Genome assembly using Nanopore-guided Long and Error-free DNA reads

Jean-Marc Aury[1], Mohammed-Amin Madoui[1], Stefan Engelen[1], Adriana Alberti[1], Caroline Belser[1], Laurie Bertrand[1], Corinne Cruaud[1], Arnaud Lemainque[1], Patrick Wincker[1]

[1]Commissariat à l’Energie Atomique (CEA), Institut de Génomique (IG), Genoscope, BP5706 Evry, France

http://www.genoscope.cns.fr/nas

Introduction

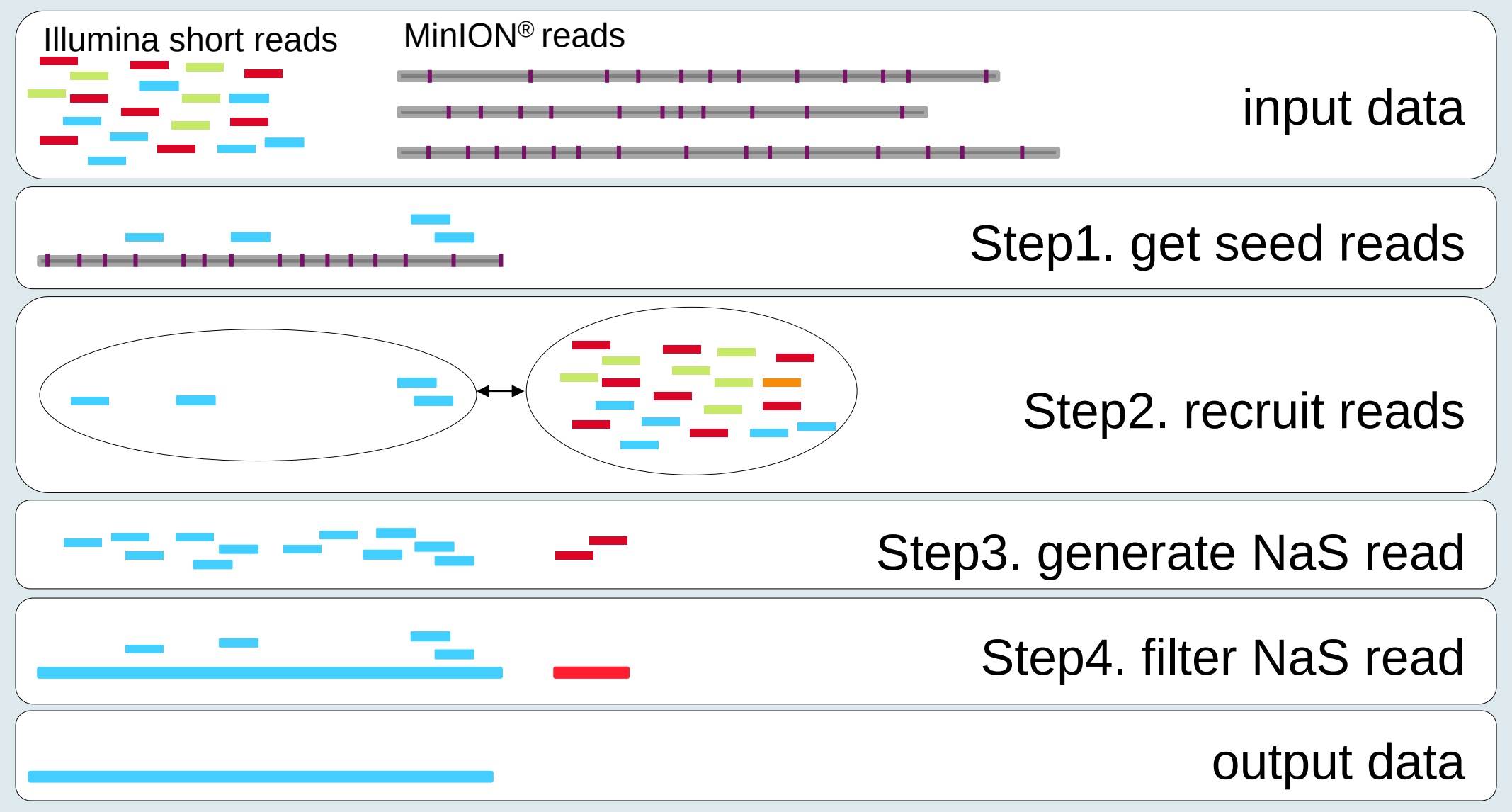
The technology of long-read sequencing now offers different alternatives to solve genome assembly problems and haplotype phasing, which can not be resolved adequately by short-read sequencing.

In 2014, Oxford Nanopore released the MinION® device, a small and low-cost single-molecule nanopore sequencer, which offers the possibility of sequencing long DNA fragments.

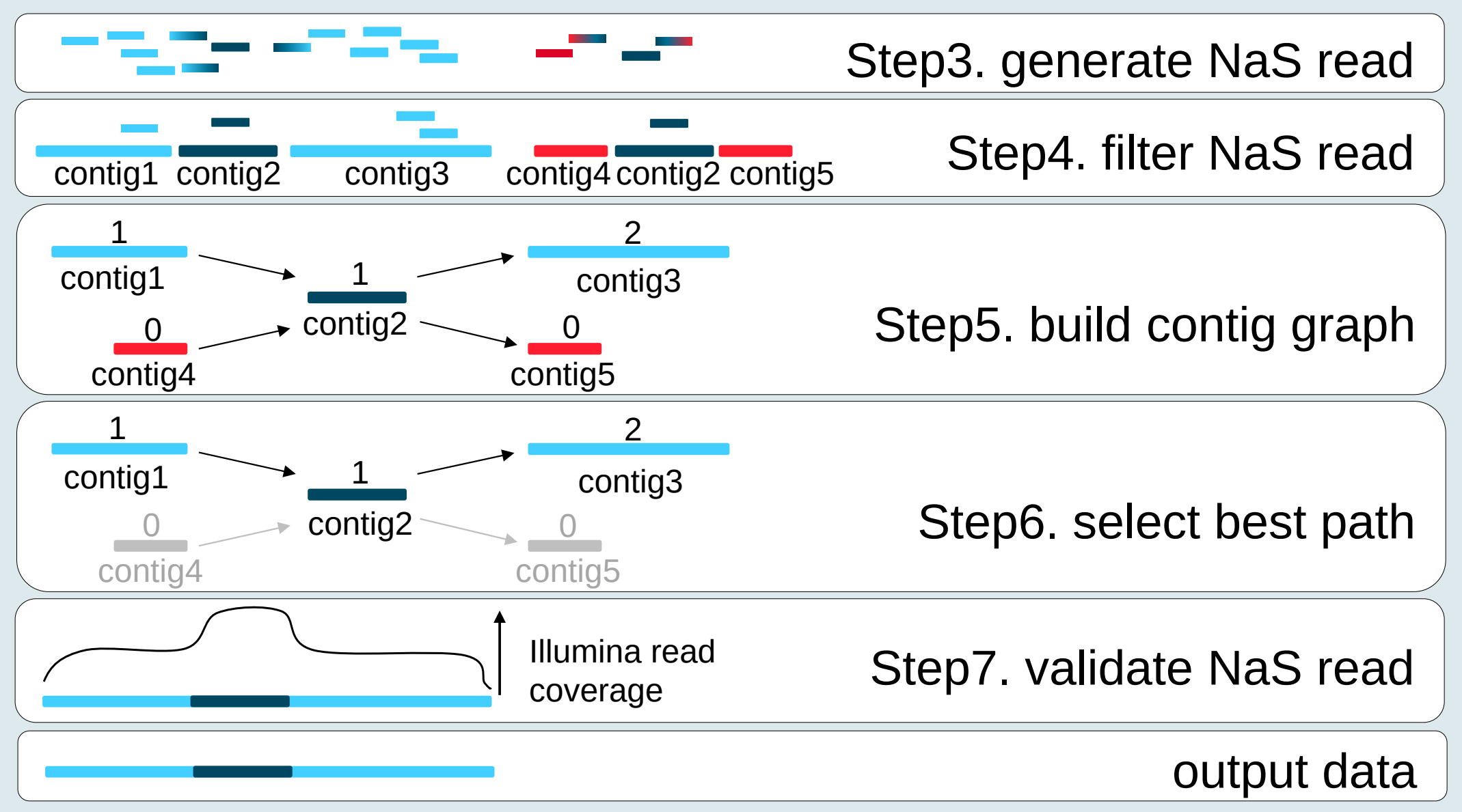
Here, we present a hybrid approach developed to take advantage of data generated using MinION ® device. Our method is able to generate NaS (Nanopore Synthetic-long) reads up to 60kb with no error and that spanned repetitive regions. We applied NaS on a well-known bacterium (*Acinetobacter baylyi ADP1*) and a small eukaryotic genome (*S. cerevisae* strain W303), and compared NaS reads and NaS assemblies with two other existing tools : Nanocorr[1] and ECtools[2].

Methods

Instead of using Illumina short reads to correct MinION® reads, we propose a method that uses the MinION® read as a template to recruit Illumina reads, and by performing a local assembly, build a high-quality synthetic read.



The NaS workflow. Inputs are the Illumina short reads and the MinION® reads (purple bars represent sequencing errors). **Step1.** Illumina reads are aligned using blat[3] on the MinION® templates to select seed-reads (blue rectangles). **Step2.** Seed-reads are used to recruit similar reads in the initial Illumina read set using compareads[4]. **Step3.** Good recruits are blue rectangles and bad recruits are red rectangles. **Step4.** OLC-based assembly (using newbler[5]) of the recruited-reads and the seed-reads. Outputted contigs (light blue and red rectangles) are then filtered using seed-read alignments. In this example, a single contig representing the final NaS read is produced.



Untangling complex regions. In the case of repetitive regions (represented by dark blue rectangles), the NaS workflow produced several contigs per MinION® template (Step3 and Step4). Indeed, the NaS read is fragmented, due to the indeterminate position of the repetitive region (contig2). **Step5.** Construction of the contig graph weighted with the seed-reads coverage of the given contig. Contig2, which represents the repetitive region, is linked to four different contigs. **Step6.** The contigs present in the path with the highest weight (contig1 – contig2 – contig3) are selected, using the Floyd-Warshall algorithm, and assembled to generate the final NaS read. **Step7.** The consistency of the synthetic NaS read is checked by aligning the initial Illumina reads set and detecting gap of coverage.

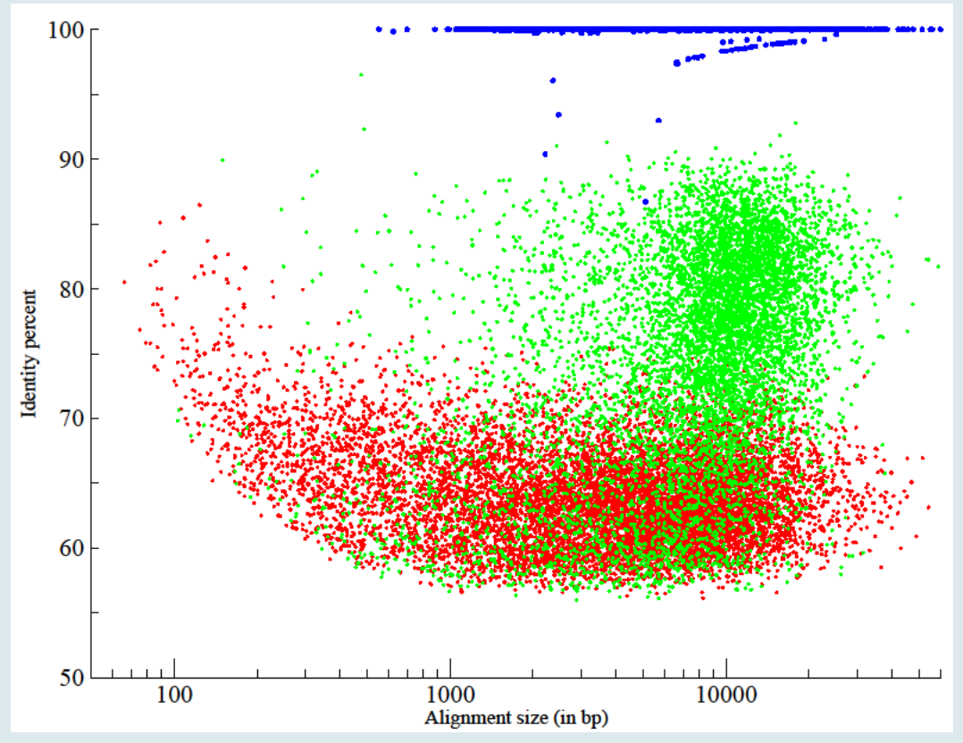
Overview of MinION® reads

Overview of the five MinION runs on Acinetobacter baylyi ADP1

	Run1	Run2	Run3	Run4	Run5
DNA library	1	2	3	4	4
DNA fragment size	8 kb	20 kb	20 kb	20 kb	20 kb
Flowcell chemistry	R7	R7	R7.3	R7.3	R7.3
Number of reads	9,241	3,990	6,052	11,957	35,252
Cumulative size (Mb)	21.4	19.3	40.8	34.5	88.9
N50 size (bp)	5,388	11,288	10,217	12,729	13,967
Average size (bp)	2,314	4,830	6,746	2,886	2,523
% of 2D reads	6.5%	13.6%	43.3%	11.6%	9.7%
% of 2D bases	14.6%	27.1%	57.1%	42.7%	44.6%

Summary statistics of the MinION reads

		1D reads	2D reads
MinION® reads aligned using LAST[6]	# reads	57,911	8,581
	# reads (>10Kb)	3,609	3,866
	Cumulative size (Mbp)	118.9	86.1
	Average size (bp)	2,052	10,033
	N50 size (bp)	11,058	12,141
	Max size (bp)	123,135	58,704
	Aligned reads	9,623 (16.6%)	7,140 (83.2%)
	Mean identity percent	56.6%	74.5%
	Max alignment size	54,158	58,656
	Error-free reads	0	0



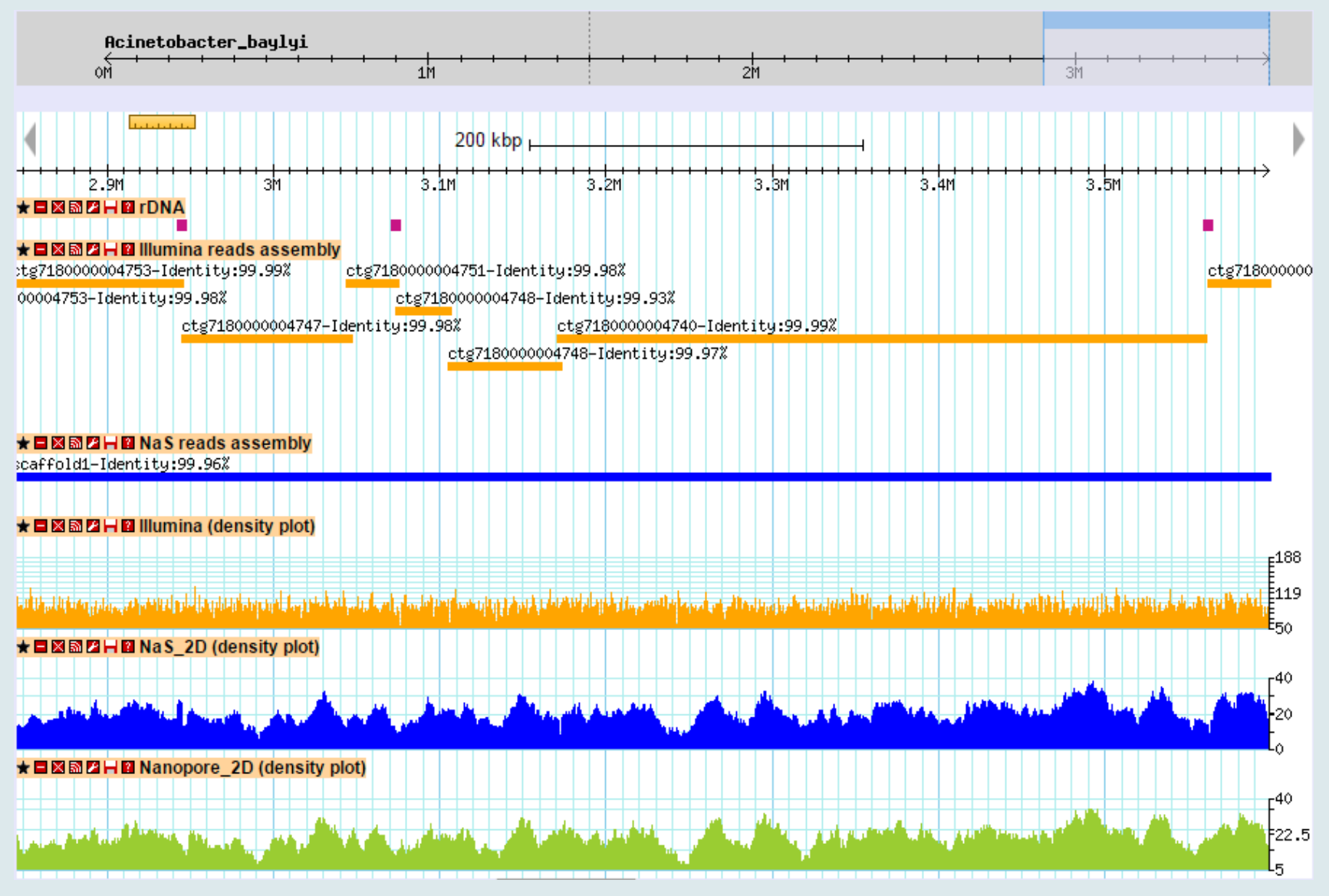
Comparison of MinION® 1D (red circles), 2D (green circles) and NaS (blue circles) reads quality.

Acinetobacter baylyi ADP1 dataset

We combined ~57X of MinION® reads with 50X of Illumina 250bp paired-end reads to produce high quality synthetic reads. To demonstrate the utility of the NaS workflow, we attempted synthetic reads assembly using the Celera asembler. Moreover, we compared NaS and two recent tools: Nanocorr[1] and ECTools[2].

Summary statistics of the MinION ® , Nanocorr, ECtools and NaS reads. Reads were aligned using bwa mem[7] with ‘-x pacbio’ option

Read set	MinION® reads	NanoCorr	ECtools	NaS
# reads	66 492	11 836	4 867	11 476
# reads >10Kb	7 475	2 915	2661	3 077
Cumulative size (coverage)	204 951 379 (57X)	67 636 754 (19X)	55 473 374(15X)	79 900 983 (22X)
Average size	3 082	5 714	11 398	6 962
N50 size	11 670	12 166	12 698	11 331
Max size	123 135	58 414	54 615	59 864
Aligned reads	16 763 (25.2%)	11 802 (99.71%)	4 867 (100%)	11 476 (100%)
Aligned bases	123 416 224 (60.2%)	67 135 095 (99.25%)	55 293 130 (99,67%)	79 838 313 (99.92%)
Mean identity percent	66.3747%	96.5665%	99.9636%	99.9847%
Perfect reads	0 (0%)	2 117 (17.93%)	4 456 (91.55%)	11 015 (95.98%)
Coverage of the reference sequence	3 598 621 (100%)	3 598 621 (100%)	3 598 621 (100%)	3 598 621 (100%)


Acinetobacter_baylyi
0M
1M
2M
3M
200 kbp
2.9M
3M
3.1M
3.2M
3.3M
3.4M
3.5M
rDNA
Illumina reads assembly
ctg7180000004753-Identity:99.99%
ctg7180000004751-Identity:99.98%
00004753-Identity:99.98%
ctg7180000004748-Identity:99.93%
ctg7180000004747-Identity:99.98%
ctg7180000004740-Identity:99.99%
ctg7180000004748-Identity:99.97%
NaS reads assembly
scaffold1-Identity:99.96%
Illumina (density plot)
188
119
50
NaS_2D (density plot)
40
20
0
Nanopore_2D (density plot)
40
22.5
5

Comparison of Illumina and NaS reads assemblies. The figure shows a capture of a 700 kb genomic region from *Acinetobacter baylyi ADP1*. The first track contains rDNA clusters 5, 6 and 7 (purple rectangles). The orange rectangles represent alignments of contigs from the Illumina-only assembly, whereas blue rectangle represents the alignment of the NaS assembly contig. The three plots represent respectively the coverage of Illumina, Nas 2D and MinION® 2D reads. We observed that breakpoints of the Illumina assembly coincide in part with rDNA clusters, in contrast with the NaS assembly which exhibits a perfect alignment.

Summary statistics of genome assemblies produced using Celera assembler[8]. Metrics were computed using Quast[9].

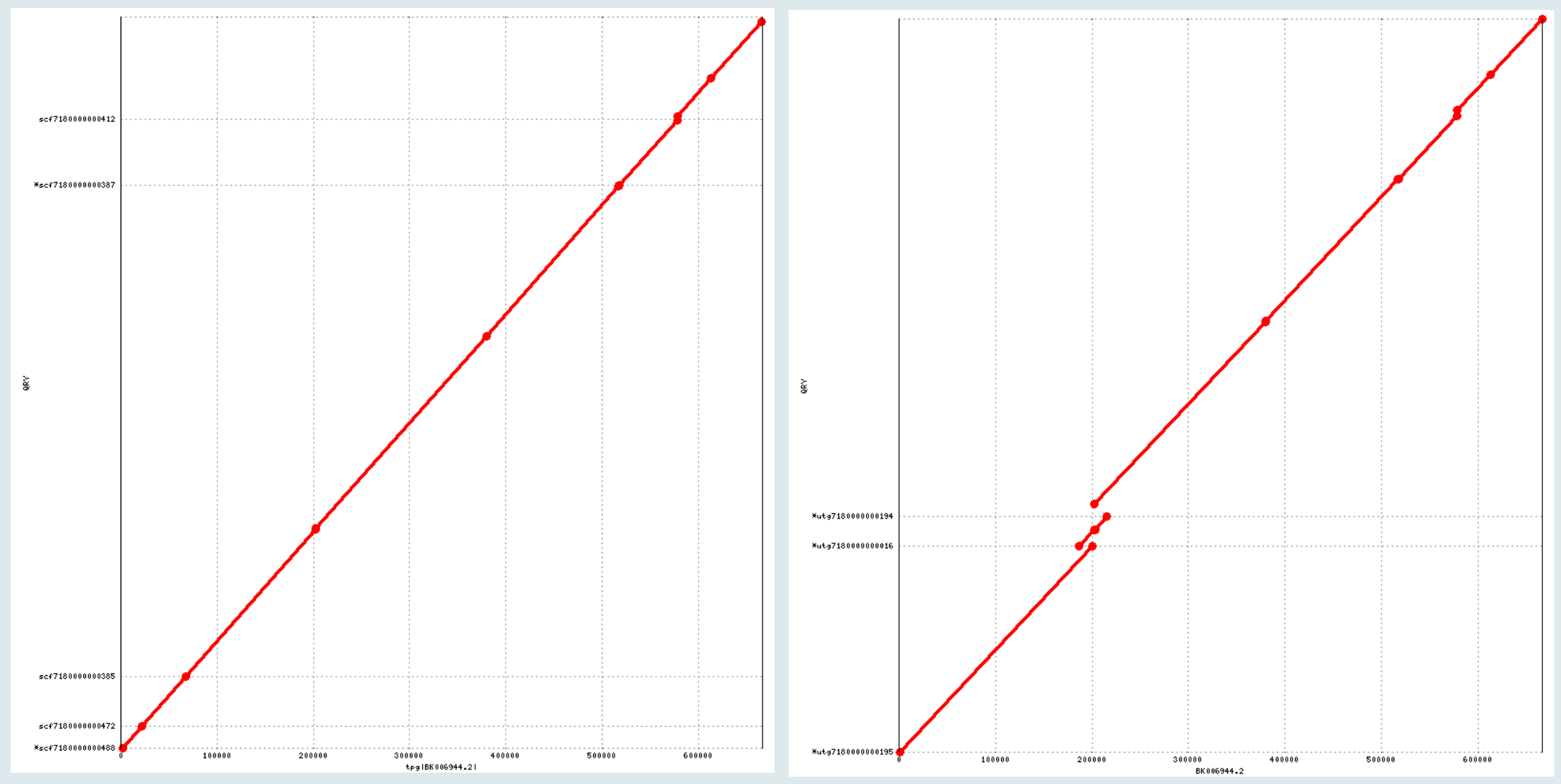
Metrics (Quast)	Illumina	NaS	Nanocorr	ECtools
#contigs	20	3	6	4
Assembly size	3 592 537	3 635 796	3 620 823	3 616 882
N50	326 117	3 609 416	3 604 474	2 468 787
L50	5	1	1	1
N90	140 386	3 609 416	3 604 474	954 595
L90	11	1	1	2
MinContigSize	3 547	9 380	1 458	54 816
MaxContigSize	520 993	3 609 416	3 604 474	2 468 787
ID%	99.99	99.99	99.98	99.97
Max aln	520 993	1 442 823	1 825 329	1 701 411
NA50	290 660	1 212 310	2 598 906	954 061
NA75	194 326	953 958	681 989	496 599
Genome fraction (%)	99.735	100	100	99.971
# misassemblies	4	2	2	3
# local misassemblies	3	2	8	6
# mismatches per 100 kbp	6.49	0.78	4.95	9.37
# indels per 100 kbp	0.33	0.44	3.11	6.75

Yeast dataset

We used the dataset provided with the nanocorr[1] tool, based on the W303 strain of *S. cerevisae*, in combination with Illumina paired-end reads and compared our results with the one obtained with Nanocorr[1].

Summary statistics of the MinION ® , Nanocorr and NaS reads. Reads were aligned on the W303 PacBio asembly, using bwa mem with ‘-x pacbio’ option

Read set	MinION® reads	NanoCorr (all MinION® reads)	NaS (template and 2D reads only)
Short read dataset	NA	30X of PE @ 300bp	50X of PE @ 250bp
# reads	267 768	105 281	71 793
# reads >10Kb	34 300	12 254	6 141
Cumulative size (coverage)	1 465Mb (117X)	488 Mb (39X)	426 Mb (34X)
Average size	5 473	4 636	5 938
N50 size	7 937	8 294	7 085
Max size	146 992	72 936	45 745
Aligned reads	68 215 (25.47%)	104 094 (98.87%)	71 614 (99.75%)
Aligned bases	411 Mb (28.02%)	475 Mb (97.27%)	424 Mb (99.47%)
Mean identity percent	55.4937%	97.5005%	99.9246%
Perfect reads	0 (0%)	3 334 (3.2%)	56 991 (79.58%)
Coverage of the reference sequence	12 353 715 (99.86%)	12 336 482 (99.72%)	12 196 844 (98.6%)



Dot-plot alignment and comparison of NaS (left) and nanocorr (right) assemblies with S288C chromosome 11

Summary statistics of genome assemblies produced using Celera assembler. Metrics were computed using Quast.

Metrics (Quast)	Illumina	Nanocorr	NaS
# contigs	6 953	204	125
Assembly size	14 910 895	14 000 895	11 845 583
GC (%)	38.71	38.64	38.14
Reference GC (%)	38.21	38.21	38.21
N50	53 444	334 484	148 384
L50	80	15	21
N90	544	20 612	45 795
L90	3 137	98	75
# misassemblies	72	161	83
# misassembled contigs	52	107	51
# local misassemblies	22	44	13
Genome fraction (%)	97.0	92.2	91.5
Duplication ratio	1.18	1.23	1.04
# mismatches per 100 kbp	91.11	72.65	36.65
# indels per 100 kbp	9.20	34.17	7.23
average id%	99.79	99.69	99.93

Conclusion

The approach we present here is an efficient method to sequence genome by combining advantages of Illumina and the new Oxford Nanopore technologies. These sequencing technologies are commercialized through two desktop instruments, the MinION® device and the MiSeq sequencer respectively, that have the advantage to be small and relatively low cost.

Our method offers the opportunity to sequence microbial or small eukaryotic genomes in a very short time, even in small facilities.

This hybrid approach presents an interesting alternative compared with standard strategies, such as SMRT of Pacific BioSciences and Illumina TruSeq Synthetic long reads. For example, our approach is straightforward in terms of library preparation, as well as laboratory and information technology infrastructure requirements.

Moreover, we demonstrated that although the Oxford Nanopore technology is a relatively new sequencing technology, currently with a high error rate, it is already useful in the generation of high-quality genome assemblies.

1. Sara Goodwin , James Gurtowski , Scott Ethe-Sayers , Panchajanya Deshpande ,Michael Schatz , W Richard McCombie: **Oxford Nanopore Sequencing and de novo Assembly of a Eukaryotic Genome**. *bioRxiv* doi: http://dx.doi.org/10.1101/013490
2. https://github.com/jgurtowski/ectools
3. Kent WJ: **BLAT--the BLAST-like alignment tool**. *Genome Res* 2002, **12**(4):656-664.
4. Maillet N, Lemaitre C, Chikhi R, Lavenier D, Peterlongo P: **Compareads: comparing huge metagenomic experiments**. *BMC bioinformatics* 2012, **13 Suppl 19**:S10.
5. http://www.454.com/products/analysis-software/
6. Kielbasa SM, Wan R, Sato K, Horton P, Frith MC: **Adaptive seeds tame genomic sequence comparison**. *Genome Res* 2011, **21**(3):487-493.
7. Li H, Durbin R: **Fast and accurate long-read alignment with Burrows-Wheeler transform**. *Bioinformatics* 2010, **26**(5):589-595.
8. Myers EW, Sutton GG, Delcher AL, Dew IM, Fasulo DP, Flanigan MJ, Kravitz SA, Mobarry CM, Reinert KH, Remington KA *et al*: **A whole-genome assembly of Drosophila**. *Science* 2000, **287**(5461):2196-2204.
9. Gurevich A, Saveliev V, Vyahhi N, Tesler G.**: QUAST: quality assessment tool for genome assemblies**. *Bioinformatics.* 2013 Apr 15;29(8):1072-5. doi: 10.1093/bioinformatics/btt086. Epub 2013 Feb 19.