Genomic landscape of somatic retrotransposons in Breast Cancer Whole-Exomes

NISHA HEMANDHAR KUMAR

Study Objectives

- The aim is to find software that can detect somatic L1 and Alu retrotransposons in cancer whole exome sequencing.
- Apply to a single patient with WES and determine the concordance
- Determine if an evolutionary tree can be built
- After the clean analysis of a single patient, the goal is to apply to a larger dataset of 96 samples, to establish a retrotransposition landscape in a cohort at high risk of relapse and early detection breast screening cohort.

Methodology

- Reviewing open software to detect L1 and Alu elements from whole exome sequencing
- Installing software and testing with the whole exome BAM files
- Concordance analysis for insertions between samples
- Identify somatic retrotransposon events and map the events to known genes (BEDtools)
- Analyse the genes for relatedness to cancer (DAVID)
- Summarize the identified somatic retrotransposons from the identified software in tabular and graphical formats

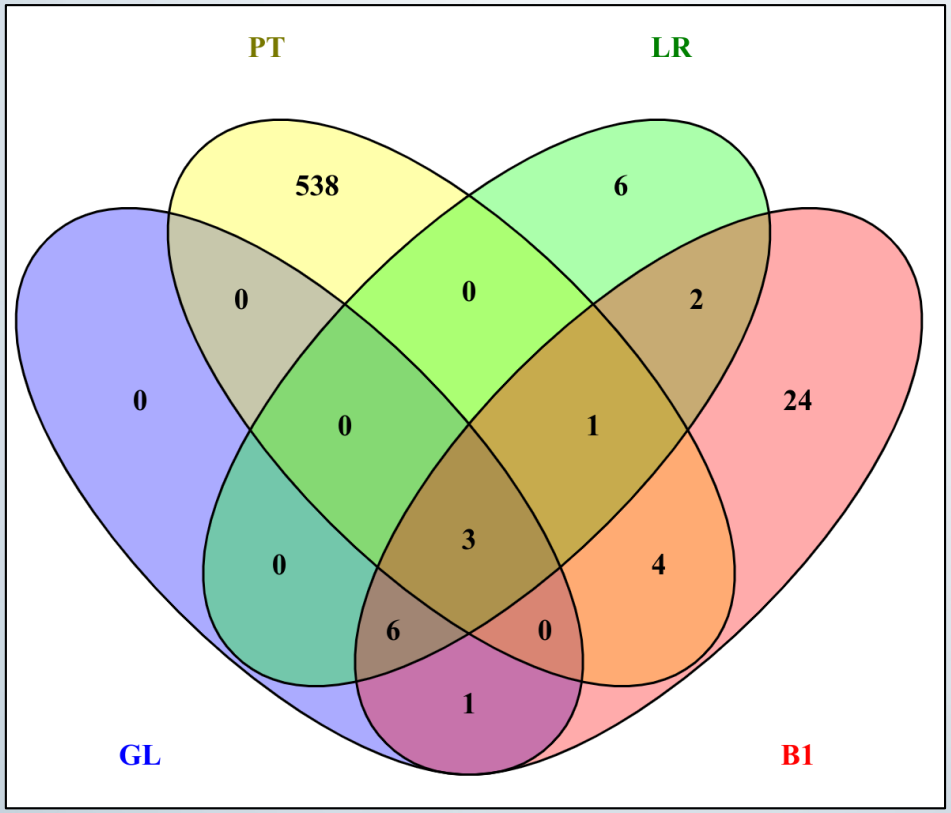
Software's for detecting TEIs in WES data

Tool	Detects in reference	Requires specific alignment	Third party tools	Approach	Implementation
Mobster	No	Yes, BWA/ Mosaik	Picard tools and MOSAIK (Version:2.1.33)	Database-based	Java
Alu-detect	No	No, but must be in BAM/ Fastq	SAMtools, BEDtools, Repeat Masker	Database-based	Python
Transposeq	No	No, but must be in BAM	SAMtools, Repeat Masker	Database-based	Java / R
mcclintock	No	No, but must be in Fastq	FastQC, RepeatMasker, R, BWA, Perl, BioPerl, SAMtools, Blat, Bowtie, faToTwoBit, twoBitToFa, BEDtools, BCFTools, Exonerate, Java	Database-based	Perl / Java
Jitterbug	No	No, but must be in BAM	pysam, pybedtools, psutil, matplotlib, matplotlib-venn, numpy	Database-based	Python
TranSurVeyor	No	Yes, BWA MEM	g++4.7.2, NumPy, PyFaidx, PySam, HtsLib	Database-free	C++, Python

Tools and Databases

- DAVID: The Database for Annotation, Visualization and Integrated Discovery
- UCSC:
  - Genome Browser: to interactively visualize genomic data
  - BLAT: rapidly align sequences to the genome
- CENSOR: classifies all known repeats
- Bedtools: allows to *intersect*, *merge*, *count*, *complement*, and *shuffle* genomic intervals from multiple files in widely-used genomic file formats such as BAM, BED, GFF/GTF, VCF.
- Samtools: Samtools is a suite of programs for interacting with high-throughput sequencing data.
- HTSlib: A C library for reading/writing high-throughput sequencing data.
- GitHub: a Git repository hosting service. While Git is a command line tool, GitHub provides a Web-based graphical interface.
- Rstudio: uses R programming language to develop statistical programs
  - ChromPlot: Visualization of genomic data in chromosomal context
- Packages in Python:
  - Matplotlib- plotting library
  - Itertools - module implements several iterator building blocks inspired by constructs from APL, Haskell, and SML.

Concordance of retrotransposon insertions in tumour-normal samples



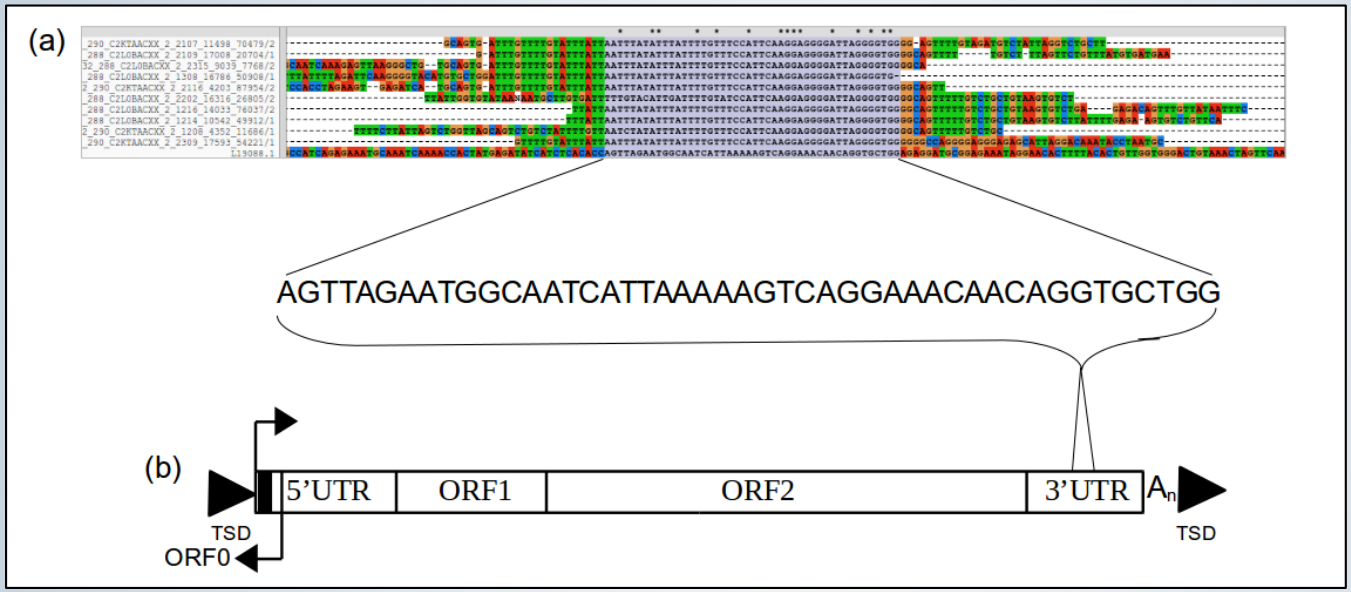
Classification and validation of retrotransposon insertions

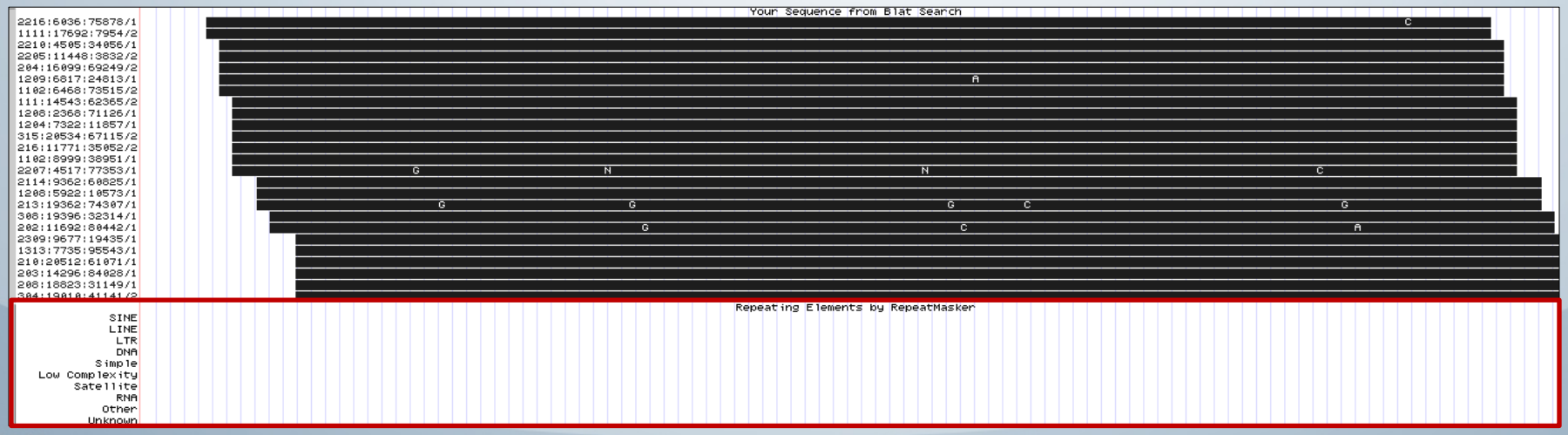
Chromosome and Insertion point	Somatic/ Germline	Gene	Germline	Primary tumour	Liver	cfDNA	MEI (Mobster)	MEI (CENSOR)
X:136655212	Somatic	ZIC3	✗	✔	✔	✔	Alu	unknown
8:89044346	Somatic	MMP16	✗	✔	✗	✔	L1	L1HS
5:6738126	Somatic	TENT4A	✗	✔	✗	✔	Alu	Alu
6:124814864	Somatic	NKAIN2	✗	✔	✗	✔	Alu	Alu
13:98119813	Somatic	RAP2A	✗	✔	✗	✔	L1	L1HS
4:187093487	Somatic	FAM149A	✗	✗	✔	✔	Alu	Alu
17:61565890	Somatic	ACE	✗	✗	✔	✔	Alu	Alu

Gene enrichment analysis using DAVID

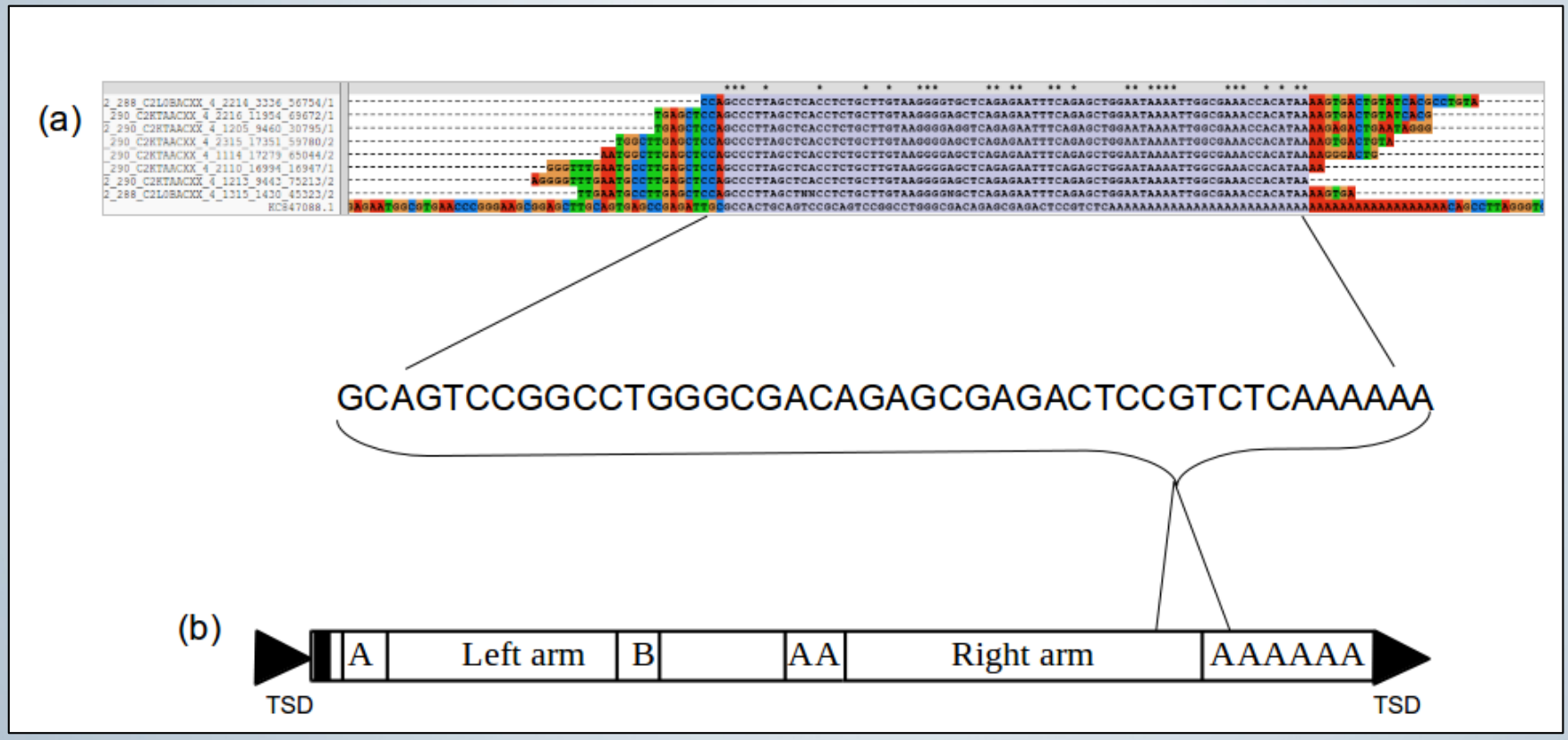
Genes	Term	Ease Score	Fisher Exact
MMP16, NKAIN2, RAP2A, ACE	Plasma Membrane	0.06	1.8E-2
ACE, MMP16	Metalloprotease	0.06	1.9E-3

L1 insertion in MMP16 a potential biomarker





Alu insertion in ACE a potential biomarker



Take home message

- Retrotransposons are mobile DNA sequence that can insert itself at a different position by using reverse transcriptase.
- cfDNA is significantly increased in the plasma of diseased patients. It was determined that cfDNA are elevated in over half of cancer patients.
- Sequencing cfDNA is the ability to sequence serially-collected and minimally-invasive plasma samples, allowing for near real-time monitoring of the tumor genome during treatment.
- Identification of the somatic retrotransposons, could likely be used as a personalized biomarkers, for analysing tumour progression and early detection of this case of metastatic breast cancer.
- Identification of SRI from whole exomes high throughput sequencing data using tools/software are time and cost-efficient


Thank
You