



Technical notes

Using Kooplex-edu for sequence analysis
a short tutorial

Overview



- ▶ Kooplex-edu provides a separate “virtual computer” for each student for each of his/her courses
- ▶ Data and tools can be uploaded to Kooplex-edu by teachers to keep necessary resources in the same place
- ▶ Assignments can be created, submitted and corrected on Kooplex-edu (with automatic collection at deadline)
- ▶ You can use Kooplex-edu for any of your courses if you need a place for programming or documenting pipelines/laboratory exercises

Find the website



<https://kooplex-edu.elte.hu>

Log in

Kooplex Reports ▾ Documentation Help ▾ **Log in** ↗

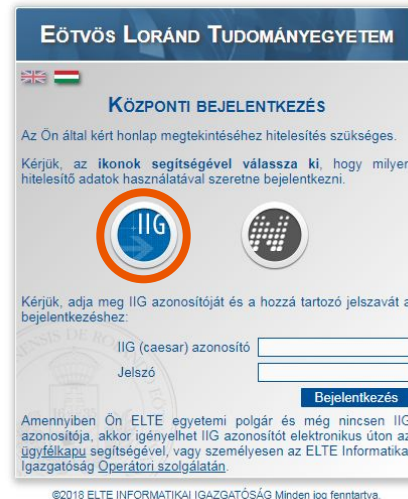
Kooplex
Educational platform at ELTE

Dive into various fields of computational physics, modeling and numerical analysis of data with the help of Jupyter notebooks.

Powered by

docker django jupyter R Studio NGINX slurm

Log in



Sign in with your caesar

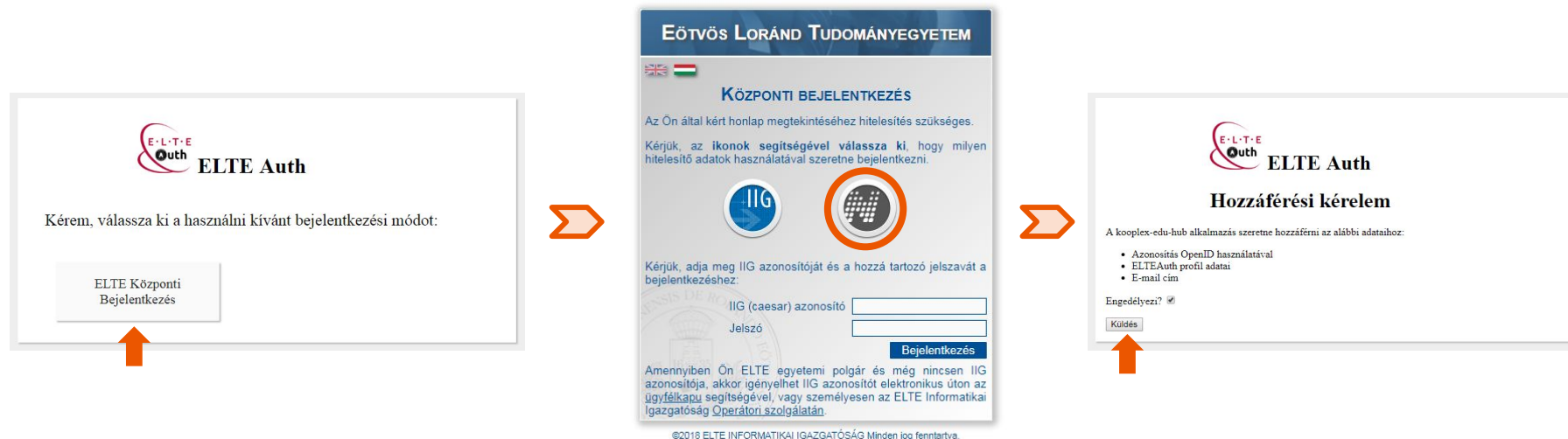
Log in



The screenshot shows the login page of Eötvös Loránd University. The header is "EÖTVÖS LORÁND TUDOMÁNYEGYETEM". Below it are the Hungarian and UK flags. The title is "KÖZPONTI BEJELENTKEZÉS". The text says: "Az Ön által kért honlap megtekintéséhez hitelesítés szükséges. Kérjük, az ikonok segítségével válassza ki, hogy milyen hitelesítő adatok használatával szeretne bejelentkezni." There are two icons: "IIG" and "Neptun". Below them, the text says: "Kérjük, adja meg IIG azonosítóját és a hozzá tartozó jelszavát a bejelentkezéshez:". There are two input fields: "IIG (caesar) azonosító" and "Jelszó". A blue button labeled "Bejelentkezés" is below the fields. At the bottom, there is a note: "Amennyiben Ön ELTE egyetemi polgár és még nincs IIG azonosítója, akkor igényelhet IIG azonosítót elektronikus úton az ügyfélfelkapu segítségével, vagy személyesen az ELTE Informatikai Igazgatóság Operátori szolgálatán". The footer says: "©2018 ELTE INFORMATIKAI IGAZGATÓSÁG Minden jog fenntartva."

Sign in with your caesar or neptun account

Log in



Sign in with your caesar or neptun account

Find the course

The screenshot shows the Kooplex Educational platform at ELTE. The header includes the Kooplex logo, navigation links (Projects, Reports, Courses, Containers), and a user greeting 'Hello Veronika!' with a dropdown arrow and a 'log off' button. The main content area features a large banner with the Kooplex logo and the text 'Educational platform at ELTE'. Below the banner, there are three informational boxes: an important note about course imports, a contact email for course setup, and a code update notice. At the bottom, there is a section titled 'Dive into various fields of computational physics, modeling and numerical analysis of data with the help of Jupyter notebooks.' which includes logos for various technologies: Docker, Django, Jupyter, RStudio, NGMX, and Slurm. The background of the website is a scenic view of the ELTE campus.

Kooplex Projects Reports Courses Containers Documentation Hello Veronika! log off

Kooplex Educational platform at ELTE

Important note: Courses are not imported automatically anymore. If you want to use notebooks for any of your courses please contact the administrators with the **course code** and the **title of the course** at kooplex@complex.elte.hu

If you don't have any courses but would like to **use Kooplex**, then please contact the administrators at kooplex@complex.elte.hu

There was a **code update** on 24. september 2019. If you notice any misbehaviour of the site, please report it to the email above! Thank you!

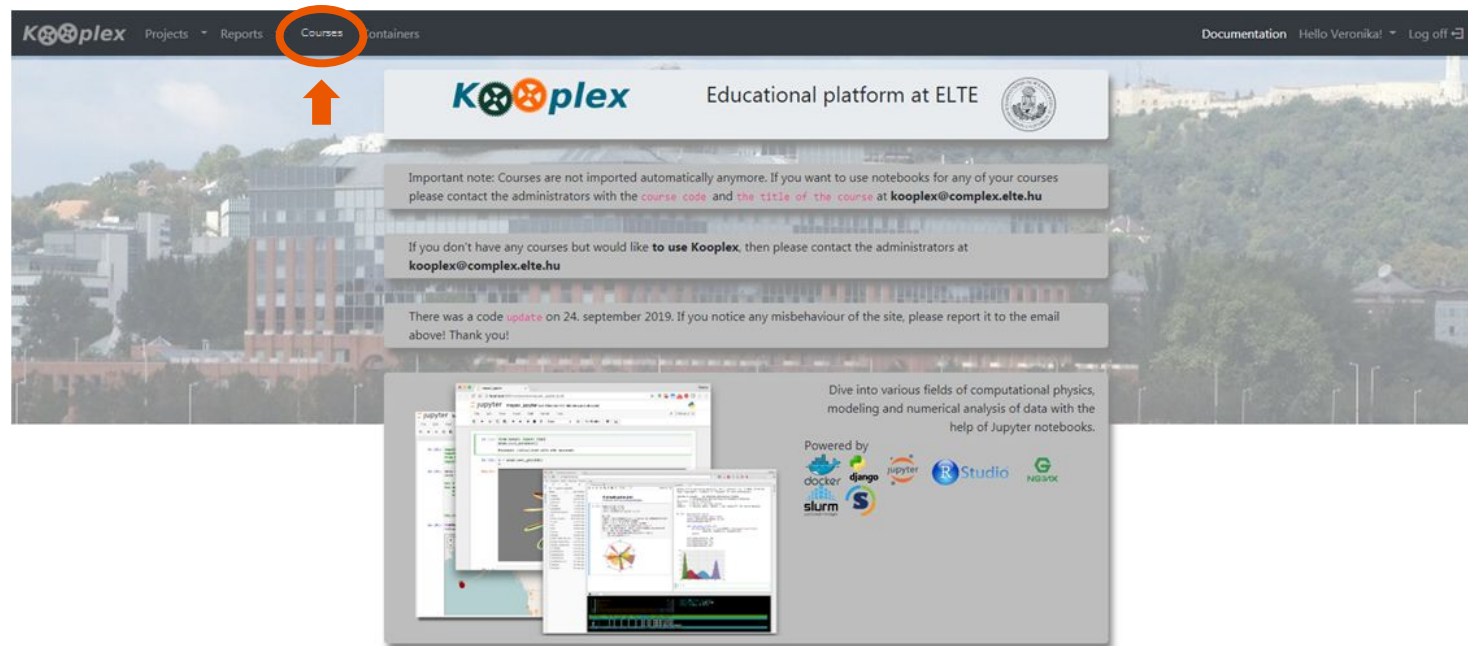
Dive into various fields of computational physics, modeling and numerical analysis of data with the help of Jupyter notebooks.

Powered by

docker, django, jupyter, RStudio, NGMX, slurm

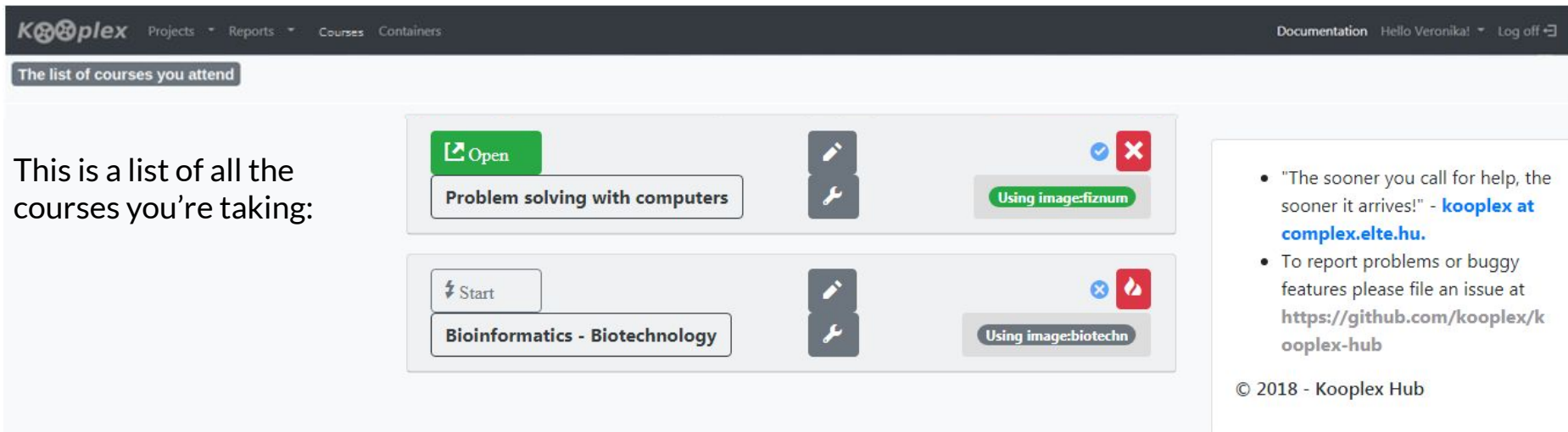
You're in!

Find the course



The screenshot shows the Kooplex website interface. The top navigation bar includes 'Kooplex', 'Projects', 'Reports', 'Courses' (circled in orange with an arrow pointing to it), and 'Containers'. The right side of the navigation bar has 'Documentation', 'Hello Veronika!', and 'Log off'. The main content area features the Kooplex logo and the text 'Educational platform at ELTE' next to the ELTE logo. Below this, there are three informational boxes: 1) 'Important note: Courses are not imported automatically anymore. If you want to use notebooks for any of your courses please contact the administrators with the **course code** and the **title of the course** at kooplex@complex.elte.hu' 2) 'If you don't have any courses but would like to **use Kooplex**, then please contact the administrators at kooplex@complex.elte.hu' 3) 'There was a **code update** on 24. september 2019. If you notice any misbehaviour of the site, please report it to the email above! Thank you!' At the bottom, there is a section titled 'Dive into various fields of computational physics, modeling and numerical analysis of data with the help of Jupyter notebooks.' which includes logos for 'Powered by' technologies: docker, slurm, jupyter, and RStudio, along with a 'G' logo.

Find the course



The screenshot shows the Kooplex web interface. The header includes the Kooplex logo and navigation links: Projects, Reports, Courses, and Containers. On the right, there are links for Documentation, a user greeting 'Hello Veronika!', and a Log off button. Below the header, a section titled 'The list of courses you attend' displays two course cards. The first card, 'Problem solving with computers', has an 'Open' button and a status indicator showing a checkmark and a red 'X'. The second card, 'Bioinformatics - Biotechnology', has a 'Start' button and a status indicator showing a red 'X' and a flame icon. Both cards include a wrench icon for settings and a button labeled 'Using image:'. To the right of the course list, a text box contains two bullet points: 'The sooner you call for help, the sooner it arrives!' - [kooplex at complex.elte.hu](https://kooplex.elte.hu), and 'To report problems or buggy features please file an issue at <https://github.com/kooplex/kooplex-hub>'. At the bottom right, the copyright notice '© 2018 - Kooplex Hub' is displayed.

This is a list of all the courses you're taking:

The list of courses you attend

Problem solving with computers

Open

Using image:fiznum

Bioinformatics - Biotechnology

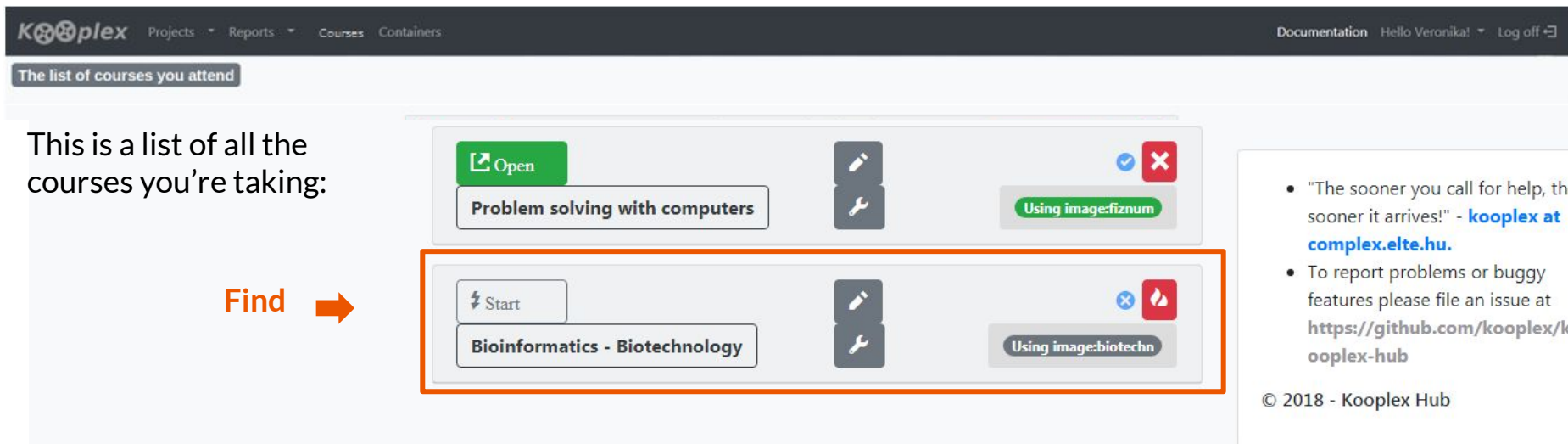
Start

Using image:biotechn

- "The sooner you call for help, the sooner it arrives!" - [kooplex at complex.elte.hu](https://kooplex.elte.hu).
- To report problems or buggy features please file an issue at <https://github.com/kooplex/kooplex-hub>

© 2018 - Kooplex Hub

Find the course



The screenshot shows the Kooplex web interface. The top navigation bar includes 'Kooplex', 'Projects', 'Reports', 'Courses', and 'Containers'. On the right, there are links for 'Documentation', 'Hello Veronika!', and 'Log off'. Below the navigation bar, a header reads 'The list of courses you attend'. The main content area displays two course cards. The first card, 'Problem solving with computers', has a green 'Open' button and a status 'Using image:fiznum'. The second card, 'Bioinformatics - Biotechnology', has a grey 'Start' button and a status 'Using image:biotechn'. This second card is highlighted with an orange border. To the left of the cards, the text 'This is a list of all the courses you're taking:' is followed by the word 'Find' in orange, with an orange arrow pointing towards the highlighted course card.

This is a list of all the courses you're taking:

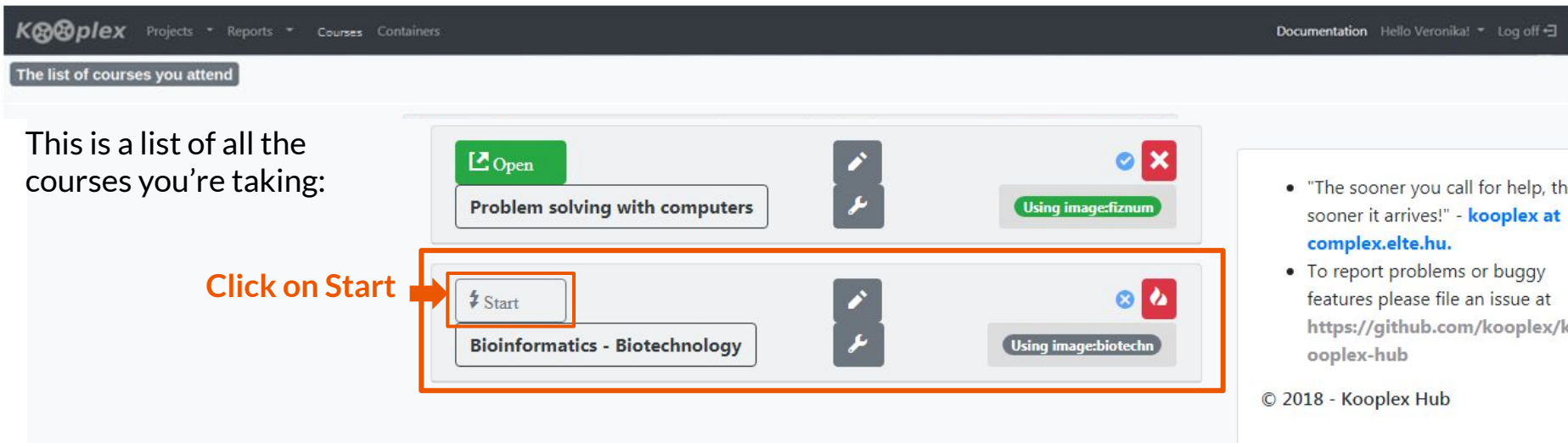
Find →

Bioinformatics - Biotechnology

- "The sooner you call for help, the sooner it arrives!" - [kooplex at complex.elte.hu](https://kooplex.elte.hu).
- To report problems or buggy features please file an issue at <https://github.com/kooplex/kooplex-hub>

© 2018 - Kooplex Hub

Find the course



The screenshot shows the Kooplex web interface. The top navigation bar includes the Kooplex logo and links for Projects, Reports, Courses, and Containers. On the right, there are links for Documentation, a user greeting (Hello Veronika!), and a Log off button. Below the navigation bar, a section titled "The list of courses you attend" displays two course cards. The first card is for "Problem solving with computers" and includes an "Open" button, a settings icon, and a status indicator showing a checkmark and a red X. The second card is for "Bioinformatics - Biotechnology" and includes a "Start" button, a settings icon, and a status indicator showing a blue X and a red flame icon. An orange box highlights the "Start" button of the second course, and an orange arrow points to it with the text "Click on Start".

This is a list of all the courses you're taking:

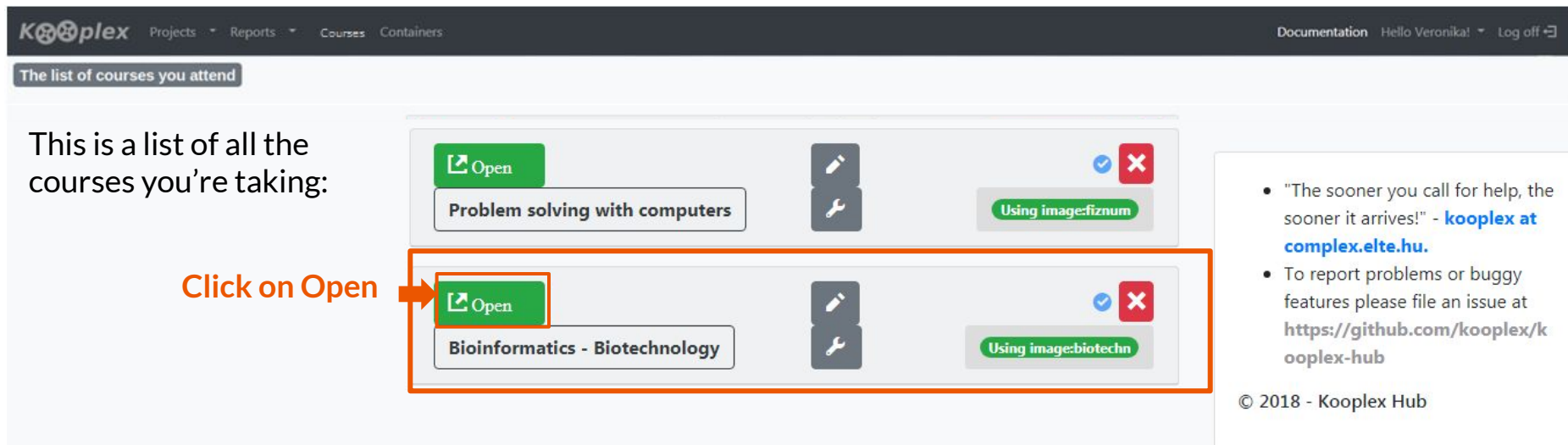
Click on Start →

- Problem solving with computers
- Bioinformatics - Biotechnology

- "The sooner you call for help, the sooner it arrives!" - [kooplex at complex.elte.hu](https://kooplex.elte.hu).
- To report problems or buggy features please file an issue at <https://github.com/kooplex/kooplex-hub>

© 2018 - Kooplex Hub

Find the course









Kooplex Projects Reports Courses Containers Documentation Hello Veronika! Log off

The list of courses you attend

This is a list of all the courses you're taking:

Click on Open →

| | | | | |
|--|--------------------------------|--|---|----------------------|
|  Open | Problem solving with computers |  |  | Using image:fiznum |
|  Open | Bioinformatics - Biotechnology |  |  | Using image:biotechn |

- "The sooner you call for help, the sooner it arrives!" - [kooplex at complex.elte.hu](https://kooplex.elte.hu).
- To report problems or buggy features please file an issue at <https://github.com/kooplex/kooplex-hub>

© 2018 - Kooplex Hub

Assignment directory

Files

Running

Clusters

Switch to JupyterLab

Conda

Nbextensions

Select items to perform actions on them.

0

/

Search

Upload

N

| | Name | Last Modified |
|--------------------------|----------|----------------|
| <input type="checkbox"/> | course | 34 minutes ago |
| <input type="checkbox"/> | yorkdir | 31 minutes ago |
| <input type="checkbox"/> | x9C | 6 months ago |
| <input type="checkbox"/> | feedback | 6 months ago |



Assignment directory

Files Running Clusters Switch to JupyterLab Conda Nbextensions

Select items to perform actions on them.

Search Upload

| | Name | Last Modified |
|----------|------|----------------|
| 0 / | | |
| course | | 34 minutes ago |
| workdir | | 31 minutes ago |
| x9C | | 6 months ago |
| feedback | | 6 months ago |

Files Running Clusters

Select items to perform actions on them.

Upload New

| | Name | Last Modified | File size |
|----------------------------|------|----------------|-----------|
| 0 / workdir | | | |
| .. | | seconds ago | |
| genomeSequencingAssignment | | 28 minutes ago | |

Assignment notebook

Files

Running

Clusters

Select items to perform actions on them.

Upload

New

0

/ workdir / genomeSequencingAssignment

Name

Last Modified

File size

| | | |
|--|---------------|---------|
| <div><div></div><div>..</div></div> | seconds ago | |
| <div><div></div><div>ANNOVARinput</div></div> | a day ago | |
| <div><div></div><div>refgenome</div></div> | a day ago | |
| <div><div></div><div>RGBAM</div></div> | a day ago | |
| <div><div></div><div>sortedBAM</div></div> | a day ago | |
| <div><div></div><div>testResults</div></div> | a day ago | |
| <div><div></div><div>NGSAnalysisPipeline.ipynb</div></div> | 2 minutes ago | 2.22 MB |



Shared files and tools



The screenshot shows the JupyterLab file browser interface. At the top, there are tabs for 'Files', 'Running', 'Clusters', 'Switch to JupyterLab', 'Conda', and 'Nbextensions'. Below the tabs, there is a prompt 'Select items to perform actions on them.' and buttons for 'Search', 'Upload', and 'New'. The main area displays a directory listing with columns for 'Name' and 'Last Modified'. The 'course' directory is circled in orange, and an orange arrow points to it with the text 'Not writable!'.

| Name | Last Modified |
|----------|----------------|
| course | 34 minutes ago |
| workdir | 31 minutes ago |
| x9l24x | 6 months ago |
| feedback | 6 months ago |

Shared files and tools

Files

Running

Clusters

Select items to perform actions on them.

Upload

New ▾

☐ 0 ▾

/ share / genomeSequencing

Name ▾

Last Modified

File size

| | | |
|--------------------------|----------------|----------------|
| <input type="checkbox"/> | .. | seconds ago |
| <input type="checkbox"/> | ANNOVARresults | an hour ago |
| <input type="checkbox"/> | BAM | an hour ago |
| <input type="checkbox"/> | FASTQ | an hour ago |
| <input type="checkbox"/> | HC_GVCFs | an hour ago |
| <input type="checkbox"/> | PoN_VCFs | an hour ago |
| <input type="checkbox"/> | PoNs | an hour ago |
| <input type="checkbox"/> | PUP | an hour ago |
| <input type="checkbox"/> | tools | an hour ago |
| <input type="checkbox"/> | VarCallResults | 17 minutes ago |

About the notebook



- ▶ The notebook uses a `python3` kernel and `bash magics` (!) to run external tools in linux bash
- ▶ In order to make the most out of jupyter notebooks, invest some time into reading `tutorials` ([beginners guide](#), [tips&tricks](#), [detailed course in Hungarian](#))

About the notebook



- ▶ The notebook uses a **python3** kernel and **bash magics** (!) to run external tools in linux bash
- ▶ In order to make the most out of jupyter notebooks, invest some time into reading **tutorials** ([beginners guide](#), [tips&tricks](#), [detailed course in Hungarian](#))
- ▶ Notebook cells are either **Markdown**

pretty text with formatting, LaTeX formulas, tables, etc.

About the notebook



- ▶ The notebook uses a `python3` kernel and `bash magics` (!) to run external tools in linux bash
- ▶ In order to make the most out of jupyter notebooks, invest some time into reading `tutorials` ([beginners guide](#), [tips&tricks](#), [detailed course in Hungarian](#))
- ▶ Notebook cells are either Markdown, `Code`

pretty text with formatting, LaTeX formulas, tables, etc.

actual code that can be run and does things

About the notebook



- ▶ The notebook uses a **python3** kernel and **bash magics** (!) to run external tools in linux bash
- ▶ In order to make the most out of jupyter notebooks, invest some time into reading **tutorials** ([beginners guide](#), [tips&tricks](#), [detailed course in Hungarian](#))
- ▶ Notebook cells are either Markdown, Code or **Raw NBConvert**

pretty text with formatting, LaTeX formulas, tables, etc.

actual code that can be run and does things

looks like code (not formatted), but cannot be run (does nothing)

About the notebook



- ▶ The notebook uses a `python3` kernel and `bash magics` (!) to run external tools in linux bash
- ▶ In order to make the most out of jupyter notebooks, invest some time into reading `tutorials` ([beginners guide](#), [tips&tricks](#), [detailed course in Hungarian](#))
- ▶ Notebook cells are either Markdown, Code or Raw NBConvert
- ▶ All cell types can be run with: `Shift+Enter`

Things you don't need to know at this point



- ▶ Kooplex-edu allows you to [submit](#) your solutions to assignments and also [collects](#) them automatically at the predefined deadline
- ▶ You can [install different python packages](#) on Kooplex-edu that you might need for your courses
- ▶ There is a python package available for every task, [learn to love python](#)
- ▶ The Jupyter Notebook is an extremely powerful tool to create whole analysis pipelines that are documented in detail and are thus reproducible
- ▶ [Reproducibility](#) should be one of the key objectives of science

Technical inquiries



kooplex@complex.elte.hu

Technical problems:

- I can't sign in.
- The website is unavailable.
- I can't open the notebook.
- I can't find the course.
- etc.

Non-technical problems:

- I don't understand what the code does.
- My code doesn't work.
- I require more hints for the task.
- etc.



Genome sequencing

Bioinformatical analysis of
Next Generation Sequencing results

with a focus on human genome sequencing and cancer

Overview



- ▶ NGS technology, a (very) brief introduction
 - PCR, short reads, base calling from image files
- ▶ Data analysis pipeline
 - ▶ Alignment
 - ▶ Preprocessing alignment files
 - ▶ Variant calling (SNVs, indels)
 - ▶ Interpreting variant files



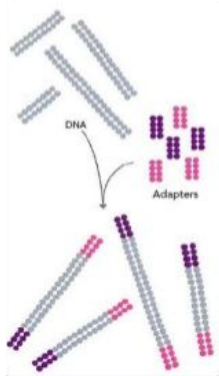
NGS technology



Preparing DNA for sequencing

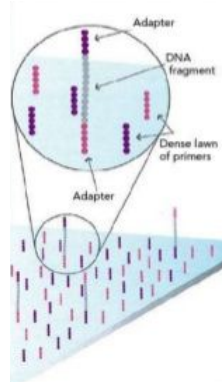
PCR + Short reads

Randomly fragment DNA after PCR and ligate adapters to both ends.



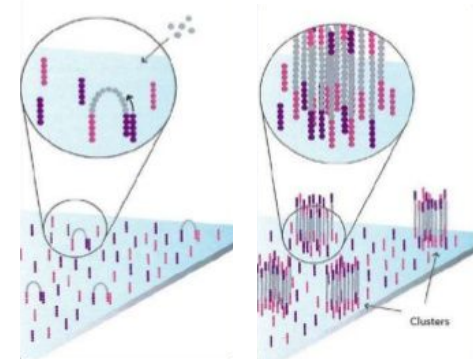
Attach to surface

Bind **single-stranded** fragments randomly to the flow cell.



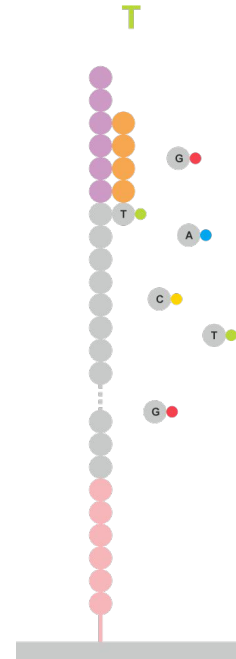
Amplification

Dense clusters of **identical** DNA are generated on the flow cell.

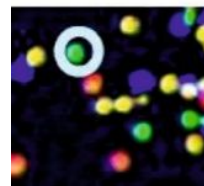


Sequencing by synthesis

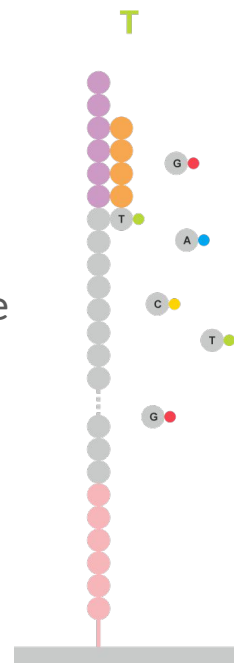
1. Flow cell is washed with a mixture of **DNA polymerase enzyme** and **fluorescently tagged ddNTPs**. (four-colour chemistry: different emission for each base)



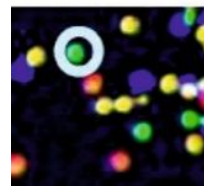
Sequencing by synthesis



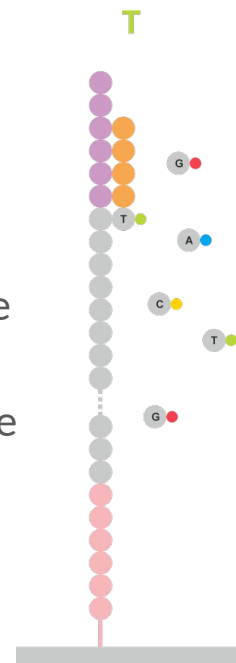
1. Flow cell is washed with a mixture of **DNA polymerase enzyme** and **fluorescently tagged ddNTPs**. (four-colour chemistry: different emission for each base)
2. **Bright colored patches** (one for each cluster) are recorded as a **photo**.



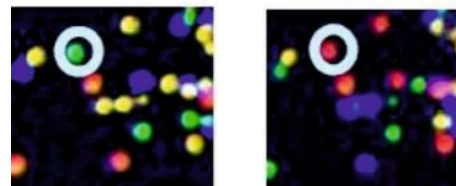
Sequencing by synthesis



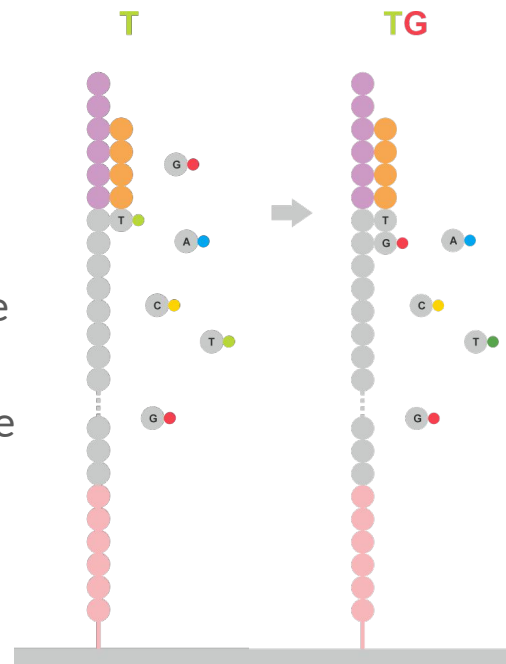
1. Flow cell is washed with a mixture of **DNA polymerase enzyme** and **fluorescently tagged ddNTPs**. (four-colour chemistry: different emission for each base)
2. **Bright colored patches** (one for each cluster) are recorded as a **photo**.
3. Fluorescent tags and sequencing terminators are washed away from the last base.



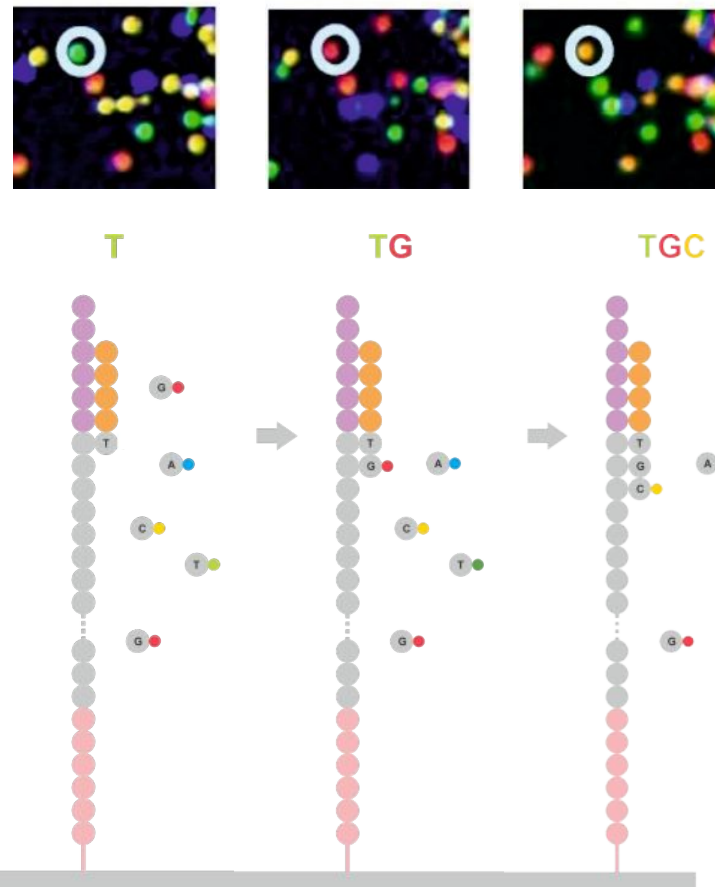
Sequencing by synthesis



1. Flow cell is washed with a mixture of **DNA polymerase enzyme** and **fluorescently tagged ddNTPs**. (four-colour chemistry: different emission for each base)
2. **Bright colored patches** (one for each cluster) are recorded as a **photo**.
3. Fluorescent tags and sequencing terminators are washed away from the last base.
4. The process is repeated again...

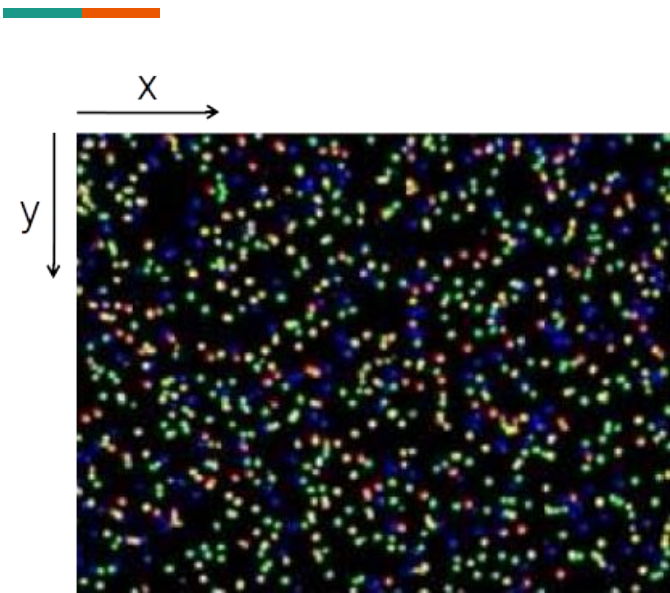


Sequencing by synthesis



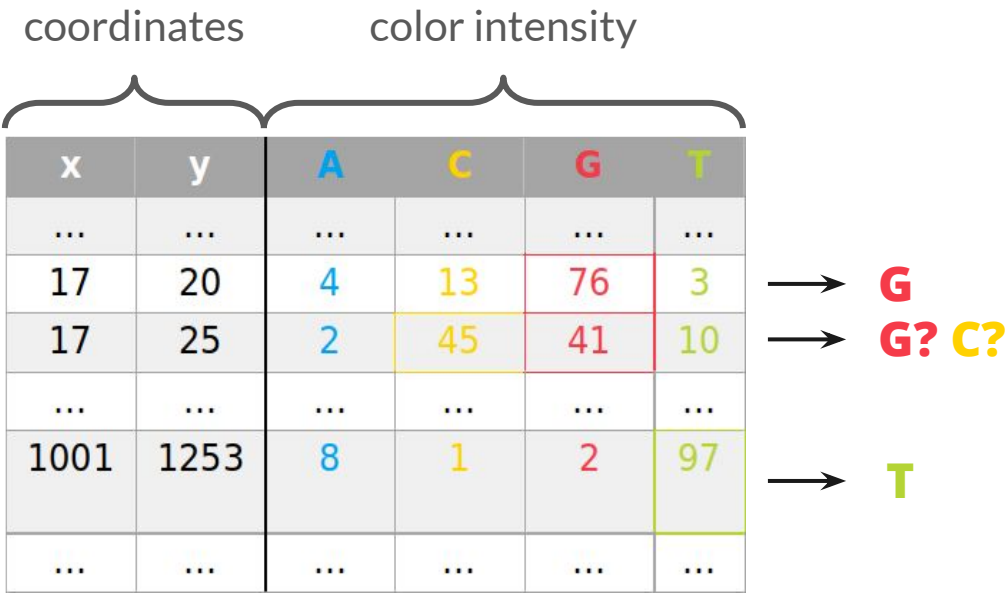
1. Flow cell is washed with a mixture of **DNA polymerase enzyme** and **fluorescently tagged ddNTPs**. (four-colour chemistry: different emission for each base)
2. **Bright colored patches** (one for each cluster) are recorded as a **photo**.
3. Fluorescent tags and sequencing terminators are washed away from the last base.
4. The process is repeated again...
5. And again...

From image to text



3D: (x,y) and time

Short read: fixed (x,y), changing time →



A A A C G T A C A C A bases
Q_A Q_A Q_A Q_C Q_G Q_T Q_A Q_C Q_A Q_C Q_A base qualities

Base calling error, base quality



The probability of calling a given base incorrectly: P

(~ high, when we have trouble deciding between the colors)

Base quality (Phred-score): $Q = -10 \log_{10} P$

(The higher Q is, the more reliable the base call.)

Output of NGS: **FASTQ format**

Convert to ASCII:

1. Round to integer value
2. Add 33

$$Q_{ASCII} = \text{round}(Q) + 33$$

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((( (**+)) %%%++) (%%%) .1***-+*'') **55CCF>>>>>CCCCCCC65
```



Data analysis

A general idea



De-novo assembly



1. **Reconstruct the whole genome** from the short reads

De-novo assembly



1. **Reconstruct the whole genome** from the short reads

_diff

is_fa

fairl

rly_d

cult.

_is_f

his_i

this_

fficu

irly_

iffic

De-novo assembly



1. **Reconstruct the whole genome** from the short reads

```
this_is_fa rly_diffic  
his_i fairl diff cult.  
_is_f irly_ fficu
```

(de-novo assembly)

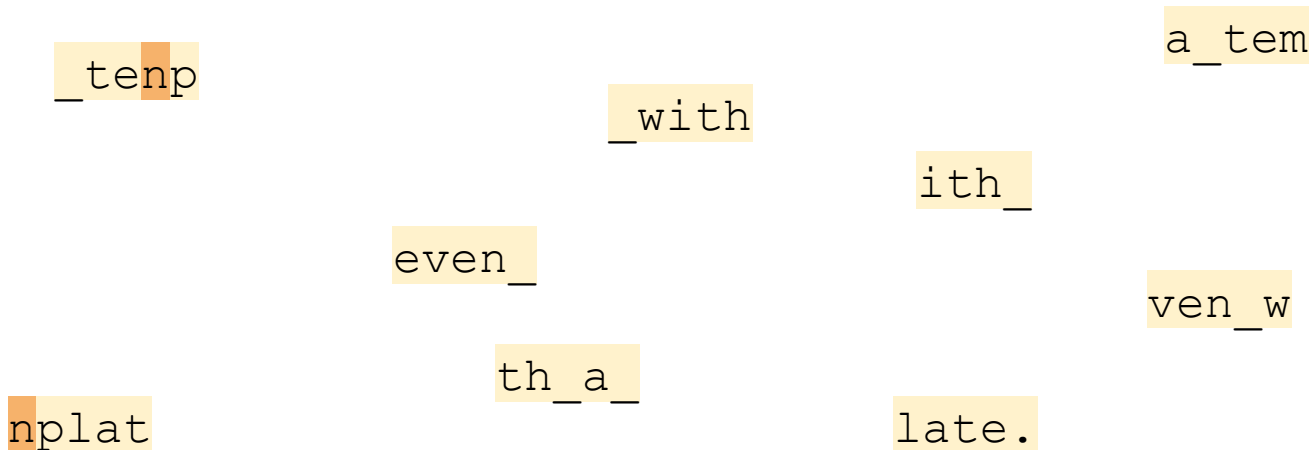
```
this_is_fairly_difficult.
```


Alignment to a reference genome



1. **Reconstruct the whole genome** from the short reads

even_with_a_template.



_tenp

_with

ith_

a_tem

even_

ven_w

th_a_

nplat

late.

Alignment to a reference genome



1. **Reconstruct the whole genome** from the short reads

even_with_a_template.

even_ ith_a tem late.

ven_w th_a nplat

_with _tenp

(alignment to reference genome)

Variant calling

1. **Reconstruct the whole genome** from the short reads

even_with_a_template.

even_ ith_a tem late.

ven_w th_a nplat

_with _tenp



m > n

(alignment to reference genome)

2. Compare the reconstructed genome(s) to a reference genome or to each other and **find differences** (variants/mutations)

Variant calling

1. **Reconstruct the whole genome** from the short reads

even_with_a_template.

even_ ith_a tem late.

ven w th_a nplat

_with _tenp

(alignment to reference genome)

m > n

Not at all trivial either!



2. Compare the reconstructed genome(s) to a reference genome or to each other and **find differences** (variants/mutations)



Data analysis

Details



Analysis pipeline: outputs and file formats



NGS

Sequences and base qualities of short reads in **random order**.

FASTQ files

Alignment

Sequences and base qualities of short reads with the **genomic position of where they fit on the reference genome**.

SAM/BAM files
Pileup files

Preprocessing

Sorted (easy to search) alignment files with sequences labelled with **read groups** and sometimes with **duplicates removed**.

BAM files

Variant calling

List of somatic and germline variants with additional information.

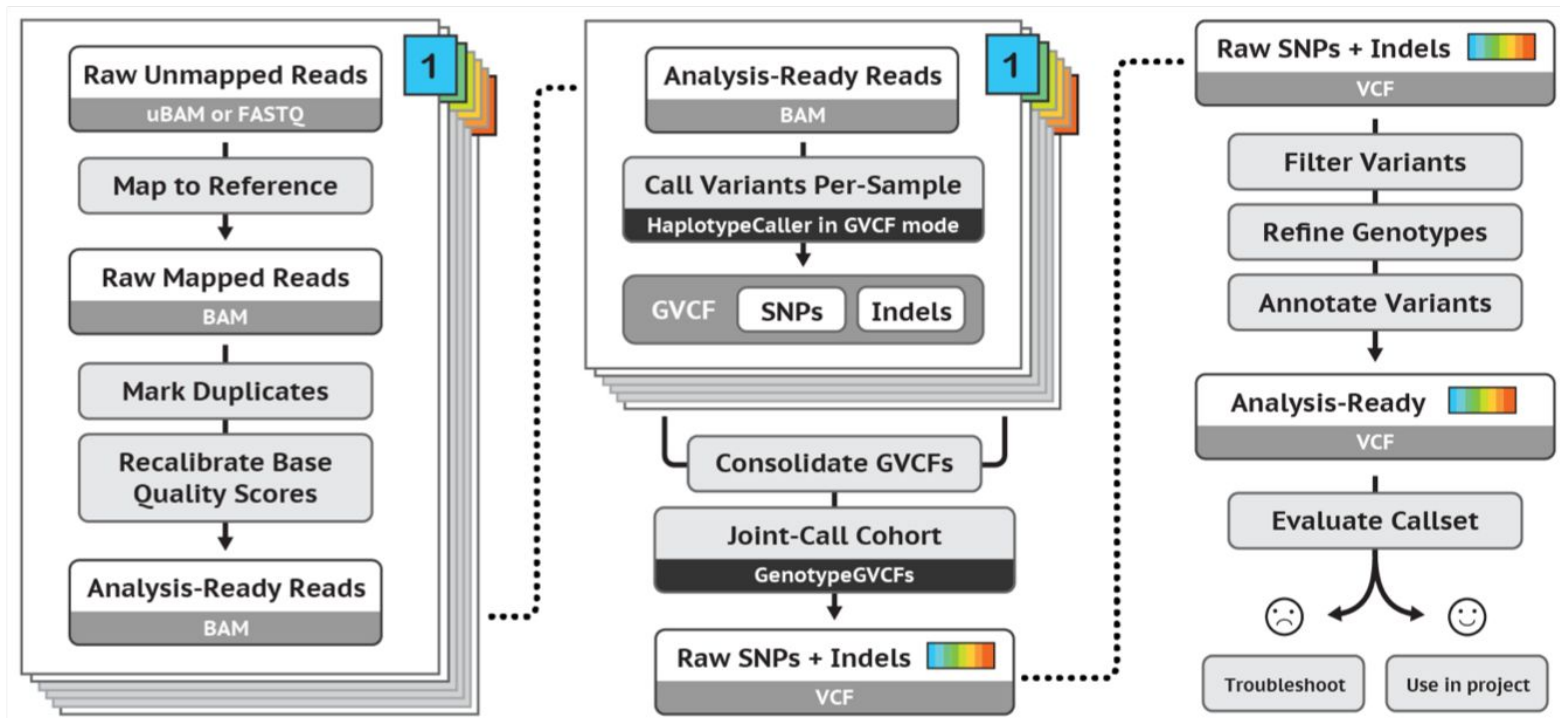
VCF files

Interpretation

List of **annotated, filtered variants, figures**, etc.

Image files
Filtered VCF files

GATK Best Practices - Main steps for Germline Cohort Data



Preparations for alignment



Not always easy!

1. Find and download the **appropriate** reference genome (**FASTA format**)
(i.e. do not align sequencing data from a chicken to the human reference genome)

e.g: <http://hgdownload.cse.ucsc.edu/goldenPath/hg38/chromosomes/>

2. Create an index file for the reference genome

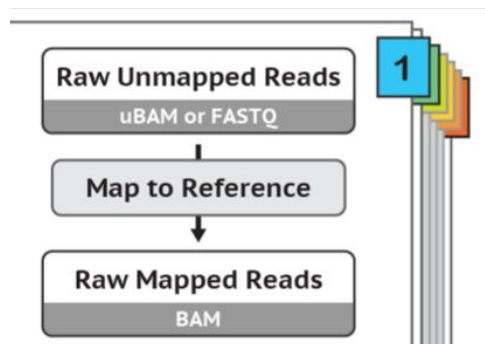
Tools: BWA, samtools

```
samtools faidx refgenome.fa  
bwa index refgenome.fa
```


Alignment to a reference genome

Input: short read sequences (and base qualities) in random order (FASTQ files or uBAM files (convert first to FASTQ files))

Goal: determining the order of the short reads by fitting them to a template (reference genome)



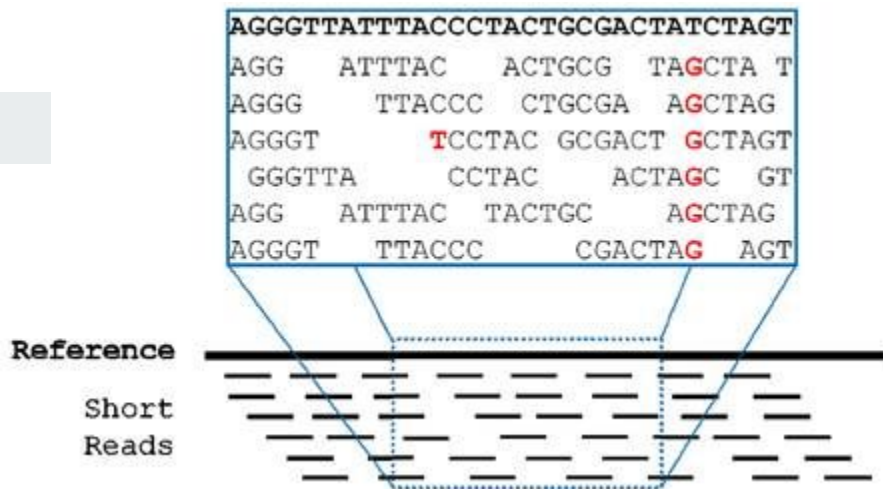
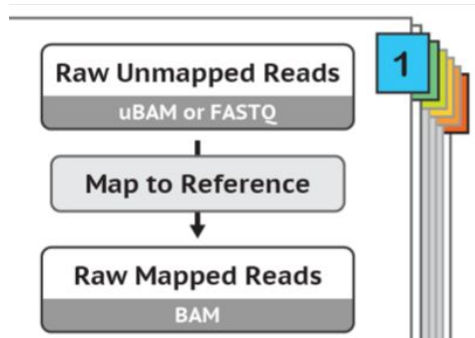
Alignment to a reference genome

Input: short read sequences (and base qualities) in random order (**FASTQ files or uBAM files** (convert first to FASTQ files))

Goal: determining the order of the short reads by fitting them to a template (**reference genome**)

Tool: BWA

```
bwa mem refgenome.fa s1.fq.gz > s1.sam
```



Input: short read sequences (and base qualities) in random order (FASTQ files or uBAM files (convert first to FASTQ files))

Goal: determining the order of the short reads by fitting them to a template (reference genome)

```
samtools view -bS s1.sam > s1.bam
```

Output: short reads with **position**
and **mapping quality** (SAM/BAM files)

Conversion tool: samtools

```

1:497:R:-272+13M17D24M      113      1      497      37      37M      15      100338662      0
CGGGTCTGACCTGAGGAGAACTGTGCTCCGCCTTCAG      0;==-=9;>>>>=>>>>>>>>>=>>>>>>>>      XT:A:U NM:i:0 SM:i:37      AM:i:0
X0:i:1 X1:i:0 XM:i:0 XO:i:0 XG:i:0 MD:Z:37
19:20389:F:275+18M2D19M      99      1      17644      0      37M      =      17919      314      TATGACTGCTAATAATACCTACACATGTTAGAACCAT
>>>>>>>>>>>>>>>><<>><<>>4:.:>:<9      RG:Z:UM0098:1 XT:A:R NM:i:0 SM:i:0 AM:i:0 X0:i:4 X1:i:0 XM:i:0 XO:i:0
XG:i:0 MD:Z:37
19:20389:F:275+18M2D19M      147      1      17919      0      18M2D19M      =      17644      -314
GTAGTACCAACTGTAAGTCCTTATCTTCATACTTTGT      ;44999;499<8<8<<<8<<<<<<<<<7<;<<<>><<      XT:A:R NM:i:2 SM:i:0 AM:i:0 X0:i:4
X1:i:0 XM:i:0 XO:i:1 XG:i:2 MD:Z:18^CA19

```

- | | | | | | | | | |
|--|--------------------------------------|---|-------|----|----------|----|------------|---|
| 1:497;R:-272+13M17D24M | 113 | 1 | 497 | 37 | 37M | 15 | 100338662 | 0 |
| CGGGTCTGACCTGAGGAGAAGTGTCTCCGCCTTCAG | 0;==--=9;>>>>=>>>>>>>>>>>>>>> | | | | | | | XT:A:U NM:i:0 SM:i:37 AM:i:0 |
| X0:i:1 X1:i:0 XM:i:0 XO:i:0 XG:i:0 MD:Z:37 | | | | | | | | |
| 19:20389:F:275+18M2D19M | 99 | 1 | 17644 | 0 | 37M | = | 17919 314 | TATGACTGCTAATAATACCTACACATGTTAGAACCAT |
| >>>>>>>>>>>>>>>><<<>><>>4:;>><9 | | | | | | | | RG:Z:UM0098:1 XT:A:R NM:i:0 SM:i:0 AM:i:0 X0:i:4 X1:i:0 XM:i:0 XO:i:0 |
| XG:i:0 MD:Z:37 | | | | | | | | |
| 19:20389:F:275+18M2D19M | 147 | 1 | 17919 | 0 | 18M2D19M | = | 17644 -314 | |
| GTAGTACCAACTGTAAGTCCTTATCTTCATACTTTGT | ;44999;499<8<<<8<<<<<<<<<<7<;<<<>><< | | | | | | | XT:A:R NM:i:2 SM:i:0 AM:i:0 X0:i:4 |
| X1:i:0 XM:i:0 XO:i:1 XG:i:2 MD:Z:18^CA19 | | | | | | | | |

- ```

1:497:R:-272+13M17D24M 113 1 497 37 37M 15 100338662 0
CGGGTCTGACCTGAGGAGAACTGTGCTCCGCCTTCAG 0;==--=9;>>>>=>>>>>>>>=>>>>>>>> XT:A:U NM:i:0 SM:i:37 AM:i:0
X0:i:1 X1:i:0 XM:i:0 XO:i:0 XG:i:0 MD:Z:37
19:20389:F:275+18M2D19M 99 1 17644 0 37M = 17919 314 TATGACTGCTAATAATACCTACACATGTTAGAACCAT
>>>>>>>>>>>>>>>>>><<<<><<>4:;>><9 RG:Z:UM0098:1 XT:A:R NM:i:0 SM:i:0 AM:i:0 X0:i:4 X1:i:0 XM:i:0 XO:i:0
XG:i:0 MD:Z:37
19:20389:F:275+18M2D19M 147 1 17919 0 18M2D19M = 17644 -314
GTAGTACCAACTGTAAGTCCTTATCTTCATACTTTGT ;44999;499<8<8<<<8<<<<<<<<<7<;<<<>><< XT:A:R NM:i:2 SM:i:0 AM:i:0 X0:i:4
X1:i:0 XM:i:0 XO:i:1 XG:i:2 MD:Z:18^CA19

```



- |                                            |     |   |                                       |    |          |    |                                    |                                       |
|--------------------------------------------|-----|---|---------------------------------------|----|----------|----|------------------------------------|---------------------------------------|
| 1:497:R:-272+13M17D24M                     | 113 | 1 | 497                                   | 37 | 37M      | 15 | 100338662                          | 0                                     |
| CGGGTCTGACCTGAGGAGAAGTGTCTCCGCCTTCAG       |     |   | 0;==--=9;>>>>=>>>>>>>>=>>>>>>>        |    |          |    | XT:A:U NM:i:0 SM:i:37              | AM:i:0                                |
| X0:i:1 X1:i:0 XM:i:0 XO:i:0 XG:i:0 MD:Z:37 |     |   |                                       |    |          |    |                                    |                                       |
| 19:20389:F:275+18M2D19M                    | 99  | 1 | 17644                                 | 0  | 37M      | =  | 17919 314                          | TATGACTGCTAATAATACCTACACATGTTAGAACCAT |
| >>>>>>>>>>>>>>><<<>><>>4:;>><9             |     |   |                                       |    |          |    | RG:Z:UM0098:1 XT:A:R NM:i:0 SM:i:0 | AM:i:0 X0:i:4 X1:i:0 XM:i:0 XO:i:0    |
| XG:i:0 MD:Z:37                             |     |   |                                       |    |          |    |                                    |                                       |
| 19:20389:F:275+18M2D19M                    | 147 | 1 | 17919                                 | 0  | 18M2D19M | =  | 17644 -314                         |                                       |
| GTAGTACCAACTGTAAGTCCTTATCTTCATACTTTGT      |     |   | ;44999;499<8<8<<<8<<<<<<<<<7<;<<<>><< |    |          |    | XT:A:R NM:i:2 SM:i:0 AM:i:0 X0:i:4 |                                       |
| X1:i:0 XM:i:0 XO:i:1 XG:i:2 MD:Z:18^CA19   |     |   |                                       |    |          |    |                                    |                                       |

- QNAME: name of the short read present in FASTQ file
- FLAG: a number (“code”) describing the alignment
- RNAME: name of the reference sequence  
(often contains chromosome)
- POS: mapping position
- MAPQ: mapping quality

|                                          |                                     |     |       |     |          |     |       |           |                                                                       |                                    |
|------------------------------------------|-------------------------------------|-----|-------|-----|----------|-----|-------|-----------|-----------------------------------------------------------------------|------------------------------------|
| 1:497                                    | R:-272+13M17D24M                    | 113 | 1     | 497 | 37       | 37M | 15    | 100338662 | 0                                                                     |                                    |
| CGGGTCTGACCTGAGGAGAAGTGCTCCGCCTTCAG      |                                     |     |       |     |          |     |       |           | 0;===9;>>>>=>>>>>>>>>>>>>>>>>>>>                                      | XT:A:U NM:i:0 SM:i:37 AM:i:0       |
| X0:i:1                                   | X1:i:0 XM:i:0 XO:i:0 XG:i:0 MD:Z:37 |     |       |     |          |     |       |           |                                                                       |                                    |
| 19:20389:F:275+18M2D19M                  | 99                                  | 1   | 17644 | 0   | 37M      | =   | 17919 | 314       | TATGACTGCTAATAATACCTACACATGTTAGAACCAT                                 |                                    |
| >>>>>>>>>>>>>><<<<><<>>4::>><9           |                                     |     |       |     |          |     |       |           | RG:Z:UM0098:1 XT:A:R NM:i:0 SM:i:0 AM:i:0 X0:i:4 X1:i:0 XM:i:0 XO:i:0 |                                    |
| XG:i:0 MD:Z:37                           |                                     |     |       |     |          |     |       |           |                                                                       |                                    |
| 19:20389:F:275+18M2D19M                  | 147                                 | 1   | 17919 | 0   | 18M2D19M | =   | 17644 | -314      |                                                                       |                                    |
| GTAGTACCACCTGTAAGTCCTTATCCTTCATACTTTGT   |                                     |     |       |     |          |     |       |           | ;44999;499<8<8<<<8<<<<<<<<<<7<;<<<>><<                                | XT:A:R NM:i:2 SM:i:0 AM:i:0 X0:i:4 |
| X1:i:0 XM:i:0 XO:i:1 XG:i:2 MD:Z:18^CA19 |                                     |     |       |     |          |     |       |           |                                                                       |                                    |



- QNAME: name of the short read present in FASTQ file
- FLAG: a number (“code”) describing the alignment
- RNAME: name of the reference sequence  
(often contains chromosome)
- POS: mapping position
- MAPQ: mapping quality
- CIGAR string: indicating alignment information

1:497:R:-272+13M17D24M 113 1 497 37 37M 15 100338662 0  
CGGGTCTGACCTGAGGAGAACTGTGCTCCGCCTTCAG 0;==-=9;>>>>=>>>>>>>>=>>>>>>>> XT:A:U NM:i:0 SM:i:37 AM:i:0  
X0:i:1 X1:i:0 XM:i:0 XO:i:0 XG:i:0 MD:Z:37  
19:20389:F:275+18M2D19M 99 1 17644 0 37M = 17919 314 TATGACTGCTAATAATACCTACACATGTTAGAACCAT  
>>>>>>>>>>>>>>>><<<<><<>>4:.>><9 RG:Z:UM0098:1 XT:A:R NM:i:0 SM:i:0 AM:i:0 X0:i:4 X1:i:0 XM:i:0 XO:i:0  
XG:i:0 MD:Z:37  
19:20389:F:275+18M2D19M 147 1 17919 0 18M2D19M = 17644 -314  
GTAGTACCAACTGTAAGTCCTTATCTTCATACTTTGT ;44999;499<8<8<<<8<<<<<<<<<<7<;<<<>><< XT:A:R NM:i:2 SM:i:0 AM:i:0 X0:i:4  
X1:i:0 XM:i:0 XO:i:1 XG:i:2 MD:Z:18^CA19

- QNAME: name of the short read present in FASTQ file
- FLAG: a number (“code”) describing the alignment
- RNAME: name of the reference sequence  
(often contains chromosome)
- POS: mapping position
- MAPQ: mapping quality
- CIGAR string: indicating alignment information
- **RNEXT**: reference sequence name for the next read

```

1:497:R:-272+13M17D24M 113 1 497 37 37M 15 100338662 0
CGGGTCTGACCTGAGGAGAAGTGTGCTCCGCCTTCAG 0;===9;>>>>=>>>>>>>>>>=>>>>>>>> XT:A:U NM:i:0 SM:i:37 AM:i:0
XO:i:1 Xl:i:0 XM:i:0 XO:i:0 XG:i:0 MD:Z:37
19:20389:F:275+18M2D19M 99 1 17644 0 37M = 17919 314 TATGACTGCTAATAATACCTACACATGTTAGAACCAT
>>>>>>>>>>>>>>>><<<>><<>>4::>><9 RG:Z:UM0098:1 XT:A:R NM:i:0 SM:i:0 AM:i:0 X0:i:4 Xl:i:0 XM:i:0 XO:i:0
XG:i:0 MD:Z:37
19:20389:F:275+18M2D19M 147 1 17919 0 18M2D19M = 17644 -314
GTAGTACCAACTGTAAGTCCTTATCTTCATACTTTGT ;44999;499<8<8<<<8<<<<<<<<<<7<;<<<>><< XT:A:R NM:i:2 SM:i:0 AM:i:0 X0:i:4
Xl:i:0 XM:i:0 XO:i:1 XG:i:2 MD:Z:18^CA19

```

- QNAME: name of the short read present in FASTQ file
- FLAG: a number (“code”) describing the alignment
- RNAME: name of the reference sequence  
(often contains chromosome)
- POS: mapping position
- MAPQ: mapping quality
- CIGAR string: indicating alignment information
- RNEXT: reference sequence name for the next read
- **PNEXT**: mapping position for next read

[illegible]

- QNAME: name of the short read present in FASTQ file
- FLAG: a number (“code”) describing the alignment
- RNAME: name of the reference sequence  
(often contains chromosome)
- POS: mapping position
- MAPQ: mapping quality
- CIGAR string: indicating alignment information
- RNEXT: reference sequence name for the next read
- PNEXT: mapping position for next read
- TLEN: length of read group

```

1:497R:-272+13M17D24M 113 1 497 37 37M 15 100338662 0
CGGGTCTGACCTGAGGAGAAGTGTCTCCGCCTTCAG 0;==--=9;>>>>=>>>>>>>>=>>>>>>>> XT:A:U NM:i:0 SM:i:37 AM:i:0
XO:i:1 Xl:i:0 XM:i:0 XO:i:0 XG:i:0 MD:Z:37
19:20389:F:275+18M2D19M 99 1 17644 0 37M = 17919 314 TATGACTGCTAATAATACCTACACATGTTAGAACCAT
>>>>>>>>>>>>>>><<<>><<>4.:>><9 RG:Z:UM0098:1 XT:A:R NM:i:0 SM:i:0 AM:i:0 XO:i:4 Xl:i:0 XM:i:0 XO:i:0
XG:i:0 MD:Z:37
19:20389:F:275+18M2D19M 147 1 17919 0 18M2D19M = 17644 -314
GTAGTACCAACTGTAAGTCCTTATCTTCATACTTTGT ;44999;499<8<8<<<8<<<<<<<<<<7<;<<<>><< XT:A:R NM:i:2 SM:i:0 AM:i:0 XO:i:4
Xl:i:0 XM:i:0 XO:i:1 XG:i:2 MD:Z:18^CA19

```

- QNAME: name of the short read present in FASTQ file
- FLAG: a number (“code”) describing the alignment
- RNAME: name of the reference sequence  
(often contains chromosome)
- POS: mapping position
- MAPQ: mapping quality
- CIGAR string: indicating alignment information
- RNEXT: reference sequence name for the next read
- PNEXT: mapping position for next read
- TLEN: length of read group
- SEQ: short read sequence

[illegible]

# Alignment to a reference genome

- QNAME: name of the short read present in FASTQ file
- FLAG: a number (“code”) describing the alignment
- RNAME: name of the reference sequence (often contains chromosome)
- POS: mapping position
- MAPQ: mapping quality
- CIGAR string: indicating alignment information
- RNEXT: reference sequence name for the next read
- PNEXT: mapping position for next read
- TLEN: length of read group
- SEQ: short read sequence
- QUAL: short read base qualities

```
1:497:R:-272+13M17D24M 113 1 497 37 37M 15 100338662 0
CGGGTCTGACCTGAGGAGAACTGTGCTCCGCCTTCAG 0;===9;>>>>=>>>>>>>>>=>>>>>>>>> XT:A:U NM:i:0 SM:i:37 AM:i:0
X0:i:1 X1:i:0 XM:i:0 XO:i:0 XG:i:0 MD:Z:37
19:20389:F:275+18M2D19M 99 1 17644 0 37M = 17919 314 TATGACTGCTAATAATACCTACACATGTTAGAACCAT
>>>>>>>>>>>>>>>>>>><<<<<<<<<<<<<<4::>>><9 RG:Z:UM0098:1 XT:A:R NM:i:0 SM:i:0 AM:i:0 X0:i:4 X1:i:0 XM:i:0 XO:i:0
XG:i:0 MD:Z:37
19:20389:F:275+18M2D19M 147 1 17919 0 18M2D19M = 17644 -314
GTAGTACCAACTGTAAGTCCTTATCTTCATACTTTGT ;44999;499<8<8<<<8<<<<<<<<<<<<<<7<;<<<<<<<< XT:A:R NM:i:2 SM:i:0 AM:i:0 X0:i:4
X1:i:0 XM:i:0 XO:i:1 XG:i:2 MD:Z:18^CA19
```

- QNAME: name of the short read present in FASTQ file
- FLAG: a number (“code”) describing the alignment
- RNAME: name of the reference sequence (often contains chromosome)
- POS: mapping position
- MAPQ: mapping quality
- CIGAR string: indicating alignment information
- RNEXT: reference sequence name for the next read
- PNEXT: mapping position for next read
- TLEN: length of read group
- SEQ: short read sequence
- QUAL: short read base qualities
- TAGS: additional optional information

1:497:R:-272+13M17D24M      113          1            497         37           37M         15           100338662         0  
CGGGTCTGACCTGAGGAGAAGTGTCGCCCTCAG      0;==-=9;>>>>=>>>>>>>>>=>>>>>>>>  
**XO:i:1** **XI:i:0** **XM:i:0** **XO:i:0** **XG:i:0** **MD:Z:37**  
19:20389:F:275+18M2D19M      99          1            17644         0            37M         =            17919         314         TATGACTGCTAATAATACCTACACATGTTAGAACCAT  
>>>>>>>>>>>>>>><<<><<>>4:::>><9      **RG:Z:UM0098:1** **XT:A:R** **NM:i:0** **SM:i:0** **AM:i:0** **XO:i:4** **XI:i:0** **XM:i:0** **XO:i:0**  
**XG:i:0** **MD:Z:37**  
19:20389:F:275+18M2D19M      147          1            17919         0            18M2D19M         =            17644         -314  
GTAGTACCAACTGTAAGTCCTTATCTTCATACTTTGT      ;44999;499<8<8<<<8<<<<<<<<<7<;<<<><<  
**XI:i:0** **XM:i:0** **XO:i:1** **XG:i:2** **MD:Z:18^CA19** **XT:A:R** **NM:i:2** **SM:i:0** **AM:i:0** **XO:i:4**

# Alignment to a reference genome

Pileup files: `samtools mpileup -f refgenome.fa [options] s1.bam > s1.pup`

Tool: samtools

```
seq1 272 T 24 ,.$. , , , , , , , , , , , . ^ + . <<<+; <<<<<<<<<<=<;<; 7<&
seq1 273 T 23 ,. , , , , , , , , , , , . A <<<; <<<<<<<<<3<=<<<;<<+
seq1 274 T 23 ,. $. , , , , , , , , , , , . 7<7;<; <<<<<<<<<=<;<; <<6
seq1 275 A 23 , $. , , , , , , , , , , , . ^ 1 . <+; 9* <<<<<<<<<=<<:
```

- name of the reference sequence (usually the chromosome)



```
samtools mpileup -f refgenome.fa [options] s1.bam > s1.pup
```

## Tool: samtools

|      |     |   |    |                                                                                        |
|------|-----|---|----|----------------------------------------------------------------------------------------|
| seq1 | 272 | T | 24 | , . \$ . . . . , / . . . . , / . . . . ^ + . < < < + ; < < < < < < < < = < ; < ; 7 < & |
| seq1 | 273 | T | 23 | , . . . . , / . . . . , / . . . . A < < < ; < < < < < < < 3 < = < < < ; < < +          |
| seq1 | 274 | T | 23 | , . \$ . . . . , / . . . . , / . . . . 7 < 7 ; < ; < < < < < < < = < ; < ; < < 6       |
| seq1 | 275 | A | 23 | , \$ . . . . , / . . . . , / . . . . ^ 1 . < + ; 9 * < < < < < < < < = < < :           |

- name of the reference sequence (usually the chromosome)
- **genomic position** on the chromosome (contig)

```
samtools mpileup -f refgenome.fa [options] s1.bam > s1.pup
```

## Tool: samtools

```
seq1 272 T 24 , . $ / / ^ + . < < < + ; < < < < < < < < = < ; < ; 7 < &
seq1 273 T 23 , / / A < < < ; < < < < < < < 3 < = < < < ; < < +
seq1 274 T 23 , . $ / / 7 < 7 ; < ; < < < < < < < = < ; < ; < < 6
seq1 275 A 23 , $ / / ^ 1 . < + ; 9 * < < < < < < < < = < < :
```

- name of the reference sequence (usually the chromosome)
- genomic position on the chromosome (contig)
- the **reference base** at the genomic position

```
samtools mpileup -f refgenome.fa [options] s1.bam > s1.pup
```

## Tool: samtools

```
seq1 272 T 24 , . $ / , / / , / ^ + . < < < + ; < < < < < < < < = < ; < ; 7 < &
seq1 273 T 23 , / , / / , / . . . A < < < ; < < < < < < < 3 < = < < < ; < < +
seq1 274 T 23 , . $ / , / / , / 7 < 7 ; < ; < < < < < < < = < ; < ; < < 6
seq1 275 A 23 , $ / , / / , / ^ 1 . < + ; 9 * < < < < < < < < = < < :
```

- name of the reference sequence (usually the chromosome)
- genomic position on the chromosome (contig)
- the reference base at the genomic position
- **coverage**: the number of short reads aligned to the genomic position

# Alignment to a reference genome

Pileup files:

```
samtools mpileup -f refgenome.fa [options] s1.bam > s1.pup
```

Tool: samtools

```
seq1 272 T 24 ,. $. / / / ^+. <<<+; <<<<<<<<<<=<;<; 7<&
seq1 273 T 23 ,. / / / A <<<; <<<<<<<<<3<=<<<; <<+
seq1 274 T 23 ,. $. / / / 7<7; <; <<<<<<<<<=<;<; <<6
seq1 275 A 23 ,. $. / / / ^1. <+; 9*<<<<<<<<<=<:
```

- name of the reference sequence (usually the chromosome)
- genomic position on the chromosome (contig)
- the reference base at the genomic position
- coverage: the number of short reads aligned to the genomic position
- the **bases aligned to the genomic position** (they originate from *different* short reads) (ref: .,)

```
samtools mpileup -f refgenome.fa [options] s1.bam > s1.pup
```

## Tool: samtools

|      |     |   |    |                                                    |                                |
|------|-----|---|----|----------------------------------------------------|--------------------------------|
| seq1 | 272 | T | 24 | , . \$ . . . . . , / . . . . . , / . . . . . ^ + . | <<<+ ; <<<<<<<<<<=< ; < ; 7<&  |
| seq1 | 273 | T | 23 | , . . . . . , / . . . . . , / . . . . . A          | <<< ; <<<<<<<<<3<=<<< ; <<+    |
| seq1 | 274 | T | 23 | , . \$ . . . . . , / . . . . . , / . . . . .       | 7<7 ; < ; <<<<<<<<=< ; < ; <<6 |
| seq1 | 275 | A | 23 | , \$ . . . . . , / . . . . . , / . . . . . ^ 1 .   | <+ ; 9*<<<<<<<<=<< :           |

- name of the reference sequence (usually the chromosome)
- genomic position on the chromosome (contig)
- the reference base at the genomic position
- coverage: the number of short reads aligned to the genomic position
- the bases aligned to the genomic position (they originate from *different* short reads) (ref: .,)
- the **base qualities** of the aligned bases

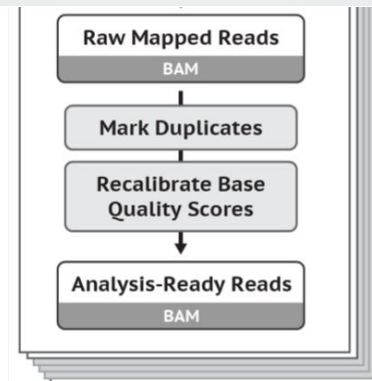
# Preprocessing alignment files

Input: alignment files (**BAM files**)

Goals:

- Sorting BAM files **Tool: samtools**

```
samtools sort -o s1_sorted.bam s1.bam
```



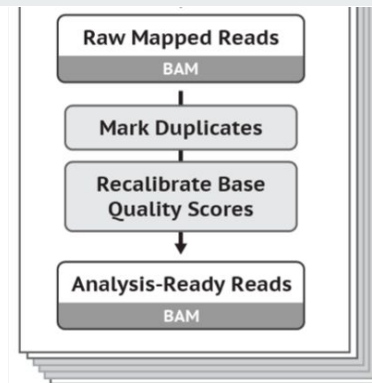
# Preprocessing alignment files

Input: alignment files (**BAM files**)

Goals:

- Sorting BAM files
- Marking duplicate reads

**Tool: Picard Tools | MarkDuplicates**



```
java -jar picard.jar MarkDuplicates \
 I=input.bam \
 O=marked_duplicates.bam \
 M=marked_dup_metrics.txt
```

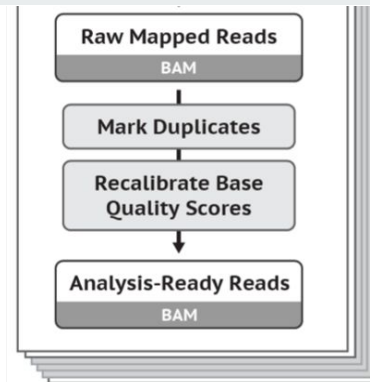
# Preprocessing alignment files

Input: alignment files (**BAM files**)

Goals:

- Sorting BAM files
- Marking duplicate reads
- Adding read groups if necessary

**Tool: Picard Tools | AddOrReplaceReadGroups**



```
java -jar picard.jar AddOrReplaceReadGroups \
 INPUT=s1_RMdup.bam OUTPUT=s1_RG.bam \
 RGLB=lib1 RGPL=illumina RGPU=unit1 \
 RGSM=1
```

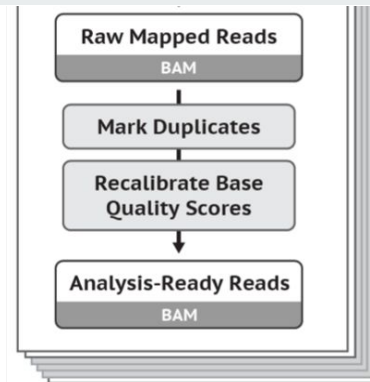


# Preprocessing alignment files

Input: alignment files (**BAM files**)

Goals:

- Sorting BAM files
- Marking duplicate reads
- Adding read groups if necessary
- Recalibrate base quality scores



**Tools: GATK Base Recalibrator, ApplyBQSR**

```
gatk BaseRecalibrator \
 -I my_reads.bam \
 -R reference.fasta \
 --known-sites sites_of_variation.vcf \
 --known-sites another/optional/setOfSitesToMask.vcf \
 -O recal_data.table
```

```
gatk ApplyBQSR \
 -R reference.fasta \
 -I input.bam \
 --bqsr-recal-file
 recalibration.table \
 -O output.bam
```

# Preprocessing alignment files

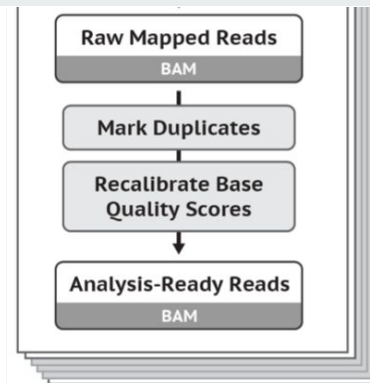
Input: alignment files (**BAM files**)

Goals:

- Sorting BAM files
- Marking duplicate reads
- Adding read groups if necessary
- Recalibrate base quality scores
- Indexing BAM files

**Tool: samtools**

```
samtools index s1_RG.bam
```



# Preprocessing alignment files

Input: alignment files (**BAM files**)

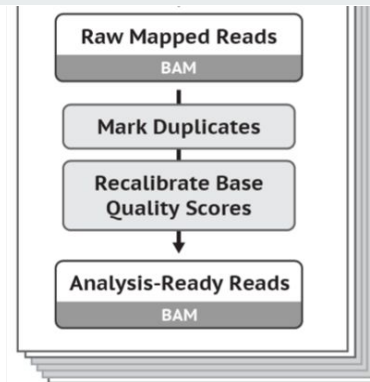
Goals:

- Sorting BAM files
- Marking duplicate reads
- Adding read groups if necessary
- Recalibrate base quality scores
- Indexing BAM files

**Tool: samtools**

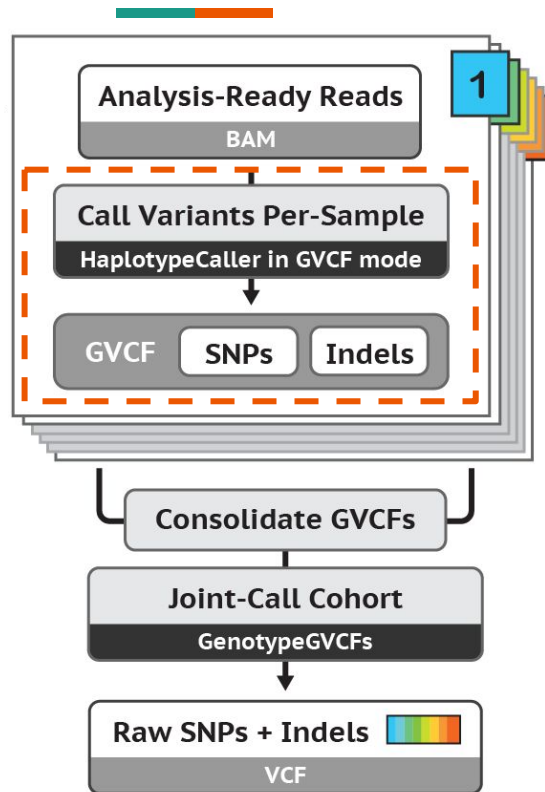
```
samtools index s1_RG.bam
```

Output: modified alignment files (**BAM files**)



# Variant calling: germline variants

Tool: GATK

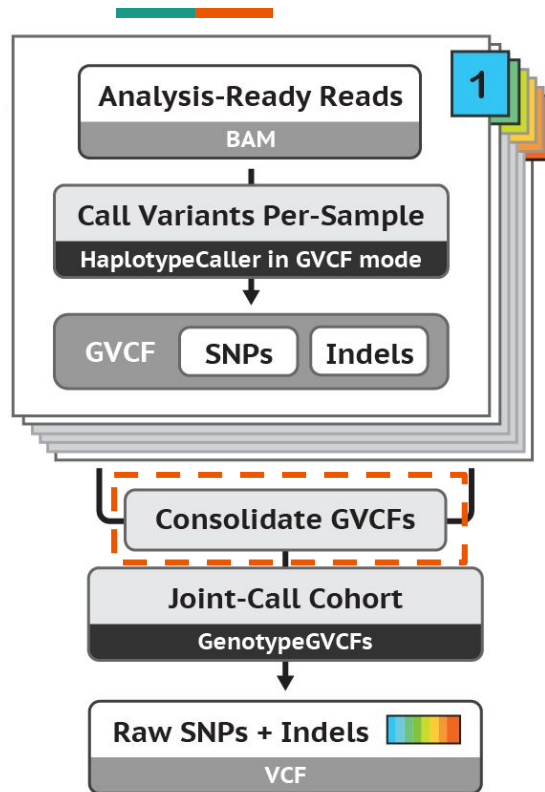


1. Analyse each sample and determine if the given genomic position has any variation (compared to reference) (one **GVCF file** per sample)

```
gatk HaplotypeCaller \
 -R refgenome.fa \
 -I s1_RG.bam \
 -O s1.raw.g.vcf.gz \
 -ERC GVCF
```

# Variant calling: germline variants

Tool: GATK

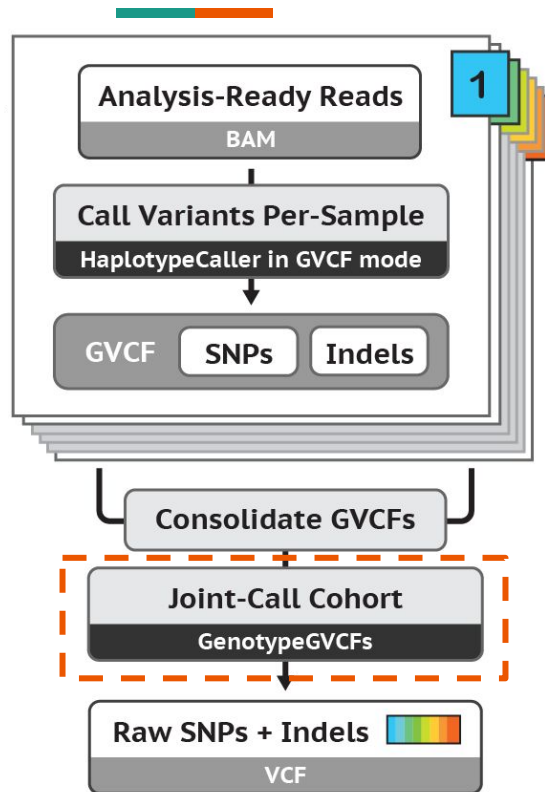


1. Analyse each sample and determine if the given genomic position has any variation (compared to reference) (one **GVCF file** per sample)
2. Combine GVCF files to a **database** (one common database for all samples)

```
gatk GenomicsDBImport \
 -V s1.raw.g.vcf.gz \
 -V s2.raw.g.vcf.gz \
 -V s3.raw.g.vcf.gz \
 [-V ...] \
 -V sn.raw.g.vcf.gz \
 --genomicsdb-workspace-path my_database \
 -L chr19
```

# Variant calling: germline variants

Tool: GATK

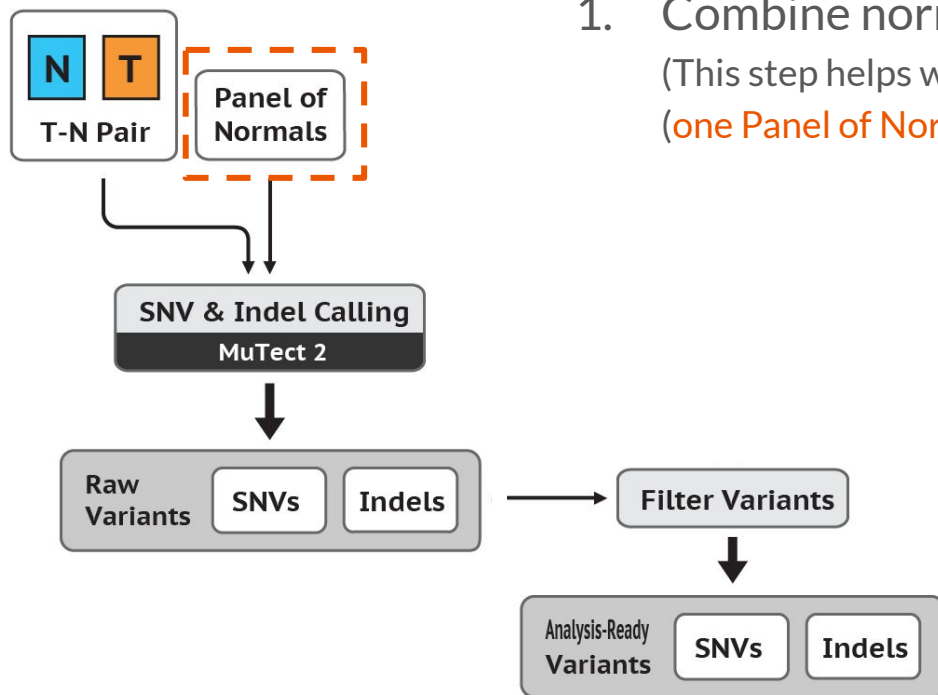


1. Analyse each sample and determine if the given genomic position has any variation (compared to reference) (one **GVCF file** per sample)
2. Combine GVCF files to a **database** (one common database for all samples)
3. Check **each genomic position** and determine the **genotype of each sample**:
  - homozygous reference (same as reference)
  - heterozygous (about half of the aligned bases are non-reference)
  - homozygous non-reference (none of the aligned bases is reference)

```
gatk GenotypeGVCFs \
-R refgenome.fa \
-V gendb://my_database \
-O jointgenotypingresults.vcf.gz
```

# Variant calling: somatic variants

Tool: GATK



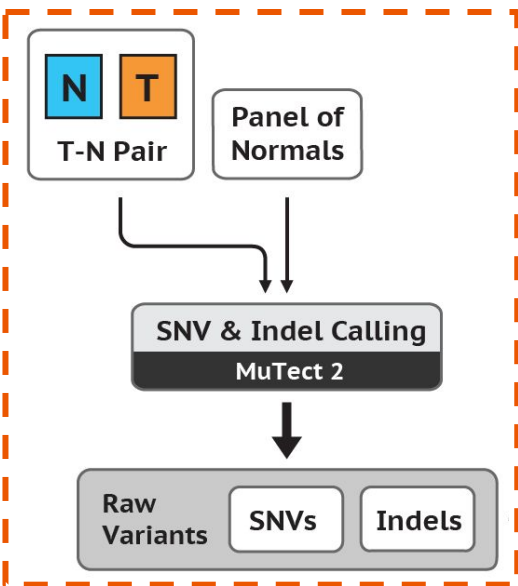
1. Combine normal samples to **Panel of Normals (PoN)**  
(This step helps with the filtering of common germline variations)  
(one Panel of Normals VCF file from many normal samples)

```
gatk Mutect2 \
 -R refgenome.fa \
 -I s1_RG.bam \
 -tumor s1_sample_name \
 -O s1_pon.vcf.gz
```

```
gatk CreateSomaticPanelOfNormals \
 -vcfs s1_pon.vcf.gz \
 -vcfs s2_pon.vcf.gz \
 [-vcfs ...]
 -vcfs sn_pon.vcf.gz \
 -O combined_pon.vcf.gz
```

# Variant calling: somatic variants

Tool: GATK



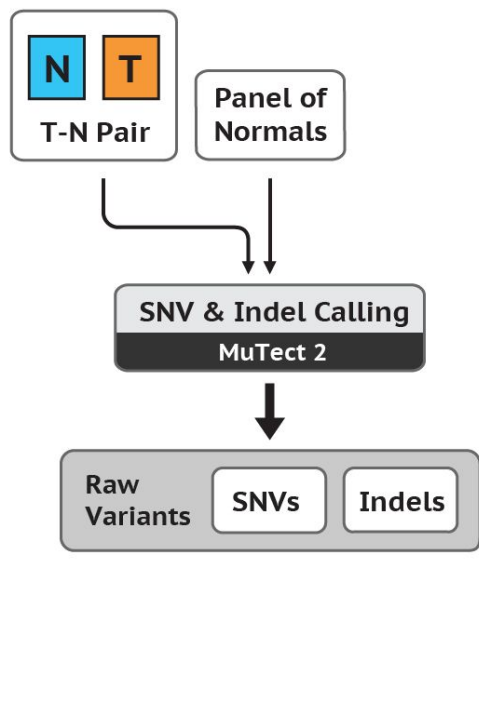
1. Combine normal samples to **Panel of Normals (PoN)**  
(This step helps with the filtering of common germline variations)  
(one Panel of Normals VCF file from many normal samples)
2. **Compare normal-tumor sample pairs** and use PoN as an additional reference  
(one VCF file containing raw variant calls for each normal-tumor sample pair)

```
gatk Mutect2 \
 -R refgenome.fa \
 -I s1_RG.bam \
 -I s2_RG.bam \
 -tumor s1_sample_name \
 -normal s2_sample_name \
 -pon combined_pon.vcf.gz \
 -L chr19 \
 -O s1_somatic_m.vcf.gz
```



# Variant calling: somatic variants

Tool: GATK

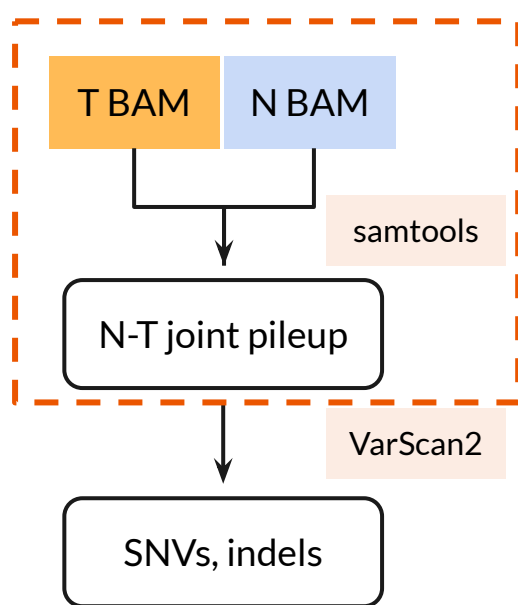


1. Combine normal samples to **Panel of Normals (PoN)**  
(This step helps with the filtering of common germline variations)  
(one Panel of Normals VCF file from many normal samples)
2. **Compare normal-tumor sample pairs** and use PoN as an additional reference  
(one VCF file containing raw variant calls for each normal-tumor sample pair)
3. Further filter variants  
(one VCF file containing final variant calls for each normal-tumor sample pair)

```
gatk FilterMutectCalls \
-V s1_somatic_m.vcf.gz \
-O s1_somatic_filtered.vcf.gz
```

# Variant calling: another method

Tool: VarScan2

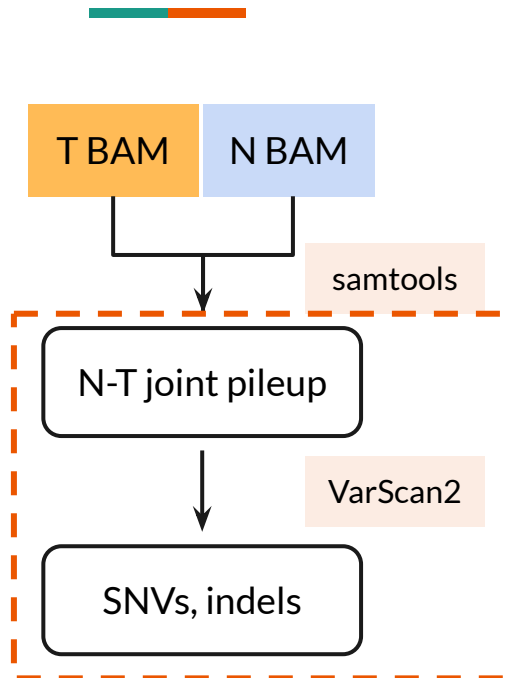


1. Generate a **joint pileup** file for each tumor normal pair

```
samtools mpileup \
-f refgenome.fa \
[options] \
s2_RG.bam s1_RG.bam > s2_s1.pup
```

# Variant calling: another method

Tool: VarScan2



1. Generate a **joint pileup** file for each tumor normal pair
2. **Call short indels and SNVs** with VarScan2  
(categorized as “somatic”, “germline” or “LOH”)

```
java -jar VarScan.v2.4.3.jar somatic \
s2_s1.pup \
1_somatic_varscan \
--mpileup 1
```

# Variant calling: VCF files

HEADER: starts with “##”, contains information about the structure of the file

```
##fileformat=VCFv4.3
##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
```

# Variant calling: VCF files

Field names: starts with “#”, name of columns

```
##fileformat=VCFv4.3
##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      | NA000001       | NA000002       | NA000003    |
|--------|---------|-----------|-----|--------|------|--------|-----------------------------------|-------------|----------------|----------------|-------------|
| 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    |

# Variant calling: VCF files

Variants: actual data lines, with columns defined above

```
##fileformat=VCFv4.3
##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 NA000001 NA000002 NA000003
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
```

# Variant calling: VCF files



Examples: how many samples were analysed?

```
##fileformat=VCFv4.3
##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
```

# Variant calling: VCF files



Examples: how many samples were analysed?

```
##fileformat=VCFv4.3
##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
```



# Variant calling: VCF files

Examples: how many samples were analysed?

```
##fileformat=VCFv4.3
##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
```

# Variant calling: VCF files



Examples: how many samples were analysed?

```
##fileformat=VCFv4.3
##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 1 2 3
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
```

# Variant calling: VCF files



Examples: was the 20:17330 variant filtered? Why?

```
##fileformat=VCFv4.3
##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
```

# Variant calling: VCF files

Examples: was the 20:17330 variant filtered? Why?

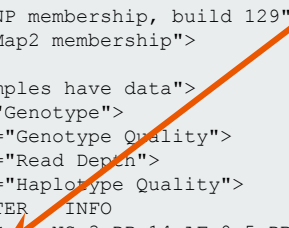
```
##fileformat=VCFv4.3
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number 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 rs6040555 A G,T 67 PASS NS=2;DP=10;AF=0.333,0.667;AA=1;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
```

# Variant calling: VCF files

Examples: was the 20:17330 variant filtered? Why?

```
##fileformat=VCFv4.3
##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 rs6040555 A G,T 67 PASS NS=2;DP=10;AF=0.333,0.667;AA=1;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
```

Yes



# Variant calling: VCF files

Examples: was the 20:17330 variant filtered? Why?

```
##fileformat=VCFv4.3
##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 rs6040555 A G,T 67 PASS NS=2;DP=10;AF=0.333,0.667;AA=1;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
```

Yes, because the quality was below 10.

# Variant calling: VCF files



Examples: what were the genotypes of the 3 samples in position 20:14370?

```
##fileformat=VCFv4.3
##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
```

# Variant calling: VCF files

Examples: what were the genotypes of the 3 samples in position 20:14370?

```
##fileformat=VCFv4.3
##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 NA000001 NA000002 NA000003
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 PASS 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
```



# Variant calling: VCF files

Examples: what were the genotypes of the 3 samples in position 20:14370?

```
##fileformat=VCFv4.3
##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">
```

Genotype ID: GT

| #CHROM | POS     | ID        | REF | ALT    | QUAL | FILTER | INFO                              | FORMAT      | NA000001       | NA000002       | NA000003     |
|--------|---------|-----------|-----|--------|------|--------|-----------------------------------|-------------|----------------|----------------|--------------|
| 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     |

# Variant calling: VCF files

Examples: what were the genotypes of the 3 samples in position 20:14370?

```
##fileformat=VCFv4.3
##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 NA000001 NA000002 NA000003
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
```

Genotype ID: GT

GT: first field in FORMAT

# Variant calling: VCF files

```
##fileformat=VCFv4.3
##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 NA000001 NA000002 NA000003
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
```

Genotype ID: GT

GT: first field in FORMAT

# Variant calling: VCF files

Examples: what were the genotypes of the 3 samples in position 20:14370?

```
##fileformat=VCFv4.3
##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">
```

0/0: homozygous reference

Genotype ID: GT

GT: first field in FORMAT

| CHROM | POS     | ID        | REF | ALT    | QUAL | FILTER | INFO                              | FORMAT      | NA000001       | NA000002       | NA000003     |
|-------|---------|-----------|-----|--------|------|--------|-----------------------------------|-------------|----------------|----------------|--------------|
| 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     |

# Variant calling: VCF files

Examples: what were the genotypes of the 3 samples in position 20:14370?

```
##fileformat=VCFv4.3
##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">
```

0/0: homozygous reference

1/0: heterozygous

Genotype ID: GT

GT: first field in FORMAT

| CHROM | POS     | ID        | REF | ALT    | QUAL | FILTER | INFO                              | FORMAT      | NA000001       | NA000002       | NA000003    |
|-------|---------|-----------|-----|--------|------|--------|-----------------------------------|-------------|----------------|----------------|-------------|
| 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    |

# Variant calling: VCF files

Examples: what were the genotypes of the 3 samples in position 20:14370?

##fileformat=VCFv4.3  
##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      | NA000001       | NA000002       | NA000003    |
|-------|---------|-----------|-----|--------|------|--------|-----------------------------------|-------------|----------------|----------------|-------------|
| 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    |

0/0: homozygous reference      1/0: heterozygous      1/1: homozygous non-reference

Genotype ID: GT

GT: first field in FORMAT

# Interpreting results: general questions



- How many mutations were detected in each sample?
  - **Note:** different variant callers result in *extremely different* variant lists
- Are these mutations **novel or already listed in available databases?** (e.g. dbSNP)
- Are the mutations located in **specific genes?** Which ones? What are the **roles of these genes?**
- What are the **consequences** of these mutations?
- Is there a **specific pattern** in mutations? (e.g. most of them are C>T)
- etc.

# Interpreting results: annotation

Input: list of variants (final **VCF files**)

Goal: search for each mutation (chrom, pos, ref, alt) in available databases

Output: annotated variant list (tables, csv files)

**Tool: ANNOVAR**

| Chr | Start     | End       | Ref | Alt   | Func.refGene | Gene.refGene | GeneDetail.refGene              | ExonicFunc.refGene         | AAChange.refGene                                                                                                                                                                                                      | Xref.refGene                  | ExAC_Freq |
|-----|-----------|-----------|-----|-------|--------------|--------------|---------------------------------|----------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------|-----------|
| 1   | 948921    | 948921    | T   | C     | UTR5         | ISG15        | NM_005101:c.-33T>C              | .                          | .                                                                                                                                                                                                                     | Immunodeficiency              | 0.941     |
| 1   | 1404001   | 1404001   | G   | T     | UTR3         | ATAD3C       | NM_001039211:c.*91G>T           | .                          | .                                                                                                                                                                                                                     | .                             | 0.054     |
| 1   | 5935162   | 5935162   | A   | T     | splicing     | NPHP4        | NM_001291594:exon17:c.1282-2T>A | .                          | .                                                                                                                                                                                                                     | Nephronophthisis              | 0.825     |
| 1   | 162736463 | 162736463 | C   | T     | intronic     | DDR2         | .                               | .                          | .                                                                                                                                                                                                                     | Spondylometaphyseal dysplasia | .         |
| 1   | 84875173  | 84875173  | C   | T     | intronic     | DNASE2B      | .                               | .                          | .                                                                                                                                                                                                                     | .                             | .         |
| 1   | 13211293  | 13211294  | TC  | -     | intergenic   | PRAMEF36P    | F dist=11566;dist=116902        | .                          | .                                                                                                                                                                                                                     | .                             | .         |
| 1   | 11403596  | 11403596  | -   | AT    | intergenic   | UBIAD1       | PTCH dist=55105;dist=135699     | .                          | .                                                                                                                                                                                                                     | .                             | .         |
| 1   | 105492231 | 105492231 | A   | ATAAA | intergenic   | LOC10012911  | dist=872538;dist=640085         | .                          | .                                                                                                                                                                                                                     | .                             | .         |
| 1   | 67705958  | 67705958  | G   | A     | exonic       | IL23R        | .                               | nonsynonymous SNV          | IL23R:NM_144701:exon9:c.G1142A:p.R381Q                                                                                                                                                                                | .                             | 0.041     |
| 2   | 234183368 | 234183368 | A   | G     | exonic       | ATG16L1      | .                               | nonsynonymous SNV          | ATG16L1:NM_198890:exon5:c.A409G:p.T137A;ATG16L1:NM_001293557:exon3:c.C2023T:p.R675W;NOD2:NM_001293557:exon7:c.G2641C:p.G881R;NOD2:NM_001293557:exon10:c.2936dupC:p.L980Pfs*2;NOD2:NM_004004:exon2:c.35delG:p.G12Vfs*2 | .                             | 0.457     |
| 16  | 50745926  | 50745926  | C   | T     | exonic       | NOD2         | .                               | nonsynonymous SNV          | NOD2:NM_001293557:exon3:c.C2023T:p.R675W;NOD2:NM_001293557:exon7:c.G2641C:p.G881R;NOD2:NM_001293557:exon10:c.2936dupC:p.L980Pfs*2;NOD2:NM_004004:exon2:c.35delG:p.G12Vfs*2                                            | Blau syndrome, A              | 0.023     |
| 16  | 50756540  | 50756540  | G   | C     | exonic       | NOD2         | .                               | nonsynonymous SNV          | NOD2:NM_001293557:exon3:c.C2023T:p.R675W;NOD2:NM_001293557:exon7:c.G2641C:p.G881R;NOD2:NM_001293557:exon10:c.2936dupC:p.L980Pfs*2;NOD2:NM_004004:exon2:c.35delG:p.G12Vfs*2                                            | Blau syndrome, A              | 0.009917  |
| 16  | 50763778  | 50763778  | -   | C     | exonic       | NOD2         | .                               | frameshift insertion       | NOD2:NM_001293557:exon3:c.C2023T:p.R675W;NOD2:NM_001293557:exon7:c.G2641C:p.G881R;NOD2:NM_001293557:exon10:c.2936dupC:p.L980Pfs*2;NOD2:NM_004004:exon2:c.35delG:p.G12Vfs*2                                            | Blau syndrome, A              | 0.013     |
| 13  | 20763686  | 20763686  | G   | -     | exonic       | GJB2         | .                               | frameshift deletion        | GJB2:NM_004004:exon2:c.35delG:p.G12Vfs*2                                                                                                                                                                              | Bart-Pumphrey syndrome        | 0.006038  |
| 13  | 20797176  | 21105944  | O   | -     | exonic       | CRYL1        | GJB6                            | frameshift deletion        | GJB6:NM_001110220:wholegene;GJB6:NM_001110221:wholegene                                                                                                                                                               | .                             | .         |
| 8   | 8887543   | 8887543   | A   | T     | exonic       | ERI1         | .                               | stoploss                   | ERI1:NM_153332:exon7:c.A1049T:p.X350L                                                                                                                                                                                 | .                             | .         |
| 8   | 8887539   | 8887539   | A   | T     | exonic       | ERI1         | .                               | stopgain                   | ERI1:NM_153332:exon7:c.A1045T:p.K349X                                                                                                                                                                                 | .                             | .         |
| 8   | 8887536   | 8887537   | AG  | GATT  | exonic       | ERI1         | .                               | frameshift substitution    | ERI1:NM_153332:exon7:c.1042_1043GATT:p.R348Dfs*2                                                                                                                                                                      | .                             | .         |
| 8   | 8887540   | 8887540   | G   | GGAA  | exonic       | ERI1         | .                               | nonframeshift substitution | ERI1:NM_153332:exon7:c.1046delinsGGAA:p.R348_K349insR                                                                                                                                                                 | .                             | .         |
| 5   | 1295288   | 1295288   | G   | A     | upstream     | TERT         | dist=126                        | .                          | .                                                                                                                                                                                                                     | .                             | .         |



# Problems



- Different tools result in **very different results**. (Sometimes there is no overlap between the list of detected variants.)
  - Which one is better? **Which results are true?**
  - *Why* do they detect different variants? What filtering steps do they use?
  - How can we compare results?
- Different tools have **very different runtimes**.
  - Is it realistic to use tools that run for weeks?
  - How many samples do we have?
  - What kind of computer do we have?

# General guidelines



- Be very patient both with the computer and yourself.
- Many commands take a lot of time to run. Think before you start running a faulty pipeline on many samples.
- You *will* come across error messages. Don't panic, Google.
- Practice makes perfect, don't give up.
- Always make sure your results make sense. (E.g. Try interchanging tumor and normal samples, do you get only a few somatic mutations in normals?)
- If you find anything strange, don't sweep it under the rug.

Good luck!



**Thank you  
for your attention!**