宏基因组二代测序数据的病毒鉴定自动化分析流程

李洋
liyang@ivdc.chinacdc.cn

中心实验室

卫生部医学病毒学重点实验室

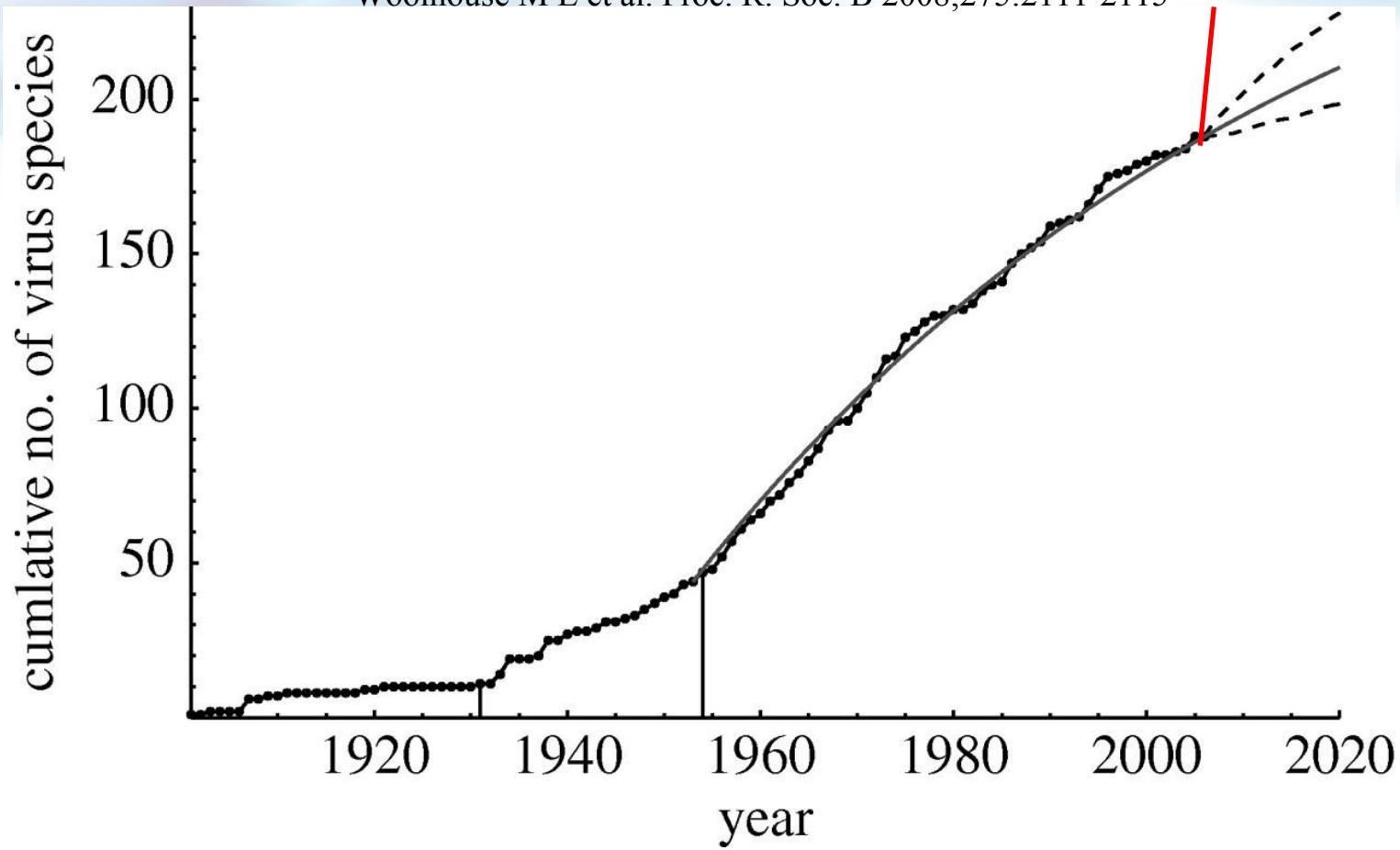
中国疾病预防控制中心病毒病预防控制所

病毒与下一代测序（NGS）

病毒鉴定自动化分析流程 (Virus Identification Pipeline)

可感染人病毒的发现时间曲线

Woolhouse M E et al. Proc. R. Soc. B 2008;275:2111-2115



随着越来越多的病毒被发现了，其中可能致病的病毒种类也随之增加。

检测技术概述

- 形态学（电镜）
- 组织学（细胞培养）
- 免疫学（ELISA,Western Blot, IF,NT,HI）
- 分子生物学（PCR,RT-PCR,qPCR,mPCR, Microarray,Sequencing……）

为什么？

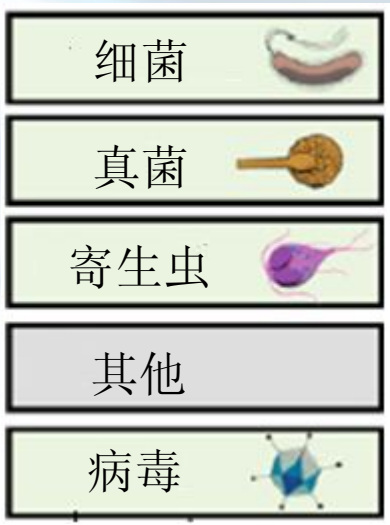
- 病毒变异
- 含量低，背景噪声大
- 新发疾病

接下来呢？

- 分离培养
- 高通量测序

临床样本

高通量测序

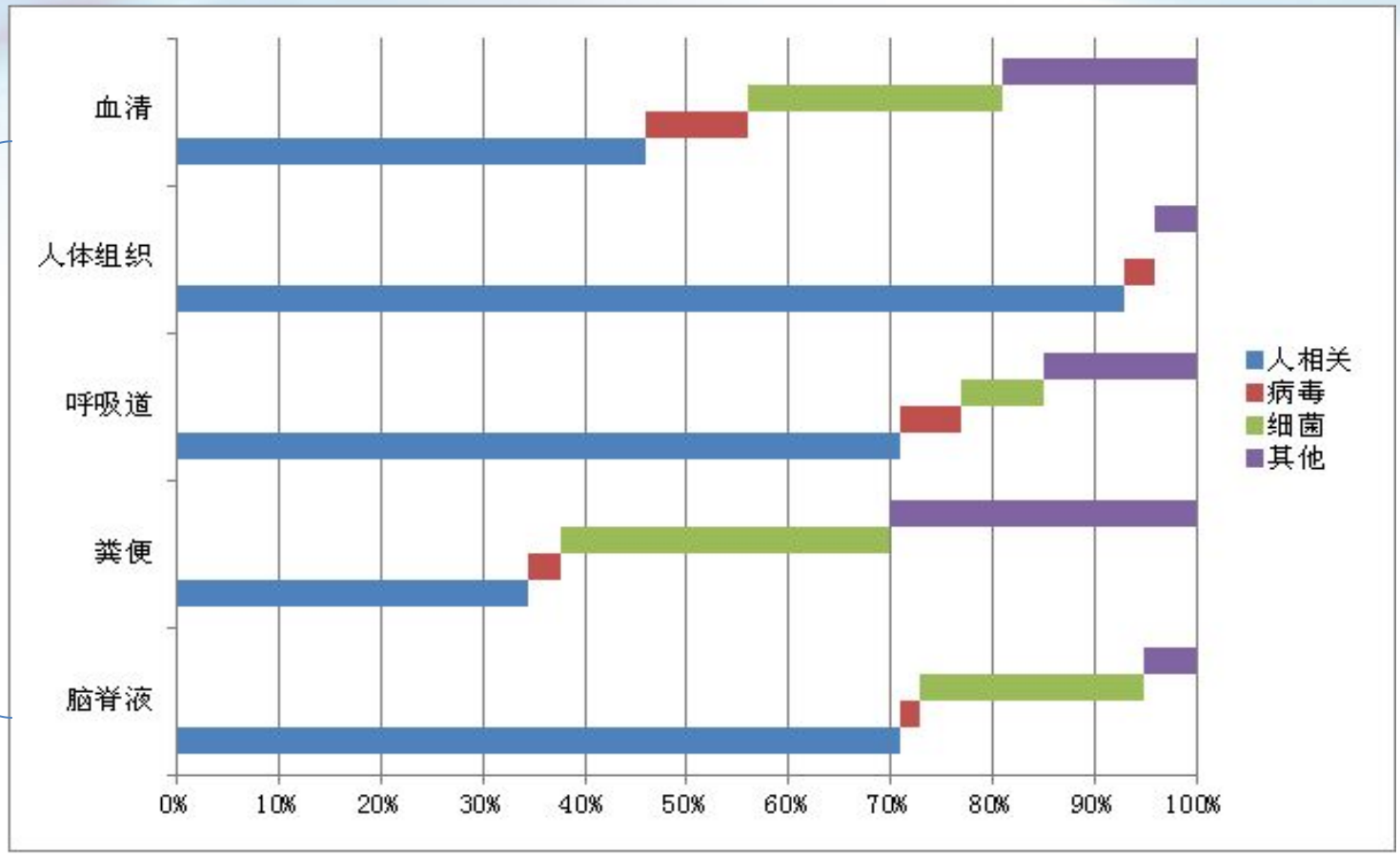


临床宏基因组

(Clinical Metagenomics)

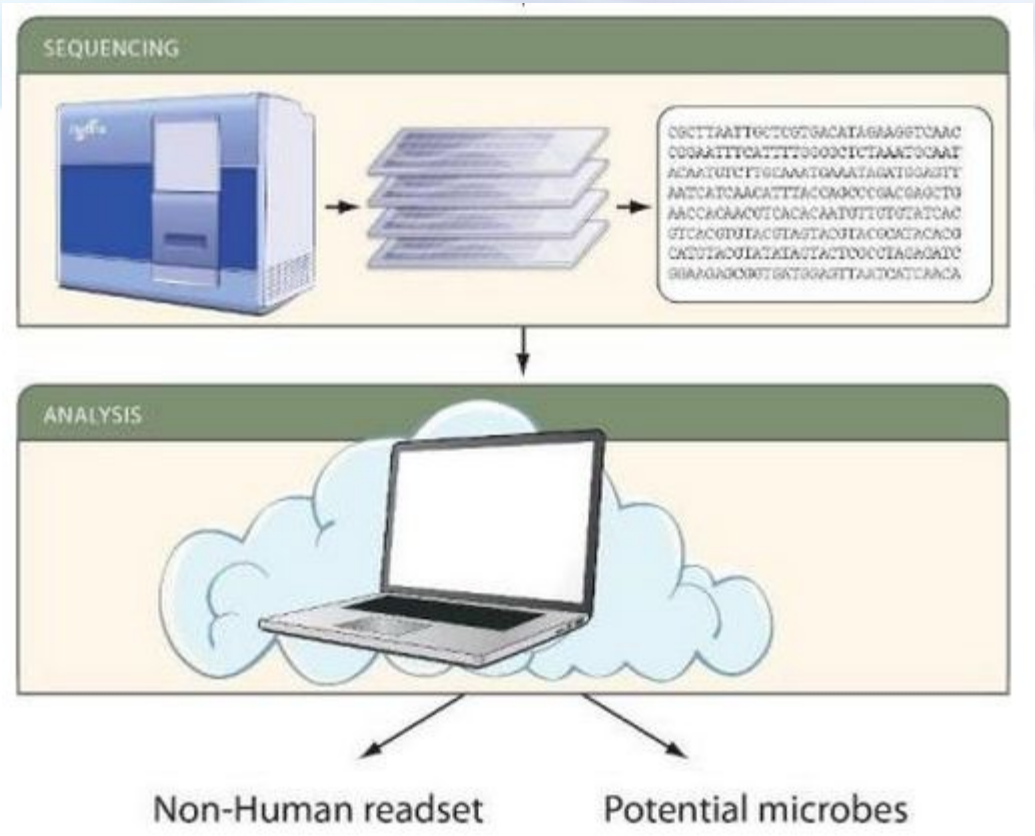
可以在不需要目标病原体的背景资料的情况下，
一次高通量测序试验就可以完成检测工作。

临床样本



病毒检测如大海捞针

主要难点



- 比对算法/分类算法
  （BLAST）
- 参考数据库
  （NT/NR）
- 结果信息？

Table 1. Summary of viruses identified in this study.

Patient code	Clinic virus ID	Virochip virus ID	Sequencing virus ID	Virus TaxID[a]	# virus reads	# initial reads	Fraction virus reads
187	DENV-2	DENV-2	Dengue virus 2	11060	4280	1.1E+06	3.9E−03
275	DENV-2	DENV-2	Dengue virus 2	11060	1511	1.6E+06	9.7E−04
282	DENV-2	DENV-2	Dengue virus 2	11060	699	1.6E+06	4.2E−04
266	DENV-2	DENV-2	Dengue virus 2	11060	135749	4.8E+06	2.8E−02
274	DENV-1	DENV-1	Dengue virus 1	11053	27	1.2E+06	2.3E−05
401[b]	HAV	HAV	Hepatitis A virus	12092	2164	1.8E+05	1.2E−02
401[b]	HAV	HAV	Hepatitis A virus	12092	4562	1.3E+06	3.5E−03
235	-	-	Human herpesvirus 6	10368	116	5.5E+06	2.1E−05
451	-	-	Human herpesvirus 6	10368	88	2.7E+06	3.2E−05
207	-	-	Human herpesvirus 6	10368	390	9.6E+06	4.1E−05
432	-	-	Human herpesvirus 6	10368	411	3.5E+06	1.2E−04
574	-	-	Human herpesvirus 6	10368	138	3.2E+06	4.4E−05
370	-	-	Human herpesvirus 6	10368	90	3.2E+06	2.9E−05
78	-	-	Human herpesvirus 6	10368	113	1.2E+06	9.8E−05

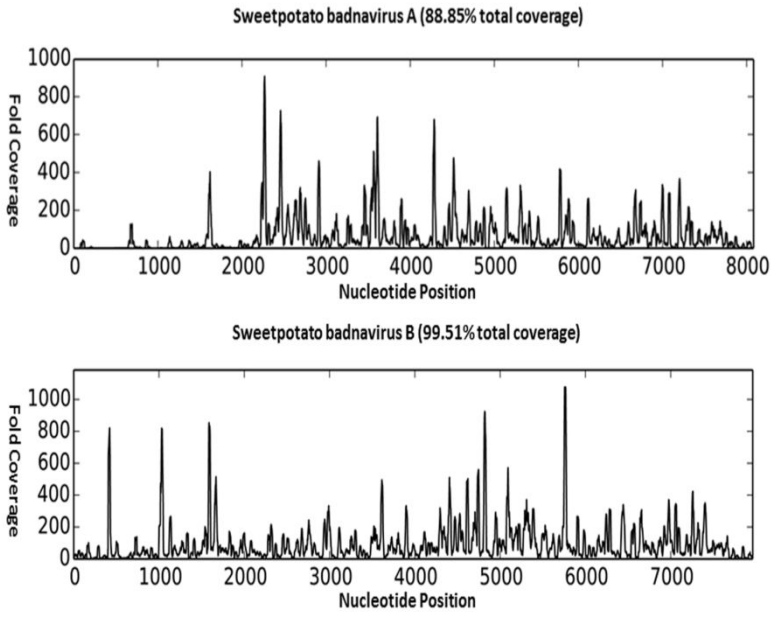
Yozwiak, N.L., Skewes-Cox, P., Stenglein, M.D., Balmaseda, A., Harris, E. and DeRisi, J.L. (2012) Virus identification in unknown tropical febrile illness cases using deep sequencing. *PLoS Negl Trop Dis*, **6**, e1485.

病毒与下一代测序（NGS）

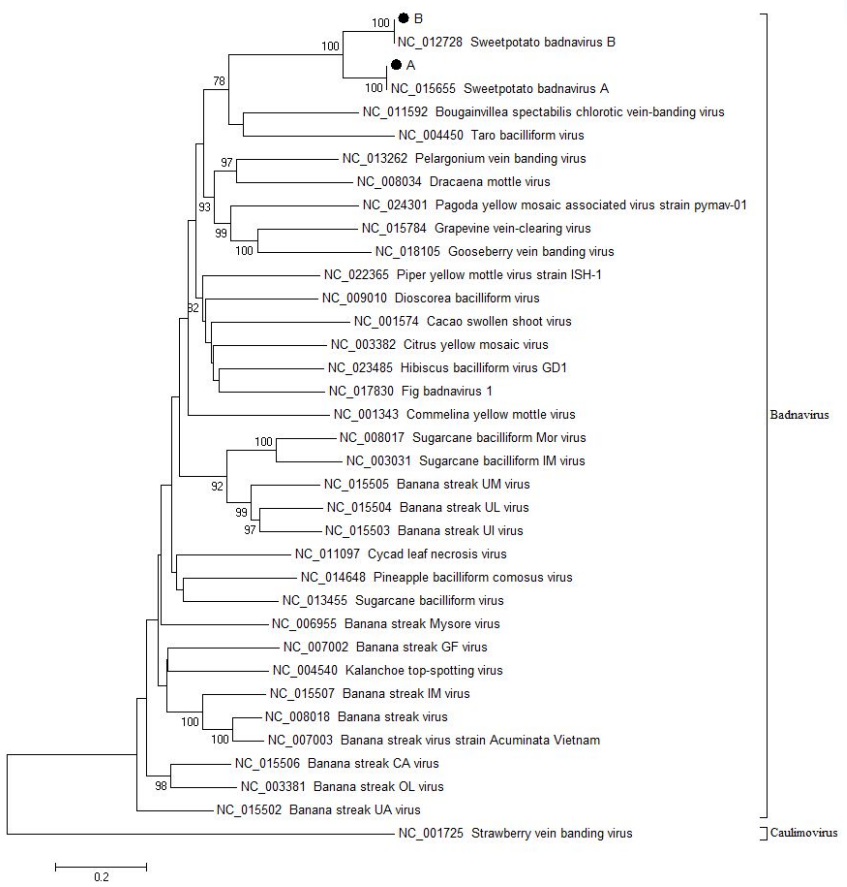
病毒鉴定自动化分析流程 (Virus Identification Pipeline)

什么样的分析结果

基因组覆盖度



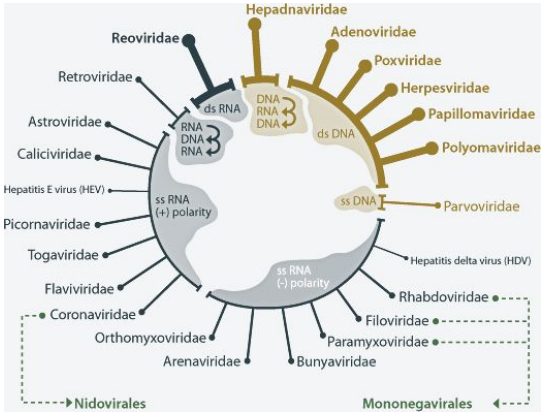
聚类分析



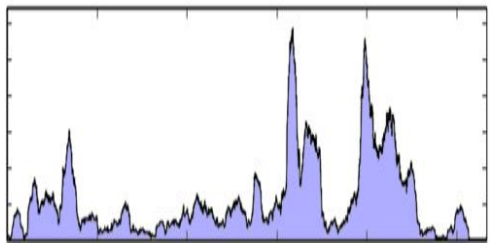
汇总报告

Virus Identification Pipeline (VIP)

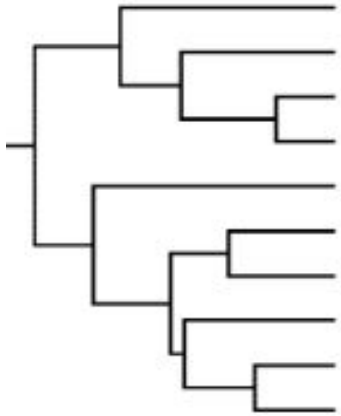
分类学



基因组覆盖度



聚类分析



高准确度分类检测

SRA	sample type	literature report	VIP_result	Coverage%	Reads
SRR453448	serum	Dengue virus 2	Dengue virus 2	99.14	64778
SRR453458	serum	Hepatitis A virus	Hepatitis A virus	57.15	2315
SRR1106123	serum	Hepatitis C virus, Hepatitis G virus	Hepatitis C virus	100.00	108274
			Hepatitis G virus	100.00	751691
SRR1106548	plasma	HIV	HIV	96.61	29528
SRR1106553	nasal swab	H1N1	H1N1	98.25	12613
SRR1106550	stool	Sapovirus, Rotavirus A	Sapovirus	96.16	13019
			Rotavirus A	92.41	786
			Adeno-associated virus 2	80.74	1836
			Human parechovirus 6	76.36	4467
			Torque teno mini virus 5	42.09	270

特异性和灵敏度结果

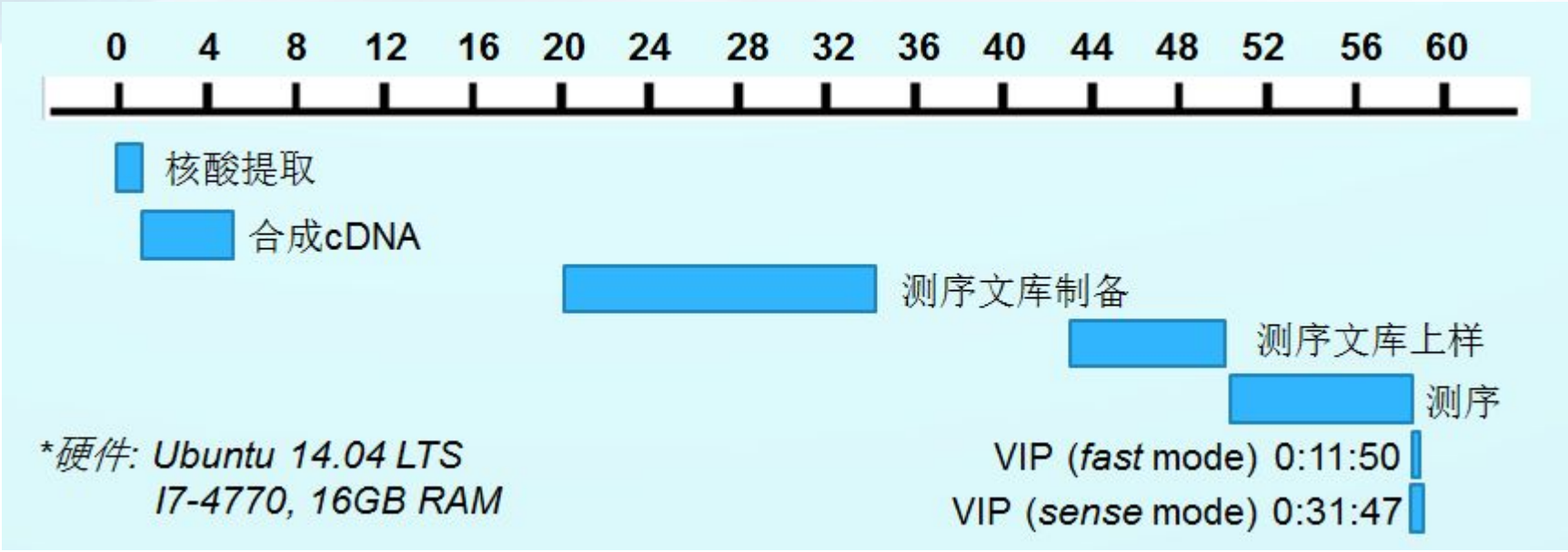
	BLAST	VIP		
SRR1106553		Tru	Neg	H1N1 (JF915184 - JF915191)
	Tru	51,509.0	89.0	
	Neg	15.0	0.0	
	BLAST	VIP		
SRR1106548		Tru	Neg	HIV (AF063223.1)
	Tru	30,108.0	51.0	
	Neg	452.0	0.0	
	BLAST	VIP		
SRR1170797		Tru	Neg	BVDV (JN380086.1)
	Tru	35,554.0	1,087.0	
	Neg	65.0	0.0	

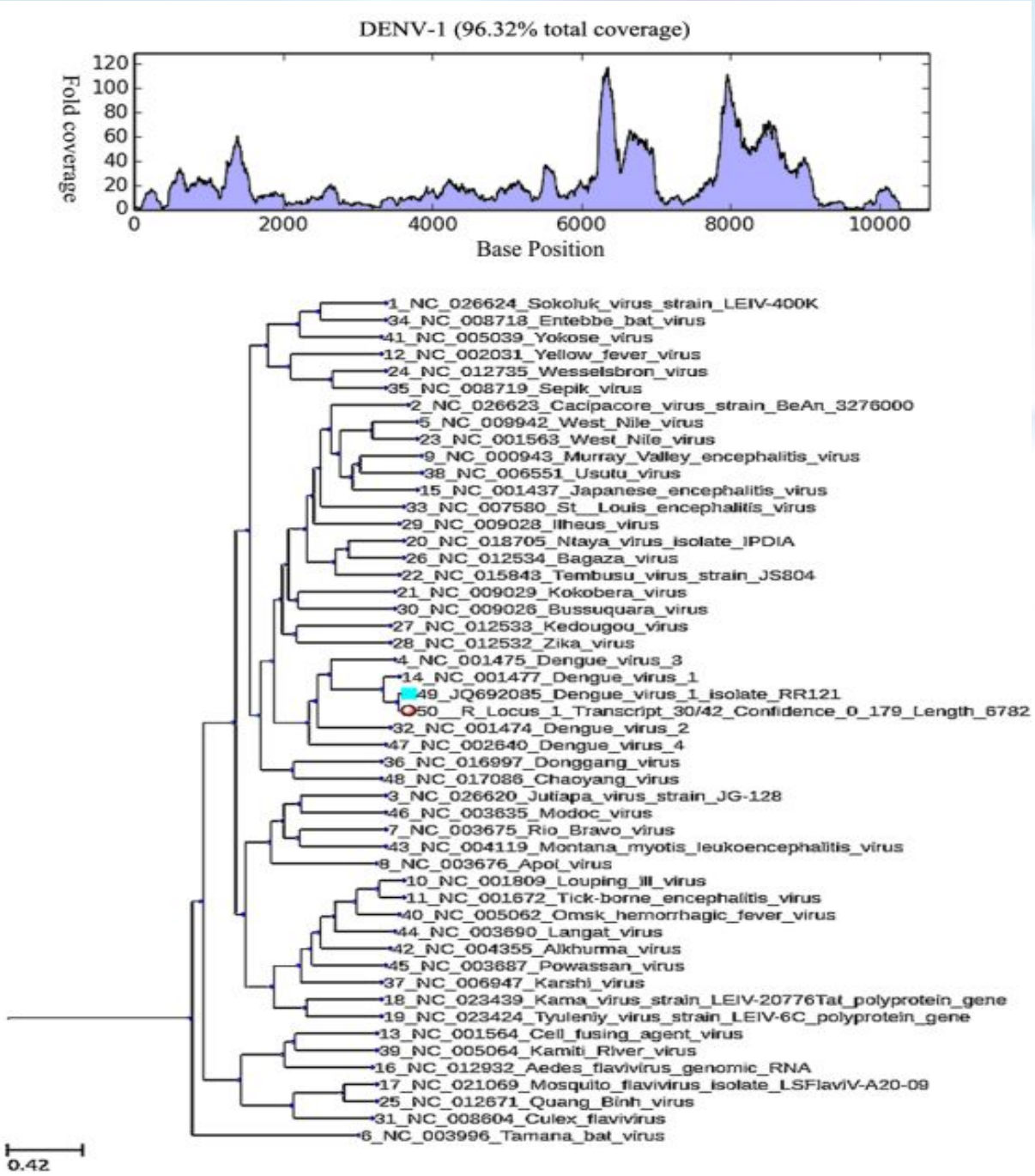
特异性和灵敏度结果

TPR = TP/(TP + FN) **FPR = FP/(FP + TN)**

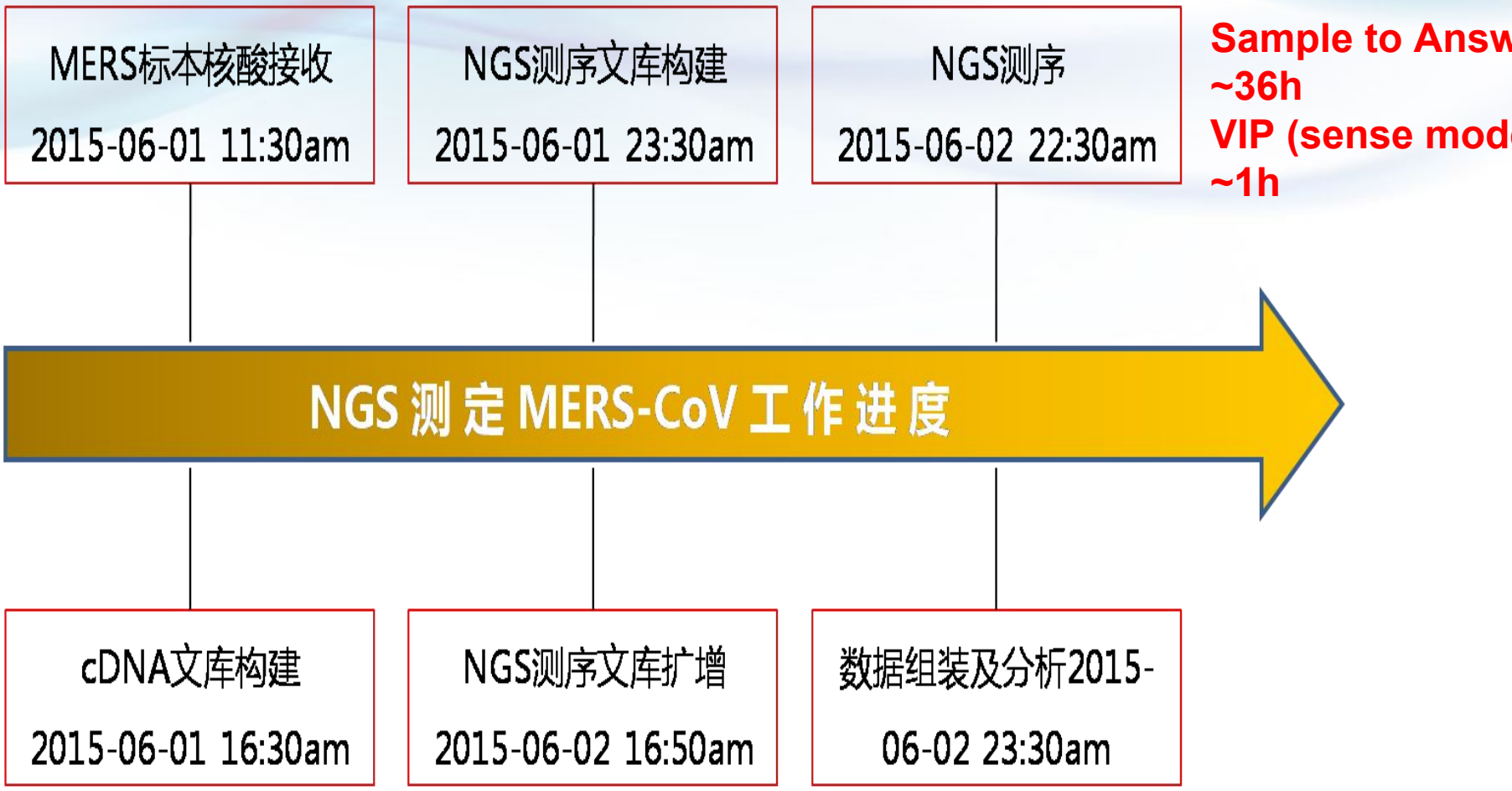
SRA	Viruses	% True Positive Rate	% False Positive Rate
SRR1170797	BVDV	97.03	100.00
SRR1106548	HIV	99.83	100.00
SRR1106553	H1N1	99.82	100.00

“十二五”考核


0 4 8 12 16 20 24 28 32 36 40 44 48 52 56 60
核酸提取
合成cDNA
测序文库制备
测序文库上样
测序
*硬件: Ubuntu 14.04 LTS
I7-4770, 16GB RAM
VIP (fast mode) 0:11:50
VIP (sense mode) 0:31:47


DENV-1 (96.32% total coverage)
Fold coverage
120
100
80
60
40
20
0
0
2000
4000
6000
8000
10000
Base Position
1_NC_026624_Sokoluk_virus_strain_LEIV-400K
34_NC_008718_Entebbe_bat_virus
41_NC_005039_Yokose_virus
12_NC_002031_Yellow_fever_virus
24_NC_012735_Wesselsbron_virus
35_NC_008719_Sepik_virus
2_NC_026623_Cacipacore_virus_strain_BeAn_3276000
5_NC_009942_West_Nile_virus
23_NC_001563_West_Nile_virus
9_NC_000943_Murray_Valley_encephalitis_virus
38_NC_006551_Usutu_virus
15_NC_001437_Japanese_encephalitis_virus
33_NC_007580_St__Louis_encephalitis_virus
29_NC_009028_Ilheus_virus
20_NC_018705_Ntaya_virus_isolate_IPDIA
26_NC_012534_Bagaza_virus
22_NC_015843_Tembusu_virus_strain_JS804
21_NC_009029_Kokobera_virus
30_NC_009026_Bussuquara_virus
27_NC_012533_Kedougou_virus
28_NC_012532_Zika_virus
4_NC_001475_Dengue_virus_3
14_NC_001477_Dengue_virus_1
49_JQ692085_Dengue_virus_1_isolate_RR121
50__R_Locus_1_Transcript_30/42_Confidence_0_179_Length_6782
32_NC_001474_Dengue_virus_2
47_NC_002640_Dengue_virus_4
36_NC_016997_Donggang_virus
48_NC_017086_Chaoyang_virus
3_NC_026620_Jutiapa_virus_strain_JG-128
46_NC_003635_Modoc_virus
7_NC_003675_Rio_Bravo_virus
43_NC_004119_Montana_myotis_leukoencephalitis_virus
8_NC_003676_Apoi_virus
10_NC_001809_Louping_ill_virus
11_NC_001672_Tick-borne_encephalitis_virus
40_NC_005062_Omsk_hemorrhagic_fever_virus
44_NC_003690_Langat_virus
42_NC_004355_Alkhurma_virus
45_NC_003687_Powassan_virus
37_NC_006947_Karshi_virus
18_NC_023439_Kama_virus_strain_LEIV-20776Tat_polyprotein_gene
19_NC_023424_Tyuleniy_virus_strain_LEIV-6C_polyprotein_gene
13_NC_001564_Cell_fusing_agent_virus
39_NC_005064_Kamiti_River_virus
16_NC_012932_Aedes_flavivirus_genomic_RNA
17_NC_021069_Mosquito_flavivirus_isolate_LSFlaviV-A20-09
25_NC_012671_Quang_Binh_virus
31_NC_008604_Culex_flavivirus
6_NC_003996_Tamana_bat_virus
0.42

MERS应急



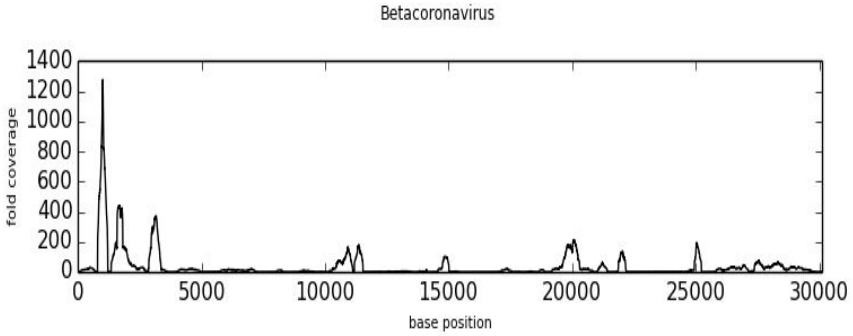
NGS Timeline

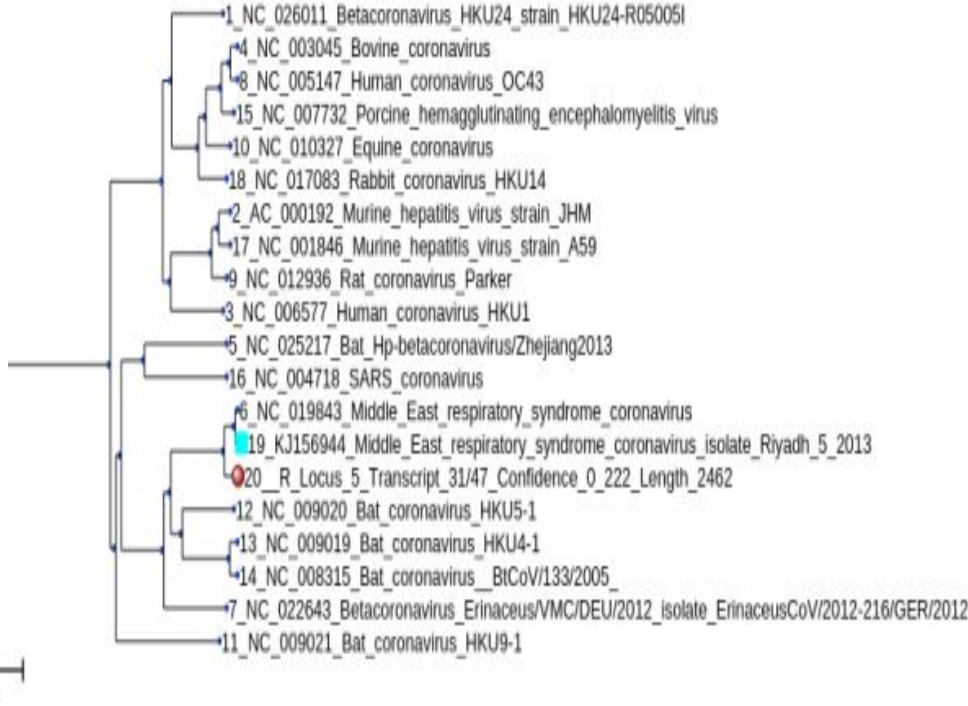
MERS应急

Species	Genus	GI	%Coverage	Reads_hit	Reads_num	Average depth of coverage
Middle East respiratory syndrome coronavirus, complete genome	Betacoronavirus	511262662	67.95	5,980	6,340	3,506.33

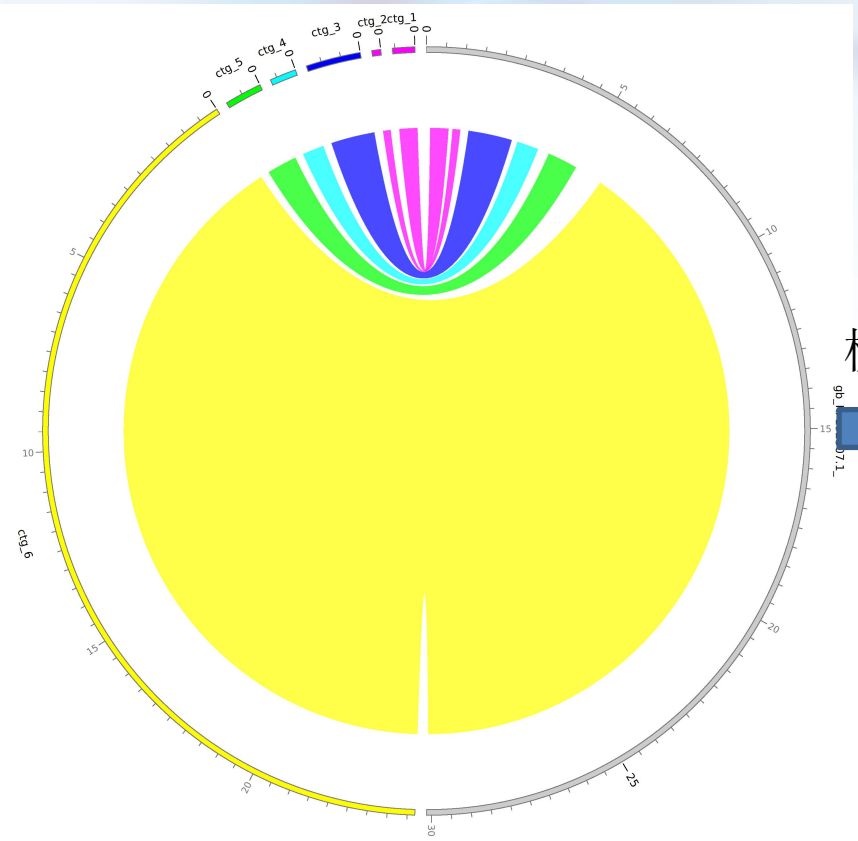
基因组覆盖度(67.95%)

进化分析

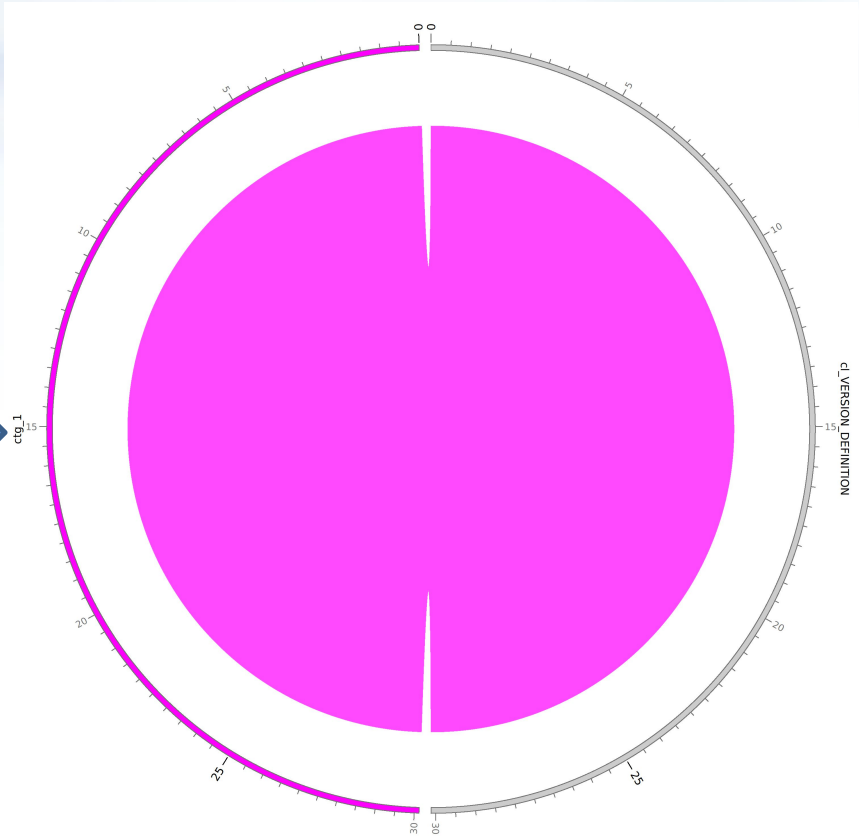




MERS GD01全基因组绘制



校准

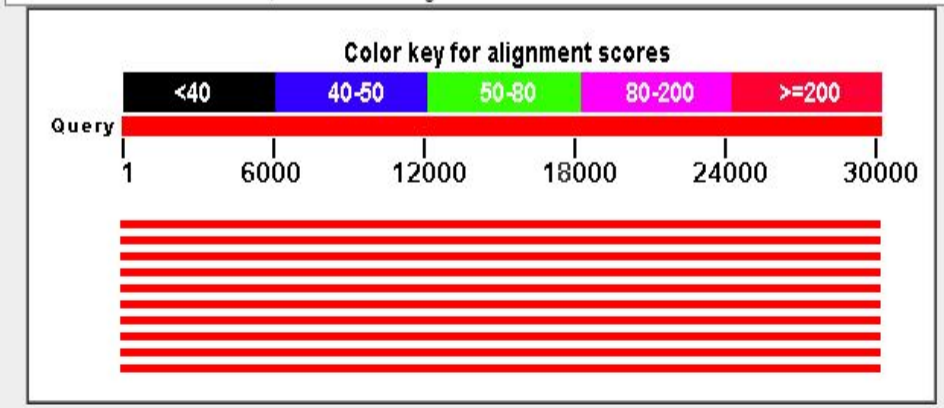


Draft genome
(95.8% coverage)

Complete genome
(100.0% coverage)

Distribution of 10 Blast Hits on the Query Sequence

Mouse over to see the defline, click to show alignments



Descriptions

Sequences producing significant alignments:

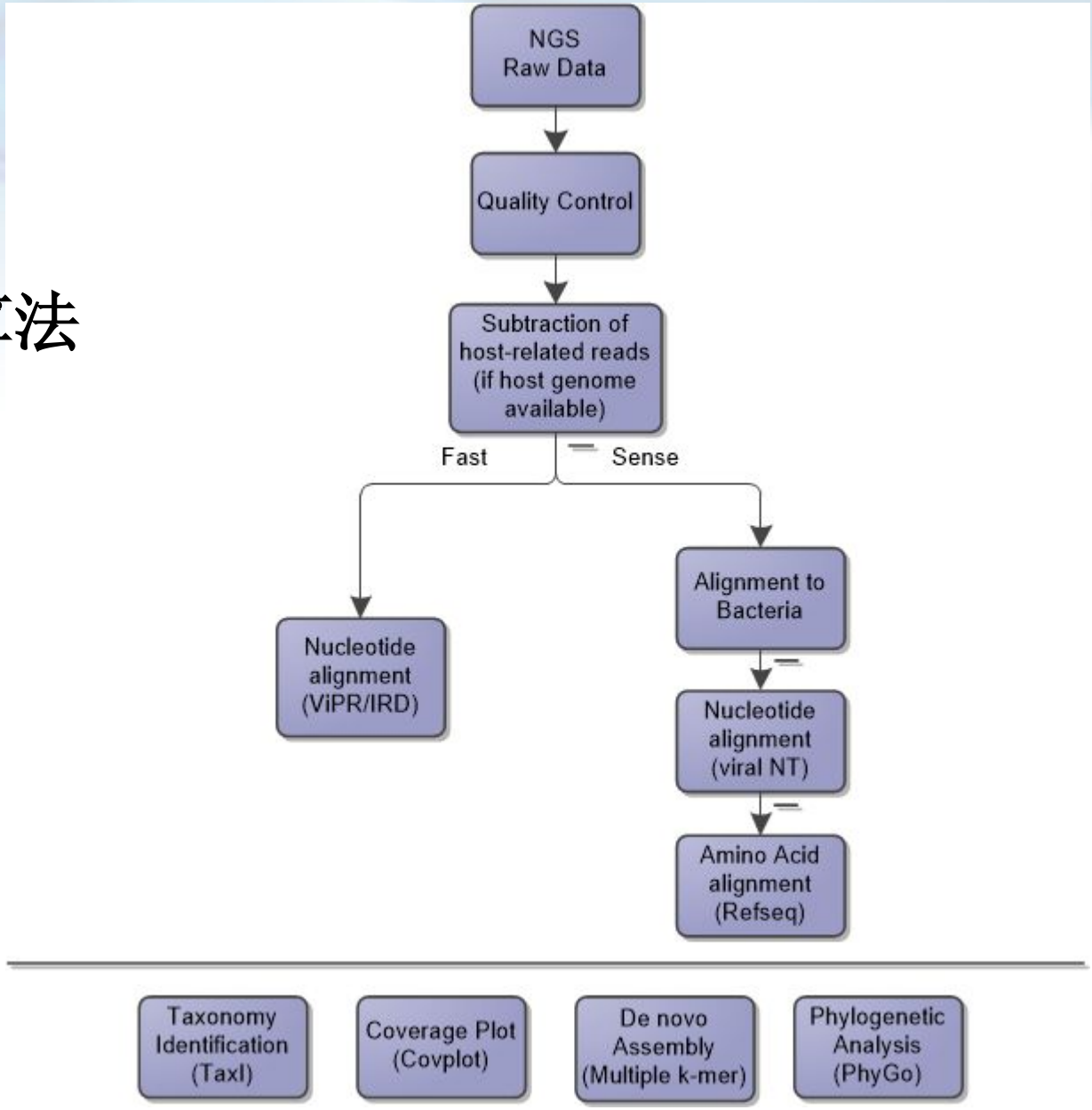
Select: All None Selected:0

Alignments Download GenBank Graphics Distance tree of results

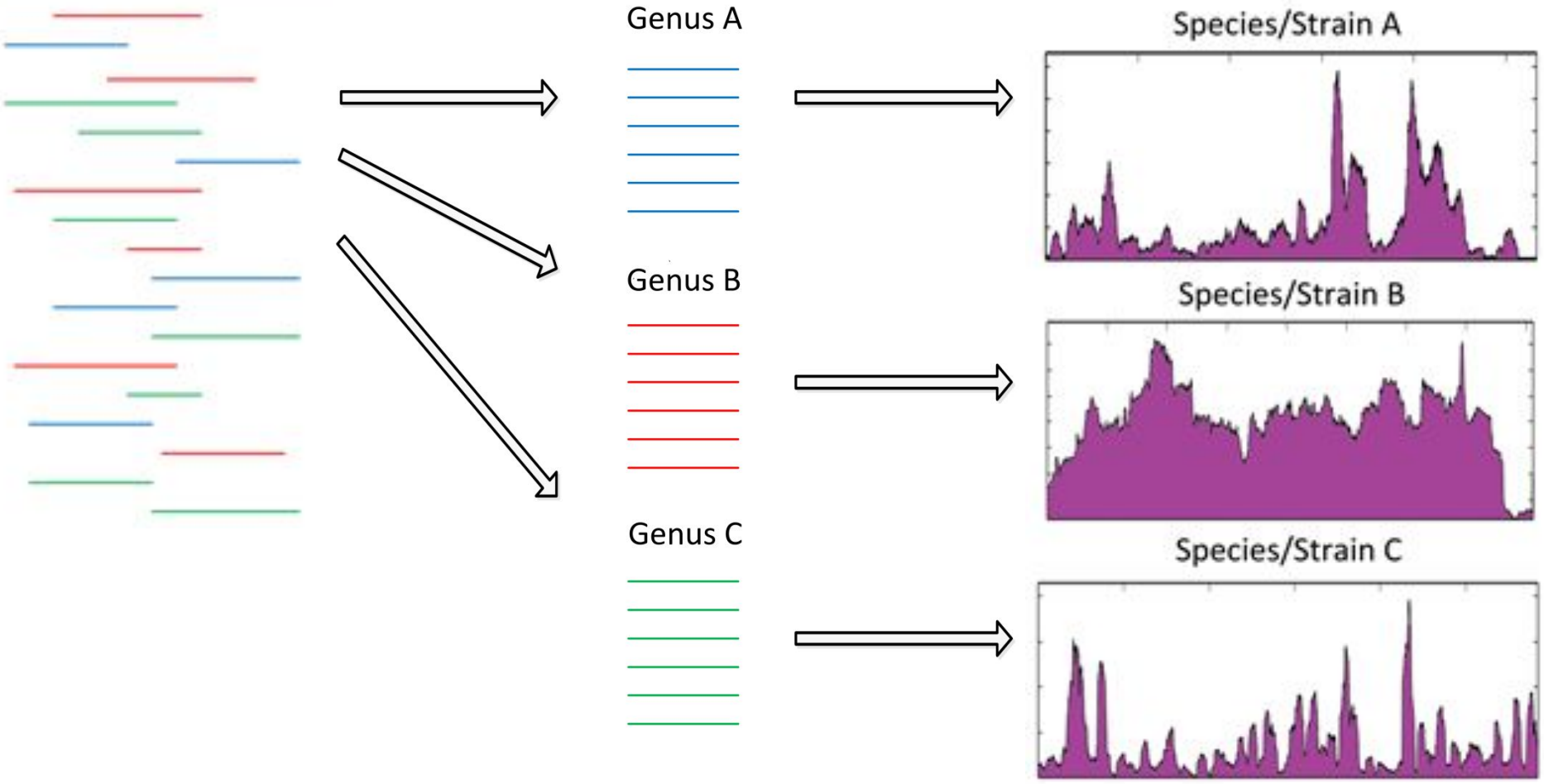
	Description	Max score	Total score	Query cover	E value	Ident	Accession
☐	Middle East respiratory syndrome coronavirus isolate Hafr-Al-Batin_1_2013, complete genome	55238	55238	99%	0.0	99%	KF600628.1
☐	Middle East respiratory syndrome coronavirus isolate Camel/Qatar_2_2014, complete genome	55197	55197	99%	0.0	99%	KJ650098.1
☐	Middle East respiratory syndrome coronavirus strain Florida/USA-2_Saudi Arabia_2014, complete genome	55184	55184	100%	0.0	99%	KJ829365.1
☐	Middle East respiratory syndrome coronavirus isolate Qatar4, complete genome	55182	55182	99%	0.0	99%	KF961222.1
☐	Middle East respiratory syndrome coronavirus isolate Florida/USA-2_Saudi Arabia_2014, complete genome	55173	55173	100%	0.0	99%	KP223131.1

VIP流程示意图

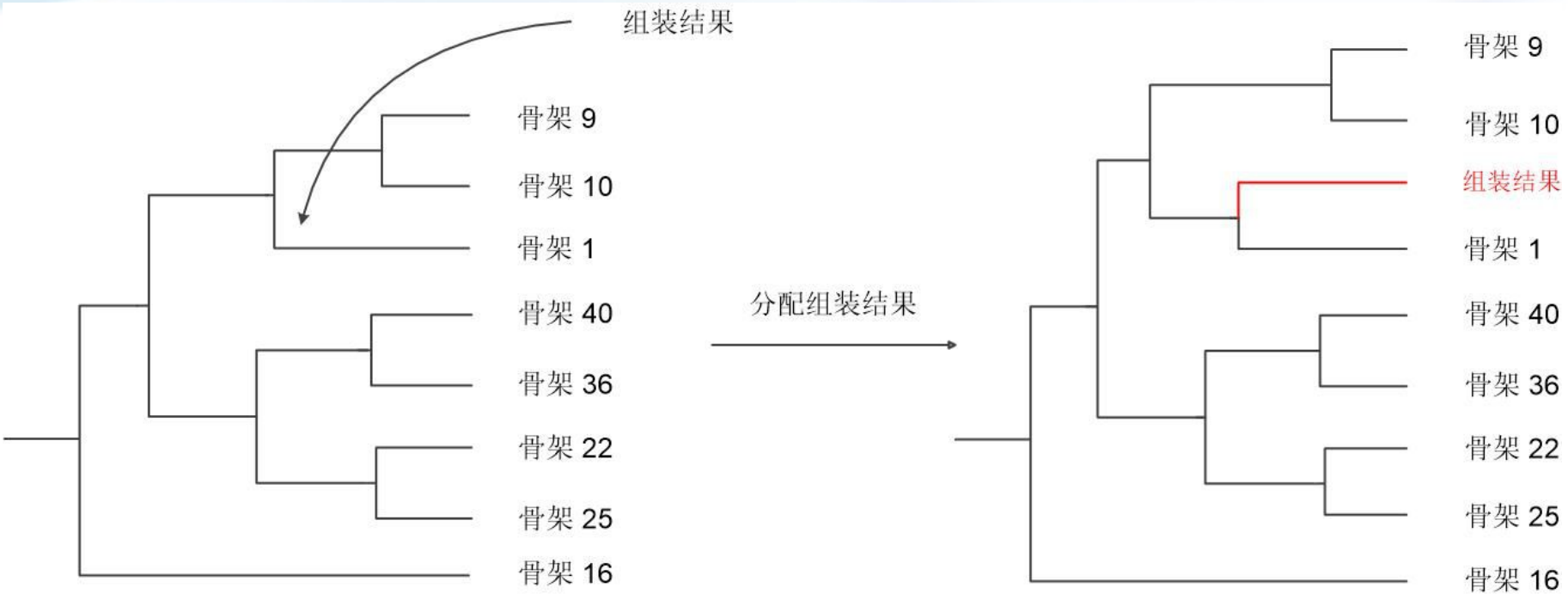
- 比对算法/分类算法
- 参考数据库
- 结果信息？



分类算法示意图


Genus A
Genus B
Genus C
Species/Strain A
Species/Strain B
Species/Strain C

进化算法示意图


组装结果
骨架 9
骨架 10
骨架 1
骨架 40
骨架 36
骨架 22
骨架 25
骨架 16
分配组装结果
骨架 9
骨架 10
组装结果
骨架 1
骨架 40
骨架 36
骨架 22
骨架 25
骨架 16

总结

中心实验室：坚持是一种美德

Thank you