What does it take to build and run a pipeline?
A CGC- Exome example

A Healthy Evaluation of the process

Durga Addepalli
NCI / Attain LLC
March 16th 2017

NIH NATIONAL CANCER INSTITUTE
Center for Biomedical Informatics
& Information Technology

NCI Cancer Genomics Cloud Pilots

Data
- Access and analyze 11,000 TCGA samples without having to download data
- Upload your own data for analysis

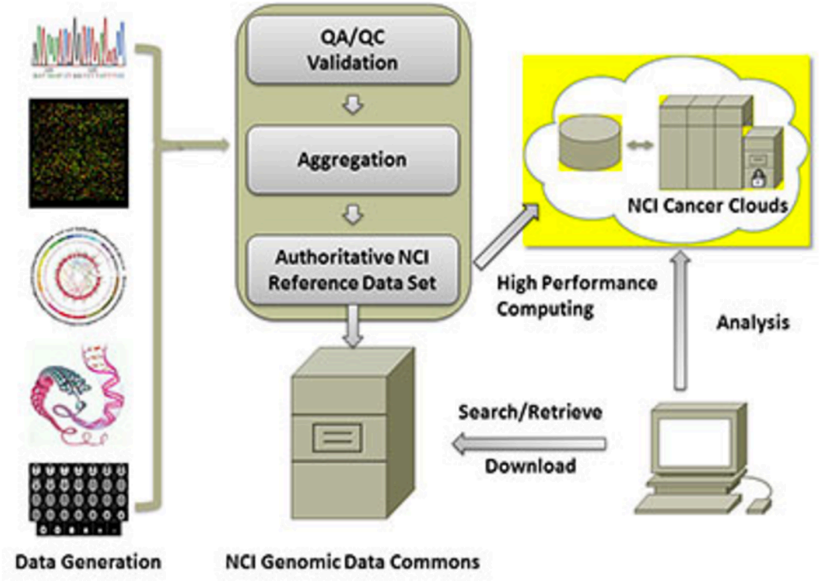
Compute
- Perform large scale analysis using the elastic compute power of commercial cloud platforms

Security
- dbGaP-authorized users can access controlled TCGA data
- Systems meet strict Federal security guidelines

Democratize access to NCI-generated genomic and related data, and to create a cost-effective way to provide scalable computational capacity to the cancer research community.

#NCICloud

Cloud Pilots provides a researcher:
- Access to large genomic data sets without need to download
- Access to popular pipelines and visualization tools
- Ability to bring their own tools and pipelines to the data
- Ability to bring their own data and co-analyze with existing 'omics' data
- Workspaces to save and share their data and results of analyses



Three NCI Genomics Cloud Pilot Awardees

Broad Institute

- PI: Gad Getz
- Google Cloud
- Firehose in the cloud & Broad best practices workflows

Institute for Systems Biology

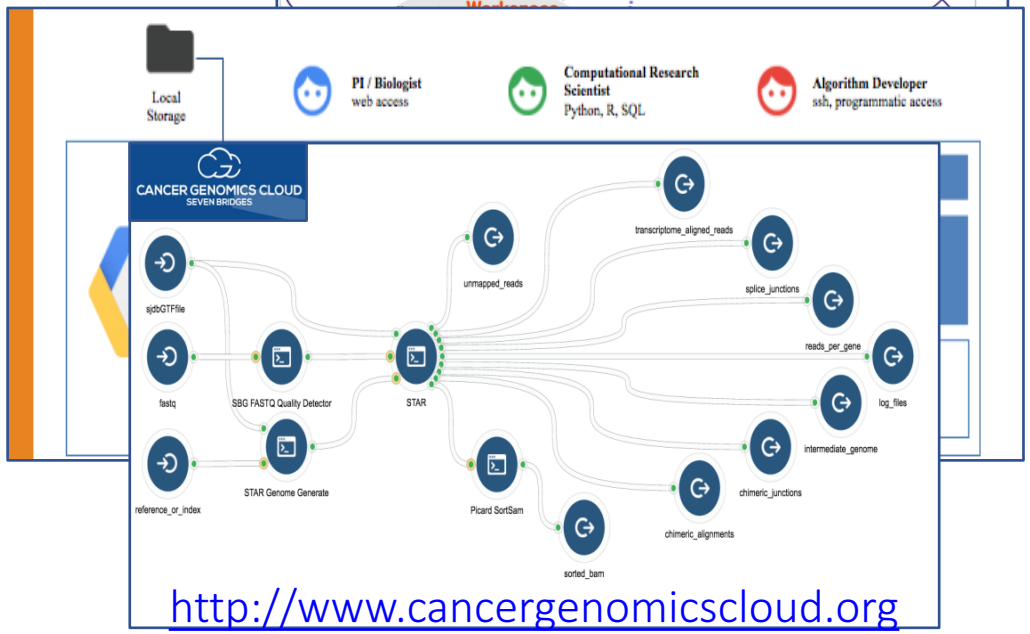
- PI: Ilya Shmulevich
- Google Cloud
- Leverage Google infrastructure; Novel query & visualization

Seven Bridges Genomics

- PI: Deniz Kural
- Amazon Web Services
- Interactive data exploration; > 200 tools & workflows

http://firecloud.org

https://isb-cgc.appspot.com/

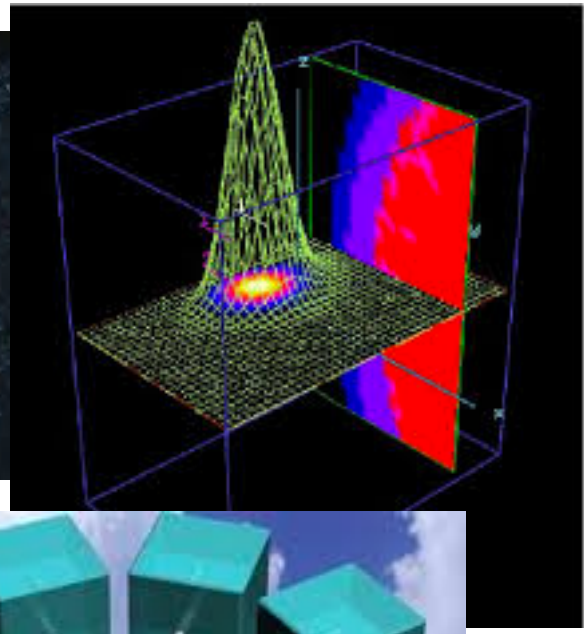


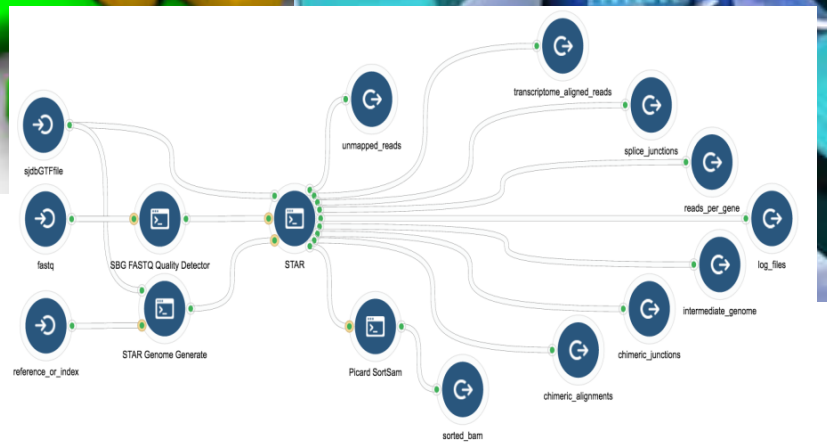
http://www.cancergenomicscloud.org

Jan 2014	Sept 2014	April 2015	Jan 2016	Sept 2016
Selection	Design/Build I	Design/Build II	Evaluation	Extension

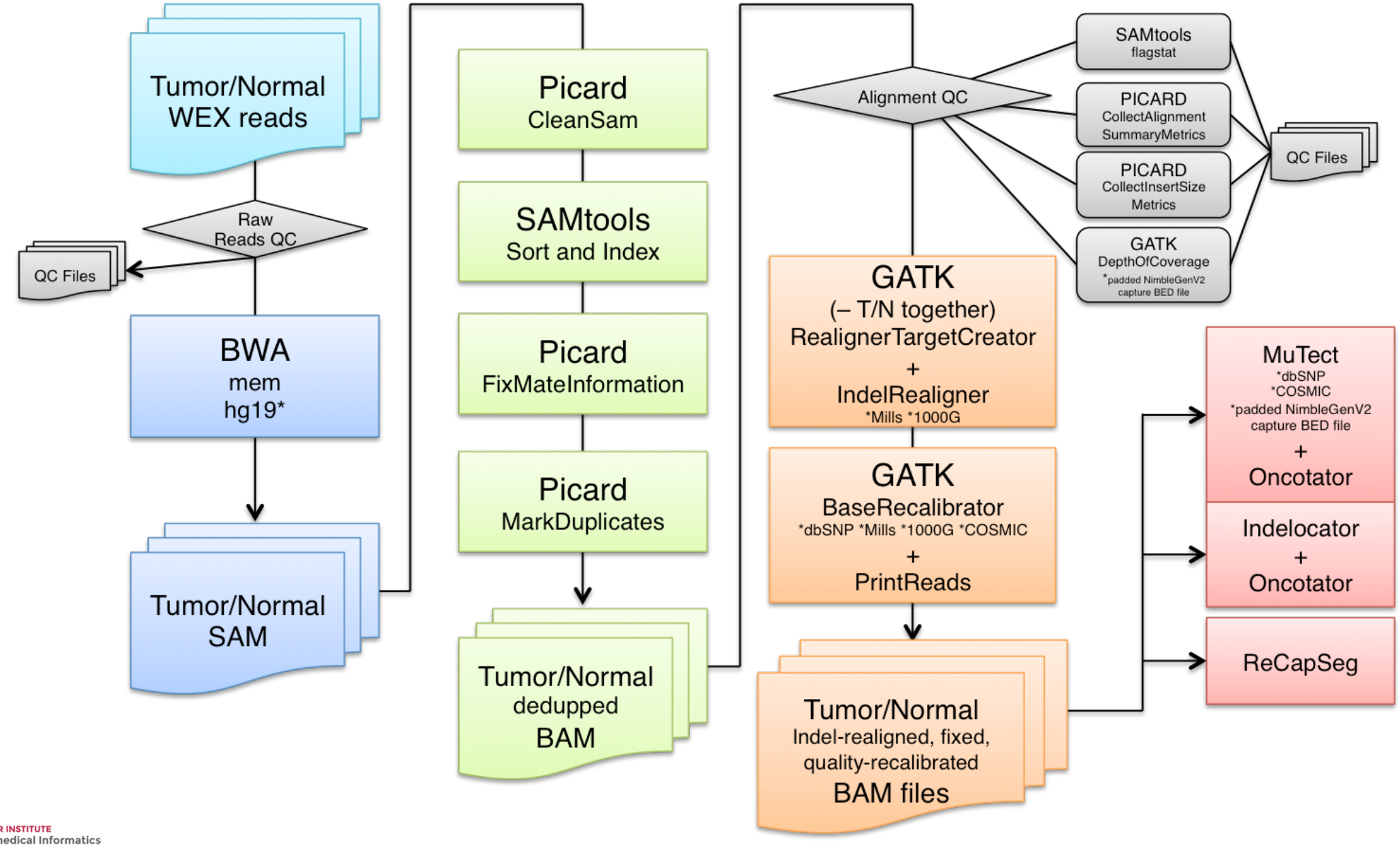
Genomics on the CGC – Advantages

- Massive data storage and protection
- Secure and scalable computational resources
- Unified repositories
- Managed sharing and collaboration
  - No need for copies, multiple uploads/downloads
- Ready to-use tools and **pipelines/workflows**
  - No installation, configuration or deployment time or effort
  - Edit the existing pipeline - Add new tools / Wrap new tools

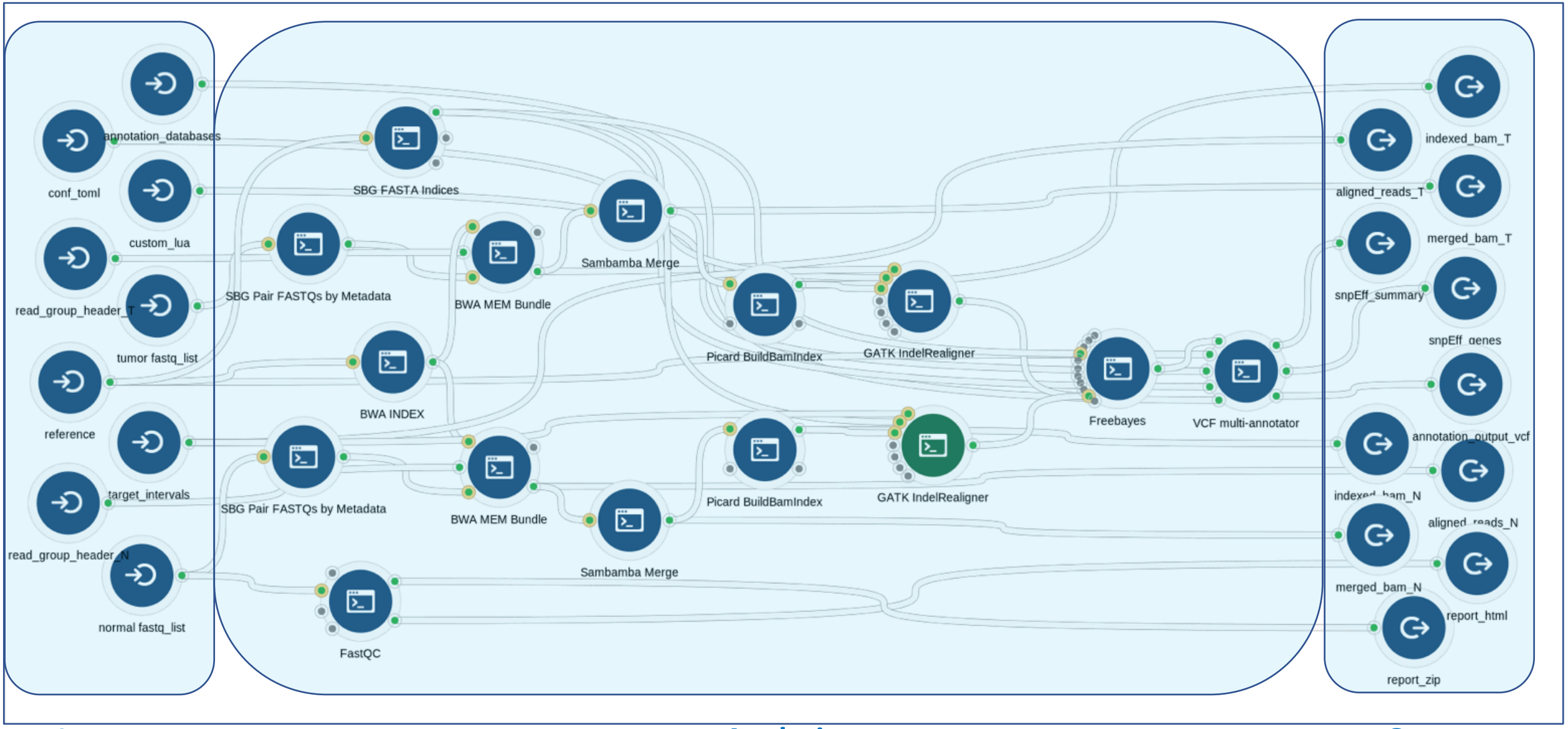




GATK: Tumor-Normal Paired Exome-Sequencing pipeline

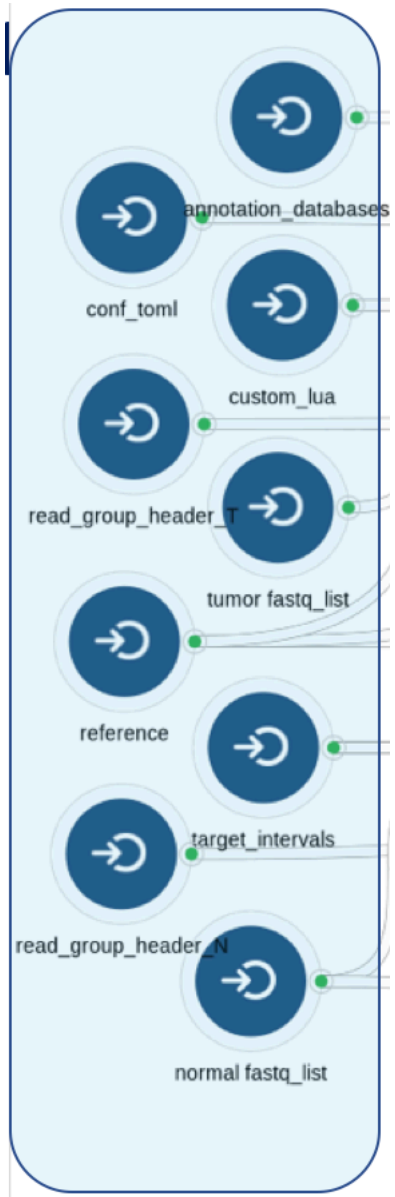


Exome Sequencing Analysis Workflow on SBG



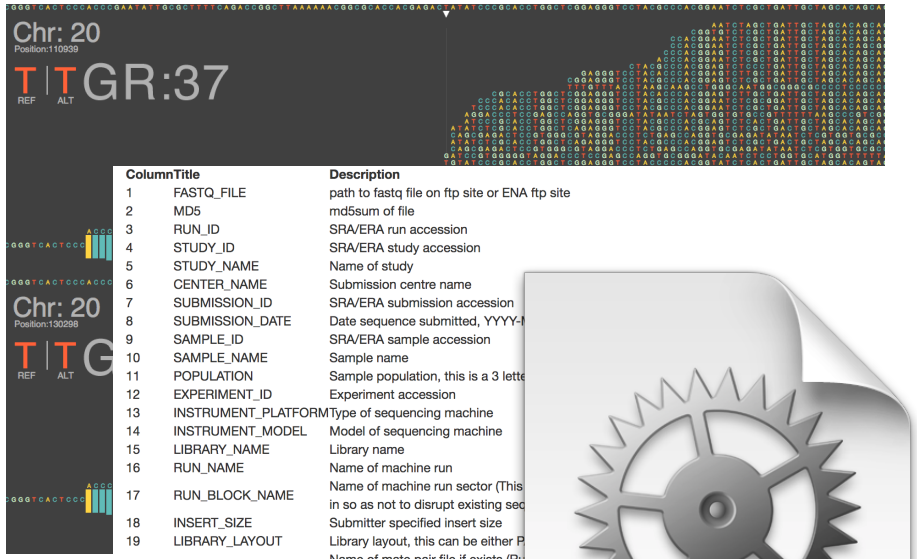
Inputs **Analysis** **Outputs**

What did I need to build the pipeline?

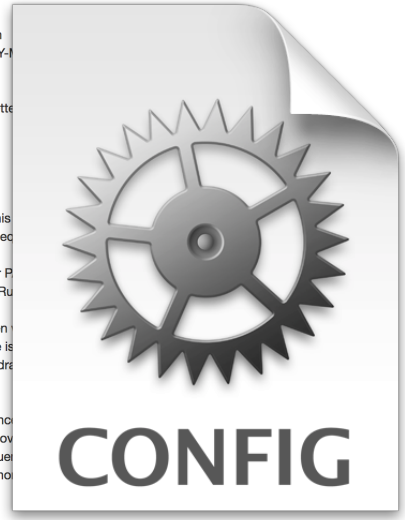


Input Supplementary files

- Reference files
- Index files
- Configuration files
- Target Intervals
- Database files



ColumnTitle		Description
1	FASTQ_FILE	path to fastq file on ftp site or ENA ftp site
2	MD5	md5sum of file
3	RUN_ID	SRA/ERA run accession
4	STUDY_ID	SRA/ERA study accession
5	STUDY_NAME	Name of study
6	CENTER_NAME	Submission centre name
7	SUBMISSION_ID	SRA/ERA submission accession
8	SUBMISSION_DATE	Date sequence submitted, YYYY-
9	SAMPLE_ID	SRA/ERA sample accession
10	SAMPLE_NAME	Sample name
11	POPULATION	Sample population, this is a 3 lette
12	EXPERIMENT_ID	Experiment accession
13	INSTRUMENT_PLATFORM	Type of sequencing machine
14	INSTRUMENT_MODEL	Model of sequencing machine
15	LIBRARY_NAME	Library name
16	RUN_NAME	Name of machine run
17	RUN_BLOCK_NAME	Name of machine run sector (This in so as not to disrupt existing sec
18	INSERT_SIZE	Submitter specified insert size
19	LIBRARY_LAYOUT	Library layout, this can be either P
20	PAIRED_FASTQ	Name of mate pair file if exists (Ru no paired fastq file)
21	WITHDRAWN	0/1 to indicate if the file has been
22	WITHDRAWN_DATE	This is generally the date the file is
23	COMMENT	comment about reason for withdra
24	READ_COUNT	read count for the file
25	BASE_COUNT	basepair count for the file
26	ANALYSIS_GROUP	the analysis group of the sequenc Project data, this includes low cov two non low coverage pilot seque strategies used by the 1000 Geno



COSMIC
Cat

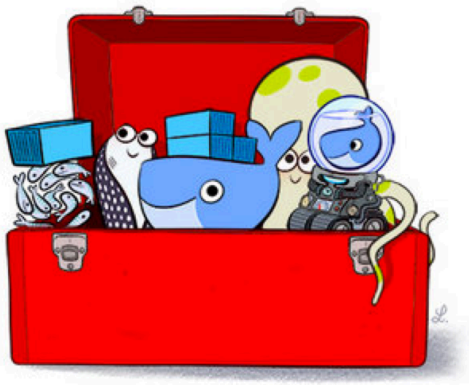
Clin
Clinically relevant

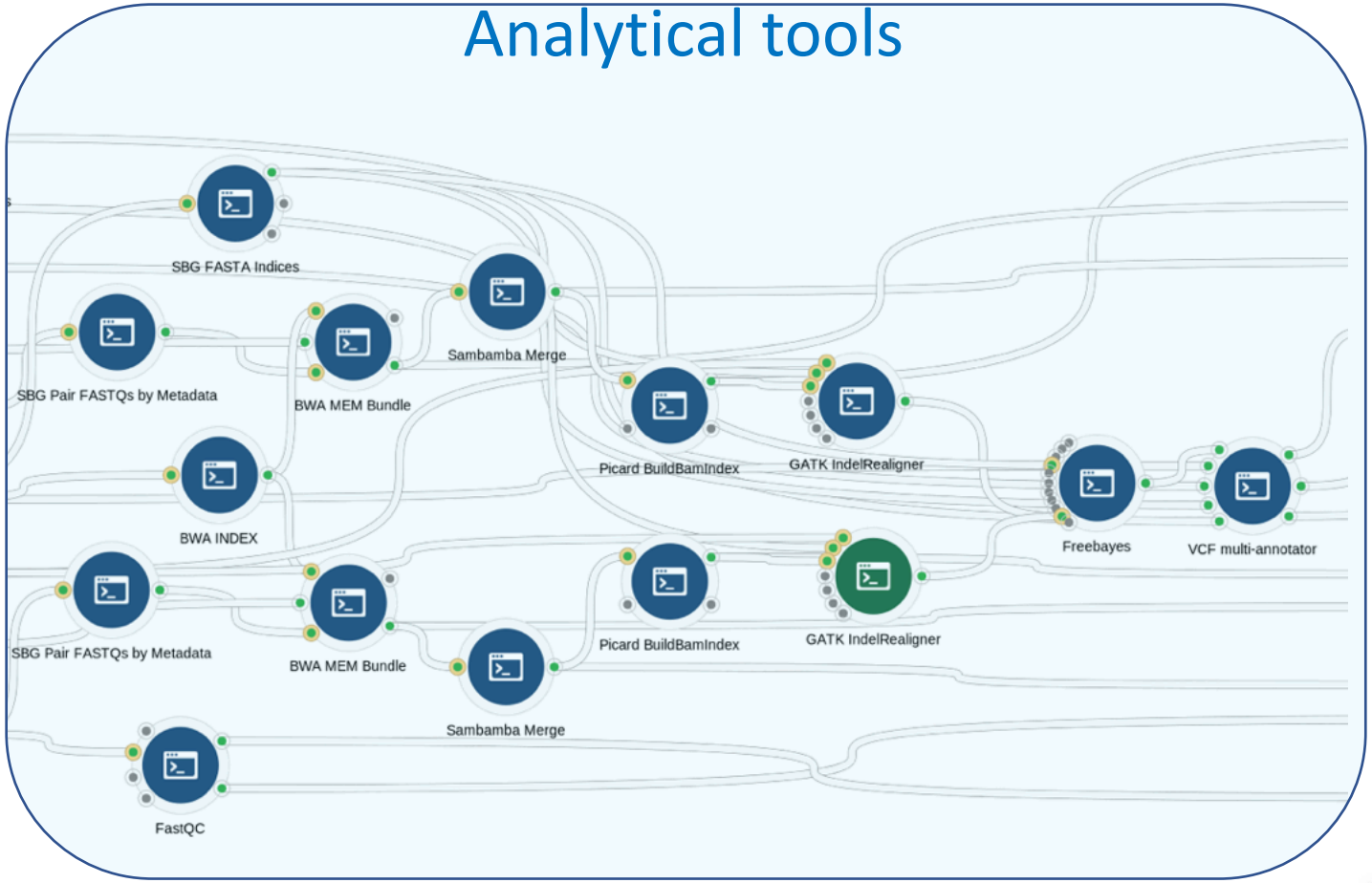
dbSNP
Organisms
dbVar

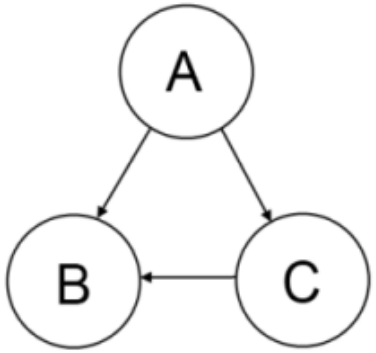
Bioinformatics knowledge required ?

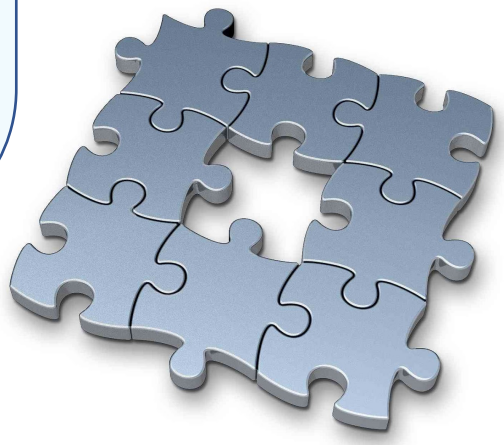
What did I need to build the pipeline?

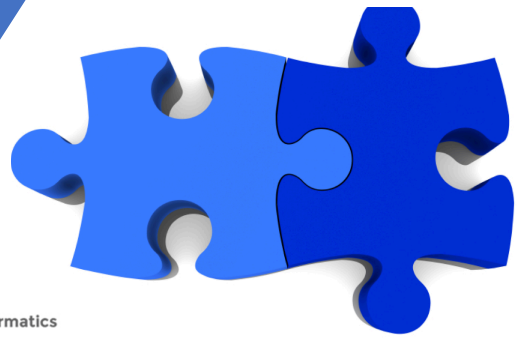
Analytical tools

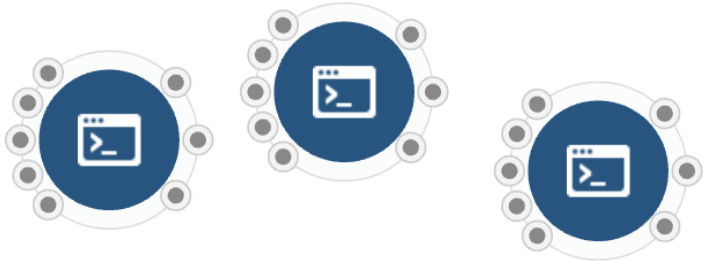






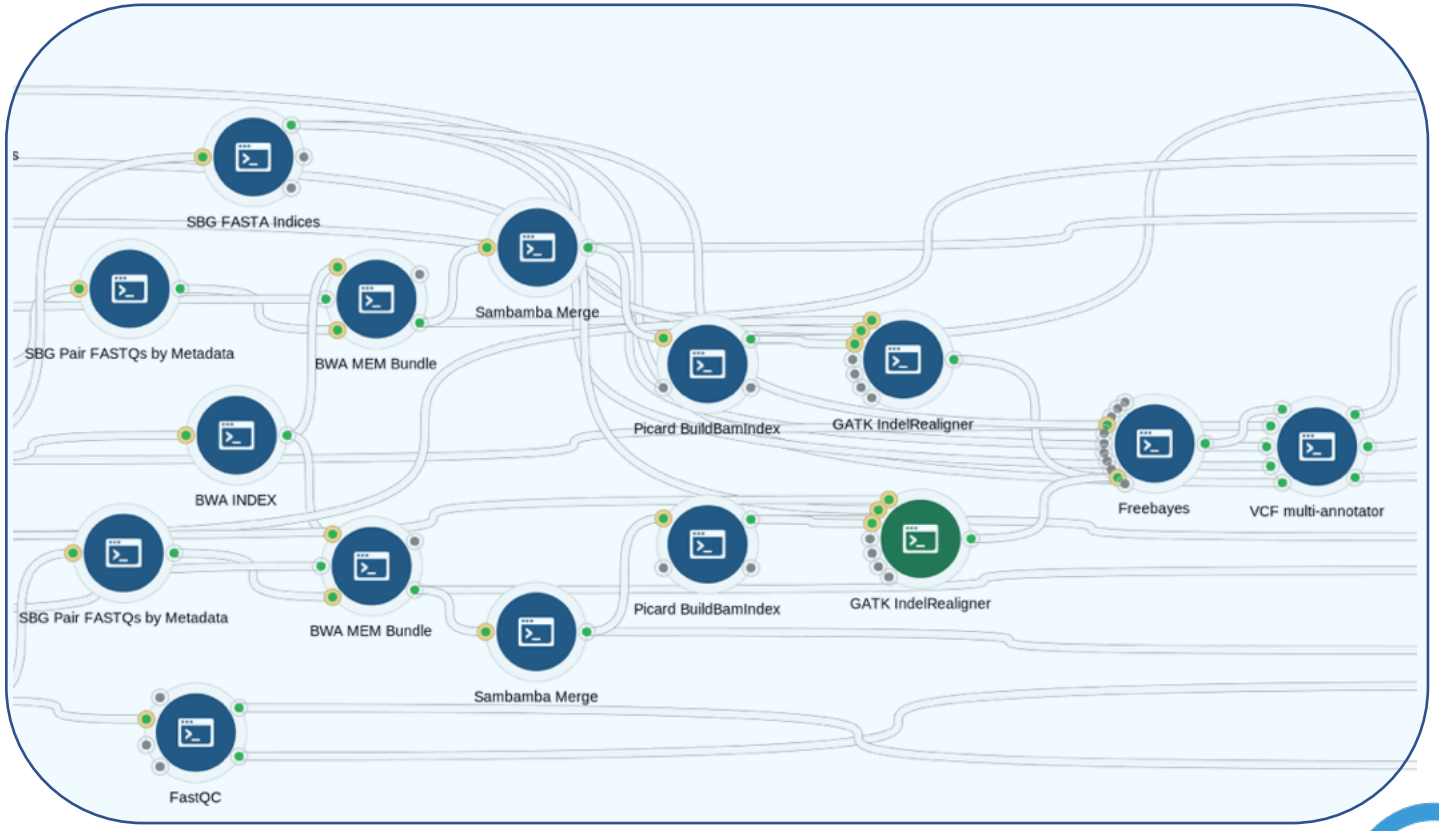




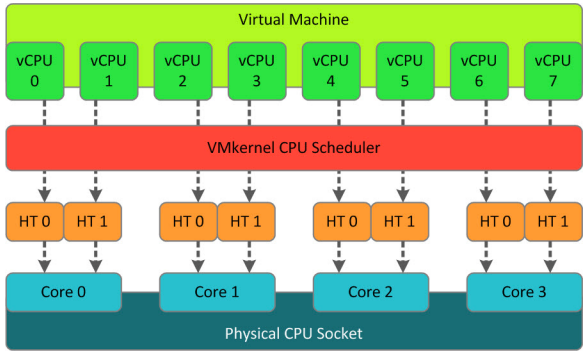


Bioinformatics knowledge required ?

What did I need to build the pipeline?



- Submit a task
  - per sample?
  - per file?
- Process in batches
- Not all steps can be parallelized
- Right instance type
- Multiple cores, Memory, Storage
- Scatter

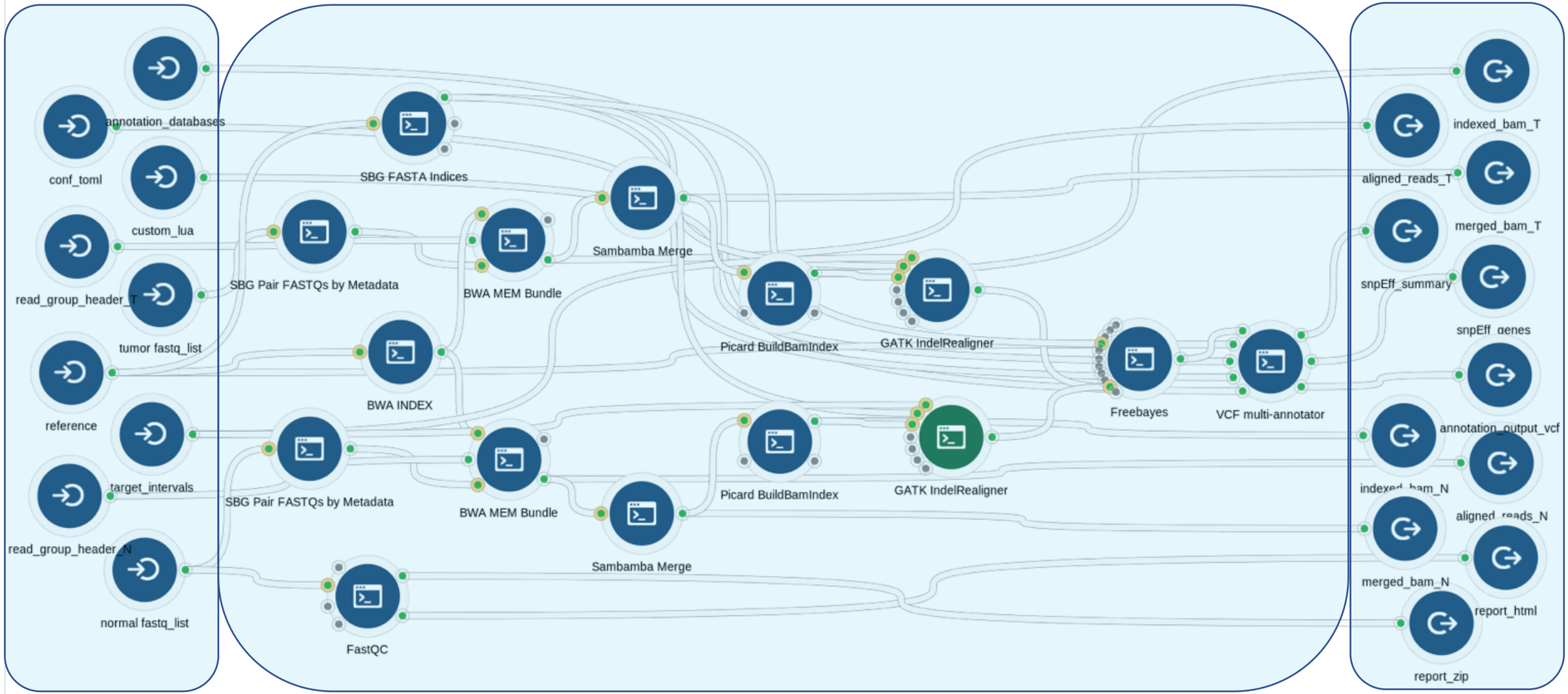


Bioinformatics knowledge required ?

Exome Sequencing Analysis Workflow on SBG

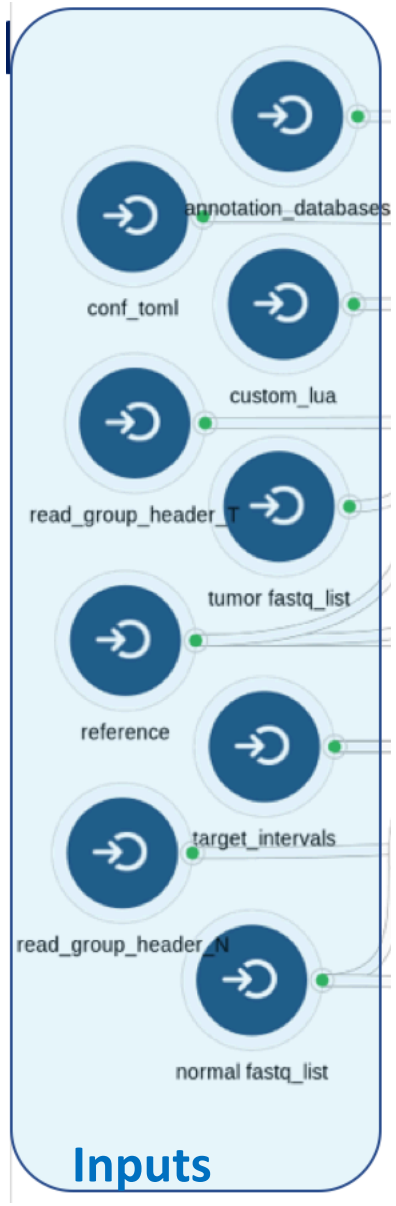
Can’t share with support team unless they have access to controlled data

What did I need to RUN the pipeline?



Inputs **Analysis** **Outputs**

What did I need to RUN the pipeline?

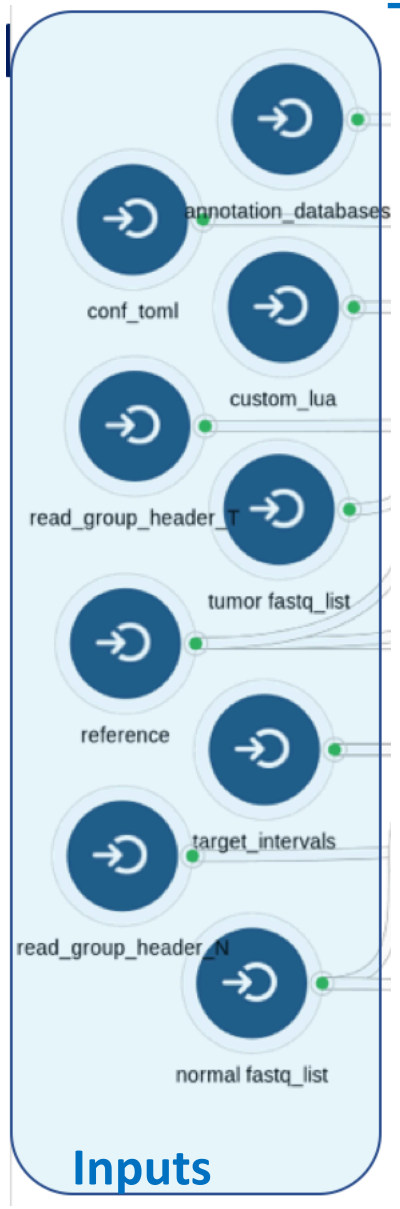


- If you are not working with TCGA or CCLE, you still need to upload your data!
- Data Transfer
  - a separate individual task by itself
  - not straight forward

Can I upload data using any data transfer protocol?

Example: T/N Exome data from TCRB, available only via SFTP

What did I need to RUN the pipeline?



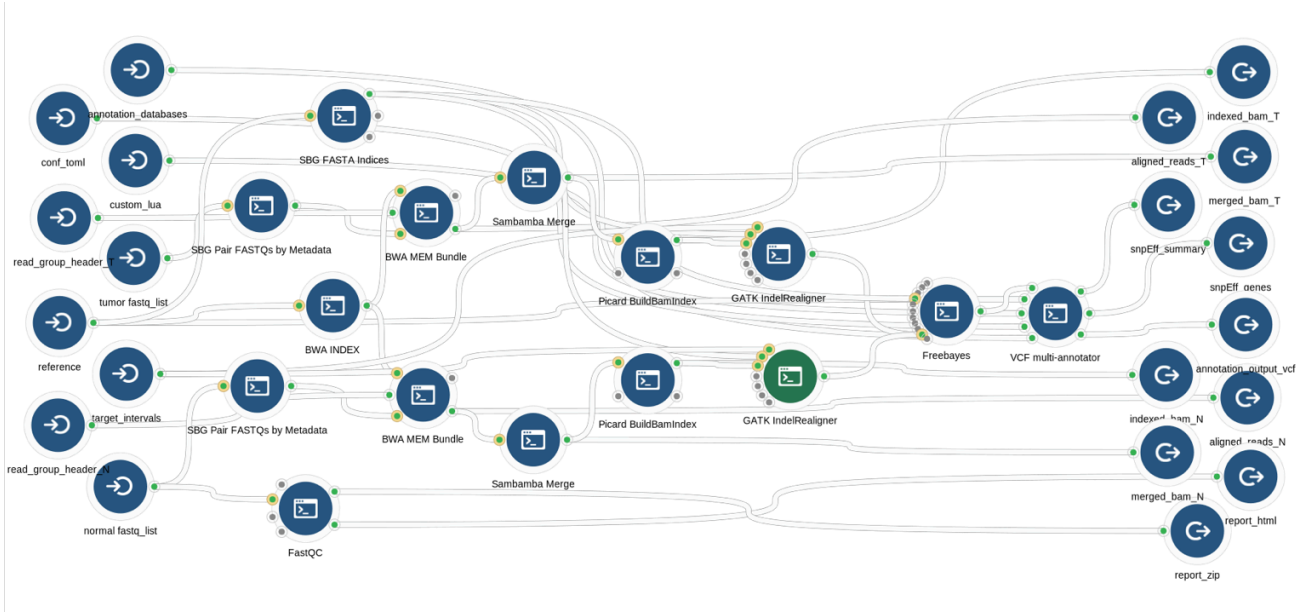
Do I need any file processing prior to input to the pipeline?

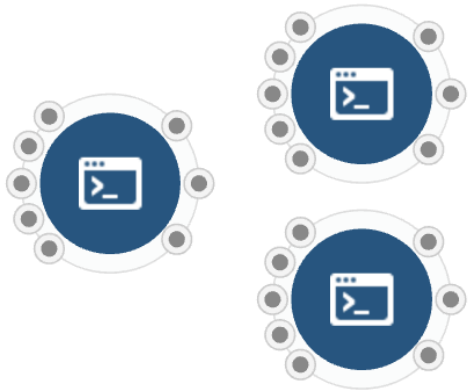
- Compressed data files
- Format Conversion
- conversion of genome coordinates
- Filters

How will the pipeline identify the T/N files for each sample??

- Metadata/Manifest file
- Read Group Header

What did I need to RUN the pipeline?

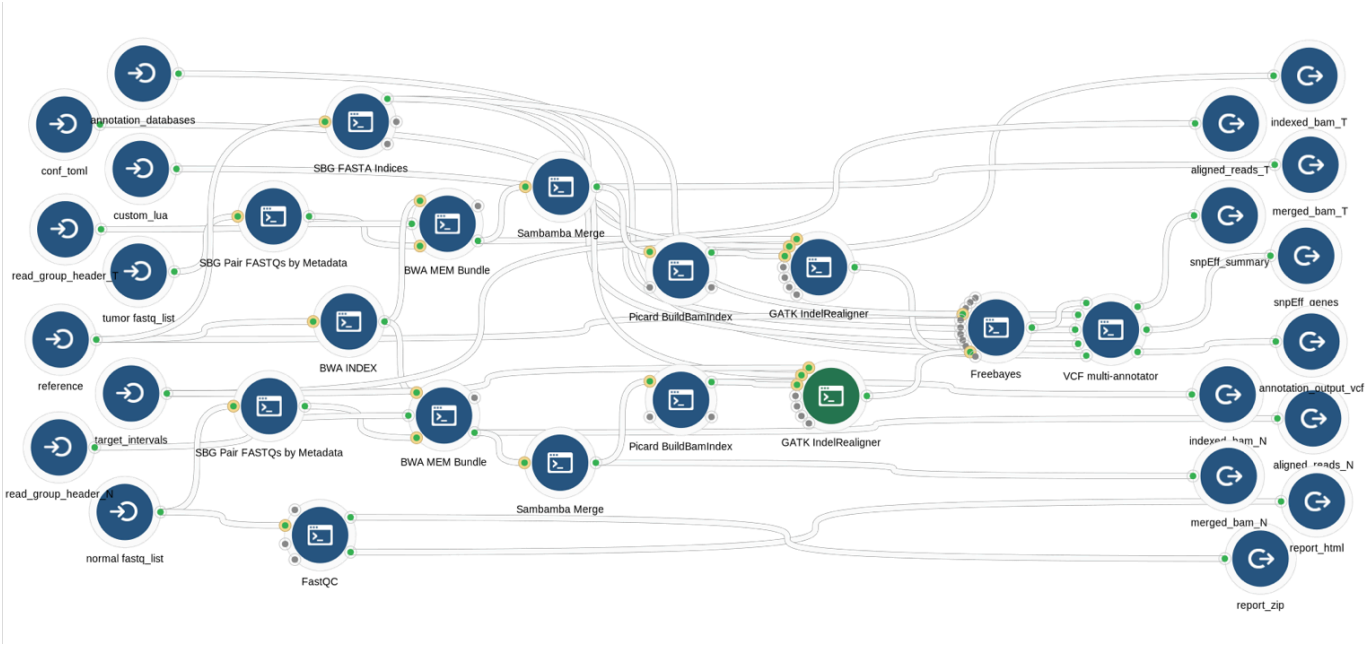




Can you use the pre-configured pipelines right away?

- Does the pipeline do **exactly what** I want to do in my analysis?
  - Edit parameters
- Does the pipeline do **everything** I want to do in my analysis?
  - Add more tools / workflows

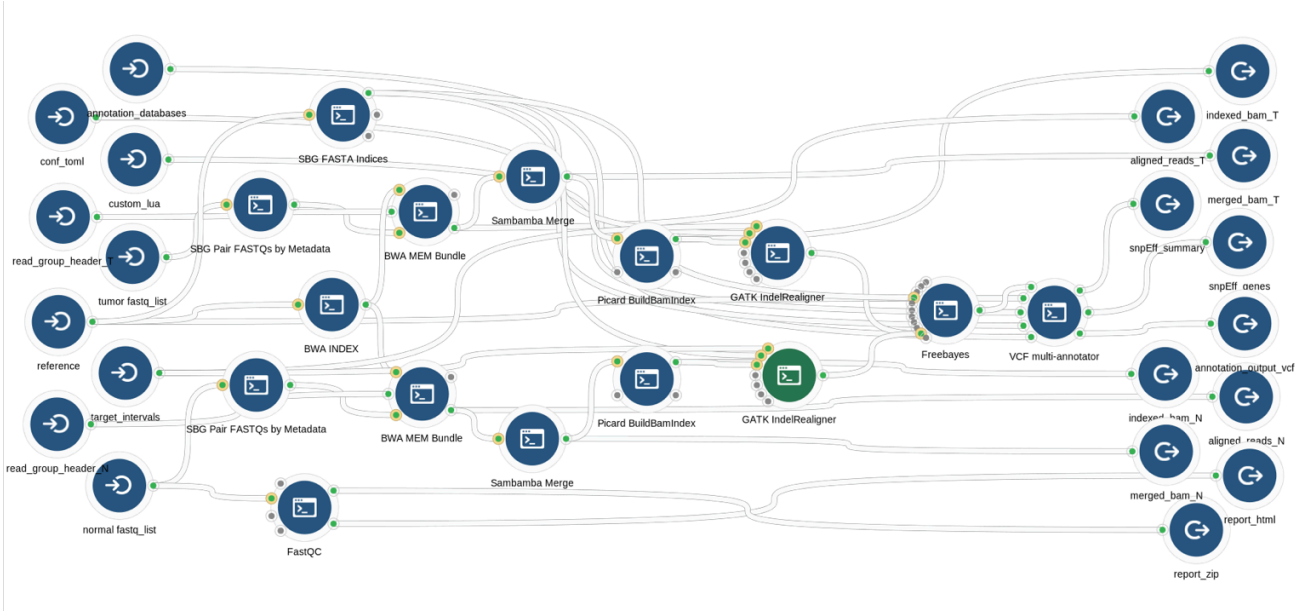
What did I need to RUN the pipeline?



How much would it cost to run the pipeline, do I have enough credits?

- Credits module/support
- Pre-Estimate
- Extra charges
  - Preprocessing
  - Format conversions
  - New tools added to the pipeline

What did I need to RUN the pipeline?



Can I take this pipeline and run it on a different cloud set-up ?

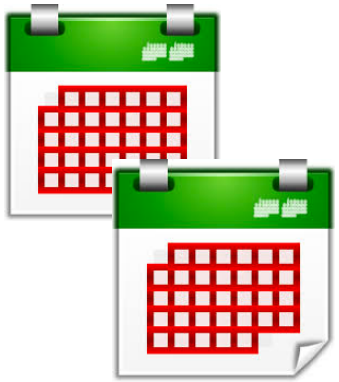
- Interoperability
- Portability

On going efforts working towards interoperability

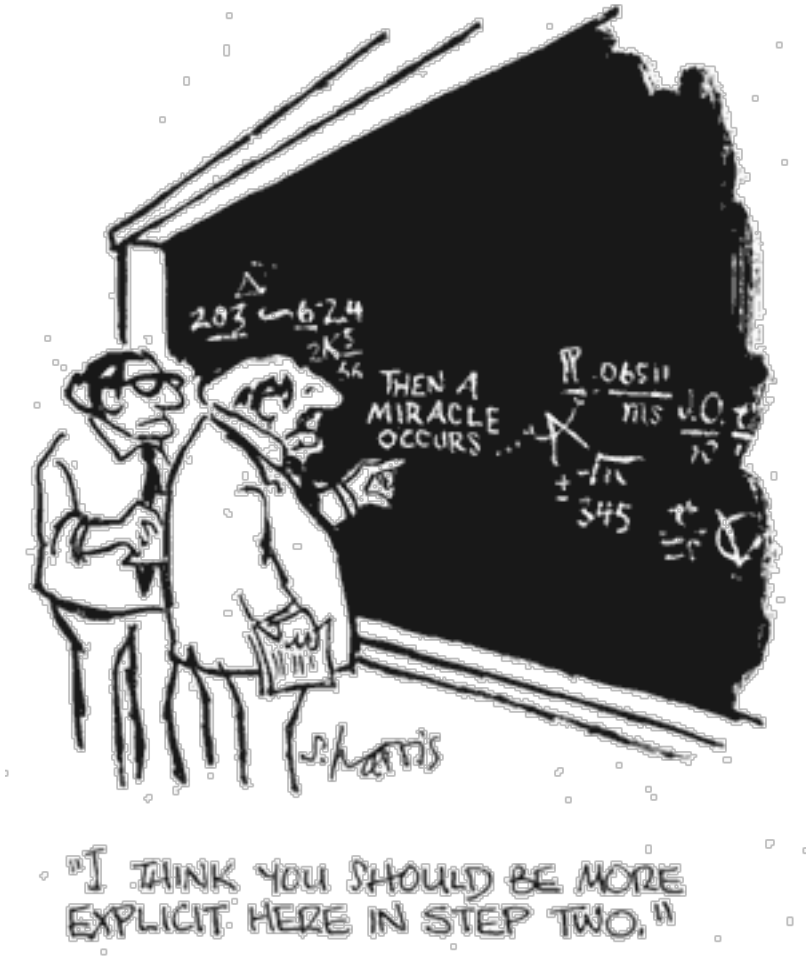
- CGC-GDC interoperability working group
- GA4GH Challenge on Container and workflow interoperability

What does it take to build and run a pipeline?

- Reliable Data
- Reliable, Secure Authorizations
- Reliable, Secure, Cost-effective Storage (possibly long term)
- Reliable, Secure, Cost-effective, Scalable Compute
- Reliable, Secure, Cost-effective, Scalable, Steady Support



Scientific Validation of a Pipeline



- Testing and fine-tuning
- Well-defined
  - Input Datasets
  - Output results
- Verifies provenance information
- Provides
  - Information/documentation,
  - limitations,
  - assumptions,
  - potential problems and solutions,
  - usability and support annotation

Biocompute Object

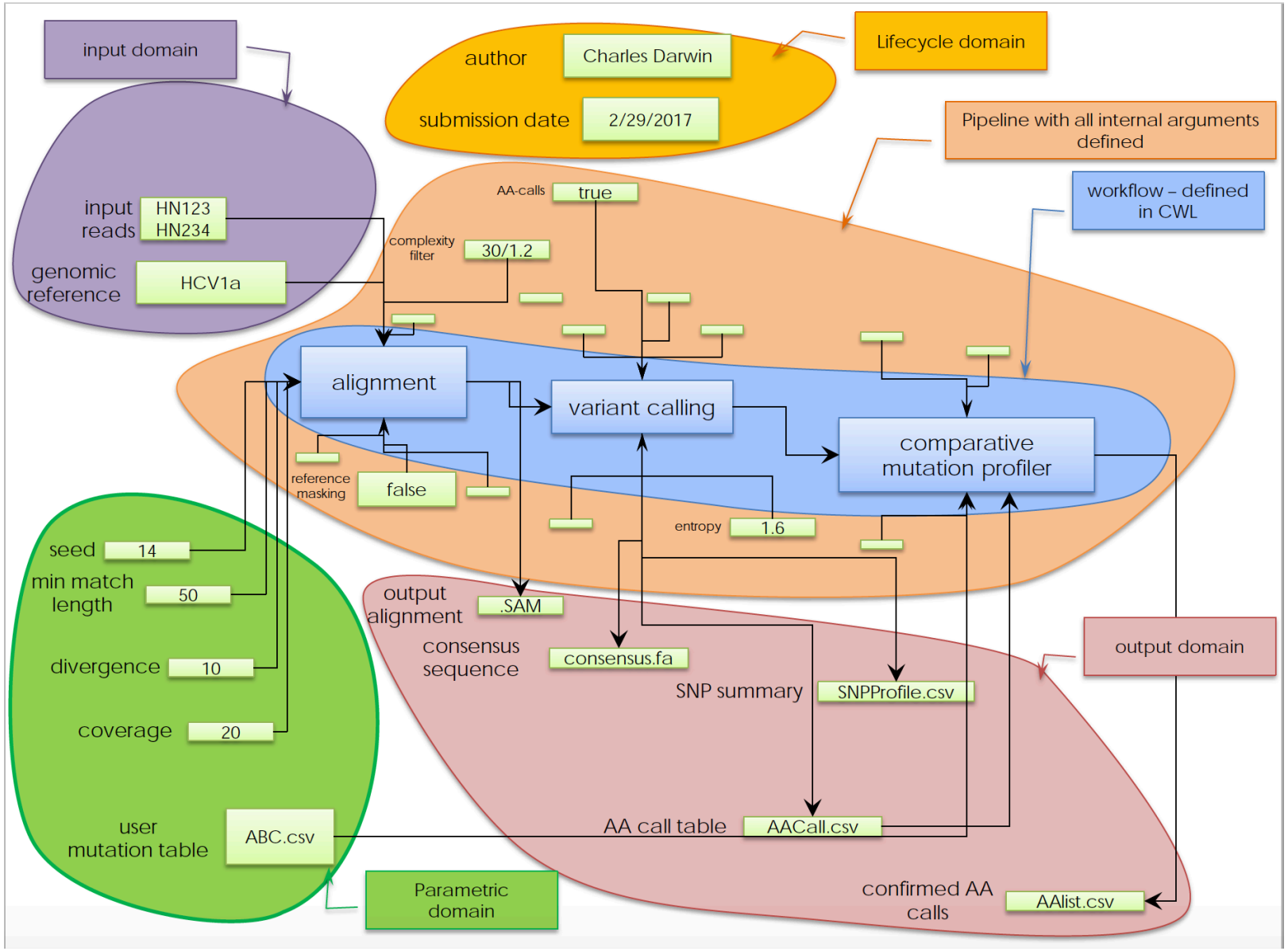


Image Curtesy Dr.Raja Mazumder, GWU

Biocompute Object – BCO Format

JSON Format

WXSwf1.json


```
{
  "sbg:revisionsInfo": [
    {
      "sbg:modifiedOn": 1481139873,
      "sbg:revision": 0,
      "sbg:modifiedBy": "Durga",
      "sbg:revisionNotes": "Copy of Durga/exome-sequencing/exomeseqanalysis02-withfreeba
    },
    {
      "sbg:modifiedOn": 1481139918,
      "sbg:revision": 1,
      "sbg:modifiedBy": "Durga",
      "sbg:revisionNotes": "Changed Name"
    },
    {
      "sbg:modifiedOn": 1481141068,
      "sbg:revision": 2,
      "sbg:modifiedBy": "Durga",
      "sbg:revisionNotes": "Output format - SortedBAM (default threads)\nDuplication - M
    },
    {
      "sbg:modifiedOn": 1481141157,
      "sbg:revision": 3,
      "sbg:modifiedBy": "Durga",
      "sbg:revisionNotes": null
    },
    {
      "sbg:modifiedOn": 1481157645,
      "sbg:revision": 4,
      "sbg:modifiedBy": "Durga",
      "sbg:revisionNotes": "added\n- append_comment"
    },
    {
      "sbg:modifiedOn": 1483475270,
```


BCO JSON Format


```
{
    "id": "obj.1481139873",
    "name": "Exome Sequence Analysis",
    "version": "5",
    "digital_signature": "",
    "verification_status": "unreviewed",
    "publication_status": "draft",
    "usability_domain": [],
    "authors": [{"name": "Durga"}],
    "description_domain": {
        "xref":[],
        "keywords": [],
        "pipeline_steps": {
            "FastQC":{
                "description": "A quality control tool for high throughput sequence data.",
                "version": "sbg:toolkitVersion:0.11.4",
                "step_number": 1,
                "package": ["http://www.bioinformatics.babraham.ac.uk/projects/fastqc/fastqc_v0.11.4.zip"],
                "input":["sbg://Durga/exome-sequencing/fastq_list.fastq.gz"],
                "output":[
          "#FastQC.report_zip",
          "#FastQC.report_html"
        ]
            },
            "SBG_Pair_FASTQs_by_Metadata":{
                "description": "Tool accepts list of FASTQ files for one sample as the input and groups them
                "version": null,
                "step_number": 1,
                "package":["https://cgc.sbgenomics.com/u/stefanristeski/group-fastqs/apps/#sbg-pair-fastqs-by
                "input":["#SBG_Pair_FASTQs_by_Metadata.fastq_list"],
                "output":["#SBG_Pair_FASTQs_by_Metadata.tuple_list"]
            },
            "BWA_INDEX":{
                "description": "BWA INDEX constructs the FM-index (Full-text index in Minute space) for the r
```

Acknowledgement

- NCI CBIIT
- Team CGC
- Seven Bridges Genomics - Cancer Genomics Cloud
  - http://www.cancergenomicscloud.org
- Broad Institute - FireCloud
  - http://firecloud.org
- Institute of Systems Biology - Cancer Genomics Cloud
  - http://cgc.systemsbiology.net
- Attain, LLC

Questions?