Calling DNA variants – SNVs, CNVs, and SVs

Steve Laurie

Variant Effect Predictor Training Course

Prague, 6th November 2017

cnag
centre nacional d'anàlisi genòmica
centro nacional de análisis genómicc

CRG
Centre
for Genomic
Regulation

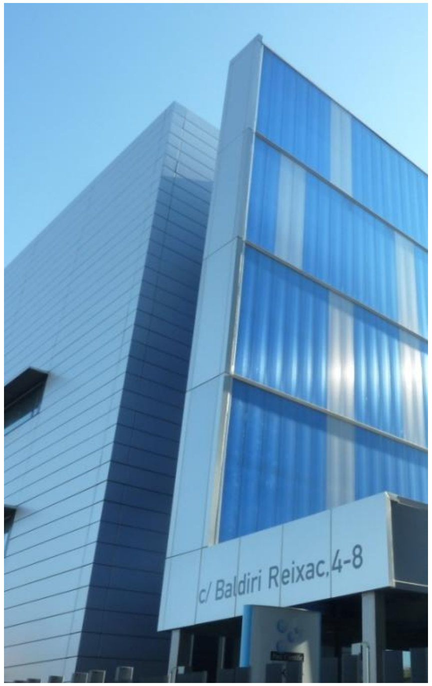
RD Connect

Calling DNA variants – SNVs, CNVs, SVs

1. What is a variant?
2. Paired End read mapping
3. Calling Single Nucleotide Variants (SNVs) and InDels
4. Calling Copy Number Variants (CNVs)
   - ➢ From Whole Genome Sequencing data
   - ➢ From Whole Exome Sequencing data
5. Calling Structural Variants (SVs)

cnag centre nacional d'anàlisi genòmica
centro nacional de análisis genómico

- Internationally recognised state-of-the-art sequencing centre situated in the Parc Científic de Barcelona. **Publically funded, not-for-profit.**
- 60 staff, **over 50% informatics/computer engineers**



Mission

Carry out projects in genome analysis that will lead to **significant improvements in people's health and quality of life**, in collaboration with the Spanish, European and International Research Community.

Research interests

- Disease Gene Identification and **Personalised Medicine**
- Cancer Genomics
- Single Cell RNAseq
- Agrogenomics and Model Organisms (e.g. genome assembly and gene prediction of various primate *spp*., Iberian Lynx, Olive …)

CNAG Genomehenge (version 2017)

Sequencing capacity

- >1000 Gbases/day = 10 human genomes per day at 30x coverage

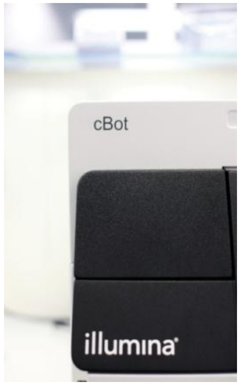
Sequencing

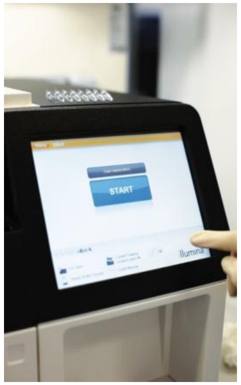
- 3 Illumina HiSeq2000
- 3 Illumina HiSeq2500
- 1 Illumina HiSeq4000
- 1 Illumina MiSeq
- 4 Illumina cBots
- 3 Oxford Nanopore MinIons

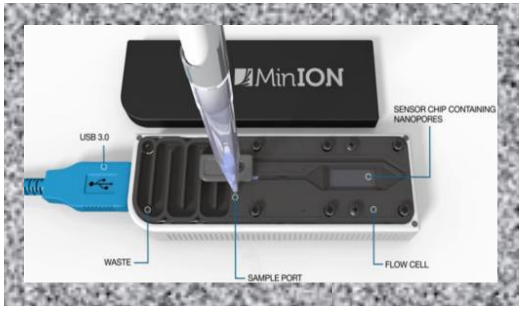
Computing

- 3552 cores
- 3.7 PB disk + 3 PB tape archive
- 35.5 TB RAM
- Barcelona SuperComputing Center - 10 x 10 Gb/s

Copyright 2005. Barcelona Supercomputing Center - BSC







cnag
centre nacional d'anàlisi genòmica
centro nacional de análisis genómico

CRG
Centre for Genomic Regulation

CNAG QA Certification

- ✓ **December 2013**

  “Illumina CSPro recognizes that CNAG provides customers with industry-leading data quality and service in genetic analysis.”

- ✓ ***May 2014***

  “CNAG has successfully completed **Agilent Certified ServicesTraining** for Target Enrichment System for NGS.”

- ✓ **December 2014**

  “ISO 9001 certified for management and performance of high throughput sequencing and genomic analysis projects and services.”

- ✓ **April 2016**

  “**ISO 17025 accreditation** for DNA & RNAAnalysis using high throughput sequencing (NGS)”

- ✓ **May 2017**

  Roche- Nimblegen SeqCap EZ Certified Service Providers
  CNAG is the first and only Nimblegen certified provider in Europe

Active participant in many international biomedical initiatives

Member of the **Global Alliance for Genomics and Health (GA4GH)**

Global Alliance
for Genomics & Health

Participation, through the National Bioinformatics Institute (INB), in **ELIXIR,** the European bioinformatics infrastructure.

elixir

Participation in the **International Human Epigenome Consortium (IHEC)**

IHEC
International Human Epigenome Consortium

Participation in the **International Cancer Genome Consortium (ICGC)**

International
Cancer Genome
Consortium

Participation in the **International Rare Diseases Research Consortium (IRDiRC)**

IRDiRC
INTERNATIONAL
RARE DISEASES RESEARCH
CONSORTIUM

Calling DNA variants – SNVs, CNVs, SVs

1. **What is a variant?**
2. Paired End read mapping
3. Calling Single Nucleotide Variants (SNVs) and InDels
4. Calling Copy Number Variants (CNVs)
   - ➢ From Whole Genome Sequencing data
   - ➢ From Whole Exome Sequencing data
5. Calling Structural Variants (SVs)

What is a **variant**?

A **variant** is any position/region in our sample **which differs from the haploid reference genome** to which we are comparing.
There are 4 basic classes:

- **Single Nucleotide Variants** (**SNVs**)
  - e.g. A → G – note **diploid** individual may be “AA”, “AG”, or “GG”

- **Short (<50nt) insertions and deletions** (**InDels**)
  - e.g. TA → TATA (**insertion** of “TA”)
  - e.g. CT → C (**deletion** of the “T” at the second position)

- **Copy Number Variants** (**CNVs**) – generally tandem duplications of typically longer regions (~1-100kb) that are often polymorphic within the population e.g. AMY1

- **Structural Variants** (**SVs**) – often larger still, and often complex in nature

What is a **variant**?

A **variant** is any position/region in our sample **which differs from the haploid reference genome** to which we are comparing.

There are 4 basic classes:

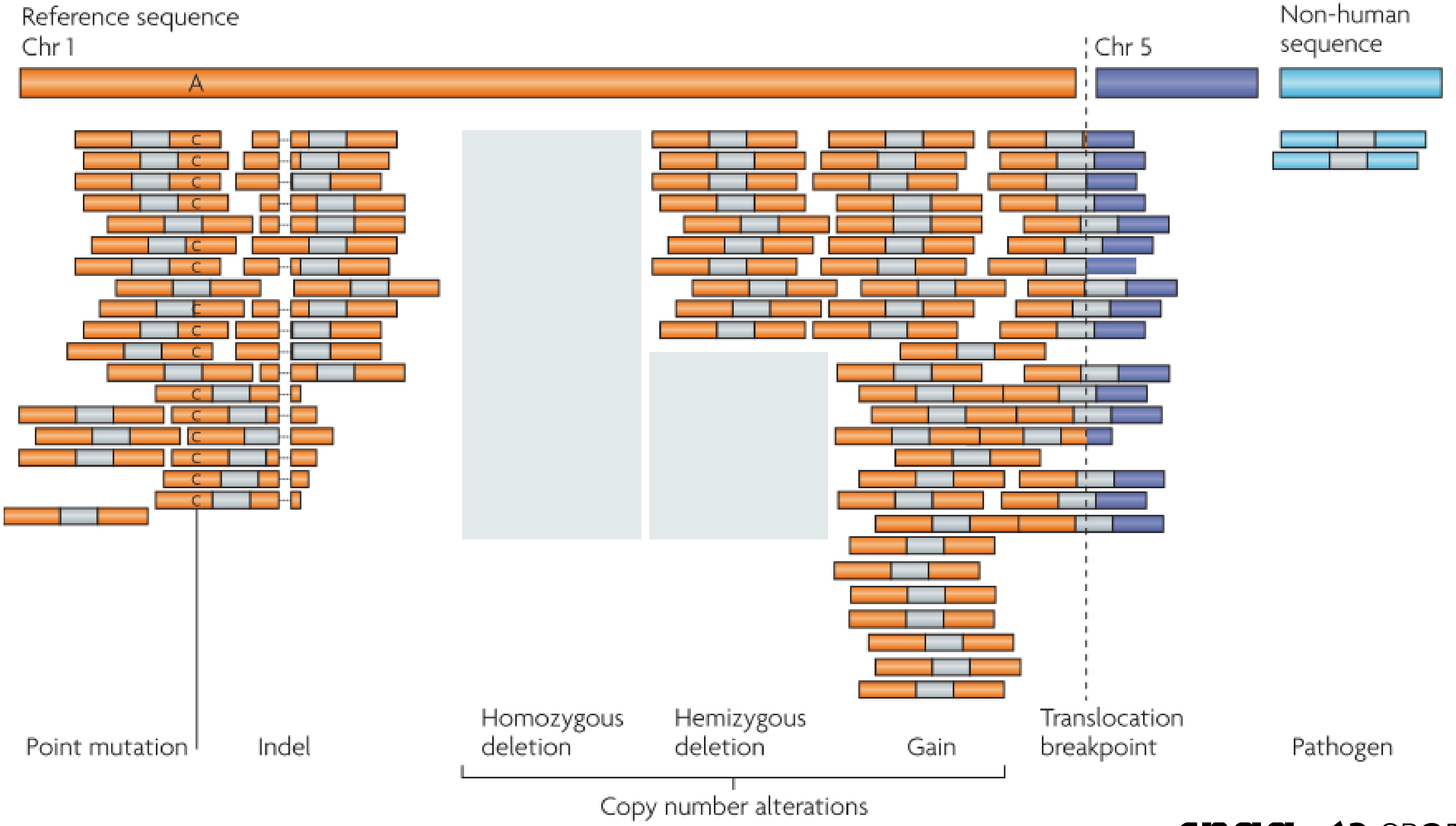
- **Single Nucleotide Variants** (**SNVs**)
  ~ 3,750,000-4,500,000 (Yuen *et al*, Nat. Neuro. 2017)

- **Short (<50nt) insertions and deletions** (**InDels**)
  ~ 700,000-1,000,000 (Yuen *et al*, Nat. Neuro. 2017)

- **Copy Number Variants** (**CNVs**) – generally tandem duplications
  ~ 11.3Mbp per individual (1kGP);
  5-9% of genome 50bp-3Mbp (Zarrei *et al*, NRG, 2015)

- **Structural Variants** (**SVs**) – often larger still, and often complex in nature
  ~ 10Mbp per individual (1kGP) – 59Mbp (English *et al*, 2015)

Summary of Variant Types

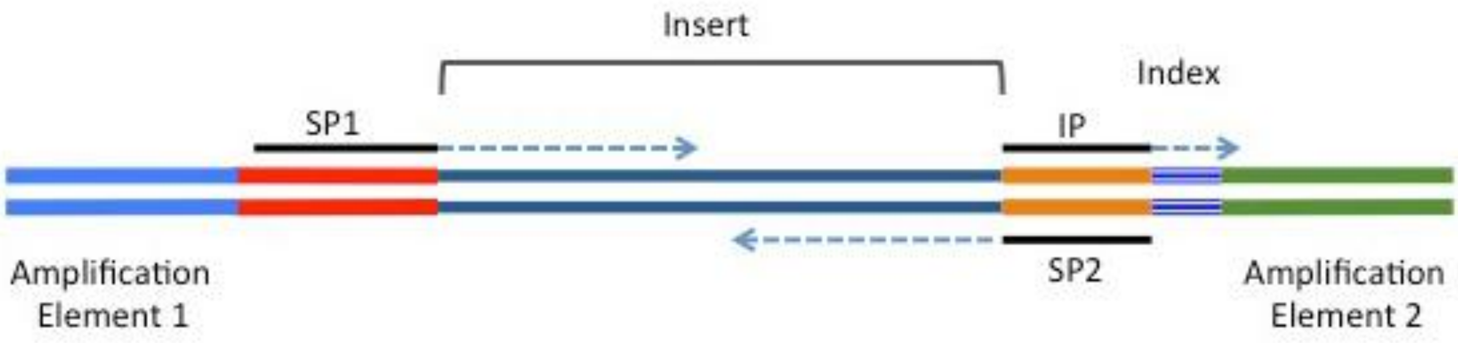


Meyerson et al. . 2010. Advances in understanding cancer genomes through second-generation sequencing. Nature Reviews Genetics 11, no. 10 (October): 685-696

Calling DNA variants – SNVs, CNVs, SVs

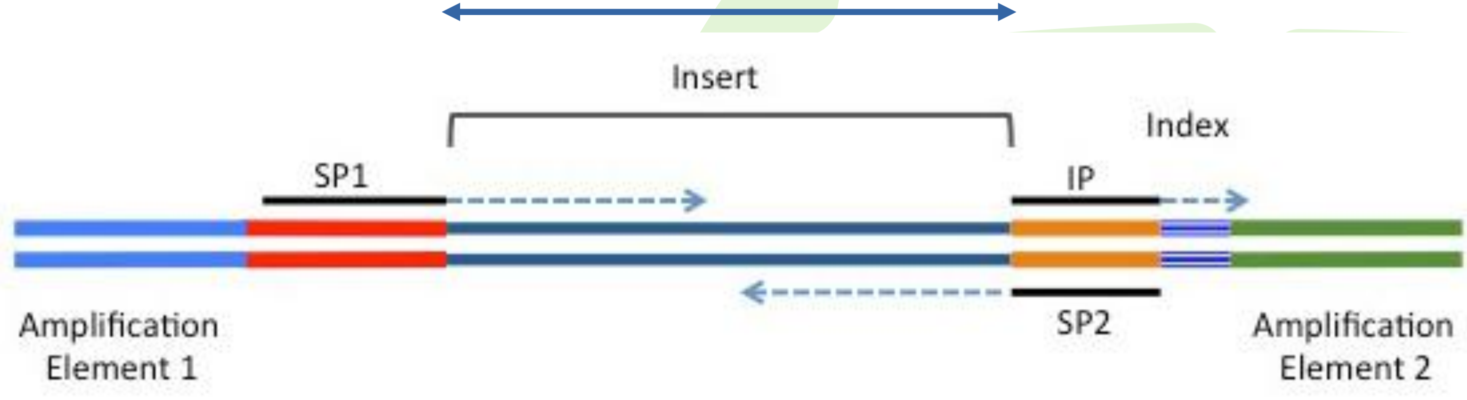
1. What is a variant?
2. **Paired End read mapping**
3. Calling Single Nucleotide Variants (SNVs) and InDels
4. Calling Copy Number Variants (CNVs)
   - ➢ From Whole Genome Sequencing data
   - ➢ From Whole Exome Sequencing data
5. Calling Structural Variants (SVs)

Paired-end Read Mapping



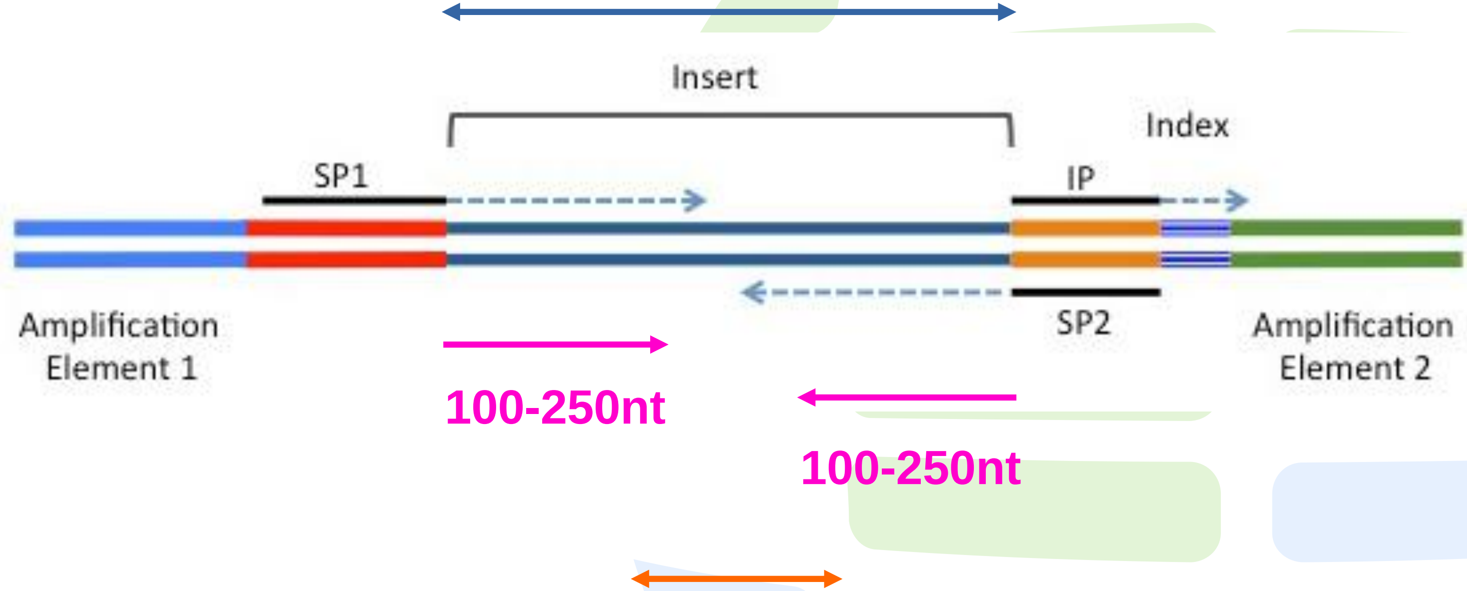
Paired-end Read Mapping

Typically 250-600bp



Paired-end Read Mapping

Typically 250-600bp

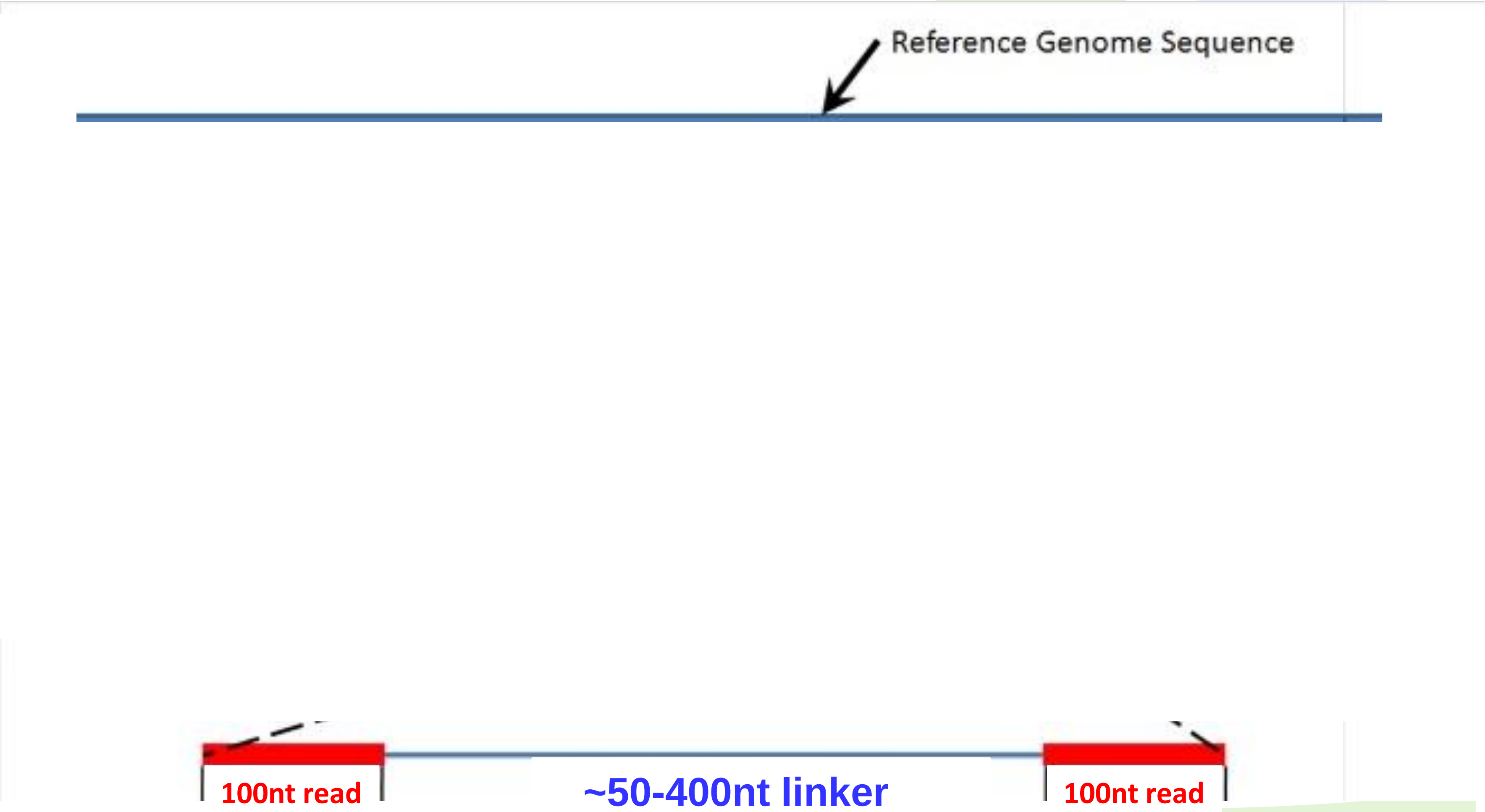


100-250nt

100-250nt

~50-100bp

Paired-end Read Mapping



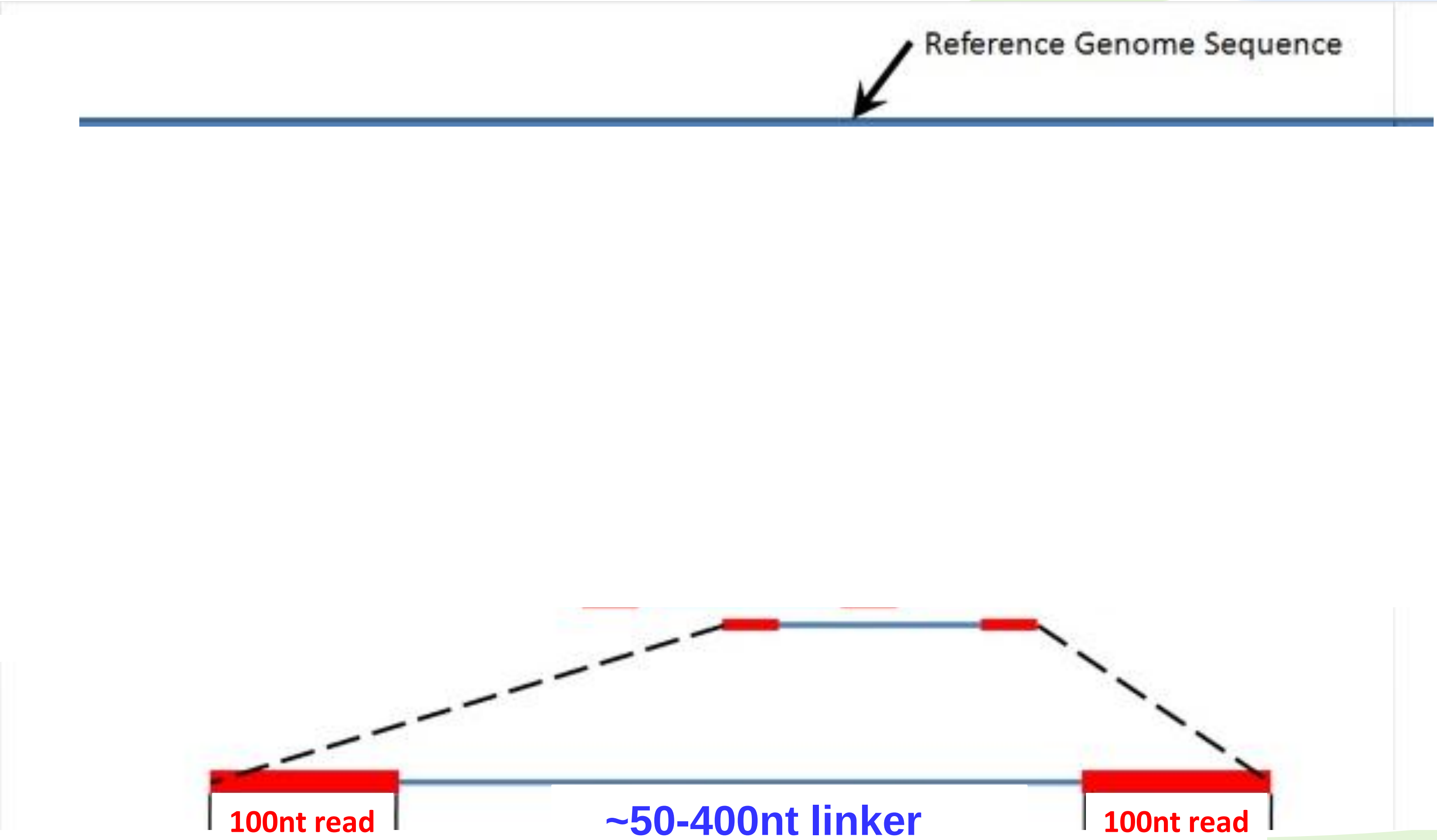
Reference Genome Sequence

100nt read

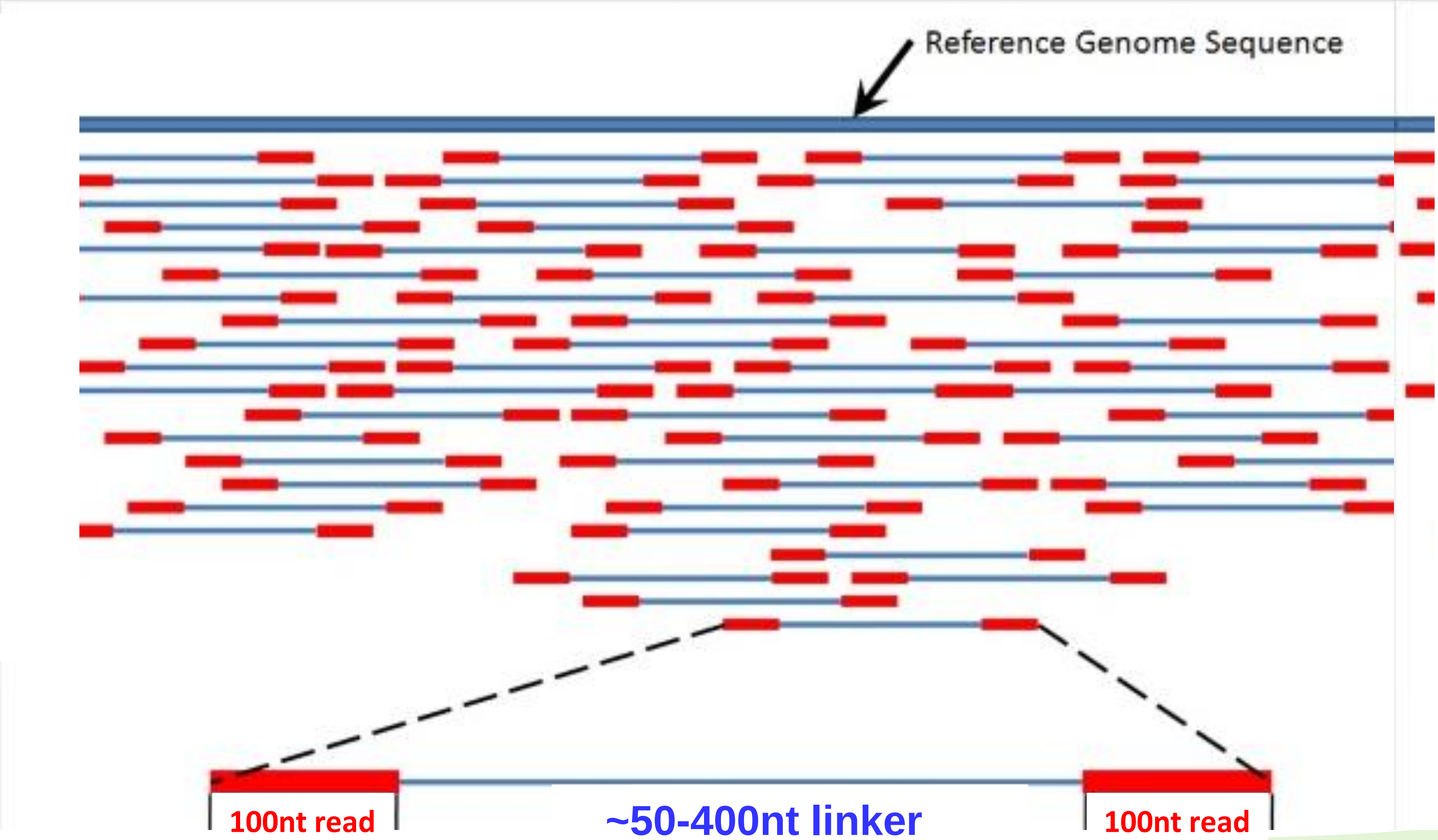
~50-400nt linker

100nt read

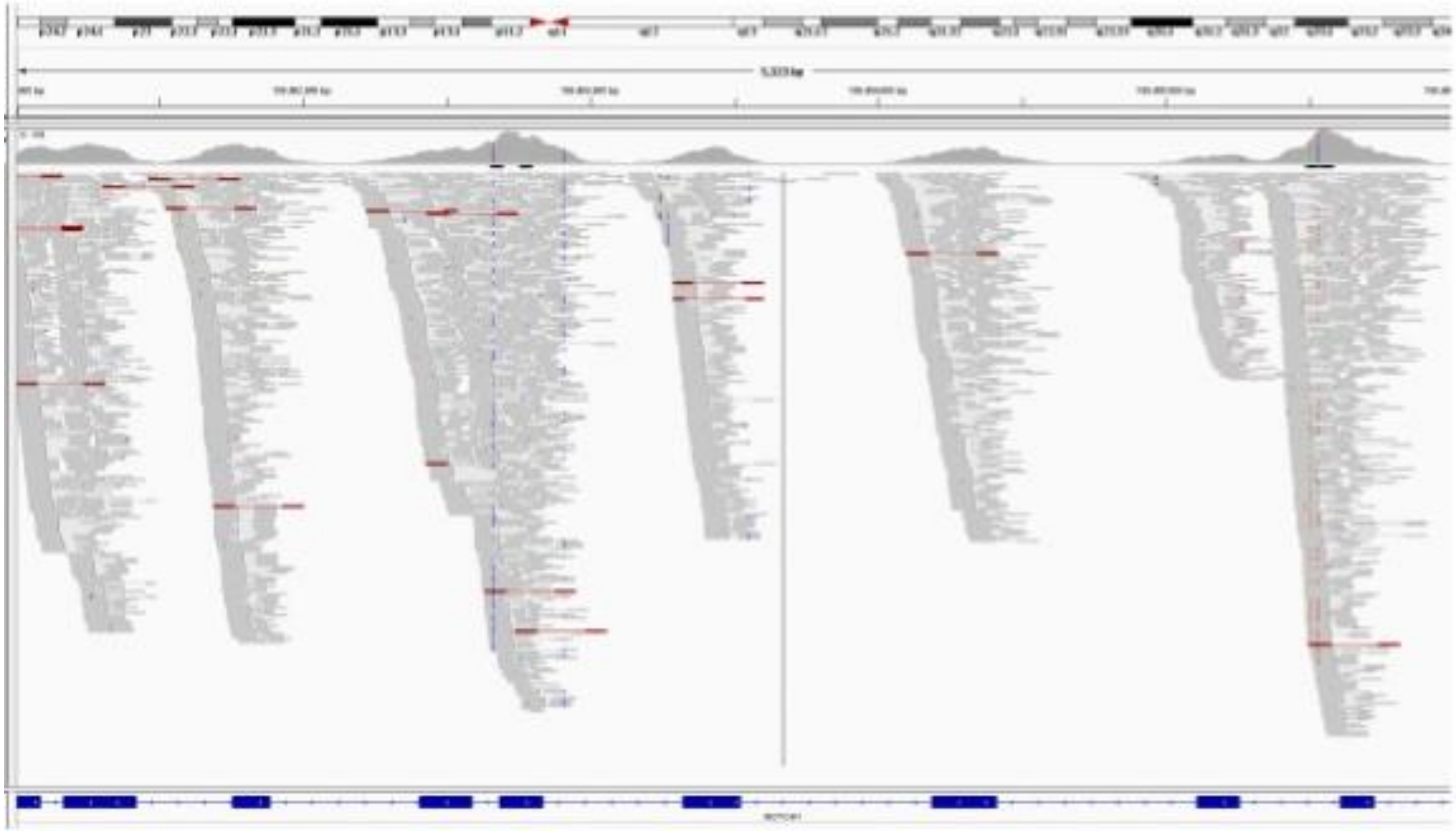
Paired-end Read Mapping



Paired-end Read Mapping



Mapped reads viewed in IGV



What is a **variant**?

A **variant** is any position/region in our sample **which differs from the haploid reference genome** to which we are comparing.
There are 4 basic classes:

- **Single Nucleotide Variants (SNVs)**
  ~ 4,000,000
- **Short (<50nt) insertions and deletions (InDels)**
  ~ 400,000
- **Copy Number Variants (CNVs)** – generally tandem duplications
  ~ 5-10% of genome
- **Structural Variants (SVs)** – often larger still, and often complex in nature
  ~ 13% of genome

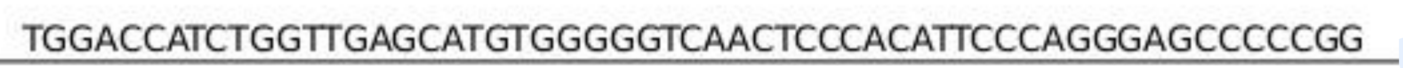
Calling DNA variants – SNVs, CNVs, SVs

1. What is a variant?
2. Paired End read mapping
3. **Calling Single Nucleotide Variants (SNVs) and InDels**
4. Calling Copy Number Variants (CNVs)
   - From Whole Genome Sequencing data
   - From Whole Exome Sequencing data
5. Calling Structural Variants (SVs)

Calling SNVs and InDels

Calling SNVs and InDels

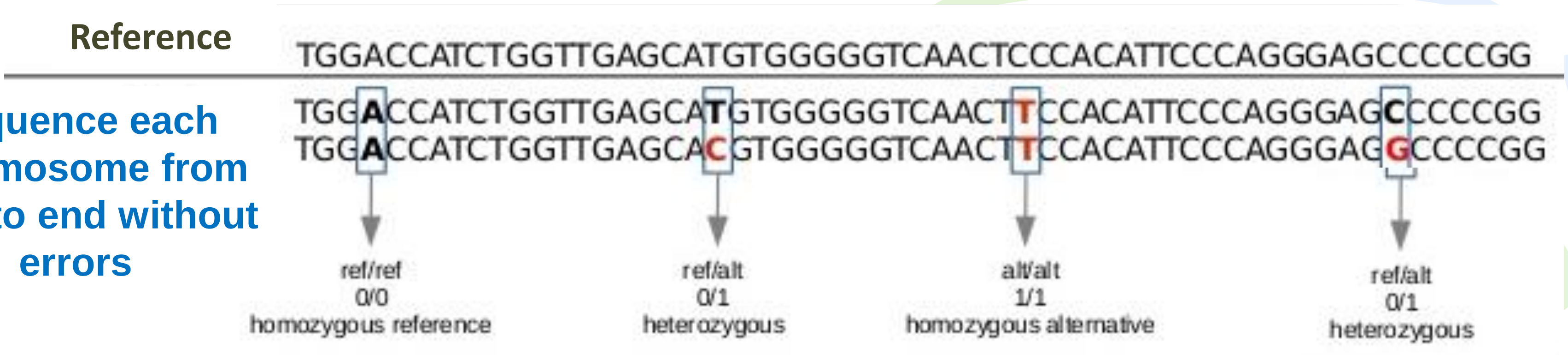
Reference



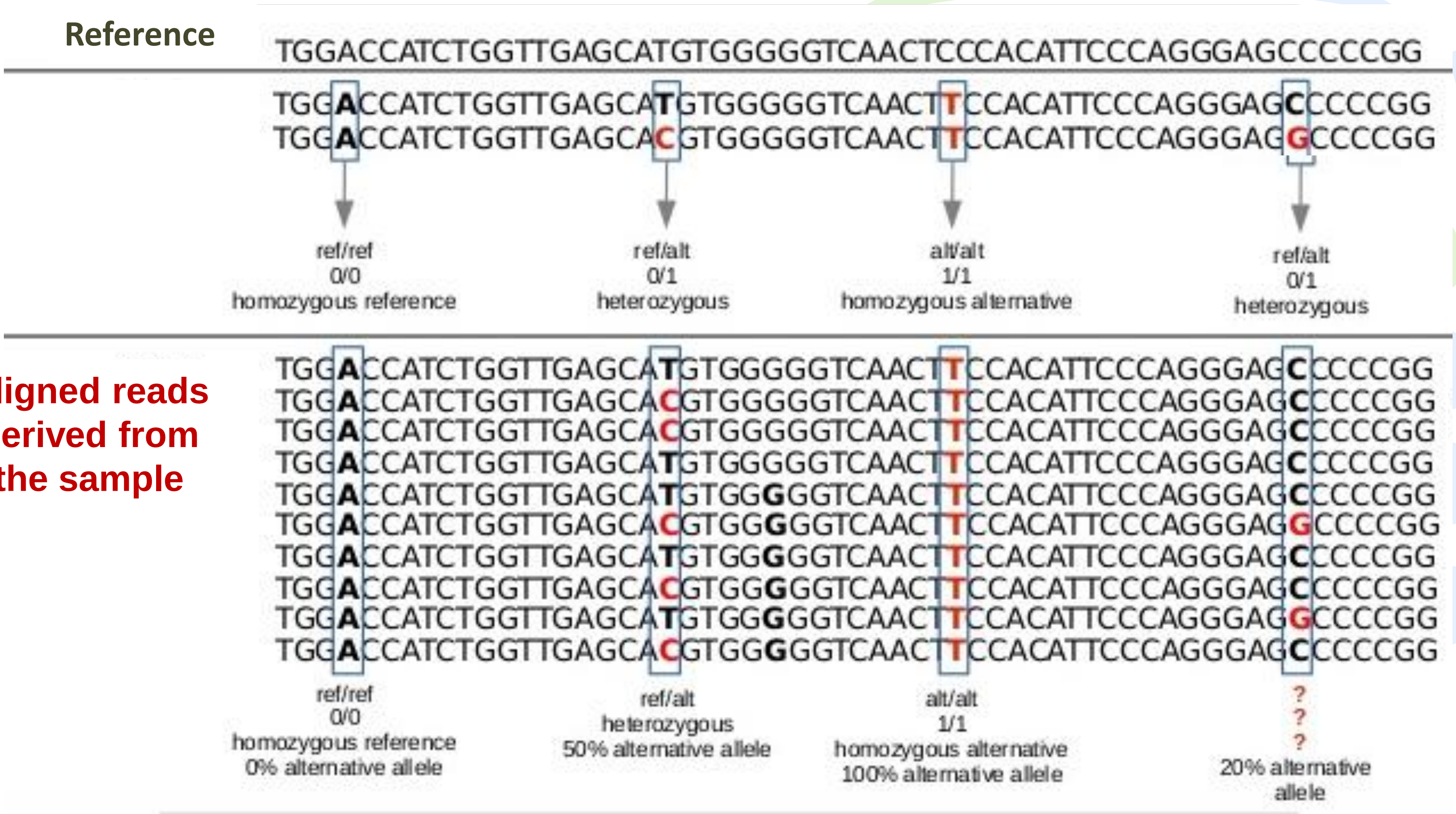
TGGACCATCTGGTTGAGCATGTGGGGGTCAACTCCCACATTCCCAGGGAGCCCCCGG

Calling SNVs and InDels – **dream future?**

Sequence each chromosome from start to end without errors


Reference
TGGACCATCTGGTTGAGCATGTGGGGGTCAACTCCCACATTCCCAGGGAGCCCCGG
TGGACCATCTGGTTGAGCATGTGGGGGTCAACTTCCACATTCCCAGGGAGCCCCCGG
TGGACCATCTGGTTGAGCACGTGGGGGTCAACTTCCACATTCCCAGGGAGGCCCCGG
ref/ref
0/0
homozygous reference
ref/alt
0/1
heterozygous
alt/alt
1/1
homozygous alternative
ref/alt
0/1
heterozygous

Calling SNVs and InDels – **back to reality**



Reference

TGGACCATCTGGTTGAGCATGTGGGGGTCAACTCCCACATTCCCAGGGAGCCCCCGG

TGGACCATCTGGTTGAGCATGTGGGGGTCAACTTCCACATTCCCAGGGAGCCCCCGG
TGGACCATCTGGTTGAGCACGTGGGGGTCAACTTCCACATTCCCAGGGAGGCCCCGG

ref/ref
0/0
homozygous reference

ref/alt
0/1
heterozygous

alt/alt
1/1
homozygous alternative

ref/alt
0/1
heterozygous

Aligned reads derived from the sample

TGGACCATCTGGTTGAGCATGTGGGGGTCAACTTCCACATTCCCAGGGAGCCCCCGG
TGGACCATCTGGTTGAGCACGTGGGGGTCAACTTCCACATTCCCAGGGAGCCCCCGG
TGGACCATCTGGTTGAGCACGTGGGGGTCAACTTCCACATTCCCAGGGAGCCCCCGG
TGGACCATCTGGTTGAGCATGTGGGGGTCAACTTCCACATTCCCAGGGAGCCCCCGG
TGGACCATCTGGTTGAGCATGTGGGGGTCAACTTCCACATTCCCAGGGAGCCCCCGG
TGGACCATCTGGTTGAGCACGTGGGGGTCAACTTCCACATTCCCAGGGAGGCCCCGG
TGGACCATCTGGTTGAGCATGTGGGGGTCAACTTCCACATTCCCAGGGAGCCCCCGG
TGGACCATCTGGTTGAGCACGTGGGGGTCAACTTCCACATTCCCAGGGAGCCCCCGG
TGGACCATCTGGTTGAGCATGTGGGGGTCAACTTCCACATTCCCAGGGAGGCCCCGG
TGGACCATCTGGTTGAGCACGTGGGGGTCAACTTCCACATTCCCAGGGAGCCCCCGG

ref/ref
0/0
homozygous reference
0% alternative allele

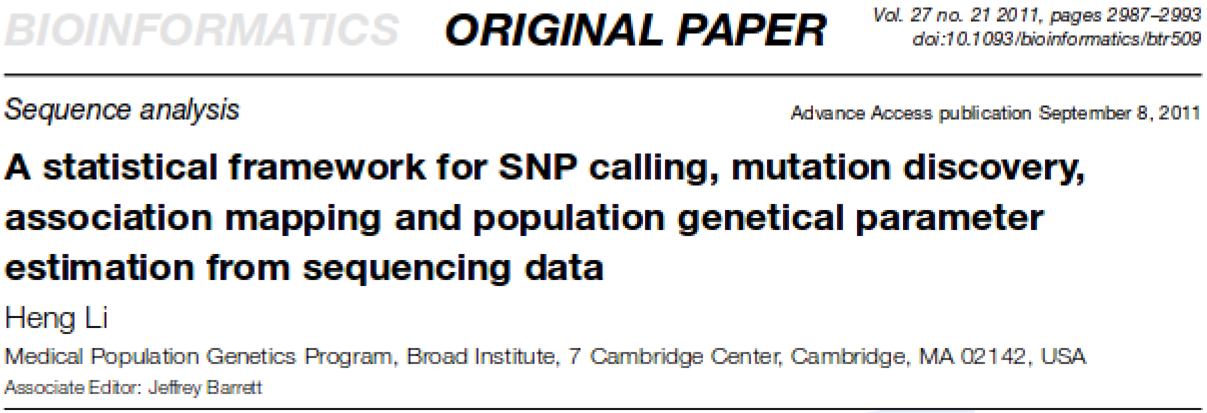
ref/alt
heterozygous
50% alternative allele

alt/alt
1/1
homozygous alternative
100% alternative allele

?
?
?
20% alternative allele

Calling SNVs & InDels

Tools for Calling SNVs & InDels



BIOINFORMATICS **ORIGINAL PAPER**

Vol. 27 no. 21 2011, pages 2987–2993
doi:10.1093/bioinformatics/btr509

Sequence analysis

Advance Access publication September 8, 2011

A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data

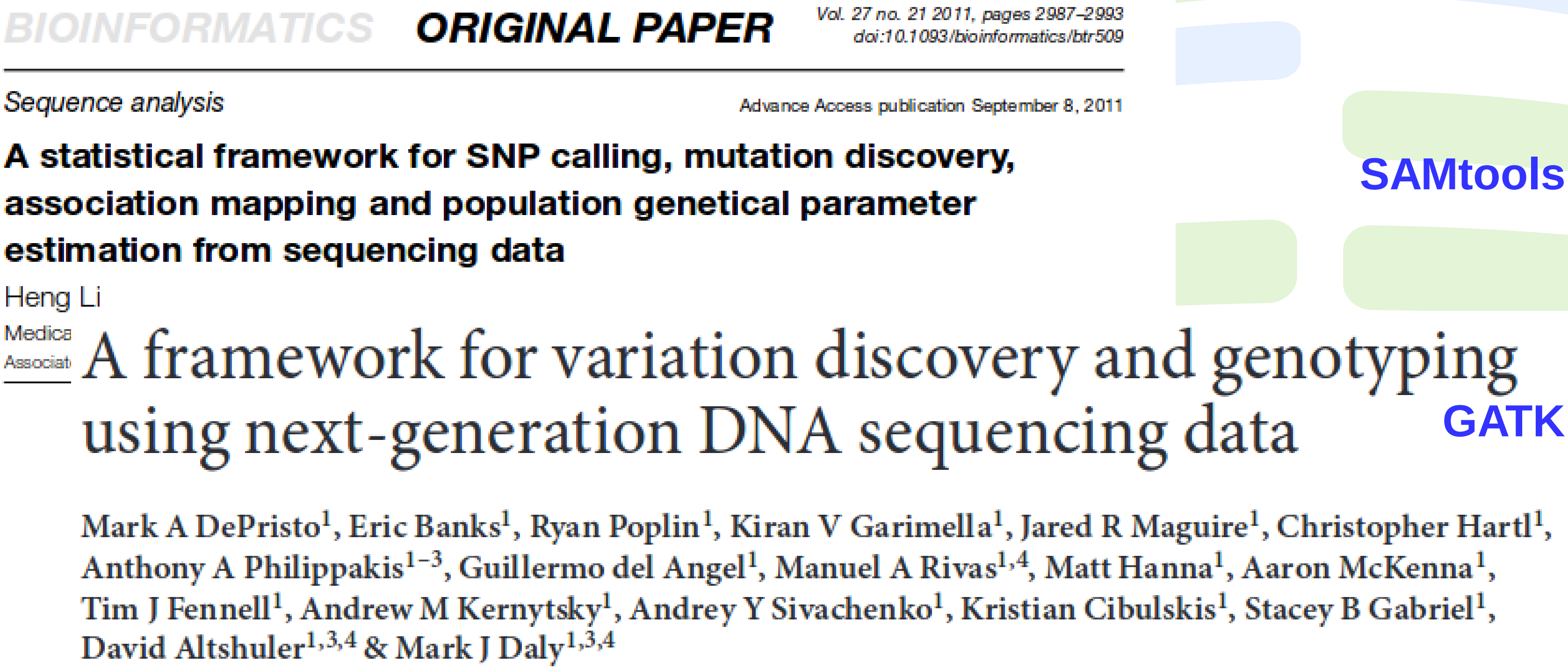
Heng Li

Medical Population Genetics Program, Broad Institute, 7 Cambridge Center, Cambridge, MA 02142, USA

Associate Editor: Jeffrey Barrett

SAMtools, 2011

Tools for Calling SNVs & InDels



BIOINFORMATICS **ORIGINAL PAPER**

Vol. 27 no. 21 2011, pages 2987–2993
doi:10.1093/bioinformatics/btr509

Sequence analysis

Advance Access publication September 8, 2011

A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data

Heng Li

Medica

Associat

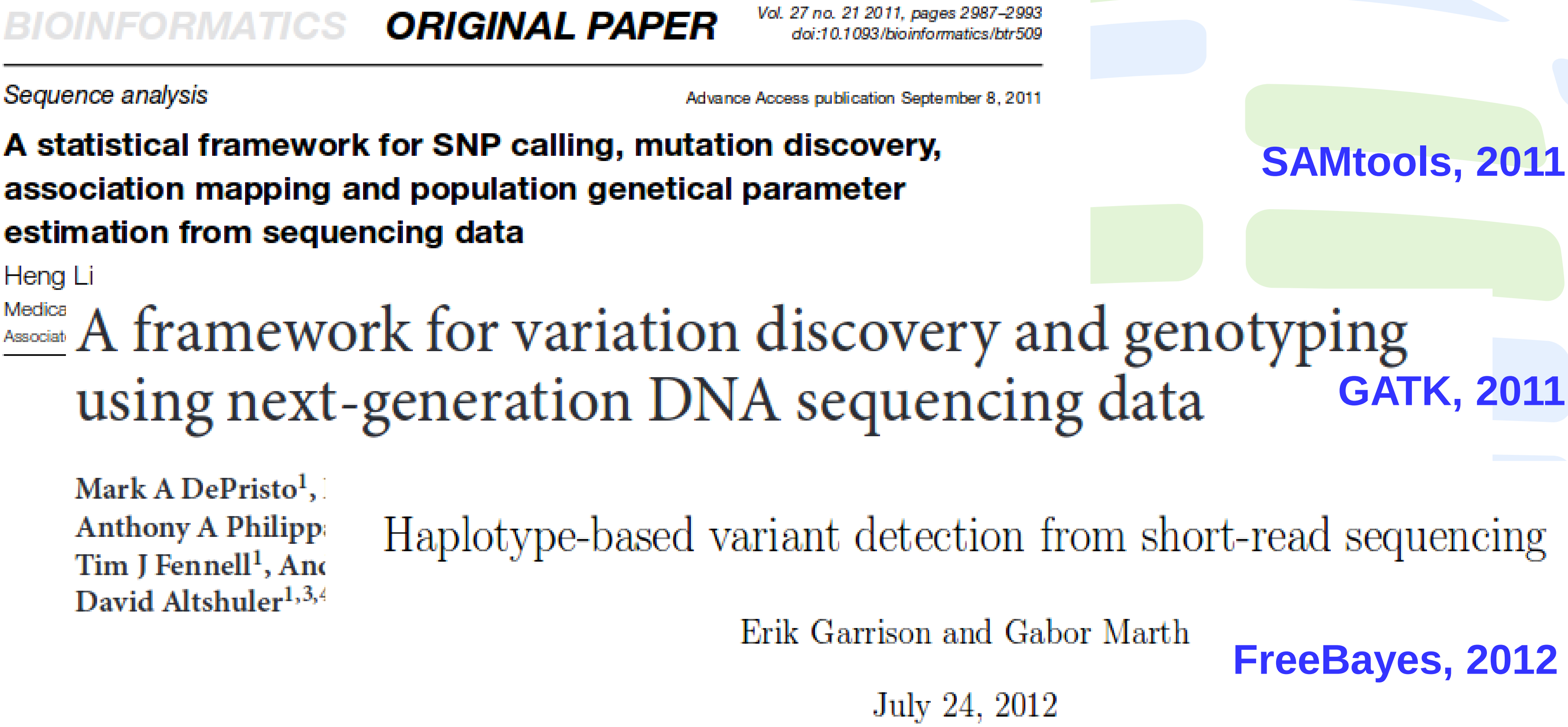
SAMtools, 2011

A framework for variation discovery and genotyping using next-generation DNA sequencing data

Mark A DePristo[1], Eric Banks[1], Ryan Poplin[1], Kiran V Garimella[1], Jared R Maguire[1], Christopher Hartl[1], Anthony A Philippakis[1–3], Guillermo del Angel[1], Manuel A Rivas[1,4], Matt Hanna[1], Aaron McKenna[1], Tim J Fennell[1], Andrew M Kernytsky[1], Andrey Y Sivachenko[1], Kristian Cibulskis[1], Stacey B Gabriel[1], David Altshuler[1,3,4] & Mark J Daly[1,3,4]

GATK, 2011

Tools for Calling SNVs & InDels



BIOINFORMATICS **ORIGINAL PAPER** *Vol. 27 no. 21 2011, pages 2987–2993 doi:10.1093/bioinformatics/btr509*

Sequence analysis Advance Access publication September 8, 2011

A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data

Heng Li

Medica

Associat

A framework for variation discovery and genotyping using next-generation DNA sequencing data

Mark A DePristo[1],
Anthony A Philipp
Tim J Fennell[1], And
David Altshuler[1,3,4]

Haplotype-based variant detection from short-read sequencing

Erik Garrison and Gabor Marth

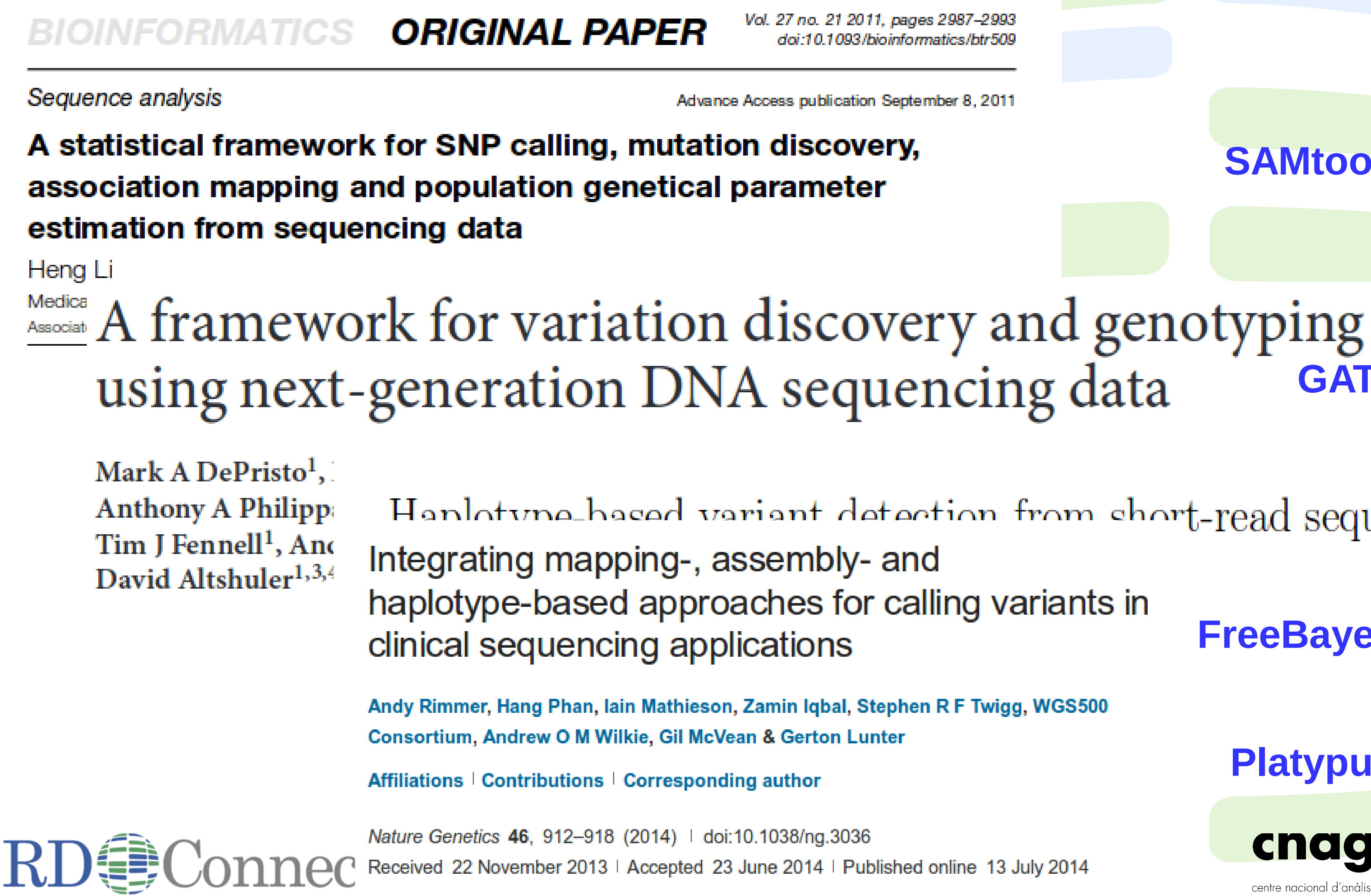
July 24, 2012

SAMtools, 2011

GATK, 2011

FreeBayes, 2012

Variant Calling Tools (SNVs & InDels)



BIOINFORMATICS ORIGINAL PAPER

Vol. 27 no. 21 2011, pages 2987–2993
doi:10.1093/bioinformatics/btr509

Sequence analysis

Advance Access publication September 8, 2011

A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data

Heng Li

Medica
Associat

A framework for variation discovery and genotyping using next-generation DNA sequencing data

Mark A DePristo[1],
Anthony A Philipp
Tim J Fennell[1], And
David Altshuler[1,3,4]

Haplotype-based variant detection from short-read sequencing

Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications

Andy Rimmer, Hang Phan, Iain Mathieson, Zamin Iqbal, Stephen R F Twigg, WGS500 Consortium, Andrew O M Wilkie, Gil McVean & Gerton Lunter

Affiliations | Contributions | Corresponding author

Nature Genetics **46**, 912–918 (2014) | doi:10.1038/ng.3036
Received 22 November 2013 | Accepted 23 June 2014 | Published online 13 July 2014

SAMtools, 2011

GATK, 2011

FreeBayes, 2012

Platypus, 2014

Calling SNVs & InDels

Variant calling tools will start by calling **every** potential variant they observe

- This will include true variants, and false-positives due to:
  - Sample quality/Library preparation issues
  - PCR artefacts
  - Sequencing errors
  - Mapping issues
  - Variant Calling algorithm issues

- Subsequently they apply a number of mechanisms to attempt to help identify the false-positives.

Calling SNVs & InDels

Variant calling tools will start by calling **every** potential variant they observe

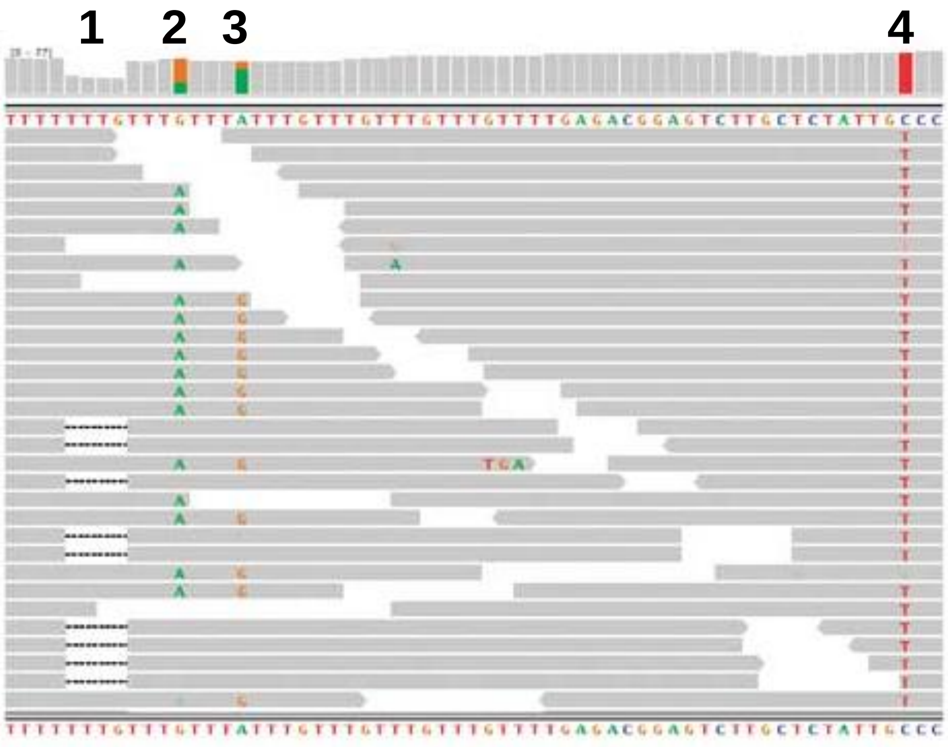
- This will include true variants, and false-positives due to:
  - Sample quality/Library preparation issues
  - PCR artefacts
  - Sequencing errors
  - Mapping issues
  - Variant Calling algorithm issues

- Subsequently they apply a number of mechanisms to attempt to help identify the false-positives
- **Currently, you will always encounter some false positives, and some false negatives**

Calling SNVs & InDels

- There are **3 key metrics** that can give us a good idea as to whether to trust a variant call
  - **Read Depth (DP)**
    - A general rule is the deeper, the better – ideally >20 supporting reads
  - **Genotype Quality (GQ)**
    - A value produced by variant calling algorithms indicating the probability that the call is wrong. Scaled from 1-99 (30 means 1/1000)
  - **Allele Balance (aka. Alternative/Beta Allele frequency)**
    - For heterozygote positions this should be close to 0.5
      - 0.25-0.75 is generally reliable
      - <0.15 or >0.85 is highly suspicious
    - For homozygote positions this should be very close to 0 or 1

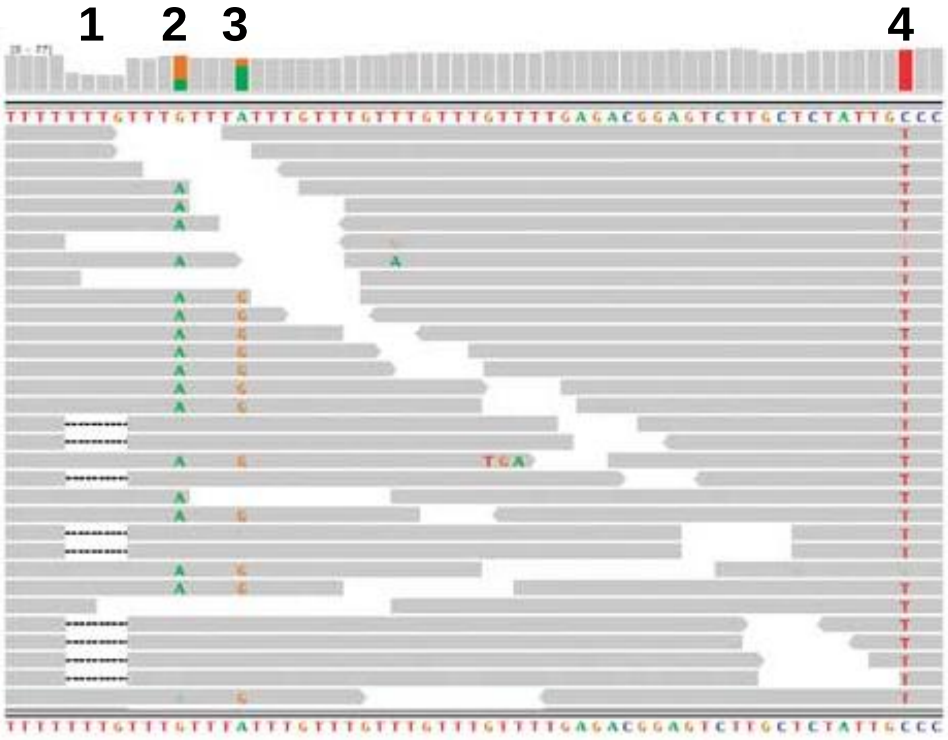
InDel identification

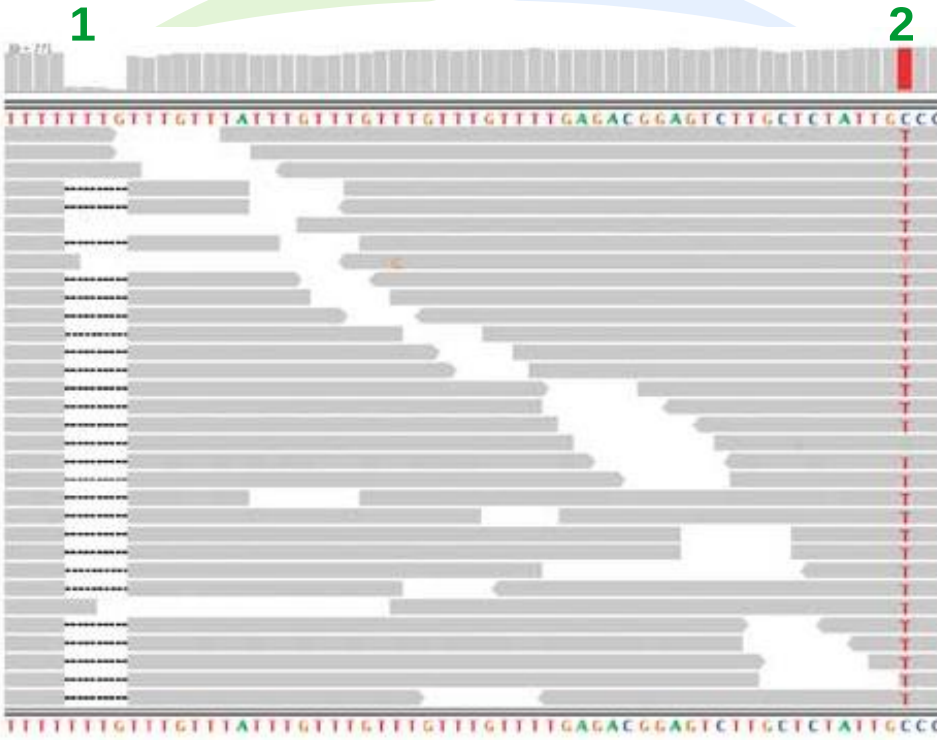

1
2
3
4
[0 - 77]
TTTTTTTGTTTGTTTATTTGTTTGTTTGTTTGTTTTTGAGACGGAGTCTTGCTCTATTGCCC

Raw BWA mapped reads

DePristo, M. *et al*. (2011)

InDel identification


1
2
3
4
TTTTTTTGTTTGTTTATTTGTTTGTTTGTTTGTTTTGAGACGGAGTCTTGCTCTATTGCCC
A
G
TGA
T

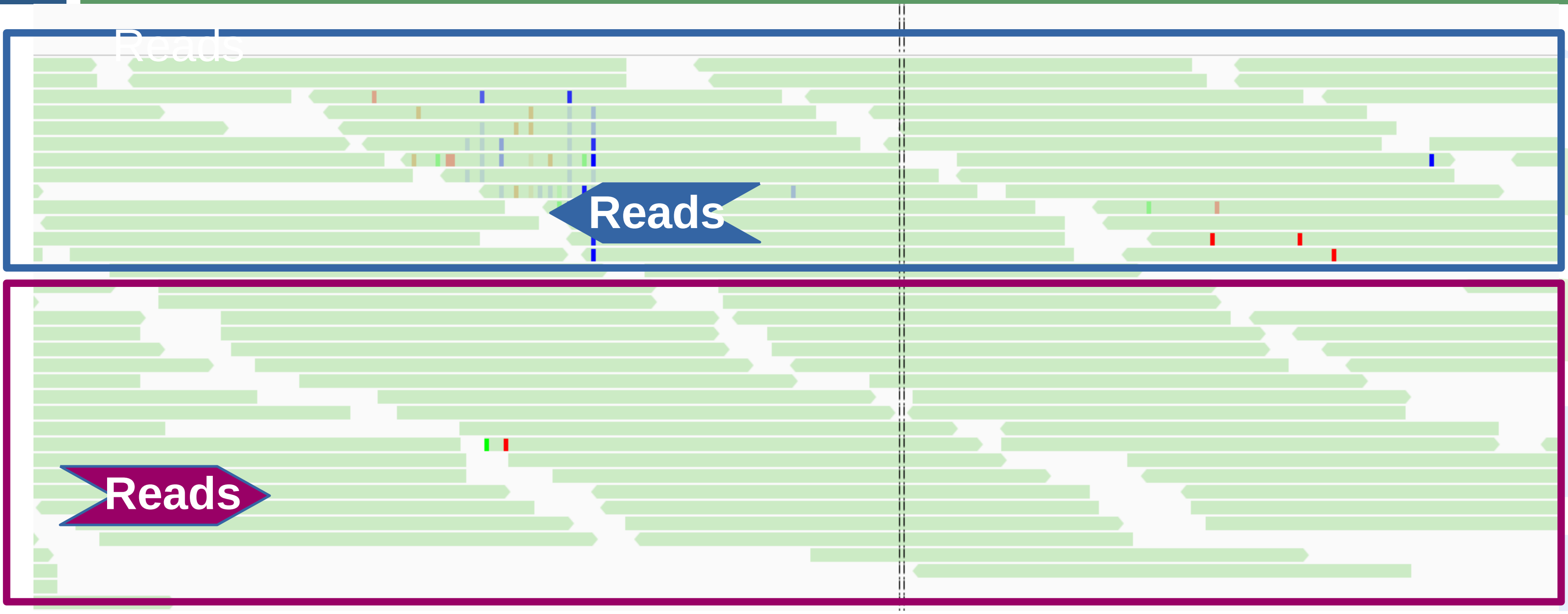

1
2
TTTTTTTGTTTGTTTATTTGTTTGTTTGTTTGTTTTGAGACGGAGTCTTGCTCTATTGCCC
T

Raw BWA mapped reads

Following GATK local realignment

DePristo, M. *et al*. (2011)

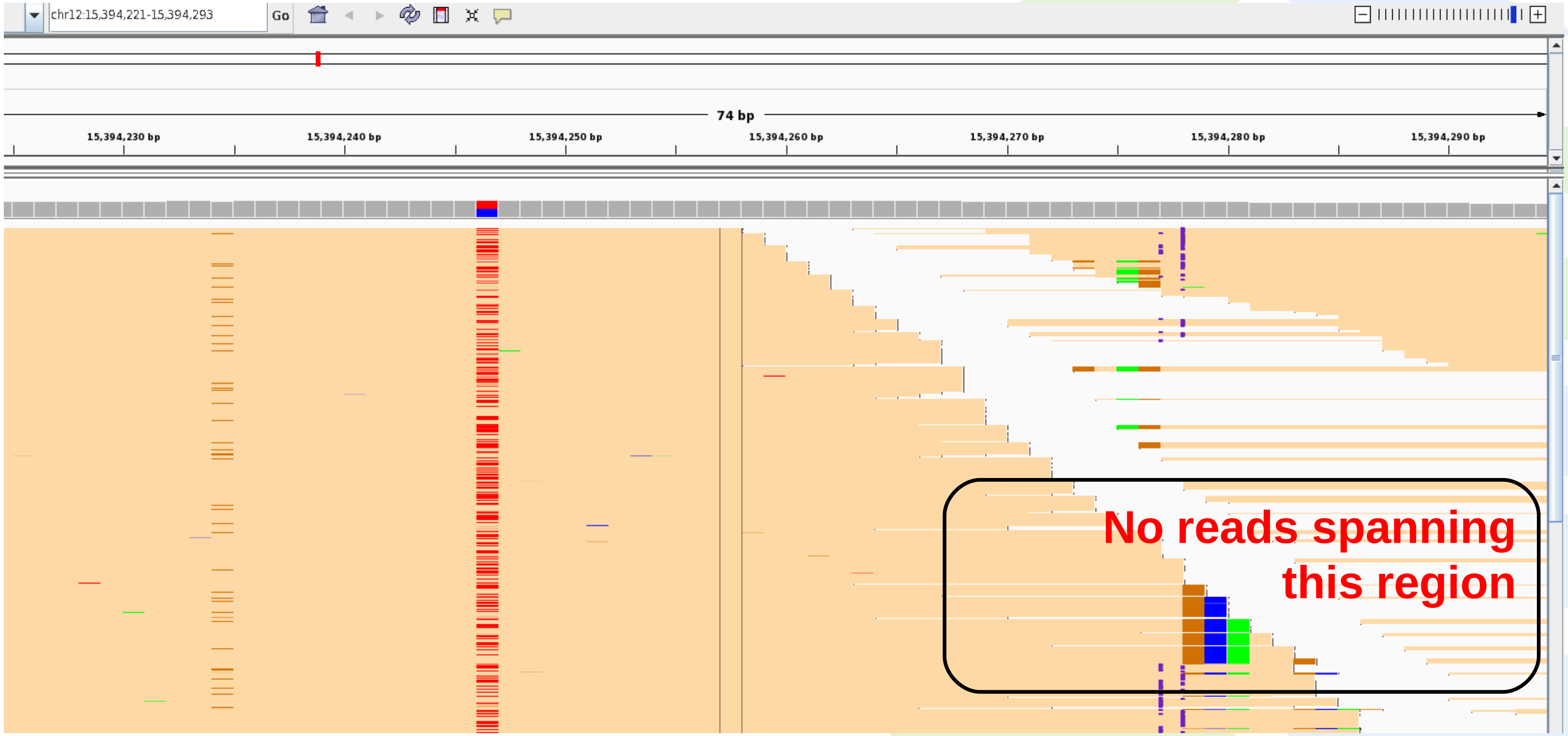
Strand Bias



GATK: FS field (Phred-scaled p-value)

SAMtools: PV4 field (p-value)

Tail Distance/Variant Position Bias



No reads spanning this region

ReadPosRankSum = 1.635

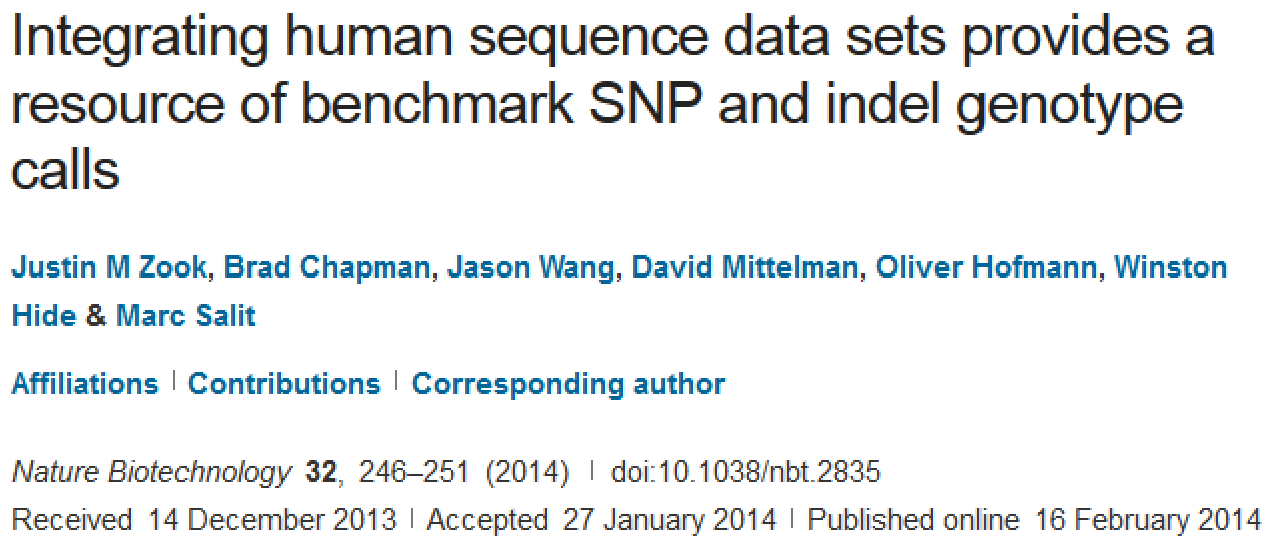
ReadPosRankSum = - 0.434

ReadPosRankSum = - 9.805

SAMtools equivalent: PV4 field (p-value)

NIST/GIAB – Analyses of NA12878

- The NIST is attempting to produce "Gold Standard" call sets for all variants in NA12878, and other samples, through integration of results from a variety of pipelines



Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls

Justin M Zook, Brad Chapman, Jason Wang, David Mittelman, Oliver Hofmann, Winston Hide & Marc Salit

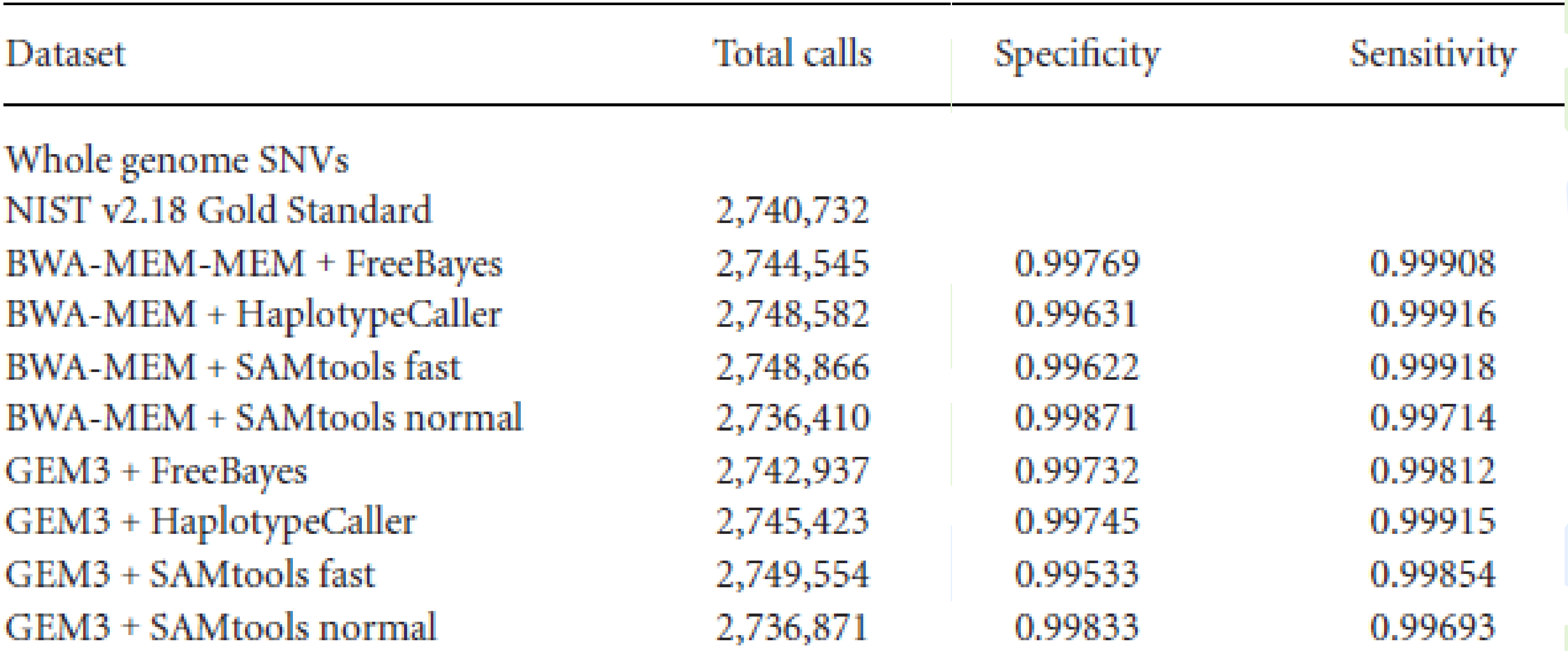
Affiliations | Contributions | Corresponding author

Nature Biotechnology **32**, 246–251 (2014) | doi:10.1038/nbt.2835

Received 14 December 2013 | Accepted 27 January 2014 | Published online 16 February 2014

Benchmarking of VC Pipelines

NA12878 50xWGS FastQs (Illumina Platinum), analysed with several pipelines. Concordance with **Gold Standard VC set** from GIAB/NIST (Zook *et al.*, 2014) **for the reliably-callable region of the genome (70%)**

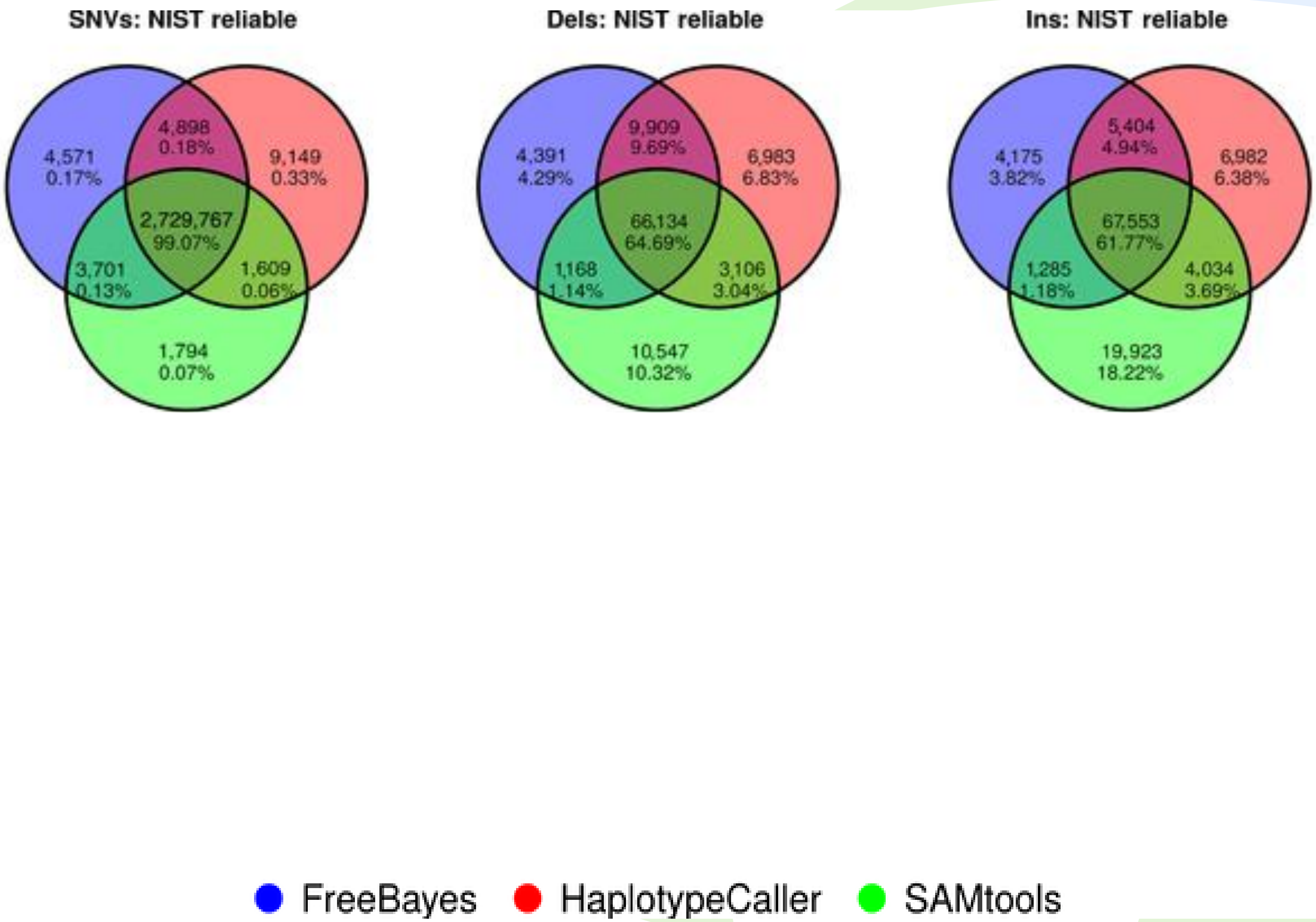


Dataset	Total calls	Specificity	Sensitivity
Whole genome SNVs			
NIST v2.18 Gold Standard	2,740,732		
BWA-MEM-MEM + FreeBayes	2,744,545	0.99769	0.99908
BWA-MEM + HaplotypeCaller	2,748,582	0.99631	0.99916
BWA-MEM + SAMtools fast	2,748,866	0.99622	0.99918
BWA-MEM + SAMtools normal	2,736,410	0.99871	0.99714
GEM3 + FreeBayes	2,742,937	0.99732	0.99812
GEM3 + HaplotypeCaller	2,745,423	0.99745	0.99915
GEM3 + SAMtools fast	2,749,554	0.99533	0.99854
GEM3 + SAMtools normal	2,736,871	0.99833	0.99693

Laurie *et al*. Human Mutation, 2016

Benchmarking of VC Pipelines

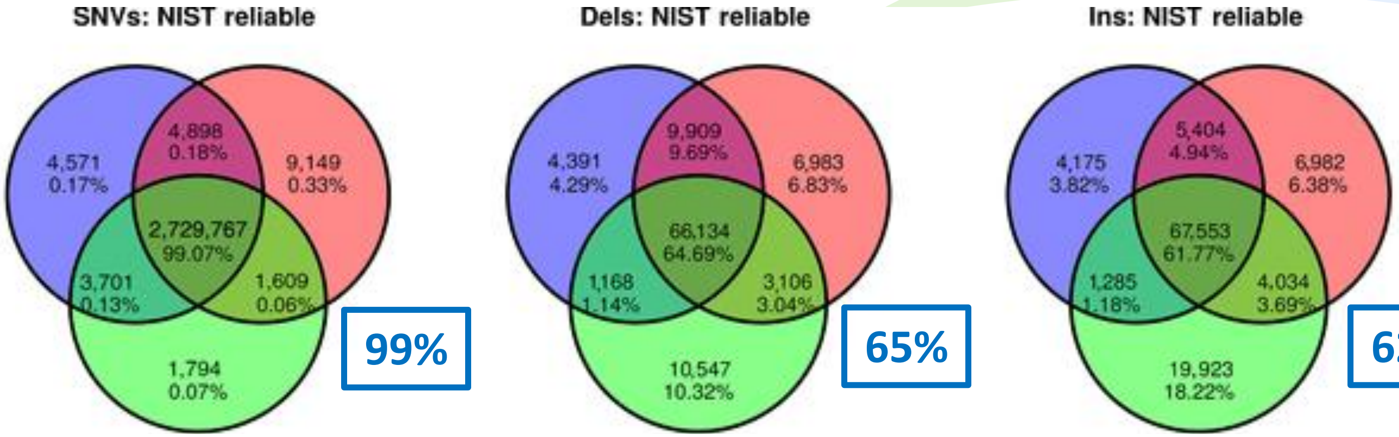
Reliably Callable



Laurie *et al*. Human Mutation, 2016

Benchmarking of VC Pipelines

Reliably Callable



99%

65%

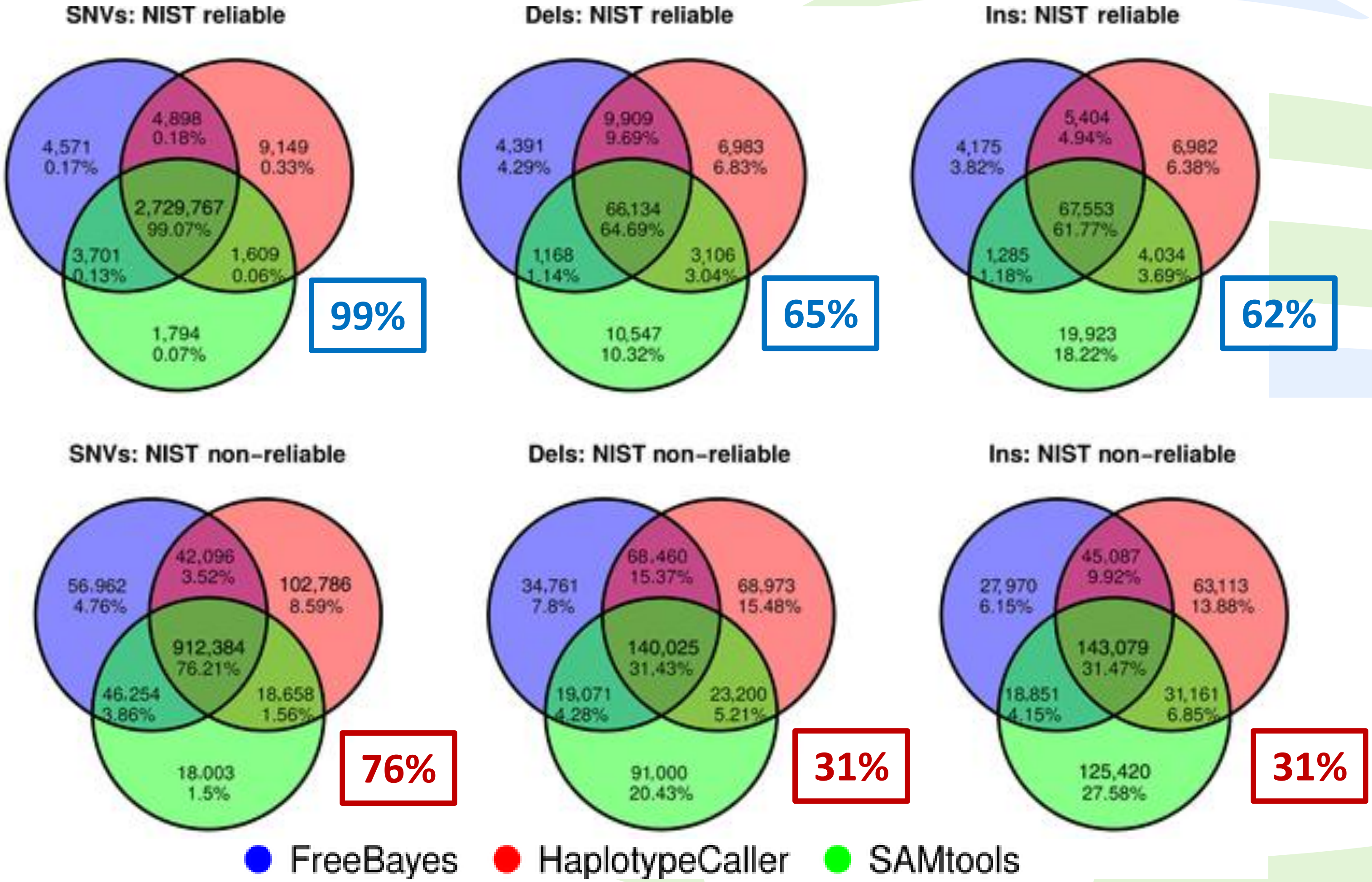
62%

● FreeBayes ● HaplotypeCaller ● SAMtools

Laurie *et al*. Human Mutation, 2016

Benchmarking of VC Pipelines

Reliably Callable



99%

65%

62%

Not Reliably Callable

76%

31%

31%

Laurie *et al*. Human Mutation, 2016

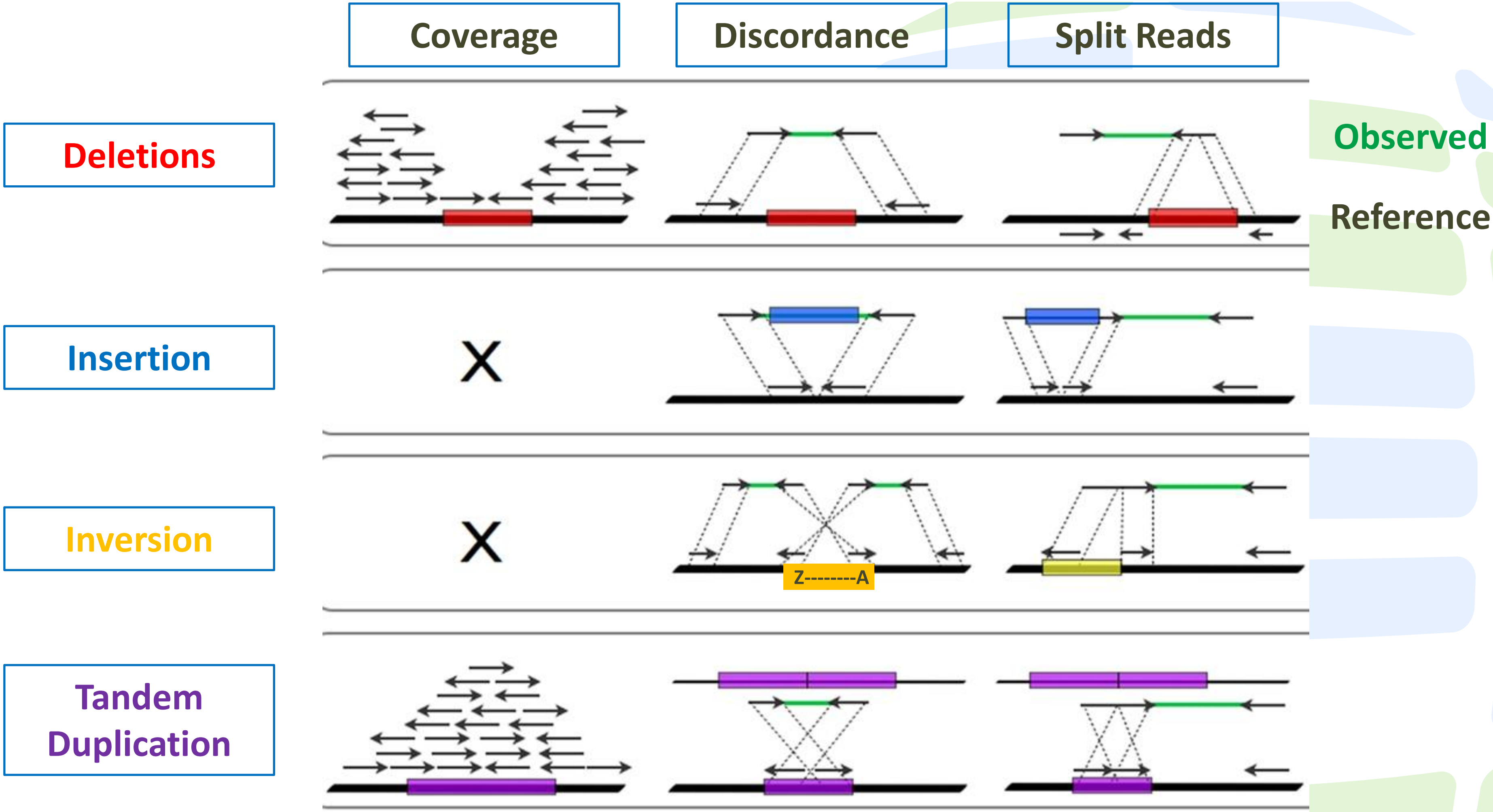
Calling DNA variants – SNVs, CNVs, SVs

1. What is a variant?
2. Paired End read mapping
3. Calling Single Nucleotide Variants (SNVs) and InDels
4. **Calling Copy Number Variants (CNVs)**
   - From Whole Genome Sequencing data
   - From Whole Exome Sequencing data
5. Calling Structural Variants (SVs)

Calling CNVs and SVs

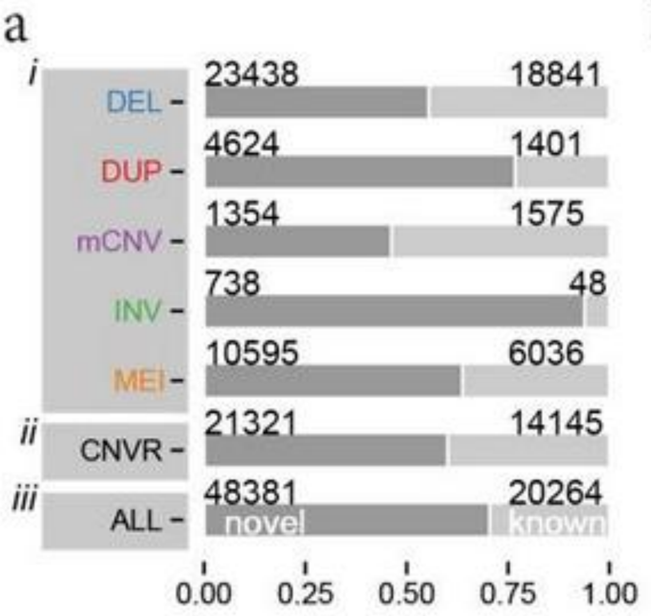
- There are **3 main classes of signal** that tools use when attempting to identify the presence of a Copy Number or Structural Variant
  - **Discordant Read Pair Mapping**
    - The gap between the two reads is significantly longer/shorter than expected ➔ Insertion or deletion respectively
    - The orientation of the reads is different from that expected ➔ inversion
  - **Split Read Mapping**
    - The ends of an **individual read** map to different locations
  - **Depth of Coverage (Read Count) Metrics**
    - The depth of coverage in a particular region is significantly more than, or less than expected ➔ copy number gain or copy number loss respectively

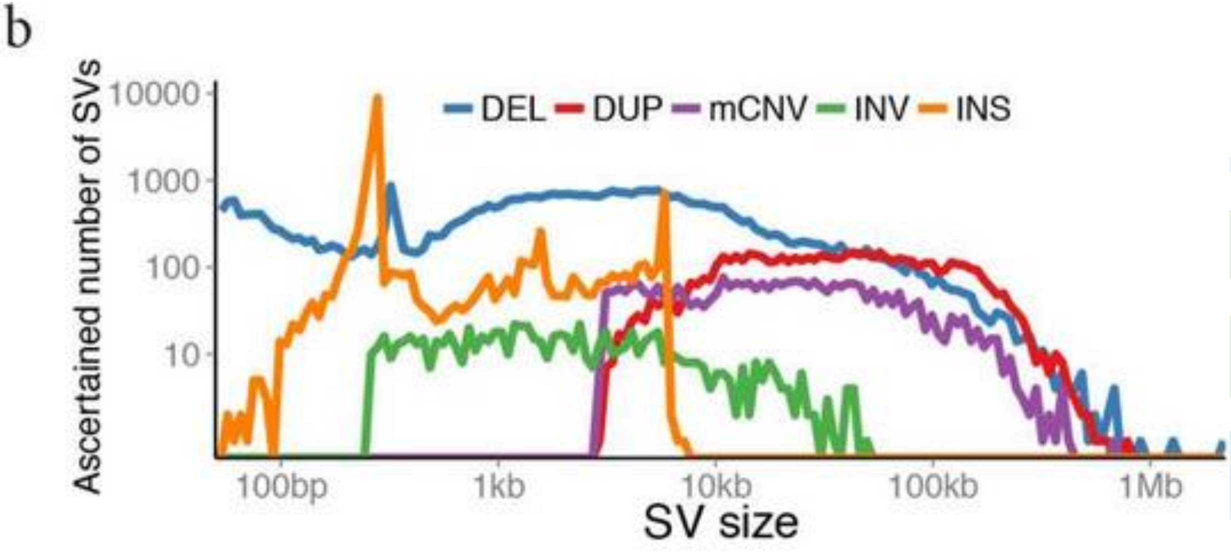
Calling CNVs and SVs – The Signals



Adapted from Tattini *et al*, 2015

Calling CNVs and SVs - overview





SV Class	Median Size	Median alleles	Median Kbp
DEL	2455	2788	5615
DUP	35890	17	518
mCNV	19466	340	11346
Inversion	1697	37	78
MEI	297	1218	691

Sudmant, P. *et al*. (2015)

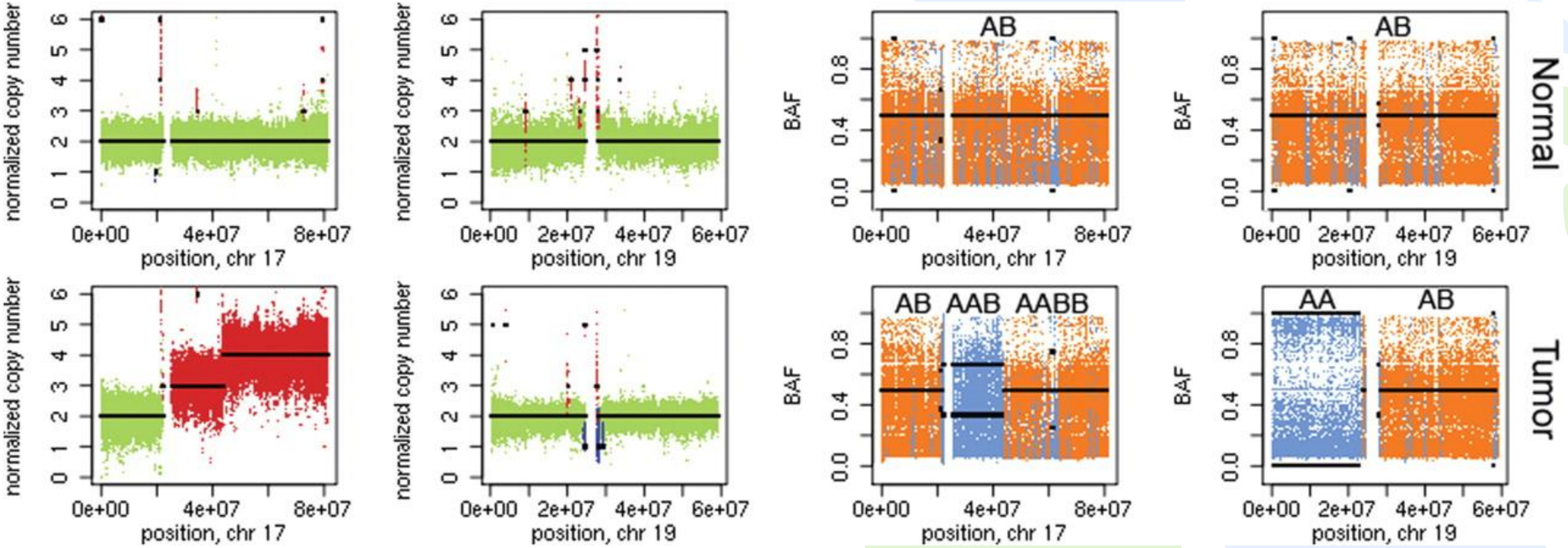
Calling CNVs from WGS data

- Popular tools include
  - cm.mops
  - CNVnator
  - Control-FreeC
  - Delly
  - ERDS
  - GenomeSTRiP
  - Lumpy

Calling CNVs from WGS data

- In general easier than for WES data
- Can typically be used on a single sample
- Account for sources of bias such as GC content, and low complexity regions
- Sensitive to stochastic coverage effects

Calling CNVs from WGS data

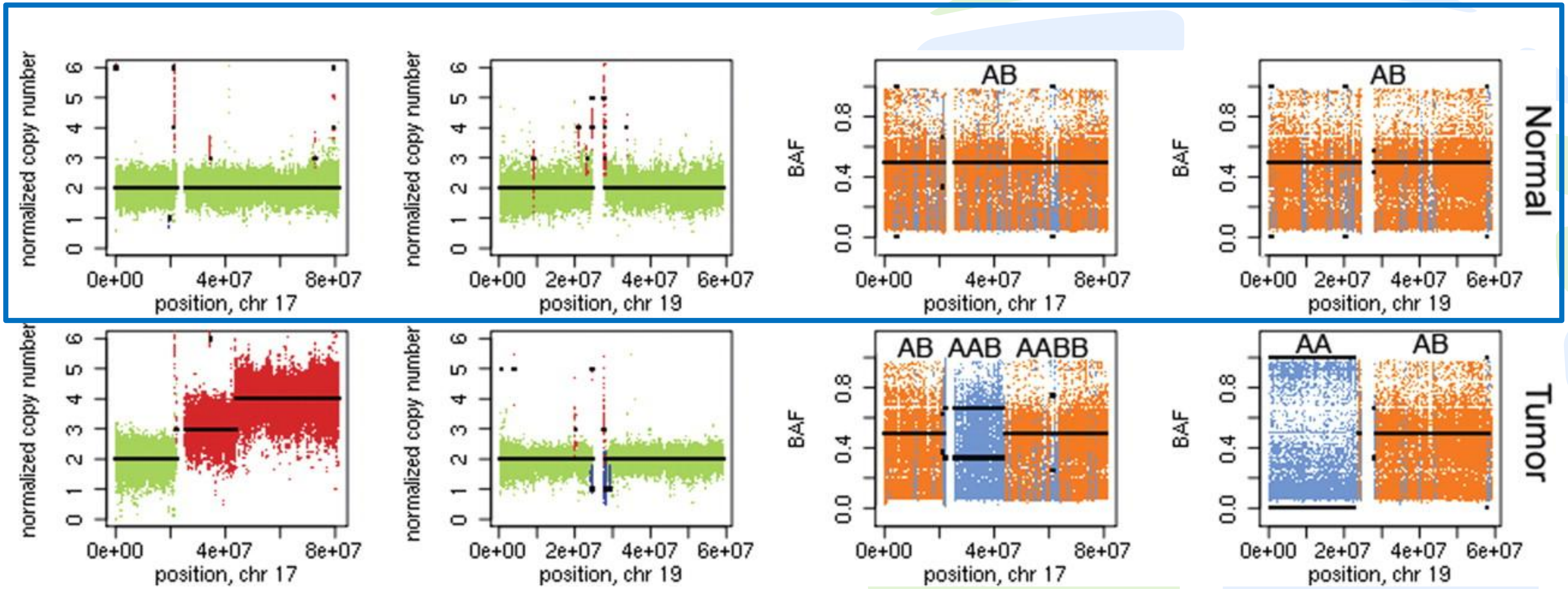


Normalised Copy Number

Beta Allele Frequency

Boeva, V. *et al*. (2012)

Calling CNVs from WGS data

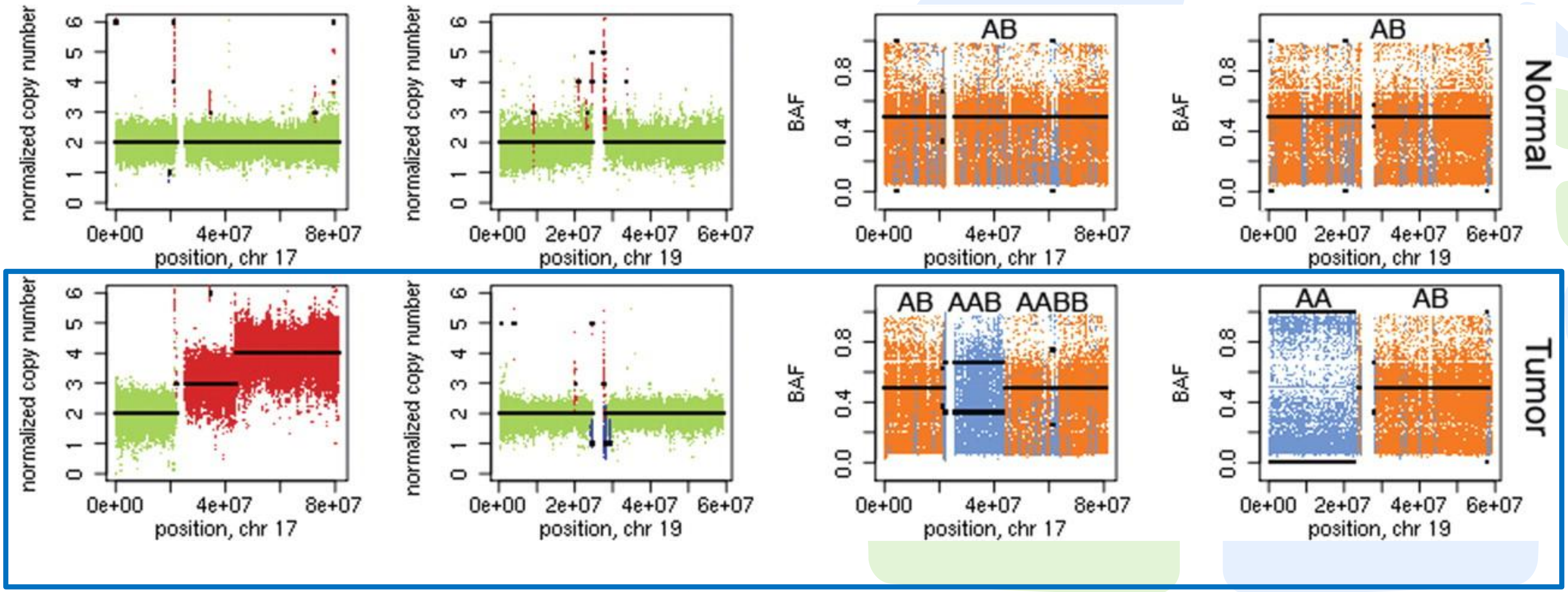


Normalised Copy Number

Beta Allele Frequency

Boeva, V. *et al*. (2012)

Calling CNVs from WGS data

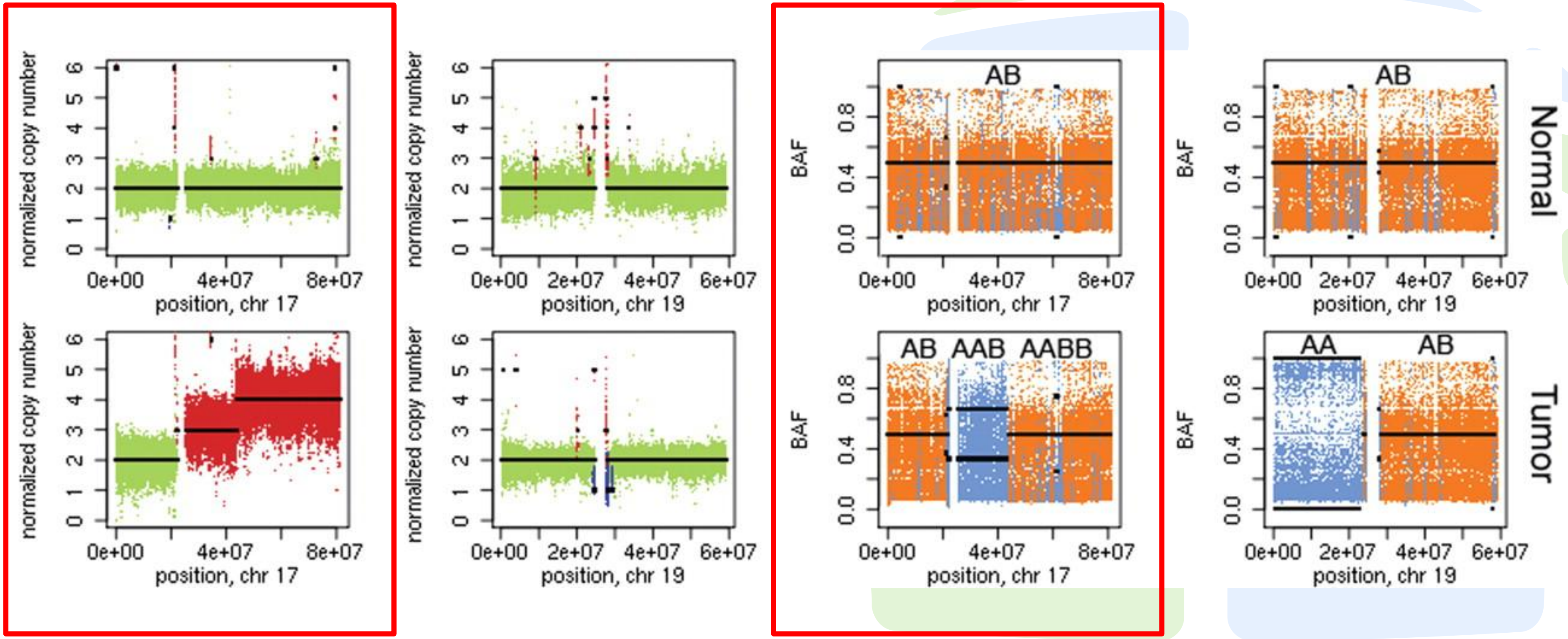


Normalised Copy Number

Beta Allele Frequency

Boeva, V. *et al*. (2012)

Calling CNVs from WGS data

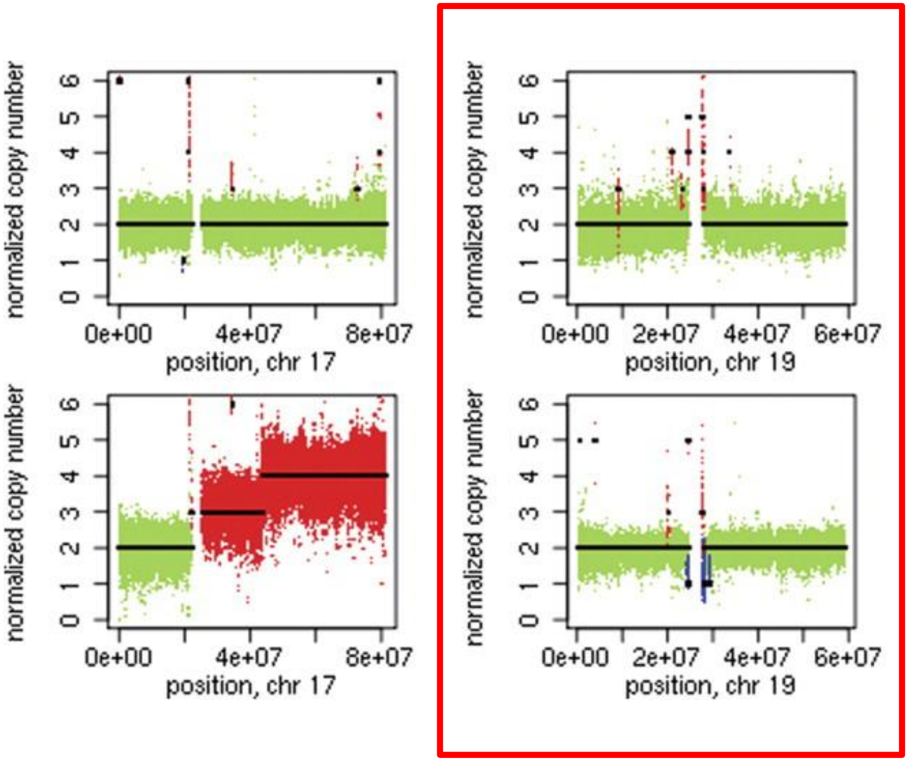


Normalised Copy Number

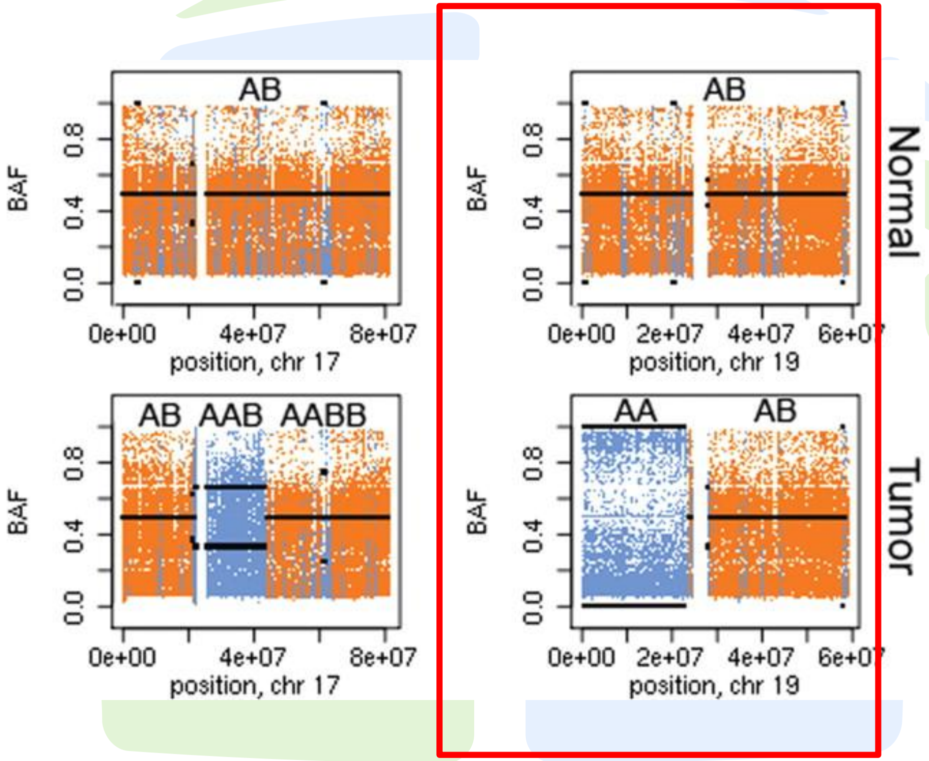
Beta Allele Frequency

Boeva, V. *et al*. (2012)

Calling CNVs from WGS data



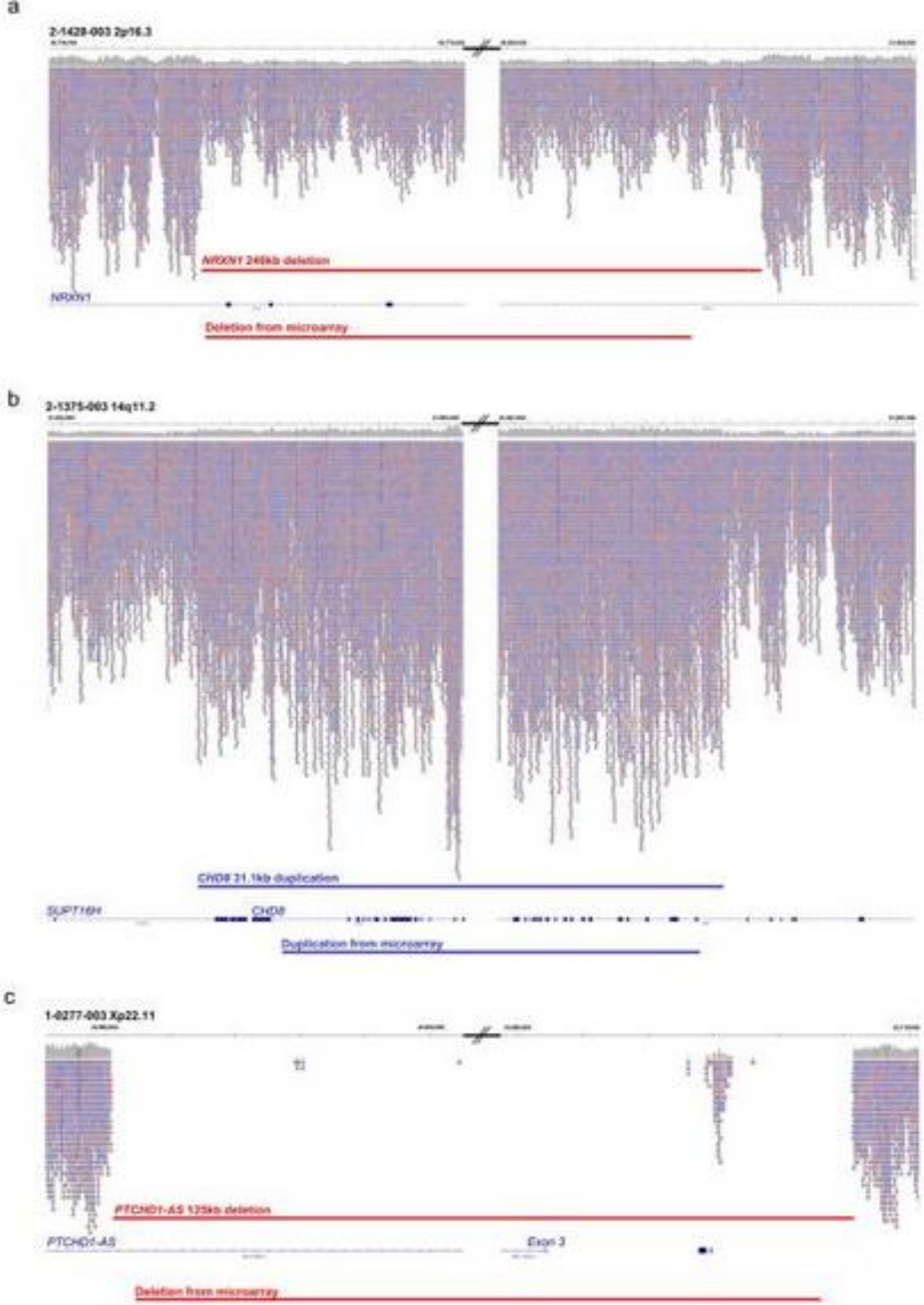
Normalised Copy Number



Beta Allele Frequency

Boeva, V. *et al*. (2012)

Calling CNVs from WGS data

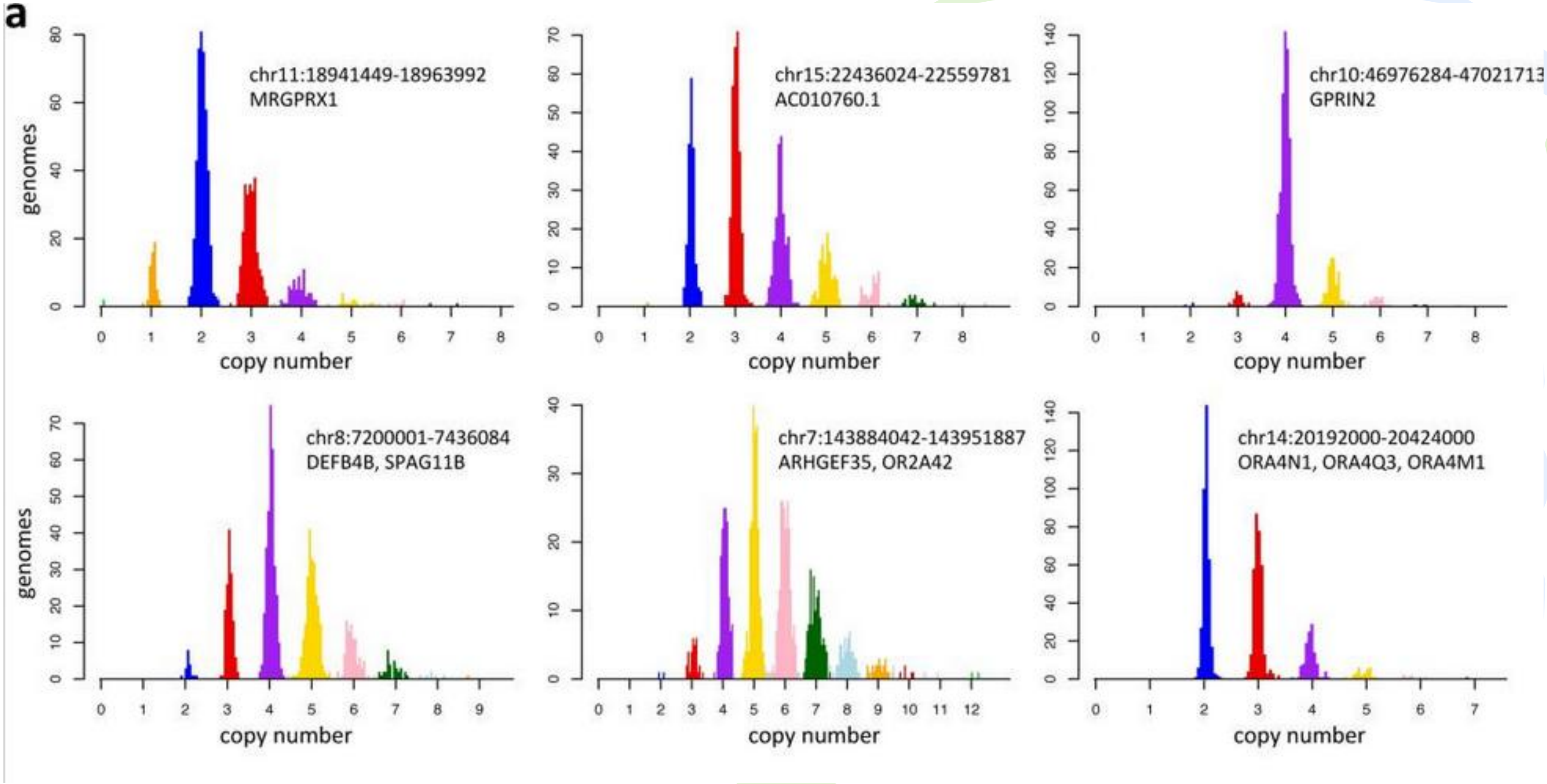


Yuen, RK *et al.* (2017)

centre nacional d'anàlisi genòmica
centro nacional de análisis genómico

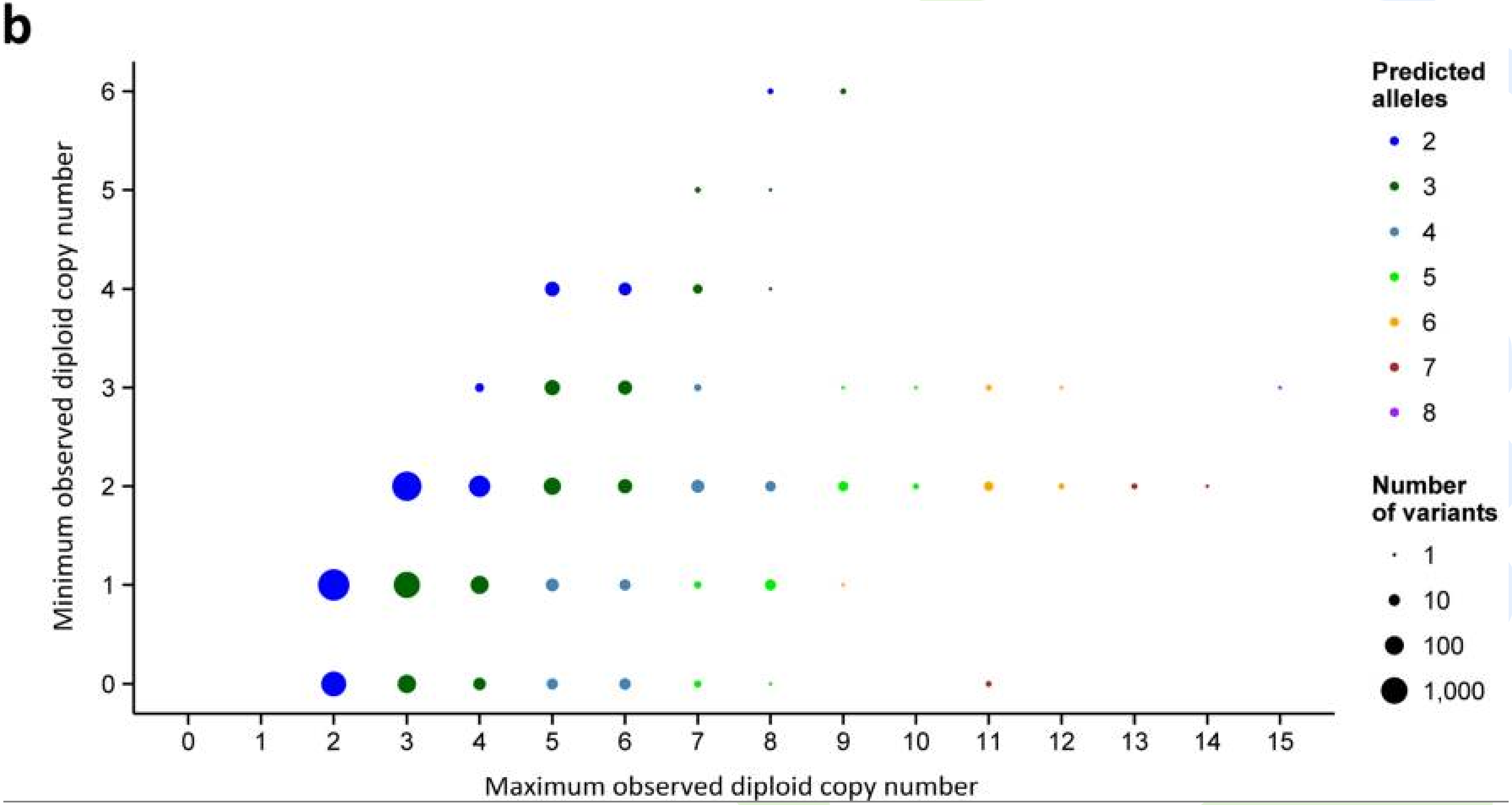
Centre for Genomic Regulation

mCNVs are segregating in the wild



Handsaker, RA *et al*. (2015)

mCNVs are segregating in the wild



Handsaker, RA *et al*. (2015)

Calling CNVs from WES data

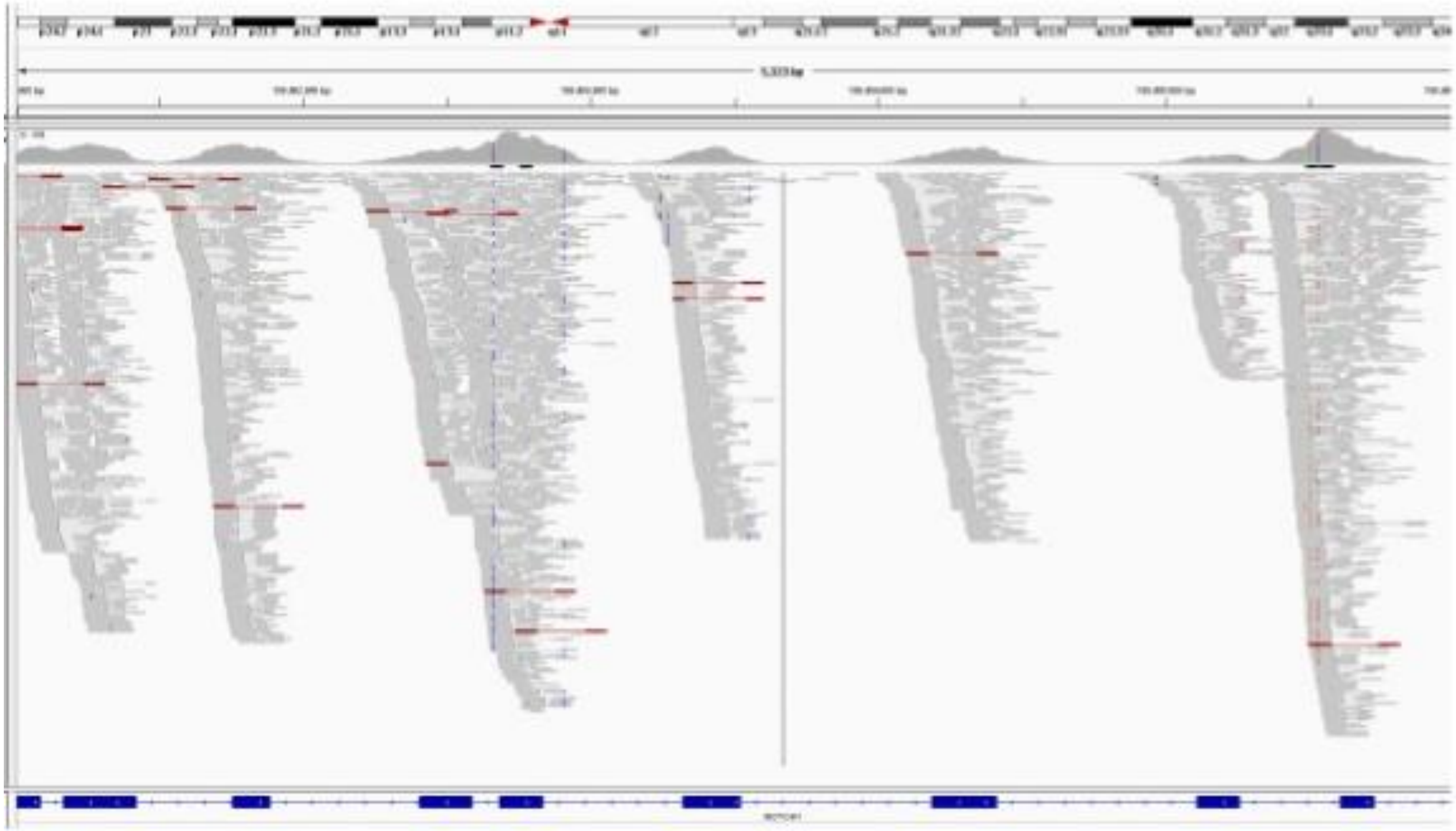
- Unlike WGS data, WES data is discontinuous meaning it is virtually impossible to detect large SVs, other than large deletions
- The discontinuity also makes it difficult, but not impossible to detect Copy Number Variants. Most tools require a minimum of 3 exons to be affected to make a reliable call
- Detection is further complicated by the fact that coverage is not uniformly distributed across the capture regions, with peaks in the middle, dropping off to the sides

Mapped reads viewed in IGV

Coverage

Reads

Exons



Calling CNVs from WES data

- Tools have to normalise, both horizontally, and vertically
  - Comparison to a reference set
  - Account for factors such as GC content, low complexity regions
  - Account for batch-type effects, by removing sources of extreme variance using PCA/SVD
- Make calls, typically using a Hidden Markov Model (HMM)
- Identify regions that appear significantly different in a specific sample when compared to the reference set

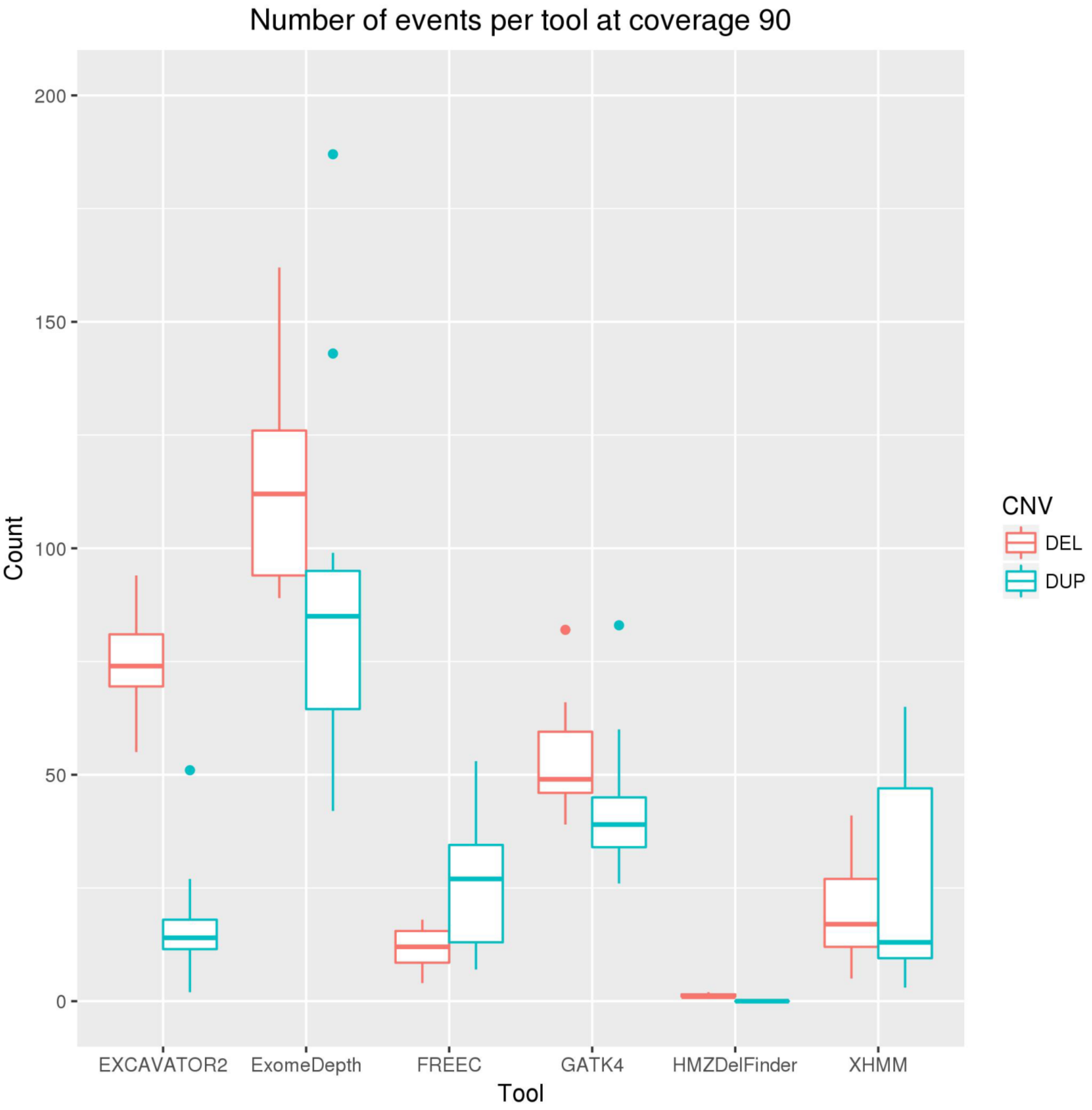
Calling CNVs from WES data

- Tools have to normalise, both horizontally, and vertically
  - Comparison to a reference set
  - Account for factors such as GC content, low complexity regions
  - Account for batch-type effects, by removing sources of extreme variance using PCA/SVD
- Make calls, typically using a Hidden Markov Model (HMM)
- Identify regions that appear significantly different in a specific sample when compared to the reference set
- Even when detected, we don’t know where they are

Calling CNVs from WES data

- Popular tools include
  - ExomeDepth (1 versus 10)
  - Conifer (All v All 8+)
  - XHMM (All v All – rare)
- Other notable options
  - Control-FreeC (ongoing development)
  - GATK-4 (Coming soon ...)
  - For all tools, **the more standardised your data, the better they will perform** i.e. Capture kit, sequencing depth, sequencing lab etc.

Calling CNVs from WES data



Sandra Rédo

Large Structural Variants – WGS

- Popular tools include
  - BreakDancer
  - cm.mops
  - CNVnator
  - Control-FreeC
  - Delly
  - ERDS
  - GenomeSTRiP
  - Lumpy
  - Pindel

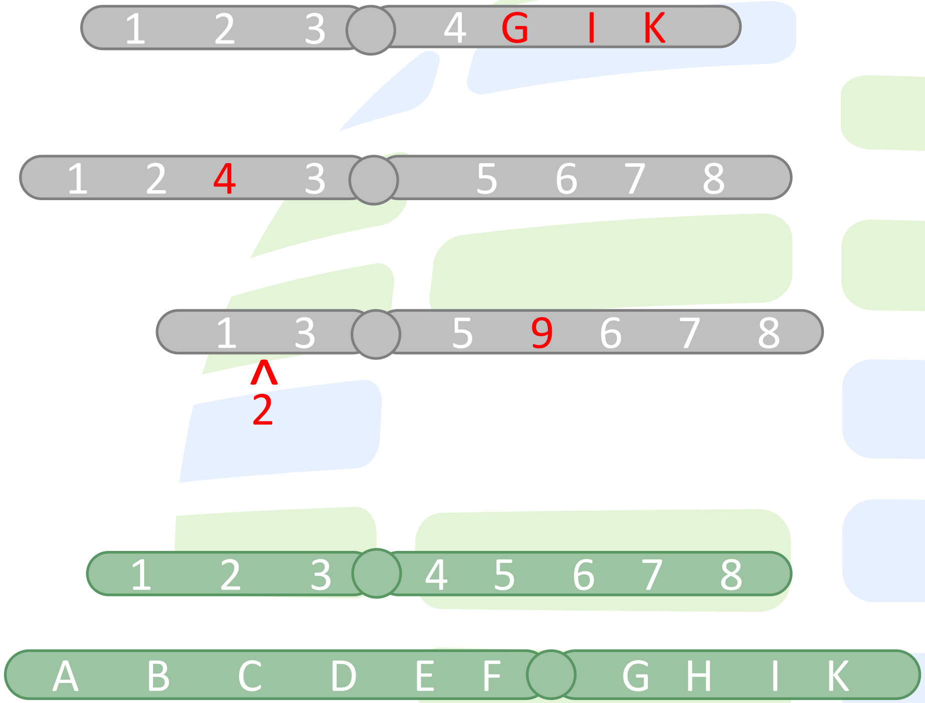
Large Structural Variant Classes

Translocation

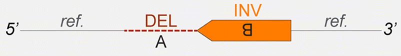
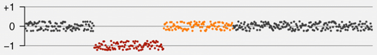
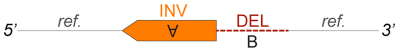
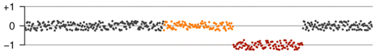
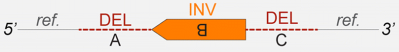
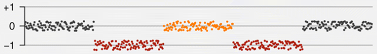
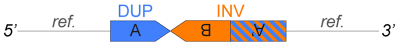
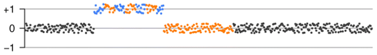
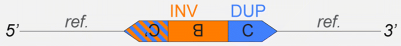
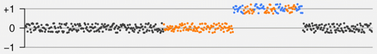
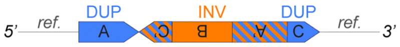
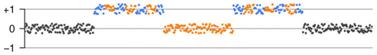
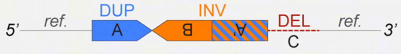
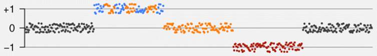
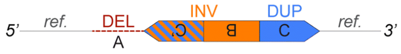
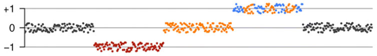
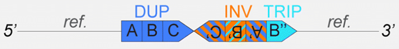
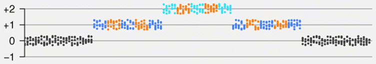
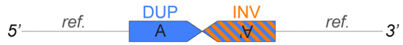
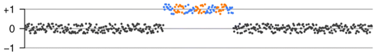
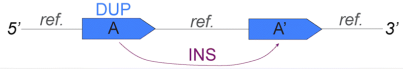
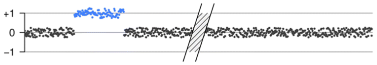
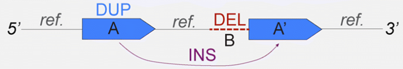
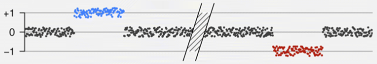
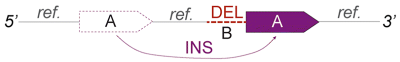
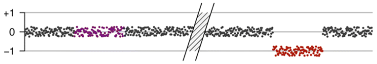
Inversion

Large Insertions and Deletions

Reference Chromosome



In principle should be easy – lots of signal ☺

cxSV Subclass	Abbreviation	SVs	Validation Rate	Total Observations	Subjects with ≥ 1 SV	Median Size (kb)	DEL DUP INS INV	Rearrangement Schematic	Simulated Copy Number Profile
Inversion with 5' Flanking Deletion	delINV	38	100% (19/19)	2,301	99.7%	12.2	🟥 🟧		
Inversion with 3' Flanking Deletion	INVdel	40	100% (19/19)	2,242	100.0%	9.3	🟥 🟧		
Paired-Deletion Inversion	delINVdel	58	96% (25/26)	2,288	98.0%	15.2	🟥 🟧	 5’ ref. DEL A INV B DEL C ref. 3’	
Inversion with 5' Flanking Duplication	dupINV	7	75% (3/4)	62	8.7%	54.6	🟦 🟧		
Inversion with 3' Flanking Duplication	INVdup	3	100% (1/1)	6	0.9%	82.9	🟦 🟧		
Paired-Duplication Inversion	dupINVdup	45	96% (27/28)	151	19.0%	112.5	🟦 🟧	 5’ ref. DUP A C INV B A DUP C ref. 3’	
Inversion with 5' Flanking Duplication and 3' Flanking Deletion	dupINVdel	6	100% (2/2)	9	1.3%	27.3	🟥 🟦 🟧	 5’ ref. DUP A INV B A DEL C ref. 3’	
Inversion with 5' Flanking Deletion and 3' Flanking Duplication	delINVdup	10	100% (5/5)	90	12.2%	67.5	🟥 🟦 🟧	 5’ ref. DEL A INV C B DUP C ref. 3’	
Inverted Duplication with Flanking Triplication	dupTRIPdup-INV	5	100% (5/5)	5	0.7%	113.9	🟦 ◆ 🟧	 5’ ref. DUP A B C INV TRIP C B A B” ref. 3’	
Inverted Repeat / Inverted Tandem Duplication	IR	11	88% (7/8)	36	5.1%	73.8	🟦 🟧		
Compound CNV	cpdCNV	22	100% (17/17)	2,085	99.4%	29.1	🟥 🟦	*Various*	
Dispersed Duplication	dDUP	10	100% (2/2)	42	6.1%	17.5	🟦 🟪		
Dispersed Duplication with Deletion	dDUPdel	9	100% (4/4)	60	8.5%	32.1	🟥 🟦 🟪	 5’ ref. DUP A ref. DEL B A’ ref. 3’ INS	
Insertion with Deletion	INSdel	4	100% (1/1)	12	1.7%	5.9	🟥 🟪	 5’ ref. A ref. DEL B A ref. 3’ INS	
Compound Insertion	cpdINS	5	100% (2/2)	251	36.0%	3.1	🟪 🟧	*Various*	
Compound Insertion with Deletion	cpdINSdel	1	NA (0/0)	5	0.7%	9.6	🟥 🟪 🟧	*Various*	
Compound/Complex Rearrangement (or Other)	CCR	15	100% (11/11)	21	3.1%	239.8	*Various*		
All cxSV	-	289	97% (150/154)	9,666 (14/subject)	100.0%	27.3	DEL: 61.8% DUP: 47.8% INV: 84.8% INS: 11.8%		

Collins, RL *et al*. (2017)

Large Structural Variants – WGS

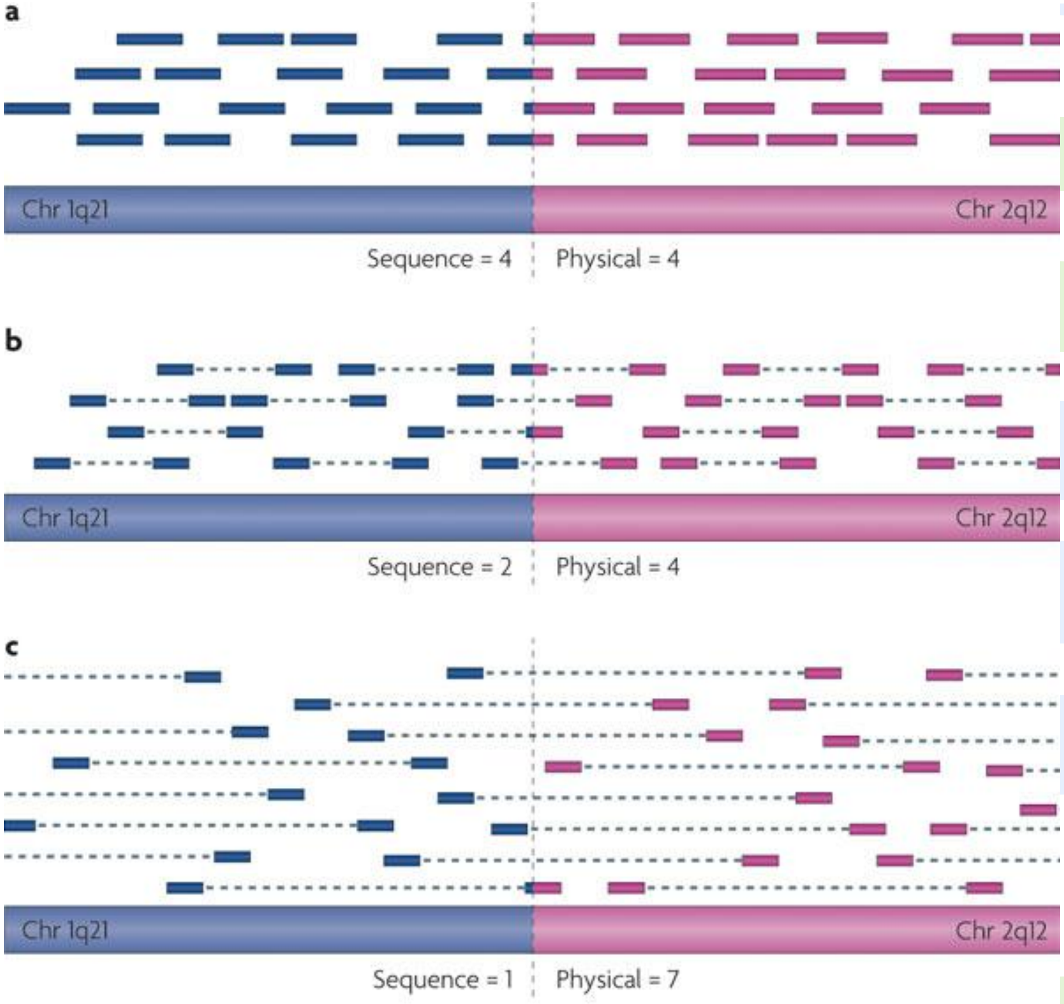
- Most tools have been tailored to best identify specific classes of SV
  - Therefore may want to use more than one tool
- More recently developed tools tend to look at more than one type of evidence, and thus can call different classes
- To optimise discovery of SVs, ideally want to use **a mix of library strategies** and/or technologies i.e. short-read and long-read simultaneously

Large Structural Variants – WGS

Single End

Paired End

"Mate Pairs"



Other interesting topics

- Assembly approaches to structural variant detection
- Long read technologies e.g. PacBio and Oxford Nanopore
- Somatic variant calling
- Balance cytogenic abnormalities

Acknowledgements

cnag
centre nacional d'anàlisi genòmica
centro nacional de análisis genómico

CRG Centre for Genomic Regulation

Management and Lab

I. Gut
M. Gut
M. Bayès
B. Fusté
Lidia Aguade
Julie Blanc
CNAG lab
CNAG admin

Data Analysis

S. Beltran
R. Tonda
M. Fernandez-Callejo
J.R. Trotta
J. Camps
S. Marco-Sole
S. Redó

If you would like to join RD-Connect, please contact **platform@rd-connect.eu**

RD-Connect: http://rd-connect.eu/ @ConnectRD

Platform: https://platform.rd-connect.eu/

Other sequencing and data analysis projects:
projectmanager@cnag.crg.eu

steven.laurie@cnag.crg.eu @SteveLaurie42