Profile of the Circulating DNA in Apparently Healthy Individuals

Julia Beck,[1] Howard B. Urnovitz,[1] Joachim Riggert,[2] Mario Clerici,[3,4] and Ekkehard Schütz[1*]

BACKGROUND: Circulating nucleic acids (CNAs) have been shown to have diagnostic utility in human diseases. The aim of this study was to sequence and organize CNAs to document typical profiles of circulating DNA in apparently healthy individuals.

METHODS: Serum DNA from 51 apparently healthy humans was extracted, amplified, sequenced via pyrosequencing (454 Life Sciences/Roche Diagnostics), and categorized by (a) origin (human vs xenogeneic), (b) functionality (repeats, genes, coding or noncoding), and (c) chromosomal localization. CNA results were compared with genomic DNA controls (n = 4) that were subjected to the identical procedure.

RESULTS: We obtained 4.5×10^5 sequences (7.5×10^7 nucleotides), of which 87% were attributable to known database sequences. Of these sequences, 97% were genomic, and 3% were xenogeneic. CNAs and genomic DNA did not differ with respect to sequences attributable to repeats, genes, RNA, and protein-coding DNA sequences. CNA tended to have a higher proportion of short interspersed nuclear element sequences ($P = 0.1$), of which Alu sequences were significant ($P < 0.01$). CNAs had a significantly lower proportion of L1 and L2 long interspersed nuclear element sequences ($P < 0.01$). In addition, hepatitis B virus (HBV) genotype F sequences were found in an individual accidentally evaluated as a healthy control.

CONCLUSIONS: Comparison of CNAs with genomic DNA suggests that nonspecific DNA release is not the sole origin for CNAs. The CNA profiling of healthy individuals we have described, together with the detailed biometric analysis, provides the basis for future studies of patients with specific diseases. Furthermore, the detection of previously unknown HBV infection suggests the capability of this method to uncover occult infections.

© 2009 American Association for Clinical Chemistry

Nucleic acids have been detected in the plasma, serum, and urine of healthy and diseased humans and animals (1). Both DNA and RNA can be isolated from serum and plasma and are commonly referred to as circulating nucleic acids (CNAs).[5]

Early work concentrated on detecting quantitative differences in circulating DNA between samples from patients with disease and samples from healthy individuals (2–4). The general diagnostic value of simple quantitative measures of circulating DNA is controversial (5–8).

Further work on the use of CNA as a diagnostic marker for neoplasia included the detection of qualitative rather than quantitative differences, such as specific oncogene mutations (9, 10), loss of heterozygosity (11–13), specific Alu amplicons (14), and methylation patterns (15) found in plasma or serum, and matching them with DNA characteristics in primary tumors (16).

Although most of the data available in the literature on the possible diagnostic uses of CNA were derived from studies of cancer patients, increases in circulating DNA have also been reported for other diseases, including trauma (17), stroke (18), Gulf War–related illnesses (19), autoimmune diseases such as systemic lupus erythematosus (20), and diabetes mellitus (4). In addition, fetal CNAs extracted from maternal plasma have served as markers in prenatal diagnostics (21), and fetal-DNA abnormalities have been linked to pregnancy-associated disorders (22, 23).

The cellular origin of the circulating DNA found in healthy individuals and the precise mechanisms by

[1] Chronix Biomedical GmbH, Goettingen, Germany; [2] Department of Transfusion Medicine, University of Goettingen, Goettingen, Germany; [3] Laboratory of Molecular Medicine and Biotechnology, Don C. Gnocchi ONLUS Foundation IRCCS, Milan, Italy; [4] Department of Biomedical Sciences and Technologies, University of Milan, Milan, Italy.
* Address correspondence to this author at: Chronix Biomedical, Goetheallee 8, 37073 Goettingen, Germany. Fax +49 551 37075726; e-mail esc@ chronixbiomedical.de.

[5] Nonstandard abbreviations: CNA, circulating nucleic acid; HBV, hepatitis B virus; WGA, whole-genome amplification; BLAST, Basic Local Alignment Search Tool; NCBI, National Center for Biotechnology Information; UTR, untranslated region; CDS, protein-coding DNA sequence.

which DNA enters the bloodstream are unknown. An early report found a correlation between plasma DNA concentrations and known markers of cell death in lung cancer patients, suggesting that at least a portion of the DNA in serum and plasma does originate from apoptotic cells (24). In favor of this hypothesis are data indicating that most of the circulating DNA in the plasma of sex-mismatched bone marrow transplant patients is of hematopoietic origin (25). Alternatively, active cellular release of newly synthesized DNA has been suggested (26–28).

A complete analysis of genomic sequences in circulating DNA, especially from healthy individuals, is not currently available. Recently, analyses were reported for 556 clones of plasma DNA obtained from healthy individuals (29). The availability of massively parallel sequencing technologies, such as the 454 Life Sciences/Roche Diagnostics GS FLX systems, allows the generation of 100 megabases of sequence information in a single experiment. For the first time, we have applied this high-throughput sequencing technology to generate an unbiased profile of the circulating DNA in healthy individuals, a profile based on an unprecedented amount of sequence information.

Materials and Methods

STUDY PARTICIPANTS

We obtained serum samples from 51 apparently healthy individuals (27 female and 24 male) between 18 and 64 years of age in the Department of Transfusion Medicine of the Georg-August University of Göttingen (n = 37) and the Don Gnocchi Foundation IRCCS repositories (n = 14). Donor samples were from excess serum from blood drawn for required serologic diagnostics in accordance with European Union regulation 98/79/EC. IRCCS samples were from apparently healthy volunteers. All donors provided written informed consent. All samples were anonymized. A previously undiagnosed hepatitis B virus (HBV) infection was found in one of the male volunteers (IRCCS); therefore, sequences obtained from this sample were excluded from the subsequent analysis of 50 apparently healthy individuals.

SAMPLING

Serum samples were collected and stored at −80 °C until further processing. Frozen serum was thawed at 4 °C, and cell debris was removed by brief centrifugation at 4000g for 20 min. Total nucleic acids were extracted from 200 μL of the supernatant with the High Pure Viral Nucleic Acid Kit (Roche Applied Science) according to the manufacturer's instructions. We also collected EDTA-anticoagulated samples of whole blood from a subgroup of the volunteers (2 females, 2 males) and extracted genomic DNA with standard protocols.

GENERATION OF RANDOM DNA LIBRARIES

We used the GenomePlex® Single Cell Whole Genome Amplification Kit (Sigma–Aldrich) according to manufacturer's instructions to amplify DNA from 1 μL of the nucleic acid solution extracted from serum. We amplified comparable amounts (0.1 ng) of genomic DNA with the same procedure. Figs. 1 and 2 in the Data Supplement that accompanies the online version of this article at http://www.clinchem.org/content/vol55/issue4 present the size distribution and the effect of whole-genome amplification (WGA). The amplified DNA preparations were sequenced directly with a GS FLX genome sequencer (454 Life Sciences/Roche Diagnostics) according to the manufacturer's instructions. Raw sequences were trimmed of sequences corresponding to the adapters and primers that were used.

SEQUENCE-ANALYSIS PIPELINE

We conducted local-alignment analyses with the BLAST program (Basic Local Alignment Search Tool) and highly stringent parameters to investigate the origins of circulating DNA molecules (30). To detect and mask repetitive elements, we used a local install of the RepeatMasker software package (Institute for Systems Biology) (31), which makes use of Repbase (version 12.09; Genetic Information Research Institute) (32). After masking repetitive elements and regions of low sequence complexity, we conducted sequential BLAST analyses for each sequence by querying databases of bacterial, viral, and fungal genomes, as well as the human genome (reference genome build 36.2). Bacterial, viral, fungal, and human genomes were obtained from the National Center for Biotechnology Information (NCBI) (ftp://ftp.ncbi.nih.gov). After each of the sequential database searches, we masked all parts of a queried sequence that produced significant hits ($e <$ 0.0001) and subsequently used the masked sequences to query the next database. To quantify the amounts of unidentified nucleotides, we counted and subtracted the masked nucleotides from the total nucleotide counts.

For each query fragment and each database search, we recorded the highest-scoring BLAST hit with a length of >50% of the query sequence in an SQL database. The highest-scoring BLAST hit was defined as the longest hit with the highest percent identity (maximum of hit length × identity). For each of the sequences, we recorded the start and stop positions for query and hit, and recorded the description of the corresponding matching subject. Repeat annotations and the respective lengths were recorded according to the output produced by the RepeatMasker software.

Table 1. Detailed definition of the analyzed genomic features.	
Functional genomic feature	**Detailed description**
Gene	Sequence annotated as gene by seq_gene file as obtained from NCBI
Pseudogene	Sequence annotated as pseudogene by seq_gene file as obtained from NCBI
RNA	All parts of a gene transcribed into RNA (according to NCBI annotation)
CDS	All parts of a gene/RNA translated into protein (according to NCBI annotation)
UTR	Parts of a gene transcribed; mRNA not translated into protein (according to NCBI annotation)
Intergenic sequence	All sequences not annotated as gene or pseudogene

The genome-annotation file for known and predicted genes (seq_gene.md) was obtained from the NCBI (download date, September 15, 2008). Only entries referring to the reference genome assembly were extracted from this file. Our evaluations of positions of hits within the genomic contigs with this annotation file led to hit counts and corresponding hit lengths within annotated genes and pseudogenes. We subdivided annotated gene sequences further into transcribed sequences [RNAs and untranslated regions (UTRs)] and protein-coding DNA sequences (CDSs) (Table 1). We normalized the observed nucleotide counts for each sample and each feature by the sample's total number of genomic hits, which was defined as the sum of nucleotides matching the human genome and repetitive-elements databases.

To exemplify the binning of the genomic elements and repeats, we used repetitive elements from the annotation file of the masked human genome available at the RepeatMasker Web site (31) in a joined query with the seq_gene.md database. Fig. 3 in the online Data Supplement presents a Venn diagram illustrating the relationships and binnings.

STATISTICAL ANALYSIS

The primary null hypothesis was equality of representation of the evaluated element in circulating DNA (observed values) and in genomic DNA processed and analyzed by the same methods (expected values). All serum nucleic acid data are presented as a ratio to the corresponding values for genomic DNA subjected to the same experimental and biometrical procedures. The Kolmogorov–Smirnov test was used to test all data sets for goodness of fit to a normal distribution. To compare observed and expected values, we used the value dispersion of the observed/expected ratio to generate Z statistics and derived the corresponding P value from the cumulative gaussian distribution function. Where applicable, we used the Bonferroni approach to correct P values for the effects of multiple testing. For parameters requiring correction for repetitive ele-

ments, statistical significance was set at P values <0.01, to account for gaussian error carry-forward effects.

Results

THE CIRCULATING GENOME

We generated 4.5×10^5 high-quality sequence reads (7.5×10^7 nucleotides total) from serum samples of 50 apparently healthy blood donors. The mean (SD) number of sequence reads per sample was 9100 (2620), of which 87% (5%) produced significant hits in one of the databases. Of these hits, 97% (4%) were assigned to be of genomic origin. The mean read length per sample was 169 (14) bp.

REPRESENTATION OF GENES, RNAs, AND CDSs

Only hits with a length >50% of the query length were considered for subsequent detailed allocations. The relative mean amounts of nucleotides matching to genes, pseudogenes, transcribed regions (annotated as RNAs and UTRs), and CDSs were calculated (observed) and compared with the mean amounts found in genomic DNA samples (expected). Because annotated genes, RNAs, and unprocessed pseudogenes contain introns and therefore repetitive elements, we corrected the expected amounts of these features for the amounts of repetitive elements they contained. This correction was necessary because all repetitive elements in the analyzed sequences could not be allocated to a unique genomic region and thus had to be masked before their use in queries to the human genome database.

Overall, a ratio of approximately 1 was found for all of the genomic features, indicating that essentially no difference existed between the circulating DNA pool and the genome in these features' representation in healthy individuals (Fig. 1).

The highest variation was observed in the representation of CDSs, UTRs, and pseudogenes in serum DNA samples (as well as in the genomic samples). In contrast, the representation of genes and RNA se-
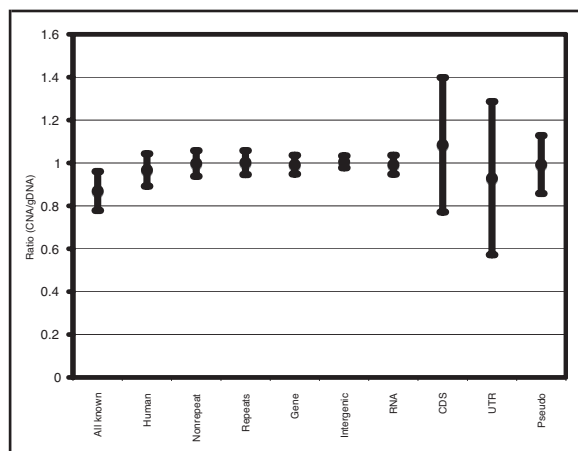
Fig. 1. Representation of sequences from different origins in the circulating DNA pool of healthy individuals, expressed as observed/expected ratios.

All hits with an expected value of <0.001 were evaluated for general assignment of nucleotides to unidentified, genomic, nonrepetitive, and repetitive sequence classes. For detailed allocation to the different genomic features, only hits >40 bp were determined. Whiskers represent 1.96× SD. gDNA, genomic DNA; Pseudo, pseudogene.

quences in the circulating DNA pool was more consistent among the samples from healthy individuals. Fig. 2, A–C, shows the correlation of serum values with genomic values in the 50 healthy individuals.

REPRESENTATION OF SINGLE GENES

The representation of serum sequences matching to annotated genes depended strongly on gene length. We compared the observed values for serum representation of sequences matching to the 4000 largest human genes with expected values and found gene length to be correlated with the gene's representation in serum {$r = 0.91$; $y = [1.11 (0.02)]x - [0.06 (0.02)]$} (95% CIs for slope and y intercept are indicated in parentheses). The observed/expected ratio was >5 in 4 genes and <0.2 in 40 genes.

CHROMOSOMAL DISTRIBUTION

Sequences obtained from serum DNA and genomic DNA were further evaluated by the chromosomal positions of their highest-scoring hits. We calculated the number of nucleotides matching to each of the human chromosomes and compared that number to chromosome length. The number of nucleotides derived from a chromosome was correlated with chromosome length for both sample types ($r^2 = 0.96$ for serum DNA samples; $r^2 = 0.93$ for genomic samples). The ratio of observed (serum) to expected (genomic) hit counts

was approximately 1 for all chromosomes with the exception of chromosome 19, for which the observed hits accounted for only 81% of the expected value ($P > 0.05$; Fig. 2D). Chromosome 19 has the highest gene density and GC content of any chromosome. Therefore, we tested whether the representation of the different chromosomes in the serum is correlated with chromosome gene density or GC content. Neither Pearson correlation ($r^2 = 0.22$) nor Spearman rank correlation ($r^2 = 0.03$) analysis revealed a significant correlation between gene density and representation of the different chromosomes in the serum. The correlation between GC content and chromosomal representation was also weak ($r^2 = 0.19$).

REPRESENTATION OF REPETITIVE ELEMENTS

Of the CNA fragments of human origin, 51.7% were of repetitive elements. A comparative analysis of the premasked human genome showed that 50.2% of the genome represents repetitive elements. We detected 51.6% repetitive elements in the sequenced genomic samples.
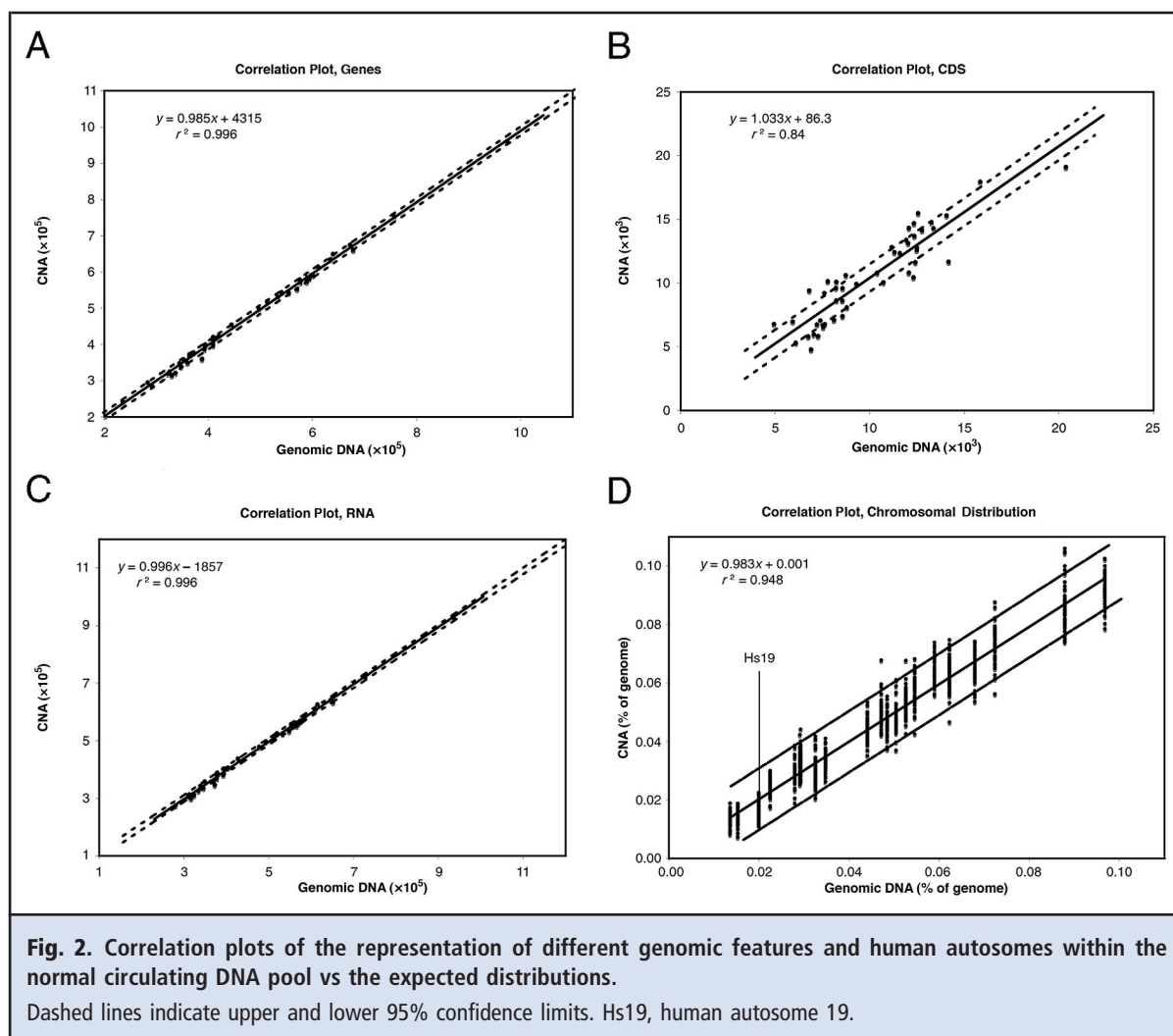
We detected repetitive elements within the circulating DNA sequences with RepeatMasker software and compared them with the amounts calculated for the genomic DNA samples.

No significant differences were detected between the genomic and CNA samples for the different classes of interspersed repeats (short interspersed nuclear elements, long interspersed nuclear elements, long terminal repeats, and DNA transposons) (Fig. 3A).

Further detailed analyses of the families and elements belonging to the most abundant repeat classes revealed an overrepresentation of *Alu* elements ($P < 0.01$; Fig. 3B). Whereas long interspersed nuclear elements L1 and L2 (Fig. 3C) were significantly underrepresented in the circulating DNA compared with the genome ($P < 0.01$), L3 elements were represented in genomic DNA and CNA samples to equivalent extents.

PROCEDURE CONTROL

We examined whether the experimental procedure or the computerized sequence analyses was responsible for any bias in the representation of repetitive elements. To estimate this bias, we ran 2 types of additional controls. For the first control experiment, we partitioned the human genome as obtained from NCBI into 175-bp segments. A second control consisted of shearing the genomic DNA by ultrasonication before the WGA reaction. We used the amounts of repetitive elements as calculated from the premasked human genome to compare the RepeatMasker results for the genomic DNA and sheared genomic DNA samples, as well as the results for the partitioned FASTA file. The premasked human genome was downloaded from the

Fig. 2. Correlation plots of the representation of different genomic features and human autosomes within the normal circulating DNA pool vs the expected distributions.

Dashed lines indicate upper and lower 95% confidence limits. Hs19, human autosome 19.

RepeatMasker Web site. Fig. 4 presents the ratios obtained in this analysis. The deviation of the amounts for the partitioned genomic FASTA sample from the nucleotide amounts of the unpartitioned genomic sequence reveals the bias that is introduced by querying short sequences alone. Deviations from 1, when seen only in the experimentally amplified genomic samples, indicate a bias that is introduced by the amplification or sequencing reactions. Close proximity of both lines indicates that shearing of the DNA before the WGA procedure introduces little additional bias. Shortening of the query sequence had no effect on nonrepetitive elements but did hinder the detection of repeats, particularly L1 elements, as seen in the divided FASTA sample. Both genomic DNA (whether sheared or of high molecular weight) and CNAs, however, showed an overrepresentation of L1 elements. This finding suggests that L1 elements are favored in the amplification reaction or in the sequencing procedure. On the con-

trary, the underestimation of L2 and L3 elements is introduced by the bioinformatics approach. Because the repetitive elements found in the DNA sequenced from serum samples were compared with the amounts in the genomic DNA samples subjected to the same experimental procedures, it is unlikely that experimental bias is a cause for these differences (Fig. 3).

We controlled the accuracy of sequence annotation in the query pipeline via the use of several representative parts of the human genome that were partitioned into 175-bp fragments as the input (total, 1.1×10^8 bp). We compared the annotation with the corresponding annotation in the seq_gene.md database and calculated an accuracy of >96% for genes, RNAs, CDSs, and UTRs.

The data in Fig. 4 reveal that the proportions of RNAs and CDSs in genomic DNA were lower than expected, a finding that appears mostly due to the WGA or sequencing procedure.
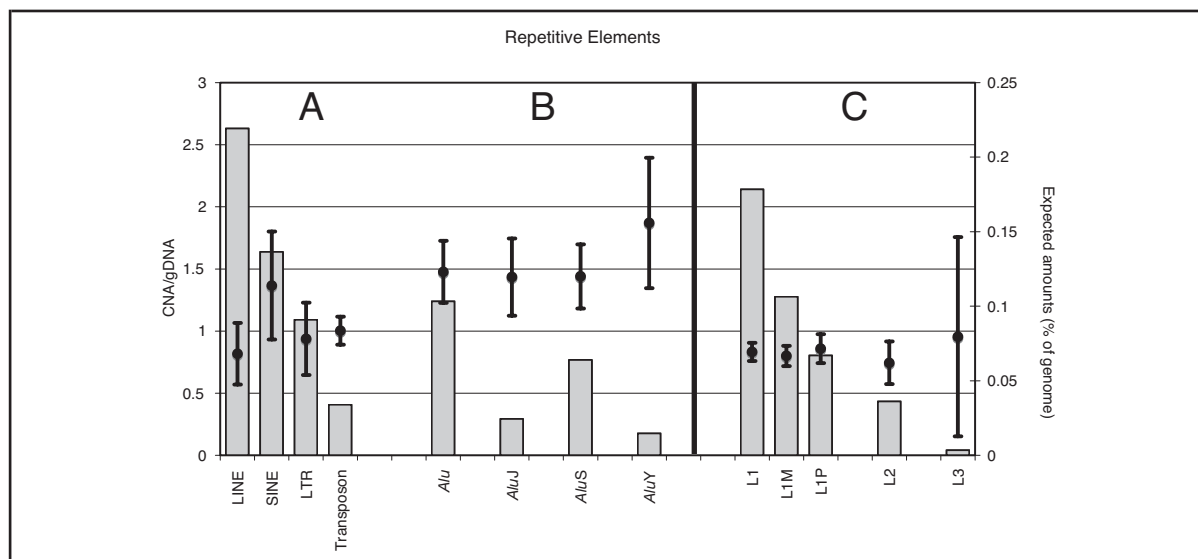
Fig. 3. Representation of repetitive elements expressed as observed/expected ratios (black dots).
Error bars represent 1.96× SD. Gray columns indicate the expected amounts as a percentage of the human genome covered by the respective repeat class/family/element (right *y* axis scale): repeat classes (A); *Alu* family belonging to the short interspersed nuclear element (SINE) class (B); families of the long interspersed nuclear element (LINE) class (C). gDNA, genomic DNA; LTR, long terminal repeat.

SEQUENCES MATCHING TO BACTERIAL AND VIRAL GENOMES
Of the total significant hits, 0.16% originated from the bacterial genomes database, and 0.02% and 0.01% were of viral and fungal origin, respectively. One of the control individuals had an undiagnosed HBV infection at sampling time. Of the total sequence data from this patient, 15.5% were HBV sequences. The complete HBV genome could be assembled from the sequence reads derived from this patient. Comparison of the consensus sequence against the known sequences of different HBV strains revealed the highest homology to HBV genotype F.

Discussion

We report sequence profiles for the circulating DNA pool in healthy individuals. The combination of random amplification of whole serum DNA isolates and high-throughput sequencing provides the first description and analysis of a large amount of unbiased sequence data. Use of a sequential BLAST-analysis pipeline allowed every fragment to be compared, not only to the endogenous genome but also to bacterial and viral genomes. An interesting finding was the detection of HBV infection in one of the volunteers, who was later determined to be an HBV carrier. This result shows that mass sequencing of serum nucleic acids can provide a powerful diagnostic approach for detecting not only disease-related endogenous CNA profiles but also blood-borne infectious agents.

Profiling of the circulating DNA present in the blood of healthy individuals provides valuable information for elucidating the origin of serum CNAs. Two sources of endogenous circulating DNA have been discussed in the literature: dying cells, whether necrotic or apoptotic, and DNA actively secreted by viable cells (27, 28, 33). Internucleosomal fragmentation of nuclear DNA occurs during the last stages of the apoptotic cascade (34, 35), and a small portion of the apoptotic genomic DNA has been experimentally shown to escape final cleavage to mono- or oligonucleotides and appear in the bloodstream or the urine (33, 36, 37). In a recently published analysis of 556 independent clones obtained from circulating DNA of healthy humans, the authors concluded that circulating DNA in plasma is derived from apoptotic cells rather than necrotic cells (29). They also reported that the number of clones derived from each chromosome was correlated with chromosome size. Our findings confirm that the representation of serum sequences is generally correlated with chromosome size, although we found a slight underrepresentation of chromosome 19. Chromosome 19 contains the most genes and has the highest amount of *Alu* elements and the highest GC content of any chromosome (38). We found no correlation of gene density or GC content with the chromosomal distribution of the sequenced fragments in either serum DNA or genomic DNA. It is
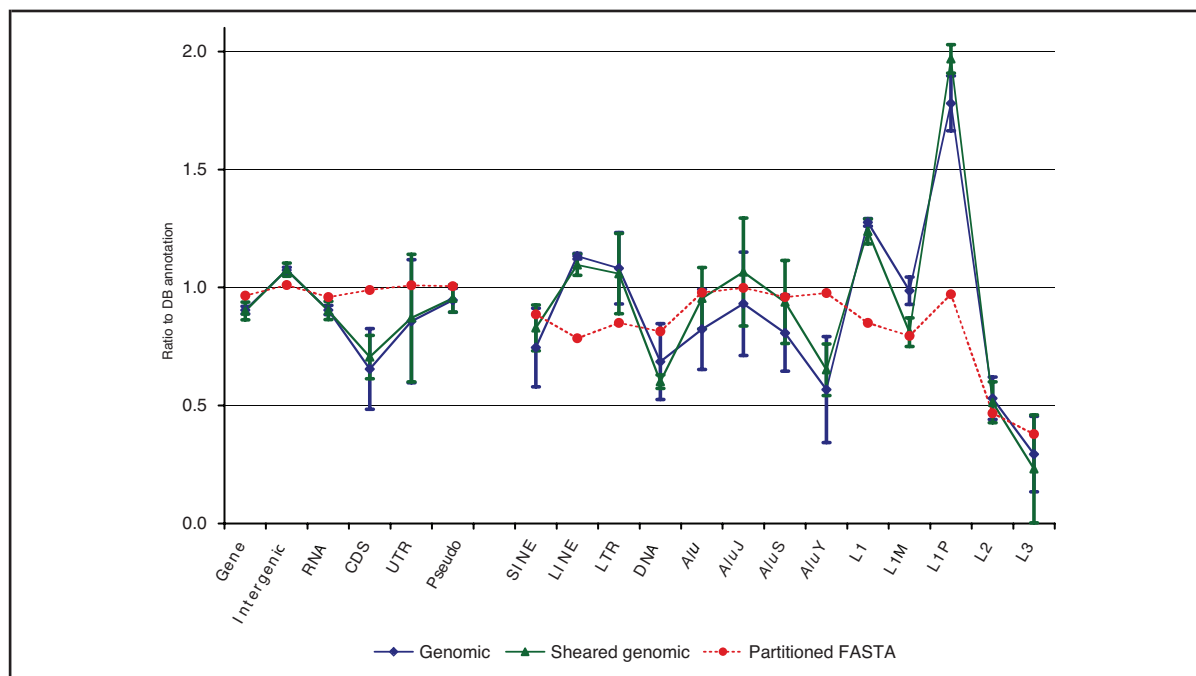
Fig. 4. Mean normalized nucleotide amounts for 4 samples of sheared genomic DNA (green triangles) or high molecular weight genomic DNA (blue diamonds) were calculated from pipeline results and divided by the normalized nucleotide amounts as calculated from the premasked human genome.

Error bars represent 1.96× SD. The human genomic sequence as downloaded from NCBI was split into 175-bp pieces and run through the repeat-masking procedure (red circles, right side). In addition, 1.1×10^8 bp were randomly selected, run through the pipeline, and directly compared with the corresponding genome annotation (red circles, left side). Shown are the normalized values calculated as above (red circles). Pseudo, pseudogene; SINE, short interspersed nuclear element; LINE, long interspersed nuclear element; LTR, long terminal repeat.

conceivable, however, that the underrepresentation of chromosome 19 in serum DNA is related to the overrepresentation of *Alu* sequences in the CNA pool of healthy individuals.

High-throughput sequencing data on the genomic distribution of cell-free DNA isolated from the plasma of pregnant women have recently been published *(39)*. The Solexa/Illumina platform was used for sequencing in this study, and isolated plasma DNA was used in the sequencing reaction without prior amplification of the DNA. The authors reported a strong bias in the representation of sequences toward GC-rich sequences for both the plasma and genomic DNA samples. First, these investigators found that the mean density of sequences matching to particular chromosomes correlated strongly with chromosomal GC content, and, second, the GC content of the sequenced fragments was, on average, approximately 10% higher than that of the sequenced human genome. The authors speculated that this bias is generated during the sequencing process *(39)*. The sequences obtained with our approach are not biased toward GC-rich regions, because the

GC-content value of 42.1% (0.2%) that we obtained is close to that of the sequenced genome (41%) *(38)*. In addition, we detected no correlation between chromosomal representation and GC content.

Suzuki and colleagues *(29)* reported that plasma samples from healthy individuals contained primarily DNA fragments of approximately 180 bp, with fragments >500 bp observed to a much lesser extent. Native CNA preparations extracted from 4 mL of serum (pooled from 3 different blood donors) on an Agilent Technologies 2100 Bioanalyzer displayed a comparable size distribution, which was not significantly altered by the WGA procedure.

The proportion of *Alu* repeat sequences relative to β-globin gene sequences has been reported to be greater in serum DNA than in lymphocyte DNA, in both healthy individuals and cancer patients *(40)*. Our results confirm this finding, because we found an overrepresentation of *Alu* elements in the CNA pool of 50 healthy individuals compared with human genomic samples. We obtained the same result when we compared the 4 CNA samples against the 4 genomic sam-

ples obtained from the same individuals. Sequences matching to *Alu* elements accounted for 11.4% (0.4%) of the total genomic hits in the CNA samples and 8.5% (0.8%) in the genomic samples.

L1 elements were found in higher proportions in genomic DNA and CNA samples than expected from the published genomic sequence, in which L1 retrotransposons account for approximately 17% of the genome *(38)*. L1Hs is the youngest branch of the L1 family, and some of its members are still retrotranscribed. It is estimated that 100 L1 copies in the human genome are still capable of retrotransposition, and approximately 10% of these active elements are classified as "hot," or highly active in an artificial culture system *(41)*. Although L1-activity potential has been shown to vary substantially between individuals *(42)*, the low number of active L1 sequences produces little quantitative effect on L1 counts. Our data indicated an overrepresentation of L1 elements in both the mean of 4 genomic samples (22.8%) and 50 serum DNA samples (19%), compared with the L1 content calculated from the premasked human genome (17.8%). From the small SDs for L1 in the sample groups (0.1% for the genomic samples and 0.8% for the serum DNA samples), we conclude that the interindividual variation in L1 cannot account for the detected differences.

Taken together, the data we have presented suggest that apoptotic genomic DNA is the major but not the sole source of CNAs in apparently healthy individuals. A circulating DNA pool consisting purely of unspecific apoptotic or necrotic nuclear DNA would have shown an even distribution over the entire genome or eventually some overrepresentation of highly histone-protected regions. Such a distribution is not seen in our data. Whether further subtle differences exist cannot be proved or excluded; deeper (high-coverage) sequencing would be needed to address such questions.

The profile of serum DNA from healthy individuals that we have presented provides baseline information, which is especially important because CNAs are increasingly recognized as valuable diagnostic biomarkers. The use of mass sequencing and bioinformatics provides the basis for new diagnostic approaches that use CNAs as biomarkers for both malignant and nonmalignant diseases.

References

1. Fleischhacker M, Schmidt B. Circulating nucleic acids (CNAs) and cancer—a survey. Biochim Biophys Acta 2007;1775:181–232.
2. Johnson PJ, Lo YM. Plasma nucleic acids in the diagnosis and management of malignant disease. Clin Chem 2002;48:1186–93.
3. Leon SA, Shapiro B, Sklaroff DM, Yaros MJ. Free DNA in the serum of cancer patients and the effect of therapy. Cancer Res 1977;37:646–50.
4. Swaminathan R, Butt AN. Circulating nucleic acids in plasma and serum: recent developments. Ann N Y Acad Sci 2006;1075:1–9.
5. Sozzi G, Conte D, Leon M, Ciricione R, Roz L, Ratcliffe C, et al. Quantification of free circulating DNA as a diagnostic marker in lung cancer. J Clin Oncol 2003;21:3902–8.
6. Wu TL, Zhang D, Chia JH, Tsao KH, Sun CF, Wu JT. Cell-free DNA: measurement in various carcinomas and establishment of normal reference range. Clin Chim Acta 2002;321:77–87.
7. Boddy JL, Gal S, Malone PR, Harris AL, Wainscoat JS. Prospective study of quantitation of plasma DNA levels in the diagnosis of malignant versus benign prostate disease. Clin Cancer Res 2005; 11:1394–9.
8. Boddy JL, Gal S, Malone PR, Shaida N, Wainscoat JS, Harris AL. The role of cell-free DNA size distribution in the management of prostate cancer. Oncol Res 2006;16:35–41.
9. Mayall F, Jacobson G, Wilkins R, Chang B. Mutations of p53 gene can be detected in the plasma of patients with large bowel carcinoma. J Clin Pathol 1998;51:611–3.
10. Sorenson GD, Pribish DM, Valone FH, Memoli VA, Bzik DJ, Yao SL. Soluble normal and mutated DNA sequences from single-copy genes in human blood. Cancer Epidemiol Biomarkers Prev 1994; 3:67–71.
11. Fujiwara Y, Chi DD, Wang H, Keleman P, Morton DL, Turner R, Hoon DS. Plasma DNA microsatellites as tumor-specific markers and indicators of tumor progression in melanoma patients. Cancer Res 1999;59:1567–71.
12. Nawroz H, Koch W, Anker P, Stroun M, Sidransky D. Microsatellite alterations in serum DNA of head and neck cancer patients. Nat Med 1996;2: 1035–7.
13. Silva JM, Dominguez G, Garcia JM, Gonzalez R, Villanueva MJ, Navarro F, et al. Presence of tumor DNA in plasma of breast cancer patients: clinicopathological correlations. Cancer Res 1999; 59:3251–6.
14. Durie BG, Urnovitz HB, Murphy WH. RT-PCR amplicons in the plasma of multiple myeloma patients—clinical relevance and molecular pathology. Acta Oncol 2000;39:789–96.
15. Korshunova Y, Maloney RK, Lakey N, Citek RW,

Bacher B, Budiman A, et al. Massively parallel bisulphite pyrosequencing reveals the molecular complexity of breast cancer-associated cytosine-methylation patterns obtained from tissue and serum DNA. Genome Res 2008;18:19–29.

16. Ziegler A, Zangemeister-Wittke U, Stahel RA. Circulating DNA: a new diagnostic gold mine? Cancer Treat Rev 2002;28:255–71.

17. Lo YM, Rainer TH, Chan LY, Hjelm NM, Cocks RA. Plasma DNA as a prognostic marker in trauma patients. Clin Chem 2000;46:319–23.

18. Rainer TH, Wong LK, Lam W, Yuen E, Lam NY, Metreweli C, Lo YM. Prognostic use of circulating plasma nucleic acid concentrations in patients with acute stroke. Clin Chem 2003;49:562–9.

19. Urnovitz HB, Tuite JJ, Higashida JM, Murphy WH. RNAs in the sera of Persian Gulf War veterans have segments homologous to chromosome 22q11.2. Clin Diagn Lab Immunol 1999;6:330–5.

20. Li JZ, Steinman CR. Plasma DNA in systemic lupus erythematosus. Characterization of cloned base sequences. Arthritis Rheum 1989;32:726–33.

21. Chim SS, Jin S, Lee TY, Lun FM, Lee WS, Chan LY, et al. Systematic search for placental DNA-methylation markers on chromosome 21: toward a maternal plasma-based epigenetic test for fetal trisomy 21. Clin Chem 2008;54:500–11.

22. Lo YM. Fetal DNA in maternal plasma: biology and diagnostic applications. Clin Chem 2000;46: 1903–6.

23. Tsui DW, Chan KC, Chim SS, Chan LW, Leung TY, Lau TK, et al. Quantitative aberrations of hypermethylated RASSF1A gene sequences in maternal plasma in pre-eclampsia. Prenat Diagn 2007;27: 1212–8.

24. Fournie GJ, Courtin JP, Laval F, Chale JJ, Pourrat JP, Pujazon MC, et al. Plasma DNA as a marker of cancerous cell death. Investigations in patients suffering from lung cancer and in nude mice bearing human tumours. Cancer Lett 1995;91: 221–7.

25. Lui YY, Chik KW, Chiu RW, Ho CY, Lam CW, Lo YM. Predominant hematopoietic origin of cell-free DNA in plasma and serum after sex-mismatched bone marrow transplantation. Clin Chem 2002;48:421–7.

26. Anker P, Stroun M, Maurice PA. Spontaneous extracellular synthesis of DNA released by human blood lymphocytes. Cancer Res 1976;36:2832–9.

27. Anker P, Mulcahy H, Chen XQ, Stroun M. Detection of circulating tumour DNA in the blood (plasma/serum) of cancer patients. Cancer Metastasis Rev 1999;18:65–73.

28. Stroun M, Maurice P, Vasioukhin V, Lyautey J, Lederrey C, Lefort F, et al. The origin and mechanism of circulating DNA. Ann N Y Acad Sci 2000;906:161–8.

29. Suzuki N, Kamataki A, Yamaki J, Homma Y. Characterization of circulating DNA in healthy human plasma. Clin Chim Acta 2008;387:55–8.

30. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25: 3389–402.

31. Smit AFA, Hubley R, Green P. RepeatMasker Open-3.0. http://www.repeatmasker.org (Accessed April 2008).

32. Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J. Repbase Update, a database of eukaryotic repetitive elements. Cytogenet Genome Res 2005;110:462–7.

33. Giacona MB, Ruben GC, Iczkowski KA, Roos TB, Porter DM, Sorenson GD. Cell-free DNA in human blood plasma: length measurements in patients with pancreatic cancer and healthy controls. Pancreas 1998;17:89–97.

34. Bicknell GR, Cohen GM. Cleavage of DNA to large kilobase pair fragments occurs in some forms of necrosis as well as apoptosis. Biochem Biophys Res Commun 1995;207:40–7.

35. Wyllie AH. Glucocorticoid-induced thymocyte apoptosis is associated with endogenous endonuclease activation. Nature 1980;284:555–6.

36. Botezatu I, Serdyuk O, Potapova G, Shelepov V, Alechina R, Molyaka Y, et al. Genetic analysis of DNA excreted in urine: a new approach for detecting specific genomic DNA sequences from cells dying in an organism. Clin Chem 2000;46: 1078–84.

37. Lichtenstein AV, Melkonyan HS, Tomei LD, Umansky SR. Circulating nucleic acids and apoptosis. Ann N Y Acad Sci 2001;945:239–49.

38. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. Nature 2001;409: 860–921.

39. Fan HC, Blumenfeld YJ, Chitkara U, Hudgins L, Quake SR. Noninvasive diagnosis of fetal aneuploidy by shotgun sequencing DNA from maternal blood. Proc Natl Acad Sci U S A 2008;105: 16266–71.

40. Stroun M, Lyautey J, Lederrey C, Mulcahy HE, Anker P. Alu repeat sequences are present in increased proportions compared to a unique gene in plasma/serum DNA: evidence for a preferential release from viable cells? Ann N Y Acad Sci 2001; 945:258–64.

41. Muotri AR, Marchetto MC, Coufal NG, Gage FH. The necessary junk: new functions for transposable elements. Hum Mol Genet 2007;16(Spec No 2):R159–67.

42. Seleme MC, Vetter MR, Cordaux R, Bastone L, Batzer MA, Kazazian HH Jr. Extensive individual variation in L1 retrotransposition capability contributes to human genetic diversity. Proc Natl Acad Sci U S A 2006;103:6611–6.