Genomics on a Shoestring Budget

Thomas Gatter

17.02.2023

Bioinformatics, Leipzig University

Oh the costs

Can we study the genome of novel organisms on a small budget?

Oh the costs

Can we study the genome of novel organisms on a small budget?

Don’t do it. You just get a lot of money or it won’t be worth it. But thats also way to much spending for this proposal. So you won’t get it here.

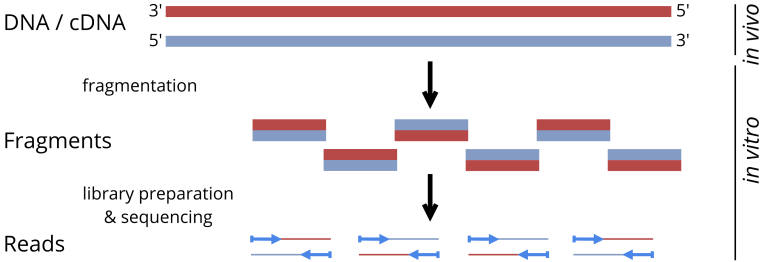
– Every Reviewer

Oh the costs

Proposed Solution:

Get a lot of Money (somehow) $\rightarrow$ Sequence with PacBio HiFi + Hi-C
$\rightarrow$ Process with Super-Computer $\rightarrow$ Profit (?)

Option 1: Short Read Sequencing

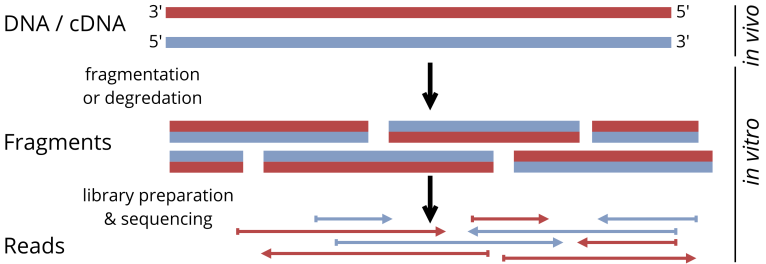


Read length: 2 × 75-150 bp

Accuracy: > 99%

Provider: Illumina, Thermofischer, Roche

Option 2: Long Read Sequencing



Read length: avrg. $>$ 10 kbp, often much longer
Accuracy: $\approx$ 90% $\rightarrow$ better with newer iterations
Provider: Oxford Nanopore(, Pacific Biosciences)

Short Read vs Long Read Assembly

Short Read Assembly

+ cheap
+ accurate
+ well established with de Bruijn or String-Graphs
- cannot resolve longer repeats

Short Read vs Long Read Assembly

Short Read Assembly

+ cheap
+ accurate
+ well established with de Bruijn or String-Graphs
- cannot resolve longer repeats

Long Read Assembly

+ resolve long complex regions
- less accurate
- more expensive
- expensive pairwise overlapping
- require high coverage

Short Read vs Long Read Assembly

Short Read Assembly

+ cheap
+ accurate
+ well established with de Bruijn or String-Graphs
- cannot resolve longer repeats

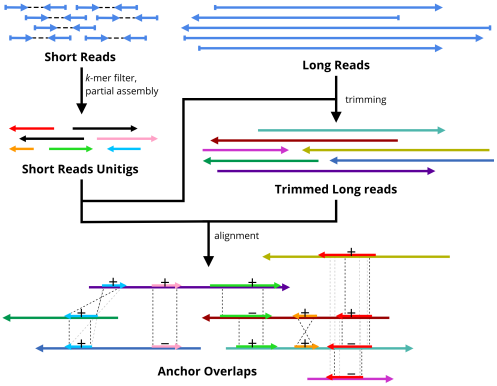
Long Read Assembly

+ resolve long complex regions
- less accurate
- more expensive
- expensive pairwise overlapping
- require high coverage

Combine both!

The LazyB Workflow

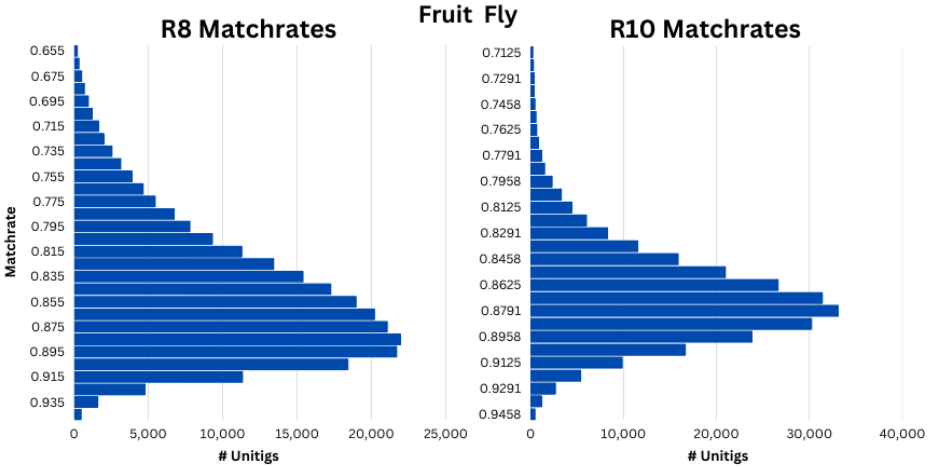
Short reads are assembled to build accurate anchors between long reads.



Then some ma(th)gic happens...

Fruit Fly

X	Tool	compl.[%]	#ctg	#MA	NA50
∼5×	LazyB	71.624	1879	68	64415
	Canu	-	-	-	-
	Wtdbg2	6.351	2293	2	-
	HASLR	24.484	1407	10	-
	DBG2OLC	25.262	974	141	-
	Wengan	81.02	2129	192	77215
∼10×	LazyB	80.111	596	99	454664
	Canu	49.262	1411	275	-
	Wtdbg2	41.82	1277	155	-
	HASLR	67.059	2463	45	36979
	DBG2OLC	82.52	487	468	498732
	Wengan	84.129	926	237	221730
∼45×	ABySS	83.628	5811	123	67970


Fruit Fly
R8 Matchrates
Matchrate
0.655
0.675
0.695
0.715
0.735
0.755
0.775
0.795
0.815
0.835
0.855
0.875
0.895
0.915
0.935
0
5,000
10,000
15,000
20,000
25,000
Unitigs
R10 Matchrates
0.7125
0.7291
0.7458
0.7625
0.7791
0.7958
0.8125
0.8291
0.8458
0.8625
0.8791
0.8958
0.9125
0.9291
0.9458
0
10,000
20,000
30,000
40,000
Unitigs

Fruit Fly - R 8 vs R10

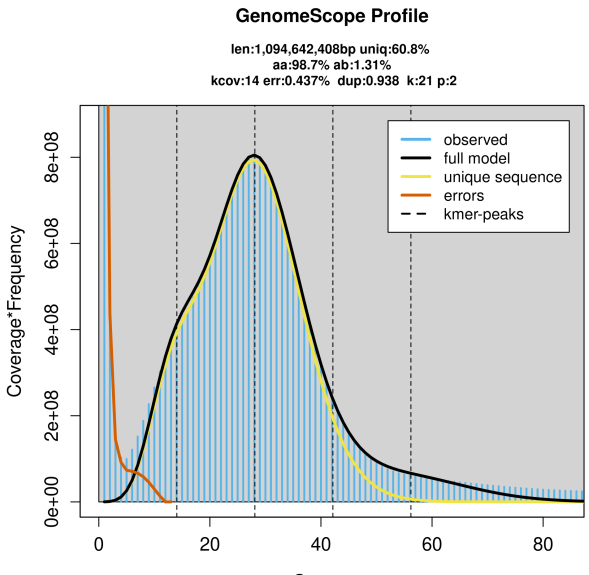
X	Tool	Chem.	compl.[%]	#ctg	#MA	NA50
∼5×	LazyB	R8	71.624	1879	68	64415
		R10	71.028	708	91	189244
	DBG2OLC	R8	25.262	974	141	-
		R10	33.413	895	161	-
	Wengan	R8	81.02	2129	192	77215
		R10	78.564	1645	140	117504
∼10×	LazyB	R8	80.111	596	99	454664
		R10	78.206	191	91	1031893
	DBG2OLC	R8	82.52	487	468	498732
		R10	87.519	230	281	1016141
	Wengan	R8	84.129	926	237	221730
		R10	83.037	483	182	528879
∼45×	ABySS		83.628	5811	123	67970

Meccus longipennis

- blood sucking bug
- native to latin america
- host for Trypanosoma cruzi
- vector for Chagas disease

Genome virtually unexplored...

Lets get practical...



Estimated genome size of 1.1 Gb

⇒

6.2× Nanopore

40× Illumina

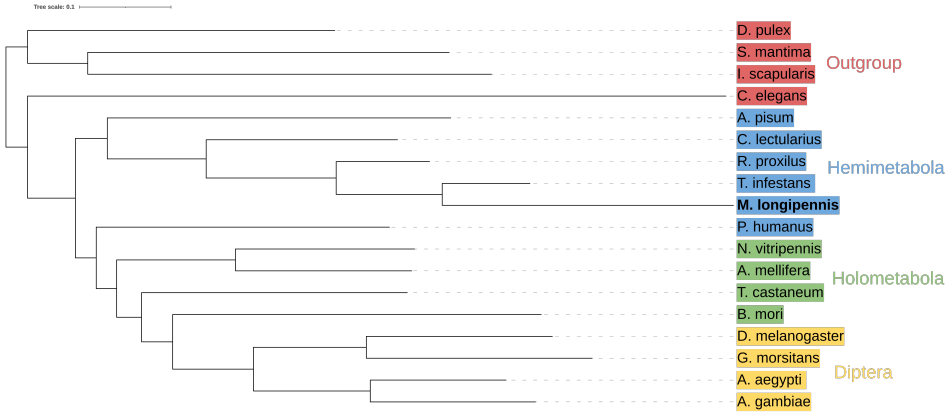
Assembly Statistics for M. longipennis

(with BUSCO Hemiptera ODB10)

Tool	#ctg	Assembled bp	N50	BUSCO C + F
DBG2OLC	-	-	-	-
Wengan	677	5,681,409	8,642	50 (2.0%) + 2 (0.1%)
HASLR	68,585	416,784,090	8,326	1280 (51.0%) + 232 (9.2%)
ABySS	695,368	893,209,008	1,582	1314 (52.4%) + 500 (19.9%)
LazyB	48,074	788,046,408	22,713	1596 (63.5%) + 181 (7.2%)

Annotation

- RepeatModeler + BRAKER2 → 19353 Proteins
- OrthoFinder for related organisms → 7592 Orthologues to R. prolixus
- including orthologues for hematophagy und immune related proteins



Thank You!

Peter Stadler
Javier T Granados Riverón

Sarah von Löhneysen
Kevin Klein
Felix Kühnl