Introduction to GDS

Stephanie Gogarten

Genomic Data Structure

CoreArray (C++ library)

- designed for large-scale data management of genome-wide variants
- data format (GDS) to store multiple array-oriented datasets in a single file

R packages

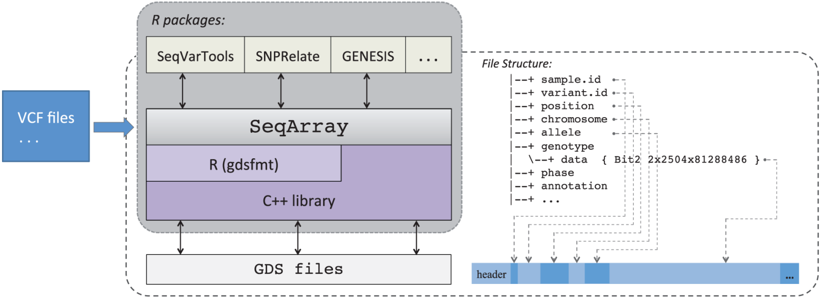
Package	Author	Description
gdsfmt	Xiuwen Zheng	R interface to CoreArray Genomic Data Structure (GDS) files
SeqArray	Xiuwen Zheng	data management of genome-wide sequence variants from Variant Call Format (VCF) files
SeqVarTools	Stephanie Gogarten	additional functions for common tasks (display genotypes, summary statistics, HWE, TiTv)
SNPRelate	Xiuwen Zheng	a parallel computing toolset for relatedness and principal component analysis
GENESIS	Matthew Conomos	relatedness in the presence of admixture, mixed models

https://github.com/zhengxwen/SeqArray/blob/master/vignettes/OverviewSlides.Rmd

GDS – Advantages

- SeqArray provides the same capabilities as VCF
- Stores data in a binary and array-oriented manner
  - efficient operations specifically designed for integers of less than 8 bits
- Genotype compression
  - 2-bit array to store alleles (95% sites are bi-allelic)
  - rare variants: highly compressed
  - 1KG, 203.5 billion genotypes, saved in 4.3G (2.26% if a byte stores a genotype)
  - zlib, lzma: decompression with relatively efficient random access

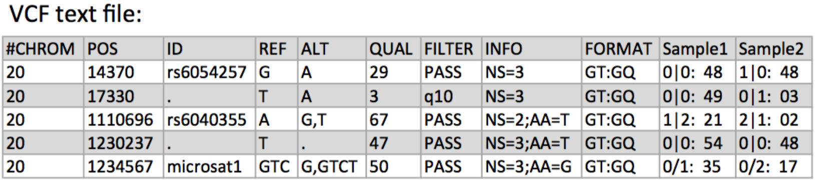
SeqArray framework



File Structure:

```
|--+ sample.id
|--+ variant.id
|--+ position
|--+ chromosome
|--+ allele
|--+ genotype
|  \--+ data  { Bit2 2x2504x81288486 }
|--+ phase
|--+ annotation
|--+ ...
```

VCF Reformatting

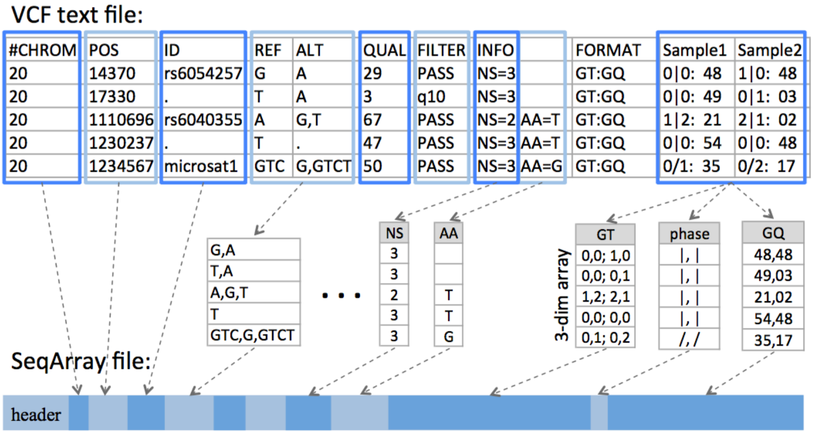


VCF text file:

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	Sample1	Sample2
20	14370	rs6054257	G	A	29	PASS	NS=3	GT:GQ	0|0: 48	1|0: 48
20	17330	.	T	A	3	q10	NS=3	GT:GQ	0|0: 49	0|1: 03
20	1110696	rs6040355	A	G,T	67	PASS	NS=2;AA=T	GT:GQ	1|2: 21	2|1: 02
20	1230237	.	T	.	47	PASS	NS=3;AA=T	GT:GQ	0|0: 54	0|0: 48
20	1234567	microsat1	GTC	G,GTCT	50	PASS	NS=3;AA=G	GT:GQ	0/1: 35	0/2: 17

- VCF consists of a header section and a data section
- each variant is stored in a line
- genotypes
  - GT: alleles + phasing states
- annotations in INFO and FORMAT fields
  - NS, AA, GQ

VCF Reformatting



VCF text file:

<table>
<tr><td>#CHROM</td><td>POS</td><td>ID</td><td>REF</td><td>ALT</td><td>QUAL</td><td>FILTER</td><td>INFO</td><td></td><td>FORMAT</td><td>Sample1</td><td>Sample2</td></tr>
<tr><td>20</td><td>14370</td><td>rs6054257</td><td>G</td><td>A</td><td>29</td><td>PASS</td><td>NS=3</td><td></td><td>GT:GQ</td><td>0|0: 48</td><td>1|0: 48</td></tr>
<tr><td>20</td><td>17330</td><td>.</td><td>T</td><td>A</td><td>3</td><td>q10</td><td>NS=3</td><td></td><td>GT:GQ</td><td>0|0: 49</td><td>0|1: 03</td></tr>
<tr><td>20</td><td>1110696</td><td>rs6040355</td><td>A</td><td>G,T</td><td>67</td><td>PASS</td><td>NS=2</td><td>AA=T</td><td>GT:GQ</td><td>1|2: 21</td><td>2|1: 02</td></tr>
<tr><td>20</td><td>1230237</td><td>.</td><td>T</td><td>.</td><td>47</td><td>PASS</td><td>NS=3</td><td>AA=T</td><td>GT:GQ</td><td>0|0: 54</td><td>0|0: 48</td></tr>
<tr><td>20</td><td>1234567</td><td>microsat1</td><td>GTC</td><td>G,GTCT</td><td>50</td><td>PASS</td><td>NS=3</td><td>AA=G</td><td>GT:GQ</td><td>0/1: 35</td><td>0/2: 17</td></tr>
</table>

G,A
T,A
A,G,T
T
GTC,G,GTCT

. . .

NS
3
3
2
3
3

AA
T
T
G

3-dim array

<table>
<tr><td>GT</td></tr>
<tr><td>0,0; 1,0</td></tr>
<tr><td>0,0; 0,1</td></tr>
<tr><td>1,2; 2,1</td></tr>
<tr><td>0,0; 0,0</td></tr>
<tr><td>0,1; 0,2</td></tr>
</table>

<table>
<tr><td>phase</td></tr>
<tr><td>|, |</td></tr>
<tr><td>|, |</td></tr>
<tr><td>|, |</td></tr>
<tr><td>|, |</td></tr>
<tr><td>/, /</td></tr>
</table>

GQ
48,48
49,03
21,02
54,48
35,17

SeqArray file:

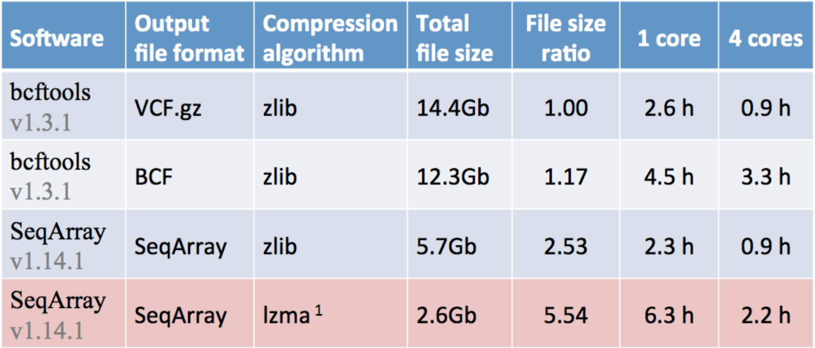
3D array has dimensions [allele, sample, variant]

Annotation storage

- Per-sample annotations
  - DP, AD, GQ, PL
  - stored as integers according to the VCF Specification (VCFv4.0)
- SeqArray
  - most integers can be saved in 1 or 2 bytes
  - variable-length integer encoding instead of 32 bits
  - storage-efficient
  - reduce compression time

Performance - file conversion and compression

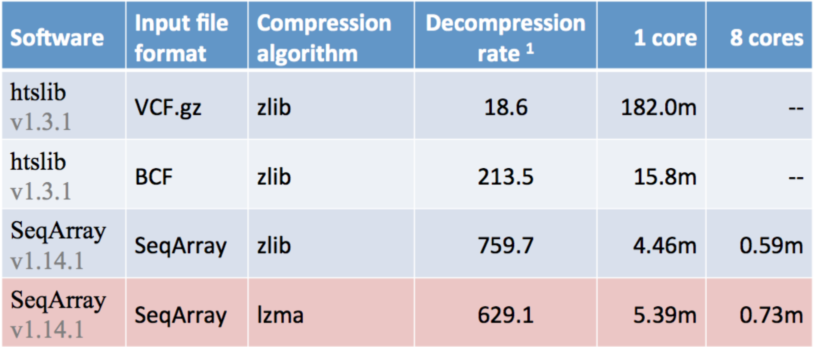
22 autosomal compressed genotype VCF files of 1000G Phase3 (2,504 samples, 81M variants) converted to a single output file



Software	Output file format	Compression algorithm	Total file size	File size ratio	1 core	4 cores
bcftools v1.3.1	VCF.gz	zlib	14.4Gb	1.00	2.6 h	0.9 h
bcftools v1.3.1	BCF	zlib	12.3Gb	1.17	4.5 h	3.3 h
SeqArray v1.14.1	SeqArray	zlib	5.7Gb	2.53	2.3 h	0.9 h
SeqArray v1.14.1	SeqArray	lzma [1]	2.6Gb	5.54	6.3 h	2.2 h

[1] the default compression algorithm used in SeqArray

Performance - genotype decompression

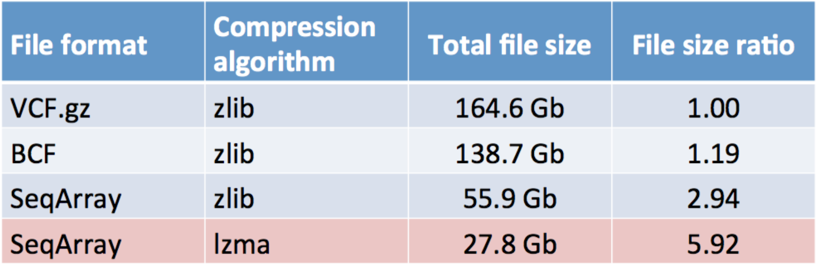


Software	Input file format	Compression algorithm	Decompression rate [1]	1 core	8 cores
htslib v1.3.1	VCF.gz	zlib	18.6	182.0m	--
htslib v1.3.1	BCF	zlib	213.5	15.8m	--
SeqArray v1.14.1	SeqArray	zlib	759.7	4.46m	0.59m
SeqArray v1.14.1	SeqArray	lzma	629.1	5.39m	0.73m

1: million genotypes per second on one core

TOPMed Storage Use

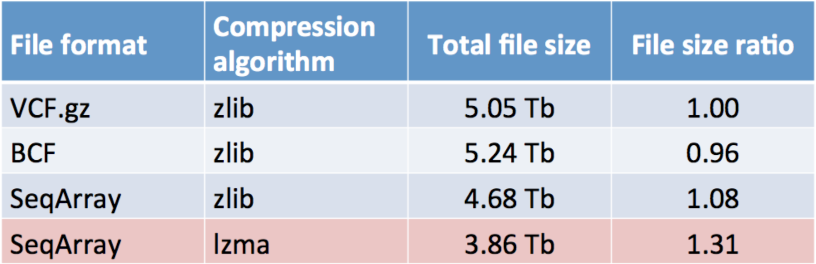
Comparison of file sizes for genotypes only: TOPMed Freeze4 (18,526 samples and 219M variants).



File format	Compression algorithm	Total file size	File size ratio
VCF.gz	zlib	164.6 Gb	1.00
BCF	zlib	138.7 Gb	1.19
SeqArray	zlib	55.9 Gb	2.94
SeqArray	lzma	27.8 Gb	5.92

TOPMed Storage Use

Comparison of file sizes for genotypes and per-sample annotations (PL, GQ, AD, OD)[1]: TOPMed Freeze2 (9,109 samples and 140M variants).



File format	Compression algorithm	Total file size	File size ratio
VCF.gz	zlib	5.05 Tb	1.00
BCF	zlib	5.24 Tb	0.96
SeqArray	zlib	4.68 Tb	1.08
SeqArray	lzma	3.86 Tb	1.31

[1] PL - Phred-scaled genotype likelihoods; AD - allelic depths; GQ - conditional genotype quality; OD - other allelic depths

SeqArray Key Functions

Function	Description
seqVCF2GDS	Reformat a VCF file
seqSetFilter	Define a subset of samples or variants
seqGetData	Get data with a defined filter
seqApply	Apply a user-defined function over array margins
seqParallel	Apply a function in parallel

SeqVarTools R package

https://github.com/smgogarten/SeqVarTools

- Combine GDS file with sample and variant annotation
- Iterator classes
- User-friendly retrieval and display of genotypes
- Summary statistics (SNV counts, het/hom, TiTv)
- HWE, Mendelian errors, concordance checking

SNPRelate R package

https://github.com/zhengxwen/SNPRelate

- LD pruning
- Kinship estimates
  - Multiple methods including KING-robust
- Principal Component Analysis
- Genetic Relatedness Matrix

GENESIS R package

https://github.com/smgogarten/GENESIS

- Relatedness and population structure
  - Select a set of unrelated samples, maximizing ancestral diversity
  - PCA on unrelated samples, projection of relatives
  - Kinship estimates adjusted for PCs
- Mixed-model association testing
  - Fit a null mixed model with fixed effects and one or more random effects (e.g., kinship, GRM)
  - Single-variant tests
  - Rare variant tests
    - Burden or SKAT
    - Sliding window or user-defined aggregation units (e.g., genes)

Extra

GDS – File Contents

```
File: SeqArray/extdata/CEU_Exon.gds (387.3K)
|--+ description   [  ] *
|--+ sample.id   { Str8 90 ZIP_ra(30.8%), 222B }
|--+ variant.id   { Int32 1348 ZIP_ra(35.7%), 1.9K }
|--+ position   { Int32 1348 ZIP_ra(86.4%), 4.6K }
|--+ chromosome   { Str8 1348 ZIP_ra(2.66%), 91B }
|--+ allele   { Str8 1348 ZIP_ra(17.2%), 928B }
|--+ genotype   [  ] *
|  \--+ data   { Bit2 2x90x1348 ZIP_ra(28.4%), 16.8K } *
|--+ phase   [  ]
|  \--+ data   { Bit1 90x1348 ZIP_ra(0.36%), 55B } *
|--+ annotation   [  ]
|  |--+ id   { Str8 1348 ZIP_ra(41.0%), 5.8K }
|  |--+ qual   { Float32 1348 ZIP_ra(0.91%), 49B }
|  |--+ filter   { Int32,factor 1348 ZIP_ra(0.89%), 48B } *
|  |--+ info   [  ]
|  |  |--+ AA   { Str8 1348 ZIP_ra(24.2%), 653B } *
|  |  \--+ HM2   { Bit1 1348 ZIP_ra(117.2%), 198B } *
|  \--+ format   [  ]
|     \--+ DP   [  ] *
|        \--+ data   { Int32 90x1348 ZIP_ra(33.8%), 160.3K }
\--+ sample.annotation   [  ]
   \--+ family   { Str8 90 ZIP_ra(34.7%), 135B }
```

Diploid Genotype Decoding

Raw 2-bit array (M_{2x3x3})

M[1, ,]: *one haploid*

	variant 1	variant 2	
sample 1	**0**	**0**	**1**
sample 2	**1**	**3**	**1**
sample 3	**3**	**3**	**3**

M[2, ,]: *the other haploid*

	variant 1	variant 2	
sample 1	**0**	**0**	**0**
sample 2	**1**	**3**	**0**
sample 3	**2**	**1**	**3**

	variant 1	variant 2
extra vector	**1**	**2**

indicating how many bits used for each variant

Genotype array (G_{2x3x2})

G[1, ,]: *one haploid*

	variant 1	variant 2
sample 1	**0**	**4**
sample 2	**1**	**7**
sample 3	**NA**	**NA**

G[2, ,]: *the other haploid*

	variant 1	variant 2
sample 1	**0**	**0**
sample 2	**1**	**3**
sample 3	**2**	**13**

```
G[,,1] := (M[,,1] == 3) ? NA : M[,,1]
G[,,2] := (M[,,2] + M[,,3]*4 == 15) ? NA : M[,,2] + M[,,3]*4
```

Diploid Genotype Decoding

Raw 2-bit array (M_{2x3x3})

M[1, ,]: *one haploid*

	variant 1	variant 2	
sample 1	**0**	**0**	**1**
sample 2	**1**	**3**	**1**
sample 3	**3**	**3**	**3**

M[2, ,]: *the other haploid*

	variant 1	variant 2	
sample 1	**0**	**0**	**0**
sample 2	**1**	**3**	**0**
sample 3	**2**	**1**	**3**

	variant 1	variant 2
extra vector	**1**	**2**

indicating how many bits used for each variant

(for an arbitrary number of unique alleles)

Genotype array (G_{2x3x2})

G[1, ,]: *one haploid*

	variant 1	variant 2
sample 1	**0**	**4**
sample 2	**1**	**7**
sample 3	**NA**	**NA**

G[2, ,]: *the other haploid*

	variant 1	variant 2
sample 1	**0**	**0**
sample 2	**1**	**3**
sample 3	**2**	**13**

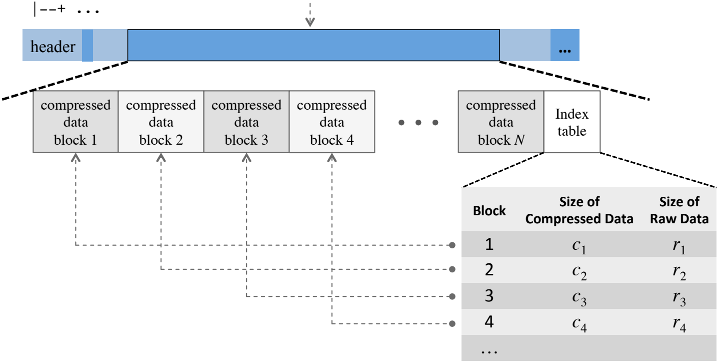
G[,,1] := (M[,,1] == 3) ? NA : M[,,1]

G[,,2] := (M[,,2] + M[,,3]*4 == 15) ? NA : M[,,2] + M[,,3]*4

Indexing Strategy

SeqArray File Structure:

```
|--+ genotype
|   \--+ data  { Bit2 2x2504x81288486 }
|--+ ...
```



Block	Size of Compressed Data	Size of Raw Data
1	c_1	r_1
2	c_2	r_2
3	c_3	r_3
4	c_4	r_4
…		