# Exome Genotyping Arrays

Gonçalo Abecasis and Benjamin Neale

# Motivation for an Exome Array

- Current sequencing studies are well powered to discover exome variants that contribute to disease (MAF > 0.1%)

- Sequencing studies may be underpowered to establish association of those variants to phenotype
  - Larger numbers of individuals must be examined to establish the effect of a variant than to discover it

- Genotyping is less expensive that exome sequencing, allowing larger sample sizes and, perhaps, power

# Caveats and limitations

- Because assay conversion rates are ~80%, genotyping arrays will only reach a subset of SNPs captured by sequencing
- Genotyping will not be effective for loci where private mutations drive the association
- The proposed arrays are heavily biased towards SNPs

# Approach to designing the array

- Collate sites and counts from a “coalition of the willing” with data from exome or genome sequencing
  - Site lists constitute preliminary analyses of unpublished data
- Cover as much variation as possible while avoiding private mutations and technical artifacts
  - Nonsynonymous variants >=3 times and in >=2 call sets
  - Splice and stop variants seen >=2 times and in >=2 call sets
  - Relaxed frequency filter for ancestries with fewer samples
- Pass quality filters, HWE $P > 10^{-6}$ (except on X)

# Approach to design (continued)

- Singleton sites excluded
  - Too many and most unlikely to be seen again
- Synonymous sites generally excluded
  - Focus is enriching for functional mutations
  - 5000 synonymous sites included as “controls”
- Selection agnostic to previous annotation with respect to clinical impact

# Content Contributors (12,031 samples)

| Contributor | Enrichment | Major Ethnicity | N | Contact |
|---|---|---|---|---|
| NHLBI Exome Sequencing Project (5 tranches) | Cardiovascular, Lung Traits, Obesity | European, African American | 4260 | D. Nickerson<br>D. Altshuler<br>S. Rich |
| Autism (2 tranches) | Autism | European | 1778 | M. Daly<br>R. Gibbs |
| GO T2D (2 tranches) | Type 2 Diabetes | European | 1618 | D. Altshuler<br>M. Boehnke<br>M. McCarthy |
| 1000 Genomes Project (2 tranches) | Random Sample | Diverse | 1128 | H. M. Kang |
| Sweden Schizophrenia Study | Schizophrenia | European | 525 | S. Purcell<br>P. Sklar |
| SardiNIA | Random Sample | European | 508 | F. Cucca<br>G. Abecasis |
| Sanger / CoLaus | Overweight, Diabetes, Fasting Glucose | European | 456 | I. Barroso |
| Cancer Genome Atlas | Cancer | European | 422 | S. Gabriel |
| T2D Genes | Diabetes | Hispanic | 362 | J. Blangero<br>G. Abecasis |
| Cancer Cohort Study | Cancer | Chinese | 327 | W. Zheng |
| Smaller Contributors (6 tranches) | T2D, Lipids, HIV, DILI, Depression, BMI | European | 647 | (next slide) |

# Content Contributors (continued)

| Contributor | Enrichment | Major Ethnicity | N | Contact |
|---|---|---|---|---|
| BMI Extremes | BMI Extremes | European | 46 | J. Hirschhorn |
| I2B2 - Major Depression | Major Depression, Major Depressive Disorder | European | 50 | R. Perlis<br>J. Smoller |
| SAEC DILI<br>(merged w/Autism tranches) | Augmentin DILI | European | 117 | M. Daly<br>A. Holden |
| Int'l HIV Controllers Study | HIV Controllers | European | 121 | P. De Bakker |
| Lipid Extremes | Lipid Extremes | European | 131 | S. Kathiresan<br>D. Rader |
| Pfizer – MGH – Broad T2D | Type 2 Diabetes Extremes of Risk | European | 182 | D. Altshuler |

- Chip design estimated to be based on
  - ~9000 samples of primarily European ancestry
  - ~2000 samples of primarily African ancestry
  - ~500 samples each of primarily Hispanic and Asian ancestry

# Coding Content (RefSeq genes)

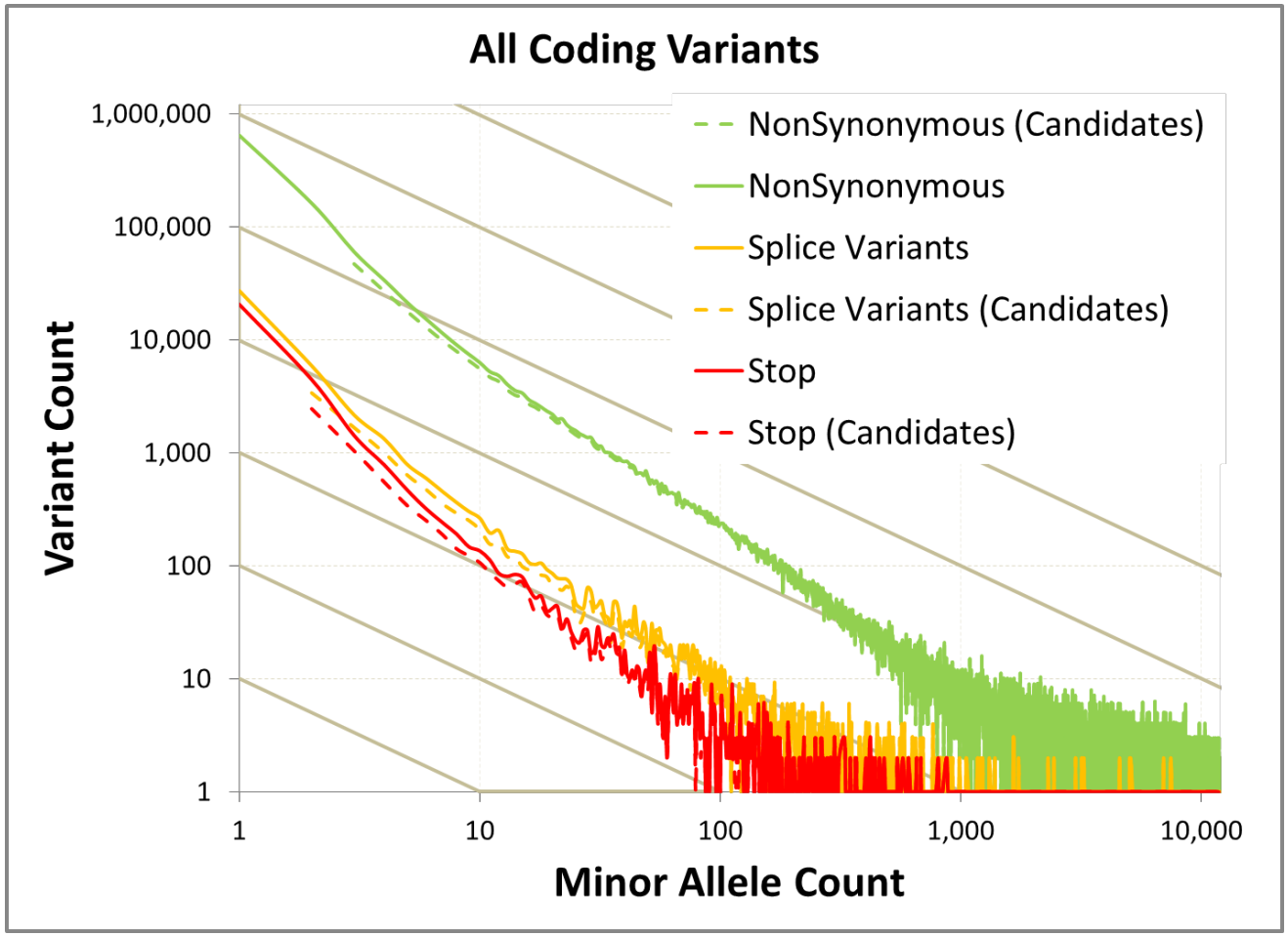
- 1,107,051 nonsynonymous variants
  - 646,888 with allele count = 1
  - 163,044 with allele count = 2
  - 297,119 with allele count > 2
  - 260,054 seen in at least 2 studies
- 44,529 splice variants
  - 27,265 with allele count = 1
  - 17,264 with allele count > 1
  - 12,662 seen in at least 2 studies
- 31,003 stop gain/loss variants
  - 20,637 with allele count = 1
  - 10,366 with allele count > 1
  - 7,137 seen in at least 2 studies

# Coding Content (RefSeq genes)

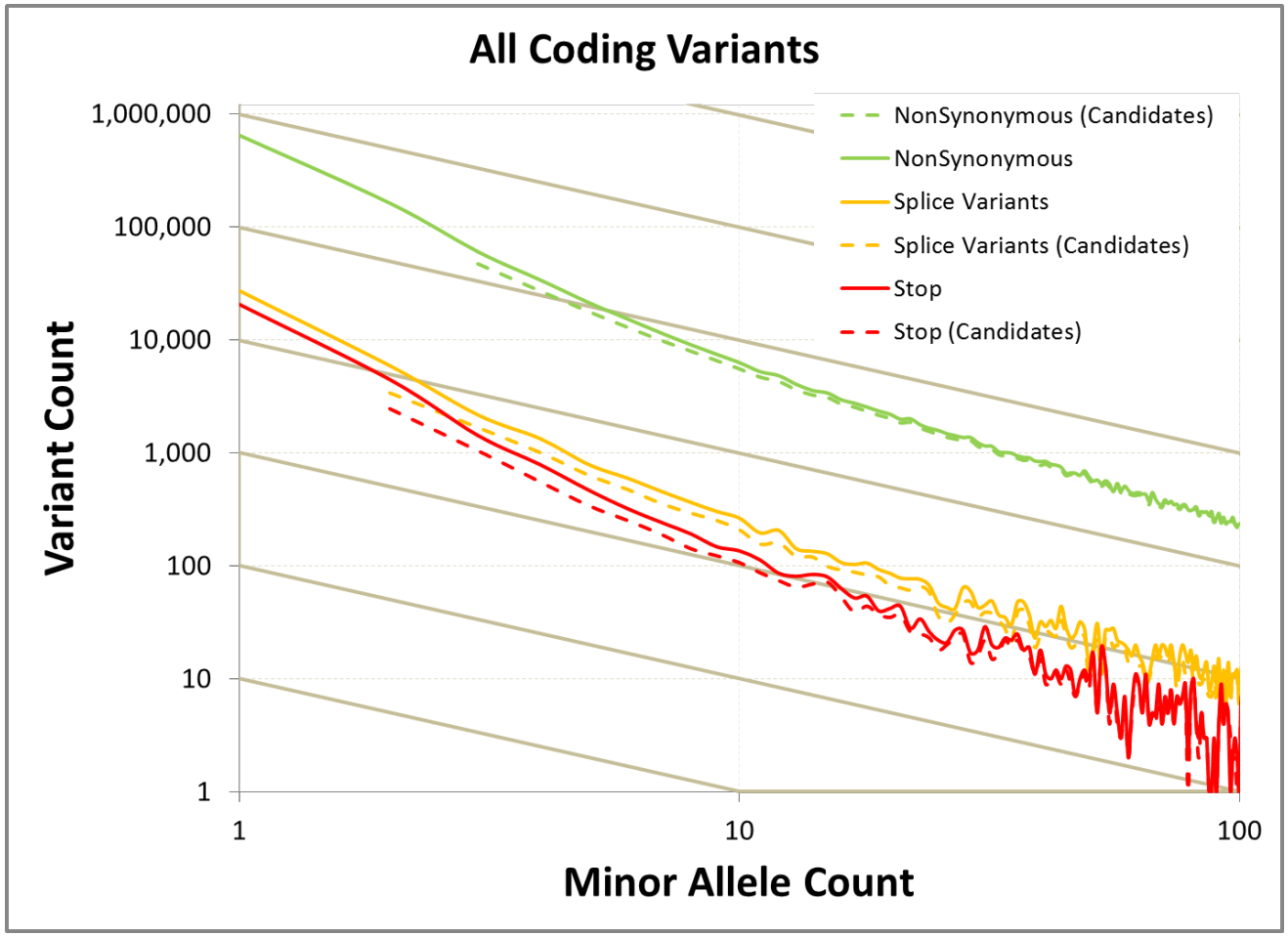
- 1,107,051 nonsynonymous variants
  - 646,888 with allele count = 1
  - 163,044 with allele count = 2
  - 297,119 with allele count > 2
  - **260,054 seen in at least 2 studies**
- 44,529 splice variants
  - 27,265 with allele count = 1
  - 17,264 with allele count > 1
  - **12,662 seen in at least 2 studies**
- 31,003 stop gain/loss variants
  - 20,637 with allele count = 1
  - 10,366 with allele count > 1
  - **7,137 seen in at least 2 studies**

We estimate the **candidate sites** account for 97-98% of the non-synonymous sites in an individual and 94-95% of the splice and stop gain/loss variants in an individual.

# Frequency Spectrum: Most Variants Are Very Rare



# Frequency Spectrum (Allele Count <100): Most Variants Are Very, Very Rare



# Additional Content

| SNP Set | Size | Contact |
|---|---|---|
| GWAS Tag SNPs | 5763 | G. Abecasis |
| Grid of Common Variants | 5710 | G. Abecasis |
| Random Synonymous | 5000 | G. Abecasis |
| AIM – African Ancestry | 3388 | G. Abecasis |
| AIM – Native American | 1000 | C. Bustamante, E. Buchard |
| HLA Tags | 2536 | P. De Bakker / S. Rich |
| ESP Requests | 1003 | NHLBI ESP |
| Fingerprint SNPs | 285 | S. Gabriel / M. Rieder |
| Micro RNA Target Sites | 285 | P. Chines |
| Mitochondrial | 246 | V. Mootha |
| Chromosome Y | 188 | J. Wilson |
| Indels | 181 | B. Neale / D. MacArthur |

# Implementing Array with Illumina

- Illumina allocated a budget of ~300,000 beads
  - Most SNPs require one bead
  - A/T and G/C SNPs require two beads
  - Tri-allelic SNPs require four beads for full determination

- Design considerations
  - Excluded sites with Illumina design score < 0.5
  - Avoided primers with multiple hits in genome
  - Favored primers extending into mRNA (versus intron)
  - Excluded sites within 5-bp of site with allele count of 100+
  - Used only one or two beads per site (even for tri-allelics)

- Array will allow ~20,000 – 30,000 custom beads

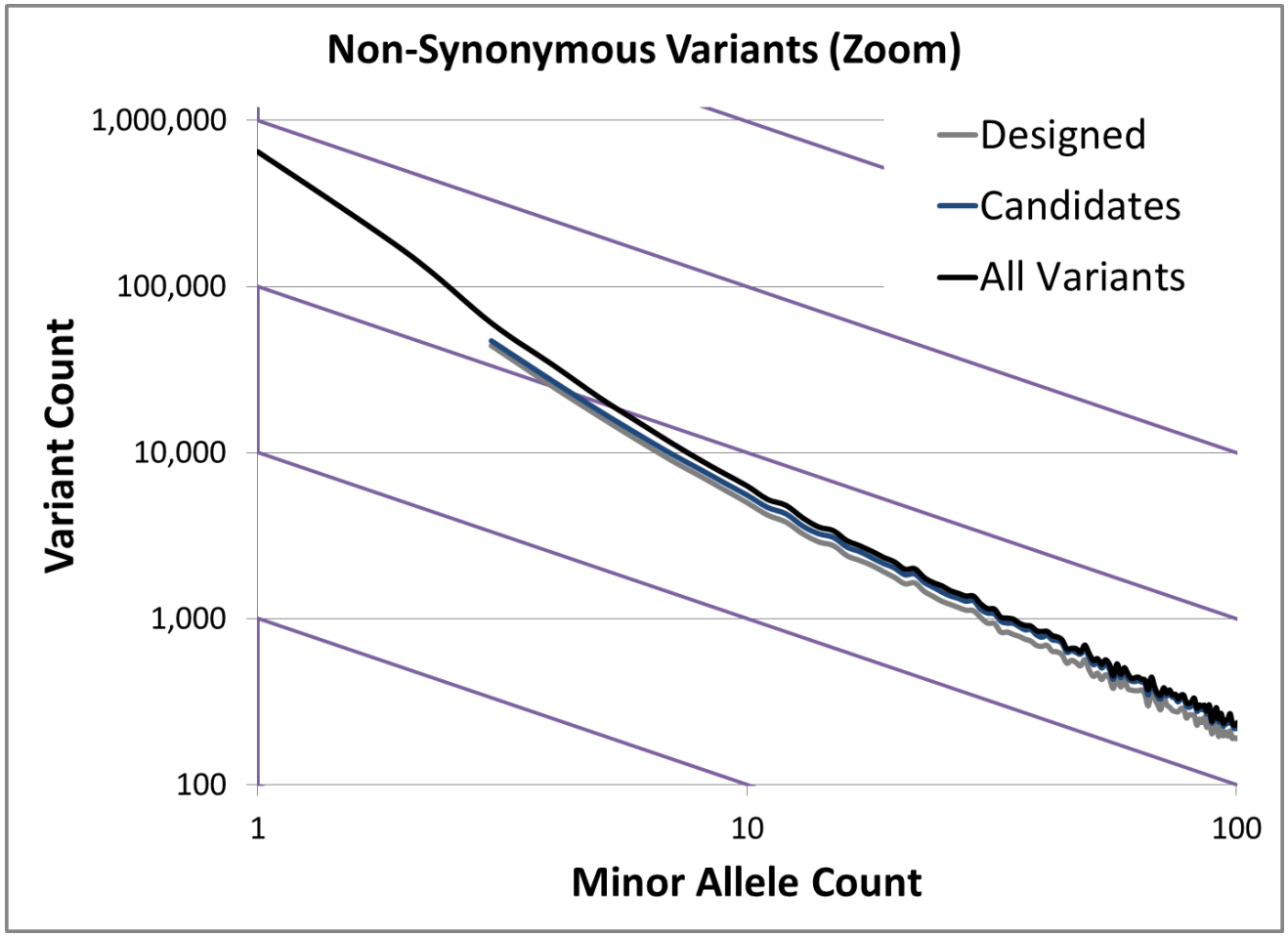
# Designed Content
# (10-15% will fail manufacturing)

| SNP Set | Candidates | Designed |
|---|---|---|
| Coding Content | 275,165 | 243,094 (+ 8,242*) |
| GWAS Tag SNPs | 5,763 | 5,325 |
| Grid of Common Variants | 5,710 | 5,286 |
| Random Synonymous | 5,000 | 4,651 (+870**) |
| AIM – African Ancestry | 3,388 | 3,241 |
| AIM – Native American | 1,000 | 998 |
| HLA Tags | 2,536 | 2,459 |
| ESP Requests | 1,003 | 846 |
| Finger Print SNPs | 285 | 259 |
| Micro RNA Target Sites | 285 | 270 |
| Mitochondrial | 246 | 246 |
| Chromosome Y | 188 | 128 |
| Indels | 181 | 181 |

* 8,242 variants seen in one study, but included to increase ethnic diversity.

** 870 variants assayed on both strands to facilitate development of genotyping methods / QC

# Design Failures
# Largely Independent of Frequency



# Proposed Illumina Pricing

| | Early Access Pricing (Prior to October 15, 2011) | Post Launch Pricing (After October 15, 2011) |
|---|---|---|
| **Exome Chip** | | |
| - List Price | $39 | $75 |
| > 5,000 samples / PO | $39 | $59 |
| > 10,000 samples / PO | $39 | $50 |
| **OmniExpress + Exome Chip** | | |
| - List Price | $279 | $299 |
| > 5,000 samples / PO | $265 | $284 |
| > 10,000 samples / PO | $149 | $155 |

These prices were committed to by Illumina as a condition of our sharing the site list

# Implementing Array with Affymetrix

- Affymetrix allocated budget of ~350,000 probes
  - Design will be based on Axiom platform
  - Will allow for querying 10,000s of indels
- Design considerations
  - Plan to favor sites extending into exon
  - Plan to exclude sites that are close to each other
- Affymetrix will assay 1000 Genome Project samples
  - Results will be provided to 1000 Genome Consortium
- Cost per sample will be $45/chip until December 31, 2011
  - Price was committed to by Affymetrix

# FAQ

- **Will these exome arrays provide information of clinical utility?**
- The truthful answer is that we don't know.
- Some of the variants included in the chip may be of clinical utility. (for example, >50 BRCA1 coding variants meet selection criteria)
- There is no data on whether chip-derived genotypes will be accurate enough to support clinical decisions.
- There is no data on the clinical utility of these variants when detected in the general population (versus at risk individuals).

# More Details

- Wiki Page with Array Description
  http://genome.sph.umich.edu/wiki/Exome_Chip_Design

- FTP Site with Proposed Array Content
  ftp://share.sph.umich.edu/exomeChip/

- Excel Files with Annotated Variant Lists (.csv)
  ftp://share.sph.umich.edu/exomeChip/ProposedContent/codingContent/
  ftp://share.sph.umich.edu/exomeChip/IlluminaDesigns/codingContent/

# Collaborators on the design team

- Goncalo Abecasis*
- David Altshuler*
- Suganti Bala
- Mike Boehnke*
- Candia Brown
- Peter Chines
- Mark Daly*
- Kyle Gaulton
- Goo Jun
- Hyun Min Kang
- Daniel MacArthur
- Mark McCarthy*
- Sean McGee
- Karen Mohlke
- Benjamin Neale
- Debbie Nickerson*
- Shaun Purcell
- Steve Rich*
- Manny Rivas
- Carlo Sidore
- Jen Stone
- Joshua Smith
- Benjamin Voight

* Steering Committee