系统生物学

天津医科大学
生物医学工程与技术学院

2016-2017 学年上学期（秋）
2013 级生信班

第二章　基因组学

伊现富（Yi Xianfu）

天津医科大学（TIJMU）
生物医学工程与技术学院

2016 年 9 月

教学提纲

1. 数据库与数据格式
   - 数据库
   - 数据格式
2. 回顾与总结
   - 总结
   - 思考题

教学提纲

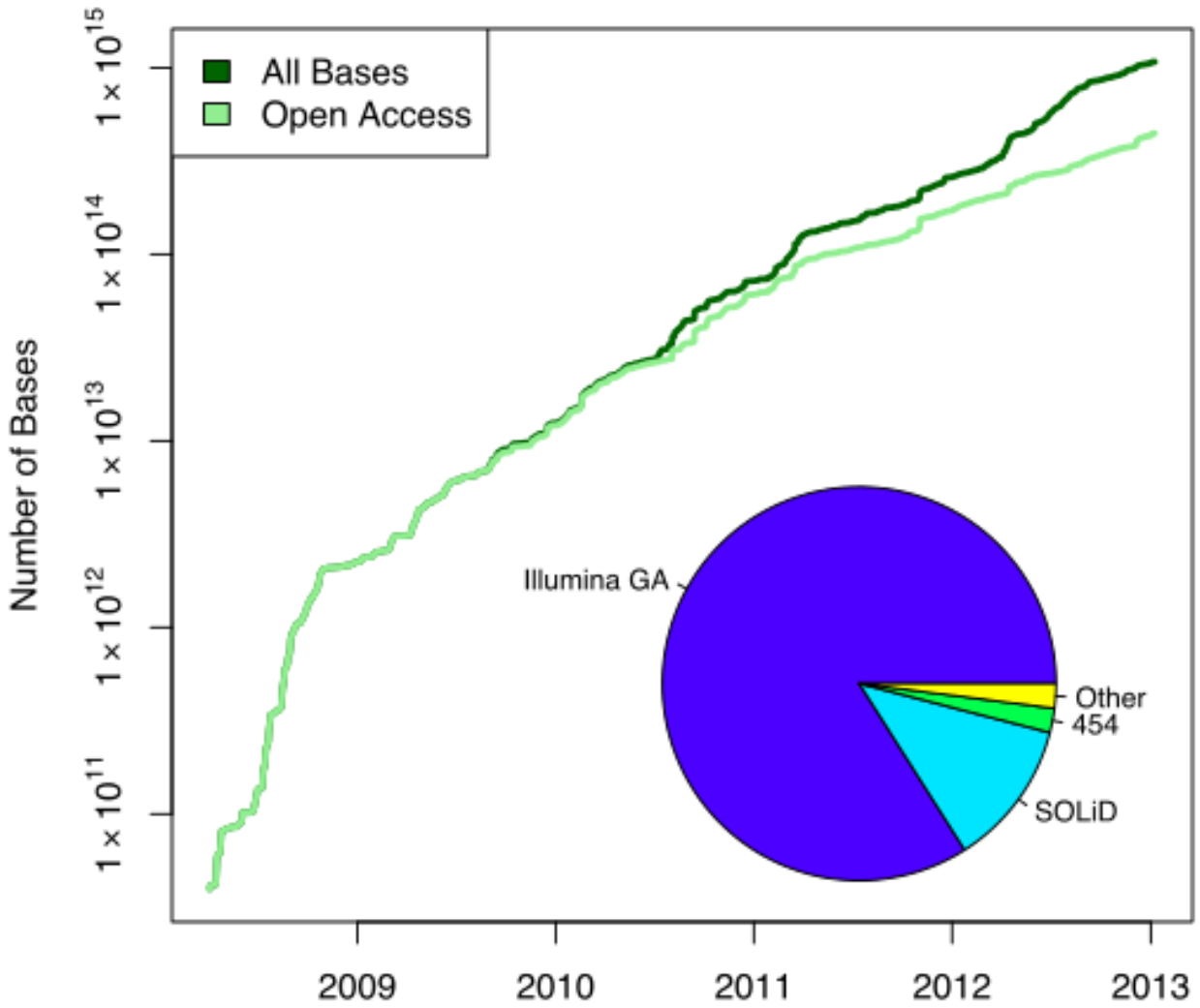
教学提纲

SRA

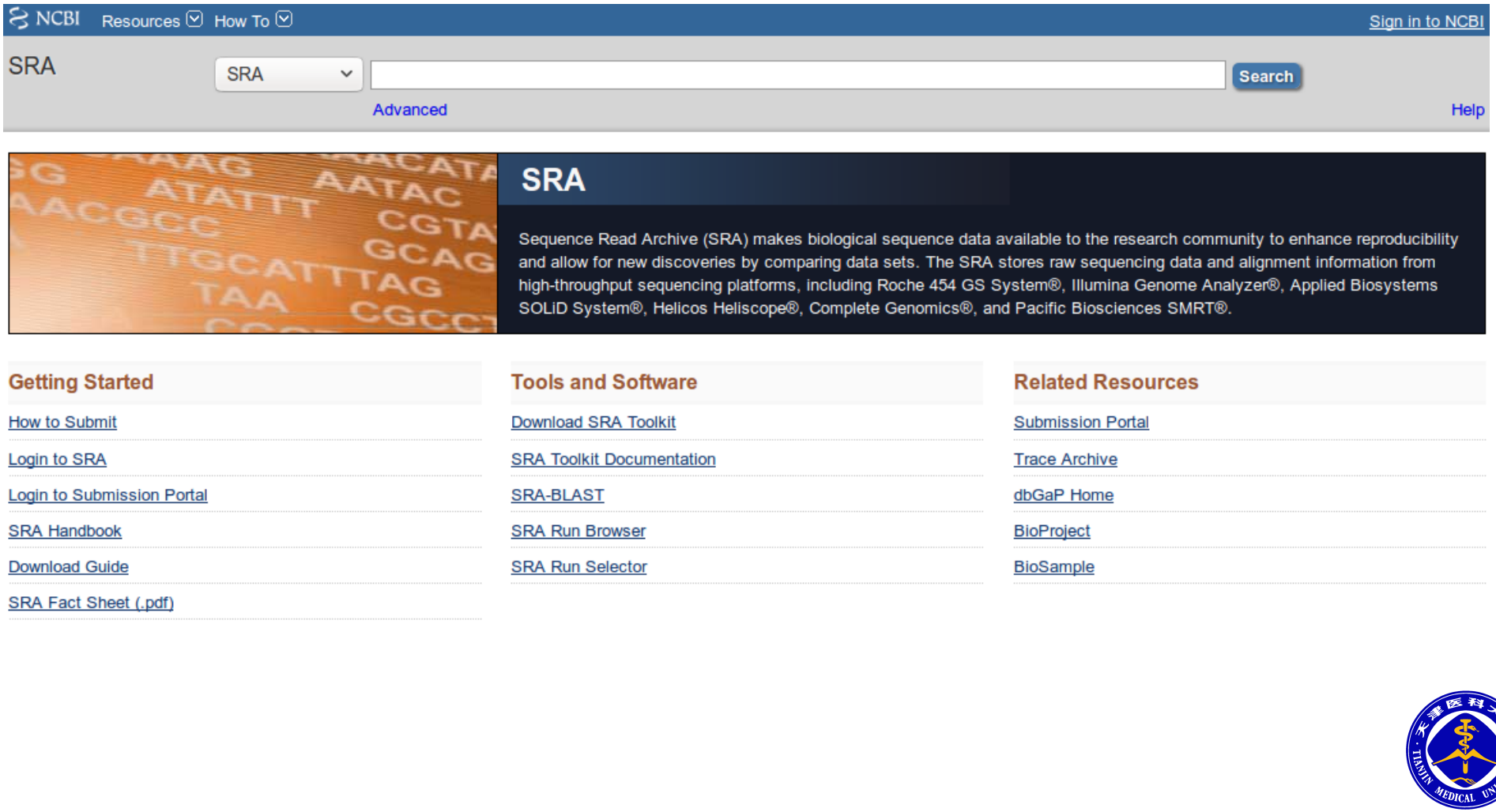
NCBI 在 2007 年底推出了 SRA 数据库，专门用于存储、显示、提取和分析高通量测序数据。

SRA 数据库，最初的命名为 Short Read Archive，现已改为 Sequence Read Archive。

Sequence Read Archive (SRA) makes biological sequence data available to the research community to enhance reproducibility and allow for new discoveries by comparing data sets. The SRA stores raw sequencing data and alignment information from high-throughput sequencing platforms, including Roche 454 GS System®, Illumina Genome Analyzer®, Applied Biosystems SOLiD System®, Helicos Heliscope®, Complete Genomics®, and Pacific Biosciences SMRT®.


All Bases
Open Access
Number of Bases
1 × 10^11
1 × 10^12
1 × 10^13
1 × 10^14
1 × 10^15
2009
2010
2011
2012
2013
Illumina GA
Other
454
SOLiD

基因组学 | 测序 | 数据库 | SRA



SRX214992: DGE sequencing of Human tumor 3
1 ILLUMINA (Illumina HiSeq 2000) run: 4.4M spots, 92.9M bases, 59.8Mb downloads

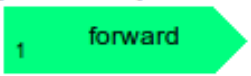
Submitted by: SYSU

Study: Homo sapiensTranscriptome or Gene expression
PRJNA185379 • SRP017786 • All experiments • All runs
show Abstract

Sample: DGE sequencing of Human tumor 3
SAMN01883035 • SRS383504 • All experiments • All runs
Organism: Homo sapiens

Library:
Instrument: Illumina HiSeq 2000
Strategy: RNA-Seq
Source: TRANSCRIPTOMIC
Selection: RANDOM
Layout: SINGLE

Spot descriptor:



Runs: 1 run, 4.4M spots, 92.9M bases, 59.8Mb

Run	# of Spots	# of Bases	Size	Published
SRR645375	4,424,492	92.9M	59.8Mb	2015-07-22

ID: 294252

Sequence Read Archive

Main | Browse | Search | Download | Submit | Documentation | Software | Trace Archive | Trace Assembly | Trace Home | Trace BLAST

Studies | Samples | Analyses | Run Browser | Run Selector | Provisional SRA

DGE sequencing of Human tumor 3 (SRR645375)

Metadata | Reads | Download

Run	Spots	Bases	Size	GC content	Published	Access Type
SRR645375	4.4M	92.9Mbp	62.7M	43.5%	2015-07-22	public

Quality graph (bigger)

This run has 1 read per spot:

L=21, 100%

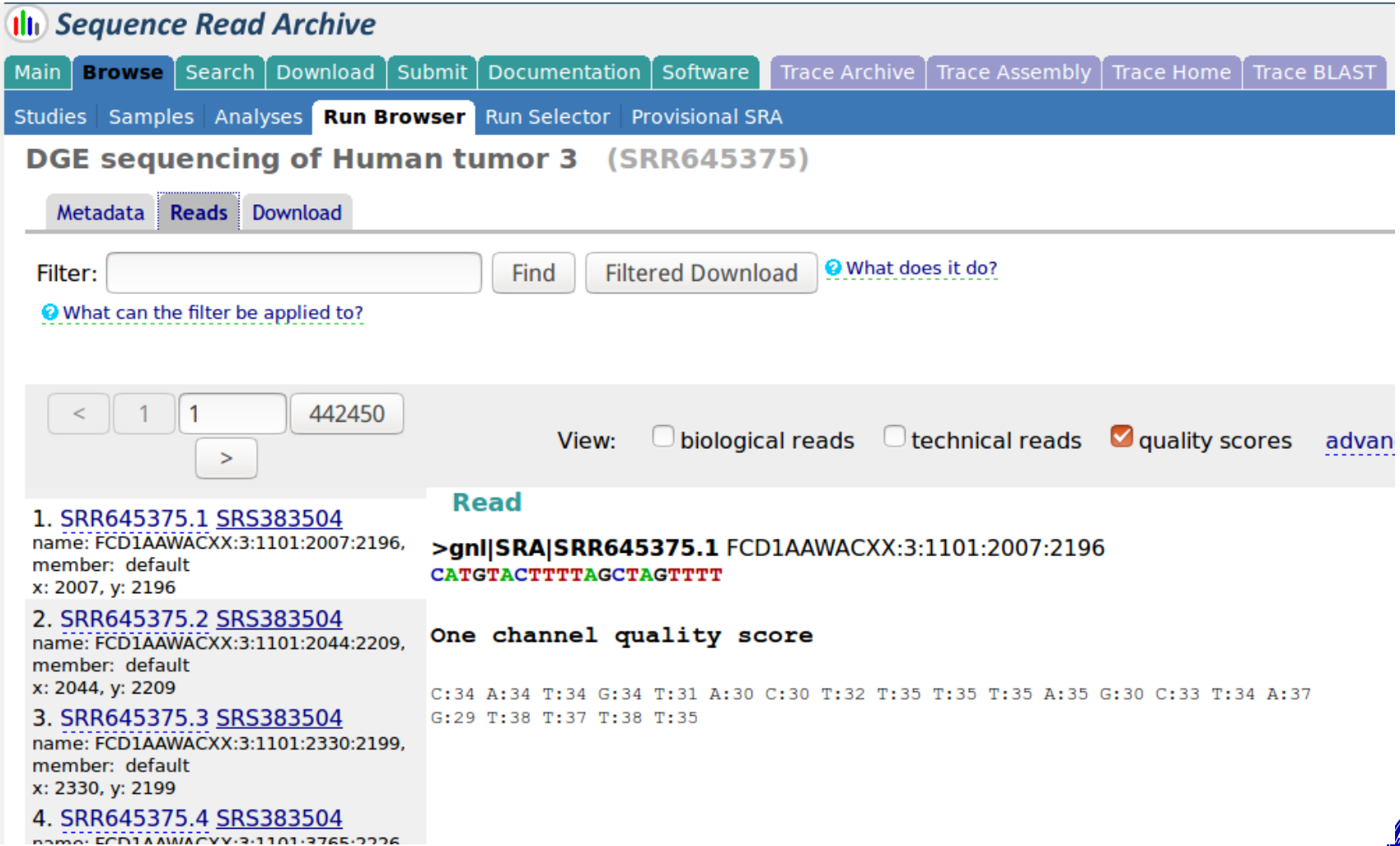
Legend

Experiment	Library					
SRX214992	Name	Platform	Strategy	Source	Selection	Layout
to BLAST		Illumina	RNA-Seq	TRANSCRIPTOMIC	RANDOM	SINGLE

Biosample	Sample Description	Organism
SAMN01883035 (SRS383504)		Homo sapiens

Bioproject	SRA Study	Title
PRJNA185379	SRP017786	Homo sapiensTranscriptome or Gene expression

Show abstract


Sequence Read Archive
Main
Browse
Search
Download
Submit
Documentation
Software
Trace Archive
Trace Assembly
Trace Home
Trace BLAST
Studies
Samples
Analyses
Run Browser
Run Selector
Provisional SRA
DGE sequencing of Human tumor 3 (SRR645375)
Metadata
Reads
Download
Filter:
Find
Filtered Download
What does it do?
What can the filter be applied to?
<
1
1
442450
>
View:
biological reads
technical reads
quality scores
advan
1. SRR645375.1 SRS383504
name: FCD1AAWACXX:3:1101:2007:2196,
member: default
x: 2007, y: 2196
2. SRR645375.2 SRS383504
name: FCD1AAWACXX:3:1101:2044:2209,
member: default
x: 2044, y: 2209
3. SRR645375.3 SRS383504
name: FCD1AAWACXX:3:1101:2330:2199,
member: default
x: 2330, y: 2199
4. SRR645375.4 SRS383504
Read
>gnl|SRA|SRR645375.1 FCD1AAWACXX:3:1101:2007:2196
CATGTACTTTTAGCTAGTTTT
One channel quality score
C:34 A:34 T:34 G:34 T:31 A:30 C:30 T:32 T:35 T:35 T:35 A:35 G:30 C:33 T:34 A:37
G:29 T:38 T:37 T:38 T:35

GEO

NCBI 的 GEO（Gene Expression Omnibus）数据库是一个非常强大的高通量数据集合，它综合了大量的芯片数据和二代测序数据，供全球科研工作者免费使用。

NCBI 的 GEO 数据库用于存储高通量的芯片实验数据，在 SRA 未建立之前，GEO 数据库也用于存储高通量测序数据。

GEO is a public functional genomics data repository supporting MIAME-compliant data submissions. Array- and sequence-based data are accepted. Tools are provided to help users query and download experiments and curated gene expression profiles.

NCBI Resources How To Sign in to NCBI

GEO Home | Documentation | Query & Browse | Email GEO

Gene Expression Omnibus

GEO is a public functional genomics data repository supporting MIAME-compliant data submissions. Array- and sequence-based data are accepted. Tools are provided to help users query and download experiments and curated gene expression profiles.

Gene Expression Omnibus

Keyword or GEO Accession Search

Getting Started

- Overview
- FAQ
- About GEO DataSets
- About GEO Profiles
- About GEO2R Analysis
- How to Construct a Query
- How to Download Data

Tools

- Search for Studies at GEO DataSets
- Search for Gene Expression at GEO Profiles
- Search GEO Documentation
- Analyze a Study with GEO2R
- GEO BLAST
- Programmatic Access
- FTP Site

Browse Content

- Repository Browser

DataSets:	3848
Series:	71107
Platforms:	16059
Samples:	1863122

Information for Submitters

- Login to Submit
- Submission Guidelines
- Update Guidelines
- MIAME Standards
- Citing and Linking to GEO
- Guidelines for Reviewers
- GEO Publications

Platforms (1)	GPL17021 Illumina HiSeq 2500 (Mus musculus)
Samples (15) ⊞ More...	GSM2176510 LSCs in gonadal adipose tissue (replicate 1) GSM2176511 LSCs in gonadal adipose tissue (replicate 2) GSM2176512 LSCs in gonadal adipose tissue (replicate 3)

Relations

BioProject	PRJNA322680
SRA	SRP075661

Download family	Format
SOFT formatted family file(s)	SOFT
MINiML formatted family file(s)	MINiML
Series Matrix File(s)	TXT

Supplementary file	Size	Download	File type/resource
GSE81842_AT_vs_NBM.txt.gz	904.8 Kb	(ftp)(http)	TXT
GSE81842_BL_vs_AT.txt.gz	901.7 Kb	(ftp)(http)	TXT
GSE81842_BL_vs_BM.txt.gz	906.4 Kb	(ftp)(http)	TXT
GSE81842_BL_vs_NBM.txt.gz	906.6 Kb	(ftp)(http)	TXT
GSE81842_BL_vs_Spl.txt.gz	884.1 Kb	(ftp)(http)	TXT
GSE81842_BM_vs_AT.txt.gz	904.1 Kb	(ftp)(http)	TXT
GSE81842_BM_vs_NBM.txt.gz	897.2 Kb	(ftp)(http)	TXT
GSE81842_Spl_vs_AT.txt.gz	881.7 Kb	(ftp)(http)	TXT
GSE81842_Spl_vs_BM.txt.gz	867.7 Kb	(ftp)(http)	TXT
GSE81842_Spl_vs_NBM.txt.gz	872.7 Kb	(ftp)(http)	TXT
GSE81842_Summary.txt.gz	271 b	(ftp)(http)	TXT
GSE81842_frm_gene_exp.diff.txt.gz	8.7 Mb	(ftp)(http)	TXT
GSE81842_frm_genes.count_tracking.txt.gz	763.1 Kb	(ftp)(http)	TXT
GSE81842_frm_genes.fpkm_tracking.txt.gz	1.4 Mb	(ftp)(http)	TXT
SRP/SRP075/SRP075661		(ftp)	SRA Study

Raw data provided as supplementary file

Processed data is available on Series record

Library strategy	RNA-Seq
Library source	transcriptomic
Library selection	cDNA
Instrument model	Illumina HiSeq 2500

Description	AT1
Data processing	(Illumina -> FastQ) & DEMULTIPLEXING - bcltofastq-1.8.4 Quality filter raw read data: Trimmomatic-0.32 Read alignment: SHRiMP_2_2_3 Data normalization and differential expression analysis: cufflinks-2.0.2.Linux_x86_64 (cuffdiff) Genome build: mm10

Submission date	May 24, 2016
Last update date	Jul 07, 2016
Contact name	Haobin Ye
E-mail	haobin.ye@ucdenver.edu
Organization name	University of Colorado
Department	Hematology
Lab	Craig Jordan Lab
Street address	12700 East 19th Ave, Room 9122
City	Aurora
State/province	Colorado
ZIP/Postal code	80045
Country	USA

Platform ID	GPL17021
Series (1)	GSE81842 Genome-wide comparison of gene expression level between leukemia stem cells in different tissues

Relations

BioSample	SAMN05172382
SRA	SRX1798522

Supplementary file	Size	Download	File type/resource
SRX/SRX179/SRX1798522		(ftp)	SRA Experiment

Raw data provided as supplementary file

千人基因组计划

千人基因组计划（1000 Genomes Project），旨在绘制迄今（截至 2011 年）最详尽、最有医学应用价值的人类基因多态性图谱，该图谱由中美英等国科研机构发起的“千人基因组计划”共同协作完成，标志着人类基因研究取得重大突破。这项计划于 2008 年启动，目前该项目拥有超过 1700 个样本，高达 200TB 数据量的 DNA 序列。2012 年开始全部数据免费对外开放。

IGSR: The International Genome Sample Resource

Providing ongoing support for the 1000 Genomes Project data

Home About Data Portal Analysis Contact Browser FAQ

Search 1000genomes

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.

Links

Announcements

IGSR Sample Collection Principles

1000 Genomes Project Publications

File formats

Software tools

Download data

Twitter

Using data from IGSR

IGSR provides open data to support the community's research efforts. You can see our terms of use in our data disclaimer.

Data portal ***beta***

We are developing a new data portal to make it easier to find and browse data in IGSR. You can use the development version to explore the data set. Let us know what you think at info@1000genomes.org.

Sample	Sex	Population	Exome	Low cov WG	High cov WG	HD genotype	Complete
HG00513	Female	CHS	●	●	●	●	
HG01112	Male	CLM	●	●	●	●	
HG00759	Female	CDX	●	●	●	●	
HG01500	Male	IBS	●	●	●	●	
HG03006	Male	BEB	●	●	●		
NA18525	Female	CHB	●	●	●	●	
NA19648	Female	MXL	●	●	●	●	

Data collections for HG00119

1000 Genomes on GRCh38 | 1000 Genomes phase 3 release | 1000 Genomes phase 1 release

Data reuse policy for 1000 Genomes on GRCh38

22 matching data files

Download the list

« Previous | Next »

Data types

- Sequence
- Alignment

Analysis groups

- Exome
- Low coverage WGS

File	Analysis group
ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR043/SRR043348/SRR043348_1.-fastq.gz	Low coverage WGS
ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR043/SRR043354/SRR043354_1.-fastq.gz	Low coverage WGS
ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR043/SRR043372/SRR043372_1.-fastq.gz	Low coverage WGS
ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR043/SRR043378/SRR043378_2.-fastq.gz	Low coverage WGS
ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR099/SRR099967/SRR099967_1.-	Exome

TCGA

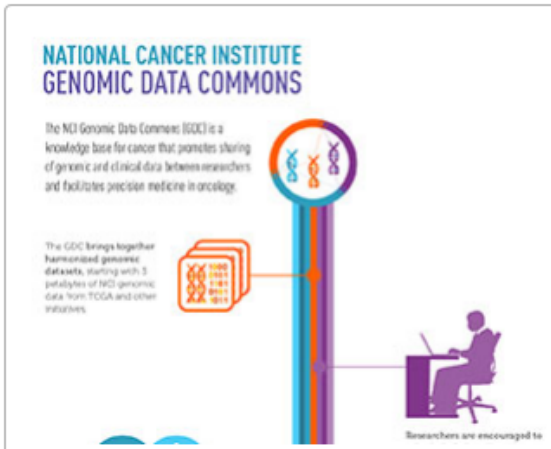
Cancer Genome Atlas（TCGA）和 International Cancer Consortium(ICGC) 是目前国际上最大的两个癌症基因信息检索数据库，共收集了 43 种癌症的超过 13 万个样本数据，此外还涉及到相关癌症基因的 mRNA/microRNA 表达谱、拷贝数变异、突变等大量的生物信息学数据。

NIH THE CANCER GENOME ATLAS
National Cancer Institute
National Human Genome Research Institute

Launch Data Portal | Contact Us | For the Media

Search Search

Home | About Cancer Genomics | Cancers Selected for Study | Research Highlights | Publications | News and Events | About TCGA



The NCI Genomic Data Commons Launches

The Genomic Data Commons (GDC) is a data sharing platform that promotes precision medicine in oncology, and it will host all of the TCGA data.

Learn More

Genomic Data Commons Launches | Analysis of Adrenocortical Carcinoma | Cancers Selected for Study | About TCGA

Research Briefs

July 2016
Longitudinal Study Charts Brain Tumor Evolution

News and Announcements

June 06, 2016
Newly launched Genomic Data Commons to facilitate

Launch Data Portal

The Genomic Data Commons (GDC) Data Portal is an interactive data system for researchers to search, download, upload, and analyze harmonized cancer genomic data sets, including TCGA.

Questions About Cancer

Visit www.cancer.gov

Call 1-800-4-CANCER

Use LiveHelp Online Chat

Multimedia Library

Images

Videos and Animations

Podcasts

NIH NATIONAL CANCER INSTITUTE GDC Data Portal

Home | Projects | Data | Analysis

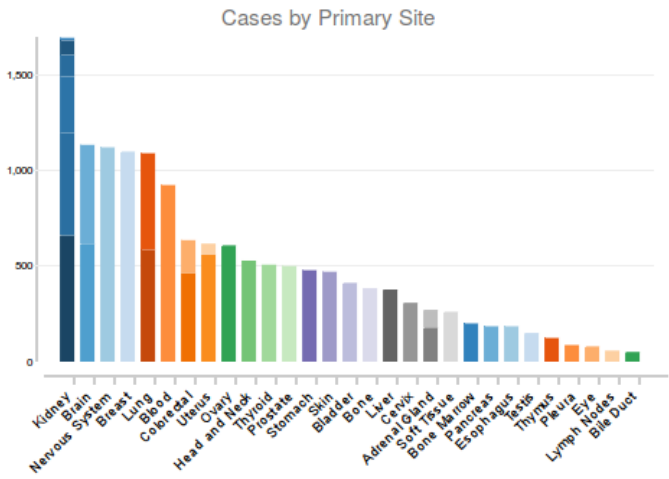
Harmonized Cancer Datasets
Genomic Data Commons Data Portal

Get Started by Exploring:

Projects | Data

Perform Advanced Search Queries, such as:

Cases of kidney cancer diagnosed at the age of 20 and below	182 Cases	1,501 Files
CNV data of female brain cancer cases	459 Cases	1,788 Files
Gene expression quantification data in TCGA-GBM project	166 Cases	522 Files



DATA PORTAL SUMMARY

Data Release 1.0 - June 6, 2016

PROJECTS	PRIMARY SITE	CASES	FILES
39	29	14,531	250,498

Infrastructure

Data is continuously being processed and harmonized by the GDC.
View GDC system statistics:

Compute Infrastructure	12,800 Cores	87.96 TB RAM
Storage Infrastructure	4.98 PB Used	5.42 PB Total

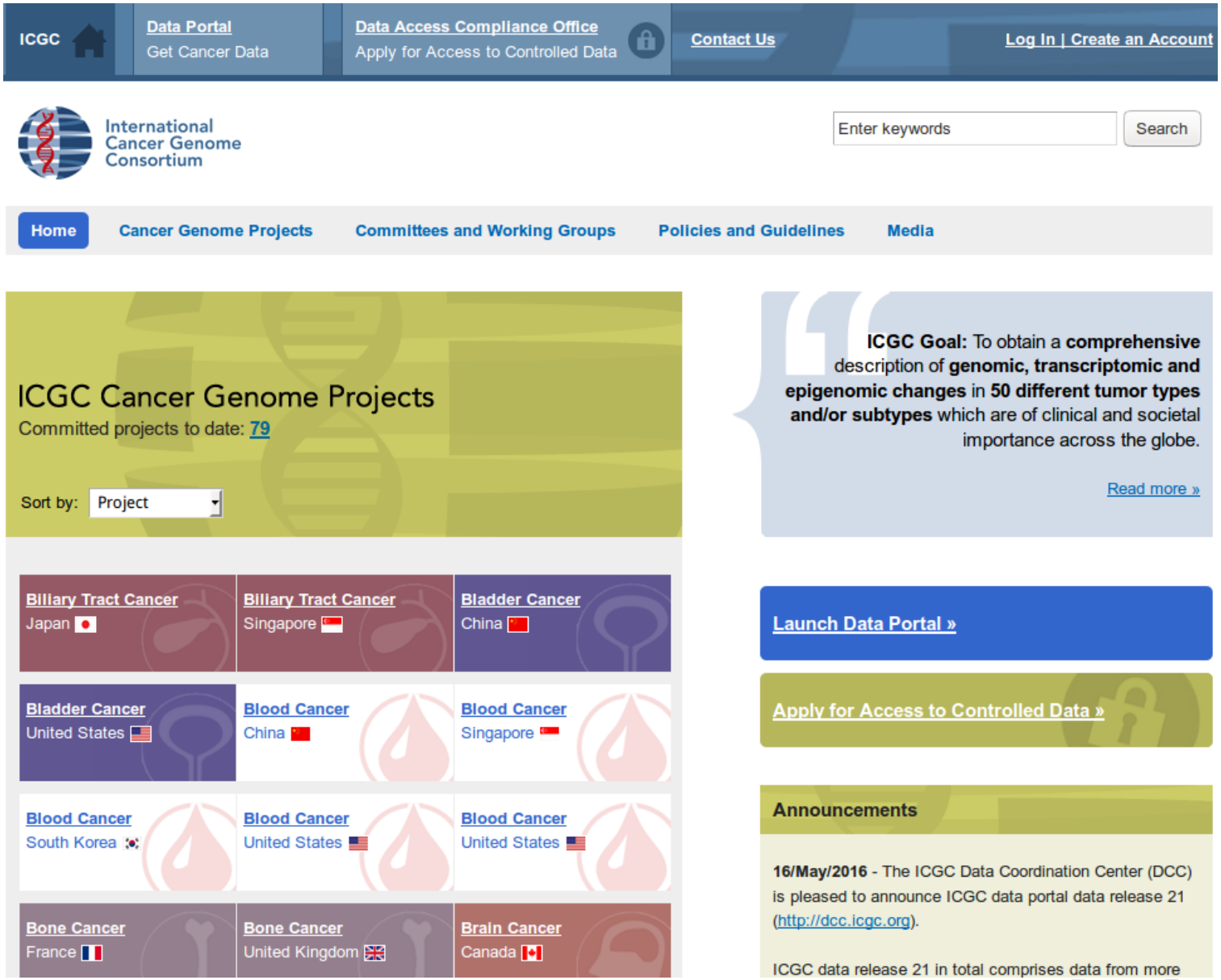
View Data Download Statistics Report »

Documentation

Learn how to use the GDC Data Portal to its full potential with common topics such as:

- Browse Data using Facet Search
- Search Data with Advanced Search Technology
- Project Based Data Availability
- Controlled Access Data

Visit the Documentation Website »


ICGC
Data Portal
Get Cancer Data
Data Access Compliance Office
Apply for Access to Controlled Data
Contact Us
Log In | Create an Account
International Cancer Genome Consortium
Enter keywords
Search
Home
Cancer Genome Projects
Committees and Working Groups
Policies and Guidelines
Media
ICGC Cancer Genome Projects
Committed projects to date: 79
Sort by: Project
Biliary Tract Cancer Japan
Biliary Tract Cancer Singapore
Bladder Cancer China
Bladder Cancer United States
Blood Cancer China
Blood Cancer Singapore
Blood Cancer South Korea
Blood Cancer United States
Blood Cancer United States
Bone Cancer France
Bone Cancer United Kingdom
Brain Cancer Canada
ICGC Goal: To obtain a comprehensive description of genomic, transcriptomic and epigenomic changes in 50 different tumor types and/or subtypes which are of clinical and societal importance across the globe.
Read more »
Launch Data Portal »
Apply for Access to Controlled Data »
Announcements
16/May/2016 - The ICGC Data Coordination Center (DCC) is pleased to announce ICGC data portal data release 21 (http://dcc.icgc.org).
ICGC data release 21 in total comprises data from more

ICGC Data Portal

Cancer Projects | Advanced Search | Data Analysis | DCC Data Releases | Data Repositories

e.g. BRAF, KRAS G12D, DO35100, MU7870, FI998, apoptosis, Cancer Gene Census, imatinib, GO:0016049

About Us

The ICGC Data Portal provides tools for visualizing, querying and downloading the data released quarterly by the consortium's member projects.

To access ICGC controlled tier data, please read these instructions.

New features will be regularly added by the DCC development team. Feedback is welcome.

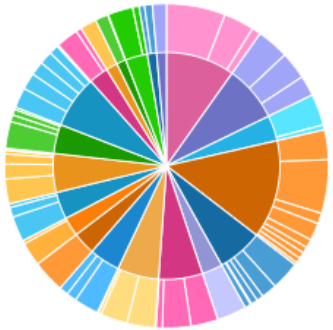
PCAWG PanCancer Analysis OF WHOLE GENOMES

The Pancancer Analysis of Whole Genomes (PCAWG) study is an international collaboration to identify common patterns of mutation in more than 2,800 cancer whole genomes from the International Cancer Genome Consortium.

Data Release 21

May 16th, 2016

Donor Distribution by Primary Site



Cancer projects	68
Cancer primary sites	21
Donors with molecular data in DCC	15,613
Total Donors	18,677
Simple somatic mutations	42,584,179

Tutorial

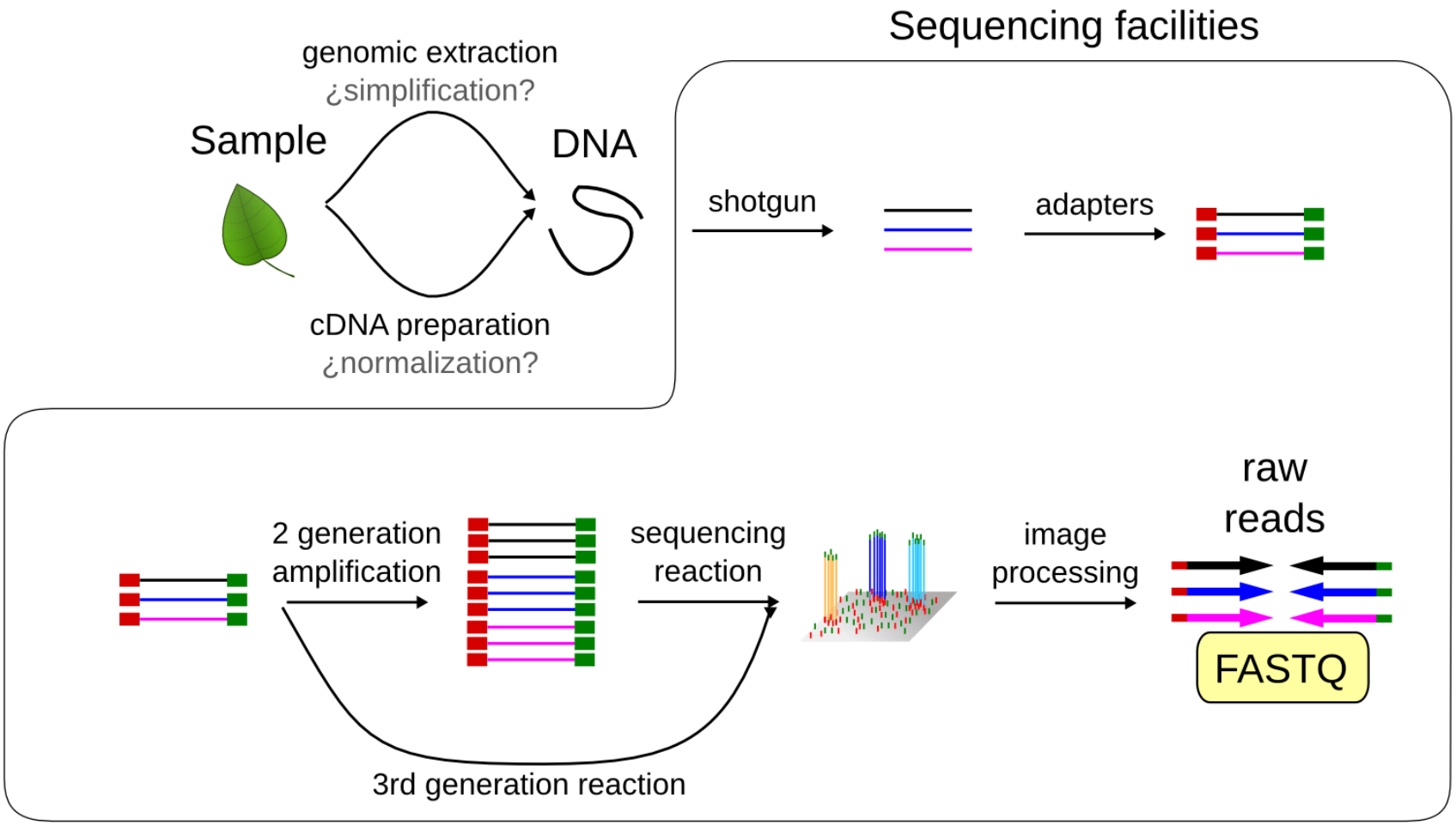
EXAMPLE QUERIES

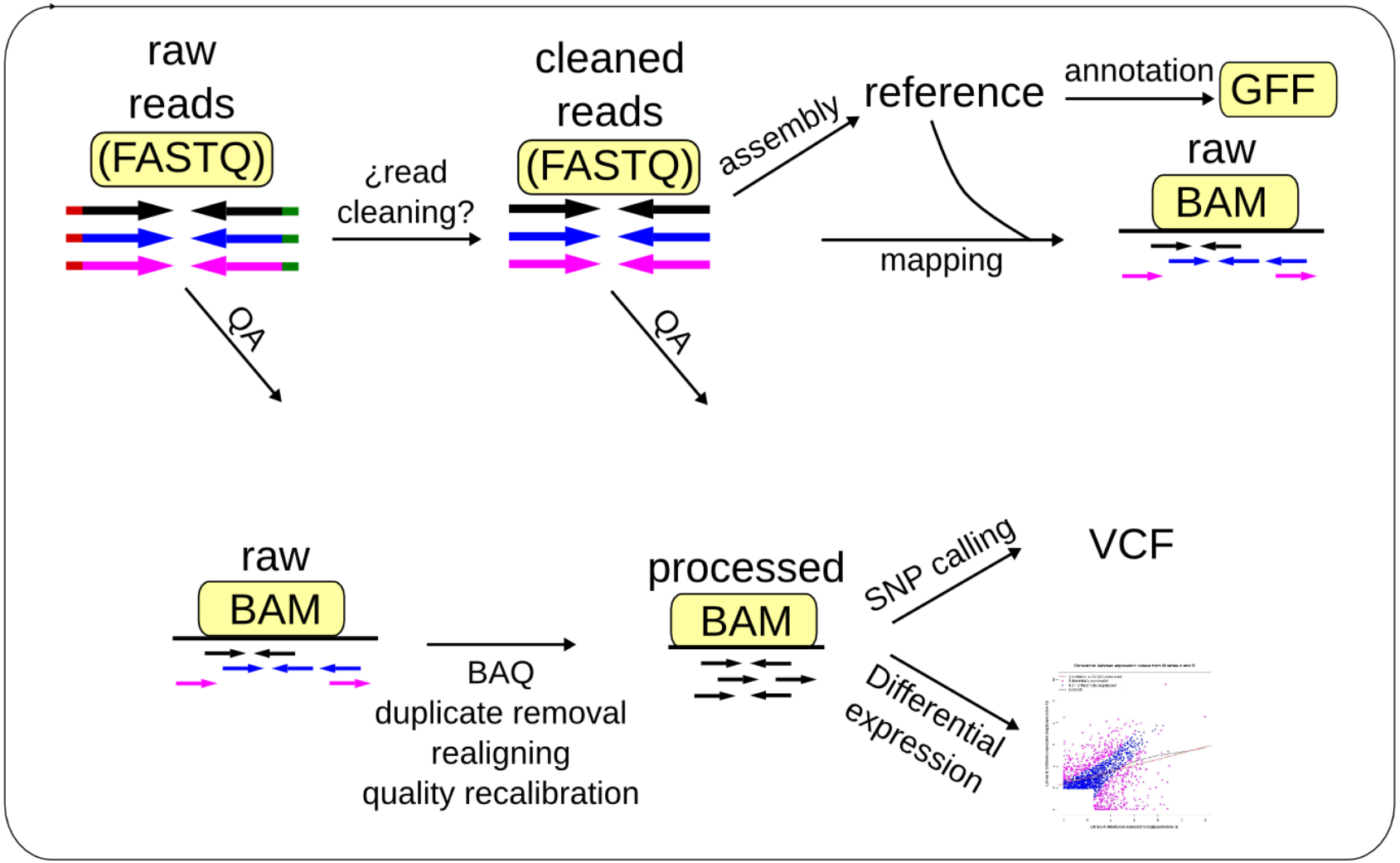
1. BRAF missense mutations in colorectal cancer
2. Most frequently mutated genes by high impact mutations in stage III malignant lymphoma
3. Brain cancer donors with frameshift mutations and having methylation data available

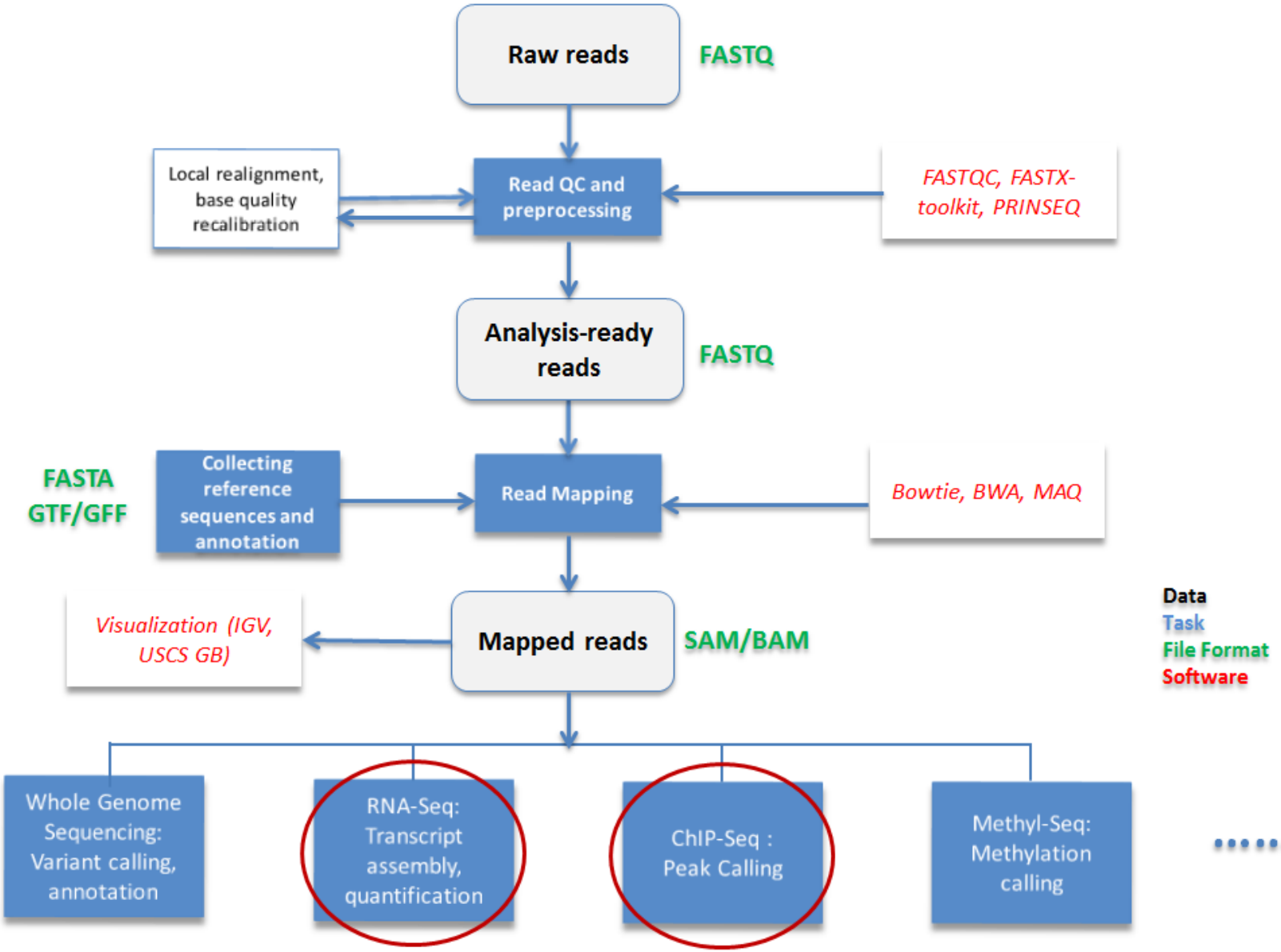
ICGC International Cancer Genome Consortium in the cloud

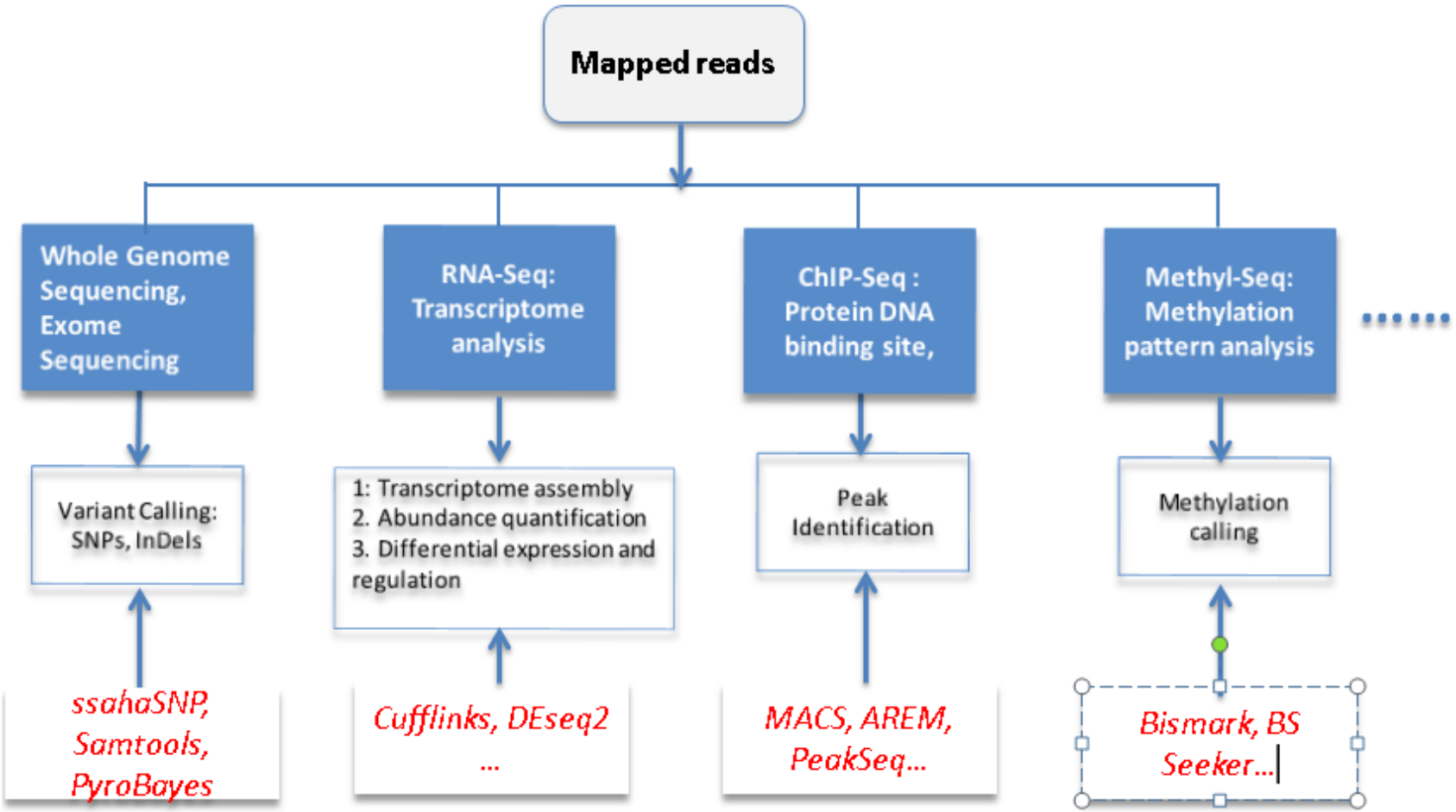
ICGC data is now available on commercial and academic compute cloud. Read more...

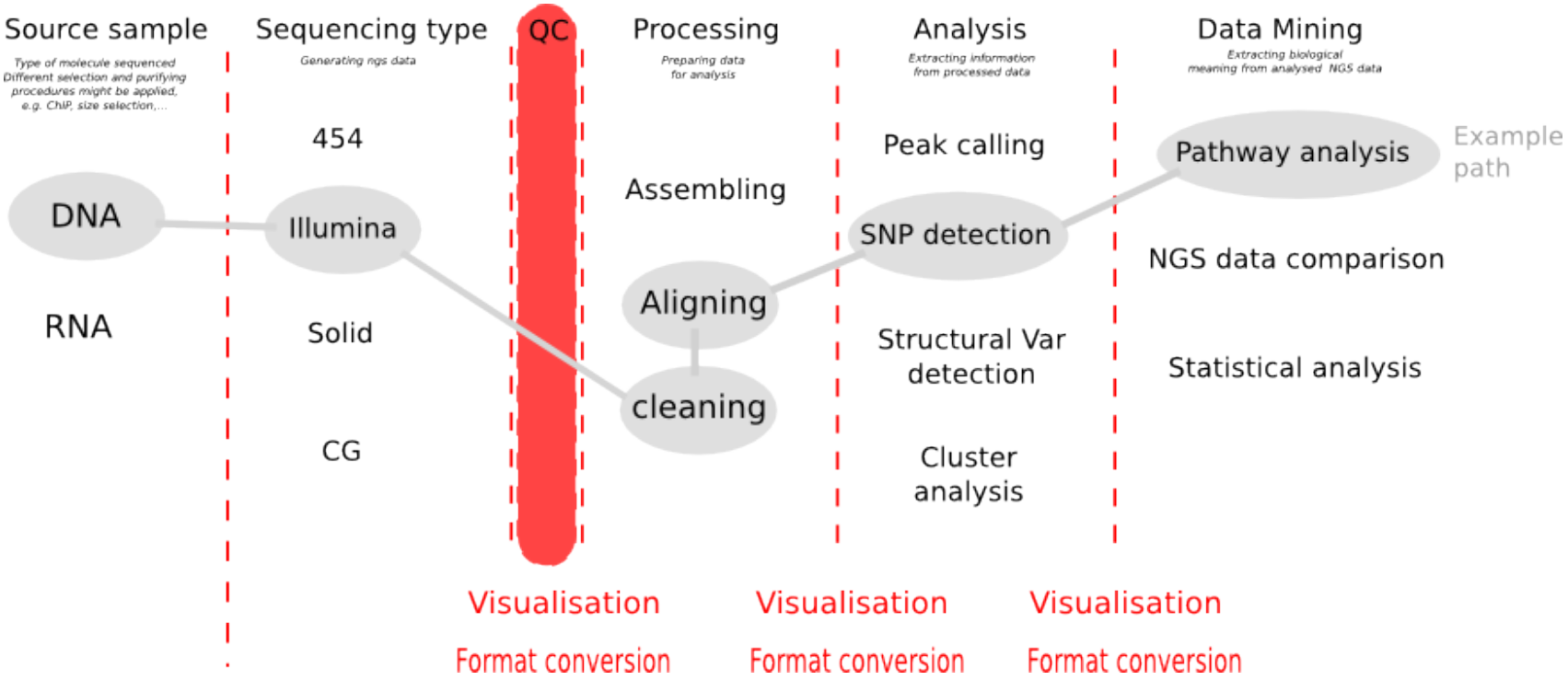
教学提纲


Sequencing facilities
genomic extraction
¿simplification?
Sample
DNA
cDNA preparation
¿normalization?
shotgun
adapters
2 generation
amplification
sequencing
reaction
image
processing
raw
reads
FASTQ
3rd generation reaction


raw
reads
(FASTQ)
¿read
cleaning?
cleaned
reads
(FASTQ)
assembly
reference
annotation
GFF
raw
BAM
mapping
QA
QA
raw
BAM
BAQ
duplicate removal
realigning
quality recalibration
processed
BAM
SNP calling
VCF
Differential
expression


Raw reads
FASTQ
Local realignment, base quality recalibration
Read QC and preprocessing
FASTQC, FASTX-toolkit, PRINSEQ
Analysis-ready reads
FASTQ
FASTA
GTF/GFF
Collecting reference sequences and annotation
Read Mapping
Bowtie, BWA, MAQ
Visualization (IGV, USCS GB)
Mapped reads
SAM/BAM
Data
Task
File Format
Software
Whole Genome Sequencing: Variant calling, annotation
RNA-Seq: Transcript assembly, quantification
ChIP-Seq : Peak Calling
Methyl-Seq: Methylation calling

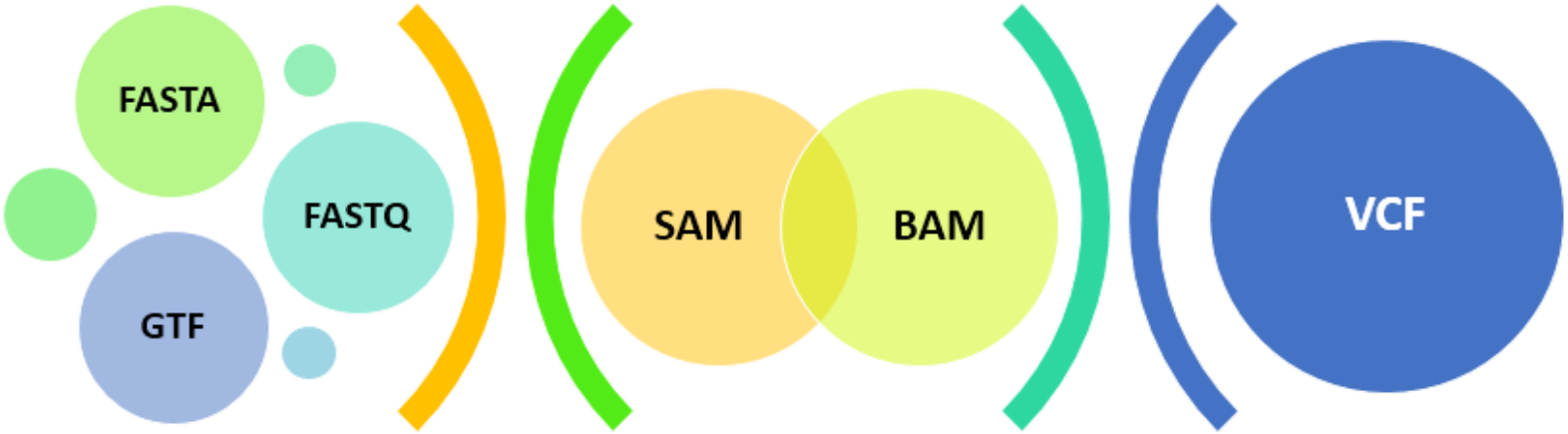

Mapped reads
Whole Genome Sequencing, Exome Sequencing
RNA-Seq: Transcriptome analysis
ChIP-Seq : Protein DNA binding site,
Methyl-Seq: Methylation pattern analysis
.....
Variant Calling: SNPs, InDels
1: Transcriptome assembly
2. Abundance quantification
3. Differential expression and regulation
Peak Identification
Methylation calling
ssahaSNP, Samtools, PyroBayes
Cufflinks, DEseq2 ...
MACS, AREM, PeakSeq...
Bismark, BS Seeker...


Source sample
Type of molecule sequenced
Different selection and purifying procedures might be applied, e.g. ChIP, size selection,...
DNA
RNA
Sequencing type
Generating ngs data
454
Illumina
Solid
CG
QC
Processing
Preparing data for analysis
Assembling
Aligning
cleaning
Analysis
Extracting information from processed data
Peak calling
SNP detection
Structural Var detection
Cluster analysis
Data Mining
Extracting biological meaning from analysed NGS data
Pathway analysis
NGS data comparison
Statistical analysis
Example path
Visualisation
Format conversion
Visualisation
Format conversion
Visualisation
Format conversion

Common NGS Data Formats

File extension	Description	Reference	Publication
.fasta	Classic DNA sequence file format	http://www.ncbi.nlm.nih.gov/blast/fasta.shtml	n/a
.ace	File format for whole-genome assemblies	Annotated in the documentation for CONSED, currently: http://www.phrap.org/consed/distributions/README.19.0.txt	Gordon, Abajian, and Green, 1998
.wig	A reference-genome indexed data series for "dense" and continuous data (such as %GC)	http://genome.ucsc.edu/goldenPath/help/wiggle.html	Haussler, 2002
.bed	A reference-genome indexed data series for "sparse" data (such as transcriptome data)	http://genome.ucsc.edu/goldenPath/help/bedgraph.html	Haussler, 2002
.tab	Tab-delimited text	N/A	n/a
.pdf	Portable document format	Either ISO-32000-1 or http://www.adobe.com/devnet/pdf/pdf_reference.html	n/a
.sam	"Sequence Alignment/Map" format	http://samtools.sourceforge.net/SAM1.pdf	Li, 2009
.bam	Binary format of .sam	http://samtools.sourceforge.net/SAM1.pdf	Li, 2009
.fastq	Combination of sequences and quality scores in one file; mainly for data from Illumina sequencers in which case quality scores have been transformed.	http://maq.sourceforge.net/fastq.shtml	Li, 2008, and Cock, 2009
.csfasta	Life Technologies SOLiD colorspace fasta file - containing color calls (0, 1, 2, 3) rather than base calls	See: http://solidsoftwaretools.com/	n/a
.qual	Per-base quality scores generated during basecalling. All but Illumina scores are scaled to estimate the probability of an incorrect base call, as is in common use for conventional sequencing as Phred quality scores.		Ewing & Green, 1998
.gff	A flexible format for annotating features (e.g. genes) on a sequence.	http://www.sanger.ac.uk/resources/software/gff/	n/a
.srf	"Short Read Format" - a new format proposed for short-read DNA sequence	http://srf.sf.net	n/a
.sff	Standard Flowgram Format (specific for Roche/454)	http://www.ncbi.nlm.nih.gov/Traces/trace.cgi?cmd=show&f=formats&m=doc&s=formats#header-global	n/a
.gtf	Gene transfer format, an alternate format to GFF for specifying gene features	http://mblab.wustl.edu/GTF22.html	n/a

For a full list, go to http://genome.ucsc.edu/FAQ/FAQformat.html


FASTA
FASTQ
GTF
SAM
BAM
VCF
Sequencing and annotation
Alignment
Variant Calling

Reference sequences

- FASTA
- 2bit

Reads

- FASTQ (FASTA with quality scores)

Alignments

- SAM (Sequence Alignment/Map format)
- BAM (Binary version of SAM)

Reference sequences

- FASTA
- 2bit

Reads

- FASTQ (FASTA with quality scores)

Alignments

- SAM (Sequence Alignment/Map format)
- BAM (Binary version of SAM)

Reference sequences

- FASTA
- 2bit

Reads

- FASTQ (FASTA with quality scores)

Alignments

- SAM (Sequence Alignment/Map format)
- BAM (Binary version of SAM)

Features, annotation, coverage, scores

- GFF3/GTF (General Feature Format, Gene Transfer Format)
- BED/bigBed (Browser Extensible Data)
- WIG/bigWig (Wiggle format)
- bedGraph

Variations

- VCF (Variant Call Format)
- BCF (Binary version of VCF)

Features, annotation, coverage, scores

- GFF3/GTF (General Feature Format, Gene Transfer Format)
- BED/bigBed (Browser Extensible Data)
- WIG/bigWig (Wiggle format)
- bedGraph

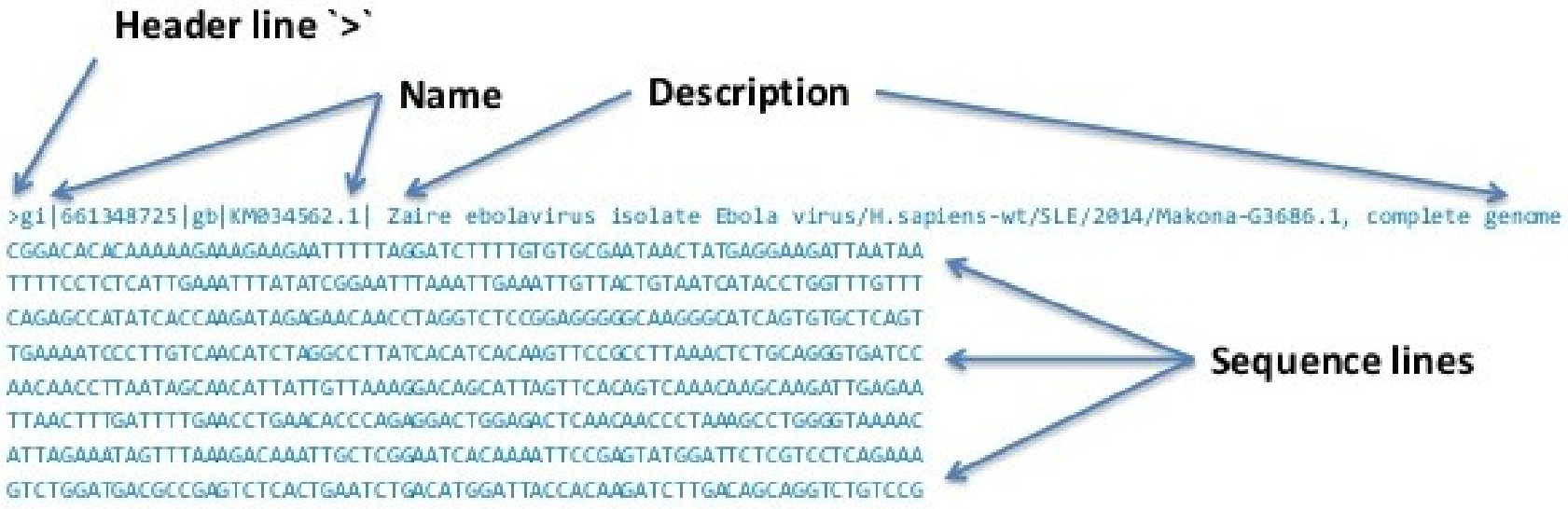
Variations

- VCF (Variant Call Format)
- BCF (Binary version of VCF)

FASTA Format

The FASTA format is a standard for displaying (nucleotide or protein) sequences in a text file. An entry for a sequence takes up two lines in the file: the first line begins with a ">" symbol, followed by the sequence description, and the second line contains the sequence itself.

```
>gi|67328264|gb|AAFC02129962.1| Bos taurus breed Hereford Con136352, whole genome shotgun sequence
CCCCCCCCCCCCCCCCCCGGGCACGTACCTGCTGGATCAGCCCCACCTGGAGCTGGGTGAGGAACAGCTG
GGGAAGGAAGCAAGCGGACAGTGAGCTGAGCCCCGGTGCCGGGTGCCGGCAGGCCCGCCCACCCTGGCCC
```



FASTQ

FASTQ format is a text-based format for storing both a biological sequence (usually nucleotide sequence) and its corresponding quality scores. Both the sequence letter and quality score are each encoded with a single ASCII character for brevity.

It was originally developed at the Wellcome Trust Sanger Institute to bundle a FASTA sequence and its quality data, but has recently become the de facto standard for storing the output of high-throughput sequencing instruments such as the Illumina Genome Analyzer.

There is no standard file extension for a FASTQ file, but .fq and .fastq, are commonly used.

- **FastA** format (everybody knows about it)
  - Header line starts with “>” followed by a sequence ID
  - Sequence (string of nt).

- **FastQ** format
  - First is the sequence (like Fasta but starting with “@”)
  - Then “+” and sequence ID (optional) and in the following line are QVs encoded as single byte ASCII codes
    - Different quality encode variants

FASTQ Format

The FASTQ format stores sequences and **Phred qualities scores** in a single file. FASTQ file uses four lines per sequence:

- Line 1 - begins with a '@' character and is followed by a sequence identifier and an optional description (like a sequence description),
- Line 2 - is the raw sequence letters,
- Line 3 - begins with a '+' character,
- Line 4 - encodes the quality values for the sequence in Line 2.

```
@WGG97JN1:192:C200YACXX:7:1101:1307:1960 1:N:0:TTAGGC
CGAGGAGCTGAGTCACAGAGCAGAAGGGGTTTCAGAGATTCGGGCTGTCCA
+
@@@FFFFFHHCFHHDHHIEGIIGIJGIHHGGHJIJIIIJIJJJGGGHFI7@
@WGG97JN1:192:C200YACXX:7:1101:1602:1991 1:N:0:TTAGGC
CTGCGGTTCCTCTCGTACTGAGCAGGATTACTAGCGCAACAACACATCATC
+
=??DD@=<AF?DFFF;EBDHCCFFG:E<D<?DFC>GGHD@BG.=@C;FGEE
```

Sequence identifier contains: WGG97JN1 the unique instrument name; 192 the run id; C200YACXX the flowcell id; 7 flowcell lane; 1101 tile number within the flowcell lane; 1307 'x'-coordinate of the cluster within the tile; 1960 'y'-coordinate of the cluster within the tile; 1 the member of a pair, 1 or 2 (paired-end or mate-pair reads only); Y if the read fails filter (read is bad), N otherwise; 0 when none of the control bits are on, otherwise it is an even number; TTAGGC index sequence.

Fastq files:

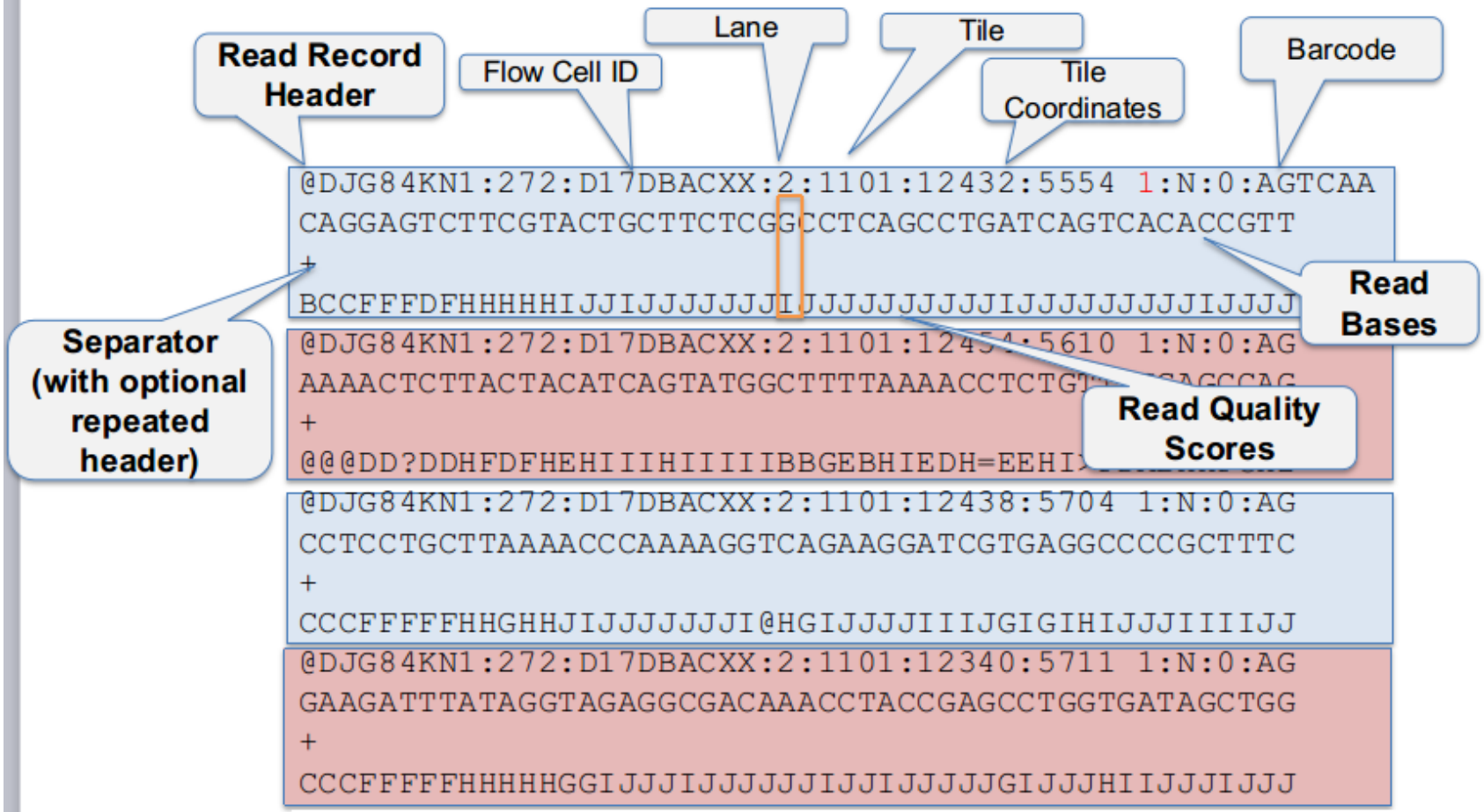
FASTQ format is a **text-based format** for storing both a biological **sequence** (usually nucleotide sequence) and its corresponding **quality scores.**

-Wikipedia

```
@SEQUENCE_ID1
ATGCGCGCGCGCGCGCGCGGGTAGCAGATGACGACACAGAGCGAGGATGCGCTGAGAGTA
GTGTGACGACGATGACGGAAAATCAGA
+
BBBBBPPPPPXXXXX^^^^^^^^^^^^^^^^^^^^^^^__^^^^^^^^^^^^_eeeeeee
[[[[[^^^^]]]]XXXXXPPPPPBBBB
```

1. **Single line ID** with at symbol ("@") in the first column.
2. There should be not space between "@" symbol and the first letter of the identifier.
3. Sequences are in multiple lines after the ID line
4. Single line with plus symbol ("+") in the first column to represent the quality line.
5. Quality ID line can have or have not ID
6. Quality values are in multiple lines after the + line

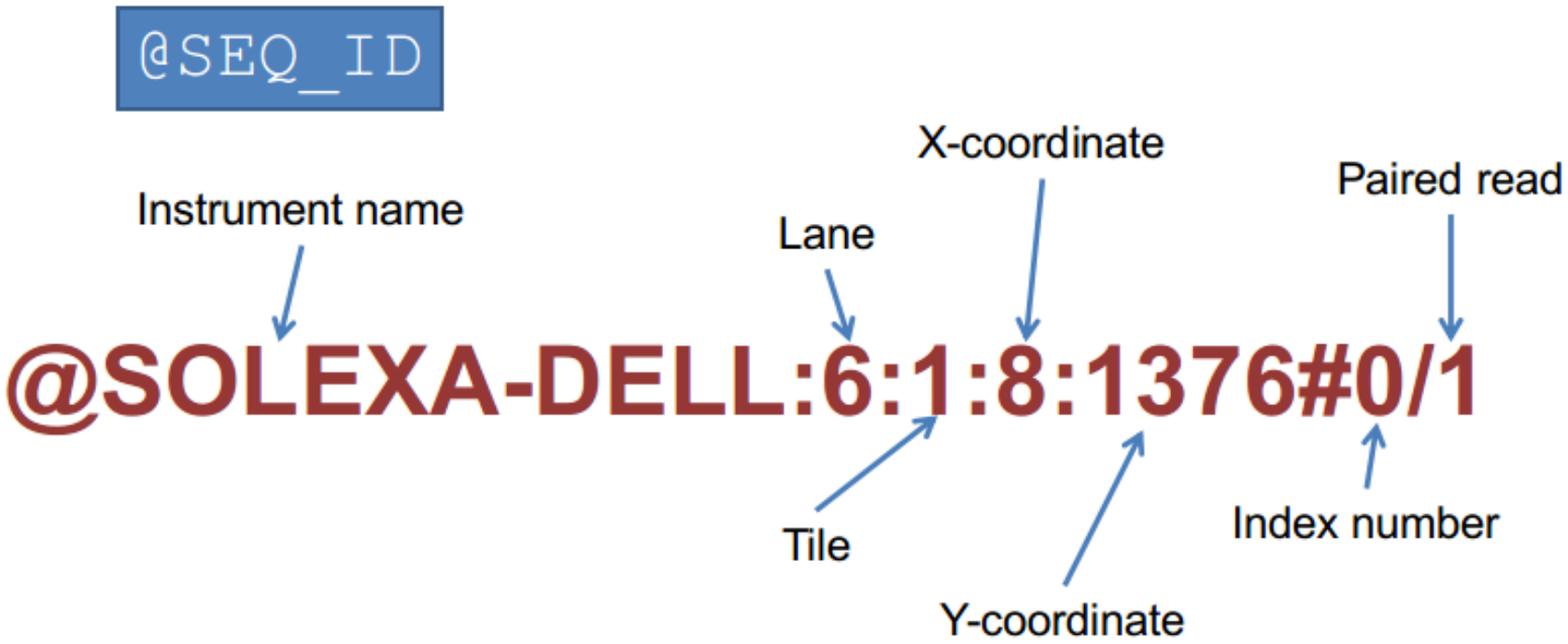
FASTQ Format (Illumina Example)



NOTE: for paired-end runs, there is a second file with one-to-one corresponding headers and reads.

(Passarelli, 2012)

Illumina sequence identifiers



Sequences from the Illumina software use a systematic identifier:

```
@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)*

Versions of the Illumina pipeline since 1.4 appear to use **#NNNNNN** instead of **#0** for the multiplex ID, where **NNNNNN** is the sequence of the multiplex tag.

With Casava 1.8 the format of the '@' line has changed:

```
@EAS139:136:FC706VJ:2:2104:15343:197393 1:Y:18:ATCACG
```

EAS139	the unique instrument name
136	the run id
FC706VJ	the flowcell id
2	flowcell lane
2104	tile number within the flowcell lane
15343	'x'-coordinate of the cluster within the tile
197393	'y'-coordinate of the cluster within the tile
1	the member of a pair, 1 or 2 *(paired-end or mate-pair reads only)*
Y	Y if the read is filtered, N otherwise
18	0 when none of the control bits are on, otherwise it is an even number
ATCACG	index sequence

Note that more recent versions of Illumina software output a sample number (as taken from the sample sheet) in place of an index sequence. For example, the following header might appear in the first sample of a batch:

```
@EAS139:136:FC706VJ:2:2104:15343:197393 1:N:18:1
```

```
@SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36
GGGTGATGGCCGCTGCCGATGGCGTCAAATCCCACC
+SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36
IIIIIIIIIIIIIIIIIIIIIIIIIIIII9IG9IC
```

Note

- ID = NCBI-assigned identifier + the original identifier from Solexa/Illumina + the read length
- fastq-dump: lost the paired-end information, concatenate sequence of the forward and reverse reads together into a non-sense
- NCBI have converted this FASTQ data from the original Solexa/Illumina encoding to the Sanger standard

Phred quality score

A quality value Q is an integer mapping of p (i.e., the probability that the corresponding base call is incorrect).

Phred quality score (the standard Sanger variant, assess reliability of a base call):

$$Q_{\text{sanger}} = -10 \log_{10} p$$

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%

基因组学 | 测序 | 数据格式 | FASTQ | Quality | Encoding

```
 SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................................
 ..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX......................
 ............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII..........................
 ..............................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ..........................
 LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL.....................................................
 !"#$%&'()*+,-./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)
```

Quality control

- quality score distribution
- GC content
- k-mer enrichment
- ...

Preprocessing

- adapter removal
- low-quality reads filtering
- ...

Processing

- alignment
- further analysis

Quality control

- quality score distribution
- GC content
- k-mer enrichment
- ...

Preprocessing

- adapter removal
- low-quality reads filtering
- ...

Processing

- alignment
- further analysis

Quality control

- quality score distribution
- GC content
- k-mer enrichment
- ...

Preprocessing

- adapter removal
- low-quality reads filtering
- ...

Processing

- alignment
- further analysis

Headers

Alignments

```
1 @HD VN:1.0  SO:unsorted
2 @SQ SN:gi|110640213|ref|NC_008253.1|    LN:4938920
3 @PG ID:bowtie2  PN:bowtie2  VN:2.1.0
4 gi|110640213|ref|NC_008253.1|_418_952_1:0:0_1:0:0_0/1   0   gi|110640213|ref|NC_
  008253.1|   418 42  70M *   0   0   CCAGGCAGTGGCAGGTGGCCACCGTCCTCTCTGCCCCCGCCAAA
  ATCACCAACCATCTGGTAGCGATGAT  2222222222222222222222222222222222222222222222222
  2222222222222222  AS:i:-3 XN:i:0  XM:i:1  XO:i:0  XG:i:0  NM:i:1  MD:Z:8G61
  YT:Z:UU
5 gi|110640213|ref|NC_008253.1|_31_476_0:0:0_0:0:0_1/1    16  gi|110640213|ref|NC_
  008253.1|   407 42  70M *   0   0   GGAAAGCAATGCCAGGCAGGGGCAGGTGGCCACCGTCCTCTCTG
  CCCCCGCCAAAATCACCAACCATCTG  2222222222222222222222222222222222222222222222222
  2222222222222222  AS:i:0  XN:i:0  XM:i:0  XO:i:0  XG:i:0  NM:i:0  MD:Z:70 YT:Z
  :UU
6 gi|110640213|ref|NC_008253.1|_210_743_2:0:0_1:1:0_2/1   0   gi|110640213|ref|NC_
  008253.1|   210 42  70M *   0   0   CATTACCACCACCATCACCATTACCACAGGAAACGGTGCGGGCT
  GACGCGTACAGGAAACACCGAAAAAA  2222222222222222222222222222222222222222222222222
  2222222222222222  AS:i:-6 XN:i:0  XM:i:2  XO:i:0  XG:i:0  NM:i:2  MD:Z:30T31A7
      YT:Z:UU
```

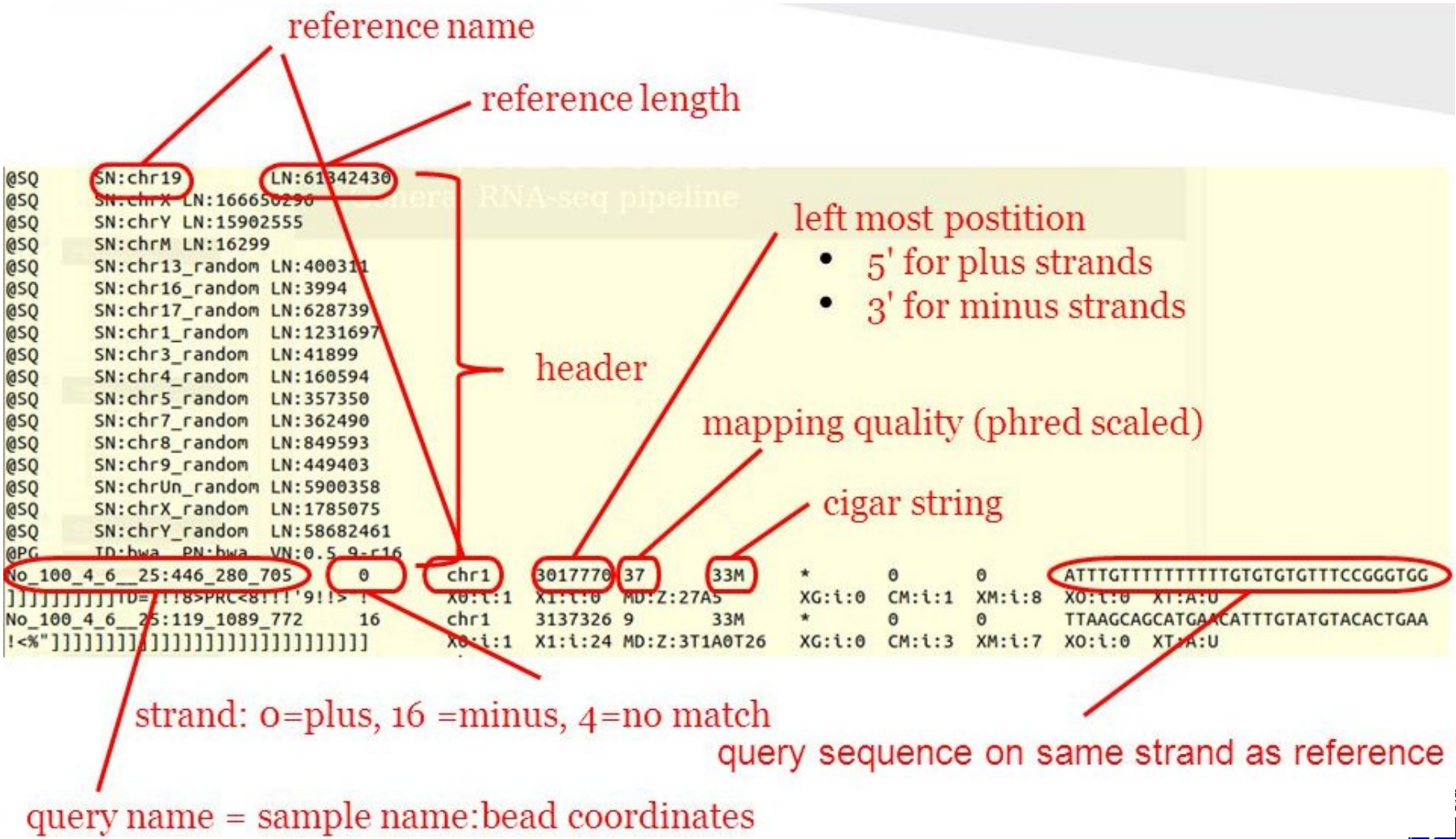
Each row describes a single alignment of a raw read against the reference genome. Each alignment has 11 mandatory fields, followed by any number of optional fields.

SAM Format Specification

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

Col	Field	Type	Regexp/Range	Brief description
1	QNAME	String	[!-?A-~]{1,254}	Query template NAME
2	FLAG	Int	[0,2^{16}-1]	bitwise FLAG
3	RNAME	String	*|[!-()+-<>-~][!-~]*	Reference sequence NAME
4	POS	Int	[0,2^{31}-1]	1-based leftmost mapping POSition
5	MAPQ	Int	[0,2^{8}-1]	MAPping Quality
6	CIGAR	String	*|([0-9]+[MIDNSHPX=])+	CIGAR string
7	RNEXT	String	*|=|[!-()+-<>-~][!-~]*	Ref. name of the mate/next read
8	PNEXT	Int	[0,2^{31}-1]	Position of the mate/next read
9	TLEN	Int	[-2^{31}+1,2^{31}-1]	observed Template LENgth
10	SEQ	String	*|[A-Za-z=.]+	segment SEQuence
11	QUAL	String	[!-~]+	ASCII of Phred-scaled base QUALity+33

```
@HD    VN:1.0
@SQ    SN:chr20 LN:62435964
@RG    ID:L1 PU:SC_1_10 LB:SC_1 SM:NA12891
@RG    ID:L2 PU:SC_2_12 LB:SC_2 SM:NA12891
read_28833_29006_6945 99  chr20 28833 20 10M1D25M = 28993 195 \
       AGCTTAGCTAGCTACCTATATCTTGGTCTTGGCCG <<<<<<<<<<<<<<<<<<<:<9/,&,22;;<<< \
       NM:i:1 RG:Z:L1
read_28701_28881_323b 147 chr20 28834 30 35M      = 28701 -168 \
       ACCTATATCTTGGCCTTGGCCGATGCGGCCTTGCA <<<<<;<<<<7;:<<<6;<<<<<<<<<<7<<<< \
       MF:i:18 RG:Z:L2
```


reference name
reference length
header
left most postition
• 5' for plus strands
• 3' for minus strands
mapping quality (phred scaled)
cigar string
strand: 0=plus, 16 =minus, 4=no match
query sequence on same strand as reference
query name = sample name:bead coordinates
@SQ SN:chr19 LN:61342430
@SQ SN:chrX LN:166650296
@SQ SN:chrY LN:15902555
@SQ SN:chrM LN:16299
@SQ SN:chr13_random LN:400311
@SQ SN:chr16_random LN:3994
@SQ SN:chr17_random LN:628739
@SQ SN:chr1_random LN:1231697
@SQ SN:chr3_random LN:41899
@SQ SN:chr4_random LN:160594
@SQ SN:chr5_random LN:357350
@SQ SN:chr7_random LN:362490
@SQ SN:chr8_random LN:849593
@SQ SN:chr9_random LN:449403
@SQ SN:chrUn_random LN:5900358
@SQ SN:chrX_random LN:1785075
@SQ SN:chrY_random LN:58682461
@PG ID:bwa PN:bwa VN:0.5.9-r16
No_100_4_6__25:446_280_705 0 chr1 3017770 37 33M * 0 0 ATTTGTTTTTTTTTTGTGTGTGTTTCCGGGTGG
]]]]]]]]]]ID=!!!8>PRC<8!!!'9!!> ! X0:i:1 X1:i:0 MD:Z:27A5 XG:i:0 CM:i:1 XM:i:8 XO:i:0 XT:A:U
No_100_4_6__25:119_1089_772 16 chr1 3137326 9 33M * 0 0 TTAAGCAGCATGAACATTTGTATGTACACTGAA
!<%"]]]]]]]]]]]]]]]]]]]]]]]]]]]] X0:i:1 X1:i:24 MD:Z:3T1A0T26 XG:i:0 CM:i:3 XM:i:7 XO:i:0 XT:A:U

Bit		Description
1	0x1	template having multiple segments in sequencing
2	0x2	each segment properly aligned according to the aligner
4	0x4	segment unmapped
8	0x8	next segment in the template unmapped
16	0x10	SEQ being reverse complemented
32	0x20	SEQ of the next segment in the template being reverse complemented
64	0x40	the first segment in the template
128	0x80	the last segment in the template
256	0x100	secondary alignment
512	0x200	not passing filters, such as platform/vendor quality controls
1024	0x400	PCR or optical duplicate
2048	0x800	supplementary alignment

Example

99 = 64 + 32 + 2 + 1

Decoding SAM flags

http://broadinstitute.github.io/picard/explain-flags.html

FLAG meaning in English	FLAG
read paired	1
read mapped in proper pair	2
read unmapped	4
mate unmapped	8
read reverse strand	16
mate reverse strand	32
first in pair	64
second in pair	128
not primary alignment	256
read fails platform/vendor quality checks	512
read is PCR or optical duplicate	1024

Most common flags:
0 (mapped, not paired, forward strand), 4 and 16.

Op	BAM	Description
M	0	alignment match (can be a sequence match or mismatch)
I	1	insertion to the reference
D	2	deletion from the reference
N	3	skipped region from the reference
S	4	soft clipping (clipped sequences present in SEQ)
H	5	hard clipping (clipped sequences NOT present in SEQ)
P	6	padding (silent deletion from padded reference)
=	7	sequence match
X	8	sequence mismatch

CIGAR string

For example:

```
RefPos:     1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19
Reference:  C  C  A  T  A  C  T  G  A  A  C  T  G  A  C  T  A  A  C
Read: ACTAGAATGGCT
```

Aligning these two:

```
RefPos:     1  2  3  4  5  6  7     8  9 10 11 12 13 14 15 16 17 18 19
Reference:  C  C  A  T  A  C  T     G  A  A  C  T  G  A  C  T  A  A  C
Read:                  A  C  T  A  G  A  A     T  G  G  C  T
```

With the alignment above, you get:

```
POS: 5
CIGAR: 3M1I3M1D5M
```

2(222..2!!!!!!!!!!(!!!!!!!!!!!!!!!!!!!!!!!!!!!!	
4422232222222223,333332222111111000000000-.---..	XT:A:U NM:i:0 X0:i:1 X1:i:0 XM:i:0 XO:i:0 XG:i:0 MD:Z:47
00*0000000*00111*1111114112222332+2222+22002244	XT:A:U NM:i:0 X0:i:1 X1:i:0 XM:i:0 XO:i:0 XG:i:0 MD:Z:47
442222220232232333333322222211111111000000000000	XT:A:U NM:i:0 X0:i:1 X1:i:0 XM:i:0 XO:i:0 XG:i:0 MD:Z:47
',,,---'...(000)000001111112+2332)22222022++222	XT:A:U NM:i:1 X0:i:1 X1:i:0 XM:i:1 XO:i:0 XG:i:0 MD:Z:7G39
44220000032222222222211111111110001000000000/	XT:A:U NM:i:0 X0:i:1 X1:i:1 XM:i:0 XO:i:0 XG:i:0 MD:Z:47 XA:Z:chr2,-22444222,47M,1;
44222222222222333322210100000000000/0----.....	XT:A:U NM:i:0 X0:i:1 X1:i:0 XM:i:0 XO:i:0 XG:i:0 MD:Z:47
.-000000000000000011141222233333222322222222244	XT:A:U NM:i:0 X0:i:1 X1:i:0 XM:i:0 XO:i:0 XG:i:0 MD:Z:47

Last, but very important, SAM field is the TAG field

Each TAG has a meaning and summarizes some aspect of the alignment.

Some tags (e.g. NM) have a predefined meaning in the format, NM is the number of mismatches between the read and the template

Other tags (e.g XT) are program specific – XT:A:U/R in BWA tells whether there is one or many “best alignments” for the read.

There are numerous predefined, or program specific tags that convey much useful information about each alignment, and alternative mappings for the reads. These tags are used when you filter alignments based on number of mismatches, or unique versus repeat, etc.

```
Coor     12345678901234  5678901234567890123456789012345
ref      AGCATGTTAGATAA**GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT

+r001/1        TTAGATAAAGGATA*CTG
+r002         aaaAGATAA*GGATA
+r003       gcctaAGCTAA
+r004                       ATAGCT..............TCAGC
-r003                             ttagctTAGGC
-r001/2                                        CAGCGGCAT
```

```
@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
```

BAM: the binary version of SAM

- SAM files are large: 1M short reads => 200MB; 100M short reads => 20GB.
- Makes sense for compression
- BAM: Binary sAM; compress using gzip library.
- Two parts: compressed data + index
- Index: random access (visualization, analysis, etc.)

BED format

- Text-based, tab-delimited format for storing signals for intervals
  - 3 required fields: chrom, chromStart, chromEnd
  - 9 optional fields: name, score, strand, thickStart, thickEnd, itemRgb, blockCount, blockSizes, blockStarts (last ones for visualization)
  - Example:

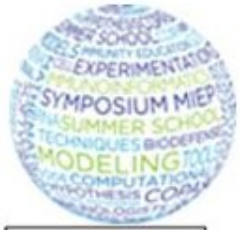
```
chr22 1000 5000 cloneA 960 + 1000 5000 0 2 567,488, 0,3512
chr22 2000 6000 cloneB 900 - 2000 6000 0 2 433,399, 0,3601
```

- There is also a binary format called BigBed with more efficient data access
- Many variations, such as the commonly-used bedGraph format with only 4 fields: chrom, chromStart, chromEnd, dataValue

chr1	817371	819837	ENSG00000177757.2_FAM87B_lincRNA	0	+
chr1	826206	827522	ENSG00000225880.5_LINC00115_lincRNA	0	-
chr1	827608	859446	ENSG00000228794.5_LINC01128_processed_transcript	0	+
chr1	868071	876903	ENSG00000230368.2_FAM41C_lincRNA	0	-
chr1	873292	874349	ENSG00000234711.1_TUBB8P11_unprocessed_pseudogene	0	+
chr1	904834	915976	ENSG00000272438.1_RP11-54O7.16_lincRNA	0	+
chr1	911435	914948	ENSG00000230699.2_RP11-54O7.1_lincRNA	0	+
chr1	914171	914971	ENSG00000241180.1_RP11-54O7.2_lincRNA	0	+
chr1	916865	921016	ENSG00000223764.2_RP11-54O7.3_lincRNA	0	-
chr1	924880	944581	ENSG00000187634.7_SAMD11_protein_coding	0	+

BED

- Developed primarily for the UCSC genome browser
- Used to store annotations on genomic coordinates
  - Annotate gene/mRNA/exon/... position
  - Annotate Transcription Factor binding sites
  - Annotate SNP genotypes
  - Annotate Gene Expression
  - Annotate ...



File formats – GTF

- GTF format
  - **G**ene **T**ransfer **F**ormat
  - Widely used format for annotated genome and transcriptome
  - Downloadable from major browser sites, e.g. UCSC, Ensembl, NCBI
  - Illumina also provides a set of annotated genomes: igenomes
    - Available through Galaxy and command line

Seqname	Source	feature	start	end	score	strand	frame	attributes
chr1	unknown	exon	3204563	3207049	.	-	.	gene_id "Xkr4"; transcript_id "NM_001011874";

Seqname	Source	Feature	Start	End	Score	Strand	Frame	Attributes
chr4	protein_coding	CDS	24053	24477	.	+	0	exon_number "1"; gene_id "FBgn0040037"; gene_name "JYalpha"; p_
chr4	protein_coding	exon	24053	24477	.	+	.	exon_number "1"; gene_id "FBgn0040037"; gene_name "JYalpha"; p_
chr4	protein_coding	CDS	24979	25153	.	+	1	exon_number "2"; gene_id "FBgn0040037"; gene_name "JYalpha"; p_
chr4	protein_coding	exon	24979	25153	.	+	.	exon_number "2"; gene_id "FBgn0040037"; gene_name "JYalpha"; p_
chr4	protein_coding	CDS	25218	25450	.	+	0	exon_number "3"; gene_id "FBgn0040037"; gene_name "JYalpha"; p_
chr4	protein_coding	exon	25218	25450	.	+	.	exon_number "3"; gene_id "FBgn0040037"; gene_name "JYalpha"; p_
chr4	protein_coding	CDS	25501	25618	.	+	1	exon_number "4"; gene_id "FBgn0040037"; gene_name "JYalpha"; p_
chr4	protein_coding	exon	25501	25621	.	+	.	exon_number "4"; gene_id "FBgn0040037"; gene_name "JYalpha"; p_
chr4	protein_coding	stop_codon	25619	25621	.	+	0	exon_number "4"; gene_id "FBgn0040037"; gene_name "JYalpha"; p_
chr4	pseudogene	exon	26994	27101	.	-	.	exon_number "7"; gene_id "FBgn0052011"; gene_name "CR32011";
chr4	pseudogene	exon	27167	27349	.	-	.	exon_number "6"; gene_id "FBgn0052011"; gene_name "CR32011";
chr4	pseudogene	exon	28371	28609	.	-	.	exon_number "5"; gene_id "FBgn0052011"; gene_name "CR32011";

GFF: a standard annotation format

- Stands for:
  - Gene Finding Format -or- General Feature Format
- Designed as a single line record for describing features on DNA sequence -- originally used for gene prediction output
- 9 tab-delimited fields common to all versions
  - seq source feature begin end score strand frame group
- The group field differs between versions, but in every case no tabs are allowed
  - GFF2: group is a unique description, usually the gene name.
    - `NCOA1`
  - GFF2.5 / GTF (Gene Transfer Format): tag-value pairs introduced, start_codon and stop_codon are required features for CDS
    - `transcript_id "NM_056789" ; gene_id "NCOA1"`
  - GFF3: Capitalized tags follow Sequence Ontology (SO) relationships, FASTA seqs can be embedded
    - `ID=NM_056789_exon1; Parent=NM_056789; note="5' UTR exon"`

```
ctg123 example gene             1050 9000 . + . ID=EDEN;Name=EDEN;Note=protein kinase

ctg123 example mRNA             1050 9000 . + . ID=EDEN.1;Parent=EDEN;Name=EDEN.1;Index=1
ctg123 example five_prime_UTR   1050 1200 . + . Parent=EDEN.1
ctg123 example CDS              1201 1500 . + 0 Parent=EDEN.1
ctg123 example CDS              3000 3902 . + 0 Parent=EDEN.1
ctg123 example CDS              5000 5500 . + 0 Parent=EDEN.1
ctg123 example CDS              7000 7608 . + 0 Parent=EDEN.1
ctg123 example three_prime_UTR  7609 9000 . + . Parent=EDEN.1

ctg123 example mRNA             1050 9000 . + . ID=EDEN.2;Parent=EDEN;Name=EDEN.2;Index=1
ctg123 example five_prime_UTR   1050 1200 . + . Parent=EDEN.2
ctg123 example CDS              1201 1500 . + 0 Parent=EDEN.2
ctg123 example CDS              5000 5500 . + 0 Parent=EDEN.2
ctg123 example CDS              7000 7608 . + 0 Parent=EDEN.2
ctg123 example three_prime_UTR  7609 9000 . + . Parent=EDEN.2

ctg123 example mRNA             1300 9000 . + . ID=EDEN.3;Parent=EDEN;Name=EDEN.3;Index=1
ctg123 example five_prime_UTR   1300 1500 . + . Parent=EDEN.3
ctg123 example five_prime_UTR   3000 3300 . + . Parent=EDEN.3
ctg123 example CDS              3301 3902 . + 0 Parent=EDEN.3
ctg123 example CDS              5000 5500 . + 1 Parent=EDEN.3
ctg123 example CDS              7000 7600 . + 1 Parent=EDEN.3
ctg123 example three_prime_UTR  7601 9000 . + . Parent=EDEN.3
```

Feature formats: **GFF3 vs. GTF**

✧ GFF3 – Gene feature format

```
Chr1  amel_OGSv3.1  gene   204921  223005  .  +  .  ID=GB42165
Chr1  amel_OGSv3.1  mRNA   204921  223005  .  +  .  ID=GB42165-RA;Parent=GB42165
Chr1  amel_OGSv3.1  3'UTR  222859  223005  .  +  .  Parent=GB42165-RA
Chr1  amel_OGSv3.1  exon   204921  205070  .  +  .  Parent=GB42165-RA
Chr1  amel_OGSv3.1  exon   222772  223005  .  +  .  Parent=GB42165-RA
```

✧ GTF – Gene transfer format

```
AB000381 Twinscan CDS         380  401  .  +  0  gene_id "001"; transcript_id "001.1";
AB000381 Twinscan CDS         501  650  .  +  2  gene_id "001"; transcript_id "001.1";
AB000381 Twinscan CDS         700  707  .  +  2  gene_id "001"; transcript_id "001.1";
AB000381 Twinscan start_codon 380  382  .  +  0  gene_id "001"; transcript_id "001.1";
AB000381 Twinscan stop_codon  708  710  .  +  0  gene_id "001"; transcript_id "001.1";
```

Always check which of the two formats is accepted by your application of choice, sometimes they cannot be swapped

BED: zero based, start inclusive, stop exclusive

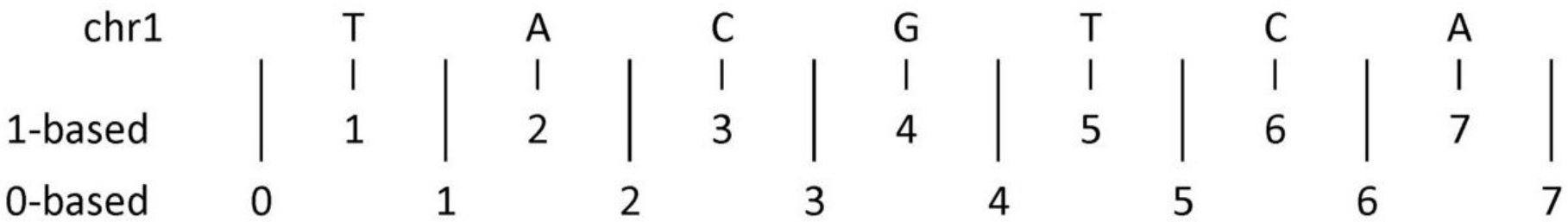
```
chr1 10491     10492     rs55998931     0     +
chr1 10582     10583     rs58108140     0     +
```

- First base on the chromosome is 0
- Length = stop - start

GTF/GFF: one based, inclusive

```
chr1 snp135Com exon 10492     10492     0.000
chr1 snp135Com exon 10583     10583     0.000
```

- First base on the chromosome is 1
- Length = stop – start+1



	1-based	0-based
Indicate a single nucleotide	chr1:4-4 G	chr1:3-4 G
Indicate a range of nucleotides	chr1:2-4 ACG	chr1:1-4 ACG
Indicate a single nucleotide variant	chr1:5-5 T/A	chr1:4-5 T/A

VCF files

- There is a file format defined for genetic variants called VCF (Variant Call Format).
  - Specification available at http://www.1000genomes.org/wiki/Analysis/Variant%20Call%20Format/vcf-variant-call-format-version-41
  - Two main sections: header and content
  - Header provides basic information of the file, and defines content attributes and filters
  - Each line in the content section represents one variant in one or more samples

VCF format

- The Variant Call Format (VCF) is the emerging standard for storing variant data.
- Originally designed for SNPs and short INDELs, it also works for structural variations.

- VCF consists of a header section and a data section.
- The header must contain a line starting with one '#', showing the name of each field, and then the sample names starting at the 10th column.
- The data section is TAB delimited with each line consisting of at least 8 mandatory fields

The FORMAT field and sample information are allowed to be absent.

Col	Field	Description
1	CHROM	Chromosome name
2	POS	1-based position. For an indel, this is the position preceding the indel.
3	ID	Variant identifier. Usually the dbSNP rsID.
4	REF	Reference sequence at POS involved in the variant. For a SNP, it is a single base.
5	ALT	Comma delimited list of alternative sequence(s).
6	QUAL	Phred-scaled probability of all samples being homozygous reference.
7	FILTER	Semicolon delimited list of filters that the variant fails to pass.
8	INFO	Semicolon delimited list of variant information.
9	FORMAT	Colon delimited list of the format of individual genotypes in the following fields.
10+	Sample(s)	Individual genotype information defined by FORMAT.

Example .VCF file

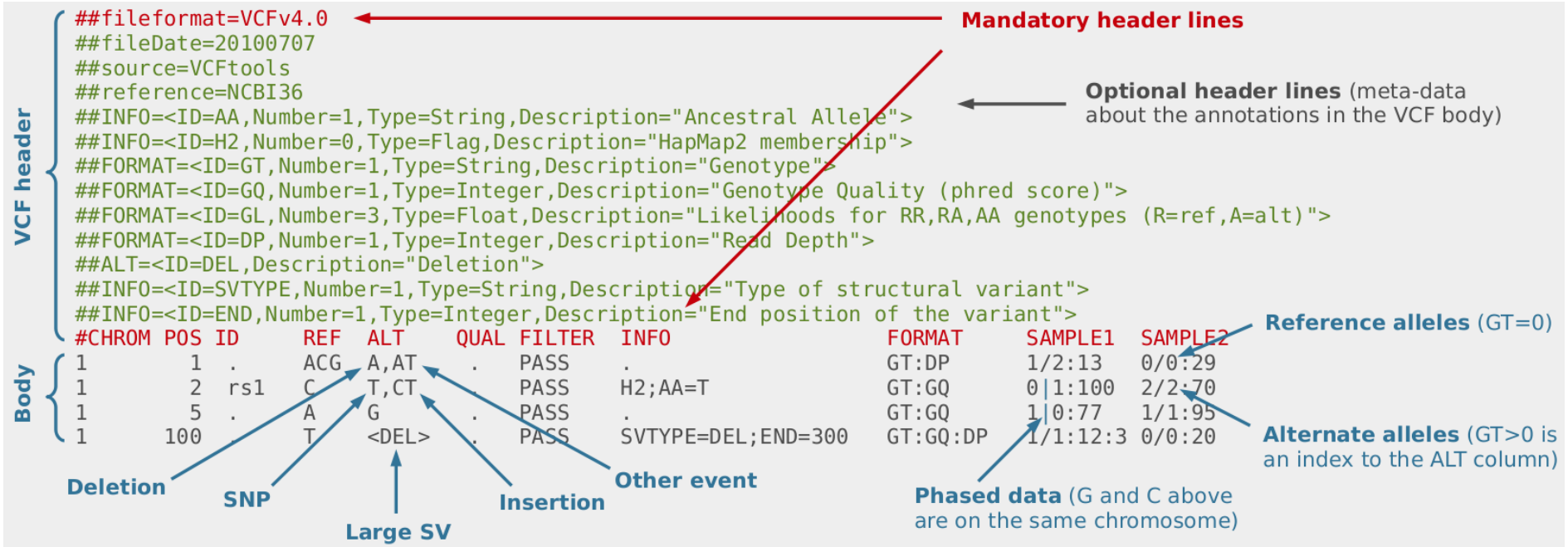
Header lines
(marked by ##):
Metadata of analysis

```
##fileformat=VCFv4.1
##fileDate=20090805
##source=myProgramV3
##reference=file:///seq/NCBI36.fasta
...
#CHROM  POS     ID          REF   ALT   QUAL   FILTER
20      14370   rs6054257   G     A     29     PASS
20      17330   .           T     A     3      q10

INFO                              FORMAT      SAMPLE1     ...
NS=2;DP=14;AF=0.5;DB;H2           GT:GQ:DP    1|0:48:8
NS=2;DP=11;AF=0.017               GT:GQ:DP    0|0:49:3
```

Data lines:
Individual variant calls

基因组学 | 测序 | 数据格式 | VCF



```
##fileformat=VCFv4.1
##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
```

教学提纲

教学提纲

知识点

- 基因组学：人类基因组计划，相关学科
- 测序技术：发展历史，三代测序技术的主要原理
- 外显子组测序：外显子组，实验步骤，主要应用
- 数据库：SRA、GEO
- 数据格式：FASTQ、SAM、BED、GFF、VCF
- 测序数据分析：常见术语，主要步骤，常用工具

技能

- 能够对测序数据进行质控和预处理
- 能够对测序数据进行完整分析
- 能够掌握常见测序数据分析软件的使用方法

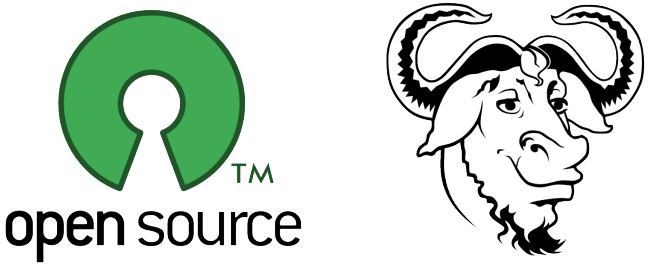
教学提纲

1. 根据自己的理解对人类基因组计划进行评价。
2. 简述 Sanger 测序法的原理。
3. 列举第二代测序方法的主要技术。
4. 简述 Illumina/Solexa 测序的基本过程。
5. 列举第三测序方法的主要技术。
6. 简述外显子组测序的流程和应用。
7. 根据实例解释 FASTQ 格式。
8. 根据实例解释 BED、GFF 和 VCF 格式。
9. 解释测序深度和覆盖度。
10. 简述测序数据分析的主要步骤。
11. 列举测序数据分析的常用工具并进行简介。

下节预告

- 回顾 DNA 测序的实验方法和数据分析步骤。
- 回顾表达芯片的实验过程和数据分析步骤。

Powered by



TEX LATEX XETEX Beamer