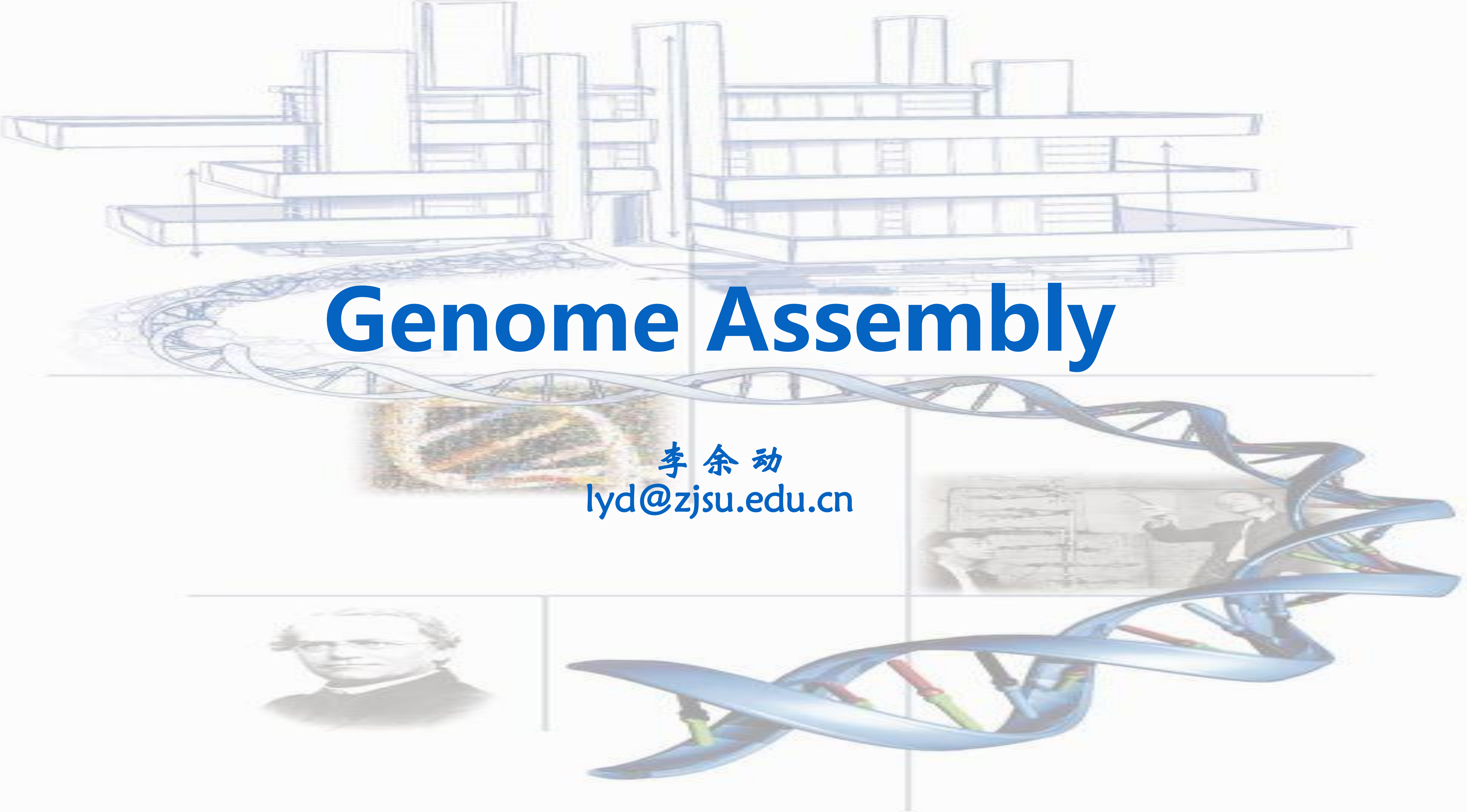


Genome Assembly

李 余 动
lyd@zjsu.edu.cn

目录

CONTENTS

●新冠病毒概述

●基因组测序策略

●基因组组装 (Genome Assembly)

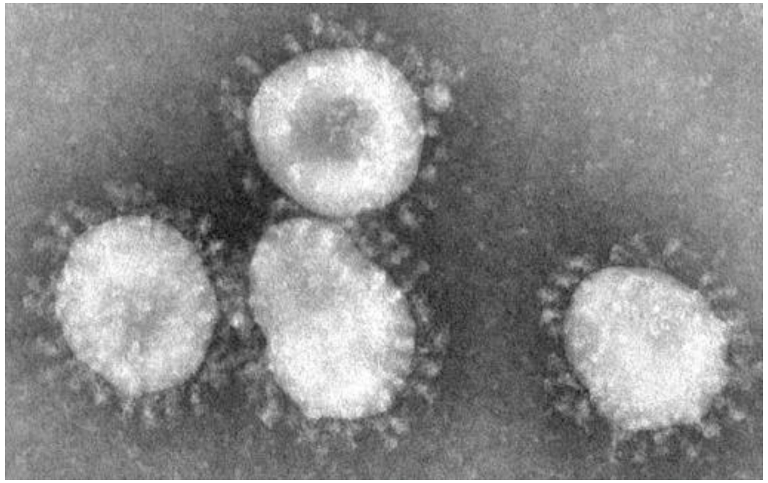
●案例：Sequencing of SARS-CoV-2

新冠病毒概述

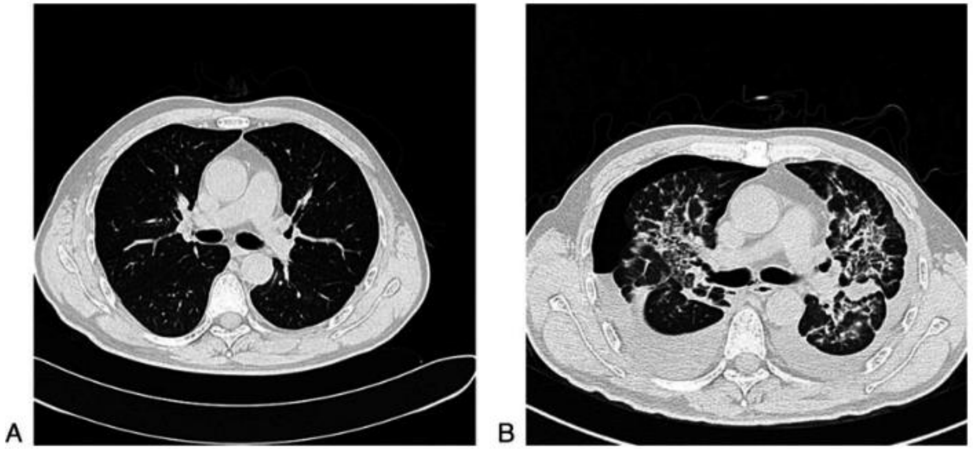
SARS and COVID-19

- 肺炎(pneumonia):
  - Severe Acute Respiratory Syndrome (SARS), 2003
  - Coronavirus disease 2019 (COVID-19)

- SARS vs. COVID-19
  - SARS was more lethal than COVID-19, 10% of people infected with SARS died.
  - There were no **asymptomatic** SARS patients. however, COVID-19 patients sometimes without symptoms, and spread faster.



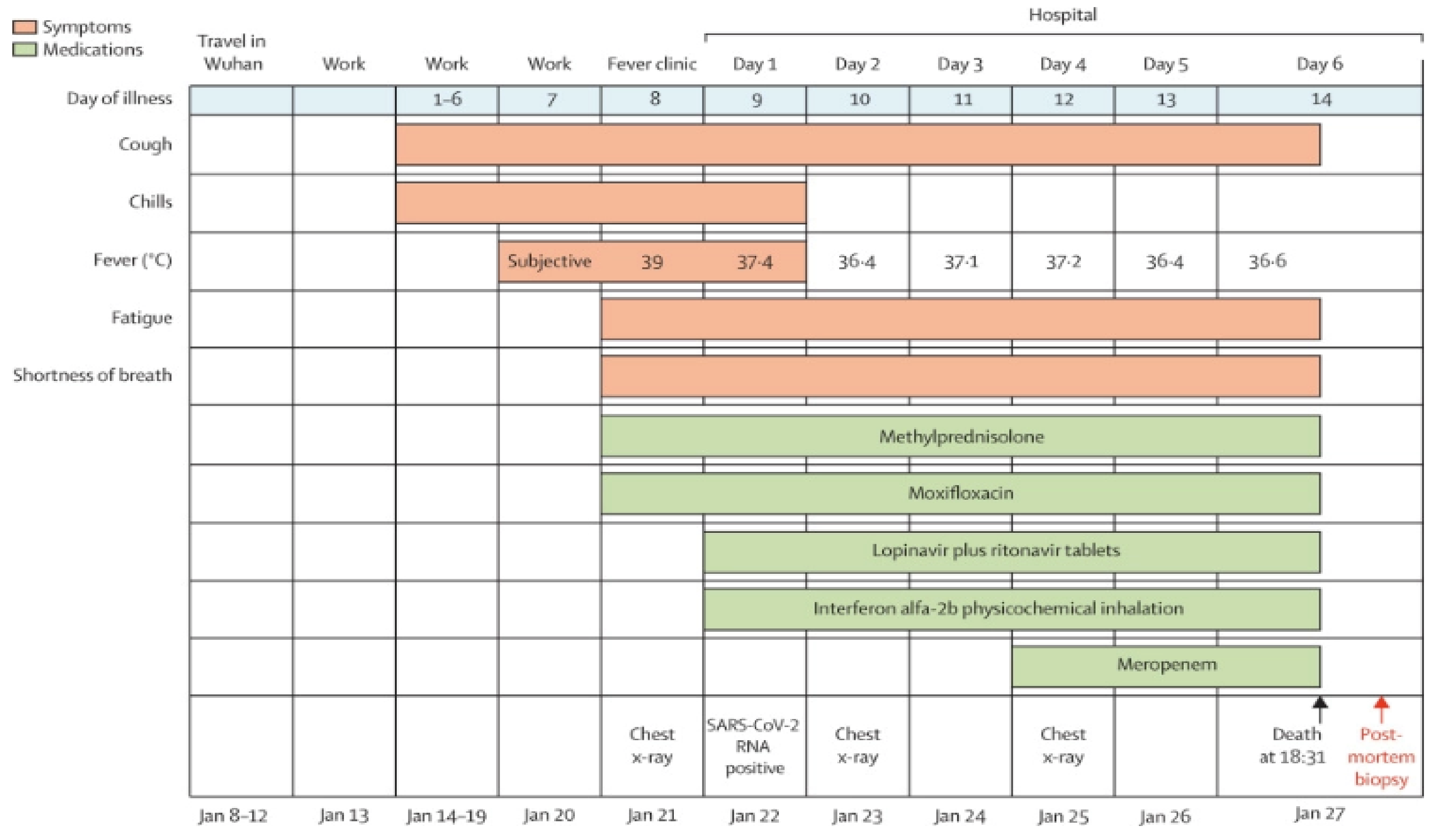
Coronavirus particles



A. Normal lung B. Lung with ground glass opacity

A COVID-19 Patient

- A 50-year-old man was admitted to a clinic in Hong Kong on Jan 21, 2020.
- Symptoms of fever, chills, cough, fatigue and shortness of breath.
- He traveled to Wuhan Jan 8–12.
- Initial symptoms of mild chills and dry cough on Jan 14 (day 1 of illness).
- Did not see a doctor and kept working until January 21.
- Chest x-ray showed multiple patchy shadows in both lungs.
- On January 22 (day 9 of illness), confirmed by reverse real-time PCR assay that the patient had COVID-19.



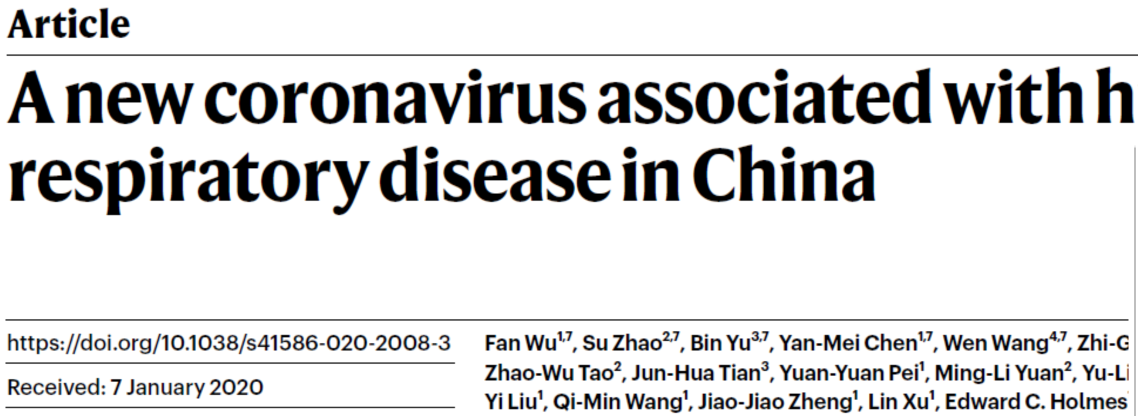
🟧 Symptoms
🟩 Medications

						Hospital					
	Travel in Wuhan	Work	Work	Work	Fever clinic	Day 1	Day 2	Day 3	Day 4	Day 5	Day 6
Day of illness			1–6	7	8	9	10	11	12	13	14
Cough			🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧
Chills			🟧	🟧	🟧	🟧					
Fever (°C)				🟧 Subjective	🟧 39	🟧 37·4	36·4	37·1	37·2	36·4	36·6
Fatigue					🟧	🟧	🟧	🟧	🟧	🟧	🟧
Shortness of breath					🟧	🟧	🟧	🟧	🟧	🟧	🟧
Methylprednisolone					🟩	🟩	🟩	🟩	🟩	🟩	🟩
Moxifloxacin					🟩	🟩	🟩	🟩	🟩	🟩	🟩
Lopinavir plus ritonavir tablets						🟩	🟩	🟩	🟩	🟩	🟩
Interferon alfa-2b physicochemical inhalation						🟩	🟩	🟩	🟩	🟩	🟩
Meropenem									🟩	🟩	🟩
					Chest x-ray	SARS-CoV-2 RNA positive	Chest x-ray		Chest x-ray		Death at 18:31 Post-mortem biopsy
	Jan 8–12	Jan 13	Jan 14–19	Jan 20	Jan 21	Jan 22	Jan 23	Jan 24	Jan 25	Jan 26	Jan 27

Figure 1 Timeline of disease course according to days from initial presentation of illness and days from hospital admission, from Jan 8–27, 2020

Xu Z, Shi L, Wang Y, Zhang J, Huang L, Zhang C et al. 2020. A Case Report: Pathological findings of COVID-19 associated with acute respiratory distress syndrome. Lancet Respiratory Medicine 8:420–22, https://doi.org/10.1016/S2213-2600(20)30076-X

Sequencing approaches for SARS-CoV-2



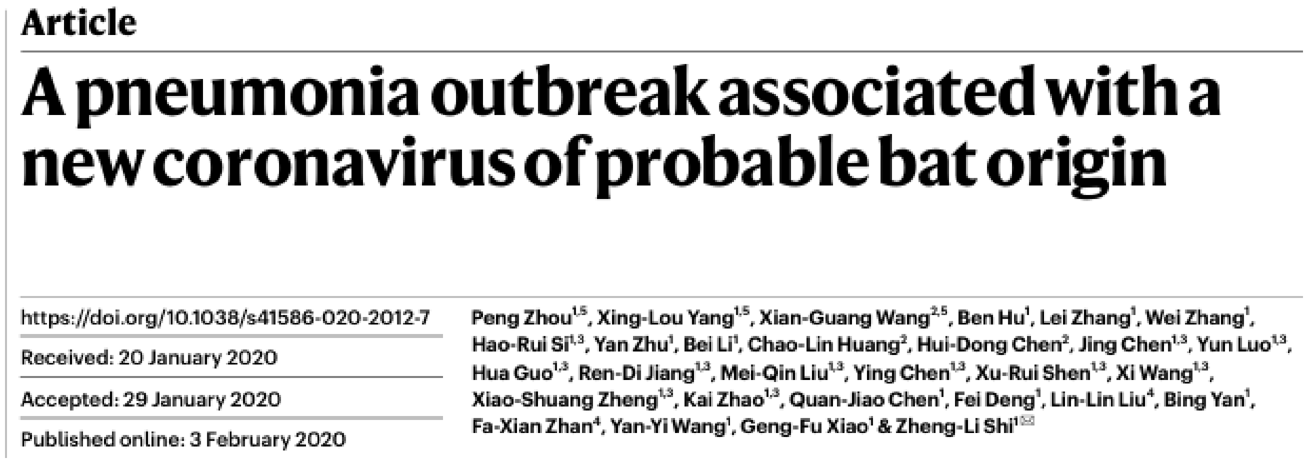
Article

A new coronavirus associated with human respiratory disease in China

https://doi.org/10.1038/s41586-020-2008-3

Received: 7 January 2020

Fan Wu[1,7], Su Zhao[2,7], Bin Yu[3,7], Yan-Mei Chen[1,7], Wen Wang[4,7], Zhi-G
Zhao-Wu Tao[2], Jun-Hua Tian[3], Yuan-Yuan Pei[1], Ming-Li Yuan[2], Yu-Li
Yi Liu[1], Qi-Min Wang[1], Jiao-Jiao Zheng[1], Lin Xu[1], Edward C. Holmes



Article

A pneumonia outbreak associated with a new coronavirus of probable bat origin

https://doi.org/10.1038/s41586-020-2012-7

Received: 20 January 2020

Accepted: 29 January 2020

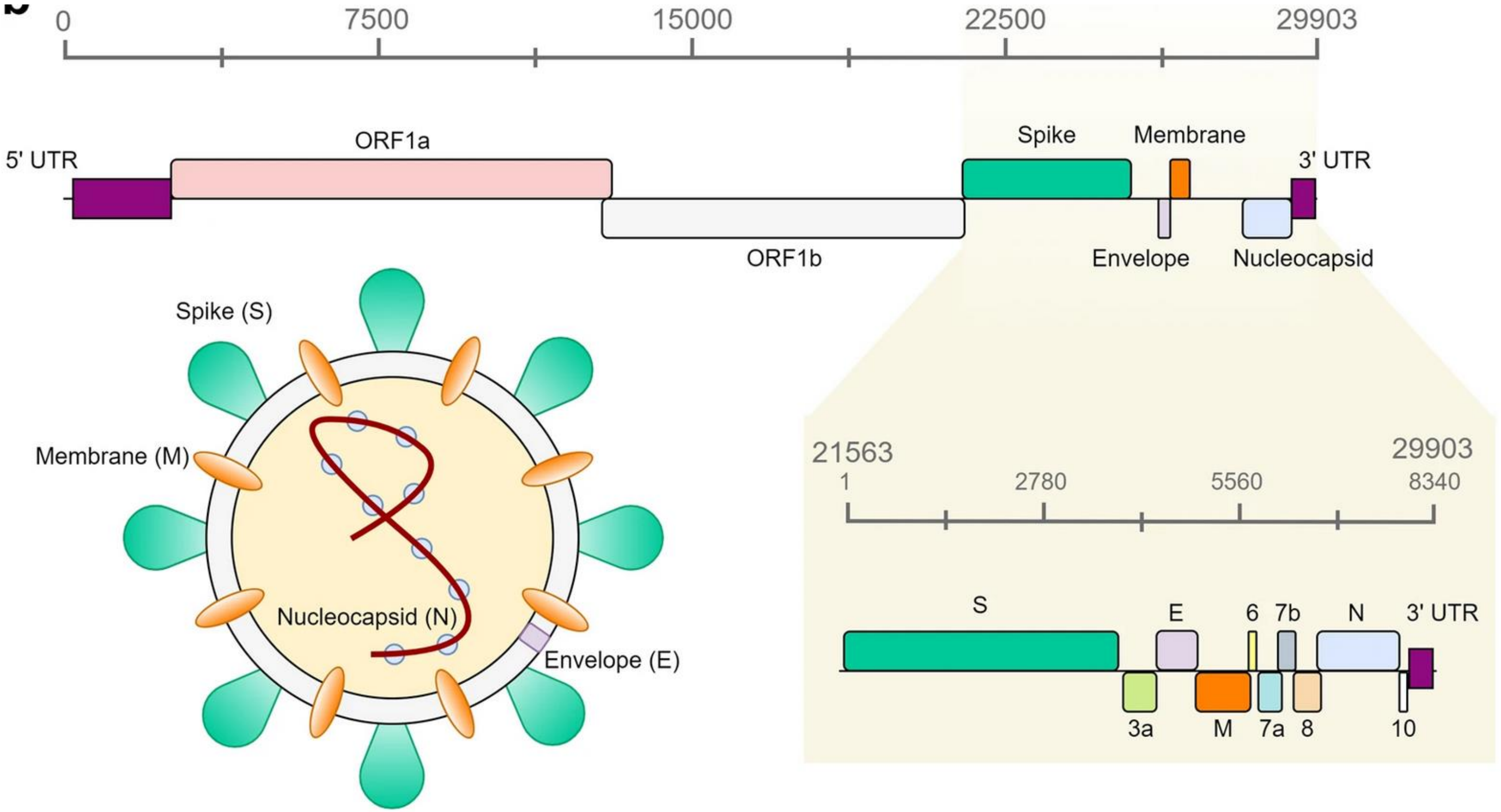
Published online: 3 February 2020

Peng Zhou[1,5], Xing-Lou Yang[1,5], Xian-Guang Wang[2,5], Ben Hu[1], Lei Zhang[1], Wei Zhang[1], Hao-Rui Si[1,3], Yan Zhu[1], Bei Li[1], Chao-Lin Huang[2], Hui-Dong Chen[2], Jing Chen[1,3], Yun Luo[1,3], Hua Guo[1,3], Ren-Di Jiang[1,3], Mei-Qin Liu[1,3], Ying Chen[1,3], Xu-Rui Shen[1,3], Xi Wang[1,3], Xiao-Shuang Zheng[1,3], Kai Zhao[1,3], Quan-Jiao Chen[1], Fei Deng[1], Lin-Lin Liu[4], Bing Yan[1], Fa-Xian Zhan[4], Yan-Yi Wang[1], Geng-Fu Xiao[1] & Zheng-Li Shi[1]✉

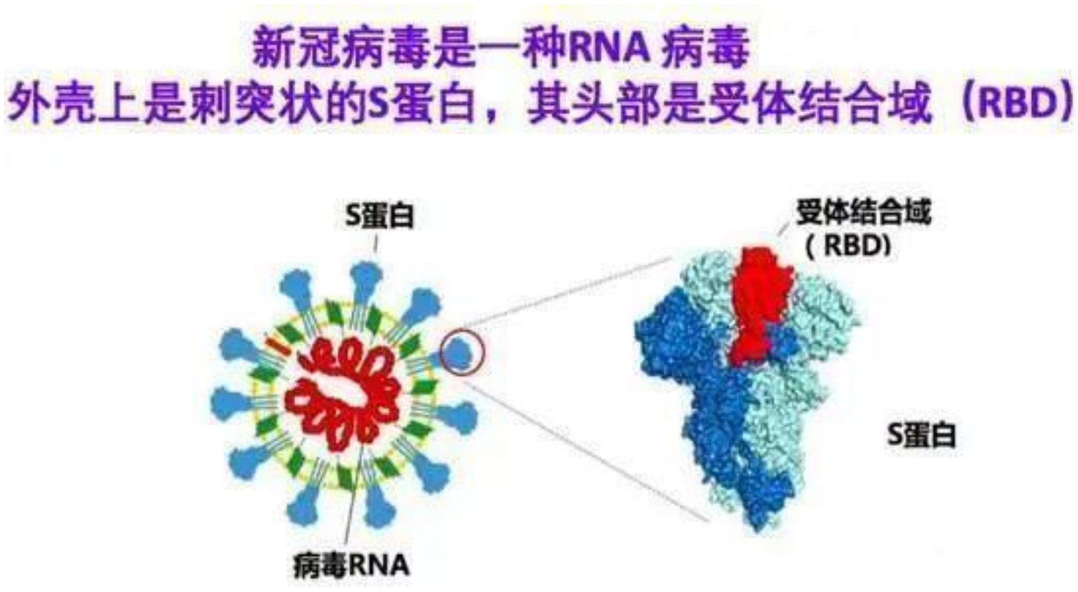
- 2020年初复旦大学与武汉病毒所研究团队分别对患者支气管肺泡灌洗液进行了**mNGS测序**，鉴定出了一株新型冠状病毒（Coronavirus）。
- 根据病毒基因组的系统发育分析发现，该病毒基因组与蝙蝠体内发现的**类严重急性呼吸综合征（SARS-like）**冠状病毒基因组的相似性高达89.1%。
- 新冠病毒的全基因组序列对新冠疫情的防疫工作具有重大意义，如设计测序引物开展核酸检测试剂盒开发，或根据新冠刺穿蛋白的基因序列进行疫苗研发等。

The Coronavirus genome

- RNA virus (single-stranded, positive-sense)
- Linear genome = ~30,000 nucleotides
- 11 coding-regions (genes)
- 12 potential gene productions

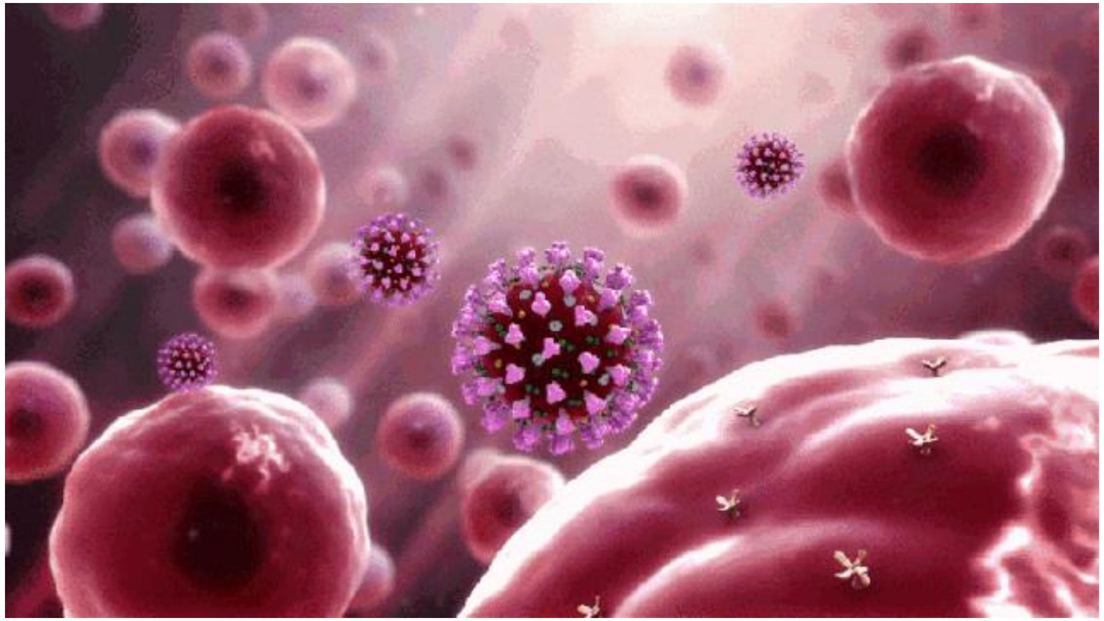


A CLOSE UP OF THE VIRUS



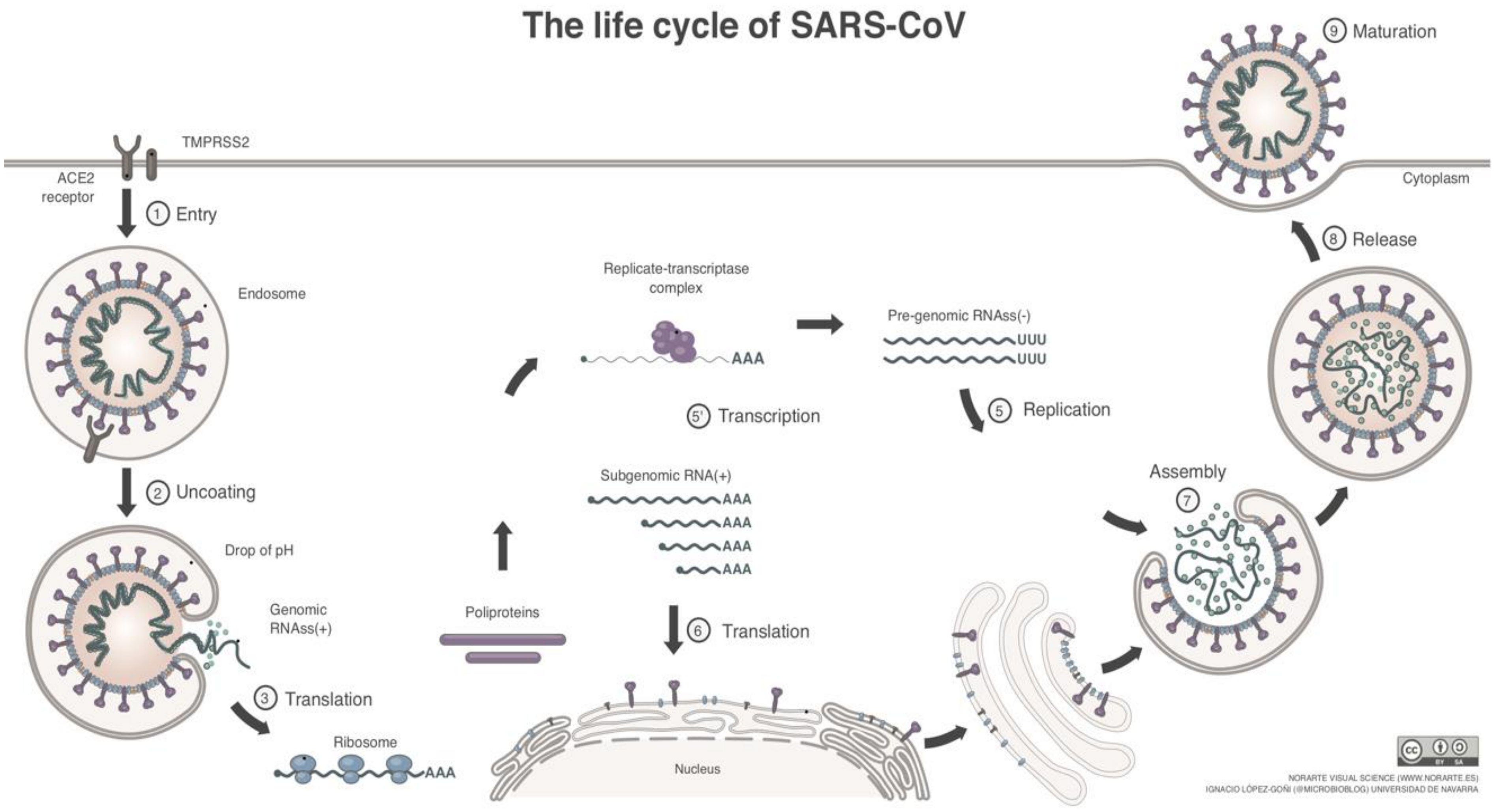
新冠病毒是一种RNA 病毒
外壳上是刺突状的S蛋白，其头部是受体结合域（RBD）

Viral spikes target angiotensin converting enzyme 2 (ACE2) receptors—found in lung pneumocytes II (for one).



The Life cycle of SARS-CoV-2

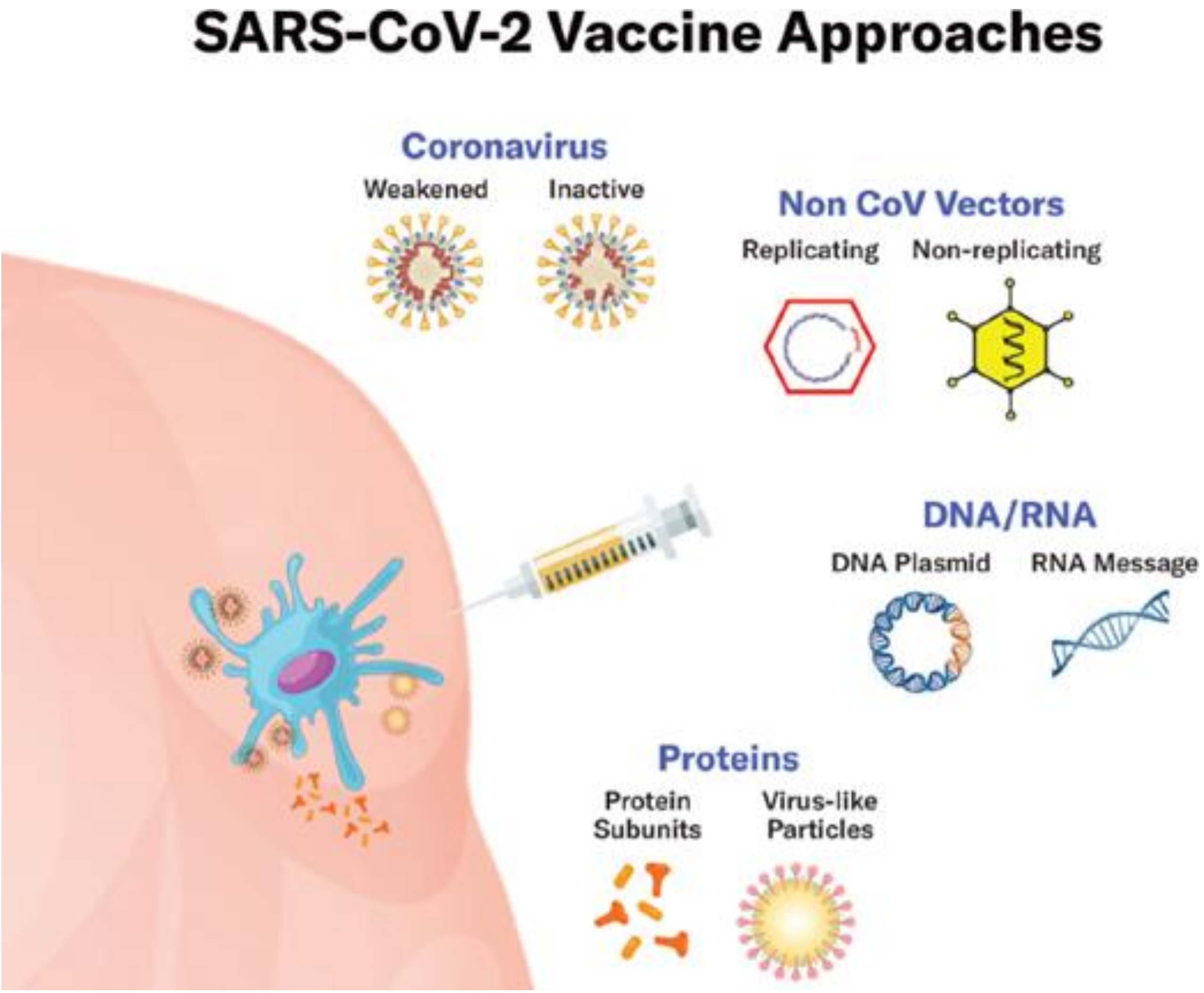
- Polyprotein
- Protease



https://upload.wikimedia.org/wikipedia/commons/c/ca/SARS-CoV-2_ciclo.svg

Virus proteins and vaccine

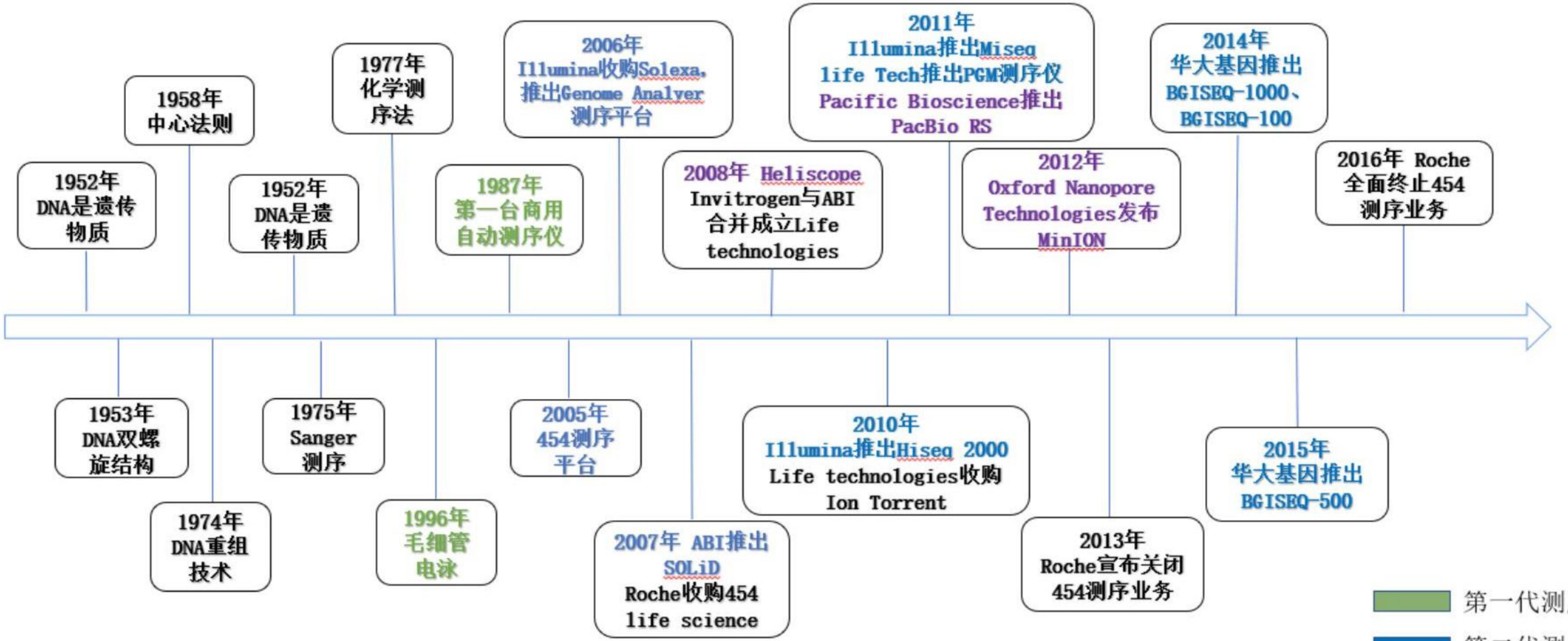
疫苗是病原体的一部分蛋白质(也叫**抗原**)，可刺激机体产生**抗体**。



- 灭活疫苗
- 减毒疫苗
- 亚单位疫苗
- 核酸疫苗
  - RNA vaccine (Pfizer, Moderna)

基因组测序策略

DNA测序技术发展历程



1952年
DNA是遗传
物质

1953年
DNA双螺
旋结构

1958年
中心法则

1974年
DNA重组
技术

1952年
DNA是遗
传物质

1975年
Sanger
测序

1977年
化学测
序法

1996年
毛细管
电泳

1987年
第一台商用
自动测序仪

2005年
454测序
平台

2006年
Illumina收购Solexa,
推出Genome Analver
测序平台

2007年 ABI推出
SOLiD
Roche收购454
life science

2008年 Heliscope
Invitrogen与ABI
合并成立Life
technologies

2010年
Illumina推出Hiseq 2000
Life technologies收购
Ion Torrent

2011年
Illumina推出Miseq
life Tech推出PGM测序仪
Pacific Bioscience推出
PacBio RS

2012年
Oxford Nanopore
Technologies发布
MinION

2013年
Roche宣布关闭
454测序业务

2014年
华大基因推出
BGISEQ-1000、
BGISEQ-100

2015年
华大基因推出
BGISEQ-500

2016年 Roche
全面终止454
测序业务

🟩 第一代测序

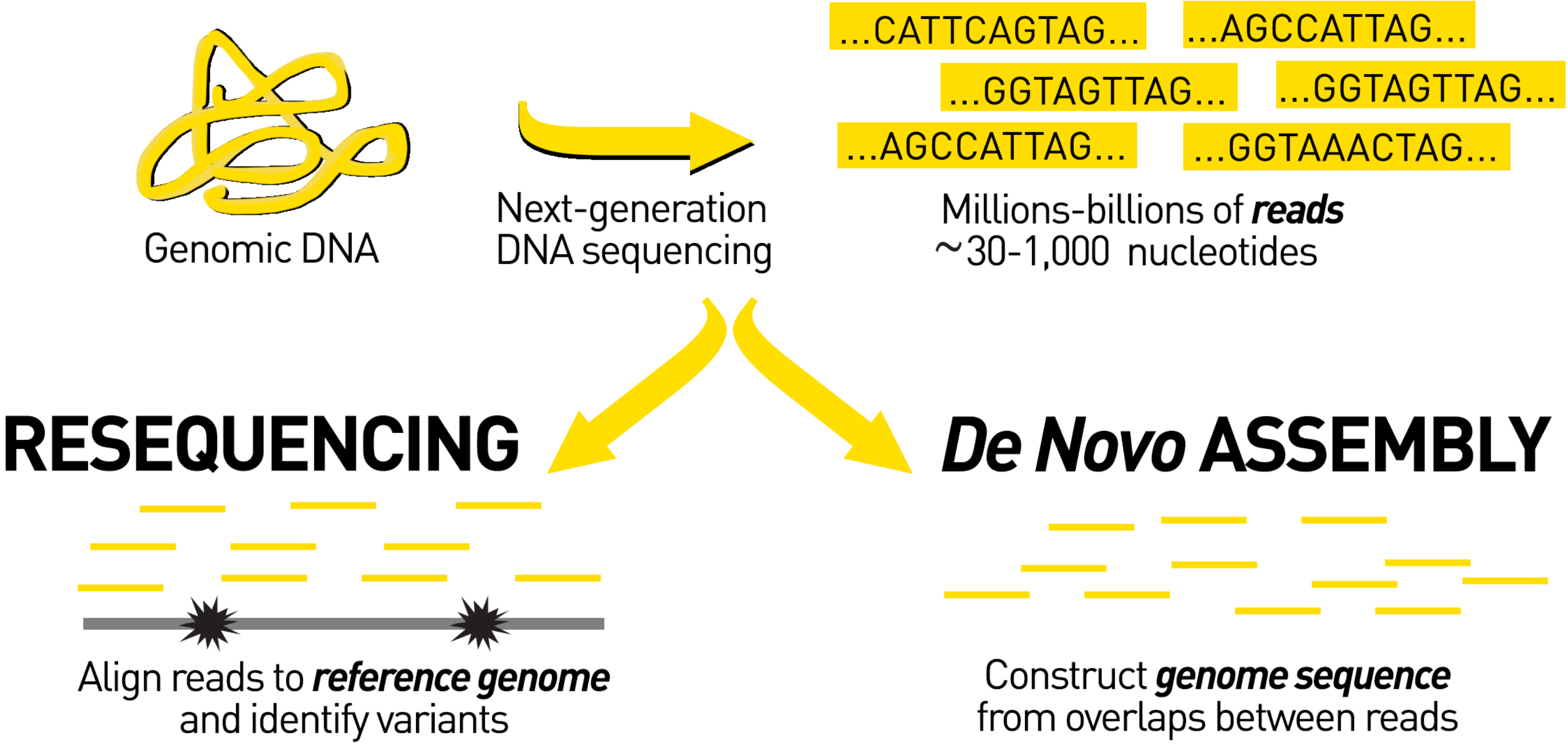
🟦 第二代测序

🟪 第三代测序

	一代	二代	三代
并行度	低<400	高>10000	
PCR	需要		不需要

全基因组测序(Whole Genome Sequencing)

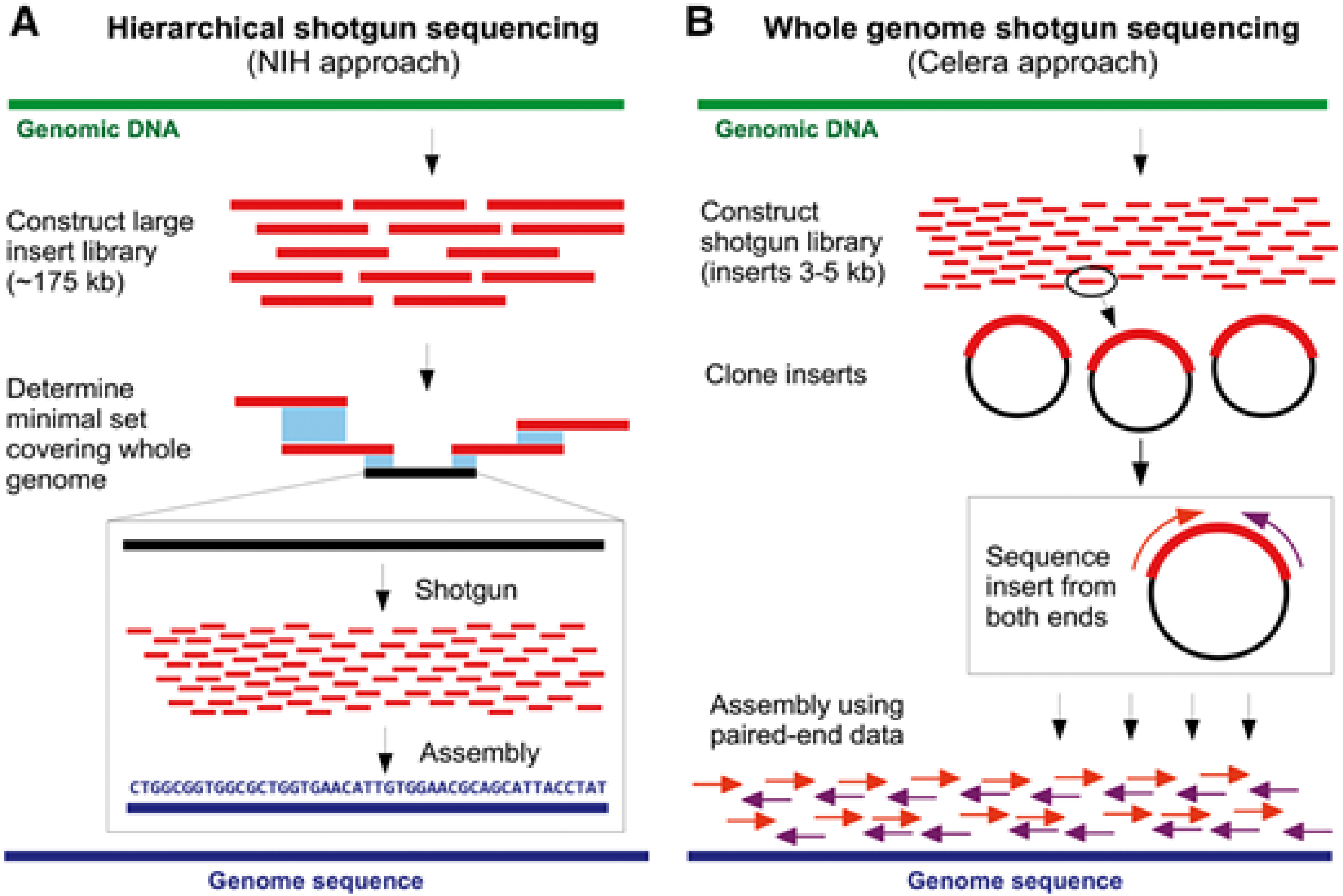
- 重测序(RESEQUENCING):将测序reads比对到参考基因组，并鉴定遗传差异。
- 从头组装(*De Novo* Assembly): 将测序reads准确地拼接成基因组。



...CATTCAGTAG...
...AGCCATTAG...
...GGTAGTTAG...
...GGTAGTTAG...
...AGCCATTAG...
...GGTAAACTAG...
Genomic DNA
Next-generation DNA sequencing
Millions-billions of *reads*
~30-1,000 nucleotides
RESEQUENCING
De Novo ASSEMBLY
Align reads to ***reference genome*** and identify variants
Construct ***genome sequence*** from overlaps between reads

基因组测序两种策略

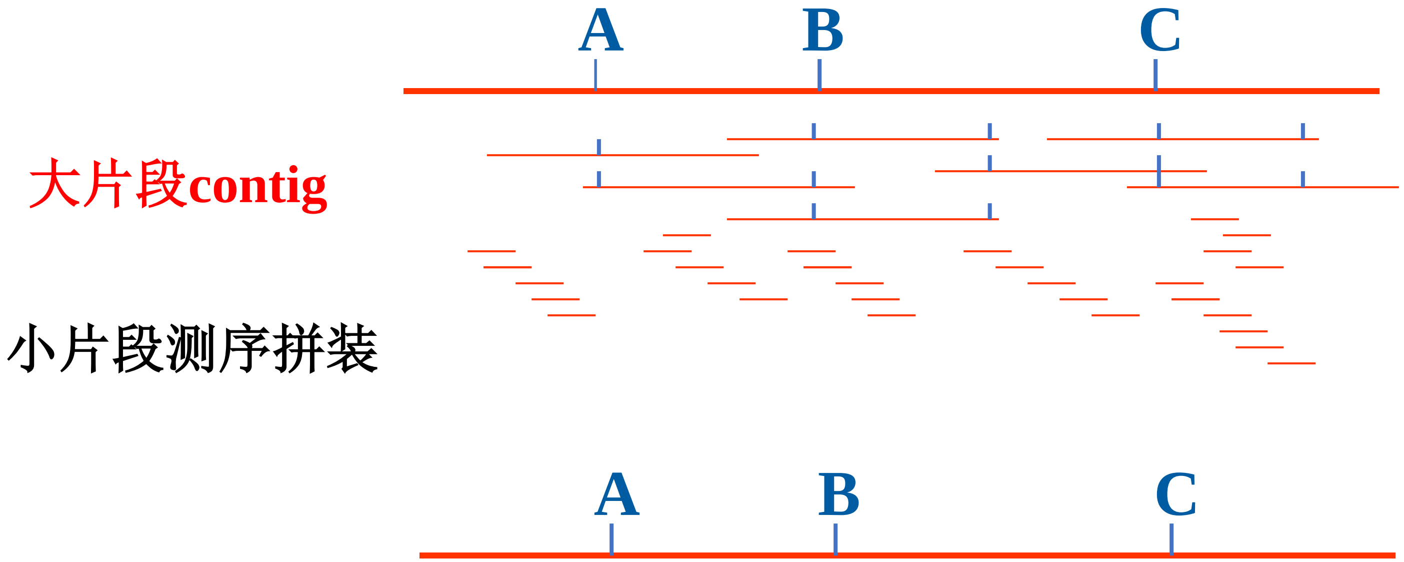
- 逐步克隆法(A)
- 全基因组鸟枪法(B)


A
Hierarchical shotgun sequencing
(NIH approach)
Genomic DNA
Construct large insert library (~175 kb)
Determine minimal set covering whole genome
Shotgun
Assembly
CTGGCGGTGGCGCTGGTGAACATTGTGGAACGCAGCATTACCTAT
Genome sequence
B
Whole genome shotgun sequencing
(Celera approach)
Genomic DNA
Construct shotgun library (inserts 3-5 kb)
Clone inserts
Sequence insert from both ends
Assembly using paired-end data
Genome sequence

Gauthier, et al., 2018

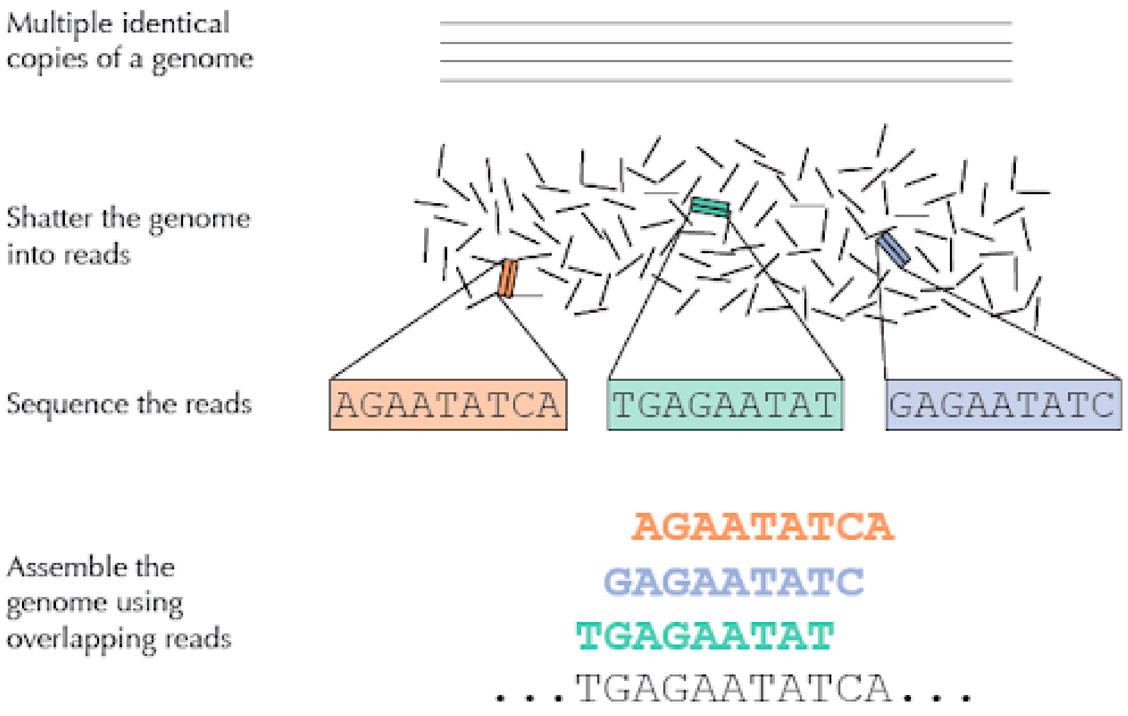
基于克隆群(contig-based)的策略

先将染色体打成比较大的片段(几十-几百Kb), 利用分子标记将这些大片段排成重叠的克隆群(Contig), 分别测序后拼装。



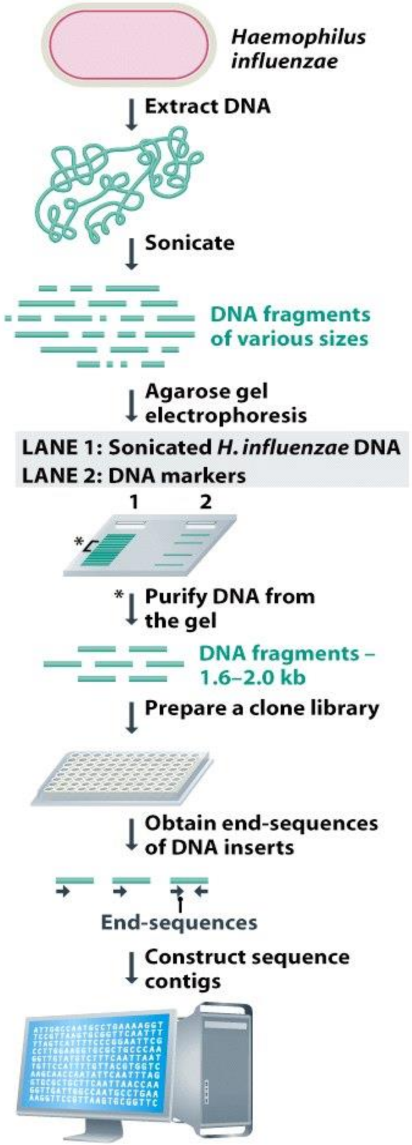
Shotgun测序策略—序列的组装

- 序列组装原理: 直接从已测序的小片段中寻找彼此重叠的测序克隆，然后依次向两侧邻接的序列延伸。



Multiple identical copies of a genome
Shatter the genome into reads
Sequence the reads
AGAATATCA
TGAGAATAT
GAGAATATC
Assemble the genome using overlapping reads
AGAATATCA
GAGAATATC
TGAGAATAT
...TGAGAATATCA...

实例: 流感嗜血杆菌基因组的测序及序列组装


Haemophilus influenzae
Extract DNA
Sonicate
DNA fragments of various sizes
Agarose gel electrophoresis
LANE 1: Sonicated *H. influenzae* DNA
LANE 2: DNA markers
1
2
*
Purify DNA from the gel
DNA fragments – 1.6–2.0 kb
Prepare a clone library
Obtain end-sequences of DNA inserts
End-sequences
Construct sequence contigs

超声波打断纯化的基因组DNA

↓

琼脂糖电泳收集1.6~2.0Kb的区段、纯化

↓

构建到质粒载体中

↓

随机挑选19687个克隆,进行28643次测序,得到可读顺序为11 631 485 bp

↓

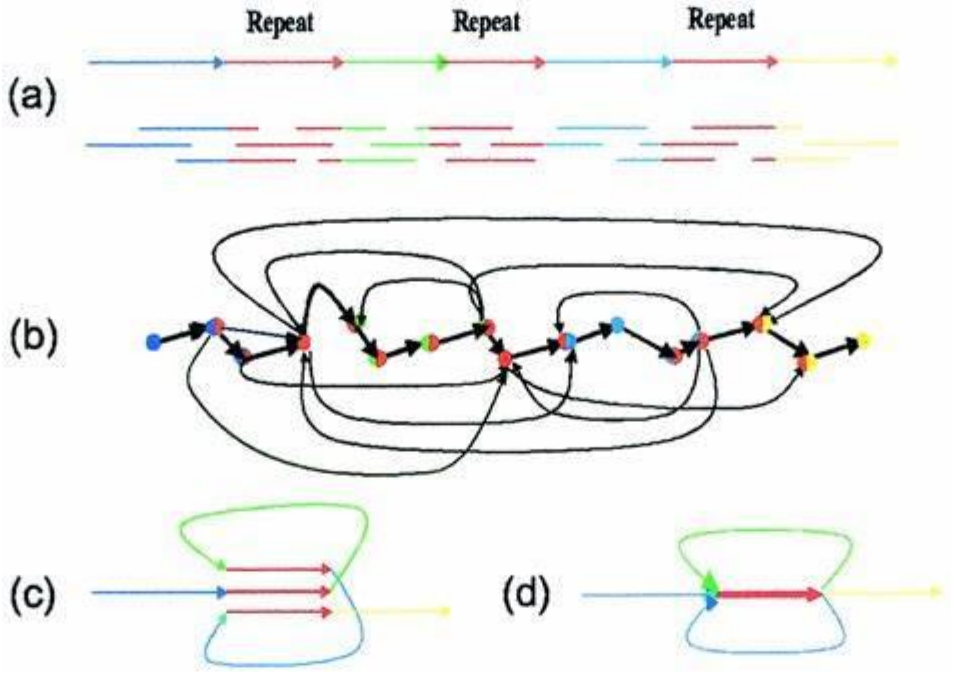
组装成140个覆盖全基因组范围的独立的顺序重叠群,

↓

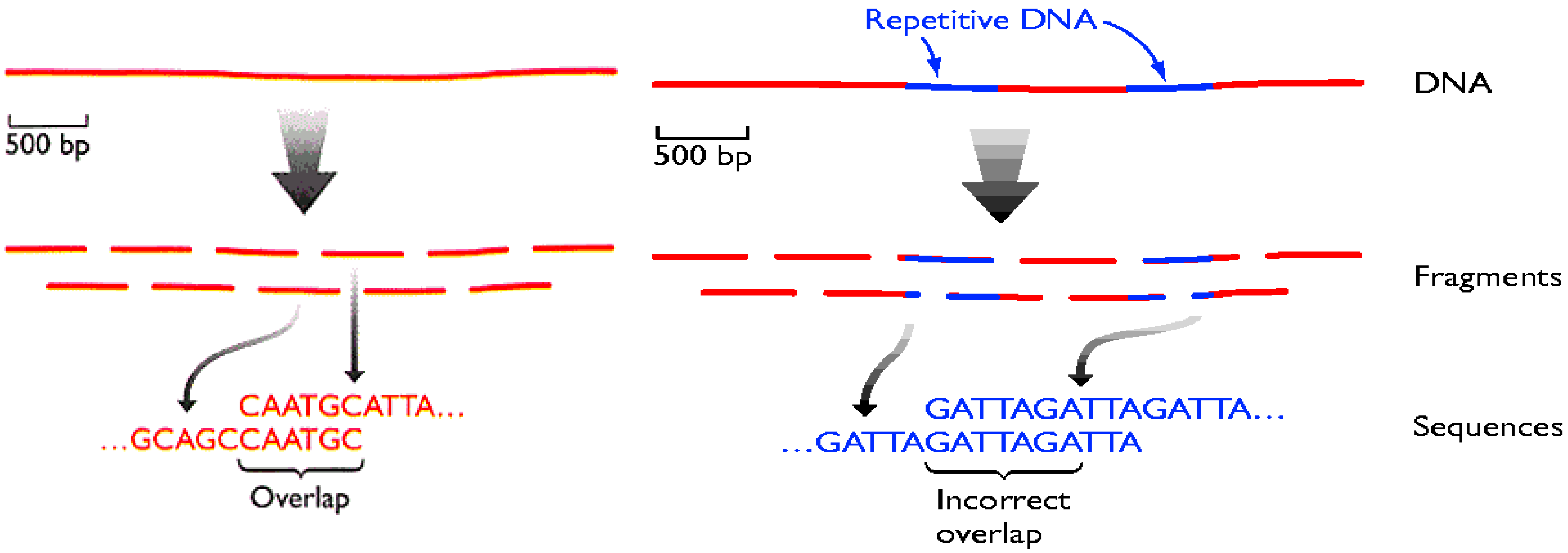
……完整基因组

Shotgun测序—拼接问题？

- 因为整个基因组太长（>1M)，而每次只能测得一个500的小片断(read)
- 问题：如何根据read恢复原始顺序?
- 类比：10本圣经，都从随机点起始剪成500个字母左右的小纸条，问：给你这么一堆小纸条，你能读出圣经来吗?
- 转成图论问题：Hamilton和Euler路径
- 但是都会拼错!



拼接错误：Repeat的存在



基因组重复序列测序策略

二代双端测序可通过构建长文库，跨过一个大片段序列的两末端。

双端测序Paired-End Sequencing

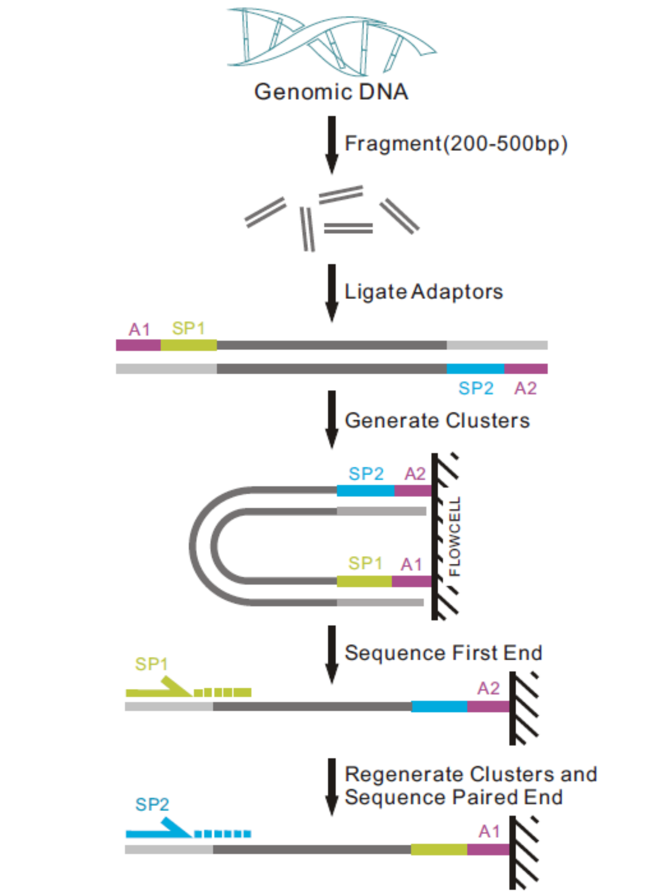
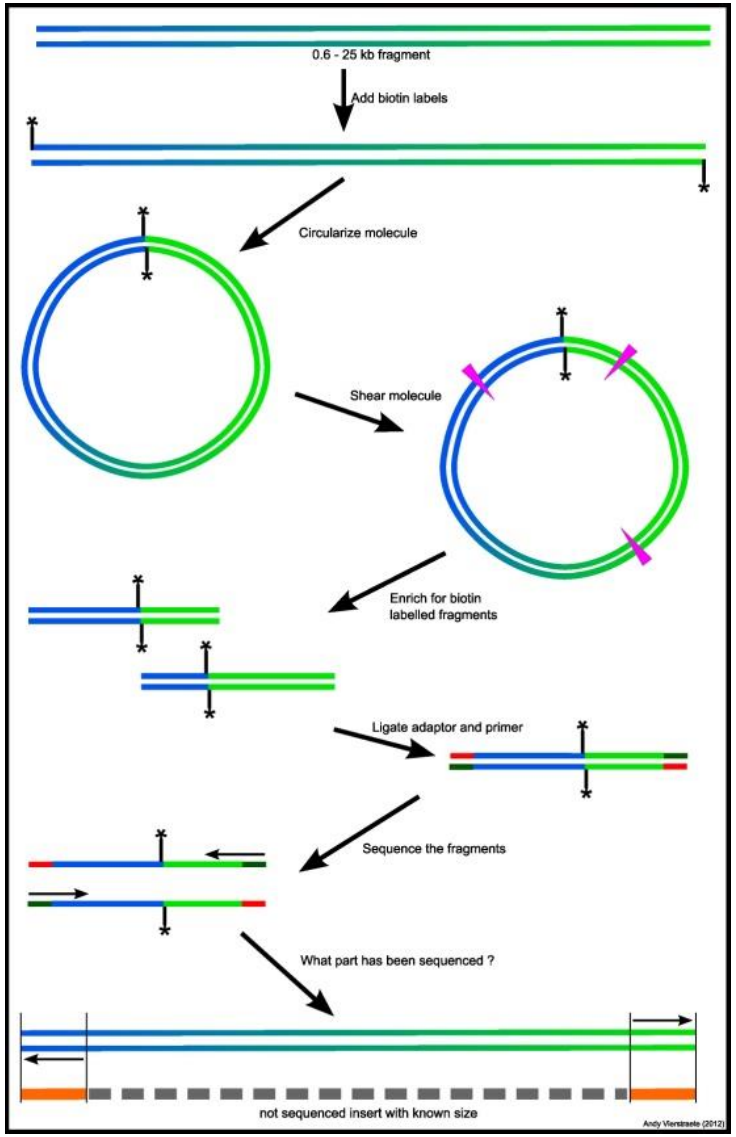


Figure 1-2-1 Pipeline of paired-end sequencing (www.illumina.com)

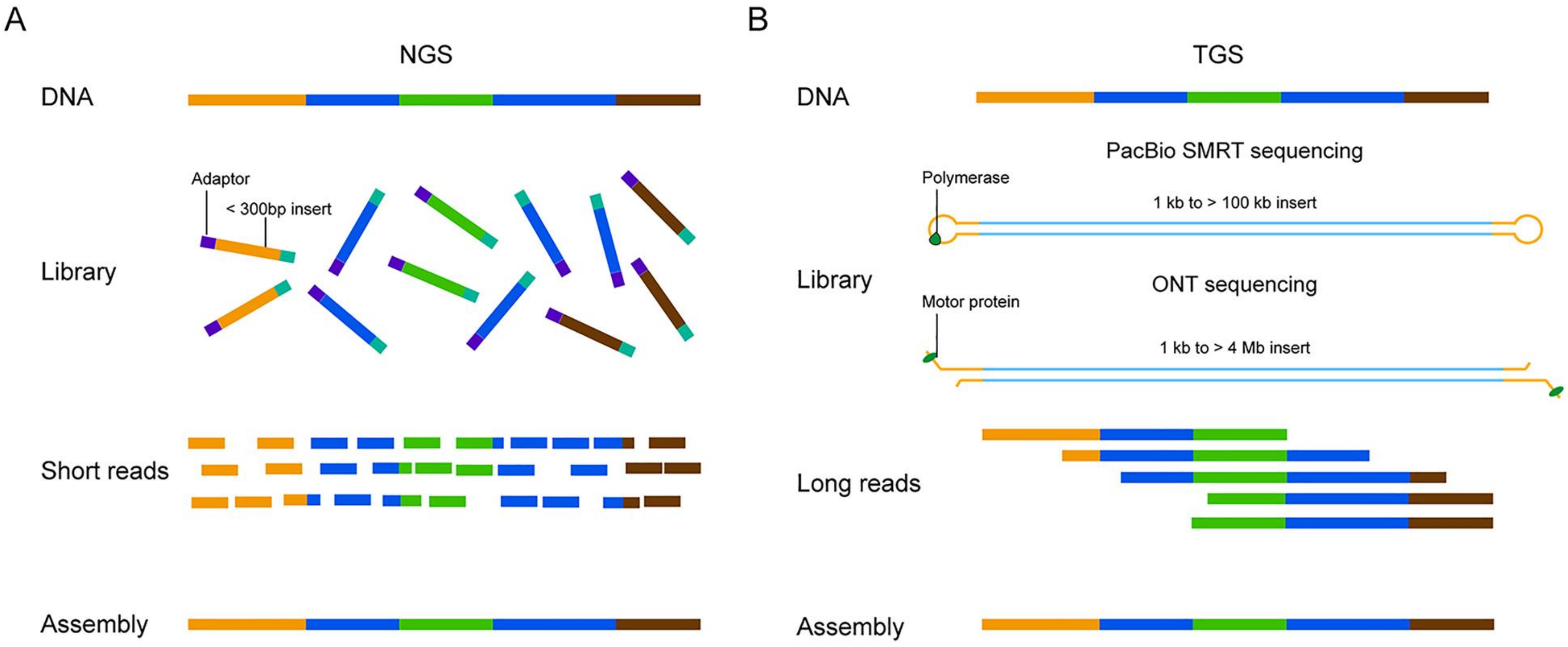
环化测序Mate-Pair Sequencing



0.6 - 25 kb fragment
Add biotin labels
Circularize molecule
Shear molecule
Enrich for biotin labelled fragments
Ligate adaptor and primer
Sequence the fragments
What part has been sequenced ?
not sequenced insert with known size
Andy Vierstraete (2012)

基因组重复序列测序策略

随着三代测序技术的进步，基因组测序一般采用二代三代组合测序策略

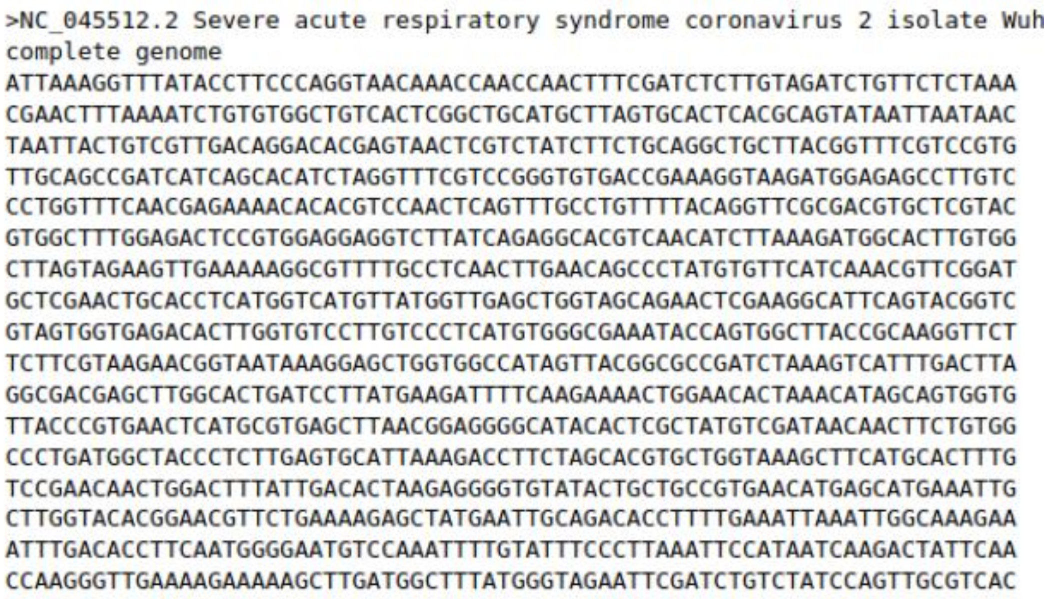


Chen Z, He X. Application of third-generation sequencing in cancer research. Medical Review. 2021 Dec 1;1(2):150–71.

Genome Assembly

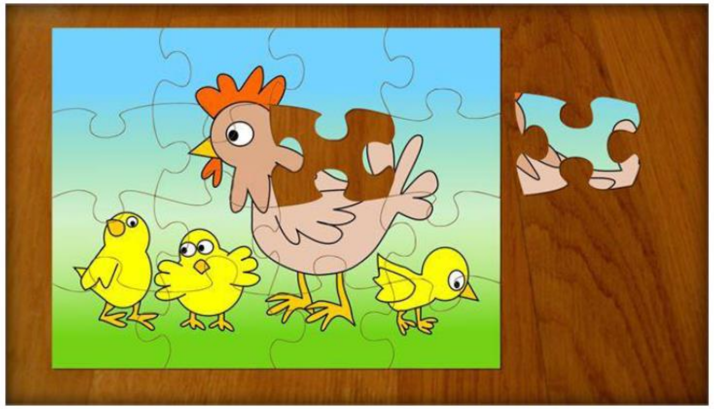
De novo Assembly

- 从头组装：把短序列组装起来，拼出一条完整的测序序列
- 从头组装类似拼图
  - The bigger the pieces the easier to reconstruct the puzzle!



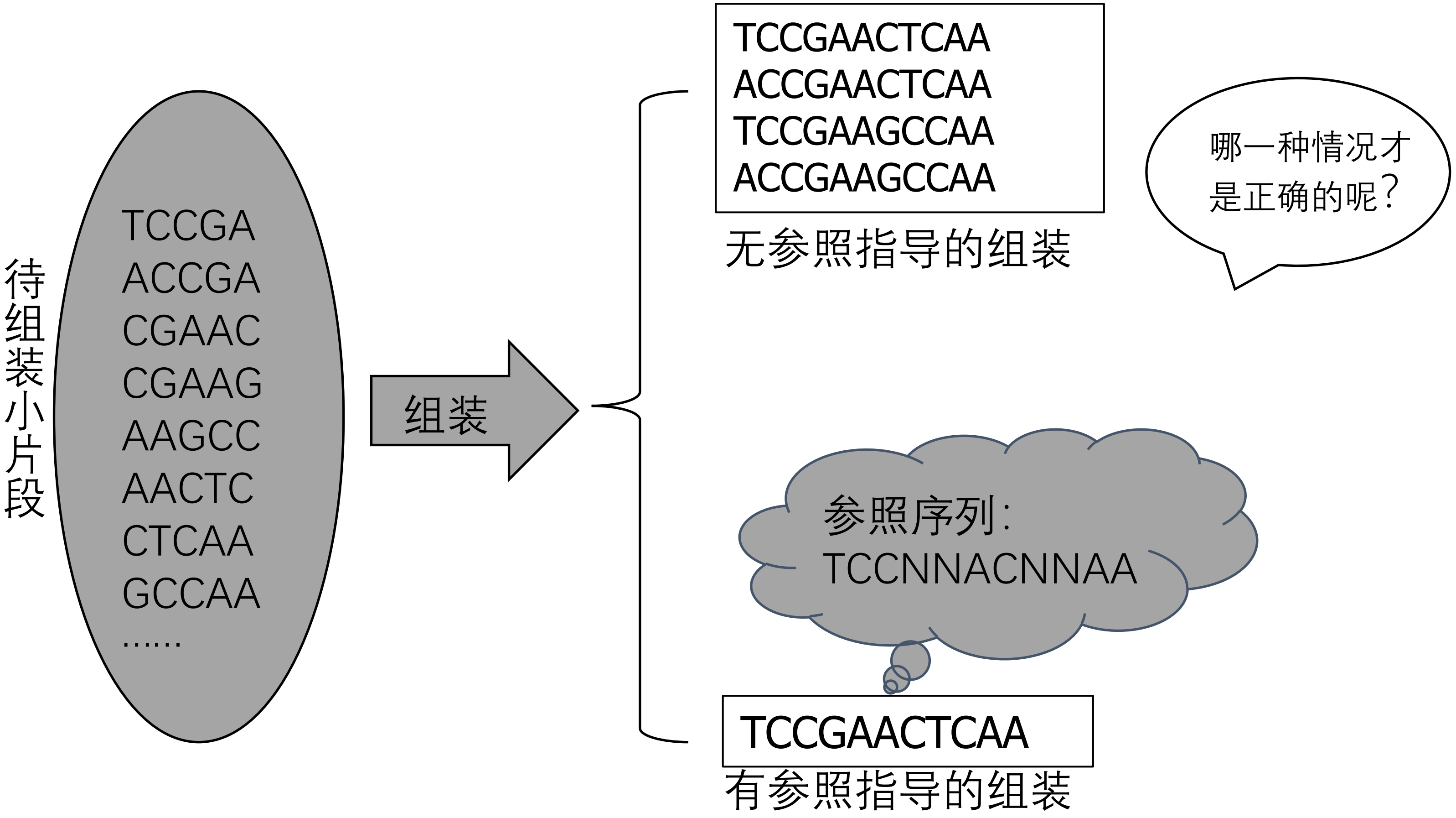
```
>NC_045512.2 Severe acute respiratory syndrome coronavirus 2 isolate Wuh
complete genome
ATTAAAGGTTTATACCTTCCCAGGTAACAAACCAACCAACTTTCGATCTCTTGTAGATCTGTTCTCTAAA
CGAACTTTAAAATCTGTGTGGCTGTCACTCGGCTGCATGCTTAGTGCACTCACGCAGTATAATTAATAAC
TAATTACTGTCGTTGACAGGACACGAGTAACTCGTCTATCTTCTGCAGGCTGCTTACGGTTTCGTCCGTG
TTGCAGCCGATCATCAGCACATCTAGGTTTCGTCCGGGTGTGACCGAAAGGTAAGATGGAGAGCCTTGTC
CCTGGTTTCAACGAGAAAACACACGTCCAACTCAGTTTGCCTGTTTTACAGGTTCGCGACGTGCTCGTAC
GTGGCTTTGGAGACTCCGTGGAGGAGGTCTTATCAGAGGCACGTCAACATCTTAAAGATGGCACTTGTGG
CTTAGTAGAAGTTGAAAAAGGCGTTTTGCCTCAACTTGAACAGCCCTATGTGTTCATCAAACGTTCGGAT
GCTCGAACTGCACCTCATGGTCATGTTATGGTTGAGCTGGTAGCAGAACTCGAAGGCATTCAGTACGGTC
GTAGTGGTGAGACACTTGGTGTCCTTGTCCCTCATGTGGGCGAAATACCAGTGGCTTACCGCAAGGTTCT
TCTTCGTAAGAACGGTAATAAAGGAGCTGGTGGCCATAGTTACGGCGCCGATCTAAAGTCATTTGACTTA
GGCGACGAGCTTGGCACTGATCCTTATGAAGATTTTCAAGAAAACTGGAACACTAAACATAGCAGTGGTG
TTACCCGTGAACTCATGCGTGAGCTTAACGGAGGGGCATACACTCGCTATGTCGATAACAACTTCTGTGG
CCCTGATGGCTACCCTCTTGAGTGCATTAAAGACCTTCTAGCACGTGCTGGTAAAGCTTCATGCACTTTG
TCCGAACAACTGGACTTTATTGACACTAAGAGGGGTGTATACTGCTGCCGTGAACATGAGCATGAAATTG
CTTGGTACACGGAACGTTCTGAAAAGAGCTATGAATTGCAGACACCTTTTGAAATTAAATTGGCAAAGAA
ATTTGACACCTTCAATGGGGAATGTCCAAATTTTGTATTTCCCTTAAATTCCATAATCAAGACTATTCAA
CCAAGGGTTGAAAAGAAAAAGCTTGATGGCTTTATGGGTAGAATTCGATCTGTCTATCCAGTTGCGTCAC
```

.fasta consensus genome



基因组装方式

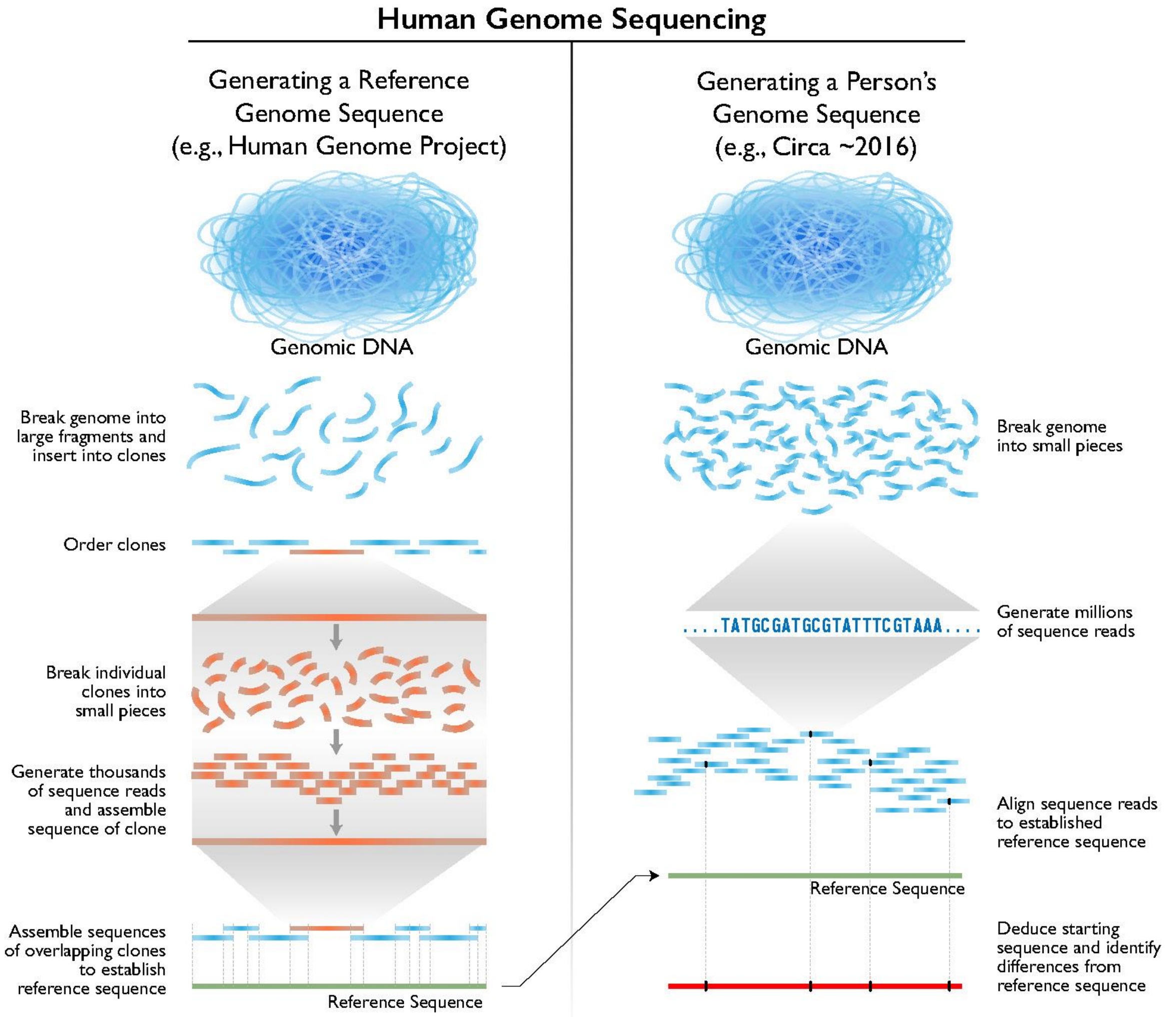
- *De novo* assembly: without reference sequences
  - 测序一个新的物种的基因组
- Reference-guided assembly: with reference sequences
  - 测序相近的物种（亚种）的基因组
  - 测序数据mapping到参考序列上，推测可能的基因组



参照序列用于指导序列片段组装

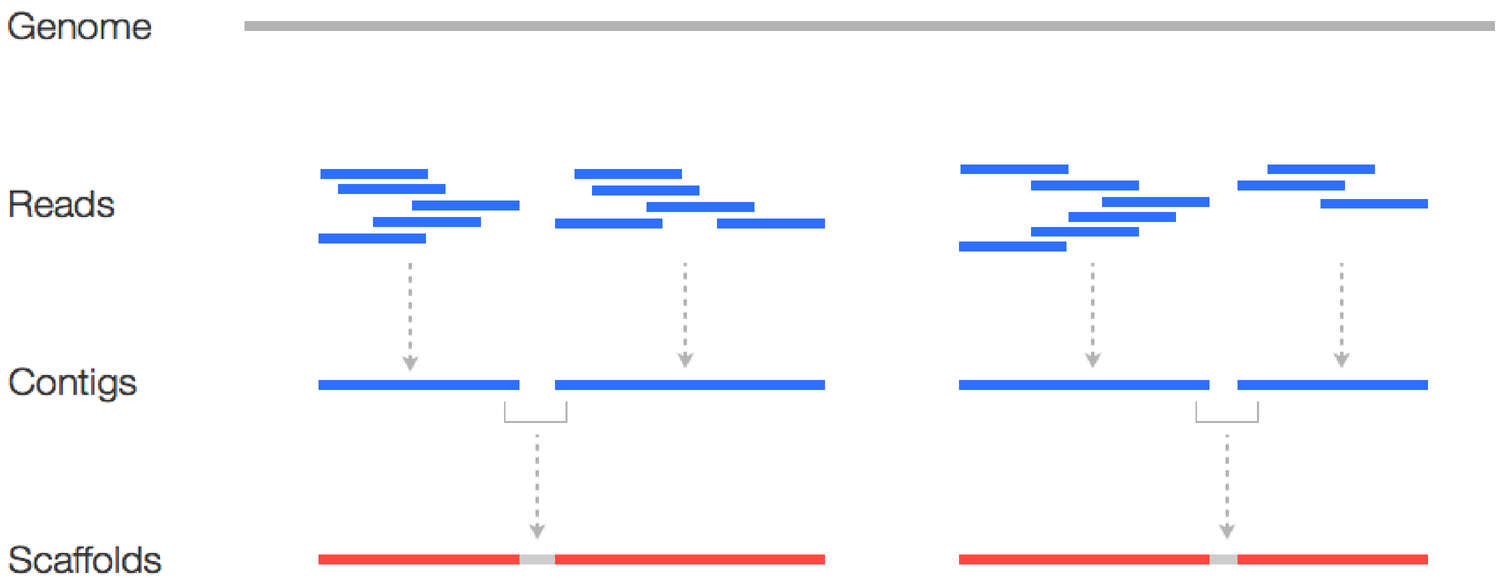
人类基因组测序方法

个人基因组测序


Human Genome Sequencing
Generating a Reference Genome Sequence (e.g., Human Genome Project)
Generating a Person's Genome Sequence (e.g., Circa ~2016)
Genomic DNA
Genomic DNA
Break genome into large fragments and insert into clones
Order clones
Break individual clones into small pieces
Generate thousands of sequence reads and assemble sequence of clone
Assemble sequences of overlapping clones to establish reference sequence
Reference Sequence
Break genome into small pieces
....TATGCGATGCGTATTTCGTAAA....
Generate millions of sequence reads
Align sequence reads to established reference sequence
Reference Sequence
Deduce starting sequence and identify differences from reference sequence

https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost

De novo assembly

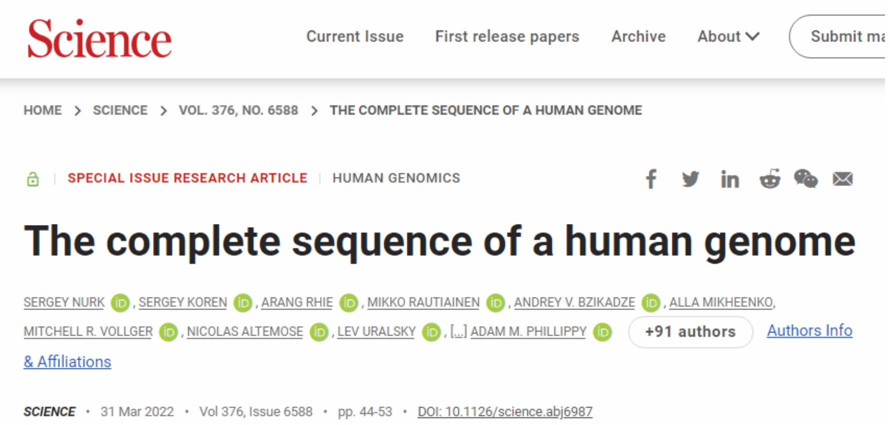


基因组组装的概念：

- Reads: 读段，即测序产生的短序列，通常一代的reads读长在一千左右，二代的reads相对较短，平均是几百bp，三代的reads较长在几千到几万bp之间；
- Contig: 重叠群，基于reads之间的overlap区域拼接成的连续序列。
- Scaffold: 获得contig按照一定顺序和方向组成scaffold，不同contig之间还有空缺(Gap)。

De novo Assembly：工作草图与完成图/精细图

- Gap filling and assembly polishing
- Draft genome (草图): 90%
- Finished genome (精细图): 99.99%



Science   Current Issue   First release papers   Archive   About   Submit ma

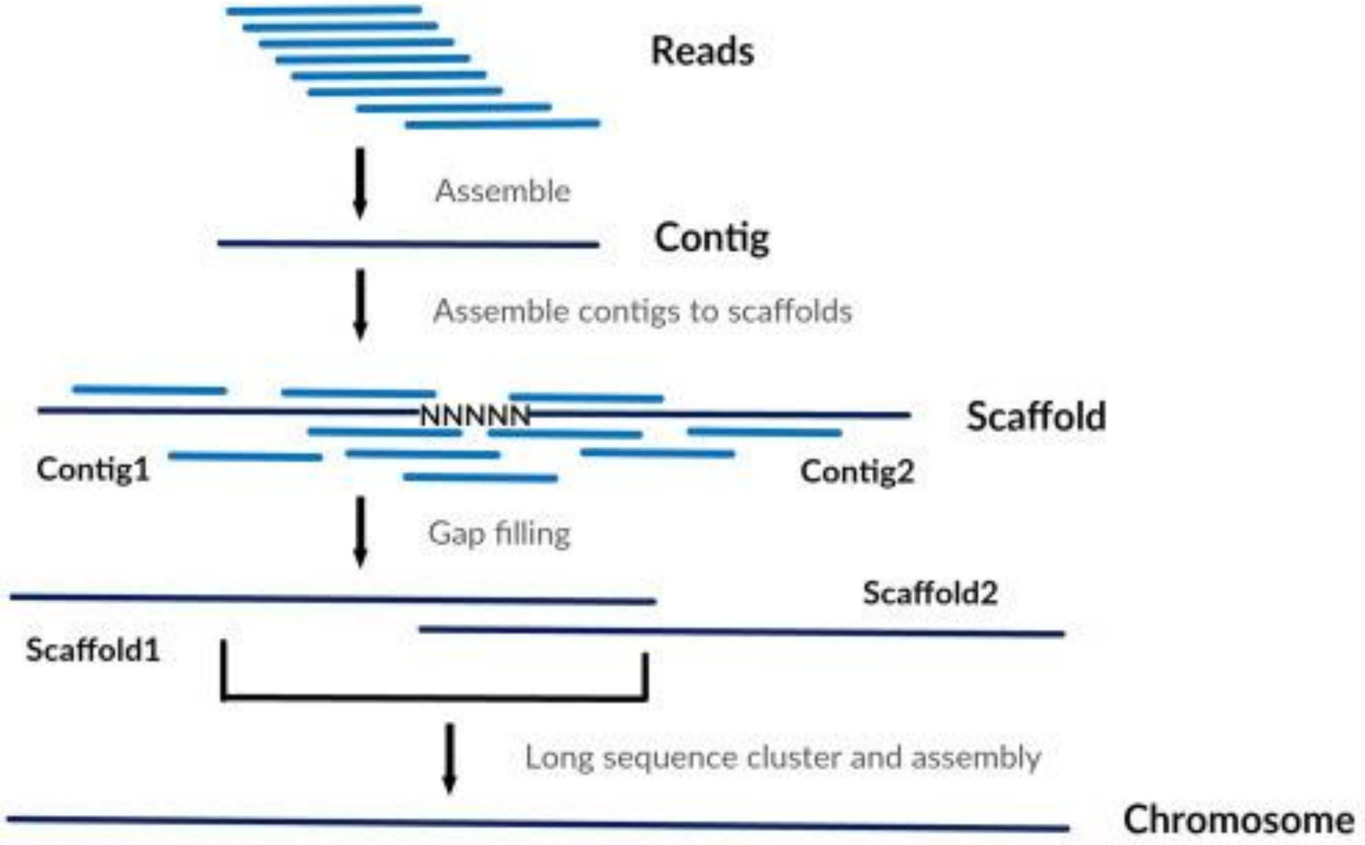
HOME > SCIENCE > VOL. 376, NO. 6588 > THE COMPLETE SEQUENCE OF A HUMAN GENOME

SPECIAL ISSUE RESEARCH ARTICLE | HUMAN GENOMICS

The complete sequence of a human genome

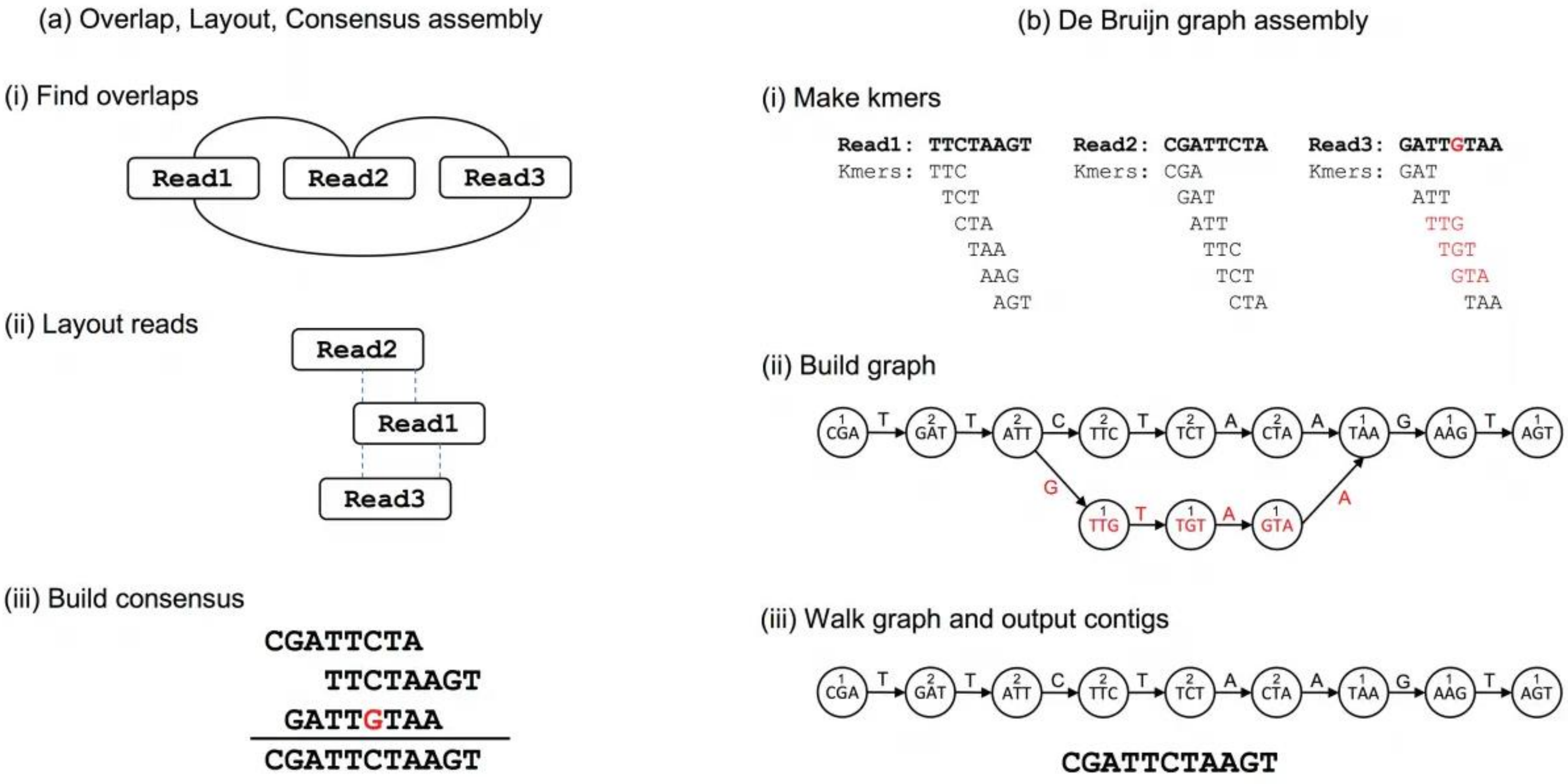
SERGEY NURK, SERGEY KOREN, ARANG RHIE, MIKKO RAUTIAINEN, ANDREY V. BZIKADZE, ALLA MIKHEENKO, MITCHELL R. VOLLGER, NICOLAS ALTEMOSE, LEV URALSKY, [...] ADAM M. PHILLIPPY +91 authors Authors Info & Affiliations

SCIENCE • 31 Mar 2022 • Vol 376, Issue 6588 • pp. 44-53 • DOI: 10.1126/science.abj6987



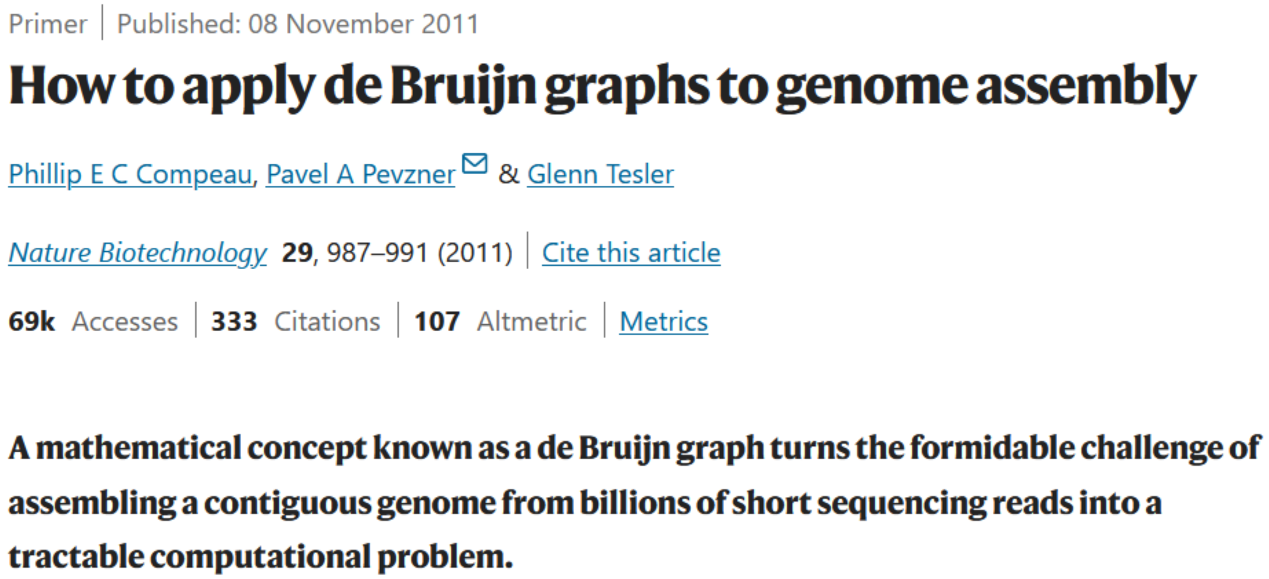
De novo Assembly的两种常用算法

- Overlap-Layout-Consensus (OLC) – 序列重叠一致性
  - Phrap
  - Newbler
  - Canu
- de Bruijn graph(DBG) – 德布鲁因图
  - SPAdes
  - Velvet
  - ABySS


(a) Overlap, Layout, Consensus assembly
(i) Find overlaps
Read1
Read2
Read3
(ii) Layout reads
Read2
Read1
Read3
(iii) Build consensus
CGATTCTA
TTCTAAGT
GATTGTAA
CGATTCTAAGT
(b) De Bruijn graph assembly
(i) Make kmers
Read1: TTCTAAGT
Kmers: TTC TCT CTA TAA AAG AGT
Read2: CGATTCTA
Kmers: CGA GAT ATT TTC TCT CTA
Read3: GATTGTAA
Kmers: GAT ATT TTG TGT GTA TAA
(ii) Build graph
1 CGA T 2 GAT T 2 ATT C 2 TTC T 2 TCT A 2 CTA A 1 TAA G 1 AAG T 1 AGT
G 1 TTG T 1 TGT A 1 GTA A
(iii) Walk graph and output contigs
1 CGA T 2 GAT T 2 ATT C 2 TTC T 2 TCT A 2 CTA A 1 TAA G 1 AAG T 1 AGT
CGATTCTAAGT

（引自Ayling et al., Briefings in Bioinformatics, 2019）

de Bruijn图算法 – K-mer选择



Primer | Published: 08 November 2011

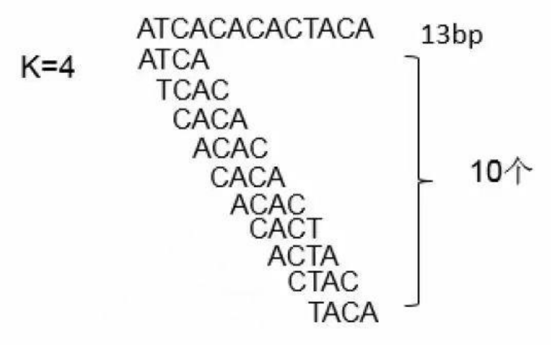
How to apply de Bruijn graphs to genome assembly

Phillip E C Compeau, Pavel A Pevzner & Glenn Tesler

Nature Biotechnology **29**, 987–991 (2011) | Cite this article

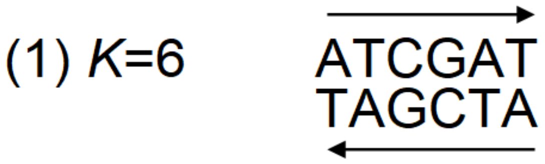
69k Accesses | **333** Citations | **107** Altmetric | Metrics

A mathematical concept known as a de Bruijn graph turns the formidable challenge of assembling a contiguous genome from billions of short sequencing reads into a tractable computational problem.

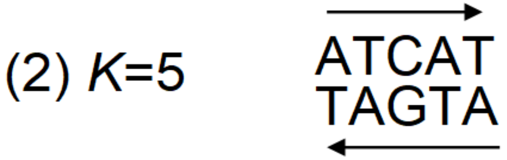


K=4
ATCACACACTACA 13bp
ATCA
TCAC
CACA
ACAC
CACA
ACAC
CACT
ACTA
CTAC
TACA
10个

- ***K*-mers指一个read中长度为k的所有子序列**，如序列ATTACGTCGA可以分成一系列3-mers（k=3）：ATT，TTA，TAC，ACG，CGT，GTC，TCG和CGA。
    - 若序列的长度为L，那么可以得到L-K+1个K-mers。
- 为防止由于出现回文序列而导致序列自身拼接的问题，*K*值应取奇数。



(1) *K*=6
ATCGAT
TAGCTA



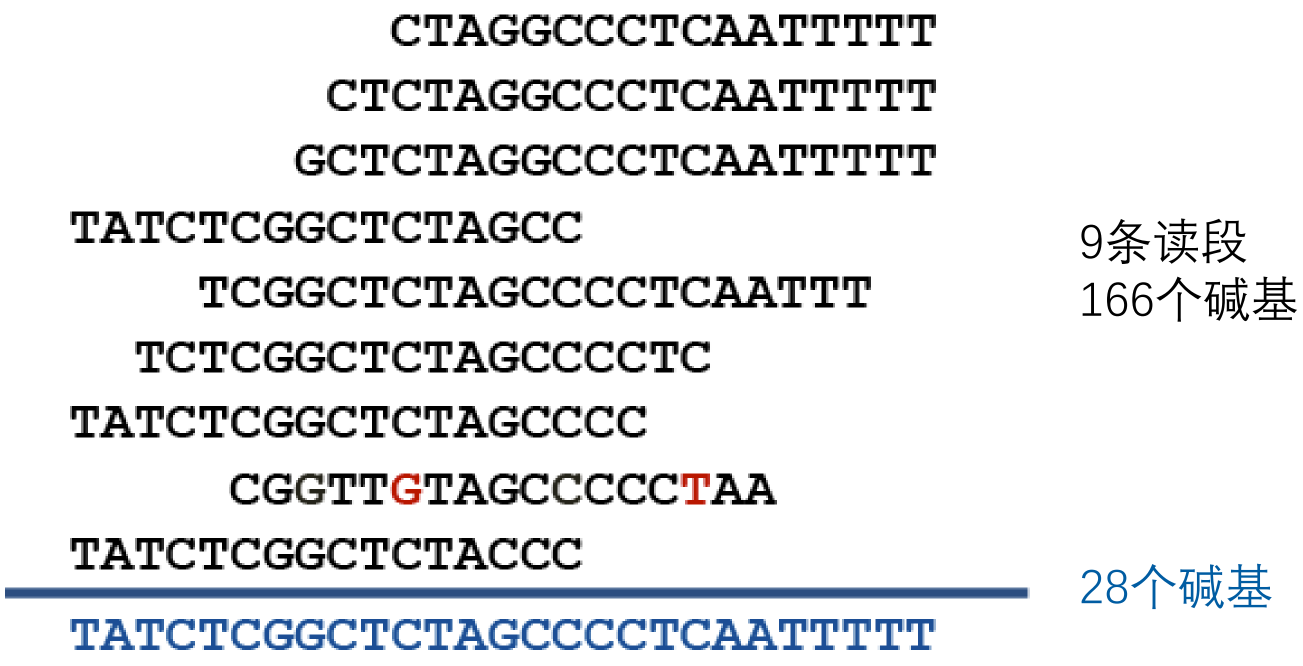
(2) *K*=5
ATCAT
TAGTA

基因组装质量评估参数

- Contig/Scaffold 数目
  - 越少越好
- N50/N90
  - 将Contigs/Scaffolds从长到短排序，逐项相加，长度达到全基因组一半(50%）或90%时对应的一段Contig/Scaffold长度；
  - 越长越好
- Coverage/Depth
  - 测序覆盖度/深度：单个碱基被测序reads覆盖的次数
- Genome coverage
  - 越长越好

测序覆盖度/深度(Coverage/Depth)

- 测序覆盖度：测序得到的总碱基数与待测基因组大小的比值。
- 测序深度：基因组上每一个位置平均覆盖读段的条数。
- 覆盖度C = n * l / L
  - n: 读段数
  - l: 读段长度
  - L：序列长度



```
          CTAGGCCCTCAATTTTT
        CTCTAGGCCCTCAATTTTT
       GCTCTAGGCCCTCAATTTTT
TATCTCGGCTCTAGCC
    TCGGCTCTAGCCCCTCAATTT
  TCTCGGCTCTAGCCCCTC
TATCTCGGCTCTAGCCCC
     CGGTTGTAGCCCCTAA
TATCTCGGCTCTACCC
───────────────────────────
TATCTCGGCTCTAGCCCCTCAATTTTT
```

9条读段
166个碱基

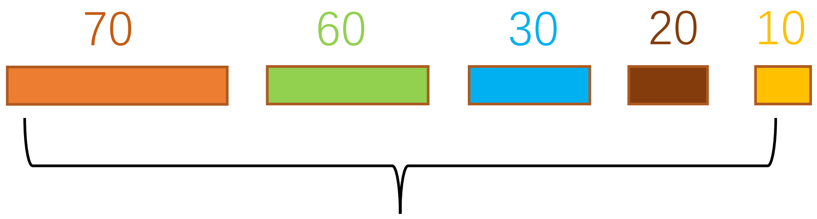
28个碱基

覆盖度C = 166 / 28 = 6x

基因组装质量评估参数

- N50 = length of the shortest contig where 50% of sum is held
- N90 = length of the shortest contig where 90% of sum is held
- L50 = number of contigs which have 50% of the genome
- L90 = number of contigs which have 90% of the genome

e.g., N50为将所有Contigs按长度从大到小排序后，使得累积长度达到总长度的一半的Contig长度



190 → 50% × 190 = 95

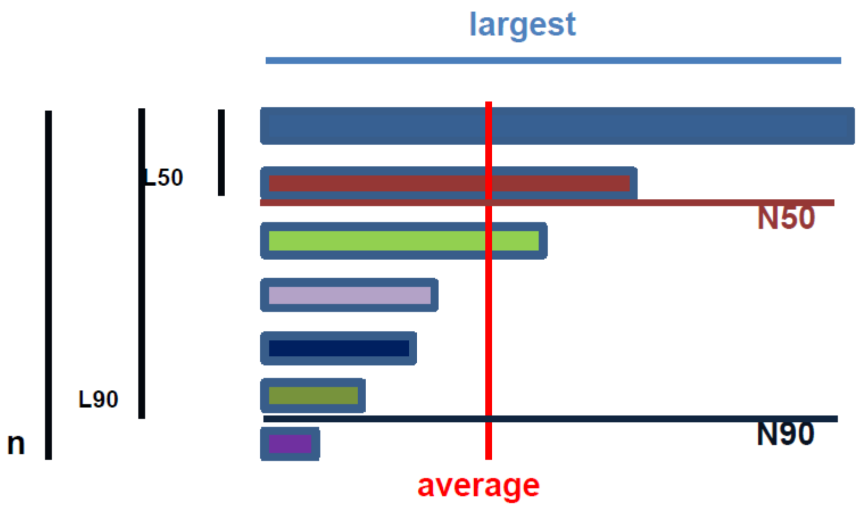
90% × 190 = 171

TotalContigLength (70) = 70 < 95

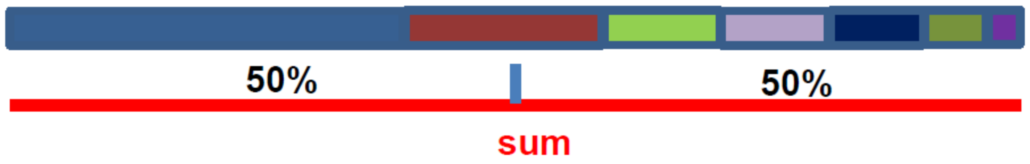
TotalContigLength (60) = 70 + 60 > 95 → N50:60

TotalContigLength (30) = 70 + 60 +30 < 171

TotalContigLength (20) = 70 + 60 +30+20 > 171 → N90:20



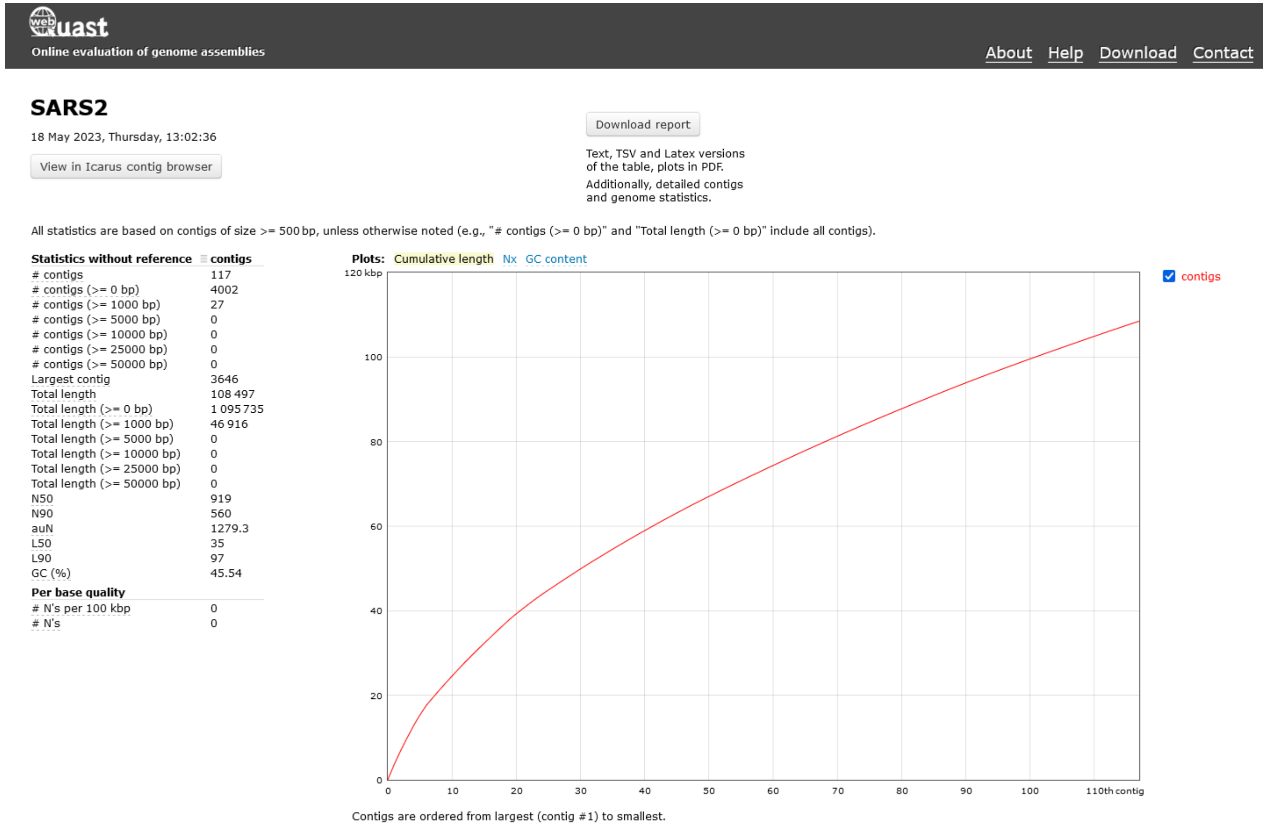
TotalContigLength(t) – total length of all contigs of length at least t



Sum: ottal length of all contigs

组装质量评估软件

- QUAST (Quality Assessment Tool for Genome Assembly)
  - https://quast.sourceforge.net
  - 根据多种指标，如组装速度、重叠群N50、最大重叠群长度等，评估从头组装软件的效果。



Online evaluation of genome assemblies

SARS2

18 May 2023, Thursday, 13:02:36

Text, TSV and Latex versions of the table, plots in PDF.

Additionally, detailed contigs and genome statistics.

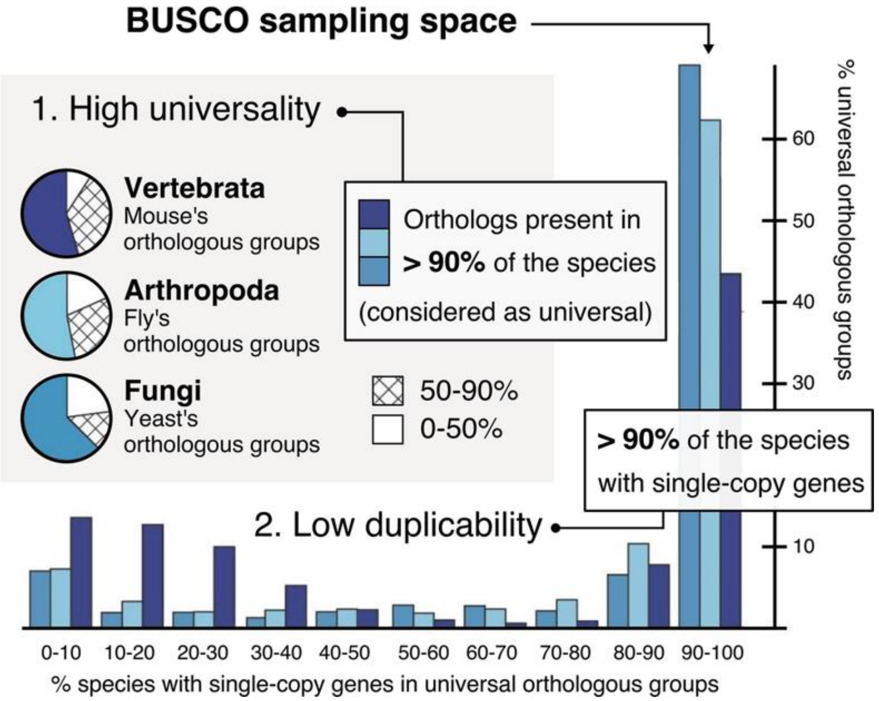
All statistics are based on contigs of size >= 500 bp, unless otherwise noted (e.g., "# contigs (>= 0 bp)" and "Total length (>= 0 bp)" include all contigs).

Statistics without reference	contigs
# contigs	117
# contigs (>= 0 bp)	4002
# contigs (>= 1000 bp)	27
# contigs (>= 5000 bp)	0
# contigs (>= 10000 bp)	0
# contigs (>= 25000 bp)	0
# contigs (>= 50000 bp)	0
Largest contig	3646
Total length	108 497
Total length (>= 0 bp)	1 095 735
Total length (>= 1000 bp)	46 916
Total length (>= 5000 bp)	0
Total length (>= 10000 bp)	0
Total length (>= 25000 bp)	0
Total length (>= 50000 bp)	0
N50	919
N90	560
auN	1279.3
L50	35
L90	97
GC (%)	45.54
Per base quality	
# N's per 100 kbp	0
# N's	0

Contigs are ordered from largest (contig #1) to smallest.

基因组完整性评估

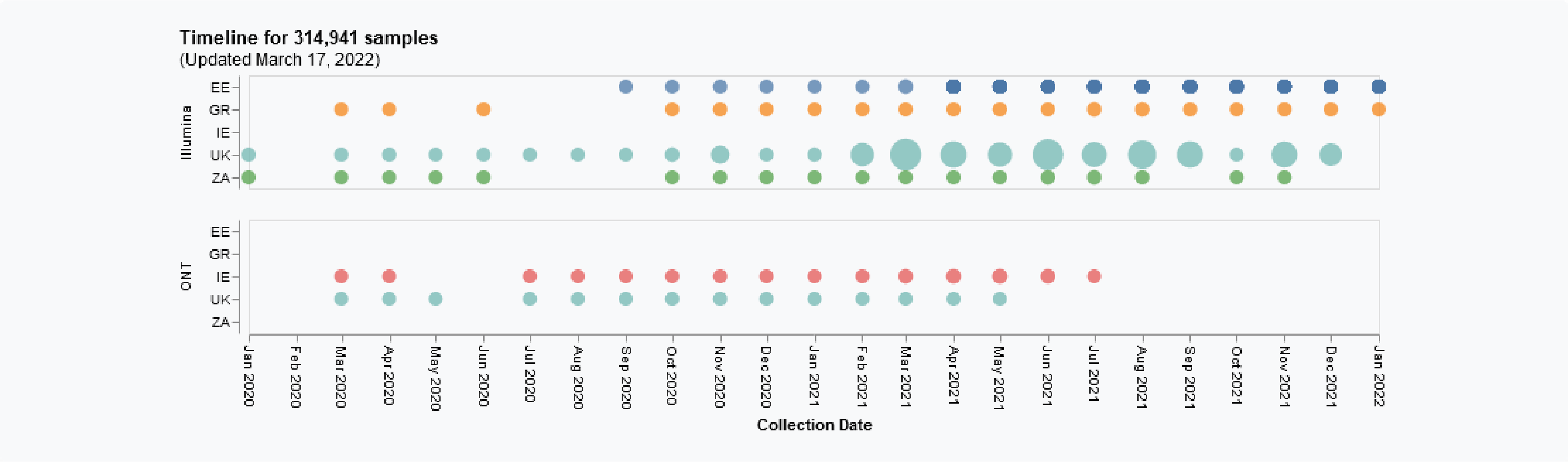
- CEGMA(Core Eukaryotic Genes Mapping Approach)
  - CEGMA选取了几个跨度较大的物种（人类、果蝇、线虫、拟南芥、酵母等），并找到这些物种中保守的458个基因。
  - 理论上，所有物种都有这458个基因，所以一个测序完成的基因组如果包含了这些基因，则表明了其基因组序列的完整性。
- BUSCO(Benchmarking Universal Single-Copy Orthologs)
  - Based on evolutionarily-informed expectations of gene content of near-universal single-copy orthologs, the BUSCO metric is complementary to technical metrics like N50.
  - BUSCO对各个物种大类分支都有一个保守基因集合（Bacteria, Eukaryota, Fungi, Plant …）
  - 网址：https://busco.ezlab.org/



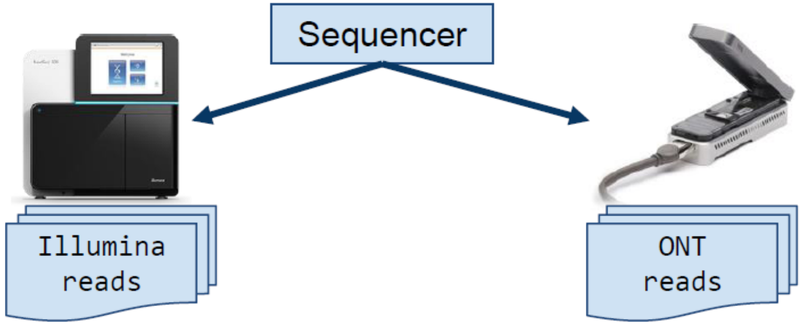
案例：Sequencing of SARS-CoV-2

Sequencing approaches for SARS-CoV-2

- Current Sequencing Sample Information (https://galaxyproject.org/projects/covid19/samples/)

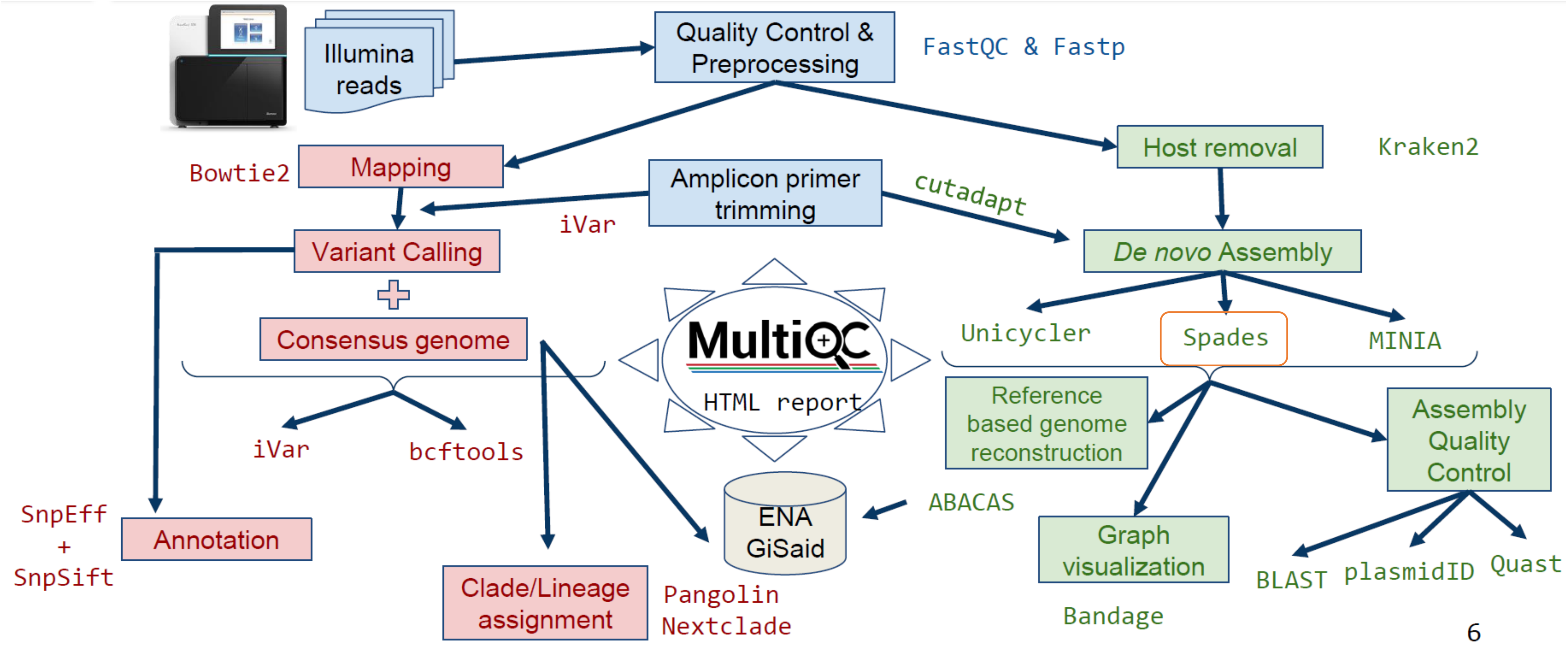


Timeline for 314,941 samples
(Updated March 17, 2022)

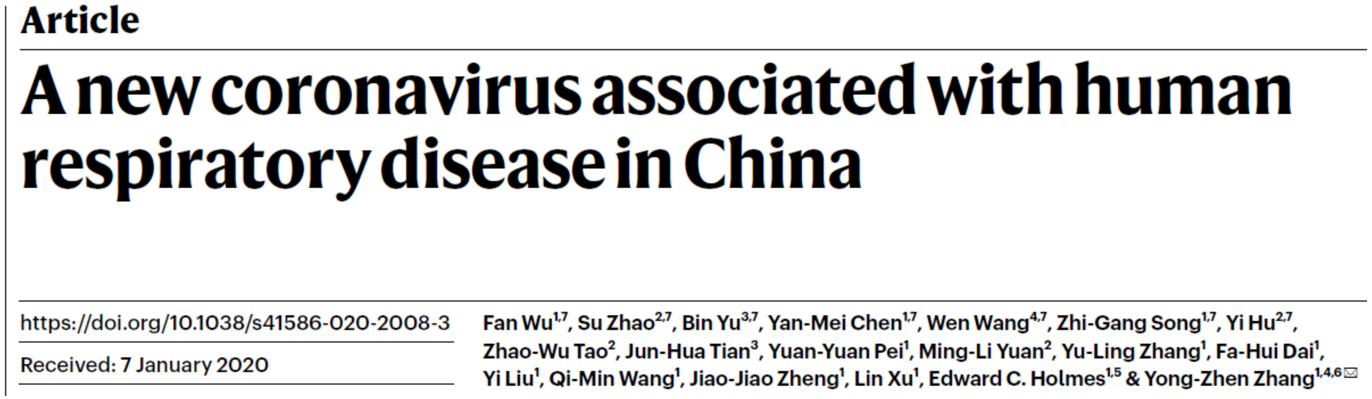


How next-generation sequencing can help identify and track SARS-CoV-2 (nature.com)

Illumina sequencing data analysis



Sequencing of SARS-CoV-2



Article

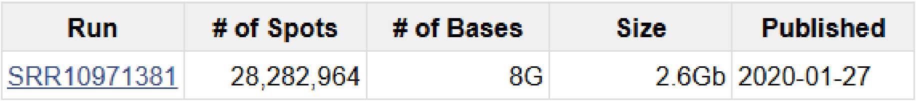
A new coronavirus associated with human respiratory disease in China

https://doi.org/10.1038/s41586-020-2008-3

Received: 7 January 2020

Fan Wu[1,7], Su Zhao[2,7], Bin Yu[3,7], Yan-Mei Chen[1,7], Wen Wang[4,7], Zhi-Gang Song[1,7], Yi Hu[2,7], Zhao-Wu Tao[2], Jun-Hua Tian[3], Yuan-Yuan Pei[1], Ming-Li Yuan[2], Yu-Ling Zhang[1], Fa-Hui Dai[1], Yi Liu[1], Qi-Min Wang[1], Jiao-Jiao Zheng[1], Lin Xu[1], Edward C. Holmes[1,5] & Yong-Zhen Zhang[1,4,6] ✉

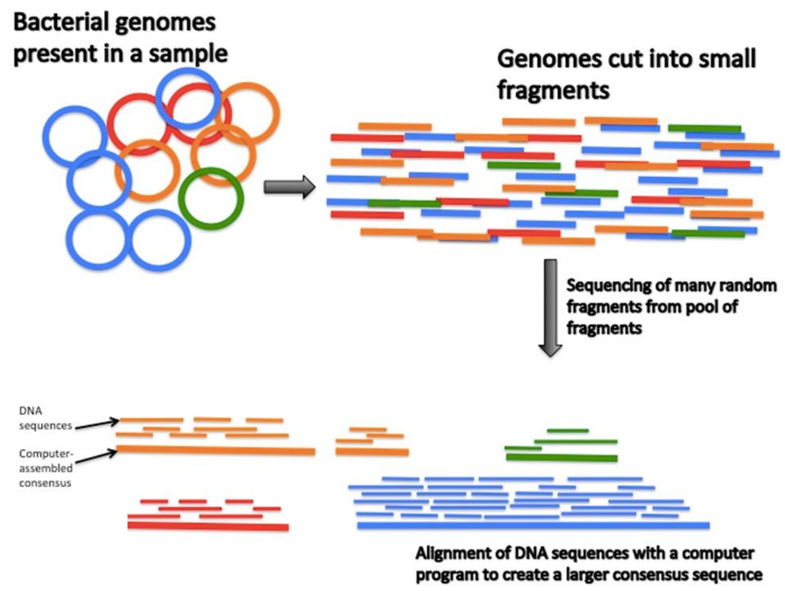
- Total RNA was extracted from the BALF sample of a patient.
  - The term **BALF** is a shorthand for **B**roncho**a**lveolar **L**avage **F**luid, which is fluid collected from a patient's lungs.
- It is a mix of the patient's RNA and the viral RNA.
  - this mixture of RNA is called a **metatranscriptome**
- Sequencing data: https://www.ncbi.nlm.nih.gov/sra/SRR10971381



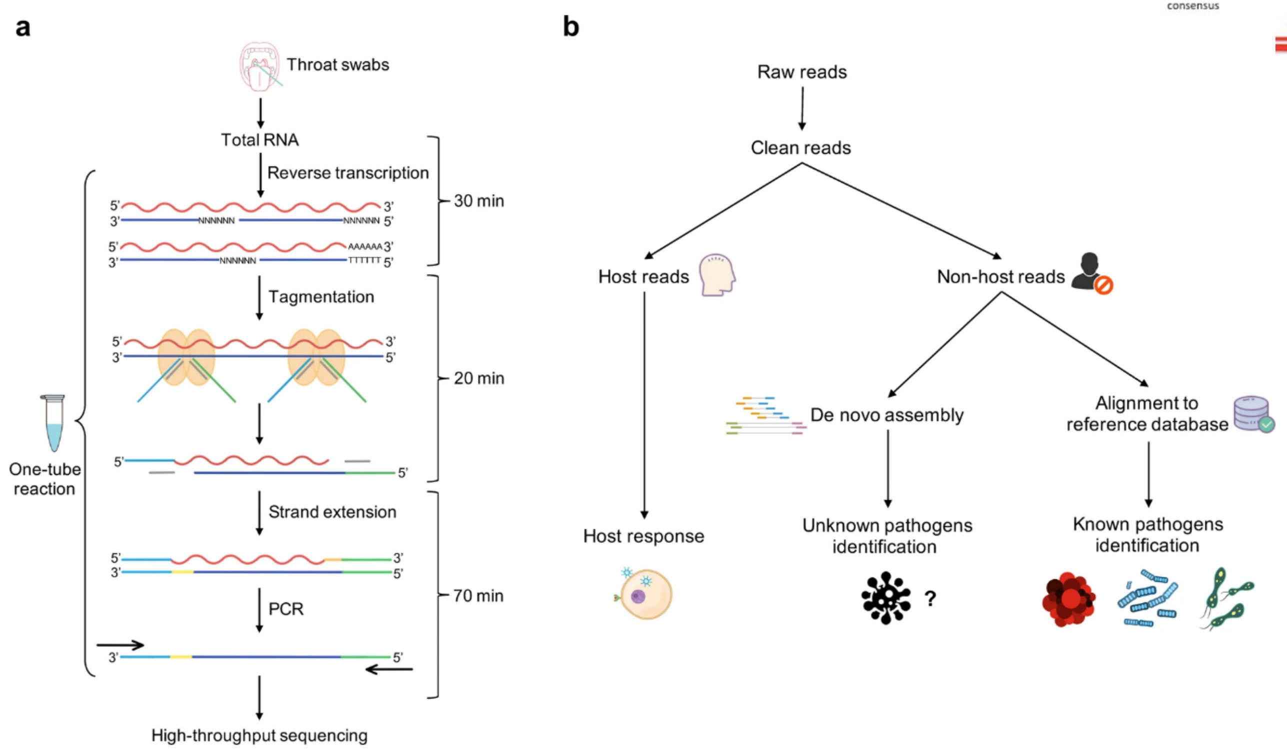
Run	# of Spots	# of Bases	Size	Published
SRR10971381	28,282,964	8G	2.6Gb	2020-01-27

Metagenomic Next-Generation Sequencing

- NGS技术在微生物检测中的应用为宏基因组测序(metagenomic Next-Generation Sequencing, mNGS) 。
  - DNA: Metagenome sequencing
  - RNA: Metatranscriptome sequencing

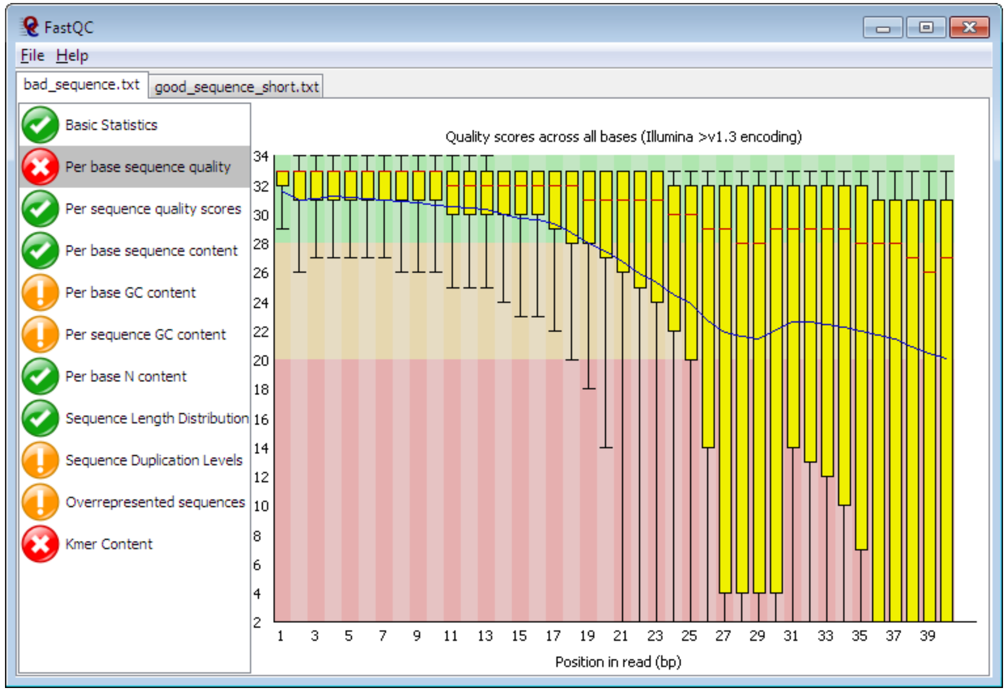

Bacterial genomes present in a sample
Genomes cut into small fragments
Sequencing of many random fragments from pool of fragments
DNA sequences
Computer-assembled consensus
Alignment of DNA sequences with a computer program to create a larger consensus sequence

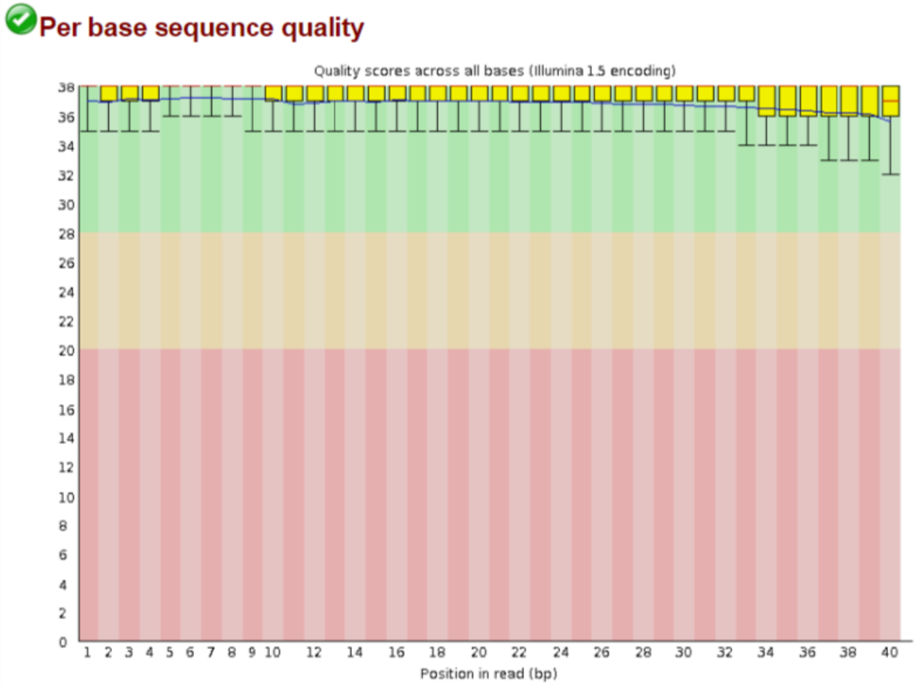
- mNGS是新冠疫情中鉴定出新冠病毒的主要技术。



数据质控-FastQC

- FastQC是目前最常用的NGS数据质量评估软件，可用于统计质量分数、GC含量、测序长度等信息
  - http://www.bioinformatics.babraham.ac.uk/projects/fastqc/


FastQC
File
Help
bad_sequence.txt
good_sequence_short.txt
✓ Basic Statistics
✗ Per base sequence quality
✓ Per sequence quality scores
✓ Per base sequence content
! Per base GC content
! Per sequence GC content
✓ Per base N content
✓ Sequence Length Distribution
! Sequence Duplication Levels
! Overrepresented sequences
✗ Kmer Content
Quality scores across all bases (Illumina >v1.3 encoding)
34
32
30
28
26
24
22
20
18
16
14
12
10
8
6
4
2
1
3
5
7
9
11
13
15
17
19
21
23
25
27
29
31
33
35
37
39
Position in read (bp)



Overview of the range of quality values across all bases at each position in the FastQ file

数据预处理-Trimmomatic

- Trimmomatic可用于过滤与切除低质量序列、接头序列，不需要的污染物序列等
  - http://www.usadellab.org/cms/uploads/supplementary/Trimmomatic/Trimmomatic-0.38.zip

- 执行命令：

```
$ java -jar Trimmomatic-0.39/trimmomatic-0.39.jar PE SRR10971381_1mini.fq
SRR10971381_2mini.fq SRR10971381_1_paired.fq SRR10971381_1_unpaired.fq
SRR10971381_2_paired.fq SRR10971381_2_unpaired.fq ILLUMINACLIP:
Trimmomatic-0.39/adapters/TruSeq3-PE.fa:2:30:10:2:True LEADING:20
TRAILING:20 MINLEN:36
```

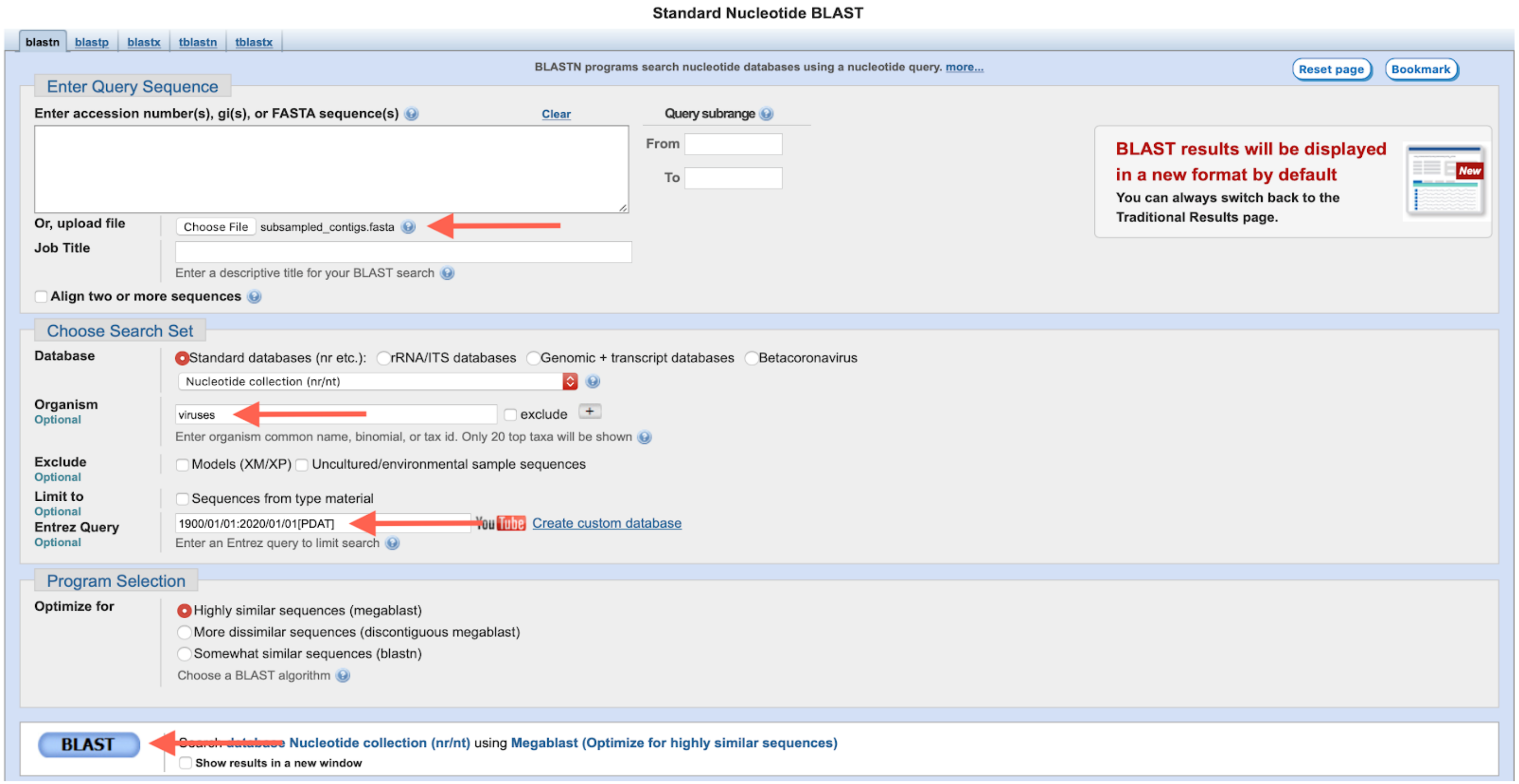
注意，在运行此java命令的目录下安装Trimmomatic，并存放测序数据！

Assembly with SPAdes

- SPAdes (St. Petersburg genome assembler)是一款非常好用的软件，而且该软件不仅支持Illumina测序数据，而且还可以用于Ion Torrent，PacBio、Sanger，Nanopore测序数据。使用简单，非常适合用于混合组装来改善拼接效果。
  - 安装说明: http://ablab.github.io/spades/installation.html
-
  ```
  $ spades.py -1 SRR10971381_1_paired.fq -2 SRR10971381_2_paired.fq --careful -o sars_illupe
  ```
  - **-o** output_dir：指定输出的文件夹；
  - **-1/-2**：单个paired-end library的左端与右端测序reads；
  - --pe1-1/--pe1-2：参数用于质量过滤后还是配对的reads，pe后的数字1为每1个文库的编号，而后跟着配对reads的编号，1是左端(forward), 而2是右端(reverse)；
  - --pe1-s：paired-end library 数据过滤后两端reads不能再配对的single数据，就在同一个--pe参数后加-s标记；
  - --careful：通过运行MismatchCorrector模块进行基因组上mismatches和short indels的修正，推荐使用此参数。
- 命令运行结束后，在输出文件夹sars_illupe中产生许多文件，其中configs.fasta是最终的组装结果。

Analyzing your assembly

- 用记事本打开contigs.fasta文件，复制第1条序列(NODE_1)做BLAST，结果显示为新冠病毒序列。
- BLAST program：BLASTN
- BLAST parameters:
  - Entrez Query: 1900/12/01:2020/02/01[PDAT] (search results before January 1st, 2020
  - Organism: viruses



Standard Nucleotide BLAST

blastn | blastp | blastx | tblastn | tblastx

BLASTN programs search nucleotide databases using a nucleotide query. more...

Reset page | Bookmark

Enter Query Sequence

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

Query subrange: From | To

Or, upload file: Choose File subsampled_contigs.fasta

Job Title: Enter a descriptive title for your BLAST search

□ Align two or more sequences

BLAST results will be displayed in a new format by default

You can always switch back to the Traditional Results page.

Choose Search Set

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

Nucleotide collection (nr/nt)

Organism (Optional): viruses □ exclude +

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): 1900/01/01:2020/01/01[PDAT] 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

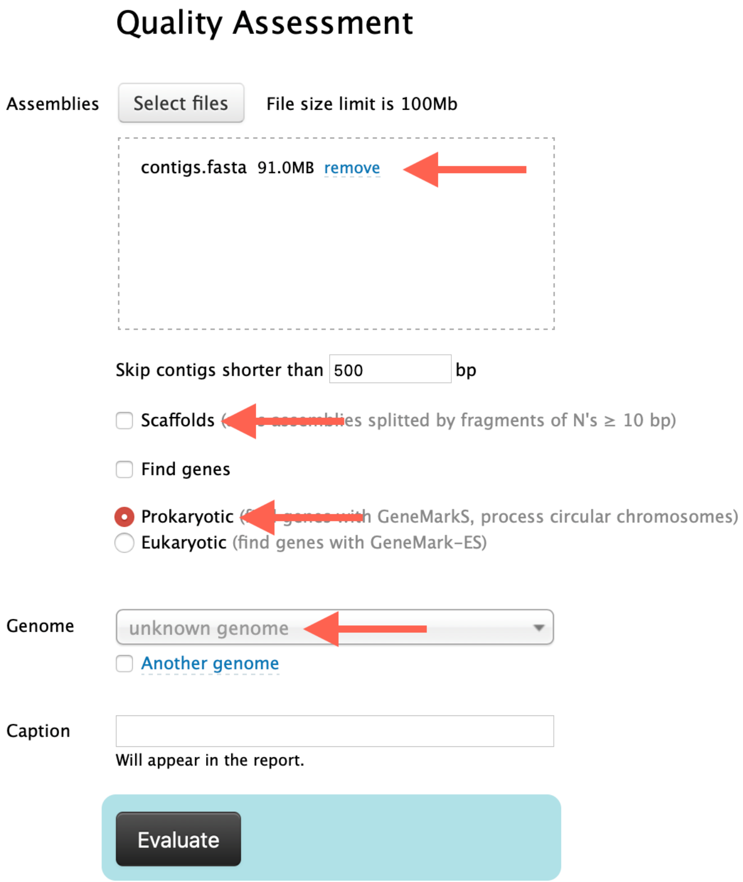
BLAST | Search database Nucleotide collection (nr/nt) using Megablast (Optimize for highly similar sequences)

□ Show results in a new window

QUAST-the Quality Assessment Tool for Genome Assembly

- QUAST本地安装与运行：
  - $wget https://downloads.sourceforge.net/project/quast/quast-5.2.0.tar.gz
  - $tar -zxvf quast-5.2.0.tar.gz
  - $./quast-5.2.0/quast.py contigs.fasta –o quast_result

- QUAST online tool: http://cab.cc.spbu.ru/quast/
  - "Scaffolds" is not checked,
  - "Prokaryotic" is selected,
  - "Genome" is set to " SARS-CoV-2”



Quality Assessment

Assemblies [Select files] File size limit is 100Mb

contigs.fasta 91.0MB remove

Skip contigs shorter than 500 bp

□ Scaffolds (assemblies splitted by fragments of N's ≥ 10 bp)

□ Find genes

● Prokaryotic (find genes with GeneMarkS, process circular chromosomes)
○ Eukaryotic (find genes with GeneMark-ES)

Genome unknown genome
□ Another genome

Caption
Will appear in the report.

[Evaluate]

Explore k-mer Length Effects

- Look at the log file generated(spades.log), what k-mer sizes were used for the assembly?
  - Default k-mer sizes were set to [21, 33, 55, 77]

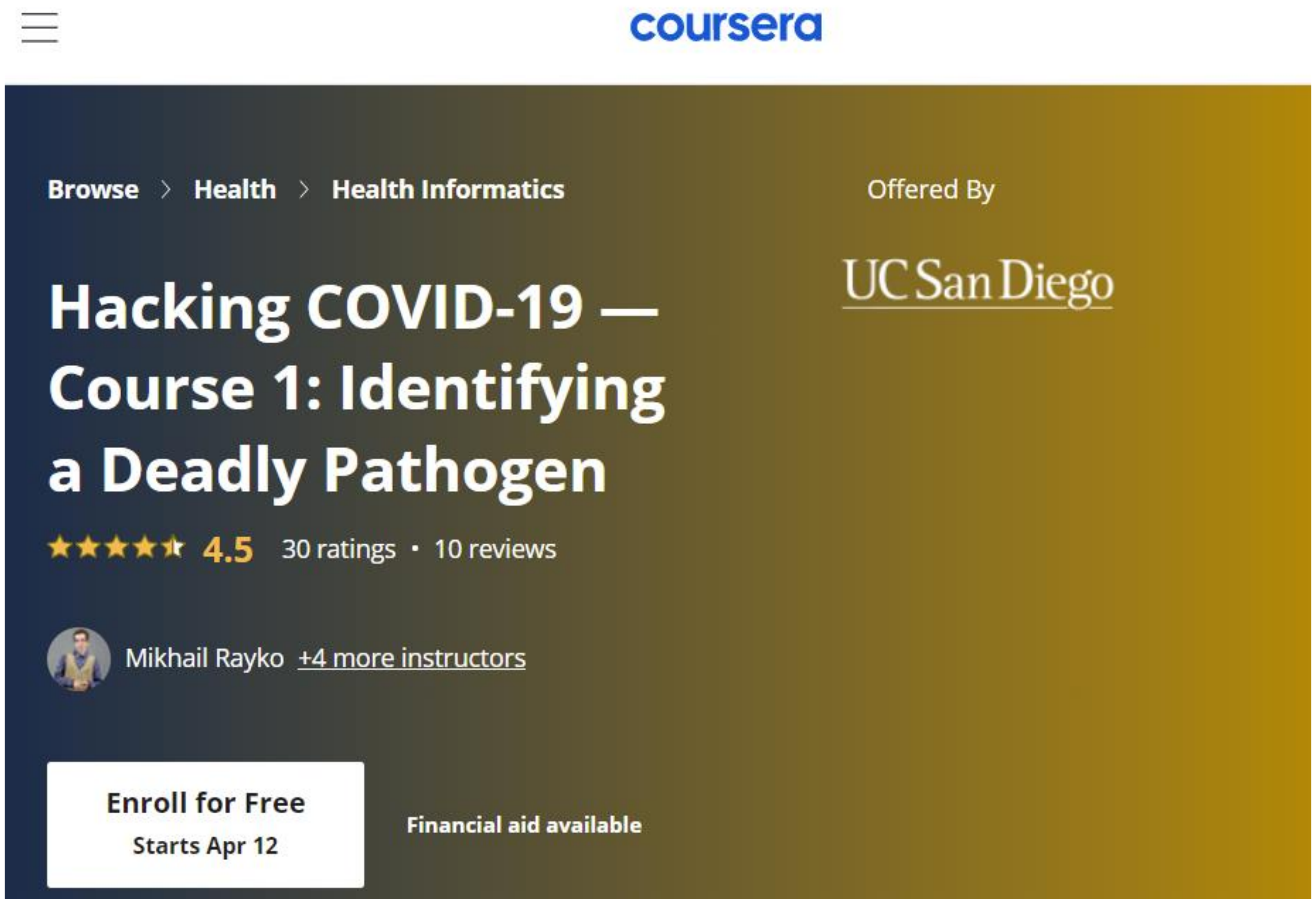
- Try to assembly with a different k-mer length:

```
$k = 127
$spades.py \
  --threads 2 \
  -k ${k} \
  -o ~/de_novo_illumina/SRR10971381-k${k} \
  -1 ~/data/illumina_pe/SRR10971381_1_paired.fq \
  -2 ~/data/illumina_pe/SRR10971381_2_paired.fq \
| tee ~/de_novo_illumina/SRR10971381-k${k}.log
```

https://university-of-adelaide-bx-masters.github.io/BIOINF-3010-7150/Practicals/short_read_assembly/short-read-assembly.html

Hacking COVID-19 — Identifying a Deadly Pathogen

- 参考教学视频：
  - https://www.coursera.org/learn/covid-19-genome-assembly



coursera

Browse › Health › Health Informatics

Offered By

UC San Diego

Hacking COVID-19 — Course 1: Identifying a Deadly Pathogen

★★★★★ 4.5 30 ratings · 10 reviews

Mikhail Rayko +4 more instructors

Enroll for Free
Starts Apr 12

Financial aid available

作业

- 根据本课提供的测序数据(SRR10971381)与分析流程，完成SARS-CoV-2基因组测序分析的数据质控(FastQC, Trimmomatic)、基因组组装与评估(SPAdes, QUAST)与病原体鉴定(BLAST)等。

ZHEJIANG GONGSHANG UNIVERSITY

1911

浙江工商大学

谢 谢

THANK YOU