Benefits of Integrative Methods on Analysis of Low Coverage NGS Data

Dragana Dudić

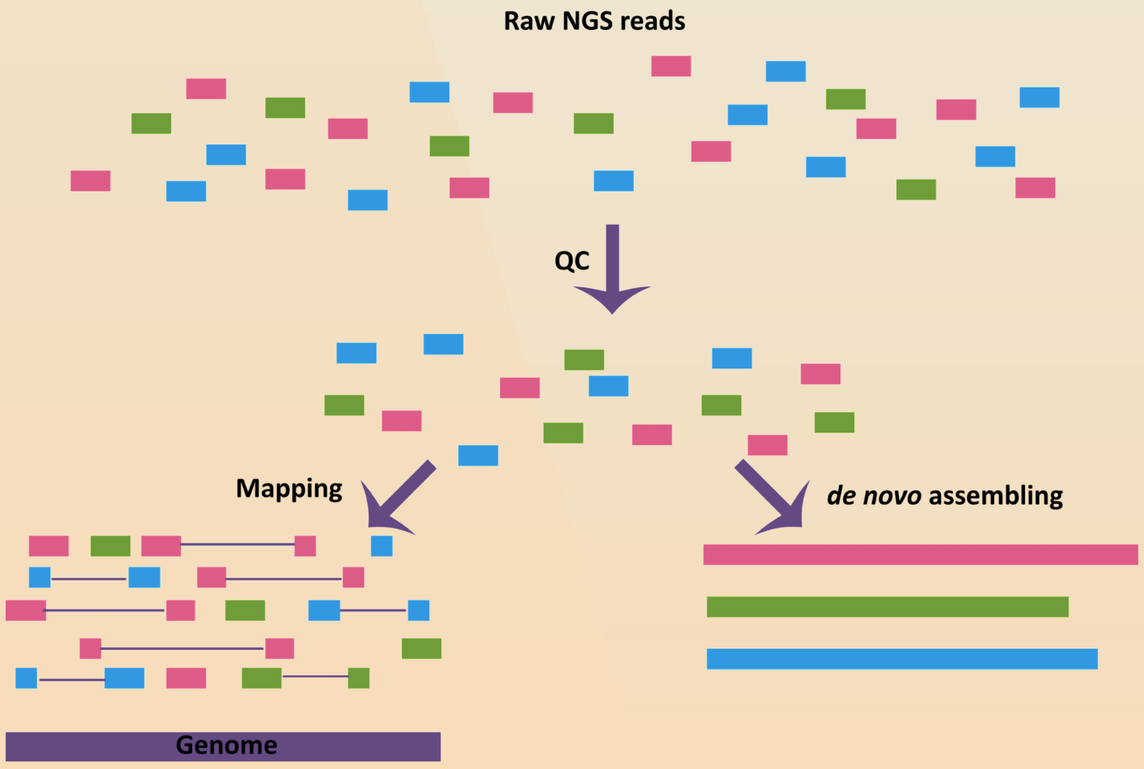
Center for Data Mining and Bioinformatics, Faculty of Agriculture, University of Belgrade, Zemun-Belgrade, Serbia

INTRODUCTION

Next generation sequencing (NGS) technologies provide a way of studying the structure of genetic material, both DNA and RNA. With repetitive nature of some genomes, a high number of multi-mapped reads and high duplication rate need to be addressed. This issue is even more critical for low coverage NGS data, when the need to get as much as possible information arises. We tested different bioinformatics tools on low coverage maize total transcriptome data, setting an optimal integrative approach to achieve the lowest duplicate reads rates with the highest percentage of reads mapped to exon regions.

MATERIAL AND METHODS

- 35-151bp PE total RNA-seq data
- Raw read depth: 0.09
- Two quality control (QC) steps:
  - QC evaluation: FastQC
  - QC preprocessing:
    - BLAST and TopHat for removing contamination
    - Trimmomatic for removing index adapters and perform sliding window trimming with average quality threshold set to 5.



- Mapping
  - Splice aware tools: TopHat2, Subread and STAR.
  - Specific parameters : insert size set to 130, standard deviation set to 50
  - Other parameters are equalized: introns up to 100,000, max 8 number of mismatches
- *de novo* assembling
  - Transcriptome data suitable tools: TransAbySS (TA), Oases (O) and Trinity (TR).
  - Parameters: k-mers ranging from 15 to 31, minimal contig length 30.
- Data processing
  - Standard method: mapping (M)
  - Integrative methods: assembling followed by mapping (AM), assembling with repetitive sequences followed by mapping (ARM), and mapping followed by assembling followed by mapping (MAM).

RESULTS

- Initial dataset: 693242 reads
- Clean dataset (after QC): 649792 reads

Assembling – determining the best k-mer

map %	TA 15	TA 17	TA 19	TA 21	TA 23	TA 25	TA 27	TA 29	TA 31
	99.0	98.6	98.8	**98.9**	**98.9**	**98.8**	98.8	98.7	98.7

avg contig	TA 15	TA 21	TA 23	TA 25
	61.6	144.9	149.1	**151.8**

map %	O 15	O 17	O 19	O 21	O 23	O 25	O 27	O 29	O 31
	44.7	45.7	49.2	87.1	91.4	**91.6**	91.4	91.3	91.1

map %	TR 15	TR 17	TR 19	TR 21	TR 23	TR 25	TR 27	TR 29	TR 31
	98.3	98.4	98.4	**98.5**	**98.6**	**98.6**	**98.6**	98.4	98.3

avg contig	TR 21	TR 23	TR 25	TR 27
	196.5	195.0	**211.6**	210.1

M method

Mapping tool	map %
Tophat2	95.4
Subread	**98.8**
STAR	**99.1**

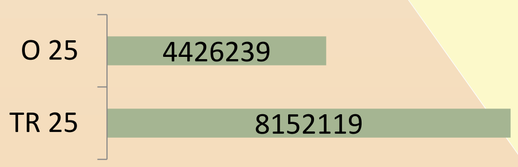
	Subread	STAR
Duplicate reads (D) %	57,22	**54,50**
Uniquely mapped reads (U) %	**23,13**	19,39
Reads mapped to exons (E) %	46,57	**57,38**
Reads mapped to introns (I) %	21,96	**12,01**

AM method

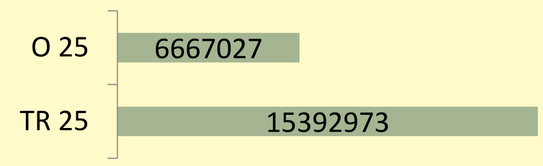
TA 25	map%	D %	U%	E%	I %
	99.05	6.61	82.53	53.89	27.74

O 25	map%	D %	U%	E%	I %
	98.94	7.45	82.58	**66.39**	**18.94**

TR 25	map%	D %	U%	E%	I %
	99.7	**5.25**	**85.31**	52.96	29.77



Exonic regions information



Total contigs length

ARM method

TA 25	map%	D %	U%	E%	I %
	99.41	6.21	83.14	51.81	32.77

O 25	map%	D %	U%	E%	I %
	98.13	9.58	79.22	61.93	21.67

TR 25	map%	D %	U%	E%	I %
	99.9	**3.33**	**90.97**	**68.81**	**16.47**

MAM method

TR 25	map%	D %	U%	E%	I %
	99.76	15.42	75.91	46.62	38.02

CONCLUSION

For low coverage NGS data derived from repetitive genomes, we advise using integrative methods like AM, ARM and MAM, depending on level of repetitiveness of sequenced data.

CONTACT

ddragana@agrif.bg.ac.rs