INTRODUCTIONTO NEXTELOW

Elixir workflows workshop - I Dec 2021 Paolo Di Tommaso - CTO Seqera Labs







• 9.30 - 9.50: Intro to Nextflow core concepts Paolo Di Tommaso, Seqera Labs



Evan Floden, Seqera Labs



10.00 - 12.00: Nextflow hands-on tutorial Luca Cozzuto, Centre for Genomic Regulation (CRG)

AGENDA

• 9.50 - 10.00: Deploy in the cloud with Nextflow Tower

QUICK BIO



- Paolo DiTommaso
- Software engineer
- Creator & maintainer of Nextflow project
- CTO & Co-founder Seqera Labs

GENOMICS WORKFLOWS

- Data analysis applications to extract information from large genomic datasets (TB)
- Embarrassingly parallelisation, can spawn 100s-100k jobs over distributed cluster
- Mash-up of many different tools and scripts
- Complex executions and dependencies graphs → very fragile ecosystem



- 70 tasks
- 55 custom scripts
- 39 software tools & libraries

Complexity \Rightarrow instability \Rightarrow not reproducible result

Steinbiss et al., Companion parassite genome annotation pipeline, DOI: 10.1093/nar/gkw292



THE PROBLEM WITH REPRODUCIBILITY

A comprehensive analysis of the usability and archival stability of omics computational tools and resources

Serghei Mangul, et al. BioRxiv **doi:** https://doi.org/10.1101/452532

We found that 26% of all omics software resources are currently *not accessible* through URLs published in the paper.

Among the tools selected 49% were deemed "difficult to install," and 28% of the tools failed to be installed due to problems in the implementation.



Nextflow enables reproducible computational workflows

Paolo Di Tommaso, Maria Chatzou, Evan W Floden, Pablo Prieto Barja, Emilio Palumbo & Cedric Notredame 🔀

NATURE BIOTECHNOLOGY VOLUME 35 NUMBER 4 APRIL 2017



Comparison of the Companion pipeline annotation of *Leishmania infantum* genome executed across different platforms *

Platform	Amazon Linux	Debian Linux	Mac OSX
Number of chromosomes	36	36	36
Overall length (bp)	32.032.223	32.032.223	32.032.223
Number of genes	7.781	7.783	7.771
Gene density	236,64	236,64	236,32
Number of coding genes	7.580	7.580	7570
Average coding length (bp)	1.764	1.764	1.762
Number of genes with multiple CDS	113	113	111
Number of genes with known function	4.147	4.147	4.142
Number of t-RNAs	88	90	88

* Di Tommaso P, et al., Nextflow enables computational reproducibility, Nature Biotech, 2017

MAIN CHALLENGES

- Scalability & parallelisation
 - Enable portability
- Guarantee reproducibility

THE **nextflow** FUNDAMENTALS FOR SCALEABLE GENOMIC WORKFLOWS

Write code in any language



Declarative parallelisation



Deploy everywhere





Isolate dependencies with containers



Version control





Open source

••• # # one-liner to convert a BAM file to a FASTQ file

samtools view file.bam \

SIMPLE EXAMPLE

awk 'BEGIN {FS="\t"} {print "@" \$1 "\n" \$10 "\n+\n" \$11}' > file.fq

NEXTFLOW PROCESS

awk 'BEGIN {FS="\t"} {print "@" \$1 "\n" \$10 "\n+\n" \$11}' > file.fq

NEXTFLOW WORKFLOW

•••

workflow {

channel.fromPath("/data/sample.bam") |

awk 'BEGIN {FS="\t"} {print "@" \$1 "\n" \$10 "\n+\n" \$11}' > file.fq

n") | bam_to_fastq

IMPLICIT PARALLELISM

•••

workflow {

channel.fromPath("/data/*.bam")

awk 'BEGIN {FS="\t"} {print "@" \$1 "\n" \$10 "\n+\n" \$11}' > file.fq

bam_to_fastq

IMPLICIT PARALLELISM

channel.fromPath("/data/*.bam")



FASTQ files

CONTAINERISATION



- Nextflow envisioned the use of software containers to fix computational reproducibility
- Mar 2014, support for Docker
- Dec 2016, support for Singularity
- Jan 2020, support for Podman







MULTI-PLATFORM















Azure















EXECUTION REPORT

X Nextflow Report Summary Resources Tasks

Nextflow workflow report [trusting_cuvier] (resumed run)

Workflow execution completed successfully!

Run times

Fri Apr 27 23:19:53 CEST 2018 - Sat Apr 28 03:18:15 CEST 2018 (completed a day ago, duration: 3h 58m 21s)

5329 succeeded

Nextflow command

nextflow run main.nf -profile crg --std_align=true --default_align=true --align_method=CLUSTALO,MAFFT -tree_method=CLUSTAL0,MAFFT_PARTTREE --seqs=/users/cn/egarriga/datasets/homfamClustalo/seqs/*.fa -refs=/users/cn/egarriga/datasets/homfamClustalo/refs/*.ref -with-report -with-trace -resume -bg

CPU-Hours	156.6 (31.5% cached, 4.6%
Launch directory	/nfs/users2/cn/egarriga/p
Work directory	/nfs/users2/cn/egarriga/p
Project directory	/nfs/users2/cn/egarriga/p
Script name	main.nf
Script ID	6ff267a42e50448d41927a6e
Workflow session	087c9bc8-e488-4311-88aa-9
Workflow profile	crg
Workflow container	<pre>cbcrg/regressive-msa:v0.2</pre>
Container engine	singularity
Nextflow version	version 0.28.2, build 4782 (06-0

[trusting_cuvier]

2849 cached

failed)

projects/dpa_cp

projects/dpa_cp/work

rojects/dpa_cp

5a9787fc

-961138c42fd6

04-2018 12:25 UTC)

EXECUTION REPORT



TIMELINE CHART

Processes execution timeline

Processes executio	numenne									
Launch time: 15 Jun 20 Elapsed time: 49m 9s	016 15:03									
downloadPoference (1)	5.7s/549.2 GB	I I	I I	1	I I	I I	I I	I I	1	1
downloadSPA (1)	59.9s / 1.2 GB	1	I.	1	I	1	1	1	1	
docompross Poforonco (1)	1.5s/-		1	I	1	I	1	1		1
indexPoforonco (1)	5.7s / 80.9 MB	:	l	i	: 1	l	:	i	:	:
ovtractSPA (1)	1m 11s / 150.5	MB	l I	I	l I	I I	1	l I	ı I	1
extractora (1)	<mark>3</mark> 6.1s/8.1	GB	I I	1	I I	1	1	I I	1	1
urini (1)	40.2s/3	31.6 MB	I	1	1	1	1	1	1	1
fitersking (1)	2m 14	<mark>4s</mark> / 8.6 GB	I	1	1	I		1		
	38m 3	31s / 8.9 GB	i I	1	L. L.	i I	1	1		1
alignBoads kms (1)	1	10 m 17s / 2 GB	l I		I I	I	1	I I	1	1
alignReads_KIIC(1)	1	1	I I	1m 52s / 53	38.9 MB	I I		I I	1	1
indexAlignment_kmc(1)	1	1		8.85	/ 36.2 MB	1	1	1	1	1
coll/variants kms (1)	1	:	1	20r	n 16s / 89.2 MB	I	:	1	: 1	:
alignPoods khmor (1)	1	1	l I	1	I I	I I	1	I I	1m 23s	/ 1.8 GB
anglikeaus_kiiner (1)	1	1	I I	1	1	I I	1	1	22	2s / 524.4 MB
indexAlignment_kimer (1)	1	1	I	1	1	I	1	1	. 2	2.2s / 36.2 MB
coll/orionts kbmsr(1)	i I	1	1	1	I	I	1	I		5m 21s / 89.1 MB
Call Valiants_KIIIIer (1)	05	10	15	20	25	30	35	40	45	50

```
main.nf — ~/projects/callings-nf
• • •
  main.nf
   log.info """\
   CALLINGS – NF v 1.0
   _____
                             ____
   genome : $params.genome
   reads : $params.reads
   variants : $params.variants
   blacklist: $params.blacklist
   results : $params.results
   gatk : $params.gatk
   GATK
               = params.gatk_launch
   genome_file = file(params.genome)
   variants_file = file(params.variants)
   blacklist_file = file(params.blacklist)
   reads_ch = Channel.fromFilePairs(params.reads)
   process '1A_prepare_genome_samtools' {
    tag "$genome.baseName"
    input:
       file genome from genome_file
     output:
       file "${genome}.fai" into genome_index_ch
    script:
    samtools faidx ${genome}
   process '1B_prepare_genome_picard' {
    tag "$genome.baseName"
     input:
        file genome from genome_file
    output:
        file "${genome.baseName}.dict" into genome_dict_ch
     script:
    PICARD=`which picard.jar`
     java -jar \$PICARD CreateSequenceDictionary R= $genome 0= ${genome.baseName}.dict
     .....
                                      LF UTF-8 Nextflow 🎾 master 🔸 🛧 主 4 files
main.nf* 93:1
```

EDITORS

		main.nf	
L)	🖹 ρορι	u.sh	r 🗆 🗠
<u> </u>	47		New Control of the Second Seco
0	49	CALLINGS - NF v 1.0	Series STREAM STREAM AND STREAM STREA
~	50		
ទទ	51	genome : \$params.genome	
20	52	reads : \$params.reads	
~	54	blacklist: \$params.blacklist	ESTIMA - ESTIMATION - ESTIMATION
B	55	results : \$params.results	¹⁰ Systems for some for ¹⁰ Systems for style and some system for some system for some
	56	gatk : \$params.gatk	Constant in the second se
Ċ?	57		Construction of the second states of the second sta
	59	/*	*** Comparison of the second of the ************************************
	60	* Parse the input parameters	Constant Constant in the state of a series
	61	*/	
	62	CATK - parama anth launch	
	63 64	denome file = file(params.genome)	March 1998 Block State States
	65	variants_file = file(params.variants)	The second free must ended on an interview of the must ended on the end of the product of the large second secon
	66	<pre>blacklist_file = file(params.blacklist)</pre>	Difference of the second secon
	67	reads_ch = Channel.fromFilePairs(params.reads)	
	68 60		The second secon
	70	/*****	Bankuroka sureko karke-
	71	* PART 1: Data preparation	
	72	*	Original and the second s
	73	* Process 1A: Create a FASTA genome index (.fai) with samtools for GATK	The second s
	74	*/	Beneficial and a second s
	76	process '1A_prepare_genome_samtools' {	
	77	tag "\$genome.baseName"	ABB/ AF TO Y ROUGER BY AN A STATE
	78		A CONTRACTOR OF A CONTRACT OF
	79	input:	The second secon
	81	Tite genome from genome_Tite	Receiver and an
	82	output:	
	83	<pre>file "\${genome}.fai" into genome_index_ch</pre>	The second secon
	84	comint.	Personal Control and Contro
	86		[Annual A. Sana and A. San
	87	<pre>samtools faidx \${genome}</pre>	
	88		and a set of a set of more and the set of th
	89	}	The second secon
	90 91		Contractory of the second second
	92	/*	
	93	* Process 1B: Create a FASTA genome sequence dictionary with Picard for GATK	
	94	*/	1. All and the second s
	95 96	process '18 prepare genome picard' {	
	97	tag "\$genome.baseName"	
	98		
	99	input:	
	100	Tile genome from genome_file	
	102	file "\${genome.baseName}.dict" into genome dict ch	
	103		
	104	script:	
	105	PICAPD-`which picard jar`	
	107	iava -iar \\$PICARD CreateSequenceDictionary R= \$genome O= \${genome.baseName}.dict	
	108	IIII	
**	109	·}	
~	110		
₽ mast	ter* 🗘 0,	Ln 1, Col 1 Spaces: 4 UTF-8 LF	Nextflow 🙂

SUMMARY

- Data analysis reproducibility is hard and it's often underestimated.
- and provides support for community and industry standards.
- deployment logic, enabling self-contained workflows.
- reproducible manner with a single command.

Nextflow does not provide a magic solution but enables best-practices

• It strictly separates the application logic from the configuration and

• Applications can be easily deployed across different environment in a

• The functional/reactive model allows applications to scale with ease.