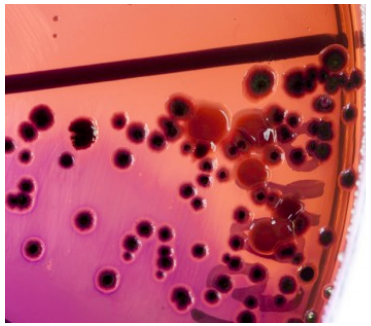
Analysis of GSC MinION run SJ_42

René Warren, March 28th, 2017



Sample background

- Non-tuberculous mycobacterium rec'd by BCCDC
  - BC patient with pneumonia
  - "pre-outbreak", likely *M. chimaera* strain
  - slow-growing, common soil & water, rarely infectious (CDC)

- *M. chimaera* caused global outbreak (SD. 2015)
  - cardiac bypass surgery patients, using spec. brand of blood heater-cooler (water circuit)
  - AU outbreak strain MC045 sequenced 2016

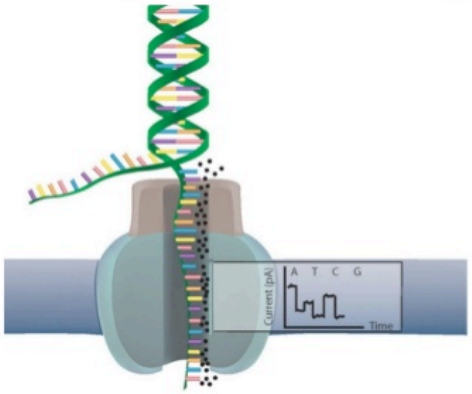
https://wwwnc.cdc.gov/eid/article/22/6/16-0045_article

Run stats

170130_A66989		
Metrichor: (Oxford Nanopore Tech [ONT] base caller)		
	Template	**2D**
Read Count	127957	
Total Yield	111.53 M Bases	36.77 M Bases
Sequence Length - Average	874 bases	652 bases
Sequence Length - Median	495 bases	385 bases
Sequence Length - Mode	309 bases	337 bases
Longest Read	71.66 K bases	10.81 K bases
QScore - Average	7.3	12.7
QScore - Median	7.6	12.8
QScore - Mode	8.2	13.1

Run courtesy:
Steve Pleasance

Nanopore - technology

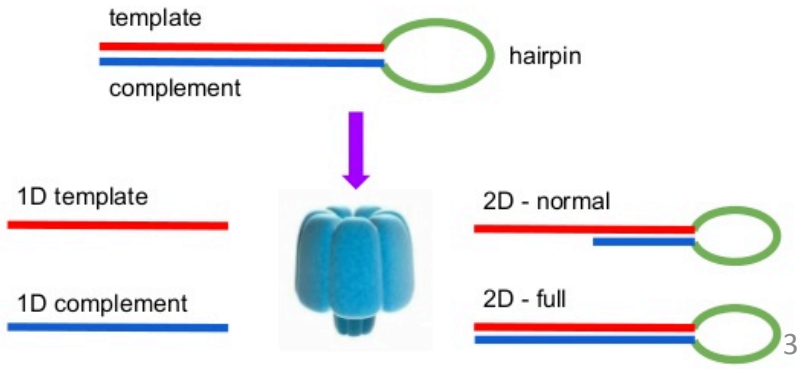


Signal is measured from 5 bases

Timing is irregular

Base modifications do alter the signal

Nanopore - reads



source: Torsten Seemann

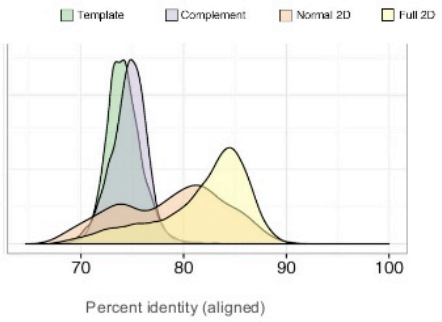
Run stats : poRe

standardized toolkit for the analysis of nanopore datasets

poRe:			
		Fail	
	Template	Complement	2D
Failed			23206
Number	74885	26822	3616
Max length	71655	85599	6369
Min length	5	5	79
Mean length	979	645	528
SD length	1267	2273	530
poRe:			
		Pass	
	Template	Complement	2D
Failed			0
Number	52775	52775	52775
Max length	10638	11443	10808
Min length	103	101	88
Mean length	725	573	661
SD length	730	608	690

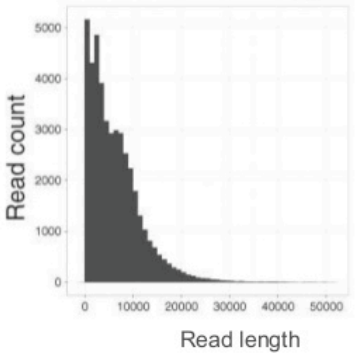
~207,000 1D reads

Nanopore - error rate



:: 5-mer errors
:: Homopolymer issues
:: Not modelling base mods yet
:: Changes with pore & motor enzyme

Nanopore - read lengths



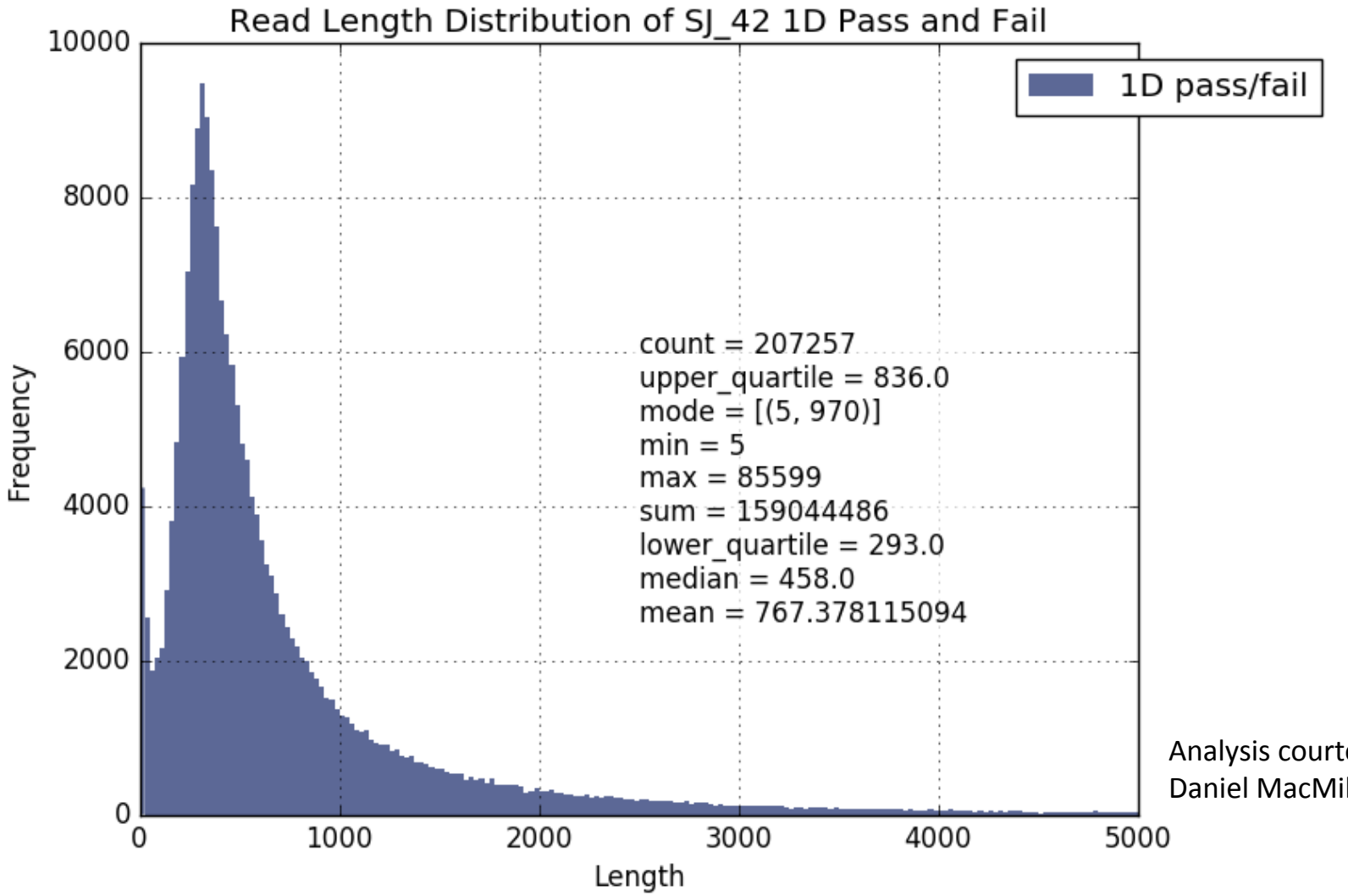
Read length is not limited by technology but by library preparation.

Can get >100kbp reads.

But not trivial to do so!

source: Torsten Seemann

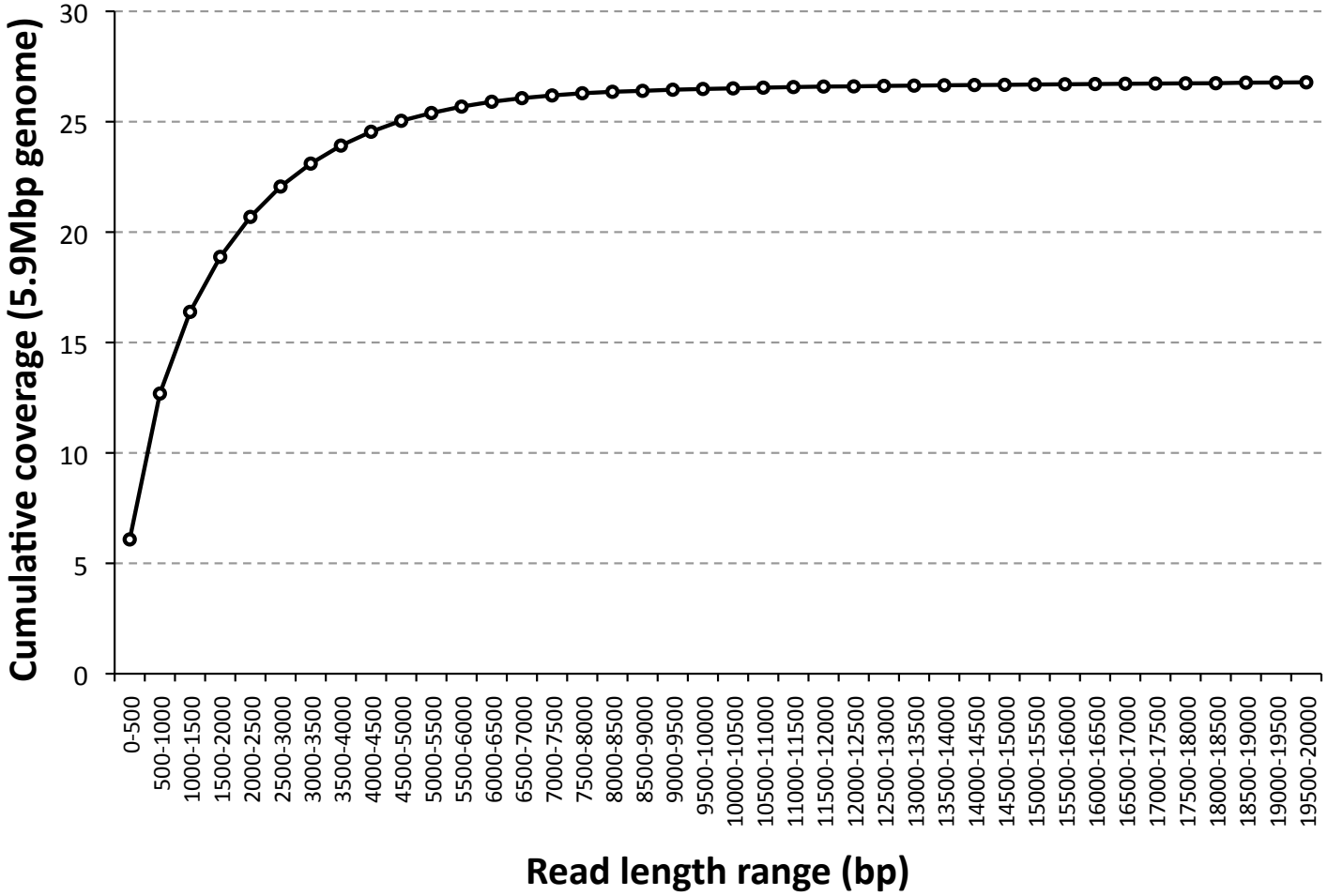
SJ_42 ONT sequence length distribution



Analysis courtesy:
Daniel MacMillan

Run stats

Size range (kbp)	Number of sequences*	Number of bases	Estimated coverage**
0-1	141,767	67,359,033	11.42
1-10	40,882	83,725,372	14.19
10-100	248	5,357,291	0.91

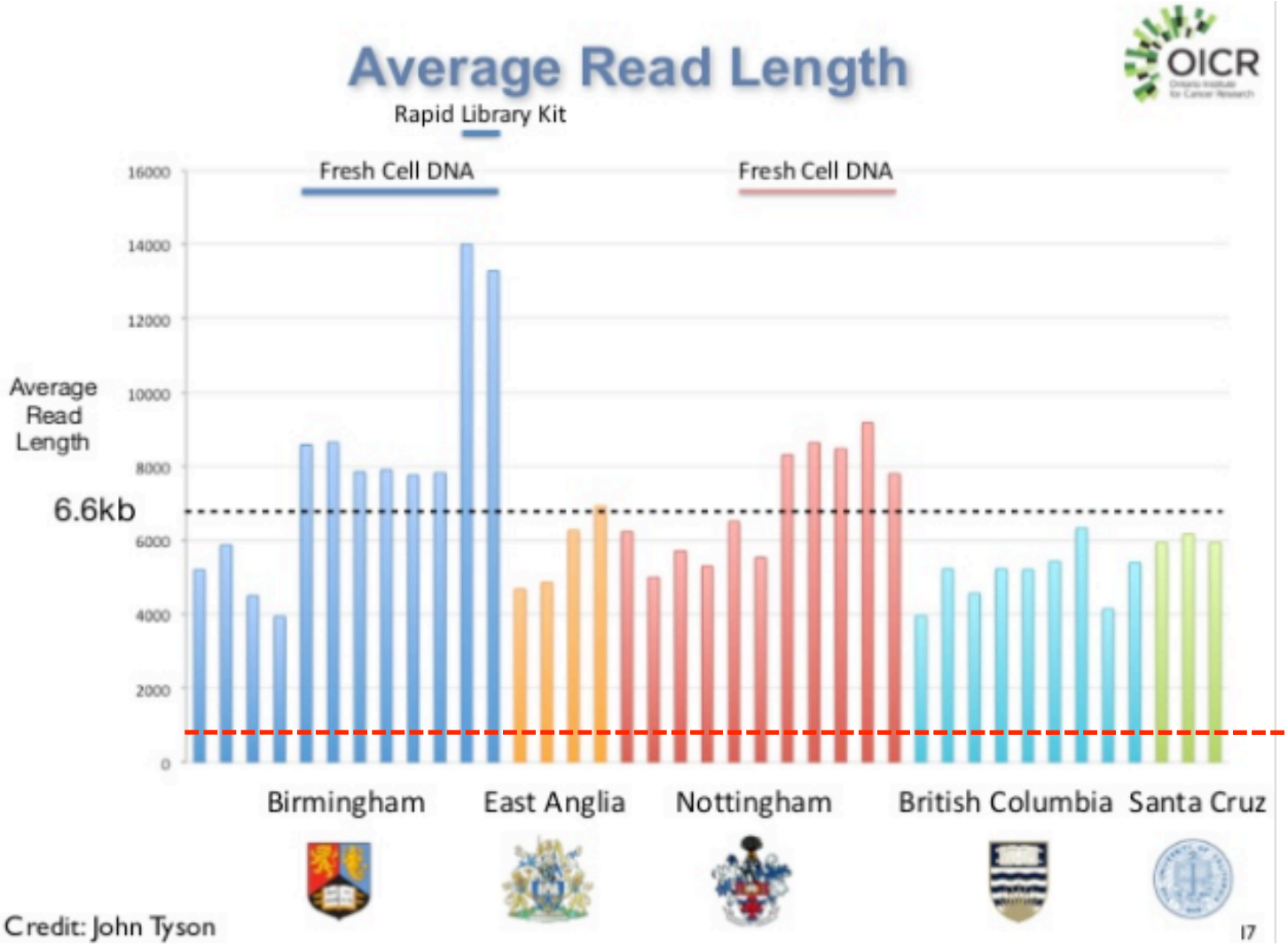


*200 bp or larger

**based on a 5.9Mbp genome

Stability of nanopores

distance from manufacture



SJ_42

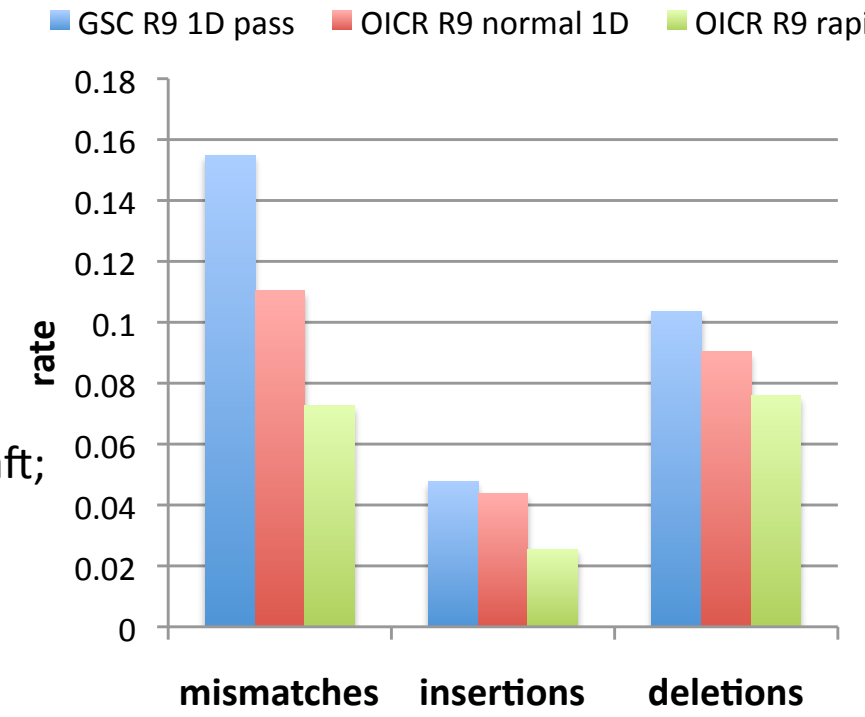
slide: Jared Simpson

ONT 1D error models (NanoSim*)

	M. chimaera	*E. coli*	
		OICR Toronto	OICR Toronto
comparisons	**GSC R9 1D pass**	**OICR R9 normal 1D**	**OICR R9 rapid 1D** **
# bases	68,465,920	668,828,202	1,481,822,568
# reads	105,550	100,554	164,458
median length	436	6,436	5,946
max length	11,443	144,661	131,969

10-fold! ↔

**aims at producing high-quality 1D reads

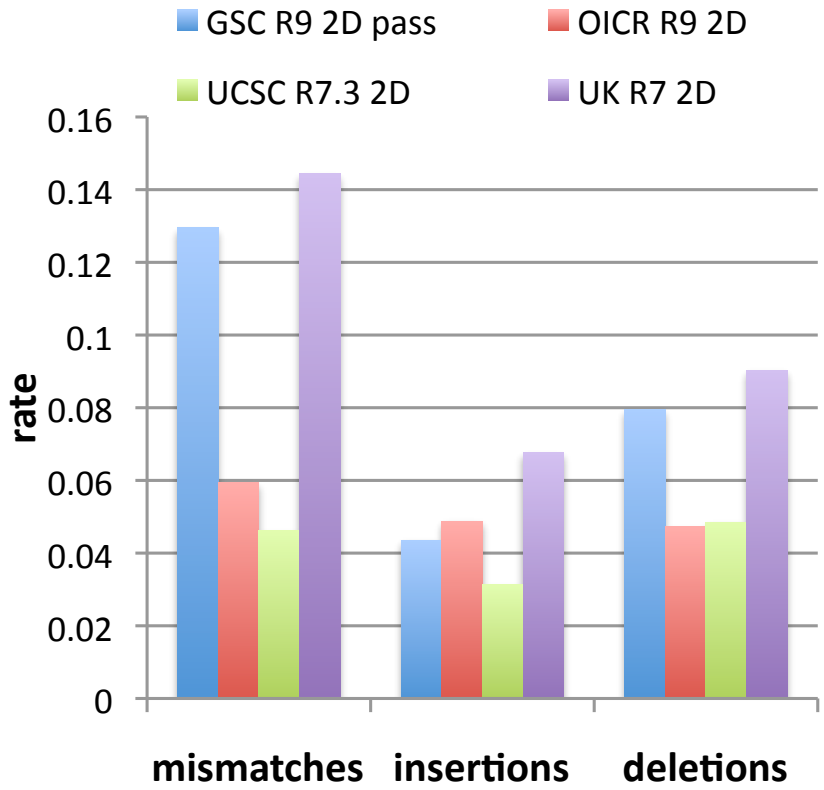


* Using reconstructed draft; difficult to unequivocally ascertain origin of errors

Analysis courtesy: Chen Yang

ONT 2D error models (NanoSim)

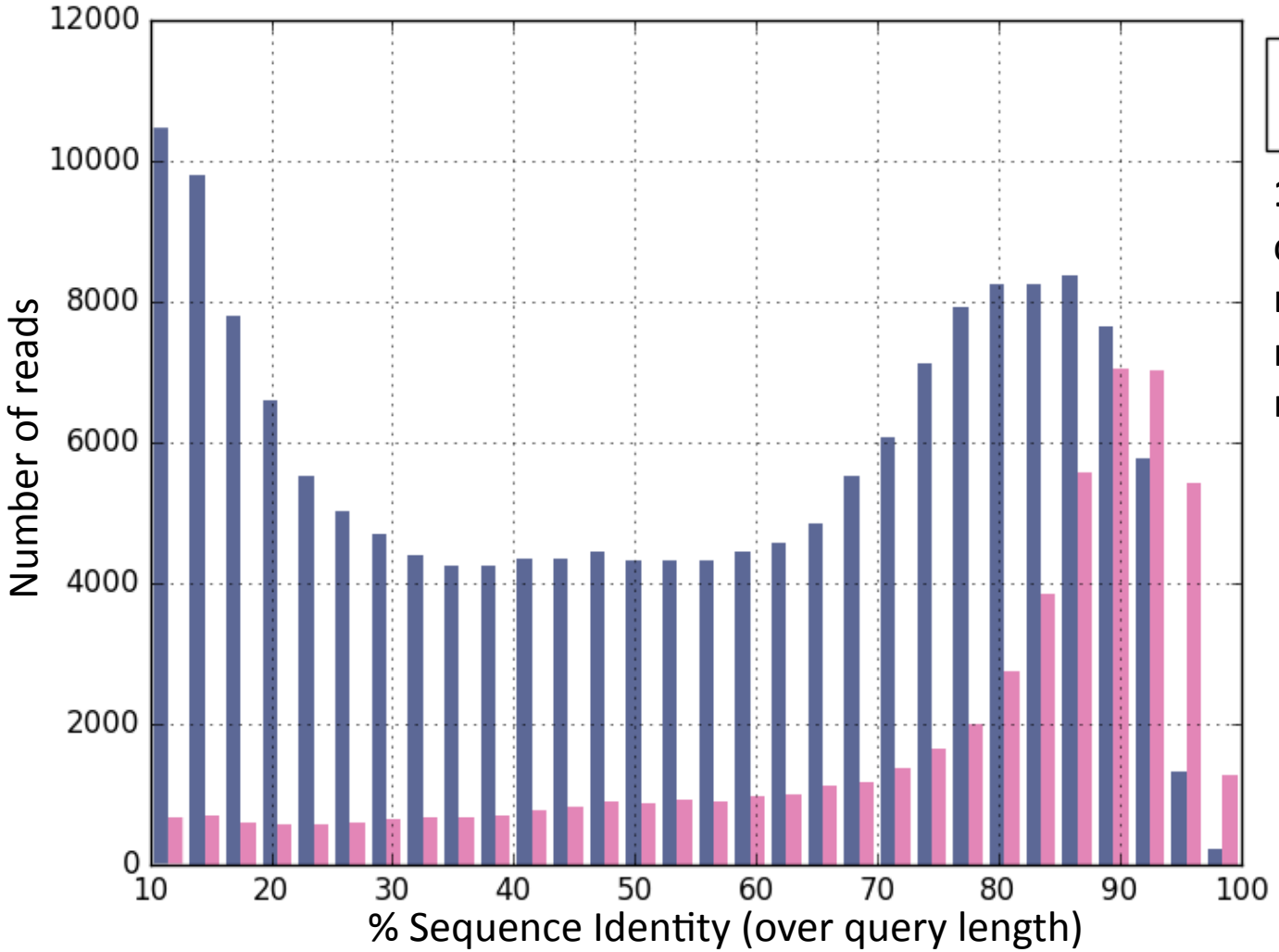
	M. chimaera	*E. coli*		
		Toronto	California	UK
comparisons	**GSC R9 2D pass**	**OICR R9 2D**	**UCSC R7.3 2D**	**UK R7 2D**
# bases	34,863,878	244,275,647	319,181,306	158,867,566
# reads	52,775	31,858	45,049	20,640
median length	440	7,603	6,906	6,503
max length	10,808	64,218	45,588	47,422



Analysis courtesy: Chen Yang

Read quality assessment

- based on final *M. chimaera* SJ_42 draft



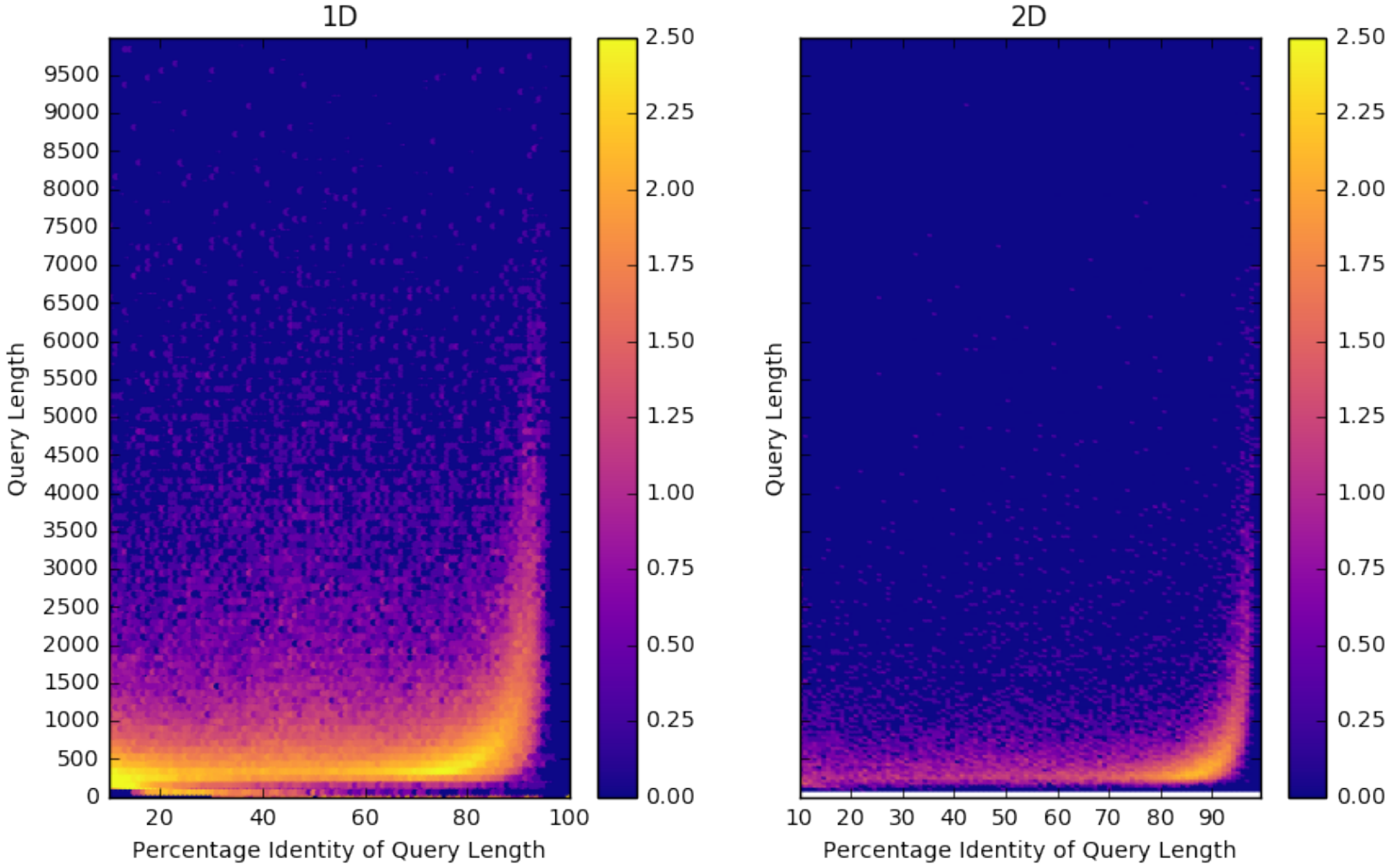
1D
count = 168953
max = 100%
median = 55.0%
mean = 52.3%

2D
count = 53,483
max = 99%
median = 84.0%
mean = 74.2%

Analysis courtesy: Daniel MacMillan

Read quality assessment

sequence identity vs. length



Analysis courtesy: Daniel MacMillan

de novo genome assembly

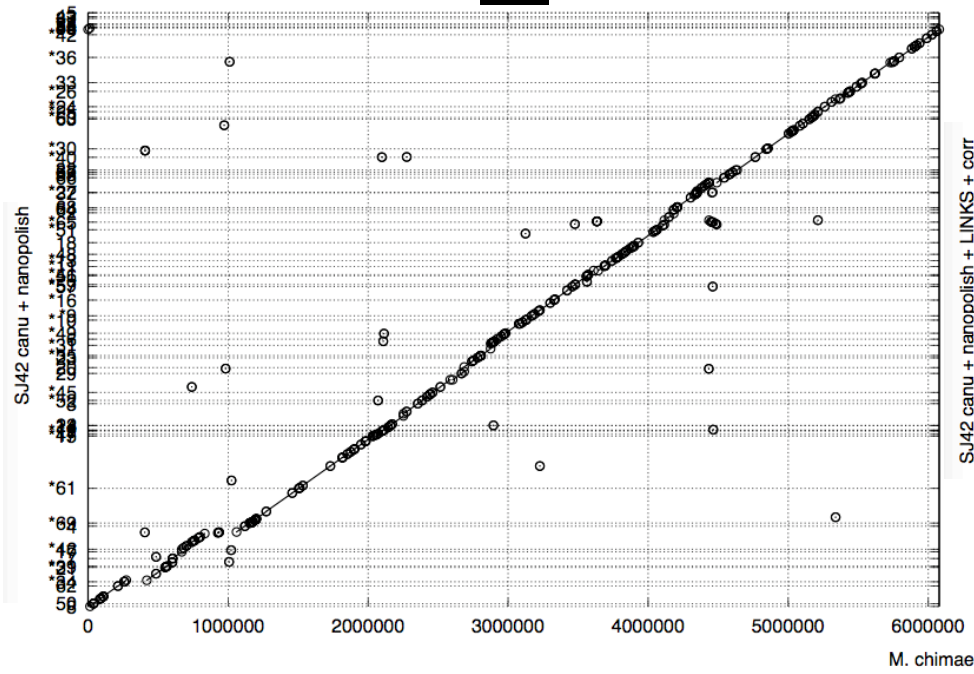
ONT-only Assembly

- canu + LINKS worked best

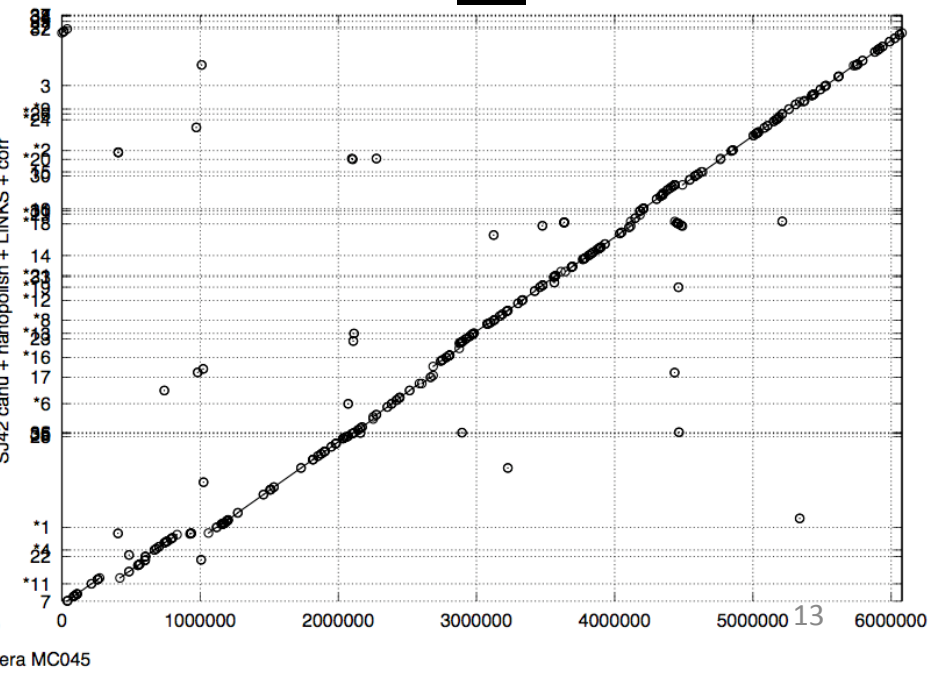
Assemblies courtesy:
Daniel MacMillan

n	n:500	n:N50	min	N80	N50	N20	E-size	max	sum	name
96	96	19	2855	43142	85571	167050	106013	227674	5689525	SJ42canu1D.fa
96	96	19	2919	44092	87699	169915	107761	231185	5777431	SJ42canu1D_racon_nanopolish.fa **A**
41	41	7	2919	128913	211887	756357	335373	780567	5777431	SJ42canu1D_racon_nanopolish_links.fa **B**

A



B



M. chimaera MC045

ONT+MiSeq Hybrid Assembly

- Matched Illumina MiSeq PE250 reads (offsite)
- *Unicycler* assembler worked best

n	n:500	L50	min	N80	N50	N20	E-size	max	sum	
43	33	4	747	209932	570557	940706	579669	1099614	5944060	*unicycler* 1D-ONT+MiSeq
43	33	4	747	209973	570557	940761	579679	1099614	5944164	+ polishing
37	27	3	747	240706	640788	1693028	932964	1693028	5944164	+ scaffolding (*LINKS*)*

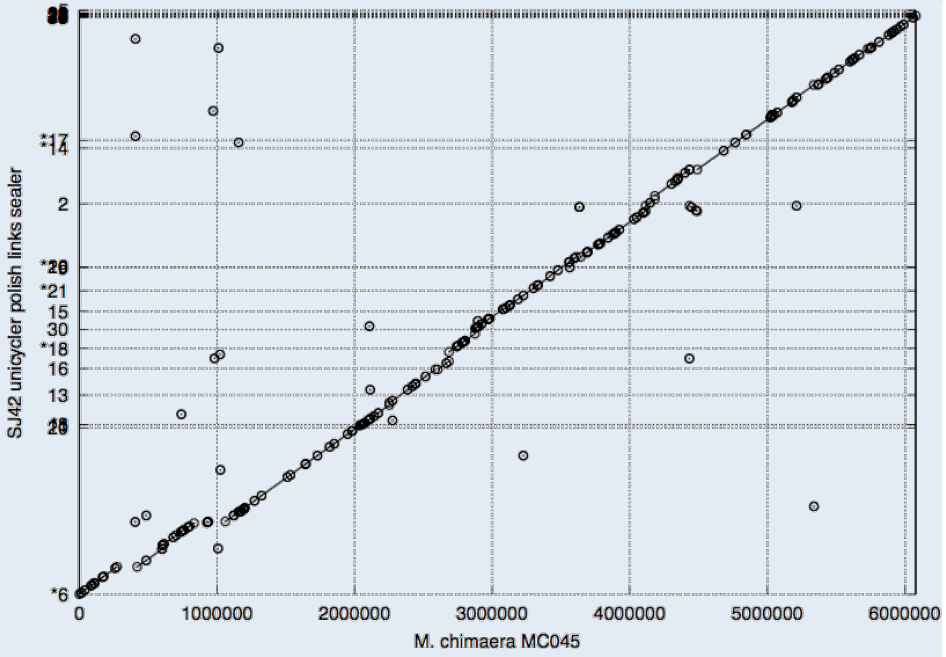
*Automated gap-filler *Sealer* closed 2 additional gaps (out of 6 predicted overlaps)
*Assembly aligns best to *M. chimaera* MC045 chromosome. 12 seqs >=500 bp possible plasmids

Assemblies courtesy: Daniel MacMillan

Final SJ_42 assembly

de novo

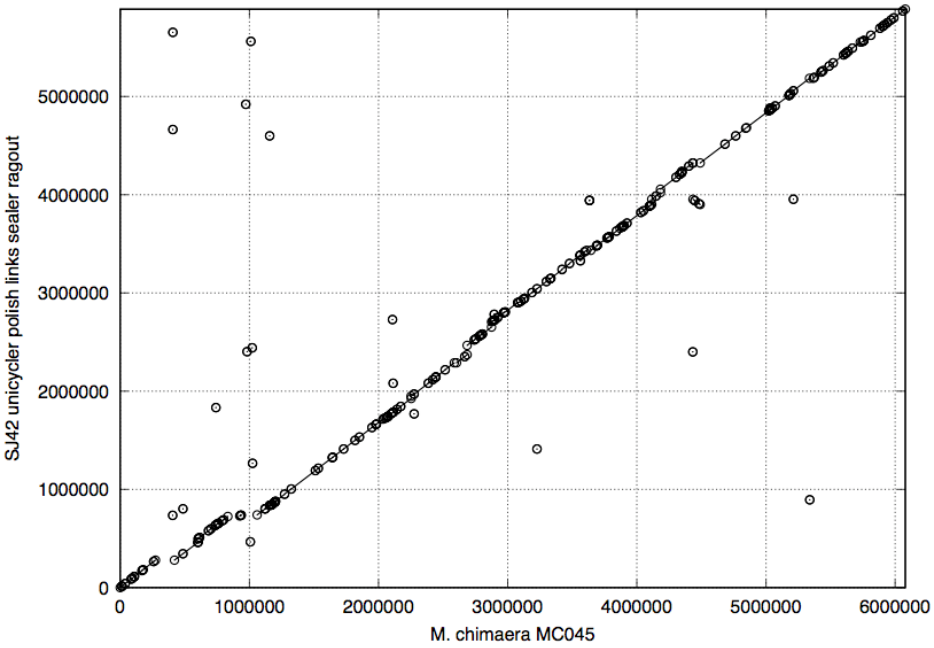
27 scaffolds (>=500 bp)
N50 : 640.8 kbp
Max : 1,693,028 bp
Reconstruction: 5,944,164 bp
No gaps, 4 predicted overlaps



Unicycler hybrid + polish > ONT + MiSeq
LINKS scaffolding > ONT
Sealer Gap closing > MiSeq

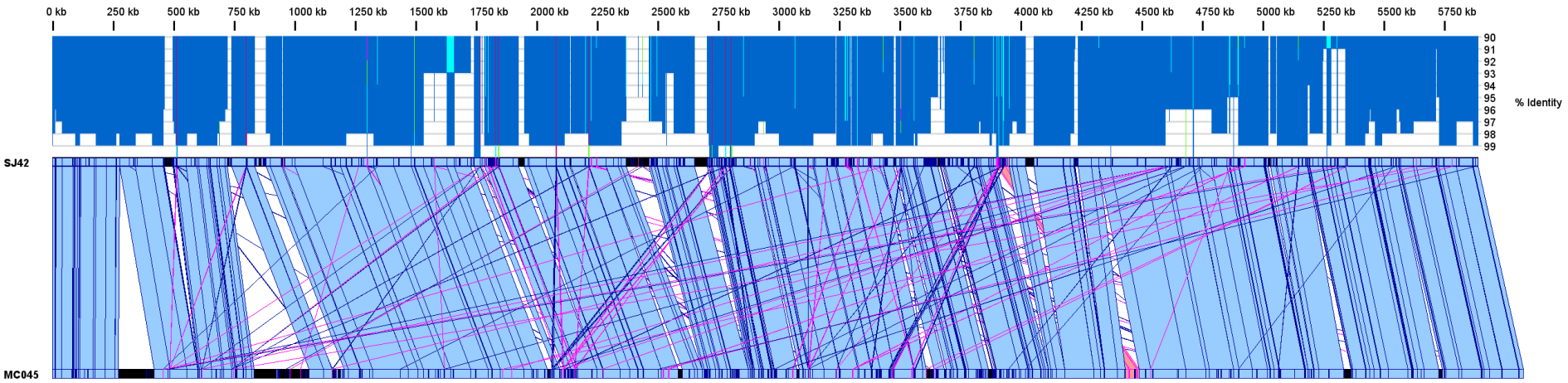
reference-based*

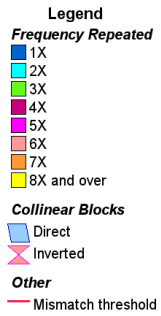
1 seq: 15 scaffolds shared with MC045 chr.
12 unplaced >= 500bp (incl. likely plasmids**) :
N50 : 16.6 kbp
Max : 33,560 bp
Reconstruction : 76,013



*applied to assembly on the left, genome structure may be different than that of MC045
**some seq may be part of the SJ42 chromosome

SJ42 vs. MC045





Mismatch threshold 10
Minimum Block Length=10
Scale=1:2500
Wed Mar 15 11:54:50 PDT 2017
XMatchView v0.2 :: rwarren@bcgsc.ca

Reference: *M. chimaera* MC045
size: 6,078,679

from SJ42 alignments:
coverage : 91.04 %
avg seq. identity over MC045 chromosome : 94.17 %

Summary SJ_42

- Reads are short (10-fold shorter)
- Base yield and accuracy on low end
- R9 run throughput (#sequences) on par with OICR's

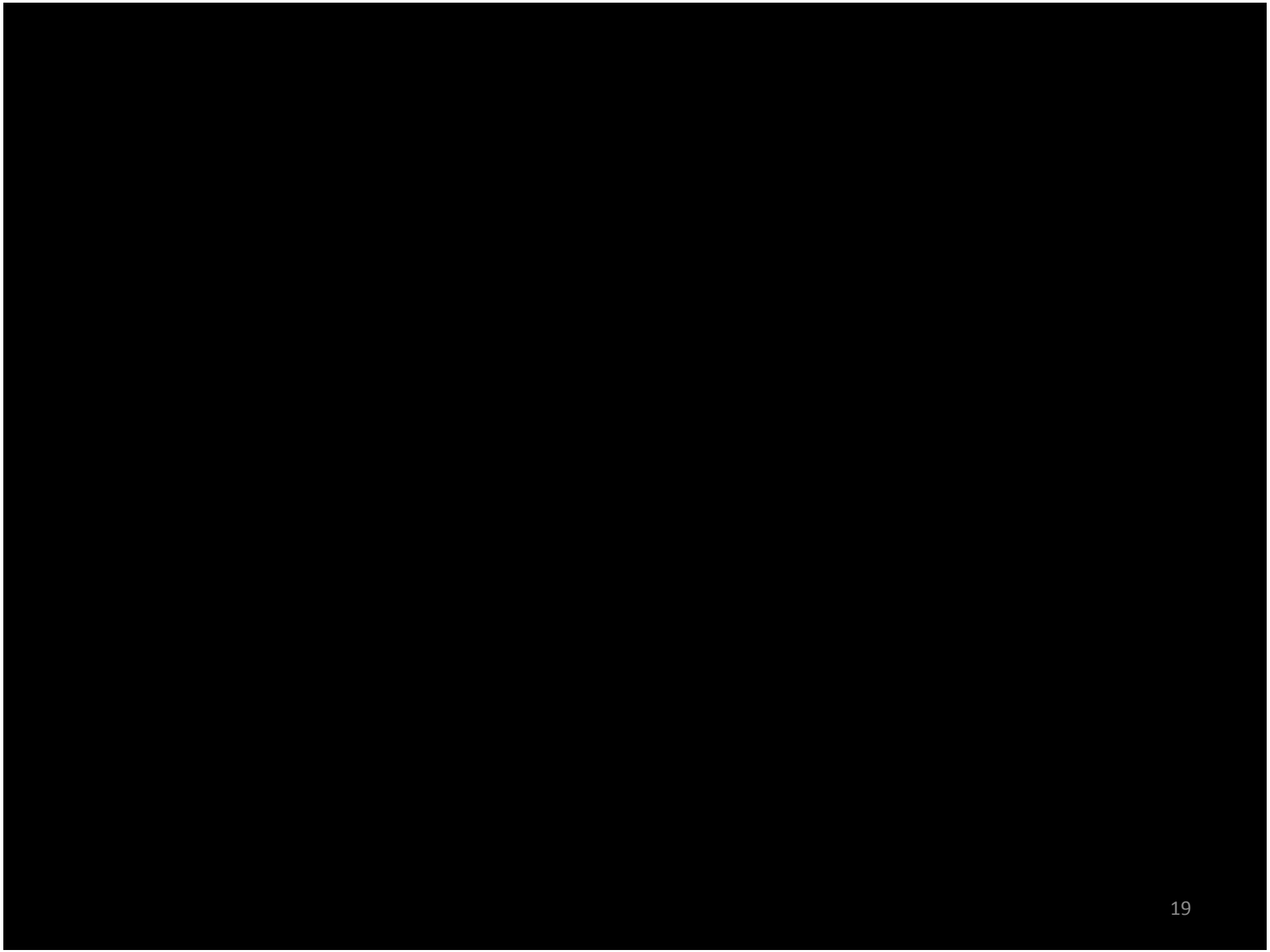
data

- Despite data limitations, *de novo* assembly worked well
  - ONT only: 41 scaffolds, N50 = 200 kbp
  - ONT+MiSeq: 27 scaffolds, N50 = 640 kbp
  - Both drafts are structurally similar to MC045 (130-210 kbp shorter)
  - MC045 and SJ_42 diverge, both harbor unique sequences
  - Seq. divergence from MC045 est. ~5%, ~91% coverage

application

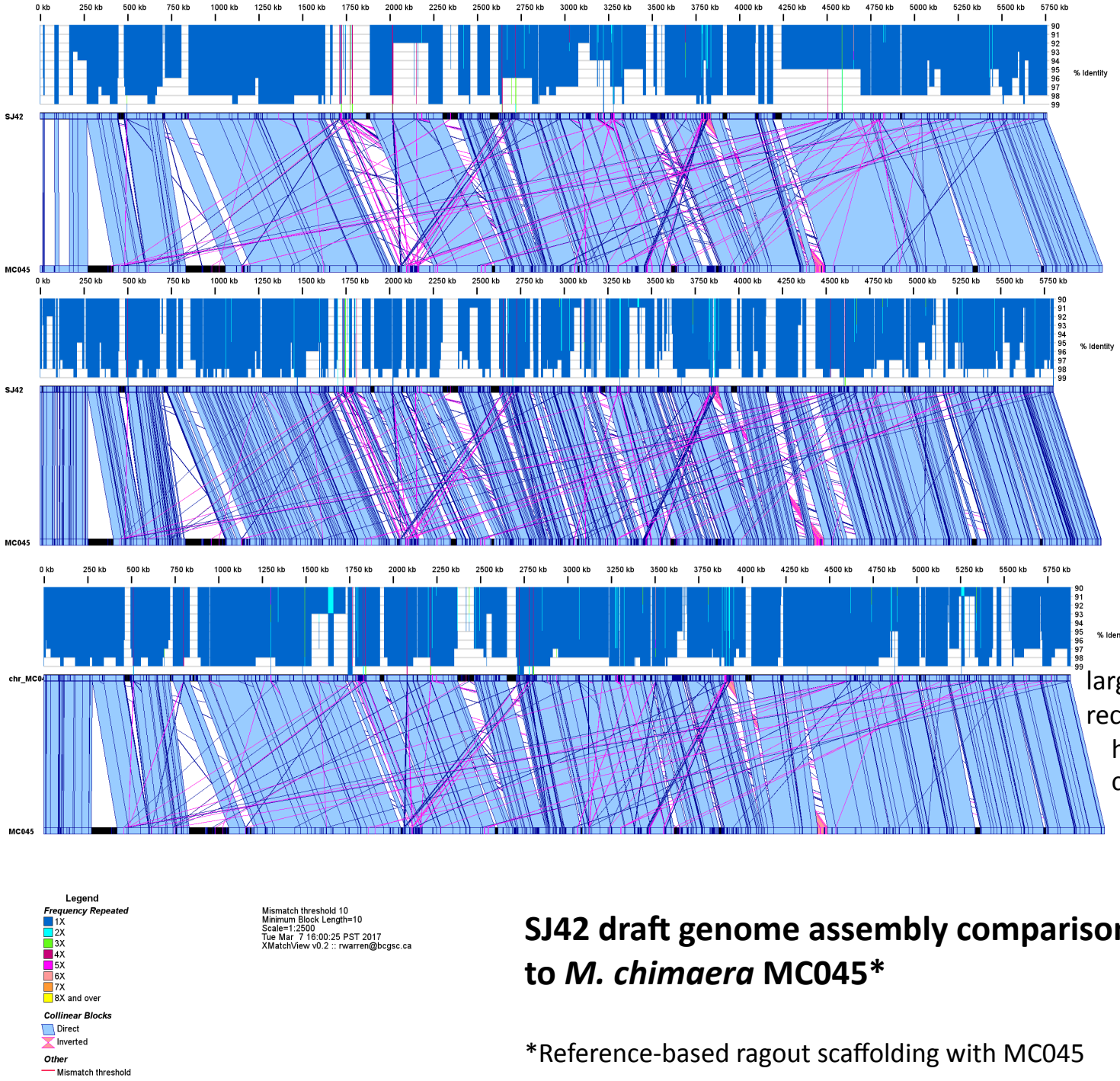
Long read technologies

- Depends on application
  - PacBio
  - ONT
  - Moleculo TruSeq (read cloud)
    - expensive for mammalian genome
    - reads still on “short” side (5-10 kbp)
  - 10XG Chromium
    - long molecule, sub 1X coverage
    - need radical bioinformatics tech development



Illumina MiSeq ABySS assemblies

n	n:500	L50	min	N80	N50	N20	E-size	max	sum	name
416	236	39	503	24342	52323	82829	57901	192012	5906821	k60/sj42_k60-scaffolds.fa
313	213	30	508	27221	61926	107595	74706	203627	5900572	k70/sj42_k70-scaffolds.fa
288	212	32	512	27350	59624	100266	68957	192133	5901912	k80/sj42_k80-scaffolds.fa
269	216	34	518	26553	55241	101093	65327	192108	5905016	k90/sj42_k90-scaffolds.fa
262	222	37	538	24422	46704	99692	59984	167121	5919365	k100/sj42_k100-scaffolds.fa
258	219	37	535	23666	48631	93023	60199	188792	5891363	k110/sj42_k110-scaffolds.fa
264	225	39	535	22729	49033	82780	58361	188763	5923126	k120/sj42_k120-scaffolds.fa
286	249	45	502	19029	45749	74909	49205	131867	5930092	k130/sj42_k130-scaffolds.fa
322	286	50	528	16222	37707	71816	44001	119887	5924043	k140/sj42_k140-scaffolds.fa
392	340	60	535	14542	28686	63965	38285	119662	5928298	k150/sj42_k150-scaffolds.fa
8	7	2	509	815	1118	3883	2214	3883	8619	k240/sj42_k240-scaffolds.fa



SJ42 draft genome assembly comparison to *M. chimaera* MC045*

*Reference-based ragout scaffolding with MC045