Code in Long Noncoding RNA

Chen-Hanson Ting

SVFIG

October 26, 2019

Summary

- Python is Forth in C
- Problems in lncRNA searches

Python is Forth in C

- Python interpreter IDLE feels like Forth, and syntax is like C.
- Python lists handle multidimensional data, mixing numbers and strings, with ease.
- File system is flexible.

Python Code

- Forth code is easy to convert to Python code.
- It is easy to write Python interpreter code in files and compile files.
- Python string and list methods are easy to invoke.
- No new functions are necessary.

Python Code

- Python can handle huge genome data in lists and strings.
- Python opens more than 4000 files simultaneously.
- 32-bit Python shows memory errors. 64-bir python does not.
- Excel is used to display data.

Python Code I

- FORMAT: names/nucleotides
- TRIM: redundancy removal
- SPLIT: data/index files
- REPEATS: find all Repeats
- SORT: sort Repeats
- PACK: pack Repeats to Pearls

Python Code II

- SELECT: remove redundant Pearls
- REDUCE: from selected Pearls reduce index file
- CLEANUP: from reduced index file cleanup format file

Python Code III

- SPLIT: split reduced format file into data / index files
- REPEATS: find all Repeats
- SORT: sort Repeats
- PACK: pack Repeats to Pearls

Python Code III

- SELECT: remove redundant Pearls
- REDUCE: from selected Pearls reduce index file
- CLEANUP: from reduced index file cleanup format file

Biggest Problem

- Duplicated lncRNA produce huge amount of bogus Repeats which invalidate matching patterns.
- It is difficult to identify and remove these nucleotide sequences.

RNA Pattern Search

Steps in lncRNA analysis:

- **Find all 20 nt repeated patterns as Repeats.**
- **Consolidate adjacent Repeats for form Pearls.**
- **Find clusters of Pearls as Necklaces.**

lncRNA Databases

Name	Size(KB)	RNA
GRCh38_ncrna.fa	64,249	67,419
LNCipedia_5_2.fasta	196,560	102,369
NONCODEv5.fa	284,922	165,911
GRCh38_cdna.fa	361,405	139,155
lncRNA_lncbook.fa	400,768	208,848

Redundancy in RNA

- There are many very long Pearls caused by redundancy in lncRNA
- Two long stretches of n nucleotides would give 2n bogus Repeats.
- Bogus Repeats cause bogus Pearls.

Redundancy Removal

- Eliminate identical lncRNA in database.
- Eliminate redundant long Pearls.
- Backtrack redundant long Pearls to respect lncRNA and remove them from database.
- Repeat pattern analysis.

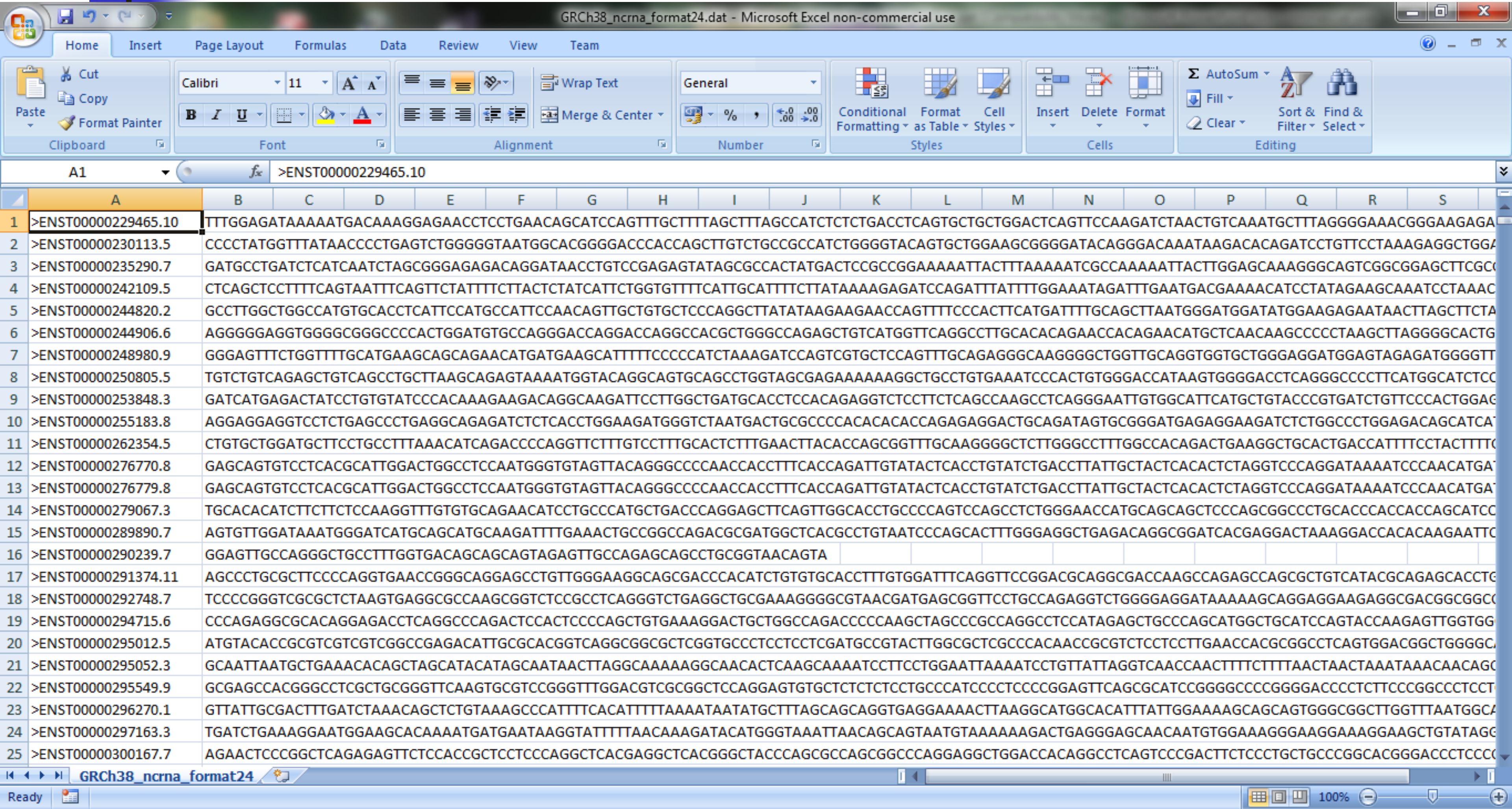
Redundancy Removal

Fasta File	lncRNA	Pearls	Select Pearls	Select lncRNA	Final Pearls
GRCh38_ncrna	47489	83626	76926	10563	35614
LNCipedia	126876	327561	294429	28927	110326
NONCODE	170767	574827	527627	51516	67019
GRCh38_cdna	177455	370062	267700	22945	109402
lncbook	268639	847583	776065	81465	320500

Redundancy Removal

- **Formatted lncRNA database.**
- **Very long Pearls**
- **Pearls file after redundancy removal.**

Formatted lncRNA Data



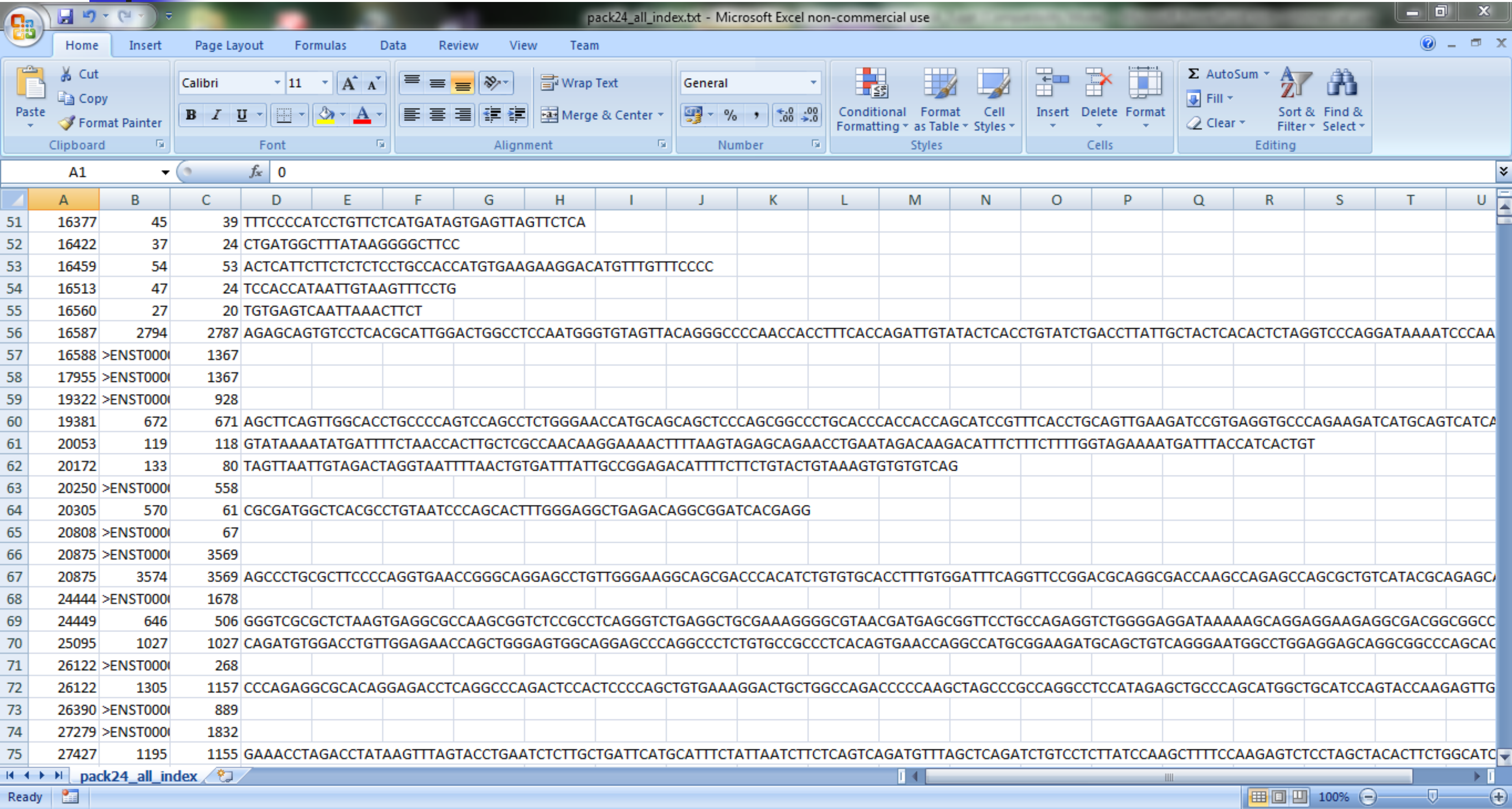
GRCh38_ncrna_format24.dat - Microsoft Excel non-commercial use

A1 | >ENST00000229465.10

	A	B
1	>ENST00000229465.10	TTTGGAGATAAAAATGACAAAGGAGAACCTCCTGAACAGCATCCAGTTTGCTTTTAGCTTTAGCCATCTCTCTGACCTCAGTGCTGCTGGACTCAGTTCCAAGATCTAACTGTCAAATGCTTTAGGGAAACGGAAGAGA
2	>ENST00000230113.5	CCCCTATGGTTTATAACCCCTGAGTCTGGGGGTAATGGCACGGGGACCCACCAGCTTGTCTGCCGCCATCTGGGGTACAGTGCTGGAAGCGGGGATACAGGGACAAATAAGACACAGATCCTGTTCCTAAAGAGGCTGGA
3	>ENST00000235290.7	GATGCCTGATCTCATCAATCTAGCGGGAGAGACAGGATAACCTGTCCGAGAGTATAGCGCCACTATGACTCCGCCGGAAAAATTACTTTAAAAATCGCCAAAAATTACTTGGAGCAAAGGGCAGTCGGCGGAGCTTCGC
4	>ENST00000242109.5	CTCAGCTCCTTTTCAGTAATTTCAGTTCTATTTTCTTACTCTATCATTCTGGTGTTTTCATTGCATTTTCTTATAAAAGAGATCCAGATTTATTTTGGAAATAGATTTGAATGACGAAAACATCCTATAGAAGCAAATCCTAAAC
5	>ENST00000244820.2	GCCTTGGCTGGCCATGTGCACCTCATTCCATGCCATTCCAACAGTTGCTGTGCTCCCAGGCTTATATAAGAAGAACCAGTTTTCCCACTTCATGATTTTGCAGCTTAATGGGATGGATATGGAAGAGAATAACTTAGCTTCTA
6	>ENST00000244906.6	AGGGGGAGGTGGGGCGGGCCCCACTGGATGTGCCAGGGACCAGGACCAGGCCACGCTGGGCCAGAGCTGTCATGGTTCAGGCCTTGCACACAGAACCACAGAACATGCTCAACAAGCCCCCTAAGCTTAGGGGCACTG
7	>ENST00000248980.9	GGGAGTTTCTGGTTTTGCATGAAGCAGCAGAACATGATGAAGCATTTTTCCCCCATCTAAAGATCCAGTCGTGCTCCAGTTTGCAGAGGGCAAGGGGCTGGTTGCAGGTGGTGCTGGGAGGATGGAGTAGAGATGGGGTT
8	>ENST00000250805.5	TGTCTGTCAGAGCTGTCAGCCTGCTTAAGCAGAGTAAAATGGTACAGGCAGTGCAGCCTGGTAGCGAGAAAAAAGGCTGCCTGTGAAATCCCACTGTGGGACCATAAGTGGGGACCTCAGGGCCCCTTCATGGCATCTCC
9	>ENST00000253848.3	GATCATGAGACTATCCTGTGTATCCCACAAAGAAGACAGGCAAGATTCCTTGGCTGATGCACCTCCACAGAGGTCTCCTTCTCAGCCAAGCCTCAGGGAATTGTGGCATTCATGCTGTACCCGTGATCTGTTCCCACTGGAG
10	>ENST00000255183.8	AGGAGGAGGTCCTCTGAGCCCTGAGGCAGAGATCTCTCACCTGGAAGATGGGTCTAATGACTGCGCCCCACACACACCAGAGAGGACTGCAGATAGTGCGGGATGAGAGGAAGATCTCTGGCCCTGGAGACAGCATCA
11	>ENST00000262354.5	CTGTGCTGGATGCTTCCTGCCTTTAAACATCAGACCCCAGGTTCTTTGTCCTTTGCACTCTTTGAACTTACACCAGCGGTTTGCAAGGGGCTCTTGGGCCTTTGGCCACAGACTGAAGGCTGCACTGACCATTTTCCTACTTTT
12	>ENST00000276770.8	GAGCAGTGTCCTCACGCATTGGACTGGCCTCCAATGGGTGTAGTTACAGGGCCCCAACCACCTTTCACCAGATTGTATACTCACCTGTATCTGACCTTATTGCTACTCACACTCTAGGTCCCAGGATAAAATCCCAACATGA
13	>ENST00000276779.8	GAGCAGTGTCCTCACGCATTGGACTGGCCTCCAATGGGTGTAGTTACAGGGCCCCAACCACCTTTCACCAGATTGTATACTCACCTGTATCTGACCTTATTGCTACTCACACTCTAGGTCCCAGGATAAAATCCCAACATGA
14	>ENST00000279067.3	TGCACACATCTTCTTCTCCAAGGTTTGTGTGCAGAACATCCTGCCCATGCTGACCCAGGAGCTTCAGTTGGCACCTGCCCCAGTCCAGCCTCTGGGAACCATGCAGCAGCTCCCAGCGGCCCTGCACCCACCACCAGCATCC
15	>ENST00000289890.7	AGTGTTGGATAAATGGGATCATGCAGCATGCAAGATTTTGAAACTGCCGGCCAGACGCGATGGCTCACGCCTGTAATCCCAGCACTTTGGGAGGCTGAGACAGGCGGATCACGAGGACTAAAGGACCACACAAGAATTC
16	>ENST00000290239.7	GGAGTTGCCAGGGCTGCCTTTGGTGACAGCAGCAGTAGAGTTGCCAGAGCAGCCTGCGGTAACAGTA
17	>ENST00000291374.11	AGCCCTGCGCTTCCCCAGGTGAACCGGGCAGGAGCCTGTTGGGAAGGCAGCGACCCACATCTGTGTGCACCTTTGTGGATTTCAGGTTCCGGACGCAGGCGACCAAGCCAGAGCCAGCGCTGTCATACGCAGAGCACCTG
18	>ENST00000292748.7	TCCCCGGGTCGCGCTCTAAGTGAGGCGCCAAGCGGTCTCCGCCTCAGGGTCTGAGGCTGCGAAAGGGGCGTAACGATGAGCGGTTCCTGCCAGAGGTCTGGGGAGGATAAAAAGCAGGAGGAAGAGGCGACGGCGGC
19	>ENST00000294715.6	CCCAGAGGCGCACAGGAGACCTCAGGCCCAGACTCCACTCCCCAGCTGTGAAAGGACTGCTGGCCAGACCCCCAAGCTAGCCCGCCAGGCCTCCATAGAGCTGCCCAGCATGGCTGCATCCAGTACCAAGAGTTGGTGG
20	>ENST00000295012.5	ATGTACACCGCGTCGTCGTCGGCCGAGACATTGCGCACGGTCAGGCGGCGCTCGGTGCCCTCCTCCTCGATGCCGTACTTGGCGCTCGCCCACAACCGCGTCTCCTCCTTGAACCACGCGGCCTCAGTGGACGGCTGGGGC
21	>ENST00000295052.3	GCAATTAATGCTGAAACACAGCTAGCATACATAGCAATAACTTAGGCAAAAAGGCAACACTCAAGCAAAATCCTTCCTGGAATTAAAATCCTGTTATTAGGTCAACCAACTTTTCTTTTAACTAACTAAATAAACAACAGC
22	>ENST00000295549.9	GCGAGCCACGGGCCTCGCTGCGGGTTCAAGTGCGTCCGGGTTTGGACGTCGCGGCTCCAGGAGTGTGCTCTCTCTCCTGCCCATCCCCTCCCCGGAGTTCAGCGCATCCGGGGCCCCGGGGACCCCTCTTCCCGGCCCTCCT
23	>ENST00000296270.1	GTTATTGCGACTTTGATCTAAACAGCTCTGTAAAGCCCATTTTCACATTTTTAAAATAATATGCTTTAGCAGCAGGTGAGGAAAACTTAAGGCATGGCACATTTATTGGAAAAGCAGCAGTGGGCGGCTTGGTTTAATGGCA
24	>ENST00000297163.3	TGATCTGAAAGGAATGGAAGCACAAAATGATGAATAAGGTATTTTTAACAAAGATACATGGGTAAATTAACAGCAGTAATGTAAAAAAGACTGAGGGAGCAACAATGTGGAAAGGGAAGGAAAGGAAGCTGTATAGG
25	>ENST00000300167.7	AGAACTCCCGGCTCAGAGAGTTCTCCACCGCTCCTCCCAGGCTCACGAGGCTCACGGGCTACCCAGCGCCAGCGGCCCAGGAGGCTGGACCACAGGCCTCAGTCCCGACTTCTCCCTGCTGCCCGGCACGGGACCCTCCC

GRCh38_ncrna_format24

Very long Pearls

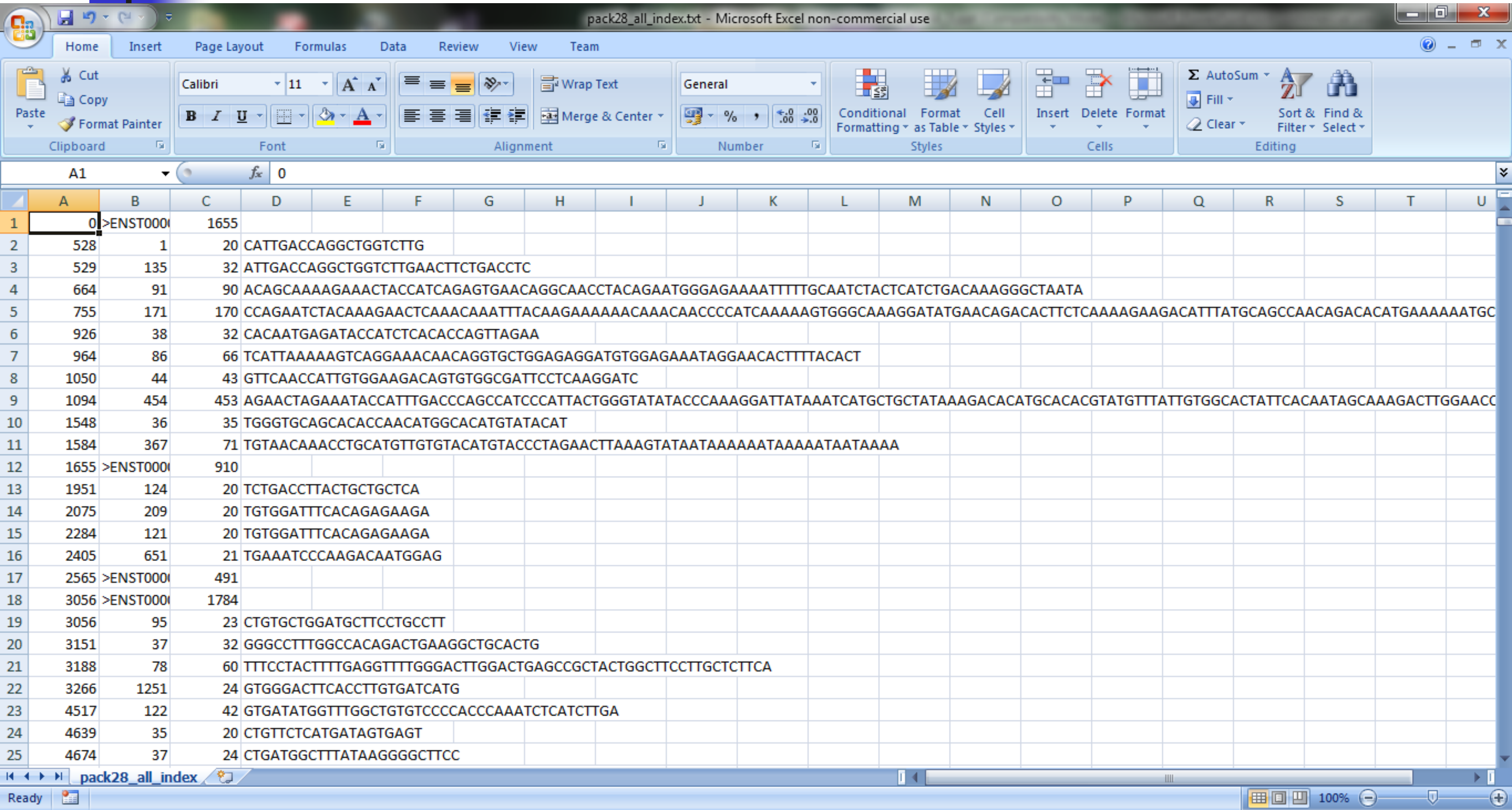


A1 fx 0

	A	B	C	D
51	16377	45	39	TTTCCCCATCCTGTTCTCATGATAGTGAGTTAGTTCTCA
52	16422	37	24	CTGATGGCTTTATAAGGGGCTTCC
53	16459	54	53	ACTCATTCTTCTCTCTCCTGCCACCATGTGAAGAAGGACATGTTTGTTTCCCC
54	16513	47	24	TCCACCATAATTGTAAGTTTCCTG
55	16560	27	20	TGTGAGTCAATTAAACTTCT
56	16587	2794	2787	AGAGCAGTGTCCTCACGCATTGGACTGGCCTCCAATGGGTGTAGTTACAGGGCCCCAACCACCTTTCACCAGATTGTATACTCACCTGTATCTGACCTTATTGCTACTCACACTCTAGGTCCCAGGATAAAATCCCAA
57	16588	>ENST000	1367	
58	17955	>ENST000	1367	
59	19322	>ENST000	928	
60	19381	672	671	AGCTTCAGTTGGCACCTGCCCCAGTCCAGCCTCTGGGAACCATGCAGCAGCTCCCAGCGGCCCTGCACCCACCACCAGCATCCGTTTCACCTGCAGTTGAAGATCCGTGAGGTGCCCAGAAGATCATGCAGTCATCA
61	20053	119	118	GTATAAAATATGATTTTCTAACCACTTGCTCGCCAACAAGGAAAACTTTTAAGTAGAGCAGAACCTGAATAGACAAGACATTTCTTTCTTTTGGTAGAAAATGATTTACCATCACTGT
62	20172	133	80	TAGTTAATTGTAGACTAGGTAATTTTAACTGTGATTTATTGCCGGAGACATTTTCTTCTGTACTGTAAAGTGTGTGTCAG
63	20250	>ENST000	558	
64	20305	570	61	CGCGATGGCTCACGCCTGTAATCCCAGCACTTTGGGAGGCTGAGACAGGCGGATCACGAGG
65	20808	>ENST000	67	
66	20875	>ENST000	3569	
67	20875	3574	3569	AGCCCTGCGCTTCCCCAGGTGAACCGGGCAGGAGCCTGTTGGGAAGGCAGCGACCCACATCTGTGTGCACCTTTGTGGATTTCAGGTTCCGGACGCAGGCGACCAAGCCAGAGCCAGCGCTGTCATACGCAGAGCA
68	24444	>ENST000	1678	
69	24449	646	506	GGGTCGCGCTCTAAGTGAGGCGCCAAGCGGTCTCCGCCTCAGGGTCTGAGGCTGCGAAAGGGGCGTAACGATGAGCGGTTCCTGCCAGAGGTCTGGGGAGGATAAAAAGCAGGAGGAAGAGGCGACGGCGGCC
70	25095	1027	1027	CAGATGTGGACCTGTTGGAGAACCAGCTGGGAGTGGCAGGAGCCCAGGCCCTCTGTGCCGCCCTCACAGTGAACCAGGCCATGCGGAAGATGCAGCTGTCAGGGAATGGCCTGGAGGAGCAGGCGGCCCAGCAC
71	26122	>ENST000	268	
72	26122	1305	1157	CCCAGAGGCGCACAGGAGACCTCAGGCCCAGACTCCACTCCCCAGCTGTGAAAGGACTGCTGGCCAGACCCCCAAGCTAGCCCGCCAGGCCTCCATAGAGCTGCCCAGCATGGCTGCATCCAGTACCAAGAGTTG
73	26390	>ENST000	889	
74	27279	>ENST000	1832	
75	27427	1195	1155	GAAACCTAGACCTATAAGTTTAGTACCTGAATCTCTTGCTGATTCATGCATTTCTATTAATCTTCTCAGTCAGATGTTTAGCTCAGATCTGTCCTCTTATCCAAGCTTTTCCAAGAGTCTCCTAGCTACACTTCTGGCATC

pack24_all_index

Redundancy Removal



A1 | 0

	A	B	C	D
1	0	>ENST000(	1655	
2	528	1	20	CATTGACCAGGCTGGTCTTG
3	529	135	32	ATTGACCAGGCTGGTCTTGAACTTCTGACCTC
4	664	91	90	ACAGCAAAAGAAACTACCATCAGAGTGAACAGGCAACCTACAGAATGGGAGAAAATTTTTGCAATCTACTCATCTGACAAAGGGCTAATA
5	755	171	170	CCAGAATCTACAAAGAACTCAAACAAATTTACAAGAAAAAACAAACAACCCCATCAAAAAGTGGGCAAAGGATATGAACAGACACTTCTCAAAAGAAGACATTTATGCAGCCAACAGACACATGAAAAAATGC
6	926	38	32	CACAATGAGATACCATCTCACACCAGTTAGAA
7	964	86	66	TCATTAAAAAGTCAGGAAACAACAGGTGCTGGAGAGGATGTGGAGAAATAGGAACACTTTTACACT
8	1050	44	43	GTTCAACCATTGTGGAAGACAGTGTGGCGATTCCTCAAGGATC
9	1094	454	453	AGAACTAGAAATACCATTTGACCCAGCCATCCCATTACTGGGTATATACCCAAAGGATTATAAATCATGCTGCTATAAAGACACATGCACACGTATGTTTATTGTGGCACTATTCACAATAGCAAAGACTTGGAACC
10	1548	36	35	TGGGTGCAGCACACCAACATGGCACATGTATACAT
11	1584	367	71	TGTAACAAACCTGCATGTTGTGTACATGTACCCTAGAACTTAAAGTATAATAAAAAATAAAAATAATAAAA
12	1655	>ENST000(	910	
13	1951	124	20	TCTGACCTTACTGCTGCTCA
14	2075	209	20	TGTGGATTTCACAGAGAAGA
15	2284	121	20	TGTGGATTTCACAGAGAAGA
16	2405	651	21	TGAAATCCCAAGACAATGGAG
17	2565	>ENST000(	491	
18	3056	>ENST000(	1784	
19	3056	95	23	CTGTGCTGGATGCTTCCTGCCTT
20	3151	37	32	GGGCCTTTGGCCACAGACTGAAGGCTGCACTG
21	3188	78	60	TTTCCTACTTTTGAGGTTTTGGGACTTGGACTGAGCCGCTACTGGCTTCCTTGCTCTTCA
22	3266	1251	24	GTGGGACTTCACCTTGTGATCATG
23	4517	122	42	GTGATATGGTTTGGCTGTGTCCCCACCCAAATCTCATCTTGA
24	4639	35	20	CTGTTCTCATGATAGTGAGT
25	4674	37	24	CTGATGGCTTTATAAGGGGCTTCC

lncRNA Database Intersections

	G_ncrna	LNCipedia	G_cdna	NONCODE	ncbook
G_ncrna	54941	22640	58	3093	12938
LNCipedia		126876	21765	17236	51787
G_cdna			177455	6089	15866
NONCODE				170767	14101
lncbook					268639

MicroRNA

- miRBase Database
  - 48,885 miRNA’s
  - 27,383 unique miRNA’s
  - 5312 human miRNA’s

miRNA in lncRNA

Fasta File	lncRNA	Match Records	Match miRNA	Match lncRNA
GRCh38_ncrna	54941	4564	2803	2937
LNCipedia	126876	6404	453	4291
GRCh38_cdna	177455	3502	405	2900
NONCODE	170767	1302	380	453
lncbook	268639	10658	532	8754

Questions?

Thank You!

Long Noncoding RNA

- **GRCh38_ncrna.fa**
  - **65,790,873 bp**
  - **67,419 lncRNA**
- European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Cambridge, United Kingdom.

Long Noncoding RNA

- **LNCipedia_5_2.fasta**
  - **192,690,141 bp**
  - **127,802 transcripts**
  - **56,946 genes**
- Ghent University - VIB, Life Sciences Research Institute in Flanders, Belgium.

Long Noncoding RNA

- **NONCODEv5_human.fa,**
  - **278,614,288 bp**
  - **165,911 lncRNA**
- Key Laboratory of Intelligent Information Processing, Advanced Computer Research Center, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China

Long Noncoding RNA

- **GRCh38_cdna.fa**
  - **316,791,371 bp**
  - **139,155 lncRNA**
- European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Cambridge, United Kingdom.

Long Noncoding RNA

- **lncRNA_lncbook.fa**
  - **405,815,189 bp**
  - **268,848 lncRNA**
- BIG Data Center, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, China