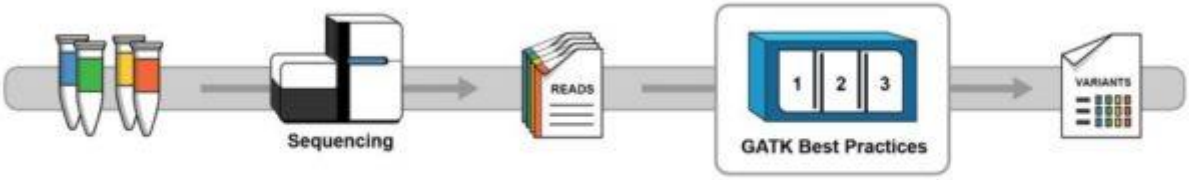
# 滴雨科技生物信息分析大数据平台

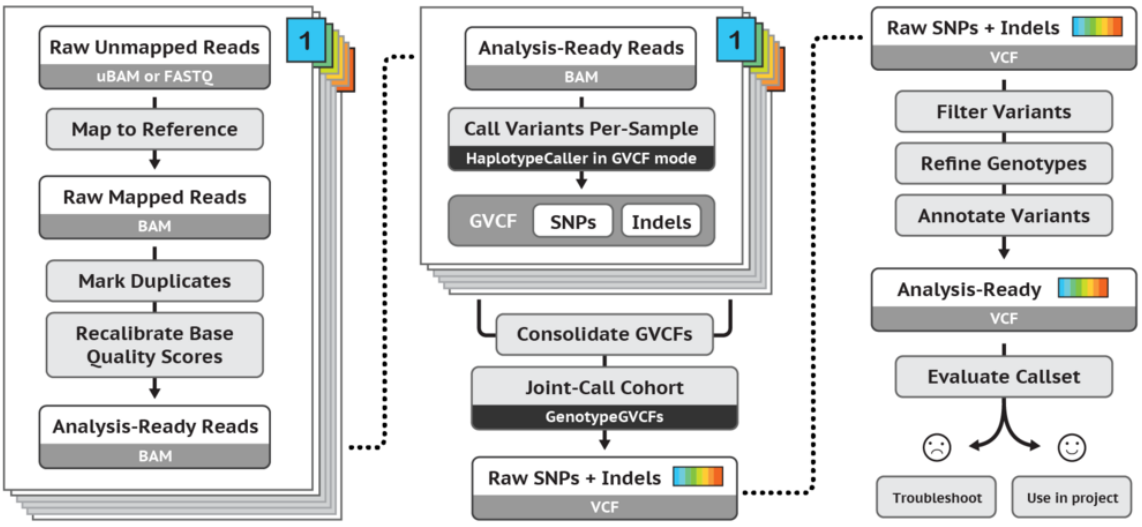
将人工智能和深度学习的技术应用于生物信息，是第三代基因分析的核心技术，滴雨科技在拥有第二代生物分析平台 GATK 的前提下，推出了基于 GOOGLE DEEPVARIANT 的第三代基因分析技术。所有的生物信息分析技术同时是在滴雨科技的第二代分布式文件系统和集群之上，是当今领先的新一代生物分析大数据平台，能支持上万节点的大数据集群。

## 基于ＧＡＴＫ的第二代技术

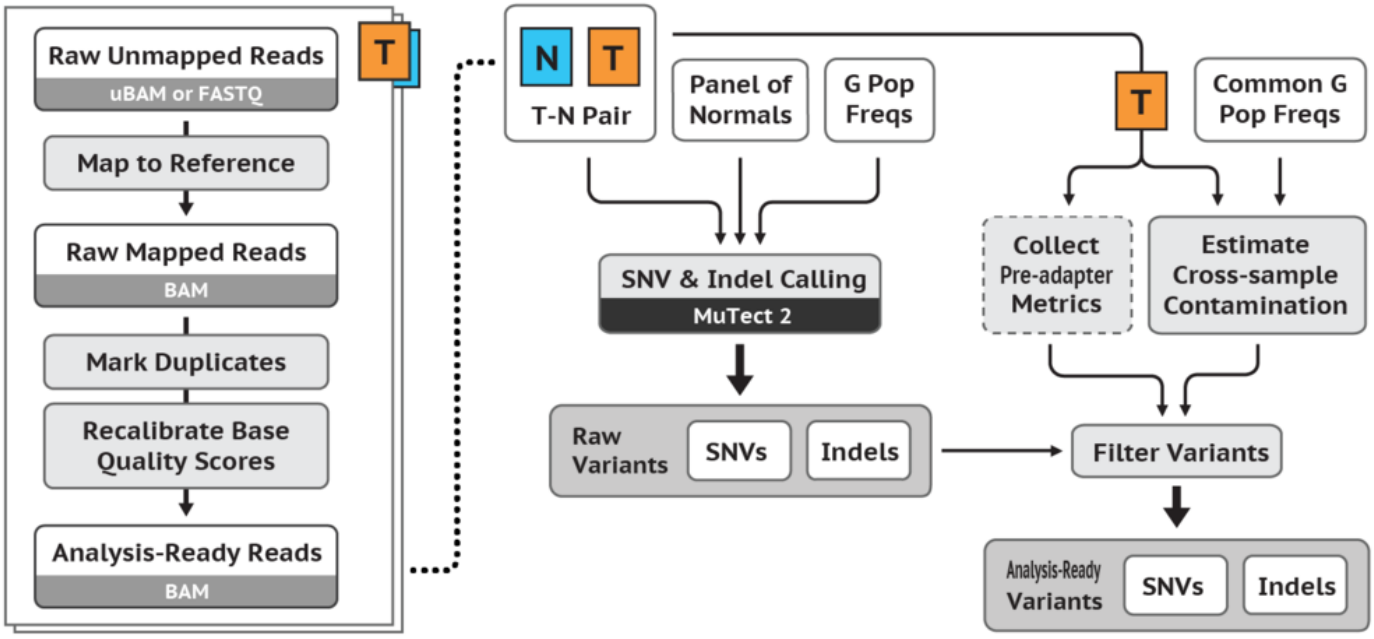
GATK 是 Genome Analysis ToolKit 的缩写，是一款从高通量测序数据中分析变异信息的软件，是目前最主流的 snp calling 软件之一。GATK 设计之初是用于分析人类的全外显子和全基因组数据，随着不断发展，现在也可以用于其他的物种，还支持 CNV 和 SV 变异信息的检测。滴雨科技定制并提供了完整的分析流程，叫做 GATK Best Practices。




目前最新版本文为 4.0.4.0，叫做 GATK4。 和之前的版本相比，GATK4 在算法上进行了优化，运行速率有所提高，而且整合了 picard 软件的功能。GATK4 基于 java 语言开发的，需要 java 1.8 版本。




Ｇｅｒｍｌｉｎｅ遗传基因组分析流程




somatic个体基因变异分析流程

GATK4 的最佳实践给出了 5 套 pipeline

1. Germline SNPs + Indels
2. Somatic SNVs + Indels
3. RNAseq SNPs + Indels
4. Germline CNVs
5. Somatic CNVs

以上五套 pipeline 可以根据研究对象是 DNA 还是 RNA 进行划分：DNA 测序（包含 1,2,4,5）和 RNA 测序（3）。可以看到，GATK 更多的是倾向于 DNA 测序数据的分析。对于 DNA 测序而言，主要识别 SNP 和 CNV 两大类型的变异，每种变异类型又有 Germline 和 Somatic 的区别。 Germline 指的是在胚胎发育早起出现的变异，这种变异会在所有细胞中广泛存在，是可以遗传给后代的变异；Somatic 指的是体细胞变异，身体特定区域或者组织中出现的变异。通常不会遗传给后代。 在所有的 pipeline 之前，都存在一个数据预处理步骤 data pre-processing。 GATK4 版本的最佳实践并不是直接给出了每个步骤对应的代码，而是给出了几套它们自己编写的流程，以供参考。这些流程以 WDL 这种 workflow 语言进行编写。

## 第三代基因测序技术

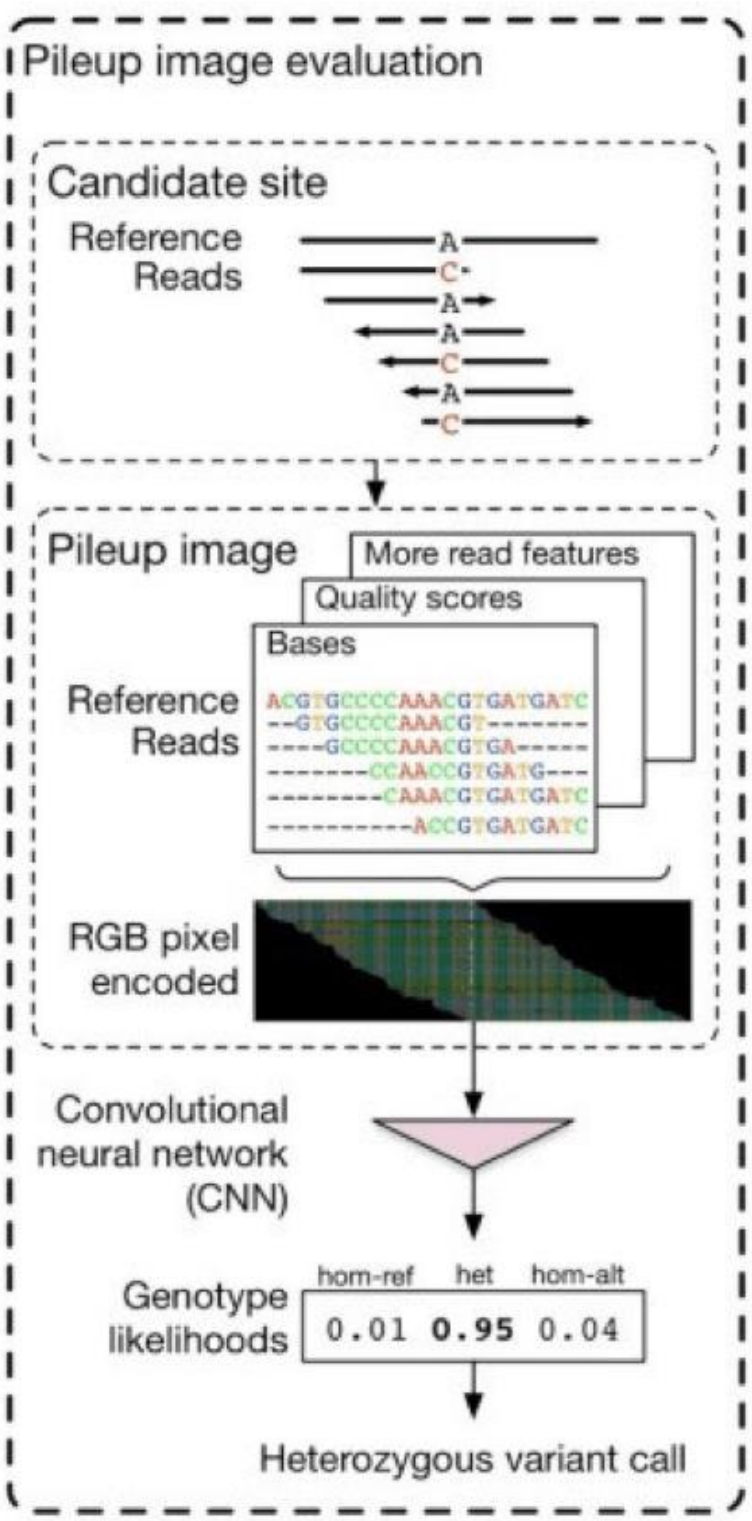
随着技术的发展，不断增强的计算能力和大数据已经使多层复杂的深度神经网络的“学习潜力”超过了传统的统计方法。

一般来说，深度学习网络中输入的是相对原始的数据。前面的图层可以从中学习较为“粗略”的特征，例如用于计算机视觉中的边缘检测。后面的图层则从中学习更高级的、抽象的信息。而神经网络的架构是这些深度神经网络顺利工作的保证，只有特定的结构才能让信息组合有意义。

通过开放源代码框架 TensorFlow、创建出优秀的机器翻译软件以及无人能敌的 AlphaGo，谷歌正将其顶尖的的深度学习技术推进到数据中心的优化使用中。

## 什么是 DeepVariant?

DeepVariant 采用了 TensorFlow 的 Inception 框架，这个框架最初是用于图像分类工作的。DeepVariant 将一个 BAM 转换成类似于基因组浏览器快照似的图像，然后根据是否有变体进行分类。理论上说，如果一个人能够在基因组浏览器中判断识别器是否是正确的，那么一个足够聪明的框架也能做到。



第一部分是收集有可能存在变异的基因序列样本。这就需要用非常灵敏的识别器找出任何有可能存在变体的区域。除此之外，DeepVariant 执行的是局部重组，是 Indel realignment 的一个更彻底的版本。最后生成多维图像传至分类器。

第二部分是用 TensorFlow 框架识别变体。将图片输入到经过训练的 Inception 中，它就能识别 SNP 和 Indel 变体的标志。

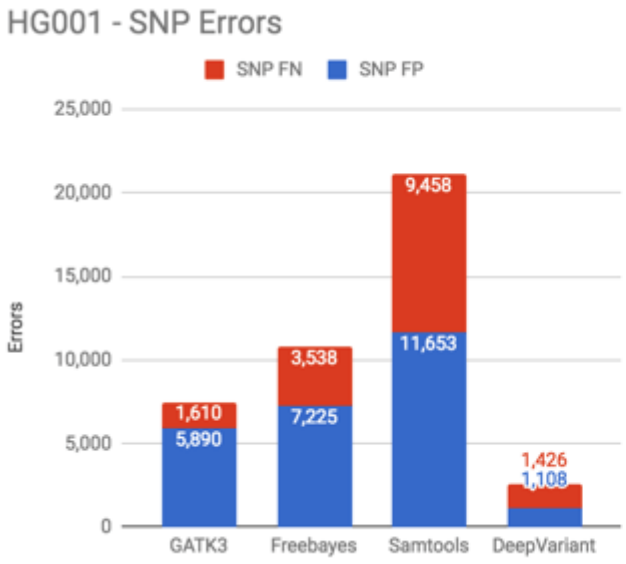
这两部分都需要强大的计算力，如果搭建了 GPU 加速的 TensorFlow 环境，识别变体的速度将会更快。若再用上谷歌专门设计的 TPU，那这一步可能会变得更快、更便宜。

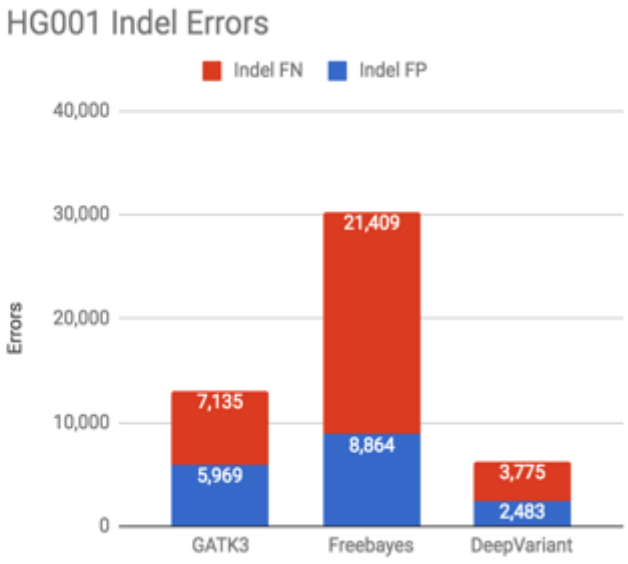
Inception 框架是一个“重量级”（heavy-weight）的深度学习架构，它在训练和应用上花费的计算成本很高。所以基因组学中的问题不能全部依赖 Inception 解决。目前在深度学习领域，为一个问题单独定制网络架构是相当费时费力的。所以使用经过验证的体系结构才是长远之计。

## DeepVariant 有多精确?

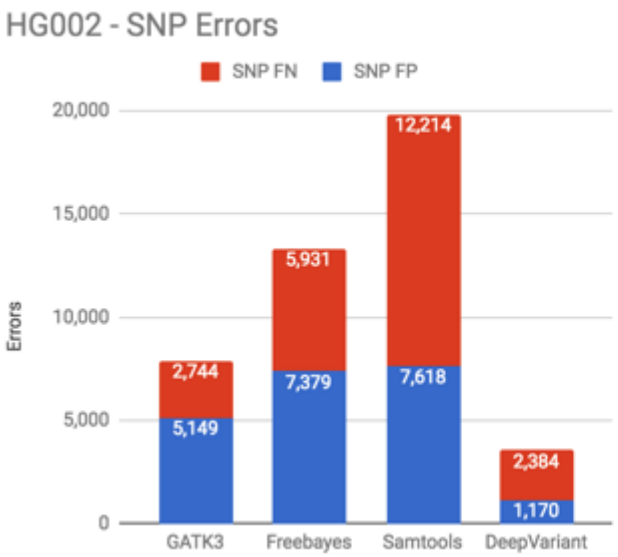
为了了解 DeepVariant 的实际表现，研究人员利用多种全基因组测序（WGS）设置，将 DeepVariant 与其他多种方式进行了比较。最终发现，DeepVariant 在各项测试中的表现均优于目前常用的方式。

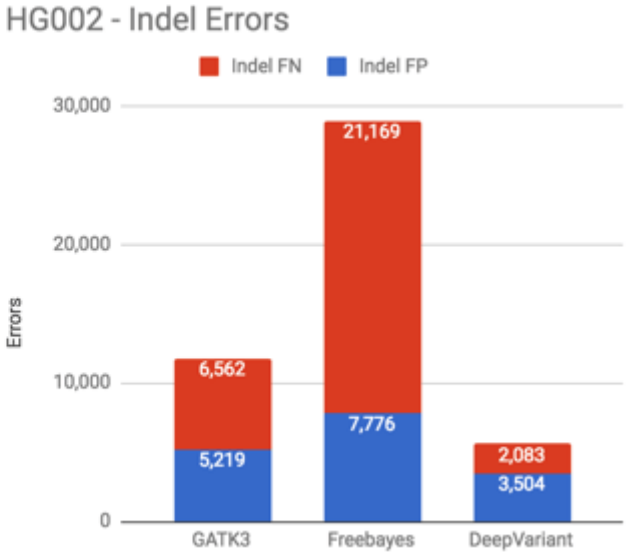
DNAnexus 的研究人员设置了三个基准：HG001、HG002 和 HG005。这些数据都是从 GIAB 的基因组中构建的。他们通过评估，能帮助客户选择最佳的分析工具。评估在 Illumia 的 hap.py 上完成。以下的图表显示了几个样本上 SNP 和 Indel 错误的数量，数字越小越好（由于 Indel 错误率高，Samtools 在 Indel 的图中没有显示)。




在这个样本上，DeepVariant 识别的 SNP 错误率明显低于其他方法，比 Samtools 的错误减少了 10 倍。在 Indel 上，DeepVariant 也是妥妥的赢家。



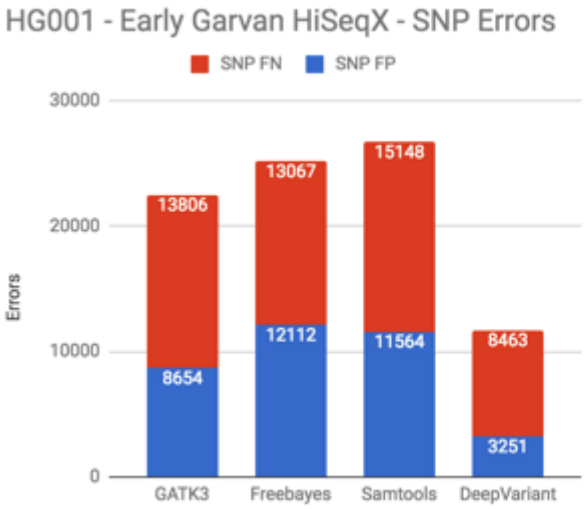

换个样本，DeepVariant 也是稳赢。

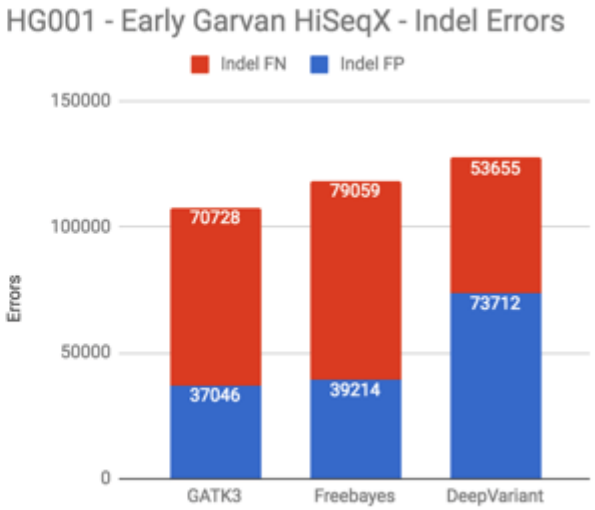
## 不同基准上的评估

做完了标准基准评估，研究人员想看看有没有能让 DeepVariant 表现不佳的样本。他们有些担心机器学习模型可能过度适应了他们的训练条件。

2014 年，Garvan 研究所通过 DNAnexus 首次公开发布 HiSeqX Genome。然而，新测序仪第一次运行测试的结果质量不如几年后生成的结果。2016 年，Garvan 为 PrecisionFDA 挑战赛提供了一个无 PCR 的 HiSeqX，作为高质量的数据集。

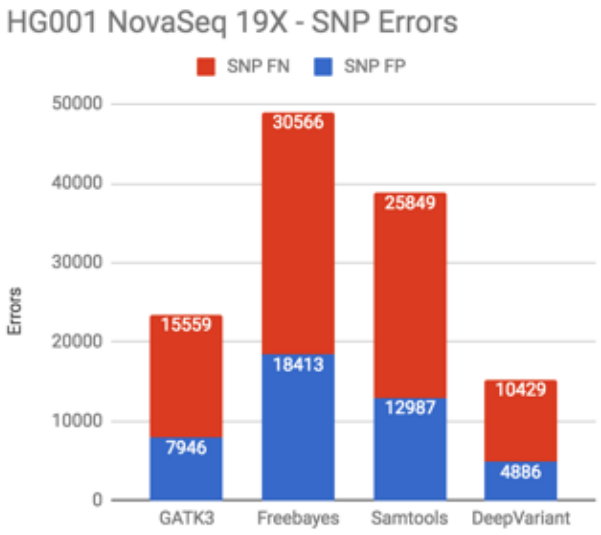
为更好地评估 DeepVariant 在不同样本上的表现，研究人员将它与其他开源方法同时应用到基因组中。

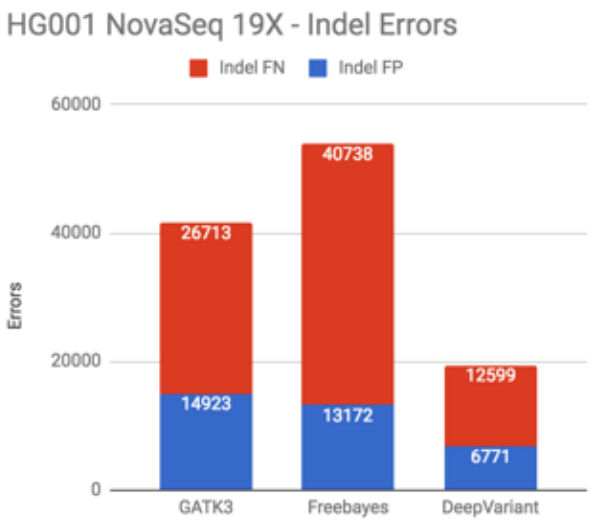



在 2014 年的 Garvan HiSeqX 中，DeepVariant 在 SNP 识别上遥遥领先。但是在 Indel 识别中表现最差，每一类都有超过 10 万个错误。

为了进一步测试 DeepVariant，研究人员将其应用至 NovaSeq 上，是 Illumia 今年新推出的测序仪器。它们使用了 BaseSpace 的 NA12878-I30 作为样本，测序深度从 35X 降到 19X。



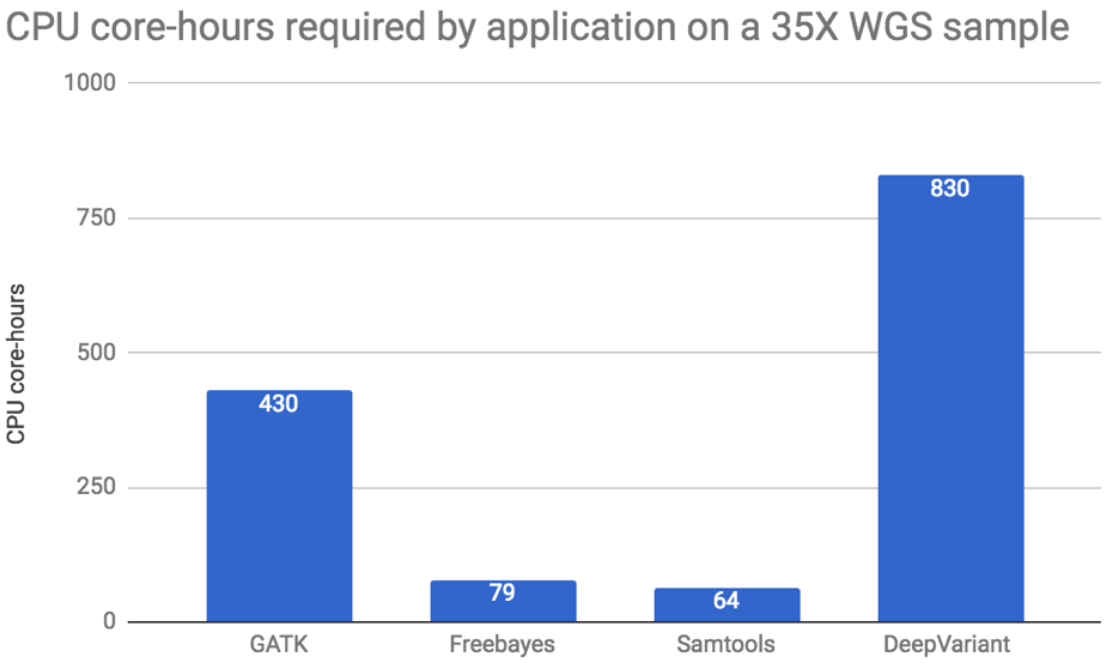

结果显示，即使在 NovaSeq 的低覆盖率这样的版本上，DeepVariant 也比其他版本优秀。现在看来，不论是什么样本，也不论在何种机器、测序深度上，DeepVariant 的表现都是最好的。

除了这里展示出来的对比图，研究人员还在 35X 的 NovaSeq 数据集、2016 版 HiSeqX Garvan 高质量样本以及 HG005 上做了对比，结果大致相同。

## DeepVariant 需要多大的计算能力?

之前提到，DeepVariant 虽然拥有高精确度，但是对计算能力也有很高的要求。虽然 GPU（或 TPU）能减轻一些负担，但所需要的计算强度仍然很高。

下图展示了在不使用 GPU 的情况下完成 HG001 的 CPU 运行时间（数字越低越好）：




不过，DNAnexus 表示他们的云平台能够以更低的成本实现云资源的广泛并行。同时运行多台机器，可以在几小时内做完原本需要 830 个小时的任务。

## 结束语

十多年来，专家一直在对下一代测序中的 SNP 和 Indel 问题进行改进。而 DeepVariant 的作者利用深度学习框架，在短短几年时间里就弥补了传统方法的不足，让基因组测序的准确度更上一层楼。

DeepVariant 真正厉害的地方不在于它能精准地识别变体（这一领域早已成熟），而是它为深度学习在生物医疗领域做出的贡献，能让科学家们在这个新兴领域迅速实现以往需要几十年才能取得的成就。

滴雨科技具有ＧＰＵ和ＴＦ ｏｎ ｓｐａｒｋ的平台，已平滑移植了第二代和第三代基于大数据平台的基因测序和分析技术，同时也是全方位的生物信息分析平台。

DEW 科技同时集成了最新的深度学习框架： Tensorflow on spark， Pytorch Biggraph，同时通过 PG-STORM 平台支持数据库的加速。这些系统由于是构建在第二代分布式文件系统之上，因而有更快的速度和超级性能，是真正能支持万级节点及生物人工智能数据大分析第三代平台。