MutaGeneSys

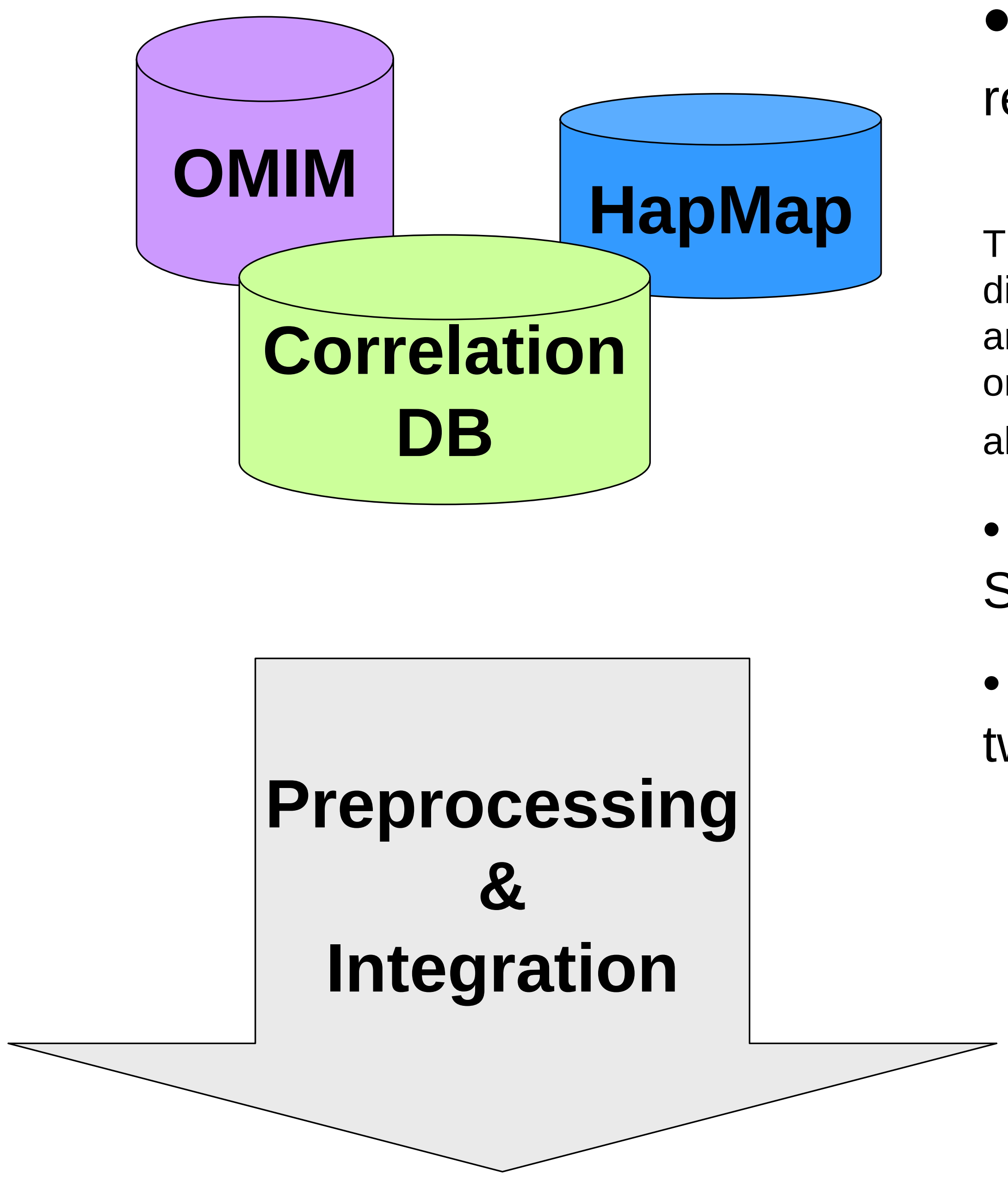
Making Diagnostic Predictions Based on Genome-Wide Genotype Data in Association Studies

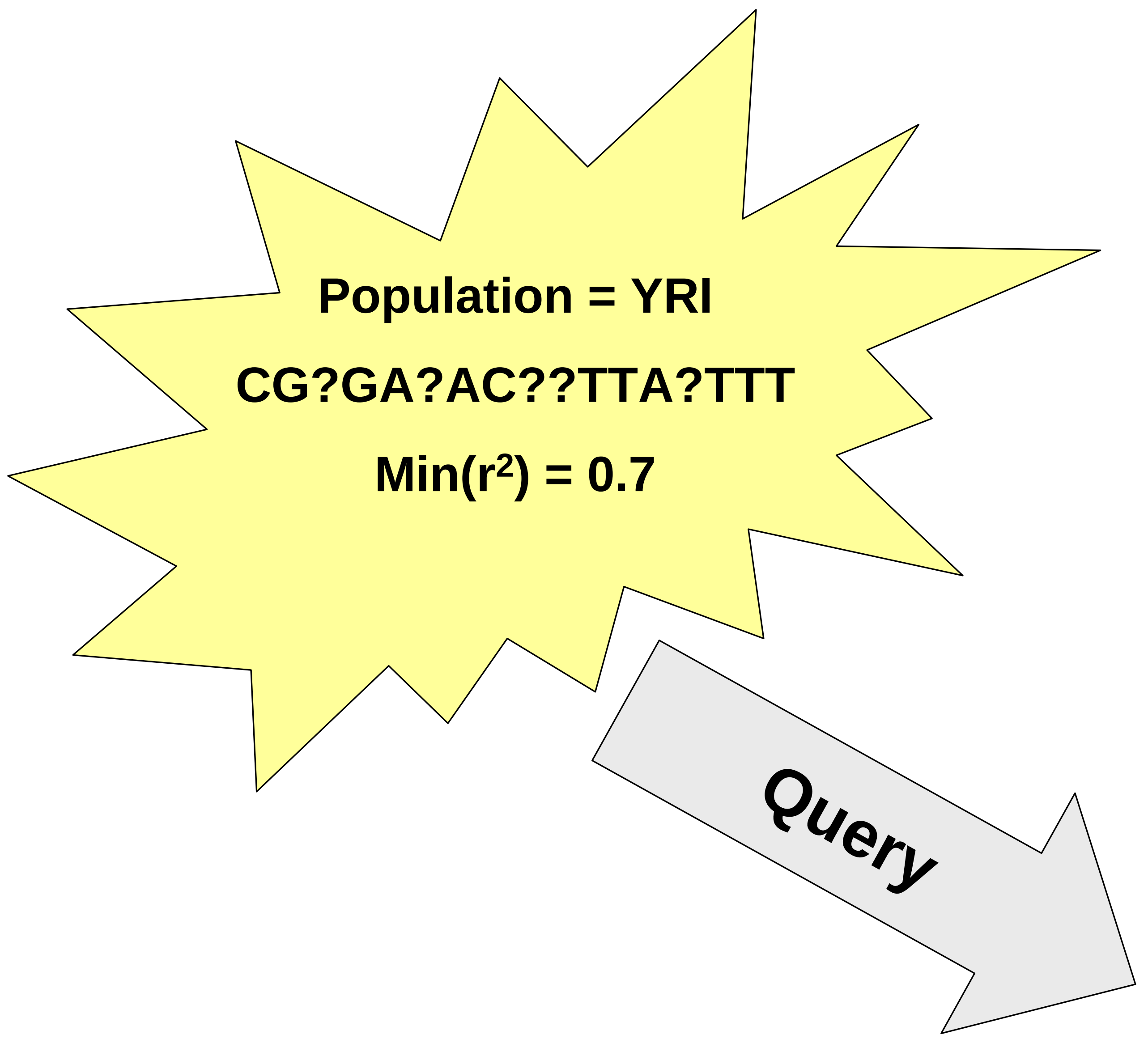
Julia Stoyanovich (Ross Lab) and Itsik Pe’er

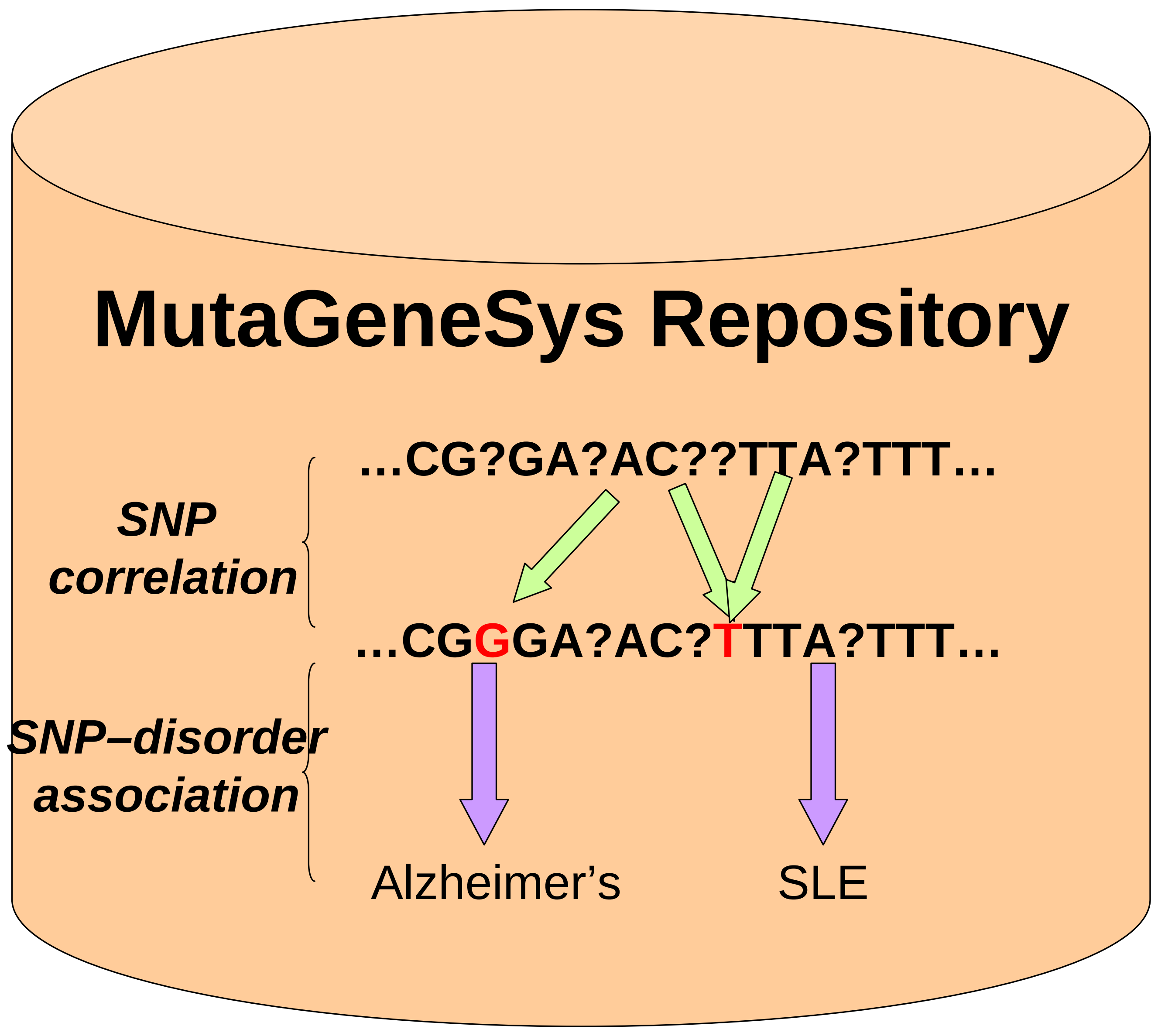
Towards Personalized Medicine

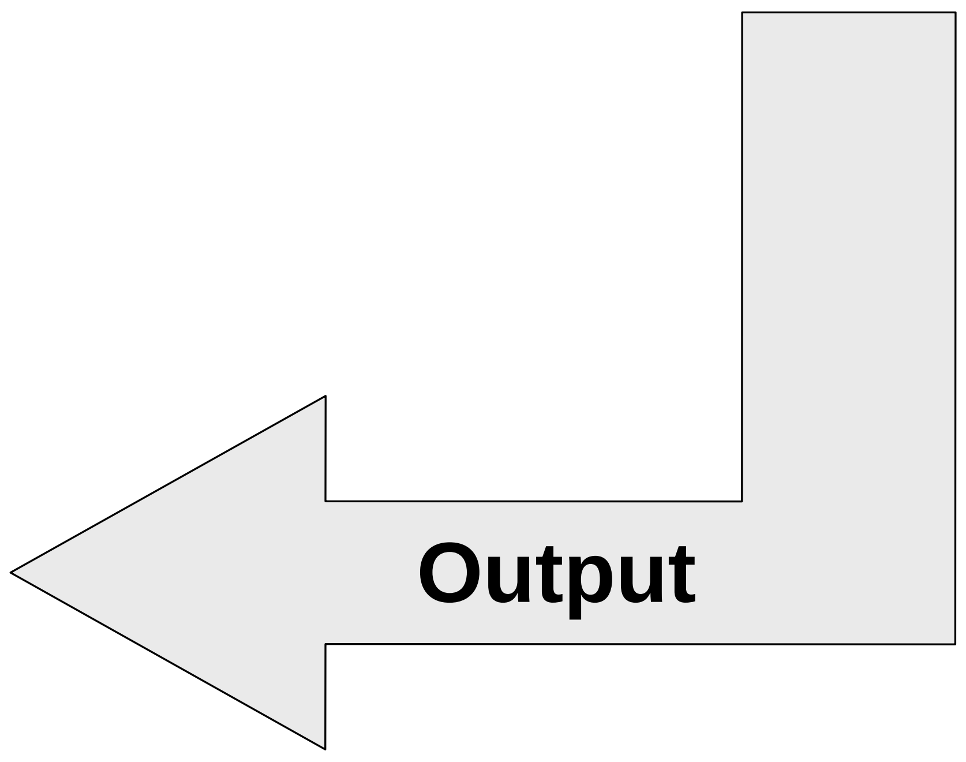
- Use individual’s genetic information for disease susceptibility prognosis.
- Genotyping still expensive (both time and $), so often only partial genetic data is available.
- Indirect association to the rescue:

SNPs have many proxies!









```
<XML>
<ROWSET>
<ROW>
<REQID>20</REQID>
<SNPID>rs6267</SNPID>
<OMIMID>116790</OMIMID>
<TITLE>;+116790 CATECHOL-O-METHYLTRANSFERASE; COMT;</TITLE>
<TS>06-DEC-06</TS>
</ROW>
<ROW>
<REQID>20</REQID>
<SNPID>rs10456057</SNPID>
<OMIMID>177900</OMIMID>
<TITLE>#177900 PSORIASIS SUSCEPTIBILITY 1; PSORS1</TITLE>
<TS>06-DEC-06</TS>
</ROW>
<ROW>
<REQID>20</REQID>
<SNPID>rs795009</SNPID>
<OMIMID>181500</OMIMID>
<TITLE>#181500 SCHIZOPHRENIA; SCZD</TITLE>
<TS>06-DEC-06</TS>
</ROW>
……
</XML>
```

Building Blocks of MutaGeneSys

- Online Mendelian Inheritance in Man (OMIM): highly reputable data, but in text form (scientific articles)

> “… Mace et al. (2005) found a significant association between a C-T SNP (rs908832) in exon 14 of the ABCA2 gene (600047) and Alzheimer disease in a large case-control study involving 440 AD patients. Additional analysis showed the strongest association between the SNP and early-onset AD (odds ratio of 3.82 for disease development in carriers of the T allele compared to controls)... “

- The International HapMap Project: complete list of SNPs, by population, with alleles and frequencies
- Genome-Wide marker correlation data: single and two-marker correlations

rs12076827 (A) + rs1572970 (A) => rs1205 (T)

Implementation Details

- Repository implemented as a relational schema in Oracle 10g
- Tables, materialized views, an API to interact with the data (e.g. from a Java program)
- Load and refresh utilities for HapMap, OMIM, IntraGenDB and standard linkage formats

Results

Large repository

- 10M SNP records
- 50M single correlations, 20M two-marker correlations
- significantly enriched SNP-disorder associations
  - 187 in OMIM
  - 1312 in MutaGeneSys

Real-time Performance

- under 5 seconds for a full genotype scan (317,000 SNPs) on a slow machine

Flexible Retrieval

- by population: CEU, YRI, JPT+CHB
- technology & resolution: Affymetrix, Illumina
- by Pearson’s correlation coefficient (r^2)

References

[1] MutaGeneSys: Making Diagnostic Predictions Based on Genome-Wide Genotype Data in Association Studies", J. Stoyanovich, I. Pe'er, Columbia University tech report, February 2007.

[2] Evaluating and improving power of whole genome products. www.cs.columbia.edu/~itsik/StandardGenotyping.htm.

[3] de Bakker et al. Transferability of tag SNPs in genetic association studies in multiple populations. *Nature Genetics*, 38:1298.1303, 2006.

[4] Pe'er, de Bakker, Maller, Yelensky, Altshuler, and Daly. Evaluating and improving power in whole genome association studies using fixed marker sets. *Nature Genetics*, 38(6):663.7, 2006.