PMWC 2017
PRECISION MEDICINE WORLD CONFERENCE

Jan 23-25 SILICON VALLEY

Assessment of variant calling pipelines for clinical diagnosis

Vandhana Krishnan
Stanford Center for Genomics and Personalized Medicine
Stanford Clinical Genomics Service

Stanford University

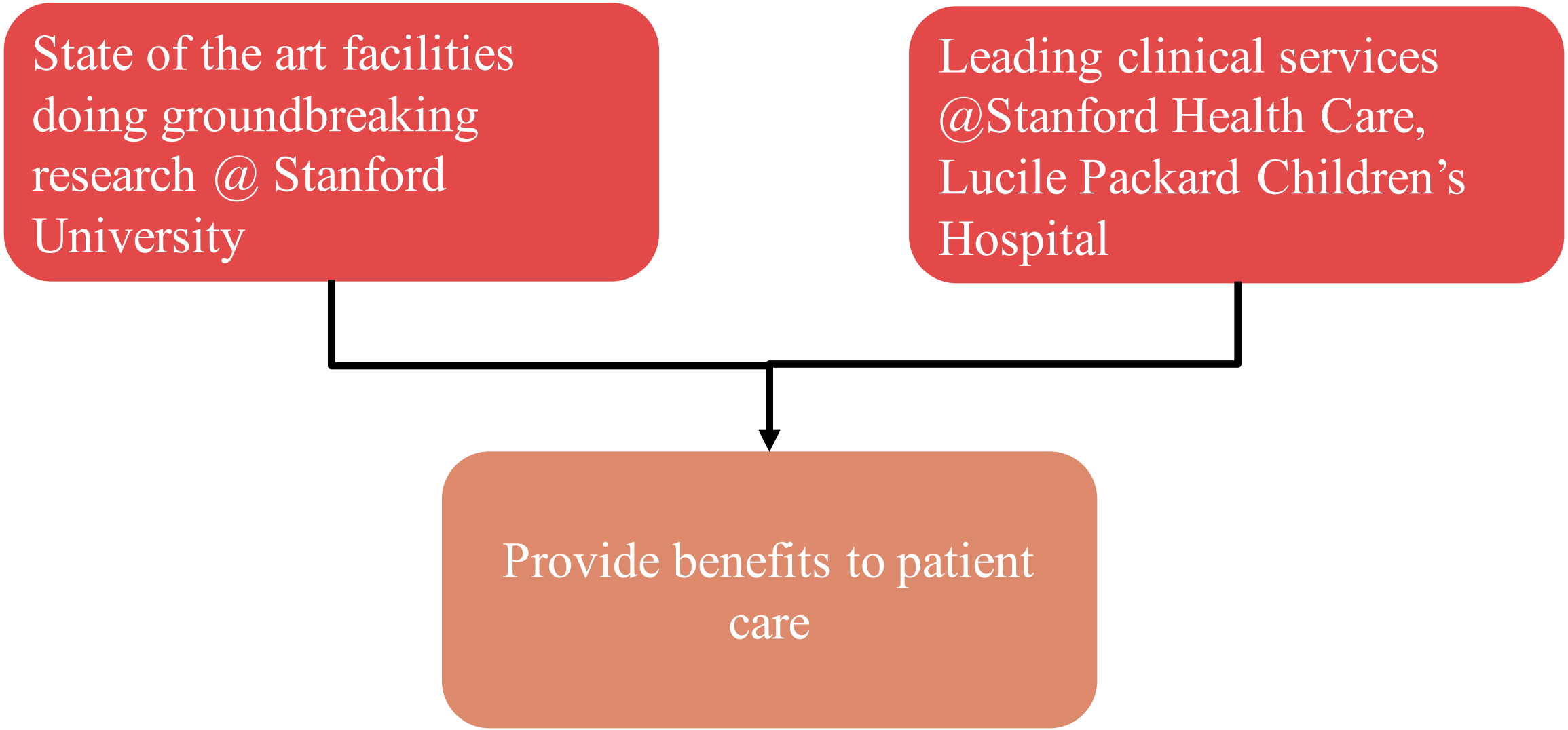
Overview

- Advancements in research towards patient care
- Challenges in translating genomics research
- Clinical Genomics Service pipeline (@Stanford Health Care)
- Accuracy benchmarking

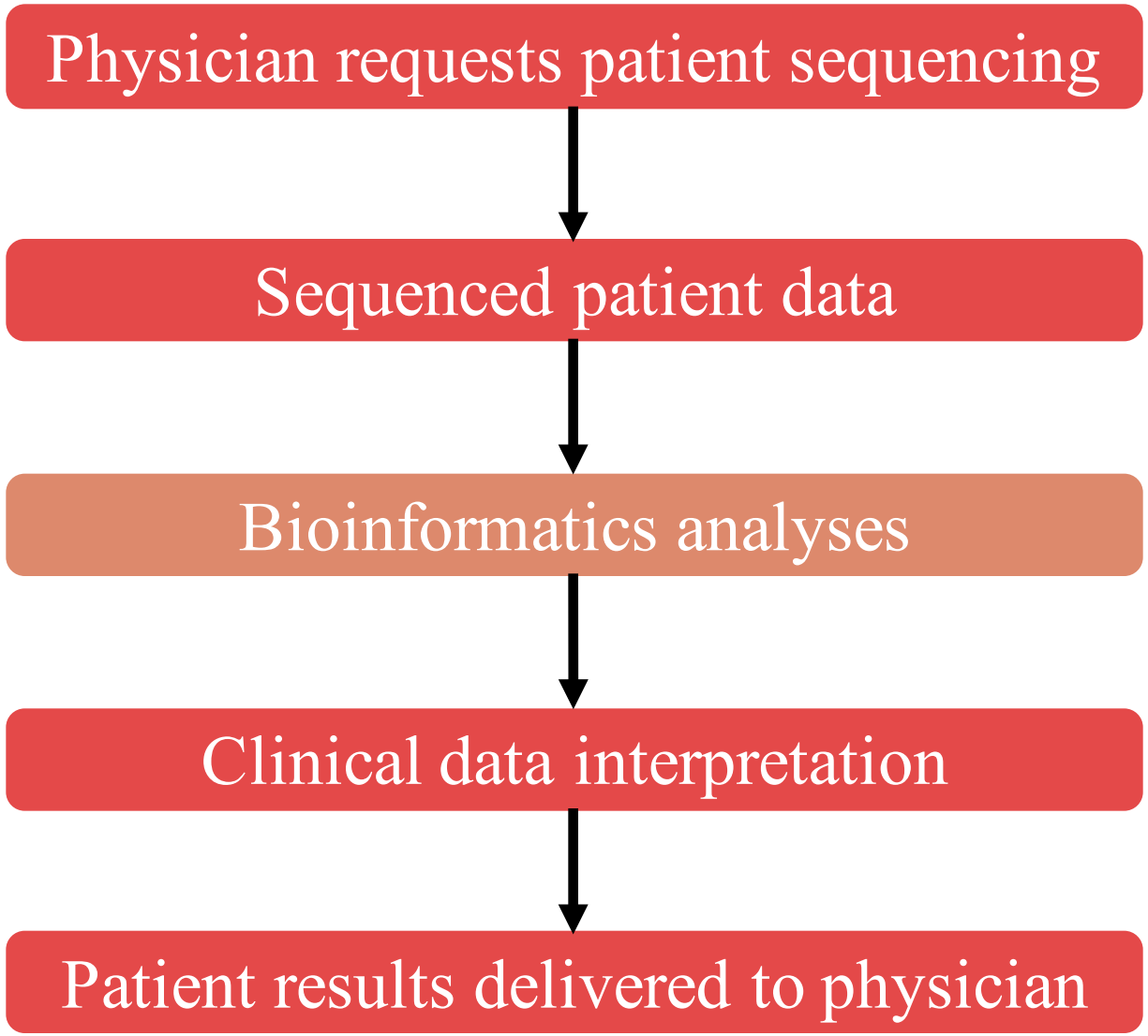
Advancements in research towards patient care

- Rare variants causing disease
- Pharmacogenomics
- Clinical research informatics (health records)
- Computational modelling
- Gene interaction networks

Translating research for patient care



Translating genomics research for patient care



Translating genomics research for patient care: Challenges

Raw data to interpretation

- Several sequencing platforms constantly evolving
  - Long reads versus short reads
  - Use one platform or a combination?
- Identification of variants
  - Choosing among multiple sequence aligners
  - Choosing among multiple variant callers
  - Testing variant calling pipelines (aligner + variant caller combos)

Challenges in translating genomics research

Raw data to interpretation

- Evaluation of variant calling pipelines (existing publications)
  - BWA + GATK-HC
  - Bowtie2 + GATK-UG / Bowtie2 + GATK-HC
  - NovoAlign + GATK-HC
  - Isaac + Isaac
  - BWA + FreeBayes
  - More combinations…

Involves testing with truth sets

Challenges in translating genomics research

Raw data to interpretation

- Performance is different for detection of SNPs and INDELs
  - Different biases on variant calling (ignoring or adding REF allele)
  - Sequencing platform used and sequence coverage
  - Specificity and sensitivity based on truth set calls

Benchmarking is required

Challenges in translating genomics research

Raw data to interpretation

- What is true can be challenging
  - Genome-in-a-Bottle (GIAB) truth sets
- Databases: ExAC, dbSNP, DGV, ClinVar, dbGaP, ...

Truth/Gold standards available - GIAB/NIST

GIAB (NA12878)	Number of bases in the high confidence region (10^9)	# Truth SNPs + Indels in GIAB high confidence bed (10^6)	#Truth SNPs in GIAB high confidence bed (10^6)	# Truth Indels in GIAB high confidence bed
NISTv2.19	~ 2.22	~ 3.15	~ 2.79	365,459
NISTv3.2.2	~ 2.53	~ 3.51	~ 3.15	358,207
NISTv3.3	~ 2.57	~ 3.56	~ 3.19	371,889

Integrated data to generate GIAB truth sets

GIAB (NA12878)	Sequencing platforms used to generate the integrated data set
NISTv2.19	Illumina Gallx, Illumina HiSeq, 454, Complete Genomics, SOLiD, Ion Torrent
NISTv3.2.2	Illumina, SOLiD, Complete Genomics, Ion Torrent
NISTv3.3	Illumina, BioNano Genomics, Nabsys, Complete Genomics, 10X Genomics, Ion Proton, Oxford Nanopore, Pacific Biosciences, SOLiD

GIAB updates (Sep 2016)

- NISTv3.3 - phased calls and phased ID's from GATK-HC
- New way to define callable sets that
  - excludes decoy and certain segmental duplications
  - includes some variant and homozygous reference calls found in >10bp homopolymers
- Includes data from Ashkenazim Jew Trio and Chinese son (Asian Trio)

CLINICAL GENOMICS SERVICE

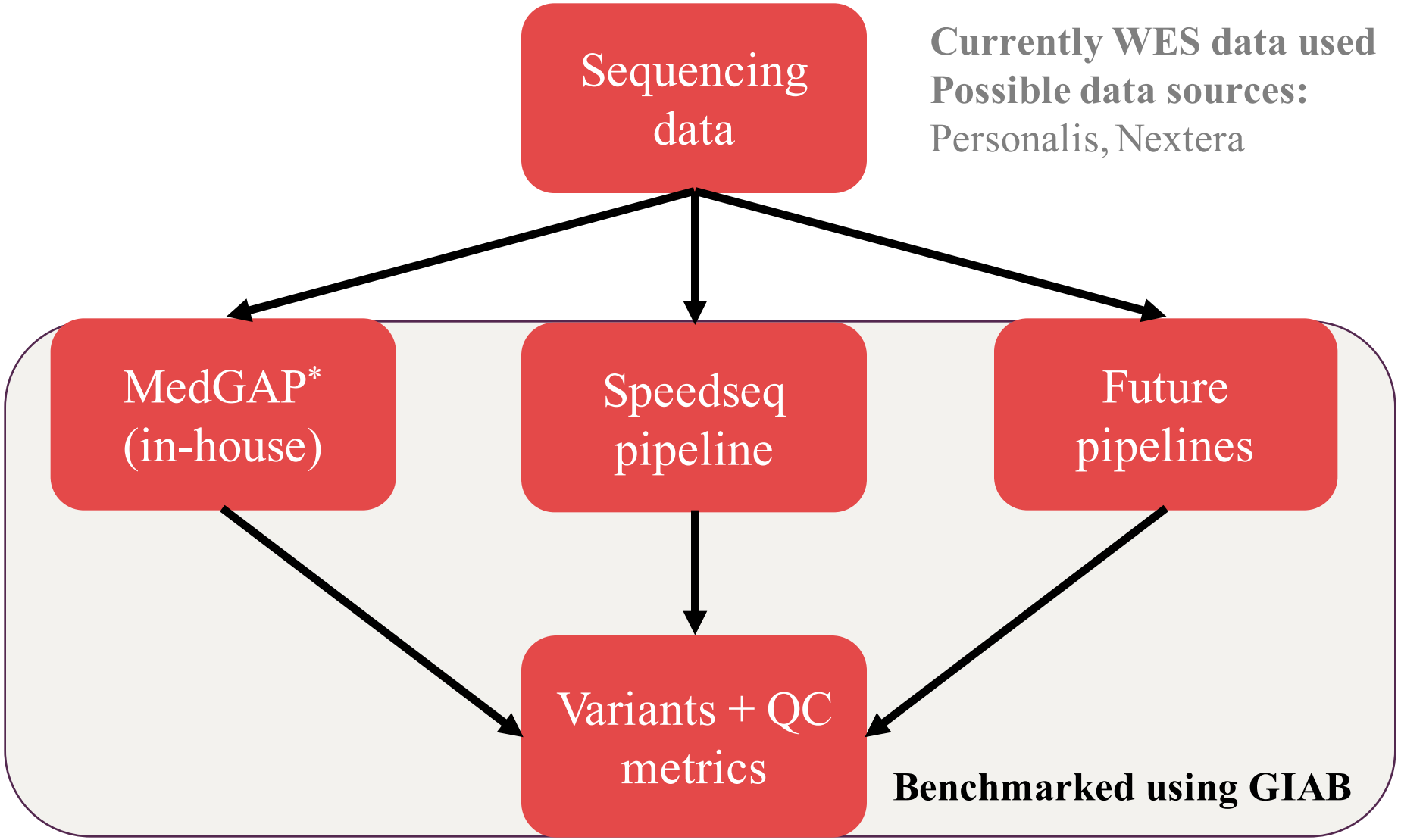
BUILDING A ROBUST PIPELINE FOR CLINICAL DIAGNOSIS

Stanford University

Choosing variant calling pipeline(s)

- MedGAP pipeline (in-house)
  - BWA for alignment
  - GATK-HC for variant calling
- Speedseq pipeline
  - BWA for alignment
  - FreeBayes for variant calling
- Future directions: Novoalign or another aligner combined with variant callers

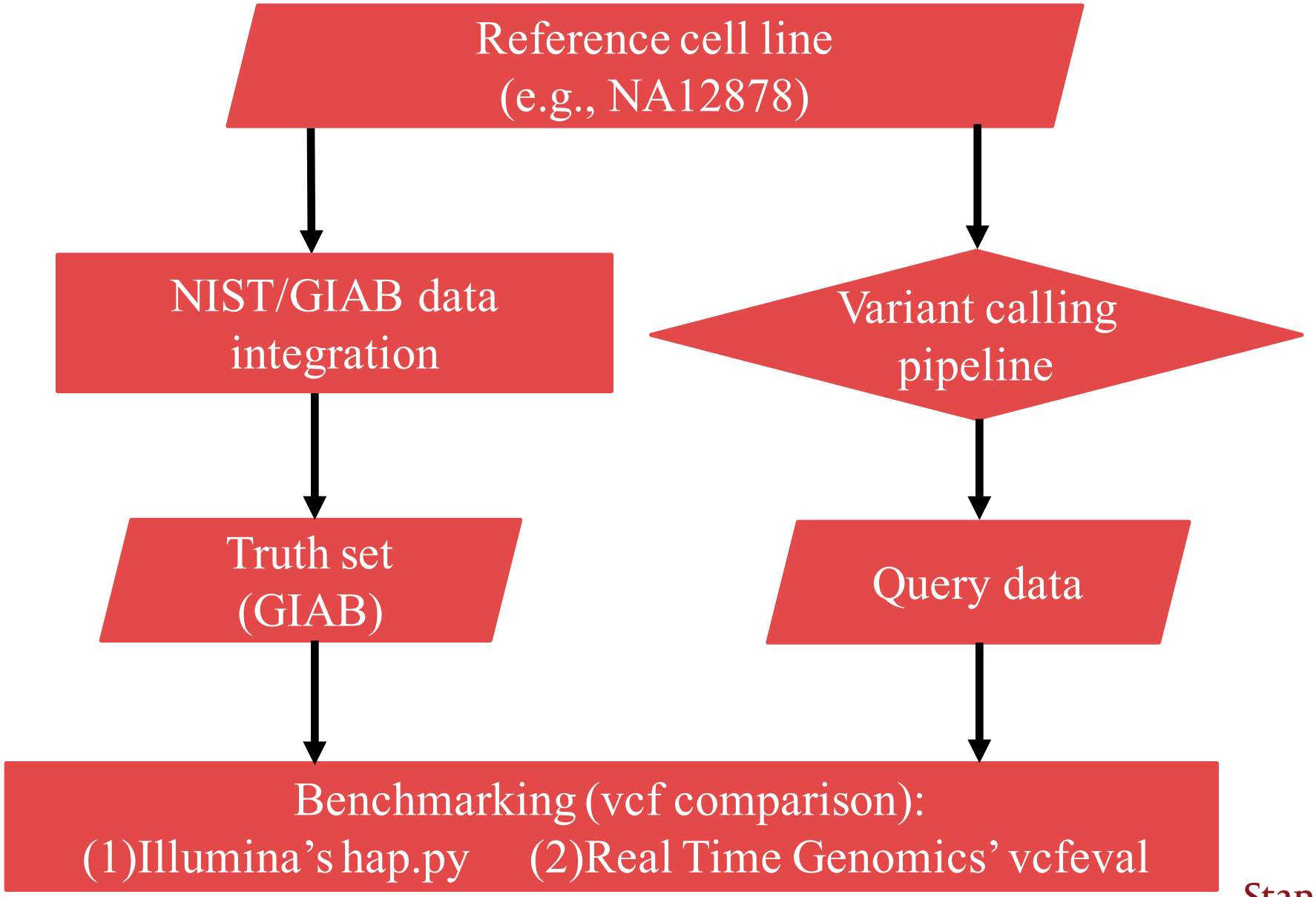
Choosing variant calling pipeline(s)



Currently WES data used
Possible data sources:
Personalis, Nextera

*Medical Genome Analysis Pipeline

Benchmarking variant calling pipeline(s)



What is unique about the pipeline?

Current clinical standards use Coding Exons, @ Stanford Health Care we also use Exons +/- 2 bases to capture rare variants in splice sites

	GIAB (NA12878)	Number of bases in the region of interest (10^6)	# Truth SNPs + Indels in GIAB high confidence bed	#Truth SNPs in GIAB high confidence bed	# Truth Indels in GIAB high confidence bed
Coding Exons	NISTv3.3	~ 30.7	19,409	18,613	436
Exons +/-2 bases	NISTv3.3	~ 69.9	62,938	57,308	5,630

What is unique about the pipeline?

- Best practices in GATK and GA4GH
  - PrecisionFDA standards - rtg and hap.py tools
- Enabling reproducibility, traceability and scalability of the computational pipeline
  - Loom (workflow engine that ensures reproducibility)
  - Docker images for each step in a loom workflow
- Optimizing specificity and sensitivity – assessing against truth set by exploring tools' parameter space

Clinical Genomics Service pipeline: Addressing Challenges

- Upgrading pipeline as versions of software/tools change
  - Variant callers
  - Benchmarking tools

  Docker helps isolate dependencies
- GIAB truth sets keep evolving
  - Frequent updates to exome pipeline (benchmarking)
- Improving specificity and sensitivity
  - Identifying variants in growing clinical databases (e.g.,ExAC) to add to the GIAB truth set

ACCURACY BENCHMARKING

TESTING THE PERFORMANCE OF VARIANT CALLING PIPELINES

Benchmarking results for NA12878 Coding Exons using hap.py

Truth set used: GIAB - NISTv3.3, Query data source: Personalis

	Variant calling pipeline	#Truth total/ #Query total	# True positives (% of truth covered)	#False negatives	#False positives
Indels	MedGAP	428 / 366	346 (80.84%)	84	20
	Speedseq	428 / 389	364 (85.05%)	64	24
SNPs	MedGAP	18,304 / 16,945	16,911 (92.39%)	1,390	34
	Speedseq	18,304 / 18,233	18,058 (98.66%)	246	173

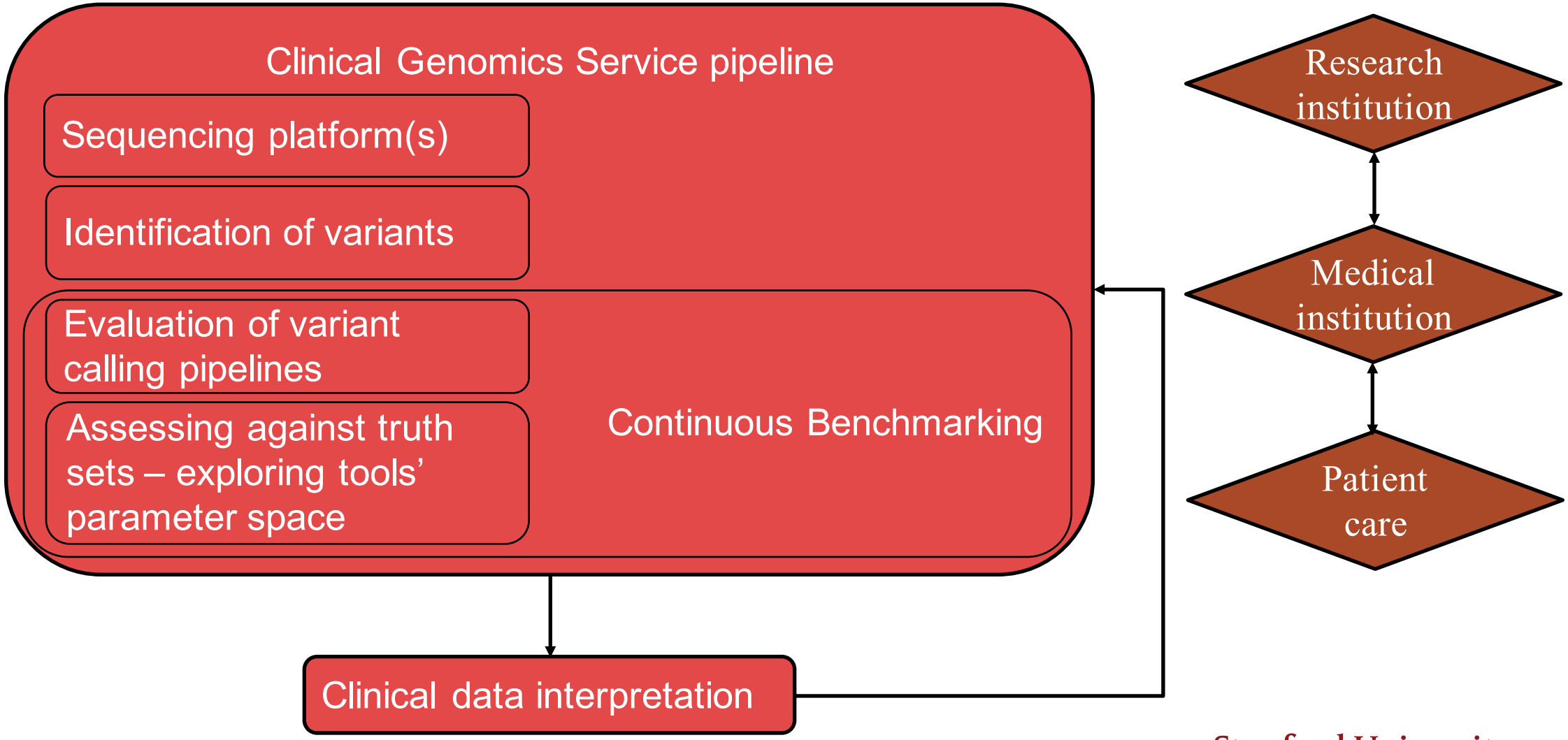
Benchmarking results for AJ Trio son Coding Exons using hap.py

Truth set used: GIAB (NA24385) - NISTv3.3, Query data source: Personalis

	Variant calling pipeline	#Truth total/ #Query total	# True positives (% of truth covered)	#False negatives	#False positives
Indels	MedGAP	419 / 320	311 (74.22%)	111	9
	Speedseq	419 / 385	377 (89.98%)	42	8
SNPs	MedGAP	17,667 / 16,198	16,180 (91.58%)	1,483	18
	Speedseq	17,667 / 17,625	17,505 (99.08%)	162	120

SUMMARY

Enabling Precision Medicine: Translational research



Acknowledgements

Stanford Center for Genomics & Personalized Medicine:

Mike Snyder (Director), Somalee Datta (Director of Bioinformatics), Isaac Liao (Software Engineer, Loom), Amin Zia (Sr. Bioinformatics Scientist)

Clinical Genomics Service @ Stanford Health Care:

Euan Ashley & Jason Merker (Directors), Sowmi Utiramerur (Director of Bioinformatics), Nathan Hammond (Sr. Scientist, Loom), Lalitha Viswanathan (Sr. Pipeline Engineer, Workflows)

Thank you!