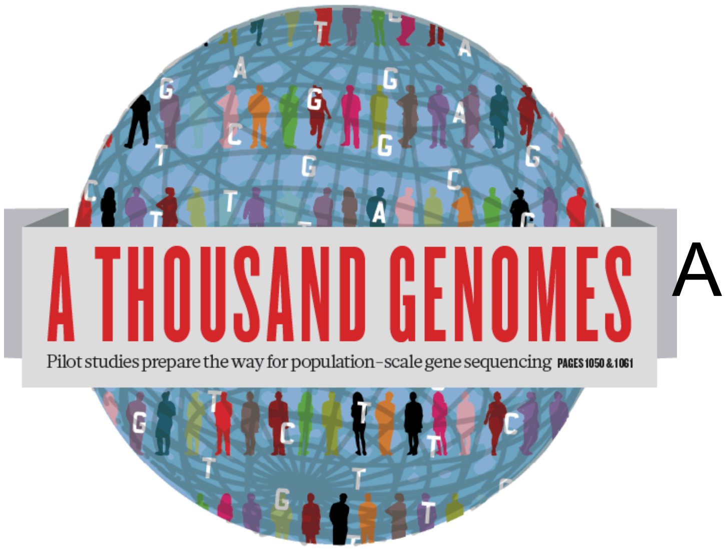

A THOUSAND GENOMES
Pilot studies prepare the way for population-scale gene sequencing PAGES 1050 & 1061

The 1000 Genomes Tutorial
A Brief History of Data and Analysis

Laura Clarke
17th February 2012

EMBL-EBI

Introduction

This Presentation should give the user an overview of the 1000 genomes project and a brief history of its data and analysis.

Summary

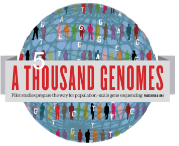
- Glossary
- Overview
- Pilot Strategies
- Main Project Design
- Sample Selection
- Hapmap, the Pilot and the Main Project
- The 1000 Genomes Timeline
- Fraction of SNPs in DbSNP overtime
- Sequence Data Evolution
- Present and Future of Project
- Pipeline Structure
- Alignment Strategies
- Variant Calling
- Integrated view of Variant Calling
- Exome versus Low Coverage Frequencies
- Functional Annotation
- Data Availability

Glossary

- **Pilot** : The 1000 Genomes project ran a pilot study between 2008 and 2010
- **Phase 1**: The initial round of exome and low coverage sequencing of 1000 individuals
- **Phase 2**: Expanded sequencing of 1700 individuals and method improvement
- **SAM/BAM**: Sequence Alignment/Map Format, an alignment format
- **VCF**: Variant Call Format, a variant format

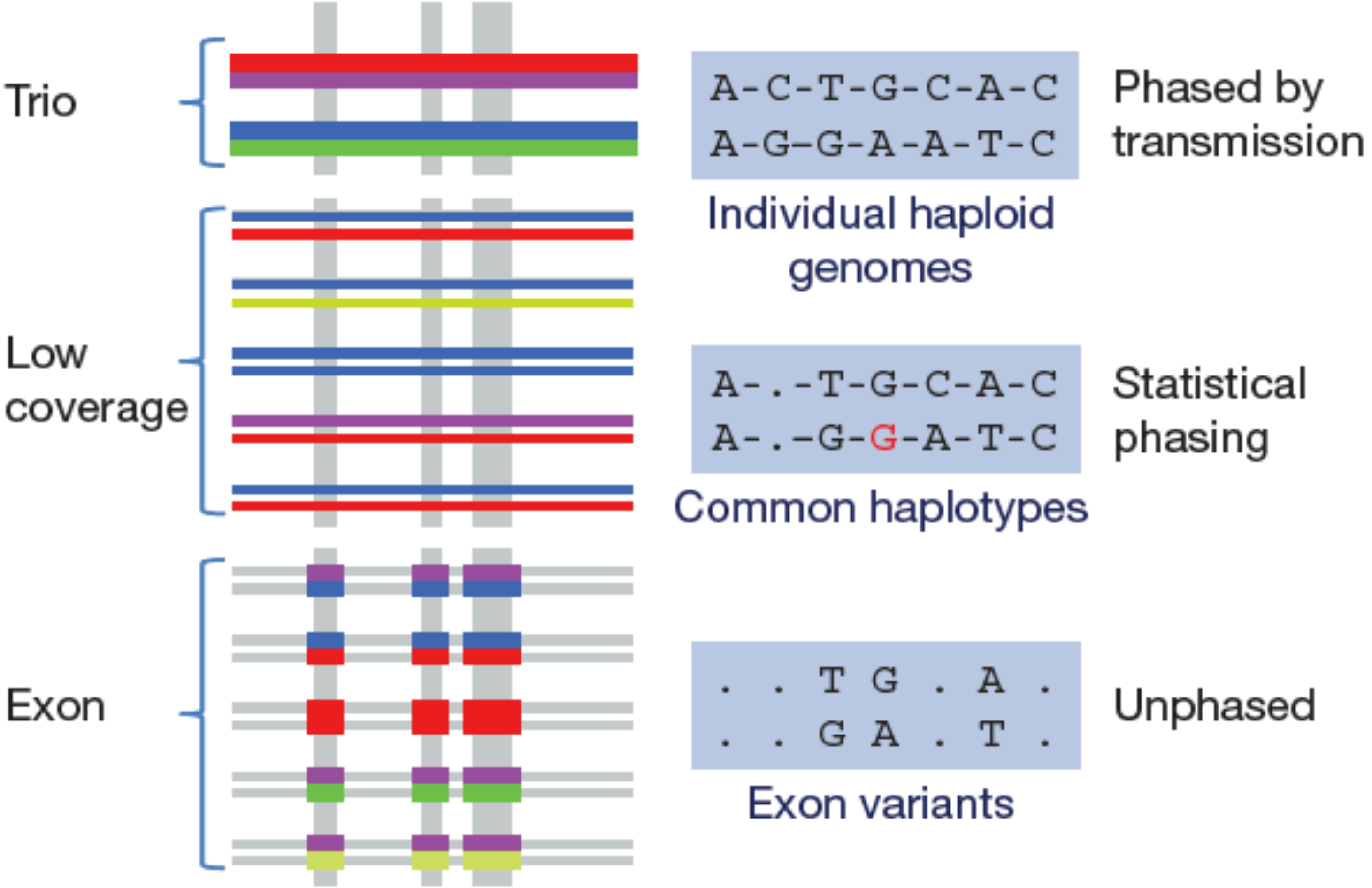
The 1000 Genomes Project: Overview

- 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



EMBL-EBI

3 pilot coverage strategies



Trio

Low coverage

Exon

A-C-T-G-C-A-C
A-G-G-A-A-T-C

Individual haploid genomes

Phased by transmission

A-.-T-G-C-A-C
A-.-G-G-A-T-C

Common haplotypes

Statistical phasing

. . T G . A .
. . G A . T .

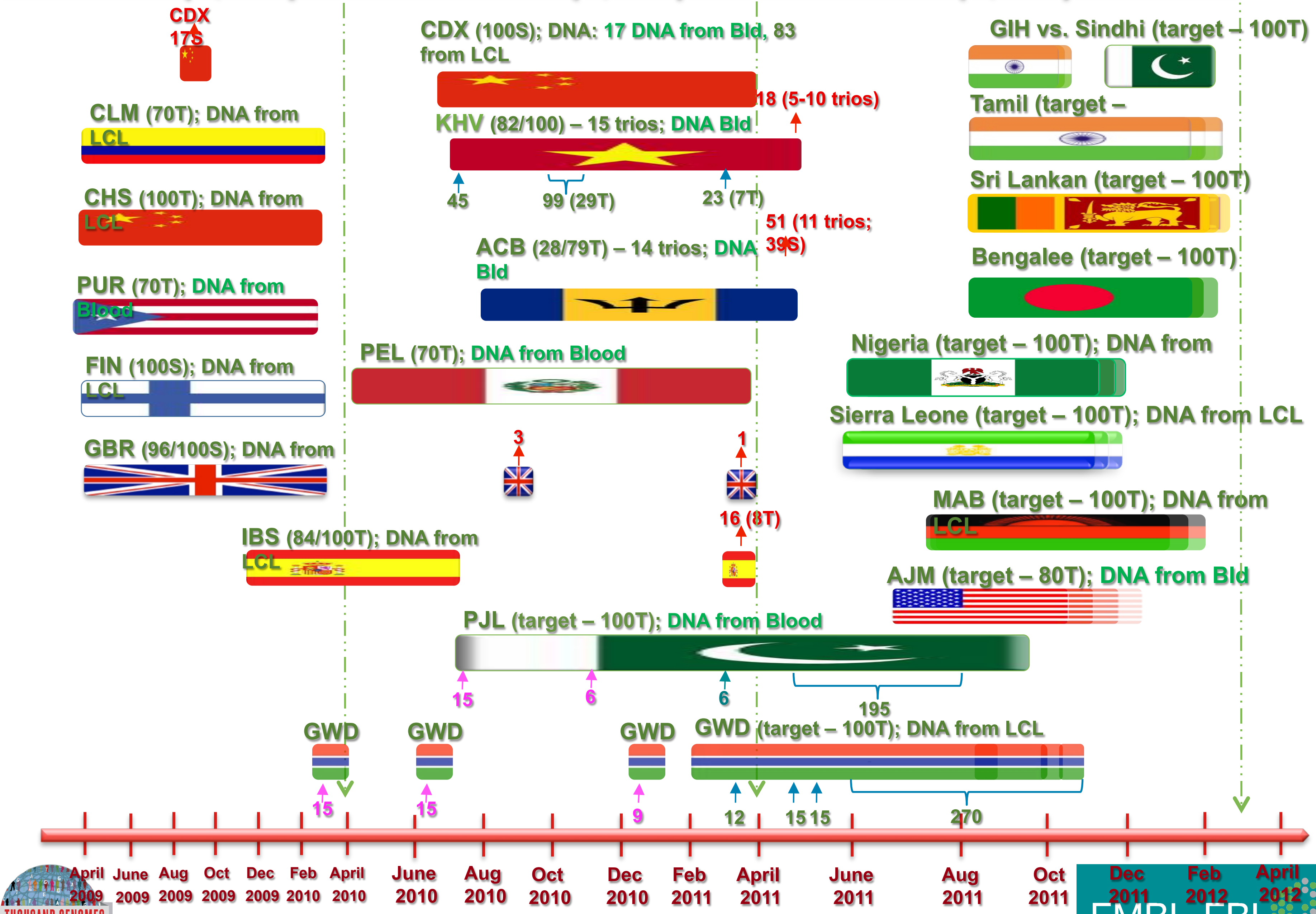
Exon variants

Unphased

Main 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 (>20x)
  - A number of deep coverage genomes to be sequenced, with details to be decided
  - High density genotyping at subsets of sites using both Illumina Omni and Affymetrix Axiom
- Phase 1 Release Integrated Variant Release has been made.

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

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

51 (11 trios; 39S)

ACB (28/79T) – 14 trios; DNA Bld

PEL (70T); DNA from Blood

16 (8T)

PJL (target – 100T); DNA from Blood

GIH vs. Sindhi (target – 100T)

Tamil (target –

Sri Lankan (target – 100T)

Bengalee (target – 100T)

Nigeria (target – 100T); DNA from

Sierra Leone (target – 100T); DNA from LCL

MAB (target – 100T); DNA from LCL

AJM (target – 80T); DNA from Bld

GWD

GWD

GWD

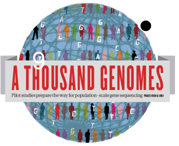
GWD (target – 100T); DNA from LCL

A THOUSAND GENOMES

EMBL-EBI

Hapmap, The Pilot Project and The Main Project

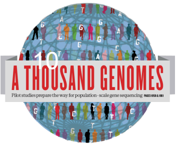
- Hapmap
  - Starting in 2002
  - Last release contained ~3m snps
  - 1400 individuals
  - 11 populations
  - High Throughput genotyping chips
- 1000 Genomes Pilot project
  - Started in 2008
  - Paper release contained ~14 million snps
  - 179 individuals
  - 4 populations
  - Low coverage next generation sequencing
- 1000 Genomes Phase 1
  - Started in 2009
  - Phase 1 release has 36.6millon snps, 3.8millon indels and 14K deletions
  - 1094 individuals
  - 14 populations
  - Low coverage and exome next generation sequencing
- 1000 Genomes Phase 2
  - Started in 2011
  - 1715 individuals
  - 19 Populations
  - Low coverage and exome next generation sequencing



EMBL-EBI

Timeline

- September 2007: 1000 Genomes project formally proposed Cambridge, UK
- April 2008: First Submission of Data to the Short Read Archive.
- May 2008: First public data release.
- October 2008: SAM/BAM Format Defined.
- December 2008: First High Coverage Variants Released.
- December 2008: First 1000 genomes browser released
- May 2009: First Indel Calls released.
- July 2009: VCF Format defined
- August 2009: First Large Scale Deletions released.
- December 2009: First Main Project Sequence Data Released.
- March 2010: Low Coverage Pilot Variant Release made
- July 2010: Phased genotypes for 159 Individuals released.
- October 2010: A Map of Human Variation from population scale sequencing is published in Nature.
- January 2011: Final Phase 1 Low coverage alignments are released
- May 2011: @1000genomes appears on Twitter
- May 2011: First Variant Release made on more than 1000 individuals
- October 2011: Phase 1 integrated variant release made



A THOUSAND GENOMES

EMBL-EBI

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%
Now	<1%

Ryan Poplin, David Altshuler

Sequencing Data Evolution

- The Project contains data from 3 different providers and multiple platforms

Platform	Min Read Length (bp)	Max Read Length (bp)
454 Roche GS FLX Titanium	70	400
Illumina GA	30	81
Illumina GA II	26	160
Illumina HiSeq	50	102
ABI Solid System 2.0	25	35
ABI Solid System 2.5	50	50
ABI Solid System 3.0	50	50

1000 Genomes Project: Present & Future

- First Phase 2 sequence release 14th November 2011
- First Phase 2 alignment release in progress
- First Phase 2 variant site release Summer 2012

- Sample collected expected end to June 2012
- Final Phase 3 Sequence release expected December 2012
- 2013 will represent finalization of 1000 genomes analysis results and final data releases

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 Site Discovery
  - Individual Genotype Assignment (sometimes part of site discovery)
  - Variant filtering / call set refinement
  - Variant reporting

Alignment Data

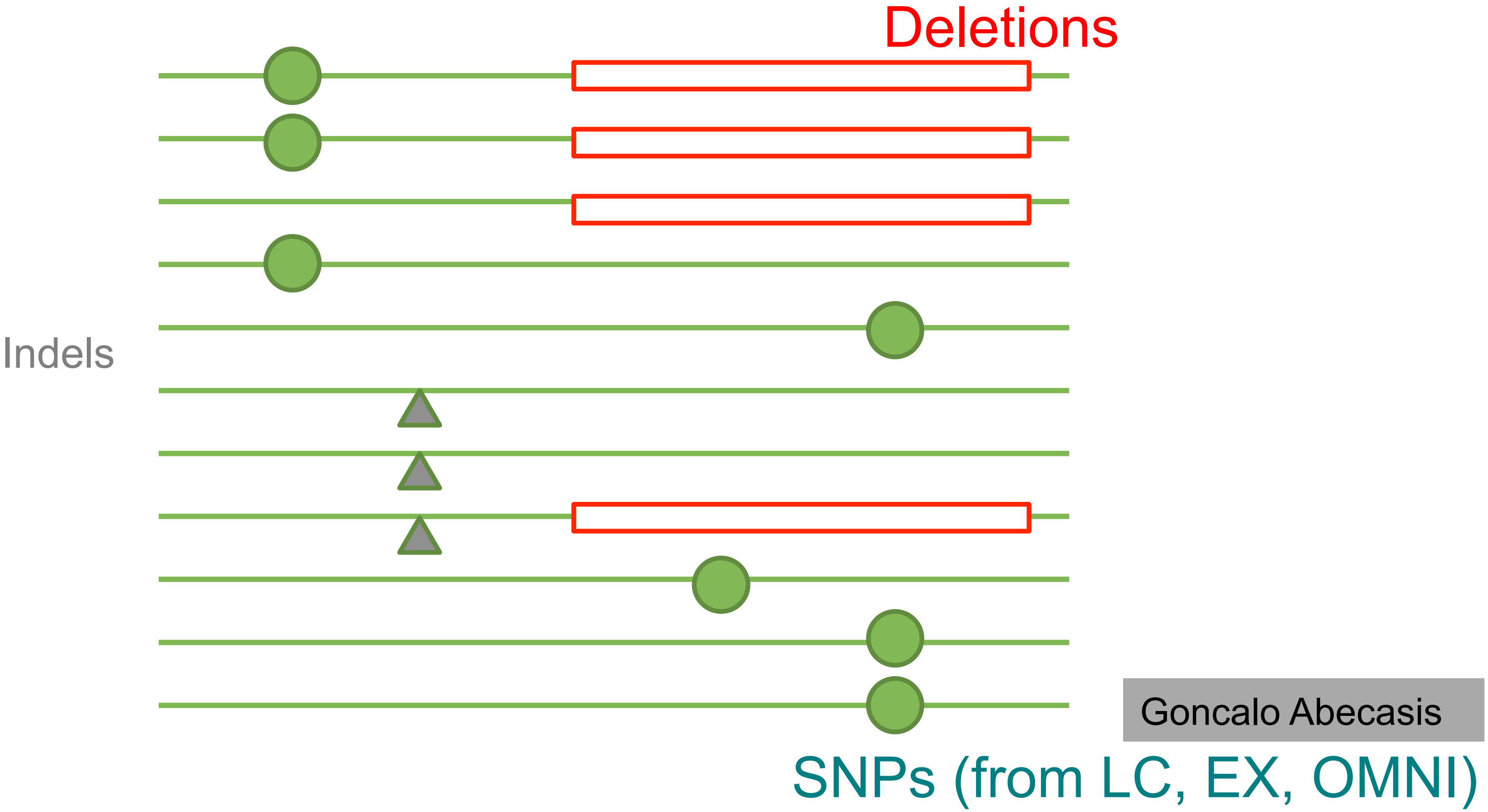
- The project has made more than 10 releases of Alignment Data
- Pilot Project
  - Aligned to NCBI36
  - Maq and Corona
  - Base Quality Recalibration done
- Phase 1
  - Aligned to GRCh37
  - BWA and Bfast
  - Indel Realignment
- Phase 2
  - Aligned to extended GRCh37
  - Improvements to Base Quality Recalibration

Variant Calling

- Early call sets used a single variant caller
- Intersect approach developed during pilot
- Variant Quality Score Recalibration (VQSR) developed for Phase 1
- Genotype Likelihoods assigned to help with genotype calling
- Integrated genotype calling based on individual variant call sets
- Phase 2 looks to improve site discovery and improve integration

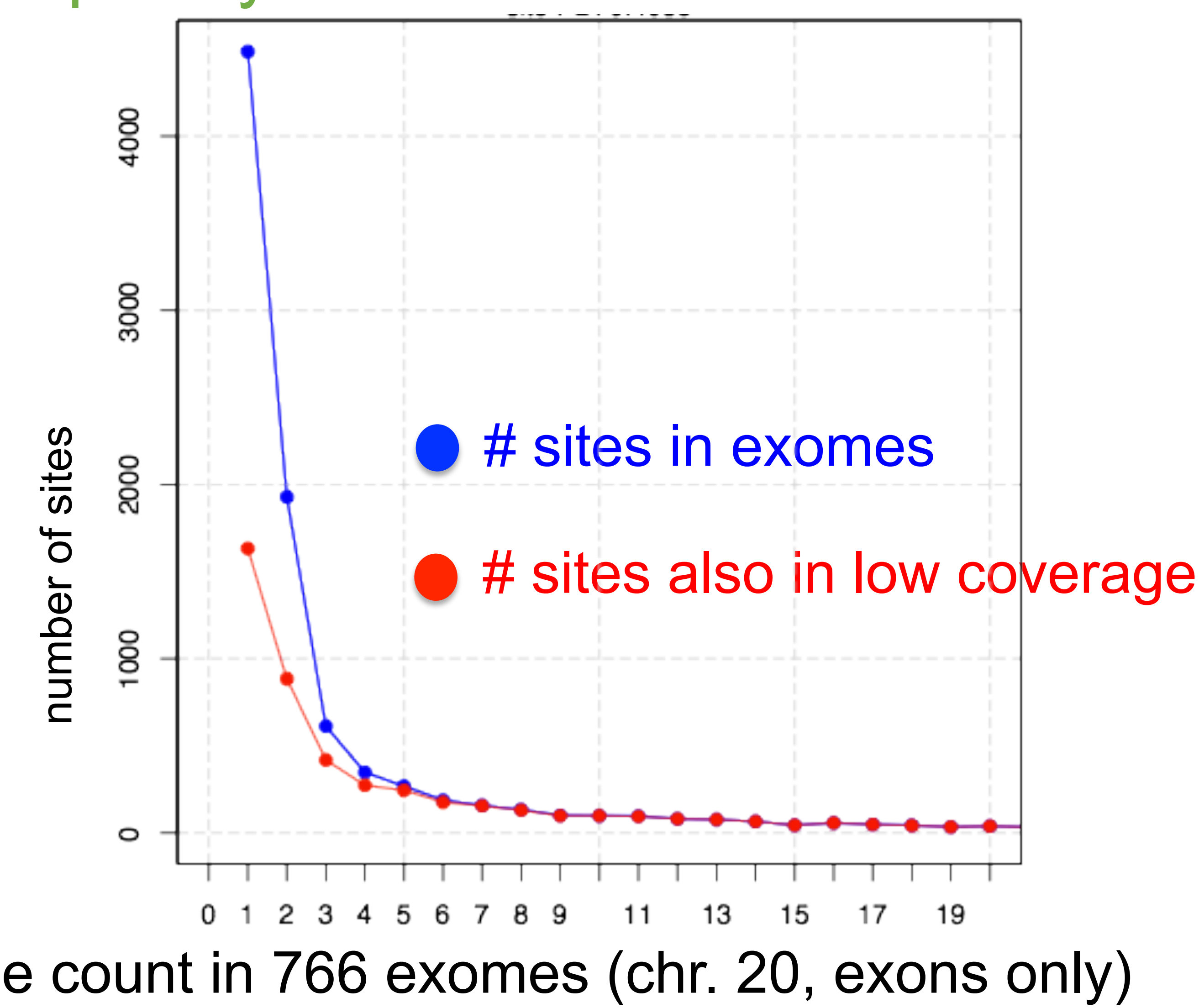
Phase 1 analysis goal: an **integrated view of human variations**

- Reconstruct haplotypes including all variant types, using all datasets



Goncalo Abecasis

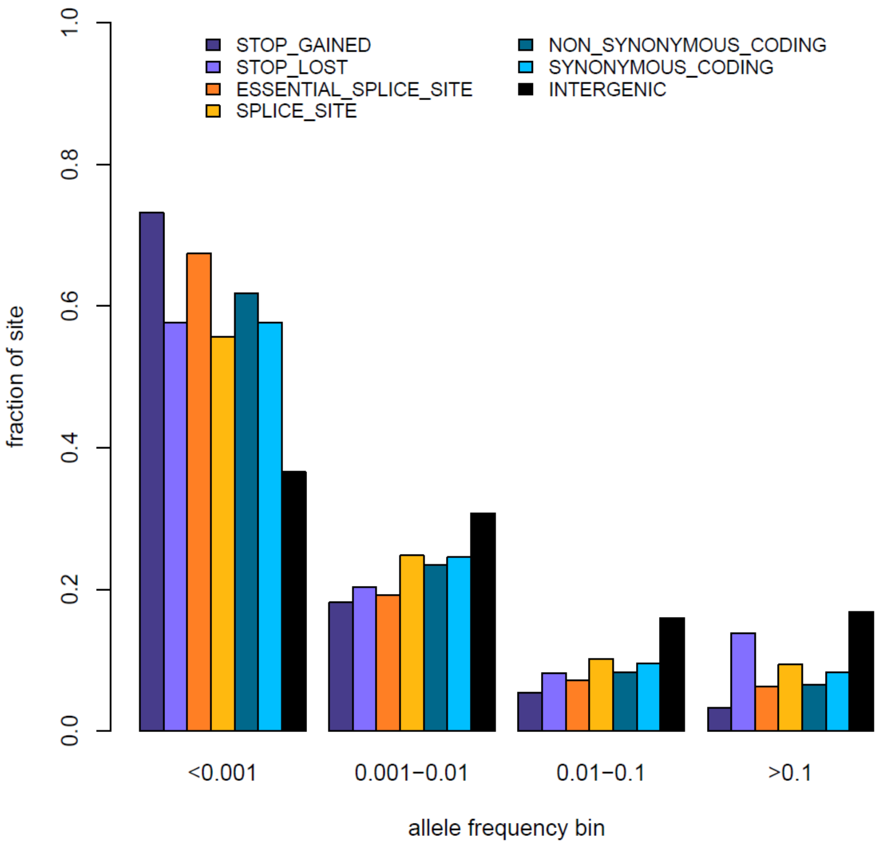
Deep coverage exome data is more sensitive to low-frequency variants



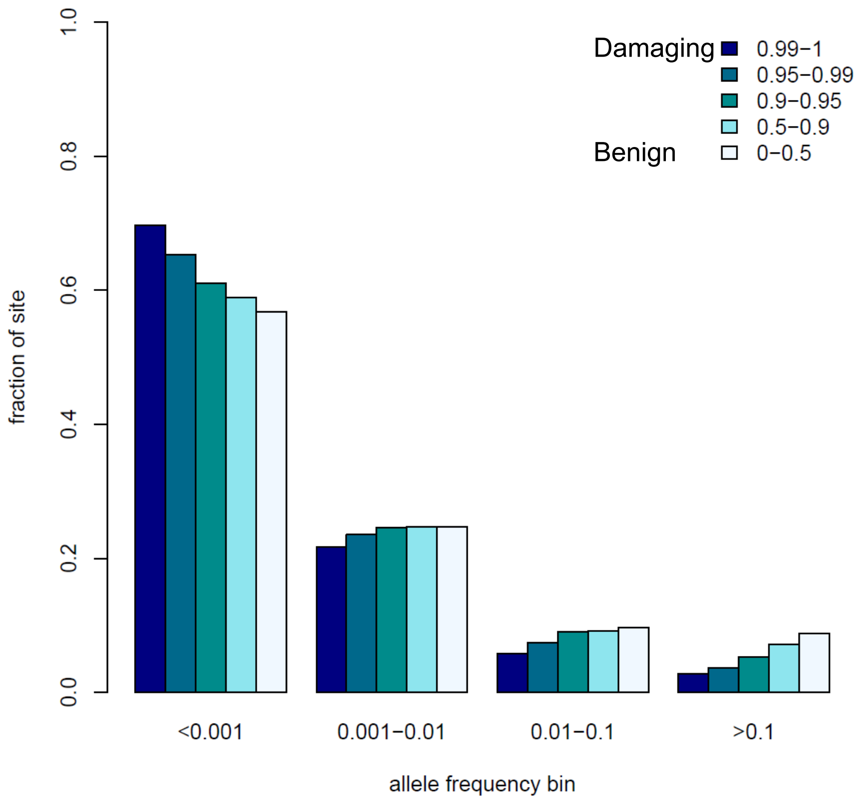
Erik Garrison

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

EMBL-EBI

Data Availability

- FTP site: ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/
  - Raw Data Files
- Web site: http://www.1000genomes.org
  - Release Announcements
  - Documentation
- Ensembl Style Browser: http://browser.1000genomes.org
  - Browse 1000 Genomes variants in Genomic Context
  - Variant Effect Predictor
  - Data Slicer
  - Other Tools

Announcements

- http://1000genomes.org
- 1000announce@1000genomes.org
- http://www.1000genomes.org/1000-genomes-annoucement-mailing-list
- http://www.1000genomes.org/announcements/rss.xml
- http://twitter.com/#!/1000genomes

Questions

Please send any questions about this presentation and any other material on our website to info@1000genomes.org

Thanks

- The 1000 Genomes Project Consortium
- Paul Flicek
- Richard Smith
- Holly Zheng Bradley
- Ian Streeter

A THOUSAND GENOMES

EMBL-EBI