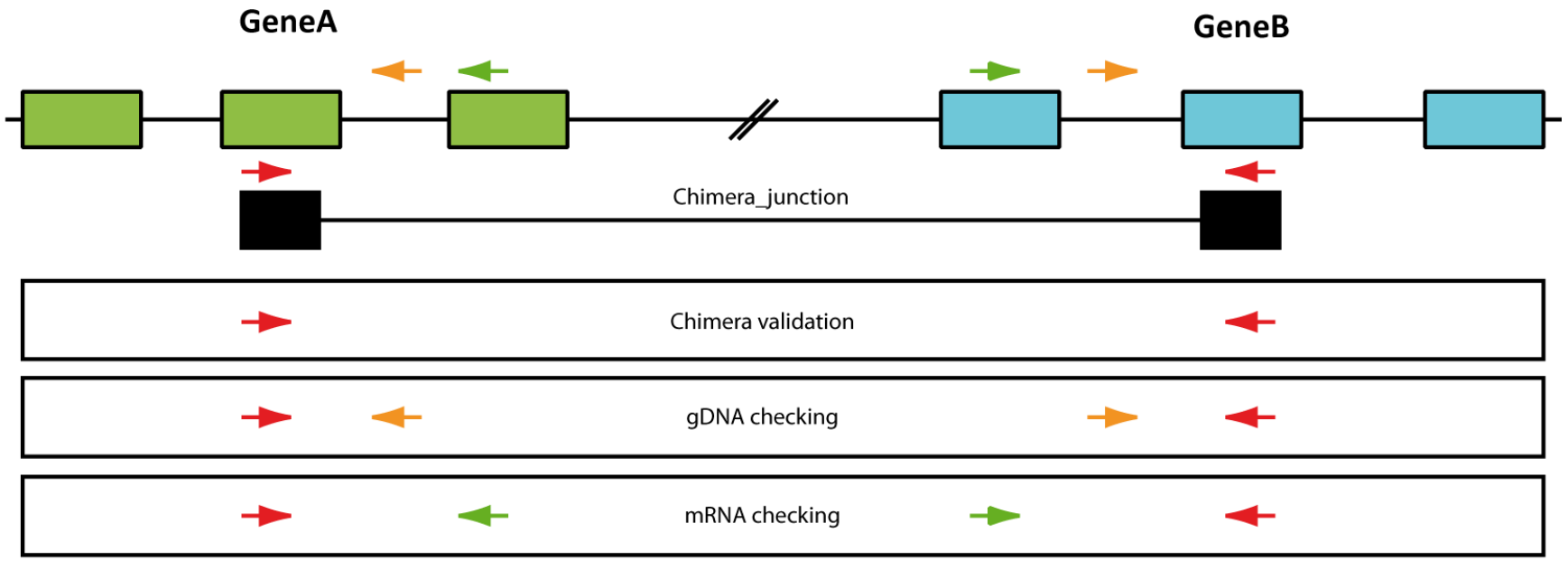
Output of the 5 programs on the melanoma dataset (junction level)



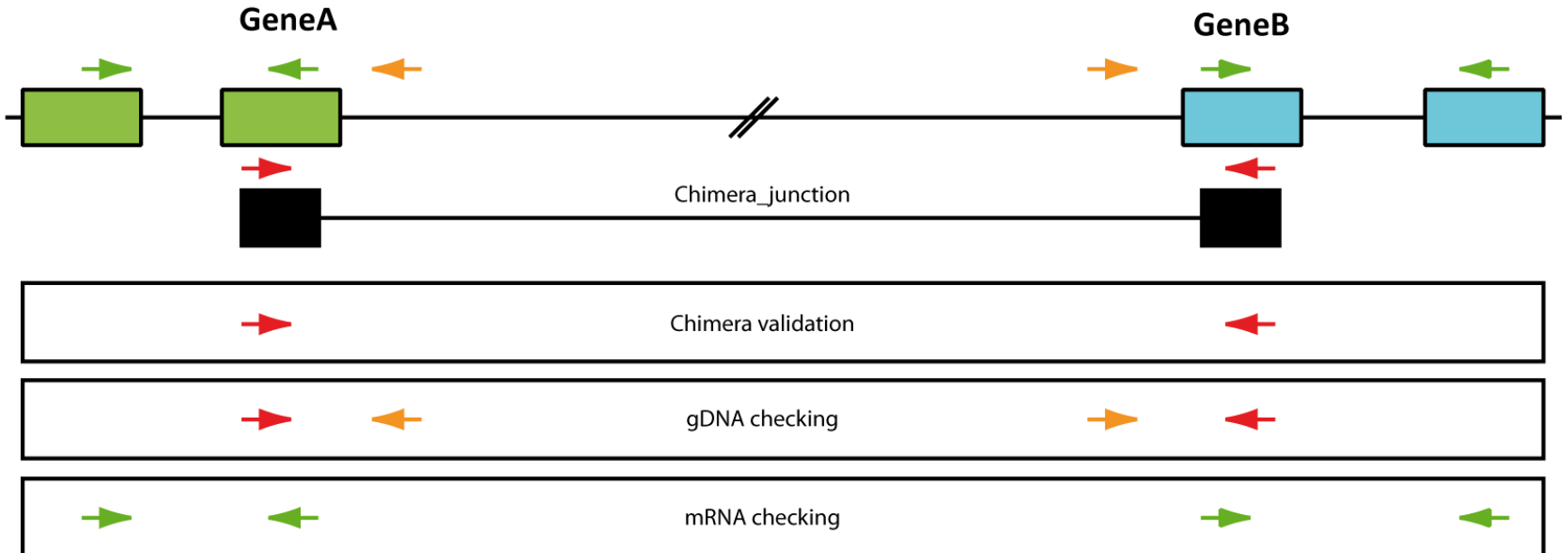
Filtering resons for chimpipe on the breast cancer dataset

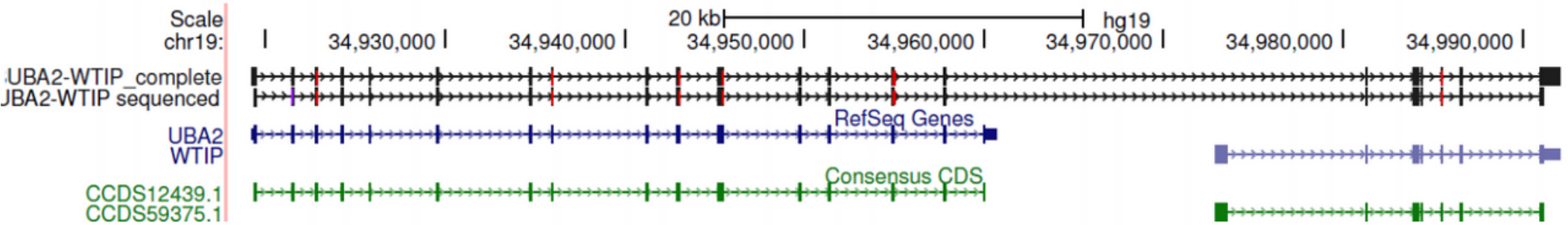


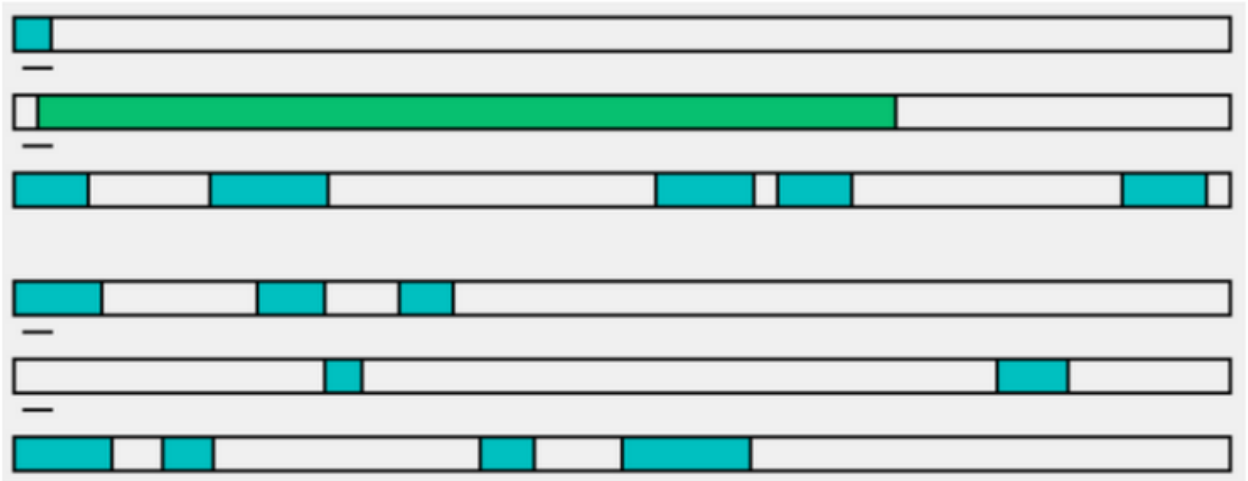
A) With internal exons



B) Without internal exons







Frame	from	to	Length
+2	71	2437	2367
-3	1682	2029	348
+3	546	872	327
+3	1773	2045	273
-3	1	271	271
-1	1	246	246
+3	3060	3284	225
+3	2103	2315	213
+3	3	209	207
-2	2712	2906	195
-1	679	858	180
-1	1063	1218	156
-3	1289	1441	153
-3	416	556	141
+1	1	108	108
-2	858	959	102

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 guees 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 remanining validated isoforms.

Chimeric junctions from Encode RNAseq experiments

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

Min.	1st Qu.	Median	Mean	3rd Qu	Max	Number of exp
0	3	12.5	16.98	27	74	108

- 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.
- <u>On the 400 highly reliable junctions:</u>
  - 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)).
- <u>On the 386 intra-chromosomal ones:</u>
  - 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):

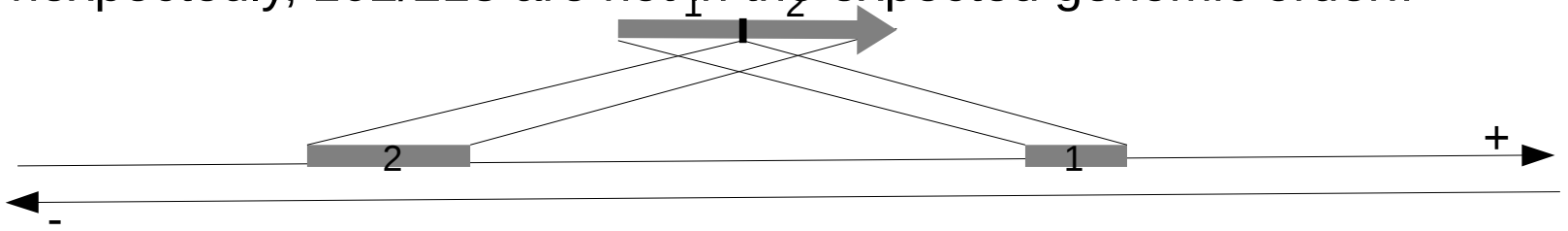
Min.	1st Qu.	Median	Mean	3rd Qu	Max	Number of junctions
0	1,424	7,708	423,800	40,230	107,000,000	386

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:

Min.	1st Qu.	Median	Mean	3rd Qu	Max
231	6,274	25,020	736,900	89,150	107,000,000

- 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):

# cases / type	Min.	1st Qu.	Median	Mean	3rd Qu	Max
102 / bad order	621	5,840	20,120	421,300	91,490	21,140,000
116 / good order	231	9,054	27,710	1,014,000	85,670	107,000,000

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.

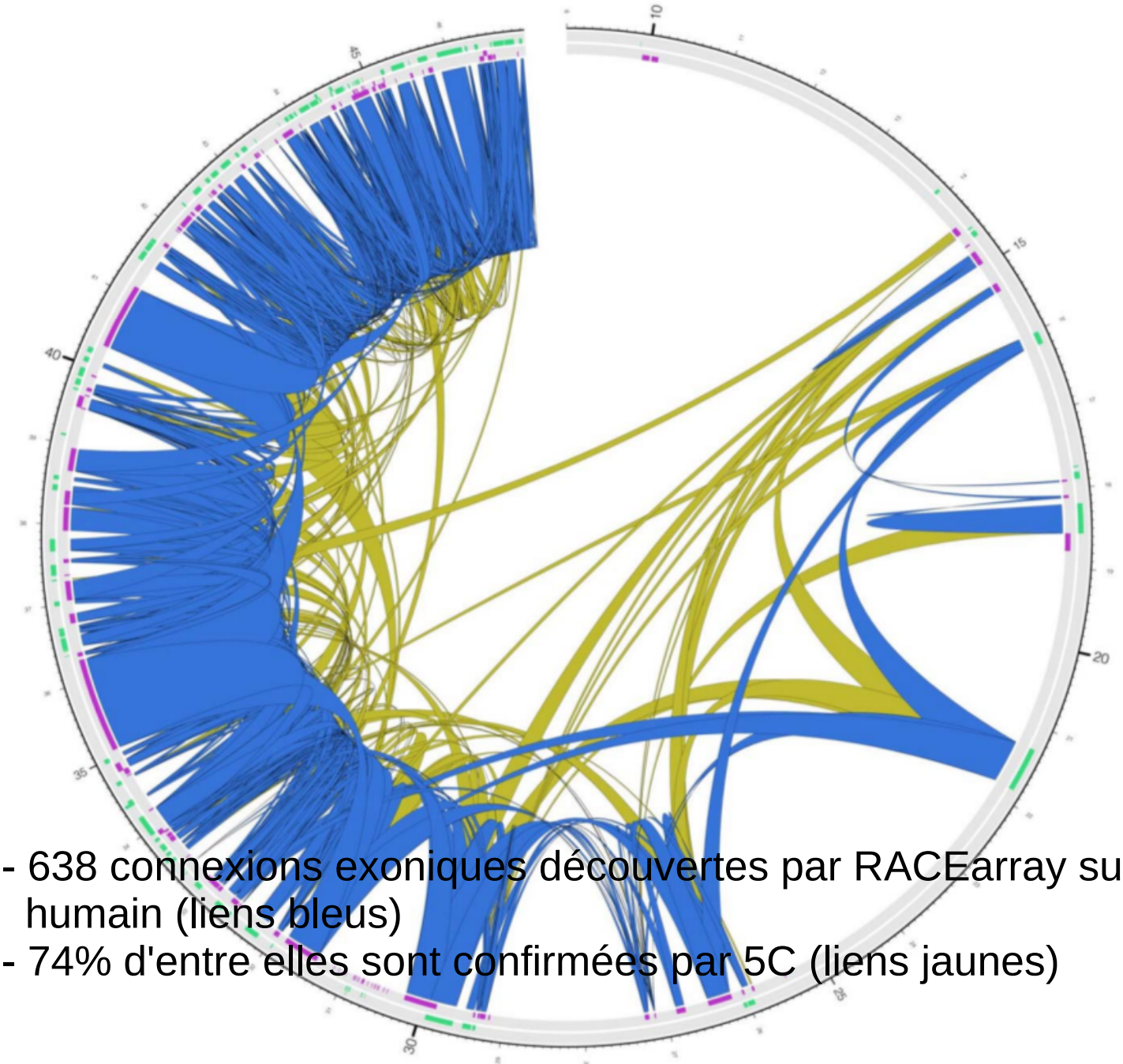
Issues with current benchmarks

- 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



RACEarray

5C

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