Perils and Promise

Peter Donnelly
The Wellcome Trust Centre for Human Genetics
University of Oxford

Challenges

1. Data access
2. Analysis methods
3. Sample sizes

Data Access

Critical that for large publicly funded sequencing efforts data should be widely available.

That means sequence data and rich phenotype data.

We need joined-up thinking aiming for a large database where we can link phenotypes with all (or most) known sequence variants.

Analysis Challenges

Methods for calling variants from sequence data at varying stages of maturity: SNPs, indels, CNVs.

Annotation of variants.

Linking sequence and (rich) phenotype data.

We do not yet know the best methods for analysing sequence data for association even with simple phenotypes (case-control status, univariate QTLs).

The Curious Incident ….

In *Silver Blaze*, Inspector Gregory said to Sherlock Holmes:

“Is there any point to which you would wish to draw my attention?”
“To the curious incident of the dog in the night-time.”
“The dog did nothing in the night-time.”
“That was the curious incident,” remarked Sherlock Holmes.

Large sequencing studies for complex traits

ESP: 7,000 exomes
GO T2D: 3,000 exomes + low pass genomes
T2D-GENES: 5,000 exomes
Autism: 4,400 exomes
Schizophrenia: 5,000 exomes
(Cancer: ~10,000 normal exomes, probably not analysed for germline variation)

Extensive imputation from 1,000 Genomes and other reference panels into very large GWAS studies.

Evidence against (widespread) “Goldilocks” mutations: mutations at low frequency (say 0.5% - 3%) with moderate or large effect sizes.

Evidence that (very) large sample sizes (+ follow-up) will be needed to uncover genes harbouring rare variants which affect disease susceptibility.

We have no idea how important rare sequence variation will be for common disease phenotypes.

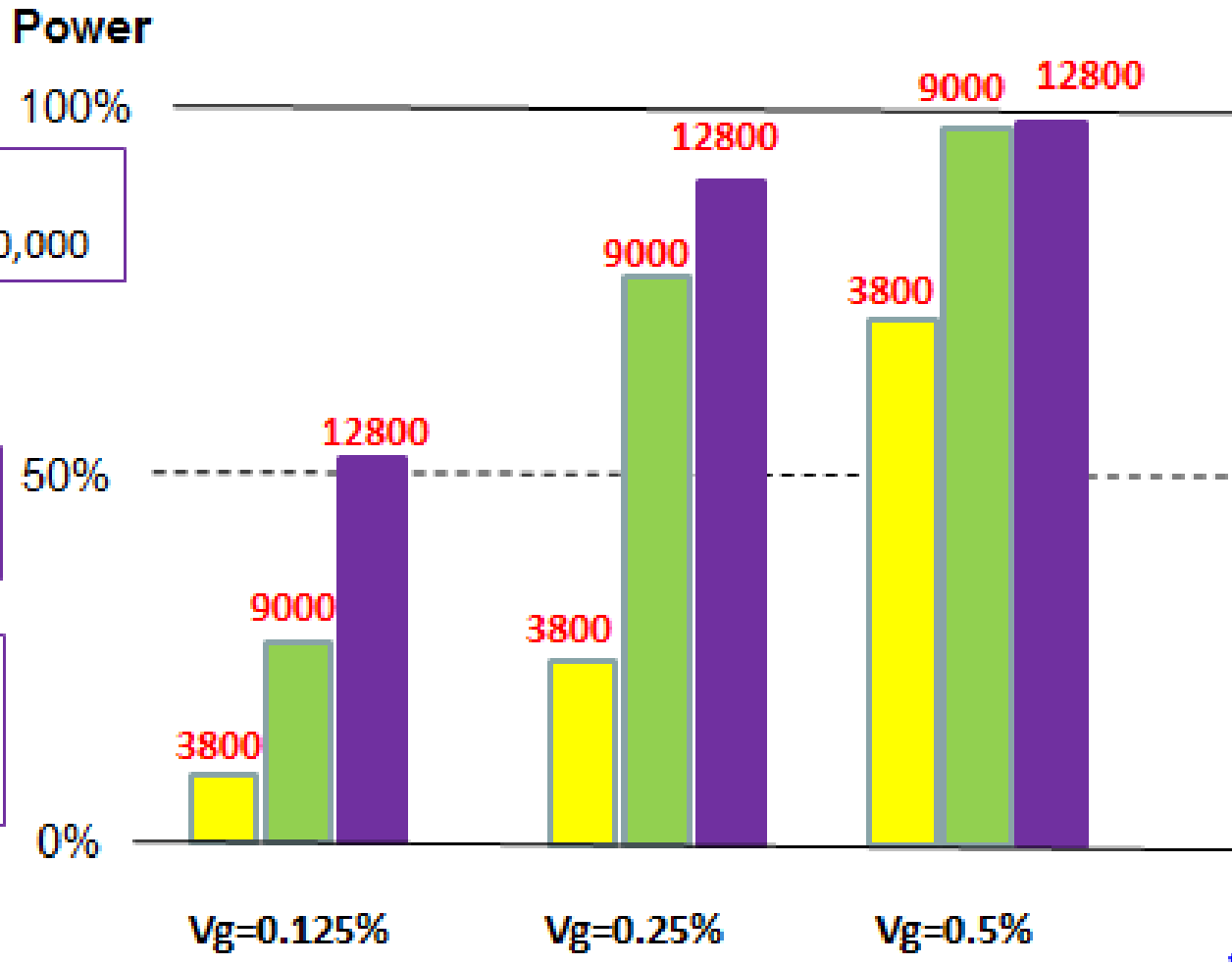
Power for exome sequencing.....

Stage 1: 3800 – 12800 exomes
Stage 2: targeted sequencing 20,000

C-alpha burden test

$\alpha_1 = 0.005$ (100 transcripts)
$\alpha_{ft} = 2.5 \times 10^{-6}$ (20k transcripts)

Power by simulation
SFS = European 12k exomes
Transcript size 1.5kb



Manny Rivas
Loukas Moutsianis

Slide courtesy of Mark McCarthy