# 1000G August release, whole genome call set

Marth lab, Boston College
November 2010

# Data

- Whole genome call set on 08/04/2010 sequence data
- Data consists of 629 samples from:
  - AFR (YRI + LWK + ASW + PUR)
  - ASN (CHB + CHS + JPT + MXL)
  - EUR (CEU + TSI + GBR + FIN + MXL + PUR)
- Calling performed on all samples simultaneously

# Data processing - pipeline

- Alignments generated using:
  - Mosaik (LS454 and Illumina) at the NCBI
  - bFast (SOLiD) at TGEN
- Base quality score recalibration (GATK)
- Duplicate marking (Picard/BCM)
- BAQ calculation from Heng Li
- Post-processing filtering based on:
  - SNP quality
  - Strand bias
  - Allele balance

# High quality SNP call set

- Comparisons with:
  - dbSNP build 129
  - HapMap 3.2
  - Pilot3 release set

| Total # | Known | Novel | %dbSNP | Known Ts/Tv | Novel Ts/Tv | Total Ts/Tv | % missed HapMap | % missed Pilot3 |
|---|---|---|---|---|---|---|---|---|
| 23,669,324 | 7,780,182 | 15,889,142 | 32.87 | 2.16 | 2.30 | 2.26 | 1.12 | 46.72 |