illumina®

Variant filtering in TruSight™ Software Suite

Gain insight into the extensive filtering options available in TruSight Software Suite to obtain a prioritized list of variants of interest.

Introduction

Whole-genome sequencing (WGS) enables a high-resolution view across the entire genome, making it a useful method for discovering causative variants of inherited disorders. However, the vast amounts of data produced by WGS present a significant bottleneck and require comprehensive data analysis tools that can efficiently translate the raw sequencing data into meaningful, interpretable results. To address this challenge, Illumina offers TruSight Software Suite, a software as a service (SaaS) analytics solution that integrates with the NovaSeq™ 6000 System and the cloud-based DRAGEN™ Bio-IT Platform to enable comprehensive, streamlined secondary and tertiary analysis workflows for next-generation sequencing (NGS) studies.

TruSight Software Suite offers a variety of default variant filtering options that can be customized to meet each user's needs. Test-level filters enable a streamlined workflow and users can add ad hoc filters while interpreting a case. This technical note outlines important considerations selecting filters in TruSight Software Suite.

Quality control: mapping and sample metrics

TruSight Software Suite provides a rich set of quality control (QC) metrics that flag samples with lower than expected quality and identify inconsistencies in the family-based analysis (Table 1). QC thresholds are set in the test definition in TruSight Software Suite. Default QC thresholds are provided and can be customized by the lab director. Sample-level QC metrics are evaluated for each proband or familial relation. In duo or trio cases, QC failures will need to be reviewed before the familial joint-genotyping analysis can begin. After the completion of familial joint-genotyping analysis, family-based metrics are evaluated. QC failures at this level will also need to be reviewed before the case transitions to "Ready for Interpretation" state. For additional details about DRAGEN joint-based calling, please see the DRAGEN Bio-IT User Guide.

In the event of a QC failure, a QC warning is provided, and the case manager can elect to QC Override the failure or Modify the sample information listed in the case (Figure 1). The QC Override option should be used if the failure is deemed acceptable. Lower than expected coverage could be considered acceptable, provided that the minimum coverage (10×) supported by the system is satisfied. As a best practice, a minimum of 30× coverage is recommended. The case manager can also elect to modify the case, where appropriate. This may include sample swaps, sample contamination, or the availability of additional sequence data. QC failures pertaining to trio concordance and sex ploidy validation should not be bypassed and require the secondary analysis pipeline to be restarted with the correct sample information. Once the case is Ready for Interpretation, TruSight Software Suite allows the case manager to filter on variants that passed QC filters in the DRAGEN platform. It is recommended that all filter schemas include the Quality filter setting, as only high-quality variants should be evaluated and reported.

Table 1: Secondary analysis QC metrics

QC metric	Default threshold
Sample-level QC metrics[a]	
Average autosomal coverage over genome	≥ 30
Mapped reads (%)	≥ 90
Q30 bases (%)	≥ 80
Q30 bases R1 (%)	≥ 85
Q30 bases R2 (%)	≥ 75
Mismatched bases R1 (%)	≤ 1.25
Mismatched bases R2 (%)	≤ 1.5
Insert length median	≥ 100
Paired reads mapped to different chromosomes (MAPQ ≥ 10) (%)	≤ 10
Autosome callability (%)	≥ 90
AT dropout	≤ 10
GC dropout	≤ 10
Normalized coverage at GCs 80-100	≥ 1
SNVs in large runs of homozygosity (≥ 3 Mb)	≥ 2
Sex ploidy validation	TRUE
Estimated sample contamination[b]	≤ 0.01
Pedigree-level QC metrics[c]	
Het/hom ratio	≤ 2.05
Ti/Tv ratio	≤ 2.5
Trio concordance (%)	≥ 99

a. Sample-level QC metrics summarize the single sample germline analysis.
b. Pedigree-level QC metrics are derived from the joint genotyping analysis.
c. Metric is optionally enabled and is not provided in the default set of metrics.

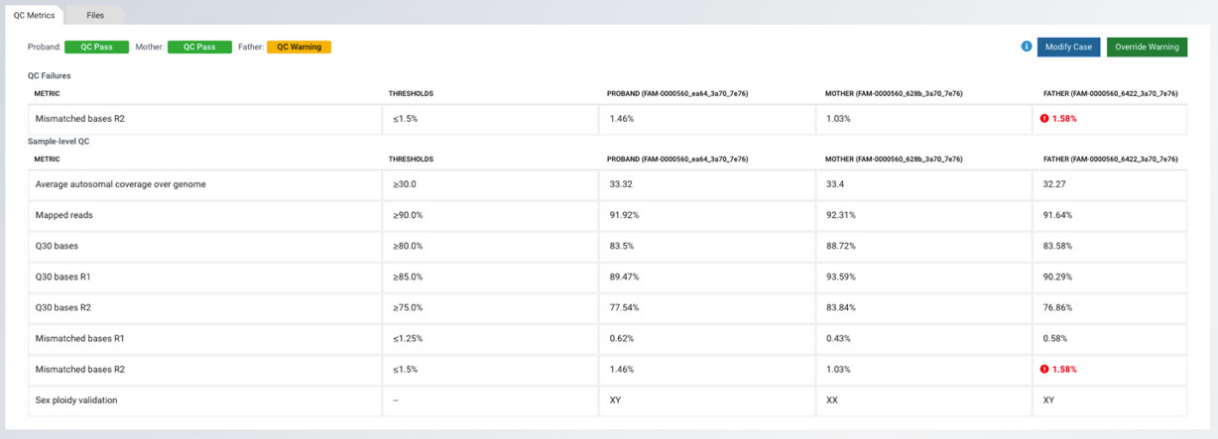
Filtering best practices

In TruSight Software Suite, users have great flexibility in filtering. Consistent filters can be applied to every sample by locking in a defined filter scheme during test configuration. Users can apply filters ad hoc while interpreting individual cases.

Family-based analysis

Best practice for performing analysis on a pedigree of related individuals includes variant calling using the DRAGEN Joint Genotyping Pipeline. This algorithm takes the variant calls from gVCFs produced for each family member and jointly genotypes them into a single gVCF file. Incorporating the variant information from each family member increases the genotype sensitivity and specificity of the proband.

TruSight Software Suite allows users to filter variants following the pattern for several modes of inheritance (MOI) such as *de novo*, dominant, and recessive. The filtering logic for these MOI are designed to be as general as possible to capture as many variants that fit the



METRIC	THRESHOLDS	PROBAND (FAM-0000560_ea64_3a70_7e76)	MOTHER (FAM-0000560_628b_3a70_7e76)	FATHER (FAM-0000560_6422_3a70_7e76)
Mismatched bases R2	≤1.5%	1.46%	1.03%	1.58%

METRIC	THRESHOLDS	PROBAND (FAM-0000560_ea64_3a70_7e76)	MOTHER (FAM-0000560_628b_3a70_7e76)	FATHER (FAM-0000560_6422_3a70_7e76)
Average autosomal coverage over genome	≥30.0	33.32	33.4	32.27
Mapped reads	≥90.0%	91.92%	92.31%	91.64%
Q30 bases	≥80.0%	83.5%	88.72%	83.58%
Q30 bases R1	≥85.0%	89.47%	93.59%	90.29%
Q30 bases R2	≥75.0%	77.54%	83.84%	76.86%
Mismatched bases R1	≤1.25%	0.62%	0.43%	0.58%
Mismatched bases R2	≤1.5%	1.46%	1.03%	1.58%
Sex ploidy validation	–	XY	XX	XY

Figure 1: TruSight Software Suite QC results for a trio— TruSight Software Suite flags QC metrics that fail with a QC warning, allowing a case manager to choose whether to QC Override the failure or Modify the sample information listed in the case.

selected MOI as possible, but allows for further filtering through the selection of additional criteria. If multiple MOI filters are combined in the same filter, the results will be linked by a logical 'OR' to enable the exploration of multiple MOI in the same query. Guidance on best practices for using these filters follows.

De novo

When both parents of an affected proband are unaffected or there is no family history of disease, it is typical to suspect either a recessive or *de novo* MOI. The primary requirements for variants to be returned by the ***de novo*** filter is that a variant is present in the proband, but not in either parent. The ***de novo*** filter is not appropriate when genomic data is not available for both parents, since a *de novo* determination cannot be made. The **Present in Proband** filter may be more suitable for non-trio scenarios. Layering of sub-options on top of the base logic applies further restrictions on the filter. For example, In Affected and Not in Unaffected options are used in pedigrees where there are affected siblings, unaffected siblings, or pedigree members other than the parents.

- In Affected option will identify variants that are not shared across affected individuals.
- Not in Unaffected option will identify variants that are not present in any unaffected individuals.
- Exclude Observed Homozygotes option will rule out any variants that are homozygous in any individual of the pedigree, which is not typical for a *de novo* scenario.
- Exclude Low Confidence will exclude variants that do not pass the *de novo* scoring for the variant caller.

Recessive

If both parents of an affected proband are unaffected, a recessive MOI may be suspected. Though notably, the father may be affected in the case of X-linked recessive inheritance for a female proband. The **Recessive** filter captures use cases where the causal variant or variants are homozygous, compound heterozygous, or hemizygous.

- The Strict option will identify only variants that show evidence of inheritance from one of the parents, so *de novo* variants will not be considered for compound heterozygous analysis.
- The Assume Complete Penetrance option will further exclude variants that are homozygous or hemizygous in any unaffected individual, since this would be expected to manifest phenotypes if complete penetrance is assumed. This will be expected in most cases (or at least manifesting mild phenotypes) but can be relaxed if phenotypes are known or suspected to be incompletely penetrant.
- The Exclude Ambiguous option further builds upon the Assume Complete Penetrance option to exclude variants that are ambiguously inherited from either parent in the absence of any *de novo* variants. This is because ambiguously inherited variants cannot form a unique compound heterozygote in the proband that isn't present in at least one of the parents, thus when assuming complete penetrance, these variants can be ruled out.
- The Shared Variant in Affected and Same Genotype in Affected options will further filter variants based on presence in other affected individuals, such as siblings. For Shared Variant in Affected, it requires that any variant that is returned or at least one variant of a compound heterozygote to be shared across affected individuals. It is expected that the same variants will be causal across multiple individuals in a pedigree, though this allows for differing genotypes, such as a homozygous inheritance for one affected individual versus a compound heterozygote for another.
- The Same Genotype in Affected option will further filter variants based on presence in other affected individuals, such as siblings. This option requires that that the variants returned share the same genotype across all affected individuals. This is likely to be the case when a recessive inheritance is occurring across multiple members of a pedigree.
- The Exclude Low Confidence option impacts *de novo* variants that are considered when the Strict option is not being used.

Dominant

If at least one parent is affected, a dominant MOI may be likely, though recessive inheritance is still possible. The primary requirement for any variants returned is that the candidate variants are present in the affected proband and the affected parent.

- The Exclude Observed Homozygotes option will further remove variants that are observed as homozygous in any individual. This is typically the case for dominant inheritance, since dominant alleles are rarely observed in a homozygous state.
- The In Affected option will further require the variants are present in other affected individuals, such as siblings.
- The Not in Unaffected option will require that the variant is not present in any unaffected individuals, which is expected in the case where the alleles are fully penetrant.

Present in Proband
In the absence of at least one of the parents in the pedigree, the exact MOI cannot be determined. When combined with options such as In Affected and/or Not in Unaffected, this will allow the identification of variants shared in affected individuals and absent from unaffected individuals. The Exclude Homozygotes option will exclude variants that are homozygous in any individual, if causal alleles are expected to be unlikely to appear in a homozygous state (eg, dominant). It is advisable that sub-options are used to limit variants to the most likely scenarios; however, they can be rolled back if a strong candidate variant is not identified to explore more rare scenarios.

Setting the variant category for a filter group

TruSight Software Suite provides users with the flexibility to apply a selection of filter criteria to each variant category supported in the software.

The selection of variant categories will impact the set of filtering criteria that can be selected for a given filter group (discussed further in the next section). The supported variant categories in TruSight Software Suite include:

- **Small variants**—Single nucleotide variants (SNVs), multiple nucleotide variants (MNVs), and insertions and deletions (indels) typically ≤ 50 bp in length.
- **Structural variants (SVs)**—Indels that are ≥ 50 bp, duplications < 10 kb, and translocations.
- **Copy number variants (CNVs)**—Regions in the genome with a higher (copy number gain) or lower (copy number loss) number of copies compared to the expected ploidy of the chromosome where they occur.
- **Runs of homozygosity (ROH)**—Long contiguous stretches of homozygosity (ie, an absence of heterozygous variants). Presence of such events can provide evidence of consanguinity and allow detection of uniparental disomy during interpretation of a trio analysis (ie, a proband and parents).
- **Short tandem repeats (STRs)**—Short, repeated regions of the genome that have been associated with a variety of disorders.
- **Mitochondrial variants**—Small variants and CNVs that occur in mitochondrial DNA (mtDNA).
- **Paralogs**—Gene copies that result from a duplication event. They can maintain the same function as the original gene, or diverge to develop novel functions.

Setting up multiple filter groups within a filter strategy

It is critical that filtering can address the identification of candidate variants across multiple variant types simultaneously, as pathogenic variants in a case may span multiple variant categories. This is particularly true when employing family-based analyses (discussed further in a later section), since compound heterozygotes can exist between both small and large variants. The logic structure of the filters is designed to be flexible to enable these use cases. Before any filtering criteria can be added to a filter strategy, users must first designate which variant categories should be returned.

Once the variant categories are selected, filter criteria specific to the variant category selection can be added. For example, if only small variants are selected, then the user can add criteria that are specific to small variants, such as small variant transcript consequences like "missense variant" or "frameshift variant." If small variants and CNVs have been selected, then small variant transcript consequences cannot be added, as it would not be applicable to all variant types. However, other filters, such as family-based analyses, can be added, since some criteria are applicable to all variant types. The need to apply these consequences in a larger plan are addressed through the addition of filter groups in the filtering tree. For the purposes of this technical note, the selection of variant categories and filtering criteria under that selection are referred to as a filter group. The logic within a filter group are joined with a logical 'AND'. This means that variants must meet all criteria within that group in order to be returned. For example, if a gnomAD population frequency filter and a coding transcript consequences filter are added for a small variant filter group, only variants that meet the allele frequency requirements and consequence requirements of that filter group are returned.

Filter groups can then be nested together to form complex filtering strategies that address many needs, such as identifying variants across multiple variant categories. Adding additional filtering groups will automatically be linked by a logical 'AND' with the initial variant category set-up at the "top-level," but they will be linked by a logical 'OR' with any additional filter groups added, such that the logic follows [Group A] AND ([Group B] OR [Group C] OR [Group D]…). This flexibility enables logic to explore multiple variant types, while simultaneously applying family-based analysis on top of the entire filter (Figure 2).

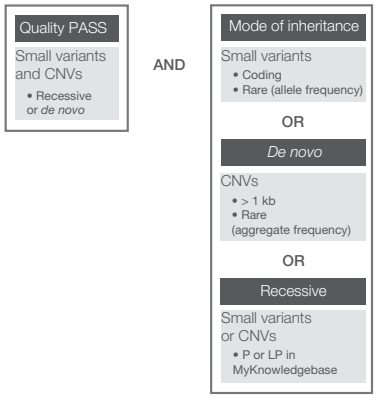


Figure 2: Example logic scheme for linking additional filters to a filter group

Filtering by annotations

TruSight Software Suite provides a rich set of variant and gene annotations to aid in filtering down to variants of interest. These include sources from publicly available databases, such as gnomAD and ClinVar. TruSight Software Suite also provides a framework for users to enter and filter on their own custom annotations. These include information related to a variant's prevalence in the population, potential involvement in human disease and phenotypes, and predicted impact to protein function.

Filtering on publicly available annotations

Consequence-based filters
A user can filter variants based on the predicted effect on all overlapping transcripts from either RefSeq or Ensembl gene annotations, which is set at the test creation step. Each different variant category has a specific set of consequence filters, and a separate filter group is required to apply the consequence filters to each different variant category in a variant set.

External annotations

- **OMIM (Online Mendelian Inheritance in Man)**—OMIM provides information about the known relationship between human phenotype and genotype. The TruSight Software Suite OMIM filter allows a user to filter genes based on if the gene is present in OMIM (with OMIM gene ID) and associated with any OMIM phenotypes. The user is also provided with the option to include or exclude provisional phenotype/genotype associations.
- **gnomAD LCR**—A filter to exclude variants inside identified low-complexity genomic regions (LCR) based on information from gnomAD.
- **gnomAD Constraint**—A set of gene-level metrics provided by gnomAD to help identify genes subject to strong selection against various classes of mutations.
  - **The loss-of-function observed/expected upper bound fraction (LOEUF)**—A conservative estimate of the observed/expected ratio of loss-of-function (LoF) variants in a gene. Low LOEUF scores indicate strong selection against predicted LoF variation in a given gene, and the suggested value for filtering is < 0.35. This score is recommended by gnomAD over the pLI score.
  - **pLI**—The probability that a gene is intolerant of a LoF mutation. The suggested value for LoF intolerant genes is ≥ 0.9.
  - **pRec**—The probability of being intolerant of homozygous, but not heterozygous, LoF variants.
  - **pNull**—The probability of being tolerant of both homozygous and heterozygous LoF variants.
  - **misZ**—Z score indicating gene intolerance of missense variants. Higher scores suggest higher intolerance of missense variants. A suggested threshold value is 3.09 ($p < 0.001$).
  - **synZ**—Z core indicating gene intolerance to synonymous variants. Higher scores suggest higher intolerance to synonymous variants.
- **Splice AI**—A score ranging from 0 to 1 indicates the probability of the variant being splice-altering. The suggested values for filtering are 0.2 (high recall), 0.5 (recommended), and 0.8 (high precision).

Filtering by gene disease or variant disease relationships

TruSight Software Suite enables the analyst to filter both genes and variants based on their relationship to human disease. These include public data sources such as OMIM, ClinVar, and ClinGen. In addition to these public resources, TruSight Software Suite enables the user to filter on their own disease associations stored in their own Knowledge Network.

- **OMIM**—See description under "external annotations."
- **Pathogenicity by ClinVar**—Users can filter variants based on the assertion of clinical significance (Benign to Pathogenic) and the level of review supporting (0 to 4 stars for review status) from ClinVar. The level of confidence for the results is expected to increase with the number of stars for review status. For example, one-star review status means "criteria provided, conflicting interpretations" and two stars correspond to "criteria provided, multiple submitters, no conflicts."
- **Haploinsufficiency and Triplosensitivity by ClinGen**—Users can filter genes by gene dosage sensitivity based on information from ClinGen. Filters for haploinsufficiency and triplosensitivity are provided with options including the level of supporting evidence and association with autosomal recessive phenotype.

Filtering on population data

Pathogenic variants often occur less frequently in the general population, therefore filtering variants based on population data can be a powerful tool. Small variants can be filtered directly based on their frequency across the population through allele frequency. Large variants, such as CNVs and SVs, can be filtered based on an aggregate frequency that is a summation of allele frequency of different events that have a high similarity, based on reciprocal overlap, with the variants observed in a case. See the TruSight Software Suite User Guide for more information.

Filtering on custom annotations

Custom annotations allow users to filter on data that is outside of the core annotations provided by TruSight Software Suite. Users can upload valid custom annotation files within a test definition in their case interpretation (see the TruSight Software Suite User Guide for technical details).

TruSight Software Suite supports two primary use cases for the addition of custom annotations:

- **Population Data**—Users can add custom annotations that specify data for use in the allele frequency and aggregate frequency filters. Users can specify small variants for specific alleles and large variants for observed variants with a genomic region. For these variants, the allele frequency and allele number can be specified by using the AlleleFrequency and AlleleNumber custom annotation categories. The description field will list the population code for the values being specified (eg, ALL, AFR, EAS, etc).
- **Filter Labels**—Users can attach labels (up to 20 characters long) to use for filtering variants. Users can specify small variants for specific alleles and large variants for observed variants with a genomic region that will match variants with a case based on annotation overlap. With a custom annotation uploaded to a test, the user will have access to a filter that allows the user to target variants that have a particular label annotated to them.

Summary

TruSight Software Suite provides a powerful interface for filtering millions of variants from a WGS dataset down to a prioritized list of key variants of interest. The software offers a broad range of variant filtering options, including publicly available and custom annotations that can be used to create a filtering plan that meets each user's needs.

Learn more

Learn more about TruSight Software Suite at www.illumina.com/products/by-type/informatics-products/trusight-software-suite.html

Illumina • 1.800.809.4566 toll-free (US) • +1.858.202.4566 tel • techsupport@illumina.com • www.illumina.com

illumina®