Identifying drug-targetable key drivers of disease

Expression data

Phenotypes

Public data
‘To capture something small you need something big’
DNA

ACGT
‘To capture something small you need something big’
‘To capture something small you needed something big’
large amounts of data now available
Goal: better diagnose and treat patients
Seven years of GWAS studies

6,054 disease associations

Teri Manolio et al: A catalog of published genome-wide association studies
Genes unknown
Pathways unknown
Cell-types unknown

Genetic risk factors

>10,000 known

Black Box

Disease

>200 diseases
**Expression quantitative trait locus (eQTL)**

### Cis-eQTL

- **Gene X**
  - SNP A/G
  - 5' promoter region → exon 1 → intron 1 → exon 2 → 3'

### Trans-eQTL

- **Gene X**
  - Coding SNP A/G
  - 5' promoter region → exon 1 → intron 1 → exon 2 → 3'

- **Protein X**
  - Amino acid change

- **Gene Y**
  - 5' promoter region → exon 1 → 3'
Far majority of genetic risk factors affect gene expression

Dubois et al, Nature Genetics 2010
Fehrmann et al, PLoS Genetics 2011
Fu et al, PLoS Genetics 2012
Westra et al, Nature Genetics 2013
Get larger sample-sizes: meta-analysis in 5,311 samples

Systemic lupus erythematosus risk factor: Chr. 7

Local expression effect: IKZF1 Chr. 7

Type 1 interferon response: (in Monocytes)

Downstream effects identified for >200 genetic risk factors

New meta-analysis ongoing in 25,000 blood samples

Downstream *trans*-eQTL effects

Westra et al, Nature Genetics 2013
**Goal**

Genome-wide association studies

**cis-eQTL mapping**

**trans-eQTL mapping**

Key driver gene identification
Possible to identify all these downstream effects?

This is not going to be possible!
- Massive sample-sizes required
- Many cell-types required
- Genotype and gene expression data required from the same samples
Lifelines Deep (1500 samples)

- Methylation
- Phenotypes
- DNA
- RNA-seq
- Microbiome
- Metabolites
The opportunities

Methylation

DNA

RNA

Gut bacteria

Organic Compounds

Phenotypes

Metabolites

P = 10^{-21}

Risk factor

Methylation gene X
Trans-meQTL meta-analysis in 3,840 samples

- 34.4% of 405,709 tested CpG sites are *cis*-meQTL (FDR < 0.05)

- 31.2% of established GWAS risk factors give *trans*-meQTL effect (FDR < 0.05). 1,907 SNPs affecting 10,141 unique CpG sites in *trans*

- *Trans*-meQTL replicate in monocytes: 95% identical allelic direction

- *Trans*-SNPs affect expression of nearby TFs, subsequent methylation of downstream targets of these TF
Trans-meQTL meta-analysis in 3,840 samples

NFkB1
4: 103,422,486-103,538,459
rs3774937
4:103,434,253

Risk factor associated to
Ulcerative Colitis

Bonder et al, BiorXiv preprint
Trans-meQTL meta-analysis in 3,840 samples

Bonder et al, BiorXiv preprint
Detecting cell-type dependent eQTLs in whole blood

**NOD2** eQTL in whole peripheral blood

- **C/C**
- **C/T**
- **T/T**

Leprosy risk SNP rs1981760

- **P = 1.11 \times 10^{-294}**

**STX3**

- **low**
- **high**

**NOD2** eQTL interaction analysis, **STX3** interacts with rs1981760

- Interaction **P = 1.1 \times 10^{-69}**

- **C/C**
- **C/T**
- **T/T**
Context specific cis-eQTL analysis in 2,116 samples

Zhernakova et al, BiorXiv preprint
Context specific cis-eQTL analysis in 2,116 samples

Zhernakova et al, BiorXiv preprint
Regulatory network reconstruction in 2,116 samples

Zhernakova et al, BioRxiv preprint
but is this relevant for my patients?
But what about patients we see?

Patient with a severe disease. You suspect a genetic cause. What do you do?

- Targeted gene panel?
- Whole exome sequencing?
- Whole genome sequencing?

Problem:
Many (rare) variants of unknown significance
Smart ways to filter?

___ gene expression?
- Rare genetic variants also have effects on gene expression

- Rationale BBMRI-NL BIOS Consortium to establish ‘Transcriptome of the Netherlands’ in 5,000 population based samples

- Generate RNA-seq data on patients. Contrast these expression values to the Transcriptome of the Netherlands.
Essential to get very accurate reference values for each gene.
Remove non-genetic expression variation

Most expression variation due to:
- Physiological state
- Metabolic state
- Environmental state

RNA blood expression when you wake up

RNA blood expression after nice diner
Strategies

- Get large sample-sizes
- Get many different cell-types
- Recycle big data
Amplifier can change many aspects of music

Setting: Activity of switch
Size of switch: Importance of switch

Wiring: Way the switch has effect
A control panel that determines gene expression?

**Setting:** State of a certain sample

**Size of switch:** Importance

**Wiring:** Effect on individual genes

**Regulatory factors:**
- Hormones,
- Transcription factors,
- Physiological factors,
- Other (external) stimuli
- Genetic variation

Fehrmann *et al*, Nature Genetics 2015
800 ‘transcriptional components’: Component 1 - 50

Components 1 - 50:
Physiology, metabolism, cell-type differences

Fehrmann et al, Nature Genetics 2015
Component 1 and 2

Transcriptional Component 1

Transcriptional Component 2

Cell Line Samples

Blood Samples

Primary Tissue Samples

TC 1

TC 2

TC 1

TC 2
Predicted gene functions: www.genenetwork.nl

Fehrmann et al, Nature Genetics 2015
GWAS on red blood cell traits:
Mean hemoglobin concentration: rs1175550*G

Blood eQTL mapping:
SMIM1: Expression Levels
AA AG GG
P < 10^{-16}

Gene function prediction:
(GeneNetwork.nl, based on 80,000 RNA microarrays)
Genes known to be involved in hemoglobin metabolism

Exome sequencing of individuals, negative for Vel bloodgroup antigen:
Homozygous 17bp deletion in SMIM1

Knock-down in zebrafish:
Reduced number of red blood cells

Van der Harst et al, Nature 2012
Cvejic et al, Nature Genetics 2013

Amounts of data integrated:
GWAS in 135,000 samples
eQTL mapping in 1,500 samples
Transcriptomics in 80,000 samples
Exome sequencing
Wet lab proof
DEPICT: New prioritisation algorithm for GWAS

697 significant adult height associations:

Wood et al, Nature Genetics 2014

DEPICT Method:
Pers et al, Nature Communications 2015

DEPICT used for:
Body mass index (Locke et al, Nature 2015)
Waist hip ratio (Shungin et al, Nature 2015)
Hypospadias (Geller et al, Nature Genetics 2014)
Lipid Levels (Surakka, Nature Genetics 2015)
Some component show weird behaviour

TC 165: Strong cytogenetic effects, high autocorrelation

Redo analysis in healthy samples, correct cancer data for healthy components
Detection cytogenetic aberration in expression data

Chromosome

Down Syndrome patient: dup 21
Identifying five chromosome duplications

Karyogram
HapMap LCL

Chromosome 4 7 9 14 21
Comparison of arrayCGH and cytogenetic RNA profiles

Fehrmann et al, Nature Genetics 2015
By recycling big data it is possible to clean data and get very accurate measurements.
Amount of cytogenetic aberrations

Tipping point at component 165

Transition to chaos in the logistic map
Crutchfield et al, 1990

Distribution identical to simulations in complexity theory
Forest fire: when will a forest burn down entirely? How many trees can you plant without the risk that everything burns down?
Complexity: Forest fire

Percentage of land filled with trees

Expected effect?
Complexity: Forest fire

Percentage of land filled with trees after forest fire

Percentage of land with living trees

Tipping point

20%
Complexity: Forest fire

Percentage of land filled with trees

Percentage of land with living trees after forest fire

Tipping point

20%

40%
Complexity: Forest fire

Percentage of land filled with trees after forest fire

- 20%
- 40%
- 80%

Percentage of land with living trees

- 0%
- 50%
- 100%
Complexity: Forest fire

Percentage of land filled with trees after forest fire:

- 20%
- 40%
- 60%
- 80%

Percentage of land with living trees:

- 0%
- 50%
- 100%
More accurate reference values for genes

TRIM51BP gene expression distribution in the Dutch population

Log$_2$ expression vs. Number of samples
Explosion of publicly available RNA-seq data

9,527 public human RNA-seq runs from ENA

Read alignment, expression quantification, normalization and PCA:
- 4,028 runs with low mapping statistics removed
- 521 expression outliers removed

4,978 samples (used for expression clustering)
Derive SNP genotypes from RNA-seq data

Public RNA-seq data (5,000 samples)

Infer genotypes

rs1136055

P = 10^{-21}

Deelen et al, Genome Medicine 2015
GATK to call genotypes and output genotype likelihoods, BEAGLE used for imputation towards Genome of the Netherlands

Deelen et al, Genome Medicine 2015
Ability to call SNP is largely dependent on expressed transcripts

Deelen et al, Genome Medicine 2015
Tissue-specific eQTL mapping for free

Deelen et al, Genome Medicine 2015
Allele specific effects for rare variants

rs12203592 – IRF4
p-value: 4.06 x 10^-53

rs72550870 – MASP2
p-value: 5.07 x 10^-15

Functional class annotation
Wilcoxon p-value: 1.36 x 10^-6

Estimated percentage alternative allele

Log2 expression
Public RNA-seq data: (5,000 samples)

Deelen et al, Genome Medicine 2015

Westra et al, Nature Genetics 2013

Genotype calling enables functional effect analysis of:

- Common variants: Expression quantitative trait loci
- Rare variants: Allele specific expression

Gene expression profile, corrected for 'transcriptional components':

Fehrmann et al, Nature Genetics 2015

Gene expression levels corrected for healthy physiological and metabolic variation

Apply methodology to Individual patients

Apply methodology to Transcriptome of the Netherlands (5,000 samples)

Patient has certain phenotypes:
- Seizures
- Short stature

Candidate causal gene

Very low TRIM51BP expression in patient

TRIM51BP gene expression distribution in the Dutch population

Candidate causal gene

TRIM51BP likely causal gene
Integration of different datasets

**Lifelines Deep**

- 1,500 samples
- Many omics levels
- Genotype data
- Extensive phenotyping

**Transcriptome of the Netherlands**

- 5,000 samples
- RNA-seq data
- Genotype data
- Methylation 450k data

**Public RNA-seq data**

- 25,000 samples
- RNA-seq data
- Genotype data
Enormous opportunities exist when recycling ‘big data’, permits gaining insight into downstream consequences of (rare) genetic variants

Workshop: how to conduct these analyses yourself:
- Pointers to the software that is available
- Identifying sample mix-ups
- Correcting for unknown confounders
- Multiple testing correction
- Allele specific expression
Acknowledgements

UMC Groningen

Juha Karjalainen
Dasha Zhernakova
Patrick Deelen
Marc Jan Bonder
Sipko van Dam
Morris Swertz

Freerk van Dijk
Niek de Klein
Urho Vosa
Annique Claringbould
Rudolf Fehrmann
Cisca Wijmenga

Peter-Bram ’t Hoen
Tonu Esko
Haije Wind

Bas Heijmans
Tune Pers

Funding

ERC
NWO
BBMRI-NL