

From: [John West](#)
To: [Richard Chen](#); [Hugo Lam](#); [Mark Pratt](#); [Scott Kirk](#)
Subject: Spreadsheet for 2pm discussion of Accuracy Differentiation
Date: Tuesday, May 22, 2012 1:17:12 PM
Attachments: [Accuracy differentiation JW 22May2012.xls](#)

All,

For our meeting this afternoon on accuracy differentiation, I have put together some notes to serve as a starting point.

They are attached in Excel. Comments are welcome.

John

ACCURACY DIFFERENTIATION

Draft for discussion, JW, 22 May, 2012

ORGANIZATION OF THIS FILE

- Company differentiation around accuracy
- Better sequencing in the laboratory
- Better variant detection & reporting
- More accurate databases, for interpretation

COMPANY DIFFERENTIATION AROUND ACCURACY

- Focus on accuracy relevant to medical interpretation
- Better understanding of the issues than anyone
- Best track record of publications on the issue
- Unbiased from a platform standpoint & able to combine platforms
- More comprehensive data on accuracy than anyone
 - World's only collection of genomes sequenced on both ILMN and CG platforms, plus arrays and karyotyping
 - Largest family pedigree sequenced to high coverage
 - Only genome sequenced on Sanger, ILMN and CGI
- Databases more accurate than those publicly available
- Able to provide a detailed quantitative view of mechanisms underlying errors
- Deep understanding of accuracy issues used to create better results
 - Not just flagging errors, or filtering out those loci, but fixing the problems
 - More insightful approaches delivery accuracy affordably

BETTER SEQUENCING IN THE LABORATORY (WHEN MANAGED BY PERSONALIS)

Focus on getting the whole medically-interpretable genome, accurately, even if more expensive

- Use insight into error types & medical content to keep this affordable

Combine data from multiple different runs of a single platform

- Combine paired-end libraries made with multiple insert lengths
- Use longer read lengths (e.g. 2 x 250 bases, when available later in 2012)
 - More expensive because only MiSeq, but clearly better
- Combine with bulk shorter-read data from HiSeq
- Substantially more efficient at split-read & junction-sequence SV detection
- Key to single-base breakpoint determination

Combine data from multiple platforms

- More experience with Illumina / Complete Genomics than anyone
- May add Ion Torrent, Oxford Nanopore
- Guided by deep understanding of differential error mechanisms in each platform
 - Not tied to any one platform
 - Use whatever it takes to get the best possible combination

Combine data from outside next-gen sequencing

- Several major areas of medical genetics are not well assayed by next gen sequencing
 - Example 1 : Diseases caused by STR-expansion (e.g. Huntington's)
 - Example 2 : Robertsonian translocations
- Personalis will combine NGS data with Non-NGS technologies to create a complete assessment
 - Add karyotyping
 - Add electrophoresis where appropriate (TBD)
 - Others
- Orthogonal technologies also provide validation of NGS results
 - Integrate NGS with array (fluorescence for SV's, in addition to genotypes)
 - Sanger follow-up to findings of specific genomes (option TBD)

Ability to create semi-custom products focused on medically interpretable parts of the genome

- Leverages Personalis advantages in content
- Custom hybridization array
- Custom pullout set
- Other assays to fix specific error types

Question : Should there be a "Personalis exome" option ?

- More comprehensive / accurate at exome price level ?

What is proprietary about this approach ?

Personalis can focus its efforts based on the world's best content (re medical interpretation)
Personalis will develop proprietary understanding about how best to combine multiple technologies
Personalis does not face the competitive & anti-trust barriers that platform companies do
Personalis' people combine deep experience with both platforms and interpretation & can leverage the two against each other
Personalis can combine work in the lab and in bioinformatics, in a way that pure informatics companies can't

BETTER VARIANT DETECTION & REPORTING

Fewer false positive SNP's due to method of generating laboratory data

Combination of paired-end insert lengths covers more of the genome uniquely

Fewer false negative SNP's due to method of generating laboratory data

More uniform coverage by combining library prep methods & platforms

Orthogonal validation of millions of SNP genotypes by array

Integrated with next gen sequencing data, not just another separate report

Better alignment, due to better reference sequence

SNP major allele ref by ethnicity (in our first product)

InDel major allele ref by ethnicity (later)

Other advances as R&D develops them:

Iterative alignment

TBD changes in SNP alleles called with Personalis reference vs public standard

Likely more improvement in non-European ethnicities

Include (eventually) changes in InDels called as well

We provide the only support available specifically for admixed genomes

Major-allele non-ethnic reference, or even more advanced options

Focused effort to align in the presence of SNP / InDel clusters, MNP's

May leverage Hugo's BreakSeq approach

May need time in the development plan

Better SNP reporting due to better reference

We report variants when sequence is a homozygous match to the public ref but that's the minor allele

Entirely missed by systems which use the public reference

Example : Factor V Leiden

Rong had a whole paper on all the disease variants in the public ref

> 1M loci where we can be different

We should calculate the average # actual loci / genome, by ethnicity

At het loci, we report both alleles, but we report the minor allele as the variant

Not the allele which is different from the public reference

Better detection of SV's

Better lab data for SV detection:

Longer reads (better for approaches based on split-read & junction-sequences)

MiSeq 2x250 or other platform

Multiple insert lengths

Electrophoretic assay of STR-expansions

Karyotyping for Robertsonian translocations

Orthogonal technologies for validation of SV's:

Fluorescence intensity data from hybridization arrays

We combine the results from five different algorithmic approaches

We test our SV algorithms by Mendelian Inheritance in high coverage whole genome family data sets

one which was sequenced with ten different paired-end libraries spanning 200 - 40,000 bases

and validate them using fluorescent intensities from high density hybridization arrays

We don't treat all SV's as novel - we have the world's best database of known SV's and their junction sequences

Detection is better when you know exactly what you are looking for

We should have a meeting to discuss how we can (easily ?) build this

Start with 1,000 Genomes result Hugo has helped create

Large data set but low coverage may make detection less certain in low MAF SV's

Others will be able to access this eventually, potentially catching up, or claiming to

Augment this with (more confident ?) SV's from:

Full coverage (30-40x) genomes (West, Altman, 40 Koreans, others we can download)
High coverage (>60x) genomes (Snyder, CEPH1463, Venter, others ?)

Better reporting of SV's

We determine the zygosity of deletions and report it
Deletions integrated with SNP report, e.g. "A-" vs "AA" inside a het deletion
We report SV's with their allele frequencies in the ethnicity matching the sample

Flagging of potential errors

Many subtle error types not recognized by others
Error mechanisms underlying differences when the same person is sequenced twice (it's not just Poisson !)
Error loci determined from deep & multi-platform sequencing of large families
Error loci determined by extensive platform comparison, both NGS/NGS and NGS/Non-NGS
Detailed understanding of compressions, and large unpublished catalog of them

MORE ACCURATE DATABASES, FOR INTERPRETATION

Cleaner databases:

Well financed, systematic manual curation to industrial QC standard
Standardized medical language hierarchy
Extensive cross checking of databases developed independently
VariMed vs HGMD
MendelDB vs OMIM
Personalis PharmGKB vs public PGKB (need to be careful in this positioning)

Databases others will not have:

Regulome
BreakSeq (esp if augmented with private Personalis data)
Compression list (described in publications but not released)
Variant data derived from a broad collection of genomes
Multiple public data sets, some processed in proprietary ways by Personalis
Access to private data sets, sequenced by others
Access to private data sets, sequenced by Personalis