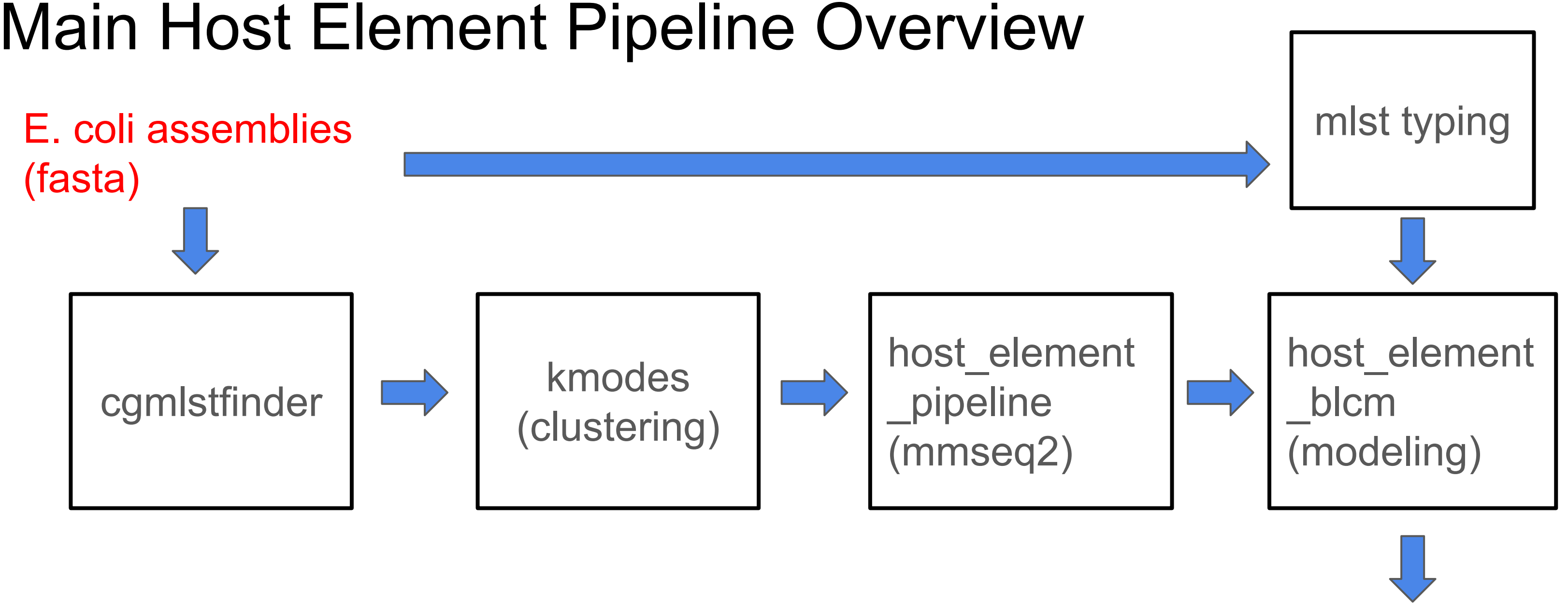
Host Element Pipeline Overview

ARAC
08/20/25

Pipeline Installation

1. git clone git@github.com:araclab/general.git
    a. This will clone the entire general repo, but what we want is just the host_element_v2 folder
2. cd Food-epidemiology/host_element_v2
3. Install the conda environments in pipeline_modules/conda_envs/
4. Modify the script pathing within each script
    a. Can create a config file in the future to make script pathing easier to change
5. Please delete/remove all other folders outside, everything you need is in host_element_v2

Main Host Element Pipeline Overview



E. coli assemblies (fasta)

mlst typing

cgmlstfinder

kmodes (clustering)

host_element_pipeline (mmseq2)

host_element_blcm (modeling)

E. coli assemblies
Host Origin Predictions

Note:
Each module operates independently and you will need to run each one one after the other manually in order

cgmlstFinder Module

Inputs: E. coli assemblies (fasta)

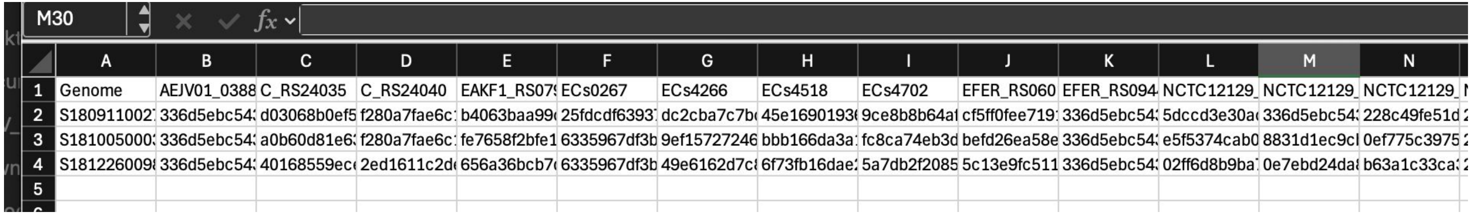
Outputs: kmodes_ready_inputfile.txt

- Rows – E. coli samples
- Columns – 1st col are genomes; rest are cgmlst calls (N=2513)
- Calls are converted into md5

Note:
cgmlstFinder is designed for E. coli, the input flag is set to ecoli in the script file

cmd: bash cgmlstFinder_Submitter.sh (Data_Folder_input) (Data_Folder_Samplelist_input) (Job_Name_input)

- **Data_Folder_input – folder containing your E. coli assemblies**
- **Data_Folder_Samplelist_input – text file, with each line being the sample/filename from the Data_Folder_Input (include the full filename with .fasta extension)**
- **Job_Name_input -- names the output folder and slurm jobs**



M30

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Genome	AEJV01_0388	C_RS24035	C_RS24040	EAKF1_RS07	ECs0267	ECs4266	ECs4518	ECs4702	EFER_RS060	EFER_RS094	NCTC12129_	NCTC12129_	NCTC12129_
2	S180911002	336d5ebc54	d03068b0ef5	f280a7fae6c	b4063baa99	25fdcdf6393	dc2cba7c7b	45e169019	9ce8b8b64a	cf5ff0fee719	336d5ebc54	5dccd3e30a	336d5ebc54	228c49fe51d
3	S181005000	336d5ebc54	a0b60d81e6	f280a7fae6c	fe7658f2bfe1	6335967df3b	9ef15727246	bbb166da3a	fc8ca74eb3d	befd26ea58e	336d5ebc54	e5f5374cab0	8831d1ec9c	0ef775c3975
4	S181226009	336d5ebc54	40168559ec	2ed1611c2d	656a36bcb7	6335967df3b	49e6162d7c	6f73fb16dae	5a7db2f2085	5c13e9fc511	336d5ebc54	02ff6d8b9ba	0e7ebd24da	b63a1c33ca
5														

kmodes Module

Inputs: kmodes_ready_inputfile.txt from cgmlstfinder module

Outputs: kmodes_cgmlst_clustering_predictions.csv

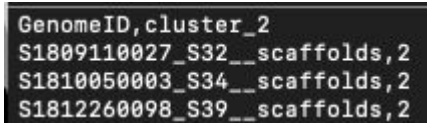
- Two column csv file:
  - GenomeID: E. coli assemblies names
  - cluster_2: kmodes clustering number (1 or 2)

FULL_sb27_training_context_kmodes_output_Cluster_2_model.pkl

This is a pre-trained kmodes model on a worldwide E. coli dataset provided by pathogen watch (https://pathogen.watch/). It uses the script files within kmodes/model_training_scripts/ to generate the model.

cmd: python kmodes_clustering_predicting.py (kmodes_ready_inputfile.txt) (FULL_sb27_training_context_kmodes_output_Cluster_2_model.pkl)

- **kmodes_ready_inputfile.txt – input file comes from the cgmlstfinder module output**
- **FULL_sb27_training_context_kmodes_output_Cluster_2_model.pkl – pretrained model (kmodes/trained_models/cluster_2/)**



```
GenomeID,cluster_2
S1809110027_S32__scaffolds,2
S1810050003_S34__scaffolds,2
S1812260098_S39__scaffolds,2
```

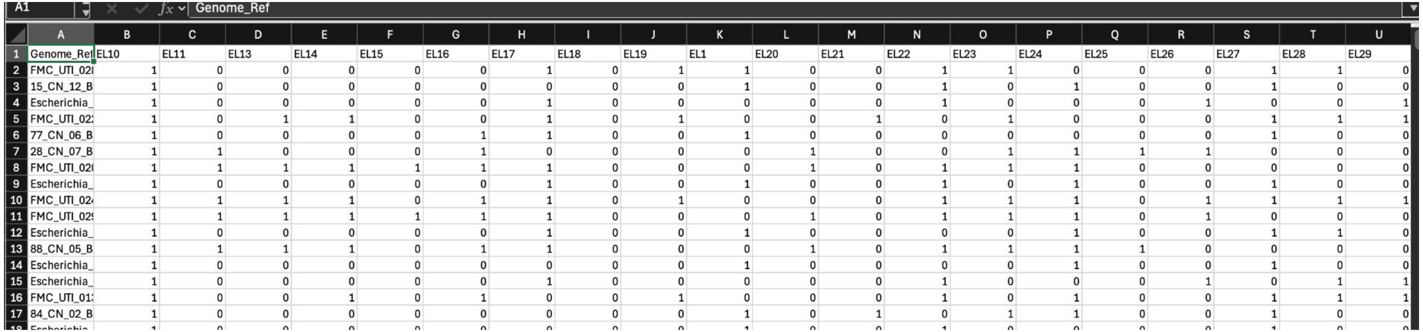
host_element_pipeline (mmseq2 caller)

Inputs: E. coli assemblies (fasta), Host label

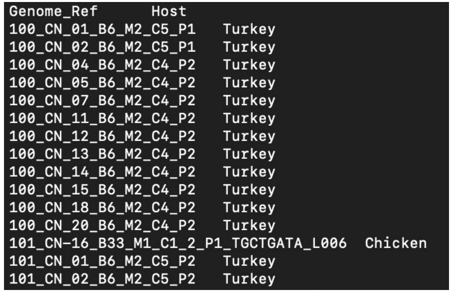
Outputs: host_element_pipeline_element_presence.tsv (Presence Absence Matrix with N=17 elements)

cmd: sbatch host_element_pipeline_Submitter.sh (Data_Folder_input) (Data_Folder_Samplelist_input) (Data_Folder_Hostlist_input) (Job_Name_input)

- **Data_Folder_input – folder containing your E. coli assemblies**
- **Data_Folder_Samplelist_input – text file, with each line being the sample/filename from the Data_Folder_Input (include the full filename with .fasta extension)**
- **Data_Folder_Hostlist_input – tab delimited textfile with two columns: Genome_Ref and Host**
  - **Genome_Ref – filename (without the .fasta extension)**
  - **Host – [Turkey, Chicken, Pork, Beef, Human], source host where E. coli was sequenced from.**
- **Job_Name_input -- names the output folder and slurm jobs**



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	Genome_Ref	EL10	EL11	EL13	EL14	EL15	EL16	EL17	EL18	EL19	EL1	EL20	EL21	EL22	EL23	EL24	EL25	EL26	EL27	EL28	EL29
2	FMC_UTI_02	1	0	0	0	0	0	1	0	1	1	0	0	1	1	0	0	0	1	1	0
3	15_CN_12_B	1	0	0	0	0	0	0	0	0	1	0	0	1	0	1	0	0	1	0	0
4	Escherichia_	1	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	1	0	0	1
5	FMC_UTI_02	1	0	1	1	0	0	1	0	1	0	0	1	0	1	0	0	0	1	1	1
6	77_CN_06_B	1	0	0	0	0	1	1	0	0	1	0	0	0	0	0	0	0	1	0	0
7	28_CN_07_B	1	1	0	0	0	1	0	0	0	0	1	0	0	1	1	1	1	0	0	0
8	FMC_UTI_02	1	1	1	1	1	1	1	0	0	0	1	0	1	1	1	0	0	0	0	0
9	Escherichia_	1	0	0	0	0	0	1	0	0	1	0	0	1	0	1	0	0	1	0	0
10	FMC_UTI_02	1	1	1	1	0	1	1	0	1	0	0	0	1	1	1	0	1	1	1	1
11	FMC_UTI_02	1	1	1	1	1	1	1	0	0	0	1	0	1	1	1	0	1	0	0	0
12	Escherichia_	1	0	0	0	0	0	1	0	0	1	0	0	0	0	1	0	0	1	1	0
13	88_CN_05_B	1	1	1	1	0	1	1	0	0	0	1	0	1	1	1	1	0	0	0	0
14	Escherichia_	1	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	1	0	0
15	Escherichia_	1	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	1	0	1	1
16	FMC_UTI_01	1	0	0	1	0	1	0	0	1	0	0	0	1	0	1	0	0	1	1	1
17	84_CN_02_B	1	0	0	0	0	0	0	0	0	1	0	1	0	1	1	0	0	1	0	0
18	Escherichia_	1	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	1	0	0



```
Genome_Ref        Host
100_CN_01_B6_M2_C5_P1   Turkey
100_CN_02_B6_M2_C5_P1   Turkey
100_CN_04_B6_M2_C4_P2   Turkey
100_CN_05_B6_M2_C4_P2   Turkey
100_CN_07_B6_M2_C4_P2   Turkey
100_CN_11_B6_M2_C4_P2   Turkey
100_CN_12_B6_M2_C4_P2   Turkey
100_CN_13_B6_M2_C4_P2   Turkey
100_CN_14_B6_M2_C4_P2   Turkey
100_CN_15_B6_M2_C4_P2   Turkey
100_CN_18_B6_M2_C4_P2   Turkey
100_CN_20_B6_M2_C4_P2   Turkey
101_CN-16_B33_M1_C1_2_P1_TGCTGATA_L006  Chicken
101_CN_01_B6_M2_C5_P2   Turkey
101_CN_02_B6_M2_C5_P2   Turkey
```

mlst typing

Will add this later to the git, but you can install and run this on your own.

Choose either mlst typers and run the E. coli assemblies to obtain mlst types.

https://github.com/tseemann/mlst

https://bitbucket.org/genomicepidemiology/mlst/src/master/

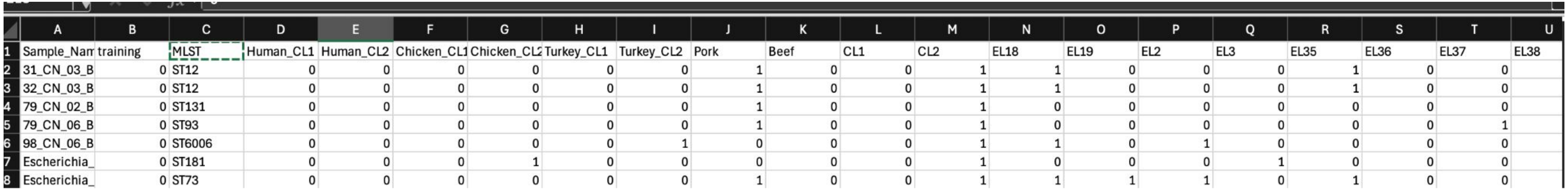
host_element_blcm (modeling)

Inputs: Modified 082025_any_element_presence_input_bigfuti_dcfuti.csv

Outputs: pred_scores.csv

Manual Steps to generate the Modified 082025_any_element_presence_input_bigfuti_dcfuti.csv

1. Copy the 082025_any_element_presence_input_bigfuti_dcfuti.csv from host_element_blcm/base_blcm_input/ folder
2. Append the respective information for each sample you want to add following the columns names from:
   a. kmodes_cgmlst_clustering_predictions.csv – kmodes module output
   b. host_element_pipeline_element_presence.tsv – host_element_pipeline module output
   c. host labels combinations
   d. MLST
   e. Training Column set to 0 (this will generate predictions for them)



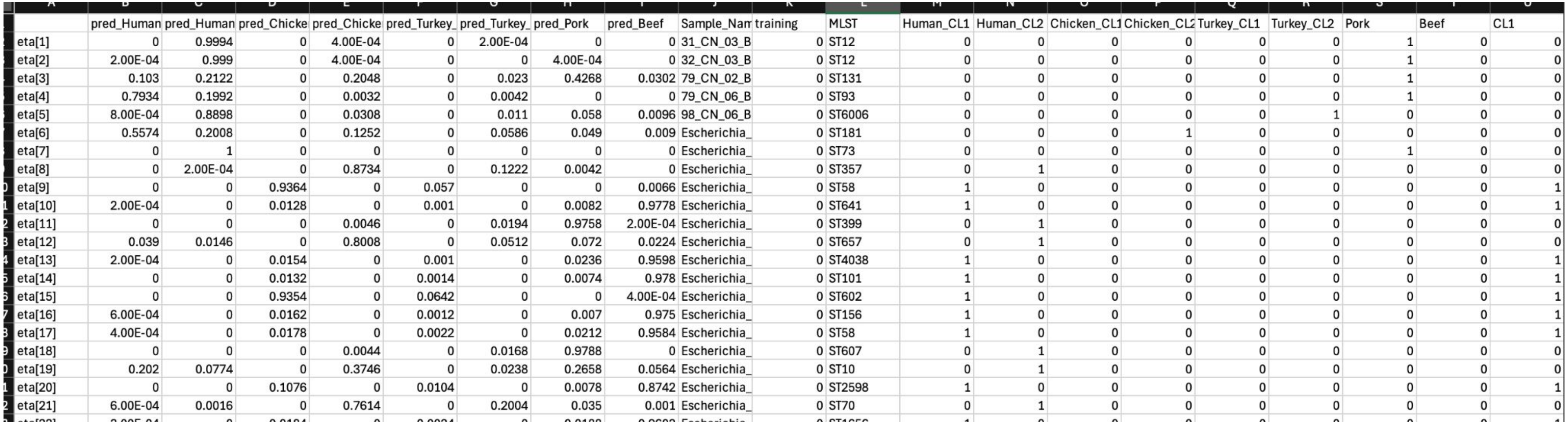
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	Sample_Nam	training	MLST	Human_CL1	Human_CL2	Chicken_CL1	Chicken_CL2	Turkey_CL1	Turkey_CL2	Pork	Beef	CL1	CL2	EL18	EL19	EL2	EL3	EL35	EL36	EL37	EL38
2	31_CN_03_B	0	ST12	0	0	0	0	0	0	1	0	0	1	1	0	0	0	1	0	0	
3	32_CN_03_B	0	ST12	0	0	0	0	0	0	1	0	0	1	1	0	0	0	1	0	0	
4	79_CN_02_B	0	ST131	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	
5	79_CN_06_B	0	ST93	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	1	
6	98_CN_06_B	0	ST6006	0	0	0	0	0	1	0	0	0	1	1	0	1	0	0	0	0	
7	Escherichia_	0	ST181	0	0	0	1	0	0	0	0	0	1	0	0	0	1	0	0	0	
8	Escherichia_	0	ST73	0	0	0	0	0	0	1	0	0	1	1	1	1	0	1	0	0	

host_element_blcm (modeling)

cmd: sbatch run_hostelement_blca.sh (Modified 082025_any_element_presence_input_bigfuti_dcfuti.csv) (Output_name)

- **Modified 082025_any_element_presence_input_bigfuti_dcfuti.csv – Follow the manual instructions to generate this input filec**
- **Output_name – names the output folder and output file names**

pred_scores.csv generated by the blcm file



	pred_Human	pred_Human	pred_Chicke	pred_Chicke	pred_Turkey_	pred_Turkey_	pred_Pork	pred_Beef	Sample_Nam	training	MLST	Human_CL1	Human_CL2	Chicken_CL1	Chicken_CL2	Turkey_CL1	Turkey_CL2	Pork	Beef	CL1
eta[1]	0	0.9994	0	4.00E-04	0	2.00E-04	0	0	31_CN_03_B	0	ST12	0	0	0	0	0	0	1	0	0
eta[2]	2.00E-04	0.999	0	4.00E-04	0	0	4.00E-04	0	32_CN_03_B	0	ST12	0	0	0	0	0	0	1	0	0
eta[3]	0.103	0.2122	0	0.2048	0	0.023	0.4268	0.0302	79_CN_02_B	0	ST131	0	0	0	0	0	0	1	0	0
eta[4]	0.7934	0.1992	0	0.0032	0	0.0042	0	0	79_CN_06_B	0	ST93	0	0	0	0	0	0	1	0	0
eta[5]	8.00E-04	0.8898	0	0.0308	0	0.011	0.058	0.0096	98_CN_06_B	0	ST6006	0	0	0	0	0	1	0	0	0
eta[6]	0.5574	0.2008	0	0.1252	0	0.0586	0.049	0.009	Escherichia_	0	ST181	0	0	0	1	0	0	0	0	0
eta[7]	0	1	0	0	0	0	0	0	Escherichia_	0	ST73	0	0	0	0	0	0	1	0	0
eta[8]	0	2.00E-04	0	0.8734	0	0.1222	0.0042	0	Escherichia_	0	ST357	0	1	0	0	0	0	0	0	0
eta[9]	0	0	0.9364	0	0.057	0	0	0.0066	Escherichia_	0	ST58	1	0	0	0	0	0	0	0	1
eta[10]	2.00E-04	0	0.0128	0	0.001	0	0.0082	0.9778	Escherichia_	0	ST641	1	0	0	0	0	0	0	0	1
eta[11]	0	0	0	0.0046	0	0.0194	0.9758	2.00E-04	Escherichia_	0	ST399	0	1	0	0	0	0	0	0	0
eta[12]	0.039	0.0146	0	0.8008	0	0.0512	0.072	0.0224	Escherichia_	0	ST657	0	1	0	0	0	0	0	0	0
eta[13]	2.00E-04	0	0.0154	0	0.001	0	0.0236	0.9598	Escherichia_	0	ST4038	1	0	0	0	0	0	0	0	1
eta[14]	0	0	0.0132	0	0.0014	0	0.0074	0.978	Escherichia_	0	ST101	1	0	0	0	0	0	0	0	1
eta[15]	0	0	0.9354	0	0.0642	0	0	4.00E-04	Escherichia_	0	ST602	1	0	0	0	0	0	0	0	1
eta[16]	6.00E-04	0	0.0162	0	0.0012	0	0.007	0.975	Escherichia_	0	ST156	1	0	0	0	0	0	0	0	1
eta[17]	4.00E-04	0	0.0178	0	0.0022	0	0.0212	0.9584	Escherichia_	0	ST58	1	0	0	0	0	0	0	0	1
eta[18]	0	0	0	0.0044	0	0.0168	0.9788	0	Escherichia_	0	ST607	0	1	0	0	0	0	0	0	0
eta[19]	0.202	0.0774	0	0.3746	0	0.0238	0.2658	0.0564	Escherichia_	0	ST10	0	1	0	0	0	0	0	0	0
eta[20]	0	0	0.1076	0	0.0104	0	0.0078	0.8742	Escherichia_	0	ST2598	1	0	0	0	0	0	0	0	1
eta[21]	6.00E-04	0.0016	0	0.7614	0	0.2004	0.035	0.001	Escherichia_	0	ST70	0	1	0	0	0	0	0	0	0