The HV31 / Omniome projects

Gavin Band

Long Read Sequencing Networking Event

NHS Central and South Genomic Medicine Service Alliance

20th July 2022

wellcome centre human genetics

MRC Weatherall Institute of Molecular Medicine

wellcome sanger institute

University of Oxford

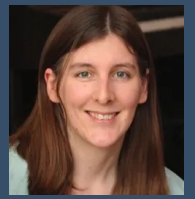
John Todd

Julian Knight

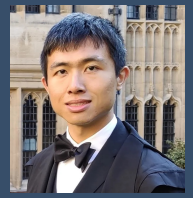
David Buck / OGC

Andy Brown

Tony Cutler



Helen Lockstone



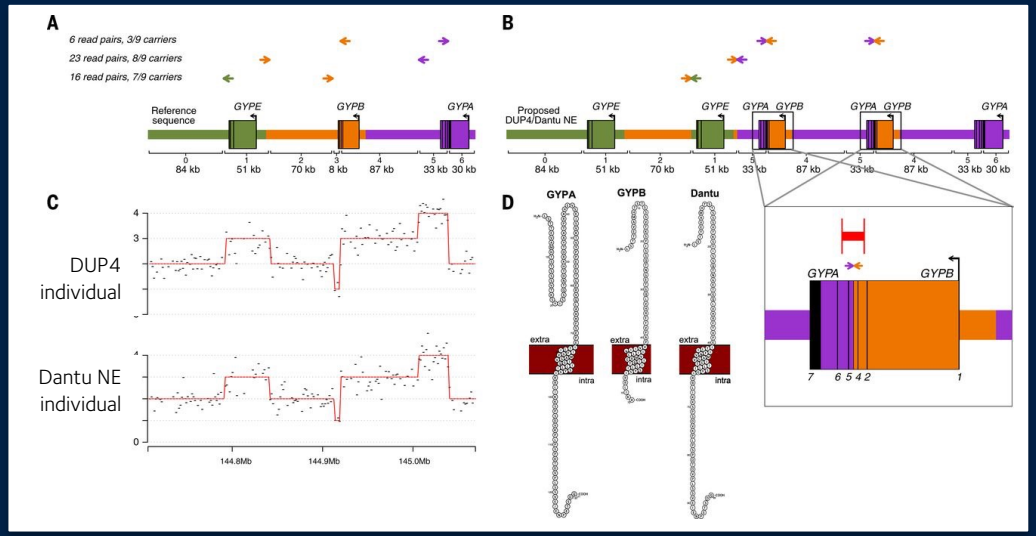
Jia-Yuan Zhang

Hannah Roberts
David Flores
Justin Whalley
Rachael Bashford-Rogers
David Smith
Amy Trebes

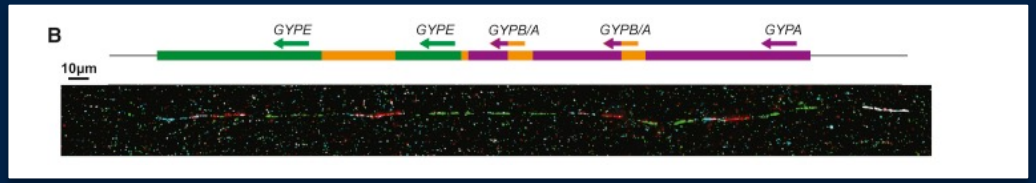
Olga Mielczarek
Barbara Xella (WIMM)
Karen Oliver (Sanger)
Connor Davidson
Paolo Piazza
+ others...

Motivation: linking genome structural variation to function and disease

- The role of genome structural variation in disease remains poorly understood.
- Often a feature of complex genomic regions such as segmental duplications.
- Example: the Dantu blood group variant protects against malaria

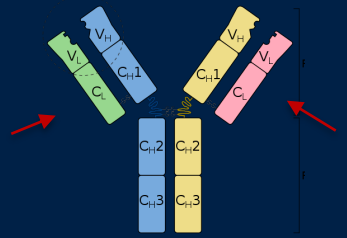


Leffler et al Science 2017

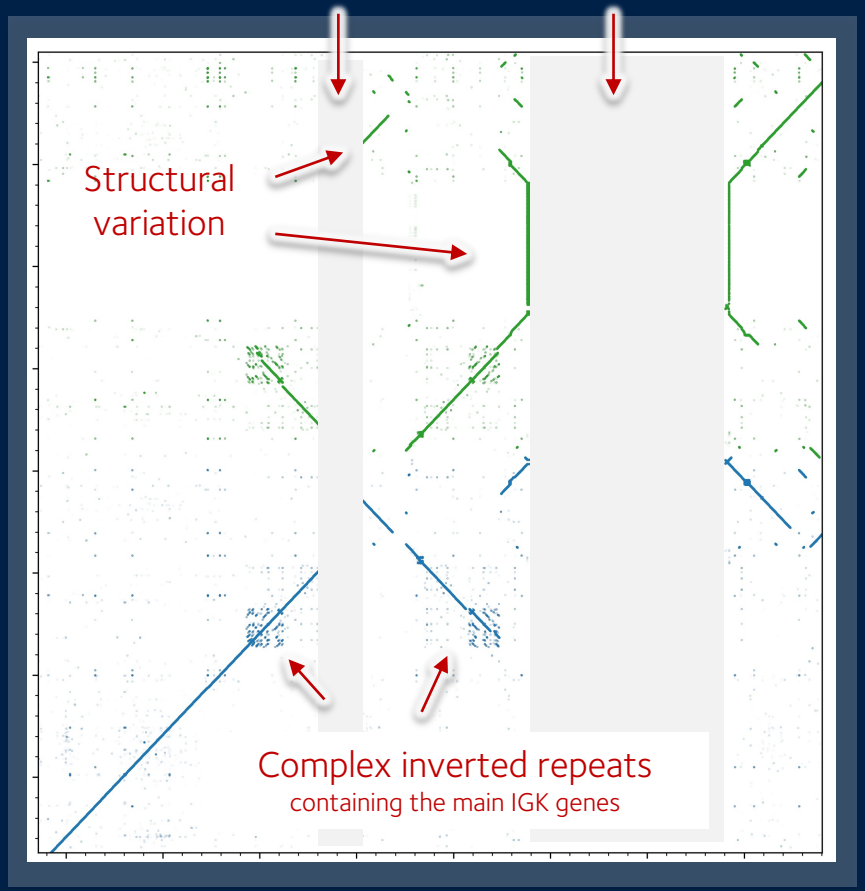


Algady et al AJHG 2018

Example:
Immunoglobulin
kappa chain locus
chromosome 2
Build 38 vs HV31 comparison



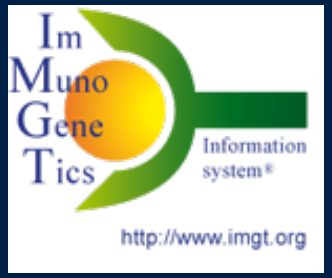
Build 38 assembly gaps (missing sequence)

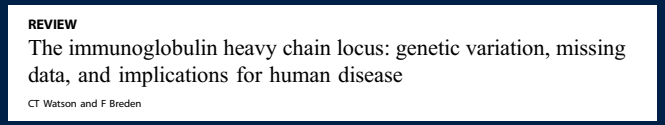


Each point is a 50-basepair segment (50-*mer*) shared identically between the two genomes

What is it about these regions?

- Targets of longstanding selection pressures due to pathogens and autoimmunity
- High genetic diversity
- Complex genome structure including multilevel repeats, mediating structural variation.

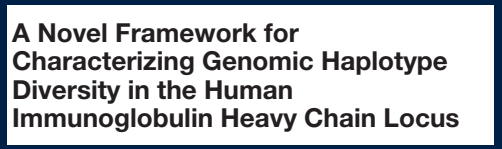



REVIEW

The immunoglobulin heavy chain locus: genetic variation, missing data, and implications for human disease

CT Watson and F Breden

Watson and Breden 2012


A Novel Framework for Characterizing Genomic Haplotype Diversity in the Human Immunoglobulin Heavy Chain Locus

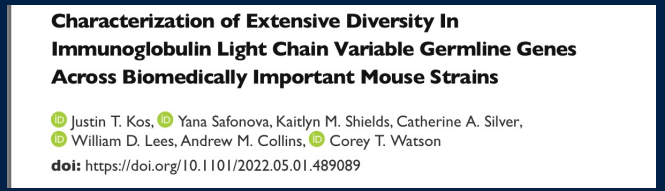
Rodriguez et al Frontiers in immunology 2020


RESEARCH ARTICLE

HUMAN GENOMICS

The complete sequence of a human genome

Nurk et al Science 2022


Characterization of Extensive Diversity In Immunoglobulin Light Chain Variable Germline Genes Across Biomedically Important Mouse Strains

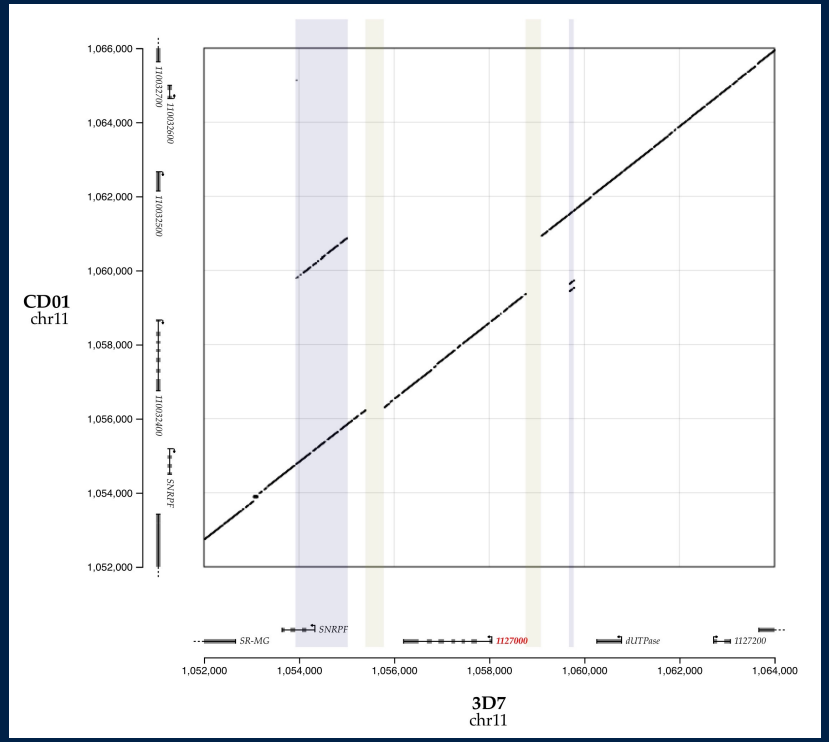
Justin T. Kos, Yana Safonova, Kaitlyn M. Shields, Catherine A. Silver, William D. Lees, Andrew M. Collins, Corey T. Watson

doi: https://doi.org/10.1101/2022.05.01.489089

Kos et al 2022

Genomic variation in these regions is poorly understood

Pathogen genomes are not immune to complex variation



A structural variant of the *P.falciparum* genome involved in human-malaria interaction

A resource of genomic data on a single healthy individual – **HV31**

Evaluate the use of 3rd-generation sequencing technologies to map structural variation in complex biomedically important regions.

Link to genome function.

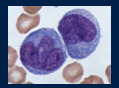
Using a single healthy European volunteer, codenamed HV31

A resource of genomic data on a single healthy volunteer - **HV31**

Short read genomic sequencing:

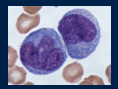
- Illumina Novaseq (~40x)
- MGI (~150x)

Long read genomic sequencing:


monocyte

- PacBio continuous long reads (~35x)
- Nanopore reads (~63x)
- PacBio 'HiFi' reads (~12x)

Linked-read genomic sequencing:


monocyte

- 10X linked reads (~40x)

Genome optical mapping:

- Bionano (~150x coverage by imaged fragments)

Data on genome function:

- RNA-seq, ATAC-seq, ChIP-seq:
  - CD4+ "helper T cell"
  - CD8+ "killer T cell"
  - CD19+ B cells
  - CD14+ monocytes

+ more data coming on newer ONT and PacBio versions.

Aim: evaluate use of these technologies, in particular for mapping structural variation.

Data available through EGA (EGAS00001005046)

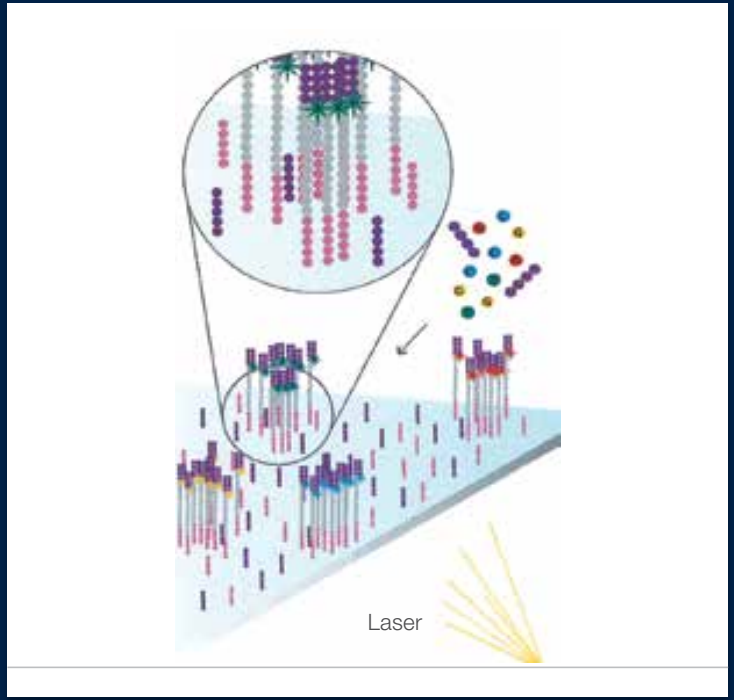
Quick survey of technologies – Illumina

Illumina short-read sequencing makes essential use of amplification of short fragments

They are sequenced in “lockstep“ by polymerase incorporation of labelled bases

Quality drops off as sequencing proceeds along the read due to molecules getting out of sync

=> Short reads (e.g. 150bp paired end)

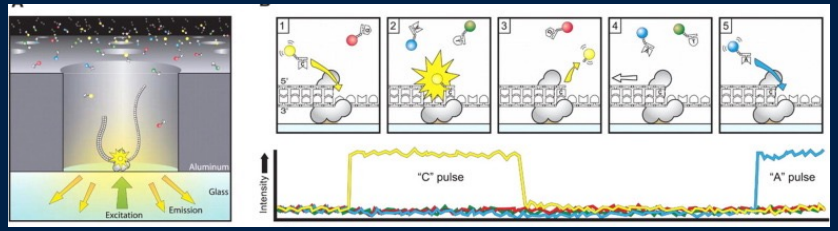


Quick survey of technologies – Long reads

PacBio continuous long read (CLR)

Up to 100s of kb

Figure: Rhoads and Fai Au 2015

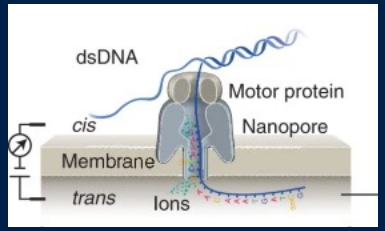


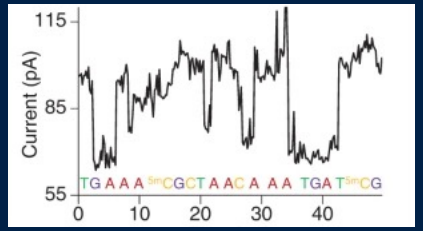
Pacbio: single DNA molecules are trapped in individual wells with polymerase, which incorporates fluorescent bases.

Nanopore (ONT)

Up to 100s of kb - or maybe even Mb

Figure: Wang et al 2021

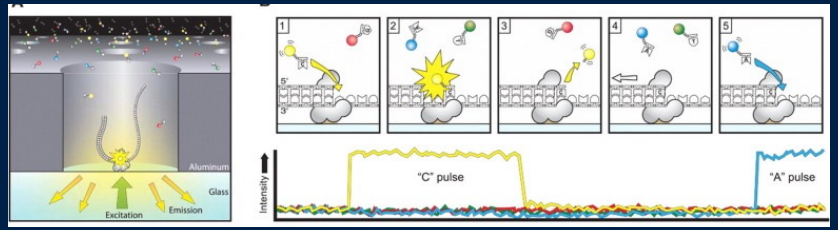




ONT: single DNA molecules are drawn through small pores. Ionic current flow is measured.

Quick survey of technologies – Long reads

Figure: Rhoads and Fai Au 2015

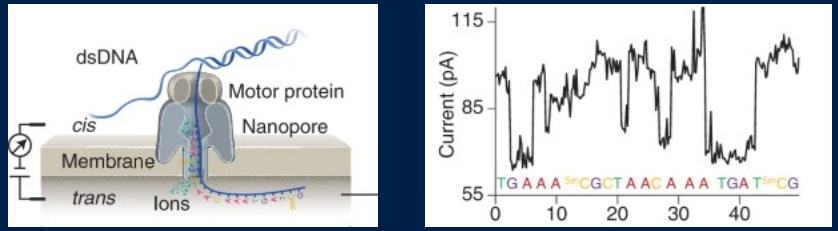


PacBio continuous long read (CLR)

Up to 100s of kb

Pacbio: single DNA molecules are trapped in individual wells with polymerase, which incorporates fluorescent bases.

Figure: Wang et al 2021



Nanopore (ONT)

Up to 100s of kb - or maybe even Mb

ONT: single DNA molecules are drawn through small pores. Ionic current flow is measured.

Pros and cons:

+ Long (100s of kb) sequencing lengths
+ Use of unamplified DNA
+ Software improvements might lead to new inference on existing data

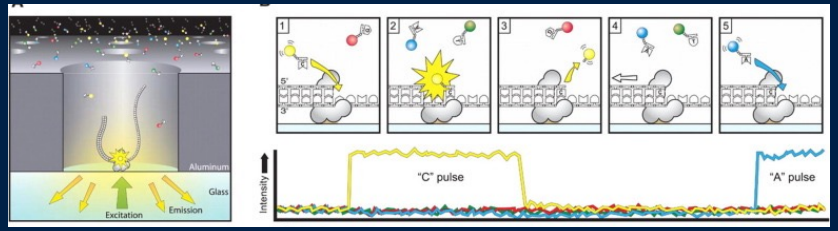
- High error rates
- Poor at calling homopolymer length

Quick survey of technologies – Long reads

PacBio continuous long read (CLR)

Up to 100s of kb

Figure: Rhoads and Fai Au 2015

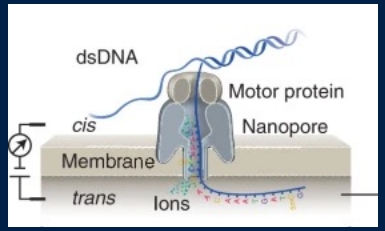


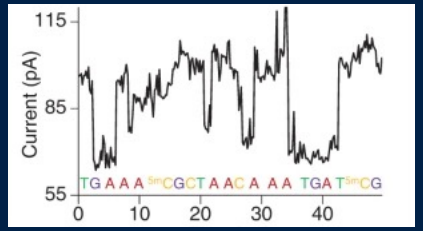
Pacbio: single DNA molecules are trapped in individual wells with polymerase, which incorporates fluorescent bases.

Nanopore (ONT)

Up to 100s of kb - or maybe even Mb

Figure: Wang et al 2021



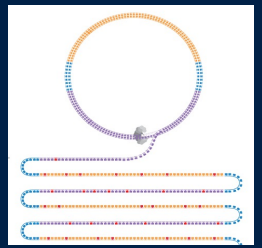


ONT: single DNA molecules are drawn through small pores. Ionic current flow is measured.

PacBio – 'HiFi' / circular consensus sequencing

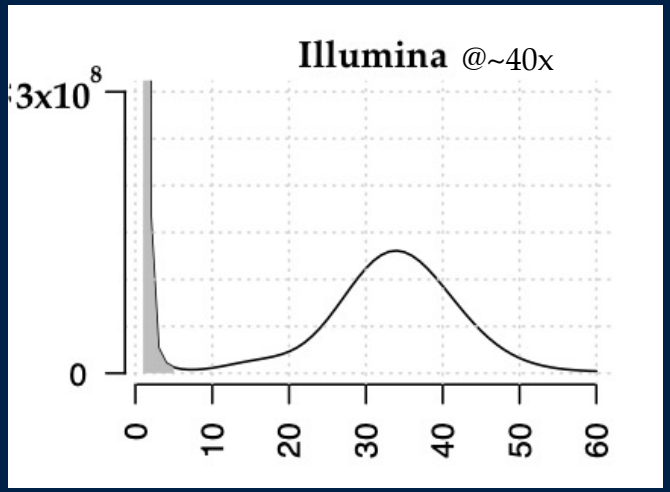
Size-selected to (say) 10kb

circularised DNA



Consensus 'HiFi' read generated in software from ~10-15 fold coverage of same molecule

A crude quality comparison



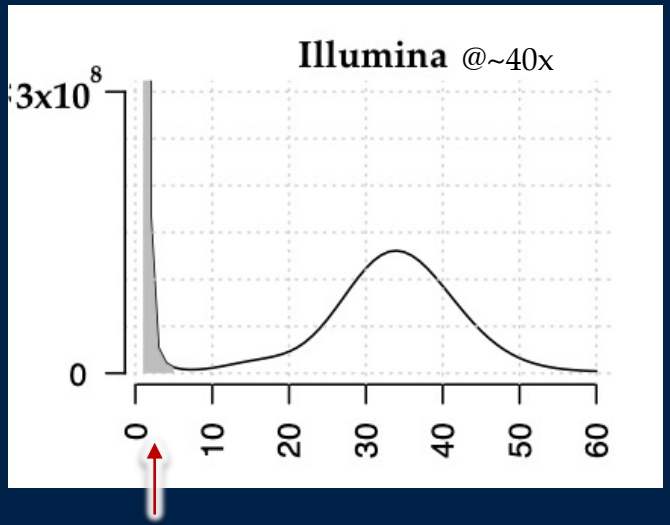
Kmer frequency histogram

Computed using jellyfish2

Count occurences of 31 bp kmers in reads. Those with low counts are probably errors.

"What proportion of sequenced kmers are correct?"

A crude quality comparison



About 8% of 31bp kmers in unfiltered reads look like **errors.**

Base error rate ~ 0.082/31 = 0.26%

i.e. ~ one error in 400bp

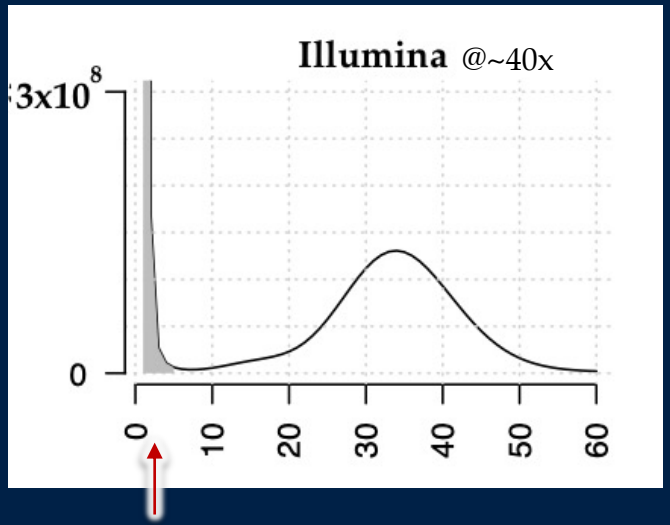
150bp PE

Kmer frequency histogram

Computed using jellyfish2:

Count occurences of 31 bp kmers in reads. Those with low counts are probably errors.

A crude quality comparison

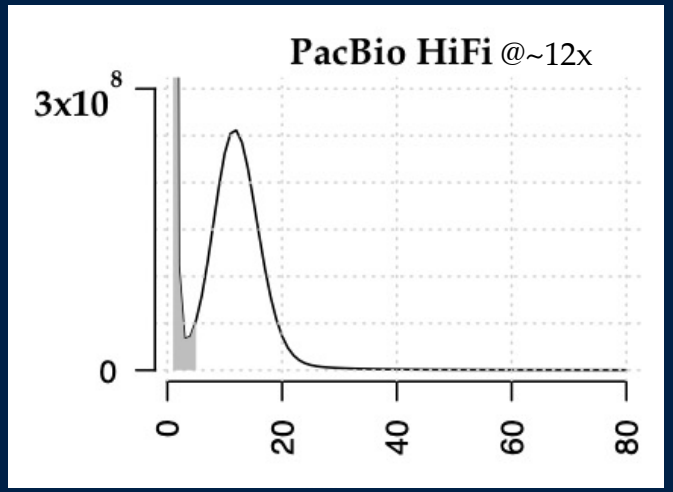


About 8% of 31bp kmers in unfiltered reads look like **errors.**

Base error rate ~ 0.082/31 = 0.26%

i.e. ~ one error in 400bp

150bp PE



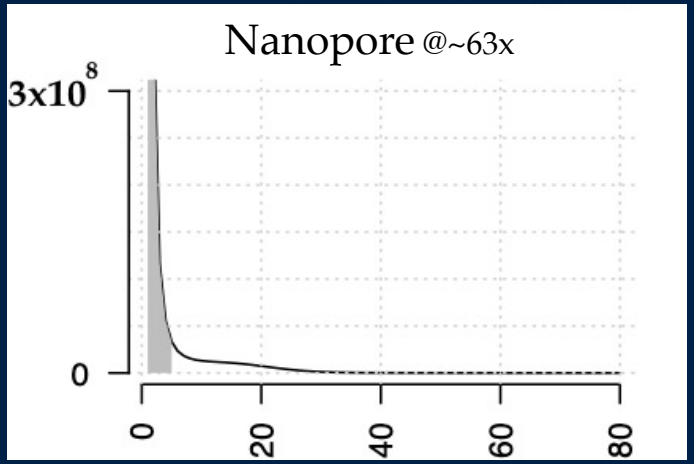
Also about 8% of 31bp kmers in unfiltered reads look like **errors.**

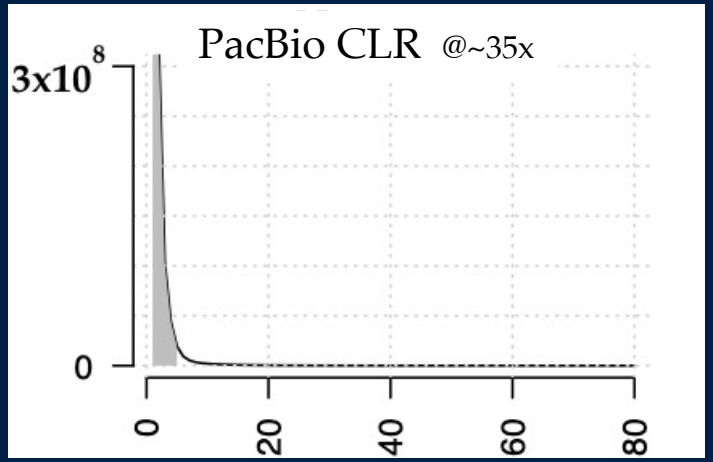
Base error rate ~ 0.078/4 = 0.25%

i.e. ~ one error in 400bp

10kb size-selected

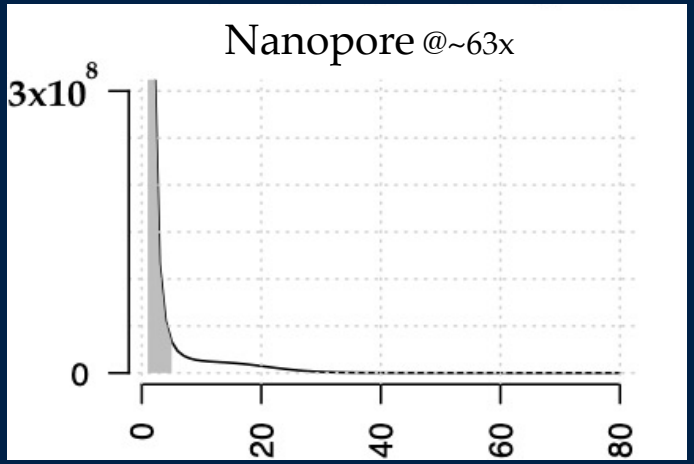
A crude quality comparison

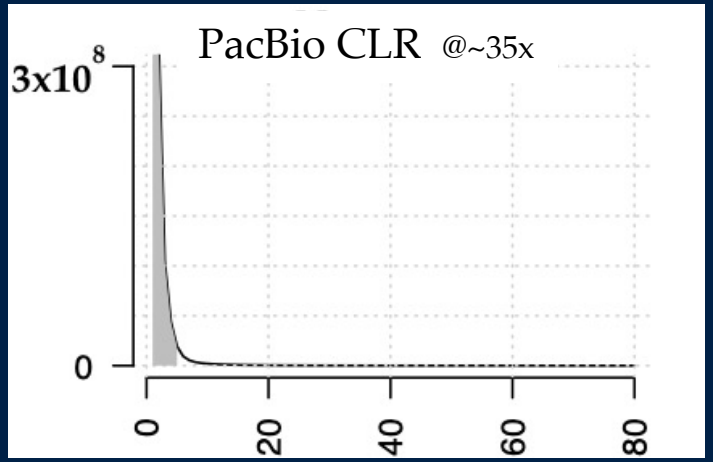




In R9 nanopore or Pacbio CLR, most kmers contain errors

A crude quality comparison



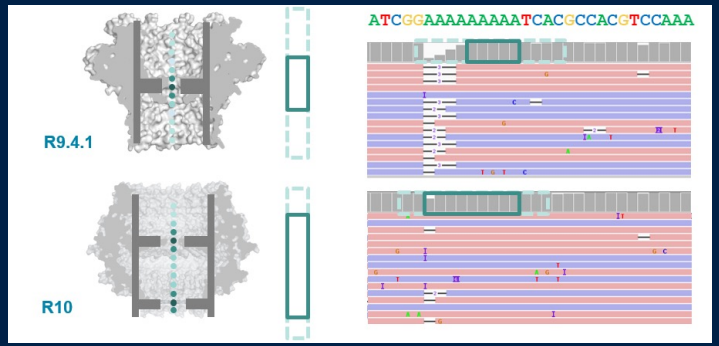


	count	type	contig_sequence	read_sequence	left_flank	right_flank	RepeatTract1	RepeatTract1_length
	<int>	<chr>	<chr>	<chr>	<chr>	<chr>	<chr>	<int>
1	43	D	T	""	TAA	TTT	T	6
2	41	D	C	""	TTT	CCT	C	3
3	40	D	C	""	CCT	CCC	C	4
4	39	D	C	""	CTT	CCT	C	3
5	38	D	C	""	CCT	CCT	C	3

	count	type	contig_sequence	read_sequence	left_flank	right_flank	RepeatTract1	RepeatTract1_length
	<int>	<chr>	<chr>	<chr>	<chr>	<chr>	<chr>	<int>
1	22	X	C	T	TTT	TTT	NA	NA
2	20	X	C	T	AGG	TGG	NA	NA
3	18	X	C	T	TGG	GAA	NA	NA
4	18	X	G	A	GCT	GGA	G	3

Most errors are in homopolymer repeats

Technology improvements may change this picture



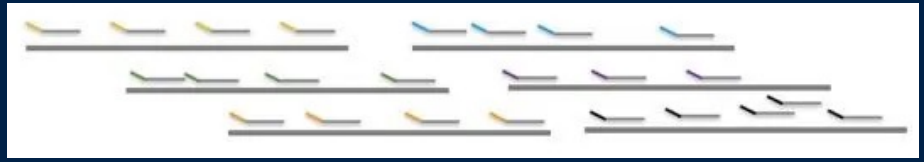
We are working to evaluate new iterations of the Oxford Nanopore and Pacbio technologies. Chemistry and software improvements may change this picture.

In regions where reads can be effectively aligned the errors can often be dealt with by consensus accuracy.

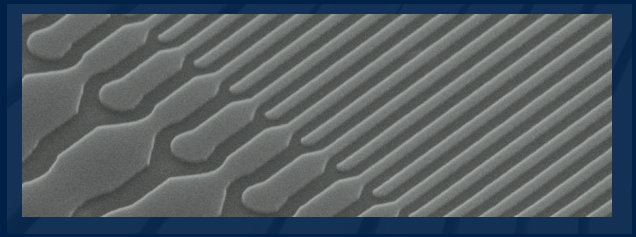
What about complex regions?

(Aside – Linked reads and optical mapping)

Linked reads = short-read sequencing, with fragments barcoded according to the long (~100kb) molecule they originate from



Optical mapping: individual long (100s of kb) molecules drawn through channels, with imaging of a specific 6bp motif.



Application: assemble core regions underlying the immune system

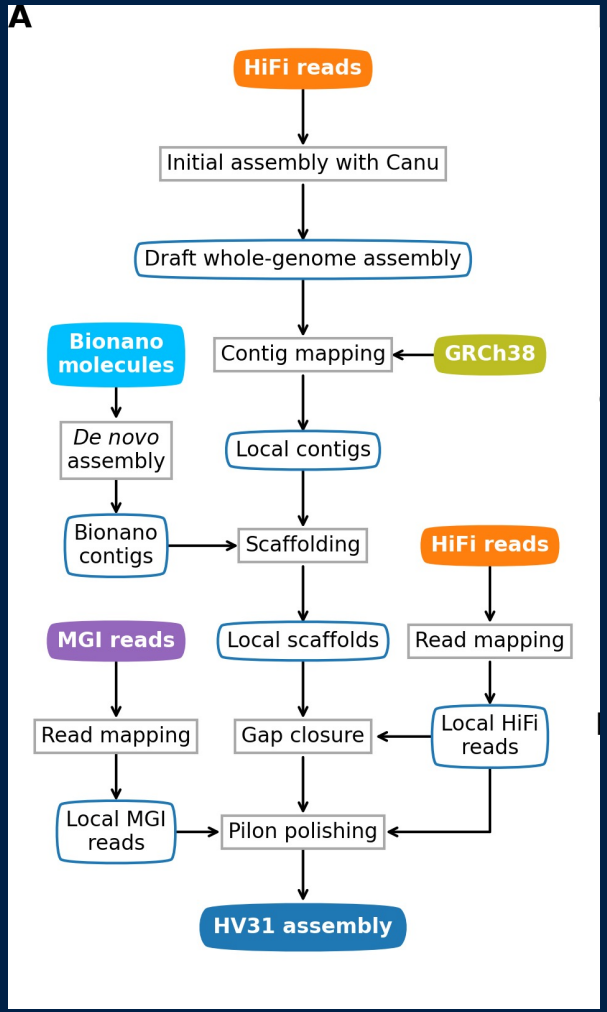
We used project data to attempt to de novo assemble eight regions that encode important components of the immune system:

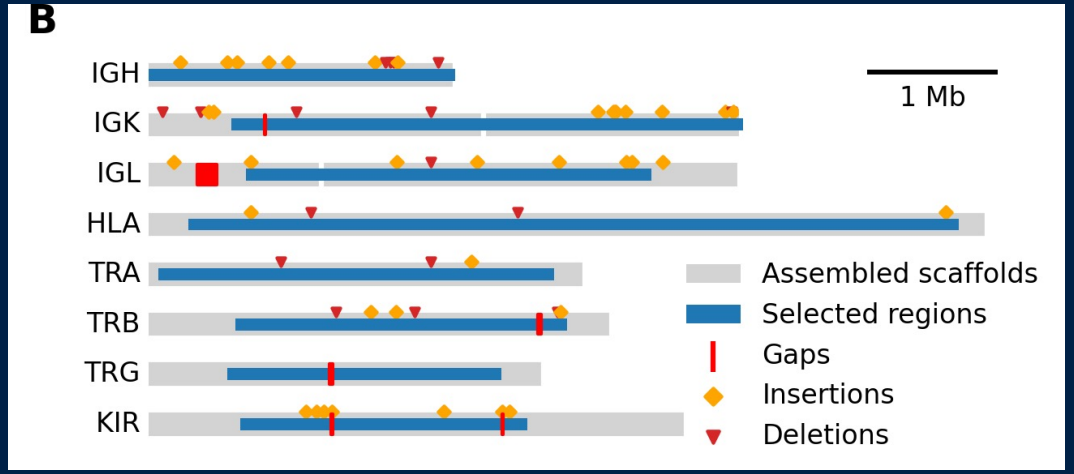
- Human Leukocyte Antigen (HLA)
- Killer cell immunoglobulin-like receptors (KIRs)
- Immunoglobulin heavy and light chains (IGH, IGK, IGL)
- T cell receptors (TRG, TRA/D, TRB)

We based the assembly around the PacBio HiFi data using other data to further improve and validate the assembly.

We then assessed heterozygous and homozygous structural variation relative to the genome reference assembly.

Zhang et al PLoS Comp Biol 2021

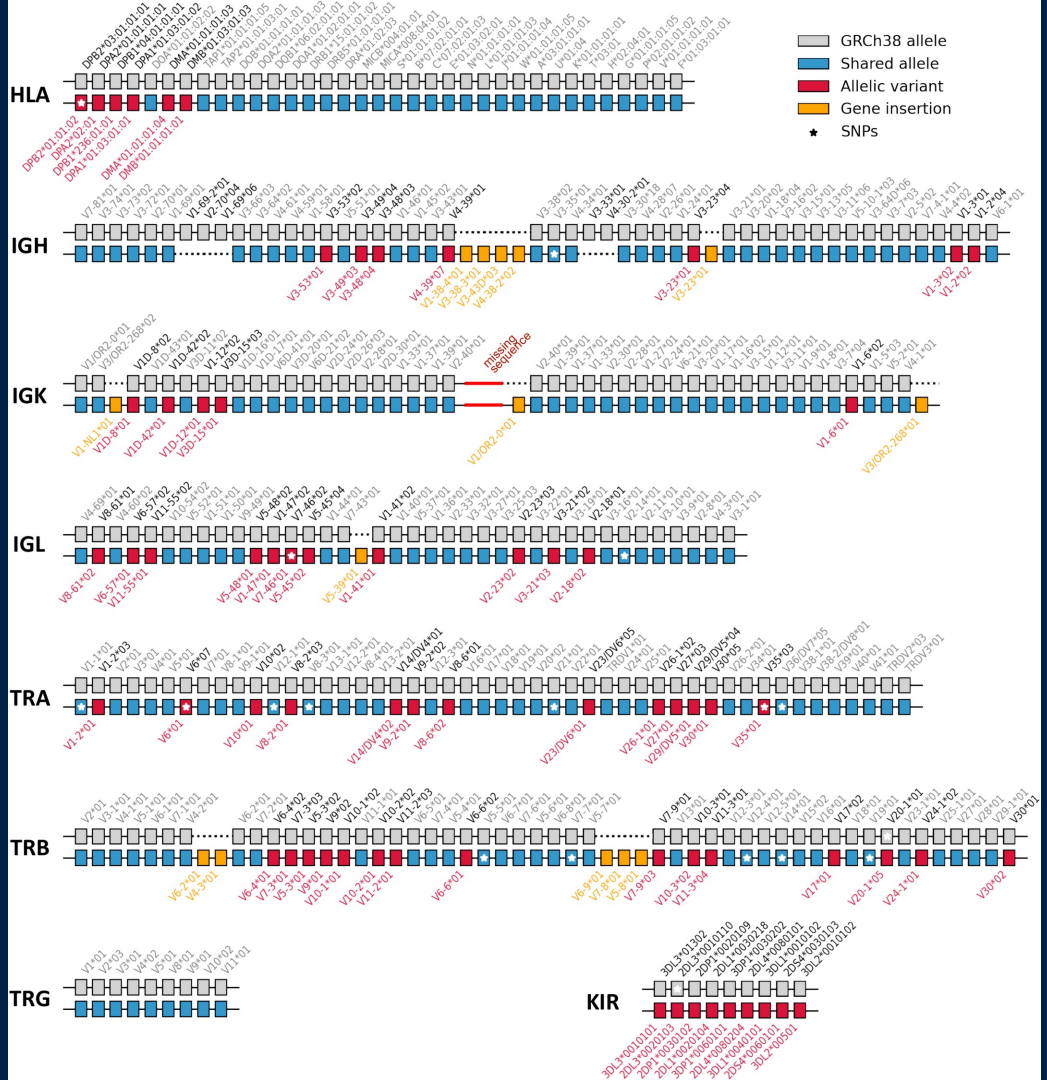




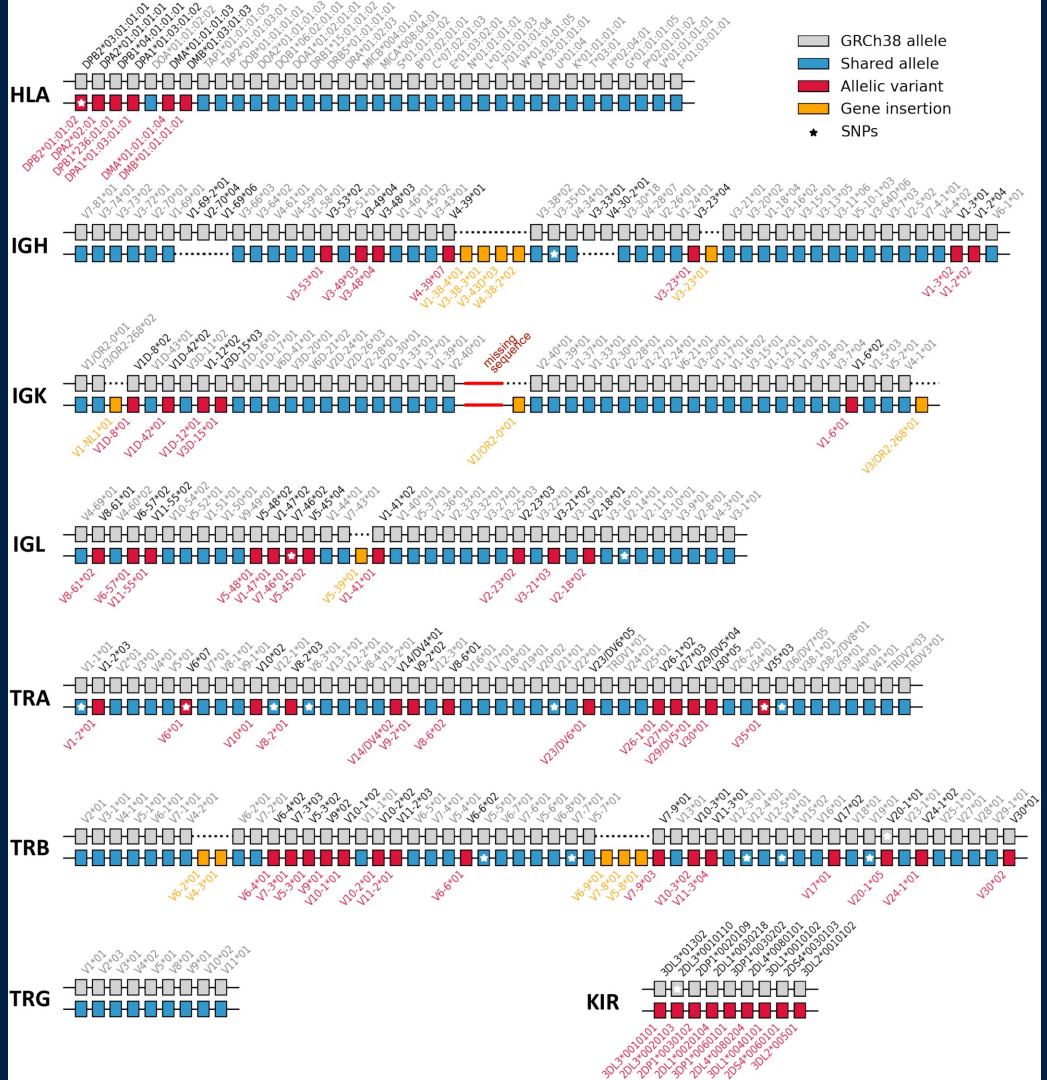
Most regions assembled in one contig (do still have contig breaks in IGK and IGKL).

Estimated error rate < 1 in 10kb - *provided* you believe we got the structure right.

Zhang et al PLoS Comp Biol 2021



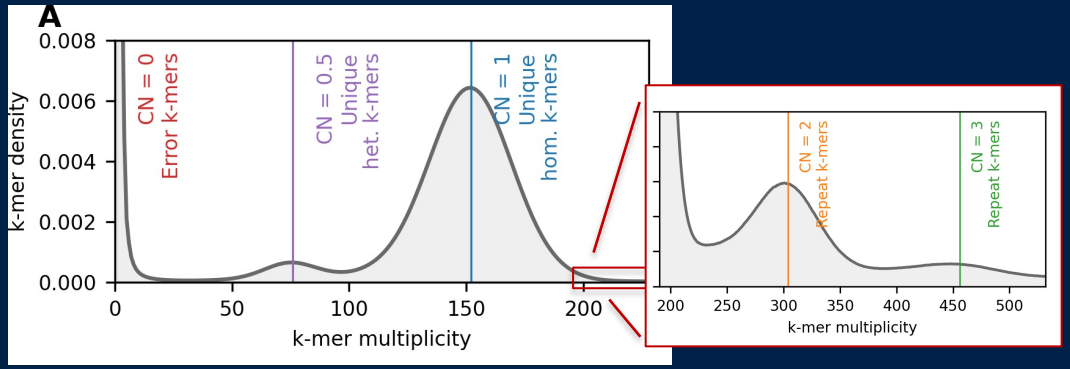
Large structural rearrangements seen in this single individual across most regions.



Large structural rearrangements seen in this single individual across most regions.

Warning: genome is diploid but assembly isn't!

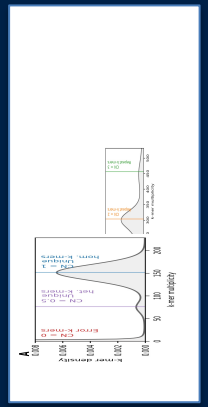
How do we know our structure is right?



Kmer histogram from short-read data as on earlier slide

Idea: compare assembly kmer multiplicity to kmer counts from all the short-read data (~150x kmer coverage) as validation.

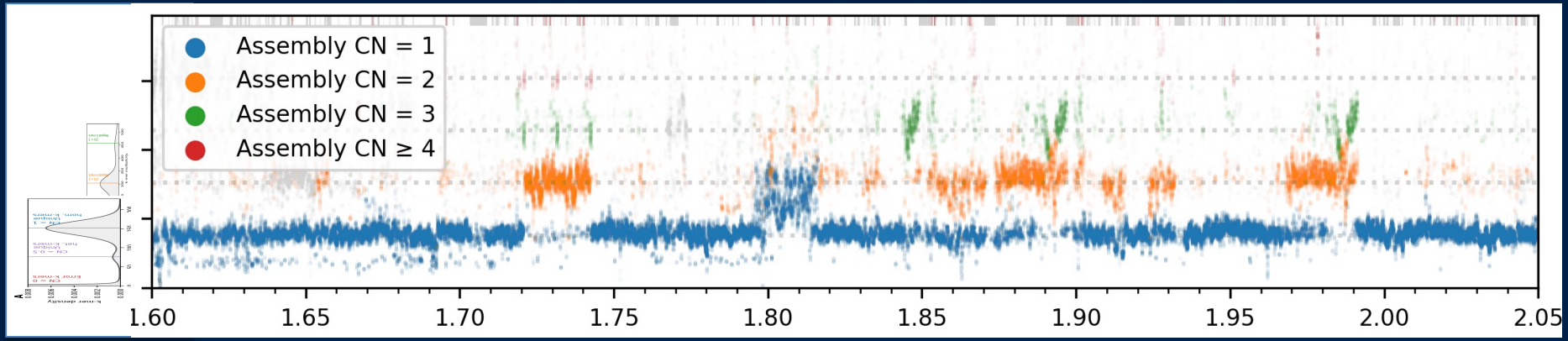
Validating the assembly structure



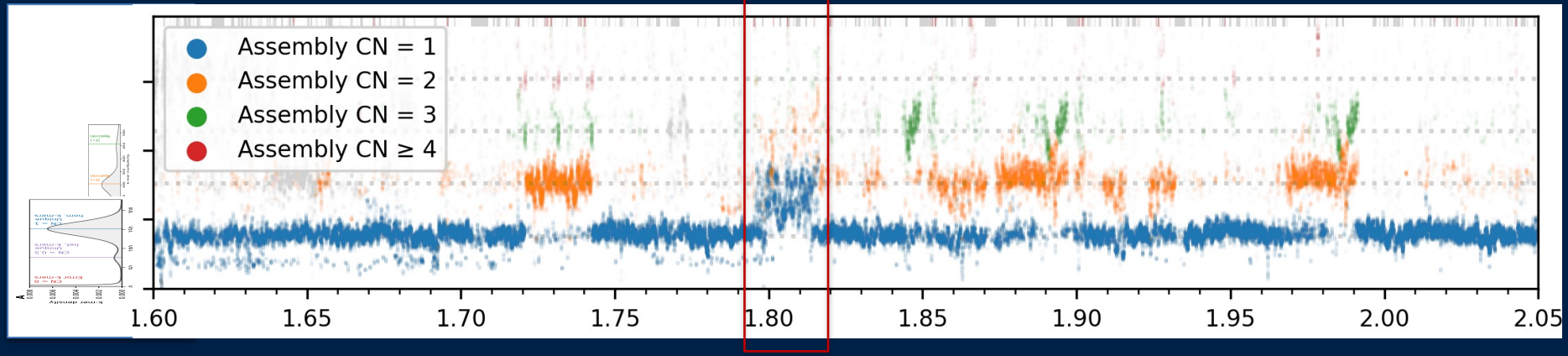
Idea: compare assembly kmer multiplicity to kmer counts from all the short-read data (~150x kmer coverage) as validation.

Validating the assembly structure

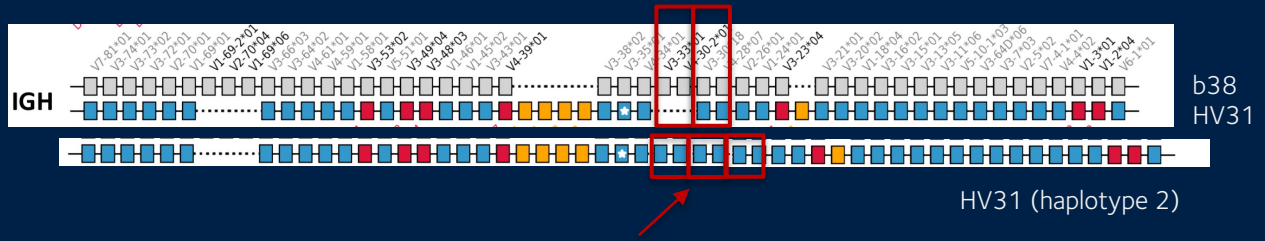
Shown below for part of IGH region



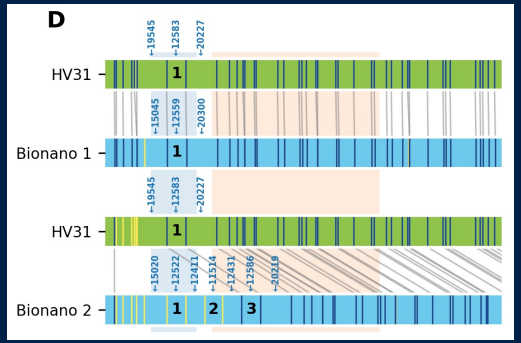
Validating the assembly structure: a heterozygous SV

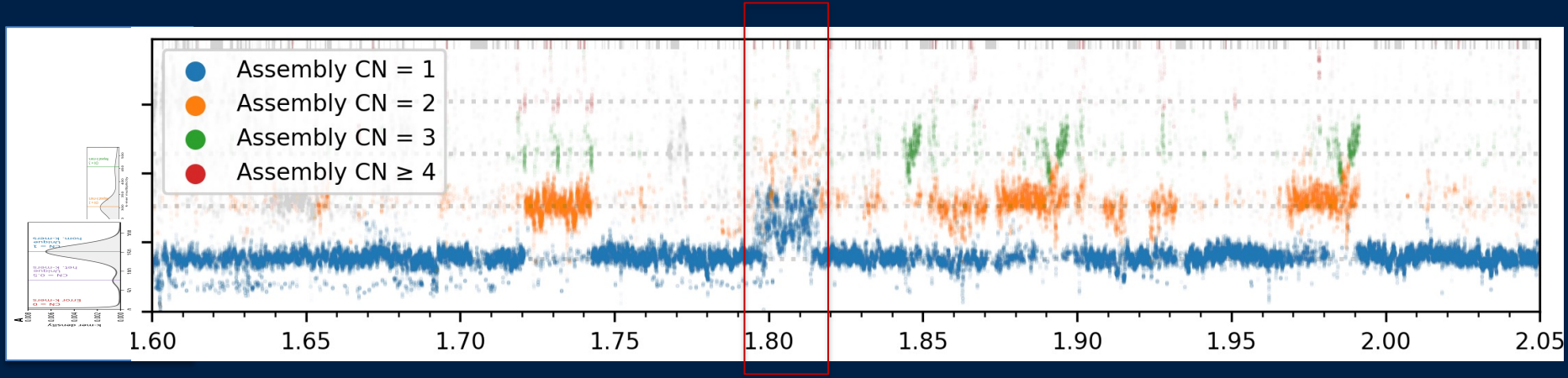


Validating the assembly structure: a heterozygous SV

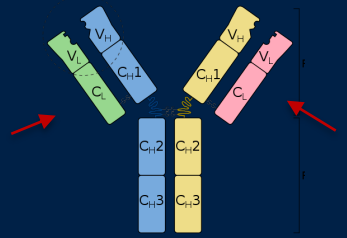


Unassembled haplotype has 3 copies

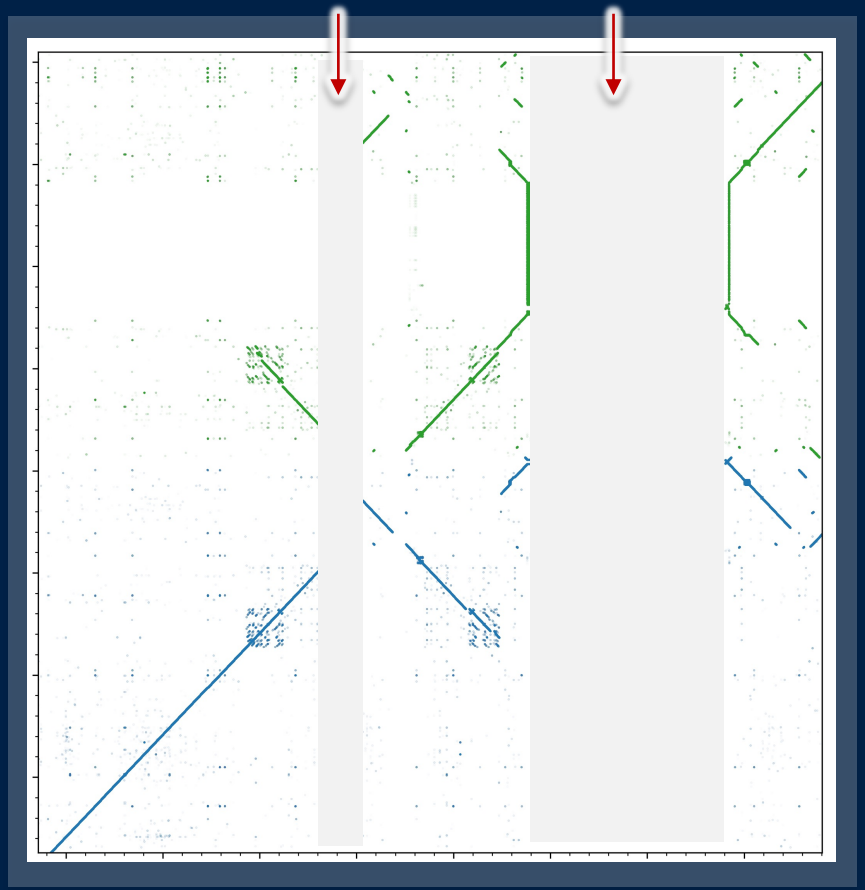




Example:
Immunoglobulin
kappa chain locus
chromosome 2
Build 38 vs HV31 comparison



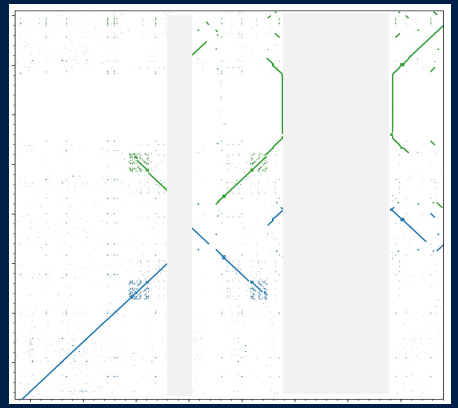
Build 38 assembly gaps (missing sequence)



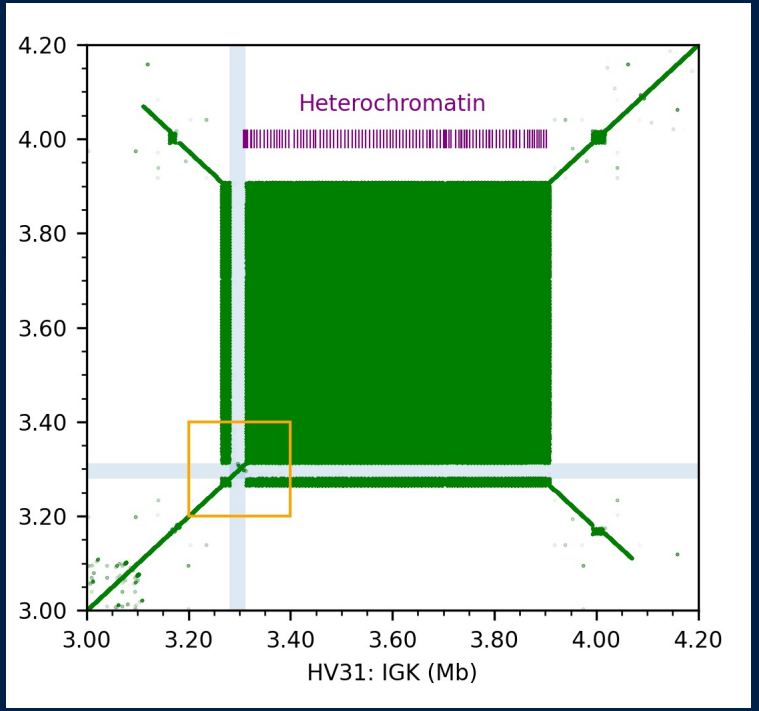
Each point is a 50-basepair segment (50-*mer*) shared identically between the two genomes

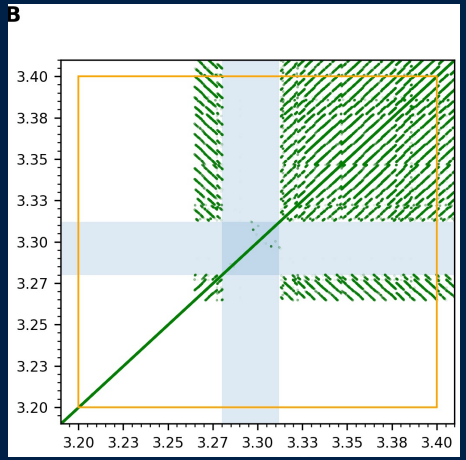
Filling the gaps – IGK region

HV31



b38





~650kb gap is filled with satellite repeats – but also unique sequence!

”115 imperfect tandem copies of 6 kb repeat units containing a 22bp signature from HSat2B family”

HV31 as a resource

Short reads:

- Illumina and CoolMPS, to ~200x

Long reads:

- PacBio - continuous long reads (~35x)
- Nanopore reads (~63x)
- PacBio - 'HiFi' reads (~12x)
- Coming soon! New ONT and Pacbio data – better accuracy + methylation!

Linked reads and optical mapping:

- Bionano optical mapping (~150x coverage by imaged fragments)
- 10X linked reads (~40x)

Functional data

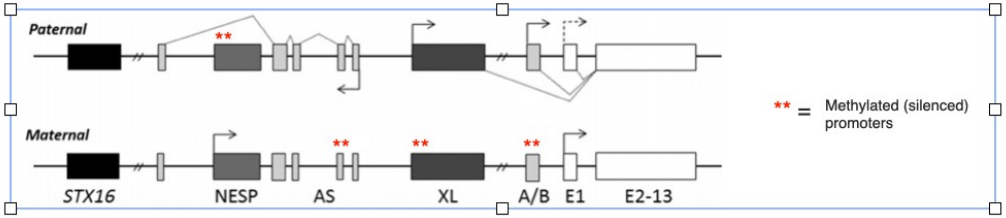
- ATAC-seq, ChIP-seq for histone modifications, RNA-seq in immune cell type
- Coming soon! Long-read transcripts.

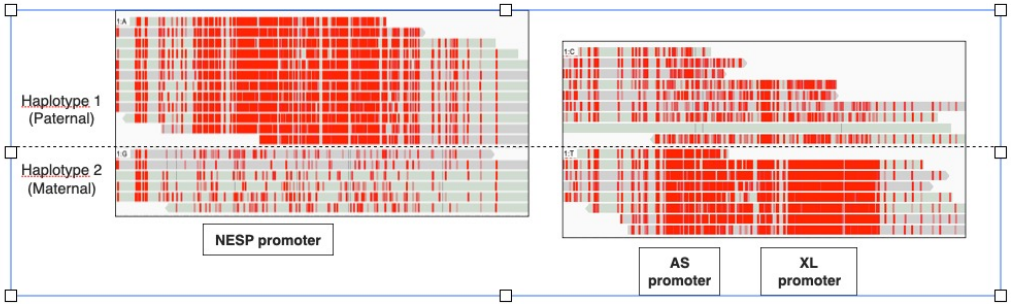
Future work…

- We are continuing to evaluate these technologies
- New iterations have better error rates and **they call methylation**
- Ongoing work to build the data into a joined-up resource, including functional data aligned to personal genomes.
- Remaining methodological issues around scaling up and resolving diploid sequence.

Pacbio methylation calls

Isoform-dependent imprinting of *GNAS*





Bréhin, A.-C., Colson, C., Maupetit-Méhouas, S., Grybek, V., Richard, N., Linglart, A., Kottler, M.-L., & Jüppner, H. (2015). Loss of Methylation at GNAS Exon A/B Is Associated With Increased Intrauterine Growth. The Journal of Clinical Endocrinology & Metabolism, 100(4), E623–E631. https://doi.org/10.1210/jc.2014-4047

wellcome centre human genetics

MRC Weatherall Institute of Molecular Medicine

wellcome sanger institute

University of Oxford

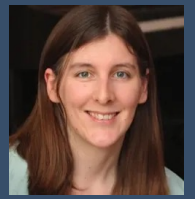
John Todd

Julian Knight

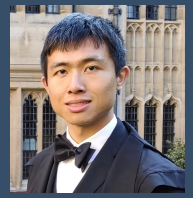
David Buck / OGC

Andy Brown

Tony Cutler



Helen Lockstone



Jia-Yuan Zhang

Hannah Roberts
David Flores
Justin Whalley
Rachael Bashford-Rogers
David Smith
Amy Trebes

Olga Mielczarek
Barbara Xella (WIMM)
Karen Oliver (Sanger)
Connor Davidson
Paolo Piazza
+ others...