

Structural variation

Programming for Biology

CSH, October 2015

Tomas Marques-Bonet
ICREA Research Professor
Institut de Biologia Evolutiva



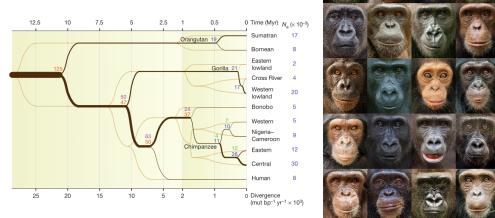
Who we are?

- Evolutionary genomics
 - Barcelona, Biomedical Research Park



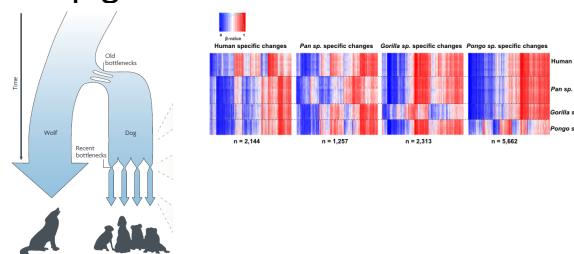
What do we do?

- Natural selection on human evolution



- Transcriptome and Epigenetics in Primates

- Canid evolution



Continuum of Genomic Variation

Single nucleotide

- Single base-pair changes

• Cpg Methylation

• Small insertions/deletions

• Mobile elements

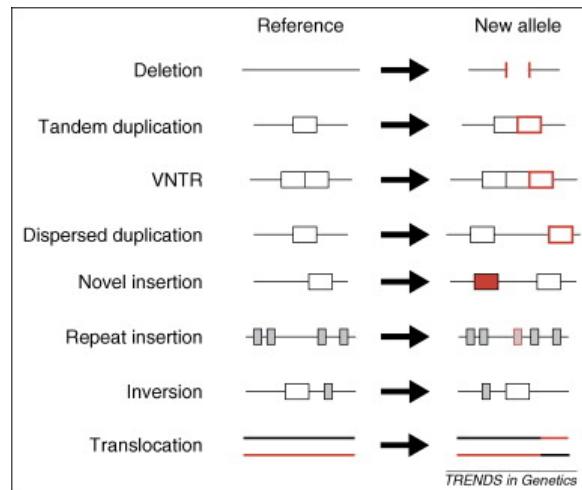
• Large-scale genomic copy number variation (>10 kb)

• Local Rearrangements

Chromosome Chromosomal variation

Copy Number Variation
Structural Variants (SV)

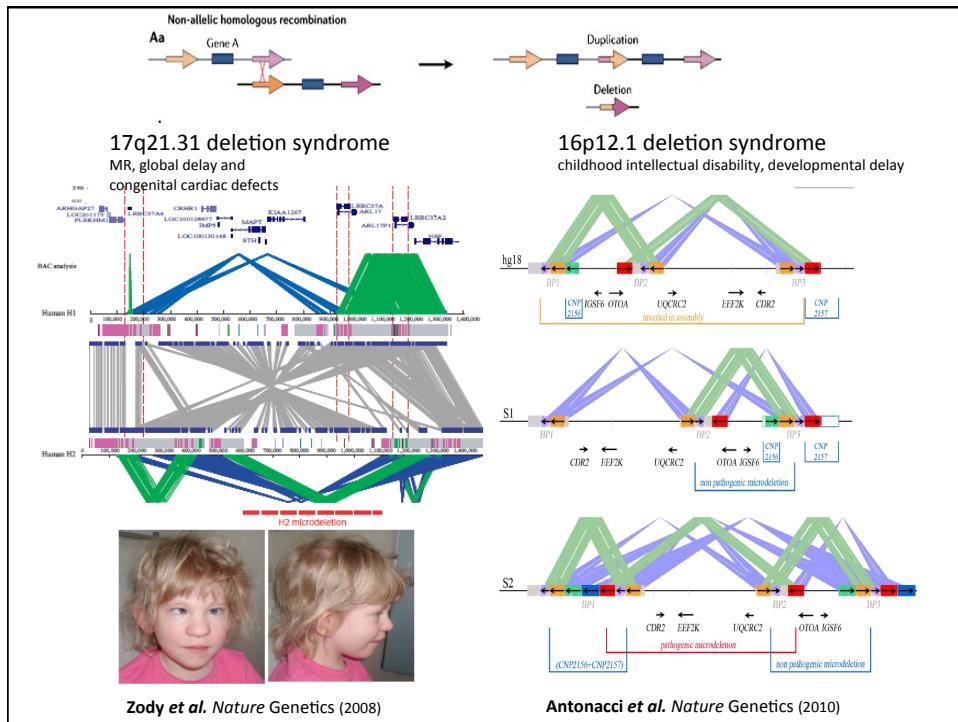
Types of Structural Variation



Hurles et al. 2008

Why Study Structural Variation?

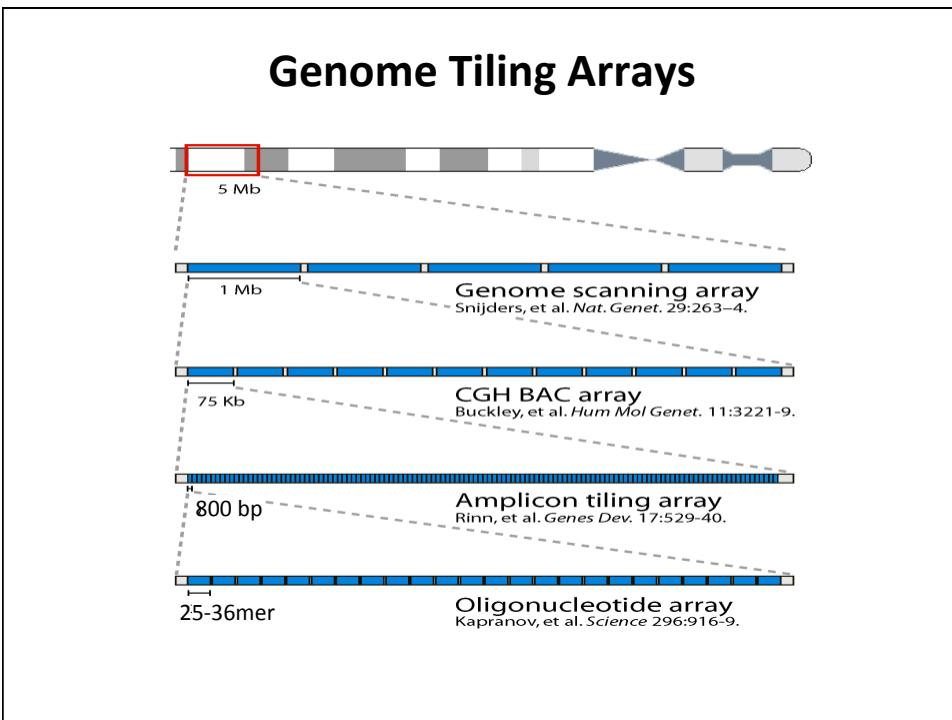
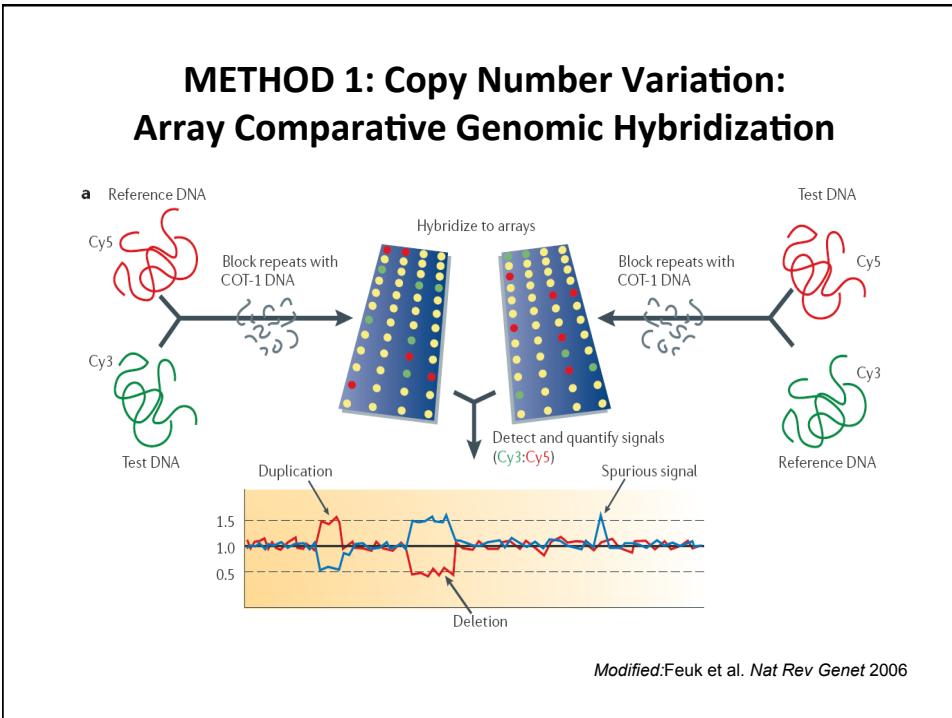
- Common in “normal” human genomes-- major cause of phenotypic variation
- Common in certain diseases, particularly cancer
- Now showing up in rare disease; autism, schizophrenia



Methods to Find SVs

Experimental approach
ArrayCGH (SNP based and genomic)

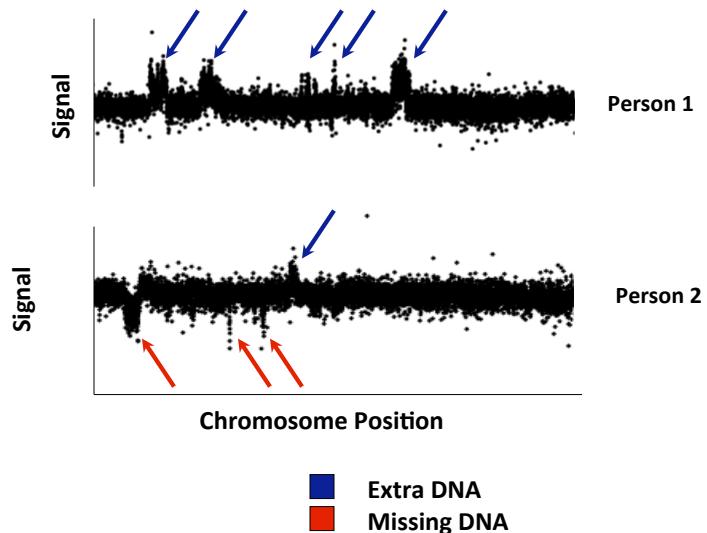
- Sequence based
- Local and *de novo* assembly
- Read pair analysis
- Read depth analysis
- Split read analysis

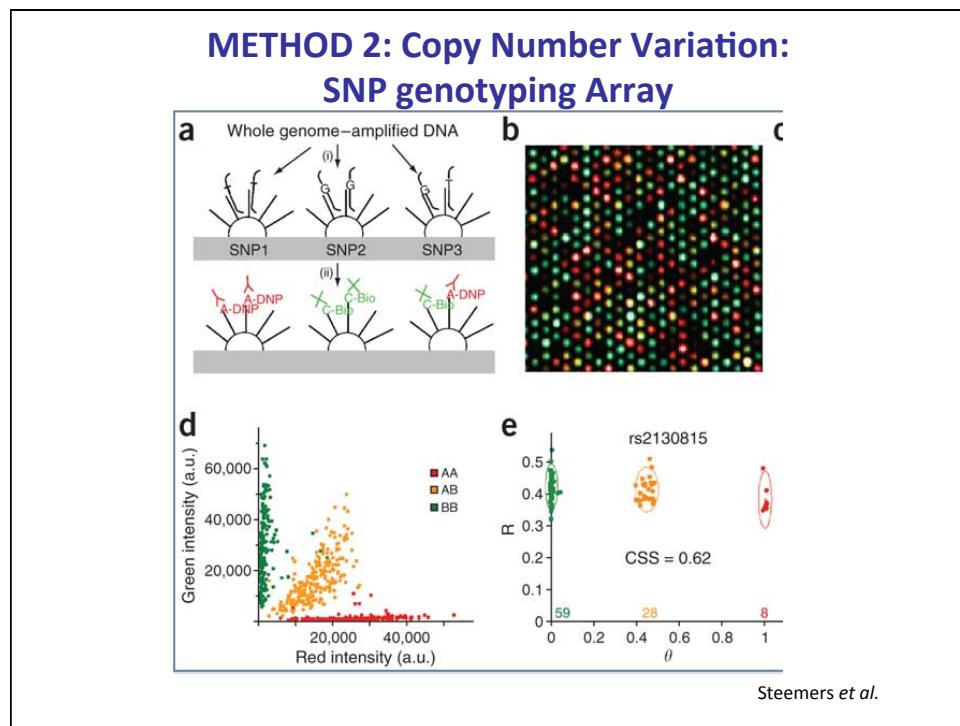
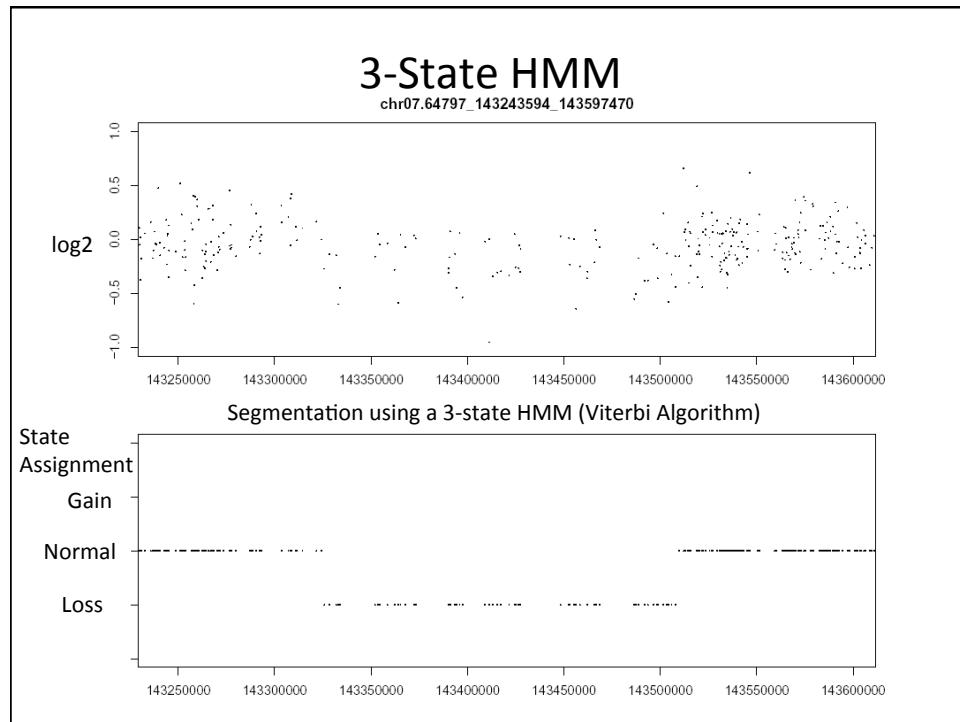


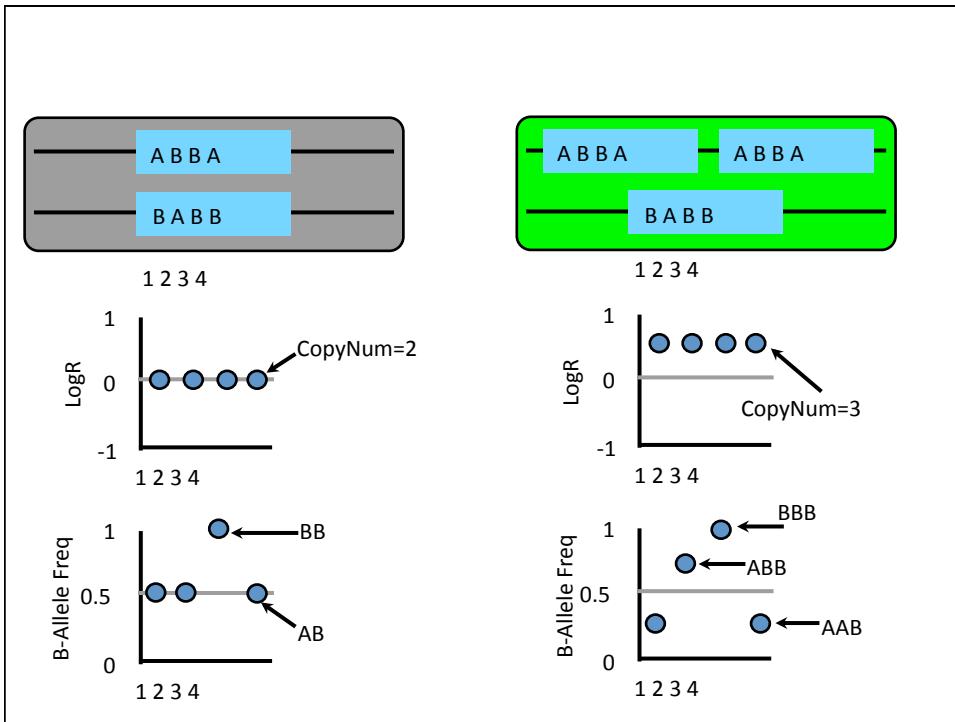
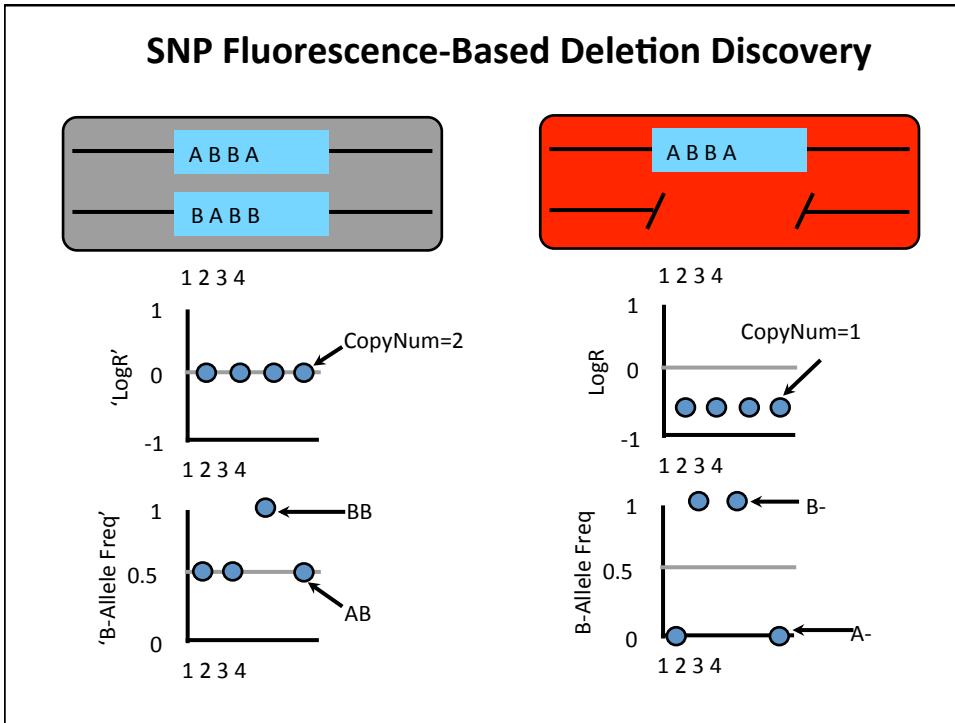
Typical Analysis Procedure

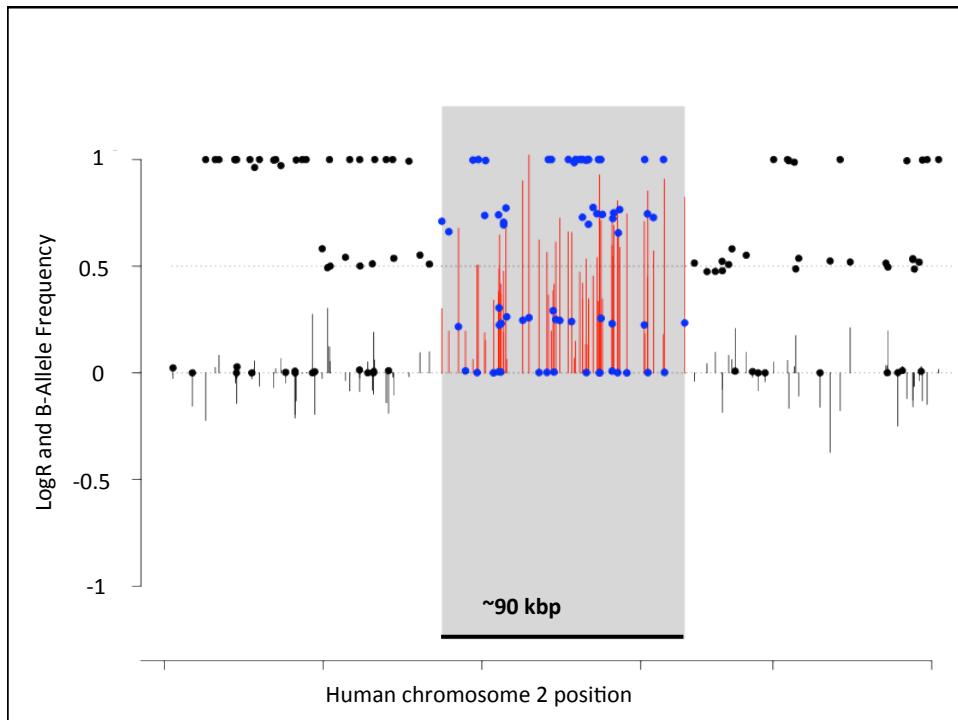
- For each probe, calculate a log2 ratio of test/reference
 - Log2 serves to center values around 0
 - Hemizygous deletion in test: $\log_2(\text{test}/\text{reference}) = \log_2(1/2) = -1$
 - Duplication in test: $\log_2(\text{test}/\text{reference}) = \log_2(3/2) = 0.59$
 - Homozygous duplication: $\log_2(\text{test}/\text{reference}) = \log_2(4/2) = 1$

Copy Number Variations in the Human Genome





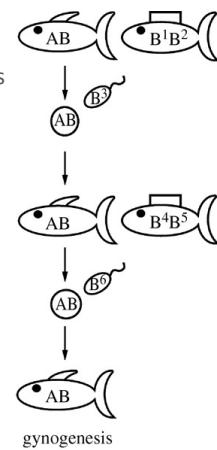




Amazon Molly (*Poecilia formosa*)

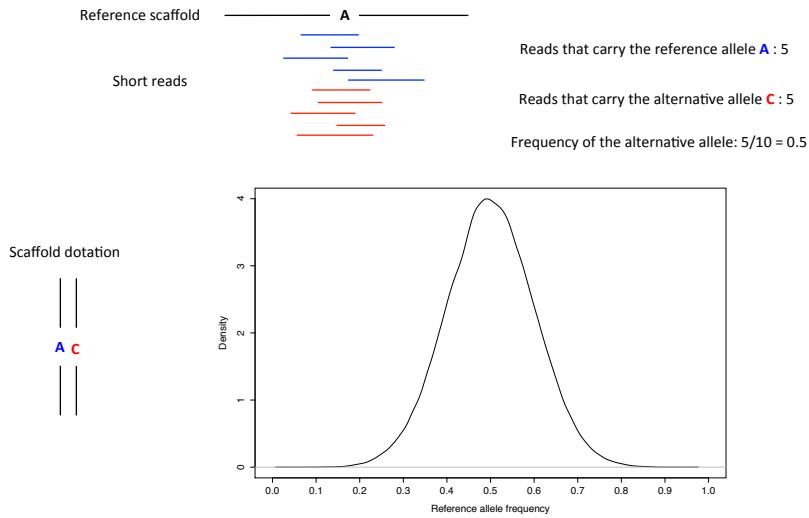
What is it that makes the Amazon molly so unique?

- The Amazon molly is one of the few extant asexual vertebrates
- All individuals in the species are females
- They reproduce by gynogenesis (sperm-dependant parthenogenesis)

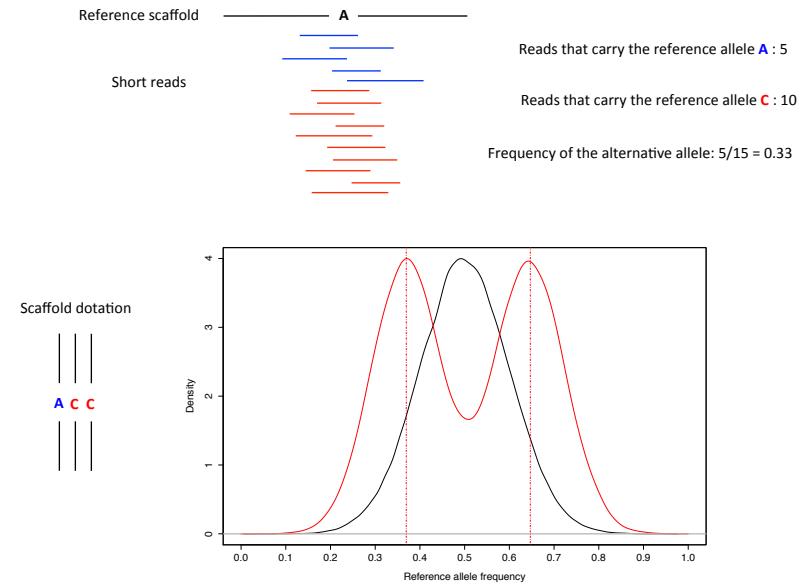


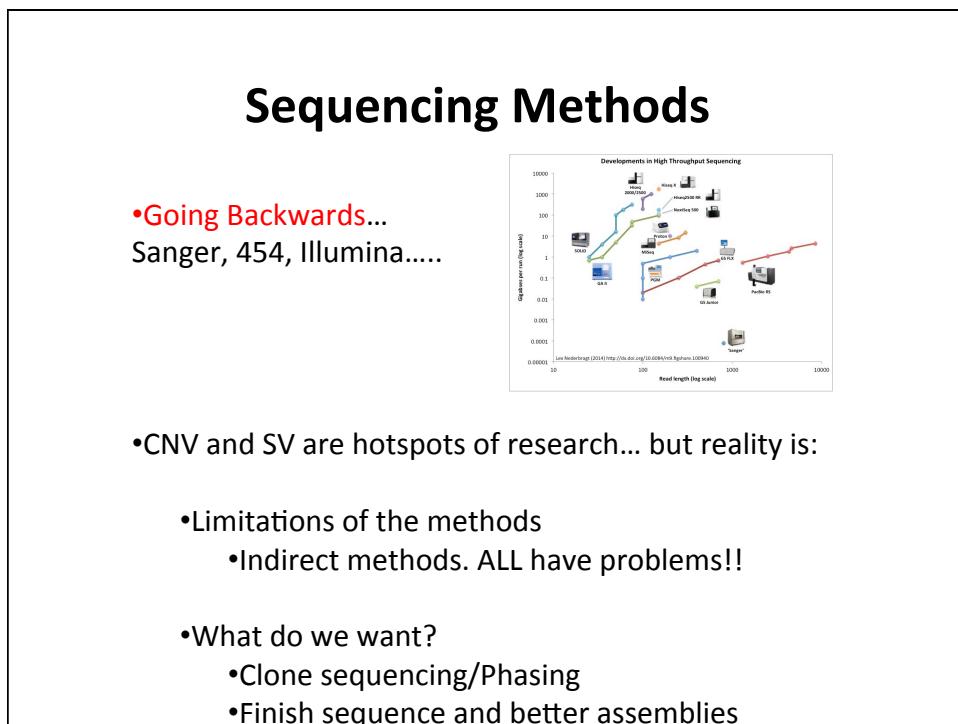
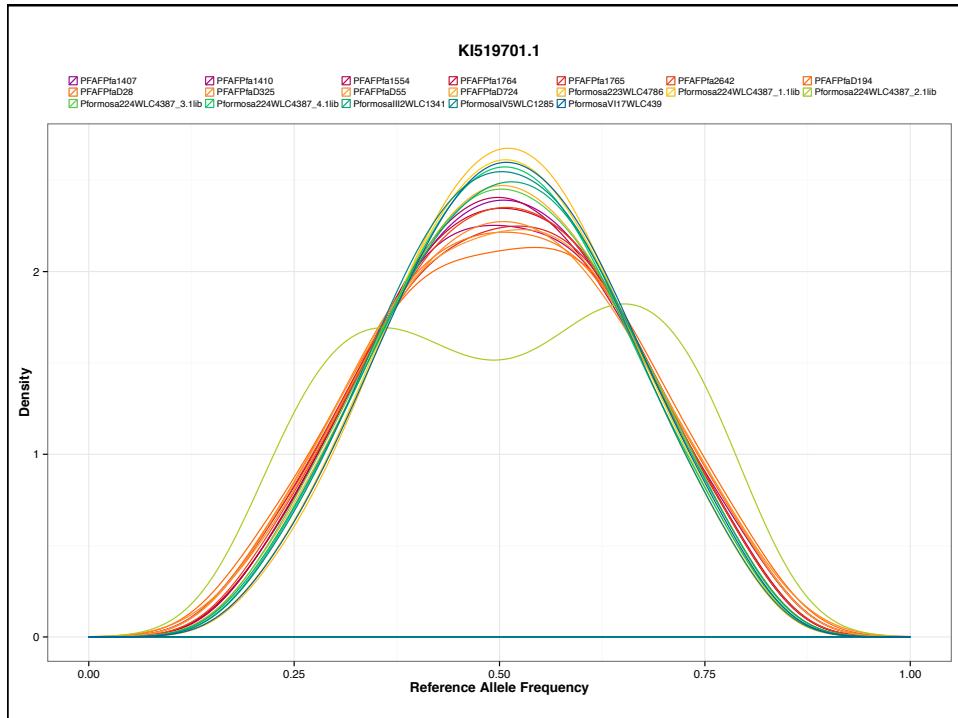
Lampert, K.P. 2006

Detecting paternal introgression



Detecting paternal introgression

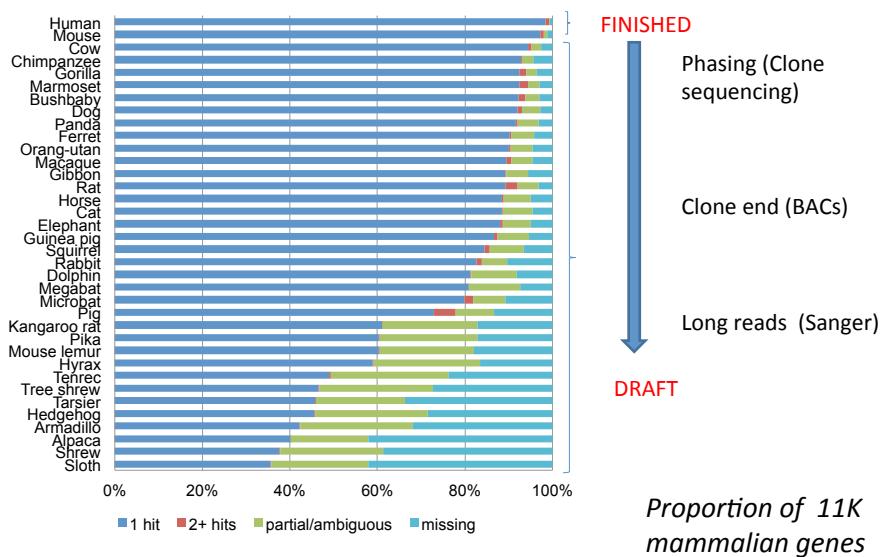




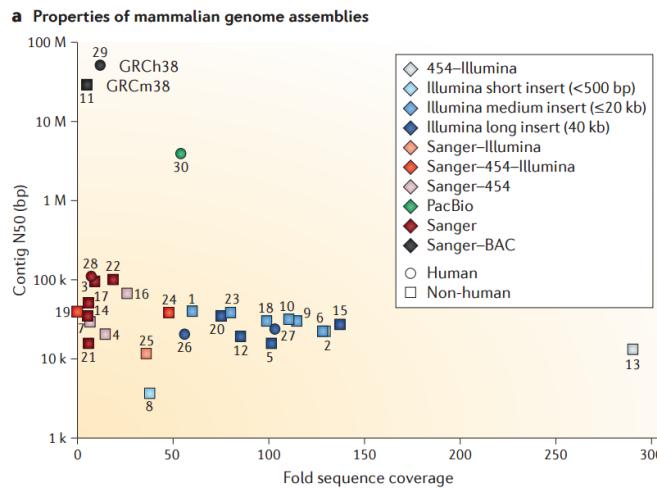
De novo assemblies

- Theory vs. Reality
- Most assemblies (even with Sanger technology!) are collapsed.

Quality of “old days” assemblies

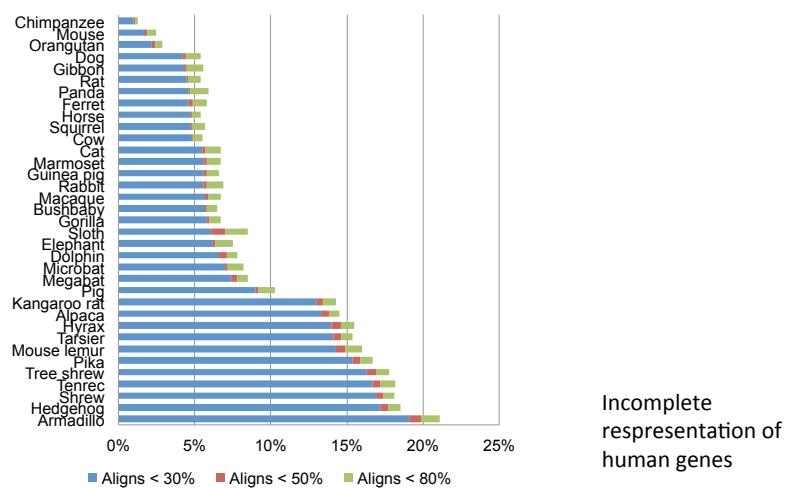


Contigs are small!



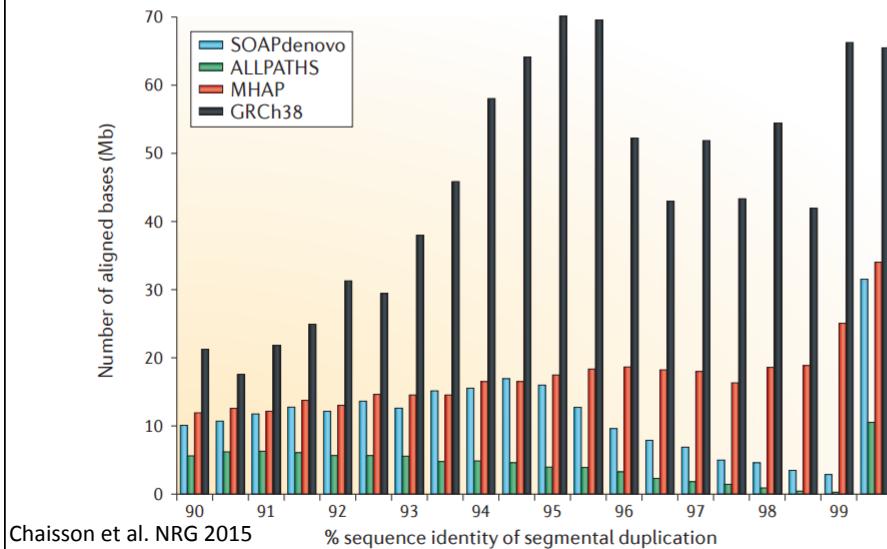
Chaisson et al. NRG 2015

Quality of assemblies (II)

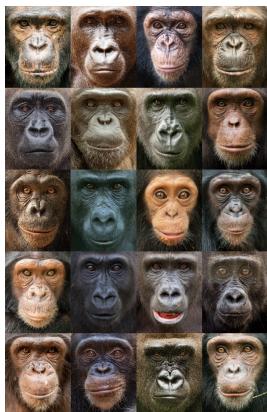


Limitations of NGS assemblies

b Duplicated sequences



Chimpanzee data to improve the reference



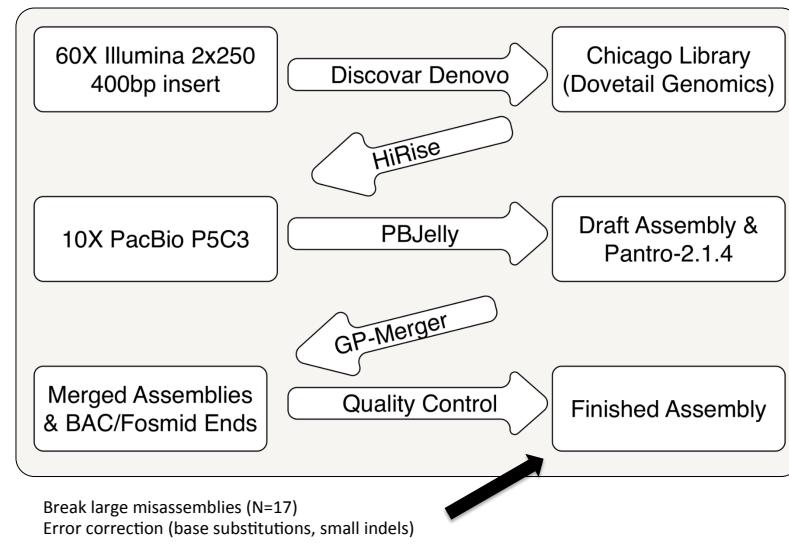
- 6x Sanger (plasmids, fosmids)
- BAC end reads
- 1000 finished BACs

New generated data

- Fosmid pool sequencing (\$\$\$\$)
- Bionano (optical mapping) (\$\$)
- 80x Illumina (short insert) (\$)
- 20 X Illumina mate pairs (\$)
- ~100X Illumina short ins size, PE overlap (\$)
- 10x PacBio from a 20k library (\$\$)*
- 3x Moleculo (\$\$)*
- 2 lanes HiC (Chicago) Dovetail (\$\$)

Wes Warren (WashU),
Andrew Sharp (Mount
Sinai),
Lars Feuk (Uppsala)

Our assembly pipe



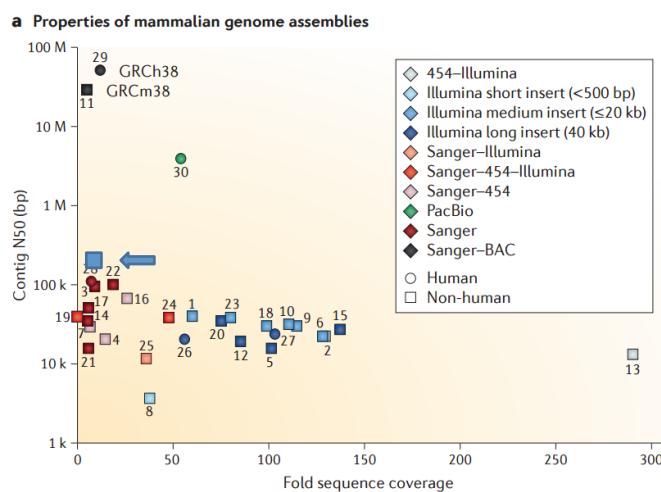
What is each data good for?

- **60 X Illumina 2x250**: Initial contiguous fragments (contigs) of the genome
- **Dovetail Genomics**: Long range information of order & orientation of contigs (scaffolds) (closed system)
- **10X PacBio**: Fill in remaining gaps (stretches of unknown sequence within scaffolds, between 2 contigs)

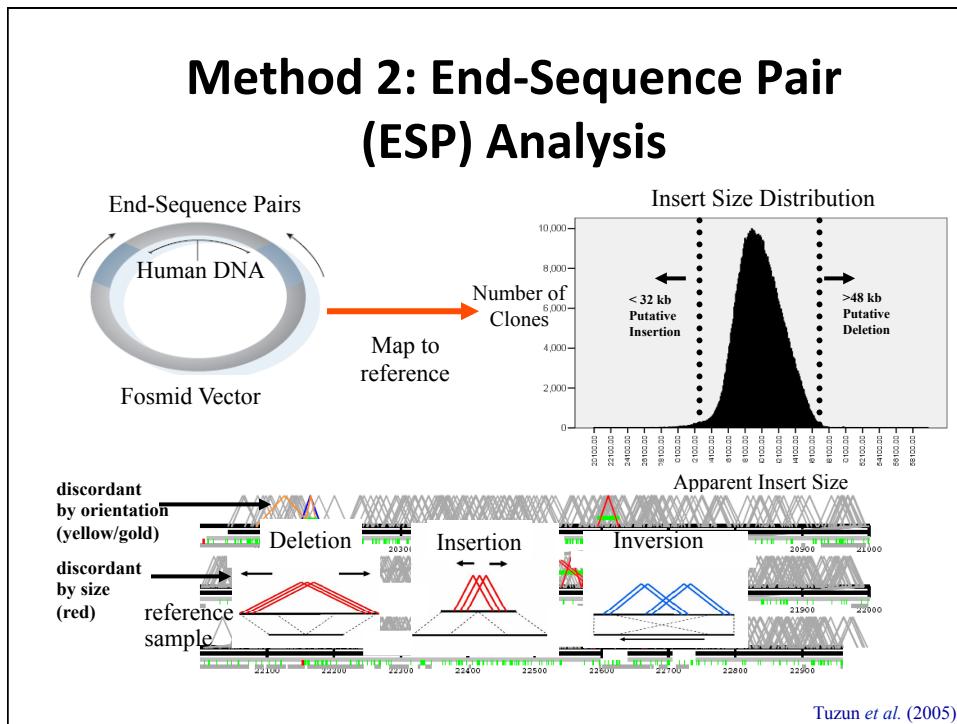
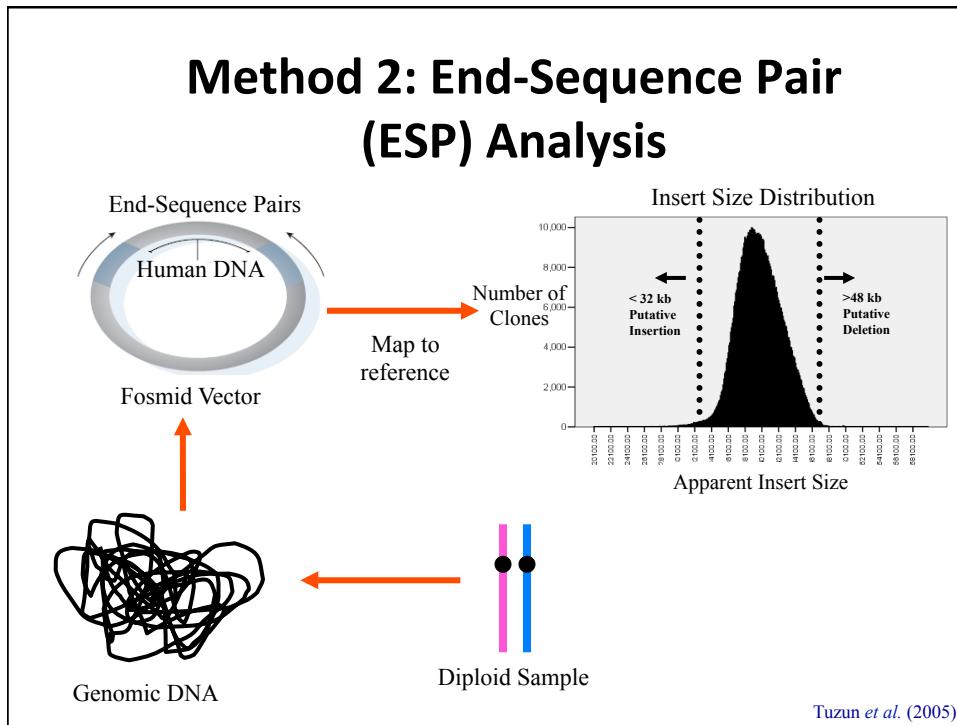
What do we get?

	Pantro-2.1.4	Pantro5V0.2 preliminary
Number of scaffolds	24,129	45,000
Scaffold N50 (bp)	8,925,874	26,673,241
Number of contigs	148,553	72,817
Contig N50 (bp)	64,231	334,510

Where will the chimp go?

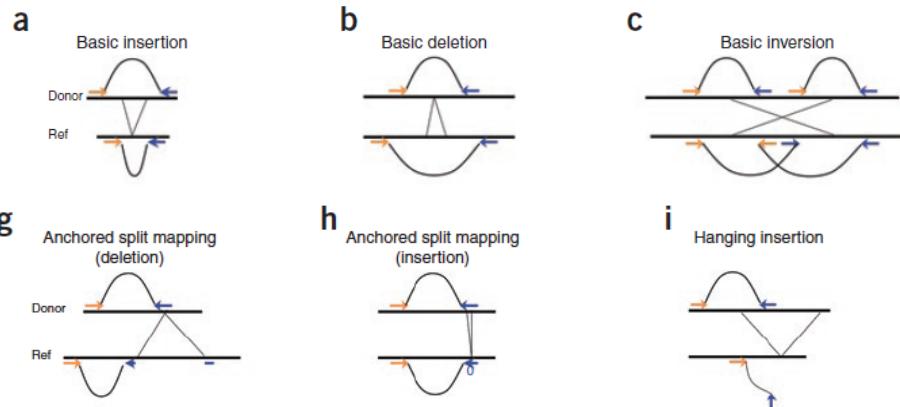


Chaisson et al. NRG 2015



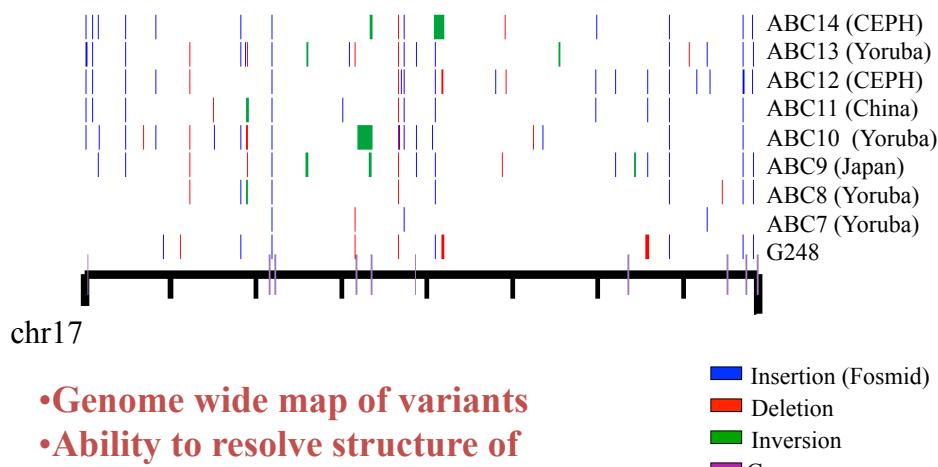
What can we find?

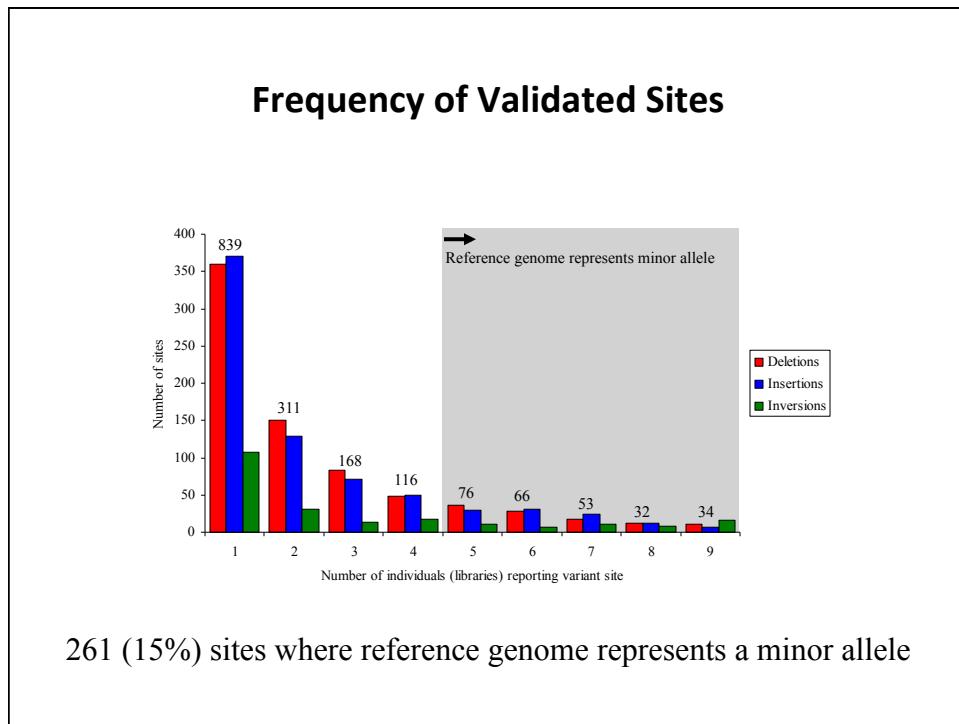
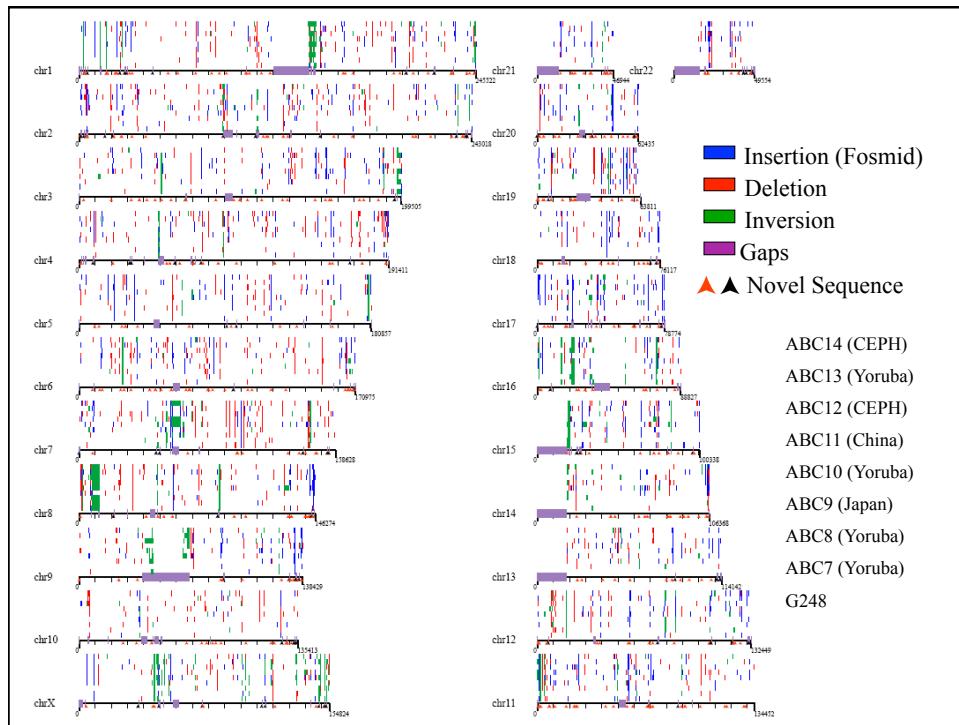
Structural variation detection:



Alkan et al. Nature Review Genetics 2011

Map of Validated Variants





Other application of PE analyses

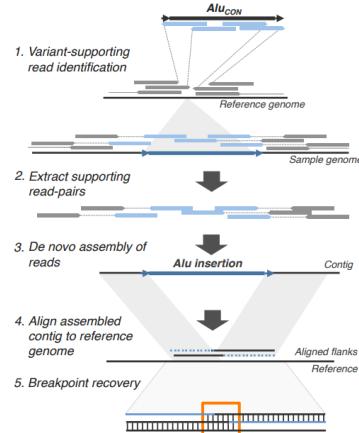
Figure 1

New Results

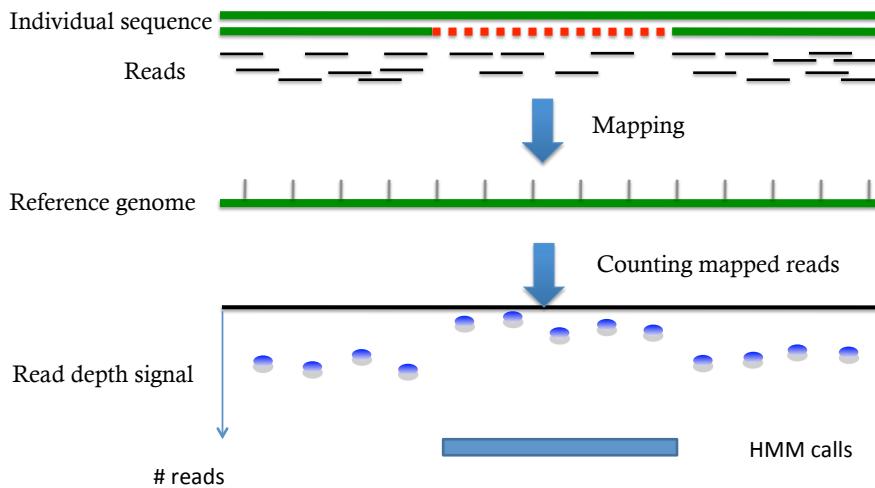
[View current version of this article](#)

Assembly of polymorphic Alu repeat sequences from whole genome sequence data in diverse humans

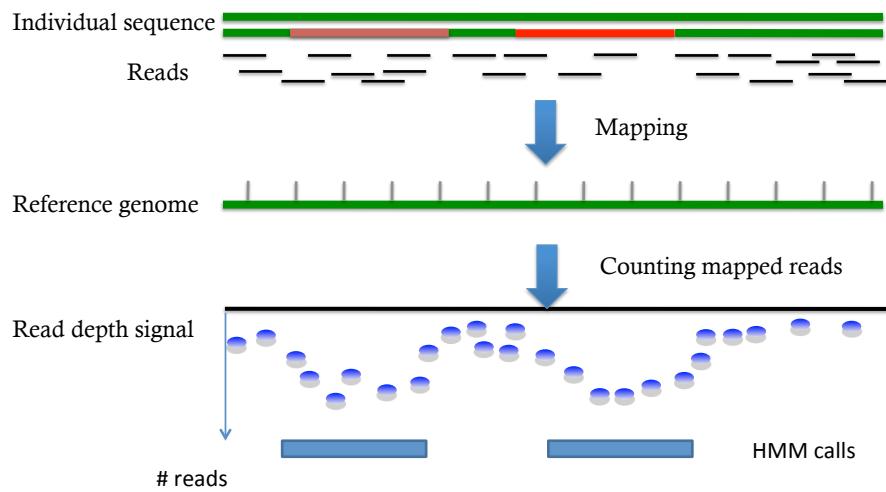
Julia H Wildschutte, Alayna A Baron, Nicolette M Diroff, Jeffrey M Kidd

doi: <http://dx.doi.org/10.1101/014977>

Method 3: Sequence Read Depth Analysis

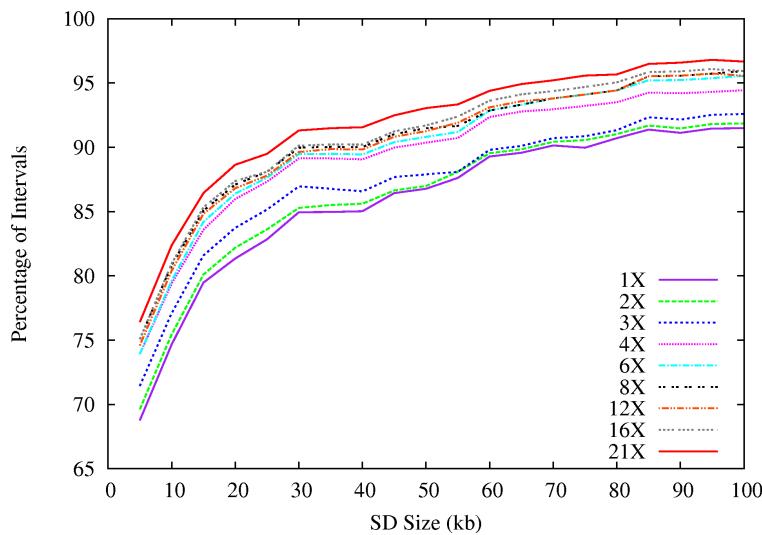


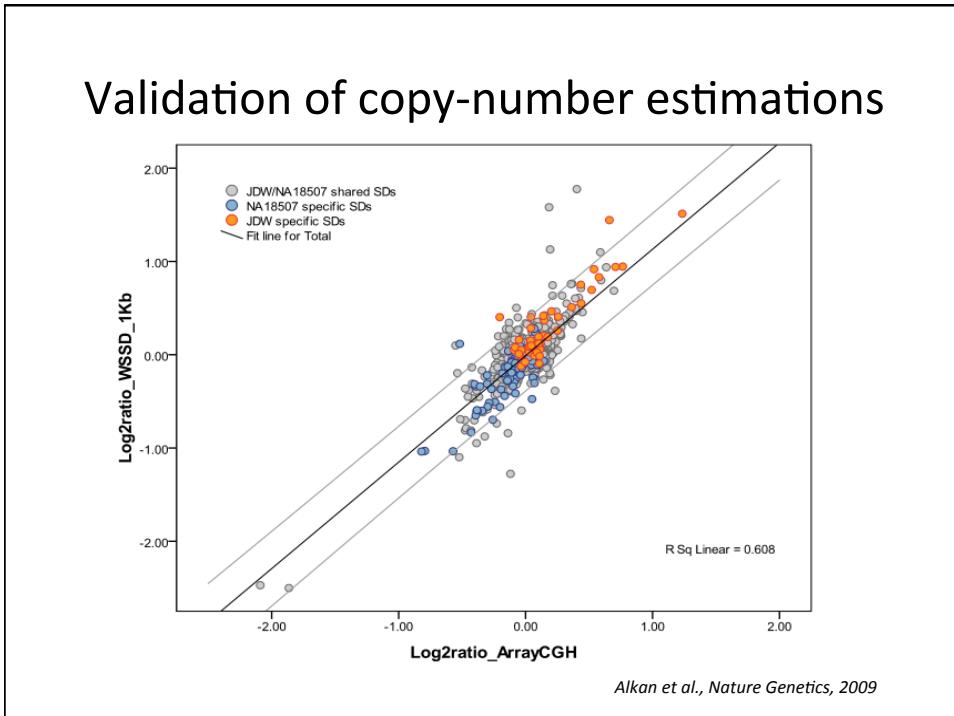
Method 3: Sequence Read Depth Analysis

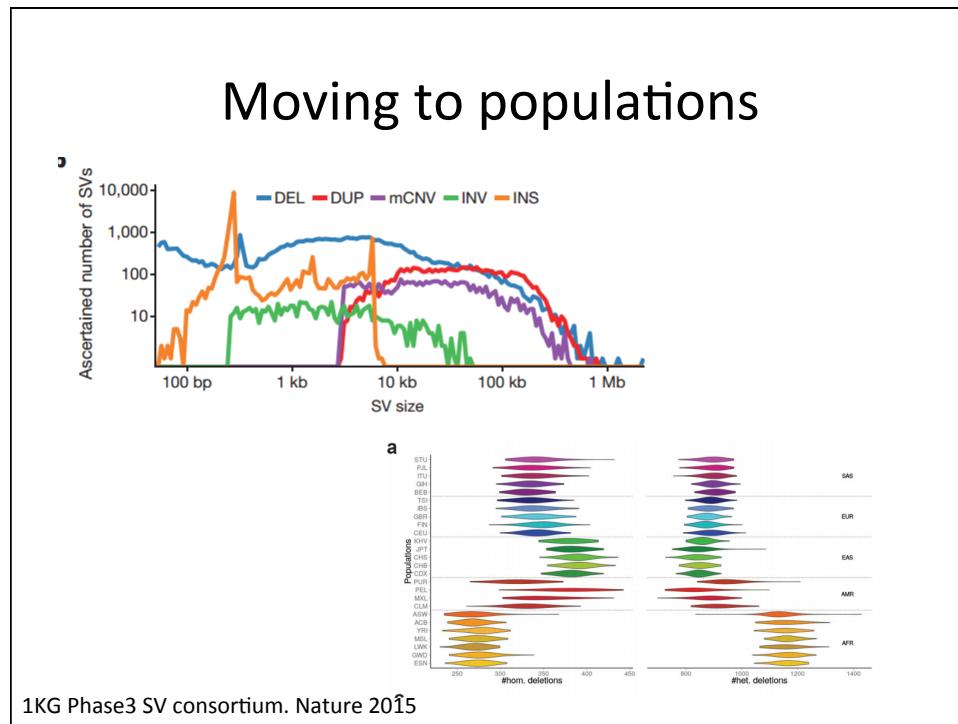
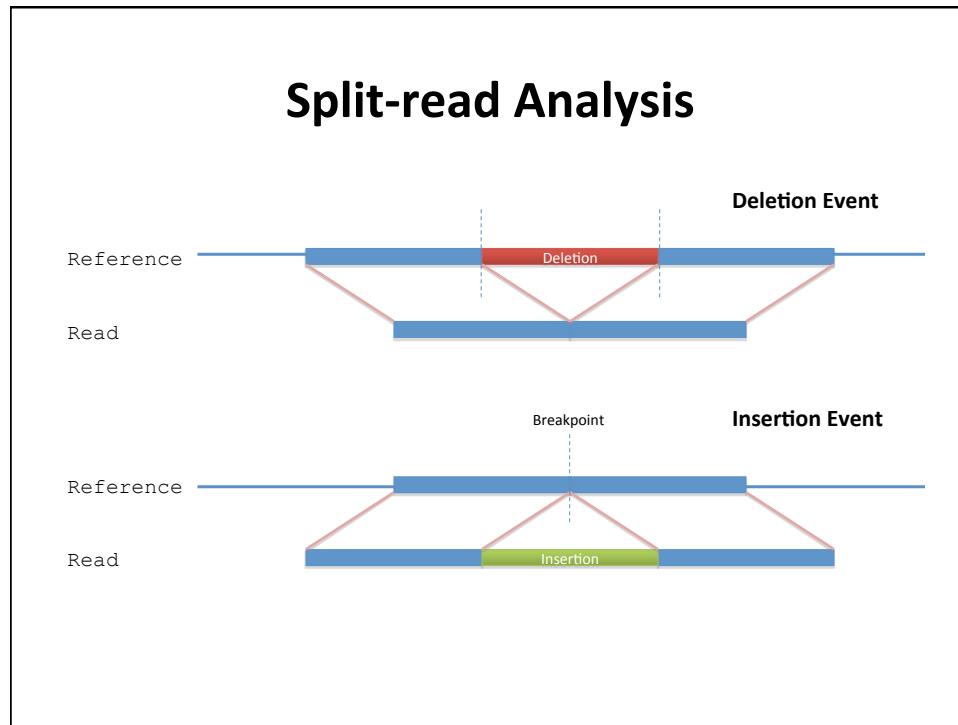


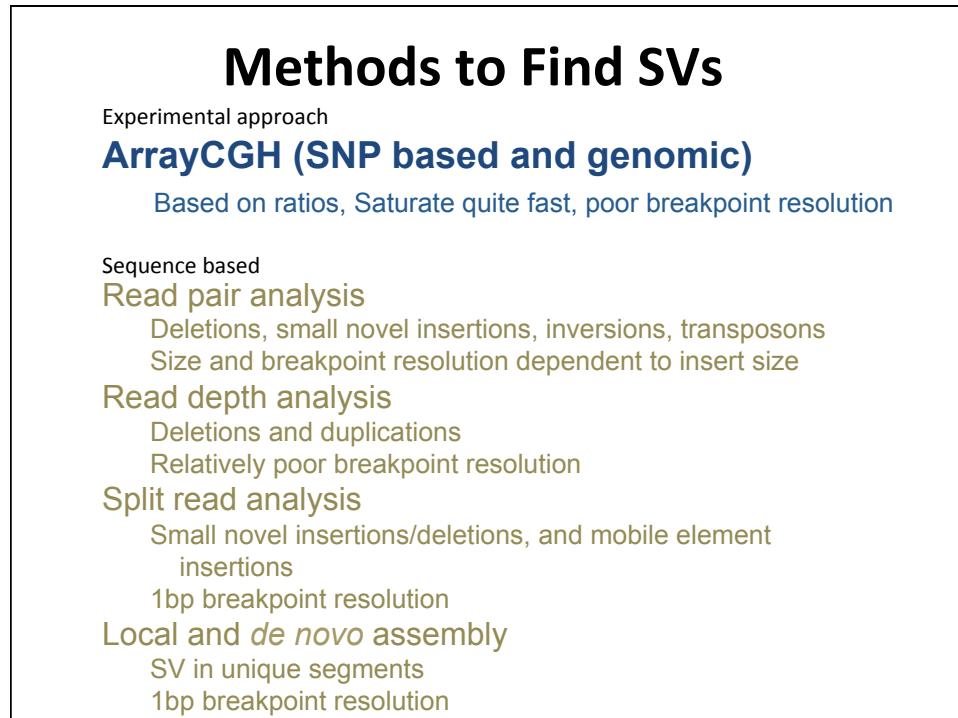
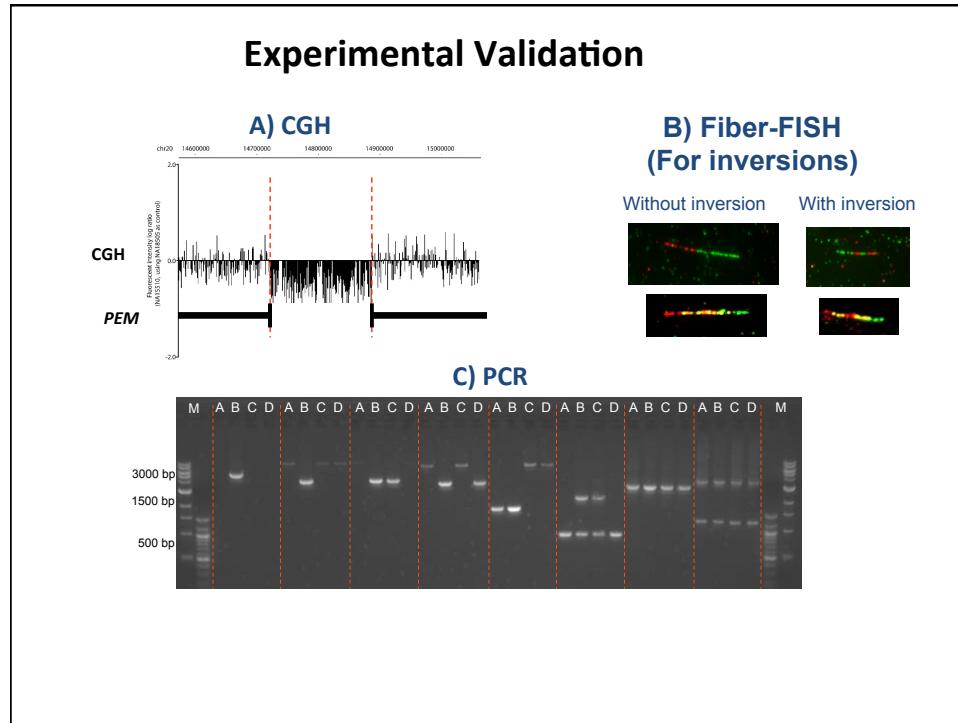
41

Sequence coverage and detection power

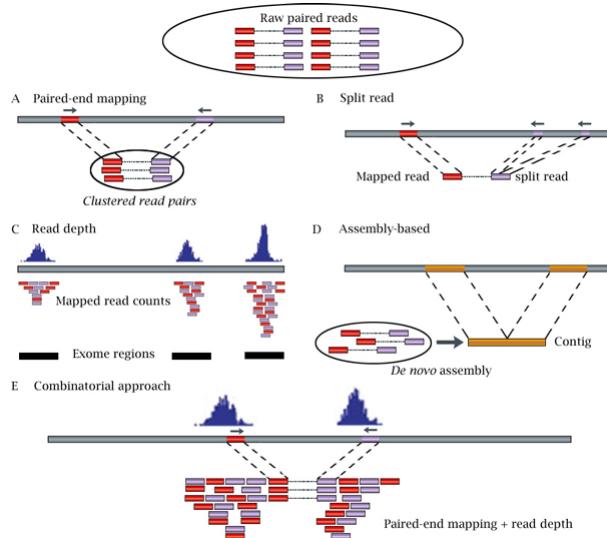








Review software



Software I

Method	Reference	Language	Control required?	Input format	GC correction	single-end/pair-end	Methodology characteristics
CNV-seq	[15]	R, perl	Yes	hits	No	single-end	statistical testing
FREEC	[21]	C	Optional	SAM,BAM,bed,etc	Optional	both	LASSO regression
readDepth	[22]	R	No	bed	Yes	both	CBS, LOESS regression
CNVnator	[23]	C	No	BAM	Yes	both	mean shift algorithm
SegSeq	[14]	Matlab	Yes	bed	No	single-end	statistical testing,CBS
EWT (RDxplorer)	[11]	R, python	No	BAM	Yes	single-end	statistical testing
cnD	[16]	D	No	SAM,BAM	No	both	HMM, Viterbi algorithm
CNVer	[17]	C	No	BAM	Yes	pair-end	maximum-likelihood, graphic flow
CopySeq	[18]	Java	No	BAM	Yes	pair-end	MAP estimator
rSW-seq	[19]	NA	Yes	NA	Yes	single-end	Smith-Waterman algorithm
CNAseg	[20]	R	Yes	BAM	No	pair-end	wavelet transform and HMM
CNAnorm	[24]	R	Yes	SAM,BAM	Yes	both	linear regression or CBS
cn.MOPS	[26]	R, C++	multiple samples	BAM or data matrix	No	both	mixture of Poissons, MAP, EM, CBS
JointSLM	[27]	R, Fortran	multiple samples	data matrix	Yes	both	HMM, ML estimator, Viterbi algorithm

doi:10.1371/journal.pone.0059128.t001

break point position estimation: readDepth = EWT>CNVnator>FREEC>CNV-seq>SegSeq;
copy number estimation: CNVnator>CNV-seq>readDepth>FREEC>EWT>SegSeq;

Zhao et al. BMC Bioinformatics 2013

Duan et al. Plos One 2013