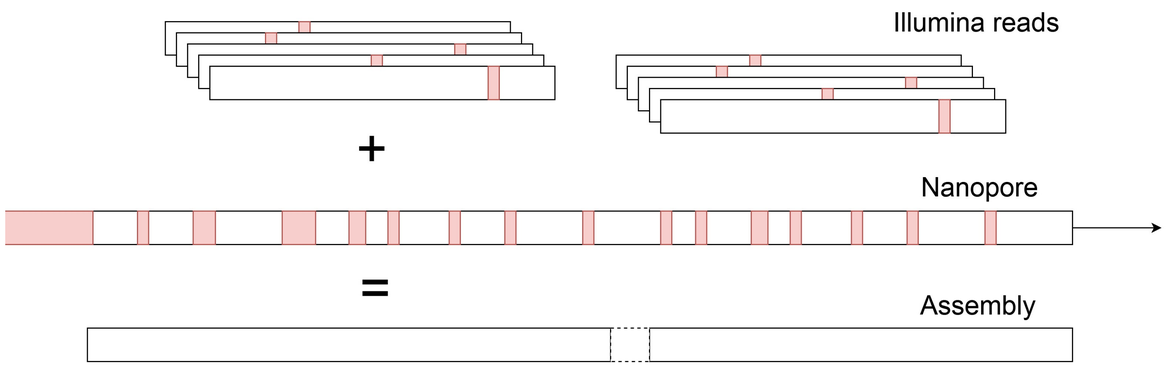
Improving metagenomic assembly using Nanopore Read-Until technology

Sergey Kazakov[1], Vladimir Ulyantsev[1], Sergey Nurk[2]

[1] ITMO University, Saint Petersburg, Russia
[2] SPbSU, Saint Petersburg, Russia

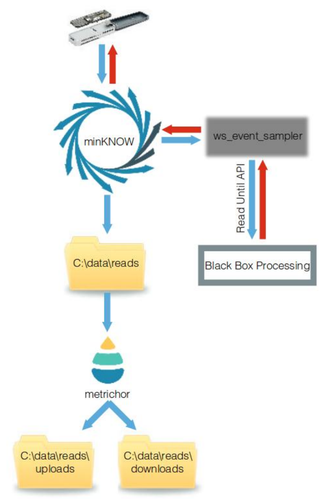
Aim

- Assembling metagenome data together – is one way to analize metagenomes
- Long Nanopore reads with short accurate Illumina reads can produce better assembly than separate ones
- **Aim** – to improve hybrid metagenomic assembly (Illumina + Nanopore) using Nanopore Read-Until technology



Read Until technology

- Online feedback interface, allowing ‘release’ of current molecule
- It enables the selective rejection of an individual reads during sequencing
- Among many incredible things, it can significantly reduce effective cost of assembly projects!



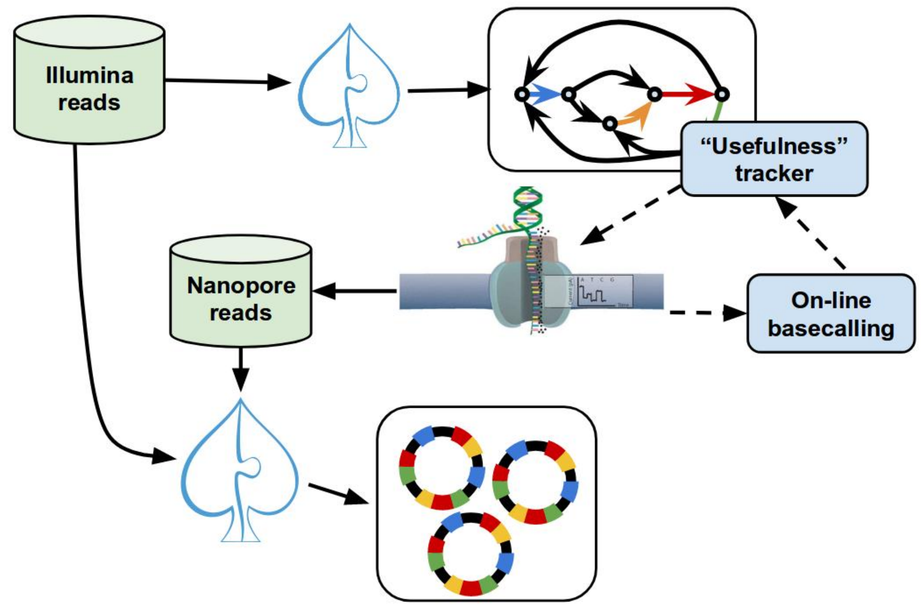
Loose et al “Real-time selective sequencing using nanopore technology”, Nature methods, 2016

Problem

- What strategy of rejecting reads will give the best set of Nanopore reads to improve hybrid assembly?

Possible strategies:
- Filtering contaminant DNA (e.g. host DNA in microbiome projects)
- Limiting coverage of highly-abundant species
- Ignoring reads **probably useless** for repeat resolution
- ...

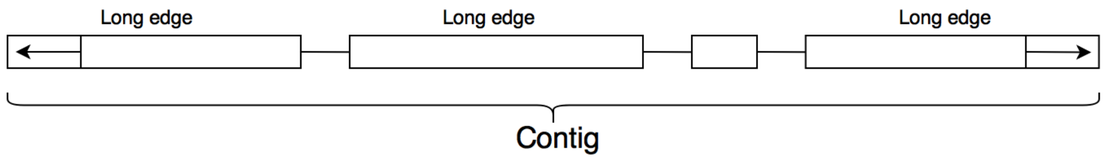
Pipeline



Experiments

1. Bacterial repeat resolution:
- Enriching for reads useful for repeat resolution in *E.coli K12*
- *Useful* – reads spanning unresolved repeat

Rejection strategy:
- Index flanks of length **L** on both sides of the contig
  a. **L** = 10 Kbp, depends on read length distribution
  b. Consider only “unique regions” (long edges in graph)
- Reject if prefix doesn’t fall in the flank or oriented within the contig
- Otherwise *accept*. **NB.** *accepted* != *useful*



2. Filtering contaminant human DNA:
- *Useful* – reads longer 1kbp NOT aligning to host reference

Rejection strategy:
- Index reference
- Reject if prefix maps to it

Results

1. E.coli K12 (R9, Loman) dataset* *with skipping reads ≤ 5 kbp

	No Filtering (Baseline)	Strategy Enabled	Enrichment
Useful reads count	33	**83**	2.52x
Useful length	11.4%	22.7%	1.99x
Accepted reads # - length	383 / 383 = 100% - 100% of signal	137 / 5051 = 2.7% - 31% of signal	

1. E.coli K12 (R9.4, UCSD) dataset*

	No Filtering (Baseline)	Strategy Enabled	Enrichment
Useful reads count	79	**195**	2.47x
Useful length	32.8%	44.4%	1.35x
Accepted reads # - length	498 / 498 = 100% - 100% of signal	331 / 3264 = 10.1% - 59% of signal	

2. Saliva metagenome (R9.4, UCSD) dataset*

	No Filtering (Baseline)	Strategy Enabled	Enrichment
Useful reads count	124	**271**	2.19x
Useful length	33.6%	**71.5%**	2.13x
Accepted reads # - length	454 / 454 = 100% - 100% of signal	337 / 1001 = 34% - 91% of signal	