Germline and Somatic Mutation Analysis: Experimental Design, Variant Calling, and Analysis

Justin Lack

NIAID Collaborative Bioinformatics Resource (NCBR) Frederick National Laboratory for Cancer Research October 11, 2018

Presentation Outline

- Experimental Design Considerations in Variant Analysis
 - Germline vs Somatic Variant Detection
 - Whole Genome vs Whole Exome Sequencing
 - Best Practices
- Variant Calling Pipelines from NCBR/CCBR
 - Pipeline Performance
 - Using Pipeliner on Biowulf



Germline vs Somatic Variation

- Germline Variant Analysis
 - Heritable genetic variation
 - Large cohort analysis -> GWAS/Burden testing (quantitative)
 - Small cohort analysis -> Candidate gene identification (qualitative)
 - Pedigree analysis -> variant/disease co-segregation
- Somatic Variant Analysis
 - Non-heritable genetic variation arising in non-germ cells
 - Tumor/Normal or tumor-only analysis
 - Somatic mosaicism (e.g., Neurofibromatosis)

• Very different expectations in terms of allele frequency distribution



Germline vs Somatic Variant Calling

• Potentially very different allele frequency expectations



Germline - ~0.5 read proportions

Somatic - ~0.3 read proportions

Germline vs Somatic Variant Calling

Potentially very different allele frequency expectations





Germline vs Somatic Variant Calling

Within-sample germline allele frequency variance is driven by sample quality and sequencing quality

- Amount of sample input
- Sequencing depth
- Read/base quality

Somatic allele frequency variance is driven by MANY more factors:

- Sample quality and sequencing quality
- Subclonality/heterogeneity
- Copy number variation
- Tumor purity







Depth Effects - Germline

- ~30X target for genome data (below)
- ~50X target for exome, due to increased depth variance



Telenti et al., 2016 PNAS

Depth Effects - Germline

- ~30X target for genome data (below)
- ~50X target for exome, due to increased depth variance



Belkadi et al., 2015 PNAS

Depth Effects - Somatic

- "Spike-in" simulation analysis of impact of tumor purity and sequencing depth
- Simulated somatic variants into real germline whole exome sequencing on the NA12878 sample

0.8 0.8 Indel Sensitivity SNV Sensitivity 0.6 0.6 0.4 0.4 T 80/N 60 0.2 0.2 T 80/N 40 T 80/N 30 • T 80/N 20 0 0 0 10 20 30 40 50 0 **Expected VAF** Tumor Depth Variable/Normal 40× 0.8 0.8 Indel Sensitivity SNV Sensitivity 0.6 0.6 0.4 T 100/N 40 0.2 0.2 T 80/N 40 T 60/N 40 • T 40/N 40

Tumor 80×/Normal Depth Variable





Expected VAF



Depth Effects - Somatic

- Conservative recommendations:
 - >50X target for germline exome
 - >100X target for somatic exome
 - Tumor purity ≥50% (ideally ≥60% for copy number calling)

0.8 SNV Sensitivity 0.6 0.4 T 80/N 60 0.2 T 80/N 40 T 80/N 30 • T 80/N 20 0 20 0 10 30 40 50 **Expected VAF**



Tumor Depth Variable/Normal 40×





Tumor 80×/Normal Depth Variable

• Exome Sequencing

- Covers ~5% of genome (depending on capture kit)
- Allows for high depth targeting
- Most reasonable option for somatic variant analysis
- Poor copy number/structural variant calling
- Genome Sequencing
 - Confidently call >85% of reference genome (hg38)
 - Confidently call copy number/structural variant calling due to reduced depth variance
 - Significantly more accurate variant (SNP/INDEL) calling relative to exome
 - Price for WGS comparable to exome for germline-only projects

- Depth variance MUCH higher for exome
- ~2-fold more variants with GQ < 20 for exome
- Read ratio for heterozygous variants significantly skewed for exome
 - Especially pronounced for INDELs



Exome Capture Considerations

- Significant capture and enrichment biases for different kits
- Illustrates issue with combining samples from multiple kits
- For germline-only analysis, WGS strongly preferred



Meienberg et al., 2015 Nucleic Acids Research

- Sure, there's bias in WES introduced due to capture, but does it significantly affect variant calling?
- Genome in a bottle (GIAB) truth sample (NA12878)
 - 50X WES and 30 WGS available from exact same sample
 - Processed both WES and WGS through identical pipelines
 - Compared both variant sets to GIAB truth set
 - Used only exonic sites targeted in WES capture for performance assessment

35,768

Exome (50X WES)98.8%TRUE-POSITIVES

RECALL	93.21%	FALSE-POSITIVES	436
F-MEASURE	95.92%	FALSE-NEGATIVES	2,607

PRECISION



	Genor
PRECISION	99.62%
RECALL	99.84%
F-MEASURE	99.73%

me (3	OX WGS)	
	TRUE-POSITIVES	38,314
	FALSE-POSITIVES	147
	FALSE-NEGATIVES	61

10001000



3X higher False Positive rate, and >40X higher False Negative rate for exome!!!

	False Negative	False Positive
	Rate	Rate
50X Whole		
Exome	0.067934853	0.011361564
30X Genome		
(Exome Sites)	0.001589577	0.003830619



Genome (30X WGS)

TRUE-POSITIVES	38,314
FALSE-POSITIVES	147
FALSE-NEGATIVES	61



Somatic Variant Calling – Considerations and Best Practices

Paired Tumor/Normal vs Tumor-only Somatic Variant Calling



Paired Tumor/Normal vs Tumor-only Somatic Variant Calling For tumor/normal calling, strong prior on variant evidence in germline sample Variant filters (site-based) Tumor Normal Panel of normal dbSNP -1 -1 samples -1 Read filters - 1 somatic mutations Proximal gap Strand bias HC + PON callset STD callset callset $L[M_f^m]P(m,f)$ $\left| \stackrel{?}{\geq} \log_{10} \delta_T \right|$ log₁₀ $L[M_0](1-P(m,f))$ Ч Triallelic site Poor mapping Candidate Variant detection statistic Variant classification Clustered Observed position in control

Paired Tumor/Normal vs Tumor-only Somatic Variant Calling



- Collection of individuals assumed to be "normal"
- Used to augment population frequency databases (e.g., ExAC, GnomAD)
 - Population databases are highly filtered and curated
 - Many segregating germline variants are missing from population databases because they occur in challenging portions of the genome to call genotypes
- Also useful for removing systematic sequencing and mapping artifacts...



- This is a tumor/normal cohort of adrenocortical carcinoma (ACC)
- Most common driver gene for ACC is known to be beta-catenin (CTNNB1)
- Somatic variant analysis for our cohort suggests RFPL4AL1 is the most frequently mutated gene
- WE'RE GONNA BE FAMOUS! PUBLISH IN NEJM! WOOHOO!





- This is a tumor/normal cohort of adrenocortical carcinoma (ACC)
- Most common driver gene for ACC is CTNNB1
- Somatic variant analysis for our cohort suggests RFPL4AL1 is the most frequently mutated gene







- Collection of individuals assumed to be "normal"
- Used to augment population frequency databases (e.g., GnomAD)
 - Population databases are highly filtered and curated
 - Many segregating germline variants are missing from population databases because they occur in challenging portions of the genome to call genotypes
- Even with matched germline, these artifacts will be prevalent, and because they are systematic, they can be widespread

- Collection of individuals assumed to be "normal"
- Used to augment population frequency databases (e.g., GnomAD)
 - Population databases are highly filtered and curated
 - Many segregating germline variants are missing from population databases because they occur in challenging portions of the genome to call genotypes
- Even with matched germline, these artifacts will be prevalent, and because they are systematic, they can be widespread
- PON that was processed in (approximately) the same way as the case samples can remove many of these artifacts

PON Development at CCBR/NCBR

- 211 unaffected spouses from diversity of NCBR/CCBR germline projects
- 445 additional germlines from "normals" in various publicly available databases
- All WES samples
 - Processed with variety of WES capture kits (Agilent, Illumina, etc.)
 - Sequenced on multiple Illumina platforms
- Processed each sample individually in PON mode in MuTect2
- Retained only variants present in >=2 samples

PON Development

- Annotated the entire raw PON with gene information using VEP
- Removed any variant in the gene region of a gene in the COSMIC v84 database
- Removed any variant in a confirmed gene from ClinVar that was annotated as Pathogenic, Potentially Pathogenic, Drug Response, and Risk Factor
- Removed any specific variant identified in ClinVar as "Associated"
- Pooled all remaining variants into a single PON

PON Performance





Paired Tumor/Normal vs Tumor-only Somatic Variant Calling

Even with a PON, false positive rate will be significantly higher for tumor-only relative to tumor/normal calling

	100-	1000 -
		and the second
	のためには、「「「「「「「」」」」	
	たいないなが認識が認識を見たいという。	ALAR - CALLER
	11日 一般を行うたい。	
		and the second
	いるのななので、「ない」のないで、「いい」」	
3		
2	- 「「「「「」」」、「「」」、「」、「」、「」、「」、「」、「」、「」、「」、「」	as and a second s
	1111日の「日本語」の第二日の11日日	
		a strategy and a stra
		and the second second
1		
	1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1	
	and the second s	
	and the state of t	
3		
• •		
- 11.11 - 11.11 - 11.11 - 11.11 - 11.11	and the state of the second	
	-	

Paired Tumor/Normal vs Tumor-only Somatic Variant Calling

Even with a PON, false positive rate will be significantly higher for tumor-only relative to tumor/normal calling

	all for the second s						1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1			Service							3	4					*					T	ĊĠ	AAM
000 -		Tu		Dr-	on		「「「「「「「「「」」」」			And the state of t	、そうない、深語ない		Construction of the second			:	•	Constitution in the second	and the second s						 	4.)				
100 -		Tu	mc	pr/	NO)ri	が温泉語を用いた		たたたちな教御時間の		軍機管理の構成などで	のないないないである	AND PROPERTY AND	ので、「「「ない」」	一、小学校の教育を行			なると、読んでのないので、	all and a second of	- 11 小学学校の	the state of the s	Sales of the second sec	The second s	Performance and	- 111日の記録	1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1	14	· · · ·	and the second	
10-						· · · · ·		•				- 10									-1- -			· · · · · · · · · · · · · · · · · · ·						
1.		2				si.			*		1										8				 			22		

Tumor Heterogeneity/Subclonality



Tumor Heterogeneity/Subclonality

- For identification of subclones, certain conditions required:
 - High tumor purity****
 - High depth/coverage (sky's the limit!)
 - Paired tumor-normal
 - WGS essentially required for adequate sensitivity and accuracy
 - WES alone inadequate for copy number segmentation, high VAF variability, too few mutational events per subclone



Tumor Heterogeneity/Subclonality

- For identification of subclones, certain conditions required:
 - High tumor purity****
 - High depth/coverage (sky's the limit!)
 - Paired tumor-normal
 - WGS essentially required for adequate sensitivity and accuracy
 - WES alone inadequate for copy number segmentation, high VAF variability, too few mutational events per subclone



FFPE vs Fresh/Frozen Tissue – 50X target depth


Somatic Variant Calling – Best Practices

- STRONGLY favor paired tumor/normal design
- For non-human samples (e.g., mouse models) without paired somatic/germline
 - >=2 control/"germline" samples
- >=100X/50X mean depth for tumor/normal samples
- Significantly higher target depth for FFPE samples
- Tumor purity >50% (ideally, >60%) for variant calling
- MUST visually verify any somatic variant of "significance"
- Subclone analysis requires WGS, high purity, and high depth

Germline Variant Calling – Considerations and Best Practices

GWAS/Burden Testing Design

"With an odds ratio (OR) = 1.4, the sample sizes required to achieve 80% power are 6,400, 54,000, and 540,000 for a MAF = 0.1, 0.01, and 0.001, respectively, if one assumes 5% disease prevalence and a significance level of 5 x 10⁻⁸. Because the number of rare variants is much larger than the number of common variants, more stringent significance levels might be required, further reducing power."

Lee et al., 2014, American Journal of Human Genetics

REVIEW

Rare-Variant Association Analysis: Study Designs and Statistical Tests

Seunggeung Lee,¹ Gonçalo R. Abecasis,¹ Michael Boehnke,¹ and Xihong Lin^{2,*}

GWAS/Burden Testing Design

- Biggest challenge is having data that is homogenous
- Cases and controls MUST have genotype data that is generated identically



Familial Sequencing Design

- Power is the primary limiting factor
- When budgets are limited, decisions have to be made about who to sequence



Familial Sequencing Design

- 3 cases, no controls
 - 3,176 candidates
- 3 cases, 1 spousal control (ethnicity matched) 1542 candidates
 - +1 spouse controls 1121 candidates
 - +1 case 525 candidates
- 3 cases, 1 related control 854 candidates
 - +1 related control 307 candidates
 - +1 case 284 candidates



Familial Sequencing Design

- 3 cases, no controls
 - 3,176 candidates
- 3 cases, 1 spousal control (ethnicity matched) 1542 candidates

F1*

F49 F3 F4 F2* F2-2 F16* F15

F£11¥£12¥E13√E

\$25F183F184F185F(18)

F186

- +1 spouse controls 1121 candidates
- +1 case 525 candidates
- 3 cases, 1 related control 854 candidates
 - +1 related control 307 candidates
 - +1 case 284 candidates

ALWAYS PERFORM ETHNICITY-AWARE FILTERING!!!!





Familial Sequencing Design

- 3 cases, no controls
 - 3,176 candidates with global allele frequency threshold of ≤0.01
 - 2,923 candidates with EUR-only!

Germline Variant Calling – Best Practices

- Whole genome strongly preferred
 - >=30X mean target depth
 - Far superior to exome for structural variants, copy number analysis, and SNP/INDEL detection
- Germline exome
 - >=50X mean depth
- For familial/trio analyses, we strongly encourage early consultation
 - Selection of samples for sequencing can be CRUCIAL to maximizing power



Other Considerations and Best Practices for Variant Analysis

Always Visualize Significant Variants

- ABSOLUTELY CRUCIAL!!
- ALVIEW (<u>https://github.com/NCIP/alview</u>)
 - Internally-developed tool for BAM/SAM visualization (Richard Finney)



Variant Analysis in Cell Lines

- Can never assume your cell lines are homogenous!
- This cell line had a subclone with an **ARID1B** loss
- Another subclone had lost the Y chromosome

Replicates of the parental cell line



Variant Analysis in Animal Models





- Mice have retained significant levels of heterozygosity
- Vary considerably in regional gene expression

Multiomic Integration

- Gene and protein expression data can be used to prioritize variants and genes
 - Especially critical in underpowered, small cohort analyses
- Germline or somatic variants that do not have expression consequences at the gene and/or pathway level can be reduced in priority
- Powerful method for prioritizing non-coding mutations in WGS

FOXA2/FOXA3 Transcription Factor Network



Other Considerations and Best Practices for Variant Analysis

- Visualization and validation is a necessity
- STRONGLY recommend WGS for any cell line or model organism to be used in quantitative analysis
 - Considerable variation exists among cell line replicates in gene and protein expression, and a major contributor to this is genetic heterogeneity
 - Model organisms can retain considerable levels of heterozygosity, even after long-term maintenance in colony
 - Back-crossing is a necessity
 - Drift causes cell lines and model organisms to randomly accumulate genetic differences from progenitors





only







CCBR Pipeliner

Variant Calling at CCBR

- Multiple Variant Calling CCBR Pipelines
 - Whole genome (germline and somatic)
 - Whole exome/targeted sequencing (germline and somatic)
 - Variants from RNAseq
- All variant calling pipelines available through CCBR_Pipeliner app
 - https://github.com/CCBR/Pipeliner
 - Just need Biowulf account and xquartz installed on our local machine
 - module load ccbrpipeliner (enter)
 - ccbrpipe.sh (enter)

	X CCBR Pipeliner			
Project Information]			
Project Id	project (Examples: CCBR-nnn,Lab	(Examples: CCBR-nnn,Labname or short project name)		
Email address	(Mandatory field: must use	(Mandatory field: must use @nih.gov email address)		
Flow Cell ID	stats (Examples: FlowCellID, La	bname, date or short project name)		
Global Settings—				
Genome: hg19	Pipeline Family: exomeseq - Set a pipe	line		
Project Description	×			



Raw RNAseq Reads Analysis-Ready RNAseq Reads RNAseq Variant Calling Map to Reference **Call Variants** Non-GATK STAR 2-pass HaplotypeCaller http://gatkforums.broadinstitute.org/gat k/discussion/3892/the-gatk-best-**Mark Duplicates** practices-for-variant-calling-on-rnaseq-Picard Raw **SNPs** Indels in-full-detail Variants FastQ Screen AH-31 S8.R1.trimmed scree AH-32_S5.R1.trimmed_scree Split'N'Trim AN-02_S4.R1.trimmed_scree AN-04 S3.R1.trimmed scree + Reassign MAPQ HS-03 S1.R1.trimmed scree HS-04_S7.R1.trimmed_screen HS-05 S6.R1.trimmed scree **Filter Variants** S-06 S14.R1.tr HS-07 S12.R1.trimmed scree HS-08_S11.R1.trimmed_screen HS-09_S10.R1.trimmed_screen HS-10_S9.R1.trimmed_scree **Base Recalibration** HS-11_S16.R1.trimmed_scree HS-13 S15.R1.trimmed scree **Contamination Summary Filtered Analysis -Ready** Mean Quality Scores **SNPs** Indels Variants **RNAseq Reads** Alignment/Read QC Variant **Read Quality** Annotations Genome Fraction Coverag AVIA + SnpEff 20 30 40 50 60 100M 30M **Duplication Stats Read Trimming** CoverageStats Summary

PRE-PROCESSING

VARIANT DISCOVERY

• All variant calling follows the same basic approach







Created with MultiOC

Dropped



Read Processing/QC





Indel realignment



1	1111116111	GITTATT	61116111	61
		the second		
			-	
	1000000		Contract of Contra	
	********		1000	
	and a second second		Comment of the	
nt			-	
:IIU	********		10 E 10	
			-	
	********			5

1.00				
			2.4	
	and the second sec			
1.1				
1				
				_



• Multiple sources of quality score bias



Read Processing/QC





Variant Calling at CCBR

Germline

- Joint genotype with GATK HaplotypeCaller
 - SNPs/short INDELs
- We have benchmarked and optimized a series of hard filters for removing errors





Variant Caller Performance and Filtering



False Positive
False Negative

False Positive False Negative

Variant Calling at CCBR

Somatic

• MuTect, MuTect2 (with hard filters), Strelka



Read Processing/QC

Read Mapping

WES/WGS Pipelines with Multiple Entrypoints

- Can be run starting from fastq reads, BAMs, or gVCFs
- Setup within raw data directory
 - Raw reads all in main directory
 - 'bams' directory containing all BAM files from any source (e.g., Dragen)
 - 'gvcfs' directory containing all gVCFs from any source (e.g., Dragen)
- Initialize as usual
 - During initialization, Pipeliner automatically symlinks BAMs and gVCFs if there is a 'bams' and/or 'gvcfs' directory
- Run as usual, but skip initialQC and start from the variant calling pipeline that is appropriate

How do our Pipelines perform?

FDA Consistency/Truth Challenges

 Sought to establish best practices for germline variant calling



The **Food and Drug Administration (FDA)** calls on the genomics community to further assess, compare, and improve techniques used in DNA testing by launching the second precisionFDA challenge.



President Obama's Precision Medicine Initiative envisions a day when an individual's medical care will be tailored in part based on their unique characteristics and genetic makeup.



The goal of the FDA's second precisionFDA challenge, similarly to the first challenge, is to continue engaging the genomics community in advancing the quality standards in order to achieve more accurate and consistent results in the context of genetic tests (related to whole human genome sequencing), advancing the goal of better personalized care. PrecisionFDA invites all innovators to take the challenge and assess their (or their favorite!) software on the supplied human datasets. Participation is voluntary, but instrumental in helping the community prepare for the coming genomic data revolution.

FDA Consistency/Truth Challenges

- GIAB genome (NA12878) provides known real data truth set
- Second genome (NA24385) provides unknown truth set
 - Eliminates "overtraining" problem
- Challenge(s):
 - 1. Pipeline determinism
 - 2. Precision/recall on a known truth (training set available)
 - 3. Precision/recall on unknown truth



FDA Consistency/Truth Challenges

• Performance of most recent pipeline version


FDA Consistency/Truth Challenges

• Performance of most recent pipeline version





Pipeline performance vs Sentieon and Dragen - Recall vs Precision (with Keyur Talsania, CCR Sequencing Facility)



FDA Hidden Treasure Challenge

- NA12878MOD generated by *in silico* modification of NA12878
- 50 Spike-in variants >=0.2 frequency
- INDELs <= 40bp
- Evaluate ability to detect *in silico* variants
- FP/FN balance for SNPs and INDELs
- Ability to accurately call allele frequencies



precisionFDA Hidden Treasures - Warm Up

July 18, 2017 through September 12, 2017

FDA Hidden Treasure Challenge

- 86 entries
- Ran the full somatic pipeline with both strict and relaxed filtering criteria

FDA Hidden Treasure Challenge

- 86 entries
- Ran the full somatic pipeline with both strict and relaxed filtering criteria
- Successfully recovered all 50 spike-in variants

50 45 40 35 20 20 20 20 20 20 20 5 0 0	378MO 378MO	betor4	filter.re 7 0 HC	eq_std Ssion1	bhetor1	3 / 8MU bletor2	dummy bhetor3	hdardPi	decoy	3 78M0 878 vrf	Test	code.vc scion 1	almap	ndardPi	n omics	10D.vcf 278.v.f	8/8.vc1 .KHC_2	KHC_1	o_node depth	filtered	al_2.vc dummv	mission	iission1	378MO	78MOD	aeaup Ily_std	3.vcf.gz	v_node	allma	3 78MO	3 78MO - based	+ GATK	+ GATK iterion 2	tioner 1 is a second seco	(- based	ntered 317091	D_conf	_rrom D tight	filtered	loD_all		7.0_VD 3.vtf. <i>e</i> z	ayes_1	quency s≙ATK	test2	test	test al 3 vr	ayes_2	Variant	tration	SNVer	55ION2 Dn rere	mosaic	tration calling	RPRISM	_resuit 	al_4.vc	
	G3GA DRAGEN_DNASeq_NA1287 DREC **	B8GA	NA12878MOD.imp.hardfilt KCCG 170	Sentieo n_DNAs eq NCH SURMISS	Saphe	DRAGEN_DNASeq_NAL28 / Saphe	dui Saphe	JLack_CCBR_Leidos_Stand	Sentieo n_DNAs eq_node	DRAGEN_DNASeq_NA1287 BIG_NGST00_kit-NA1287		NA12878MOD.imp.v2.reco VW submissi	Sentieo n_DNAs eq_all	JLack_CCBR_Leidos_Stand	Beijing Institute of Gen c	BIG_CelerSNP-NA12878MO	Novoalign GATK	Novoalign GATK	Sentieon_UNAseq_aiimap_ GATK Hap lo typeCaller de	GATK Hap b typeCaller u nfilt	JSLee_NA12878MOD_final_ dui	sub mi	MySubmiss DRAGEN DNASed NA1287	DRAGEN_DNASeq_NA1287	Sentieo n_Exp-NA12878	senteon_Exp_node Sentieon_TNscope_Tonly	BIG_m yP ip eli ne-NA 12878.v	Sentieon_TNscope_Tonly	Sentieon_TNscope_Tonly_8	DRAGEN_DNASeq_NA1287	DRAGEN_DNASeq_NA1287 SeasLab with GATK-b	Bowtie2 + C	Bowtie2 + C	Submiss	SeqsLab with GATK-b	Challenge_submission_201	ADIscan_NA12878MOD_	ADIscan NA12878MOD	ADIscan_NA12878MOD_filt	ADIscan_NA12878MOI HUIP⊸	Yotta3_GRCh37_liftover_fr	KCCG_1.7.0 BIG_HXG-NA12878.v	Novoalign Freebay	DeepVariant_LowFrequ			ISI AA NA12878MOD final	Novoalign Freebay	DeepVa	SGI_4th_filtra	BWA-MEM + GATK + S	SGI 3rd filtration	- HULP-m	SGI_2nd_filtr Whole-exome variant c	SKESA_SRPI	NCB1_Magic_ThierryMieg_N	JSLee_NA12878MOD_final_ Yotta2_GRCh37_liftover_fr	

0.1 0

DeepVariant DRAGEN_DNASeq_N... JLack_CCBR_Leidos_S... NA12878MOD.imp.h.. ADIscan_NA12878M... MySubmission2 VW_submission_1 Sentieon_DNAseq_n... SGL_2nd_filtration ADIscan_NA12878M... SGL_1st_result ADIscan_NA12878M... JLack_CCBR_Leidos_S... JSLee_NA12878MOD... Yotta3_GRCh37_lifto... JSLee_NA12878MOD... Sentieon_TNscope_1... SGI_4th_filtration HULP-mosaic Challenge_submissio...I SKESA_SRPRISM Sentieon_DNAseq_st. KCCG_1.7.0_HC NA12878MOD.imp.vc NA12878MOD.imp.v. MySubmission1 DRAGEN_DNASeq_N. DRAGEN_DNASeq_N. SIB SGI_3rd_filtration_rere NCBI_Magic_Thierry. JSLee_NA12878MOD. NCH_SUBMISSION2 Yotta2_GRCh37_lifto. DeepVariant_LowFre. DRAGEN_DNASeq_N. NCH_SUBMISSION1 NA12878MOD.imp.v. Sentieon_DNAseq_all. Sentieon_DNAseq_all. ADIscan_NA12878M Whole-exome variant. SeqsLab with GATK-. DRAGEN_DNASeq_N. Yotta1_GRCh37_lifto. BWA-MEM + GATK +. Sentieon_Exp_noded. mySubmission_GATK Beijing Institute of SeqsLab with GATK Sentieon_TNscope_ Sentieon_TNscope_ Sentieon_TNscope_ DRAGEN_DNASeq_ DRAGEN_DNASeq_ NovoalignFreebayes_ NovoalignGATKHC_ NovoalignFreebayes NovoalignGATKHC **INDEL F-score** Sentieon_DNAseq_nod... SGI_1st_result MySubmission2 BIG_NGSToolkit-... ADlscan_NA12878MOD.. NA12878MOD.imp.v3.r.. SGI_3rd_filtration_rere Yotta3_GRCh37_liftover... NCBI_Magic_ThierryMi... BIG_HXG-NA12878.vcf.gz Sentieon_Exp_nodedup... SGI_4th_filtration BWA-MEM + GATK +... SKESA_SRPRISM Sentieon_DNAseq_allm... GATK HaplotypeCaller... SeqsLab with GATK-... Yotta1_GRCh37_liftover... VW_submission_1 Bowtie2 + GATK... MySubmission1 GATK HaplotypeCaller... NA12878MOD.imp.v2.r.. B15GATK36 B8GATK36 Saphetor3 Saphetor2 test dummy dummy Challenge_submission_.. mySubmission_GATK test2 test Yotta2_GRCh37_liftover.. JSLee_NA12878MOD_fi.. NA1. D, NA12878MOD.imp.vcf.gz Experimental Unfiltered ADIscan_NA12878MOD. SIB Fest ADIscan_NA12878MOD. ADIscan_NA12878MOD. Sentieon_TNscope_Ton. Sentieon_TNscope_Ton. HULP-germ HULP-mosaic DRAGEN_DNASeq_NA1. DRAGEN_DNASeq_NA1. DRAGEN_DNASeq_NA1. DeepVariant DRAGEN_DNASeq_NA1. DRAGEN_DNASeq_NA1. DeepVariant_LowFrequ. NA12878MOD.imp.hard. NovoalignFreebayes_1 JLack_CCBR_Leidos_Sta Saphetor4 exome variant JSLee_NA12878MOD_fi Saphetor1 Submission1 DRAGEN_DNASeq_NA1. KCCG 1.7.0 HC Sentieon_DNAseq_allm. G3GATK36 NCH_SUBMISSION2 JLack_CCBR_Leidos_Sta. NovoalignGATKHC_2 SGI_2nd_filtration submission SeqsLab with GATK-. NCH_SUBMISSION1 Beijing Institute of. BIG_CelerSNP-. Sentieon_TNscope_Ton. entieon_TNscope_Ton. Bowtie2 + GATK Sentieon_DNAseq_std-Sentieon_Exp-BIG_CelerSNP. BIG_myPipeline NovoalignFreebayes_ JSLee_NA12878MOD_ NovoalignGATKHC_ KCCG_1.7.0_ DRAGEN_DNASeq_ Whole-

SNP F-score

B8GATK36 G3GATK36

GATK..

Saphetor4

B15GATK36

BIG_HXG

BIG_CelerSNP. CelerSNP.

ЪВ

KCCG_1.7.0_VD submission

dummy

Saphetor1 Bowtie2 + GATK.

Saphetor2 Saphetor3 dummy test2 test

test

HULP-germ

GATK..

Exp

Sentieon_

Bowtie2 + GATK.

BIG_myPipeline.

Performance - F-score

BIG_NGSToolkit-

Test

Submission1

Experimental.

0.9 0.8 0.7 0.6 0.5 0.4 0.3 0.2

A Pipeliner Demo...

Questions?

GATK - Variant Quality Score Recalibration (VQSR)

