**GTX Compressor使用报告**  
**（技术预览版）**

人和未来生物科技有限公司

2017年1月

摘要

自2010年以来，随着新一代测序技术的发展，更大数量级的基因组数据产出日渐增加（从GB，TB级到PB，EB级。Illumina公司最新的推出的HISEQ X10测序仪3天内测序约1.8TB的测序数据）。大规模的基因组数据的分析和管理正在成为推动生命科学创新的重要源泉。由于基因测序规模庞大，其传输与存储都需要消耗大量的时间与经济成本，制约了生物大数据发展。

GTX Compressor是Genetalks公司GTX Lab实验室开发的通用数据压缩打包系统，可以对任意基因测序数据的目录进行高压缩率的快速打包，形成单个压缩数据文件，以方便存档与远程传输、校验。GTX Compressor可以以超过114MB/s的速度将接近200GB大小的文件在29分钟内压缩到原大小的13%，而对于X10等只有7个质量数的FASTQ数据，其压缩率更可以达到5.5%。

**GTX Compressor提供“直压上云”功能**。考虑商业使用时，用户不仅需要将测序产生的海量数据存储于本地，更迫切地寻求将数据快速稳定传输至云端的能力。 GTX Compressor的数据压缩引擎允许用户直接将fastq文件压缩存储到亚马逊AWS平台或者阿里云OSS平台，并保持与本地压缩相同的压缩速度与压缩效率。

# 系统特点

该数据打包压缩系统的特点：

* + **高压缩比：**采用Context Model压缩技术，配合多种优化的预测模型，平衡系统并发度与内存资源消耗后，能达到极高的压缩率。
  + **高性能：**GTX compressor充分发挥了CPU的并发性以及新型Haswell CPU体系结构与AVX2、BMI2等指令集的计算能力，使得在普通服务器上的压缩速度，最高能够以接近114MB/s的输入流量输入数据并压缩完毕。
  + **高速直压上云：** GTX compressor支持直压上云和从云端直接解压下载功能。普通的20核服务器，通过百兆Intenet线路，可以在短短30分钟内稳定地将200GB Fastq文件的直压上云。

# 软件操作手册（技术预览版）

## 2.1 命令行说明

执行 ./gtz -h，输出命令行帮助说明

|  |
| --- |
| USAGE:  ./gtz [-f] [--endpoint <string>] [--timeout <string>]  [--secret-access-key <string>] [--access-key-id <string>] [-b  <string>] [-s <string>] [-t] [-n <string>] [-l <string>] [-i]  [-d] [--delete] [-a] [-g <number>] [-o <string>] [--] [--version]  [-h] <file names> ... |

通用选项说明：

-h：输出以上命令行帮助信息

--version：输出gt\_compress程序的版本号

压缩选项说明：

-f, --force ： 强制删除容器内的object

--endpoint ： 指定阿里云OSS平台的访问域名和数据中心

--timeout ： 指定上传超时阀值

--access-key-id : 指定云平台用户ID

--secret-access-key： 指定云平台用户密钥

-i：压缩时增加索引，主要用于在压缩文件中快速检索fastq文件的某段内容，该选项会降低压缩速度

-a：追加模式，本次压缩的内容会追加到压缩文件中

-g：分组加速压缩，分组越多，需要的cpu和内存越多，压缩速度越快。不指定该值时，程序会根据cpu和内存自动选择最优值

-o：指定压缩文件名，不指定时，默认为out.gtz

file\_name：需要压缩的文件或目录, 若不指定，则从标准输入中读入数据

解压选项说明：

-d,--decode : 解压模式

-f, --force ： 强制删除容器内的object

--endpoint ： 指定阿里云OSS平台的访问域名和数据中心

--timeout ： 指定下载超时阀值

--access-key-id : 指定云平台用户ID

--secret-access-key： 指定云平台用户密钥

-t,--terminal : 解压数据输出至标准输出

-i,--index : 解压模式下，启用索引功能

-l, --line : 指定需检索的起始行，起始行在fastq中的位置为（该值-1)\*4

-n, --number : 指定需要检索输出的行数，输出的行数为该值\*4，负数为向后检索

-o：指定输出文件名，使用-n或-l时需要指定该选项，否则不需要该选项

file\_name：需要压缩的文件或目录, 若不指定，则从标准输入中读入数据

## 2.2 压缩和解压示例

为了与阿里云OSS平台和亚马逊AWS平台命令保持一致，-o指定上传文件路径或者-d指定文件解压路径方式如下：

oss://bucket /directory/filename.gtz

s3://bucket / directory / filename.gtz

第一部分oss/s3为用户要上传的云平台

第二部分bucket为用户要上传到的bucket

第三部分directory为用户在bucket中存储数据的目录

第四部分filename.gtz为用户在目录中存储的压缩文件名称

### 2.2.1直压上阿里云的 OSS与从OSS直接解压

通过将云平台的相关信息（access-key-id, secret-access-key, endpoint）配置成环境变量的方式，压缩软件可以自动读取。例如：

|  |
| --- |
| export access\_key\_id=xxxxxx  export secret\_access\_key=xxxxxx  export endpoint=xxxxxx （该环境变量只有上传至OSS时才需设置） |

若不在环境变量中设置密钥，也可以通过gtz命令行参数具体指定。

**注：以下示例，均以环境变量已经设置为前提。**

#### 直压上传至阿里云OSS

./gtz -o oss://gt-compress/dest.gtz source.fastq

或者，若fastq文件是gzip压缩过的（如：source.fastq.gz），可以通过下面方式边gzip解压、边gtz重压缩、边直传：

zcat source.fastq.gz | ./gtz -o oss://gt-compress/dest.gtz

#### 从阿里云OSS直接解压

./gtz -d oss://gt-compress/dest.gtz

### 2.2.2直压上AWS S3与从S3直接解压

通过将云平台的相关信息（access-key-id, secret-access-key）配置成环境变量的方式，压缩软件可以自动读取。例如：

|  |
| --- |
| export access\_key\_id=xxxxxx  export secret\_access\_key=xxxxxx |

若不在环境变量中设置密钥，也可以通过gtz命令行参数具体指定。

**注：以下示例，均以环境变量已经设置为前提。**

#### 直压上传至AWS S3

./gtz -o s3://gt-compress/dest.gtz source.fastq

或者，若fastq文件是gzip压缩过的（如：source.fastq.gz），可以通过下面方式边gzip解压、边gtz重压缩、边直传：

zcat source.fastq.gz | ./gtz -o s3://gt-compress/dest.gtz

#### 从AWS S3直接解压

./gtz -d s3://gt-compress/dest.gtz

### 2.2.3 压缩至本地磁盘

#### 压缩过程

./gtz -o /gt-compress/dest.gtz source.fastq

或者，若fastq文件是gzip压缩过的（如：source.fastq.gz），可以通过下面方式边gzip解压、边gtz重压缩、边直传：

zcat source.fastq.gz | ./gtz -o /gt-compress/dest.gtz

#### 从本地文件解压

./gtz -d /gt-compress/dest.gtz