Genome Assembly Statistics

Chumpol Ngamphiw, Ph.D.

National Center for Genetics Engineering and Biotechnology (BIOTEC)
National Science and Technology Development Agency (NSTDA)

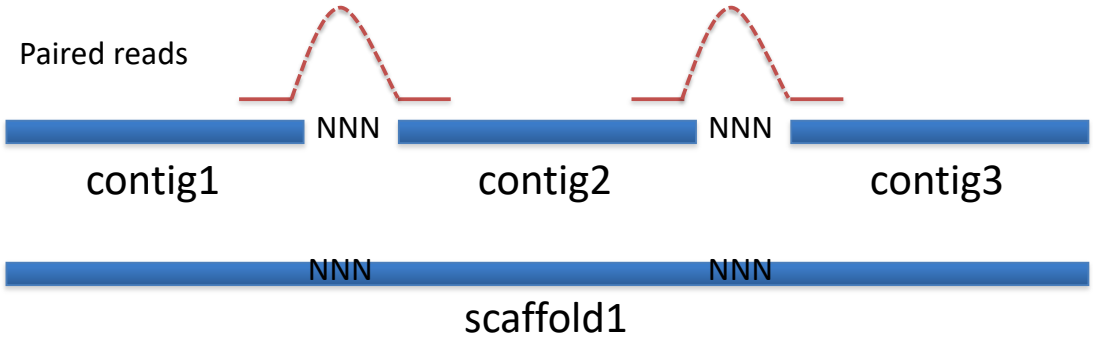
"Genome assembly and annotation" : August 6-9, 2018 @ KUPS

Outline

* Genome Assembly output
* Metrics for Genome Assembly
* QUAST : quality assessment tool for genome assemblies
* How to compare the quality of assemblers ?

Contig & Scaffold

- Contig : (from contiguous) is a set of overlapping DNA segments that together represent a consensus region of DNA
- Scaffold : several contigs stitched together with NNNs in between



https://en.wikipedia.org

Contigs : FASTA-format

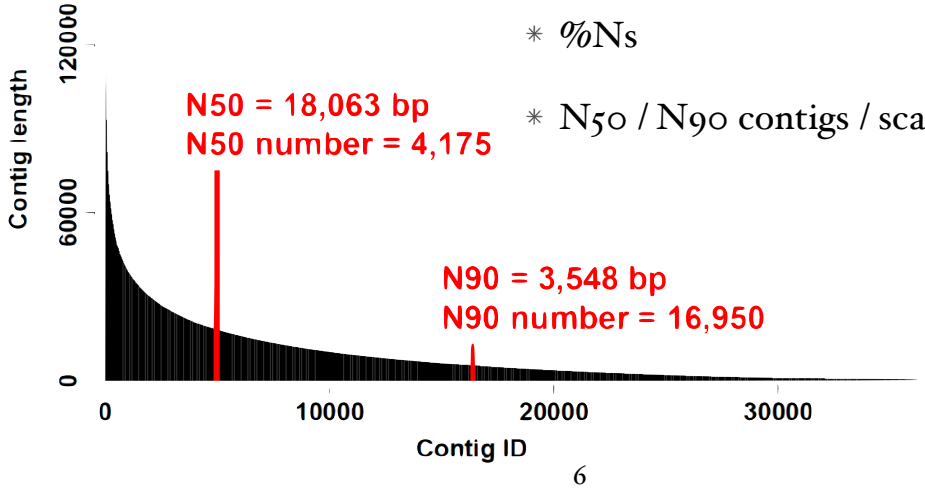
```
>contig_2001
AGCACCTAGAGCAGGATGGGAGGTCTCTCCTTGCTGTGGCAGAGGCAGATCTCCTTTCCC
AACACCTAGCAGTATGAACTAGTGAGCTCCTGACTGTTTTCCAGTGGTAATGAGGTGTGA
CCCGCTGCAGCTGCACACTGAATTCTCTCAGTTCCCCGAGGCCAGCCCAGCAGTGTGGGC
AATGCTTTGTTTGTGTGCTGTTGACCATTCC
>contig_2002
GTCTGCACTGGGAATGCCCCCTGGAGCAGAACCATTGCCATGGATAAGGACACTACATTT
CCTGGTGTTAAGGTGAATATAACCTCCAGGTTAAGGATGACATTAATTTCAATTACAGCT
TGCCTCTTGTAAGCTAAGCAGTTAATCAACAAGCTATACTGTGACTACACCCTTAGATCA
ATAGCTGGGAAAACATCACCTCCCCCAAATACTCCACCTCTTAACTGCACTCTTTGAAAG
AAGTACAGGCCAGAGTTTAGCTGATCCATCCCTGTGGCTAATCGTCCTGCTTACAAGCTG
CAATATTTTTTAAAACCAGACAATTGGTAGAGGTTTAAACATCAGCCAAGCTGTTCAATT
TACAGCAGGTTAAGCATTCCTGAAACTGTGATCACTGATATATTTGGGTCAGTCAGATGT
CTTGTTAGTGCTT
>contig_2003
ACAAACAAAACAAAATAAAACAAAGGAAACAAGCAAAAAAAACCATCATACAATCCCATG
TGTCCAAGAGCTTTACTGTGAAATCAACTATGGAGTCAAAACAATAGAAAAGCTTCCAGA
TTTCTGTATTCCAGGCTGAGACAAGTTTGTAAATACTTCCAGAAATTGCCAACAAGCCTG
CAGGGTAACATCTCTAATGCACACCTCCCTGATACGAAATGCAGAGCACCTTAACTTCTT
CAGCCCTCCCCCAGTCACAACCAGCTATAAATCCTGCCCTTCACTTGTTGGAATATCTCA
TCATAAGGGAAGCATTTTTTAGGCTGAGAAATACAAATCCACCTTGACGGAGCCGGTCAG
GCATATACATGGGCTATGCTGCTGATAGGTTTGTACCAAGCACTCCTAGTGTGAGAATAA
```

scaffold : FASTA-format

```
>scaffold_1
AGCACCTAGAGCAGGATGGGAGGTCTCTCCTTGCTGTGGCAGAGGCAGATCTCCTTTCCC
AACACCTAGCAGTATGAACTAGTGAGCTCCTGACTGTTTTCCAGTGGTAATGAGGTGTGA
CCCGCTGCAGCTGCACACTGAATTCTCTCAGTTCCCCGAGGCCAGCCCAGCAGTGTGGGC
AATGCTTTGTTTGTGTGCTGTTGACCATTCCNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
GTCTGCACTGGGAATGCCCCCTGGAGCAGAACCATTGCCATGGATAAGGACACTACATTT
CCTGGTGTTAAGGTGAATATAACCTCCAGGTTAAGGATGACATTAATTTCAATTACAGCT
TGCCTCTTGTAAGCTAAGCAGTTAATCAACAAGCTATACTGTGACTACACCCTTAGATCA
ATAGCTGGGAAAACATCACCTCCCCCAAATACTCCACCTCTTAACTGCACTCTTTGAAAG
AAGTACAGGCCAGAGTTTAGCTGATCCATCCCTGTGGCTAATCGTCCTGCTTACAAGCTG
CAATATTTTTTAAAACCAGACAATTGGTAGAGGTTTAAACATCAGCCAAGCTGTTCAATT
TACAGCAGGTTAAGCATTCCTGAAACTGTGATCACTGATATATTTGGGTCAGTCAGATGT
CTTGTTAGTGCTTNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
ACAAACAAAACAAAATAAAACAAAGGAAACAAGCAAAAAAAACCATCATACAATCCCATG
TGTCCAAGAGCTTTACTGTGAAATCAACTATGGAGTCAAAACAATAGAAAAGCTTCCAGA
TTTCTGTATTCCAGGCTGAGACAAGTTTGTAAATACTTCCAGAAATTGCCAACAAGCCTG
CAGGGTAACATCTCTAATGCACACCTCCCTGATACGAAATGCAGAGCACCTTAACTTCTT
CAGCCCTCCCCCAGTCACAACCAGCTATAAATCCTGCCCTTCACTTGTTGGAATATCTCA
TCATAAGGGAAGCATTTTTTAGGCTGAGAAATACAAATCCACCTTGACGGAGCCGGTCAG
GCATATACATGGGCTATGCTGCTGATAGGTTTGTACCAAGCACTCCTAGTGTGAGAATAA
```


Metrics for Genome Assembly

- Number of contigs / scaffolds
- Total size of contigs / scaffolds
- Longest contig / scaffold
- %Ns
- N50 / N90 contigs / scaffolds length



N50 - a measure of contiguity

N50 = contigs of this size or larger include 50 % of the assembly

>contig1		
TTTATGTCCGTAGCATGTAGACATATGGCAGCATG	35 bp	35
>contig2		
AGTCTTGAGCCGAATTCGTGGCATG	25 bp	35+25=60 (>50)
>contig3		
GTTGGAGCTATTCAGCGTAC	20 bp	N50 = 25 bp
>contig4		
ACAAATGATC	10 bp	
>contig5		
CGCTTCGAAC	10 bp	
	100 bp total	

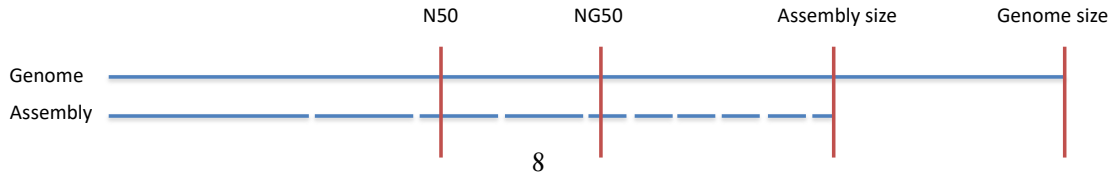
50% of total = 50

L50 = number of contigs that include 50% of the assembly. Here, L50 = 2

Other measurements : N90, L90

NG50 : compared with genome size rather than assembly size

- N50 is calculated in the context of the assembly size rather than genome size
- Comparisons of N50 values derived from assemblies of significantly different length are usually not informative, even if for the same genome
- The new measure called NG50 is defined by the authors of the Assemblathon competition
- NG50 is the same as N50 except that it is 50% of the known or estimated genome size
- More meaningful comparisons between different assemblies
- In the typical case that the assembly size is not more than genome size then NG50 will not be more than the N50
- LG50 is the number of contigs that include 50% of the genome



QUAST: quality assessment tool for genome assemblies

BIOINFORMATICS APPLICATIONS NOTE Vol. 29 no. 8 2013, pages 1072–1075
doi:10.1093/bioinformatics/btt086

Genome analysis Advance Access publication February 19, 2013

QUAST: quality assessment tool for genome assemblies

Alexey Gurevich[1,*], Vladislav Saveliev[1], Nikolay Vyahhi[1] and Glenn Tesler[2]

[1]Algorithmic Biology Laboratory, St. Petersburg Academic University, Russian Academy of Sciences, St. Petersburg 194021, Russia and [2]Department of Mathematics, University of California, San Diego, La Jolla, CA 92093-0112, USA

Associate Editor: Michael Brudno

ABSTRACT

Summary: Limitations of genome sequencing techniques have led to dozens of assembly algorithms, none of which is perfect. A number of methods for comparing assemblers have been developed, but none is yet a recognized benchmark. Further, most existing methods for comparing assemblies are only applicable to new assemblies of finished genomes; the problem of evaluating assemblies of previously unsequenced species has not been adequately considered. Here, we present QUAST—a quality assessment tool for evaluating and comparing genome assemblies. This tool improves on leading assembly comparison software with new ideas and quality metrics. QUAST can evaluate assemblies both with a reference genome, as well as without a reference. QUAST produces many reports, summary tables and plots to help scientists in their research and in their publications. In this study, we used QUAST to compare several genome assemblers on three datasets. QUAST tables and plots for all of them are available in the Supplementary Material, and interactive versions of these reports

popular sequencing strategies (including sequencing platforms and assembly software) for plant genomes. Plantagora has a well-designed interface to browse their database of evaluation results. Researchers may run the Plantagora assessment tool on their own assembly, but the results cannot be viewed through the friendly user-interface; instead, the user has to parse a large log file.

The Assemblathon competition (Earl *et al.*, 2011) compared 41 *de novo* assemblies on >100 evaluation metrics. The Assemblathon assessment scripts are freely available, but they are highly focused on the genomes used in the competition, and normal users cannot easily apply them to other genomes.

Another freely available genome assembly assessment tool is GAGE (Salzberg *et al.*, 2011). In Salzberg *et al.* (2011), it was used to evaluate several leading genome assemblers on four datasets. GAGE evaluates a set of metrics, including different types of misassembly errors (inversions, relocations and

QUAST

- **Metrics**
  - **1. Contig sizes**
    - *No. of contigs*
    - *Largest contig*
    - *Total length*
    - *Nx* : $(0 \leq x \leq 100)$ The largest contig length, L, such that using contigs of $length \geq L$ accounts for at least $x\,\%$ of the bases of the assembly
    - *NGx*, Genome *Nx*: The contig length such that using equal or longer length contigs produces $x\,\%$ of the length of genome

QUAST Metrics (cont)

* **2. Misassemblies and structural variations**
  * Evaluate them only with respect to a known reference genome
  * *No. of misassemblies*
  * *No. of misassembled contigs*
  * *Misassembled contigs length*
  * *No. of unaligned contigs*
  * *No. of ambiguously mapped contigs*

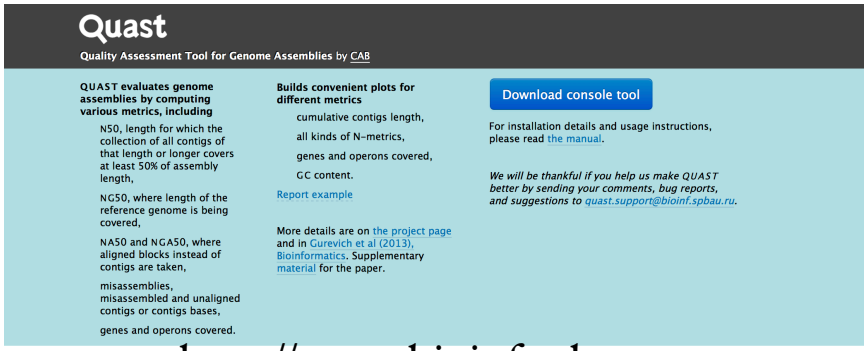
QUAST Metrics (cont)

* **3. Genome representation and its function elements**
  * Most of these require a reference genome
    * *Genome fraction (%) :* the total number of aligned bases in reference, divided by genome size
    * *Duplication ratio :* the total number of aligned bases in the assembly (total length - unaligned contigs length), divided by total number of aligned bases in the reference
    * *GC (%) :* total number of G and C in the assembly, divided by the total length of assembly (without reference genome)
    * *No. of mismatches per 100kb*
    * *No. of indels per 100kb*
    * *No. of genes* (complete an partial)
    * *No. of operons*
    * *No. of predicted genes*

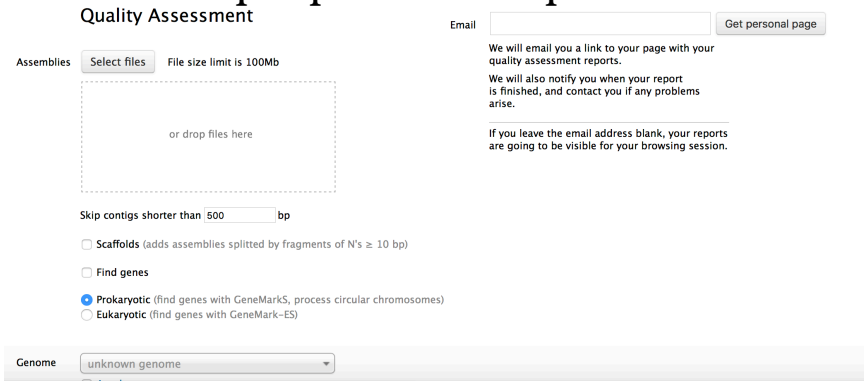
QUAST metrics (cont)

- **4. Variations of N50 based on aligned blocks**
  - Require a reference genome
    - *NAx (A stands for aligned, x range from 0 - 100)*
      - Break contigs into aligned blocks (if there are unaligned regions within a contig, these regions are removed , and the contig is split into blocks
      - Compute the Nx statistics based on these blocks instead of on the original contigs
    - *NGAx*
      - Break contigs into aligned blocks
      - Compute the NGx statistics on these blocks

QUAST : Web Interface



http://quast.bioinf.spbau.ru



QUAST Command line

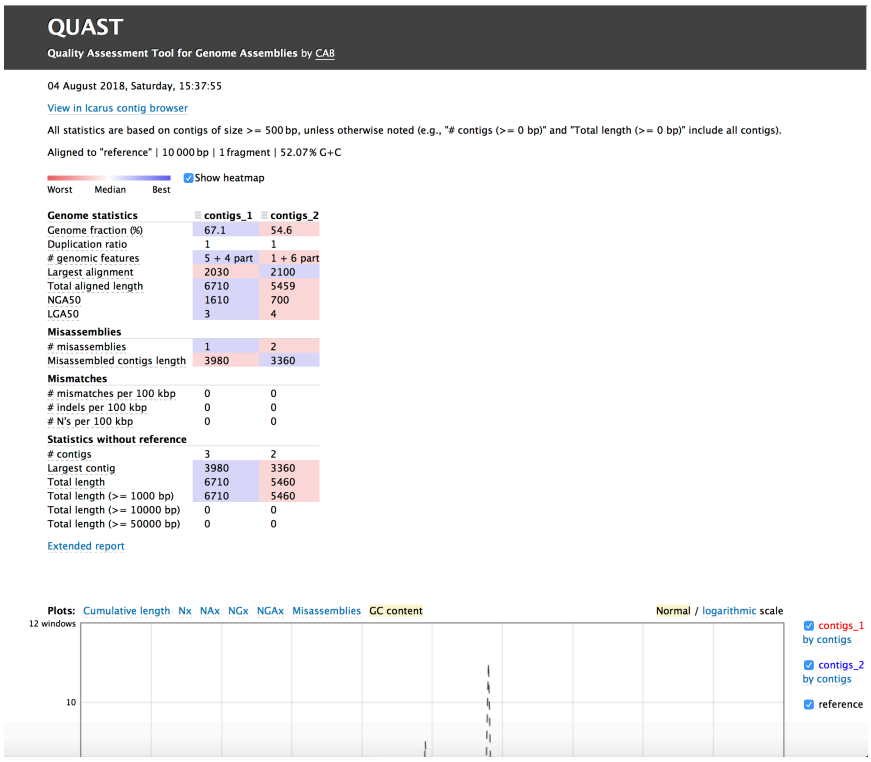
```
[user@agcipher ~]$ /share/apps/quast-5.0.0/quast.py test_data/contigs_1.fasta
test_data/contigs_2.fasta -R test_data/reference.fasta.gz -G test_data/
genes.gff -o Test_QUAST
WARNING: Option -G is deprecated! Please use --features to specify a file with
genomic features.
If you want QUAST to extract only a specific genomic feature from the file,
you should prepend the filepath with the feature name and a colon, for
example:
--features CDS:genes.gff --features transcript:transcripts.bed
Otherwise, all features would be counted:
--features genes.gff

/share/apps/quast-5.0.0/quast.py test_data/contigs_1.fasta test_data/
contigs_2.fasta -R test_data/reference.fasta.gz -G test_data/genes.gff -o
Test_QUAST

Version: 5.0.0

System information:
  OS: Linux-3.10.0-514.26.2.el7.x86_64-x86_64-with-centos-7.4.1708-Core
(linux_64)
  Python version: 3.5.5
  CPUs number: 144

Started: 2018-08-04 18:19:39
```


QUAST : Evaluation results

All statistics are based on contigs of size >= 500 bp, unless otherwise noted (e.g., "# contigs (>= 0 bp)" and "Total length (>= 0 bp)" include all contigs).

Assembly	contigs_1	contigs_2
# contigs (>= 0 bp)	3	4
# contigs (>= 1000 bp)	3	2
# contigs (>= 5000 bp)	0	0
# contigs (>= 10000 bp)	0	0
# contigs (>= 25000 bp)	0	0
# contigs (>= 50000 bp)	0	0
Total length (>= 0 bp)	6710	5870
Total length (>= 1000 bp)	6710	5460
Total length (>= 5000 bp)	0	0
Total length (>= 10000 bp)	0	0
Total length (>= 25000 bp)	0	0
Total length (>= 50000 bp)	0	0
# contigs	3	2
Largest contig	3980	3360
Total length	6710	5460
Reference length	10000	10000
GC (%)	51.28	52.44
Reference GC (%)	52.07	52.07
N50	3980	3360
NG50	1610	2100
N75	1610	2100
L50	1	1
LG50	2	2
L75	2	2
# misassemblies	1	2
# misassembled contigs	1	1
Misassembled contigs length	3980	3360
# local misassemblies	0	0
# scaffold gap ext. mis.	0	0
# scaffold gap loc. mis.	0	0
# unaligned mis. contigs	0	0
# unaligned contigs	0 + 0 part	0 + 0 part

QUASTT : report in html



QUASTT

Quality Assessment Tool for Genome Assemblies by CAB

04 August 2018, Saturday, 15:37:55

View in Icarus contig browser

All statistics are based on contigs of size >= 500 bp, unless otherwise noted (e.g., "# contigs (>= 0 bp)" and "Total length (>= 0 bp)" include all contigs).

Aligned to "reference" | 10 000 bp | 1 fragment | 52.07% G+C

Worst Median Best — Show heatmap

Genome statistics	contigs_1	contigs_2
Genome fraction (%)	67.1	54.6
Duplication ratio	1	1
# genomic features	5 + 4 part	1 + 6 part
Largest alignment	2030	2100
Total aligned length	6710	5459
NGA50	1610	700
LGA50	3	4
Misassemblies		
# misassemblies	1	2
Misassembled contigs length	3980	3360
Mismatches		
# mismatches per 100 kbp	0	0
# indels per 100 kbp	0	0
# N's per 100 kbp	0	0
Statistics without reference		
# contigs	3	2
Largest contig	3980	3360
Total length	6710	5460
Total length (>= 1000 bp)	6710	5460
Total length (>= 10000 bp)	0	0
Total length (>= 50000 bp)	0	0

Extended report

Plots: Cumulative length Nx NAx NGx NGAx Misassemblies GC content — Normal / logarithmic scale

12 windows

10

contigs_1 by contigs

contigs_2 by contigs

reference

QUAST

- **Comparing assemblers**
  - Comparison of assemblies of a single-cell sample of E.coli (for contigs >= 200 bp)

Assembler	No. of contigs	NGA50 (bp)	Largest (bp)	Total (bp)	Genome fraction (%)	No. of misassemblies	No. of complete genes
EULER-SR	610	26 580	140 518	4 306 898	86.54	19	3442
E+V-SC	396	32 051	132 865	4 555 721	93.58	**2**	3816
IDBA-UD	**283**	90 607	**224 018**	4 734 432	95.90	9	4030
SOAPdenovo	817	16 606	87 533	4 183 037	81.36	6	3060
SPAdes	532	**99 913**	211 020	4 975 641	**96.99**	11	**4071**
Velvet	310	22 648	132 865	3 517 182	75.53	**2**	3121
Velvet-SC	617	19 791	121 367	4 556 809	93.31	**2**	3662

The best value for each column is indicated in bold.

Q & A