

The smallest commonly deleted region encompassed 1.8 Mbp (chr12:14746099-16521376) and contained 16 genes. This suggests that besides haploinsufficiency of *CDKN1B*, another tumor suppressor gene may be

involved.¹⁹ Of note, two samples (TP166, TP34) harbored breakpoints directly upstream of the promiscuous fusion partner *ETV6*, which is also heterozygously deleted in these samples. This could be an indicator that

Table 2. Homozygous deletions: a list of all homozygous deletions detected in our data set and the genes affected.

Sample	Region	Start-end	Gene symbol / Ref seq (NCBI)	Gene name
TP168	8p23.2	4917841-4975677	No gene	
TP25	20p12.1	14985065-15034862	<i>MACROD2 (NM_080676.5)</i>	MACRO domain containing 2 isoform 1
TP34	8p11.23	38923786-38959748	<i>PLEKHA2 (NM_021623.1)</i> <i>HTRA4 (NM_153692.2)</i>	Pleckstrin homology domain containing family A HtrA serine peptidase 4
	8p11.21	40000344-40068118	No gene	
	9p12	42937610-43863873	No gene	
	9q31.1	102041415-102711737	<i>BC038565</i>	
	22q11.1	15962234-16119107	<i>IL17R (NM_014339)</i>	Interleukin 17 receptor precursor
		CECR6 (NM_031890.2) CECR5 (NM_033070.2) CECR1 (NM_177405.1)		Cat eye syndrome chromosome region candidate 6 Cat eye syndrome chromosome region candidate 5 Cat eye syndrome critical region protein 1
TP35	6q12	67068322-67105350	No gene, <i>CNP</i>	
TP4	3p13	71276619-71682261	<i>FOXP1 NM_032682.4</i>	Forkhead box P1 isoform 1

Table 3. Regions of acquired uniparental disomy or allelic imbalance.

Sample	Chromosome	Copy number	Start	End	Length (Mbp)
TP166	17q11.1-q25.3	2	22691152	78598059	55.91
TP168	12q21.33-q24.33	2	91163204	132387995	41.22
TP21	5q23.1-q23.2	2	115257555	123818130	8.56
	17q11.2-q25.3	2	27219405	78598059	51.38
TP28	4q13.3-q24	2	70683042	102463716	31.78
	17q21.2-q25.3	2	36884548	78598059	41.71
TP34	8p11.22-22.3	variable	39639901	105716979	66.08
TP37	2p25.1-22.3	2	7952368	32254818	24.30
	3p25.3-21.31	2	10996002	45379802	34.38
	3q23-25.1	2	141171610	151356745	10.19
	3q28-q29	2	191479491	197803570	6.32
	13q31.3-33.1	2	90823104	102071466	11.25
	20p13-12.3	2	169537	7213069	7.04
	20p11.1-q13.33	2	26170699	60036888	33.87
	22q12.3-13.31	2	35263007	43891222	8.63
TP41	8p23.3-p12	2*	180568	36965267	36.78
	11p15.4-q25	2*	3568345	134437775	130.87
	13q11-q34	2*	18380972	114092980	95.71
	17p13.3-p11.2	2*	18901	21459693	21.44
	17q24.3-25.3	2*	68346414	78598059	10.25
TP43	22q11.22-12.3	2*	19767837	30681028	10.91
	5q11.2-q13.2	2	57251185	71082801	13.83
TP57	5q13.1-q14.2	2	67311044	82317258	15.01
	16p13.2-p12.3	2	8822406	16884402	8.06
	18q22.3-q23	2	70382731	75795156	5.41
TP651	2p21	2	42208606	46496503	4.29
	20q11.21-q11.23	3	30400408	35663935	5.26
	20q13.12	3 or 4	43326890	45079291	1.75
	20q13.31-q13.32	3 or 4	54307144	56564113	2.26
	20q13.33	3 or 4	59395068	62111653	2.72

A list of all regions of acquired UPD (i.e. loss of heterozygosity despite a genomic copy number of two) detected by molecular allelotyping with 250K SNP arrays in our T-PLL samples. *occurred in a sample with tetraploidy.

ETV6 may be the target of these two deletions. As reported for other hematologic malignancies, the presence of fusions involving the *ETV6* gene is often associated with a deletion or lack of expression of the other *ETV6* allele.^{26,27}

A common heterozygous deletion on chromosome 13

occurred in six samples (33%) measuring 1.5 Mbp (chr13: 48879065-50392988). This region is commonly deleted in chronic lymphocytic leukemia and contains two microRNA (*miR-15a* and *miR-16-1*),²⁸ which regulate a significant number of cancer-related genes.²⁹

Chromosome 17 often has genomic alterations in can-

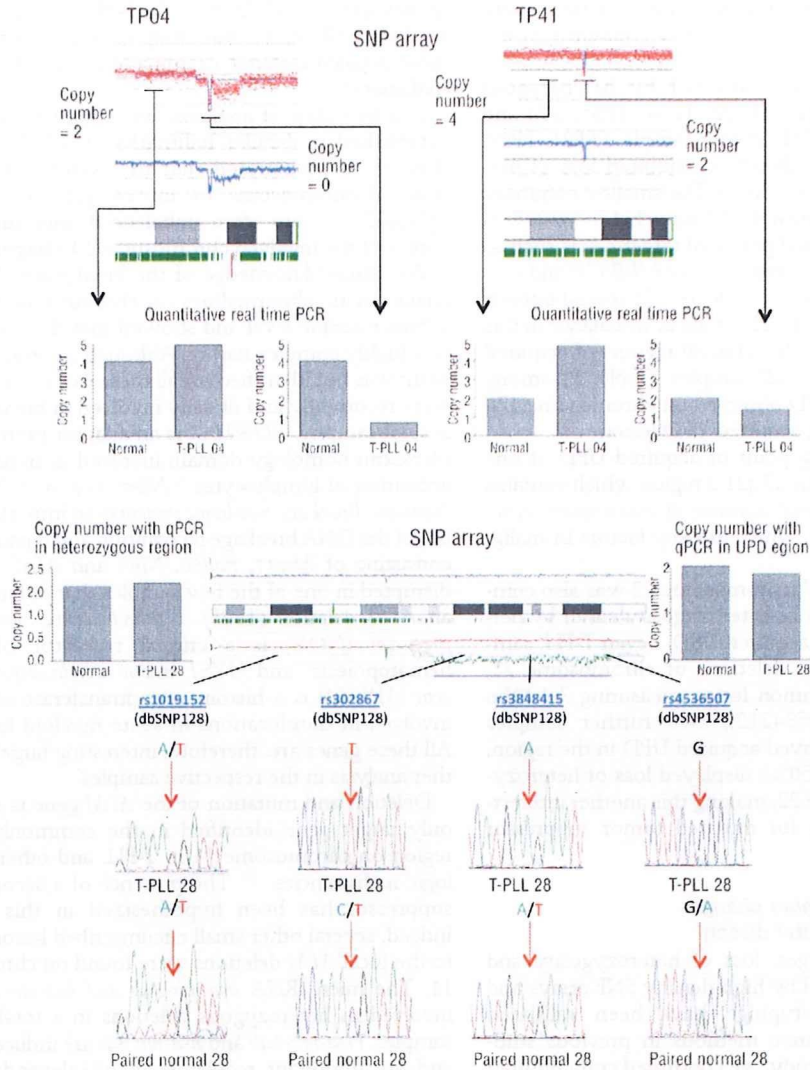


Figure 3. Validation of copy number analysis and acquired UPD. (A) Copy number results obtained by SNP arrays were validated by performing quantitative real-time PCR on genomic DNA of samples containing putative deletions of the *FOXP1* gene. Chromosome views of chromosome 3 are displayed. In sample TP04, *FOXP1* was putatively homozygously deleted (left image, *FOXP1* locus = copy number 0, adjacent regions = copy number: two), while in sample TP41, *FOXP1* was deleted heterozygously in a tetraploid setting (right image, *FOXP1* locus = copy number: two, adjacent regions = copy number: four). Quantitative real-time PCR in the corresponding T-PLL samples and their matched normal DNA confirmed the copy number states estimated by the allelotyping software. (B) Acquired UPD detected by SNP arrays was validated using sample TP28 on chromosome 17. The chromosome view of chromosome 17 in sample TP28 is depicted. Acquired UPD is present on chromosome 17q, as visualized by the divergence of the estimated allele specific copy number (red and green lines in the lower panel of the image), indicating the duplication of one allele and concomitant loss of the other allele and loss of heterozygosity, as evidenced by the abrupt absence of heterozygous SNP calls (vertical green bars directly below the cytoband image). Chromatographs of sequenced SNPs within the acquired UPD region and the adjacent heterozygous region in T-PLL DNA (TP28) and matched normal DNA (paired normal DNA) show that SNPs were homozygous in the T-PLL sample in the acquired UPD region and heterozygous in the matched normal DNA. In the adjacent region, which did not display acquired UPD, the sequenced SNPs were heterozygous in both T-PLL and matched normal samples. Quantitative real-time PCR confirmed a copy number state of two in all regions sequenced in each of the samples.

cer. In total, chromosome 17 was affected by deletion, amplification or acquired UPD in 12 samples. Regions of chromosome 17p were heterozygously deleted in seven cases and affected by an acquired UPD in sample TP41, which featured tetrasomy of most of its chromosomes but showed a copy number of two with deletion of one allele on chromosome 17p. Therefore, eight samples (44%) showed loss of heterozygosity of various lengths on chromosome 17p and in five of these samples, the tumor suppressor *p53* was contained in the affected regions.

Chromosome 17q was affected by heterozygous deletions in five samples (TP22, TP34, TP35, TP4 and TP799) and acquired UPD in four samples (TP21, TP28, TP166 and TP41) and therefore exhibited loss of heterozygosity in nine cases (50%). The smallest commonly deleted region spanned 4.87 Mbp (chr17: 73729578-78599918) and contained potential target genes such as *baculoviral IAP repeat containing protein 5 (BIRC5)* and *suppressor of cytokine signaling 3 (SOCS3)*. Of special interest is the recurrent acquired UPD or allelic imbalance in this region in four samples. In total, 30 regions of acquired UPD were detected in all samples (Table 3); among these, the acquired UPD observed on chromosome 17q was the only recurring acquired UPD lesion.

Of note, the starting point of acquired UPD of one sample (TP28) lies in the 17q21.2 region which contains the *signal transducer and activator of transcription genes STAT5 A/B and STAT3*, which are key factors in malignant transformation.³⁰

Like chromosome 17, chromosome 22 was also commonly affected by loss of heterozygosity either by heterozygous deletion or acquired UPD. Seven T-PLL samples had heterozygous deletions of chromosome 22, with the smallest common lesion measuring 7.4 Mbp (chr.22: 22180211-29584212); two further samples (TP37 and TP41) displayed acquired UPD in the region, so that nine samples (50%) displayed loss of heterozygosity on chromosome 22, making this another interesting site for screening for mutated tumor suppressor genes.

Validation of copy number changes and acquired uniparental disomy

Copy number changes, loss of heterozygosity and acquired UPD detected by high density SNP arrays and molecular allelokaryotyping²¹ have been validated extensively by alternative methods in previous studies.^{27,31} In the current study, we confirmed copy number results by quantitative real-time PCR on the genomic DNA of the putatively deleted region and the adjacent region to the lesion in patients and matched normal samples in accordance to a method specifically designed to analyze genomic DNA by real-time PCR.³² We confirmed a homozygous deletion of the *FOXP1* gene on chromosome 3p in sample TP4 and a heterozygous deletion of this gene in sample TP41 (Figure 3A). Acquired UPD was confirmed on chromosome 17 in sample TP28 by sequencing SNPs in the region displaying acquired UPD as compared to the adjacent heterozygous region on the same chromosome (Figure 3B).

Discussion

In this study, we aimed to utilize the increased interrogational power of high density SNP arrays and molecular allelokaryotyping²¹ to refine the understanding of known genomic lesions and discover new ones present in the malignant cells of patients with T-PLL. With high quality genomic DNA from 18 T-PLL patients analyzed by 250K SNP arrays, this study currently represents the most detailed genomic examination of this hematologic malignancy.

In a first step of analysis, we sought to re-evaluate established molecular hallmarks of T-PLL. By doing this, we demonstrated that the well-known disruptions of chromosome 14, *inv(14)(q11q34)* or *t(14;14)(q11;q32)*,^{4,12,15} are often unbalanced, indicating excess copies of the involved chromosome 14q fragments.

We refined knowledge of the breakpoints leading to characteristic abnormalities on chromosome 8^{18,24,30} at a submicroscopic level and showed that they were often of a highly complex nature. Although a common breakpoint was not identified for all these cases, certain genes were recurrently and directly involved in breakpoints in several samples. *PLEKHA2* is an adaptor protein with a pleckstrin homology domain involved in signaling after activation of lymphocytes.³² *Nibrin isoform 2 (NBN)*, the *Nijmegen Breakage Syndrome protein* is an important member of the DNA breakage recognition and repair complex consisting of *Mre11*, *Rad50*, *Nbs1* and *ATM*¹⁶ and was disrupted in one of the few samples that did not exhibit an *ATM* deletion (TP25). *Nephroblastoma overexpressed precursor (NOV)* is a critical regulator of human hematopoiesis³³ and *MYST histone acetyltransferase monocytic (MYST3)* is a histone acetyltransferase commonly involved in translocations in acute myeloid leukemia.³⁴ All these genes are, therefore, interesting targets for further analysis in the respective samples.

Deletion and mutation of the *ATM* gene is so far the only target gene identified in the commonly deleted region on chromosome 11 in T-PLL and other hematologic malignancies.^{6,7,18} The presence of a second tumor suppressor has been hypothesized in this region.³¹ Indeed, several other small circumscribed lesions similar to the focal *ATM* deletions were found on chromosome 11. The micro RNA *hsa-mir-34b* and *hsa-mir-34c* were involved in heterozygous deletions in a total of eight samples. *Hsa-mir-34b* and *hsa-mir-34c* are induced by *p53* and are important regulators in *p53*-dependent pathways³⁵ and may represent interesting targets, as reduced expression of *mir-34s* has been found in several tumors.³⁶⁻³⁸ Furthermore, the oncogenes *ETS1* and *FLI1* were encompassed in small confined deletions in two samples. Disruption of both of these genes has been determined to be an initiating event in malignant transformation of hematologic diseases and solid tumors.^{39,40}

In the subsequent search for new common genomic lesions in T-PLL, the *FOXP1* gene was detected in a small homozygous deletion in one sample and was heterozygously deleted in two other samples. *FOXP1* is a member of the FOX family of transcription factors and is involved in the development of the heart, lungs and lym-

phocytes.⁴¹ Deletion and loss of expression in breast cancer confers a worse prognosis⁴² and this gene is targeted by recurrent chromosome translocations in mucosal-associated lymphoid tissue (MALT) lymphoma.⁴³ We confirmed the deletion of this gene by quantitative real-time PCR. Although screening for mutations of this gene in samples containing heterozygous deletions showed no alterations in this respect, the *FOXP1* gene could also be involved in a fusion gene, as recently detected in acute lymphoblastic leukemia.⁴⁴ The underlying mechanism for accumulated gene fusions in T-PLL could be increased aberrant V(D)J recombination due to mutation of the *ATM* gene.⁴⁵ The concept of perturbed V(D)J recombination has recently also been shown to be responsible for common deletions of the *Ikaros* gene in Philadelphia chromosome-positive acute lymphoblastic leukemia.⁴⁶

Newly detected common lesions with recurrence in six or more samples were found on chromosomes 5p, 12p, 13q, 17 and 22. Loss of heterozygosity, either by heterozygous deletion or acquired UPD, in nine of 18 samples on both chromosome 17 and chromosome 22 suggests common lesions specific to T-PLL. Loss of chromosome 17p is a common phenomenon in chronic lymphocytic leukemia; it is known to confer a worse prognosis and bad response to chemotherapy, possibly through disruption of the *p53* pathway.⁴⁷ However, loss of heterozygosity of chromosome 17q at such high frequency has not been reported for leukemic diseases. This finding was further corroborated by the observation of a common region of acquired UPD in four of our samples on chromosome 17q. UPD can either arise through several mechanisms at the level of the gametes such as trisomy

rescue, compensatory UPD or gametic complementation or can develop due to a somatic recombinational event.⁴⁸ The UPD regions detected in our experiments are most probably acquired isodisomy that evolved due to somatic recombination events shown in our validation of acquired UPD, which demonstrated that loss of heterozygosity was only detectable in the tumor sample but not in the matched control. While acquired UPD has previously been described as a new genomic lesion in T-PLL,¹³ this is the first study to report a recurring (n=4) acquired UPD lesion on chromosome 17q.

In conclusion, the use of high density SNP arrays to genotype T-PLL has refined our knowledge of established genomic alterations and revealed numerous new candidate lesions by directly pinpointing affected genes for ongoing functional studies to elucidate the pathogenesis of T-PLL.

Authorship and Disclosures

DN analyzed the data, carried out validation experiments and wrote the paper. ELT and MHS designed the study, acquired samples and wrote the paper. NK designed this study, analyzed the data, acquired samples and wrote the paper. TA analyzed the data, MJD designed the study and acquired samples, WKH wrote the paper, SO designed the study, performed SNP array experiments, molecular allelotyping and acquired samples, HPK designed the study, analyzed the data and wrote the paper.

The authors reported no potential conflicts of interest.

References

- Dearden CE. T-cell prolymphocytic leukemia. *Med Oncol* 2006;23:17-22.
- Dearden CE, Matutes E, Cazin B, Tjonnfjord GE, Parreira A, Nomdedeu B, et al. High remission rate in T-cell prolymphocytic leukemia with CAMPATH-1H. *Blood* 2001;98:1721-6.
- Keating MJ, Cazin B, Coutre S, Bihiray R, Kovacsics T, Langer W, et al. Campath-1H treatment of T-cell prolymphocytic leukemia in patients for whom at least one prior chemotherapy regimen has failed. *J Clin Oncol* 2002;20:205-13.
- Pekarsky Y, Hallas C, Croce CM. Molecular basis of mature T-cell leukemia. *JAMA* 2001;286:2308-14.
- Stern MH, Soulier J, Rosenzweig M, Nakahara K, Canki-Klain N, Aurias A, et al. MTCP-1: a novel gene on the human chromosome Xq28 translocated to the T cell receptor alpha/delta locus in mature T cell proliferations. *Oncogene* 1993;8:2475-83.
- Stankovic T, Taylor AM, Yuille MR, Vorechovsky I. Recurrent ATM mutations in T-PLL on diverse haplotypes: no support for their germline origin. *Blood* 2001;97:1517-8.
- Stilgenbauer S, Schaffner C, Litterst A, Liebisch P, Gilad S, Bar-Shira A, et al. Biallelic mutations in the ATM gene in T-prolymphocytic leukemia. *Nat Med* 1997;3:1155-9.
- Maljaei SH, Brito-Babapulle V, Hioms LR, Catovsky D. Abnormalities of chromosomes 8, 11, 14, and X in T-prolymphocytic leukemia studied by fluorescence in situ hybridization. *Cancer Genet Cytogenet* 1998;103:110-6.
- Sorour A, Brito-Babapulle V, Smedley D, Yuille M, Catovsky D. Unusual breakpoint distribution of 8p abnormalities in T-prolymphocytic leukemia: a study with YACS mapping to 8p11-p12. *Cancer Genet Cytogenet* 2000;121:128-32.
- Stoppa-Lyonnet D, Soulier J, Lauge A, Dastot H, Garand R, Sigaux F, et al. Inactivation of the ATM gene in T-cell prolymphocytic leukemias. *Blood* 1998;91:3920-6.
- Stoppa-Lyonnet D, Lauge A, Sigaux F, Stern MH. No germline ATM mutation in a series of 16 T-cell prolymphocytic leukemias. *Blood* 2000;96:374-6.
- Teitell MA. The TCL1 family of oncoproteins: co-activators of transformation. *Nat Rev Cancer* 2005;5:640-8.
- Despoux G, Joiner M, Le Toriellac E, Weil R, Stern MH. The TCL1 oncoprotein inhibits activation-induced cell death by impairing PKCtheta and ERK pathways. *Blood* 2007;110:4406-16.
- Herling M, Patel KA, Teitell MA, Konopleva M, Ravandi F, Kobayashi R, et al. High TCL1 expression and intact T-cell receptor signaling define a hyperproliferative subset of T-cell prolymphocytic leukemia. *Blood* 2008;111:328-37.
- Shiloh Y. Ataxia-telangiectasia and the Nijmegen breakage syndrome: related disorders but genes apart. *Annu Rev Genet* 1997;31:635-62.
- Lavin MF. ATM and the Mre11 complex combine to recognize and signal DNA double-strand breaks. *Oncogene* 2007;26:7749-58.
- Lee JH, Paull TT. Activation and regulation of ATM kinase activity in response to DNA double-strand breaks. *Oncogene* 2007;26:7741-8.
- Durig J, Bug S, Klein-Hitpass L, Boes T, Jons T, Martin-Subero JJ, et al. Combined single nucleotide polymorphism-based genomic mapping and global gene expression profiling identifies novel chromosomal imbalances, mechanisms and candidate genes important in the pathogenesis of T-cell prolymphocytic leukemia with *inv(14)(q11q32)*. *Leukemia*

- 2007;21:2153-63.
19. Le Toriellac E, Despouy G, Pierron G, Gaye N, Joiner M, Bellanger D, et al. Haploinsufficiency of CDKN1B contributes to leukemogenesis in T-cell prolymphocytic leukemia. *Blood* 2008;111:2321-8.
 20. Dutt A, Beroukhir R. Single nucleotide polymorphism array analysis of cancer. *Curr Opin Oncol* 2007;19:43-9.
 21. Yamamoto G, Nannya Y, Kato M, Sanada M, Levine RL, Kawamata N, et al. Highly sensitive method for genomewide detection of allelic composition in nonpaired, primary tumor specimens by use of Affymetrix single-nucleotide-polymorphism genotyping microarrays. *Am J Hum Genet* 2007;81:114-26.
 22. Nannya Y, Sanada M, Nakazaki K, Hosoya N, Wang L, Hangaishi A, et al. A robust algorithm for copy number detection using high-density oligonucleotide single nucleotide polymorphism genotyping arrays. *Cancer Res* 2005;65:6071-9.
 23. Weksberg R, Hughes S, Moldovan L, Bassett AS, Chow EW, Squire JA. A method for accurate detection of genomic microdeletions using real-time quantitative PCR. *BMC Genomics* 2005;6:180.
 24. Soulier J, Pierron G, Vecchione D, Garand R, Brizard F, Sigaux F, et al. A complex pattern of recurrent chromosomal losses and gains in T-cell prolymphocytic leukemia. *Genes Chromosomes Cancer* 2001;31:248-54.
 25. Kobayashi H, Montgomery KT, Bohlander SK, Adra CN, Lim BL, Kucherlapati RS, et al. Fluorescence in situ hybridization mapping of translocations and deletions involving the short arm of human chromosome 12 in malignant hematologic diseases. *Blood* 1994;84:3473-82.
 26. Bohlander SK. ETV6: a versatile player in leukemogenesis. *Semin Cancer Biol* 2005;15:162-74.
 27. Kawamata N, Ogawa S, Zimmermann M, Kato M, Sanada M, Hemminki K, et al. Molecular allelotyping of pediatric acute lymphoblastic leukemias by high-resolution single nucleotide polymorphism oligonucleotide genomic microarray. *Blood* 2008;111:776-84.
 28. Calin GA, Dumitru CD, Shimizu M, Bichi R, Zupo S, Noch E, et al. Frequent deletions and down-regulation of micro-RNA genes miR15 and miR16 at 13q14 in chronic lymphocytic leukemia. *Proc Natl Acad Sci USA* 2002;99:15524-9.
 29. Calin GA, Cimmino A, Fabbri M, Ferracin M, Wojcik SE, Shimizu M, et al. MiR-15a and miR-16-1 cluster functions in human leukemia. *Proc Natl Acad Sci USA* 2008;105:5166-71.
 30. Constantinescu SN, Girardot M, Pecquet C. Mining for JAK-STAT mutations in cancer. *Trends Biochem Sci* 2008;33:122-31.
 31. Lehmann S, Ogawa S, Raynaud SD, Sanada M, Nannya Y, Ticchioni M, et al. Molecular allelotyping of early-stage, untreated chronic lymphocytic leukemia. *Cancer* 2008;112:1296-305.
 32. Allam A, Marshall AJ. Role of the adaptor proteins Bam32, TAPP1 and TAPP2 in lymphocyte activation. *Immunol Lett* 2005;97:7-17.
 33. Gupta R, Hong D, Iborra F, Sarno S, Enver T. NOV (CCN3) functions as a regulator of human hematopoietic stem or progenitor cells. *Science* 2007;316:590-3.
 34. Troke PJ, Kindle KB, Collins HM, Heery DM. MOZ fusion proteins in acute myeloid leukaemia. *Biochem Soc Symp* 2006:23-39.
 35. Hermeking H. p53 enters the microRNA world. *Cancer Cell* 2007;12:414-8.
 36. Welch C, Chen Y, Stallings RL. MicroRNA-34a functions as a potential tumor suppressor by inducing apoptosis in neuroblastoma cells. *Oncogene* 2007;26:5017-22.
 37. Chang TC, Wentzel EA, Kent OA, Ramachandran K, Mullendore M, Lee KH, et al. Transactivation of miR-34a by p53 broadly influences gene expression and promotes apoptosis. *Mol Cell* 2007;26:745-52.
 38. Bommer GT, Gerin I, Feng Y, Kaczorowski AJ, Quick R, Love RE, et al. p53-mediated activation of miRNA34 candidate tumor-suppressor genes. *Curr Biol* 2007;17:1298-307.
 39. Dittmer J. The biology of the Ets1 proto-oncogene. *Mol Cancer* 2003;2:29.
 40. Truong AH, Ben-David Y. The role of Fli-1 in normal cell function and malignant transformation. *Oncogene* 2000;19:6482-9.
 41. Koon HB, Ippolito GC, Banham AH, Tucker PW. FOXP1: a potential therapeutic target in cancer. *Expert Opin Ther Targets* 2007;11:955-65.
 42. Fox SB, Brown P, Han C, Ashe S, Leek RD, Harris AL, et al. Expression of the forkhead transcription factor FOXP1 is associated with estrogen receptor α and improved survival in primary human breast carcinomas. *Clin Cancer Res* 2004;10:3521-7.
 43. Du MQ. MALT lymphoma: recent advances in aetiology and molecular genetics. *J Clin Exp Hematop* 2007;47:31-42.
 44. Kawamata N, Ogawa S, Zimmermann M, Niebuhr B, Stocking C, Sanada M, et al. Cloning of genes involved in chromosomal translocations by high-resolution single nucleotide polymorphism genomic microarray. *Proc Natl Acad Sci USA* 2008;105:11921-6.
 45. Bredemeyer AL, Huang CY, Walker LM, Bassing CH, Sleckman BP. Aberrant V(D)J recombination in ataxia telangiectasia mutated-deficient lymphocytes is dependent on nonhomologous DNA end joining. *J Immunol* 2008;181:2620-5.
 46. Mullighan CG, Miller CB, Radtke I, Phillips LA, Dalton J, Ma J, et al. BCR-ABL1 lymphoblastic leukaemia is characterized by the deletion of Ikaros. *Nature* 2008;453:110-4.
 47. Zenz T, Dohner H, Stilgenbauer S. Genetics and risk-stratified approach to therapy in chronic lymphocytic leukemia. *Best Pract Res Clin Haematol* 2007;20:439-53.
 48. Robinson WP. Mechanisms leading to uniparental disomy and their clinical consequences. *Bioessays* 2000;22:452-9.

Robust Model Selection for Classification of Microarrays

Ikumi Suzuki¹, Takashi Takenouchi¹, Miki Ohira², Shigeyuki Oba^{3,4} and Shin Ishii^{1,3}

¹Graduate School of Information Science, Nara Institute of Science and Technology, Takayama, Ikoma, Nara 630-0192, Japan. ²Division of Biochemistry, Chiba Cancer Center Research Institute, Chiba 260-8717, Japan. ³Graduate School of Informatics, Kyoto University, Gokasho, Uji, Kyoto 611-0011, Japan. ⁴PRESTO, Japan Science and Technology Corporation.

Abstract: Recently, microarray-based cancer diagnosis systems have been increasingly investigated. However, cost reduction and reliability assurance of such diagnosis systems are still remaining problems in real clinical scenes. To reduce the cost, we need a supervised classifier involving the smallest number of genes, as long as the classifier is sufficiently reliable. To achieve a reliable classifier, we should assess candidate classifiers and select the best one. In the selection process of the best classifier, however, the assessment criterion must involve large variance because of limited number of samples and non-negligible observation noise. Therefore, even if a classifier with a very small number of genes exhibited the smallest leave-one-out cross-validation (LOO) error rate, it would not necessarily be reliable because classifiers based on a small number of genes tend to show large variance. We propose a robust model selection criterion, the min-max criterion, based on a resampling bootstrap simulation to assess the variance of estimation of classification error rates. We applied our assessment framework to four published real gene expression datasets and one synthetic dataset. We found that a state-of-the-art procedure, weighted voting classifiers with LOO criterion, had a non-negligible risk of selecting extremely poor classifiers and, on the other hand, that the new min-max criterion could eliminate that risk. These findings suggest that our criterion presents a safer procedure to design a practical cancer diagnosis system.

Keywords: gene expression, cancer diagnosis, mini-chip microarrays, supervised analysis

Cancer Informatics 2009;7 141–157

This article is available from <http://www.la-press.com>.

© the authors, licensee Libertas Academica Ltd.

This is an open access article distributed under the terms of the Creative Commons Attribution License (<http://www.creativecommons.org/licenses/by/2.0>) which permits unrestricted use, distribution and reproduction provided the original work is properly cited.



1. Introduction

Microarray technology¹ has been applied to predict prognosis of cancer patients by comparing gene expression profiles in cancer tissue samples, and its predictive power has been demonstrated for many types of cancers.²⁻⁵ The prognosis prediction systems based on microarrays have been expected to be new efficient bio-markers that enable personalized cancer medicine.⁶ We consider, in this paper, two problems in expanding the use of microarray-based prediction systems in real clinical scenes, namely, observation cost and reliability.⁷

To reduce the observation cost without losing reliability, there have been several efforts to design diagnosis systems involving small numbers of specially selected genes. Recently, specialized diagnostic microarrays harboring small numbers of genes, to say tens or hundreds genes, are developed based on a supervised analysis with a dataset taken by a full microarray system involving thousands or tens of thousands of genes.^{5,8,9} Measurement cost per patient becomes smaller by reducing the number of genes that is involved in such a system. If number of spots on a chip is fixed, more spots corresponding to a single gene can be included in a chip, which enables more reliable measurement by averaging multiple spots of same genes, and/or more efficient measurement by diagnosing multiple patients simultaneously in a single chip.⁸ Manufacturing cost of a chip can be reduced by designing mini-chip harboring small number of spots.⁵

To achieve a reliable predictor, a well-known trade-off problem exists even if the above-mentioned issue of observation cost is omitted; we should select as large a number of informative genes and as small a number of non-informative genes as possible. We often need a certain number of genes to gain prediction accuracy, partly because multiple informative genes tend to provide different kinds of information which are complementary to each other for the prediction, and partly because, even when a set of multiple genes provides identical information, observation noise can be reduced by averaging them. On the other hand, since the prediction error increases when non-informative genes are included, we need to reduce the number of non-informative genes, putting the observation cost aside. These two demands are a trade-off because the process of determining whether

each gene is informative or non-informative itself is not always reliable enough, due to non-negligible noise and a limited number of observations.

In summary, our goal can be stated as to achieve a reliable predictor based on as few genes as possible, which is accomplished in a supervised analysis with the following three processes:

- a gene selection process,
- a supervised learning process that constructs predictors based on a labeled set of expression data of the selected genes, and
- an assessment process for the constructed candidate predictors.

There have been many options proposed for the first two processes, and comparisons of their combinations were made from the viewpoint of prediction error rates on test datasets, namely generalization performances.^{10,11} In the present study, we use the following two procedures that were applied in the previous study.¹²

- Weighted voting (WV) classifier¹³ with gene selection based on absolute t-score (T-WV)
- Linear-kernel support vector machine (SVM)¹⁴ with recursive elimination of genes that have the smallest contribution to current classification performance (R-SVM).¹⁵

These procedures construct multiple candidate predictors with various numbers of genes included in the predictors. Since their prediction performances for independent test datasets depend on the number of genes, their assessment is crucial.

In the assessment process, the prediction performance of each candidate predictor is estimated based on the training data, and good estimation is obtained by reducing the estimation bias and the variance. Since the true performance on independent data in the future is unknown, we should select the best predictor with less bias and smaller variance of the estimated performance. In general, the bias-variance trade-off problem is inherent to all statistical models used for prediction, especially in the classification framework.^{16,17} For prognosis prediction by microarray, several past studies focused on reducing the estimation biases of the prediction error rates in determining the best model¹⁸⁻²⁰ because inclusion of biases could lead to over-estimation of



the classification performance of the proposed system. The cross-validation (CV) technique is used widely for predicting true classification error rate in samples that are not included in either the training or the test sample sets. Among the CV methods, the leave-one-out cross-validation technique (LOO) is often used because of its small bias.¹⁸ These studies, however, paid little attention to the variances of estimated classification error rates.

The estimated variances in the assessment process are important for practical applications. Even if a classifier has a sufficiently low error rate accompanied instead by large variance in prediction, it suffers from a high risk of having a large actual error rate when applied to unknown test samples.²¹ The LOO criterion sometimes selects a classifier involving a very small number of genes, or even a single gene. Although the single-gene classifier fits the ‘as few genes as possible’ criterion, classifiers involving redundant genes tend to exhibit lower noise and provide better prognosis.⁹ Several recent methods consider the estimated error rate variances,^{21–24} and unsupervised methods^{25,26} also minimize the variance of the model by focusing on the stability of the signatures instead of on the supervised class labels. However, there has been no discussion from the viewpoint of mini-chip design, namely, to explore a reliable predictor based on as few genes as possible.

In the present study, we consider both the bias and the variance of performance estimation so as to achieve a reliable predictor. We applied a bootstrap sampling method to estimate the distribution of possible error rates, with bias and variance, and propose a min-max criterion to obtain a stable classifier. We conducted a simulation study and found that the min-max criterion tends to select better candidate predictors than the LOO criterion, especially when the number of samples is small. We then compared two supervised analysis procedures, T-WV and R-SVM, and showed that T-WV achieves reliable predictors with a small number of genes, indicating that T-WV with the min-max criterion is desirable for our purpose of obtaining a reliable predictor with as few genes as possible.

2. Methods

2.1. Notations

Let $x_i = (x_{i1}, \dots, x_{iM})$ be a vector of the M -dimensional gene expression profile of the i -th sample, and y_i

a binary class label $y_i \in \{-1, 1\}$ representing the binary status of the i -th sample, for example, tumor or non-tumor. The numbers of samples in the negative ($y_i = -1$) and positive ($y_i = 1$) classes are denoted as n_n and n_p , respectively. Suppose that we have a dataset $D = \{d_i | i = 1, \dots, N\}$ including N samples, where $d_i = (x_i, y_i)$ is a pair of input (expression) and output (class label) of the i -th sample. By applying a supervised machine learning method to the dataset D , we construct a discriminant function $h(x | D)$ such that we predict a label $\hat{y}(x')$ for a new input x' by

$$\hat{y}(x') = \begin{cases} 1 & \text{if } h(x' | D) \geq 0 \\ -1 & \text{if } h(x' | D) < 0 \end{cases} \quad (1)$$

2.2. T-WV method

The WV method is a typical supervised machine learning method that employs the top k significant genes. Since the significance of the j -th gene is defined according to the following t-score, the entire procedure is referred to as the T-WV method,

$$t_j = \frac{\bar{x}_{pj} - \bar{x}_{nj}}{\sqrt{1/n_p + 1/n_n} S_j}, \quad (2)$$

where \bar{x}_{pj} and \bar{x}_{nj} are the average expression levels of the j -th gene over the training samples labeled 1 and -1 , respectively, and S_j^2 is the pooled within-class variance of the j -th gene,

$$S_j^2 = \frac{\sum_{i:y_i=-1} (x_{ij} - \bar{x}_{nj})^2 + \sum_{i:y_i=1} (x_{ij} - \bar{x}_{pj})^2}{n_n + n_p - 2}. \quad (3)$$

The genes are ranked according to the absolute value of $|t_j|$, and the top-ranked k genes are selected as significant genes so that the set of these genes is denoted as C_k . The discriminant function obtained by the T-WV method is then constructed as

$$h_k(x | D) = \frac{1}{k} \sum_{j \in C_k} t_j (x_j - \bar{x}_j), \quad (4)$$

where $\bar{x}_j \equiv \frac{1}{N} \sum_j x_{ij}$ is the average expression level of the j -th gene in the training samples.

In the discriminant function h_k , the difference between the j -th gene expression and its average is weighted by its significance, i.e. the t-score. Note that



the function h_k depends on the number k of significant genes, and thus we need to specify k appropriately.

2.3. R-SVM method

R-SVM is another typical supervised machine learning method, which was developed to select important genes for SVM classification.¹⁵ An R code package is publicly available at <http://www.hsph.harvard.edu/bioinfocore/R-SVM.html>. The discriminant function of a linear SVM is defined as

$$h_k(x' | D) = (w \cdot x') + b = \sum_{i=1}^N \alpha_i y_i (x_i \cdot x') + b, \quad (5)$$

where x' is a new input expression vector and x_i is the i -th sample expression vector in the training dataset. α_i and b are parameters to be determined so that training data points with different class labels are classified with the largest margin. $x \cdot x' = \sum_{j=1}^M x_j x'_j$ denotes the inner product. Each element of w , w_j , is defined as

$$w_j = \sum_{i=1}^n \alpha_i y_i x_{ij}, \quad (6)$$

the absolute value $|w_j|$ of which represents the significance weight of the j th gene in the current discriminant function.

As in the T-WV method, the classification performance of SVM also depends on gene subset selection. R-SVM applies a recursive feature elimination (RFE) procedure.²⁷ In RFE, less significant genes in the current discriminant function are recursively eliminated, and the next discriminant function is constructed based on the new, smaller set of genes. Consequently, a sequence of discriminant functions with decreasing numbers of genes is constructed. Thus, the prediction performance of each discriminant function h_k depends on the number k of significant genes, which causes the same problem as in T-WV, i.e. setting an appropriate number k . In the following section, we describe a common way to set the number of genes in both T-WV and R-SVM.

2.4. LOO model selection

T-WV and R-SVM, both produce many candidate classifiers, from which we should select the best one by an assessment process. Although the true performance of a classifier is measured as classification accuracy

on an unknown dataset given in the future, we should instead estimate the performance using the dataset obtained in the assessment process. Note that we refer to each candidate in the assessment process as a *model*, to clarify that we are assessing all procedures used to construct a classifier rather than assessing solely the classifier. In T-WV and R-SVM, a model is characterized by the number of significant genes that it includes.

The LOO procedure has been widely used to estimate, or predict, the future performance of a classifier. In LOO, a classifier h is built using each leave-one-out dataset D^{-i} , $i = 1, \dots, N$; that is, the i -th sample d_i is excluded in the training procedure from the dataset D , and becomes a validation sample. The classification performance of h is assessed using the validation sample. After the assessments for d_1, \dots, d_N , the LOO error rate of the classifier h , $G_{\text{LOO}}(h | D)$, is calculated as the averaged error rate

$$G_{\text{LOO}}(h | D) = \frac{1}{N} \sum_{i=1}^N I(y_i h(x_i | D^{-i}) < 0), \quad (7)$$

where $I(R)$ denotes the indicator function that takes a value of one if condition R holds, and is otherwise zero. When we select the number k of significant genes by

$$h_k^{\text{LOO}} = \underset{k}{\operatorname{argmin}} G_{\text{LOO}}(h_k | D), \quad (8)$$

this model selection is said to be based on the LOO criterion.

2.5. Resampling bootstrap method

It is known that the error rates used to estimate the LOO procedure are nearly unbiased. Molinaro et al¹⁸ compared estimated generalization error rates between different resampling methods and showed that LOO had the smallest bias for a simulation dataset and a real microarray dataset. However, LOO has a tendency to include large variance, despite its small bias,²⁸ because classifiers constructed based on the leave-one-out datasets, D^{-i} , are quite similar to each other, whereas the data points used for validation vary widely. The large variance of the error rate estimation leads to a high risk of selecting a classifier whose 'true' performance is poor, and this risk becomes



higher as the number of candidates becomes larger. When we assess the performance of many candidate classifiers with large variances, some of the candidates often exhibit remarkably low errors, even if their true performance is poor. This is the same problem as overfitting, which was originally found in parametric learning especially when there are many parameters to be learnt. Therefore, it is important to reduce the estimation variance to obtain a robust classifier.

We applied a bootstrap method to simulate possible variation of the given dataset and to obtain the distribution of LOO error rates over the range of that variation. We generated bootstrap datasets $\{D^{*b} \mid b = 1, \dots, B\}$, in which each bootstrap dataset is defined as

$$D^{*b} = \{d_r^{*b} = (x_r^{*b}, y_r^{*b}) \mid r = 1, \dots, N-1\}, \quad (9)$$

where d_r^{*b} is randomly sampled with replacements from the LOO dataset D^{-i} . The single validation sample d_i is evaluated by the classifiers that were trained by different datasets D^{*b} , leading to a set of LOO error rates: $G_{\text{LOO}}(h_k^{*1} \mid D^{*1}), G_{\text{LOO}}(h_k^{*2} \mid D^{*2}), \dots, G_{\text{LOO}}(h_k^{*B} \mid D^{*B})$. h_k^{*b} , $b = 1, \dots, B$, is given by Eq. (4) after replacing the dataset D with the bootstrap dataset D^{*b} . This set of LOO error rates is considered to be a distribution of G_{LOO} and provides a guideline to determine the number of genes used in the T-WV classifier.

2.6. Min-max model selection

Using the simulated distribution of LOO error rates, $\{G_{\text{LOO}}(h_k^{*b} \mid D^{*b})\}_{b=1}^B$, we defined a risk score called a min-max criterion,

$$G_{\text{BOOT}}(h_k \mid D) = \text{Per95}(\{G_{\text{LOO}}(h_k^{*b} \mid D^{*b})\}_{b=1}^B), \quad (10)$$

where ‘Per95’ denotes the 95th percentile of the set of values. Based on this risk score, an appropriate model (i.e. the number of genes, k) is selected as

$$h_k^{\text{BOOT}} = \underset{k}{\text{argmin}}\{G_{\text{BOOT}}(h_k \mid D)\}. \quad (11)$$

We considered the 95th percentile with the number of bootstrap $B = 100$ as the representative of possible high error rates for each model with different numbers

of genes. The 95th percentile is a robust criterion to estimate the risk of selecting a bad model against the possibly asymmetric nature of the error rate distribution.

Our approach is referred to as the ‘‘min-max’’ selection criterion because we minimized the risk of selecting a model for which the expected prediction error rate was almost the maximum in the distribution of possibilities. This min-max model selection is likely to refuse classifiers for which the estimated error rates are distributed with a large variance, even if LOO shows the lowest error rate from a single dataset. Therefore, the min-max criterion reduces the instability stemming from the variation of possible future datasets that could be simulated by random sampling from a large pool of samples.

In other words, the min-max criterion assumes an underlying game between an analyzer and nature. A dataset is given by nature, and a model is selected by an analyzer. For the analyzer to achieve stability, one good idea is to minimize the risk (Eq. (11)), which stems from the possibility that nature could provide a bad situation (and hence the classifier has been over-trained) (Eq. 10).

The number 95 of the percentile and number of bootstrap $B = 100$ were determined arbitrarily by considering trade-offs between computation time, estimation variance of the percentile point, and appropriateness as a representative of high error rates:

- The computation time is proportional to the number of bootstrappings.
- Estimation variance is a monotonic function of both the percentile number and the number of bootstrappings. Namely, the variance becomes large as the percentile number diverges from 50 and as the number of bootstrappings is small.
- The criterion should evaluate possible high error rates even when the distribution of bootstrap samples is asymmetric.

We did not select the 50th percentile, i.e. the median, because of the third reason above; we attempted to obtain a safe classifier rather than to show good average performance. Although the 99th percentile could be another representative of possible high error rates, we rejected it, because it relies on 1% of bootstrap samples, and will therefore lead to high variance especially with small B . The estimation



variance of each percentile of the bootstrap error rate can be evaluated in terms of the standard deviation of the corresponding order statistic if the distribution of error rates is known. Table 1 shows the standard deviations (SDs) of several percentiles when the distribution of error rates is a standard normal distribution. These SDs are proportional to the SD of the distribution of error rates, implying that the SDs of the percentiles can represent their variation well even for non-normal distributions.

3. Results

3.1. Results for real datasets

We evaluated our method using four published real gene expression profile datasets:

- **Breast cancer**
van't Veer et al³ obtained gene expression microarray data for approximately 5,000 genes for 78 + 19 breast cancer tissue samples. The samples were classified into favorable and unfavorable samples: patients with recurrence-free survival in five years and those with recurrence in five years, respectively. The authors trained supervised classifiers using 78 samples (34 favorable and 44 unfavorable samples), which we call the training dataset, and tested using 19 independent samples (7 favorable and 12 unfavorable samples), which we call the test dataset. The same group also provided a larger dataset consisting of 295 samples.²⁹ Among the 295 samples, 32 samples were also included in the former dataset³ and 10 samples were censored in five years; hence, we used the remaining 253 (192 favorable and 61 unfavorable) samples for the second test dataset.
- **Colon cancer**
The colon cancer dataset³⁰ contains microarray expression data for 2,000 genes for 62 colon tissues.

Table 1. Estimated standard deviations of bootstrap percentiles. Bold type marks the setting which we used in the current study.

	B = 100	B = 500	B = 1000
99th	0.315	0.171	0.120
95th	0.216	0.095	0.067
90th	0.172	0.077	0.054
50th	0.125	0.056	0.040

Among the 62 tissue samples, 40 and 22 were labeled as “tumor” and “normal,” respectively, and these were used as the labels to be predicted.

- **Neuroblastoma (NBL)**
The NBL dataset⁵ consists of microarray expression data for 5,180 genes for 136 patients. Among the 136 samples, 25 and 102 were labeled as “favorable” and “unfavorable” patients, respectively, according to their status at 24 months after diagnosis, and these were used as the labels to be predicted. The remaining nine samples of unknown status at 24 months after diagnosis were omitted.
- **Breast cancer Affymetrix (Affymetrix)**
Wang et al³¹ analyzed 286 breast cancer patients with an Affymetrix chip harboring 22,283 genes. Among the 286 patients, 183 and 93 were labeled as favorable and unfavorable, respectively, and these were used as the labels to be predicted. We omitted 10 samples which were censored in five years. Although this dataset concerned breast cancer, we did not consider relationship between this set and the breast cancer datasets at the top of this list because these two datasets were assembled by entirely different systems and hence had fairly different characters in distribution. Considering different systems of microarrays together may be an important issue, but is beyond the scope of the current study.

For each of the above four datasets, we trained T-WV and R-SVM classifiers with various numbers of genes using the training samples, and assessed their classification errors in terms of LOO, 3-, 5- and 10-fold-CV, and min-max criteria. In the case of the breast cancer dataset with large numbers of test samples,^{3,29} we also assessed their classification errors in the test datasets.

Figure 1 shows the results for the breast cancer dataset. The results with the T-WV classifier (left panel), indicated characteristic behaviors of the three criteria to assess the classification error rate, LOO (dashed line), 3-fold-CV (dotted line), and the proposed min-max criterion (solid line at the top of the blue area). The 90% interval of LOO error rates (blue area), which was estimated by the resampling bootstrap method, describes the estimation variance of error rates. The LOO error rate profile showed the

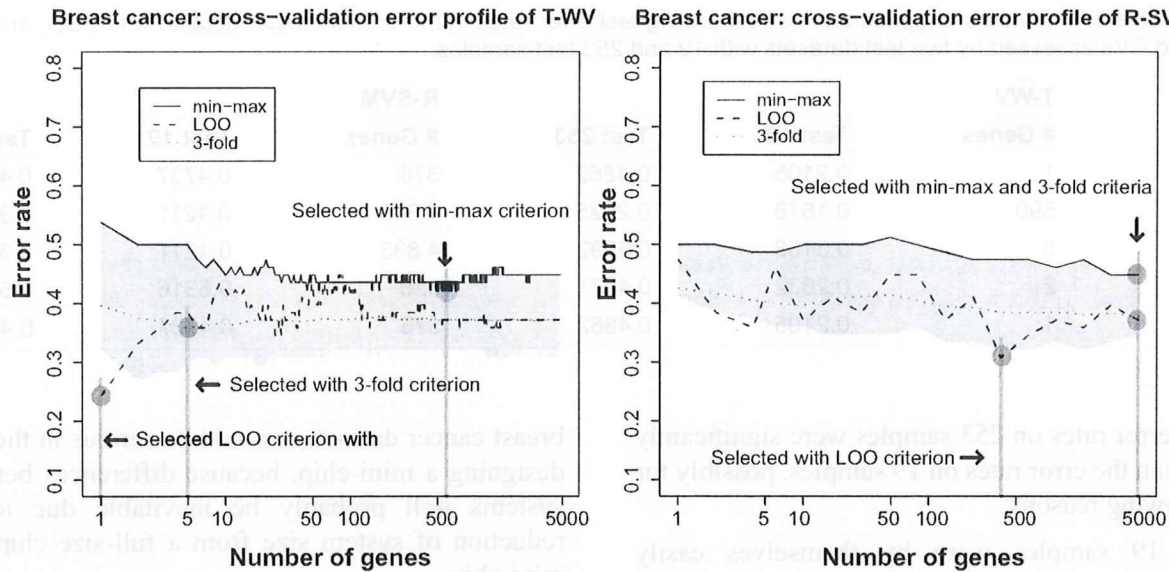


Figure 1. Estimated classification errors in the breast cancer dataset. The left and right panels show the results obtained with the T-WV and R-SVM methods, respectively. The vertical and horizontal axes denote classification error rates estimated by various criteria and the number of genes included in each classifier, respectively. The 90% interval of resampling bootstrap of the estimated classification errors at each number of genes is denoted by blue areas. The classification errors estimated by the three criteria, min-max criterion (solid line on the top of blue area), LOO error rate (dashed line), and 3-fold-CV error rate (dotted line), are plotted against different numbers of genes. Vertical lines indicate the numbers of genes selected by the three criteria.

lowest value with a small number of genes, $k = 1$, so that $k = 1$ was selected as the best number of genes by the LOO criterion. On the other hand, the 90% interval of the bootstrap distribution at $k = 1$ exhibited a large width in the error rate, and the 95th percentile error rate was above the chance level 0.5, suggesting large risk of the $k = 1$ classifier falling into a poor predictor around the chance level. Also, the LOO error rate at $k = 1$ was below both the 5th percentile and the 3-fold-CV error rate, indicating that the low LOO error rate at $k = 1$ could have been obtained by chance. The 3-fold-CV showed a smoother profile than those obtained by the LOO, and stayed in the midst of the 90% interval. The 3-fold-CV criterion selected a classifier with $k = 5$ where the 90% interval was narrower than that at $k = 1$. We also calculated 5- and 10-fold-CVs and observed similar curves to that of the 3-fold-CV. The proposed min-max criterion, i.e. the 95th percentile, selected a larger number of genes, $k = 590$. The LOO and 3-fold-CV error rates at $k = 590$ were higher than those at $k = 1$ and $k = 5$; however, we expected that the classifier of $k = 590$ would have a lower risk of being a poor predictor than those at $k = 1$ and $k = 5$.

In the right panel of Figure 1, a similar comparison is shown between LOO, 3-fold-CV, and the min-max

criteria with the R-SVM classifier. The LOO criterion showed an instability similar to that of T-WV, so that the lowest LOO error rate at $k = 376$ seems to have been obtained by chance. All criteria selected larger numbers of genes than in the cases of T-WV classifiers.

In Table 2, test error rates of the selected predictors were assessed using two test datasets with 19 and 253 samples, where five criteria (LOO, min-max, and 3-, 5- and 10-fold-CVs) with two classifiers (T-WV and R-SVM) are compared. The min-max criterion outperformed the other criteria, LOO and k-fold-CVs, on both test sets. The LOO exhibited poor performance with 19 test samples and worse with 253 test samples whose test error rate was around the chance level. Intuitively, this result pointed out a defect of the LOO criterion in terms of the risk of taking a poor classifier, which has already been suggested by the 90% interval shown in Figure 1. The 3-, 5- and 10-fold-CVs achieved better performance in test error rates than LOO, but worse than the min-max criterion. T-WV tended to exhibit lower error rates than R-SVM with smaller numbers of genes, although we cannot conclude the general superiority of T-WV based on this single example.



Table 2. Selected numbers of genes and corresponding test error rates in the breast cancer dataset with LOO, min-max, and k-fold CVs assessed by two test datasets with 19 and 253 test samples.

	T-WV			R-SVM		
	# Genes	Test 19	Test 253	# Genes	Test 19	Test 253
LOO	1	0.2105	0.4862	376	0.4737	0.4664
min-max	590	0.1578	0.2925	4,833	0.4211	0.3992
3-fold	5	0.3158	0.3992	4,833	0.4211	0.3992
5-fold	2	0.2632	0.4071	626	0.6316	0.5217
10-fold	1	0.2105	0.4862	376	0.4737	0.4664

Test error rates on 253 samples were significantly worse than the error rates on 19 samples, possibly for the following reasons:

- The 19 samples were by themselves easily classified.
- The number of samples (19) was too small to reproduce the error rate with low variance.
- The test data of 253 samples were gathered from different populations from those for the training data of 78 samples and the other test data of 19 samples.
- The microarray measurement system differed between the two sets of data.

The considerations above will be important when designing mini-chips based on training datasets. Although the last reason, difference in microarray systems, may not be very serious in the case of this

breast cancer dataset, it would be serious in the case designing a mini-chip, because differences between systems will probably be inevitable due to the reduction of system size from a full-size chip to a mini-chip.

We compared three criteria, LOO, min-max, and 3-fold-CV, with the two classifiers T-WV and R-SVM on the other three datasets (NBL, colon cancer and breast cancer Affymetrix) in Figures 2, 3 and 4, respectively. From the total comparisons over Figures 1–4, we observed the following tendencies:

- Although the error rates estimated by LOO fluctuate as the number of genes increases, they stay mostly within the 90% interval. This suggests that the LOO estimation of the tuned number of genes includes a large variance and the character

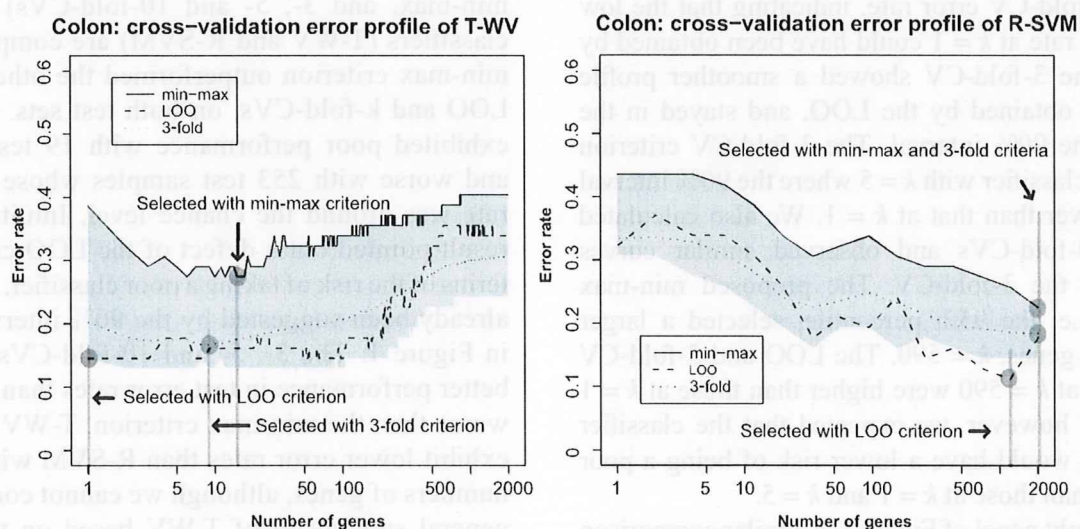


Figure 2. Estimated classification errors in the colon cancer dataset. See Figure 1 legend for details.

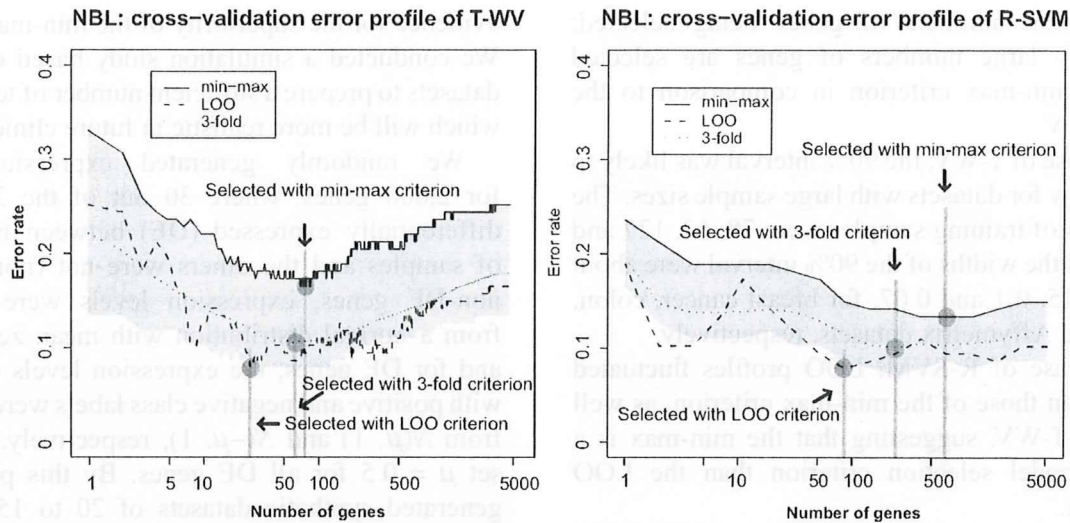


Figure 3. Estimated classification errors in the NBL dataset. See Figure 1 legend for details.

of the variance is well captured by the estimated 90% interval.

- In contrast to the fluctuating profile of LOO error rates, the profiles of the 3-fold-CV and the 95th percentile (G_{boot}) exhibit smoother curves. This suggests a more stable character for the 3-fold-CV and the min-max criterion than the LOO criterion.
- With T-WV, the 90% confidence interval was likely to be wide when the number of genes was small, $k < 10$, indicating that prediction based on too few genes is risky; we occasionally get a model with poor performance. The 95th percentile

is likely to show a higher error rate for a smaller number of genes, e.g. $k < 10$, than for a large number of genes. Thus, the min-max criterion based on the 95th percentile can avoid risky prediction so that a smaller error rate is achieved on average.

- The 3-fold-CV profile stayed almost in the middle of the 90% interval and showed a similar curve to the 95th percentile. However, there was difference between the 3-fold-CV and the 95th percentile in the range of 90% interval, which was prominent in T-WV with small numbers of genes, $k < 10$. The 3-fold-CV and the min-max criterion lead

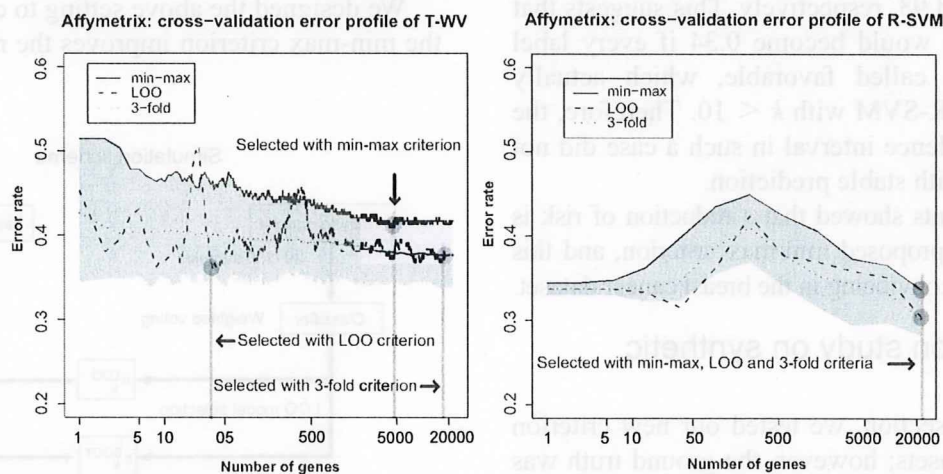


Figure 4. Estimated classification errors in the Affymetrix dataset. See Figure 1 legend for details. Note that the errors of R-SVM with $k < 10$ are not reliable because of unbalanced numbers of labeled samples.

to different numbers of genes being selected; relatively large numbers of genes are selected by the min-max criterion in comparison to the 3-fold-CV.

- In the case of T-WV, the 90% interval was likely to be narrow for datasets with large sample sizes. The numbers of training samples were 78, 62, 127 and 276, and the widths of the 90% interval were about 0.15, 0.15, 0.1 and 0.07, for breast cancer, colon, NBL and Affymetrix datasets, respectively.
- In the case of R-SVM, LOO profiles fluctuated more than those of the min-max criterion, as well as with T-WV, suggesting that the min-max is a better model selection criterion than the LOO criterion.
- Whereas the best performance was comparable between R-SVM and T-WV, a larger number of genes was required to achieve the best performance by R-SVM than by T-WV. Thus, T-WV employing a relatively small number of genes is more suitable for practical clinical applications, which is consistent with a previous finding.¹²
- The confidence intervals for R-SVM were likely to be narrower than those for T-WV, implying that SVM, as a large margin classifier, is more stable against observation noise than T-WV. Even though we are not interested here in classifiers with a large number of genes, say $k > 1,000$, this finding may be important for applications other than mini-chip construction.
- The Affymetrix data set was unbalanced, with the numbers of favorable and unfavorable samples being 183 and 93, respectively. This suggests that the error rate would become 0.34 if every label prediction is called favorable, which actually occurred for R-SVM with $k < 10$. Therefore, the narrow confidence interval in such a case did not correspond with stable prediction.

The experiments showed that a reduction of risk is achieved by the proposed min-max criterion, and this was particularly convincing in the breast cancer dataset.

3.2. Simulation study on synthetic datasets

In the previous section, we tested our new criterion on four real datasets; however, the ground truth was unknown and the number of samples was limited in many cases, which prevented us from obtaining strong

evidence for the superiority of the min-max criterion. We conducted a simulation study based on artificial datasets to prepare a sufficient number of test samples, which will be more realistic in future clinical studies.

We randomly generated expression profiles for 2,000 genes, where 30 out of the 2,000 were differentially expressed (DE) between two classes of samples and the others were not (non-DE). For non-DE genes, expression levels were generated from a normal distribution with mean zero, $N(0,1)$, and for DE genes, the expression levels of samples with positive and negative class labels were generated from $N(\mu, 1)$ and $N(-\mu, 1)$, respectively, where we set $\mu = 0.5$ for all DE genes. By this process, we generated synthetic datasets of 20 to 150 samples for training, and 1,000 samples for testing, where the numbers of samples with the two class labels were set to be equal.

The proposed simulation scheme is illustrated in Figure 5. For each training dataset, the candidate classifiers involving various numbers of genes were trained and assessed, and the best numbers of genes were selected by the LOO and the min-max criteria, where the number B of the bootstrap in the min-max procedure was set at 100. The performance of the finally selected classifier was then assessed by a test dataset with 1,000 samples. We repeated this process with a randomly generated training dataset and assessed the corresponding test error rates by using a test dataset of 1,000 samples. The distributions of the test error rates were compared between different conditions.

We designed the above setting to clarify how well the min-max criterion improves the model selection.

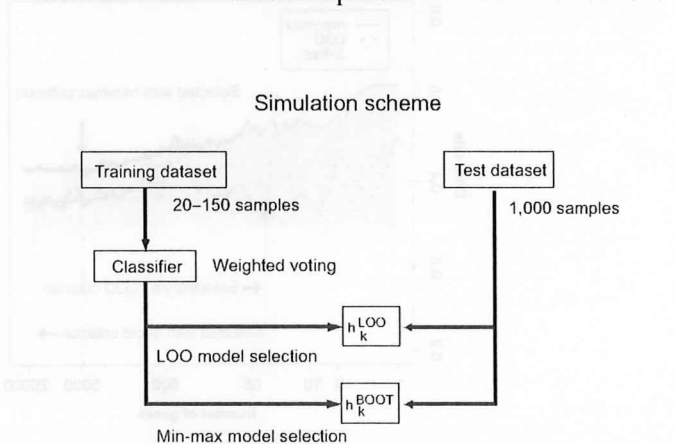


Figure 5. Setting of the simulation experiment.



The number of test datasets was set sufficiently large, and is commonly used in various settings of the other features to reduce the variance of error rates that stems from random sampling of the test dataset. The number of DE genes (30) and the strength of differential expression ($\mu = 0.5$) were determined to examine typical situations that arise in realistic cases. We omitted other realistic features of datasets that may arise such as variation in the number of DE genes, strength μ , and the proportion of numbers of positive and negative samples, because they had shown no significant effect in our preliminary experiments. We also omitted correlations of gene expression patterns

between DE genes because such correlations would not affect either T-WV or R-SVM.

Figure 6 shows the distributions of test error rates of the T-WV classifiers selected by LOO and min-max, with 20, 50, 100 and 150 training samples. We found that there were certain levels of variance for both criteria, and the variance was larger for smaller numbers of samples. LOO sometimes showed much worse results than min-max, as indicated by the points in the bottom-right area of each panel in Figure 6. Note that the number of test samples, 1,000, was so large that there was no significant increase in sampling variance. Table 3 shows the means and

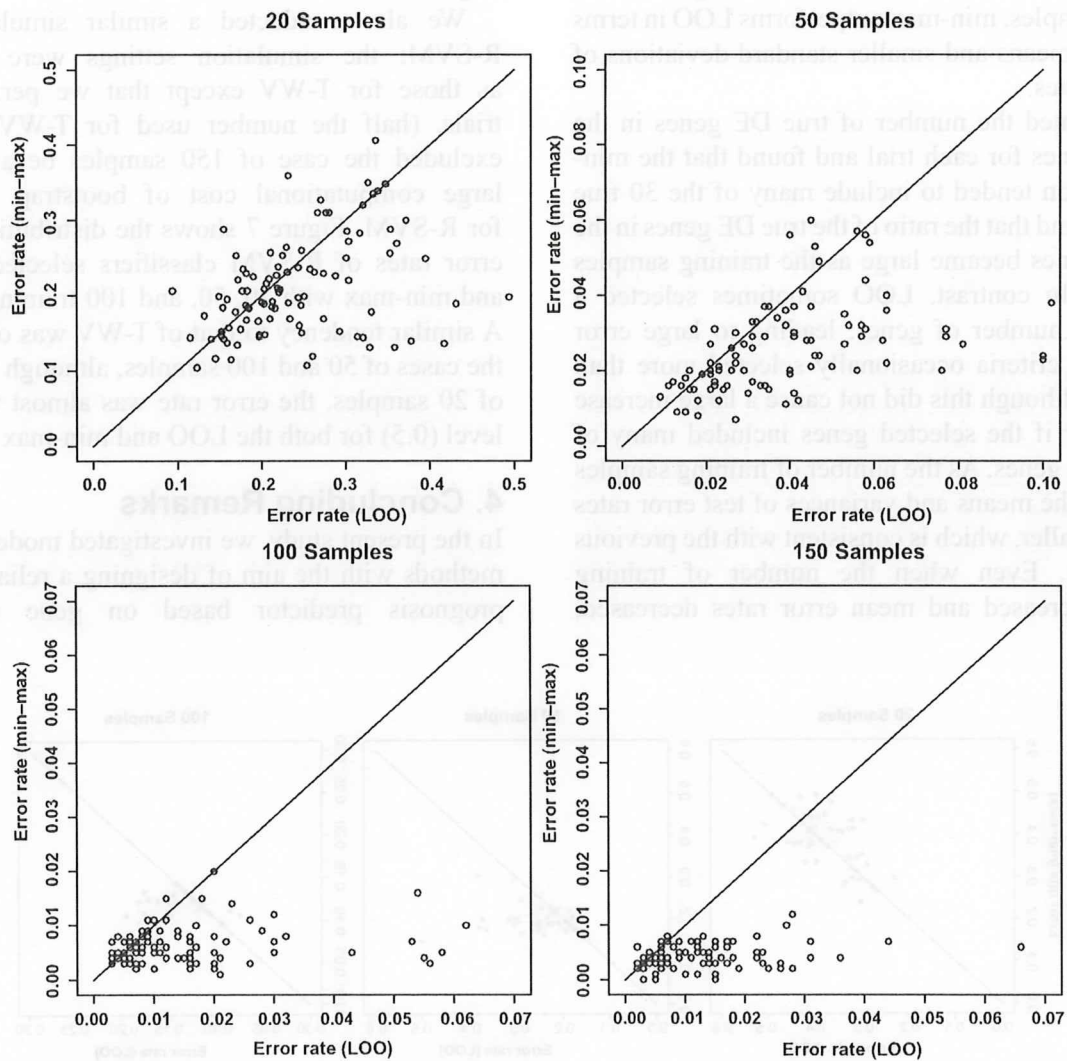


Figure 6. Distribution of test error rates of T-WV. The vertical and horizontal axes denote the test error rates of classifiers selected by the min-max and LOO criteria, respectively. The results from 100 trials of random sampling of 20, 50, 100 and 150 samples are shown in the four panels.

Table 3. Test error rate of simulation dataset.

Number of training samples	Selection criterion	Mean	Standard deviation
20	LOO	0.241	0.077
	min-max	0.210	0.064
50	LOO	0.042	0.024
	min-max	0.026	0.012
100	LOO	0.015	0.013
	min-max	0.006	0.003
150	LOO	0.012	0.010
	min-max	0.004	0.002

standard deviations of test error rates of the classifiers selected by LOO and min-max. Through 20–150 training samples, min-max outperforms LOO in terms of smaller means and smaller standard deviations of test error rates.

We counted the number of true DE genes in the selected genes for each trial and found that the min-max criterion tended to include many of the 30 true DE genes, and that the ratio of the true DE genes in the selected genes became large as the training samples increased. In contrast, LOO sometimes selected a very small number of genes, leading to large error rates. Both criteria occasionally selected more than 30 genes, although this did not cause a large increase in the error if the selected genes included many of the true DE genes. As the number of training samples increased, the means and variances of test error rates became smaller, which is consistent with the previous observation. Even when the number of training samples increased and mean error rates decreased,

however, the test error rates of LOO still showed larger variance than those of min-max.

We also conducted a similar simulation with R-SVM; the simulation settings were the same as those for T-WV except that we performed 50 trials, (half the number used for T-WV), and we excluded the case of 150 samples because of the large computational cost of bootstrap simulation for R-SVM. Figure 7 shows the distributions of test error rates of R-SVM classifiers selected by LOO and min-max with 20, 50, and 100 training samples. A similar tendency to that of T-WV was observed in the cases of 50 and 100 samples, although in the case of 20 samples, the error rate was almost the chance level (0.5) for both the LOO and min-max criteria.

4. Concluding Remarks

In the present study, we investigated model selection methods with the aim of designing a reliable cancer prognosis predictor based on gene expression

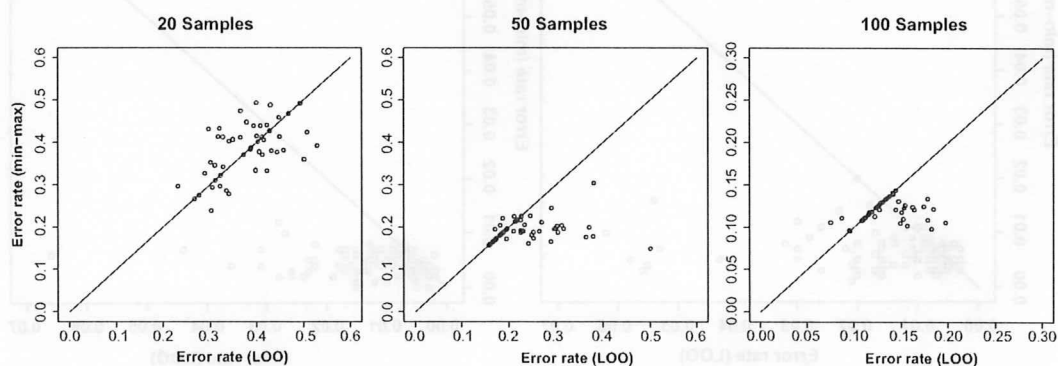


Figure 7. Distribution of test error rates of R-SVM. The vertical and horizontal axes denote the test error rates of classifiers selected by the min-max and LOO criteria, respectively. The results from 50 trials of random sampling of 20, 50 and 100 samples are shown in the three panels.



microarrays involving as small a number of genes as possible. We assessed possible variation in prediction error rate of each microarray-based predictor by simulating a distribution of classification error rates via a resampling bootstrap method. Accordingly, we proposed a novel min-max criterion to select a predictor from multiple candidates. In numerical comparisons that used real and synthetic datasets, we showed that the conventional LOO estimation of their error rates resulted in large variances; consequently, the LOO criterion had a large risk of choosing inappropriate classifiers that would exhibit extremely poor prediction performance. In contrast, we showed the stability of the min-max criterion relative to well-established statistical criteria including the LOO. We also compared two different supervised analysis procedures, T-WV and R-SVM, and found that, in general, T-WV performed the best when it involved a small or moderate number of genes in contrast to that R-SVM performed the best when it involved almost all genes, although the mean and variance of the best possible performances were not always significantly different between those achieved with T-WV and R-SVM. Thus, overall, we concluded to recommend T-WV with the min-max criterion, which satisfied our demand; the most reliable predictor involving as small a number of genes.

It should be important to note that, we proposed our procedure to select a set of genes for designing a good predictor of cancer prognosis, rather than for determining a set of genes which have statistically significant relationship to the prognosis; these purposes are different from each other in general. In other words, the 'robust' model selection is meant to lower the risk to select an extremely poor predictor, rather than to select a stable set of genes. In fact, different research groups reported prognosis prediction systems with different sets of genes based on different sets of microarray data for the same type of cancer.⁶ The microarray-based predictors for breast cancer, were designed with 70 and 76 genes by two different research groups,^{3,31} respectively, and these gene sets had only three genes in common. Namely, the selected sets of genes were not stable at all, however, the 70 gene-based diagnosis system of breast cancer have been verified by increasingly large number of new patients and authorized by Food and Drug Administration in USA.⁶ In our own numerical experiments, we also

observed that number of common genes tended to be small between any gene sets that were selected based on different datasets generated by resampling bootstrap (data not shown), although we achieved good predictors in vast amount of the cases as we had shown. Thus, it should be emphasized that such an instable selection of gene subsets did not necessarily cause a poor predictor as long as the predictor was selected by a robust model selection method.

Once a prediction system based on a small number of genes is developed, the system can be transferred not only to mini-chip microarrays but also to other easy accessible devices such as quantitative real-time polymerase-chain-reaction (RT-PCR) analysis,³² which would be tractable if only tens of genes were targeted. Robust model selection methods, like the proposed one, will be needed especially when we consider such a transfer work between different measurement devices because large bias is often expected between different devices. In general, when a procedure is designed to be robust against measurement variance, such a method is also robust against an unknown bias which would appear like in the above transfer; thus, our min-max criterion will be used for this purpose.

In order to design a practical tool for real scenes in clinical cancer therapy, new demands in informatics can always arise. As we had seen in this study, although past efforts in informatics tended to pursue good performances in average, minimizing risk to catch poor predictor against possible variability in cancer diagnosis systems becomes a next issue. There are few methods to directly seek such risk minimization as long as we know. Reducing cost by selecting relevant genes based on high-dimensional gene expression profile is a relatively well-investigated field of research. However, the combination of the cost and reliability is not investigated well. Thus, there must be room to develop a novel supervised classification method that satisfies these demands for designing mini-chip systems, and future studies in cancer informatics should proceed to such directions.

Authors Contributions

IS performed the experiments and wrote the manuscript. IS and SO proposed the main idea that the variance influenced the performance of the classifiers. TT contributed to the construction of the simulation scheme and the development of the



variance estimation methods. SI provided advice on the min-max strategy and supervised the present study. MO provided several topics concerning real cancer therapy and future directions. All five authors participated in the preparation of the final manuscript.

Acknowledgement

We are grateful to Dr. Ian Smith for his careful proofreading of our manuscript. Thanks are also due to anonymous reviewers for their critical reviews and useful suggestions.

Disclosure

The authors report no conflicts of interest.

References

- Brown PO, Botstein D. Exploring the new world of the genome with DNA microarrays. *Nature Genetics*. 1999;21:33–7.
- Alizadeh AA, Eisen MB, Davis EE, et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*. 2000;403:503–11.
- van't Veer L, Dai H, van de Vijver MJ, et al. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*. 2002;415:530–6.
- Iizuka N, Oka M, Yamada-Okabe H, et al. Oligonucleotide microarray for prediction of early intrahepatic recurrence of hepatocellular carcinoma after curative resection. *Lancet*. 2003;361:923–9.
- Ohira M, Oba S, Nakamura Y, et al. Expression profiling using a tumor-specific cDNA microarray predicts the prognosis of intermediate risk neuroblastomas. *Cancer Cell*. 2005;7:337–50.
- van't Veer L, Bernards R. Enabling personalized cancer medicine through analysis of gene-expression patterns. *Nature*. 2008;415:564–70.
- Draghici S, Khatri P, Eklund AC, et al. Reliability and reproducibility issues in DNA microarray measurements. *Trends in Genetics*. 2006;22:101–9.
- Glas AM, Floore A, Delahaye LJM, et al. Converting a breast cancer microarray signature into a high-throughput diagnostic test. *BMC Genomics*. 2006;7:278.
- Jaeger J, Spang R. Selecting normalization genes for small diagnostic microarrays. *BMC Bioinformatics*. 2006;7:388.
- Dudoit S, Fridlyand J, Speed T. Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association*. 2002;97:77–87.
- Li T, Zhang C, Ogihara MA comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression. *Bioinformatics*. 2004;20:2429–37.
- Ramaswamy S, Tamayo P, Rifkin R, et al. Multiclass cancer diagnosis using tumor gene expression signatures. *Proceedings of the National Academy of Sciences of the United States of America*. 2001;98:15149–54.
- Golub T, Slonim D, Tamayo P, et al. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*. 1999;286:531–7.
- Vapnik VN. *The nature of statistical learning theory*. Springer, New York, 2000.
- Zhang X, Lu X, Shi Q, et al. Recursive SVM feature selection and sample classification for mass-spectrometry and microarray data. *BMC Bioinformatics*. 2006;7:197.
- Friedman JH. On bias, variance, 0/1-loss, and the curse-of-dimensionality. *Data Mining Knowledge Discovery*. 1997;1:55–77.
- Borgne YL. Bias variance trade-off characterization in a classification. what differences with regression? Technical report, ULB, 2005.
- Molinaro AM, Simon R, Pfeiffer RM. Prediction error estimation: a comparison of resampling methods. *Bioinformatics*. 2005;21:3301–7.
- Varma S, Simon R. Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*. 2006;7:91.
- Wood IA, Visscher PM, Mengersen KL. Classification based upon gene expression data: bias and precision of error rates. *Bioinformatics*. 2007;23:1363–70.
- Braga-Neto UM, Dougherty E. Is cross-validation valid for small-sample microarray classification? *Bioinformatics*. 2004;20:374–80.
- Fu WJ, Carroll RJ, Wang S. Estimating misclassification error with small samples via bootstrap cross-validation. *Bioinformatics*. 2005;21:1979–86.
- Xu Q, Hua J, Braga-Neto U, et al. Confidence intervals for the true classification error conditioned on the estimated error. *Technology in Cancer Research and Treatment*. 2006;5:579–89.
- Jiang W, Varma S, Simon R. Calculating confidence intervals for prediction error in microarray classification using resampling. *Statistical Applications in Genetics and Molecular Biology*. 2008;7.
- Davis CA, Gerick F, Hintermair V, et al. Reliable gene signatures for microarray classification: assessment of stability and performance. *Bioinformatics*. 2006;22:2356–63.
- Haibe-Kains B, Desmedt C, Loi S, et al. *Computational Intelligence in Clinical Oncology: Lessons Learned from an Analysis of a Clinical Study*, volume 122. Springer-Verlag Berlin/Heidelberg, 2008.
- Guyon I, Weston J, Barnhill S, et al. Gene selection for cancer classification using support vector machines. *Machine Learning*. 2002;46:389–422.
- Hastie T, Tibshirani R, Friedman JH. *The Elements of Statistical Learning*. Springer, New York, 2001.
- van de Vijver MJ, He YD, van't Veer L, et al. A gene-expression signature as a predictor of survival in breast cancer. *The New England Journal of Medicine*. 2002;347:1999–2009.
- Alon U, Barkai N, Notterman DA, et al. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences of the United States of America*. 1999;96:6745–50.
- Wang Y, Kljij J, Zhang Y, et al. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet*. 2005;365:671–9.
- Lossos IS, Czerwinski DK, Alizadeh AA, et al. Prediction of survival in diffuse large-B-cell lymphoma based on the expression of six genes. *New England Journal of Medicine*. 2004;350:1828–37.



Supplementary Figures for “Robust Model Selection for Classification of Microarrays”

Ikumi Suzuki, Takashi Takenouchi, Miki Ohira, Shigeyuki Oba, and Shin Ishii

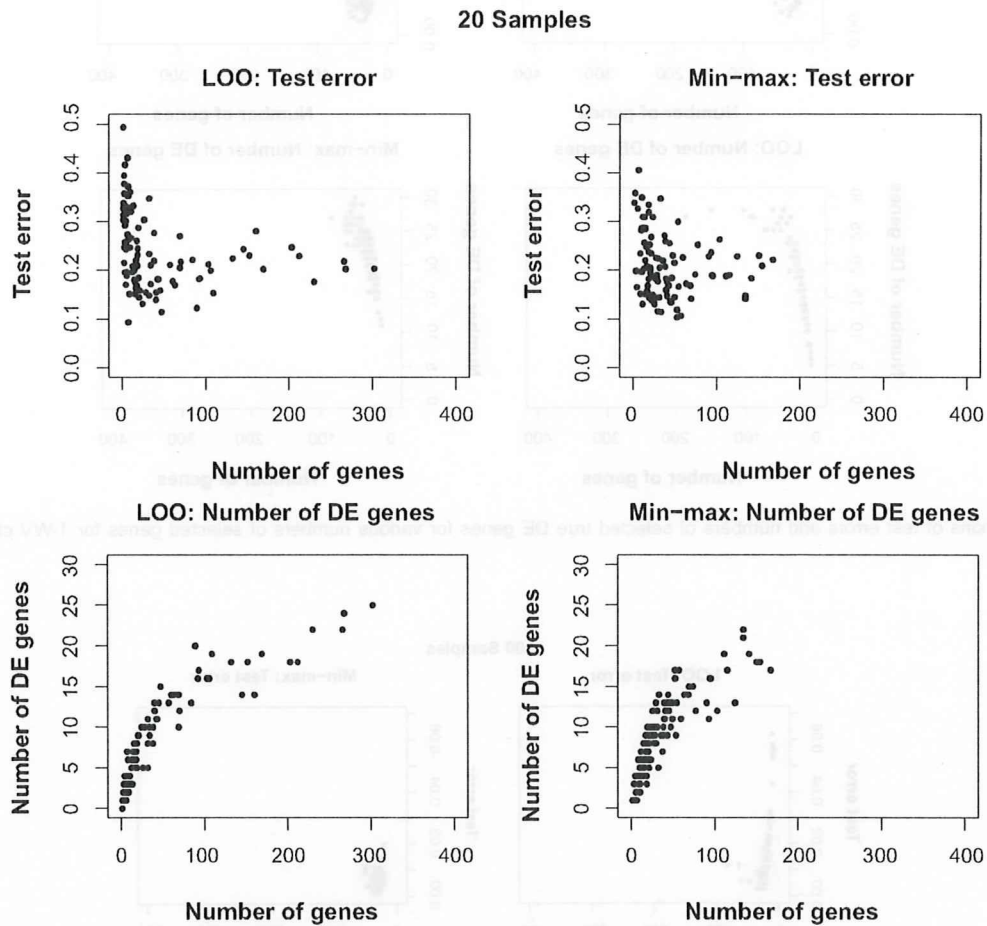


Figure S1. Distributions of test errors and numbers of selected true DE genes for various numbers of selected genes for T-WV classifiers based on 20 artificial samples. Each point denotes one of 100 trials in each setting. Horizontal axes denote the number of genes selected by either LOO or min-max criterion. The vertical axes in the top two panels and the bottom two panels denote the test error estimated by 1000 test samples and the number of true DE genes in the selected set of genes, respectively.