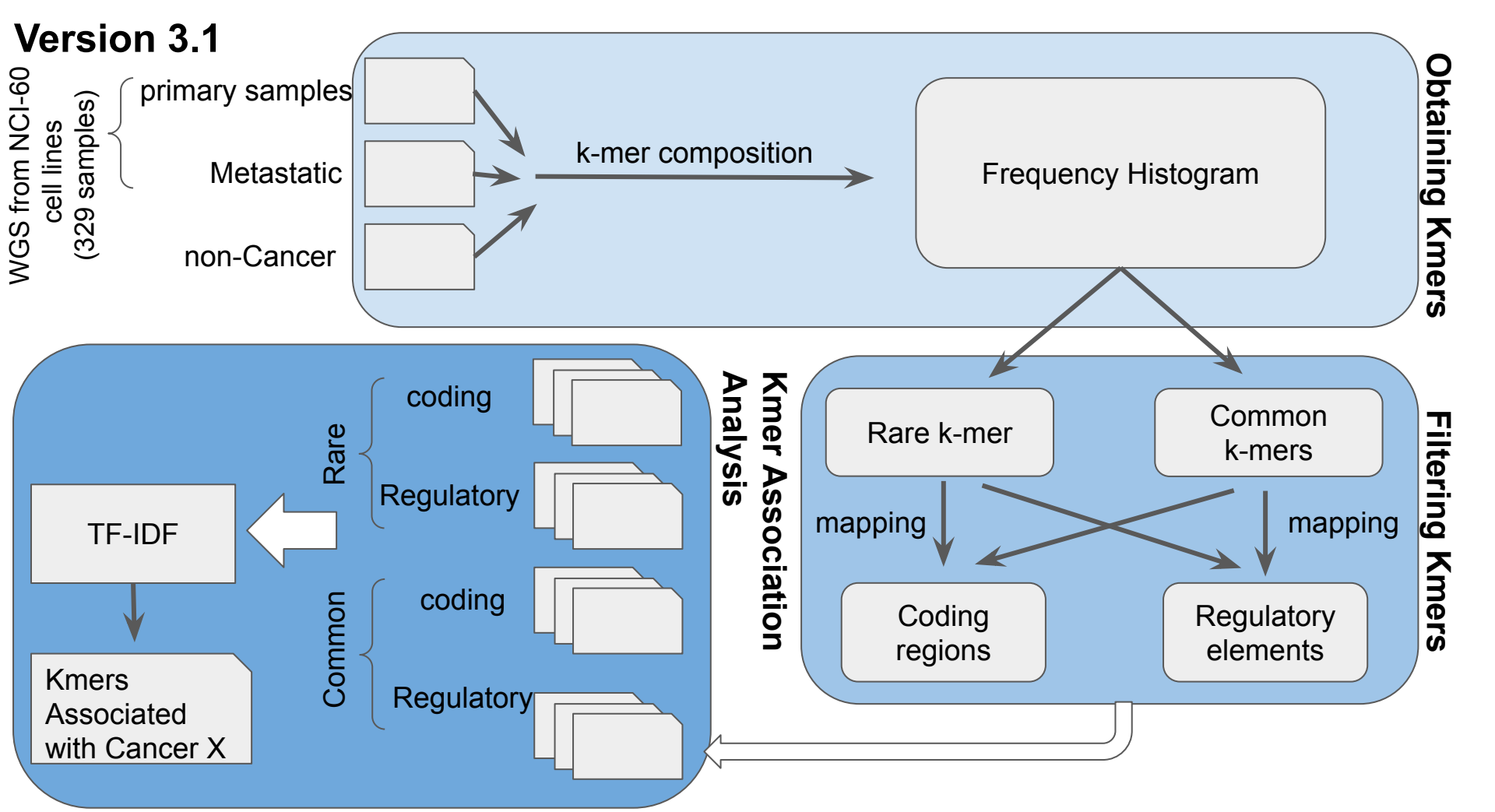
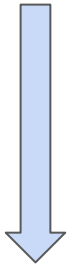
kvar

Github: https://github.com/collaborativebioinformatics/kvar


Version 3.1
WGS from NCI-60 cell lines (329 samples)
primary samples
Metastatic
non-Cancer
k-mer composition
Frequency Histogram
Obtaining Kmers
Rare k-mer
Common k-mers
mapping
mapping
Coding regions
Regulatory elements
Filtering Kmers
Kmer Association Analysis
Rare
coding
Regulatory
Common
coding
Regulatory
TF-IDF
Kmers Associated with Cancer X

Working in Parallel (Defining Inputs/Outputs)

Obtaining Kmers (Rohit)



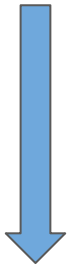
INPUT: Fastq file

EXAMPLE: <Standard Fastq File>

OUTPUT: Frequency (count) of each kmer

EXAMPLE: GCTGATCGAC, 2000 / AVG COV per sample
GCTAAACGAC, 0.004 (for example)

Filtering Kmers (Daniela)



INPUT: Frequency (count) of each kmer EXAMPLE: GCTGATCGAC, 0.0005
GCTAAACGAC, 0.004

OUTPUT: Filter Freq. AND 4x <Standard Bam files>

EXAMPLE: GCTGATCGAC, 0.0004 AND <Standard Bam file>

Kmer Association Analysis (Dreycey)

INPUT: Files with kmers and counts, seperated by region/etc

EXAMPLE: GCTGATCGAC, 0.0004

OUTPUT: A matrix or CSV with TF-IDF calc. AND flat file with disease-associated kmers

EXAMPLE: GCTGATCGAC, TF-IDF score

GOALS / RESPONSIBILITIES

Team Lead - Ben Busby

Writer(s) - Rohit, Danliela

SysAdmins - Dreycey, Quek

- Download primary data -- Quek
  - Find **QUICK** controls -- Quek (using primary vs metastatic)
- Implement kmer counting -- Rohit
  - First look into using NibSV - https://github.com/fritzsedlazeck/nibSV
  - If too cumbersome
    - meryl - https://github.com/marbl/meryl
    - Jellyfish - https://github.com/gmarcais/Jellyfish
    - KMC - https://github.com/refresh-bio/KMC
- Find **appropriate** (non-cancer) controls -- Daniela (Ben can help)!
- Set up “overlap analysis” -- Daniela
- Set up “tertiary analysis” -- Dreycey