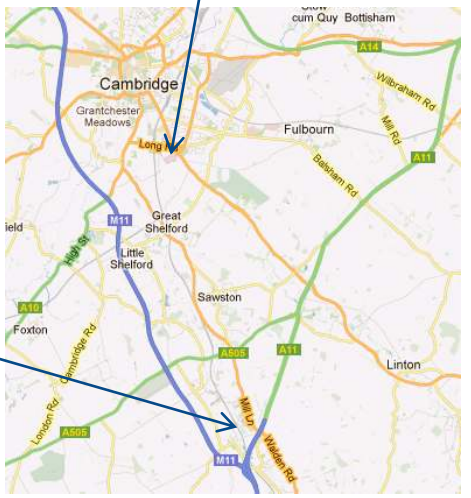
SMRT Sanger:
A Smorgasbord Tapas of SMRT Sequencing at the Sanger Institute

Mike Quail, Senior Staff Scientist WTSI
mq1@sanger.ac.uk

Wellcome Trust Sanger Institute

LMB

WTSI



A community of organisations

Research

Learning & Engagement

Enterprise & Innovation

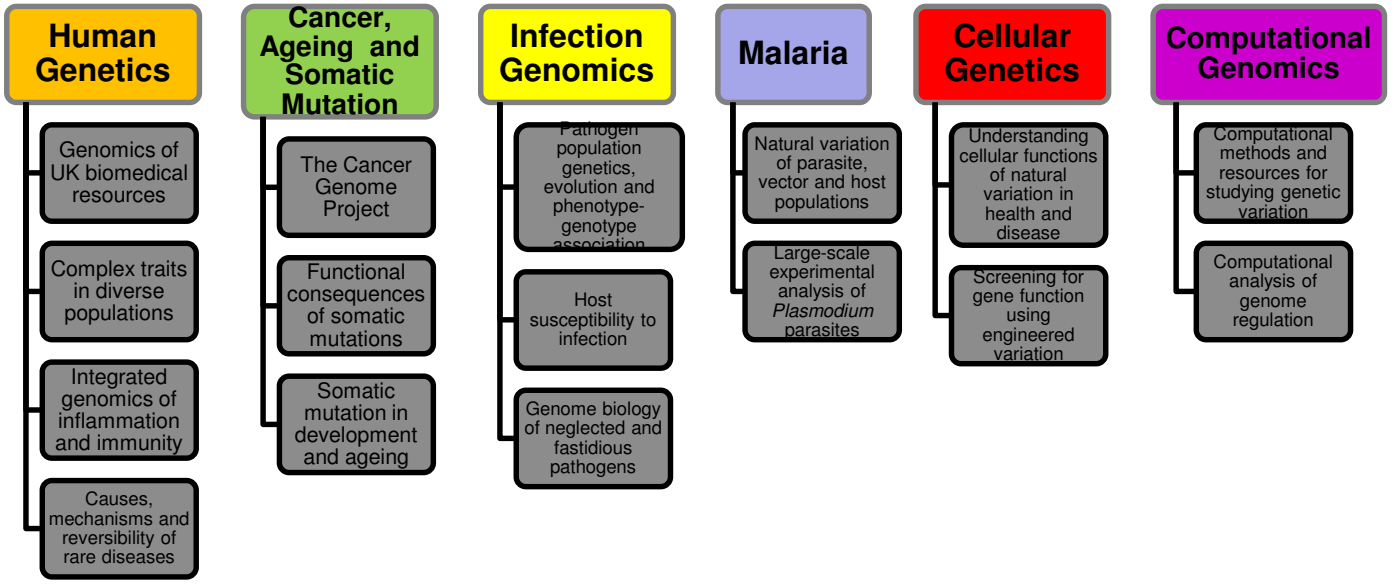
Ogilvie Building

Home of sequencing operations

Wellcome Trust Sanger Institute
People, budget and platforms

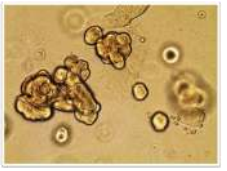
- 1300 people
- Operating budget: ~£100m per year
- Centre for large-scale biomedical science research
- Largest DNA sequencing facility in Europe
- One of the largest biological data-centres in Europe
- High-throughput cell culture facility
- Largest mouse vivarium in Europe

Strategic areas of science



Scientific Operations

- DNA pipelines
- Animal Facility and Mouse Pipeline
- Cellular Genetics and Phenotyping Facility
- Single Cell Genomics
- Information Technology



Instruments: production

2

10

10

5

©2011, Illumina Inc. All rights reserved.

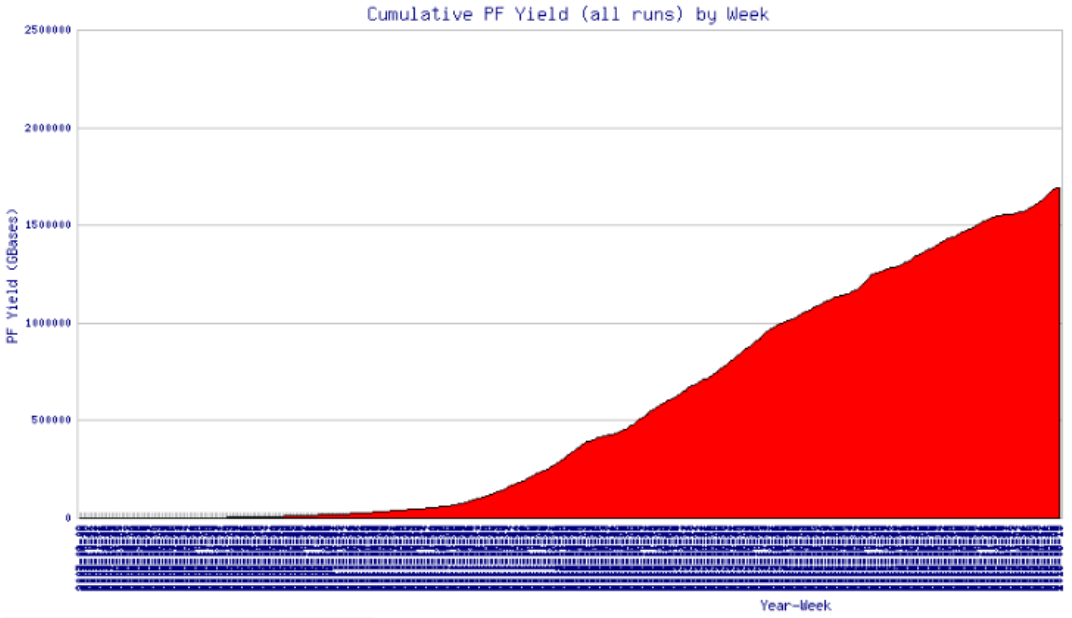
Cumulative yield

>~ 4Pb , Average monthly yield = >100Tb

Statistical Data by Week

Statistics by Run | Statistics by Week | Instrument Statistics| Error Rate for Latest Runs

Yield | Cum. Yield | Error | Avg. Yield per Read Lane



Cumulative PF Yield by Week

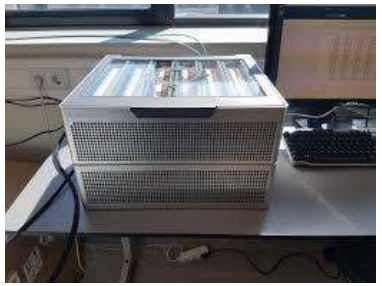
Year-Week	Cumulative PF Yield(GBases)
2016-06-19	1693449.86

PacBio RS II

Sequel

Sequel

Long Read Technologies



Michelle Smith

Kim Judge

PacBio RS II

The PacBio Team

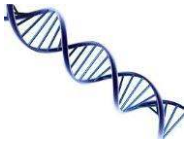
Karen Oliver

Jason Skelton

Craig Corton

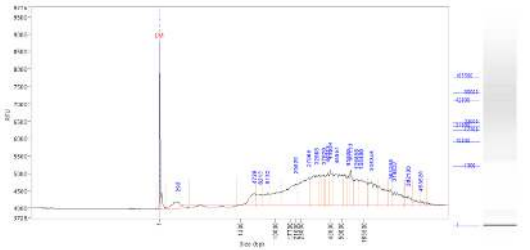
Plus Emma Betteridge

What we do in the lab - QC



&

Sequel samples



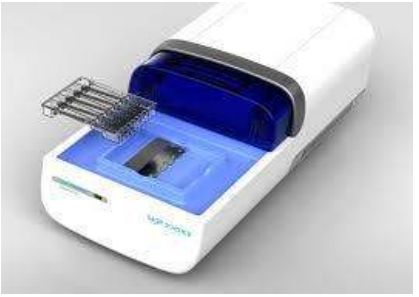
What we do in the lab – Library prep

RSII

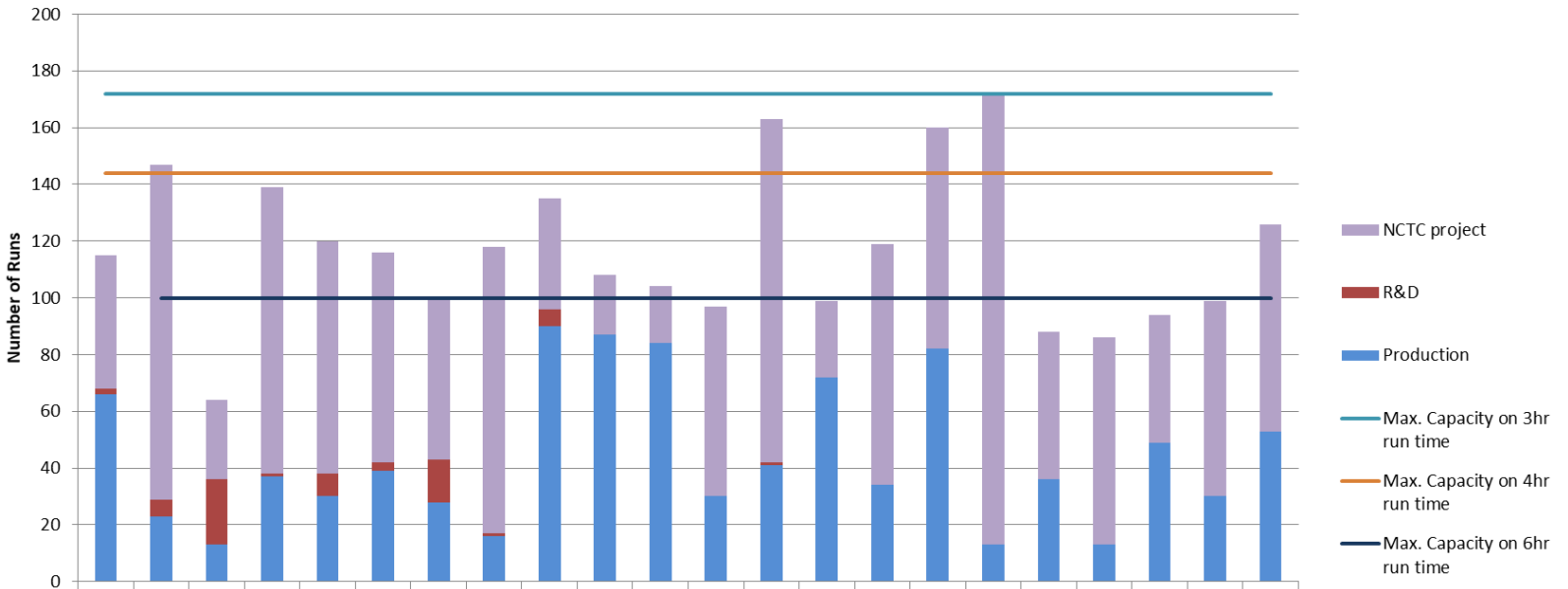
Library Prep

Sequel

Library Prep



PacBio RSII Sequencing number of SMRT cells processed by month



	Nov-15	Dec-15	Jan-16	Feb-16	Mar-16	Apr-16	May-16	Jun-16	Jul-16	Aug-16	Sep-16	Upto 25th Oct	26.10.16-24.11.16	25.11.16-19.12.16	20.12.16-25.01.17	26.01.17-22.02.17	23.02.17-27.03.17	28.03.17-24.04.17	25.04.17-24.05.17	25.05.17-26.06.17	27.06.17-25.07.17	26.07.17-25.08.17
NCTC project	47	118	28	101	82	74	57	101	39	21	20	67	121	27	85	78	159	52	73	45	69	73
R&D	2	6	23	1	8	3	15	1	6	0	0	0	1	0	0	0	0	0	0	0	0	
Production	66	23	13	37	30	39	28	16	90	87	84	30	41	72	34	82	13	36	13	49	30	53
Max. Capacity on 3hr run time	172	172	172	172	172	172	172	172	172	172	172	172	172	172	172	172	172	172	172	172	172	172
Max. Capacity on 4hr run time	144	144	144	144	144	144	144	144	144	144	144	144	144	144	144	144	144	144	144	144	144	144
Max. Capacity on 6hr run time		100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100

Sequencing 3000 Bacteria From Public Health England's National Collection of Type Cultures

wellcome trust sanger institute

PACIFIC BIOSCIENCES

NCTC National Collection of Type Cultures

Operated by Public Health England

NCPV National Collection of Pathogenic Viruses

Operated by Public Health England

Nick Grayson (Sanger)
Sarah Alexander (PHE)

@NCTC_3000

NCTC

NCTC
National Collection of Type Cultures
Operated by Public Health England

- Founded in 1920

- >5000 bacteria > 300 species

- 900 type strains

- Medical / Veterinary importance

- Historical / Scientific interest

- www.phe-culturecollections.org.uk

Robert Koch

Alexander Fleming

NCTC Website

Bacteria Collection: Neisseria weaveri

NCTC National Collection of Type Cultures
Operated by Public Health England

NCTC Number:	NCTC 13585
Current Name:	Neisseria weaveri
Original Strain Reference:	Danderyd 6209
Other Collection No:	CCUG 30381
Other Names:	Formerly Neisseria parelongata
Type Strain:	No
Family:	Neisseriaceae
Hazard Group (ACDP):	2
Release Restrictions:	Terms & Conditions of Supply of Microbial Pathogens: Safety
Conditions for growth on solid media:	Chocolate blood agar, 37°C, CO2, 1 day
Isolated From:	Human wound (dog bite)
Whole Genome Sequence:	http://www.ebi.ac.uk/ena/data/view/ERS627823
Annotated Genome:	ftp://ftp.sanger.ac.uk/pub/project/pathogens/NCTC3000/d...
Accession Date:	Jan 2012
Authority:	Holmes et al. 1993 (VP)
Depositor:	Culture Collection, University of Goteborg, Sweden (CCUG)
Biosafety Responsibility:	It is the responsibility of the customer to ensure that their facilities comply with biosafety regulations

Formats	Price
⊙ Ampoule (Bacteria)	
In Stock	£150.00

Quantity

Add to cart

Click for Delivery Prices

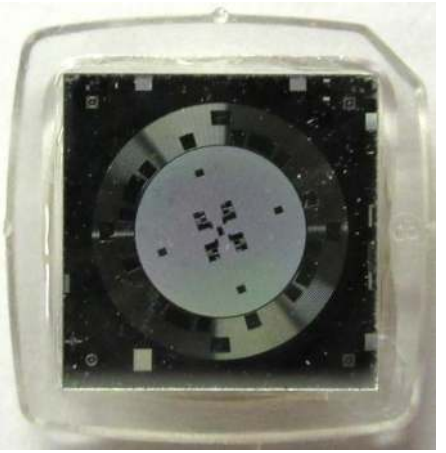
Credit Account orders accepted or email / fax your order to us.

Certificate of Analysis

Bacterial Sequencing

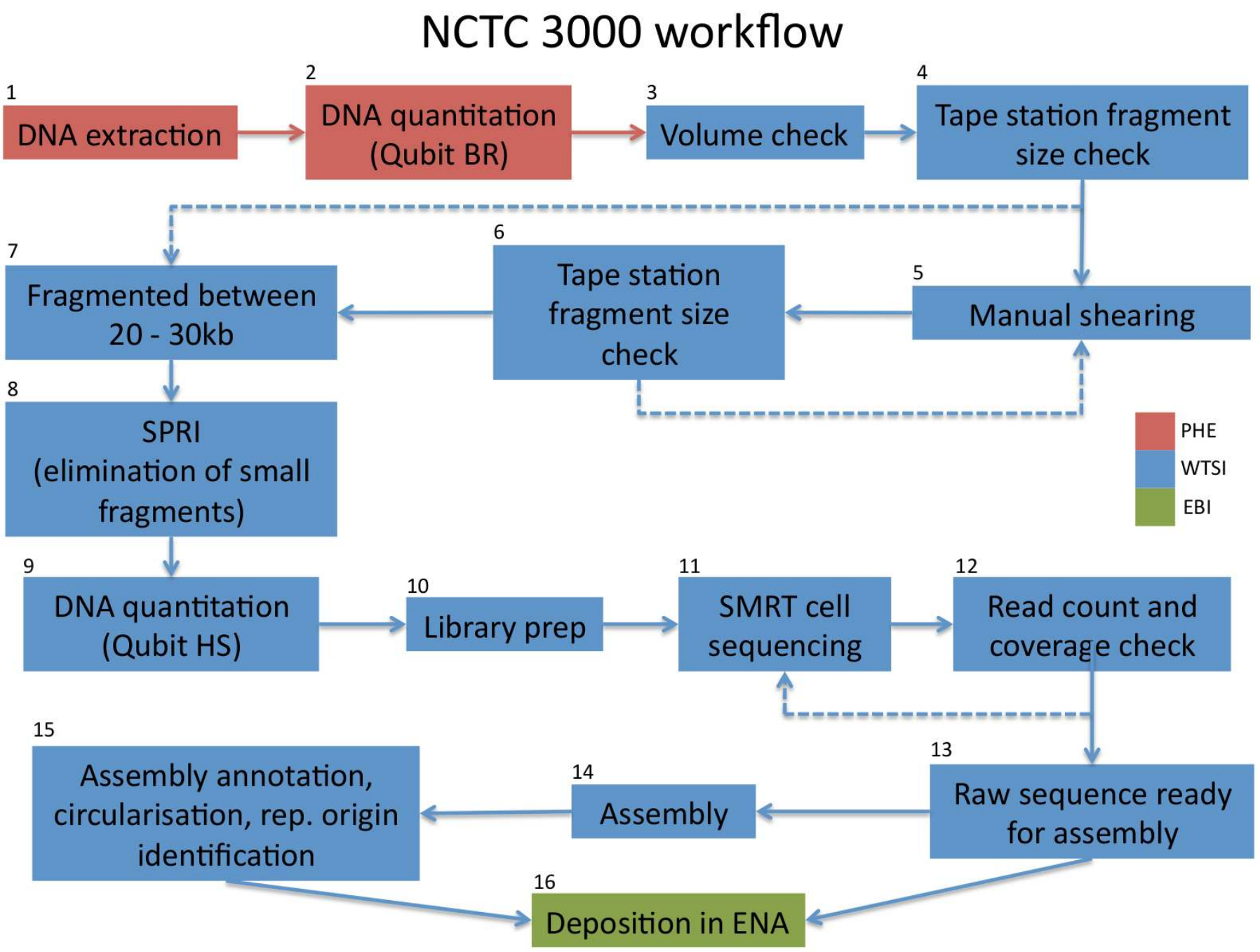
- PacBio RSII
- Fragment to ~15kb
- 400Mb / SMRT Cell
- 1.07 SMRT cells / genome
- 20/week
- Robust to GC %

New Ogilvie Building

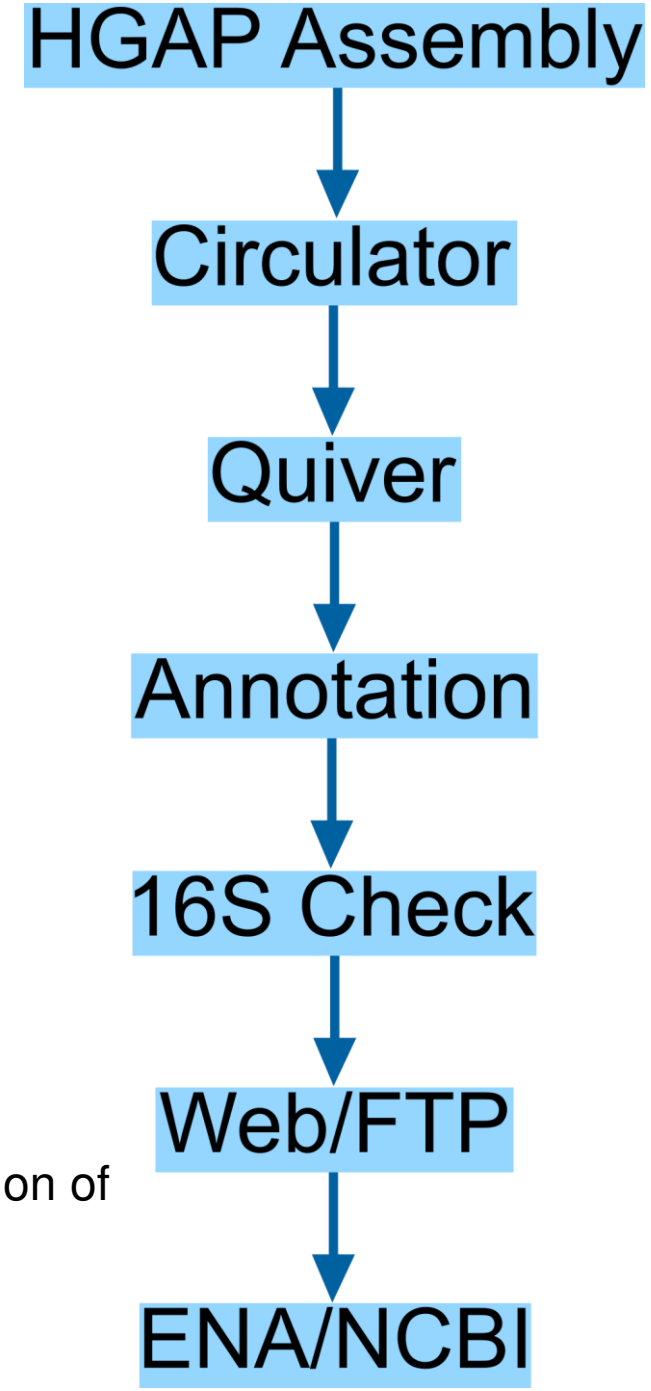


1st SMRT Cell

PacBio RSII


NCTC 3000 workflow
1
DNA extraction
2
DNA quantitation (Qubit BR)
3
Volume check
4
Tape station fragment size check
5
Manual shearing
6
Tape station fragment size check
7
Fragmented between 20 - 30kb
8
SPRI (elimination of small fragments)
9
DNA quantitation (Qubit HS)
10
Library prep
11
SMRT cell sequencing
12
Read count and coverage check
13
Raw sequence ready for assembly
14
Assembly
15
Assembly annotation, circularisation, rep. origin identification
16
Deposition in ENA
PHE
WTSI
EBI

Analysis Pipeline



Circlator: automated circularization of genome assemblies using long sequencing reads. Hunt *et* al. Genome Biology 2015 16:294

NCTC1

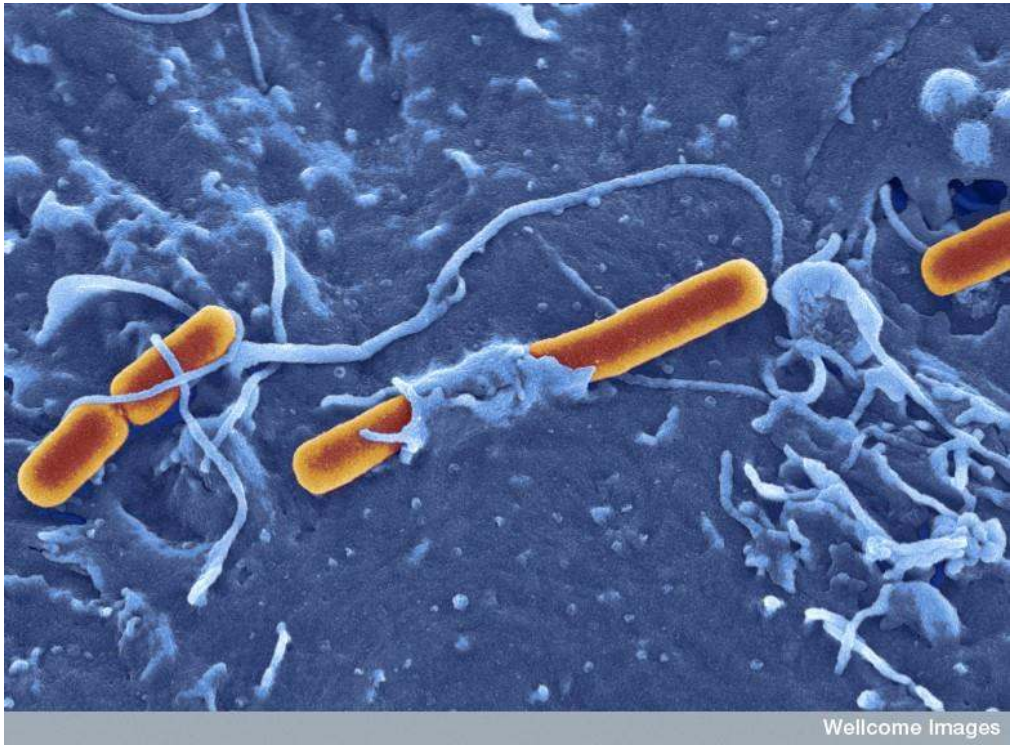
- Collected 1915

- Shigella Flexneri

- Antibiotic Resistant

- 98% genes conserved

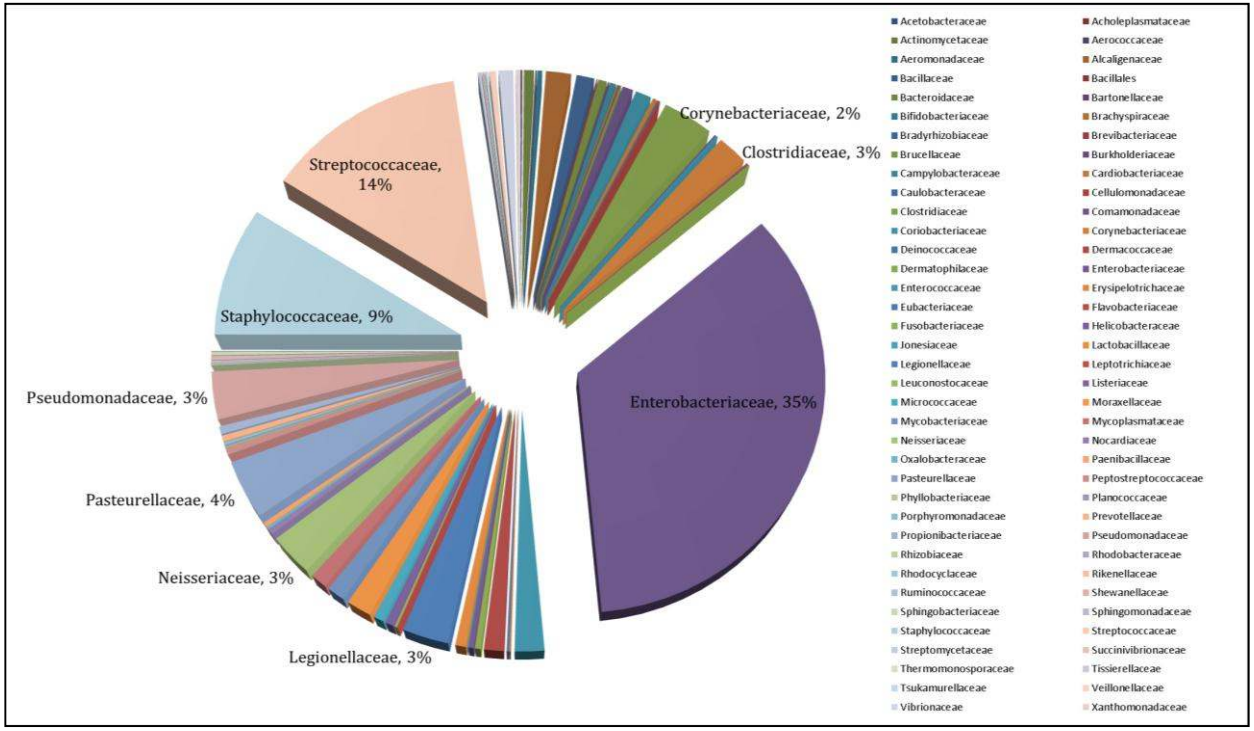
- Vaccine development

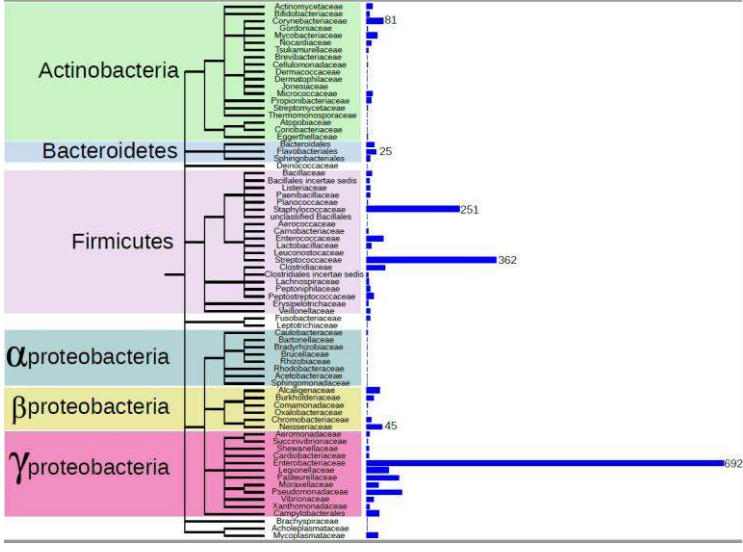


The extant World War 1 dysentery bacillus NCTC1: a genomic analysis. Kate S Baker *et* al. Lancet Vol. 384, No. 9955, p1691–1697, 8/11/2014

Bacteria Sequenced To Date

- NCTC3000 Project:
  - 2169 strains
  - 832 species
  - 75 families represented
  - 85.4% Type strain completed
- Genomes:
  - 1255 publically available
  - 56.5% single contig
  - 23.1% - Evidence of =>1 plasmids

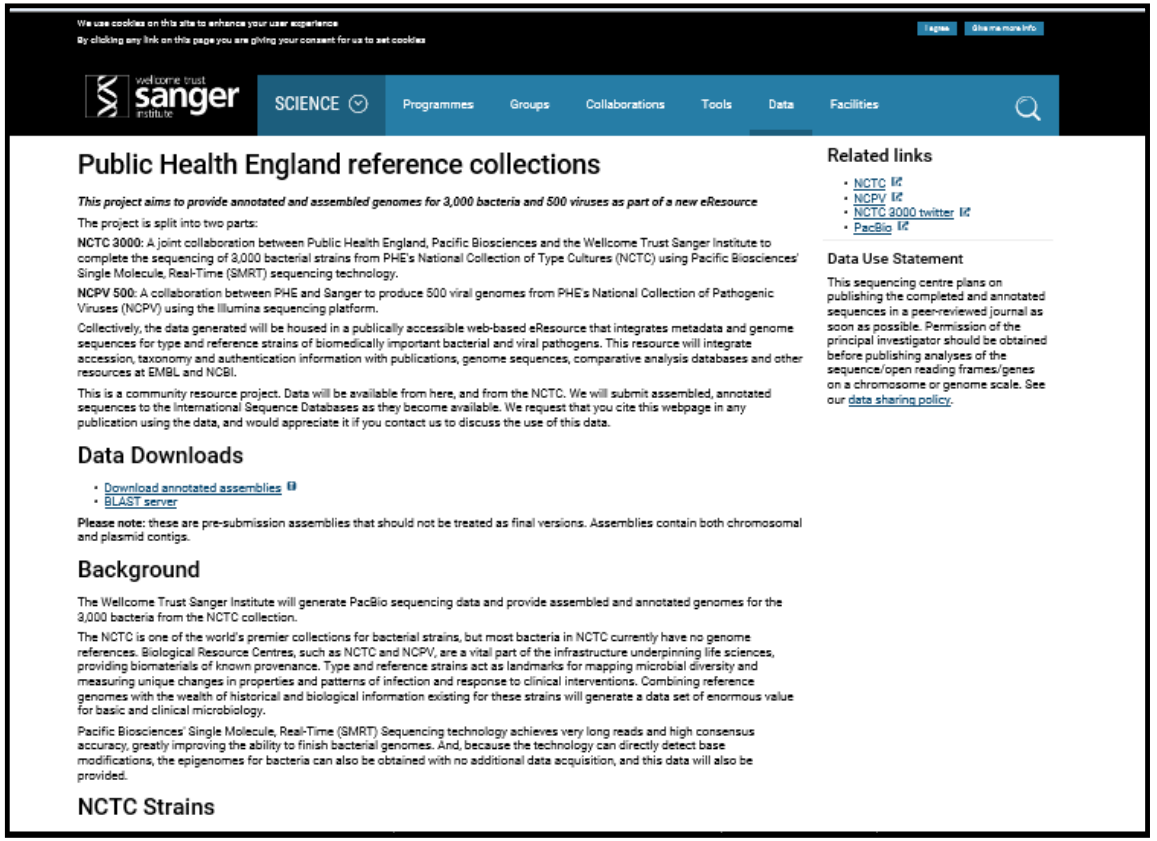




http://www.sanger.ac.uk/resources/downloads/bacteria/nctc/

NCTC3000 – Data Sharing

- Data regularly uploaded on the WTSI website
- http://www.sanger.ac.uk/resources/downloads/bacteria/nctc/#t_2

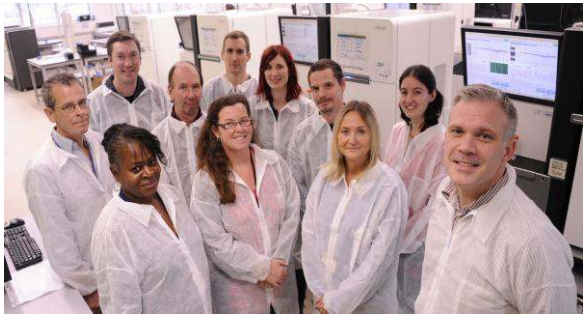


@NCTC_3000

Acknowledgements

Sanger Institute

- Julian Parkhill
- Nancy Holroyd
- Jane Lomax
- Gemma Langridge
- Liz McMinn
- Theresa Feltwell
- Kate Baker
- Alison Mather
- Richard Rance
- Jacqui Brown
- Kirsty Ambridge
- Karen Oliver
- Craig Corton



Public Health England

- Sarah Alexander
- Julie Russell

NCPV

- Karen Buttigieg
- Edward Burnett

NCTC

- Mohammed Fazal
- Trisha Rawat

Nick.grayson@sanger.ac.uk

@nickegrayson

Sequel

Scaling up the generation of reference quality genomes across a range of vertebrate diversity

Iliana Bista

wellcome trust sanger institute

@ilianailiana3

Aims

Richard Durbin
Sanger VGP Lead

- Sanger Institute core funded project 2016-17 aiming to sequence 50-100 vertebrate genomes
- Generate novel reference quality genomes
- Evaluate sequencing technologies for reference sequence quality assemblies
- Engage with international Genome10K / VGP consortium

Iliana Bista

Target species

We are part of the international Vertebrate Genomes Project, with an initial goal of producing reference quality genome assemblies of a representative of every vertebrate order.

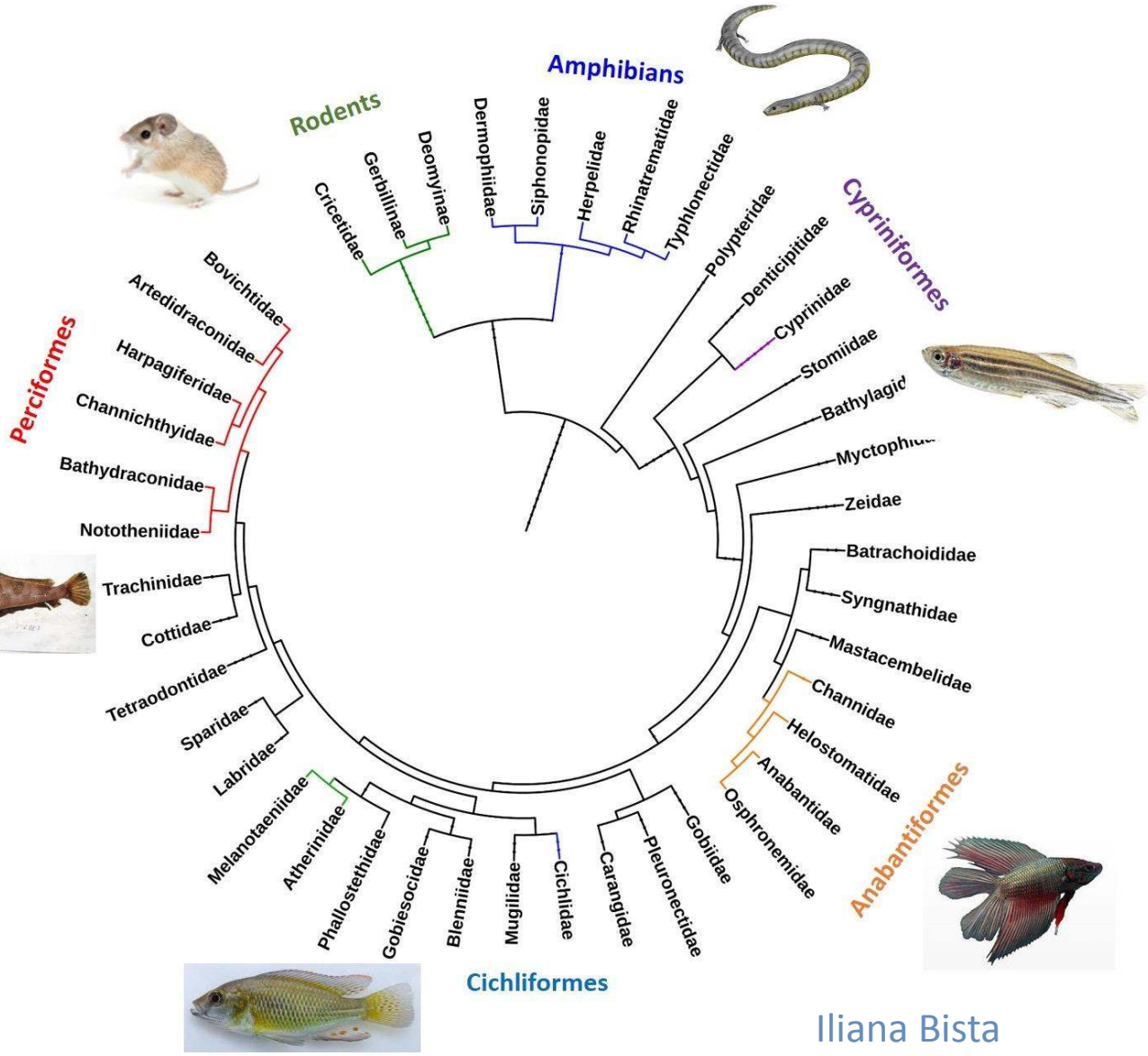
1. Fish

- Representatives of ray-finned fish orders
- The Antarctic radiation of notothenioid fish
- The Malawi cichlid radiation
- The anabantoid group (gouramis)
- Strains of zebrafish and related cyprinids

2. Caecilian amphibians

3. Rodents

- selected species



Iliana Bista

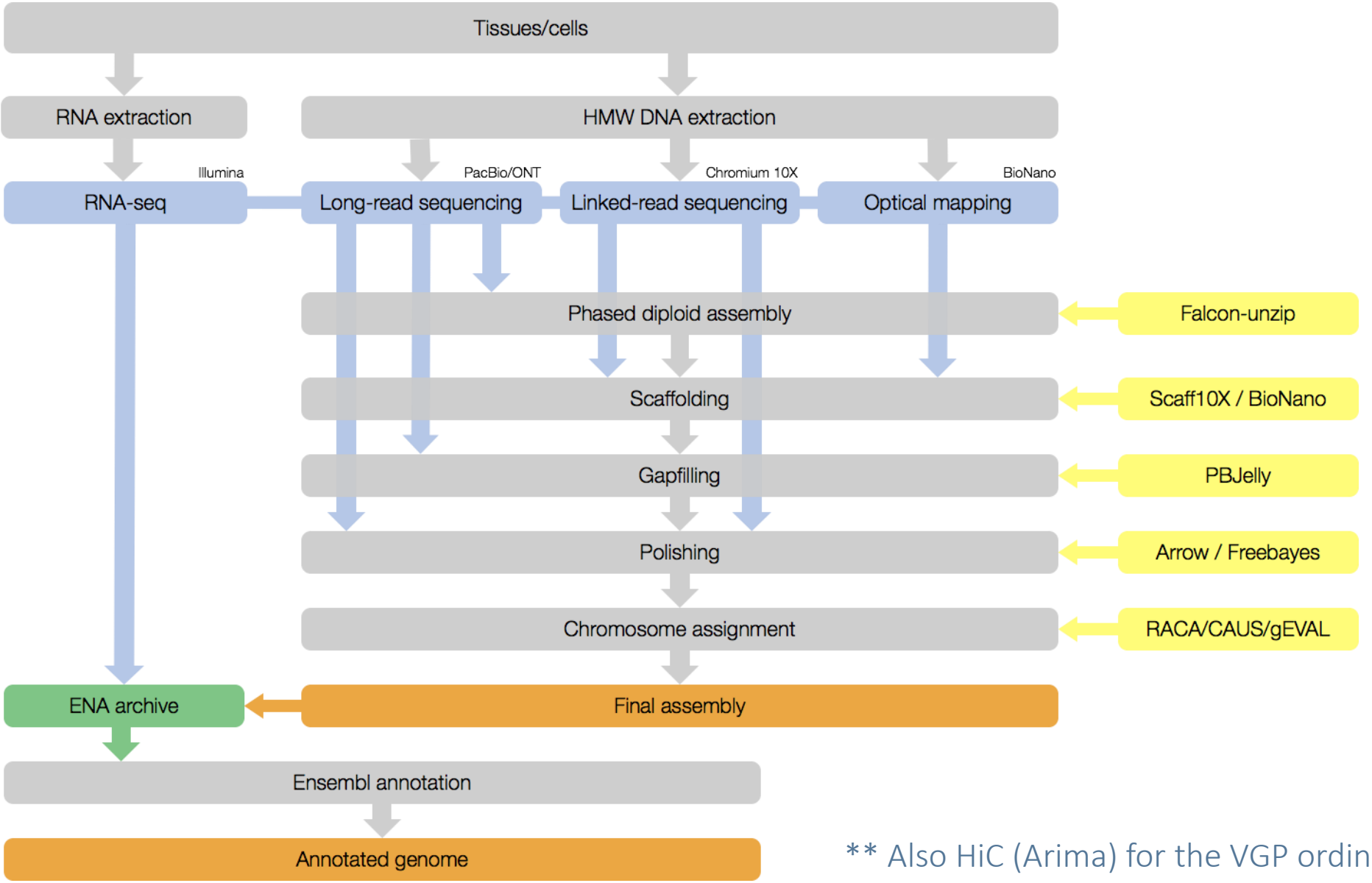
Data collection

- Collection of high quality tissue – flash frozen
  - Iliana has explored ethanol-fixed quick frozen samples for easier collection
- High Molecular Weight DNA extraction
  - Agarose plugs
  - Magnetic beads based extraction protocols

- PacBio Sequel (~60x)
- 10X Genomics - Chromium
- Bionano Irys (Saphyr when available)
- Arima HiC
- Oxford Nanopore GridION

- Ordinal project = PacBio+10X+BioNano Saphyr+Arima

Workflow



** Also HiC (Arima) for the VGP ordinal project

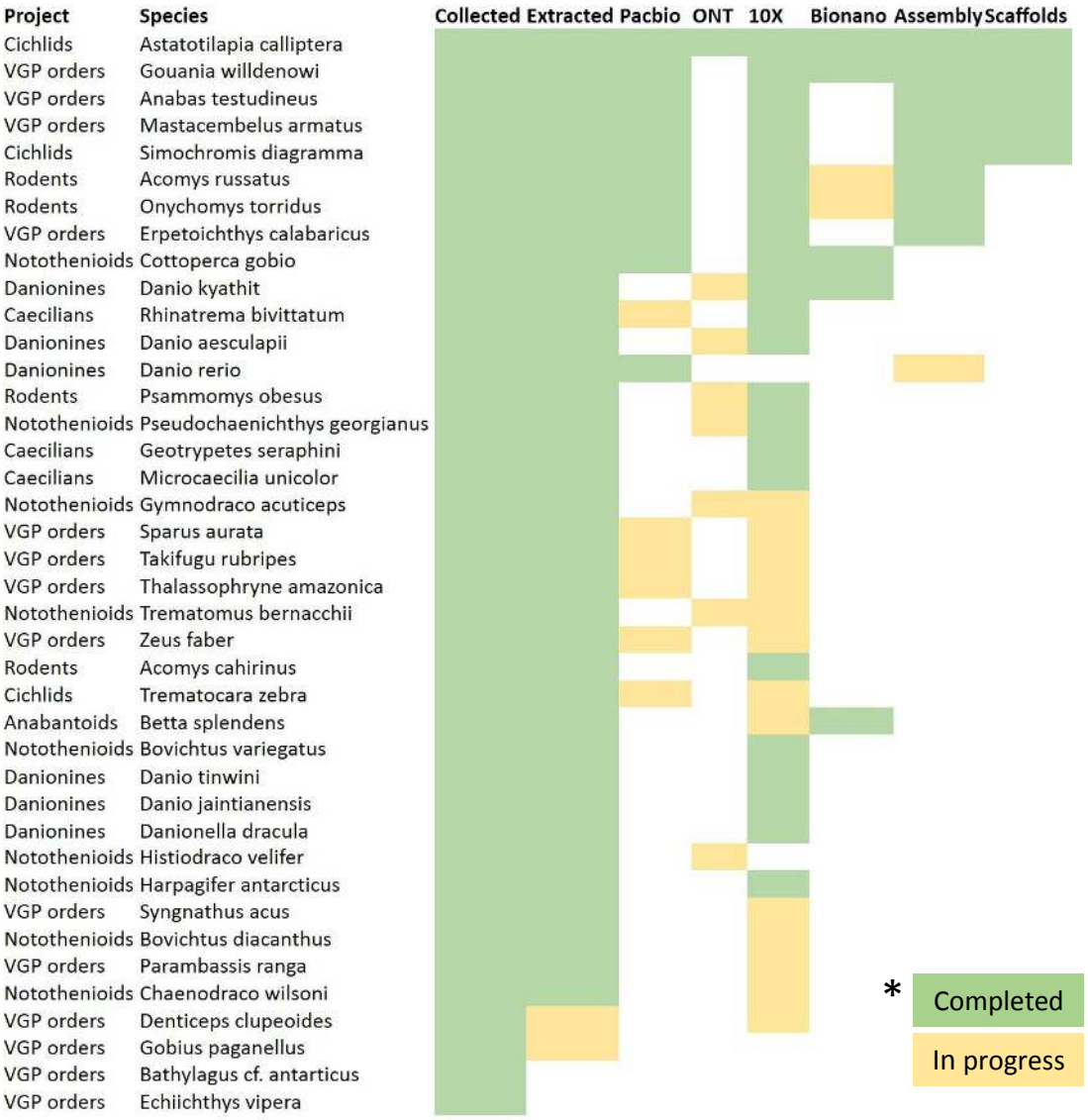
3.4.2q40 reference quality standard

- Project target of:
  - >1Mb contig N50 and
  - >10Mb scaffold N50
  - Chromosome (or syntenic block) assignment of >90% through synteny or where possible genetic maps
  - Average quality >Q40

Progress summary

- **PacBio**: 110 SMRT cells, 3.6 Gb/cell
- **10X Genomics**: data for 43 species
- **HiSeqX**: data for 36 species
- Complete **BioNano maps** for 4 species

- Chromosome group level assignment for the first 4 genomes – soon to be released



Current state of our assemblies

		Dazzler+Falcon			PacBio miniasm			Merged and Scaffolded					Gapfilled					Polishing	
		CtgN50 (Mb)	#Ctgs	Length (Gb)	CtgN50 (Mb)	#Ctgs	Length (Gb)	CtgN50 (Mb)	#Ctgs	ScaffN50 (Mb)	#Scaffs	Length (Gb)	**CtgN50 (Mb)**	**#Ctgs**	**ScaffN50 (Mb)**	**#Scaffs**	**Length (Gb)**	Arrow	10X
fAnaTes1	*	1.83	1,985	0.577	2.67	1,794	0.600	1.89	1,576	16.51	612	0.566	**3.49**	**1,150**	**16.56**	**612**	**0.573**	Y	Y
fAstCal1		2.47	1,089	0.877	1.06	5,068	1.04	2.47	1087	11.24	352	0.877	**3.33**	**914**	**11.27**	**352**	**0.883**	Y	
fCotGob3	*	0.158	10,731	0.644	1.32	1,641	0.646	1.10	3,575	3.39									
fErpCal1	*	0.506	19,458	3.48	0.837	13,339	3.77	0.925	12,813	1.26									
fGouWil2	*	0.402	8,201	0.975	0.903	6,181	1.13	0.434	6,679	7.11	2,216	0.931	**0.822**	**4,673**	**7.10**	**2,216**	**0.974**	Y	Y
fMasArm1	*	1.16	1,566	0.576	4.38	1,854	0.627	1.16	1,553	10.54	690	0.576	**2.80**	**1,051**	**10.67**	**690**	**0.583**	Y	Y
fSimDia1		1.22	3,301	0.848	1.09	5,447	1.01	1.24	3,294	6.10	1,593	0.848	**1.81**	**2,827**	**6.16**	**1,593**	**0.864**	Y	Y
fSpaAur1	*	0.785	4,560	0.834	1.72	2,863	0.890	0.785	4,553	1.90	2,880	0.834							
mAcoRus1		0.368	13,067	2.19	5.09	3,077	2.43												
mOncTor1		1.00	7,678	2.40	1.85	8,904	2.65												
fMasArm1	*	4.99	421	0.576	4.38	1,854	0.627	4.99	420	12.77	196	0.576	6.25	**349**	12.75	**196**	**0.578**	Y	

fAstCal1, *Astatotilapia calliptera*, Eastern Happy Malawi cichlid

fAnaTes1: *Anabas testudineus*, climbing perch

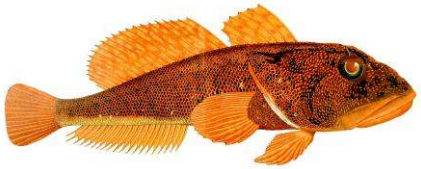
fGouWil2: *Gouania willdenowi*, blunt-snouted clingfish

fMasArm1: *Mastacembelus armatus*, tire track eel

fSimDia1: *Simochromis diagramma*, Tanganyika cichlid

Current state of our assemblies

		Dazzler+Falcon			PacBio miniasm			Merged and Scaffolded					Gapfilled					Polishing	
		CtgN50 (Mb)	#Ctgs	Length (Gb)	CtgN50 (Mb)	#Ctgs	Length (Gb)	CtgN50 (Mb)	#Ctgs	ScaffN50 (Mb)	#Scaffs	Length (Gb)	**CtgN50 (Mb)**	**#Ctgs**	**ScaffN50 (Mb)**	**#Scaffs**	**Length (Gb)**	Arrow	10X
fAnaTes1	*	1.83	1,985	0.577	2.67	1,794	0.600	1.89	1,576	16.51	612	0.566	**3.49**	**1,150**	**16.56**	**612**	**0.573**	Y	Y
fAstCal1		2.47	1,089	0.877	1.06	5,068	1.04	2.47	1087	11.24	352	0.877	**3.33**	**914**	**11.27**	**352**	**0.883**	Y	
fCotGob3	*	0.158	10,731	0.644	1.32	1,641	0.646	1.10	3,575	3.39									
fErpCal1	*	0.506	19,458	3.48	0.837	13,339	3.77	0.925	12,813	1.26									
fGouWil2	*	0.402	8,201	0.975	0.903	6,181	1.13	0.434	6,679	7.11	2,216	0.931	**0.822**	**4,673**	**7.10**	**2,216**	**0.974**	Y	Y
fMasArm1	*	1.16	1,566	0.576	4.38	1,854	0.627	1.16	1,553	10.54	690	0.576	**2.80**	**1,051**	**10.67**	**690**	**0.583**	Y	Y
fSimDia1		1.22	3,301	0.848	1.09	5,447	1.01	1.24	3,294	6.10	1,593	0.848	**1.81**	**2,827**	**6.16**	**1,593**	**0.864**	Y	Y
fSpaAur1	*	0.785	4,560	0.834	1.72	2,863	0.890	0.785	4,553	1.90	2,880	0.834							
mAcoRus1		0.368	13,067	2.19	5.09	3,077	2.43												
mOncTor1		1.00	7,678	2.40	1.85	8,904	2.65												
fMasArm1	*	4.99	421	0.576	4.38	1,854	0.627	4.99	420	12.77	196	0.576	6.25	**349**	12.75	**196**	**0.578**	Y	



fCotGob3: ***Cottoperca gobio*****, Channel bull blenny**

fSpaAur1: ***Sparus aurata*****, gilt head sea bream**

fErpCal1: ***Erpetoichthys calabaricus*****, reedfish**

mAcoRus1: ***Acomys russatus*****, golden spiny mouse**

mOncTor1: ***Onychomys torridus*****, southern grasshopper mouse**

Current state of our assemblies

		Dazzler+Falcon			PacBio miniasm			Merged and Scaffolded					Gapfilled					Polishing	
		CtgN50 (Mb)	#Ctgs	Length (Gb)	CtgN50 (Mb)	#Ctgs	Length (Gb)	CtgN50 (Mb)	#Ctgs	ScaffN50 (Mb)	#Scaffs	Length (Gb)	**CtgN50 (Mb)**	**#Ctgs**	**ScaffN50 (Mb)**	**#Scaffs**	**Length (Gb)**	Arrow	10X
fAnaTes1	*	1.83	1,985	0.577	2.67	1,794	0.600	1.89	1,576	16.51	612	0.566	**3.49**	**1,150**	**16.56**	**612**	**0.573**	Y	Y
fAstCal1		2.47	1,089	0.877	1.06	5,068	1.04	2.47	1087	11.24	352	0.877	**3.33**	**914**	**11.27**	**352**	**0.883**	Y	
fCotGob3	*	0.158	10,731	0.644	1.32	1,641	0.646	1.10	3,575	3.39									
fErpCal1	*	0.506	19,458	3.48	0.837	13,339	3.77	0.925	12,813	1.26									
fGouWil2	*	0.402	8,201	0.975	0.903	6,181	1.13	0.434	6,679	7.11	2,216	0.931	**0.822**	**4,673**	**7.10**	**2,216**	**0.974**	Y	Y
fMasArm1	*	1.16	1,566	0.576	4.38	1,854	0.627	1.16	1,553	10.54	690	0.576	**2.80**	**1,051**	**10.67**	**690**	**0.583**	Y	Y
fSimDia1		1.22	3,301	0.848	1.09	5,447	1.01	1.24	3,294	6.10	1,593	0.848	**1.81**	**2,827**	**6.16**	**1,593**	**0.864**	Y	Y
fSpaAur1	*	0.785	4,560	0.834	1.72	2,863	0.890	0.785	4,553	1.90	2,880	0.834							
mAcoRus1		0.368	13,067	2.19	5.09	3,077	2.43												
mOncTor1		1.00	7,678	2.40	1.85	8,904	2.65												
fMasArm1	*	4.99	421	0.576	4.38	1,854	0.627	4.99	420	12.77	196	0.576	6.25	**349**	12.75	**196**	**0.578**	Y	

- **Scaffolding here is with scaff10x + synteny with cross_genome (Zemin Ning)**
- **Short-molecule length for fSpaAur1 limits scaffolding**
- **Short-molecule length for fMasArm1 perhaps limits scaffolding, but still reached 10Mb target**
- **Too low-coverage 10X for fErpCal1 limits scaffolding**
- **Bottom fMasArm1 is using the Falcon-unzip assembly from the DNAnexus pipeline**

Sequel QC

https://github.com/VertebrateResequencing/SEQUELstats

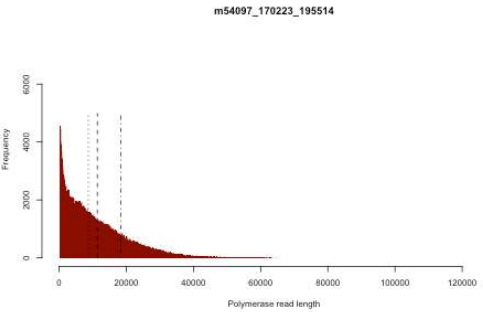
- Inputs are the subreads and scraps BAM files from each SMRTcell of the Sequel, i.e.
  - <smrtcell>.subreads.bam
  - <smrtcell>.scraps.bam

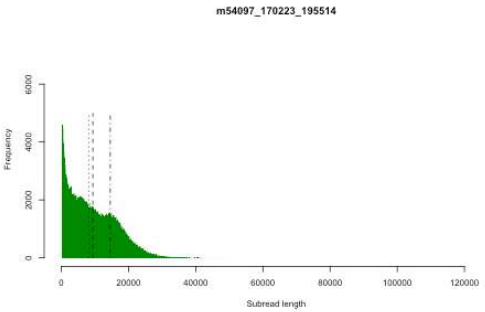
- Parses tags and readnames in the BAM files to determine ZMW IDs, classification of the reads etc.

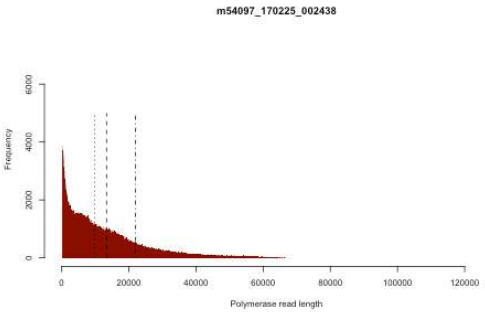
`http://vertreseq.internal.sanger.ac.uk:8000/vgp/`

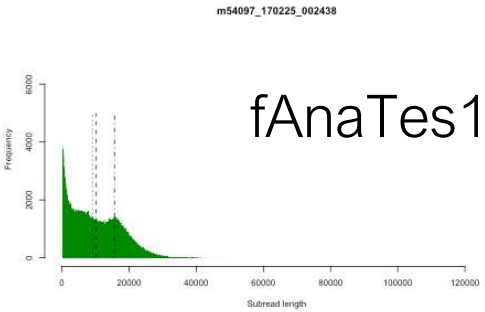
Polymerase and subread length distribution

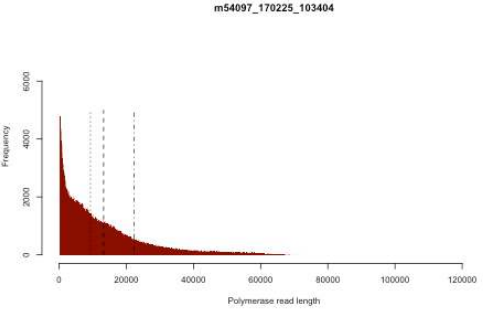
fAnaTes1

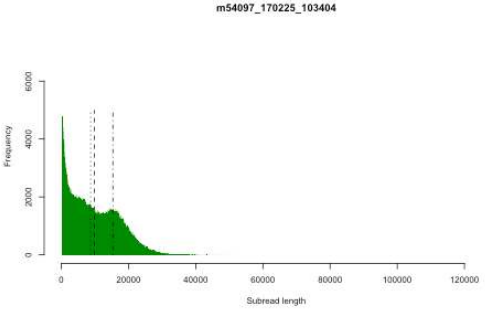


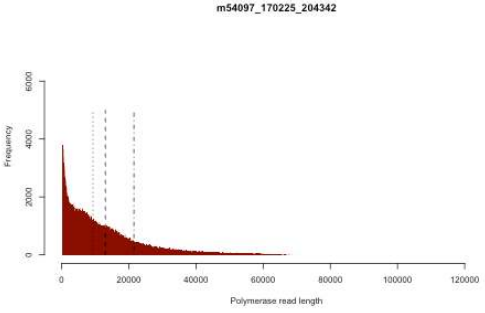


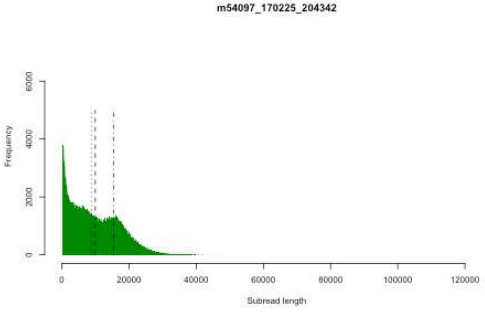


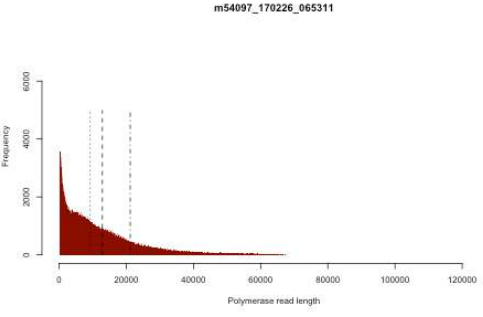


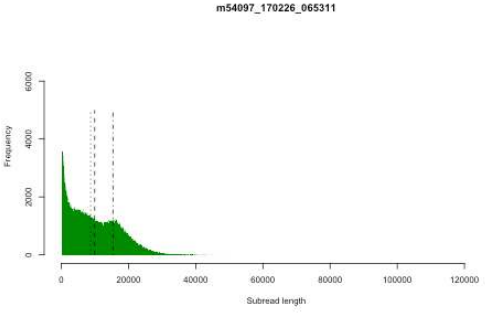


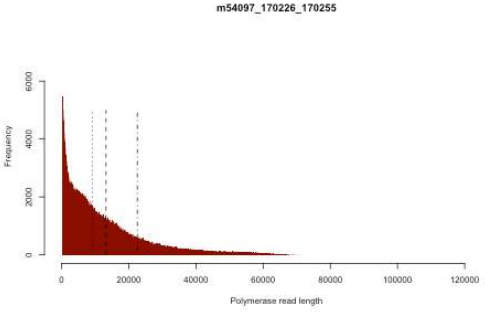


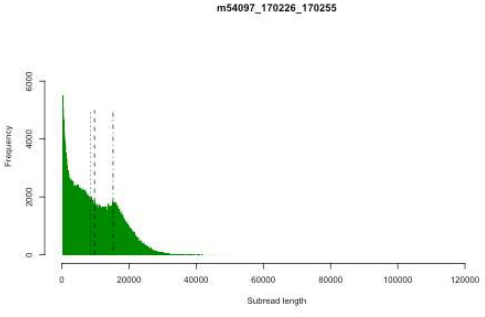


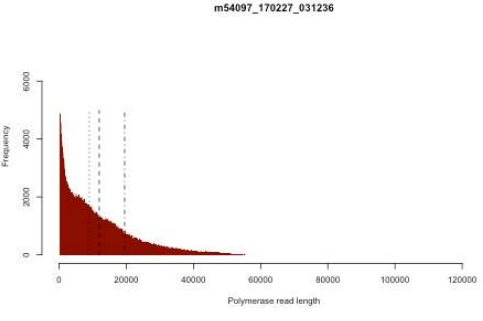


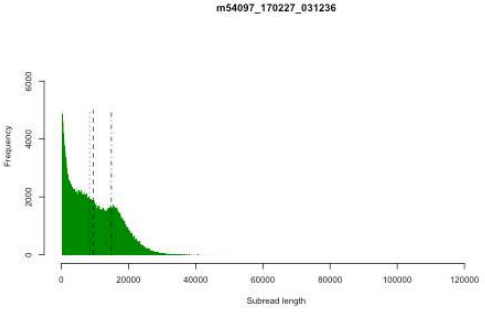


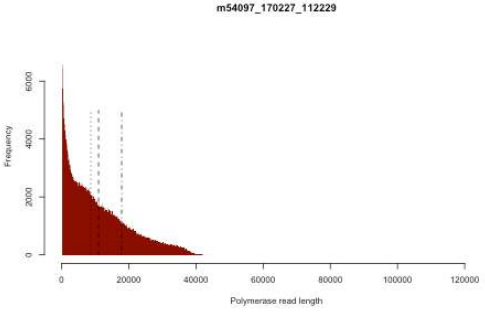


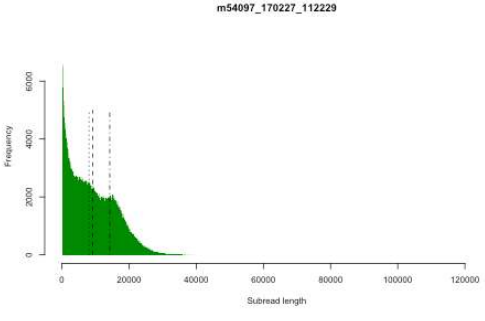




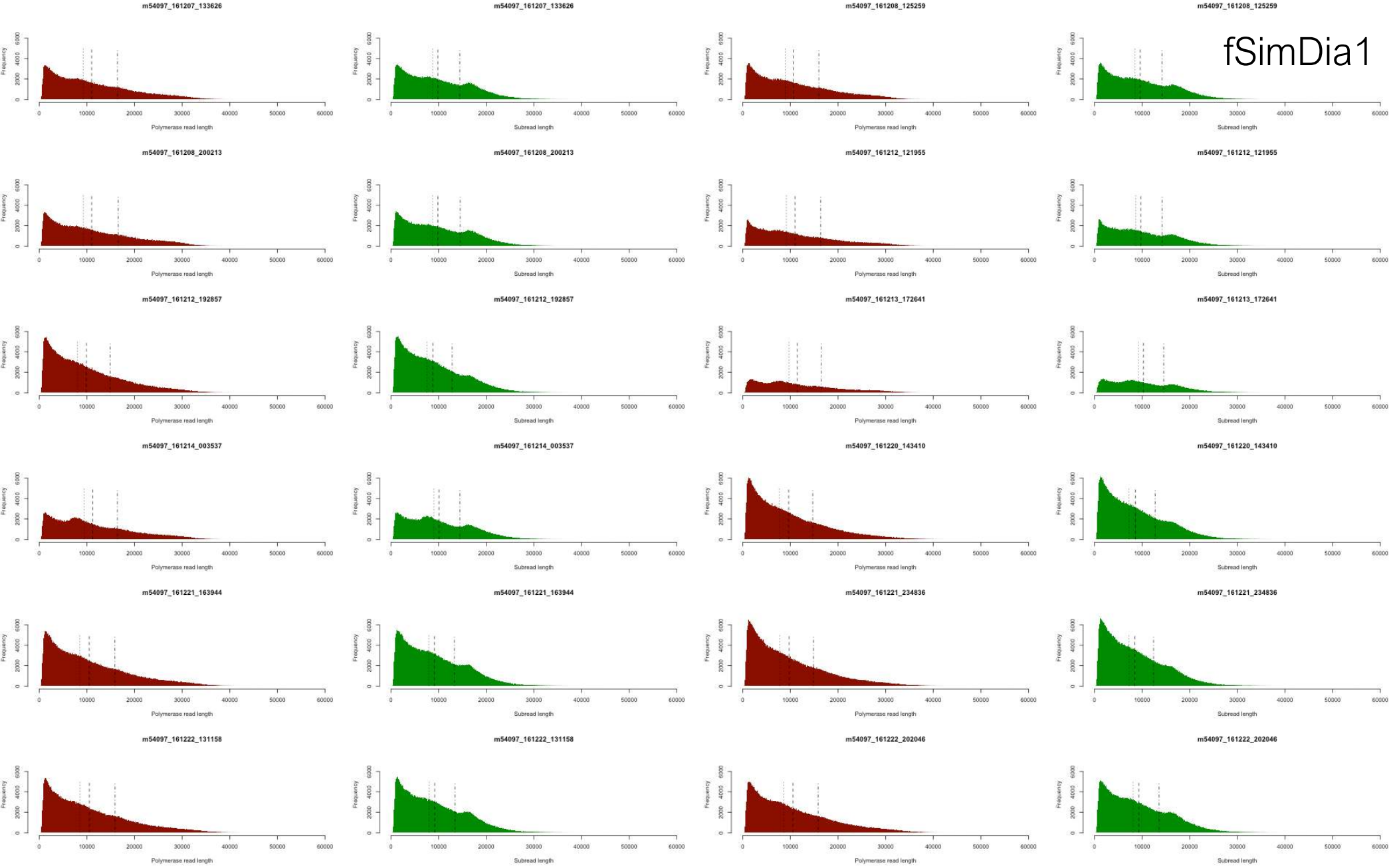








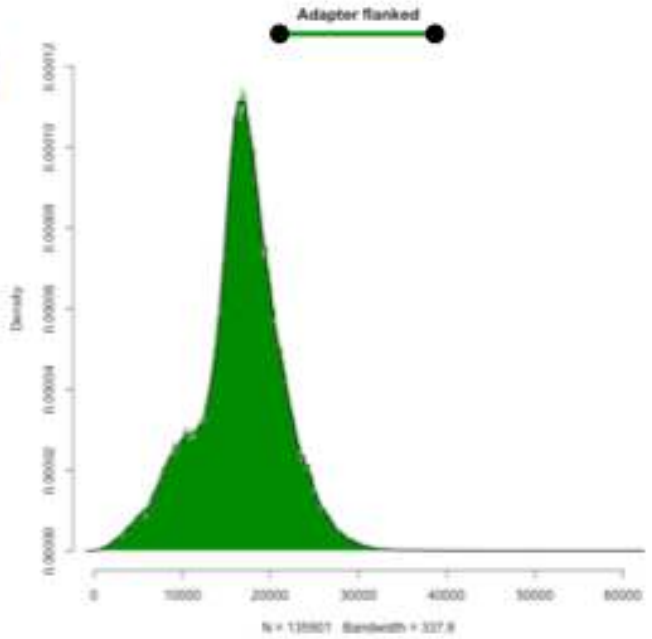
Polymerase and subread length distribution

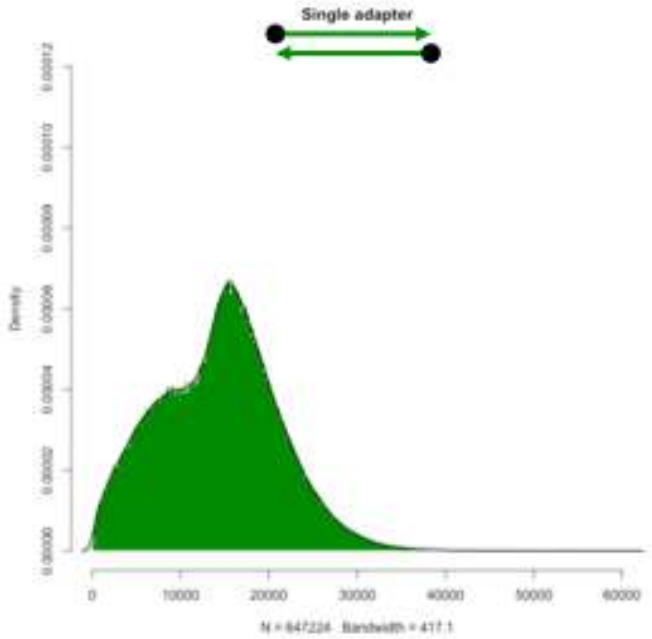


Estimated library size distribution

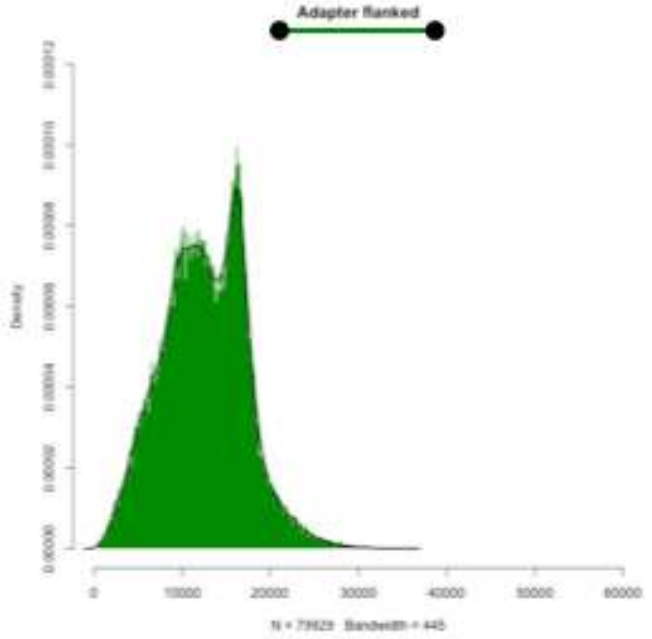
(Longest subread only)

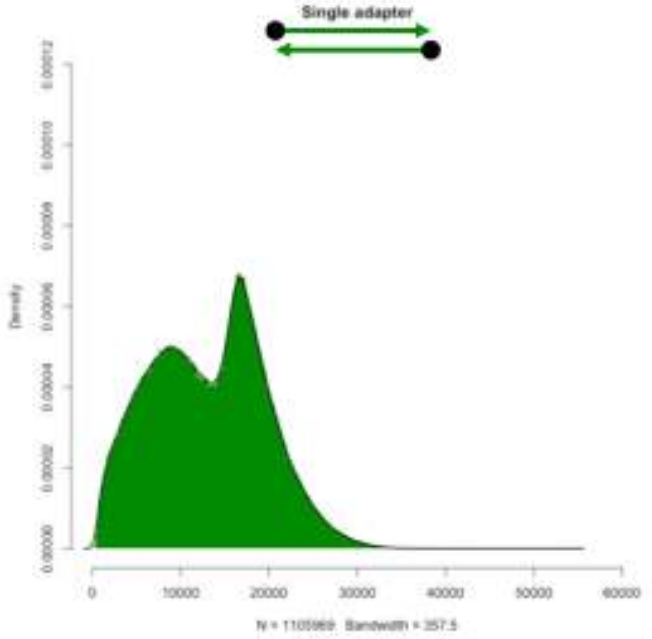
fAnaTes1





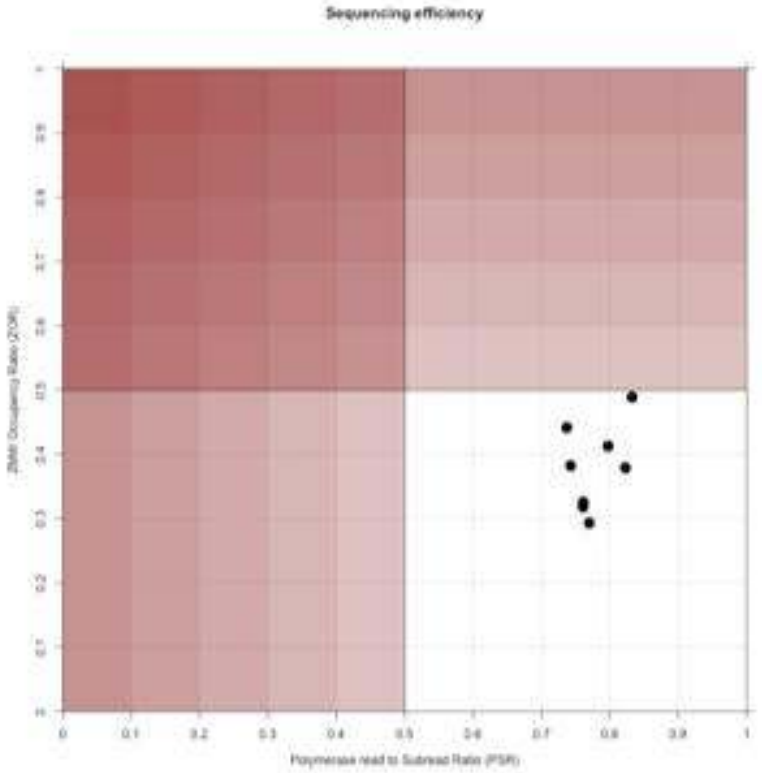
fSimDia1



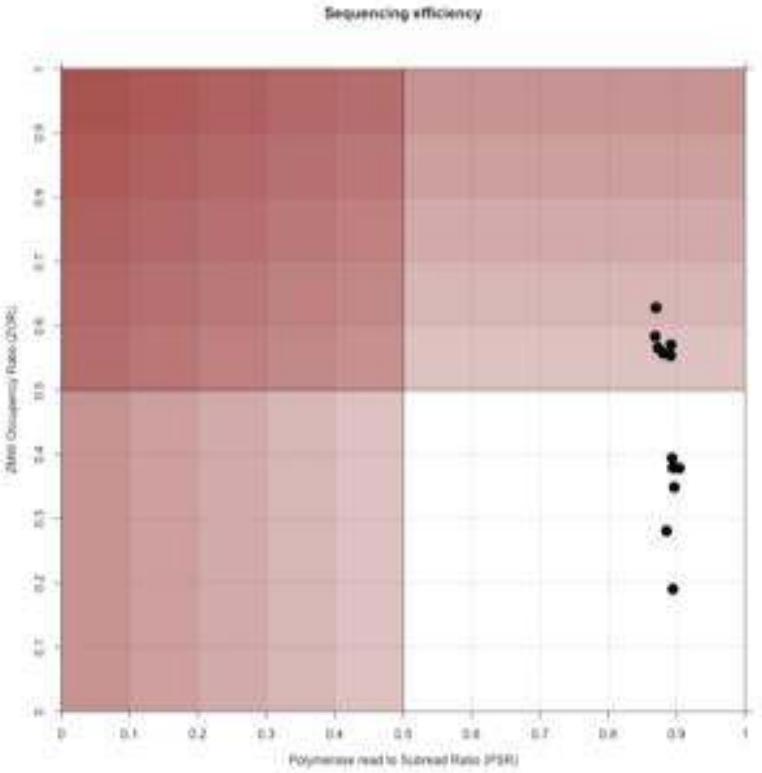


SMRTcell efficiency

fAnaTes1



fSimDia1



"ZOR" = # of ZMWs in HQ data / # of ZMWs per SMRTcell

"PSR" = Sum of longest subreads (nt) / Sum of polymerase reads (nt)

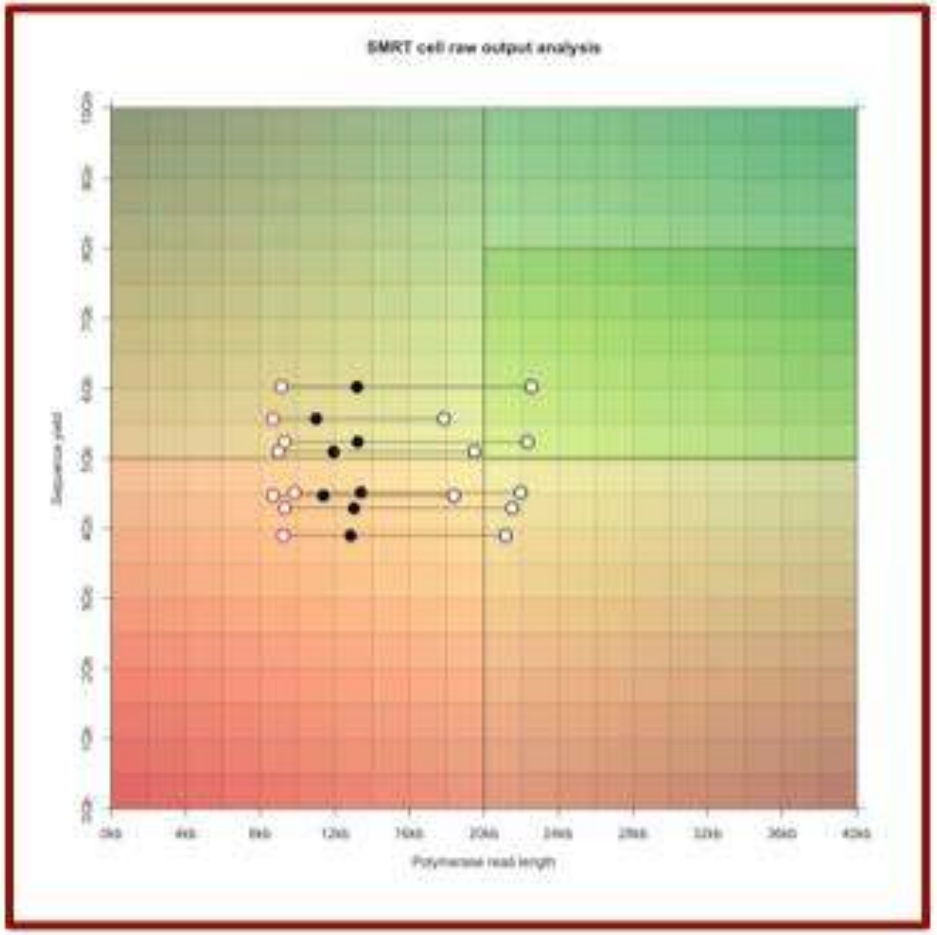
ZOR: How many of the ZMWs in the SMRTcell contained at least one DNA fragment for sequencing. Values >0.5 usually indicate overloading.

PSR: For assembly a PSR=1 would be ideal as in that case the fragment would have been read only once without reaching the end.

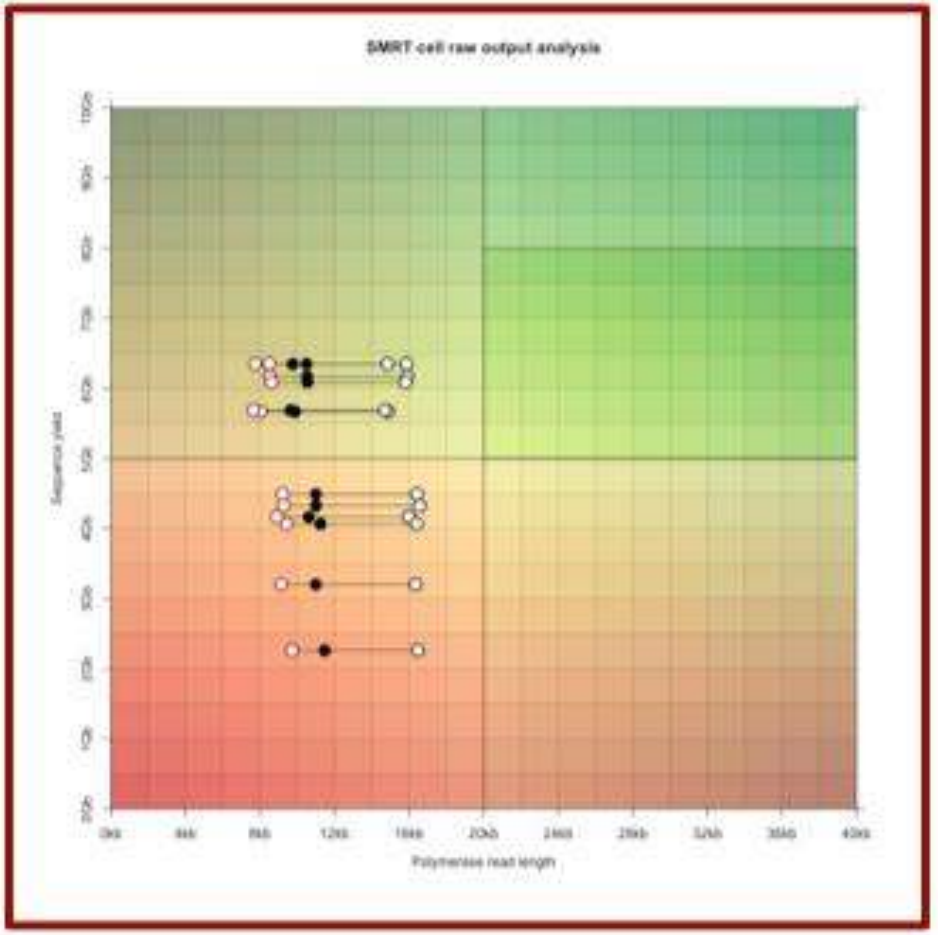
Yield (polymerase)

Median
Mean
N50

fAnaTes1

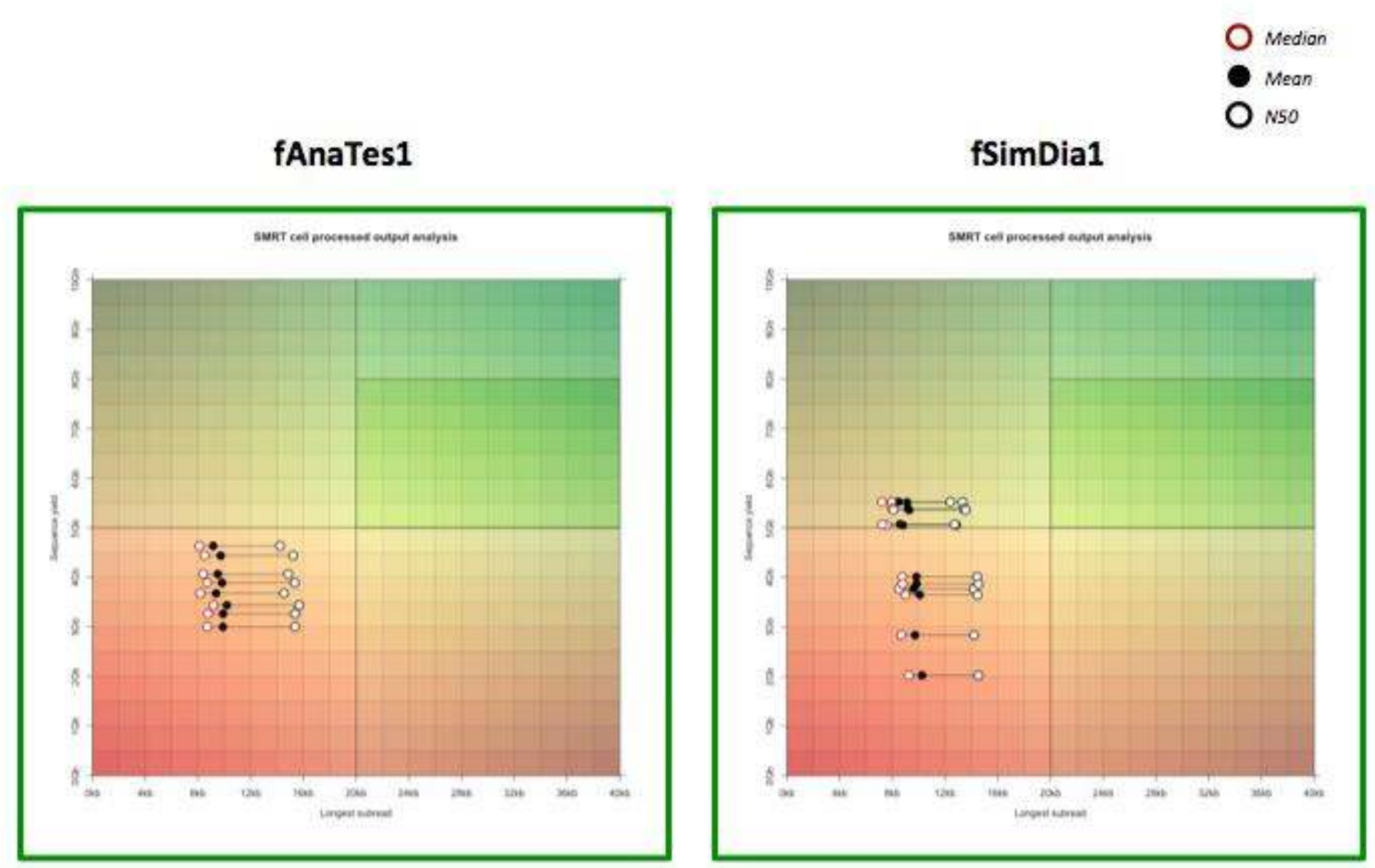


fSimDia1



(Polymerase reads)

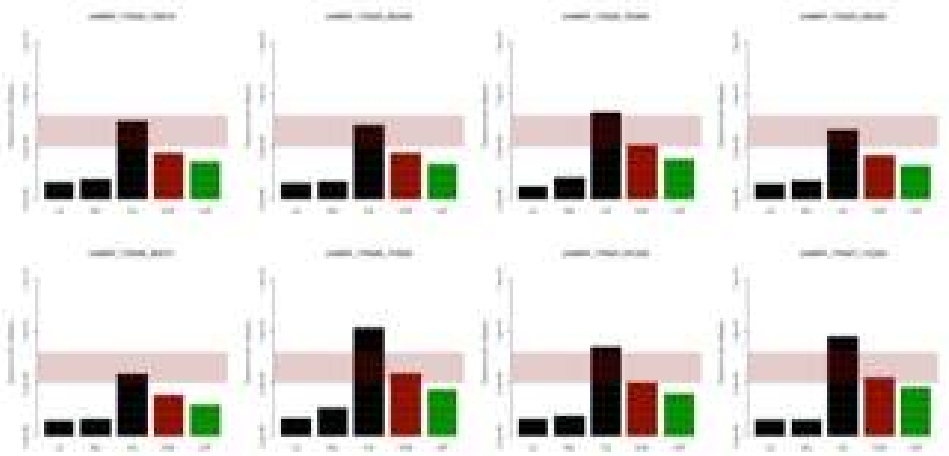
Yield (subreads)



(Longest subreads)

Run yield efficiency

fAnaTes1

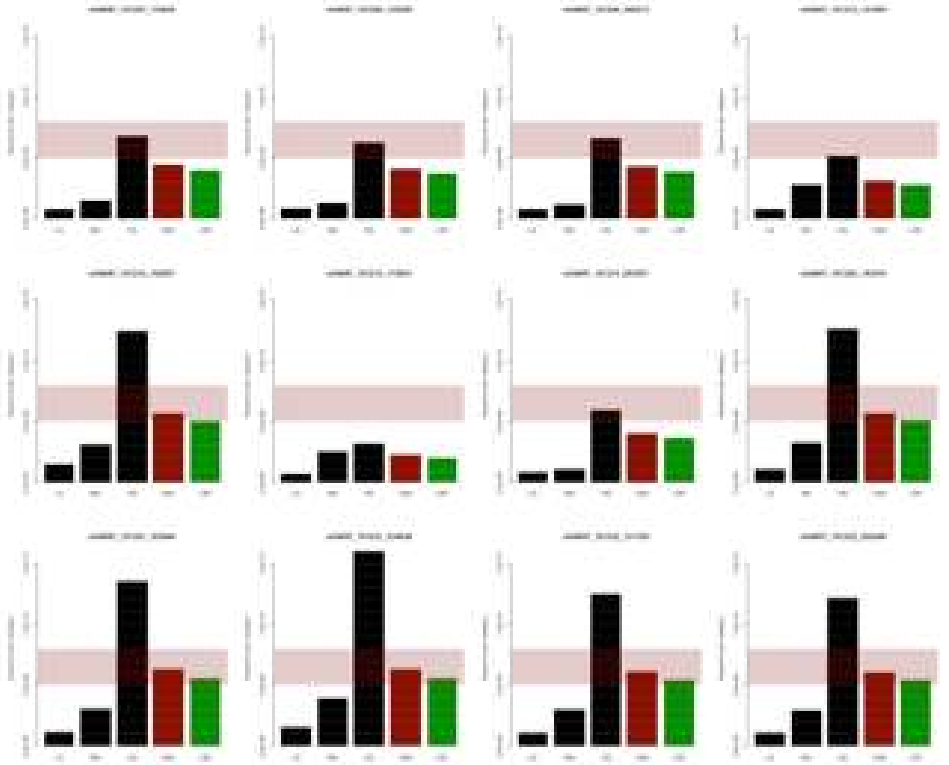


Total raw signal

Sum of high quality regions (i.e. "polymerase reads") in "HQ" data

Sum of longest subread per polymerase read in "HQ" data

fSimDia1

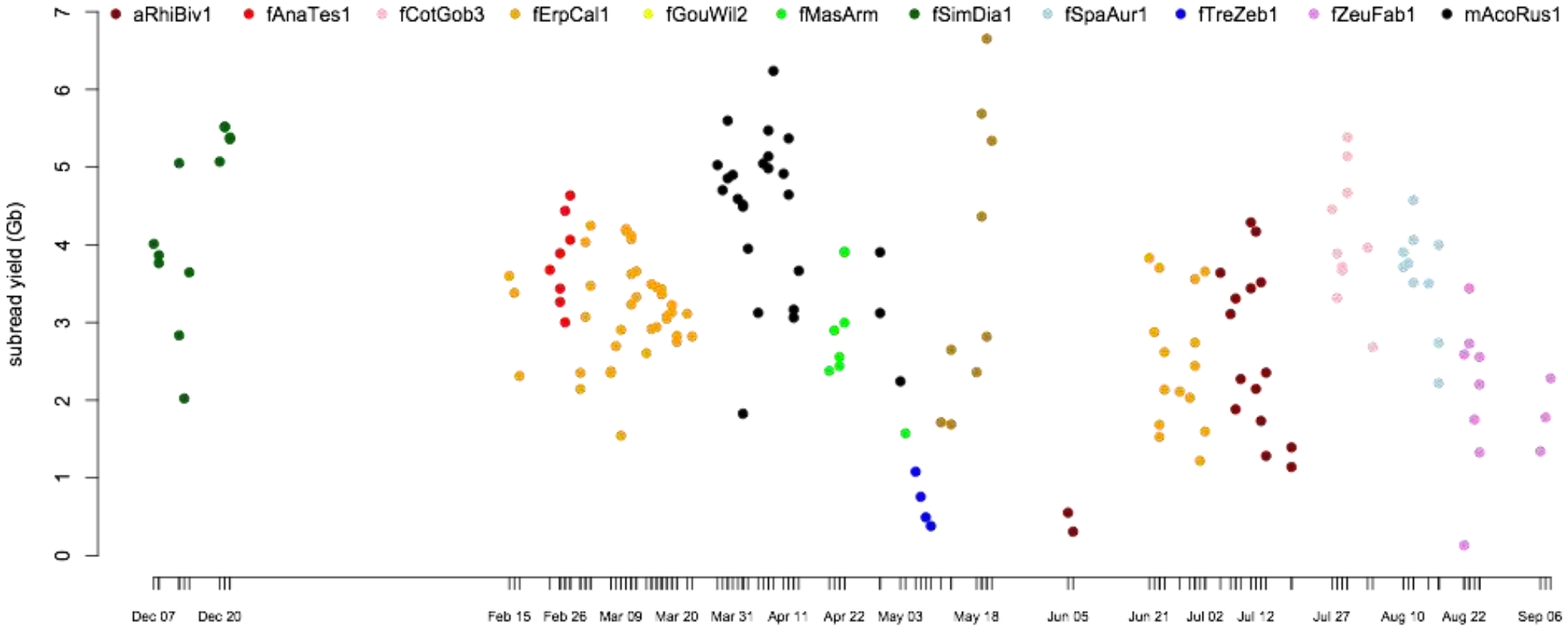


"LQ" = no high quality region found in raw signal

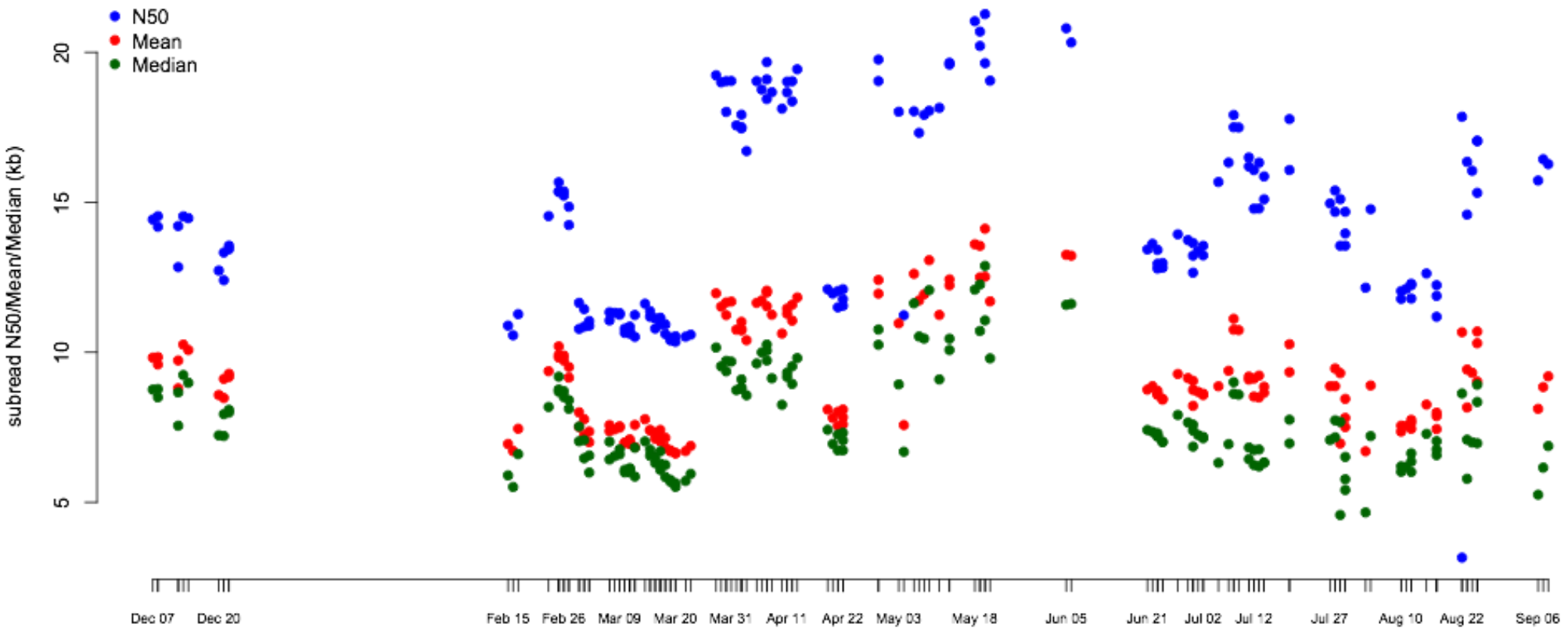
"MQ" = high quality region found but all contained subreads were discarded

"HQ" = high quality region found, at least one accepted subread

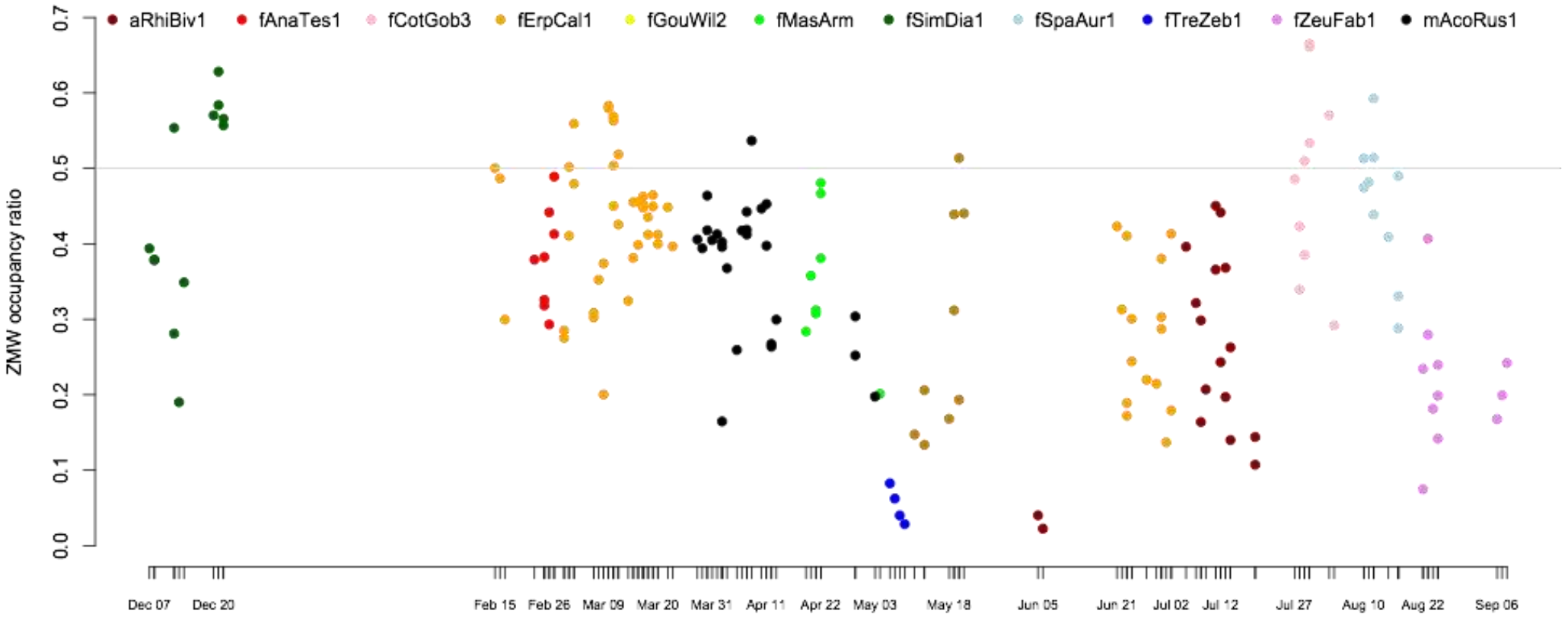
PacBio subread yield over time



PacBio subread length over time

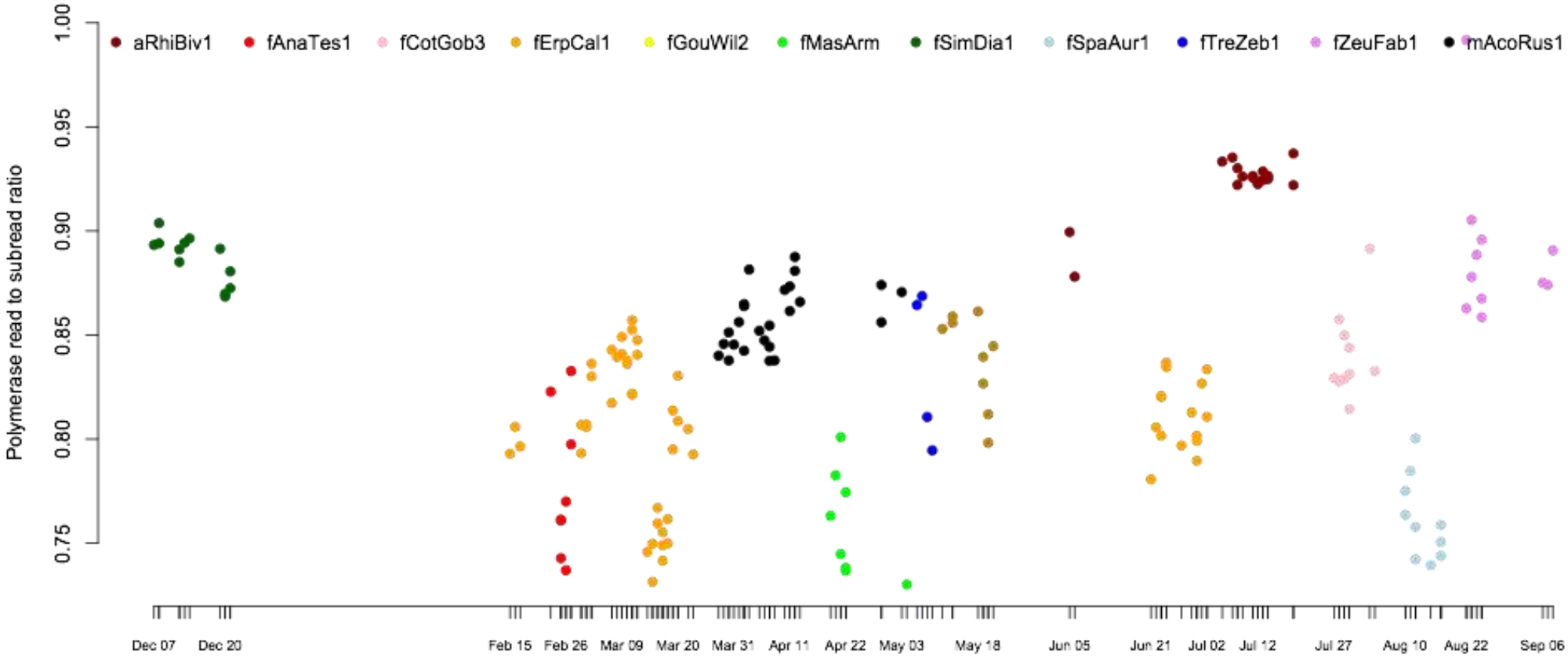


PacBio ZMW occupancy ratio over time



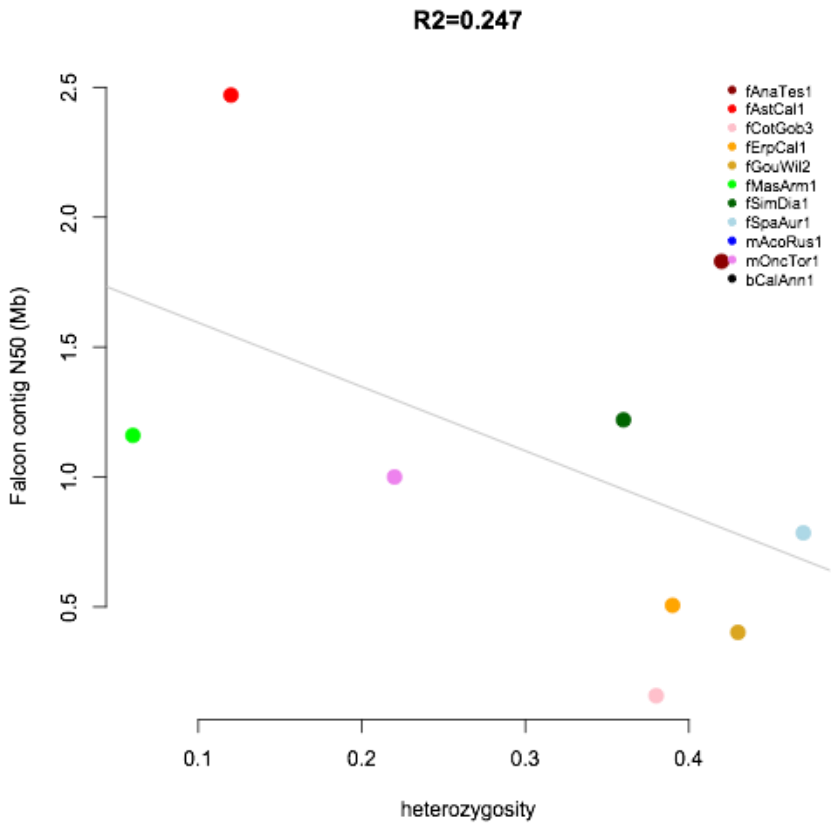
i.e. how many of the ZMWs in the SMRTcell contained at least one DNA fragment for sequencing. Values >0.5 usually indicate overloading.

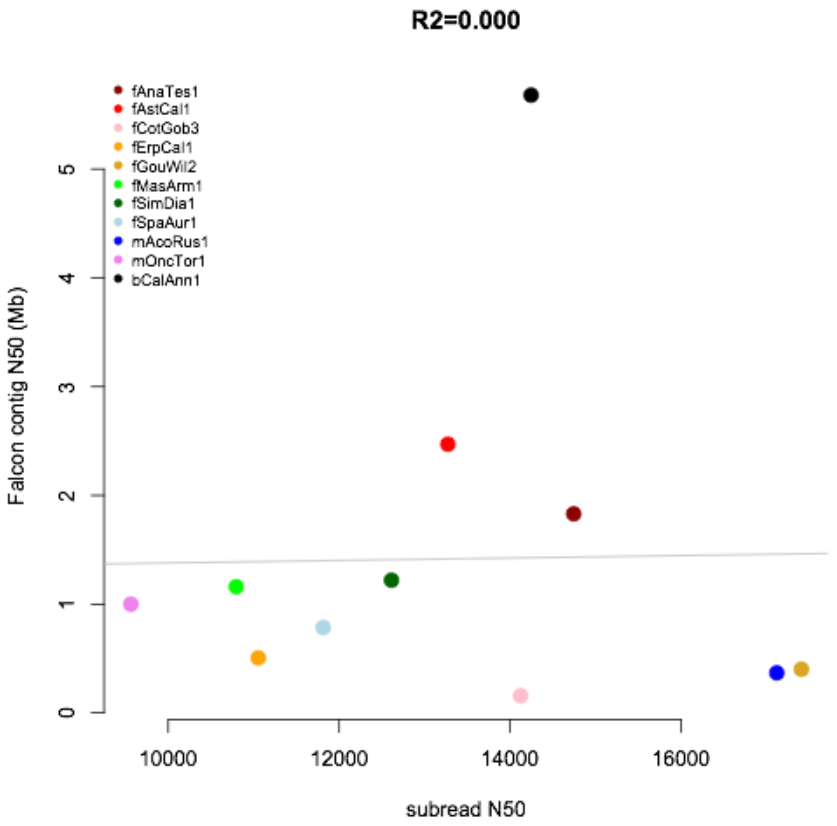
Polymerase to subread ratio over time



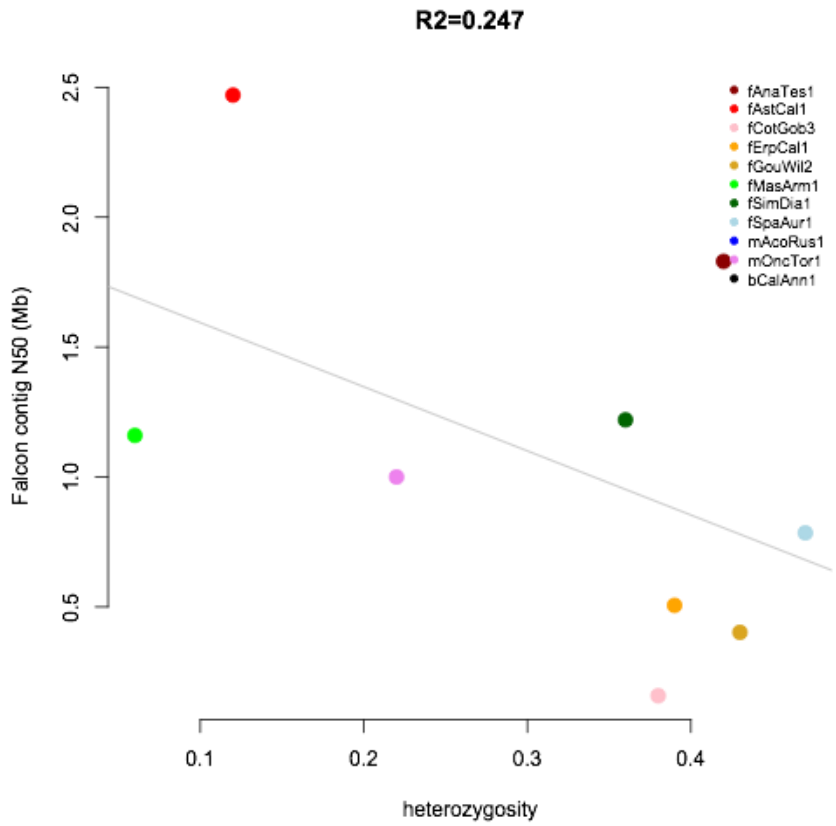
For assembly a PSR=1 would be ideal as in that case the fragment would have been read only o

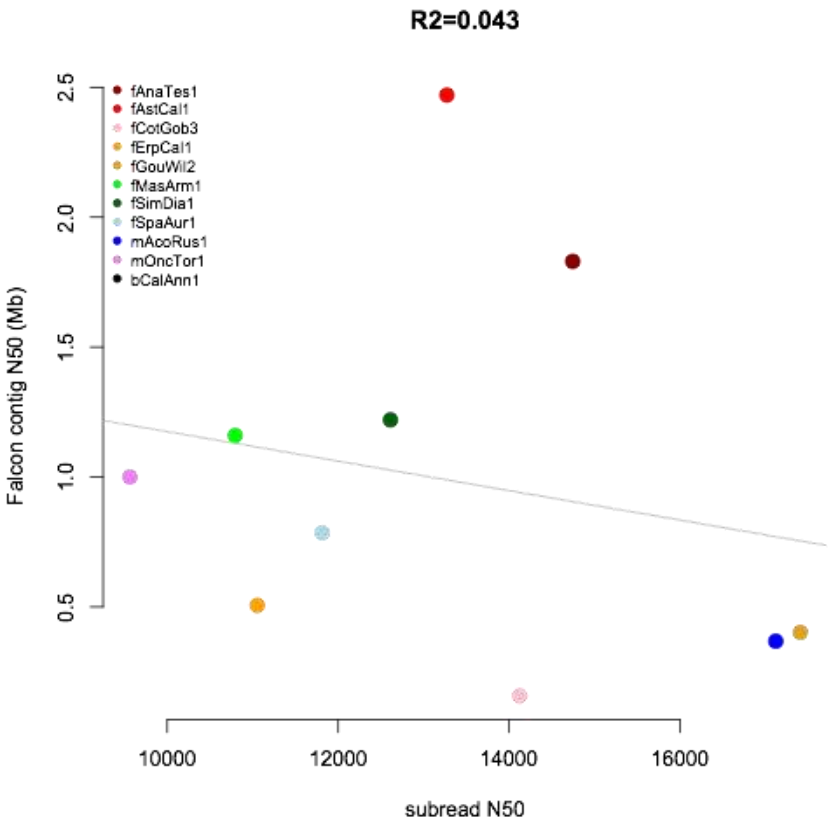
Contig N50 correlations



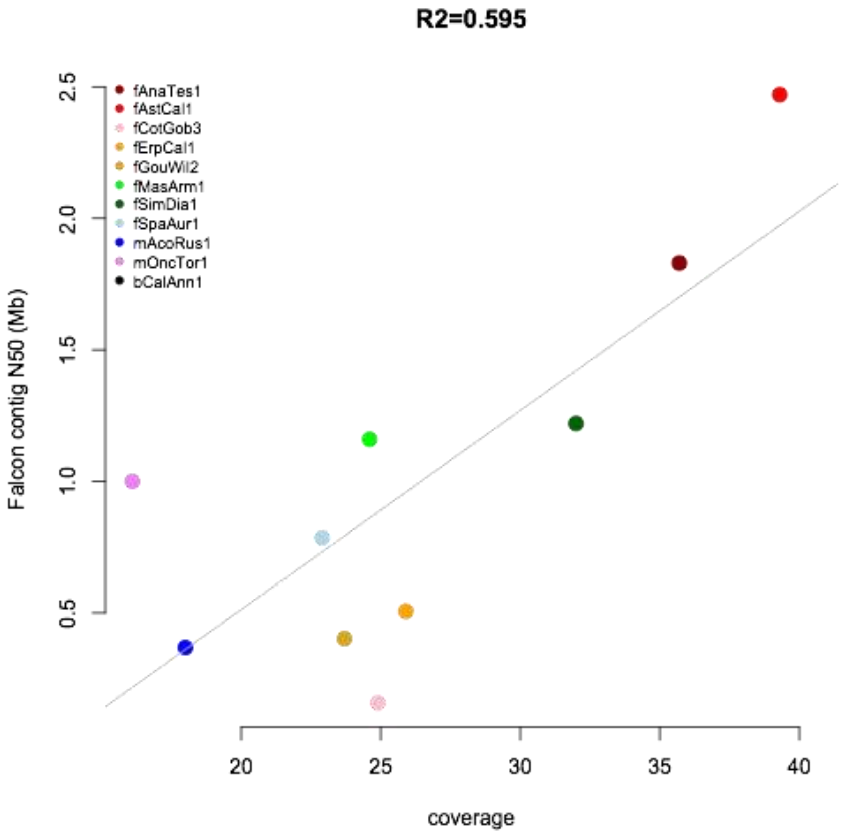


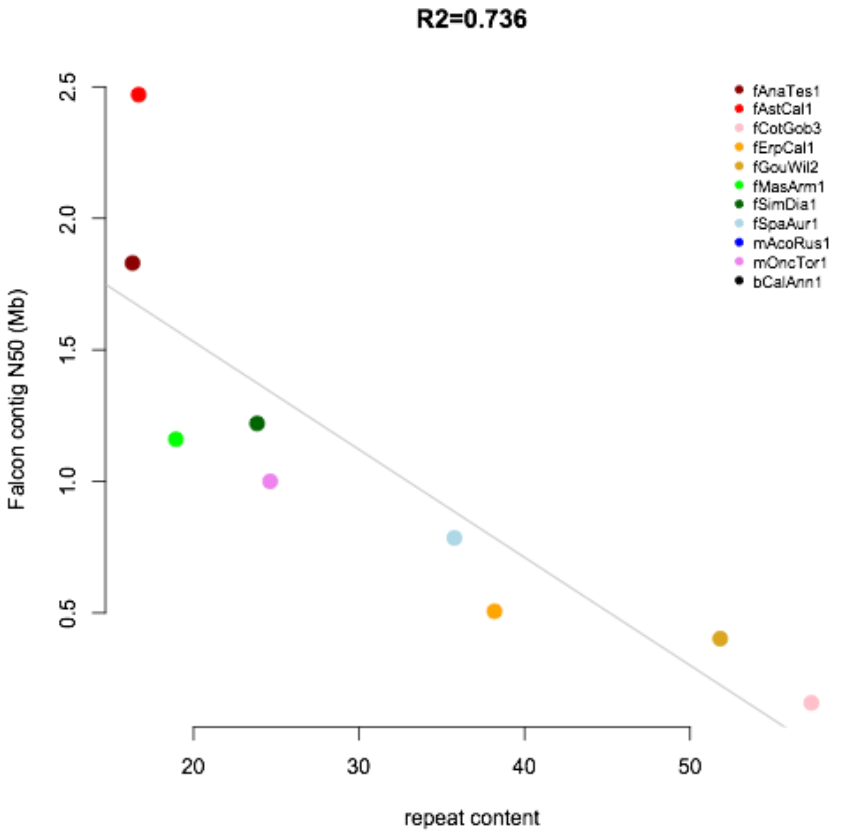
Contig N50 correlations





Contig N50 correlations





- Correlation with coverage quite sensitive to how you calculate coverage, which is sensitive to how one estimates genome size (see later)

Acknowledgements

EBI
Thomas Keane

Notothenioids
Melody Clark
Christina Cheng
Bill Detrich
John Postlethwait
Thomas Desvignes

International Collaboration
Ventakesh Byrappa
Gene Myers

Caecilians
Mark Wilkinson

Rodents
David Thybert

Cichlids
George Turner
Martin Genner
Axel Meyer
Walter Salzburger
Milan Malinsky
Alexandra Tyers
Antonia Ford
Craig Albertson

Cyprinids
Lukas Ruber
Ralf Britz
Braedan McCluskey
Andrew Whiteley
Bill Trevarrow
Uwe Irion
Elisabeth Busch-Nentwich
Christiane Nuesslein Vollhard

Acknowledgements

Wellcome Trust Sanger Institute

Richard Durbin
Iliana Bista
Kerstin Howe
Zemin Ning
William Chow
Petr Danecek
Francesca Giordano
Dirk-Dominik Dolle
Jingtao Li
Eric Miska
Hannes Svardal
Marcus Klarqvist
Mike Quail
Michelle Smith
Karen Oliver
Kim Judge
Craig Corton
Jason Skelton
Carol Scott
Steven Leonard
Daniel Mead

wellcome trust sanger institute

UNIVERSITY OF CAMBRIDGE

http://www.sanger.ac.uk/science/data/vertebrate-genomes-project

Acknowledgements - Collaborators

EBI
Thomas Keane

Notothenioids
Melody Clark
Christina Cheng
Bill Detrich
John Postlethwait
Thomas Desvignes

International Collaboration
Ventakesh Byrappa
Gene Myers

Caecilians
Mark Wilkinson

Rodents
David Thybert

Cichlids
George Turner
Martin Genner
Axel Meyer
Walter Salzburger
Milan Malinsky
Alexandra Tyers
Antonia Ford
Craig Albertson

Cyprinids
Lukas Ruber
Ralf Britz
Braedan McCluskey
Andrew Whiteley
Bill Trevarrow
Uwe Irion
Elisabeth Busch-Nentwich
Christiane Nuesslein
Vollhard

The MHC Diversity in Africa Resource

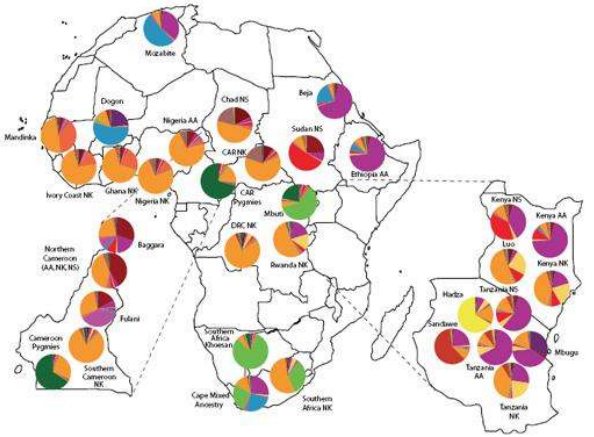
a roadmap to understanding HLA diversity in Africa

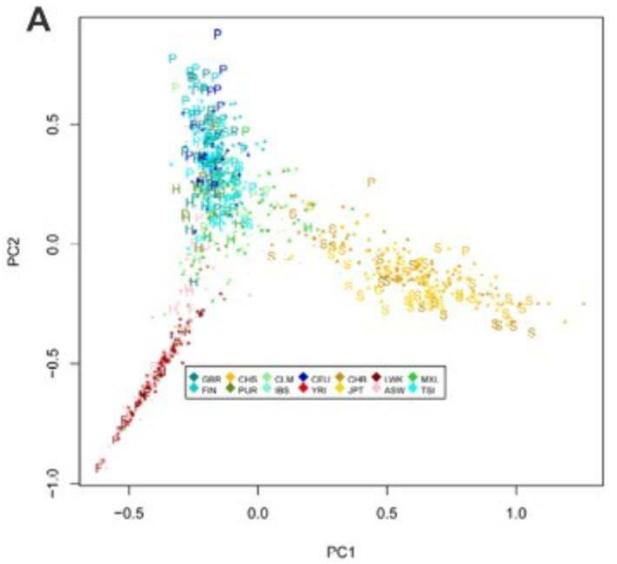
Tarryn Porter and Martin Pollard

PI: Manj Sanhu

Introduction

- Africa - genetically most diverse region in the world
- Evolutionary adaptation to selective pressures
- HLA - one of the most genetically diverse regions
- High HLA diversity relative to Europeans
- Understudied in Africa
  - X sequences in IMGT European
  - X sequences from Africa





Combination of approaches to develop the largest reference resource of African HLA sequences

PacBio based pipeline for typing HLA alleles

- PCR approach (commercial solution, GenDX) long range (3.4-5.8kb)- Class I & Class II
- Sequence- RS-II and Sequel
- Analysis- PacBio software and own purpose built in-house typing software
- Additional validation by commercial sequencing providers

Whole Genome Sequencing of African genome

- PacBio 50x (Sequel) currently 30x
- Illumina 10x Genomics
- BioNano Irys and Saphyr

Imputation Panel for medical genetics

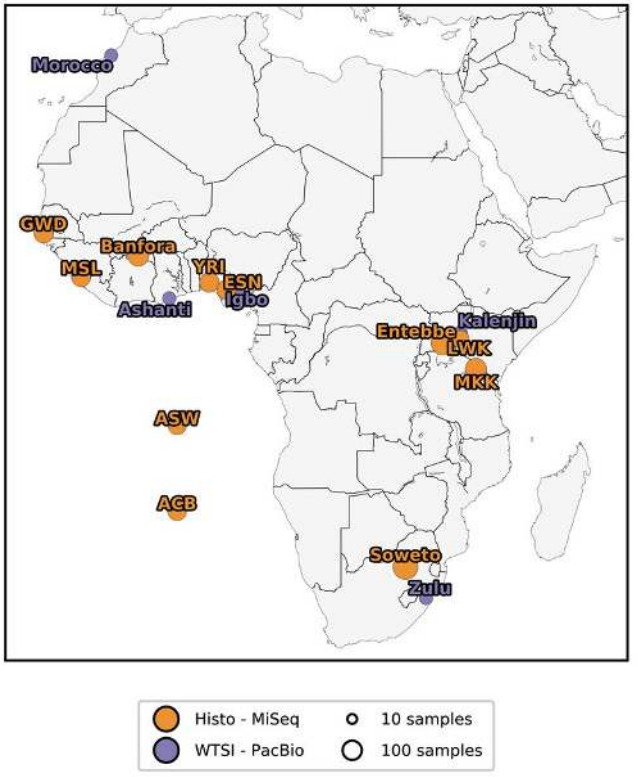
HLA Diversity in Africa panel

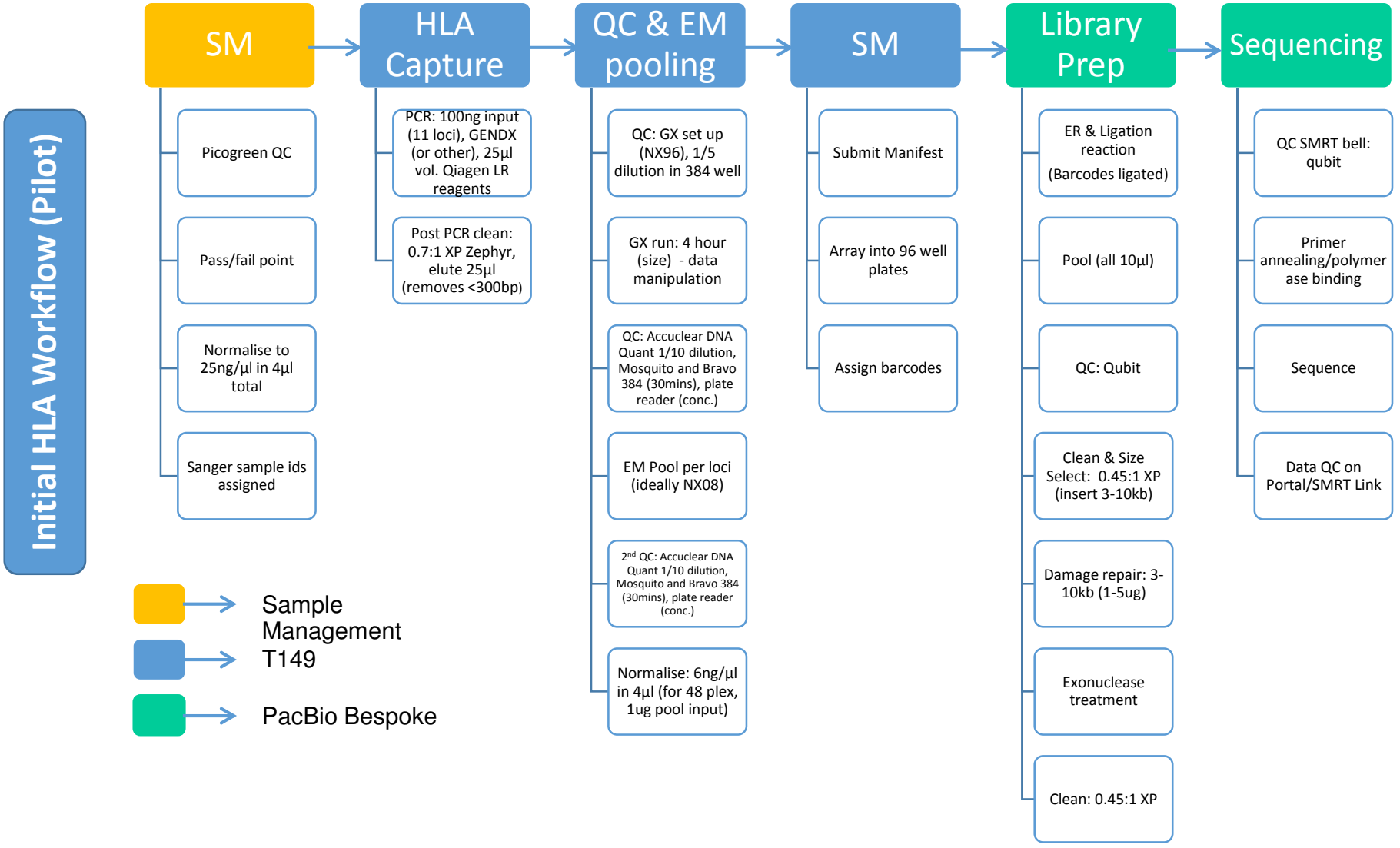
Expected high diversity

- Majority of alleles have an allele frequencies below 10%.
- Important functional & clinically significant alleles observed

1,831 HLA sequences
16 diverse populations

Population	Samples
Urban – Entebbe, Uganda	330
Urban – Banfora, Burkina Faso	167
Urban – Soweto, South Africa	204+196
Morrocan Expat – Netherlands	25
Igbo – Nigeria	25
Kalenjin – Kenya	25
Ashanti – Ghana	25
Zulu – South Africa	25
Yoruba – Ibadan, Nigeria (YRI)	110
Esan – Nigeria (ESN)	99
Maasai – Kinyawa, Kenya (MKK)	166
Gambian – Western Division, The Gambia (GWD)	112
Mende – Sierra Leone (MSL)	84
Americans of African Ancestry, SW USA (ASW)	62
Afro Caribbean – Barbados (ACB)	79
Luhya – Webuye, Kenya (LWK)	97




Initial HLA Workflow (Pilot)
SM
Picogreen QC
Pass/fail point
Normalise to 25ng/µl in 4µl total
Sanger sample ids assigned
HLA Capture
PCR: 100ng input (11 loci), GENDX (or other), 25µl vol. Qiagen LR reagents
Post PCR clean: 0.7:1 XP Zephyr, elute 25µl (removes <300bp)
QC & EM pooling
QC: GX set up (NX96), 1/5 dilution in 384 well
GX run: 4 hour (size) - data manipulation
QC: Accuclear DNA Quant 1/10 dilution, Mosquito and Bravo 384 (30mins), plate reader (conc.)
EM Pool per loci (ideally NX08)
2nd QC: Accuclear DNA Quant 1/10 dilution, Mosquito and Bravo 384 (30mins), plate reader (conc.)
Normalise: 6ng/µl in 4µl (for 48 plex, 1ug pool input)
SM
Submit Manifest
Array into 96 well plates
Assign barcodes
Library Prep
ER & Ligation reaction (Barcodes ligated)
Pool (all 10µl)
QC: Qubit
Clean & Size Select: 0.45:1 XP (insert 3-10kb)
Damage repair: 3-10kb (1-5ug)
Exonuclease treatment
Clean: 0.45:1 XP
Sequencing
QC SMRT bell: qubit
Primer annealing/polymerase binding
Sequence
Data QC on Portal/SMRT Link
Sample Management
T149
PacBio Bespoke

Whole Genome Sequencing on Sequel

Large Insert Libraries	>20 kb Library (8 SMRT cells)	>40 kb Library (1 SMRT cell)
Yield (GB)	1 - 8	8.1
Polymerase Read Avg. (KB)	14.3	13.6
Polymerase Read N50 (KB)	23.8	23.3
Longest Subread Avg. (KB)	11.5	12.8
Longest Subread N50 (KB)	18.8	21.8

- Agarose plug extraction
- 5GB average yield, 8GB with newer SMRT cells
- Magbead loading
- Good results, next steps- diffusion loading and new sequencing chemistry

Whole Genome Sequencing on Sequel

Large Insert Libraries	>20 kb Library (8 SMRT cells)	>40 kb Library (1 SMRT cell)	2.1a	2.1b
Yield (GB)	1 - 8	8.1	2.76	3.85
Polymerase Read Avg. (KB)	14.3	13.6	19405	15505
Polymerase Read N50 (KB)	23.8	23.3	33250	27750
Longest Subread Avg. (KB)	11.5	12.8	18044	14613
Longest Subread N50 (KB)	18.8	21.8	30250	25750

- Agarose plug extraction
- 5GB average yield, 8GB with newer SMRT cells
- Magbead loading
- Good results, next steps- diffusion loading and new sequencing chemistry

Challenges and Observations

HLA PacBio PCR pipeline

- Current commercial method £££££ (primers and barcoded adapters)
- Variable PCR yields and sizes can create downstream equimolar pooling difficulties
- Automation possible but challenging
- Chimeras
- Highly diverse, understudied populations = many novel alleles (Repeat PCR as validation)

Issues largely overcome by barcoding at PCR step (primer design and optimisation required)

DNA input and quality

- Multiple collaborators often with limited DNA available
- Potential viral risk requires CL3 extraction
- Old or degraded DNA not suitable for long read platforms

25 Genomes Project

Overview

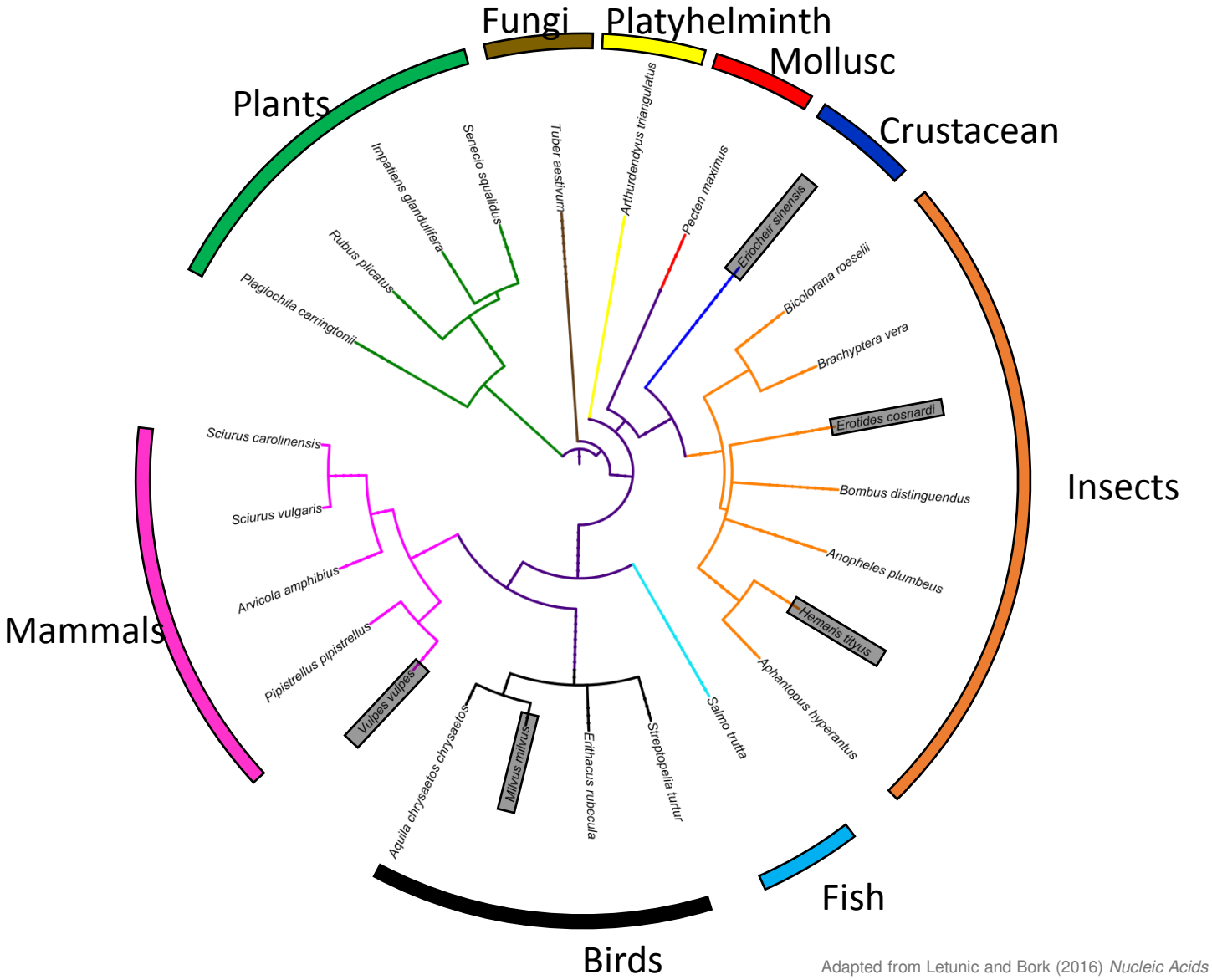
- 25 reference quality genomes (32-48Gb)
- Combined sequencing technologies
  - 50x Illumina plus 10x Genomics
  - 50x Long read
- Varied organisms
  - Vertebrates
    - Mammals
    - Birds
    - Reptiles
  - Invertebrates
    - Insects
    - Molluscs
  - Plants
  - Fungi

20+5 Species decided in house

Cryptic	Dangerous	Iconic	Flourishing	Floundering
Brown Trout	British Mosquito	Golden Eagle	Roesel's Bush-cricket	Red Squirrel
Carrington's Featherwort	King Scallop*	Red Fox	Oxford Ragwort	Water Vole
Narrow-bordered Bee Hawkmoth	Indian Balsam	Blackberry	Red kite	Turtle Dove
Common Pipistrelle	New Zealand Flatworm*	European robin	Ringlet butterfly	Cosnard's Net-winged Beetle*
Summer truffle	Chinese Mitten Crab	Great yellow bumblebee	Grey squirrel	Northern February red stonefly

Source not yet in contact
Sample at Sanger
Source agreed

*subject to change



Adapted from Letunic and Bork (2016) *Nucleic Acids Res* **doi: 10.1093/nar/gkw290**

I'm a Scientist Get me out of here

Search site

LOG IN

Home | Teachers | Scientists | 25 Genomes | June 2017 Zones | About | Contact

Students talk with scientists

A free online event where school students meet and interact with scientists. It's an **X Factor-style competition** between scientists, where students are the judges.

Students challenge the scientists over fast-paced online text-based live **CHAT**s. They **ASK** the scientists anything they want, and **VOTE** for their favourite scientist to win a prize of £500 to communicate their work with the public.

Ask Chat Vote

Login to I'm a Scientist UK

Log in using the username and password we sent you!

Username:

Your username

Password:

Your password

LOG IN

Lost your password?

The Next Event: 6 - 17 November 2017

IASGMOOH

Cryptic
Abyssal Grenadier
Baltic Clam
Brachiopod
Common Starfish
Orkney Vole
Scottish Crossbill
Snake Pipefish
The Naval Shipworm
Twisted-Wing Fly

Dangerous
Asian Hornet
Blue-Green Algae
Daubenton'S Bat
Giant Hogweed
Leathery Sea Squirt
Oak Apple Gallwasp
Turkey Oak

Iconic
lesser-spotted catshark [mermaids purse]
Barn Owl
Common Crane
Emperor Dragonfly
Hazel Dormouse
Jeremy, Brown Garden Snail
Orange-Tailed Mining Bee
Scotch Thistle
St Kilda Wren

Floundering
Barbastelle Bat
Eurasian Otter
European Flat Oyster
Glow Worm
Lundy Cabbage
Scottish Wildcat
Spotted Flycatcher
Strapwort
The Scaly Cricket

Flourishing
Beaver
Canada Goose
Cirl Bunting
Danish Scurvygrass
Fen Raft Spider
Pill Millipede
Small Red-Eyed Damselfly
Tree Lichen

25 Genomes

Search site

LOG IN

Home | Teachers | Species Champions

25 Genomes: Mapping Nature's Code

The Wellcome Sanger Institute is one of the world's leading genomics research centres. It was the UK home for the Human Genome Project – which revealed the entire DNA code of life.

In celebration of its 25th Anniversary, the Institute is going to decode the DNA of 25 UK species for the first time…**and you can vote to decide 5 of them!**

From 6th November – 8th December, a team of scientists and wildlife experts will each champion a species they think should have its entire DNA sequence decoded.

We need you to…

- read their pitches
- get to know the species
- vote for the ones whose genetic code you want to see fully unravelled for the first time

The winning species in each Zone will have it's entire DNA sequence decoded by the Wellcome Sanger Institute!

Sign up for updates

The public vote opens on the **6th November**, when you can vote for the species you want to see have its entire DNA sequenced decoded.

Add your email to receive an update when you can vote for your favourite species.

Enter Email

Confirm Email

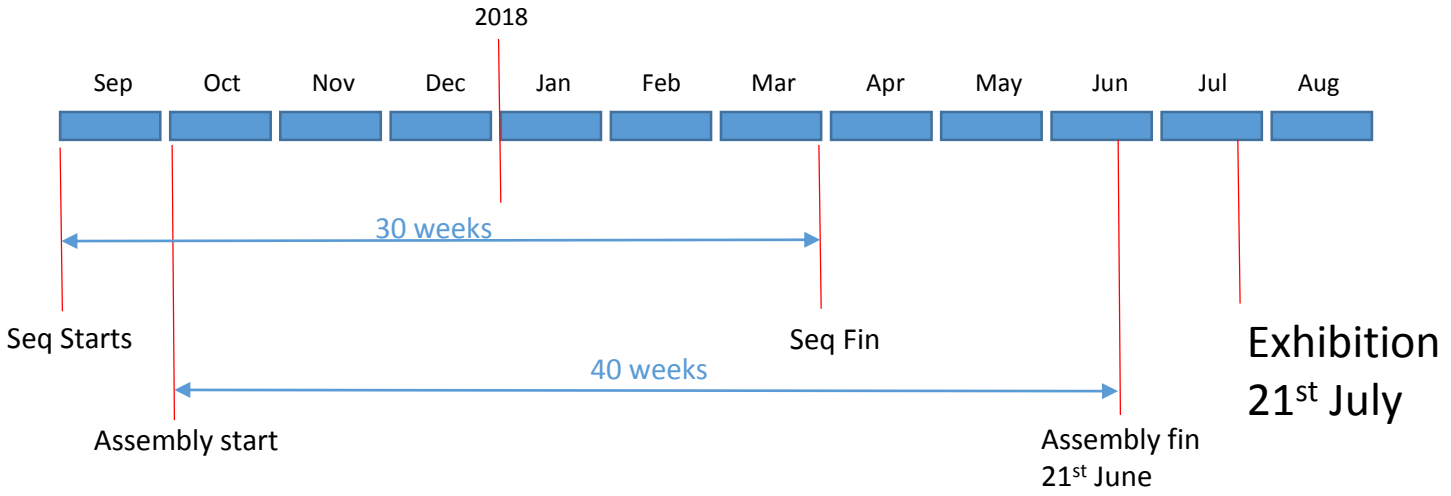
If you'd like to receive updates on the competition give us your email address, and we'll let you know when voting is open.

SUBMIT

Zone News

Other Activities

- Public vote to determine 5 species
  - IASGMOOH
  - October-December
  - Species Champions
- Citizen Science for Annotation
  - IRIS
  - App
- Springwatch
  - Sequencing in the field
- Exhibition
  - Conference centre


2018
Sep
Oct
Nov
Dec
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
30 weeks
Seq Starts
Seq Fin
40 weeks
Exhibition
21st July
Assembly start
Assembly fin
21st June

10x GENOMICS®

buglife

NOTTINGHAM TRENT UNIVERSITY

Centre for Ecology & Hydrology
NATURAL ENVIRONMENT RESEARCH COUNCIL

The James Hutton Institute

ACES

Bumblebee Conservation Trust

eagle genomics

rspb

wildwood

APHA SCIENTIFIC

VILLE DE GENÈVE
SITE OFFICIEL

THE UNIVERSITY of EDINBURGH
The Royal (Dick) School of Veterinary Studies

Cefas

nurture

UNIVERSITY OF LINCOLN

THE wildlife TRUSTS
Lancashire, Manchester & N Merseyside

Mycorrhizal Systems Ltd

Royal Botanic Garden Edinburgh

NATURAL HISTORY MUSEUM

EMBL-EBI

PACBIO

UCD DUBLIN

LSTM
LIVERPOOL SCHOOL OF TROPICAL MEDICINE

With Special Thanks to:

- Tarryn Porter

- Karen Oliver

- Craig Corton

- Jason Skelton

- Shane McCarthy

- Daniel Mead