Tables

Table 1: Exon Statistics for years >= 2016

Name	Total species	Exon number	Gene number	Gene Length	Exon per Gene
Bacteria	92287	N/A	$4.3k \pm 1.5k$	890 ± 64	N/A
Fungi	90	$32.3k \pm 1.8k$	$10k \pm 3.5k$	$1.6k \pm 171$	2.9 ± 1.3
Archaea	338	N/A	$2.9k \pm 0.9k$	851 ± 31	N/A
Viridiplantae	46	$385k \pm 155k$	$43k \pm 21k$	$4.1k \pm 1.3k$	9.2 ± 1.9
Metazoas	185	$462k \pm 280k$	$24.9k \pm 10.3k$	$23k \pm 11.8k$	17.7 ± 6.4
Ascomycota	70	$28.4k \pm 13.7k$	$10.4k \pm 3.1k$	$1.6k \pm 142$	2.5 ± 0.8
eudicotyledons(dicots)	37	$397k \pm 167k$	$45k \pm 22k$	$3.8k \pm 688$	9 ± 1.3

Table 2: Exon Statistics for years < 2016

Name	Total species	Exon number	Gene number	Gene Length	Exon per Gene
Bacteria	51537	N/A	$3.8k \pm 1.5k$	885 ± 65	N/A
Fungi	194	$29k \pm 20k$	$9.2k \pm 3.5k$	$1.6k \pm 254$	2.8 ± 1.5
Archaea	474	*N/A*	$2.9k \pm 0.8k$	855 ± 40	N/A
Viridiplantae	61	$273k \pm 153k$	$32k \pm 17k$	$4.1k \pm 2.3k$	8 ± 2.5
Metazoas	262	$314k \pm 211k$	$22.3k \pm 9.6k$	$22k \pm 12k$	13.4 ± 5.4
Ascomycota	143	$25.2k \pm 14.3k$	$9.5k \pm 3.1k$	$1.6k \pm 205$	2.4 ± 1
eudicotyledons(dicots)	41	$328k \pm 133k$	$38k \pm 16k$	$4k \pm 1.4k$	8.6 ± 1.3

Table 3: List of top three most used assembly programs for Metazoa (Year > =2016)

Kingdom	Program Name	species	Total length	Scaffold-count	ScaffoldN50	ContigCount	ContigN50
Metazoa	SOAPdenovo	21	1B ± 0.8B	$38k \pm 49k$	7.8M ± 11M	86k ± 66k	98k ± 208k
	AllPaths	48	0.9B ± 0.7B	7.1k ± 7k	4.3M ± 1.4M	33k ± 38k	188k ± 335k
	Newbler	7	0.8B±0.9B	3.3k ± 2.2k	877k ± 910k	56k±80k	75k±60k

Table 4: List of top three most used assembly programs for Metazoa (Year < 2016)

Kingdom	Program Name	species	Total length	Scaffold-count	ScaffoldN50	ContigCount	ContigN50
Metazoa	SOAPdenovo	98	$1.2B \pm 0.7B$	$40k \pm 38k$	$4.5M \pm 13M$	$116k \pm 79k$	$42k \pm 48k$
	AllPaths	54	$1.5B \pm 1.1B$	$11k \pm 13k$	$7.4M \pm 9.7M$	$119k \pm 97k$	$38k \pm 32k$
	Newbler	18	$0.9B \pm 0.9B$	$87k \pm 117k$	$2.1M \pm 2.3M$	$133k \pm 157k$	$34k \pm 27k$

Table 5: Kingdoms and average summary statistics for their genome assemblies (Years > =2016)

Tax ID	Name	Species	Total length	Scaffold-count	ScaffoldN50	ContigCount	ContigN50
2	Bacteria	92290	$4.3M \pm 1.6M$	66 ± 78	$0.9M \pm 1.4M$	132 ± 176	$0.39M \pm 0.86M$
4751	Fungi	90	$29M \pm 15M$	139 ± 159	$1.3M \pm 0.9M$	360 ± 688	$0.78M \pm 1M$
2157	Archaea	338	$2.9M \pm 0.98M$	52 ± 40	$0.38M \pm 0.43M$	74 ± 121	$0.53M \pm 71M$
33090	Viridiplantae	46	$0.97B \pm 0.88B$	$9.1k \pm 18.3k$	$31M \pm 49M$	$38k \pm 43k$	$1.8M \pm 4.9M$
33208	Metazoas	185	$1.2B \pm 0.95B$	$20.6k \pm 43.7k$	$22M \pm 36M$	$53k \pm 77k$	$2.5M \pm 7.9M$
71240	eudicotyledons(dicots)	37	$0.91B \pm 0.76B$	$6.4k \pm 10.6k$	$26M \pm 50M$	$40k \pm 44k$	$1.6M \pm 4.3M$

Table 6: Kingdoms and average summary statistics for their genome assemblies (Years <= 2015)

Tax ID	Name	Species	Total length	Scaffold Count	ScaffoldN50	ContigCount	ContigN50
2	Bacteria	51962	$3.8M \pm 1.6M$	45 ± 82	$1.3M \pm 1.5M$	126 ± 177	$0.27M \pm 0.55M$
4751	Fungi	202	$29M \pm 17M$	341 ± 699	$2M \pm 1.7M$	858 ± 1433	$0.55M \pm 0.75M$
2157	Archaea	470	$2.9M \pm 1M$	17 ± 16	$1.35M \pm 1.17M$	110 ± 126	$0.38M \pm 0.7M$
33090	Viridiplantae	67	$0.62B \pm 0.68B$	$22.9k \pm 46.6k$	$14.7M \pm 24.9M$	$52.5k \pm 71.6k$	$0.47M \pm 1.8M$
33208	Metazoas	295	$1.3B \pm 1B$	$37.4k \pm 64.2k$	$7.2M \pm 14M$	$118.6k \pm 119k$	$0.13M \pm 1.2M$
71240	eudicotyledons(dicots)	46	$0.754B \pm 0.750B$	$26.3k \pm 53.5k$	$17M \pm 27M$	$58.8k \pm 74k$	$0.3M \pm 1.6M$

Table 7: Comparison between MongoDB and BoaG

Feature	MongoDB	BoaG
Lines of Code	larger	smaller because it abstracts details of data analysis
Data generation time	longer due to the larger file	faster because of Binary file
Data file	JSON is 2.7 times larger than raw data	Hadoop Sequence file 5 times smaller than raw data
Schema Flexibility	Yes. Supports semi-structured data	Yes. Schema and compiler can be modified
MapReduce	Yes	Yes

Table 8: Domain types for Genomics data in BoaG

Type	Attributes	Details
Genome	taxid	Taxonomy ID of each species
	refseq	Refseq ID of the GFF file
	Sequence	List of sequence reads in each GFF file[29].
	AssemblerRoot	List of assembly programs associated with this genome
Sequence	accession	Accession number
	header	Header of Sequence
	FeatureRoot	List of features including exon,gene,mRNA, and CDS associated with this sequence
	seq	Actual DNA sequences from FASTA files
FeatureRoot	refseq	This field shows the key ID
	feature	This field is the list of features associated with this ID
Feature	accession	Accession code of the Sequence
	seqid	Sequence ID
	source	A text qualifier that describes the algorithm or procedure that generated this feature.
	ftype	Type of the feature
	start	starting point of the feature
	end	End point of the feature
	score	Score of the feature. This is a floating point number.
	strand	+ and - for positive and negative strand respectively
	phase	Phase of the feature. The phase is one of the integers 0, 1, or 2
	Attribute	List of attributes for each feature
	parent	Shows the parent of the attribute
Attribute	id	Attribute ID
	tag	Attribute tag including gbkey etc.
	value	Value of the tag
AssemblerRoot	Assembler	List of assembly programs
	total-length	Total length or genome size (base pair)
	total-gap-length	Total gap length after genome assembly
	scaffold-N50	Scaffold N50 metric
	scaffold-count	Scaffold count metric
	contig-N50	Contig N50 metric
	contig-count	Contig count metric
Assembler	name	Assembly program used to assemble the genome
	desc	Program attributes: program name, program version, etc.

Table 9: The BoaG aggregators list

Aggregator	Description
MeanAggreagtor	Calculates the average
MaxAggreagtor	Finds the maximum value
SumAggregator	Calculates the sum of the emitted values to the reducer
MinAggregator	Finds the minimum value
TopAggregator	Takes an integer argument and returns the top elements for the given argument
StDevAggregator	Calculates the standard deviation