Nascent Transcript Identification Using CHM13
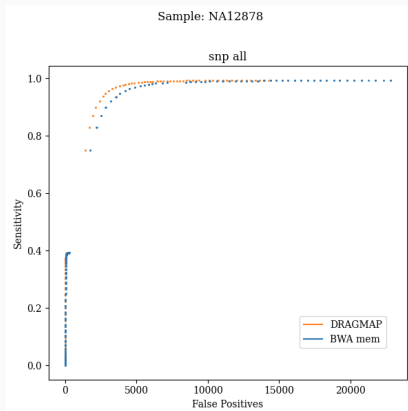
Edmund Miller

2021-10-25 Mon

Cell Ranger

- Wrapped up modules with test data
- Pending addition to nf-core/scRNA-seq

DRAGMAP

- Competes with BWA-MEM
- Pulled out of the DRAGEN-GATK pipeline
- First version of DRAGMAP has not been optimized for speed.(1.10x-1.74x longer than BWA)
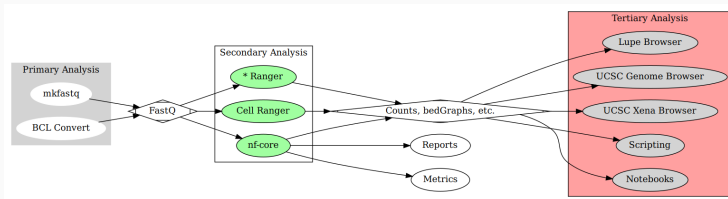
DRAGMAP

- Fewer false positives

Kedro

Where we left off

Kedro

- Kedro is an open-source Python framework for creating reproducible, maintainable and modular data science code.
- Data Catalog
- Jupyter integration
- Quick pipeline creation
- Versioning
- Testing

Kedro Data Catalog

```
bikes:
    type: pandas.CSVDataSet
    filepath: data/01_raw/bikes.csv

motorbikes:
    type: pandas.CSVDataSet
    filepath: s3://your_bucket/data/02_intermediate/con
    credentials: dev_s3
    versioned: True
    load_args:
        sep: ','
        skiprows: 5
        na_values: ['#NA', NA]
```

Usage

```
bikes = catalog.load("bikes")
motorbikes = catalog.load("motorbikes")
# pandas dataframe
bikes.head()
```
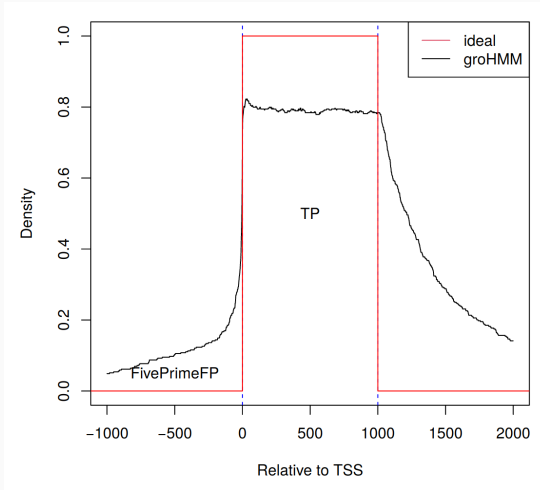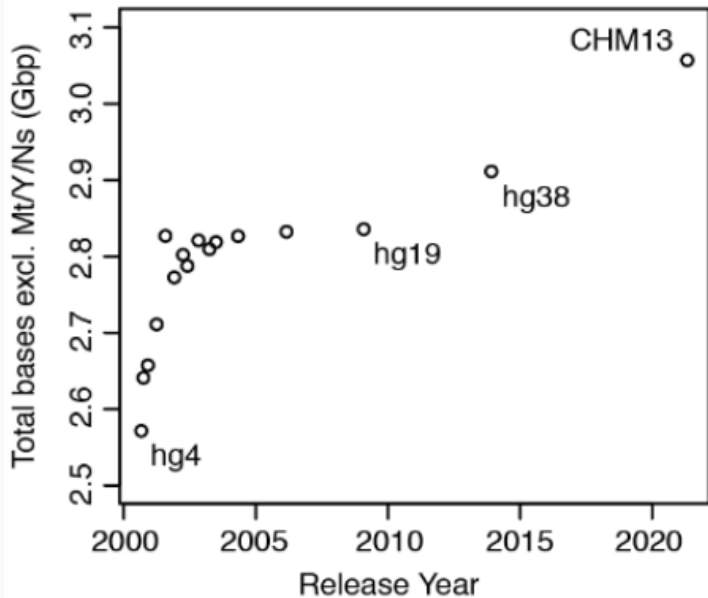
nascent

Updates

- groHMM works from any gtf(important for chm-13)
- Update to new Native dsl2 syntax
- Added DRAGMAP

Example Transcript Density Plot

CHM13

- Used PacBio's HiFi, and Nanopore's "ultra-long" reads to resolve complex forms of structural variation and gaps in GRCh38
- CHM13 Cell line a complete hydatidiform mole (CHM) cell line, "essentially haploid nature"