BIOINFORMATICS

LAB2: 1st Exercise

PACIELLO Giulia
May 5-6th 2016

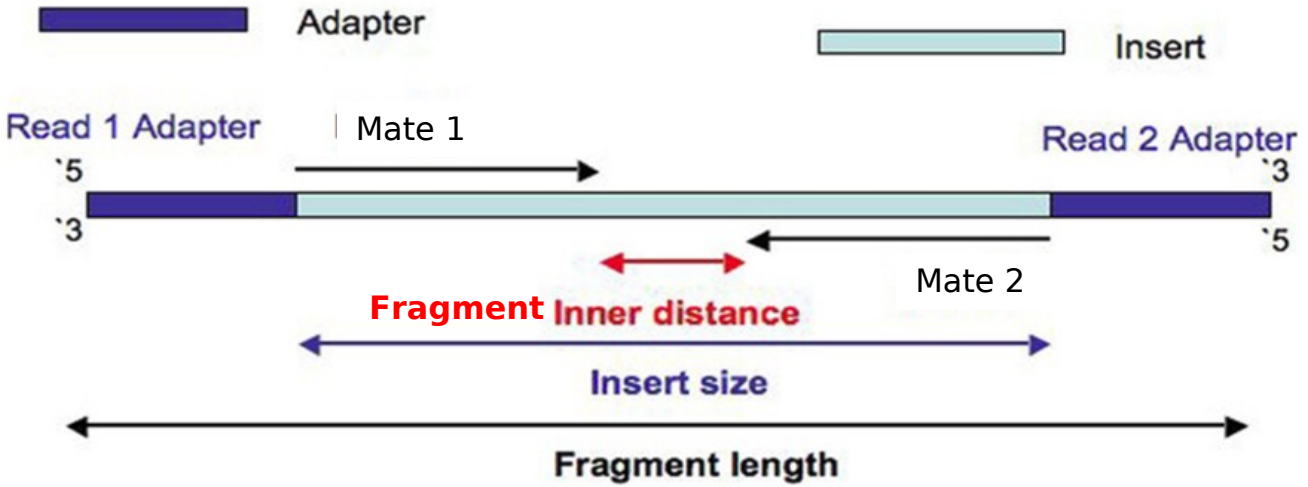
REMINDER (1)

- READ: Elementary sequence output of Next Generation Sequencing Machines.

  ATCGTCGATCCAGTCCCGGTACGTCGTCCCGGCT

- SINGLE-END SEQUENCING: Allows users to sequence only one end of a DNA/RNA fragment.

- PAIRED-END SEQUENCING: Allows users to sequence both ends of a DNA/RNA fragment. Paired-end sequencing facilitates detection of genomic rearrangements, as well as gene fusions and novel transcripts.



REMINDER (2)

- FASTQ FILE: Text-based format file storing biological sequences and their corresponding quality scores.

```
@HWI-1KL152:175:C3D02ACXX:1:1101:1206:2168/1
GGGCACCAAAAACTTTATTGCCAGTGGTAGTTCTGGCAAGGCCTGCATCCAAATAGCAGGTGAAGGCACCTGGCTGACCATCAATGCTTTCCACATTGTAT
+
@1=DDFFFGHHHHJJJIIJIJJJJGHIHIJIJIJJJIJIJJIJJJJJJJJJJFIJIJJJIIJJJJIJJIHHFGFFBECEECECCDCDDEDCCDDDDDDDDD
@HWI-1KL152:175:C3D02ACXX:1:1101:1302:2077/1
TGCGTAGGGATGGGAGGGCGATGAGGACTAGGATGATGGCGGGCAGGATAGTTCAGACGGTTTCTATTTCCTGAGCGTCTGAGATGTTCAGATCGGAAGAG
+
111A+2@DD+CDDBAEDAE@0:D9BD98BDB39<DEDA##########################################################################
@HWI-1KL152:175:C3D02ACXX:1:1101:1375:2092/1
GGGGCCCCTCAGAATGATATTTGGCCTCACGGGAGGACATAGCCTATGAAGGCTGTTGCTATAGTTGCAAGCAGGAGGATAATGCCGATGTTTCAGGTTTC
+
?11ADDADHFHHFIIIG@HHHIHIGIGGHIFIJGGGIIIICGIJIFHIIIJIJEGIBAHEHEHHACEFFFFBEC9?@B=CDACDDDD/<B<CDDDACAACA
@HWI-1KL152:175:C3D02ACXX:1:1101:1422:2104/1
GGTGGTTCTGCCTTGCTGCTGGTACTGATAACTTCTTGCTTCAGTTCATCTACAATGATCTTTCCCTCTAAATCCCAGATCTTGATGCTGGGGCCTGTGGC
+
B+1ADDFFHHHHHJJJJJJJJIJGHIIJIJJJJJJJJJJJJJJJJJIJJJJIJJIIIJJJIJJJJJIIIJJJGIJJJJGIIEHHGHHHHFFFFFDDDDDDDDDD
@HWI-1KL152:175:C3D02ACXX:1:1101:1358:2111/1
GGGCATTGTGTTAAGCTGGGGGAAGTTGCAGCTTGTCCTGAATTACATGGAGAGGAATAAAGTGGGTGAGTAGGAGGAGCCATTTCATCTGAAAGCCTGAG
+
@1:BDEFFHHHHHJJJJIGJJJGHIJIIJJJJIJJJIJIJIJIJJJIJJHHGGIGJJJJJJHHHH=BEFA:ACE@BBBDDCCDEEEEEDDCDDCCDDBDD
```

Fastq File storing mates 1

```
@HWI-1KL152:175:C3D02ACXX:1:1101:1206:2168/2
ATTTGTCAGATTGCTTATGCCCGTATAGAGGGGGATATGATAGTCTGCGCAGCGTATGCACACGAACTGCCAAAATATGGTGTGAAGGTTGGCCTGACAAA
+
@@BFFDFDHHHHHJJJIJIIIIJFGCGIHFHJEIIDGIIGGIGIJIIJJJJIHHHFEFFEEDEDDDDDDDDDACADDEEDCDABC@CDCDDDDDCDCDCD<
@HWI-1KL152:175:C3D02ACXX:1:1101:1302:2077/2
TAACATCTCAGACGCTCAGGAAATAGAAACCGTCTGAACTATCCTGCCCGCCATCATCCTAGTCCTCATCGCCCTCCCATCCCTACGCAAGATCGGAAGCG
+
;@@DAD>D>FABBCD:?DDE<ED<<FHEG@FF8?;;DHECE99DFDHC@A@8@5=CCD@AACE??BE);7@=ADD?;;9:<AA3(92<<9<3:@#######
@HWI-1KL152:175:C3D02ACXX:1:1101:1375:2092/2
GCAAAATTAACCCCCTAATAAAATTAATTAACCACTCATTCATCGACCTCCCCACCCCATCCAACATCTCCGCATGATGAAACTTCGGCTCACTCCTTGGC
+
@@@FFDDDHHHHHJIJJJIJJIIHIJJGIIJJIGIGHGHGIIIIGIIGE?DGIFJ8FHHGHHFHHFDBBEE?CDDBDCDD>CDAC:?BBBDD>CC<3:@C?
@HWI-1KL152:175:C3D02ACXX:1:1101:1422:2104/2
GCTTCTGGAGGCAAGGATGGCCAGGCCATGTTATGGGATCTCAACGAAGGCAAACACCTTTACACGCTAGATGGTGGGGACATCATCAACGCCCTGTGCTT
+
CCCFFFFFHHHHHHJJJHIJJJJJJJJJJJJJIJIIIJJGIIIJJJIJIJEIIIJIJJJJJIJHHHHHFFDDDDDDCDDDDDDDDDDDDDCBDDDDDDCDDD
@HWI-1KL152:175:C3D02ACXX:1:1101:1358:2111/2
CCAGTGGGGTTTATCGAGGGGACCAGAGGGGCCTCAGGCTTTCAGATGAAATGGCTCCTCCTACTCACCCACTTTATTCCTCTCCATGTAATTCAGGACAA
+
BC@DDDFFHFHHHHJJJIJJDGHGIIIGGJJJGIIJIIJIGIJHGHEHFFFFEFDEDEDEDDDCDDD<CBD8?ADD:>C:@C>A>A>>CDDEEDC@@C@ADB
```

Fastq File storing mates 2

- FASTA FILE: Text-based format file storing biological sequences without quality information.

```
>HWI-1KL152:175:C3D02ACXX:1:1101:1206:2168/1
GGGCACCAAAAACTTTATTGCCAGTGGTAGTTCTGGCAAGGCCTGCATCCAAATAGCAGGTGAAGGCACCTGGCTGACCATCAATGCTTTCCACATTGTAT
>HWI-1KL152:175:C3D02ACXX:1:1101:1302:2077/1
TGCGTAGGGATGGGAGGGCGATGAGGACTAGGATGATGGCGGGCAGGATAGTTCAGACGGTTTCTATTTCCTGAGCGTCTGAGATGTTCAGATCGGAAGAG
>HWI-1KL152:175:C3D02ACXX:1:1101:1375:2092/1
GGGGCCCCTCAGAATGATATTTGGCCTCACGGGAGGACATAGCCTATGAAGGCTGTTGCTATAGTTGCAAGCAGGAGGATAATGCCGATGTTTCAGGTTTC
>HWI-1KL152:175:C3D02ACXX:1:1101:1422:2104/1
GGTGGTTCTGCCTTGCTGCTGGTACTGATAACTTCTTGCTTCAGTTCATCTACAATGATCTTTCCCTCTAAATCCCAGATCTTGATGCTGGGGCCTGTGGC
>HWI-1KL152:175:C3D02ACXX:1:1101:1358:2111/1
GGGCATTGTGTTAAGCTGGGGGAAGTTGCAGCTTGTCCTGAATTACATGGAGAGGAATAAAGTGGGTGAGTAGGAGGAGCCATTTCATCTGAAAGCCTGAG
```

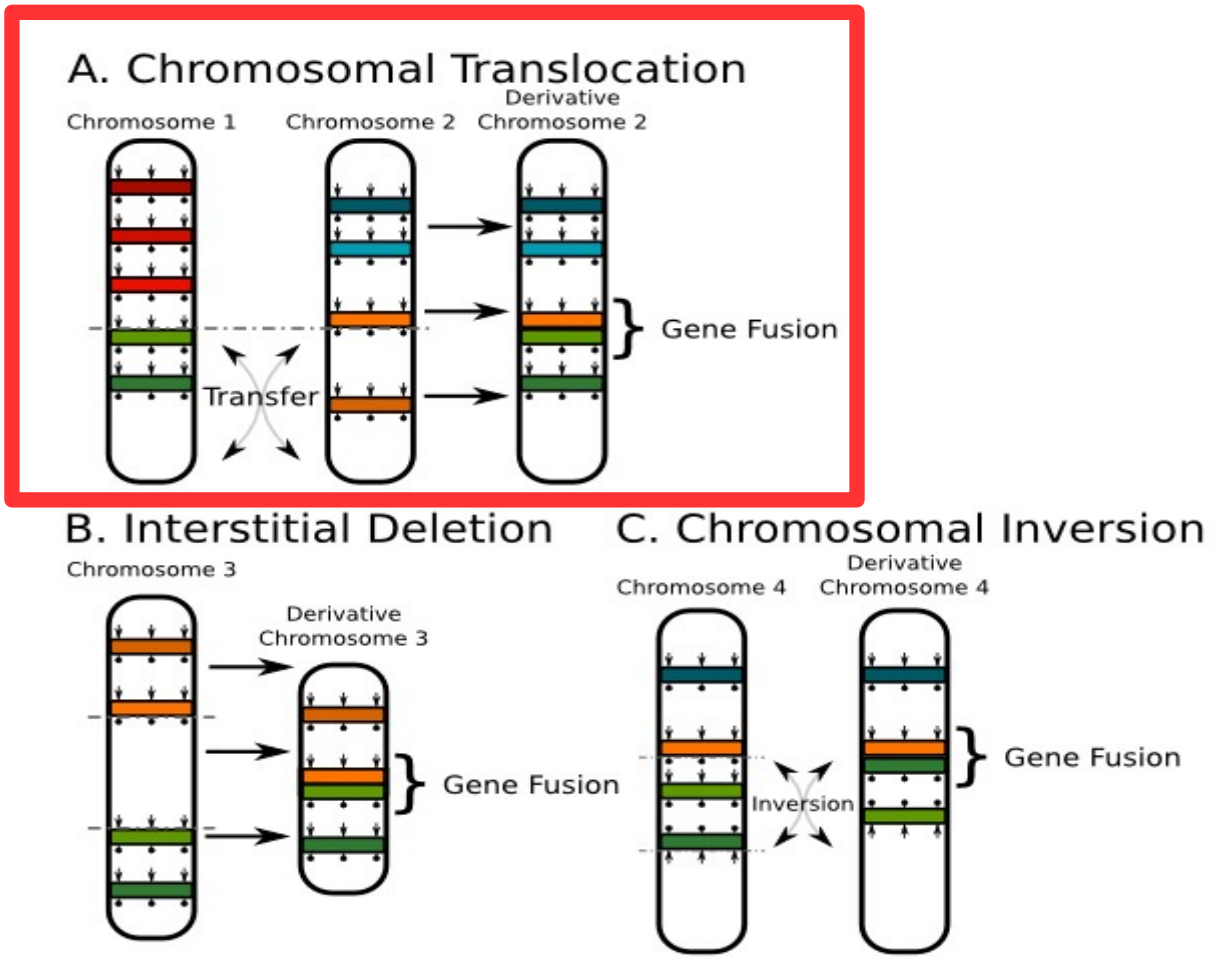
Fasta File storing mates 1

```
>HWI-1KL152:175:C3D02ACXX:1:1101:1206:2168/2
ATTTGTCAGATTGCTTATGCCCGTATAGAGGGGGATATGATAGTCTGCGCAGCGTATGCACACGAACTGCCAAAATATGGTGTGAAGGTTGGCCTGACAAA
@HWI-1KL152:175:C3D02ACXX:1:1101:1302:2077/2
TAACATCTCAGACGCTCAGGAAATAGAAACCGTCTGAACTATCCTGCCCGCCATCATCCTAGTCCTCATCGCCCTCCCATCCCTACGCAAGATCGGAAGCG
>HWI-1KL152:175:C3D02ACXX:1:1101:1375:2092/2
GCAAAATTAACCCCCTAATAAAATTAATTAACCACTCATTCATCGACCTCCCCACCCCATCCAACATCTCCGCATGATGAAACTTCGGCTCACTCCTTGGC
>HWI-1KL152:175:C3D02ACXX:1:1101:1422:2104/2
GCTTCTGGAGGCAAGGATGGCCAGGCCATGTTATGGGATCTCAACGAAGGCAAACACCTTTACACGCTAGATGGTGGGGACATCATCAACGCCCTGTGCTT
>HWI-1KL152:175:C3D02ACXX:1:1101:1358:2111/2
CCAGTGGGGTTTATCGAGGGGACCAGAGGGGCCTCAGGCTTTCAGATGAAATGGCTCCTCCTACTCACCCACTTTATTCCTCTCCATGTAATTCAGGACAA
```

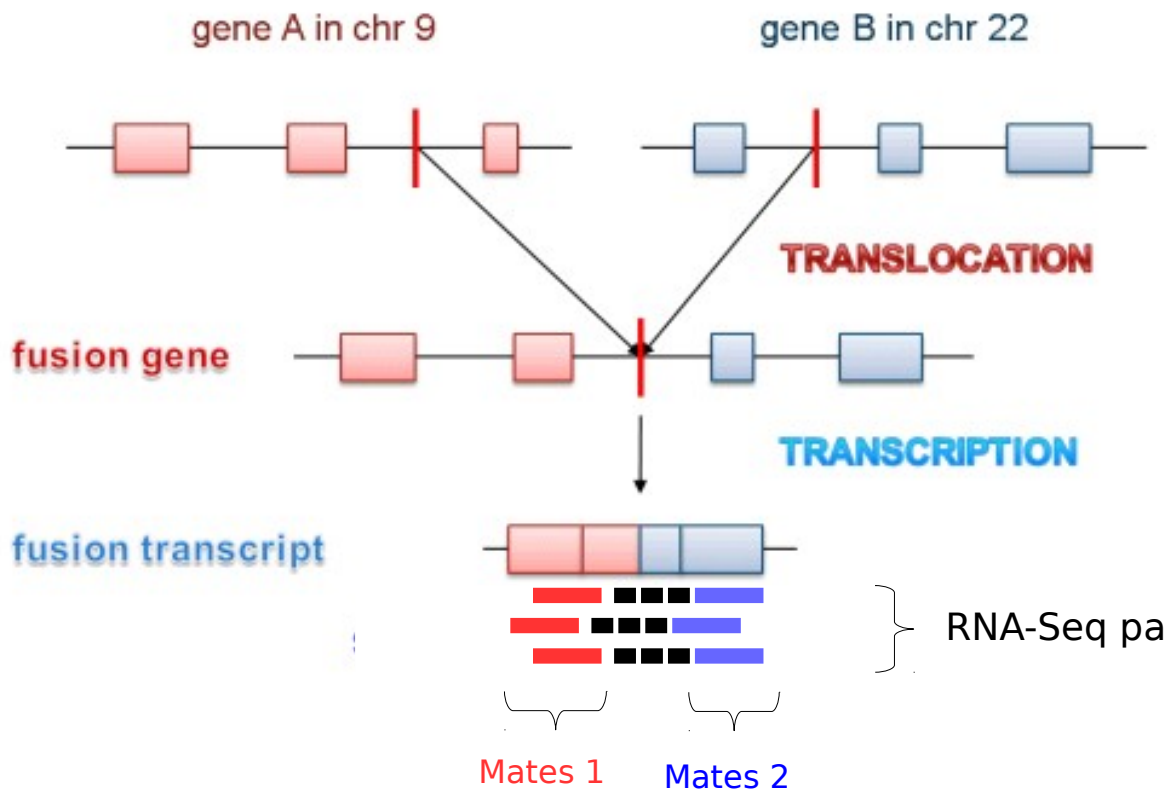
Fasta File storing mates 2

REMINDER (3)

- <u>FUSION GENE:</u> A fusion gene is a hybrid gene formed by two previously separated genes. At the DNA level these genes can originate from three different mechanisms.



REMINDER (4)

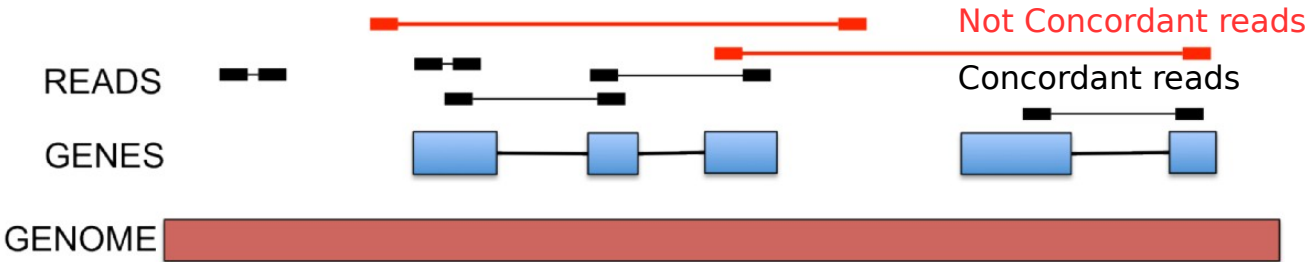

gene A in chr 9
gene B in chr 22
TRANSLOCATION
DNA Level
fusion gene
TRANSCRIPTION
RNA Level
fusion transcript
RNA-Seq paired-end Reads
Mates 1
Mates 2

CHIMERASCAN WORKFLOW (1)

- CREATE ALIGNMENT INDEX
  Chimerascan indexer uses the bowtie-build indexing scheme from Bowtie to build the index of the reference genome and transcriptome on which reads will be mapped.

- PREPARE READS FOR ALIGNMENT:
  1)All reads quality scores are converted from Phred to Sanger format
  Phred = $-10\log_{10}P$ where P is the base-calling error probability.
  Sanger = Phred+22
  2)All reads identifier are converted from an arbitrarily long string to a number.

- ALIGN PAIRED-END READS:
  Paired-end reads are aligned to the genome and the transcriptome reference taking advantage of Bowtie aligner. Both the mates belonging to a pair-end read must align on the reference within a distance fixed by the user (Concordant reads).

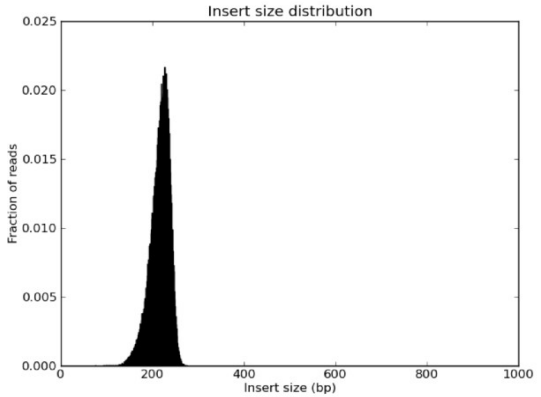


Concordant reads: Align as a pair on the genome (within a fixed distance) or on the transcriptome within a single gene.
Not Concordant reads: Span multiple genes or large genomic distance.

CHIMERASCAN WORKFLOW (2)

- CREATE A SORTED/INDEXED BAM FILE
  Alignments of concordant reads, originally stored in a SAM file, are sorted and compressed to produce an indexed BAM file.

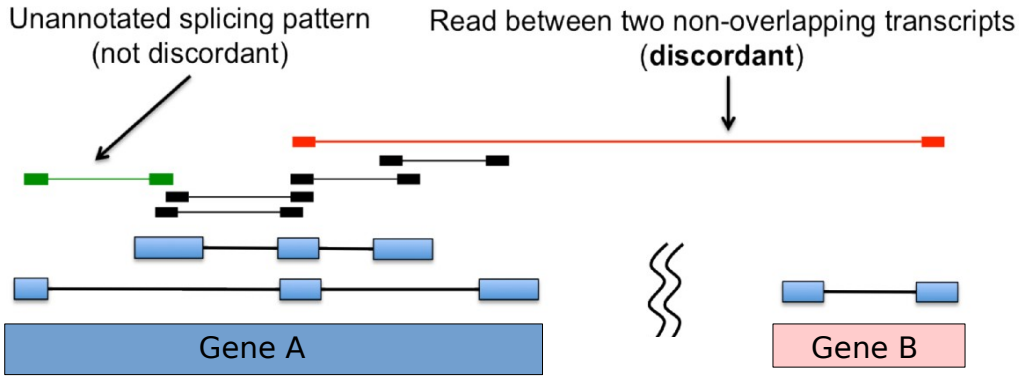
- ESTIMATE FRAGMENT SIZE DISTRIBUTION
  Aligned reads are parsed in order to compute the empirical distribution of inner distances. Only uniquely mapped reads are used to sample this distribution.



- REALIGN INITIALLY UNMAPPED READS
  Reads that have not been aligned concordantly are trimmed into smaller segments and realigned on the reference genome and transcriptome.

CHIMERASCAN WORKFLOW (3)

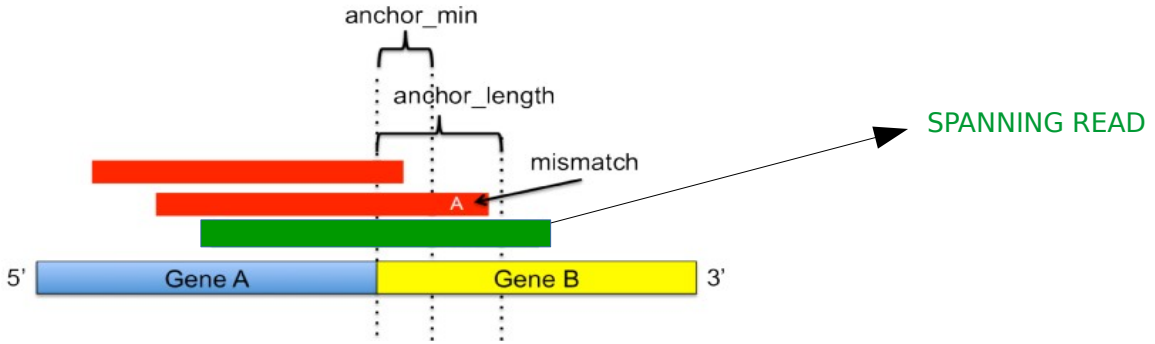
- DISCOVER DISCORDANT READS
  The trimmed alignments are scanned for evidence of discordant reads. A discordant read is defined if:
  1)The two mates do not align to the genome within a user-specified inner distance;
  2)The pair does not align to a single transcript;
  3)The two mates do not map on different transcripts of the same gene.



- NOMINATE CHIMERAS
  ChimeraScan clusters discordant reads producing a list of putative chimeric candidates (A). For each putative chimeric transcript ChimeraScan build a virtual reference by extracting from the genome its sequence (B).

CHIMERASCAN WORKFLOW (4)

- CANDIDATE SPANNING READS REALIGNMENT AND SPANNING READS DEFINITION
  Discordant reads with trimmed alignments bordering a junction or unmapped reads whose mates align to a predicted chimera are realigned on the obtained putative chimeric transcripts sequences. Spanning reads are defined accordingly to a minimum anchor length and a maximum number of mismatches in the anchor region.



- FILTERING CHIMERAS
  In order to ensure high reliability of the detection:
  1) Chimeras with very low coverage are filtered;
  2) Chimeras with fragment sizes far outside the range of the distribution are filtered;
  3) Known false positives are filtered.

- REPORTING CHIMERAS
  ChimeraScan produces a tabular text file describing each chimera, and optionally generates a user-friendly HTML page with links to detailed descriptions of the chimeric genes.

EXERCISE OBJECTIVES

1) Identify gene fusions in the sample dataset_mate*.fq using Chimerascan Tool;

2) Check on COSMIC database if the identified gene fusions have been previously reported:

http://cancer.sanger.ac.uk/cancergenome/projects/cosmic/

3) Reconstruct, starting from the so called spanning reads provided by Chimerascan tool for the detected gene fusions, the longest consensus region sequence for the fusion. This activity has to be performed by writing an ad-hoc python script.

DATA (downloaded from the Course Material page)

1)**The INDEX Folder containing:**
- *align_index.fa*: The reference file. It is in fasta format and stores some portions of sequences belonging to human chr 3, 11, 18 and 21.
For computational reasons we previously produced all the index files needed for Chimerascan run (*align_index.1.ebwt,alighn_index.2.ebwt,align_index.3.ebwt,align_index.4.ebwt,align_index.fa.fai,align_index.rev.1.ebwt,align_index.rev.2.ebwt*) thanks to Chimerascan_index.py.

2)**The input files:**
- *dataset_mate1.fq* and *dataset_mate2.fq*: These files contain respectively mate1 and mate2 sequences of the sample under investigation.

CHIMERASCAN RUN

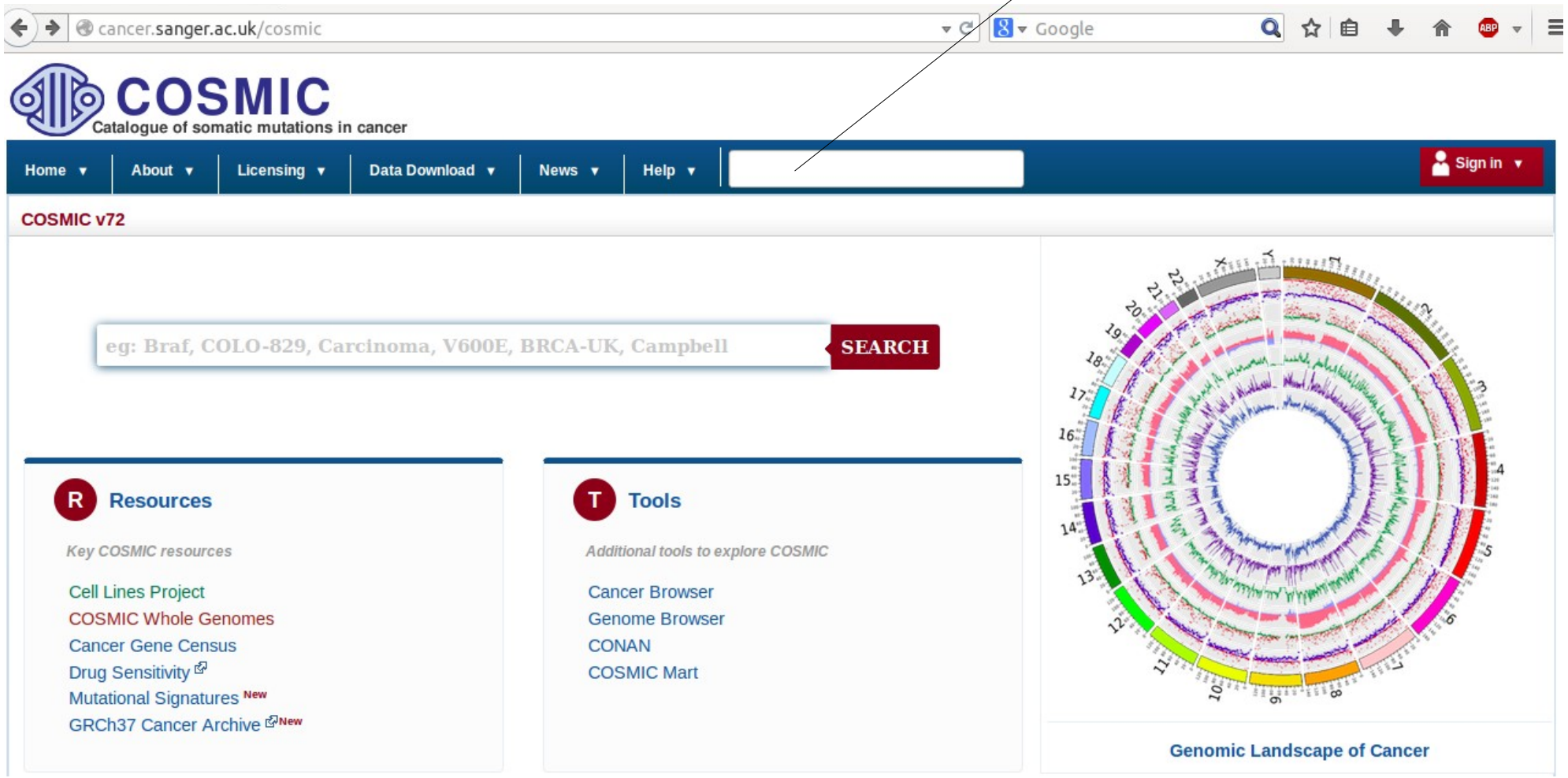
1)Create a folder directory to store Chimerascan output (chimera_out dir)

2)Run chimerascan_run.py program:
python /opt/chimerascan-0.4.5/chimerascan/chimerascan_run.py path_to_indexfolder/INDEX/ /path_to_datasetfolder/dataset_mate1.fq /path_to_datasetfolder/dataset_mate2.fq /path_to_outputdirectory/chimera_out

3) Try to change Chimerascan running parameters:
https://code.google.com/p/chimerascan/wiki/Running

COSMIC DB SEARCH

Insert here Gene Name



CHIMERASCAN OUTPUT ELABORATION

Starting from chimeras.bedpe file write a program to elaborate the output in order to obtain from the spanning reads a breakpoint consensus sequence:

BEFORE

GCGGAGGCGGAGGGCGAGGGGCGGGGAGCGCCGCCTGGAGCGCGGCAG|GAAGC
CGGAGGCGGAGGGCGAGGGGCGGGGAGCGCCGCCTGGAGCGCGGCAG|GAAGCC
CGGAGGCGGAGGGCGAGGGGCGGGGAGAGCCGCCTGGAGCGCGGCAG|GAAGCC
CGGCGGCGGAGGGCGAGGGGCGGGGAGCGCCGCCTGGAGCGCGGCCG|GAAGCC
GGAGGCGGAGGGCGAGGGGCGGGGAGCGCCGCCTGGAGCGCGGCAG|GAAGCCT
GGAGGCGGAGGGCGAGGGGCGGGGAGAGCCGCCTGGAGCGCGGCAG|GAAGCCT
GGTGGCGGAGGGCGAGGGGCGGGGAGCGCCGCCTGGAGCGCGGCAG|GAAGCCT
GGAGGCGGAGGGCGAGGGGCGGGGAGCGCCGCCTGGAGCGCGGCAT|GAAGCCT
GGAGGCGGAGGGCGAGGGGCGGCGAGCGCCGCCTGGAGCGCGGCAG|GGAGCCT
GAGGCGGAGGGCGAGGGGCGGGGAGCGCCGCCTGGAGCGCGGCAG|GAAGCCTT
AGGCGGAGGGCGAGGGGCGGGGAGCGCCGCCTGGAGCGCGGCAG|GAAGCCTTA
GCGGAGGGCGAGGGGCGGGGAGCGCCGCCTGGAGCGCGGCAG|GAAGCCTTATC
CGGAGGGCGAGGGGCGGGGAGCGCCGCCTGGAGCGCGGCAG|GAAGCCTTATCA
GGAGGGCGAGGGGCGGGGAGCGCCGCCTGGAGCGCGGCAG|GAAGCCTTATCAG
AGGGCGAGGGGCGGGGAGCGCCGCCTGGAGCGCGGCAG|GAAGCCTTATCAGTT
GGGCGAGGGGCGGGGAGCGCCGCCTGGAGCGCGGCAG|GAAGCCTTATCAGGTC
GCGAGGGGCGGGGAGCGCCGCCTGGAGCGCGGCAG|GAAGCCTTATCAGTTGTG
GGGCGGGGAGCGCCGCCTGGAGCGCGGCAG|GAAGCCTTATCAGTTGTGAGTGA
GGCGGGGAGCGCCGCCTGGAGCGCGGCAG|GAAGCCTTATCAGTTGTGAGTGAG

AFTER

GCGGAGGCGGAGGGCGAGGGGCGGGGAGCGCCGCCTGGAGCGCGGCAG|
GAAGCCTTATCAGTTGTGAGTGAG