University of Leicester

The University of Nottingham
UNITED KINGDOM · CHINA · MALAYSIA

UK Biobank Lung Exome Variant Evaluation (UK BiLEVE)

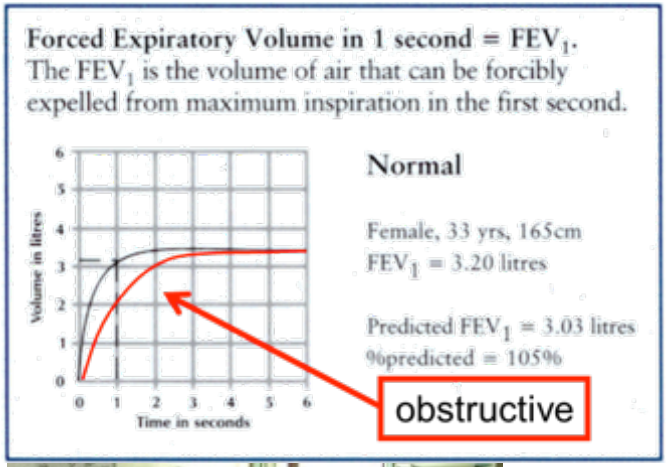
Martin Tobin
Professor of Genetic Epidemiology & Public Health,
MRC Senior Clinical Fellow,
Honorary Consultant in Public Health

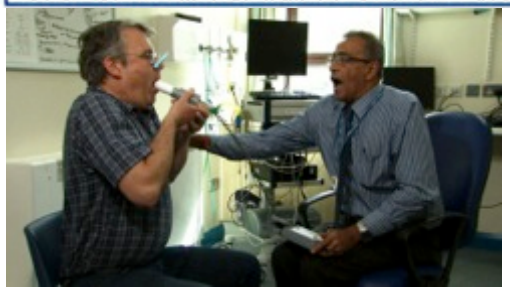
UK Biobank Frontiers Meeting
26th June 2014

Chronic obstructive pulmonary disease (COPD)

- 900,000 diagnosed
- 5th biggest UK killer

  - Environmental / behavioural risk
  - Genetic:
    - α1-antitrypsin mutations (1-2% patients)
    - genome-wide association studies (GWAS)

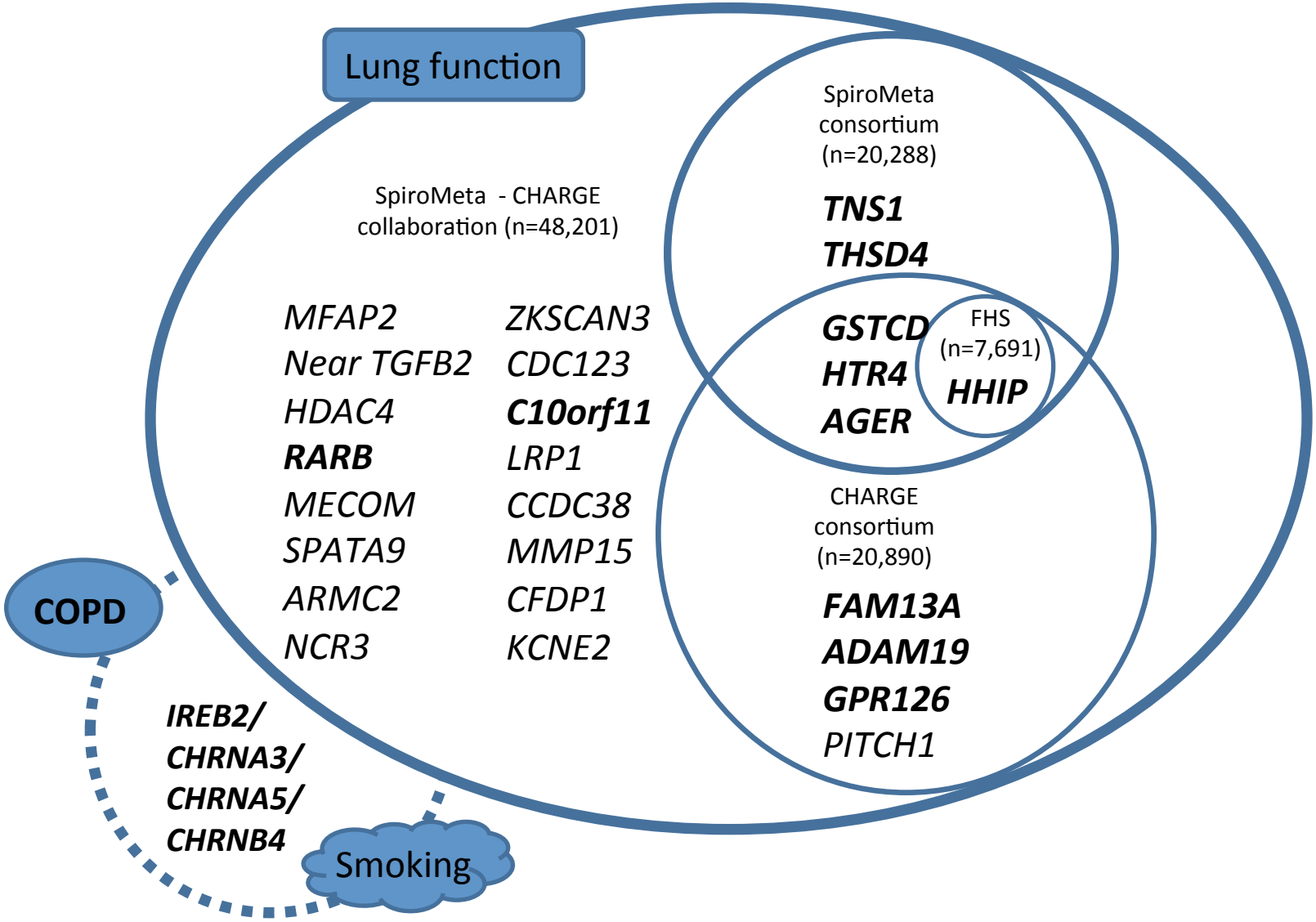




Lung function

- FEV_1= forced expired volume in 1 second (ml)
- FVC = forced vital capacity
- FEV_1/FVC ratio (%)
- Used for diagnosing COPD & for grading COPD severity
- Lung function measures are reliable and heritable

Key genes associated with FEV_1 or FEV_1/FVC & **COPD (bold)**, & the 15q25 locus associated with smoking & COPD



Genetic architecture of lung function

Of the associations at 26 loci:

- proteins/pathways not previously reported in lung disease
  - none driven by tobacco addiction, no smoking interaction
  - most probably affect lung development
- common variants, variable localisation of signals
- up to 7.5% of variance in FEV_1/FVC

Resources

Results (2.5m SNPs): **http://www.gwascentral.org/**

- Repapi et al. Nature Genetics 2010 Jan;42(1):36-44.
- Soler Artigas et al. Nature Genetics 2011 Sep 25;43(11):1082-90
- Soler Artigas et al, Thorax 2012;67:271-273 (review)
- Wain et al. Clin Exp Allergy 2012;42(8):1176-82 (review)
- Loth et al. Nature Genetics 2014, June 15.

Public outreach: **http://sse.royalsociety.org/2012/exhibits**

UK BiLEVE: Aims

Phenotypes define by extremes of the lung function distribution

- heavy smokers
- non-smokers

Common SNPs
MAF>5%

Low frequency SNPs
MAF 1-5%

Rare SNPs
MAF <1%

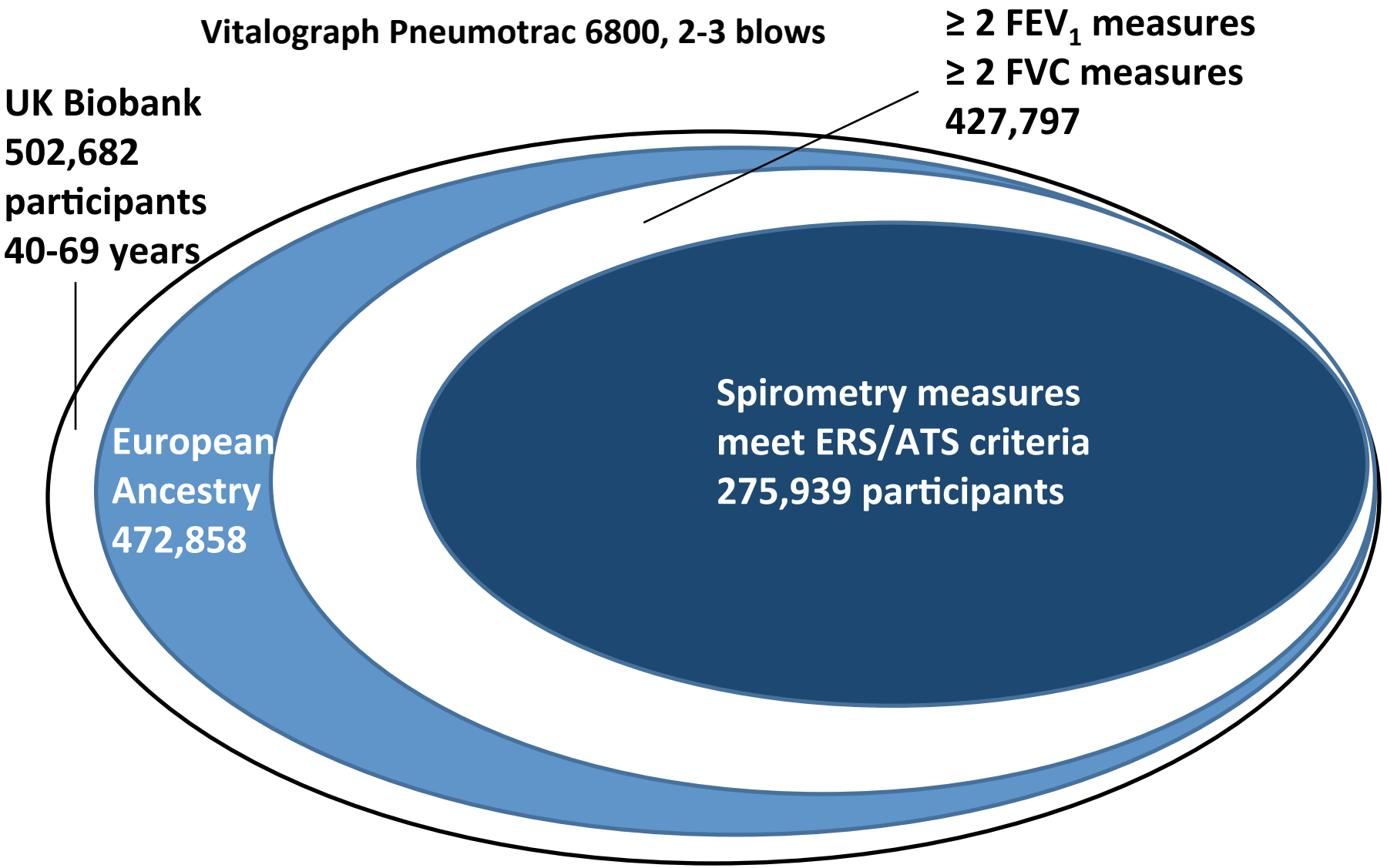
Putative functional variation

Novel regions

Fine map known regions

Independent signals in known regions

UK BiLEVE sampling frame

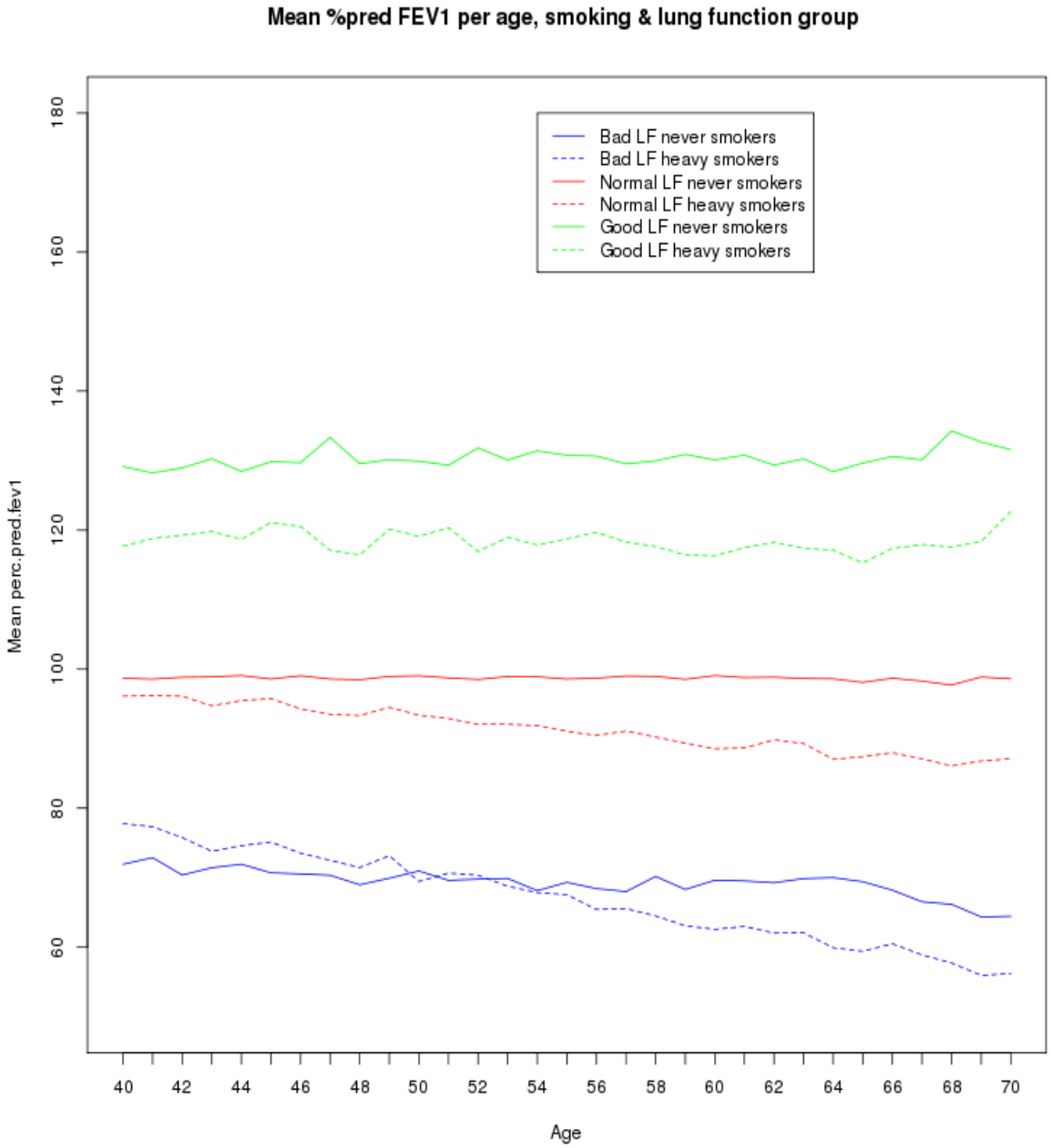


UK BiLEVE design: sampling 50,000 participants

- % predicted FEV_1 calculated using UK Biobank healthy non-smokers as a reference population
- adjusted for height separately in age-sex strata
- "Heavy smokers" - mean 35 pack-years

Smoking status	**% predicted FEV_1**		
	low	**middle**	**high**
Never smokers	10,000	10,000	5,000
Heavy smokers	10,000	10,000	5,000

GOLD Stage 2+ COPD: N=4893


Mean %pred FEV1 per age, smoking & lung function group
Bad LF never smokers
Bad LF heavy smokers
Normal LF never smokers
Normal LF heavy smokers
Good LF never smokers
Good LF heavy smokers
Mean perc.pred.fev1
60
80
100
120
140
160
180
40
42
44
46
48
50
52
54
56
58
60
62
64
66
68
70
Age

Study conduct: outline

- Phenotype data cleaning & sample selection
- UK Biobank DNA extraction
- Randomisation of samples for genotyping
- Array specification and tendering
  - custom Affymetrix array
  - genotyping by Affymetrix in Santa Clara
    - 11 batches, each of 50 plates

Initial design: Customised exome array

- cost-effective coverage of rare, putative functional variants
- subset of common variants informed by genome-wide studies of lung function

UK BiLEVE custom genotyping array design

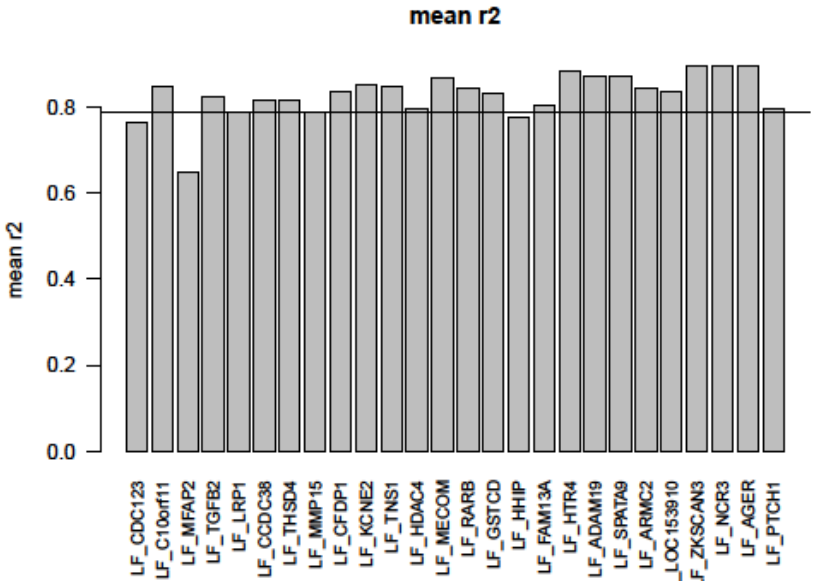
- **Genome-wide "exome chip" style content : 129K**
- Affymetrix exome chip, Illumina exome chip, ExAC European exomes (sequencing data from the Broad Institute)
  - Missense – MAF>0.2% in UK population ~80K*
  - Loss of function – MAF >0.02% in UK population ~32K *
  - Disease-causing – HGMD ~21K *
    - » *categories overlap

- **Genome-wide Imputation "grid" – 642K**
  - Improved imputation in MAF 1-5% range
    - 246K CEU MAF 5-50%
    - 103K EUR MAF 5-50%
    - 293K EUR MAF 1-5%

UK BiLEVE custom respiratory content

- **~9,000 SNPs**
- Optimise coverage of regions of genome-wide significant association with lung function, COPD, asthma, idiopathic pulmonary fibrosis, smoking behaviour
- Variants showing $P<10^{-4}$ in previous genome-wide association studies of lung function, COPD and asthma

Coverage of low-frequency (MAF 1-5%) variants in 26 regions associated with lung function



QC overview

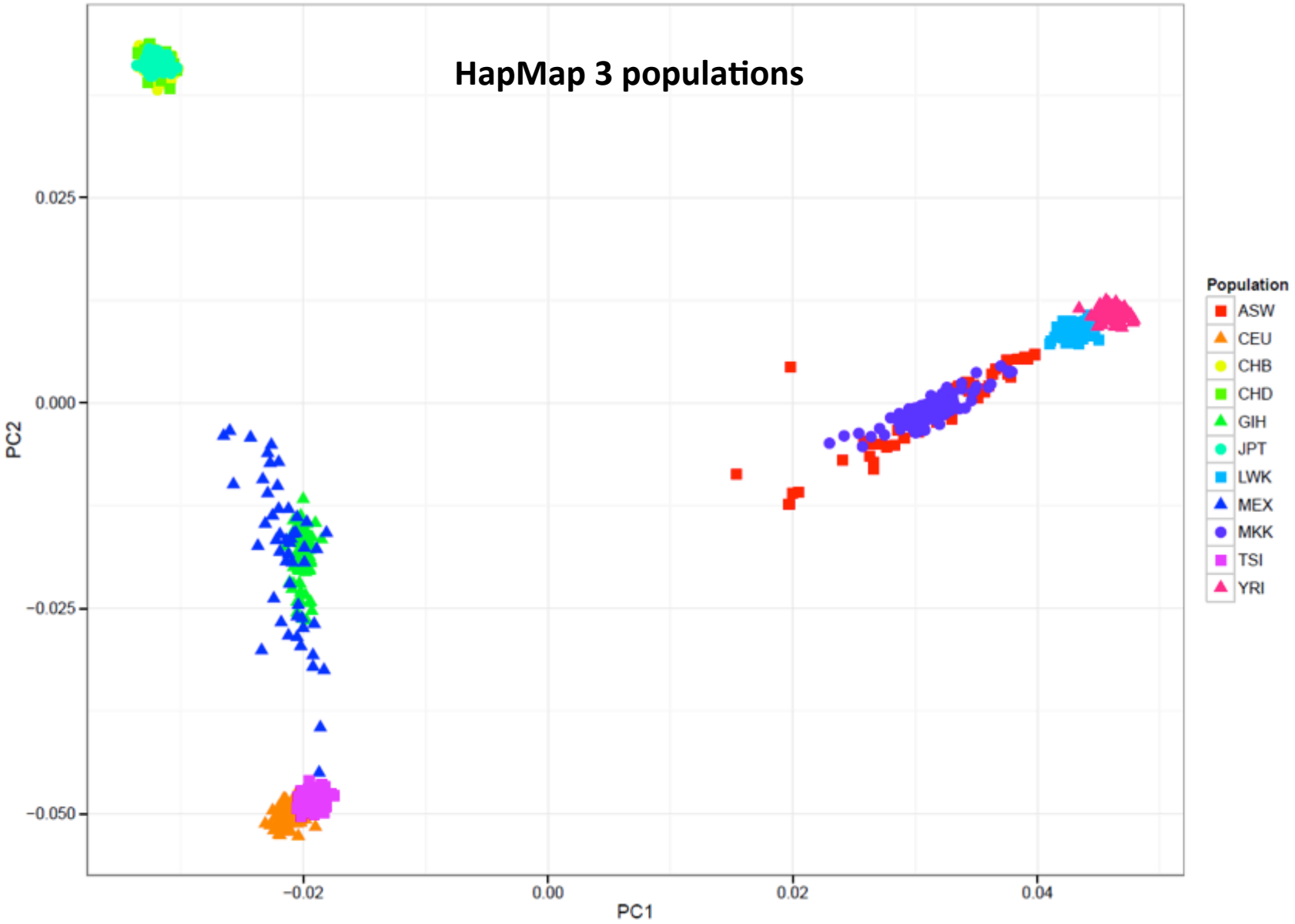
Sample filter	Removed	Remaining
No filters	0	50,561
DNA quality (dQC)	10	50,551
Initial clustering CR<97%	31	50,520
Sex mismatch	125	50,395
Final clustering CR<95%	1	50,394
Heterozygosity outlier	333	50,061
Unintended duplicates	17	50,044
Intended duplicates	481	49,563
PCA outliers	104	49,459
Withdrawn participant	1	**49,458**

Repeat UK BiLEVE samples:

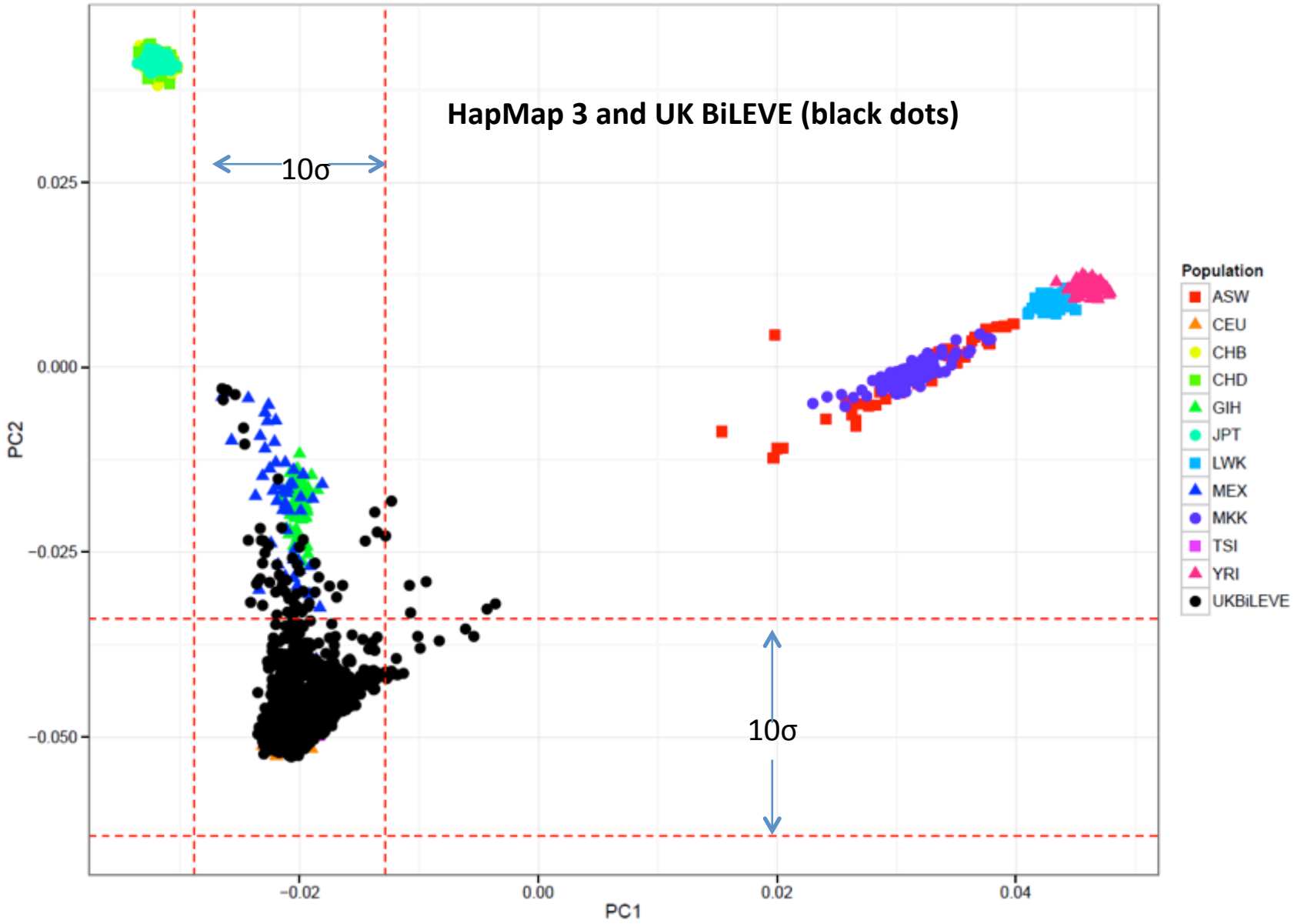
- Median reproducibility 99.95%
- Minimum reproducibility 99.70%.

762,260 SNPs of the 807,411 SNPs on the array (96.9%) passed QC in 9 or more of the 11 batches

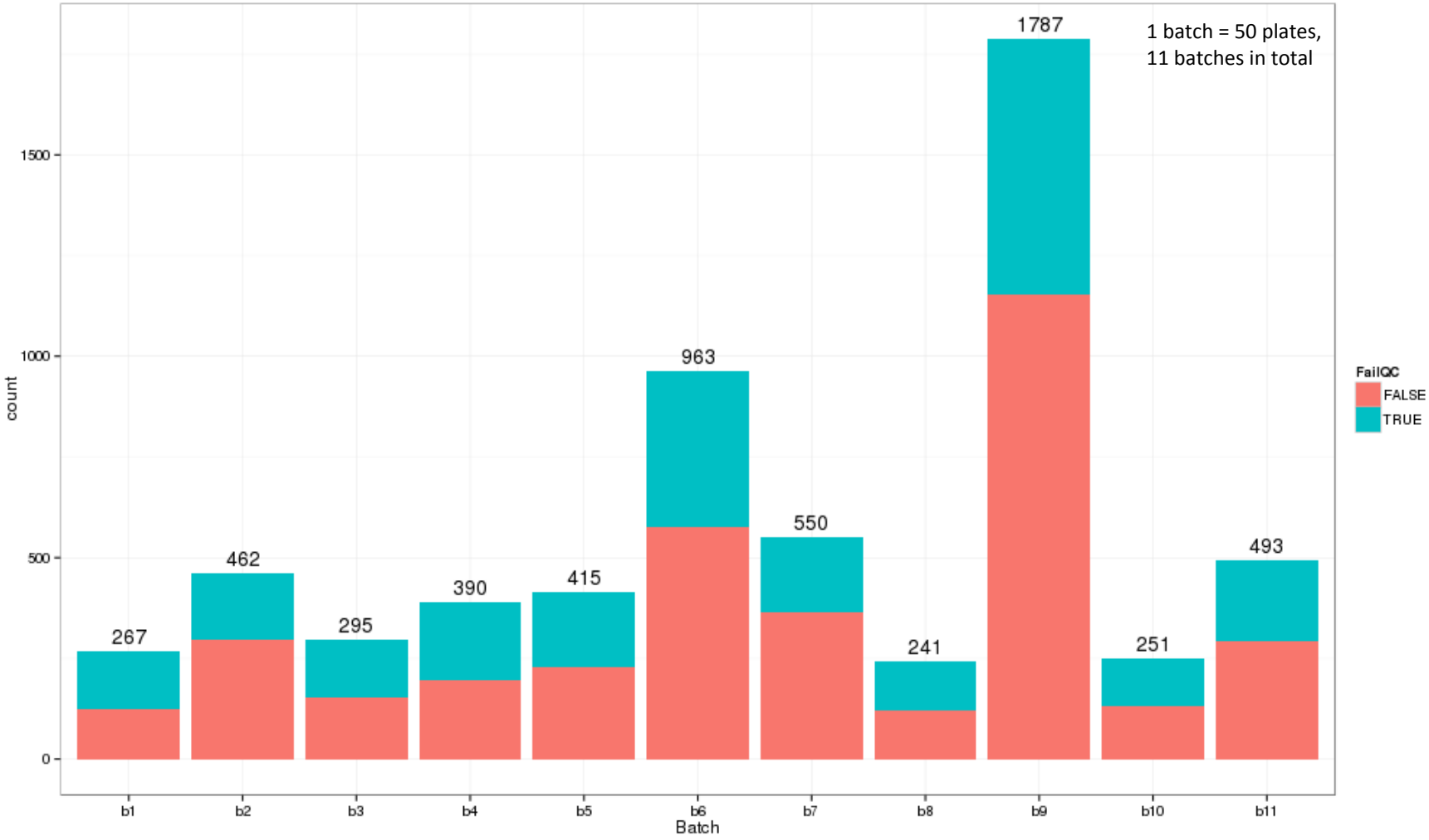
Ancestry: Plot of first 2 principal components



Ancestry: Plot of first 2 principal components



Number of probes showing a plate effect per batch (chisq. $P < 10^{-6}$) and how many would fail a QC filter (HWE $P < 10^{-6}$, Call rate < 95%)



Association with 26 variants previously reported as associated[1] with FEV_1 or FEV_1/FVC

Odds ratio (95% confidence limits) per 1 standard deviation change in effect-size weighted risk score; P-value

FEV_1 risk score	Heavy Smokers	Never Smokers
Good lung function vs normal	0.89 (0.86, 0.92) $P = 1.0\times10^{-10}$	0.88 (0.85, 0.91) $P = 1.9\times10^{-13}$
Bad lung function vs normal	1.10 (1.07, 1.13) $P = 1.7\times10^{-11}$	1.13 (1.09, 1.16) $P = 2.1\times10^{-16}$

1. Soler Artigas et al. Nature Genetics 2011 Sep 25;43(11):1082-90

Association with 26 variants previously reported as associated[1] with FEV_1 or FEV_1/FVC

Odds ratio (95% confidence limits) per 1 standard deviation change in effect-size weighted risk score; P-value

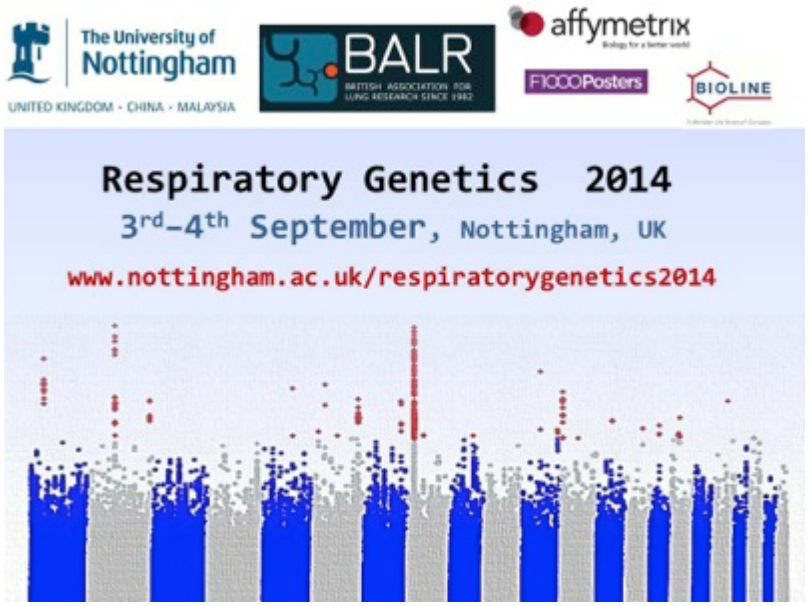
FEV_1 risk score	Heavy Smokers	Never Smokers
Good lung function vs normal	0.89 (0.86, 0.92) $P = 1.0\times10^{-10}$	0.88 (0.85, 0.91) $P = 1.9\times10^{-13}$
Bad lung function vs normal	1.10 (1.07, 1.13) $P = 1.7\times10^{-11}$	1.13 (1.09, 1.16) $P = 2.1\times10^{-16}$
Bad lung function vs good	**1.23 (1.19, 1.27) $P = 1.1\times10^{-31}$**	**1.28 (1.24, 1.32) $P = 5.0\times10^{-44}$**

1. Soler Artigas et al. Nature Genetics 2011 Sep 25;43(11):1082-90

Summary of progress

- 807,000 UK BiLEVE array
  - shares 96% of content with subsequent UK Biobank array
  - we have shared genotype data and shared methods with UK Biobank
  - high quality genotype data
  - good imputation performance
    - 1000g: >22m variants after QC
    - 1000g + UK10K: >31m variants after QC
  - re-calling of very rare genotypes
  - association testing – single variant and gene-based
- Data will be deposited with UK Biobank
- Findings will be made publicly available

Presentation of UK BiLEVE findings



The Genomics of Common Diseases 2014
September 17-20, 2014, Bolger Center, Potomac, MD, USA

BSGM
The British Society for Genetic Medicine

Liverpool, 22-24 September 2014

THE AMERICAN SOCIETY OF HUMAN GENETICS
discover. educate. advocate.

ASHG 2014
San Diego, CA | October 18-22

Acknowledgements

UK BiLEVE consortium: **Ian Hall**, Martin Tobin, Louise Wain, David Strachan. Steering Group: Anna Hansell, Richard Hubbard, Ian Pavord, Neil Thomson, Panos Deloukas. Analysis Group: Andrew Morris, Ele Zeggini, Jonathan Marchini

Analyses: Nick Shine, Maria Soler Artigas, Vicki Jackson, Ioanna Ntalla, James Cook

The unpublished research in the UK BiLEVE consortium study was conducted using **UK Biobank** http://www.ukbiobank.ac.uk/
We thank all scientists, staff and participants contributing to UK Biobank and also thank:
Peter Donnelly, Mark McCarthy and the UK Biobank SNP design group

Affymetrix

MRC Medical Research Council Leading science for better health

SpiroMeta consortium
CHARGE consortium
UK COPD Exome chip consortium

Potential extra slide

UK BiLEVE design: custom genotyping array

Genome-wide genotyping array ~807,000 variants (Affymetrix)

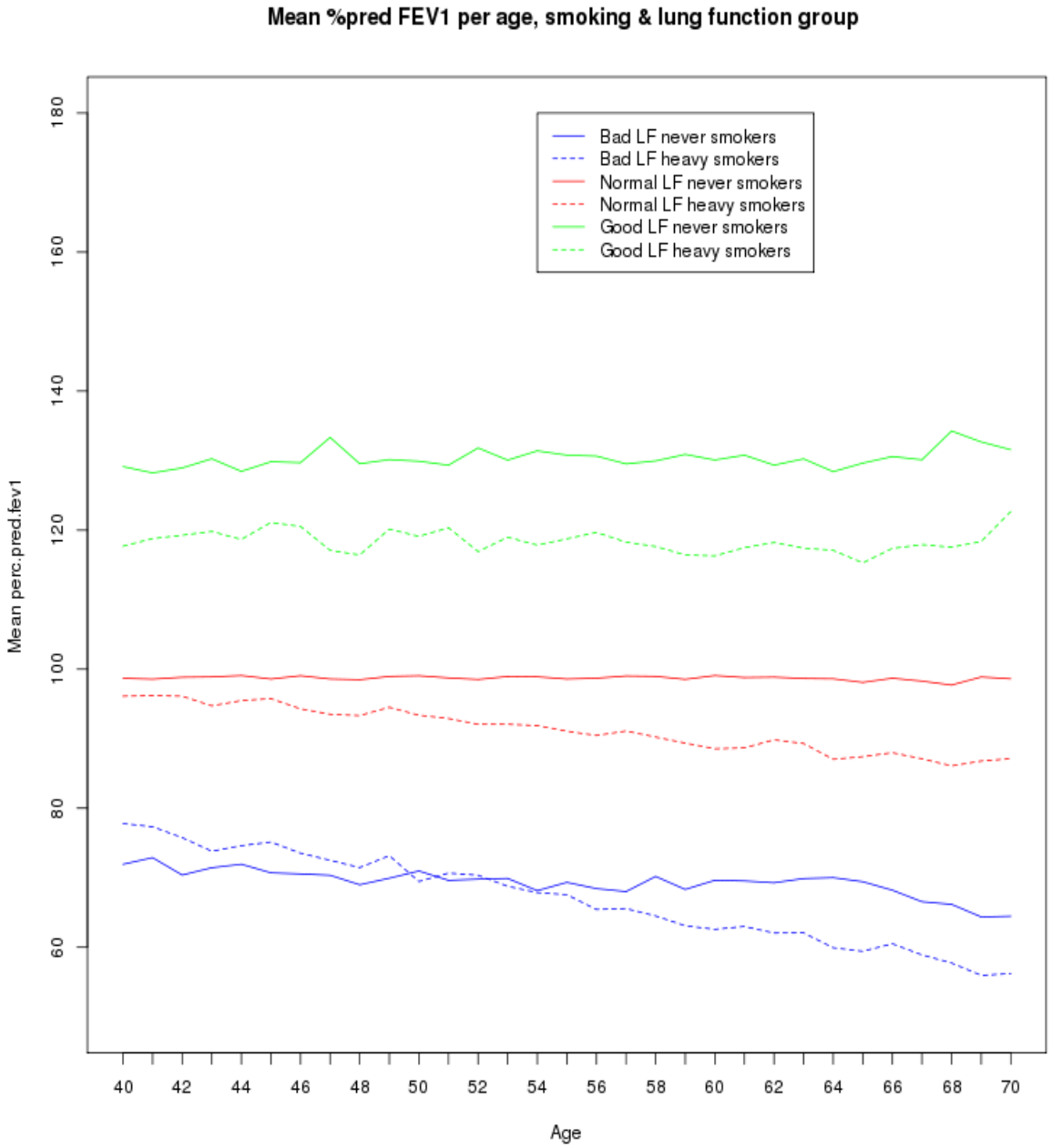
Category	Notes	Markers on Array
GWAS	EUR coverage 1% <= MAF <= 50%	**642k**
Lung Function	Lung function, COPD, asthma, smoking, other	**9k**
Rare Coding	Missense, LoF, HGMD Disease-causing	**129k**
eQTL	Multiple sources	19k
Other	GWAS hits, HLA/KIR, Chr Y & MT, eQTL, ADME	18k
CNV	CNV regions of interest	2k

UK BiLEVE array: imputation based coverage of common and low-frequency variation

MAF range	# markers	# r^2>0.8	Coverage (# r^2>0.8/# markers)	Mean r^2 (all markers)
5-50%	6,777,946	6,107,870	**0.901**	**0.919**
1-5%	2,950,263	1,993,498	**0.676**	**0.787**

UK BiLEVE array imputation coverage of ~9.7K 1000 genomes variants in EUR population (union of CEU, GBR, FIN, IBS, TSI), data provided by Affymetrix

Improved imputation of low frequency variants with a larger reference panel


Mean %pred FEV1 per age, smoking & lung function group
Bad LF never smokers
Bad LF heavy smokers
Normal LF never smokers
Normal LF heavy smokers
Good LF never smokers
Good LF heavy smokers
Mean perc.pred.fev1
60
80
100
120
140
160
180
40
42
44
46
48
50
52
54
56
58
60
62
64
66
68
70
Age