

Implementation of a copy number analysis pipeline for shallow sequencing in Galaxy framework

Van Lieshout S., Scheinin I., Sie D., Fijneman R.J.A., Ylstra B. Department of Pathology, VU University Medical Center, Amsterdam, The Netherlands

Get (

Introduction

DNA copy number aberrations are a hallmark of cancer and can be quantified by shallow wholegenome sequencing (WGS). A robust method has been developed ⁽¹⁾ that detects copy number aberrations. This method is currently being integrated into the Galaxy platform ⁽³⁾ within the Center for Translational Molecular Medicine - Translational Research IT (CTMM TraIT) project ⁽⁴⁾.

- ✓ highly concordant with array CGH at considerably lower cost ⁽¹⁾
- ✓ also available as a **Bioconductor** package, QDNAseq⁽²⁾

Methods

Sequence reads are **binned and counted in non** overlapping windows of 15kb and a combined LOESS correction for mappability and GC content is applied. Then a **comprehensive**

- the popular BAM format as input
- reports results in a clear and concise HTML **based view** within Galaxy itself
- ✓ R data structure file for downstream analysis available
- Zipped archive with all output

Galaxy		Ana	alyze Data Workflow Shared Data - Visualization -	Fm	bedd	led re	nor	t
ols 🔤	Load Data							
search tools	8	QDNAse	eq Results	7		Unnamed history		
<u>t Data</u> nd Data		About this an	alysis		=	1.2 GB 32: QDNAseq: test	Q 🗞 🗩	
t-Over ct Manipulation			es access to all results. To have a local copy	ı just download tr	ne zipfile	31: QDNAseq: test	• • ×	
<u>ter and Sort</u> in, Subtract and Group	2	with all output (6 Settings	.25 MB)			25: HCT116_GTGGCC_L008_R		
nvert Formats tract Features			•			<u>24:</u>	• 🖋 🗙	
ch Sequences		setting	value			HCT116_GCCAAT_L008_R1	L_001.bam	
ch Alignments		AnalysisName	test			12: flagstat on data 10	۰ 🖋 🗙	
<u>t Genomic Scores</u> erate on Genomic Inte	rvale	BinSize (kb)	15			<u>10: Bowtie2 on data 9:</u>	● # ×	
itistics	IVais	undoSD	1			aligned reads		
aph/Display Data	=	InputBam	dataset_5846.dat (HCT116_GCCAAT_L008	R1 001.bam)		<u>9: fastq_example_1000.fq</u>	۲	
gional Variation		InputBam	dataset 5847.dat (HCT116 GTGGCC L008			2: fasta_example.fa		
<u>Itiple regression</u> Itivariate Analysis		pacban				<u>1:</u>	● 	
olution		Output files				<u></u> <u>bam_example_chr22.bam</u>		
tif Tools								
STA manipulation		<u>15kbp_QDNAseqCopyNumbers.rds</u>						

filtering excludes genomic regions from both ENCODE project blacklists and a novel blacklist we have developed based on sequence depth of 38 individuals from the 1000 Genomes project. Procedures for copy number detection have been **optimized for** use in combination with DNA isolated from **formalin-fixed paraffin**embedded (FFPE) samples.



PASTA manipulation NGS: QC and manipulation NGS: Mapping NGS: RNA Analysis NGS: SAM Tools NGS: GATK Tools (beta) NGS: Simulation Phenotype Association QDNAseq Suite NGS: Annotation Workflows	QC	15kbp_QDNAseqRe list with segm Results: overv Sample HCT116_GCCAAT ONASE out this an page provide	ented/called regions for each sample, 0 MB iew L008_R1_001.bam eq Results					
Settings								
	set	ting	value					
	Ana	alysisName	test					
	Bin	Size (kb)	15					
	und	doSD	1					
	Inn	Bam	dataset_5846.dat (HCT116_GCCAAT_L008_R1_001.bam)					
Εσ=0.1	$52, \hat{\sigma}_{\Delta} = 0.161$	Bam	dataset_5847.dat (HCT116_GTGGCC_L008_R1_001.bam)					
		ut files						
		DNAseq ob ODNAseq	CopyNumbers.rds oject with copy numbers per bin, 7.93 MB Regions.rds mented/called regions for each sample, 0 MB					

and formalin-fixed specimens by whole-genome sequencing: improved correction of systematic biases and exclusion of problematic regions (submitted).

- 2. <u>http://www.bioconductor.org/</u>
- 3. <u>http://galaxyproject.org</u>
- 4. <u>http://www.ctmm-trait.nl/</u>