Nascent Transcript Identification Using CHM13

Edmund Miller

2022-03-02 Wed

nascent

Updated to nascent v1

- Updated Multiqc report with v1 metrics
    - Great spot checking for data mining
- Finished up homer transcript identification
- Started on dREG(Going to have to package it up)
    - Kinda makes me want to build our own transcript identification model

DeepVariant Approach

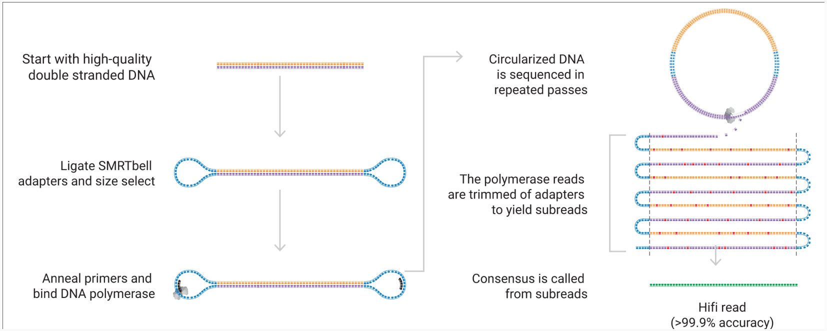
- Looking Through DeepVariant’s Eyes | DeepVariant Blog
- They’re trying to identify single bases
- We’re just looking at the bigger picture
- CNNs are great at this

CHM13 Refresher

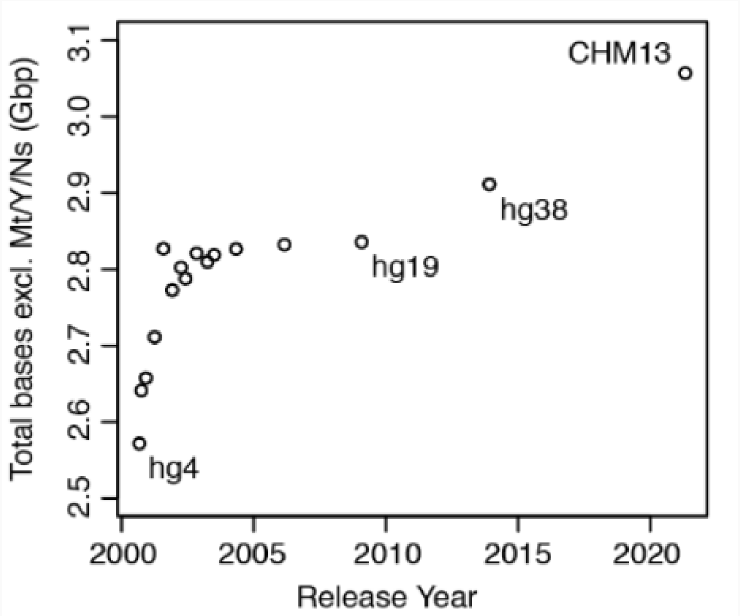
CHM13

- 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"

Pacbio HiFi Reads

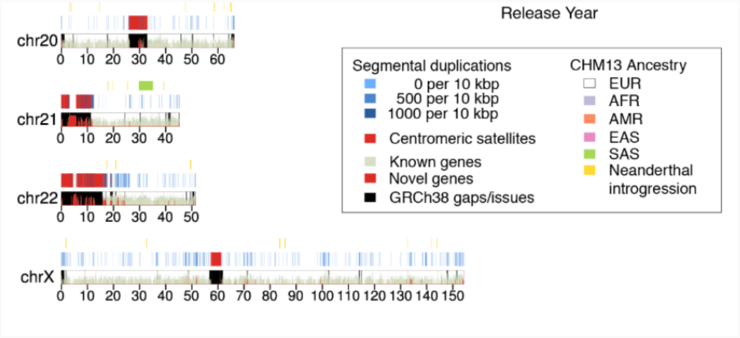

Start with high-quality double stranded DNA
Ligate SMRTbell adapters and size select
Anneal primers and bind DNA polymerase
Circularized DNA is sequenced in repeated passes
The polymerase reads are trimmed of adapters to yield subreads
Consensus is called from subreads
Hifi read (>99.9% accuracy)

CHM13



CHM13
hg38
hg19
hg4
Total bases excl. Mt/Y/Ns (Gbp)
2.5
2.6
2.7
2.8
2.9
3.0
3.1
2000
2005
2010
2015
2020
Release Year

CHM13



Release Year
chr20
0 10 20 30 40 50 60
chr21
0 10 20 30 40
chr22
0 10 20 30 40 50
chrX
0 10 20 30 40 50 60 70 80 90 100 110 120 130 140 150
Segmental duplications
0 per 10 kbp
500 per 10 kbp
1000 per 10 kbp
Centromeric satellites
Known genes
Novel genes
GRCh38 gaps/issues
CHM13 Ancestry
EUR
AFR
AMR
EAS
SAS
Neanderthal introgression

CHM13 Results

The intersection of Homer identified peaks and centromeres

- Near identical number of reads mapped(There's only 5% more to align to)
- Had to use hg19 centromeres for the intersection
- There were no hits
- Reviewed the multiqc and found there was an issue with the homer identification in CHM13(18% and 6% efficiency)
- Confirmed with bedtools intersect call to aligned reads for centromeres

Tertiary analysis Best practices

Some Goals

- Reproducibility
- Easy for others to read the code
- Easy for “exploratory analysis”
- Scales (Cluster or cloud submission)
    - Avoid `for` loops
- Could I teach this to someone in a Summer semester? (Applied Genomics)

Shiny New Datascience tools popping up

- Kedro
- Ploomber - Build data pipelines. FAST.
- The Julia Programming Language

Computational Biology is a beautiful mess

The first step of any bioinformatics project is to define a new file format, incompatible with all previous ones.

- Those all looked exciting. Until I needed to align some histones to hg38.
- Why waste a bunch of time rewriting scripts that already worked?

Code complexity aside

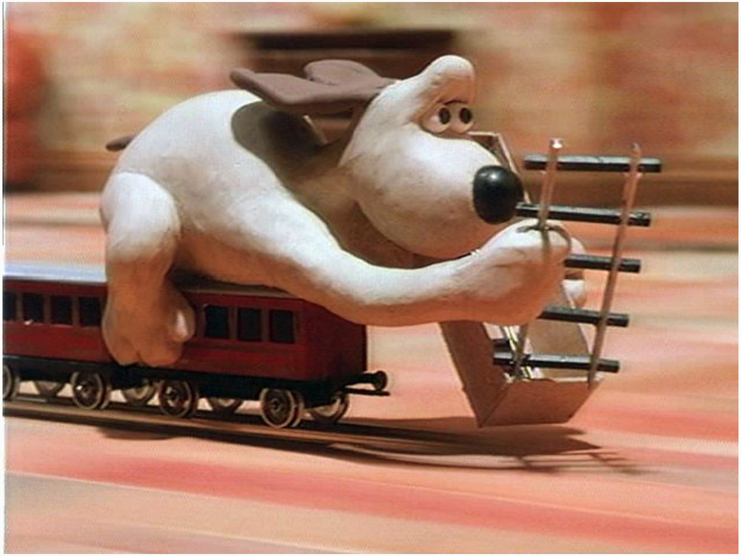
```
def bar():
    x = 1
    if x == 2:
        print("Success")

def foo():
    evens = [2, 4, 6, 8, 10]
    odds = [1, 3, 5, 7, 9]
    for x in evens:
        for y in odds:
            product = x * y
            if product % 2 == 0:
                print "Product result is even"
            if product % 5 == 0:
                print "Product is divisible by 5"
            if product % 3 == 0:
                print "Product is divisible by 3"
```

- McCabe's Cyclomatic Complexity

What if I just used nextflow?

- I was afraid it would be too "heavy"
    - Making container images
    - Boilerplate
- Somewhat Bioinformatics specific

Reality



Some mental shifts

- Just use conda environments for scripts
- Feedback of using the results directory for creating exploratory scripts
- Utilize nf-core modules
- Don't try to make features/job submission one size fits all(Leave that to nf-core)
    - Allowed me to cut down on a lot of the boilerplate
- Could a new grad student reproduce this and pick up where I left off?

How do we compare versions of things?

```
# move to the tag v1
git checkout v1
# store results in v1 directory
nextflow run . --outdir ./results/v1

# move to the tag v2
git checkout v2
# store results in v2 directory
nextflow run . --outdir ./results/v2
```

Some examples

- qbic-pipelines/rnadeseq: Differential gene expression analysis and pathway analysis of RNAseq data
- qbic-pipelines/bamtofastq
- Functional-Genomics-Lab/eRNA-complementarity