Next-generation sequence characterization of complex genome structural variation

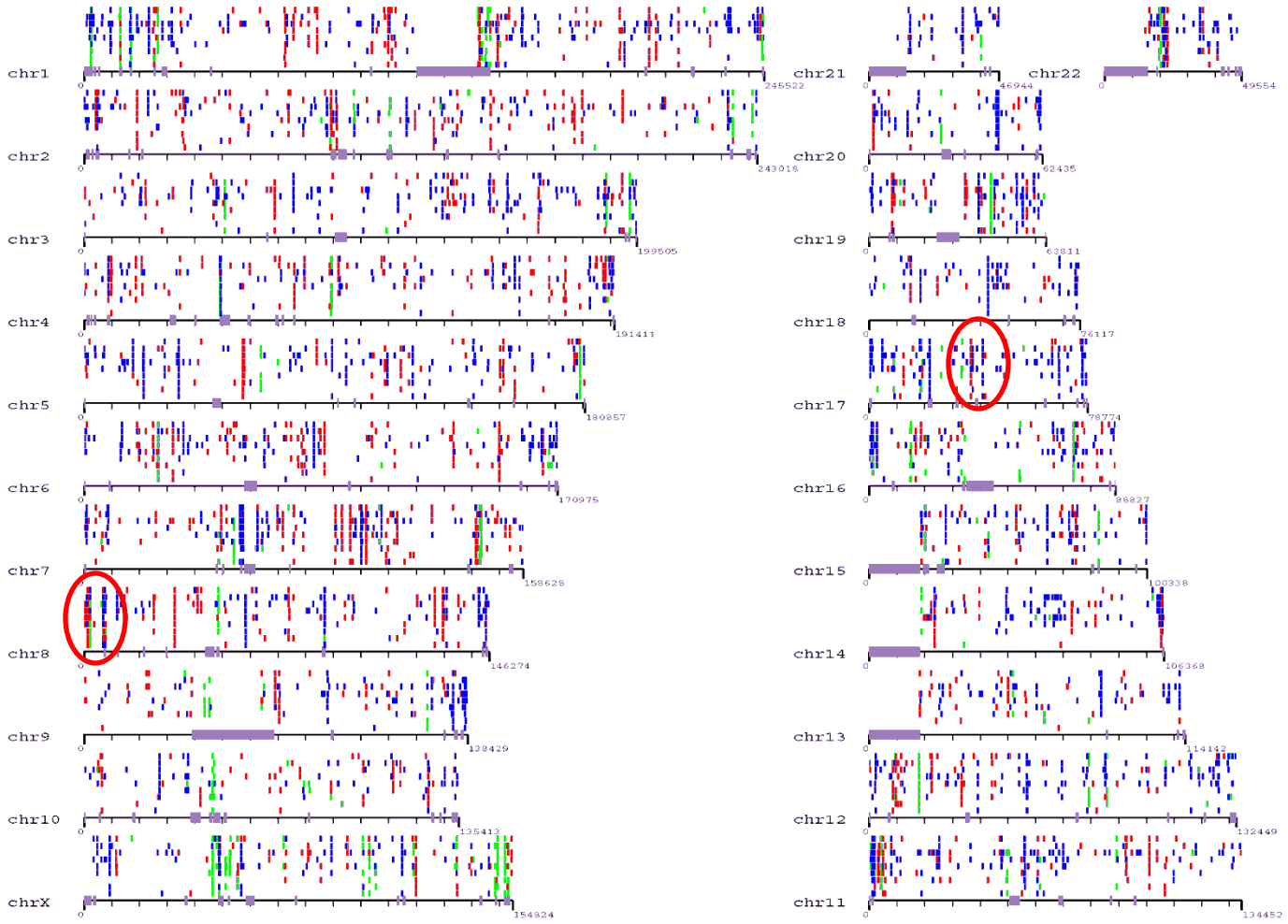
Can Alkan
Department of Genome Sciences,
University of Washington
Seattle, WA, USA

Size range of genetic variation

- Single nucleotide (SNPs)
- Up to ~50bp (small indels, microsatellites)
- >50bp to several megabases (**structural variants)**:
  - Deletions
  - Insertions
    - Novel sequence
    - Mobile elements (*Alu*, L1, SVA, etc.)
  - Segmental Duplications
    - Duplications of size ≥ 1 kbp and sequence similarity ≥ 90%
      - Tandem or Interspersed
  - Inversions
  - Translocations
- Chromosomal changes

CNVs

Human genome structural variation

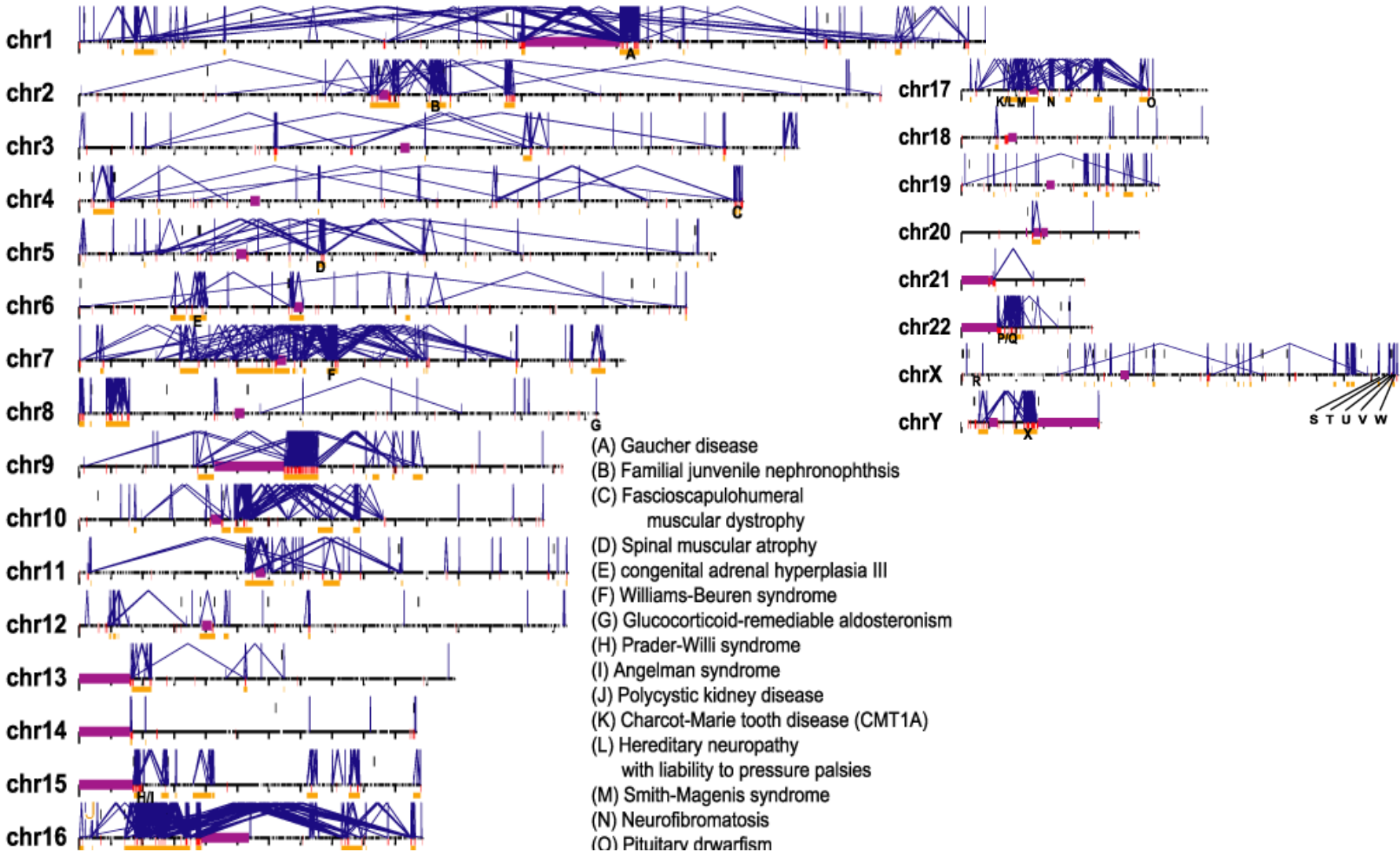


Insertion
Deletion
Inversion
Gaps

ABC14 (CEPH)
ABC13 (Yoruba)
ABC12 (CEPH)
ABC11 (China)
ABC10 (Yoruba)
ABC9 (Japan)
ABC8 (Yoruba)
ABC7 (Yoruba)
G248

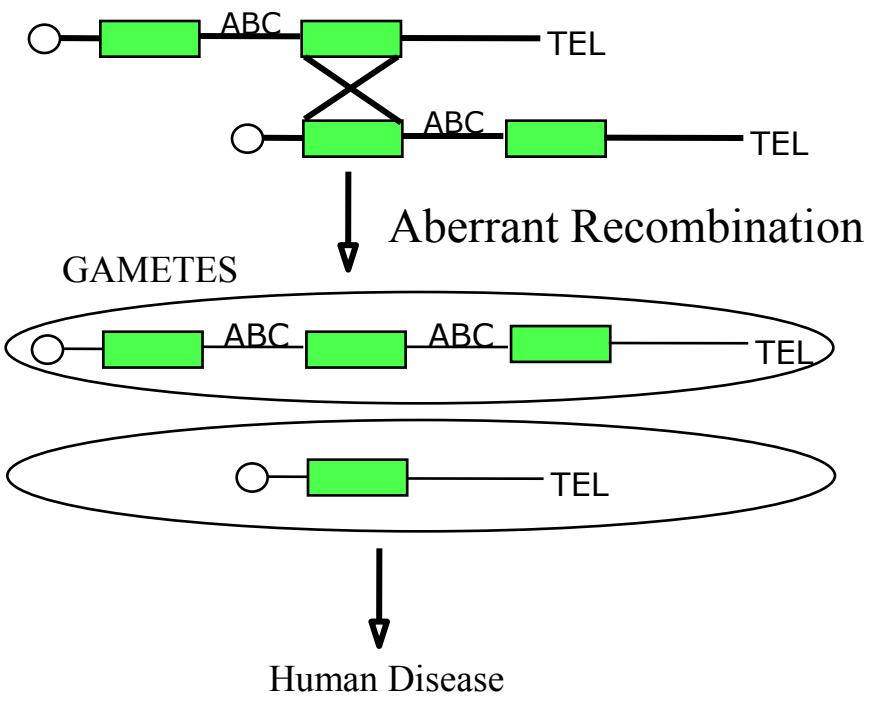
Kidd et al., Nature, 2008

Duplications and CNV hotspots



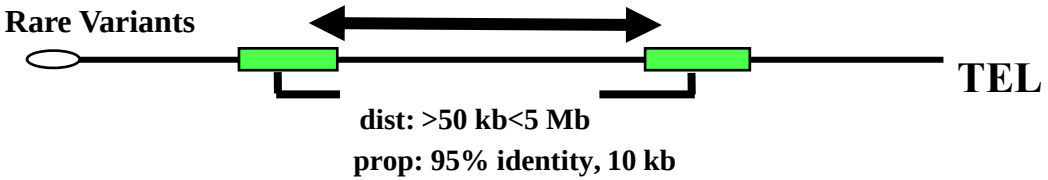
Bailey et al., Science, 2002

CNVs and disease



Triplosensitive, Haploinsufficient and Imprinted Genes

- 17q21.31 Deletion (Sharp et al. *Nat Gen* 2006)
  - ~1% mental retardation in European populations
- 15q24.1 Deletion (Sharp et al. *HMG* 2007)
  - autism spectrum disorder/MR/growth deficiency
- 17q12 Deletion (Mefford et al. *AJHG* 2007)
  - 20% of patients with renal disease due to hypodysplastic kidneys and 36% of children with MODY-5
- 1q21.1 Deletion (Mefford et al. *NEJM* 2008)
  - 0.4% of mental retardation and 0.3% schizophrenia.



Multi-copy CNP and disease

CNP	SD/SV (kb / % identity)	Variation	Disease/Effect	
C4A/C4B	32.8 / 99.1	decrease	lupus	Yang, 2007
DEFB4.103,104	310 / 99.4	increase	psoriasis	Hollox, 2008
		decrease	Crohn's disease	Fellerman, 2006
CCL3L1	64 / 99.8	decrease	HIV susceptibility	Gonzalez, 2005
FCGR3B	**	decrease	glomerulonephritis	Aitman, 2006; Fanciulli, 2008
IRGM	**	deletion	Crohn's disease	McCarroll, 2008
			alternative splicing	Bekpen, 2009

** correspond to more ancient primate segmental duplications

- Previous work largely array based
  - Biased against duplicated regions due to signal saturation
  - Limited to CNVs (no balanced events)
  - Give only (relative) copy number
    - No information on variant sequence or organization
- **Goal: systematically discover and characterize genomic structural variation and *copy*, *content*, and *structure* of segmental duplications.**

Genome-wide SV Discovery Approaches

Hybridization-based

- Iafrate et al., 2004, Sebat et al., 2004
- SNP microarrays: McCarroll *et al*., 2008, Cooper *et al.*, 2008, Itsara *et al.*, 2009
- Array CGH: Redon *et al.* 2006, Conrad *et al.,* 2010, Park *et al.,* 2010, WTCCC, 2010

Single molecule analysis

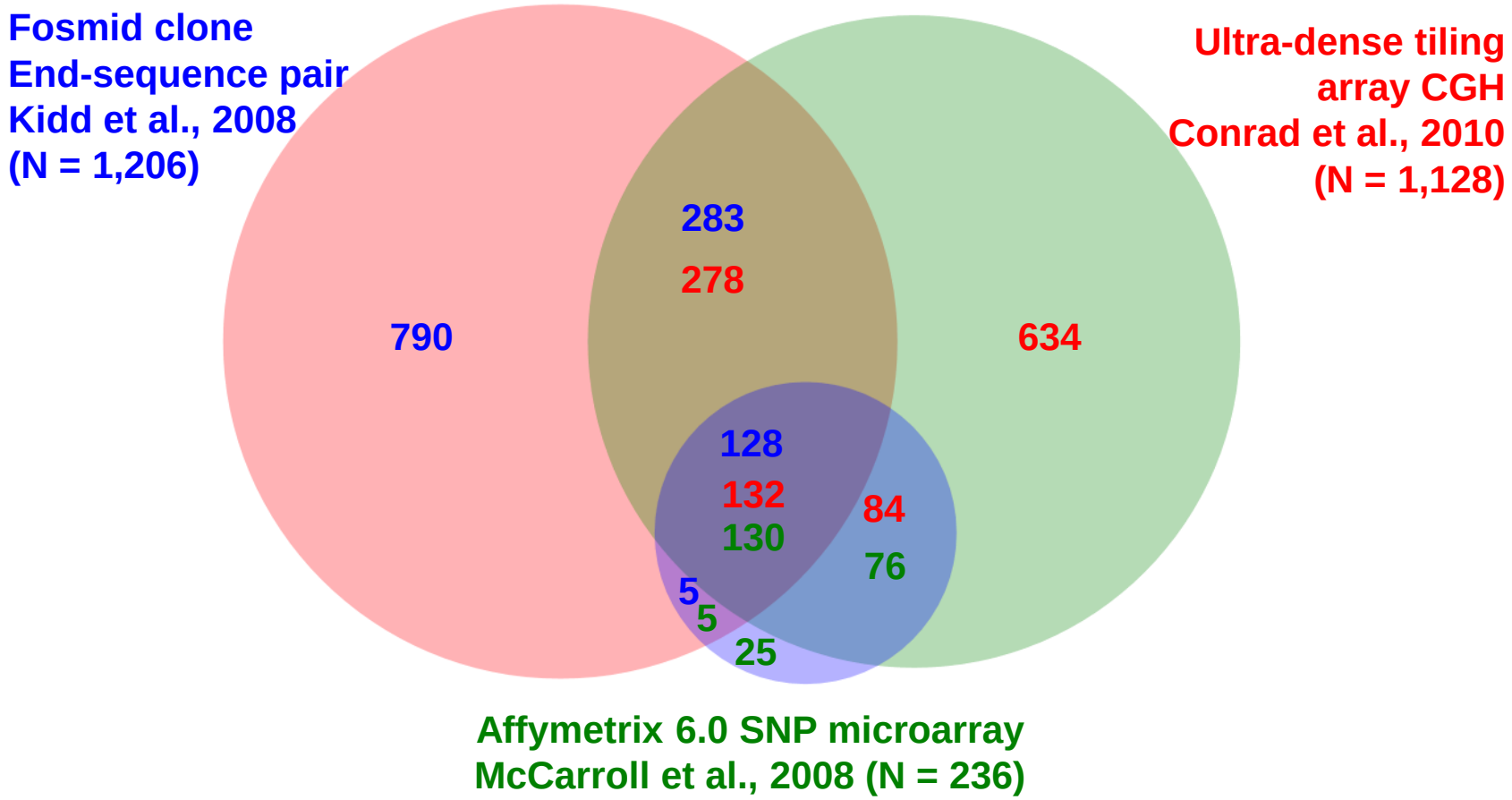
- Optical mapping: Teague et al., 2010

Sequencing-based

- Read-depth: Bailey et al, 2002
- Fosmid ESP: Tuzun *et al.* 2005, Kidd *et al.* 2008
- Sanger sequencing: Mills *et al.*, 2006
- Next-gen sequencing: Korbel *et al.* 2007, Yoon *et al.*, 2009, Alkan et al., 2009, Hormozdiari *et al.* 2009, Chen *et al.* 2009,
  - 1000 Genomes Project

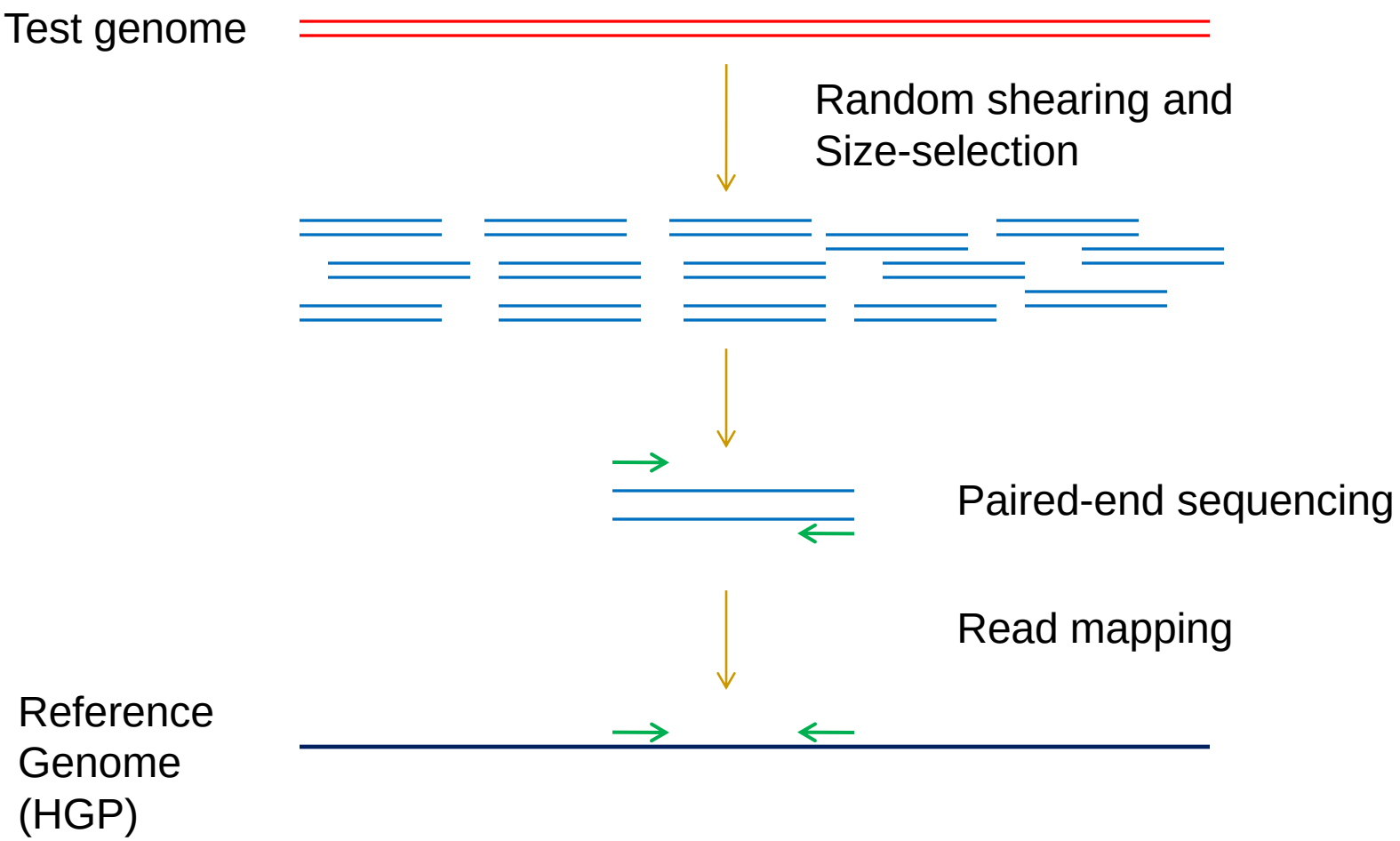
Detection diversity

Gains & Losses > 5 Kbp in the same 5 individuals



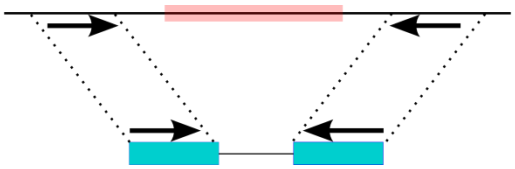
Kidd et al. Cell, 2010

Resequencing Genomes

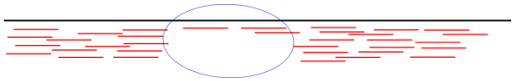


Sequence signatures of structural variation

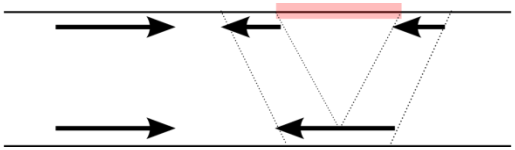
- Read pair analysis
  - Deletions, small novel insertions, inversions, transposons
  - Size and breakpoint resolution dependent to insert size



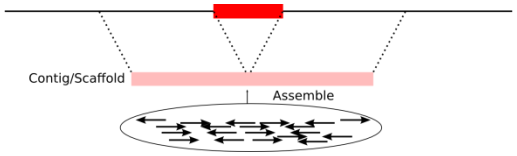
- Read depth analysis
  - Deletions and duplications only
  - Relatively poor breakpoint resolution



- Split read analysis
  - Small novel insertions/deletions, and mobile element insertions
  - 1bp breakpoint resolution



- Local and *de novo* assembly
  - SV in unique segments
  - 1bp breakpoint resolution



SV callers in the 1000 GP

Read Pair
U. Washington: VariationHunter (Illumina)
WUGSC: BreakDancer (Illumina)
ABI/LifeTech: Corona Light (SOLiD)
Yale: PEMer (Roche/454)
Wellcome-Trust: (unnamed) (Illumina)
BGI: (unnamed) (Illumina)

Read Depth
U. Washington: WSSD (Illumina)
Yale: CNVnator (Illumina)
UCSD/CSHL: EWT/RDXplorer (Illumina)
AECOM: (unnamed) (Illumina

Split Read
Leiden/WTSI: Pindel (Illumina)
Yale: (unnamed) (Roche/454)

Read Pair + Depth
Boston College: Spanner (Illumina)
Broad: Genome STRiP (Illumina)

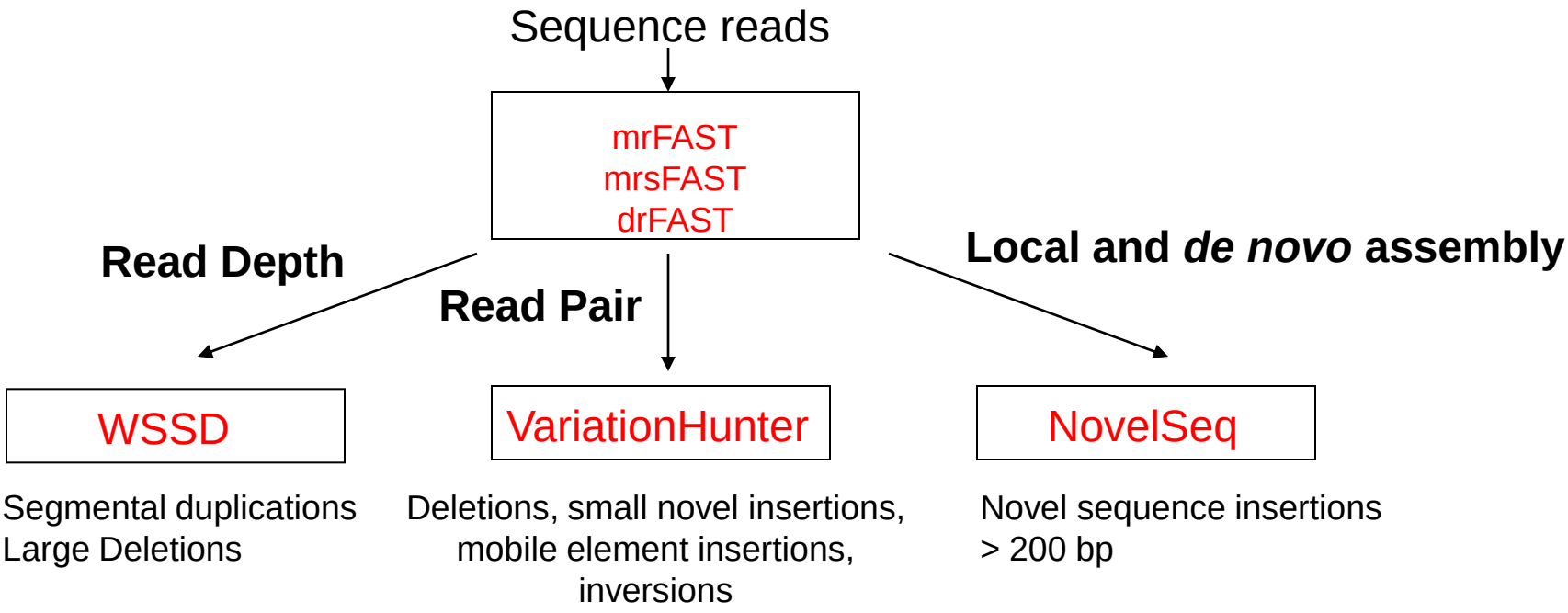
Assembly
U. Washington: NovelSeq (Illumina)
BGI: SOAPdenovo (Illumina)
EBI: Cortex (Illumina)

Breakpoint assembly
WUGSC: TIGRA (Illumina)

Upcoming from 1000Genomes: Delly/Invy (EMBL Heidelberg) + updates to the above
Non-1000G: MoDIL, MOGuL, CommonLAW, CNVer + GASV + HYDRA + *many more*
See Alkan et al., Nat Rev Genet, 2011, Mills et al. Nature 2011, and Medvedev et al., Nat Methods, 2009

Challenges and algorithms for NGS

- Short read mapping artifacts:
  - Discard non-unique: lose sensitivity and detection power.
  - Try to utilize all information: higher sensitivity, less specificity.
    - *micro-read Fast Alignment Search Tools*



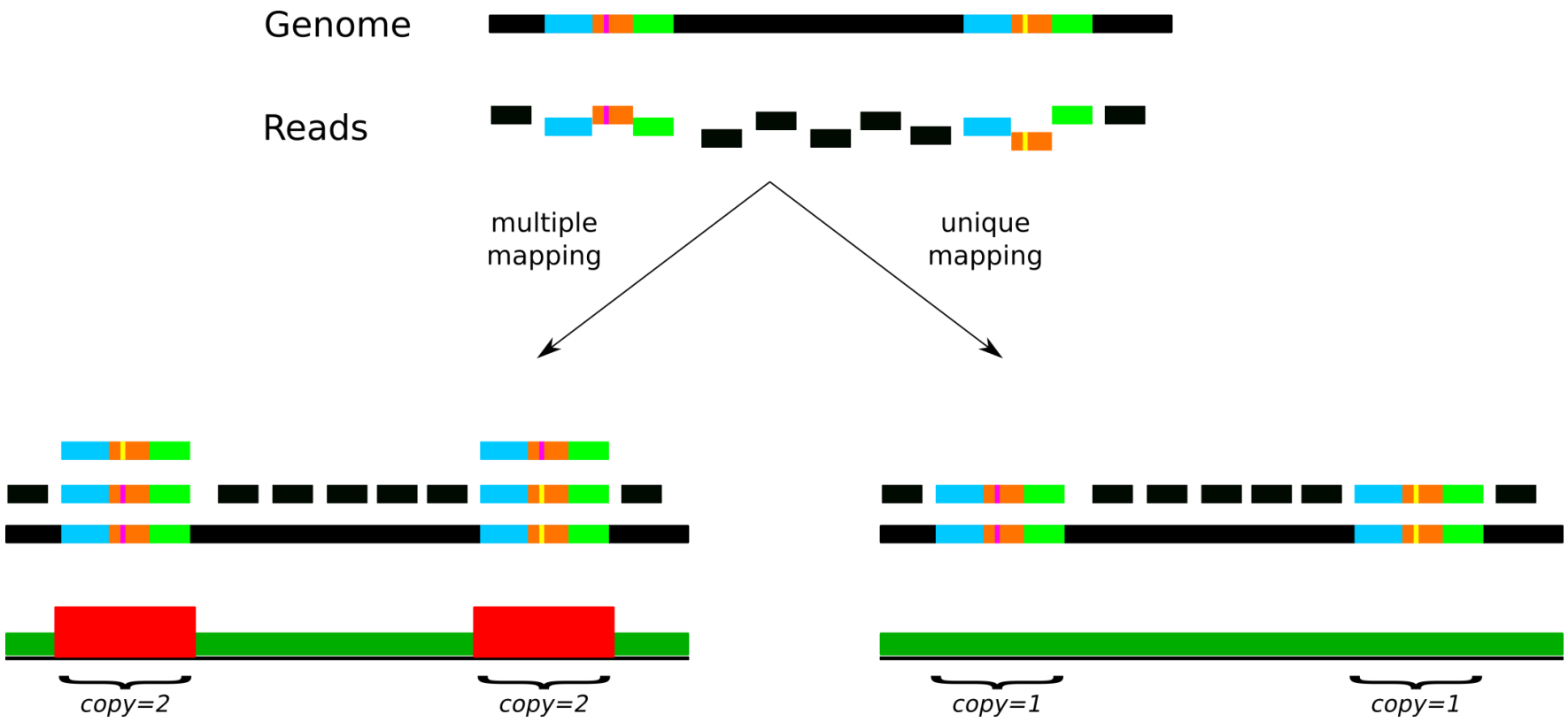
[mr(s)|dr]FAST algorithms

- *mr(s)FAST*: Designed for reads generated with the Illumina platform
  - ***mrFAST:*** Small indels up to 4 bp are supported
  - ***mrsFAST:*** Mismatch-only version for increased mapping speed
- *drFAST:* Di-base (color-space) read version for the SOLiD platform (Hamming distance only)
- Guaranteed sensitivity within user-specified edit (Hamming for mrsFAST and drFAST) distance threshold *d*:
  - $d \leq (read\_length / k) - 1$; [k=12]
- **Iterative search:** ***All*** locations and underlying sequence variation are returned
  - "best" mapping option for single-location mapping available; minimize edit distance & paired-end span size closest to median

http://mrfast.sourceforge.net
http://drfast.sourceforge.net

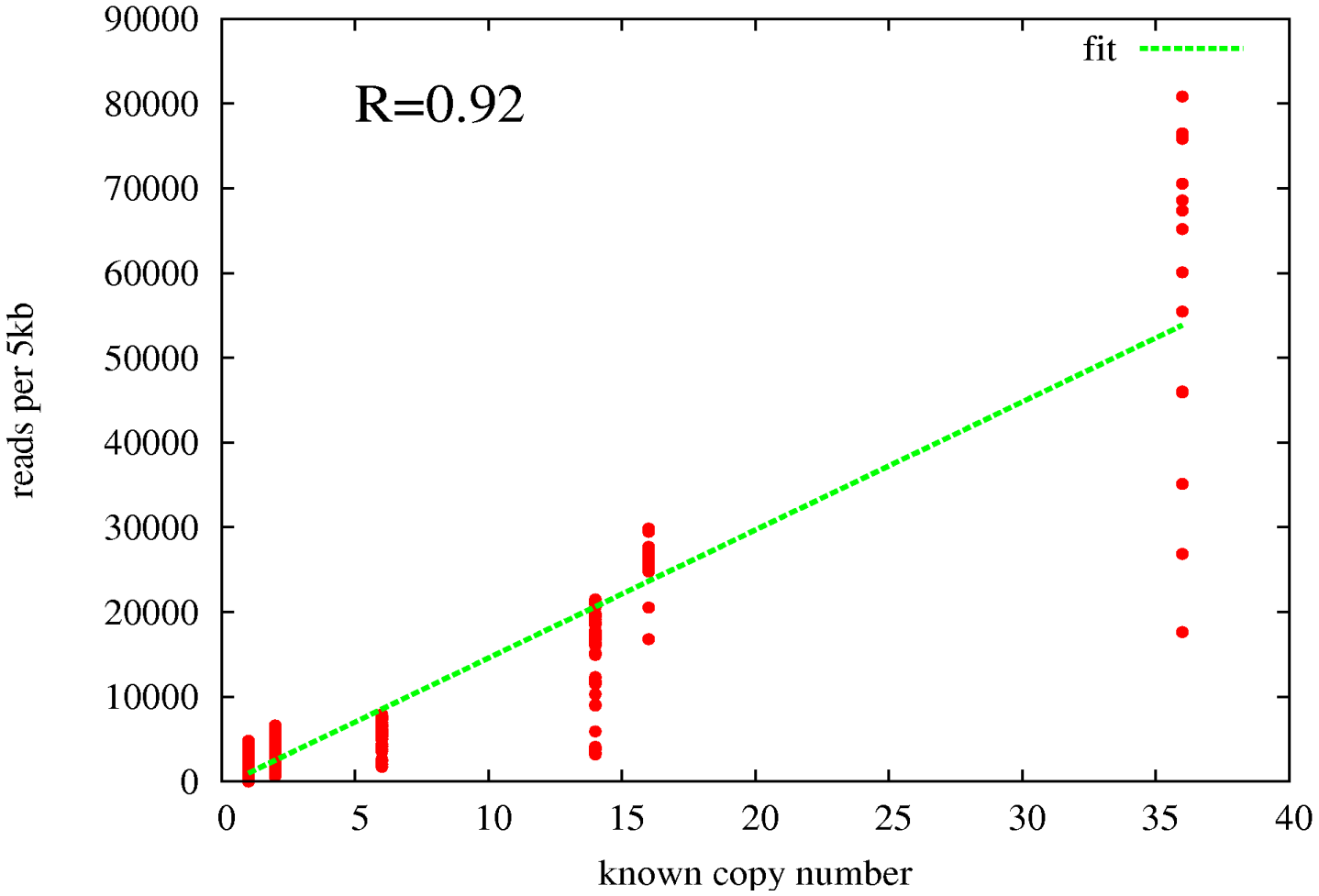
Alkan et al., Nature Genetics 2009; Hach et al., Nature Methods 2010; Hormozdiari et al., Bioinformatics 2011

Multiple vs. unique mapping



Modified from Chiang & McCarroll, Nat Biotech, 2009

1) Read depth - copy number correlation



Alkan et al., Nature Genetics, 2009

Personal duplication maps

- Next-gen specific error correction (GC% normalization, resolving repeats)

Individual	#WGS Reads	Read Length	Technology	
Jim Watson	74,198,831	266bp*	454 FLX	Wheeler 2008
HapMap NA18507 (YRI)	1,776,928,308	36bp	Illumina	Bentley 2008
YH (Han Chinese)	1,315,249,404	35bp	Illumina	Wang 2008

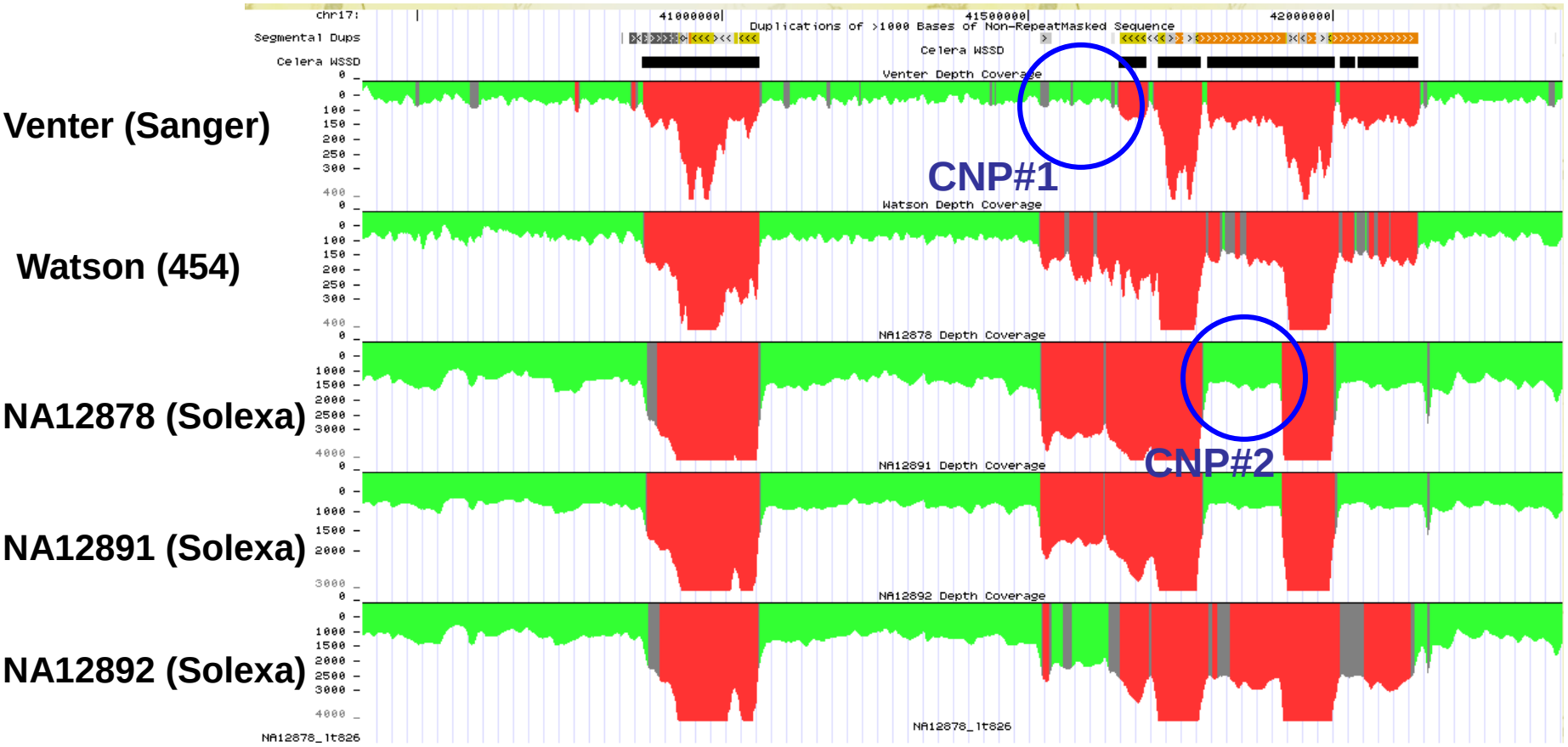
* Rendered into 509,667,772 reads of length 36bp

- Copy numbers predicted in 1kb non-overlapping windows
- 3.8% (662/17601) genes show a difference of at least one copy
  - 113/662 validated with arrayCGH
  - Most others are in high-copy regions, biased against aCGH
- Absolute copy number counts and characterization of the content of duplications made possible for the **first time**

http://mrcanavar.sourceforge.net

Alkan et al., Nature Genetics, 2009

Personal duplication maps

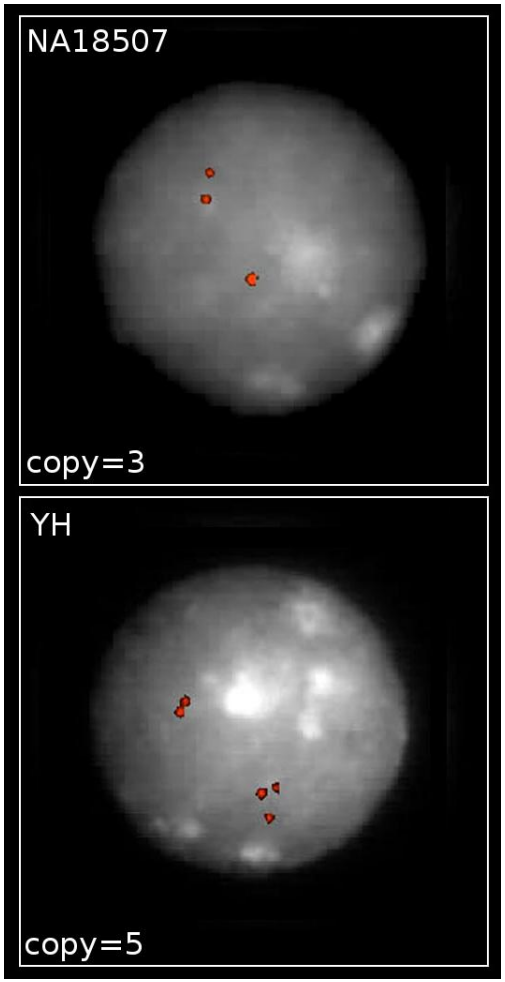


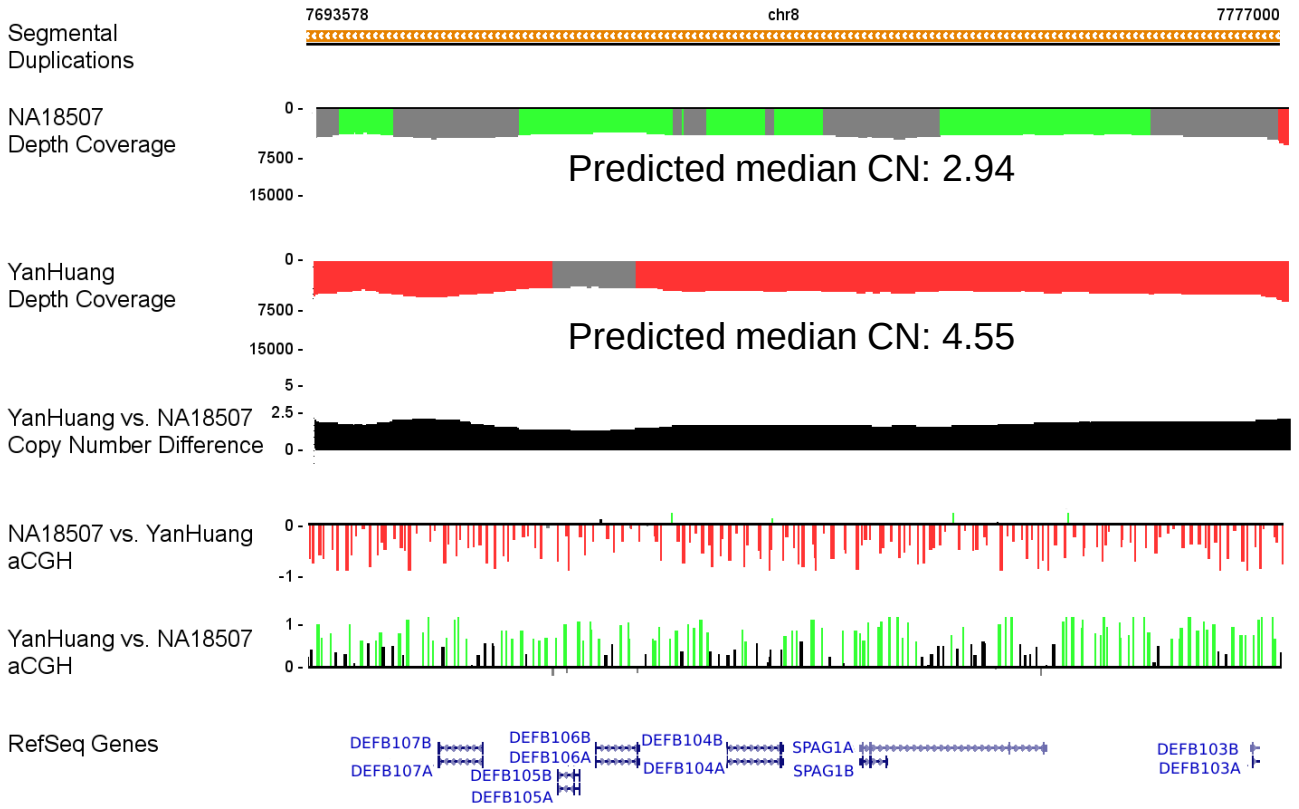
- **Two known ~70 kbp CNPs, CNP#1 duplication absent in Venter but predicted in Watson and NA12878, CNP#2 present mother but neither father or child**

http://mrcanavar.sourceforge.net

Alkan et al., Nature Genetics, 2009

FISH: defensin

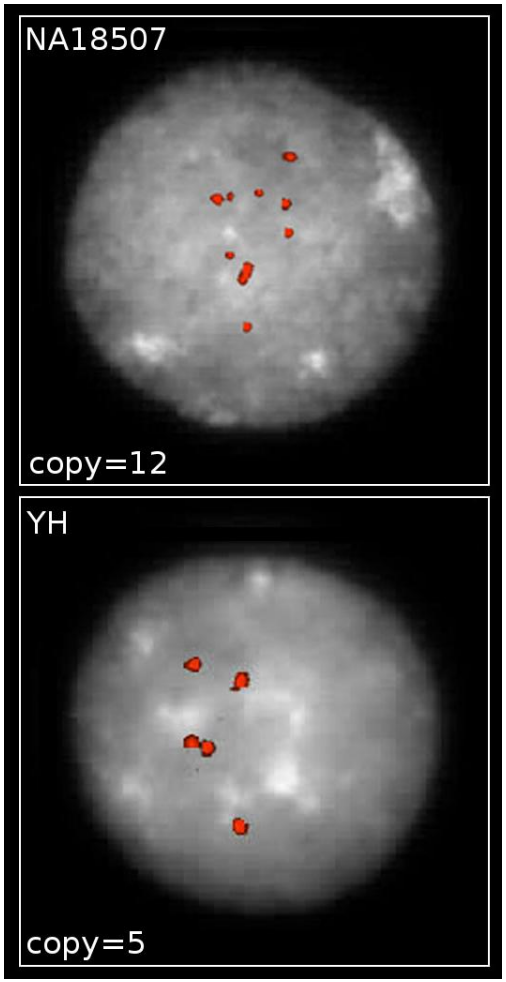


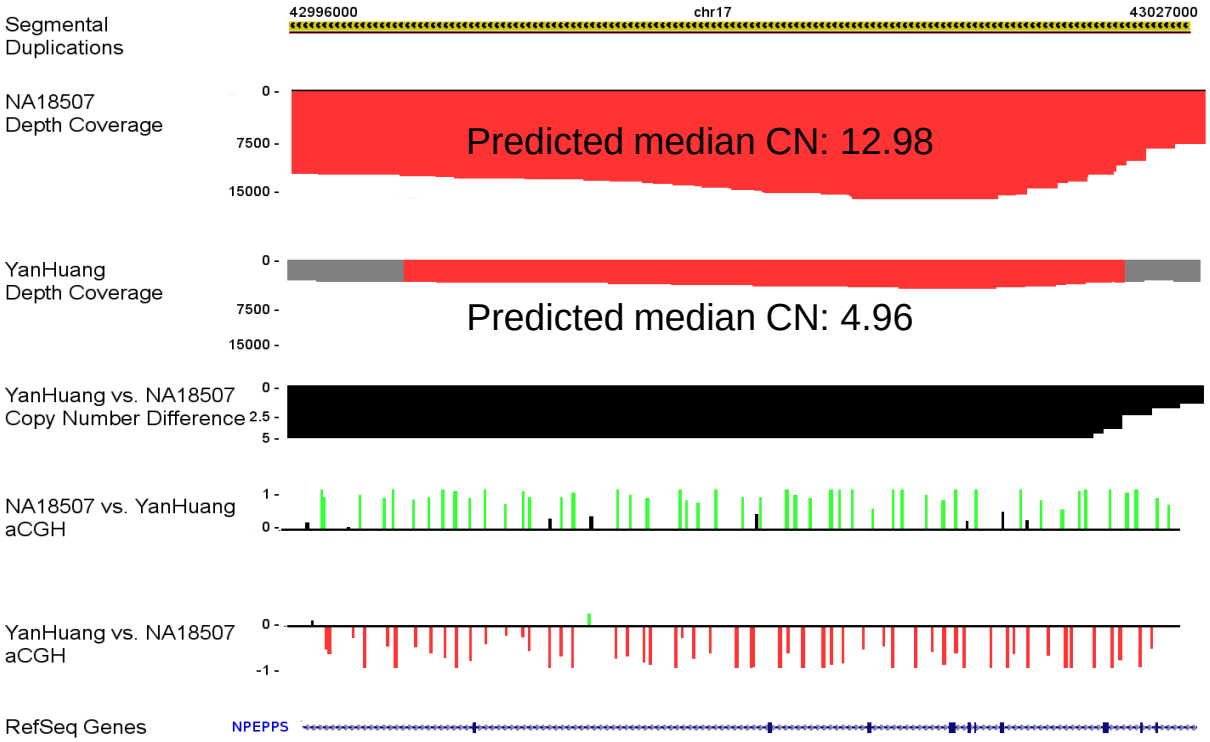


Associated with psoriasis and Crohn's disease

Alkan et al., Nature Genetics, 2009

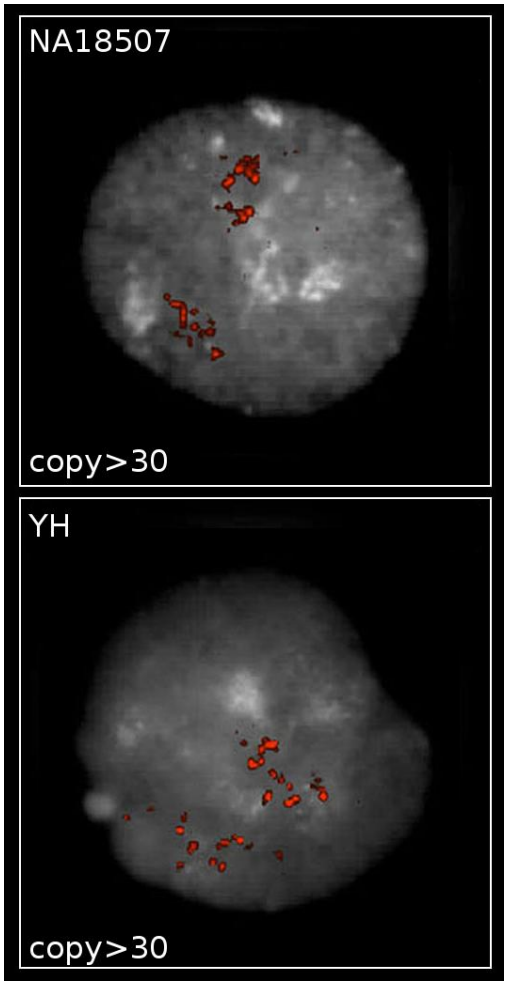
FISH validation: *NPEPPS*

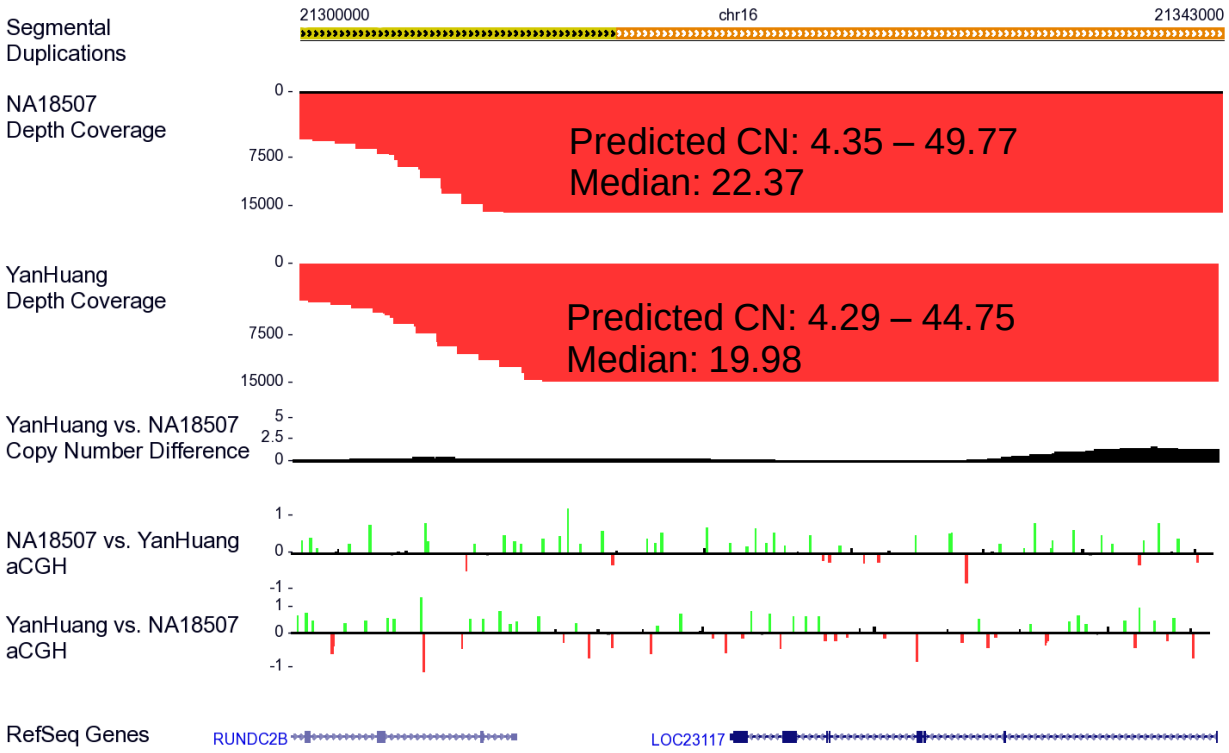




Alkan et al., Nature Genetics, 2009

FISH: Morpheus gene family



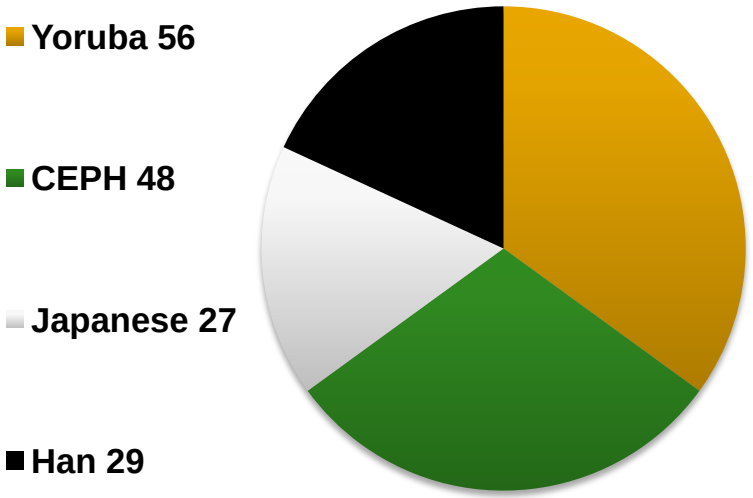


Primate-specific duplication

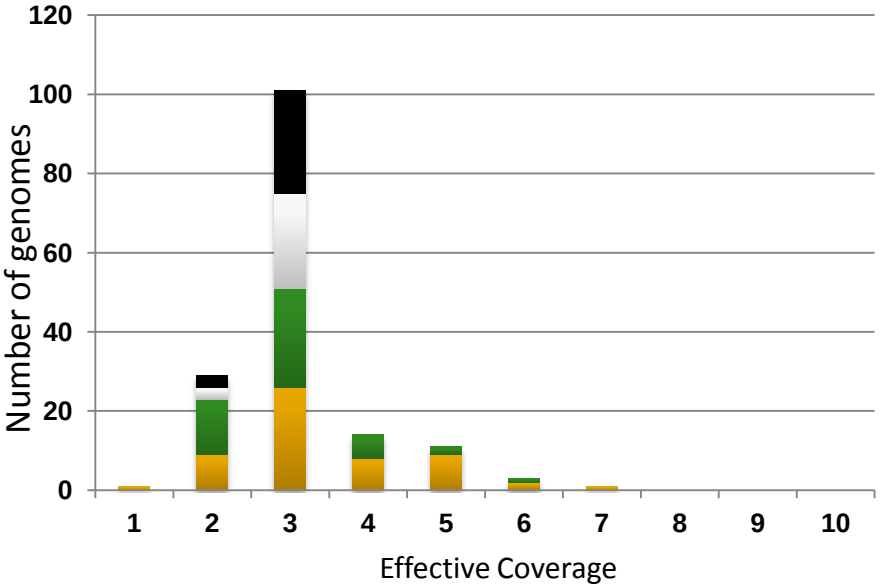
Alkan et al., Nature Genetics, 2009

Scaling up: 1000 Genomes and more

Individuals sequenced in Pilot 1



Histogram of Pilot 1 Illumina effective coverage



Individuals sequenced in Pilot 2

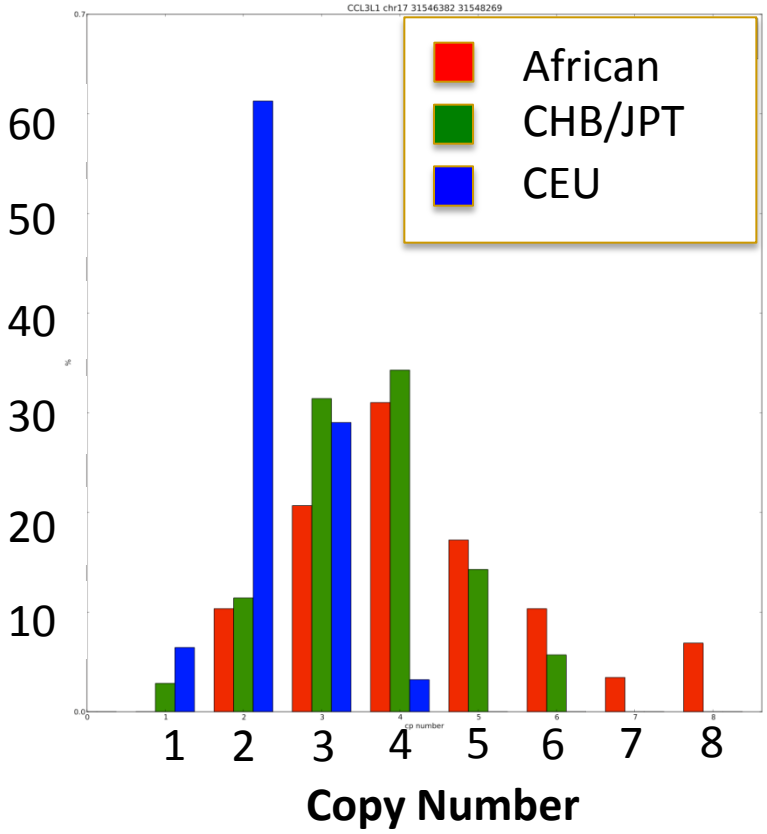
ID	Effective Coverage	Population
NA19240	24	YORUBA
NA19239	19	YORUBA
NA19238	13	YORUBA
NA12891	21	CEPH
NA12892	18	CEPH
NA12878	22	CEPH

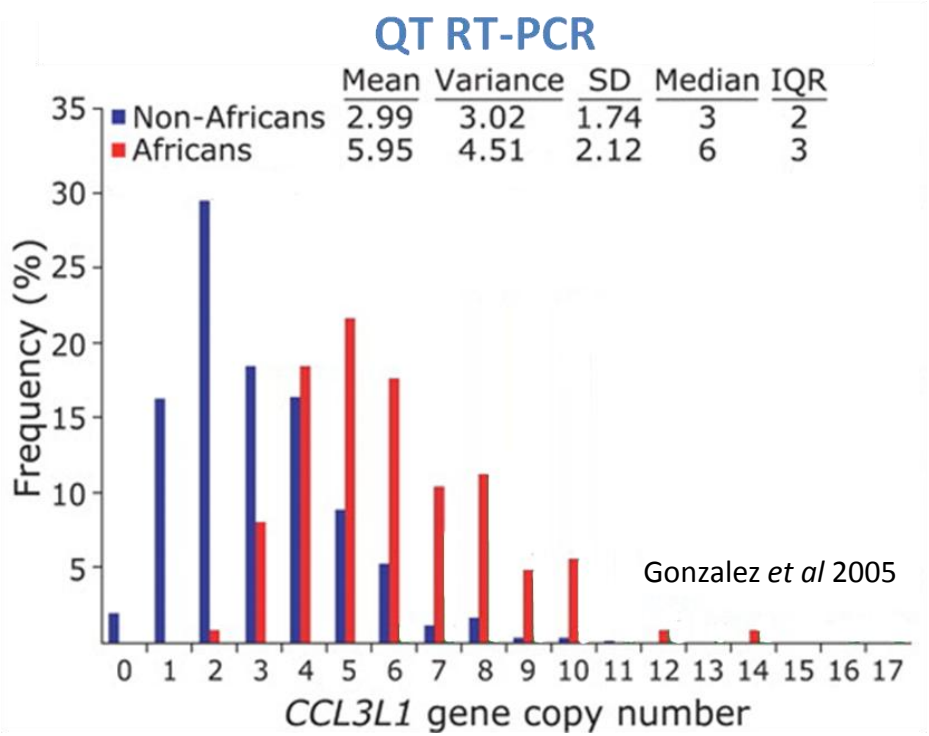
Other Genomes

ID	Effective Coverage	Population
YH-1	22	HAN CHINESE°
NA18507	29	YORUBA ⌘
NA18506	30	YORUBA *
NA18508	25	YORUBA *
KOREAN	12	KOREAN ❖

Sudmant, Kitzman, et al., Science, 2010

Increased *CCL3L1* CN in Africans

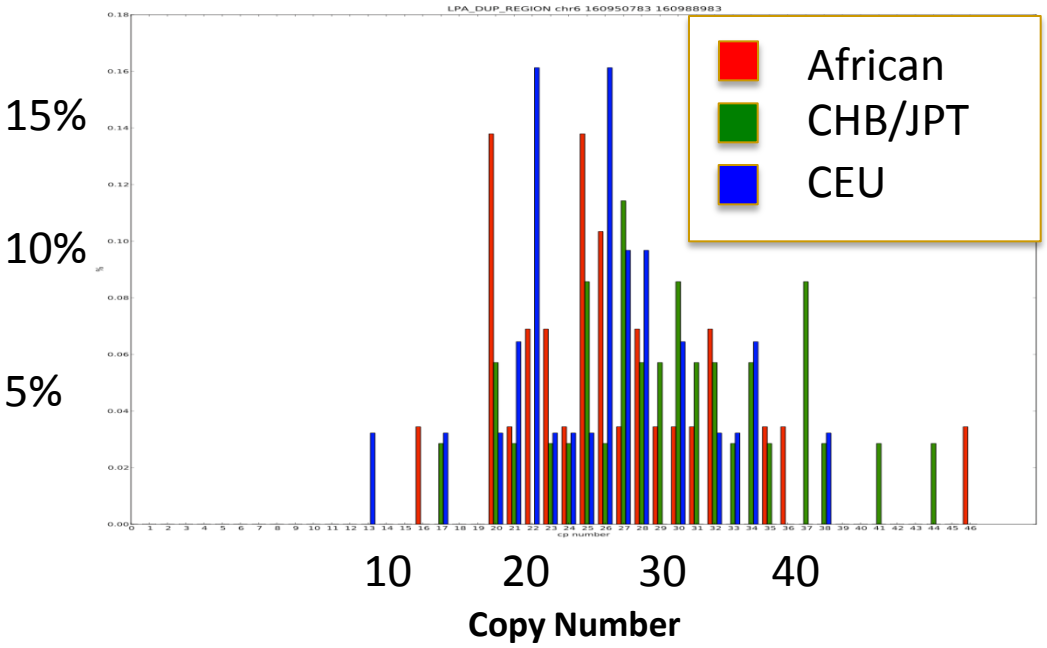




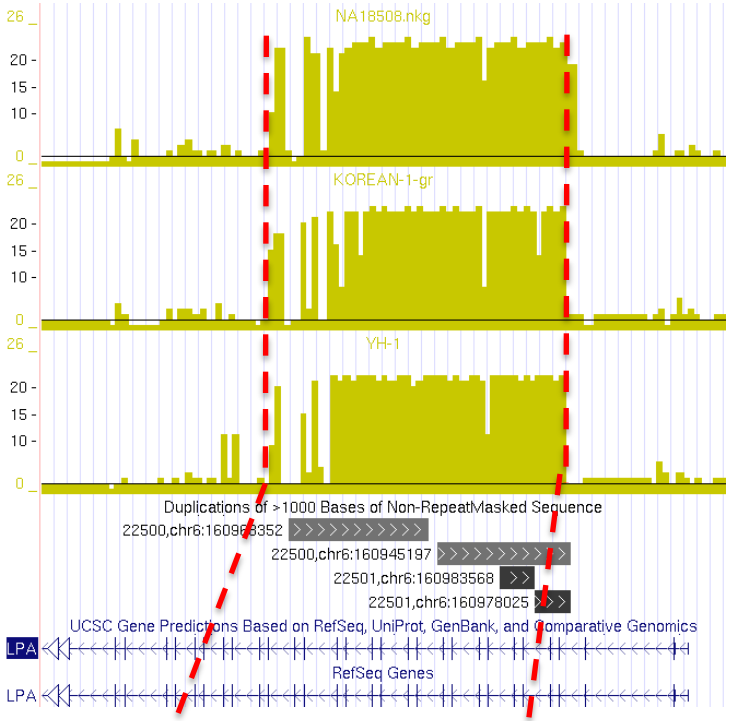
Associated with HIV susceptibility

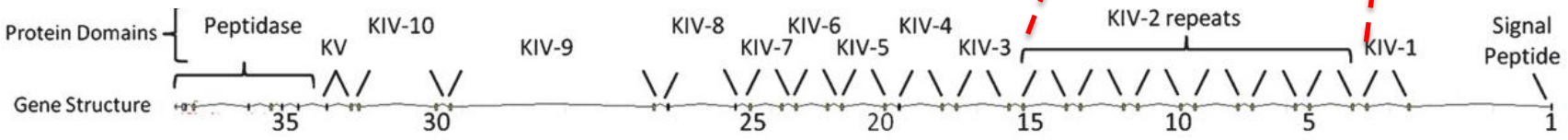
Sudmant, Kitzman, et al., Science, 2010

Smaller *LPA* structure in Africans



LPA Kringle IV repeat region

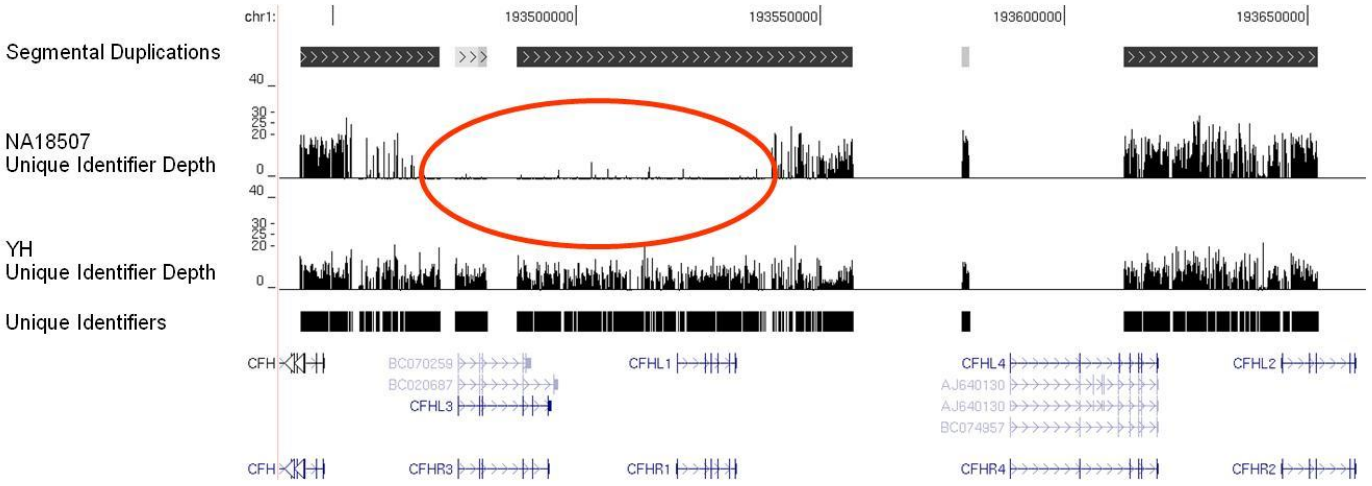




More copies: protective against coronary heart disease

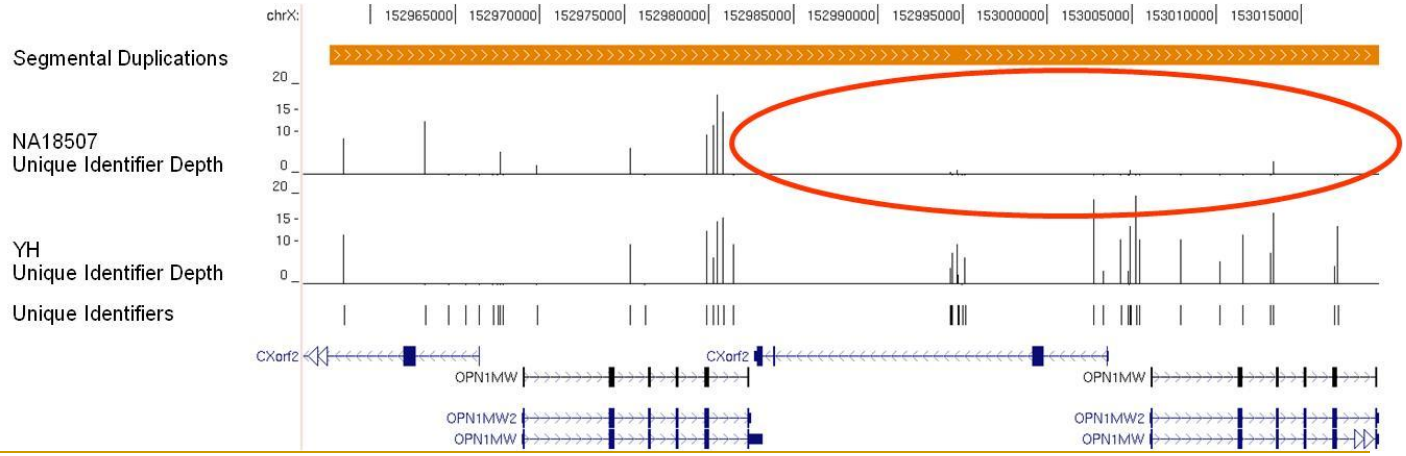
Sudmant, Kitzman, et al., Science, 2010

Differentiating Paralogous Genes



Associated with psoriasis and Crohn's disease

CFHR



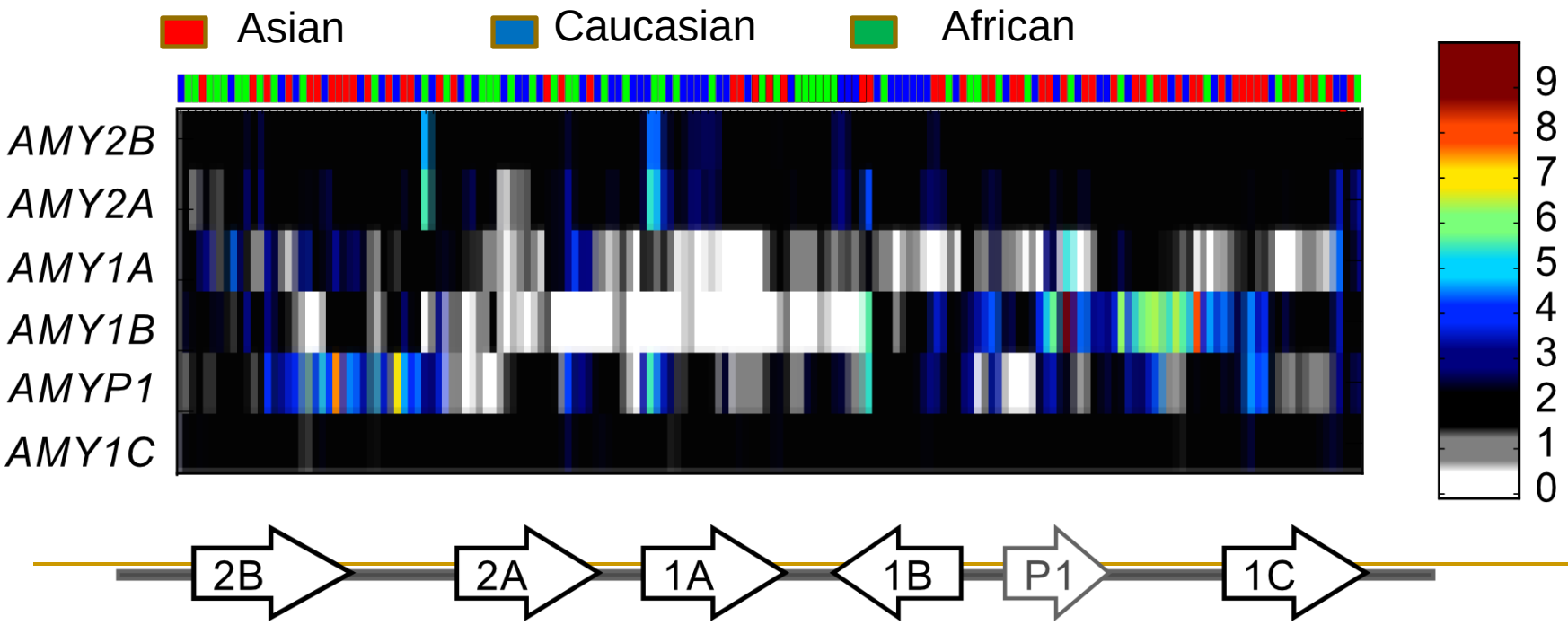
Associated with color blindness

opsin

Alkan et al., Nature Genetics, 2009

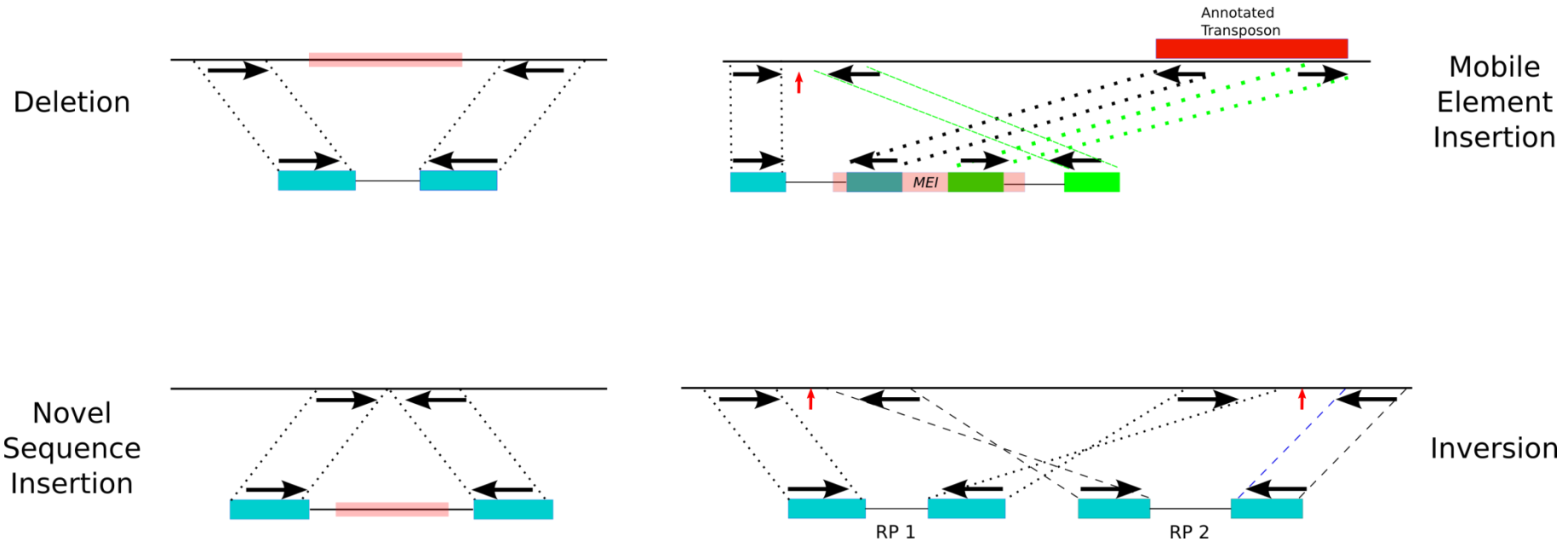
Singly Unique Identifiers (SUNs)

```
Copy 1  ATACTAGGCATATAATATCCGACGATATACATATAGATGTTAG
Copy 2  ATGCTAGGCATGTAATATCCGACGACATACATATACATGTTAG
Copy 3  ATACTAGGCATATAACATCCGACGATATACATATACATGTTAG
Copy 4  ATGCTACGCATATAATATCCCACGATATACATATACATGTTAG
Copy 5  ATGCTACGCATATAATATCCGACGATATACATATACATGATAG
Copy 6  ATACTAGGCATGTAATATCCGACGATATAC- -ATACATGTTAG
```



2) Read pair analysis: VariationHunter

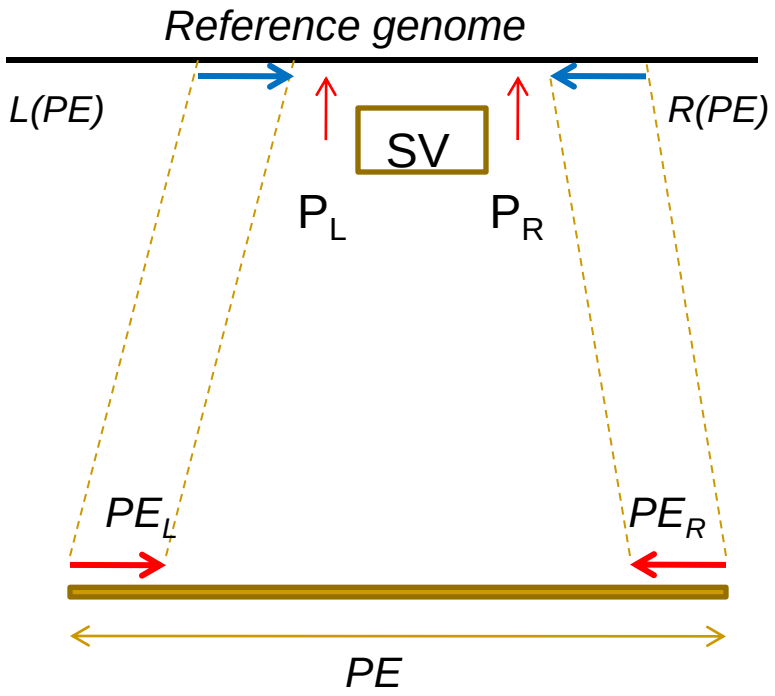
- VariationHunter: Maximum parsimony approach; using all **discordant** map locations; finds an optimal set of SVs through a combinatorial (greedy) algorithm based on approximation to ***set-cover***



http://variationhunter.sourceforge.net (coming soon)

Hormozdiari, Alkan, et al, Genome Res. 2009

Definitions



Paired-end read

PE:= (PE_L, PE_R)

PE-Alignment

(PE, L(PE), R(PE), O(PE))

O(PE): mapping orientation:

- "+/-": normal
- "+/+" or "-/-": inversion
- "-/+": tandem duplication

$SV = (P_L, P_R, L_{min}, L_{max})$

F Hormozdiari, C Alkan, EE Eichler, SC Şahinalp, Genome Research, 2009

Mathematical model

Let L_{min}, L_{max} be *minimum* and *maximum* size of the predicted variant

A Structural Variation is defined by event:

$$SV = (P_L, P_R, L_{min}, L_{max})$$

A PE-Alignment APE=(PE, L(PE), R(PE), O(PE)) supports an insertion $SV = (P_L, P_R, L_{min}, L_{max})$ if:

$$L(PE) \leq P_L$$

$$R(PE) \geq P_R$$

$$L_{min} \geq \Delta_{min} - (R(PE) - L(PE))$$

$$L_{max} \leq \Delta_{max} - (R(PE) - L(PE))$$

F Hormozdiari, C Alkan, EE Eichler, SC Şahinalp, Genome Research, 2009

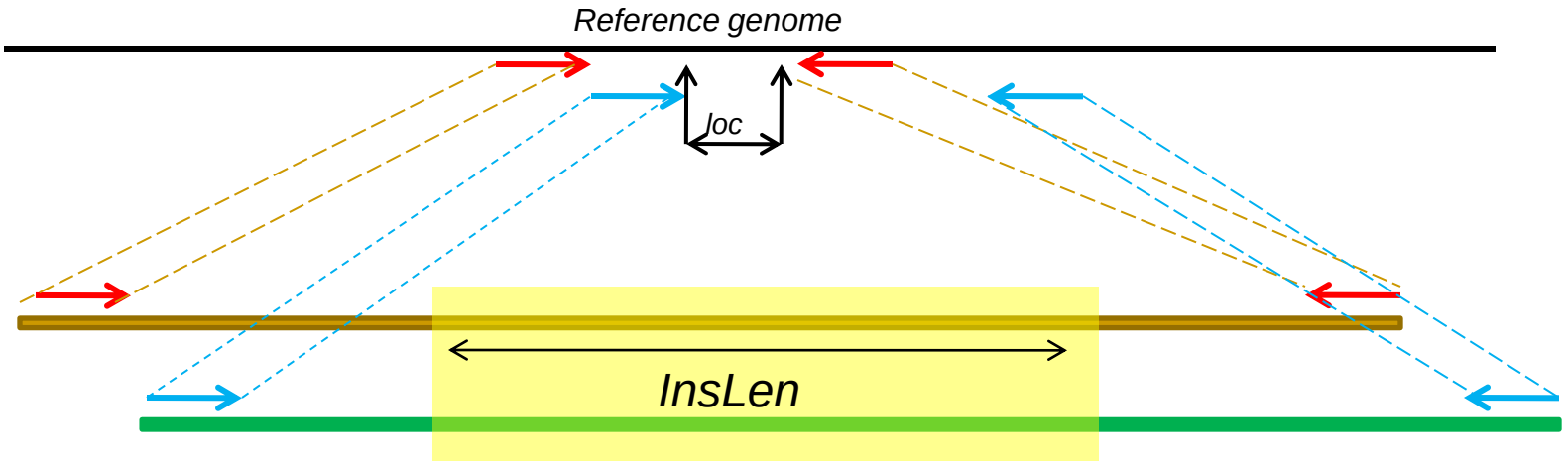
Valid clusters

A set of PE-Alignments that support the same structural variation event SV

A cluster **C** is a *valid cluster* supporting insertions if:

$$\exists loc, \forall APE \in C : L(APE) < loc < R(APE)$$

$$\exists InsLen, \forall APE \in C : \Delta_{\min} - (R(APE) - L(APE)) < InsLen < \Delta_{\max} - (R(APE) - L(APE))$$



F Hormozdiari, C Alkan, EE Eichler, SC Şahinalp, Genome Research, 2009

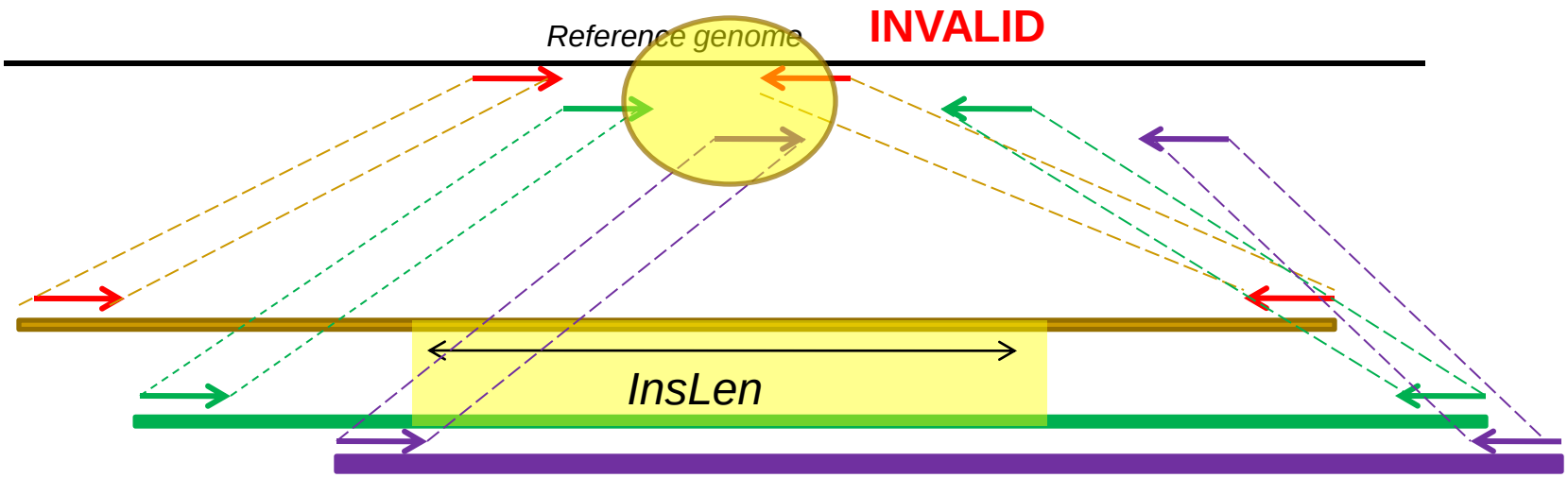
Valid clusters

A set of PE-Alignments that support the same structural variation event SV

A cluster **C** is a *valid cluster* supporting insertions if:

$$\exists loc, \forall APE \in C : L(APE) < loc < R(APE)$$

$$\exists InsLen, \forall APE \in C : \Delta_{\min} - R(APE) + L(APE) < InsLen < \Delta_{\max} - R(APE) + L(APE)$$



F Hormozdiari, C Alkan, EE Eichler, SC Şahinalp, Genome Research, 2009

Maximal Valid Clusters for Insertions

A *Maximal Valid Cluster* is a valid cluster that no additional APE can be added without violating the validity of the cluster

1. Find all the Maximal sets of overlapping paired-end alignments
2. For each maximal set S_k found in Step 1, find all the maximal subsets s_i in S_k that the insertion size (*InsLen*) they suggest is overlapping
3. Among all the sets s_i found in Step 2, remove any set which is a proper subset of another chosen set

Problem: Among all the maximal valid clusters, which ones are correct?
Aim: Assign a single PE-Alignment to all paired-end reads

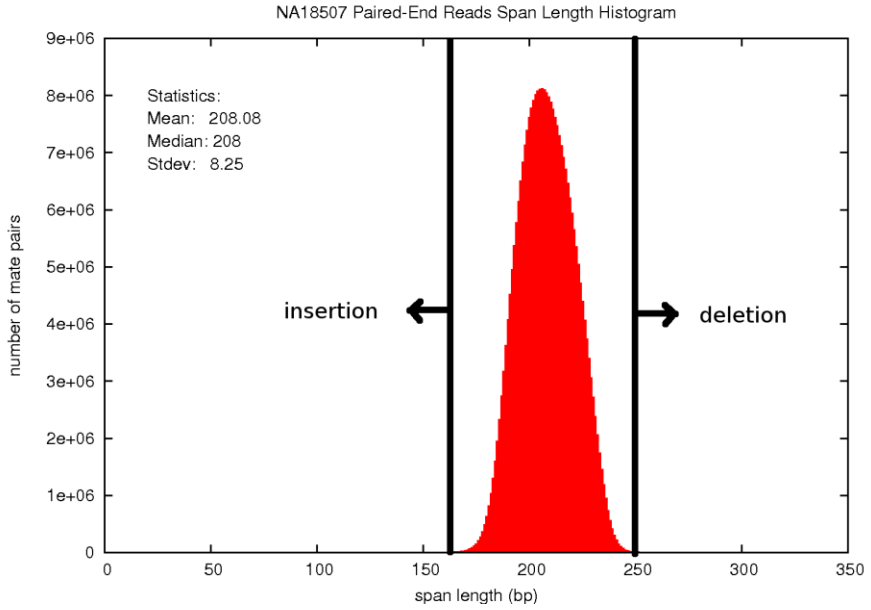
Maximum Parsimony Structural Variation

- Find a *minimum* number of *SVs* such that all the paired-end reads are covered
  - Similar to SET-COVER problem
  - Greedy algorithm. Approximation factor ***O(log(n))***

F Hormozdiari, C Alkan, EE Eichler, SC Şahinalp, Genome Research, 2009

Case Study: NA18507

- NA18507: Yoruba male, sequenced with the Illumina Genome Analyzer
  - 42X sequence coverage
  - Read length: 36bp; average insert size: 208bp, standard deviation: 8.25bp

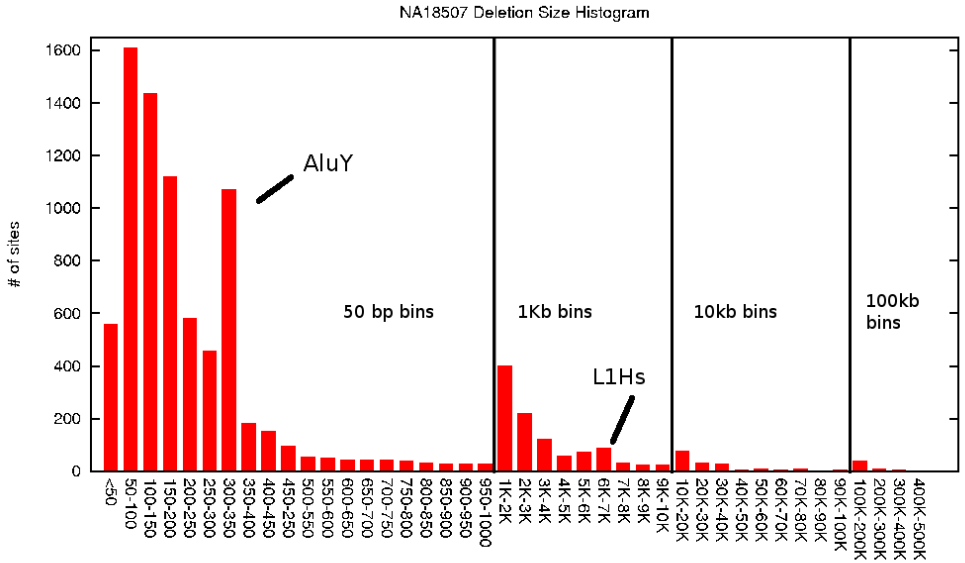


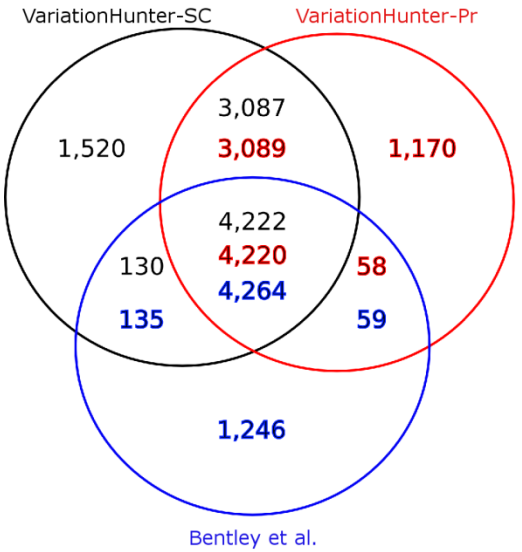
	Validated (Kidd 2008)	VariationHunter-SC		VariationHunter-Pr		Bentley 2008	
		Predicted	Overlap	Predicted	Overlap	Predicted	Overlap
Deletion	143	8,959	85	8,537	85	5,704	67
Inversion	82	504	23	181	11	*NA*	*NA*
Insertion	*NA*	5,575	*NA*	7,142	*NA*	*NA*	*NA*

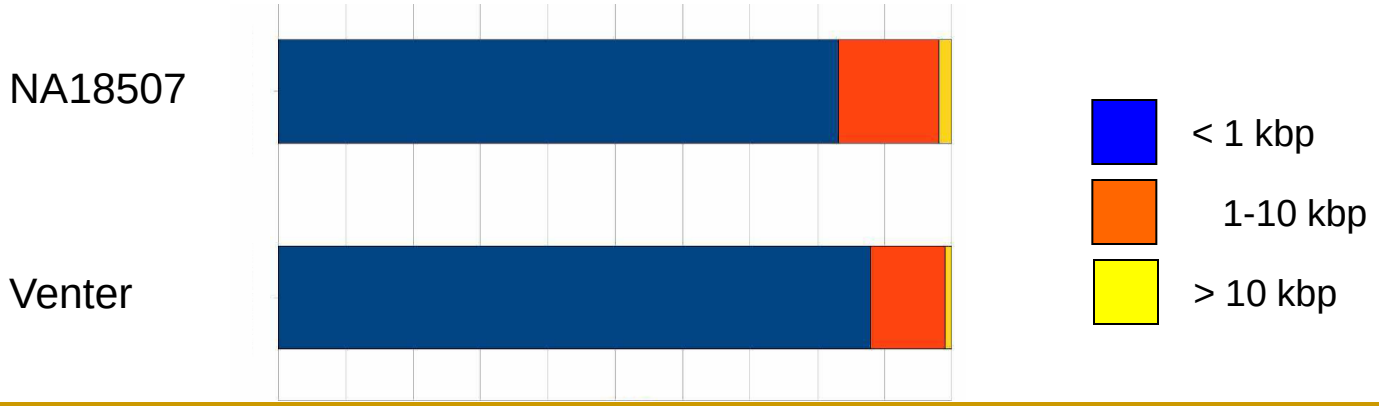
* predicted insertions < 100bp

Hormozdiari, Alkan, et al, Genome Res. 2009

Case Study: NA18507





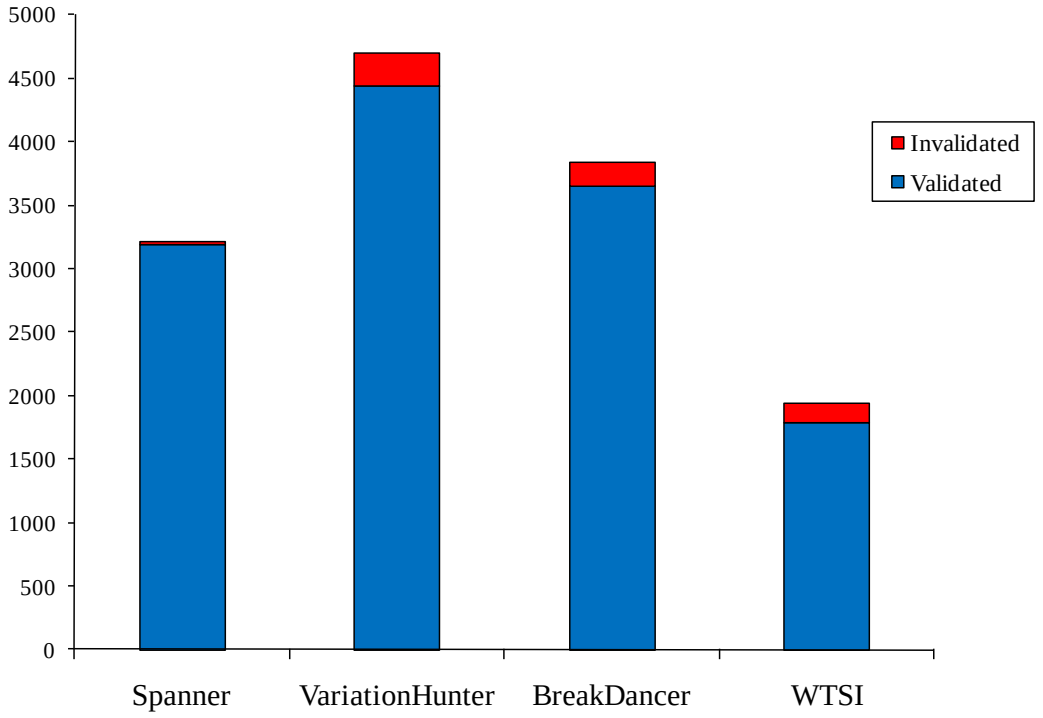


Hormozdiari, Alkan, et al, Genome Res. 2009

1000 Genomes: Validation

YRI trio: NA19238, NA19239, NA19240
Deletion calls with Illumina read pair analysis
Validation: arrayCGH and PCR

- Higher sensitivity than unique-location based methods
- Higher false discovery rate
  - Assumes complete reference genome
  - Paralogous variants that do not exist in the reference genome will be invalidated



VariationHunter for *Alu* insertions

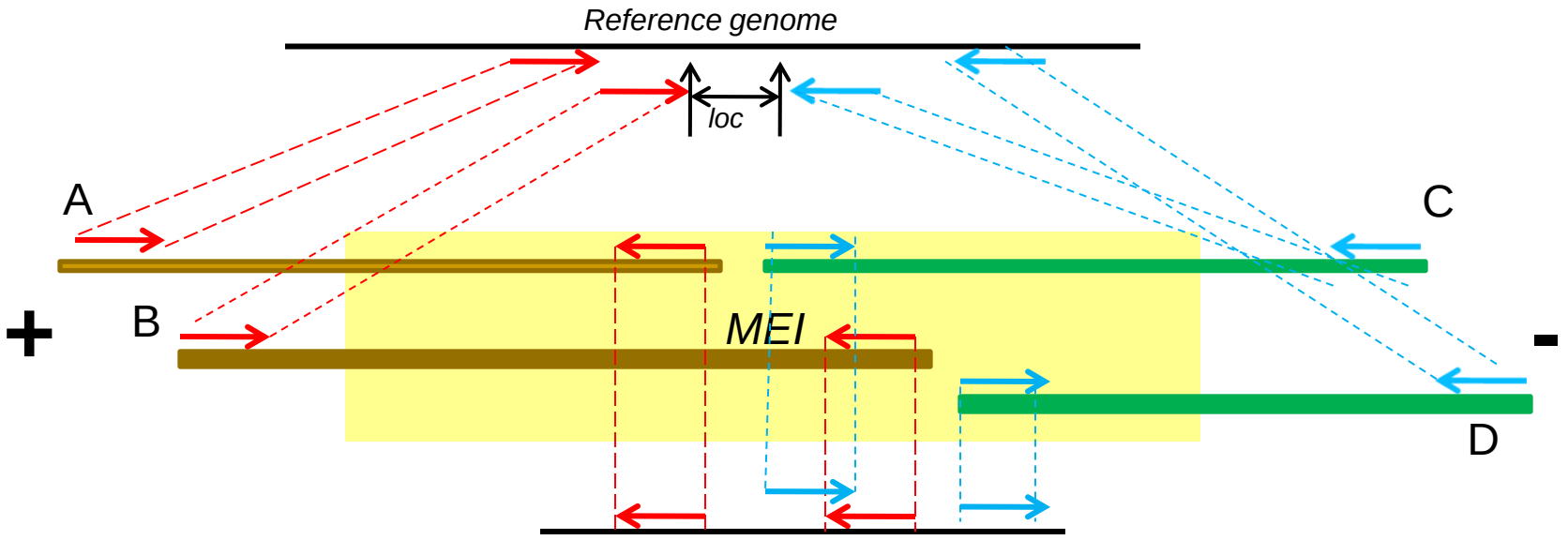
- 8 high-depth genomes
  - One trio (YRI: NA18506, NA18507, NA18508)

Individual	Population	Seq. Coverage	Phys. Coverage	# *Alu*	Novel
NA18506	YRI	40.1x	255x	1,720	1,280
NA18507	YRI	27.1x	157x	1,579	1,144
NA18508	YRI	37x	214x	1,744	1,293
NA10851	CEU	22x	160x	1,282	781
AK1	Korean	22.5x	49x	909	582
YH	Han Chinese	11.4x	27x	1,160	698
KB1	Khoisan	21x	25x	457	313
HGDP01029	Khoisan	4x	12x	307	214
Non Redundant Total				4,342	3,432

- 63/64 (98%) sites tested with PCR are validated.
- 1,437/4,342 (33.1%) map within genes (RefSeq May 2010)

Hormozdiari, Alkan et al., Genome Res. 2011

MEI sequence signature

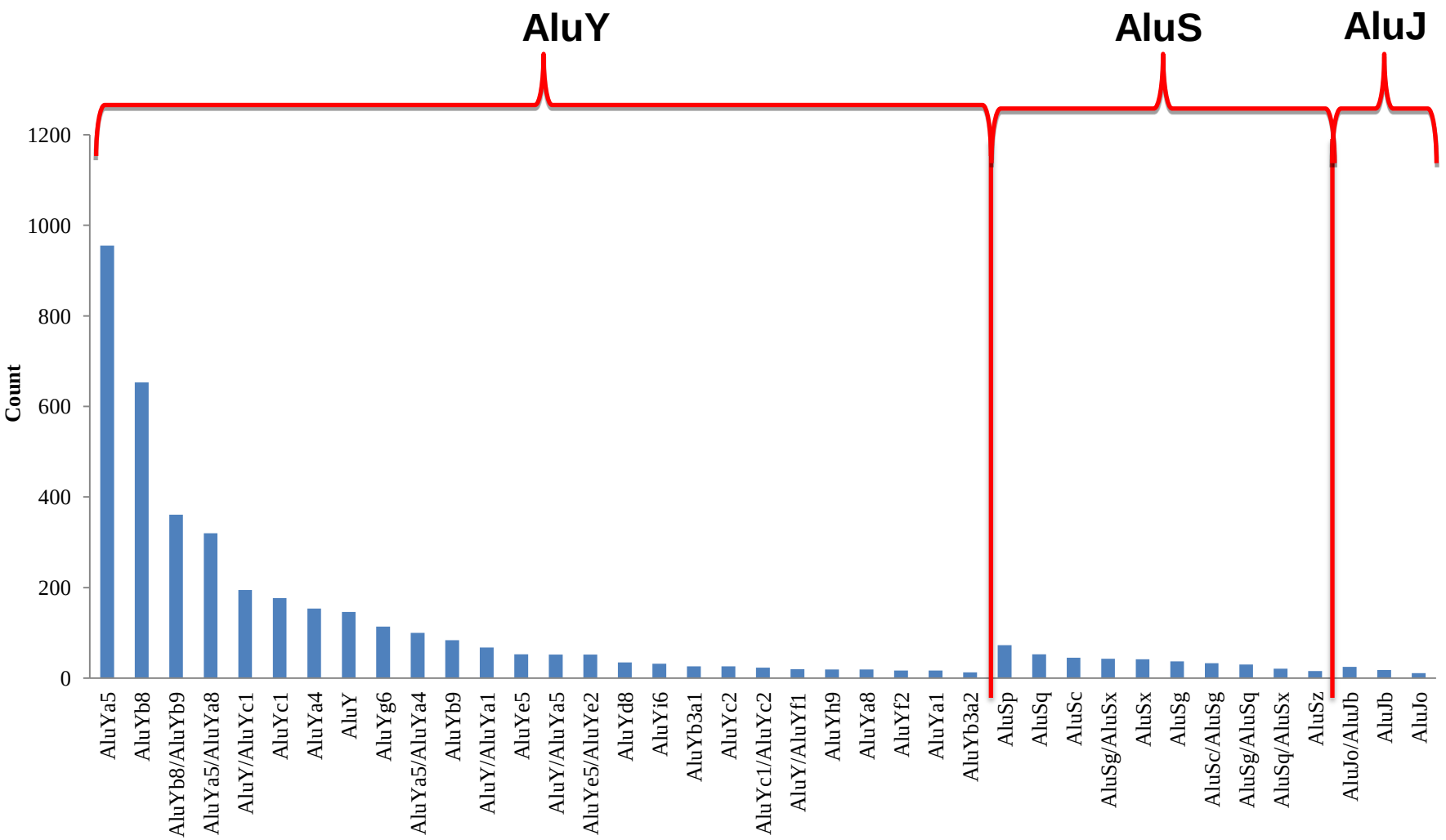


- Strand rules: MEI-mapping "+" reads and MEI mapping "-" reads should be in different orientations:
  - +/- and -/+ clusters; or +/+ and -/- clusters (inverted MEI)
- Span rules: A=(A1, A2); B=(B1, B2); C=(C1, C2); D=(D1, D2)
  - |A1-B1| ~ |A2-B2| and |C1-D1| ~ |C2-D2| (simplified; we have 8 rules)
- Location and 2-breakpoint rule:

$$\exists loc, \forall PE : RightMost(+) < loc < LeftMost(-)$$

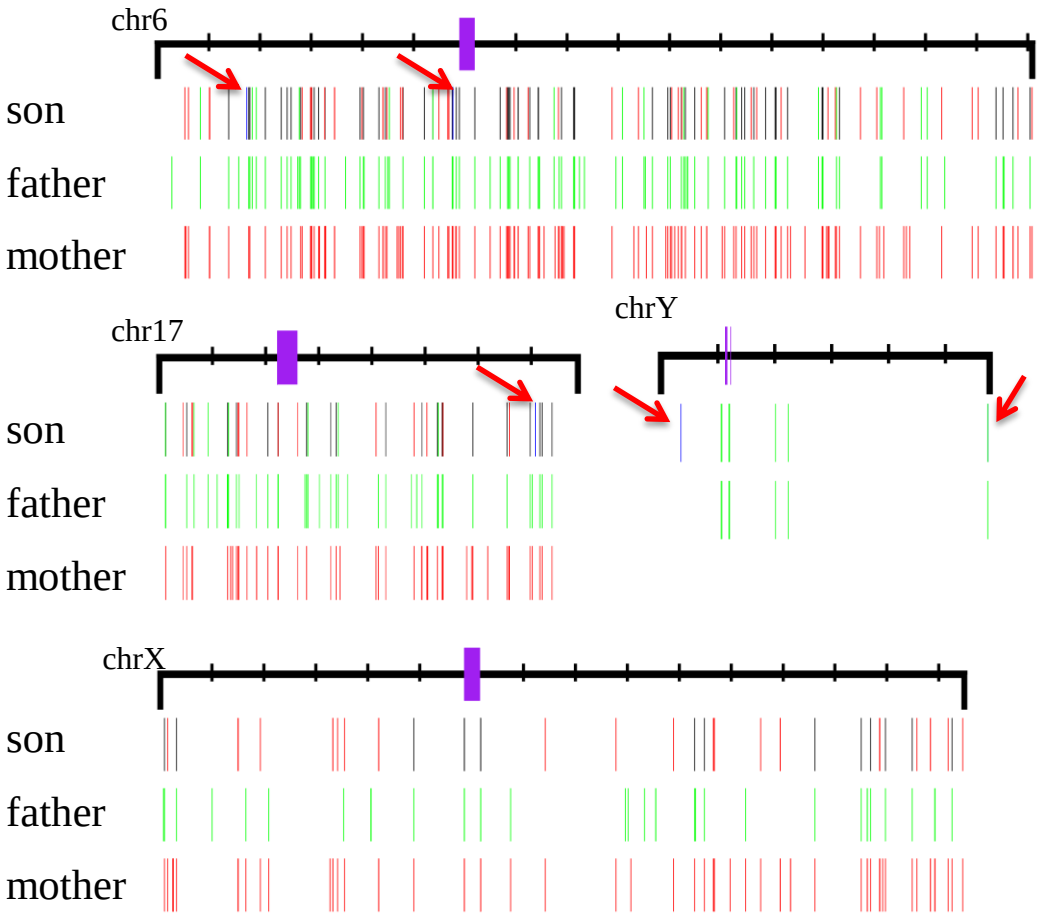
Hormozdiari et al., Bioinformatics 2010

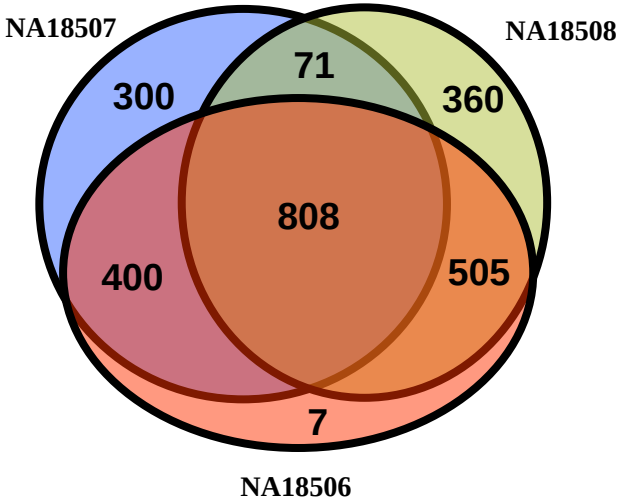
Alu subfamilies

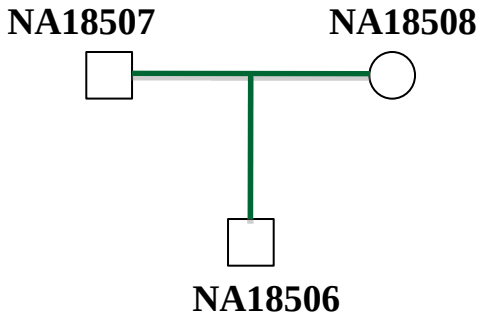


Hormozdiari, Alkan et al., Genome Res. 2010

Familial transmission of *Alu*s

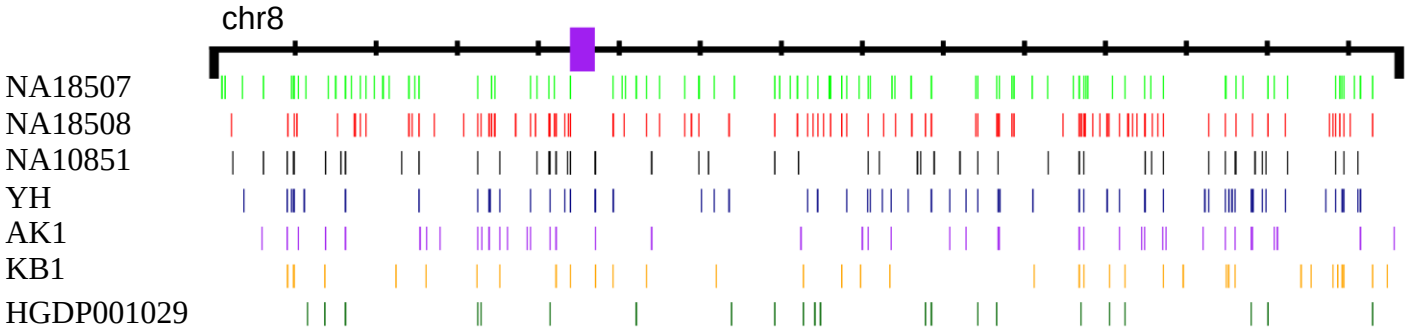


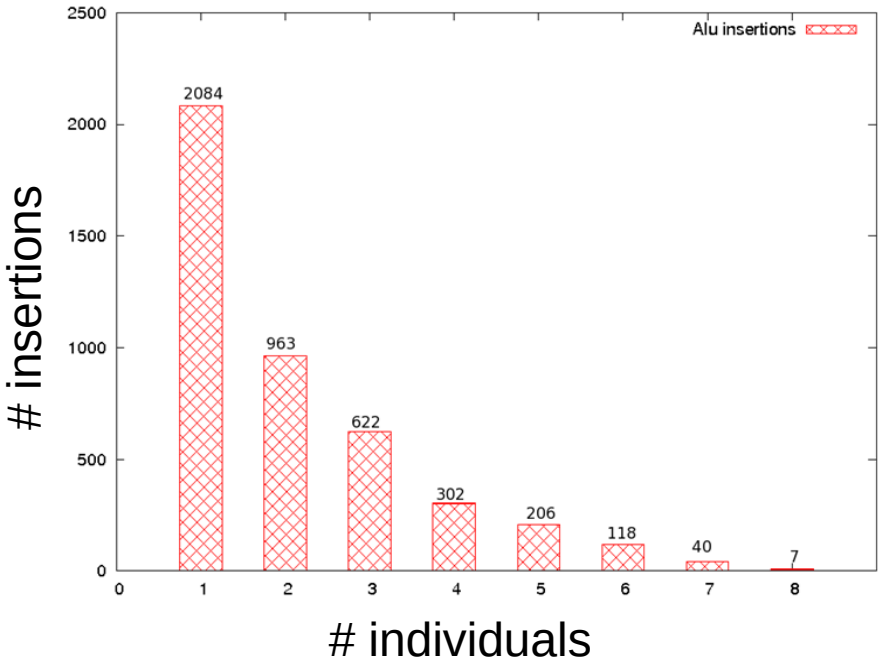


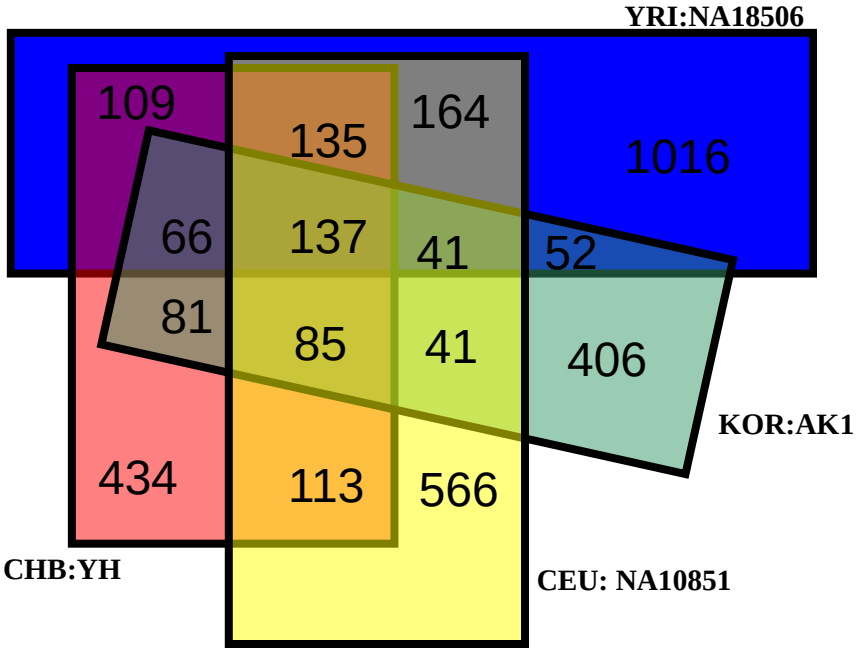


Hormozdiari, Alkan et al., Genome Res. 2010

Common *Alus*



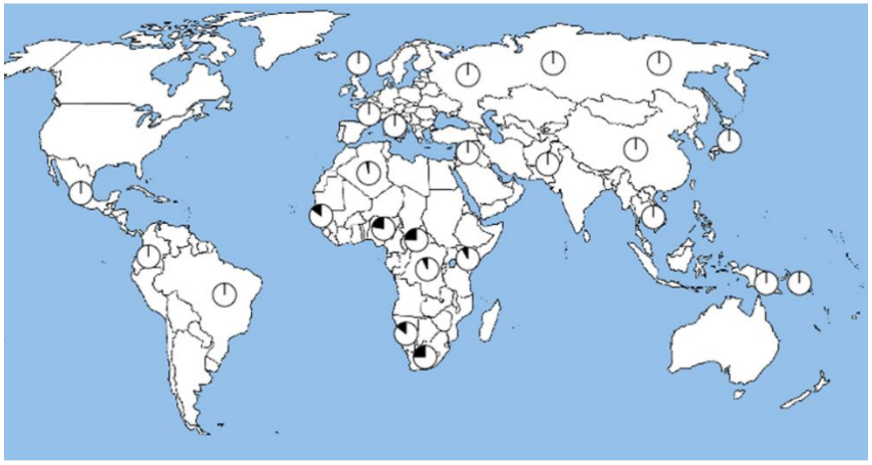


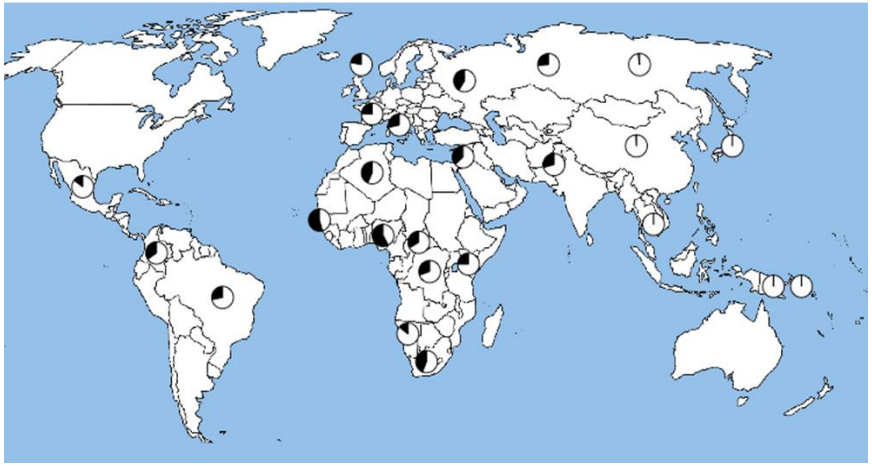


Hormozdiari, Alkan et al., Genome Res. 2010

Alu polymorphism

- *in silico* genotyping with 1000GP:
  - 10.5% (21/201; chr1) are significantly stratified (Fst>0.2),
  - → ~400 markers for population genetics analyses
  - 18/21 show increased allele frequency in the YRI when compared to either the ASN or CEU populations
- PCR genotyping of 3 *Alu* insertions in 1,058 individuals from 52 populations (HGDP)



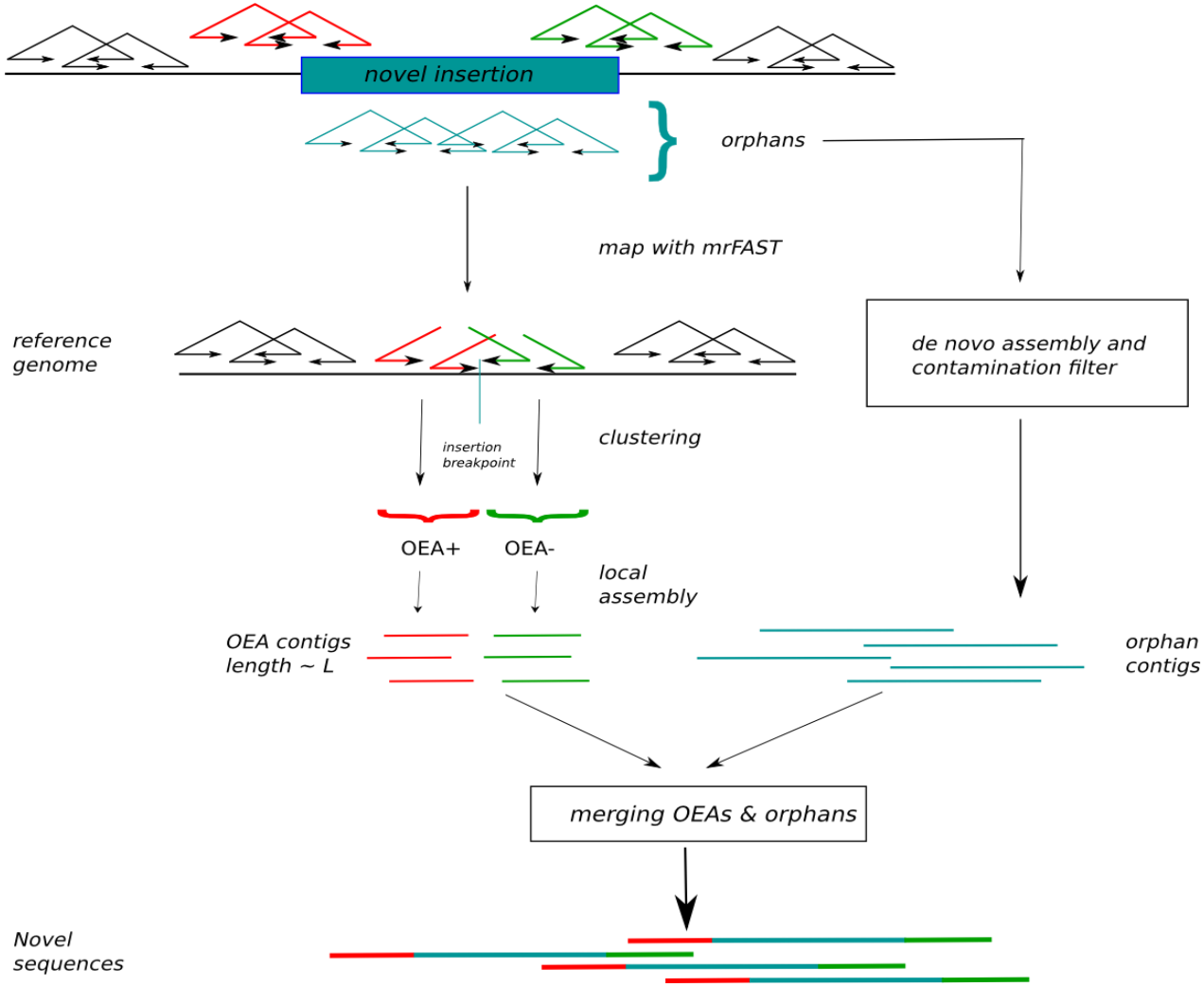


Hormozdiari, Alkan et al., Genome Res. 2010

Finding "novel" sequences

- DNA sequences that have no representation in the reference genome assembly
  - Excludes duplications & common repeats
- Two major NGS-based methods:
  - *Whole genome de novo assembly*
    - *ALLPATHS-LG, SOAPdenovo, ABySS, Cortex, Velvet, Euler, etc.*
      - *Compute and memory intensive*
  - *Local de novo assembly using mapping information*
    - *Poor man's method: Going through the trash that the mapper left*

3) Local Assembly: NovelSeq

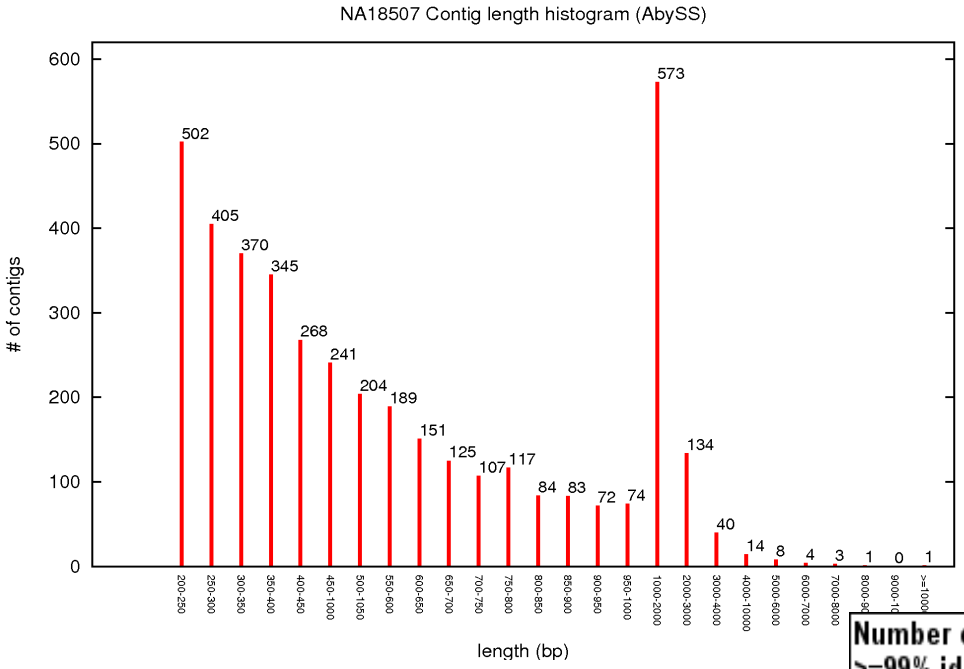


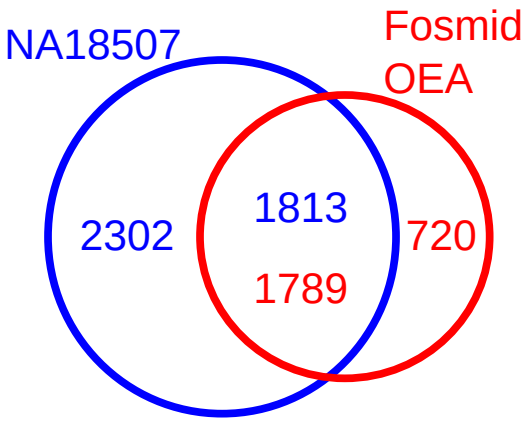
http://compbio.cs.sfu.ca/strvar.htm

Hajirasouliha et al., Bioinformatics 2010

NovelSeq: NA18507

4,154 contigs (≥200bp)
Total 2.9 Mb sequence that is not in the reference assembly
N50 size: 955 bp



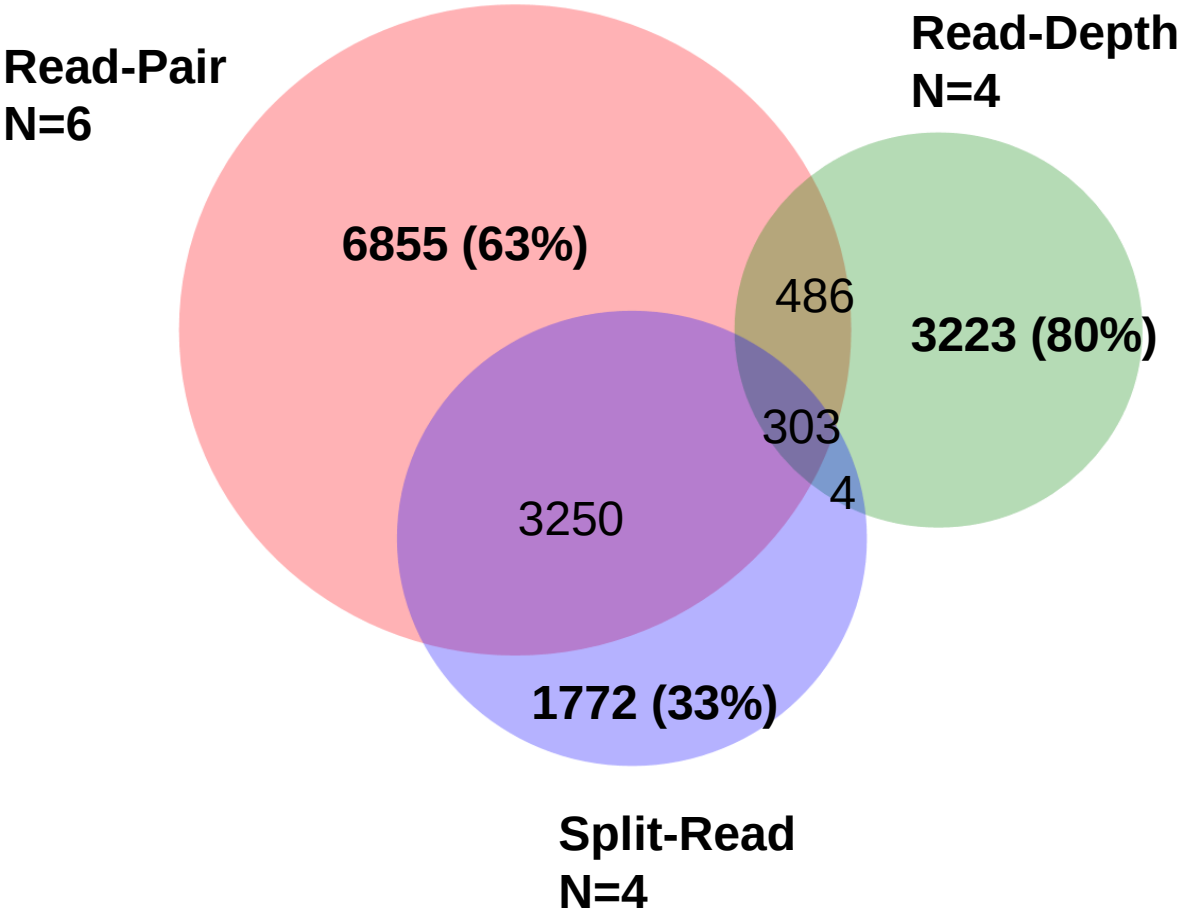


Number of total contigs, and contigs map to various sequence databases with >=99% identity

	Total	Build36	Venter WGS	HuRef	NA18507 Fosmid WGS
>=1 kb	779	13	744	43	337
>=200bp (<1Kb)	3336	110	2957	69	654

Hajirasouliha et al., Bioinformatics 2010

No method is comprehensive



1000G Consortium, Nature 2010, Mills et al., Nature 2011

Summary

- Next-generation sequencing technologies
  - Promises to replace array based methods; but:
    - Entire spectrum of structural variation is not yet detected
    - Most current studies target only CNVs in relatively less complex areas of the genome
    - Different sequencing platforms present different error models
  - Need better methods to
    - Identify ***inversions*** and ***translocations***
    - Discover SVs in repeat- and duplication-rich regions
    - Accurately characterize ***copy***, ***content***, and ***structure*** of structural variants
- Long term goal: **accurate** *de novo* assemblies to detect a broad range of variants

Acknowledgements

UW
Evan E. Eichler
Jeffrey M. Kidd
Tomas Marques-Bonet
Peter Sudmant
Jacob Kitzman
Gözde Aksay
Carl Baker
Francesca Antonacci
Santhosh Girirajan
Farhad Hormozdiari
(now at UCLA)
Maika Malig
Emre Karakoç

SFU
S. Cenk Şahinalp
Fereydoun Hormozdiari
Iman Hajirasouliha
Faraz Hach

1000 Genomes Project Structural Variation Analysis Subgroup

EBI

Paul Flicek
Rasko Leinonen

Baylor College of Medicine
Richard A. Gibbs

NCBI

Martin Shumway

Carnegie Mellon University
Onur Mutlu