Getting Genomics Done with R

Stephen A. Sefick

2017-03-28



Stephen A. Sefick Science Some Will 2017-03-28 1 /

Outline

Introduction

2 VCF Annotations

Using R/Bioconductor to filter vcf

4 Exercise (HW 7)

Topic

Introduction

VCF Annotations

Using R/Bioconductor to filter vcf

4 Exercise (HW 7)

• How to arrive at analysis ready variants?

4 / 23

Stephen A. Sefick Gesting Genomies Done with R 2017-03-28

- How to arrive at analysis ready variants?
- GATK HaplotypeCaller is "permissive"



Stephen A. Sefick Salum Gramma Done with E 2017-03-28

- How to arrive at analysis ready variants?
- GATK HaplotypeCaller is "permissive"
 - False Positives

2017-03-28

Stephen A. Sefick Cetting Genomics Done with R

- How to arrive at analysis ready variants?
- GATK HaplotypeCaller is "permissive"
 - False Positives
- How to separate True/False Positives

Stephen A. Sefick Country Generales Bone with IC 2017-03-28

- How to arrive at analysis ready variants?
- GATK HaplotypeCaller is "permissive"
 - False Positives
- How to separate True/False Positives
 - classification/machine learning

Stephen A. Sefick Getting Genomics Done with R

- How to arrive at analysis ready variants?
- GATK HaplotypeCaller is "permissive"
 - False Positives
- How to separate True/False Positives
 - classification/machine learning
 - filtering

- How to arrive at analysis ready variants?
- GATK HaplotypeCaller is "permissive"
 - False Positives
- How to separate True/False Positives
 - classification/machine learning
 - filtering
 - both

- How to arrive at analysis ready variants?
- GATK HaplotypeCaller is "permissive"
 - False Positives
- How to separate True/False Positives
 - classification/machine learning
 - filtering
 - both
- What software to use?

- How to arrive at analysis ready variants?
- GATK HaplotypeCaller is "permissive"
 - False Positives
- Output
 Output
 Output
 Description
 Description
 Output
 Description
 - classification/machine learning
 - filtering
 - both
- What software to use?
 - R

- How to arrive at analysis ready variants?
- GATK HaplotypeCaller is "permissive"
 - False Positives
- How to separate True/False Positives
 - classification/machine learning
 - filtering
 - both
- What software to use?
 - R
 - Bioconductor Project

CRAN

- CRAN
- Bioconductor- additional software repository

- CRAN
- Bioconductor- additional software repository
- Open Source Software For Bioinformatics

- CRAN
- Bioconductor- additional software repository
- Open Source Software For Bioinformatics
 - DNA micro-array

- CRAN
- Bioconductor- additional software repository
- Open Source Software For Bioinformatics
 - DNA micro-array
 - sequence

- CRAN
- Bioconductor- additional software repository
- Open Source Software For Bioinformatics
 - DNA micro-array
 - sequence
 - SNP

- CRAN
- Bioconductor- additional software repository
- Open Source Software For Bioinformatics
 - DNA micro-array
 - sequence
 - SNP
 - etc.

- CRAN
- Bioconductor- additional software repository
- Open Source Software For Bioinformatics
 - DNA micro-array
 - sequence
 - SNP
 - etc.
- started 2001

- CRAN
- Bioconductor- additional software repository
- Open Source Software For Bioinformatics
 - DNA micro-array
 - sequence
 - SNP
 - etc.
- started 2001
- 1294 user contributed packages

Installation similar to CRAN packages (automatic dependency resolution)

Stephen A. Sefick Gellung Conomics Done with E 2017-03-28 6

- Installation similar to CRAN packages (automatic dependency resolution)
 - Install/update Bioconductor

Stephen A. Sefick County Communications with E 2017-03-28 6 /

- Installation similar to CRAN packages (automatic dependency resolution)
 - Install/update Bioconductor
 - source("https://bioconductor.org/biocLite.R")

Stephen A. Sefick Sessing Sessing Votes with II 2017-03-28 6 /

- Installation similar to CRAN packages (automatic dependency resolution)
 - Install/update Bioconductor
 - source("https://bioconductor.org/biocLite.R")
 - biocLite()

6 / 23

Stephen A. Sefick Sauma Computer Units with E. 2017-03-28

- Installation similar to CRAN packages (automatic dependency resolution)
 - Install/update Bioconductor
 - source("https://bioconductor.org/biocLite.R")
 - biocLite()
 - Install packages

6 / 23

Stephen A. Sefick Grownia Done with IS 2017-03-28

- Installation similar to CRAN packages (automatic dependency resolution)
 - Install/update Bioconductor
 - source("https://bioconductor.org/biocLite.R")
 - biocLite()
 - Install packages
 - biocLite("VariantAnnotation")

Stephen A. Sefick Caving Generalist Date with R 2017-03-28

- Installation similar to CRAN packages (automatic dependency resolution)
 - Install/update Bioconductor
 - source("https://bioconductor.org/biocLite.R")
 - biocLite()
 - Install packages
 - biocLite("VariantAnnotation")
- 2 Documentation familiar and excellent

Stephen A. Sefick Cetting Genomics Done with R 2017-03-28

- Installation similar to CRAN packages (automatic dependency resolution)
 - Install/update Bioconductor
 - source("https://bioconductor.org/biocLite.R")
 - biocLite()
 - Install packages
 - biocLite("VariantAnnotation")
- Occumentation familiar and excellent
- In addition, most packages have vignettes

- Installation similar to CRAN packages (automatic dependency resolution)
 - Install/update Bioconductor
 - source("https://bioconductor.org/biocLite.R")
 - biocLite()
 - Install packages
 - biocLite("VariantAnnotation")
- Occumentation familiar and excellent
- In addition, most packages have vignettes
 - Vignettes are short "how-tos"

- Installation similar to CRAN packages (automatic dependency resolution)
 - Install/update Bioconductor
 - source("https://bioconductor.org/biocLite.R")
 - biocLite()
 - Install packages
 - biocLite("VariantAnnotation")
- 2 Documentation familiar and excellent
- In addition, most packages have vignettes
 - Vignettes are short "how-tos"
- Most object oriented providing consistent "feel" for bioconductor packages

Stephen A. Sefick Genomics Done with R. 2017-03-28

- Installation similar to CRAN packages (automatic dependency resolution)
 - Install/update Bioconductor
 - source("https://bioconductor.org/biocLite.R")
 - biocLite()
 - Install packages
 - biocLite("VariantAnnotation")
- 2 Documentation familiar and excellent
- In addition, most packages have vignettes
 - Vignettes are short "how-tos"
- Most object oriented providing consistent "feel" for bioconductor packages
 - CRAN mix of no OO, S3, S4, and newer R6

Stephen A. Sefick Getting Genomics Done with R

- Installation similar to CRAN packages (automatic dependency resolution)
 - Install/update Bioconductor
 - source("https://bioconductor.org/biocLite.R")
 - biocLite()
 - Install packages
 - biocLite("VariantAnnotation")
- 2 Documentation familiar and excellent
- In addition, most packages have vignettes
 - Vignettes are short "how-tos"
- Most object oriented providing consistent "feel" for bioconductor packages
 - CRAN mix of no OO, S3, S4, and newer R6
 - Flexibility is R's strength and weakness

6 / 23

Stephen A. Sefick Gaussia Gaussia Done with S 2017-03-28

- Installation similar to CRAN packages (automatic dependency resolution)
 - Install/update Bioconductor
 - source("https://bioconductor.org/biocLite.R")
 - biocLite()
 - Install packages
 - biocLite("VariantAnnotation")
- 2 Documentation familiar and excellent
- In addition, most packages have vignettes
 - Vignettes are short "how-tos"
- Most object oriented providing consistent "feel" for bioconductor packages
 - CRAN mix of no OO, S3, S4, and newer R6
 - Flexibility is R's strength and weakness
- 1 https://bioconductor.org/

Topic

Introduction

2 VCF Annotations

Using R/Bioconductor to filter vcf

Exercise (HW 7)



HaplotypeCaller includes/calculates annotations

Stephen A. Sefick Cetting Conomics Done with R

- HaplotypeCaller includes/calculates annotations
 - VCFs have a number of annotations

Stephen A. Sefick Caronic Concepts With R 2017-03-28

- HaplotypeCaller includes/calculates annotations
 - VCFs have a number of annotations
 - e.g., genotype quality, depth, etc.

Stephen A. Sefick Sching Geronnics Done with E 2017-03-28 8

- HaplotypeCaller includes/calculates annotations
 - VCFs have a number of annotations
 - e.g., genotype quality, depth, etc.
- Annotations change

Stephen A. Sefick Garanies Dane with K 2017-03-28

- HaplotypeCaller includes/calculates annotations
 - VCFs have a number of annotations
 - e.g., genotype quality, depth, etc.
- Annotations change
 - e.g., HaplotypeCaller's gVCFs with Reference Genotype Quality

Stephen A. Sefick Genting Genomics Done with R 2017-03-28

- HaplotypeCaller includes/calculates annotations
 - VCFs have a number of annotations
 - e.g., genotype quality, depth, etc.
- Annotations change
 - e.g., HaplotypeCaller's gVCFs with Reference Genotype Quality
 - Custom annotations

Stephen A. Sefick Genomics Done with R. 2017-03-28

- HaplotypeCaller includes/calculates annotations
 - VCFs have a number of annotations
 - e.g., genotype quality, depth, etc.
- Annotations change
 - e.g., HaplotypeCaller's gVCFs with Reference Genotype Quality
 - Custom annotations
- How do we use annotations?

- HaplotypeCaller includes/calculates annotations
 - VCFs have a number of annotations
 - e.g., genotype quality, depth, etc.
- Annotations change
 - e.g., HaplotypeCaller's gVCFs with Reference Genotype Quality
 - Custom annotations
- 4 How do we use annotations?
 - understanding the data

- HaplotypeCaller includes/calculates annotations
 - VCFs have a number of annotations
 - e.g., genotype quality, depth, etc.
- Annotations change
 - e.g., HaplotypeCaller's gVCFs with Reference Genotype Quality
 - Custom annotations
- 4 How do we use annotations?
 - understanding the data
 - filtering

- HaplotypeCaller includes/calculates annotations
 - VCFs have a number of annotations
 - e.g., genotype quality, depth, etc.
- Annotations change
 - e.g., HaplotypeCaller's gVCFs with Reference Genotype Quality
 - Custom annotations
- 4 How do we use annotations?
 - understanding the data
 - filtering
 - classification/machine learning

Stephen A. Sefick Getting Genomics Don

8 / 23

- HaplotypeCaller includes/calculates annotations
 - VCFs have a number of annotations
 - e.g., genotype quality, depth, etc.
- Annotations change
 - e.g., HaplotypeCaller's gVCFs with Reference Genotype Quality
 - Custom annotations
- 4 How do we use annotations?
 - understanding the data
 - filtering
 - classification/machine learning
- Non-model systems analysis ready variants

2017-03-28

Stephen A. Sefick Getting Genomics Done with R

- HaplotypeCaller includes/calculates annotations
 - VCFs have a number of annotations
 - e.g., genotype quality, depth, etc.
- Annotations change
 - e.g., HaplotypeCaller's gVCFs with Reference Genotype Quality
 - Custom annotations
- 4 How do we use annotations?
 - understanding the data
 - filtering
 - classification/machine learning
- Non-model systems analysis ready variants
 - hard-filtered call set

- HaplotypeCaller includes/calculates annotations
 - VCFs have a number of annotations
 - e.g., genotype quality, depth, etc.
- Annotations change
 - e.g., HaplotypeCaller's gVCFs with Reference Genotype Quality
 - Custom annotations
- 4 How do we use annotations?
 - understanding the data
 - filtering
 - classification/machine learning
- Non-model systems analysis ready variants
 - hard-filtered call set
 - Bootstrapped Variant Quality Score Re-calibration (VQSR) call set using HF as training/truth data

8 / 23

Stephen A. Sefick Caving Generalist Date with R 2017-03-28

- HaplotypeCaller includes/calculates annotations
 - VCFs have a number of annotations
 - e.g., genotype quality, depth, etc.
- Annotations change
 - e.g., HaplotypeCaller's gVCFs with Reference Genotype Quality
 - Custom annotations
- 4 How do we use annotations?
 - understanding the data
 - filtering
 - classification/machine learning
- Non-model systems analysis ready variants
 - hard-filtered call set
 - Bootstrapped Variant Quality Score Re-calibration (VQSR) call set using HF as training/truth data
- Today concentrate on hard-filtering SNPs

4□ > 4回 > 4 = > 4 = > = 900

Context specific

- Context specific
 - SNPs

- Context specific
 - SNPs
 - INDELs

- Context specific
 - SNPs
 - INDELs
- GATK hard-filtering recommendations

- Context specific
 - SNPs
 - INDELs
- GATK hard-filtering recommendations
- These are recommendations, developed with human data, and might/likely will need to be modified based on the data

Variant	#	Annotation	Remove If
Both	1	DP	min=empirical; max=5 or 6 sigma
	2	GQ (or RGQ)	empirical
SNP	3	QD	< 2.0
	4	MQ	< 40
	5	FS	> 60
	6	SOR	> 3.0
	7	MQRankSum	< -12.5
	8	ReadPosRankSum	< -8.0

Min - no empirical guidance- look at plots

- Min no empirical guidance- look at plots
- 2 Max 5 or 6x the standard deviation

- Min no empirical guidance- look at plots
- Max 5 or 6x the standard deviation
- How many reads cover a position

10 / 23

Stephen A. Sefick Saving Senomies Vone with E. 2017-03-28

- Min no empirical guidance- look at plots
- 2 Max 5 or 6x the standard deviation
- How many reads cover a position
 - GATK Caveat- slightly different from raw depth of coverage

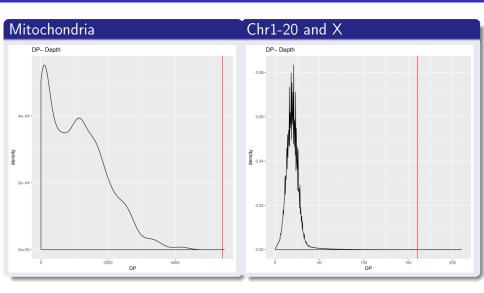
10 / 23

Stephen A. Sefick Couring Conomics Done with R 2017-03-28

- Min no empirical guidance- look at plots
- Max 5 or 6x the standard deviation
- How many reads cover a position
 - GATK Caveat- slightly different from raw depth of coverage
 - QC of the caller will remove reads (MAQ)

10 / 23

Stephen A. Sefick Gesting Genomics Done with R 2017-03-28

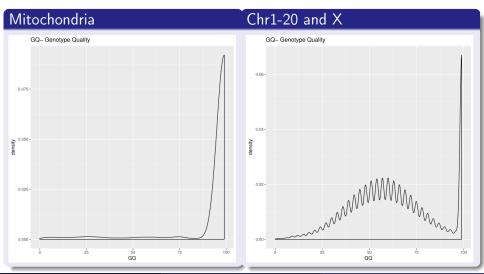


 Stephen A. Sefick
 Gowing Committee 20th of the C

2 Genotype quality (GQ); Reference GQ (RGQ)

Phred scaled probability of incorrect genotype

20 - 0.01; 30 - 0.001; 40 - 0.0001

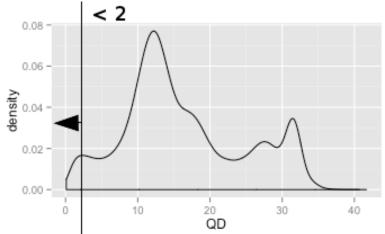


Stephen A. Sefick Gatting Genomics Done with R 2017-03-28

12 / 23

3 Variant quality/allele depth (QD)

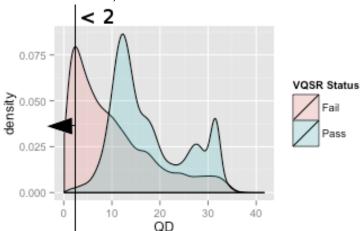
- Variant Quality (QUAL) is the phred scaled probability that the variant is wrong.
- allele depth is actual depth of each observed allele (How many actual reads; in contrast to DP).



13 / 23

3 Variant quality/allele depth (QD)

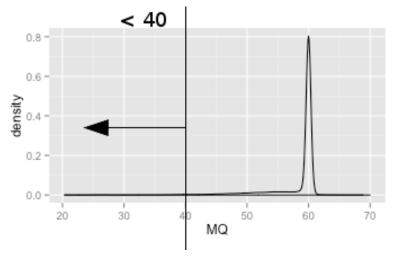
- Variant Quality (QUAL) is the phred scaled probability that the variant is wrong.
- allele depth is actual depth of each observed allele (How many actual reads; in contrast to DP).



€ 99€

4 Root mean square mapping quality (MQ)

• phred scaled probability that the mapping position is wrong

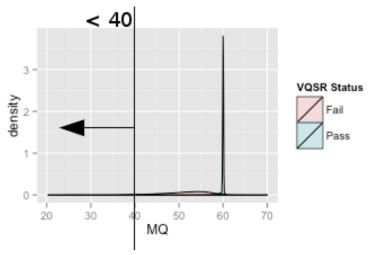


14 / 23

Stephen A. Sefick Ground Commission William 2017-03-28

4 Root mean square mapping quality (MQ)

• phred scaled probability that the mapping position is wrong

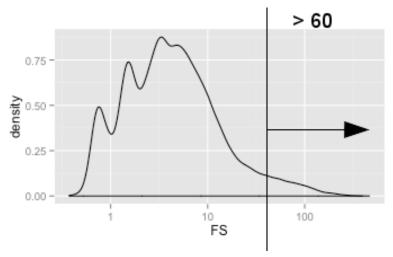


↓□→ ←□→ ←□→ ←□→ □ ♥♀○

14 / 23

5 Fisher strand bias (FS)

phred scaled probability ALT on forward or reverse strand more or less than REF

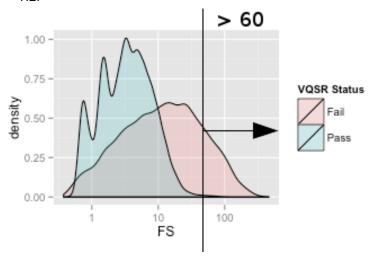


15 / 23

Stephen A. Sefick Grand Constitution Stephen A. Sefick 2017-03-28

5 Fisher strand bias (FS)

phred scaled probability ALT on forward or reverse strand more or less than REF

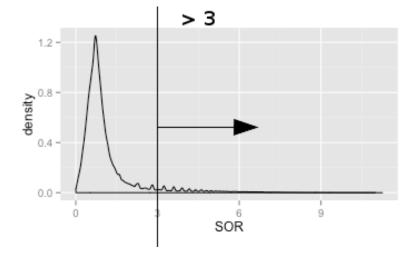


15 / 23

Stephen A. Sefick Group Consules Dans William 2017-03-28

6 Strand odds ratio (SOR)

- similar to FS, but updated for high coverage (NGS)
 - Ratio of reads that cover both alleles

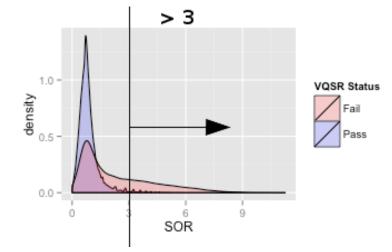


16 / 23

Stephen A. Sefick Crasing Genomies Done with R. 2017-03-28

6 Strand odds ratio (SOR)

- similar to FS, but updated for high coverage (NGS)
 - Ratio of reads that cover both alleles

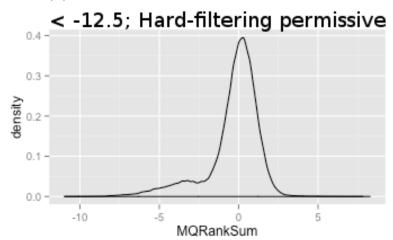


イロト イ御 ト イミト イミト (章) からぐ

16 / 23

7 MQ rank sum test (MQRankSum)

- test compares MAQ ALT to REF
 - (-) Alt lower MAQ
 - (+) Ref lower MAQ



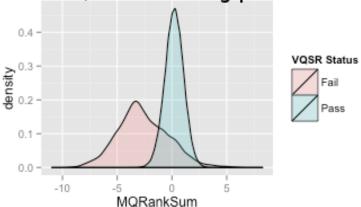
Stephen A. Sefick Galling Gallonia Dans Will E 2017-03-28

17 / 23

7 MQ rank sum test (MQRankSum)

- test compares MAQ ALT to REF
 - (-) Alt lower MAQ
 - (+) Ref lower MAQ

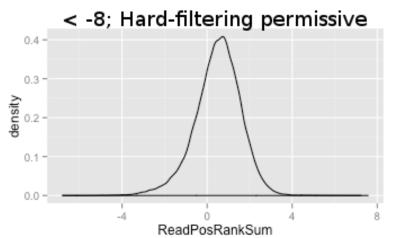
< -12.5; Hard-filtering permissive



Stephen A. Sefick Carini, Garantia Concernity 2017-03-28 17 / 23

8 Read position rank sum test (ReadPosRankSum)

- test for positional effects
 - (-) Alt close to end of read
 - (+) Ref close to end of read



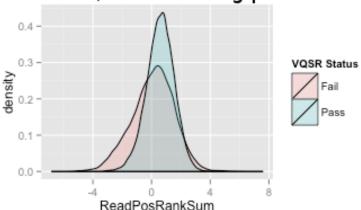
Stephen A. Sefick Garding Congress Days with E 2017-03-28

18 / 23

8 Read position rank sum test (ReadPosRankSum)

- test for positional effects
 - (-) Alt close to end of read
 - (+) Ref close to end of read

< -8; Hard-filtering permissive



Stephen A. Sefick Getting Genomics Done with R 2017-03-28 18 / 23

Hard-filtering Summary (SNPs and INDELS)

Variant	Annotation	Remove If
Both	DP	min=empirical; max=5 or 6 sigma
	GQ (or RGQ)	empirical
SNP	QD	< 2.0
	MQ	< 40
	FS	> 60
	SOR	> 3.0
	MQRankSum	< -12.5
	ReadPosRankSum	< -8.0
INDELs	QD	< 2.0
	ReadPosRankSum	< -20.0
	InbreedingCoeff (> 10 samples)	< -8.0
	FS	< 200.0
	SOR	> 10.0

 Stephen A. Sefick
 Cashing Generality Density (I)
 2017-03-28
 19 / 23

Topic

Introduction

VCF Annotations

3 Using R/Bioconductor to filter vcf

4 Exercise (HW 7)

• Could write a script in favorite language.

 Stephen A. Sefick
 Octobre Stephen A. Sefick
 2017-03-28
 21 / 23

- Could write a script in favorite language.
 - ullet Know exactly what you did (+)

21 / 23

Stephen A. Sefick Ground Communication Will R 2017-03-28

- Could write a script in favorite language.
 - Know exactly what you did (+)
 - Time spent engineering software (-)

21 / 23

Stephen A. Sefick Goodings Genomics (Denocard) 13: 2017-03-28

- Could write a script in favorite language.
 - Know exactly what you did (+)
 - Time spent engineering software (-)
- Hard Work already done

21 / 23

Stephen A. Sefick Geology Genomics Done with Its 2017-03-28

- Could write a script in favorite language.
 - Know exactly what you did (+)
 - Time spent engineering software (-)
- Hard Work already done
 - Bioconductor

21 / 23

Stephen A. Sefick Gatting Genomics Done with 8 2017-03-28

- Could write a script in favorite language.
 - Know exactly what you did (+)
 - Time spent engineering software (-)
- Hard Work already done
 - Bioconductor
 - VariantAnnotation

21 / 23

Stephen A. Sefick Gatting Genomics Done with R 2017-03-28

- Could write a script in favorite language.
 - Know exactly what you did (+)
 - Time spent engineering software (-)
- Hard Work already done
 - Bioconductor
 - VariantAnnotation
 - general parsing and filtering

- Could write a script in favorite language.
 - Know exactly what you did (+)
 - Time spent engineering software (-)
- Hard Work already done
 - Bioconductor
 - VariantAnnotation
 - general parsing and filtering
- Consistent interface

- Could write a script in favorite language.
 - Know exactly what you did (+)
 - Time spent engineering software (-)
- Hard Work already done
 - Bioconductor
 - VariantAnnotation
 - general parsing and filtering
- Consistent interface
 - Learn 1 piece of software and reuse

- Could write a script in favorite language.
 - Know exactly what you did (+)
 - Time spent engineering software (-)
- Hard Work already done
 - Bioconductor
 - VariantAnnotation
 - general parsing and filtering
- Consistent interface
 - Learn 1 piece of software and reuse
- Custom filters

- Could write a script in favorite language.
 - Know exactly what you did (+)
 - Time spent engineering software (-)
- Hard Work already done
 - Bioconductor
 - VariantAnnotation
 - general parsing and filtering
- Consistent interface
 - Learn 1 piece of software and reuse
- Custom filters
 - flexible annotations (e.g., RGQ)

- Could write a script in favorite language.
 - Know exactly what you did (+)
 - Time spent engineering software (-)
- Hard Work already done
 - Bioconductor
 - VariantAnnotation
 - general parsing and filtering
- Consistent interface
 - Learn 1 piece of software and reuse
- Custom filters
 - flexible annotations (e.g., RGQ)
 - New annotations just "show up"

Topic

Introduction

VCF Annotations

Using R/Bioconductor to filter vcf

4 Exercise (HW 7)



Extract, Filter, and Plot

- Exercise folder on asc
 - Scripts: 1_initial_annotation_plot.sh; 2_filter_and_plot.sh
 - Data: D PseudoFS14 16
 - UsefulBioinformaticScripts
- Edit "Variables" in 1_initial_annotation_plot.sh

```
script_dir=${HOME}/Exercise/UsefulBioinformaticScripts
data_dir=${HOME}/Exercise/D_PseudoFS14_16
out_dir=${HOME}/Exercise
```

- save script and run
- Inspect graphs and decide upon filtering thresholds
- add variable definitions in 2 to 2_filter_and_plot.sh
- Fdit

```
##Filtering Parameters
##filtering Parameters
##this is
$(script_dir)/filter_SNPs_GATK_hard_filter.CHUNKS.R -I $(out_dir)/$(vcf1).gz -T
$(out_dir)/$(vcf1).gz.tbi -0 $(out_dir)/$(vcf1).filtered.vcf -C 10000 --QD=2
--FS=60 --SOR=3 --MQRankSum=-8 --min_Depth=4 --max_Depth=32 --Genotype_Quality=20
$(script_dir)/filter_SNPs_GATK_hard_filter.CHUNKS.R -I $(out_dir)/$(vcf2).gz -T
$(out_dir)/$(vcf2).gz.tbi -0 $(out_dir)/$(vcf2).filtered.vcf -C 10000 --QD=-FS=60 --SOR=3 --MQRankSum=-8 --min_Depth=4 --max_Depth=32 --Genotype_Quality=20
```

- save script and run
- inspect graphs and write up.