Microsoft Research
Summit 2022

An introduction to ag-genomics

Angels de Luis Balaguer

Outline

- Introduction: The social impact of ag-genomics
- Data analysis for genetic variant identification
- Running variant calling jobs in the cloud
- Demo: running a workflow on Cromwell on Azure
- Tertiary analysis: from SNPs to agricultural answers

Introduction: the social impact of ag-genomics

Feeding a growing population

The world population is expected to grow to 9 billion by 2050

Feeding the population will require massive increases in global food production

Factors such as climate change, pests, and plant disease will further hinder that goal

An agricultural revolution that uses genomics and bioinformatics and advanced computing resources will play a key role in addressing this challenge

Genetic diversity can be key

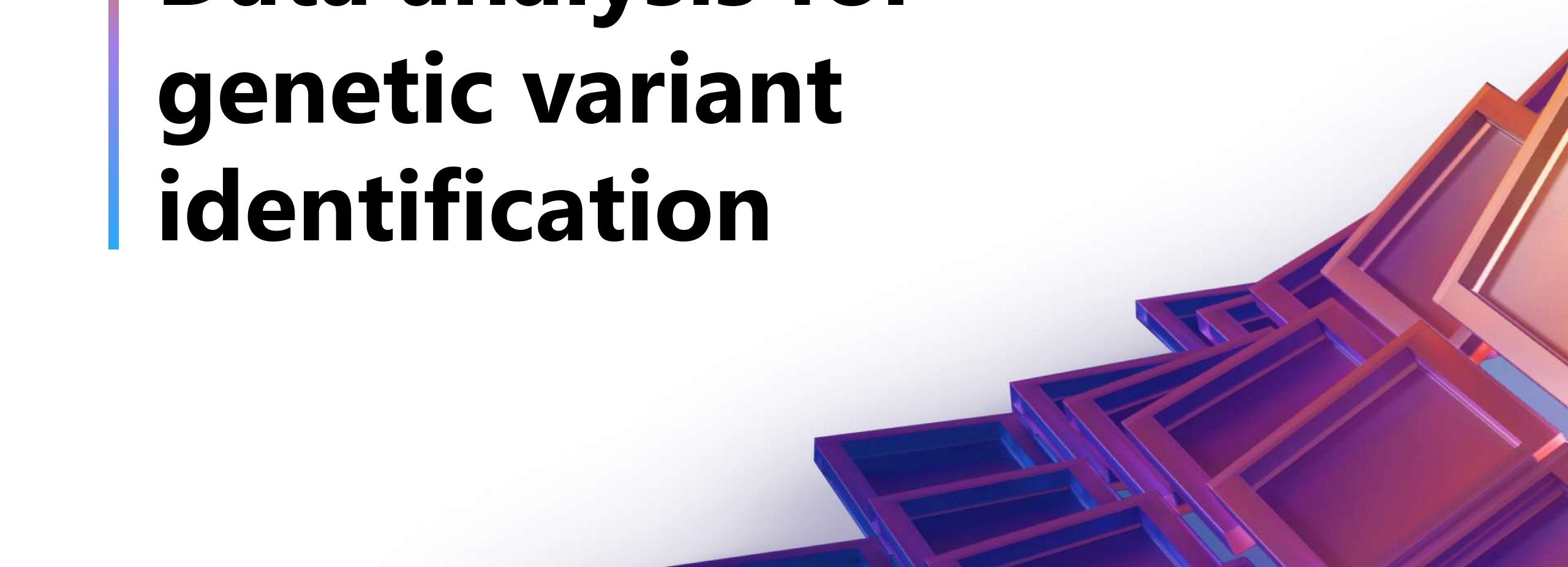
- Over centuries, breeding has increased the nutritional content of domesticated crops, but this has brought its own challenges:
  - Fewer variants are cultivated, so there is less genetic variety in modern crops.
  - Therefore, crops are more susceptible to pests, diseases, and a changing climate.
- At the core of the solution to tackle this issue is genomics:
  - Analyzing the genomes of many different varieties, old ones and new ones, can allow us to transfer genes back to current varieties.

The genomics solution

- The number of targets for crop improvement are limited due to a lack of understanding of biological process involving genes, pathways, networks, and their interactions with environmental factors.
- We can integrate multidisciplinary approaches, including genomics, agriculture, and computer science, to identify targets that both bear potential for beneficial agronomic traits and are suitable for genome editing.

- Seedbanks and genebanks are collections where agricultural species are stored and maintained.
- For the most part, they have not been sequenced and analyzed
- Performing studies to genetically characterize entire gene-banks can be a rich source of information to reintroduce genomic diversity
- Sequencing and analyzing seed and gene banks can lead to rich genetic information

Data analysis for genetic variant identification



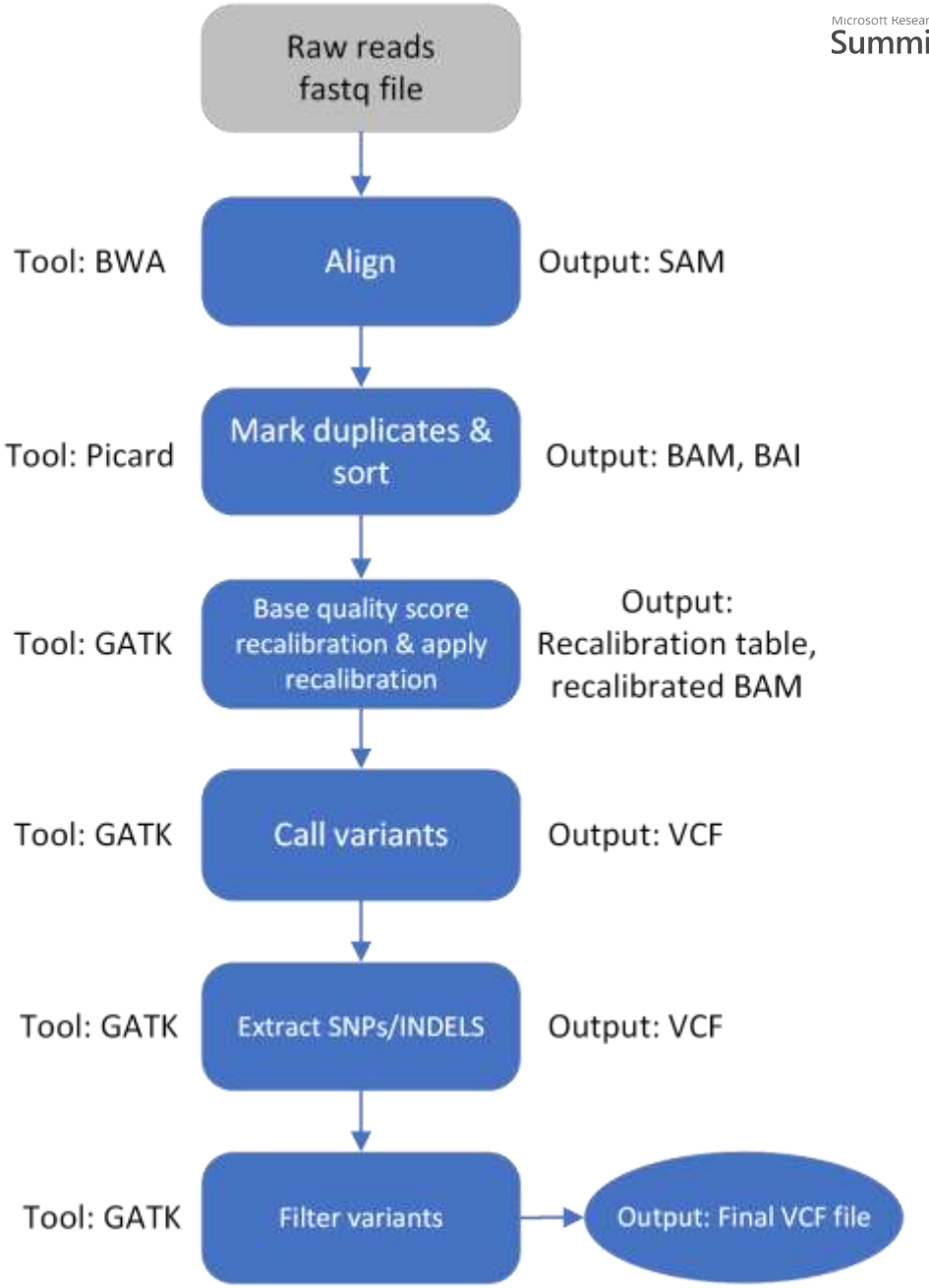
Variant calling: a key step to approach many of the big challenges

Genetic variant identification is a recurrent activity needed to solve big challenges in genomics

Reference	T	T	A	C	A	A	G	Phenotype
Individual 1	T	A	T	C	A	T	G	Disease
Individual 2	T	T	T	G	G	A	G	Disease
Individual 3	G	T	A	C	G	A	G	No disease
Individual 4	G	T	A	C	C	G	A	No disease

Best practices for calling variants

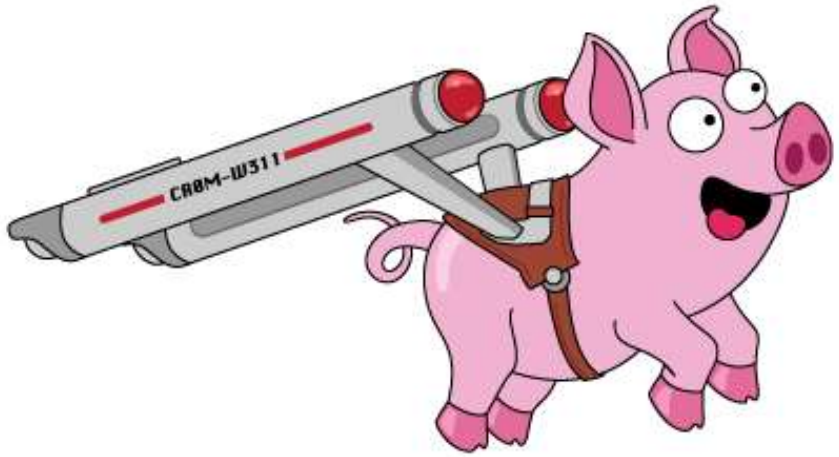
Example: schematic of a basic variant calling pipeline for 1 sample based on BWA and GATK



Running variant calling jobs in the cloud

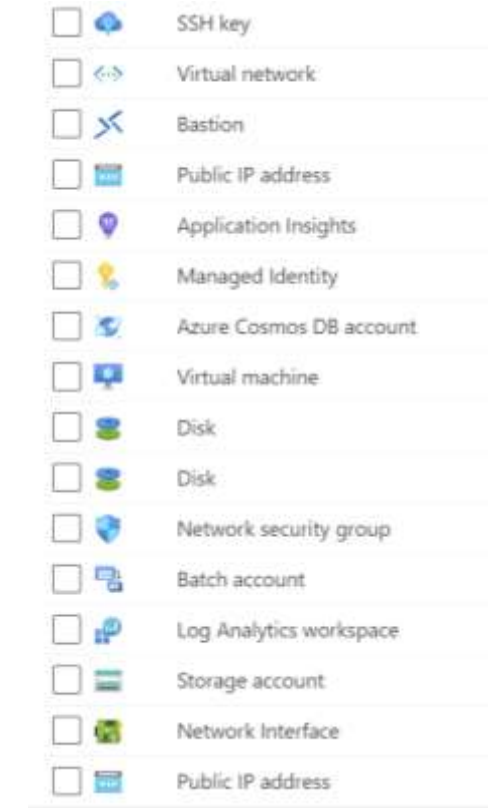
Cromwell

- Open-source workflow management system for scientific workflows
- Orchestrates the computing tasks needed for genomics analysis
- Originally developed by the Broad Institute
  - Used in the GATK Best Practices genome analysis pipeline
- Supports running scripts at various scales, including local machine, local computing cluster, and cloud.



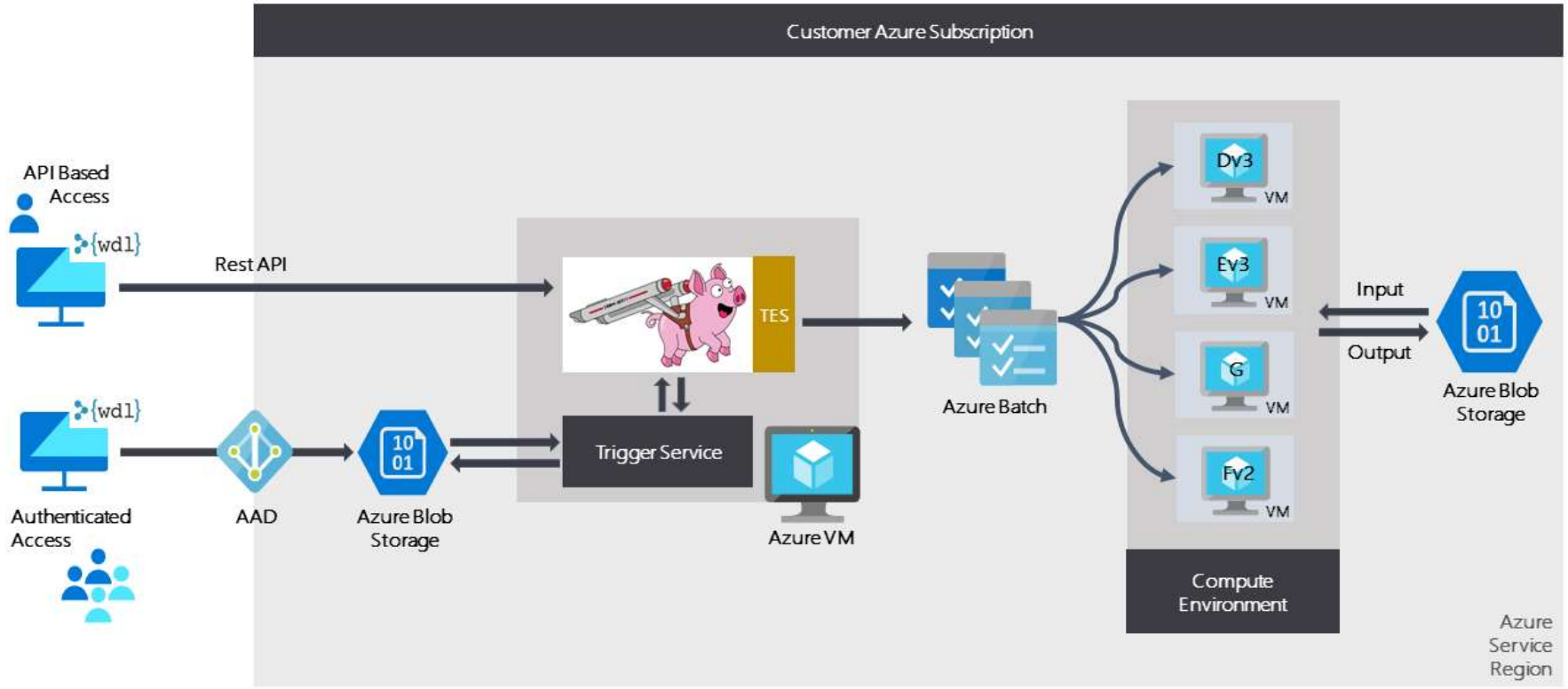
Cromwell on Azure

Cromwell on Azure configures all Azure resources needed to run workflows through Cromwell on the Azure cloud



Cromwell's backend

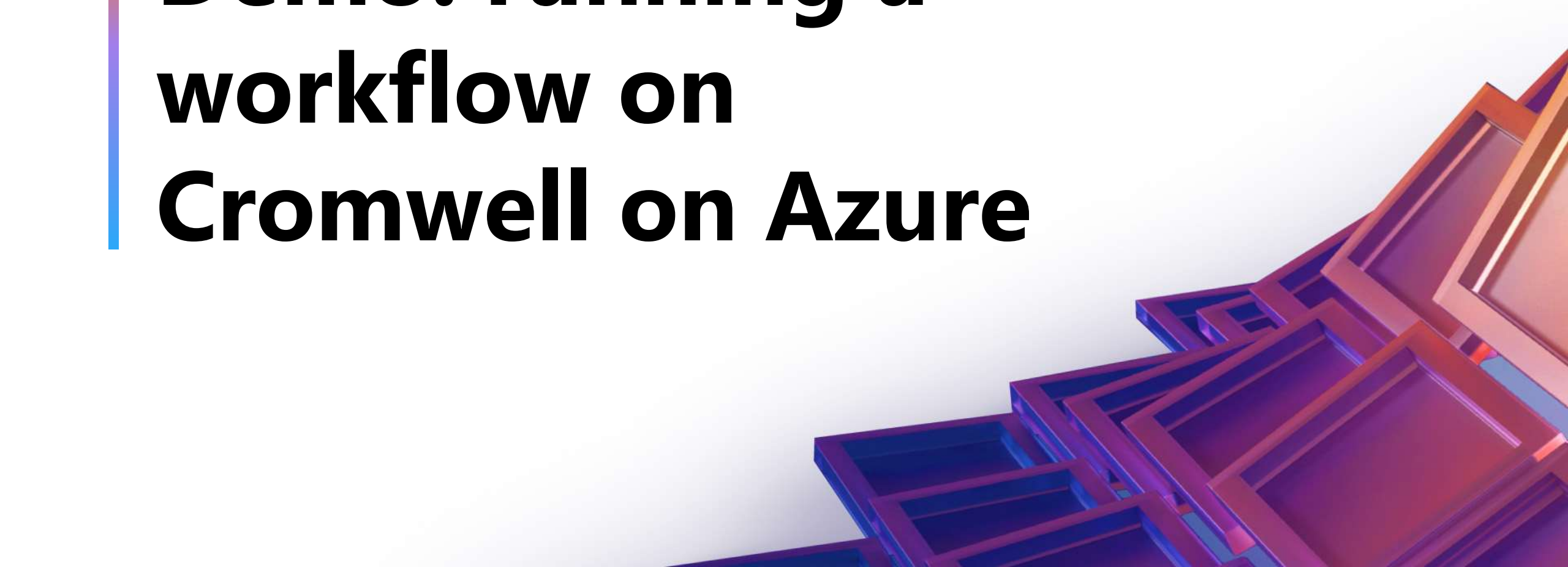
Uses the GA4GH TES backend for orchestrating the tasks that create a workflow



Cromwell workflows

- Cromwell workflows can be written using either the WDL or the CWL scripting languages
  - The Workflow Description Language (WDL) is a way to specify data processing workflows with a human-readable and -writeable syntax. WDL makes it straightforward to define analysis tasks, chain them together in workflows, and parallelize their execution.
  - The Common Workflow Language (CWL) is a standard for describing computational data-analysis workflows. Development of CWL is focused particularly on serving the data-intensive sciences, such as Bioinformatics, Medical Imaging, Astronomy, Physics, and Chemistry.

Demo: running a workflow on Cromwell on Azure



Sequence file format conversion

fastq2bam.wdl

```
task task1 {
  String input_arg1

  command {
    bash_command
  }
  output {

  }
  runtime {
  docker: 'docker-container'
  }
}

workflow fastq2bam {
  call task1
}
```

fastq2bam.inputs.json

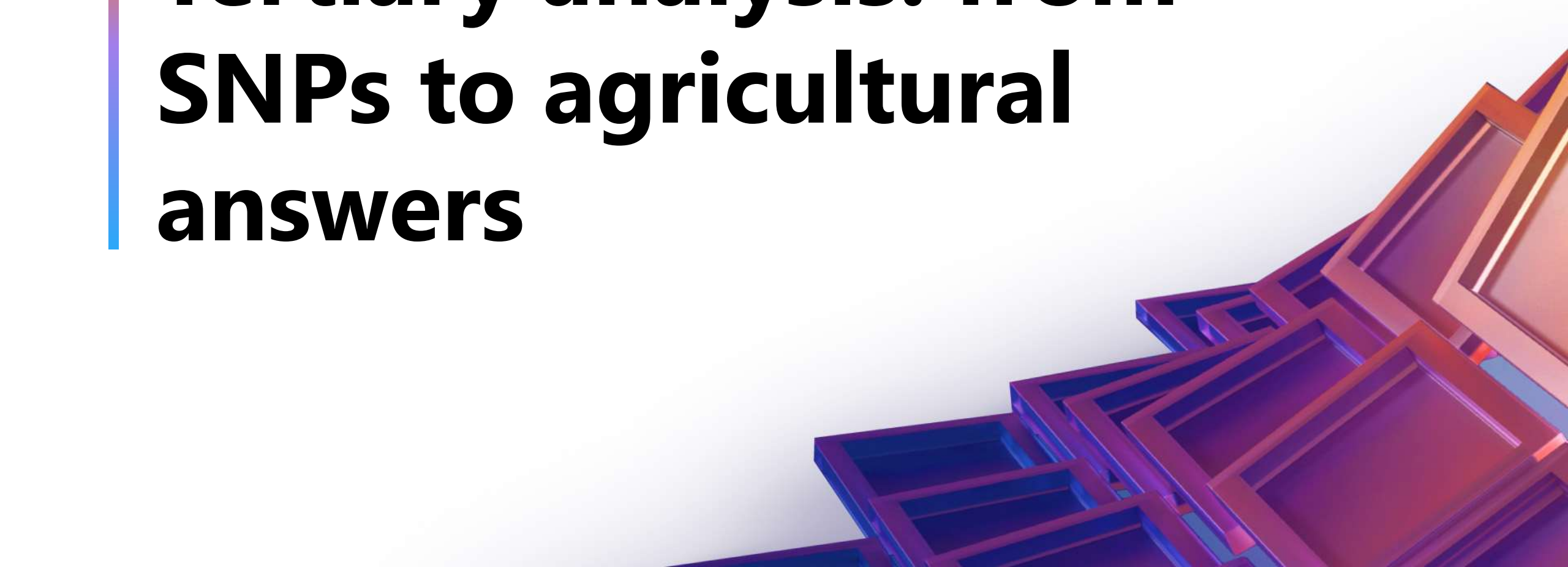
```
{
  "fastq2bam.input_arg1":
"argument_1",
  "fastq2bam.fastq_1":
"/path/to/forwardRead.fastq",
  " fastq2bam.fastq_2":
"/path/to/reverseRead.fastq"
}
```

fastq2bam.trigger.json

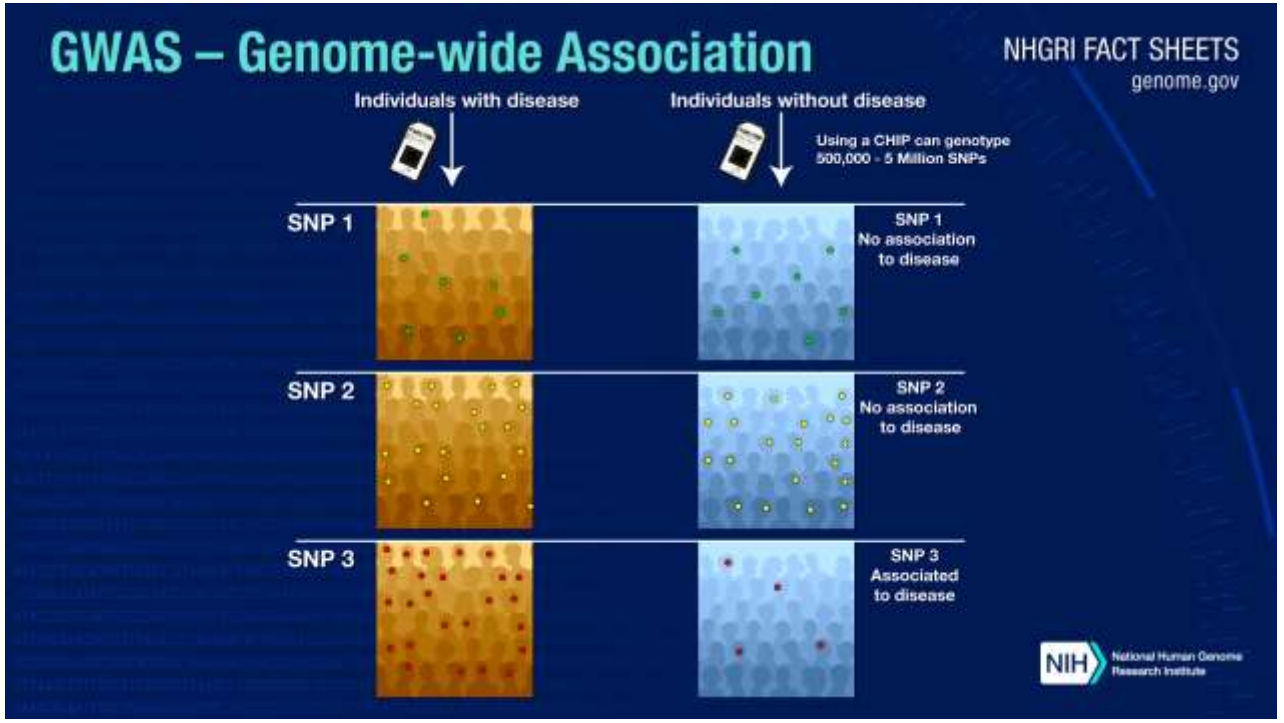
```
{
  "WorkflowUrl":
"/path/to/fastq2bam.wdl",
  "WorkflowInputsUrl":
"/path/to/fastq2bam.inputs.json",
  "WorkflowOptionsUrl": null,
  "WorkflowDependenciesUrl": null
}
```

https://github.com/microsoft/seq-format-conversion-azure

Tertiary analysis: from SNPs to agricultural answers



From SNPs to agricultural answers



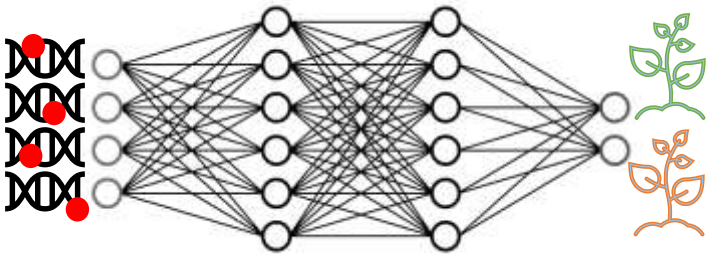
Towards AI genomics

Genomic data has been used by plant/animal breeders to perform genomic selection of desirable traits

Statistical methods have been used to do this selection, either from genomics data alone or in combination with environmental data

Statistical methods are limited by the linear relationship assumption among the features

AI approaches can overcome this limitation due to their ability to model complex non-linear interactions



Microsoft Research
Summit 2022

Thank you