Estimating risk profiles for common diseases from environmental and genetic factors

Cathryn Lewis
King’s College London

Contents

• Introduction to genetic prediction
• Estimating disease risks
• Implications
Introduction to genetics: 1
DNA structure

DNA has a double-helix structure like a twisted ladder. The steps of the ladder are bases...

ADENINE (A) IS A BASE.
THYMINE (T) IS A BASE.
CYTOSINE (C) IS A BASE.
GUANINE (G) IS A BASE.

www.onlineeducation.net/dna

Introduction to genetics: 2
DNA differences

What makes us different?
These differences control our hair colour, our height, and the diseases we will get

99.9% OF OUR DNA SEQUENCE IS THE SAME AS OTHER HUMANS'.

This 0.1% DNA DIFFERENCE between us may have to do with the number of nucleotides in a person’s DNA.
When DNA is copied into a new life, the nucleotides are either gained or lost in the process. This gain or loss results in our differences.
Inherited genetic mutations

Single gene disorders
- Huntington’s disease
- Cystic fibrosis
- Breast cancer genes: BRCA1, BRCA2

Complex disease: contributions from genetic and environmental factors

Examples: asthma, breast cancer, heart disease, autism, arthritis, migraine, obesity, diabetes, stroke

Most diseases that have a major economic, social and health burden
Genetic variation:
Single nucleotide polymorphism (SNP)

....TGGACATGCA....
....TGGACCTGCA....

Alleles A and C are present in the population

Genotype: carried by an individual, on paternal and maternal inherited chromosomes

....TGGACATGCA....  ....TGGACATGCA....  ....TGGACCTGCA....
....TGGACATGCA....  ....TGGACCTGCA....  ....TGGACCTGCA....

Genotype: AA  AC  CC

Identifying SNPs that increase risk of disease

Genotype SNP with A, C alleles:  AA  AC  CC

Cases – affected with disease  Controls – not affected with disease

More AC and CC genotypes in cases than in controls
Indicates that carrying C allele increases risk of disease
Genetic association studies

- Test 500k SNPs across genome for differences between cases and controls
- Identify panels of SNPs that control risk of disease
- Each SNP: odds ratio of disease, frequency in population
- For any individual, can calculate genetic risk profile across these SNPs

Breast cancer genetics

<table>
<thead>
<tr>
<th>Name of SNP</th>
<th>Gene location</th>
<th>Risk allele</th>
<th>Odds Ratio, by number of risk alleles</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs2981579</td>
<td>FGFR2</td>
<td>A</td>
<td>0  1  1.35  1.82</td>
</tr>
</tbody>
</table>

To combine relative risk across SNPs: multiply odds ratio for genotype

Product of odds ratios = 1 x 1.28 x 1.42 x 1.31 x 1 = 2.38
Rescale so OR is compared to an ‘average’ member of the population
Distribution of genetic risk in the population

Decreased risk
Carry few risk alleles

Baseline risk

Increased risk
Carry many risk alleles

How useful is this information?
Research programme:
Disease risk estimation for combining genetic and environmental risk factors

• Developed new statistical methodology
  – Combining genetic and environmental risk factors
  – Incorporating confidence intervals

• Issued software program REGENT

• Evaluated utility of risk prediction for common diseases
Risk modelling: genetic factors

- Risk SNP characterised by
  - Minor allele frequency (MAF), \( p \)
  - Odds ratio for each minor allele \((1, g, g^2)\)
- Disease prevalence \( r \)
- Assume risks are multiplicative across SNPs

\[ N \text{ SNPs, with genotype } k_i = 0, 1, 2, i=1, \ldots, N \]

\[ P(D | k_1, k_2, \ldots, k_N) = r \prod_{k=1}^{N} g_i^{k_i} / \left( 1 + (g_i - 1)p_i \right)^2 \]

Risk modelling: environmental factors

- Environmental risk factors \((M)\), each with
  - OR \( h_j \)
  - Confidence interval
  - Exposure prevalence, \( e[j] = 0, 1 \)
- Risk component relative to individual with no exposure is:

\[ \prod_{j=1}^{M} h_{i[j]}^{e[j]} \]

Model assume environmental risks are independent
Risk modelling: confidence intervals

Disease risk estimated using multiplicative model between
Genetic risk factors
Environmental risk factors

Calculate empiric confidence intervals for an individual genotype

Type 2 diabetes risk SNPs

<table>
<thead>
<tr>
<th>SNP</th>
<th>Allele frequency</th>
<th>OR</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs5215</td>
<td>0.35</td>
<td>1.14</td>
</tr>
<tr>
<td>rs7901695</td>
<td>0.31</td>
<td>1.37</td>
</tr>
<tr>
<td>rs4430796</td>
<td>0.47</td>
<td>1.10</td>
</tr>
</tbody>
</table>

Frayling et al., 2007
### Three SNPs: Type 2 Diabetes

The table below summarizes the risk categories for different combinations of three SNPs in Type 2 Diabetes.

<table>
<thead>
<tr>
<th>Combination Number</th>
<th>SNP Number</th>
<th>Population Frequency</th>
<th>Rel. Risk (Rebased)</th>
<th>Rel. Risk Quartiles</th>
<th>Risk Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0 0 0</td>
<td>0.0565</td>
<td>0.6636</td>
<td>0.5649</td>
<td>Low</td>
</tr>
<tr>
<td>2</td>
<td>0 0 1</td>
<td>0.1002</td>
<td>0.7299</td>
<td>0.6395</td>
<td>Low</td>
</tr>
<tr>
<td>3</td>
<td>1 0 0</td>
<td>0.0609</td>
<td>0.7565</td>
<td>0.6482</td>
<td>Low</td>
</tr>
<tr>
<td>4</td>
<td>0 0 2</td>
<td>0.0444</td>
<td>0.8029</td>
<td>0.6607</td>
<td>Average</td>
</tr>
<tr>
<td>5</td>
<td>1 0 1</td>
<td>0.1079</td>
<td>0.8321</td>
<td>0.7385</td>
<td>Average</td>
</tr>
<tr>
<td>6</td>
<td>2 0 0</td>
<td>0.0164</td>
<td>0.8624</td>
<td>0.7010</td>
<td>Average</td>
</tr>
<tr>
<td>7</td>
<td>0 1 0</td>
<td>0.0506</td>
<td>0.9091</td>
<td>0.7609</td>
<td>Average</td>
</tr>
<tr>
<td>8</td>
<td>1 0 2</td>
<td>0.0479</td>
<td>0.9153</td>
<td>0.7579</td>
<td>Average</td>
</tr>
<tr>
<td>9</td>
<td>2 0 1</td>
<td>0.0291</td>
<td>0.9486</td>
<td>0.7866</td>
<td>Average</td>
</tr>
<tr>
<td>10</td>
<td>0 1 1</td>
<td>0.0900</td>
<td>1.0900</td>
<td>0.8876</td>
<td>Average</td>
</tr>
<tr>
<td>11</td>
<td>1 1 0</td>
<td>0.0547</td>
<td>1.0364</td>
<td>0.9026</td>
<td>Average</td>
</tr>
<tr>
<td>12</td>
<td>2 0 2</td>
<td>0.0129</td>
<td>1.0435</td>
<td>0.8297</td>
<td>Average</td>
</tr>
<tr>
<td>13</td>
<td>0 1 2</td>
<td>0.0399</td>
<td>1.1000</td>
<td>0.9112</td>
<td>Average</td>
</tr>
<tr>
<td>14</td>
<td>1 1 1</td>
<td>0.0970</td>
<td>1.1400</td>
<td>0.9288</td>
<td>Average</td>
</tr>
<tr>
<td>15</td>
<td>2 1 0</td>
<td>0.0147</td>
<td>1.1815</td>
<td>0.9738</td>
<td>Average</td>
</tr>
<tr>
<td>16</td>
<td>0 2 0</td>
<td>0.0114</td>
<td>1.2455</td>
<td>1.0094</td>
<td>Average</td>
</tr>
<tr>
<td>17</td>
<td>1 1 2</td>
<td>0.0430</td>
<td>1.2540</td>
<td>1.0609</td>
<td>Average</td>
</tr>
<tr>
<td>18</td>
<td>2 1 1</td>
<td>0.0281</td>
<td>1.2996</td>
<td>1.0950</td>
<td>Average</td>
</tr>
<tr>
<td>19</td>
<td>0 2 1</td>
<td>0.0202</td>
<td>1.3700</td>
<td>1.1394</td>
<td>Moderate</td>
</tr>
<tr>
<td>20</td>
<td>1 2 0</td>
<td>0.0123</td>
<td>1.4198</td>
<td>1.1634</td>
<td>Moderate</td>
</tr>
<tr>
<td>21</td>
<td>2 1 2</td>
<td>0.0116</td>
<td>1.4296</td>
<td>1.1452</td>
<td>Moderate</td>
</tr>
<tr>
<td>22</td>
<td>0 2 2</td>
<td>0.0090</td>
<td>1.5070</td>
<td>1.1937</td>
<td>Moderate</td>
</tr>
<tr>
<td>23</td>
<td>1 2 1</td>
<td>0.0218</td>
<td>1.5616</td>
<td>1.3149</td>
<td>Moderate</td>
</tr>
<tr>
<td>24</td>
<td>2 2 0</td>
<td>0.0033</td>
<td>1.6186</td>
<td>1.2709</td>
<td>Moderate</td>
</tr>
<tr>
<td>25</td>
<td>1 2 2</td>
<td>0.0097</td>
<td>1.7180</td>
<td>1.3723</td>
<td>Moderate</td>
</tr>
<tr>
<td>26</td>
<td>2 2 1</td>
<td>0.0004</td>
<td>1.7605</td>
<td>1.4225</td>
<td>Moderate</td>
</tr>
<tr>
<td>27</td>
<td>2 2 2</td>
<td>0.0026</td>
<td>1.9585</td>
<td>1.4956</td>
<td>Moderate</td>
</tr>
</tbody>
</table>

Three SNPs: Type 2 Diabetes

**Average risk: 68%**

**Elevated risk: 10%**

**Reduced Risk: 22%**

No high risk genotypes

Different Risk Categories
Crohn's disease risk estimation

REGENT software

- R package
  - [http://cran.r-project.org/web/packages/REGENT/](http://cran.r-project.org/web/packages/REGENT/)
- Population distribution of disease risk and risk categories
- Individual-level risk assessment
- Genetic risk factors (SNP genotypes) and environmental risk factors (multilevel)

Crouch, Goddard, Lewis, EJHG, 2012
Genetic risk profile

- Case studies of three adult-onset disorders:
  - Coronary artery disease
  - Colorectal cancer
  - Type 2 diabetes

- Identified SNPs most strongly associated with disease

- Modelled genetic profiles in the population through simulation

- Assessed ability of model to identify individuals at high risk of disease

Receiver operating characteristic curve

<table>
<thead>
<tr>
<th>Hypothised</th>
<th>True group</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

TP: True Positive
FP: False Positive
FN: False Negative
TN: True Negative

AUC: area under curve
 Measure of discrimination of test

Sensitivity = 1 - specificity

Receiver operating characteristic curve

AUC = Area under curve = P(correctly assign disease status to a case control pair)
## Genetic risk assessment

<table>
<thead>
<tr>
<th>Disease</th>
<th>No. SNPs modelled</th>
<th>Area under curve</th>
<th>Proportion of population at increased risk</th>
<th>Lifetime risks</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>OR &gt; 2</td>
<td>OR &gt; 3</td>
</tr>
<tr>
<td>Coronary artery disease</td>
<td>25</td>
<td>0.60</td>
<td>1.5%</td>
<td>0.0%</td>
</tr>
<tr>
<td>Colorectal cancer</td>
<td>10</td>
<td>0.59</td>
<td>0.7%</td>
<td>0.0%</td>
</tr>
<tr>
<td>Type 2 diabetes</td>
<td>19</td>
<td>0.60</td>
<td>1.7%</td>
<td>0.0%</td>
</tr>
</tbody>
</table>

### Odds ratios: genetic v. conventional risk factors

<table>
<thead>
<tr>
<th>Disease</th>
<th>Top 5% of genetic risk</th>
<th>Family history (affected sibling)</th>
<th>Epidemiological &amp; risk factors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coronary artery disease</td>
<td>1.7</td>
<td>3.2</td>
<td>Total cholesterol Smoking 3.1 1.9</td>
</tr>
<tr>
<td>Colorectal cancer</td>
<td>1.6</td>
<td>5.1</td>
<td>Smoking Obesity 1.3 1.5</td>
</tr>
<tr>
<td>Type 2 diabetes</td>
<td>1.7</td>
<td>3.5</td>
<td>Obesity 2.5</td>
</tr>
</tbody>
</table>
Summary

• *Scientific* strides in identifying the inherited genetic variants that affect disease risk

• Very limited prediction available from current findings
  – Incomplete knowledge of polygenic component of disease
  – Causal genetic variants are unknown

• Better prediction comes from
  – Family history
  – Environmental risk factors (smoking, body mass index)
  – Pre-clinical factors (blood pressure, cholesterol levels)

Acknowledgements

**King’s College London**

• Graham Goddard
• Daniel Crouch
• Jane Yarnall

Funding
“Prediction is very difficult, especially about the future”

Niels Bohr