ANÁLISIS DE ENSAMBLAJES

Laura González, MSc

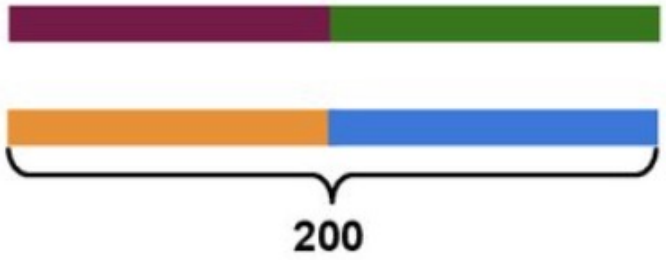
Romain Guyot, PhD

Universidad de los Andes

SGR
Sistema General de Regalías

EN LA VIDA REAL …

Assembly A



N50 = 200
# misassemblies = 2

Assembly B

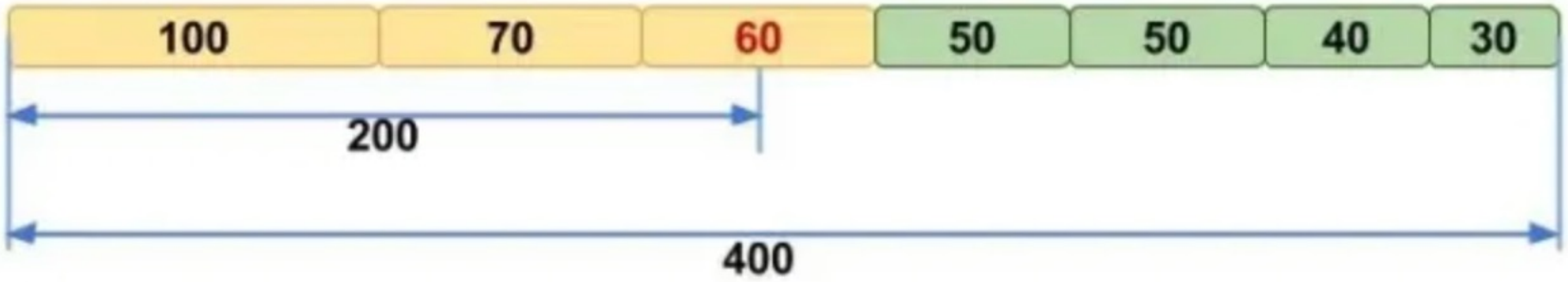
N50 = 100
# misassemblies = 0

CALIDAD DE ENSAMBLAJE

Tamaño de genoma: 400
contigs: 7
Contig más largo: 100

https://www.molecularecologist.com/2017/03/29/whats-n50/

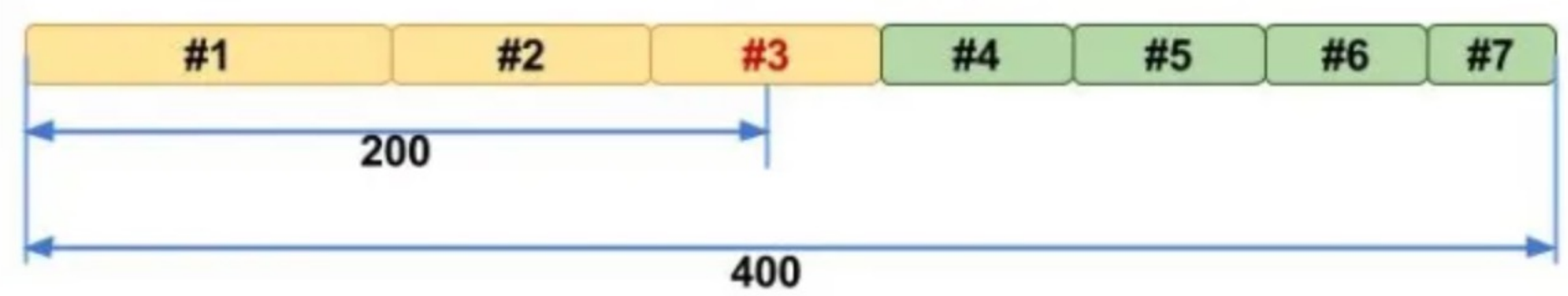
N50 – NG50



Tamaño del contig que suma la mitad del tamaño del ensamblaje o del genoma de referencia

https://www.molecularecologist.com/2017/03/29/whats-n50/

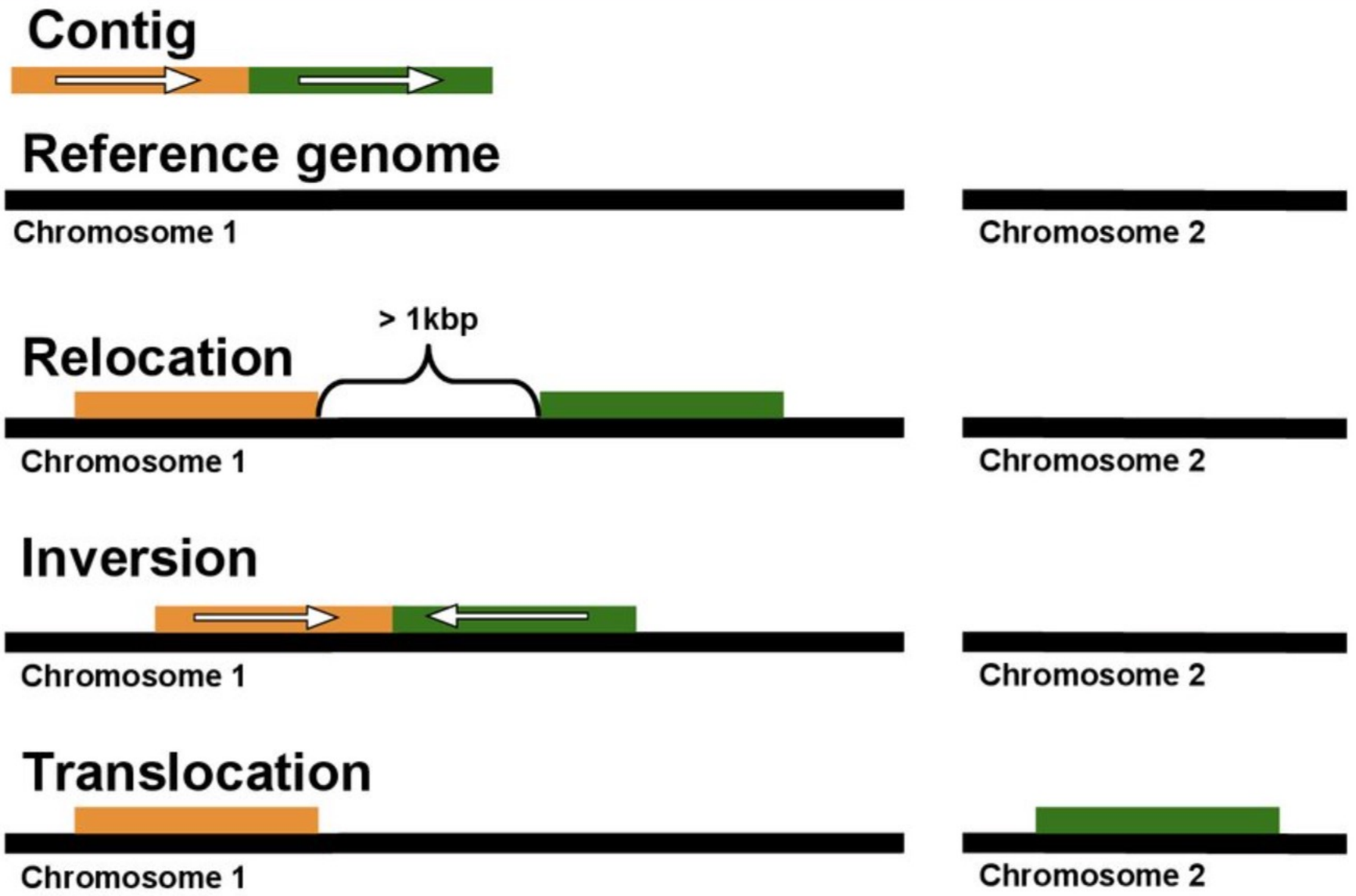
L50 – LG50



Número de contigs que suman la mitad del tamaño del ensamblaje o del genoma de referencia

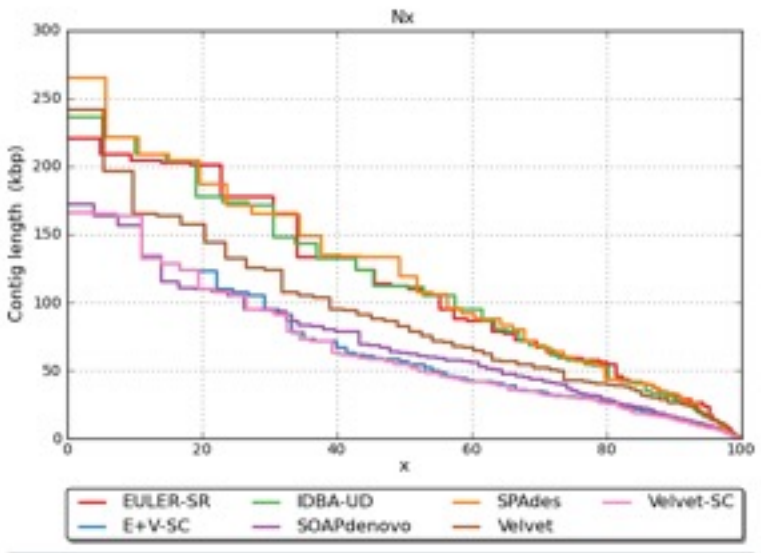
https://www.molecularecologist.com/2017/03/29/whats-n50/

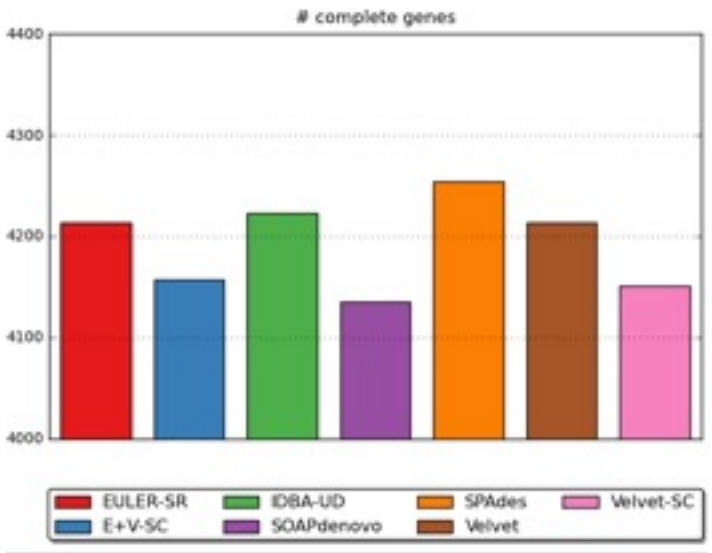
MISASSEMBLIES

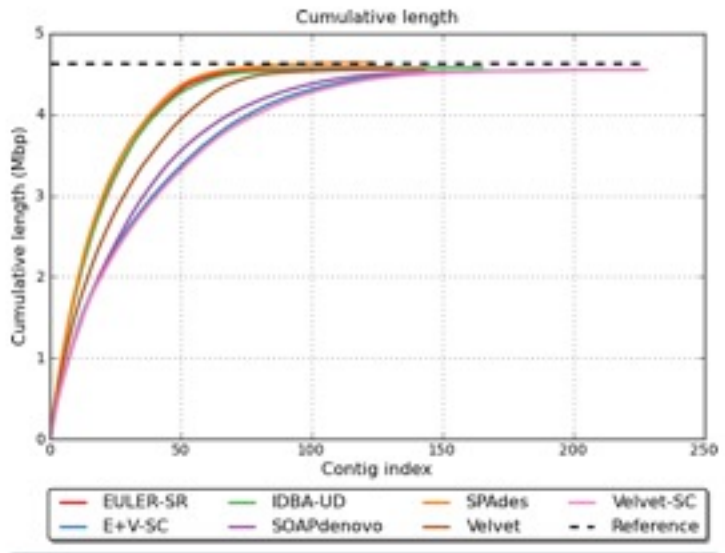


QUAST

Samples of QUAST plots:







https://en.ppt-online.org/686477

QUAST

quast.bioinf.spbau.ru

MetaQUAST report for assemblies of the MH0045 sample from MetaHIT (Qin et al., 2010)

Download report as metahit.tar.gz

09 November 2015, Monday, 20:03:39

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.)

Rows show values for the whole assembly (column name) vs. combined reference (concatenation of input references).
Clicking on a row with + sign will expand values for contigs aligned to each of input references separately.
Note that some metrics (e.g. # contigs) may not sum up, because one contig may be aligned to several references and thus, counted several times.
All metrics that depend on the reference length (such as NG50, LG50, etc), plus the GC % are not calculated for the combined reference.
The combined reference is just a concatenation of all available reference genomes of the species, presumably represented in the metagenomic dataset, but not necessarily the real content.
So it might miss many correctly assembled species, and therefore it doesn't make sense to apply the size and the GC content of the combined reference for assembly evaluation.

Reference size: 306 971 432 bp

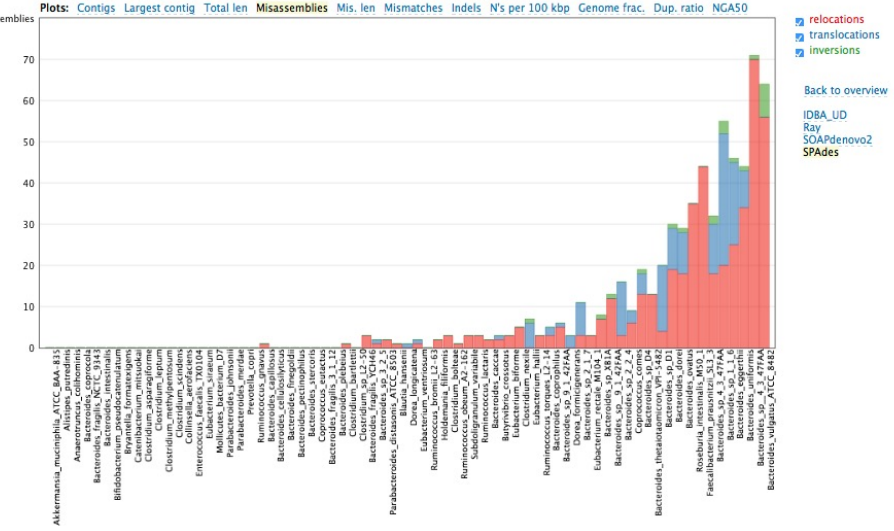
Reference	Size, bp	GC, %
Akkermansia_muciniphila_ATCC_BAA-835	2 664 102	55.76
Alistipes_putredinis	2 550 678	53.27
Anaerotruncus_colihominis	3 719 688	54.18
Bacteroides_caccae	5 493 117	42.83
Bacteroides_capillosus	4 241 076	59.11
Bacteroides_cellulosilyticus	7 694 202	43.05
Bacteroides_coprocola	2784	45.19
Bacteroides_coprophilus	4 041 504	45.72
Bacteroides_dorei	6 060 928	42.2
Bacteroides_eggerthii	4 611 535	44.71
Bacteroides_finegoldii	5 124 109	42.5
Bacteroides_fragilis_3_1_12	5 530 115	43.62
Bacteroides_fragilis_NCTC_9343	5 205 140	43.19
Bacteroides_fragilis_YCH46	5 277 274	43.27
Bacteroides_intestinalis	4 605 106	43.54
Bacteroides_ovatus	7 010 996	42.3
Bacteroides_pectinophilus	29 332	36.96
Bacteroides_plebeius	4 421 924	44.31
Bacteroides_sp_1_1_6	6 760 735	43.02
Bacteroides_sp_2_1_7	5 180 144	45.08
Bacteroides_sp_2_2_4	7 101 224	42.13
Bacteroides_sp_3_2_5	5 116 282	43.17
Bacteroides_sp_4_3_47FAA	5 442 925	42.7
Bacteroides_sp_9_1_42FAA	5 622 644	42.33
Bacteroides_sp_D1	5 974 559	41.88
Bacteroides_sp_D4	5 538 248	41.75
Bacteroides_sp_X81A	5 976 145	41.89
Bacteroides_sp__4_3_47FAA	5 442 925	42.7
Bacteroides_sp__9_1_42FAA	4 684 745	42.2
Bacteroides_stercoris	4 102 660	45.93
Bacteroides_thetaiotaomicron_VPI-5482	6 260 361	42.84
Bacteroides_uniformis	4 835 507	46.49
Bacteroides_vulgatus_ATCC_8482	5 163 189	42.2
Bifidobacterium_pseudocatenulatum	2 313 752	56.38
Blautia_hansenii	3 058 721	38.99
Bryantella_formatexigens	4 548 960	49.55
Butyrivibrio_crossotus	2 496 039	37.75
Catenibacterium_mitsuokai	2 671 313	36.82
Clostridium_asparagiforme	6 417 332	55.6
Clostridium_bartlettii	2 972 256	28.84
Clostridium_bolteae	6 538 460	49.39
Clostridium_leptum	3 270 209	50.19
Clostridium_methylpentosum	3 478 423	51.82
Clostridium_nexile	3 995 628	40.090
Clostridium_scindens	1 631 609	46.03
Clostridium_sp_L2-50	2 954 616	41.37
Collinsella_aerofaciens	2 439 869	60.55
Coprococcus_comes	3 242 215	42.49
Coprococcus_eutactus	3 102 987	43.09
Dorea_formicigenerans	3 843 583	40.340
Dorea_longicatena	2 915 433	41.44
Enterococcus_faecalis_TX0104	3 156 478	37.270
Eubacterium_biforme	2 517 763	33.79
Eubacterium_hallii	3 290 996	38.19
Eubacterium_rectale_M104_1	3 698 419	40.550
Eubacterium_siraeum	2 664 035	44.97
Eubacterium_ventriosum	2 870 795	34.92
Faecalibacterium_prausnitzii_SL3_3	3 214 418	55.65
Holdemania_filiformis	3 932 923	50.18
Mollicutes_bacterium_D7	3 561 737	31.37
Parabacteroides_distasonis_ATCC_8503	4 811 379	45.06
Parabacteroides_johnsonii	4 629 061	45.13
Parabacteroides_merdae	4 458 741	45.25
Prevotella_copri	3 512 473	44.85
Roseburia_intestinalis_M50_1	4 143 550	42.41
Ruminococcus_bromii_L2-63	2 249 085	41.39
Ruminococcus_gnavus	3 501 911	42.88

Worst Median Best — Show heatmap

Statistics without reference	IDBA_UD	Ray	SOAPdenovo2	SPAdes
# contigs	31 224	10 327	36 468	40 546
# contigs (>= 0 bp)	46 096	195 402	208 740	92 463
# contigs (>= 1000 bp)	15 638	5490	13 068	19 235
# contigs (>= 5000 bp)	2243	1273	1037	2950
# contigs (>= 10000 bp)	1079	630	196	1255
# contigs (>= 25000 bp)	452	161	7	426
# contigs (>= 50000 bp)	182	36	0	146
Largest contig	305 144	99 107	40 707	189 063
Total length	80 325 286	30 411 921	46 741 224	92 397 329
Total length (>= 0 bp)	85 398 219	59 853 665	82 244 277	106 967 180
Total length (>= 1000 bp)	69 223 529	27 080 646	30 720 336	77 823 828
Total length (>= 5000 bp)	42 843 090	18 289 015	8 400 340	44 989 853
Total length (>= 10000 bp)	34 930 908	13 755 677	2 800 864	33 477 263
Total length (>= 25000 bp)	25 310 756	6 553 349	223 453	20 919 132
Total length (>= 50000 bp)	16 008 349	2 346 322	0	11 409 912
N50	6111	8131	1525	4692
N75	1696	2279	814	1525
L50	1756	790	7020	3200
L75	8674	2651	17 886	12 282
GC (%)	...	...	...	...
Misassemblies				
# misassemblies	1132	407	831	1240
# relocations	306	92	70	251
# translocations	84	32	22	90
# inversions	35	2	6	15
# interspecies translocations	707	281	733	884
# possibly misassembled contigs	1031	416	352	1623
# misassembled contigs	857	299	683	936
Misassembled contigs length	10 448 260	4 115 772	911 826	10 780 557
# local misassemblies	313	1217	10 977	287
# structural variations	108	50	56	99
# fully unaligned contigs	20 053	6857	21 087	28 661
Fully unaligned length	38 837 027	15 942 170	28 626 572	52 077 982
# partially unaligned contigs	3031	624	2213	4358
# with misassembly	284	87	395	294
# both parts are significant	742	203	250	1215
Partially unaligned length	11 712 726	2 343 203	1 502 717	10 309 003
Mismatches				
# mismatches	355 477	141 995	219 630	498 523
# indels	12 524	3729	4227	18 364
Indels length	35 483	9447	9641	63 455
# mismatches per 100 kbp	904.95	1054.68	888.21	1401.84
# indels per 100 kbp	31.88	27.7	17.09	51.64
# short indels	11 033	3383	3899	15 225
# long indels	1491	346	328	3139
# N's	191 558	634 779	1 743 687	1 316 794
# N's per 100 kbp	238.48	2087.27	3730.51	1425.14
Genome statistics				
Genome fraction (%)	12.796	4.386	8.055	11.585
Duplication ratio	1.044	1.094	1.039	1.046
Largest alignment	179 515	72 570	21 339	108 559
NGA50	...	...	...	...
NG50	...	...	...	...
NG75	...	...	...	...
LG50	...	...	...	...
LG75	...	...	...	...
LGA50	...	...	...	...
NGA75	...	...	...	...
LGA75	...	...	...	...

Short report

Plots: Contigs Largest contig Total len Misassemblies Mis. len Mismatches Indels N's per 100 kbp Genome frac. Dup. ratio NGA50



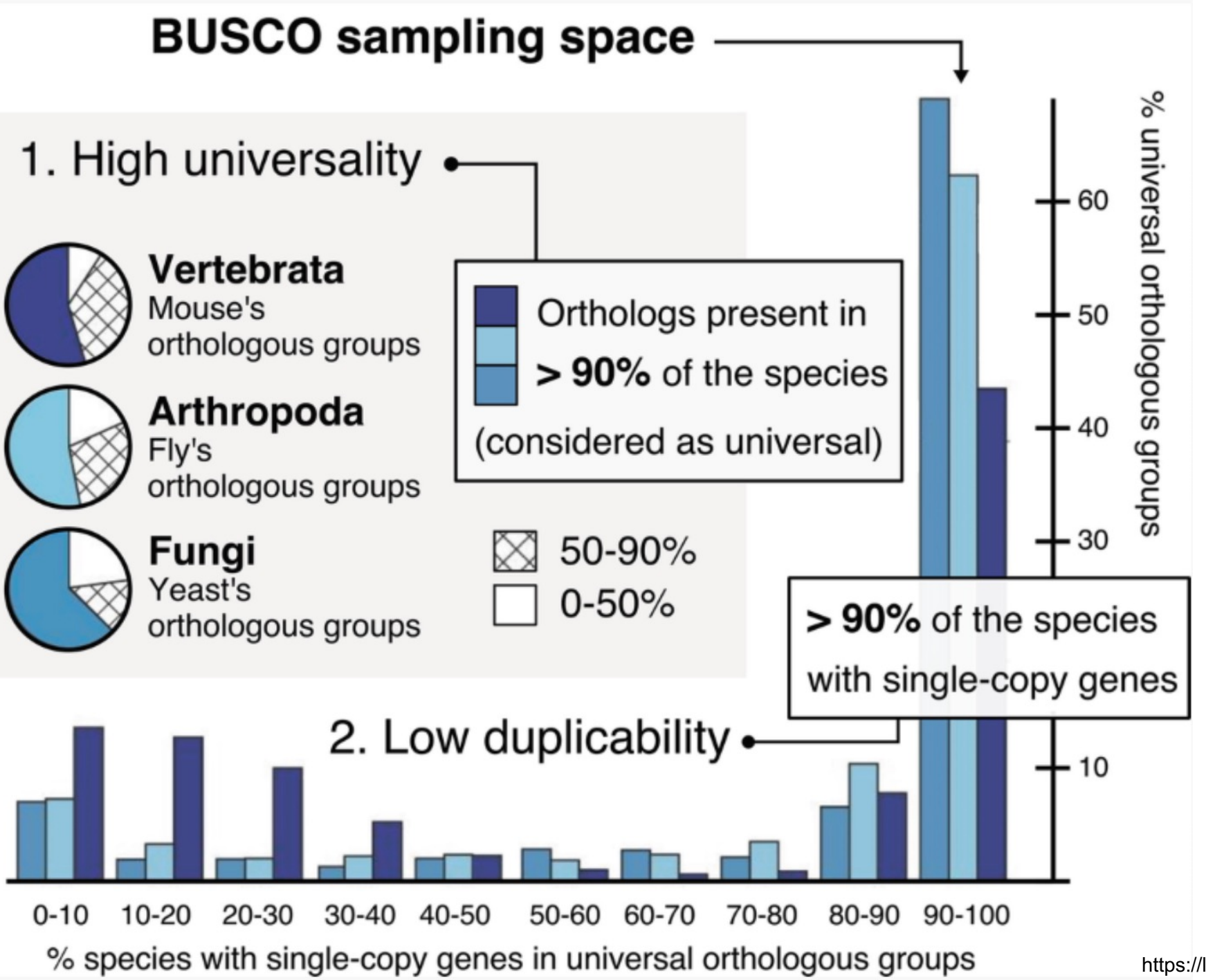
Back to overview

IDBA_UD
Ray
SOAPdenovo2
SPAdes

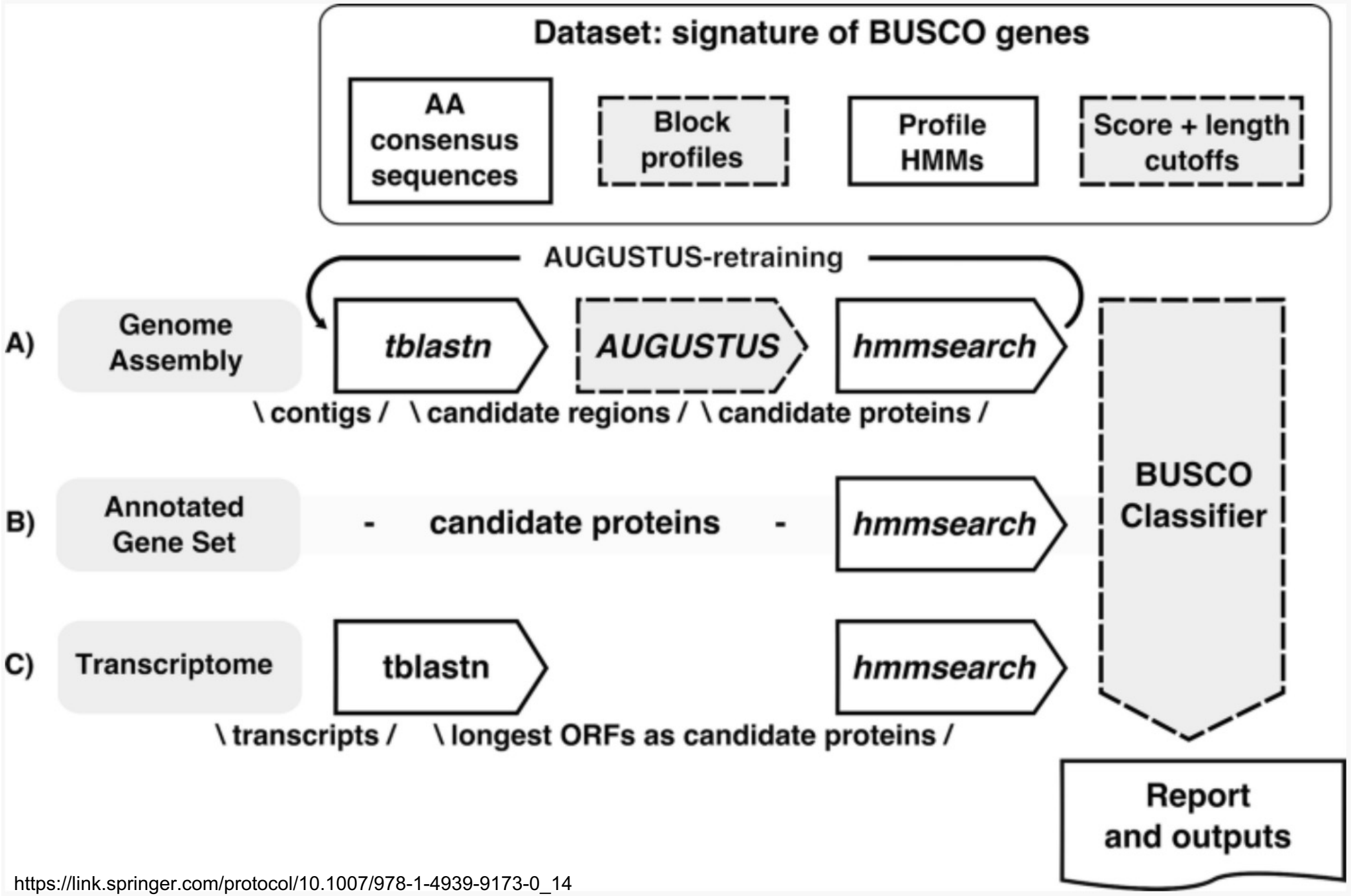
References are ordered starting from the best average value among all assemblies.

77

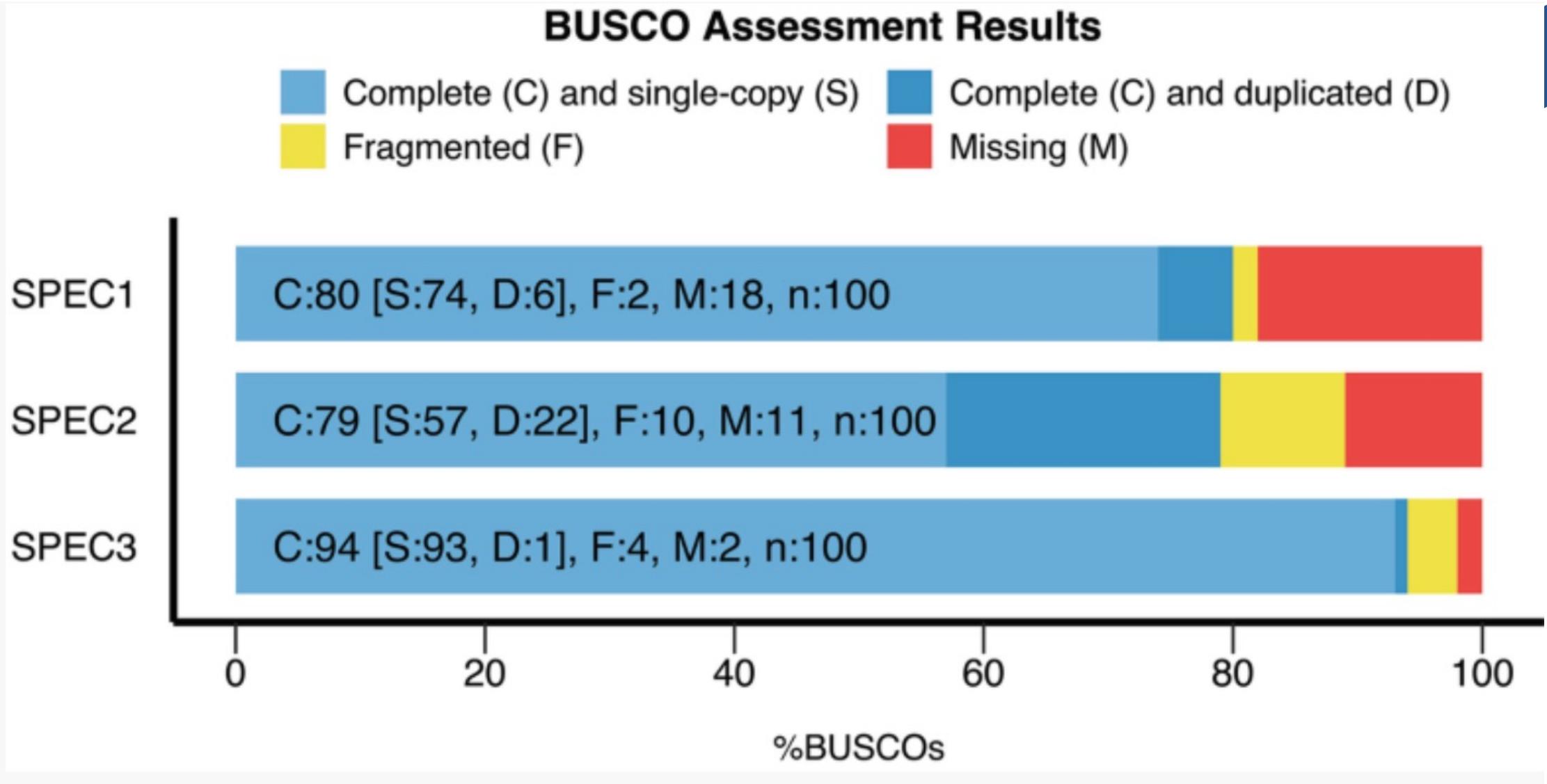
COMPLETITUD



https://link.springer.com/protocol/10.1007/978-1-4939-9173-0_14

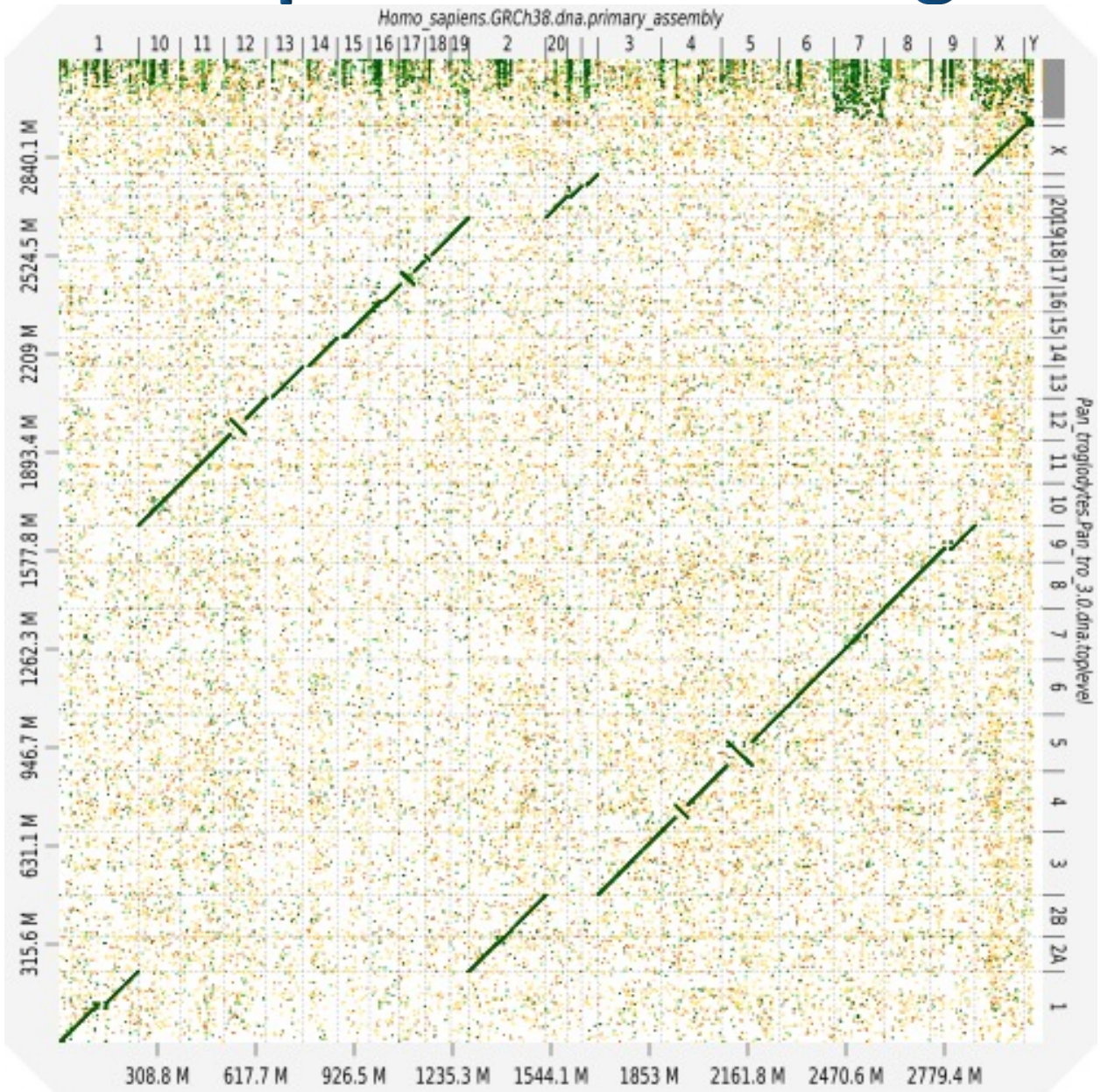


https://link.springer.com/protocol/10.1007/978-1-4939-9173-0_14



https://link.springer.com/protocol/10.1007/978-1-4939-9173-0_14

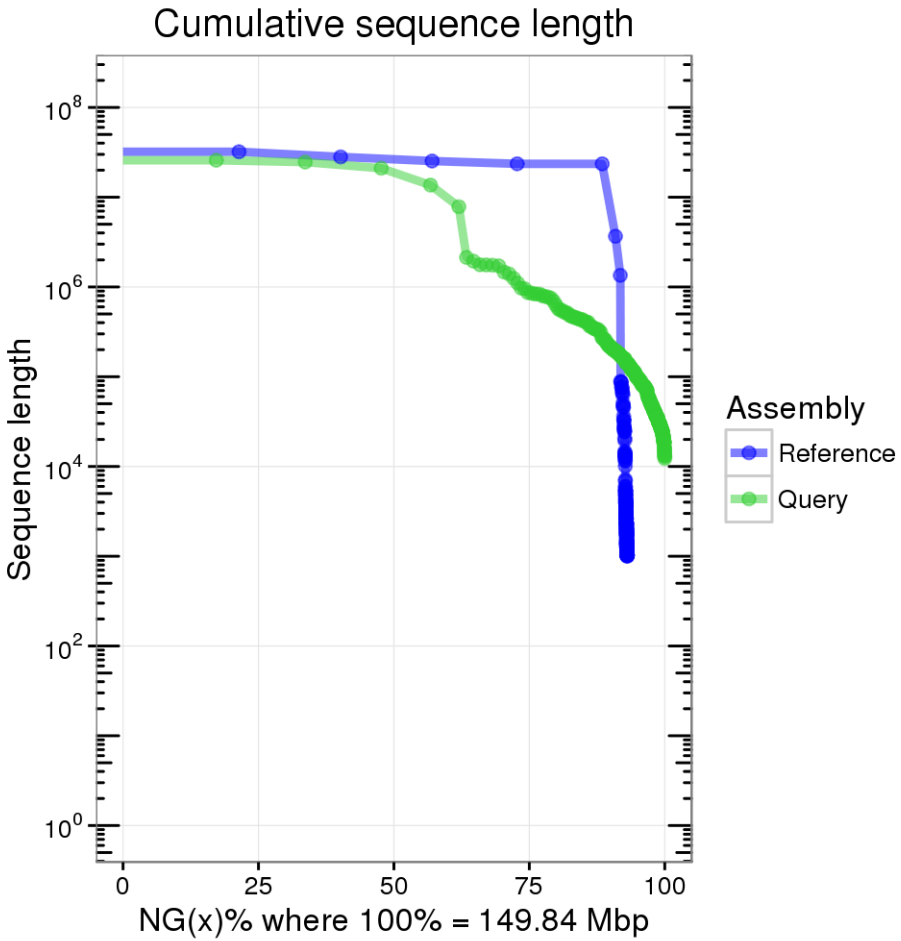
Comparación con un genoma de referencia



Gepard
Mummer
D-genie

https://dgenies.toulouse.inra.fr/result/chimp_vs_human

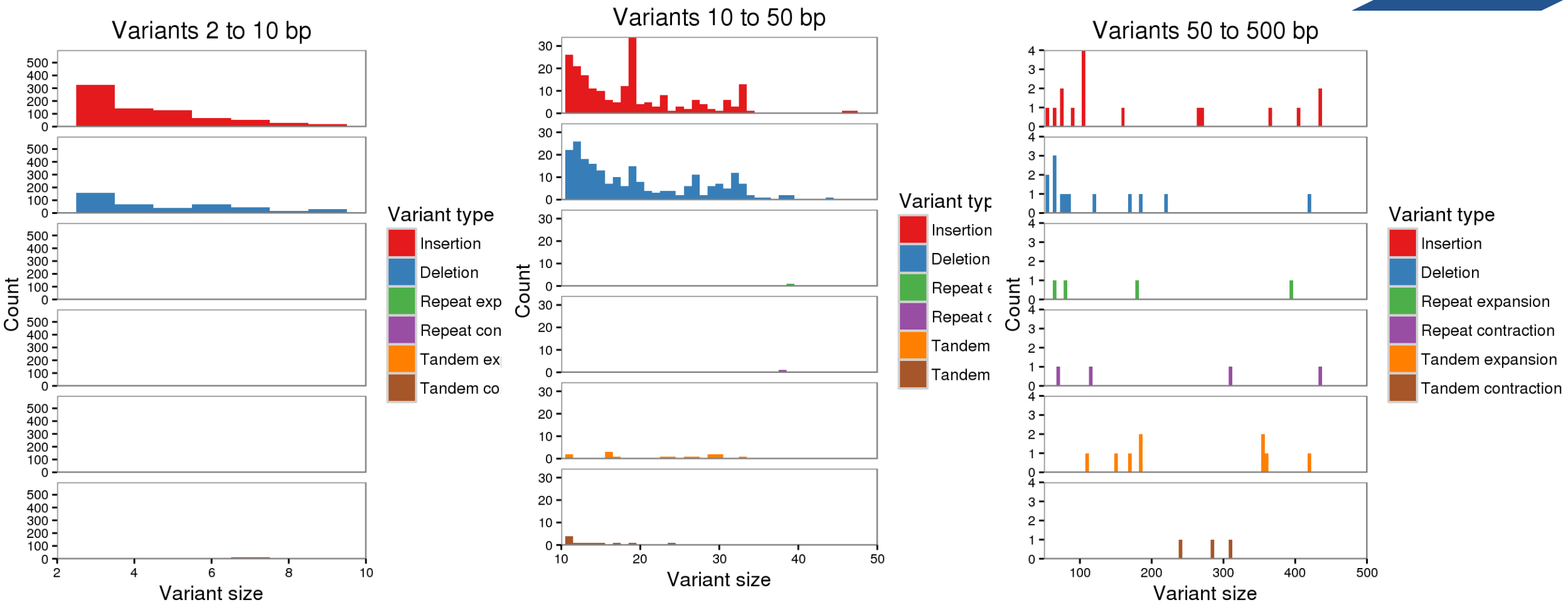
Comparación con un genoma de referencia



Assemblytics

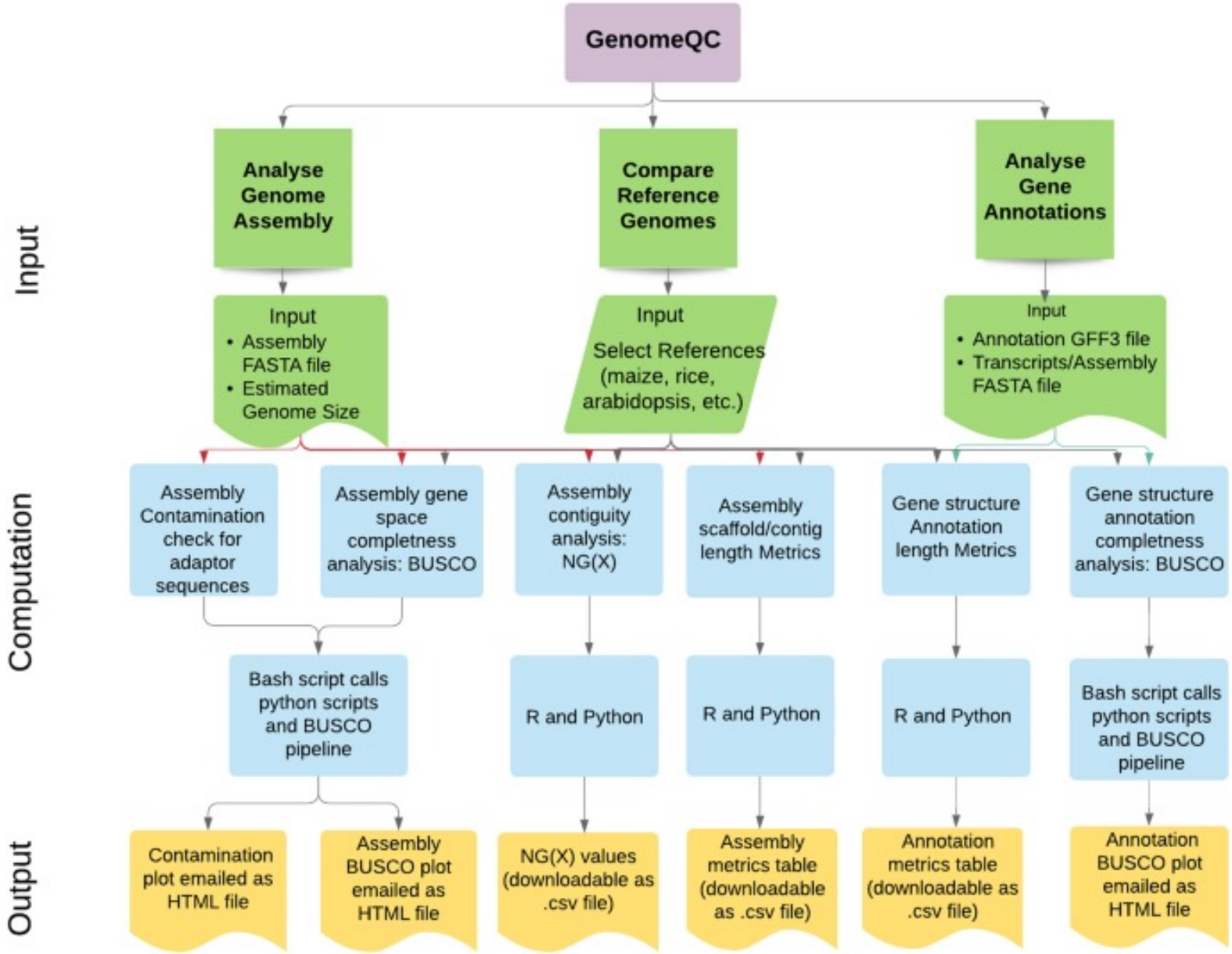
http://assemblytics.com

Comparación con un genoma de referencia

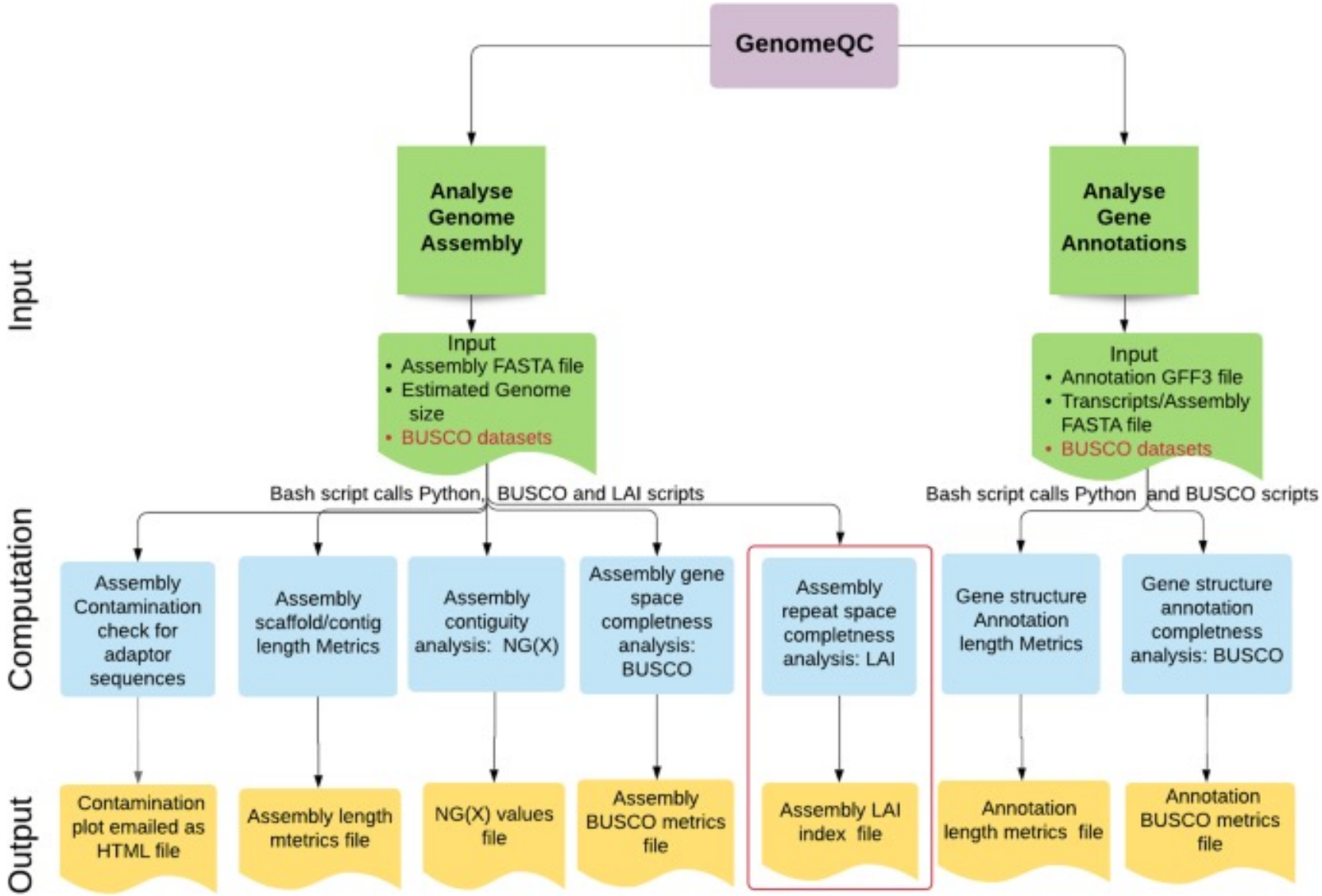


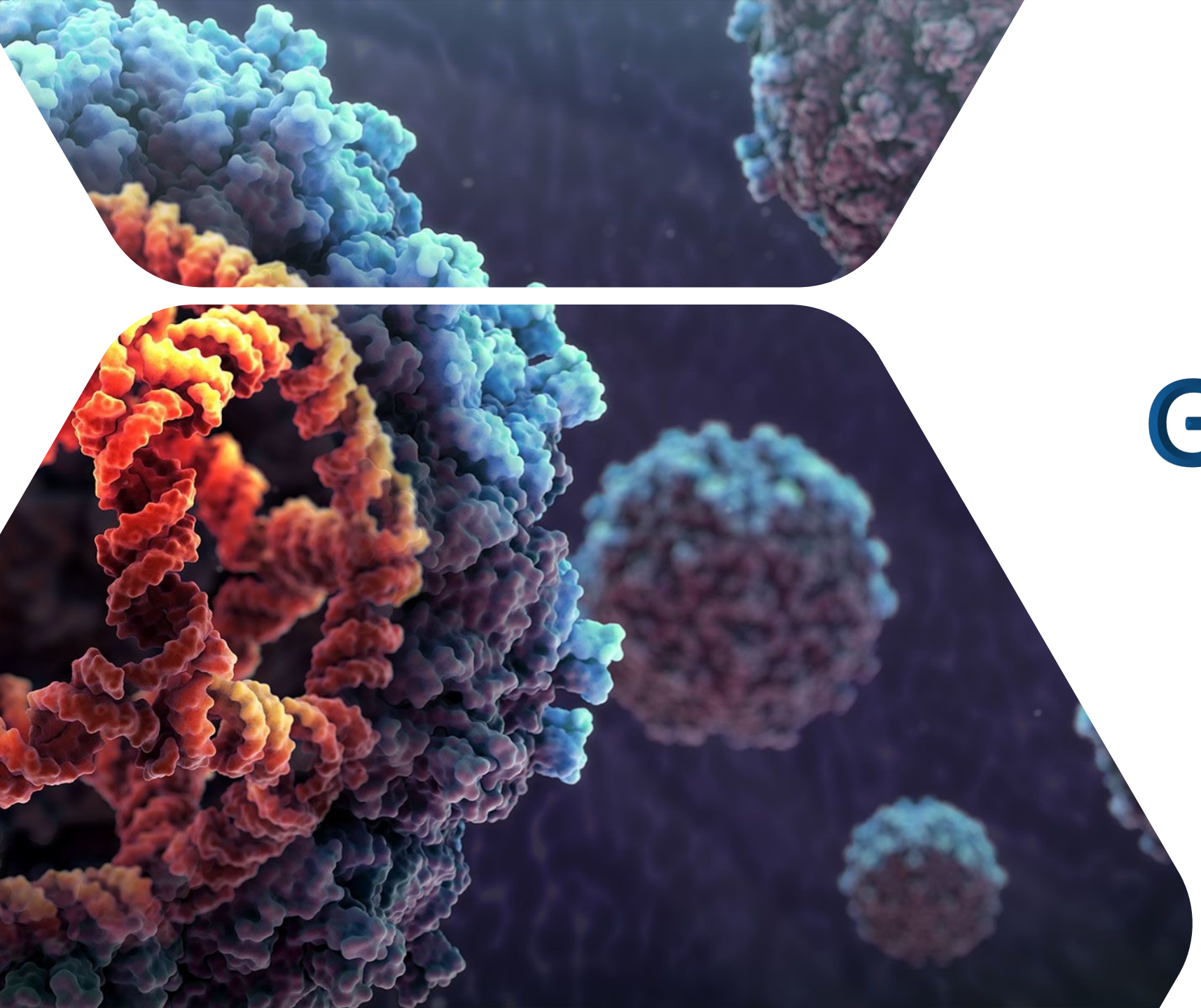
http://assemblytics.com

GenomeQC


Input
Computation
Output
GenomeQC
Analyse Genome Assembly
Compare Reference Genomes
Analyse Gene Annotations
Input
• Assembly FASTA file
• Estimated Genome Size
Input
Select References (maize, rice, arabidopsis, etc.)
Input
• Annotation GFF3 file
• Transcripts/Assembly FASTA file
Assembly Contamination check for adaptor sequences
Assembly gene space completness analysis: BUSCO
Assembly contiguity analysis: NG(X)
Assembly scaffold/contig length Metrics
Gene structure Annotation length Metrics
Gene structure annotation completness analysis: BUSCO
Bash script calls python scripts and BUSCO pipeline
R and Python
R and Python
R and Python
Bash script calls python scripts and BUSCO pipeline
Contamination plot emailed as HTML file
Assembly BUSCO plot emailed as HTML file
NG(X) values (downloadable as .csv file)
Assembly metrics table (downloadable as .csv file)
Annotation metrics table (downloadable as .csv file)
Annotation BUSCO plot emailed as HTML file

GenomeQC


GenomeQC
Analyse Genome Assembly
Analyse Gene Annotations
Input
Computation
Output
Input
• Assembly FASTA file
• Estimated Genome size
• BUSCO datasets
Input
• Annotation GFF3 file
• Transcripts/Assembly FASTA file
• BUSCO datasets
Bash script calls Python, BUSCO and LAI scripts
Bash script calls Python and BUSCO scripts
Assembly Contamination check for adaptor sequences
Assembly scaffold/contig length Metrics
Assembly contiguity analysis: NG(X)
Assembly gene space completness analysis: BUSCO
Assembly repeat space completness analysis: LAI
Gene structure Annotation length Metrics
Gene structure annotation completness analysis: BUSCO
Contamination plot emailed as HTML file
Assembly length mtetrics file
NG(X) values file
Assembly BUSCO metrics file
Assembly LAI index file
Annotation length metrics file
Annotation BUSCO metrics file



Universidad de los Andes

GRACIAS

SGR
Sistema General de Regalías