Data repositories and File formats

Gustavo Adolfo Silva-Arias Ph.D

Technische Universität München

Bogotá, 24 Agosto 2021

Overview

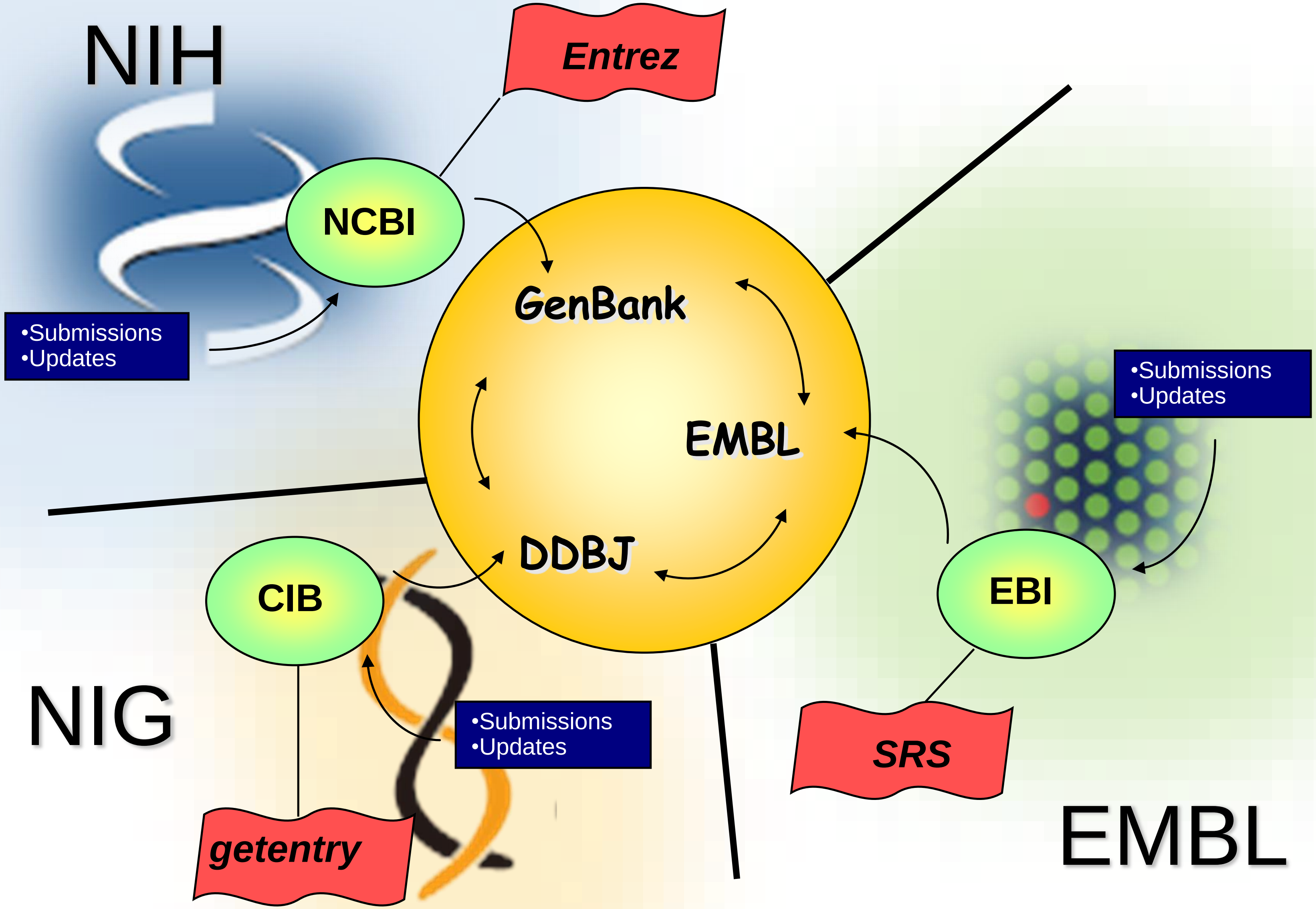
- Part 1
  - Associate any NGS related study publication with the databases
  - Link the raw data files with the study
  - Download all the relevant data

- Part 2
  - Familiarize with basic data file formats and understand the information

Data repositories

- The National Center for Biotechnology Information (NCBI)
  - Genbank
  - Sequence Read Archive (SRA)
- EMBL-EBI
  - European Nucleotide Archive (ENA)
  - Ensembl
  - UniProt
- DNA DataBank of Japan (DDBJ)
- … among many others
- https://www.nature.com/sdata/policies/repositories
- https://en.wikipedia.org/wiki/List_of_biological_databases

The International Sequence Database Collaboration



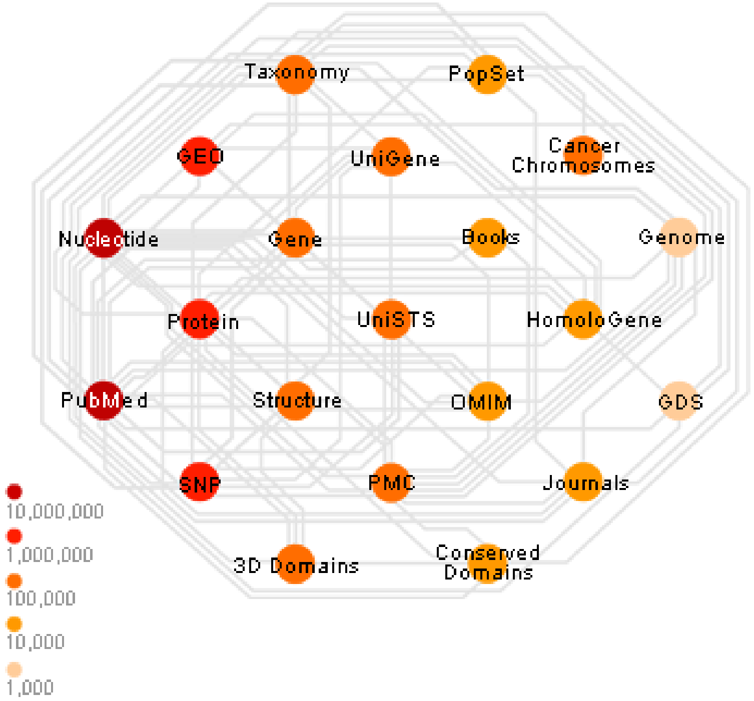
Data repositories - NCBI

Accepts submissions of:
- Bibliographic records (publication)
- Primary research data (nucleotide sequences for an organism/gene)

Organizes the information into databases, maintains them, makes them available to the world

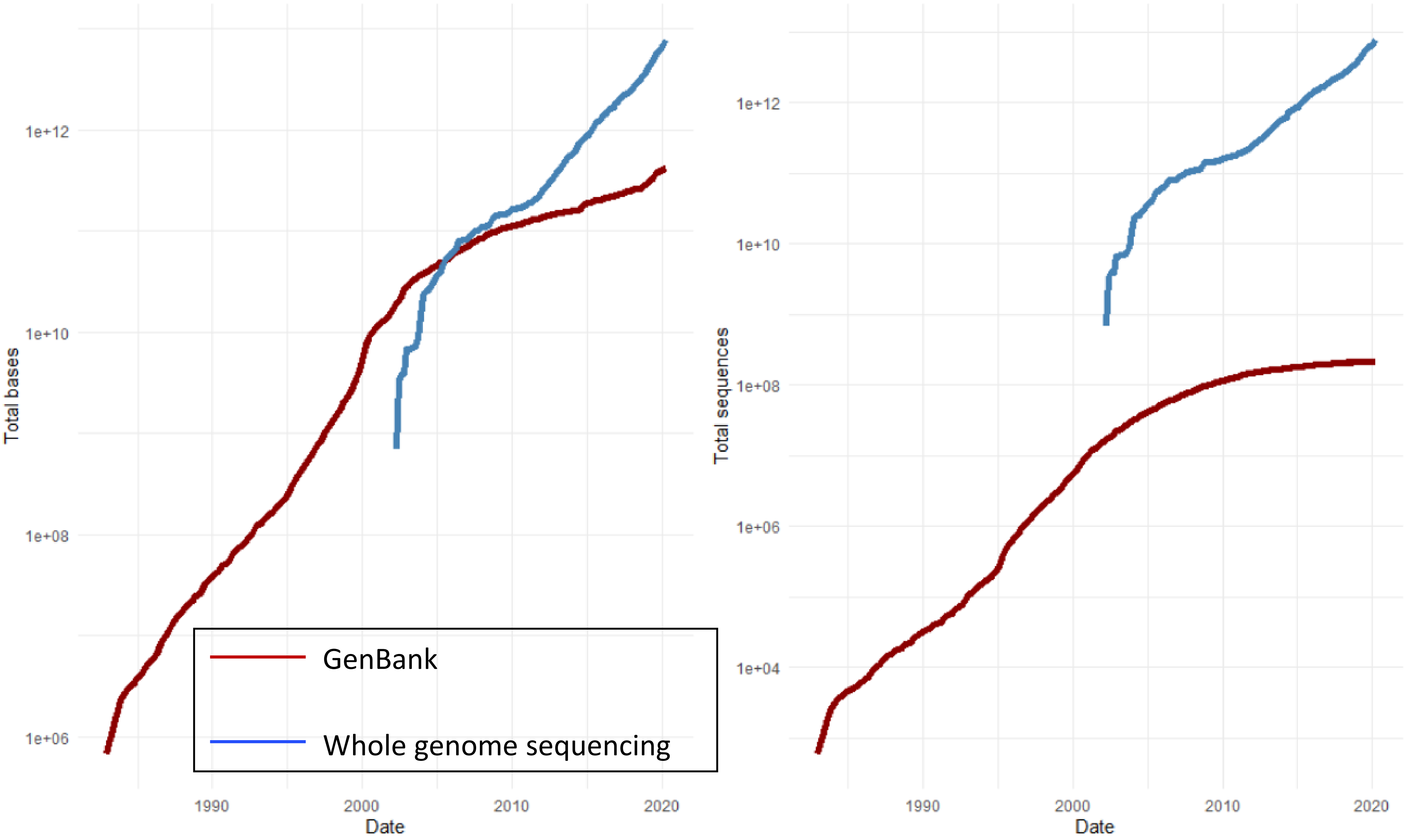
Develops software to retrieve and analyze the data conducts basic research to make new biological discoveries

NCBI databases
https://www.ncbi.nlm.nih.gov/guide/all/



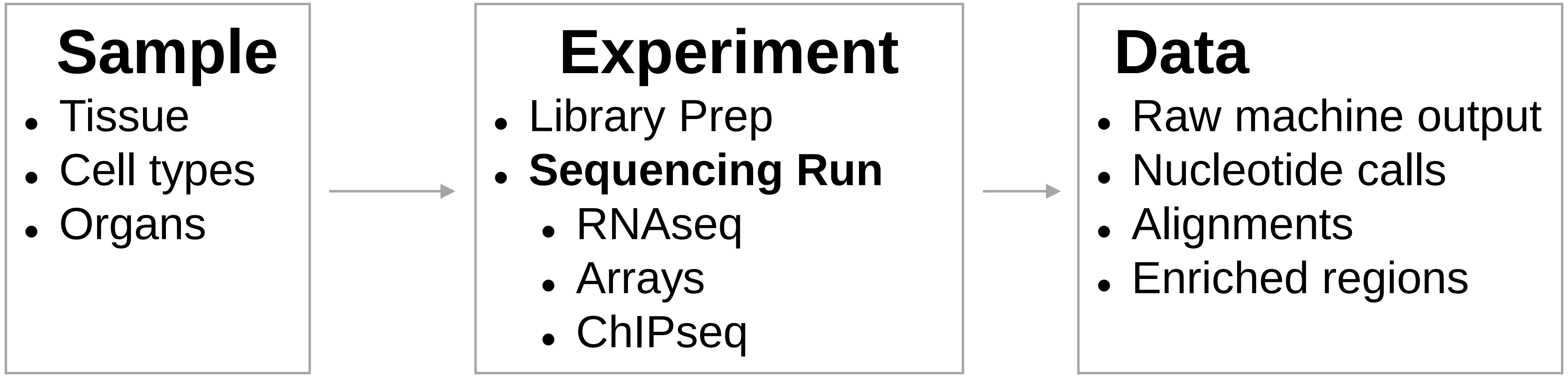
https://www.ncbi.nlm.nih.gov/Web/Search/entrezfs.html

Data repositories - NCBI



Source: https://www.ncbi.nlm.nih.gov/genbank/statistics/

Data repositories - Basic structure



Sample

- Tissue
- Cell types
- Organs

Experiment

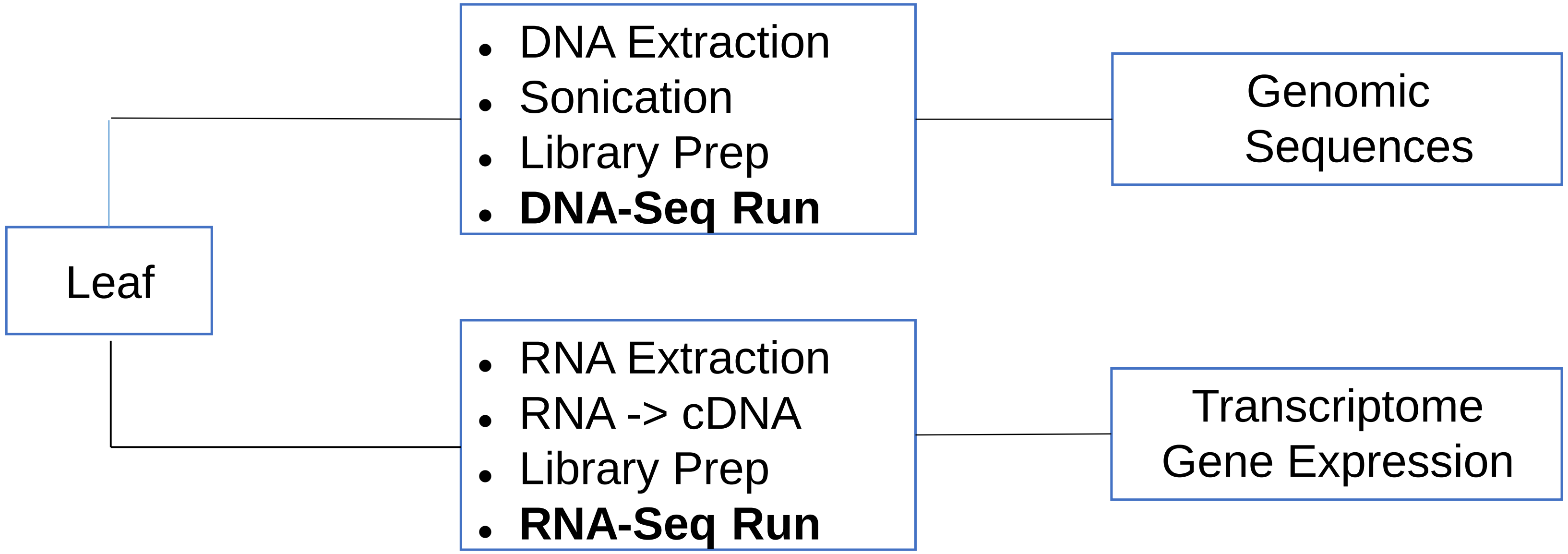
- Library Prep
- **Sequencing Run**
  - RNAseq
  - Arrays
  - ChIPseq

Data

- Raw machine output
- Nucleotide calls
- Alignments
- Enriched regions

Sample → **E**xperiment → **D**ata

Data structure - Basic projects



Leaf

- DNA Extraction
- Sonication
- Library Prep
- **DNA-Seq Run**

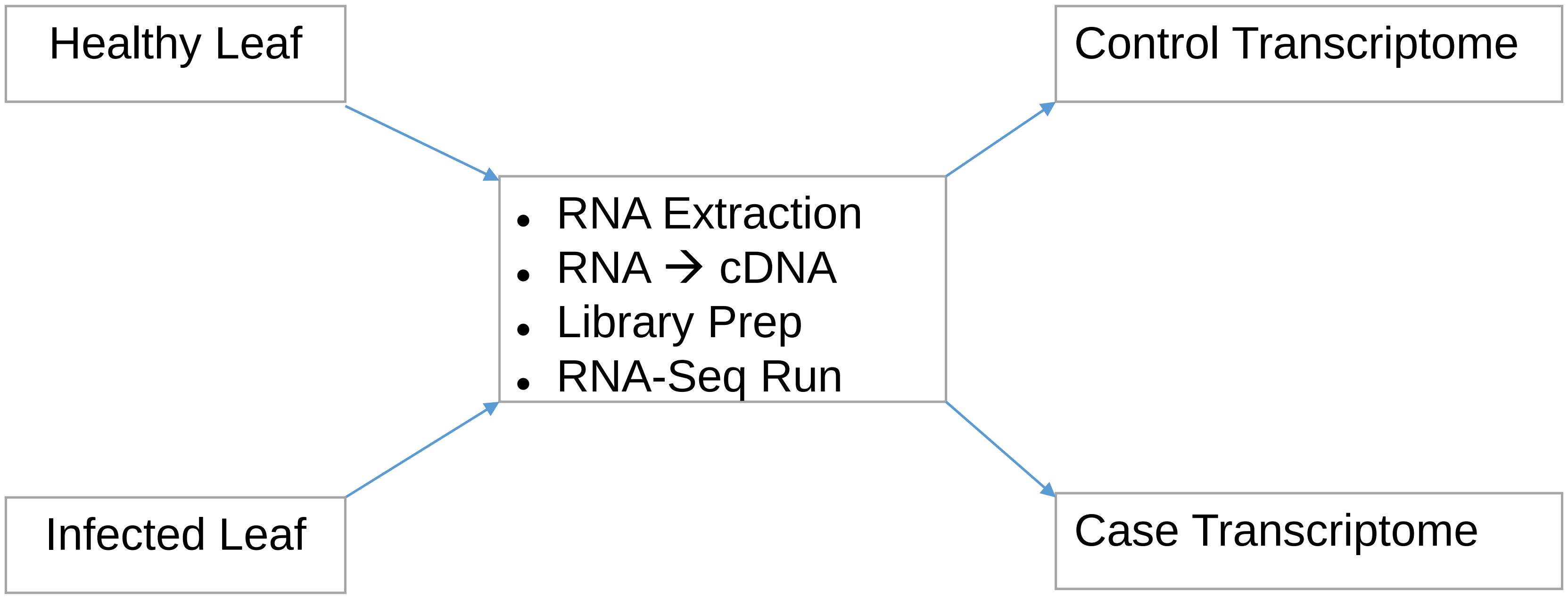
Genomic Sequences

- RNA Extraction
- RNA -> cDNA
- Library Prep
- **RNA-Seq Run**

Transcriptome Gene Expression

Sample → **Experiment** → **Data**

Data structure - Basic projects



Healthy Leaf

Infected Leaf

- RNA Extraction
- RNA → cDNA
- Library Prep
- RNA-Seq Run

Control Transcriptome

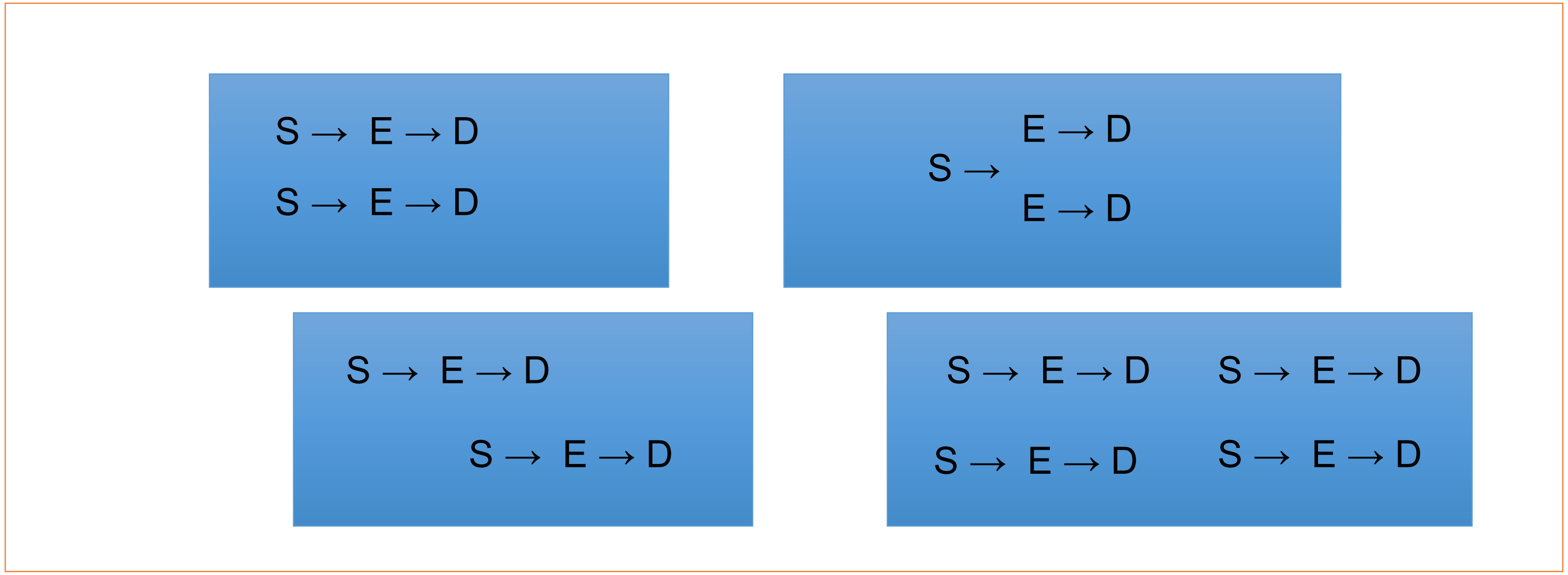
Case Transcriptome

Sample → **Experiment** → **Data**

Data structure - Complex projects

BioProjects

- Initiative
- Organizations/Consortium
- Many studies in one big project



S → E → D

S → E → D

E → D

S →

E → D

S → E → D

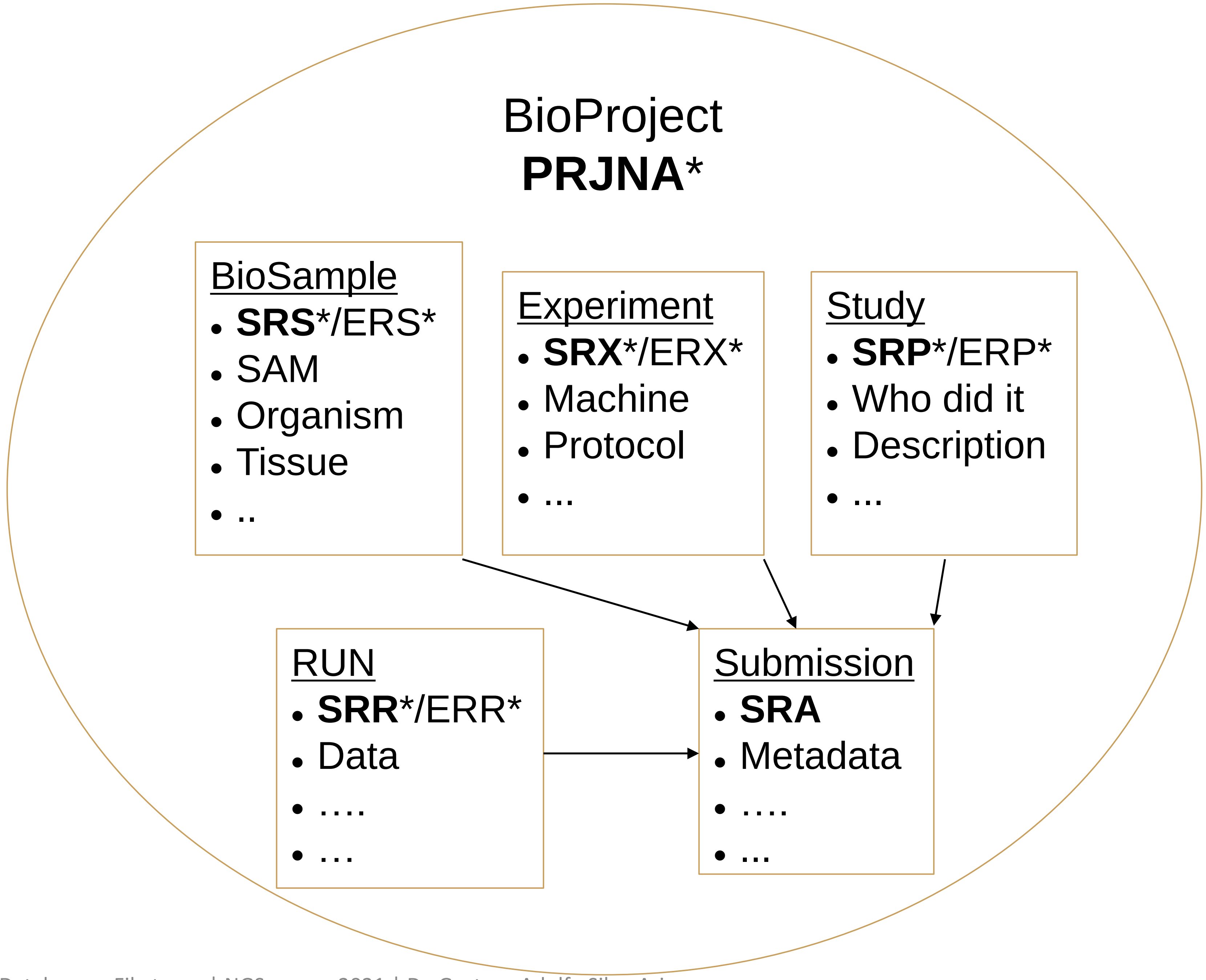
S → E → D

S → E → D

S → E → D

S → E → D

S → E → D



BioProject

PRJNA*

BioSample

- **SRS***/ERS*
- SAM
- Organism
- Tissue
- ..

Experiment

- **SRX***/ERX*
- Machine
- Protocol
- …

Study

- **SRP***/ERP*
- Who did it
- Description
- …

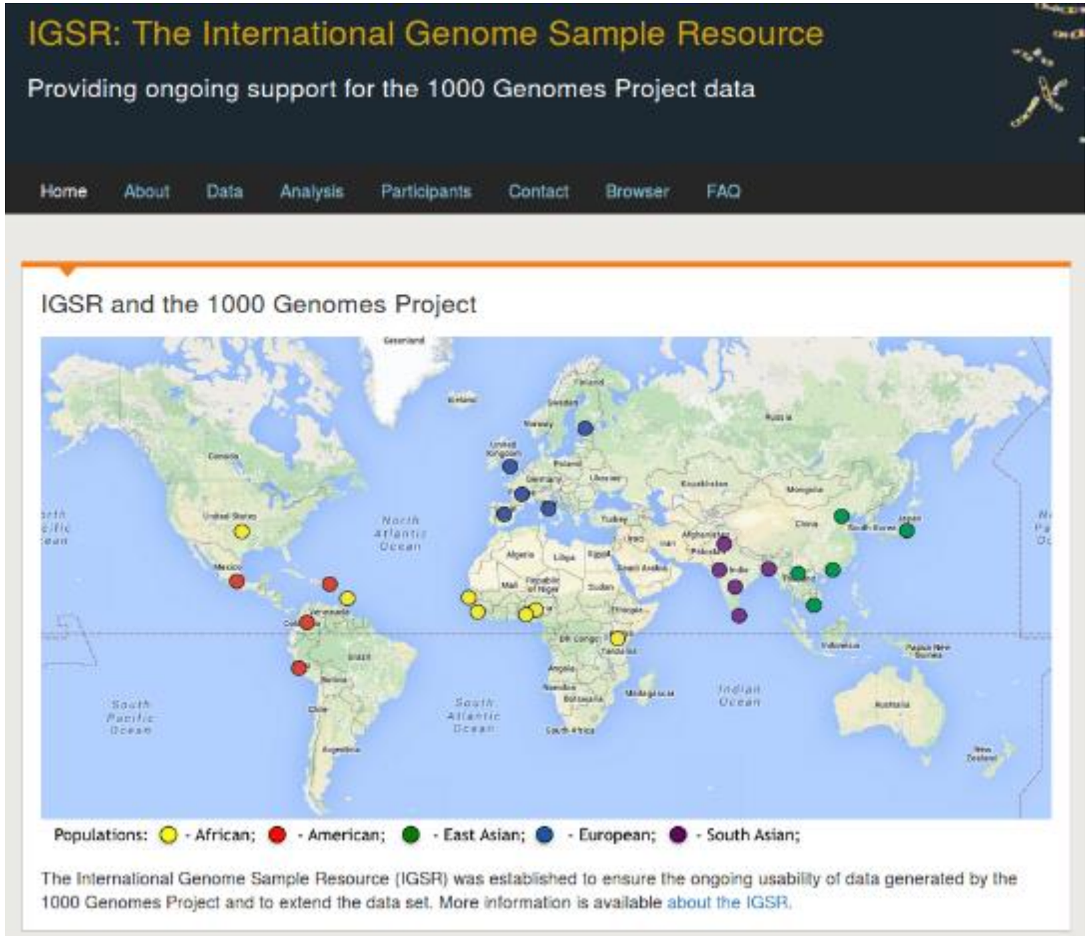
RUN

- **SRR***/ERR*
- Data
- ….
- …

Submission

- **SRA**
- Metadata
- ….
- …

Data structure - Complex projects
1000 genomes bioproject



IGSR: The International Genome Sample Resource

Providing ongoing support for the 1000 Genomes Project data

IGSR and the 1000 Genomes Project

Populations: ● - African; ● - American; ● - East Asian; ● - European; ● - South Asian;

The International Genome Sample Resource (IGSR) was established to ensure the ongoing usability of data generated by the 1000 Genomes Project and to extend the data set. More information is available about the IGSR.

Accession : PRJNA28889

http://www.1000genomes.org/

Different SRA ID types

- **Study** (SRP)– A study is a set of experiments and has an overall goal.
- **Experiment** (SRX) – An experiment is a consistent set of laboratory operations on input material with an expected result.
- **Sample** (SRS)– An experiment targets one or more samples. Results are expressed in terms of individual samples or bundles of samples as defined by the experiment.
- **Run** (SRR)– Results are called runs. Runs comprise the data gathered for a sample or sample bundle and refer to a defining experiment.
- **Submission** (SRA) – A submission is a package of metadata and/or data objects and a directive for what to do with those objects.

Source : http://www.ncbi.nlm.nih.gov/books/NBK47533/
Also: http://www.ncbi.nlm.nih.gov/books/NBK56913/

Data structure - Complex projects

Tomato genome

nature
THE TOMATO GENOME
Sequencing the culinary staple and its closest wild relative from South America
THE GREAT ESCAPE
WRITE OF SPRING
REPAIRS OF THE HEART

BioSample: SAMN02981290

BioProjects:

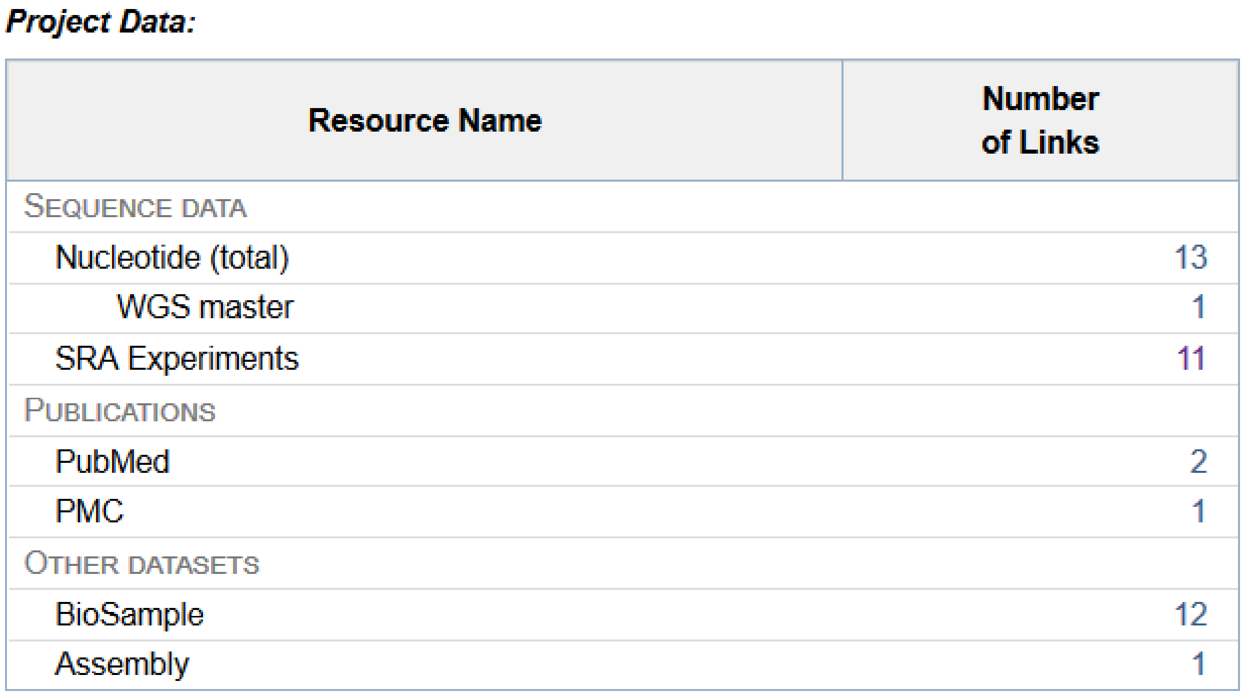
PRJNA66163 Solanum lycopersicum
strain: Heinz 1706

PRJNA119 Solanum lycopersicum
cultivar:Heinz 1706

Source: https://www.sgn.cornell.edu/organism/Solanum_lycopersicum/genome

Data structure - Complex projects
Tomato genome

BioProject:
PRJNA119

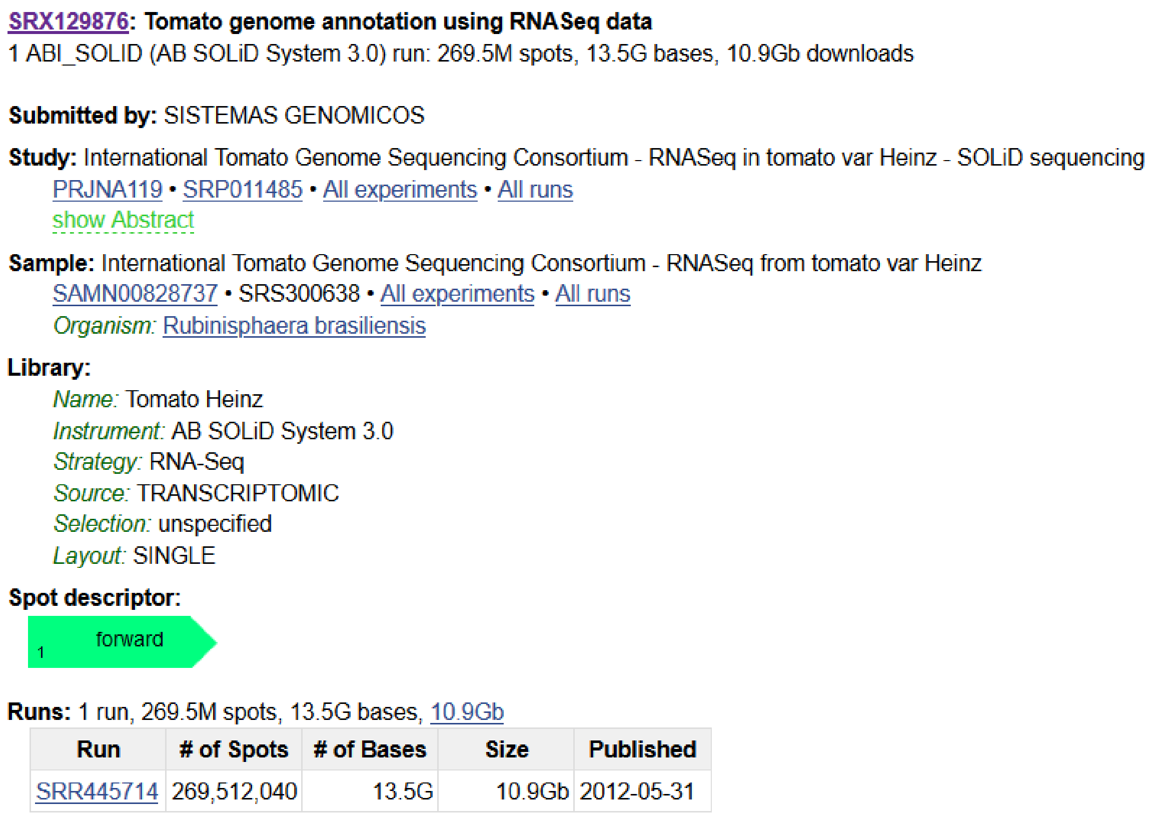


Project Data:

Resource Name	Number of Links
SEQUENCE DATA	
Nucleotide (total)	13
WGS master	1
SRA Experiments	11
PUBLICATIONS	
PubMed	2
PMC	1
OTHER DATASETS	
BioSample	12
Assembly	1

Data structure - Complex projects
Tomato genome

Experiment:
SRX129876



SRX129876: Tomato genome annotation using RNASeq data
1 ABI_SOLID (AB SOLiD System 3.0) run: 269.5M spots, 13.5G bases, 10.9Gb downloads

Submitted by: SISTEMAS GENOMICOS

Study: International Tomato Genome Sequencing Consortium - RNASeq in tomato var Heinz - SOLiD sequencing
PRJNA119 • SRP011485 • All experiments • All runs
show Abstract

Sample: International Tomato Genome Sequencing Consortium - RNASeq from tomato var Heinz
SAMN00828737 • SRS300638 • All experiments • All runs
Organism: Rubinisphaera brasiliensis

Library:
Name: Tomato Heinz
Instrument: AB SOLiD System 3.0
Strategy: RNA-Seq
Source: TRANSCRIPTOMIC
Selection: unspecified
Layout: SINGLE

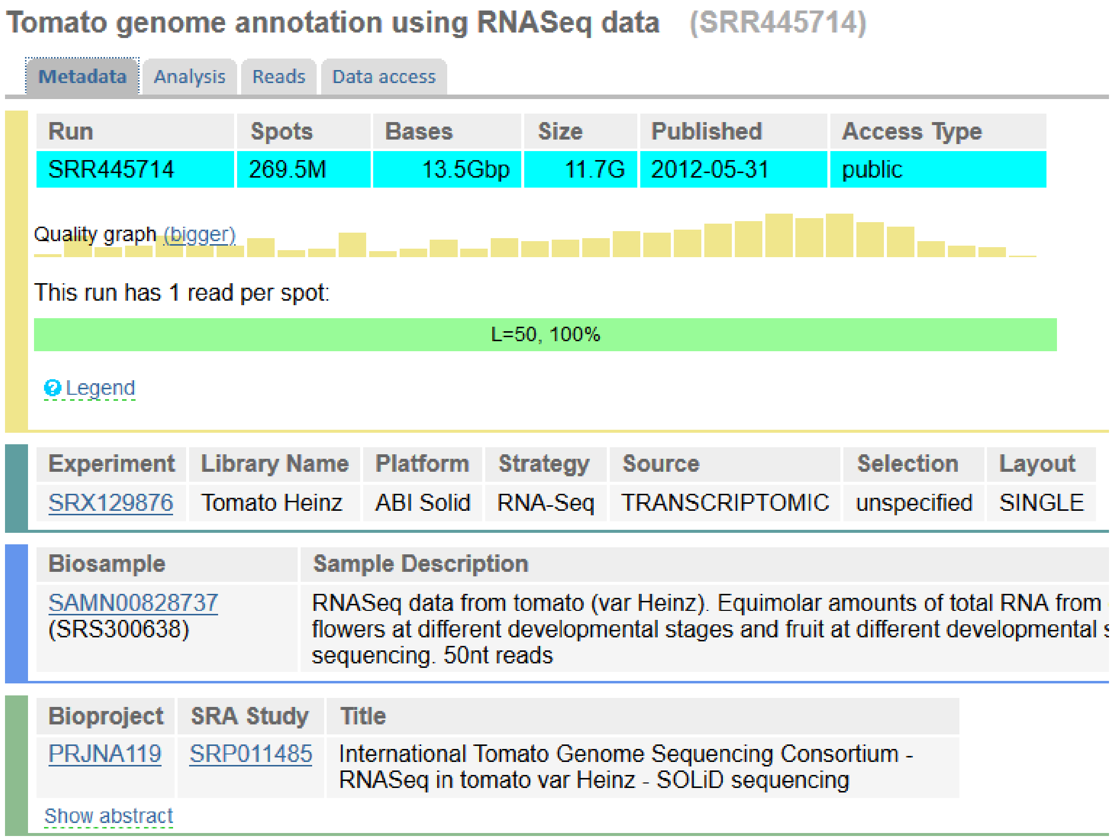
Spot descriptor:

Runs: 1 run, 269.5M spots, 13.5G bases, 10.9Gb

Run	# of Spots	# of Bases	Size	Published
SRR445714	269,512,040	13.5G	10.9Gb	2012-05-31

Data structure - Complex projects
Tomato genome

Data:
SRR445714



Tomato genome annotation using RNASeq data (SRR445714)

Run	Spots	Bases	Size	Published	Access Type
SRR445714	269.5M	13.5Gbp	11.7G	2012-05-31	public

Quality graph (bigger)

This run has 1 read per spot:

L=50, 100%

Legend

Experiment	Library Name	Platform	Strategy	Source	Selection	Layout
SRX129876	Tomato Heinz	ABI Solid	RNA-Seq	TRANSCRIPTOMIC	unspecified	SINGLE

Biosample	Sample Description
SAMN00828737 (SRS300638)	RNASeq data from tomato (var Heinz). Equimolar amounts of total RNA from flowers at different developmental stages and fruit at different developmental s sequencing. 50nt reads

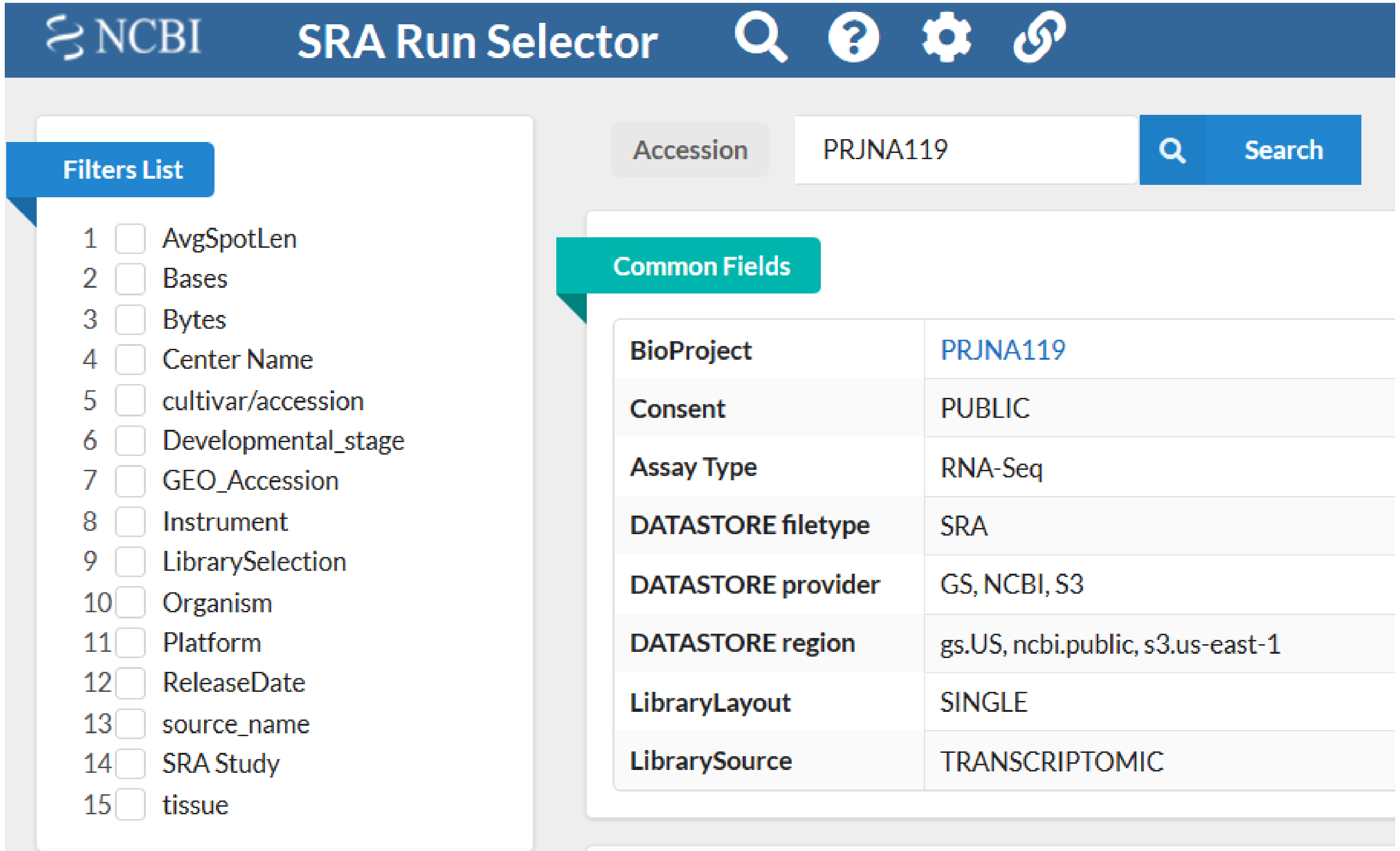
Bioproject	SRA Study	Title
PRJNA119	SRP011485	International Tomato Genome Sequencing Consortium - RNASeq in tomato var Heinz - SOLiD sequencing

Show abstract

Data structure - Complex projects

Tomato genome

SRA Run Selector



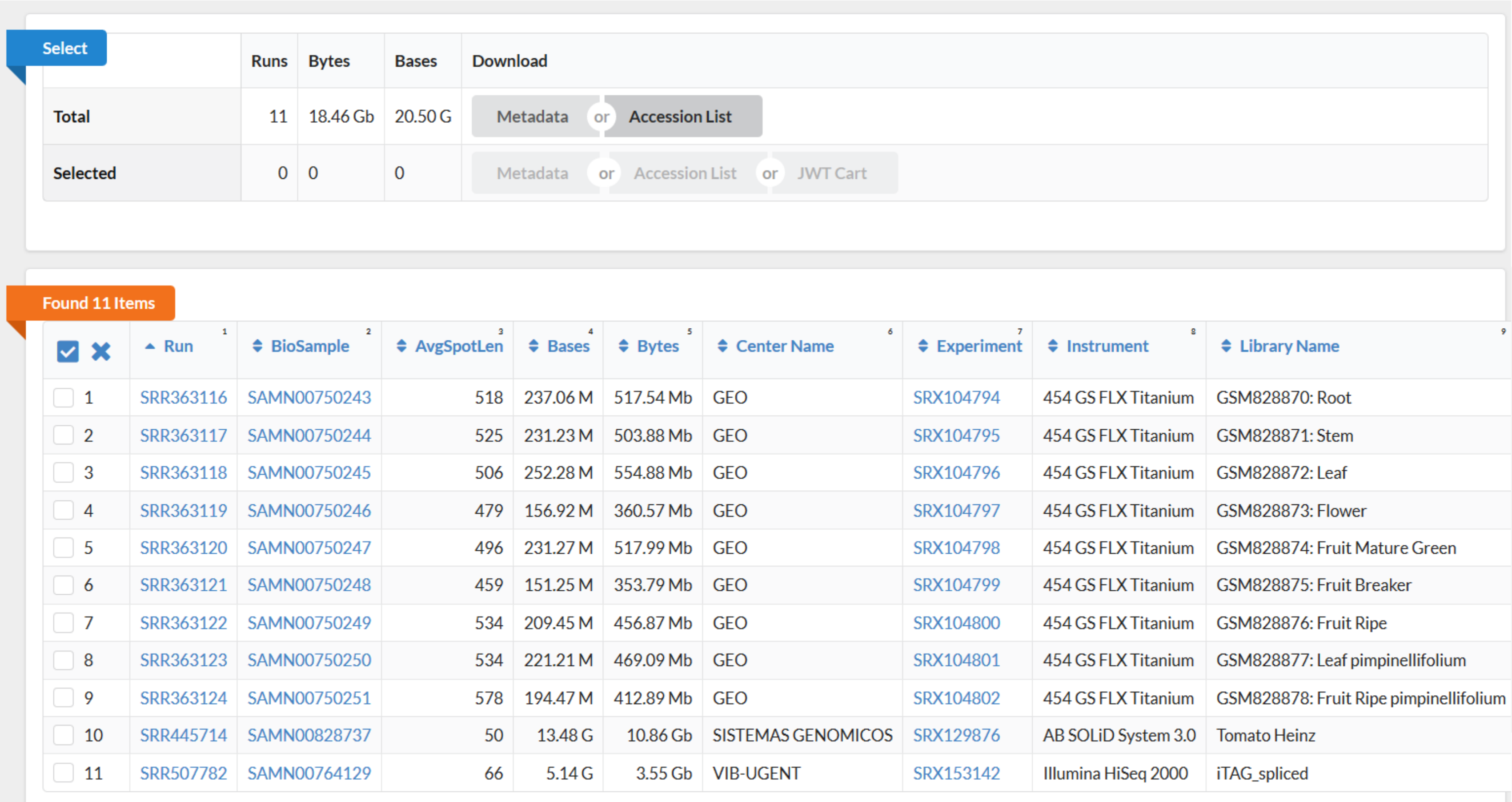
PRJNA119

1 □ AvgSpotLen
2 □ Bases
3 □ Bytes
4 □ Center Name
5 □ cultivar/accession
6 □ Developmental_stage
7 □ GEO_Accession
8 □ Instrument
9 □ LibrarySelection
10 □ Organism
11 □ Platform
12 □ ReleaseDate
13 □ source_name
14 □ SRA Study
15 □ tissue

BioProject	PRJNA119
Consent	PUBLIC
Assay Type	RNA-Seq
DATASTORE filetype	SRA
DATASTORE provider	GS, NCBI, S3
DATASTORE region	gs.US, ncbi.public, s3.us-east-1
LibraryLayout	SINGLE
LibrarySource	TRANSCRIPTOMIC

Data structure - Complex projects
Tomato genome

SRA Run Selector



Select	Runs	Bytes	Bases	Download
Total	11	18.46 Gb	20.50 G	Metadata or Accession List
Selected	0	0	0	Metadata or Accession List or JWT Cart

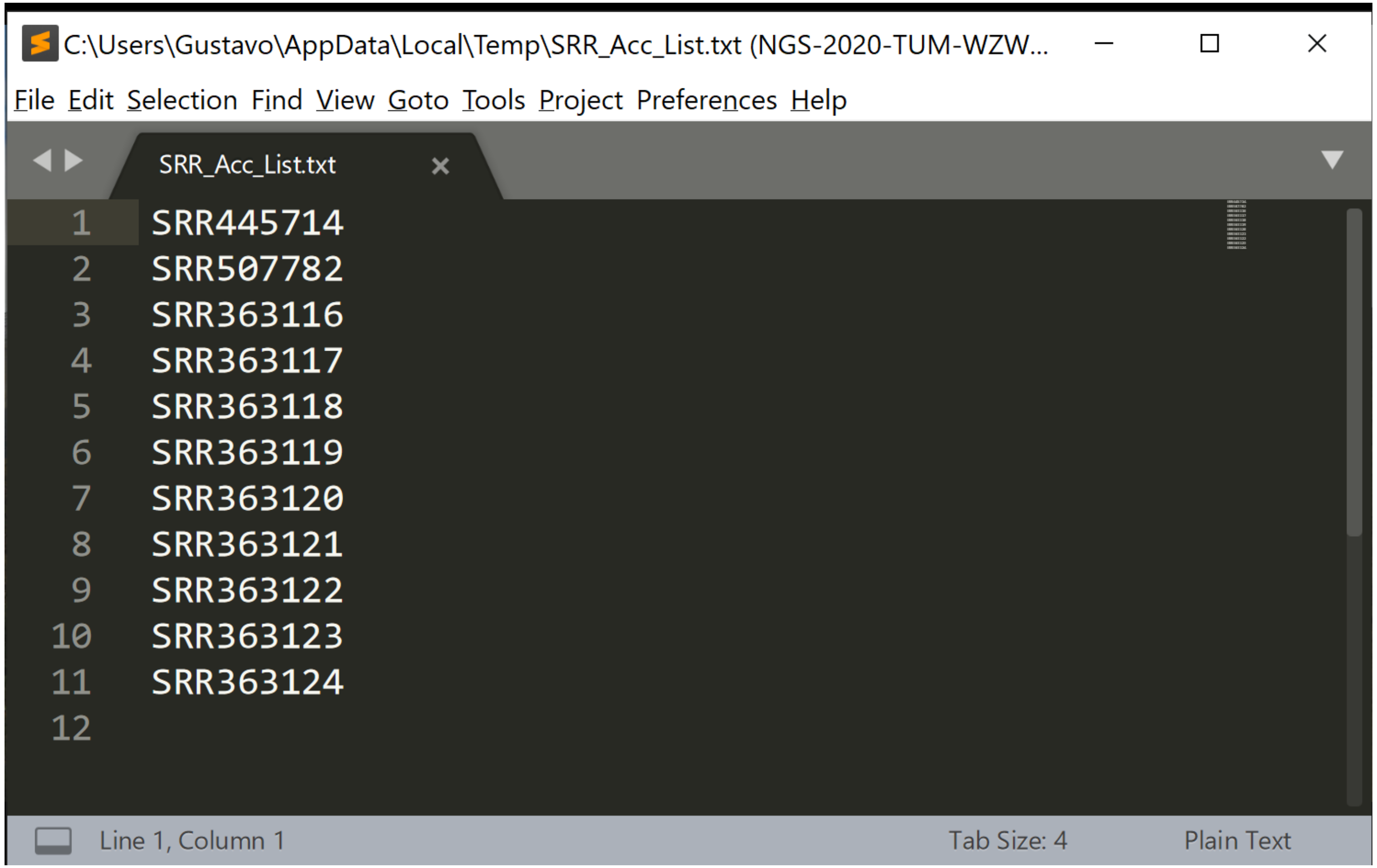
Found 11 Items

☑ ✖	Run	BioSample	AvgSpotLen	Bases	Bytes	Center Name	Experiment	Instrument	Library Name
□ 1	SRR363116	SAMN00750243	518	237.06 M	517.54 Mb	GEO	SRX104794	454 GS FLX Titanium	GSM828870: Root
□ 2	SRR363117	SAMN00750244	525	231.23 M	503.88 Mb	GEO	SRX104795	454 GS FLX Titanium	GSM828871: Stem
□ 3	SRR363118	SAMN00750245	506	252.28 M	554.88 Mb	GEO	SRX104796	454 GS FLX Titanium	GSM828872: Leaf
□ 4	SRR363119	SAMN00750246	479	156.92 M	360.57 Mb	GEO	SRX104797	454 GS FLX Titanium	GSM828873: Flower
□ 5	SRR363120	SAMN00750247	496	231.27 M	517.99 Mb	GEO	SRX104798	454 GS FLX Titanium	GSM828874: Fruit Mature Green
□ 6	SRR363121	SAMN00750248	459	151.25 M	353.79 Mb	GEO	SRX104799	454 GS FLX Titanium	GSM828875: Fruit Breaker
□ 7	SRR363122	SAMN00750249	534	209.45 M	456.87 Mb	GEO	SRX104800	454 GS FLX Titanium	GSM828876: Fruit Ripe
□ 8	SRR363123	SAMN00750250	534	221.21 M	469.09 Mb	GEO	SRX104801	454 GS FLX Titanium	GSM828877: Leaf pimpinellifolium
□ 9	SRR363124	SAMN00750251	578	194.47 M	412.89 Mb	GEO	SRX104802	454 GS FLX Titanium	GSM828878: Fruit Ripe pimpinellifolium
□ 10	SRR445714	SAMN00828737	50	13.48 G	10.86 Gb	SISTEMAS GENOMICOS	SRX129876	AB SOLiD System 3.0	Tomato Heinz
□ 11	SRR507782	SAMN00764129	66	5.14 G	3.55 Gb	VIB-UGENT	SRX153142	Illumina HiSeq 2000	iTAG_spliced

Data structure - Complex projects

Tomato genome

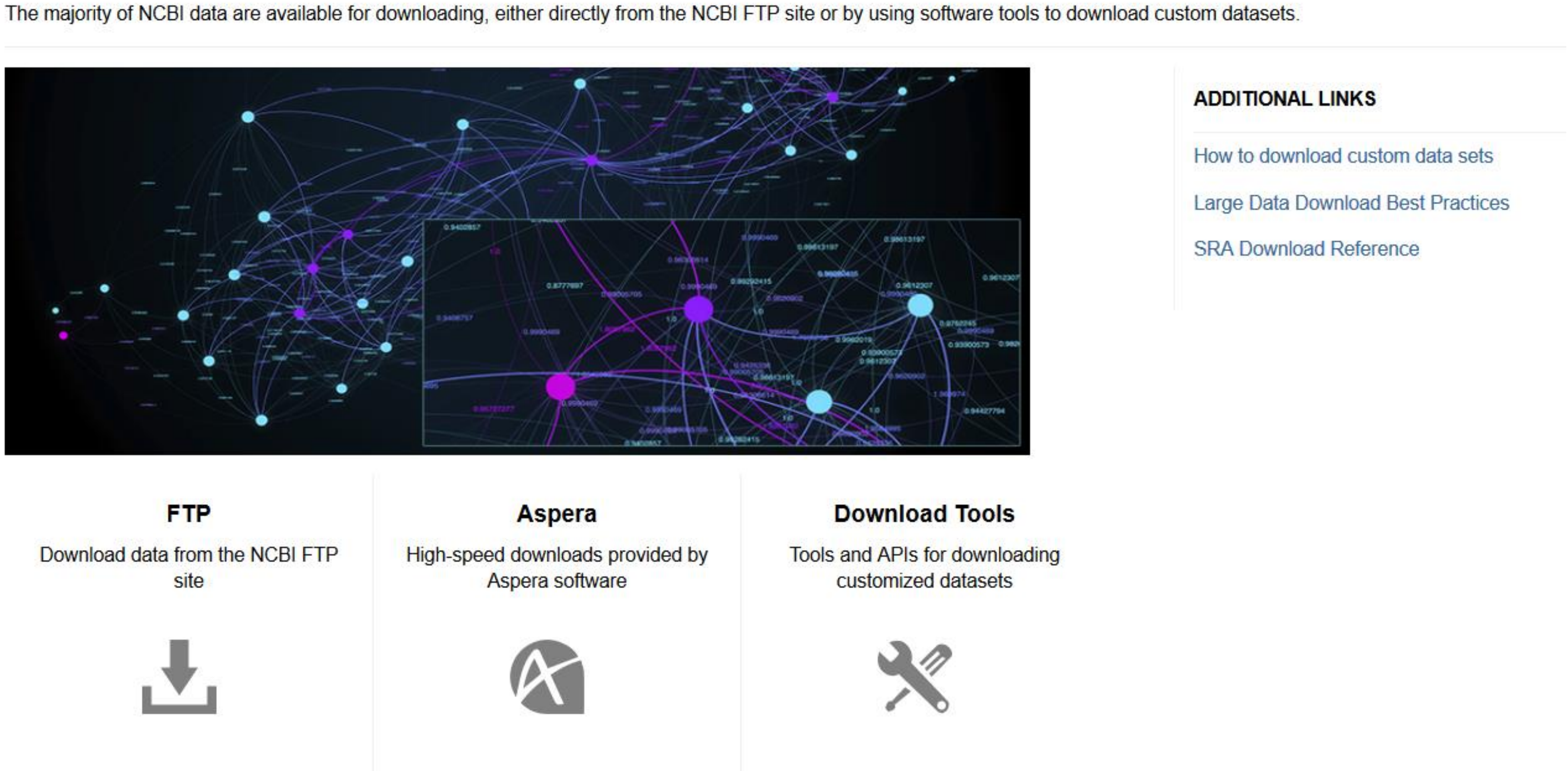
SRA Run Selector



C:\Users\Gustavo\AppData\Local\Temp\SRR_Acc_List.txt (NGS-2020-TUM-WZW...
File Edit Selection Find View Goto Tools Project Preferences Help
SRR_Acc_List.txt
Line 1, Column 1
Tab Size: 4
Plain Text

```
SRR445714
SRR507782
SRR363116
SRR363117
SRR363118
SRR363119
SRR363120
SRR363121
SRR363122
SRR363123
SRR363124
```

Download NGS data



The majority of NCBI data are available for downloading, either directly from the NCBI FTP site or by using software tools to download custom datasets.

ADDITIONAL LINKS

How to download custom data sets

Large Data Download Best Practices

SRA Download Reference

FTP

Download data from the NCBI FTP site

Aspera

High-speed downloads provided by Aspera software

Download Tools

Tools and APIs for downloading customized datasets

https://www.ncbi.nlm.nih.gov/home/download/

Download NGS data

- Direct Download (ftp, http, Aspera) (Browser or command line)

```
wget ftp://ftp.sra.ebi.ac.uk/…/SRR4454118_1.fastq.gz
```

- Direct Download from EBI/DDBJ

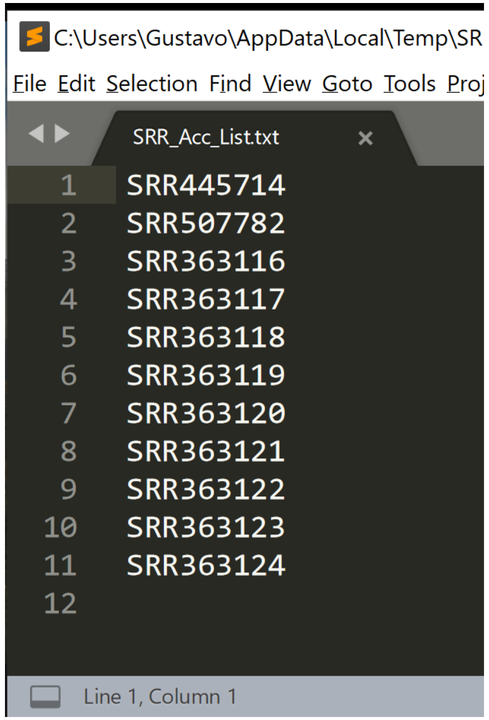
- **sra-toolkit** software has a command fastq-dump

Command	Argument	Input	Output
fastq-dump	-h		Print help
fastq-dump		SRR_ID	Download entire file
fastq-dump	-X <number>	SRR_ID	Download N spots
fastq-dump	--skip-technical	SRR_ID	Do not include technical reads
fastq-dump	-Z	SRR_ID	Print to terminal
fastq-dump	-F	SRR_ID	Get original id
fastq-dump	--split-files	SRR_ID	Print read pairs in separate files

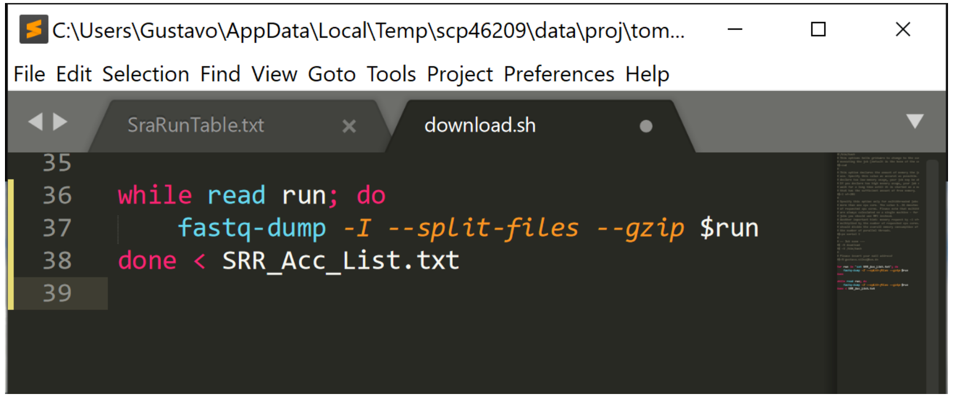
Download NGS data

- **sra-toolkit**

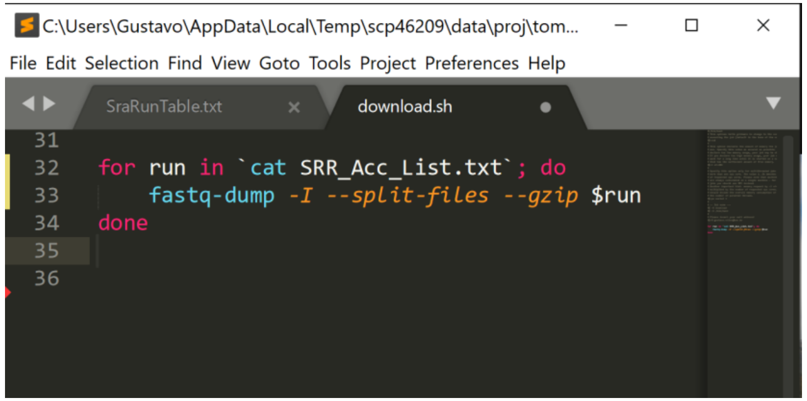
Fastq-dump (download fastq files, pair end in separated files, compressed)


C:\Users\Gustavo\AppData\Local\Temp\SR
File Edit Selection Find View Goto Tools Proj
SRR_Acc_List.txt
Line 1, Column 1

```
SRR445714
SRR507782
SRR363116
SRR363117
SRR363118
SRR363119
SRR363120
SRR363121
SRR363122
SRR363123
SRR363124

```


C:\Users\Gustavo\AppData\Local\Temp\scp46209\data\proj\tom...
File Edit Selection Find View Goto Tools Project Preferences Help
SraRunTable.txt
download.sh

```

while read run; do
    fastq-dump -I --split-files --gzip $run
done < SRR_Acc_List.txt

```


C:\Users\Gustavo\AppData\Local\Temp\scp46209\data\proj\tom...
File Edit Selection Find View Goto Tools Project Preferences Help
SraRunTable.txt
download.sh

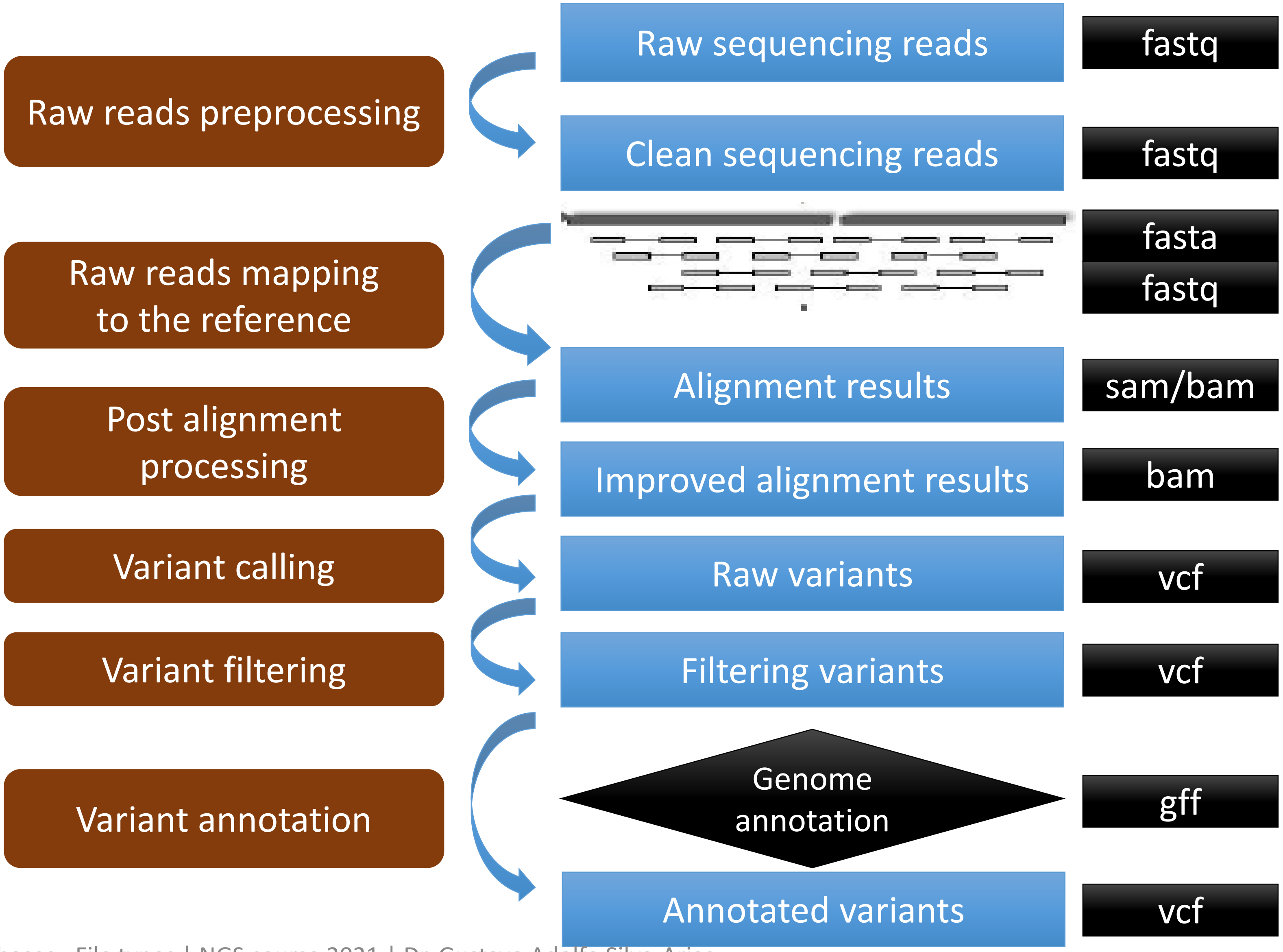
```

for run in `cat SRR_Acc_List.txt`; do
    fastq-dump -I --split-files --gzip $run
done


```

Overview

- Part 1
  - Associate any NGS related study publication with the databases
  - Link the raw data files with the study
  - Download all the relevant data

- Part 2
  - Familiarize with basic data file formats and understand the information

NGS data processing workflow



Raw reads preprocessing: Raw sequencing reads (fastq) → Clean sequencing reads (fastq)

Raw reads mapping to the reference: fasta, fastq → Alignment results (sam/bam)

Post alignment processing: → Improved alignment results (bam)

Variant calling: → Raw variants (vcf)

Variant filtering: → Filtering variants (vcf)

Variant annotation: Genome annotation (gff) → Annotated variants (vcf)

Data files overview

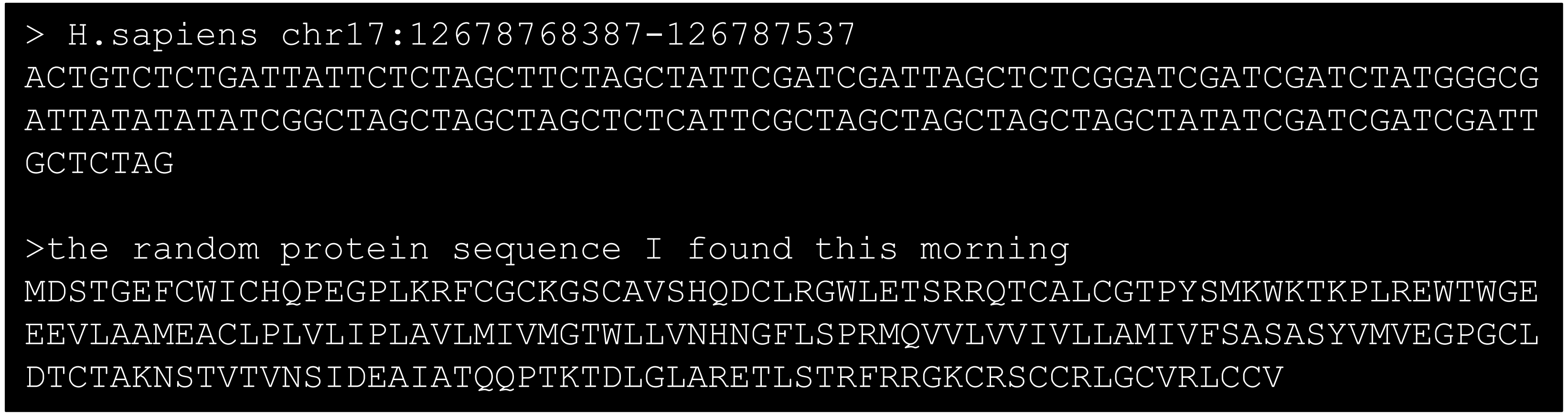
Fasta – reference sequence
Fastq – unprocessed reads
Sam – aligned reads to the reference
Bam – binary (compressed) SAM file
BED – browser format (store genomic regions)
GFF/GTF – annotations
VCF – variant calls

FASTA format

- Nucleotide or peptide sequence
- Simple structure
  - 2 lines per sequence

> Header
>
> Sequence

- Multiple sequences per file



```
> H.sapiens chr17:12678768387-126787537
ACTGTCTCTGATTATTCTCTAGCTTCTAGCTATTCGATCGATTAGCTCTCGGATCGATCGATCTATGGGCG
ATTATATATATCGGCTAGCTAGCTAGCTCTCATTCGCTAGCTAGCTAGCTAGCTATATCGATCGATCGATT
GCTCTAG

>the random protein sequence I found this morning
MDSTGEFCWICHQPEGPLKRFCGCKGSCAVSHQDCLRGWLETSRRQTCALCGTPYSMKWKTKPLREWTWGE
EEVLAAMEACLPLVLIPLAVLMIVMGTWLLVNHNGFLSPRMQVVLVVIVLLAMIVFSASASYVMVEGPGCL
DTCTAKNSTVTVNSIDEAIATQQPTKTDLGLARETLSTRFRRGKCRSCCRLGCVRLCCV
```

FASTQ format

Standard format for high-throughput sequencing instruments

4 lines per sequence (read)

@Header

Sequence

+

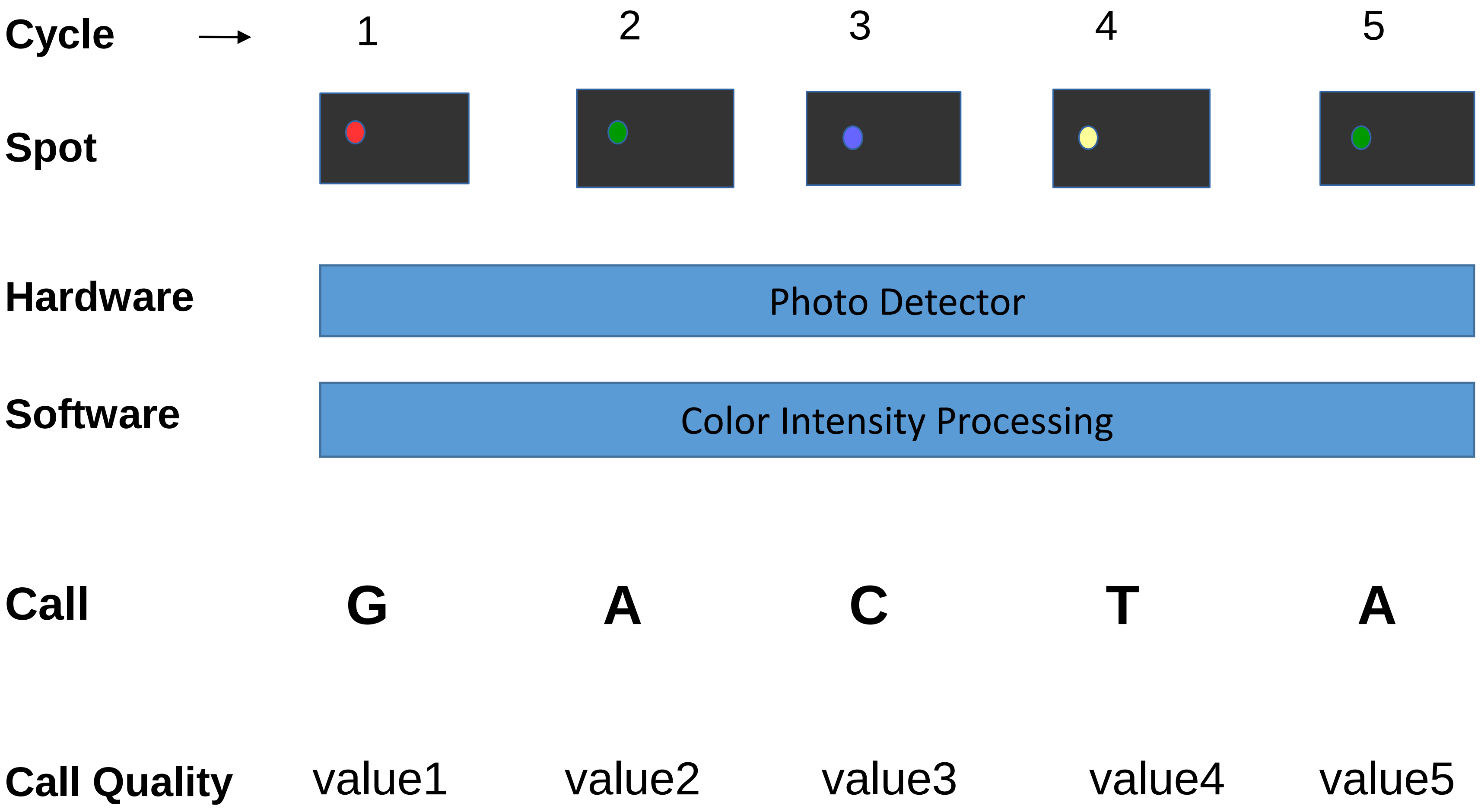
quality

Multiple sequences per file

```
@GWNJ-0850:627:GW190820000:5:1101:12033:1450 1:N:0:NCTCCTGA+NGGCTATA
CTTTTTCCTCGAGTATCTTTTGGAGGCGATTCTTTTTTTGAACTTGCTTTTTTTTTTGAGATCTACACGGTAGATTCAA
+
?@;DDDDDHDC<D<AEEHIGIII+<B@F?@FFGEHGIIII(77.7=7=AEHBBBB?=B8823(>A>(985++5(:@AC4
```

FASTQ format

Quality data

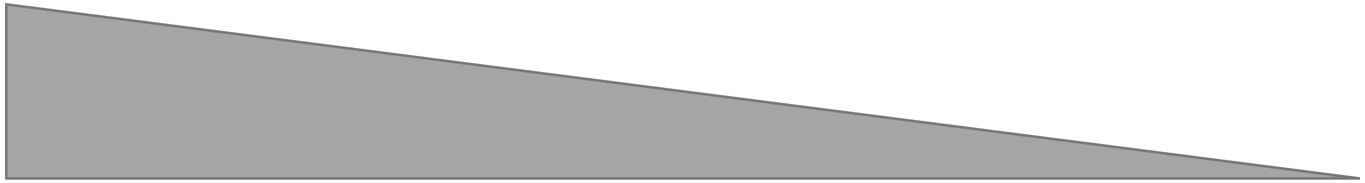


Cycle →	1	2	3	4	5
Spot	🔴	🟢	🔵	🟡	🟢
Hardware	Photo Detector				
Software	Color Intensity Processing				
Call	**G**	**A**	**C**	**T**	**A**
Call Quality	value1	value2	value3	value4	value5

FASTQ format
Quality data

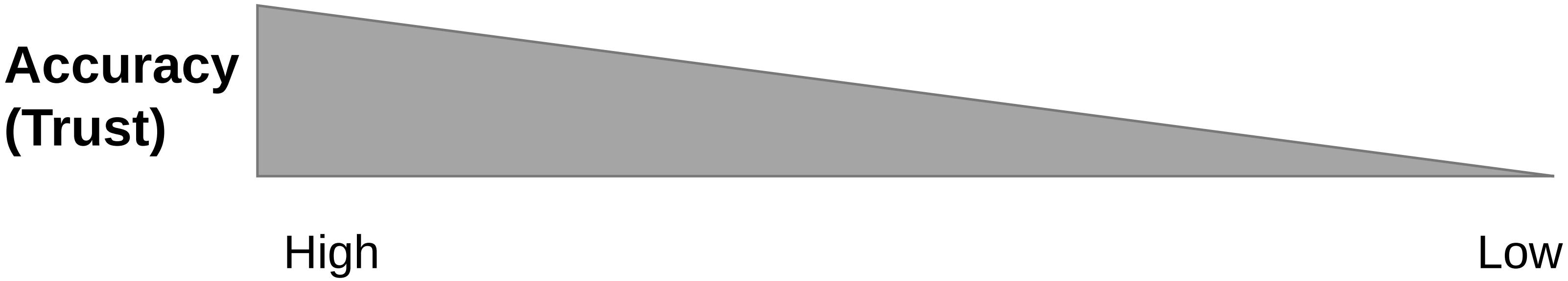
Call	**G**	**A**	**C**	**T**	**A**
Call Quality	value1	value2	value3	value4	value5
What User Wants	Awesome	Very good	Good	OK	BAD

Accuracy



FASTQ format

Quality data

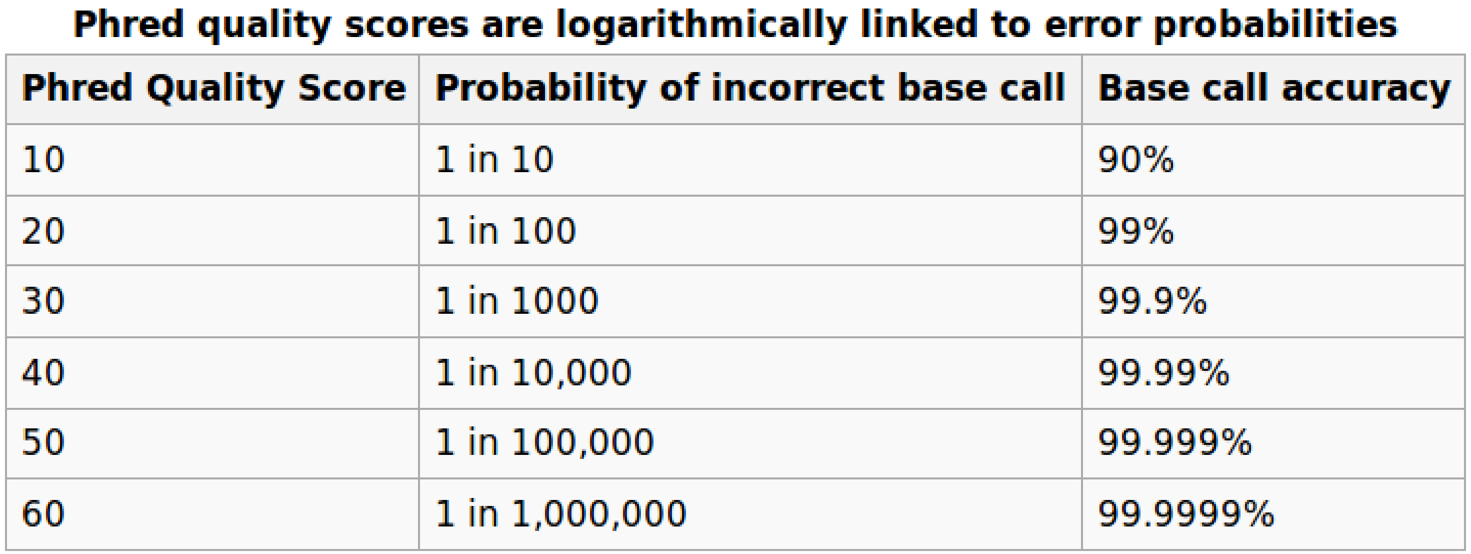


- Higher the value higher the trust
- **Higher the value higher the probability that call is correct**
- Amenable to statistical and probabilistic methods
- Common across all studies/platforms/machines
- Universally accepted
- Easily encoded/printed in a file

Phred Score

Quality data

- Denoted by letter **Q**
- **Q = - 10 log$_{10}$ P**
- **P**: probability of error or the call being wrong



Phred quality scores are logarithmically linked to error probabilities

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%
60	1 in 1,000,000	99.9999%

Phred scores: Phil Green‘s group, originally for Sanger reads. Ewing et al. (1998) Genome Res. 8:175-186

https://en.wikipedia.org/wiki/Phred_quality_score
https://www.illumina.com/documents/products/technotes/technote_Q-Scores.pdf

Source: Wikipedia

Sequence data and Phred scores together

- Encoding ~ printing the phred scores along with base calls in a file.
- Nucleotides are typically available as a fasta file
- Quality scores could be added to the fasta file?
- Cumbersome and space consuming

```
>read1
ATGC
>read1
10 20 30 40
```

Sequence data and Phred scores together

- … better solution
- Put calls and quality scores and one below another

```
>read1
ATGC
10 20 30 40
```

Encode ~ Encrypt

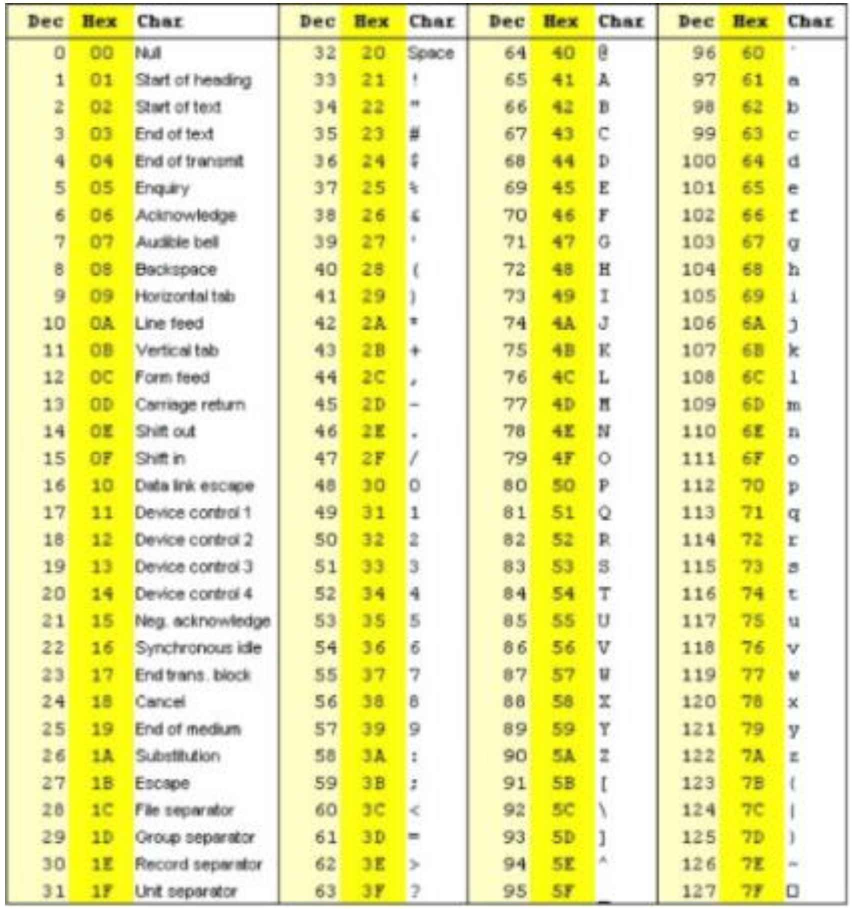
```
10 = +
20 = 5
30 = ?
40 = I
```

→

```
>read1
ATGC
+5?I
```

ASCII code

- **Decimal**
- 10 12 34 39 40 23 4 7 17 22 19 20 35 12 3 18 29 30 11 5 18 22
- Add 33 : 43 45 67 72 73 56 37 40 50 55 52 53 45 36 51 62 63 44 38 51 55
- **ASCII**
- +-CHI8%(2745-$3>?,&37



Dec	Hex	Char	Dec	Hex	Char	Dec	Hex	Char	Dec	Hex	Char
0	00	Null	32	20	Space	64	40	@	96	60	`
1	01	Start of heading	33	21	!	65	41	A	97	61	a
2	02	Start of text	34	22	"	66	42	B	98	62	b
3	03	End of text	35	23	#	67	43	C	99	63	c
4	04	End of transmit	36	24	$	68	44	D	100	64	d
5	05	Enquiry	37	25	%	69	45	E	101	65	e
6	06	Acknowledge	38	26	&	70	46	F	102	66	f
7	07	Audible bell	39	27	'	71	47	G	103	67	g
8	08	Backspace	40	28	(	72	48	H	104	68	h
9	09	Horizontal tab	41	29	)	73	49	I	105	69	i
10	0A	Line feed	42	2A	*	74	4A	J	106	6A	j
11	0B	Vertical tab	43	2B	+	75	4B	K	107	6B	k
12	0C	Form feed	44	2C	,	76	4C	L	108	6C	l
13	0D	Carriage return	45	2D	-	77	4D	M	109	6D	m
14	0E	Shift out	46	2E	.	78	4E	N	110	6E	n
15	0F	Shift in	47	2F	/	79	4F	O	111	6F	o
16	10	Data link escape	48	30	0	80	50	P	112	70	p
17	11	Device control 1	49	31	1	81	51	Q	113	71	q
18	12	Device control 2	50	32	2	82	52	R	114	72	r
19	13	Device control 3	51	33	3	83	53	S	115	73	s
20	14	Device control 4	52	34	4	84	54	T	116	74	t
21	15	Neg. acknowledge	53	35	5	85	55	U	117	75	u
22	16	Synchronous idle	54	36	6	86	56	V	118	76	v
23	17	End trans. block	55	37	7	87	57	W	119	77	w
24	18	Cancel	56	38	8	88	58	X	120	78	x
25	19	End of medium	57	39	9	89	59	Y	121	79	y
26	1A	Substitution	58	3A	:	90	5A	Z	122	7A	z
27	1B	Escape	59	3B	;	91	5B	[	123	7B	{
28	1C	File separator	60	3C	<	92	5C	\	124	7C	|
29	1D	Group separator	61	3D	=	93	5D	]	125	7D	}
30	1E	Record separator	62	3E	>	94	5E	^	126	7E	~
31	1F	Unit separator	63	3F	?	95	5F	_	127	7F	□

Phred to ASCII

- Depends on encoding
- Sanger Encoding
  - Add 33 to the phred score and convert the number to character
  - Subtract 33 from the ascii code of the character
- Illumina encoding < 1.8 add 64
- Illumina encoding 1.8+ add 33

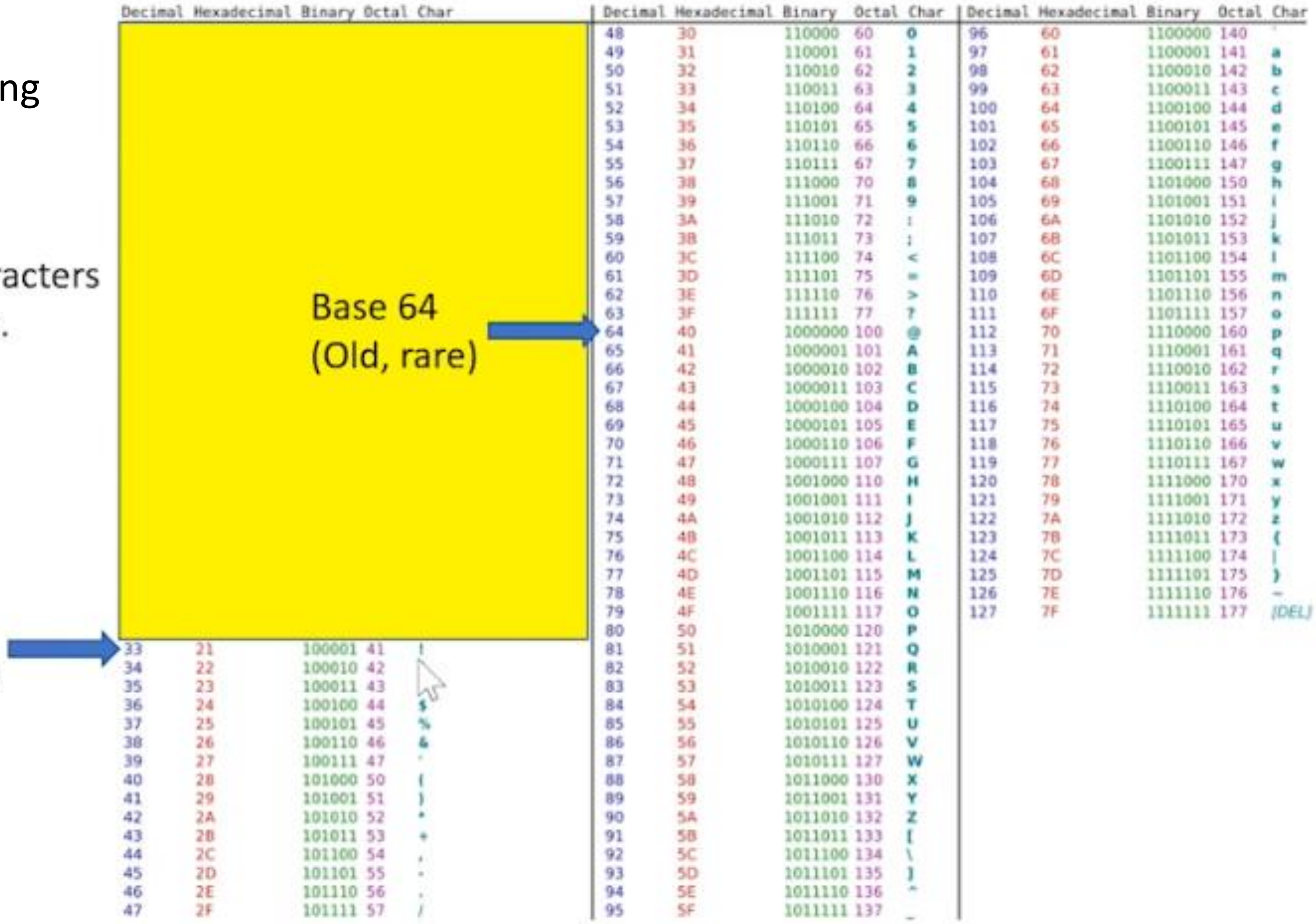
- **Software like FASTQC will tell you the encoding**

Phred to ASCII

- Encoding

These characters don't print.

Base 33 (Typical)

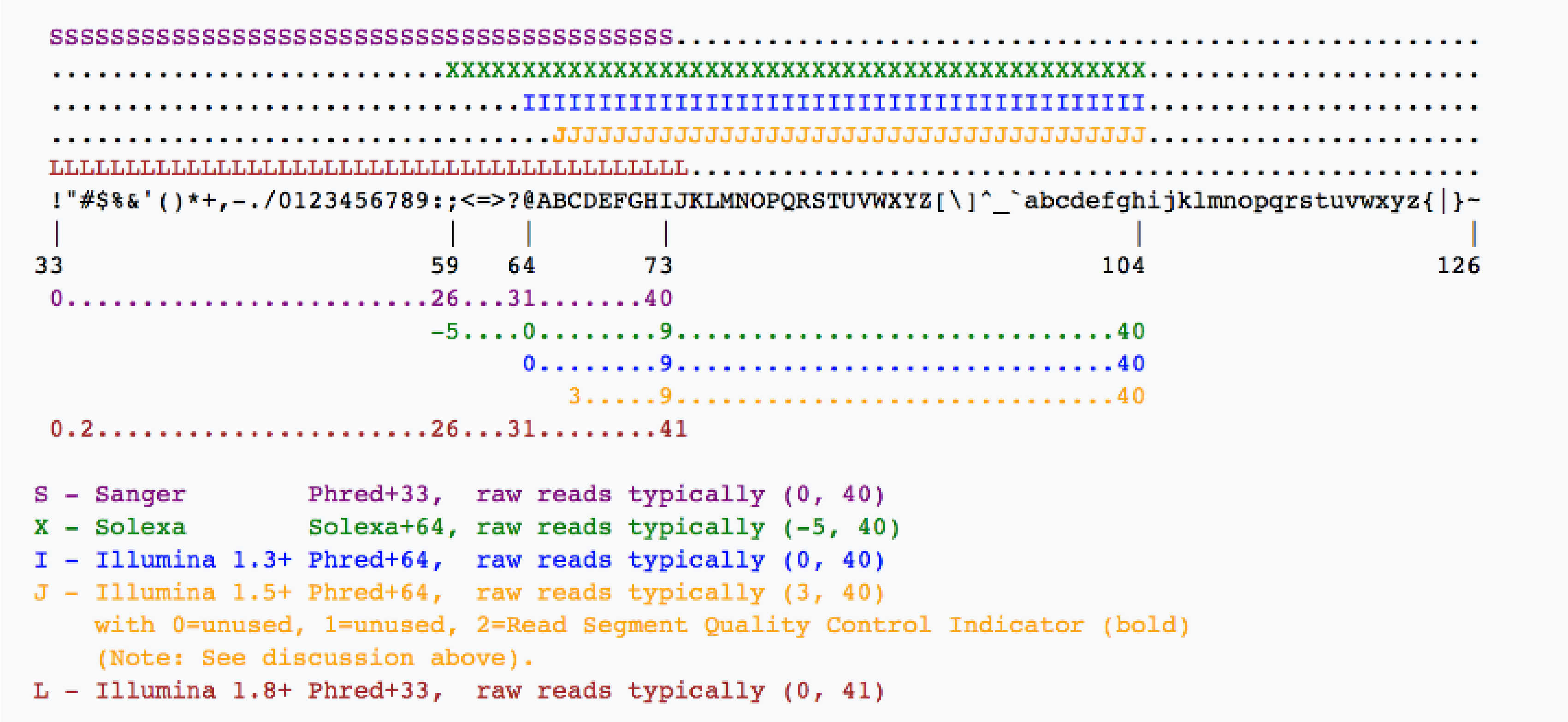


Base 64 (Old, rare)

Decimal	Hexadecimal	Binary	Octal	Char
33	21	100001	41	!
34	22	100010	42	
35	23	100011	43	
36	24	100100	44	$
37	25	100101	45	%
38	26	100110	46	&
39	27	100111	47	'
40	28	101000	50	(
41	29	101001	51	)
42	2A	101010	52	*
43	2B	101011	53	+
44	2C	101100	54	,
45	2D	101101	55	-
46	2E	101110	56	.
47	2F	101111	57	/
48	30	110000	60	0
49	31	110001	61	1
50	32	110010	62	2
51	33	110011	63	3
52	34	110100	64	4
53	35	110101	65	5
54	36	110110	66	6
55	37	110111	67	7
56	38	111000	70	8
57	39	111001	71	9
58	3A	111010	72	:
59	3B	111011	73	;
60	3C	111100	74	<
61	3D	111101	75	=
62	3E	111110	76	>
63	3F	111111	77	?
64	40	1000000	100	@
65	41	1000001	101	A
66	42	1000010	102	B
67	43	1000011	103	C
68	44	1000100	104	D
69	45	1000101	105	E
70	46	1000110	106	F
71	47	1000111	107	G
72	48	1001000	110	H
73	49	1001001	111	I
74	4A	1001010	112	J
75	4B	1001011	113	K
76	4C	1001100	114	L
77	4D	1001101	115	M
78	4E	1001110	116	N
79	4F	1001111	117	O
80	50	1010000	120	P
81	51	1010001	121	Q
82	52	1010010	122	R
83	53	1010011	123	S
84	54	1010100	124	T
85	55	1010101	125	U
86	56	1010110	126	V
87	57	1010111	127	W
88	58	1011000	130	X
89	59	1011001	131	Y
90	5A	1011010	132	Z
91	5B	1011011	133	[
92	5C	1011100	134	\
93	5D	1011101	135	]
94	5E	1011110	136	^
95	5F	1011111	137	_
96	60	1100000	140	`
97	61	1100001	141	a
98	62	1100010	142	b
99	63	1100011	143	c
100	64	1100100	144	d
101	65	1100101	145	e
102	66	1100110	146	f
103	67	1100111	147	g
104	68	1101000	150	h
105	69	1101001	151	i
106	6A	1101010	152	j
107	6B	1101011	153	k
108	6C	1101100	154	l
109	6D	1101101	155	m
110	6E	1101110	156	n
111	6F	1101111	157	o
112	70	1110000	160	p
113	71	1110001	161	q
114	72	1110010	162	r
115	73	1110011	163	s
116	74	1110100	164	t
117	75	1110101	165	u
118	76	1110110	166	v
119	77	1110111	167	w
120	78	1111000	170	x
121	79	1111001	171	y
122	7A	1111010	172	z
123	7B	1111011	173	{
124	7C	1111100	174	|
125	7D	1111101	175	}
126	7E	1111110	176	~
127	7F	1111111	177	[DEL]

Phred to ASCII

- Encoding in different platforms

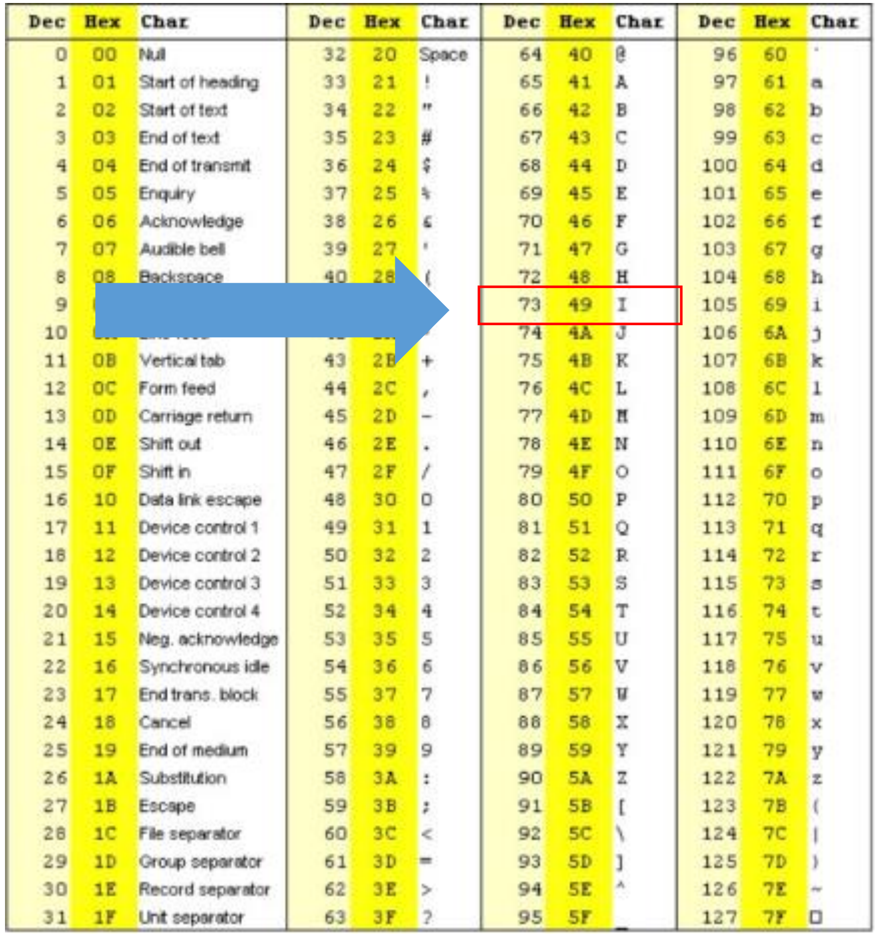


```
 SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................................
 ..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX......................
 ...............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII......................
 .................................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ......................
 LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL....................................................
 !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
 |                         |    |        |                              |                     |
33                        59   64       73                            104                   126
 0.........................26...31.......40
                          -5....0........9.............................40
                                0........9.............................40
                                   3.....9.............................40
 0.2.......................26...31........41

S - Sanger        Phred+33,  raw reads typically (0, 40)
X - Solexa        Solexa+64, raw reads typically (-5, 40)
I - Illumina 1.3+ Phred+64,  raw reads typically (0, 40)
J - Illumina 1.5+ Phred+64,  raw reads typically (3, 40)
    with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)
    (Note: See discussion above).
L - Illumina 1.8+ Phred+33,  raw reads typically (0, 41)
```

Phred to ASCII

- Illumina 1.9 uses **ASCII-33**, i.e. Illumina **quality score of 40** becomes

- 40 + 33 = **73** : ”**I**”



Dec	Hex	Char	Dec	Hex	Char	Dec	Hex	Char	Dec	Hex	Char
0	00	Null	32	20	Space	64	40	@	96	60	`
1	01	Start of heading	33	21	!	65	41	A	97	61	a
2	02	Start of text	34	22	"	66	42	B	98	62	b
3	03	End of text	35	23	#	67	43	C	99	63	c
4	04	End of transmit	36	24	$	68	44	D	100	64	d
5	05	Enquiry	37	25	%	69	45	E	101	65	e
6	06	Acknowledge	38	26	&	70	46	F	102	66	f
7	07	Audible bell	39	27	'	71	47	G	103	67	g
8	08	Backspace	40	28	(	72	48	H	104	68	h
9	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	73	49	I	105	69	i
10	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	74	4A	J	106	6A	j
11	0B	Vertical tab	43	2B	+	75	4B	K	107	6B	k
12	0C	Form feed	44	2C	,	76	4C	L	108	6C	l
13	0D	Carriage return	45	2D	-	77	4D	M	109	6D	m
14	0E	Shift out	46	2E	.	78	4E	N	110	6E	n
15	0F	Shift in	47	2F	/	79	4F	O	111	6F	o
16	10	Data link escape	48	30	0	80	50	P	112	70	p
17	11	Device control 1	49	31	1	81	51	Q	113	71	q
18	12	Device control 2	50	32	2	82	52	R	114	72	r
19	13	Device control 3	51	33	3	83	53	S	115	73	s
20	14	Device control 4	52	34	4	84	54	T	116	74	t
21	15	Neg. acknowledge	53	35	5	85	55	U	117	75	u
22	16	Synchronous idle	54	36	6	86	56	V	118	76	v
23	17	End trans. block	55	37	7	87	57	W	119	77	w
24	18	Cancel	56	38	8	88	58	X	120	78	x
25	19	End of medium	57	39	9	89	59	Y	121	79	y
26	1A	Substitution	58	3A	:	90	5A	Z	122	7A	z
27	1B	Escape	59	3B	;	91	5B	[	123	7B	{
28	1C	File separator	60	3C	<	92	5C	\	124	7C	|
29	1D	Group separator	61	3D	=	93	5D	]	125	7D	}
30	1E	Record separator	62	3E	>	94	5E	^	126	7E	~
31	1F	Unit separator	63	3F	?	95	5F	_	127	7F	□

FASTQ format

- Each read is **4** lines
- Read starts with a character @ followed by the read descriptor
- Sequence follows in the second line
- Third line is reserved for additional info
- Fourth line is the Phred score encoding

- Read pairs are typically in different files

```
@GWNJ-0850:627:GW190820000:5:1101:12033:1450 1:N:0:NCTCCTGA+NGGCTATA
CTTTTTCCTCGAGTATCTTTTGGAGGCGATTCTTTTTTTGAACTTGCTTTTTTTTTTGAGATCTACACGGTAGATTCAA
+
?@;DDDDDHDC<D<AEEHIGIII+<B@F?@FFGEHGIIII(77.7=7=AEHBBBB?=B8823(>A>(985++5(:@AC4
```

FASTQ format

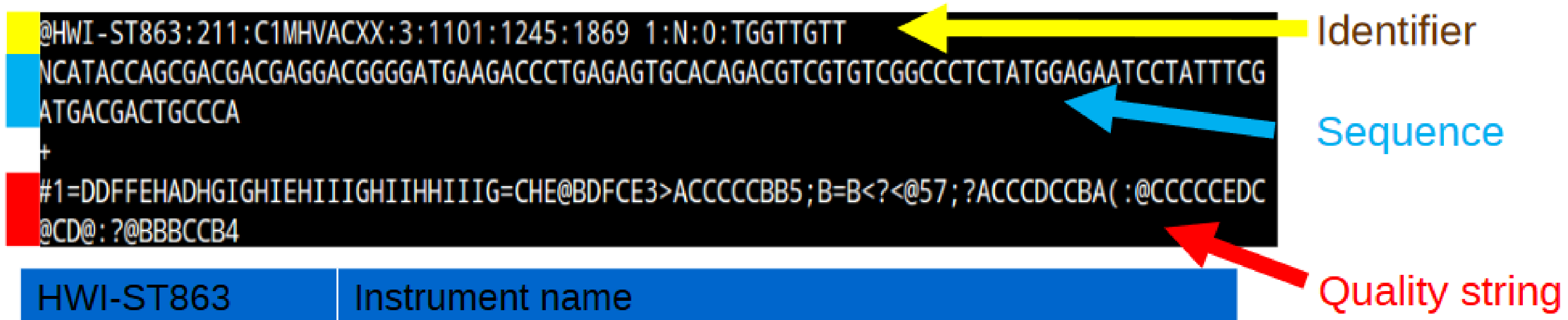
- Store calls (ATGC …)
- Store Phred scores (Encoded)

- Store Machine make/ID
- Store Flowcell id for each spot
- Store coordinates of the spot
- Store additional info (Seq names ….)
- Easily parsed and stored.

```
@GWNJ-0850:627:GW190820000:5:1101:12033:1450 1:N:0:NCTCCTGA+NGGCTATA
CTTTTTCCTCGAGTATCTTTTGGAGGCGATTCTTTTTTTGAACTTGCTTTTTTTTTTGAGATCTACACGGTAGATTCAA
+
?@;DDDDDHDC<D<AEEHIGIII+<B@F?@FFGEHGIIII(77.7=7=AEHBBBB?=B8823(>A>(985++5(:@AC4
```

FASTQ format

- Header

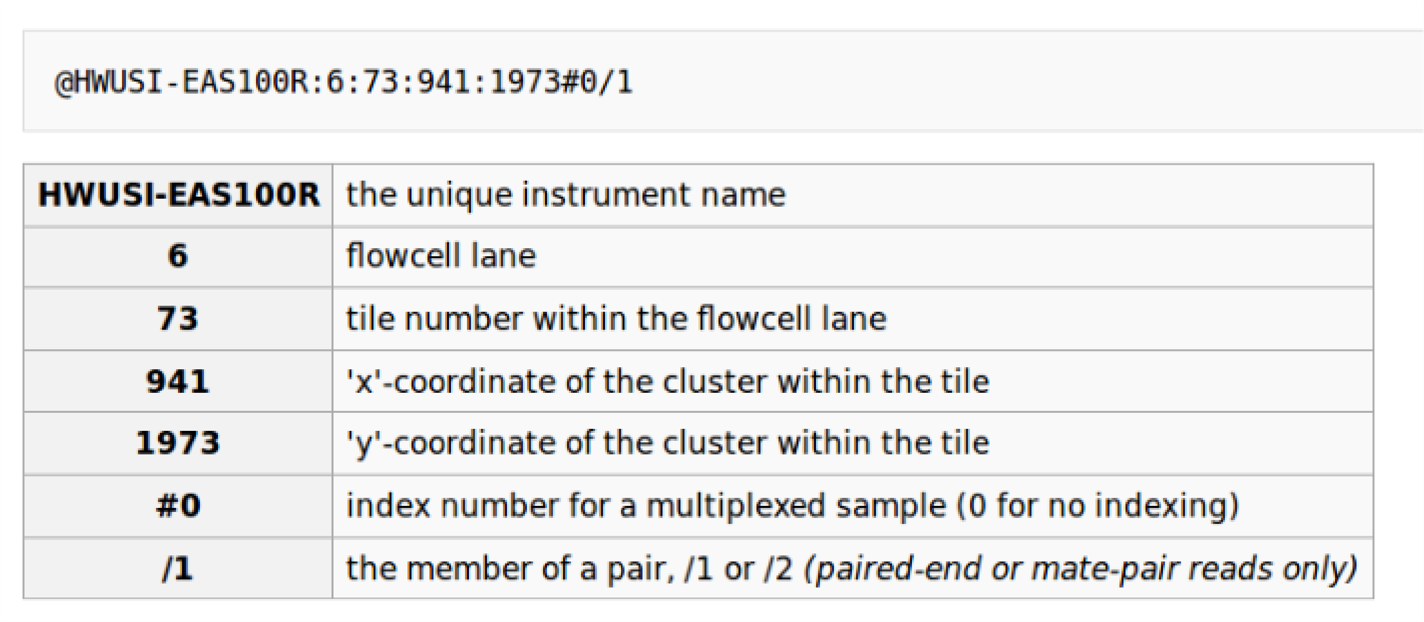


```
@HWI-ST863:211:C1MHVACXX:3:1101:1245:1869 1:N:0:TGGTTGTT
NCATACCAGCGACGACGAGGACGGGGATGAAGACCCTGAGAGTGCACAGACGTCGTGTCGGCCCTCTATGGAGAATCCTATTTCG
ATGACGACTGCCCA
+
#1=DDFFEHADHGIGHIEHIIIGHIIHHIIIG=CHE@BDFCE3>ACCCCCBB5;B=B<?<@57;?ACCCDCCBA(:@CCCCCEDC
@CD@:?@BBBCCB4
```

HWI-ST863	Instrument name
211	Run id
C1MHVACXX	Flowcell id
3	Flowcell lane
1101	Tile number of flowcell lane
1245	'x'-coordinate of the cluster within the tile
1869	'y'-coordinate of the cluster within the tile
1	the member of a pair, 1 or 2
N	Y if the read is filtered, N otherwise
0	0 when none of the control bits are on, otherwise it is an even number
TGGTTGTT	Index sequence

FASTQ format

- Old header format



```
@HWUSI-EAS100R:6:73:941:1973#0/1
```

HWUSI-EAS100R	the unique instrument name
6	flowcell lane
73	tile number within the flowcell lane
941	'x'-coordinate of the cluster within the tile
1973	'y'-coordinate of the cluster within the tile
#0	index number for a multiplexed sample (0 for no indexing)
/1	the member of a pair, /1 or /2 *(paired-end or mate-pair reads only)*

Source: Wikipedia

SAM format

- **SAM** : Sequence Alignment/Map
- **BAM** : Binary Alignment/Map (binary SAM)

- Used for: aligned reads
- Multiple *tab* delimited columns
- It is flexible enough to store **all the alignment information** generated by various alignment programs
- It allows most of the operations on the alignment to work on a stream without loading the whole alignment into memory
- It allows the file to be **indexed by genomic position** to efficiently retrieve all reads aligning to a locus

SAM format

Two sections

- Header section, each line begins with “@”–Several record types

  5 fields

- Alignment section

  11 mandatory fields (columns)

SAM format

- **Header section**

- @HD The header line. The first line if present.

  VN* Format version

- @SQ Reference sequence dictionary. The order of @SQ lines defines the alignment sorting order

  SN* Reference sequence name

  LN* Reference sequence length

- @RG Read group. Unordered multiple @RG lines are allowed.

  ID* Read group identifier

- @PG Program

  ID* Program record identifier

- @CO One-line text comment

SAM format

- Header section + Alignment section (first line)

```
@HD     VN:1.5  GO:none SO:coordinate
@SQ     SN:Spenn-ch01   LN:109333515
@SQ     SN:Spenn-ch02   LN:59803892
@SQ     SN:Spenn-ch03   LN:75414019
@SQ     SN:Spenn-ch04   LN:77197300
@SQ     SN:Spenn-ch05   LN:77991103
@SQ     SN:Spenn-ch06   LN:60730942
@RG     ID:LA2932_28.CA1PNANXX.3.CAGATC PU:CA1PNANXX.3.CAGATC   LB:LA2932_28
SM:LA2932_28    PL:ILLUMINA
@PG     ID:MarkDuplicates       VN:1.119 CL:picard.sam.MarkDuplicates …
@PG     ID:bwa  VN:0.7.12-r1039 CLbwa mem -M -t 24 Spenn2.fa …
@PG     ID:GATK IndelRealigner  VN:3.7-0-gcfedb67       …
HISEQ:202:CA1PNANXX:4:1311:19476:42830  163     Spenn-ch01      13      0
4S99M23S        =       124     228
TTATGGCCAACCGGATGCATAGACAAGGTCTTGACGGACGTCCACAAAAAAATTTGCCATTTTTGATGTCGGAATCCGG
ATCACCCAGAAAATGGTTTGCTATGTCACACGGAAATCGTTAAAATG
BBBBB<FFFFFFFFBFFFFFFFFFBFFFF<F<FFFF</FFBBFFFFFFFFFFFFBF/BFF/<F/FF<</<<FBBB/BFFFF
FFFFFFFFFFB<FFFBFFFFFFFFF<7/7//<7/7F//777/BFFFFB  MC:Z:117M4S
MD:Z:46G1G34A15 PG:Z:MarkDuplicates     RG:Z:LA2932_28.CA1PNANXX.3.CAGATC
NM:i:3  MQ:i:0  AS:i:84 XS:i:90
…
```

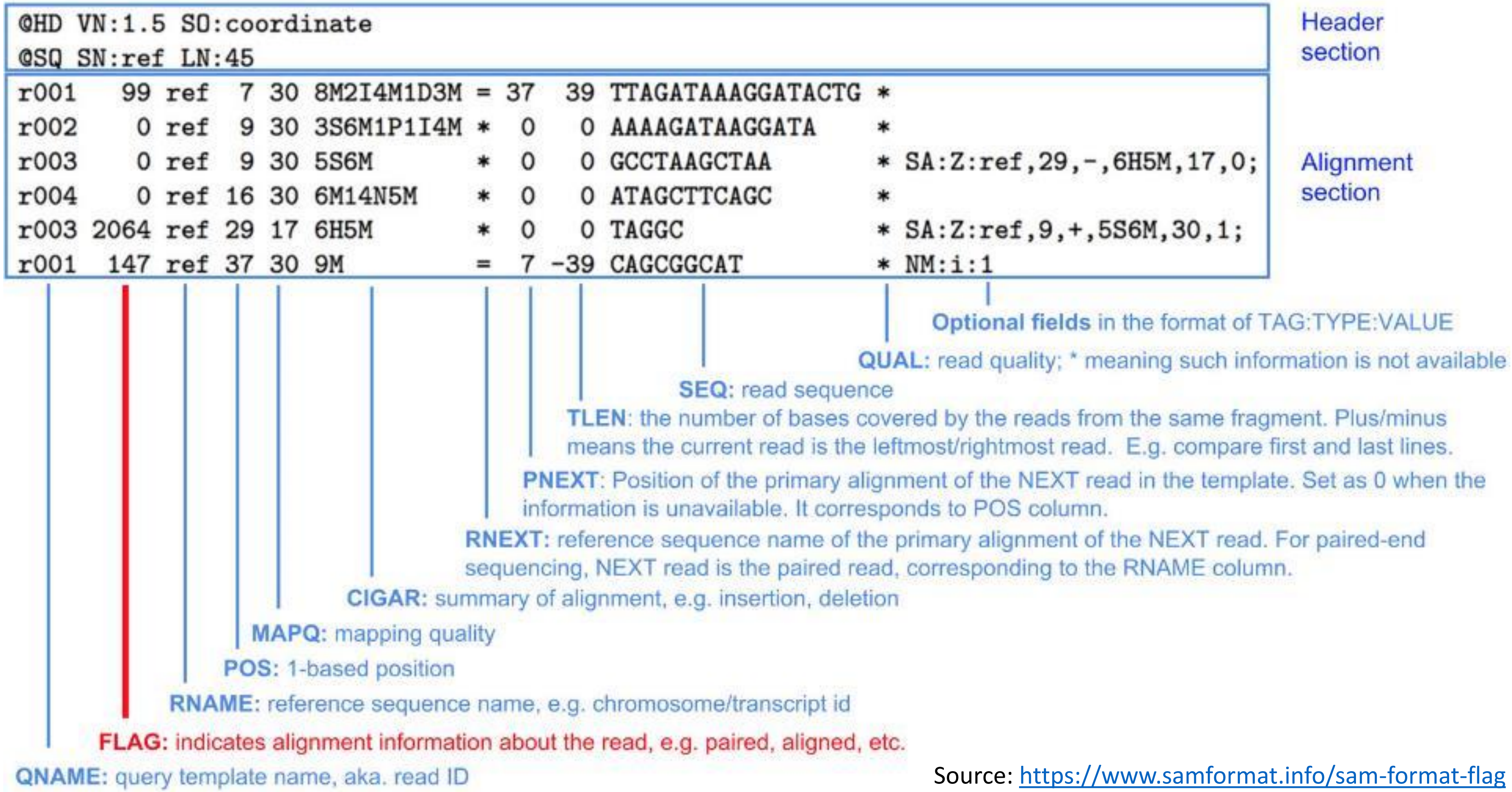
Source: Wikipedia

SAM format

- Alignment section

HISEQ:202:CA1PNANXX:4:1311:19476:42830	Query name: shared by pair-end mates
163	Flag value: Decimal > Binary > Multiple True/False values
Spenn-ch01	Chromosome/Contig where the read aligned
13	Position on chromosome/contig where the read aligned
0	Alignment confidence (Phred)
4S99M23S	CIGAR string
=	Chomosome/Contig where mate aligned (= if same)
124	Position on chromosome/contig where the mate aligned
228	Lenght of the reference sequence read aligned to
TTATGGCCAACCGGATGCATAGACAAGGTCTTGACG…	Read sequence
BBBBB<FFFFFFFFBFFFFFFFFFBFFFF<F<FFFF<…	Read quality score (same as fastq file)
MC:Z: MD:Z: PG:Z: RG:Z: NM:i:3 MQ:i:0 AS:i:84 XS:i:90	Optional tags: http://samtools.github.io/hts-specs/SAMtags.pdf

SAM format



```
@HD VN:1.5 SO:coordinate
@SQ SN:ref LN:45
r001    99 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG *
r002     0 ref  9 30 3S6M1P1I4M *  0   0 AAAAGATAAGGATA    *
r003     0 ref  9 30 5S6M       *  0   0 GCCTAAGCTAA       * SA:Z:ref,29,-,6H5M,17,0;
r004     0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCAGC       *
r003  2064 ref 29 17 6H5M       *  0   0 TAGGC             * SA:Z:ref,9,+,5S6M,30,1;
r001   147 ref 37 30 9M         =  7 -39 CAGCGGCAT         * NM:i:1
```

Optional fields in the format of TAG:TYPE:VALUE

QUAL: read quality; * meaning such information is not available

SEQ: read sequence

TLEN: the number of bases covered by the reads from the same fragment. Plus/minus means the current read is the leftmost/rightmost read. E.g. compare first and last lines.

PNEXT: Position of the primary alignment of the NEXT read in the template. Set as 0 when the information is unavailable. It corresponds to POS column.

RNEXT: reference sequence name of the primary alignment of the NEXT read. For paired-end sequencing, NEXT read is the paired read, corresponding to the RNAME column.

CIGAR: summary of alignment, e.g. insertion, deletion

MAPQ: mapping quality

POS: 1-based position

RNAME: reference sequence name, e.g. chromosome/transcript id

FLAG: indicates alignment information about the read, e.g. paired, aligned, etc.

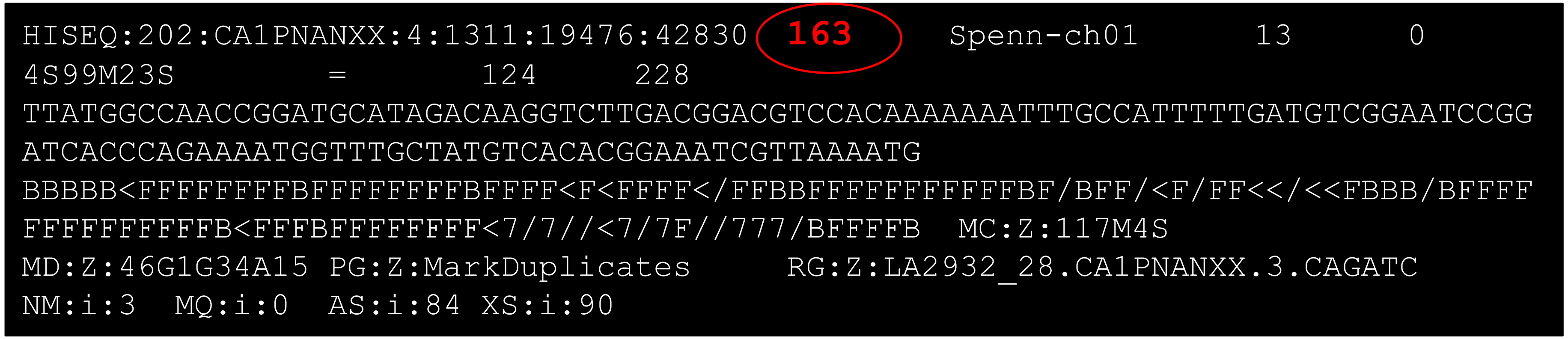
QNAME: query template name, aka. read ID

Source: https://www.samformat.info/sam-format-flag

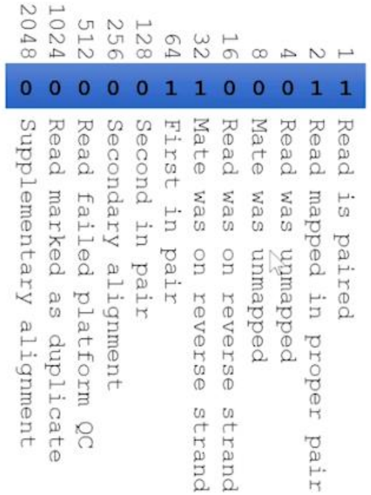
More details: https://samtools.github.io/hts-specs/SAMv1.pdf

SAM format

- bitwise flag



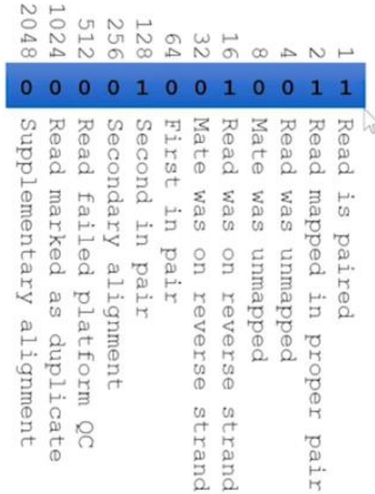
```
HISEQ:202:CA1PNANXX:4:1311:19476:42830 163     Spenn-ch01      13       0
4S99M23S         =         124      228
TTATGGCCAACCGGATGCATAGACAAGGTCTTGACGGACGTCCACAAAAAAATTTGCCATTTTTGATGTCGGAATCCGG
ATCACCCAGAAAATGGTTTGCTATGTCACACGGAAATCGTTAAAATG
BBBBB<FFFFFFFFBFFFFFFFFFBFFFF<F<FFFF</FFBBFFFFFFFFFFFFBF/BFF/<F/FF<</<<FBBB/BFFFF
FFFFFFFFFFB<FFFBFFFFFFFFF<7/7//<7/7F//777/BFFFFB  MC:Z:117M4S
MD:Z:46G1G34A15 PG:Z:MarkDuplicates     RG:Z:LA2932_28.CA1PNANXX.3.CAGATC
NM:i:3   MQ:i:0   AS:i:84 XS:i:90
```



2048	0	Supplementary alignment
1024	0	Read marked as duplicate
512	0	Read failed platform QC
256	0	Secondary alignment
128	0	Second in pair
64	1	First in pair
32	1	Mate was on reverse strand
16	0	Read was on reverse strand
8	0	Mate was unmapped
4	0	Read was unmapped
2	1	Read mapped in proper pair
1	1	Read is paired

1 + 2 + 32 + 64 =

99



2048	0	Supplementary alignment
1024	0	Read marked as duplicate
512	0	Read failed platform QC
256	0	Secondary alignment
128	1	Second in pair
64	0	First in pair
32	0	Mate was on reverse strand
16	1	Read was on reverse strand
8	0	Mate was unmapped
4	0	Read was unmapped
2	1	Read mapped in proper pair
1	1	Read is paired

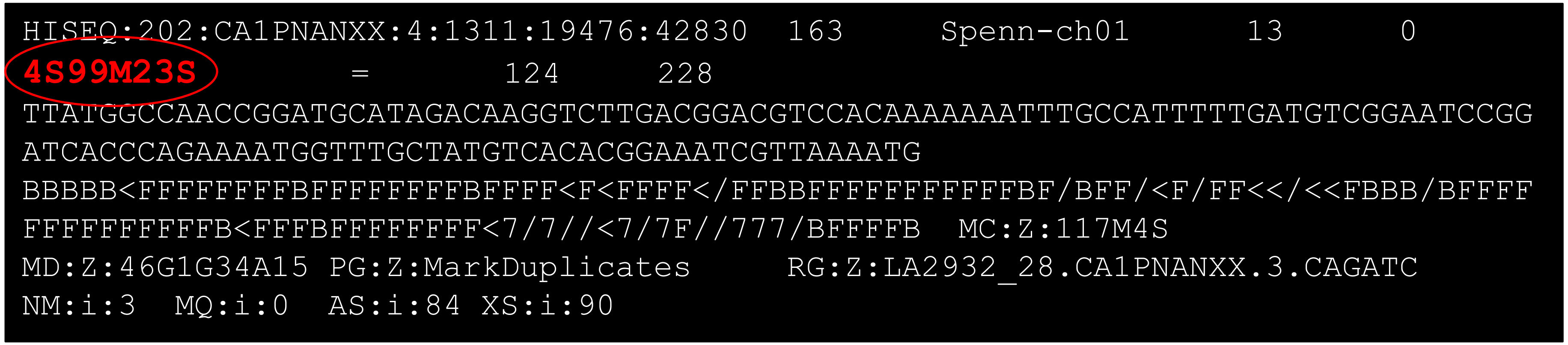
1 + 2 + 16 + 128 =

147

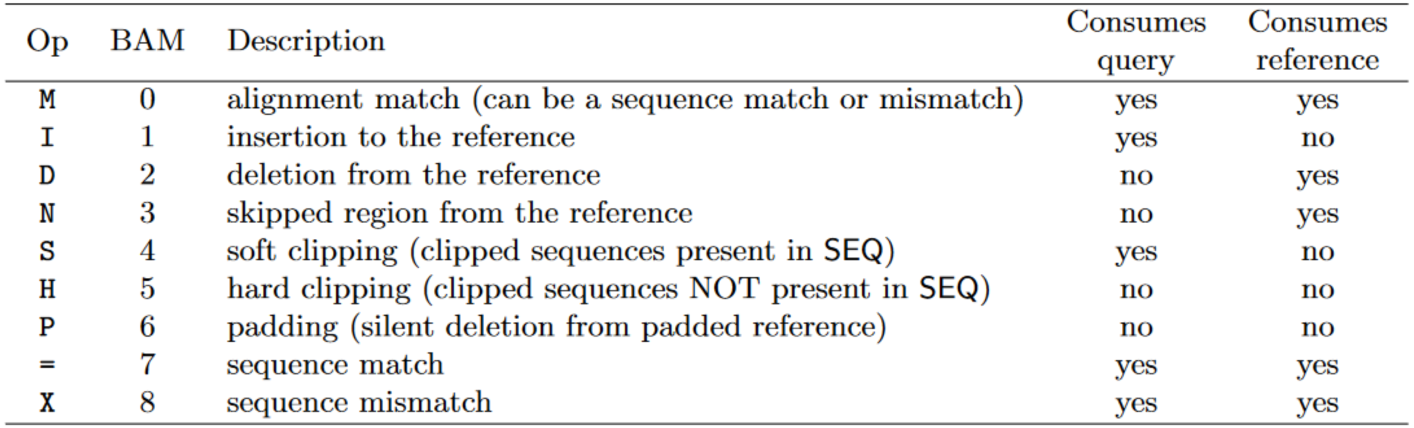
Decoding SAM flags https://broadinstitute.github.io/picard/explain-flags.html

SAM format

- CIGAR string – Describes how the read align to the reference



```
HISEQ:202:CA1PNANXX:4:1311:19476:42830  163     Spenn-ch01      13       0
4S99M23S          =        124      228
TTATGGCCAACCGGATGCATAGACAAGGTCTTGACGGACGTCCACAAAAAAATTTGCCATTTTTGATGTCGGAATCCGG
ATCACCCAGAAAATGGTTTGCTATGTCACACGGAAATCGTTAAAATG
BBBBB<FFFFFFFFBFFFFFFFFFBFFFF<F<FFFF</FFBBFFFFFFFFFFFFBF/BFF/<F/FF<</<<FBBB/BFFFF
FFFFFFFFFFB<FFFBFFFFFFFF<7/7//<7/7F//777/BFFFFB  MC:Z:117M4S
MD:Z:46G1G34A15 PG:Z:MarkDuplicates     RG:Z:LA2932_28.CA1PNANXX.3.CAGATC
NM:i:3  MQ:i:0  AS:i:84 XS:i:90
```



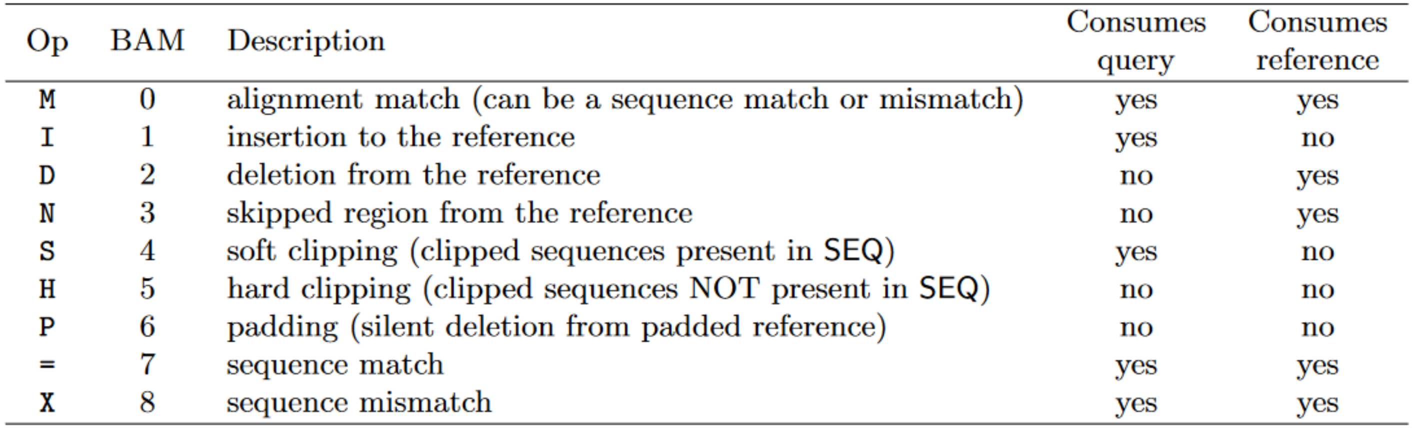
Op	BAM	Description	Consumes query	Consumes reference
M	0	alignment match (can be a sequence match or mismatch)	yes	yes
I	1	insertion to the reference	yes	no
D	2	deletion from the reference	no	yes
N	3	skipped region from the reference	no	yes
S	4	soft clipping (clipped sequences present in SEQ)	yes	no
H	5	hard clipping (clipped sequences NOT present in SEQ)	no	no
P	6	padding (silent deletion from padded reference)	no	no
=	7	sequence match	yes	yes
X	8	sequence mismatch	yes	yes

Source: https://samtools.github.io/hts-specs/SAMv1.pdf

SAM format

- CIGAR string – Describes how the read align to the reference

```
Reference: ATGAAGGATAGTGATACTCTAGAGGG
Read: ACGAATAGTGATACTCGGGTAGAGGG
```



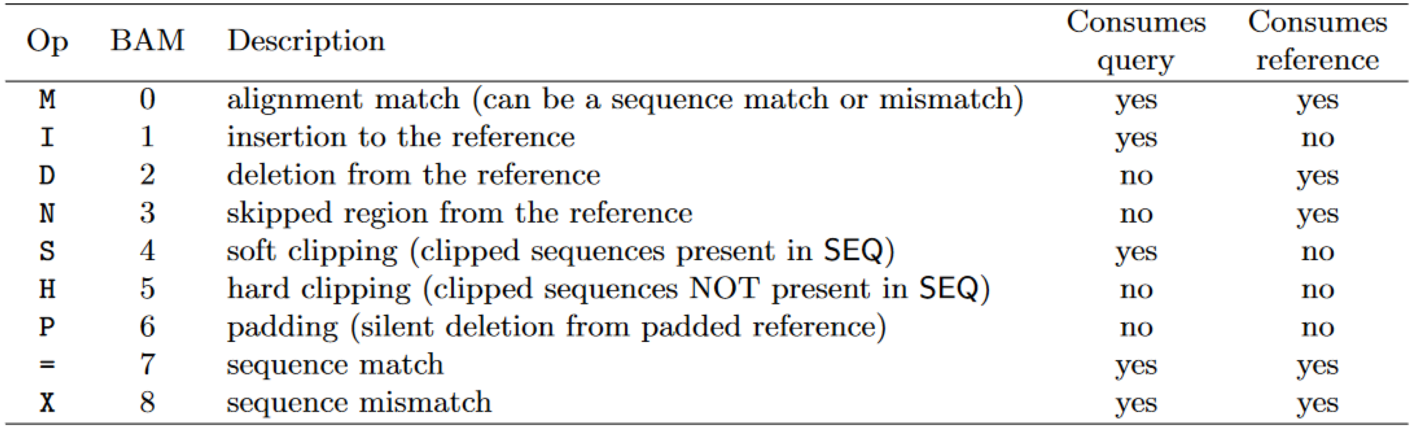
Op	BAM	Description	Consumes query	Consumes reference
M	0	alignment match (can be a sequence match or mismatch)	yes	yes
I	1	insertion to the reference	yes	no
D	2	deletion from the reference	no	yes
N	3	skipped region from the reference	no	yes
S	4	soft clipping (clipped sequences present in SEQ)	yes	no
H	5	hard clipping (clipped sequences NOT present in SEQ)	no	no
P	6	padding (silent deletion from padded reference)	no	no
=	7	sequence match	yes	yes
X	8	sequence mismatch	yes	yes

Source: https://samtools.github.io/hts-specs/SAMv1.pdf

SAM format

- CIGAR string – Describes how the read align to the reference

```
Reference:   ATGAAGGATAGTGATACTC---TAGAGGG
Read:        ACGAA---TAGTGATACTCGGGTAGAGGG
CIGAR:       5M3D11M3I7M
```

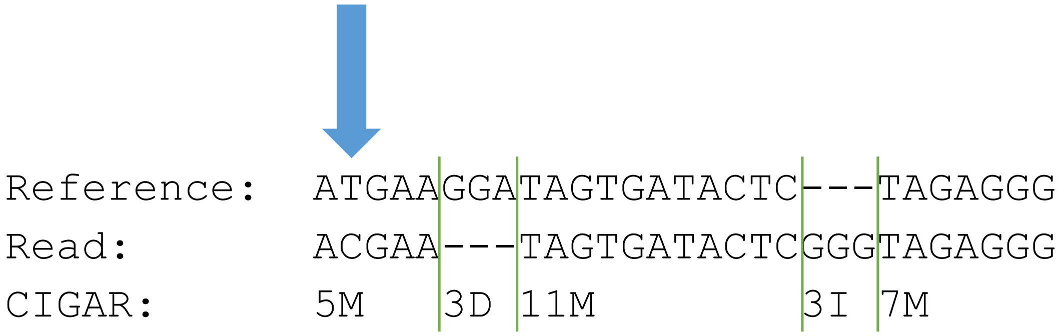


Op	BAM	Description	Consumes query	Consumes reference
M	0	alignment match (can be a sequence match or mismatch)	yes	yes
I	1	insertion to the reference	yes	no
D	2	deletion from the reference	no	yes
N	3	skipped region from the reference	no	yes
S	4	soft clipping (clipped sequences present in SEQ)	yes	no
H	5	hard clipping (clipped sequences NOT present in SEQ)	no	no
P	6	padding (silent deletion from padded reference)	no	no
=	7	sequence match	yes	yes
X	8	sequence mismatch	yes	yes

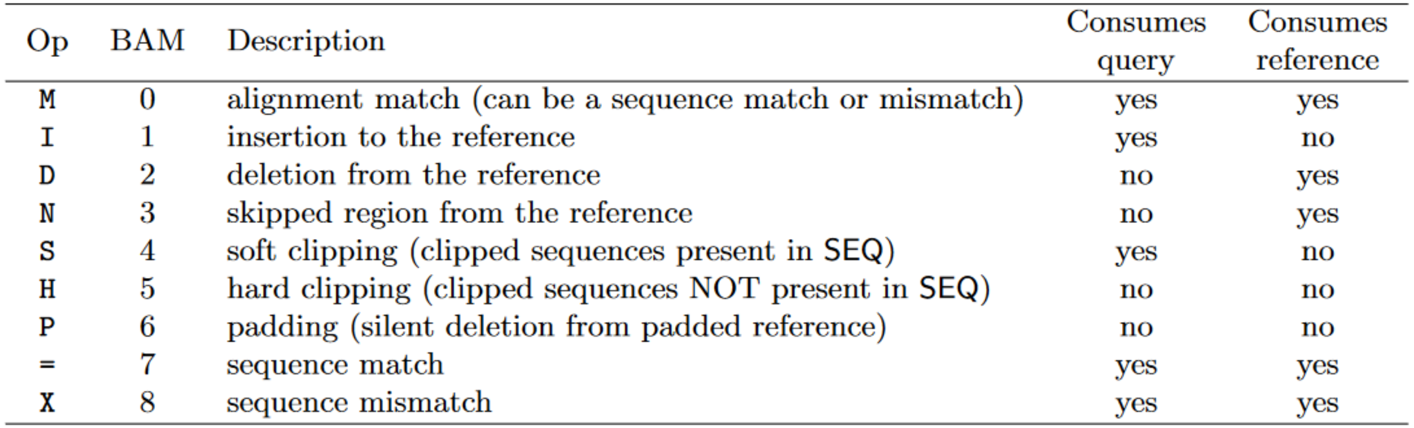
Source: https://samtools.github.io/hts-specs/SAMv1.pdf

SAM format

- CIGAR string – Describes how the read align to the reference



```
Reference:   ATGAAGGATAGTGATACTC---TAGAGGG
Read:        ACGAA---TAGTGATACTCGGGTAGAGGG
CIGAR:       5M   3D 11M        3I 7M
```

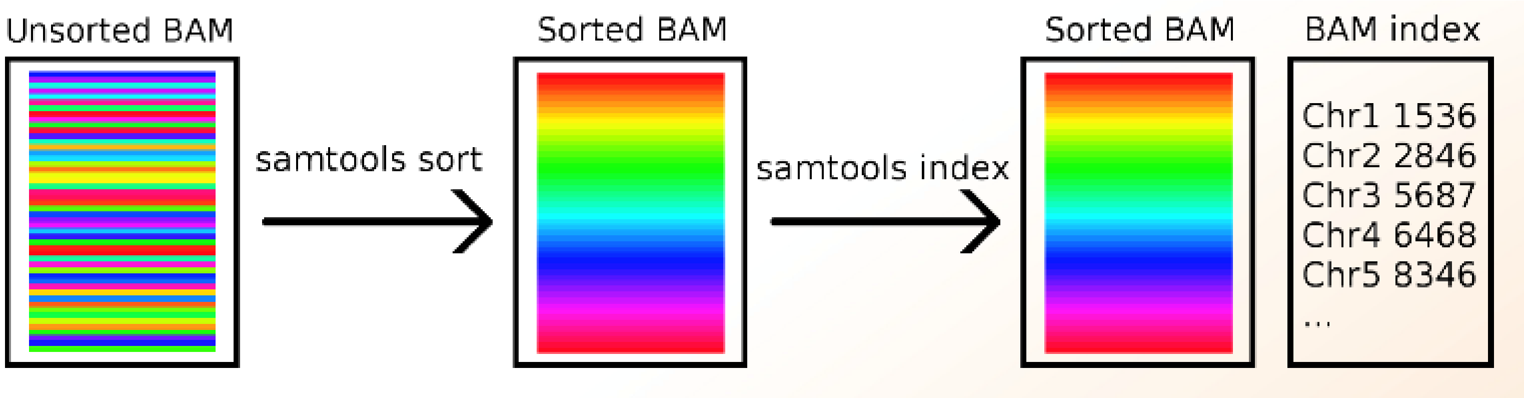


Op	BAM	Description	Consumes query	Consumes reference
M	0	alignment match (can be a sequence match or mismatch)	yes	yes
I	1	insertion to the reference	yes	no
D	2	deletion from the reference	no	yes
N	3	skipped region from the reference	no	yes
S	4	soft clipping (clipped sequences present in SEQ)	yes	no
H	5	hard clipping (clipped sequences NOT present in SEQ)	no	no
P	6	padding (silent deletion from padded reference)	no	no
=	7	sequence match	yes	yes
X	8	sequence mismatch	yes	yes

Source: https://samtools.github.io/hts-specs/SAMv1.pdf

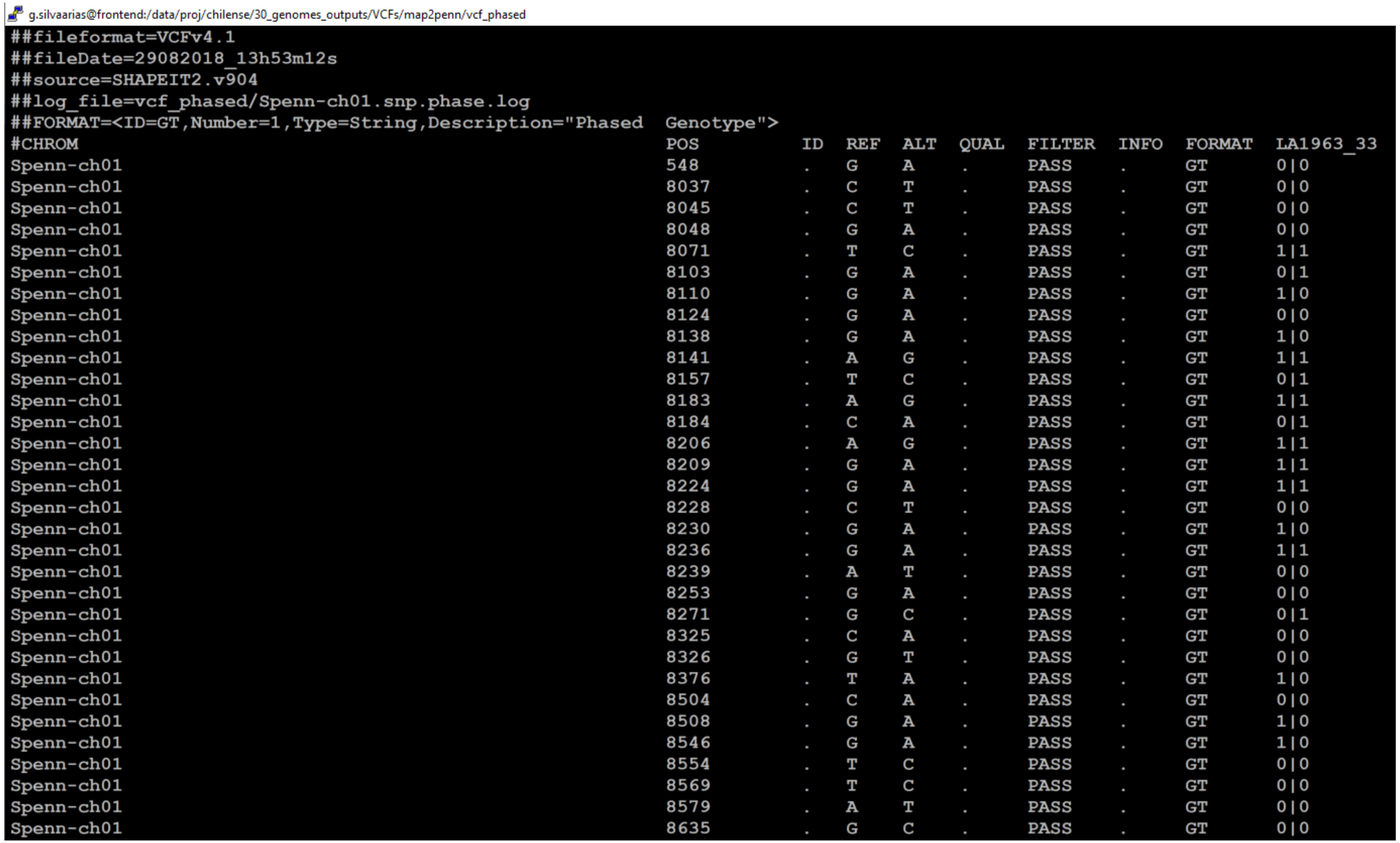
BAM format

- Binary SAM
- Used for: aligned reads
- **25% of the size**
- **SAMtools** to convert
- **.bai** = BAM index



VCF format

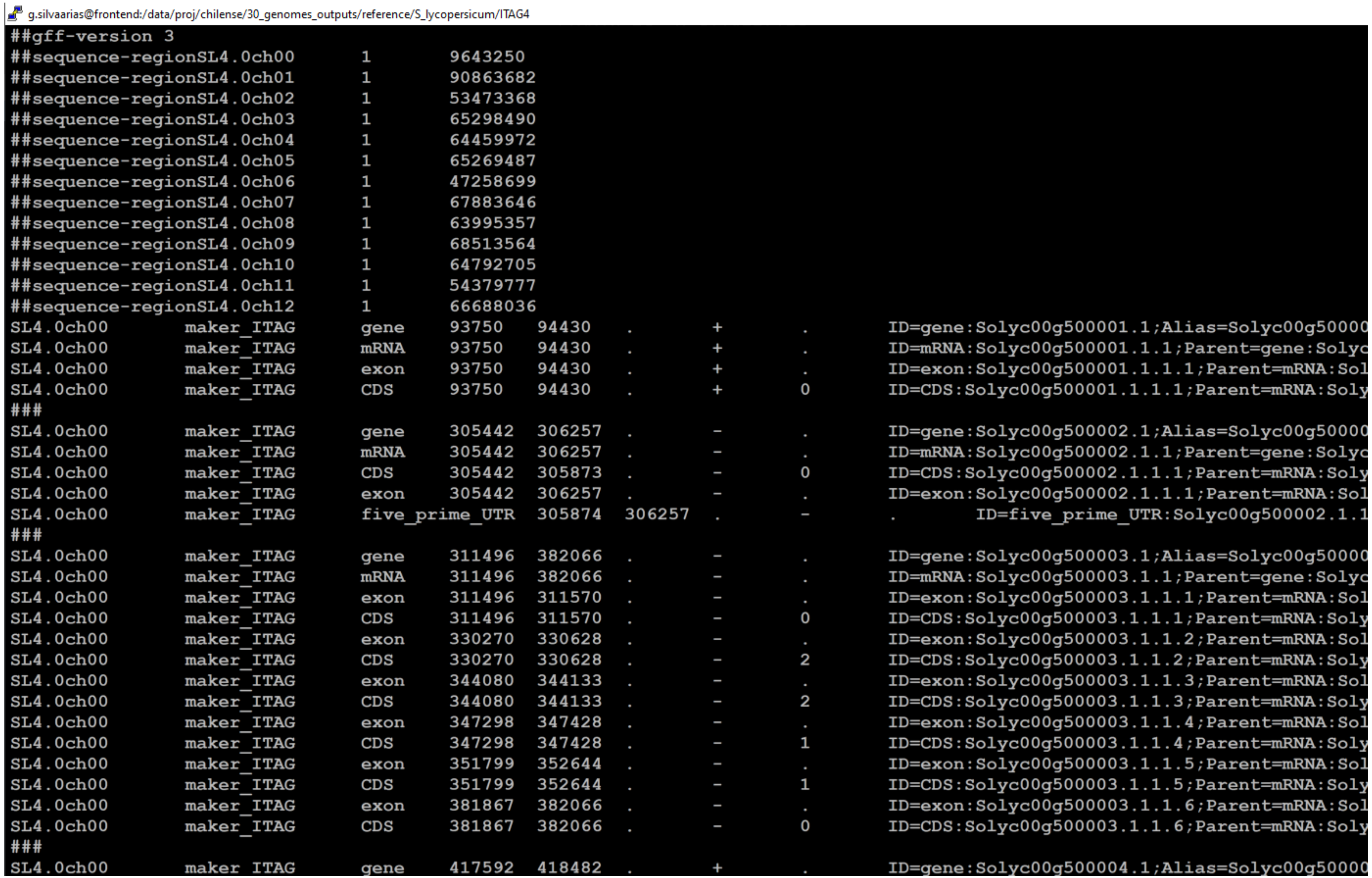
- Variant call format


g.silvaarias@frontend:/data/proj/chilense/30_genomes_outputs/VCFs/map2penn/vcf_phased

```
##fileformat=VCFv4.1
##fileDate=29082018_13h53m12s
##source=SHAPEIT2.v904
##log_file=vcf_phased/Spenn-ch01.snp.phase.log
##FORMAT=<ID=GT,Number=1,Type=String,Description="Phased Genotype">
#CHROM                                                    POS          ID  REF  ALT  QUAL  FILTER  INFO  FORMAT  LA1963_33
Spenn-ch01                                                548          .   G    A    .     PASS    .     GT      0|0
Spenn-ch01                                                8037         .   C    T    .     PASS    .     GT      0|0
Spenn-ch01                                                8045         .   C    T    .     PASS    .     GT      0|0
Spenn-ch01                                                8048         .   G    A    .     PASS    .     GT      0|0
Spenn-ch01                                                8071         .   T    C    .     PASS    .     GT      1|1
Spenn-ch01                                                8103         .   G    A    .     PASS    .     GT      0|1
Spenn-ch01                                                8110         .   G    A    .     PASS    .     GT      1|0
Spenn-ch01                                                8124         .   G    A    .     PASS    .     GT      0|0
Spenn-ch01                                                8138         .   G    A    .     PASS    .     GT      1|0
Spenn-ch01                                                8141         .   A    G    .     PASS    .     GT      1|1
Spenn-ch01                                                8157         .   T    C    .     PASS    .     GT      0|1
Spenn-ch01                                                8183         .   A    G    .     PASS    .     GT      1|1
Spenn-ch01                                                8184         .   C    A    .     PASS    .     GT      0|1
Spenn-ch01                                                8206         .   A    G    .     PASS    .     GT      1|1
Spenn-ch01                                                8209         .   G    A    .     PASS    .     GT      1|1
Spenn-ch01                                                8224         .   G    A    .     PASS    .     GT      1|1
Spenn-ch01                                                8228         .   C    T    .     PASS    .     GT      0|0
Spenn-ch01                                                8230         .   G    A    .     PASS    .     GT      1|0
Spenn-ch01                                                8236         .   G    A    .     PASS    .     GT      1|1
Spenn-ch01                                                8239         .   A    T    .     PASS    .     GT      0|0
Spenn-ch01                                                8253         .   G    A    .     PASS    .     GT      0|0
Spenn-ch01                                                8271         .   G    C    .     PASS    .     GT      0|1
Spenn-ch01                                                8325         .   C    A    .     PASS    .     GT      0|0
Spenn-ch01                                                8326         .   G    T    .     PASS    .     GT      0|0
Spenn-ch01                                                8376         .   T    A    .     PASS    .     GT      1|0
Spenn-ch01                                                8504         .   C    A    .     PASS    .     GT      0|0
Spenn-ch01                                                8508         .   G    A    .     PASS    .     GT      1|0
Spenn-ch01                                                8546         .   G    A    .     PASS    .     GT      1|0
Spenn-ch01                                                8554         .   T    C    .     PASS    .     GT      0|0
Spenn-ch01                                                8569         .   T    C    .     PASS    .     GT      0|0
Spenn-ch01                                                8579         .   A    T    .     PASS    .     GT      0|0
Spenn-ch01                                                8635         .   G    C    .     PASS    .     GT      0|0
```

GFF/GTF format

- Genome annotation



g.silvaarias@frontend:/data/proj/chilense/30_genomes_outputs/reference/S_lycopersicum/ITAG4

```
##gff-version 3
##sequence-regionSL4.0ch00      1       9643250
##sequence-regionSL4.0ch01      1       90863682
##sequence-regionSL4.0ch02      1       53473368
##sequence-regionSL4.0ch03      1       65298490
##sequence-regionSL4.0ch04      1       64459972
##sequence-regionSL4.0ch05      1       65269487
##sequence-regionSL4.0ch06      1       47258699
##sequence-regionSL4.0ch07      1       67883646
##sequence-regionSL4.0ch08      1       63995357
##sequence-regionSL4.0ch09      1       68513564
##sequence-regionSL4.0ch10      1       64792705
##sequence-regionSL4.0ch11      1       54379777
##sequence-regionSL4.0ch12      1       66688036
SL4.0ch00       maker_ITAG      gene    93750   94430   .       +       .       ID=gene:Solyc00g500001.1;Alias=Solyc00g50000
SL4.0ch00       maker_ITAG      mRNA    93750   94430   .       +       .       ID=mRNA:Solyc00g500001.1.1;Parent=gene:Solyc
SL4.0ch00       maker_ITAG      exon    93750   94430   .       +       .       ID=exon:Solyc00g500001.1.1.1;Parent=mRNA:Sol
SL4.0ch00       maker_ITAG      CDS     93750   94430   .       +       0       ID=CDS:Solyc00g500001.1.1.1;Parent=mRNA:Soly
###
SL4.0ch00       maker_ITAG      gene    305442  306257  .       -       .       ID=gene:Solyc00g500002.1;Alias=Solyc00g50000
SL4.0ch00       maker_ITAG      mRNA    305442  306257  .       -       .       ID=mRNA:Solyc00g500002.1.1;Parent=gene:Solyc
SL4.0ch00       maker_ITAG      CDS     305442  305873  .       -       0       ID=CDS:Solyc00g500002.1.1.1;Parent=mRNA:Soly
SL4.0ch00       maker_ITAG      exon    305442  306257  .       -       .       ID=exon:Solyc00g500002.1.1.1;Parent=mRNA:Sol
SL4.0ch00       maker_ITAG      five_prime_UTR  305874  306257  .       -       .       ID=five_prime_UTR:Solyc00g500002.1.1
###
SL4.0ch00       maker_ITAG      gene    311496  382066  .       -       .       ID=gene:Solyc00g500003.1;Alias=Solyc00g50000
SL4.0ch00       maker_ITAG      mRNA    311496  382066  .       -       .       ID=mRNA:Solyc00g500003.1.1;Parent=gene:Solyc
SL4.0ch00       maker_ITAG      exon    311496  311570  .       -       .       ID=exon:Solyc00g500003.1.1.1;Parent=mRNA:Sol
SL4.0ch00       maker_ITAG      CDS     311496  311570  .       -       0       ID=CDS:Solyc00g500003.1.1.1;Parent=mRNA:Soly
SL4.0ch00       maker_ITAG      exon    330270  330628  .       -       .       ID=exon:Solyc00g500003.1.1.2;Parent=mRNA:Sol
SL4.0ch00       maker_ITAG      CDS     330270  330628  .       -       2       ID=CDS:Solyc00g500003.1.1.2;Parent=mRNA:Soly
SL4.0ch00       maker_ITAG      exon    344080  344133  .       -       .       ID=exon:Solyc00g500003.1.1.3;Parent=mRNA:Sol
SL4.0ch00       maker_ITAG      CDS     344080  344133  .       -       2       ID=CDS:Solyc00g500003.1.1.3;Parent=mRNA:Soly
SL4.0ch00       maker_ITAG      exon    347298  347428  .       -       .       ID=exon:Solyc00g500003.1.1.4;Parent=mRNA:Sol
SL4.0ch00       maker_ITAG      CDS     347298  347428  .       -       1       ID=CDS:Solyc00g500003.1.1.4;Parent=mRNA:Soly
SL4.0ch00       maker_ITAG      exon    351799  352644  .       -       .       ID=exon:Solyc00g500003.1.1.5;Parent=mRNA:Sol
SL4.0ch00       maker_ITAG      CDS     351799  352644  .       -       1       ID=CDS:Solyc00g500003.1.1.5;Parent=mRNA:Soly
SL4.0ch00       maker_ITAG      exon    381867  382066  .       -       .       ID=exon:Solyc00g500003.1.1.6;Parent=mRNA:Sol
SL4.0ch00       maker_ITAG      CDS     381867  382066  .       -       0       ID=CDS:Solyc00g500003.1.1.6;Parent=mRNA:Soly
###
SL4.0ch00       maker_ITAG      gene    417592  418482  .       +       .       ID=gene:Solyc00g500004.1;Alias=Solyc00g50000
```

Hands-on …

NGS wiki

https://github.com/gsilvaarias/NGS2021-AGROSAVIA/wiki/02.-Bases-de-datos-y-formatos-de-archivos-NGS

1st part: Download sequence data and familiarize with SRA data repository

2nd part: Dive into different file formats, read on your own, try to dissect information (focus on fastq and SAM)