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 3 月

修订记录

版本号	修订内容	修订日期
gtz_0.2.0	1、 增加了新算法，性能开销小，默认使用该算法 2、 增加了对压缩文件的 md5 校验，上云过程也增加了 md5 校验 3、 增加了压缩目录功能，并提供抽取解压压缩包中某些文件功能 4、 变更了命令参数，-t 改为-c，增加了-e 和-list 项	2017/3/20

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 [--list] [--mixing] [-e <string>] [-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

-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

--list : 列出压缩包中所有的压缩文件名，与-d 参数一起使用

-e, --extract : 解压压缩包中指定的压缩文件，文件名之间用冒号:分割，与-d 参数一起使用

-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

-c,--stdout ： 解压数据输出至标准输出

-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
```

2.2.4 压缩目录

- 压缩过程

例如，test 目录下有 source1.fastq，source2.fastq，source3.fastq 三个文件，压缩目录的命令如下：

```
./gtz -o dest.gtz test
```

- 解压整个压缩包

```
./gtz -d dest.gtz
```

- 查看压缩包的文件名

列出压缩包中所有压缩文件的名称

```
./gtz -list -d dest.gtz
```

- 解压某些文件

例如，解压 dest.gtz 中 test 目录下的 source1.fastq 和 source2.fastq 两个压缩文件，命令行如下：

./gtz -e “test/source1.fastq:test/source2.fastq” -d dest.gtz