IgDiscover analysis of adaptive immune receptor germline gene diversity

Martin Corcoran

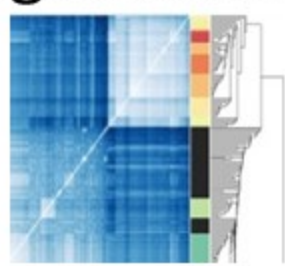
Gunilla Karlsson Hedestam Group
Karolinska Institute

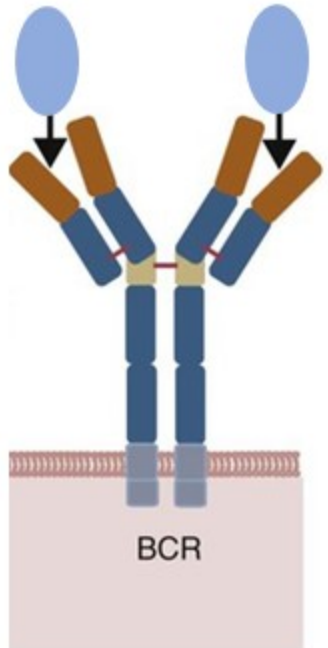
Germline Inference

Computational Identification of germline Ig or TCR sequences from expressed Rep-Seq data

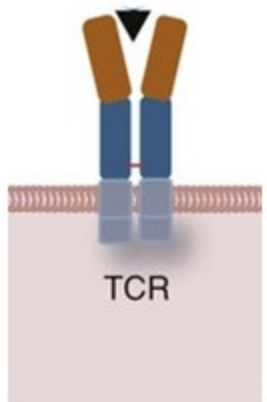
Isolation of naïve germline sequences that are used in multiple independent antibody or TCR rearrangements

IgDiscover

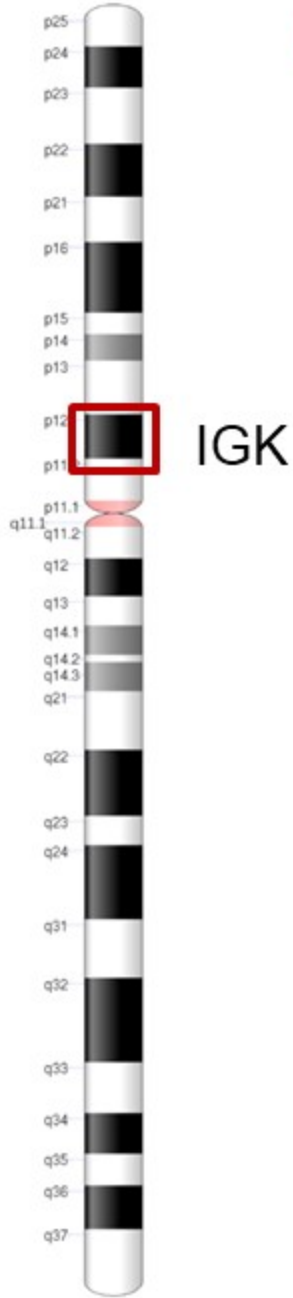




Adaptive immune receptor loci are among the most polymorphic in the human genome

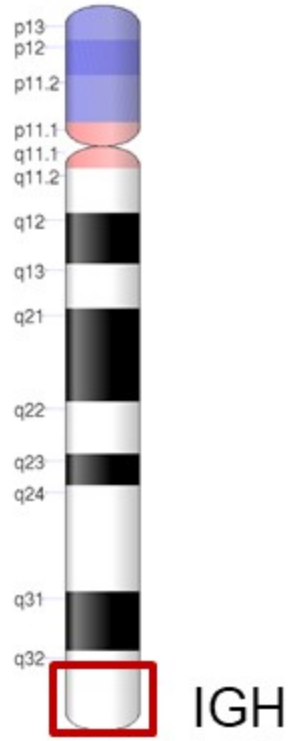


Genomic localization



IGK

Chr. 2



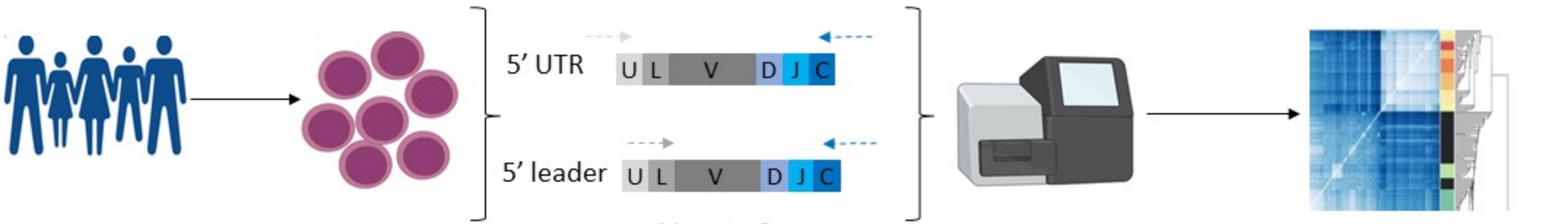
IGH

Chr. 14

IGL

Chr. 22

IgDiscover library production



PBMCs

Create libraries from expressed recombined repertoires

Illumina sequencing (2 X 300 bp)

IgDiscover analysis

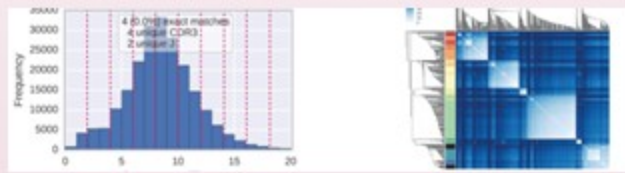
- Germline gene inference
- Known and Novel alleles
- Genomic Validation

IgDiscover analysis overview

Rep-Seq Library > MiSeq 2 x 300 bp

High frequency alleles

- Assignment of sequences to closest reference database sequence
- Windowed and linkage clustering



- Discovery of germline candidates

  Number of unique CDR3s, J genes, Germline Filtering
- Replacement of reference database with newly created database

V-gene "Starting" Database

Low frequency alleles

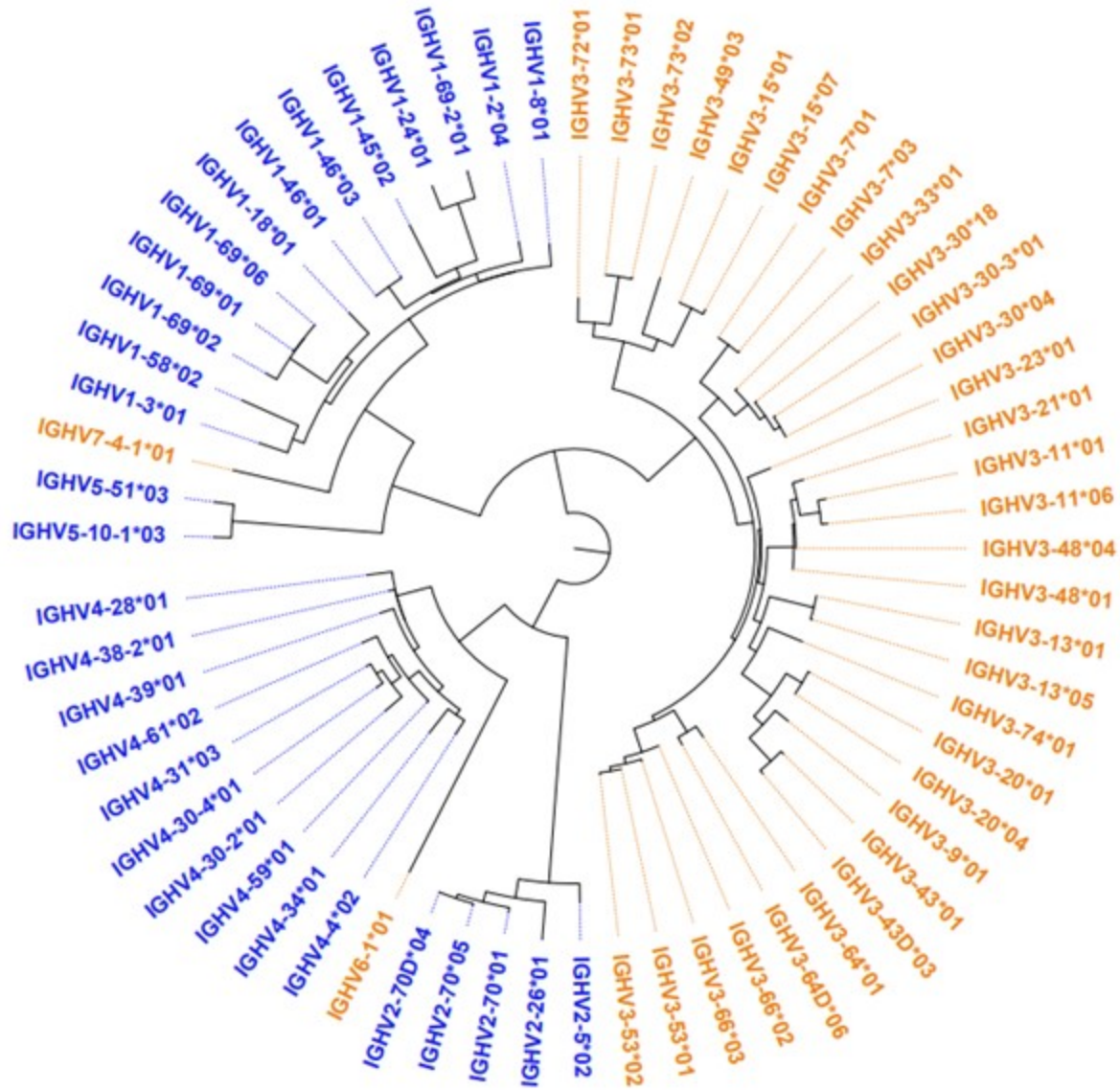
- Further iterations, replacement of previous database, whitelisting

Replacement Database Iteration *n*

Individualized database

Corcoran, Phad et al. Nature Communications 2016

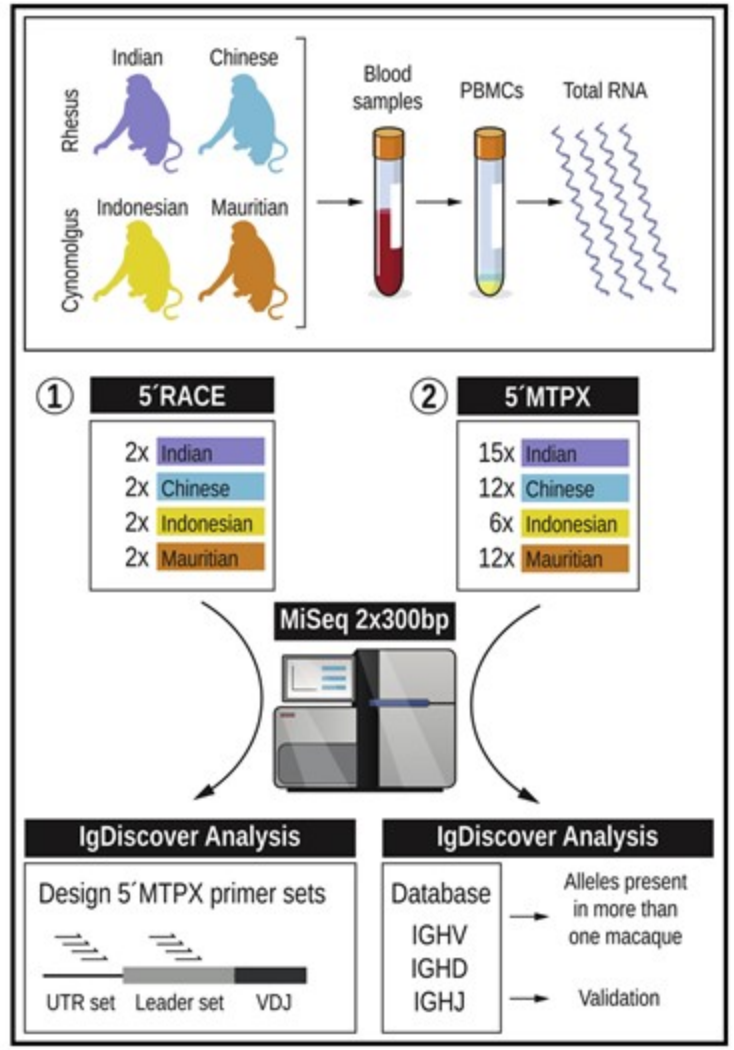
Individualized IGHV Genotype



Individualized database usage

- Accuracy of SHM estimation
- Lineage tracing of diverse antibodies
- Identification of gene/allele content to enable association studies

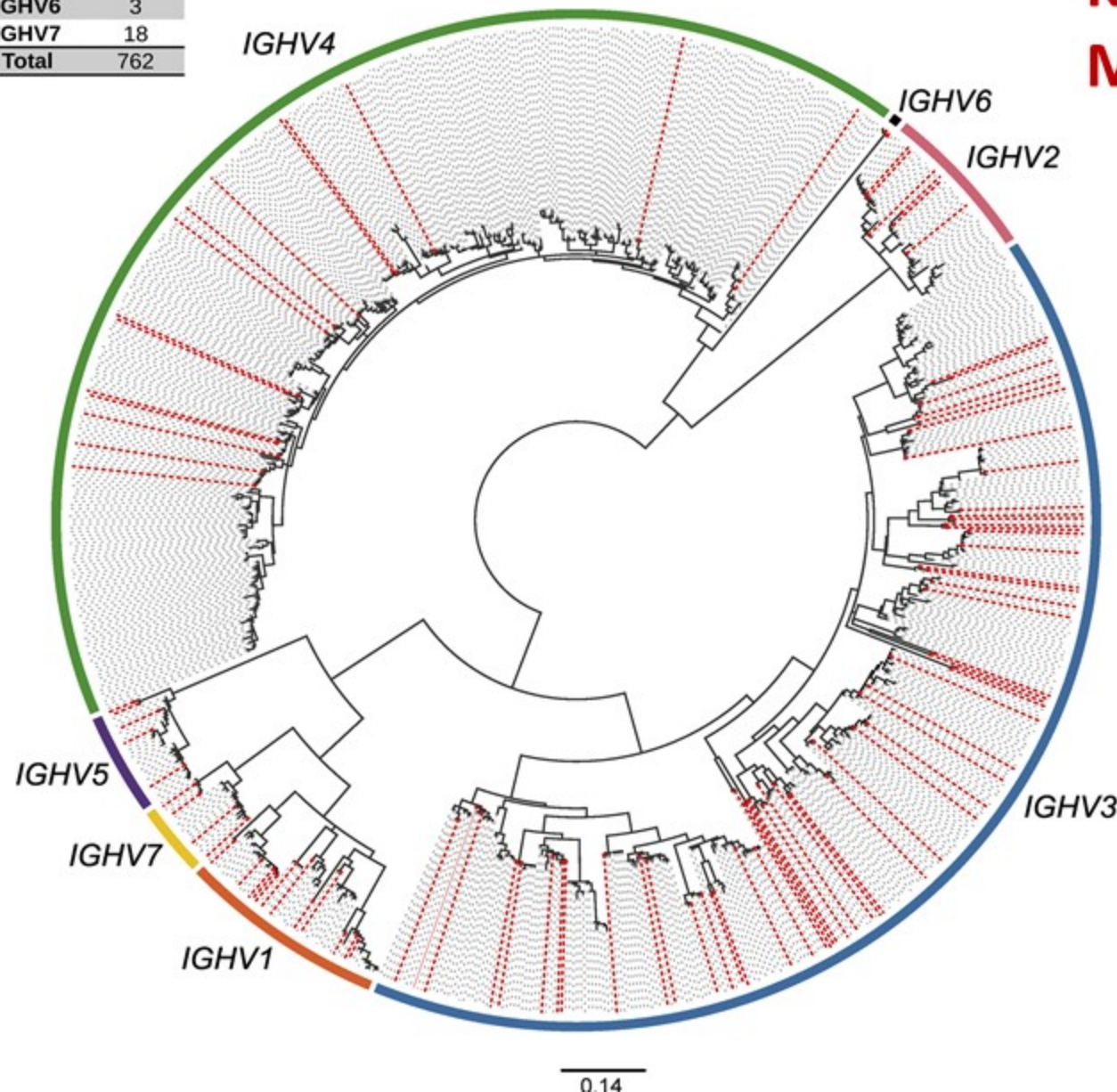
Using Germline Inferral to assemble a Rep-Seq suitable database

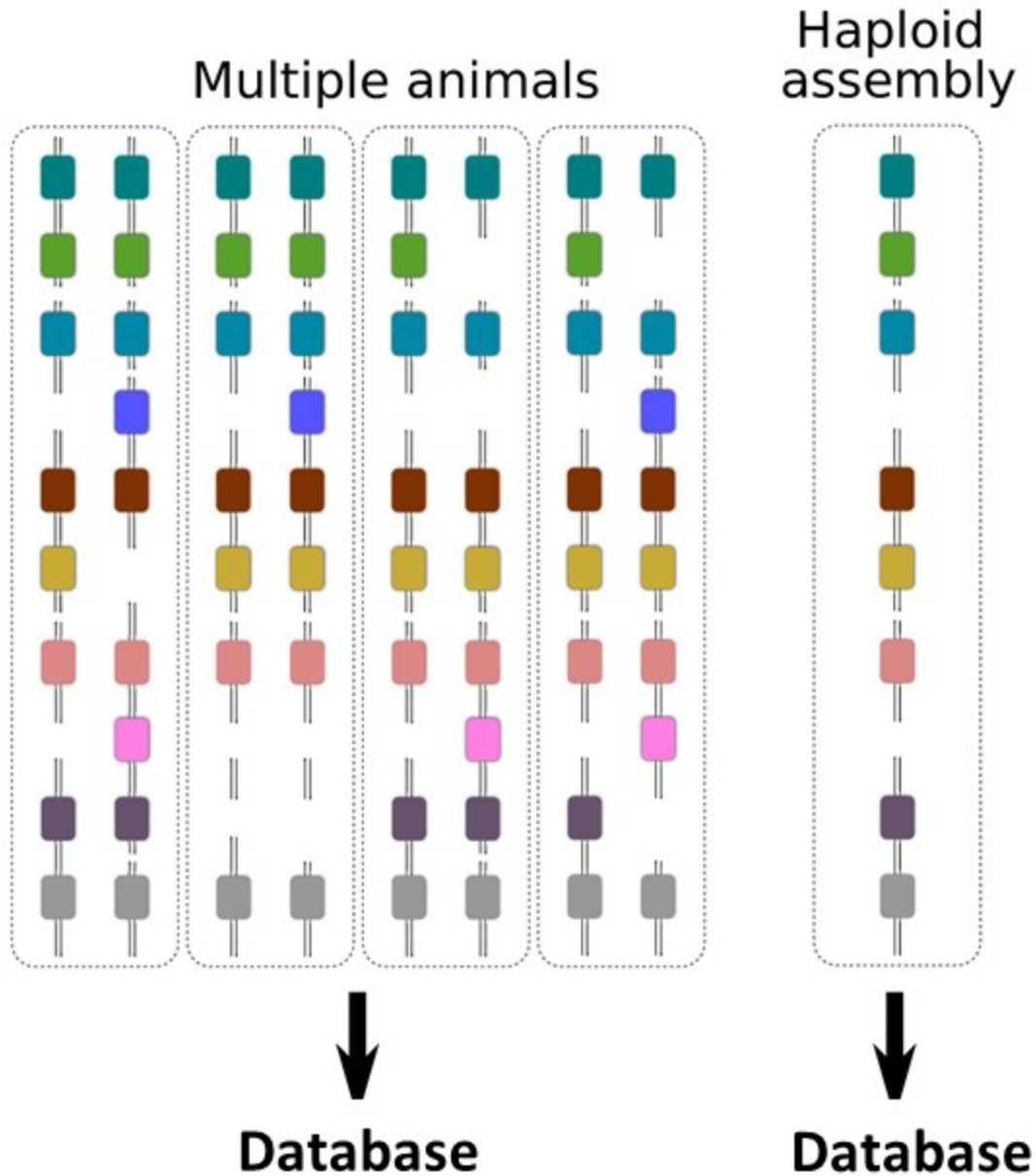


Vázquez Bernat N, et al. Rhesus and cynomolgus macaque immunoglobulin heavy-chain genotyping yields comprehensive databases of germline VDJ alleles. Immunity. 2021

Rhesus V alleles	
IGHV1	48
IGHV2	39
IGHV3	300
IGHV4	330
IGHV5	24
IGHV6	3
IGHV7	18
Total	762

KIMDB
Karolinska Institute
Macaque database




Multiple animals
Haploid assembly
Database
Database

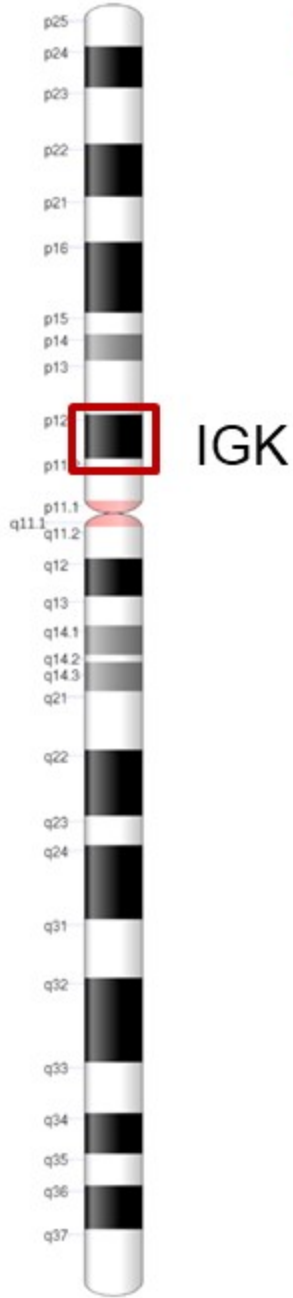
Inferred germline database

- Highly comprehensive in terms of germline sequences
- Problems with gene nomenclature remain (non-human species)

IgDiscover updated version

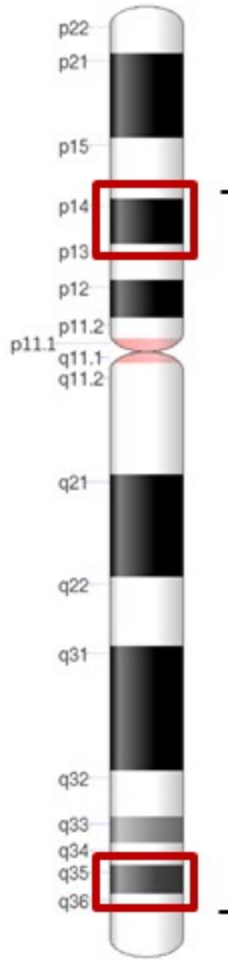
- MIARR compliant
- TCR discovery
- Improved haplotyping facility
- Genotype module – core count
- J and D discovery

Genomic localization



IGK

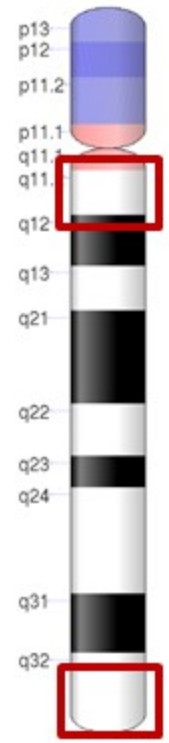
Chr. 2



TRGV

TRBV

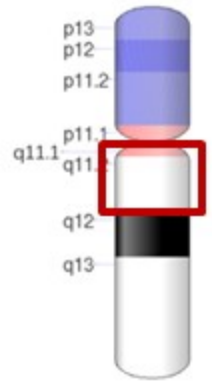
Chr. 7



TRAV +
TRDV

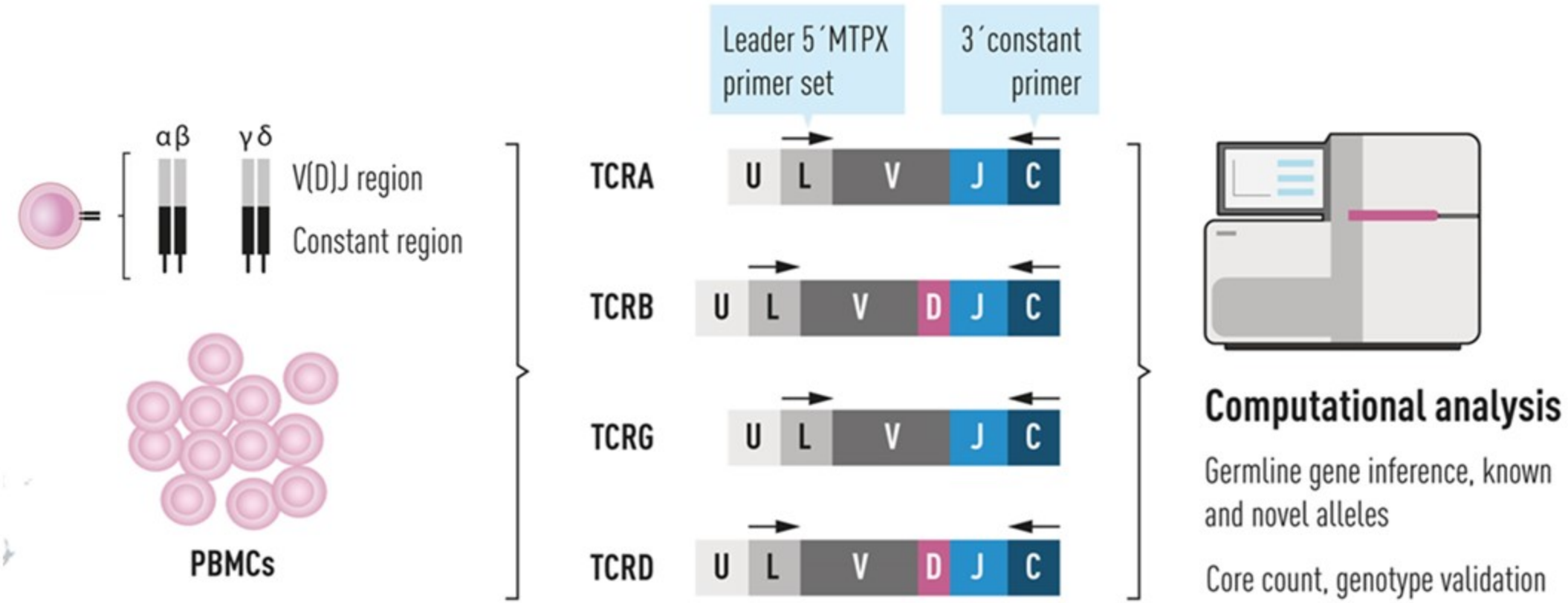
IGH

Chr. 14

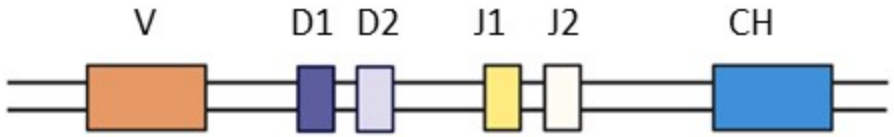


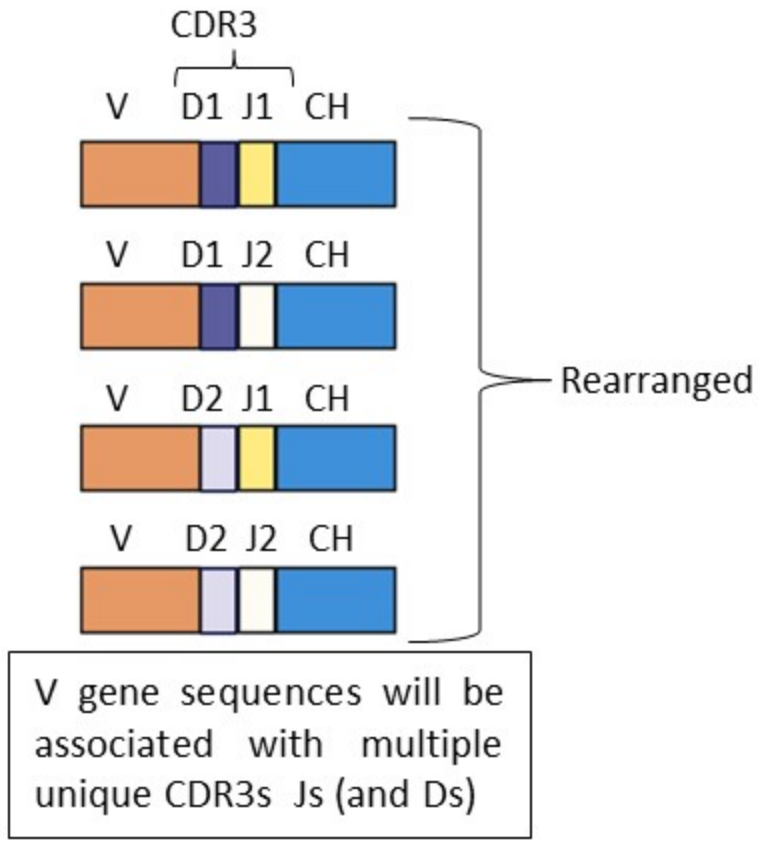
IGL

Chr. 22


αβ
γδ
V(D)J region
Constant region
PBMCs
Leader 5´MTPX primer set
3´constant primer
TCRA
TCRB
TCRG
TCRD
U
L
V
D
J
C
Computational analysis
Germline gene inference, known and novel alleles
Core count, genotype validation

TCR Germline Discovery





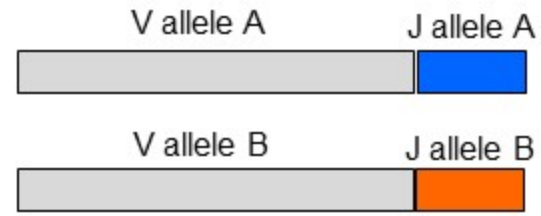
TCR specific

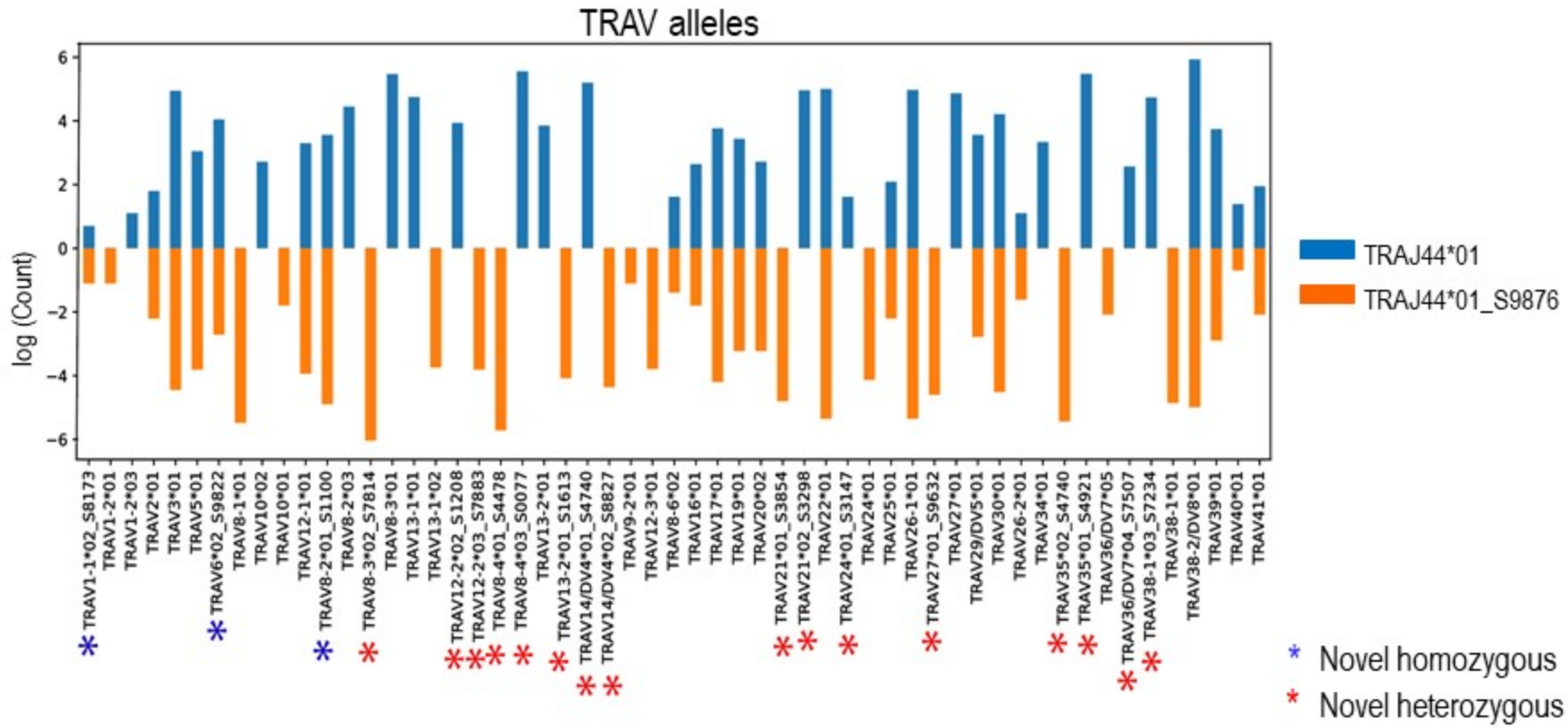
Validation

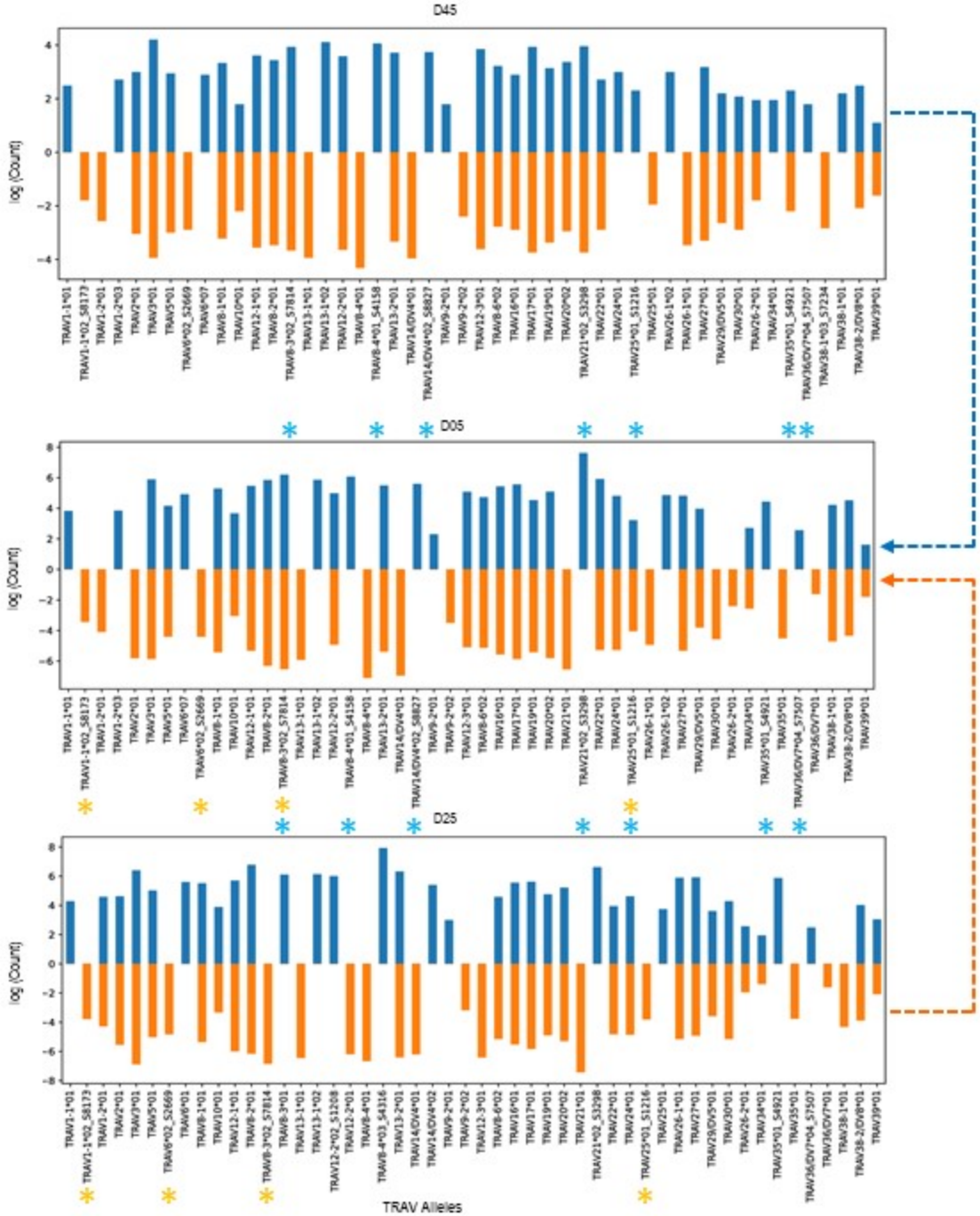
- PCR and Sanger sequencing
- Haplotype analysis
- corecount genotype analysis

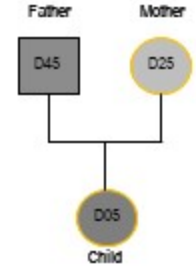
Validation by inferred haplotype analysis

V allele status can be determined through linkage to heterozygous J alleles





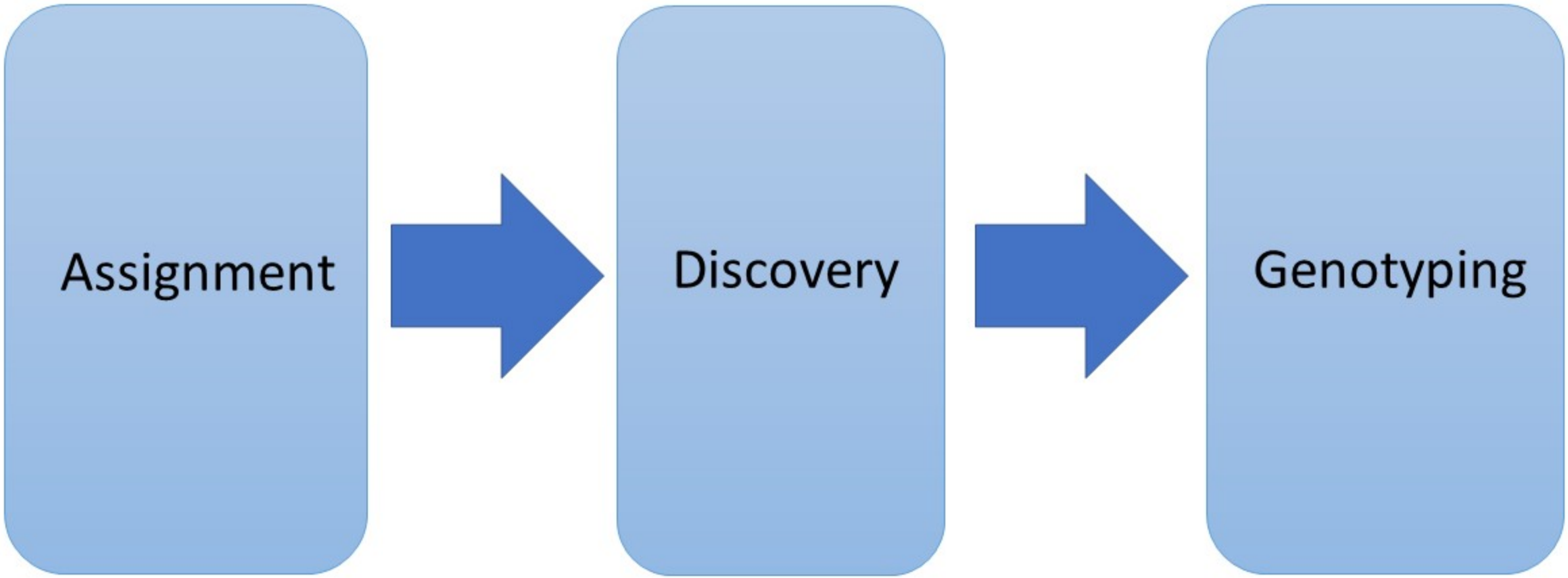




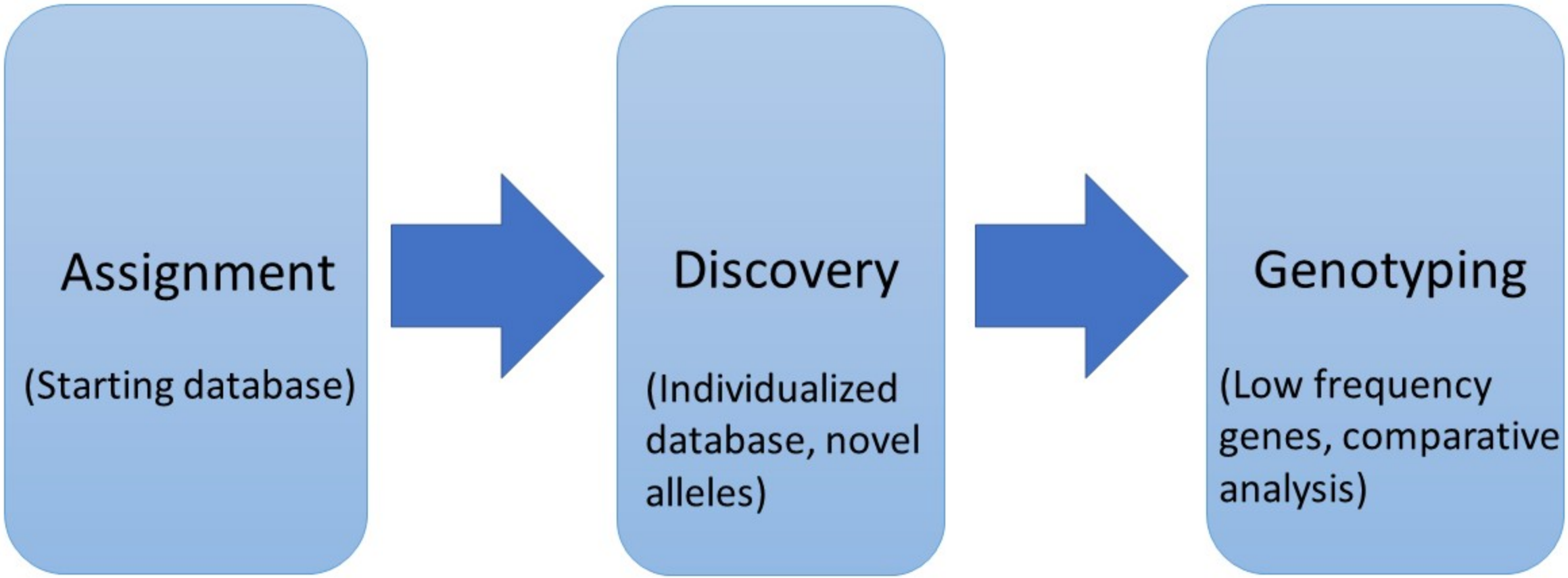
Familial analysis

- 9 novel alleles inherited
- 7 novel alleles from father
- 4 novel alleles from mother
- 2 novel alleles from both

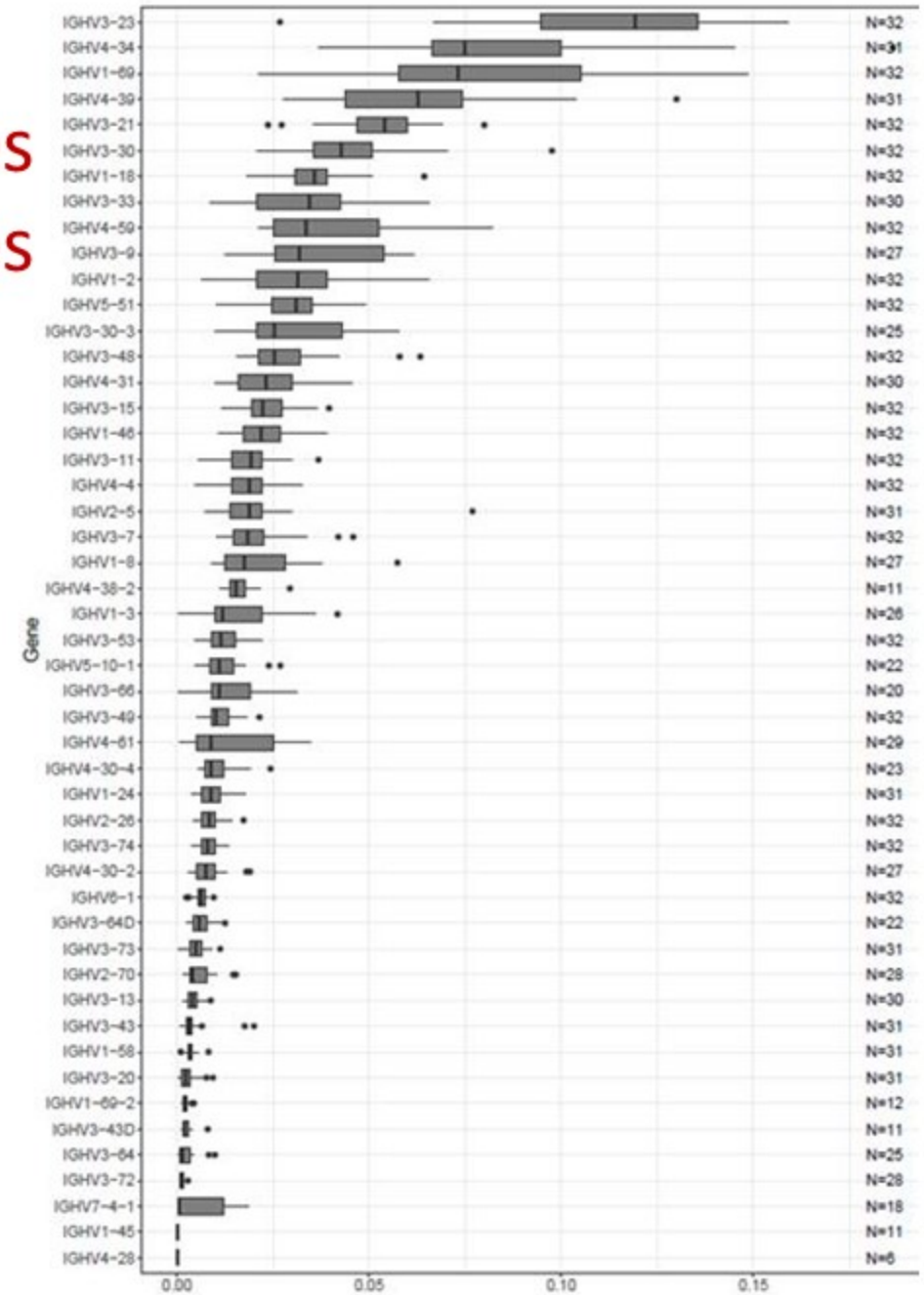
Genotype comparison corecount analysis module



Genotype comparison corecount analysis module



Unmutated sequences in human IgM libraries



Genotype comparison
corecount analysis module

- Most recombined VDJ sequences will not contain full length Vs, Ds or Js
- They will contain a 'core' of the appropriate V, D and J allele
- Using a modified 'core' database allows highly accurate identification of even low frequency genes and alleles from the filtered.tab output of IgDiscover.
- core sequence = full length allele minus nucleotides frequency lost during recombination

Reference database sequence

>IGHD2-21*01
AGCATATTGTGGTGGTGATTGCTATTCC
>IGHD2-21*02
AGCATATTGTGGTGGTGACTGCTATTCC

Corecount database sequence

>IGHD2-21*01
CATATTGTGGTGGTGATTGCTATT
>IGHD2-21*02
CATATTGTGGTGGTGACTGCTATT

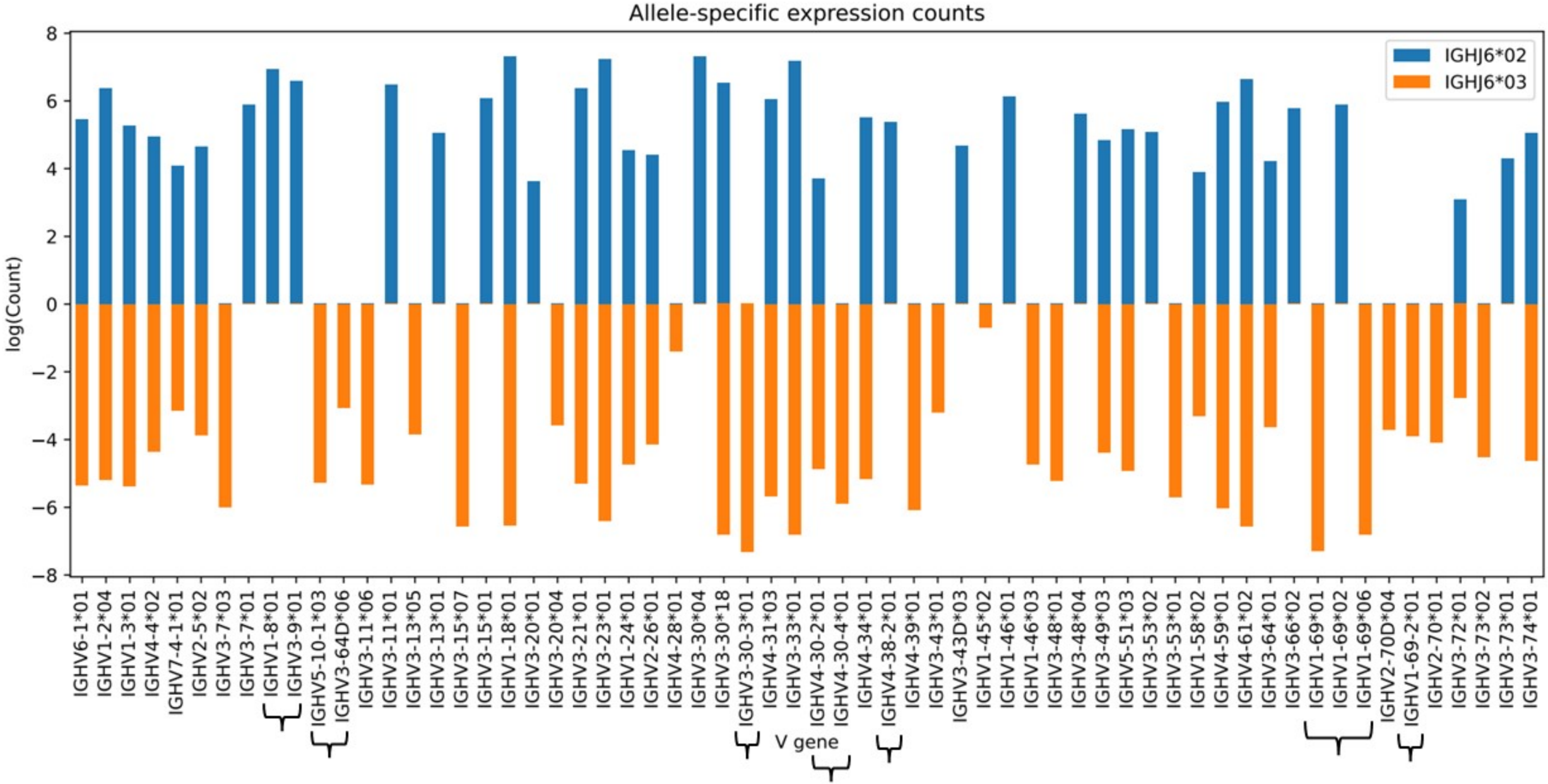
Genotype comparison corecount analysis module

Requirements

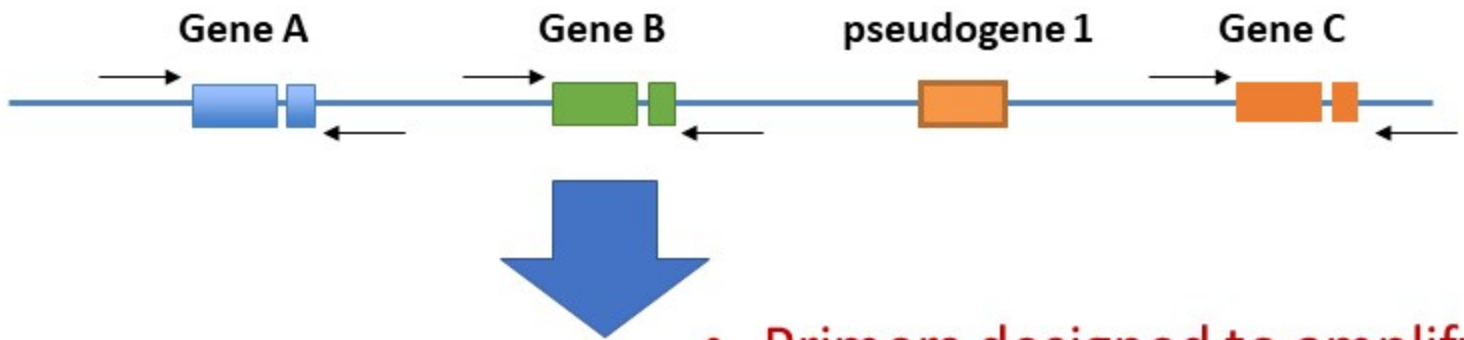
- Comprehensive database
- Rep-Seq library with sufficient unmutated sequences (IgM, IgD, IgK, IgL, TRA, TRB, TRD, TRG)
- High quality sequence library (MiSeq, 2 x 300 V3, avoid high clustering density)

Testing the accuracy of the corecount genotyping

- Analysis of subject of ”known” genotype
- High gene content individual


Allele-specific expression counts
log(Count)
8
6
4
2
0
−2
−4
−6
−8
IGHJ6*02
IGHJ6*03
IGHV6-1*01
IGHV1-2*04
IGHV1-3*01
IGHV4-4*02
IGHV7-4-1*01
IGHV2-5*02
IGHV3-7*03
IGHV3-7*01
IGHV1-8*01
IGHV3-9*01
IGHV5-10-1*03
IGHV3-64D*06
IGHV3-11*06
IGHV3-11*01
IGHV3-13*05
IGHV3-13*01
IGHV3-15*07
IGHV3-15*01
IGHV1-18*01
IGHV3-20*01
IGHV3-20*04
IGHV3-21*01
IGHV3-23*01
IGHV1-24*01
IGHV2-26*01
IGHV4-28*01
IGHV3-30*04
IGHV3-30*18
IGHV3-30-3*01
IGHV4-31*03
IGHV3-33*01
IGHV4-30-2*01
IGHV4-30-4*01
IGHV4-34*01
IGHV4-38-2*01
IGHV4-39*01
IGHV3-43*01
IGHV3-43D*03
IGHV1-45*02
IGHV1-46*01
IGHV1-46*03
IGHV3-48*01
IGHV3-48*04
IGHV3-49*03
IGHV5-51*03
IGHV3-53*02
IGHV3-53*01
IGHV1-58*02
IGHV4-59*01
IGHV4-61*02
IGHV3-64*01
IGHV3-66*02
IGHV1-69*01
IGHV1-69*02
IGHV1-69*06
IGHV2-70D*04
IGHV1-69-2*01
IGHV2-70*01
IGHV3-72*01
IGHV3-73*02
IGHV3-73*01
IGHV3-74*01
V gene

Genomic validation of all expressed sequences



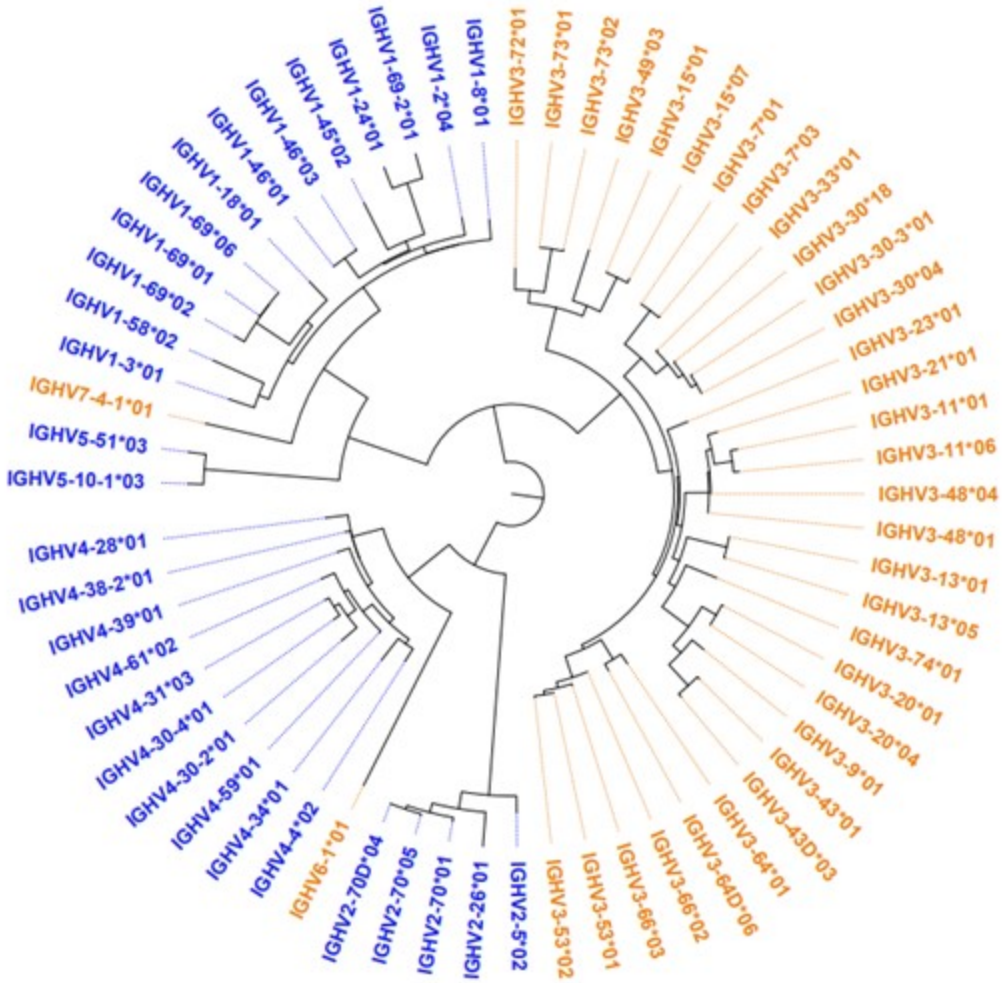
50 IGHV genes, 64 alleles

27 IGHD genes

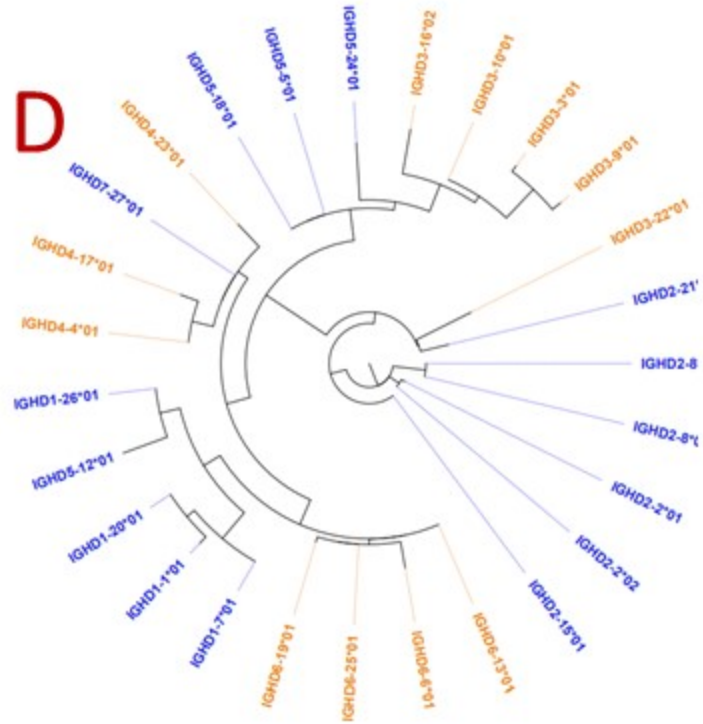
6 IGHJ genes, 7 alleles

- Primers designed to amplify all expressed genes
- Products cloned into plasmids and Sanger sequenced
- All expressed V genes, D genes and J genes targeted

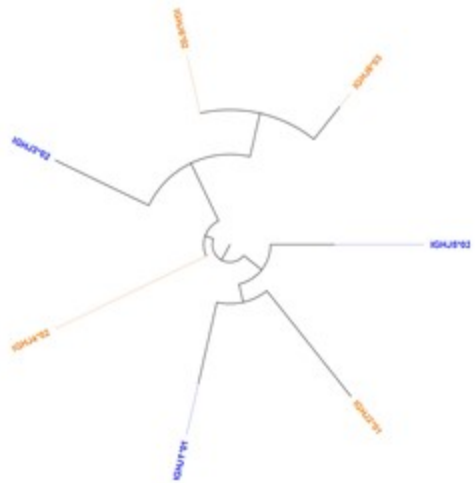
V (50 genes, 64 alleles)



D



J



- Sanger sequenced genotype showed 100% identity to the corecount genotype output

Corecount genotyping summary

- Performed on filtered output of IgDiscover
- Batch analysis is **rapid**
- **Filtering is simplified** (allelic ratio, expected frequency, CDR3 diversity)
- **Highly accurate in genotyping low frequency genes**
- Can **discriminate between alleles with 3' end variation**

Conclusion

- Diversity is extensive in outbred species
- Involves both allele and gene content
- Combination of Rep-Seq and genomic methods help define diversity
- Databases will become more comprehensive and accurate
- High throughput, rapid and consistent methods will enable future association studies

Acknowledgements

Karlsson Hedestam lab
Sanjana Narang
Mateusz Kaduk
Mark Chernyshev
Marco Mandolesi

Christopher Sundling
Anna Färnert

Mateusz Kaduk

Sanjana Narang

Gunilla Karlsson Hedestam

Mark Chernyshev

Marco Mandolesi

Karolinska Institutet

National Institutes of Health

Vetenskapsrådet

CHAVD

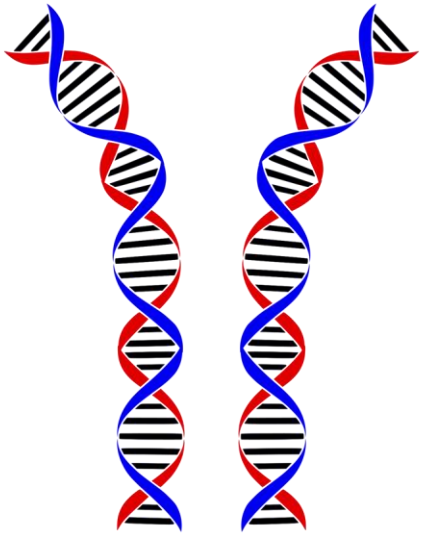
erc

THANK YOU!

Revealing Ig/TR germline variations by AIRR-seq analysis

VDJbase

RAbHIT



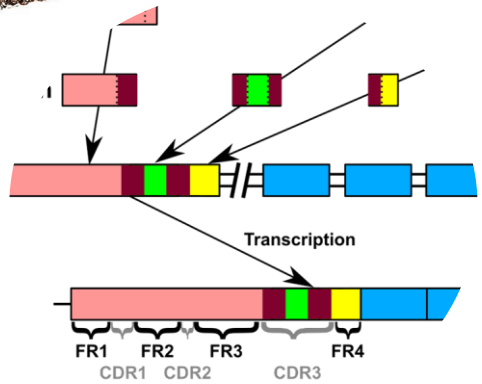
Ayelet Peres
Prof. Gur Yaari
Bar Ilan university

BINA
Bar-Ilan Institute
of Nanotechnology
& Advanced Materials
Bar-Ilan University

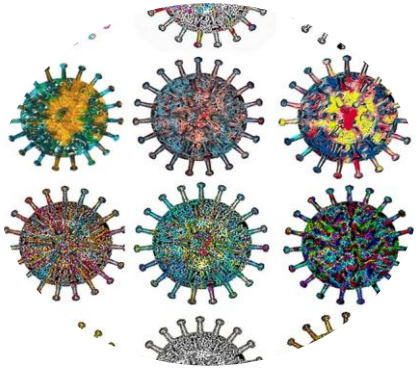
הפקולטה להנדסה
ע"ש אלכסנדר קופקין
אוניברסיטת בר־אילן

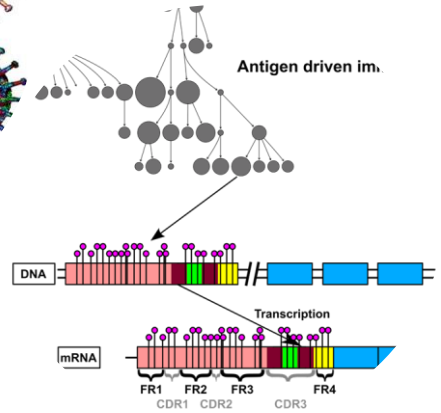
Adaptive immune system

Genetics

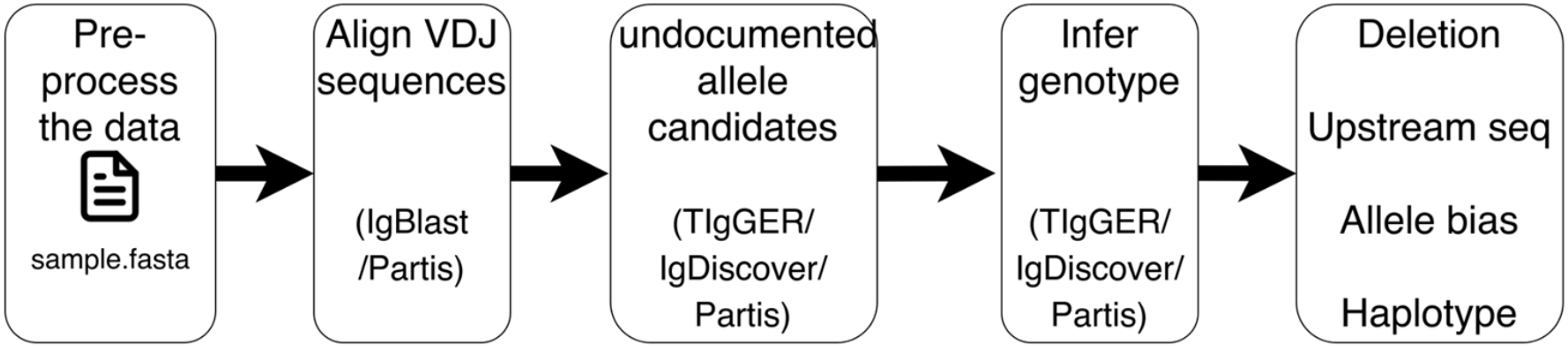


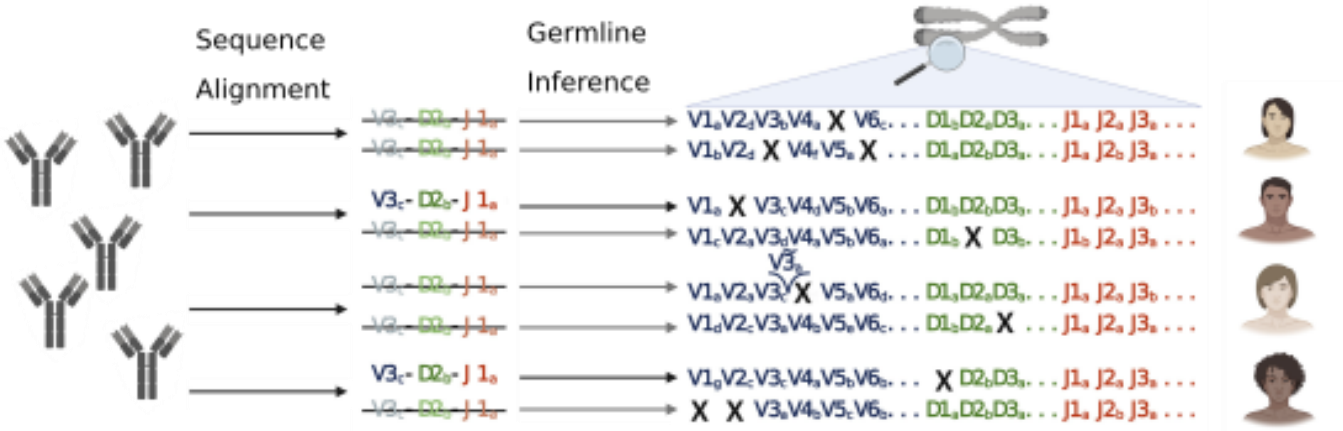
Environment





Repertoire inference steps





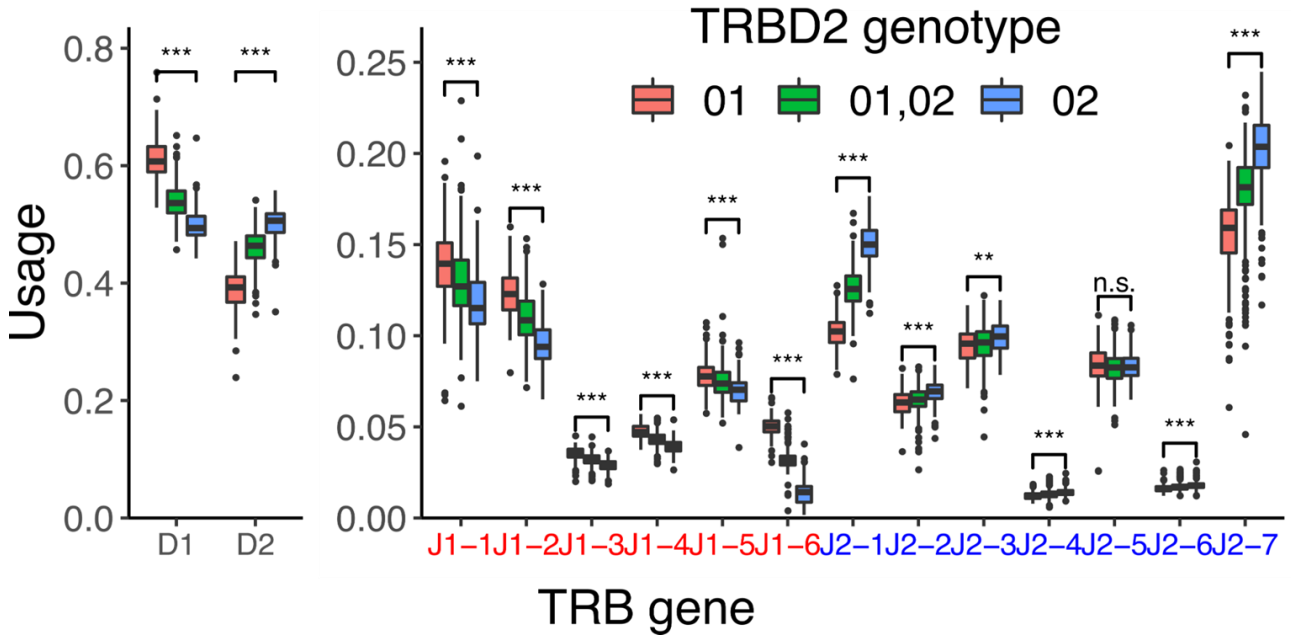
Collins et al. 2020

Undocumented allele inference

Chain	Known Alleles	Undocumented discovered	
IGH	287	25	Mikocziova at el. 2020, NAR
IGK	64	25	Mikocziova and Peres at el. 2021, Iscience
IGL	75	23	
TRB	64	38	Omer and Peres at el. 2021, genome medicine

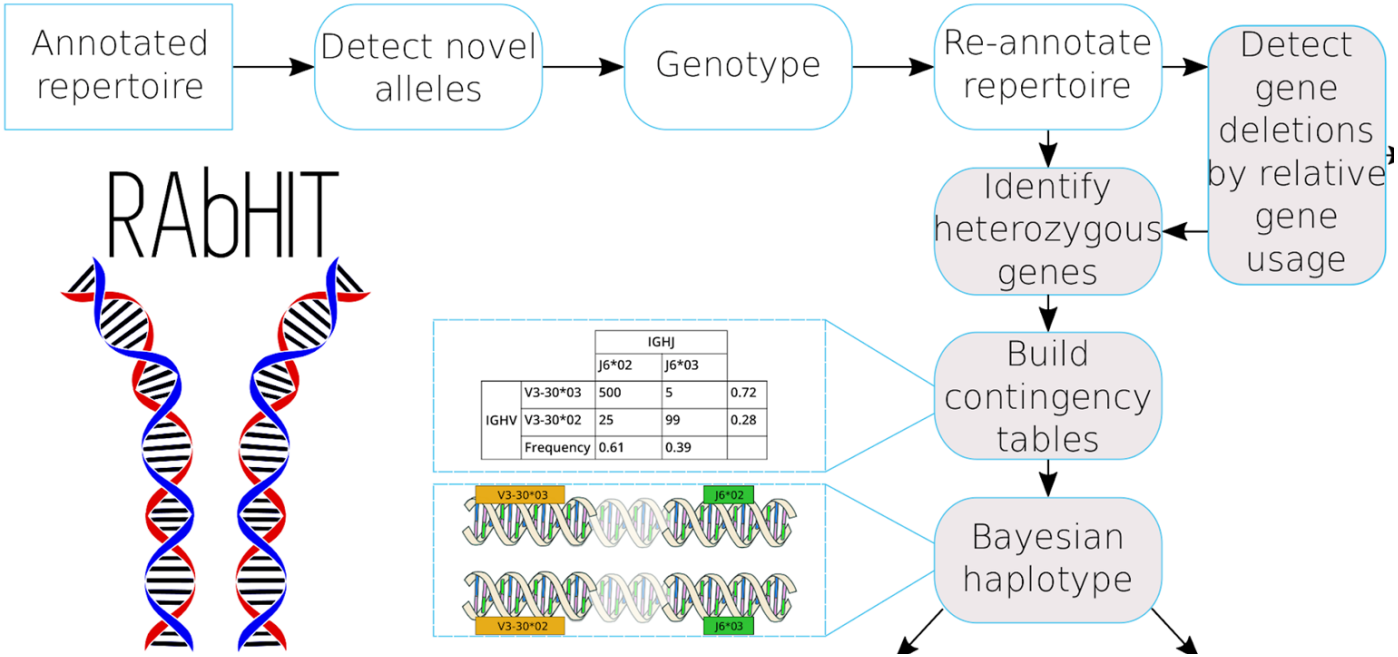
Genotype influence [Omer* and Peres* et al., 2021 *Genome Medicine*]

```
TRBD2*01    GGGACTAGCGGGGGGG
TRBD2*02    GGGACTAGCGGGAGGG
            **********.***
```

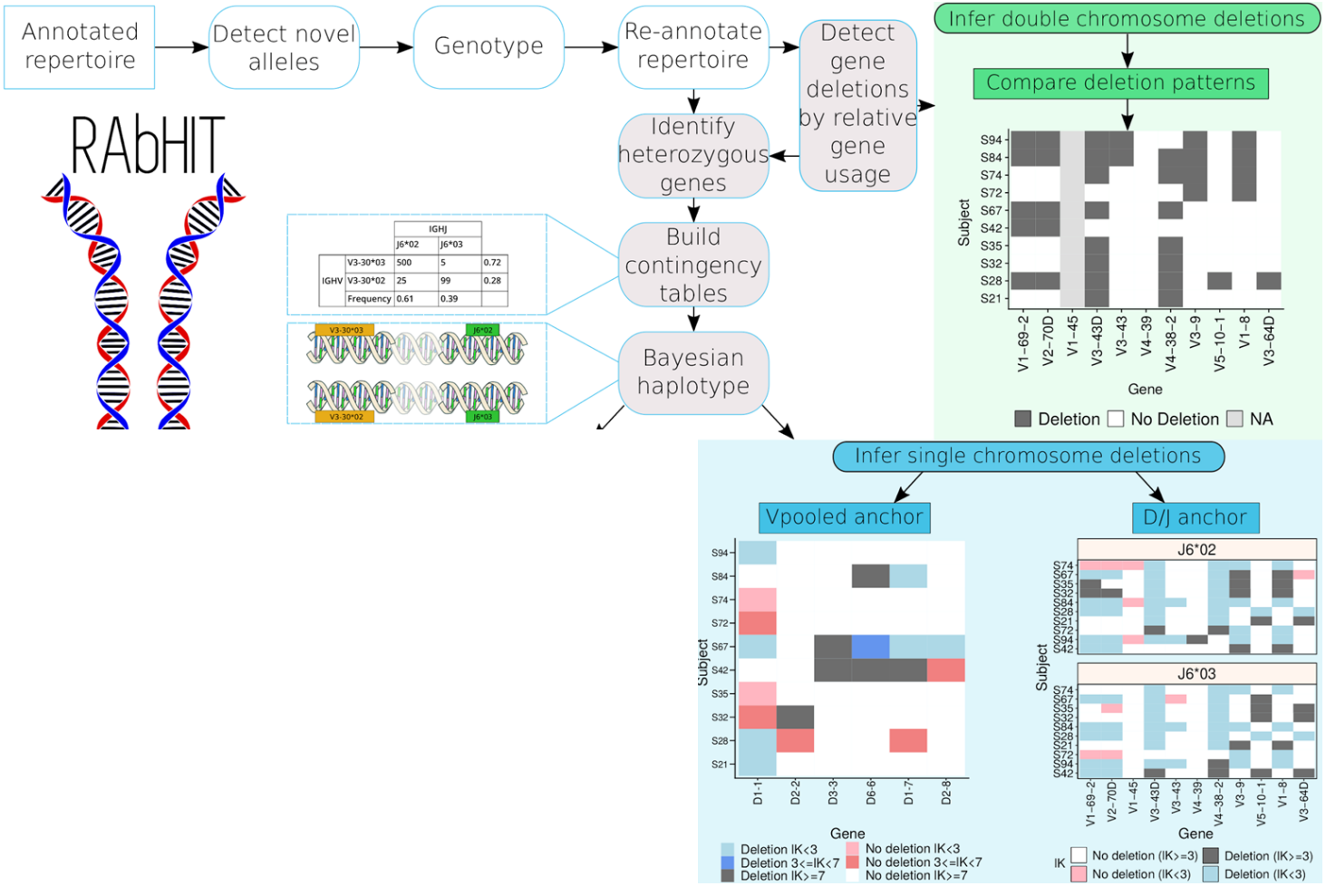


Strong bias between TRBD2 genotype and TRBJ1-6 usage

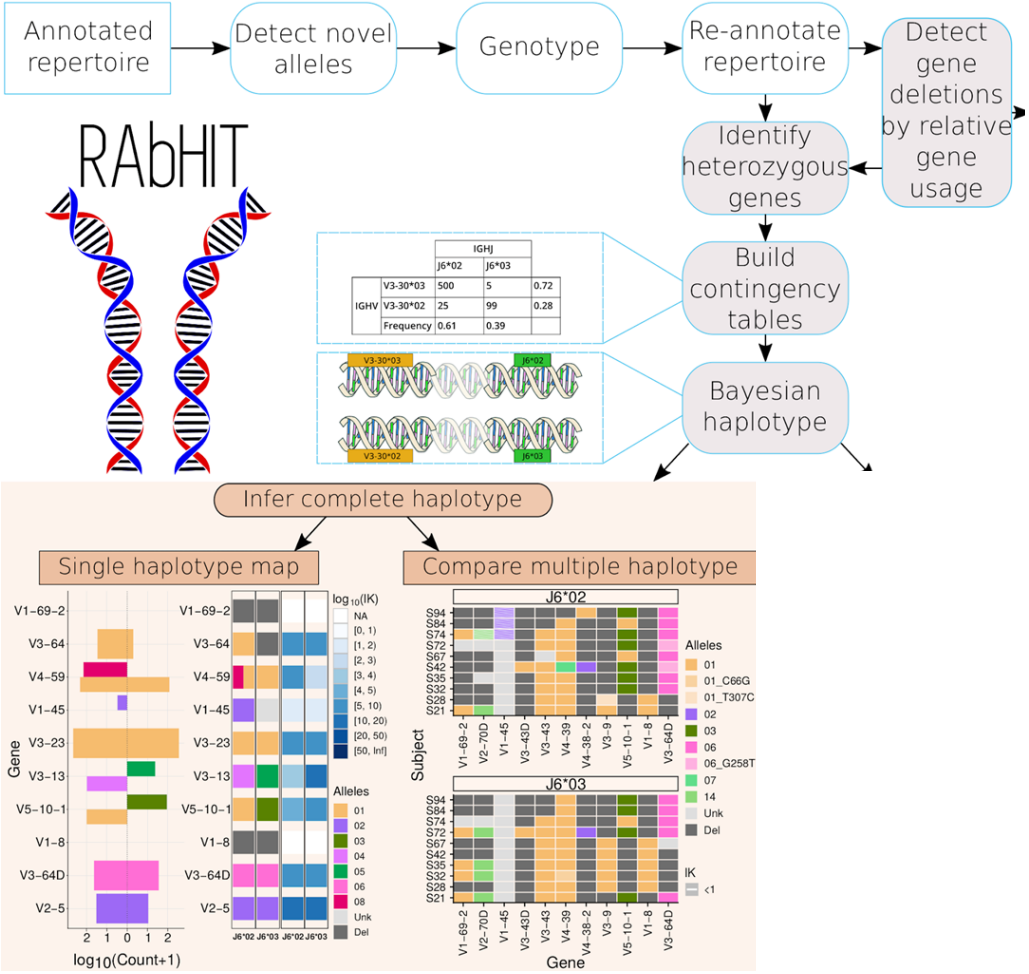
RAbHIT: R Antibody Haplotype Inference Tool [Peres et al., 2019 *Bioinformatics*]



RAbHIT: R Antibody Haplotype Inference Tool [Peres et al., 2019 *Bioinformatics*]

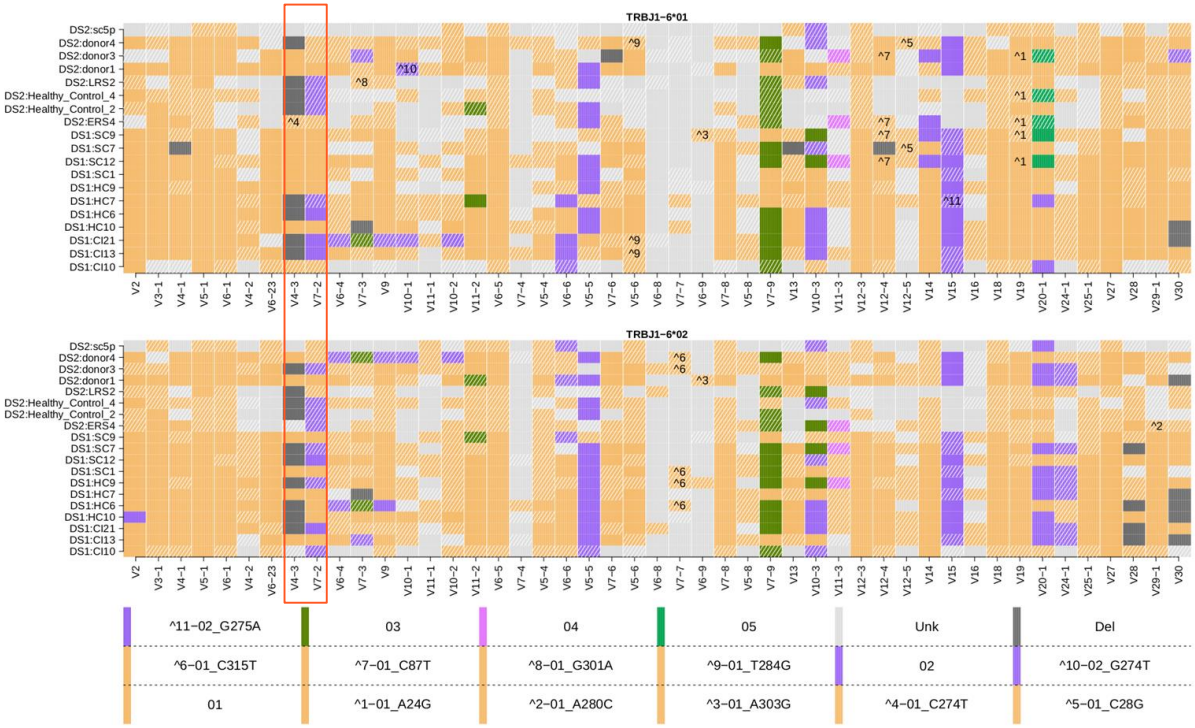


RAbHIT: R Antibody Haplotype Inference Tool [Peres et al., 2019 *Bioinformatics*]



Haplotype inference

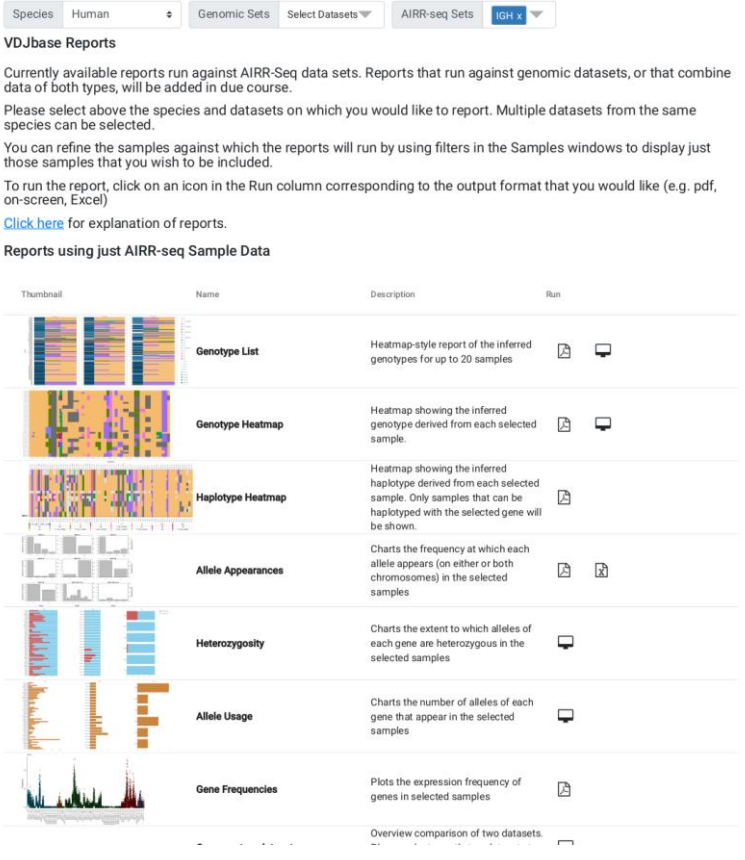
- J genes as anchors (TRBJ1-6)
- Allele patterns and deletion associations
- Linkage between TRBV7-2*02 and TRBV4-3 deletion



VDJbase.org [Omer*, Shemesh*, Peres* et al., 2020 *NAR*]

- Publicly available Ig and TRB database search engine
- Describing complete sets of genotype and haplotype data
- Gene usage data

- Available for local installation
- Allows you to privately add your own datasets
- VDJbase processing pipeline is available in dockerhub [peresay/suite]



Future plans

- Adapting TigGER, RAbHIT and VDJbase to other species.
- Adding more projects to VDJbase.
- Adding and improving comparative visualizations.
- Creating interface between different servers and VDJbase (e.g OGRDB)

Acknowledgements

Yaari Lab

- ★ Gur Yaari
- ★ Pazit Polak
- ★ Aviv Omer
- ★ Moriah Gidoni

Oslo Group

- ★ Ludvig Sollid
- ★ Victor Greiff
- ★ Ivana Mikocziova
- ★ Omri snir
- ★ Ida Lindeman

Watson Lab

- ★ Corey T Watson
- ★ Oscar L Rodriguez

IARC Group

- ★ William Lees
- ★ Andrew M Collins
- ★ Martin Corcoran
- ★ Mats Ohlin

Detecting immunoglobulin heavy chain locus genetic variation by targeted long-read sequencing

Oscar L. Rodriguez, PhD
Watson Lab
Postdoctoral fellow | Department of Biochemistry and Molecular Genetics
University of Louisville School of Medicine

AIRR Community webinars
February 10, 2022

UofL SCHOOL OF MEDICINE

Outline of presentation

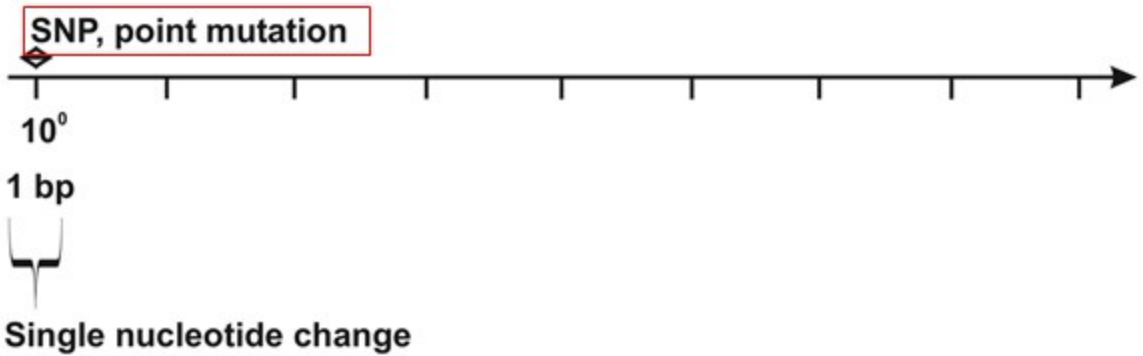
1) Introduction
    a) Genetic variation in the immunoglobulin heavy chain locus (IGH)
    b) Advantages of long-read sequencing over short-read sequencing
2) Framework for resolving the immunoglobulin heavy locus in a high throughput fashion using long read sequencing
3) Resolving the immunoglobulin heavy chain locus in a cohort with adaptive immune repertoire sequencing
4) Resolving the T-cell receptor loci using long read sequencing
5) Application of framework to the immunoglobulin loci in rhesus macaque
6) Conclusion

Genetic variation defined by genetic differences between individuals

Person 1: ATGCTAGCTAGTCAG
Person 2: ATGCTACCTAGTCAG

Genetic variation ranges in size from base differences to large structural variants

Person 1: ATGCTAGCTAGTCAG
Person 2: ATGCTACCTAGTCAG

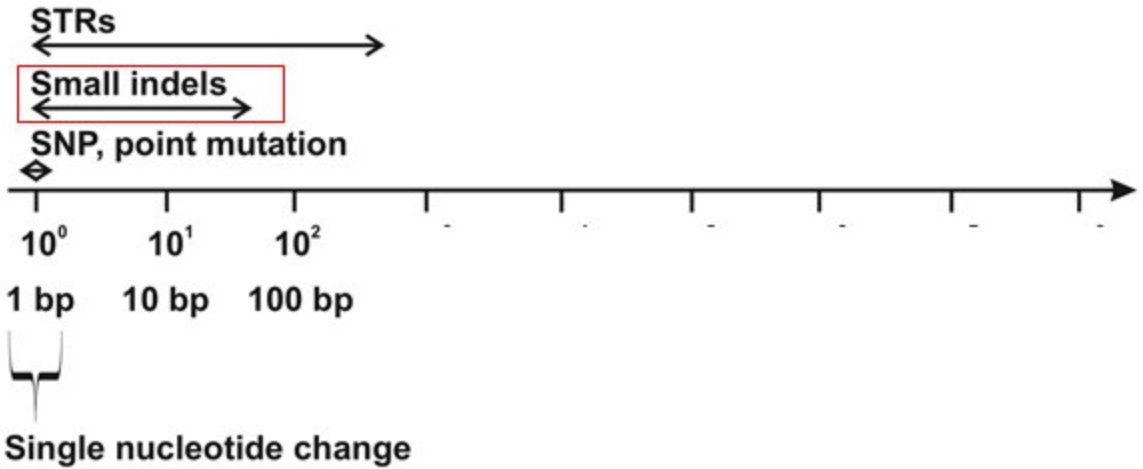


Adopted from Pollex et al

Genetic variation ranges in size from base differences to large structural variants

Person 1: ATGCTAGCTAGTCAG
Person 2: ATGCTACCTAGTCAG

Person 1: ATGCTAGCTAGTCAG
Person 2: ATGCTA———TCAG



Adopted from Pollex et al

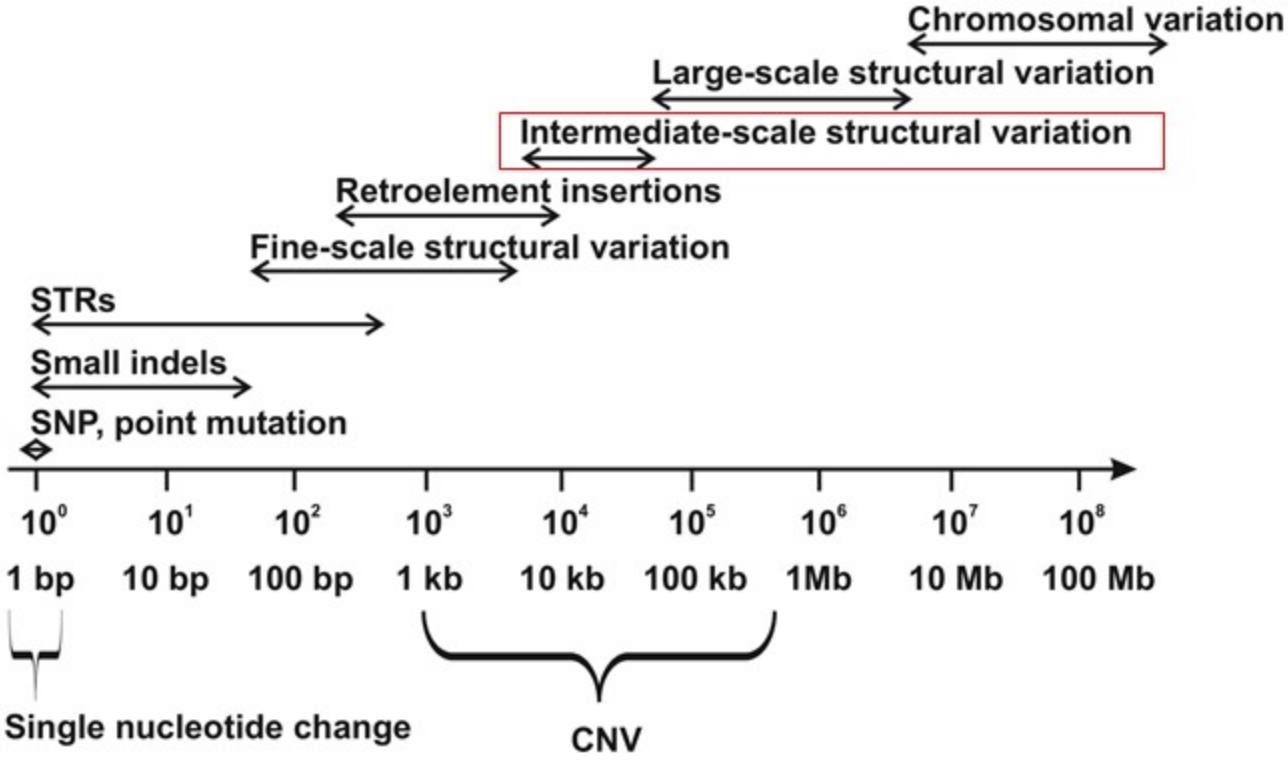
Genetic variation ranges in size from base differences to large structural variants

Person 1: ATGCTAGCTAGTCAG
Person 2: ATGCTACCTAGTCAG

Person 1: ATGCTAGCTAGTCAG
Person 2: ATGCTA——TCAG

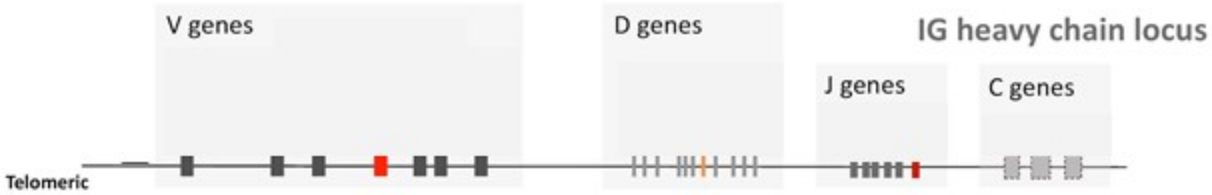
Person 1: ATGCTAGCTAGTCAG
Person 2: ATGCTACCTAGTCAG

ATGCTACCTAGTCAGATGCTAC
CTAGTCAGATGCTACCTAGTCA
GATGCTACCTAGTCAGATGCTA
CCTAGTCAGATGCTACCTAGTC
AGAGTCAGATGCTACCTAGTCA
GAGTCAGATGCTACCTAGTCAG
AGTCAGATGCTACCTAGTCAGA
CAGTCAGATGCTACCTAGTCAG
AGTCAGAT

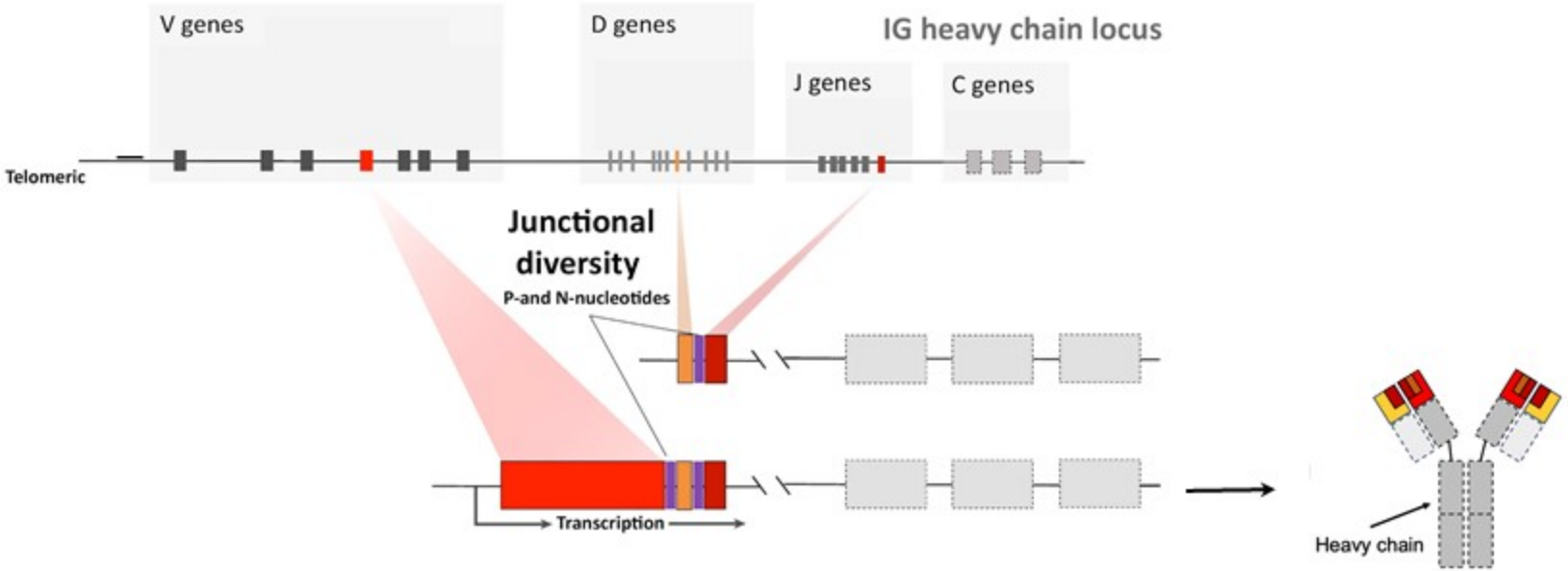


Adopted from Pollex et al

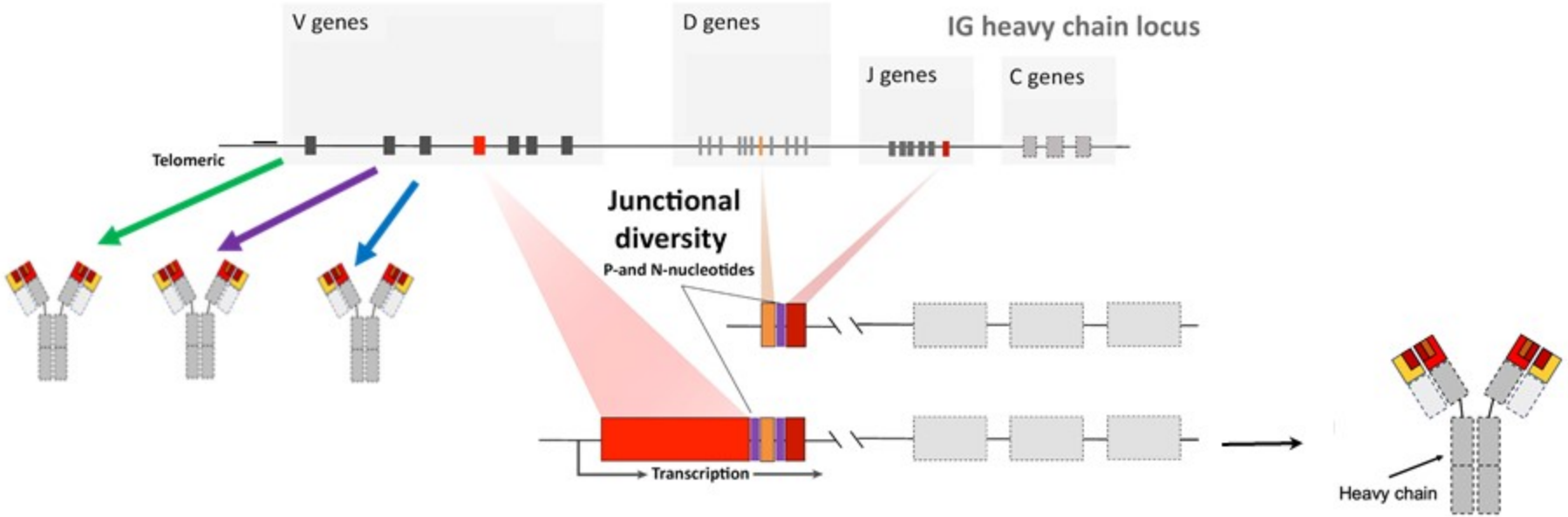
IGH locus is located at the telomeric end of chromosome 14



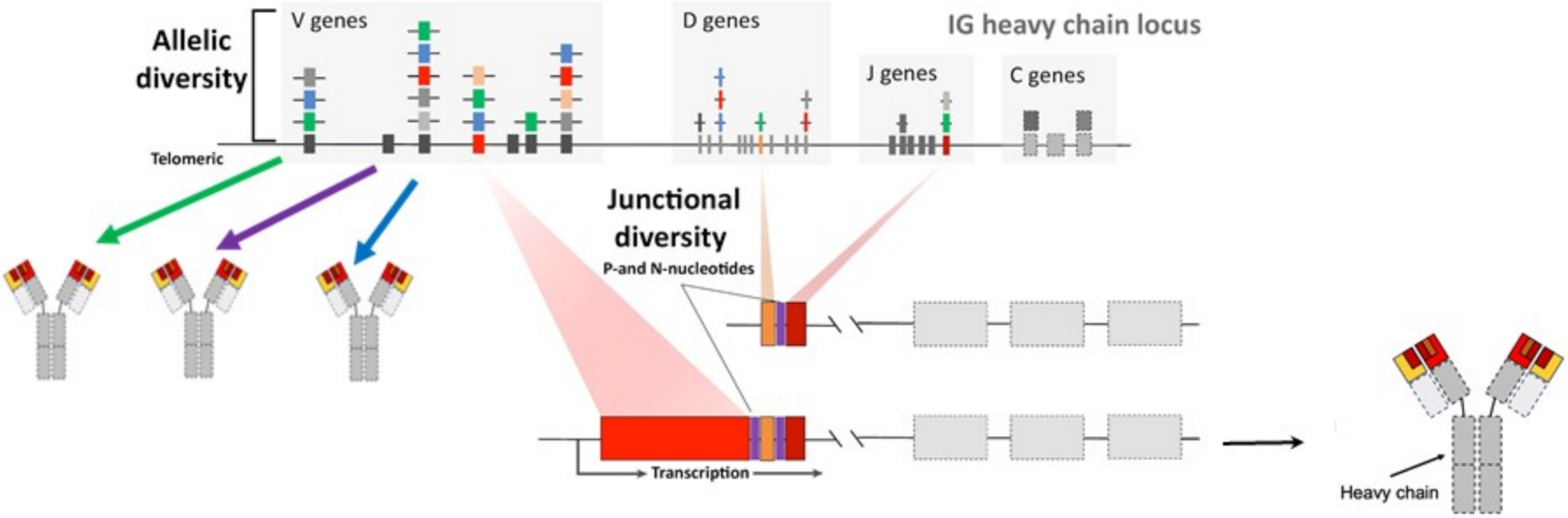
Rearranged DNA gets transcribed and translated to produce the heavy chain of antibodies



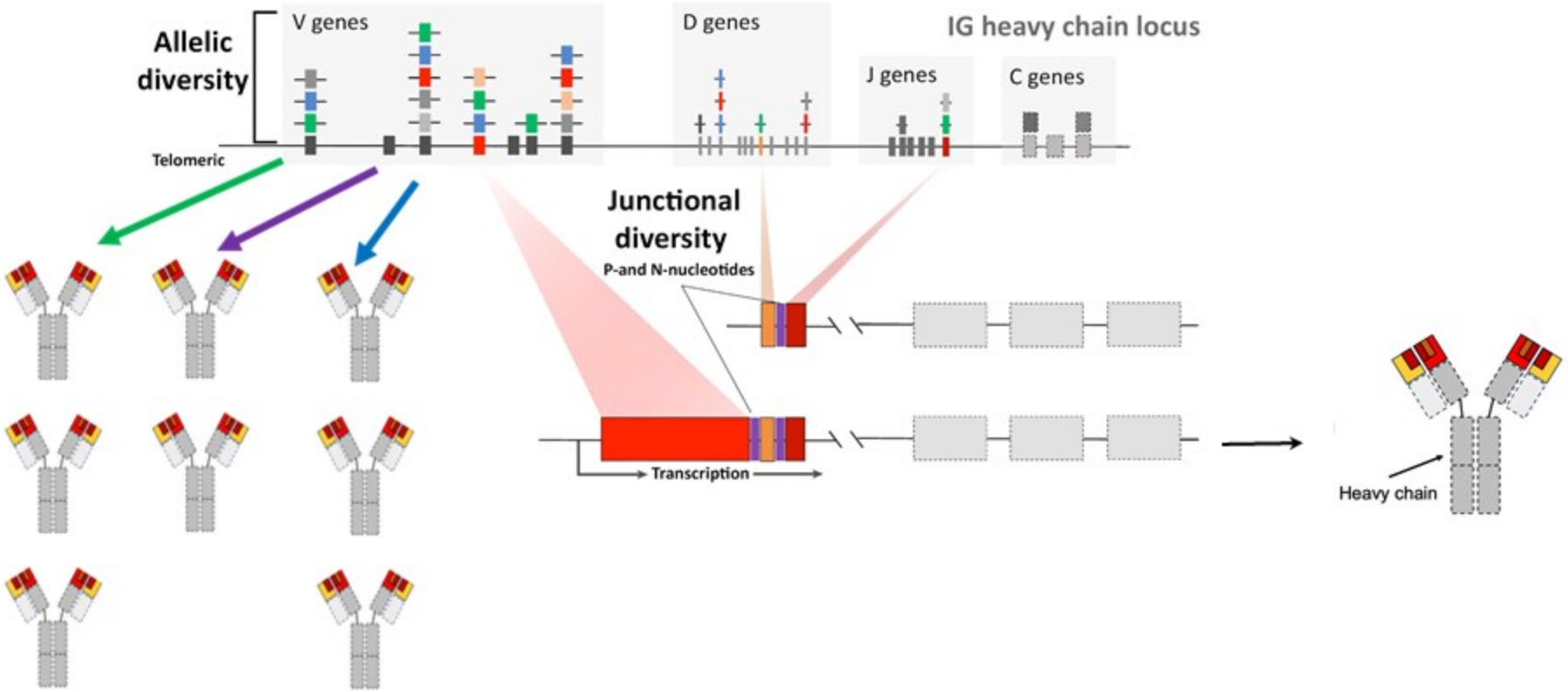
By selecting different genes, B cells can create a diverse antibody repertoire



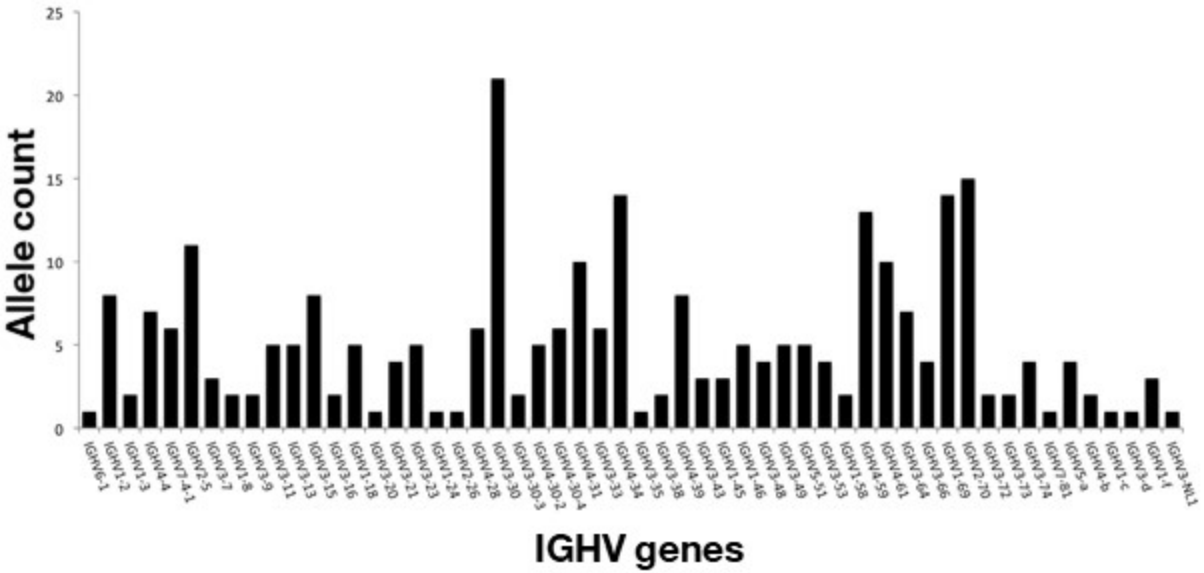
At the population level, there also exist haplotype diversity



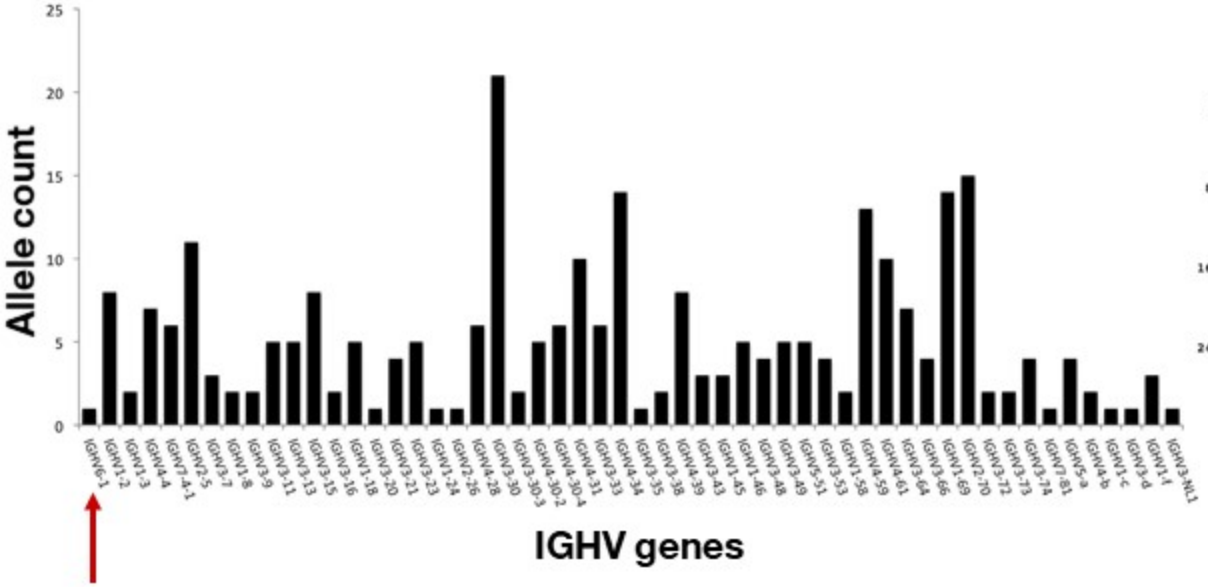
At the population level, there also exist haplotype diversity

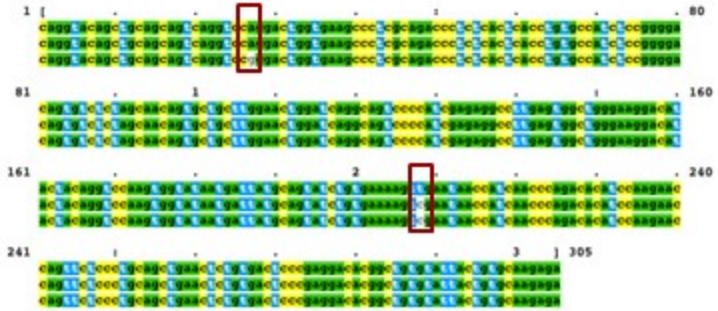


Current data shows extensive allelic diversity



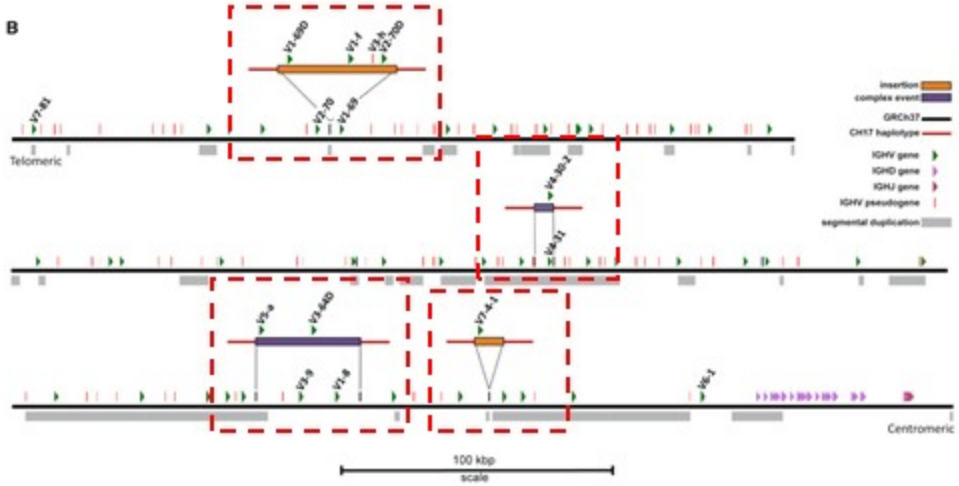
Current data shows extensive allelic diversity





IGHV6-1*01
IGHV6-1*02
IGHV6-1*03

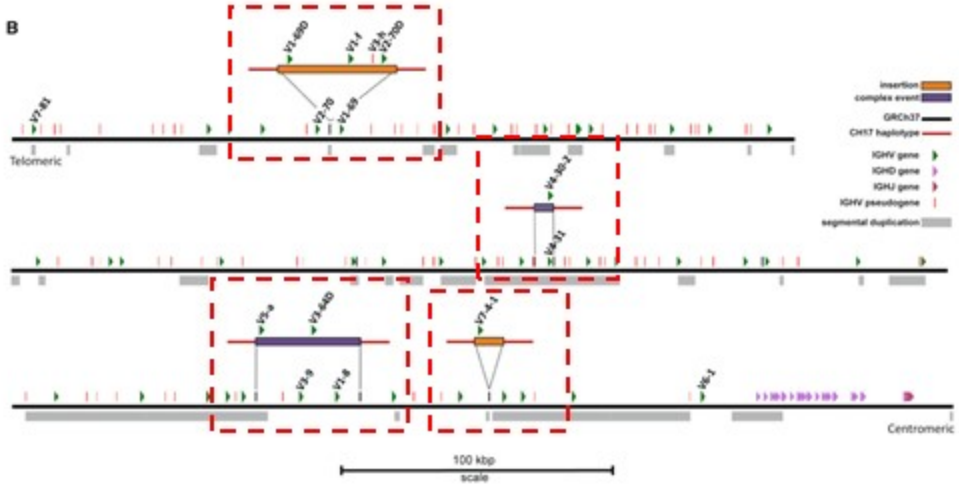
Large number of structural variants in IGH



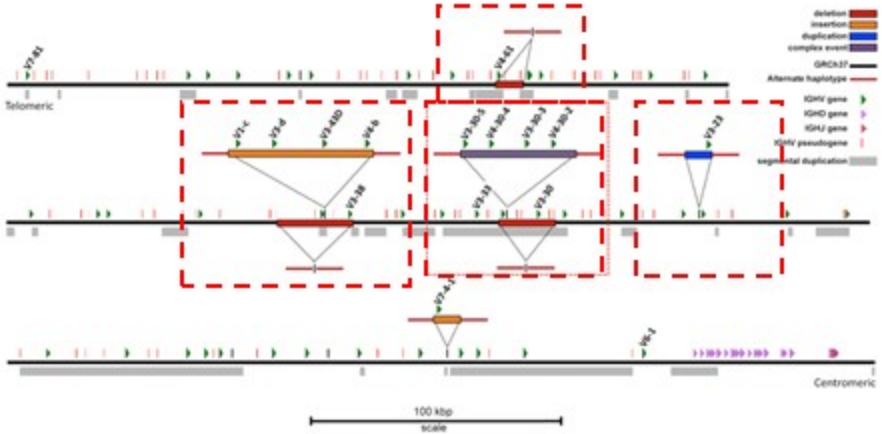
Structural variants identified in a single individual

Watson et al. AJHG 2013

Large number of structural variants in IGH



Structural variants identified in a single individual



Structural variants identified in a many individuals across different populations

Watson et al. AJHG 2013

Large number of structural variants in IGH

Table 1. CNVs Identified from BAC and Fosmid Clones

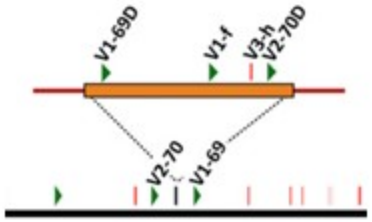
Individual	Population	CNV Type	IGHV Genes Included in CNV[a]	GRCh37 Outer-Start (Breakpoint)[b]	GRCh37 Outer-End (Breakpoint)[b]	Event Size (~kbp)
CH17	nd	Insertion	V1-69D, V1-f, V3-h, V2-70D (gain)	107174927	107174941	46.6
CH17	nd	Complex event	V4-30-2 (gain) V4-31 (loss)	106804332	106810878	6.5[c]/48.8[d]
CH17	nd	Complex event	V5-a, V3-64D (gain) V3-9, V1-8 (loss)	106531320	106569343	38[c]/37.7[e]
CH17	nd	Insertion	V7-4-1 (gain)	106483362	106484225	9.5
NA12156	CEPH	Deletion	V4-39, V3-38 (loss)	106866357	106899042	32.7
NA15510 and NA19240	nd and Yoruba	Insertion	V1-c, V3-d, V3-43D, V4-b (gain)	106877146	106877535	61.1
NA18555 (haplotype A)	Han Chinese	Complex event	V3-30-5, V4-30-4, V3-30-3, V4-30-2 (gain)	106804332	106804333	49.2
NA18555 (haplotype B)	Han Chinese	Deletion	V4-31, V3-30 (loss)	106786254	106811213	24.9[c]/73.9[d]
NA18507	Yoruban	Complex event[g]	V4-30-4, V3-30-3 (gain) V3-30 (loss)	106784242	nd	25.2
NA18502	Yoruban	Complex event[g]	V3-30-5 (gain) V3-33, V4-31 (loss)	nd	106820685	24.7
NA18956 and NA12156	Japanese and CEPH	Duplication	V3-23D (gain)	106716650	106727861	10.8
NA19240 and NA12878	Yoruban and CEPH	Insertion	V7-4-1 (gain)	106483362	106484225	9.5

Watson et al. AJHG 2013

Large number of structural variants in IGH

Table 1. CNVs Identified from BAC and Fosmid Clones

Individual	Population	CNV Type	IGHV Genes Included in CNV[a]	GRCh37 Outer-Start (Breakpoint)[b]	GRCh37 Outer-End (Breakpoint)[b]	Event Size (~kbp)
CH17	nd	Insertion	V1-69D, V1-f, V3-h, V2-70D (gain)	107174927	107174941	46.6
CH17	nd	Complex event	V4-30-2 (gain) V4-31 (loss)	106804332	106810878	6.5[c]/48.8[d]
CH17	nd	Complex event	V5-a, V3-64D (gain) V3-9, V1-8 (loss)	106531320	106569343	38[c]/37.7[e]
CH17	nd	Insertion	V7-4-1 (gain)	106483362	106484225	9.5
NA12156	CEPH	Deletion	V4-39, V3-38 (loss)	106866357	106899042	32.7
NA15510 and NA19240	nd and Yoruba	Insertion	V1-c, V3-d, V3-43D, V4-b (gain)	106877146	106877535	61.1
NA18555 (haplotype A)	Han Chinese	Complex event	V3-30-5, V4-30-4, V3-30-3, V4-30-2 (gain)	106804332	106804333	49.2
NA18555 (haplotype B)	Han Chinese	Deletion	V4-31, V3-30 (loss)	106786254	106811213	24.9[c]/73.9[d]
NA18507	Yoruban	Complex event[g]	V4-30-4, V3-30-3 (gain) V3-30 (loss)	106784242	nd	25.2
NA18502	Yoruban	Complex event[g]	V3-30-5 (gain) V3-33, V4-31 (loss)	nd	106820685	24.7
NA18956 and NA12156	Japanese and CEPH	Duplication	V3-23D (gain)	106716650	106727861	10.8
NA19240 and NA12878	Yoruban and CEPH	Insertion	V7-4-1 (gain)	106483362	106484225	9.5

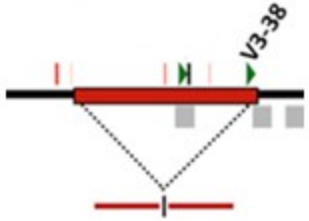


Watson et al. AJHG 2013

Large number of structural variants in IGH

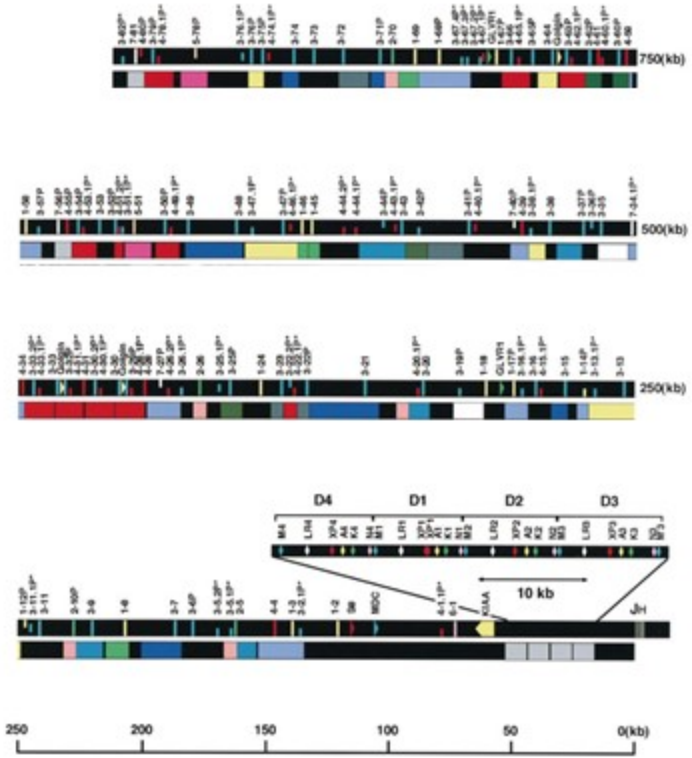
Table 1. CNVs Identified from BAC and Fosmid Clones

Individual	Population	CNV Type	IGHV Genes Included in CNV[a]	GRCh37 Outer-Start (Breakpoint)[b]	GRCh37 Outer-End (Breakpoint)[b]	Event Size (~kbp)
CH17	nd	Insertion	V1-69D, V1-f, V3-h, V2-70D (gain)	107174927	107174941	46.6
CH17	nd	Complex event	V4-30-2 (gain) V4-31 (loss)	106804332	106810878	6.5[c]/48.8[d]
CH17	nd	Complex event	V5-a, V3-64D (gain) V3-9, V1-8 (loss)	106531320	106569343	38[c]/37.7[e]
CH17	nd	Insertion	V7-4-1 (gain)	106483362	106484225	9.5
NA12156	CEPH	Deletion	V4-39, V3-38 (loss)	106866357	106899042	32.7
NA15510 and NA19240	nd and Yoruba	Insertion	V1-c, V3-d, V3-43D, V4-b (gain)	106877146	106877535	61.1
NA18555 (haplotype A)	Han Chinese	Complex event	V3-30-5, V4-30-4, V3-30-3, V4-30-2 (gain)	106804332	106804333	49.2
NA18555 (haplotype B)	Han Chinese	Deletion	V4-31, V3-30 (loss)	106786254	106811213	24.9[c]/73.9[d]
NA18507	Yoruban	Complex event[g]	V4-30-4, V3-30-3 (gain) V3-30 (loss)	106784242	nd	25.2
NA18502	Yoruban	Complex event[g]	V3-30-5 (gain) V3-33, V4-31 (loss)	nd	106820685	24.7
NA18956 and NA12156	Japanese and CEPH	Duplication	V3-23D (gain)	106716650	106727861	10.8
NA19240 and NA12878	Yoruban and CEPH	Insertion	V7-4-1 (gain)	106483362	106484225	9.5



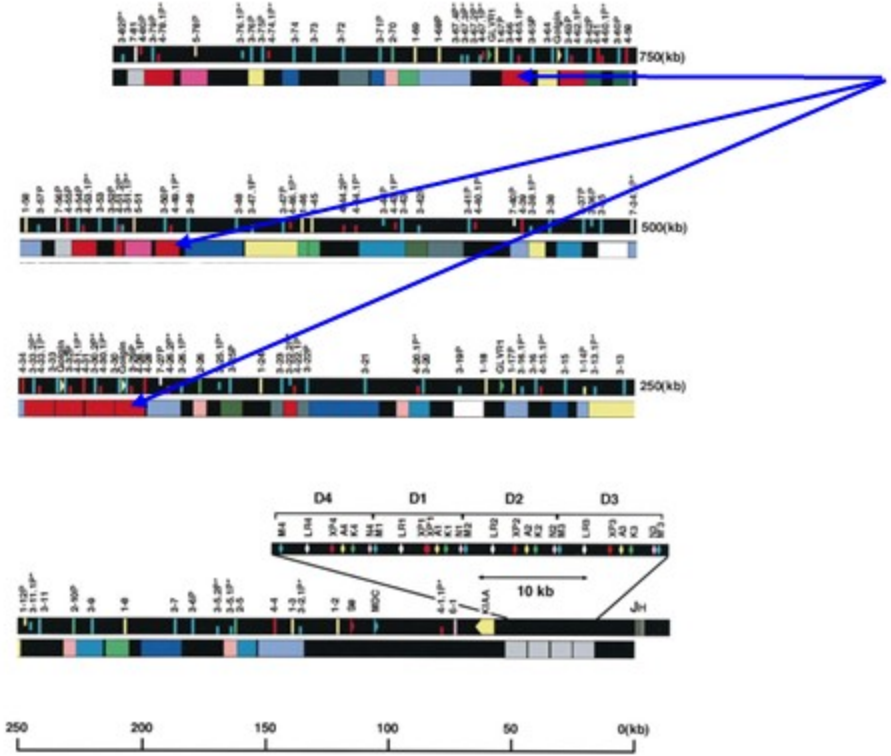
Watson et al. AJHG 2013

Complexity of locus has hindered previous sequencing methods to genotype IGH



Matsuda et al. J. Exp. Med 1998

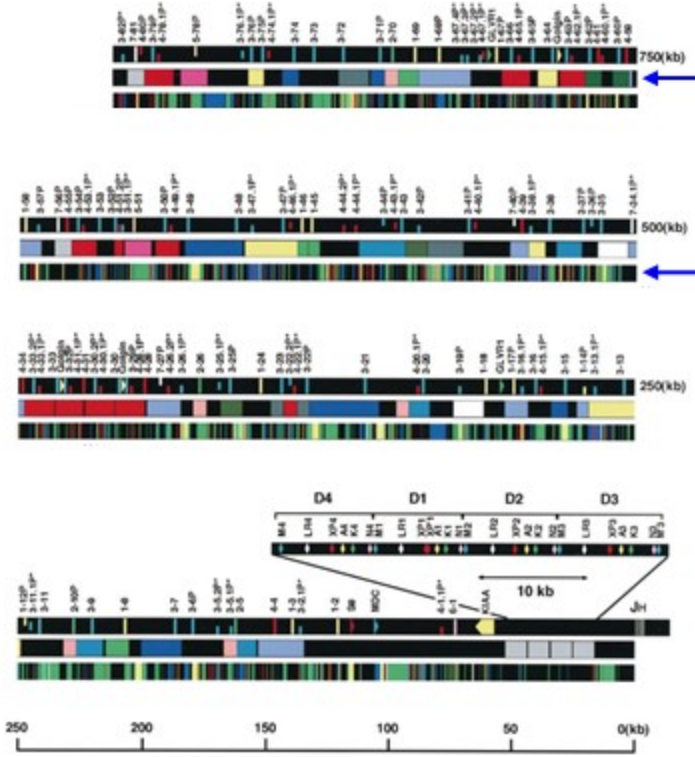
Complexity of locus has hindered previous sequencing methods to genotype IGH



Long highly identical sequences

Matsuda et al. J. Exp. Med 1998

Complexity of locus has hindered previous sequencing methods to genotype IGH



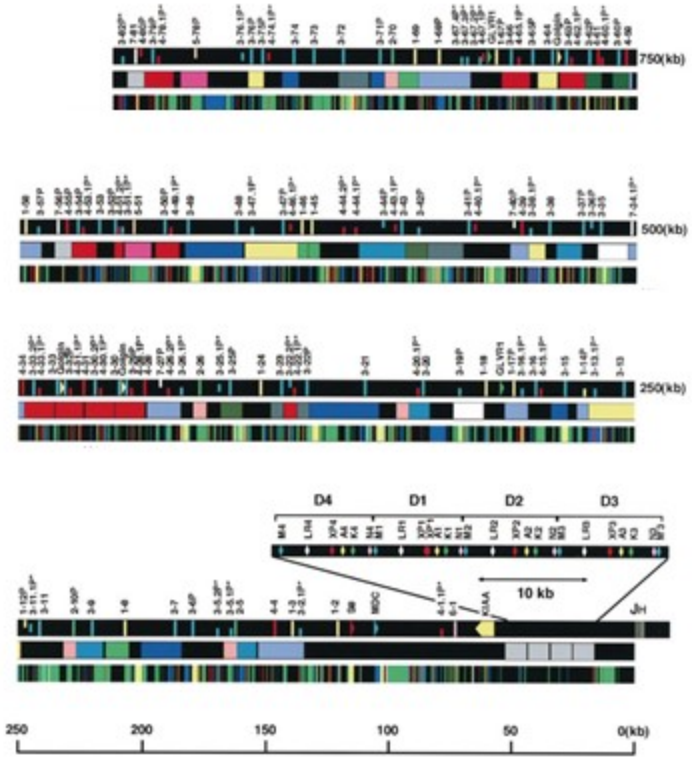
Long highly identical sequences

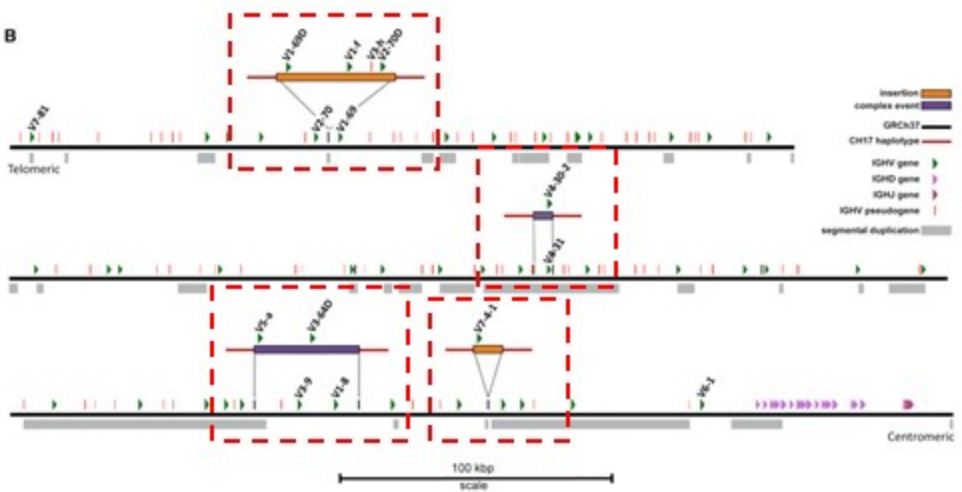
Different repeat classes

1. Alu
2. MIR
3. LINE
4. Simple repeats

Matsuda et al. J. Exp. Med 1998

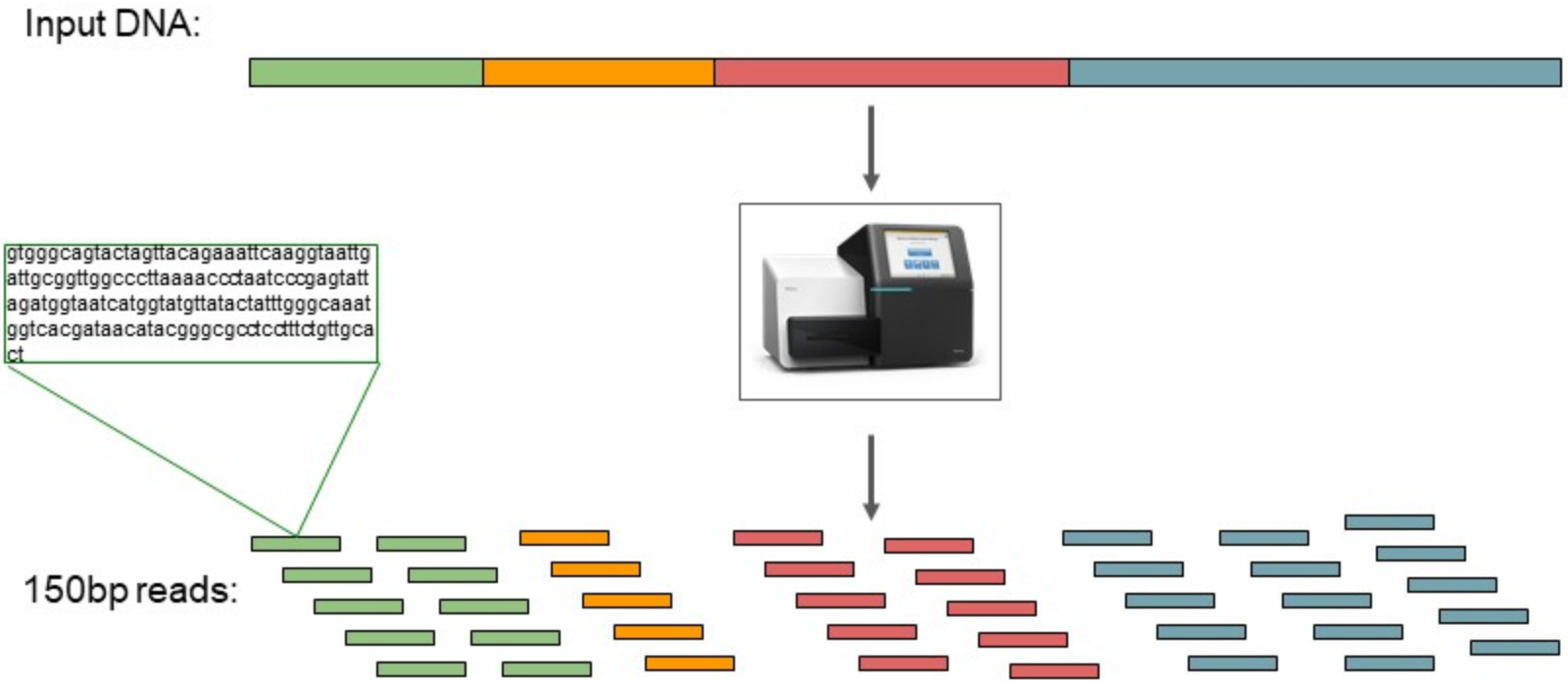
Complexity of locus has hindered previous sequencing methods to genotype IGH



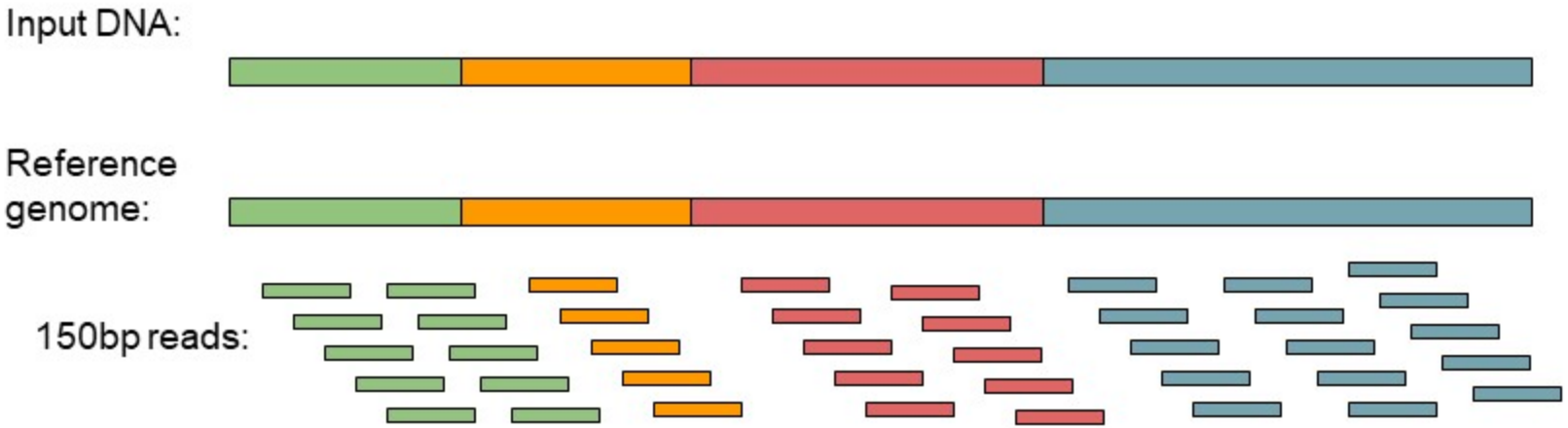


Matsuda et al. J. Exp. Med 1998

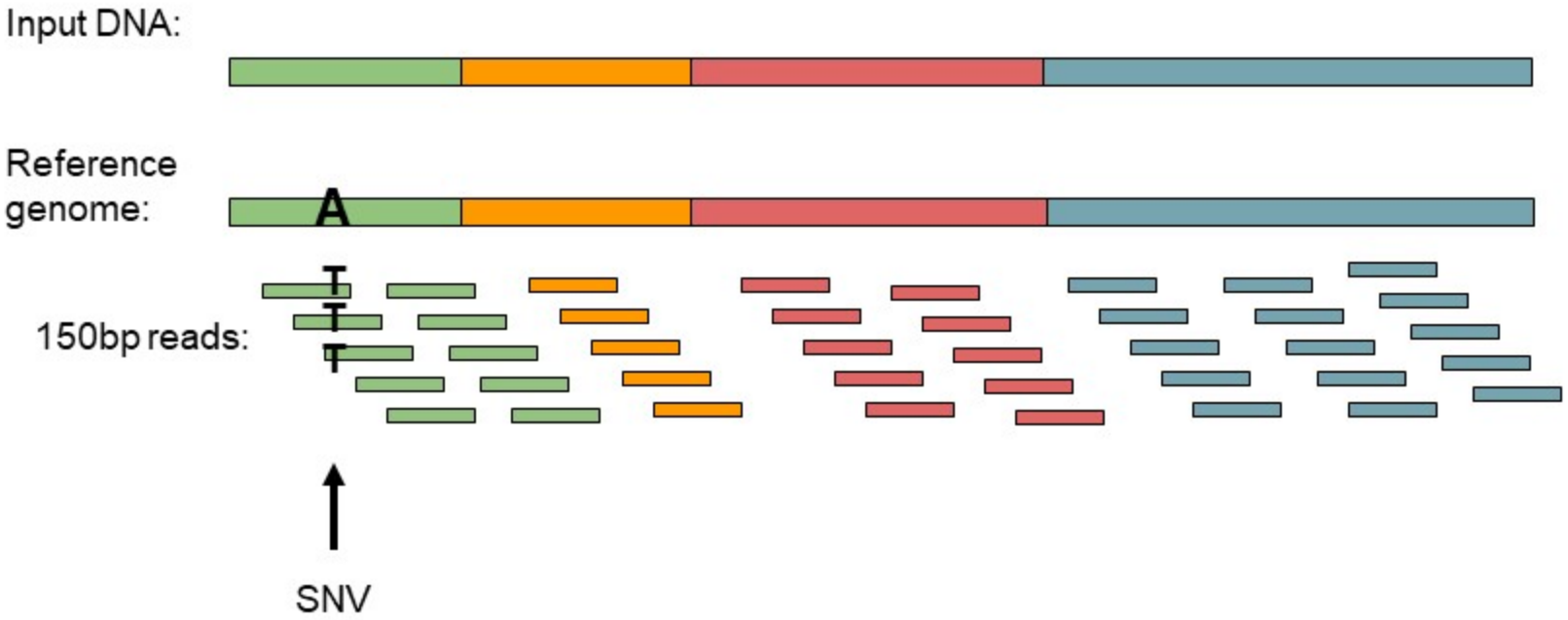
Next-generation sequencing technology generates reads from 50 to 250 bases



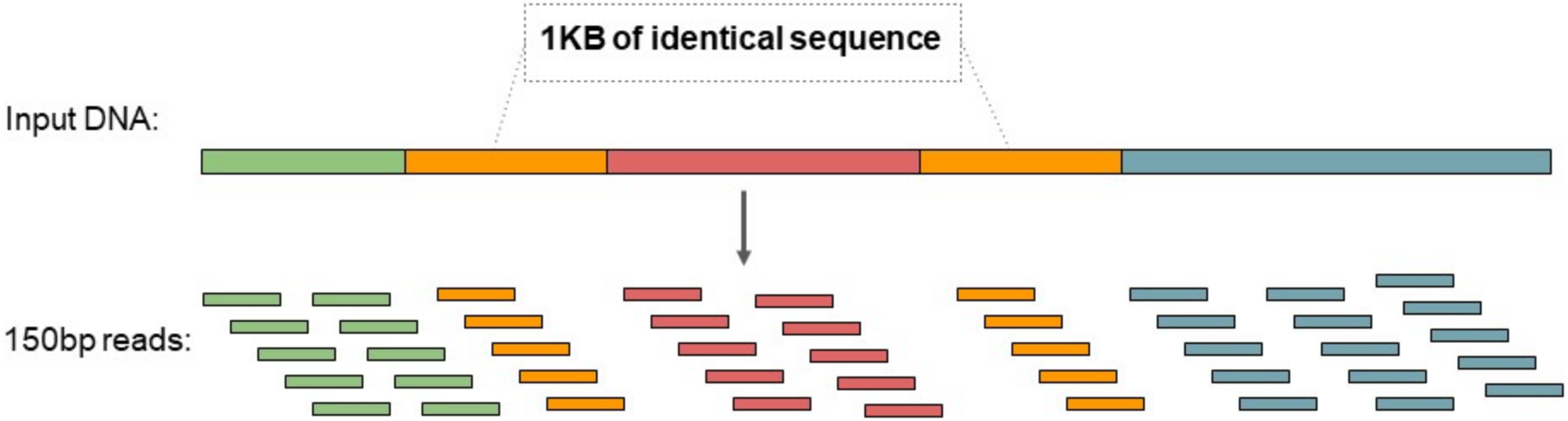
Reads are aligned to a reference genome



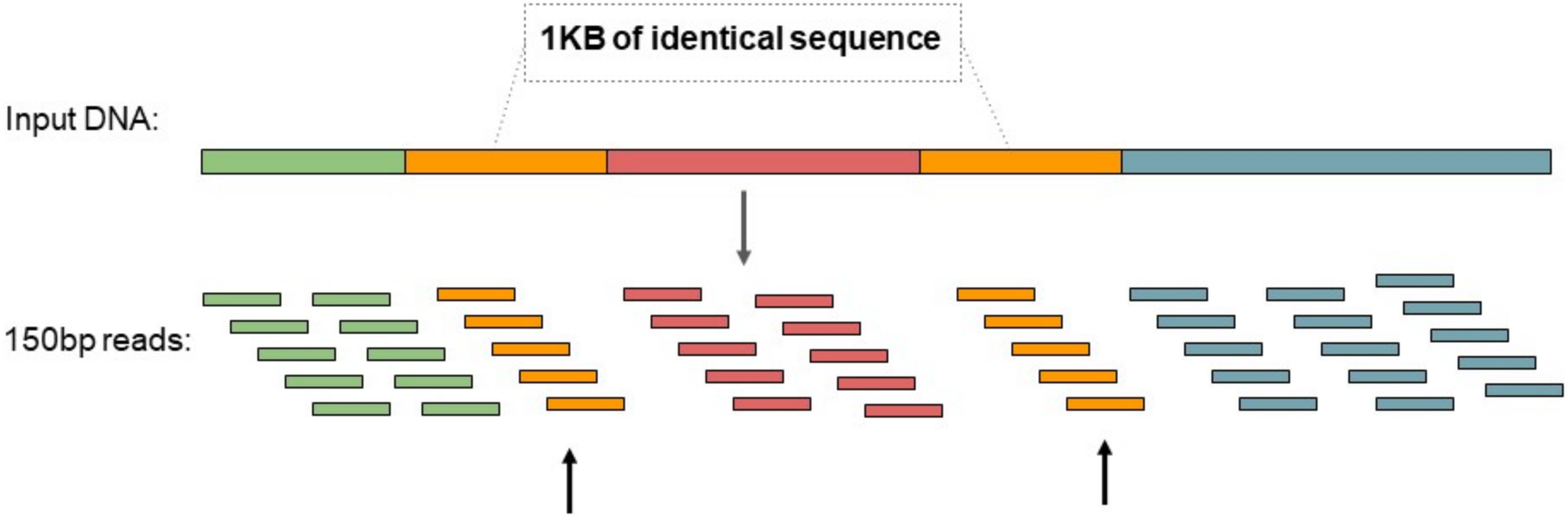
Difference between reads and reference genome correspond to genetic variants



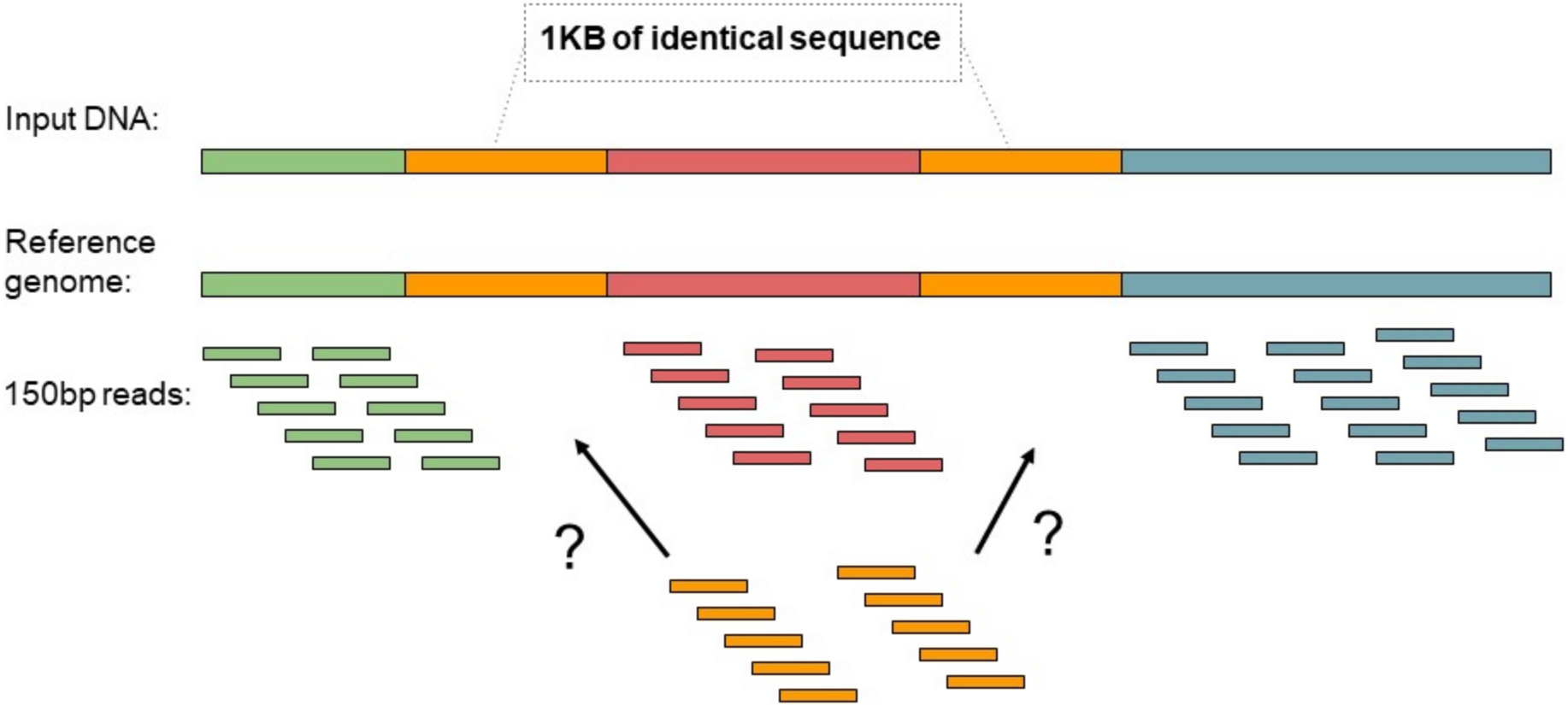
Limitation of next-generation sequencing technology



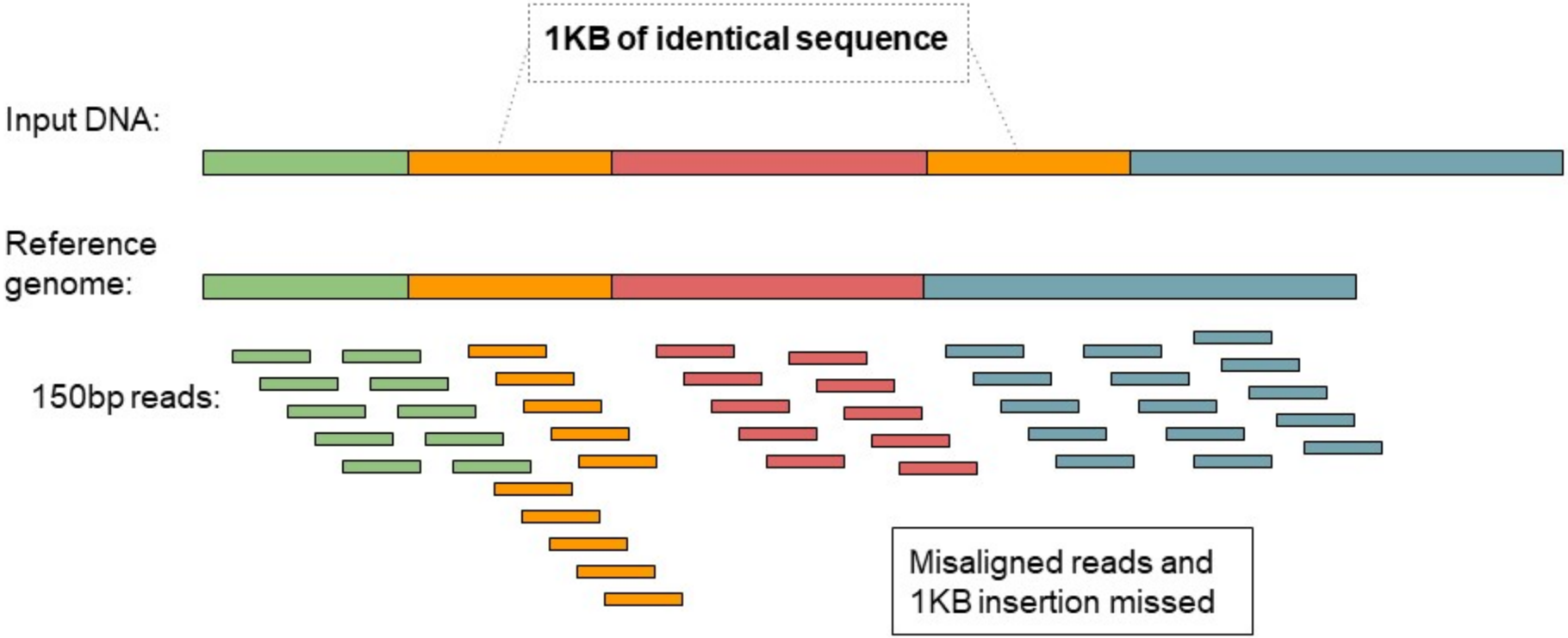
Input DNA with repetitive or identical sequence generates reads with identical sequence



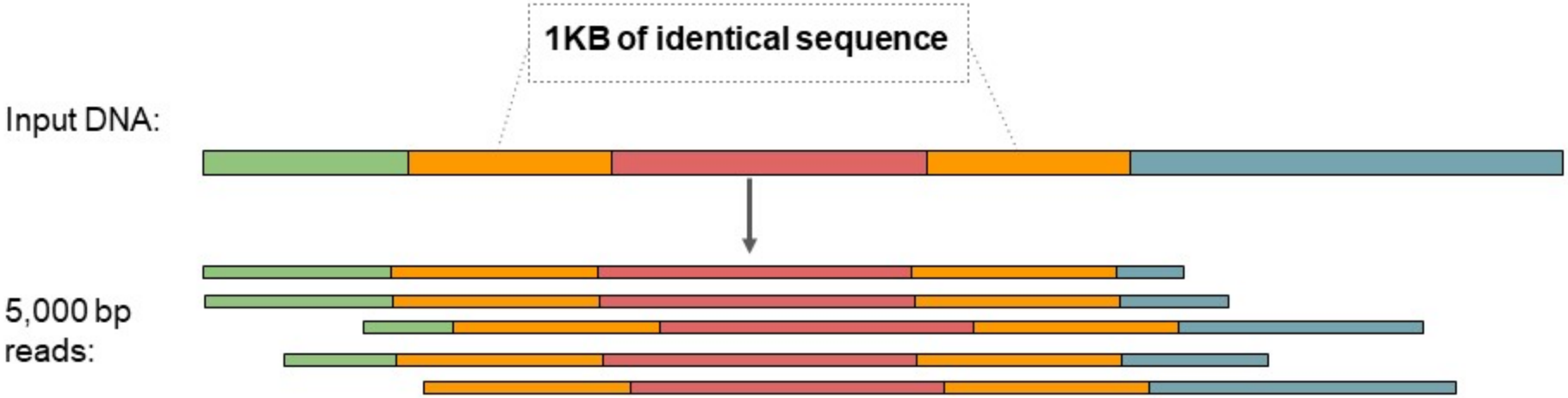
Aligning reads from repetitive sequence misses genetic variation



Aligning reads from repetitive sequence misses genetic variation

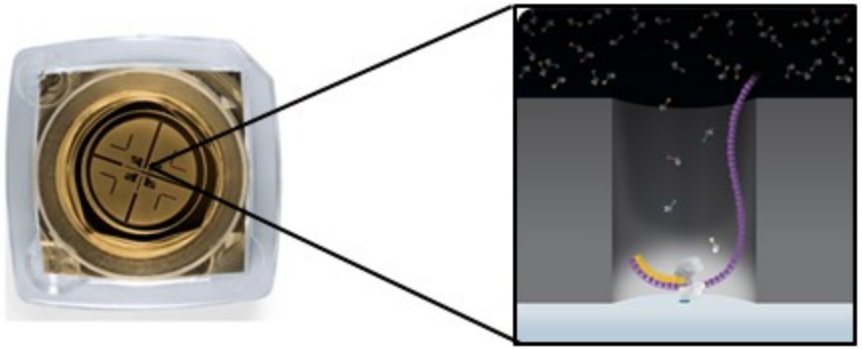


Long-read sequencing can detect more genetic variants

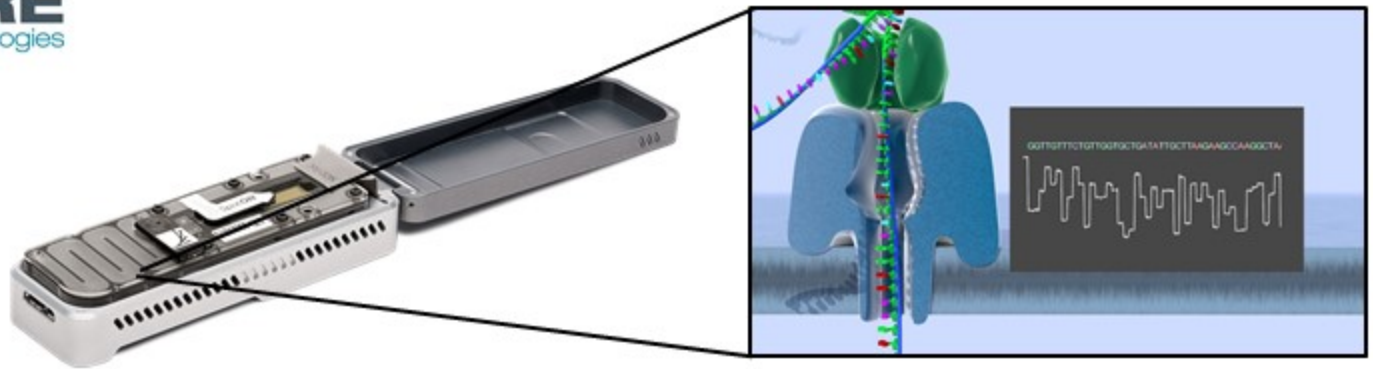


Third generation sequencing technologies are capable of generating long reads

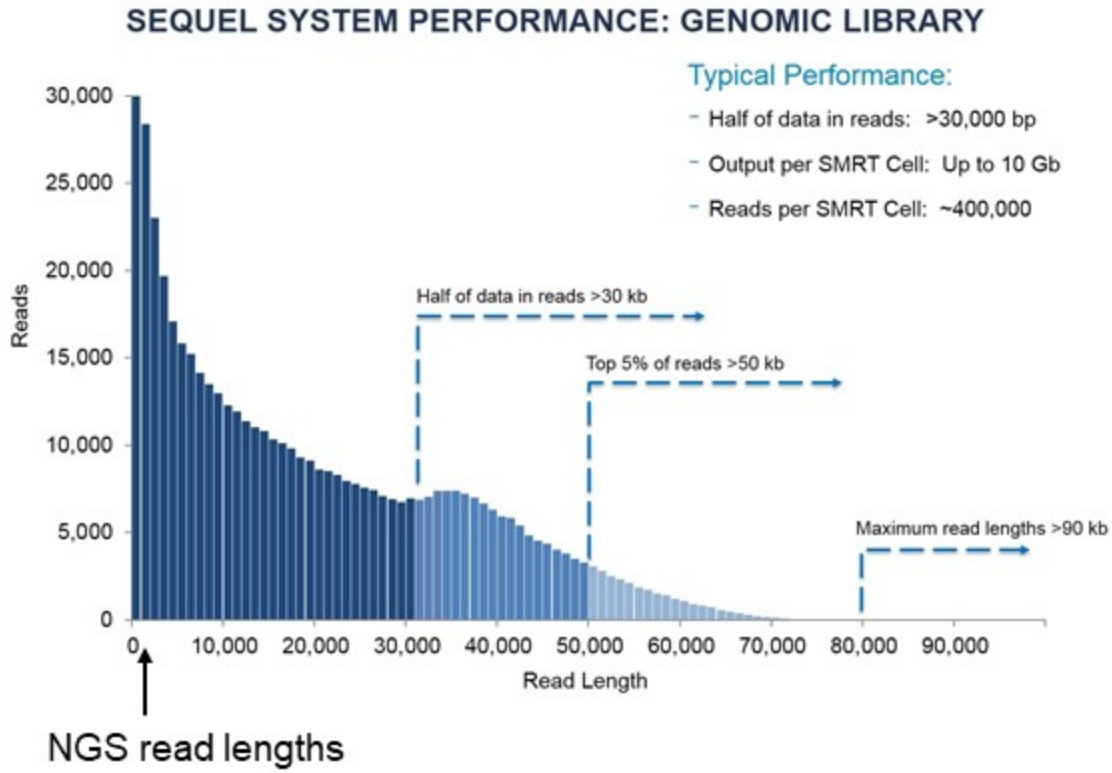
PACBIO®



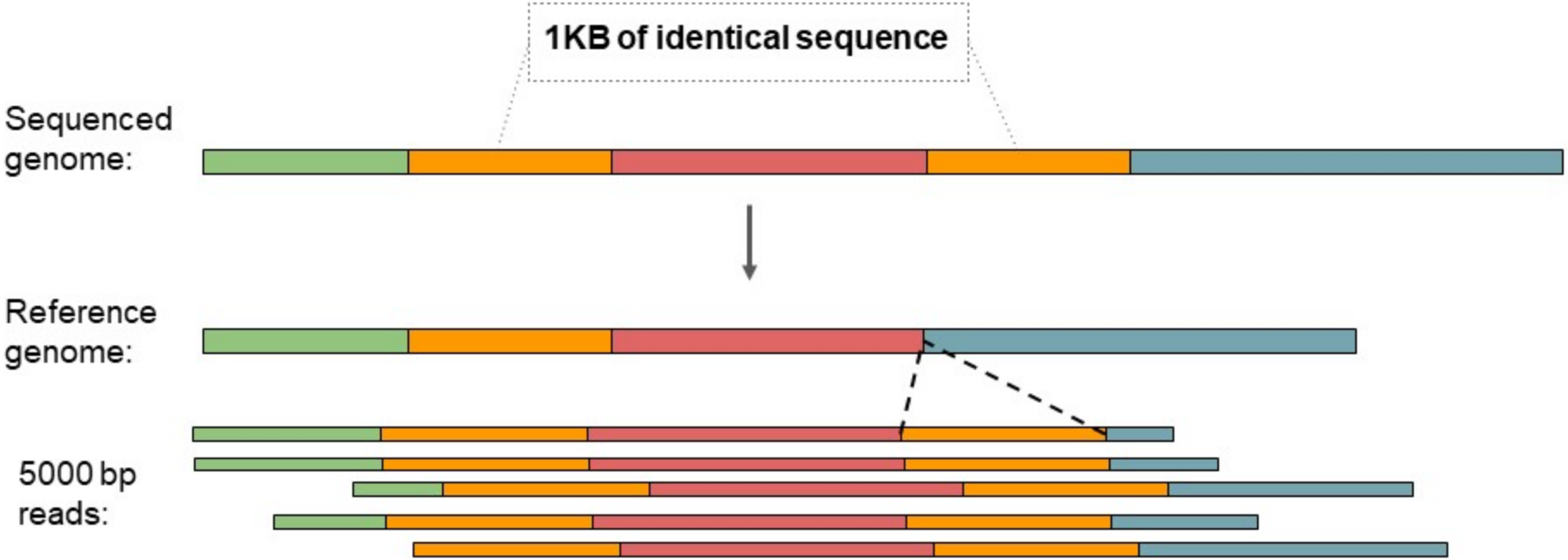
Oxford NANOPORE Technologies



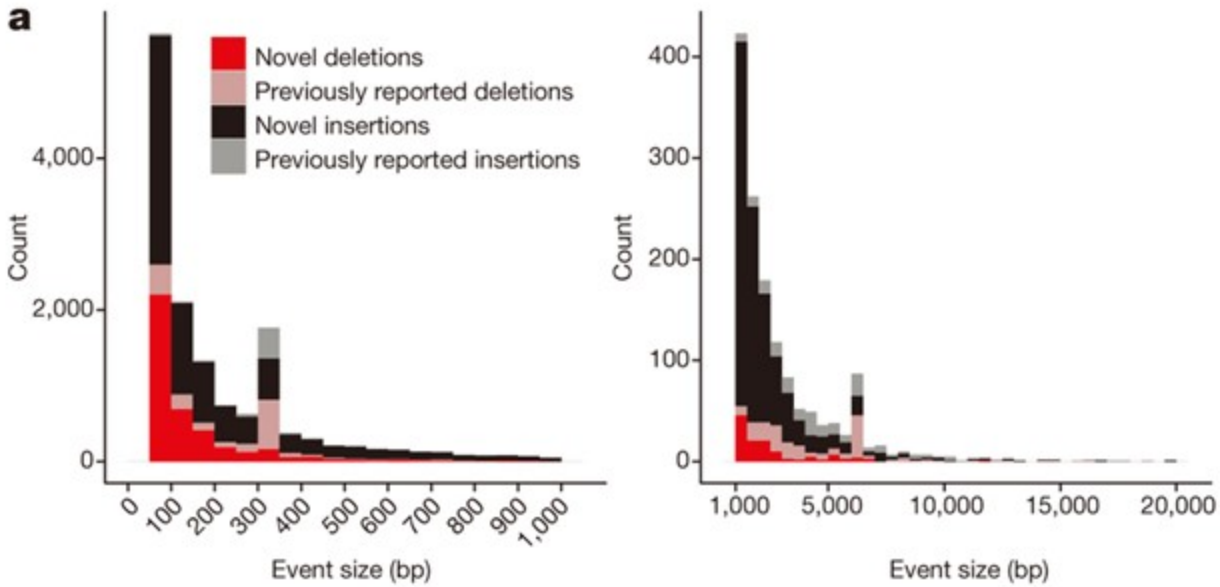
Third generation sequencing technologies are capable of generating long reads



Aligning long reads to the reference identifies previously missed insertion

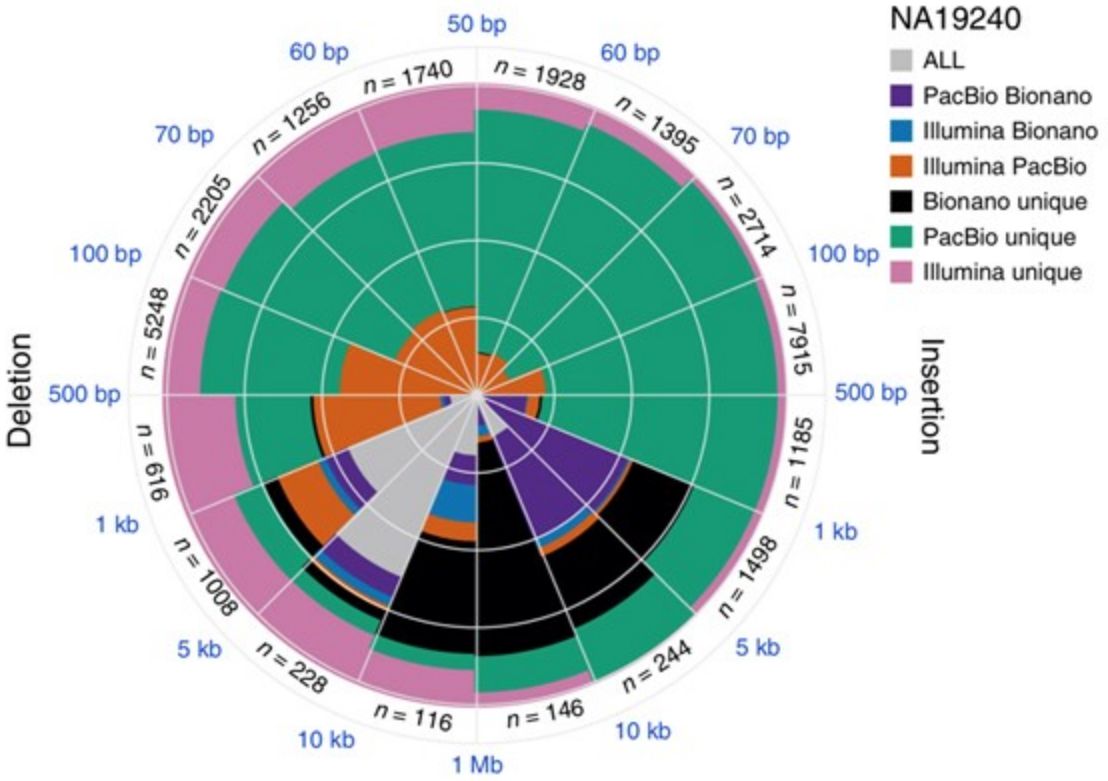


Large number of novel deletions and insertions identified in complete hydatidiform mole (CHM)



Chaisson et al. Nature Letters 2015

Large number of SVs only detected by long read sequencing methods

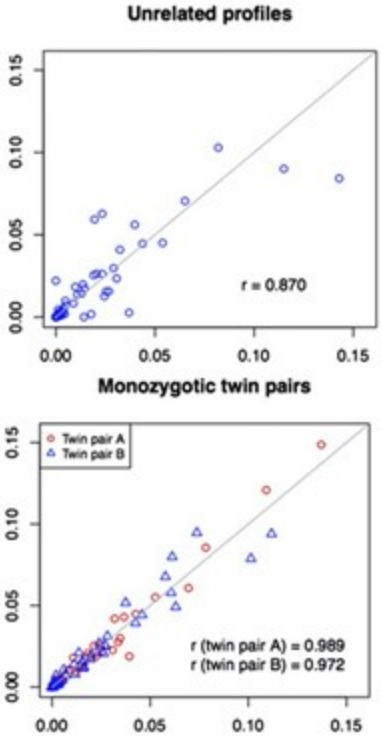


1. ~83% of insertions are being missed by short-read-calling algorithms.
2. 3 to 7 fold more variation detected than with just Illumina-based methods

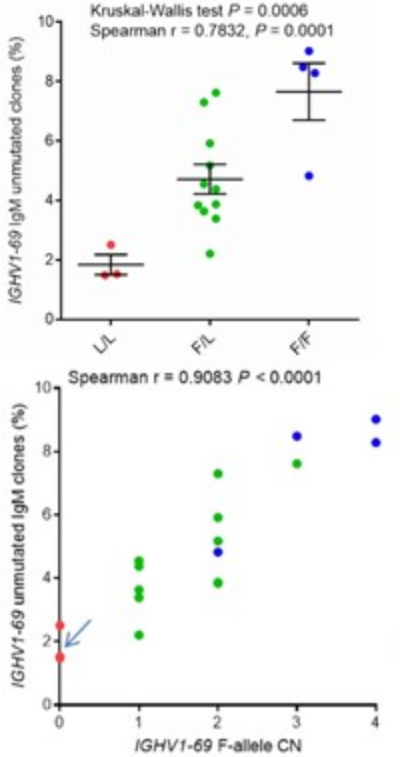
Chaisson et al. Nature Comm 2019

Evidence that genetic diversity affects the antibody repertoire

Twin studies



Functional studies



GWAS

Seven diseases/traits

- Kawasaki disease
- HDL cholesterol
- Rheumatic heart disease
- Blood protein levels
- Alzheimer's disease (late onset)
- Reaction time
- Myopia (age of diagnosis)

Identification of novel susceptibility Loci for kawasaki disease in a Han chinese population by a genome-wide association study. Tsai et al, 2011 PloS One

Exploration of haplotype research consortium imputation for genome-wide association studies in 20,032 Generation Scotland participants. Nagy R et al, 2017 Genome Medicine

Association between a common immunoglobulin heavy chain allele and rheumatic heart disease risk in Oceania. Parks T et al, 2017 Nature Communications

Genomic atlas of the human plasma proteome. Sun BB et al, 2018 Nature

Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. Lambert JC et al, 2013 Nature Genetics

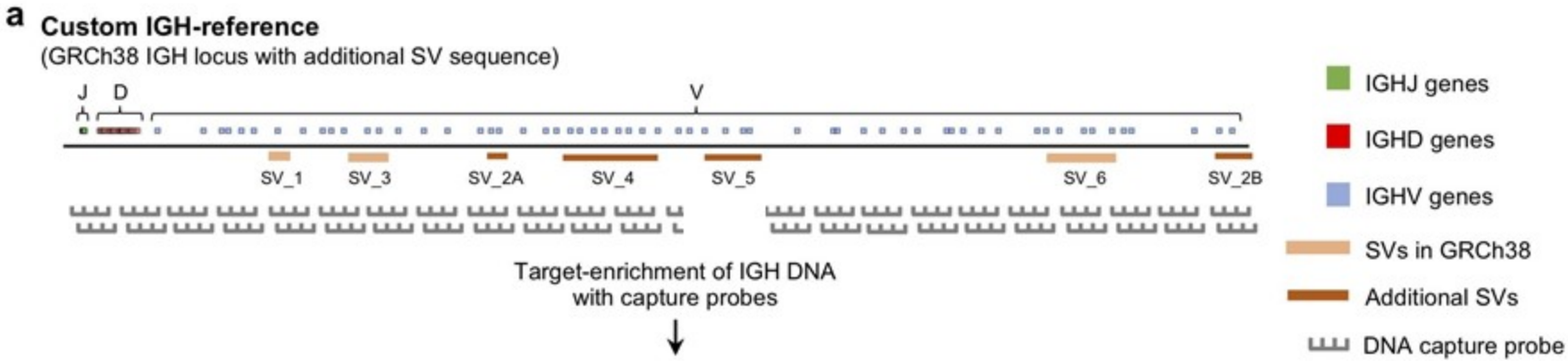
Study of 300,486 individuals identifies 148 independent genetic loci influencing general cognitive function. Davies G et al, 2018 Nature Communications

Genome-wide association meta-analysis highlights light-induced signaling as a driver for refractive error. Tedja MS et al, 2018 Nature Genetics

Resolving the IGH locus in a higher throughput and scalable fashion using long read sequencing

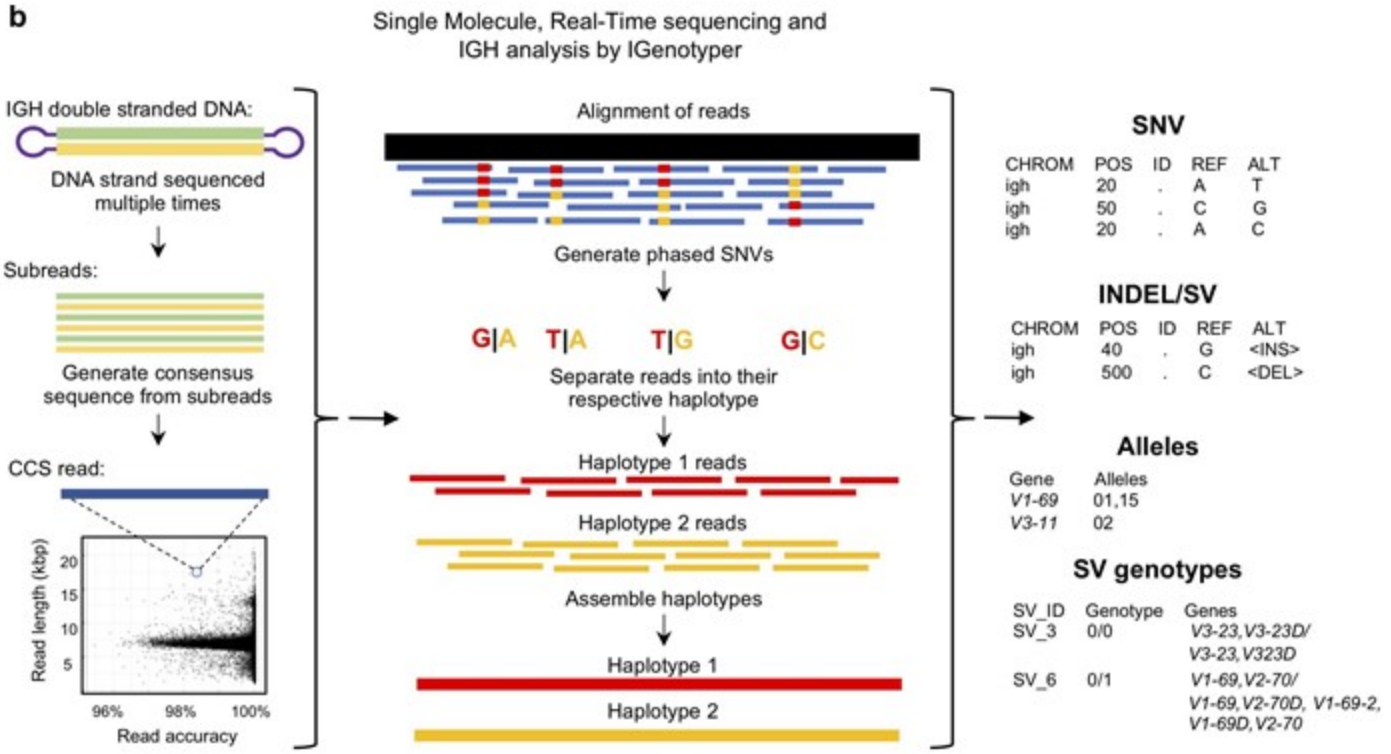
O. L. Rodriguez, W. S. Gibson, T. Parks, M. Emery, J. Powell, M. Strahl, G. Deikus, K. Auckland, E. E. Eichler, W. A. Marasco, R. Sebra, A. J. Sharp, M. L. Smith, A. Bashir, C. T. Watson, A novel framework for characterizing genomic haplotype diversity in the human immunoglobulin heavy chain locus, *Frontiers in Immunology*. (2020)

Specifically targeting the IGH locus



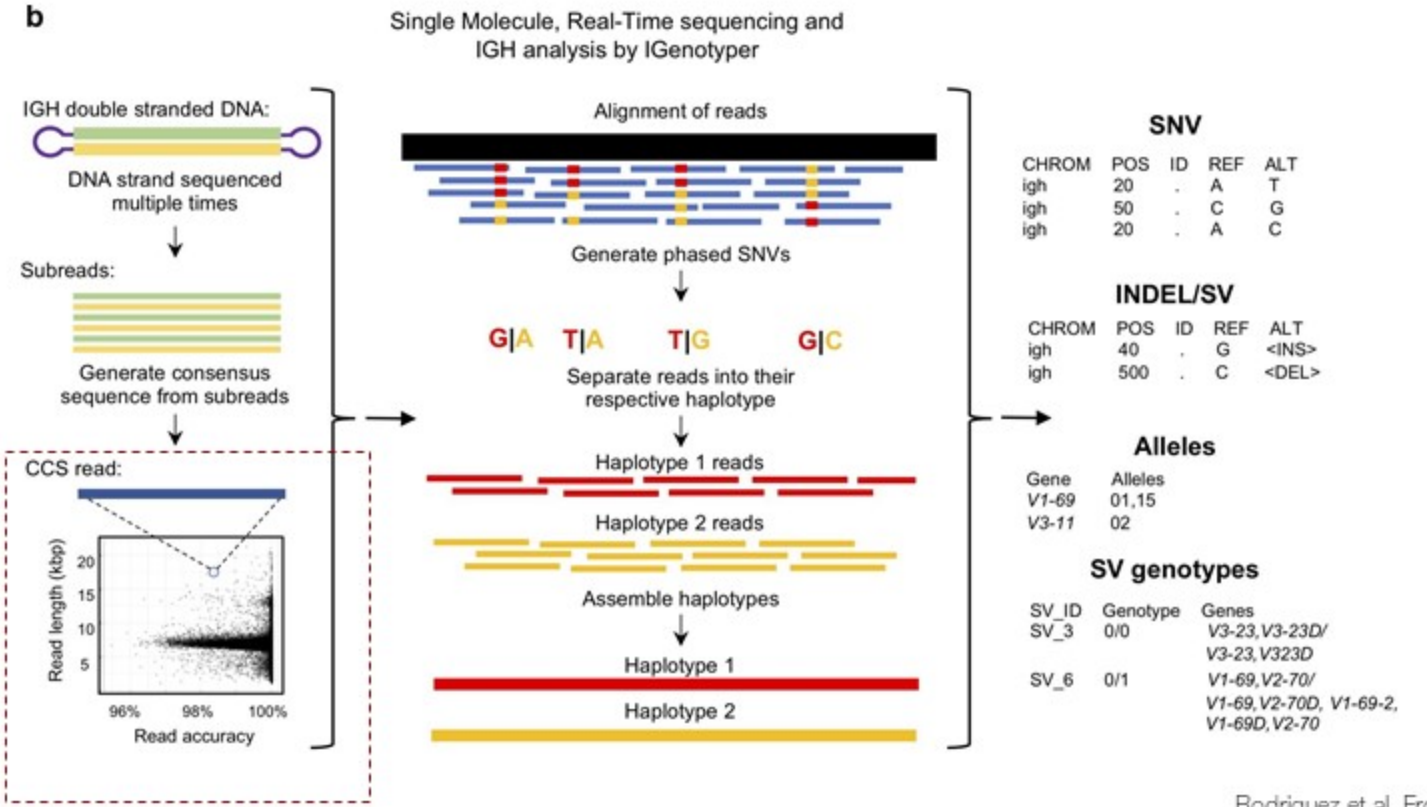
Rodriguez et al. Frontiers in Imm 2020

Long read sequencing of targeted DNA



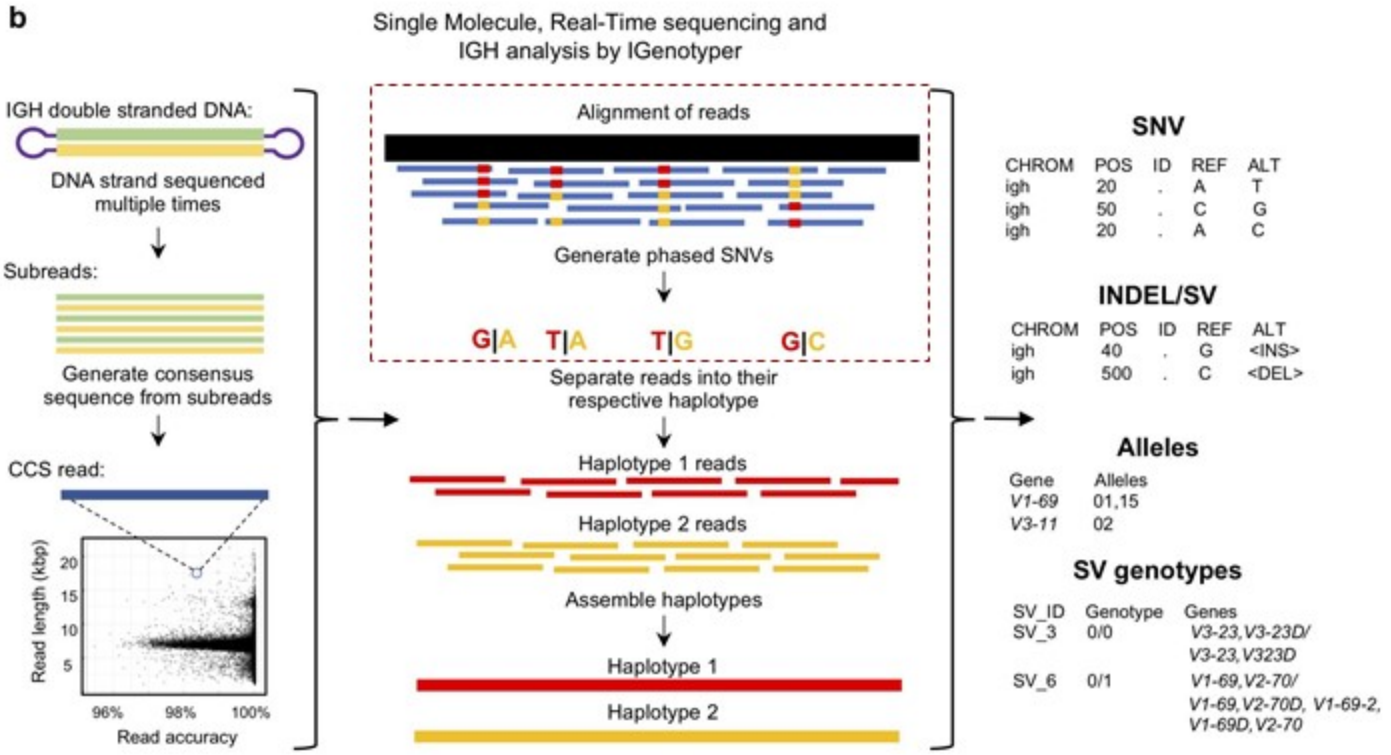
Rodriguez et al. Frontiers in Imm 2020

Input to IGenotyper are the highly accurate CCS/HiFi reads



Rodriguez et al. Frontiers in Imm 2020

First step in IGenotyper is to detect phased SNVs



Rodriguez et al. Frontiers in Imm 2020

Input reads are aligned to the reference genome

Rodriguez et al. Bioinformatics 2020

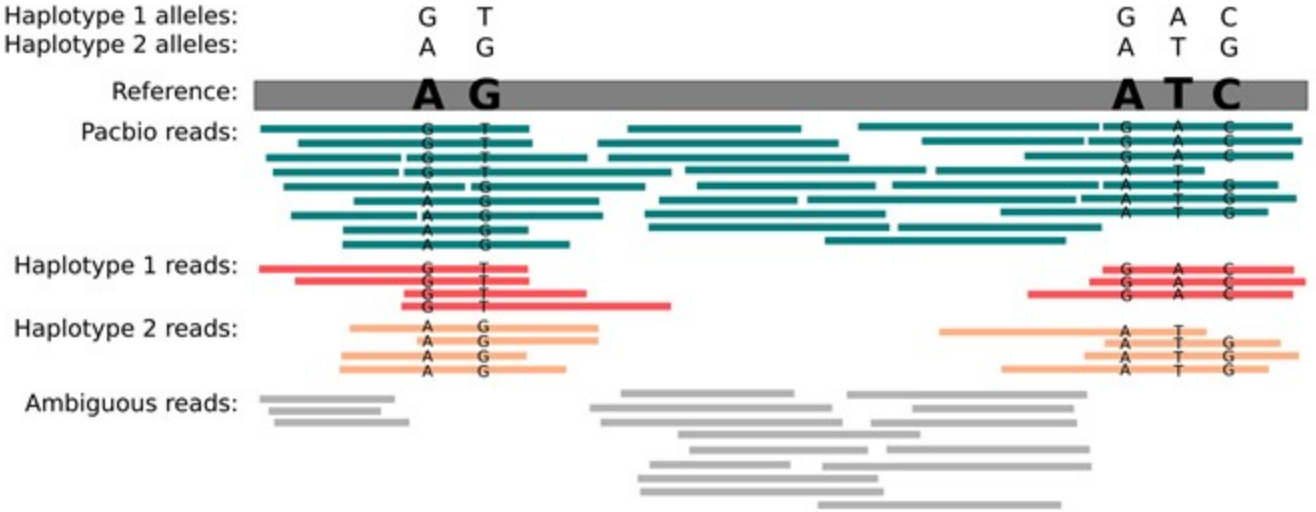
Phased heterozygous SNVs detected using WhatsHap

Rodriguez et al. Bioinformatics 2020

Phased heterozygous SNVs are used to partition reads into their respective haplotype

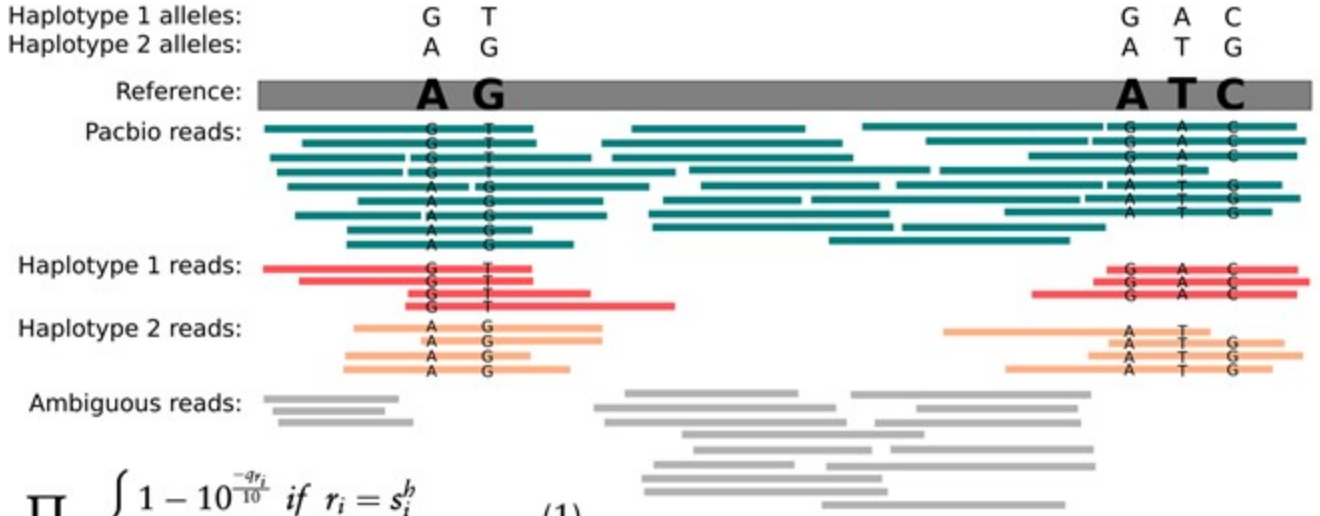
Rodriguez et al. Bioinformatics 2020

PacBio reads with haplotype-specific bases are partitioned into their respective haplotype



Rodriguez et al. Bioinformatics 2020

PacBio reads with haplotype-specific bases are partitioned into their respective haplotype

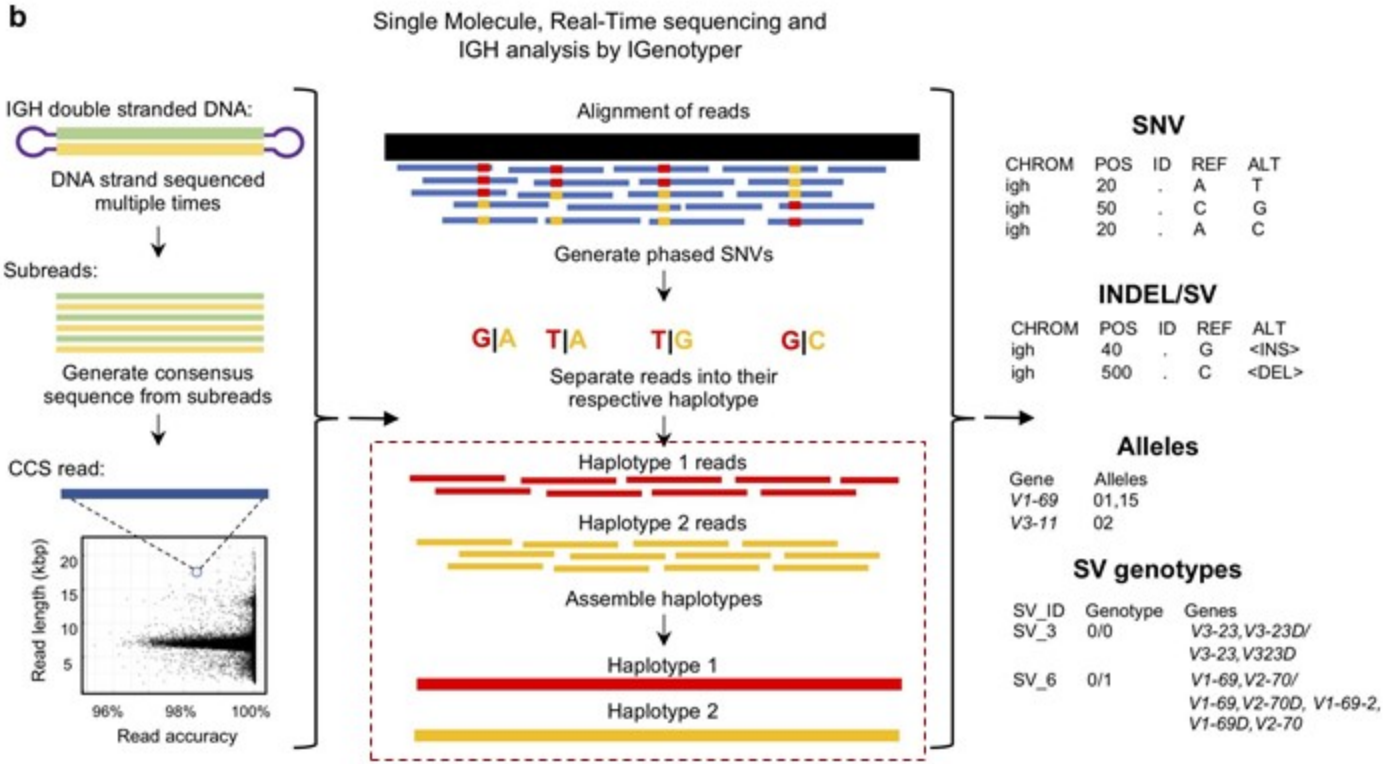


$$P(r|\ h_r = h) = \prod_{(s_i^1, s_i^2) \in S_r} \begin{cases} 1 - 10^{\frac{-q_{r_i}}{10}} & \text{if } r_i = s_i^h \\ 10^{\frac{-q_{r_i}}{10}} & \text{otherwise} \end{cases} \quad (1)$$

$$h_r = \begin{cases} 1 \text{ if } \dfrac{P(r|h_r = 1)}{P(r|h_r = 1) + P(r|h_r = 2)} > \tau \\ 2 \text{ if } \dfrac{P(r|h_r = 2)}{P(r|h_r = 1) + P(r|h_r = 2)} > \tau \\ 0 \text{ otherwise,} \end{cases} \quad (2)$$

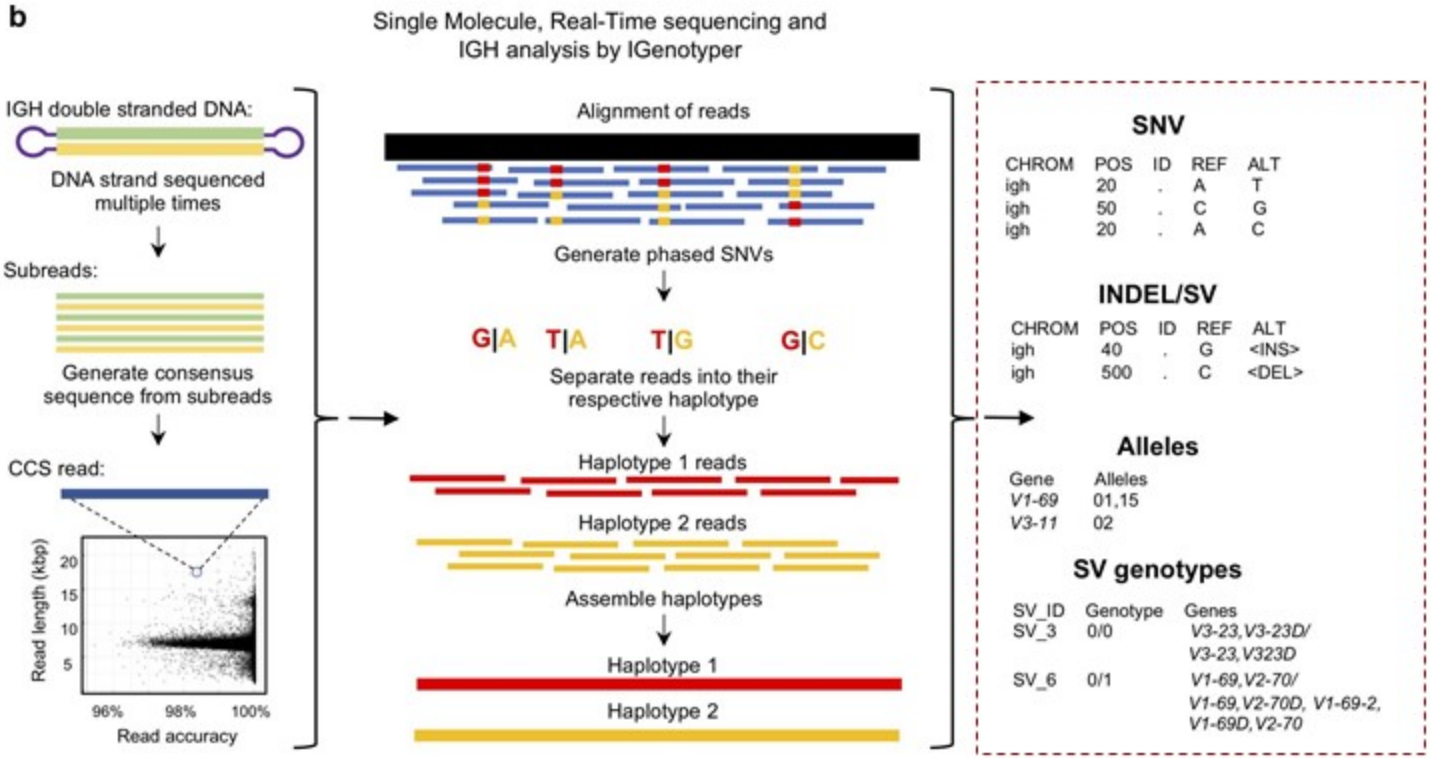
Rodriguez et al. Bioinformatics 2020

IGenotyper produces haplotype-specific assemblies



Rodriguez et al. Frontiers in Imm 2020

Haplotype-resolved assemblies are used to detect genetic variation



Rodriguez et al. Frontiers in Imm 2020

After assembling each haplotype, next step is to detect indels and SVs

Multiple Sequence Alignment

```
Ref:   CCCTTGTCCGACTTTACACCAAGGCGAGTTATTTTAAGCAGAGAGCGTCATCTA-------TCCGAAAGATAT
Hap1:  CCCTTGTCCGAC-----------CCGAGTTATTTTAAGCAGAGAGCGTCATCTA-------TCCGAAAGATAT
Hap2:  CCCTTGTCCGAC-----------GCGAGTTATTTTAAGCAGAGAGCGTCATCTAATGGAGATCCGAAAGATAT
```

Hidden Markov Model

14 modeled states:
DELETION:
- 1/1, 0|1, 1|0

INSERTION:
- 1/1, 0|1, 1|0

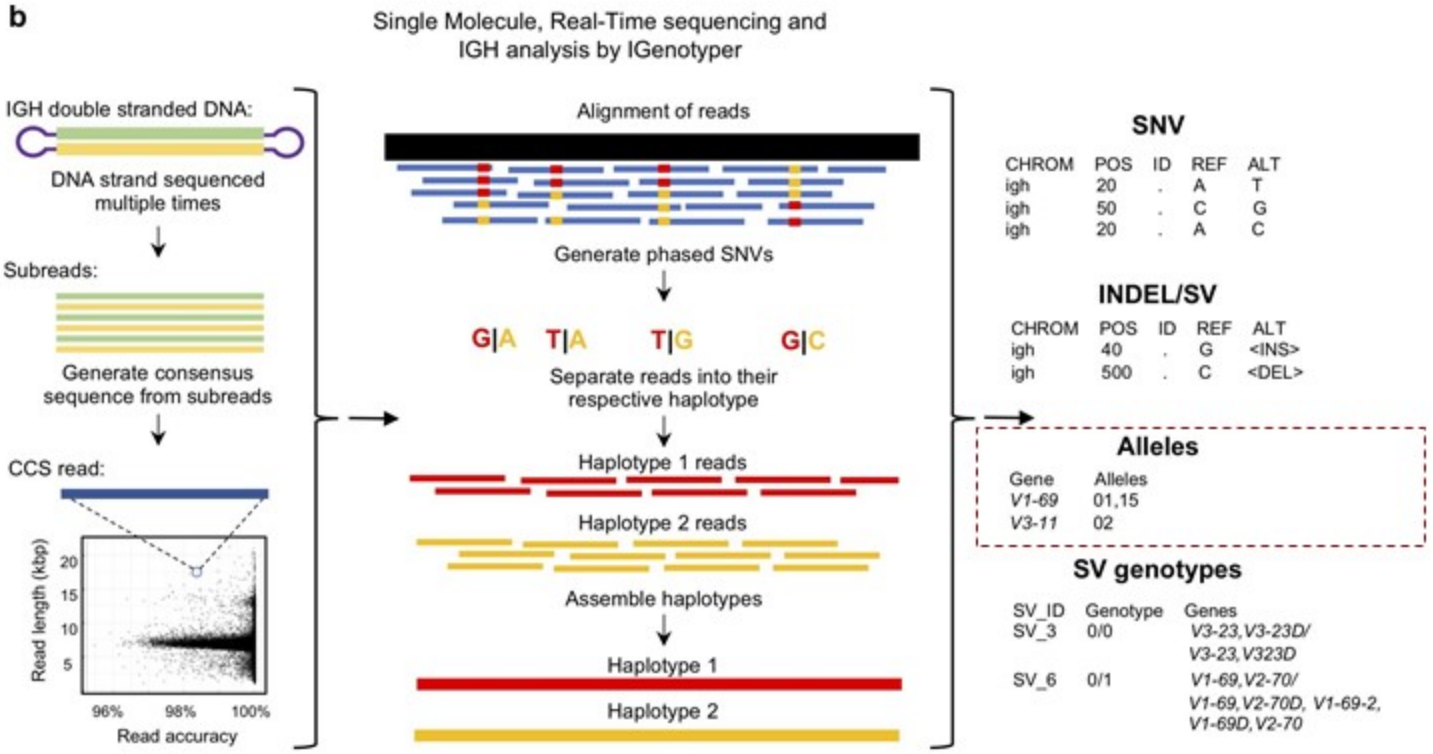
COMPLEX DELETIONS:
- 1/1, 0|1, 1|0

COMPLEX INSERTIONS:
- 1/1, 0|1, 1|0

NORMAL
COMPLEX STATES

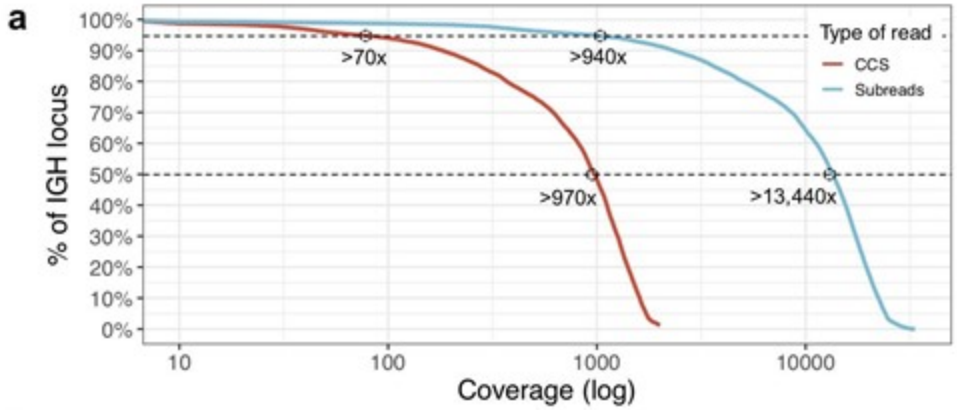
Rodriguez et al. Bioinformatics 2020

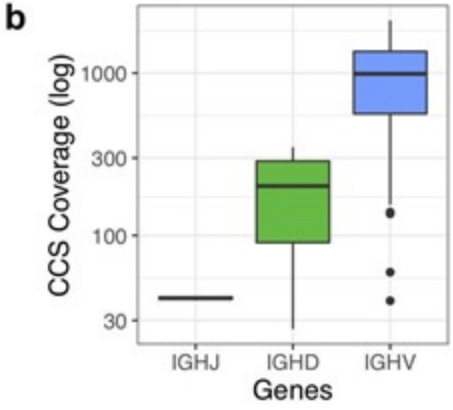
Genes and alleles are directly annotated from assemblies



Rodriguez et al. Frontiers in Imm 2020

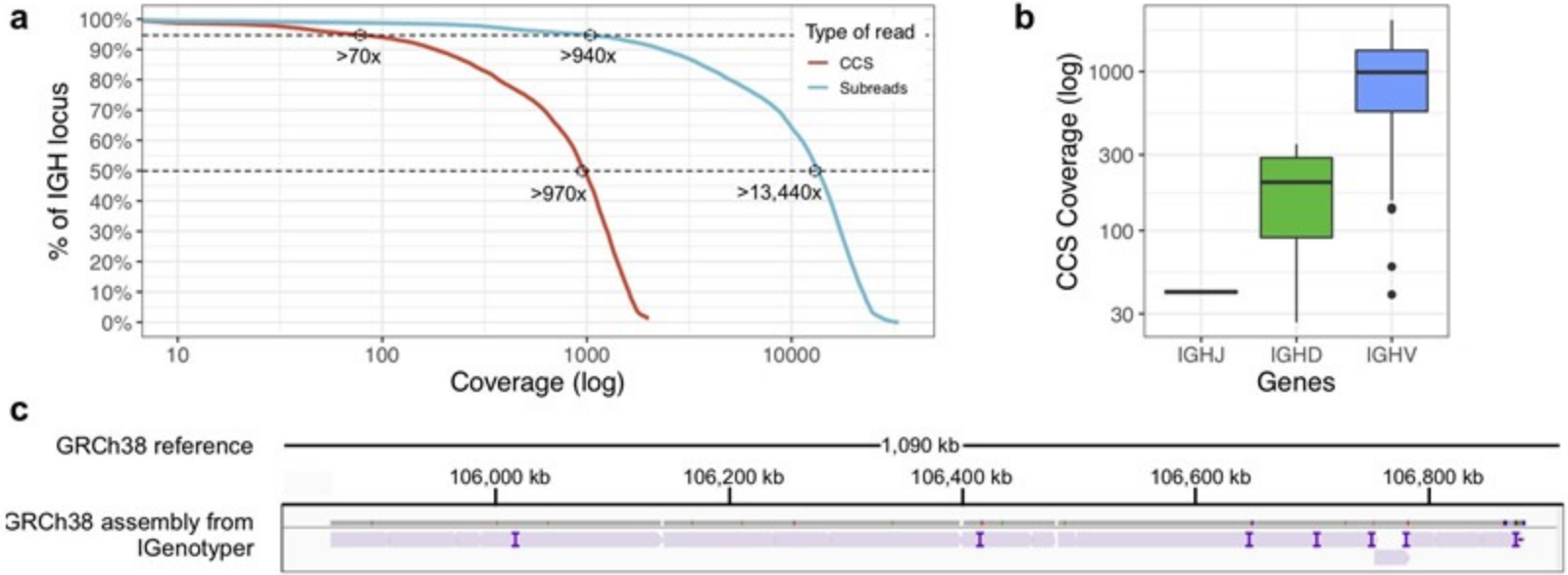
IGH locus is captured at an extremely high coverage





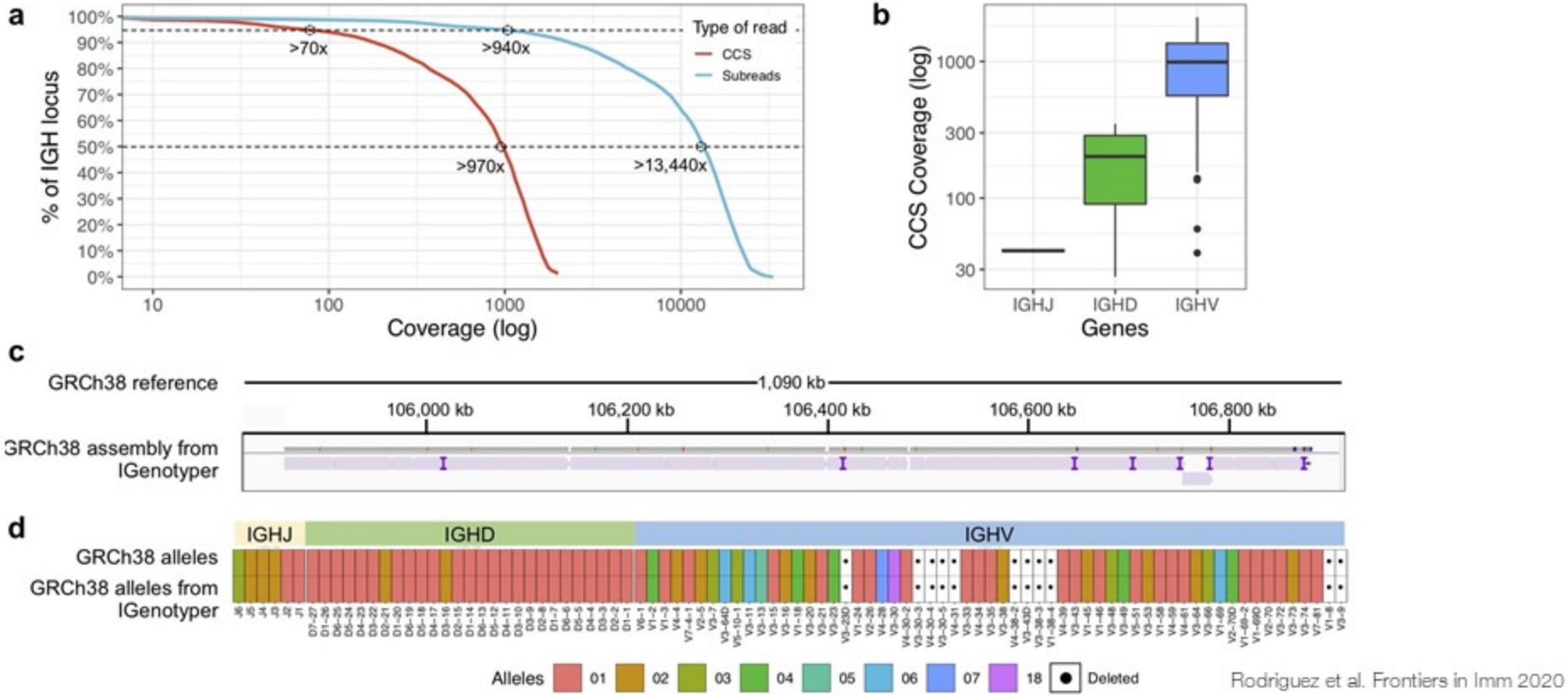
Rodriguez et al. Frontiers in Imm 2020

IGenotyper assembled 98.7% of IGH with only 37 errors producing an accuracy > 99.99%



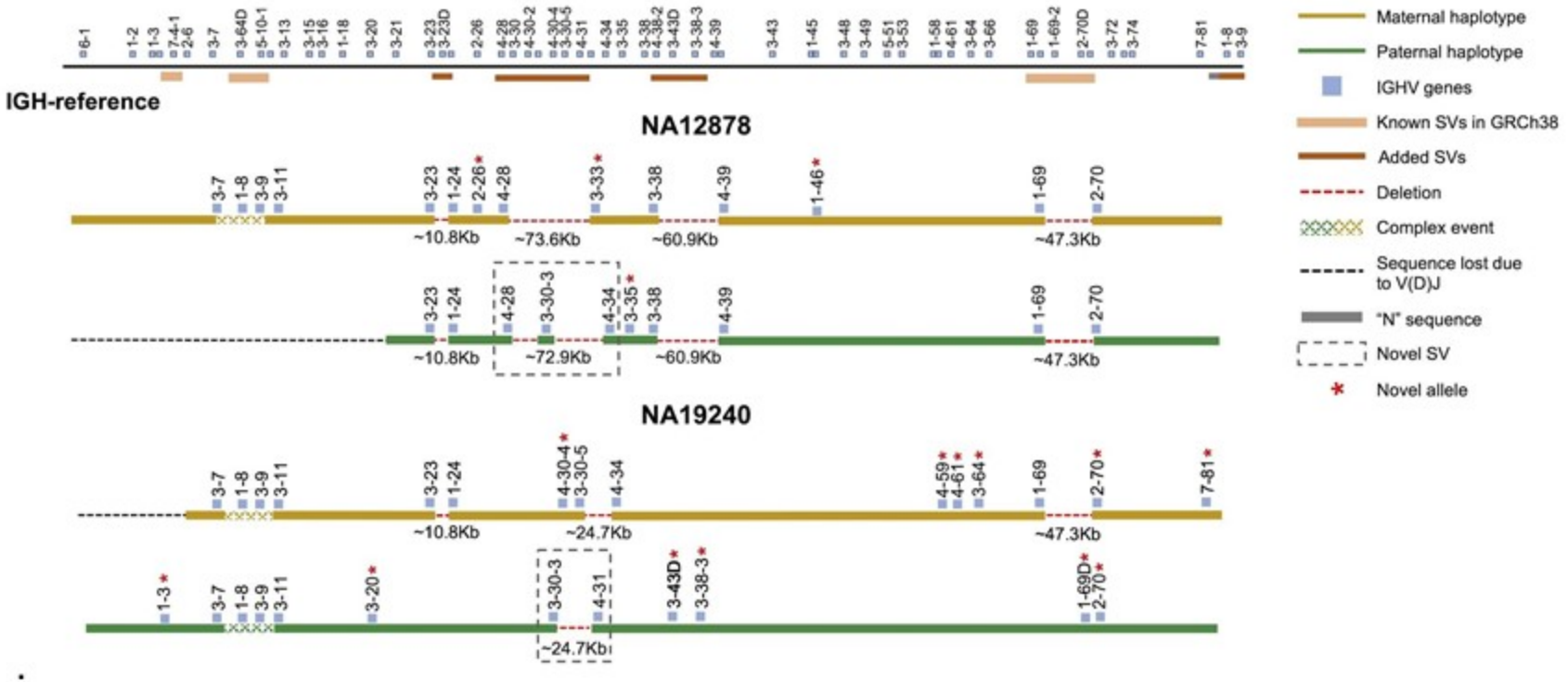
Rodriguez et al. Frontiers in Imm 2020

All alleles are correctly assigned



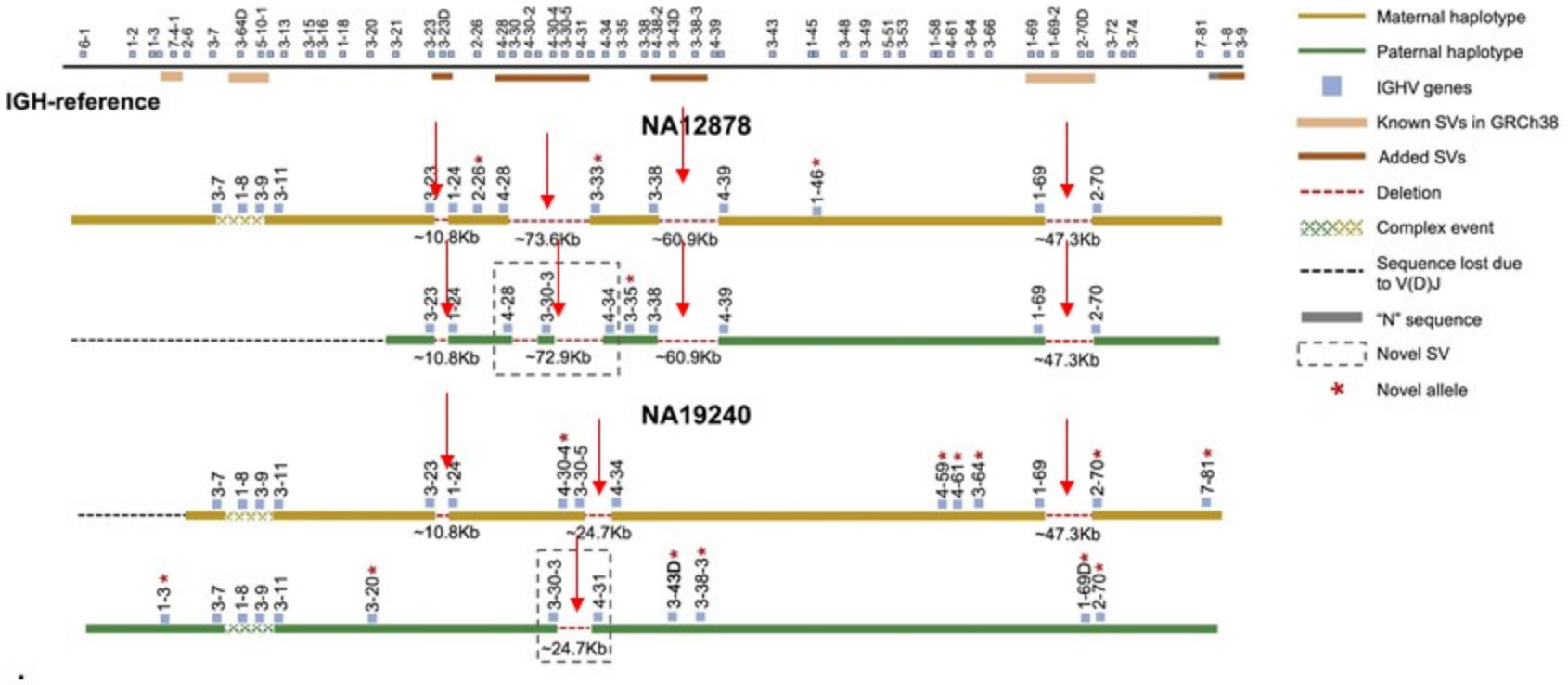
Rodriguez et al. Frontiers in Imm 2020

Resolving the IGH locus in two diploid samples



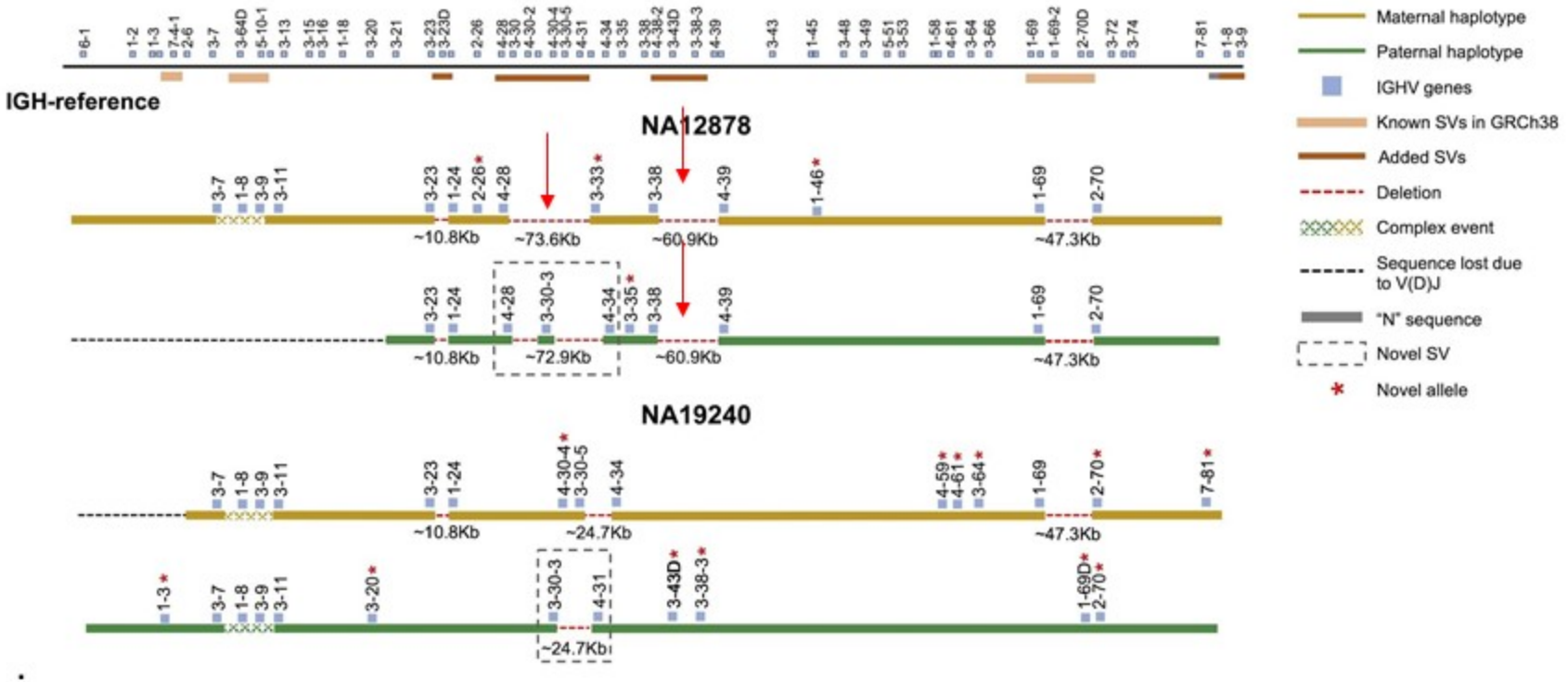
Rodriguez et al. Frontiers in Imm 2020

Resolving the IGH locus in two diploid samples identifies several SVs



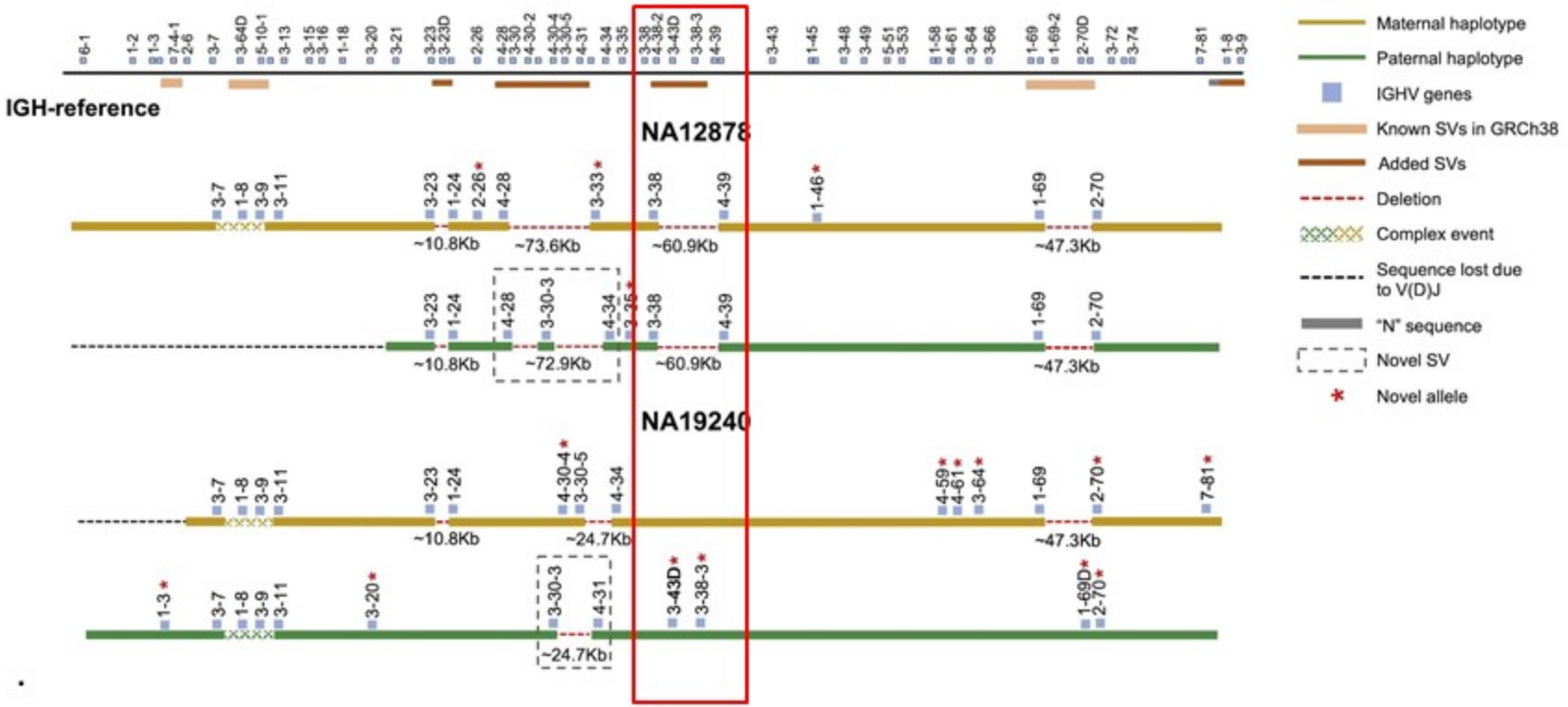
Rodriguez et al. Frontiers in Imm 2020

Resolving the IGH locus in two diploid samples identifies large SVs



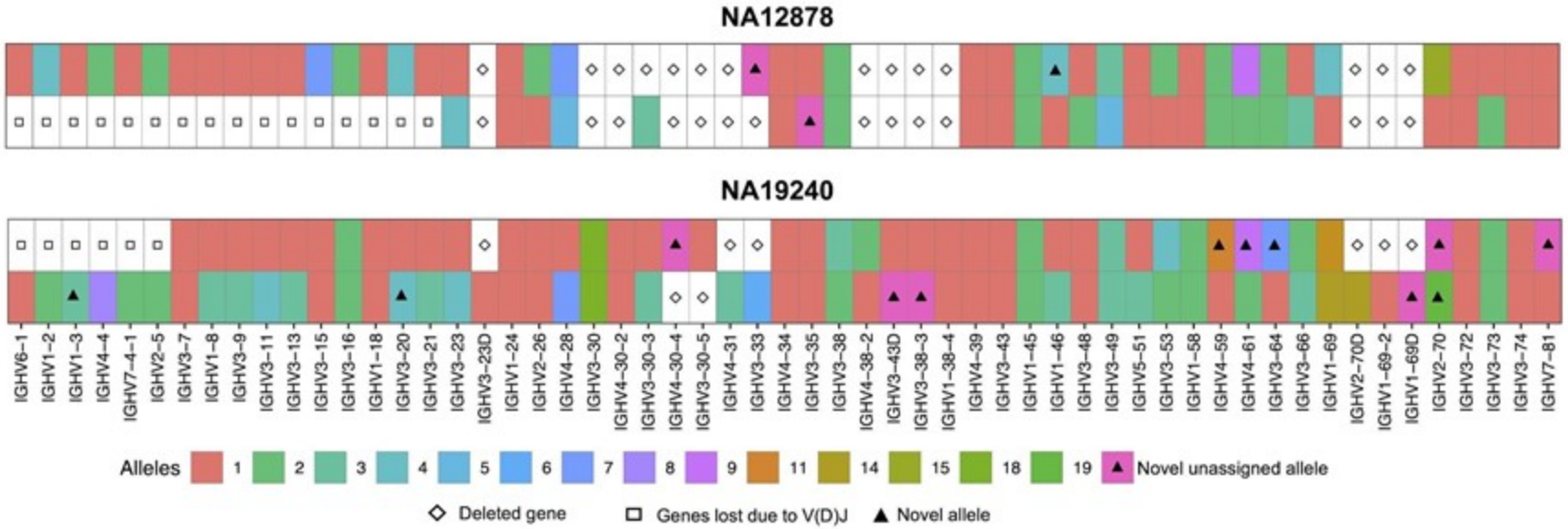
Rodriguez et al. Frontiers in Imm 2020

Resolving the IGH locus in two diploid samples identifies inter-individual haplotype variation



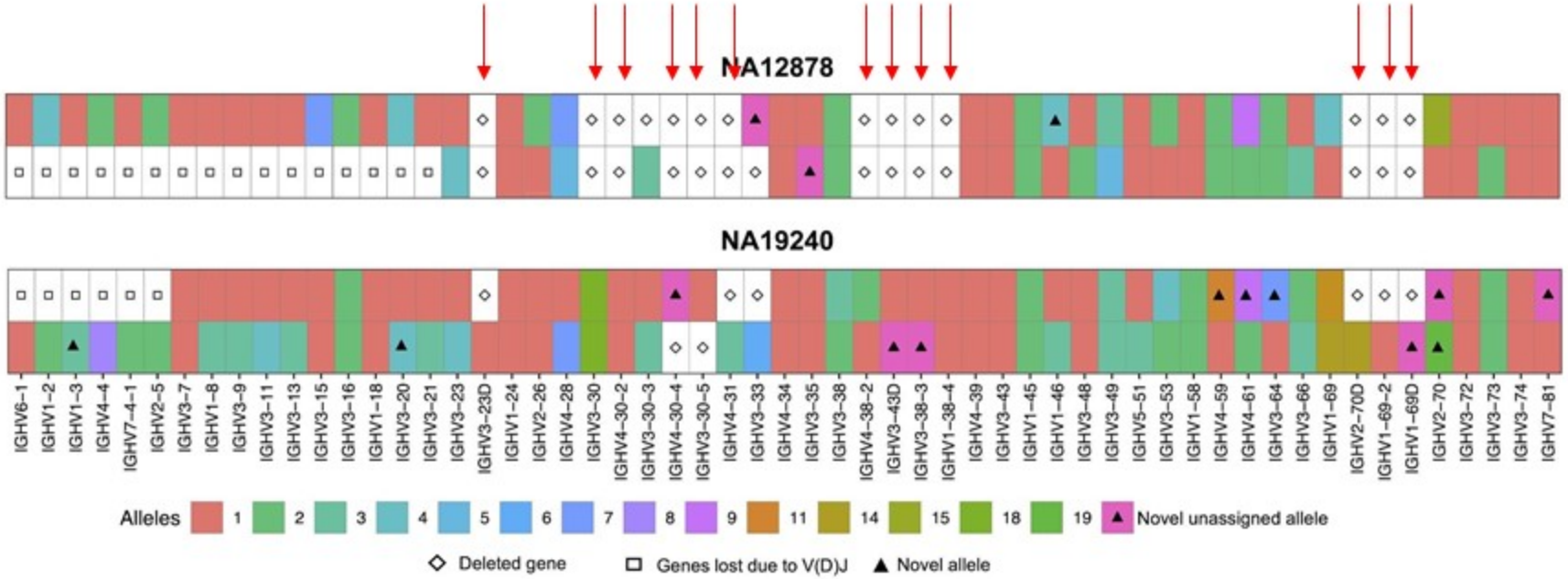
Rodriguez et al. Frontiers in Imm 2020

Alleles from each haplotype are resolved using the IGH assemblies



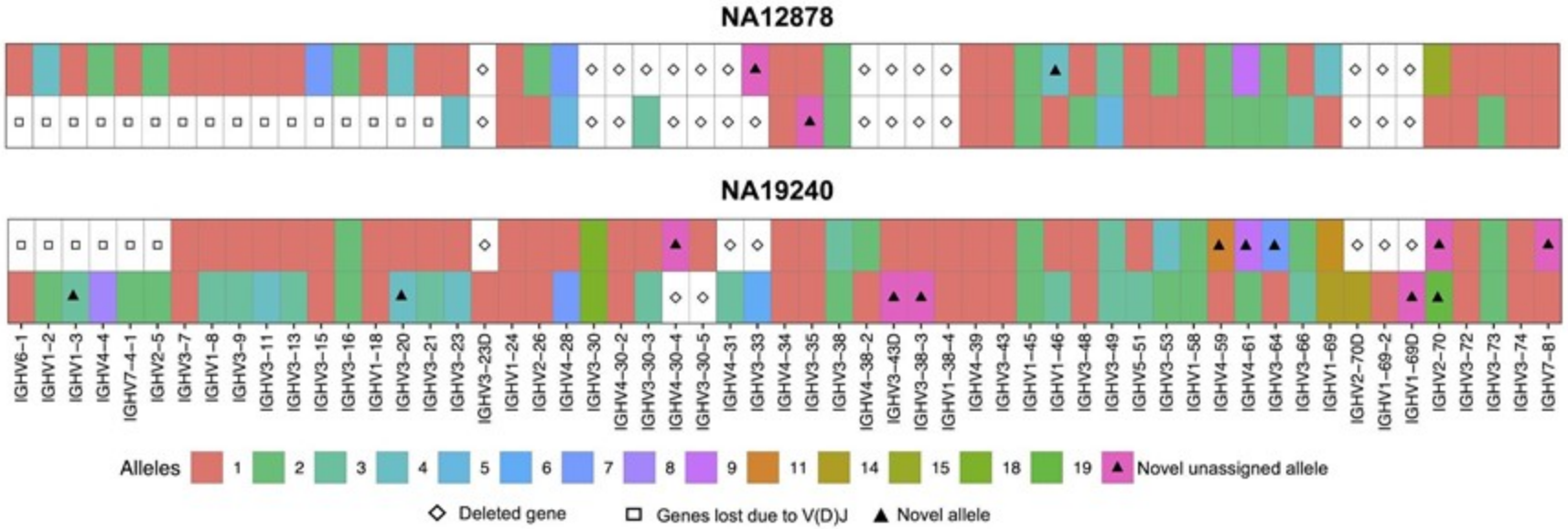
Rodriguez et al. Frontiers in Imm 2020

Alleles from each haplotype are resolved using the IGH assemblies



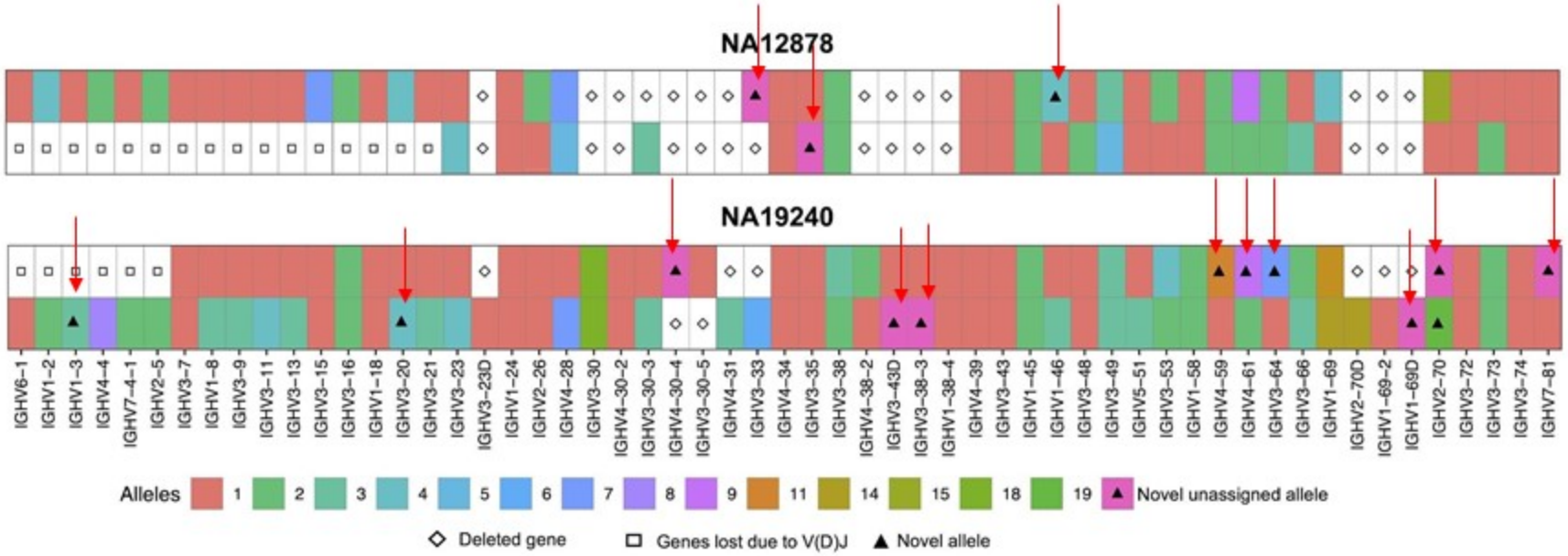
Rodriguez et al. Frontiers in Imm 2020

Alleles from each haplotype are resolved using the IGH assemblies



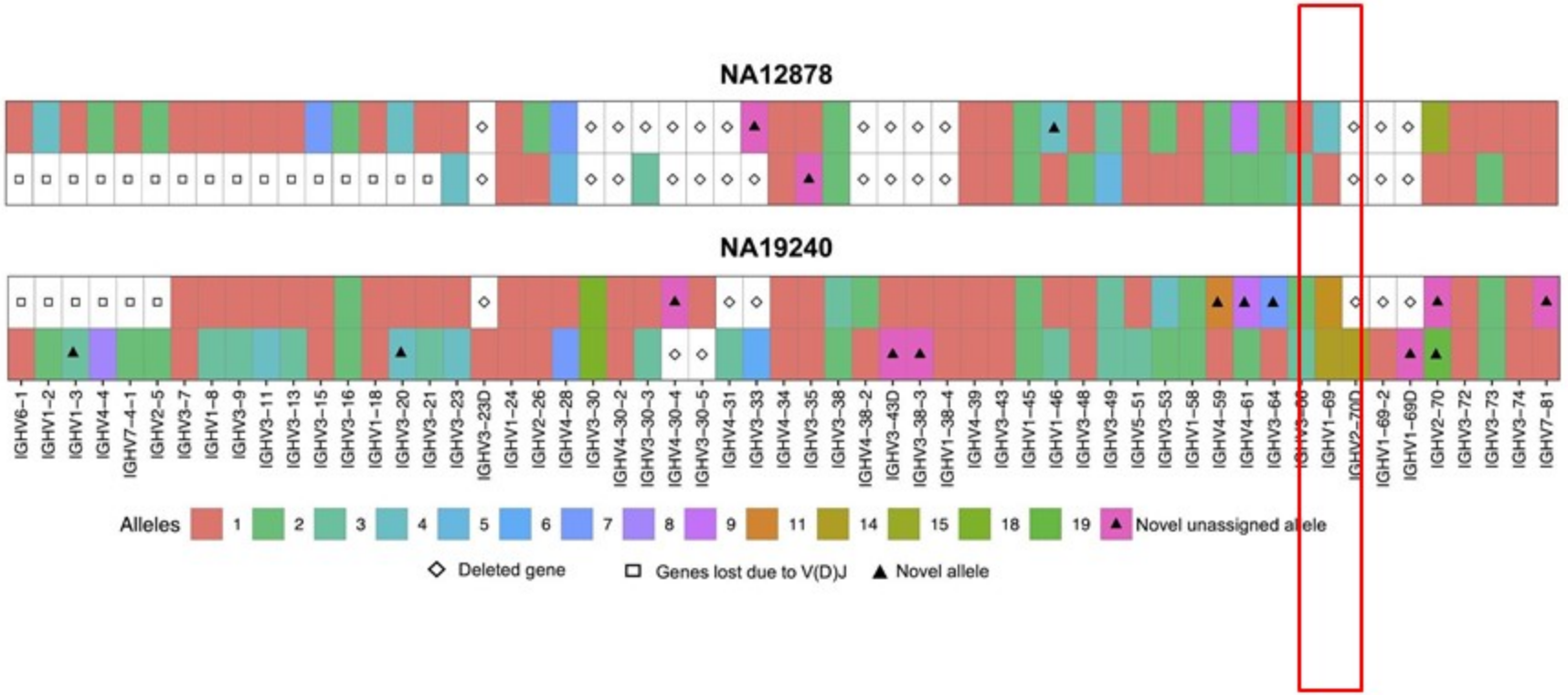
Rodriguez et al. Frontiers in Imm 2020

Alleles from each haplotype are resolved using the IGH assemblies



Rodriguez et al. Frontiers in Imm 2020

Alleles from each haplotype are resolved using the IGH assemblies



Rodriguez et al. Frontiers in Imm 2020

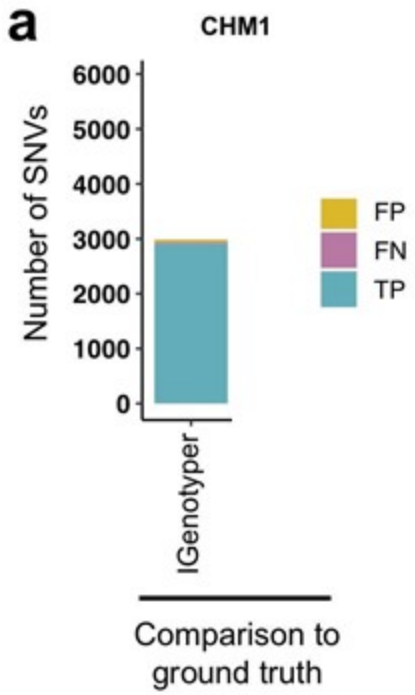
IGH locus is accurately resolved

TABLE 1 | Assembly statistics and evaluation of the accuracy of the haplotype-specific assemblies.

Sample	Contigs (n)	Assembly size (bp)	Assembly validation		
			Concordance with fosmids (SMRT sequencing)	Concordance with BACs or fosmids (Sanger sequencing)	Concordance with Pilon/Illumina
CHM1	16	1,026,385	NA	99.996%	NA
NA19240	38	1,829,616	99.996%	99.99%	99.99%
NA12878	45	1,442,310	99.995%	100.0%	99.99%

Rodriguez et al. Frontiers in Imm 2020

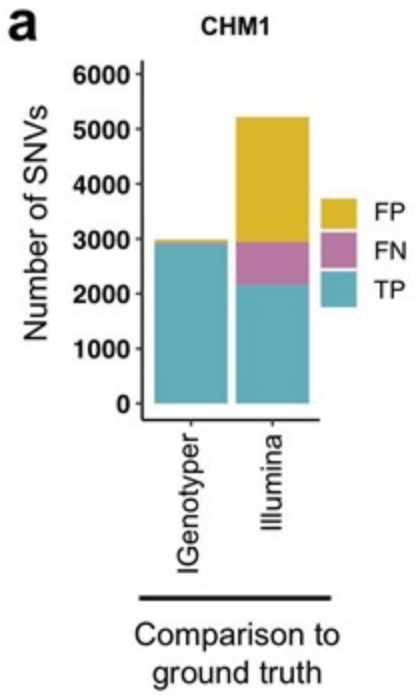
IGenotyper SNVs are more accurate than short-read data SNVs



1. 2,958 SNVs using IGenotyper
    a. 2,912 (99.0%) true SNVs
    b. 46 false-positive SNVs

Rodriguez et al. Frontiers in Imm 2020

IGenotyper SNVs are more accurate than short-read data SNVs

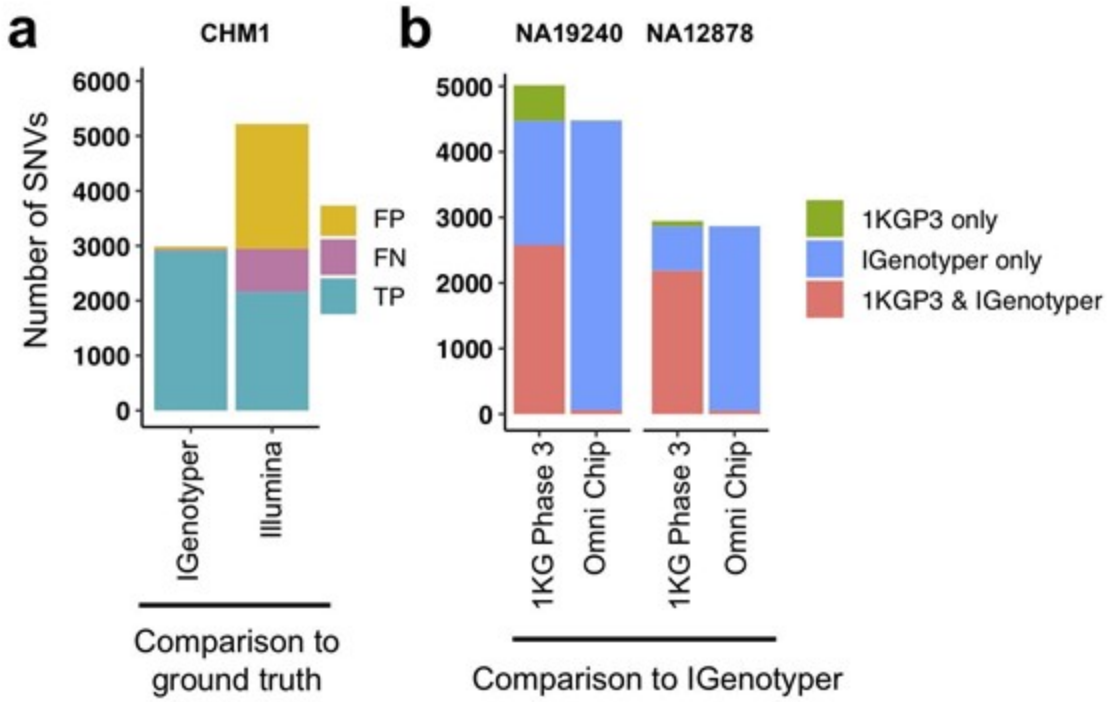


1. 2,958 SNVs using IGenotyper
    a. 2,912 (99.0%) true SNVs
    b. 46 false-positive SNVs

1. 4,433 SNVs from short read data
    a. 2,159 (49%) true SNVs
    b. 2,274 false-positive SNVs

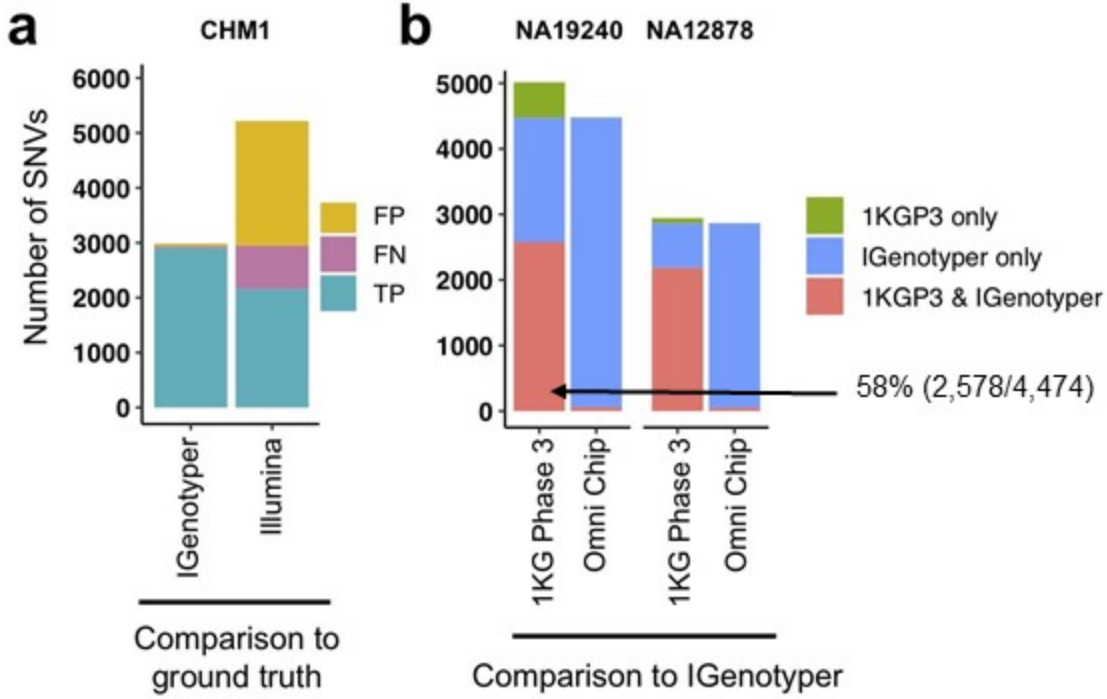
Rodriguez et al. Frontiers in Imm 2020

IGenotyper SNVs are more accurate than short-read data SNVs



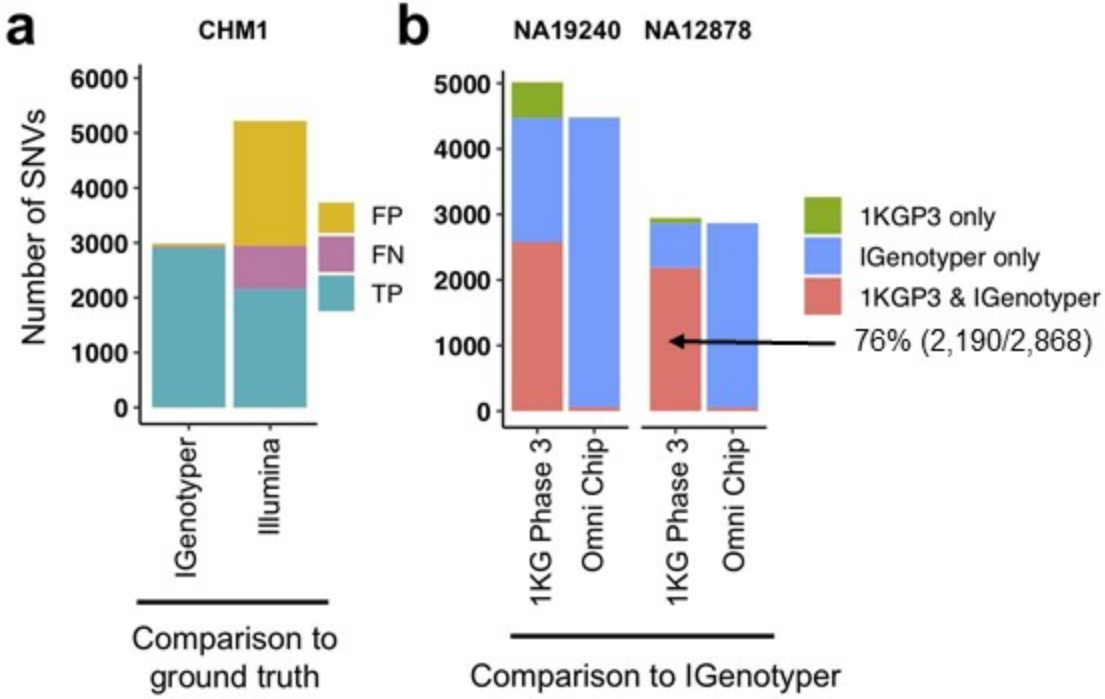
Rodriguez et al. Frontiers in Imm 2020

IGenotyper SNVs are more accurate than short-read data SNVs



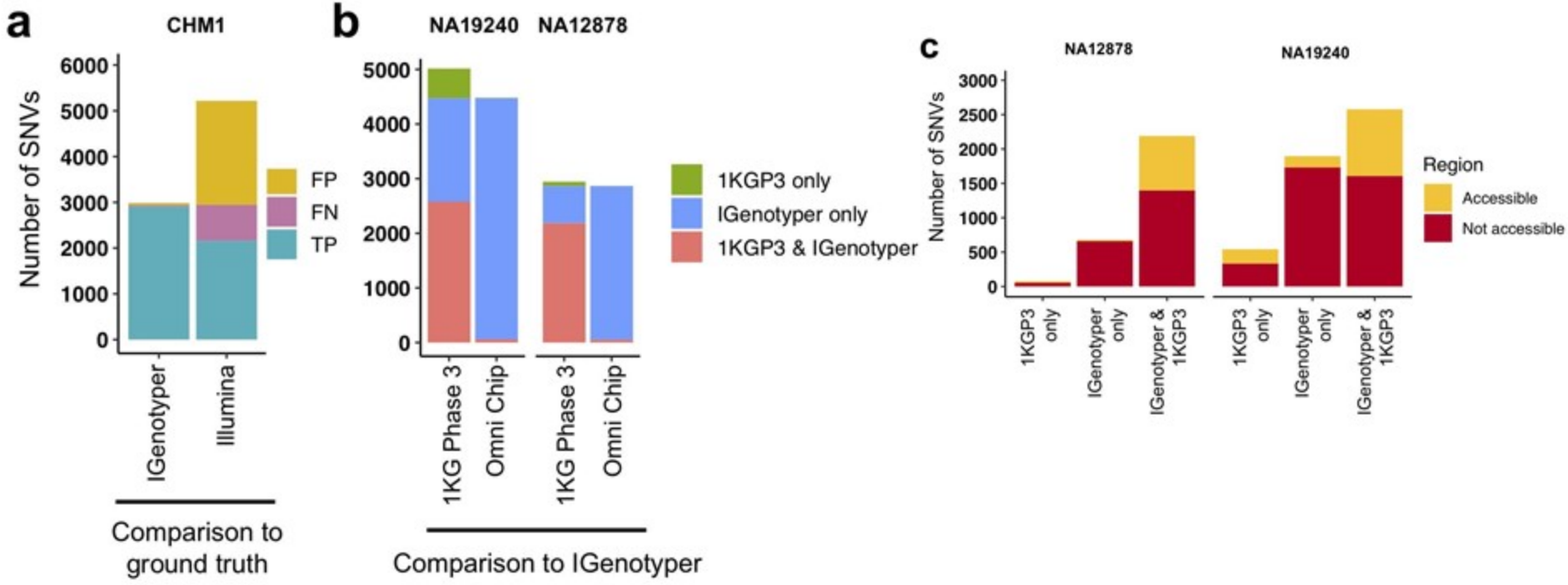
Rodriguez et al. Frontiers in Imm 2020

IGenotyper SNVs are more accurate than short-read data SNVs



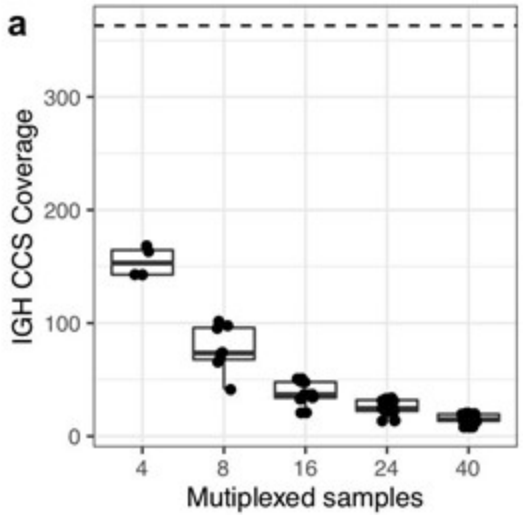
Rodriguez et al. Frontiers in Imm 2020

SNVs found just by IGenotyper were within inaccessible regions



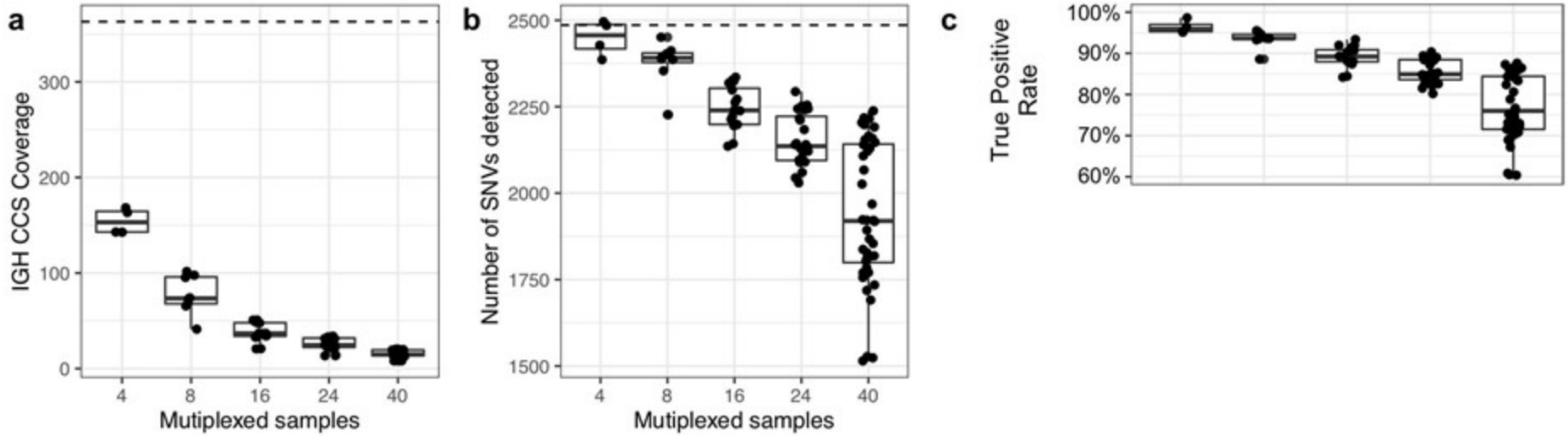
Rodriguez et al. Frontiers in Imm 2020

Coverage decreases when multiplexing samples



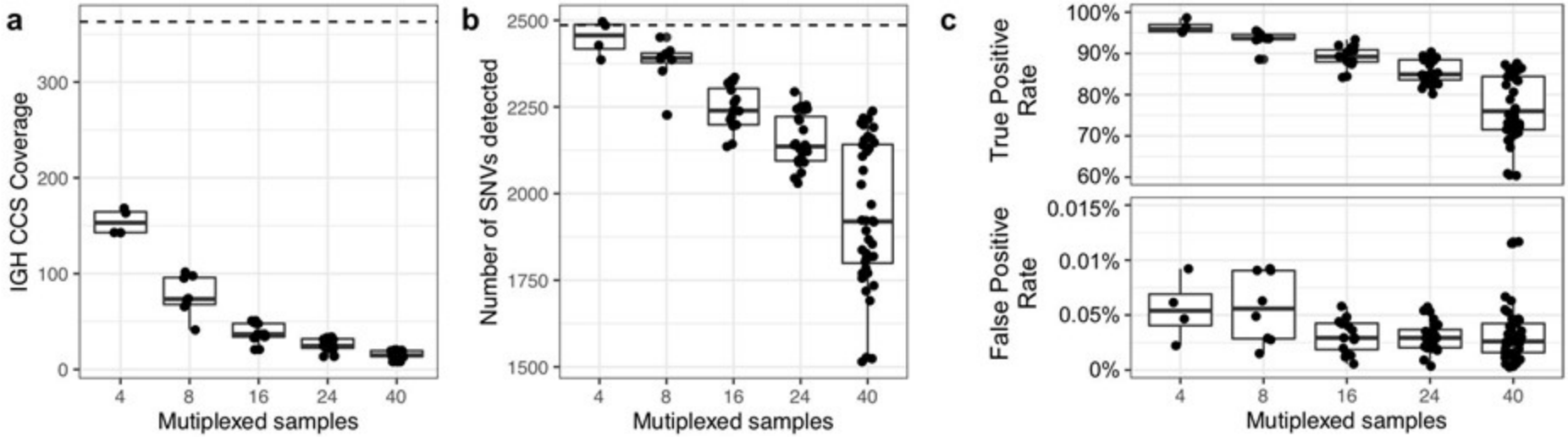
Rodriguez et al. Frontiers in Imm 2020

Less SNVs are detected when multiplexing samples



Rodriguez et al. Frontiers in Imm 2020

However, the false positive rate stays consistent



Rodriguez et al. Frontiers in Imm 2020

Output from IGenotyper

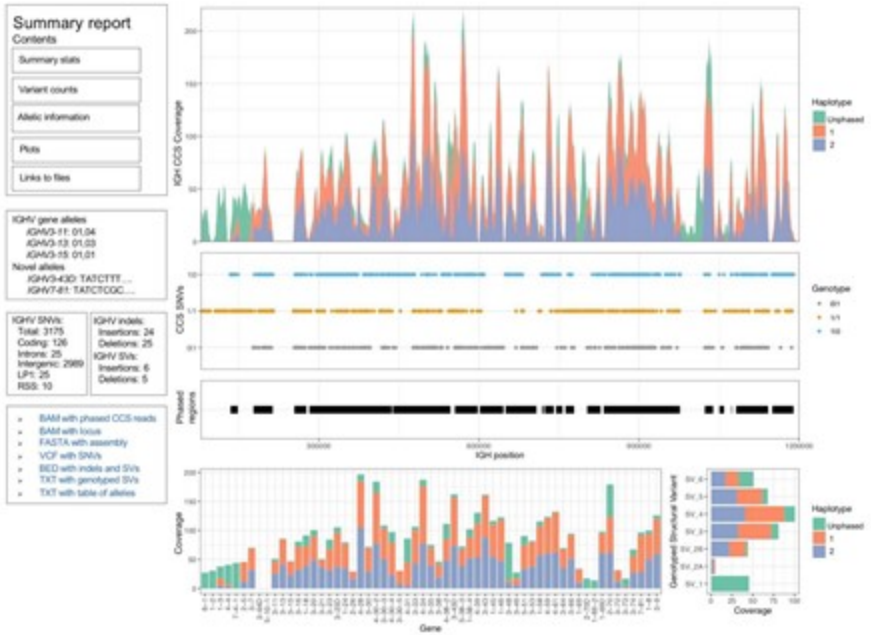
Output directories

alignments alleles assembly logs plots preprocessed report.html tmp variants

Directories	Description
`<output>/alignments`	Alignments of CCS, subreads and contigs (phased and unphased)
`<output>/assembly`	Assembly of IGH locus
`<output>/variants`	SNVs, indels and SVs
`<output>/alleles`	Alleles in sample
`<output>/logs`	Log files with input parameters
`<output>/tmp`	Temporary files. Could be deleted.

Output files

1. `alignments/`
   i. `ccs_to_ref*` : CCS reads aligned to reference
   ii. `contigs_to_ref*` : All assembled contigs aligned to refernece
   iii. `igh_contigs_to_ref*` : IGH assembled contigs aligned to igh reference
2. `assembly/`
   i. `contigs.fasta` : All assembled contigs
   ii. `igh_contigs.fasta` : IGH assembled contigs
3. `alleles/`
   i. `assembly_alleles.bed` : Alleles extracted from the assembly for each gene
   ii. `assembly_genes.fasta` : Fasta sequence from the assembly for each gene/allele
   iii. `ccs_alleles.bed` : Alleles extracted from the CCS reads for each gene
   iv. `ccs_genes.fasta` : Fasta sequence from the CCS reads for each gene/allele
4. `logs/`
   i. `gene_cov.txt` : Haplotype coverage for each gene
5. `variants/`
   i. `snvs_phased_from_ccs.vcf` : Phased SNVs detected from the CCS reads
   ii. `snvs_assembly.vcf` : SNVS detected from the assembly
   iii. `indel_assembly.bed` : Indels detected from the assembly
   iv. `sv_assembly.bed` : SVs detected from the assembly
   v. `phased_blocks.txt` : Phased haplotype blocks

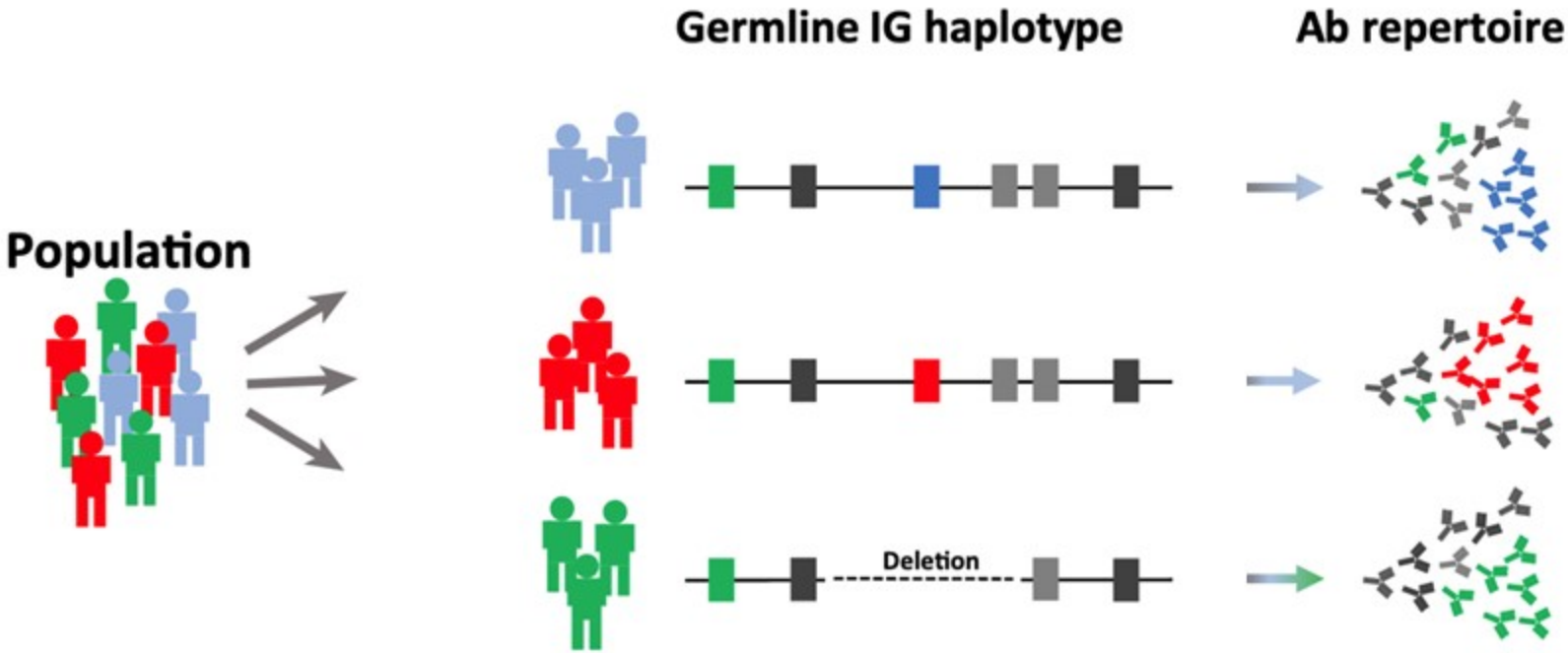


Rodriguez et al. Frontiers in Imm 2020

Outline of presentation

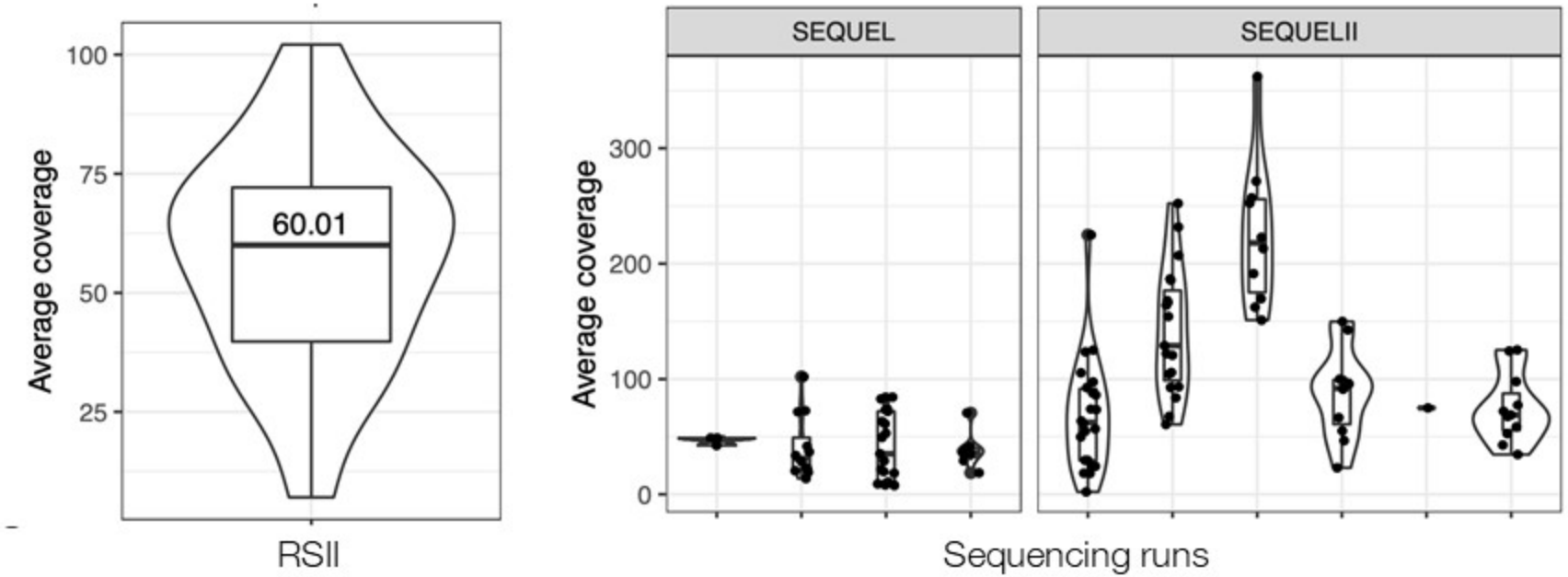
1) Introduction
    a) Genetic variation in the immunoglobulin heavy chain locus (IGH)
    b) Advantages of long-read sequencing over short-read sequencing
2) Framework for resolving the immunoglobulin heavy locus in a high throughput fashion using long read sequencing
3) Resolving the immunoglobulin heavy chain locus in a cohort with adaptive immune repertoire sequencing data
4) Resolving the T-cell receptor locus using long read sequencing
5) Application of framework to the immunoglobulin heavy chain locus in rhesus macaque
6) Conclusion and perspective

Using framework on a study with a cohort of 154 individuals with Ab repertoire data

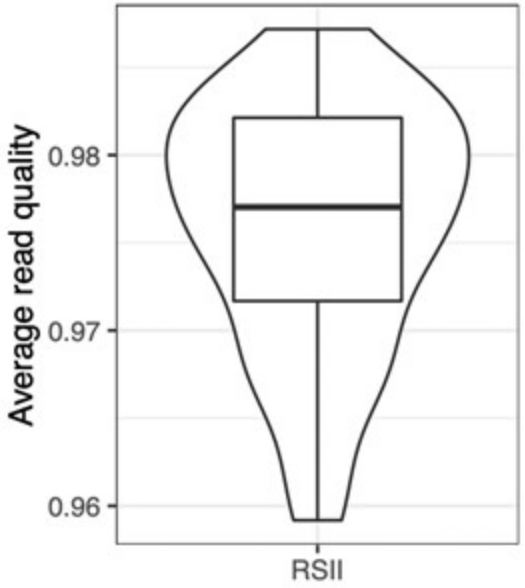


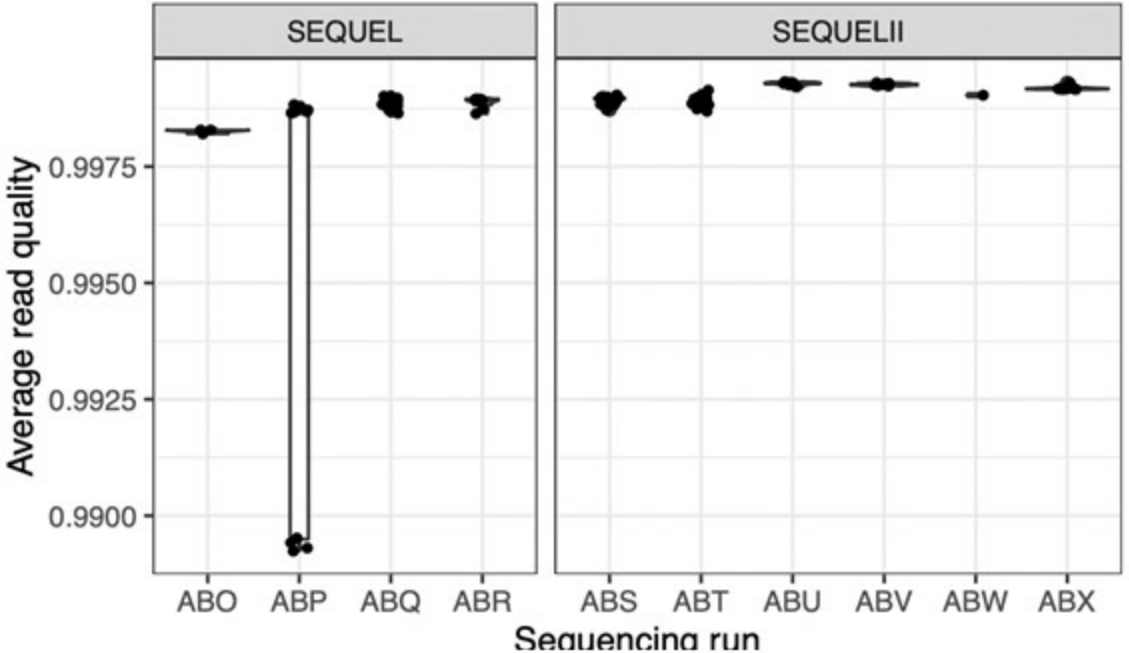
Rodriguez et al. *In preparation*

Average long-read sequencing coverage across cohort was 76X

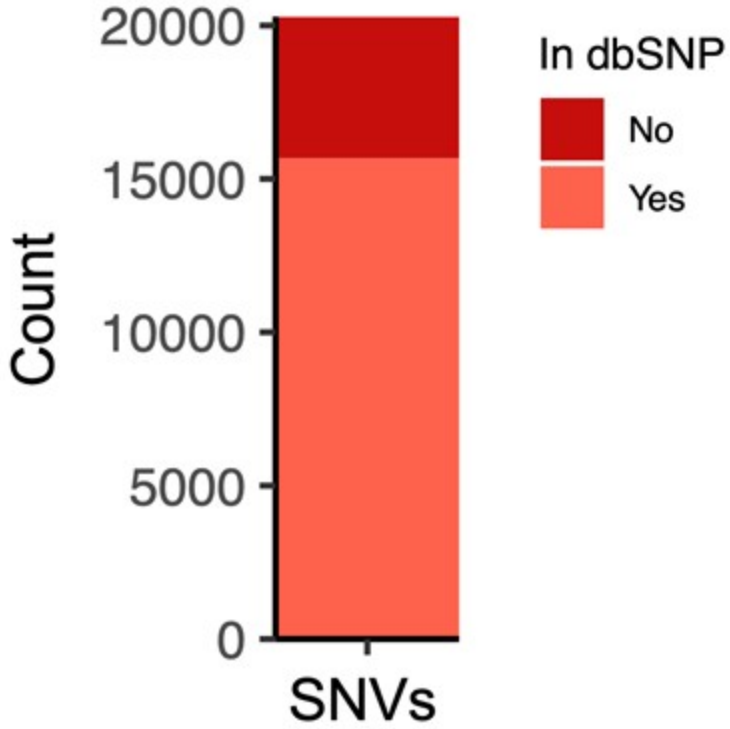


Largest improvement is observed in read quality



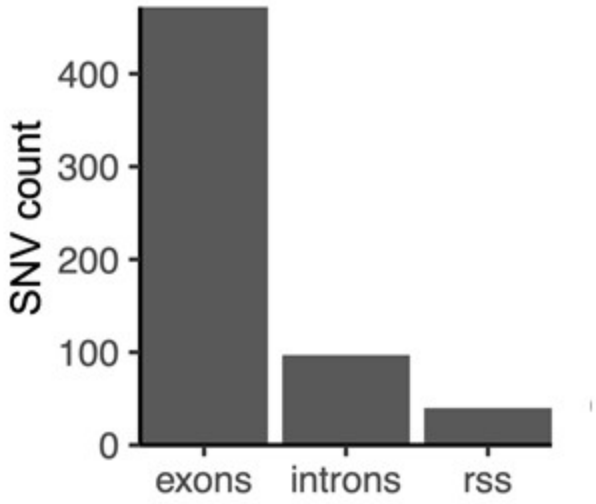


A total of 20,510 SNVs were found in one or more individuals

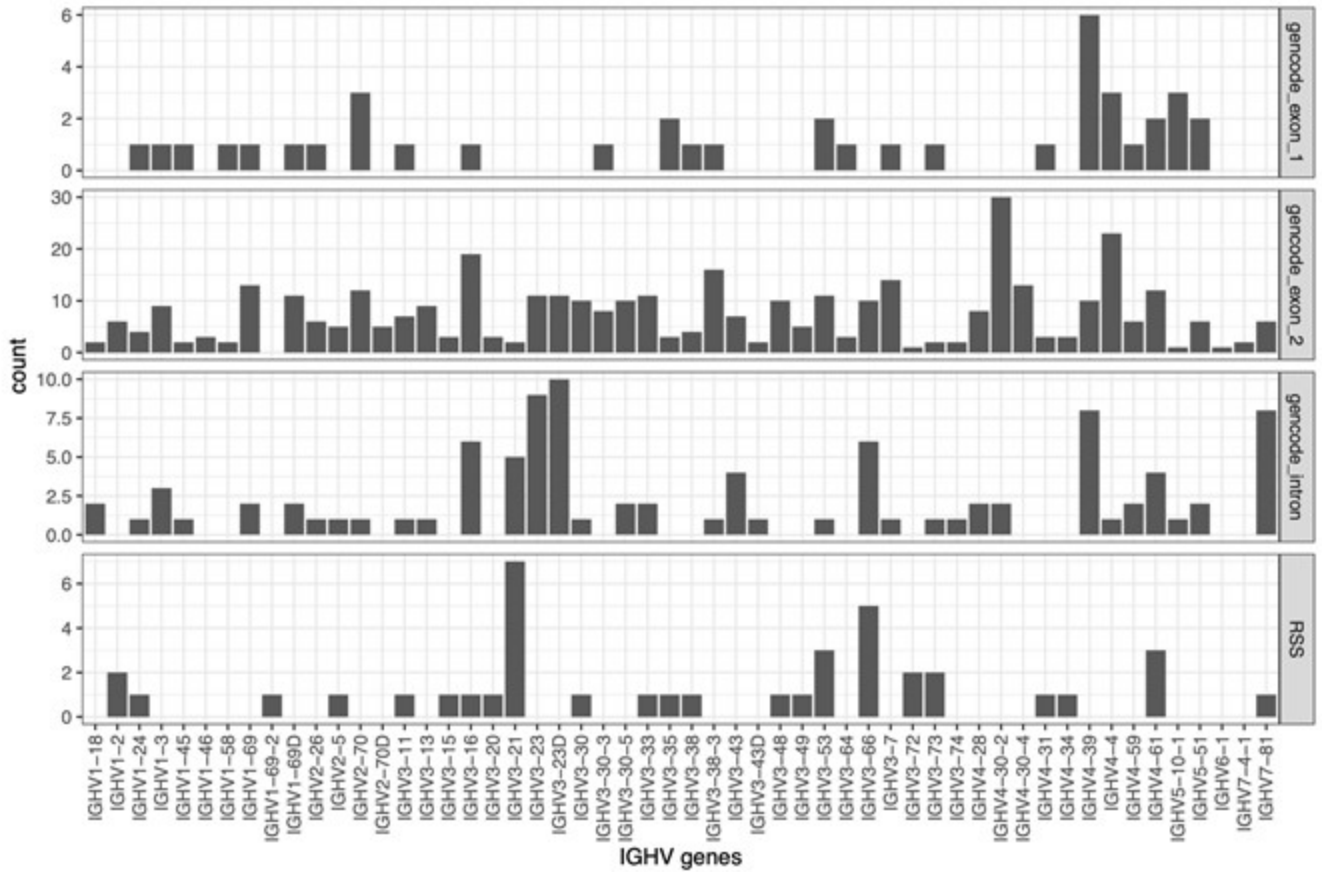


4,625 (23%) SNVs were not in the complete dbSNP release 153

SNVs are found in different gene components



SNVs are found in different gene components



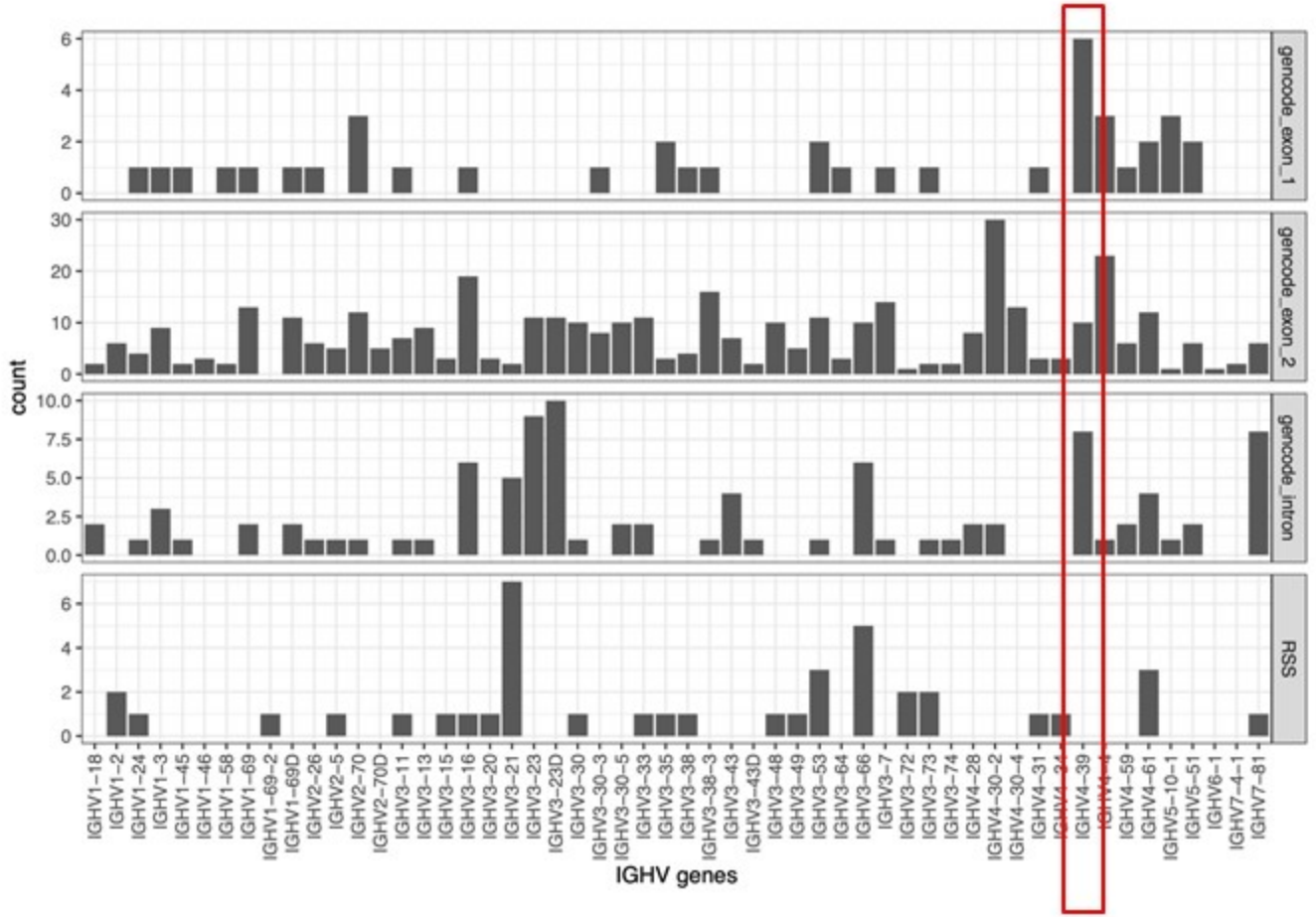
L-PART1

Gene alleles

Introns

RSS

SNVs are found in different gene components



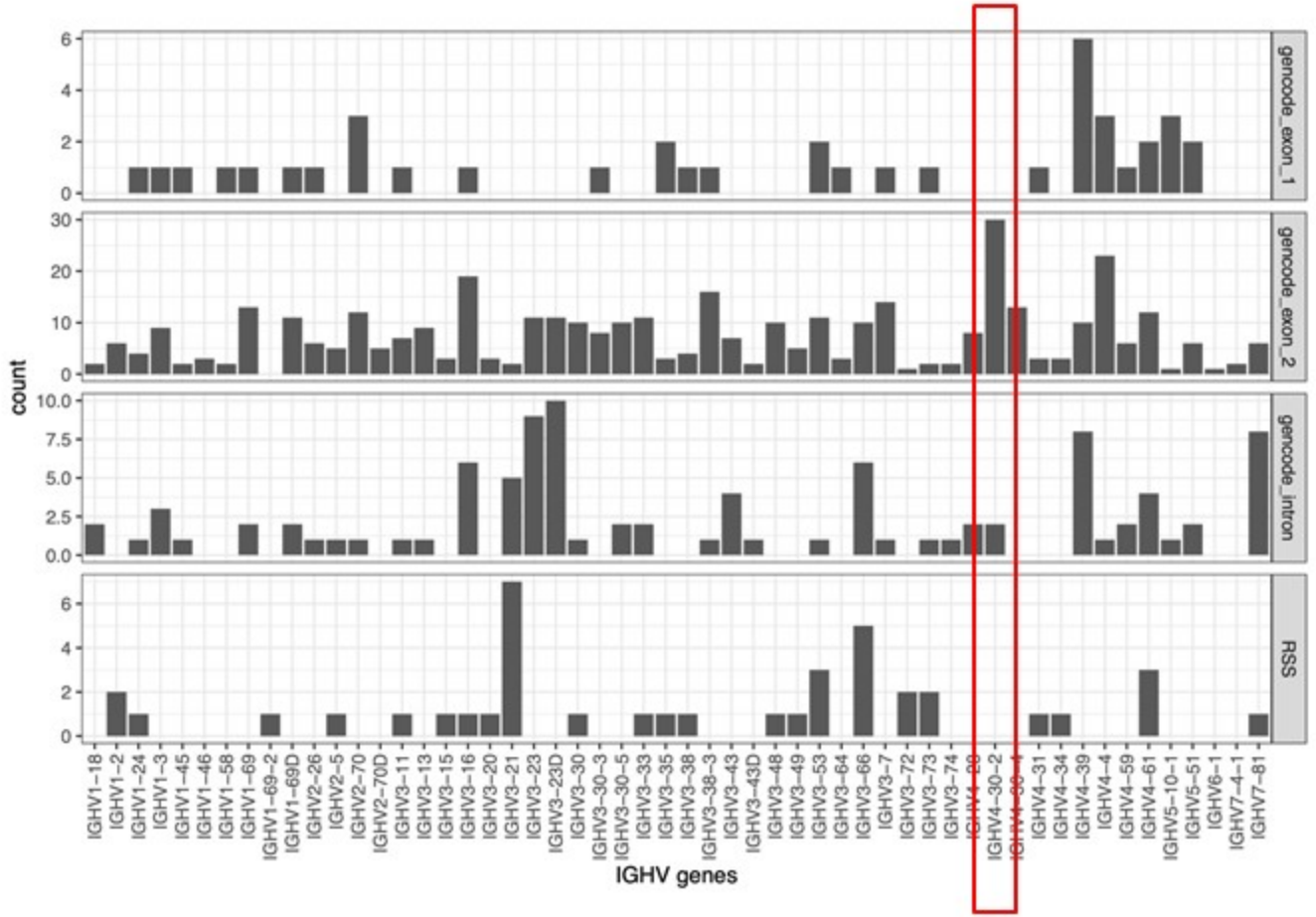
L-PART1

Gene alleles

Introns

RSS

SNVs are found in different gene components



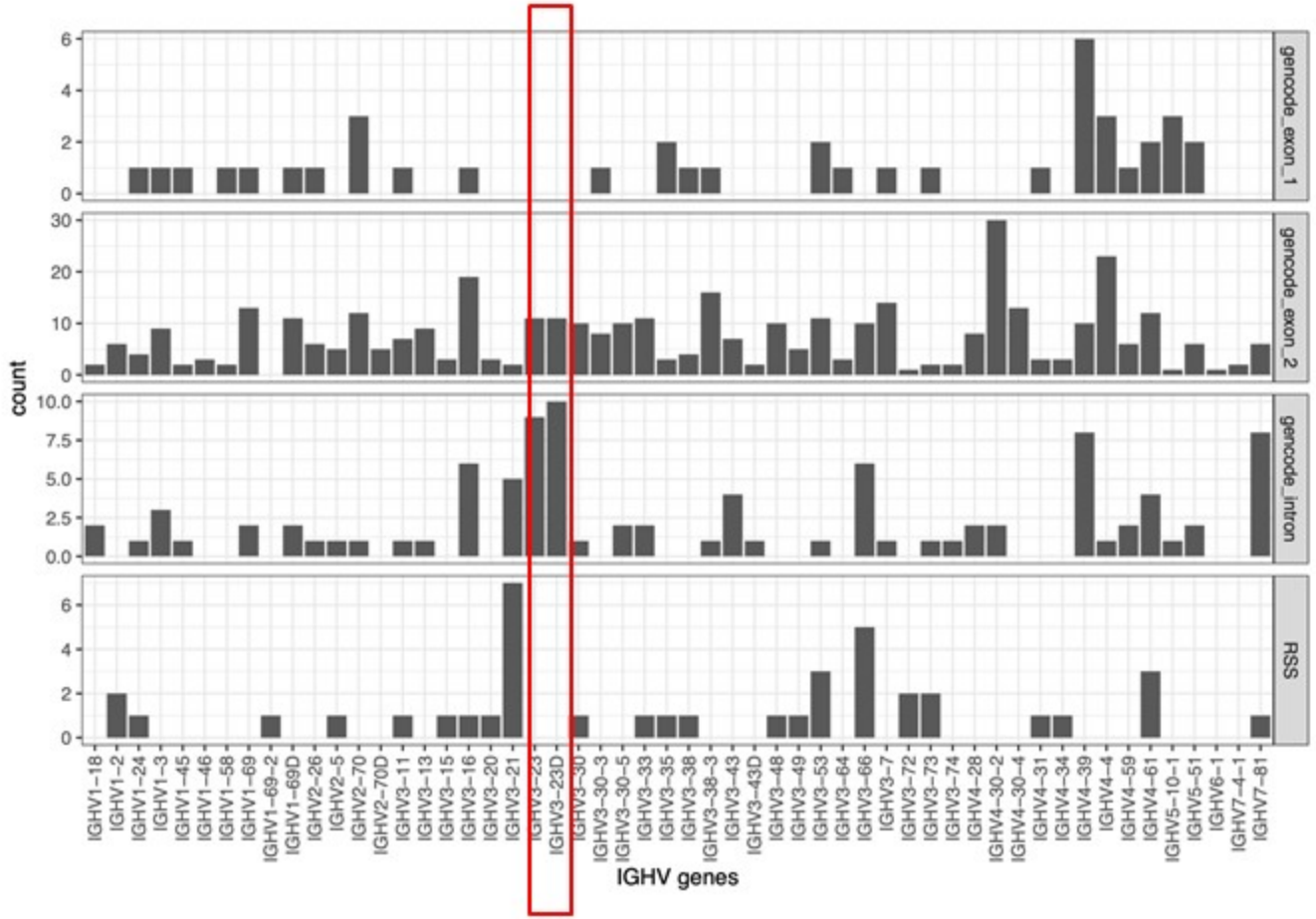
L-PART1

Gene alleles

Introns

RSS

SNVs are found in different gene components



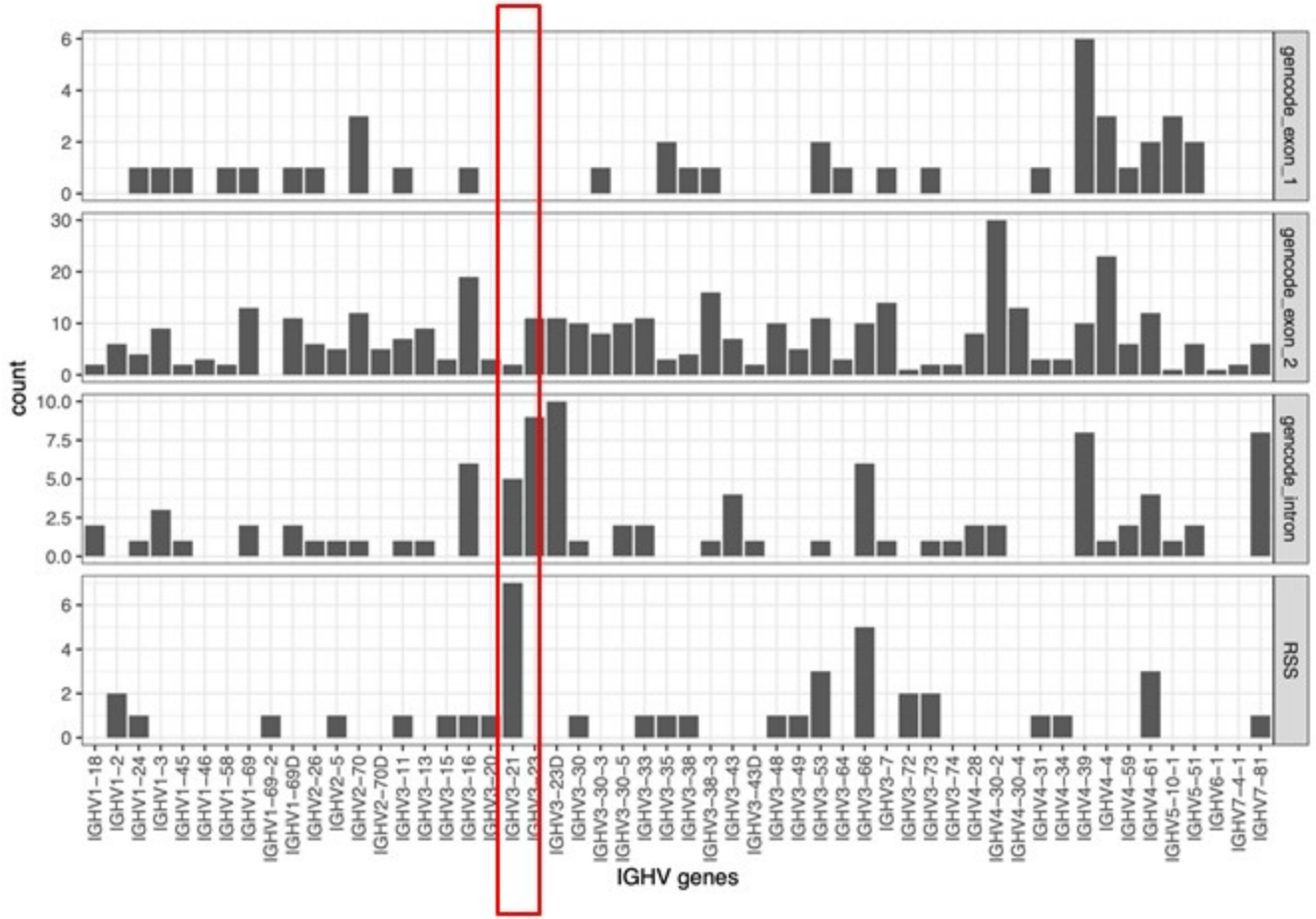
L-PART1

Gene alleles

Introns

RSS

SNVs are found in different gene components



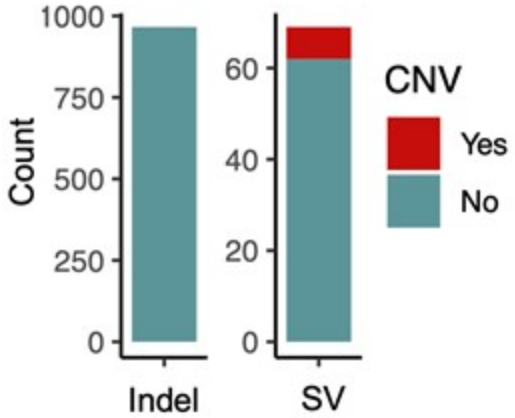
L-PART1

Gene alleles

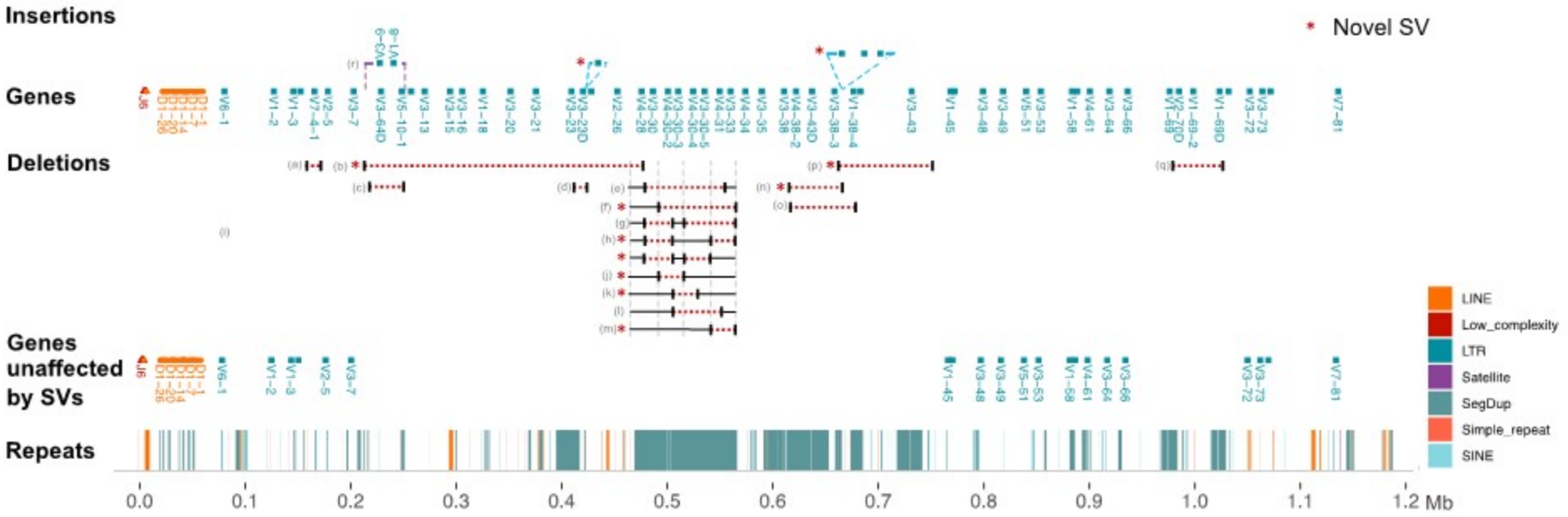
Introns

RSS

Large number of indels (966) and SVs (71) were also identified

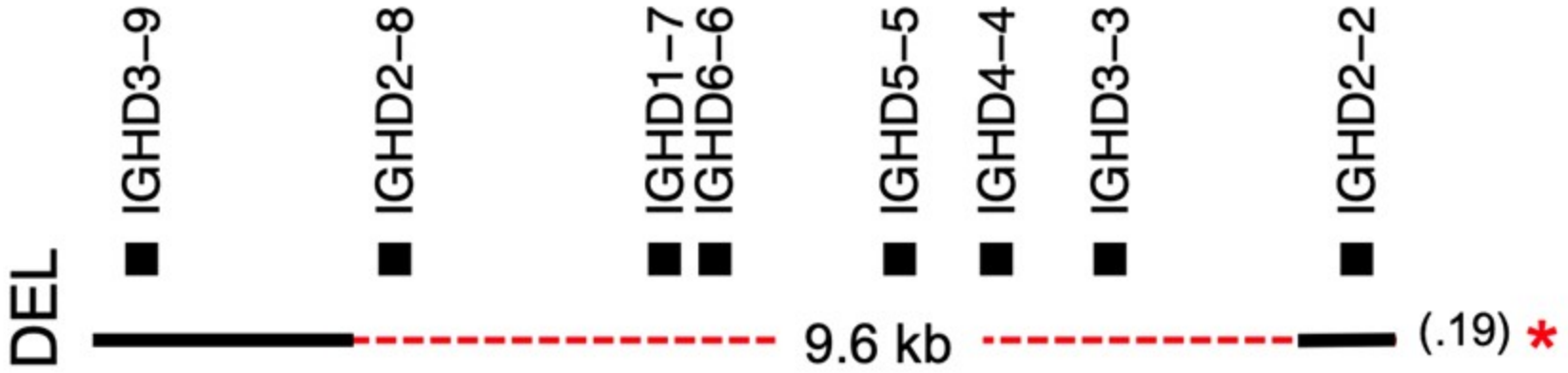


8 large SVs genotyped that affect gene copy number

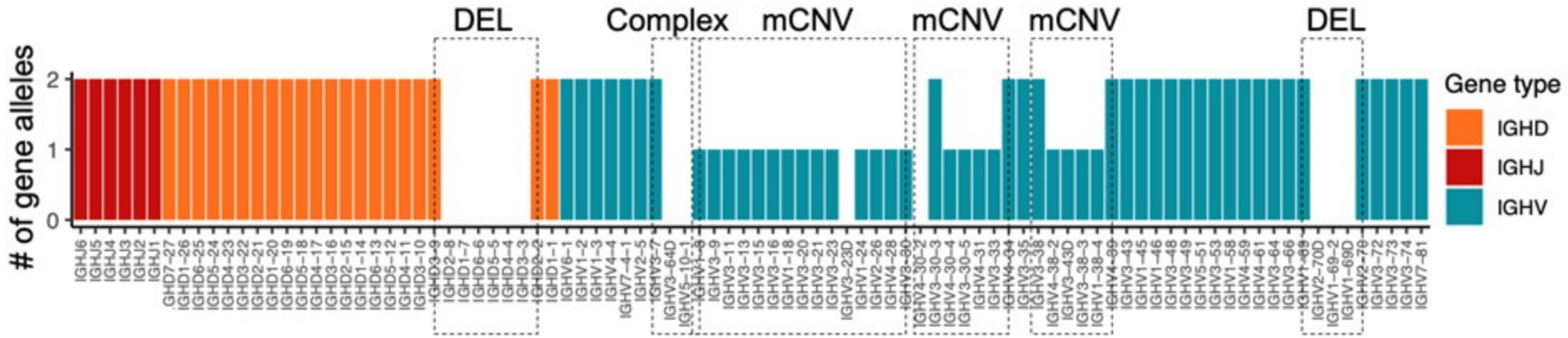


SVs affecting gene copy number were also resolved

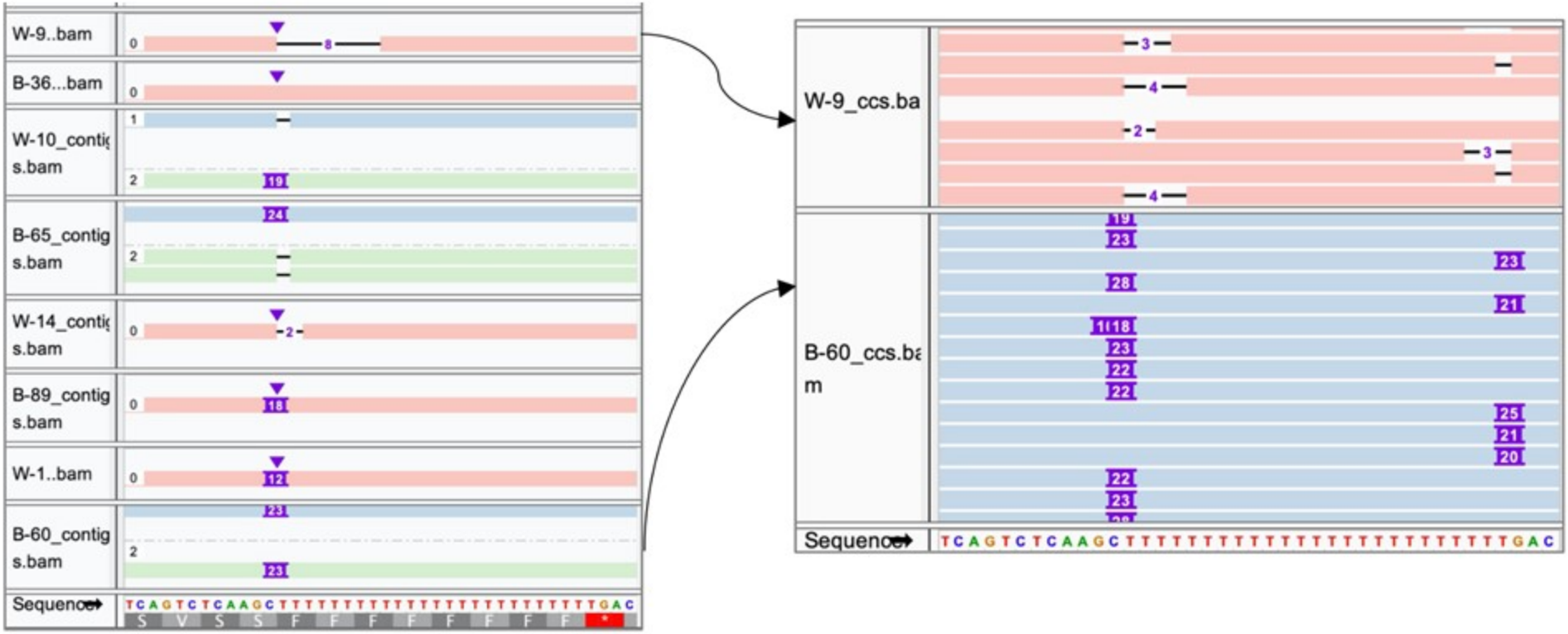
SVs affecting gene copy number were also resolved



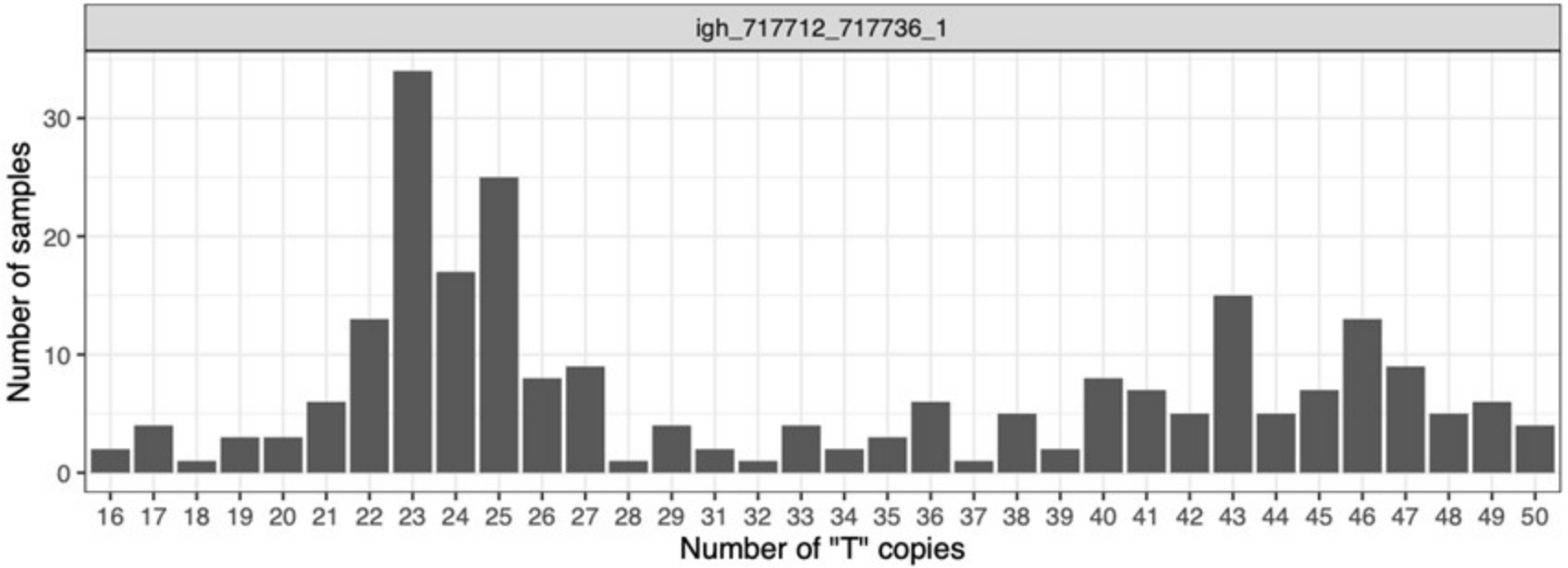
SVs affecting gene copy number were also resolved



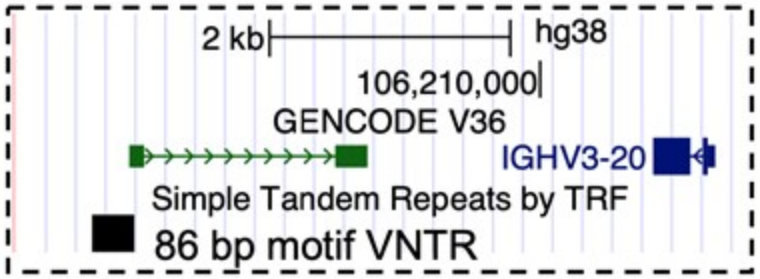
Large number of indels (966) and SVs (71) were also identified

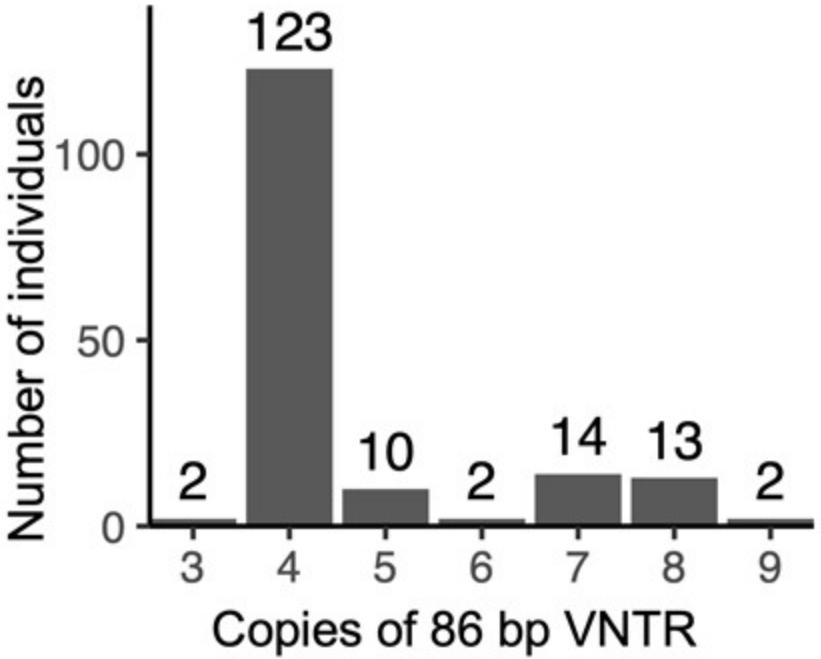


Large number of indels (966) and SVs (71) were also identified

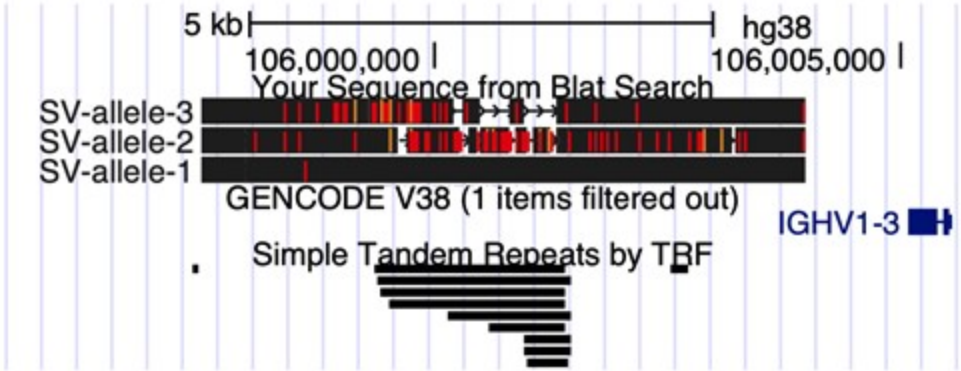


Large number of indels (966) and SVs (71) were also identified

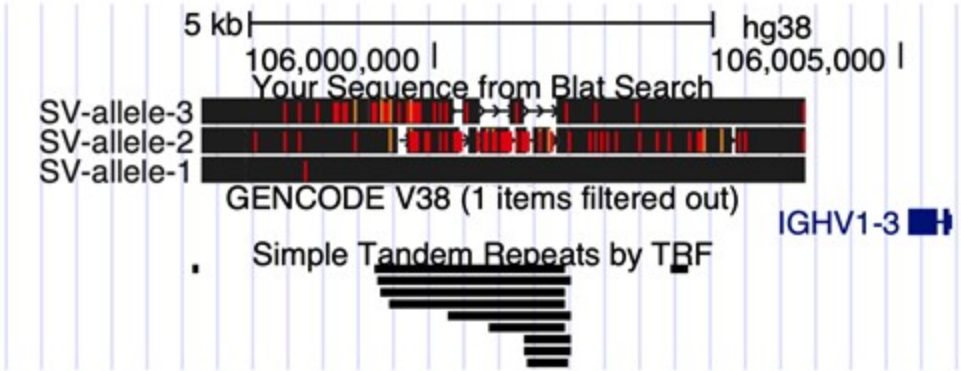




Large number of indels (966) and SVs (71) were also identified



Large number of indels (966) and SVs (71) were also identified



Consensus motif sequence

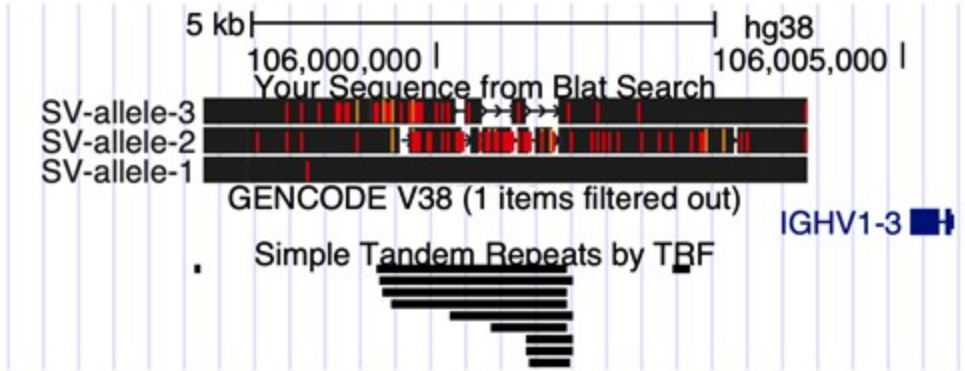
```
(3)  CTGGTGGTTC-TGAGCG-CCCC
(2)  -TGGTGGTTCCTGAGCACCCCC
(1)  -TGGTG--TCCTGAGCGCCCCC
```

Large number of indels (966) and SVs (71) were also identified

Consensus motif sequence

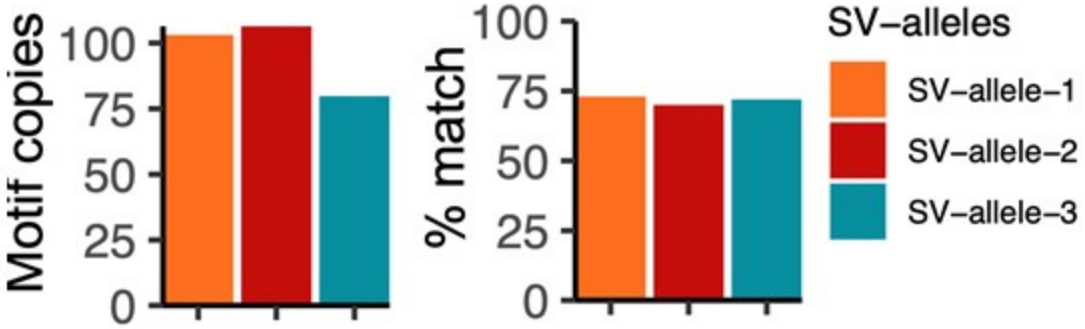
```
(3)  CTGGTGGTTC-TGAGCG-CCCC
(2)  -TGGTGGTTCCTGAGCACCCCC
(1)  -TGGTG--TCCTGAGCGCCCCC
```

Large number of indels (966) and SVs (71) were also identified

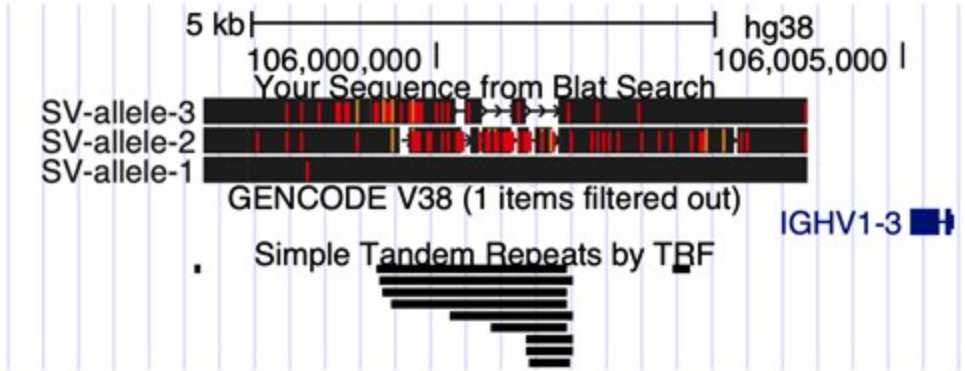


Consensus motif sequence

```
(3)  CTGGTGGTTC-TGAGCG-CCCC
(2)  -TGGTGGTTCCTGAGCACCCCC
(1)  -TGGTG--TCCTGAGCGCCCCC
```

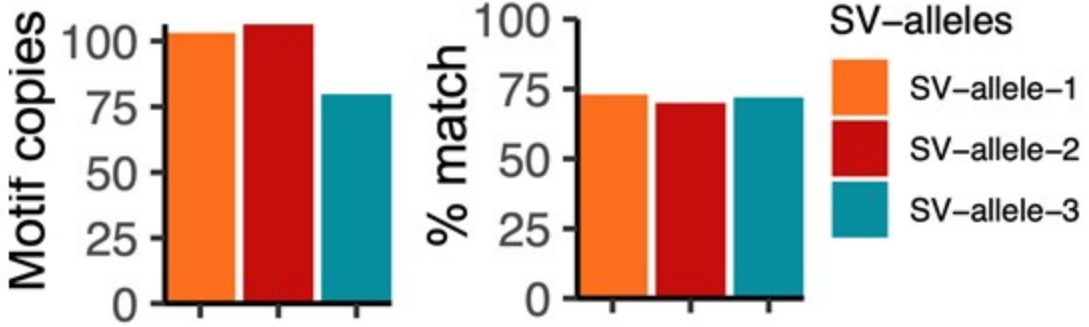


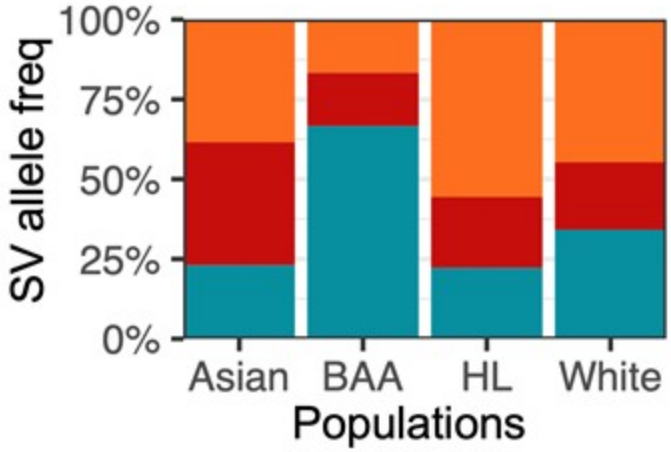
Large number of indels (966) and SVs (71) were also identified



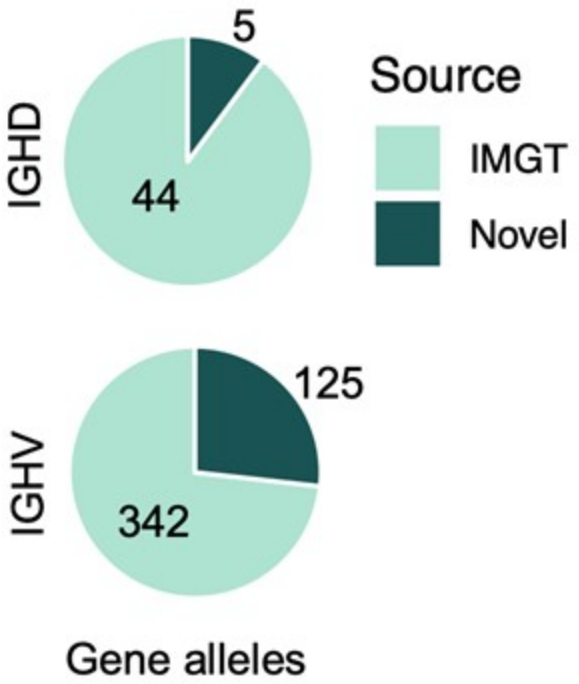
Consensus motif sequence

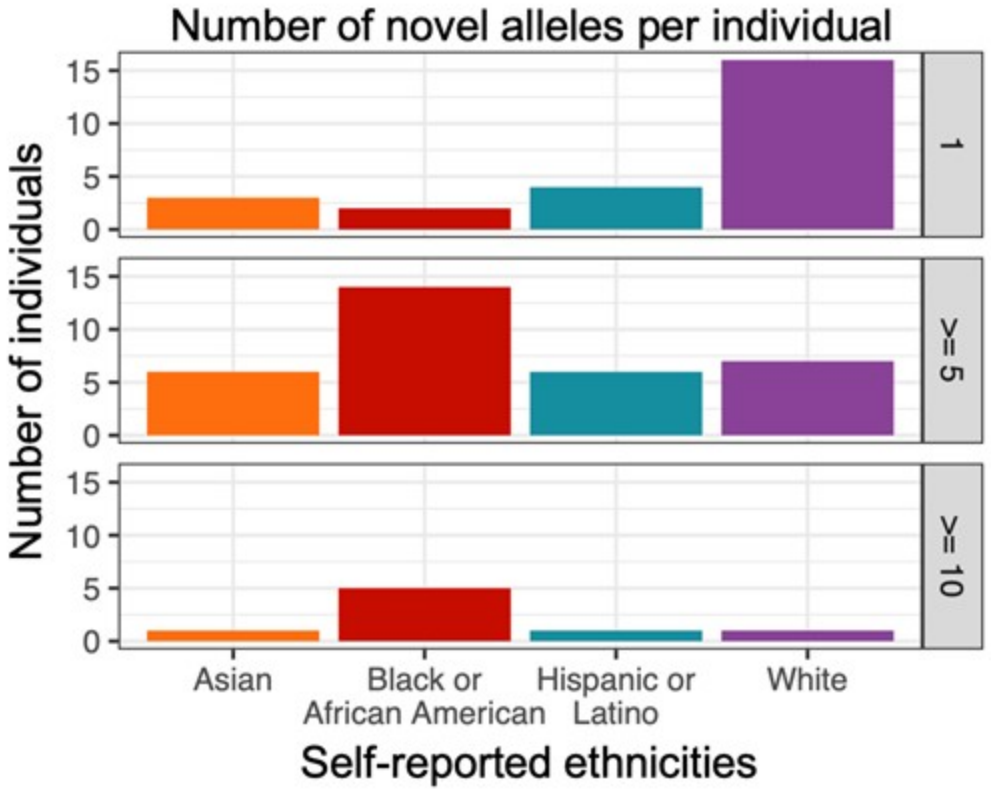
```
(3)  CTGGTGGTTC-TGAGCG-CCCC
(2)  -TGGTGGTTCCTGAGCACCCCC
(1)  -TGGTG--TCCTGAGCGCCCCC
```





A large number of novel alleles were identified

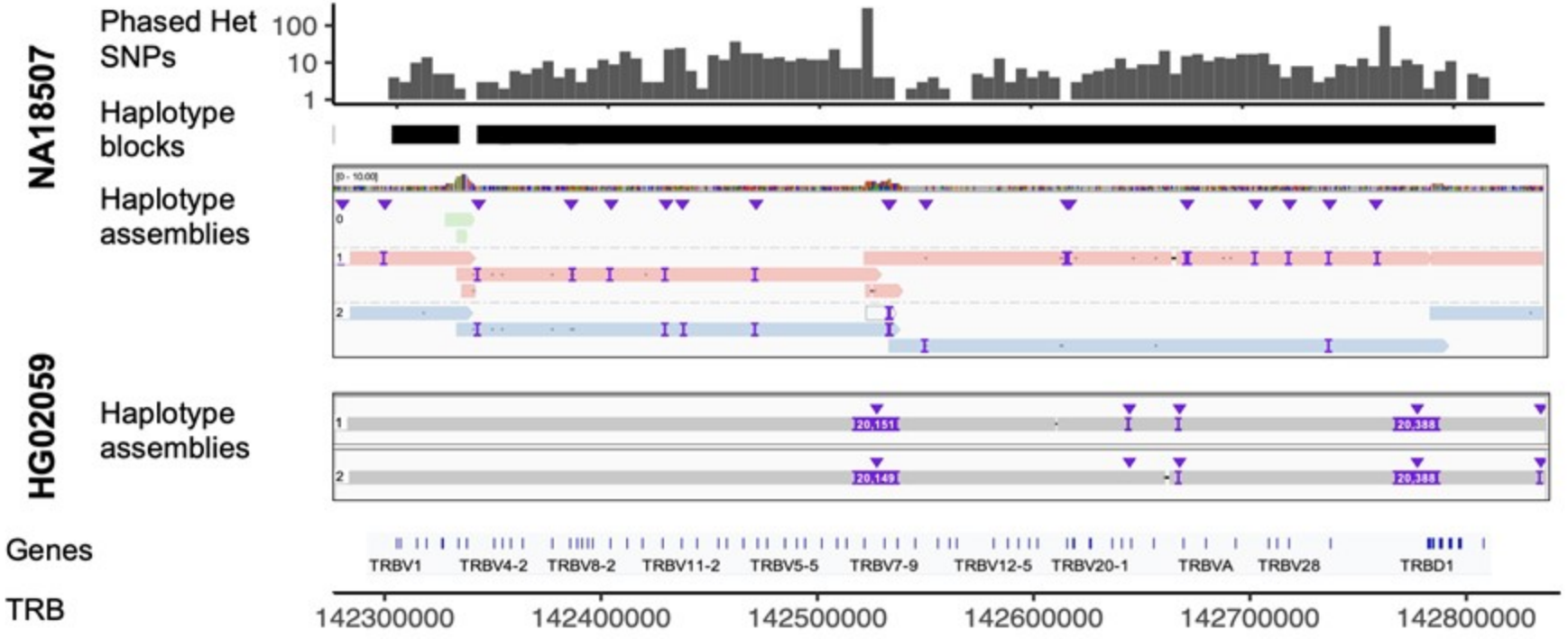




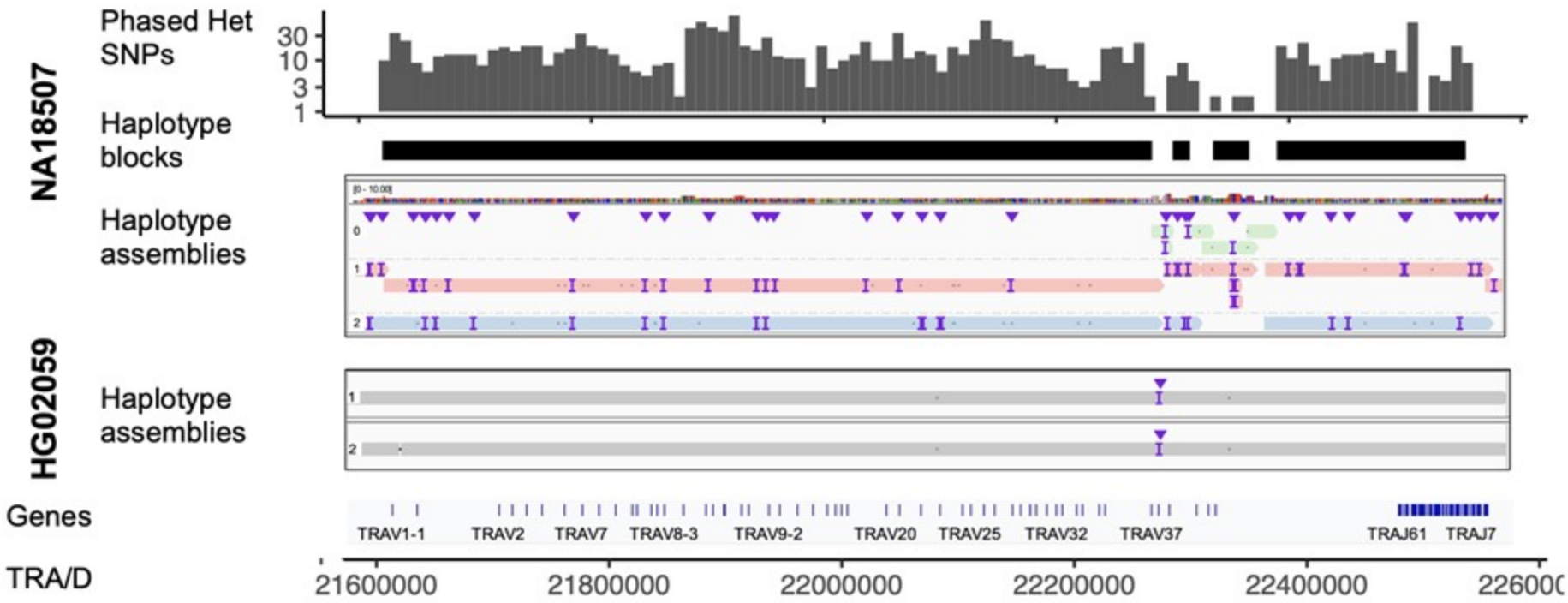
Outline of presentation

1) Introduction
    a) Genetic variation in the immunoglobulin heavy chain locus (IGH)
    b) Advantages of long-read sequencing over short-read sequencing
2) Framework for resolving the immunoglobulin heavy locus in a high throughput fashion using long read sequencing
3) Resolving the immunoglobulin heavy chain locus in a cohort with adaptive immune repertoire sequencing data
4) Resolving the T-cell receptor locus using long read sequencing
5) Application of framework to the immunoglobulin heavy chain locus in rhesus macaque
6) Conclusion

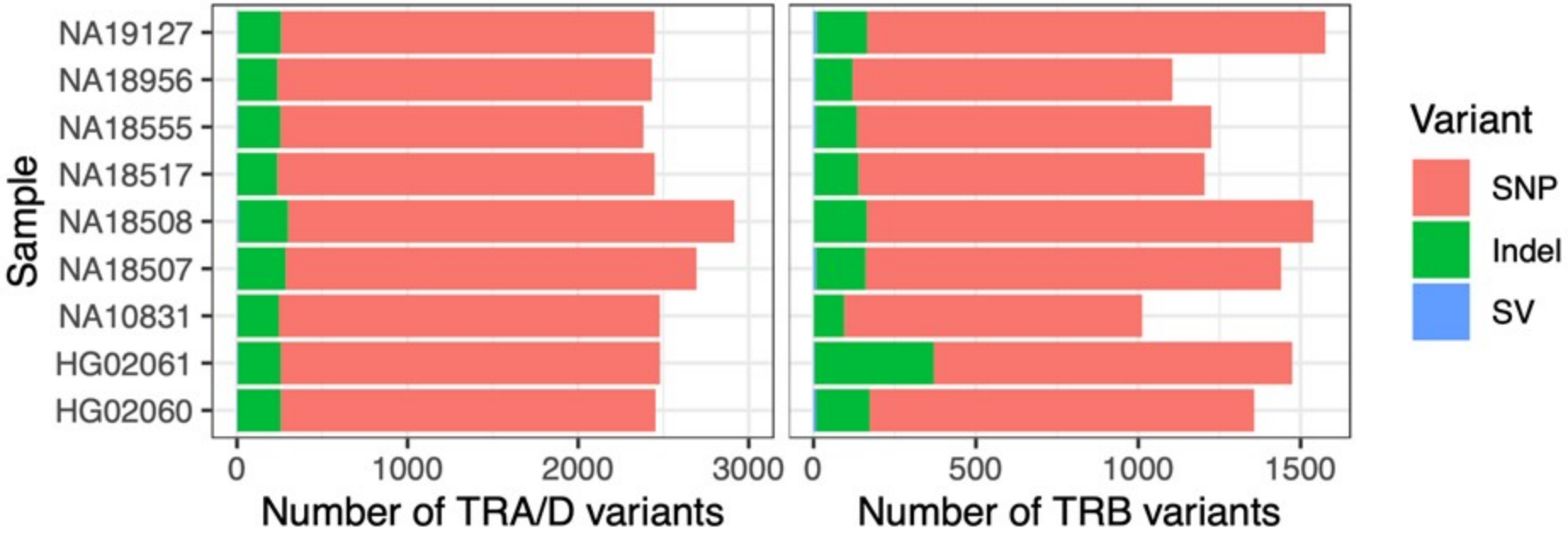
Resolving the TCR loci using long-read sequencing



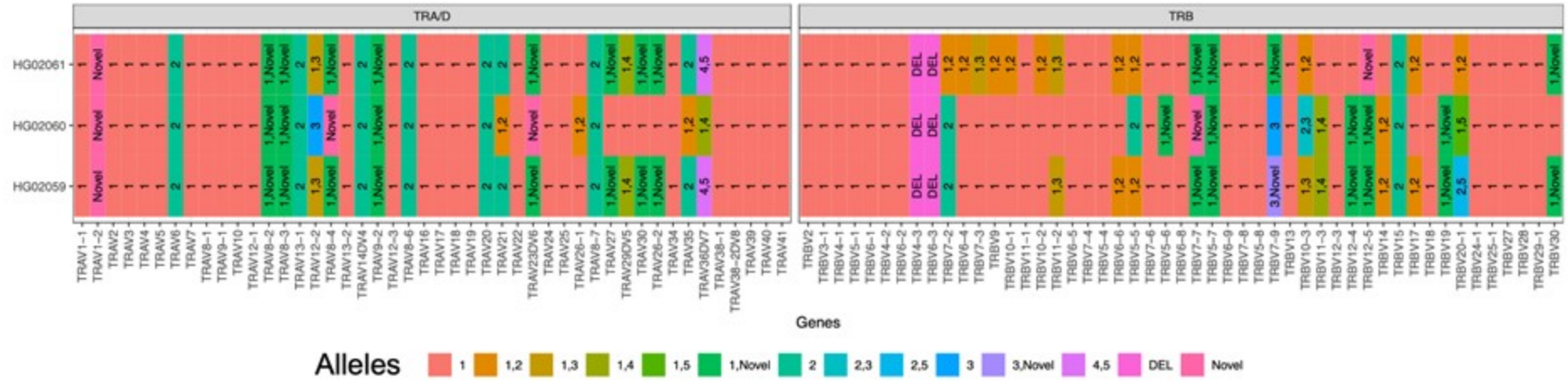
Resolving the TCR loci using long-read sequencing



SNPs, indels and SVs detected from the assemblies



Genes in TRA/D and TRB genotyped



Outline of presentation

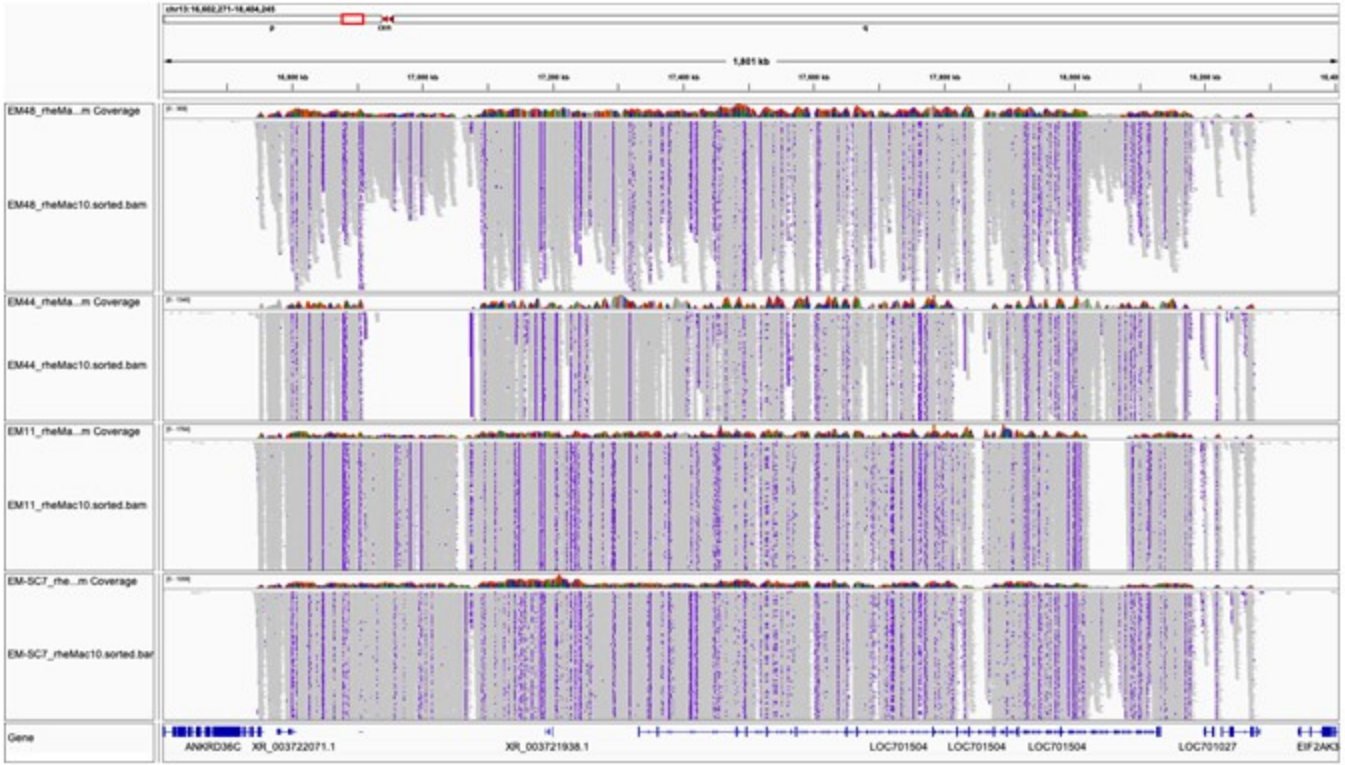
1) Introduction
    a) Genetic variation in the immunoglobulin heavy chain locus (IGH)
    b) Advantages of long-read sequencing over short-read sequencing
2) Framework for resolving the immunoglobulin heavy locus in a high throughput fashion using long read sequencing
3) Resolving the immunoglobulin heavy chain locus in a cohort with adaptive immune repertoire sequencing data
4) Resolving the T-cell receptor locus using long read sequencing
5) Application of framework to the immunoglobulin heavy chain locus in rhesus macaque
6) Conclusion

Added non-human functionality to IGenotyper

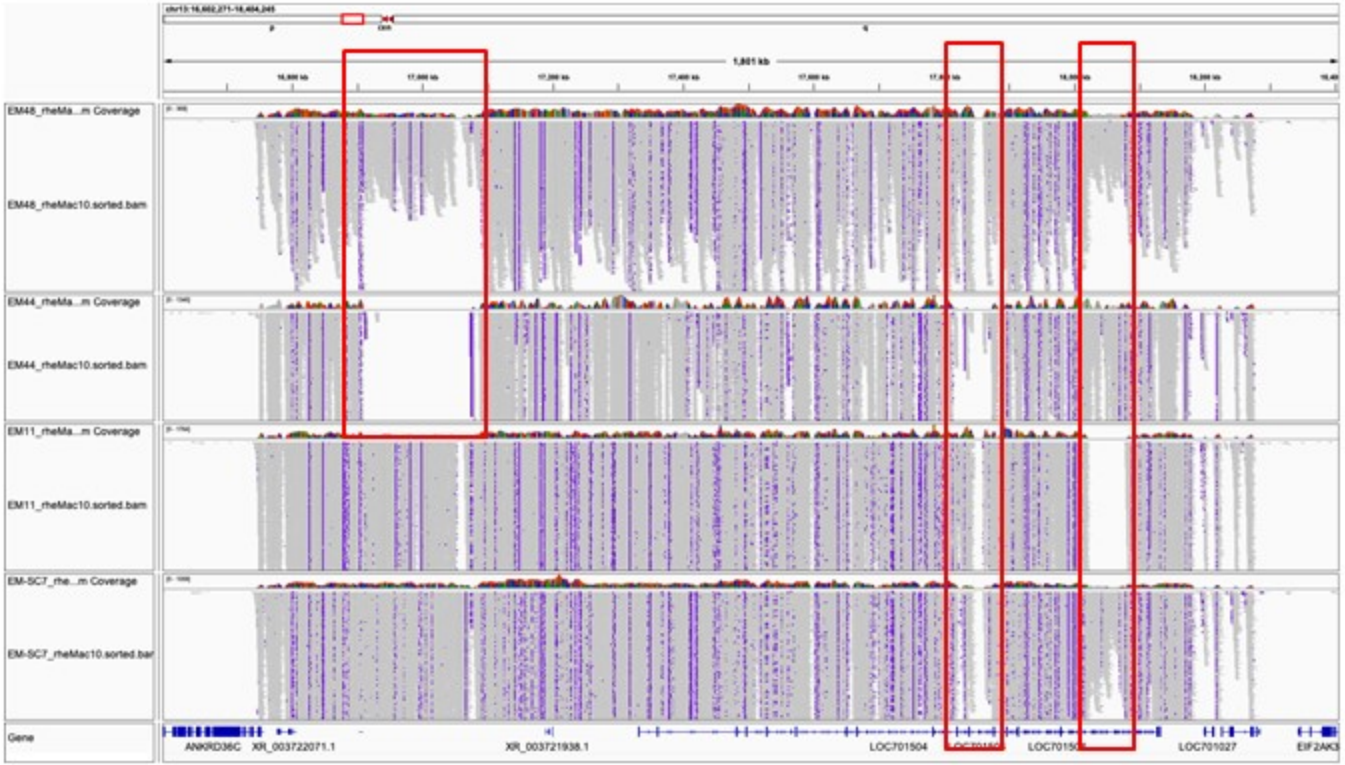
```
IG phase \
    --sample EM48 \
    --rhesus \
    demultiplex.bc1009--bc1009.bam \
    EM48_igenotyper
```

Collaboration:
Melissa Smith, PhD
Steve Bosinger, PhD

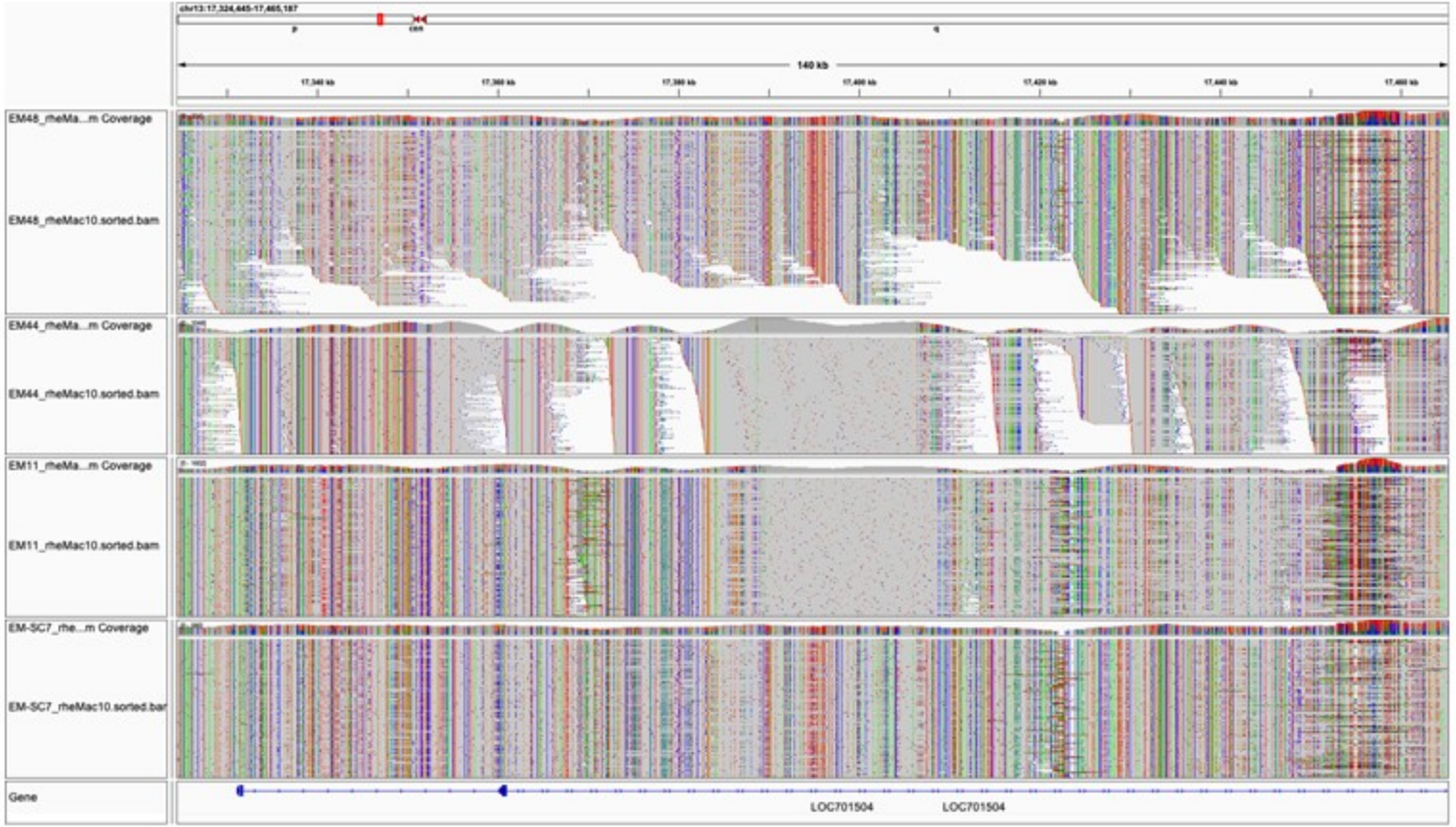
Preliminary targeted long-read sequencing coverage on 4 out 34 rhesus macaque individuals



Large structural variants detected in IGK from coverage profile



Large amount SNVs present



Conclusions

- ➔ Using a targeted approach and long read sequencing allows the resolution of the IG and TCR loci in humans and extendable to rhesus macaque
- ➔ Large amount of novel genetic variation in IGH was identified
  - ◆ Initial application to two samples identified 2 novel SVs and 16 novel alleles
  - ◆ SNVs were more accurately detected compared to NGS data
- ➔ Multiplexing allows for the application in a large cohort
- ➔ Large cohort identified even more novel genetic variation and alleles

Acknowledgments

University of Louisville
Corey Watson, PhD
Allison Silver, BS
Kaitlyn Shields, BS
William Gibson, BS

Melissa Smith, PhD

Harvard University
Wayne A. Marasco, MD, PhD

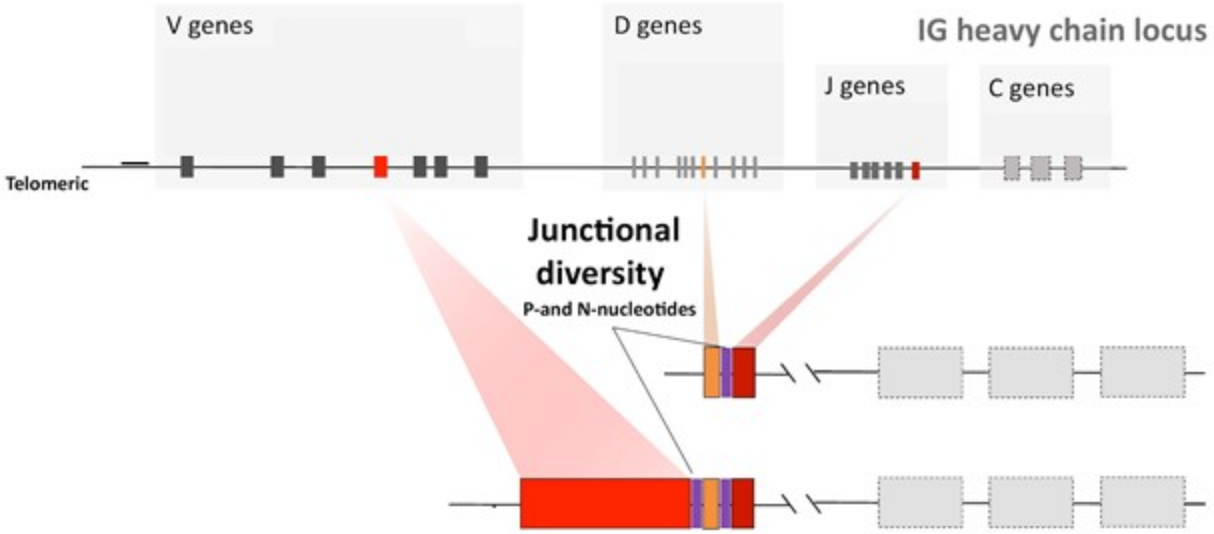
La Jolla Institute for Immunology
Shane Crotty, PhD

Emory University
Steve Bosinger, PhD
Adam Ericsen, PhD
Amit Upadhyay, PhD

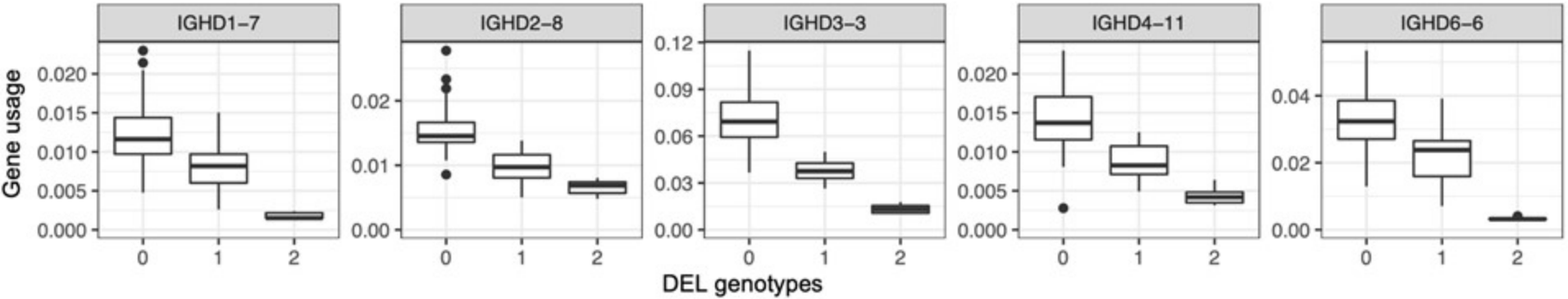
github.com/oscarlr/IGenotyper

oscarlr.github.io
louisville.edu/research/watsonlab

DNA rearranges during the development of B cells



Using Ab repertoire data we correlated gene usage with genetic genotypes



Large number of structural variants in IG

Table 1. CNVs Identified from BAC and Fosmid Clones

Individual	Population	CNV Type	IGHV Genes Included in CNV[a]	GRCh37 Outer-Start (Breakpoint)[b]	GRCh37 Outer-End (Breakpoint)[b]	Event Size (~kbp)
CH17	nd	Insertion	V1-69D, V1-f, V3-h, V2-70D (gain)	107174927	107174941	46.6
CH17	nd	Complex event	V4-30-2 (gain) V4-31 (loss)	106804332	106810878	6.5[c]/48.8[d]
CH17	nd	Complex event	V5-a, V3-64D (gain) V3-9, V1-8 (loss)	106531320	106569343	38[c]/37.7[e]
CH17	nd	Insertion	V7-4-1 (gain)	106483362	106484225	9.5
NA12156	CEPH	Deletion	V4-39, V3-38 (loss)	106866357	106899042	32.7
NA15510 and NA19240	nd and Yoruba	Insertion	V1-c, V3-d, V3-43D, V4-b (gain)	106877146	106877535	61.1
NA18555 (haplotype A)	Han Chinese	Complex event	V3-30-5, V4-30-4, V3-30-3, V4-30-2 (gain)	106804332	106804333	49.2
NA18555 (haplotype B)	Han Chinese	Deletion	V4-31, V3-30 (loss)	106786254	106811213	24.9[c]/73.9[d]
NA18507	Yoruban	Complex event[g]	V4-30-4, V3-30-3 (gain) V3-30 (loss)	106784242	nd	25.2
NA18502	Yoruban	Complex event[g]	V3-30-5 (gain) V3-33, V4-31 (loss)	nd	106820685	24.7
NA18956 and NA12156	Japanese and CEPH	Duplication	V3-23D (gain)	106716650	106727861	10.8
NA19240 and NA12878	Yoruban and CEPH	Insertion	V7-4-1 (gain)	106483362	106484225	9.5

PICTURE

Watson et al. AJHG 2013

Testing capture plus IGenotyper on sample from Rheumatic heart disease GWAS study

ARTICLE

Received 15 Sep 2016 | Accepted 15 Feb 2017 | Published 11 May 2017

DOI: 10.1038/ncomms14946 OPEN

Association between a common immunoglobulin heavy chain allele and rheumatic heart disease risk in Oceania

Tom Parks[1], Mariana M. Mirabel[2], Joseph Kado[3,4], Kathryn Auckland[1], Jaroslaw Nowak[5], Anna Rautanen[1], Alexander J. Mentzer[1], Eloi Marijon[2,6], Xavier Jouven[2,6], Mai Ling Perman[4], Tuliana Cua[7], John K. Kauwe[8], John B. Allen[8], Henry Taylor[9], Kathryn J. Robson[10], Charlotte M. Deane[5], Andrew C. Steer[11,12,*], Adrian V.S. Hill[1,*] & for the Pacific Islands Rheumatic Heart Disease Genetics Network[†]

GWAS SNVs were a combination of array, short-read data and imputed SNVs

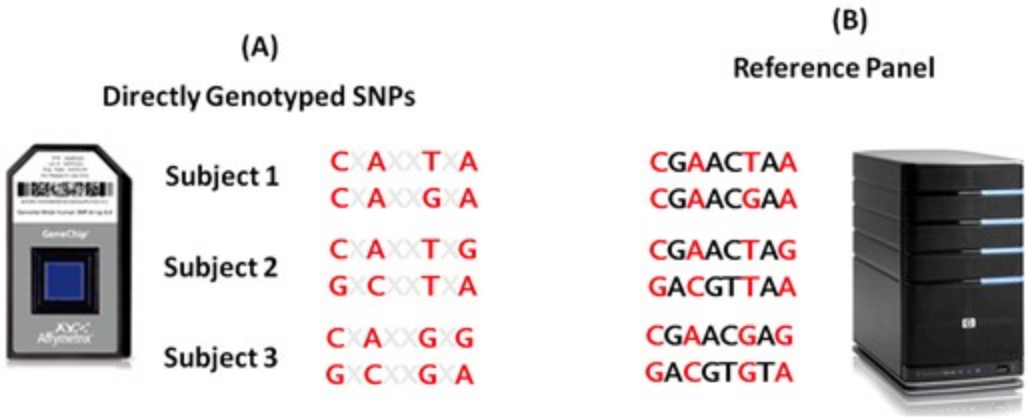
(A)

Directly Genotyped SNPs

Subject 1	C A T A	
	C A G A	
Subject 2	C A T G	
	G C T A	
Subject 3	C A G G	
	G C G A	

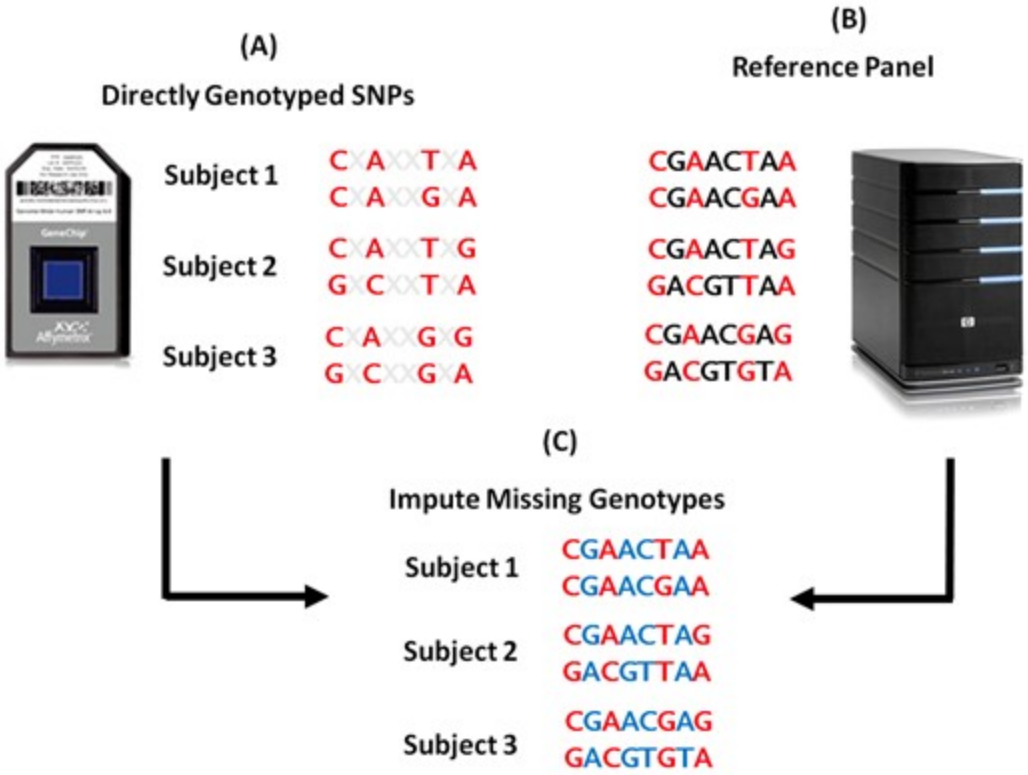
Wood et al. PloS 2013

GWAS SNVs were a combination of array, short-read data and imputed SNVs



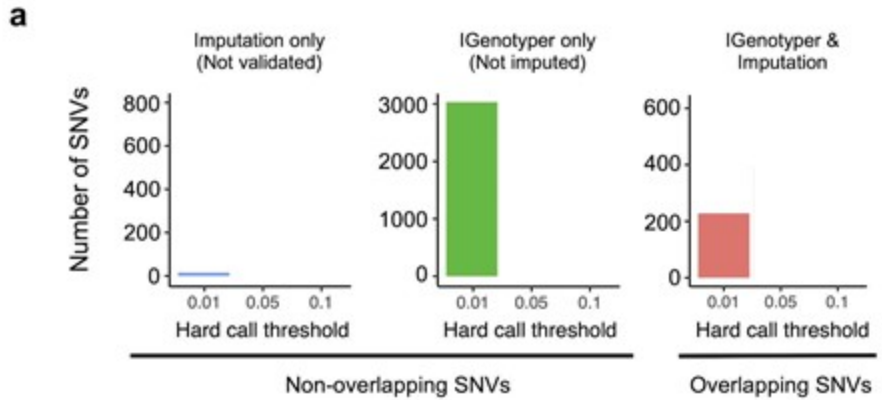
Wood et al. PloS 2013

GWAS SNVs were a combination of array, short-read data and imputed SNVs

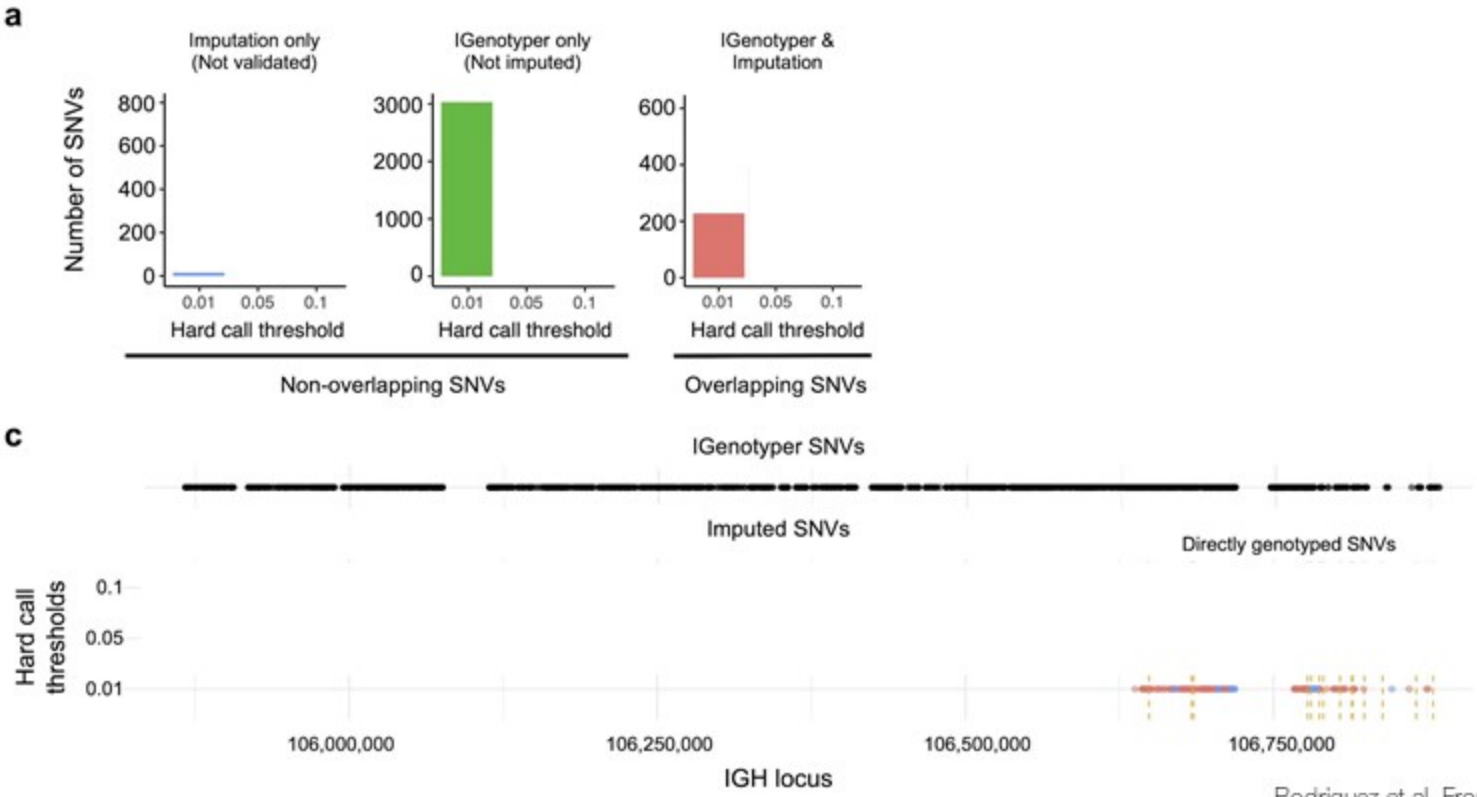


Wood et al. PloS 2013

Reducing imputation threshold to allow more imputed SNVs increased number of FP SNV

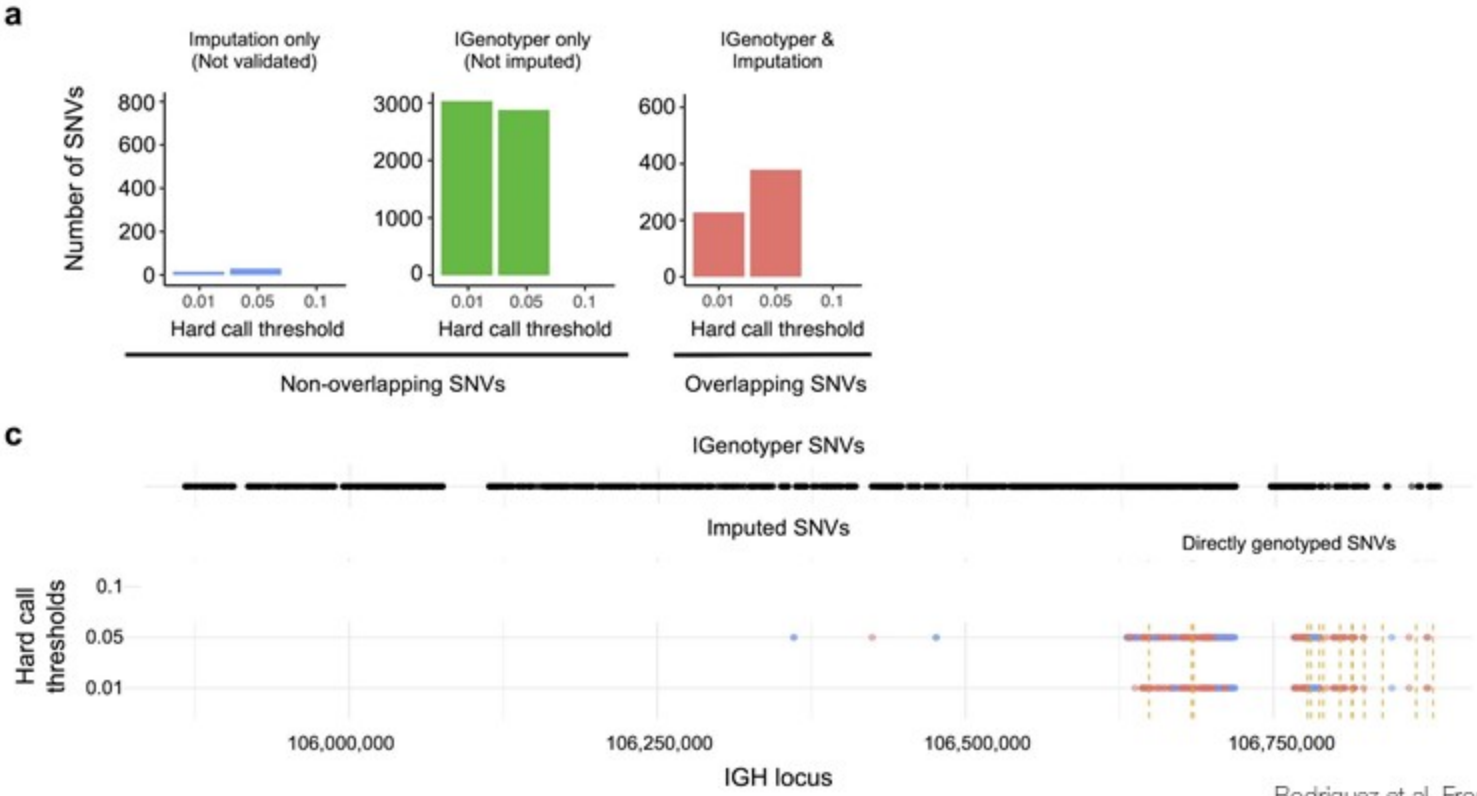


Reducing imputation threshold to allow more imputed SNVs increased number of FP SNV



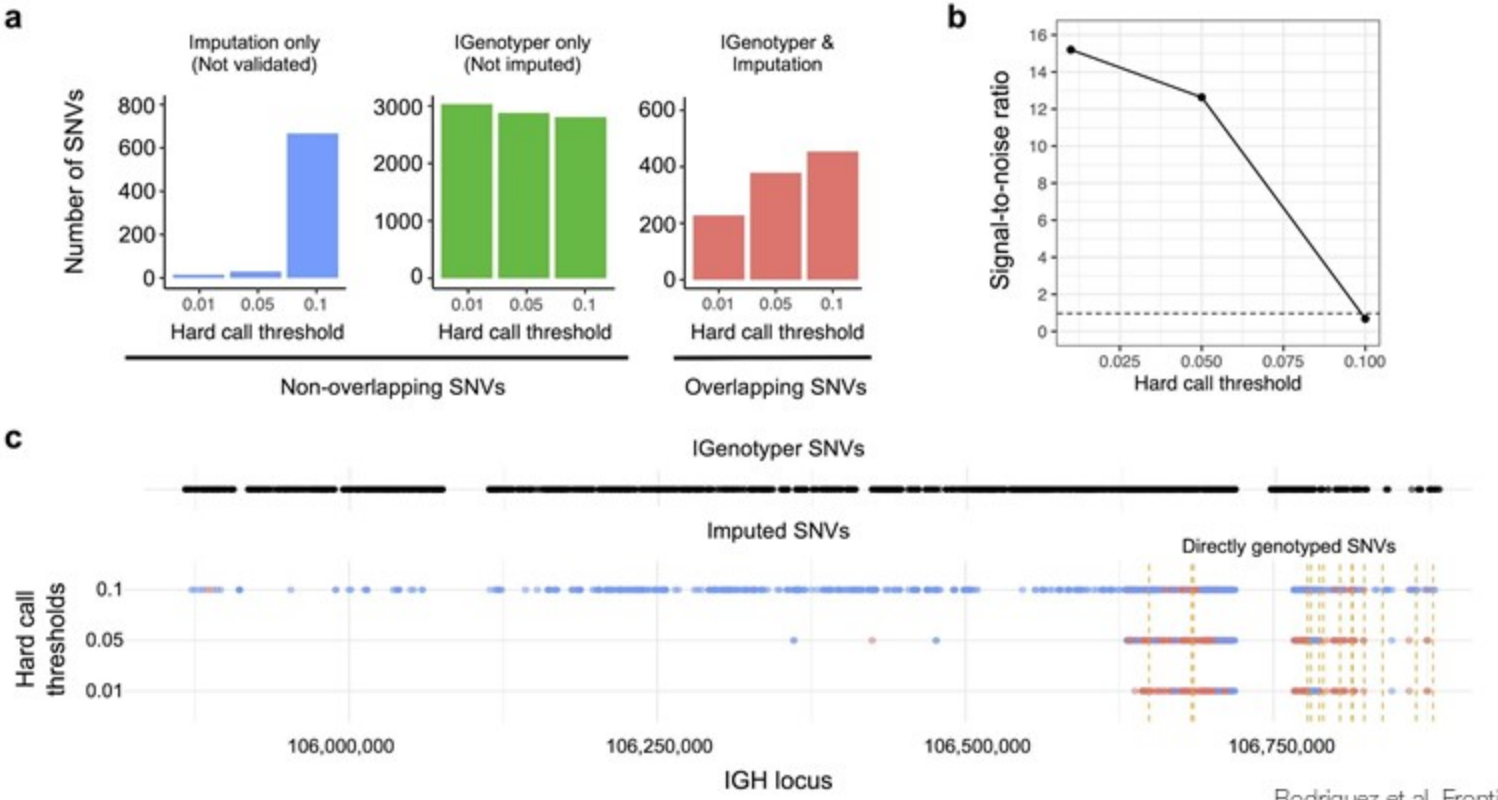
Rodriguez et al. Frontiers in Imm 2020

Reducing imputation threshold to allow more imputed SNVs increased number of FP SNV



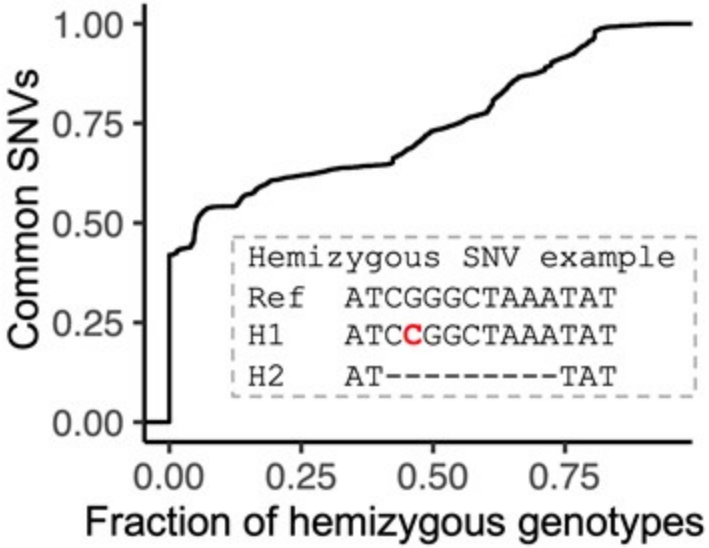
Rodriguez et al. Frontiers in Imm 2020

Reducing imputation threshold to allow more imputed SNVs increased number of FP SNV

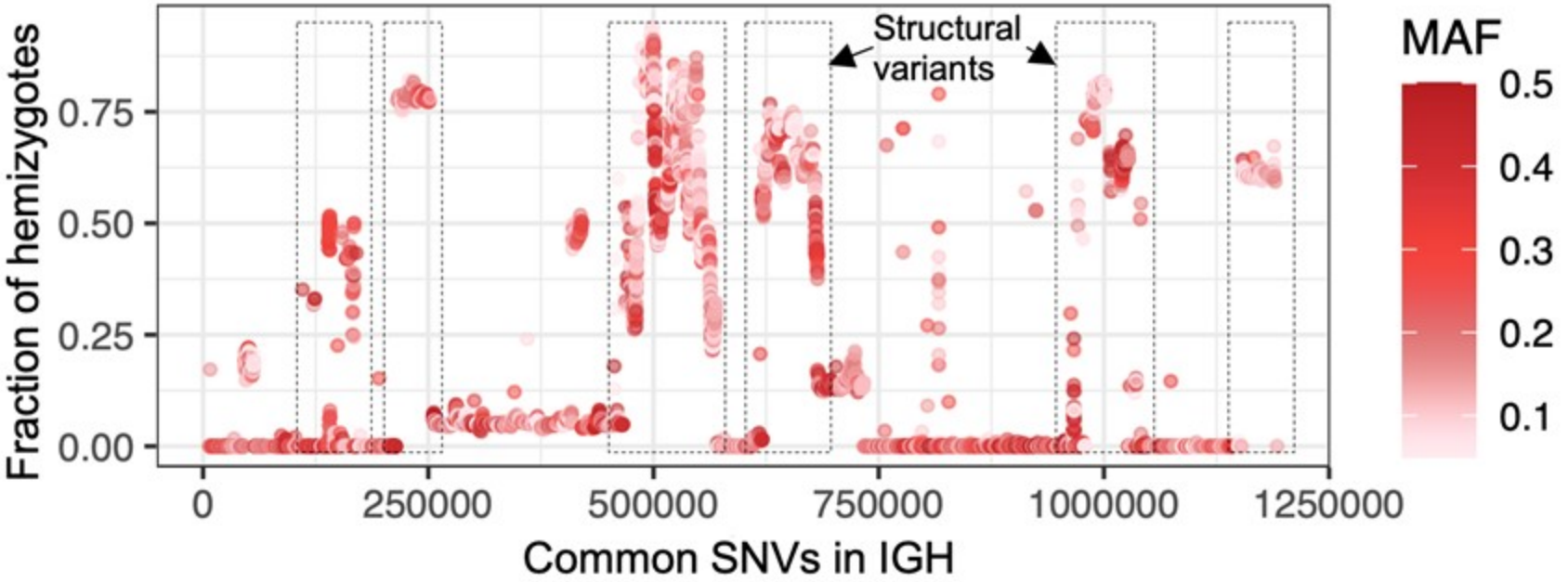


Rodriguez et al. Frontiers in Imm 2020

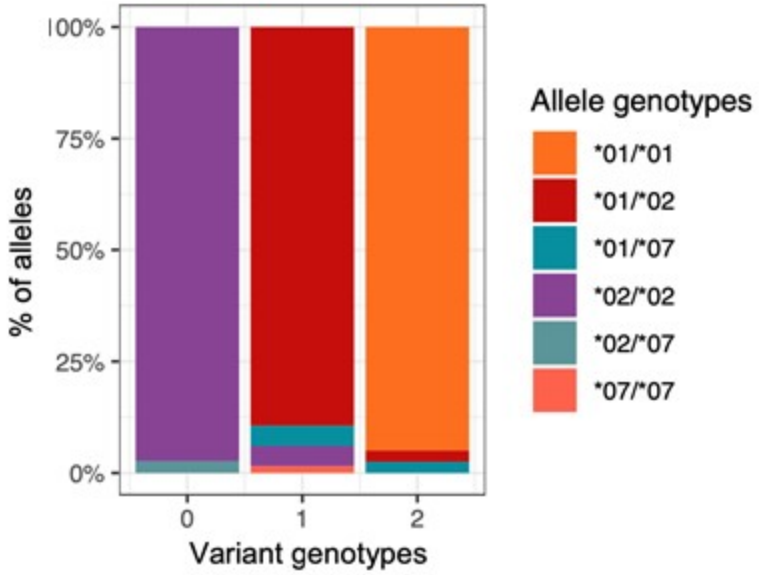
Large amount of hemizygous genotypes



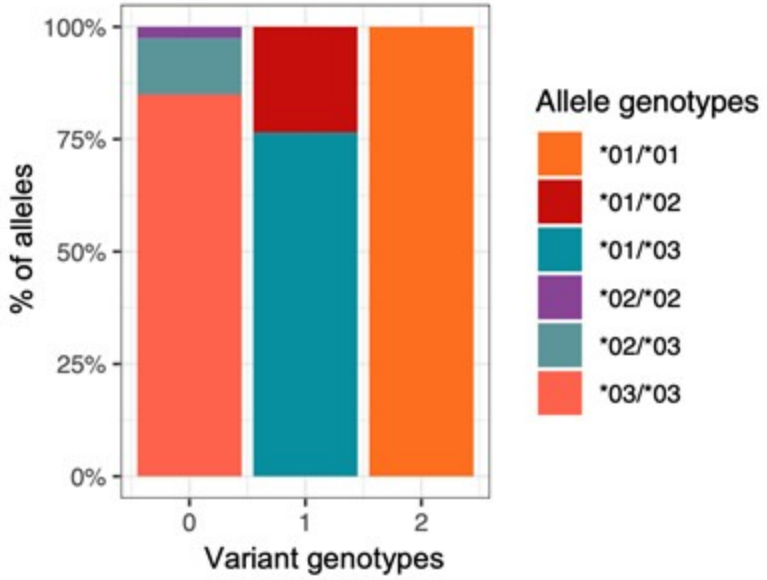
Large amount of hemizygous genotypes



Variants associated with gene usage contained different alleles across genotype groups

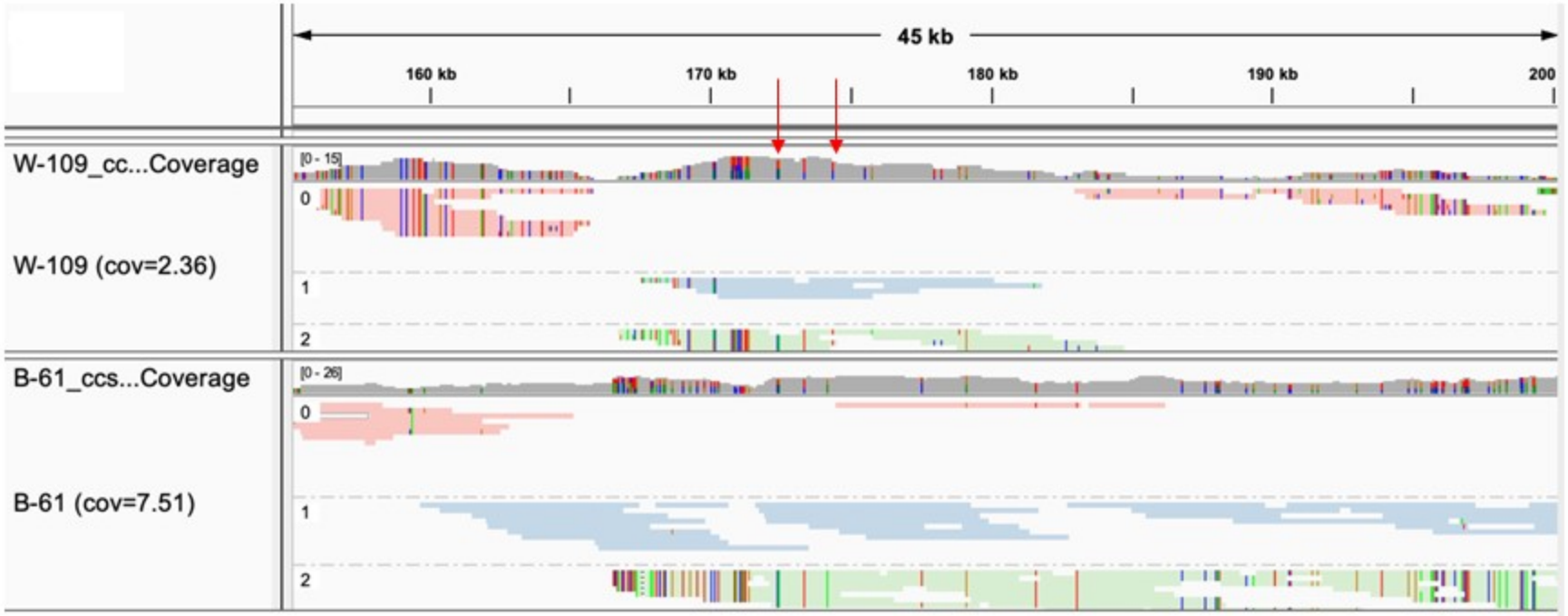


IGHV3-64



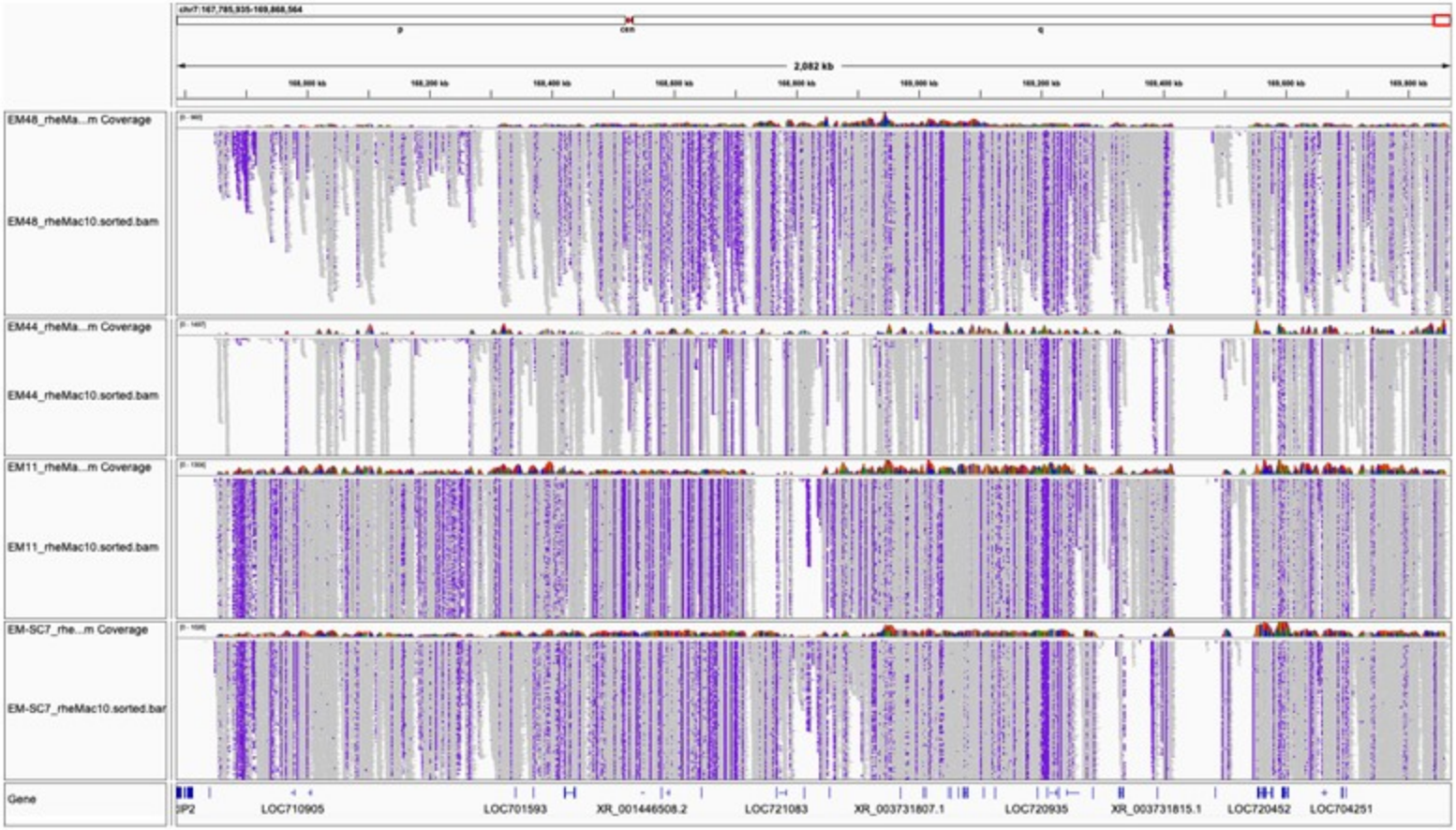
IGHV3-66

SNVs genotyped in samples with lower locus-wide coverage



Almost all SNVs were supported by CCS reads

IGH locus-wide coverage of long-read data



SVs present in IGK