

Informatics tools for next-generation sequencing analysis



Gabor Marth
Boston College Biology

Next-Generation Sequencing Meeting
Barcelona
October 1-3, 2009

New sequencing technologies...

Illumina



Capillary (e.g. AB 3730)



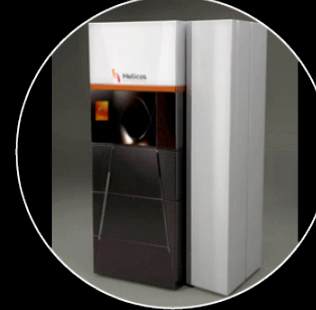
Roche 454



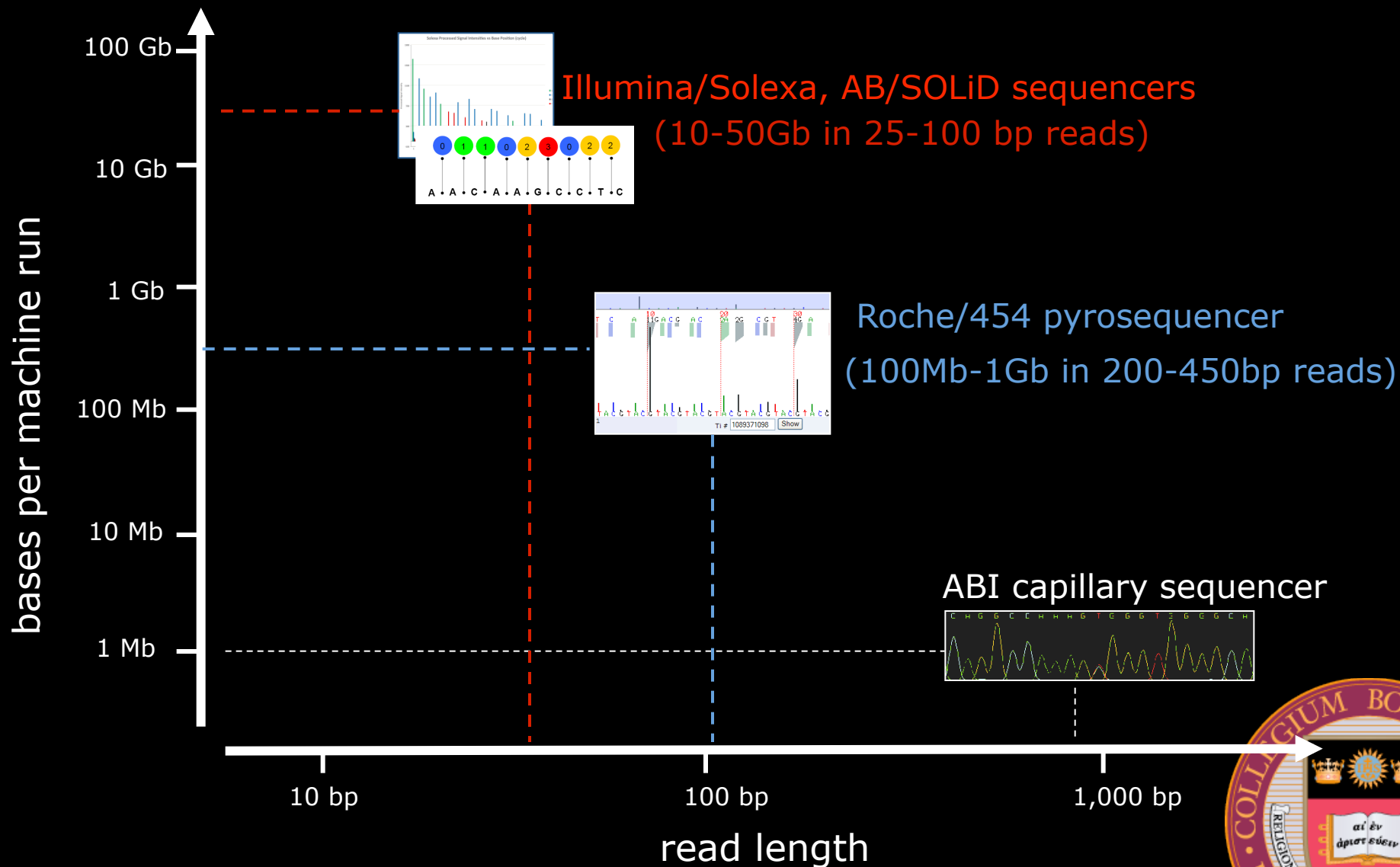
AB SOLiD



Helicos



... offer vast throughput ...



... & enable personal genome sequencing

nature Vol 452|17 April 2008|doi:10.1038/nature06884

LETTERS

The complete genome of a human individual

parallel DNA sequencing

David A. Wheeler^{1*}, Maithreya Venkatesh², Wen He², Yi-Ju Chen², Vinod M. Rao², Cynthia L. Turcotte², Gerard P. Smith², Lynne Nazareth¹, Xiang Qin¹, David A. Wheeler¹, & Jonathan M. Rothberg^{2†}

Home About Partners Data Contact Wiki

nature

ARTICLES

The diploid genome of a human individual

Jun Wang^{1,2,3,4*}, Wei Wang^{1,3*}, Junqing Zhang¹, Jun Li¹, Juanbin Zhou¹, Huiqing Liang¹, Zhenglin Du¹, Dongsheng Hu¹, Ines Hellmann⁵, Michael Inouye⁶, Guoqing Li¹, Zhentao Yang¹, Guoqing Li¹, Dawei Li¹, Peixiang Ni¹, Jue Ruan¹, Jianguo Zhang¹, Jia Ye¹, Lin Fang¹, Shuang Yang¹, Fang Chen^{1,7}, Li Li¹, Guohua Yang^{1,2}, Zhuo Li¹, Xiaoli Li¹, Richard Durbin⁸, Lars Bolund^{1,11}, Jun Wang^{1,2,3,4*}, Wei Wang^{1,3*}, Junqing Zhang¹, Jun Li¹, Juanbin Zhou¹, Huiqing Liang¹, Zhenglin Du¹, Dongsheng Hu¹, Ines Hellmann⁵, Michael Inouye⁶, Guoqing Li¹, Zhentao Yang¹, Guoqing Li¹, Dawei Li¹, Peixiang Ni¹, Jue Ruan¹, Jianguo Zhang¹, Jia Ye¹, Lin Fang¹, Shuang Yang¹, Fang Chen^{1,7}, Li Li¹, Guohua Yang^{1,2}, Zhuo Li¹, Xiaoli Li¹, Richard Durbin⁸, Lars Bolund^{1,11}, Jun Wang^{1,2,3,4*}, Wei Wang^{1,3*}, Junqing Zhang¹, Jun Li¹, Juanbin Zhou¹, Huiqing Liang¹, Zhenglin Du¹, Dongsheng Hu¹, Ines Hellmann⁵, Michael Inouye⁶, Guoqing Li¹, Zhentao Yang¹, Guoqing Li¹, Dawei Li¹, Peixiang Ni¹, Jue Ruan¹, Jianguo Zhang¹, Jia Ye¹, Lin Fang¹, Shuang Yang¹, Fang Chen^{1,7}, Li Li¹, Guohua Yang^{1,2}, Zhuo Li¹, Xiaoli Li¹, Richard Durbin⁸, Lars Bolund^{1,11}

1000 Genomes

A Deep Catalog of Human Genetic Variation

INTERNATIONAL CONSORTIUM ANNOUNCES THE 1000 GENOMES PROJECT

Major Sequencing Effort Will Produce Most Detailed Map Of Human Genetic Variation to Support Disease Studies

An international research consortium has been formed to create the most detailed and medically useful picture to date of human genetic variation. The 1000 Genomes Project will involve sequencing the genomes of at least a thousand people from around the world. The project will receive major support from the [Wellcome Trust Sanger Institute](#) in Hinxton, England, the [Beijing Genomics Institute Shenzhen](#) in China and the [National Human Genome Research Institute](#) (NHGRI), part of the [National Institutes of Health](#) (NIH).

Drawing on the expertise of multidisciplinary research teams, the 1000 Genomes Project will develop a new map of the human genome that will provide a view of biomedically relevant DNA variations at a resolution unmatched by current resources. As with other major human genome reference projects, data from the 1000 Genomes Project will be made swiftly available to the worldwide scientific community through freely accessible public databases.

OPEN ACCESS Freely available online PLOS BIOLOGY

The Diploid Genome Sequence of an Individual Human

David A. Wheeler¹, Maithreya Venkatesh², Wen He², Yi-Ju Chen², Vinod M. Rao², Cynthia L. Turcotte², Gerard P. Smith², Lynne Nazareth¹, Xiang Qin¹, David A. Wheeler¹, & Jonathan M. Rothberg^{2†}

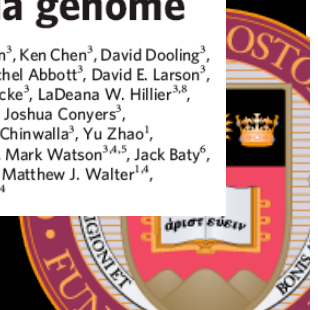
Genetics and Genomic Biology, The Hospital for Sick Children, and Molecular and Computer Science and Engineering, University of California San Diego, La Jolla, California, United States of America, Institut de Barcelona, Barcelona, Catalonia, Spain

It was produced from ~32 million random DNA fragments assembled into 4,528 scaffolds, comprising 2,810 million-fold coverage for any given region. We developed a pipeline for identification and comparison of alternate alleles within this reference genome. We identified 12.3 Mb of variants (of which 1.5 Mb are polymorphisms (SNPs), 53,823 block substitutions (2–206 bp), 559,473 homozygous indels (1–82,711 bp), 90 copy number variation regions. Non-SNP DNA variation at these sites they involve 74% of all variant bases. This suggests an alternative diploid genome structure. Moreover, 44% of genes were not covered by our assembly strategy, we were able to span 1.5 Gb of the genome.

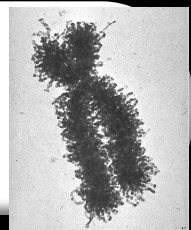
Vol 456|6 November 2008|doi:10.1038/nature07485

Genetically diverse human genome

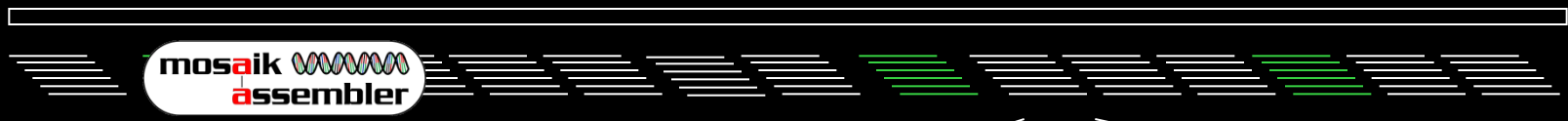
Michael D. McLellan³, Ken Chen³, David Dooling³, David A. Wheeler¹, Lisa Cook³, Rachel Abbott³, David E. Larson³, David E. Larson³, Devin Locke³, LaDeana W. Hillier^{3,8}, Jarret Glasscock³, Joshua Conyers³, David Gordon⁹, Asif Chinwalla³, Yu Zhao¹, David E. Larson³, H. Tomasson^{1,4}, Mark Watson^{3,4,5}, Jack Baty⁶, Anand Mehta^{4,5}, Matthew J. Walter^{1,4}, Richard K. Wilson^{2,3,4}



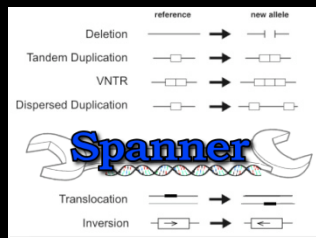
The re-sequencing informatics pipeline



REF
IND



(ii) read mapping



```

GTT*ACT*GTCGTTGT*AA*TACTCC*AA*CGATGCTTTTCAGGG*TC*ATAAAGAT
GTT*ACT*GTCGTTGT*AA*TACTCC*AA*CGATGCTTTTCAGGG*TC*ATAAAGAT
tt*act*gtt*aa*ga*at*ca*ca*aa*gt*tt*aa*aa*cc*ca*aa*aa*cc*cc*cc*cc*
gTT*ACT*GtcGTTGT*AA*TACTCC*AA*cgatgCTTTTCAGGG*tc*cc*ATAAAGAT
GTT*ACT*GTCGTTGT*AA*TACTCC*AA*CGATGCTTTTCAGGG*TC*ATAAAGAT
gtt*act*gc gttg aa*tactcc*aa*cgatgcttttcagg*tc*cc*ATAAAGAT
GTT*ACT*GTCGTTGT*AA*TACTCC*AA*CGATGCTTTTCAGGG*TC*ATAAAGAT
GTT*ACT*GTCGTTGT*AA*TACTCC*AA*CGATGCTTTTCAGGG*TC*ATAAAGAT
gtt act*gtcgtt*gt*aa tactcc*aa cgatgcttttcagg*tc*cc ATAAGAT
GTT*a t*gcgTTGT*AA*TACTCC*AA*CGATGCTTTTCAGGG*TC*ATAAAGAT
gtt*act*gcgTTgt aa*tactcc a*cgatgCTtttcagg*TC*cc ataaagat
gtt*act*gtcgtt*gt SACCAATCTAA*AAATACCTGTGA *ATAAAGAT
GTT*act*gtcgtt*gt SACCAATCTAA*AAATACCTGTGA *ATAAAGAT
gtt*ac * t*cgTTGT *ATAAAGAT
GTT*act g cgtt*gt *ATAAAGAT
gtt*a t*gtcgtt*gt *ataaagat
GTT*act*gcgTTgt*aa*ga*at*ca*ca*aa*gt*tt*aa*aa*cc*ca*aa*aa*cc*cc*cc*cc*
GTT*ACT*GTCGTTGT*AA*TACTCC*AA*CGATGCTTTTCAGGG*TC*ATAAAGAT
gtt*act*gcgTTgt*aa*ga*at*ca*ca*aa*gt*tt*aa*aa*cc*ca*aa*aa*cc*cc*cc*cc*
  
```

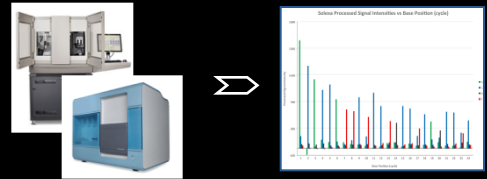
(iv) SV calling

IND



(iii) SNP and short INDEL calling

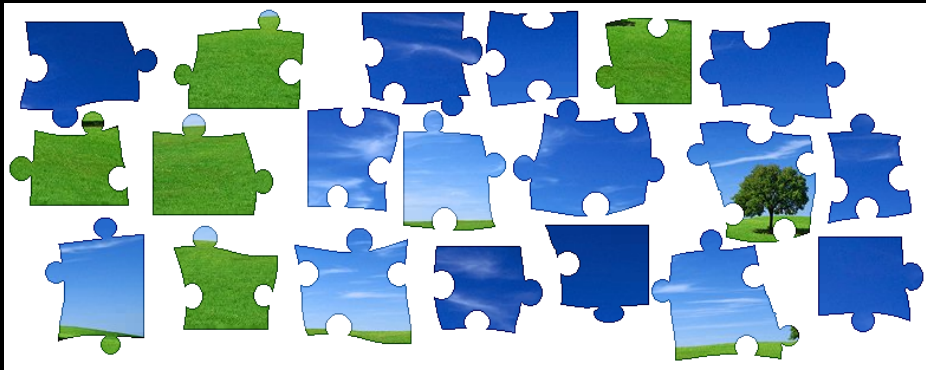
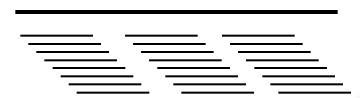
(i) base calling



(v) data viewing, hypothesis generation



Read mapping is like a jigsaw



...you get the pieces...

... and they give you the picture on the box

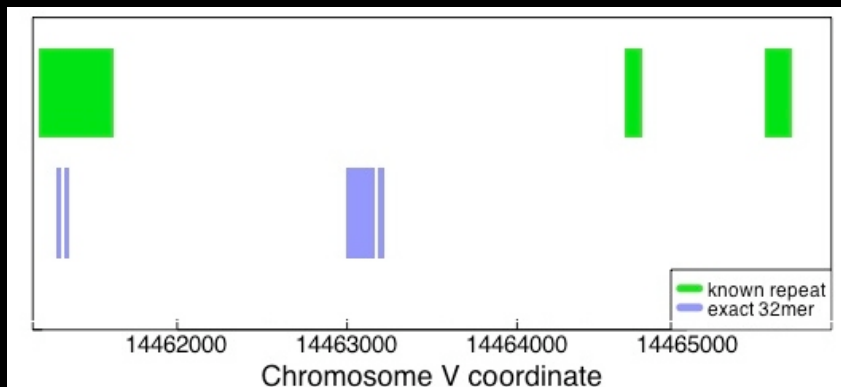
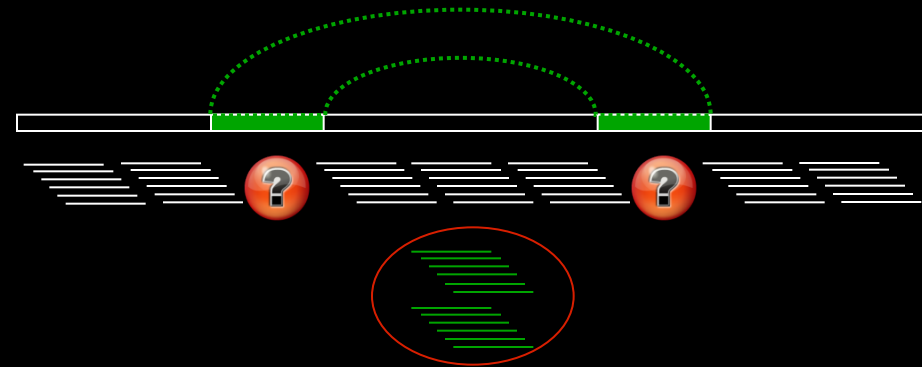


Unique pieces are easier to place than others...



Multiply-mapping reads

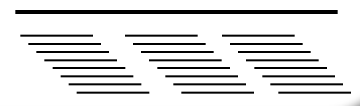
- Reads from repeats cannot be uniquely mapped back to their true region of origin



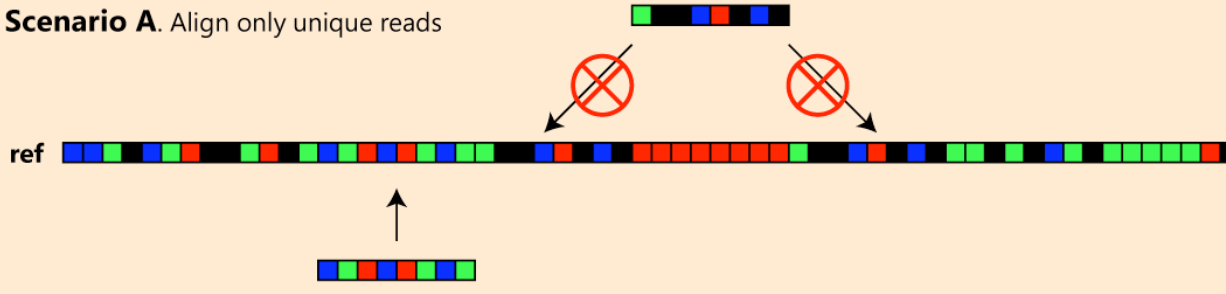
- “Traditional” repeat masking does not capture repeats at the scale of the read length



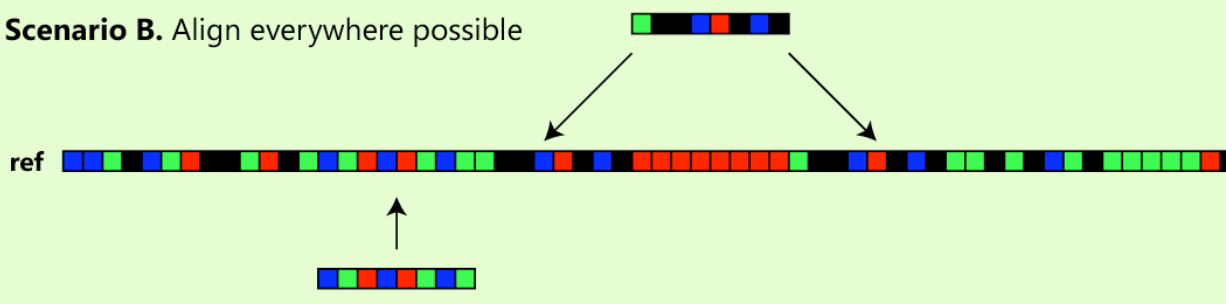
Dealing with multiple mapping



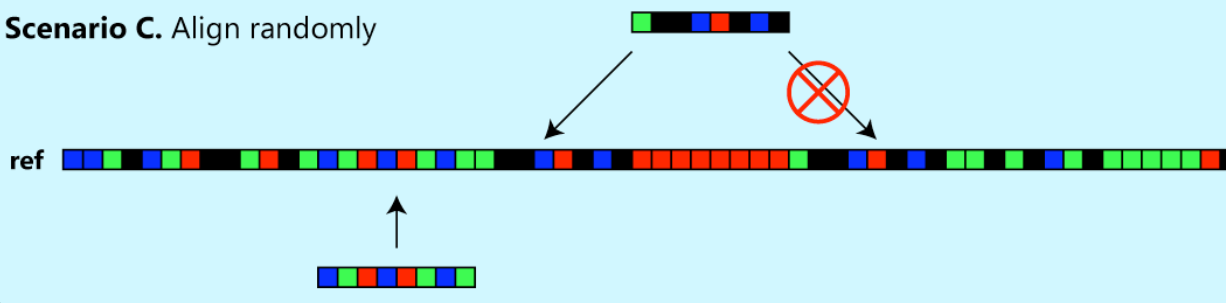
Scenario A. Align only unique reads



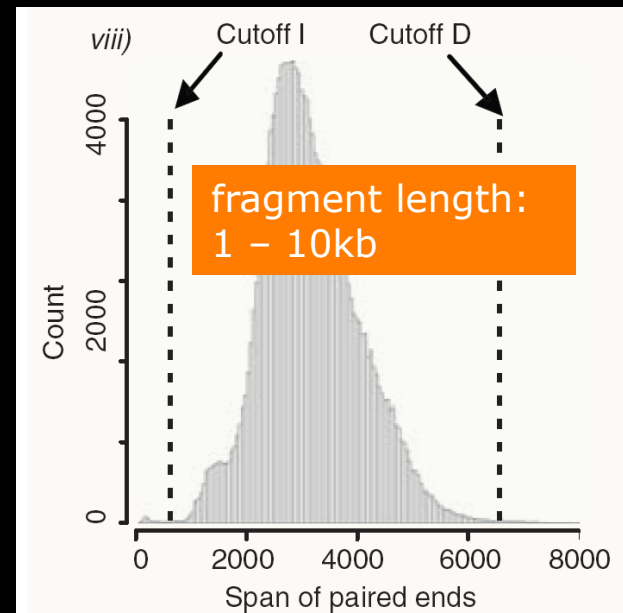
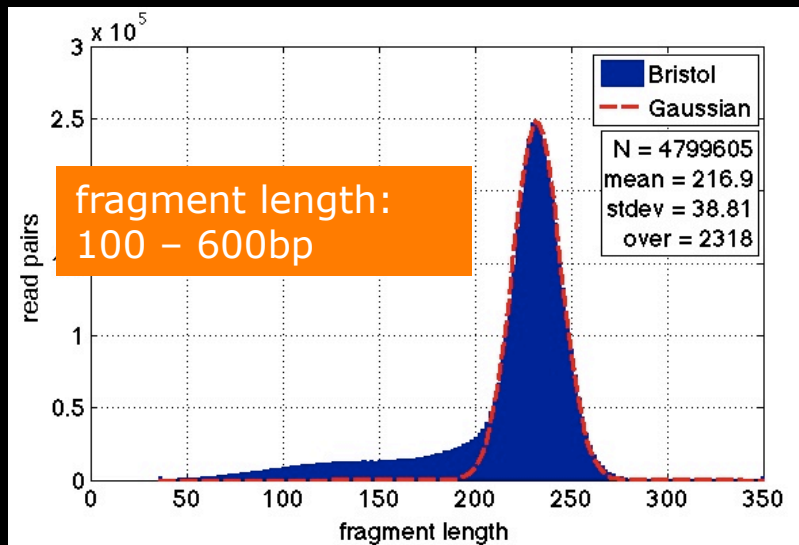
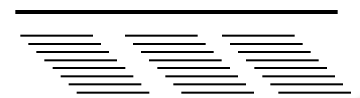
Scenario B. Align everywhere possible



Scenario C. Align randomly



Paired-end (PE) reads

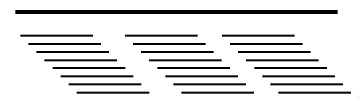


Korbel *et al.* **Science** 2007

PE reads are now the standard for whole-genome short-read sequencing



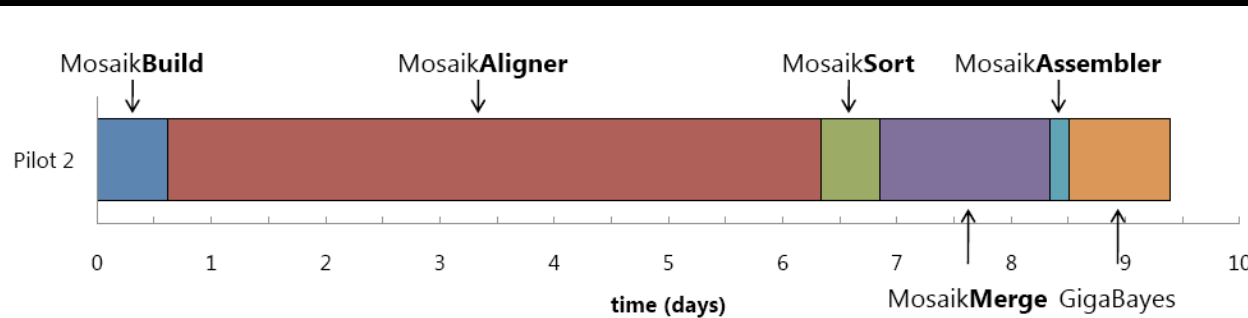
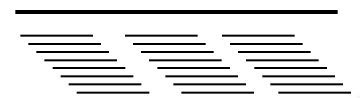
Gapped alignments (for INDELs)



```
tttatttaggctgagcaataatag
tttatttaggctgagcaataatag
tttatttaggctgagcaataatag
tttatttaggctgagc**taatagacg
      ttaggctgagcaataatagacg
        aggctgagc**taatagacg
          aggctgagc**taatagacg
            gctgagc**taatagacg
              tgagc**taatagacg
                tgagc**taatagacg
                  tgagcaataatagacg
                    gagc**taatagacg
                      gagc**taatagacg
                        gagc**taatagacg
                          agcaataatagacg
                            gc**taatagacg
                              taatagacg
                                agacg
                                  gacg
```



The MOSAIK read mapper



mosaik 
assembler

Michael Strömberg

Pilot 2 (January 2009)

6 billion Illumina & 454 reads
9 node cluster
Alignment time: 6 days
169 reads/core/second

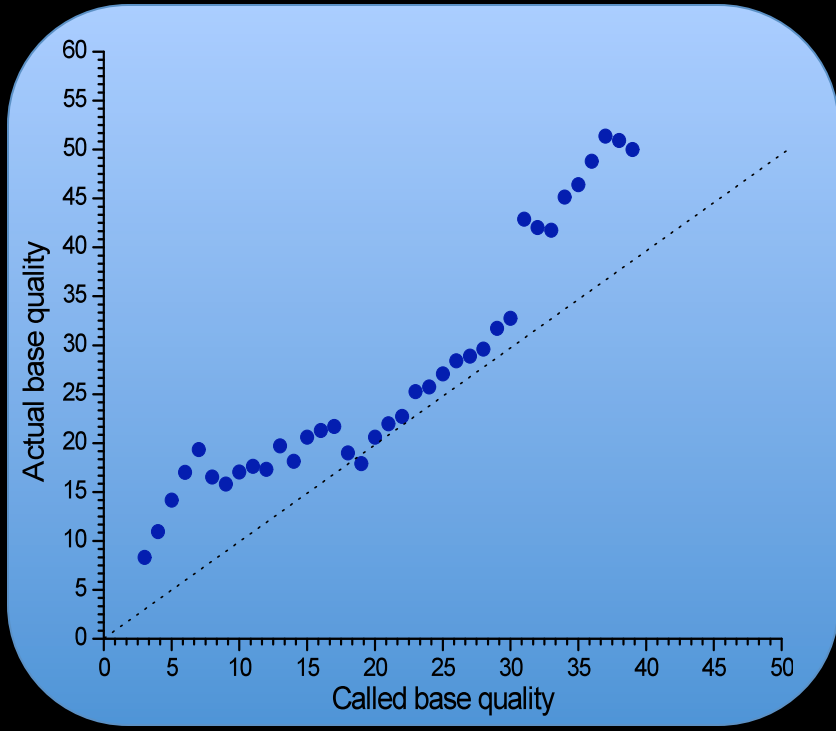
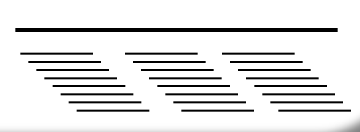
Pilot 1 (August 2009)

28 billion Illumina & 454 reads
31 node cluster
Alignment time: 7 days
195 reads/core/second (+16%)

- gapped mapper
- option to report multiple map locations
- aligns 454, Illumina, SOLiD, Helicos reads
- works with standard file formats (SRF, FASTQ, SAM/BAM)

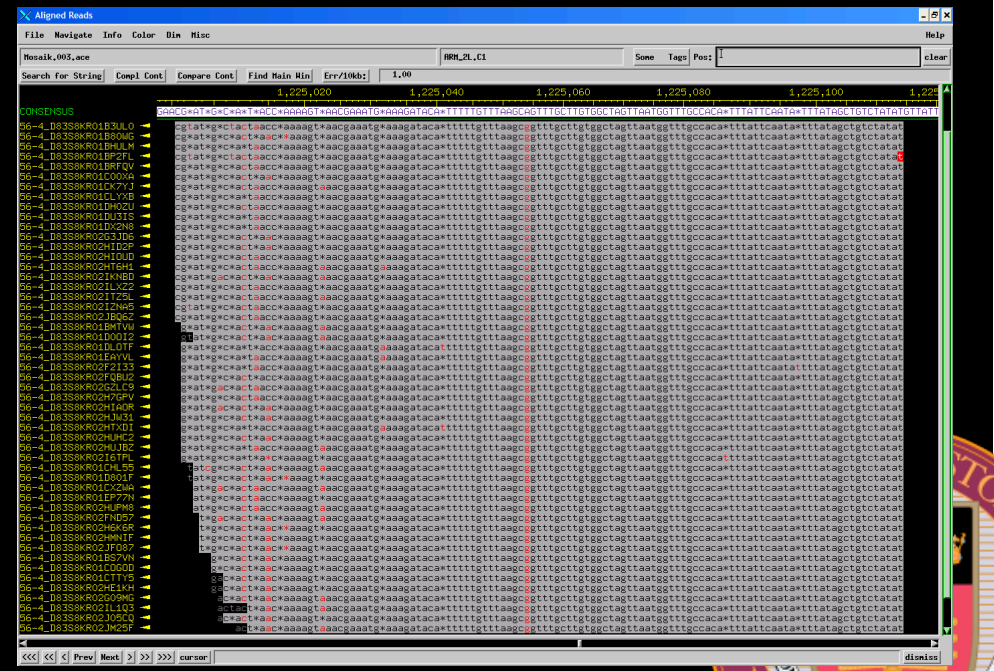


Alignment post-processing

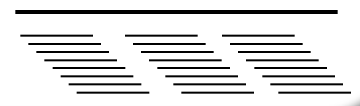


- quality value re-calibration

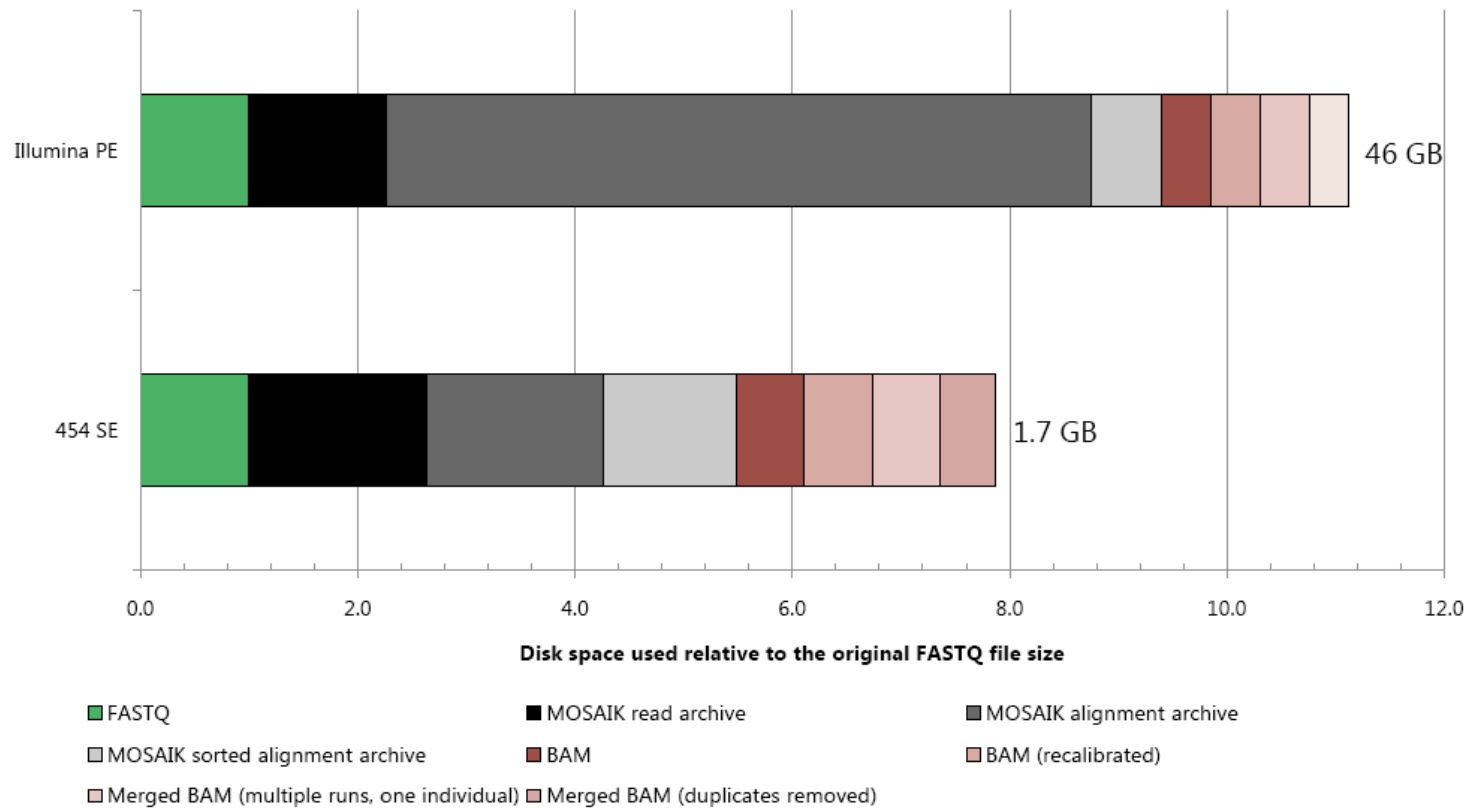
- duplicate fragment removal



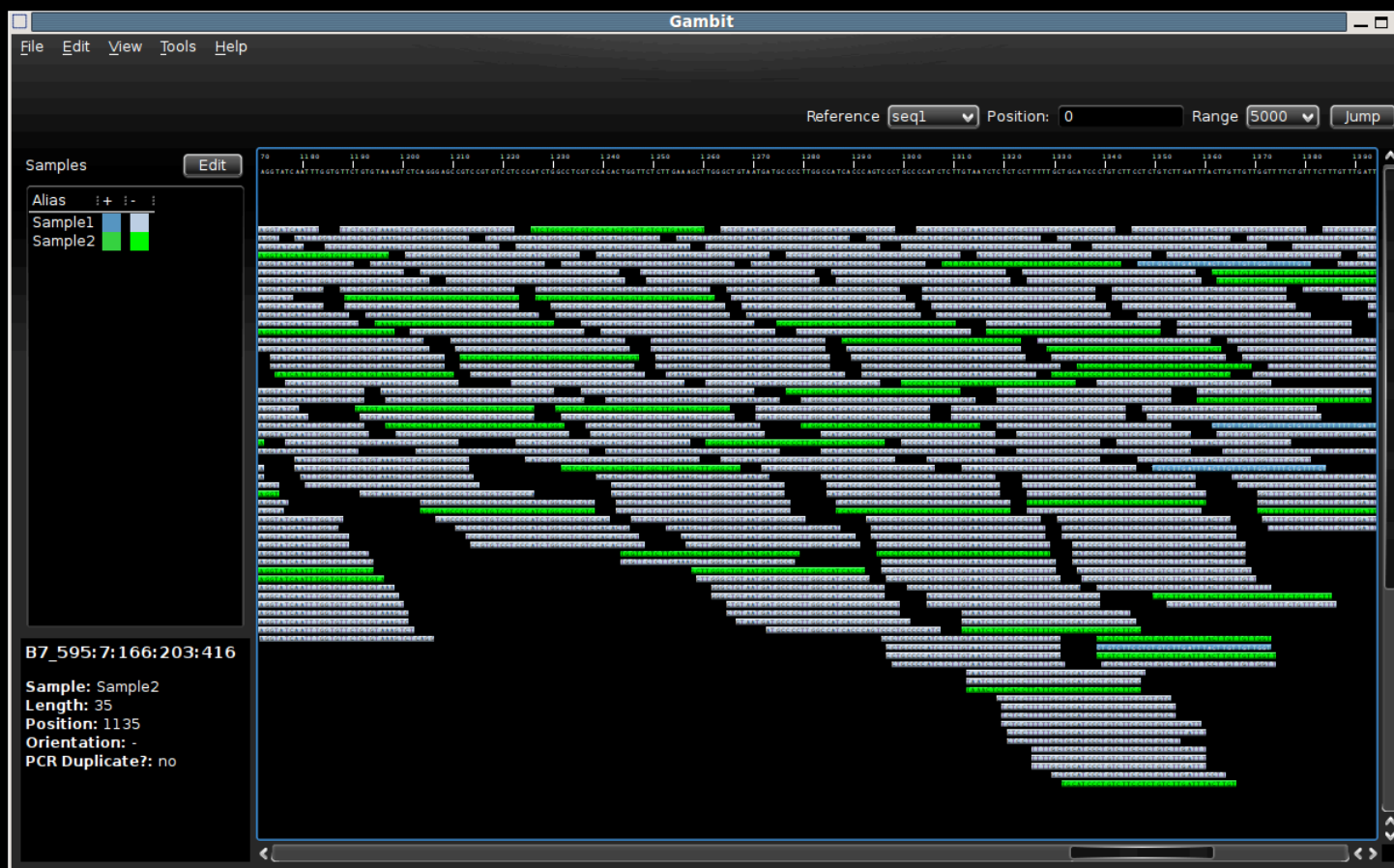
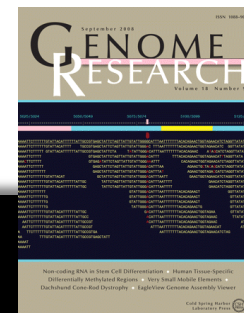
Data storage requirements



Pipeline Disk Usage



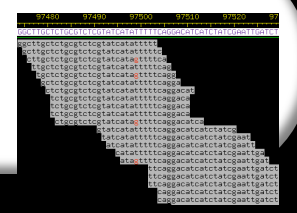
Alignment visualization



- too much data – indexed browsing
- too much detail – color coding, show/hide



SNP calling: old problem, new data



```

GTT*ACT*GTCGTTGT*AA*TACTCC*A*CGATGTCTTTCAGGG*TCTCC*ATAAAGAT
GTT*ACT*GTCGTTGT*AA*TACTCC*A*CGATGTCTTTCAGGG*TCTCC*ATAAAGAT
*tt*act*gtaatggaatactcatgaagtgttaagggctcaaaaagaagcctccggcctt
gTT*ACT*GtcGTTGT*AA*TACTCC*a*cgatgtCTTTCAGGG*tctcc*atAAAGat
GTT*ACT*GTCGTTGT*AA*TACTCC*A*CGATGTCTTTCAGGG*TCTCC*ATAAAGAT
tgt*act*gaaagtgc*aa*tactCc*a*cgATGTctttcaGGG*TCTcc*aTAAAGAT
GTT*ACT*GTCGTTGT*AA*TACTCC*A*CGATGTCTTTCAGGG*TCTCC*ATAAAGAT
GTT*ACT*GTCGTTGT*AA*TACTCC*A*CGATGTCTTTCAGGG*TCTCC*ATAAAGAT

```

$$\Pr(G_1, G_2, \dots, G_n | B) = \frac{\prod_{i=1}^n \left[\sum_{\forall T^k} \Pr(B_i | T_i^k) \Pr(T_i^k | G_i) \right] \Pr(G_1, G_2, \dots, G_n)}{\sum_{\forall G^l} \left\{ \prod_{i=1}^n \left[\sum_{\forall T^k} \Pr(B_i | T_i^k) \Pr(T_i^k | G_i^l) \right] \Pr(G_1^l, G_2^l, \dots, G_n^l) \right\}}$$

```

GTT*ACT*GTCGTTGT*AA*TACTCC*A*CGATGTCTTTCAGGG*TCTCC*ATAAAGAT
GTT*ACT*GTCGTTGT*AA*TACTCC*A*CGATGTCTTTCAGGG*TCTCC*ATAAAGAT
gtt*act*gTCgttgt*AA*TACTcC*a*cAATgtctttcaggg*tctcC*ATaaaGAT

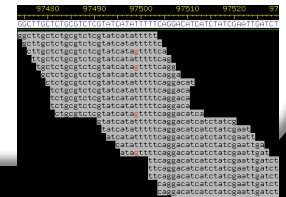
```

↑
sequencing error

↑
polymorphism



SNP calling in multi-sample read sets

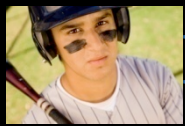


a
a
c
c

$$P(B_1 = aacc | G_1 = aa)$$

$$P(B_1 = aacc | G_1 = cc)$$

$$P(B_1 = aacc | G_1 = ac)$$



a
a
a
a
c

$$P(B_i = aaaac | G_i = aa)$$

$$P(B_i = aaaac | G_i = cc)$$

$$P(B_i = aaaac | G_i = ac)$$



c
c
c
c

$$P(B_n = cccc | G_n = aa)$$

$$P(B_n = cccc | G_n = cc)$$

$$P(B_n = cccc | G_n = ac)$$

Prior($G_1, \dots, G_i, \dots, G_n$)

$$P(G_1 = aa | B_1 = aacc; B_i = aaaac; B_n = cccc)$$

$$P(G_1 = cc | B_1 = aacc; B_i = aaaac; B_n = cccc)$$

$$P(G_1 = ac | B_1 = aacc; B_i = aaaac; B_n = cccc)$$

$$P(G_i = aa | B_1 = aacc; B_i = aaaac; B_n = cccc)$$

$$P(G_i = cc | B_1 = aacc; B_i = aaaac; B_n = cccc)$$

$$P(G_i = ac | B_1 = aacc; B_i = aaaac; B_n = cccc)$$

$$P(G_n = aa | B_1 = aacc; B_i = aaaac; B_n = cccc)$$

$$P(G_n = cc | B_1 = aacc; B_i = aaaac; B_n = cccc)$$

$$P(G_n = ac | B_1 = aacc; B_i = aaaac; B_n = cccc)$$

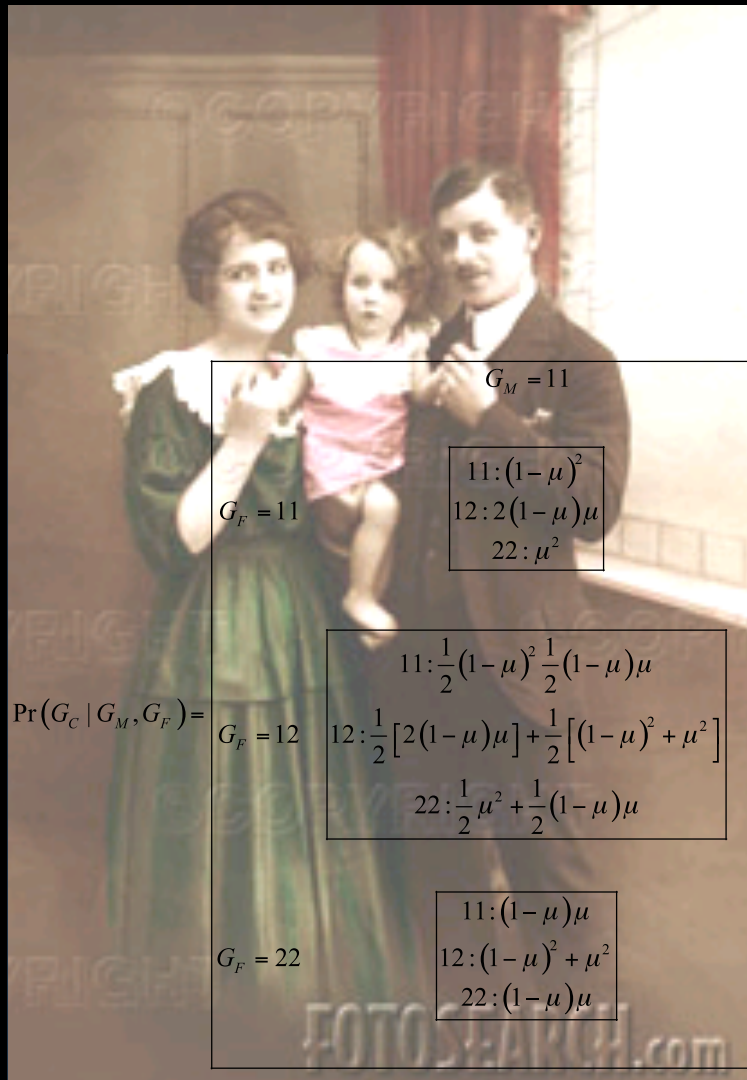
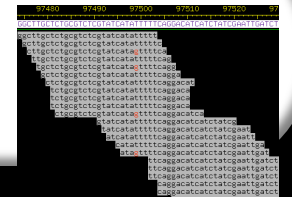
“genotype likelihoods”

“genotype probabilities”

↓
P(SNP)

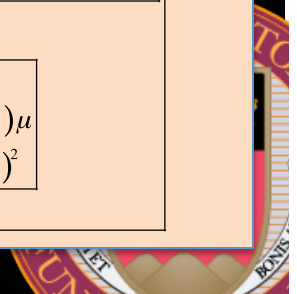


Trio sequencing

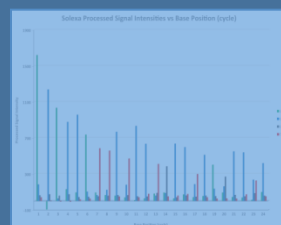
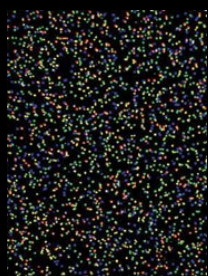
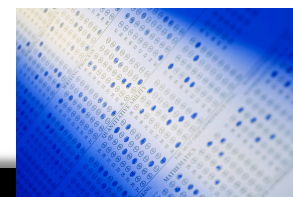


- the child inherits one chromosome from each parent
- there is a small probability for a *de novo* (germ-line or somatic) mutation in the child

	$G_M = 11$	$G_M = 12$	$G_M = 22$
$G_F = 11$	$\begin{matrix} 11: (1-\mu)^2 \\ 12: 2(1-\mu)\mu \\ 22: \mu^2 \end{matrix}$	$\begin{matrix} 11: \frac{1}{2}(1-\mu)^2 + \frac{1}{2}(1-\mu)\mu \\ 12: \frac{1}{2}[2(1-\mu)\mu] + \frac{1}{2}[(1-\mu)^2 + \mu^2] \\ 22: \frac{1}{2}\mu^2 + \frac{1}{2}(1-\mu)\mu \end{matrix}$	$\begin{matrix} 11: (1-\mu)\mu \\ 12: (1-\mu)^2 + \mu^2 \\ 22: (1-\mu)\mu \end{matrix}$
$G_F = 12$	$\begin{matrix} 11: \frac{1}{2}(1-\mu)^2 \frac{1}{2}(1-\mu)\mu \\ 12: \frac{1}{2}[2(1-\mu)\mu] + \frac{1}{2}[(1-\mu)^2 + \mu^2] \\ 22: \frac{1}{2}\mu^2 + \frac{1}{2}(1-\mu)\mu \end{matrix}$	$\begin{matrix} 11: \frac{1}{4}(1-\mu)^2 + \frac{1}{2}[(1-\mu)\mu] + \frac{1}{4}\mu^2 \\ 12: \frac{1}{4}[2(1-\mu)\mu] + \frac{1}{2}[(1-\mu)^2 + \mu^2] + \frac{1}{4}[2(1-\mu)\mu] \\ 22: \frac{1}{4}\mu^2 + \frac{1}{2}(1-\mu)\mu + \frac{1}{4}(1-\mu)^2 \end{matrix}$	$\begin{matrix} 11: \frac{1}{2}(1-\mu)\mu + \frac{1}{2}\mu^2 \\ 12: \frac{1}{2}[(1-\mu)^2 + \mu^2] + \frac{1}{2}[2(1-\mu)\mu] \\ 22: \frac{1}{2}(1-\mu)\mu + \frac{1}{2}(1-\mu)^2 \end{matrix}$
$G_F = 22$	$\begin{matrix} 11: (1-\mu)\mu \\ 12: (1-\mu)^2 + \mu^2 \\ 22: (1-\mu)\mu \end{matrix}$	$\begin{matrix} 11: \frac{1}{2}(1-\mu)\mu + \frac{1}{2}\mu^2 \\ 12: \frac{1}{2}[(1-\mu)^2 + \mu^2] + \frac{1}{2}[2(1-\mu)\mu] \\ 22: \frac{1}{2}(1-\mu)\mu + \frac{1}{2}(1-\mu)^2 \end{matrix}$	$\begin{matrix} 11: \mu^2 \\ 12: 2(1-\mu)\mu \\ 22: (1-\mu)^2 \end{matrix}$



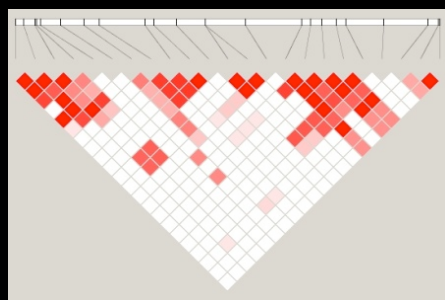
Standard data formats



```
>B_TITR_1_1_668_35 TIME: Tue Feb 20 02:26:06 2007
ATATCGGATGACACAATATGGGAGGTGAC
>B_TITR_1_2_843_403 TIME: Tue Feb 20 02:26:06 2007
TGTAGCTTTTCATGACAATTTTATAGGTGT
```

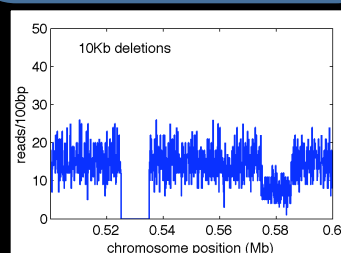
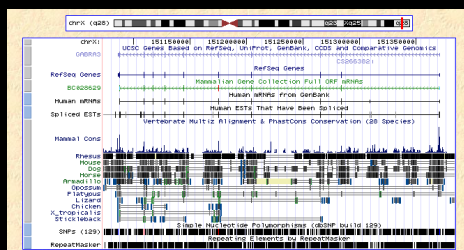
SRF/FASTQ

```
B_TITR_1_1_668_35
27 27 27 27 27 27 27 27 27 27 27 27 27 27 27 27 27 27 27 27 27 27 27 27 27 27
>B_TITR_1_2_843_403
26 26 26 26 26 26 26 26 26 26 26 26 26 26 26 26 26 26 26 26 26 26 26 26 26 26
>B_TITR_1_3_618_922
```



GLF/VCF

SAM/BAM



The 1000 Genomes Project

1000 Genomes

A Deep Catalog of Human Genetic Variation

Pilot 1

1. **To evaluate the use of low-redundancy genome sequencing to characterize single nucleotide and copy number variants, discovering all variants with frequency > 5% in the original HapMap samples.**

This pilot will evaluate the utility of low-redundancy genomic sequence from many individuals, using the new sequencing technologies, including paired-end reads, for discovering SNP and structural variants and inferring haplotypes. These data will guide evaluation and development of methods for imputation from incomplete sequence data. In total 180 samples (60 unrelated samples from each of the HapMap CEU, YRI, and CHB+JPT populations) would be sequenced to a coverage depth of 2X of high quality mapped bases (1080 Gb total), and the resulting data analyzed to discover SNP and copy number variants.

Pilot 2

2. **To evaluate the effect of coverage depth on project goals, based on deep sequencing of two sets of trio samples.**

This pilot will evaluate the relationship between coverage depth and the yield of variation data, based on genomic sequence from a few individuals. A high level of redundancy will provide a solid basis for assessing the coverage needed for discovering variants, inferring haplotypes, imputing non-typed variants, and using paired-end reads for finding structural variants. Two trios (6 samples), one from each of the HapMap YRI and CEU panels, would be sequenced to a coverage depth of 20X of high quality mapped bases (360 Gb total).

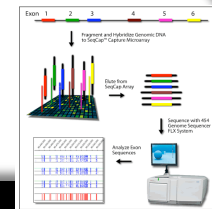
Pilot 3

3. **To develop and evaluate technologies to perform targeted sequencing of exons and other functional elements at genome-wide scale, and pilot deep sequencing in more than 1,000 DNA samples.**

This pilot will develop and evaluate technologies to capture specific genomic regions and discover variants. It will provide data on the frequency distribution of rare variants, and in combination with other data enable the study of haplotype patterns around rare alleles. It will thus guide development of algorithms to impute less common alleles from SNP data. In total, 1000-2000 gene regions and conserved elements would be sequenced at 20X of high quality mapped bases in 1085-1536 samples (109-307 Gb).



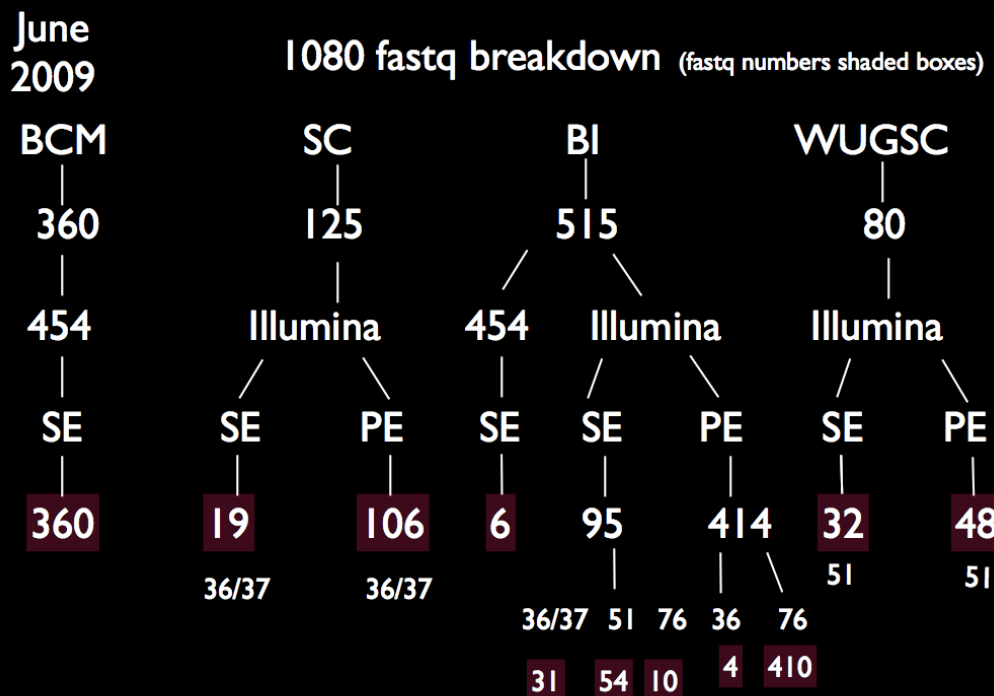
1000G Pilot 3 – exon sequencing



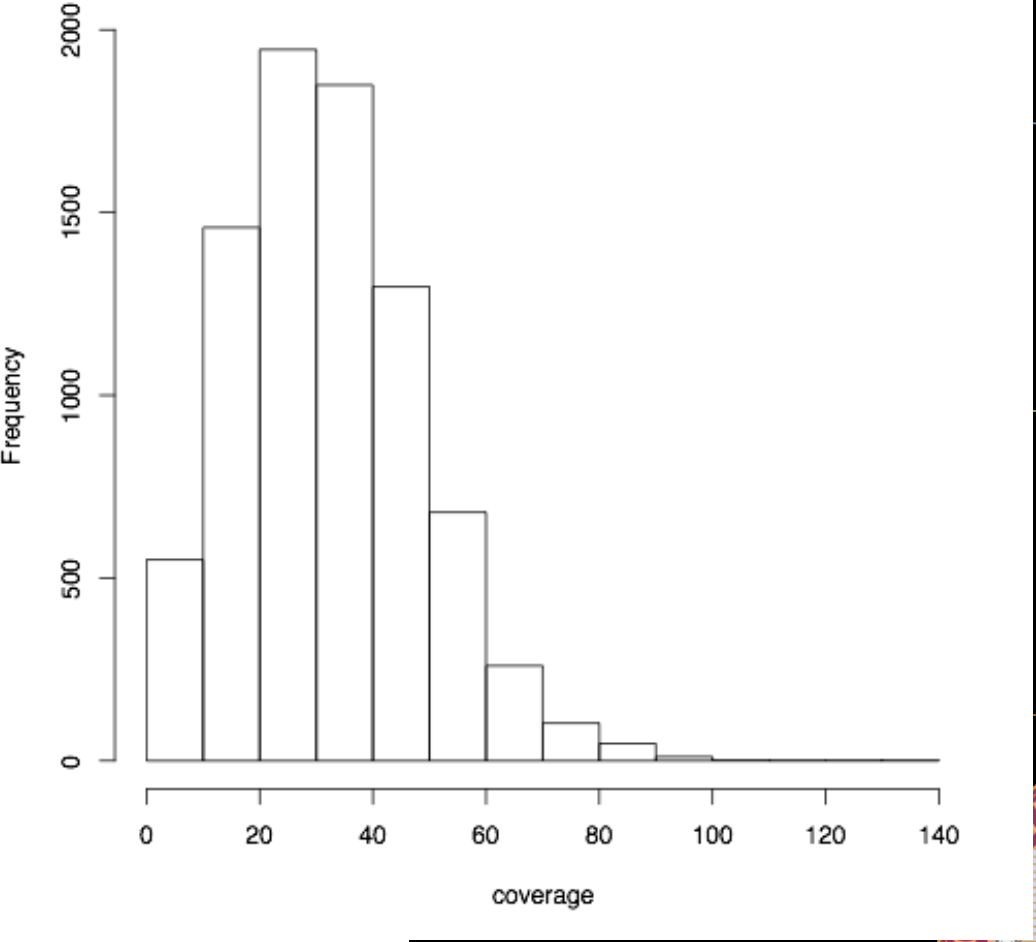
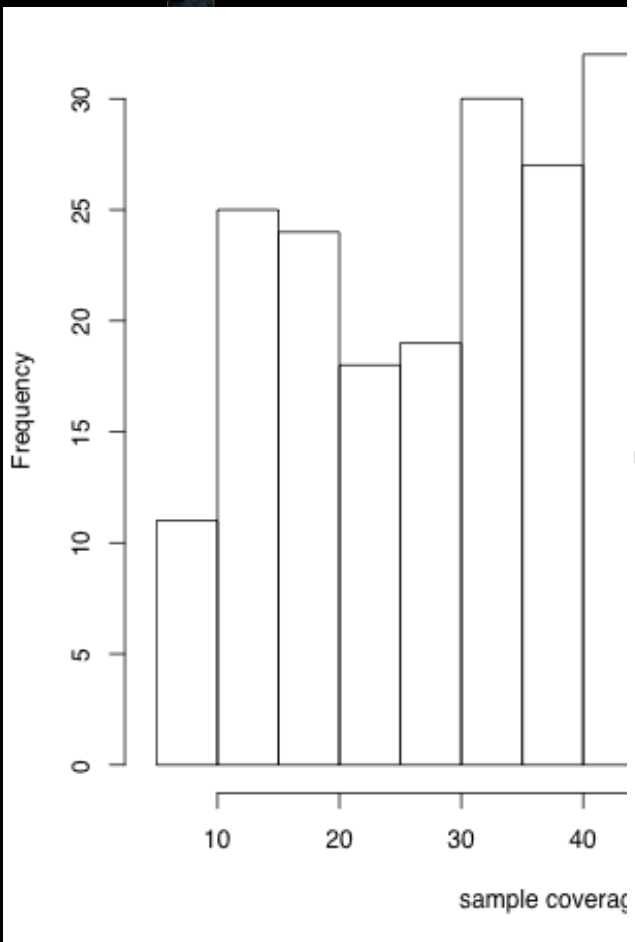
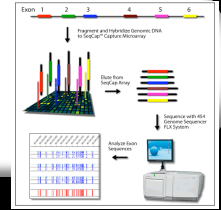
1000 Genomes

A Deep Catalog of Human Genetic Variation

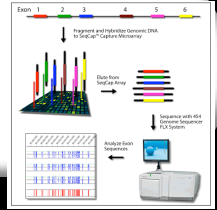
- Targets:
 - 1K genes / 10K targets
- Capture:
 - Solid / liquid phase
- Sequencing:
 - 454 / Illumina
 - SE / PE
- Data producers:
 - Baylor
 - Broad
 - Sanger
 - Wash. U.
- Informatics methods:
 - Multiple read mapping & SNP calling programs



Coverage varies



On/off target capture



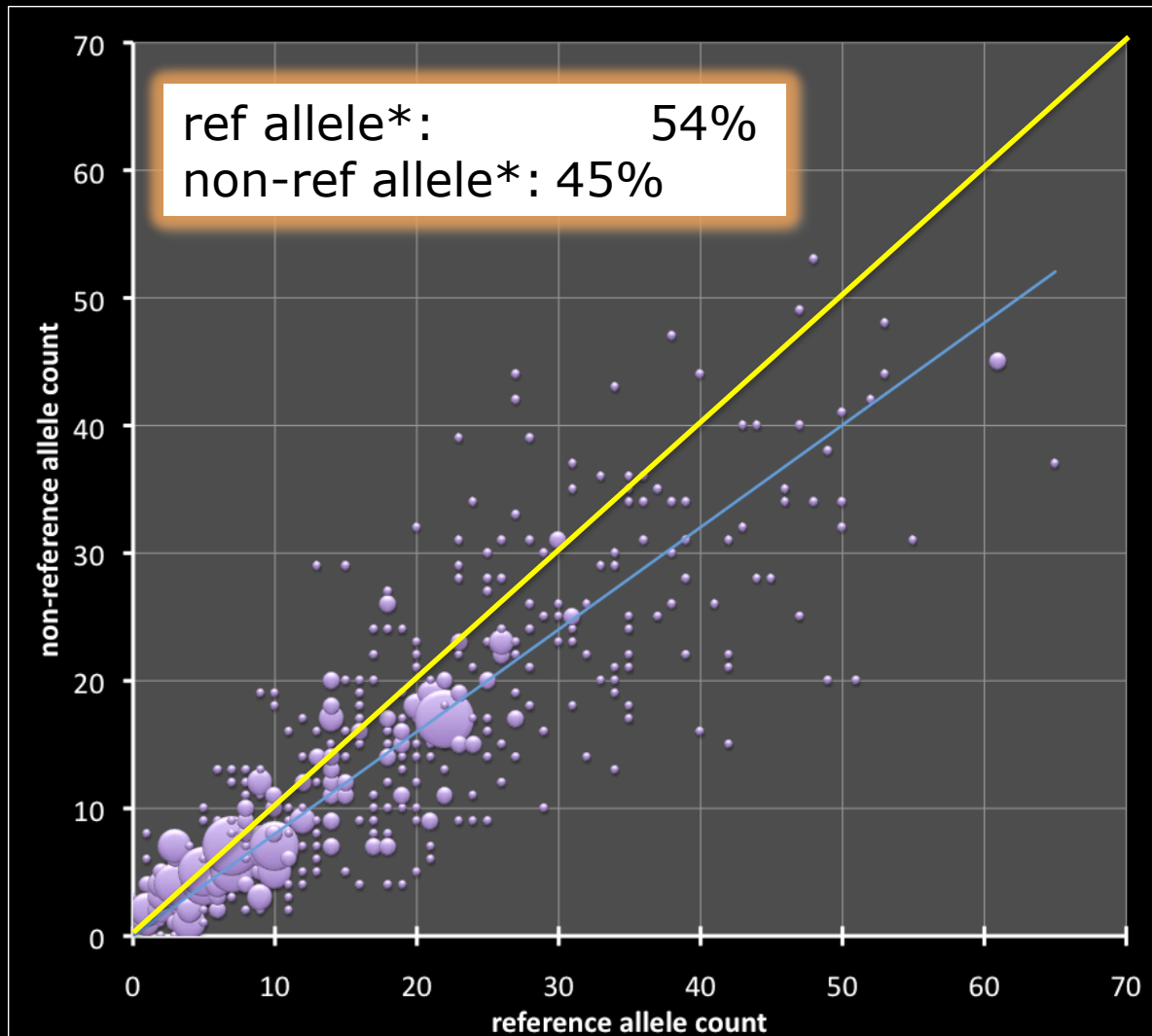
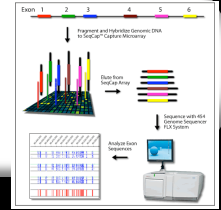
Target region



SNP
(outside target region)



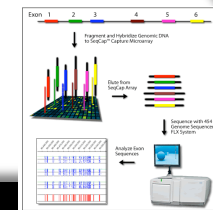
Reference allele bias



(*) measured at 450 het HapMap 3 sites overlapping capture target regions in sample NA07346

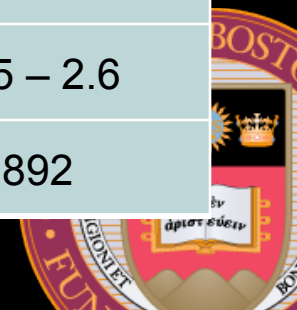


SNP calling findings



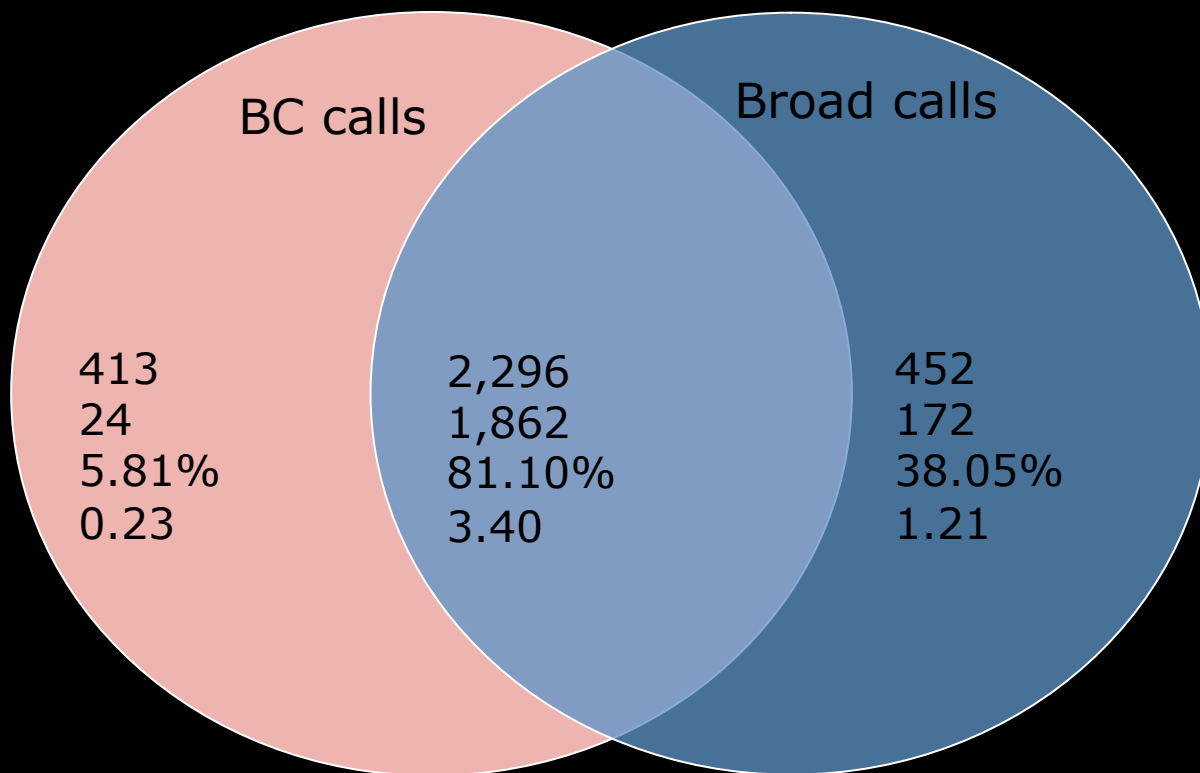
- based on a method comparison / testing exercise
- 80 samples drawn from the 4 Centers
- read mapping / SNP calling by the Baylor pipeline (BCM/454 data); the Broad and the BC pipelines (all 80 samples)

	BCM/454	BI/SLX	WUGSC/SLX	SC/SLX
# Samples	32	23	16	11
<read depth> per sample	35 X	62 X	117 X	51 X
# SNPs called	7,200 – 8,400	4,500 – 4,700	3,700	3,500 – 3,700
% dbSNPs	39 - 55	65 - 72	68	75 - 85
Ts/Tv(#SNP)	1.7 – 2.6	1.9 – 2.3	2.3	2.5 – 2.6
# Novel SNPs	3,998	1,550	1,947	892



Overlap between call sets

SNP calls:
dbSNPs:
% dbSNPs:
Ts/Tv ratio:



The 1000G Structural Variation Discovery Effort

1000 Genomes

A Deep Catalog of Human Genetic Variation

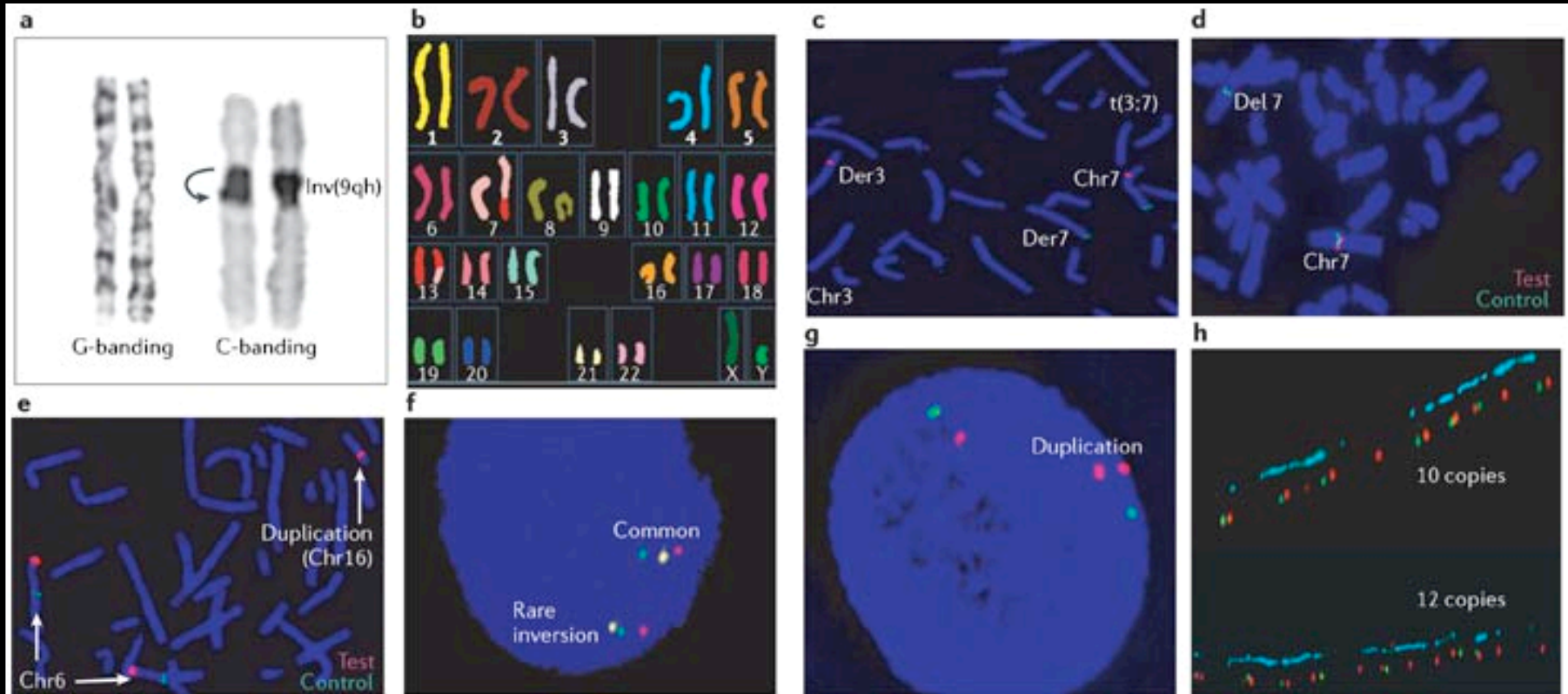
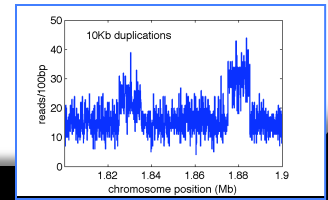
Primary goals:

1. Discover variants (SNPs, copy-number variants, insertions, deletions, other structural variants).

As a genomic project the resource should provide completeness; **the resource should include almost all accessible variants with allele frequencies as low as 1% across the genome and 0.1-0.5% in gene regions.** Currently the common SNPs are mostly known; the additional sequencing will be especially valuable for the discovery and characterization of many more rare variants and structural variants.



Structural variation detection

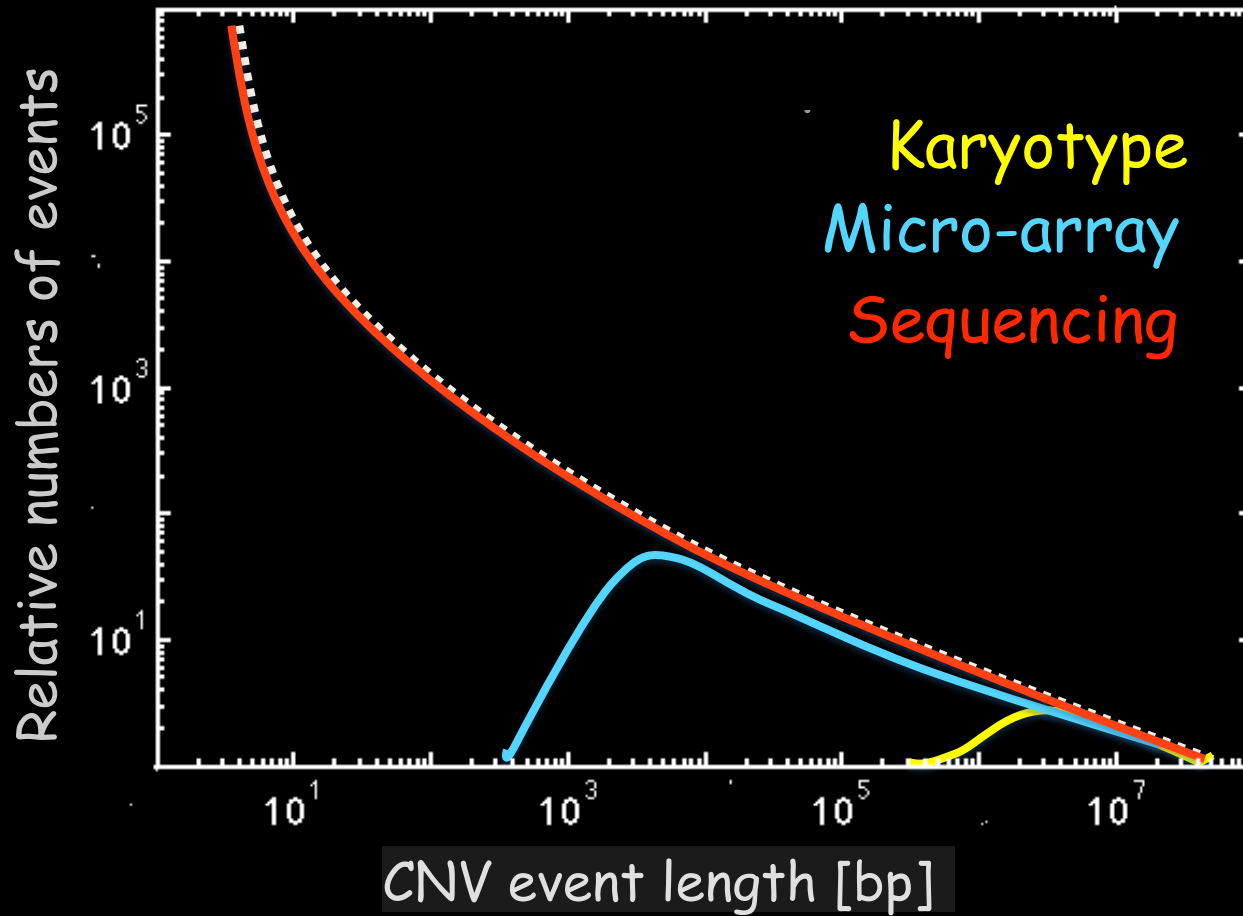
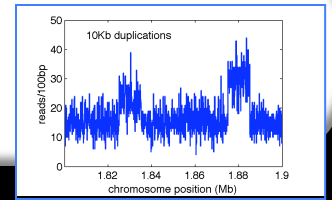


Copyright © 2006 Nature Publishing Group
Nature Reviews | Genetics

Feuk *et al.* Nature Reviews Genetics, 2006

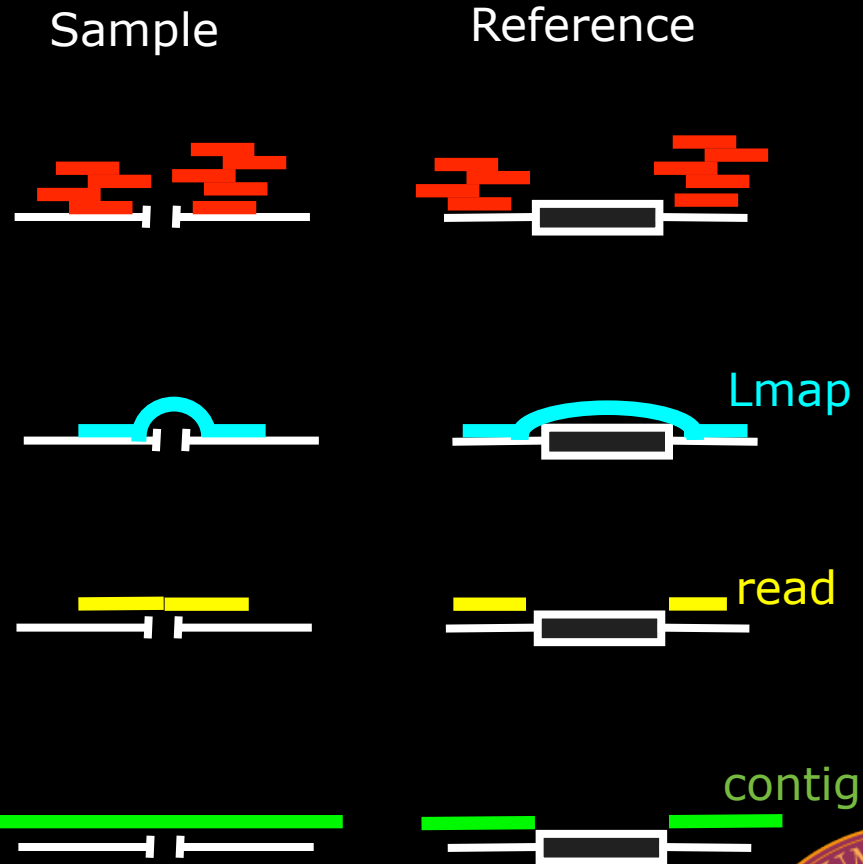


SV detection – resolution



Detection Approaches

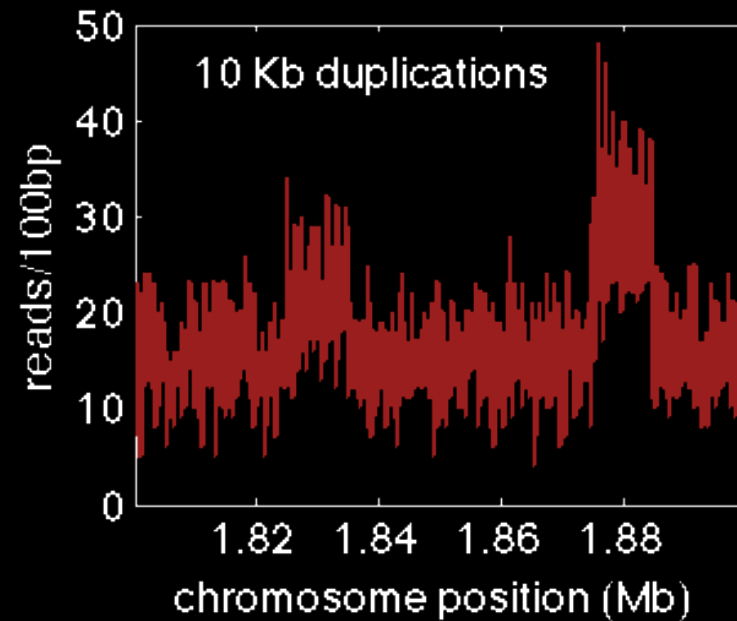
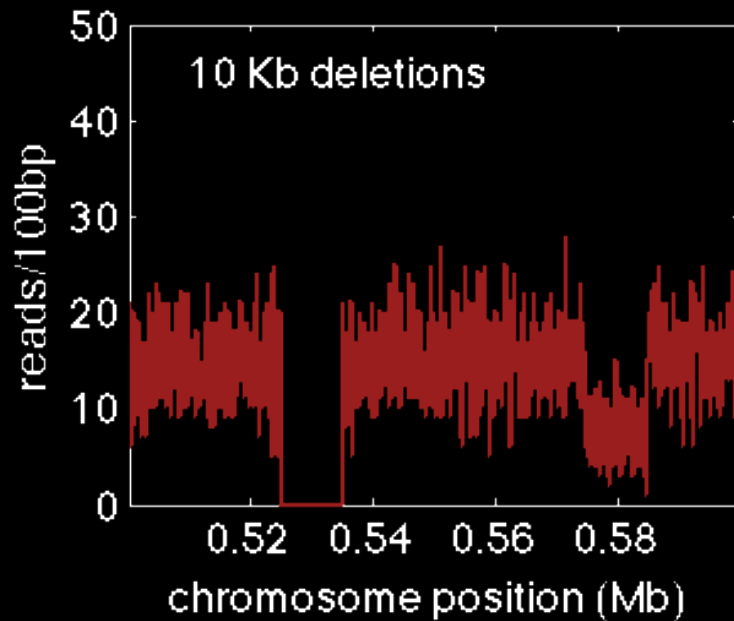
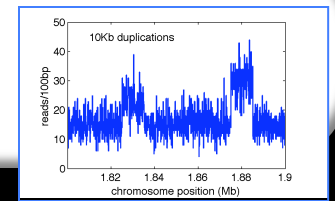
- Read Depth:
good for big CNVs
- Paired-end:
all types of SV
- Split-Reads
good break-point
resolution
- deNovo Assembly
~ the future



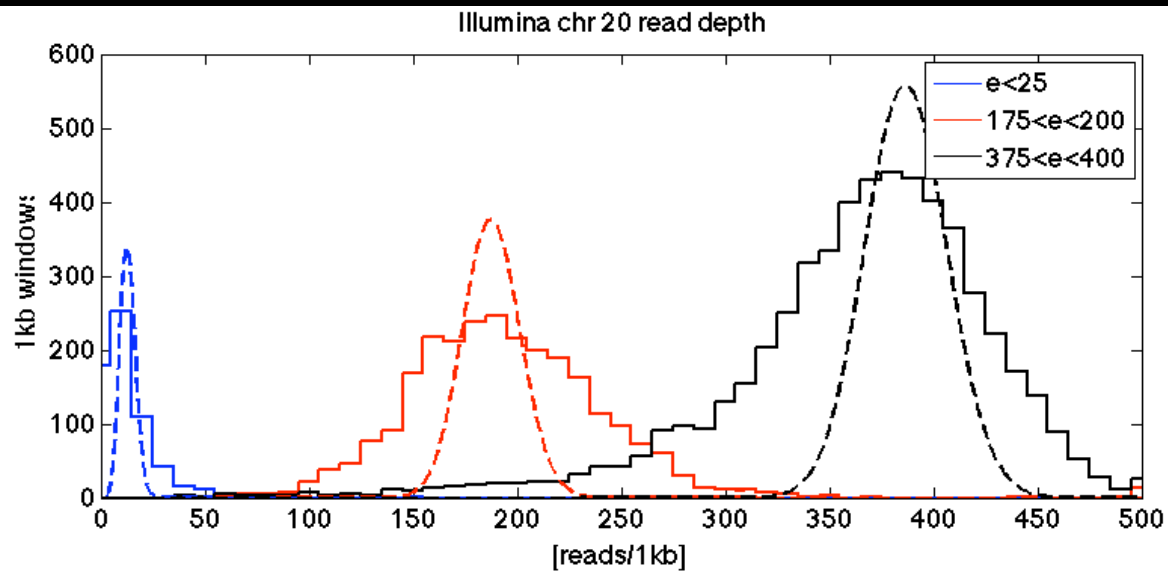
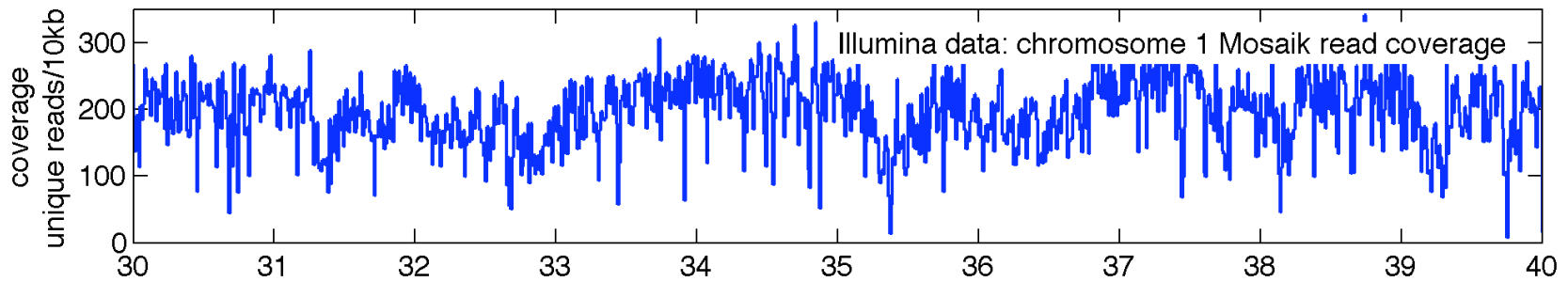
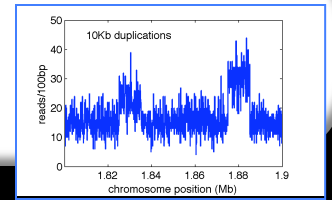
SV slides courtesy of Chip Stewart, Boston College



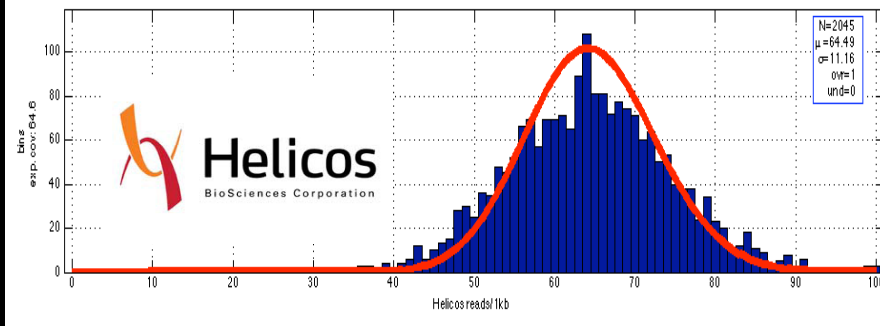
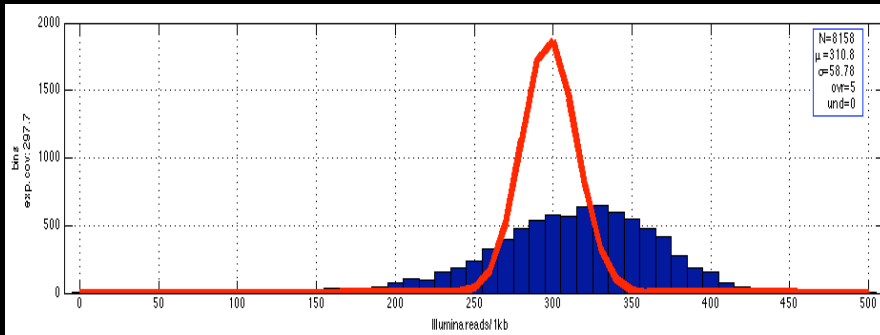
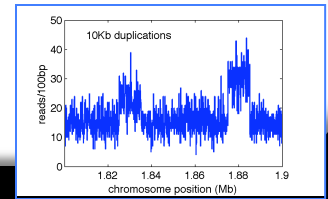
Read depth (RD)



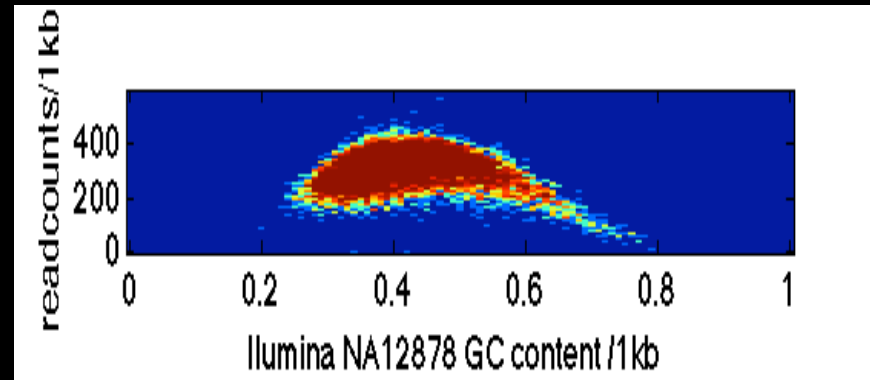
Statistical & systematic biases



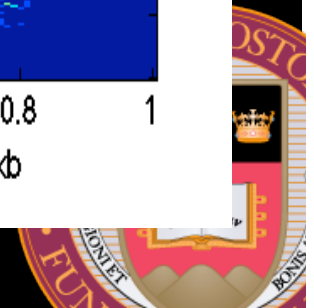
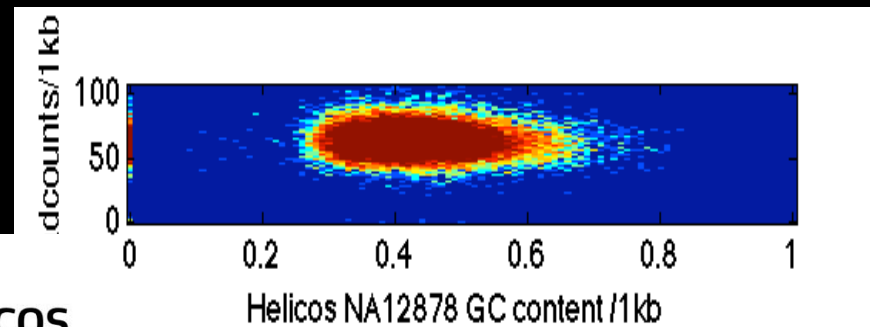
Single molecule sequencing?



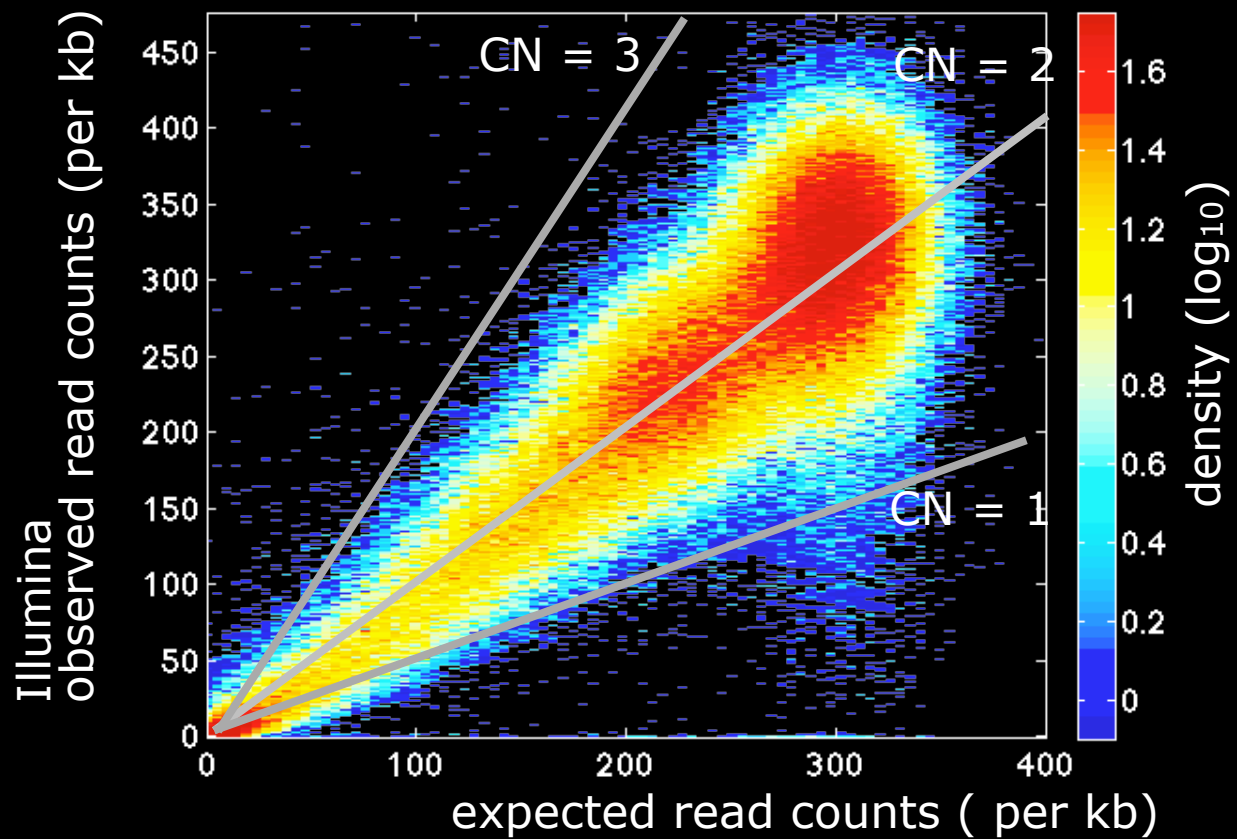
GC Bias



Coverage bias

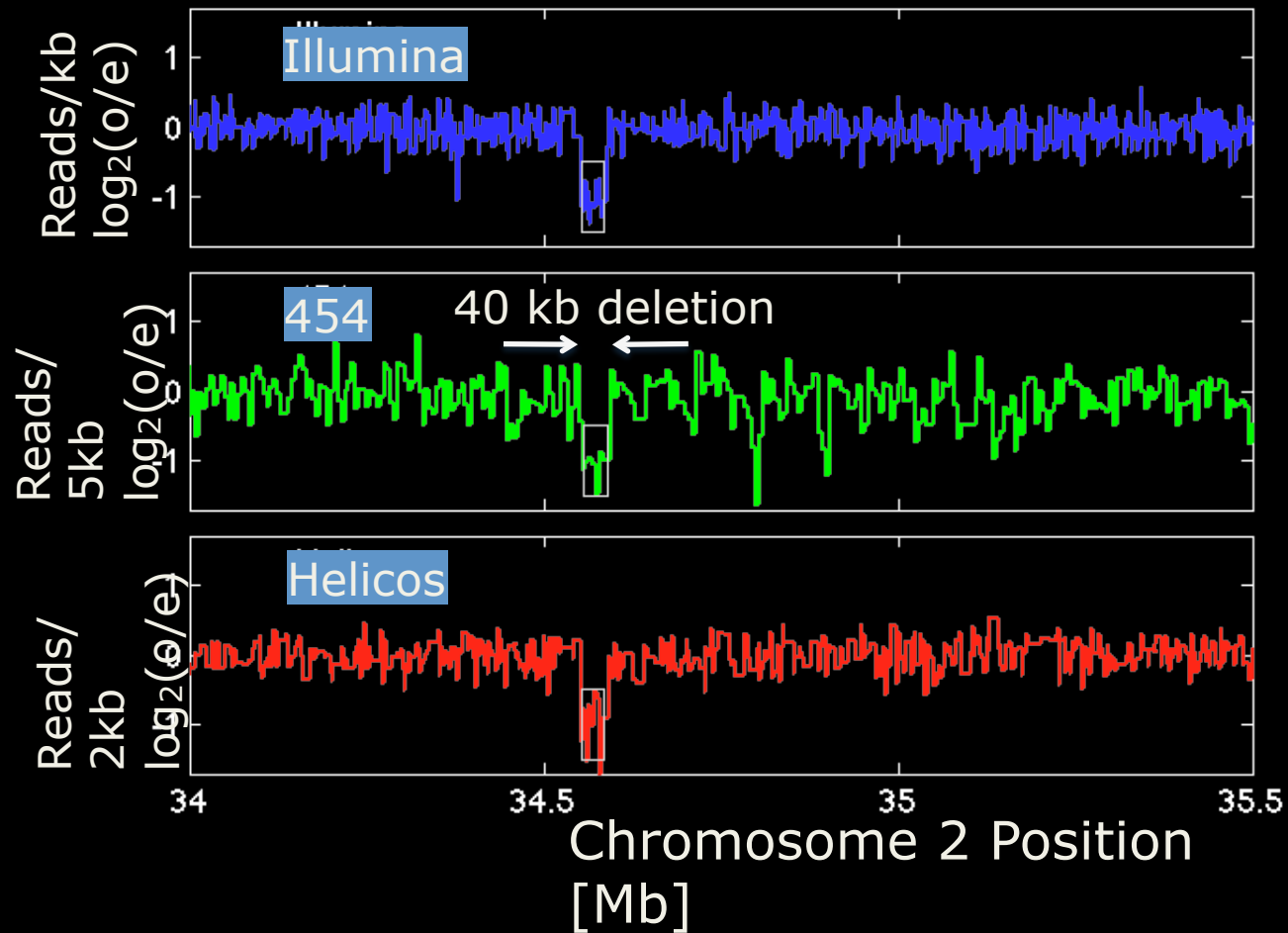


RD resolution

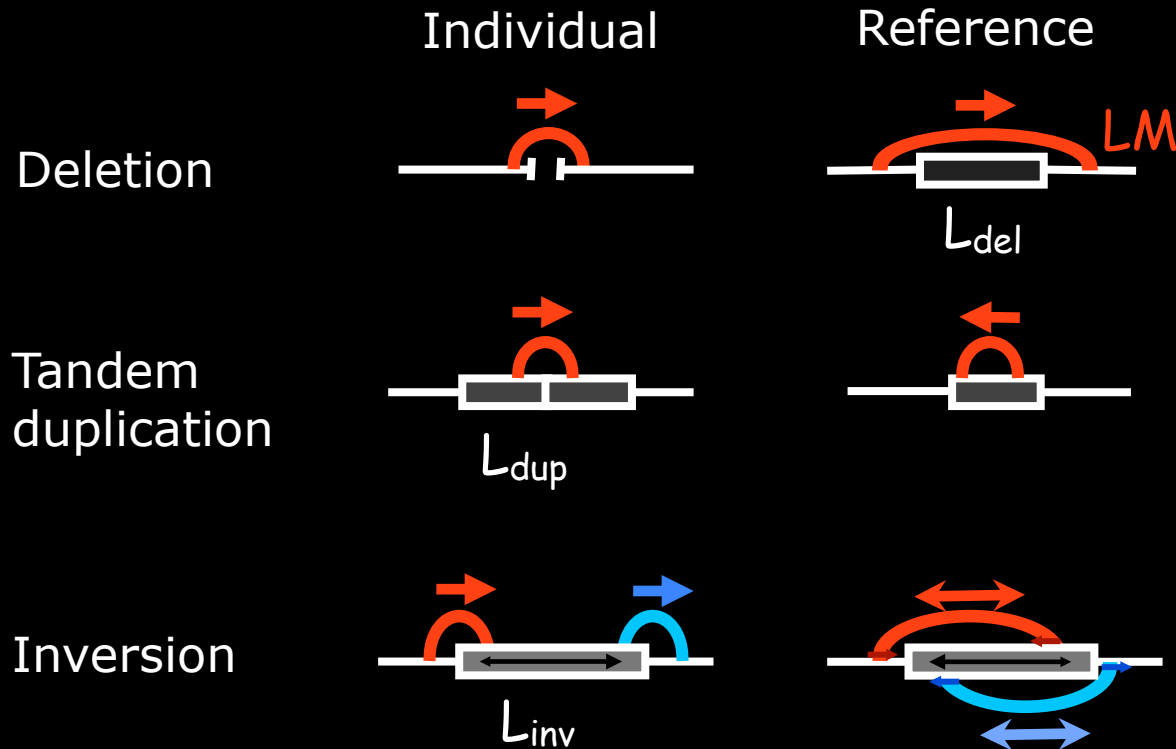


CNV events detected with RD

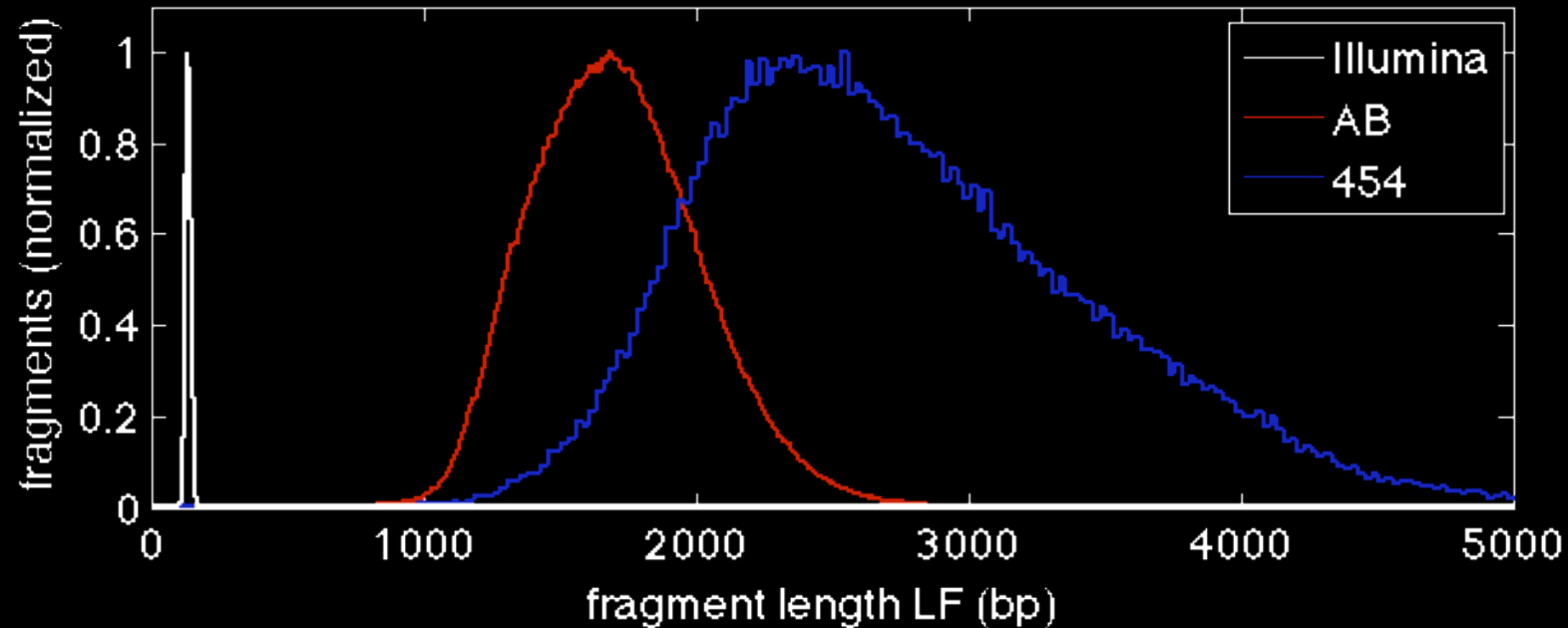
individual "NA12878"



SV detection with PE read map positions



Fragment length distributions



- long fragments \sim better fragment coverage and sensitivity to large events (454)
- tighter distributions \sim better breakpoint resolution and sensitivity for shorter events (Illumina)



The SV/CNV event display

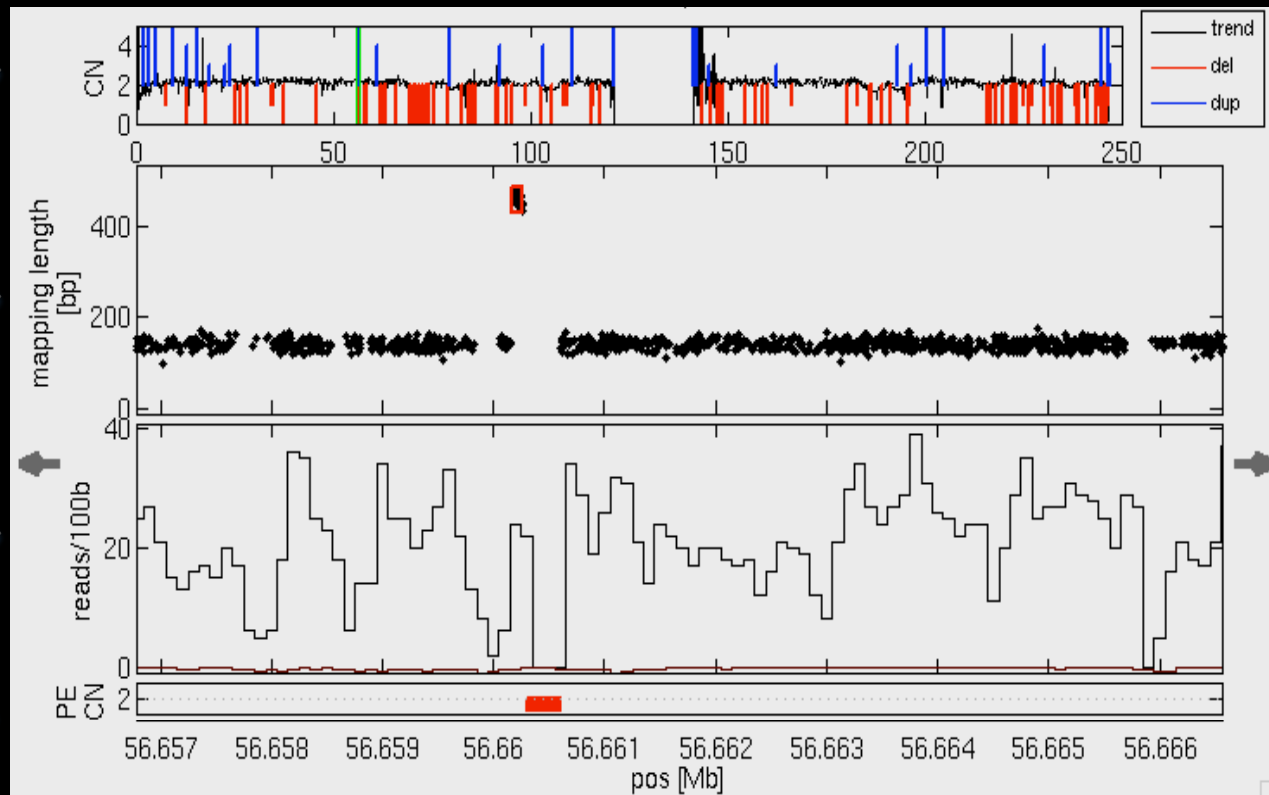
300 bp deletion in chromosome of NA12878 by Illumina paired-end data from the 1000 Genomes project

chromosome
overview

fragment
lengths

read
depth

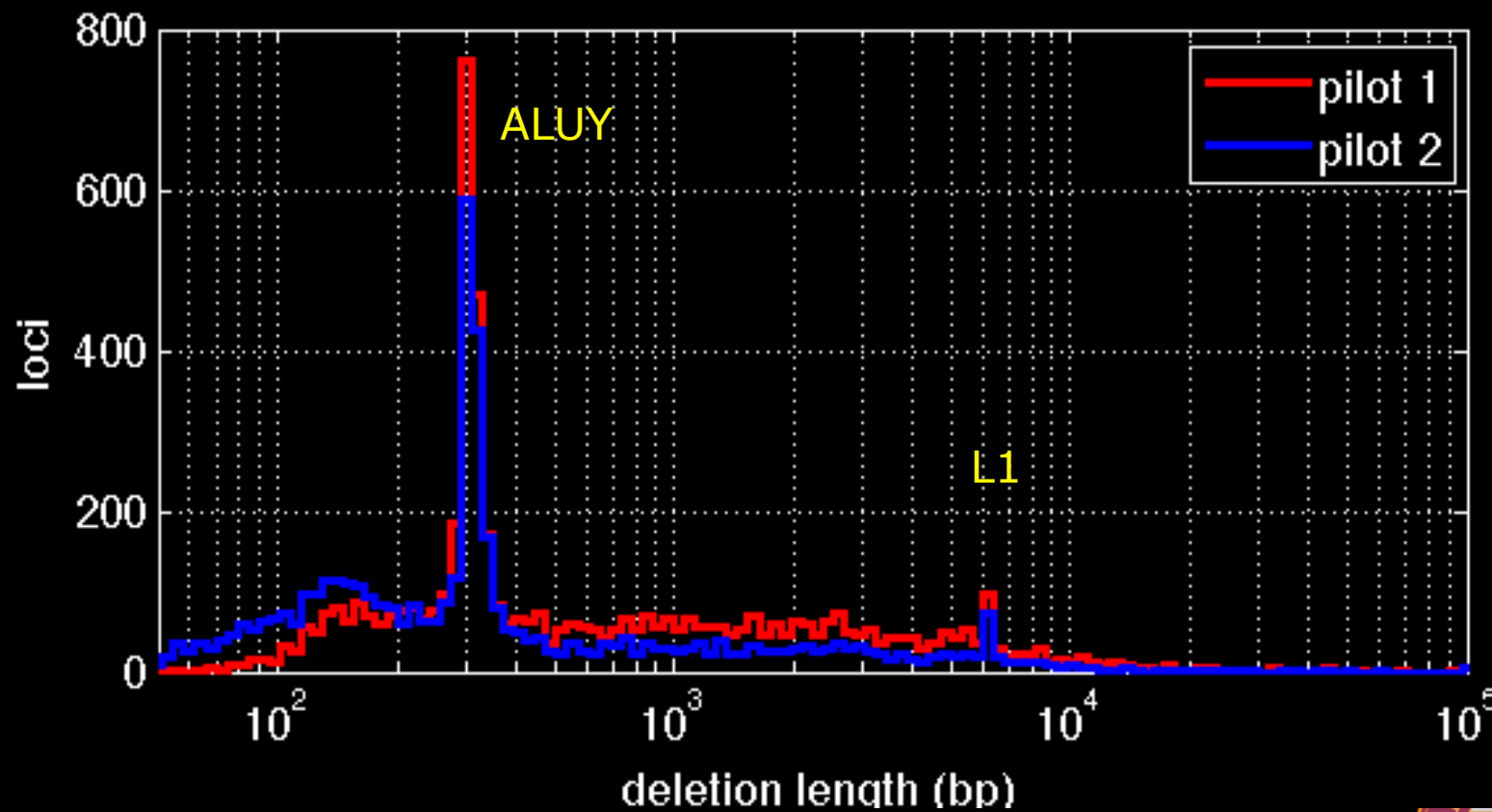
event
track



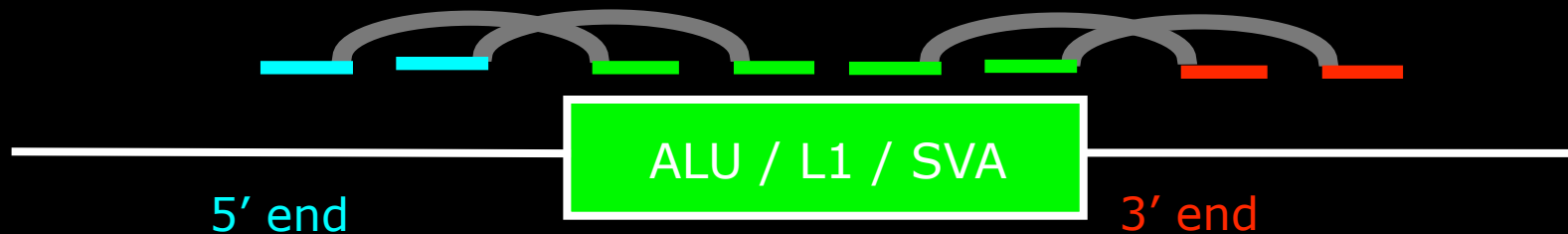
Chip Stewart



Deletion event lengths



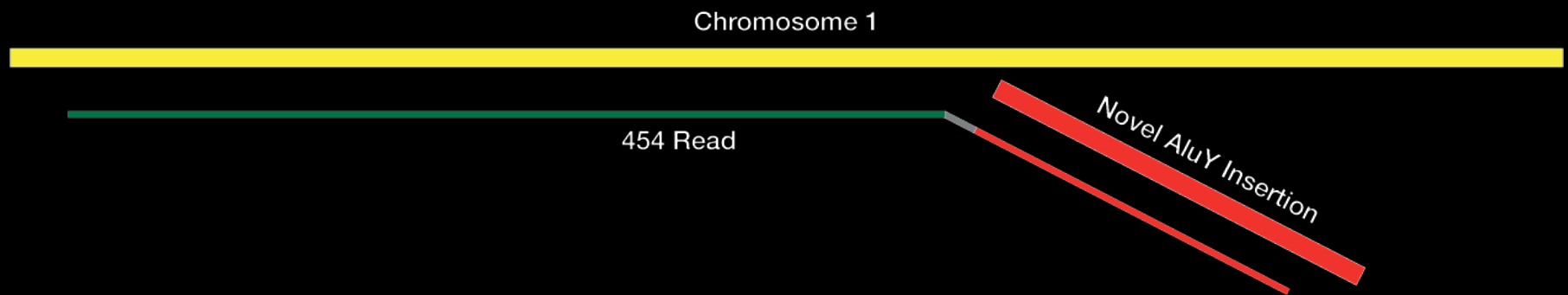
Mobile element insertions: **PE reads**



- Used with short-read data (Illumina, in our case)
- Detect clusters of 5' & 3' read pairs with one end mapping to a mobile element
- Clusters far from annotated elements are candidate insertion events



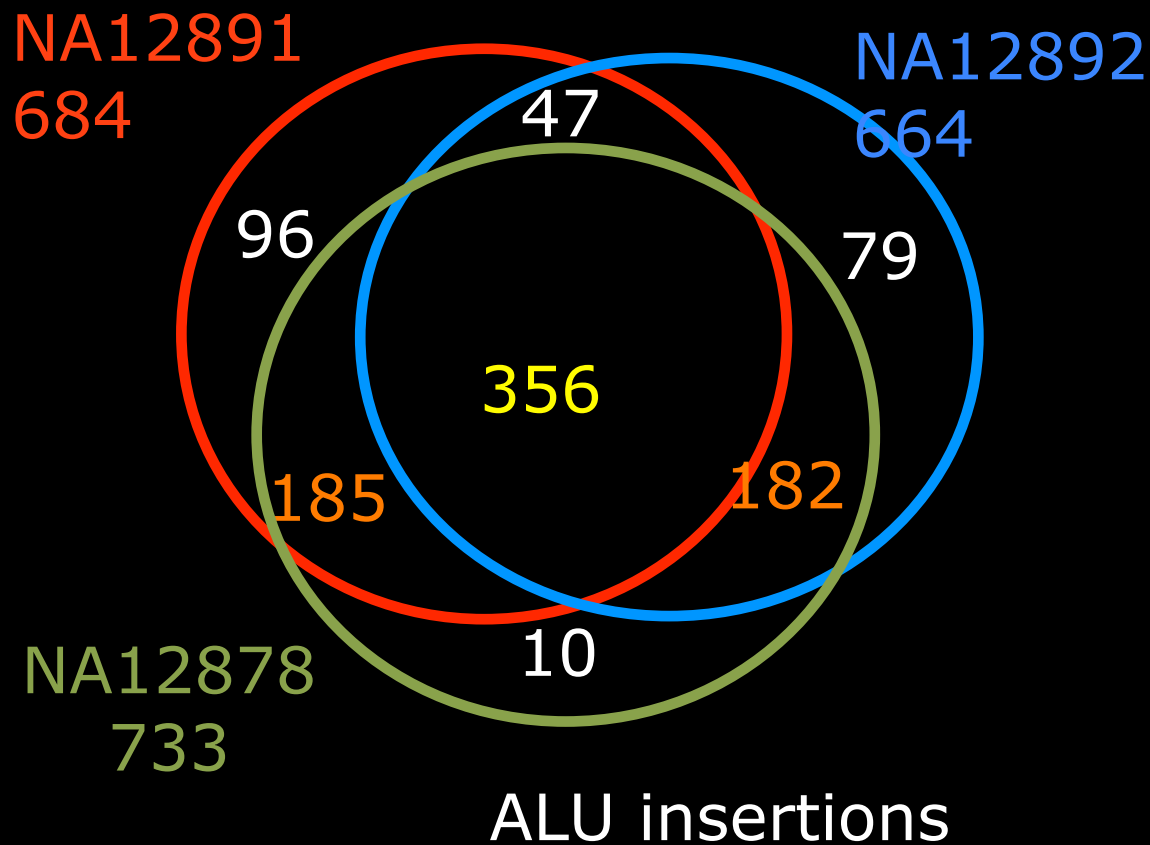
Mobile element insertions: **Split reads**



- Requires longer reads (454)
- Reads “mapping into” mobile element not present in the reference genome sequence are candidate insertion events



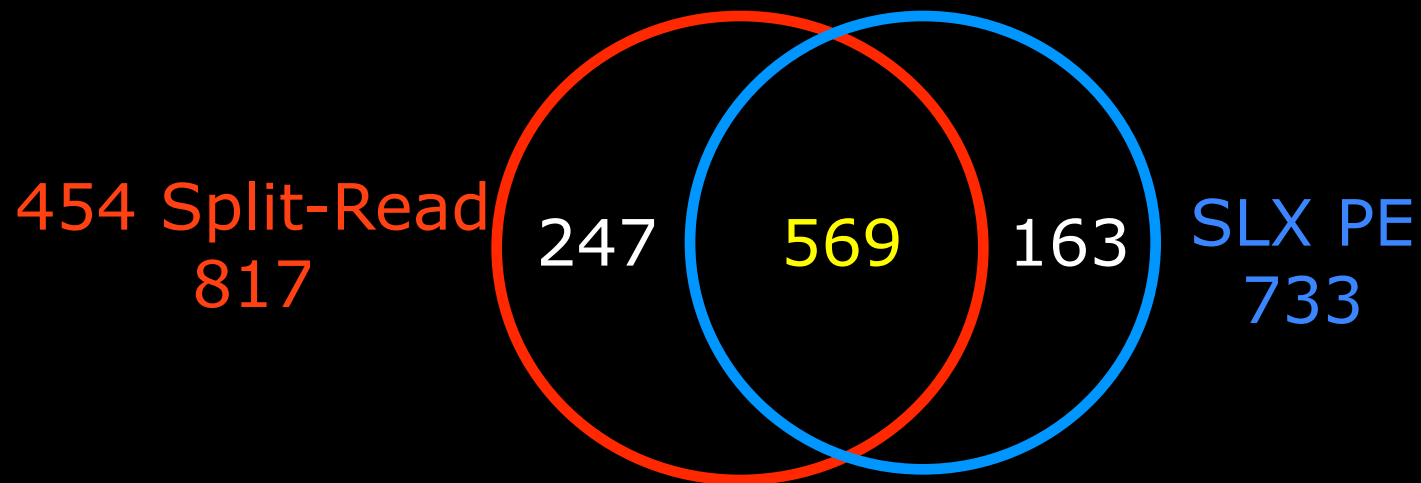
Mobile element insertions: **trio members**



Detection in 1000 G Pilot 2 **CEU trio PE Illumina data**



Mobile element insertions: **PE vs. Split-reads**



ALU insertions in NA12878

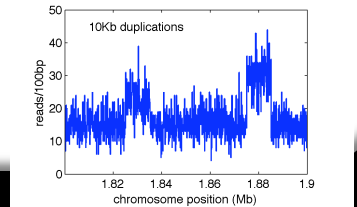


BC event lists in 1000 Genomes data

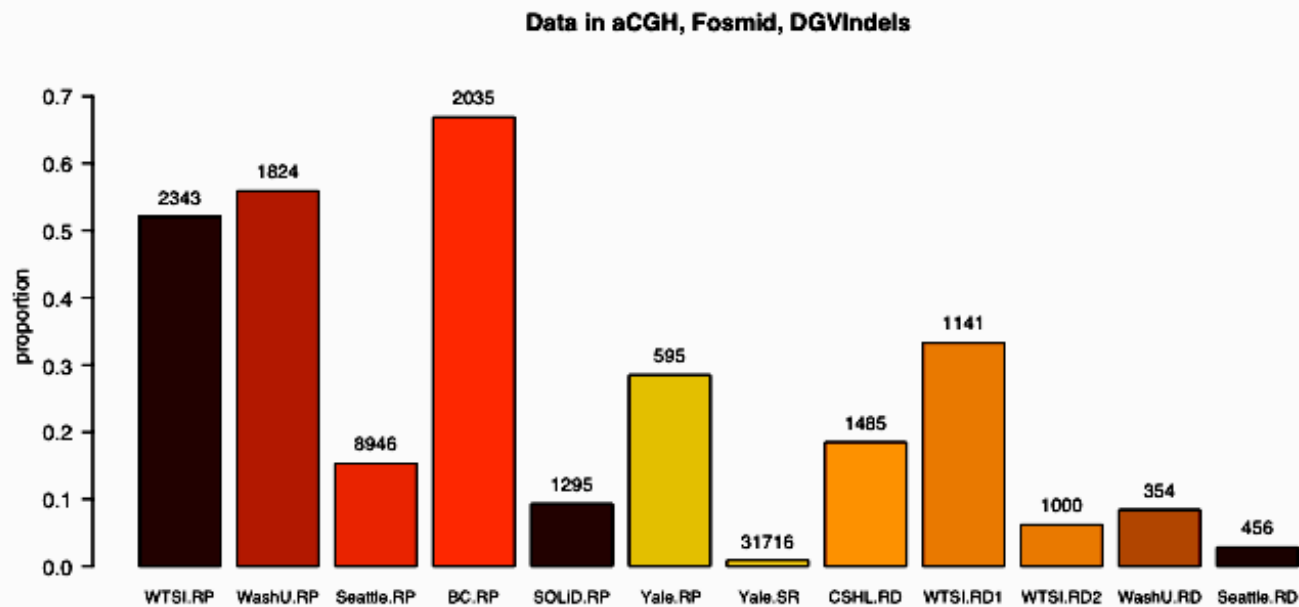
SV type	Pilot 1 140 samples low coverage	Pilot 2 6 samples high coverage
deletions	5,555	4,718
tandem duplications	540	406
mobile element insertions	3,276	2,013



SV calls / validation in 1000G datasets



Deletions validated in either aCGH, Fosmids or DGVIndels (overlap $\geq 50\%$)

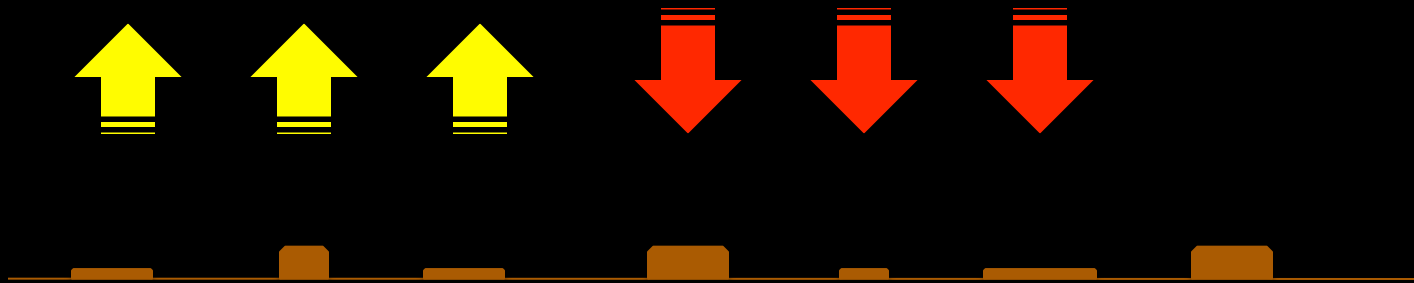
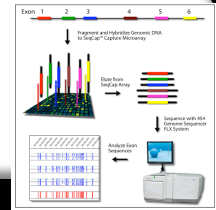


Klaudia Walter, Matt Hurles

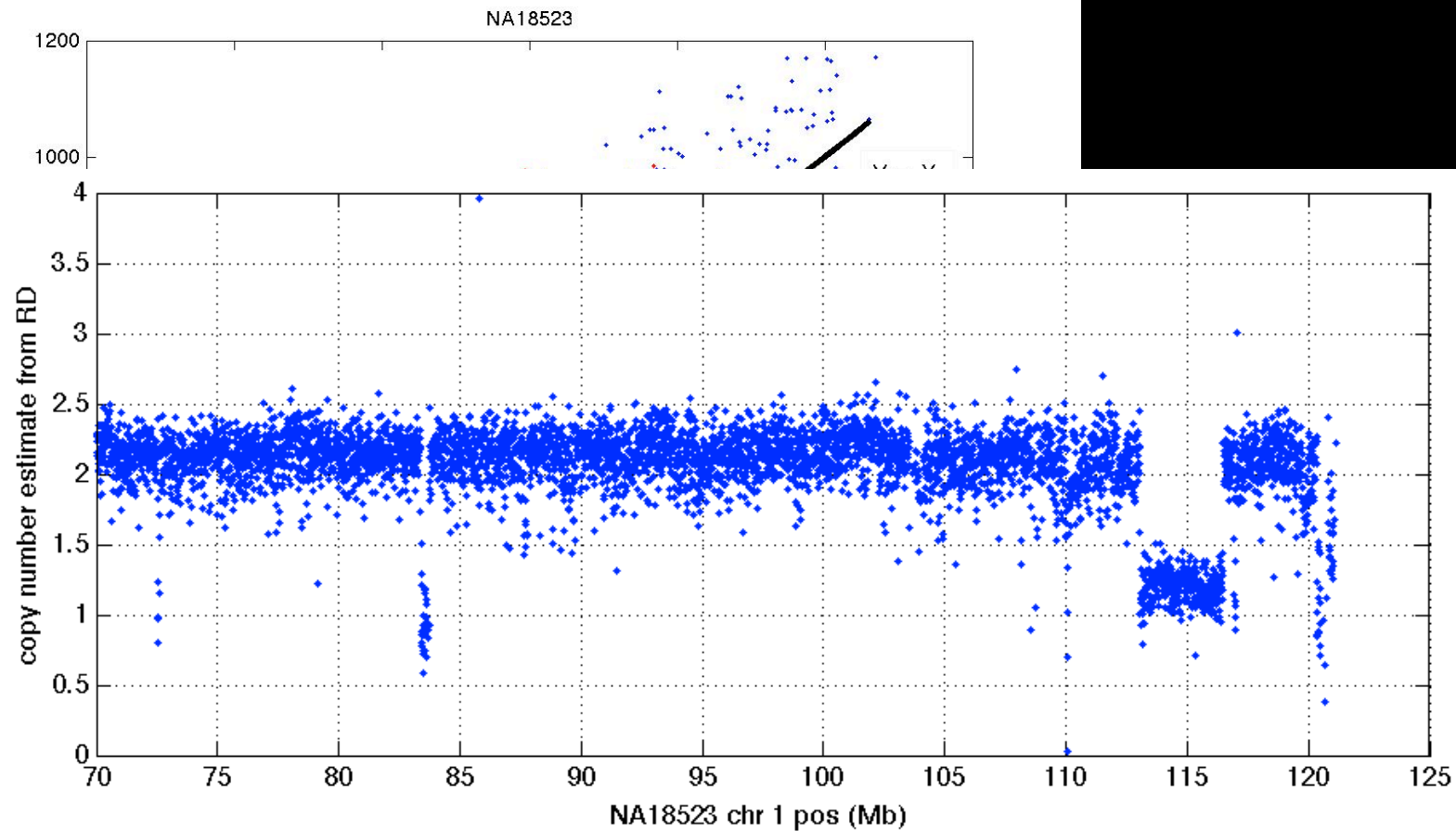
Overlap of NA12878 Deletions with aCGH, Fosmid data and DGVIndels



Can we find exon CNVs in Pilot 3 data?



SVs in exon sequencing data



Software access

THE MARTHLAB : SOFTWARE RELEASE



Welcome

This is the site for the beta release of our suite of analysis tools for next-generation sequencing machines. If you are a beta tester, you should have received the appropriate credentials to download the software, example data sets and relevant documentation. We respectfully request that you do not distribute any of the software, data or documentation to other parties.

Access

You will be able to access our beta software and serve as a beta tester by clicking on the following link: [obtain download instructions and credentials](#). This form will request that you fill out your contact information. After this an automatically generated email will be sent to your email address with download instructions and credential information.

Software components

The following software is included in the "downloadable" packages below.

- PyroBayes: Base caller for 454 pyrosequencing reads
- MOSAIK: Reference Guided Read Aligner / Assembler
- GigaBayes: Short-read polymorphism detection software

http://bioinformatics.bc.edu/marthlab/Software_Release



Credits

Elaine Mardis



Andy Clark



Aravinda Chakravarti



Michael Egholm



Scott Kahn



Francisco de la Vega



Patrice Milos

John Thompson



R01 HG004719

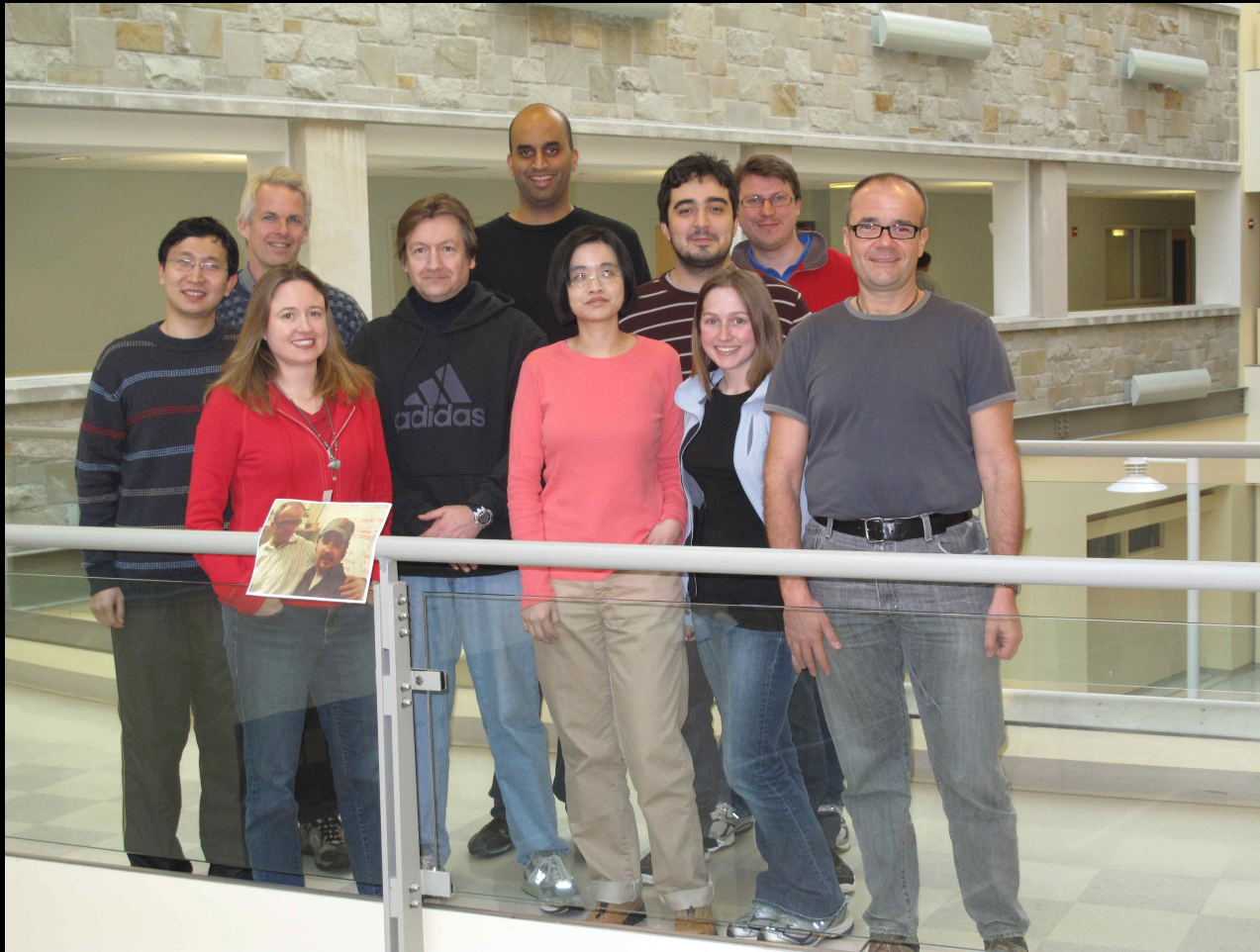
R01 HG003698

R21 AI081220

RC2 HG5552



Lab



Several positions are available:
grad students / postdocs / programmers

