ImmunoID NeXT
Analysis Pipeline Documentation

DRAFT

Document 101-172, Revision C

Table of Contents

Pipeline Overview **3**
Workflow Diagram **4**
Pipeline Tools **5**
Annotation Types and Databases **6**
Quality Control (QC) Reports **6**
Data Deliverables **7**
Directory Structure **7**
DNA Pipeline Reports 9
RNA Pipeline Reports 10
Neoantigen (and Immunogenomics) Reports 11
TCR Reports 11
Results Summary (Quality Control) Reports 12
Sequencing and Alignment **12**
Sequencing Data 12
Alignment Data 12
Small Variants: Single Nucleotide Variants (SNVs) and Insertion/Deletions (Indels) **13**
Individual Sample SNV and Indel Calling 13
Somatic SNV and Indel Calling 13
SNV and Small Indel Annotation Report 13
Annotation Reports 14
Copy Number Alterations (CNAs) **16**
Gene Fusions **16**
Gene Expression Analysis **16**
HLA **17**
Interactive Results Summary (or QC) Reports **17**
QC Report for Normal and Tumor DNA Exomes **17**
Sample and Run Information 17
Sequencing Information 18
Alignment Information 18
Variant Calling 19
Somatic Variant Calling and Annotation 20
Somatic Cancer Annotation 21
QC Report for Tumor RNA Transcriptome **21**
Sample and Run Information 21
Sequencing Information 22
Alignment Information 22
Variant Calling 23
Somatic Variant Calling and Annotation 24
Somatic Cancer Annotation 25
RNA Fusion Metrics 25
Gene Expression Metrics 26
Attribution **26**
Appendix **28**

DRAFT

About this Document:
This document contains details for the Personalis ImmunoID NeXT, powered by the NeXT Exome and NeXT Transcriptome analysis pipelines.

Pipeline Overview

The Personalis Cancer DNA and RNA pipelines use best-of-breed third-party tools and internally-developed proprietary algorithms in a robust, validated workflow. Together with the ImmunoID NeXT Exome and Transcriptome assays, our products provide highly accurate alignments and variant outputs. The Personalis Cancer DNA Pipeline covers a complete spectrum of somatic variant types: single nucleotide variants (SNVs), short insertions and deletions (indels), copy number alterations (CNAs), and gene fusions.

The Personalis Cancer DNA pipeline extends the Personalis Core DNA pipeline with an expanded feature set and workflow, integrating >20 public and proprietary tools designed specifically for somatic cancer analyses. The annotation workflow integrates content from several somatic mutation reference databases such as COSMIC and ClinVar.

The Personalis Cancer RNA pipeline performs a number of analyses: SNV and Indel calling in the RNA data, gene fusion detection, gene expression analysis, and T-cell receptor (TCR)/B-cell receptor (BCR) repertoire analysis. The gene fusion detections are filtered based on over 40 important features and 6 cancer fusion gene databases, identifying the most likely gene fusion candidates. Finally, expression levels for genes in each sample are accurately measured by counting the reads mapped to each gene's preferred transcript.

The Personalis annotation engine provides a comprehensive, integrated solution to annotate SNVs and indels. Personalis updates, integrates, and version controls these databases on a regular basis. The breadth of databases allows Personalis to provide a wide variety of annotations for variants and genes.

Personalis creates comprehensive quality control (QC) reports, including raw sequencing data statistics, read alignment metrics, and variant call counts for each sample analyzed. The Personalis Cancer DNA pipeline augments the QC report to include details of the somatic analysis and advanced metrics important for cancer research and biomarker identification for immuno-/precision oncology applications.

For the Cancer RNA pipeline, transcriptome summary statistics are provided, including: splice site classification, chromosome mappings, mapping specificity, gene element mapping metrics, and transcript coverage. Additionally, several gene fusion and cancer-associated variant summary statistics are presented in the report, including occurrences of gene fusion events that had been previously identified in cancer or healthy samples, numbers of aligned reads supporting fusion events, and small variant effect summaries, and cancer pathway associations.

The ImmunoID NeXT Platform requires a tumor (DNA and RNA) sample with a matched normal (DNA only) sample:

- The matched normal DNA sample allows analysis of:
    - Variants in the Tumor DNA sample against those in the Normal DNA sample to remove germline variants
    - Variants in the Tumor RNA sample against those in the Normal DNA sample to identify novel, expressed mutations

DRAFT

Workflow Diagram

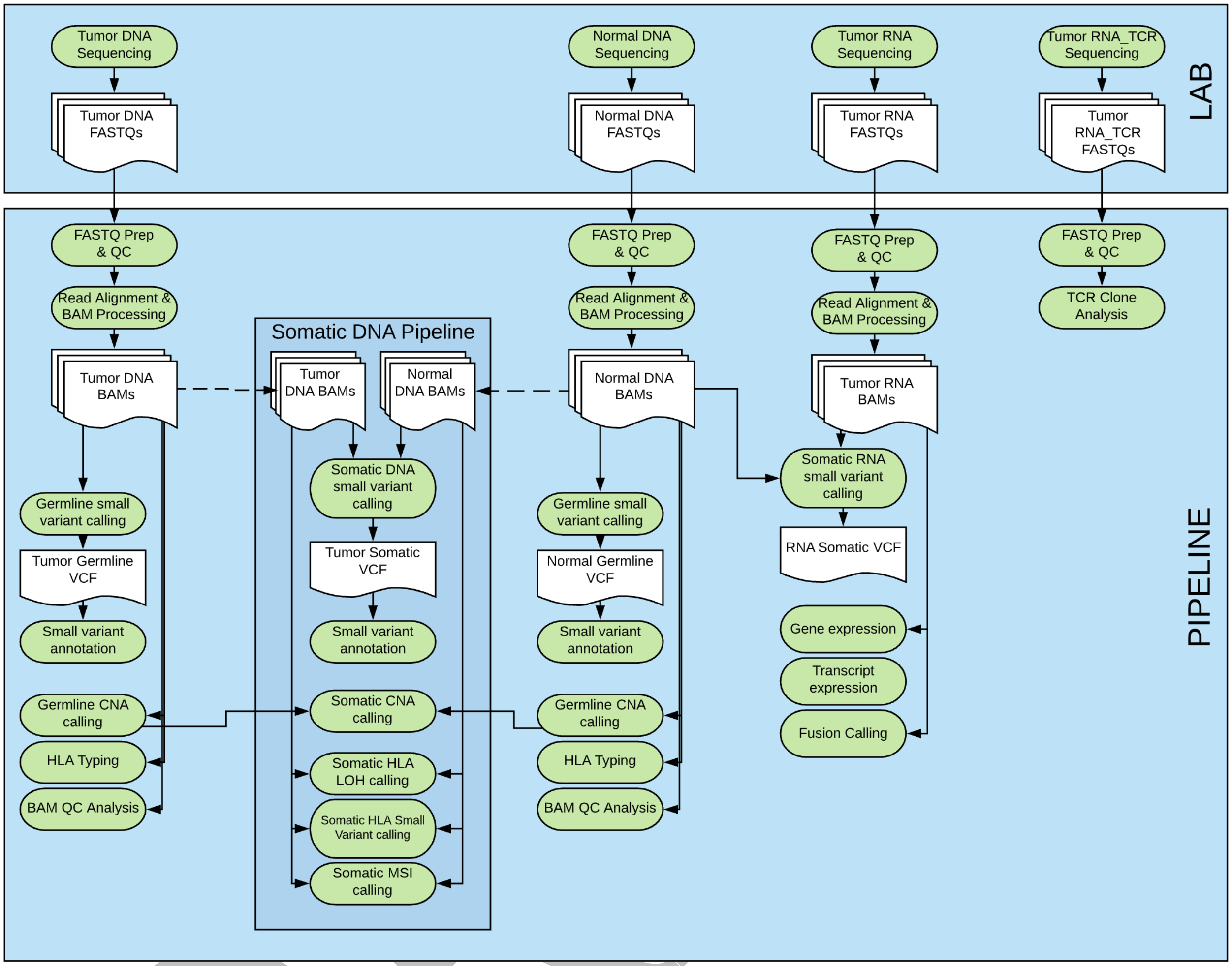


Figure 1: High-level ImmunoID NeXT laboratory and pipeline workflow and data products.

Pipeline Tools

Major software tools employed by our DNA Core pipeline:

Category	Tools	Description	URL
Alignment	BWA	Burrows Wheeler Aligner	http://bio-bwa.sourceforge.net/
SNV & Indel Detection	Sentieon	Optimized reimplementation of GATK	https://www.sentieon.com/
Alignment & Variant Processing	BEDtools	Comparing genomics features	http://bedtools.readthedocs.io/en/latest/
	Tabix	Indexing genome position files	https://github.com/samtools/tabix
	SAMtools	BAM processing	http://samtools.sourceforge.net/
	Novocraft	BAM processing	http://www.novocraft.com/
	Sentieon	Realignment and recalibration	https://www.sentieon.com/
	Personalis tool	Somatic VCF aggregator and filtering	
Annotation	Personalis tool	Functional annotation for SNV and indels	
	snpEFF	Variant effect prediction tool	http://snpeff.sourceforge.net/SnpEff_manual.html#input
HLA Typing	Personalis tool	Determine Class-I and Class-II HLA alleles	
QC and Statistics	FastQC	FASTQ QC statistics	https://github.com/s-andrews/FastQC
	Personalis tool	QC and statistics tool	
	Personalis QC Reporting Engine	Generates visual QC Reports	

Major software tools employed by our Cancer DNA pipeline:

Category	Tools	Description	URL
Somatic SNV & Indel Detection	MuTect	Somatic SNV caller	https://www.broadinstitute.org/cancer/cga/mutect
	Vardict	Somatic Indel caller	https://github.com/AstraZeneca-NGS/VarDictJava
	Picard	Alignment processing	https://broadinstitute.github.io/picard/
Alignment & Variant Processing	BEDtools	Comparing genomics features	http://bedtools.readthedocs.io/en/latest/
	Tabix	Indexing genome position files	https://github.com/samtools/tabix
	SAMtools	BAM processing	http://samtools.sourceforge.net/
	Novocraft	BAM processing	http://www.novocraft.com/
	Sentieon	Realignment and recalibration	https://www.sentieon.com/
	Personalis tool	Somatic VCF aggregator and filtering	
	Personalis tool	Functional annotation for SNV and indels	
HLA	Personalis tool	Somatic variant detection in HLA genes	
	Personalis tool	Somatic HLA LOH detection	
CNA	Personalis tool	Somatic CNA detection	
MSI	MSIsensor	Measure microsatellite instability	https://github.com/ding-lab/msisensor
Annotation	snpEFF	Variant effect prediction tool	http://snpeff.sourceforge.net/SnpEff_manual.html#input
	FastQC	Read QC	http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
	Personalis QC Reporting Engine	Generates visual interactive QC Reports	
QC and Statistics	Personalis tool	QC and statistics tool	

Major software tools employed by our Cancer RNA pipeline:

Category	Tools	Description	URL
Alignment	STAR	Splice-aware read alignment	https://github.com/alexdobin/STAR
SNV & Indel Detection	GATK	SNV and Indel genotyper	https://software.broadinstitute.org/gatk/

DRAFT

Expression (RNA)	Personalis tool	Normalized gene expression	
	Personalis tool	Normalized transcript expression	
Gene Fusion (RNA)	Fusion Catcher	Gene fusion detection	https://github.com/ndaniel/fusioncatcher
TCR/BCR	MiXCR	NGS data-derived TCR/BCR repertoire analyzer	https://mixcr.readthedocs.io/en/master/#
Somatic SNV & Indel Detection	MuTect*	Somatic SNV caller	https://www.broadinstitute.org/cancer/cga/mutect
	Vardict	Somatic Indel caller	https://github.com/AstraZeneca-NGS/VarDictJava
	Picard	Alignment processing	https://broadinstitute.github.io/picard/
Alignment & Variant Manipulation	BEDtools	Comparing genomics features	http://bedtools.readthedocs.io/en/latest/
	Tabix	Indexing genome position files	https://github.com/samtools/tabix
	SAMtools	BAM processing	http://samtools.sourceforge.net/
	SortMeRNA	Ribosomal read filtering	https://github.com/biocore/sortmerna
	GATK	Realignment and recalibration	https://software.broadinstitute.org/gatk/
	Personalis tool	Somatic VCF aggregator and filtering	
	Personalis tool	Functional annotation for SNV and indels	
Annotation	snpEFF	Variant effect prediction tool	http://snpeff.sourceforge.net/SnpEff_manual.html#input
	Personalis QC Reporting Engine	Generates visual interactive QC Reports	
QC and Statistics	Personalis tool	QC and statistics tool	
	FastQC	Read QC	http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Annotation Types and Databases

The Personalis annotation engine uses publicly available databases to annotate SNVs and indels. The annotation databases are listed in the following table.

Annotation Type	Databases	Description
Gene Annotations	RefSeq Ensembl (gene fusions only)	Gene annotations from major gene databases. Includes genomic location of exons, introns, alternative transcripts, protein coding regions, gene symbols, and pseudogenes.
Population Frequencies	ExAC 1000 Genomes NHLBI GO-ESP6500 Exomes	Allele frequency annotations are derived from multiple sources. We include frequencies from both large and small sub-populations across multiple ethnicities.
Mutational Impact	SIFT GERP++ MutationTaster snpEff	Annotations on the mutational impact of variants is derived from multiple methods.
SNPs & Indels	dbSNP	Genomic location, mutation type, and protein coding change.
Variants in Cancer	COSMIC Cancer Gene Census	Annotations on variants and genes known to be contributing factors in cancer are derived from the COSMIC and Cancer Gene Census databases.

Quality Control (QC) Reports

The Personalis Pipeline generates comprehensive quality control (QC), including: raw sequencing data statistics, read alignment metrics, and variant call counts, for each sample analyzed. The pipeline provides an extensive list of standard QC metrics to ensure that each component of the pipeline has run as expected. The key metrics for assessing run quality are shown in the 'Sequencing Information' and 'Alignment Information' sections.

Individual sample-level data are presented in an interactive report that includes detailed tables of standard metrics with links to the more complex supporting data. Somatic QC and summary statistics are also included in a report with cancer-relevant tables and plots depicting genome and chromosome level information. Data contained in the QC reports is also provided in a text-based format.

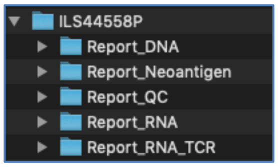
DRAFT

Data Deliverables

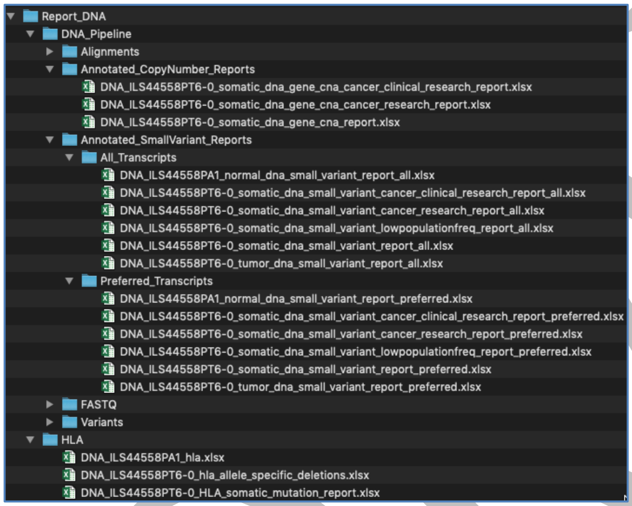
Personalis provides data outputs from each step of the pipeline including sequencing, analysis, and annotation. Wherever possible, Personalis adheres to community standard specifications such as the binary sequence alignment/map (BAM) and variant call format (VCF) files.

Directory Structure

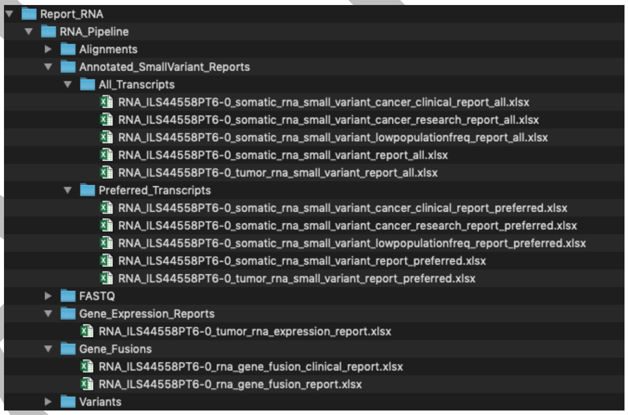
Data is returned with a directory structure to keep the results organized logically. When samples are analyzed individually, directories are nested under a root folder with a <*tumor_sample*> name, which then contains separate directories for DNA, RNA, and Results Summary/QC Reports. Multiple sample analyses have files that contain integrated analysis outputs and are grouped accordingly. This is an example of customer-facing view of the deliverable for a tumor/normal pair with a tumor sample named "`ILS44558P`":



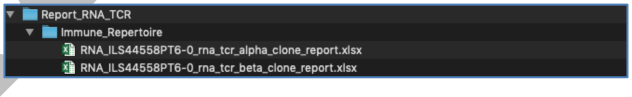
```
▼ ILS44558P
    ▶ Report_DNA
    ▶ Report_Neoantigen
    ▶ Report_QC
    ▶ Report_RNA
    ▶ Report_RNA_TCR
```



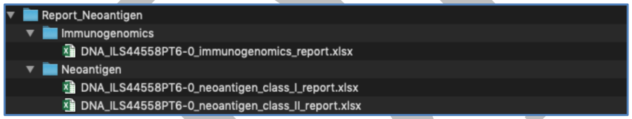
```
▼ Report_DNA
    ▼ DNA_Pipeline
        ▶ Alignments
        ▼ Annotated_CopyNumber_Reports
            DNA_ILS44558PT6-0_somatic_dna_gene_cna_cancer_clinical_research_report.xlsx
            DNA_ILS44558PT6-0_somatic_dna_gene_cna_cancer_research_report.xlsx
            DNA_ILS44558PT6-0_somatic_dna_gene_cna_report.xlsx
        ▼ Annotated_SmallVariant_Reports
            ▼ All_Transcripts
                DNA_ILS44558PA1_normal_dna_small_variant_report_all.xlsx
                DNA_ILS44558PT6-0_somatic_dna_small_variant_cancer_clinical_research_report_all.xlsx
                DNA_ILS44558PT6-0_somatic_dna_small_variant_cancer_research_report_all.xlsx
                DNA_ILS44558PT6-0_somatic_dna_small_variant_lowpopulationfreq_report_all.xlsx
                DNA_ILS44558PT6-0_somatic_dna_small_variant_report_all.xlsx
                DNA_ILS44558PT6-0_tumor_dna_small_variant_report_all.xlsx
            ▼ Preferred_Transcripts
                DNA_ILS44558PA1_normal_dna_small_variant_report_preferred.xlsx
                DNA_ILS44558PT6-0_somatic_dna_small_variant_cancer_clinical_research_report_preferred.xlsx
                DNA_ILS44558PT6-0_somatic_dna_small_variant_cancer_research_report_preferred.xlsx
                DNA_ILS44558PT6-0_somatic_dna_small_variant_lowpopulationfreq_report_preferred.xlsx
                DNA_ILS44558PT6-0_somatic_dna_small_variant_report_preferred.xlsx
                DNA_ILS44558PT6-0_tumor_dna_small_variant_report_preferred.xlsx
        ▶ FASTQ
        ▶ Variants
    ▼ HLA
        DNA_ILS44558PA1_hla.xlsx
        DNA_ILS44558PT6-0_hla_allele_specific_deletions.xlsx
        DNA_ILS44558PT6-0_HLA_somatic_mutation_report.xlsx
```



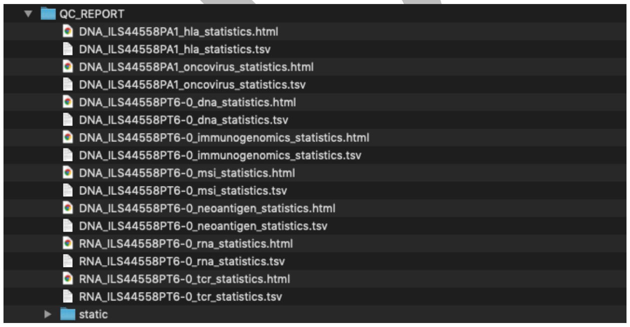
```
▼ Report_RNA
    ▼ RNA_Pipeline
        ▶ Alignments
        ▼ Annotated_SmallVariant_Reports
            ▼ All_Transcripts
                RNA_ILS44558PT6-0_somatic_rna_small_variant_cancer_clinical_report_all.xlsx
                RNA_ILS44558PT6-0_somatic_rna_small_variant_cancer_research_report_all.xlsx
                RNA_ILS44558PT6-0_somatic_rna_small_variant_lowpopulationfreq_report_all.xlsx
                RNA_ILS44558PT6-0_somatic_rna_small_variant_report_all.xlsx
                RNA_ILS44558PT6-0_tumor_rna_small_variant_report_all.xlsx
            ▼ Preferred_Transcripts
                RNA_ILS44558PT6-0_somatic_rna_small_variant_cancer_clinical_report_preferred.xlsx
                RNA_ILS44558PT6-0_somatic_rna_small_variant_cancer_research_report_preferred.xlsx
                RNA_ILS44558PT6-0_somatic_rna_small_variant_lowpopulationfreq_report_preferred.xlsx
                RNA_ILS44558PT6-0_somatic_rna_small_variant_report_preferred.xlsx
                RNA_ILS44558PT6-0_tumor_rna_small_variant_report_preferred.xlsx
        ▶ FASTQ
        ▼ Gene_Expression_Reports
            RNA_ILS44558PT6-0_tumor_rna_expression_report.xlsx
        ▼ Gene_Fusions
            RNA_ILS44558PT6-0_rna_gene_fusion_clinical_report.xlsx
            RNA_ILS44558PT6-0_rna_gene_fusion_report.xlsx
        ▶ Variants
```



```
▼ Report_RNA_TCR
    ▼ Immune_Repertoire
        RNA_ILS44558PT6-0_rna_tcr_alpha_clone_report.xlsx
        RNA_ILS44558PT6-0_rna_tcr_beta_clone_report.xlsx
```



```
▼ Report_Neoantigen
    ▼ Immunogenomics
        DNA_ILS44558PT6-0_immunogenomics_report.xlsx
    ▼ Neoantigen
        DNA_ILS44558PT6-0_neoantigen_class_I_report.xlsx
        DNA_ILS44558PT6-0_neoantigen_class_II_report.xlsx
```



```
▼ QC_REPORT
    DNA_ILS44558PA1_hla_statistics.html
    DNA_ILS44558PA1_hla_statistics.tsv
    DNA_ILS44558PA1_oncovirus_statistics.html
    DNA_ILS44558PA1_oncovirus_statistics.tsv
    DNA_ILS44558PT6-0_dna_statistics.html
    DNA_ILS44558PT6-0_dna_statistics.tsv
    DNA_ILS44558PT6-0_immunogenomics_statistics.html
    DNA_ILS44558PT6-0_immunogenomics_statistics.tsv
    DNA_ILS44558PT6-0_msi_statistics.html
    DNA_ILS44558PT6-0_msi_statistics.tsv
    DNA_ILS44558PT6-0_neoantigen_statistics.html
    DNA_ILS44558PT6-0_neoantigen_statistics.tsv
    RNA_ILS44558PT6-0_rna_statistics.html
    RNA_ILS44558PT6-0_rna_statistics.tsv
    RNA_ILS44558PT6-0_tcr_statistics.html
    RNA_ILS44558PT6-0_tcr_statistics.tsv
    ▶ static
```

DRAFT

© Personalis, Inc. All Rights Reserved

Specifically, results are arranged in five major directories:

1. **DNA Reports:** Contains the raw sequence read data in FASTQ format, alignments in BAM format, variants in VCF format, and annotated small variant and copy number reports in tabular text formats. Also contains HLA reports in tabular text formats when analyzed.
2. **RNA Reports:** Contains the raw sequence read data in FASTQ format, alignments in BAM format, variants in VCF format, gene expression reports in tabular formats, gene fusion reports in tabular formats, and annotated small variant reports in tabular formats.
3. **TCR Reports (from RNA):** Contains tabular report files with TCR alpha and beta clonotype data.
4. **Neoantigen Reports:** Contains tabular report files with variants in immunogenomic genes and predicted neoantigens.
5. **QC Reports:** Tables containing statistical data about the samples, their sequencing data, and the analytical results.

In general, the Personalis pipeline generates the following outputs:

Directory Name	File Type(s)	Deliverable
FASTQ	FASTQ	Raw sequencing data
Alignments	BAM, BAI	Recalibrated sequence alignments and index files
Variants	VCF	Sample-based somatic variant calls (SNVs and small indels)
Annotated_SmallVariant_Reports	XLSX, TSV	Variant and gene-level annotations for SNVs and indels
QC_REPORT	QC and Summary Statistics Reports	Interactive, sample-based QC and summary statistics and result overview pages for advanced analytic modules
Documentation	PDF	Descriptions of Personalis pipelines and deliverables

DRAFT

DRAFT

DNA Pipeline Reports

Results from the DNA Pipeline are found within the top-level directory, entitled “Report_DNA_<id>.”

Directory	File Name	Description
DNA_pipeline/ Alignments	DNA_${tumor_sample}_tumor_dna_aligned_recal_sorted.bam, .bai	BWA aligned, GATK-recalibrated alignments (.bam) and index (.bai) for the tumor sample
	DNA_${normal_sample}_normal_dna_aligned_recal_sorted.bam, .bai	BWA aligned, GATK-recalibrated alignments (.bam) and index (.bai) for the normal sample
	tsv	Tab-separated value files for each output listed here
DNA_pipeline/ Annotated_Smal Variant_Reports	DNA_${tumor_sample}_somatic_dna_small_variant_report.xlsx	This is the main somatic annotated small variant report. This file contains an annotated report of all of the detected somatic SNVs and indels.
	DNA_${tumor_sample}_somatic_dna_small_variant_lowpopulationfreq_report.xlsx	This file contains an annotated report of the detected somatic SNV and indels that • are present at <1% allele frequency in all control population datasets (such as 1000Genomes, ExAC, and ESP), and • are present at ≥5% tumor allele frequency in the sample
	DNA_${tumor_sample}_somatic_dna_small_variant_cancer_research_report.xlsx	This file contains an annotated report of the detected somatic SNV and indels that • are present at <1% allele frequency in all control population datasets (such as 1000Genomes, ExAC, and ESP), • are present at ≥5% tumor allele frequency in the sample, • result in a moderate or high effect on protein function, and • are present in the Personalis Research Cancer Gene List
	DNA_${tumor_sample}_somatic_dna_small_variant_cancer_clinical_research_report.xlsx	This file contains an annotated report of the detected somatic SNV and indels that • are present at <1% allele frequency in all control population datasets (such as 1000Genomes, ExAC, and ESP), • are present at ≥2% tumor allele frequency in the sample, • result in a moderate or high effect on protein function, and • are present in the Personalis Clinical Cancer Gene List
	DNA_${tumor_sample}_tumor_dna_small_variant_report.xlsx	Annotations for SNV and indels in the tumor sample, including both somatic and non-somatic variants
	DNA_${normal_sample}_normal_dna_small_variant_report.xlsx	Annotations for SNV and indels in the normal sample, which are interpreted as germline variants
DNA_Pipeline / Annotated_Copy Number_Reports	DNA_$(tumor_sample)_somatic_dna_gene_cna_report.xlsx	This file contains an annotated report of the detected somatic copy number alterations
	DNA_$(tumor_sample)_somatic_dna_gene_cna_cancer_research_report.xlsx	This file contains an annotated report of the detected somatic copy number alterations that • are present in the Personalis Research Cancer Gene List.
	DNA_$(tumor_sample)_somatic_dna_gene_cna_cancer_clinical_research_report.xlsx	This file contains an annotated report of the detected somatic copy number alterations that • are present in the Personalis Clinical Cancer Gene List
DNA_pipeline/ Annotated_Small Variant_Reports/tsv		Tab-separated value files for each output listed here
DNA_pipeline/ FASTQ	DNA_${normal_sample}_normal_dna_reads1.fastq.gz DNA_${normal_sample}_normal_dna_reads2.fastq.gz	Contains paired-end reads for the normal sample
	DNA_${tumor_sample}_tumor_dna_reads1.fastq.gz DNA_${tumor_sample}_tumor_dna_reads2.fastq.gz	Contains paired-end reads for the tumor sample
DNA_pipeline/Variants	DNA_${normal_sample}_normal_dna.vcf	This file contains all of the SNV and indel called by the Personalis pipeline in the provided normal sample alone.
	DNA_${tumor_sample}_somatic_dna.vcf	This file contains somatic SNVs and indels called by MuTect and Vardict and filtered with the Personalis analysis tool, MVP.
	DNA_${tumor_sample}_tumor_dna.vcf	This file contains all of the SNV and indel called by the Personalis pipeline in the provided tumor sample alone.
QC_REPORT	DNA_${tumor_sample}_dna_statistics.html	html file linking to the interactive DNA report

	DNA_${tumor_sample}_dna_statistics.tsv	tsv version of the DNA QC report
	RNA_${tumor_sample}_rna_statistics.html	html file linking to the interactive RNA report
	RNA_${tumor_sample}_rna_statistics.tsv	tsv version of the RNA QC report
QC_REPORT/ static		Supporting content for interactive reports and advanced analytic modules
HLAs	{sample_name}_hla.xlsx {sample_name}_hla.tsv	.tsv and .xlsx files listing HLA locus and allele information
	{sample_name}_hla_somatic_mutations_report.xlsx	.xlsx file listing HLA genes in which somatic mutations (SNVs and/or indels) have been detected
	{sample_name}_hla_allele_specific_deletions.xlsx	.xlsx file listing HLA genes in which an allele-specific deletion has occurred, resulting in loss of heterozygosity (LOH) at that locus

RNA Pipeline Reports

Results from the RNA Pipeline are found folder within the top-level directory, entitled "Report_RNA_<id>."

Directory	File Name	Description
RNA_pipeline/ Alignments	RNA_${tumor_sample}_tumor_rna_aligned_recal_sorted.bam, .bai	STAR aligned, GATK-recalibrated alignments (.bam) and index (.bai) for the tumor sample
	RNA_${tumor_sample}_tumor_rna_aligned.sorted.bam, .bai	STAR aligned reads {.bam} and index (.bai) for the normal sample
RNA_pipeline/ Annotated_Small Variant_Reports	RNA_${tumor_sample}_somatic_rna_small_variant_report.xlsx	This is the main somatic annotated small variant report. This file contains an annotated report of all of the detected somatic SNVs and indels.
	RNA_${tumor_sample}_somatic_rna_small_variant_lowpopulationfreq_report.xlsx	This file contains an annotated report of the detected somatic SNV and indels that • are present at <1% allele frequency in all control population datasets (such as 1000Genomes, ExAC, and ESP), and are present at ≥5% tumor allele frequency in the sample
	RNA_${tumor_sample}_somatic_rna_small_variant_cancer_research_report.xlsx	This file contains an annotated report of the detected somatic SNV and indels that • are present at <1% allele frequency in all control population datasets (such as 1000Genomes, ExAC, and ESP), • are present at ≥5% tumor allele frequency in the sample, • result in a moderate or high effect on protein function, and are present in the Personalis Research Cancer Gene List
	RNA_${tumor_sample}_somatic_rna_small_variant_cancer_clinical_research_report.xlsx	This file contains an annotated report of the detected somatic SNV and indels that • are present at <1% allele frequency in all control population datasets (such as 1000Genomes, ExAC, and ESP), • are present at ≥2% tumor allele frequency in the sample, • result in a moderate or high effect on protein function, and are present in the Personalis Clinical Cancer Gene List
	RNA_${tumor_sample}_tumor_rna_small_variant_report.xlsx	Annotations for SNV and indels in the tumor sample, including both somatic and non-somatic variants
RNA_pipeline/ Annotated_Small Variant_Reports/tsv		Each tab of the Excel files cited above is provided as a text file with tab separated values
RNA_pipeline/ Gene_Expression_Reports	RNA_${tumor_sample}_tumor_rna_expression_report.xlsx	.xlsx file listing the gene expression level of all genes, providing normalized metrics including FPKM (fragments per kilobase per million), CPM (counts per million), and TPM (transcripts per million), as well as each gene's expression-based percentile among all genes and whether it is expressed or not
RNA_pipeline/ Gene_Expression_Reports/tsv		Each tab of the Excel files cited above is provided as a text file with tab separated values
RNA_pipeline/ FASTQ	RNA_${tumor_sample}_tumor_rna_reads1.fastq.gz	Read is either 'reads1' or 'reads2', and describes the paired-end segment of the reads.
	RNA_${tumor_sample}_tumor_rna_reads2.fastq.gz	Read is either 'reads1' or 'reads2', and describes the paired-end segment of the reads.

DRAFT

RNA_pipeline/ Variants	RNA_${tumor_sample}_somatic_rna.vcf	This file contains somatic SNVs and indels called by the somatic pipeline and filtered with the Personalis analysis tool, MVP.
	RNA_${tumor_sample}_tumor_rna.vcf	This file contains all of the SNV and indel called by the Personalis pipeline in the provided tumor sample alone.
RNA_pipeline/ Gene_Fusions	RNA_${tumor_sample}_rna_gene_fusion_report.xlsx	This file contains all of the fusions called by the Personalis pipeline
	RNA_${tumor_sample}_rna_gene_fusion_clinical_report.xlsx	This file contains detected fusions with at least one gene partner of known cancer clinical significance (not limited to known fusion significance)
RNA_pipeline/ Gene_Fusions/tsv		Each tab of the Excel files cited above is provided as a text file with tab separated values
QC_REPORT	DNA_${tumor_sample}_dna_statistics.html (.tsv)	html file linking to the interactive DNA report (and text version)
	RNA_${tumor_sample}_rna_statistics.html (.tsv)	html file linking to the interactive RNA report (and text version)
QC_REPORT/ static		raw content for html files

Neoantigen (and Immunogenomics) Reports

The Neoantigen Reports contain lists of predicted neoantigens in the tumor specimen based on the expressed somatic variants (small variants and gene fusions) detected, while the Immunogenomics Reports characterize critical areas of tumor and immune biology such as the antigen processing machinery (APM), human leukocyte antigens (HLA), checkpoint modulation, tumor escape mechanisms, the adaptive and innate immune response, and all reports are contained in the top-level directory, entitled “Report_Neoantigen_<id>.”

Directory	File Name	Description
Neoantigen	DNA_${tumor_sample}_neoantigen_class_I_report.xlsx	Spreadsheet containing a list of class I MHC neoantigens predicted based on expressed somatic variants detected in the tumor and their MHC-binding affinity
	DNA_${tumor_sample}_neoantigen_class_II_report.xlsx	Spreadsheet containing a list of class II MHC neoantigens predicted based on expressed somatic variants detected in the tumor and their MHC-binding affinity
	tsv/DNA_${tumor_sample}_neoantigen_class_I_report_Fusions.tsv	Tab-separated text file containing a list of class I MHC neoantigens predicted based on expressed somatic fusions detected in the tumor and their MHC-binding affinity
	tsv/DNA_${tumor_sample}_neoantigen_class_II_report_Fusions.tsv	Tab-separated text file containing a list of class II MHC neoantigens predicted based on expressed somatic fusions detected in the tumor and their MHC-binding affinity
	tsv/DNA_${tumor_sample}_neoantigen_class_I_report_SNV_Indel.tsv	Tab-separated text file containing a list of class I MHC neoantigens predicted based on expressed somatic small variants detected in the tumor and their MHC-binding affinity
	tsv/DNA_${tumor_sample}_neoantigen_class_II_report_SNV_Indel.tsv	Tab-separated text file containing a list of class II MHC neoantigens predicted based on expressed somatic small variants detected in the tumor and their MHC-binding affinity
Immunogenomics	DNA_${tumor_sample}_immunogenomics_report.xlsx	Spreadsheet containing a list of all immunogenomics genes in which a somatic variant is detected in the tumor specimen (if any)
	tsv/*	Tab-separated text files containing a list of all immunogenomics genes in which a somatic variant is detected in the tumor specimen (if any)

TCR Reports

The TCR Reports contain clone counts and frequencies for the TCR alpha and beta chains based on NeXT Transcriptome sequencing data and are contained in the top-level directory, entitled “Report_RNA_TCR_<id>.”

Directory	File Name	Description
Immune_Repertoire	RNA_${tumor_sample}_rna_tcr_alpha_clone_report.xlsx	Spreadsheet containing a list of all unique TCR alpha clonotype nucleotide sequences, clone counts, and clone frequencies, as well as their respective V and J gene segments and amino acid sequences with all data derived from NeXT Transcriptome data
	RNA_${tumor_sample}_rna_tcr_beta_clone_report.xlsx	Spreadsheet containing a list of all unique TCR beta clonotype nucleotide sequences, clone counts, and clone frequencies, as well as their respective

DRAFT

		V, D, and J gene segments and amino acid sequences with all data derived from NeXT Transcriptome data
	tsv/RNA_${tumor_sample}_rna_tcr_alpha_clone_report.tsv	Tab-separated text file containing a list of all unique TCR alpha clonotype nucleotide sequences, clone counts, and clone frequencies, as well as their respective V and J gene segments and amino acid sequences, with all data derived from NeXT Transcriptome data
	tsv/RNA_${tumor_sample}_rna_tcr_beta_clone_report.tsv	Tab-separated text file containing a list of all unique TCR beta clonotype nucleotide sequences, clone counts, and clone frequencies, as well as their respective V, D, and J gene segments and amino acid sequences, with all data derived from NeXT Transcriptome data

Results Summary (Quality Control) Reports

The Results Summary/Quality Control (QC) Reports display a variety of QC metrics, statistics, and advanced analytics relating to multiple ImmunoID NeXT analytics modules and are contained in the top-level directory titled "Report_QC_<id>."

Directory	File Name	Description
QC_REPORT	DNA_${tumor_sample}_dna_statistics.html/.tsv	html/tsv file linking to the interactive report displaying DNA sequencing QC metrics and statistics
	RNA_${tumor_sample}_rna_statistics.html/.tsv	html/tsv file linking to the interactive report displaying RNA sequencing QC metrics and statistics
	DNA_${tumor_sample}_hla_statistics.html/.tsv	html/tsv file linking to the interactive report displaying HLA typing, somatic mutations, and allele-specific deletions information
	DNA_${tumor_sample}_immunogenomics_statistics.html/.tsv	html/tsv file linking to the interactive report displaying gene-level expression (TPM), variant type (SNVs, indels, fusions), variant expression, DNA and RNA allelic fraction, as well as variant effect impact information associated with variants occurring specific genes that play a critical role in immuno-oncology-related processes
	RNA_${tumor_sample}_msi_statistics.html/.tsv	html/tsv file linking to the interactive report displaying MSI status information determined by the stability status of five canonical loci as well as by the proportion of all exome-wide MSI-related loci that are found to be unstable
	DNA_${tumor_sample}_oncovirus_statistics.html/.tsv	html/tsv file linking to the interactive report displaying oncoviruses that were detected in the tumor specimen
	RNA_${tumor_sample}_tcr_statistics.html/.tsv	html/tsv file linking to the interactive report displaying TCR alpha and TCR beta-related information including clonality and the top ten clonotypes detected in the tumor specimen
	DNA_${tumor_sample}_neoantigen_statistics.html/.tsv	html/tsv file linking to the interactive report displaying the top predicted neoantigens in the tumor specimen as well as neoantigen burden and tumor mutational burden (TMB)
QC_REPORT/static		Supporting content for interactive reports and advanced analytic modules

DRAFT

Sequencing and Alignment

Sequencing Data

Original sequences are converted from the sequencer's proprietary format into the standard FASTQ format with standard (Sanger) Phred-scale+33 quality scores. Sequence data is formatted according to the MAQ FASTQ format: http://maq.sourceforge.net/fastq.shtml.

Alignment Data

The ImmunoID NeXT Exome/Transcriptome Analysis Pipeline aligns reads to the hs37d5 reference genome. The pipeline performs alignment, duplicate removal, and base quality score recalibration using best practice guidelines recommended by the Broad Institute. The pipeline uses novosort for duplicate removal and Genome Analysis Toolkit (GATK) to improve sequence alignment and to recalibrate base quality scores (BQSR). This provides more accurate quality scores by correcting for variation in quality with machine cycle and sequence context. (See http://gatkforums.broadinstitute.org/discussion/44/base-quality-score-recalibration-bqsr for discussion.) Aligned sequence data is then returned in the BAM format according to the SAM specification: http://samtools.sourceforge.net/SAM1.pdf

Small Variants: Single Nucleotide Variants (SNVs) and Insertion/Deletions (Indels)

Individual Sample SNV and Indel Calling

GATK's HaplotypeCaller module provides the pipeline's core set of germline SNV calls and their accompanying quality metrics. The pipeline then uses GATK's variant quality score recalibration module, which stratifies SNVs by their likelihood of representing false positive calls, and in-house SNV accuracy software, which incorporates both genomic context and sequence alignment information into a model that corrects miscalled variants.

All calls are made on BAM files that have been recalibrated by GATK's BAM processing tools. Variant calls are reported in VCF files, which also include standard metrics such as average mapping quality and statistics describing consistency of each variant call with the diploid genome model.

Personalis returns SNV and small Indel calls in a standard VCF file and adheres to the VCF format specification version 4.1, as detailed by the 1000 Genomes Project: http://www.internationalgenome.org/wiki/Analysis/Variant_Call_Format/vcf-variant-call-format-version-41

Somatic SNV and Indel Calling

The Personalis Cancer DNA and RNA Pipelines integrate open source, commercial, and proprietary tools to produce a set of somatic SNVs – that is, variants that are present in the tumor, but not in the matched normal sample. All calls are made on BAM files that have been recalibrated by GATK's BAM processing tools.

The somatic SNV caller utilizes a Bayesian classifier approach. It evaluates alignment files from both tumor and matched normal samples individually and simultaneously to determine the likelihood of a somatic variant at each nucleotide position. Somatic DNA calls are made by assessing the DNA Tumor mapped reads in the context of the DNA Normal mapped reads; somatic RNA calls are made by assessing the RNA Tumor mapped reads in the context of the DNA Normal mapped reads. Somatic SNVs are further filtered to remove variants in dbSNP, COSMIC, and a pool of normals from MuTect. A somatic indel caller is used to call small somatic insertions or deletions with a similar approach, but for small insertions or deletions (<50bp) at a particular position.

Somatic SNV and indel calls are combined and analyzed through a comprehensively tested set of filters based on i) alignment metrics, such as sequence coverage and read quality, ii) positional features, such as proximity to a gap region, and iii) likelihood of presence in normal tissue.

SNV and Small Indel Annotation Report

The Personalis Variant Annotation Report contains extensive annotations for the detected SNV and small indel variants. This file can be used to quickly retrieve the variety of annotations Personalis provides and for annotation-based filtering of variants. Annotations at three levels of granularity:

1. **Variant-level** annotations are specific to the variant. Examples of variant-level annotations include affected coding or ncRNA genes, genomic location, mutational effect, genetic elements affected, problematic regions, predicted impact scores, dbSNP rsIDs, population frequencies, disease annotations, and others.

2. **Gene-level** annotations describe the gene in which the variant is associated. Examples of gene-level annotations include pathways, frequency of mutations across tumor types, presence in Cancer Gene Census, and others. These gene annotations are based on Entrez GeneID and Gene Symbol, allowing for rapid cross-referencing with other gene-based resources.

3. **Variant- + Transcript-level** annotations define the role of the variant within a specific transcript. Examples of transcript-level annotations include functional class, codon change, and effect impact.

DRAFT

DRAFT

All variants are primarily reported in the context of preferred transcripts. Further annotations are provided against all transcripts to suit those cases when further detail is needed. To select a preferred transcript, we initially select the transcript with the most clinical evidence in cancer. In occurrences where there are multiple transcripts with equal evidence, we select the one that is most cited in COSMIC. In a few cases where there is still more than one candidate transcript, the transcript with the longest CDS length is selected.

Annotation Reports

In addition to full annotations of every variant detected, the Personalis Cancer DNA Pipeline returns annotated variant reports that are more targeted for cancer analysis: variants seen at low frequency in the normal population, variants present in an extensive list of cancer genes, and variants with moderate or highly debilitating effects on gene function.

The small variants that are of most relevance to cancer analyses are often those falling within genes previously known to be involved in cancer and cancer pathways. The Cancer Gene Census from the Sanger Institute is a set of ~600 genes that are thought to play a role in cancer development. Expanding upon this limited set, the Personalis Research Cancer Gene List is a comprehensively curated list of over 1,400 cancer genes, including genes with important therapeutic implications as well as genes with accumulating evidence of importance in tumor biology.

Variant Annotation Report Column Descriptions

Columns of data returned in the file are organized as shown in the table below. Various classes of annotation are provided including:

1. **Gene symbol and location -** HGNC Gene symbols and chromosomal location. This set of annotations describes genomic features within which the variant occurs. Genomic features include genes, transcripts, predicted transcripts, and the predicted effect of the variant on the coding region, if applicable. This set also includes cytoband locations and dbSNP identifiers, where available.

2. **Effect -** Describes the putative structural and functional consequences of the variant. Where applicable and available, the Ensembl transcript ID is provided, as well as the exon within which the variant occurs.

3. **Population frequencies -** Used to assess the rarity of variants in the population at large. Population frequency is often used as a filter when the rarity of the phenotype is known. For example, if the phenotype in question were very rare, it would be unexpected for a common variant in the population to be causal of the phenotype. Observed frequencies of the alternate allele are provided for 1000 Genomes data and ExAC projects.

4. **Impact scoring -** These annotations utilize a variety of algorithms to estimate the mutational impact of variations on gene function. Each algorithm uses a different approach, so their estimates will vary.

5. **Cancer annotations**_- These annotations include variant- and gene-level information with regards to associations with cancer, including whether this variant or gene has been seen previously in cancer studies and what particular tumor types are most likely to harbor this event.

Variant Annotation Report Column Descriptions

Column Name	Data Type	Description	Information Keyed On
Variant ID	integer	Unique identifier for a variant. This identifier is only unique within this sample (e.g. 1)	variant location
Sequence	Alpha-numeric	Chromosome or sequence upon which the variant was identified (*e.g.*, 19 or HG1_patch)	variant location
POS	integer	The position of the variant as defined by the VCF (*e.g.*, 94234)	variant location
REF	character string	The reference sequence for this variant as described in the VCF (*e.g.*, G)	variant location
ALT	character string	The alternate sequence for this variant as described in the VCF (*e.g.*, A)	variant location
Quality Score	decimal	The variant quality scored assigned by VSQR from the VCF (*e.g.*, 345)	variant location

DRAFT

Total Read Depth	integer	Number of high quality (>map Q20) reads at this position. (*e.g.*, 200)	variant location
Reads Supporting REF	integer	Number of high quality (>map Q20) reads supporting the reference allele (e.g. 100)	variant location
Reads Supporting ALT	integer	Number of high quality (>map Q20) reads supporting the alternate allele (e.g. 100)	variant location
Allelic Fraction	decimal	allelic fraction of variant in a background of the reference allele	
Genomic Variant	string	The HGVS description of the genomic variant (*e.g.*, g.11182171G>A)	variant location
Gene Symbol	string	HGNC symbol for the gene associated with the variant (*e.g.*, ABL1)	variant location
NCBI Gene ID	integer	Gene ID provided by NCBI (*e.g.*, 2)	variant location
Transcript ID	string	The RefSeq accession.version for the transcript used for variant analysis	variant location
Preferred Transcript	string	The RefSeq accession.version for the transcript used for variant analysis. Personalis uses a curated list of transcripts, which is based on the number of times a transcript (accession.version) is referred to in COSMIC. If not present in COSMIC, the default transcript would be the one corresponding to the longest CDS.	variant location
Transcript Biotype	string	The biotype of the transcript (*e.g.*, coding or ncRNA)	variant location
Transcript Variant	string	The variant as described in transcript coordinates, only applicable for small variants (*e.g.*, c.2641G>A)	variant location
Protein ID	string	The RefSeq accession.version for the protein used for variant analysis (*e.g.*, NP_012345.1)	variant location
Protein Variant	string	Description of the variant at the protein level (*e.g.*, p.G872S)	variant location
Variant Effect	string	The effect the variant has on the associated protein sequence. (*e.g.*, MISSENSE_VARIANT) (From SnpEFF.)	allele + transcript
Variant Effect Impact	string	The predicted impact of this variant (*e.g.*, HIGH)._(From SnpEFF.)	allele + transcript
Functional Class	string	The functional class of this variant (*e.g.*, nonsense). (From SnpEFF.)	allele + transcript
Codon Change	string	The variant in the context of the codon (*e.g.*, tCa/tGa). (From SnpEFF.)	allele + transcript
Exon Number	integer	The exon number that variant is found in, with respect to the transcript	allele + transcript
COSMIC Mutation ID	integer	The identifier assigned by COSMIC (51441)	allele
COSMIC Amino Acid Change	string	Amino acid change at this variant as descripted by COSMIC (*e.g.*, p.A3V).	allele
COSMIC Transcript Change	string	Transcript change for this variant (*e.g.*, c.54A>G). Note, this can differ from the transcript variant above due to the use of another transcript in COSMIC	allele
Seen as Somatic	string	Identifies if this variant has been observed previously in the COSMIC database as a somatic variant (yes/no)	gene
Seen as Germline	string	Identifies if this variant has been observed previously in the COSMIC database as a germline variant (yes/no)	gene
Cancer Gene Census	string	Identifies if this gene is in the cancer gene census list (yes/no)	gene
ExAC Total	float	Allele frequency in ExAC/GnomAD for all populations	allele
1KG Total	float	Allele frequency in the total 1000 Genomes population	allele
ESP6500 Total	float	Allele frequency in total ESP6500 population	allele
GERP	decimal	GERP provides an estimate of evolutionary constraint (GERP_RS) based on multiple sequence alignments. Larger scores indicate higher conservation. Cooper et al., 2005. (dbNSFP)	allele
SIFT Score	decimal	SIFT predicts functional effect based on the degree of conservation from alignments of closely related sequences. Scores less than 0.05 are predicted to be damaging, higher scores indicate tolerance. Kumar et al., 2009. (dbNSFP)	allele

MutationTaster Pred	string	Mutation Taster uses a Bayesian approach to classify the disease potential of a variant using all known disease causing variants in HGMD Pro and >6.8 million variants from the 1000 Genomes project. Schwarz JM et al., 2010. (dbNSFP)
		A = disease_causing_automatic; D = disease_causing
		N = polymorphism; P = Polymorphism automatic
			allele

MutationTaster Score	string	The probability that the MutationTaster prediction of the disease potential is correct	allele
CADD Phred	decimal	CADD assesses the deleteriousness of SNVs and small indels by integrating annotations from a variety of sources and comparing observed vs. simulated variants, resulting in a scaled-phred score. Higher scores indicate higher likelihood that the variant is deleterious. Kircher et al., 2013. (dbNSFP)	allele
dbSNP ID	integer	dbSNP identifier for this variant (*e.g.*, 185523638)	variant location
dbSNP Build	integer	dbSNP build in which this variant first appeared (*e.g.*, 135)	variant location

DRAFT

Notes:

1. Annotations from COSMIC and RefSeq may be based on different transcript and protein models, so they may not match.

Copy Number Alterations (CNAs)

The Personalis pipeline identifies somatic copy number alterations (CNAs) with the use of a proprietary tool known as CNAState. CNA events are called based on deviations from the normalized per-exon read-depth level across genes.

Gene Fusions

The Personalis pipeline identifies gene fusions with FusionCatcher. The pipeline begins by filtering reads, including 3' trimming, removal of adapter sequence, trimming of poly tails, and removal of reads with short tandem repeats, poor sequencing quality, ribosomal sequence, and/or bacterial/viral sequence. Following filtering, FusionCatcher identifies gene fusions through alignment of FASTQ RNA-seq reads using STAR, BOWTIE, and BLAT. We utilize BLAST alignment of each putative fused nucleotide sequence to identify and reject False Positive fusion events. In the clinical fusion report, we apply a level of evidence filter: passed fusion events must contain at least 10 read pairs that span the fusion junction (although this requirement is waived for a whitelist of known clinical fusion gene pairs such as EML-ALK). The clinical fusion report also only contains events where at least one of the two genes is in a list of known clinical cancer genes. Finally, we utilize certain quality tags applied by FusionCatcher to reject calls from our clinical report. We also return a research fusion report where none of this filtering is applied.

Supporting reads for each final fusion event are provided in the supporting read files (separated by whether the fusion was identified through BOWTIE, STAR, or BLAT).

Gene Expression Analysis

Gene-based expression is calculated for each reference gene as the union of all annotated exons. An intersection nonempty approach is utilized where reads are allowed to either span or hop introns and extend into another gene region. However, for reads to be counted they must be uniquely attributable to a single gene. In other words, when a read is mapped to a region which contains two gene annotations, a portion of the read must have a flanking region which maps only to a single gene to be counted to which it will be singularly attributed. Reads that ambiguously map to two genes are not counted, as it is impossible to decipher which gene they represent.

Using an in-house algorithm, we take raw strand-specific counts per gene table generated by the STAR aligner and compute normalized expression values, including CPM (Counts per Million mapped reads), FPKM (Fragments per Kilobase per Million mapped reads), and TPM (transcripts per million) for genes in a given assay. The final report includes the following data:

Column	Description	Calculation
Gene Symbol	NCBI Gene symbol	N/A
NCBI Gene ID	NCBI Gene Identifier	N/A
RNA-Seq Raw Counts	Star generated Raw Counts mapping to the Gene	N/A
FPKM	Fragments Per Kilobase per Million Mapped Reads	$10^9*X_i/(N*L_i)$, where • X_i is counts for a given gene • L_i is the length of the gene CDS • N is total number of mapped reads
CPM	Counts Per Million mapped reads	$10^6*X_i/N$, where • X_i is counts for a given gene • N is total number of mapped reads
TPM	Transcripts Per Million	$10^6*(X_i/L_i)/\Sigma(X_j/L_j)$, where • X_i is counts for a given gene • L_i is the length of the gene CDS

		- $\Sigma(X_j/L_j)$ is the sum of the ratio of counts to length for all genes in the assay
Percentile Rank	Percent rank of the gene's TPM among all other genes' TPM	N/A
Is Expressed	A binary estimate of whether a gene is expressed or not, using a TPM ≥2 cutoff	N/A

HLA

In certain configurations, HLA locus and allele information is included in the form of its own tab in the Results Summary (or QC) Report and as separate tsv and xlsx file outputs for use by downstream antigen presentation applications.

Example output:

HLA	Allele 1	Allele 2
A	A*11:01:01	A*34:02:01
B	B*27:05:02	B*82:01
C	C*01:02:01	C*03:02:02
DPA	DPA1*01:03:01	DPA1*02:02:02
DPB	DPB1*04:02:01	DPB1*01:01:01
DQA	DQA1*02:01	DQA1*01:02:01
DQB	DQB1*06:02:01	DQB1*02:02:01
DRB1	DRB1*07:01:01	DRB1*15:03:01
DRB3	nocall	nocall
DRB4	DRB4*01:01:01	DRB4*03:01N
DRB5	DRB5*01:01:01	DRB5*01:01:01

Somatic mutations and/or loss of heterozygosity (LOH) events impacting any HLA Class I genes are also detailed in HLA tab of the Results Summary Reports.

DRAFT

Interactive Results Summary (or QC) Reports

These reports include detailed statistics generated during the sequencing and pipeline analysis of the sample. For the Cancer DNA and Cancer RNA tabs, the reports are categorized into the following sections: alignment, fusion, variant annotation, and cancer gene filtered variant annotation. Details of each section and the terminology used are included below.

The small triangles on the left of each section indicate they can be expanded to display additional tables or data and graphs. For convenience, an "Open All" button is also provided to expand all tables in a section.

The Results Summary Reports are provided in both html and tsv format. They contain sequencing and alignment quality metrics as well as summary statistics of the somatic analyses.

Results Summary (QC) Report for Normal and Tumor DNA Exomes

Sample and Run Information

Display Name	Definition	Additional Details and Interpretation

Analysis Mode	Type of somatic analysis performed: Tumor/Normal	Tumor/Normal mode uses the matched normal sample provided to filter out possible germline and contamination from somatic SNV calls.
Tumor sample	Name of tumor sample used for analysis.	
Matched normal sample	Name of matched normal sample used for analysis.	
Pipeline versions	Version of the Personalis pipeline run for the sample	
Annotation version	Version of the Personalis Annotation used for the data analysis	
Platform version	Personalis Lab Assay version	
Reference assembly	Reference assembly used	

Sequencing Information

The Sequencing Information section includes a summary of basic sequencing statistics generated for each sample as well as predicted gender and blood type for the normal sample, if present.

Display Name	Definition	Additional Details and Interpretation
Read length (bp)	Number of bases in a read.	If a sample has reads of different lengths, the lengths will be listed as a comma delimited set.
Total reads	Total number of reads that were sequenced and passed fastq filters.	This number is proportional to sequence coverage. Higher numbers per sample lead to better variant calling.
Total bases	Total number of bases sequenced.	This number is proportional to sequence coverage. Higher numbers per sample lead to better variant calling.
Average base quality	Mean base quality score for reads that pass filter. Calculated after GATK base recalibration.	This value should be >30 (Q30) for a good run. Lower values indicate systematic sequencing problems.
Sex chromosome count	Counts of X and Y chromosomes.	A normal male is represented as XY and a normal female is represented as XX. Outside range indicates sex chromosome aneuploidy.
Predicted sex	Predicted sex based on chromosome count in normal sample.	
Predicted blood type	A, B, AB, and O phenotypes.	U indicates unknown status.
Percent contamination in Normal	Percentage of predicted contamination by other samples using only matched normal.	
Percent contamination in Tumor	Percentage of predicted contamination by other samples, using both tumor and normal.	This is the mean predicted contamination across all chromosomes. This does not include normal-in-tumor contamination. See the description of ConTest for more information.

QC Plots for Normal and Tumor DNA Exomes

Quality Control graphs are included to visually display common metrics of run quality for each read across both the Normal and Tumor DNA Samples:

- Quality Scores by Read Position
- Average Quality per Read
- Base Identity by Read Position
- GC Distribution Over All Sequences

Alignment Information

The Alignment Statistics section includes detailed statistics generated during the alignment stage of the Personalis Pipeline analysis. A series of statistics and graphs are displayed that indicate various aspects of the quality of the alignment.

Display Name	Definition	Additional Details and Interpretation
Average read depth	Mean coverage across the genome (not including N bases) based on mapped reads.	A normalized measure of the number of mapped reads, representing the average number of reads at a given genomic position. Low alignment coverage can cause poor variant calls and likely means poor library prep or poor enrichment in exome capture protocols.

DRAFT

Percent mapped reads	Percentage of the reads that map to the genome	The percentage of all reads that map to the reference genome compared to the total of all reads that pass raw quality filters.
Average Mapping quality	Mean mapping quality score.	Overall quality of the read mapping. A low score usually indicates poor quality sequencing, either due to systematically poor base calling or due to a large number of repetitive reads and/or unmappable reads.
Percent Duplicate read pairs	The percent of reads marked as duplicates.	Indicates sample prep problems if high. A low number means that unique rather than redundant molecules were sequenced. Expect < 2% for whole genome and 8-20% for exome. Higher values could indicate excessive amplification of library DNA.
Capture Specificity	Fraction of mapped reads that fall in targeted genomic regions.	The fraction of all mapped reads that align within the assay target capture regions.
Insert size	Mean and standard deviation are included for the size of the insert.	An alignment-based metric that reflects the fragment size chosen during library prep. This can be compared to the bioanalyzer measurement (physical measurement rather than the calculated measurement) to look at potential cluster creation biases. A low standard deviation indicates that insert sizes were consistent.

QC Plots for Normal and Tumor DNA Exomes

- Fraction of genome at specified depth
- Mapping quality distribution
- Insert size distribution

Read Mapping Table for Normal and Tumor DNA Exomes

Display Name	Definition	Additional Details and Interpretation
Total reads	Total number of pass-filter reads sequenced.	This number is proportional to sequence coverage. Higher numbers per sample lead to better variant calling. Only reads that pass Illumina chastity filters are counted.
Mapped	Number of reads mapped one or more times to the reference sequence.	If the ratio of mapped reads to total reads generated (Percent mapped reads) is low, it is suggestive of a systematic problem such as DNA contamination or poor quality sequencing. Upstream problems such as sample prep or sequencing reagents can lead to this effect. A ratio of 0.9 or higher generally indicates excellent overall run health.
Unmapped reads	Total number of reads that could not be mapped.	Equal to Total reads minus Mapped reads. A ratio of Unmapped reads to Total reads of 0.1 or lower generally indicates excellent overall run health. See Mapped reads for more information.

Anomalous Read Pair Alignments

Display Name	Definition	Additional Details and Interpretation
Inter-chromosomal	Number of read pairs for which one read maps to one chromosome and whose paired read maps to another chromosome.	Unusually high values could indicate sample prep problems in library creation, especially with mate pairs. Artifacts such as chimeric reads can lead to elevated values.
Orphaned reads	The number of reads mapped to the reference sequence where the paired read does not map.	A simple metric to identify gross failures during sequencing, such as bad base calls on all read 2's for a specific flowcell. This category should include less than 1% of all reads. In a healthy run, orphaned reads occur due to natural features of a genome such as repeats or structural variation.

Variant Calling

The Variant Calling Statistics section includes detailed statistics generated during the variant calling stage of the Personalis Pipeline analysis. A series of statistics are displayed that indicate various aspects of variant calling performed.

Summary Small Variants for Normal and Tumor DNA Exomes

Display Name	Definition	
Total Calls	Total number of small variants detected	Low indicates not enough coverage.
Calls in a public database	Number of SNPs detected that were in dbSNP	Low values mean that the sequenced individual has many novel SNPs which is commonly observed in cancer samples or when sequencing an individual from a population with unusual genetic diversity. Could be contamination as well if low. High indicates didn't call with enough sensitivity

DRAFT

DRAFT

Heterozygous/ Homozygous ratio	Ratio of heterozygous to homozygous SNPs	Abnormal values are associated with low sequence coverage.
Ti/Tv ratio	Ratio of transition (Ti) to transversion (Tv) substitutions for a set of SNPs.	This metric is useful for quality control. A low value is generally associated with high false positive rate. A high value indicates lots of false negatives, which can occur when variant call filtering is overly stringent. Expect a range depending upon platform.

SNV Statistics for Normal and Tumor DNA Exomes

A summary table shows the counts of SNVs that change from each Reference Base to each Alternate Base for each of the Normal and Tumor samples independently.

Tumor/Normal Concordance

Display Name	Definition	Additional Details
Variants called in both Normal and Tumor	Total number of variants common to both Normal and Tumor samples	The total number of variants that are present in both the tumor and normal samples
Variants unique to Normal	Number of variants unique to the Normal sample	The total number of variants that are present in normal but not in the tumor sample
Variants unique to Tumor	Number of variants unique to the Tumor sample	The total number of variants that are present in tumor but not in the normal sample

Somatic Variant Calling and Annotation

Summary Small Variants

Display Name	Definition
Somatic variants	SNVs, indels
Somatic variants per Mb	Number of somatic variants per Mega Base of DNA
Non-synonymous Somatic Variants	Total number of non-synonymous somatic mutations (i.e. mutations that alter the amino acid)
Non-synonymous Somatic Variants per Mb	Number of non-synonymous somatic mutations (i.e. mutations that alter the amino acid) per megabase of DNA. This number is commonly used to describe tumor mutational burden.
Ti/Tv ratio	Ratio of number of transitions to number of transversions for detected SNVs. See above.

SNV Statistics for Somatic DNA variants

A summary table shows the counts of somatic SNVs that change from each Reference Base to each Alternate Base.

Mutation Signature Plots

A set of plots are provided to show the percent of each type SNV change within the context of its neighboring nucleotides.

Functional Annotation

Display Name	Definition
Nonsense	Number of variants that change a codon to a stop codon
Missense	Number of variants that change a codon to produce a different amino acid
Silent	Number of variants that do not result in the change to the amino acid
None	Number of variants that do not have a functional annotation of nonsense, missense, or silent

Effect Annotation

Display Name	Definition
Codon change plus codon deletion	Number of variants that cause a single codon change resulting in one or more codon deletions, such as when deletion of whose size is a multiple of three that occurs within a codon (not at the boundary).
Downstream	Number of variants that are downstream of an annotated gene
Frame shift	Number of variants (insertions or deletions) that shifts the reading frame
Intragenic	Number of variants that occur within a gene, but no transcripts are annotated for the gene
Intron	Number of variants that occur in an intron
Non-coding exon	Number of variants that occur in the non-coding exons
Nonsynonymous coding	Number of variants that cause a codon change that results in an amino acid change
Other	Number of variants that do not have another listed effect annotation

Splice site acceptor	Number of variants that occurs in the splice acceptor site (defined as two bases before the exon start, except for the first exon)
Splice site donor	Number of variants that occurs in the predicted splice donor site (two bases after coding exon end, except for the last exon)
Splice site region	Number of variants within the region of the splice site, either within 1-3 bases of the exon or 3-8 bases of the intron
Start gained	Number of variants in 5'UTR region that produces a three-base sequence that can be a START codon
Stop gained	Number of variants that causes a stop codon
Synonymous coding	Number of variants that causes a codon change that does NOT result in an amino acid change
Upstream	Number of variants that occur within 5000 bases upstream of a gene
UTR 3 prime	Number of variants that occur in the 3' untranslated region (3' UTR)
UTR 5 prime	Number of variants that occur in the 5' untranslated region (5'UTR)

Predicted Effect

Display Name	Definition
Mutation Taster	Mutation Taster predicts the disease-causing potential of the variants and classifies them into 4 categories: Disease Causing Automatic, Disease Causing, Polymorphism, Polymorphism Automatic
LRT	Likelihood Ration Test (LRT) method that identifies deleterious mutations.

Somatic Cancer Annotation

Mutational Effect Impact

Display Name	Definition
All somatic variants	Number of somatic variants that have different mutational impacts (high, moderate, low, and modifier)
All somatic variants in Personalis Research Cancer Gene List.	Number of somatic variants that have different mutational impacts (high, moderate, low, and modifier) in the Personalis Research Cancer Gene List.

Variant Filtering

Display Name	Definition
All Somatic variants	Number of somatic variants that have an allele fraction (AF) greater than 5%
Low population frequency somatic variants	Total number of somatic SNV and indels detected in the sample at an AF of >=5% that are also present in the 1000 Genomes, ExAC and ESP populations at <=1%.
Low pop somatic variants in COSMIC	Total number of somatic SNV and indels detected in the sample at an AF of >=5% that are also present in the population at <=1% and present in COSMIC.
Low pop, damaging, and in Personalis Research Cancer Gene List.	Total number of somatic SNV and indels detected in the sample at an AF of >=5% that are also present in the population at <=1% and fall in Personalis Research Cancer Gene List.
Low pop, damaging, and in Personalis Clinical Cancer Gene List.	Total number of somatic SNV and indels detected in the sample at an AF of >=2% that are present in the population at <=1% and fall in the Personalis Clinical Cancer Gene List. (>1,000X coverage in 248 cancer driver genes confers greater sensitivity to variants >=2%.)

Results Summary (QC) Report for Tumor RNA Transcriptome

Sample and Run Information

Display Name	Definition	Additional Details and Interpretation
Analysis Mode	Tumor/Normal	Tumor RNA Transcriptome is interpreted in the context of the DNA Normal Exome as a matched sample to filter out possible germline and contamination from somatic SNV.
Tumor sample	Name of tumor sample used for analysis.	RNA Tumor sample used for transcriptome
Matched normal sample	Name of matched normal sample used for analysis.	DNA Normal sample used for exome
Pipeline versions	Version of the Personalis pipeline run for the sample	
Annotation version	Version of the Personalis Annotation used for the data analysis	
Platform version	Personalis Lab Assay version	
Reference assembly	Reference assembly used	

DRAFT

DRAFT

Sequencing Information

The Sequencing Information section includes a summary of basic sequencing statistics generated for the Tumor RNA sample.

Display Name	Definition	Additional Details and Interpretation
Read length (bp)	Number of bases in a read.	If a sample has reads of different lengths, the lengths will be listed as a comma delimited set.
Total read pairs	Total number of read pairs that were sequenced and passed fastq filters.	This number is proportional to sequence coverage. Higher numbers per sample lead to better variant calling.
Total bases	Total number of bases sequenced.	This number is proportional to sequence coverage. Higher numbers per sample lead to better variant calling.
Average base quality	Mean base quality score for reads that pass filter. Calculated after GATK base recalibration.	This value should be >30 (Q30) for a good run. Lower values indicate systematic sequencing problems.

QC Plots for Tumor RNA Transcriptome

Quality Control graphs are included to visually display common metrics of run quality for each read across both the Normal and Tumor DNA Samples:

- Quality Scores by Read Position
- Average Quality per Read
- GC Distribution Over All Sequences

Alignment Information

This section includes a summary of Star Alignment Metrics generated for each sample.

Display Name	Definition	Additional Details and Interpretation
Number of reads post rRNA removal	The number of reads which remain once rRNA reads are removed.	rRNA reads create noise in the data and should not be used in transcriptome analysis.
Number of reads that map to the reference genome	Total number of reads that mapped to the reference genome.	Only main chromosomes are considered [1-22, X, Y, MT]
Percent mapped read pairs	Percentage of all read pairs that map to reference assembly.	Based on all valid read pairs.
Average mapped read pair length	Average length of mapped read pairs.	This is the mean summed length of the left and right read pairs (if paired-ended) or single read (if single-ended).
Number of total splice sites	The number of splice sites identified by STAR aligner.	This is the total number of splice sites which were observed in any reads by the alignment algorithm.
Number of annotated splice sites	The number of splice sites which were known from gene annotation file that were observed in the reads	This number is dependent on the gene annotation file. Splice sites which have been previously observed are more likely to be true positives.
Number GT-AG splice sites	The number of GT-AG splice sites which were observed in the reads.	The GT-AG splice site is considered canonical and accounts for the vast majority of splice site sequences (99%).
Number GC-AG splice sites	The number of GC-AT splice sites which were observed in the reads.	The GC-AG splice site will be observed far less than GT-AG sites.
Number AT-AC splice sites	The number of GC-AT splice sites which were observed in the reads.	The GC-AT splice site will be observed far less than GT-AG sites.
Number of non-canonical splice sites	The number of non-canonical splice sites which were observed in the reads.	Non-canonical splicing is performed by a minor spliceosome (non-U2-dependant splicing)
Mismatch rate per base	The rate at which mismatches occur in aligned reads.	This metric indicates what percentage of the time a single base in an aligned does not match the reference genome.
Deletion rate per base	The rate at which deletions occur in aligned reads.	This metric indicates what percentage of the time a deletion has occurred in a read compared to the reference genome.
Mean deletion length	The mean deletion length that occurs in aligned reads.	This metric indicates how long the average deletions are that occur in reads compared to the reference genome.
Insertion rate per base	The rate at which insertions occur in aligned reads.	This metric indicates what percentage of the time an insertion has occurred in a read compared to the

		reference genome.
Mean insertion length	The mean insertion length that occurs in aligned reads.	This metric indicates how long the average insertions are that occur in reads compared to the reference genome.

Mapping Occurrence Metrics

This is a summary of mapping occurrence metrics, which describes the number of times an individual read pair maps to the reference assembly, for the Tumor RNA sample. This includes mapping occurrence, counts, and percentage of that class and above represents of the total read count.

Mapping Chromosome Metrics

This table describes the number of reads and percentage of total mapped read pairs from this sample that align to each chromosome.

Mapping Gene Element Metrics

This section provides a table describing how reads map relative to gene elements.

Display Name	**Definition**
Exon	Reads mapping to exonic regions
Intron	Reads mapping to intronic regions
Intergenic	Reads which map to regions in-between gene elements
Promoter	Reads mapping to promoter regions
UTR-3	Reads which map to 3' UTR regions
UTR-5	Reads which map to 5' UTR regions

Anomalous Read Pair Alignments

Display Name	**Definition**	**Additional Details and Interpretation**
Inter-chromosomal	Number of read pairs for which one read maps to one chromosome and whose paired read maps to another chromosome.	Unusually high values could indicate sample prep problems in library creation, especially with mate pairs. Artifacts such as chimeric reads can lead to elevated values.
Orphaned reads	The number of reads mapped to the reference sequence where the paired read does not map.	A simple metric to identify gross failures during sequencing, such as bad base calls on all read 2's for a specific flowcell. This category should include less than 1% of all reads. In a healthy run, orphaned reads occur due to natural features of a genome such as repeats or structural variation.

QC Plots for the Tumor RNA Transcriptome

A Read Depth Profile shows a histogram by depth of coverage with the fraction of target with Depth > X, where X is in number of reads that span that nucleotide location in the reference assembly.

A Transcript Coverage Plot presents the mean read coverage across all annotated transcripts. For this plot, the transcript distance has been normalized to represent a percentage of distance through all transcripts from 5' to 3'. The y-axis represents the mean normalized coverage considering all annotated transcripts. A perfectly flat line crossing the entire plot at 1, would represent that the coverage is perfectly even across each base position in all transcripts. However, as the coverage at the beginning and end of transcripts is always lower than the middle, due to sequencing initiation and termination, there is always a bowed appearance. A long region of higher normalized coverage symbolizes more even sequencing coverage.

Variant Calling

The Variant Calling Statistics section includes detailed statistics generated during the variant calling stage of the Personalis Pipeline analysis. A series of statistics are displayed that indicate various aspects of variant calling performed.

Summary Small Variants for Normal DNA Exome and Tumor RNA Transcriptome

Display Name	**Definition**	
Total Calls	Total number of small variants detected	Low indicates not enough coverage.

DRAFT

Calls in a public database	Number of SNPs detected that were in dbSNP	Low values mean that the sequenced individual has many novel SNPs which is commonly observed in cancer samples or when sequencing an individual from a population with unusual genetic diversity.
Heterozygous/ Homozygous ratio	Ratio of heterozygous to homozygous SNPs	Abnormal values are associated with low sequence coverage.
Ti/Tv ratio	Ratio of transition (Ti) to transversion (Tv) substitutions for a set of SNPs.	This metric is useful for quality control. A low value is generally associated with high false positive rate. A high value indicates lots of false negatives, which can occur when variant call filtering is overly stringent. Expect a range depending upon platform.

SNV Statistics for Normal and Tumor DNA Exomes

A summary table shows the counts of SNVs that change from each Reference Base to each Alternate Base for each of the Normal and Tumor samples independently.

Tumor/Normal Concordance

Display Name	Definition	Additional Details
Variants called in both Normal DNA and Tumor RNA	Total number of variants common to both Normal DNA and Tumor RNA samples	The total number of variants that are present in both the tumor and normal samples
Variants unique to Normal DNA	Number of DNA variants unique to the Normal sample	The total number of variants that are present in normal but not in the tumor sample
Variants unique to Tumor RNA	Number of RNA variants unique to the Tumor sample	The total number of variants that are present in tumor but not in the normal sample

Somatic Variant Calling and Annotation

Summary Small Variants

Display Name	Definition
Somatic variants	SNVs, indels
Somatic variants per Mb	Number of somatic variants per Mega Base of DNA
Non-synonymous Somatic Variants	Total number of non-synonymous somatic mutations (i.e. mutations that alter the amino acid)
Non-synonymous Somatic Variants per Mb	Number of non-synonymous somatic mutations (i.e. mutations that alter the amino acid) per megabase of DNA. This number is commonly used to describe tumor mutational burden.
Ti/Tv ratio	Ratio of number of transitions to number of transversions for detected SNVs. See above.

SNV Statistics for Somatic DNA variants

A summary table shows the counts of somatic SNVs that change from each Reference Base to each Alternate Base.

Mutation Signature Plots

A set of plots are provided to show the percent of each type SNV change within the context of its neighboring nucleotides.

Functional Annotation

Display Name	Definition
Nonsense	Number of variants that change a codon to a stop codon
Missense	Number of variants that change a codon to produce a different amino acid
Silent	Number of variants that do not result in the change to the amino acid
None	Number of variants that do not have a functional annotation of nonsense, missense, or silent

Effect Impact Annotation – Somatic RNA

The list of effect types that have a given impact is included in the Appendix.

Display Name	Definition
High	Count of the somatic RNA variants with High impact
Moderate	Count of the somatic RNA variants with Moderate impact
Low	Count of the somatic RNA variants with Low impact
Modifier	Count of the somatic RNA variants with Modifier impact

DRAFT

Effect Annotation

Display Name	Definition
Codon Change plus codon deletion	Number of variants that cause a single codon change resulting in one or more codon deletions, such as when deletion of whose size is a multiple of three that occurs within a codon (not at the boundary).
Codon insertion	Number of variants that cause one or more codons to be inserted
Downstream	Number of variants that are downstream of an annotated gene
Frame shift	Number of variants (insertions or deletions) that shifts the reading frame
Intragenic	Number of variants that occur within a gene, but no transcripts are annotated for the gene
Intron	Number of variants that occur in an intron
Non-coding exon	Number of variants that occur in the non-coding exons
Nonsynonymous coding	Number of variants that cause a codon change that results in an amino acid change
Other	Number of variants that do not have another listed effect annotation
Splice site acceptor	Number of variants that occurs in the splice acceptor site (defined as two bases before the exon start, except for the first exon)
Splice site donor	Number of variants that occurs in the predicted splice donor site (two bases after coding exon end, except for the last exon)
Splice site region	Number of variants within the region of the splice site, either within 1-3 bases of the exon or 3-8 bases of the intron
Start gained	Number of variants in 5'UTR region that produces a three-base sequence that can be a START codon
Stop gained	Number of variants that causes a stop codon
Synonymous coding	Number of variants that causes a codon change that does NOT result in an amino acid change
Upstream	Number of variants that occur within 5000 bases upstream of a gene
UTR 3 prime	Number of variants that occur in the 3' untranslated region (3' UTR)
UTR 5 prime	Number of variants that occur in the 5' untranslated region (5'UTR)

Predicted Effect

Display Name	Definition
Mutation Taster	Mutation Taster predicts the disease-causing potential of the variants and classifies them into 4 categories: Disease Causing Automatic, Disease Causing, Polymorphism, Polymorphism Automatic
LRT	Likelihood Ration Test (LRT) method that identifies deleterious mutations.

Somatic Cancer Annotation

Mutational Effect Impact

Display Name	Definition
All somatic variants	Number of somatic variants that have different mutational impacts (high, moderate, low, and modifier)
All somatic variants in Personalis Research Cancer Gene List.	Number of somatic variants that have different mutational impacts (high, moderate, low, and modifier) in the Personalis Research Cancer Gene List.

Variant Filtering

Display Name	Definition
All Somatic variants	Number of somatic variants that have an allele fraction (AF) greater than 5%
Low population frequency somatic variants	Total number of somatic SNV and indels detected in the sample at an AF of >=5% that are also present in the 1000 Genomes, ExAC and ESP populations at <=1%.
Low pop somatic variants in COSMIC	Total number of somatic SNV and indels detected in the sample at an AF of >=5% that are also present in the population at <=1% and present in COSMIC.
Low pop, damaging, and in Personalis Research Cancer Gene List.	Total number of somatic SNV and indels detected in the sample at an AF of >=5% that are also present in the population at <=1% and fall in Personalis Research Cancer Gene List.
Low pop, damaging, and in Personalis Clinical Cancer Gene List.	Total number of somatic SNV and indels detected in the sample at an AF of >=2% that are present in the population at <=1% and fall in the Personalis Clinical Cancer Gene List. (>1,000X coverage in 248 cancer driver genes confers greater sensitivity to variants >=2%.)

RNA Fusion Metrics

Fusion Discovery Summary

Display Name	Definition	Additional Details and Interpretation
Number of total fusions found	The total number of final filtered fusions identified in this sample	This represents the number of fusions which passed fusion filtering. Unlike the preliminary list, these fusion events are more likely to be true positives.

DRAFT

Number of unique fusion site pairs found	The number unique sets of final filtered fusions identified in this sample	As fusion events can be identified by either BOWTIE or BLAT based approaches, some events will be identified by both and will be listed twice in the final set. The number takes this into consideration and does not count these events twice.
Number of clinical gene list fusions found	The number of fusions which have known clinical relevance	Genes present in the Personalis Clinical Cancer Gene List.
Number of fusions previously observed in healthy samples	The number of fusions that have been observed before in healthy samples	Fusion events that have been observed in healthy samples are more likely to be false positives. These events have been seen in at least one healthy sample, and should be more closely scrutinized.
Number of fusions previously observed in TCGA database	The number of fusions that have been observed in the TCGA database	Genes present in the TCGA database.

Fusion Supporting Reads Summary

Display Name	Definition	Additional Details and Interpretation
Mean number of spanning read pairs per fusion	The mean number of reads pairs or clusters which bridge the fusion event.	Paired-end reads map to separate locations in the genome. When they map to distant regions, they can bridge fusion events. This count represents the number of read pairs which bridge a fusion event.
Mean number of unique spanning reads per fusion	The mean number of unique individual reads which split a fusion event.	Individual reads can map uniquely to two different locations in the genome, for example exon junctions. However, when these positions are distantly apart and on separate genes they can represent fusion events. This count represents the number of individual reads in which one portion of the read maps to one gene and the other potion to a different gene.
Mean longest anchor found per fusion	The mean length of the longest anchor event observed for each fusion event.	A fusion event anchor is the sequence which uniquely identifies a portion of a read to a particular location in the genome. The mean longest anchor would be the mean value across all final identified fusion events for the longest uniquely mapping read.

Fusion Distance Metrics Summary

Display Name	Definition	Additional Details and Interpretation
Number of short Intrachromosomal fusions (less than 100 kb)	The number of fusion events which occur in the same chromosome and span less than 100 kb.	When both partners of a fusion event occur on the same chromosome, the events are intrachomosomal.
Number of short Intrachromosomal fusions (more than 100 kb)	The number of fusion events which occur in the same chromosome and span more than 100 kb.	When both partners of a fusion event occur on the same chromosome, the events are intrachomosomal.
Number of interchromosomal fusions	The number of fusion events which occur on different chromosomes.	When both partners of a fusion event occur on different chromosomes, the events are interchomosomal.

Gene Expression Metrics

Display Name	Definition	Additional Details and Interpretation
Number of expressed genes	Number of expressed genes	TPM ≥ 2 is the cutoff for determining expression.
Number of unexpressed genes	Number of unexpressed genes	TPM ≥ 2 is the cutoff for determining expression.
Number of expressed cancer relevant genes	Number of expressed cancer relevant genes	Expressed and present in the Research Cancer Gene List.
Number of expressed clinical cancer genes	Number of expressed clinical cancer genes	Expressed and present in the Clinical Cancer Gene List.

Attribution

Further details on some of the databases and tools used in the analysis pipeline are described here.

1. MutationTaster
   Schwarz JM, Rödelsperger C, Schuelke M, Seelow D.

DRAFT

MutationTaster evaluates disease-causing potential of sequence alterations.
Nature Methods. 2010 Aug;7(8):575-6.

2. dbNSFP
    a. http://varianttools.sourceforge.net/Annotation/DbNSFP
    b. Liu X, Jian X, and Boerwinkle E.
    dbNSFP: a lightweight database of human non-synonymous SNPs and their functional predictions.
    Human Mutation. 2011 32:894-899
    c. Liu X, Jian X, and Boerwinkle E.
    dbNSFP v2.0: A Database of Human Nonsynonymous SNVs and Their Functional Predictions and Annotations.
    Human Mutation. 2013 34:E2393-E2402.

DRAFT

Appendix

Effect and Impact Definitions

Variant_Effect and Variant_Effect_Impact column annotations use the following definitions for effect and impact of the variant.

Effect	Impact	Note
3_prime_UTR_variant	MODIFIER	Variant occurs in the 3' untranslated region (3' UTR)
5_prime_UTR_premature_start_codon_gain_variant	LOW	A variant in 5'UTR region produces a three-base sequence that can be a START codon.
5_prime_UTR_variant	MODIFIER	Variant occurs in the 5' untranslated region (5' UTR)
disruptive_inframe_deletion	MODERATE	One codon is changed and one or more codons are deleted Example: A deletion of whose size is a multiple of three, that occurs within a codon
disruptive_inframe_insertion	MODERATE	One codon is changed and one or many codons are inserted Example: An insertion whose size is a multiple of three, that occurs within a codon
downstream_gene_variant	MODIFIER	Variant occurs downstream of a gene (default length: 5,000 bases)
frameshift_variant	HIGH	Insertion or deletion that causes a frame shift, i.e., size is not a multiple of 3
inframe_deletion	MODERATE	One or many codons are deleted, such as a 3-base deletion at a codon boundary
inframe_insertion	MODERATE	One or many codons are inserted, such as an insertion whose size is a multiple of three that occurs at a codon boundary
initiator_codon_variant	LOW	Variant causes the start codon to be mutated into an alternative start codon (the new codon produces a different amino acid).
intron_variant	MODIFIER	Variant occurs in an intron. Technically, hits no exon in the transcript.
missense_variant	MODERATE	Variant causes a codon change that results in an amino acid substitution
non_coding_exon_variant	LOW	Variant occurs in a non-coding portion of the exon
splice_acceptor_variant	HIGH	The variant occurs in the predicted splice acceptor site (defined as two bases before exon start, except for the first exon).
splice_donor_variant	HIGH	The variant occurs in the predicted splice donor site (defined as two bases after coding exon end, except for the last exon).
splice_region_variant	LOW	A sequence variant in which a change has occurred within the region of the splice site, either within 1-3 bases of the exon or 3-8 bases of the intron.
start_lost	HIGH	Variant causes start codon to be mutated into a non-start codon.
stop_gained	HIGH	Variant causes a STOP codon
stop_lost	HIGH	Variant causes stop codon to be mutated into a non-stop codon
stop_retained_variant	LOW	Variant causes stop codon to be mutated from one stop codon into another
synonymous_variant	LOW	Variant causes a codon change but does not result in an amino acid substitution
upstream_gene_variant	MODIFIER	Variant occurs upstream of a gene (default length: 5,000 bases)

Definition of terms used in the Small Variant QC section

Display Name	Definition	Additional Details and Interpretation
Analysis Mode	Type of somatic analysis performed: Tumor/Normal	Tumor/Normal mode uses the matched normal sample provided to filter out possible germline and contamination from somatic SNV calls.
Matched Normal	Name of matched normal sample used for analysis	
Somatic Variants	Total number of somatic SNV and indels detected	This value can vary across tumor types and individual tumors. Some tumors exhibit a hypermutator phenotype whereas others can be mutationally-silent in terms of small variants.
Frequency of Somatic Variant per Mb	Frequency of somatic variant per megabase	In general, cancers are thought to have a mutation rate of 1 per Mb, but individual tumors deviate from this depending on their specific etiology and genomic profile.
Somatic SNVs	Total number of somatic SNVs detected	
Somatic Indels	Total number of somatic Indels detected	Somatic Indel detection remains an active area of research in the community due to a number of reasons, including the inherent ambiguity between genomic regions prone to both sequencing error as well as true somatic mutation.
Transitions	Total number of base transitions	Transitions is a type of mutation where a purine is changed to another purine or a pyrimidine is changed to another pyrimidine.
Transversions	Total number of base transversions	Transversion is a type of mutation where a purine is changed to a pyrimidine or vice versa.

DRAFT

SNV Transition/Transversion Ratio	Ratio of number of transitions to number of transversions for detected SNVs.	This value can vary across tumor types and individual tumors. Exogenous factors, such as UV-radiation and tobacco, can inflate and deflate this number depending on their mutational signatures. See Mutational Signature Plot for further details.
Percent Contamination	Percentage of predicted contamination by other samples	This is the mean predicted contamination across all chromosomes. This value is important to note if sample contamination is a concern. This does not include normal-in-tumor contamination. See the description of ConTest for more information.
Tumor and Matched Normal Concordance	Comparison of the number of SNV/Indels detected in tumor and normal sample individually	The number of variants called that are common to both tumor and normal samples confirms the single-individual origin of the two samples. If there is high discordance between the two, it would suggest that the tumor and normal samples analyzed were extracted from two different individuals.
Mutational Signature Plot	Histogram of single base mutations on background of adjacent bases	Tumors contain specific base mutations, where even the two adjacent bases surrounding the mutated base affect the frequency of mutation. Specific signatures in this context are often observed for exogenous factors, such as UV-radiation and tobacco smoke.

Definition of terms used in the Variant Annotation section

Display Name	**Definition**	**Additional Details and Interpretation**
Functional Class	Table showing the number of somatic SNVs falling in each type of functional annotation.	If the variant is present in a protein-coding gene region, functional annotation for the variant (e.g. "MISSENSE", "NONSENSE", etc.) is counted. See the AnnoL description for more information on Functional Class annotation.
Variant Effect	Table showing the number of somatic SNV and Indels falling in each category of variant effect.	If no protein effect is predicted, the relative location of the variant within the gene or genomic element is given (e.g. Downstream, Intron, Exon, etc.). If a protein effect is predicted this prediction will be listed (e.g. Nonsynonymous coding etc.). For a full list of the possible effects in this column see the "Effect and Impact Definitions" section below.
Predicted Mutation Effect	Number of SNV/Indels in each effect category	Table showing counts of SNV and Indel effects as assessed by three different tools.

Gene Fusion Column Definitions

Column Name	**Data Type**	**Annotation Type**	**Description**
Gene1 NCBI GeneID	Character string	Location	NCBI Gene ID for 5' fusion partner
Gene1 (5' fusion partner)	Character string	Location	Gene symbol for 5' fusion partner
Gene2 NCBI GeneID	Character string	Location	NCBI Gene ID for 3' fusion partner
Gene2 (3' fusion partner)	Character string	Location	Gene symbol for 3' fusion partner
Predicted fusion effect	Character string	Location	See Fusion Annotation Definitions table below
Fusion point for Gene1	Character string	Location	Chromosomal position of the 5' end of fusion junction (chromosome:position:strand)
Fusion point for Gene2	Character string	Location	Chromosomal position of the 3' end of fusion junction (chromosome:position:strand)
Fusion description	Character string	Description	Fusion gene annotations (Values described in detail below)
Predicted fused protein	Character string	Description	The inferred fusion junction (the asterisk sign marks the junction point)
Method	Character string	Origin	Aligning method used for mapping the reads and finding the fusion genes. Here are two methods used which are: i) BOWTIE: only BOWTIE aligner is used for mapping the reads on the genome and exon-exon fusion junctions, and ii) BOWTIE+BLAT: BOWTIE aligner is used for mapping reads on the genome and BLAT is used for mapping reads for finding the fusion junction.
Common mapping reads (count)	Integer	Coverage	Count of reads mapping simultaneously on both genes which form the fusion gene. This is an indication how similar are the DNA/RNA sequences of the genes forming the fusion gene (i.e. what is their homology because highly homologous genes tend to appear show as candidate fusion genes). In case of completely different sequences of the genes involved in forming a fusion gene then here it is expected to have the value zero.

DRAFT

Spanning pairs	Integer	Coverage	Count of pair-end reads supporting the fusion
Spanning unique reads	Integer	Coverage	Count of unique reads (i.e. unique mapping positions) mapping on the fusion junction
Longest anchor found	integer	Coverage	Longest anchor (hangover) found among the unique reads mapping on the fusion junction

Fusion Annotation Definitions

Fusion description	**Data Type**	**Description**
antisense	Character string	One or both genes is a gene coding for antisense RNA
banned	Character string	Fusion gene is on a list of known false positive fusion genes. *A candidate fusion gene having this label has a very high probability of being a false positive.*
chimerdb2	Character string	Known fusion gene from the ChimerDB database (please use ChimerDB2 database for more information regarding the fusion gene)
conjoing	Character string	Known conjoined genes (that is fusion genes found in samples from healthy patients) from the ConjoinG database (please use ConjoinG database for more information regarding the fusion gene). *A candidate fusion gene having this label has a very high probability of being a false positive in case that one looks for fusion genes specific to a disease.*
cosmic	Character string	Known fusion gene from the COSMIC database (please use COSMIC database for more information regarding the fusion gene)
cacg	Character string	Known conjoined genes (that is fusion genes found in samples from healthy patients) from the CACG database (please see CACG database for more information). *A candidate fusion gene having this label has a very high probability of being a false positive in case that one looks for fusion genes specific to a disease.*
cgp	Character string	known fusion gene from the CGP database (please use CGP database for more information regarding the fusion gene)
ctd_gene	Character string	one gene or both genes is CTD gene (that is that the gene name starts with CTD-). *A candidate fusion gene having this label has a very high probability of being a false positive.*
distance1000bp	Character string	both genes are on the same strand and they are less than 1 000 bp apart. *A candidate fusion gene having this label has a very high probability of being a false positive.*
distance10kbp	Character string	both genes are on the same strand and they are less than 10 000 bp apart. *A candidate fusion gene having this label has a higher probability than expected of being a false positive.*
distance100kbp	Character string	both genes are on the same strand and they are less than 100 000 bp apart. *A candidate fusion gene having this label has a higher probability than expected of being a false positive.*
duplicates	Character string	Both genes involved in the fusion gene are paralog for each other. For more see duplicated genes database (DGD). *A candidate fusion gene having this label has a higher probability that expected of being a false positive.*
ensembl_fully_ overlapping	Character string	the genes forming the fusion gene are fully overlapping according to Ensembl database. *A candidate fusion gene having this label has a very high probability of being a false positive.*
ucsc_fully_ overlapping	Character string	the genes forming the fusion gene are fully overlapping according to UCSC database. *A candidate fusion gene having this label has a very high probability of being a false positive.*
refseq_fully_ overlapping	Character string	the genes forming the fusion gene are fully overlapping according to RefSeq NCBI database. *A candidate fusion gene having this label has a very high probability of being a false positive.*
healthy	Character string	fusion gene has been seen in a healthy sample. *A candidate fusion gene having this label has a very high probability of being a false positive in case that one looks for fusion genes specific to a disease.*
known_fusion	Character string	known fusion gene which has been previously published (i.e. it is not a novel fusion gene). Publications were mined from literature.
matched-normal	Character string	candidate fusion gene (which is supported by paired reads mapping on both genes and also by reads mapping on the junction point) was found also in the matched normal sample given as input to the command line option '--normal'
partial-matched-normal	Character string	candidate fusion gene (which is supported by paired reads mapping on both genes but *no* reads were found which map on the junction point) was found also in the matched normal sample given as input to the command line option '--normal'. This is much weaker than matched-normal.
lincrna	Character string	one or both genes is a lincRNA
Mirna	Character string	one or both genes is a miRNA
Mt	Character string	one or both genes are situated on mitochondrion. *A candidate fusion gene having this label has a very high probability of being a false positive.*
no_protein_product	Character string	one or both genes have no known protein product

DRAFT

oncogene	Character string	one gene or both genes are a known oncogene
pair_pseudo_genes	Character string	one gene is the other's pseudogene. *A candidate fusion gene having this label has a very high probability of being a false positive.*
paralogs	Character string	both genes involved in the fusion gene are paralog for each other (most likely this is a false positive fusion gene). *A candidate fusion gene having this label has a very high probability of being a false positive.*
ensembl_partially_ overlapping	Character string	the genes forming the fusion gene are partially overlapping (on same strand or on different strands) according the Ensembl database.
ucsc_partially_ overlapping	Character string	the genes forming the fusion gene are partially overlapping (on same strand or on different strands) according the UCSC database.
refseq_partially_ overlapping	Character string	the genes forming the fusion gene are partially overlapping (on same strand or on different strands) according the RefSeq NCBI.
pseudogene	Character string	one or both of the genes is a pseudogene
readthrough	Character string	Whether the fusion gene is a read-through event (that is both genes forming the fusion are on the same strand and there is no known gene situated in between); Please note that many of read-through fusion genes might be false positive fusion genes due to errors in Ensembl database annotation (for example, one gene is annotated in Ensembl database as two separate genes). *A candidate fusion gene having this label has a high probability of being a false positive.*
ribosomal_protein	Character string	one or both gene is a gene encoding for ribosomal protein
rp11_gene	Character string	one gene or both genes is RP11 gene (that is that the gene name starts with **RP11-**). *A candidate fusion gene having this label has a very high probability of being a false positive.*
rp_gene	Character string	one gene or both genes is RP?? gene (that is that the gene name starts with **RP??-**) where '?' is a digit. *A candidate fusion gene having this label has a very high probability of being a false positive.*
Rrna	Character string	one or both genes is a rRNA. *A candidate fusion gene having this label has a very high probability of being a false positive.*
ensembl_same_ strand_overlapping	Character string	the genes forming the fusion gene are fully/partially overlapping and are both on the same strand according to Ensembl database. *A candidate fusion gene having this label has a very high probability of being a false positive (this is most likely and alternative splicing event).*
ucsc_same_strand_ overlapping	Character string	the genes forming the fusion gene are fully/partially overlapping and are both on the same strand according to UCSC database. *A candidate fusion gene having this label has a very high probability of being a false positive (this is most likely and alternative splicing event).*
refseq_same_strand_ overlapping	Character string	the genes forming the fusion gene are fully/partially overlapping and are both on the same strand according to RefSeq NCBI database. *A candidate fusion gene having this label has a very high probability of being a false positive (this is most likely and alternative splicing event).*
short_distance	Character string	both genes are on the same strand and they are less than X bp apart, where X is set using the option '--dist-fusion' and by default it is 200 000 bp. *A candidate fusion gene having this label has a higher probability than expected of being a false positive.*
similar_reads	Character string	both genes have the same reads which map simultaneously on both of them (this is an indicator of how similar are the sequences of both genes; ideally this should be zero or as close to zero as possible for a real fusion). *A candidate fusion gene having this label has a very high probability of being a false positive.*
similar_symbols	Character string	both genes have the same or very similar gene names (for example: RP11ADF.1 and RP11ADF.2). *A candidate fusion gene having this label has a very high probability of being a false positive.*
Snorna	Character string	one or both genes is a snoRNA
Snrna	Character string	one or both genes is a snRNA
Tcga	Character string	Known fusion gene form the TCGA database
Ticdb	Character string	known fusion gene from the TICdb database (please use TICdb database for more information regarding the fusion gene)
Trna	Character string	one or both genes is a tRNA
Yrna	Character string	one or both genes is a Y RNA

DRAFT

Cancer Annotation

Display Name	Definition	Additional Details and Interpretation
Somatic Variants in COSMIC	Total number of somatic SNV and indels detected that are present in the COSMIC database.	The number of variants that have previously been identified and collated in the COSMIC database can give a sense of commonly observed mutations in cancer; some of these, however, have been seen as germline or passenger events and may not have contributed to tumorigenesis.
Somatic Variants in Cancer Gene Census Genes	Total number of Cancer Gene Census genes that contain detected somatic SNVs/Indels.	Cancer Gene Census genes have been curated as particularly important in cancer biology; the more of these canonical genes that are affected by mutation, the more likely it is that driver mutations are identified.
Somatic Variants in Genes with Drug Associations	Total number of genes that have DrugBank associations that contain detected somatic SNVs/Indels.	Using the DrugBank database, drugs that are associated with particular mutated genes in this sample are noted. One drug may be associated with multiple genes. This list is not limited to cancer therapeutics.
Mutation Effect Impact	Table showing number of variants falling in each category of mutational impact, for both all genes and Cancer Gene Census genes.	For genes containing a somatic mutation within the given sample, their mutation effect impacts are categorized and counted. The same information is displayed for Cancer Gene Census genes. The number of HIGH-impact mutations, especially within Cancer Gene Census genes, can give an estimate of the number of possible driver mutations observed.
Low Population Frequency Variants	Total number of somatic SNV and indels detected that are present in the population at <=1%.	Rare variants, present at very low or no frequencies in the population, are more commonly implicated in tumor progression. Frequency across populations is assessed using large population-based studies, such as 1000Genomics and ESP.
Low Population Frequency Variants in COSMIC	Total number of somatic SNV and indels detected that are present in the population at <=1% and present in COSMIC.	Rare variants (see above) that have been previously described in the COSMIC database are likely to be of some importance in cancer.
Low Population Frequency Variants in Cancer Gene Census Genes	Total number of somatic SNV and indels detected that are present in the population at <=1% and fall in Cancer Gene Census genes.	Rare variants present in Cancer Gene Census genes are more likely to be of direct relevance to cancer.
Low Population Frequency Variants in Personalis Cancer Genes	Total number of somatic SNV and indels detected that are present in the population at <=1% and fall in Personalis Cancer Genes.	Rare variants present in the expanded Personalis Cancer Gene database genes are more likely to be of relevance to cancer, either directly or through investigational mechanisms.

DRAFT