Using NIST Genome in a Bottle Reference Samples to Validate an NGS Analysis Pipeline with NextGENe® and Geneticist Assistant® Software

Megan McCluskey, Kevin LeVan, Edward Bouton, Shouyong Ni, CS Jonathan Liu

Introduction

The Genome in a Bottle (GIAB) Consortium hosted by NIST provides reference materials to facilitate the adaptation of next generation sequencing tests into clinical practice. The reference materials, based on NA12878 DNA from Coriell, are a valuable tool for validating an NGS pipeline as the pipeline results can be compared with the highly confident small variant (SNP and Indel) calls from GIAB.

NextGENe and Geneticist Assistant software provide a complete user-friendly analysis pipeline for NGS data from the raw data off of the sequencer to curated variant calls and variant interpretation. GIAB samples can be used to validate the optimal analysis and reporting settings by specifically evaluating high confidence variant calls. NextGENe software offers flexible variant calling and reporting settings for filtering variant calls and Geneticist Assistant NGS Workbench can be used to compare NextGENe's calls with a VCF of the known calls for the sample from GIAB. In this way settings can be adjusted to eliminate potential false negatives while minimizing false positives.

The workflow below describes a step-by-step process to set up and validate NGS analysis with NextGENe. NextGENe also provides the option to save settings and automate processing to run unattended. Import to Geneticist Assistant can also be automated to provide a complete seamless pipeline. The data discussed below is based on GIAB v2.18.

Methodology

FASTQ files are loaded in NextGENe software for alignment and variant calling. NextGENe's Mutation Report lists all detected variants and can be filtered to remove false positives. The Mutation Report can also be used to identify potential false negatives by comparing the detected variants in the report with a VCF containing the expected variants for the sample from GIAB. NextGENe's Mutation Report in VCF format is then imported to Geneticist Assistant where variant calls can be further compared with the expected variant calls within the GIAB VCF.

NextGENe Project Setup

1. On the first page of NextGENe's Project Wizard select the appropriate Instrument type, SNP/Indel Discovery as the Application type and Sequence alignment under Steps. Click Next.
2. If applicable, click "Format Conversion" to convert FASTQ files to FASTA format.
3. In the Format Conversion Tool click "Add" to load FASTQ files. Settings for quality filtering can be adjusted as needed. Click "OK" to begin conversion.
4. Under Sample files in the Project Wizard click "Load" to load the *_converted.fasta file(s) from the format conversion, or a BAM file.
5. Under Reference files click "Preloaded" to load the genome reference.
6. Under Output click "Set" to choose the output folder name and location. Click "Next."
7. On the Alignment page, settings for alignment and variant calling can be adjusted. Click "Finish" then "Run NextGENe" to begin analysis.

Figure 1: Variant detection thresholds can be adjusted as appropriate

Project Review in NextGENe Viewer

1. Once the analysis is completed the project is displayed in the NextGENe Viewer. Click the Report Display icon ( ) to display the Mutation Report at the bottom or right side of the Viewer.
2. Go to Reports > Mutation Report > Mutation Report Settings or click the Report Settings icon ( ) to view the display and filter option for the report.
   a. Filter tab, Annotation sub-tab - filters can be set relative to annotation such as filtering out non-coding and/or synonymous variants.
   b. Filter tab, Bias sub-tab - filters can be set based on directional bias, which can be indicative of false positive calls.
   c. Filter tab, Score sub-tab - filters can be set based on a variant confidence score.
   d. Filter tab, ROI sub-tab – load a BED file for targeted sequencing to remove off-target variants, or load a VCF containing expected calls for the GIAB sample to check for false negatives.

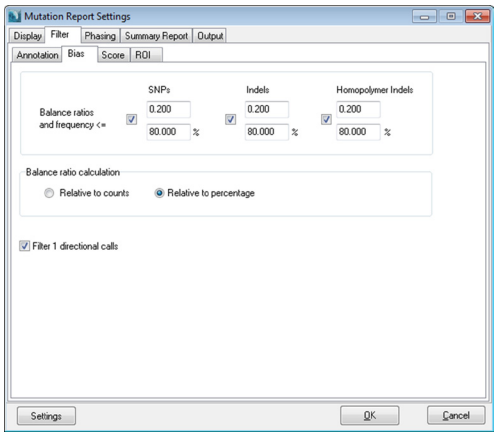

Figure 2: Directional bias filters can be applied to the Mutation Report to remove false positive calls.

3. A summary for the Mutation Report containing information such as the total number of variants included in the report can be accessed through the Reports menu in the Viewer, by selecting Reports > Mutation Report > Mutation Report Summary.
4. To compare the NextGENe variant calls with the expected calls from GIAB, on the Filter tab, ROI sub-tab, select “Filter by ROI”, click “Set” to load a VCF file containing the expected calls from GIAB for your targeted regions, and select “Include negative positions within ROI”. The full list of variant calls from GIAB can be found at ftp://ftp-trace.ncbi.nih.gov/giab/ftp/data/NA12878/variant_calls/NIST/. It is recommended to edit the VCF from GIAB to limit to variants in regions targeted by your panel. This can help to identify false negatives and can be an indication whether the settings used are correctly identifying all known variants in the sample or if settings should be adjusted. Open-source tools such as Bedtools (http://bedtools.readthedocs.org) can be used for manipulating the VCF.

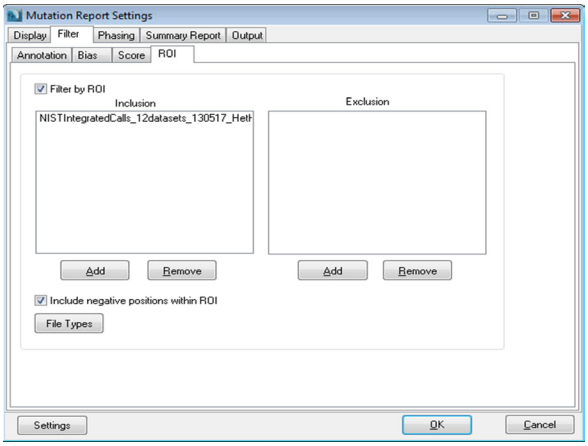

Figure 3: The Mutation Report Settings can be configured to report every position within the GIAB VCF in order to easily identify false negatives.

Index	Chrom	Pos	Coverage	Ref	Alt%	Overall Score	Mutation Call: Relative To CDS	Gene	CDS	dbSNP	Amino Acid Change
65	2	85893741	180	G	42.22	17.8	c.428C>CT	SFTPB	4	rs1130866	p.T143IT
66	2	85894308	334	G	56.29	20.0	c.232-8C>AC	SFTPB		rs3024798	
67	2	85895338	432	T	45.14	20.5	c.5A>AC	SFTPB	1	rs2077079	p.H2PH
68	2	87015633	446			0.0		CD8A		rs71378806	
69	2	87015634	444	G	52.03	18.9	c.656+19_656+20in	CD8A		rs149291342	
70	2	108608648	495	A	47.68	20.8	c.265A>AG	SLC5A7	2	rs1013940	p.I89IV
71	2	127809840	659	C	49.77	20.9	c.1137G>GT	BIN1	12	rs61748155	p.G379GG

Figure 4: Variants that are included in the VCF loaded on the Filter tab, ROI sub-tab but not called in the sample (false negatives) are shown with green text and the Mutation Call column is blank.

5. To save the Mutation Report in VCF format, go to Reports > Mutation Report > Save VCF Report (filtered).

Geneticist Assistant Import and Review

1. Configure Reference.
   a. Download the Human_37_*.7z from the SoftGenetics ftp folder and extract to any location on your computer.
   b. In the Geneticist Assistant Settings dialog, Directories tab, click "…" to select the folder location where the extracted Human 37 folder is located.
   c. The GIAB annotation can be added to Geneticist Assistant by copying the GIAB VCF into the Human 37 folder.
2. Create a panel for the targeted regions.
   a. Go to Panels > Manage Panels.
   b. Click "New Panel".
   c. Click "…" to load the BED file for the panel. Click "OK".
   d. The BED file is loaded in the Manage Panels dialog. Click "OK" to create the panel.
3. Import VCF data.
   a. Go to File > New Run.
   b. Enter a Run Name and specify the Required Settings.
   c. Click "Select Variant and/or Coverage Files…" to browse to and select the VCF file saved from the NextGENe Viewer. You can also include a BAM file for the sample in order to review coverage statistics for your panel regions.
   d. Click "OK" to submit the data.
   e. The Current Jobs tab will display the import progress.
4. Review Data
   a. Once the data import is complete the new run will be included in the list of runs on the Runs tab. Double-click the run to open the Run tab for the run.
   b. On the Run tab, double-click the sample name to view the list of variants included in the VCF file for the sample.
   c. Double-click any variant in the list to view detailed information for the variant, including the GIAB annotation at this position.
   d. On the Variant tab that opens a pane labeled with the name of the GIAB VCF is displayed, showing the Ref and Alt alleles for the position according to GIAB
   e. You can add this information to the variant table on the Sample tab to allow quick review for all variants by right-clicking in this pane and selecting "Add Values to Variant Table".

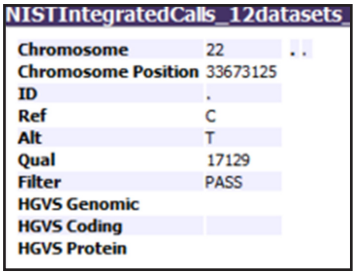

Figure 5: The detailed Variant tab for each variant displays a pane showing annotation for the position from the GIAB VCF.

Variants of 'MSP_Prec1_NIST3_100915EC_Mutation_Report_Filtered_newsettings':

ID	Chr : ChrPos	Rs	Ref	GIAB Ref	Alt	GIAB Alt	Ref AA	Alt AA	HGVS Coding
79	1:982994	rs10267	T	T	C	C	Phe	Phe	NM_198576.3:c.3558T>C
80	1:984302	rs9442391	T	T	C	C	Thr	Thr	NM_198576.3:c.4161T>C
81	1:987200	rs9803031	C	C	T	T			NM_198576.3:c.5651+5C>T
82	1:990280	rs4275402	C	C	T	T	Asp	Asp	NM_198576.3:c.6057C>T
83	1:22191454	rs897471	G	G	A	A	Ala	Val	NM_005529.5:c.4508C>T
84	1:22206649	rs989994	T	T	C	C	Asn	Ser	NM_005529.5:c.2294A>G
85	1:22206942	rs1874793	G	G	A	A	Ala	Ala	NM_005529.5:c.2109C>T
86	1:22207235	rs1874792	T	T	C	C	Met	Val	NM_005529.5:c.1912A>G
87	1:22214127	rs2229478	A	A	G	G	Leu	Leu	NM_005529.5:c.744T>C
88	1:22217108	rs2501260	G	G	A	A	Phe	Phe	NM_005529.5:c.324C>T
89	1:22263640		A		C				NM_005529.5:c.63+8T>G
90	1:40773123	rs2228567	G	G	C	C	Leu	Val	NM_001852.3:c.1003C>G

Figure 6: Annotation from the GIAB VCF such as the Ref and Alt Alleles can be added to the variant table allowing viewing of information for all variants in the sample in a single view.

Discussion

The procedure described above provides a step-by-step "from scratch" project setup workflow. NextGENe also includes the NextGENe AutoRun Tool which can be used to quickly set up batch analysis using pre-saved settings files to process multiple samples unattended. Exporting the NextGENe results to Geneticist Assistant can also be automated.

Acknowledgements

We would like to thank Jason Park, Midori Mitui, and Eric Crossley of Children's Medical Center Dallas for contributing their data and providing feedback.