

From: [Scott Kirk](#)
To: [John West](#); [Christian Haudenschild](#); [Richard Chen](#); [Hugo Lam](#); [Rong Chen](#); [Mark Pratt](#)
Subject: PGS v1 Requirements draft (rev 1.10)
Date: Wednesday, October 10, 2012 12:51:06 PM
Attachments: [PGS_v1.10_RemainingRequirements.docx](#)

Hi,

V1.10 is attached which includes a few edits and additional items from the accuracy discussion yesterday as well as a number of updates from the discussion this morning which covered secs 8,9,10,11, and part of 6 on SVs.

Thanks,
Scott



Personalis Genome Service

Remaining Requirements for v1.0

Draft v1.9

Date	Version	Author	Description
9/30/12	V1	RC, HL, MP, SK, ROC, GC, MC	First consolidated version
10/1/12	V1.1	ROC, SK, HL	Updated SV and pipeline requirements post discussion with Hugo. Merged in first round of changes from Rong
10/2/12	V1.2	RC, SK, ROC, MP	Second round of changes from Rong First round of changes from Mark.
10/3/12	V1.3	SK	Reorganization and miscellaneous edits and some priority tagging
10/4/12	V1.4	HL	Pipeline performance testing
10/5/12	V1.5	SK, MP, ROC, HL, RC	Priority tagging and miscellaneous edits. Additional refinement of exome plus requirements.
10/7/12	V1.6	SK, MP	Further refinement of exome + requirements
10/7/12	V1.7	SK, RC, MP, HL, ROC	Revisions post VA gap analysis – VA gaps is now traced to the numbering convention. Try not to change numbers.
10/8/12	V1.8	SK	Included mockup graphics in various sections and updated accuracy section some.
10/8/12	V1.9	SK	Additions and edits from accuracy discussion

Table of Contents

1	Key Differentiators.....	3
2	Scope	3
3	Key Dependencies (if any).....	3
4	Format and Priority Definitions	4
5	Exome Plus Requirements.....	4
6	Structural Variants Requirements	10
7	Additional Pipeline Requirements:.....	18
8	Build a repository of public genomes/exomes to calculate population frequencies, serve as controls, test our pipeline, and identify disease mechanisms.....	21

9	Curate proprietary databases with quality control.....	23
10	Additional Annotation requirements	26
11	Case/control discovery module	30
12	Database accuracy study	33
13	Reports Requirements	36
14	Performance:	52
15	Additional testing requirements	53
16	Additional Accuracy Related Features	55
17	Commercial	59

1 Key Differentiators

- Exome Plus support:
 - Pulldowns for long and short read sequencing that include custom content. Provides a differentiated laboratory component in our offering covering our proprietary content outside of standard Exome assays as well as regions with known accuracy issues.
- End to End SV support:
 - SV is an area where most competitors do very little. Offering more here significantly differentiates us. We will be also producing our own frequency data as well as deriving it from public data sets.
- Database accuracy study:
 - Provides a check on the quality of our proprietary databases
 - Provides marketing content around our proprietary databases
- Base Pipeline and Annotation engine.

2 Scope

- V1 of PGS
- Document also highlights features that critical for VA proposal

3 Key Dependencies (if any)

- VA proposal

- Exome Plus:
 - Custom pulldown orders from external manufacturers (Agilent, ILMN).
- End to End SV support:
 - external tools
 - test data
 - databases
- Database accuracy study:
 - tools to map the variants and phenotypes across proprietary and public databases

4 Format and Priority Definitions

Each requirement should include:

- **Requirement statement:** Statement of what the requirement is. Can be simple descriptive sentences or in user based story format (as an [persona], I want [feature], so that I can [do something]). Independent of priority, “shall” generally means required and “may” means optional.
- **Additional Description:** provide additional details and technical requirements as needed.
- **Definition of Done:** what criteria need to be satisfied for the requirement to be considered “done”. Helps to define how requirement is tested.
- **Priority:** see definitions.
- **Unique identifier:** will be added when we enter the requirements into Jira.

Priorities definitions:

- **P1** – Must have. Critical for the release.
- **VA** – required for VA proposal
- **P2** - Nice to Have. Implement if time allows.
- **P3** - Future: not required, but is perhaps desirable in the future. May require some work to make them easier to implement later on.
- **P4** - Deferred: the feature is deferred until a future/next release.

5 Exome Plus Requirements

5.1 Exome plus extends the standard exome to areas of medical significance outside the exome including regulatory regions, pharmGKB, Varimed, HGMD, and mendelian variants.

Priority: P1

Area	Priority	Comments
------	----------	----------

Varimed	P1	
PharmGKB	P1	
Regulome tier 1	P1	
Mendelian gene exons	P1	
HGMD	P1	
Regulome tier 2	P2	
Gene panels	P2	Could be a P2, it's next up if time allows.

5.2 Exome plus fixes the following types of problematic regions in areas of medical significance in both exomes and whole genomes:

Priority: P3

Additional Description:: for the first release of this, work is to be scoped through the lens of the clinical exome panels. For example, focus on fixing problematic regions in the cardiomyopathy, FH, LongQT syndrome panels.

Area	Priority	Comments
Degenerate Alignment	P2	use long reads to improve these areas
GC rich regions	P2	can we improve coverage in GC rich regions (more iterative than the other three and implies more research)
Unstable expanding repeat regions/VNTRs/STRs	P2	
Content intersecting SVs	P2	
Content intersecting compressions	P2	
Phased exons(?)	P2	
P2	P2	
HLA typing	P2	
Allelotyping, unstable expanding repeats	P2	

5.3 Design exome plus to augment established low coverage regions across the interpretable exome to achieve more uniform coverage and thus higher quality.

Priority: P2

Additional Description::

- High GC and poorly mapped (degenerately mapped) regions
- Solution is longer reads and a GC specific process.

5.4 Implement first pass Exome Plus to Achieve 5.1, 5.2, 5.3

5.4.1 An exome supplement pulldown (ESP) to cover missed and low coverage areas in standard NGS exome sequencing shall be developed.

Priority:P1

Read length is 2x150 (?) and content shall include coverage of:

Area	Priority	What it covers/details
VariMed	P1	
pharmGKB	P1	Sarah's set of 7800 genes and any SNP within vicinity of these genes
HGMD genes	P1	3502 HGMD genes
Low coverage	?	
High GC content	?	
Missed splice sites	?	
Homopolymer flanks	P2	
Important Regulome 1 – overlapped with medical variants	P1	
Important Regulome 2	P1	
Highly conserved regions	P1	
Hgmd, omim, varimed, mendelDB overlap	P1	cover key disease related variants in our proprietary databases as part of the pullout.
Intronic variants (possible differentiator)	P1	
HGMD 3400 disease causing genes	P1	
ILMN Truseq exome expanded by 50 bases(?), trimmed down by low and zero coverage areas.	P1	

5.4.2 An exome long read pulldown (LRP) set to cover areas that are interpretable, but challenging with standard NGS exome sequencing shall be developed.

Priority:P3

Read length is 2x250. Is therefore dependent upon selection, integration, and testing of a long read aligner.

Area	Priority	What it covers/details
areas in OMIM gene, pharmGKB, HGMD gene sets with mapping problems.	P3	zero coverage regions
mapping problems where assembly and compression is an issue hitting anything interpretable.	P3	
HLA regions	P2	
Unstable repeat dz	P3	
Exomic seg dups	P3	
Exomic compressions	P3	
Content sv breakpoints	P3	
Exomic long strs	P3	
Pharmgkb problems	P3	

Common exomic indels	P3	
Suspect medical snps	P3	
Small, high value sets	P3	
3 whole compressions on OMIM genes	P3	
amylase family	P3	
SMN1	P3	
SMN2	P3	

5.4.3 An exome High GC Supplement (HGCS) to cover regions of high GC content known to be problematic in NGC exome sequencing shall be developed.

Priority:P3

Content shall include coverage of:

Area	Priority	What it covers/details
Exomic high GC regions	P3	Including high GC SNPs
First exons	P3	Very high GC. Poor coverage now in ILMN v3 chemistry is an improvement over the past, so not necessarily a knock against the method.

5.5 Exome+ pulldowns shall be evaluated to determine how well various experimental strategies work and to allow marketing communication about product performance and quality.

5.5.1 The overall quality of the Exome+ pulldown assay will be evaluated and summarized by measuring the following characteristics:

Priority:P1

We may want to add these to 5.11 for measuring within each type of content region as they could be different? They could also vary by vendor selected.

<<Need to determine spec numbers to test against>>

Metric	Target	Min. Acceptable
On/Off target quality /enrichment efficiency (specificity)	Base coverage of XX% at >= 10x coverage depth. <XX>	XX% at >= 10x coverage depth. <XX>
Sensitivity	SNPs:<XX> Indels:<XX> SVs:<XX>	SNPs:<XX> Indels:<XX> SVs:<XX>
Uniformity of coverage	10, 20, 30x??<XX>	<XX>
Reproducibility	95%<XX>	<XX>
Other measures?		

5.5.2 The performance of Exome+ will be evaluated for each of the defined areas of content by measuring the following characteristics:

Priority:P1

<<Need to determine spec numbers to test against>>

Metric	Target	Min. Acceptable
Comparison to standard exome results	<XX>	<XX>
On/Off target quality /enrichment efficiency (specificity)	Base coverage of XX% at >= 10x coverage depth. <XX>	XX% at >= 10x coverage depth. <XX>
Sensitivity	SNPs:<XX> Indels:<XX> SVs:<XX>	SNPs:<XX> Indels:<XX> SVs:<XX>
Uniformity of coverage	10, 20, 30x??<XX>	<XX>
Coverage of content region	10, 20, 30x?<XX>	<XX>
Sensitivity of detecting specific medically relevant loci (per panel basis?)	SNPs:<XX> Indels:<XX> SVs:<XX>	SNPs:<XX> Indels:<XX> SVs:<XX>
Specificity of detecting specific medically relevant loci (per panel basis?)	SNPs:<XX> Indels:<XX> SVs:<XX>	SNPs:<XX> Indels:<XX> SVs:<XX>
Reproducibility	95%<XX>	<XX>

5.5.3 To improve our ability to measure quality and performance of the Exome + pulldowns, produce specialized reports and analytics that summarize performance characteristics for the following areas:

For example: a report on “here’s our performance on HGMD variants: coverage, variant detections, mean confidence”

Possibly include ref calls at critical variant loci. Nice to know you have an affirmed null detection of some nasty variant.

Area	Priority	Comments
Varimed	P1	
pharmGKB	P1	
Regulome tier 1	P1	
Regulome tier 2	P1	
Mendelian gene exons	P1	
Blood typing	P1	Double check VA proposal
HLA typing	P2	
Allelotyping, unstable expanding repeats	P3	
Content intersecting SVs	P3	
Content intersecting compressions	P3	
Phased exons(?)	P3	
Gene panels	P3	Could be a P2, it’s next up if time allows.

5.6 Overall cost of exome plus needs to be significantly less than a whole genome. (What is the target number?)

Priority: P1

5.6.1 Produce a detailed cost model of the entire exome plus workflow and use to determine what parameters we can work with to impact cost/sample.

Priority:P1

Additional Description:: Determine what parameters can be worked with. It's not necessarily content that reduces cost. Multiplexing, for example, is a possibility to reduce cost/sample.

5.6.2 Develop an itemized value assessment of content to show the marginal value of the data in various content sets.

Priority: P1

5.7 The clinical exome panel must “finish” the well established gene panels for cardiomyopathy, FH, LongQT syndrome.

Priority: P3

Additional Description:: First release is a research (exome plus) and doesn't include panels. Second release: clinical (exome panel)

5.7.1 For the cardiomyopathy, FH, LongQT syndrome. clinical panels scope what the current accuracy is.

Priority: P3

5.7.2 Determine the data combination and weighting strategy(??)

?Provide the ability to combine and weight data to ??? Not sure what this is.

Priority:

5.7.3 The pipeline analysis shall be enabled to make calls on combined data sets for single genome/exome/supplement (how is this different than multisample support?)

Priority:P1

5.7.4 Provide the ability to allelotype unstable expanding repeat diseases.

Priority:P3

5.7.5 Provide the ability to perform reassembly of SDs, compressions, and SVs to ???

Priority:P1

5.7.6 Provide the ability to detect a sample's HLA type from exome + data.

Priority:P1

5.8 Integrate support for exome+ workflow(s) into pipeline to enable analysis, variant calling, and annotation of the data in an automated fashion for multiple samples.

Priority: P1

5.8.1 A long read aligner shall be enabled to support 2x250 reads from the LRP analysis.

Priority:P3

5.8.2 GATK-lite version currently implemented in the pipeline shall be examined to determine it's compatibility with Exome + data.

Priority:P1

Definition of Done:

6 Structural Variants Requirements

6.1 Pipeline shall detect the following SV types:

We do most of these already. It's a matter of improvements and doing really well with long deletions and stay with what we have for the others for V1.

6.1.1 CNV

Priority:P2, VA

6.1.2 Insertion >50 bp (long insertion)

Priority:P2, VA

6.1.3 deletion >50 (long deletion)

Priority:P1, VA

Feel we can get the most gain by doing well with long deletions.

Prioritize those that are medically interpretable.

Exome data deletion detection.

6.1.4 inversion

Priority:P2, VA

6.1.5 interchromosomal translocation (potential place we can stand out)

Priority:P2, VA

6.1.6 Intrachromosomal translocation

Priority:P2, VA

6.1.7 VNTR (detect STR, get some VNTR with Breakseq)

Priority:P2, VA

6.1.7.1 Integrate LobSTR for STR detection

Priority:P2, VA

Contributors: Hugo, Dan, Nan

Additional Description: Detecting low complexity SVs is one of the most difficult SV detections. No known products are particularly targeting on detecting STR comprehensively. By using LobSTR, particularly in conjunction with our long reads capture, we shall detect the most STRs.

1. Integrate LobSTR as an SV detection tool in the SV module of the pipeline
 - a) Installing dependent modules on the server and in the pipeline
 - b) Adding a new SV algorithm to the pipeline

c) Integrate the STR calls with existing SV call set

Definition of Done:

6.1.8 MEI

6.1.8.1 *Implement MEI (Mobile Element Insertion) detection which can be implicated in human disease such as schizophrenia (?)*

Priority:P1, VA

Contributors: Hugo, Jing, Dan

Additional Description: **Detecting repetitive elements has always been a big challenge in the field, particularly with relatively short reads; however, MEI has been shown in various study before causing different diseases and somatic differences. There is currently no known (publicly) available tool for MEI detection.** For Complete, they have MEI detection but the quality is unknown. We shall develop a new detection tool with the best algorithms detecting mobile element insertions such as L1 and ALUs.

Looks for gene disruption.

ILMN data on BWA needs it's own algorithm written. Start from algorithms which have been published and rewrite for ILMN/BWA data.

There may be a way to take what we're doing already as a first pass implementation for identifying "existing" MEI. A more involved implementation would consider "denovo" MEI.

LOE is 1.5 Mo (1FTE) + validation time for more involved implementation.

Definition of Done:

6.1.9 Polyploidy/aneuploidy

Priority:P3

Session with Sarah and Gemma to determine what we should be scanning for.

Definition of Done:

6.1.10 SVs shall be flagged in the GFF file output by the pipeline.

6.2 Pipeline shall improve the accuracy for detecting the above SVs at increased sensitivity and specificity compared to Complete or other competitive solutions (including validation).

Priority:P1, VA

Improve to add a “p” value to the detection replacing the high/low classes that currently exist.

Testing:

Definition of Done:

6.2.1 Provide the ability to trade off between sensitivity and specificity on a per run/study basis.

Priority:P1, VA

Additional Description:

For example in discovery research we may want to emphasize sensitivity over specificity while for clinical applications specificity may be more important. ROC curves shall be constructed to measure.

Definition of Done:

6.2.2 Specifically assess accuracy and performance for medically important genes that are the focal point of our disease panels. For example: cardiomyopathy/HCM, long QT syndrome, Familial hypercholesterolemia, and the associated pharmacogenomic regions.

Priority:P1, VA

Definition of Done:

6.2.3 Generate latest breakpoint library from 1KG to improve accuracy of all types of SVs.

Priority:P1, VA

Contributors: Hugo, Jing

Additional Description: **With the collaboration with 1KG, particularly the SV group (Lam et. al. >4yrs experience), Yale (Gerstein’s lab, SV expert) and EBI (Jan Korbel, the SV leader in 1KG),** a latest SV library from the 1000 genomes shall be generated with breakpoint and validation information. We shall also run breakseq and possibly other SV tools to generate stringent breakpoints. Improves both sensitivity and specificity.

Definition of Done:

6.2.4 Local reassembly to be more precise in resolving breakpoints and to validate the SVs.

Priority:P1, VA

Contributors: Hugo, Jing

Additional Description: **None of the existing SV callers perform local reassembly. SVs detected by certain algorithms are also of low resolution where breakpoints are unclear.** Performing local reassembly in breakpoint regions to determine more breakpoints and to validate the SVs.

Definition of Done:

6.2.5 Report breakpoint resolution when possible to better assess the downstream effect on gene function

Priority:P1, VA

Contributors:Hugo, Jing

Dependency: 6.2.3, 6.2.4

Additional Description: **Many of the SVs detected nowadays are not of breakpoint resolution.** Algorithms shall be developed to report the breakpoint information for as many SVs as possible, e.g. by doing local reassembly in breakpoint regions.

Definition of Done:

6.2.6 Report zygosity for SVs when possible to better assess impact on gene function and disease, P1

Priority:P1, VA

Contributors: Hugo, Dan

Dependency: 6.2.5

Additional Description: **Thus far, none of the SV algorithms report the SV zygosity.** For the SVs detected by the pipeline, algorithms shall be developed to report the SV genotype with zygosity such as homozygous/heterozygous deletions/insertions.

Definition of Done:

6.2.7 Refine SV merging algorithms to better annotate SV and increase precision of merging overlapping SVs (for example better detection of compound heterozygosity or overestimate on size of SVs)

Priority:P1, VA

Contributors:Hugo, Dan

Dependency: 6.2.6

Additional Description: **SV merging is critical as SVs detected by the algorithms are sometimes broken into fragments or overestimated and multiple algorithms also report SVs largely overlap; however, no known algorithm or software can accurately consolidate, merge and then catalogorize these SVs.** Without phasing information, it is particularly hard to do it accurately. Algorithms shall be developed to refine the current SV merging in the pipeline with the aid of multiple orthogonal SV algorithms, the zygosity information and the SV breakpoint/frequency data sets we are generating.

Definition of Done:

6.2.8 Provide an SV characterization tool to understand the formation mechanism of SVs (targeted to discovery researcher)

Priority:P2

Additional Description: **Inference of the ancestral state of an SV locus is crucial for relating SV surveys to primate genome evolution and population genetics. Inference of its formation mechanism is crucial for understanding how the SV was generated, which may then intersect with exons of genes or that lead to gene fusion events causing diseases. Characterizing remains a challenging problem until our first publication (Lam et. al. 2010).** Based on our previous experience, we shall develop an in-house tool to infer SV formation mechanism and ancestral state.

Definition of Done:

6.3 Sensitivity and specificity for each type of SV shall be tested via the following methods

Main focus on long deletions for V1 (P1)

6.3.1 Create simulation data sets to model different types of SVs

Priority:P2

Contributors: Hugo, Ming, Mark, Jason

Additional Description: **There is currently no known robust SV accuracy/error model.** Using an existing simulation box or creating our own to generate simulation data (e.g. sequence reads) to model different types of SV and to test their detection. If possible, a p-value should be generated and assigned to the SVs being called.

Definition of Done:

6.3.2 Comparison with SV results generated with orthogonal technologies (for example Complete Genomics data we've had sequenced).

Priority:P2

Manual validation of SV detected across multiple algorithms.

Definition of Done:

6.3.3 Describe relative performance statistics (sensitivity and specificity) for our SV detection algorithm versus:

6.3.3.1 Illumina seq + each individual SV detection algorithm (Pindel, etc)

6.3.3.2 Complete genomics pipeline

6.3.4 Comparison to 1000 genomes and available results in the literature.

Priority:P2

As a golden data set to start working with.

Definition of Done:

6.3.5 Perform validations of detected SVs with PCR methods to validate accuracy and resolve discordance

Priority:P2

Contributors: Hugo, Jing, Shujun

Additional Description: For regions not or badly validated by public such as 1KG and are in content regions of interest/importance, PCR shall be performed to validate the existence and zygosity of the SVs.

Definition of Done:

6.3.6 Asses the impact of PCR or Gel free sequencing on SV detection

Priority:P3

Contributors: Hugo, Jing, Shujun

Additional Description: PCR free might change the GC of sequences and gel free might change the fragment size, which may have impact on various detection algorithms.

Definition of Done:

6.3.7 Finalize the new breakseq by testing full-scale using 1KG data to fine tune the sensitivity and performance of the algorithm(s)

Priority:P1, VA

Contributors: Hugo, Jing

Dependency: 6.2.3

Additional Description: **We are the author of BreakSeq (Lam et. al.) and a new version with zygosity information has just been released at Personalis.** We will be testing the new version of breakseq on 1KG data (~1000 samples) and fine tune the sensitivity and performance.

Definition of Done:

6.4 Build a unified and comprehensive SV/CNV database for downstream annotation

Priority:P1, VA (ask them)

Contributors: Hugo, Dan, Jing, Ming

Dependency: 6.2.5-6.2.8

Potentially focus on deletions at first.

Additional Description: **Personalis will be the first one with a cleaned, unified, comprehensive, and largest SV database with different level of resolutions, different studies, validation data, highest amount of breakpoints, breakpoint formation mechanisms, ancestral states, individual zygosity information where possible, comprehensive frequency data as well as possible phenotypic data.**

1. Public call set (Hugo, Jing, Dan)
 - a. Gather SVs with different level of resolutions (e.g. breakseq library, 1KG CNVs, 1KG SVs, etc)

- b. Develop algorithms to resolve the differences and enhance the quality of heterogeneous SVs and CNVs
- 2. Internal call set (Hugo, Dan, Ming)
 - a. Run SV module of the pipeline on our data sets to produce our own call set with possible zygosity and frequency data
 - i. Using uk10k as well (6000 exomes) – get allele frequency data for SVs.
 - ii. Look at the 25 genomes that we have.
 - iii. Grow these over time (aggregated stats from everything we run)
 - iv. Systematic approach to pulling in data to generate the information such as frequency
- 3. Consolidate the call sets to generate the highest resolution data possible (Hugo, Dan)
- 4. Validate with public datasets (e.g. 1KG), local re-assembly, and internal experiments where necessary (e.g. important content regions) (Hugo, Jing)
- 5. Generate frequency data and doing statistical validation such as hardy weinberg. (Hugo, Dan)
- 6. Correlate the final, cleaned, unified database with any phenotypic data available. (Hugo, Dan, Jing)

Definition of Done:

6.5 Detect SV in exome regions and assess accuracy to strengthen our ability to “finish” exomes (esp. for clinical application of exome+). This is also a key differentiator because nobody else is doing this commercially.

Priority:P1

Additional Description: **Detecting SV in exome regions are known to be hard and existing SV detection tools usually don't work with exome data; however, SVs in the exome regions are critical, e.g. indels in BRCA genes. Also, no known exome product is known to generate SV calls.** We shall detect CNV and SV from exome data using the latest exome-specific algorithms as well as algorithms developed in-house.

Definition of Done:

6.6 Choose an SV file format and multi-sample representation that works well for multi-algorithms, multi-sample, and downstream analyses.

Priority:P1

Contributors:Hugo, Dan, Steve, Rong

Additional Description: **There is currently no known good file format for representing SVs, particularly for multi samples. 1KG is using VCF, but its known to be not optimal.**

1. Currently GFF for every sample. (Dan)
2. Come up with a format or a way that works well for multi-algorithms and multi-sample, as well as the downstream analyses (Hugo, Dan)
3. Changing the pipeline accordingly (Dan)
4. Think about what a multisample file should have in order to work for downstream analyses. GVF is a possible option for representing multisamples. (Rong, Steve)

Definition of Done:

6.7 Create stats, QC, and annotation report for SV to deliver results.

Priority:P1, VA

Contributors: Hugo, Dan, Rong, Michael, Steve

Additional Description:

1. For VA we'll be annotating the GFF file for each individual sample.
2. In addition to annotated GFF, for v1 consider creating a multisample representation/merged SVs for annotation and comparison across samples. (Risk in this lies in the "fuzzy" criteria for merging across samples and annotating based on overlap with gene). Need to explain methodology and validate.
3. Variant (stat) report: We are currently reporting SV association with different genomic elements (e.g. repeat elements, genes) (Hugo, Dan)
 - a. elements (e.g. with other public datasets like DGV)
4. Gene report: Annotate disease for genes overlapped with SV (Rong, Michael, Steve) (see 6.5)

Definition of Done:

6.8 Create report for SV associations with diseases (for annotation) to deliver results

Priority:P1, VA?

Contributors: Rong, Steve, Michael, Hugo

Additional Description:

1. dbVar (Rong, Steve)
 - a) Investigate and download dbVar
 - b) Parse dbVar into MySQL
 - c) Separate SV in clinical and normal samples
 - d) Connect ID, genomic coordinate, and disease
 - e) Design and implement dbVar report

2. Generate report for diseases associated with SVs

Definition of Done:

6.9 Create report for statistically significant SV associations with genes, pathways, disease in case/control.

Priority:P2

Dependency: 6.7

Research project, but would be a large differentiator.

7 Additional Pipeline Requirements:

7.1 Provide the ability to determine relatedness, ethnicity, and sex to improve downstream annotation.

Priority:P1, VA for Sex determination. P2 for relatedness and ethnicity.

Contributors: Hugo, Dan, Jing

Additional Description:

Consult with Carlos and he may have suggestions or some things we could use.

Definition of Done:

7.2 Provide the ability to determine blood type from sequence data to improve ability to identity match sequence to sample record for the VA proposal.

Priority: P1, VA

Definition of Done:

7.3 Provide the ability to determine HLA type to improve downstream annotation (especially for Exome +)

Priority:P2

Definition of Done:

7.4 Determine how the raw data should be delivered (e.g. file structures, formats, and transfer/devices)

Priority:P1, VA

Additional Description: Look at Complete and ILMN outputs we've received.

Definition of Done:

7.5 Test browser compatibility for output

Priority:P1

Contributors: Hugo, Ming

Additional Description: We will be testing IE8+, Firefox, Chrome, and maybe Safari for all HTML outputs. Prioritize Firefox and Chrome.

Definition of Done:

7.6 Test compatibility with public browsing tools and document how to use them. Tools include Broad-IGV, ??other browsing favorites?? Part of a larger interaction design.

Priority:P1

Definition of Done:

7.7 Get GATK2.0 license when available and consider performing update to the full GATK2.0 version.

Priority:P1

Contributors: Hugo, Dan, Scott

Additional Description:

Full GATK update in the fall when Broad releases the version from beta.

Mostly consists of integration work.

Definition of Done:

7.8 Unique Personalis IDs shall be assigned to all variants generated.

Priority:P1

Contributors: Hugo, Dan, Michael

Additional Description: Generate unique Personalis variant IDs for all variants generated. Particularly important for the variants that don't have RSIDs.

Definition of Done:

7.9 Integrate analysis and annotation pipeline with our sample tracking/LIMS system from Genologics to pull necessary metadata to automate analysis and report generation.

Priority:P1

Contributors: Hugo, Ming, Nan

Additional Description: At a minimum includes sample names, case/control, sex, ethnicity, age as well as customer information for creating business elements reports. Additional information that is provided is contained in Christian's DB layout and Genologics documentation.

Definition of Done:

7.10 Pipeline shall provide the ability to cross-validate sequencing with array results – multiplatform support.

Priority:P1,VA

Contributors: Hugo, Dan, Jing

Additional Description:

1. On samples run at ILMN the array is run alongside the sequencing and we get the data back from them. Need specifics of the results returned.
2. Use for validating SNPs and Genotypes

3. Take array to call genotype and SV by looking at fluorescent intensity (CNV partition module from ILMN)
4. Get a report on CNV for array data and combine with sequencing data to see where they match and don't match to validate SVs.
5. On medically related SVs
6. Automate via Genome Studio command line interface – API
7. Results shall be summarized in the at-a-glance summary pages.

Definition of Done:

7.11 Provide the ability to work with CNVision with more than two algorithms.

Priority: P1, VA

Additional Description:: CNVision is an opensource tool that is designed to simplify predicting and analyzing CNVs, from running prediction algorithms and combining their results to visualizing the raw data and designing qPCR primers for confirmation.

Definition of Done:

7.12 The pipeline shall integrate a long read aligner for analysis of reads > 200 bp.

Priority:P1

Contributors: Hugo, Dan, Jing, Ming, Nan-before

Additional Description:

1. work to do informed by long read aligner investigation work
 - a. generate accuracy performance by using simulated data for both BWA and long read aligners.
 - b. select the best-of-breed long read aligners
2. integrate long read aligners into the pipeline
3. modify pipeline modules to adapt to new alignment results (e.g. new/missing annotations)

Definition of Done:

7.13 Apply HG20 patches to reference to improve alignment and variant calling by applying the latest changes.

Priority:P3

Contributors: Mark/Hugo

Additional Description: may be new think here

Definition of Done:

7.14 Include versioning information in headers of pipeline output files

Priority:P1

Contributors: Hugo, Dan

Additional Description: Adding pipeline version (which can infer the versions of different tools we used) to the headers of the output files.

Definition of Done:

7.15 Document a pipeline tool update strategy.

Priority:P1

Contributors:

Description

1. How frequently would we update, how, test process?
2. rev 6 month planned cycles?
3. maybe less or as needed?

Definition of Done:

7.16 Provide a more manageable way to access ref calls than the giant VCF file.

Priority: P1, VA

Additional Description: Create a BED file to represent all the reference regions. Replaces the need to go to the giant VCF file. Provided as an alternative to get reference calls from a more manageable sized file.

Definition of Done:

8 Build a repository of public genomes/exomes to calculate population frequencies, serve as controls, test our pipeline, and identify disease mechanisms.

Priority: P1

Contributors: Rong, Michael, Sarah, Gemma, Steve, Anil

Additional Description: see breakdown below

Definition of Done:

8.1 Usage is defined as:

- a. Calculate population frequency
 - i. SNPs
 - ii. indels
 - iii. SV
 - iv. CNVs
- b. Build the largest set of control genomes/exomes
- c. Test our pipeline and case/control discovery module
- d. Identify disease genes/mechanism to build stories for our product launch and conference presentations, such as AGBT.
- e. Determine how/if we can get breakpoint data for many of the SVs that are present.

8.2 Implement a scalable directory to host all public genomes/exomes data on our internal servers and define archiving process.

Priority:P1

Additional Description:

- a. /research/data/genome
- b. Sample tracking systems
 - i. Currently in dropbox
 - ii. Move into database in future

Definition of Done:

8.3 For UK10K: 6000 healthy genomes and 4000 exomes with extreme phenotypes

Priority:P1, VA

Additional Description: actual numbers are smaller considering the download rates and a rolling window of what we can analyze. We may be able to get VCF files.

- a. Healthy controls
 - i. 740 samples with 2,026 whole genome bam files (HighSeq 2000, GAI, 6× coverage) from Avon Longitudinal Study of Parents and Children. These families have a vast amount of genetic and environmental information over the years, which could potentially become the perfect case to convince VA to investigate genetic-environmental interaction. 15T disk space required.
 - ii. 1,713 samples with 6,004 whole genome bam files (HighSeq 2000, GAI, 6× coverage). UK's only twin registry of 11,000 identical and non-identical twins between the ages of 16 and 85 years. Most detailed clinical adult register in the world. Study the genetic and environmental aetiology of age-related complex traits and diseases. 48T disk space required
- b. Neurodevelopment Sample Sets (up to 3000 whole exomes)
- c. Obesity Sample Sets (2000 whole exomes)
- d. Rare Diseases Sample Sets (1000 whole exomes)
 - i. Severe Insulin Resistance
 - ii. Congenital muscular dystrophies and congenital myopathies.
 - iii. Ocular coloboma
 - iv. Congenital Heart Disease data set.
 - v. Ciliopathies
 - vi. Familial INtellectual Disability
 - vii. Congenital Hypothyroidism (CH) due either to dysgenesis or dyshormonogenesis; and patients with Resistance to Thyroid hormone (RTH)
 - viii. Familial Hypercholesterolemia

Definition of Done:

8.4 ESP_GO data (from dbGAP)

Priority:P1

Definition of Done:

<<Michael to update with details.>>

8.5 Already downloaded:

8.5.1 Korean genomes (40 whole genomes, high coverage)

Review how well covered these are.

Priority: P1, VA for providing the frequencies from these genomes as an annotation.

Will need to be put through the pipeline to achieve this.

Currently have 8 for V1 (P1). May process 2 more to get to 10. (P2)

8.5.2 1000 genomes (1092 samples, 14 ethnicities, 4 population categories, low coverage whole genome plus high coverage exomes). Need genders.

8.5.2.1 V3 now available – determine the frequencies for unrelated individuals.

Need to scope effort to remove related individuals from the dataset. Lower priority.

Priority: P2

8.5.2.2 Use the genomes as a control for Varimed report.

Priority: P1

Part of validating C/C Discovery. Use as part of Varimed report in a real world case.

8.5.3 Complete Genomics diversity panel (just variants (VCF files))

8.5.4 CEPH family panel (from Complete Genomics)

8.5.5 SRA genomes

8.6 Fine tune MySQL parameters and test for use with very large datasets.

Priority:P1

Definition of Done: will it index.

9 Curate proprietary databases with quality control

9.1 Improve paper selection process for Varimed curation to increase sensitivity, specificity, and to enable upto-dateness.

Priority:P1

Definition of Done:

9.1.1 Develop tools to build medline database with a bi-monthly update.

Priority:P1

Contributors: Anil, Rong

Additional Description:

- a. Apply and get the latest medline files with over 21 million abstracts with medical subject heading (Mesh), key words, title, and other 60 annotations
- b. Develop tools to build medline database periodically
- c. Database schema for nine tables
 - i. Citations
 - ii. citations_chemicals
 - iii. citations_databanks
 - iv. citations_keywords
 - v. citations_mesh_headings
 - vi. databanks
 - vii. journals
 - viii. authors
 - ix. chemicals

Definition of Done:

9.1.2 Build a tool to identify latest papers from medline and other resources for Varimed curation bi-monthly (Anil, Rong)

Priority:P1

Contributors: Anil, Rong

Additional Description:

- a. Build positive and negative sets of papers
- b. Identify lists of features, such as Mesh, keywords, abstracts, title, journal, to identify to separate the positive and negative lists of papers
- c. Build a tool to mine over 21 million abstracts in medline and identify full lists of papers on human genetic studies on disease/drug response/phenotypes
- d. Cover all papers from our competitors, such as GWAS catalog, 23 & Me, knome
- e. Purchase papers on several journals that contains a lot of genetic studies but we do not have access to the full text, including but not limited to
 - a. Schizophrenia research
 - b. Hypertension research : official journal of the Japanese Society of Hypertension

Definition of Done:

9.2 Develop process for targeted new content identification for our “medical disease panels” including HCM, longQT, and FH.

Priority:P1

Contributors: Anil, Rong
Additional Description:

This may include:

1. getting regular updates from our disease area experts
2. automated alerts sent to our curation teams
3. augmenting our current varimed article selection algorithm
4. quantitative assessment coverage of papers for panels.

Definition of Done:

9.3 Improve annotation and ability to assess accuracy (compared to other public resources) of medical variants in our proprietary databases

Priority:P1

Definition of Done:

9.3.1 Build a tool to identify dbSNP IDs and chromosome locations to

Priority:P1

Contributors: Steve, Rong

Additional Description:

- a. Identify dbSNP IDs and HGMD accession for MendelDB variants using HGVS in HGMD, including insertion and deletion
- b. Identify dbSNP IDs and HGMD accession for MendelDB variants using amino acid changes in HGMD and dbSNP137
- c. Identify non-standard formats in the DNA change /protein change columns in MendelDB and send feedback to Sarah and Gemma
- d. Evaluate the performance of dbSNP identification tool
- e. Identify the variants that failed the identification

Definition of Done:

9.3.2 Identify dbSNP IDs and chromosome locations for all medical variants in Varimed

Priority:P1

Definition of Done:

9.3.3 Identify dbSNP IDs and chromosome locations for all medical variants in MendelDB (This is required to annotate genomes using MendelDB)

Priority:P1

Definition of Done:

9.3.4 Map phenotypes in MendelDB to enable comparison with public databases, such as HGMD, OMIM

Priority:P1

Additional Description:

- a. Manually add OMIM ID into MendelDB gene tab by curators
- b. Map OMIM IDs to UMLS, Mesh, ICD IDs

Definition of Done:

9.4 For each Mendelian gene/disease, calculate the number of mutations of loss of function, and gain of functions as an additional way to assess pathogenicity of novel mutations.

Priority:P1

Additional Description: If all known Mendelian mutations are loss of functions, novel mutations with gain of function in this gene are unlikely to cause the disease.

Definition of Done:

10 Additional Annotation requirements

10.1 Enable additional annotations from public databases that fill current competitive gaps.

Priority: P1

Additional Description: sub items below.

Definition of Done:

10.1.1 Versioning all public databases individually and as a collection.

10.1.1.1 Develop a content release process (for example freeze date on content prior to release)

10.1.1.2 Add download dates and creation dates for all public databases in R & D

10.1.1.3 Version control on all databases in the production for every release using mysqldump

10.1.2 Documentation of all public databases

10.1.3 Annotate variants using dbSNP 137

10.1.3.1 Parse dbSNP 137 into database

10.1.3.2 Develop scripts to annotate dbSNP ID using dbSNP 137

10.1.3.2.1 Use chromosome, location, and allele to annotate variants with dbSNP IDs

10.1.3.2.2 How to annotate when there are multiple dbSNP IDs matched for the same variant? (not that common)

10.1.3.2.3 How to annotate when there are multiple variant types for the same dbSNP ID?

10.1.3.2.4 How to annotate when there are multiple genes associated with the same variant? For example, rs539 is associated with MEF2BNB and MEF2BNB-MEF2B. How to concatenate multiple rows of gene annotations into the same variant?

~~10.1.4 More public databases promised in the VA proposal~~

~~Priority: P1, VA~~

~~Contributors:~~

~~Additional Description:~~

~~<<to be filled in>>~~

~~Definition of Done:~~

10.1.5 Additional databases shall be added to the annotation engine.

Priority: P1, VA where noted

Contributors: Rong, Michael

Additional Description:

Databases to add include:

1. GNF Tissue Expression Body Atlas, NCI-60 Cell Line Expression Atlas from BioGPS: <http://biogps.org/#goto=welcome>

2. BIND: <http://www.ncbi.nlm.nih.gov/pubmed/12519993>
3. DIP (Database of Interacting Proteins): <http://dip.doe-mbi.ucla.edu/dip/Main.cgi>
4. MINT: <http://mint.bio.uniroma2.it/mint/Welcome.do>
5. VIRUSMINT: <http://mint.bio.uniroma2.it/virusmint/Welcome.do>
6. BioGRID (Biological General Repository for Interaction Datasets): <http://thebiogrid.org/>
7. TarBase: <http://diana.cslab.ece.ntua.gr/?sec=databases>
8. Regulome (VA)
9. HGMD (VA)
10. COSMIC (VA)
11. Most DBs will drop into the engine (figure a day or so)
12. Cancer Gene Census (VA)
13. CHOP CNV (VA)
14. GENCODE (VA)
15. Human Genome Deversity Cell Lin Panel (HGDP-CEPH) (VA)
16. Korean Personal Genome Project frequencies (Personalis generated) (VA)
17. UK10K frequencies (Personalis generated) (VA)

Definition of Done:

10.1.6 Annotation Launcher 2.0 (AnnoL) update.

Priority:P1, VA

Contributors: Michael, Rong

Additional Description: With the change to snpEff, databases formerly annotated by ANNOVAR need to be implemented in Annotation Wrapper.

Databases to add include:

1. Cytoband (UCSC)
2. Transcription factor binding sites (TRANSFAC) obtainable from UCSC
3. miRNA (miRBase): <http://www.mirbase.org/>
4. snoRNA (snoRNABase): <http://www-snorna.biotoul.fr/> (latest obtainable from UCSC)
5. Small RNA Targets (TargetScanS): <http://www.targetscan.org/>
6. EvoFold: <http://users.soe.ucsc.edu/~jsp/EvoFold/> obtainable from UCSC
7. DGV
8. SegDups (UCSC)
9. RepeatMasker (UCSC)
10. dbSNP137
11. dbSNP130
12. Complete Genomics frequencies

- a. Need to be pre-computed
 - b. create mysql table
 - c. Annotation
- 13. NHLBI GO-ESP Frequencies
 - a. Need to be pre-computed
 - b. create mysql table
 - c. Annotations
- 14. 1000 Genomes Frequencies
 - a. Need to be pre-computed (by superpop., subpop., and for all)
 - b. create mysql table
 - c. annotations

Definition of Done:

10.1.7 Integrate SNPEff into the pipeline as a replacement for ANNOVAR.

Priority: P1, VA

Contributors: Michael, Hugo, Rong

Additional Description: Integrating SNPEff into the pipeline and passing the annotated data to annotation wrapper which will annotate with the missing databases.

Definition of Done:

Comparison with ANNOVAR results in progress

Additional snpEff annotations:

- 1. dbNSFP
- 2. GWAS Catalog

Add ability to import snpEff results into MySQL and integrate with general annotation reports from Annotation Wrapper.

10.1.8 Pipeline shall convert annotation engine raw output into the report designs

Priority: P1, VA

Additional Description: Annotation input is a tab delimited file (raw data)

Definition of Done:

10.2 Support for splicing in annotation.

Priority: P2

Additional Description:

Mutation taster may have built in capabilities to deal with this already.

SNPEff also provides some of these capabilities. Need a gap analysis on if this is sufficient.

Definition of Done:

10.3 Include versioning information in headers of annotation and report output files

Priority: P1, VA

Contributors: Hugo, Dan

Additional Description: Adding a version to the headers of the output files.

Definition of Done:

10.4 Document a database tool update strategy.

Priority: P1, VA

Description

4. Database “container”
 - a. Public DBs
 - b. Proprietary DBs
5. How frequently would we update, how, test process?
6. rev 6 month planned cycles?
7. maybe less or as needed.

Definition of Done:

10.5 Create a tracking board for databases and when they update to plan revision cycles.

Priority: V1

Additional Description: Prioritize the important databases to stay current with.

Repurpose licensing spreadsheet list.

Definition of Done:

11 Case/control discovery module

11.1 Provide additional collapsing methods to complete

Priority: P2

Contributors: Alex, Anil, Rong

Additional Description:

Believe that Madsen-Browning is best in class (already implemented), but we acknowledge that some bioinformaticians may prefer their “favorite” algorithm. Also enables additional choices for analyzing data sets to see if one can get better results.

Definition of Done:

11.1.1 Develop R libraries for CMC collapsing methods

Priority: P3

Additional Description:

Common (frequency > 1%) variants are thrown out. Equal weighting between all variants. Only for Mendelian diseases.

Definition of Done:

11.1.2 Develop R libraries for CAST collapsing methods

Priority: P3

Additional Description:

Very different than Madsen-Browning. Based on linear regression.

Definition of Done:

11.2 Complete Varimed report implementation

Priority: P1

Contributors: Rong, Anil

Additional Description: see subs below

Definition of Done:

11.2.1 Add Varimed to collapse based on allele direction in gene and pathway reports

Priority: P1

Additional Description:

Use Varimed to decide direction. Possibly a large differentiator.

Definition of Done: for testing – collapsing with and without Varimed and see if you get better biological results.

11.2.2 Calculate disease risks for all individuals in the case/control groups and 1000 genomes with matched ethnicities

Priority: P1

Contributors: Anil, Rong

Definition of Done:

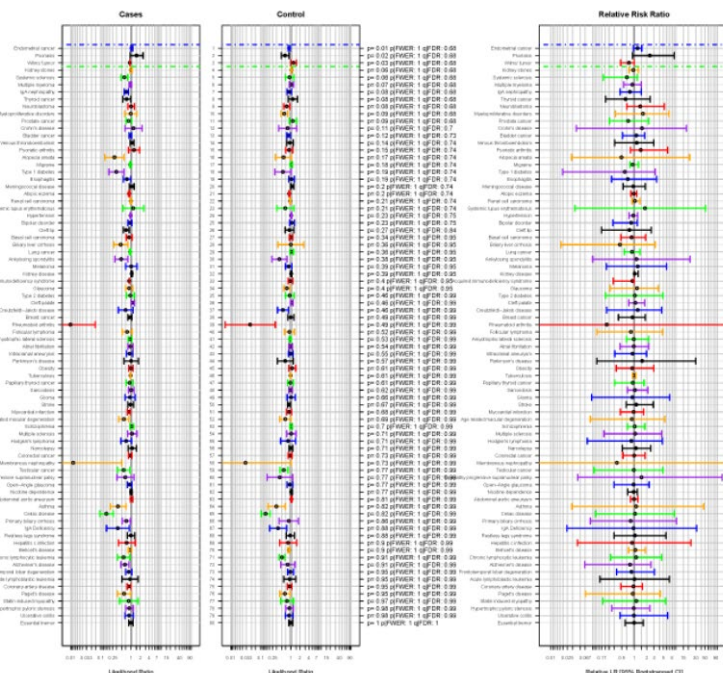
11.2.3 Population-level graphics across all diseases (Forest, circos plot)

Priority: P1

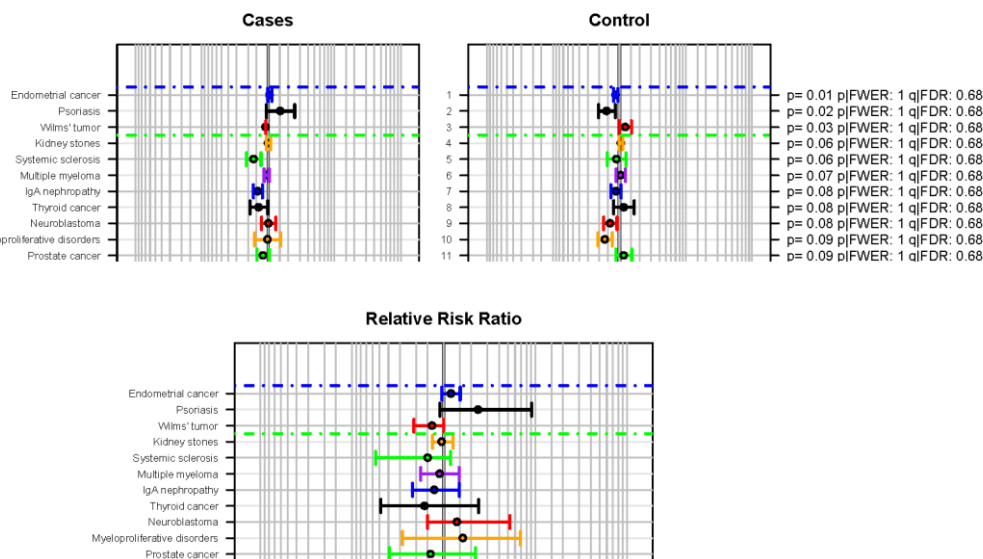
Contributors: Anil, Rong

Definition of Done:

Example Forest Plot



Example Forest Plot, Blowup



11.2.4 Calculate effect sizes in case-control studies method of ranking by disease, pathway and variant.

Priority:P1

Contributors: Anil, Rong

Definition of Done:

11.2.5 RiskOGRAM across all complex diseases

Priority:P1

Contributors: Anil, Rong

Definition of Done:

11.2.6 Graphics displaying significant or top 16 diseases

Priority:P1

Contributors: Anil, Rong

Definition of Done:

11.3 Complete PharmGKB pathway report implementation

Priority: P1

Contributors: Rong, Anil

Definition of Done:

11.3.1 Power analysis to recommend sample sizes needed to detect significant genes/gene set/pathway/disease

Priority:P3 depends on LOE to become P2

Contributors: Anil, Rong

Definition of Done:

12 Database accuracy study

12.1 Compare Mendelian variant databases to commonly used databases at variant, gene, and phenotype levels to substantiate claim that we have best in class mendelian database.

Priority: P1, VA(?)

Definition of Done:

12.1.1 Points of comparison shall include

12.1.1.1 Comparing Accuracy (for example: for the same variants in multiple dbs ours is more accurate in terms of pathogenicity, <<fill in more here>>, as measured by <<fill in more here>>)

Priority:P1

Additional Description:

- a. risk allele frequencies

- b. Allele frequencies likely not enough to catch all errors – need alternatives.

Definition of Done:

12.1.1.2 Depth/detail for each variant (for example: number and type of columns in db)

Priority:P1

Definition of Done:

12.1.1.3 Comprehensiveness or coverage (for example: we cover more genes, variants, diseases than other dbs)

Priority:P1

Additional Description:

1. Sheer variants covered
2. Gaps in coverage of particular content areas
3. Create venn diagram

Definition of Done:

12.1.1.4 Determine the “Upto-dateness” of databases

Priority:P1

Definition of Done:

12.1.2 Build a comprehensive Mendelian variant database from all sources, including MendelDB, GTR (genetic test registry) with test, GTR without test, HGMD Mendelian, HGMD others, OMIM Mendelian, OMIM QTL, OMIM Others

Priority:P1

Definition of Done:

12.1.3 Compare Mendelian variant databases from all sources at variant level

Priority:P1

Definition of Done:

12.1.4 Compare Mendelian variant databases from all sources at gene level

Priority:P1

Definition of Done:

12.1.5 Compare Mendelian variant databases from all sources at phenotype level

Priority:P1

Definition of Done:

12.2 Database wide - identify errors in public databases

Priority:P1

Definition of Done:

Dependency: 12.3 (and others?)

Need criteria for identifying errors.

12.3 Compare the number of variants called in "medically relevant genes" and the number of "medically relevant variants" from multiple genomes (including AMY)

Priority:P1

Additional Description:

- a. Using OMIM/HGMD
- b. Using our larger set of Mendelian databases + frequency/susceptibility filters + MendelDB.
- c. Determine how many we get out with our process of manual review

Definition of Done:

12.4 Follow-up on the variants that fell out in the Personalis analysis

Priority:P1

Additional Description:

- a. Report how many variants changed based on manual curation vs. frequency vs. susceptibility filters.
- b. Perform a similar analysis for Varimed using public databases (gwas catalog) as the comparison.
- a. Alex: John Ioannidis – OMIM profiles – a way of showing importance of timeliness of data.

Definition of Done:

12.5 Determine how controlled the vocabularies are between our databases

Priority:P1

Additional Description:

- a. Fill in OMIM ID for MendelDB
- b. level of connection and interoperability between databases
- c. using mendelDB as a “gold standard” – where it overlaps with hgmd and others. Compare in a sampled test.

Definition of Done:

12.6 Perform a curation check by having MendelDB curators, curate Varimed

Priority:P2

Additional Description:

- a. Varimed quality assessment program
- b. Use part time content scientists to reduplicate varimed work for a month to assess errors/accuracy, etc.
- c. "Search terms" across databases as an example of searching for information (lack of standardization, changing standards,...)

Definition of Done:

12.7 Test end to end with replicate data

Priority:P2

Additional Description:

- a. A histogram of number of missed Mendelian variants across 1,094 individuals from 1000 genome project if using public reference genome sequence. We can create a histogram using Mendelian variants from MendelDB, HGMD, and OMIM individually.
- b. A histogram of number of incorrect Mendelian variants across 1,094 individuals if using HGMD, OMIM. We can define incorrect Mendelian variants using $\text{freq} > 0.35$, OMIM phenotypes with susceptibility.

Definition of Done:

12.8 Translate comparisons into marketing content.

Priority:P1

Additional Description: why we're better.

13 Reports Requirements

<<additional revising and details ongoing in this section, these are not set yet>>

13.1 Refine graphics such that we end up with attractive publishable graphs

Priority:P1

Contributors: Hugo, Dan, Ming, Anil, Rong, Steve, Michael, others, consultant?

Additional Description:

1. Take outputs to next step of how end user would use them – the common sense 90% of things people want to see
2. Interfacing/compatibility with 3rd party tools – at the file level?
3. Generate the Circos plot as a summary?
4. A consultant to generate graphics from the data?

Definition of Done:

13.2 Reporting capabilities shall support representing summary and quality information for studies where 100s of samples are present.

Priority:P1

Additional Description:

This is a more likely scenario than a few samples to view numerically in detail.

Definition of Done:

13.3 Produce a top-level page that includes the major biological insights/findings as the front page so the customer sees the page first.

Priority:P1

Definition of Done:

13.3.1 The top level page shall include a section on the top variants discovered to highlight significant variants for the customer.

Priority:P1

Definition of Done:

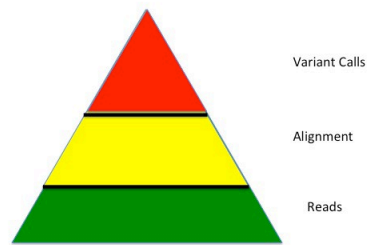
13.3.1.1 For top variants discovered a “summary” triangle may be produced that gives a snap-shot of the underlying quality behind each read.

Priority:P1

Additional Description:

The idea is for the triangle to be split into good (green), questionable (yellow), and bad (red) color codings with variants stacked on top of alignment stacked on top of reads quality. For example, triangles might look like this:





Definition of Done:

13.3.2 Create sample-specific variant report for each of top 20 genes in the enriched gene report

Priority:P1

Definition of Done:

13.3.3 Create sample-specific variant report for each of top 20 pathways in the enriched pathway report

Priority:P1

Definition of Done:

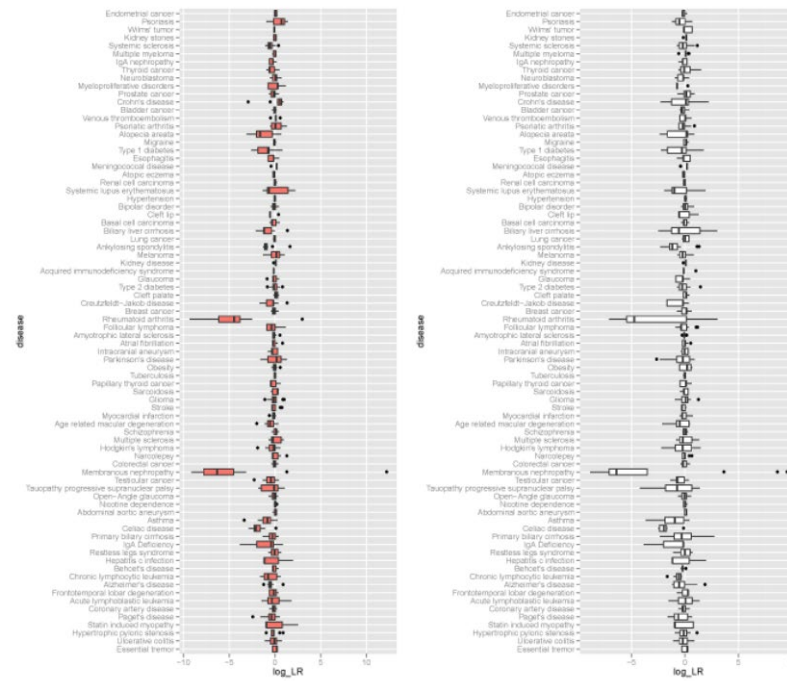
13.3.4 Include Risk-o-Grams for top identified diseases

Priority:P1

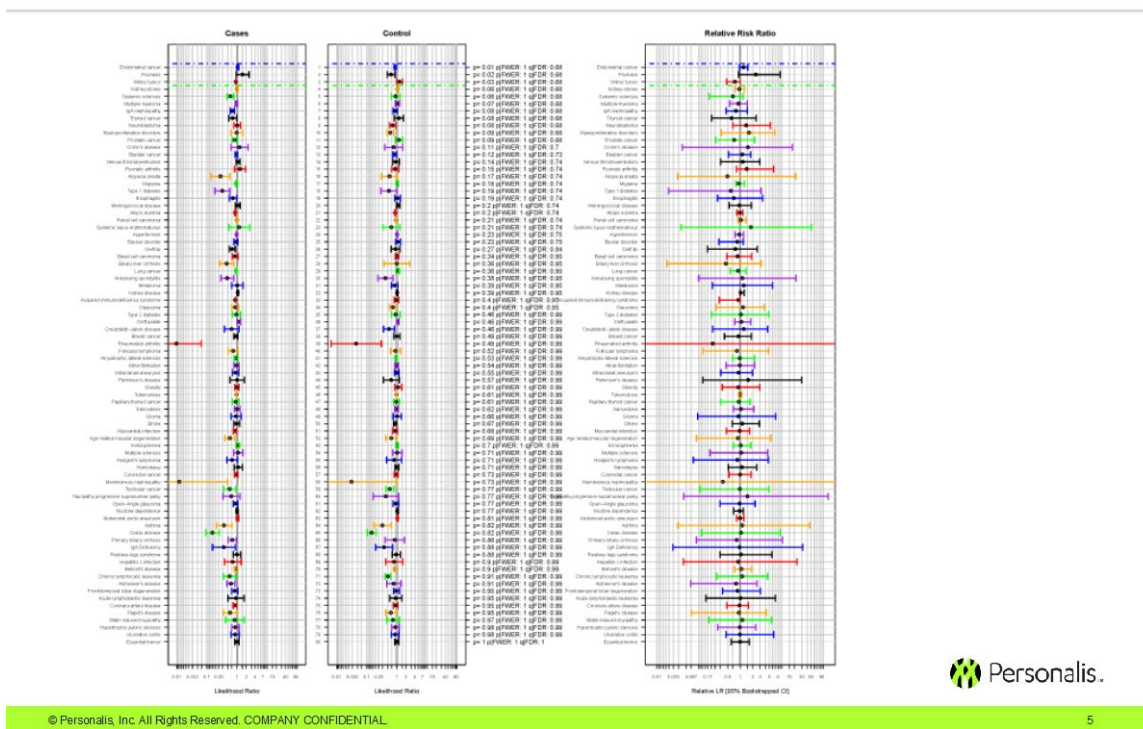
Additional Description:

- a. RiskOGRAM for individual disease
- b. Pathway, such as PharmGKB pathway
- c. Combined RiskOGRAM across all complex diseases

Example Box Plot



Example Forest Plot



Definition of Done:

13.3.5 Demographic tables may be shown that summarize the makeup of the samples in the study.

Priority:P1

Definition of Done:

13.3.6 Integrate links to all reports for case/control to simplify access

Priority:P1

Additional Description:

- variant annotation report
- rare functional variant reports
- enriched gene reports
- enriched pathway reports
- enriched PharmGKB pathway reports
- enriched MSigDB gene set reports
- recessive, dominant, and additive risk models
- varimred report

13.3.7 Integrate links to QC metrics and stats pages.

Priority:P1

Definition of Done:

13.4 QC metrics and stats page changes to improve ability to view the information.

13.4.1 QC metrics and stats shall be separated into different sections to simplify viewing them (they are used for different things).

Move the “at-a-glance” page to a separate QC section.

QC metrics are a subset of items used to diagnose problems whereas other metrics are simply statistics about the particular run.

QC R+A
QC VAR
Stats R+A
Stats Var

13.4.2 QC metrics and stats will be shown in aggregate to better summarize the information presented to the user.

13.4.2.1 A graphical representation for “all samples” shall be included to summarize the data.

13.4.2.1.1 % or # passed

13.4.2.1.2 % or # needing review

13.4.2.1.3 % or # excluded from further analysis (for various reasons to be detailed)

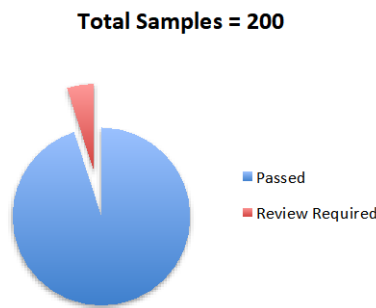
13.4.2.1.4 % or # included in final analysis.

Project ID	0081283
Project Name	VA
Project Desc	Batch of 200 samples from VA project
Samples Processed	200
Sequencing Platform	Illumina HiSeq
Date	09-12-2012

QC Sample Overview

A total of 200 samples were processed with 195 passing QC. 5 of the samples were reviewed in detail for possible quality issues, with 2 ultimately excluded from downstream analysis based on discussion with the customer.

	#	%
Passed	190	95%
Review Required	5	2.5%



13.4.2.2 “Normal ranges” for metrics will be provided to serve as a point of comparison for the customer to better evaluate their data.

Priority: P1, VA

Additional Description:

- Determines the good, questionable, bad categories for each metric and/or stat.
- Could be “normal ranges” (aka “acceptable ranges”) and/or “study ranges”

- An “acceptable” or “normal” range is within Personalis norms or a range of acceptability such that analysis continues. “Study ranges” are the ranges across the samples provided by the customer.

Definition of Done:

13.4.3 Per sample QC metrics and stats details shall be graphically shown to allow the user to examine the data per sample graphically.

Priority: P1

Additional Description:

Could be broken out by sample in a graphical, heatmap like “Master matrix”.

Samples retain their color coding (good, questionable, bad)

For example:

Project ID	0081283
Project Name	VA
Project Desc	Batch of 200 samples from VA project
Samples Processed	200
Sequencing Platform	Illumina HiSeq
Date	09-12-2012

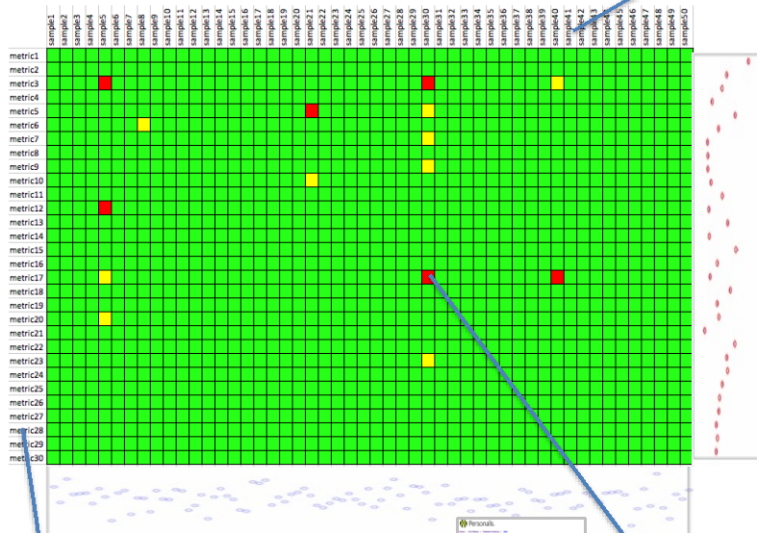
QC Sample Summary

Filter:

All

Passed

Review Req



Show Current Table

Mouse over or click to get explanation of metric (and normal range)

Mouse over or click for actual value (and normal range)

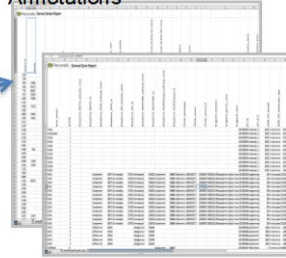
Definition of Done:

13.4.4 All of the per sample numbers behind the graphical “heat map” representations are available as “downloadable” tables, but are not shown to the user in results pages because there is a tremendous amount of statistical information.

Customer Summary Report
At-A-Glance Statistics & QC Report



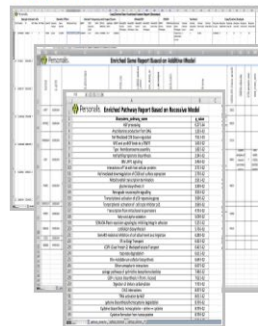
General Variant Report with Annotations



Detailed Read, Align & Var Statistics & QC Reports



Case-Control Analysis Reports



13.4.4.1 Filtering mechanisms may be developed to provide a means to display the matrix in different ways. For example, only showing failures or particular metrics.

13.5 Provide standard graphics that the customer expects to see (and are useful).

Priority: P1

Additional Description:: <<needs more detail>>

Definition of Done:

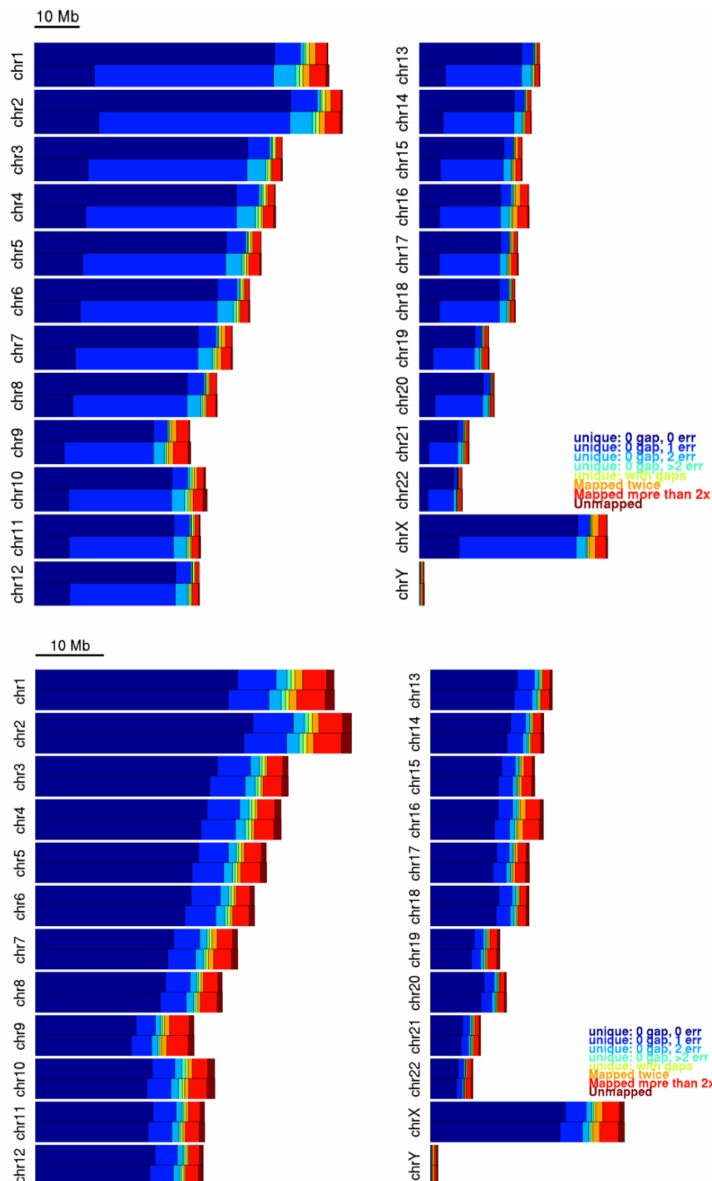
13.6 Provide novel, actionable, and publishable graphics summarizing QC stats.

13.6.1 BAM QC graphics and stats incorporation in PGS pipeline outputs (see Jira)(Jason, Dan)

Priority: P1

Definition of Done:

13.6.1.1 Include the Read1/Read2 “rainbow plots”:



13.6.2 For remaining summary statistics pages include improvements detailed below as appropriate. These currently reside in Jira and are based on the initial prototype report design.

Priority: P1

Definition of Done:

13.6.2.1 Jira entries for initial prototype design:

Issue Type	Key	Summary	Priority
Improvement	PSP-896	Improvements to Report Statistics (tables)	N/A
Sub-task	PSP-897	PSP-896 Put standard deviation along side the mean for average mapping qual and insert size (quick)	1

Sub-task	PSP-898	PSP-896 Align number to the right unless spanning multiple columns (quick)	1
Sub-task	PSP-899	PSP-896 Add average read length (low)	1
Sub-task	PSP-900	PSP-896 Add soft clipping read lengths (low)	3
Sub-task	PSP-901	PSP-896 Add percentages (high, quick)	1
Sub-task	PSP-902	PSP-896 Make top table narrower in single-sample level file (quick)	3
Sub-task	PSP-903	PSP-896 Remove gapped alignment metric from BWA (quick)	1
Sub-task	PSP-904	PSP-896 Remove, or rename FailQC to Failed BWA QC (quick)	1
Sub-task	PSP-905	PSP-896 Derive mismatch rate based (and expected mismatch rate) for mapped reads (low)	3
Sub-task	PSP-906	PSP-896 Replace duplication rate with diversity score (what is diversity score??) (low)	3
Sub-task	PSP-907	PSP-896 Report list of contaminants that were queried for in the "overrepresented sequences" table (low)	1
Sub-task	PSP-908	PSP-896 Callable bases (high priority, but dependent upon accuracy team)	1
Sub-task	PSP-909	PSP-896 Provide interpretation of "mapped to other chromosome", indicating how many of these reads represent real translocations (low)	3
Sub-task	PSP-910	PSP-896 Remove kmer content table unless compelling reason to keep (low)	3
Improvement	PSP-911	Global changes to graphs in summary reports	N/A
Sub-task	PSP-912	PSP-911 Add legends where appropriate (quick)	1
Sub-task	PSP-913	PSP-911 Bigger Fonts (low)	1
Sub-task	PSP-914	PSP-911 Create both small and hi-res images, where hi-res images are publication-ready (low)	2
Sub-task	PSP-915	PSP-911 Improve appearance of small images (low)	2
Sub-task	PSP-916	PSP-911 Reorder graphs (by accessibility, by importance?) (quick)	1
Improvement	PSP-917	Graphs to add to summary report	N/A
Sub-task	PSP-918	PSP-917 Coverage \[probably as a cumulative graph\] (high)	1
Sub-task	PSP-919	PSP-917 Coverage in regions of interest (low)	3
Sub-task	PSP-920	PSP-917 Coverage vs. GC (check out GC bias tools, possibly use empirical GC from genome) (med)	2
Sub-task	PSP-921	PSP-917 Graph of "usable" lengths (not sure what this is, possible for use in interpreting phasing efficiency) (low)	?

Sub-task	PSP-922	PSP-917 Cumulative graph of insert size (low to high and high to low) on log scale. Alternatively, just report percentiles. (med)	2
Sub-task	PSP-923	PSP-917 Cumulative graph of quality scores (high to low) (med)	1
Sub-task	PSP-924	PSP-917 Quality scores as heatmap? (low)	3
Sub-task	PSP-925	PSP-917 Cumulative quality score from gatk recal (low)	3
Improvement	PSP-926	Add xy grid to Per Base Sequence quality graph	1
Improvement	PSP-927	Combine Per Base N content with Per Base Sequence in summary report graphs	1
Improvement	PSP-928	Populate demo graphs with real data	1
Improvement	PSP-929	Add line for empirical GC content (sliding window across genome) to per sequence GC content graph (low)	3
Improvement	PSP-930	Duplication level graph: Add line for "expected" (based on sampling reads from the genome / real datasets) (low)	3
Improvement	PSP-931	Duplication level graph: Log scale (discussion with MJ suggests not worthwhile) (quick)	?
Improvement	PSP-932	Duplication level graph: remove "1" (quick)	1
Improvement	PSP-933	Per base GC content graph: o 0-100 y axis, 0-60 axis, or y-axis different for each dataset?	1
Improvement	PSP-934	Per Sequence Quality Scores graph: make graph cumulative (starting at highest quality value and summing as quality drops) (low)	3
Improvement	PSP-935	mapping quality score graph: Leave out 255 or place elsewhere in plot (low)	3
Improvement	PSP-936	Remove read length plot from lane-level report page (unless replace with soft clipping graph) (low)	3
Improvement	PSP-937	Include Personalis ranges in relevant graphs	1
Improvement	PSP-938	• Include customer submission ranges in relevant graphs (i.e. if customer submitted 100 samples, graph ranges for these samples)	1
Improvement	PSP-939	In sample-level pages, output Pass/Warn?/Fail for lane-level metrics	1
Improvement	PSP-940	Integrate report generation with LIMS-like system to populate information from database	1
Improvement	PSP-941	Add contamination detection capabilities (cross-sample and contamination from other organisms \[eukaryotic / bacterial / viral\])	3
Improvement	PSP-942	Provide full explanation of each statistic - also provide context for how it is used. (ie. what you're trying to get out of that metric)	1

Improvement	PSP-943	Optimize the report layout	1
Improvement	PSP-944	• Settle on logical hierarchy of alignment statistics (i.e. which statistics are broken down into sub-categories that sum to the first statistic)	1
Improvement	PSP-945	develop style standards for reports	1
Improvement	PSP-946	Include Run Time Analysis statistic from LIMS system, or derive some of these statistics from fastq read names	3
Bug	PSP-947	Base recalibration graph currently limited to chr1 data (aggregate over all chromosomes) (med)	1
Bug	PSP-948	Duplication level stats currently limited to chr1 data (aggregate over all chromosome) (med)	1
Bug	PSP-949	Can't currently calculate coverage if have multiple read lengths in lane-level bam files. Ask accuracy team (Jason) to build into bamqc?	3
Improvement	PSP-950	Settle on a form of delivery for the report. It should be easy to edit and include presentation quality figures and tables.	1
Improvement	PSP-951	How will custom reference be presented (converted to reference genome alleles, or compared against a separate ref genome run?)	1
Improvement	PSP-952	enhance graphs with informative comparisons.	1
Improvement	PSP-953	Customer summary report improvements	n/a
Sub-task	PSP-954	PSP-953 separate customer and instution	1
Sub-task	PSP-955	PSP-953 Move receipt date of sample(s) to be next to the start and end date	1
Sub-task	PSP-956	PSP-953 List if a sample is a case or a control.	1
Sub-task	PSP-957	PSP-953 Push version numbers to a seperate page.	1
Sub-task	PSP-958	PSP-953 Provide a broader description on what the elements of the analysis are.	1
Sub-task	PSP-959	PSP-953 Include what "optional" elements of an analysis were run - tie to the line items of the PO	1
Sub-task	PSP-960	PSP-953 With samples include if the passed incoming QC (y/n)	1
Sub-task	PSP-961	PSP-953 For sample info seperate out what the customer reported to us and what we determined.	1
Improvement	PSP-962	Figure out how to aggregate multiple samples on to graphs	2
Improvement	PSP-963	Determine how we will navigate report when there are 100s of samples.	1
Improvement	PSP-964	When statistics reported by any page include a +- standard deviation, change the statistic title to indicate that the +- value is standard devotion	2

Improvement	PSP-965	For duplication rate reporting, include expected duplication rate based on coverage	2
Improvement	PSP-966	Indicate that average base quality is calculated prior to recalibration (can grab before and after numbers from new GATK-lite output)	1
Improvement	PSP-967	Diagnostic tool for insert size as relates to coverage???	3
Improvement	PSP-968	When comparing to Personalis Normal ranges, avoid absolute numbers, like sequence total. Better to compare to normalized numbers or to values such as "sequences per run"	1
Improvement	PSP-969	When comparing to Personalis Normal ranges, better to replace vague metrics with metrics that have direct, diagnostic interpretability (e.g. "expected cluster density")	1
Improvement	PSP-970	Where relevant, instead of plus or minus, give a "target" i.e. should be higher than, or should be lower than	1
Improvement	PSP-971	Coverage --- should talk about determinants of coverage -- GC content, etc.	1
Improvement	PSP-972	Determine how EACH metric is to be presented, visualized, and compared to a Personalis Range (mean +- std, or maybe a hard cutoff)	1
Improvement	PSP-973	need a cross-customer range (aggregate statistics from the customers submission), since these ranges may be a more relevant comparison point than the Personalis ranges	1
Improvement	PSP-974	Improve linking of report and case/control	1
Improvement	PSP-975	Determine how visualizations (e.g. IGV plots) could be put into the reports.	1

14 Performance:

14.1 Performance profiling/benchmarking

14.1.1 Conduct performance profiling of the workflow (align, call, annotate, ...) to benchmark the performance of the system in end to end usage.

Priority: P1, VA

Additional Description: What is the performance of each step in a particular configuration with a defined data set.

Definition of Done:

14.1.2 Conduct per server load tests to benchmark the performance of the system on defined hardware.

Priority: P1, VA

Contributors:

Additional Description: How many WGs and exomes can be run on a single R910 and what is the impact on the performance profile?

Definition of Done:

14.1.3 Conduct configuration testing (testing performance profile with configuration changes) to benchmark the performance of the system in a defined configuration.

Priority:P1, VA

Contributors:

Additional Description: How do configuration changes impact the performance profile?

Definition of Done:

14.1.4 Determining optimal size for input FASTQ files (chunk size)

Priority:P1, VA

Definition of Done:

14.2 Throughput measurements to collect:

Priority:P1, VA

Definition of Done:

14.2.1 Measured WGS per 910 throughput per day in set configurations

14.2.2 Measured Exome per 910 throughput per day in set configurations

14.2.3 Measured Exome+ per 910 throughput per day in set configurations

14.3 Desired throughput characteristics:

14.3.1 Capability to run 30 genomes/day by Q1 2013

14.3.2 Capability to run 100 genomes/day by Q2 2013

14.3.3 What are the desired throughput numbers for exomes?

14.4 Calculate data footprint measurements for defined genome and exome data sets to better understand space requirements for storage system.

Priority: P1,VA

Definition of Done:

15 Additional testing requirements

15.1 A production release process shall be defined for releasing builds of software.

Priority: P1, VA

Definition of Done:

15.2 Performance tests shall be run on the pipeline to inform throughput calculations and requirements for performance improvements.

Priority: P1, VA

Contributors: Hugo, Ming, Nan, Dan

Description and Definition of Done:

- Testing the performance of the pipeline on both single- and multiple-sample WGS and WES
 - Measuring metrics run time and memory usage in different stages of the pipeline running on a single server and across multiple server:
 - alignment
 - sorting/binning
 - cleaning
 - variant calling
 - annotation (public)
1. Whole genome sequencing (~40x) test run
 - a. Single server testing
 - i. Single-sample
 - Hyperthreading with local storage

- Hyperthreading with NFS storage
 - Non-Hyperthreading with local storage
 - Non-Hyperthreading with NFS storage
 - b. Multiple server testing (4 servers)
 - i. Single-sample
 - HT/Non-HT with NFS storage, depending on single server performance results
 - ii. Multi-sample (10 samples)
 - HT/Non-HT with NFS storage, depending on single server performance results
- 2. Whole exome sequencing (~70x) test run
 - a. Single server testing
 - i. Single-sample
 - HT/Non-HT with local/NFS storage, depending on WGS performance results
 - b. Multiple server testing (4 servers)
 - i. Single-sample
 - HT/Non-HT with NFS storage, depending on WGS performance results
 - ii. Multiple sample (10 samples)
 - HT/Non-HT with NFS storage, depending on WGS performance results
- 3. Summarizing test results to inform throughput calculations and requirements for performance improvements

15.3 A QC testing framework shall be built to support testing, defect tracking, and resolution.

Priority: P1, VA

Contributors: Ming

Definition of Done:

15.4 Design a case/control studies and test the PGS from alignment to case/control reports

Priority: P1, VA

Definition of Done:

15.5 Test automation?

Priority: P2

Definition of Done:

16 Additional Accuracy Related Features

16.1 Provide a corrected reference that improves alignment and variant calling.

Priority: P1, VA

16.1.1 Provide a non-ethnic major allele reference (NEMAR) “corrected reference” that improves alignment and variant calling.

Priority: P1, VA (option to discuss with VA)

Additional Description: (from Rong’s NEMAR slides)

Build NEMAR reference genome sequence using 1000 genome project and evaluate its Impact on medical variants

- Build a Personalis reference genome sequence using 1,092 whole genome sequences across diverse sub-population groups from 1000 genome project
- Investigate the impact of Personalis reference genome sequences from several different methods
- Build a comprehensive database of interpretable medical variants from all proprietary and public resources (including 559,229 variants from 10 proprietary and public resources)
- Investigate the impact on the interpretable medical variants using Personalis reference genome sequences from different methods

P1: NEMAR v1.0 released

- 1) Use major alleles that are consensus across European, African, Admixed American, East Asian sub-populations
- 2) Replace 1,121,131 SNPs, including 34,731 medical variants

P2: Four ethnicity-specific reference genome sequences

- 1) European: replace 1,832,848 SNPs, including 62,659 medical variants
- 2) African: replace 1,994,105 SNPs, including 65,388 medical variants
- 3) Admixed American: replace 1,834,260 SNPs, including 62,700 medical variants
- 4) East Asian: replace 1,995,556 SNPs, including 68,029 medical variants

P3: Four ethnicity-specific reference genome sequences, including non-SNP variants

- 1) European: replace 2,004,245 variants, including 125,928 non-SNP variants
- 2) African: replace 2,058,557 variants, including 130,379 non-SNP variants
- 3) Admixed American: replace 2,005,245 variants, including 126,573 non-SNP variants
- 4) East Asian: replace 2,183,034 variants, including 138,544 non-SNP variants

Definition of Done: Testing includes generation of statistics on NEMAR for before/after comparisons and testing w/ a/b tools to determine utility of NEMAR analysis.

16.1.2 Apply HG20 patches to reference to improve alignment and variant calling by applying the latest changes.

Priority:P3

Contributors: Mark/Hugo

Additional Description: may be new think here, possibly apply to NEMAR reference itself?
Definition of Done:

16.1.3 Provide a custom “personalized” reference that improves alignment and variant calling.

Priority: P3

Additional Description:

Definition of Done: Testing includes generation of statistics on for before/after comparisons and testing w/ a/b tools to determine utility of analysis.

16.2 Improve A/B comparison tool to enable additional accuracy assessments for indels, SV, and CNV and improve our ability to detect problem areas.

Priority: P1, VA

Definition of Done:

16.2.1 Indels

Priority: P1

Definition of Done:

16.2.2 SV

Priority: P1

Definition of Done:

16.2.3 CNV

Priority: P1

Definition of Done:

16.2.4 Develop a genome to use as a standard for A/B comparisons.

16.2.4.1 Complete the “bronze genome” to use as a standard comparison genome.

Priority: P1

Additional Description: This genome becomes the “B” that we compare everything to. “Bronze genome” currently includes two Broad, CG data, array data, and Personalis generated data.

Definition of Done:

16.2.4.2 Complete the “Silver genome” to use as a standard comparison genome.

Priority: P2

Additional Description: Will additionally include the 2x250 WGS and WES data.

Definition of Done:

16.2.4.3 Complete the “Gold genome” to use as a standard comparison genome.

Priority: P3

Additional Description: Will include “silver” and additional inheritance and phasing from entire CEPH pedigree. (Michael/Rong to help with phasing?)
Definition of Done:

16.3 Improve A/B comparison by rebuilding error classification using V3, gel free, pcr free modern data so we’re working with the latest data a customer would be using.

Priority: P2

Definition of Done:

16.4 Provide the ability to flag QC metrics that are outside of normal (or “acceptable”) ranges as potentially problematic and actionable.

Priority: P1, VA

Definition of Done:

16.5 Generate normal ranges (as defined in QC pages) for WGS, WES, and Exome + using V3 chemistry generated data so we’re comparing to numbers generated using the latest chemistry.

Priority: P1, VA

Definition of Done:

16.5.1 Generate normal ranges with and without using NEMAR reference.

Priority: P1, VA

Definition of Done:

16.5.2 Generate normal ranges from a diverse panel of genomes

Priority: P1, VA (to be determined)

Definition of Done:

16.6 Provide detailed explanations for how to act on flagged metrics.

Priority: P1

Definition of Done:

16.7 Correct genotypes in homopolymer flank regions to improve accuracy of results.

Priority: P1, VA

Additional Description: includes upstream-only base call around homopolymers.

Definition of Done:

16.7.1 A homopolymer flank genotyping algorithm shall be implemented and tested

Priority: P1, VA

Definition of Done:

16.7.2 Assess impact of homopolymer flank genotyping algorithm on exome and medically relevant regions.

Priority: P1

Definition of Done:

16.8 Provide a framework and method of error propagation into complex disease and mendelian analysis.

Priority: P2

Contributors: Mark, Alex M., Gabor, Jason

Definition of Done:

16.8.1 Translation to a series of cutoffs (for example three tier system: green, yellow, red)

Priority: P2

Definition of Done:

16.8.2 Rework the categories for classifying issues (next-generation of flagging)

Priority: P2

Definition of Done:

16.8.3 Provide estimates of reproducibility, probability, and confidence.

Priority: P2

Additional Description:

- a. Monte carlo analysis of analytical uncertainty for risk loci.
- b. Test all genotypes against BAM evidence
- c. Possible BAM crawler work.

Definition of Done:

16.8.4 Pass these values through and make information available to downstream interpretation

Priority: P2

Definition of Done:

16.9 Provide a mechanism for quantitate recalibration of quality and new and improved (more specific) error flags

Priority: P3

Definition of Done:

16.9.1 Provide additional quality metrics to prioritize the variants.

Priority: P3

Definition of Done:

16.9.2 Evaluate the utility of the Personalis quality metrics vs. what is available out of the box by VQSR (in GATK)(or others).

Priority: P3

Definition of Done:

16.9.3 The problematic regions accuracy flags shall be tested to determine the correctness of the flags.

Priority: P3

Definition of Done:

16.9.4 Raw read error model.

Priority: P3

Definition of Done:

16.9.5 Exon based coverage stats database.

Priority: P3

Definition of Done:

16.10 Provide the ability to follow SVs through pedigrees to improve detection rates and ability to characterize.

Priority:

Additional Description: single base indels could be a good place to start.

Definition of Done:

16.11 Investigate moving toward a contig based system rather than chromosomal based alignment system.

Priority:

Additional Description:

Definition of Done:

16.12 Create a database of variants detected from samples we've run (or had run) to enable quality checks on variants

Priority: P2

Additional Description: includes storing alleles and tracking which alleles have been seen at each position (for all levels of quality). Could be especially useful for SVs.

Dependency in SV section.

<<needs to be detailed out more>>

Definition of Done:

17 Commercial

17.1 Content that describes the functionality of PGS shall be provided.

Priority:P1

Additional Description:

There is a balance in what is to be detailed, but at a minimum include details of the pipeline analysis methods for alignment, aggregation, and variant calling, annotation and the databases used, case/control reports and determinations

of high priority variants in the reports. Distinguishing features shall be highlighted.

Definition of Done:

- 17.2 To determine possible throughput scenarios, estimates of the expected throughput of the system broken down by the number of genomes or exomes that can be analyzed (alignment through annotation) in a given period of time (per day, per week) per R 910 system shall be provided.**

Priority:P1

Definition of Done:

- 17.3 A list of Personalis related publications shall be produced and reviewed.**

Priority:P1

Additional Description: Publication list includes employee authorships as well as publications supporting use of the public tools and methodologies in the PGS.

Definition of Done:

- 17.4 Estimates of the sensitivity and specificity of variant detection shall be provided.**

Priority:P1

Definition of Done:

- 17.5 Comparisons between results generated with full complement of Personalis' databases and other databases such as OMIM and HGMD shall be provided.**

Priority:P1

Definition of Done:

- 17.6 Customer facing descriptions and explanations of the various QC metrics, annotation columns, and report columns shown to customers in the final PGS report pages shall be provided.**

Priority:P1

Additional Description: Descriptions include a definition of what the metric/column means, it's source, and what interpretations (if any) can be made from it (why does the customer care about this number).

Definition of Done:

17.7 Customer facing slide decks and data presentations that summarize and support the analysis system, annotations, and reporting methods with example data analyses shall be produced.

Priority:P1

Additional Description: A dual purpose for such presentations is training of Personalis staff for consistent customer messaging.

Definition of Done:

17.8 Summary qc pages, annotations, and reports shall consistently use a common color scheme, Personalis logos, and style/theme (where possible).

Priority:P1

Definition of Done:

17.9 A support function/process for customers shall be outlined including chain of contact, defect submission, and results delivery/re-delivery mechanisms.

Priority:P1

Definition of Done:

17.10 Customer facing disclaimers and privacy statements shall be produced and reviewed.

Priority:P1

Definition of Done:

17.11 Standard terms and conditions shall be produced and reviewed by development.

Priority:P1

Definition of Done:

17.12 FTO and licensing determinations for tools and databases used to produce content that is released to customers shall be documented.

Priority:P1

Definition of Done: