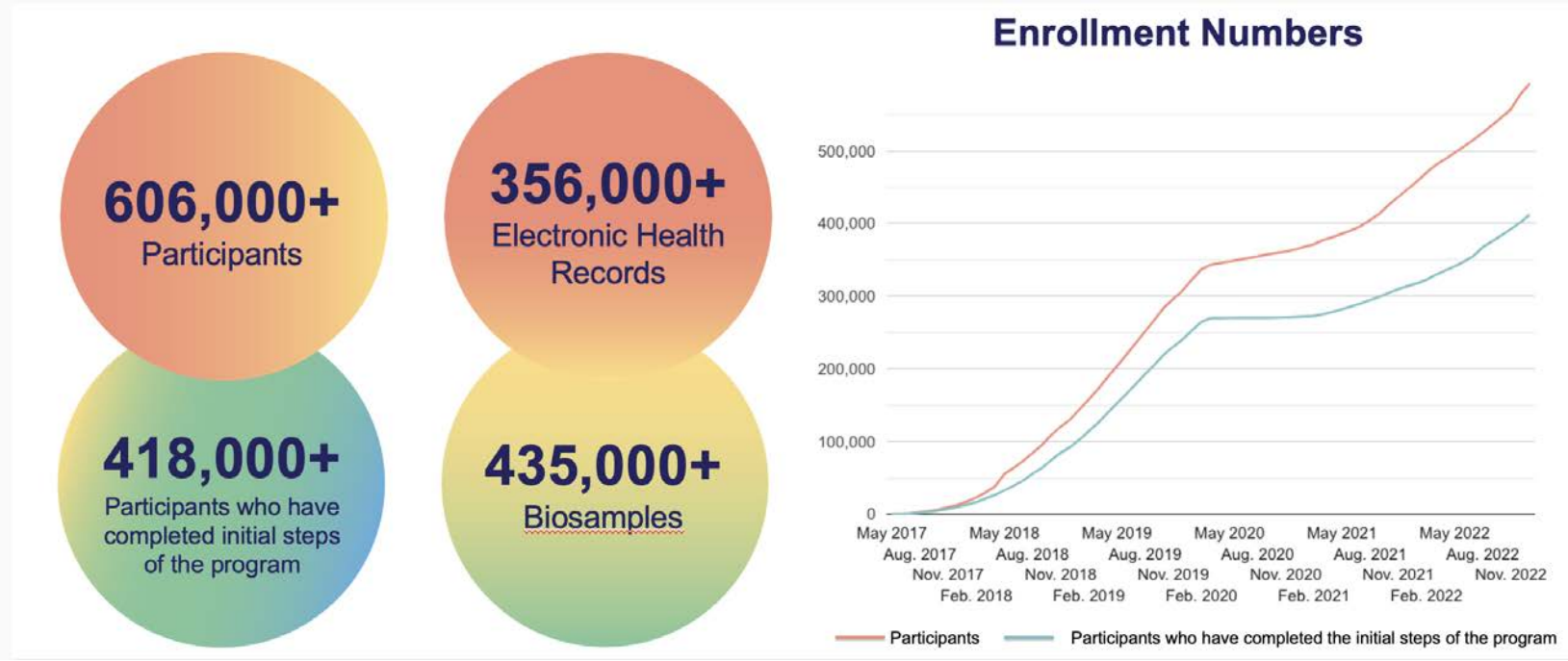
National Institute of Environmental Health Sciences
Your Environment. Your Health.

Cloud Computing for All of Us

Alison Motsinger-Reif, PhD

All of Us Cohort

- US-based cohort, goal to enroll over 1 million participants
  - Genetic data for all participants
- Focused on populations that are underrepresented in biomedical research (UBR)
  - >80% of participants fall into a UBR category



All of Us Cohort

- US-based cohort, goal to enroll over 1 million participants
  - Genetic data for all participants
- Focused on populations that are underrepresented in biomedical research (UBR)
  - >80% of participants fall into a UBR category

Data Type		Participant Counts
Electronic Health Records	Conditions	227,740
	Drug Exposures	214,040
	Labs/Measurements	227,280
	Procedures	221,860
Whole Genome Sequencing dataset		98,560
Physical Measurements		331,300
Fitbit Measurements		12,880
Social Determinants of Health Survey responses		57,620
Lifestyle Survey responses		372,380

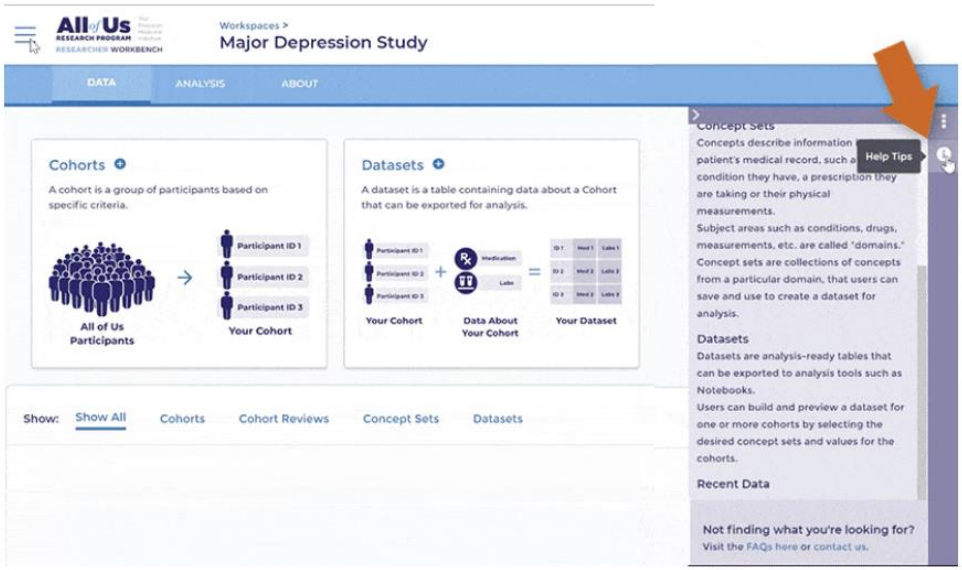
Computing in All of Us

Researcher Workbench

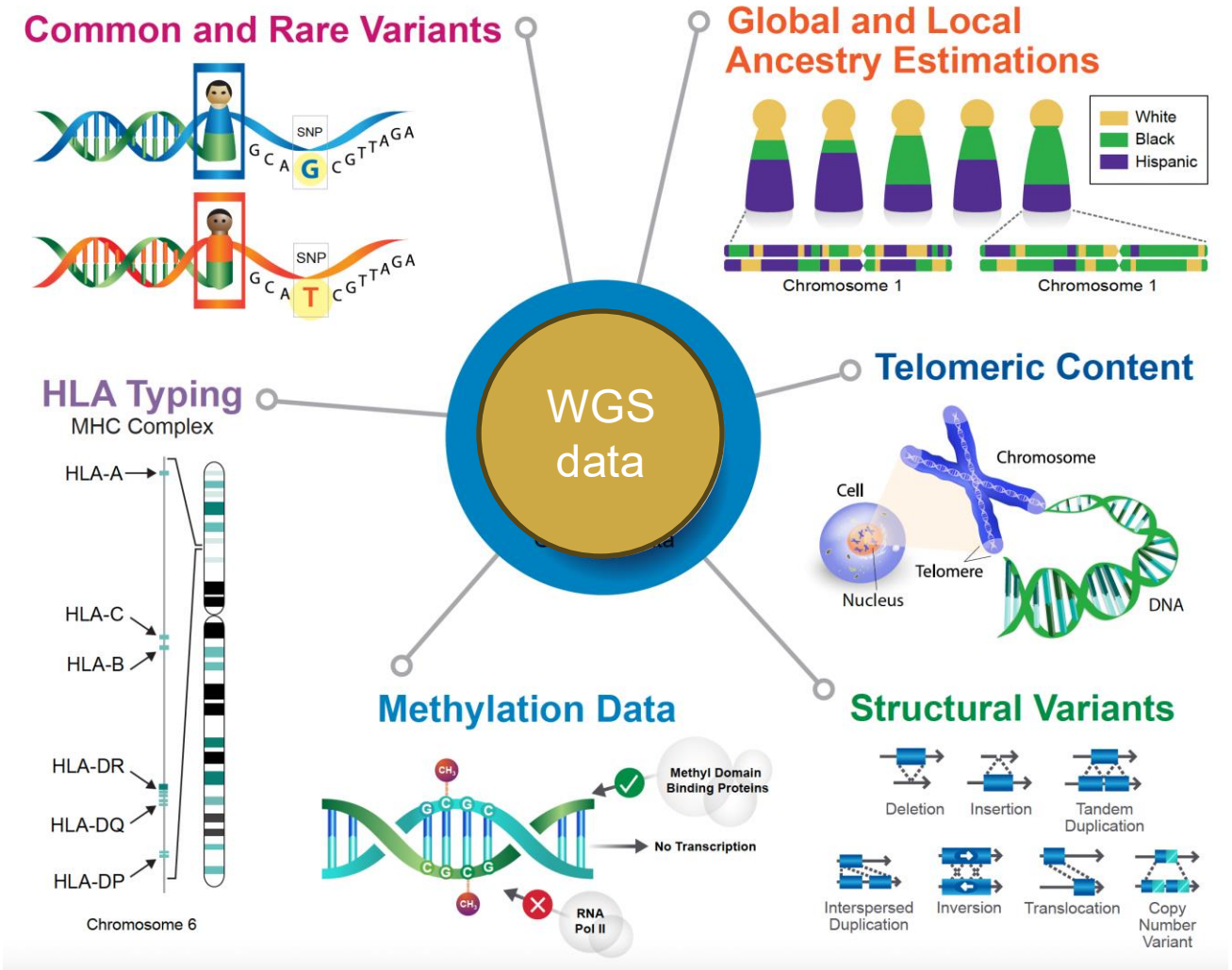
- Cloud-based platform where registered researchers can access Registered and Controlled Tier data
- Cannot download data → bring the compute to the data

Workspaces

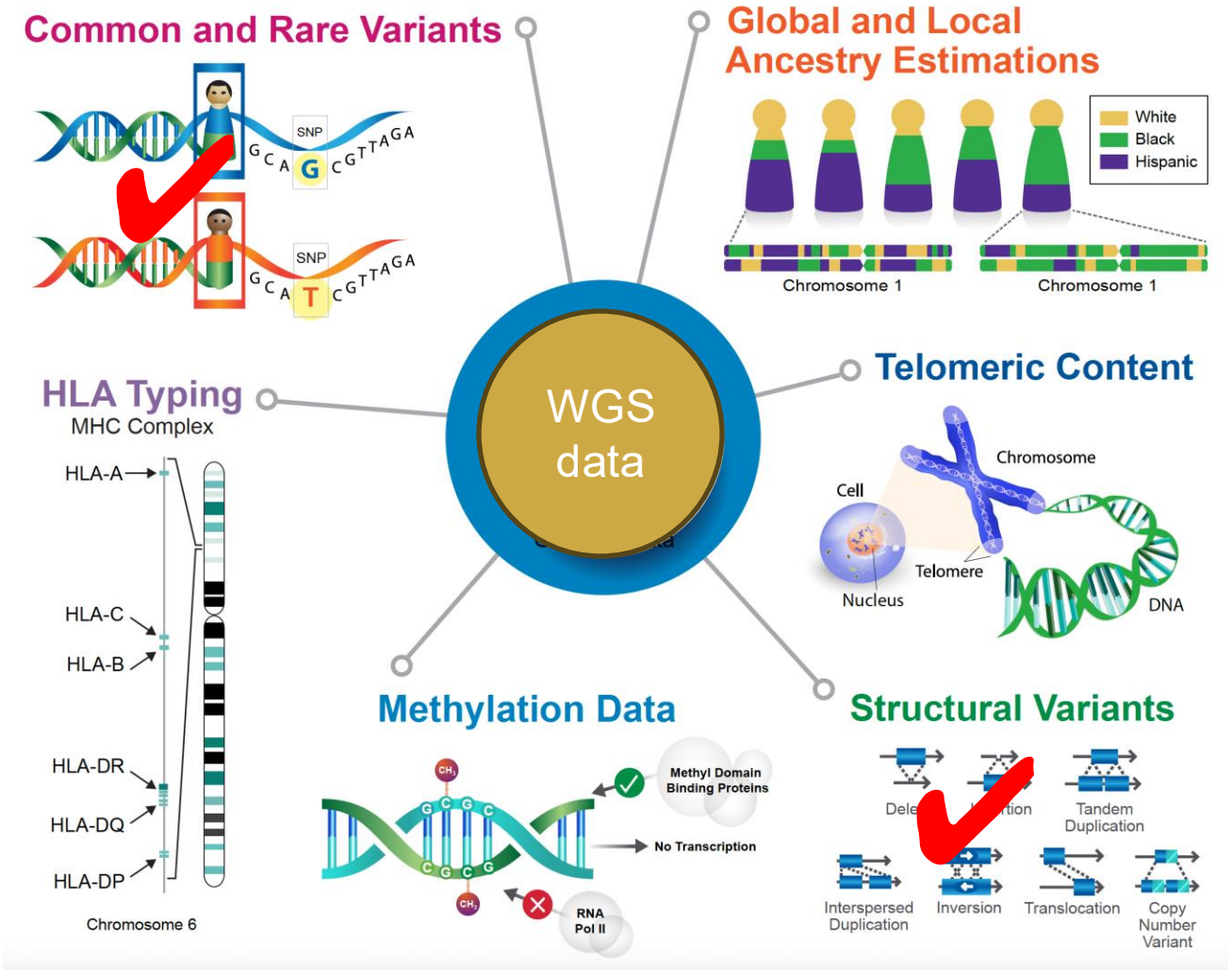
- Registered researchers use workspaces to access, store, and analyze data for specific research projects.



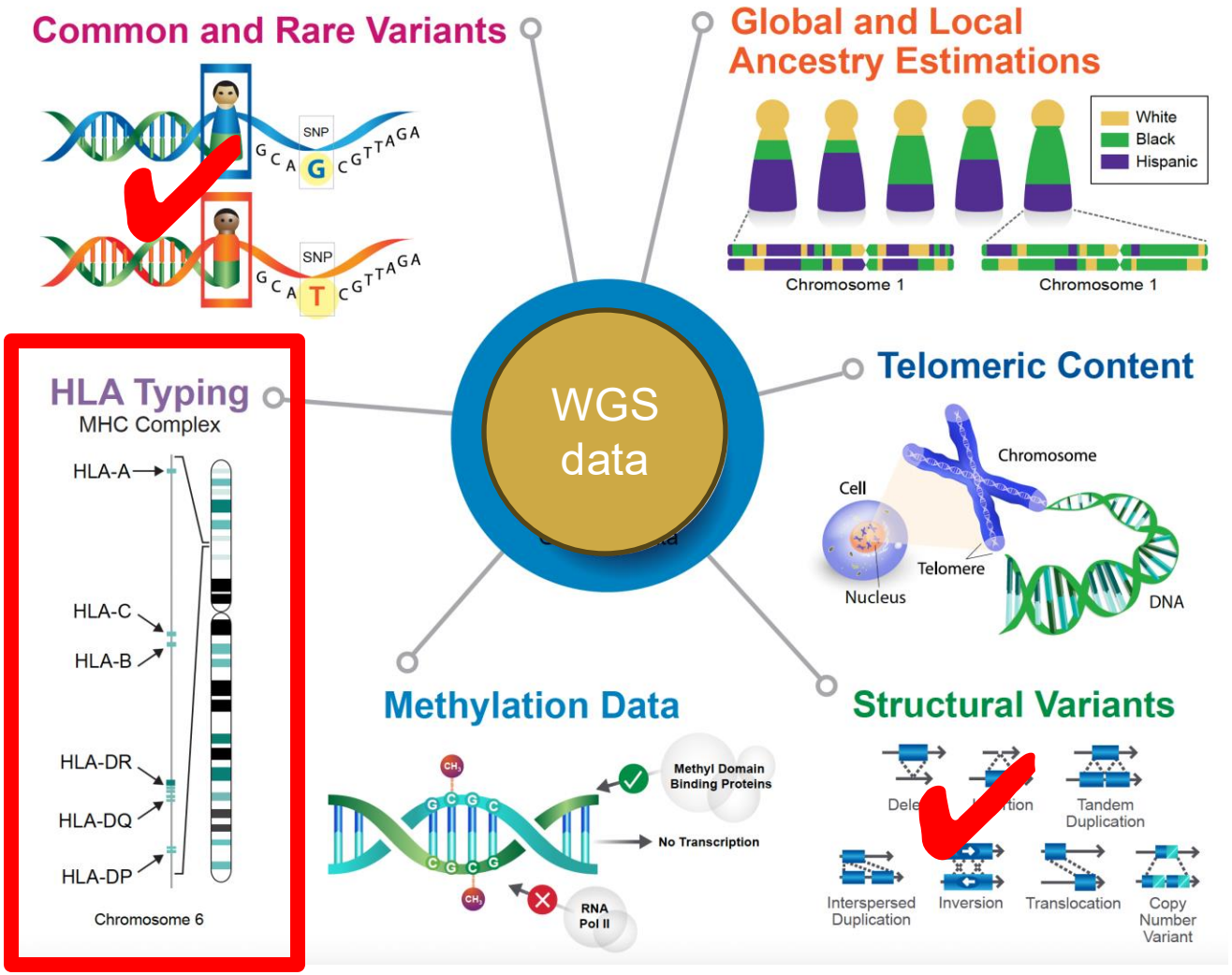
Types of Variants from Whole Genome Sequencing



Types of Variants from Whole Genome Sequencing

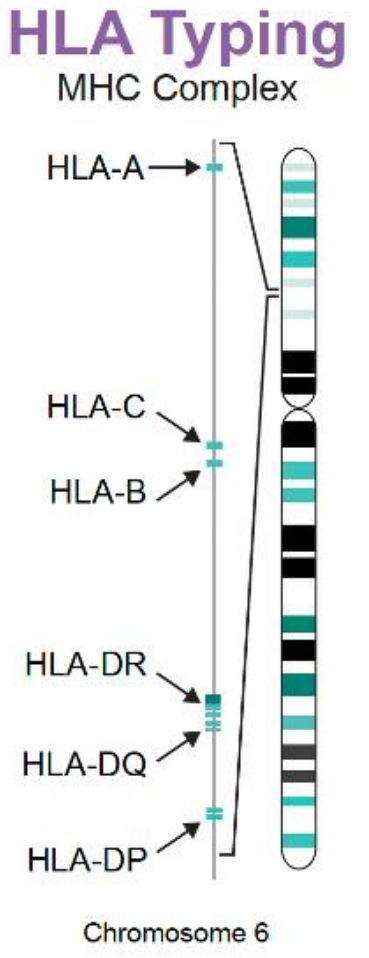


Types of Variants from Whole Genome Sequencing



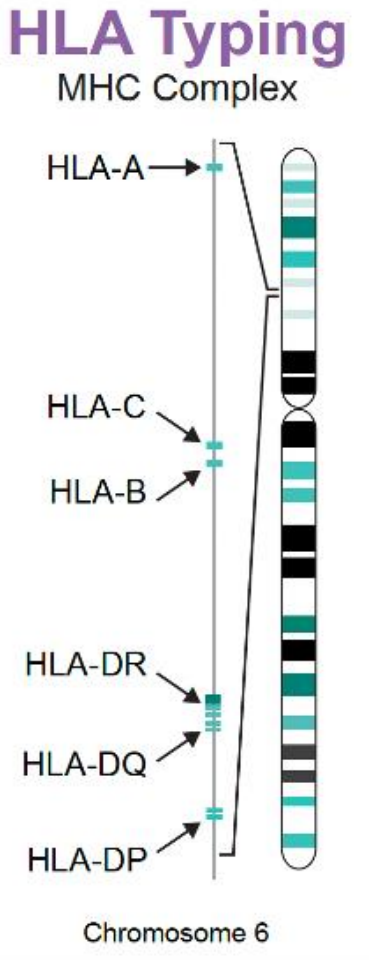
The importance of HLA

- HLA genes are crucial for the immune system.
- They help present foreign substances (antigens) to immune cells.
- HLA matching is vital for successful organ transplantation.
- Certain HLA variants are linked to autoimmune diseases.
- HLA genes affect susceptibility to infectious diseases.
- They influence how individuals respond to specific medications



The importance of HLA

- HLA genes are crucial for the immune system.
- They help present foreign substances (antigens) to immune cells.
- HLA matching is vital for successful organ transplantation.
- Certain HLA variants are linked to autoimmune diseases.
- HLA genes affect susceptibility to infectious diseases.
- They influence how individuals respond to specific medications



Goal: Test for association of HLA genotypes and common, complex diseases

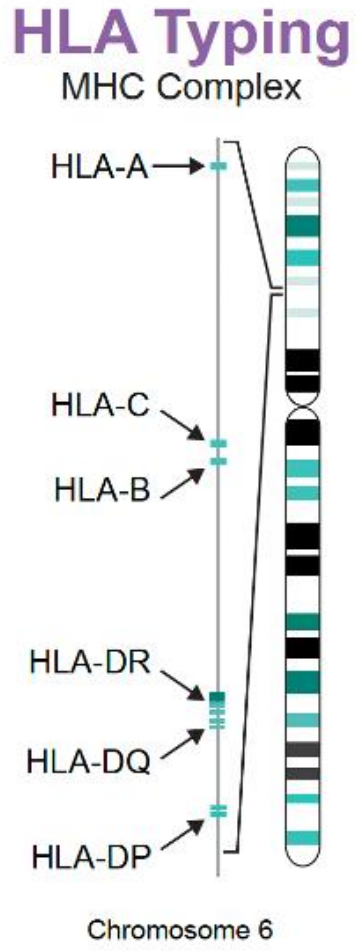
Workflow for Calling HLA variants

Assembly of HLA Alleles: Kourami

- Utilized Kourami for graph-guided assembly
- Generated four- and six-digit HLA alleles using WGS data
- Built reference panel from known HLA alleles in IPD-IMGT/HLA project database
- Assembled peptide-binding domain sequences for HLA genes
- Modified HLA graph for best paths with phasing information
- Filtered HLA alleles for clarity and accuracy

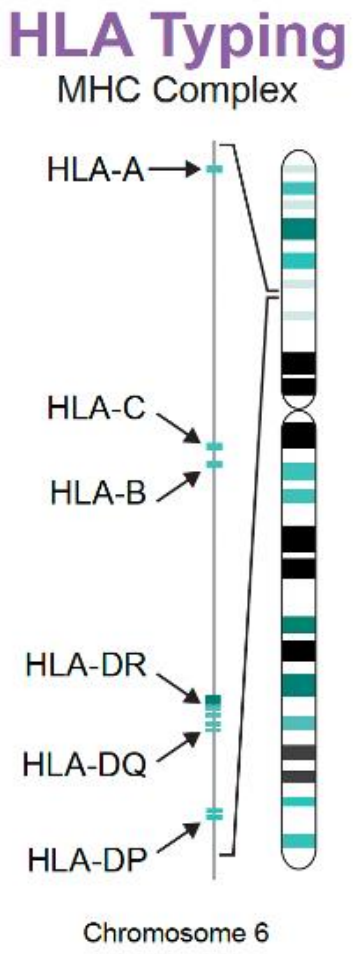
Amino Acid Position Inference: CookHLA and HATK

- Used CookHLA and HATK for HLA imputation
- Inferred amino acid residues
- Employed 1000 Genomes Project reference panels
- Conducted logistic regression for analysis



Variant Calling is Computationally Expensive

- **Data Volume:** Massive amounts of raw sequencing data, often in gigabytes or terabytes
- **Alignment:** Complex alignment of billions of short reads to a reference genome
- **Variant Detection:** Sophisticated algorithms needed for variant identification
- **Genome Complexity:** Highly complex genomes with repetitive regions and structural variants
- **Error Correction:** Correcting sequencing errors adds computational effort
- **Quality Control:** Ensuring data accuracy through extensive quality control steps
- **Population-Scale Studies:** Analyzing large-scale studies increases computational demands



Association Analysis

Linear and Logistic Regression:

- Linear regression used for continuous trait-genetic variant associations
- Logistic regression employed for binary trait-genetic variant associations
- Both methods handle confounding factors and population stratification
- Crucial for uncovering the genetic basis of complex diseases

Association Analysis

Challenges of Logistic Regression:

- Suitable for binary/categorical outcomes in machine learning and statistics
- Computationally complex, challenging for large datasets
- Requires more time and computing resources than linear regression
- Involves exponentials, divisions, and gradient calculations
- Linear regression is typically faster and more computationally efficient

Association Analysis

- Developed a fast regression (FastReg) algorithm
  - an iterative Fisher's scoring matrix that dramatically reduces need for the computing resources
  - For example, for 5,000 individuals and 10 million variants, Plink (the most commonly used GWAS software) requires approximately 12 minutes and 80 workers to conduct logistic GWAS.
  - Our FastReg algorithm can conduct the same analysis in less than half the time using a pre-formed indexable data structure (HDF5)
  - Is flexible for data types and model matrices outside genetic data use cases.

Project Goals

- Perform variant calling for the HLA regions for All of Us participants
  - Testing timing now
  - Goal 10K participants
- Use FastReg to conduct association analyses with HLA variants and ~1600 common complex diseases
- Support and distribute FastReg

Project Team

- Dr. David Fargo, Director of Environmental Science Cyberinfrastructure
- Greg Stamper, Computer Systems Analyst
- Matt Jordan, NIEHS Information Systems Security Officer
- Dr. Adam Burkholder, Computer Systems Analyst
- Dr. John House, Staff Scientist, BCBB
- Dr. Matthew Wheeler, Staff Scientist, BCBB

Questions

Alison.motsinger-reif@nih.gov