unibz

Research 21: Specialized and Domain-Specific Data Management

Relational Data Models for Genetic VCF data

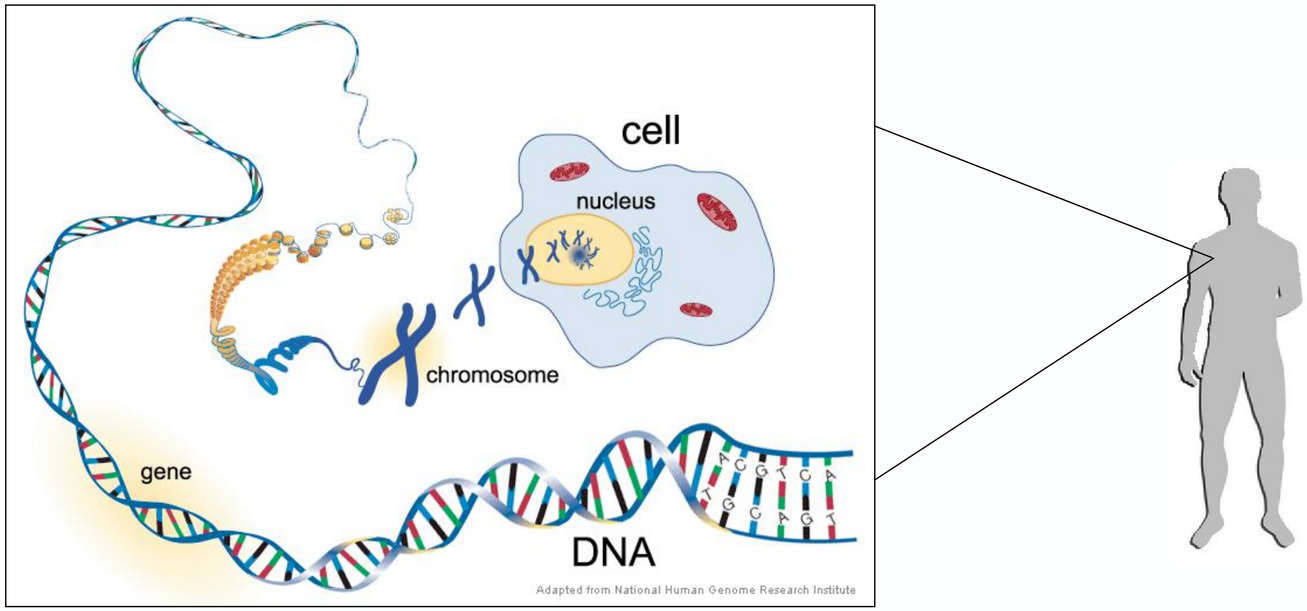
Mohamed Sabri Hafidi Ozan Kahramanoğulları Anton Dignös Johann Gamper

Free University of Bozen-Bolzano, Italy

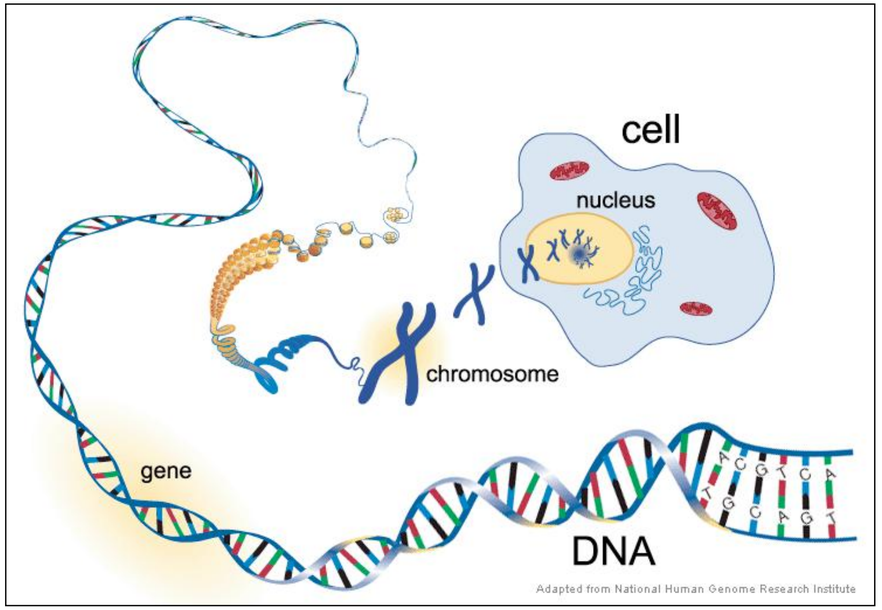
VLDB 2025

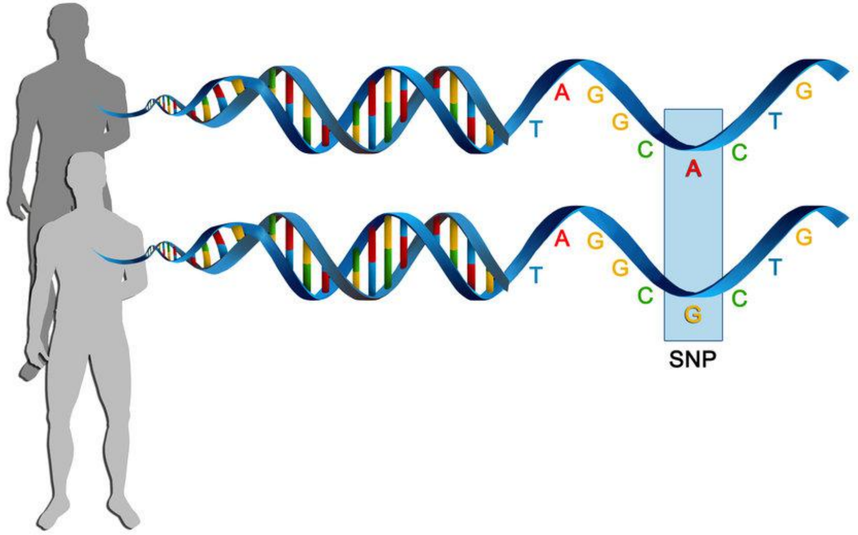
This work was funded in part by the Autonomous Province of Bozen-Bolzano, call Joint Projects South Tyrol – Germany, project DyHealthNet – CUP: I53C22002980003.

Understanding Genetic Data



Understanding Genetic Data





The Variant Call Format - VCF

Standardized text file for genomic sequence variations

Records differences from a reference genome

Includes metadata, variant annotations, and sample genotypes

Flexible and extensible: Dedicated fields allow rich annotations

VCF File

Meta-Information

```
##fileformat=VCFv4.4
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,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
```

VCF File

```
##fileformat=VCFv4.4
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,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">
```

Header

#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

VCF File

```
##fileformat=VCFv4.4
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,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
```

Fixed Fields

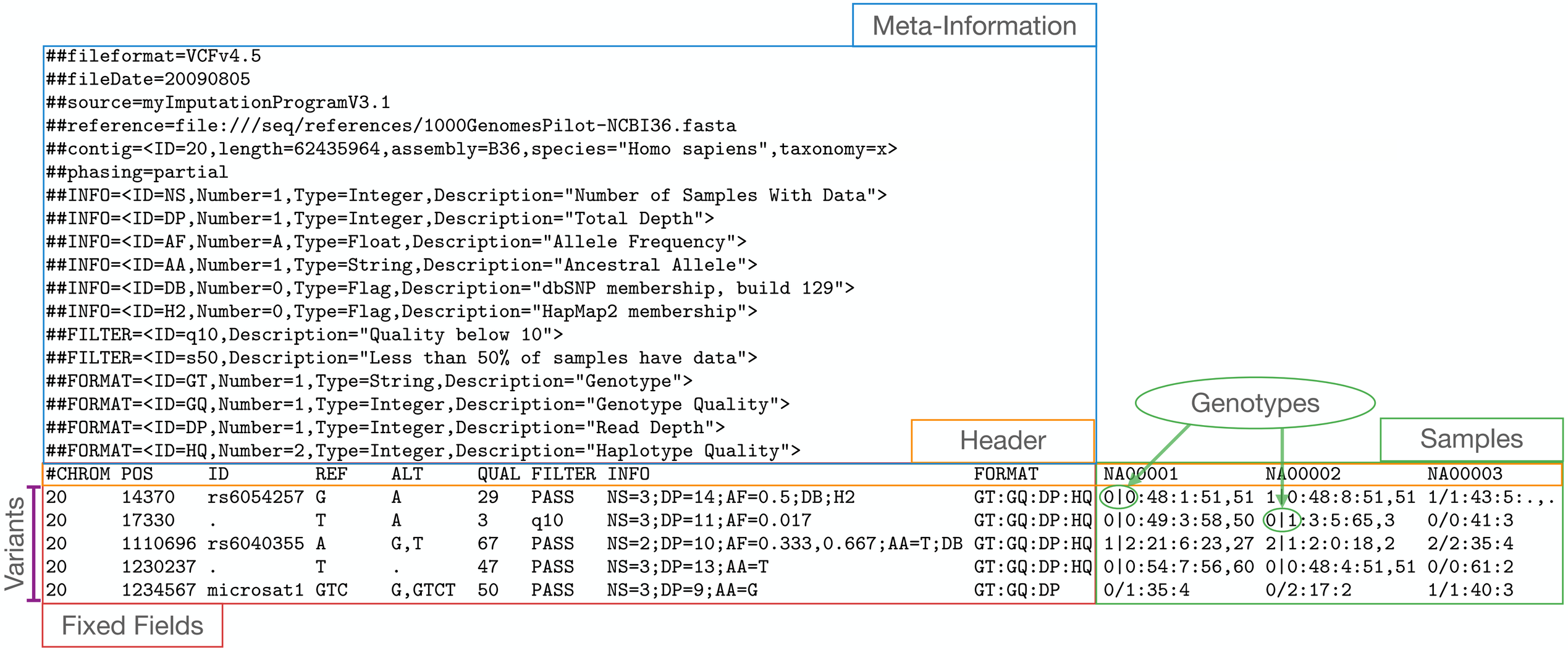
VCF File

```
##fileformat=VCFv4.4
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,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">
```

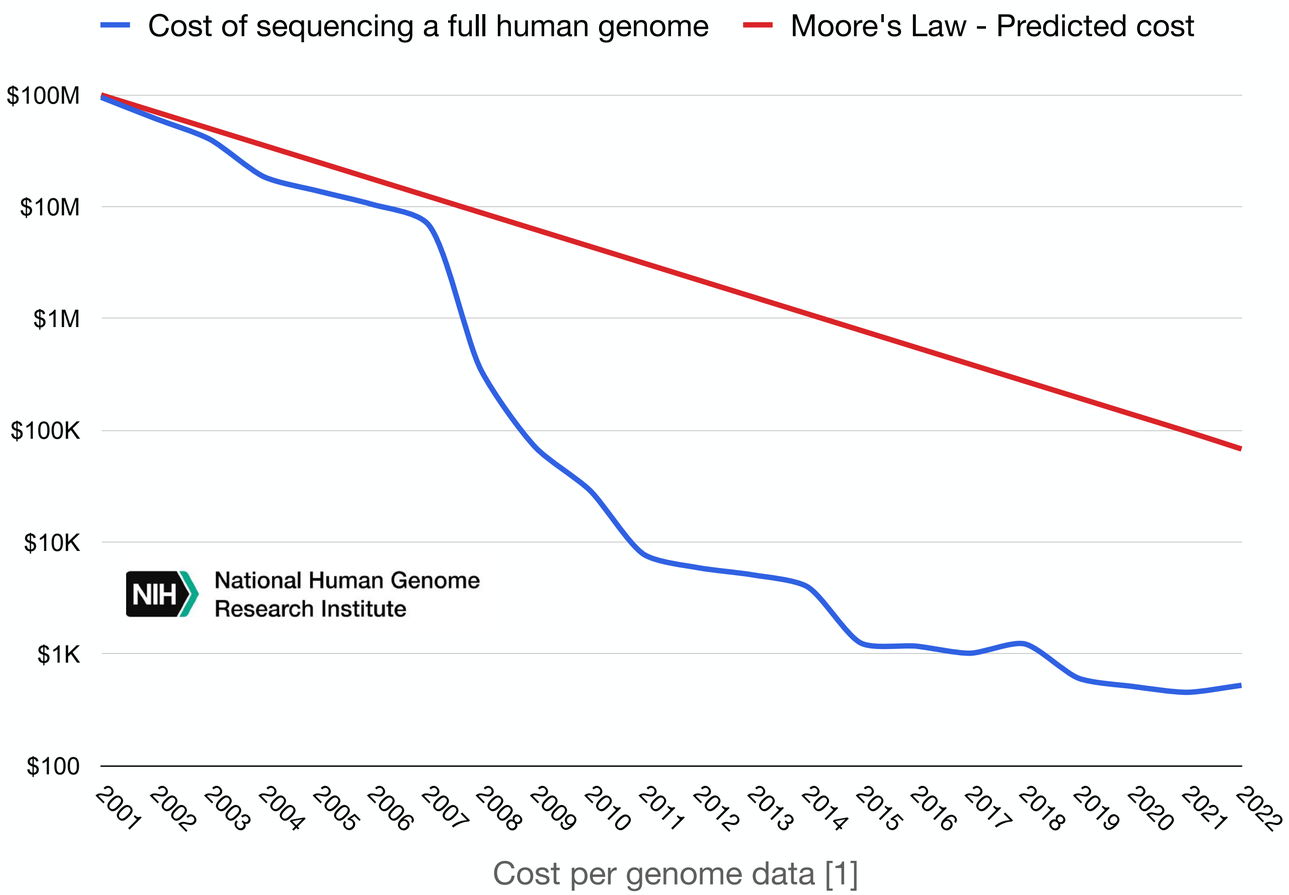
Samples

#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

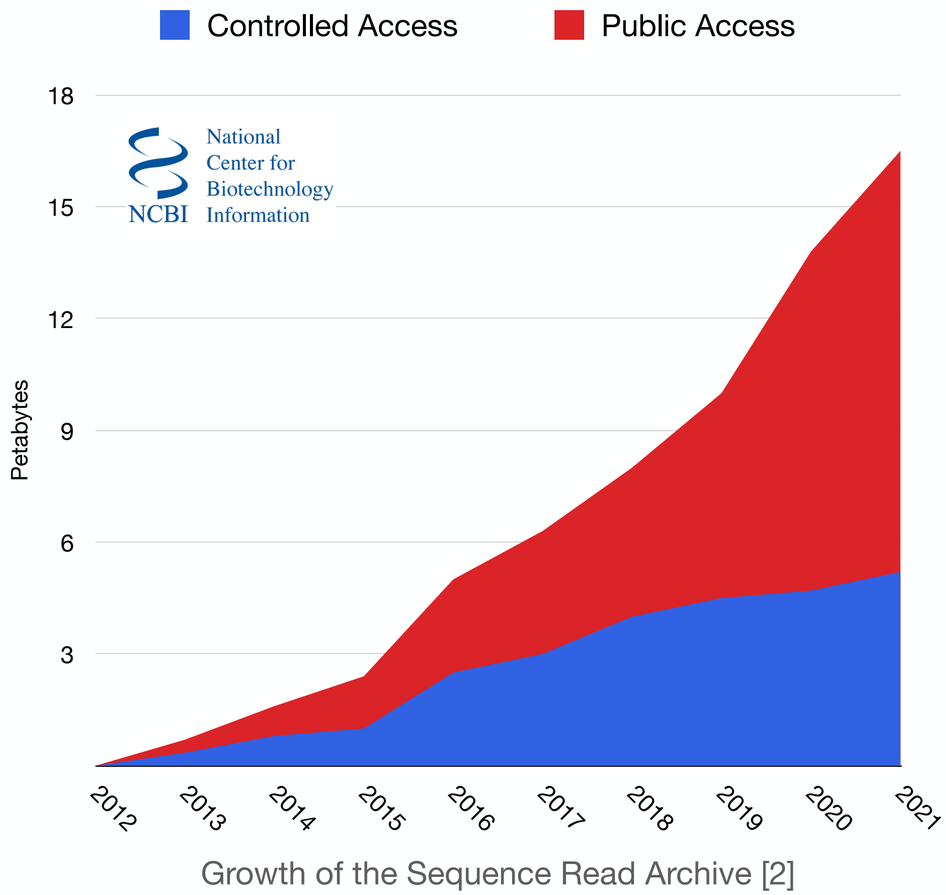
VCF File


Meta-Information
##fileformat=VCFv4.5
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,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">
Header
Genotypes
Samples
#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
Variants
Fixed Fields

Genomic Data Explosion



Cost per genome data [1]



Growth of the Sequence Read Archive [2]

Genomic Data Bottlenecks

VCF files’ text structure and parsing impede efficient database querying

Growing genomic datasets outpace legacy pipelines

Limited interoperability hinders multi-modal analysis

Complex analyses require stitching together disparate tools

Related Work

BCFtools & Tabix: industry-standard command-line tools – *Danecek et al. 2021*

TileDB-VCF: specialized database system for genetic data – *Papadopoulos et al. 2016*

Genotypic Data in Relational Databases: different representations – *Lichtenwalter et al. 2017*

Integrating Variants and Ontologies in a Document Database – *Liu et al. 2019*

Our Approach

Core Idea: Transform VCF into structured relational tables

Objective:

- Scalable storage for petabytes of data
- Flexible SQL-based querying
- Seamless multi-omics (beyond VCF) integration

Challenge: Mapping semi-structured VCF to relational schemas

Main Contributions

Novel data models for storing VCF data in RDBMS

Comprehensive performance benchmarking: competitive performance

Open-source implementation

Design Highlights

Wide: each row represents a variant and each column represents a sample

Narrow: each row represents a single variant-sample pair

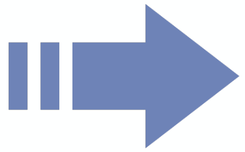
Array Plain: store the genotype data in an array

Full JSON: store the sample ID and its genotype in a JSON

Wide & Wide JSON

Samples

NA00001	NA00002	NA00003
0|0:48:1:51,51	1|0:48:8:51,51	1/1:43:5:.,.
0|0:49:3:58,50	0|1:3:5:65,3	0/0:41:3
1|2:21:6:23,27	2|1:2:0:18,2	2/2:35:4
0|0:54:7:56,60	0|0:48:4:51,51	0/0:61:2
0/1:35:4	0/2:17:2	1/1:40:3



vcf_wide_samples_chunk_1

ln	na00001	na00002	na00003	...
1	0|0:48:1:51,51	1|0:48:8:51,51	1/1:43:5:.	
2	0|0:49:3:58,50	0|1:3:5:65,3	0/0:41:3	
3	1|2:21:6:23,27	2|1:2:0:18,2	2/2:35:4	...
4	0|0:54:7:56,60	0|0:48:4:51,51	0/0:61:2	
5	0/1:35:4	0/2:17:2	1/1:40:3	

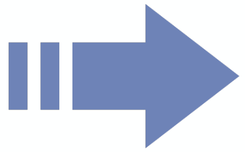
vcf_wide_json_samples_chunk_1

ln	na00001	...
1	{"DP": 1, "GQ": 48, "GT": "0/0", "HQ": "51,51"}	
2	{"DP": 3, "GQ": 49, "GT": "0/0", "HQ": "58,50"}	
3	{"DP": 6, "GQ": 21, "GT": "1/2", "HQ": "23,27"}	...
4	{"DP": 7, "GQ": 54, "GT": "0/0", "HQ": "56,60"}	
5	{"DP": 4, "GQ": 35, "GT": "0/1"}	

Narrow & Narrow JSON

Samples

NA00001	NA00002	NA00003
0|0:48:1:51,51	1|0:48:8:51,51	1/1:43:5:.,.
0|0:49:3:58,50	0|1:3:5:65,3	0/0:41:3
1|2:21:6:23,27	2|1:2:0:18,2	2/2:35:4
0|0:54:7:56,60	0|0:48:4:51,51	0/0:61:2
0/1:35:4	0/2:17:2	1/1:40:3



vcf_narrow_samples

ln	s_id	genotypes
1	NA00001	0|0:48:1:51,51
1	NA00002	1|0:48:8:51,51
1	NA00003	1/1:43:5:.
2	NA00001	0|0:49:3:58,50
2	NA00002	0|1:3:5:65,3
2	NA00003	0/0:41:3
3	NA00001	1|2:21:6:23,27
3	NA00002	2|1:2:0:18,2
3	NA00003	2/2:35:4
4	NA00001	0|0:54:7:56,60
4	NA00002	0|0:48:4:51,51
4	NA00003	0/0:61:2
5	NA00001	0/1:35:4
5	NA00002	0/2:17:2
5	NA00003	1/1:40:3
...	...	...

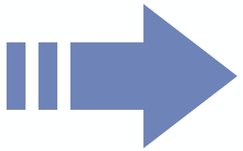
vcf_narrow_json_samples

ln	s_id	genotypes
1	NA00001	{"DP": 1, "GQ": 48, "GT": 0/0, "HQ": 51,51}
1	NA00002	{"DP": 8, "GQ": 48, "GT": 1/0, "HQ": 51,51}
1	NA00003	{"DP": 5, "GQ": 43, "GT": 1/1}
2	NA00001	{"DP": 3, "GQ": 49, "GT": 0/0, "HQ": 58,50}
2	NA00002	{"DP": 5, "GQ": 3, "GT": 0/1, "HQ": 65,3}
2	NA00003	{"DP": 3, "GQ": 41, "GT": 0/0}
3	NA00001	{"DP": 6, "GQ": 21, "GT": 1/2, "HQ": 23,27}
3	NA00002	{"DP": 0, "GQ": 2, "GT": 2/1, "HQ": 18,2}
3	NA00003	{"DP": 4, "GQ": 35, "GT": 2/2}
4	NA00001	{"DP": 7, "GQ": 54, "GT": 0/0, "HQ": 56,60}
4	NA00002	{"DP": 4, "GQ": 48, "GT": 0/0, "HQ": 51,51}
4	NA00003	{"DP": 2, "GQ": 61, "GT": 0/0}
5	NA00001	{"DP": 4, "GQ": 35, "GT": 0/1}
5	NA00002	{"DP": 2, "GQ": 17, "GT": 0/2}
5	NA00003	{"DP": 3, "GQ": 40, "GT": 1/1}
...	...	...

Array Plain & Array JSON

Samples

NA00001	NA00002	NA00003
0|0:48:1:51,51	1|0:48:8:51,51	1/1:43:5:.,.
0|0:49:3:58,50	0|1:3:5:65,3	0/0:41:3
1|2:21:6:23,27	2|1:2:0:18,2	2/2:35:4
0|0:54:7:56,60	0|0:48:4:51,51	0/0:61:2
0/1:35:4	0/2:17:2	1/1:40:3



vcf_array_indices

s_id	e_id
NA00001	1
NA00002	2
NA00003	3
...	...

vcf_array_genotypes

ln	genotypes
1	["0|0:48:1:51,51", "1|0:48:8:51,51", "1/1:43:5:.", ...]
2	["0|0:49:3:58,50", "0|1:3:5:65,3", "0/0:41:3", ...]
3	["1|2:21:6:23,27", "2|1:2:0:18,2", "2/2:35:4", ...]
4	["0|0:54:7:56,60", "0|0:48:4:51,51", "0/0:61:2", ...]
5	["0/1:35:4", "0/2:17:2", "1/1:40:3", ...]

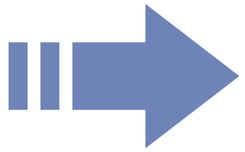
vcf_array_json_genotypes

ln	genotypes
1	[{"DP": 1, "GQ": 48, "GT": "0/0", "HQ": "51,51"}, {"DP": 8, "GQ": 48, "GT": "1/0", "HQ": "51,51"}, ...]
2	[{"DP": 3, "GQ": 49, "GT": "0/0", "HQ": "58,50"}, {"DP": 5, "GQ": 3, "GT": "0/1", "HQ": "65,3"}, ...]
3	[{"DP": 6, "GQ": 21, "GT": "1/2", "HQ": "23,27"}, {"DP": 0, "GQ": 2, "GT": "2/1", "HQ": "18,2"}, ...]
4	[{"DP": 7, "GQ": 54, "GT": "0/0", "HQ": "56,60"}, {"DP": 4, "GQ": 48, "GT": "0/0", "HQ": "51,51"}, ...]
5	[{"DP": 4, "GQ": 35, "GT": "0/1"}, {"DP": 2, "GQ": 17, "GT": "0/2"}, ...]

Full JSON

Samples

NA00001	NA00002	NA00003
0|0:48:1:51,51	1|0:48:8:51,51	1/1:43:5:.,.
0|0:49:3:58,50	0|1:3:5:65,3	0/0:41:3
1|2:21:6:23,27	2|1:2:0:18,2	2/2:35:4
0|0:54:7:56,60	0|0:48:4:51,51	0/0:61:2
0/1:35:4	0/2:17:2	1/1:40:3



vcf_json_samples

ln	sample_genotype
1	{{"s_id": "NA00001", "genotype": {"DP": "1", "GT": "0|0", "GQ": "48", "HQ": "51,51"}}, ...}
2	{{"s_id": "NA00001", "genotype": {"DP": "3", "GT": "0|0", "GQ": "49", "HQ": "58,50"}}, ...}
3	{{"s_id": "NA00001", "genotype": {"DP": "6", "GT": "1|2", "GQ": "21", "HQ": "23,27"}}, ...}
4	{{"s_id": "NA00001", "genotype": {"DP": "7", "GT": "0|0", "GQ": "54", "HQ": "56,60"}}, ...}
5	{{"s_id": "NA00001", "genotype": {"DP": "4", "GT": "0/1", "GQ": "35"}}, ...}

Performance Evaluation

Data Access: Retrieves a range of variants for a given set of sample IDs

Sample Filtering: Data access + Filter on the sample’s genotype

Variant Filtering: Data access + Filter on the variant’s fixed fields

Comparison baseline

BCFtools + Tabix:

- Standard Command-Line Toolkit
- File-Based Operations
- **Tabix** for genomic regions indexing
- Widely adopted approach

TileDB-VCF:

- Specialized Array Database
- Optimized for Genomic Data
- API-Driven Integration
- Modern DBMS Baseline

Compared to the overall best **Array Plain**

Performance Evaluation

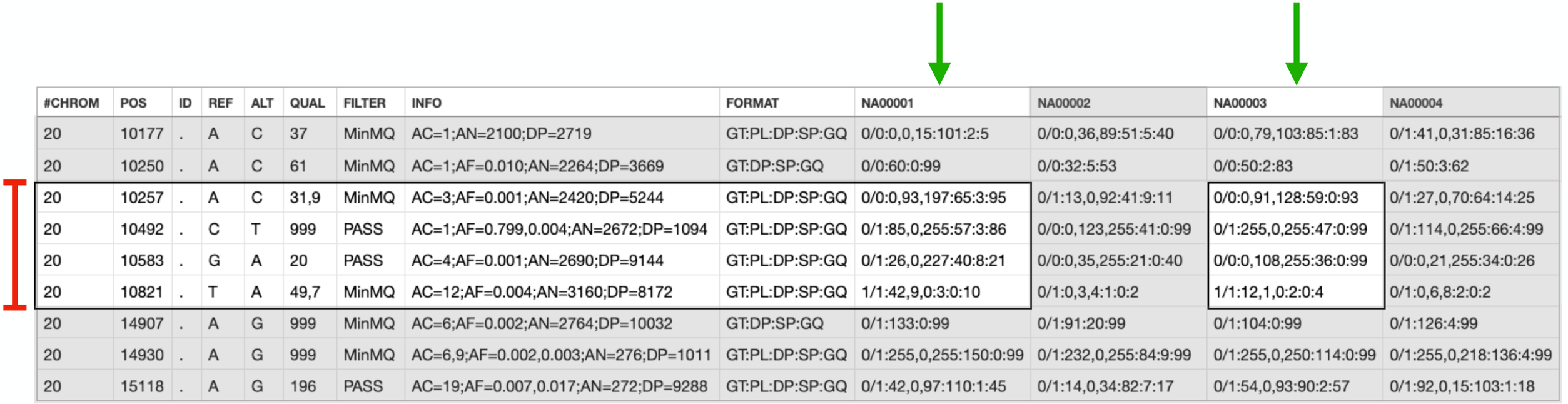
Data Access: Retrieves a range of variants for a given set of sample IDs

Sample Filtering: Data access + Filter on the sample’s genotype

Variant Filtering: Data access + Filter on the variant’s fixed fields

Data Access

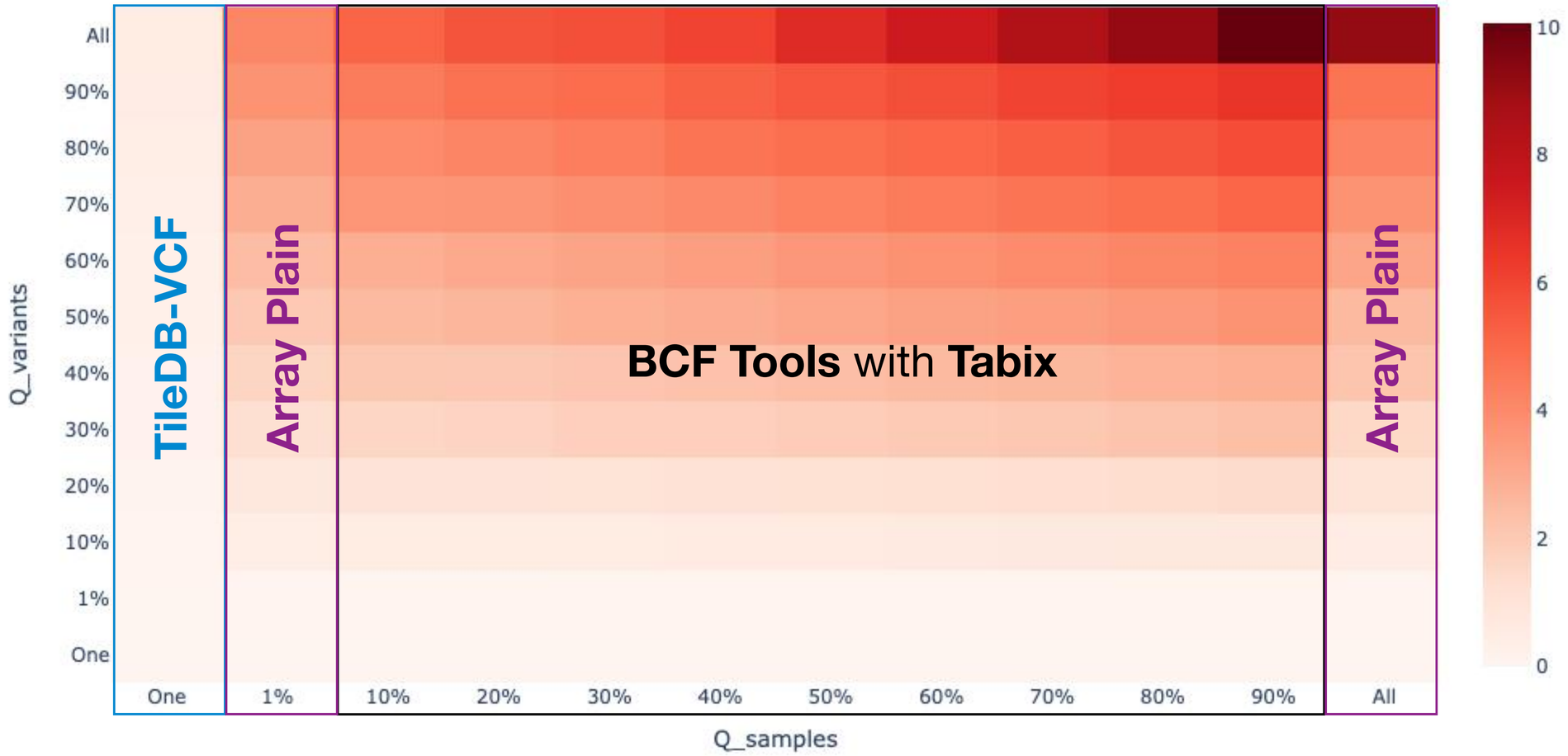
Retrieves a range of variants for a given set of sample IDs



#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	NA00001	NA00002	NA00003	NA00004
20	10177	.	A	C	37	MinMQ	AC=1;AN=2100;DP=2719	GT:PL:DP:SP:GQ	0/0:0,0,15:101:2:5	0/0:0,36,89:51:5:40	0/0:0,79,103:85:1:83	0/1:41,0,31:85:16:36
20	10250	.	A	C	61	MinMQ	AC=1;AF=0.010;AN=2264;DP=3669	GT:DP:SP:GQ	0/0:60:0:99	0/0:32:5:53	0/0:50:2:83	0/1:50:3:62
20	10257	.	A	C	31,9	MinMQ	AC=3;AF=0.001;AN=2420;DP=5244	GT:PL:DP:SP:GQ	0/0:0,93,197:65:3:95	0/1:13,0,92:41:9:11	0/0:0,91,128:59:0:93	0/1:27,0,70:64:14:25
20	10492	.	C	T	999	PASS	AC=1;AF=0.799,0.004;AN=2672;DP=1094	GT:PL:DP:SP:GQ	0/1:85,0,255:57:3:86	0/0:0,123,255:41:0:99	0/1:255,0,255:47:0:99	0/1:114,0,255:66:4:99
20	10583	.	G	A	20	PASS	AC=4;AF=0.001;AN=2690;DP=9144	GT:PL:DP:SP:GQ	0/1:26,0,227:40:8:21	0/0:0,35,255:21:0:40	0/0:0,108,255:36:0:99	0/0:0,21,255:34:0:26
20	10821	.	T	A	49,7	MinMQ	AC=12;AF=0.004;AN=3160;DP=8172	GT:PL:DP:SP:GQ	1/1:42,9,0:3:0:10	0/1:0,3,4:1:0:2	1/1:12,1,0:2:0:4	0/1:0,6,8:2:0:2
20	14907	.	A	G	999	MinMQ	AC=6;AF=0.002;AN=2764;DP=10032	GT:DP:SP:GQ	0/1:133:0:99	0/1:91:20:99	0/1:104:0:99	0/1:126:4:99
20	14930	.	A	G	999	MinMQ	AC=6,9;AF=0.002,0.003;AN=276;DP=1011	GT:PL:DP:SP:GQ	0/1:255,0,255:150:0:99	0/1:232,0,255:84:9:99	0/1:255,0,250:114:0:99	0/1:255,0,218:136:4:99
20	15118	.	A	G	196	PASS	AC=19;AF=0.007,0.017;AN=272;DP=9288	GT:PL:DP:SP:GQ	0/1:42,0,97:110:1:45	0/1:14,0,34:82:7:17	0/1:54,0,93:90:2:57	0/1:92,0,15:103:1:18

Performance Evaluation

Data Access: Heatmap of runtime (min)



Dataset ~13k samples and ~95k variants

Performance Evaluation

Data Access: Retrieves a range of variants for a given set of sample IDs

Sample Filtering: Data access + Filter on the sample’s genotype

Variant Filtering: Data access + Filter on the variant’s fixed fields

Sample Filtering

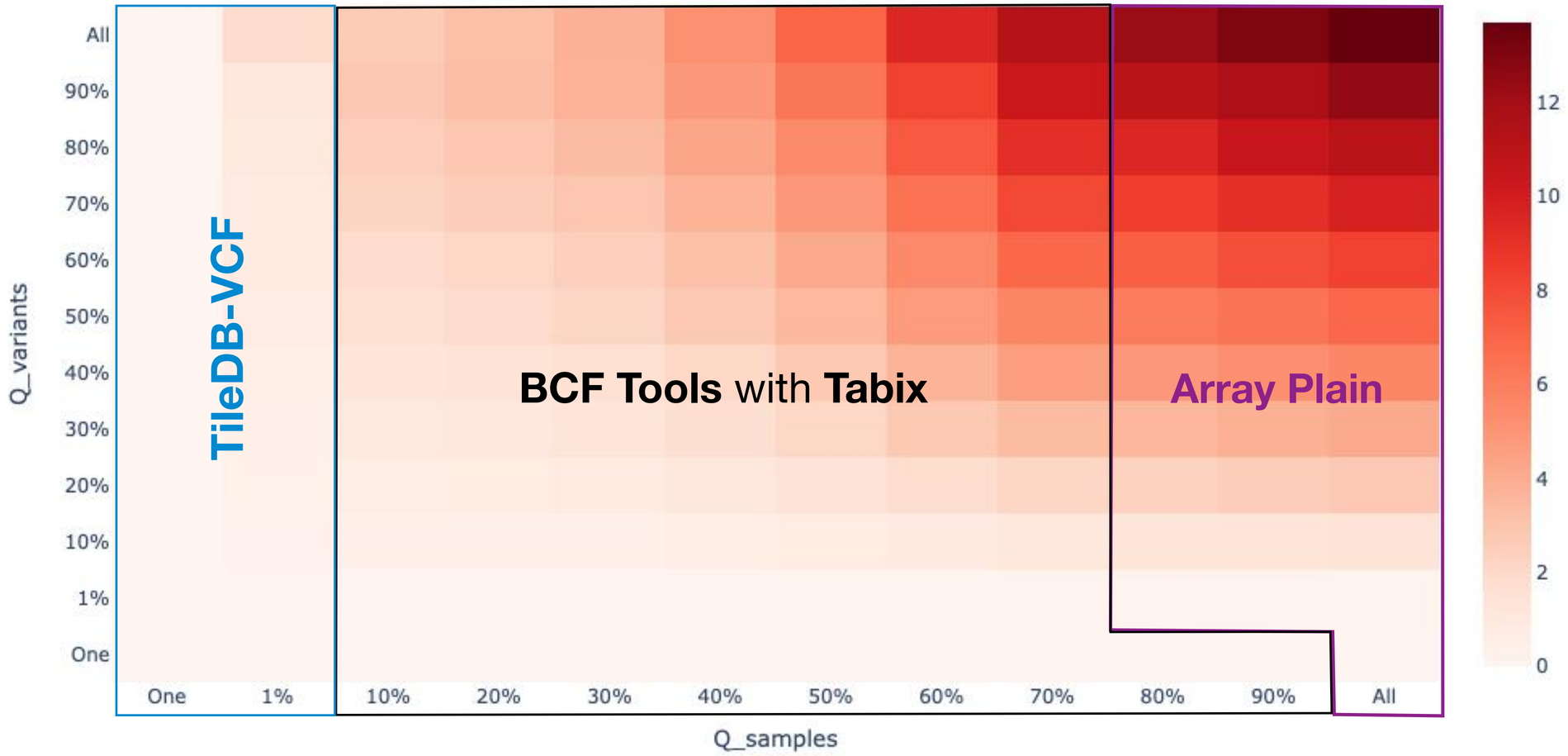
Select **Samples** that are **homozygous** at **position 10257**

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	NA00001	NA00002	NA00003	NA00004
20	10177	.	A	C	37	MinMQ	AC=1;AN=2100;DP=2719	GT:PL:DP:SP:GQ	0/0:0,0,15:101:2:5	0/0:0,36,89:51:5:40	0/0:0,79,103:85:1:83	0/1:41,0,31:85:16:36
20	10250	.	A	C	61	MinMQ	AC=1;AF=0.010;AN=2264;DP=3669	GT:DP:SP:GQ	0/0:60:0:99	0/0:32:5:53	0/0:50:2:83	0/1:50:3:62
20	10257	.	A	C	31,9	MinMQ	AC=3;AF=0.001;AN=2420;DP=5244	GT:PL:DP:SP:GQ	0/0:0,93,197:65:3:95	0/1:13,0,92:41:9:11	0/0:0,91,128:59:0:93	0/1:27,0,70:64:14:25
20	10492	.	C	T	999	PASS	AC=1;AF=0.799,0.004;AN=2672;DP=1094	GT:PL:DP:SP:GQ	0/1:85,0,255:57:3:86	0/0:0,123,255:41:0:99	0/1:255,0,255:47:0:99	0/1:114,0,255:66:4:99
20	10583	.	G	A	20	PASS	AC=4;AF=0.001;AN=2690;DP=9144	GT:PL:DP:SP:GQ	0/1:26,0,227:40:8:21	0/0:0,35,255:21:0:40	0/0:0,108,255:36:0:99	0/0:0,21,255:34:0:26
20	10821	.	T	A	49,7	MinMQ	AC=12;AF=0.004;AN=3160;DP=8172	GT:PL:DP:SP:GQ	1/1:42,9,0:3:		1/1:12,1,0:2:0:4	0/1:0,6,8:2:0:2
20	14907	.	A	G	999	MinMQ	AC=6;AF=0.002;AN=2764;DP=10032	GT:DP:SP:GQ	0/1:133:0:99	0/1:91:20:99	0/1:104:0:99	0/1:126:4:99
20	14930	.	A	G	999	MinMQ	AC=6,9;AF=0.002,0.003;AN=276;DP=1011	GT:PL:DP:SP:GQ	0/1:255,0,255:150:0:99	0/1:232,0,255:84:9:99	0/1:255,0,250:114:0:99	0/1:255,0,218:136:4:99
20	15118	.	A	G	196	PASS	AC=19;AF=0.007,0.017;AN=272;DP=9288	GT:PL:DP:SP:GQ	0/1:42,0,97:110:1:45	0/1:14,0,34:82:7:17	0/1:54,0,93:90:2:57	0/1:92,0,15:103:1:18

Homozygous

Query performance

Sample Filtering: Heatmap of runtime (min)



Dataset ~13k samples and ~95k variants

Performance Evaluation

Data Access: Retrieves a range of variants for a given set of sample IDs

Sample Filtering: Data access + Filter on the sample’s genotype

Variant Filtering: Data access + Filter on the variant’s fixed fields

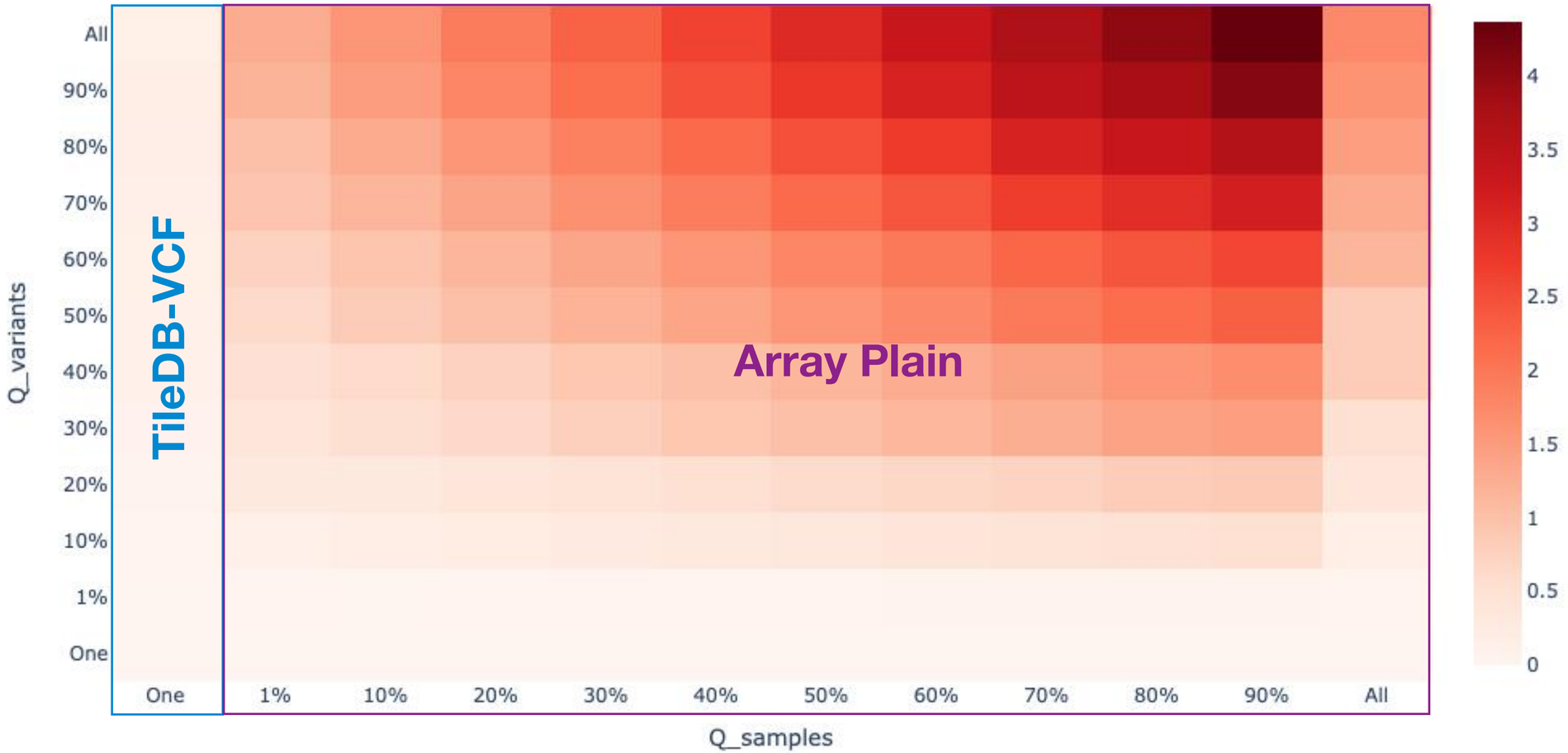
Variant Filtering

Select Variants where **AF ≥ 0,01** across all samples

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	NA00001	NA00002	NA00003	NA00004
20	10177	.	A	C	37	MinMQ	AC=1;AN=2100;DP=2719	GT:PL:DP:SP:GQ	0/0:0,0,15:101:2:5	0/0:0,36,89:51:5:40	0/0:0,79,103:85:1:83	0/1:41,0,31:85:16:36
20	10250	.	A	C	61	MinMQ	AC=1;AF=0.010;AN=2264;DP=3669	GT:DP:SP:GQ	0/0:60:0:99	0/0:32:5:53	0/0:50:2:83	0/1:50:3:62
20	10257	.	A	C	31,9	MinMQ	AC=3;AF=0.001;AN=2420;DP=5244	GT:PL:DP:SP:GQ	0/0:0,93,197:65:3:95	0/1:13,0,92:41:9:11	0/0:0,91,128:59:0:93	0/1:27,0,70:64:14:25
20	10492	.	C	T	999	PASS	AC=1;AF=0.799,0.004;AN=2672;DP=1094	GT:PL:DP:SP:GQ	0/1:85,0,255:57:3:86	0/0:0,123,255:41:0:99	0/1:255,0,255:47:0:99	0/1:114,0,255:66:4:99
20	10583					PASS	AC=4;AF=0.001;AN=2690;DP=9144	GT:PL:DP:SP:GQ	0/1:26,0,227:40:8:21	0/0:0,35,255:21:0:40	0/0:0,108,255:36:0:99	0/0:0,21,255:34:0:26
20	108					MQ	AC=12;AF=0.004;AN=3160;DP=8172	GT:PL:DP:SP:GQ	1/1:42,9,0:3:0:10	0/1:0,3,4:1:0:2	1/1:12,1,0:2:0:4	0/1:0,6,8:2:0:2
20	14907	.	A	G	999	MinMQ	AC=6;AF=0.002;AN=2764;DP=10032	GT:DP:SP:GQ	0/1:133:0:99	0/1:91:20:99	0/1:104:0:99	0/1:126:4:99
20	14930	.	A	G	999	MinMQ	AC=6,9;AF=0.002,0.003;AN=276;DP=1011	GT:PL:DP:SP:GQ	0/1:255,0,255:150:0:99	0/1:232,0,255:84:9:99	0/1:255,0,250:114:0:99	0/1:255,0,218:136:4:99
20	15118	.	A	G	196	PASS	AC=19;AF=0.007,0.017;AN=272;DP=9288	GT:PL:DP:SP:GQ	0/1:42,0,97:110:1:45	0/1:14,0,34:82:7:17	0/1:54,0,93:90:2:57	0/1:92,0,15:103:1:18

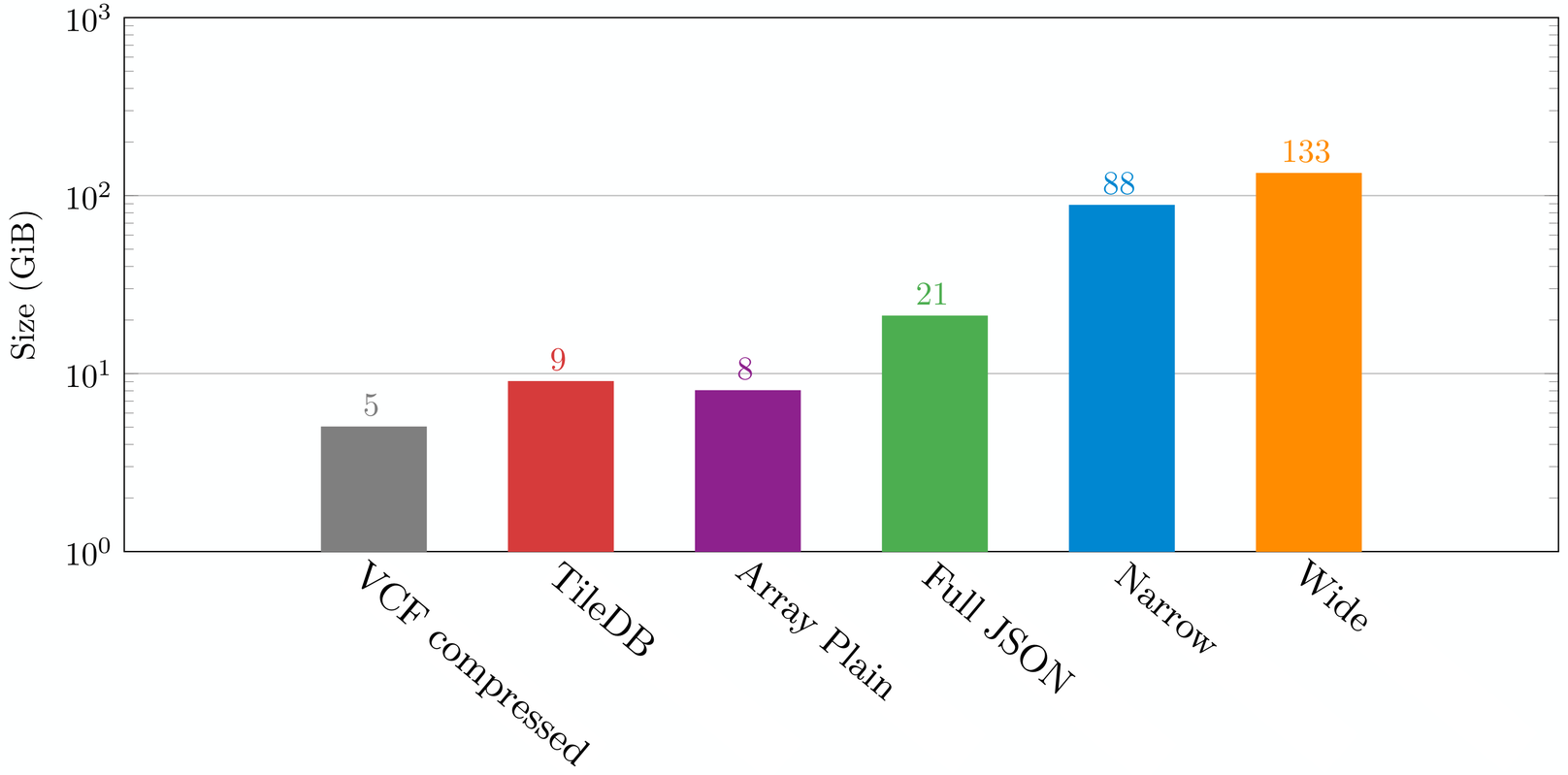
Performance Evaluation

Variant Filtering: Heatmap of runtime (min)



Dataset ~13k samples and ~95k variants

Storage requirements



Insights & Trade-offs

Competitive performance: RDBMS models rival specialized tools

No one-size-fits-all solution: Optimal model depends on query workload

Relational models offer **superior scalability** for large VCF datasets

Trade-offs are **workload-dependent**:

- Storage needs
- Query selectivity
- Dataset size

Array Plain model standout: Best overall runtime and storage

Future work

Model optimization: Hybrid data models

Advanced indexing strategies

Integration with multi-omics pipelines

Scalability & cloud deployment design

unibz

Thank you!

Mohamed Sabri Hafidi

hmohamedsabri@unibz.it

Source code: