Raw Reads to Results: Building Next-Generation Bioinformatics Pipelines

Haider Hassan[1], Jessica Corrado [1], Alissa Cait, Erica Scott, Erik Tillman[2], Dan Ryder[1,3]
[1] Bridge Informatics, Inc., 160 Alewife Brook Pkwy, Cambridge, MA 02138; [2]Akero Therapeutics, 601 Gateway Blvd, Suite 350
South San Francisco, CA 94080; [3]Corresponding author and Chief Executive Officer of Bridge Informatics, Inc.

Variant Detection in Whole-Genome and Whole-Exome Sequencing

What is variant detection in WGS and WES?

- Whole-genome sequencing (**WGS**) analyses an individual's entire genome, while whole-exome sequencing (**WES**) focuses only on protein-coding regions (exons).
- Variant detection in these approaches identifies genetic differences (SNPs, insertions/deletions, structural variants) that may contribute to disease, traits, or biological functions.

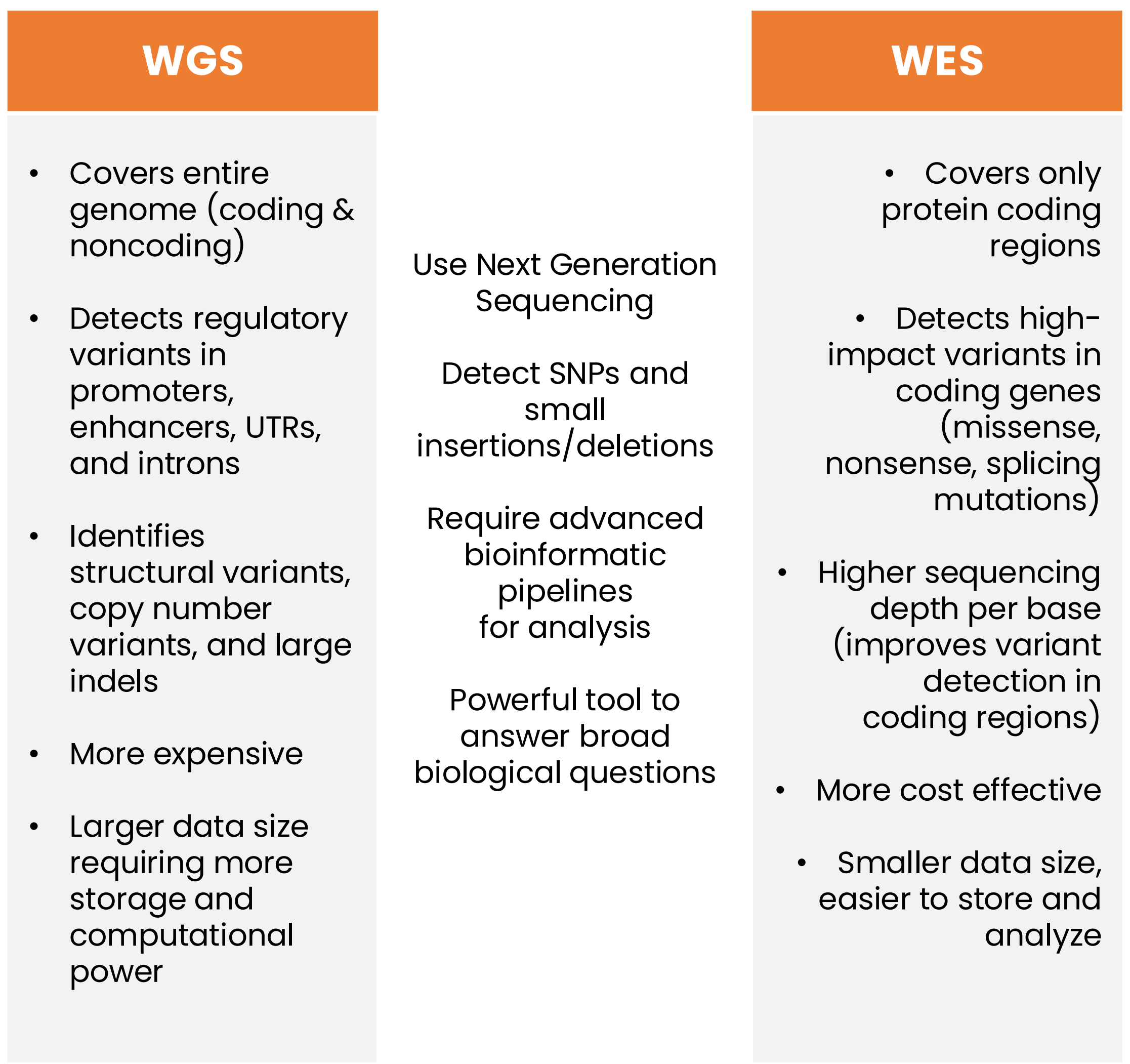

WGS
Covers entire genome (coding & noncoding)
Detects regulatory variants in promoters, enhancers, UTRs, and introns
Identifies structural variants, copy number variants, and large indels
More expensive
Larger data size requiring more storage and computational power
Use Next Generation Sequencing
Detect SNPs and small insertions/deletions
Require advanced bioinformatic pipelines for analysis
Powerful tool to answer broad biological questions
WES
Covers only protein coding regions
Detects high-impact variants in coding genes (missense, nonsense, splicing mutations)
Higher sequencing depth per base (improves variant detection in coding regions)
More cost effective
Smaller data size, easier to store and analyze

Figure 1. Venn diagram comparing variant detection in WGS and WES.

Industry-recognized for being powerful tools, WGS and WES are often used to answer broad biological questions (Figure 2). However, WGS and WES generate vast amounts of genomic data, posing technical challenges in data handling, processing, and analysis. **A well-designed bioinformatics pipeline can address these issues.**

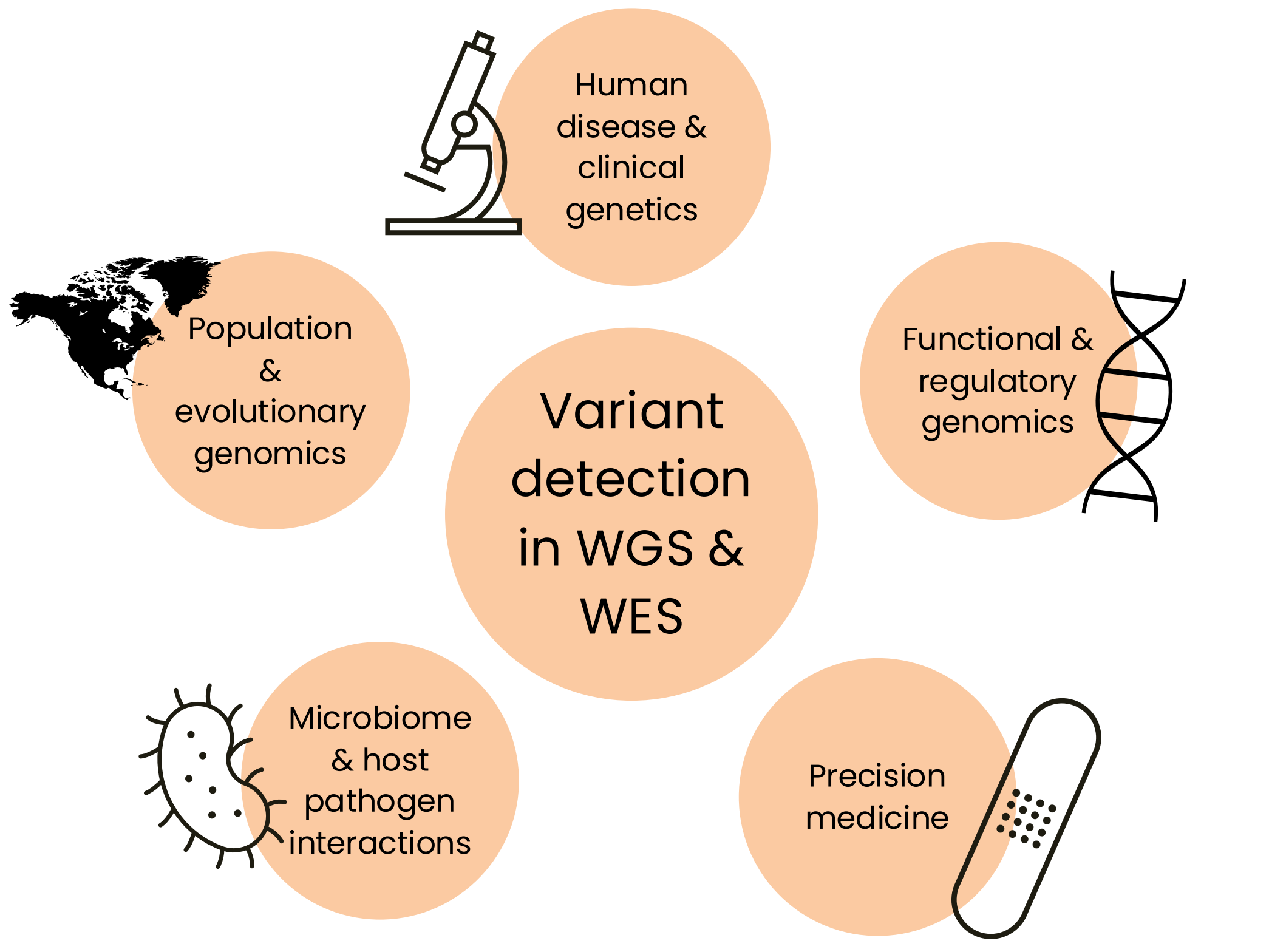
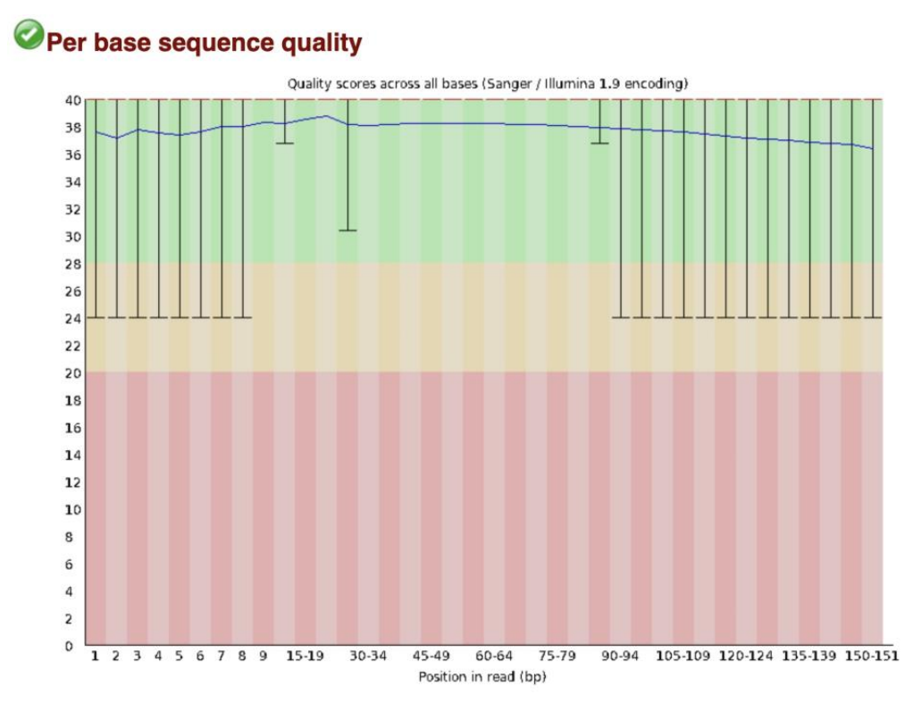

Human disease & clinical genetics
Population & evolutionary genomics
Variant detection in WGS & WES
Functional & regulatory genomics
Microbiome & host pathogen interactions
Precision medicine

Figure 2. Biological insights from whole genome sequencing and whole-exome sequencing. A bioinformatics pipeline designed for variant detection in WES and WGS can be applied to a wide range of biological and clinical research questions.

Bioinformatics Pipeline for Variant Detection

- **Modular & Scalable Design:** Integrates containerization (Docker, Singularity) with workflow managers (Nextflow, Snakemake) for efficient high-throughput sequencing analysis.
- **User-Friendly & Fast:** Enables non-bioinformaticians to process sequencing data to a finalized VCF file in under 10 minutes of user time.
- **Optimized for Performance:** Incorporates automated QC, alignment, variant calling (GATK, GLIMPSE2), and cloud-based GPU acceleration for rapid, large-scale analysis.

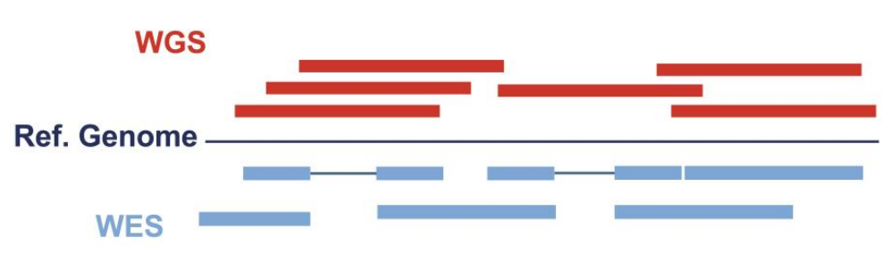
Variant Detection: Computational Workflow



FASTQC file showing confidence in sequencing calling across an Illumina read

STEP 1: PREPROCESSING

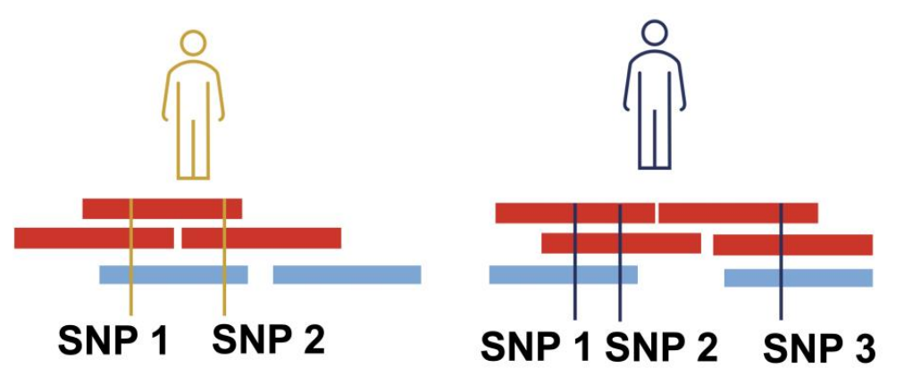
- **Tools:** FastQC, MultiQC.
- **Processes:**
  - Quality control of raw FastQ files.
  - Adapter trimming (Trimmomatic, Cutadapt).
  - Read quality assessment.



Alignment of WGS (red) or WES (blue) reads onto a reference genome

STEP 2: ALIGNMENT

- **Tools:** BWA, STAR, HISAT2.
- **Processes:**
  - Mapping reads to reference genomes (e.g., GRCh38).
  - Sorting and deduplication of BAM files (samtools, Picard).



Variant calling from WES and WGS data for different individuals (or can be applied to different cells)

STEP 3: VARIANT CALLING & ANNOTATION

- **Tools:** GATK, GLIMPSE2, bcftools, snpEff.
- **Processes:**
  - Variant detection (GATK HaplotypeCaller) for high-coverage whole exome data.
  - Genome sequencing imputation and variant detection (GLIMPSE2) for low-coverage whole genome data.
  - Filtering and annotation using the database dbSNP.

STEP 4: WORKFLOW INTEGRATION

- **Workflow Managers:** Nextflow, Snakemake.
- **Processes:**
  - Configuration files for pipeline parameters and resource allocation.
  - Secure storage and transfer of outputs (AWS S3, Google Cloud Storage).

STEP 5: SCALABILITY & VALIDATION

Cloud Integration: Dynamically scalable computing resources for large datasets.

- **Validation:** High concordance with benchmark datasets (GIAB).

Results: Understanding genetic variation using WGS and WES

Our analysis reveals genetic differentiation patterns. The clustering suggests distinct genetic subgroups, which may reflect ancestry, demographic history, or disease-associated genetic differences.

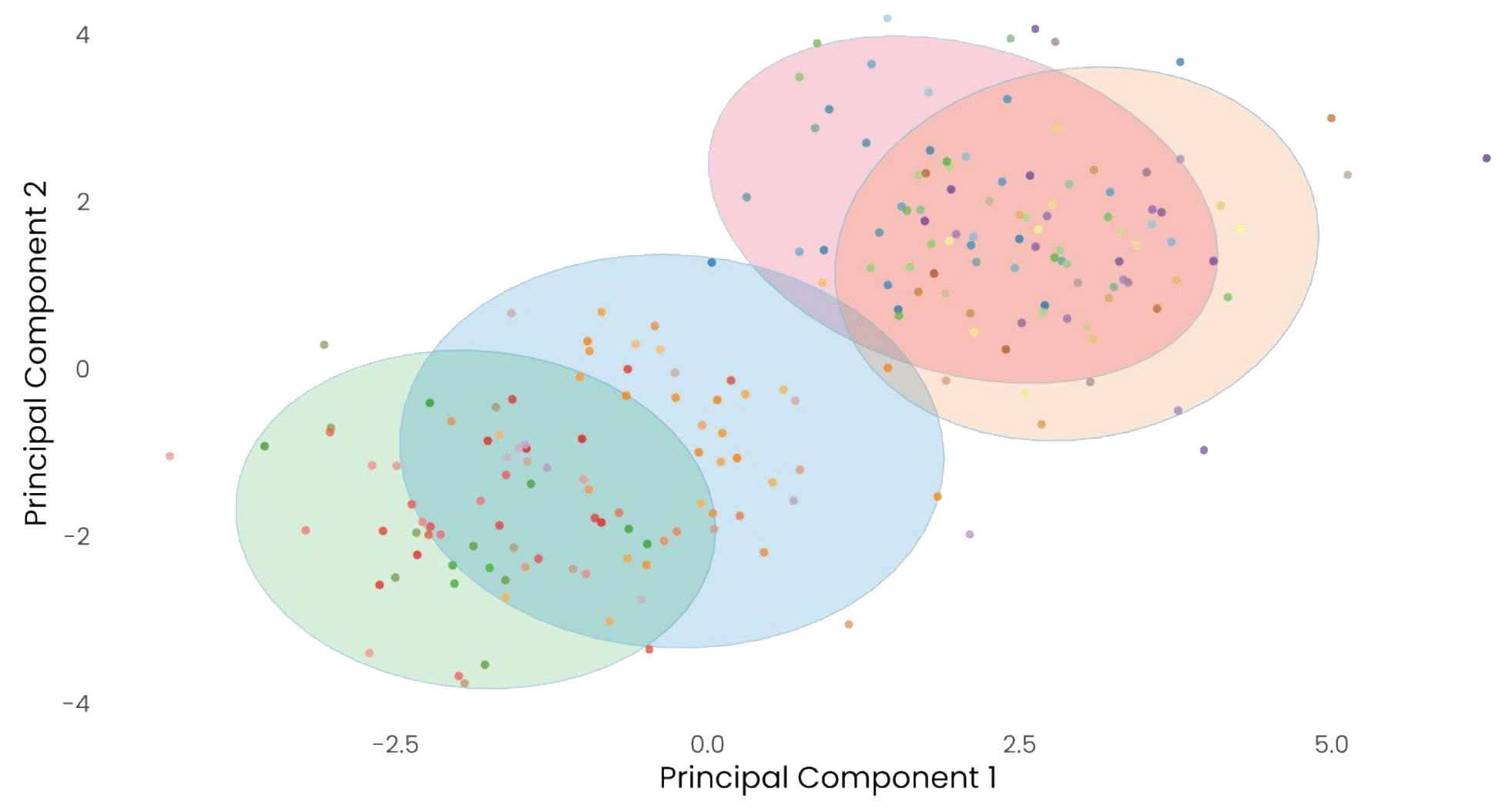


Figure 3. Principal Component Analysis (PCA) of Genetic Variation from Whole-Genome Sequencing (WGS) and Whole-Exome Sequencing (WES). The PCA plot clusters individuals based on genetic similarity, highlighting population structure. Each point represents an individual, with colors distinguishing unique individuals and ellipses representing major population clusters.

Pipeline Improvements Over Manual Process

- **Optimized & Parallel Processing:** Pipelines automate multi-step sequencing analysis (e.g., quality control, alignment, variant calling) and execute tasks in **parallel**, reducing total computation time.
- **Minimal User Intervention:** Automated workflows handle **data processing, software dependencies, and error handling**, significantly cutting down manual effort.
- **Scalable & Efficient:** Cloud-based execution dynamically distributes workloads across multiple processors, ensuring **faster analysis with minimal hands-on time** for researchers.

80% **less effort** **~50%** **overall time saved**

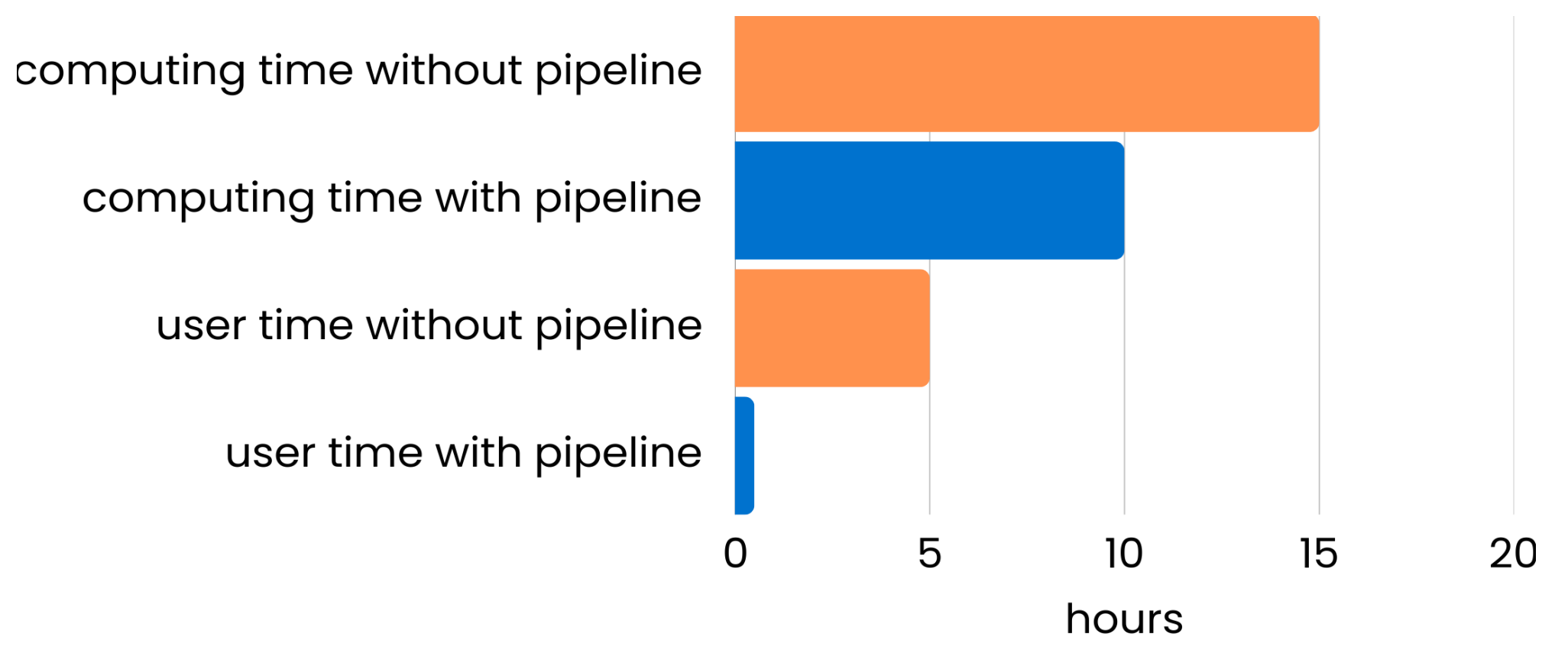


Figure 4. Automated Genomic Analysis: Bioinformatics pipelines reduce computing time by running tasks in parallel and minimizing manual intervention. Cloud-based scalability and containerization further enhance efficiency, allowing researchers to process sequencing data with minimal hands-on time.

	Manual analysis	Using pipeline
Advanced bioinformatics skills	Required	Not required
Version control	none	Highest possible
Reproducibility	Not guaranteed	To the highest standard

Pipeline Summary

Reproducibility: Containerization ensured consistency across environments.

Efficiency: Automated processes reduced time to results by 40%.

Scalability: Cloud resources optimized for large-scale datasets.

Accuracy: Concordance analysis confirmed robust and reliable variant calls.

Let's connect!

linkedin.com/company/bridge-informatics

dan.ryder@bridgeinformatics.com

Acknowledgments

We want to thank Akero Therapeutics for collaborating with us on this project and poster. In particular, we want to thank Erik Tillman, Akero's Director of Translational Biology and Pharmacology, for his collaboration.