



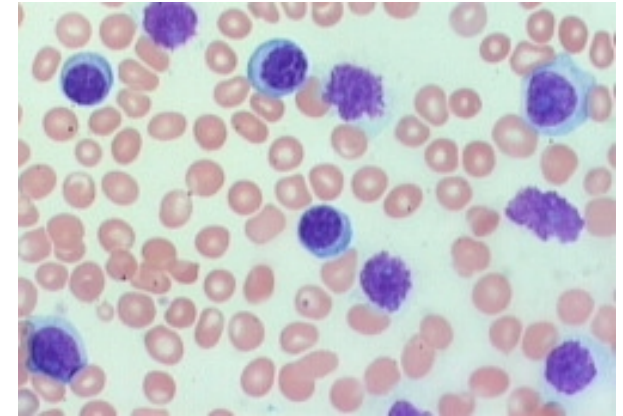
# Predicting Cancer Susceptibility from Single-Nucleotide Polymorphism Data: A Case Study in Multiple Myeloma

**Michael J. Waddell**

Laboratory of Dr. C. David Page  
Departments of Computer Science and  
Biostatistics and Medical Informatics  
University of Wisconsin

# Background

- **Multiple Myeloma** (MM) is a uniformly fatal malignancy of the plasma cells.
- MM occurs with relatively **high frequency** in older adults (0.035% of the US population aged 70+)
- MM occurs with much **lower frequency** in younger adults (0.002% of the US population aged 30–54)



Source: <http://seer.cancer.gov>



## Background

- We hypothesize that those diagnosed with MM at a young age have a **genetic predisposition** to the disease
- To study this, we would like to have the **complete genome sequences** for a large number of patients.
- Currently, it is not possible to **quickly** obtain the sequence of a patient's complete genome.
- However, it **is** possible to quickly obtain a patient's "SNP pattern."



## SNPs

- Part of the genetic variation among individuals is the **cumulative** effect of variations at a number of single-base locations within the genome.
- These locations are known as **SNPs** (Single Nucleotide Polymorphisms)
- A “**SNP pattern**” consists of the particular DNA bases present at a large number of SNP positions.



## Benefits of Using SNPs

- A person's SNP pattern is highly unlikely to **change** over time or as a result of disease.
- SNP data can be collected from **any tissue** in the body (not just from diseased tissue)
- This allows a larger number of samples to be obtained (especially controls) since **less invasive** procedures are used

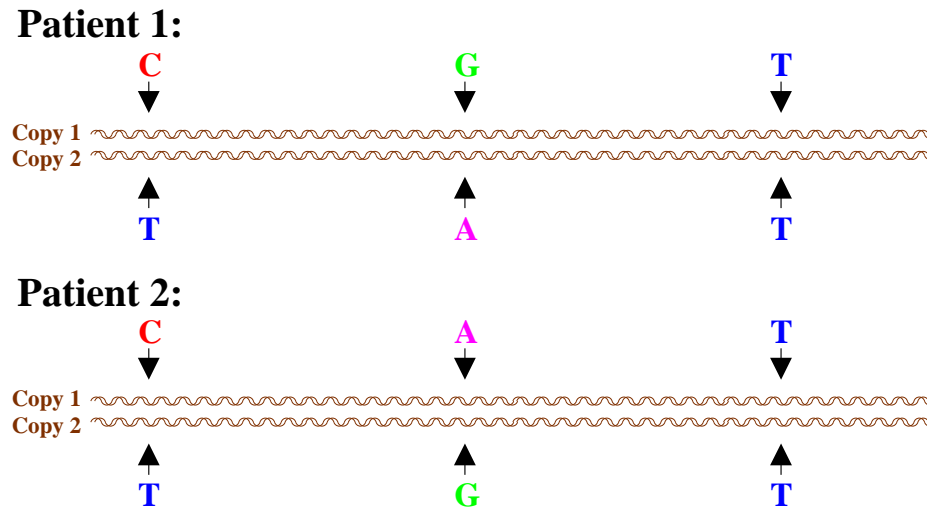


## 3 Challenges of Using SNPs

1. There are now over **one million** SNPs known
  - SNP data contain measurements for only a **small fraction** of known SNPs (typically a few thousand)
2. SNP data commonly contain **missing values**
  - This can **adversely affect** many algorithms used for classification tasks

# 3 Challenges of Using SNPs

## 3. SNP data are “unphased.”



	SNP 1		SNP 2		SNP 3		...	class
Patient 1	C	T	A	G	T	T	...	Diseased
Patient 2	C	T	A	G	T	T	...	Healthy
...	...	...	...	...	...	...	...	...



# Computational Approach

1. Divide the data (patient samples) into two (or more) **classes**
2. Use modeling algorithms to generate a classifier, or **model**, of the disease
3. Validate this model using **cross-validation** (leave-one-out)

*(This is also the approach we used in a related study using gene-expression microarray data with Multiple Myeloma)*



## Data Set

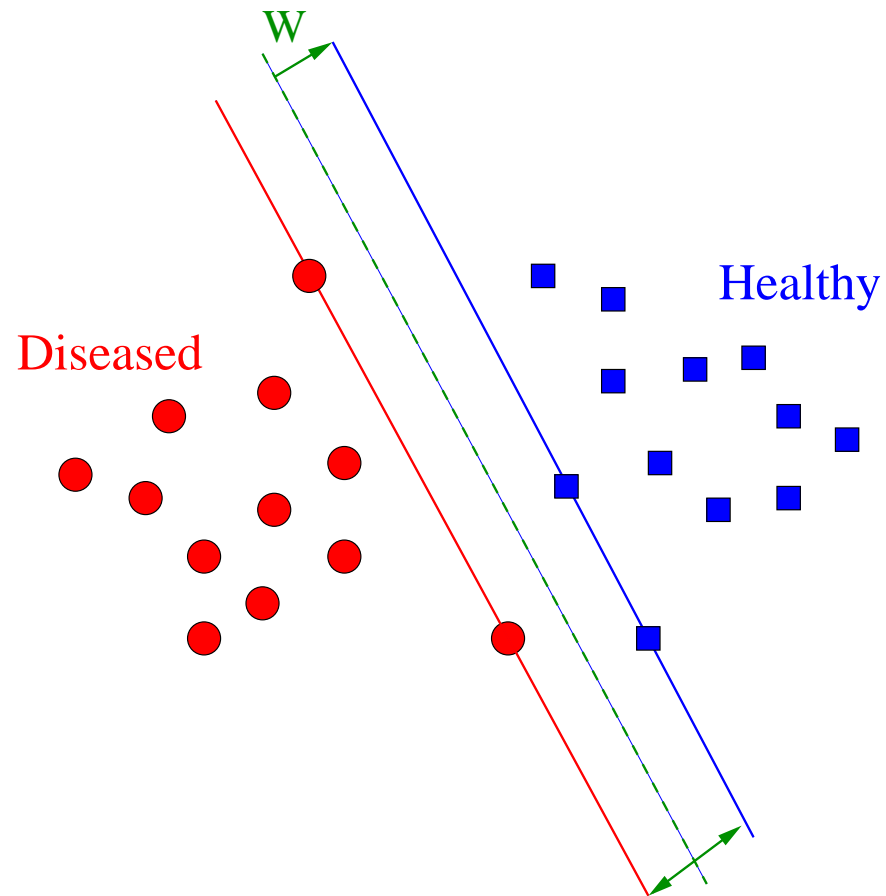
- “Unphased” SNP data for **80** patients (based on 3000 SNPs)
  - 40 “**old**” patients: diagnosed with MM after age 70
  - 40 “**young**” patients: diagnosed with MM before age 40
- The SNPs were selected to give good **overall coverage** of the human genome — **not** based on prior knowledge of MM



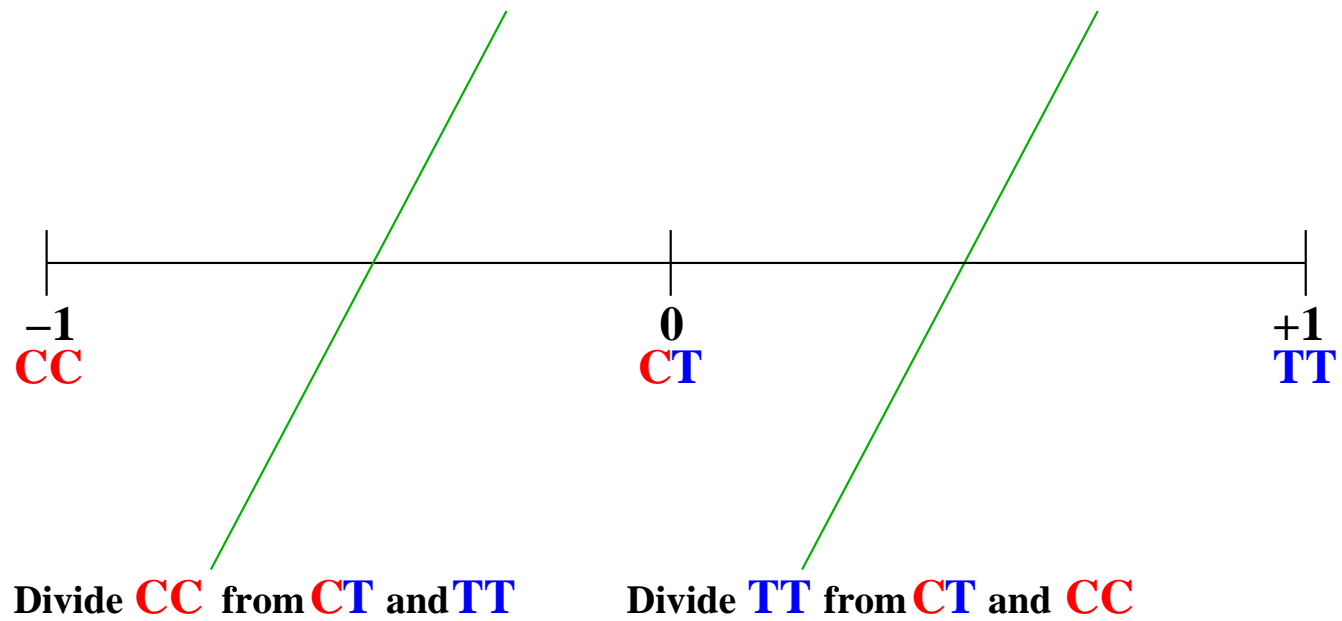
## Feature Selection

- **Feature selection** is commonly used to improve performance of modeling algorithms
- It is common practice, **though incorrect**, to perform feature selection once over the whole data set
- This practice leads to information **“leaking”** from the validation set into the training set
- To avoid this common pitfall, the top 10% of SNPs were chosen by information gain **on each fold** prior to model construction

# Support Vector Machines



# SVMs Require Continuous Features





# Results

		Predicted	
		Old	Young
Actual	Old	31	9
	Young	14	26

Our approach yields an accuracy estimate of **71%** by leave-one-out cross validation.  
This is **significantly** better than random guessing ( $p < 0.01$ )



## Conclusions

- Our accuracy of **71%** is not as high as we have obtained using microarray data
- However, this prediction is based **only** on SNP data (which are not affected by disease progression)
- Also, our SNP coverage was **relatively sparse** (only 3000 SNPs were used)
- Thus, we conclude that SNP data **do** provide predictive ability for cancer susceptibility



## Future Work

- **Expand** the study to include more patients in order to further validate these results
- Use a **denser coverage** of the genome than the 3000 SNPs used
- Extract **insight** from the SVM model as to which genes are most important
- **Compare** these genes with those that we found in gene-expression microarray experiments for MM



# Acknowledgments

---

- **David Page**  
University of Wisconsin
- **Fenghuang Zhan, Bart Barlogie and John Shaughnessy, Jr.**  
University of Arkansas for Medical Sciences
- **Computation and Informatics in Biology and Medicine (CIBM)  
Training Program**
- **National Library of Medicine**  
Grant No. 1T15LM007359-01
- **SVM<sup>light</sup>**  
<http://svmlight.joachims.org>