Finding functional elements of gene-less ecDNA: Origin of Replication, TATA boxes, and BLAST-ing k-mers

by Carlos, Sai, Clark, and Ashish

Our Team

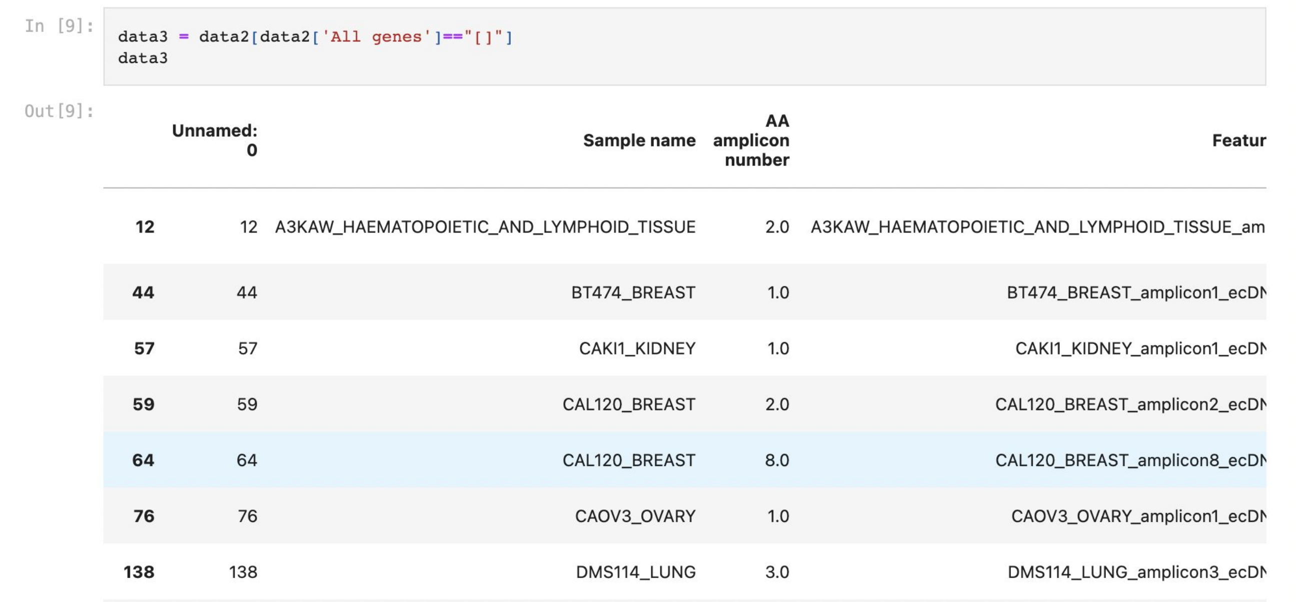
3rd year working at Han Lab

4th year - Bioinformatics working at Dr.Briggs Lab

4th year working at CMM

4th year working at Saier Lab

Data Pre-Processing



In [9]:

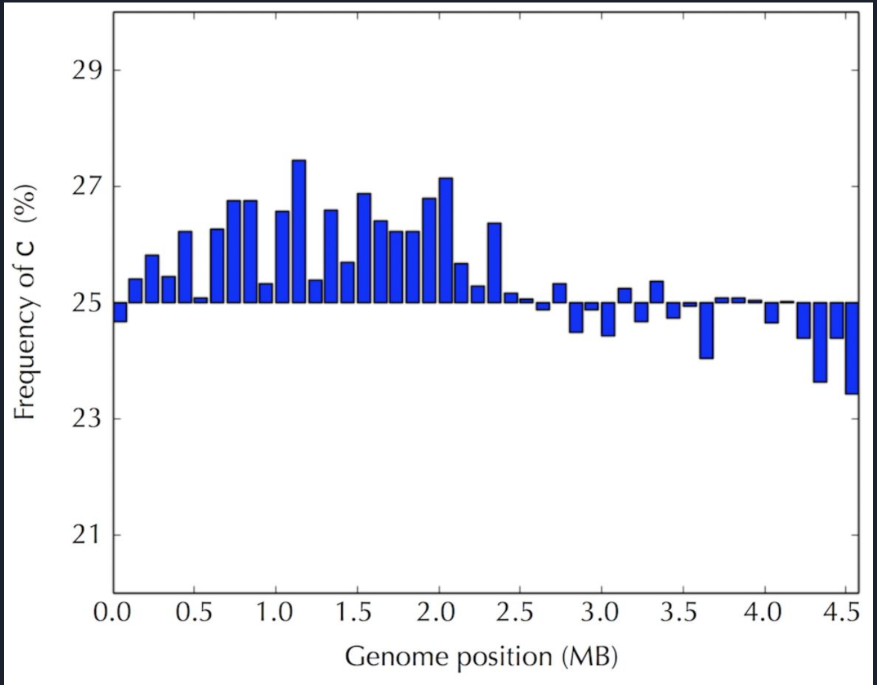
```
data3 = data2[data2['All genes']=="[]"]
data3
```

Out[9]:

	Unnamed: 0	Sample name	AA amplicon number	Featur
12	12	A3KAW_HAEMATOPOIETIC_AND_LYMPHOID_TISSUE	2.0	A3KAW_HAEMATOPOIETIC_AND_LYMPHOID_TISSUE_am
44	44	BT474_BREAST	1.0	BT474_BREAST_amplicon1_ecDN
57	57	CAKI1_KIDNEY	1.0	CAKI1_KIDNEY_amplicon1_ecDN
59	59	CAL120_BREAST	2.0	CAL120_BREAST_amplicon2_ecDN
64	64	CAL120_BREAST	8.0	CAL120_BREAST_amplicon8_ecDN
76	76	CAOV3_OVARY	1.0	CAOV3_OVARY_amplicon1_ecDN
138	138	DMS114_LUNG	3.0	DMS114_LUNG_amplicon3_ecDN

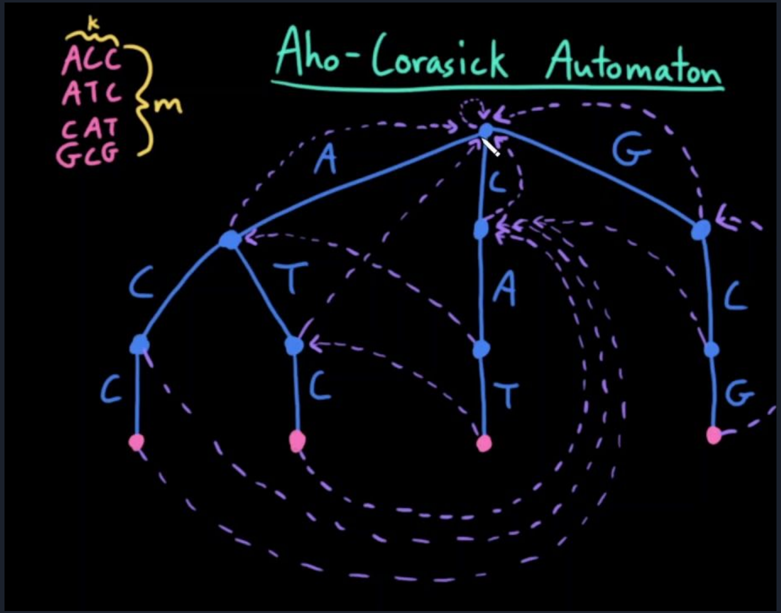
Tools: Python, Jupyter notebooks, Pandas

Methodology - Where are the origins of replication?



- Biological Principle of “Skew”
- Frequency of GC content can help find potential origins of replication
- Script that finds these origins was developed and used
- Tools: Jupyter notebook, python

Methodology - Part 2 (Finding TATA boxes)



- TATA boxes can have different patterns (Ex: TATATAAG, TATAA)

- Build the automaton and run a search on various chr3 txt files

- Compare actual hits against the E-vals for different strings
- Tools: Python, Jupyter notebooks

Methodology - Part 3 (Blasting Common k-mers)

Let’s say we have a Sequence

ACGTTGCATGTCGCATGATGCATGAGAGCT

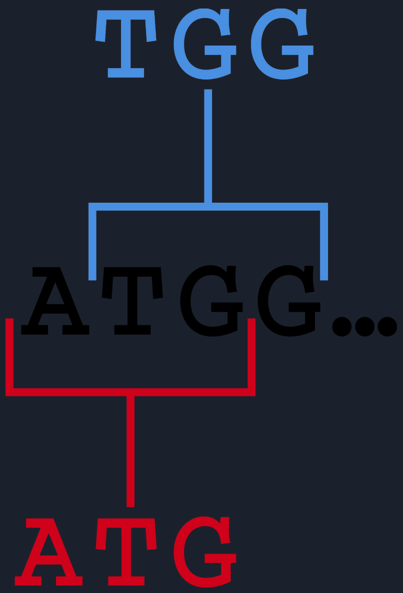
We choose K to find a kmer

Then we go from position 1 to 2 to 3 and find the K-mer from that position

Then we store the K mer and each of its occurrence

Finally we output the k mers with the most occurrences

Tools: Jupyter notebooks, Python, BLAST



Most frequent K mer , K = 3

ACGTTGCATGTCGCATGATGCATGAGAGCT

ACG - 1

Most frequent K mer , K = 3

ACGTTGCATGTCGCATGATGCATGAGAGCT

ACG - 1

CGT - 1

Most frequent K mer , K = 3

ACGTTGCATGTCGCATGATGCATGAGAGCT

ACG - 1

CGT - 1

GTT - 1

Most frequent K mer , K = 3

ACGTTGCATGTCGCATGATGCATGAGAGCT

ACG - 1

CGT - 1

GTT - 1

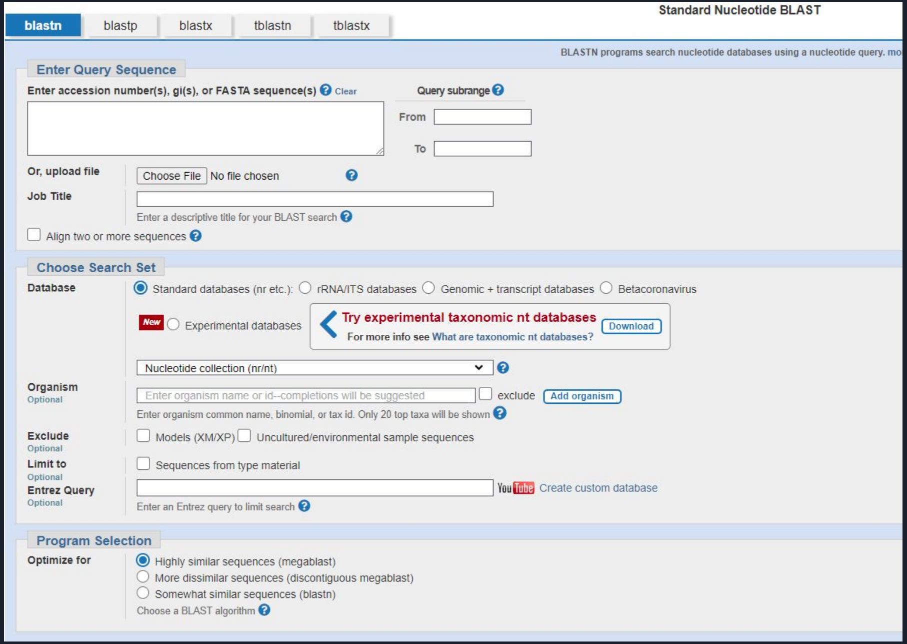
TTG - 1

ATG - 4

These is the most frequent 3-mer.

Cont

With this Kmers , we searched it in BLAST



blastn | blastp | blastx | tblastn | tblastx

Standard Nucleotide BLAST

BLASTN programs search nucleotide databases using a nucleotide query. mo

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) Clear

Query subrange

From

To

Or, upload file Choose File No file chosen

Job Title

Enter a descriptive title for your BLAST search

□ Align two or more sequences

Choose Search Set

Database ● Standard databases (nr etc.): ○ rRNA/ITS databases ○ Genomic + transcript databases ○ Betacoronavirus

New ○ Experimental databases

Try experimental taxonomic nt databases Download

For more info see What are taxonomic nt databases?

Nucleotide collection (nr/nt)

Organism Optional Enter organism name or id--completions will be suggested □ exclude Add organism

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown

Exclude Optional □ Models (XM/XP) □ Uncultured/environmental sample sequences

Limit to Optional □ Sequences from type material

Entrez Query Optional YouTube Create custom database

Enter an Entrez query to limit search

Program Selection

Optimize for

● Highly similar sequences (megablast)

○ More dissimilar sequences (discontiguous megablast)

○ Somewhat similar sequences (blastn)

Choose a BLAST algorithm

Results - Part 1

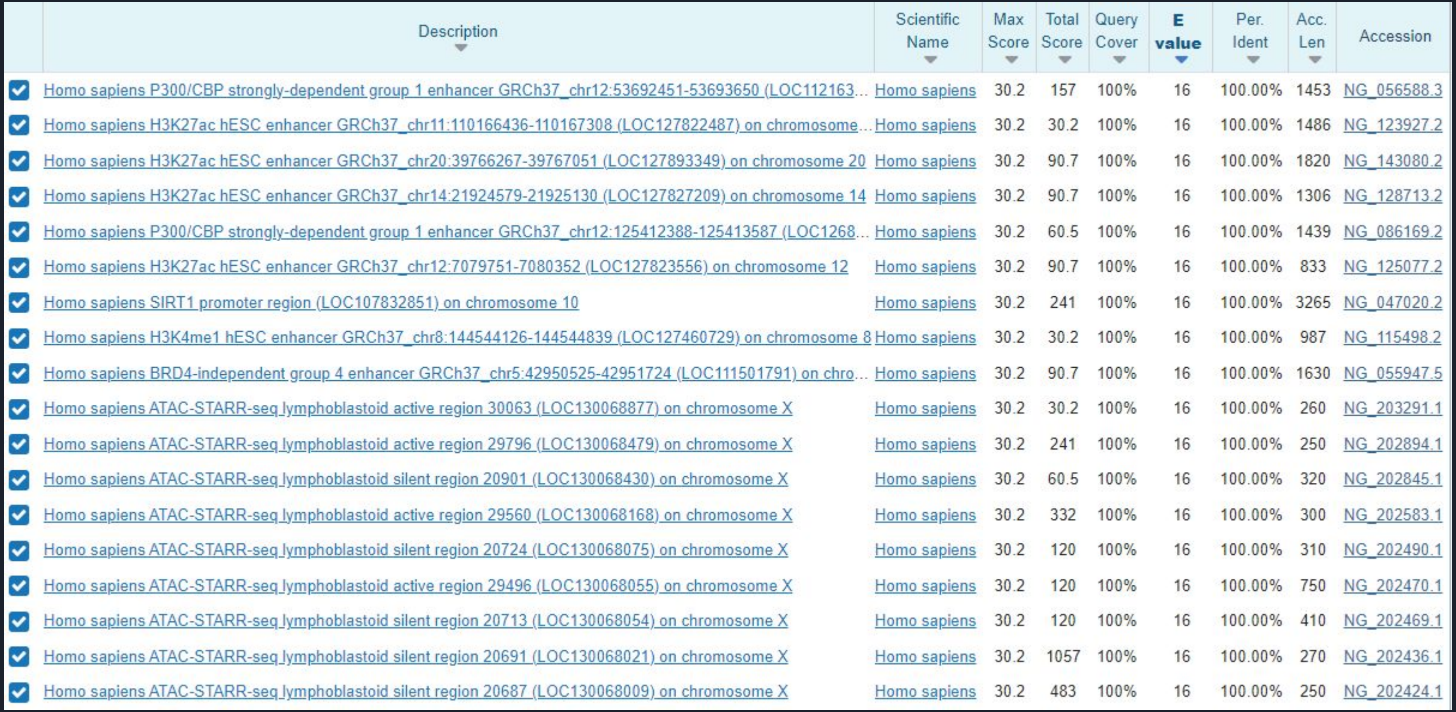
Sample	Origin at ecDNA	Origin at chromosome
A3KAW_HAEMATOPOIETIC_AND_LYMPHOID_TISSUE	75612	61013233
CAOV3_OVARY_amplicon1_ecDNA_1	13894	173376533
MFE280_ENDOMETRIUM_amplicon8_ecDNA_1	51129	122740309
MFE319_ENDOMETRIUM_amplicon1_ecDNA_1	101241	168954522
SKOV3_OVARY_amplicon1_ecDNA_1	148342	55704726

Results- Part 2 (Finding TATA boxes)

- TATATATA showed up *~30 to ~50* times more than we would expect in all the samples
- Other sequences like TATATAAG showed up at most 7x more than expectation
  - Most other TATA box variations did not show up significantly more than expectation
- These sequences are usually involved with DNA replication, why are they in these non-genic samples?
  - These are most likely remnants from the original linear DNA sequence
- How can they promote cancer?
  - ecDNA can share regulatory elements and “activate oncogenes on another ecDNA”

Results Summary - Part 3

1. Kmers seem to be of all A’s and all T’s
2. Emerged are enhancers or regulatory elements related to gene expression and control



	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
☑	Homo sapiens P300/CBP strongly-dependent group 1 enhancer GRCh37_chr12:53692451-53693650 (LOC112163...	Homo sapiens	30.2	157	100%	16	100.00%	1453	NG_056588.3
☑	Homo sapiens H3K27ac hESC enhancer GRCh37_chr11:110166436-110167308 (LOC127822487) on chromosome...	Homo sapiens	30.2	30.2	100%	16	100.00%	1486	NG_123927.2
☑	Homo sapiens H3K27ac hESC enhancer GRCh37_chr20:39766267-39767051 (LOC127893349) on chromosome 20	Homo sapiens	30.2	90.7	100%	16	100.00%	1820	NG_143080.2
☑	Homo sapiens H3K27ac hESC enhancer GRCh37_chr14:21924579-21925130 (LOC127827209) on chromosome 14	Homo sapiens	30.2	90.7	100%	16	100.00%	1306	NG_128713.2
☑	Homo sapiens P300/CBP strongly-dependent group 1 enhancer GRCh37_chr12:125412388-125413587 (LOC1268...	Homo sapiens	30.2	60.5	100%	16	100.00%	1439	NG_086169.2
☑	Homo sapiens H3K27ac hESC enhancer GRCh37_chr12:7079751-7080352 (LOC127823556) on chromosome 12	Homo sapiens	30.2	90.7	100%	16	100.00%	833	NG_125077.2
☑	Homo sapiens SIRT1 promoter region (LOC107832851) on chromosome 10	Homo sapiens	30.2	241	100%	16	100.00%	3265	NG_047020.2
☑	Homo sapiens H3K4me1 hESC enhancer GRCh37_chr8:144544126-144544839 (LOC127460729) on chromosome 8	Homo sapiens	30.2	30.2	100%	16	100.00%	987	NG_115498.2
☑	Homo sapiens BRD4-independent group 4 enhancer GRCh37_chr5:42950525-42951724 (LOC111501791) on chro...	Homo sapiens	30.2	90.7	100%	16	100.00%	1630	NG_055947.5
☑	Homo sapiens ATAC-STARR-seq lymphoblastoid active region 30063 (LOC130068877) on chromosome X	Homo sapiens	30.2	30.2	100%	16	100.00%	260	NG_203291.1
☑	Homo sapiens ATAC-STARR-seq lymphoblastoid active region 29796 (LOC130068479) on chromosome X	Homo sapiens	30.2	241	100%	16	100.00%	250	NG_202894.1
☑	Homo sapiens ATAC-STARR-seq lymphoblastoid silent region 20901 (LOC130068430) on chromosome X	Homo sapiens	30.2	60.5	100%	16	100.00%	320	NG_202845.1
☑	Homo sapiens ATAC-STARR-seq lymphoblastoid active region 29560 (LOC130068168) on chromosome X	Homo sapiens	30.2	332	100%	16	100.00%	300	NG_202583.1
☑	Homo sapiens ATAC-STARR-seq lymphoblastoid silent region 20724 (LOC130068075) on chromosome X	Homo sapiens	30.2	120	100%	16	100.00%	310	NG_202490.1
☑	Homo sapiens ATAC-STARR-seq lymphoblastoid active region 29496 (LOC130068055) on chromosome X	Homo sapiens	30.2	120	100%	16	100.00%	750	NG_202470.1
☑	Homo sapiens ATAC-STARR-seq lymphoblastoid silent region 20713 (LOC130068054) on chromosome X	Homo sapiens	30.2	120	100%	16	100.00%	410	NG_202469.1
☑	Homo sapiens ATAC-STARR-seq lymphoblastoid silent region 20691 (LOC130068021) on chromosome X	Homo sapiens	30.2	1057	100%	16	100.00%	270	NG_202436.1
☑	Homo sapiens ATAC-STARR-seq lymphoblastoid silent region 20687 (LOC130068009) on chromosome X	Homo sapiens	30.2	483	100%	16	100.00%	250	NG_202424.1

Future Directions

- Go deeper in understanding the role of TATA boxes in ecDNA and their impact on cancer promotion.
- Exploring alternative methods of aligning functional elements for sample analysis instead of blasting k-mers.
- Knowing origin of replication, can we target these through novel therapeutics to mitigate replication?
- How can we effectively target these ecDNAs, by knowing the nature of the elements that found through BLAST,?
- How to combine the findings from the analysis of TATA boxes, frequent words, origin of replication to generate hypotheses about the potential functions of ecDNA in biological pathways?

Thank you!

Works Cited and Code:

- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5881399/
- https://www.cancer.gov/ccg/blog/2022/interview-ecdna
- https://github.com/cfg00/BENG182-Project
- http://bioinformaticsalgorithms.org
- https://pandas.pydata.org/docs/
- Data Structures - Niema Moshiri and Liz Izhikevich - 2016