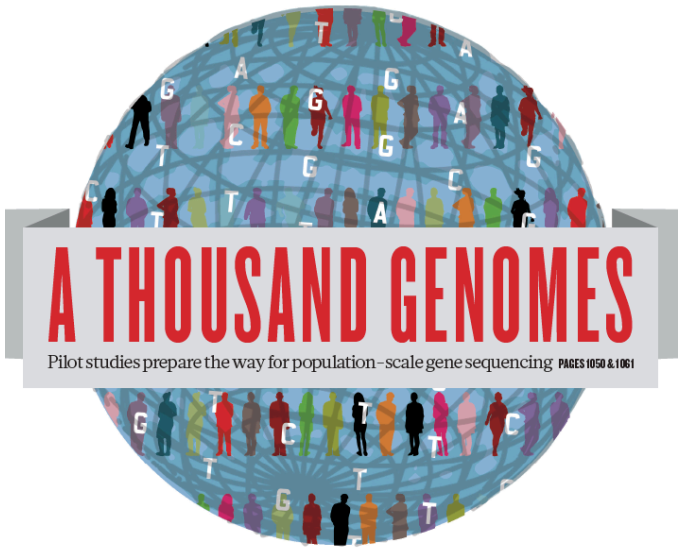


The 1000 Genomes Project Tutorial

ICHG 2011

Montreal, Quebec, Canada

October 13, 2011

- International project to construct a foundational data set for human genetics
  - Discover virtually all common human variations by investigating many genomes at the base pair level
  - Consortium with multiple centers, platforms, funders
- Aims
  - Discover population level human genetic variations of all types (95% of variation > 1% frequency)
  - Define haplotype structure in the human genome
  - Develop sequence analysis methods, tools, and other reagents that can be transferred to other sequencing projects

Agenda

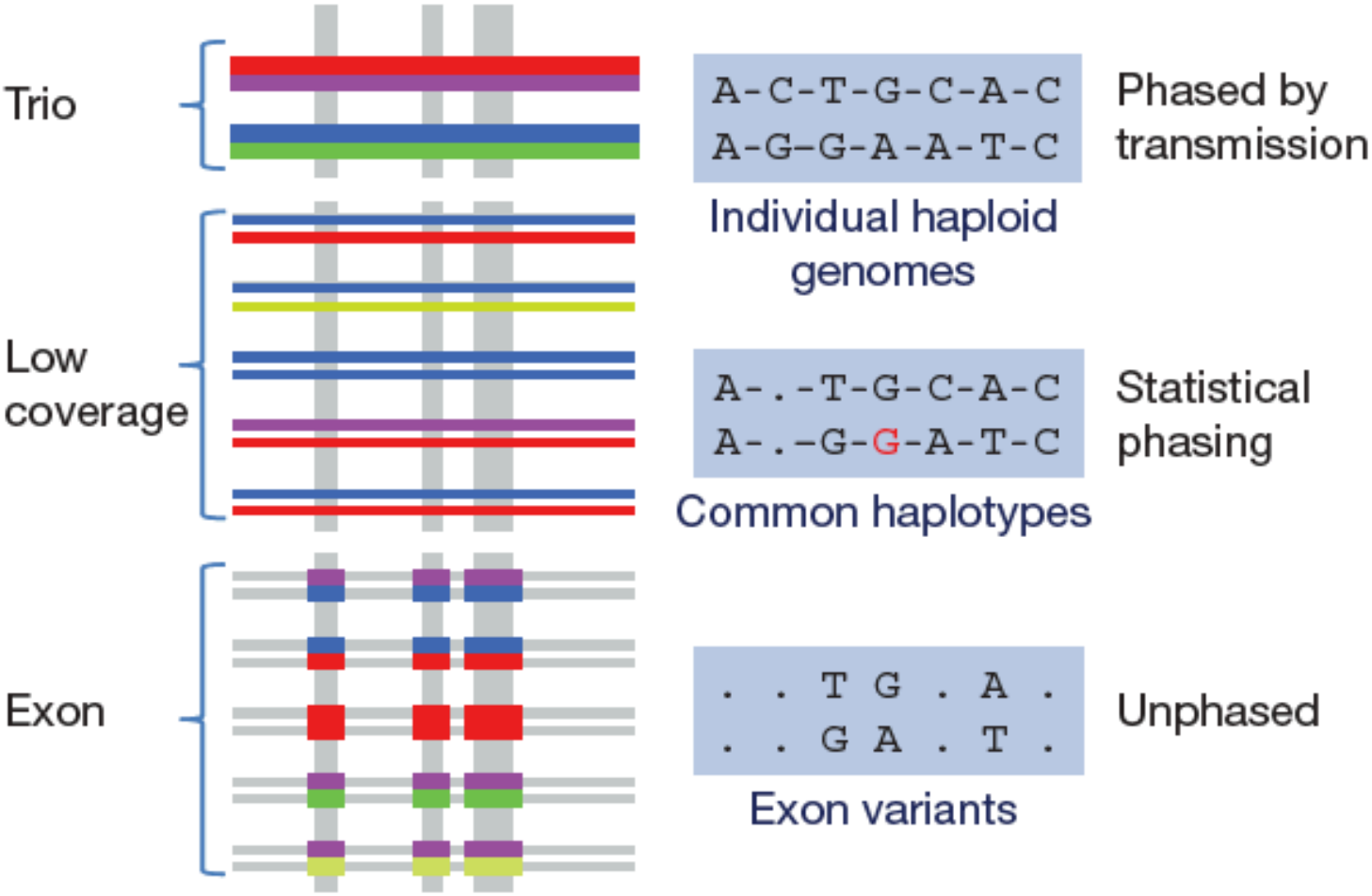
Time	Topic	Presenter	Presenter affiliation
7:30	Description of 1000 Genomes data	Gabor Marth, D.Sc.	Boston College, Boston, MA
7:55	How to access the data	Paul Flicek, D.Sc.	EMBL European Bioinformatics Inst., Hinxton, Cambridge, UK
8:20	Lessons in variant calling and genotyping	Hyun Min Kang, Ph.D.	Univ. of Michigan, Ann Arbor, MI
8:40	Structural variants	Ryan Mills, Ph.D.	Brigham and Women's Hospital, Boston, MA
9:00	Imputation in GWAS studies	Bryan Howie, Ph.D.	Univ. of Chicago, Chicago, IL
9:20	Q&A	-	-

The 1000 Genomes Project Datasets

Gabor T. Marth
Boston College Biology Department

1000 Genomes Project Tutorial
Montreal, Quebec, Canada
October 13, 2011

3 pilot coverage strategies



Pilot results published

ARTICLE

doi:10.1038/nature09534

A map of human genome variation from population-scale sequencing

The 1000 Genomes Project Consortium*

Marth et al. Genome Biology 2011, 12:R84
http://genomebiology.com/2011/12/9/R84

Genome Biology

OPEN ACCESS Freely available online

PLoS GENETICS

A Comprehensive Map of Mobile Element Insertion Polymorphisms in Humans

Chip Stewart[1☯], Deniz Kural[1☯], Michael P. Strömberg[1☯], Jerilyn A. Walker[2], Miriam K. Konkel[2], Adrian M. Stütz[3], Alexander E. Urban[4], Fabian Grubert[4], Hugo Y. K. Lam[4], Wan-Ping Lee[1], Michele Busby[1], Amit R. Indap[1], Erik Garrison[1], Chad Huff[5], Jinchuan Xing[5], Michael P. Snyder[4], Lynn B. Jorde[5], Mark A. Batzer[2], Jan O. Korbel[3], Gabor T. Marth[1]*, 1000 Genomes Project[¶]

1 Department of Biology, Boston College, Chestnut Hill, Massachusetts, United States of America, 2 Department of Biological Sciences, Louisiana State University, Baton Rouge, Louisiana, United States of America, 3 Genome Biology Unit, European Molecular Biology Laboratory, Heidelberg, Germany, 4 Department of Genetics, Stanford University, Stanford, California, United States of America, 5 Department of Human Genetics, Eccles Institute of Human Genetics, University of Utah, Salt Lake City, Utah, United States of America

PNAS

Demographic history and rare allele sharing among human populations

Simon Gravel[a], Brenna M. Henn[a], Ryan N. Gutenkunst[b], Amit R. Indap[c], Gabor T. Marth[c], Andrew G. Clark[d], Fuli Yu[e], Richard A. Gibbs[e], The 1000 Genomes Project[e], and Carlos D. Bustamante[a,1]

[a]Department of Genetics, Stanford University School of Medicine, Stanford, CA 94305-5120; [b]Department of Molecular and Cellular Biology, University of Arizona, Tucson, AZ 85721; [c]Department of Biology, Boston College, Chestnut Hill, MA 02467; [d]Department of Molecular Biology and Genetics, Cornell University, Ithaca, NY 14853; and [e]Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX 77030

Edited by Michael Lynch, Indiana University, Bloomington, IN, and approved June 3, 2011 (received for review December 24, 2010)

RESEARCH Open Access

The functional spectrum of low-frequency coding variation

Gabor T Marth[1*], Fuli Yu[2†], Amit R Indap[1†], Kiran Garimella[3†], Simon Gravel[4†], Wen Fung Leong[1†], Chris Tyler-Smith[5†], Matthew Bainbridge[2], Tom Blackwell[6], Xiangqun Zheng-Bradley[7], Yuan Chen[5], Danny Challis[2], Laura Clarke[7], Edward V Ball[8], Kristian Cibulskis[3], David N Cooper[8], Bob Fulton[9], Chris Hartl[3], Dan Koboldt[9], Donna Muzny[4], Richard Smith[7], Carrie Sougnez[3], Chip Stewart[1], Alistair Ward[1], Jin Yu[2], Yali Xue[5], David Altshuler[3], Carlos D Bustamante[4], Andrew G Clark[10], Mark Daly[3], Mark DePristo[3], Paul Flicek[7], Stacey Gabriel[3], Elaine Mardis[9], Aarno Palotie[5], Richard Gibbs[2] and the 1000 Genomes Project

ARTICLE

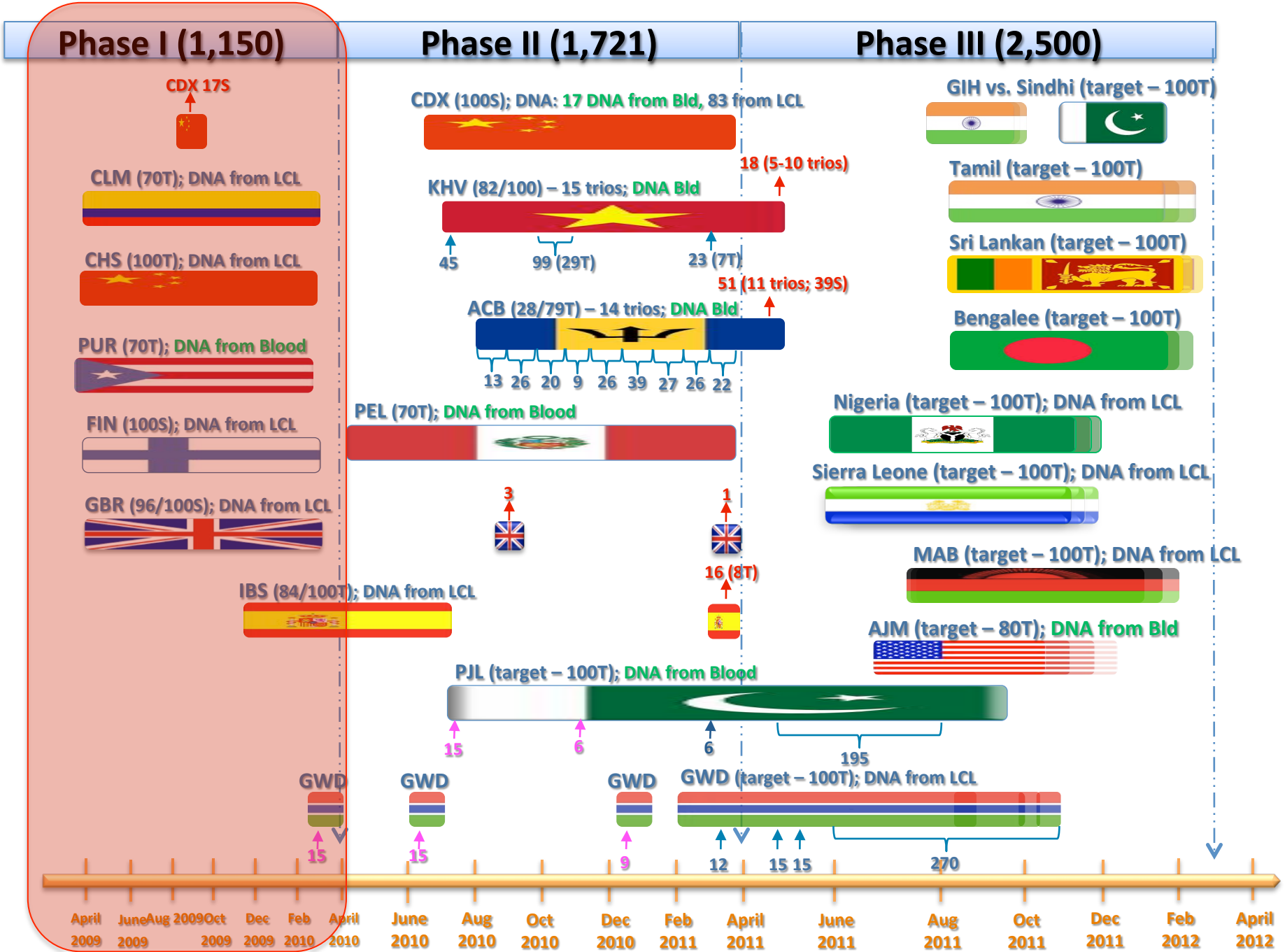
doi:10.1038/nature09708

Mapping copy number variation by population-scale genome sequencing

Ryan E. Mills[1*], Klaudia Walter[2*], Chip Stewart[3*], Robert E. Handsaker[4*], Ken Chen[5*], Can Alkan[6,7*], Alexej Abyzov[8*], Seungtai Chris Yoon[9*], Kai Ye[10*], R. Keira Cheetham[11], Asif Chinwalla[5], Donald F. Conrad[2], Yutao Fu[12], Fabian Grubert[13], Iman Hajirasouliha[14], Fereydoun Hormozdiari[14], Lilia M. Iakoucheva[15], Zamin Iqbal[16], Shuli Kang[15], Jeffrey M. Kidd[6], Miriam K. Konkel[17], Joshua Korn[4], Ekta Khurana[8,18], Deniz Kural[3], Hugo Y. K. Lam[13], Jing Leng[8], Ruiqiang Li[19], Yingrui Li[19], Chang-Yun Lin[20], Ruibang Luo[19], Xinmeng Jasmine Mu[8], James Nemesh[4], Heather E. Peckham[12], Tobias Rausch[21], Aylwyn Scally[2], Xinghua Shi[1], Michael P. Stromberg[3], Adrian M. Stütz[21], Alexander Eckehart Urban[13,27], Jerilyn A. Walker[17], Jiantao Wu[3], Yujun Zhang[2], Zhengdong D. Zhang[8], Mark A. Batzer[17], Li Ding[5,22], Gabor T. Marth[3], Gil McVean[23], Jonathan Sebat[15], Michael Snyder[13], Jun Wang[19,24], Kenny Ye[20], Evan E. Eichler[6,7], Mark B. Gerstein[8,18,25], Matthew E. Hurles[2], Charles Lee[1], Steven A. McCarroll[4,26], Jan O. Korbel[21] & 1000 Genomes Project†

Finalized project design

- Based on the result of the pilot project, we decided to collect data on 2,500 samples from 5 continental groupings
  - Whole-genome low coverage data (>4x)
  - Full exome data at deep coverage (>50x)
  - A number of deep coverage genomes to be sequenced, with details to be decided
  - Hi-density genotyping at subsets of sites
- Moved from the Pilot into Phase 1 of the project


Phase I (1,150)
Phase II (1,721)
Phase III (2,500)
CDX 17S
CLM (70T); DNA from LCL
CHS (100T); DNA from LCL
PUR (70T); DNA from Blood
FIN (100S); DNA from LCL
GBR (96/100S); DNA from LCL
IBS (84/100T); DNA from LCL
CDX (100S); DNA: 17 DNA from Bld, 83 from LCL
18 (5-10 trios)
KHV (82/100) – 15 trios; DNA Bld
45
99 (29T)
23 (7T)
51 (11 trios; 39S)
ACB (28/79T) – 14 trios; DNA Bld
13 26 20 9 26 39 27 26 22
PEL (70T); DNA from Blood
3
1
16 (8T)
PJL (target – 100T); DNA from Blood
15
6
6
195
GWD
15
GWD
15
GWD
9
GWD (target – 100T); DNA from LCL
12
15 15
270
GIH vs. Sindhi (target – 100T)
Tamil (target – 100T)
Sri Lankan (target – 100T)
Bengalee (target – 100T)
Nigeria (target – 100T); DNA from LCL
Sierra Leone (target – 100T); DNA from LCL
MAB (target – 100T); DNA from LCL
AJM (target – 80T); DNA from Bld
April 2009
June 2009
Aug 2009
Oct 2009
Dec 2009
Feb 2010
April 2010
June 2010
Aug 2010
Oct 2010
Dec 2010
Feb 2011
April 2011
June 2011
Aug 2011
Oct 2011
Dec 2011
Feb 2012
April 2012

Phase I data

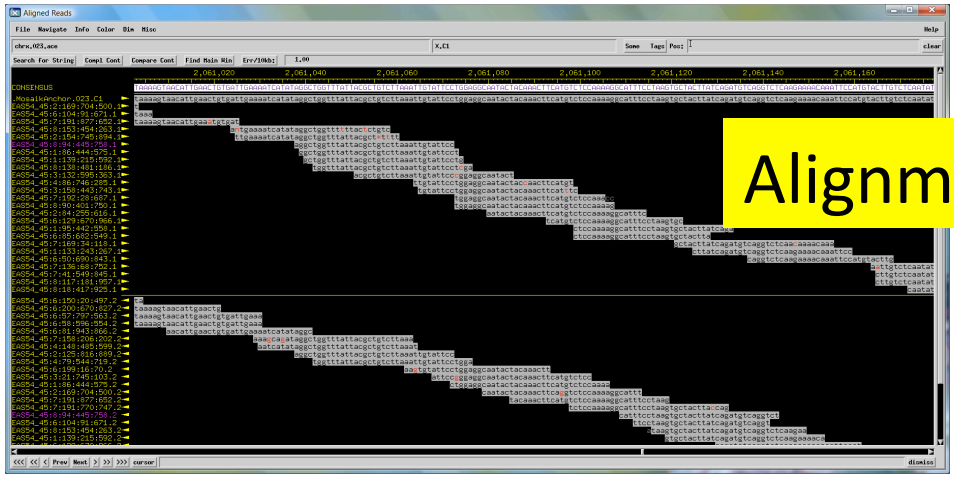
- Samples from 14 populations: ASW, CEU, CHB, CHS, CLM, FIN, GBR, IBS, JPT, LWK, MXL, PRU, TSI, YRI

Dataset	Low coverage whole genome	Deep coverage whole exome
# samples	1,094	1,128
Sequencing technologies	Illumina, SOLiD, 454	Illumina, SOLiD
Primary alignments (BAMs)	BWA, BFAST	MOSAIK, BFAST
Second alignments (BAMs)	MOSAIK	BWA, MOSAIK
Read coverage	4-8X per sample	≥70% of targets with ≥20X coverage in every sample

Raw data & read alignment delivery

```
@IL11_266:1:1:395:231/1
CCAACCACAACACAAAAAACACAAGCAAC
+
@@AAAAA?<>@@>?:475;A6?384,>5
@IL11_266:1:1:399:301/1
CAAAAAAAAAAAGAAGTACGAGATACGACACATCAC
+
;@AAAA>5;>@C67'&2?&7<&7&@1/1408=19::
```

Reads: FASTQ

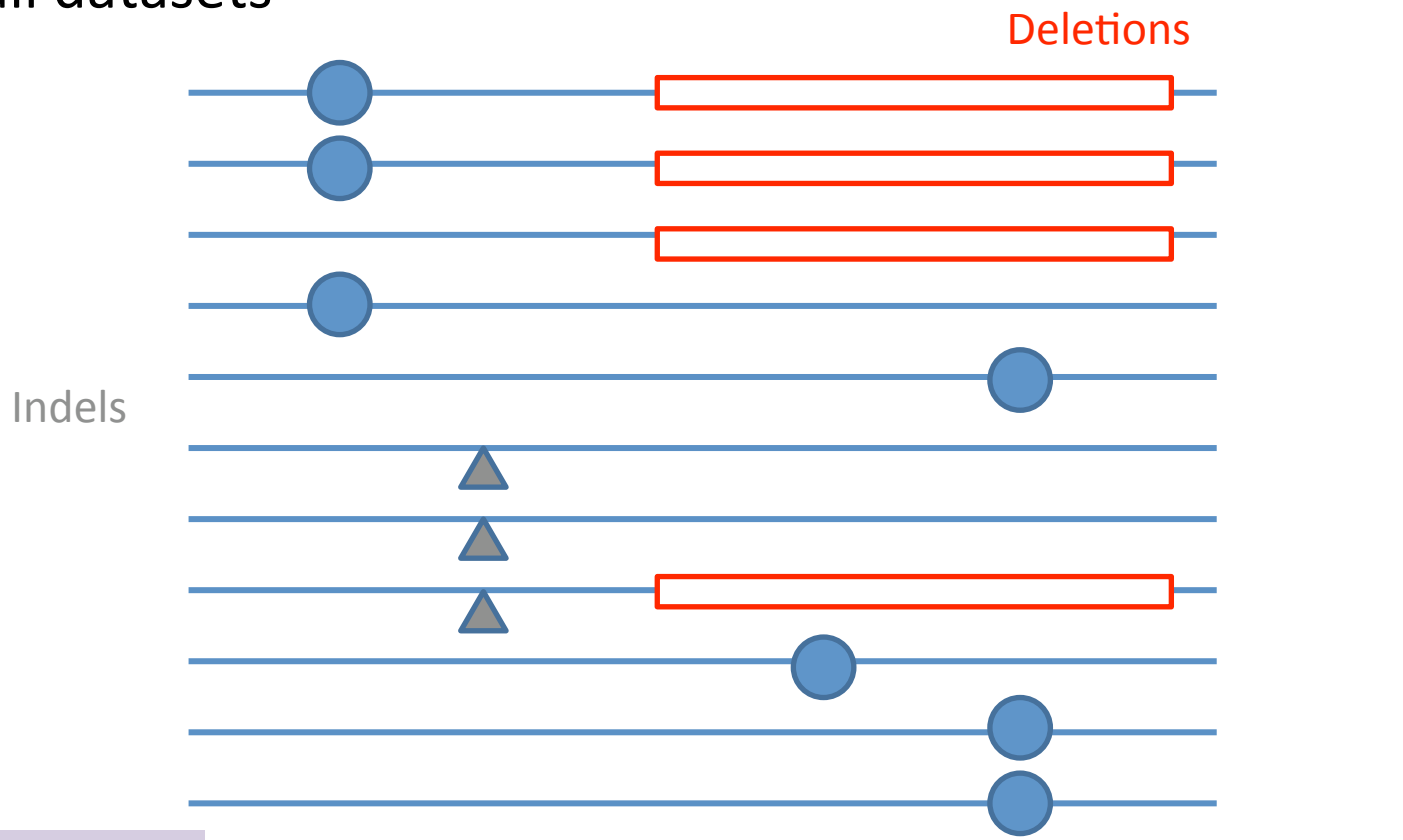


Alignments: BAM

ftp://ftp.1000genomes.ebi.ac.uk

Phase 1 analysis goal: an **integrated view of human variations**

- Reconstruct haplotypes including all variant types, using all datasets



Goncalo Abecasis

Pipelines for data processing and variant calling

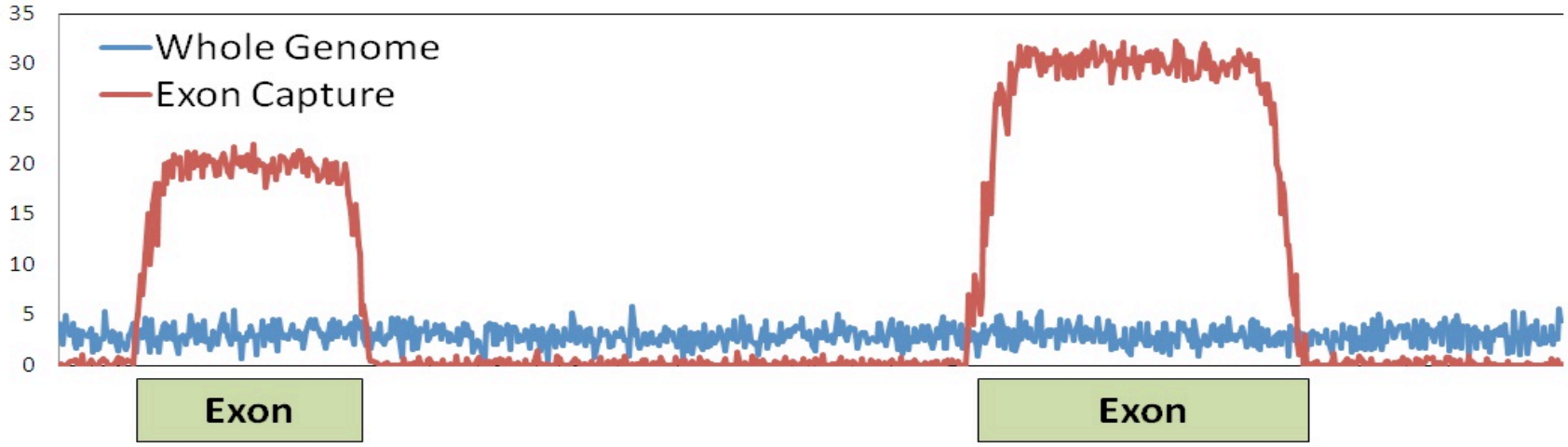
- Tens of analysis groups have contributed
- Individual pipelines and component tools vary
- Typical main steps:
  - Read mapping
  - Duplicate filtering
  - Base quality value recalibration
  - INDEL realignment
  - Variant calling (sites)
  - Sample genotype calling (sometime part of variant calling)
  - Variant filtering / call set refinement
  - Variant reporting

SNPs

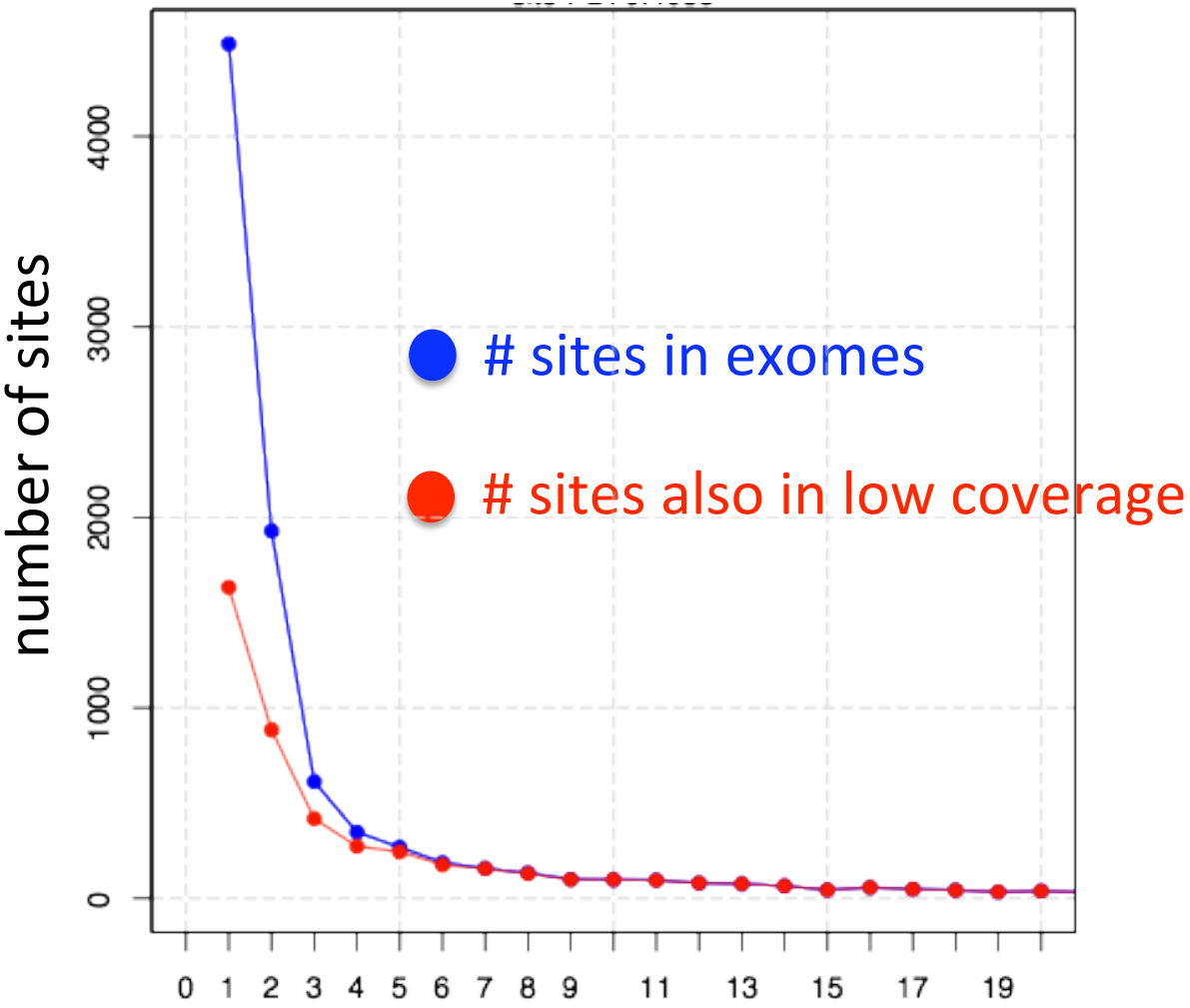

97490
97500
97510
975
TGCGTCTCGTATCATATTTTTCAGGACATCATCTATCC
:gcgtctcgtatcatattttt
:gcgtctcgtatcatatttttc
:gcgtctcgtatcatagttttca
:gcgtctcgtatcatatttttcag
:gcgtctcgtatcatagttttcagg
:gcgtctcgtatcatatttttcagga
:gcgtctcgtatcatatttttcaggacat
:gcgtctcgtatcatatttttcaggaca
:gcgtctcgtatcatatttttcaggaca
:gcgtctcgtatcatatttttcaggaca
:gcgtctcgtatcatagttttcaggacatca
gtatcatatttttcaggacatcatctatcg

SNP calls

Dataset	Contributing datasets	Consensus method	#SNPs	# Novel SNPs	Novel Ts/Tv	%ONMI poly (sensitivity)	%OMNI mono (FDR)
Low coverage	BC, BCM, BI, NCBI, UM	VQSR	**37.9M**	**29.65M**	2.16	98.4	1.80
Exome/ Illumina	BC, BCM, BI, Cornell, UM	SVM	**598K**	**468K**	2.74	98.01	1.97
Exome/SOLiD	BC, BCM, UM	SVM	**356K**	**243K**	2.91	90.67	1.29



Deep coverage exome data is more sensitive to low-frequency variants

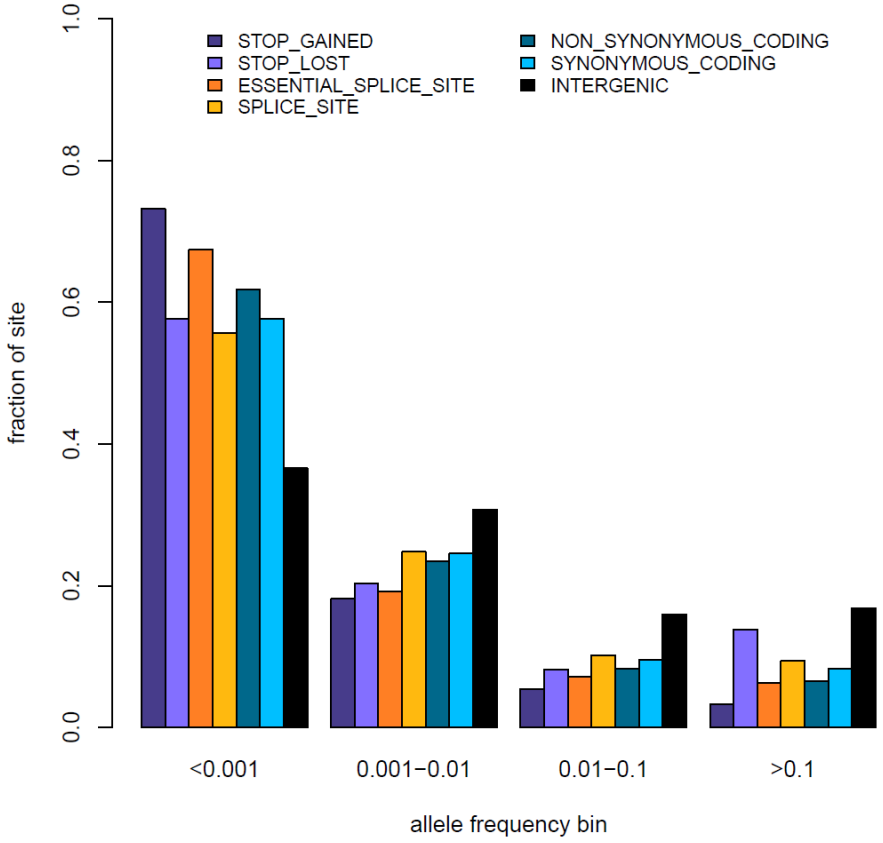


Erik Garrison

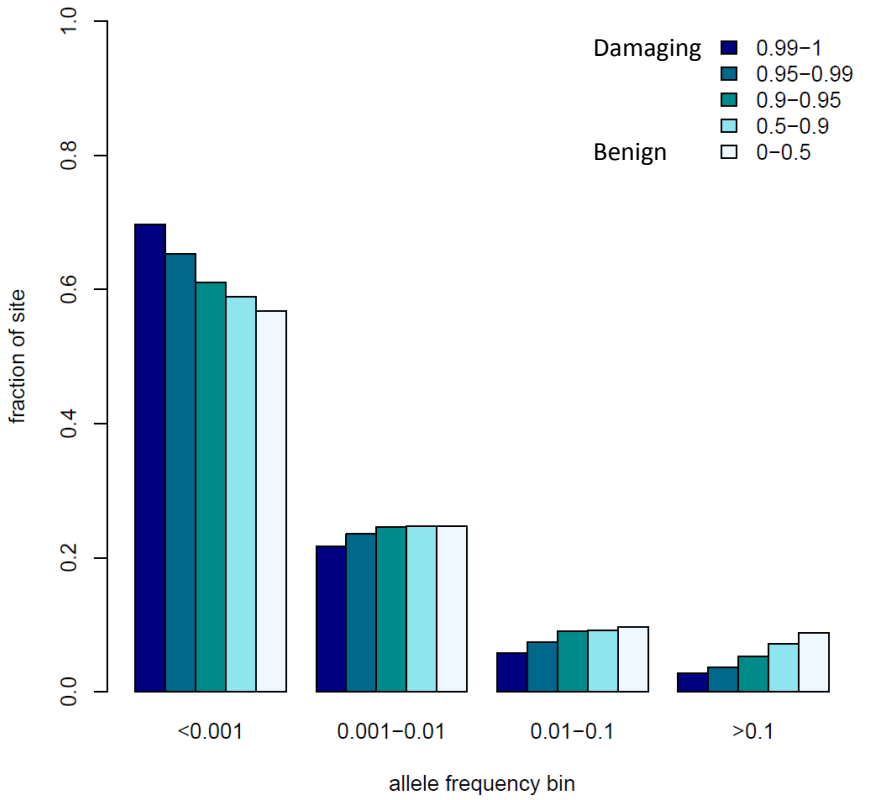
Allele count in 766 exomes (chr. 20, exons only)

Newly discovered SNPs are mostly at low frequency and enriched for functional variants

Functional category



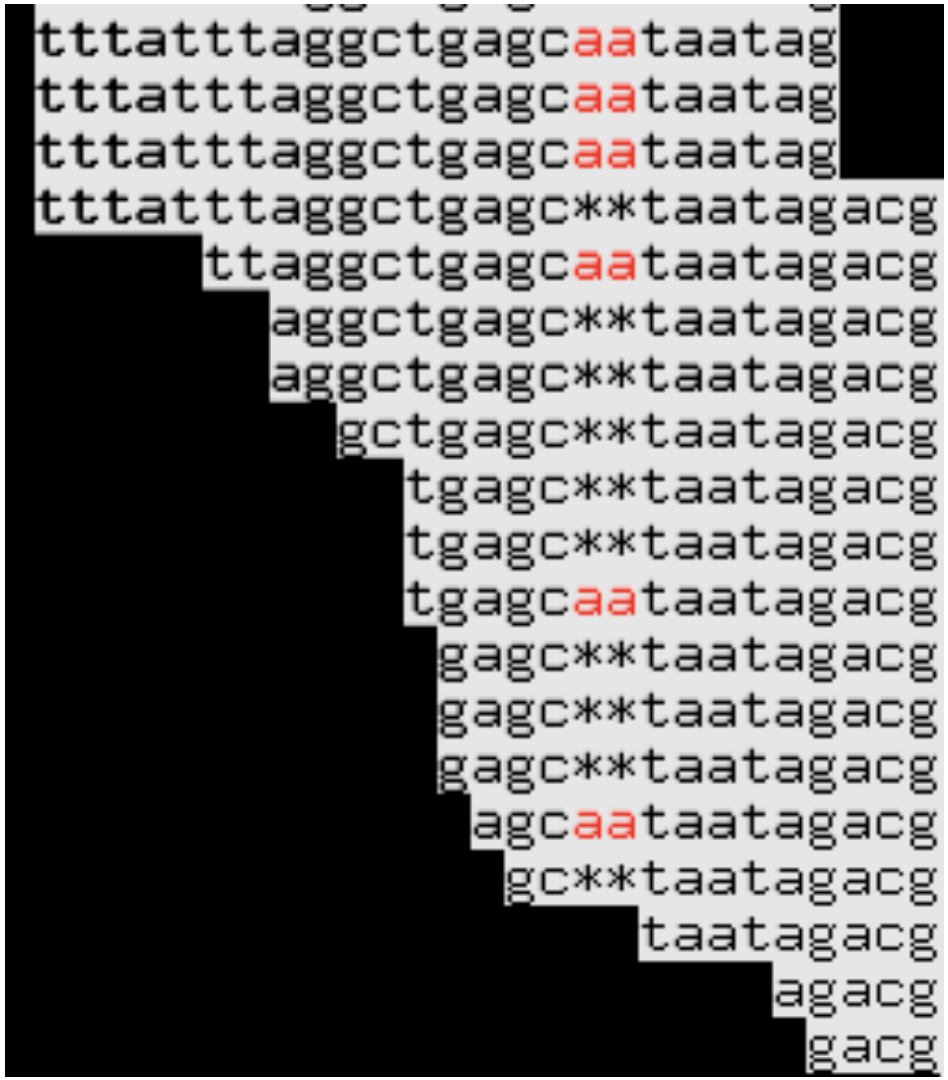
Non-synonymous: Condel score



Presentation on using the data for GWAS by Brian Howie

Enza Colonna, Yuan Chen, Yali Xue

INDELs



GATTAGAATCGCAATTAAA(
GATTAGAAT*GCAATTAAA(

AGTTTCTCT***TTCTTACAG
AGTTTCTCTGCTTTCTTACAG

CGAATTAGA******GCAAA
CGAATTAGACTTAGAGCAAA

TCTCAAAAAAAAAAAAAAAAAAGTGT
TCTCAAAAAAAAAAAAAAAAA*GTGT

)AAA**********************GA
)AAAAAAAAAAAAAAAAAAAAAAAAAGA

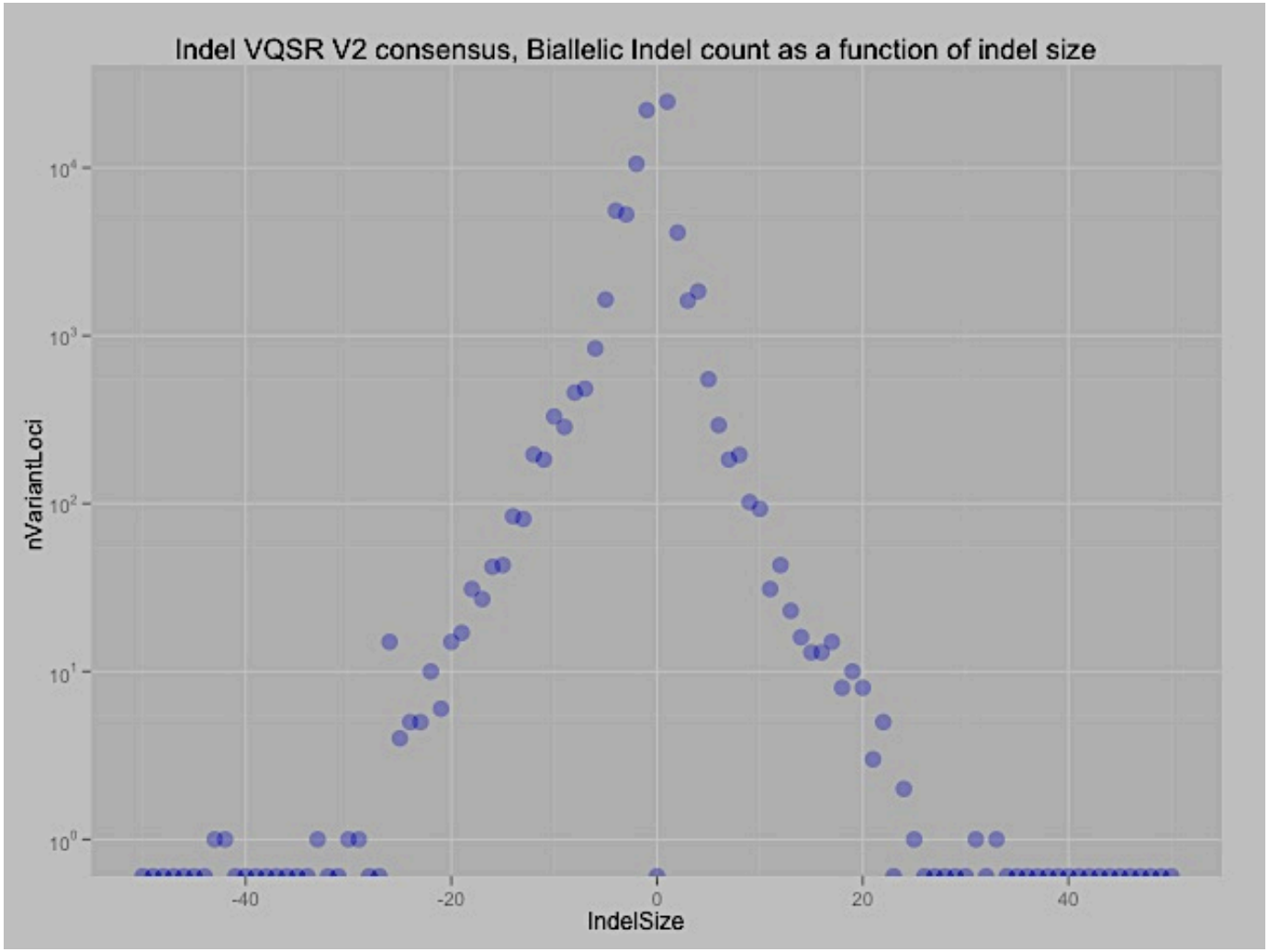
TGTGTGTGTGTGTGTGTGTATTTAAAAAACTAGG
TGTGTGTGTGTGTGTG**TATTTAAAAAACTAGG

INDEL calls

Dataset	Contributing datasets	Consensus method	#INDELs
Low coverage	BC, BI, DI, OX, SI	VQSR	**5.5M**
Exome/Illumina	BC, BCM, BI	N.A.	**6.5 – 10.2K**
Exome/SOLiD	BCM	N.A.	**4.2 – 5.0K**

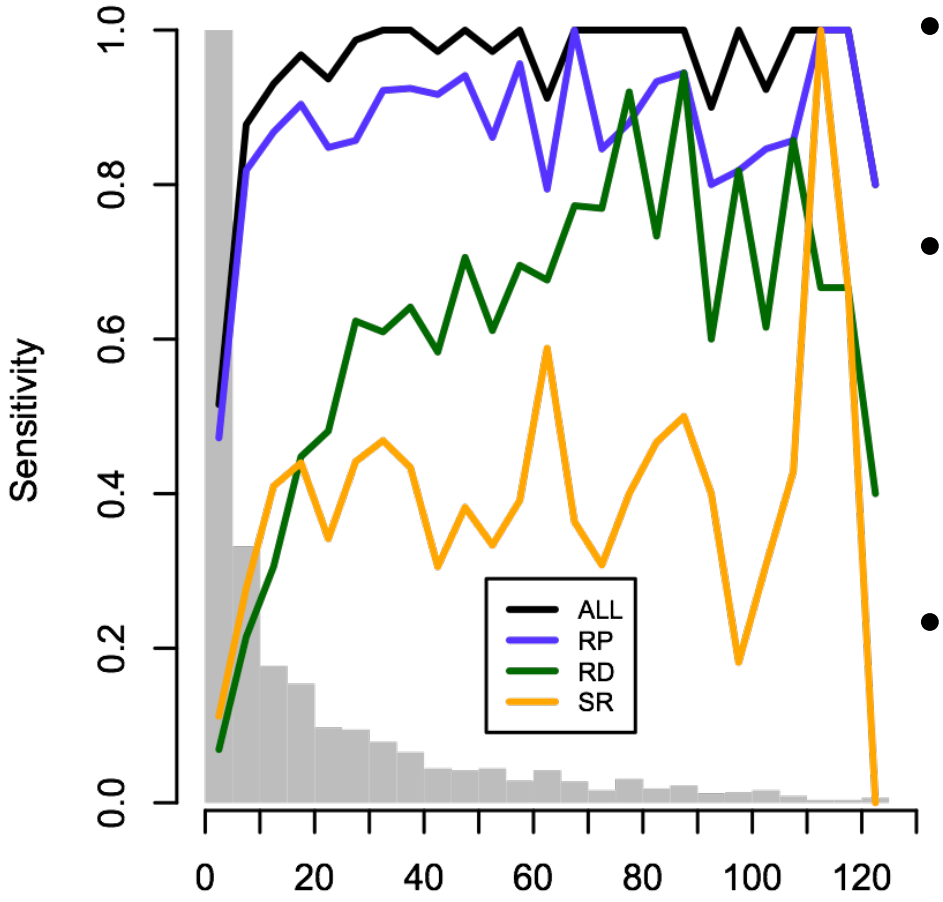
Guillermo Angel

INDEL length



Guillermo Angel

Finding structural variants



- Discovery with a number of different methods
- Several types (e.g. deletions, tandem duplications, mobile element insertions) now detectable with high accuracy
- We are pulling in new types for the Phase I data (inversions, *de novo* insertions, translocations)

Presentation on structural variations by Ryan Mills

SNP validations (low coverage data)

	Total	Polymorphic	Monomorphic	No Call	Confirmation Rate	Failure Rate
All Sites	**300**	**282**	**12**	**6**	**0.959**	**0.020**
Called in Validation Samples	**287**	**276**	**5**	**6**	**0.982**	**0.021**
Singletons	70	65	3	2	0.956	0.029
MAF<0.01*	134	131	2	1	0.985	0.007
0.01<MAF<0.05	33	33	0	0	1.000	0.000
MAF>0.05	50	47	0	3	1.000	0.060

*Excludes singletons

Danny Challis, Eric Banks

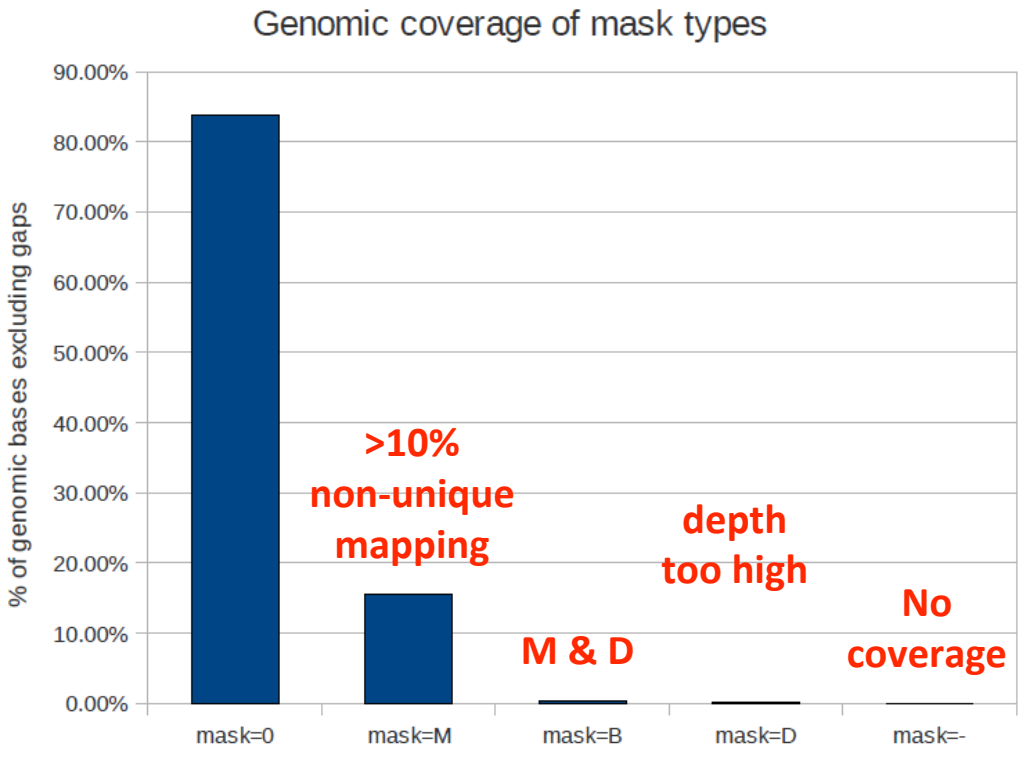
Genotypes are accurate

- Average low coverage depth is ~5x
- We obtain genotypes by sharing data between samples (using imputation-related methods)

Genotype	HomRef	Het	HomAlt	Overall
Error rate	0.16%	0.76%	0.39%	0.37%

- Genotypes are expected to be even more accurate after integration of multiple variant sources

Accessible fraction of genome



- In the Pilot data, we found that >80% of the human genome reference was accessible for SNP variant calling
- We are currently re-evaluating this fraction for the Phase 1 data (which used longer reads)
- We are developing methods to estimate the fraction for other variants (especially INDELs)

Goncalo Abecasis

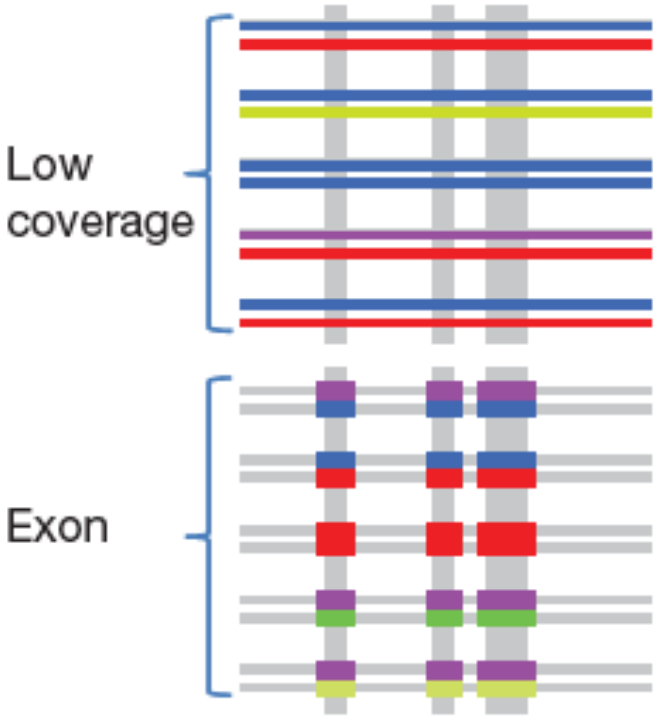
Variant call delivery

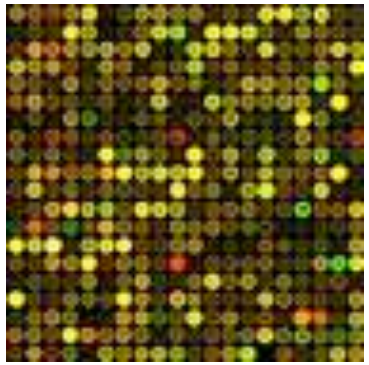
Format: VCF

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	NA00001	NA00002
20	14370	rs6054257	G	A	29	0	NS=3;DP=14;AF=0.5;DB;H2	GT:GQ:DP:HQ	0|0:48:1:51,51	1|0:48:8:51,51
20	17330	.	T	A	3	q10	NS=3;DP=11;AF=0.017	GT:GQ:DP:HQ	0|0:49:3:58,50	0|1:3:5:65,3
20	1110696	rs6040355	A	G,T	67	0	NS=2;DP=10;AF=0.333,0.667;AA=T;DB	GT:GQ:DP:HQ	1|2:21:6:23,27	2|1:2:0:18,2
20	1230237	.	T	.	47	0	NS=3;DP=13;AA=T	GT:GQ:DP:HQ	0|0:54:7:56,60	0|0:48:4:51,51
20	1234567	microsat1	G	D4,IGA	50	0	NS=3;DP=9;AA=G	GT:GQ:DP	0/1:35:4	0/2:17:2

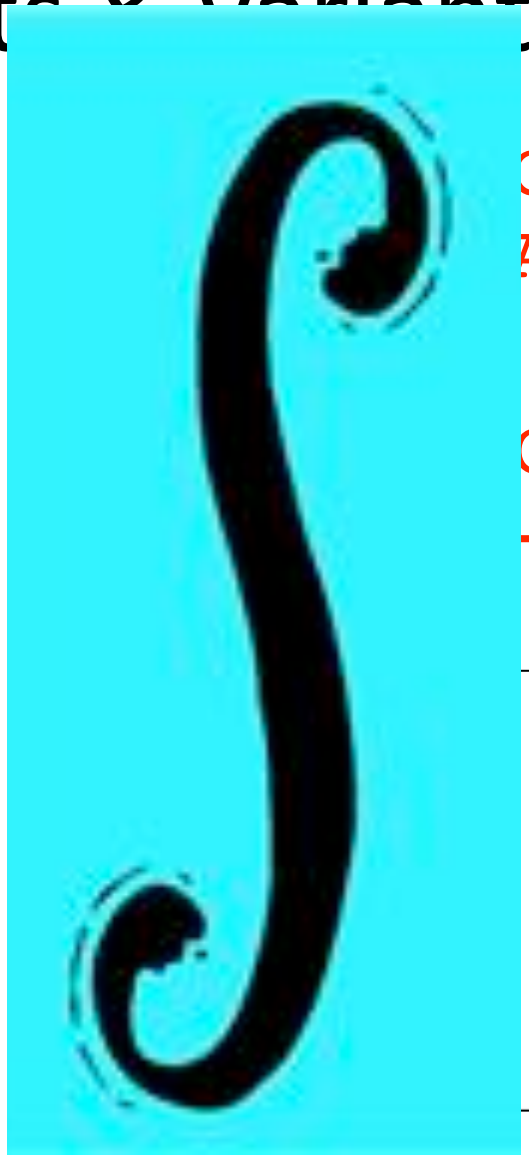
ftp://ftp.1000genomes.ebi.ac.uk

Datasets & variant types





SNP array data

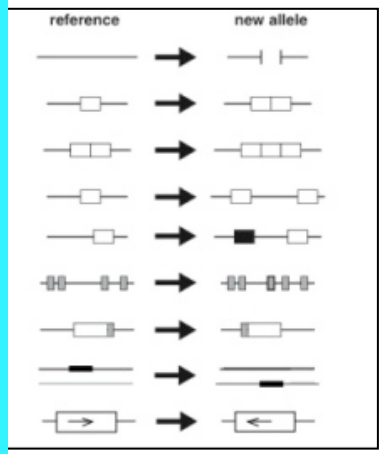


CTGAG
ATGAG

SNP

CCTGAG
--TGAG

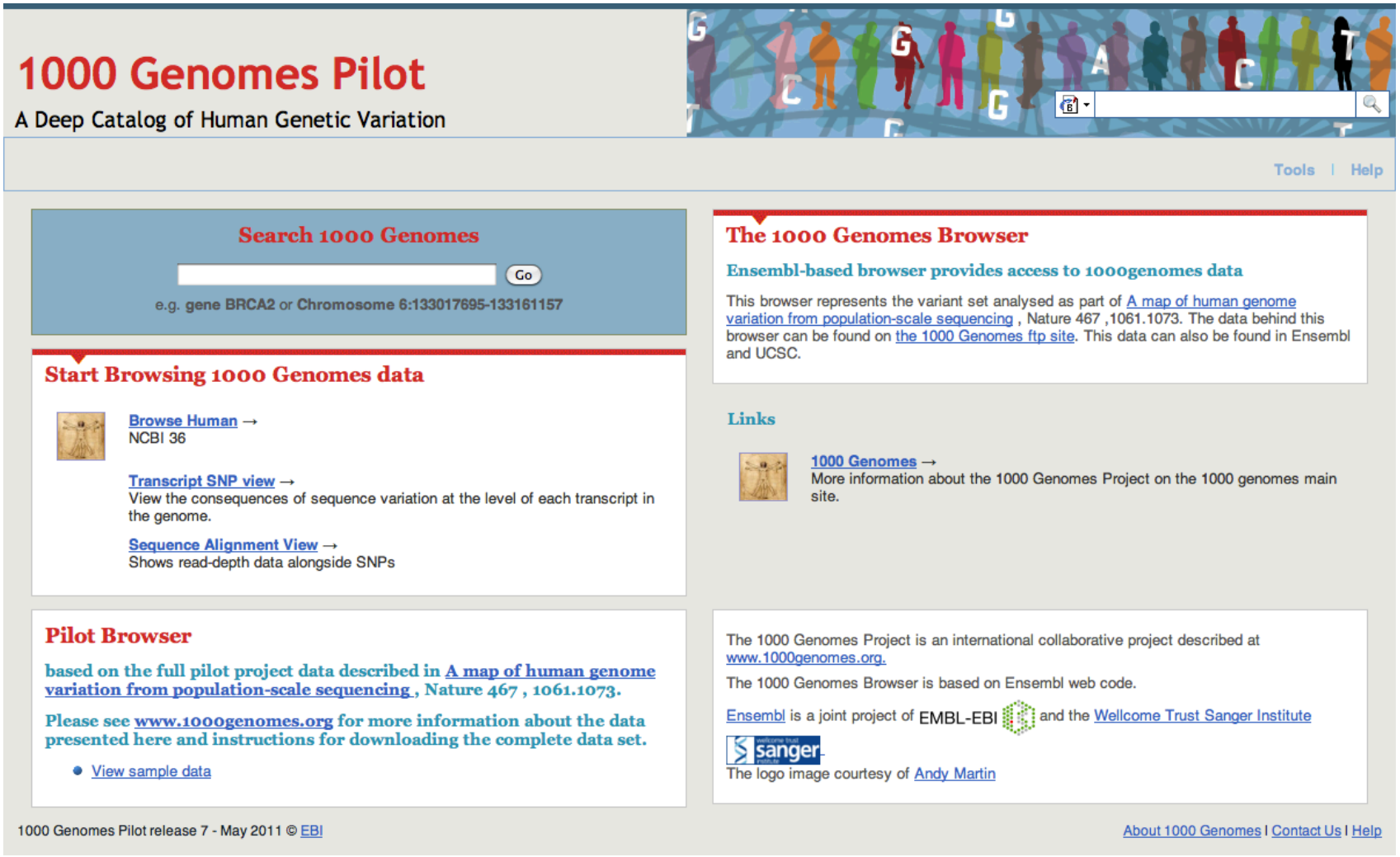
INDEL



SV

Presentation on integration by Hyun Min Kang

Data delivery

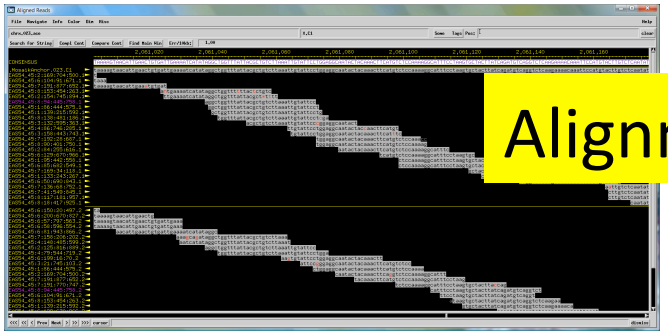


Presentation on data access by Paul Flicek

The 1000GP is a driver for method and tool development

- New data formats (SAM/BAM, VCF) developed by the 1000GP are now adopted by the entire genomics community
- Tools (read mappers e.g. BWA, MOSAIK, etc; variant callers including those for SVs)
- Data processing protocols (BQ recalibration, duplicate read removal, etc.)
- Imputation and haplotype phasing methods

Tools for analyzing & manipulating 1000G data



Alignments: SAM/BAM

- samtools: http://samtools.sourceforge.net/
- BamTools: http://sourceforge.net/projects/bamtools/
- GATK: http://www.broadinstitute.org/gsa/wiki/index.php/The_Genome_Analysis_Toolkit

```
#CHROM POS      ID         REF ALT     QUAL FILTER INFO                              FORMAT      NA00001        NA00002
20     14370    rs6054257  G   A       29   0      NS=3;DP=14;AF=0.5;DB;H2           GT:GQ:DP:HQ 0|0:48:1:51,51 1|0:48:8:51,51
20     17330    .          T   A       3    q10    NS=3;DP=11;AF=0.017               GT:GQ:DP:HQ 0|0:49:3:
20     1110696  rs6040355  A   G,T     67   0      NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:
20     1230237  .          T   .       47   0      NS=3;DP=13;AA=T                   GT:GQ:DP:HQ 0|0:54:7:
20     1234567  microsat1  G   D4,IGA  50   0      NS=3;DP=9;AA=G                    GT:GQ:DP    0/1:35:4
```

Variants: VCF

- VCFTools: http://vcftools.sourceforge.net/
- VcfCTools: https://github.com/AlistairNWard/vcfCTools

Project timeframe (approximate)

- Phase 1
  - Raw data, alignments available
  - Integrated variant set available
  - Phase 1 analysis paper by end of 2011
- Phase 2
  - Raw data mid-December 2011
  - Read mapping, variant calling early 2012
- Phase 3
  - Samples end March 2012
  - Data Summer 2012
  - Call sets end of 2012, Final paper 2013?
- End of the project

Richard Durbin

Fraction of variant sites present in an individual that are NOT already represented in dbSNP

Date	Fraction not in dbSNP
February, 2000	98%
February, 2001	80%
April, 2008	10%
February, 2011	2%
October 2011 (now)	<1%

Ryan Poplin, David Altshuler