华大基因 BGI

基于数加MaxCompute的极速全基因组数据分析

黄树嘉
华大基因 基因组学数据专家

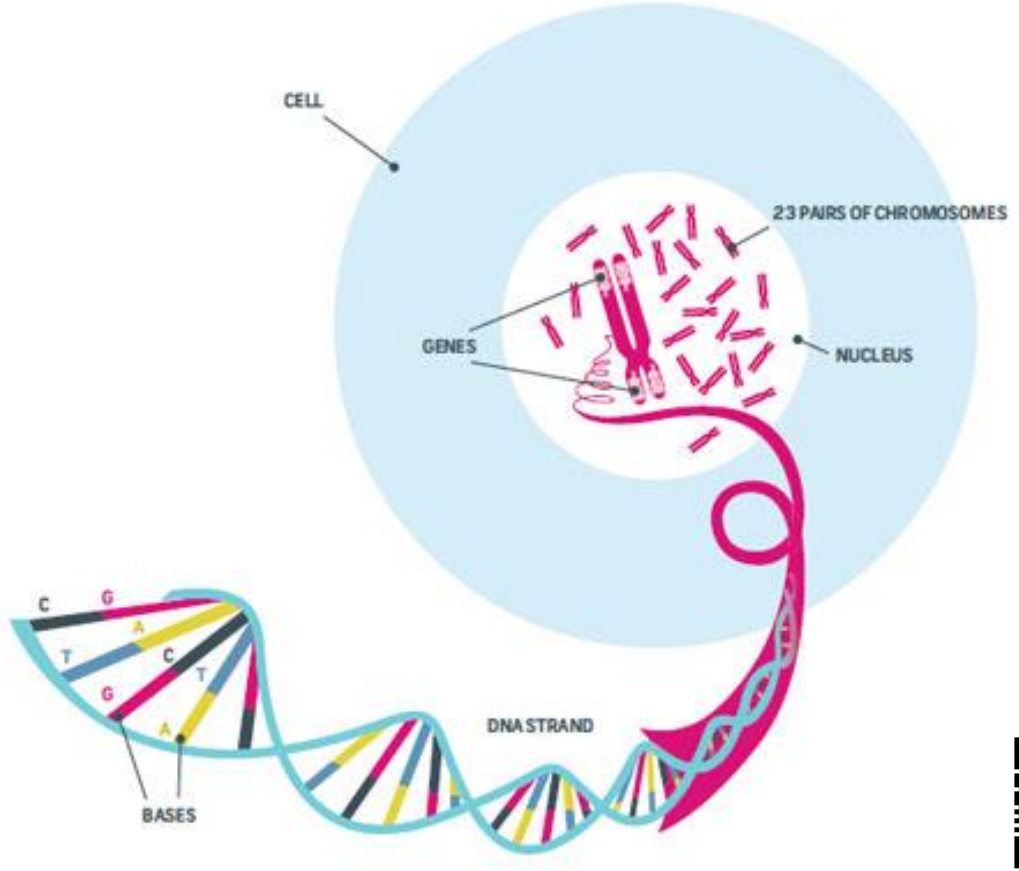
主办单位：杭州 Alibaba Group 阿里巴巴集团 战略合作伙伴：intel

扫码观看大会视频

目 录
content

扫码观看大会视频

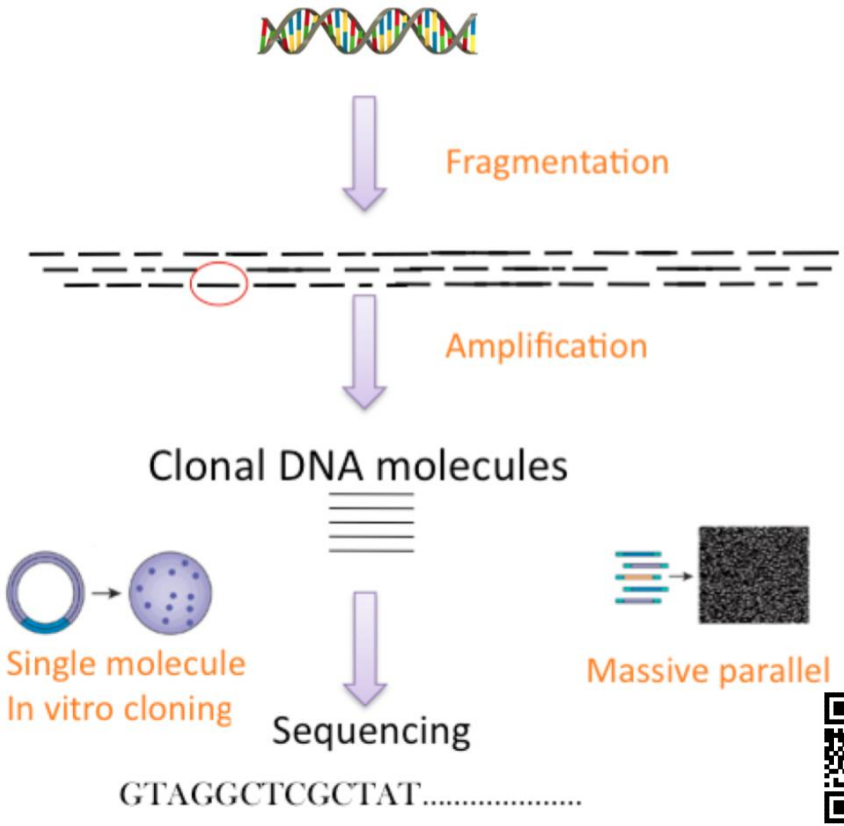
什么是基因



基因，生命的基本因素，是人类和其他生物的基础遗传物质

什么是基因测序


BGISEQ-500


Fragmentation
Amplification
Clonal DNA molecules
Single molecule
In vitro cloning
Massive parallel
Sequencing
GTAGGCTCGCTAT.....

一个人

一生的基因数据

10TB = 0.1TB + 0.7TB + 2TB + 3TB + X TB

基因组 转录组 表观组 宏基因组 其他

扫码观看大会视频

基因数据分析的过程

Sequencing

Reads

Alignment

Variant calling

Variants, QC, Analysis

Variant interpretation

External/Internal knowledge integration

Knowledge

1 ° Analysis

2 ° Analysis

3° Analysis

扫码观看大会视频

传统单机分析流程的挑战

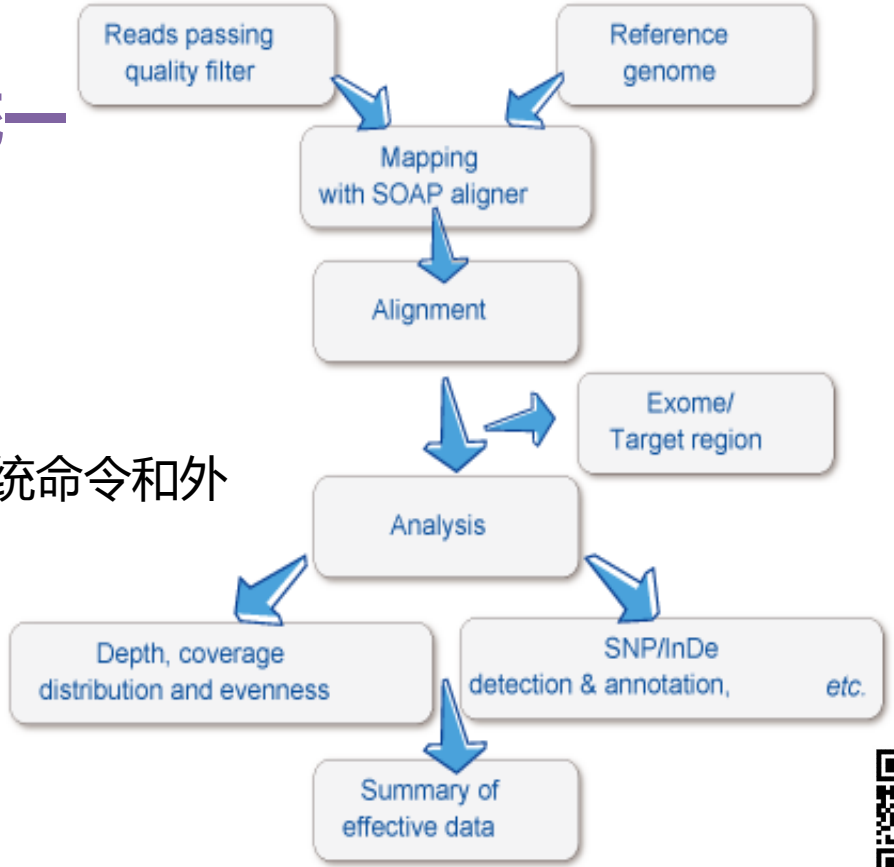
挑战1：流程繁杂，标准难统一

分析流程特点：

1.多个分析步骤

2.每个步骤都会包含很多分析脚本，系统命令和外部工具

3.工具要被反复手动部署到计算集群



挑战2：命令行操作、交互性差

```
Kernel command line: block2mtd.block2mtd=/dev/hda2,131072,rootfs root=/dev/mtdbl
ock0 rootfstype=jffs2 init=/etc/preinit  noinitrd console=ttyO0 console=ttyS0,384
00n8 reboot=bios
Found and enabled local APIC!
Enabling fast FPU save and restore... done.
Enabling unmasked SIMD FPU exception support... done.
Initializing CPU#0
PID hash table entries: 32 (order: 5, 128 bytes)
Detected 1991.657 MHz processor.
Console: colour VGA+ 80x25
console [ttyO0] enabled
console [ttyS0] enabled
Dentry cache hash table entries: 1024 (order: 0, 4096 bytes)
Inode-cache hash table entries: 1024 (order: 0, 4096 bytes)
Memory: 5112k/8128k available (1497k kernel code, 2624k reserved, 597k data, 196
k init, 0k highmem)
virtual kernel memory layout:
    fixmap  : 0xfffb9000 - 0xfffff000   ( 280 kB)
    vmalloc : 0xc1000000 - 0xfffb7000   (1007 MB)
    lowmem  : 0xc0000000 - 0xc07f0000   (   7 MB)
      .init : 0xc0313000 - 0xc0344000   ( 196 kB)
      .data : 0xc027653c - 0xc030bcfc   ( 597 kB)
      .text : 0xc0100000 - 0xc027653c   (1497 kB)
Checking if this processor honours the WP bit even in supervisor mode...Ok.
Calibrating delay using timer specific routine.. 4047.64 BogoMIPS (lpj=20238210)
```

扫码观看大会视频

挑战3：时间长

一次测序的数据产出

测序仪	一次测序的数据总产量	一次测序的Reads（Billion）	测序读长（bp）	测序时间周期
HiSeq 3000	750GB	2.1-2.5	PE 150	3.5 days
HiSeq 4000	1.5TB	4.3-5.0	PE 150	3.5 days

分析一个人的基因组~120G数据，往往需要3天以上的时间。

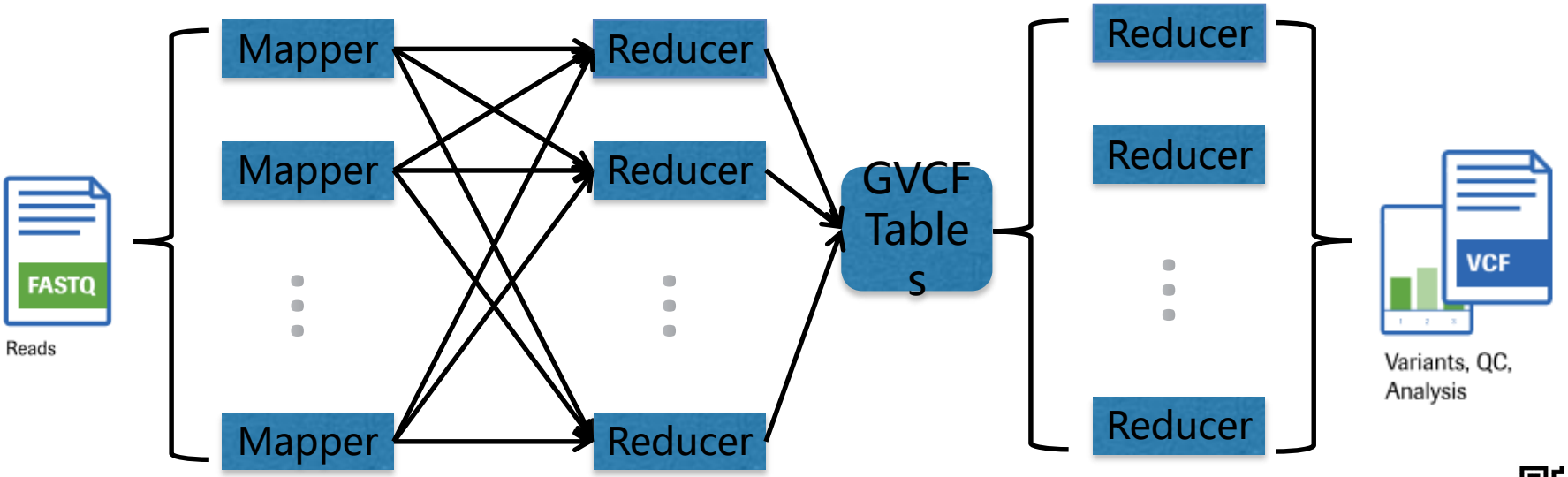
数据的解读跟不上数据的产出。

	时间（小时）
传统HPC集群	~72 (3.0 days)
单个节点计算	140 (5.8 days)

扫码观看大会视频

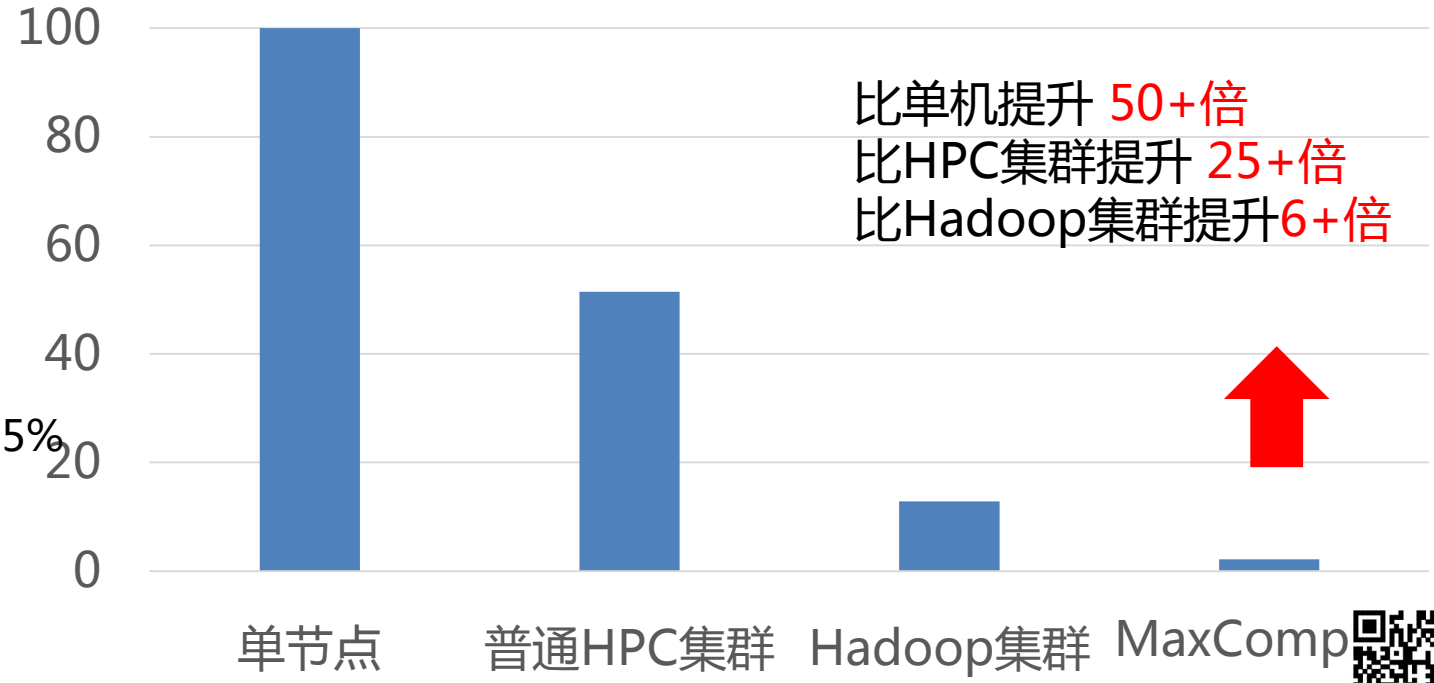
基于MaxCompute的方案

MaxCompute分布式计算



单个基因组分析实现50+倍的加速

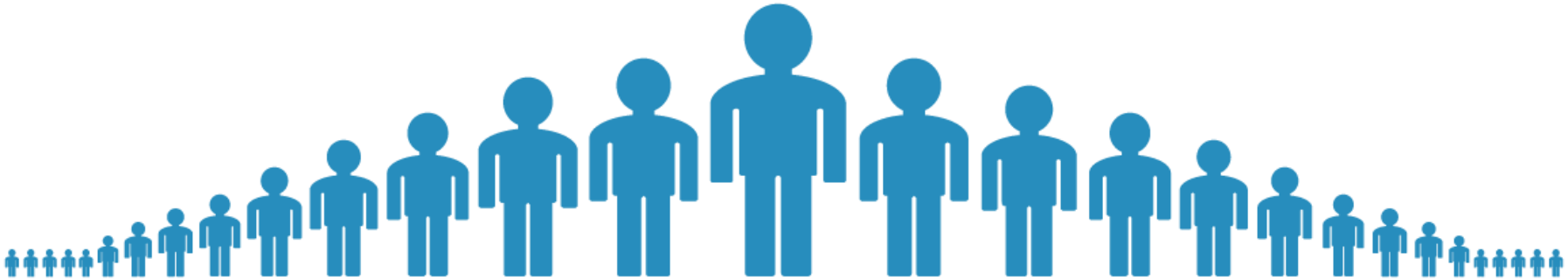
- 120G数据
- ~3个小时
- 精确度: 99.57%
- Recall: 98.53%
- F-Measure: 99.05%



扫码观看大会视频

更快：50个全基因组分析

数据来源于华大基因内部已有成果发表的项目



扫码观看大会视频

2大步骤, 70000+任务，41.5小时

2 steps

70000+ Jobs

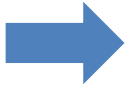
41.5 hours

50 min/genome

扫码观看大会视频

海量的计算，从原始数据到精确变异

2T FASTQ



21G VCF

华大基因
BGI

云栖社区
yq.aliyun.com

2016 The Computing Conference

THANKS

huangshujia@genomics.cn

2016 杭州·云栖大会
THE COMPUTING CONFERENCE

扫码观看大会视频