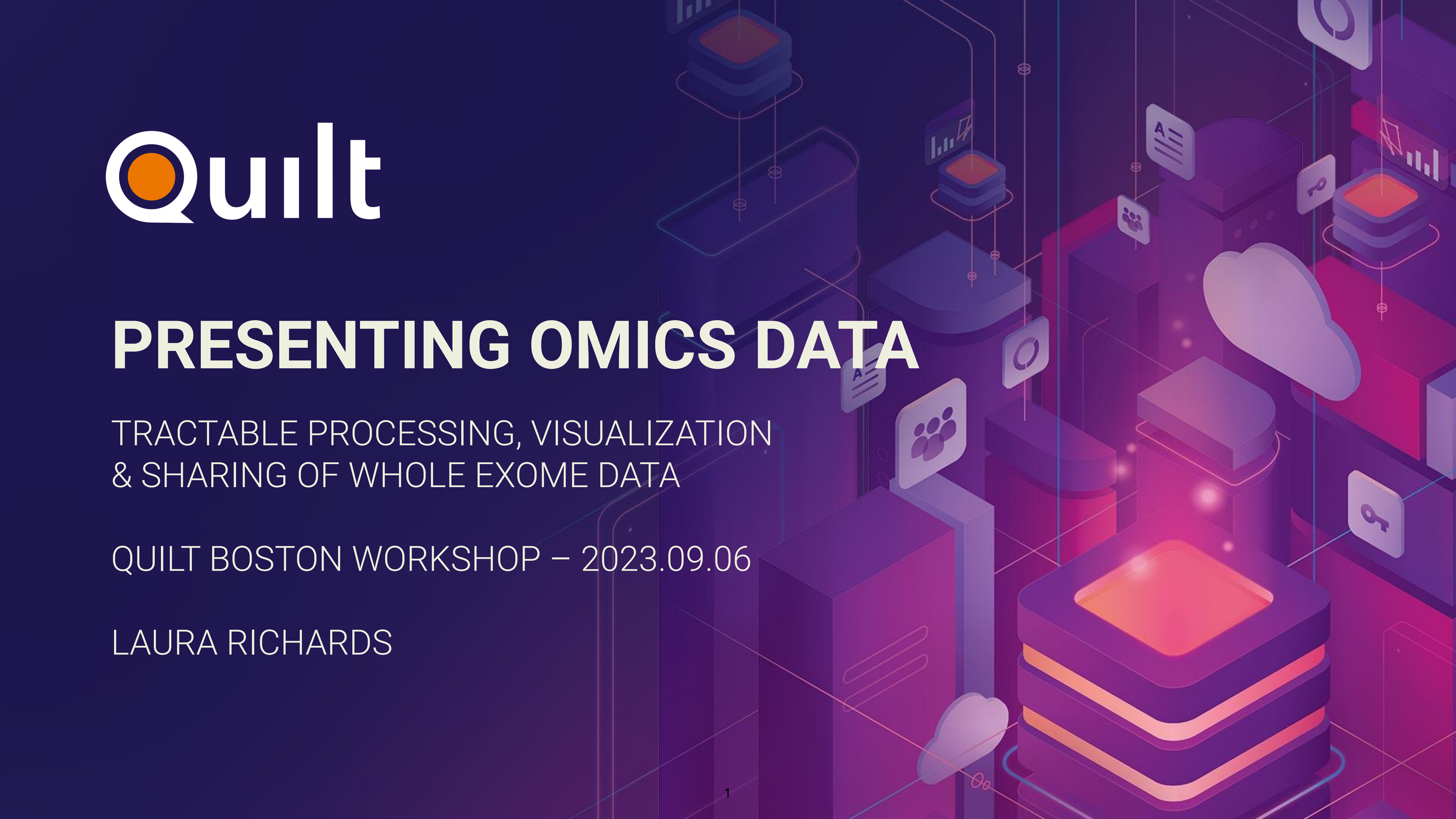


PRESENTING OMICS DATA

TRACTABLE PROCESSING, VISUALIZATION & SHARING OF WHOLE EXOME DATA

QUILT BOSTON WORKSHOP – 2023.09.06

LAURA RICHARDS

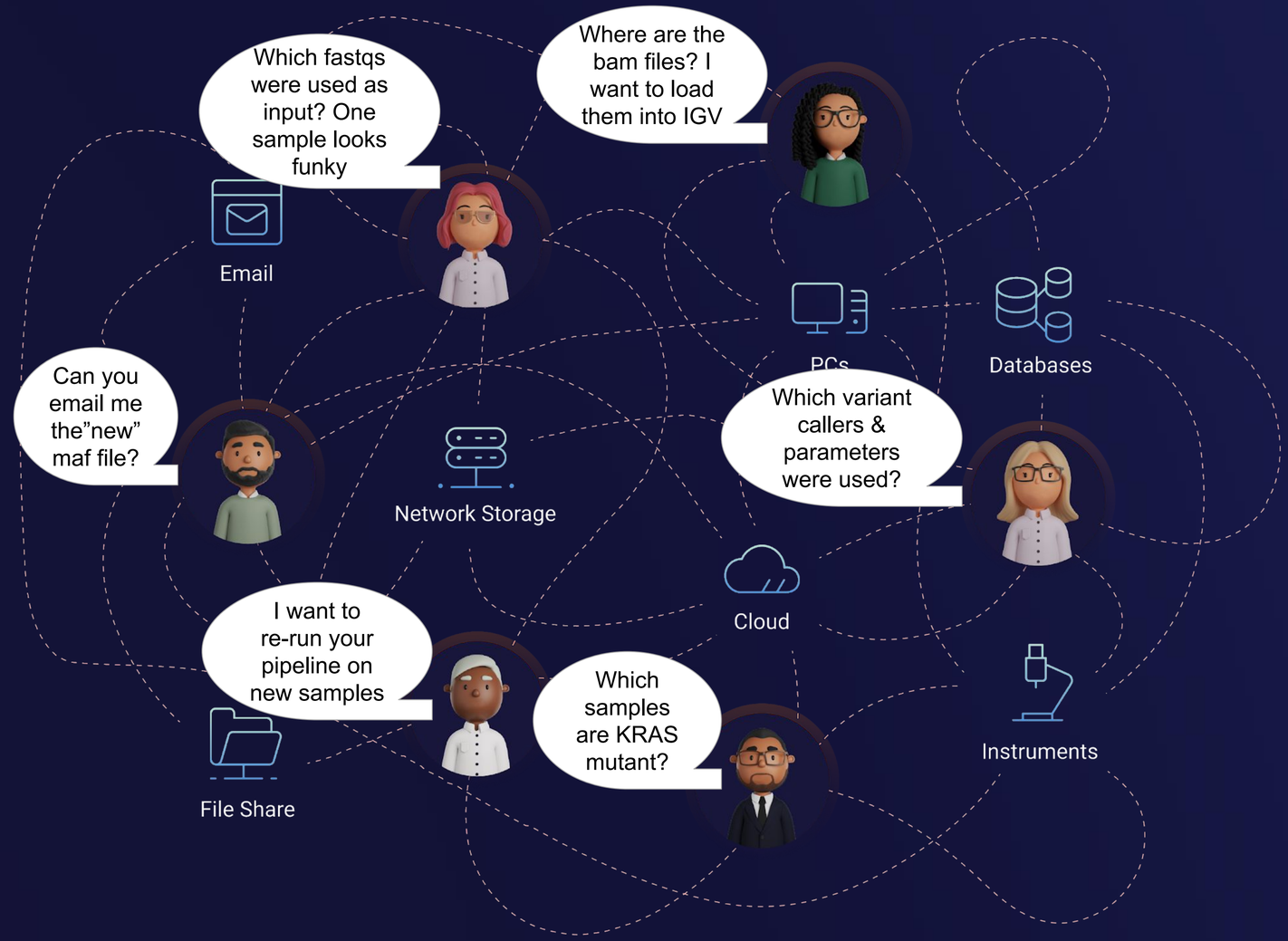
HELLO!

- Computational Biologist
  - Princess Margaret Cancer Centre (Toronto, ON)
  - Celsius Therapeutics (Cambridge, MA)

- Focus: Genomics to discover new therapeutic targets in oncology
  - Single cell RNA-sequencing
  - Whole exome & genome sequencing
  - Genome-wide CRISPR screens

PROBLEM: WES DATA IS TRADITIONALLY INACCESSIBLE

- Whole exome sequencing (WES) data is big!
- Pipeline output often accessible by command line only
- Challenging to share data across all steps of the WES pipeline – from alignment to bespoke analyses



WES DATA IS USED ACROSS DIVERSE RESEARCH TEAMS

	Sample Acquisition & Processing	Sequencing	Preprocessing & QC	Discovery, Interpretation & Validation
Teams	Project Management Clinical Operations Platform	Platform NGS Sequencing Core	Data Engineering Computational Biology Platform	Computational Biology Biologists Business Development
Use	Sample Tracking Track Data Production	Generate Data Transfer Data Catalog Data	Run Pipelines Flag Failed Samples Evaluate & Revise Protocol	Analysis Mechanistic Hypotheses Discover Novel Assets
Data	Sample Metadata Sample Sheets Mutations (vcfs, mafs)	Sample Sheets Raw Data (bcl, fastqs)	FastQC, MultiQC QC Metrics (depth …) Alignments (bams)	Custom Plots & Files Mutations (vcfs, mafs) Alignments (bams)

Users with variable programming abilities interact with WES data across all steps of workflow

SOLUTION: PACKAGE & SHARE WES DATA WITH QUILT

PROCESSING PIPELINE

Raw Data (fastqs, bams ...)

Execution Commands & Logs

Pipeline Outputs

Findable, Accessible, Browsable

QUILT PACKAGE

- Versioning
- Shareable Links
- Interactive Visualizations

DOWNSTREAM ANALYSES

Jupyter Notebooks

Sample Metadata

Custom Files

Analysis Results

CASE STUDY: PROCESSING & ANALYZING CCLE WES DATA

Raw WES Data

- Data Source: Cancer Cell Line Encyclopedia (CCLE)
- 9 KRAS-mutant samples from 3 cancer indications
- Fastqs downloaded from Sequence Read Archive (SRA)

(Pre-)Processing Pipeline

- Nextflow nf-core/sarek pipeline, tumor-only mode
- Alignment & Preprocessing
- Variant Calling & Annotation
  - mutect2, strelka, bcftools, freebayes, VEP
- Summary Reports

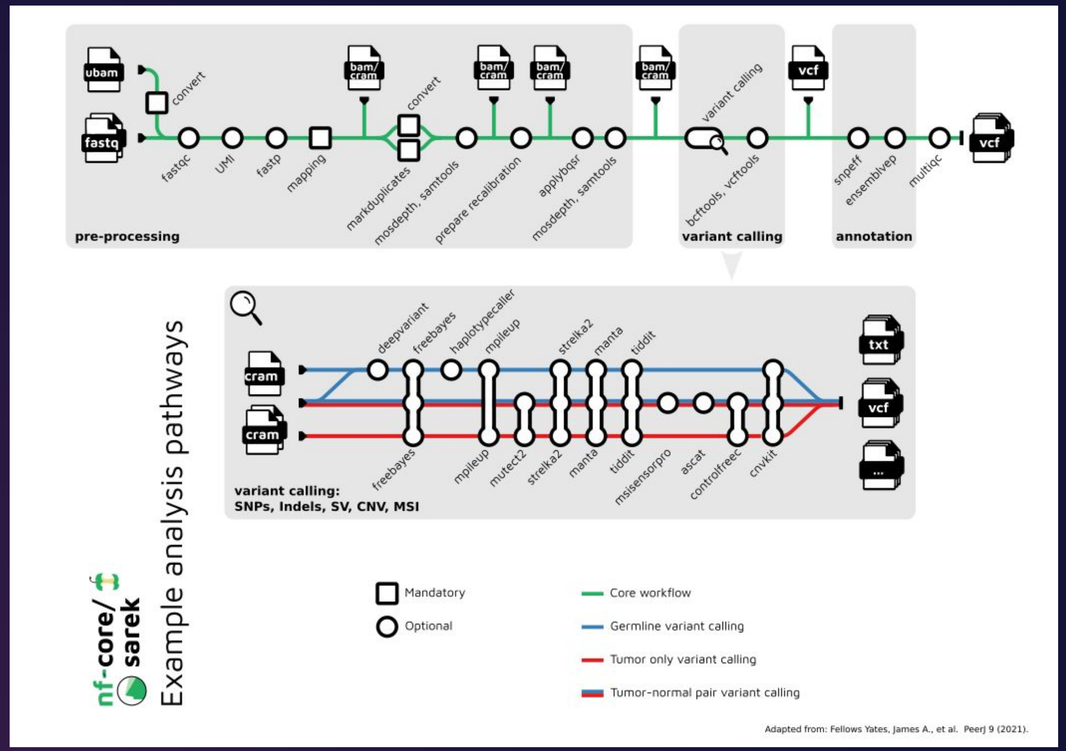
Downstream Analysis

- Convert vcfs to mafs
- Filter & query variants
- Custom analyses & outputs
- Inspect alignments & mutations
- Plot sample metadata and mutation metics

Quilt packages enables efficient data access, tracking & sharing across *all stages* of the WES workflow

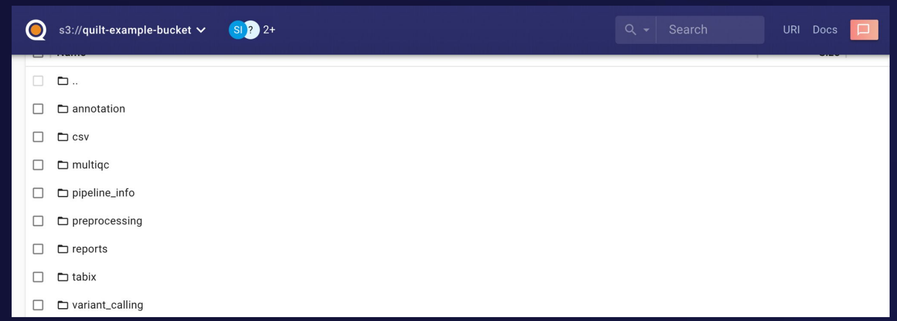
PACKAGING NEXTFLOW OUTPUTS WITH NF-QUILT

Re-process CCLE data with nf-core/sarek



https://nf-co.re/sarek/3.2.3

Outputs packaged with nf-quilt



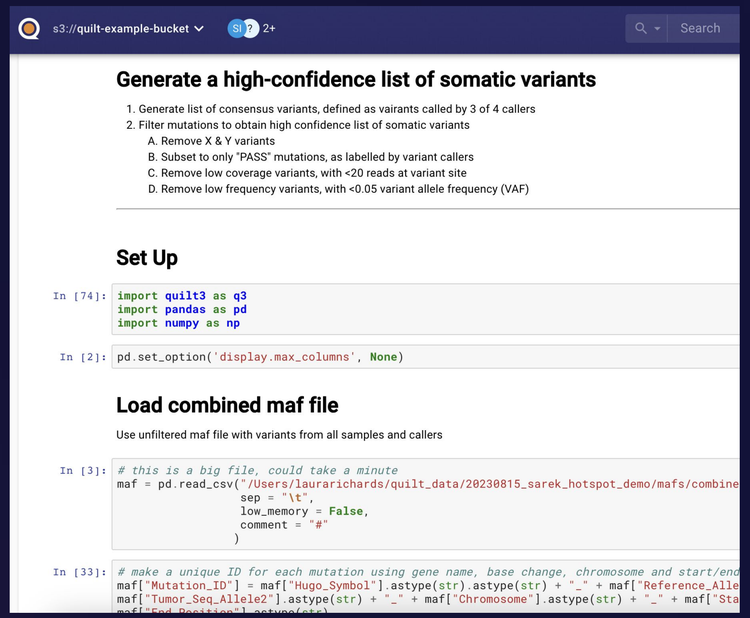
https://docs.quiltdata.com/examples/nextflow

HARMONIZED PIPELINE OUTPUTS, NOTEBOOKS & RESULTS

"How does fastq X look?"

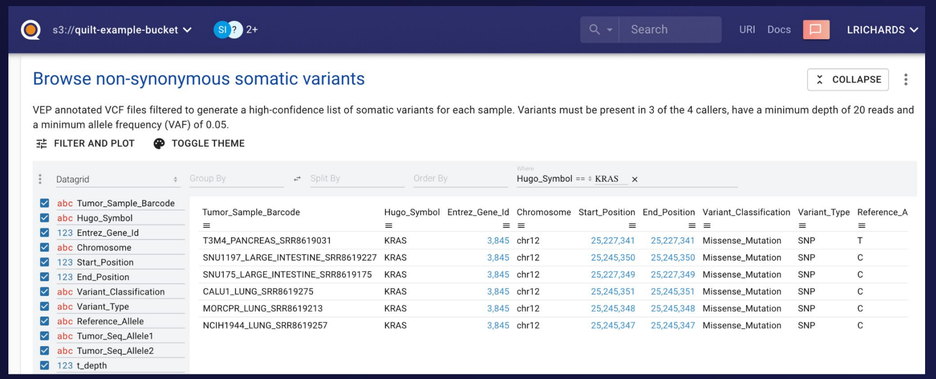
https://quilt-example-bucket.s3.amazonaws.com/wes/result/multiqc/multiqc_report.html?versionId=j41NK60ggyiNi7npG.TXxjMII9qLiSan

"How were mutations filtered?"



https://quilt-example-bucket.s3.amazonaws.com/wes/result/notebooks/3_filter_somatic_variants.ipynb?versionId=Safrp27FnAwuo.RRj4MeTb0EQAD3.Qrc

"Which samples are KRAS mutant?"



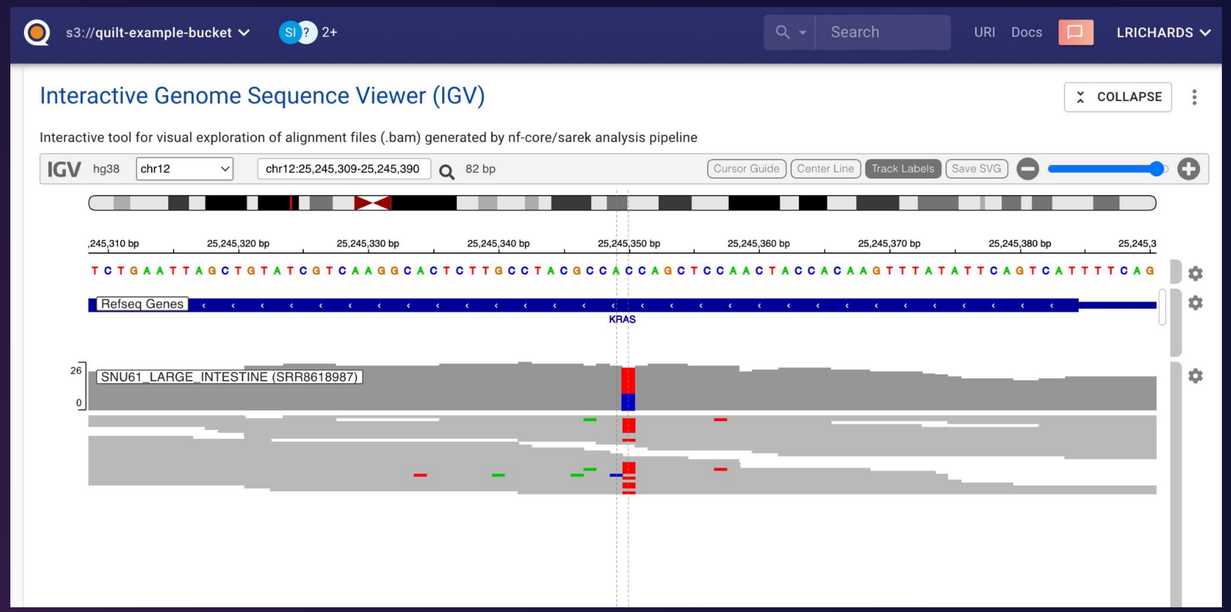
https://quilt-example-bucket.s3.amazonaws.com/wes/result/mafs/combined_VEP.ann.FILTERED.NONSYN.maf.tsv?versionId=sE_jS.sgEDHQ8Qt.aTsRYFURgPIChLrO

Preview pipeline outputs (htmls, bams, vcfs ..) & custom notebooks in the same package
Search, share, browse & plot results without ssh-ing, emailing or installing extra software

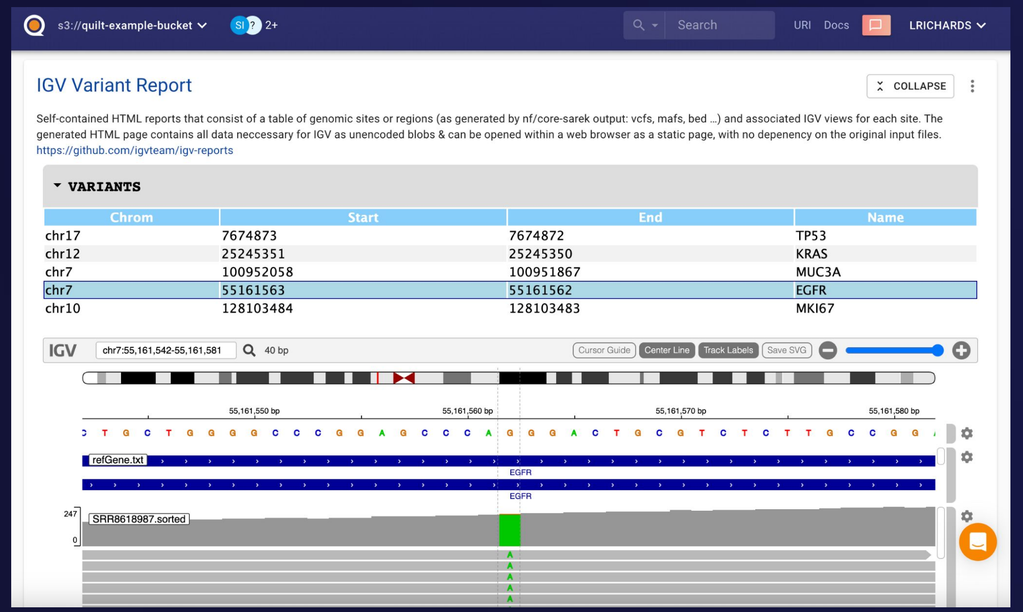
BROWSE GENOMIC DATA DIRECTLY IN QUILT PACKAGES

Interactive visualization of alignments & variants with IGV directly in Quilt package

bam files from nf-core/sarek



variants (vcf, maf …) from nf-core/sarek



No need to remember bam locations, just browse to Quilt Package
Enables quick lookups & easy access across teams

PLACEHOLDER FOR DEMO

- https://demo.quilt.guru/b/quilt-example-bucket/tree/wes/result/

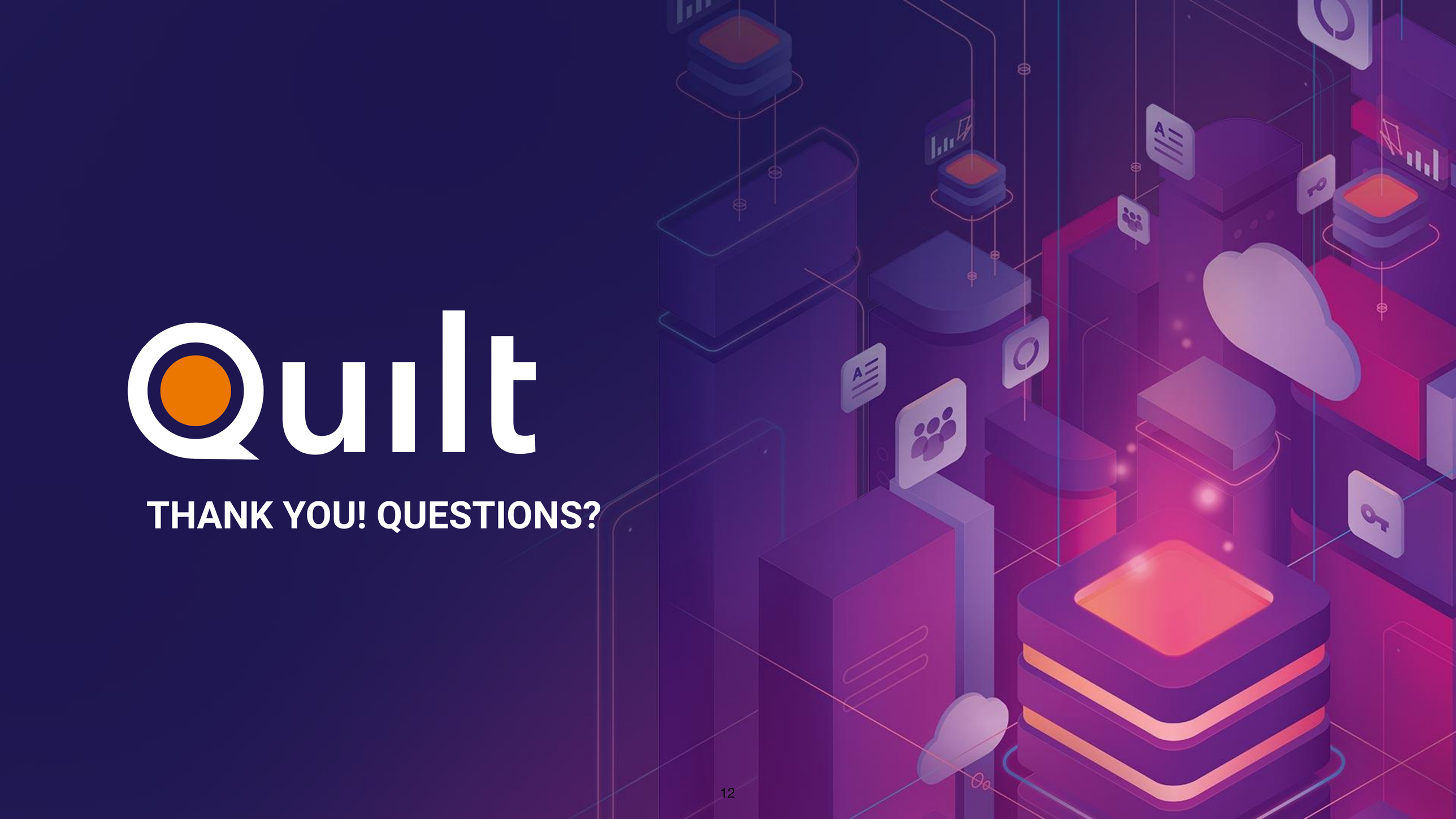
- Highlight…
  - Input data, sarek outputs (bams for example)
  - Jupyter notebook can render
  - README & Sarek log
  - MultiQC front & centre, don’t need to search nested pipeline outs
  - Loaded bams from sarek alignment outs into interactive IGV, search your fav variants, explain how to read IGV
  - Generate a variant report using vcfs or custom files from analysis, where select variants of interest, for example these in oncogenes, Ex missense mutation in EGFR
  - Search KRAS mutations
  - Plot sample metadata
    - Number mutations per cell line (Y Bar, Group By == “Sample ID”, Mutation Count)
    - Number mutations per indication (X/Y Scatter, Tumor Type, Mutation Count)
    - Relationship between fraction genome altered & mutations (XY, Doubling time, fraction genome altered)

QUILT PACKAGES MAXIMIZE UTILITY OF -OMICS DATA

Quilt package highlights from a computational biology perspective….

- Makes “big data” manageable, digestible & usable
- Everything in one place (data, run commands, logs, reports, visualizations …)
- Send shareable links to the exact version of data to colleagues (and find your data!)
- Facilitate quick lookups in notebooks & data files
- Minimize meeting preparation time, no more slides – just browse Quilt Package
- Maximize discovery and interpretation by reducing barrier to access

Quilt Packages for WES data are **Accessible**, **Searchable**, **Versioned**,
Shareable & **Interactive** across all team members, regardless of programming ability



THANK YOU! QUESTIONS?