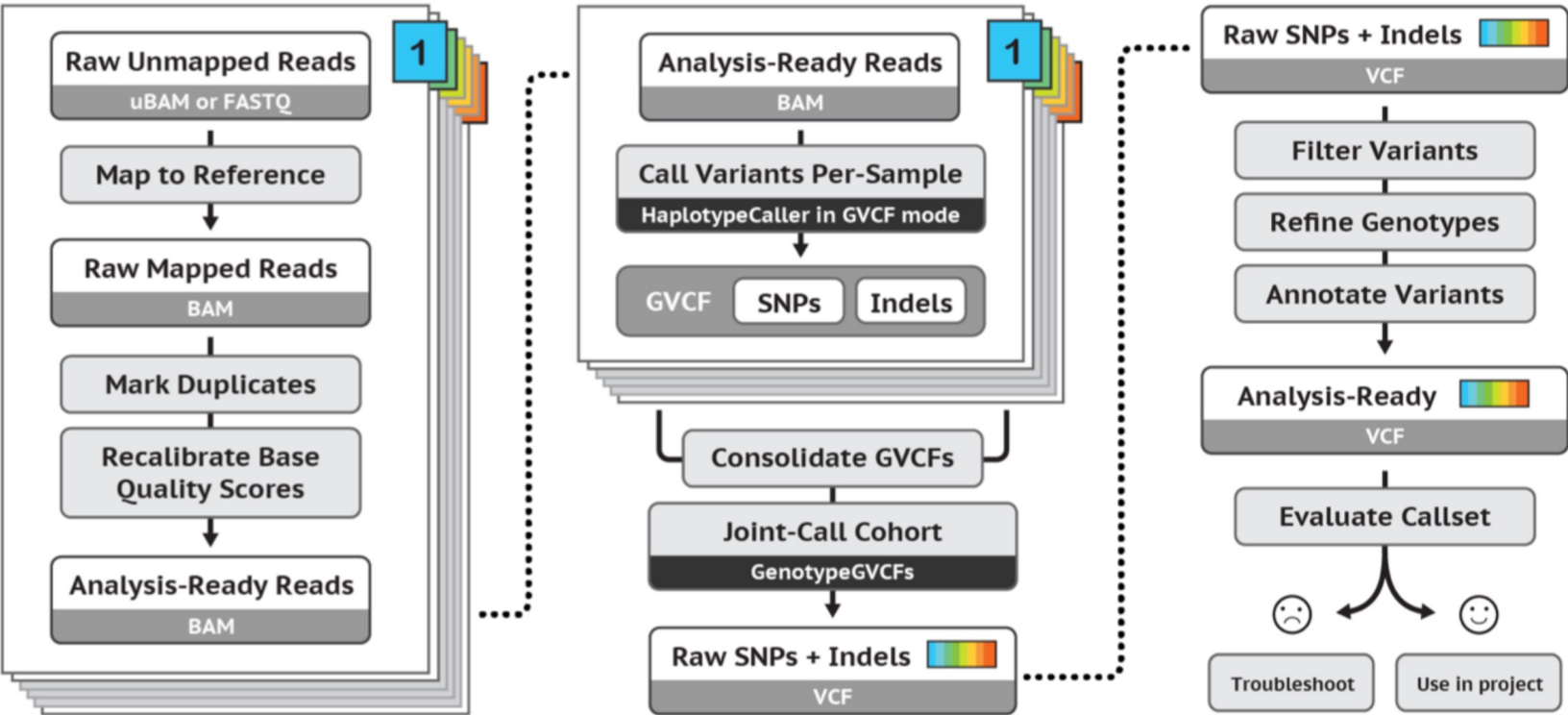


# NGS - variant analysis

Filtering and evaluation

# GATK workflow

Yesterday

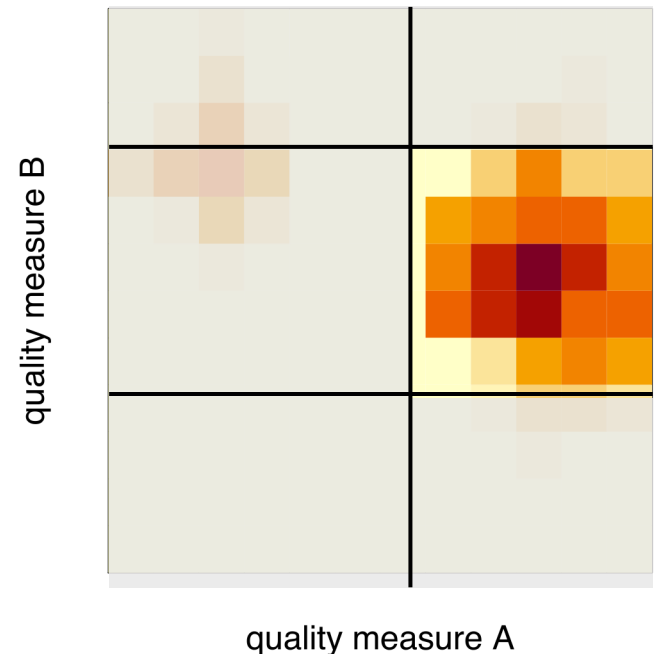


# Important QC info

- Mapping quality
- Depth
- Strand-bias
- ..

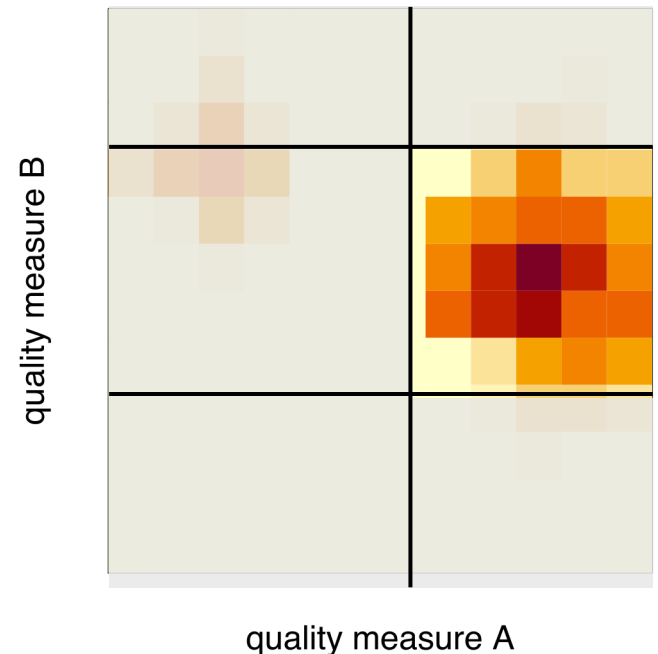
# Filtering

- Hard filtering: take thresholds of each measure
- Machine-learning-based: VQSR



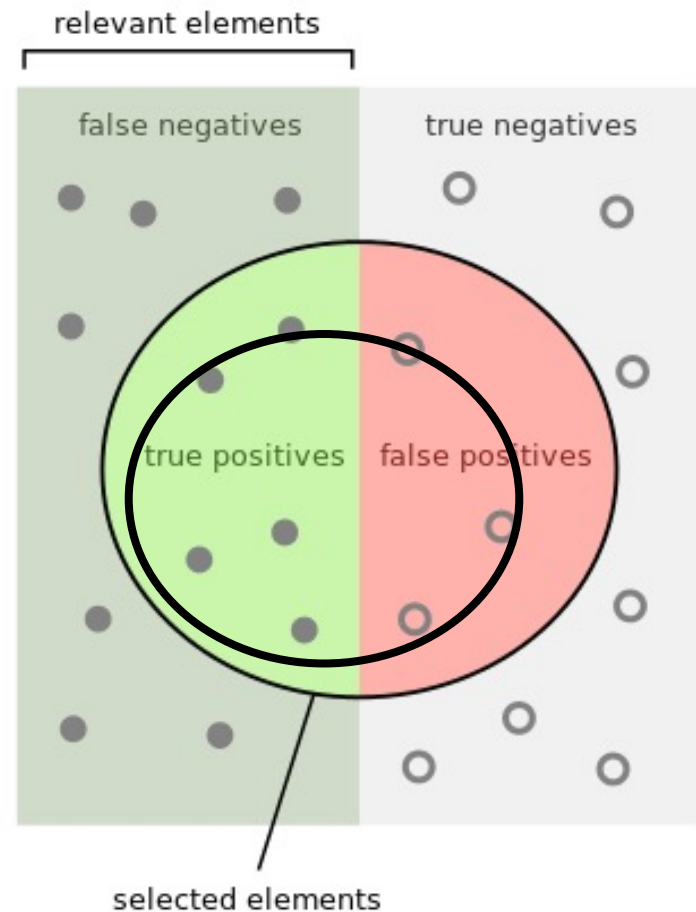
# VQSR

- Better performance compared to hard filtering, but you will need:
  - Truth-set
  - 'Large' dataset (whole genome or 30 whole exomes)



# Evaluation

- **Precision:** how many of the selected variants were true variants?
- **Recall:** how many of the true variants were selected?



How many selected items are relevant?

$$\text{Precision} = \frac{\text{true positives}}{\text{true positives} + \text{false positives}}$$

How many relevant items are selected?

$$\text{Recall} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}$$

# Quiz question 9