NOVOPlasty: De novo assembly of organelle genomes from whole genome data

Nicolas Dierckxsens[1]*, Patrick Mardulyn[1,2] and Guillaume Smits[1,2,3]

[1]Interuniversity Institute of Bioinformatics in Brussels (IB2), Université Libre de Bruxelles and Vrije Universiteit Brussel, Triomaan CP 263, 1050 Brussels, Belgium,
[2]Evolutionary Biology and Ecology Unit, CP 160/12, Faculté des Sciences, Université Libre de Bruxelles, Av. F. D. Roosevelt 50, B-1050 Brussels, Belgium, [3]Genetics, Hôpital Universitaire des Enfants Reine Fabiola, Université Libre de Bruxelles, Brussels, Belgium and [4]Center for Medical Genetics, Hôpital Erasme, Université Libre de Bruxelles, Route de Lennik 808, 1070 Brussels, Belgium
*To whom correspondence should be addressed. Tel: +32 0472 986806; Email: nicolasdierckxsens@hotmail.com

Reference validation

An accurate reference genome is essential to validate the quality of the assemblies. For the organelle genomes of *Oryza sativa*, *Arabidopsis thaliana* and *Homo sapiens* there are reference genomes available in GenBank (1). Nevertheless these reference genomes will not always be perfect because of small variances between individuals. Therefore each reference genome had to be validated manually by aligning (using bowtie2 (2)) all the reads of each dataset to the respective reference genome. Each variance was visually detected using Tablet (3) and corrected in the GenBank reference.

Homo sapiens

No variance was found between GenBank entry X93334.1 and our dataset.

Arabidopsis thaliana

The same GenBank entry AP000423.1 was used for both datasets SRR1174256 and SRR1810277. Two deletions of 1 bp were detected for SRR1174256 and one insertion of 1 bp for SRR1810277. All three were variations in the length of Single Nucleotide Repeats (SNR). The visual detection of the insertion is shown in Figure S1.

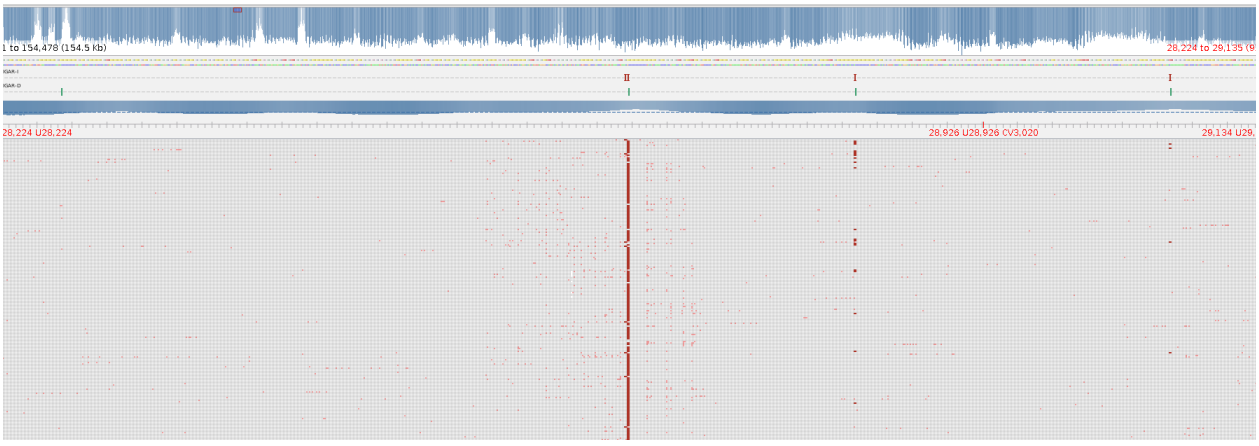

Figure S1 | Visual detection of a single nucleotide insertion in a T homopolymer between dataset SRR1810277 and GenBank entry AP000423.1

Oryza sativa

No variance was found between GenBank entry KM103369.1 and dataset SRR1328237. Between GenBank entry KM088022.1 and dataset ERR477442, there was a difference of 19 nucleotides, consisting out of 10 Single-Nucleotide Polymorphisms (SNP's), 3 single bp insertions, a 4 bp long deletion and 2 single bp deletions. Figure S2 show an example of a SNP and Figure S3 shows the 4 bp long deletion.

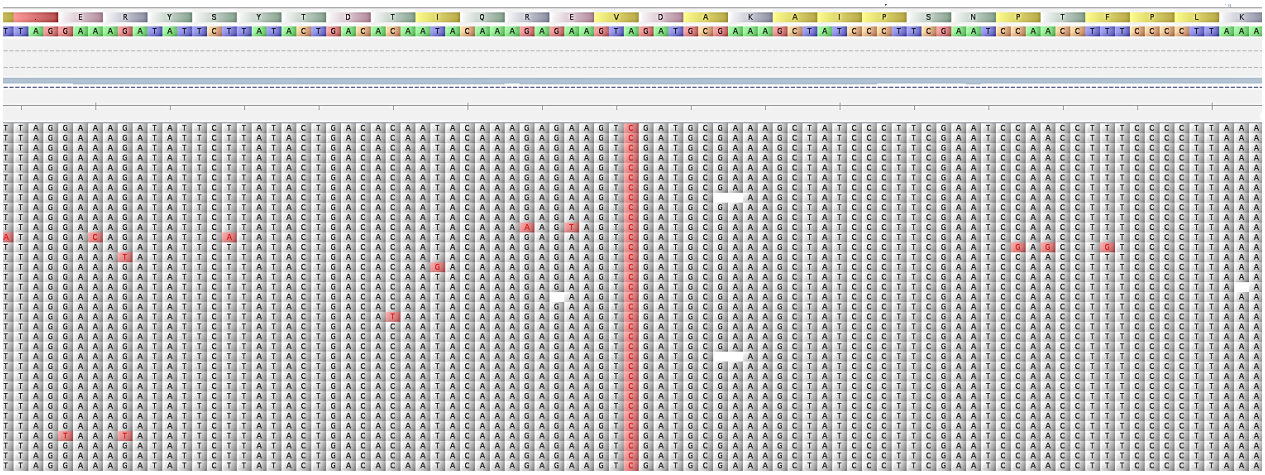

Figure S2 | Visual detection of a SNP (A to C) between dataset ERR477442 and GenBank entry KM088022.1

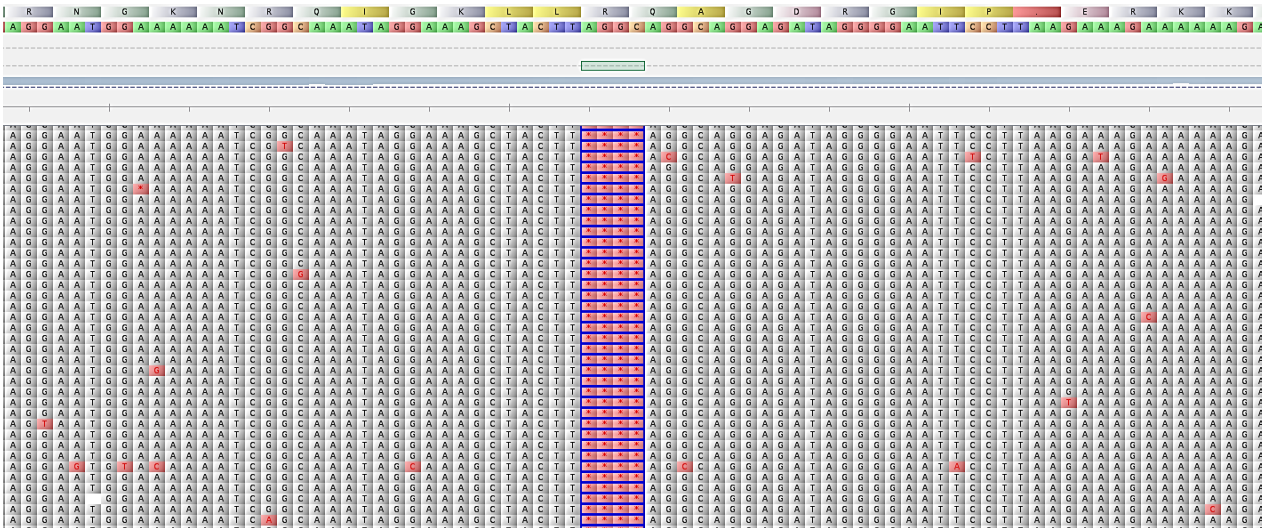

Figure S3 | Visual detection of a 4 bp deletion (AGGC) between dataset ERR477442 and GenBank entry KM088022.1

Avicennia marina assembly comparison

There is no reference available for *Avicennia marina*, therefore we visually compared the differences between the CLC and the NOVOPlasty assemblies. This was done by realigning (with bowtie2) all the reads from the dataset to both of the assembly results and visually inspect the alignment with Tablet. The CLC assembly contained seven SNP's and was missing 1043 bp in comparison to the NOVOPlasty assembly. All missing regions and the 7 SNP's were confirmed by inspecting the alignment.

Figure S4 | Visual confirmation of 4 SNP's between the CLC assembly and the *Avicennia marina* dataset.

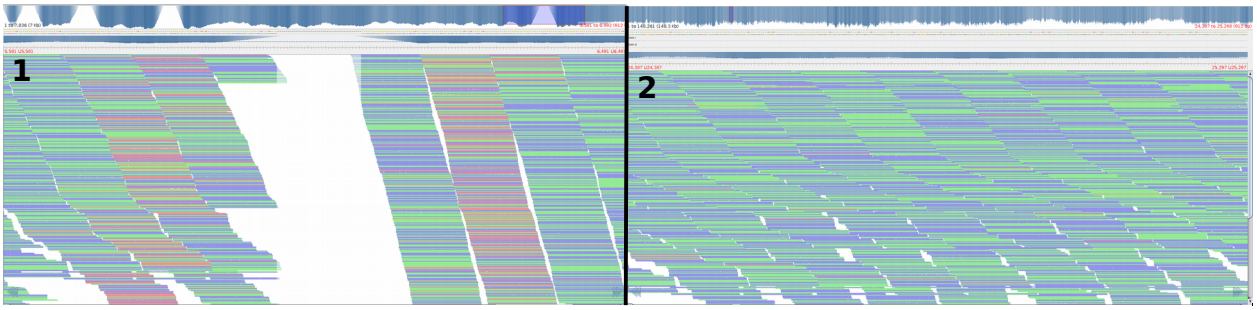

Figure S5 | Visual confirmation of 27 bp gap in the CLC assembly. Reads colored in blue or green are successfully paired together, reads in red are unpaired. (1) CLC assembly. (2) NOVOPlasty assembly.

Benchmarking results

The benchmarking study consists out of 2 mitochondrial genomes and 5 chloroplast genomes. The complete set of results can be found in the following 7 tables. Table S8 contains the results of the *G. intermedia* dataset with a read length of 126 bp and an average coverage depth of 301. These results were excluded from the benchmarking study, due to lack of memory capacity for the MIRA and MITObim assemblies.

Table S1 | Benchmarking results for the assembly of the *Avicennia marina* chloroplast.

***Avicennia marina* chloroplast**						
		NOVOPlasty	MIRA	MITObim	SOAPdenovo2	CLC
Duration	(min)	**6**	169	549	10	11
System+user time	(min)	**6**	319	1792	24	111
Memory	(GB)	7,4	21,7	5,2	17	**2,3**
Disk space	(GB)	**0,1**	30	24,5	**0,1**	2,2
Total contigs		**1**	172	87	47 449	249 654
Chloroplast contigs(scaffolds)		**1**	40	1	95	22(8)
Genome coverage		**100%**	65%	99,63%	86%	99,3%
Mismatches	(bp)	**0**	462		220	7
Ambiguous nucleotides (bp)		4	198		**0**	**0**
Accuracy		**100%**	99,52%	99,98%	99,83%	99,99%

Table S2 | Benchmarking results for the assembly of the *Oryza sativa* chloroplast (dataset SRR1328237).

***Oryza sativa* (SRR1328237) chloroplast**						
		NOVOPlasty	MIRA	MITObim	SOAPdenovo2	CLC
Duration	(min)	**10**	89	18	84	21
System+user time	(min)	**10**	212	36	102	71
Memory	(GB)	4,4	12	**1,3**	14	1,5
Disk space	(GB)	**0,1**	27	3,3	0,9	1,7
Total contigs		**1**	525	3	261 800	186 057
Chloroplast contigs(scaffolds)		**1**	54	3	203	3(7)
Genome coverage		**100%**	31,82%	8,30%	39%	99,96%
Mismatches	(bp)	**0**	136	2	222	34
Ambiguous nucleotides (bp)		1	231	234	**0**	**0**
Accuracy		**100%**	99,68%	99,98%	99,58%	99,97%

Table S3 | Benchmarking results for the assembly of the *Oryza sativa* chloroplast (dataset ERR477442).

***Oryza sativa* (ERR477442) chloroplast**						
		NOVOPlasty	MIRA	MITObim	SOAPdenovo2	CLC
Duration	(min)	5	61	6	**4**	9
System+user time	(min)	5	115	**7**	11	52
Memory	(GB)	4,3	11	**0,8**	11	1,3
Disk space	(GB)	**0,1**	20,5	0,9	0,2	0,9
Total contigs		**1**	1080	9	13 263	80 695
Chloroplast contigs(scaffolds)		**1**	35	9	77	7(3)
Genome coverage		**100%**	32,88%	4,82%	85%	99,97%
Mismatches	(bp)	**0**	37	142	27	2
Ambiguous nucleotides (bp)		**0**	157	154	**0**	0
Accuracy		**100%**	99,92%	97,90%	99,98%	**100%**

Table S4 | Benchmarking results for the assembly of the *Arabidopsis thaliana* chloroplast (dataset SRR1174256).

***Arabidopsis thaliana* (SRR1174256) chloroplast**						
		NOVOPlasty	MIRA	MITObim	SOAPdenovo2	CLC
Duration	(min)	14	217	1389	43	**6**
System+user time	(min)	**14**	367	2001	56	60
Memory	(GB)	7,6	28	14	9	**1,3**
Disk space	(GB)	**0,1**	38	75	0,6	2,5
Total contigs		**1**	65 220	1	35 686	26 104
Chloroplast contigs(scaffolds)		**1**	41	0	200	8 (3)
Genome coverage		**100%**	33,05%	0%	56,19%	99,98%
Mismatches	(bp)	1	1022	/	**0**	4
Ambiguous nucleotides	(bp)	1	367	/	**0**	**0**
Accuracy		**100%**	98%	/	**100%**	**100%**

Table S5 | Benchmarking results for the assembly of the *Arabidopsis thaliana* chloroplast (dataset SRR1810277).

***Arabidopsis thaliana* (SRR1810277) chloroplast**						
		NOVOPlasty	MIRA	MITObim	SOAPdenovo2	CLC
Duration	(min)	9	220	3224	21	**6**
System+user time	(min)	**9**	398	4977	34	55
Memory	(GB)	6,9	27,5	23,5	10	**1,4**
Disk space	(GB)	**0,1**	123	126	0,6	1,8
Total contigs		**1**	72 115	398	13 614	97 204
Chloroplast contigs(scaffolds)		**1**	34	70	265	8 (3)
Genome coverage		**100%**	18,83%	68%	94%	99,98%
Mismatches	(bp)	**1**	119	43	232	10
Ambiguous nucleotides	(bp)	**0**	19	65	**0**	**0**
Accuracy		**100%**	99,59%	99,96%	99,84%	99,99%

Table S6 | Benchmarking results for the assembly of the *Homo sapiens* mitochondrion.

***Homo sapiens* mitochondrion**							
		NOVOPlasty	MIRA	MITObim	SOAPdenovo2	CLC	ARC
Duration	(min)	**4**	152	51	25	38	21
System+user time	(min)	**4**	745	51	69	269	33
Memory	(GB)	10	34,1	1,5	51	4,5	**0,6**
Disk space	(GB)	**0,1**	71	0,2	0,6	2,6	7
Total contigs		**1**	1161	**1**	4415	273 126	16
Chloroplast contigs(scaffolds)		**1**	**1**	**1**	2	**1**	16
Genome coverage		**100%**	99,98%	**100%**	**100%**	**100%**	42%
Mismatches	(bp)	**0**	1	**0**	**0**	**0**	/
Ambiguous nucleotides	(bp)	**0**	4	**0**	**0**	**0**	**/**
Accuracy		**100%**	99,99%	**100%**	**100%**	**100%**	/

Table S7 | Benchmarking results for the assembly of the *Gonioctena intermedia* mitochondrion.

***Gonioctena intermedia* mitochondrion**							
		NOVOPlasty	MIRA	MITObim	SOAPdenovo2	CLC	ARC
Duration	(min)	**11**	536	4777	19	51	586
System+user time	(min)	**11**	1451	8077	55	347	382
Memory	(GB)	15	57,6	63,4	27	5,1	**1,9**
Disk space	(GB)	**0,1**	144	418	0,9	3	12
Total contigs		**2**	3434	2221	3199	173 117	2502
Mitochondrial contigs		**1**	**1**	2	14	**1**	48
Genome coverage (complete mt)		92,75%	**93,66%**	93,29%	75,25%	89,96%	85,39%
Mismatches	(bp)	**0**	39	74	3	**0**	2
Ambiguous nucleotides (bp)		5	194	197	**0**	**0**	**0**
Accuracy		**100%**	99,77%	99,56%	99,98%	**100%**	99,99%
Genome coverage (minus tandem repeat)		**99,98%**	98,97%	98,95%	83,57%	99,85%	94,83%
Mismatches	(bp)	**0**	**0**	12	3	**0**	2
Ambiguous nucleotides (bp)		3	170	173	**0**	**0**	**0**
Accuracy		**100%**	**100%**	99,93%	99,97%	**100%**	99,99%

Table S8 | Assembly results for the mitochondrion of a second dataset of *Gonioctena intermedia* (excluded from the benchmark).

***Gonioctena intermedia* mitochondrion 126 bp**							
		NOVOPlasty	MIRA	MITObim	SOAPdenovo2	SOAPdenovo2 (coverage depth -50%)	CLC
Duration	(min)	**14**			75	26	213
System+user time	(min)	**14**			162	68	1114
Memory	(GB)	27			54	29	**10,1**
Disk space	(GB)	**0,1**			5	1,8	5,3
Total contigs		**1**			1 796 373	552 116	590 150
Mitochondrial contigs		**1**	OUT	OUT	14	20	8(3)
Genome coverage (complete mt)		**93,32%**	OF	OF	27,74%	46,80%	88,70%
Mismatches	(bp)	**0**	MEMORY	MEMORY	14	22	13
Ambiguous nucleotides (bp)		3			**0**	**0**	2
Accuracy		**100%**			99,74%	99,74%	99,92%
Genome coverage (minus tandem repeat)		**99,99%**			30,80%	51,97%	99,85%
Mismatches	(bp)	**0**			14	22	13
Ambiguous nucleotides (bp)		1			**0**	**0**	2
Accuracy		**100%**			99,74%	99,74%	99,92%

Extra Figures

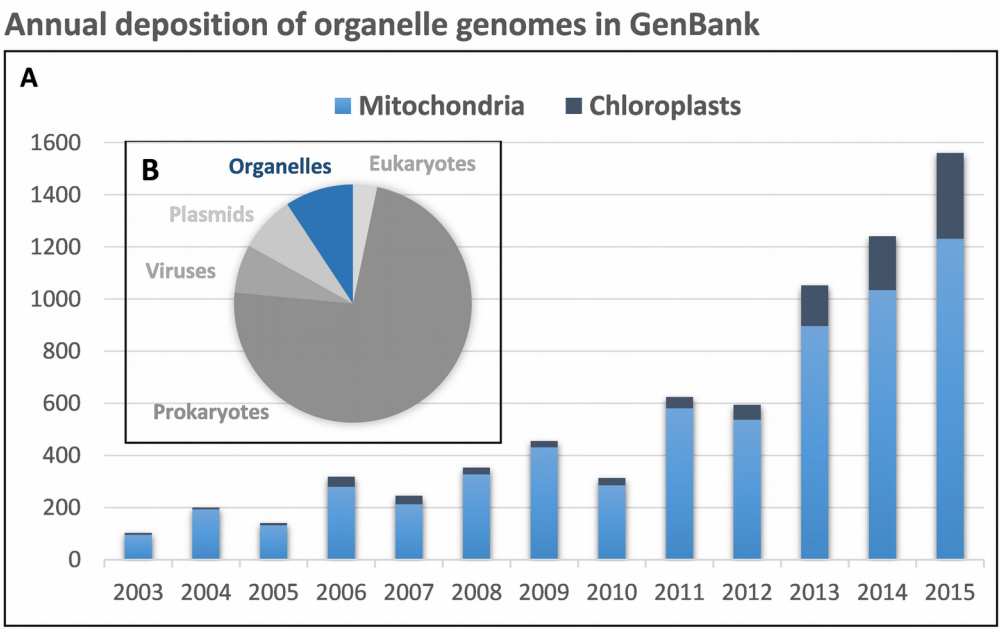

Figure S6 | Organelle genomes in GenBank. (A) Total number of deposited genomes (15704) in GenBank as of 14 March 2016. (B) Annual deposition of mitochondrial and chloroplast genomes in GenBank since 2003. Statistics from the National Center for Biotechnology Information Genome Resources (https://www.ncbi.nlm.nih.gov/genome/browse/ (13 March 2016, date last accessed)).

References

1. Benson, D.A., et al. (2013) GenBank. Nucleic Acids Research, 41, D36D42.

2. Langmead,B., Salzberg,S. (2012) Fast gapped-read alignment with Bowtie2. Nature Methods, 9, 357-359.

3. Milne,I., Stephen,G., Bayer,M., Cock,P.J.A., Pritchard,L., Cardle,L., Shaw,P.D. and Marshall,D. (2013) Using Tablet for visual exploration of second-generation sequencing data. Briefings in Bioinformatics, 14(2), 193-202.