Viral Integration

Edmund Miller

2022-07-05 Wed

nf-core/viralintegration

Background

- Started out wondering if we could look for Viral Integration in 1000 Genomes, inspired by OncoDB
- Original: `bam2fastq` unaligned reads => nf-core/sarek (WGS) with a custom reference
- Met Robert Allaway, Principal Scientist @ Sage Bionetwork on nf-core

Background

Approaches

- MetaPhlAn (Robert)
- CTAT - Virus Integration Finder (ctat-VIF) (Alyssa Briggs)
- BLASTBox (Edmund)

Why are we interested in Viral Integration?

- Besides aberrant DNA methylation or RNA expression, another major cause of cancer is viral infection (Oncoviruses)
- Identification of oncovirus-related gene expression changes helps better understand the mechanisms underlying virus-induced cancers
- Lack of current research in the area utilizing new data

CTAT - Virus Integration Finder

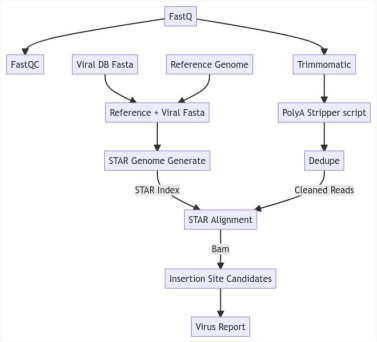
CTAT - Virus Integration Finder

- Capture evidence of virus-matching reads (consistent with an infection)
- Identify and quantify evidence for viral insertion sites in the human genome
- Provide interactive visualizations for evidence of virus-mapped reads and virus insertion sites

CTAT - Virus Integration Finder

- Trinity Cancer Transcriptome Analysis Toolkit - CWL
- Takes DNA-seq or RNA-seq
- Requires a CTAT genome lib supplemented with viral genomes (They provide a virusdb fasta).

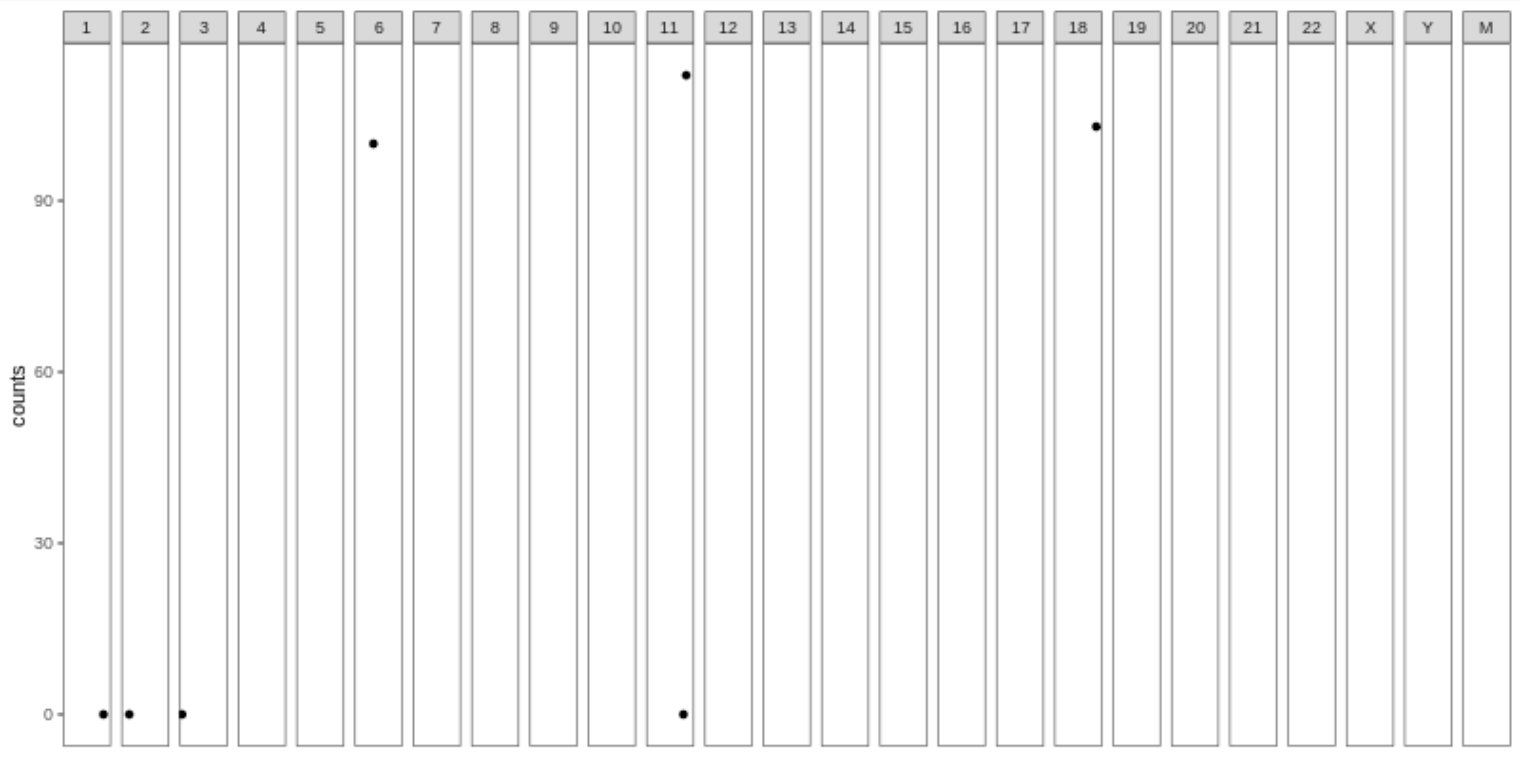
CTAT - DAG


FastQ
FastQC
Viral DB Fasta
Reference Genome
Trimmomatic
Reference + Viral Fasta
PolyA Stripper script
STAR Genome Generate
Dedupe
STAR Index
Cleaned Reads
STAR Alignment
Bam
Insertion Site Candidates
Virus Report

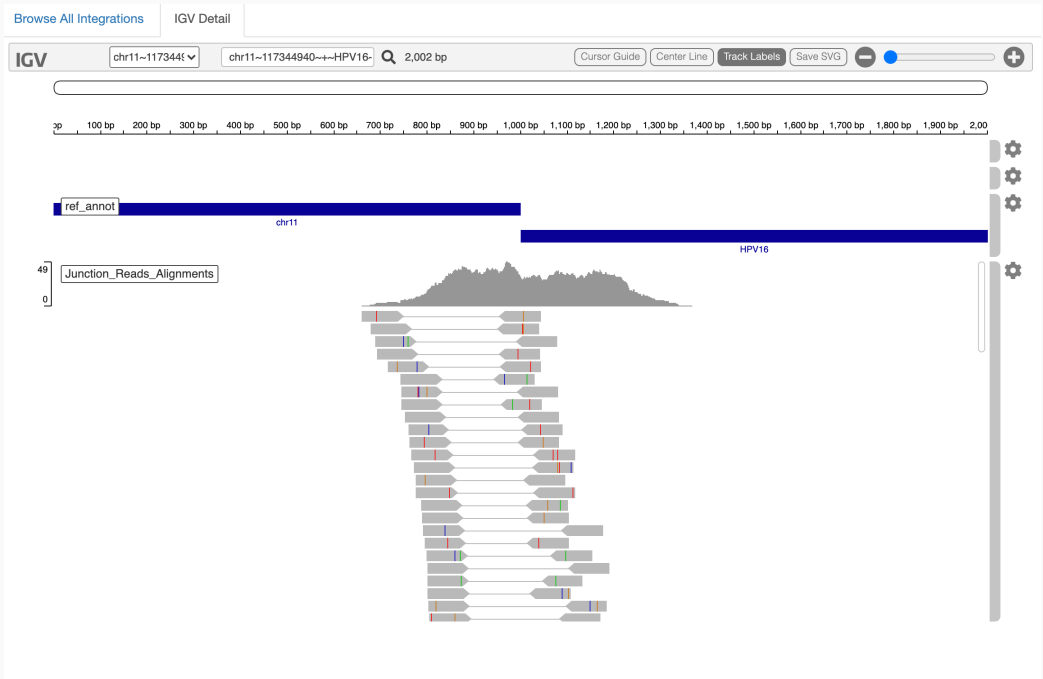
CTAT - Virus Integration Finder

chrA	coordA	orientA	chrB	coordB	orientB	prelim.prim
chr11	117344940	+	HPV16	1215	+	Split
chr18	73320805	+	HPV16	7409	-	Split
chr6	64935774	+	HPV16	4034	-	Split
HPV16	4029	+	chr3	937922	+	Split
HPV16	1215	-	chr11	108577428	+	Split
chr2	32018732	+	HPV16	1215	+	Split
chr1	218562612	+	HPV16	1215	+	Split

CTAT - Insertion Site Candidates



CTAT - Virus Insertion Viewer



CTAT - Virus Infection Evidence Viewer

- There may not be strong evidence for virus insertions
- Detects reads aligning to the target virus sequence

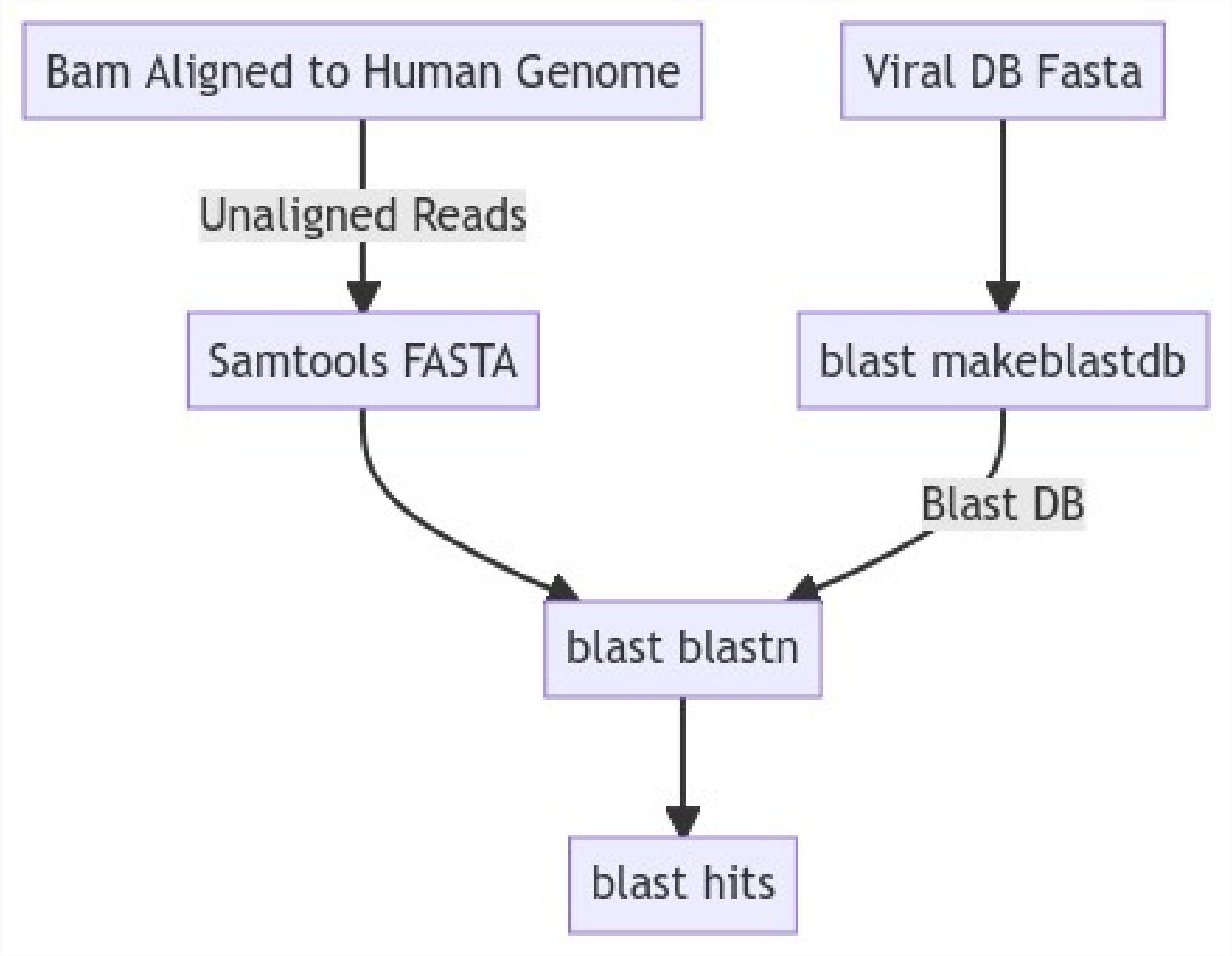
CTAT - Virus Infection Evidence Viewer

BLASTBox

BLASTBox

- Main goal was to avoid realigning thousands of high-coverage samples, by pulling out unaligned reads
- Found a similar study of TCGA data

BLASTBox - DAG



PRIMARY RESEARCH Open Access

Viral expression associated with gastrointestinal adenocarcinomas in TCGA high-throughput sequencing data

Daria Salyakina[1*] and Nicholas F Tsinoremas[1,2]

Viral expression associated with GIA Highlights (2013)

- Goal: Determine whether the presence of a virus is significantly associated with tumors
- 59% of screened gastrointestinal adenocarcinomas (GIA) were positive for at least one virus
- Used a similar approach, BAM => BLAST

BLASTBox Targets

- 1000 Genomes 30x WGS
- TCGA WGS
- dbGaP

Initial Questions to ask

- What parts of the sequence got integrated?
- What parts of the sequence got tossed out?
- Which one is potentially toxic/oncogenic