HiSeq Analysis Software (HAS) v2.1 Release Notes

HiSeq Analysis Software (HAS) v2.1

For Human Whole Genome Resequencing Analysis

January 30, 2017

Introduction

These Release Notes detail the key changes to software components for HiSeq Analysis Software (HAS) v2.1 since the package containing HAS v2.0. This is an optional update that provides numerous feature and performance enhancements over HAS v2.0. These features and performance enhancements will not be made available as part of HAS v2.0 and as such customers wishing to obtain these features should upgrade to HAS v2.1

If you are upgrading from a version prior to HAS v2.0, please review the release notes for HAS v2.0 for a list of features and bug fixes introduced in that version.

HAS v2.1 is a software package for analyzing sequencing data generated by Illumina HiSeq sequencing systems. The software leverages a suite of proven algorithms to detect genomic variants comprehensively and accurately. HAS v2.1 is a complete package with a range of variants, including Single Nucleotide Variants (SNV), Indels, Structural Variants (SV) and Copy Number Variants (CNV) for tumor and normal samples. HAS v2.1 utilizes the base calls and quality scores generated by Real-Time Analysis (RTA) software during the sequencing run to analyze data rapidly for high-throughput whole-genome sequencing analysis.

HAS v2.1 uses the Isaac Aligner and Strelka Germline Variant Caller to provide both aligned and unaligned reads and variants.

For structural variants, HAS v2.1 uses 2 complementary approaches:

- Read depth analysis by Canvas. See Canvas (Copy Number Variant Caller)
- Discordant paired-end analysis by Manta. See Manta (Structural Variant Caller)

HAS v2.1 supports a multistep analysis workflow that is amenable to using a simple command line with recommended default settings, which are specified in the sample sheet.

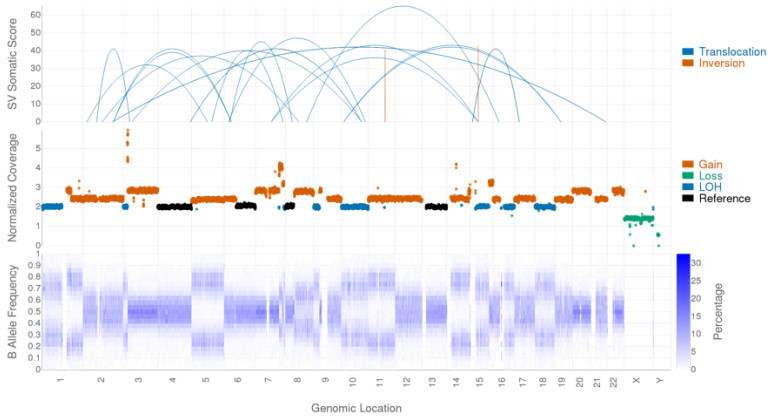
The software components included in this package are detailed below.

I. New Features:

- The aligner now soft-clips the 3' and 5' ends of reads
    - Reads that have a poor match to the reference (using the **--clip-semialigned** command-line argument) are soft-clipped. This leads to a lower mismatch rate, and a very small decrease in the number of aligned bases.
- Variant annotation is now available for the current version of the Human reference genome (as well as for hg19 / GRCh37). The particular FASTA file tested was downloaded from NCBI (NCBI/GRCh38Decoy).
    - Most annotations are derived from Ensembl database version 81.
    - Some metrics have been updated to account for additional reference contigs - for instance, coverage stats exclude decoys and EBV.
- Variants are scored using a trained model
    - Incorporates depth normalization to handle high coverage regions
    - Scoring of het-alt indels has been improved
- Starling now does short-range phasing of adjacent SNVs occurring on the same chromosome(s) into an MNP (Multiple Nucleotide Polymorphism). For example, a record in the .vcf file with REF value AC and ALT value GA would be an MNP. This helps ensure that variants affecting the same codon are annotated as a single event (with the correct amino acid change).
- Allosome filtering / annotation: Sex chromosome ploidy can be passed in via the PloidyX / PloidyY columns in the sample sheet, or determined from coverage (if the sample is XX or XY). SNVs and indels on chrY are filtered with PloidyConflict if run on a female (Y count = 0) sample. Similarly, SNVs on chrX/chrY are reported as hemizygous or flagged with the PloidyConflict filter for a male (X count = 1, Y count = 1) sample. An exception is that the PAR region on chrX is always treated as diploid, since the corresponding region on chrY is N-masked.
- The gVCF now includes explicit output for an updated set of clinical indels, a set of sites (including for hg38) where the reference base is a minor allele, and for a collection of pharmacogenomics (PGx) sites. The somatic vcf output from Strelka now explicitly genotypes a collection of known oncogenic sites.
- SNV/indel calling now uses a haploid model as appropriate for chrX/chrY.
- SNV/indel calls now include a PL tag in the output .vcf file.
- SNV/indel calling on the mitochondrial chromosome now detects variants at any frequency (down to ~5%), rather than using a diploid model.
- The .vcf header reports the tool name as "Starling" rather than "Isaac Variant Caller". (This is to avoid confusion, given that Starling is a separate open source tool from the Isaac aligner, and sometimes one uses Isaac but not Starling or vice versa)

- Default somatic SNVs calling noise parameters changes to allow for more aggressive inclusion of mid-range signal-noise loci.
- Introduced novel filter method for somatic SNV calls.
    - An empirically estimated model in the form of Random Forrest classifier is used for identifying a highly confident somatic calls from the entire candidate set reported in the VCF.
    - The model makes use of the following feature variables when for scoring a single somatic loci: QSS_NT, N_FDP_RATE, T_FDP_RATE, N_SDP_RATE, T_SDP_RATE, N_DP_RATE, TIER1_ALLELE_RATE, MQ_SCORE, MQ_ZERO_RATE, SNVSB, ReadPosRankSum, ALTMAP, ALTPOS (see VCF header for exact definition of feature values).
- **Format change:** The header of the somatic subtraction .vcf file uses the **sample IDs** as the column headers for the tumor and normal samples.
    - Note: Older pipeline versions used the string NormalID_$X and TumorID_$Y where $X and $Y are the sample IDs)
- The Strelka output now includes non-PF calls, not just variants which pass filters (PASS in the FILTER column of the .vcf file)
- Updated somatic SV caller to better handle highly degraded FFPE samples.
    - Changes ensure that the variant caller completes for these samples and reduces false positives in the context of high spurious chimera rates.
- Somatic SV caller can now detect translocations from split reads alone, without requiring a minimum paired read count, improving recall when average DNA fragment length is low.
- Reads containing "N"'s and low-quality bases are correctly handled by the SV assembler now, increasing the fraction of basepair resolution SV calls.
- Germline SV caller labels large SVs without paired read support with a new filter "NoPairSupport".
- Germline SV caller has higher large insertion recall.
- Manta has been updated for greater robustness on hg38 where many small contigs are present
- For both somatic and germline CNV calling, **reference** copy number calls (with no SVTYPE specified) are included in the .vcf file for regions with no copy number abnormalities detected
- For both somatic and germline CNV calling, the allosome copy number from the normal is used in labeling calls as reference (REF) or variant (GAIN/LOSS) calls. For XX samples, the reference copy number is 2 for chrX and 0 for chrY. For XY samples, the reference copy number is 1 for both chrX and chrY - with the exceptions of the PAR regions on chrX (which are N-masked on chrY), which have reference copy number 2.
- Somatic caller changed from SENECA to Canvas (Isaac Copy Number Variant Caller)
    - SV and CNV calls are now merged into a common SV .vcf file in the Tumor/Normal analysis output (as well as in the WGS analysis output). The SV calls (from Manta)

include a call for both the normal and the tumor columns in the .vcf file; the CNV calls (from Canvas) for tumor/normal analysis report information just for the tumor sample (the normal column is left blank).
    - Formatting of somatic CNV calls now matches that of germline CNV calls: CN tag in the sample column indicates the copy number. MCC tag in the sample column indicates the major chromosome count. Example: If CN is 2 and MCC is 2, this is an interval of apparent copy-neutral LOH (two copies of the same chromosome). A quality score for each copy number call is included in the QUAL column
    - Updates to CNV calling methods provide greater robustness to FFPE artifacts, and improved ability to select the correct baseline across a range of samples
    - The somatic CNV .vcf header includes several sample-level metrics: EstimatedTumorPurity, OverallPloidy, EstimatedChromosomeCount (see the Canvas documentation for more details)
- Germline CNV calling now leverages b allele frequency information (as in somatic calling). A plot of coverage across the genome (as well as the backing data) is included in the WGS workflow output.
    - Germline copy number calling output includes MCC (Major Chromosome Count), to distinguish between different states - e.g. three tetraploid states where the major chromosome count is 2, 3, or 4. Germline CNV calling now uses an improve model for assigning quality scores to each call (variant or reference).
- CNV calling now can call copy numbers > 10. CNV reporting now leaves out MCC (Major Chromosome Count) for segments where no b allele frequency information is available. Fixed a rare case where incorrect MCC values were reported for reference regions.
- Allosome (chrX+chrY) ploidy detection has been updated for greater accuracy.
- Germline CNV calling now leverages b allele frequency information (as in somatic calling).
    - A plot of coverage across the genome (as well as the backing data) is included in the WGS workflow output.
    - Germline copy number calling output includes MCC (Major Chromosome Count), to distinguish between different states - e.g. three tetraploid states where the major chromosome count is 2, 3, or 4.
- Germline CNV calling now uses an improve model for assigning quality scores to each call (variant or reference).
- The coverage plot y axis is now capped, to avoid unreadable scaling for samples with extreme amplifications (20+ copies).
- A new plot (see example below) includes coverage levels, b-allele frequencies and structural variants. The legacy Circos plot has been discontinued.



- The annotation tool has been upgraded from IONA (Illumina on-node annotation) to Illumina Annotation Engine
    - Improved speed and accuracy compared to IONA (removed issues related to legacy VEP scripts)
    - Updated annotation database version from Ensembl database version 72 to version **81**
    - Positions where the reference base is the minor allele are (as before) being excluded from block-compression and annotated with the RefMinor tag in the .vcf file; the corpus of RefMinor sites is now based upon phase 3 data of the 1000 genomes project (including chrY sites), rather than phase 1.
    - PhyloP is reported as conservation scores instead of PhastCon.
    - Annotation is now available for hg38 (reference folder NCBI/GRCh38Decoy)
- The workflow now includes calling of triplet repeat expansions using ExpansionHunter. The list of loci called is specified in the Annotation/ExpansionHunterDiseaseSpec folder for the reference genome.
    - This caller is disabled by default; it is invoked if (and only if) the setting RunExpansionHunter is set to true in the [Settings] section of the sample sheet.
- The workflow now can invoke an **HLA typer** to identify the most likely genotypes for six HLA genes.
    - The output is a tab delimited text file with name of the form SampleID_S#.HLA.txt with 4 columns: GeneName, Allele1, Allele2, Rank. This file lists the 10 most likely pairs of alleles for each of 6 HLA genes.
    - HLA typing is disabled by default; it is invoked if (and only if) the setting RunHLATyping is set to true in the sample sheet.
- Dates (and times) in output reports now use ISO-8601 formatting (e.g. YYYY-MM-DD), to avoid confusion between DD/MM/YYYY or MM/DD/YYYY.
- The WGS workflow now includes a table of coverage (and callability) by exon.

II. Defect Repairs:

- Fixed a rare case where incorrect MCC values were reported for reference regions.
- Fixed an issue where some variants in gvcf are inappropriately flagged with LowGQX filter
    - Confusing filter status: In rare cases, some variants are tagged with the LowGQX filter although their GQX value is above the stated filter threshold. These variants are overwhelmingly calls which should in fact be filtered (i.e. the future fix is not to clear this flag, but to filter them for the right reasons). This appears to affect ~0.5% of variants in a whole-genome .vcf file.
- Fixed an issue with NSv2 relating to poor variant filtering performance on samples at >= 70x coverage.
    - There's a known limitation of the Starling VQSR scoring model in NSv2 such that it would perform poorly on high-coverage samples. This does not affect the variant calls themselves, but whether the variants pass filters (have PASS in the FILTER column of the .vcf file) or not. NSv2 has an ad hoc work-around such that it uses the old rule-based filtering on samples > 70x; the negative impact can be seen in the 80x and 130x runs (but the performance with the VQSR that is in NSv2 would be much worse). Note that this issue does not impact high-coverage tumor samples, which do not go through variant calling using starling (instead, strelka is used for subtractive somatic variant calling). For normal samples with unusually high coverage (e.g. two lanes of a HiSeq X without multiplexing), the issue can be worked around by downsampling the data. In particular: This setting (in the [Settings] section of the sample sheet) excludes ¼ of the tiles (skipping one swath for one surface):
    - **ExcludeTiles,2200-2224**
    - This setting excludes 1/3 of the tiles (skipping one third of each swath on each surface):
    - ExcludeTiles,1117-1124+1217-1224+2117-2124+2217-2224
- Fixed an issue where an incorrect filter label added for homozygous deletions (NSv2/starling 2.1.4.x)
    - The wrong label is applied to FILTER for homozygous deletions (they are labeled LowGQXHetDel rather than the intended LowGQXHomDel), with an accompanying error in the header definition of LowGQXHetDel. This will no doubt lead to some confusion, but those calls that should not be PASS have some value in the FILTER field, and the correct FILTER value could be determined by other entries in a record (esp GT 1/1).
- Fixed an issue In rare cases, for tumor samples with a very large number of novel SNPs and indels, causing annotation failure. (IONA relies on VEP for annotation of novel variants, and VEP may hang indefinitely in these cases). The workflow completes (albeit slowly) but the somatic SNVs and indels are not annotated.

III. Known Issues:

- None.

Category	Software	Version (HAS 2.1)
Aligner	Isaac	03.15.09.04
Variant Caller (Germline SNV/Indel)	Starling	2.3.13
Variant Caller (Tumor/normal SNV/indel)	Strelka	2.3.13
Variant Caller (SVs)	Manta	0.27.2
Variant Caller (CNVs)	Canvas	1.3.0
ROH caller	Consanguinity	0.1
Variant Caller (Repeat Expansions)	ExpansionHunter	1.5.10
Annotation	Nirvana (Variant Annotator)	1.3.2
Annotation Database		Ensembl database v81
Metrics	Puma Metrics	1.0.8.0
Metrics	PUMA	00.15.07.08
Other	mono	4.0.2
Other	bcl2fastq	2.17.1.14
Other	InterOp	1.3.1
Other	scramble	1.13.10