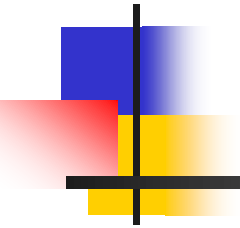
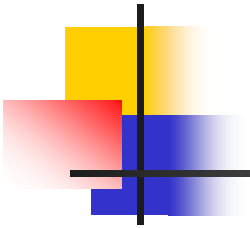
Pattern Search in Noncoding RNA



Chen-Hanson Ting

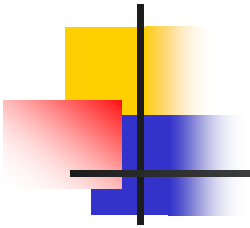
SVFIG

August 24, 2019



Summary

- Bioinformatics
- Long noncoding RNA
- RNA analysis
- Computer farm
- RNA analysis results



Old Bioinformatics

DNA chromosomes

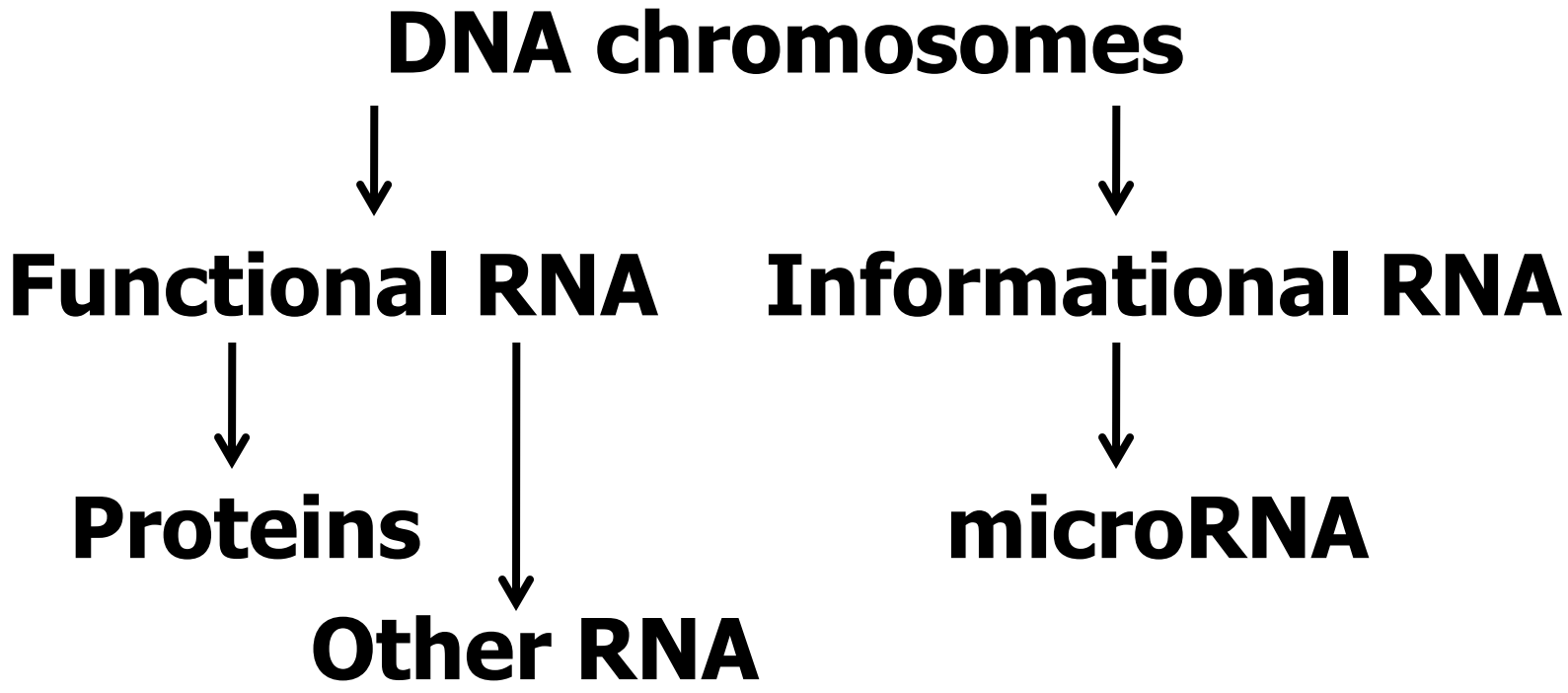
↓

Coding DNA produce coding RNA

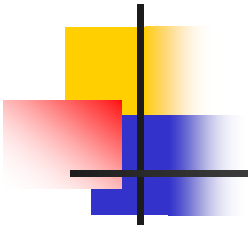
↓

Coding RNA produce proteins

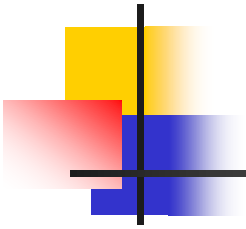
New Bioinformatics



Old RNA Bioinformatics



- mRNAs, messenger RNA to produce proteins
- tRNA, transfer RNA
- rRNA, ribosomal RNA
- lncRNA, long non-coding RNA, 200 bp or more
- miRNA, microRNA 18-22 bp
- siRNA, snoRNA, piRNA, srRNA



New RNA Bioinformatics

- **Functional RNA**
  - **mRNAs, to produce proteins**
  - **tRNA, transfer RNA**
  - **rRNA, ribosomal RNA**
  - **Other functional RNA**
- **Informational RNA**
  - **Pearls, microRNA**
  - **Necklaces, microRNA clusters**

lncRNA Databases

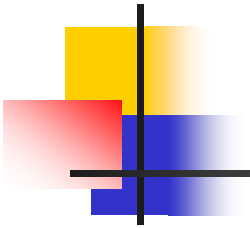
- LNCipedia 5.2 , 192,690,141 bp
- GRCh38_ncrna, 65,790,873 bp
- GRCh8_cdna, 316,791,371 bp

Long Noncoding RNA

- **LNCipedia 5.2 - Aug 2, 2018**
  - **127,802 transcripts**
  - **56,946 genes**
  - **192,690,141 bp**
- Ghent University - VIB, Life Sciences Research Institute in Flanders, Belgium

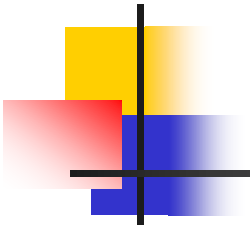
Long Noncoding RNA

- **GRCh38.p12, EMBL-EBI**
  - **GRCh38_ncrna, 65,790,873 bp**
  - **GRCh8_cdna, 316,791,371 bp**
- European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Cambridge, United Kingdom



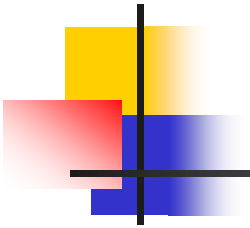
RNA Analysis

- RNA file is split into a pure data file and an index file.
- All repeated 20 bp patterns are extracted from the data file.
- Patterns are sorted and packed to variable-length pearls.
- Index file is inserted back to identify necklaces.



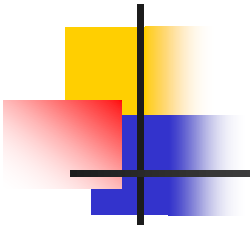
RNA Analysis

- Identifying repeated 20 bp patterns is easy to say than done.
- 4096 threads are produced pointing to one of the 6 bp patterns.
- Repeated patterns beginning with this 6 bp pattern are exhaustively searched in each thread, and written to a text file.



RNA Analysis

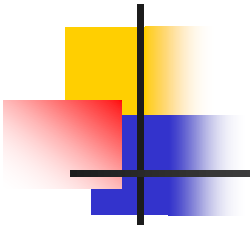
- 4096 files are sorted and redundant entries are removed.
- Entries within a 10M bp block are extracted from each file and combined into a chunk file.
- Index file is merged into chunk files to identify pearls and necklaces.



RNA Analysis

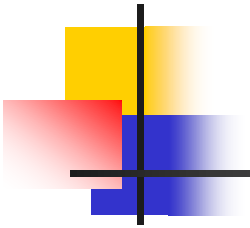
- For the largest 300 Mbp cDNA file, it took 15 minutes to process a thread, 1000 hours for 4K threads, and 40 days for human cDNA.
- I set up a computer farm with 6 PCs, and finished the analysis in 5 days.

Computer Farm



RNA Analysis

- Forth is used to do the heavy lifting, thread processing.
- Python is used to do light weight sorting and packing.
- Excel is used to do data analysis and display.

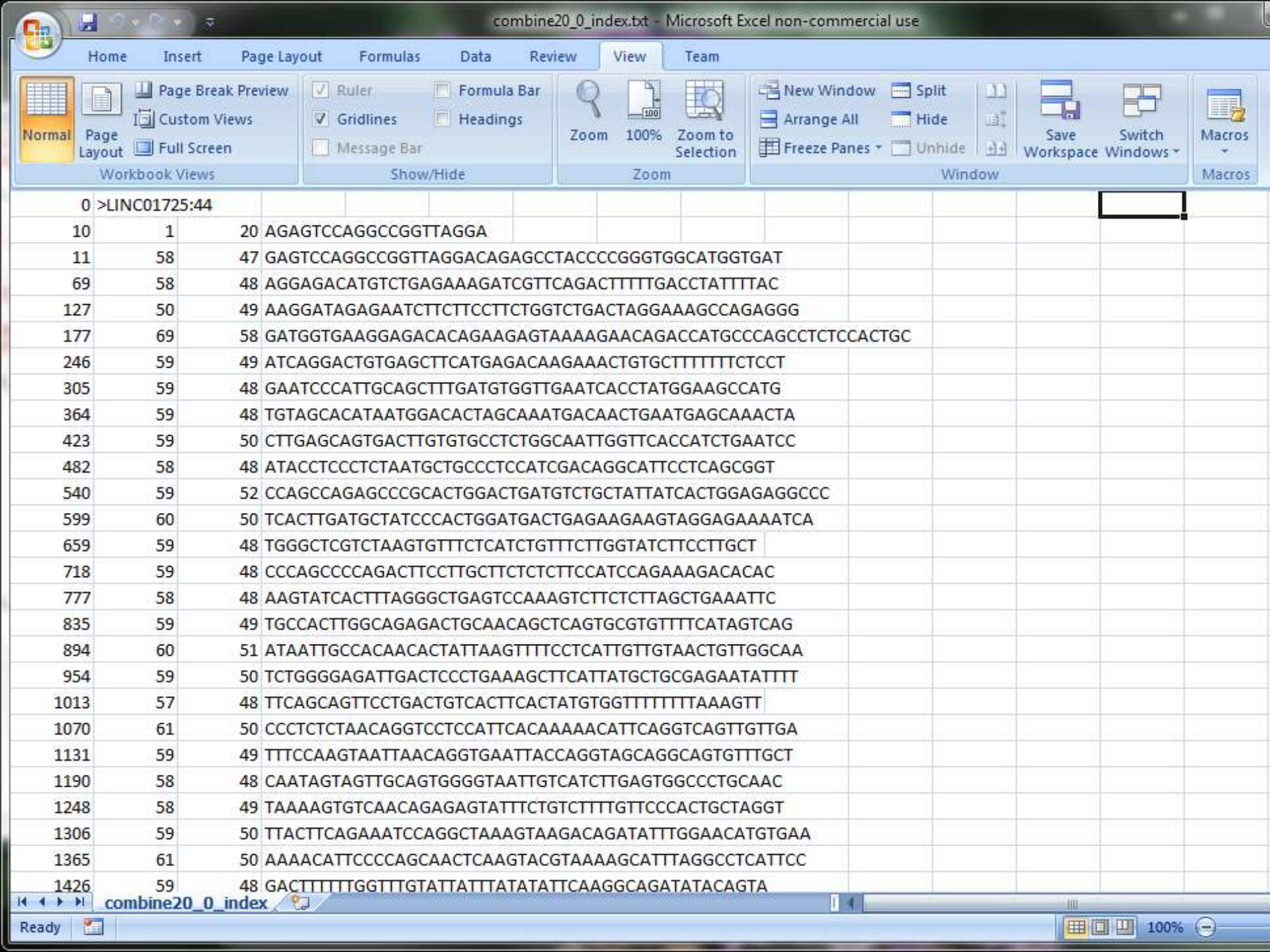


RNA Analysis

- I was pleasantly surprised at that Python was able to open 4096 thread files simultaneously and extract chunks of data for sorting and packing.

RNA Analysis Results

- **LNCipedia_5_2, however, produced nothing but pearls and necklaces.**
- **GRCh38_ncrna and GRCh38_cdna yielded occasional pearls and necklaces.**
- **GRCh38_ncrna is a small subset of LNCipedia_5_2.**



0	>LINC01725:44		
10	1	20	AGAGTCCAGGCCGGTTAGGA
11	58	47	GAGTCCAGGCCGGTTAGGACAGAGCCTACCCCGGGTGGCATGGTGAT
69	58	48	AGGAGACATGTCTGAGAAAGATCGTTCAGACTTTTTGACCTATTTTAC
127	50	49	AAGGATAGAGAATCTTCTTCCTTCTGGTCTGACTAGGAAAGCCAGAGGG
177	69	58	GATGGTGAAGGAGACACAGAAGAGTAAAAGAACAGACCATGCCCAGCCTCTCCACTGC
246	59	49	ATCAGGACTGTGAGCTTCATGAGACAAGAAACTGTGCTTTTTTTCTCCT
305	59	48	GAATCCCATTGCAGCTTTGATGTGGTTGAATCACCTATGGAAGCCATG
364	59	48	TGTAGCACATAATGGACACTAGCAAATGACAACTGAATGAGCAAACTA
423	59	50	CTTGAGCAGTGACTTGTGTGCCTCTGGCAATTGGTTCACCATCTGAATCC
482	58	48	ATACCTCCCTCTAATGCTGCCCTCCATCGACAGGCATTCCTCAGCGGT
540	59	52	CCAGCCAGAGCCCGCACTGGACTGATGTCTGCTATTATCACTGGAGAGGCCC
599	60	50	TCACTTGATGCTATCCCACTGGATGACTGAGAAGAAGTAGGAGAAAATCA
659	59	48	TGGGCTCGTCTAAGTGTTTCTCATCTGTTTCTTGGTATCTTCCTTGCT
718	59	48	CCCAGCCCCAGACTTCCTTGCTTCTCTCTTCCATCCAGAAAGACACAC
777	58	48	AAGTATCACTTTAGGGCTGAGTCCAAAGTCTTCTCTTAGCTGAAATTC
835	59	49	TGCCACTTGGCAGAGACTGCAACAGCTCAGTGCGTGTTTTCATAGTCAG
894	60	51	ATAATTGCCACAACACTATTAAGTTTTCCTCATTGTTGTAACTGTTGGCAA
954	59	50	TCTGGGGAGATTGACTCCCTGAAAGCTTCATTATGCTGCGAGAATATTTT
1013	57	48	TTCAGCAGTTCCTGACTGTCACTTCACTATGTGGTTTTTTTAAAGTT
1070	61	50	CCCTCTCTAACAGGTCCTCCATTCACAAAAACATTCAGGTCAGTTGTTGA
1131	59	49	TTTCCAAGTAATTAACAGGTGAATTACCAGGTAGCAGGCAGTGTTTGCT
1190	58	48	CAATAGTAGTTGCAGTGGGGTAATTGTCATCTTGAGTGGCCCTGCAAC
1248	58	49	TAAAAGTGTCAACAGAGAGTATTTCTGTCTTTTGTTCCCACTGCTAGGT
1306	59	50	TTACTTCAGAAATCCAGGCTAAAGTAAGACAGATATTTGGAACATGTGAA
1365	61	50	AAAACATTCCCCAGCAACTCAAGTACGTAAAAGCATTTAGGCCTCATTCC
1426	59	48	GACTTTTTGGTTTGTATTATTTATATATTCAAGGCAGATATACAGTA

RNA Analysis Results

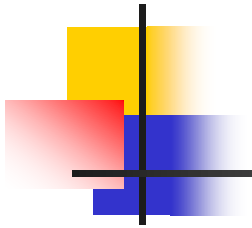
- **Scan combine20_0_index.txt file**

RNA Analysis Results

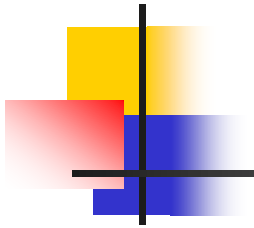
- Lots of consecutive ~60 bp pearls, and hence necklaces.
- Lots of necklaces with ~30 bp pearls.
- Lots of very long RNA stretches, thousands of bp.

RNA Analysis Results

- Are these ~60 bp pearls real?
- The necklaces are too good to be true.
- These wall-to-wall necklaces are not present in GRCh38_ncrna file, nor in GRC38_cdna file. Why?



Questions?



Thank You!