PUBLIC ACCESS

GENOMED4ALL

D8.1
Genomics data standardization plan

GENOMED4ALL receives funding from the European Union's Horizon 2020 Research and Innovation programme under Grant Agreement No. 101017549

D8.1

Genomics data standardization plan

Revision **v0.1**

Work package	WP8
Task	T8.1
Due date	30-06-2021
Deliverable lead	CRG – Centre for Genomic Regulation
Version	V1.0
Authors	Babita Singh (CRG)
Reviewers	Anne Laure Phamhung-D'Alexandry (APHP) Carlotta Cattaneo (ICH)

Abstract

Minimum set of recommended standards for the analysis and sharing of the genomics dataset for hematology and oncology to be utilized in GenoMed4All project. Further, guidelines are also described to share phenotypic and other metadata collected from different clinical partners. A brief description of data harmonization and sharing using GA4GH Phenopackets and FHIR standards, evaluation of best suited exchange format for genomics data and how to adapt and calibrate standardized data on local sites according to evaluation of clinical data set and genomics interface.

Keywords

Genomics data standards, VCF, SAM/BAM standards, Phenopackets, FHIR, Beacon

This project has received funding from the European Union's Horizon 2020 Research and Innovation programme under Grant Agreement No. **101017549**

Document revision history

Version	Date	Description of change	Contributor(s)
v0.1	25-06-2021	First version	Babita Singh (CRG)
v0.2	26-06-2021	1st round of peer review	Anne Laure Phamhung-D'Alexandry (APHP)
v0.3	28-06-2021	2nd round of peer review	Carlotta Cattaneo (ICH)
v1.0	30-06-2021	Final review	Annelore Hermann (UPM)

Disclaimer

Copyright notice

Project co-funded by the European Commission in the H2020 Programme

Nature of the deliverable		R
Dissemination level		
PU	Public, fully open. e.g., website	✔
CL	Classified information as referred to in Commission Decision 2001/844/EC	
CO	Confidential to GenoMed4All project and Commission Services	

*** Deliverable types:**

R: document, report (excluding periodic and final reports).
DEM: demonstrator, pilot, prototype, plan designs.
DEC: websites, patent filings, press and media actions, videos, etc.
OTHER: software, technical diagrams, etc.

This project has received funding from the European Union's Horizon 2020 Research and Innovation program under Grant Agreement No. **101017549**

Table of contents

Executive summary **5**
1 Genomics data standardization recommendations **6**
1.1 Prior art 6
1.2 Genomics dataset 6
1.2.1 Raw reads, mapped reads and alignment data files 6
1.2.2 Unmapped sequence data files 7
1.2.3 Mapping data files (SAM/BAM) Standards 7
1.2.4 Variant calling data files 9
1.2.5 Further readings and recommendations 14
2 Phenotypes, Metadata and Clinical Data exchange **15**
2.1 Recommended Model for Metadata Sharing 16
3 Recommendations for Genomics and Metadata sharing **19**
3.1 Beacon 19
4 References **20**

List of figures

Figure 1. A typical DNA sequencing pipeline **[4]** 7
Figure 2. An example of SAM file 8
Figure 3. An example of VCF file 10
Figure 4. Example of a structured line 11
Figure 5. Example of structural variant 13
Figure 7. An example of phenopacket describing the pathogenicity of a variant 16
Figure 8. An example of FHIR resource of a patient 17

List of tables

Table 1. SAM/BAM files – format description 9
Table 2. VCF files – Data Lines Fixed fields 12
Table 3. VCF files – Genotype fields 12

This project has received funding from the European Union's Horizon 2020 Research and Innovation program under Grant Agreement No. **101017549**

Abbreviations

HDs	Hematological diseases
MDS	Myelodysplastic Syndromes
MM	Multiple Myeloma
SCD	Sickle Cell Disease
FAIR	Findable, accessible, interoperable, reusable
FHIR	Fast Healthcare Interoperability Resources
GA4GH	Global Alliance for Genomics and Health
EGA	European Genomics Archive
SAM	Sequence Alignment/Map format
BAM	Binary Alignment/Map format
VCF	Variant Calling Format
PXF	Phenotype Exchange Format

This project has received funding from the European Union's Horizon 2020 Research and Innovation program under Grant Agreement No. **101017549**

Executive summary

This deliverable aims to provide a minimum set of standard guidelines for genomics, phenotypic and metadata dataset generated and/or utilized in GenoMed4All consortium. This set of guidelines will also be useful when sharing genomics data generated by different clinical partners outside the consortium, in order to set up common data model and format. An effective data standard is necessary to assimilate healthcare information from different resources and make them interoperable for federated learning purposes.

Further, special emphasis needs be to given for associated phenotypic and metadata generated alongside the genomics dataset of the patient. Such data needs to be properly anonymized, harmonized and produced in machine-readable formats in order to extract maximum information for the development of AI models. We have also mentioned in brief the use cases to adopt such model, current available common data model system such as FHIR (Fast Healthcare Interoperability Resources) and Phenopackets for metadata sharing, recommended implementation of FAIR principles (findable, accessible, interoperable, reusable) [7] on the dataset in use and Beacon v2 for genomics and metadata data sharing.

This project has received funding from the European Union's Horizon 2020 Research and Innovation program under Grant Agreement No. **101017549**

1 Recommendations for Genomics data standardization

1.1 Prior art

The Global Alliance for Genomics and Health (**GA4GH**) is an international consortium that is currently leading most of the efforts in the creation of standards for genomics data [1]. GA4GH proposes standards and good practices for responsible collection, storage and sharing of genomics data. In Europe, European Genomics Archive (EGA) is one of the main leaders for implementing GA4GH data standards and guiding the forefront for others **[2]**. This document will cover the GA4GH guidelines and standards recommended for good data practices on omics dataset, along with EGA's own experience of handing sensitive genomics and clinical dataset.

1.2 Genomics dataset

Genomics dataset can be obtained through different file formats based on the phases. The data is generated, such as, raw reads generated in FASTQ format, alignment data files obtained after mapping raw reads, post-processed files obtained after tertiary data analysis such as variant calling. We will describe the recommended standard for the files obtained after each phase.

1.2.1 Raw reads, mapped reads and alignment data files

Next generation genome sequencing technologies such as Illumina, PacBio and Oxford Nanopore generates several million to billion short (75 to 300 base pairs) or long read sequences (500 bp to current record of 2.3Mb **[3]**) that is typically obtained in raw reads FASTQ format.

The bioinformatics pipeline for a typical DNA sequencing strategy involves to align these raw FASTQ reads to reference genome (such as human) using genome sequencing alignment tool of choice (Bowtie, STAR, TopHat etc) as shown in **Figure 1 [4]**. The sequence alignment process assigns a genomic position to the short/long reads obtained from the sequencer to where they mapped in the reference genome, along with other metadata fields. The aligned sequences and the related metadata are then stored in a Sequence Alignment Mapping (SAM/BAM) or CRAM file format (**Figure 1**). This is usually followed by 'tertiary analysis' of the data using downstream algorithms that consume the BAM file to identify a range of genetic alterations, including single nucleotide variants, insertions and deletions (indels), and tumor mutation burden **[4]**.

This project has received funding from the European Union's Horizon 2020 Research and Innovation program under Grant Agreement No. **101017549**

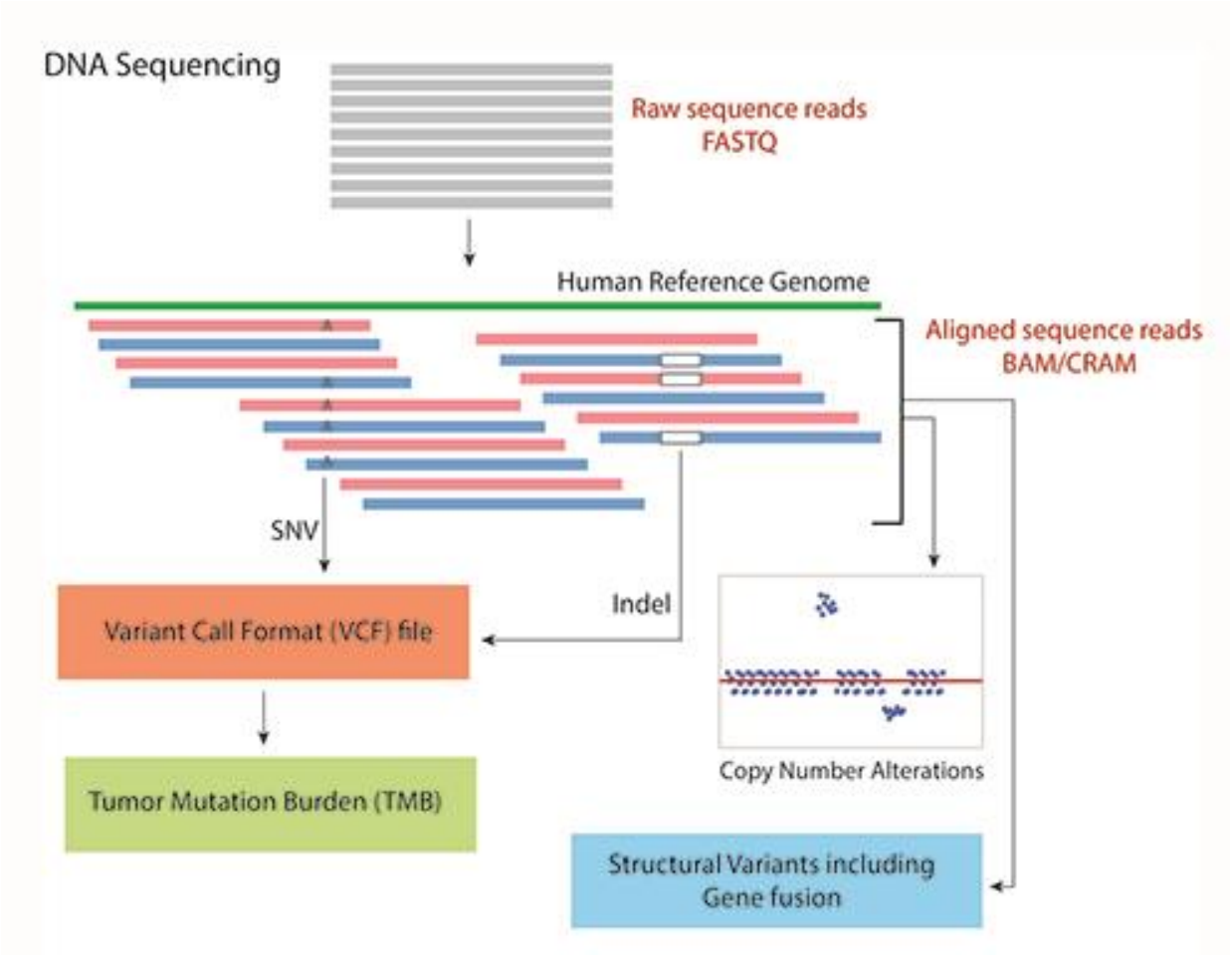

Figure 1. A typical DNA sequencing pipeline **[4]**

1.2.2 Unmapped sequence data files

In the scope of this deliverable, we do not define or endorse any dedicated unaligned sequence data format. However, we recommend storing such data in one of the alignment formats (SAM, BAM, or CRAM) with the unmapped flag set.

1.2.3 Mapping data files (SAM/BAM) Standards

For the primary aligned files, SAM/BAM format is the recommended format to share **[5]**. SAM stands for Sequence Alignment/Map format and the binary version of SAM format is known as BAM files. It is a TAB-delimited text format consisting of a header section, which is optional, and an alignment section. If present, the header must be prior to the alignments. Header lines start with '@', while alignment lines do not. Each alignment line has 11 mandatory fields for essential alignment information such as mapping position, and variable number of optional fields for flexible

This project has received funding from the European Union's Horizon 2020 Research and Innovation program under Grant Agreement No. **101017549**

or aligner specific information. Following GA4GH guidelines, this document recommends following the specification for version 1.6 of the SAM and BAM formats. It is also recommended that each SAM and BAM file specifies the version being used via the @HD VN tag (as shown in **Figure 2**).

```
@HD     VN:1.0  SO:coordinate
@SQ     SN:chr20        LN:64444167
@PG     ID:TopHat       VN:2.0.14       CL:/srv/dna_tools/tophat/tophat -N 3 --read-edit-dist 5 --read-rea
lign-edit-dist 2 -i 50 -I 5000 --max-coverage-intron 5000 -M -o out /data/user446/mapping_tophat/index/chr
20 /data/user446/mapping_tophat/L6_18_GTGAAA_L007_R1_001.fastq
HWI-ST1145:74:C101DACXX:7:1102:4284:73714        16      chr20   190930  3       100M    *       0       0
       CCGTGTTTAAAGGTGGATGCGGTCACCTTCCCAGCTAGGCTTAGGGATTCTTAGTTGGCCTAGGAAATCCAGCTAGTCCTGTCTCTCAGTCCCCCCTCT
C    BBDCCDDCCDDDDCDDDDDDCDCCCDBC?DDDDDDDDDDDDDDDCCDCDDDDDDDDDDCCCCEDDDC?DDDDDDDDDDDDDDDDDDDDDBDHFFFFDC@@
    AS:i:-15        XM:i:3  XO:i:0  XG:i:0  MD:Z:55C20C13A9 NM:i:3  NH:i:2  CC:Z:=  CP:i:55352714   HI:i:0
HWI-ST1145:74:C101DACXX:7:1114:2759:41961        16      chr20   193953  50      100M    *       0       0
       TGCTGGATCATCTGGTTAGTGGCTTCTGACTCAGAGGACCTTCGTCCCCTGGGGCAGTGGACCTTCCAGTGATTCCCCTGACATAAGGGGCATGGACGA
G    DCDDDDEDDDDDDDCDDDDDDDCCCDDDCDDDDDEEC>DFFFEJJJJJJIGJJJJIHGBHHGJIJJJJJJGJJJIJJJJJIHJJJJJJHHHHHFFFFFCCC
    AS:i:-16        XM:i:3  XO:i:0  XG:i:0  MD:Z:60G16T18T3 NM:i:3  NH:i:1
HWI-ST1145:74:C101DACXX:7:1204:14760:4030        16      chr20   270877  50      100M    *       0       0
       GGCTTTATTGGTAAAAAAGGAATAGCAGATTTAATCAGAAATTCCCACCTGGCCCAGCAGCACCAACCAGAAAGAAGGGAAGAAGACAGGAAAAAACCA
C    DDDDDDDDDCCDDDDDDDDDDEEEEEEEFFFEFFEGHHHHFGDJJIHJJIJIJJJIIIIGGFJJIHIIIIJJJJJJIGHHFAHGFHJHFGGHFFFDD@BB
    AS:i:-11        XM:i:2  XO:i:0  XG:i:0  MD:Z:0A85G13    NM:i:2  NH:i:1
HWI-ST1145:74:C101DACXX:7:1210:11167:8699        0       chr20   271218  50      50M4700N50M     *       0
       0       GTGGCTCTTCCACAGGAATGTTGAGGATGACATCCATGTCTGGGGTGCACTTGGGTCTCCGAAGCAGAACATCCTCAAATATGACCTCTCG
accepted_hits.sam
```

Figure 2. An example of SAM file

1.2.3.1 Advantages of SAM/BAM files

- Optimised and flexible to store all the alignment information generated by various read-mapping tools.
- Preferred format to be generated by read-mapping tools or converted from existing alignment formats.
- Generates compact file size.
- Allows the file to be indexed by genomic position to efficiently retrieve all reads aligned to a locus.
- Most NGS sequencing tools accepts this format as is, for further visualization purposes or tertiary analysis.

1.2.3.2 Brief format description of SAM/BAM files

Each alignment file has 11 mandatory fields (columns) for essential alignment information such as mapping position, and variable number of optional fields for flexible or aligner specific information:

This project has received funding from the European Union's Horizon 2020 Research and Innovation program under Grant Agreement No. **101017549**

#	Field
1	Read Name
2	SAM flag
3	Chromosome id (if read is has no alignment, there will be a "*" here)
4	Genomic Position (1-based index, "left end of read")
5	CIGAR string (describes the position of insertions/deletions/matches in the alignment, encodes splice junctions etc)
6	Name of mate (mate pair information for paired-end sequencing, often "=")
7	Position of mate (mate pair information)
8	Template length
9	Read Sequence
10	Read Quality
11	Program specific Flags (for eg. AS is an alignment score, NH is a number of reported alignments that contains the query in the current record)

Table 1. SAM/BAM files – format description

1.2.3.3 Converting BAM to SAM and vice versa

BAM files are non-readable, compressed file that needs to be converted back to SAM format to perform analysis or quick visualisation. Likewise, for storage purposes, SAM files should be converted to BAMs. For such purpose, Samtools provides multiple methods to handle SAM/BAM files, for eg. 'samtools view' command converts an unreadable alignment in binary BAM format to a human readable SAM format. Further functionalities provided by Samtools are recommended to explore at http://www.htslib.org/

Detailed information on SAM files specification 1.6 can be accessed at http://samtools.github.io/hts-specs/SAMv1.pdf

1.2.3.4 Indexing BAM

Indexing aims to achieve fast retrieval of alignments overlapping a specified region without going through the whole alignments. BAM must be sorted by the reference ID and then the leftmost coordinate before indexing. Samtools 'index' can be utilised to generate BAM file index, which generates a file with the same name followed by a suffix *'.bai'*. Both original BAM file and corresponding *'.bai'* needs to be stored at same location.

1.2.4 Variant calling data files

The Variant Call Format (VCF) specifies the format of a text file used in bioinformatics for storing gene sequence variations. VCF files are usually obtained during tertiary analysis utilising the

This project has received funding from the European Union's Horizon 2020 Research and Innovation program under Grant Agreement No. **101017549**

aligned BAM/SAM files that were used to call the mutations, variants or genomic regions that differ from the reference genome. It consists of first few 'headers' lines that begins with "##" and then the body with the variants and genotypes information for each sample. The mandatory columns for header are the first and the last one displaying the "fileformat" and the 8 mandatory columns (#CHROM POS ID REF ALT QUAL FILTER INFO). Optional header lines may also include meta-information and must be key=value pairs. It is strongly encouraged that information lines describing the INFO, FILTER and FORMAT entries used in the body of the VCF file be included in the meta-information section (fig 3).

In order to ensure interoperability across platforms, VCF compliant implementations must support both LF ("\n") and CR+LF ("\r\n") newline conventions.

```
##fileformat=VCFv4.3
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS     ID        REF    ALT     QUAL FILTER INFO                              FORMAT      NA00001        NA00002        NA00003
20     14370   rs6054257 G      A       29   PASS   NS=3;DP=14;AF=0.5;DB;H2           GT:GQ:DP:HQ 0|0:48:1:51,51 1|0:48:8:51,51 1/1:43:5:.,.
20     17330   .         T      A       3    q10    NS=3;DP=11;AF=0.017               GT:GQ:DP:HQ 0|0:49:3:58,50 0|1:3:5:65,3   0/0:41:3
20     1110696 rs6040355 A      G,T     67   PASS   NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27 2|1:2:0:18,2   2/2:35:4
20     1230237 .         T      .       47   PASS   NS=3;DP=13;AA=T                   GT:GQ:DP:HQ 0|0:54:7:56,60 0|0:48:4:51,51 0/0:61:2
20     1234567 microsat1 GTC    G,GTCT  50   PASS   NS=3;DP=9;AA=G                    GT:GQ:DP    0/1:35:4       0/2:17:2       1/1:40:3
```

Figure 3. An example of VCF file

1.2.4.1 Meta-information lines

File meta-information is included after the ## string and must be key=value pairs. Meta-information lines are optional, but if they are present then they must be completely well-formed. Note that BCF, the binary counterpart of VCF, requires that all entries are present. It is recommended to include meta-information lines describing the entries used in the body of the VCF file.

All structured lines that have their value enclosed within "<>" require an ID which must be unique within their type. For all of the structured lines (##INFO, ##FORMAT, ##FILTER, etc.), extra fields can be included after the default fields. For example:

This project has received funding from the European Union's Horizon 2020 Research and Innovation program under Grant Agreement No. **101017549**

```
##INFO=<ID=ID,Number=number,Type=type,Description="description",Source="description",Version="128">
```

Figure 4. Example of a structured line

In the above example (fig 4), the extra fields of “Source” and “Version” are provided. Optional fields must be stored as strings even for numeric values. It is recommended in VCF and required in BCF that the header includes tags describing the reference and contigs backing the data contained in the file. These tags are based on the SQ field from the SAM spec; all tags are optional (see the VCF example above).

Meta-information lines can be in any order with the exception of ‘fileformat‘, which must come first.

fileformat	Details
##fileformat=VCFv4.3	A single ‘fileformat’ line is always required, must be the first line in the file, and details the VCF format version number. For example VCF version 4.3

INFO fields are described as follows (first four keys are required; source and version are recommended):

fileformat	Details
##fileformat=INFO	Integer, Float, Flag, Character, and String

1.2.4.2 Data Lines Fixed fields

The data lines are tab delimited. There are 8 fixed fields per record. Fixed fields are:

#	Field	Details
1	CHROM	Chromosome: An identifier from the reference genome or an angle-bracketed ID String (“”) pointing to a contig in the assembly file (cf. the ##assembly line in the header). (String, no whitespace permitted, Required).
2	POS	Position: The reference position, with the 1st base having position 1. Positions are sorted numerically, in increasing order, within each reference sequence CHROM. It is permitted to have multiple records with the same POS
3	ID	Identifier: Semicolon-separated list of unique identifiers where available. If this is a dbSNP variant the rs number(s) should be used. If there is no identifier available, then the MISSING value should be used. (String, no whitespace or semicolons permitted, duplicate values not allowed.)

This project has received funding from the European Union’s Horizon 2020 Research and Innovation program under Grant Agreement No. **101017549**

4	REF	Reference base(s): Each base must be one of A,C,G,T,N (case insensitive). Multiple bases are permitted.
5	ALT	Alternate base (s): alternate non-reference alleles, comma separated. These alleles do not have to be called in any of the samples. If there are no alternative alleles, then the missing value should be used. Base Strings made up of the bases A,C,G,T,N,*, (case insensitive)
6	QUAL	Quality: Phred-scaled quality score for the assertion made in ALT. (Numeric)
7	FILTER	Filter status: PASS if this position has passed all filters, i.e., a call is made at this position. Otherwise, if the site has not passed all filters, a semicolon-separated list of codes for filters that fail
8	INFO	Additional information: INFO fields are encoded as a semicolon-separated series of short keys with optional values in the format: =[,data]. (String, no whitespace, semicolons, or equals-signs permitted; commas are permitted only as delimiters for lists of values).

Table 2. VCF files – Data Lines Fixed fields

For detailed explanation please see guidelines described in VCFv4.2 specification here: https://samtools.github.io/hts-specs/VCFv4.2.pdf

1.2.4.3 Genotype fields

Genotype fields or GT refers to the most likely genotype of the sample. The alleles are separated by / or |. For diploid organisms, it has **0** value for reference allele and **1** for the alternate allele (non-reference allele). If genotype information is present in a VCF file, then the same types of data must be present for all samples (**Table 3**). First a FORMAT field is given specifying the data types and order (colon-separated alphanumeric String). This is followed by one field per sample, with the colon-separated data in this field corresponding to the types specified in the format. The first sub-field must always be the genotype (GT) if it is present. There are no required sub-fields. Genotype description to specify the distribution of alleles in the samples are as follows:

Genotype	Description
0/0	The sample is a homozygous reference
0/1	The sample is heterozygous (carries both reference and alternate alleles)
1/1	The sample is a homozygous alternate
./.	No genotype called or missing genotype

Table 3. VCF files – Genotype fields

This project has received funding from the European Union's Horizon 2020 Research and Innovation program under Grant Agreement No. **101017549**

1.2.4.4 Structural Variants INFO keys

Structural variants (SVs) are large genomic alterations, where large is typically (and somewhat arbitrarily) defined as encompassing at least 50 bp. These genomic variants are typically classified as deletions, duplications, insertions, inversions, and translocations describing different combinations of DNA gains, losses, or rearrangements **[6]**. In symbolic alternate alleles for imprecise structural variants, the ID field indicates the type of structural variant, and can be a colon-separated list of types and subtypes. ID values are case sensitive strings and must not contain whitespace or angle brackets.

For precise variants, END is POS + length of REF allele – 1 and for the imprecise variants the corresponding best estimate.

```
##INFO=<ID=SVTYPE,Number=1,Type=String,Description="Type of structural variant">
```

Figure 5. Example info field of structural variant

The first level type must be one of the following:

- DEL – Deletion relative to the reference
- INS – Insertion of novel sequence relative to the reference
- DUP – Region of elevated copy number relative to the reference
- INV – Inversion of reference sequence
- CNV – Copy number variable region (may be both deletion and duplication)
- BNV – Breakend

The CNV category should not be used when a more specific category can be applied. Reserved subtypes include:

- DUP:TANDEM – Tandem duplication
- DEL:ME – Deletion of mobile element relative to the reference
- INS:ME – Insertion of mobile element relative to the reference

1.2.4.5 Other VCF aspects to be considered

- The VCF format version should follow at least 4.1 format specification (vcf 4.1 spec).
- The header should include the reference genome and the **md5 checksum** used which can be added manually or through the variant caller pipelines, following GA4GH **refget** specs (GA4GH_ref_schema).
- Add information about which software was used for the variant calling and which command was run. This info can usually be extracted from the VICF header, metadata file or publications associated.

This project has received funding from the European Union's Horizon 2020 Research and Innovation program under Grant Agreement No. **101017549**

1.2.4.6 VCF Normalization

VCF normalization consists in representation of genetic variants in an unambiguous and concise way allowing comparison between them. This is a big problem when dealing with situations like duplicate removal, variant filtering and multiallelic sites. Currently, there is no standard way to normalize VCF and at the EGA, we use both bcftools (bcftools norm) and GA4GH normalization guidelines (GA4GH_norm_guide). If the normalization is performed on the VCFs, it is recommended to include this information in the file when sharing.

1.2.4.7 Recommendation for associated information in VCF files

- Good metadata makes data reusable. It is recommended to add non-identifiable metadata information in an additional file such as phenotype, age, gender, heterozygosity etc.
- Information about the type of VCF such as panel, exome, whole genome etc should be provided.
- Although info regarding the calling process may be present in the VCF header, it is recommended to submit more detailed information such as reference to the variant calling pipelines/workflow used to call the variants.
- Add information about the sequencing platform.

1.2.5 Further readings and recommendations

VCFv4.3.tex is the canonical specification for the Variant Call Format and its textual (VCF) and binary (BCF) encodings, while VCFv4.1.tex and VCFv4.2.tex describe their predecessors. VCFv4.4.draft.tex is a working draft of the upcoming version of VCF format and is under active revision. It is recommended to stay updated with most recent guidelines by VCF community.

1.2.5.1 File encryption

crypt4gh.tex is the canonical specification of the crypt4gh format which can be used to wrap existing file formats in an encryption layer.

1.2.5.2 Transfer protocols

- Htsget.md describes the **hts-get** retrieval protocol, which enables parallel streaming access to data shared across multiple URLs or files.
- Refget.md enables access to reference sequences using an identifier derived from the sequence itself.

This project has received funding from the European Union's Horizon 2020 Research and Innovation program under Grant Agreement No. **101017549**

2 Phenotypes, Metadata and Clinical Data exchange

Human genomics dataset is accompanied by clinical phenotypes and other metadata such as patient's medical history, sex, age, diagnosis etc that is utilised to study association between certain genomic feature and its effect. Such data are sensitive and requires adequate privacy measures before sharing.

Work Package 2 and 5 of GenoMed4All consortium is dedicated to provide strict privacy & anonymization protocols for the genomics and clinical data used in this consortium. However, the challenge does not end here, the systematic lack of standardised protocols for interoperable data and lack of FAIR data principles (Findable, accessible, interoperable, reusable) **[7]** in biological data community is one of the main bottlenecks.

In order to protect and disseminate such valuable information and to mitigate the risk of 'tower of babel', Global Alliance for Genomics and Health (GA4GH) has joined hands with other genomics big data repositories such as European Genome-phenome Archive (EGA) and the database of Genotypes and Phenotypes (dbGaP) to spread awareness on both storing and consented reutilization of data of genetic and phenotypic origin in FAIR way, utilizing specialised strategies to facilitate the sharing of clinical data.

An open standard produced under the guidance of GA4GH is the Phenopackets standard (http://phenopackets.org/) for sharing disease and phenotype information. A 'PhenoPacket' provide information models to successfully exchange clinical information between different levels of complexity ie. it enables high level clinical phenotype information to be exchanged with deep clinical phenotype information **[8]**. It links phenotype descriptions with disease, patient and genetic information, thus, enabling clinicians or researchers to build more complete models of diseases.

Possible use-cases for common data model:

- A consensus in metadata model is important during the development phase of the project.
- A researcher wants to develop new algorithm and searches for data matching the specific features that algorithm is requesting (clinical/phenotype, genomic, etc). Otherwise, a researcher expects a synthetic dataset as a 'seed' with the proper data model (or proper data interface API), against which they can develop and test the algorithm.
- A researcher wants to develop and upload their own algorithm into the platform so that it can enrich a catalogue for the community. In this case scenario, a common data model framework needs to be agreed beforehand.
- A researcher that has their algorithm (that is not GenoMed4All compliant) but would like to train this algorithm with the specifications given by GenoMed4All, against new data set to re-enforce its performances.

This project has received funding from the European Union's Horizon 2020 Research and Innovation program under Grant Agreement No. **101017549**

Above points raise the concern for a common data model pre-requisite for different model training and datasets discovery. Therefore, it is highly recommended to develop a data model "contract" prior to developing AI model. This contract could be used during the discovery phase ie. for the platform search for compatible edge SoR (source of records) for a given algorithm. This requires to define a reference data model associated to an algorithm, as well as during the model development phase, for the researcher to understand which features they must use, and which data model/interface it must by compliant with.

2.1 Recommended Model for Metadata Sharing

2.1.1.1 Phenotype Exchange Format (PXF) files

Phenopackets are represented as PXF (Phenotype Exchange Format) files (fig 7), which may be encoded in JSON or YAML. Each packet associates a list of phenotypic abnormalities with a disease and patient, including details about age, sex, onset, and evidence. Standard ontologies are used to ensure interoperability between diverse sources, to simplify text-mining, and to enable machine reasoning. Further, its open standard makes it easy to adapt to other languages, systems and applications.

Figure 6. An example of phenopacket describing the pathogenicity of a variant

This project has received funding from the European Union's Horizon 2020 Research and Innovation program under Grant Agreement No. **101017549**

Further reading suggested:
https://phenopackets-analysis.readthedocs.io/_/downloads/en/latest/pdf/

2.1.1.2 FHIR

FHIR (Fast Healthcare Interoperability Resources) Specification is another standard for exchanging healthcare information electronically **[9]**. FHIR is built on previous data format standards HL7 international and it supports json, xml, nd-json and rdf as data exchange format. It uses web-based suite of API technology, including a HTTP-based RESTful protocol. A FHIR resource can be an individual packet of information that include metadata, text, or particular data elements, but can also be bundled into collections that create clinical documents. Applications can be plugged into a basic EHR operating system and feed information directly into the provider workflow, avoiding pitfalls of document-based exchange, which often requires provider to access data separately.

Figure 7. An example of FHIR resource of a patient[1]

Currently, FHIR is more equipped for exchanging EHR based clinical data information but not mature at the genomics level when compared to Phenopackets. On the other hand, Phenopackets

[1] Source: https://wiki.galenhealthcare.com/index.php/HL7_FHIR

This project has received funding from the European Union's Horizon 2020 Research and Innovation program under Grant Agreement No. **101017549**

is equipped for genomics data use cases, but not well integrated with EHR system. However, both systems are evolving fast, in particular to genomics datasets and it is recommended to stay connected with the community during the course of this project. Current development phase for both Phenopackets and FHIR do not focus on specifically on rare diseases, however, both communities are open for collaborations and therefore, GenoMed4All can contribute to the community with specific use-cases. A mapping of Phenopackets objects to FHIR already exist (Geno/pheno FHIR working group) and we recommend the usage of both FHIR and Phenopackets data standards as it suits the project, along with a strong contribution in the future towards the Phenopackets community.

This project has received funding from the European Union's Horizon 2020 Research and Innovation program under Grant Agreement No. **101017549**

3 Recommendations for Genomics and Metadata sharing

3.1 Beacon

One of the main bottlenecks in human genomics research is the lack of tools for federated discovery of identifiable genomics data that requires tight privacy controls, while making such data available to the community. Global Alliance for Genomics and Health (GA4GH) initiated the Beacon project **[10]** for the federated discovery of genomic data in biomedical research and clinical applications with secure guidelines.
A Beacon is a simple genomics variant discovery tool by aggregating worldwide genomics dataset under one umbrella. In the time of personalised medicines, inclusive diagnostics, prognostic and therapeutic strategies, the Beacon project aims at solving the problem of genomics data sharing through enabling the search of genomic variants and associated information without jeopardising the privacy of the dataset. This way, any hospital or research entity can choose to 'beaconize' their dataset without compromising the privacy or the ownership of the dataset. Further, Beacon provide various access levels and controls over one's dataset, for example, same hospital can choose different data access levels based on whether a requester works inside the hospital or as an outsider.

Beacon is agnostic to FHIR or Phenopackets and therefore, can query and export both formats through search API. Soon to be released Beacon Version2 goes beyond the core variant identification search to add more layer of information such as biosamples, phenotypes and additional metadata discovery, including other clinical information. Beacon version 2 also supports both studies defined and user-defined cohorts queries.

Further features of Beacon v2 that can be utilised for GenoMed4All's purpose:

- More informative queries, like filtering by gender or age.
- Simplified data permission and access process, e.g. who to contact or which are the data use conditions.
- An option to jump to another system where the data could be accessed, e.g. if the Beacon is for internal use of the hospital, to provide the Id of the EHR of the patients having the mutation of interest.
- Annotations about the variants found, among which the expert/clinician conclusion about the pathogenicity of a given mutation in a given individual or its role in producing a given phenotype.

This project has received funding from the European Union's Horizon 2020 Research and Innovation program under Grant Agreement No. **101017549**

4 References

[1] Knoppers, B.M. (2014). Framework for responsible sharing of genomic and health-related data. HUGO J 8, 3
https://doi.org/10.1186/s11568-014-0003-1

[2] Saunders, G., Baudis, M., Becker, R. et al. (2019). Leveraging European infrastructures to access 1 million human genomes by 2022. Nat Rev Genet 20, 693–701 https://doi.org/10.1038/s41576-019-0156-9

[3] Payne, A., Holmes, N., Rakyan, V. et al., (2018). Whale watching with BulkVis: A graphical viewer for Oxford Nanopore bulk fast5 files bioRxiv 312256
doi: https://doi.org/10.1101/312256

[4] Roy, S., Next-Generation Sequencing Bioinformatics Pipelines (https://www.aacc.org/cln/articles/2020/march/next-generation-sequencing-bioinformatics-pipelines)

[5] Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics (Oxford, England), 25(16), 2078–2079.
https://doi.org/10.1186/s13059-019-1828-7

[6] Mahmoud, M., Gobet, N., Cruz-Dávalos, D.I. et al. (2019). Structural variant calling: the long and the short of it. Genome Biol 20, 246.
https://doi.org/10.1186/s13059-019-1828-7

[7] Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., et al (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific data, 3, 160018. https://doi.org/10.1038/sdata.2016.18

[8] Dolman, L., Page, A., Babb, L., Freimuth, R. R., Arachchi, H., et al (2018). ClinGen advancing genomic data-sharing standards as a GA4GH driver project. Human mutation, 39(11), 1686–1689. https://doi.org/10.1002/humu.23625

[9] Lehne, M., Luijten, S., Vom Felde Genannt Imbusch, P., Thun, S. (2019). The Use of FHIR in Digital Health - A Review of the Scientific Literature. Stud Health Technol Inform.267, 52-58.
https://doi:10.3233/SHTI190805

[10] Fiume, M., Cupak, M., Keenan, S. et al. (2019). Federated discovery and sharing of genomic data using Beacons. Nat Biotechnol 37, 220–224.
https://doi.org/10.1038/s41587-019-0046-x

This project has received funding from the European Union's Horizon 2020 Research and Innovation program under Grant Agreement No. **101017549**

GENOMED4ALL

genomed4all.eu

@genomed4all

/genomed4all

GENOMED4ALL receives funding from the European Union's Horizon 2020 Research and Innovation programme under Grant Agreement No. 101017549