

From: [Mark Pratt](#)
To: [John West](#)
Subject: accuracy strategy notes
Date: Tuesday, May 15, 2012 9:29:07 AM
Attachments: [AccuracyStrategyNotes110514.pptx](#)

The first 5-6 slides are current. The later slides are from a deck I had been building in Feb/March.

Sorry about the state but I couldn't edit them down last night.

^^

Personalis Accuracy Program Facets

1. Ground Truth / Golden Genomes program

- Concordance and pedigree analysis over multiple platforms to independently assess errors
- Develop “gold standard” genomes using any accessible means to resolve conflicts and extend coverage
 - Alternative and complimentary assays
 - Targeted laboratory and bioinformatics analysis to resolve high impact errors
- Prepare standard output for test pipeline comparison

2. Genome comparison tool

- Basic tool for binary low-level genome-wide comparison of nearly identical genomic analysis
- Tabulate and categorize differences and problems
- Used for pipeline development, “golden” genome comparison and other development accuracy activities

3. Problem Flagging and Error model

- Identify errors using golden genomes, array discordance, etc
- Annotate comparison of sample and historical high density QC metrics
- Annotate for known and suspected “problem” areas
- Train algorithms classify and flagging errors (and later to resolve discordant data)

4. Complimentary assays and analytical methods for production use

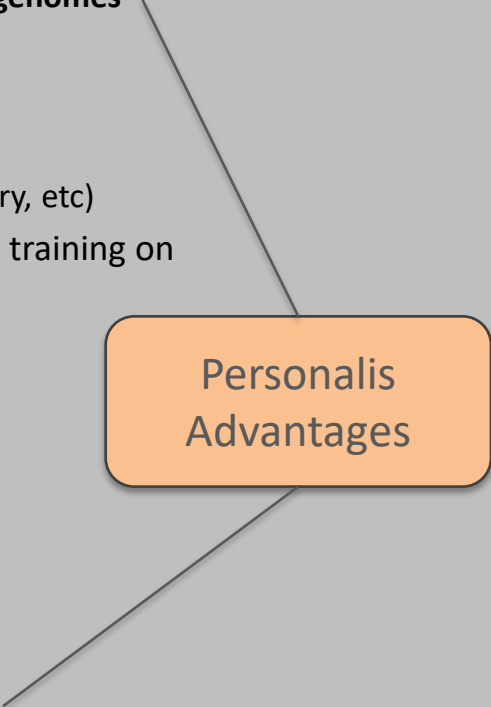
- Develop customized libraries and protocols to improve sensitivity and accuracy on content:
 - Custom content GT/CNV/Breakpoint array
 - Targeted pullout and tailored insert sequencing libraries
 - Exome, RNAseq

5. Develop Processes for Targeted Solutions

- Research pipeline for resolution of specific problems, e.g. a known SD in a high value content area
 - Long range sequencing techniques (FISH, single molecule methods)
 - Customized realignment or reassembly
 - Customized local reference

Error Analysis Using Golden Genomes

- Start with validated modern (current) genome analysis tools
 - Harnesses world-wide genomics community (e.g. GATK variant discovery framework)
 - Carefully engineered, particularly on SNP calling and Q scores
 - Rapidly evolving ...
- **Use multiple genomes and assays to produce a set of high quality “Golden” genomes**
 - Use replicates and pedigrees to develop discordance and error data sets
 - Database errors and high error rate regions
 - Compute detailed QC metrics
 - Database high density QC statistics (coverage, Q, FWD/REV asymmetry, etc)
- Develop genome-scale error model and annotation by segmenting with and training on
 - Known errors and error ranges from golden genome set analysis
 - Local and aggregate QC metrics
 - Database of known or suspected regions of difficulty
 - SDs, STRs, pseudogenes, palindromes
 - Alignment degeneracy metrics
 - GC content, problem motifs
- Use segmentation and error model to annotate and correct confidence
 - Machine learning approach
- **Prioritize analysis and other development activities by value of content**



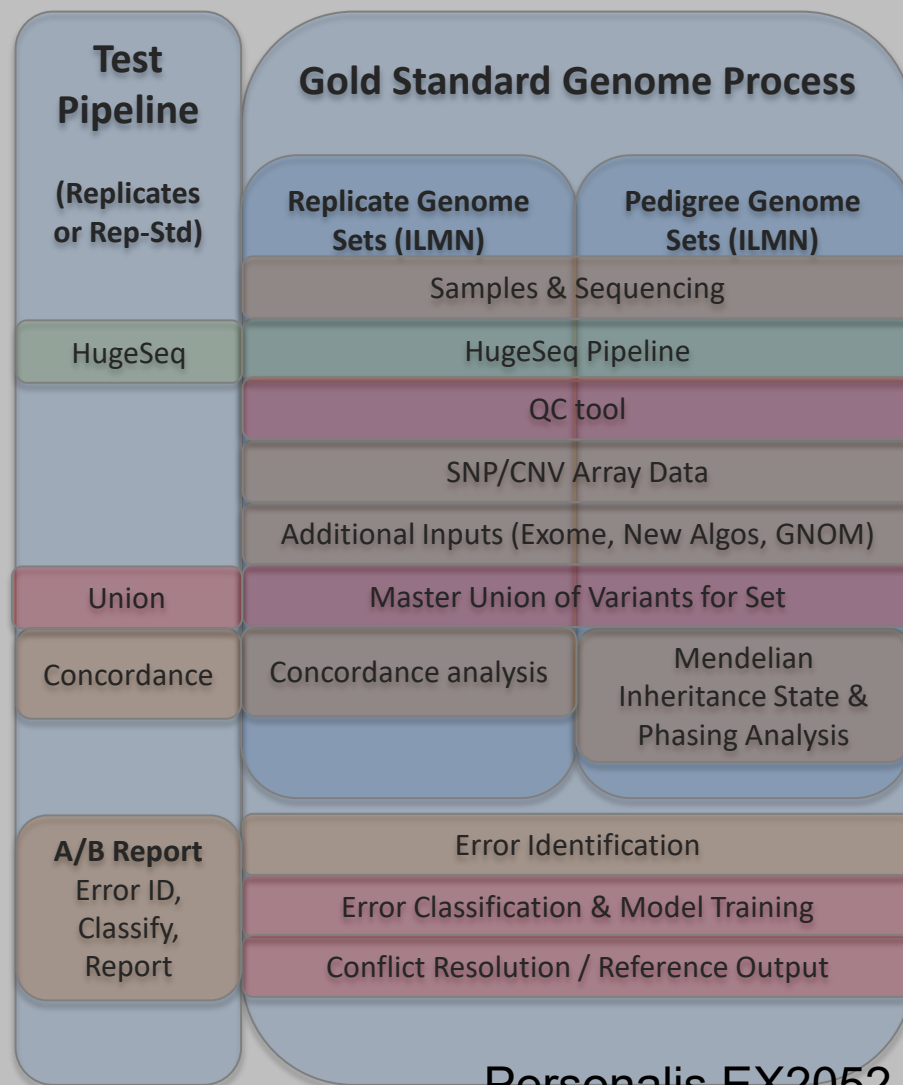
Personalis
Advantages

Process for 2-Genome Comparison

Development Test Pipeline & Golden Genome Comparison

A mechanism to comprehensively compare pipeline outputs is essential to accuracy improvements as well as maintenance and rapid development of pipeline components. Top level requirements:

- Exhaustive comparison of locus-level variant and reference calls between two genome analyses.
- Segment discordance by Personalis content and known problem areas.
- Compare quality indicators such as coverage and variant qualities.
- Produce high level concordance scoring summary.
- Compare pipeline replicates or pipeline v. golden genome.



A/B Genome Comparison Output

(Comparison of ostensibly identical analyses)

High Level

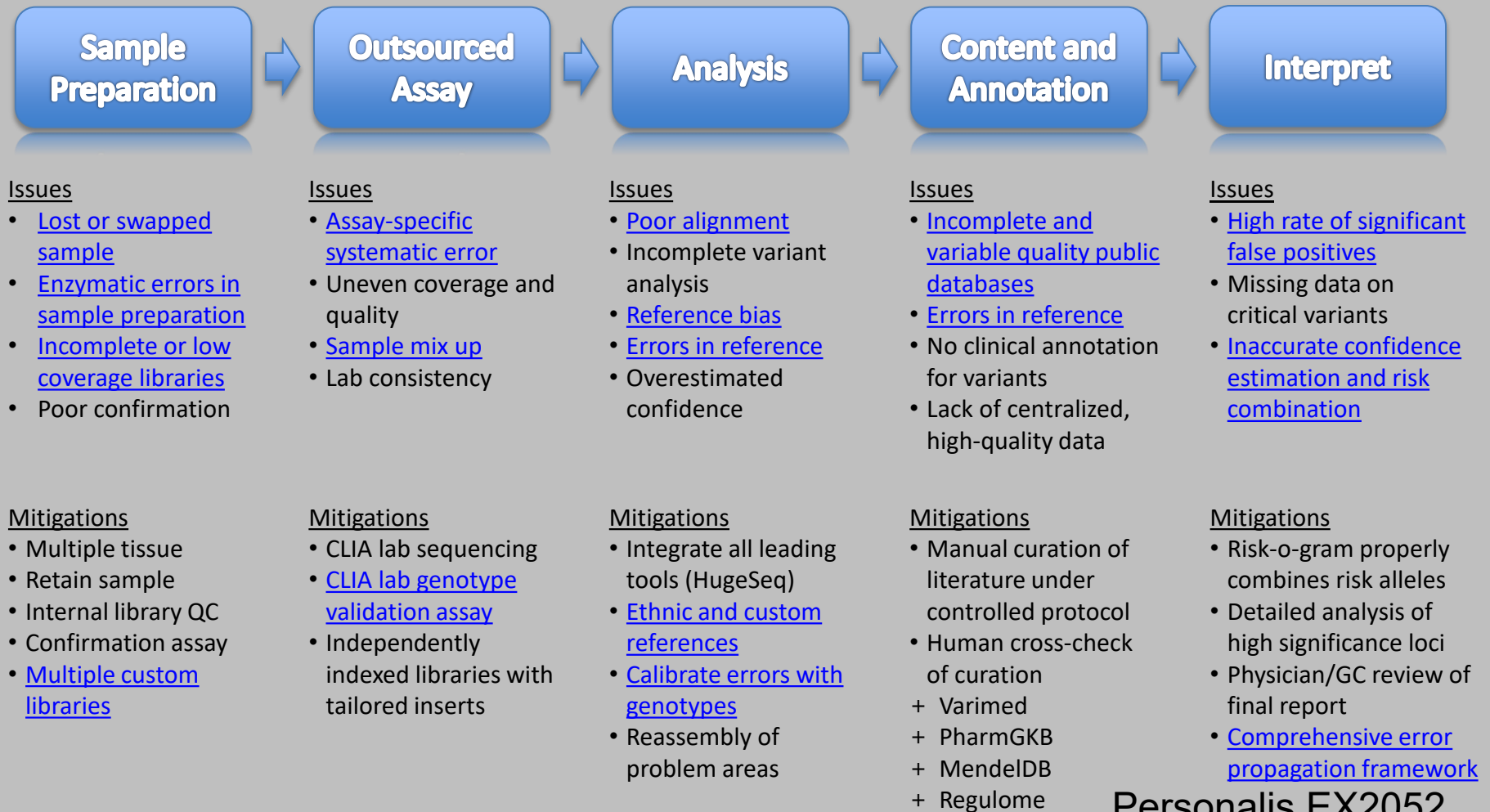
- $\Delta\%$ genome coverage
- $\Delta\#$ and $\Delta\%$ reads mapped
- $\Delta\#$ and $\Delta\%$ of variants by type
- Comparison of coverage distributions
- Comparison of Q distributions
 - Δs at specified thresholds e.g. {10,20,30}
- Comparison of Q by read position
- Comparison of InDel/SV sizes

Low Level

- Discordant SNPs & Indels
 - Segment by content regions
 - Segment by genomic regions
 - Segment by sequence context
 - Segment by QC parameters
- Discordant CNV/SVs
 - ID common variants
 - Compare parameters
 - Segment by sequence context
 - Segment by genomic regions
 - Segment by QC parameters

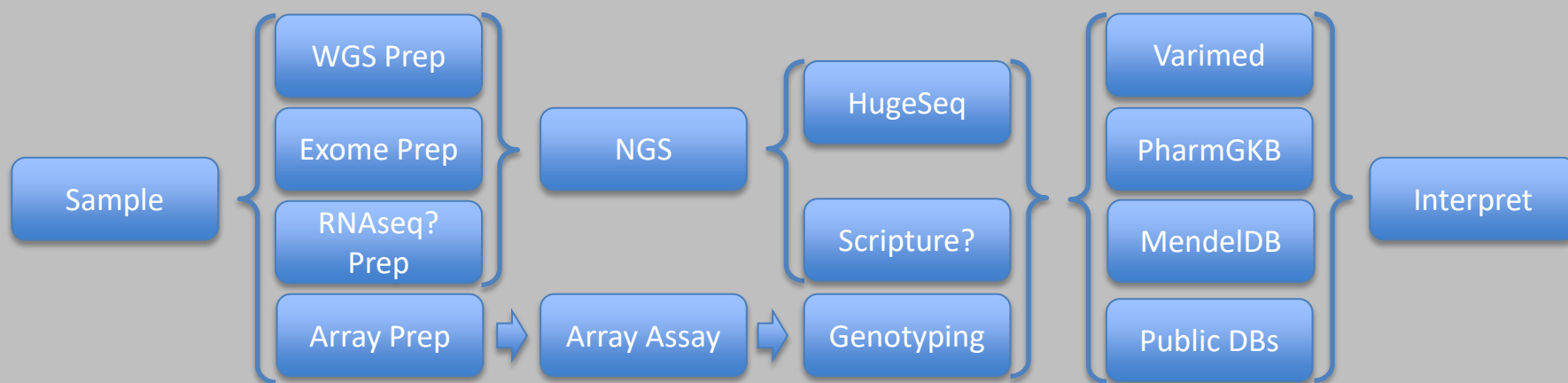
Personalis Analytical Pipeline

Accuracy Focal Points



Personalis Pipeline

Accuracy Program Activities



Enzymology (Completeness, Error floor, Platform combination)

Analysis & Tools (Accuracy/sensitivity assessment & model, QC, Annotation)

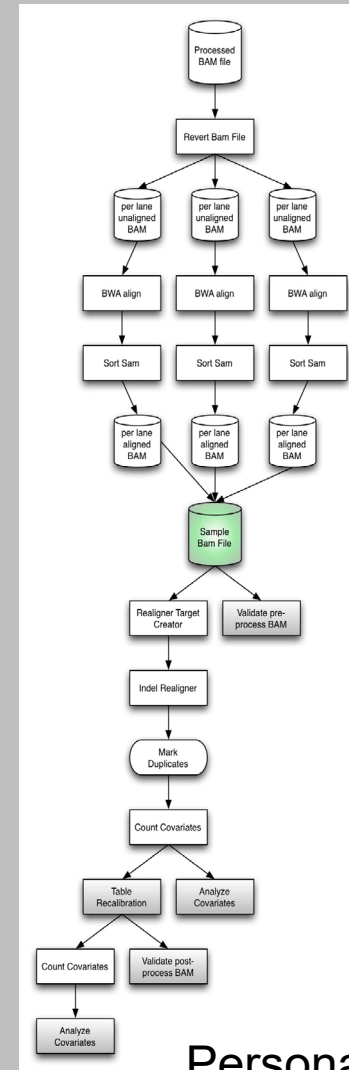
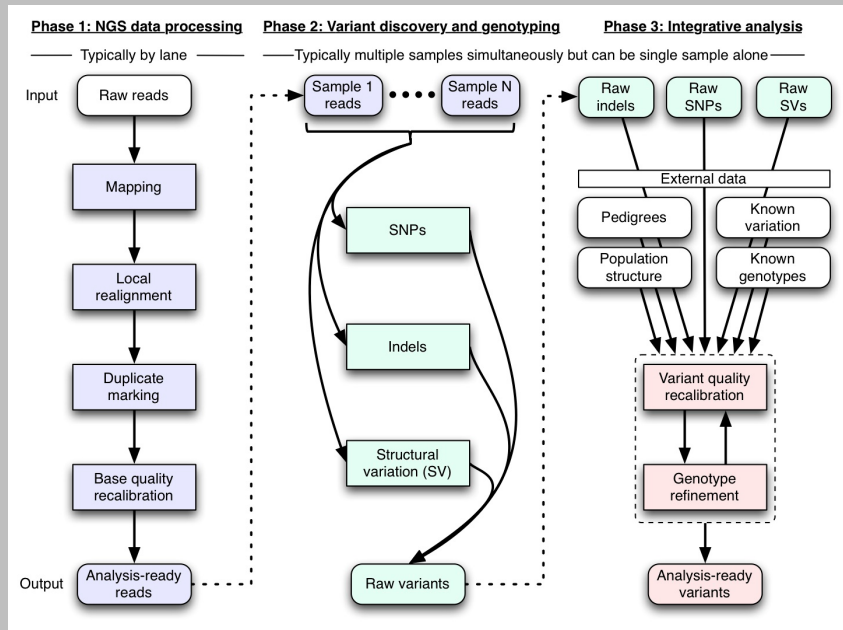
Algorithms (Error framework, Alignment and Assembly)

Bioinformatics (Error modes, Error propagation, Error model, Platform combination)

Personalis EX2052

GATK Variant Calling/Recalibration Pipeline

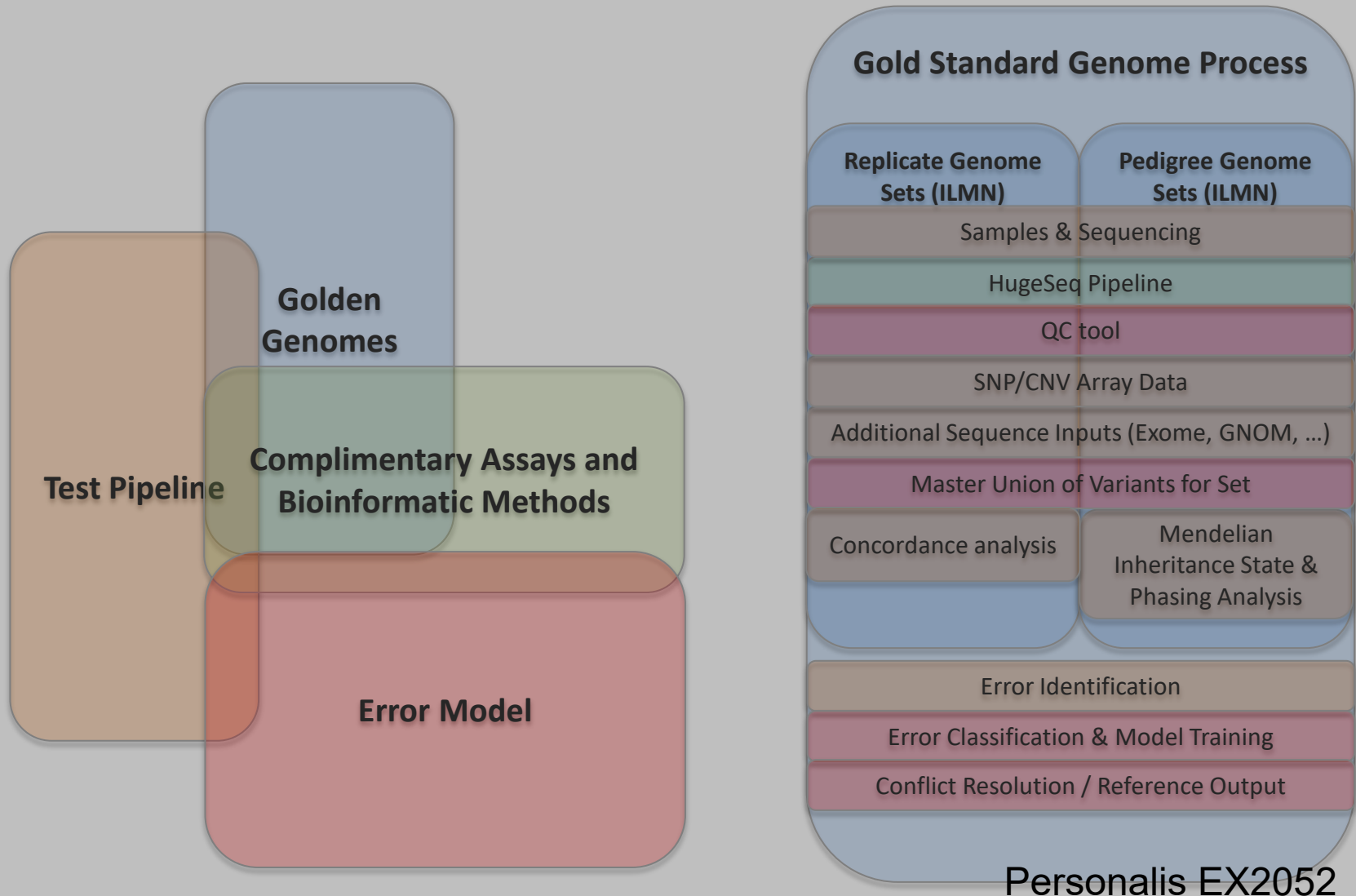
First round for golden genomes



Accuracy Program Facets

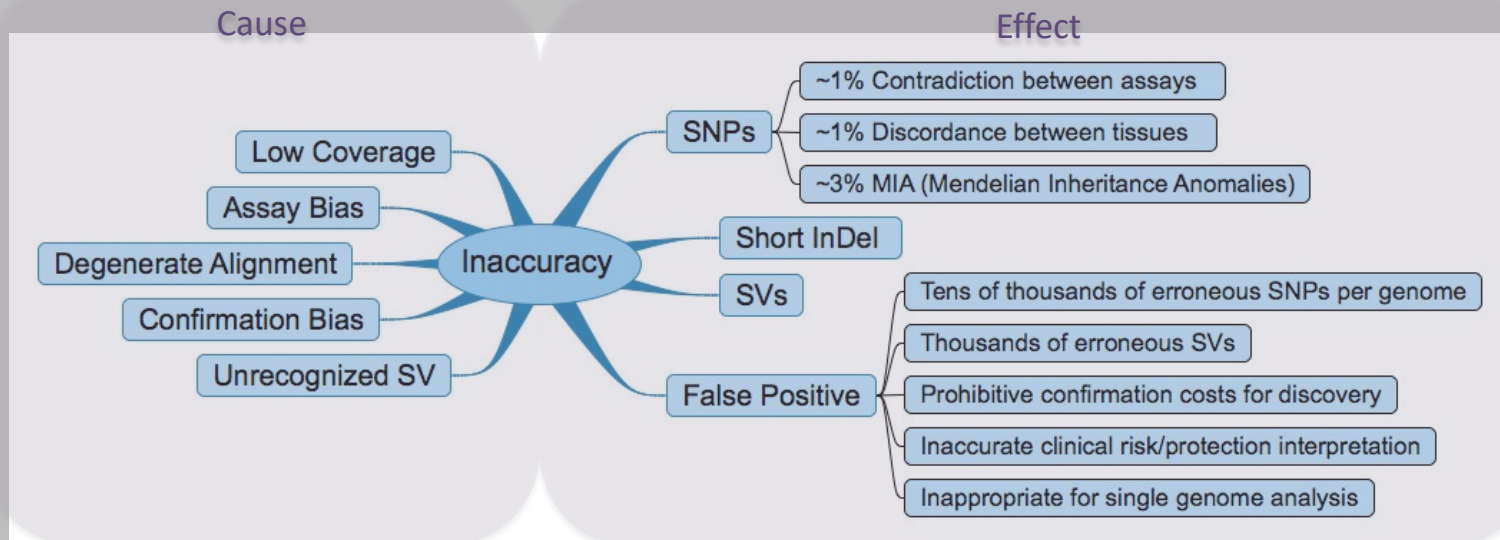
Test, Fix, Ground truth

a



Genome Scale Accuracy

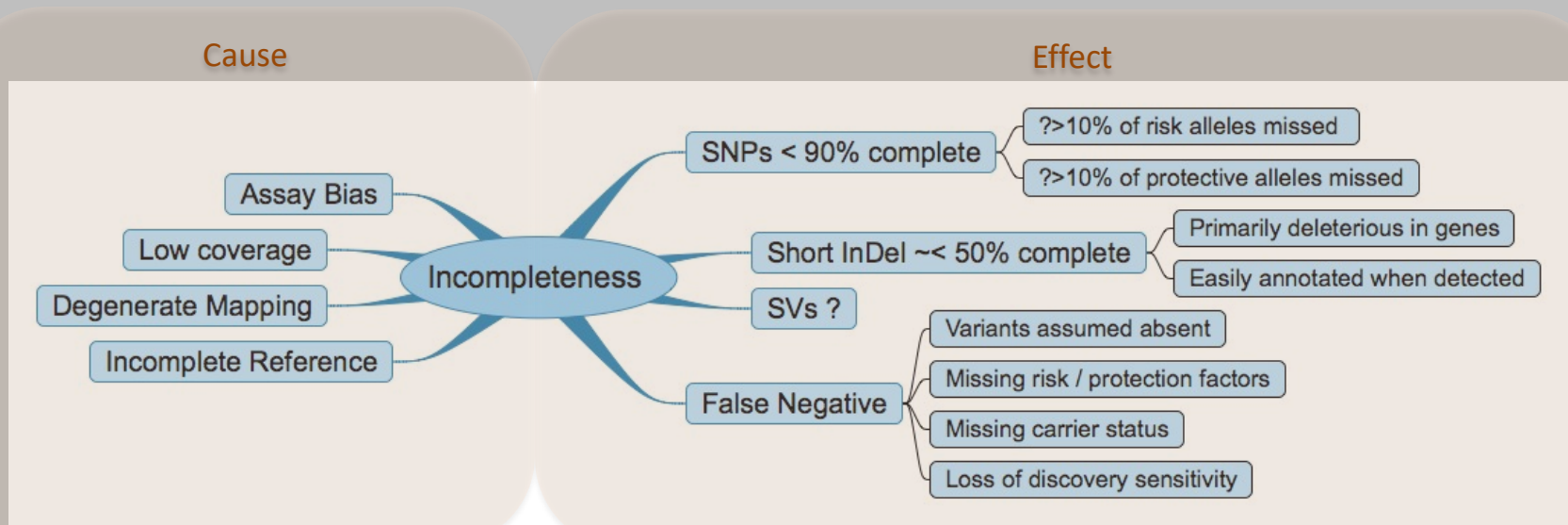
(Miscalls, Mischaracterization and Mitigation)



- No single platform has satisfactory accuracy for individual genome-wide interpretation
- High false-positive rate for SNPs make NGS challenging for clinical interpretation and inefficient for discovery
- Inaccurate characterization of structural variants leads to incorrect interpretation
- + By combining complimentary assays and improving sequence analyses, Personalis improves both accuracy quality assessment to standards unavailable elsewhere
- + Accurate and empirically tested assessment of analytical confidence is a cornerstone of the Personalis clinical interpretation.

Genome Scale Incompleteness

Causes, Consequences, and Corrections



- No single method has satisfactory completeness for individual genome-wide interpretation
- Even in large sample sets, systematic completeness issues degrade discovery sensitivity
- You can't assess what you can measure
- + By combining complimentary assays and improving sequence analyses, Personalis extends genomic coverage beyond anything available today providing the most complete genome-scale analysis available at any price
- + When information is absent, Personalis properly accounts for this implicit uncertainty in clinical interpretation and risk assessment.

Real knowledge is to know the extent of one's ignorance

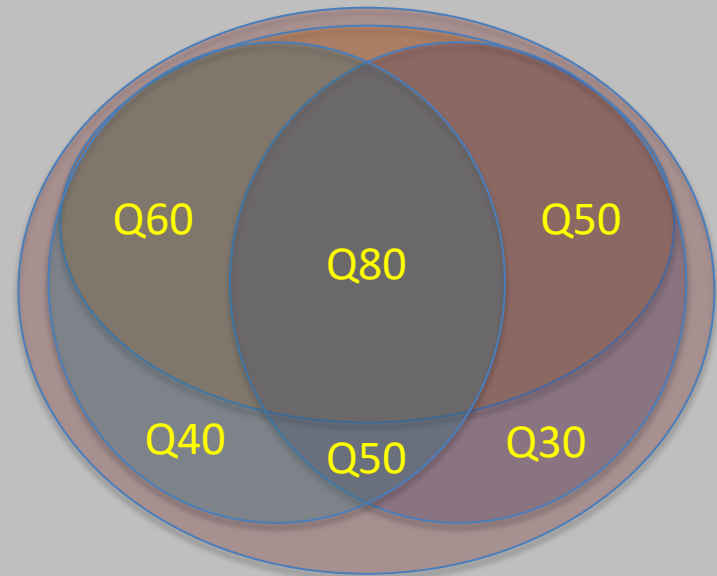
Complimentary Platforms Improves Accuracy and Completeness

State-of-the-art sequencing platforms have clinically challenging accuracy

- Poor raw variant calling accuracy and consistency
 - ~3% error from MIE analysis (Dewey, et al, 2011)
 - ~1% variation between tissues (Lam, et al, 2011)
 - ~1.5% validation failure (Abecasis, 1000 Genomes)
- Poor concordance between platforms
 - 10% discordant SNVs (%2 contradictory), 73% discordant SVs between ILMM & GNOM (Lam, et al. 2011)
 - 30% validation failure of indels (Abecasis, 1000 Genomes)
- Poor coverage or reproducibility within platform
 - Sacrifice 5% of genome to achieve Q50 run-to-run concordance (Ajay, et al 2011)
 - 80-85% accessible (Abecasis, 1000 Genomes)

These error and completeness shortfalls are problematic for individual analysis and can lead to inappropriate clinical interpretations.

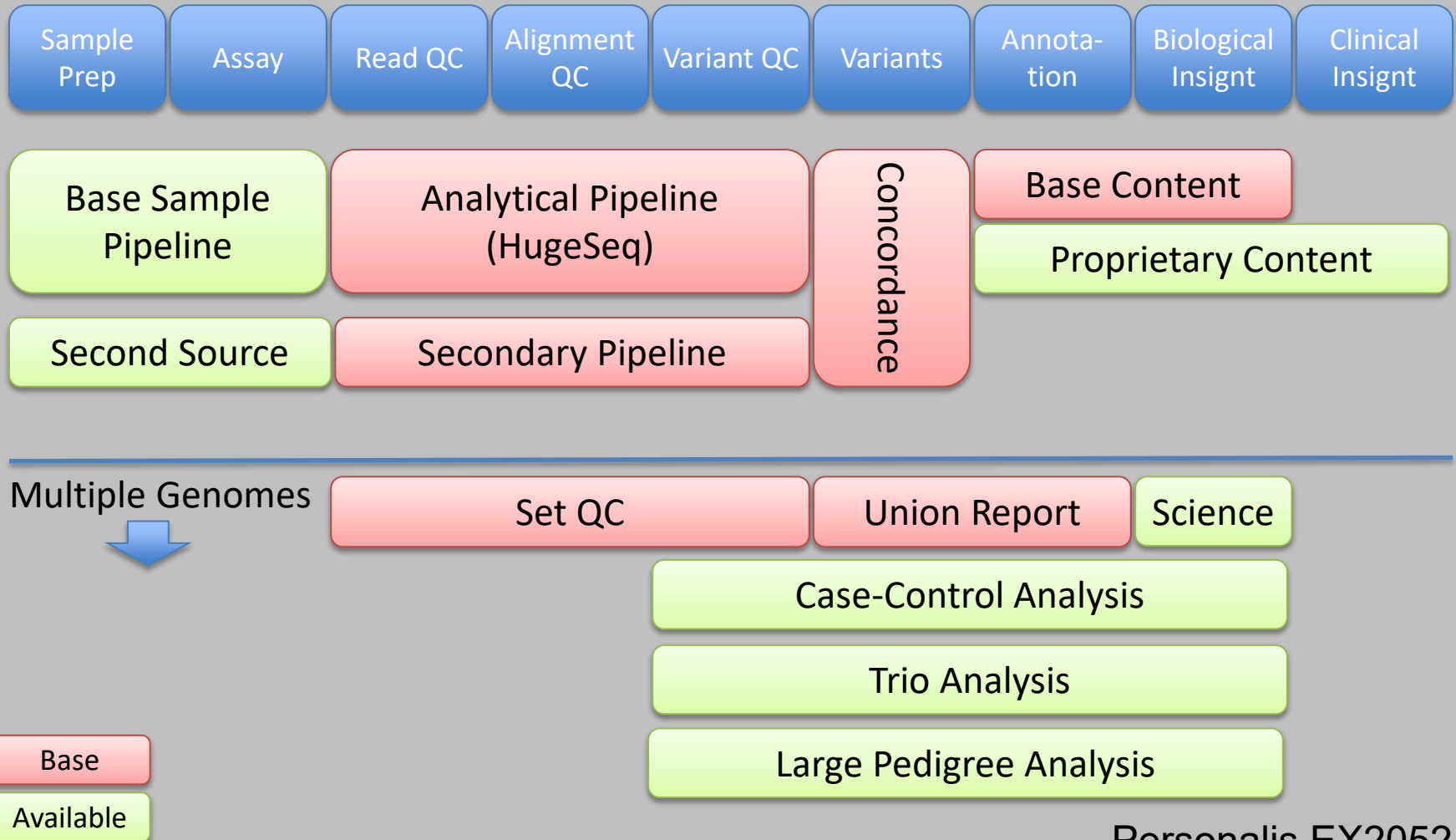
Personalis's multi-platform approach extends coverage and resolves discordant measurements by mapping platform limits of performance.



- High coverage PE DNAseq
 - Distributed over insert sizes – pullout combinations
 - Ultra-high coverage PE ExomeSeq
- 1M Custom SNP array for confirmation and calibration
- RNAseq to corroborate or resolve complex issues?

Personalis EX2052

Pipeline Modules



Personalis sets the Quality Standard in Genome Scale Analysis

Mark Pratt
16 February, 2011

Industry Leading Quality at Every Step of Personalis Genome Interpretation

- Curated Content
Manually curated content using controlled processes and cross-checking
- Complimentary Platforms
Multiple assays combined to extend coverage, corroborate and calibrate results
- Customized and Extended References
Ethnic or custom reference alignment improves accuracy & coverage
- Full Spectrum Variant Identification
Industry leading combination of SNVs, structural and copy number variants
- Validated Analytical Uncertainty Model
Confidence model based on empirical calibration of errors
- Accurate Confidence Assessment of Clinical Interpretation
Combination of biological and analytical uncertainties used in complete model

Sequence Alignment Ambiguity

Issues:

- Degenerate read alignment is suspected to be the primary driver in high sequencing error rates
Approximately 100,000 SNPs per genome are miscalled in standard service sequencing
- Much (10-15%) of genome is poorly accessed by sequencing because of repeat structures
Sequence and hybridization array data are substantially incomplete for both discovery and risk analysis

Approach:

1. Improve reference sequence on SNPs (a) then InDels and other SVs (b) to reduce number of alignment mismatches and improve specificity
2. Employ multiple insert length libraries to resolve (some) degeneracies and enable assembly
3. Improve alignment algorithms and attempt local assemblies
4. Incorporate very-long-read technology (PacBio/ONT) for resolution of long degenerate regions

Concerns:

1. Only modest improvement noted in correcting just SNVs in major allele reference (a) and substantial further development needed to incorporate InDels and other SVs into custom references (b)
2. Significant facilities, development and product costs involved in multi-library preparation
3. Development investment and outcome uncertainty for algorithm development
4. Long read technologies are currently expensive, error prone and poorly commercialized

Curated Content (Placeholder, Gemma WIP)

Issues:

- Unavailable clinical information in public databases
- Poor quality control in publicly available databases

Approach:

- Manually curated proprietary databases
- Human QC cross-check

Concerns:

- Labor intensive and slow

Customized Personalis Genotype/CNV Array

Issues:

- Poor concordance between SNP arrays and sequencing data
Few percent discordance, concordance predicts accuracy, element of error model
- Known set of high value loci, poor sensitivity with commercial sequencing
Missing data on variants in Varimed, PharmGKB will materially impact risk assessment

Approach:

1. Develop custom genotype & CNV array to baseline error model, corroborate and backfill sequence data
2. Include all accessible SNPs in Varimed, PharmGKB, MendelDB, Regulome
3. Include genome-wide QC and CNV loci to aid in sequence interpretation and error calibration

Concerns:

1. Adds \$250? baseline cost to product
2. Exposes Personalis content to array manufacturer (Illumina?)
3. Unknown effort and utility of comprehensive error model

Improved Reference Genome

Propagation of Uncertainty

Multi-library Sequencing

Content?

Poor content drives poor interpretation regardless of data type or quality.

Personalis sets standards in completeness and quality with

→ VariMed

→ PharmGKB

→ MendelDB

- Qualified manual curation
- Certificated checklist (?)
- Certified cross-check and QC (?)
- Administration (?)

Customized and Extended References

Existing reference contains numerous rare alleles and structural variants

- Requires mismatches in alignments increasing ambiguity
- Increases error through reference bias

➔ Reference is part of the problem

Personalis's team has pioneered the use of ethnic references to improve alignment performance.

- Customized references
- Diploid reference

Full Spectrum Variant Identification at Industry Leading Accuracy

- HugeSeq
 - Best of class combination of variant detection

Validated Analytical Uncertainty Model

Poorly calibrated confidence values from variant calling algorithms can result in inappropriate confidence of clinical interpretation.

Existing platforms primarily calibrate on self-consistency and small genome sequencing. These approaches neglect or underestimate the effect of large systematic effects that affect typical human samples resulting from a range of issues from complex repeat structures in human genomes to variability in sample quality.

Filtering is a typical approach improving error rates of sequencing data but often comes at an unacceptable cost of reduced sensitivity and negative bias in clinical interpretation.

Personalis has developed a variant confidence model model derived from data quantifying both precision and accuracy across many samples and technology platforms

- Sample-to-sample, run-to-run, lab-to-lab and coverage based precision model for platforms
- Cross-platform discordance prediction and conflict resolution (??)
- Characterizing and predicting excess “Mendelian inheritance anomalies” in families
- Calibration of systematic error resulting from degenerate alignment

Accurate Confidence Assessment of Clinical Interpretation

Personalis has developed a statistical framework to carry forward all analytical information to interpretation

- *This framework maintains maximum statistical power of data, and*
- *correctly distinguishes between modest confidence variant detection and missing data leading to reduction in false negative bias in interpretation....*