BIO392
Bioinformatics of Genome Variations

File Formats | Strage Sizes | Genome Editions

Michael Baudis **UZH SIB**
Computational Oncogenomics

Genomic File Formats

Types | Sizes | Use Cases

What is a PB, for human genomes?

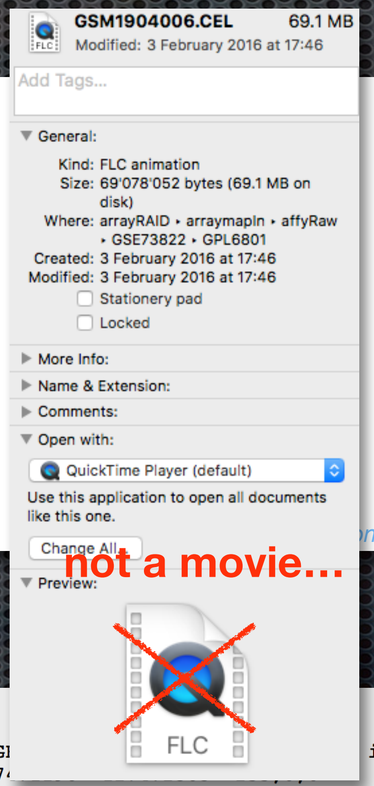
It depends...

- 2 bits per base are sufficient to encode TCGA
  - using 00, 01, 10, 11
  - [TCGA]{3’000’000’000}
  - 2 * 3 * 10^9b = 6, 000, 000, 000 b
  - perfect genome (no overhead): ~715 MB
  - 1PB => ~1’400’000 genomes
- according to Swiss online store (Sep 2021) ~35'000CHF (65x16TB á CHF550)
- this is less than a PhD position per year in Switzerland …
- (real costs are 2x that, + duplication, facilities, service … => ~500’000CHF)
- **However: A single 30x BAM file => 100GB**
- Still: 500’000CHF => 1PB => 10’000 genomes => 50CHF/genome (BAM format)

Bioinformatics: **File Formats**

- **text** or **binary** file formats, optimised for specific types of biological data
- examples from genomics:
  - **BAM** - compressed binary version of Sequence Alignment/ Map (SAM)
  - **BED** (Browser Extensible Data) -flexible way to define the data lines in an genome browser annotation tracks
  - **VCF** (Variant Call Format)

- Axt format
- BAM format
- BED format
- BED detail format
- bedGraph format
- barChart and bigBarChart format
- bigBed format
- bigGenePred table format
- bigPsl table format
- bigMaf table format
- bigChain table format
- bigWig format
- Chain format



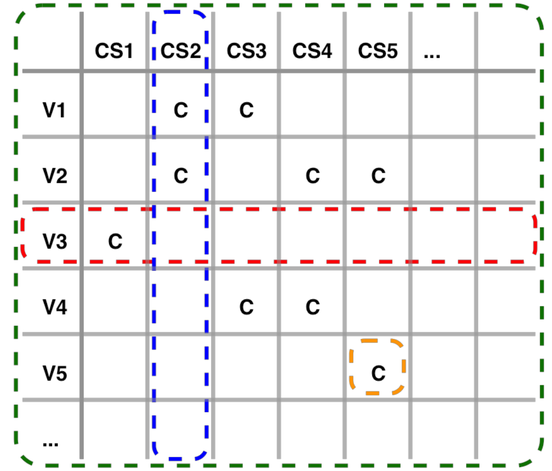
- CRAM format
- GenePred table format
- GFF format
- GTF format
- HAL format
- MAF format
- Microarray format
- Net format
- Personal Genome SNP format
- PSL format
- VCF format
- WIG format

me.ucsc.edu/FAQ/FAQformat.html

```
browser position chr7:127471196-127495720
browser hide all
track name="ItemRGBDemo" description="Item RGB                                    itemRgb="On"
chr7    127471196   127472363   Pos1  0  +  127                                    
chr7    127472363   127473530   Pos2  0  +  127472363   127473530   255,0,0
chr7    127473530   127474697   Pos3  0  +  127473530   127474697   255,0,0
chr7    127474697   127475864   Pos4  0  +  127474697   127475864   255,0,0
chr7    127475864   127477031   Neg1  0  -  127475864   127477031   0,0,255
chr7    127477031   127478198   Neg2  0  -  127477031   127478198   0,0,255
chr7    127478198   127479365   Neg3  0  -  127478198   127479365   0,0,255
chr7    127479365   127480532   Pos5  0  +  127479365   127480532   255,0,0
chr7    127480532   127481699   Neg4  0  -  127480532   127481699   0,0,255
```

BED file example

The VCF file format
Standard for variant representation



Example

VCF header

```
##fileformat=VCFv4.0
##fileDate=20100707
##source=VCFtools
##reference=NCBI36
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality (phred score)">
##FORMAT=<ID=GL,Number=3,Type=Float,Description="Likelihoods for RR,RA,AA genotypes (R=ref,A=alt)">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##ALT=<ID=DEL,Description="Deletion">
##INFO=<ID=SVTYPE,Number=1,Type=String,Description="Type of structural variant">
##INFO=<ID=END,Number=1,Type=Integer,Description="End position of the variant">
```

Mandatory header lines

Optional header lines (meta-data about the annotations in the VCF body)

Body

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	SAMPLE1	SAMPLE2
1	1	.	ACG	A,AT	.	PASS	.	GT:DP	1/2:13	0/0:29
1	2	rs1	C	T,CT	.	PASS	H2;AA=T	GT:GQ	0|1:100	2/2:70
1	5	.	A	G	.	PASS	.	GT:GQ	1|0:77	1/1:95
1	100	.	T	<DEL>	.	PASS	SVTYPE=DEL;END=300	GT:GQ:DP	1/1:12:3	0/0:20

Reference alleles (GT=0)

Alternate alleles (GT>0 is an index to the ALT column)

Deletion

SNP

Large SV

Insertion

Other event

Phased data (G and C above are on the same chromosome)

Variant
Call
Format

- stores the results of a single or multiple interpretations of genome sequencing datasets, in comparison to a reference genome
- standard format for file-based storage of human genome variants

Global Alliance for Genomics & Health

Task: Estimate Storage Requirements for 1000 Genomes

How much computer storage is required for 1000 Genomes

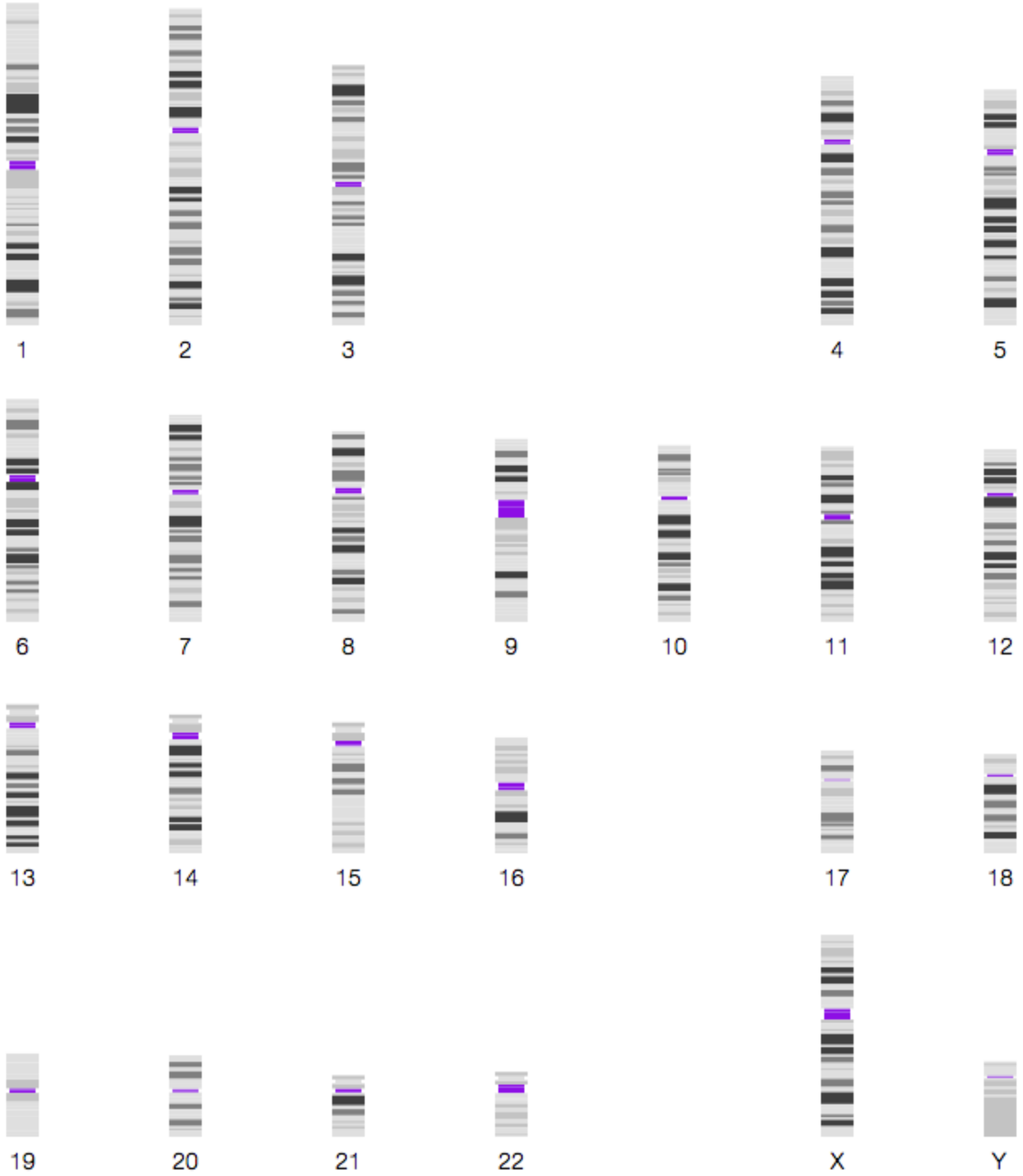
- WES & WGS
- Different file formats
  - SAM
  - BAM
  - VCF
  - FASTA
- Associated costs
  - Cost afctors
  - Raw Storage costs
- Familiarize with VCF format

  ➡ specification in article collection

IBM-storage-unit-3500-Schiphol-1957

Genome Editions

Sizes | positions | mappings



Chromosome	Basepair length (GRCh38)
1	248'956'422
2	242'193'529
3	198'295'559
4	190'214'555
5	181'538'259
6	170'805'979
7	159'345'973
8	145'138'636
9	138'394'717
10	133'797'422
11	135'086'622
12	133'275'309
13	114'364'328
14	107'043'718
15	101'991'189
16	90'338'345
17	83'257'441
18	80'373'285
19	59'128'983
20	64'444'167
21	46'709'983
22	50'818'468
X	156'040'895
Y	57'227'415
	3'080'419'480

genome.ucsc.edu

cytoBand_UCSC_hg18.txt

chr1	0	2300000	p36.33	gneg
chr1	2300000	5300000	p36.32	gpos25
chr1	5300000	7100000	p36.31	gneg
chr1	7100000	9200000	p36.23	gpos25
chr1	9200000	12600000	p36.22	gneg
chr1	12600000	16100000	p36.21	gpos50
chr1	16100000	20300000	p36.13	gneg
chr1	20300000	23800000	p36.12	gpos25
chr1	23800000	27800000	p36.11	gneg
chr1	27800000	30000000	p35.3	gpos25
chr1	30000000	32200000	p35.2	gneg
chr1	32200000	34400000	p35.1	gpos25
chr1	34400000	39600000	p34.3	gneg
chr1	39600000	43900000	p34.2	gpos25
chr1	43900000	46500000	p34.1	gneg
chr1	46500000	51300000	p33	gpos75
chr1	51300000	56200000	p32.3	gneg
chr1	56200000	58700000	p32.2	gpos50
chr1	58700000	60900000	p32.1	gneg
...	...	...	...	...
chrX	130300000	133500000	q26.2	gpos25
chrX	133500000	137800000	q26.3	gneg
chrX	137800000	140100000	q27.1	gpos75
chrX	140100000	141900000	q27.2	gneg
chrX	141900000	146900000	q27.3	gpos100
chrX	146900000	154913754	q28	gneg
chrY	0	1700000	p11.32	gneg
chrY	1700000	3300000	p11.31	gpos50
chrY	3300000	11200000	p11.2	gneg
chrY	11200000	11300000	p11.1	acen
chrY	11300000	12500000	q11.1	acen
chrY	12500000	14300000	q11.21	gneg
chrY	14300000	19000000	q11.221	gpos50
chrY	19000000	21300000	q11.222	gneg
chrY	21300000	25400000	q11.223	gpos50
chrY	25400000	27200000	q11.23	gneg
chrY	27200000	57772954	q12	gvar

Cytogenetic band sizes

chromosome	band start position	band stop position	cytogenetic band	staining intensity	band size
chr6	63400000	63500000	q11.2	gneg	100000
chr15	64900000	65000000	q22.32	gpos25	100000
chr17	22100000	22200000	p11.1	acen	100000
chrX	65000000	65100000	q11.2	gneg	100000
chrY	11200000	11300000	p11.1	acen	100000
chr17	35400000	35600000	q21.1	gneg	200000
chr3	44400000	44700000	p21.32	gpos50	300000
chr3	51400000	51700000	p21.2	gpos25	300000
chr9	132500000	132800000	q34.12	gpos25	300000
chr13	45900000	46200000	q14.13	gneg	300000
chr15	65000000	65300000	q22.33	gneg	300000
chr1	120700000	121100000	p11.2	gneg	400000
chr8	39500000	39900000	p11.22	gpos25	400000
chr9	72700000	73100000	q21.12	gneg	400000
chr16	69400000	69800000	q22.2	gpos50	400000
chr19	43000000	43400000	q13.13	gneg	400000
chr9	70000000	70500000	q13	gneg	500000
chr20	41100000	41600000	q13.11	gneg	500000
...	...	...	...	...	...
chr9	51800000	60300000	q11	acen	8500000
chrX	76000000	84500000	q21.1	gpos100	8500000
chr11	76700000	85300000	q14.1	gpos100	8600000
chr13	77800000	86500000	q31.1	gpos100	8700000
chr7	77400000	86200000	q21.11	gpos100	8800000
chr8	29700000	38500000	p12	gpos75	8800000
chr3	14700000	23800000	p24.3	gpos100	9100000
chr5	82800000	91900000	q14.3	gpos100	9100000
chr6	104800000	113900000	q21	gneg	9100000
chrX	120700000	129800000	q25	gpos100	9100000
chr9	60300000	70000000	q12	gvar	9700000
chr1	212100000	222100000	q41	gpos100	10000000
chr1	128000000	142400000	q12	gvar	14400000
chr1	69500000	84700000	p31.1	gpos100	15200000
chrY	27200000	57772954	q12	gvar	30572954

Positional genomic data has to be evaluated in the context of the correct edition

Chromosome	Basepairs 2003 (HG16)	Basepairs 2006 (HG18)	Basepairs 2009 (HG19)	Basepairs 2013 (GRCh38)	HG16 => HG19
1	246'127'941	247'249'719	249'250'621	248'956'422	2'828'481
2	243'615'958	242'951'149	243'199'373	242'193'529	-1'422'429
3	199'344'050	199'501'827	198'022'430	198'295'559	-1'048'491
4	191'731'959	191'273'063	191'154'276	190'214'555	-1'517'404
5	181'034'922	180'857'866	180'915'260	181'538'259	503'337
6	170'914'576	170'899'992	171'115'067	170'805'979	-108'597
7	158'545'518	158'821'424	159'138'663	159'345'973	800'455
8	146'308'819	146'274'826	146'364'022	145'138'636	-1'170'183
9	136'372'045	140'273'252	141'213'431	138'394'717	2'022'672
10	135'037'215	135'374'737	135'534'747	133'797'422	-1'239'793
11	134'482'954	134'452'384	135'006'516	135'086'622	603'668
12	132'078'379	132'349'534	133'851'895	133'275'309	1'196'930
13	113'042'980	114'142'980	115'169'878	114'364'328	1'321'348
14	105'311'216	106'368'585	107'349'540	107'043'718	1'732'502
15	100'256'656	100'338'915	102'531'392	101'991'189	1'734'533
16	90'041'932	88'827'254	90'354'753	90'338'345	296'413
17	81'860'266	78'774'742	81'195'210	83'257'441	1'397'175
18	76'115'139	76'117'153	78'077'248	80'373'285	4'258'146
19	63'811'651	63'811'651	59'128'983	59'128'983	-4'682'668
20	63'741'868	62'435'964	63'025'520	64'444'167	702'299
21	46'976'097	46'944'323	48'129'895	46'709'983	-266'114
22	49'396'972	49'691'432	51'304'566	50'818'468	1'421'496
X	153'692'391	154'913'754	155'270'560	156'040'895	2'348'504
Y	50'286'555	57'772'954	59'373'566	57'227'415	6'940'860
	3'070'128'059	**3'080'419'480**	**3'095'677'412**	**3'088'781'199**	**18'653'140**

Genome Liftover

Moving between genome editions

- different genome editions lead to shifting positions of defined elements such as genes
- local regions are frequently stable between editions
- shifts from change in regional lengths are defined in "chain files"
- chain files serve as guides for positional remapping using liftover methods
- Task: Read up on liftover techniques (starting w/ our article) & explore resources and other applications

SOFTWARE TOOL ARTICLE

REVISED **segment_liftover : a Python tool to convert segments between genome assemblies [version 2; referees: 2 approved]**

Bo Gao [1,2], Qingyao Huang[1,2], Michael Baudis [1,2]

[1]Institute of molecular Life Sciences, University of Zürich, Zürich, CH-8057, Switzerland
[2]Swiss Institute of Bioinformatics, University of Zürich, Zürich, CH-8057, Switzerland