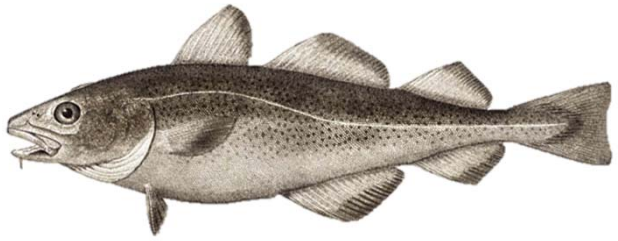
De novo assembly of teleost fishes using PacBio sequencing data



Ole K. Tørresen

UiO : Centre for Ecological and Evolutionary Synthesis
University of Oslo

Acknowledgments

CEES
Centre for Ecological and Evolutionary Synthesis

Kjetill S. Jakobsen
Sissel Jentoft
Lex Nederbragt

NSC
NORWEGIAN SEQUENCING CENTRE

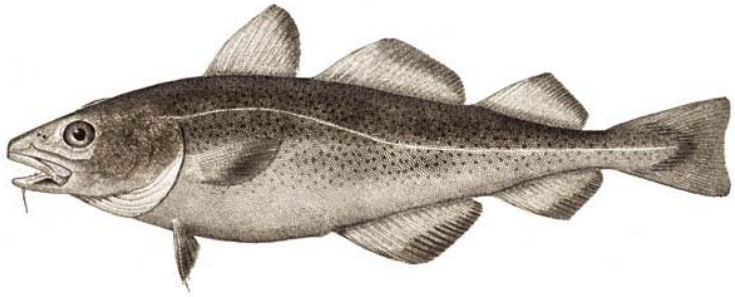
Sequencing team NSC

Bastiaan Star
Marine S. O. Brieuc
Monica Solbakken
Michael Matschiner
Martin Malmstrøm
Bastiaan Star
William Brynildsen
and others

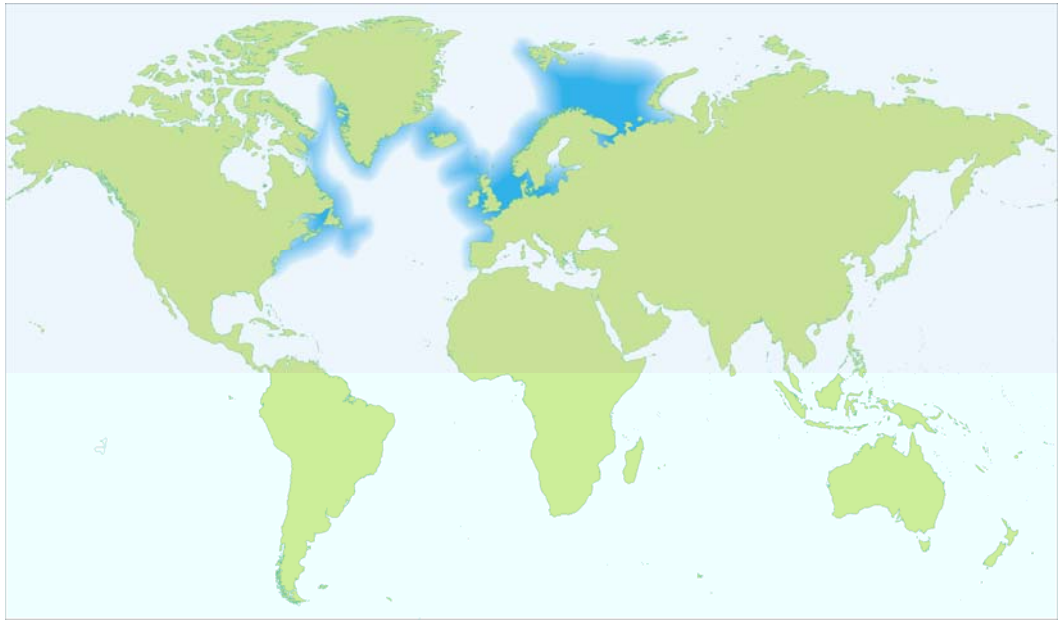
University of Oslo

Atlantic cod

- 0.6-1.2 m long
- 40 kg
- Live up to 25 years
- 180 000 tons exported for €800 million so far this year

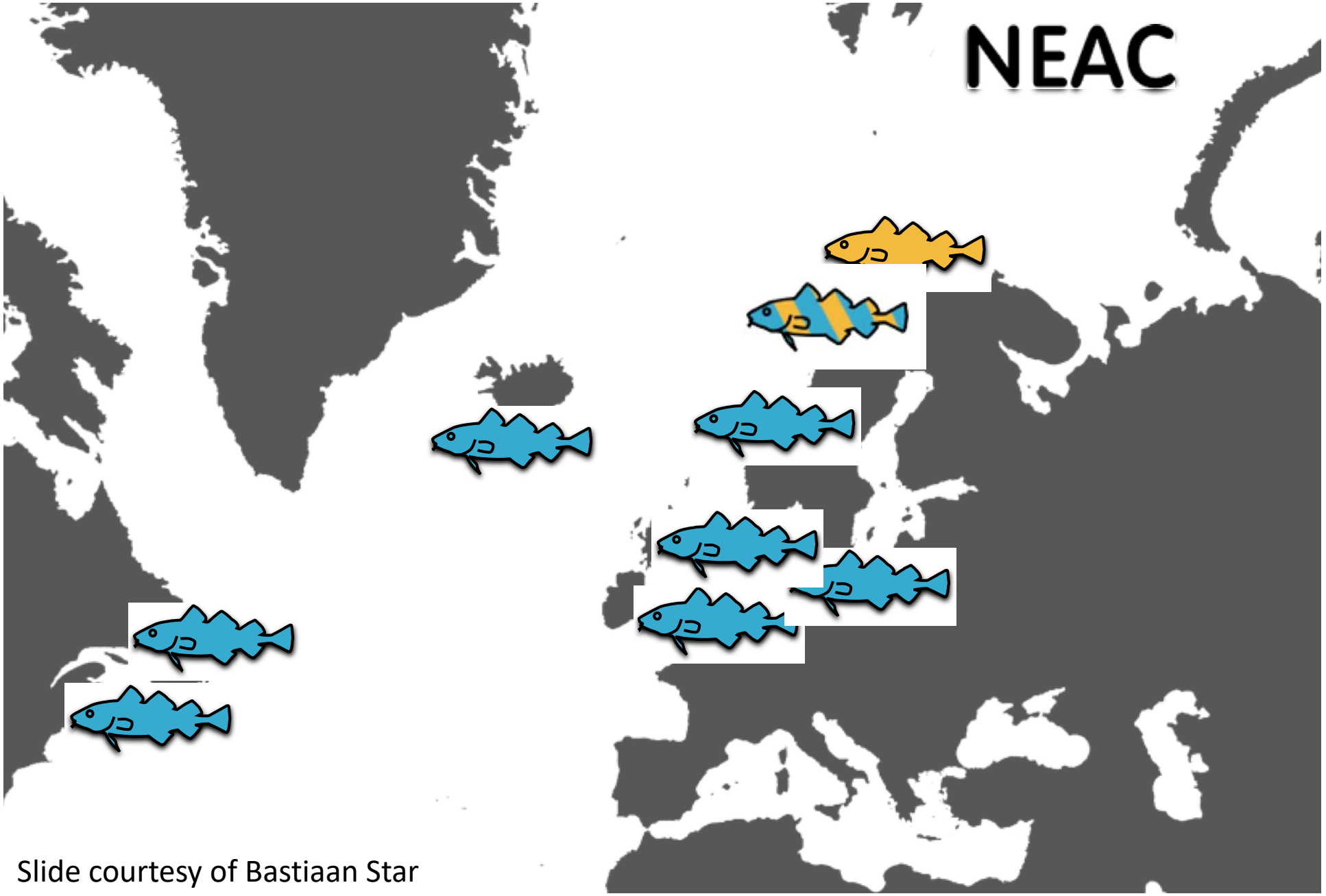


Public domain

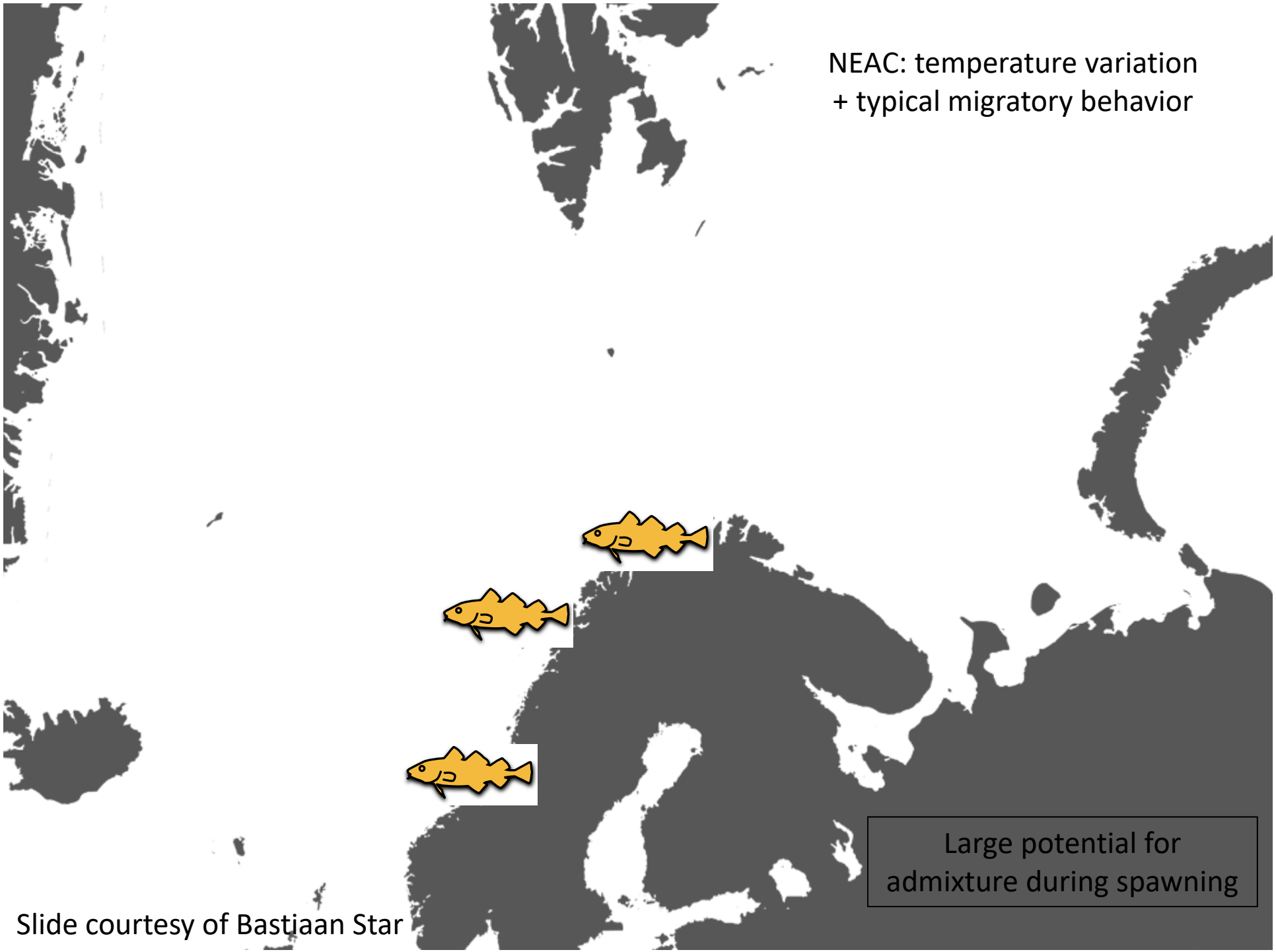


By Aotearoa, CC BY-SA 3.0

http://seafood.no/markedsinnsikt/nokkeltall/ and Wikipedia



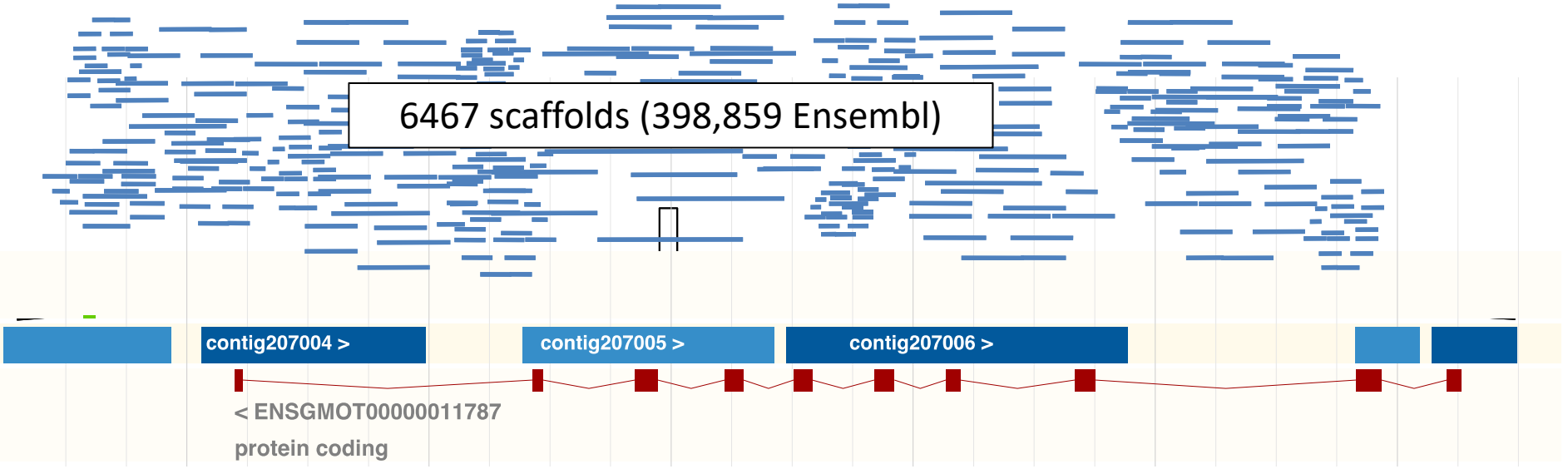
Slide courtesy of Bastiaan Star



Slide courtesy of Bastiaan Star

Fishing in Lofoten

Cod genome assembly phase 1



35% gap bases

Phase 1 results (Ensembl version)
Scaffold N50 0.69 Mbp (0.14)
Contig N50 3.9/2.8 kbp (2.3)

Slide courtesy of Lex Nederbragt

LETTER

doi:10.1038/nature10342

The genome sequence of Atlantic cod reveals a unique immune system

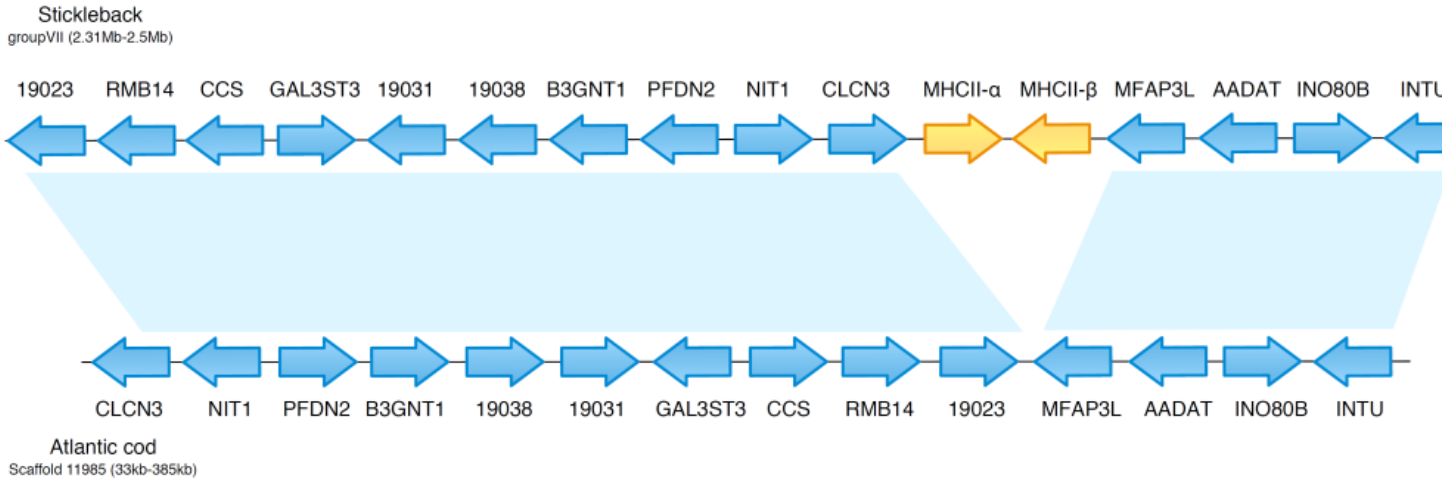
Bastiaan Star[1], Alexander J. Nederbragt[1], Sissel Jentoft[1], Unni Grimholt[1], Martin Malmstrøm[1], Tone F. Gregers[2], Trine B. Rounge[1], Jonas Paulsen[1,3], Monica H. Solbakken[1], Animesh Sharma[4], Ola F. Wetten[5,6], Anders Lanzén[7,8], Roger Winer[9], James Knight[9], Jan-Hinnerk Vogel[10], Bronwen Aken[10], Øivind Andersen[11], Karin Lagesen[1], Ave Tooming-Klunderud[1], Rolf B. Edvardsen[12], Kirubakaran G. Tina[1,13], Mari Espelund[1], Chirag Nepal[4,8], Christopher Previti[8], Bård Ove Karlsen[14], Truls Moum[14], Morten Skage[1], Paul R Berg[1] Tor Gjøen[15] Heiner Kuhl[16] Jim Thorsen[17] Ketil Malde[12] Richard Reinhardt[16] Lei Du[9] Steinar D Johansen[14,18]

The findings

Adaptive immunity

Missing

- major histocompatibility complex (MHC) class II
- CD4
- invariant chain

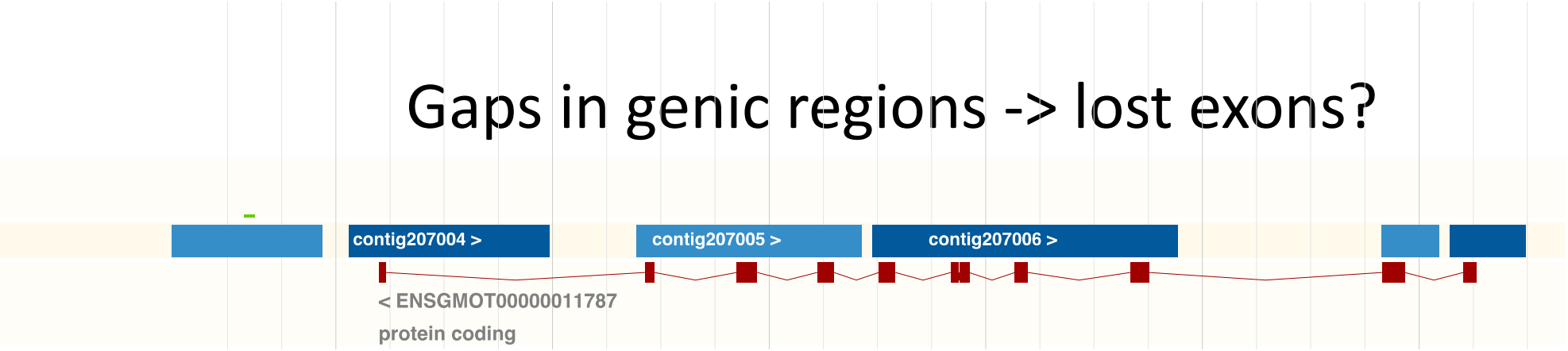


Putative MHCII region in Atlantic cod vs. Stickleback

Slide courtesy of Monica Solbakken

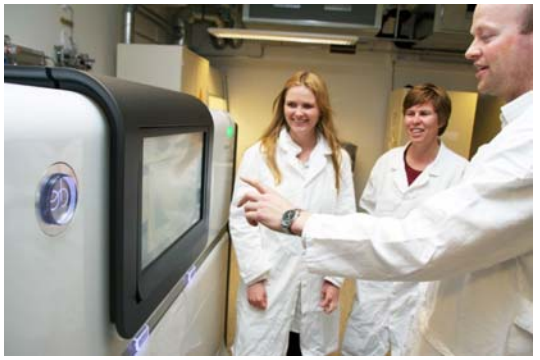
Consequences

Gaps in genic regions -> lost exons?

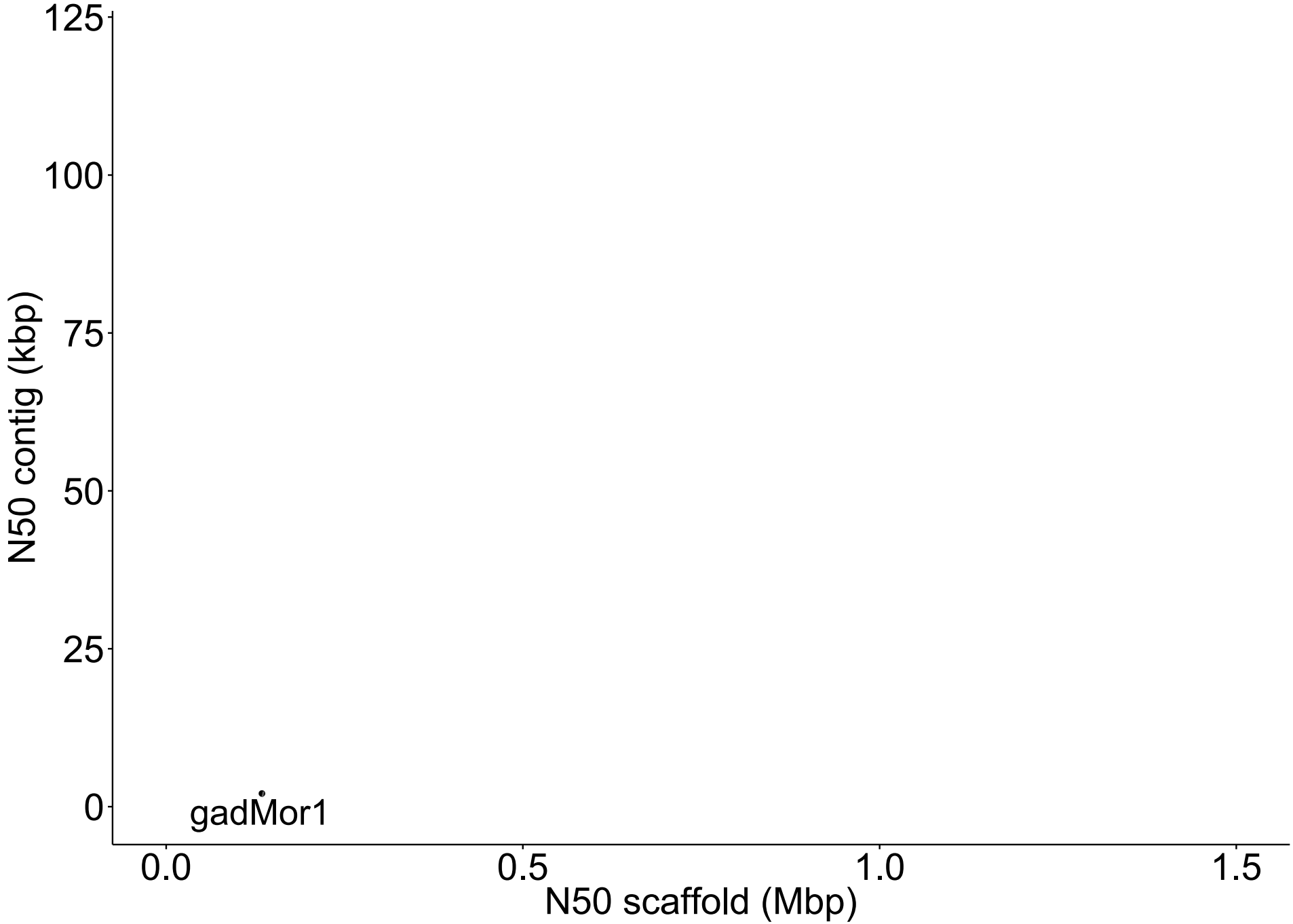


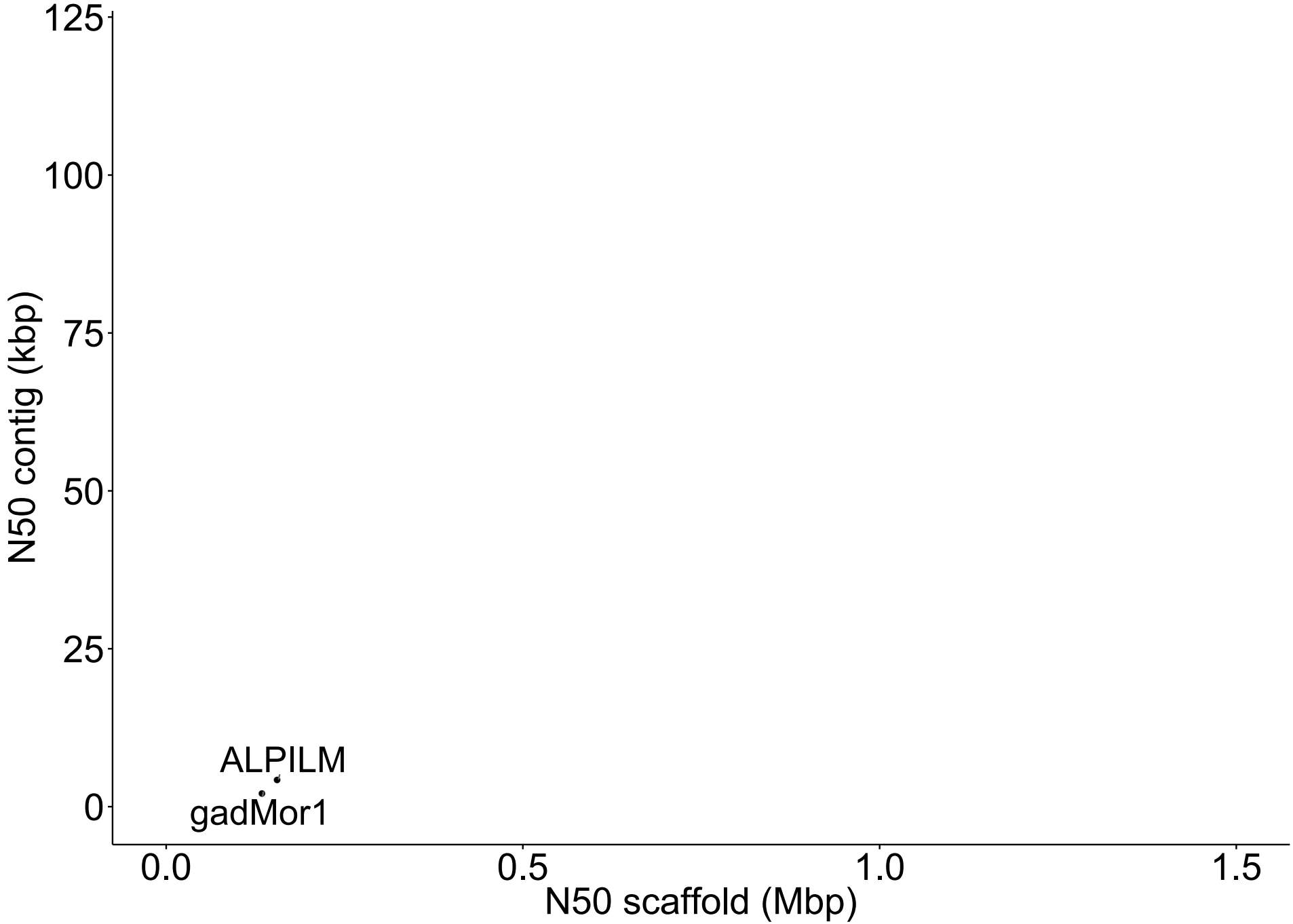
Slide courtesy of Lex Nederbragt

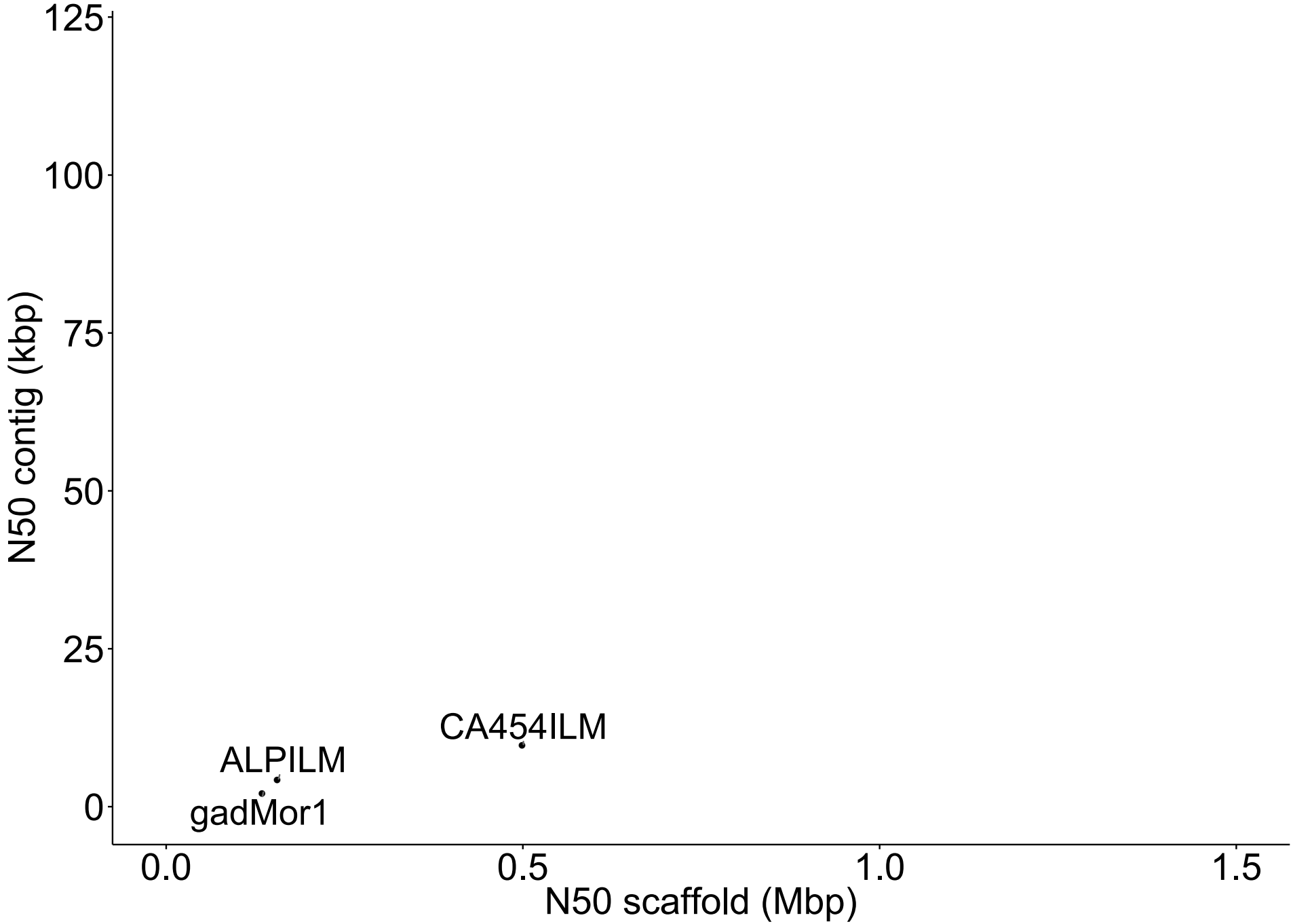
Cod genome assembly phase 2: Improvement

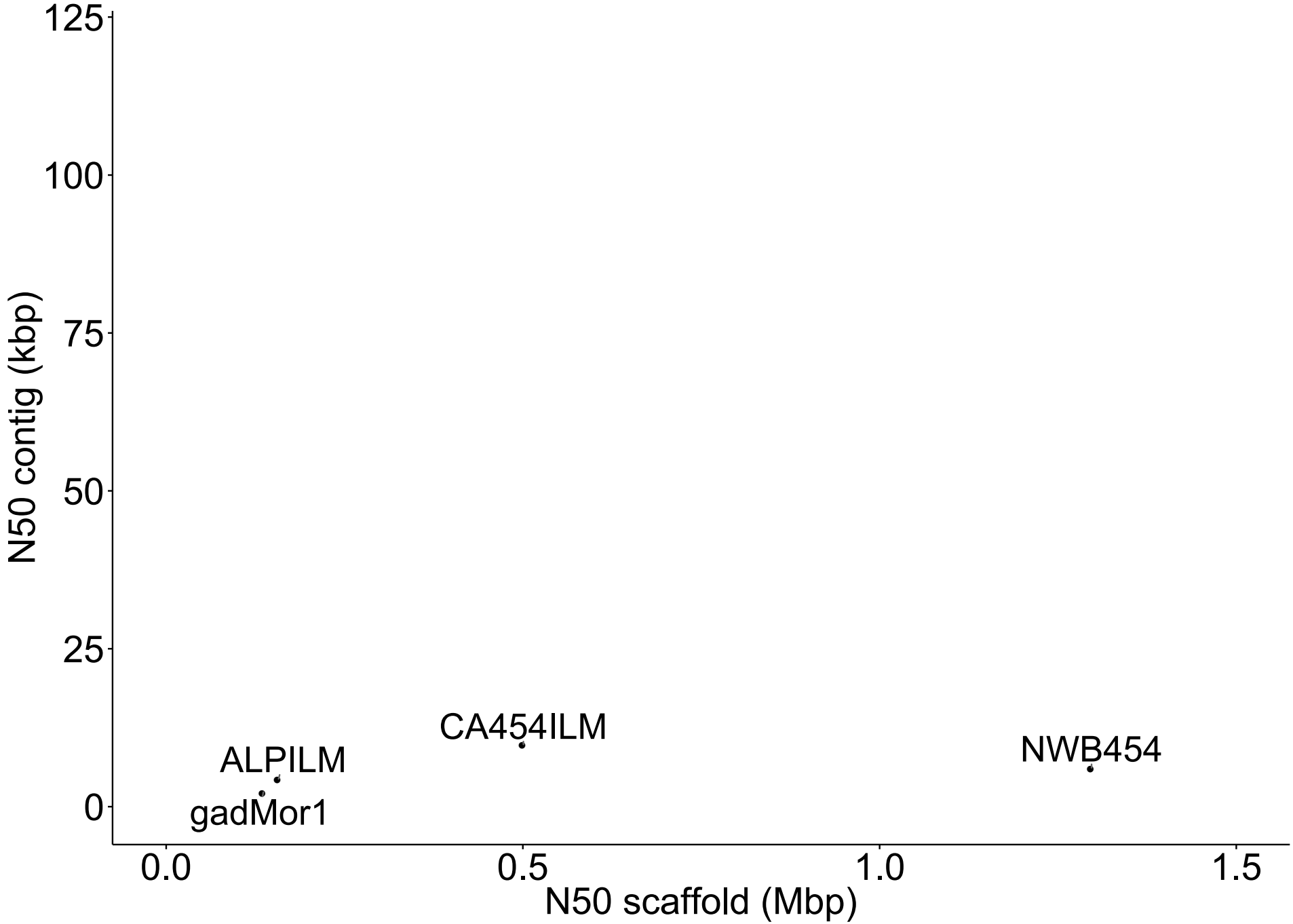
Existing data	**New data**	**New data**
		
454 sequencing	Illumina sequencing	PacBio sequencing
40x	300x	20x

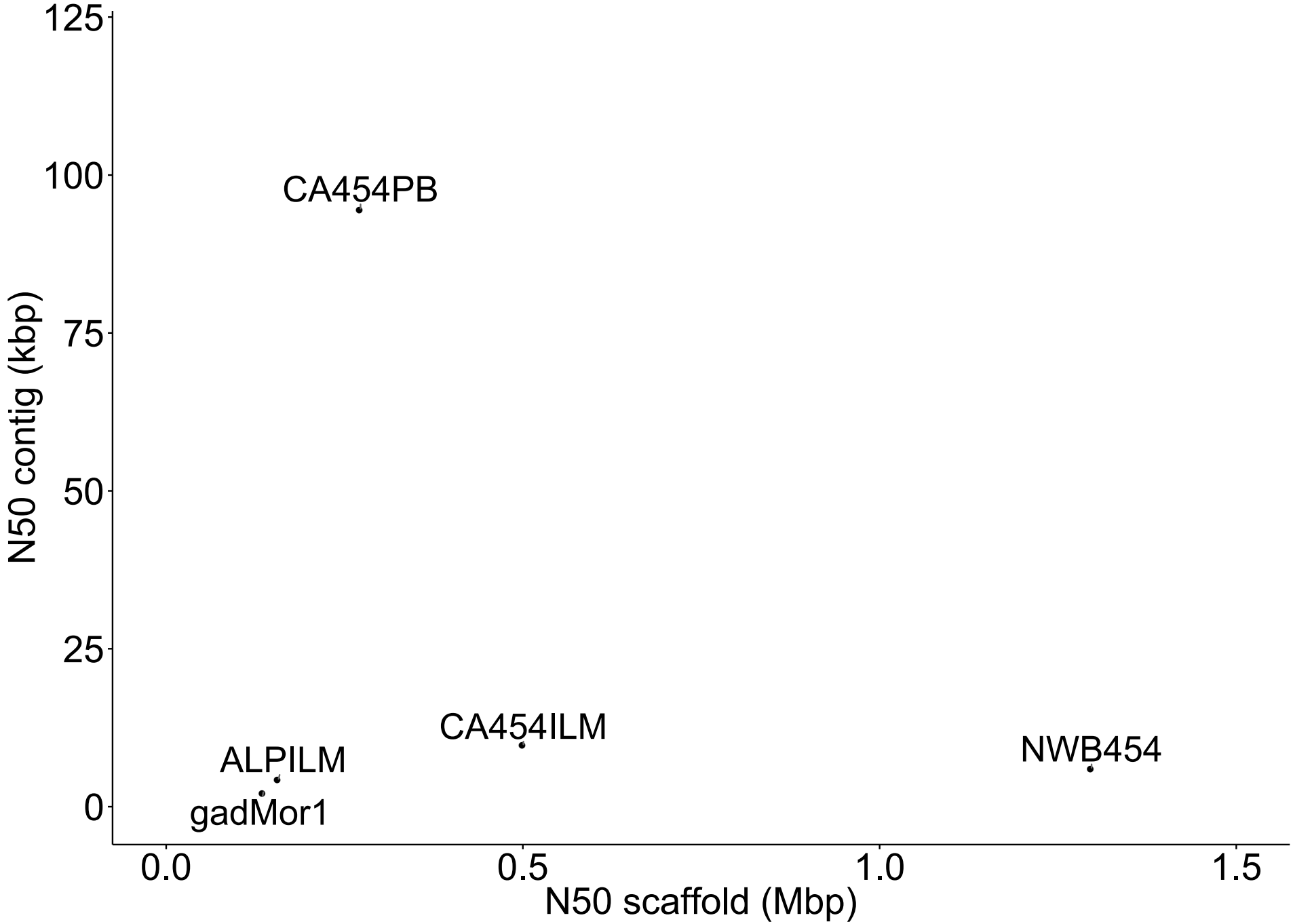
Slide courtesy of Lex Nederbragt

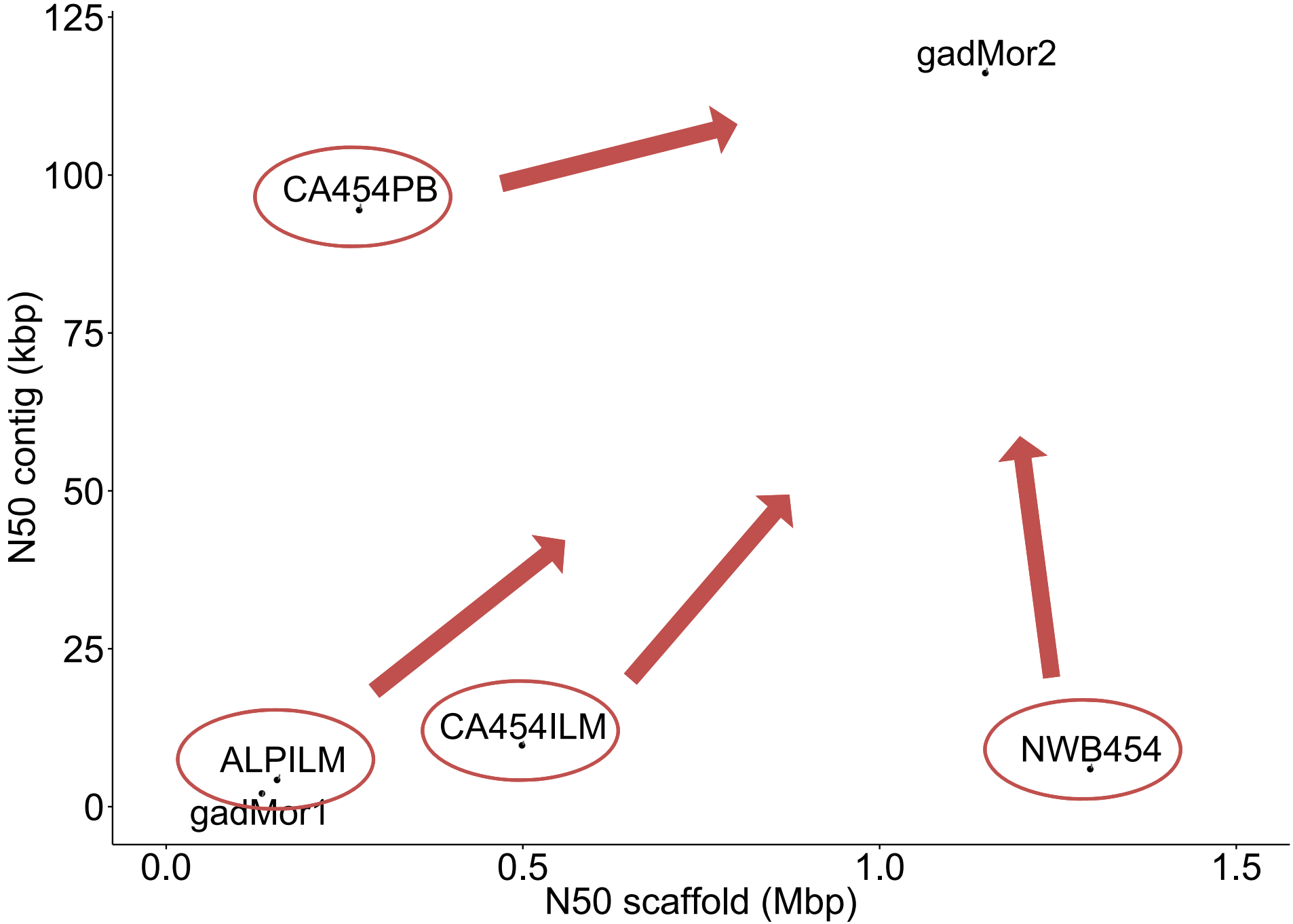

125
100
75
50
25
0
N50 contig (kbp)
gadMor1
0.0
0.5
1.0
1.5
N50 scaffold (Mbp)


125
100
75
50
25
0
N50 contig (kbp)
ALPILM
gadMor1
0.0
0.5
1.0
1.5
N50 scaffold (Mbp)


125
100
75
50
25
0
N50 contig (kbp)
CA454ILM
ALPILM
gadMor1
0.0
0.5
1.0
1.5
N50 scaffold (Mbp)


125
100
75
50
25
0
N50 contig (kbp)
ALPILM
CA454ILM
NWB454
gadMor1
0.0
0.5
1.0
1.5
N50 scaffold (Mbp)

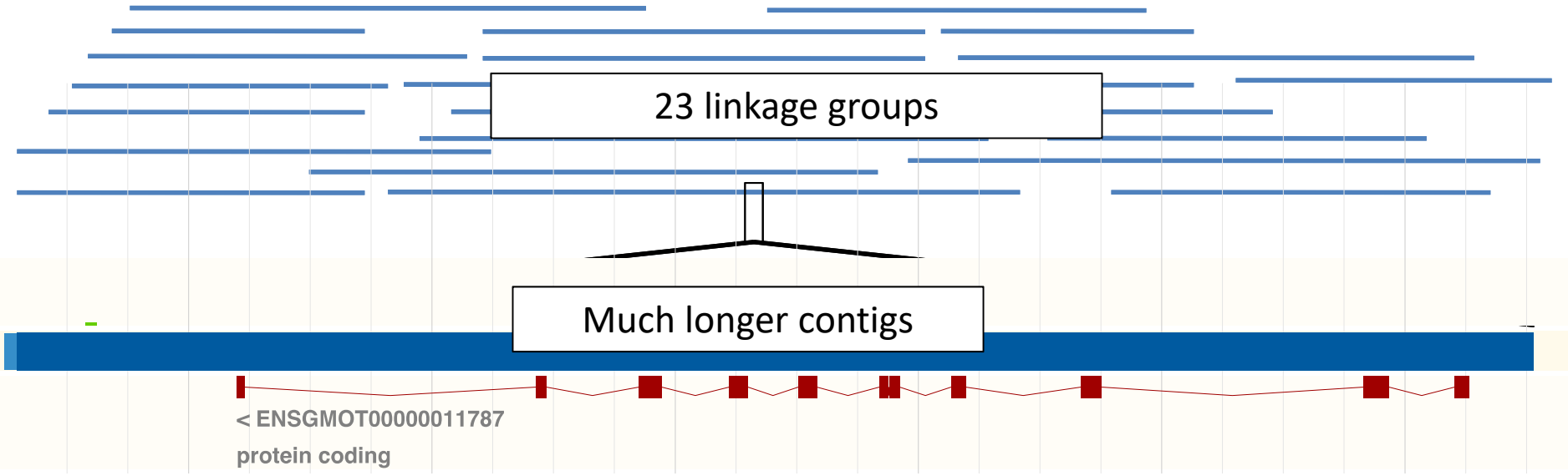

125
100
75
50
25
0
N50 contig (kbp)
CA454PB
CA454ILM
ALPILM
gadMor1
NWB454
0.0
0.5
1.0
1.5
N50 scaffold (Mbp)


125
100
75
50
25
0
N50 contig (kbp)
gadMor2
CA454PB
ALPILM
gadMor1
CA454ILM
NWB454
0.0
0.5
1.0
1.5
N50 scaffold (Mbp)

Combining long and short reads

- Attempts at correction fragmented PacBio reads
- Uncorrected PacBio assembled together with 454/Illumina reads
- First published genome assembly using uncorrected PacBio reads

What we got

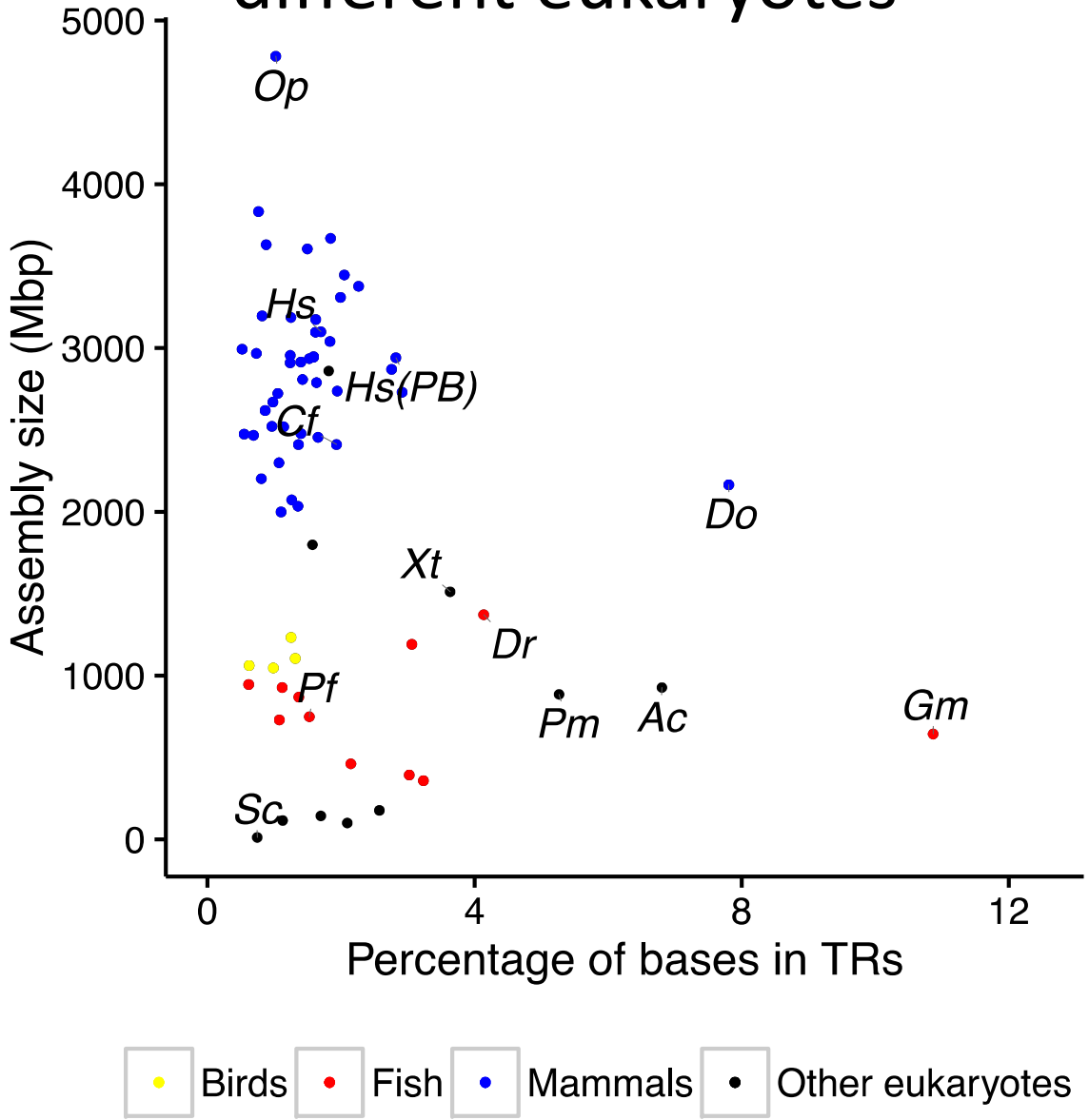


1.7 % gap bases

Phase 2 results
Scaffold N50 1.15 Mbp
Contig N50 116 kbp

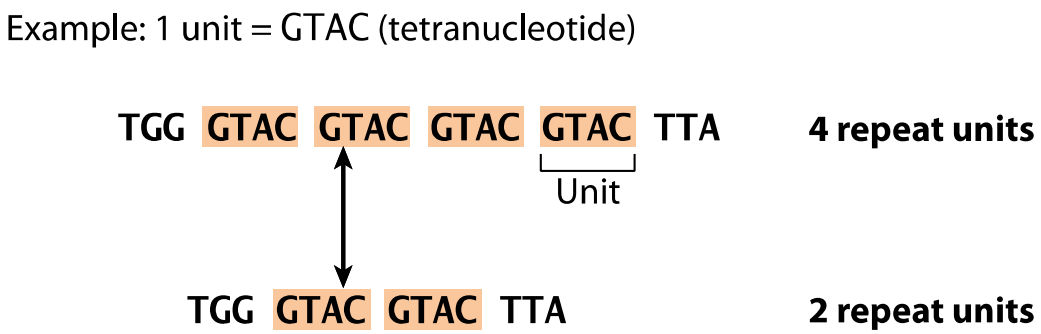
Slide courtesy of Lex Nederbragt

Amount of (short) tandem repeats in different eukaryotes

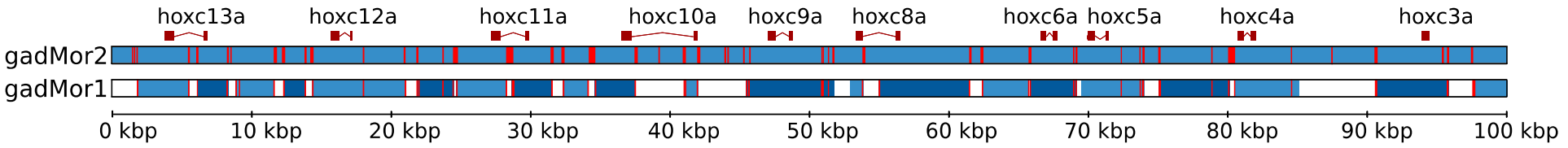


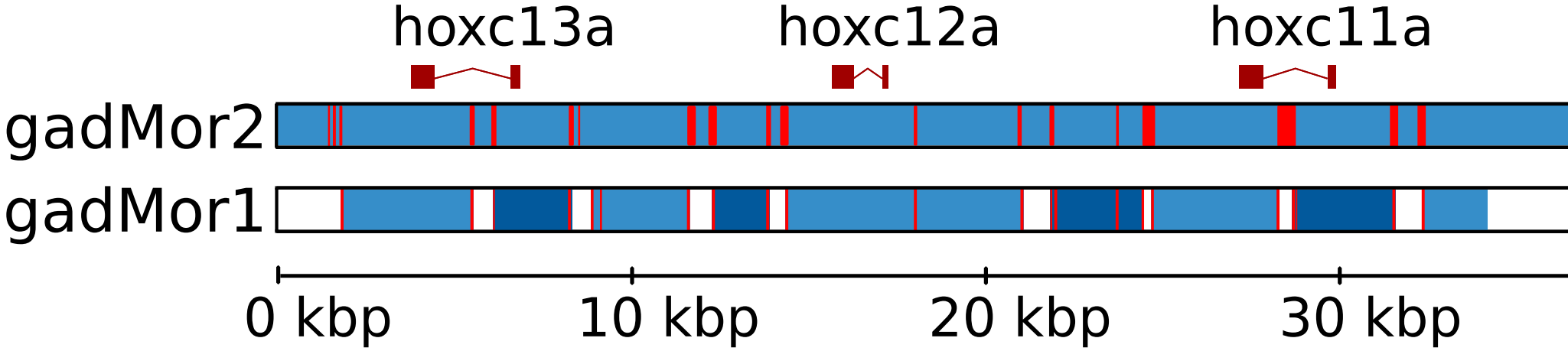
Short tandem repeats - microsatellites

- Tandem repeated short patterns
- ACTACTACTACTACTACT (ACT*6)
- Highly mutable

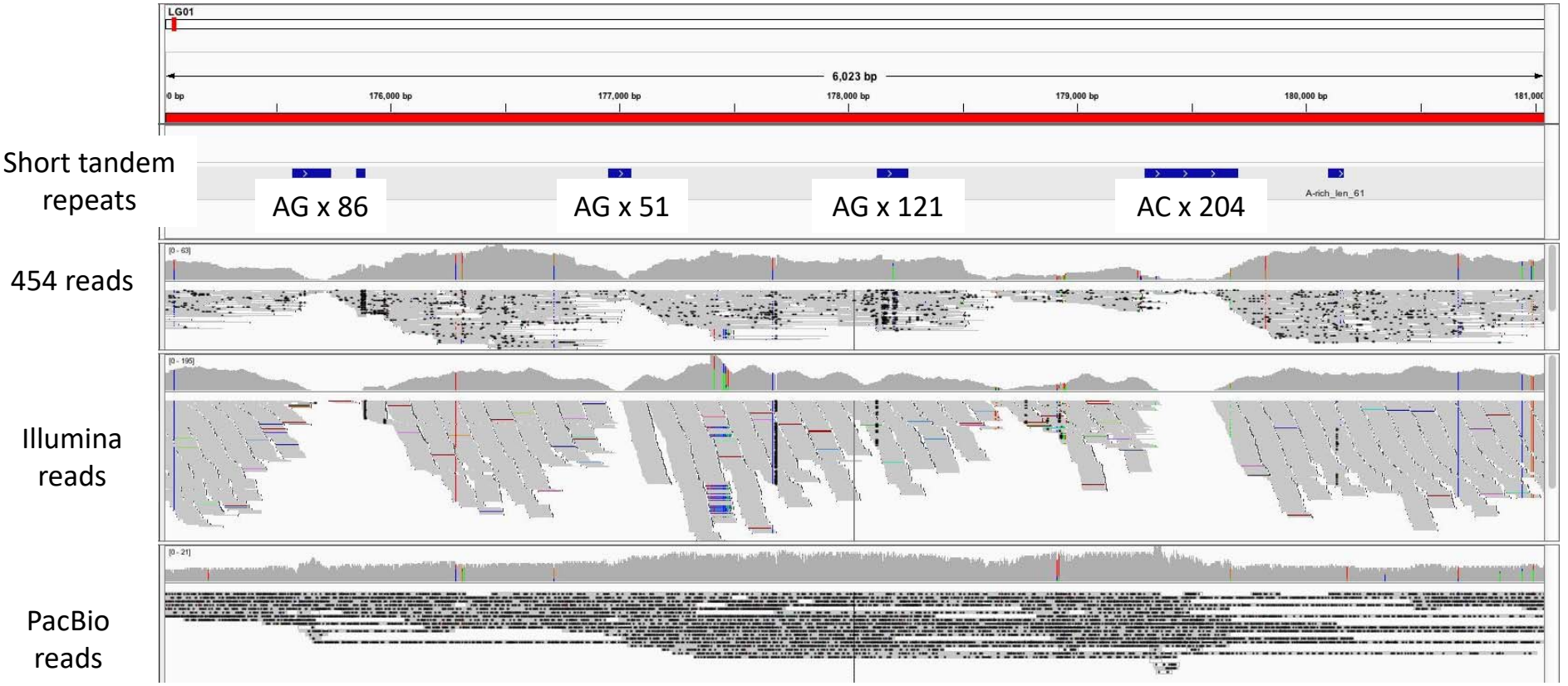


Comparison of gadMor1 and gadMor2 in the hoxc cluster





Exploring the assembly



172 bp 102 bp 242 bp 408 bp

Slide courtesy of Lex Nederbragt

The haddock

- 14 million years since last common ancestor with Atlantic cod
- Also a commercially important species
- Does it also contain substantial amount of STRs?

Sequencing and assembly strategy

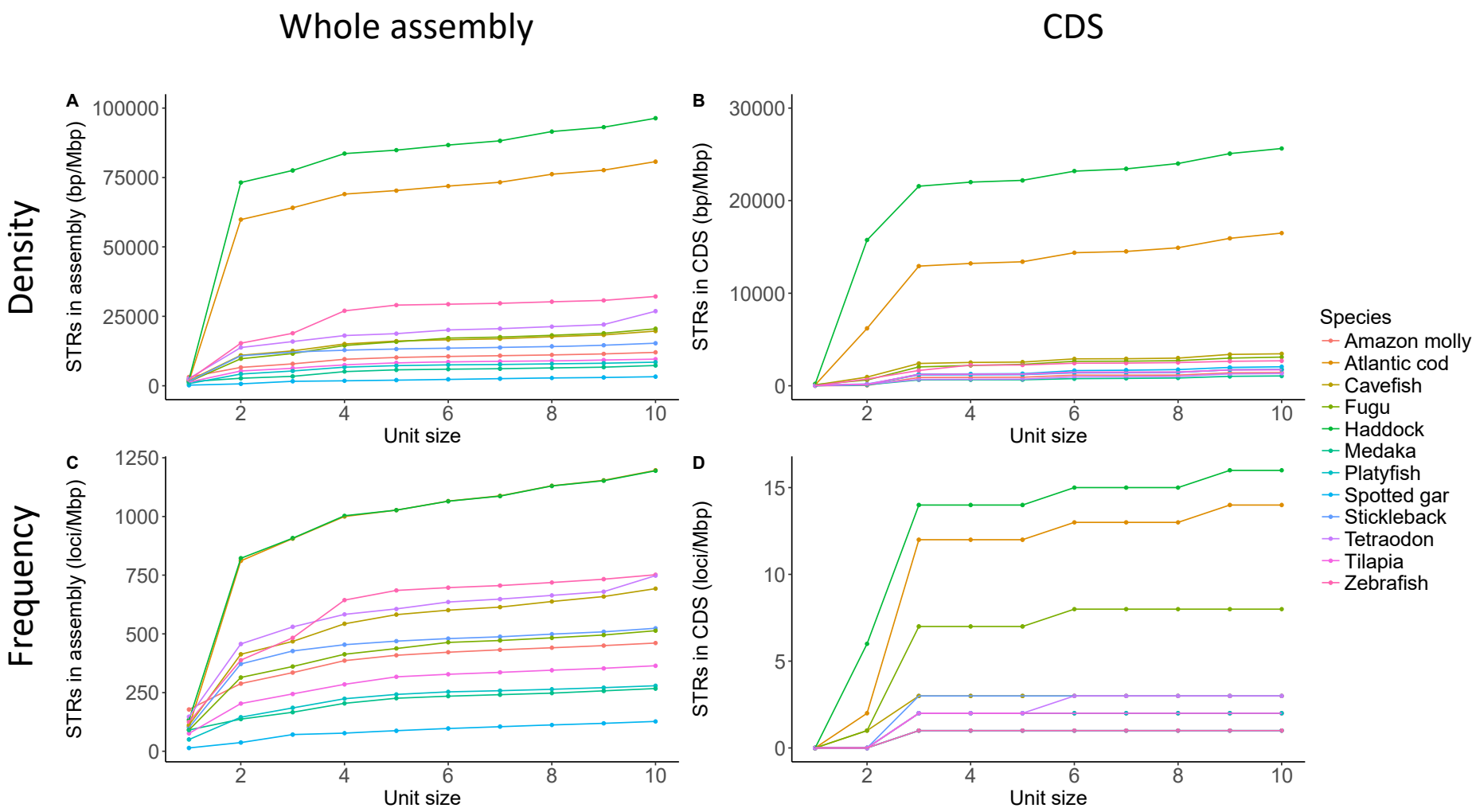
- 25X coverage PacBio P6C4
- 150X Illumina 150 bp paired reads
- Assembled together using Celera Assembler, uncorrected, but trimmed PacBio reads

Assembly statistics of haddock

	melAeg	GM_CA454PB	gadMor2
Length assembly (Mbp)	653	681	644
N50 scaffold (kbp)	209	272	1,150
N50 contig (kbp)	78	95	116
CEGMA complete (% of 458 genes)	439 (96 %)	431 (94 %)	435 (95 %)
BUSCO single	4,041 (88 %)[1]	3,819 (83 %)[1]	4,160 (91 %)[1]
BUSCO duplicated	128 (2.8 %)[1]	117 (2.6 %)[1]	127 (2.8 %)[1]
BUSCO fragmented	203 (4.4 %)[1]	359 (7.8 %)[1]	139 (3.0 %)[1]
BUSCO missing	212 (4.6 %)[1]	289 (6.3 %)[1]	158 (3.4 %)[1]

[1]% of 4584 genes

Cumulative density and frequency of STRs



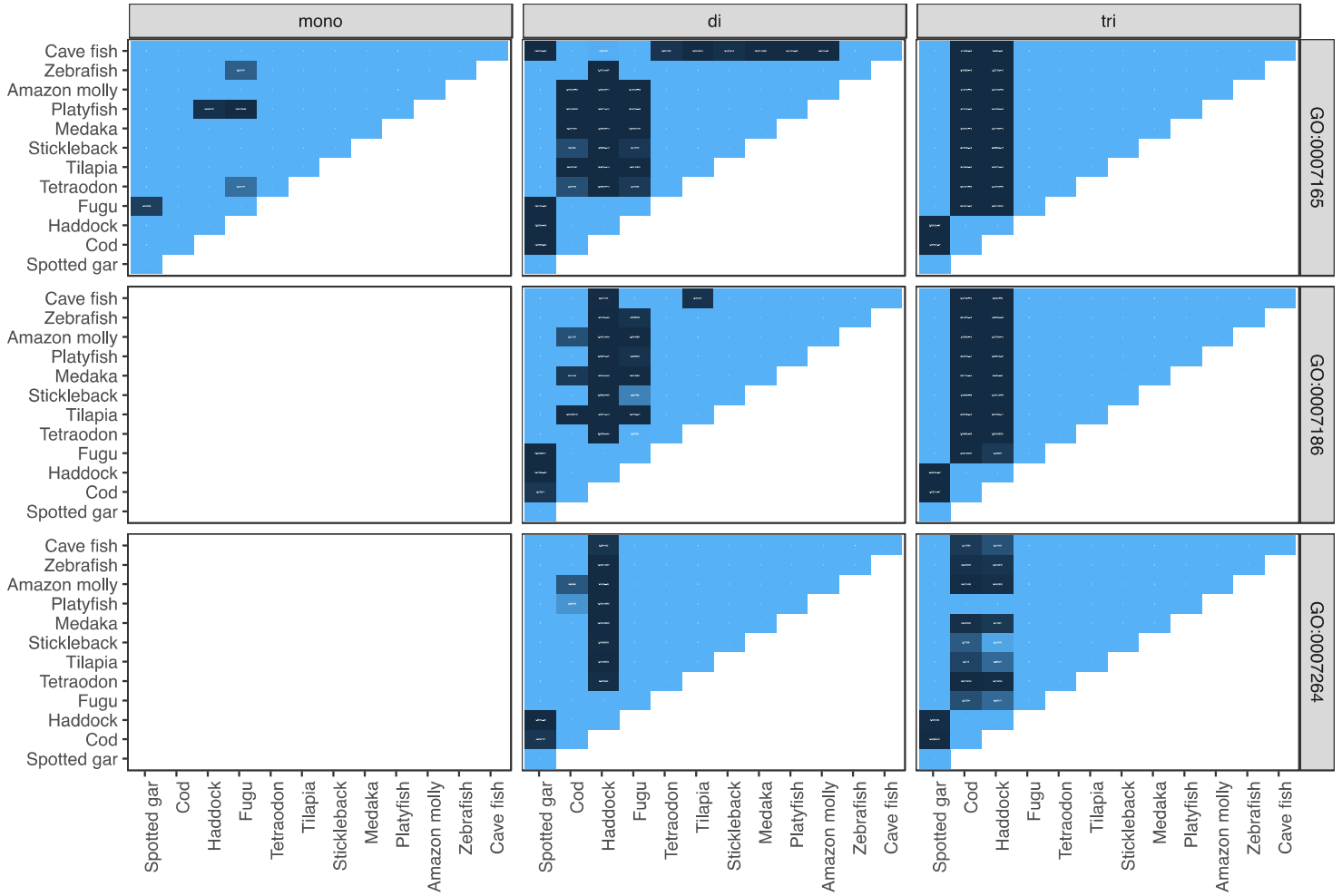
Which genes contain STRs?

- Grouped genes into functional groups (Gene Ontology)
- For each species pair and functional group (e.g. involved in signal transduction), Fisher's exact test:

	Species A	Species B	*Sum rows*
Genes with STRs	100	40	140
Total genes	150	160	310
Sum columns	250	200	450

- Corrected for multiple testing

Comparisons between species



signal transduction

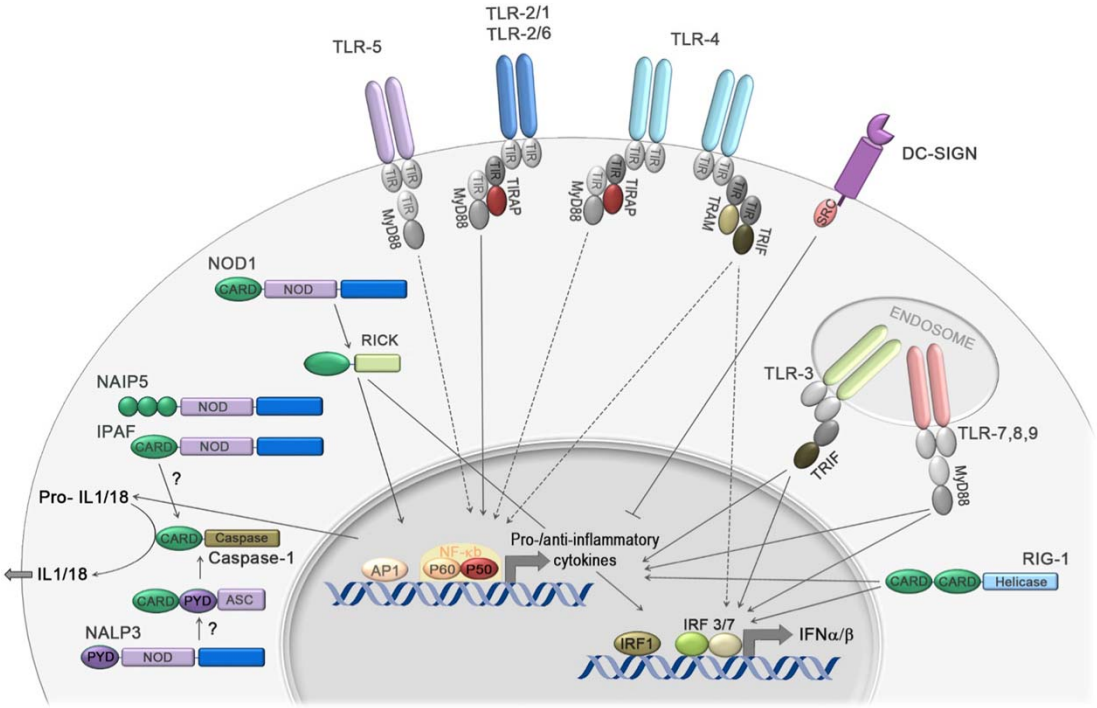
G-protein coupled receptor signaling pathway

small GTPase mediated signal transduction

Haddock and cod have significant more STRs in genes involved in signal transduction

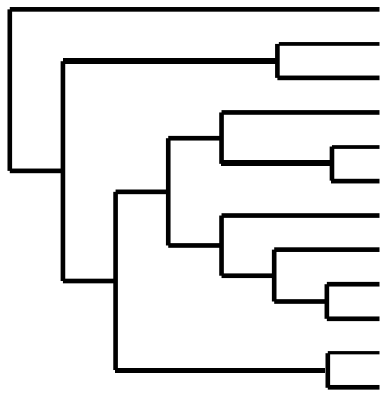
Innate immune-related genes

- TLRs, NLRs, RLRs and CLRs
- Sense molecules associated with damage or microbes (pathogens)
- Activates inflammatory and immune response



Müller et al 2011

NLR proteins are expanded in teleosts

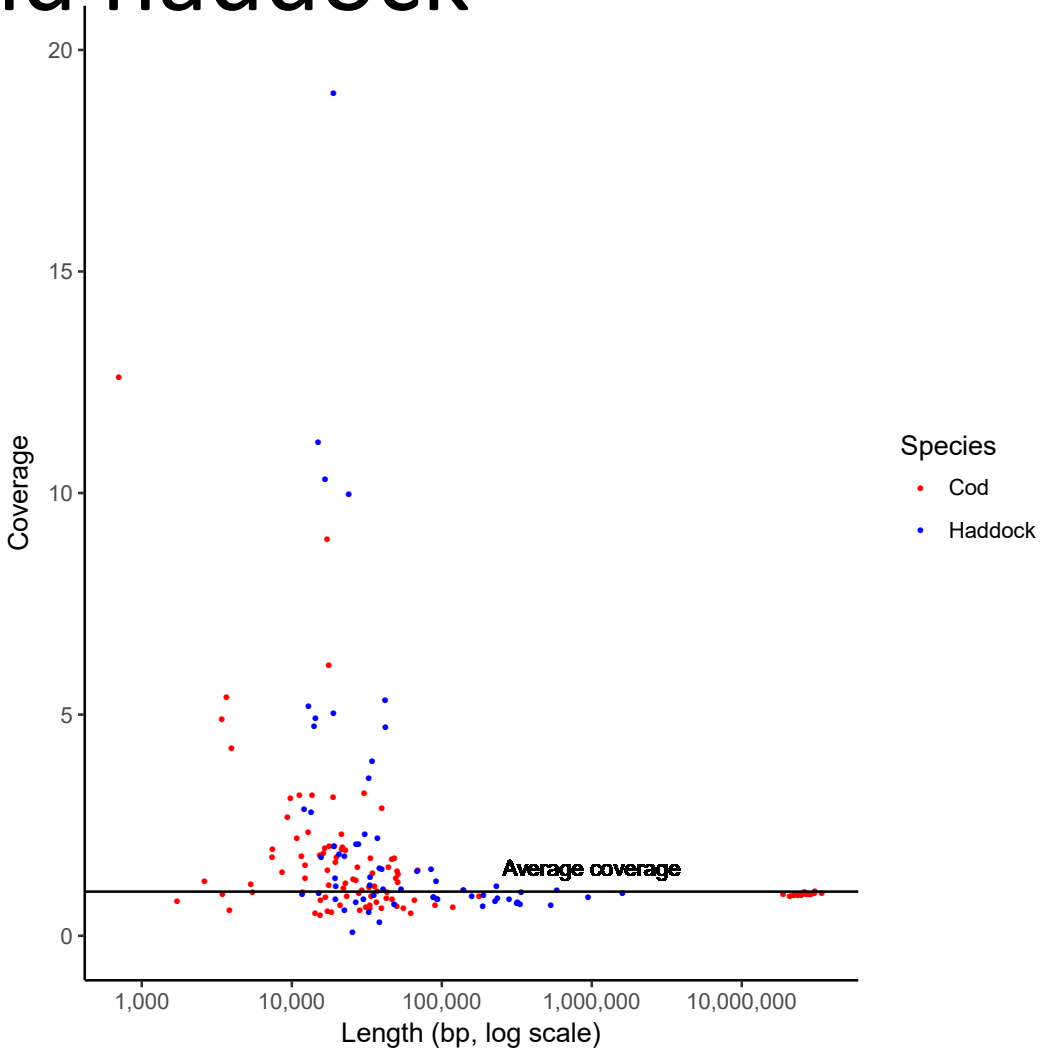


	Predicted proteins		Genome assembly	
	# NACHT	# FISNA	# NACHT (placed)	# FISNA (placed)
Spotted gar	34	15	32 (19)	16 (2)
Cavefish	97	90	107 (NA)	115 (NA)
Zebrafish	348	335	420 (380)	401 (361)
Stickleback	87	234	320 (52)	326 (46)
Fugu	42	38	76 (NA)	106 (NA)
Tetraodon	19	7	36 (10)	46 (10)
Tilapia	112	107	208 (NA)	266 (NA)
Medaka	26	18	58 (39)	93 (53)
Amazon molly	94	82	120 (NA)	111 (NA)
Platyfish	29	14	120 (NA)	111 (NA)
Atlantic cod	133	137	178 (93)	191 (90)
Haddock	36	41	59 (NA)	71 (NA)

- Genome annotation likely underrepresent the actual numbers

Sequences with *NLRs* are collapsed in cod and haddock

- Multiple copies of a gene are like repeats
- The genome assembly likely underrepresents the number of NLRs



Summary

- Atlantic cod and haddock have a high density and frequency of STRs in both coding and non-coding regions, likely of evolutionary consequences
- Likely large expansions in the NLRs
- PacBio sequencing data crucial for investigating this

More research is needed

- Present genome assemblies cannot properly represent expanded immune genes nor STRs
- What is the mechanism behind expansion of genes?
- What is the variation in populations of both STRs and immune genes?

#1MbCtgClub

The Assemblathon and 3 others Retweeted

Emily Hatas @EmilyHatas · Mar 28

Plant genome assemblies starting to come in on the Sequel System, all with contig N50 >1Mb #ABRF2017 @PacBio **#1MbCtgClub**

***DE NOVO* GENOME ASSEMBLIES ON SEQUEL**




	A. thaliana	Soybean	Rice	*Sedum alfredii* HZ	*Sedum alfredii* NHZ
# Sequel SMRT Cells	2	7	10	10	10
Assembly size	123 Mb	902 Mb	409 Mb	236 Mb	397 Mb
Contig N50	10.4 Mb	1.2 Mb	1.8 Mb	1.08 Mb	1.26 Mb

1 21 17

The future

- Generated 70X coverage of PacBio reads (P6C4 and Sequel) for Atlantic cod
- Preliminary assembly contig N50 66 kbp ☹
- Looking into both 10X Genomics' Chromium and Dovetail's Chicago and Hi-C for scaffolding
- Tissue without degraded DNA is an issue

- PacBio or Nanopore reads together with 10X linked-reads is a powerful combination