Evidence of Influence of Genomic DNA Sequence on Human X Chromosome Inactivation.

September 2006
Wang Z, Willard H-F, Mukherjee S, Furey T-S.
PLOS Comp. Biology 2(9) e113, 0979-0988

Seminar in Bioinformatics

Kõressaar Triinu
Friday, 13th october

Tartu
2006

Definition of X inactivation (also lyonization)

The phenomenon in a female mammals by which one X chromosome (either the maternally or paternally derived X) is randomly (by chance) inactivated in an early embryonic cell (dosage compensation), with fixed inactivation of that same X in all cells descended from that cell.

The process of X-inactivation

- Randomly chosen X-chromosome
- Transcription of XIST-gene
- Initial inactivation
- Formation of heterochromatin
- Maintenance of X-inactivation
  - modification of histone
  - methylation of DNA

Some details about the process of X-inactivation

- Transcriptional inactivation is not complete (~15% of human genes in inactive X-chromosome are actively transcribed)

- The distribution of genes that escape inactivation is nonrandom

- Two region on X-chromosome: X-conserved region (XCR) and X-added region (XAR)

- Genes that are subject to or escape from inactivation are clustered within the XAR

Genomic landmarks

- “way stations” that aid the spreading of X-inactivation
- Three research studies concluded:
    - long interspersed nuclear element 1 (LINE 1, or L1) retrotransposons are enriched in the vicinity of genes that are subject to inactivation (1998, 2000)
    - mammalian-wide interspersed repeat (MIR) elements and CpG islands were significantly depleted in regions that escape inactivation (2003).
    - 'GATA' repeats are enriched in a region of X-chromosome where all genes escape from X-inactivation (2006)

<u>The shortcomings of the current work:</u>

- The number of potential sequence features analyzed were limited
- The number of genes analyzed were limited

Whether a gene escapes or is subject to inactivation is thought to be determined epigenetically

In this work, the authors show that the DNA sequence surrounding genes that escape inactivation is significantly different from the sequence surrounding genes that are subject to inactivation.

Multiple sequence features may influence X inactivation in an interdependent fashion. Determining these factors and their possibly combinatorial nature thus presents a complex problem.

Data used in this study

- the complete set of human genes of known X inactivation status

- 310 repeat families and subfamilies, CpG islands, all 64 three-base and 1,024 five-base sequences were extracted from X-chromosome sequence in 2-, 5-, 10-, 20-, 50-, and 100-kb windows from surrounding their transcription start sites

- 73 escaping and 375 subject genes (XAR: 50e, 60s; XCR: 23e, 315s): totally 16,788 primary sequence features were included into analysis

- final dataset was a matrix with features as columns and genes as rows.

Wilcoxon rank-sum test

- To determine which features, if any, have different distributions in the genome sequence surrounding genes subject to inactivation as compared with those that escape inactivation (1st analyses)
- Test statistic W (weight) for every feature was calculated as:

$$W_j = (m_{j,i \in e} - m_{j,i \in s}) \times r_j$$

where $m_{j,i \in e}$ is the median rank for the escaping genes, $m_{j,i \in s}$ is the median rank for subject genes, r_j is the Pearson correlation of the *j*th feature to X-inactivation status

- p-value was calculated by randomly permuting gene labels 1,000 times and calculating a weight for each permutation

Wilcoxon rank-sum test

- to provide a measure of the false discovery rate q-values were calculated

(Storey and Tibshirani (2003) Statistical significance for genomewide studies. PNAS 100-16)

- *the false positive rate* – the rate that a truly null features are called significant
- *the FDR* – the rate that significant features are truy null
- a false positive rate of 5% means that on average 5% of the truly null features in the study will be called significant. A FDR of 5% means that among all features called significant, 5% of these are truly null on average.
- **971 significant features at q<0.02 and 2,345 features at q<0.05**

- May these results reflect the unique evolutionary history of the X chromosome rather than a specific relationship to X inactivation?

- Genes with different X inactivation statuses within XAR alone were compared (2nd analyses)
- **1,506 significant features at q < 0.02 and 3,336 at q <0.05**

- five unique strata that essentially separate the ancestral X chromosome sequence (XCR, strata 1–2) from the sequence added later were analyzed (XAR, strata 3-5) (3rd analyses)
- **100 significant features found at q< 0.15 and 449 at q< 0.20**

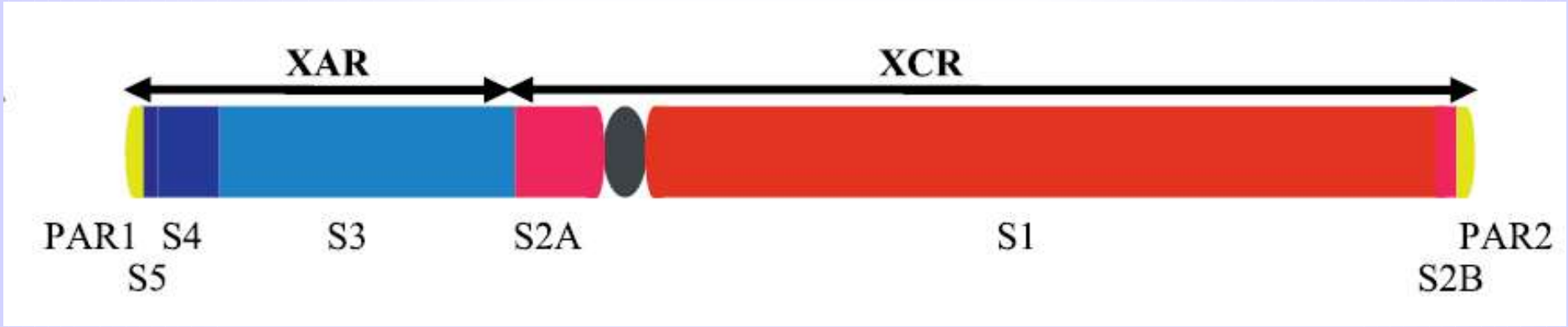


Fig1A. A schematic drawing of the X chromosome delineating each evolutionary stratum

From these data, they conclude that significant features commonly identified in these 3 analyses most likely *represent global differences between the genomic environment of escaping and subject genes and not simply regional differences*

Significant features

- L1s and MIRs in regions surrounding the TSS of genes subject to X inactivation

- Alu elements in regions surrounding the TSS of genes that escape inactivation

- The distributions of features show significant differences in multiple window sizes (especially 50-kb and 100-kb windows, located both upstream and downstream of the transcription start site) - suggesting that the larger genomic environment may be most relevant for determining X inactivation status

- 10 repeat sequence features are significant chromosomewide and in the XAR ($q < 0.05$) and also within stratum 3 ($q < 0.2$)

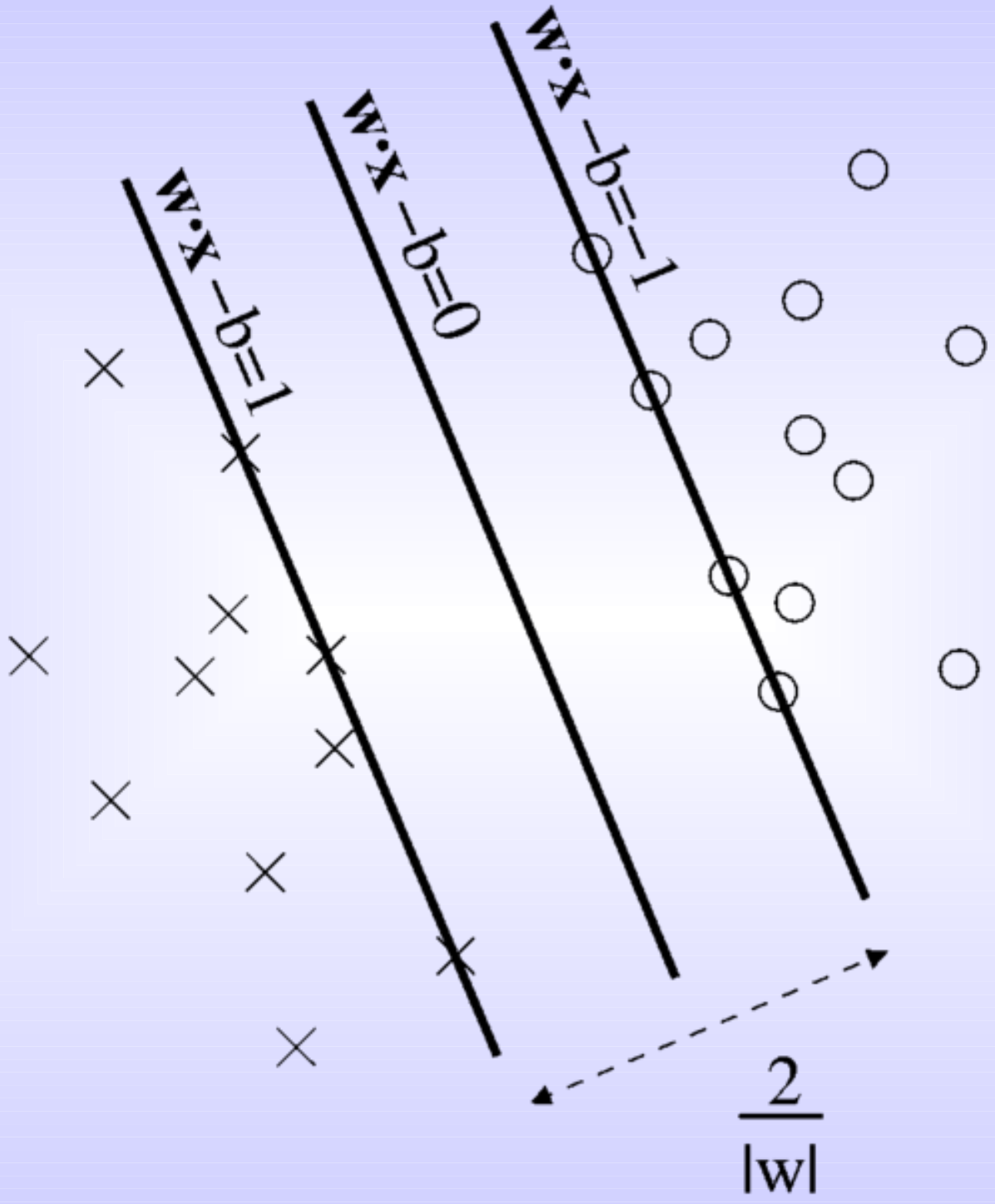
Significant features

- The concentration of several 3-base and 5-base sequences is different between the two classes of genes (around escaping genes 3- and 5-mers are GC rich, around subject genes 3- and 5-mers are AT rich)
- Escaped genes:
  - top 12 3-mers with respect to rank-sum values ($q<0.012$) are CGT/ACG
  - top eight and 43 of the top 51 5-mers contain CGT/ACG
  - neither GC content nor CpG island content is significantly different between the two sets of genes
  - particular types of GC-rich sequences (CGT/ACG motifs) are important for escaping X inactivation
- Subject genes:
  - The case of escaped genes can be concluded to subject genes also

How to accurately discriminate between the two classes of genes?

- Linear support vector machine (SVM) classifiers constructed using primary DNA sequence features are used to correctly predict the X inactivation status

- SVM (a learning method) is a technique for data classification:
  - Whether we could separate multidimensional data points ('neatly') by hyperplane?
  - Simultaneously minimize the empirical classification error and maximize the geometric margin


w•x –b=1
w•x –b=0
w•x –b=1
2
|w|

Training and Testing Data

- 110 genes on XAR (50e+60s) with known X inactivation status
    - 5,596 repeat, 3-mer and 5-mer sequence features derived from 50- and 100kb window

- Gene groups were created by calculating the overlap of every genes' 100-kb upstream (downstream) regions with neighboring genes' upstream (downstream) regions. A total of 62 groups were created for the 110 XAR genes

- y-homology features: pseudoautosomal genes (1, 1, 1), genes with functioning Y-homologs (0, 1, 1), genes with Y-linked pseudogenes (0, 0, 1), and genes with no apparent Y-homolog or pseudogene (0, 0, 0)

SVM classification and recursive feature selection

- Each instance (data point) in the training set contains one "target value" and several "attributes" (features):
  - Each gene/EST was represented by a feature vector
  - Each gene/EST was labeled as either escaping or subject to XCI
  - Data points as $\{(x_1,c_1),(x_2,c_2),..,(x_n,c_n)\}$, where x_i is the m-dimensional vector, c_i is the label

- Cross-validation procedure for SVM classifiers:
  - In an iterative fashion linear SVM models were trained on all the genes (groups of genes) except one
  - The resulting SVM classifier was used to predict the inactivation status of the held-out gene or all genes in the held-out group
  - Prediction accuracy was calculated based on results for all the genes in the set

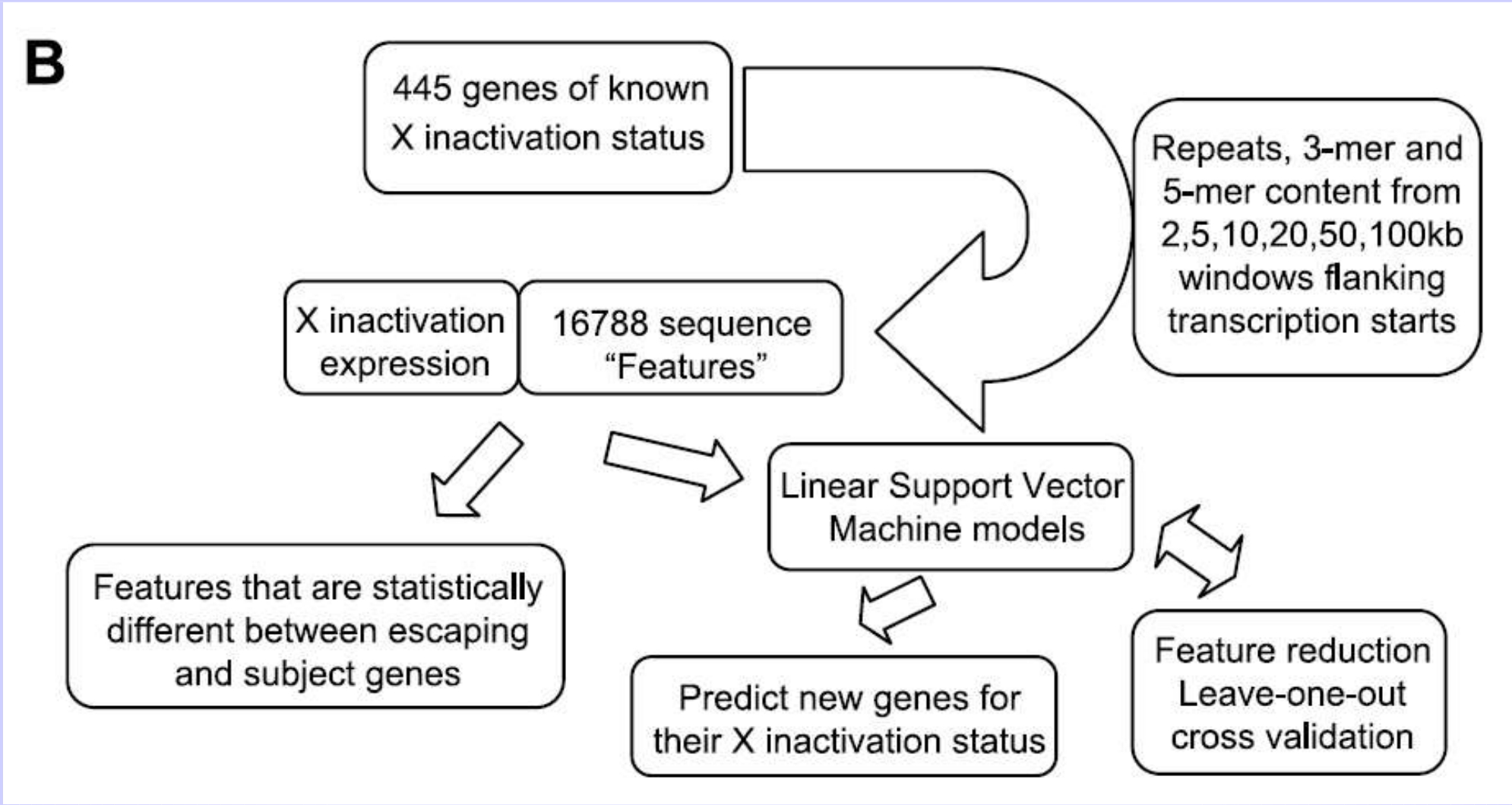


Fig1B. Strategy for statistical analysis and SVM training and classification.

Results

Table 2. Classification Accuracy for XAR Genes, XAR ESTs, and XCR Genes Using 5,596 Features from 50-kb and 100-kb Windows around Transcription Start Sites of the Genes

Training Set	Accuracy	Escape	Subject	Total
XAR (all genes)	Grouped genes CV	84% (42/50)	80% (48/60)	82% (90/110)
	Leave-one-out	76% (38/50)	85% (51/60)	81% (89/110)
	EST prediction	62% (8/13)	100% (10/10)	78% (18/23)
	Leave-one-out with Y-homology	70% (35/50)	85% (51/60)	78% (86/110)
XAR (without "border genes")	Leave-one-out	78% (28/36)	93% (43/46)	87% (71/82)
	EST prediction	46% (6/13)	100% (10/10)	70% (16/23)
	XCR prediction	17% (4/23)	92% (289/315)	87 % (293/338)

Features important for classification

- to retrieve the most important features by the SVM, the recursive feature selection process was performed 100 times

- in each iteration a random selected set of 2/3 of the XAR nonborder genes was used

- 53 features perform as well as the full complement of features

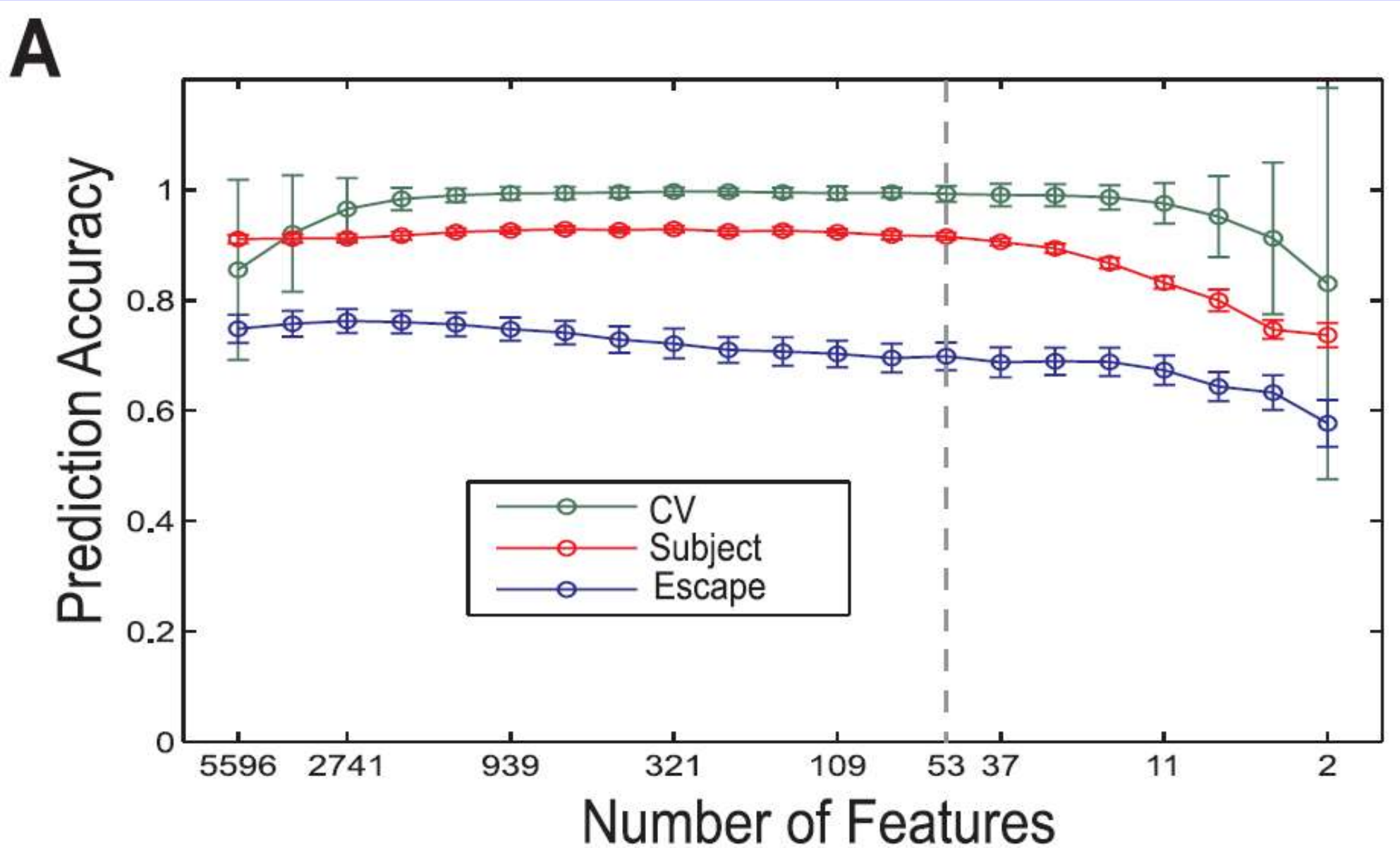
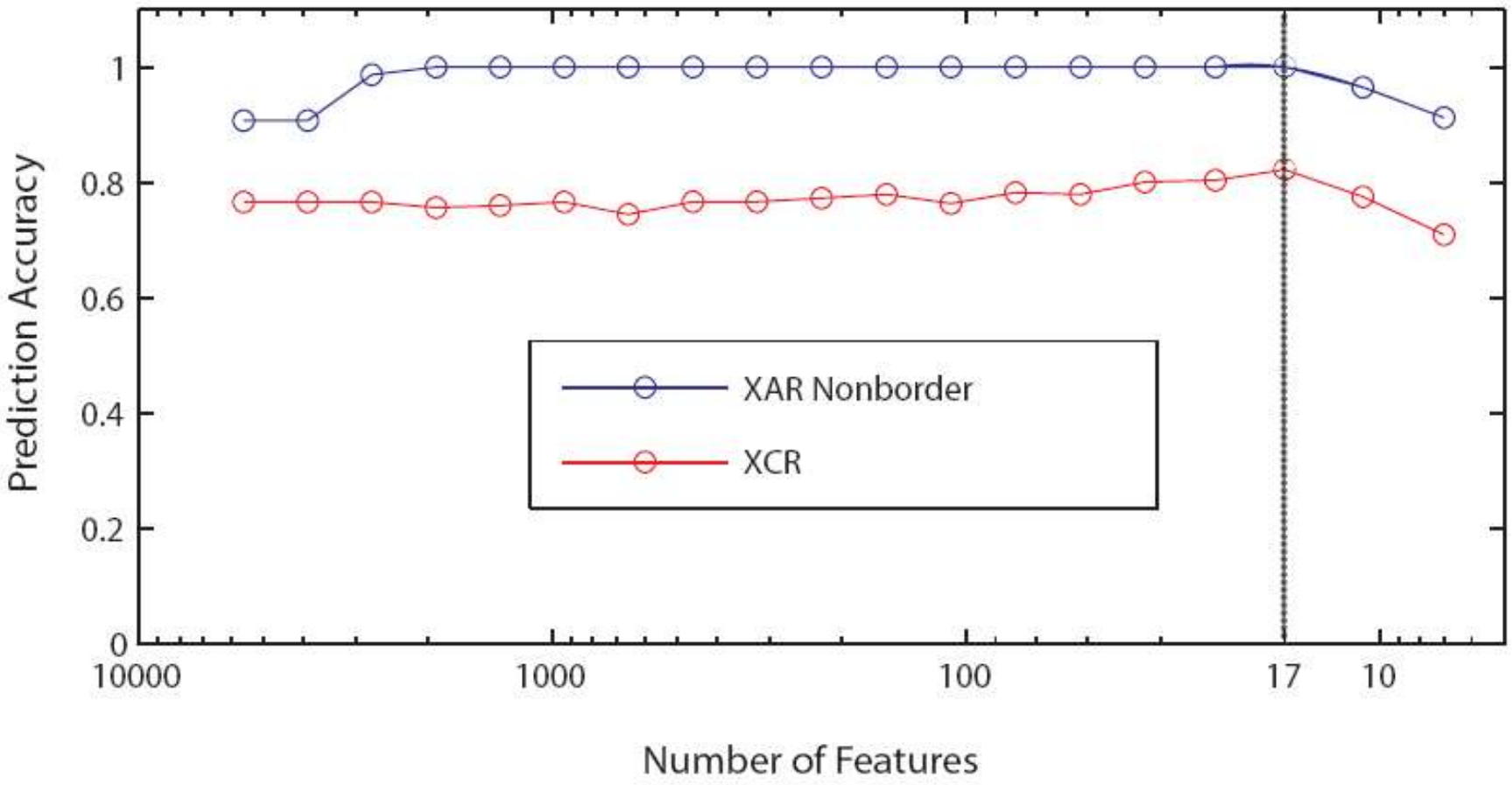


Fig 2A. Recursive Feature Reduction across the XAR Nonborder Genes

A set of 12 Features Can Accurately Classify X Inactivation



1. Feature reduction experiment until a set of 17 features
2. Hierarchical clustering and principle component analysis to refine the selected set

SVM classifier using 12 features was constructed

Table 4. Classification Accuracy for XAR Genes, XAR ESTs, and XCR Genes Using a Reduced Set of 12 Features and XAR Nonborder Genes

Dataset/Accuracy	Escape	Subject	Total
XAR leave-one-out	89% (32/36)	89% (41/46)	89% (73/82)
XAR ESTs	54% (7/13)	90% (9/10)	70% (16/23)
XCR	22% (5/23)	85% (268/315)	81% (273/338)

SVM classifier using 12 features was constructed

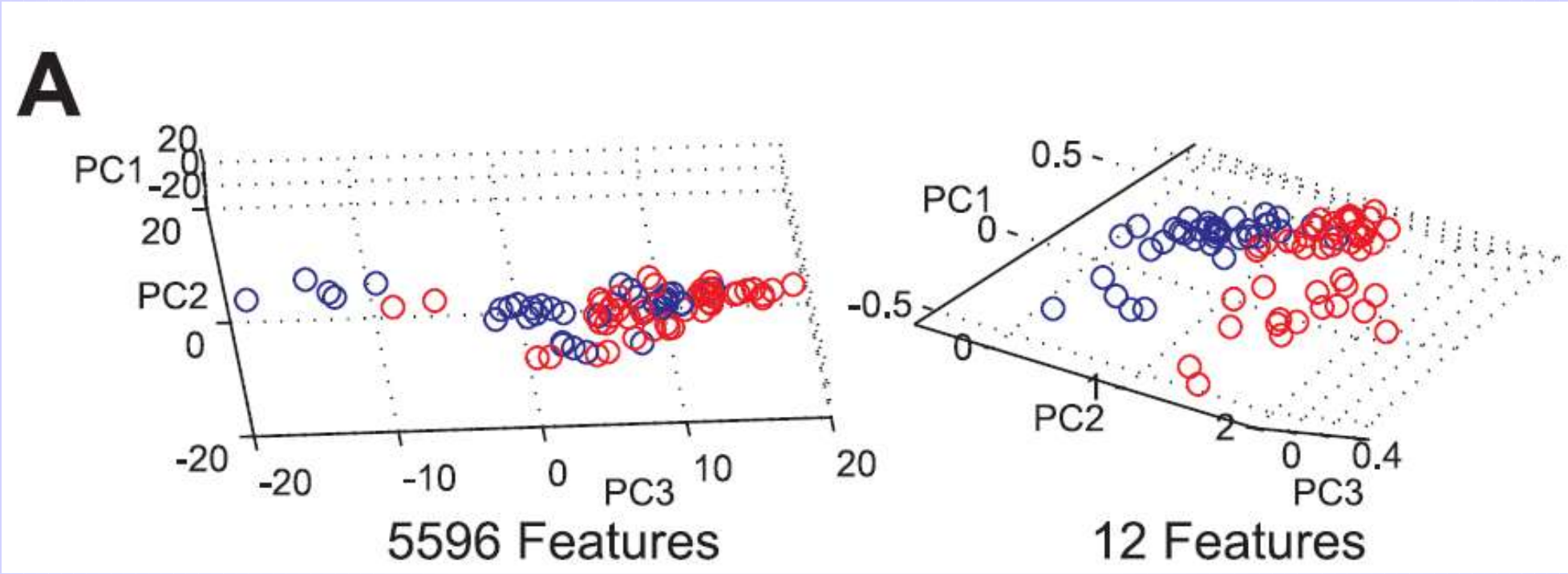


Fig 3A. The significance of the 12 selected features. The three best principle components among all 5,596 features for 50-kb and 100-kb windows (left) and the selected 12 features (right) for the 82 nonborder genes are shown projected onto a 3-D graph. Escaping genes are represented as blue circles and subject genes as red circles.

Confidence of SVM Classifiers

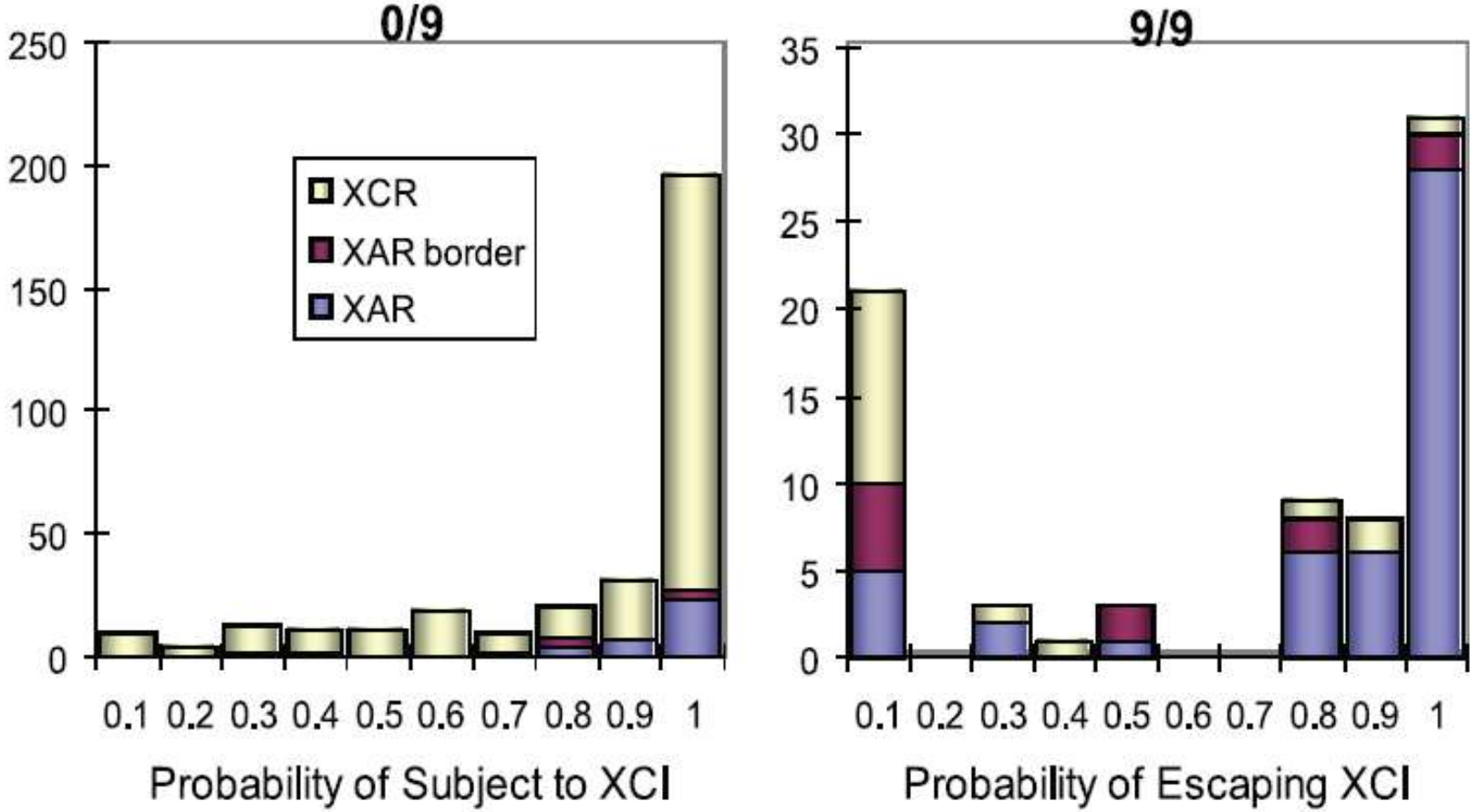


Figure 5. The Distribution of SVM Prediction Probabilities for Genes with Known X Inactivation Status

These histograms summarize the prediction probabilities of genes that are either (A) subject to inactivation (expressed in zero of nine somatic cell hybrids) or (B) escape from inactivation (expressed in nine of nine hybrids) [12]. Genes from the XCR, XAR border genes, and nonborder XAR genes coupled with XAR ESTs are represented by different colors. XCI, X chromosome inactivation.

Summary

- Based solely on primary DNA sequence, linear SVM classifiers can correctly predict 80% of all the genes

- Most, if not all, of the information necessary to determine X inactivation status is embedded in primary DNA sequence

- Information about XCI can be represented by as few as 12 sequence features.

The main article this seminar was based on:

Evidence of Influence of Genomic DNA Sequence on Human X Chromosome Inactivation. (2006). Wang Z, Willard HF, Mukherjee S, Furey TS. *PLOS Comput. Biol*, 2(9)

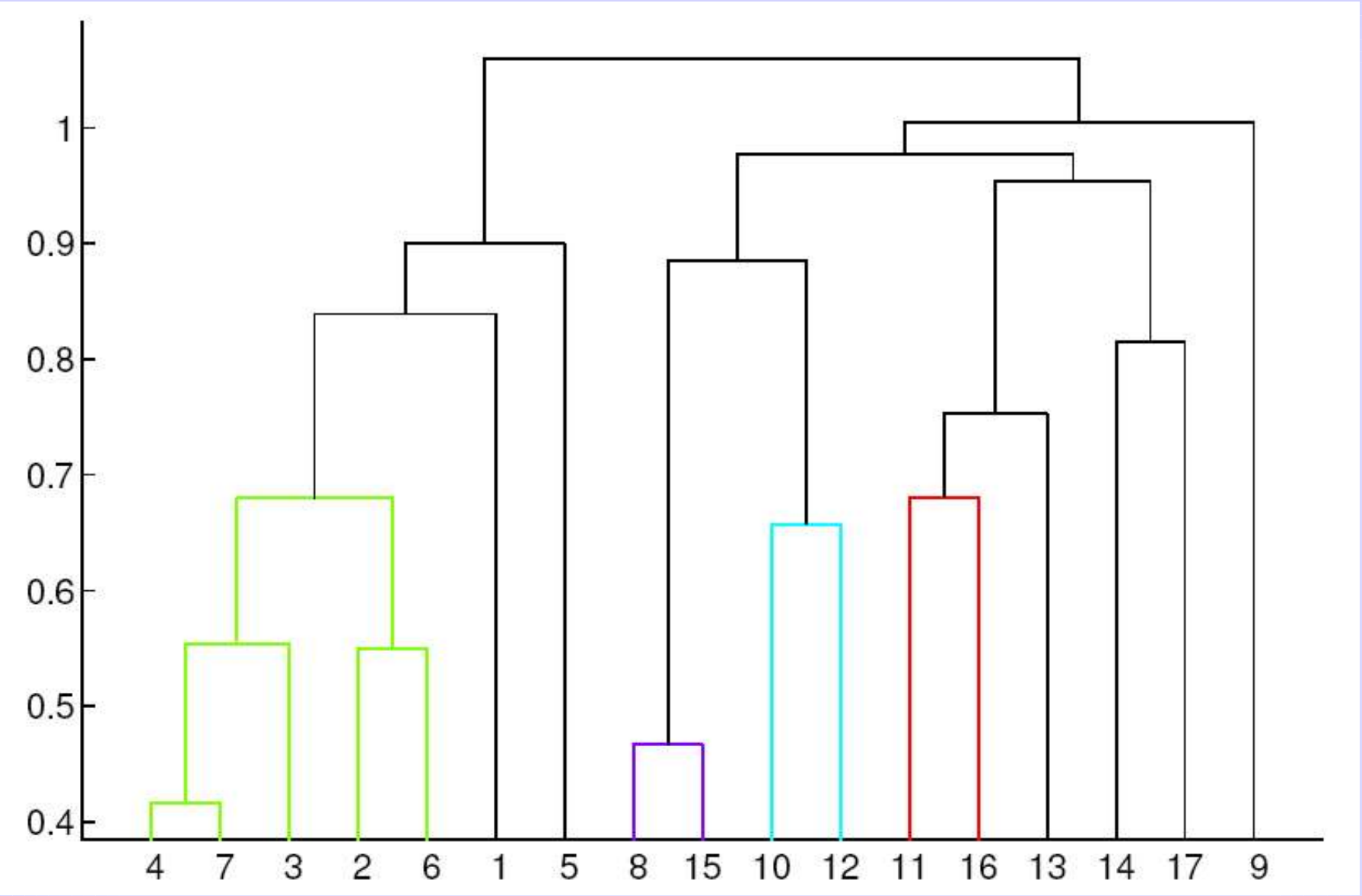


Fig. Threshold 0.7 was used to get 12? features from 17