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

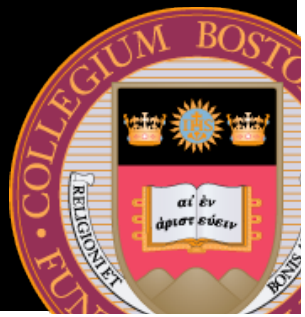
Capillary (e.g. AB 3730)

Roche 454

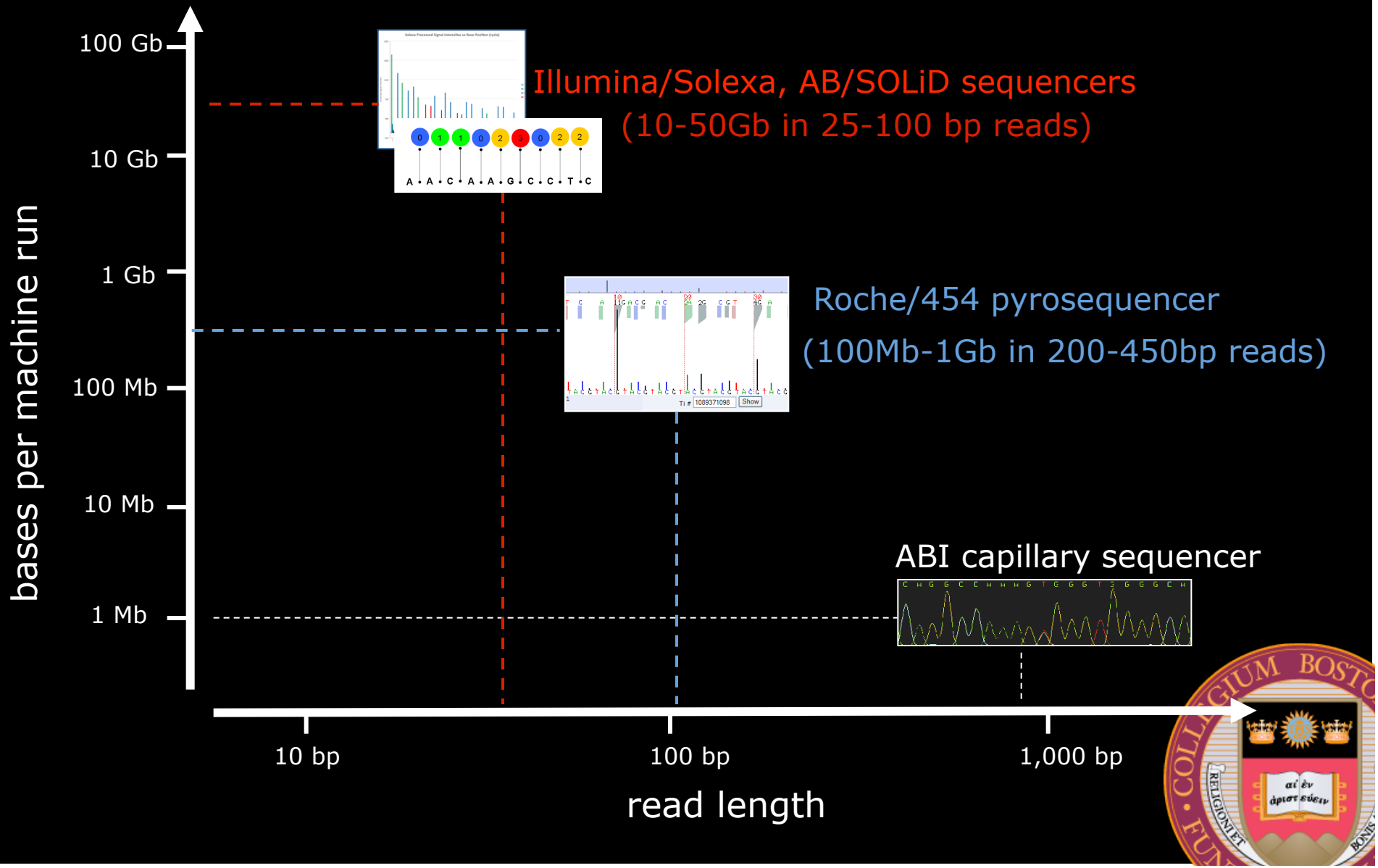
AB SOLiD

Helicos

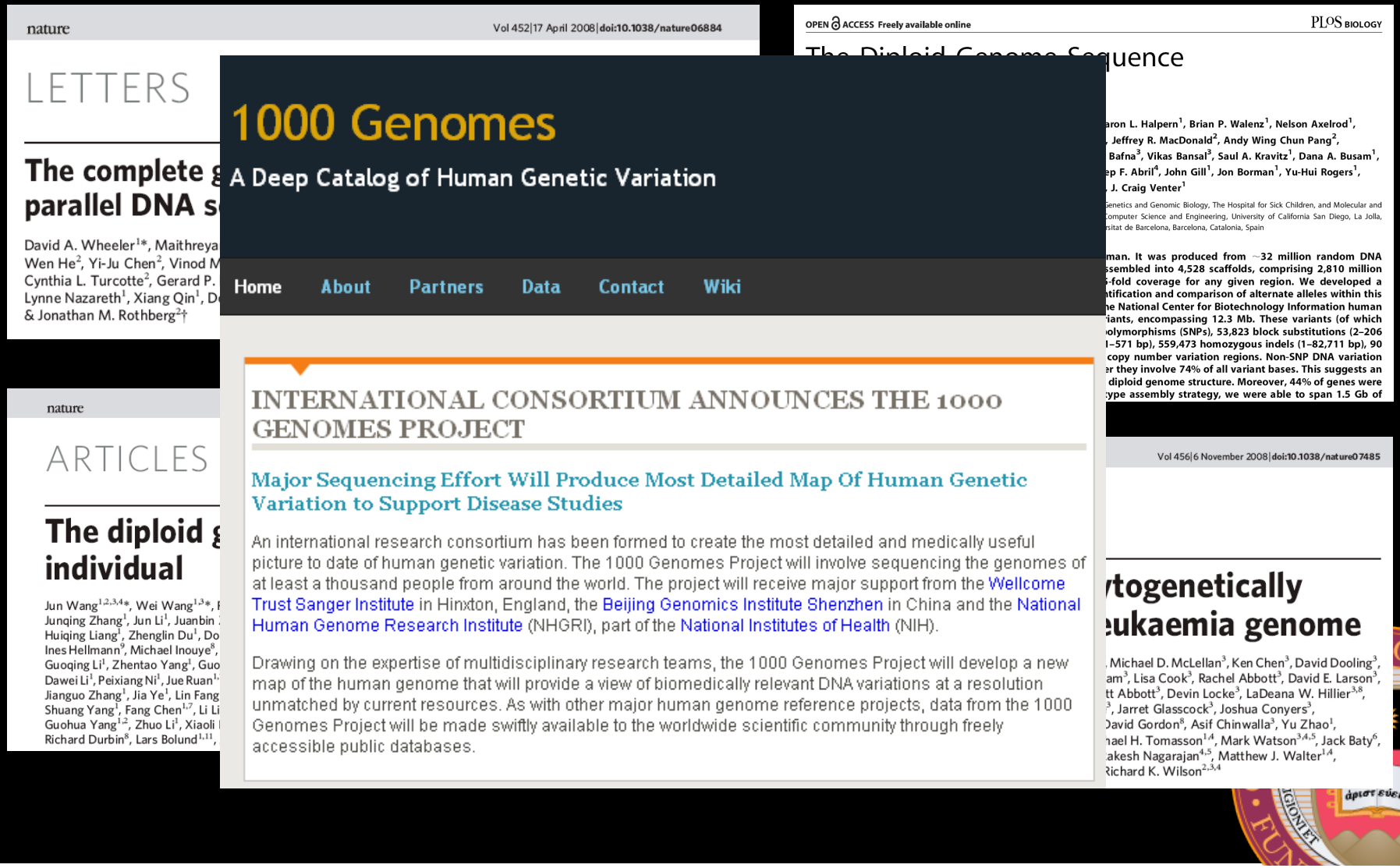
PACIFIC BIOSCIENCES™



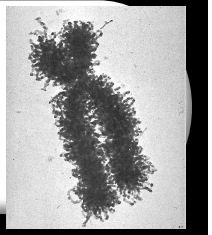
… offer vast throughput …



… & enable personal genome sequencing



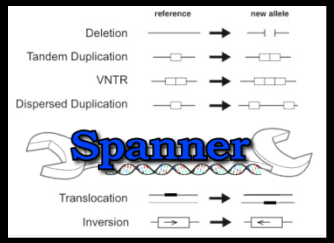
The **re-sequencing** informatics pipeline



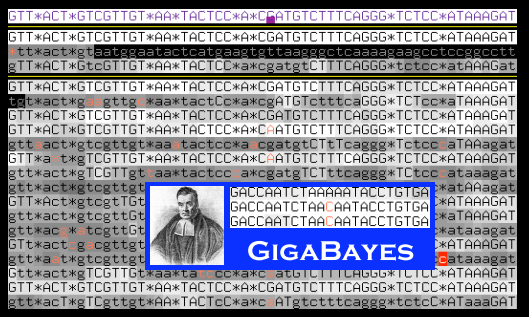
REF
IND

mosaik assembler

(ii) read mapping



(iv) SV calling

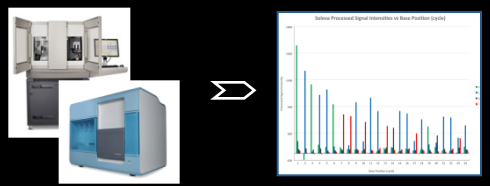


(iii) SNP and short INDEL calling

IND

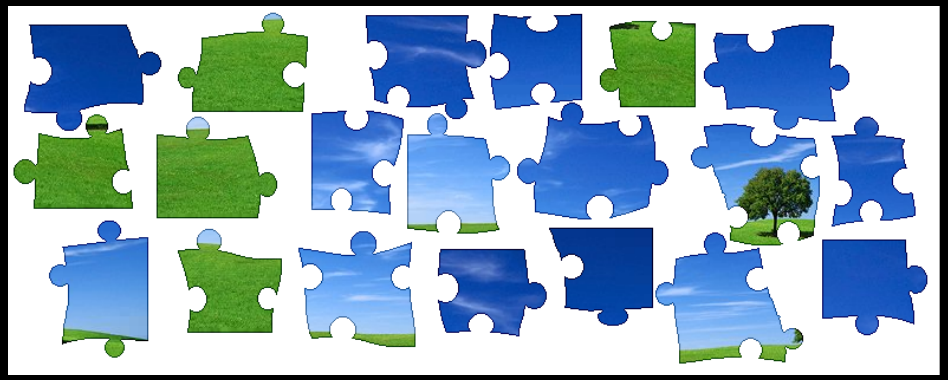
PYROBAYES

(i) base calling



(v) data viewing,
hypothesis generation

Read mapping is like a jigsaw

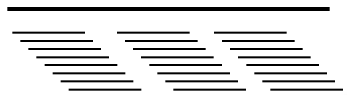


…you get the pieces…

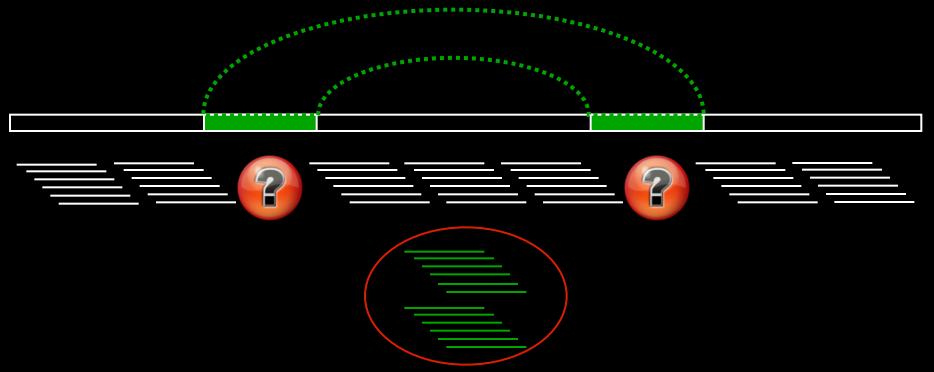
… and they give you the picture on the box

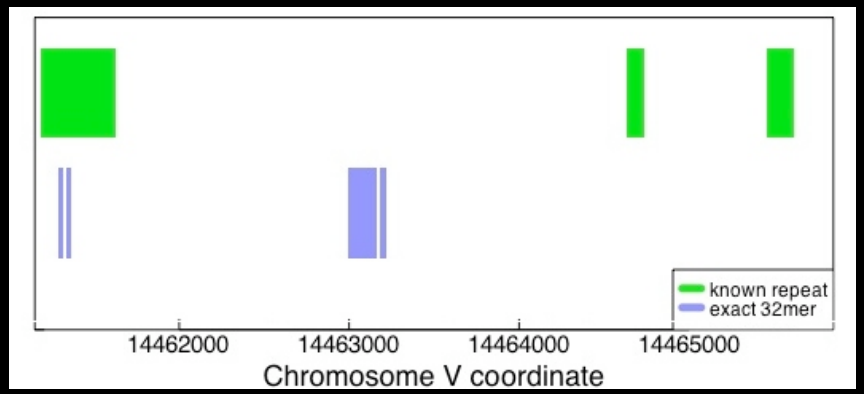
Unique pieces are easier to place than others…

Multiply-mapping reads

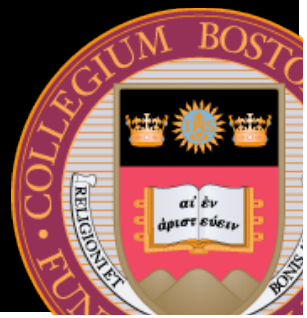


- Reads from repeats cannot be uniquely mapped back to their true region of origin

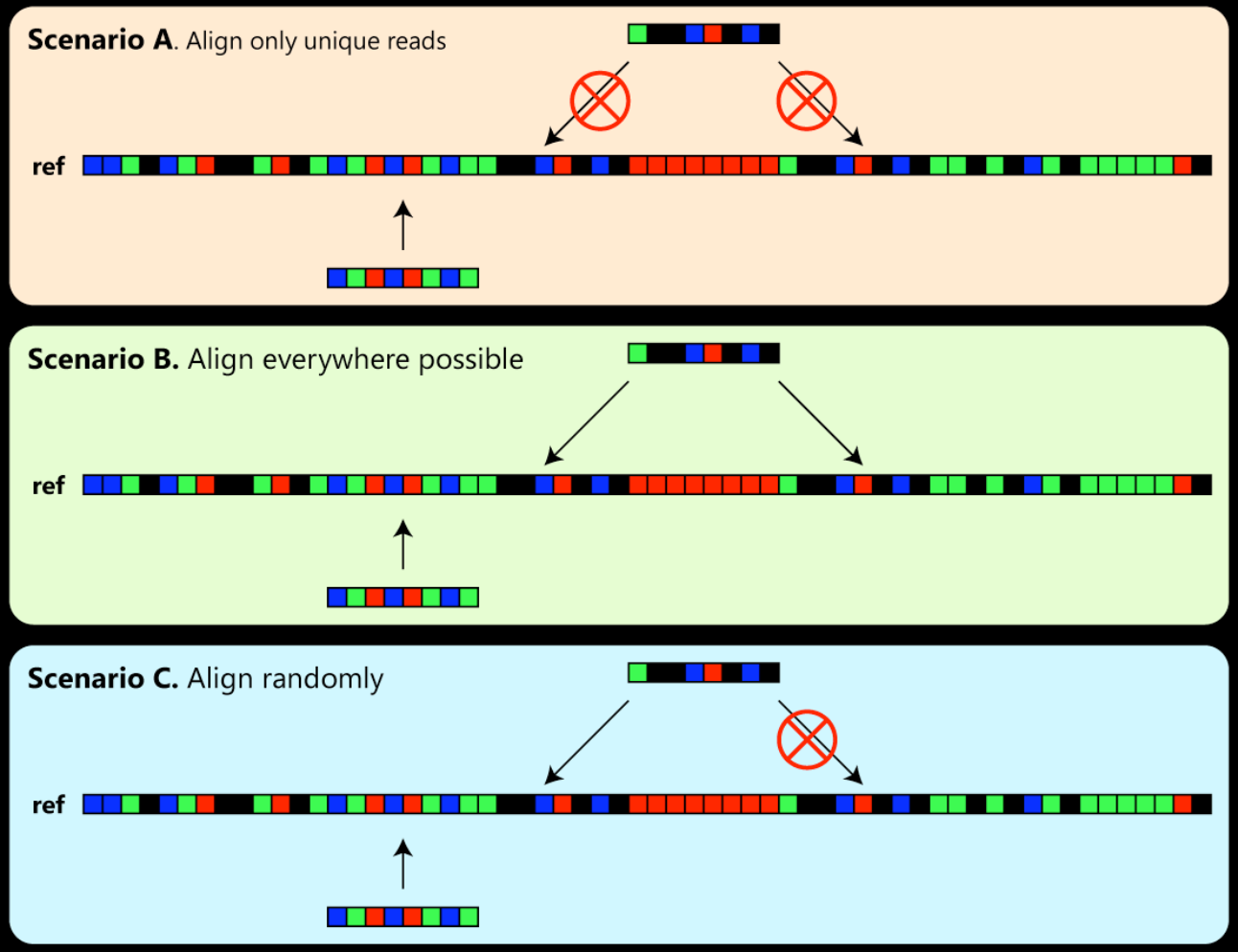




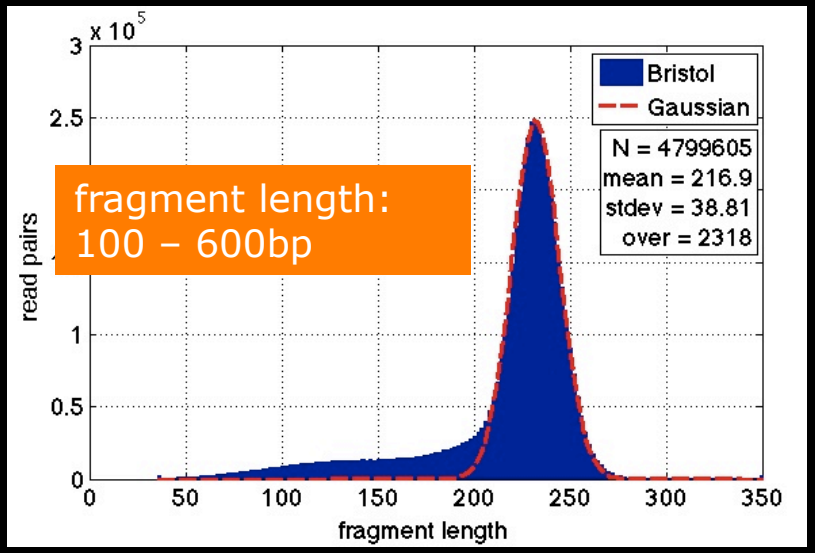
- "Traditional" repeat masking does not capture repeats at the scale of the read length

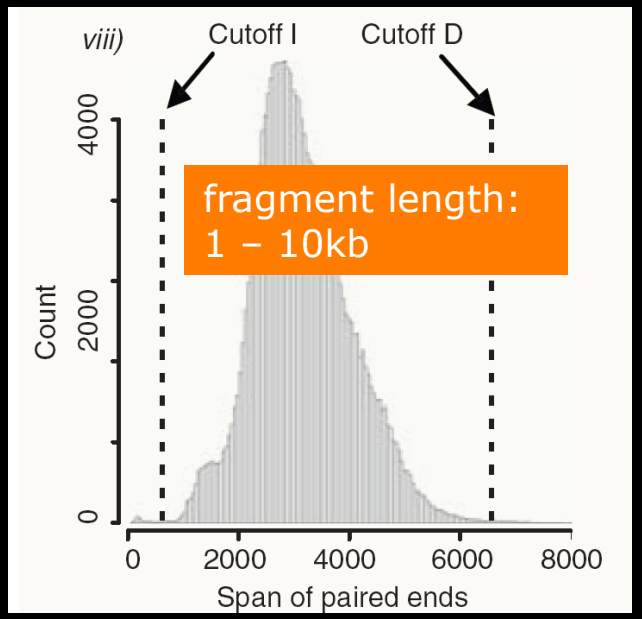


Dealing with multiple mapping



Paired-end (PE) reads



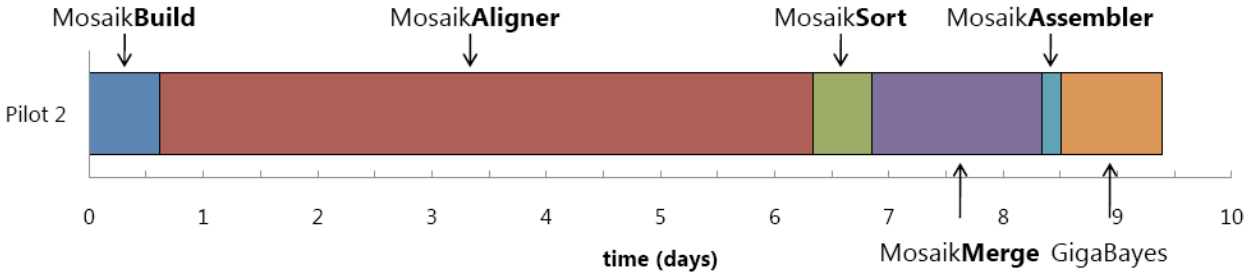


Korbel *et al*. **Science** 2007

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

Gapped alignments (for INDELs)

The MOSAIK read mapper



Pilot 2 (*January 2009*)

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

Pilot 1 (*August 2009*)

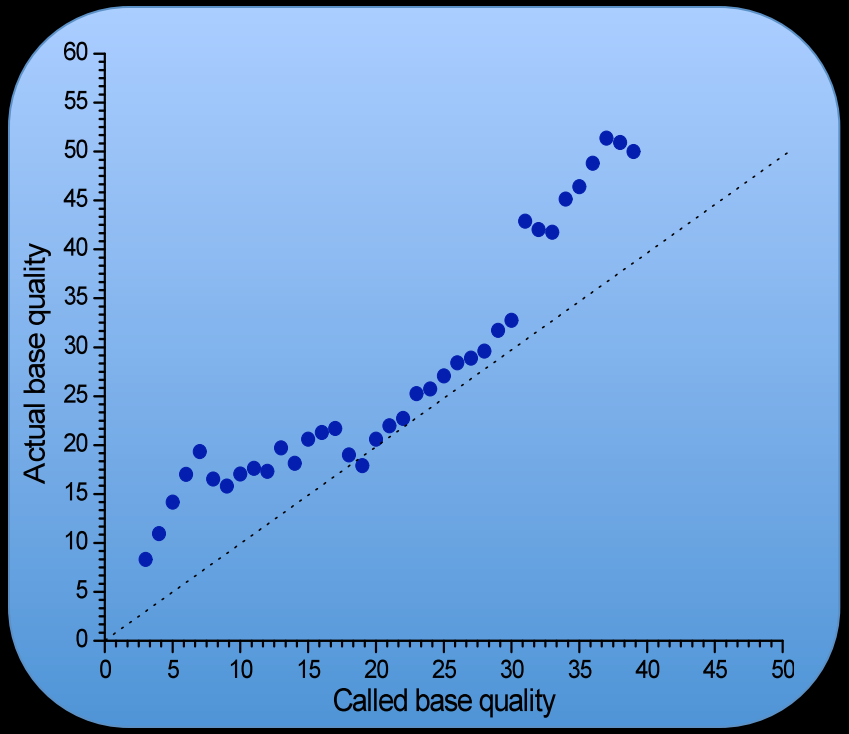
28 billion Illumina & 454 reads
31 node cluster
Alignment time: 7 days
195 reads/core/second (+16%)

mosaik assembler

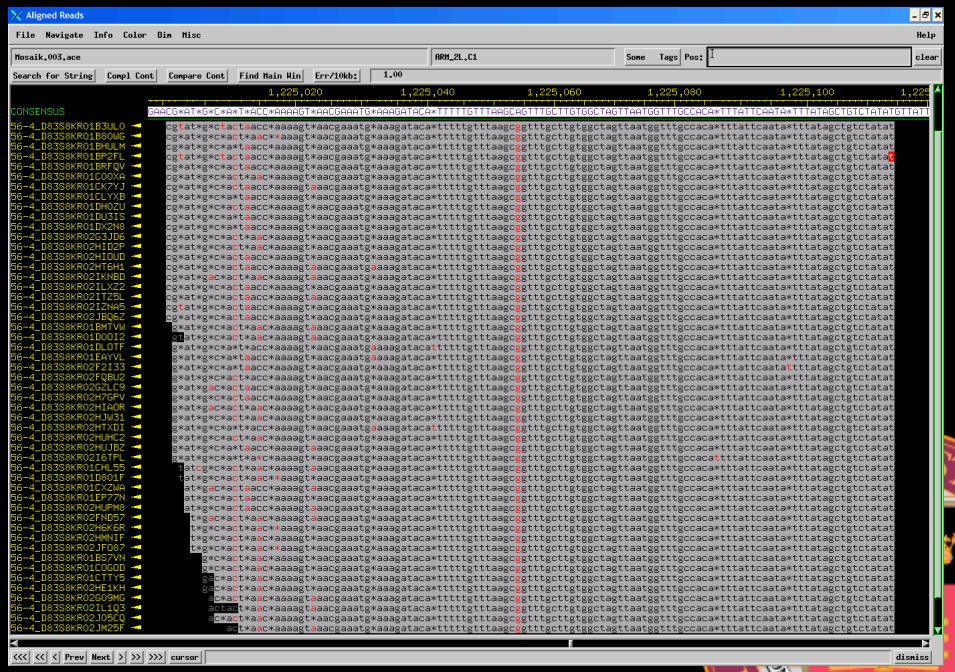
Michael Strömberg

- 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

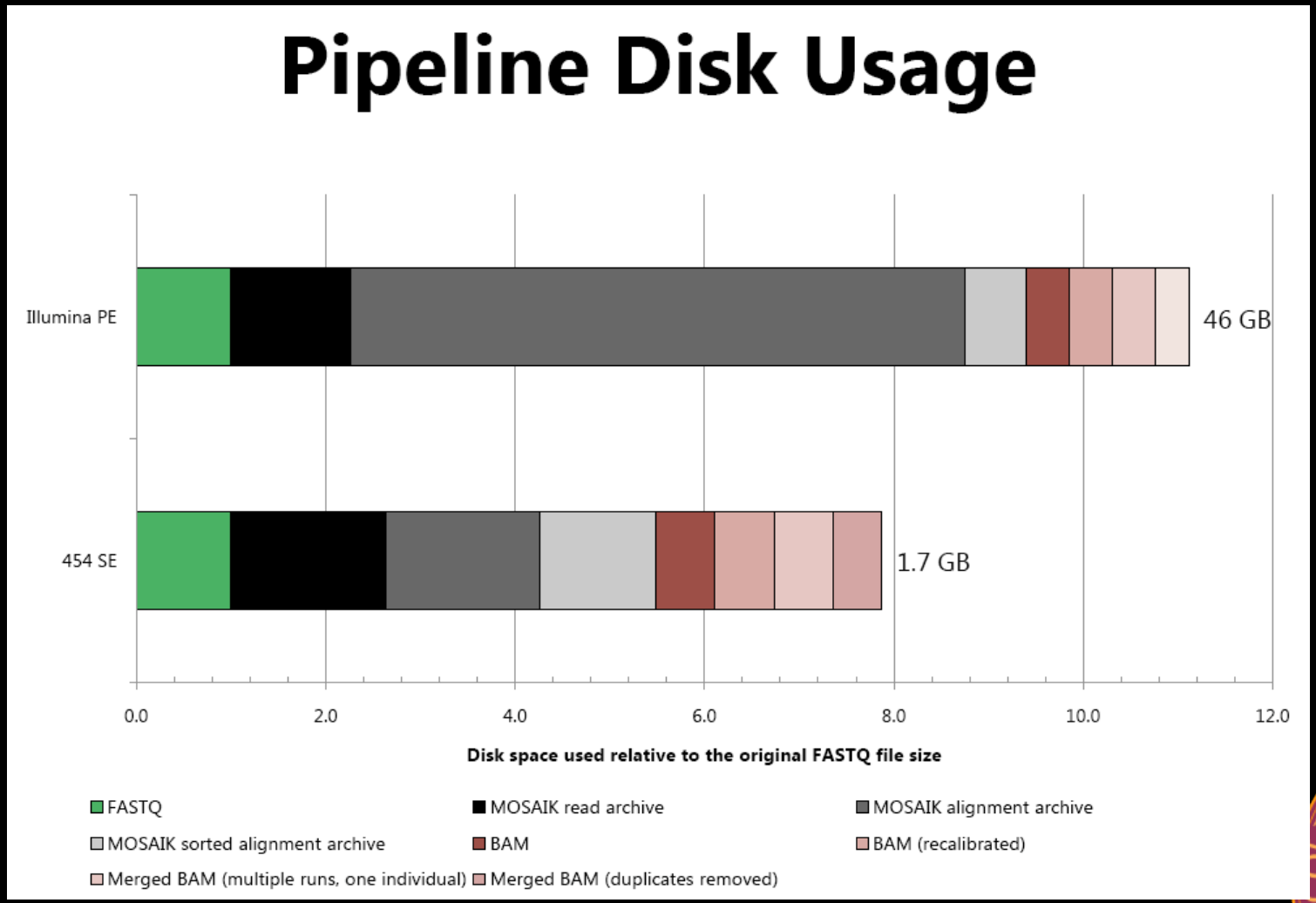


- quality value re-calibration

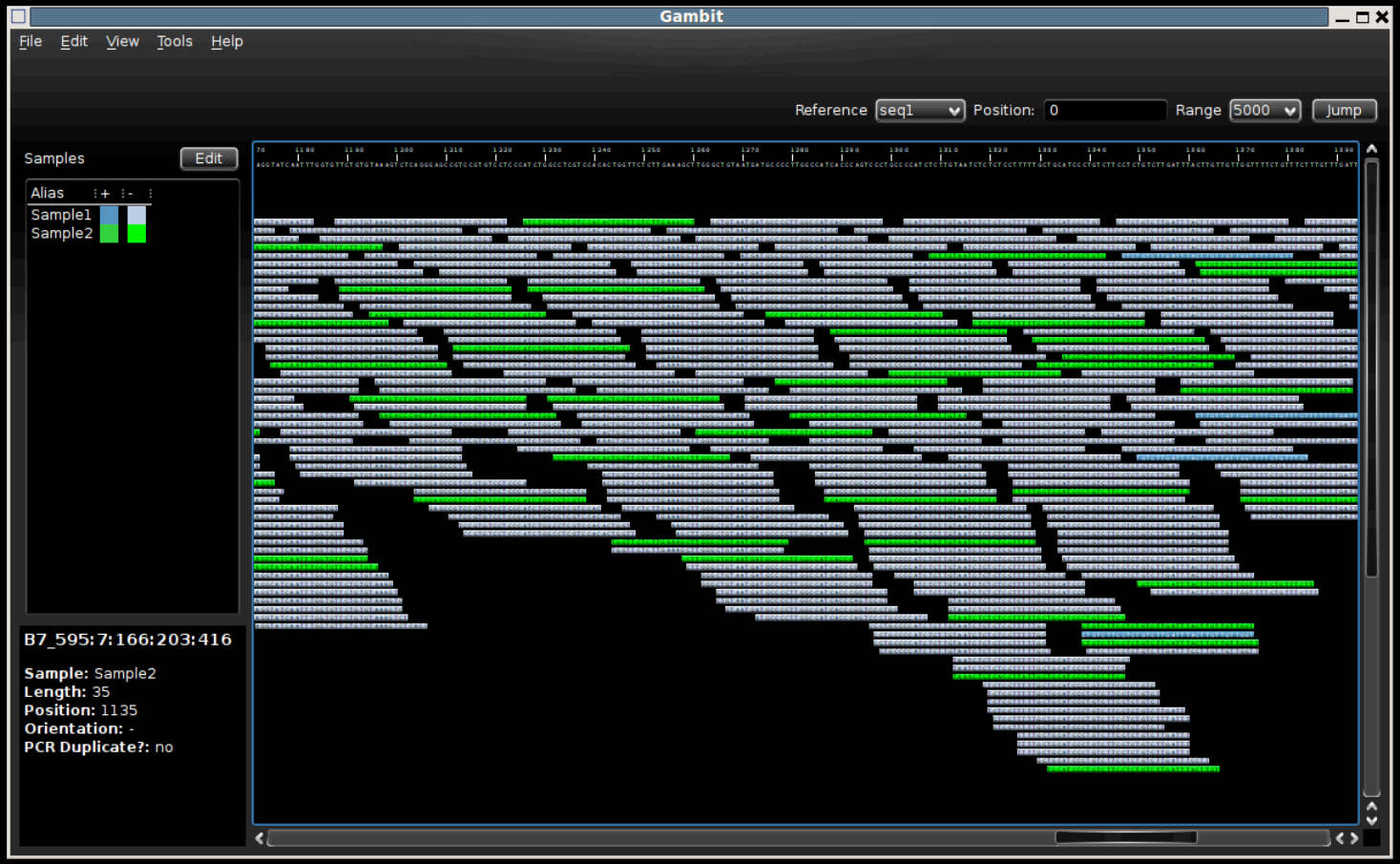


- duplicate fragment removal

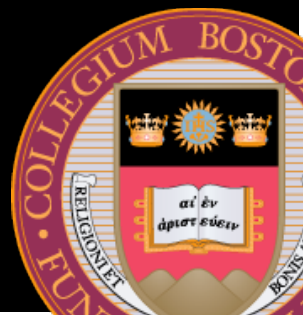
Data storage requirements



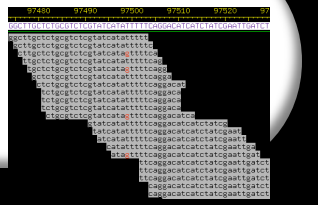
Alignment visualization



- 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*gtaatggaatactcatgaagtgttaagggctcaaaagaagcctccggcctt
gTT*ACT*GtcGTTGT*AA*TACTCC*a*cgatgtCTTTCAGGG*tctcc*atAAAGat

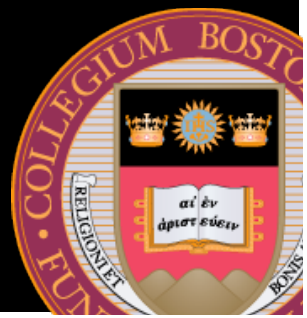
GTT*ACT*GTCGTTGT*AA*TACTCC*A*CGATGTCTTTCAGGG*TCTCC*ATAAAGAT
tgt*act*gaagttgc*aa*tactCc*a*cgATGTctttcaGGG*TCTcc*aTAAAGAT
GTT*ACT*GTCGTTGT*AA*TACTCC*A*CGATGTCTTTCAGGG*TCTCC*ATAAAGAT

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

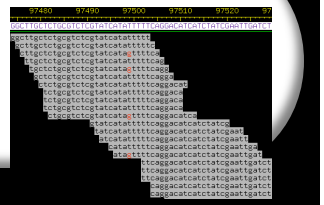
GTT*ACT*GTCGTTGT*AA*TACTCC*A*CGATGTCTTTCAGGG*TCTCC*ATAAAGAT
gtt*acT*gTCgttgt*AA*TACTcC*a*caTgtctttcaggg*tctcC*ATaaaGAT

sequencing error

polymorphism

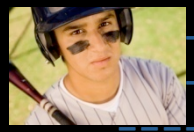


SNP calling in multi-sample read sets



--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-----

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

"genotype likelihoods"

$Prior(G_1,..,G_i,..,G_n)$

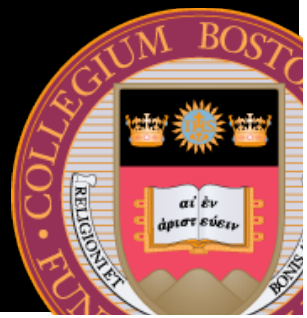
$P(G_1=aa|B_1=aacc;\ B_i=aaaac;\ B_n=cccc)$
$P(G_1=cc|B_1=aacc;\ B_i=aaaac;\ B_n=cccc)$
$P(G_1=ac|B_1=aacc;\ Bi=aaaac;\ B_n=cccc)$

$P(G_i=aa|B_1=aacc;\ B_i=aaaac;\ B_n=cccc)$
$P(G_i=cc|B_1=aacc;\ B_i=aaaac;\ B_n=cccc)$
$P(G_i=ac|B_1=aacc;\ Bi=aaaac;\ B_n=cccc)$

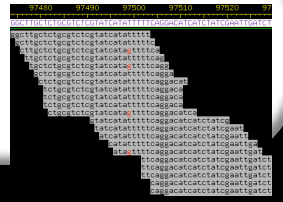
$P(G_n=aa|B_1=aacc;\ B_i=aaaac;\ B_n=cccc)$
$P(G_n=cc|B_1=aacc;\ B_i=aaaac;\ B_n=cccc)$
$P(G_n=ac|B_1=aacc;\ Bi=aaaac;\ B_n=cccc)$

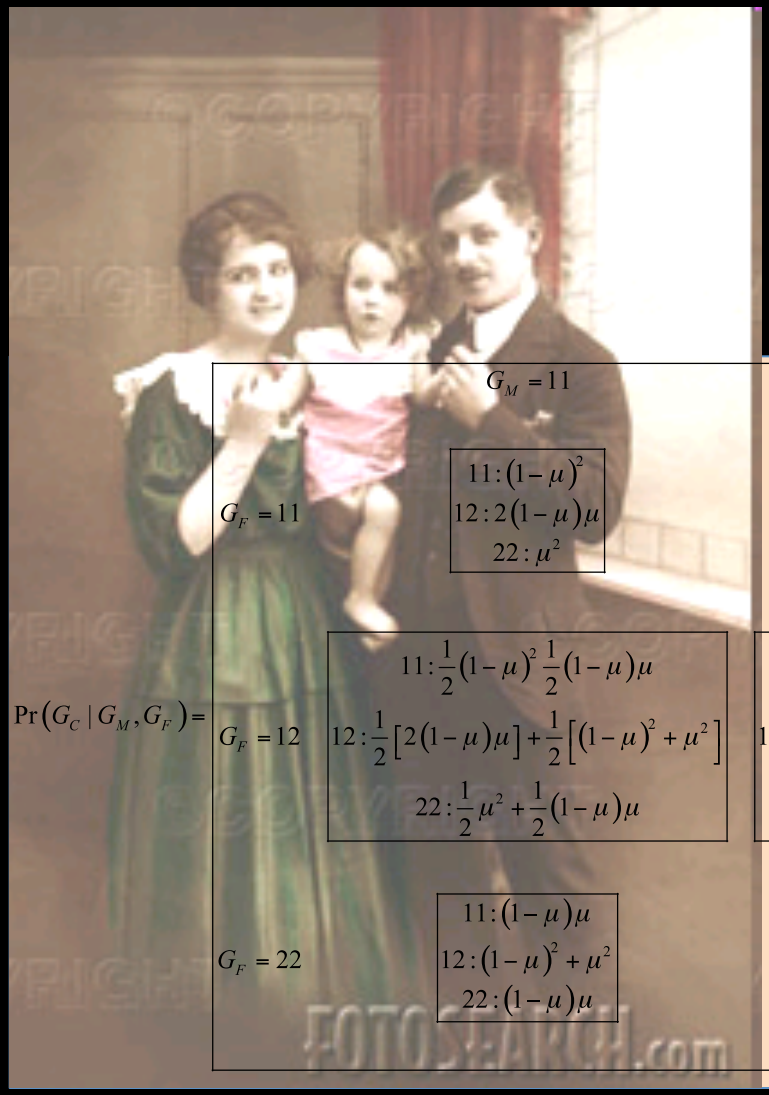
"genotype probabilities"

P(SNP)



Trio sequencing



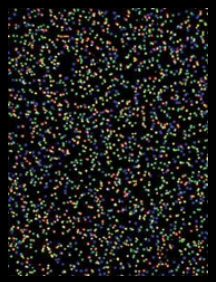


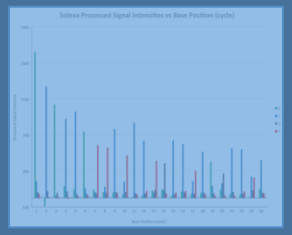
- the child inherits one chromosome from each parent
- there is a small probability for a *de novo* (germ-line or somatic) mutation in the child

$$\Pr(G_C \mid G_M, G_F) =$$

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

Standard data formats

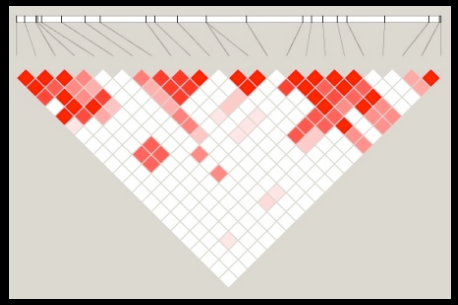




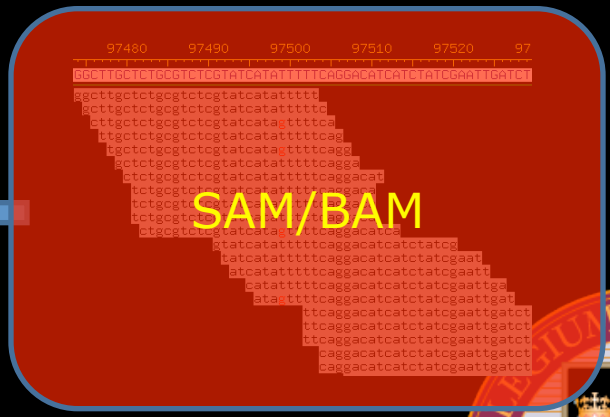
```
>B_TITR_1_1_668_35 TIME: Tue Feb 20 02:26:06 2007
ATATCGGATGACACAATATGGGAGGTTGAC
>B_TITR_1_2_843_403 TIME: Tue Feb 20 02:26:06 2007
TGTAGCTTTTCATGACAATTTTATAGGTGT
```

SRF/FASTQ

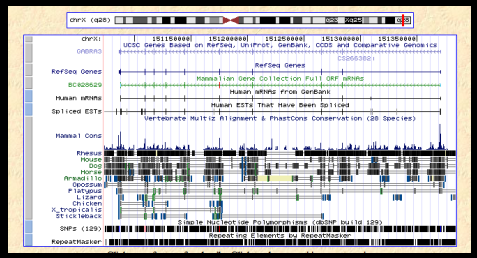
```
>B_TITR_1_1_668_35
27 27 27 27 27 27 27 27 27 27 27 27 27 27 27 27 27 27 27 27 27 27 27
>B_TITR_1_2_843_403
26 26 26 26 26 26 26 26 26 26 26 26 26 26 26 26 26 26 26 26 26 26 26
>B_TITR_1_3_618_922
```

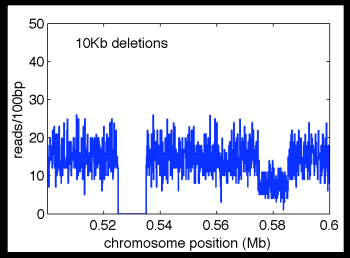


GLF/VCF



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.**

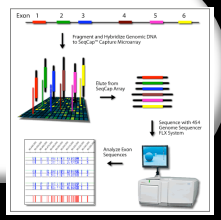
   This pilot will evaluate the relationship between coverage depth and the yield of variation data, based on genomic sequence from a few individuals. A high level of redundancy will provide a solid basis for assessing the coverage needed for discovering variants, inferring haplotypes, imputing non-typed 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).

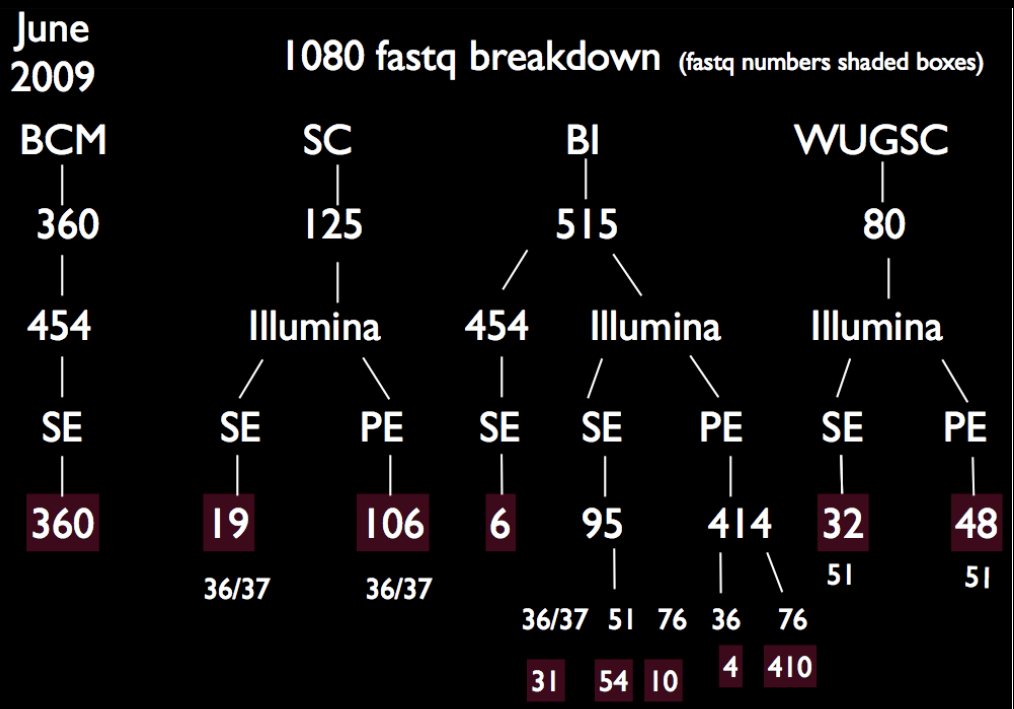
1000G Pilot 3 – exon sequencing



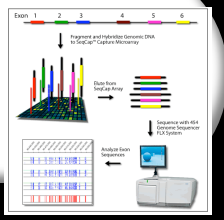
1000 Genomes

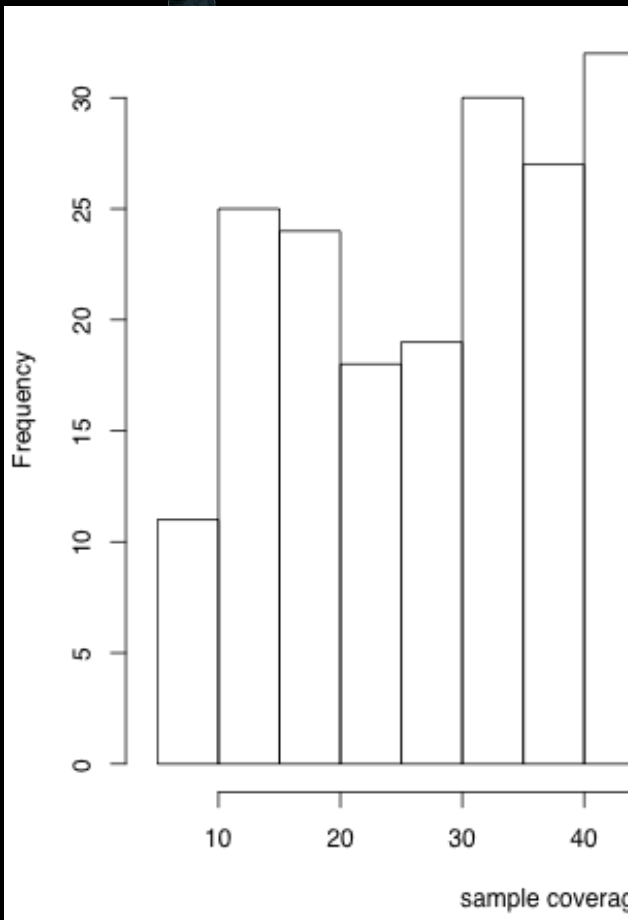
A Deep Catalog of Human Genetic Variation

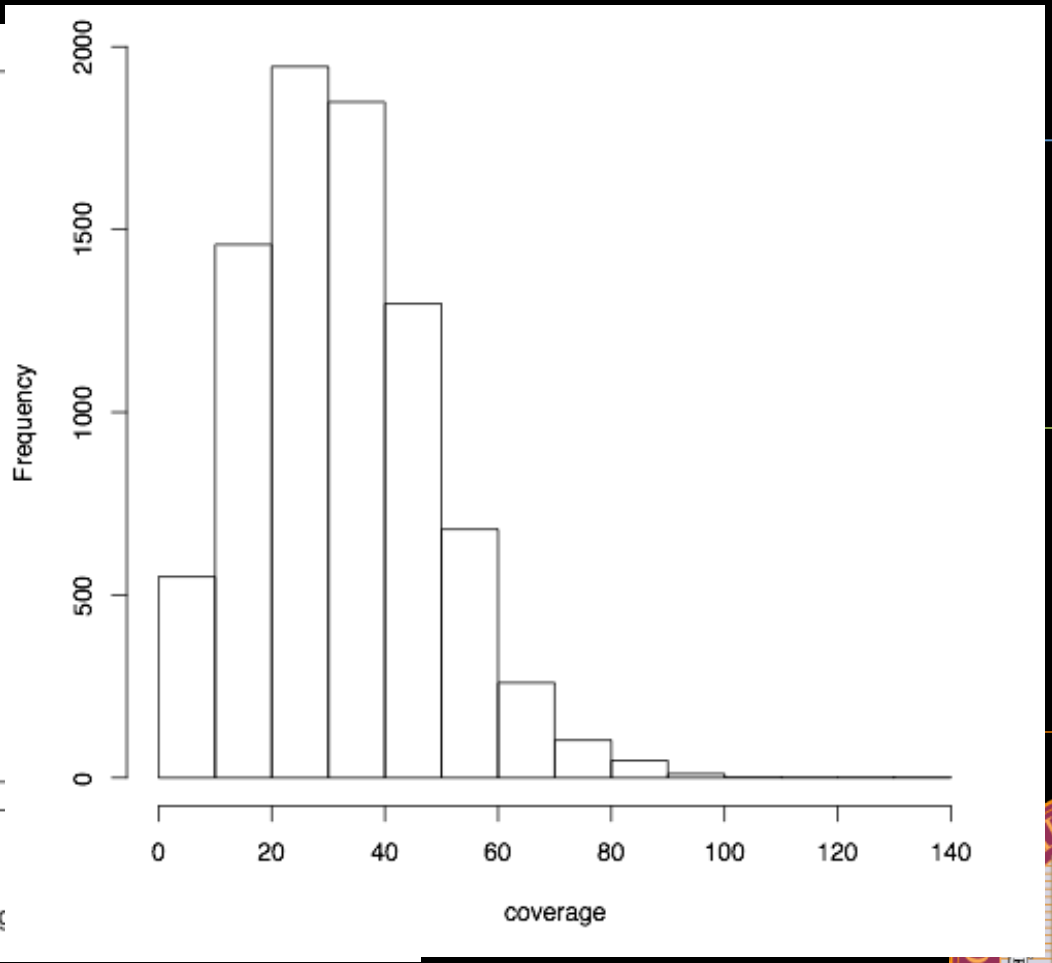
- Targets:
  - 1K genes / 10K targets
- Capture:
  - Solid / liquid phase
- Sequencing:
  - 454 / Illumina
  - SE / PE
- Data producers:
  - Baylor
  - Broad
  - Sanger
  - Wash. U.
- Informatics methods:
  - Multiple read mapping & SNP calling programs



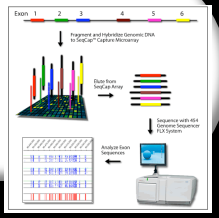
Coverage varies

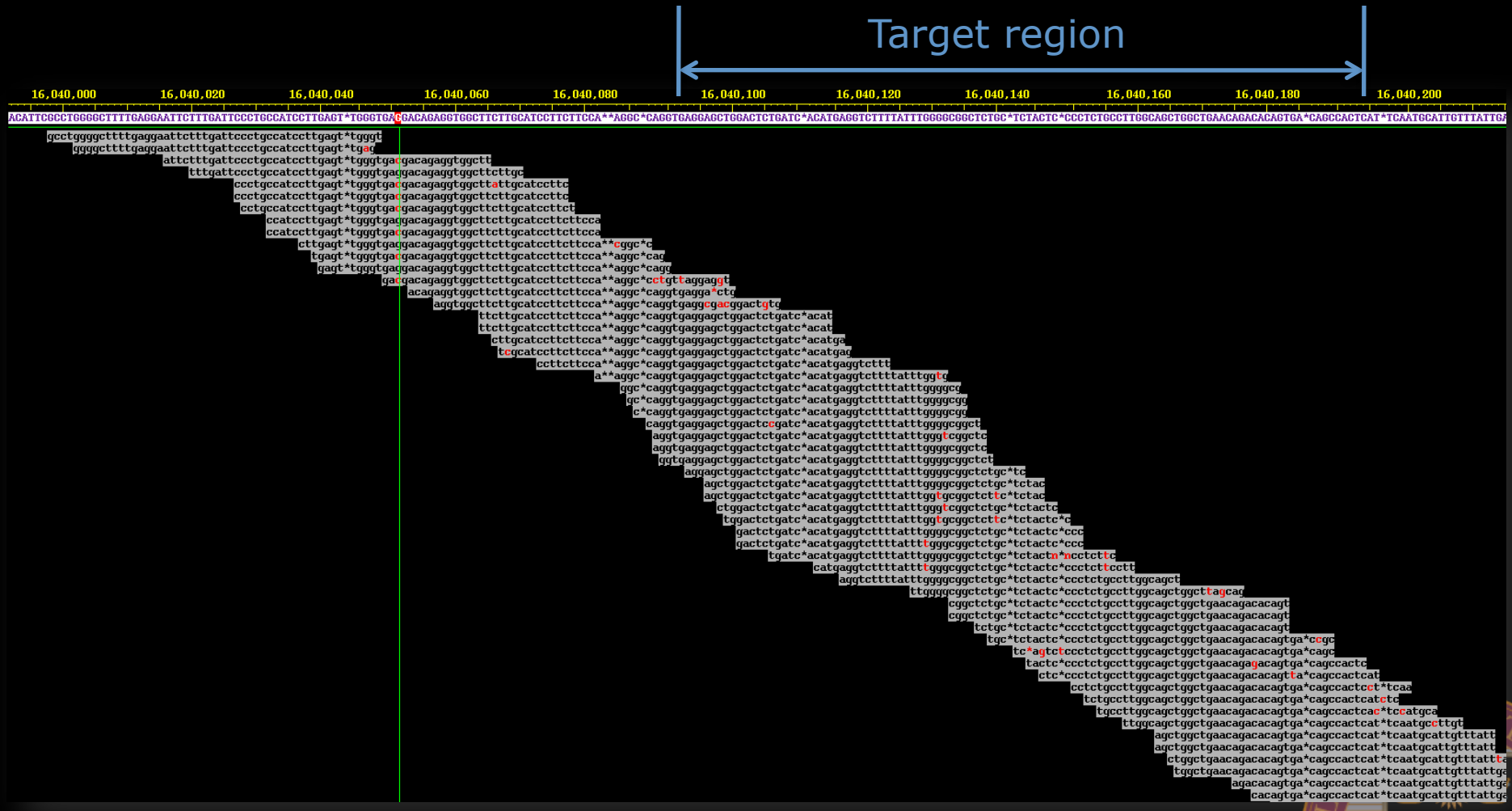






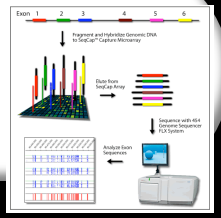
On/off target capture

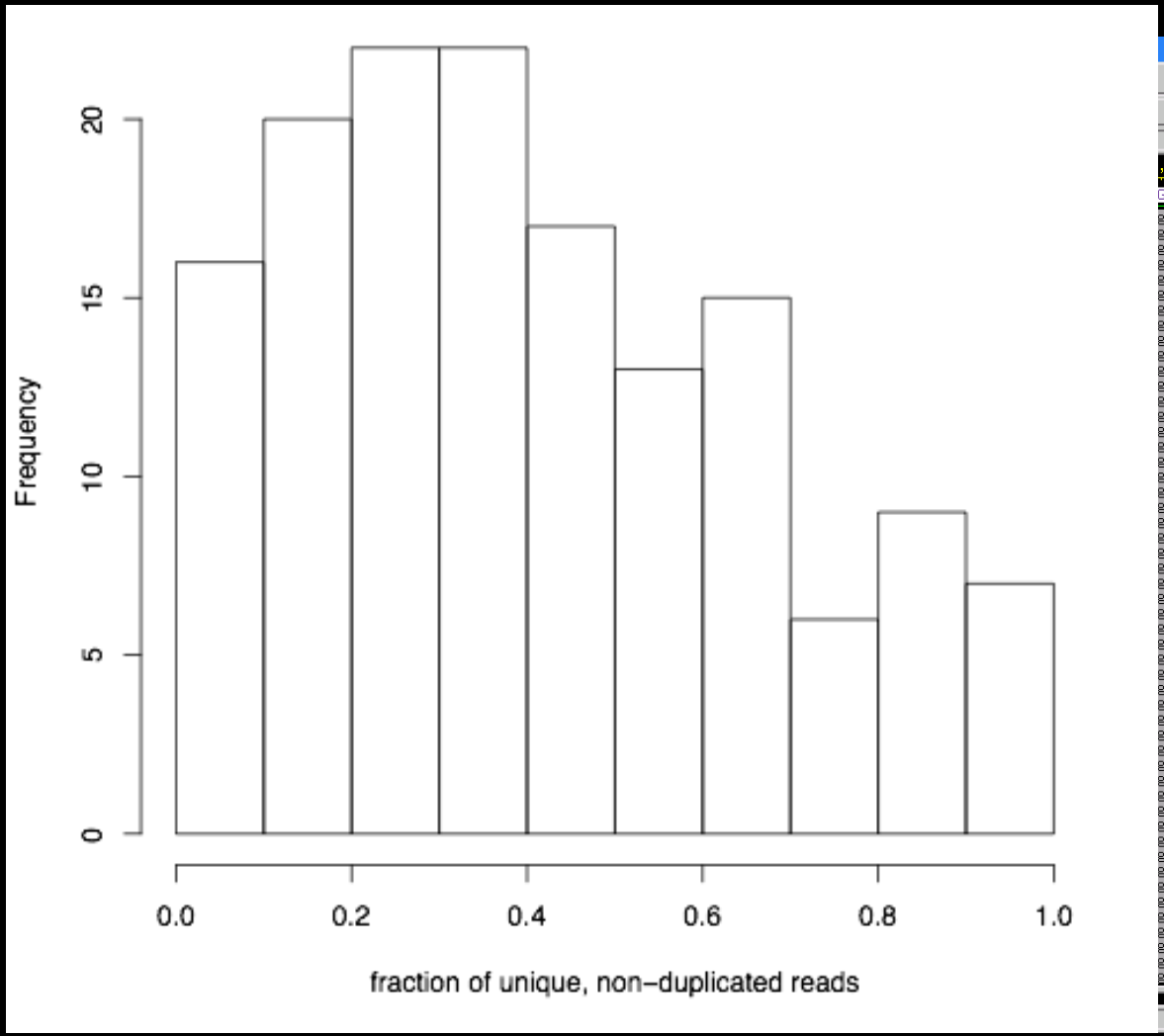


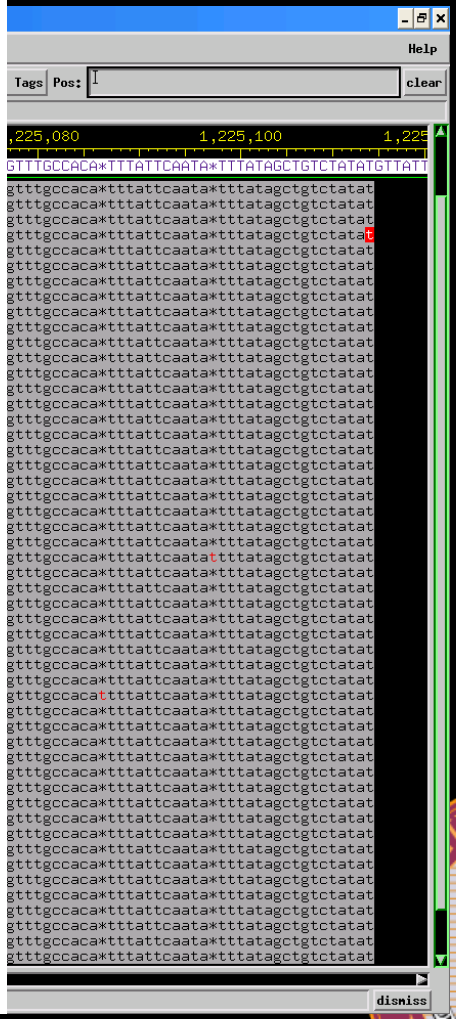


SNP
(outside target region)

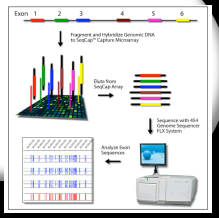
Fragment duplication

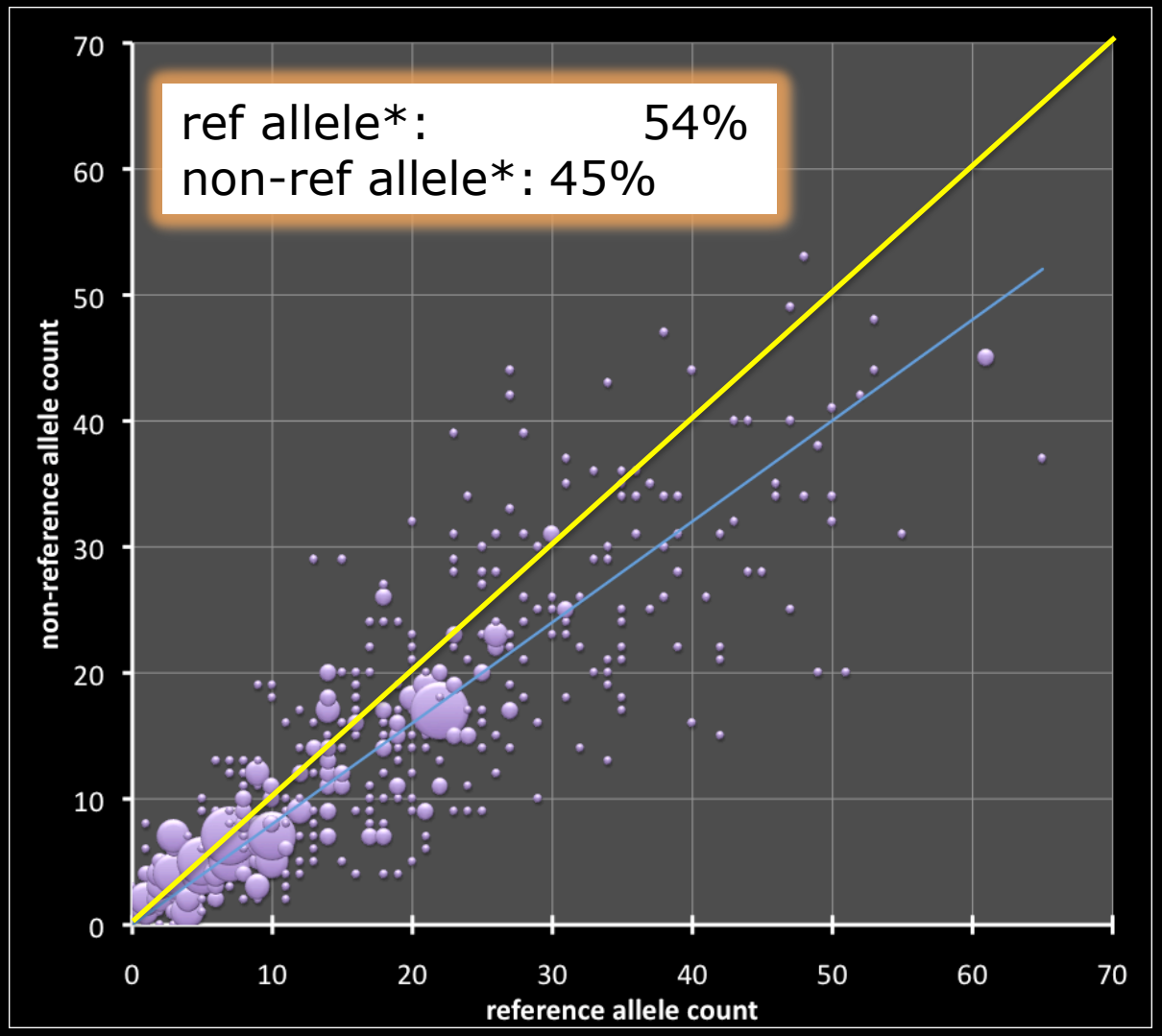






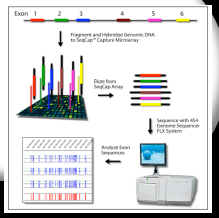
Reference allele bias





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

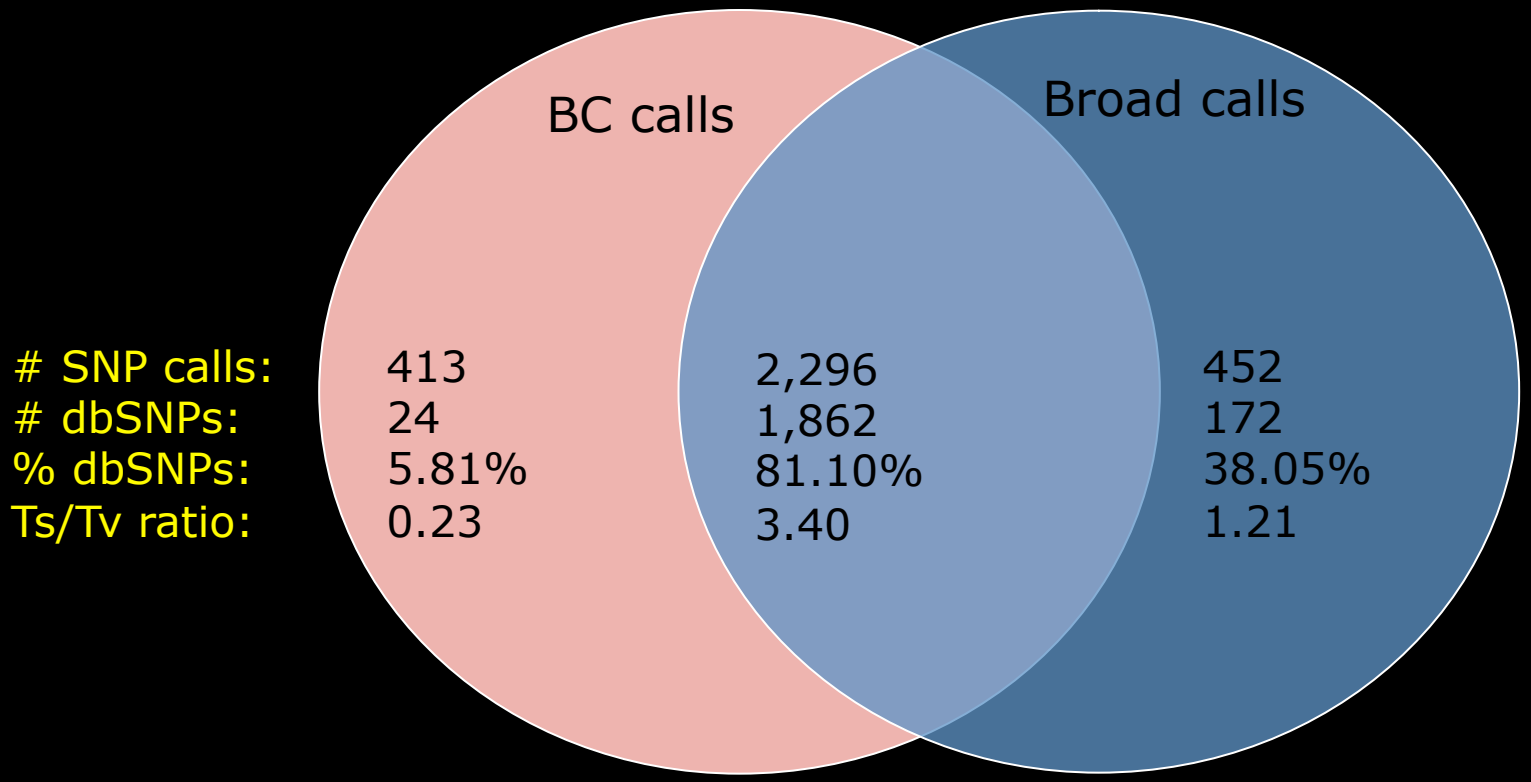
SNP calling findings



- 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



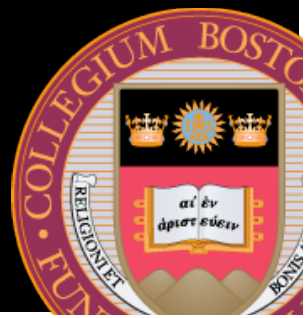
The 1000G Structural Variation Discovery Effort

1000 Genomes
A Deep Catalog of Human Genetic Variation

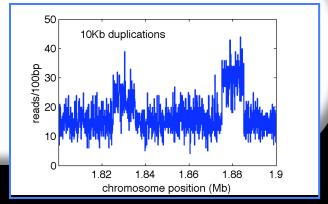
Primary goals:

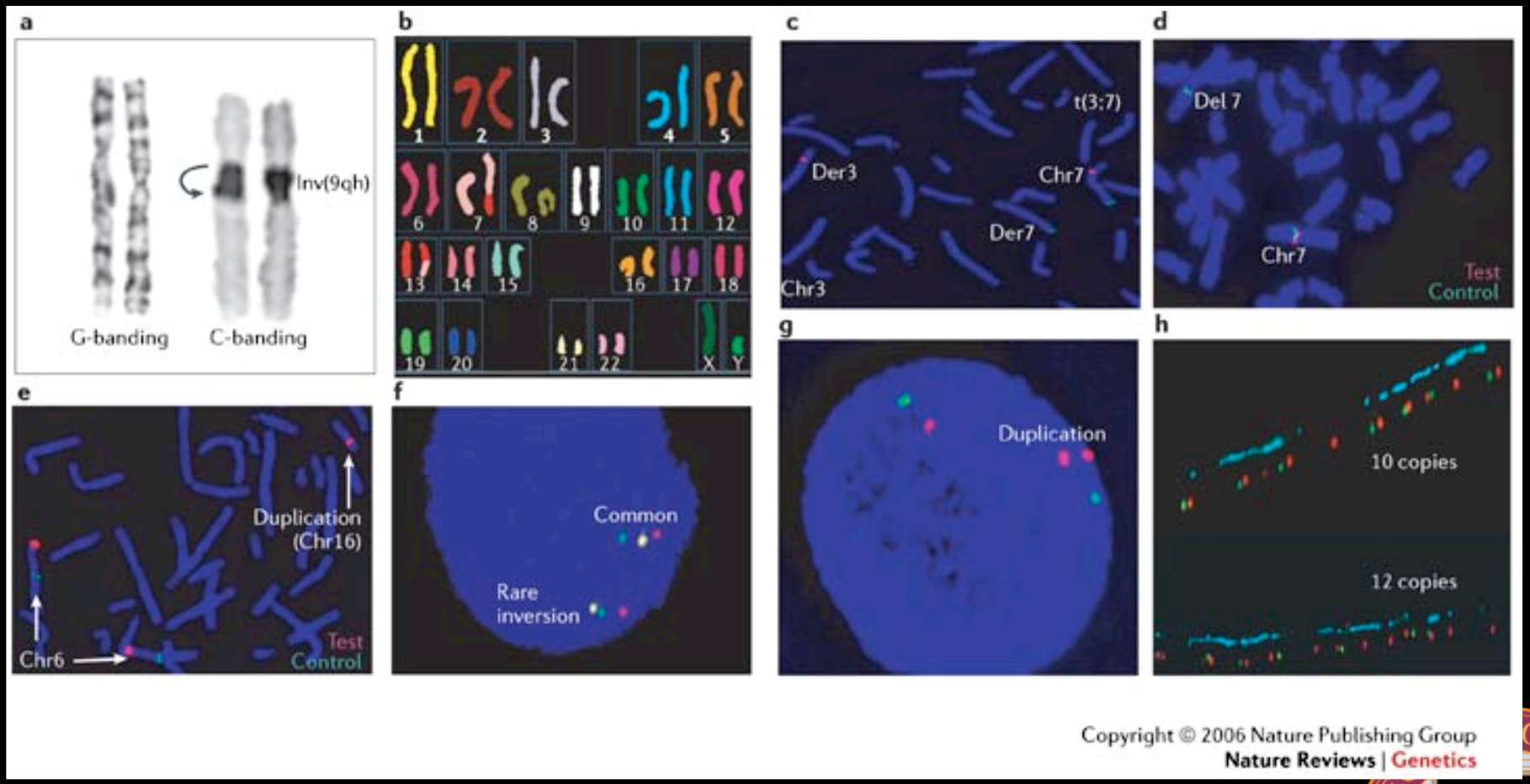
1. Discover variants (SNPs, copy-number variants, insertions, deletions, other structural variants).

   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.



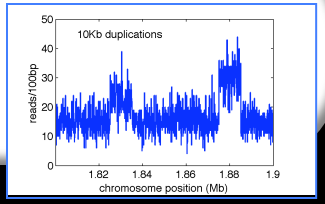
Structural variation detection

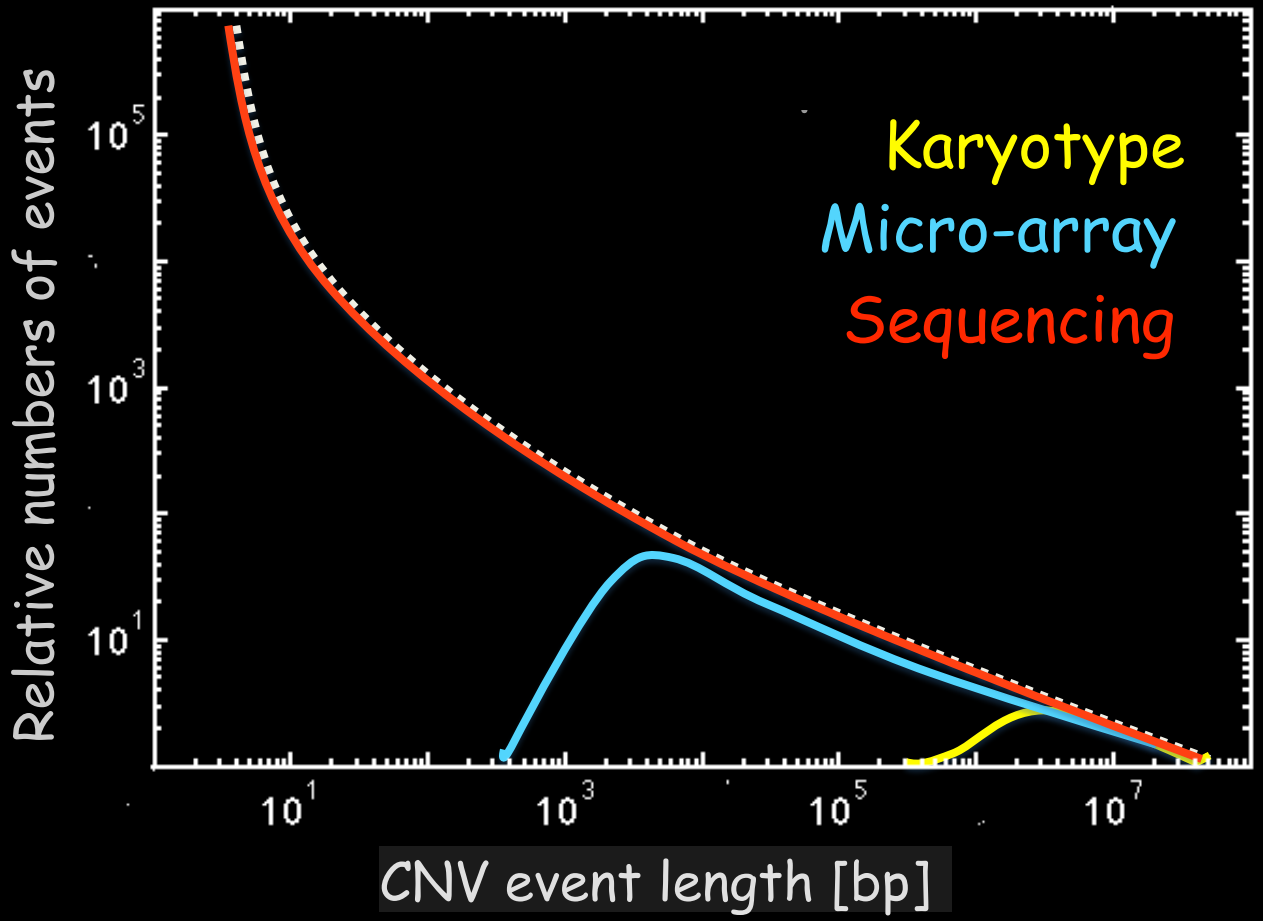




Feuk *et al.* Nature Reviews Genetics, 2006

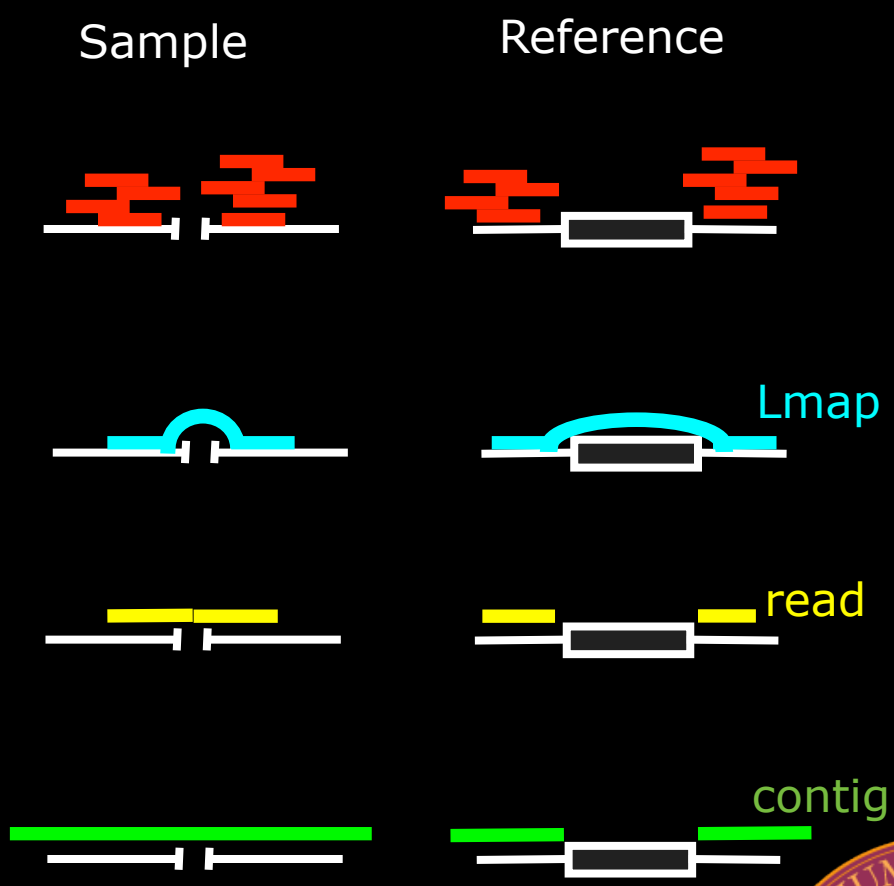
SV detection – resolution





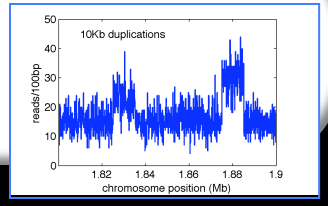
Detection Approaches

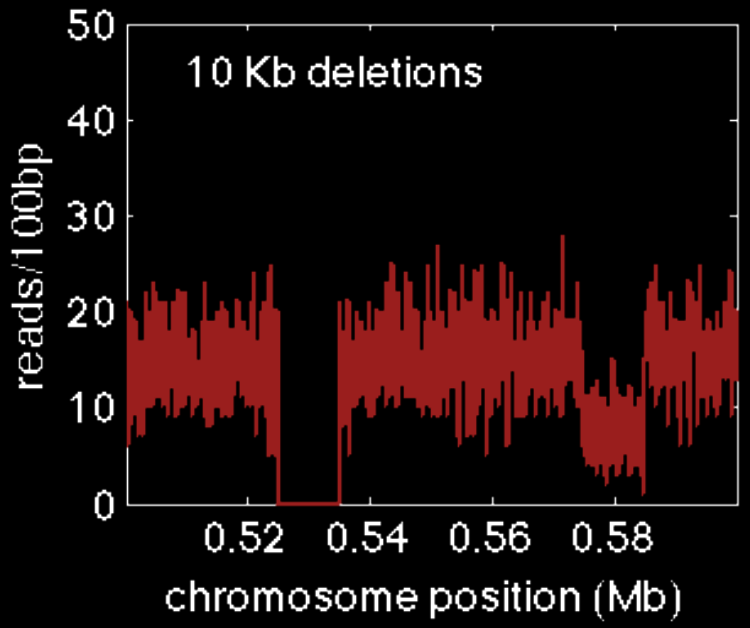
- Read Depth:
  good for big CNVs
- Paired-end:
  all types of SV
- Split-Reads
  good break-point
  resolution
- deNovo Assembly
  ~ the future

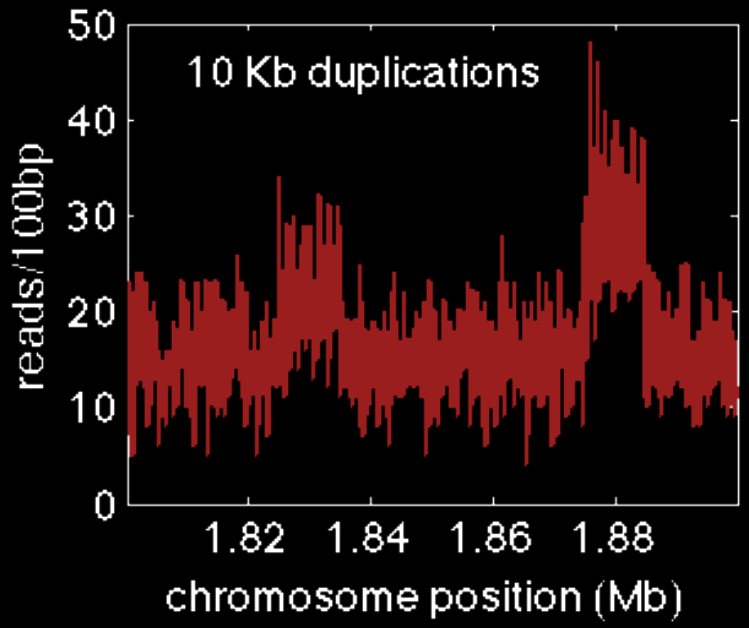


SV slides courtesy of Chip Stewart, Boston College

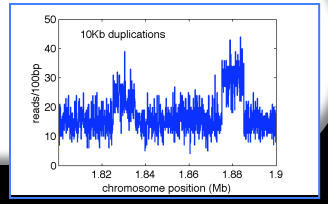
Read depth (RD)

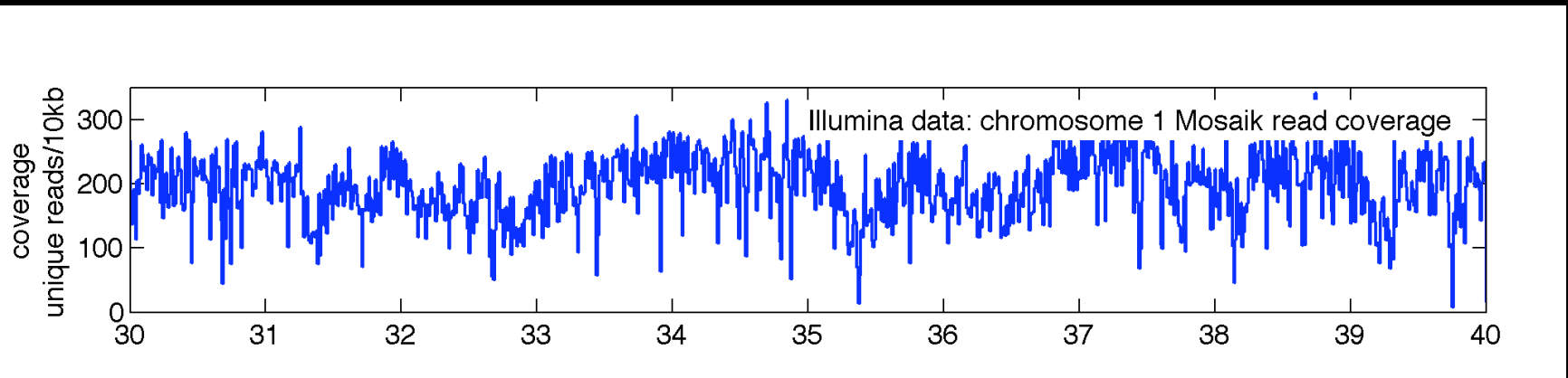


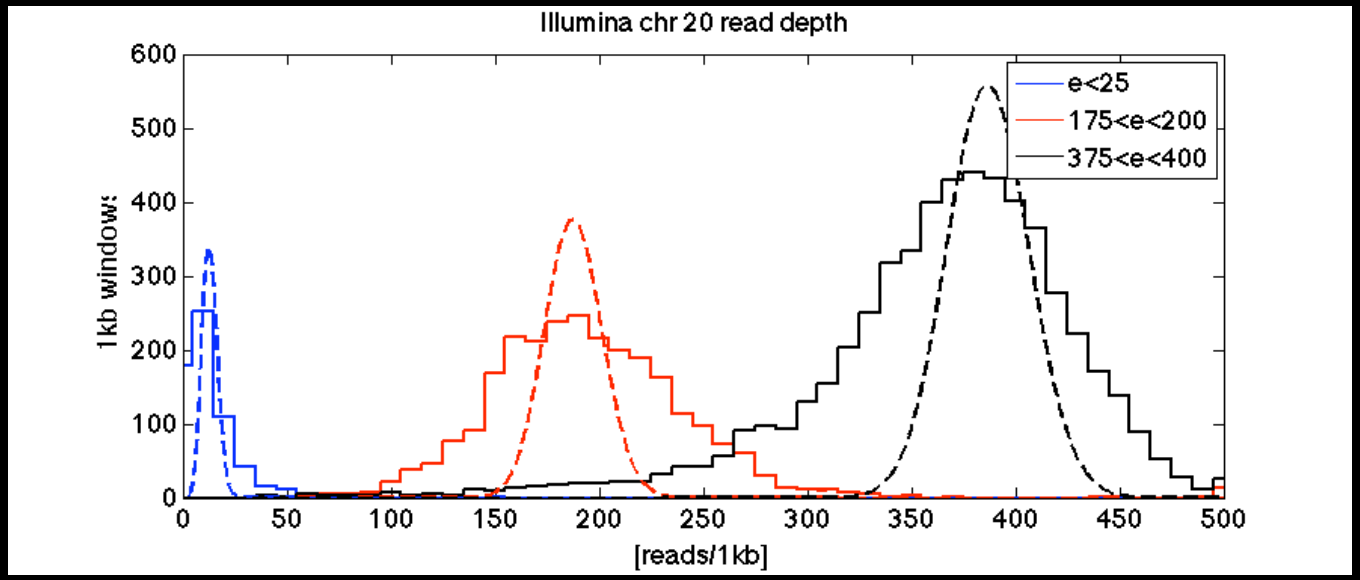




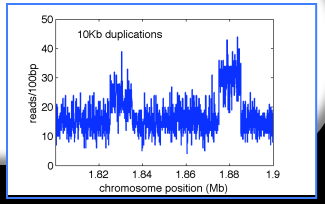
Statistical & systematic biases

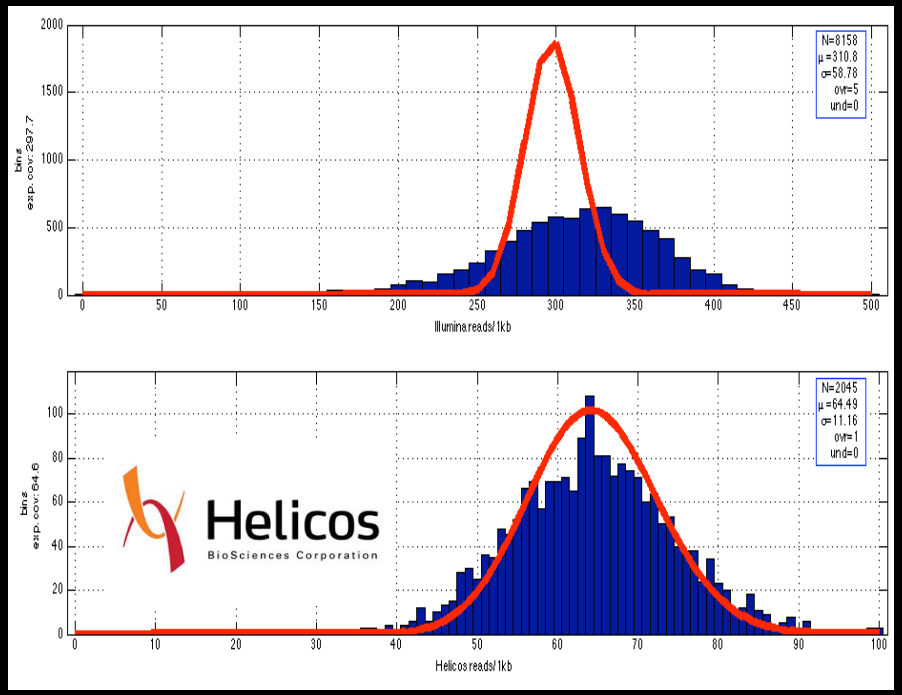






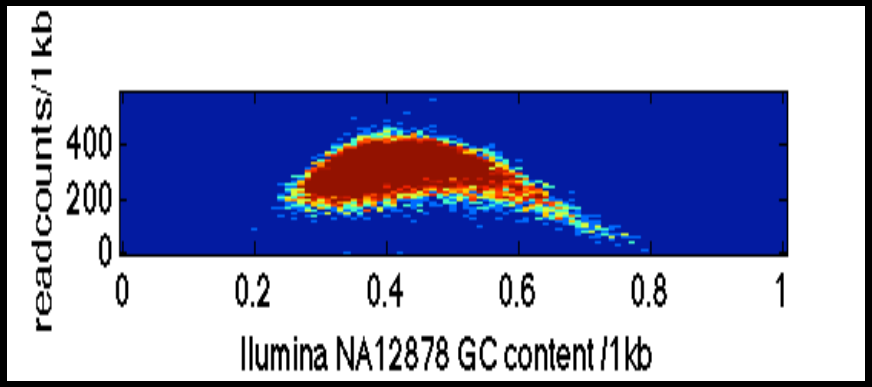
Single molecule sequencing?

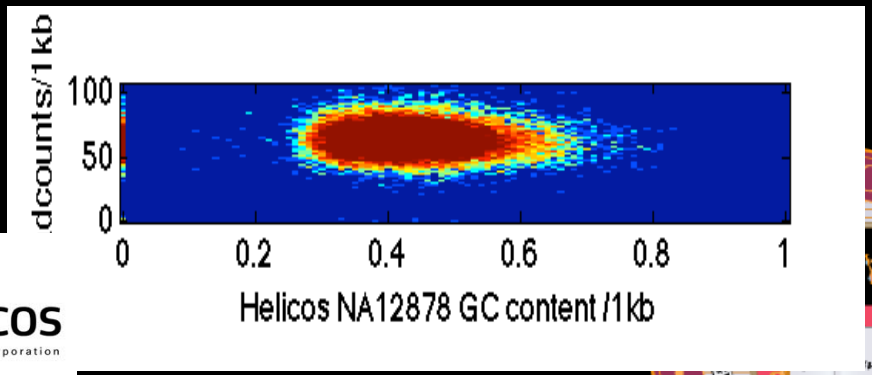




Coverage bias

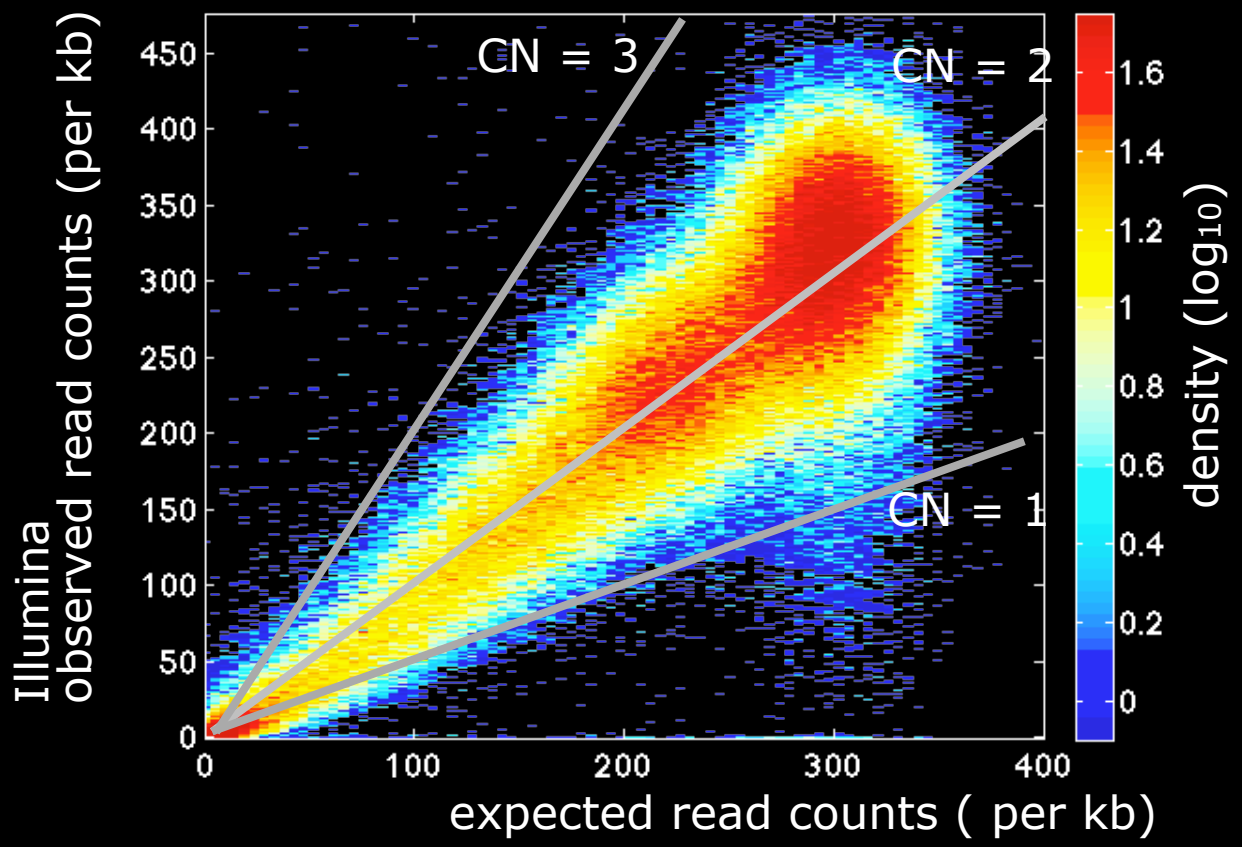
GC Bias





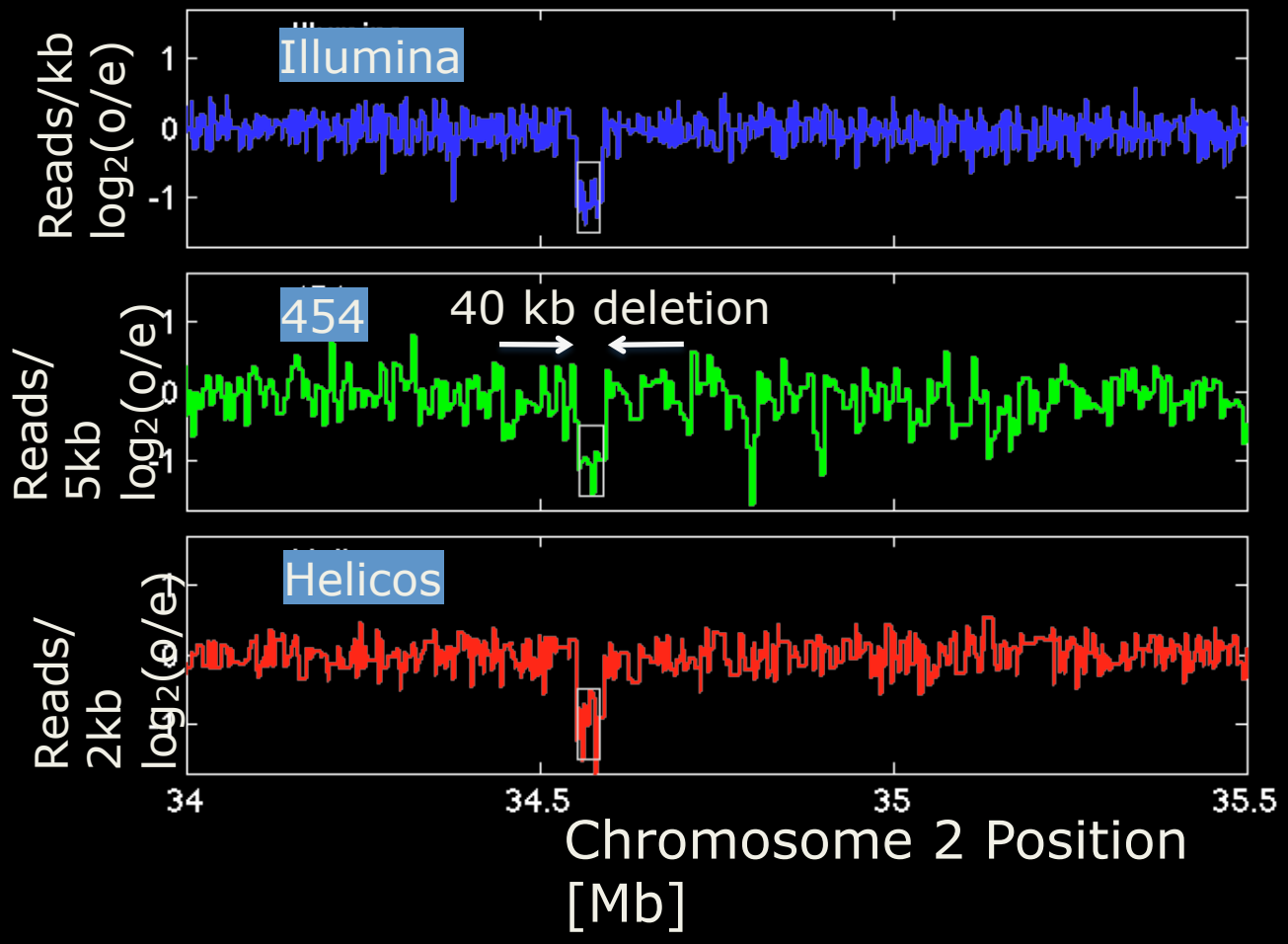
Helicos
BioSciences Corporation

RD resolution

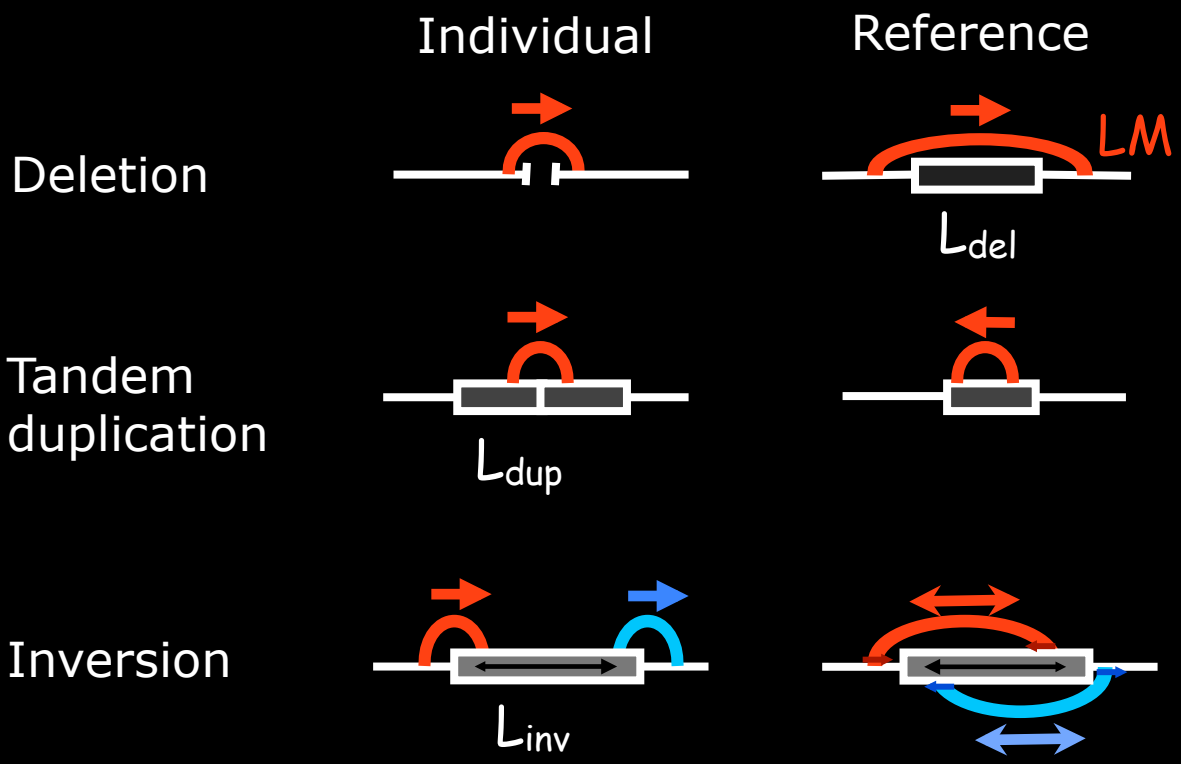


CNV events detected with RD

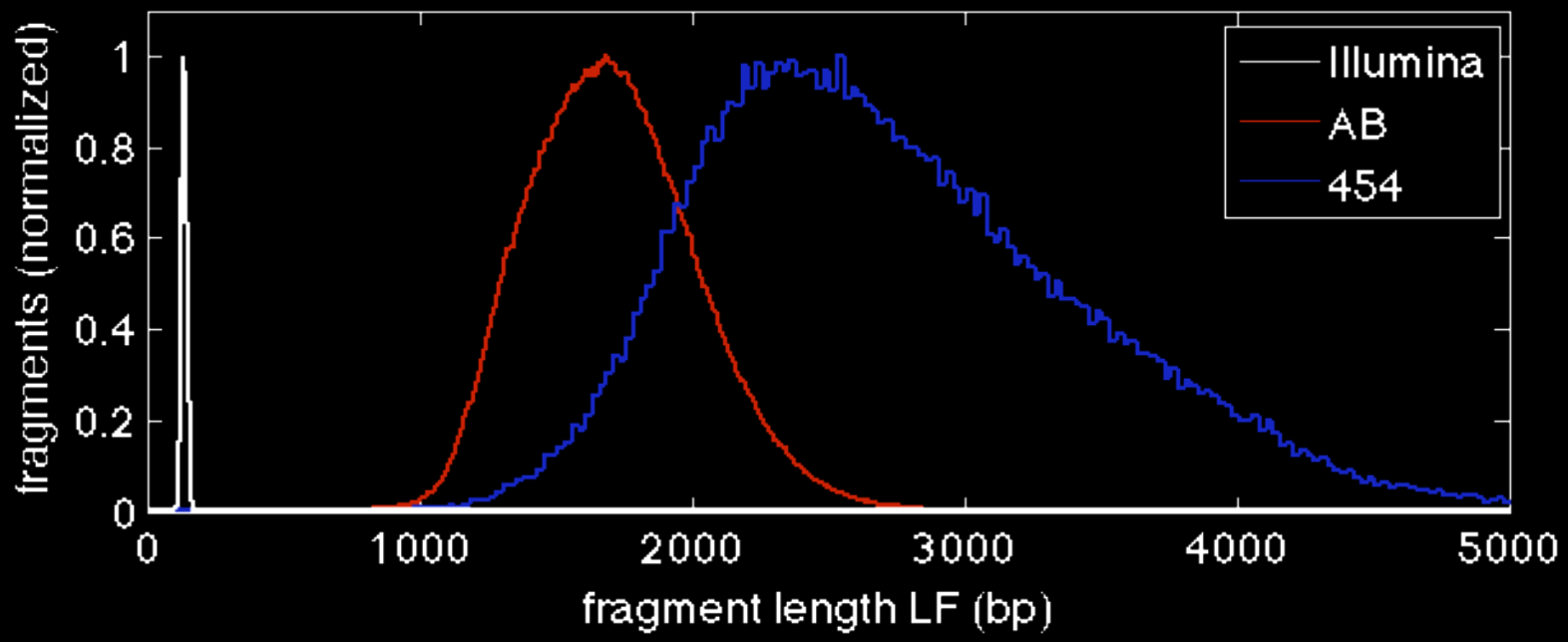
individual “NA12878”



SV detection with **PE read map positions**



Fragment length distributions



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

The SV/CNV event display

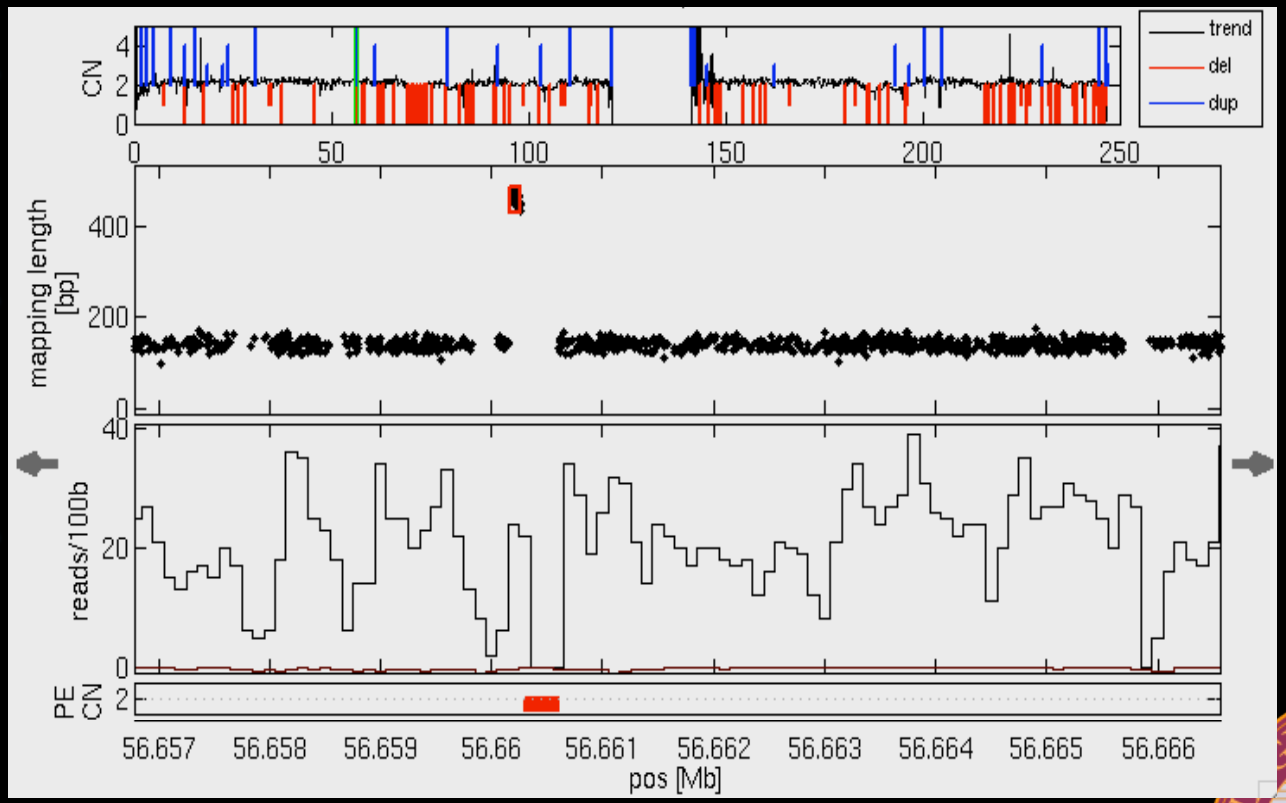
300 bp deletion in chromosome of NA12878 by Illumina paired-end data from the 1000 Genomes project

chromosome overview

fragment lengths

read depth

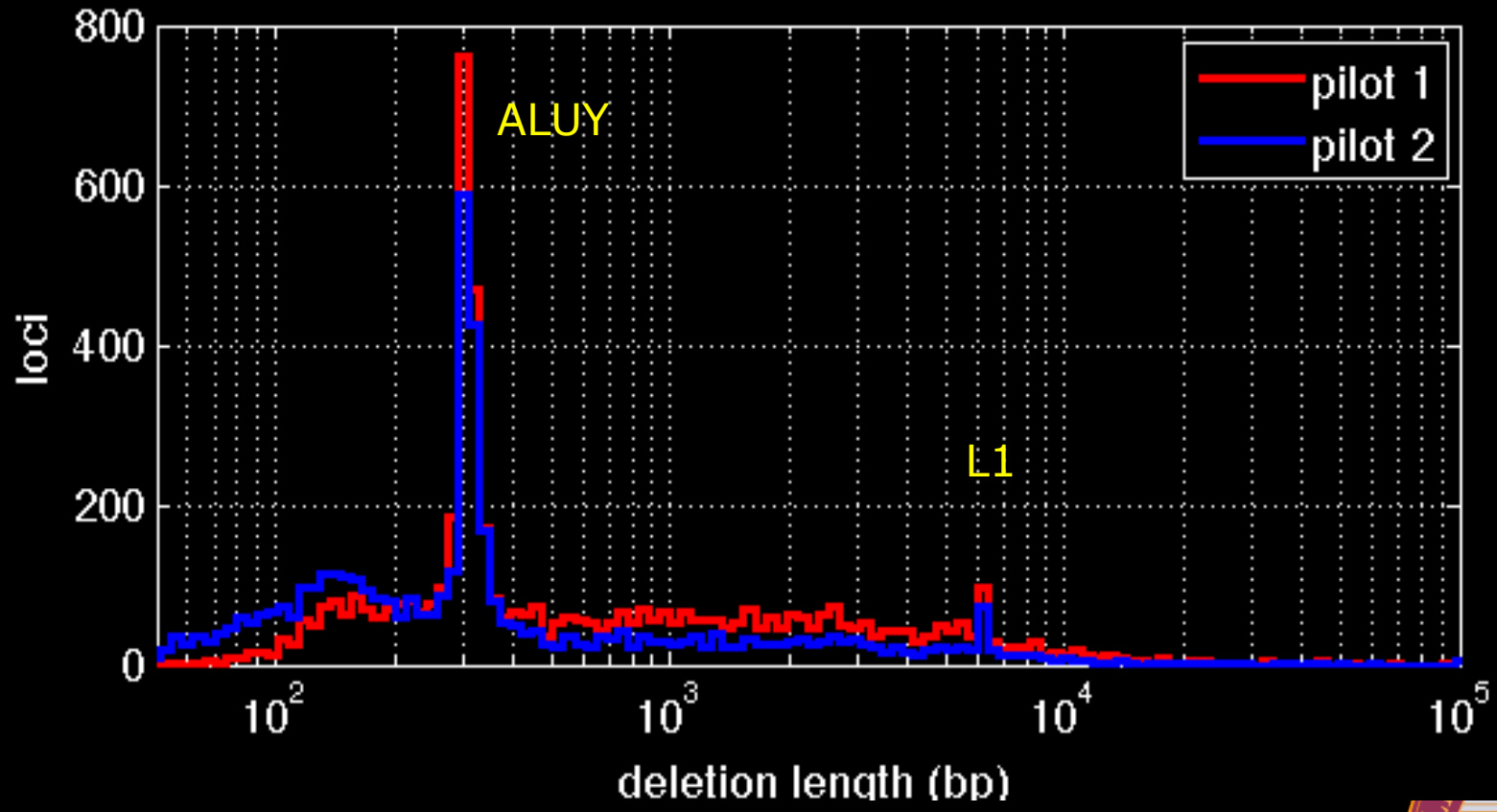
event track



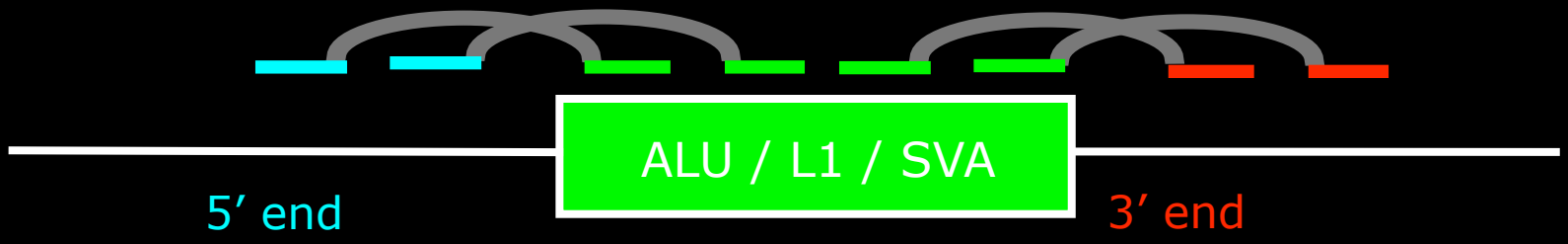
Spanner

Chip Stewart

Deletion event lengths

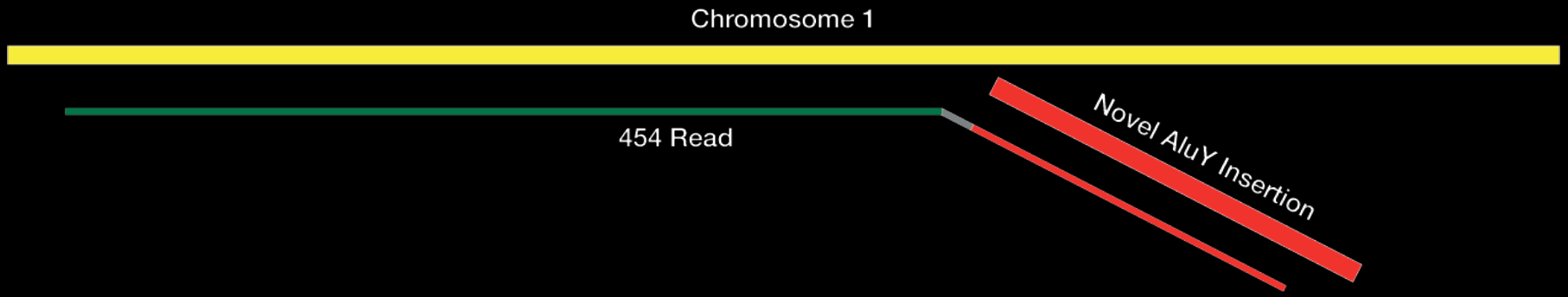


Mobile element insertions: **PE reads**



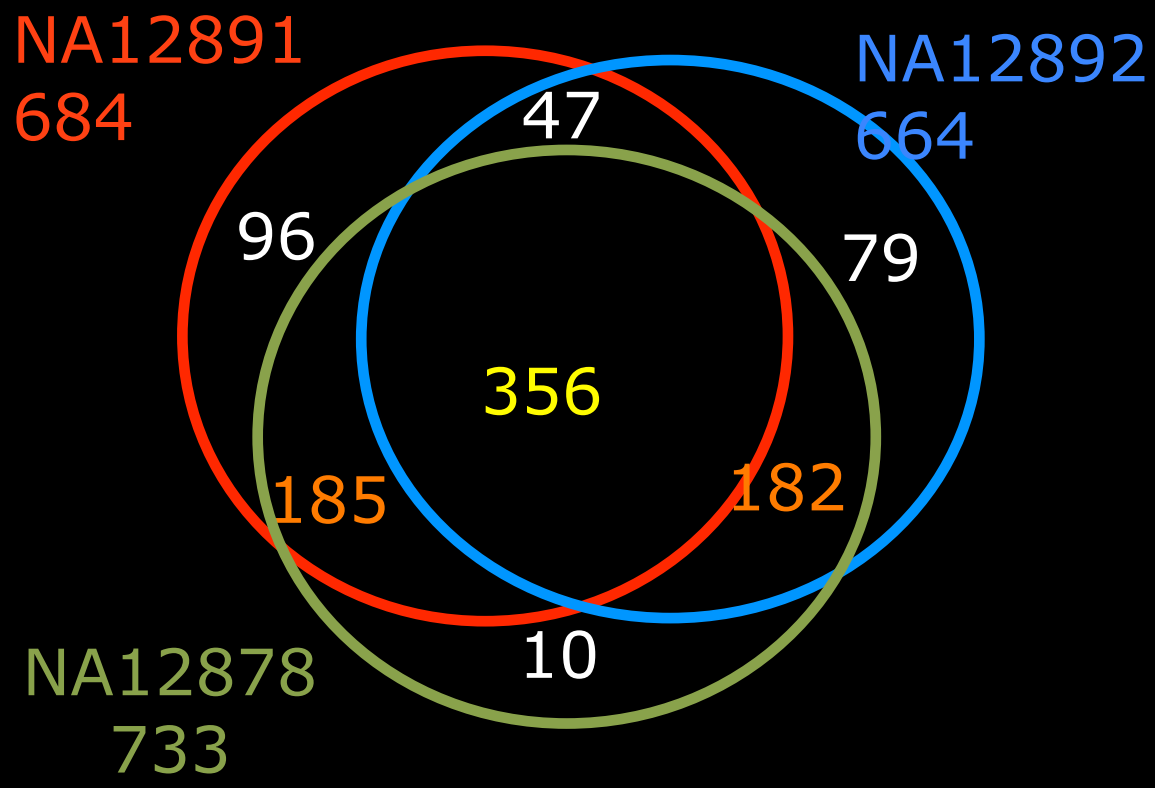
- Used with short-read data (Illumina, in our case)
- Detect clusters of 5’ & 3’ read pairs with one end mapping to a mobile element
- Clusters far from annotated elements are candidate insertion events

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

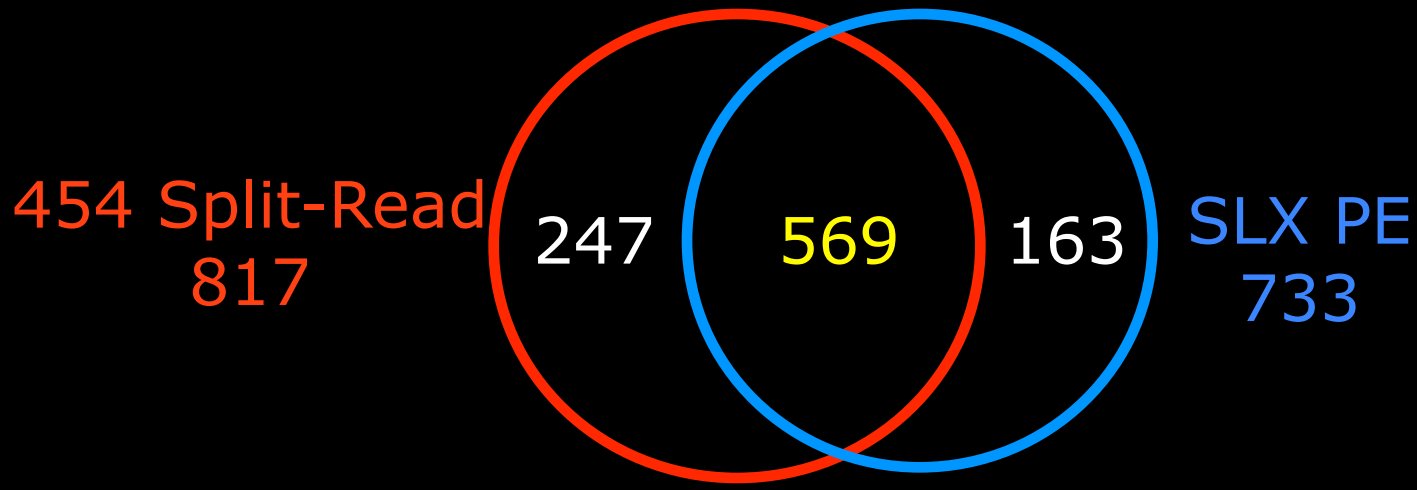
Mobile element insertions: **trio members**



ALU insertions

Detection in 1000 G Pilot 2 CEU trio PE Illumina data

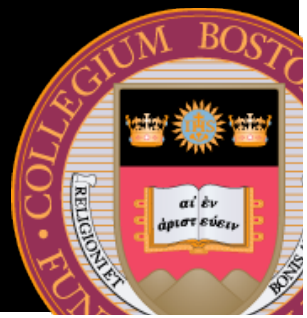
Mobile element insertions: **PE vs. Split-reads**



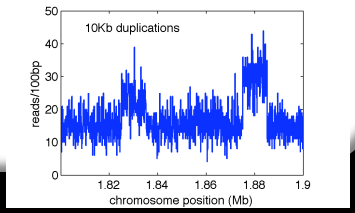
ALU insertions in NA12878

BC event lists in 1000 Genomes data

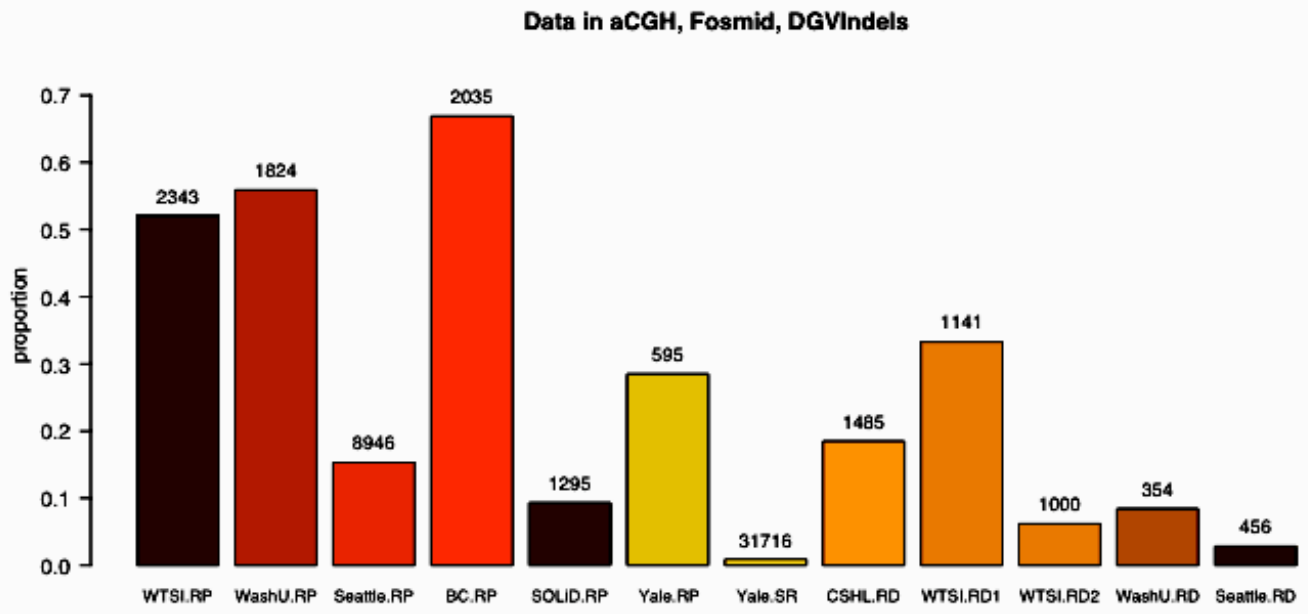
SV type	Pilot 1 140 samples low coverage	Pilot 2 6 samples high coverage
deletions	5,555	4,718
tandem duplications	540	406
mobile element insertions	3,276	2,013



SV calls / validation in 1000G datasets



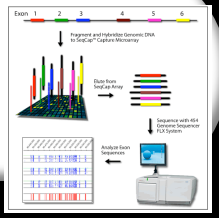
Deletions validated in either aCGH, Fosmids or DGVIndels (overlap $\geq$ 50%)

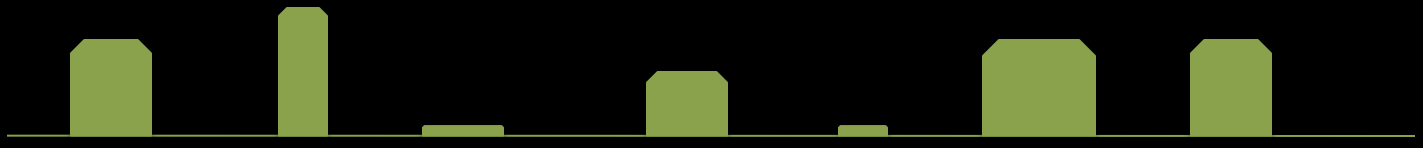


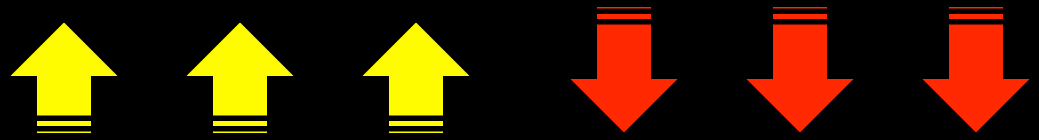
Klaudia Walter, Matt Hurles

Overlap of NA12878 Deletions with aCGH, Fosmid data and DGVIndels

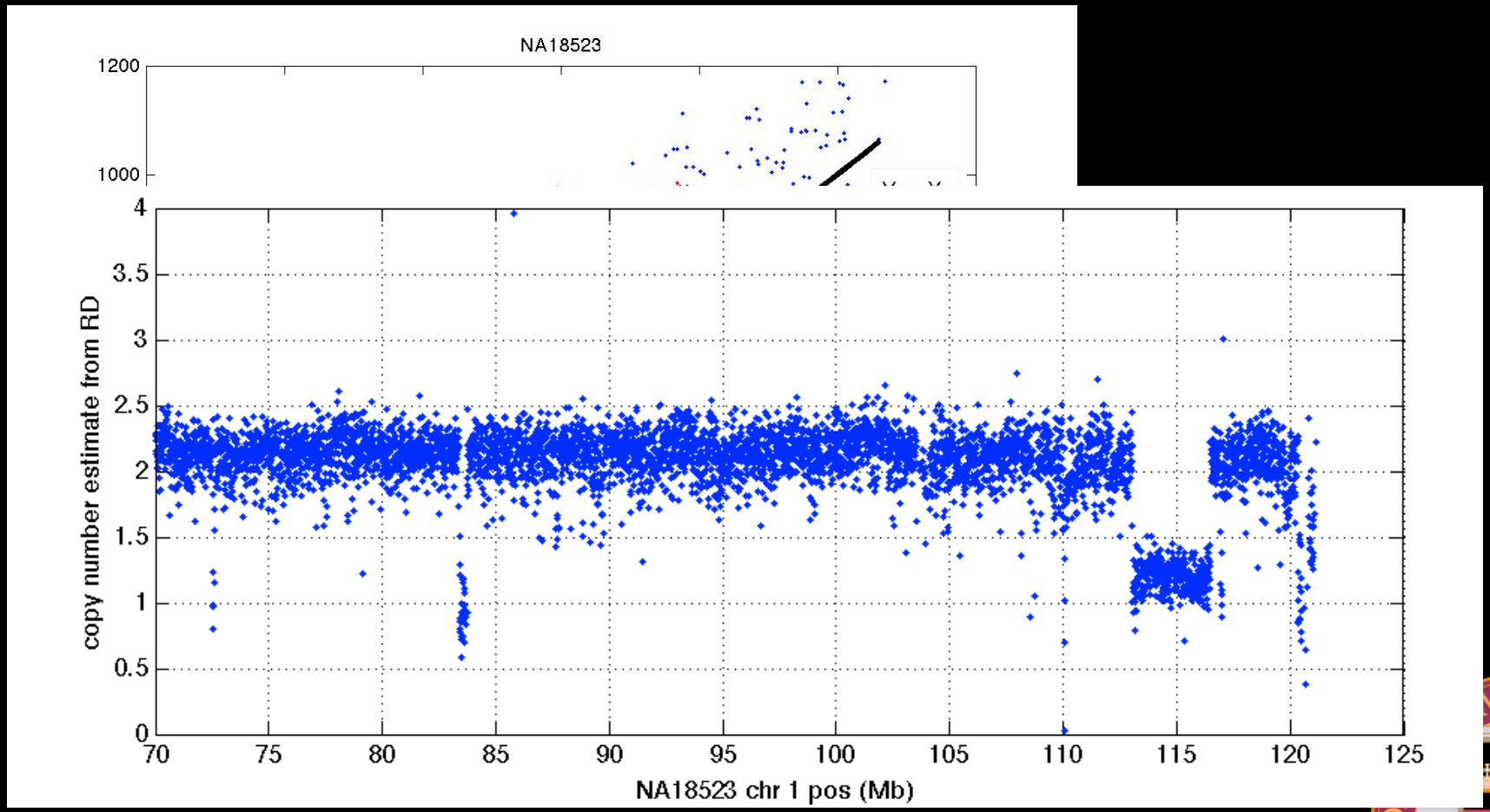
Can we find exon CNVs in Pilot 3 data?







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 / postodocs / programmers