Complex Genetics: From GWAS to HT Sequencing

André G Uitterlinden
Genetic Laboratory
Department of Internal Medicine
Department of Epidemiology
Department of Clinical Chemistry
www.glimdna.org

Why should we study DNA variation?

- **Mechanism**: understand cause of disease
- **Treatment**: finding new potential drug targets
- **Diagnostics**: understand how DNA variation contributes to variation in:
  - Risk of disease (vulnerability):
    - "personalized medicine"
  - "Response-to-treatment" (medication, diet):
    - "pharmacogenetics-pharmacogenomics"
**Determine “Functional” Effects of DNA Polymorphisms**

**DNA polymorphism**

- **Organizational Level:**
  - mRNA: level, stability, splicing/isoforms
  - Protein: level, stability, isoforms, protein-protein
  - Cells: e.g., transcriptional activity, e.g., Cell growth inhibition
  - Humans: -Serum parameters, -Intervention

**Association with disease:** > 70 yrs follow-up!

---

**“SINGLE NUCLEOTIDE POLYMORPHISM” = “SNP”**

- Frequent in the Genome:
  > 30 million SNPs, in/del, CNV
  (dbSNP, HapMap, 1000G)
- Frequent in the Population:
  > 5% = polymorphism
  0.01 - 5% = rare variant
  <0.01% = mutation

---

**“The Human Genome Project”**

- *26 June 2000: Press conference Bill Clinton & Tony Blair: “working draft”, 95% gesequenced*  

>>Cheaper and Faster!!

Costs: $ 2.7 milliard (instead of $ 3 billion estimated costs)

---

**What will DNA tell us about this...**

Stain in a dress
SNPs, alleles, genotypes and haplotypes

SNP = Single Nucleotide Polymorphism

Genotype

Haplotype Allele

Chromosomes:
- from Father
- from Mother

Single Nucleotide Polymorphisms (SNPs) are common and have subtle effects.

- AACCG
  - C
  - ATAAGG
  - ..TTGGC
  - G
  - TATTCC
  - ..AACCG
  - T
  - ATAAGG
  - ..TTGGC
  - A
  - TATTCC

DNA: C677T protein: Ala222Val

Population frequency:
- 65%
- 35%

This is what happens when there are **NO POLYMORPHISMS**

Why is the study of polymorphisms important?
- Evolution
- Disease
- Forensics

Twin Studies Demonstrate Heritability

Heritable diseases and traits:
- Diabetes
- Rheumatoid arthritis
- Breast cancer
- Lung cancer
- Osteoarthritis
- BMI
- Menopause
- Weight
- Height
- Menarche
- Infidelity
- Cholesterol
- Entrepreneurship
- Uric acid
- Paget’s Disease
- Infectious disease susceptibility
- Depression
- Ankylosing spondylitis
- Eye colour
- Myocardial Infarction
- Osteoporosis
- Skin colour
- Longevity
- Stroke
- Eye diseases
- Baldness
- Telomere length
- Smoking behaviour
- Etc.
The influence of “technology-push”

Time needed for genotyping 1 SNP in 7,000 DNA samples of the Rotterdam Study

<table>
<thead>
<tr>
<th>Year</th>
<th>Time Needed</th>
<th>Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>1996</td>
<td>6 months</td>
<td>RFLP, Epp tubes</td>
</tr>
<tr>
<td>1999</td>
<td>3 months</td>
<td>RFLP, 96-well plates</td>
</tr>
<tr>
<td>2001</td>
<td>1 week</td>
<td>SBE, 384-well plates</td>
</tr>
<tr>
<td>2003</td>
<td>1 day</td>
<td>Taqman (manual)</td>
</tr>
<tr>
<td>2004</td>
<td>6 hrs</td>
<td>Taqman, Caliper pipetting robot</td>
</tr>
<tr>
<td>2005</td>
<td>3 hrs</td>
<td>Taqman, Deerac, “Fast” PCR</td>
</tr>
<tr>
<td>2007</td>
<td>6 sec</td>
<td>Illumina 550K array, 600 DNAs/week</td>
</tr>
<tr>
<td>2010</td>
<td>&lt; 0.6 sec</td>
<td>Illumina HiSeq2000 Sequencers</td>
</tr>
</tbody>
</table>

Genetic Architecture of Complex Diseases/Traits: Study designs to identify “risk” alleles

Genetic architecture of traits

<table>
<thead>
<tr>
<th>Effect Size</th>
<th>Frequency Genetic Variant</th>
</tr>
</thead>
<tbody>
<tr>
<td>small</td>
<td>rare, monogenic</td>
</tr>
<tr>
<td>big</td>
<td>(very) few examples</td>
</tr>
<tr>
<td></td>
<td>common, complex</td>
</tr>
</tbody>
</table>

Genetic Association Analysis (1)

Case-control design

Test for “association” by counting variants of a (candidate) gene.

Compare allele frequencies: A = wild-type allele; B = risk-allele

<table>
<thead>
<tr>
<th>Genotype</th>
<th>CONTROL-group</th>
<th>DISEASE-group</th>
</tr>
</thead>
<tbody>
<tr>
<td>AAAA</td>
<td>70%</td>
<td>50%</td>
</tr>
<tr>
<td>AA</td>
<td>30%</td>
<td>50%</td>
</tr>
<tr>
<td>BB</td>
<td>50%</td>
<td>50%</td>
</tr>
</tbody>
</table>

Genetic Association Analysis (2)

Quantitative Trait analyses

Humans are diploid: compare characteristics by their genotype

Genotype | mean Femoral Neck BMD |
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>AA</td>
<td>0.82±0.12</td>
</tr>
<tr>
<td>AB</td>
<td>0.80±0.13</td>
</tr>
<tr>
<td>BB</td>
<td>0.78±0.13</td>
</tr>
</tbody>
</table>

Population (-based sample)
A single-centre, prospective population-based cohort study, started 1990
- Base-line cohort = 7,983 men and women of age ≥ 55 yrs
- In 2007: 4 Follow-up measurements: ~1500 per subject each time
- Ethnically homogeneous: 99% Caucasian
- Computerized GP + pharmacy monitoring
- Study determinants and prevalence/incidence of chronic and disabling disease in the elderly: CVD, Neurodegenerative Disease, Endocrine diseases, Locomotor disease (osteoporosis, osteoarthritis), Eye
  End 2004:
  - 1200 coronary heart disease
  - 800 stroke
  - 1300 fractures
  - 1000 maculopathy
  - 800 dementia
- ~12,000 DNA samples available:
  1990: ERGO base-line/ RSI : n=7,000
  2000: ERGO plus/ RSII : n=3,000 (55+)
  2004: ERGO young/ RSIII : 3,500 (45+)

“ERGO”: The Rotterdam Study

Fracture is a “complex” phenotype:
Clinical Expression: Fracture Risk
- Bone Strength
- Impact Force
- Fall Risk
- (Osteoporosis)
  - BMD
  - Quality
  - Geometry

Risk Factors:
- DNA mutations and polymorphisms

Environmental factors: diet, exercise, sun exposure, ...

Environmental influences can differ between populations!

HOLLAND → BELGIUM
Geographical distance: <100km

Dietary Calcium intake
- > 1100 mg/day
- < 500 mg/day

GR, 11β-HSD
Cortisol
TSHR, Dio1, Dio2, Dio3, Mct8
IGF/GH

VDR, DBP
Thyroid Hormone
ERα, ERβ, Aromatase
LH, LHR, GnRH

Vitamin D

Genetic determinants of osteoporosis?

Matrix molecules
- Col1a1
- osteocalcin
- ANISQ, LOX

homocysteine

TGFB/BMP/Wnt-signalling
- TGFβ, LRP5/6
- BMP2, ERβ
- SOST

MTHFR, MS, MTRR, CBS, THYS

Thyroid Hormone
- TSHR, Dio1, Dio2, Dio3, MCT8
- IGF/I, IGFBP3

IGF/GH

Estrogen

VDR, DBP
Thyroid Hormone
ERα, ERβ, Aromatase
LH, LHR, GnRH

Vitamin D

Genetic determinants of osteoporosis?

Matrix molecules
- Col1a1
- osteocalcin
- ANISQ, LOX

homocysteine

TGFB/BMP/Wnt-signalling
- TGFβ, LRP5/6
- BMP2, ERβ
- SOST

MTHFR, MS, MTRR, CBS, THYS

Thyroid Hormone
- TSHR, Dio1, Dio2, Dio3, MCT8
- IGF/I, IGFBP3

IGF/GH

Estrogen

VDR, DBP
Thyroid Hormone
ERα, ERβ, Aromatase
LH, LHR, GnRH

Vitamin D

Genetic determinants of osteoporosis?

Matrix molecules
- Col1a1
- osteocalcin
- ANISQ, LOX

homocysteine

TGFB/BMP/Wnt-signalling
- TGFβ, LRP5/6
- BMP2, ERβ
- SOST

MTHFR, MS, MTRR, CBS, THYS

Thyroid Hormone
- TSHR, Dio1, Dio2, Dio3, MCT8
- IGF/I, IGFBP3

IGF/GH

Estrogen

VDR, DBP
Thyroid Hormone
ERα, ERβ, Aromatase
LH, LHR, GnRH

Vitamin D

Genetic determinants of osteoporosis?
Candidate gene association analysis “in practice”:

Many controversial & ir-reproducible results because of:
- Small sample size
- Ill-defined choice of polymorphisms
- Lack of standardized genotyping
- Lack of standardized phenotype data
- Publication bias

>> How to improve?
- Combine study populations (across Europe, globally): meta-analysis
- Rationalise choice of polymorphisms: functionality, haplotypes
- Standardize genotyping methods: reference DNA plate
- Standardize phenotypes across populations: meta-analysis individual level data
- Run prospective meta-analyses

Grades of Evidence

Very Good
- Collaborative prospective meta-analysis of individual level data in consortia
- Meta-analysis of published data
- >2 large studies (n > 1000 each)
- 1-3 smaller studies

Not so Good
- 1 small study (n<500)

“GENOMOS”
a large-scale, multi-centre study for prospective meta-analyses of osteoporosis candidate gene variants

“Genetic Markers for Osteoporosis”
EU FP5 sponsored: 3 mio euro
Jan 2003 – Jan 2007

Genes analysed:
- ESR1
- COLIA1
- VDR
- TGFb
- LHRP&4

Total number of subjects (early 2006):
26,264
18,405 women
7,859 men
6,498 fractures
2,380 vertebral fx

EUROPE by prejudice........(according to USA)
(From: Yanko Tsvetkov, alphadesignar.com)
GENOMOS RESULTS (March 2008)

<table>
<thead>
<tr>
<th>GENE</th>
<th>SNP n</th>
<th>Sample n</th>
<th>BMD</th>
<th>FX</th>
</tr>
</thead>
<tbody>
<tr>
<td>ESR1*</td>
<td>3</td>
<td>18,917</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>COL1</td>
<td>1</td>
<td>20,766</td>
<td>0.15 SD</td>
<td>0.15 SD</td>
</tr>
<tr>
<td>VDR</td>
<td>5</td>
<td>26,242</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>TGFb</td>
<td>5</td>
<td>28,924</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>LRP6*</td>
<td>2</td>
<td>37,760</td>
<td>0.15 SD</td>
<td>0.15 SD</td>
</tr>
<tr>
<td>LRP5</td>
<td>1</td>
<td>37,760</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

**PUBLICATIONS:**
- Ioannidis et al., JAMA 2004
- Uitterlinden et al., Ann Int Med 2006
- Langdahl et al., Bone 2008
- van Meure et al., JAMA 2008

* = hit in GEFOS 1st GWAS metas-analysis

---

**EU-FP7 project: GEFOS (2008-2012)**

Number of subjects:
- GENOMOS: >150,000
- of which GWAS: 40,000
- www.gefos.org

---

**Genome-Wide Association Study (GWAS)**

**DATA ANALYSIS (e.g., PLINK):**
- Replication
- Illumina          Affymetrix
- Genome-Wide Association Study (GWAS)
- Select SNPs
- DNA collection: e.g. 1000 cases vs. 1000 controls
- Each dot is one SNP in, e.g, 2000 subjects
- Combine GWAS
- Meta-Analysis of all data

**GEFOS HYPOTHESIS-FREE GWAS**

LUMBAR SPINE BMD

- Rotterdam Study
- ERF Study
- Twins UK
- deCODE Genetics
- Framingham Study

N=5,000

Rivadeneira et al., Nat Genet., 2009
The "CHARGE" consortium

"Cohorts for Heart and Aging Research in Genomic Epidemiology"
Approved by participating cohort studies Jan 31, 2008, Boston, USA

<table>
<thead>
<tr>
<th>Cohort</th>
<th>Acronym</th>
<th>Contact (RSC)</th>
<th>N</th>
<th>Genotyping platform</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age, Gene/Environment, Susceptibility-Reykjavik</td>
<td>AGES</td>
<td>Tamara Harris Vilmundur Gudnadason</td>
<td>5.000</td>
<td>Illumina 317K</td>
</tr>
<tr>
<td>Atherosclerotic Risk in Communities</td>
<td>ARIC</td>
<td>Eric Boerwinkle</td>
<td>15.000</td>
<td>Affymetrix 6.0</td>
</tr>
<tr>
<td>Cardiovascular Health Study</td>
<td>CHS</td>
<td>Bruce Psaty</td>
<td>5.000</td>
<td>Illumina 317K</td>
</tr>
<tr>
<td>Framingham Heart Study</td>
<td>FHS</td>
<td>Chris O'Donnell</td>
<td>10.000</td>
<td>Affymetrix 550K</td>
</tr>
<tr>
<td>Rotterdam Study</td>
<td>RS</td>
<td>Albert Hofman André Uitterlinden</td>
<td>12.000</td>
<td>Illumina 550K, 610K</td>
</tr>
<tr>
<td>TOTAL GWAS</td>
<td></td>
<td></td>
<td>47.000</td>
<td></td>
</tr>
</tbody>
</table>

- Join GWA data for meta-analysis
- Many "Phenotype Working Groups (PWG)" (>30)
- Analyses/Publications arranged at PWG level; RSC oversees

GWAS for "height" in GIANT consortium

- 180,000 subjects
- 180 loci identified
- 10-15% variance explained
Ongoing Efforts in Genome-wide Genetics

- **Unanswered Questions.....:**
  - Causative SNP ? Causative gene ? Mechanism ?
  - Limited explained variance per trait/disease : ....“dark matter”

- **The Hunt for Genetic “Dark Matter” .....:**
  - Other types of genetic variation :
    - Rare variants (<5%, <1%, <0.1%, etc...) 
    - Copy Number Variations 
    - DNA Methylation patterns (450K ~95% CpG islands)
  - Interaction:
    - Gene-Gene and Gene-Environment

- **Technological Developments.....:**
  - High Throughput Sequencing :
    - 1000 genomes sequence project (imputation, rare variants)
    - CHARGE exome project (n > 4,000)
    - BBMRI Dutch GenomeNL (n=250 trio’s)

- **Published Genome-Wide Associations through 9/2010.....:**
  - >1,000 published GWAS hits at p<5x10^-8 for >160 traits

- **“YIELD” OF GWAS: QUANTITATIVE TRAITS**

- **September 2009 - Oktober 2010....:**
  - application-based library preparation
  - Roche Position
  - illumina®
  - Applied Biosystems
  - Genomic Genomics
Set-up for HT Sequencing
Genetic Lab, Internal Medicine

*Purchased 3x Illumina GAIIx (with support from the Dean, the dept Internal Medicine, the Genetic Lab)
*Installed April 2010
*Operational by summer 2010

3x Illumina GAIIx
• 2 x 100 bp reads
• 10 days
• 50 Gbp per machine
• 100 Gbp total

Computational set-up for HT Sequencing
(Internal Medicine, Genetic Lab, operational October 2010)

Illumina Compute
Isilon storage
• 180 TB raw
• 120 TB redundant

Erasmus MC network
"GRIMP":
Grid computing
3,000-6,000 cores

Dell compute
• 128 hard cores/256 virtual cores
• 6 GB/hard core

EDRA Storage
50-100 Tb

HT Sequencing set-up at the Genetic Lab
Erasmus MC Rotterdam (summer 2010)

*exchanged 3 GAII for 2 HiSeq2000
* Higher throughput and lower costs

Illumina HiSeq2000 (January 2011):
• 2 flowcells per machine
• 2 x 100 bp reads
• 8 days
• 150 Gbp per flowcell
• 300 Gbp per machine
• 600 Gbp with 2 HiSeq2000

Maximal number of samples per month (January 2011)

<table>
<thead>
<tr>
<th>Application</th>
<th>3 x GAIIx</th>
<th>2 x HiSeq2000</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exome @ 40X</td>
<td>270</td>
<td>1200</td>
</tr>
<tr>
<td>Transcriptome @ 12.5 Gb</td>
<td>35</td>
<td>192</td>
</tr>
<tr>
<td>Whole Genome @ 30X</td>
<td>4-5</td>
<td>24</td>
</tr>
</tbody>
</table>
### Consumable costs per sample (Jan 2011)

<table>
<thead>
<tr>
<th>Application</th>
<th>3 x GAIIx</th>
<th>2 x HiSeq2000*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exome @ 40X</td>
<td>€1100</td>
<td>€750</td>
</tr>
<tr>
<td>Transcriptome @ 12.5 Gb</td>
<td>€2500</td>
<td>€700</td>
</tr>
<tr>
<td>Whole genome @ 30X</td>
<td>€18,000</td>
<td>€5,000</td>
</tr>
</tbody>
</table>

*Prices subject to change by project volume and time

### Rotterdam Study HT Sequencing Projects... (November 2010)

**-Full Exome Sequencing** (funding by NGI-NCHA, NWO, BBMRI)
- *part of “CHARGE-S” effort:
  - >4,000 exomes across 5 cohorts: Framingham, CHS, ARIC, AGES, Rotterdam Study
- *2,000 samples (random set of RS-I; start december 2010)
- *Finish full exome in complete RS-I n=6500 (or switch to full genome?)
- *Case collections will be added depending on funding

**-Full RNA sequencing in RS blood RNA samples**
- *24 samples done (random set RS-III)
- *75 more, in progress

**-3 x 50 full genomes (trio design) for BBMRI-NL “GenomeNL”**:
- *HiSeq2000, 12x, outsourced to BGI

### DNA Sequence Data in the Rotterdam Study

An historical perspective: 1995 - 2010

<table>
<thead>
<tr>
<th>Approach</th>
<th>Genome Coverage (per person):</th>
<th>Samples Genotyped:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Candidate polymorphisms</td>
<td>n = 300 bp (SNPs/VNTRs/telomers)</td>
<td>12,000</td>
</tr>
<tr>
<td>SNP arrays 550K/610K</td>
<td>0.1% n = 2,500,000 bp (SNPs)</td>
<td>12,000</td>
</tr>
<tr>
<td>RNA arrays</td>
<td>n = &gt; 12,000 transcripts (blood cells)</td>
<td>1,000</td>
</tr>
<tr>
<td>Exome Sequencing</td>
<td>1% n = 31,000,000 bp</td>
<td>2,000 – 6,000</td>
</tr>
<tr>
<td>RNA sequencing</td>
<td>n = ?</td>
<td>24</td>
</tr>
<tr>
<td>Full Genome sequencing</td>
<td>100% n = 3,300,000,000 bp</td>
<td>(~500)</td>
</tr>
</tbody>
</table>

### Commercial SNP analysis

Recreational Genomics is “not a medical necessity”...

<table>
<thead>
<tr>
<th>[Dec 2008, Caucasian male (AU)]</th>
<th>deCODE</th>
<th>23andMe</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rheumatoid Arthritis; RR =</td>
<td>1.86 (7 SNPs)</td>
<td>0.60 (6 SNPs)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>GENE:</th>
<th>SNP rs-number (genotype):</th>
</tr>
</thead>
<tbody>
<tr>
<td>HLA DRB1</td>
<td>- rs6457617 (CT)</td>
</tr>
<tr>
<td>PADH4</td>
<td>- rs1120386 (AG)</td>
</tr>
<tr>
<td>PTPN62</td>
<td>rs2478651 (GG) rs2478651 (G/G)</td>
</tr>
<tr>
<td>MMEL1</td>
<td>- rs2890745 (CT)</td>
</tr>
<tr>
<td>Eg3</td>
<td>rs2327032 (AA) rs2327032 (A/A)</td>
</tr>
<tr>
<td>TRAF1/CT</td>
<td>rs13192841 (GG) rs13192841 (G/G)</td>
</tr>
<tr>
<td>IL23L21</td>
<td>rs6822844 (GG) rs6822844 (G/G)</td>
</tr>
<tr>
<td>ST1A4</td>
<td>rs7574865 (GT) rs7574865 (G/T)</td>
</tr>
</tbody>
</table>
Courses in Complex Genetics........

3rd International 3-day Symposium
9-11 March 2011
Erasmus MC, Rotterdam, The Netherlands

2nd “Genetics for Dummies” Course
2-day Introductory Course
MoIMed, MGC & NIHES
September 2011
Erasmus MC, Rotterdam, The Netherlands

8th “SNP COURSE”
5-day Advanced Course
MoIMed, MGC & NIHES
November 2011
Erasmus MC, Rotterdam, The Netherlands

Contact: WWW.MOLMED.NL