# Genomic Technologies and Associated Challenges

Jim Mullikin
Genome Technology Branch
NHGRI

# Overview

- SNPs
- HapMap
- WGG (whole genome genotyping)
- Directed Sequencing
- New Sequencing Technologies

# Polymorphisms

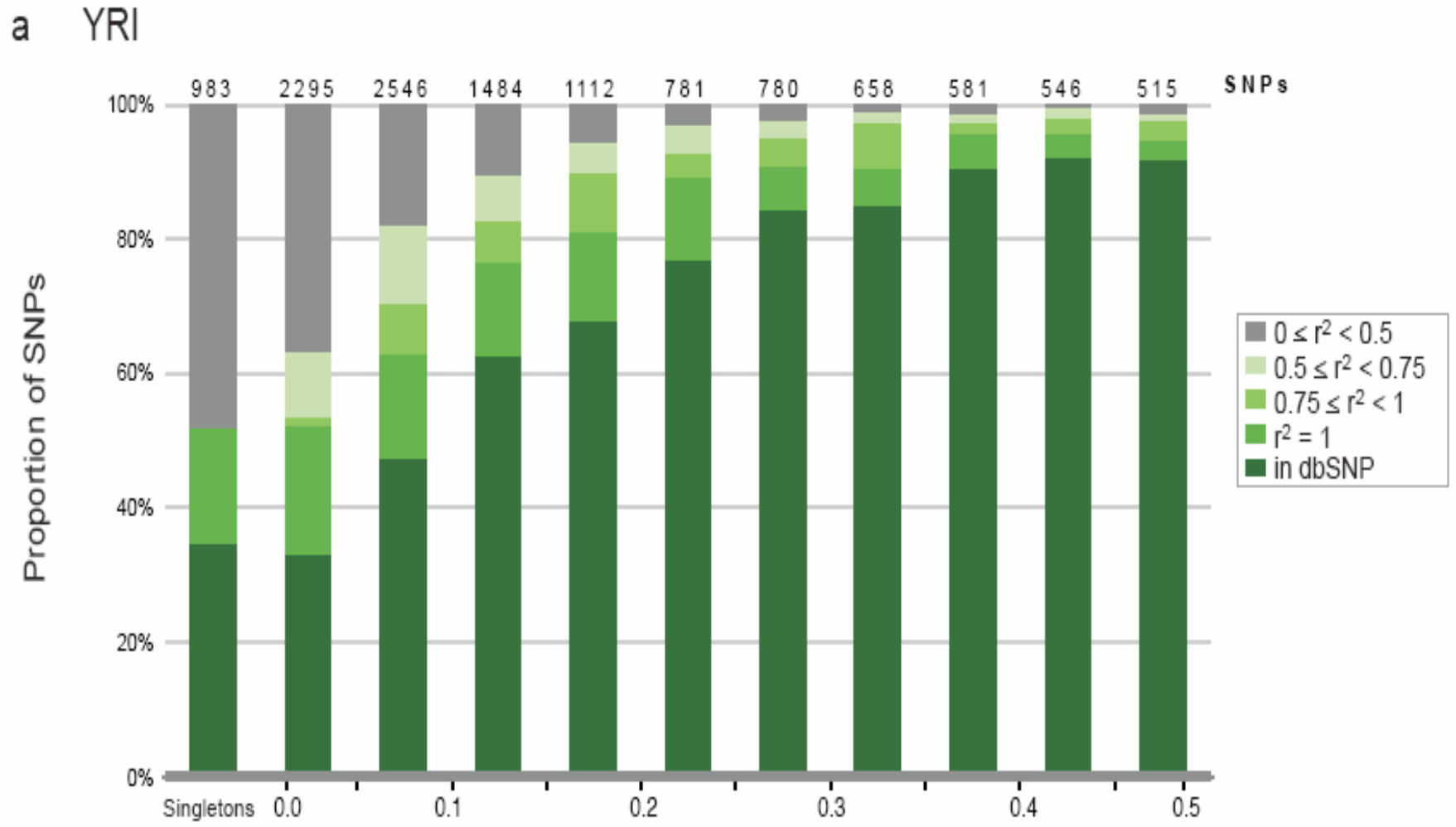
- Ascertainment
- Depth of discovery
- Deletion/Insertion Polymorphisms
- Structural Variants/Polymorphisms
  - 48 HapMap samples each having
    - Fosmid end sequencing
    - 0.4X sequence coverage
    - 6X physical coverage
- Copy Number Variants/Polymorphisms

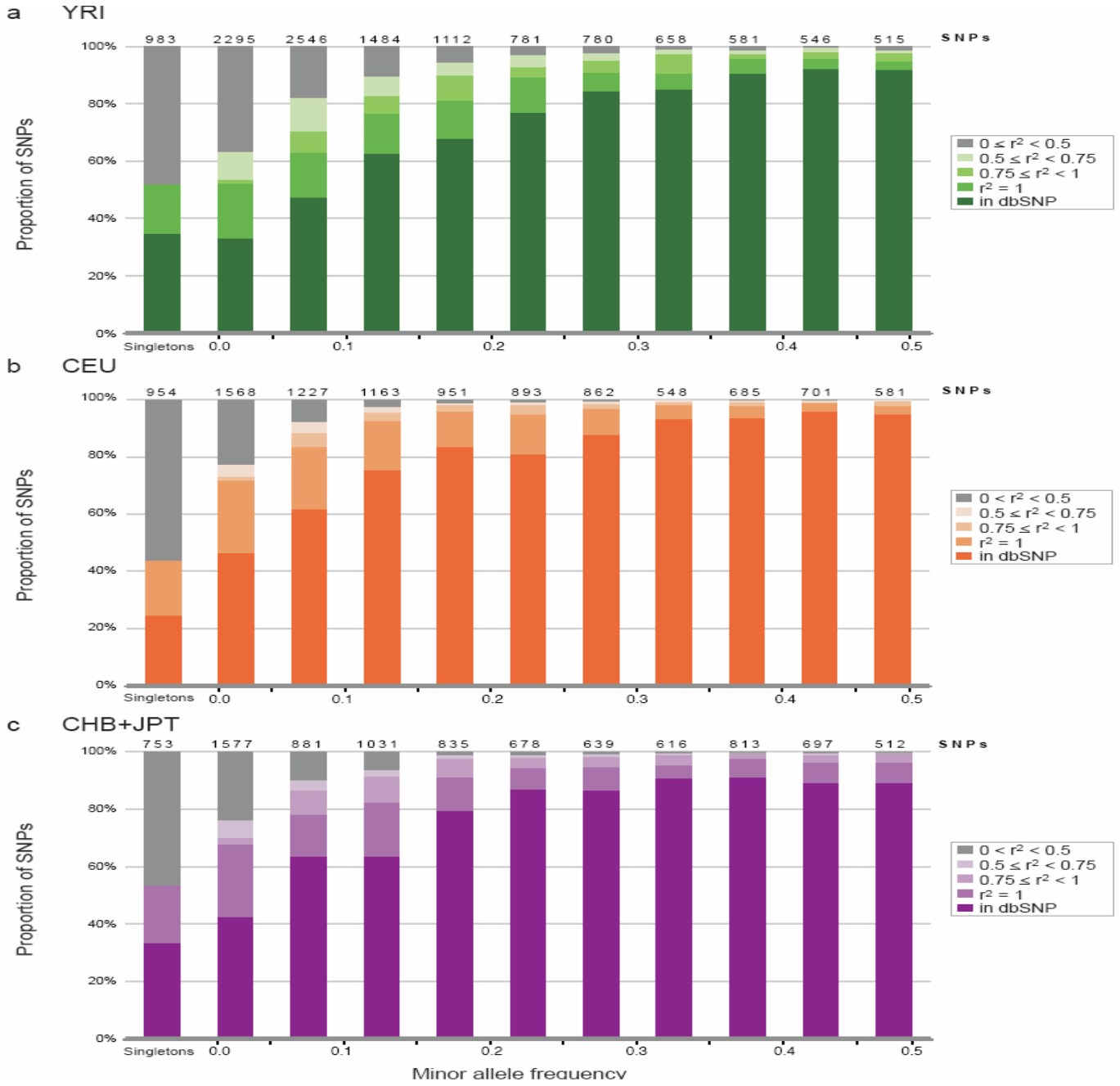

(www.hapmap.org)

- Goal: To construct a haplotype map across the entire genome
  - 270 individuals (Nigerians, Japanese, Chinese and Northern Europen)

- Phase 1: completed 03/01/2005
  - 1,000,000 common SNPs (≥ 5%) genotyped: 1 per ~5 kb
- Phase 2: completed 10/28/2005
  - ~4,000,000 common SNPs (>5%) genotyped: 1 per ~1.5 kb

- A few hundred thousand SNPs will be needed to capture common variation across the entire genome (2005-2006)
  - A framework for comprehensive candidate gene and genome-wide association studies
  - Between 500,000 SNPs (Asian, European panels) and 1,000,000 SNPs for African panels


a
YRI
983
2295
2546
1484
1112
781
780
658
581
546
515
SNPs
100%
80%
60%
40%
20%
0%
Proportion of SNPs
0 ≤ r² < 0.5
0.5 ≤ r² < 0.75
0.75 ≤ r² < 1
r² = 1
in dbSNP
Singletons
0.0
0.1
0.2
0.3
0.4
0.5


a
YRI
983 2295 2546 1484 1112 781 780 658 581 546 515
SNPs
100%
80%
60%
40%
20%
0%
Proportion of SNPs
Singletons 0.0 0.1 0.2 0.3 0.4 0.5
0 ≤ r² < 0.5
0.5 ≤ r² < 0.75
0.75 ≤ r² < 1
r² = 1
in dbSNP
b
CEU
954 1568 1227 1163 951 893 862 548 685 701 581
SNPs
100%
80%
60%
40%
20%
0%
Proportion of SNPs
Singletons 0.0 0.1 0.2 0.3 0.4 0.5
0 < r² < 0.5
0.5 ≤ r² < 0.75
0.75 ≤ r² < 1
r² = 1
in dbSNP
c
CHB+JPT
753 1577 881 1031 835 678 639 616 813 697 512
SNPs
100%
80%
60%
40%
20%
0%
Proportion of SNPs
Singletons 0.0 0.1 0.2 0.3 0.4 0.5
Minor allele frequency
0 < r² < 0.5
0.5 ≤ r² < 0.75
0.75 ≤ r² < 1
r² = 1
in dbSNP

# Whole Genome Genotyping

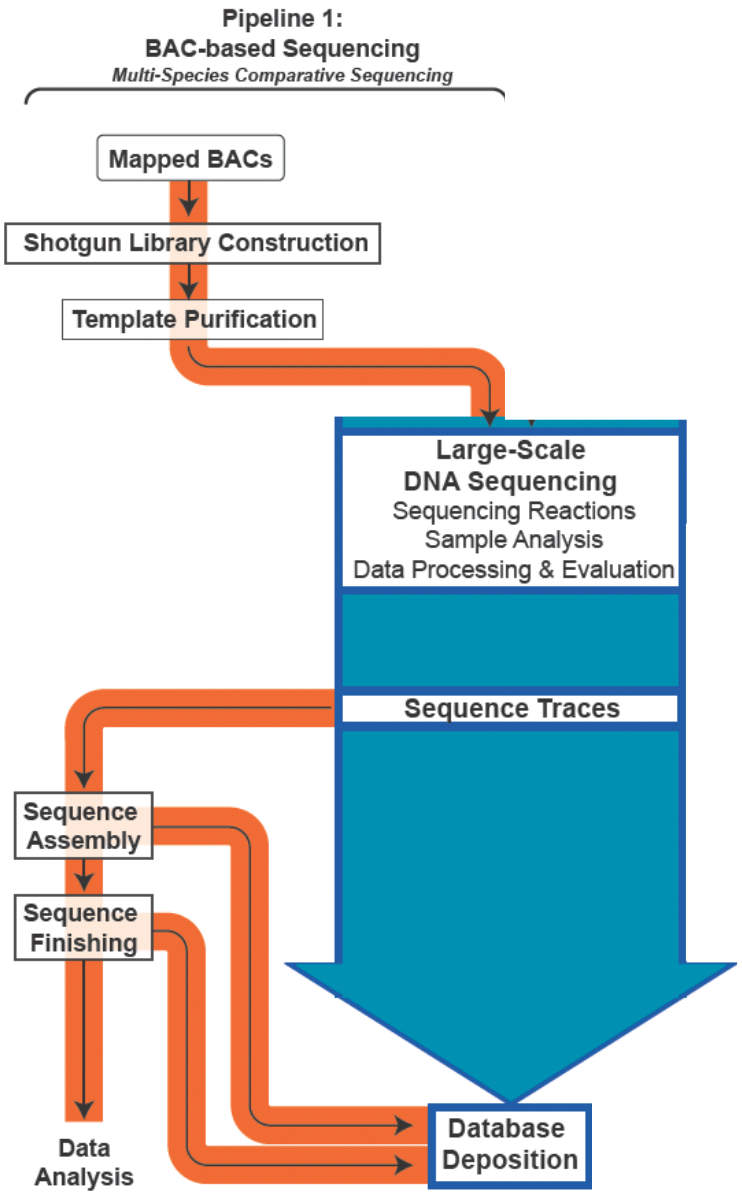
- Affymetrics
- Illumina
- Perlegen

# Directed Sequencing

- DNA samples, e.g. cases/controls
- Target regions
  - Association study
  - Candidate genes
  - Hybrid approach
- Sanger sequencing
- Variant detection

# Experimental Design

- Collaborator
  - Medical background & rationale
  - Sample collection
  - Controls
  - Gene(s)/Region(s) of interest
    - Conserved
    - Coding
    - Regulatory
    - Everything


Pipeline 1:
BAC-based Sequencing
Multi-Species Comparative Sequencing
Mapped BACs
Shotgun Library Construction
Template Purification
Large-Scale
DNA Sequencing
Sequencing Reactions
Sample Analysis
Data Processing & Evaluation
Sequence Traces
Sequence
Assembly
Sequence
Finishing
Data
Analysis
Database
Deposition

# Sequence Analysis

- Germ-line
  - Susceptibility/outcome
- Somatic analysis
  - Cancer
- Comparative analysis
  - Molecular evolution
  - Insight into sequences of significance

# New Sequencing Technologies

- 454
- Solexa
- Viable today?
  - but will be “tomorrow”…
- Enormous data generation / data overload

# Issues in Sequence Analysis

| | |
|---|---|
| **Rare Variants** | |
| Family Studies | Are There Enough? |
| Functional Analysis | Very Slow! |
| Annotation issues | Database? |
| Population-specific issues | Database? |
| **Comparison with altered tissue** | |
| Duplicate effort | Parallel analysis |
| **Copy Number Variation** | |
| Annotation issues | Database? |