



Application of Bioinformatics in Cancer Epigenetics

HOWARD H. YANG AND MAXWELL P. LEE

*Laboratory of Population Genetics, National Cancer Institute,
Bethesda, Maryland 20892, USA*

ABSTRACT: With the completion of the human genome sequence and the advent of high-throughput genomics-based technologies, it is now possible to study the entire human genome and epigenome. The challenge in the next decade of biomedical research is to functionally annotate the genome, epigenome, transcriptome, and proteome. High-throughput genome technology has already produced massive amounts of data including genome sequences, single nucleotide polymorphisms, and microarray gene expression. Our ability to manage and analyze data needs to match the speed of data acquisition. We will summarize our studies of allele-specific gene expression using genomic and computational approaches and identification of sequence motifs that are signature of imprinted genes. We will also discuss about how bioinformatics can facilitate epigenetic researches.

KEYWORDS: genomics; bioinformatics; cancer; epigenetics; genomic imprinting

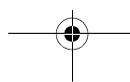
INTRODUCTION

Genetic variation in humans is largely caused by DNA polymorphism and differences in gene expression. A biological role has been identified for differential allelic expression associated with X-inactivation and genomic imprinting. Mendelian inheritance assumes that genes from maternal and paternal chromosomes contribute equally to human development. X chromosome inactivation silences gene expression from one of the two X chromosomes, thus providing an exception to Mendelian inheritance.¹ In addition, approximately 50 human autosomal genes are known to be imprinted and thus are expressed from only one chromosome.² However, it is unknown whether variations in allelic gene expression affect only the X chromosome and imprinted genes or whether they affect human genes generally. Recently, a group from Johns Hopkins University reported that 6 out of 13 genes show significant difference in gene expression between the two alleles and that this variation in allelic gene expression was transmitted by Mendelian inheritance.³ They had previously shown that the allelic variation in the APC gene expression plays a critical role in colon cancer.⁴ It will be interesting to know if genetic variations, especially

Address for correspondence: Maxwell P. Lee, Laboratory of Population Genetics, National Cancer Institute, 41 Library Drive D702C, Bethesda, MD 20892. Voice: 301-435-1536; fax: 301-402-9325.

leemax@mail.nih.gov

Ann. N.Y. Acad. Sci. 1020: 1–10 (2004). © 2004 New York Academy of Sciences.
doi: 10.1196/annals.1310.008



regulatory single nucleotide polymorphisms (SNPs), contribute to common diseases including cancer.

Genomic imprinting is an unusual mechanism of gene regulation that results in preferential expression of one specific parental allele of a gene. Abnormal imprinting can cause human diseases such as Beckwith-Wiedemann syndrome, Prader-Willi syndrome, or Angelman's syndrome.⁵⁻⁷ Loss of imprinting (LOI) is often associated with human cancers.^{8,9} Although the exact mechanism of genomic imprinting is still largely unknown, differentially methylated CpG islands, imprinted antisense transcripts, and insulators may play important roles in the regulation of imprinting.¹⁰⁻¹² Most of the imprinted genes are located in the imprinting domains.¹³ However, some genes in the imprinting domain can escape imprinting regulation.¹⁴ Many imprinted genes are scattered throughout the human genome. Therefore, it is likely that local *cis*-elements as well as chromatin structure control genomic imprinting. Since patterns of gene regulation and the corresponding regulatory elements are often conserved across species, sequence comparison between human and mouse is a powerful approach to identify regulatory sequences.¹⁵ Such comparative sequence analysis has already identified a number of conserved sequences and novel imprinted genes in human 11p15¹⁶ and Dlk1-Gtl2 loci.^{17,18}

The current release of human Unigene (build #162) contains 4,472,210 EST clones that are in 123,995 clusters: 16,069 of these Unigene clusters contain at least 33 EST clones. Genes with multiple ESTs can be used to deduce information about digital gene expression.¹⁹ Computational methods have been used to identify SNPs in redundant EST clones.²⁰⁻²² We have also used EST database to mine allele-specific gene expression.²³

GENOME-WIDE ANALYSIS OF ALLELE-SPECIFIC GENE EXPRESSION

A biological role has been identified for differential allelic expression associated with X-inactivation and genomic imprinting; however, a large-scale analysis of differential allelic expression of human genes has not been carried out. The HuSNP chip was designed for simultaneous typing of 1494 SNPs of the human genome. We adapted the HuSNP chip system to study allele-specific gene expression.²⁴

Affymetrix only provided software for genotyping using the HuSNP chip. We decided to develop the following computational method to quantify allele-specific gene expression. We extracted the intensity values for each probe from the .CEL files generated by Affymetrix MAS 4.0. The .CEL files contain the fluorescent intensity values for each of the probes. The HuSNP chip contains 16 probes for each SNP locus. Four of the 16 probes match perfectly to allele A, 4 to allele B, 4 have 1 mismatch to allele A, and the other 4 have 1 mismatch to allele B. Allele A and allele B represent the two alleles of the SNP. Each probe contains 20 nucleotides. The centers of the nucleotide probes are located at positions -4, -1, 0, and 1 relative to the SNP. The 4 mismatch probes are identical to the perfect match probes, except for 1 mismatched base, which is always located in the center of the probe. The value for each probe pair was computed by subtracting the mismatch intensity from the perfect match intensity. A *t* test was used to calculate a *P* value for the presence of signal (intensity greater than 0) for each allele of each SNP. We considered a signal

to be present if at least one allele had signal ($P < .01$, t test). Affymetrix defines a miniblock as a group of 4 probes that include a perfect match probe for allele A (PMA), a mismatch probe for allele A (MMA), a perfect match probe for allele B (PMB), and a mismatch probe for allele B (MMB). We set $(PMA - MMA) = 50$ if $(PMA - MMA)$ is less than 50 for each miniblock. Similarly, baseline for allele B was set at 50. An allele A fraction, defined as $f = (PMA - MMA)/(PMA - MMA + PMB - MMB)$, was computed for each miniblock, and the mean of the allele A fraction f from miniblocks was computed for each SNP. The gene expression difference between the two alleles from a heterozygous individual can be quantified using the ratio of allele A/allele B, computed from $f/(1 - f)$. For each chip, we have intensities from two scans called scan A and scan B. Generally, we used the intensity values from scan A. We used the intensity values from scan B if the t test showed that both alleles have no signal in scan A, while at least one of the alleles from scan B had signal. The ratio was further normalized by the ratio of genomic DNA for the SNP. We analyzed a set of HuSNP chip data from 7 individuals and found that 39 SNPs were heterozygous in at least 5 individuals. We computed the 95% confidence interval for the allelic ratio of genomic DNA for each of these 39 SNPs, and the average confidence interval was between 0.5 and 2.0. This value was used to select those genes that show significant difference in the expression between the two alleles.

In order to measure allele-specific gene expression quantitatively, we first needed to find out (1) which of the SNPs on the chip are located in transcribed regions and (2) whether the system can measure allele-specific expression accurately. Using blast searches and annotations in dbSNP, we found that 1063 of the SNPs are located in transcribed regions. To address the second issue, we used our computational method to extract the fluorescent intensity for each probe from an Affymetrix output file and quantify the ratio of expression of the two alleles. To assess the precision of the system, we performed experiments in duplicate for both genomic DNA and for polyA RNA from 3 fetuses. We found that the correlation between the repeated experiments was very high, with average Pearson correlation coefficients of 0.98 ($P < .001$) for genomic DNA and 0.95 ($P < .001$) for RNA. We then performed genotyping and allele-specific gene expression in kidney and liver from 7 fetuses. Genotype calls were obtained using the Affymetrix MAS 4.0 software, and quantitative allele-specific gene expression was computed using the method that we have developed. To be included in our analysis, each SNP had to meet the following criteria: (1) at least one fetus is heterozygous for the SNP; (2) the SNP is among the 1063 mapped within a transcribed region; and (3) the gene containing the SNP is expressed in kidney or liver. We found that 603 SNPs met all three criteria and 326 (54%) of which showed preferential expression of one allele; for 170 genes, there was at least a 4-fold difference in expression between the two alleles in at least one sample. Some of these 170 genes are imprinted (i.e., *SNPRN*, *IPW*, *HTR2A*, and *PEG3*). The genomic locations of all SNPs on the Affymetrix HuSNP chip were identified. Some of the genes showing differential allelic expression are clustered in the same genomic region, and some are in imprinted domains. *HTR2A*, *LOC51131*, and *FLJ13639* are located at 13q14, and all three show mono-allelic expression. *SNPRN*, *IPW*, and *LOC145622* are in the imprinted domain at 15q12, and all three genes preferentially express one allele. However, the majority of the genes that show preferential expression of one allele are scattered on different chromosomes, indicating that allelic variation is very common throughout the human genome. Our

studies demonstrate that allele-specific gene expression is common and thus may play a significant role in human genetic variation.

COMPUTATIONAL ANALYSIS OF ALLELE-SPECIFIC GENE EXPRESSION

We developed a computational method by data mining of Unigene database to predict differential allelic gene expression and imprinted genes.²³ A schematic diagram of the computational method is shown in FIGURE 1.

For each SNP in a cDNA library, we could observe ESTs containing either allele A, allele B, or both alleles. D1 designates data that only allele A is observed in the cDNA library. If only allele A was observed, we can calculate the probability of genotype of the cDNA library as $P_{AA} = 1/(1 + 0.5^{n-1})$, $P_{AB} = 0.5^{n-1}/(1 + 0.5^{n-1})$, and $P_{BB} = 0$. Similarly, D3 designates data that only allele B is observed in the cDNA library. If only allele B was observed, we can calculate the probability of genotype of the cDNA library as $P_{BB} = 1/(1 + 0.5^{n-1})$, $P_{AB} = 0.5^{n-1}/(1 + 0.5^{n-1})$, and $P_{AA} = 0$. If both allele A and allele B were observed, $P_{AB} = 1$ and $P_{AA} = P_{BB} = 0$. For the population allele frequency, $Q_A = P_{AA} + 0.5P_{AB}$. From Hardy-Weinberg equilibrium, we obtained $Q_{AB} = 2Q_A(1 - Q_A)$. A significant reduction in observed heterozygosity (P_{AB}) compared to the expected heterozygosity (Q_{AB}) is computed using the Z-statistics.

Imprinted genes and mono-allelic genes differentially express one allele. To model allele-specific gene expression, it is assumed that each cDNA library represents an individual and all libraries (or the sum of available libraries) constitute a population. If both A and B alleles of an SNP in a gene X are represented in a cDNA library, the individual is heterozygous at the SNP in the gene X. If only allele A is represented in the cDNA library (FIG. 1, D1), the genotype for that individual could be either AA or AB. The probability that the individual is AA or AB can be inferred using Bayes' rule.

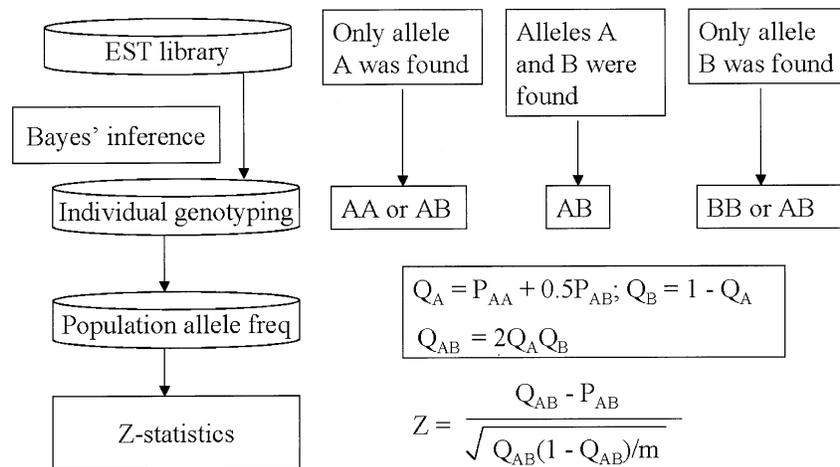


FIGURE 1. Computational analysis of genes preferentially expressing one allele.

We consider the following three kinds of allele observations D for an SNP from a library:

- D1: the allele A appeared n times in the library, or
- D2: the alleles A and B appeared n_1 and n_2 times, respectively, in the library, or
- D3: the allele B appeared n times in the library.

With the uniform prior, the posterior probability of genotypes AA and AB with the observation of n EST clones containing the allele A was calculated as

$$P_{AA|D1} = 1/(1 + 0.5^{n-1}) \text{ and } P_{AB|D1} = 0.5^{n-1}/(1 + 0.5^{n-1}).$$

Similarly, the posterior probability of genotypes BB and AB with the observation of n EST clones containing the allele B was calculated as

$$P_{BB|D3} = 1/(1 + 0.5^{n-1}) \text{ and } P_{AB|D3} = 0.5^{n-1}/(1 + 0.5^{n-1}).$$

When both alleles appeared at least once in the library,

$$P_{AB|D2} = 1, P_{AA|D2} = 0, \text{ and } P_{BB|D2} = 0.$$

Genotype frequencies, P_{AA} , P_{AB} , and P_{BB} , were estimated from individual genotypes. The allele frequency in the population is calculated as $Q_A = P_{AA} + 0.5P_{AB}$ and $Q_B = 1 - Q_A$. The expected heterozygote frequency based on the Hardy-Weinberg equilibrium distribution is calculated as $Q_{AB} = 2Q_AQ_B$. P_{AB} tends to be lower than Q_{AB} for imprinted genes and genes displaying mono-allelic expression. This behavior can be analyzed using Z-statistics described in FIGURE 1. Bayes' inference of genotypes and the computation of Z-statistics are two different procedures. The computational results from Bayes' inference are used in computing Z-statistics. The approach was applied to a data set based on SNP data from Buetow *et al.*²⁰ We have taken several steps to ensure that high-quality EST clones are used in our data set. The EST clones and SNPs must meet the following three criteria to be included in our data set: (1) Phred quality score of an EST clone is equal to or greater than 20; (2) SNP score is equal to or greater than 0.99;²⁰ (3) SNPs are mapped to Locuslink. This data set consists of 112,812 records for 19,312 unique SNPs.

The difference between P_{AB} and Q_{AB} was calculated for each SNP using Z-statistics. The probability of differential allele-specific expression is indicated by the P value for each SNP. Fifty of 19,312 SNPs in the data set are in known imprinted genes. The validity of the computational method was tested by determining if SNPs in imprinted genes had small P values, that is, within the top 1% (194 out of 19,312) of SNPs ordered according to increasing P value. Four SNPs in imprinted genes were in the top 1% of the data set: 3 in *IGF2* and 1 in *PEG3*. This finding is highly significant ($P = .0016$ in one-sided Fisher's exact test). Interestingly, when ESTs in tumor tissue libraries were used to populate the data set, only 1 of these 4 SNPs was in the top 1% of differentially expressed genes. This is consistent with the hypothesis that LOI occurs during tumorigenesis. Bayes' rule was used to infer the individual genotype frequencies. As a comparison, we consider the following non-Bayesian rule in the inference:

$$P_{AA|D1} = 1, P_{AB|D1} = P_{BB|D1} = 0, \text{ and}$$

$$P_{AB|D2} = 1, P_{AA|D2} = P_{BB|D2} = 0, \text{ and}$$

$$P_{BB|D3} = 1, P_{AA|D3} = P_{AB|D3} = 0.$$

When we replaced the Bayes' rule by the non-Bayesian rule, the SNPs in known imprinted genes had higher P values and were not in the top 1% of SNPs in the data set.

An alternative method for identifying imprinted genes was also developed. In this case, allele-specific gene expression is analyzed in libraries from heterozygotes. This approach identified 165 SNPs with differential allele-specific expression ($P < .05$, binomial test) and 2 of them were in known imprinted genes ($P = .0681$ in one-sided Fisher's exact test). Thus, this alternative method performs less well than the former method, although it may seem more intuitive.

An initial validation experiment demonstrated that 2 of 18 genes selected from the top 1% showed mono-allelic gene expression in fetal kidney and fetal liver using MALDI-TOF. Thus, we demonstrated the potential utility of this computational method in identifying differential allelic gene expression and novel imprinted genes.

SEQUENCE MOTIFS OF IMPRINTED GENES

We set out to identify novel sequence motifs that are associated with imprinted genes. Regulatory elements tend to locate on the conserved sequences.¹⁵ Thus, we searched conserved sequences between human and mouse imprinted genes using PipMaker program.²⁵ Genomic sequences of 41 imprinted genes (including their 10-kb upstream and 10-kb downstream sequences) were retrieved from <ftp://ftp.ncbi.nih.gov/genomes/>. We were able to find both human and mouse sequences for 36 imprinted genes, 24 of which were used as a training set and 12 of which were used as a testing set. The PipMaker program was used to align human and mouse genomic DNA sequences. We used the MEME program²⁶ to search motifs in the conserved noncoding sequences among the human imprinted genes. This analysis identified 16 motifs. Motifs 1–4 are located in the upstream regions of the imprinted genes, while motifs 5–8 and motifs 9–16 are located in the downstream and intron regions of the imprinted genes, respectively. We then used MAST program²⁷ to search the presence of these motifs in the 24 imprinted genes as well as 128 non-imprinted genes, which were identified in our previous study.²⁴ Fifteen of the 16 motifs were found to be significantly associated with the 24 imprinted genes ($P < .05$, Fisher's exact test).

It has been suggested that imprinted genes share some common features.^{16,28} Based on the distribution of the motifs among the 24 imprinted genes and the 128 nonimprinted genes, we developed a logistic regression model that was able to distinguish imprinted genes from nonimprinted genes. We initially had 16 motifs as predictor variables for the model. However, when all 16 motifs were used to build a logistic regression model, the iteration process to find the coefficients of the model was not convergent. We excluded motifs 4, 5, and 15 because their P values in Fisher's exact test were greater than 0.01. We also excluded motif 11 since it was underrepresented in the imprinted genes. We started a model with 12 motifs. An input vector to the model is a feature vector for a gene indicating whether each of these 12 motifs is associated with this gene. The response of the model is the probability of the gene being an imprinted gene. We performed the stepwise model selection by

minimizing the AIC criterion and found the optimal 6 motifs (motifs 3, 7, 10, 12, 13, and 16) as input variables for a logistic regression model to score imprinted genes. As we reduce the number of the predictor variables from 12 to 6, the AIC of the corresponding model drops from 43.2 to 34. The minimum AIC for the model with 5 motifs is 37. Thus, the 6 motifs are optimal predictors from the AIC point of view. In fact, we computed AIC for every possible subset of the 12-motif set. The 6-motif set (3, 7, 10, 12, 13, 16) has the minimum AIC. The estimated model is as follows:

$$P = 1/[1 + \exp(7.1 - 4.8 * M3 - 12.2 * M7 - 4.2 * M10 - 4.9 * M12 - 12.1 * M13 - 12 * M16)].$$

Our model correctly assigned 127 out of the 128 nonimprinted genes and 22 out of the 24 imprinted genes in the training set. The accuracy, sensitivity, and specificity of the model are 98%, 92%, and 99%, respectively. To further validate the model, we performed an open test on the 12 imprinted genes, which were set aside as a testing set as described. The model is able to assign high probability scores to 8 of the 12 imprinted genes.

BIOINFORMATICS APPROACH TO EPIGENOME RESEARCH

Extensive bioinformatics infrastructure is needed to fully integrate data from genome sequences, epigenome, experiments, analysis, and knowledge, and to apply these data to improve our understanding of cancer. We have built software tools and databases to support information-driven cancer research (FIG. 2). We have been developing robust software pipelines that can seamlessly integrate heterogeneous data from external and internal sources. Object-oriented programming has been used for the pipelines, which are implemented using Perl and Java. We stored data either in flat files or in a relational database such as Oracle to allow efficient storage, retrieval, and updating of the data. We built our bioinformatics system by leveraging

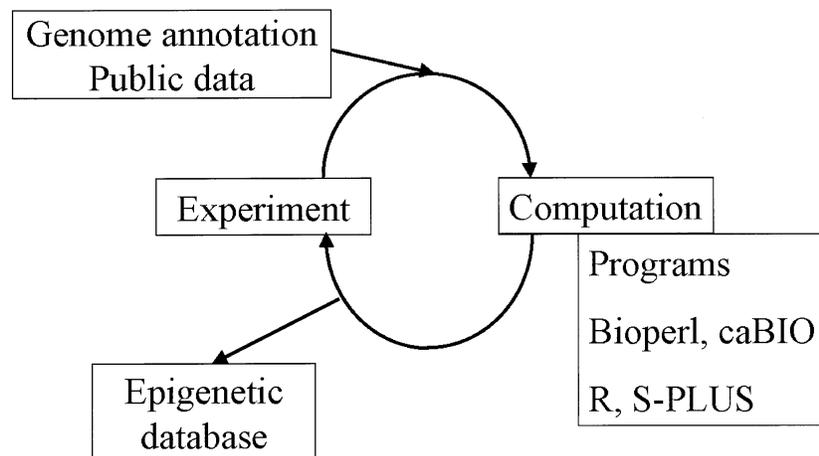
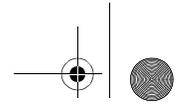


FIGURE 2. An integrated system to study cancer epigenetics.



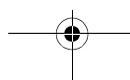
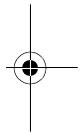
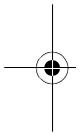
the existing infrastructures such as Bioperl (<http://bio.perl.org/>), NCBI (<http://www.ncbi.nlm.nih.gov/>), and caBIO (<http://ncicb.nci.nih.gov/core/caBIO/>). Our primary focus is to build a computation engine for biological knowledge discovery and bioinformatics system to manage the data and computation. We use statistical packages such as R, SPLUS, and SAS as the core, complemented by customized programs. We develop modules that automatically fetch data from external sources, preprocess the data, and format the input and output. Preprocessing data include data filtering, normalization, and merging relevant properties. Some examples of the application of statistical computation have been described in the previous sections. We are interested in extracting patterns, trends, and relationships of molecular activities in five levels from massive data sets using cluster analysis, classification, and regression techniques. The five levels are genome (sequence, SNP, and mutation), epigenome (methylation, imprinting, and chromatin), RNA (gene expression and RNA splicing), protein (protein expression and biochemical properties), and functions (cancer phenotype and cellular functions). It is critical to choose the correct features and models to quantify molecular processes in these levels. The learning problems for these models can be either unsupervised or supervised. The unsupervised learning methods such as clustering methods and self-organization maps will enable us to verify the existing functions and to discover new functions of genes. We can apply unsupervised learning methods to extract useful features that characterize the molecular activities. On the other hand, the supervised learning methods such as linear or nonlinear regression models, artificial neural networks, ensemble methods, and support vector machines will enable us to code biological knowledge into the models. A successful modeling depends critically on the strategy used for model selection when a huge number of factors are related to the process that we are trying to quantify. We can use Akaike's information criterion (AIC) (an example was shown in the previous section), Bayesian information criterion (BIC), Minimum Message Length (MML), and Minimum Description Length (MDL) methods for model selection. With the advanced machine learning methods and the model selection strategies, we will find better models for biological knowledge discovery.

SUMMARY

In conclusion, challenge in the next decade of biomedical research lies in the functional annotation of the genome, epigenome, and proteome in interaction networks. The success of this mission critically depends on integration between experiments and data management. The experimental design and execution should fully utilize existing data and knowledge, while computational analysis should be based on experimental data and supported by the data and should suggest new experiments. The continued cycles of experiments and computations will be the approach to study cancer in the systems biology era.

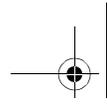
ACKNOWLEDGMENTS

We wish to thank Ying Hu for many stimulating discussions about the application of computation and statistics to cancer epigenetics.



REFERENCES

1. GARTLER, S.M. & M.A. GOLDMAN. 2001. Biology of the X chromosome. *Curr. Opin. Pediatr.* **13**: 340–345.
2. TYCKO, B. & I.M. MORISON. 2002. Physiological functions of imprinted genes. *J. Cell. Physiol.* **192**: 245–258.
3. YAN, H., W. YUAN, V.E. VELCULESCU *et al.* 2002. Allelic variation in human gene expression. *Science* **297**: 1143.
4. YAN, H., Z. DOBBIE, S.B. GRUBER *et al.* 2002. Small changes in expression affect predisposition to tumorigenesis. *Nat. Genet.* **30**: 25–26.
5. NICHOLLS, R.D., J.H. KNOLL, M.G. BUTLER *et al.* 1989. Genetic imprinting suggested by maternal heterodisomy in nondeletion Prader-Willi syndrome. *Nature* **342**: 281–285.
6. CLAYTON-SMITH, J. & M.E. PEMBREY. 1992. Angelman syndrome. *J. Med. Genet.* **29**: 412–415.
7. MANNENS, M., J.M. HOOVERS, E. REDEKER *et al.* 1994. Parental imprinting of human chromosome region 11p15.3-pter involved in the Beckwith-Wiedemann syndrome and various human neoplasia. *Eur. J. Hum. Genet.* **2**: 3–23.
8. RAINIER, S., L.A. JOHNSON, C.J. DOBRY *et al.* 1993. Relaxation of imprinted genes in human cancer. *Nature* **362**: 747–749.
9. OGAWA, O., M.R. ECCLES, J. SZETO *et al.* 1993. Relaxation of insulin-like growth factor II gene imprinting implicated in Wilms' tumour. *Nature* **362**: 749–751.
10. SUTCLIFFE, J.S., M. NAKAO, S. CHRISTIAN *et al.* 1994. Deletions of a differentially methylated CpG island at the SNRPN gene define a putative imprinting control region. *Nat. Genet.* **8**: 52–58.
11. WUTZ, A., O.W. SMRZKA, N. SCHWEIFER *et al.* 1997. Imprinted expression of the Igf2r gene depends on an intronic CpG island. *Nature* **389**: 745–749.
12. LEE, M.P., M.R. DEBAUN, K. MITSUYA *et al.* 1999. Loss of imprinting of a paternally expressed transcript, with antisense orientation to KVLQT1, occurs frequently in Beckwith-Wiedemann syndrome and is independent of insulin-like growth factor II imprinting. *Proc. Natl. Acad. Sci. USA* **96**: 5203–5208.
13. FEINBERG, A.P. 1999. Imprinting of a genomic domain of 11p15 and loss of imprinting in cancer: an introduction. *Cancer Res.* **59**: 1743s–1746s.
14. LEE, M.P., S. BRANDENBURG, G.M. LANDES *et al.* 1999. Two novel genes in the center of the 11p15 imprinted domain escape genomic imprinting. *Hum. Mol. Genet.* **8**: 683–690.
15. WASSERMAN, W.W., M. PALUMBO, W. THOMPSON *et al.* 2000. Human-mouse genome comparisons to locate regulatory sites. *Nat. Genet.* **26**: 225–228.
16. ONYANGO, P., W. MILLER, J. LEHOCZKY *et al.* 2000. Sequence and comparative analysis of the mouse 1-megabase region orthologous to the human 11p15 imprinted domain. *Genome Res.* **10**: 1697–1710.
17. CHARLIER, C., K. SEGERS, D. WAGENAAR *et al.* 2001. Human-ovine comparative sequencing of a 250-kb imprinted domain encompassing the callipyge (clpg) locus and identification of six imprinted transcripts: DLK1, DAT, GTL2, PEG11, antiPEG11, and MEG8. *Genome Res.* **11**: 850–862.
18. PAULSEN, M., S. TAKADA, N.A. YOUNGSON *et al.* 2001. Comparative sequence analysis of the imprinted Dlk1-Gtl2 locus in three mammalian species reveals highly conserved genomic elements and refines comparison with the Igf2-H19 region. *Genome Res.* **11**: 2085–2094.
19. STRAUSBERG, R.L., K.H. BUETOW, M.R. EMMERT-BUCK & R.D. KLAUSNER. 2000. The cancer genome anatomy project: building an annotated gene index. *Trends Genet.* **16**: 103–106.
20. BUETOW, K.H., M.N. EDMONSON & A.B. CASSIDY. 1999. Reliable identification of large numbers of candidate SNPs from public EST data. *Nat. Genet.* **21**: 323–325.
21. IRIZARRY, K., V. KUSTANOVICH, C. LI *et al.* 2000. Genome-wide analysis of single-nucleotide polymorphisms in human expressed sequences. *Nat. Genet.* **26**: 233–236.
22. MARTH, G.T., I. KORF, M.D. YANDELL *et al.* 1999. A general approach to single-nucleotide polymorphism discovery. *Nat. Genet.* **23**: 452–456.



23. YANG, H.H., Y. HU, M. EDMONSON *et al.* 2003. Computation method to identify differential allelic gene expression and novel imprinted genes. *Bioinformatics* **19**: 952–955.
24. LO, H.S., Z. WANG, Y. HU *et al.* 2003. Allelic variation in gene expression is common in the human genome. *Genome Res.* **13**: 1855–1862.
25. SCHWARTZ, S., Z. ZHANG, K.A. FRAZER *et al.* 2000. PipMaker—a Web server for aligning two genomic DNA sequences. *Genome Res.* **10**: 577–586.
26. BAILEY, T.L. & C. ELKAN. 1994. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **2**: 28–36.
27. BAILEY, T.L. & M. GRIBSKOV. 1998. Combining evidence using *P*-values: application to sequence homology searches. *Bioinformatics* **14**: 48–54.
28. GREALLY, J.M. 2002. Short interspersed transposable elements (SINEs) are excluded from imprinted regions in the human genome. *Proc. Natl. Acad. Sci. USA* **99**: 327–332.

