wellcome centre human genetics

NUFFIELD DEPARTMENT of MEDICINE

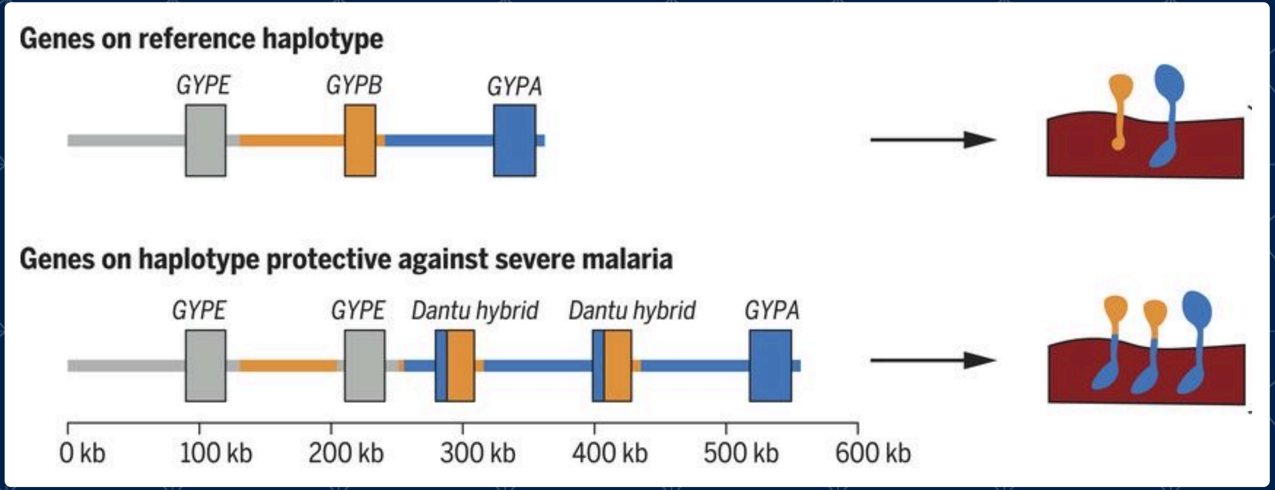
UNIVERSITY OF OXFORD

Long read accuracy and genome assembly

Gavin Band

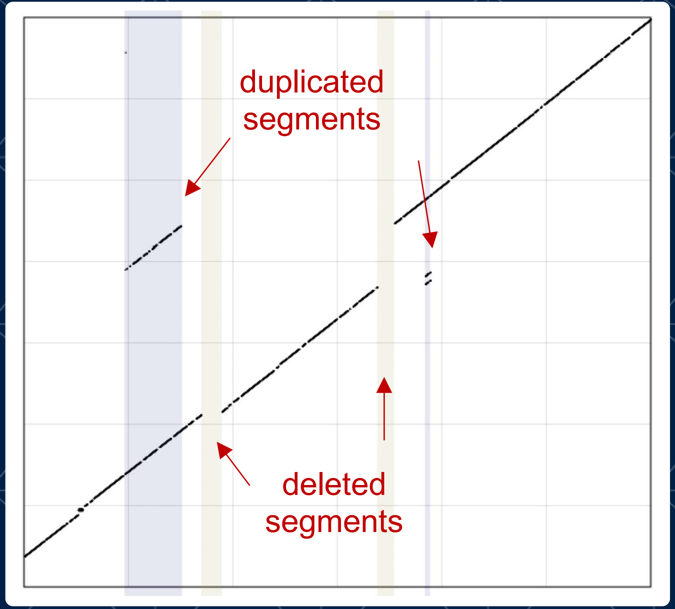
LR CASe Detectives meeting
Septembr 7th 2023

Motivation: structural variation in hosts and pathogens



“Resistance to malaria through structural variation of red blood cell invasion receptors”
2017

“Malaria protection due to sickle haemoglobin depends on parasite genotype”
2021



Another *P.falciparum* genome

P.falciparum reference genome

Many questions

- What is the structure of the variant?
- What is their functional effect?
- What is their phenotypic impact?
- How are they evolving?
- What other variants segregate?
- How can we genotype them?

Talk outline

1. How accurate are recent long-read platforms?

2. Two genome assembly applications

The HV31 omniome project: data

Genomic data:

- Illumina and MGI short-read data, to ~200x

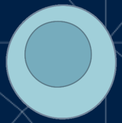
- PacBio ‘continuous long reads’ (Sequel II), to ~35x
- PacBio ‘HiFi’ reads (Sequel II and IIe) to ~24x
- **New!!** PacBio ‘HiFi’ reads (Revio), to ~57x

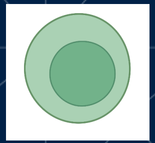
- Oxford Nanopore Technologies R9.4.1, to ~63x
- **New!!** ONT R10.4.1 data, to ~69x

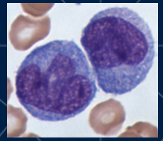
- 10X linked-reads (to ~40x)
- MGI stLFR linked reads

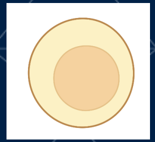
- BioNano optical mapping, to ~150x coverage by fragments

Functional data:


(B cells)


(T helper cells)


(Monocytes)


(Cytotoxic T)

- RNA-seq for gene expression
- ATAC-seq for chromatin accessibility
- CHiP-seq for histone modifications

- Methylation (from long read datasets)

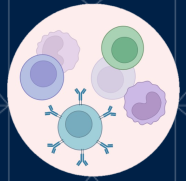
All data is, or will be available through the EGA: **EGAS00001005046**

Image credit: Dr Graham Beards - Own work, CC BY-SA 3.0

Andrew Brown
Julian Knight lab
Connor Davison

PBMCs

stored in foetal calf serum



↓

DNA extraction
Qiagen Gentra Puregene Kit

↙ ↘

Simon Mayes
Philipp Reschender
Tonya McSherry
Rosemary Sinclair-Dokos

Oxford NANOPORE Technologies

Sequencing using
5 x Promethion flowcells
To approx 67x depth

PacBio

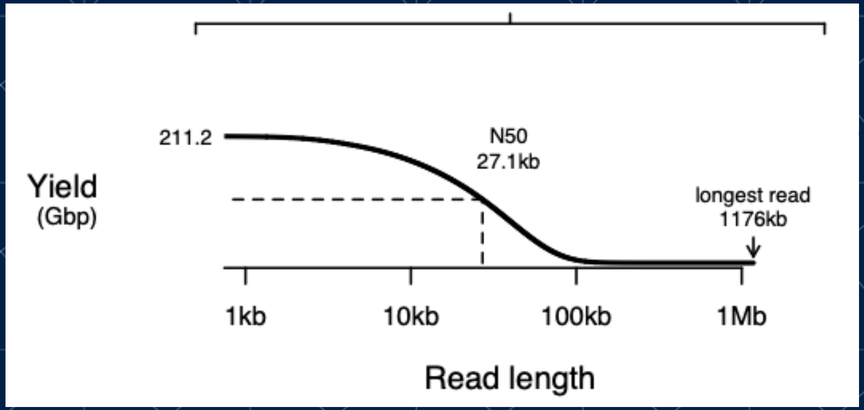
Riki Aydeniz
Eirini Maria Lampraki
Mike Eberle
Cillian Nolan

Sequencing using
2 x Revio SMRT cell
To approx 60x depth

Analysis by our team @ Oxford

Read length comparison

Simplex reads

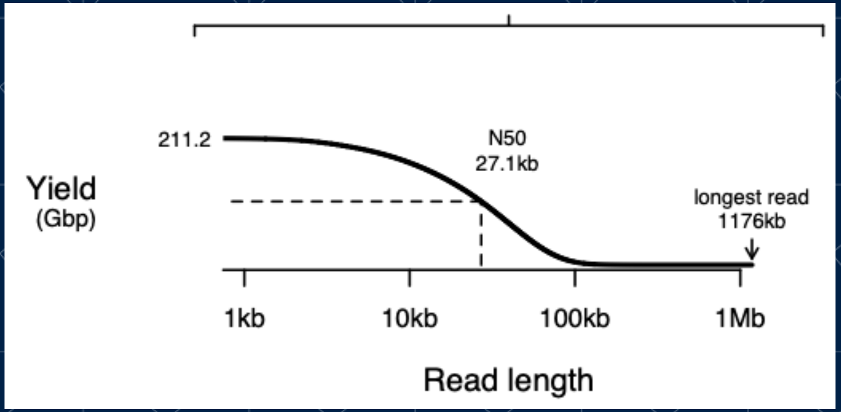


In this expt, most nanopore reads were 1-100kb long…

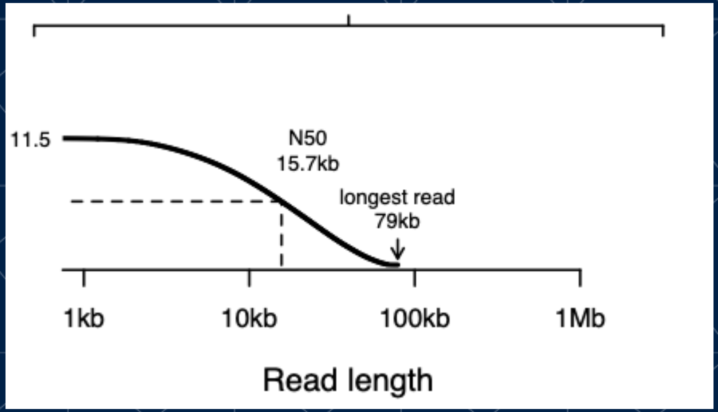
Nanopore R10.4.1

Read length comparison

Simplex reads



Duplex reads
about 5% of total reads



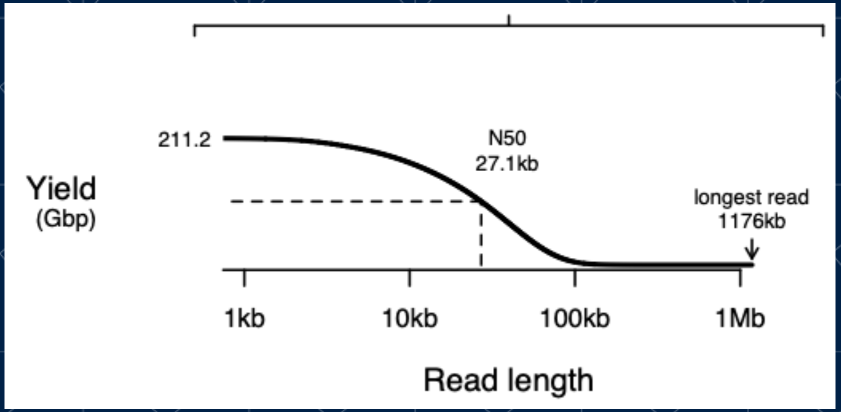
Nanopore R10.4.1

In this expt, most nanopore reads were 1-100kb long…

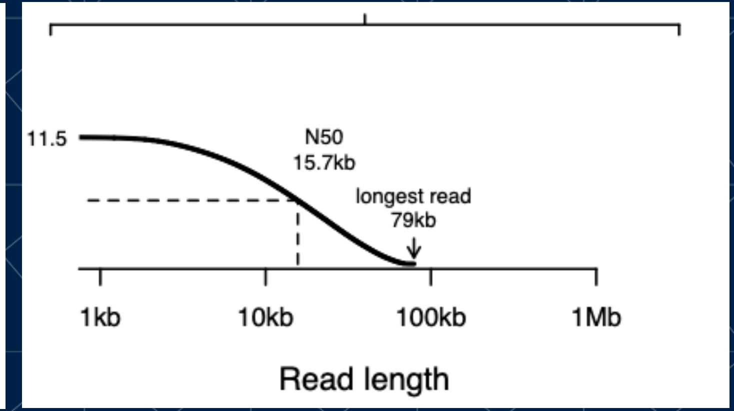
and duplex reads were slightly shorter

Read length comparison

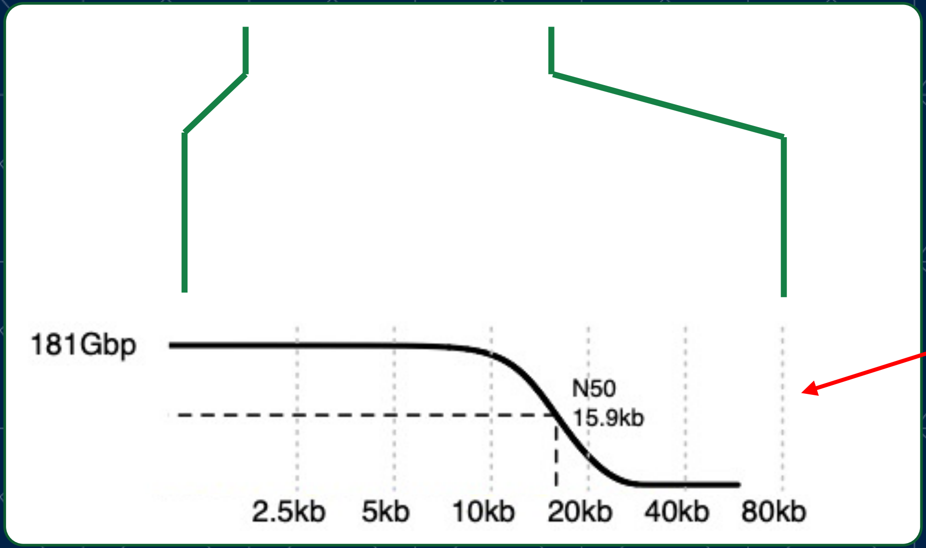
Simplex reads



Duplex reads
about 5% of total reads



Nanopore R10.4.1



Pacbio Revio

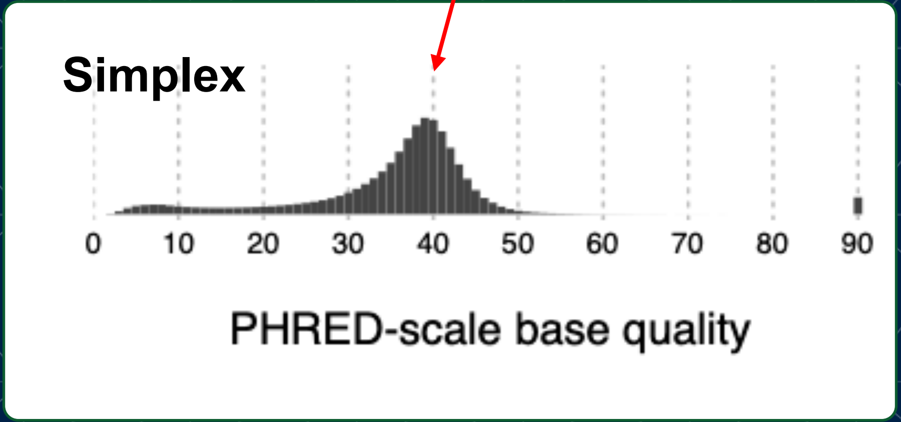
Pacbio reads are **shorter,** on average than nanopore reads.

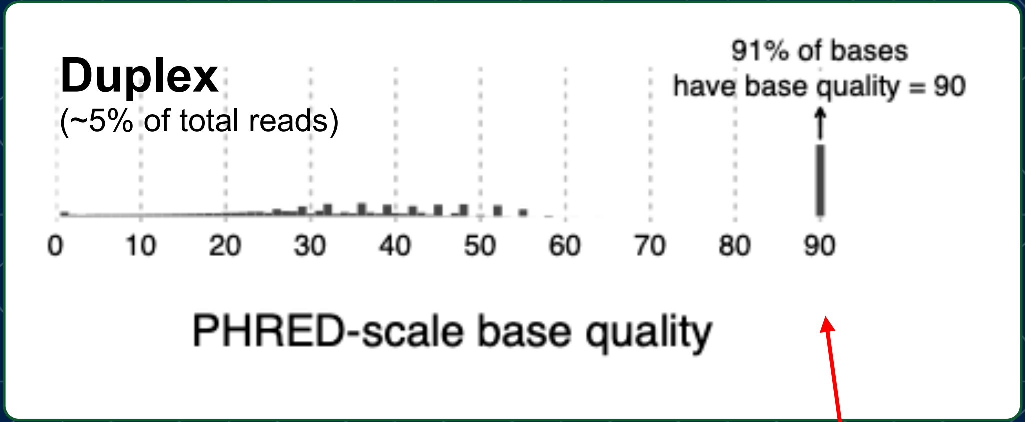
About 10-20kb long

Base quality comparison

E.g. bq 40 predicts 1 error per 10,000 bases

Nanopore
R10.4.1

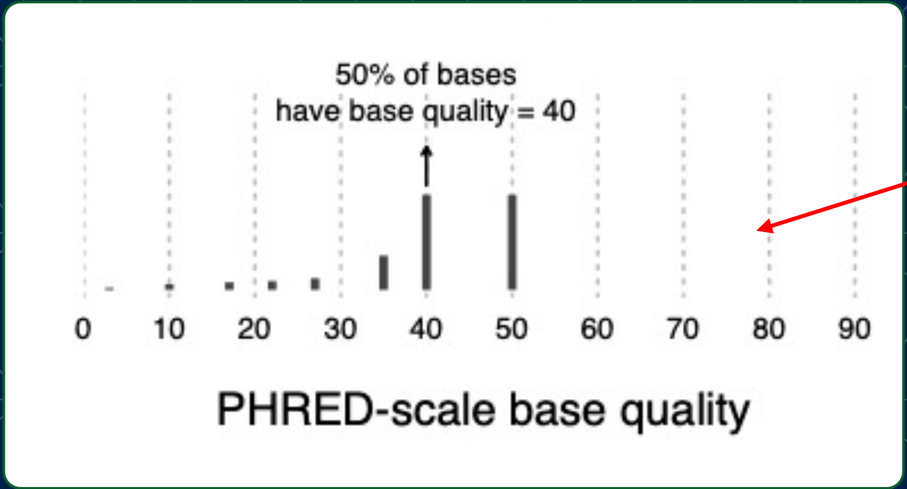




!

bq 90 predicts 1 error per billion bases

Pacbio
Revio



Pacbio base qualities are similar to nanopore simplex but compressed into discrete set of values to make the files smaller.

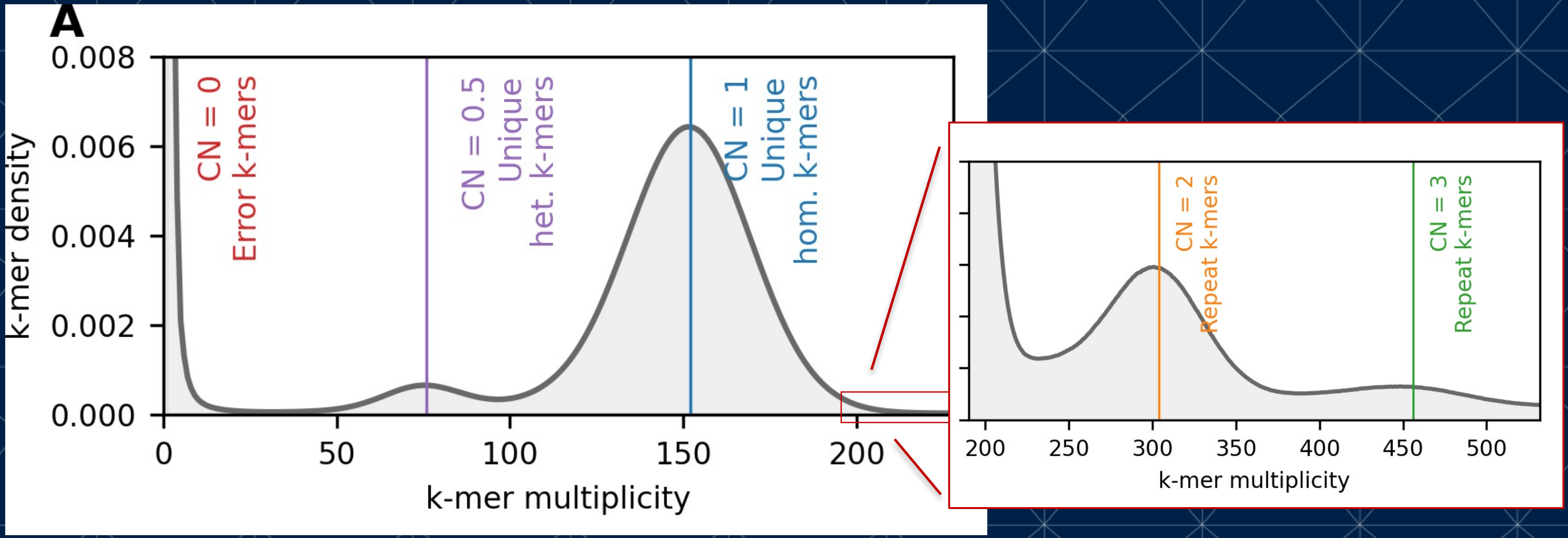
Two ways to measure error rates

1. Measure *kmer accuracy* using a set of known true kmers

2. Measure base accuracy based on alignment to a reference

Measuring kmer accuracy

Method: learn the set of true HV31 k-mers from short reads…

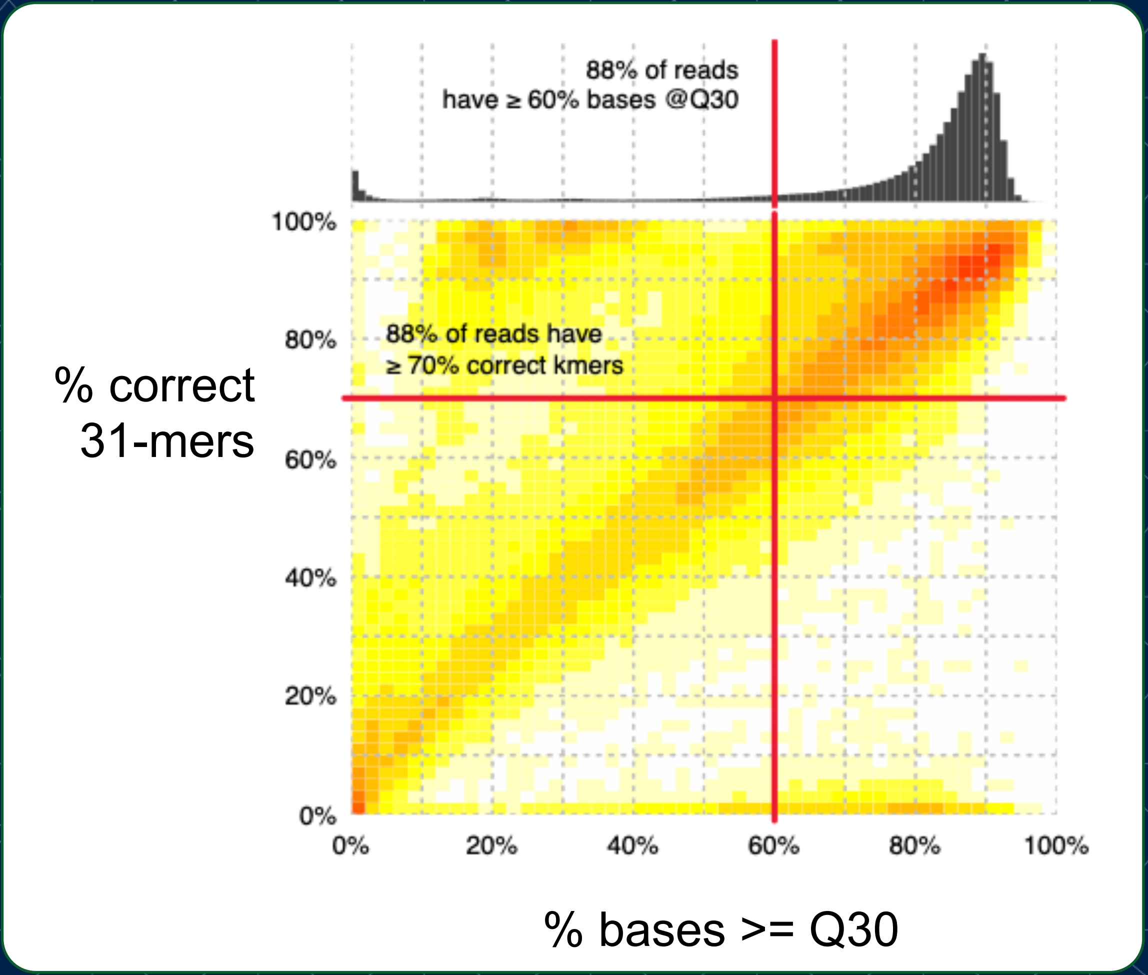


Histogram of k-mer multiplicity observed in Illumina, MGI, MGI CoolMPS, 10X and Sequel II data.

…and for each long read, count the number of true HV31 kmers (k=31)

* This histogram is mainly based on short reads, but also includes the 2019 Sequel II data.

K-mer accuracy vs. predicted accuracy



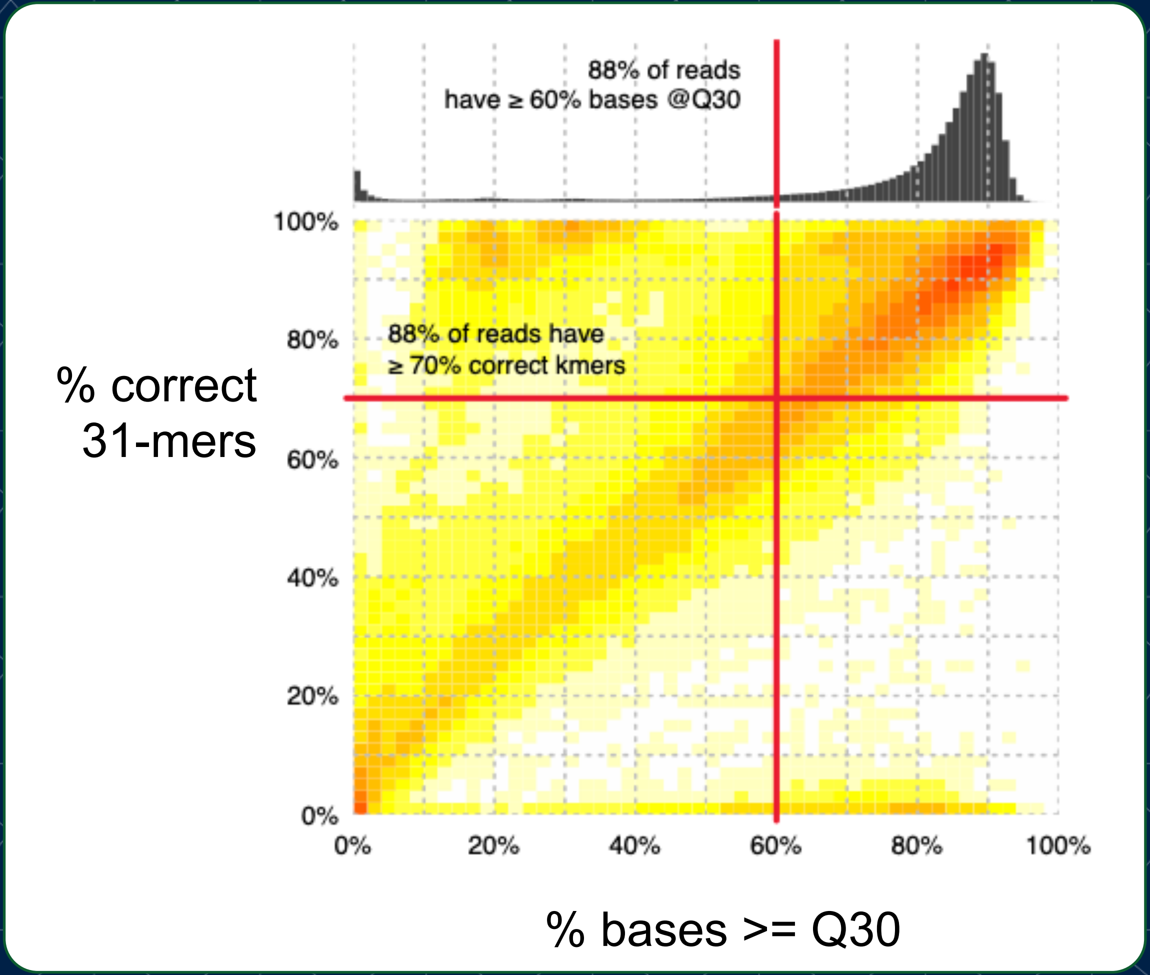
Nanopore R10.4.1 (simplex)

Nanopore simplex has a roughly linear relationship between the quality predicted by base quality scores (x axis) and the observed quality (y axis)

…as measured by accurate kmer rates.

Still about 20% of reads are poor quality.

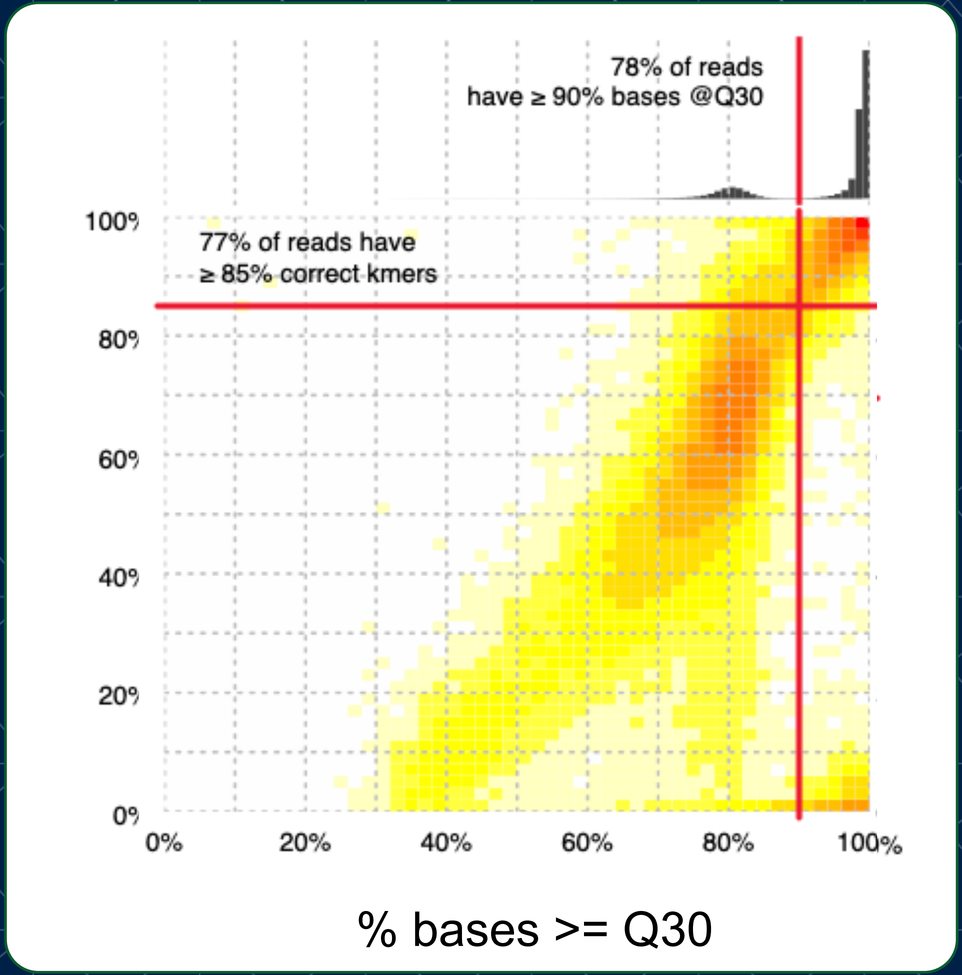
K-mer accuracy vs. predicted accuracy



88% of reads have ≥ 60% bases @Q30

88% of reads have ≥ 70% correct kmers

Nanopore R10.4.1 (simplex)

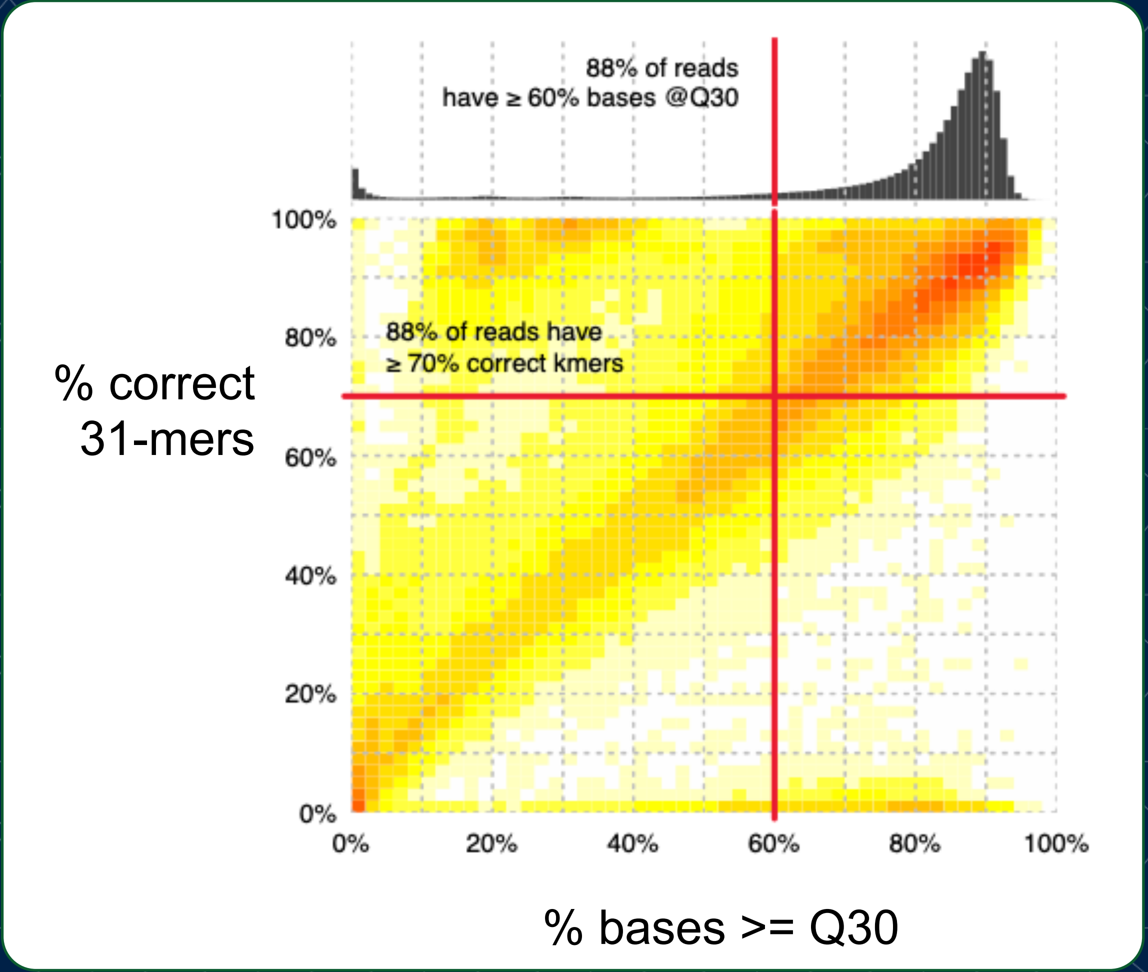


78% of reads have ≥ 90% bases @Q30

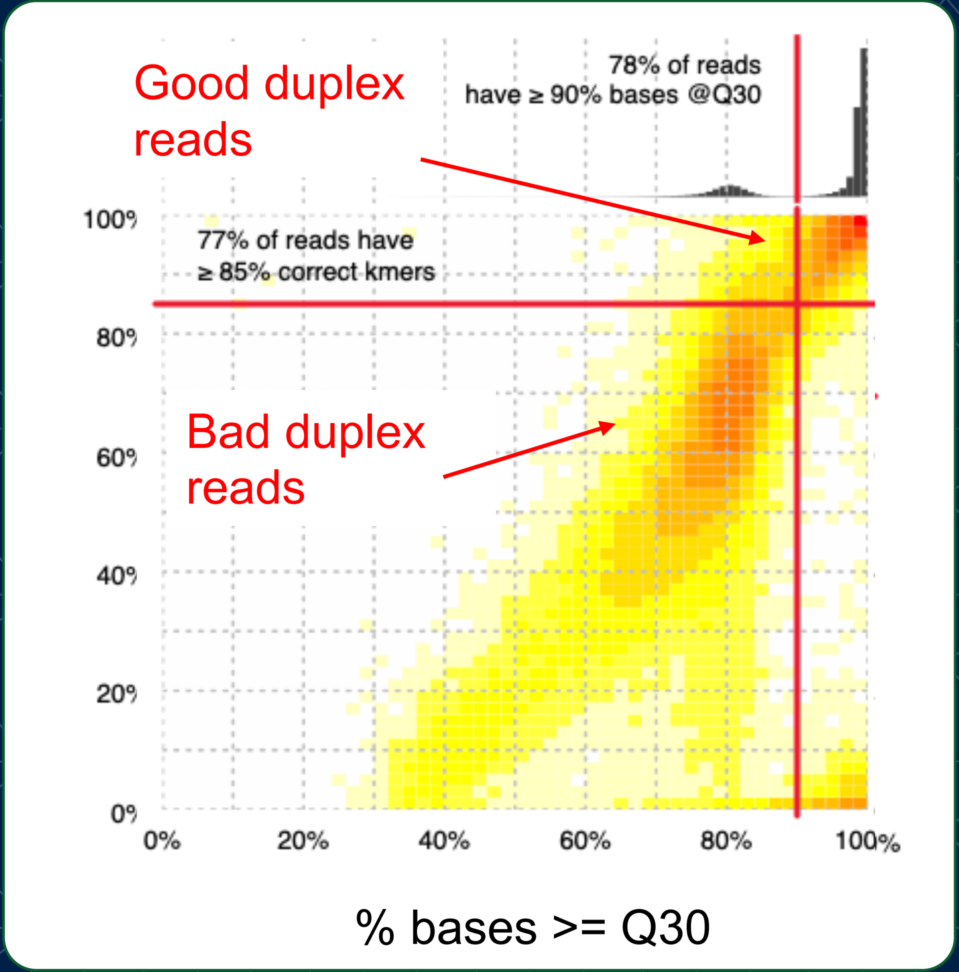
77% of reads have ≥ 85% correct kmers

Nanopore R10.4.1 (duplex)

K-mer accuracy vs. predicted accuracy

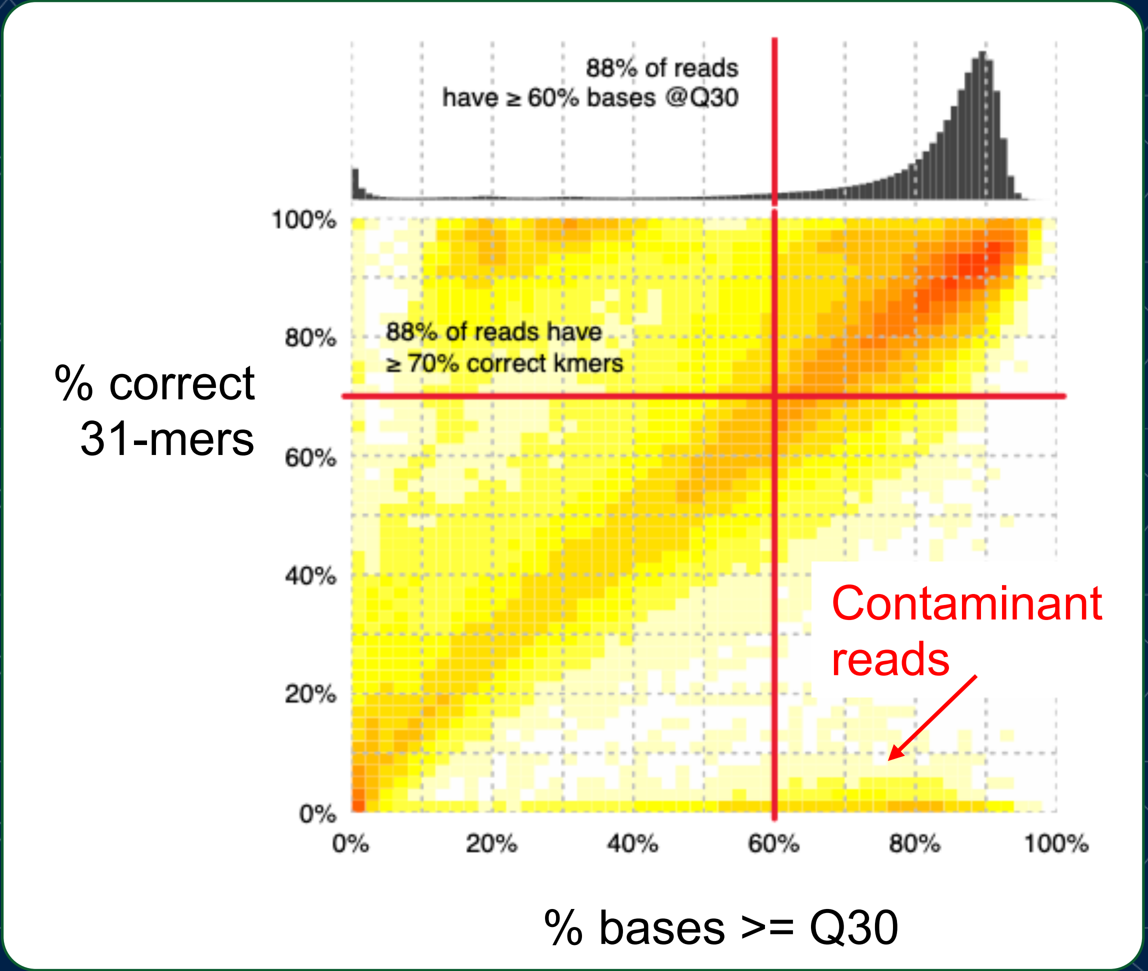


Nanopore R10.4.1 (simplex)

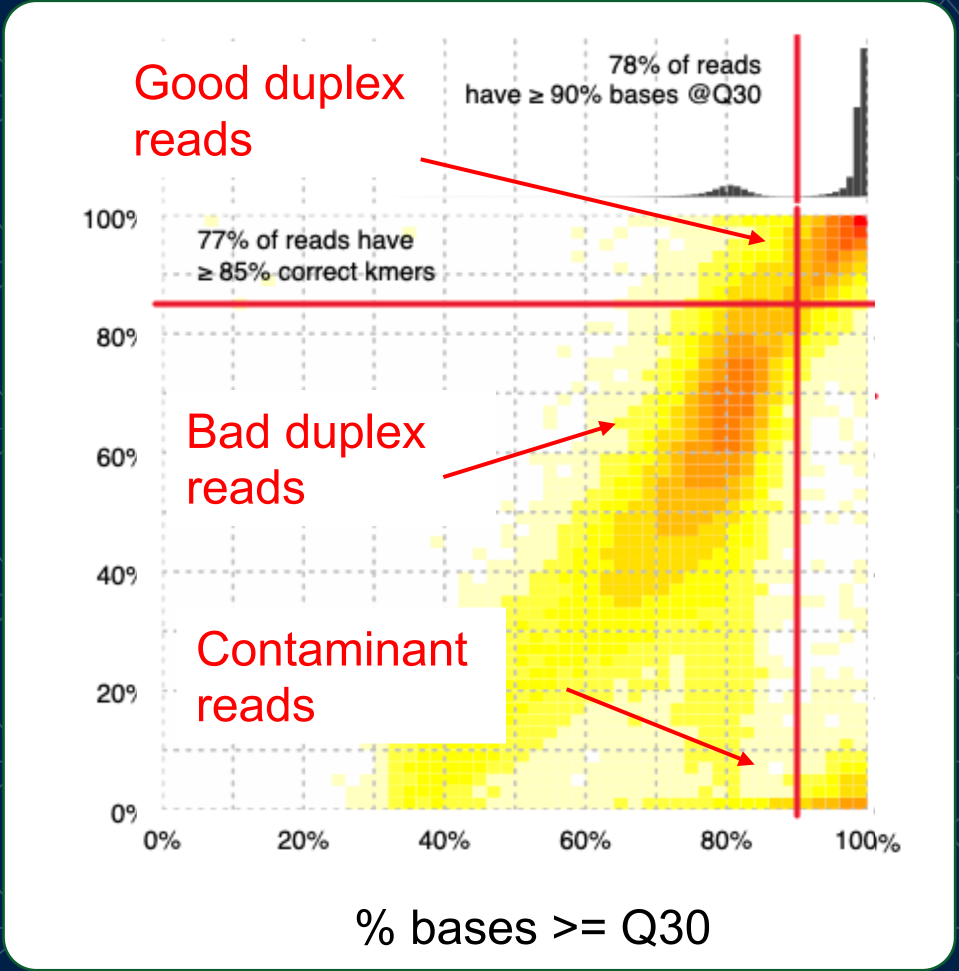


Nanopore R10.4.1 (duplex)

K-mer accuracy vs. predicted accuracy

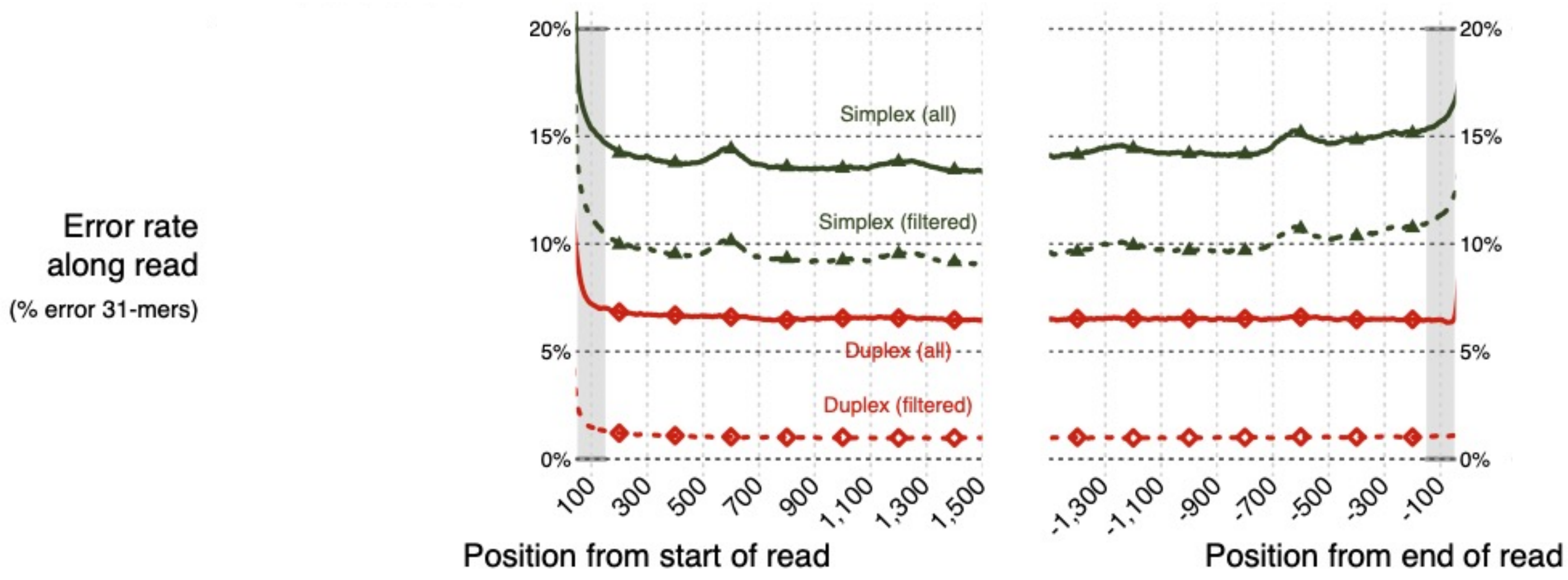


Nanopore R10.4.1 (simplex)

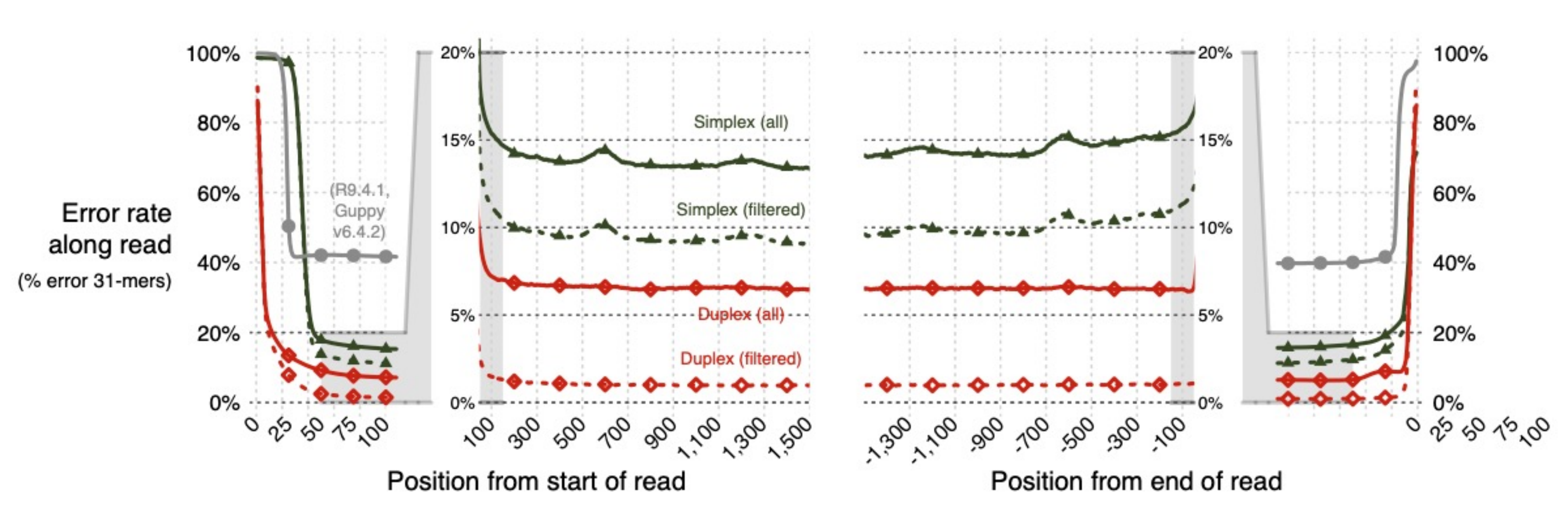


Nanopore R10.4.1 (duplex)

Accuracy along the read



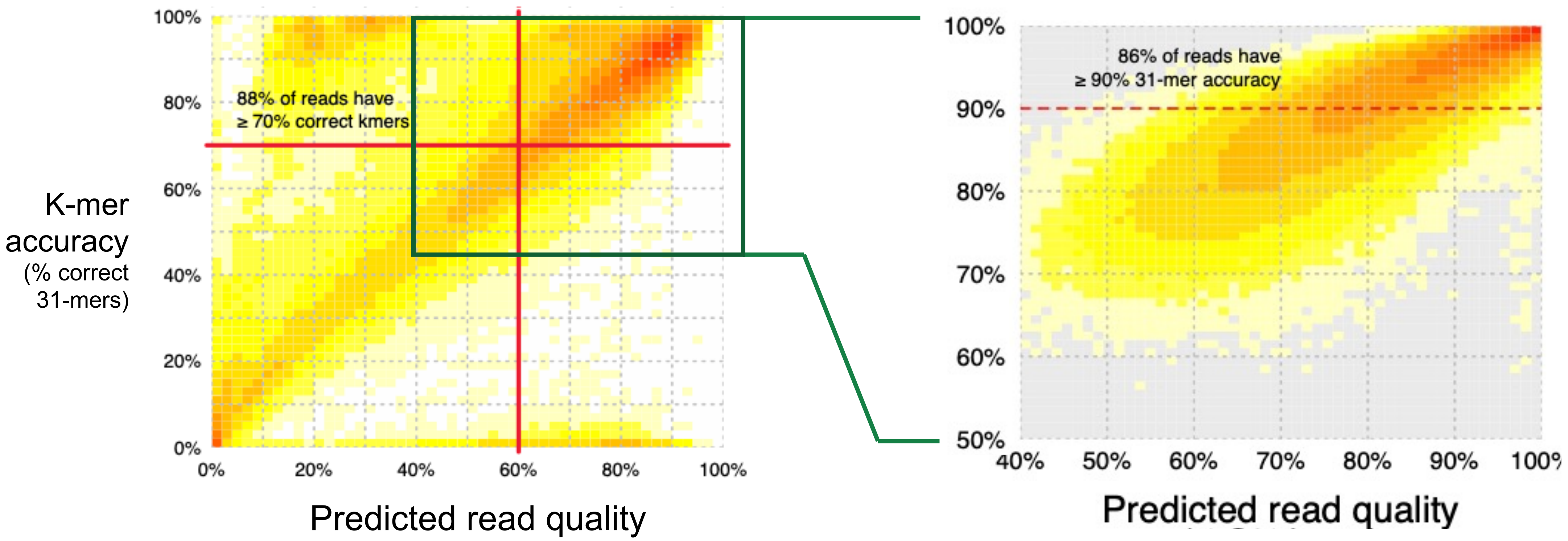
Accuracy along the read



R10.4.1 is better than R9.4.1, especially after filtering.
Filtered duplex data has stupendously low error rates across most of the read.
Prominent read-end artifacts due to adapters (that might not be completely removable)

(Also, note the weird error bumps every 600bp…)

K-mer accuracy vs. predicted accuracy

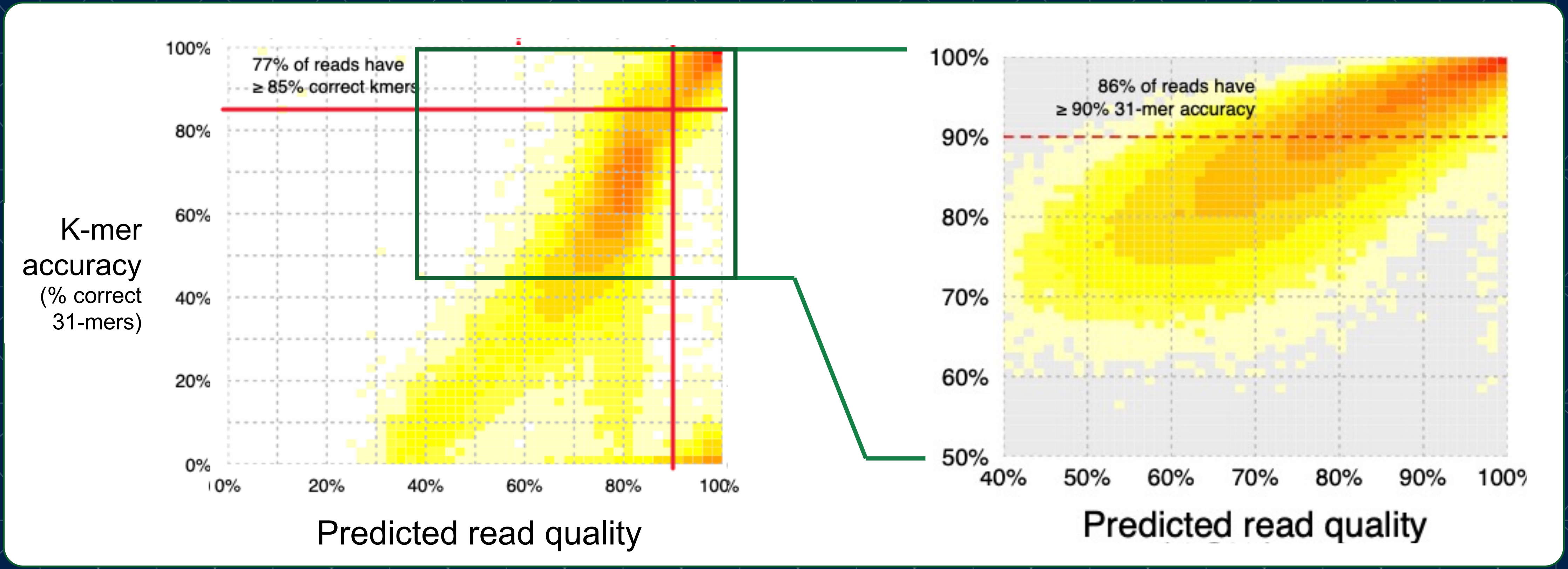


Nanopore R10.4.1
simplex

Pacbio Revio

Measuring predicted quality as: % bases >= Q30

K-mer accuracy vs. predicted accuracy

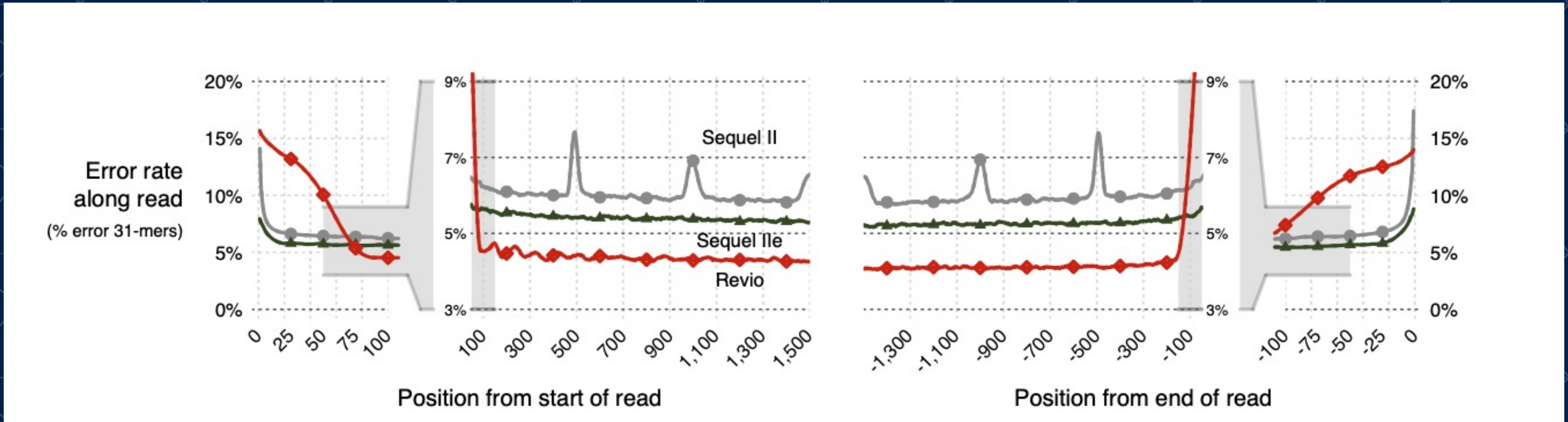


Nanopore R10.4.1
duplex

Pacbio Revio

Measuring predicted quality as: % bases >= Q30

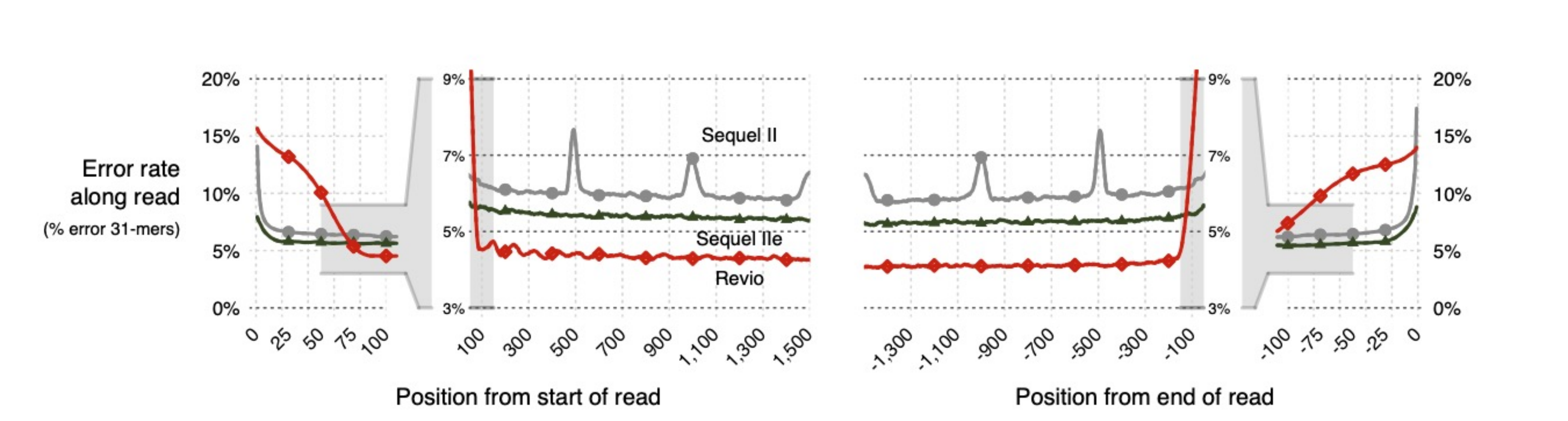
Accuracy along the read



Our Revio data **also** shows elevated rates at the end of reads - !!
But improves upon Sequel Iie across most of the read length

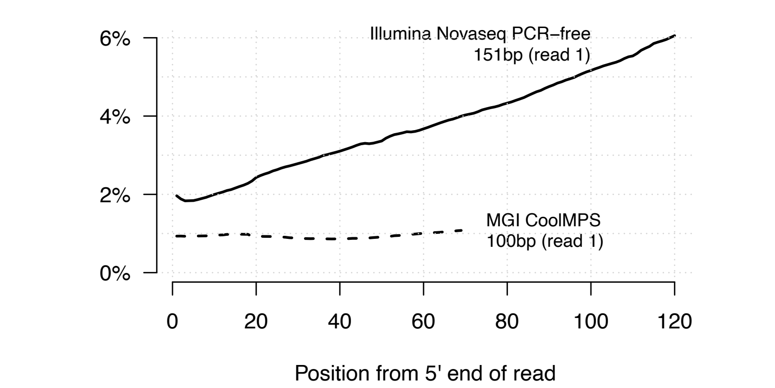
(Meanwhile our older Sequel II data has weird, unexplained ‘bumps’ every 500bp.)

Accuracy along the read



Error rates comparable to some Illumina data

Though some short-read datasets are better



WELLCOME CENTRE *for* HUMAN GENETICS

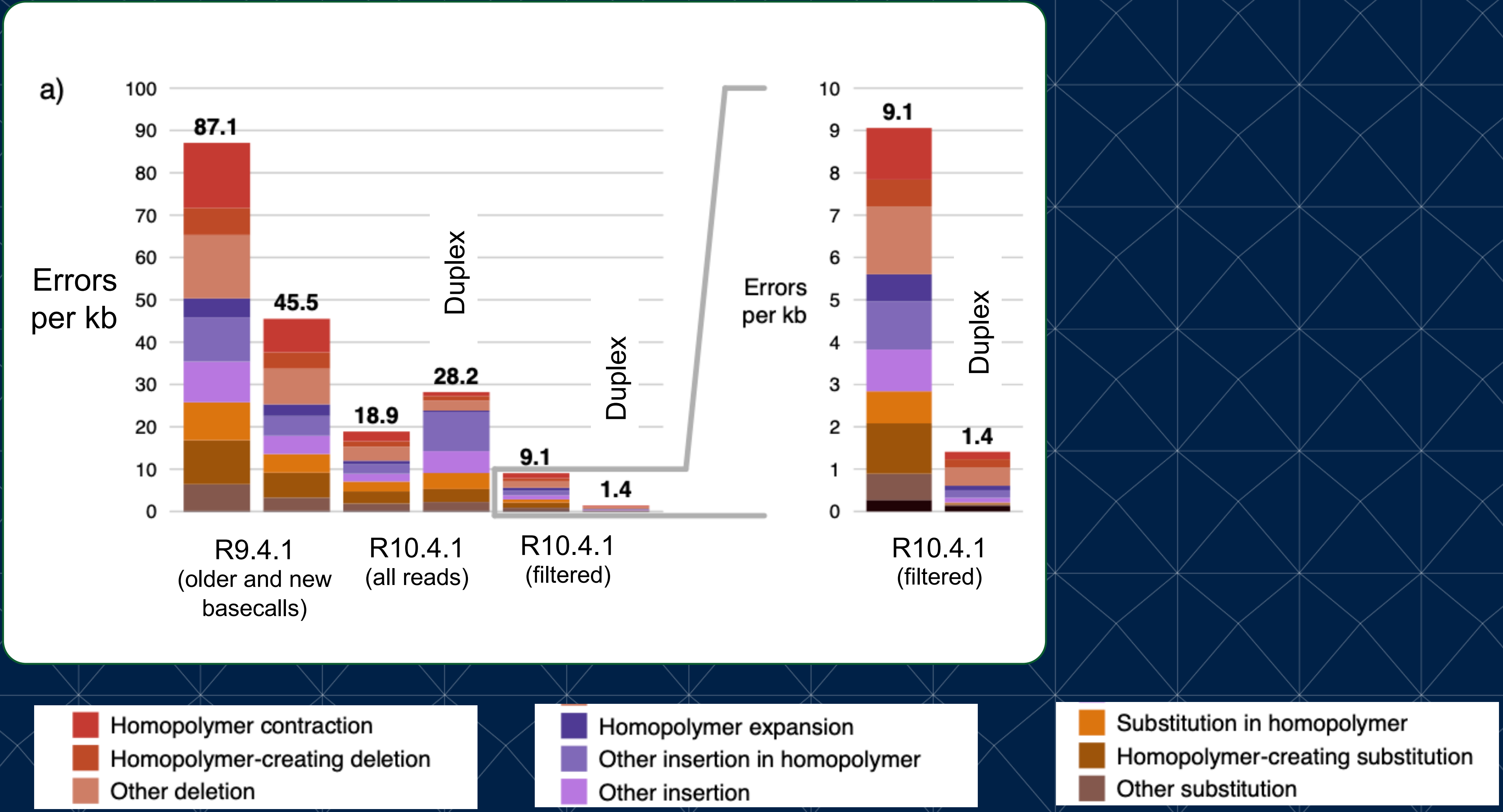
Summary

- Nanopore R10.4.1 data improves over R9.4.1 data.
- Nanopore still noisy and has a few artifacts
- Pacbio Revio also improves over Sequel IIe across most of the read
- Nanopore duplex reads are somewhat comparable to Pacbio reads – maybe better after filtering, but are only 4-5% of data
- Both platforms have annoying-looking read-end effects.

Alternate approach: **align** to a reference sequence, **mask out** true variation and repetitive sequence

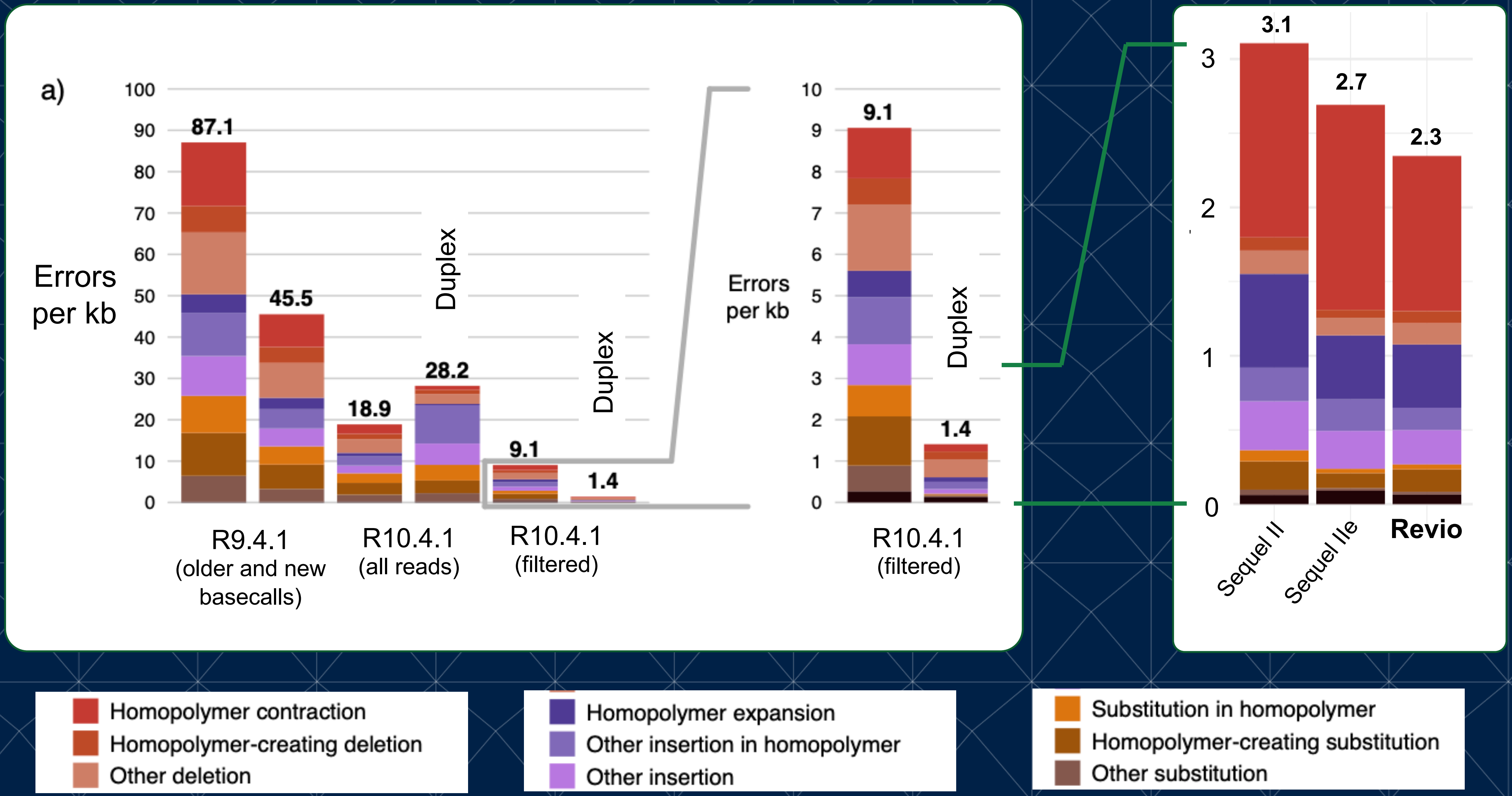
We use T2T assembly, mask out SNPs, INDELS, and SVs from HV31 data, and satellite arrays, segdups, repeat-masked elts.

Nanopore

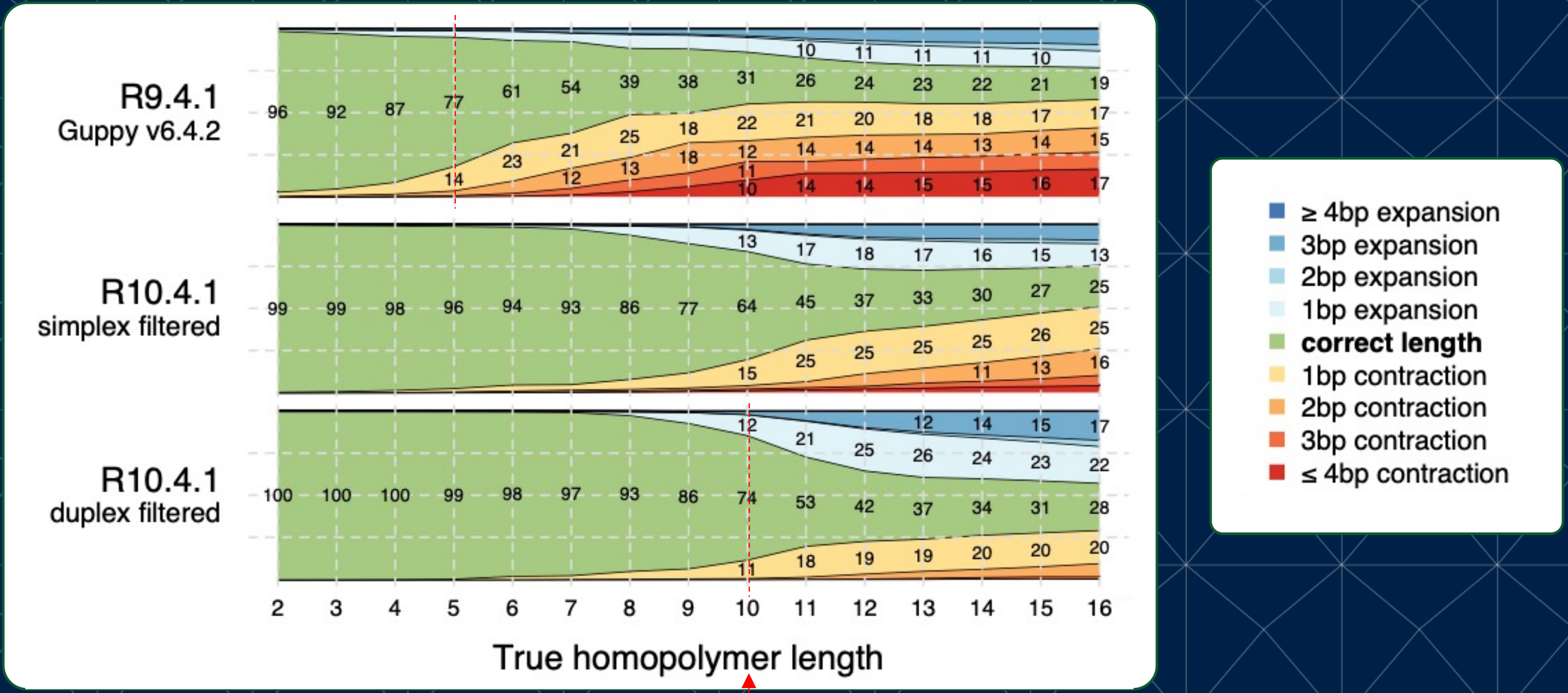

a)
Errors per kb
100
90
80
70
60
50
40
30
20
10
0
87.1
45.5
18.9
28.2
9.1
1.4
Duplex
Duplex
R9.4.1
(older and new basecalls)
R10.4.1
(all reads)
R10.4.1
(filtered)
Errors per kb
10
9
8
7
6
5
4
3
2
1
0
9.1
1.4
Duplex
R10.4.1
(filtered)
Homopolymer contraction
Homopolymer-creating deletion
Other deletion
Homopolymer expansion
Other insertion in homopolymer
Other insertion
Substitution in homopolymer
Homopolymer-creating substitution
Other substitution

Nanopore

Pacbio



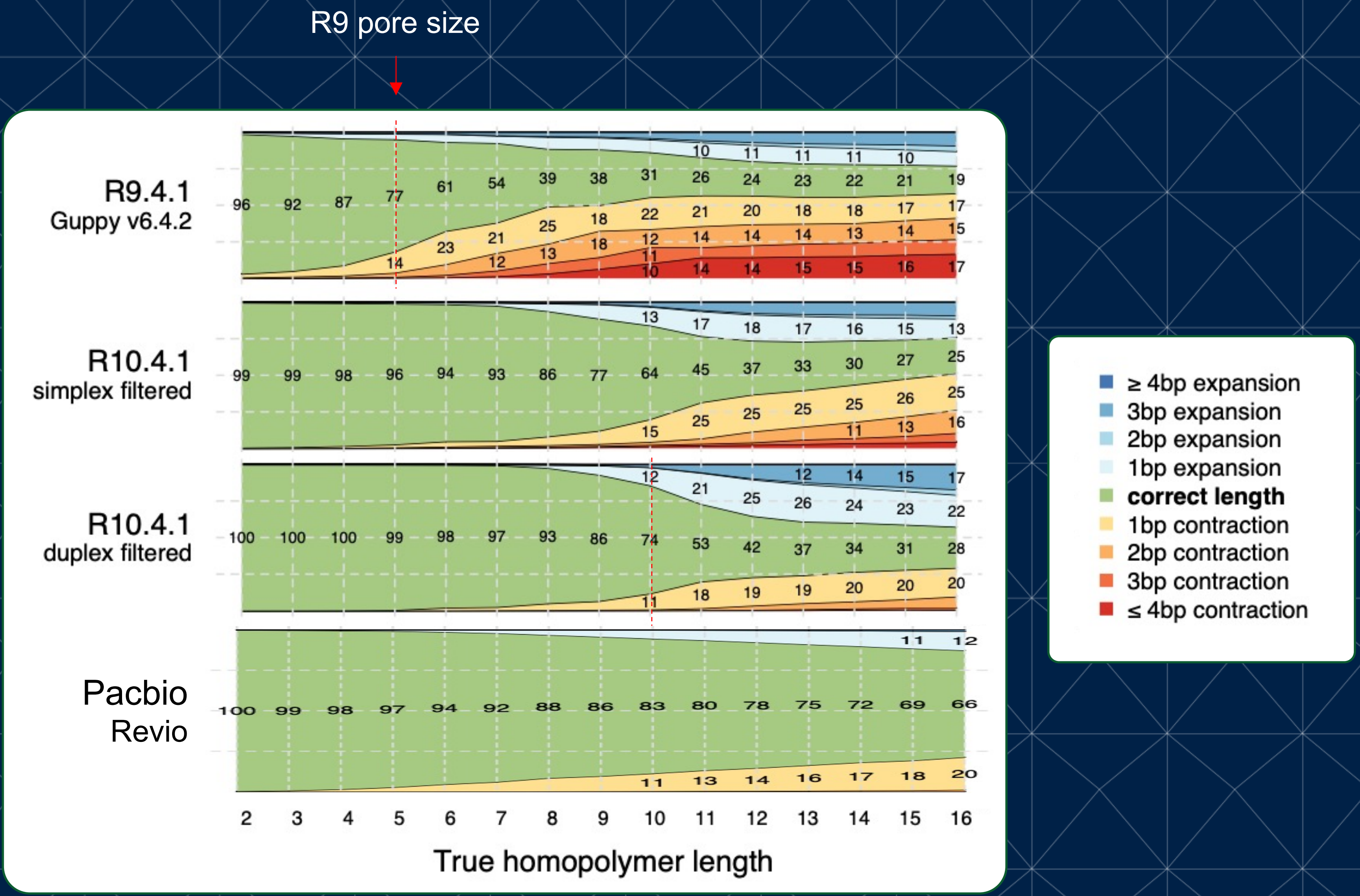
R9 pore size



R10 pore size

Nanopore homopolymer length calling still drops off above pore size…

(but R10 pore size is larger)



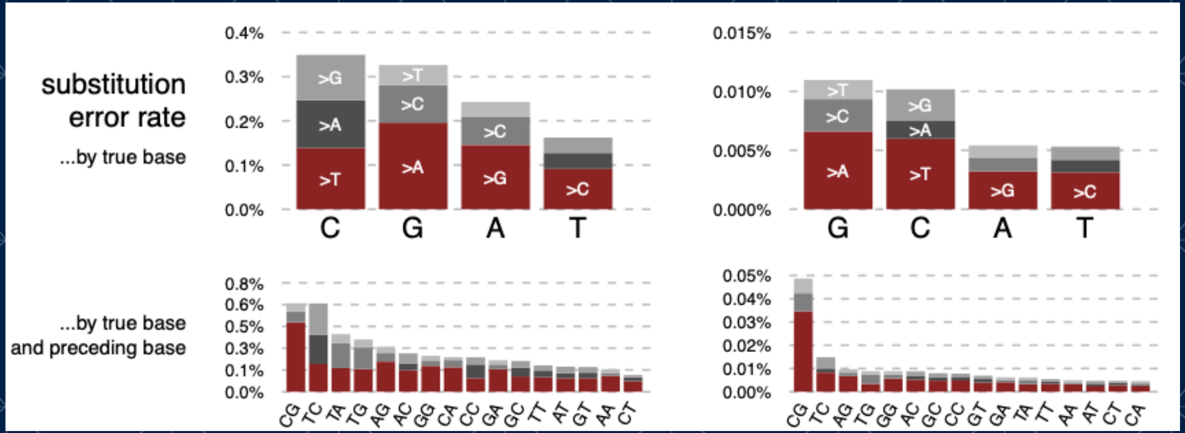
Nanopore homopolymer length calling still drops off above pore size…

(but R10 pore size is larger)

Pacbio calls longer homopolymers better

still only ~60-70% accuracy for longest lengths

Subtle substitution biases are also present

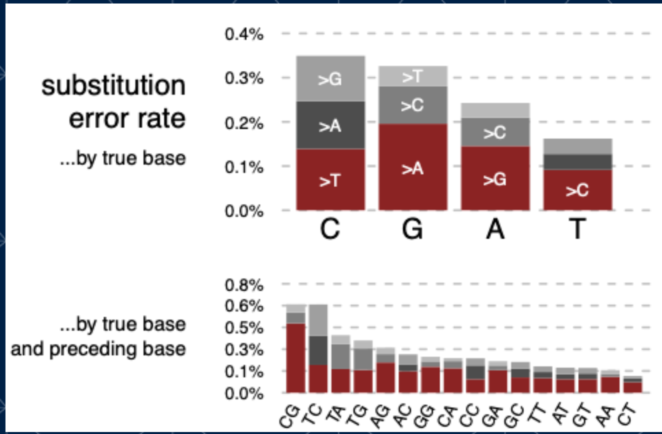


Nanopore tends to make transition-like errors
(A<->G and C <-> T).

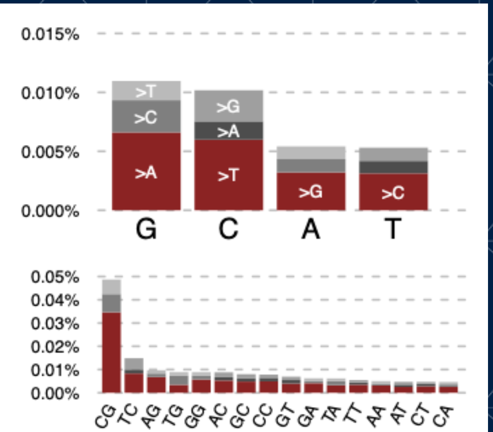
CpG sites appear to have a particularly high substitution rates.
But the absolute rate is still low.

Subtle substitution biases are also present

Simplex



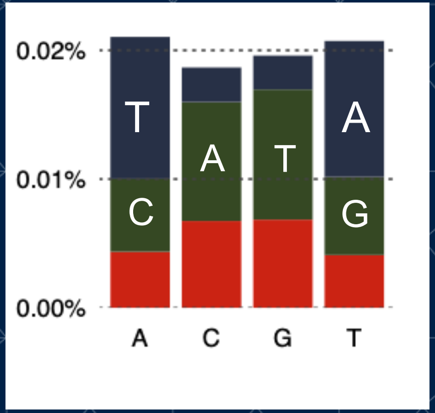
Duplex



Nanopore makes substitutions of C and G bases, and tends to make transition-like errors (A<->G and C <-> T).

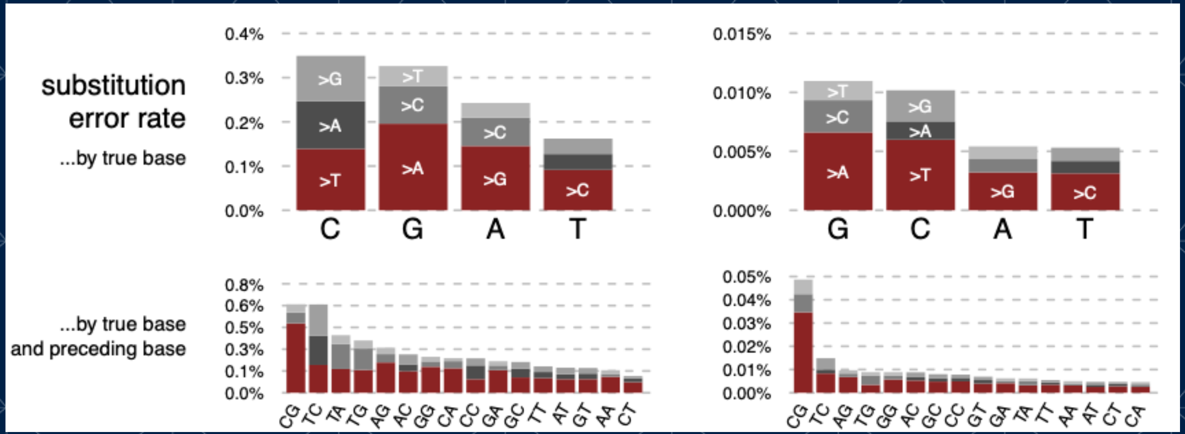
CpG sites appear to have a particularly high substitution rates.
But the absolute rate is still low.

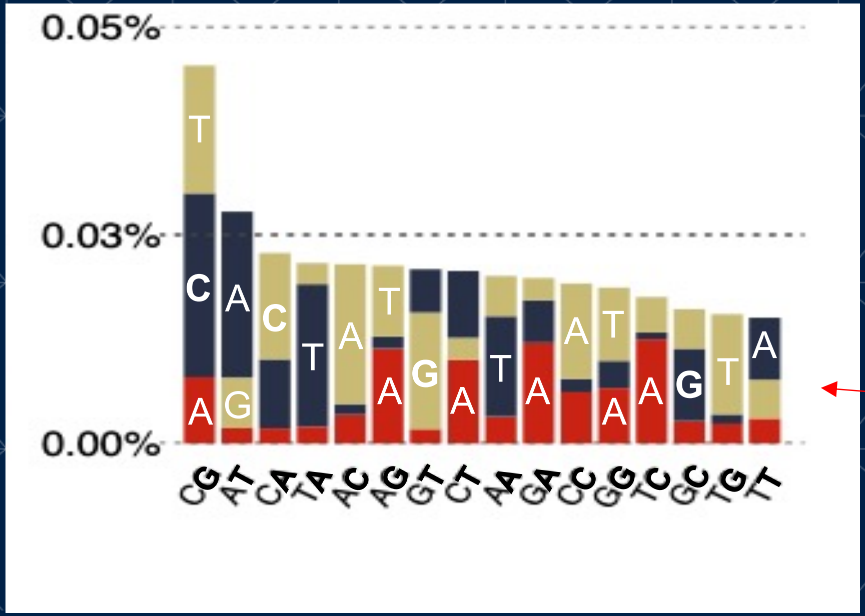
Pacbio



Pacbio makes more substitutions at A and T bases and tends to miscall to A or T.

Subtle substitution biases are also present





Moreover both platforms appear to have elevated substitution rates at CpG sites

Summary

New revisions of ONT and Pacbio data are both fantastic.

Nanopore requires more downstream work to filter / process.

Duplex reads look very exciting, if low throughput can be overcome.

Costs

For this experiment we 5 Promethion flowcells and 2 Revio SMRT cells were used.

For ONT, the list cost places the consumables cost at £2,700 - £4,050 flowcell cost, depending on order volumes, plus possibly £500 for library reagents. However you might only need 3 flowcells with current version (because it runs at a faster rate), so perhaps £2,120 - £2,930 in total

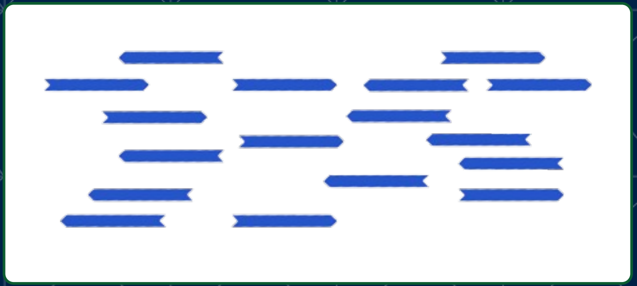
For Pacbio, it’s a bit unclear to me but two flowcells might cost ~£2,000 with library prep on the order of £500 (I think - very ballpark.), so £2,500 in total.

In other words - the costs look very similar to me.

Note these costs do **not** include equipment, service, personnel or additional reagent costs.

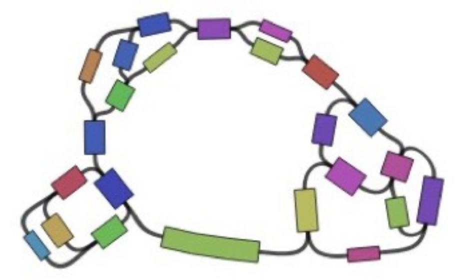
Genome assembly application 1

A haplotype-resolved assembly with functional data



PacBio Sequel II/IIe
ONT R10.4.1

Verkko

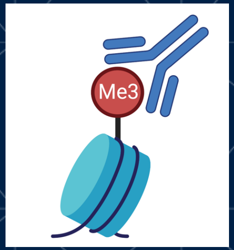


Multiple approaches
(BubbleGun, Linked reads, kmer approach),
WhatsHap, HapCut2

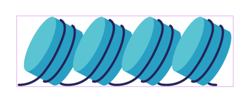
Methylation

Jia-Yuan Zhang

RNA-seq
(expression)



ChIP-seq
(For histone modifications)



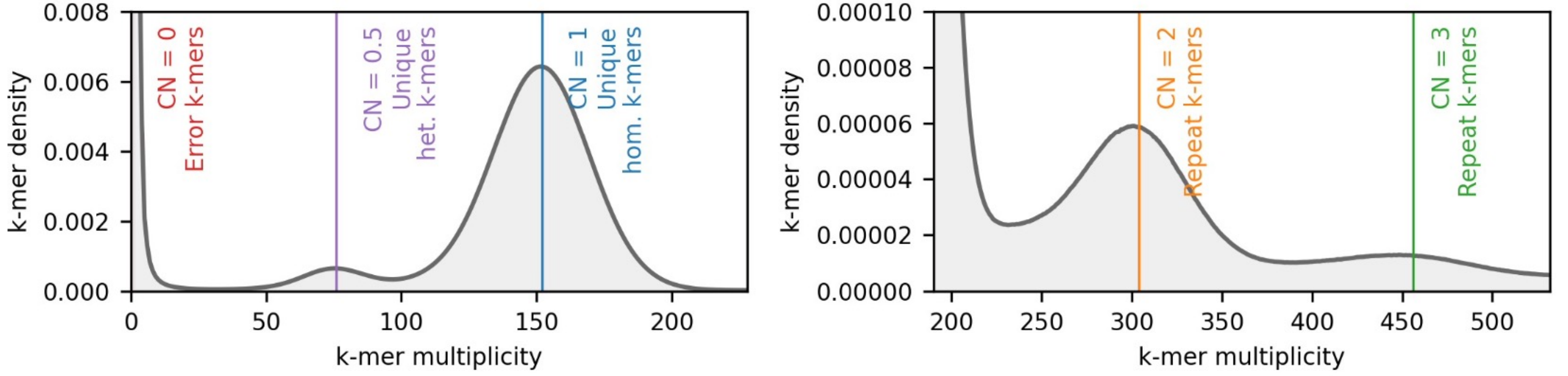
ATAC-seq
(detects open chromatin)

Align and resolve phase

Haplotype 1
Haplotype 2

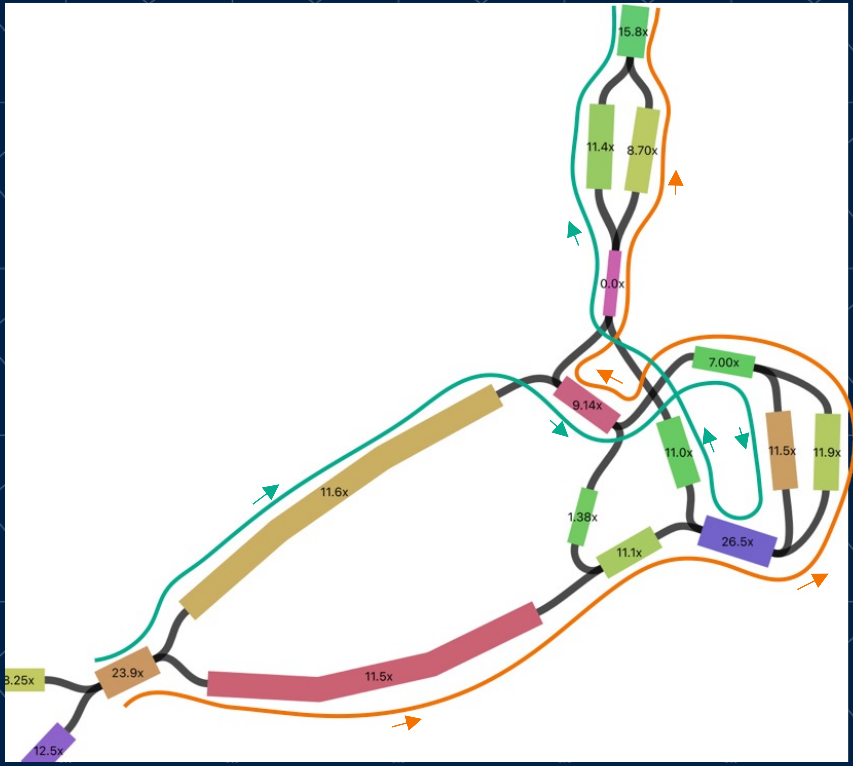
Phased ‘omniome”
reflecting immune cell types

Jia-Yuan Zhang

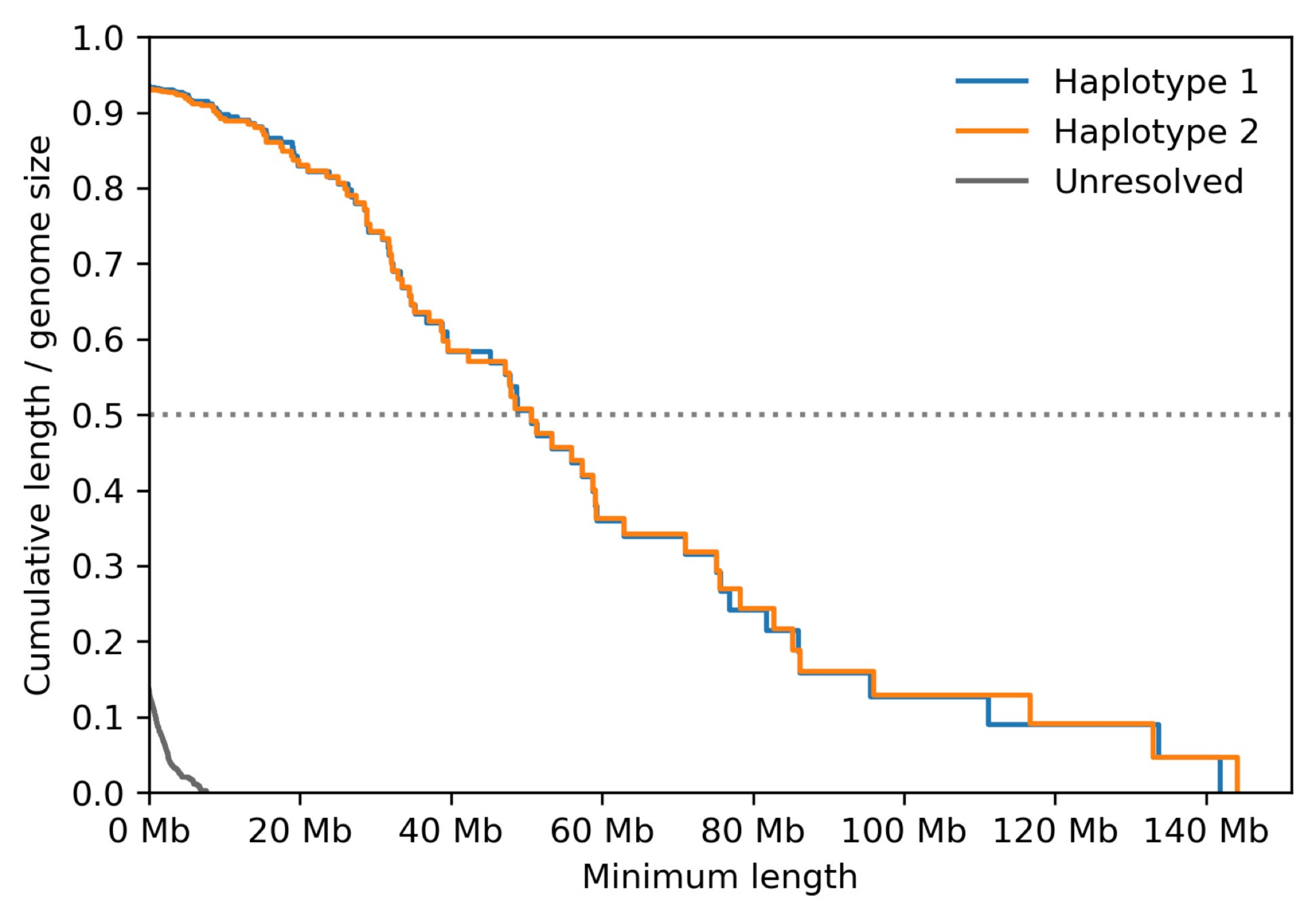


Example: a segmental duplication at TCAF1/2 locus

Not fully resolved in the Verkko assembly graph.



Use an empirical model of the k-mer distribution to probabilistically resolve the most-likely pair of haplotypes.

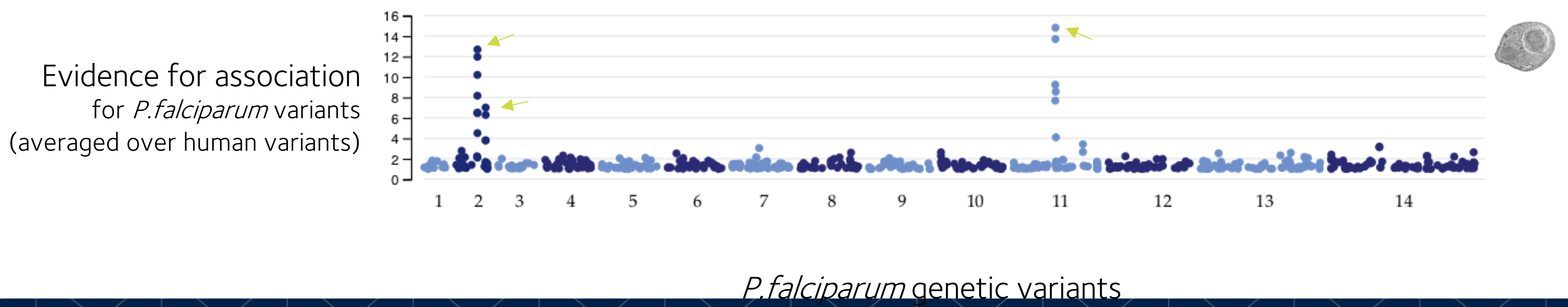


A ~50Mb ‘phased’ NG50

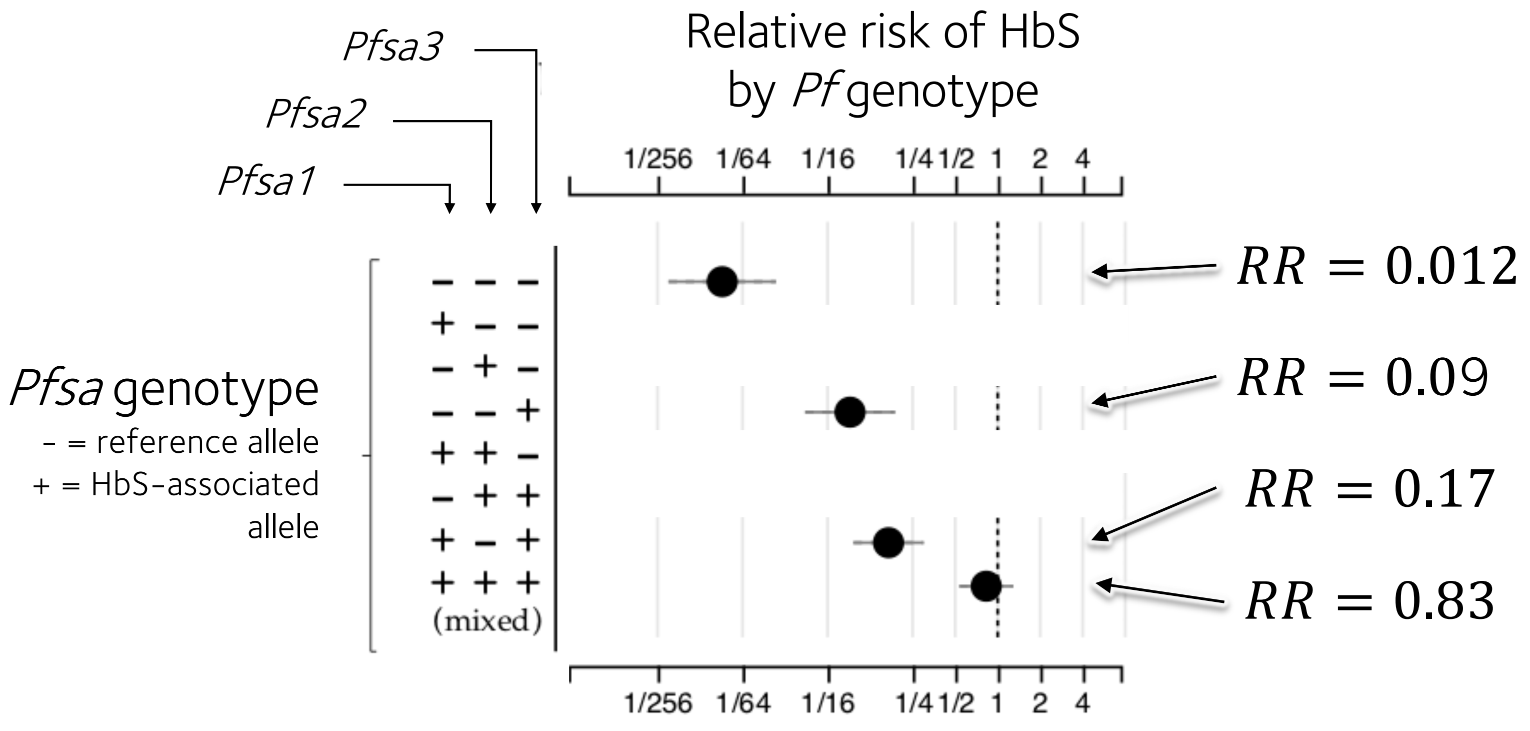
(50% of assembly bases are in phased contigs of 50Mb or greater)

Genome assembly application 2: resolving malaria structural variants involved in host-parasite interactions

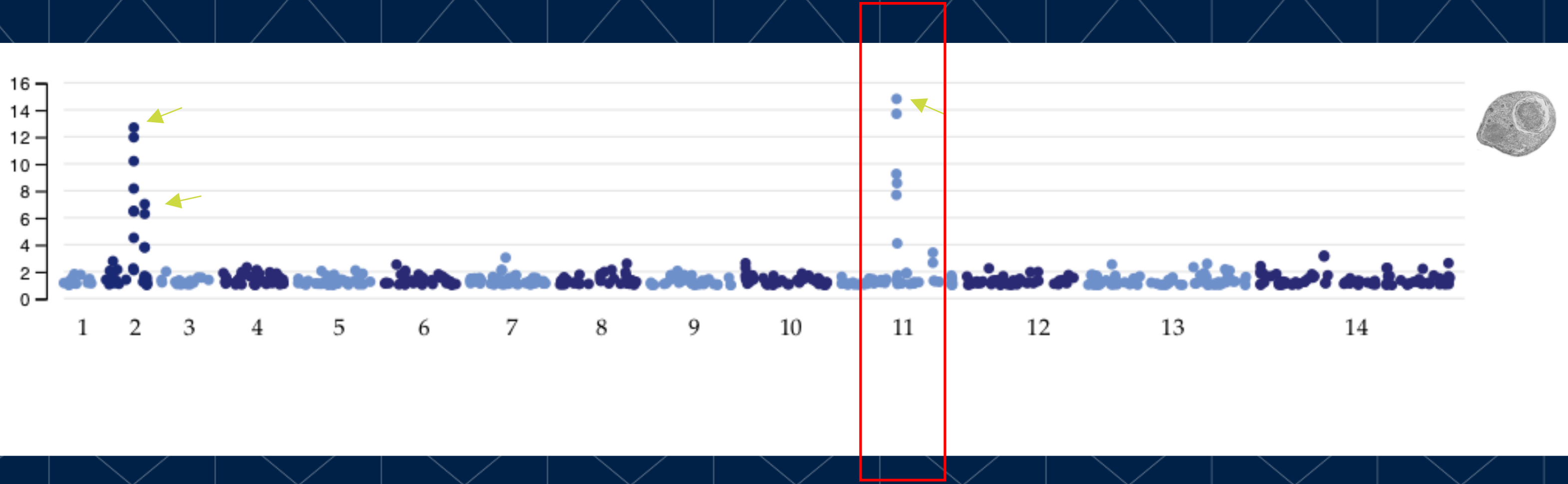
Three regions of the *Pf* genome are associated with sickle hamoglobin



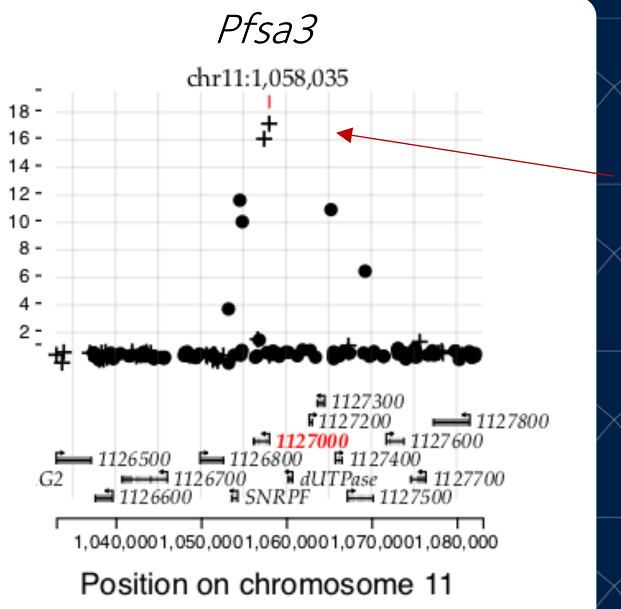
HbS appears to give very strong protection against reference-like parasites, but maybe hardly any against + + + parasites



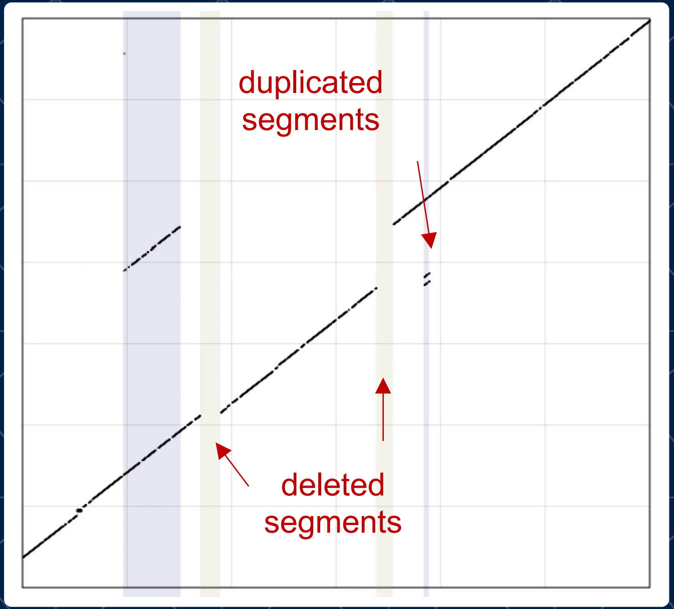
Evidence for association
for *P.falciparum* variants
(averaged over human variants)



Evidence for
association
for *P.falciparum*
variants
with HbS



The top SNPs are non-synonymous changes.
However they also appear to be linked to a surrounding structural variant, and are associated with increase transcription.



Reference parasite

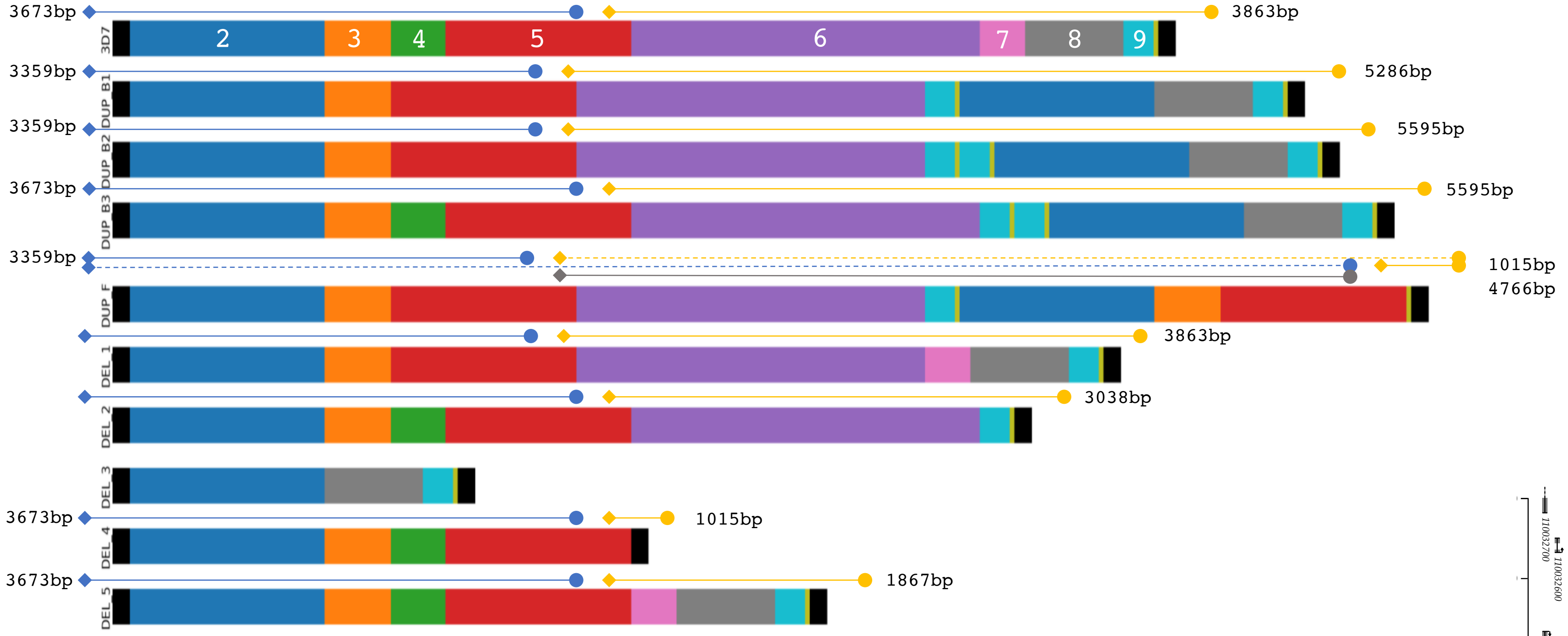
Attempt 1: Nanopore-based amplicon sequencing

Annie Forster

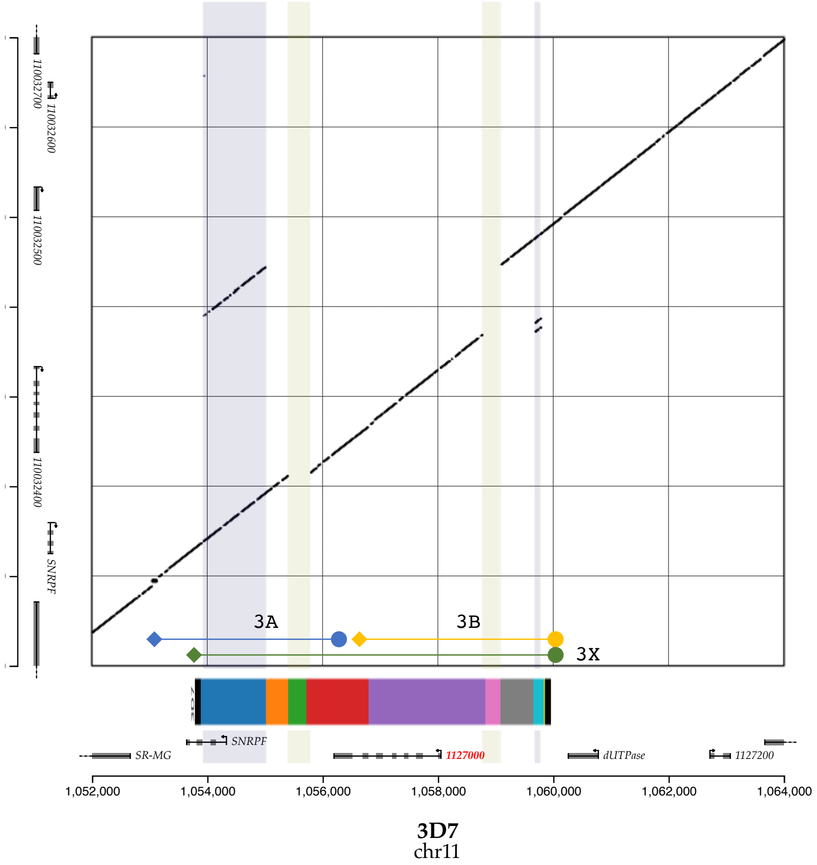
Jason Hendry

Mariateresa de Cesare

Anna Jeffresy

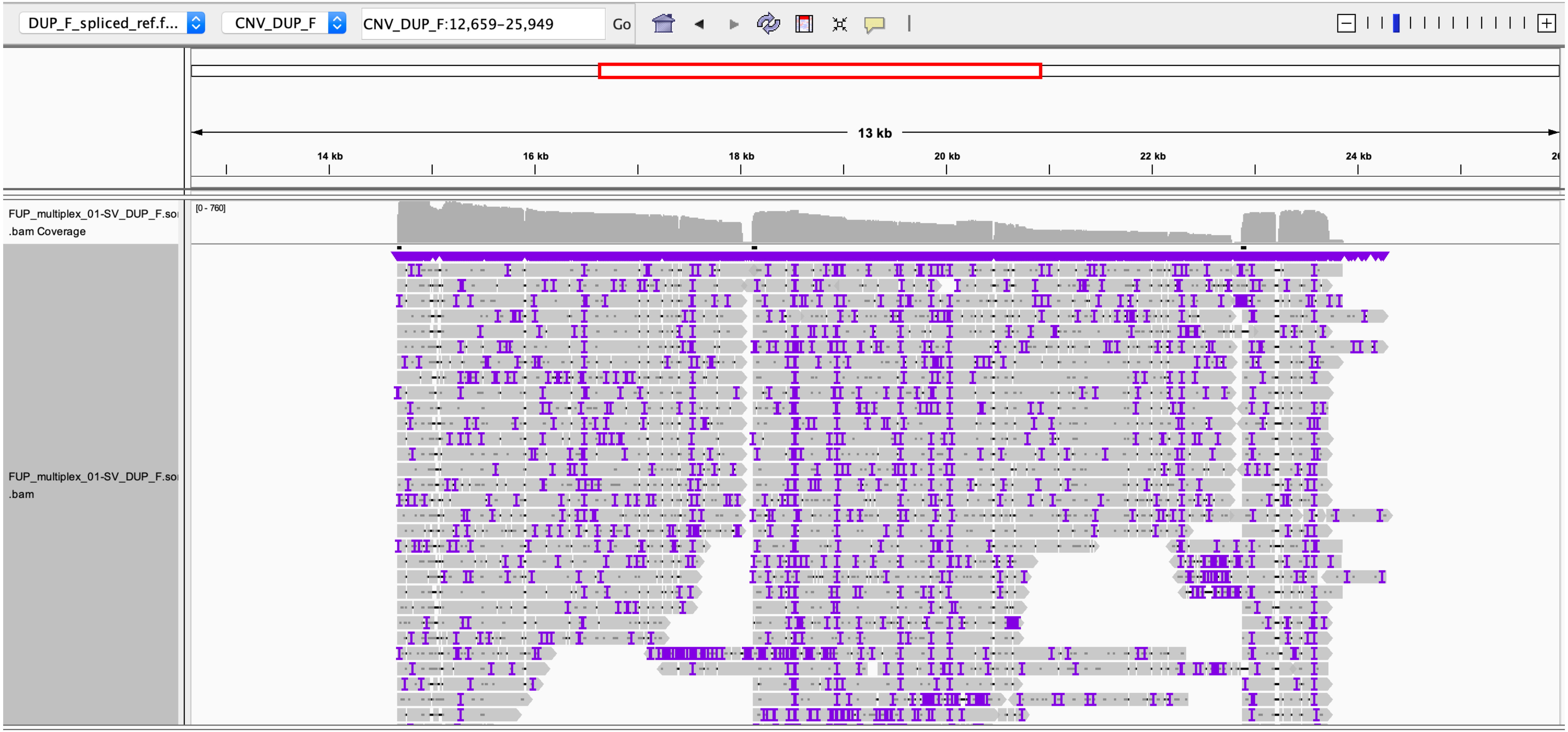


Analysis of short read data (MalariaGEN PF6) revealed there are multiple structural types segregating.



Annie Forster

Nanopore amplicon sequencing:
Jason Hendry
Mariateresa de Cesare
Anna Jeffreys



DUP_F_spliced_ref.f...

CNV_DUP_F

CNV_DUP_F:12,659-25,949

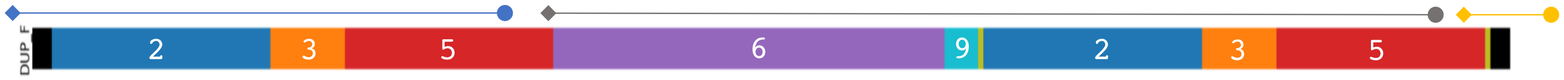
Go
13 kb
14 kb
16 kb
18 kb
20 kb
22 kb
24 kb
FUP_multiplex_01-SV_DUP_F.sor .bam Coverage
[0 - 760]
FUP_multiplex_01-SV_DUP_F.sor .bam

FUP_multiplex01 aligned to a mock-up DUP_F reference. Looks like there are three fragments as predicted! It's a bit difficult to count length but roughly they seem to be...

1: 3350bp
2: 4651bp
3: 979bp?

Predicted lengths were:

Pfsa3A – 3,359bp
Hybrid – 4,766bp
Pfsa3B – 1,015bp



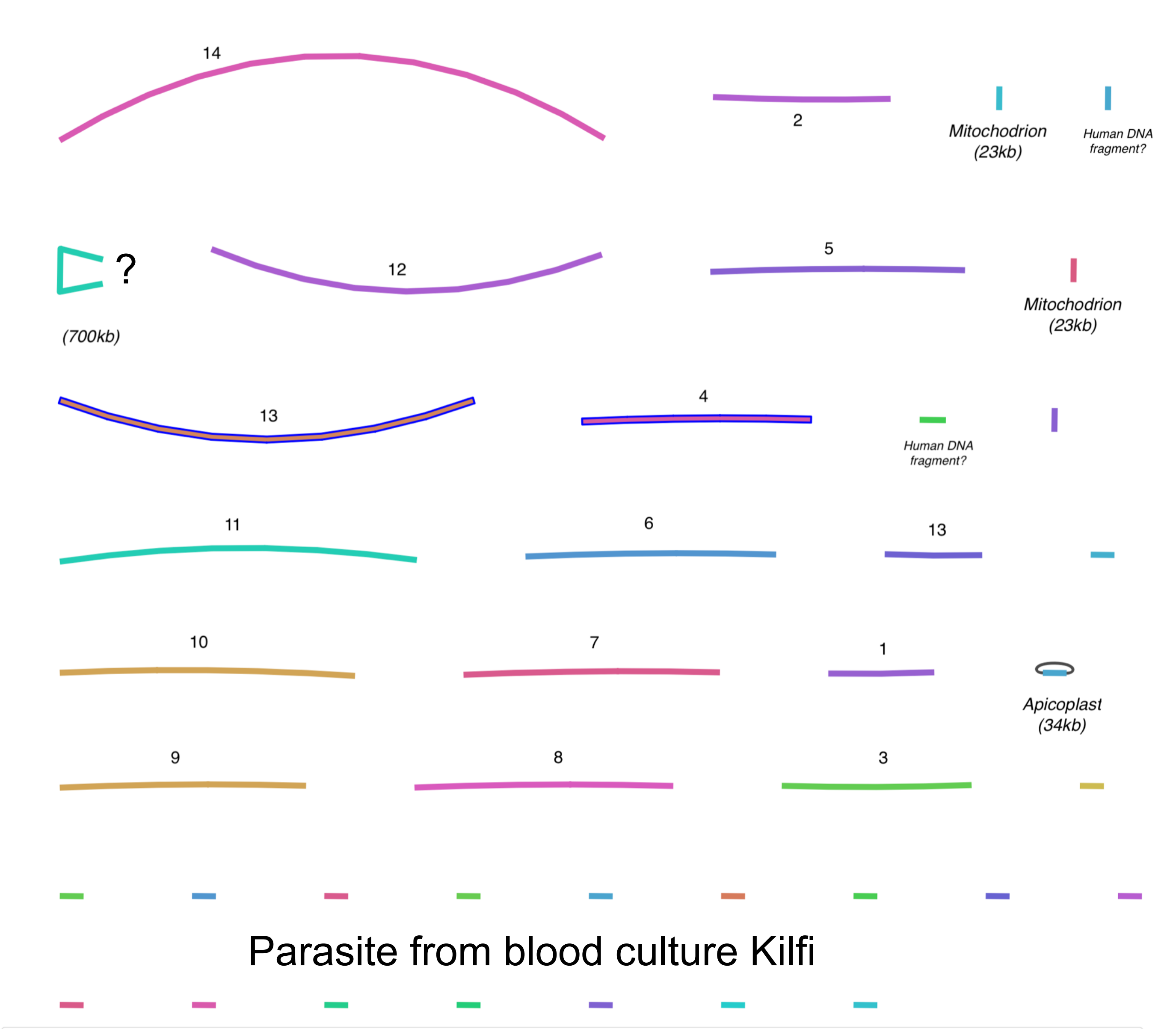
2nd attempt: Pacbio whole-genome sequencin

3D7
FUP-H
4 Kenyan parasites
2 Gambian parasites
1 parasite from single-cell sorting

Carried out by James Docker and Amy Trebes, Oxford Genomics Centre for a test of new fragmentation protocol.

Worked amazingly well

Alex Macharia, Patrick



Parasite from blood culture Kilfi

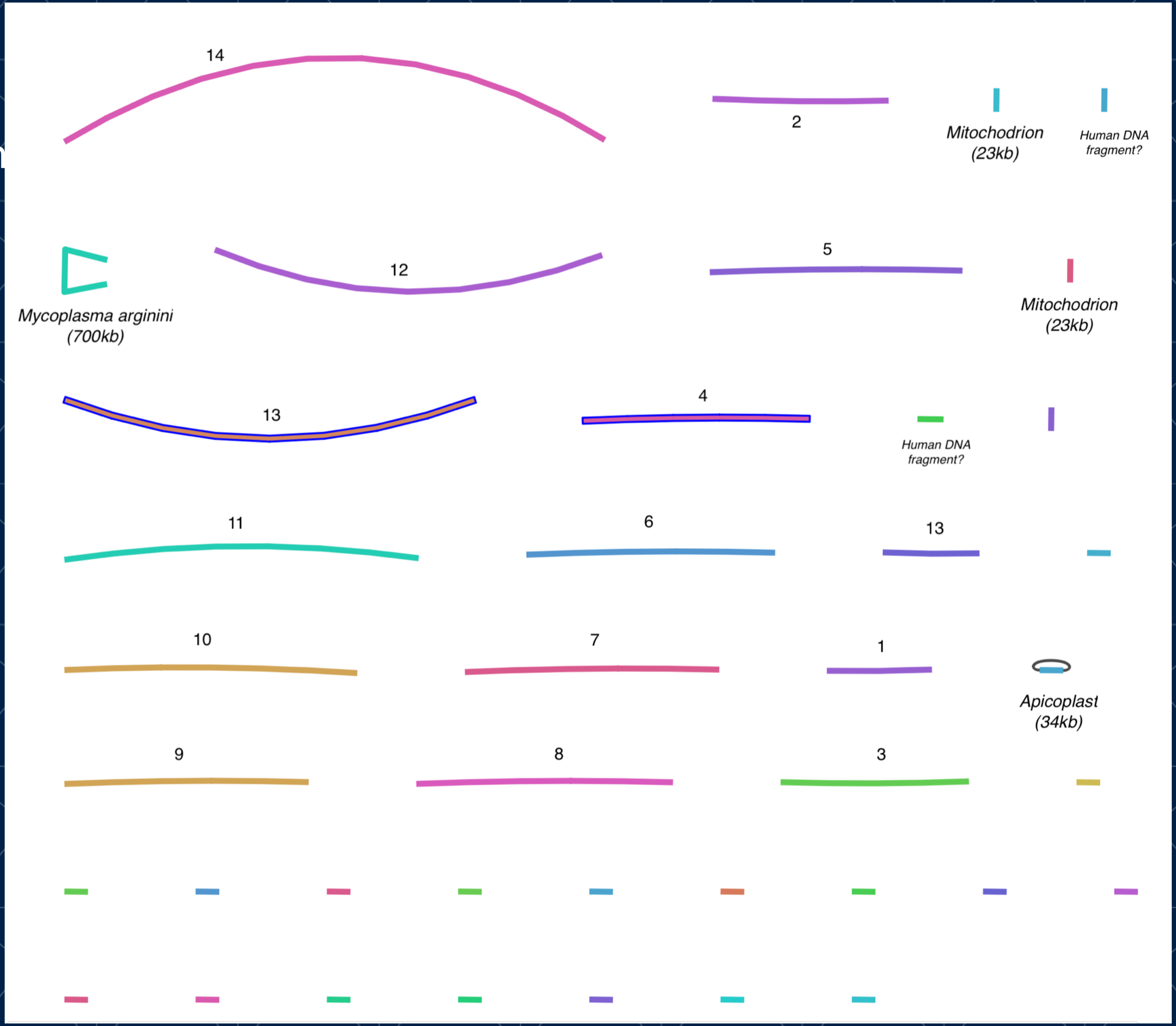
Process: hifiasm -> BandageNG

2nd attempt: Pacbio whole-genome sequencin

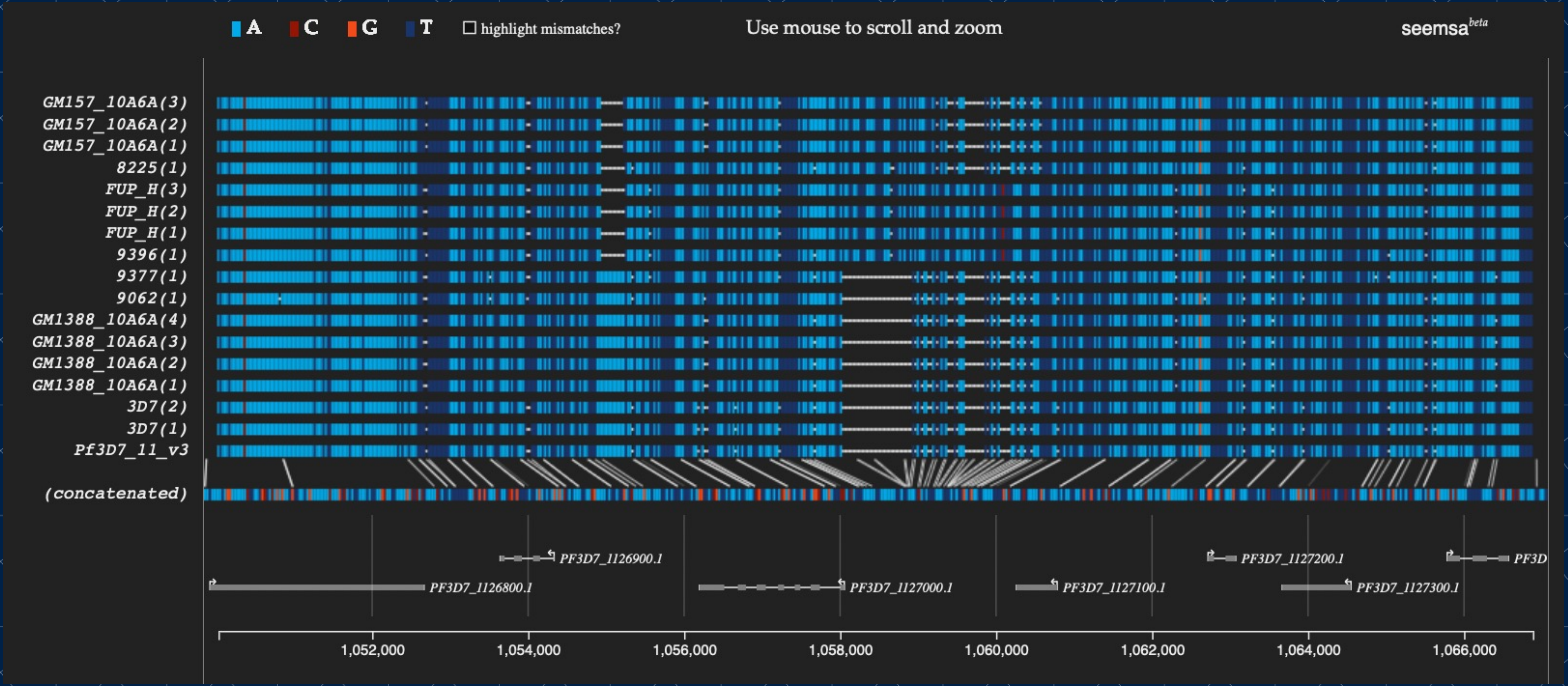
3D7
FUP-H
4 Kenyan parasites
2 Gambian parasites
1 parasite from single-cell sorting

Carried out by James Docker and Amy Trebes, Oxford Genomics Centre for a test of new fragmentation protocol.

Worked amazingly well



Hifiasm -> BandageNG



Multiple sequence alignment of *P.falciparum* whole genomes

Acknowledgments

wellcome centre human genetics

John Todd
Julian Knight
Andrew Brown
Tony Cutler
Connor Davison
Jia-yuan Zhang
David Smith
Annie Forster
Qijing Shen
Jason Hendry
Hitomi Kuwabara
David Buck
Paolo Piazza
Helen Lockstone

KEMRI-Wellcome Kilifi, Kenya
Alexander W. Macharia
Patrick Ombati
Silvia Kariuki

CIMR:
Julian Rayner

OGC:
Amy Trebes
James Docker
David Buck

PacBio

Riki Aydeniz
Eirini Maria Lampraki
Mike Eberle
Cillian Nolan

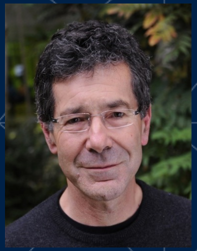
Oxford NANOPORE Technologies

Simon Mayes
Philipp Reschender
Tonya McSherry

MalariaGEN

Dominic Kwiatkowski
Ellen Leffler
Kirk Rockett

…and HV31.



Dominic
Kwiatkowski
1953-2023

"We thank the patients and staff at the Paediatric Department of the Royal Victoria Hospital in Banjul, Gambia, and at Kilifi County Hospital and the KEMRI-Wellcome Trust Research Programme, Kilifi, Kenya for their help with this study."

MalariaGEN
GENOMIC EPIDEMIOLOGY NETWORK

Thanks!