Sequence complexity of cDNA transcribed from a diverse mRNA population

Jeffrey Van Ness and William E. Hahn

Department of Anatomy, University of Colorado School of Medicine, Denver, CO 80262, USA

Received 8 July 1980

ABSTRACT

Mouse liver poly(A)$^+$ mRNA was reverse transcribed using oligo-p(dT) or random oligonucleotides as primers to yield cDNA about equal to the mass of the template RNA. The size profile of the oligo-p(dT)-primed cDNA was similar to that of the template RNA. RNA or cDNA driven saturation annealing of labeled single copy genomic DNA (scDNA) showed that 2% of the scDNA was complementary in either case indicating the sequence complexity of cDNA was equivalent to that of the template mRNA. These results establish for the first time that cDNA represents essentially all of the sequence complexity of a diverse template RNA population in which individual mRNA species are present in vastly different concentrations. RNA driven hybridization of the cDNA showed that about 80% of the cDNA mass represents most of the sequence complexity of the template RNA. Also, kinetics of this hybridization indicate a complexity of 58,000 kb for the template RNA, a value similar to that obtained by scDNA hybridization. We conclude that appropriately characterized cDNA probes can be used to make valid qualitative and quantitative comparisons of the complex, infrequent class mRNAs of different cells and tissues.

INTRODUCTION

Although full length cDNA copies of several specific messages have been synthesized (1,2), it has not been proven that the diversity and frequency distribution of all RNA in a complex population are represented in respective cDNA. In various cells and tissues most of the sequence complexity is contained in the mRNA species of the infrequent abundance class (3,4). mRNA species of the infrequent class comprise individually a very small amount (in some instances $1 \times 10^{-3}$% or less) of the total mRNA mass. It has not been rigorously demonstrated that the sequence complexity in this class is fully represented by readily measurable amounts of hybridizable cDNA. It is possible that a large number of different mRNA species do not serve, or serve inefficiently, as template for reverse transcription, owing perhaps to secondary structure (5). Also, rare species, if transcribed with relatively low efficiency, may be degraded before reverse transcription occurs when inhibition of RNase contaminating the reverse transcriptase is incomplete. If this is generally the case, the use of cDNA probes in making "complete" qualitative comparisons between complex mRNA populations in different cells and tissues would be invalid.
Nucleic Acids Research

If cDNAs primed by oligo-p(dT) do not equal the length of the template RNA a further problem arises if, as shown by Meyuhas and Perry (4), complex class mRNA species are generally above average in molecular weight. In the instance where reverse transcription is prematurely terminated, relatively less of the mass of the larger molecules would be represented in the cDNA. Thus an underestimate of proportion of total mRNA mass comprised by the complex class, and a corresponding overestimate of the relative mass of the abundant, low complexity class would be obtained from the kinetics of mRNAs-cDNA hybridization.

These possible problems with cDNA would be eliminated if it were shown that length and sequence complexity of cDNA synthesized from a diverse mRNA population was equal to that of the RNA. Therefore we synthesized cDNA from a complex population of mouse liver poly(A)$^+$mRNA molecules. Oligo-p(dT) or random oligonucleotides served as primers. The modal size and size distribution of the cDNA synthesized by the oligo-p(dT) priming method was shown to be similar to that of messenger RNA used as template. The sequence complexity of the cDNAs and respective template RNA were compared directly by measuring their complementarity with single copy genomic DNA (scDNA). Our results indicate that the sequence complexity of the cDNAs equals that of the template RNA. Kinetic data and quantitation of the yield of cDNA obtained from a known quantity of template imply that the copy frequency of the various cDNA sequences is similar to that of the respective RNA sequences.

METHODS
Isolation of Mouse Liver Polysomes

8-10 mouse (Swiss/Webster) livers, gall bladders removed, were placed in an ice slurry of 7% w/v sucrose (Schwarz-Mann, RNase-free), 0.15% Triton X-100, 0.25 M KC1, 10 mM MgCl$_2$, 50 mM Tris-HCl (pH 7.2) and 0.1% diethylpyrocarbonate. Livers were minced and then homogenized in above buffer by 6 strokes in a Dounce apparatus using an "A" pestle. The homogenate was centrifuged at 5000 x g for 5 min at 0°C to pellet nuclei. The 5000 x g supernatant was then subjected to 2, 10 min 20,000 x g centrifugations. Deoxycholate was added to 1% w/v and the supernatant was then layered onto six 10-45% sucrose gradients containing buffer as above without Triton X-100. Gradients were centrifuged for 2.5 hours at 26,000 rpm in an SW-27 rotor (Beckman) at 0°C. Material sedimenting faster than 100S was pooled. RNA was extracted by the method of Perry et al (6).

Isolation of Poly(A)$^+$mRNA

Poly(A)$^+$mRNA was isolated by oligo(dT)-cellulose chromatography essentially as previously described (7). The yield of poly(A)$^+$mRNA was 1.0-1.1% of the total polysomal RNA. 97% of the poly(A) was recovered from polysomal RNA as determined.
by hybridization with $^3$H-poly(U). The modal size of the poly(A)$^+$mRNA was approximately 1500 nt, similar to a previously reported value (8).

**Synthesis of cDNA**

The conditions for the synthesis of $^3$H-dCTP labeled or unlabeled cDNA from liver poly(A)$^+$mRNA using either random oligo-deoxynucleotides or oligo-p(dT) as primers were similar to those described by Kacian and Myers (9). Reaction mixtures contained 50 mM Tris-HCl (pH 8.0), 150 mM KCl, 8 mM MgCl$_2$, 4 mM Na pyrophosphate, 0.5 mM dithiothreitol, 0.4 mM each of dATP, dGTP and dCTP, 100 ug actinomycin D, 6A$_{260}$ units of oligo-p(dT)$_{12-18}$ or 35A$_{260}$ units of random oligodeoxynucleotides (10), 135 ug poly(A)$^+$mRNA and 150 units AMV reverse transcriptase (J. Beard, Life Sciences, Inc.) in a volume of 2 ml. To label part of the above cDNAs, 50 ul of each of the reaction mixtures (minus reverse transcriptase) were added to 2x10$^{-4}$ moles of dessicated $^3$H-dCTP (approx. 25 Ci/mmole). Four units of reverse transcriptase were then added. After incubation for 75 min at 37°C the reactions were stopped by the addition of Na$_2$EDTA and SDS. E. coli RNA and phage fd DNA were added as carrier and each mixture was then extracted with phenol/chloroform pH 7.2. The cDNA was excluded from Sephadex G-100 in 0.1 M NaCl, 10 mM Tris-HCl (pH 7.4), 1 mM EDTA. To hydrolyze the RNA moiety, the cDNA-mRNA was placed in 0.1 N NaOH for 24 hours at 25°C, neutralized with HCl and precipitated with ethanol after the addition of yeast tRNA.

**Electrophoresis and Shearing of cDNA**

The size of the cDNAs was determined by electrophoresis in alkaline agarose gels (11). Fragments of SV40 DNA generated by digestion with Hae III served as the markers. Both the oligo-p(dT) and random oligonucleotide primed cDNAs in Tris-buffered 0.05 M NaCl were sheared in a pressure cell to a modal size of about 350 nt (Fig. 1). Shearing of the cDNA increases the likelihood that the length of the RNA driver will remain greater than tracer despite cleavage of the RNA during incubation at high temperature.

**Annealing of Nucleic Acids**

Hybridization reaction mixtures using scDNA tracer contained 0.12 or 0.4 Na phosphate (pH 6.8, equal molar mixture mono and dibasic sodium phosphate), 0.5% SDS and 5 mM EDTA. The concentration of the RNA was from 0.05 to 9 ug/ul depending on the desired values of Cot. Cot values for RNA driven reactions in the presence of 0.6 M Na$^+$ were normalized for effect of salt concentration as previously described (3). Conditions for cDNA driven annealing of scDNA were the same as above and Cot values were normalized to 0.18 M Na$^+$ by reference to values pertaining to DNA reassociation (12).

Hybridization mixtures, in which cDNA was used as the tracer, contained 2x10$^{-2}$...
µg/µl to 6 µg/µl RNA, 0.12 M Na-phosphate, SDS and EDTA as above. Reaction aliquots were sealed in glass capillary tubes containing 10^3-10^4 cpm of tracer DNA and incubated at 64-67°C. EMC virus RNA driven hybridization of respective cDNA under the same conditions was used as a complexity standard (99 L. mole^-1 sec^-1, 13). Hybrids and duplexes were assayed by the S_1 nuclease/DEAE-81 filter method (14) as modified by Van Ness et al (3). If nucleic acid level was low, as in the case of samples containing only tracer, sheared double strand DNA or E. coli RNA was added (0.04 µg/µl) as protection against possible double strand hydrolysis.

### Determination of Extent of ^3^H-scDNA Self Renaturation

The renaturation of the ^3^H-scDNA in hybridization mixtures was monitored by the low-salt RNase procedure (15) in which RNA/DNA hybrids are discriminated from DNA/DNA duplexes by digestion of the RNA thus rendering the hybridized scDNA sensitive to S_1 nuclease. Correction for self-annealing of the scDNA tracer in cDNA driven annealing was made with reference to a parallel system in which E. coli RNA (no apparent measurable homology with mouse scDNA) was substituted for cDNA. This parallel control system was used in place of a system containing tracer DNA alone because self-annealing of the scDNA tracer was found to occur somewhat faster in the absence of nucleic acids than in their presence at the concentration used to drive the annealing of scDNA. The amount of self-annealed scDNA tracer was then determined as for the RNA driven reactions. Values for self-renatured scDNA tracer, relative to Cot were reproducible using the E. Coli RNA system and duplicated the values for self-renaturation monitored by the RNase-release method applied in the case of RNA driven hybridization (see above). Self annealed tracer constituted about half of the S_1 nuclease resistance after an incubation period sufficient to achieve saturation of the scDNA with the cDNA drivers.

### RESULTS

#### Synthesis, Yield and Size of cDNA

The rate of incorporation of ^3^H-dCTP was similar using either oligo-p(dT) or random oligo-deoxyribonucleotide primers. A plateau in incorporation was reached after an incubation of 75 minutes. To determine whether any double-strand cDNA was synthesized (foldbacks), labeled cDNA was assayed for S_1 resistance. Only 0.5% of the cDNA was resistant as measured by binding to filter discs of DEAE-cellulose. In addition, when the cDNA was incubated under annealing conditions, formation of S_1 nuclease resistant duplex did not occur. Therefore cDNA anticomplementary to RNA was apparently not synthesized.

From the 135 µg of the poly(A)^+ mRNA 130 µg of oligo-p(dT)-primed, as determined by absorbance at 260 nm, was obtained. A yield of 110 µg of random
oligonucleotide-primed cDNA was obtained from the same amount of mRNA. This corresponds to an efficiency of transcription of 96% and 80% for oligo-p(dT) and random oligonucleotide primed template respectively, assuming 100% recovery of the cDNA from the transcription mixtures (recoveries of labeled cDNA from the synthesis mixture were 98-100%; data not shown). Thus, a mass of cDNA almost equal to the amount of RNA template was synthesized using the oligo-p(dT) primer. In the case of random oligo-nucleotide priming, the maximum fractional mass of cDNA synthesized from a given mass of template equals \(1-(1/n+1)\) where \(n\) equals the average number of random priming sites on the template molecule. Since each RNA molecule was primed three times on average (see below), a maximal transcriptional efficiency of 75% would be expected. The value of 80% obtained experimentally is close to the predicted value. Estimation of transcriptional efficiency based on incorporation of labeled deoxynucleotides was unreliable owing to imprecise knowledge of the specific activity of the precursors.

The size distribution before and after shearing of both preparations of cDNA is shown in Figure 1. Modal size of the oligo-(dT) primed cDNA was 1350 nt, similar to

![Figure 1. Electrophoresis in alkaline gels of agarose of cDNA transcribed from liver poly(A)+mRNA primed by random deoxyoligonucleotides (O--O) or oligo-p(dT) (A--A). Solid symbols pertain to respective cDNAs after shearing. Arrows indicate position of labeled "marker" DNAs.](image)
the 1500 nt modal size of the template RNA (see Methods). The number average size of the cDNA synthesized from random oligonucleotide primed RNA was about 410 nt indicating each mRNA was primed, on average, at 3 different sites, when a mass ratio of 10:1 of primer to template at the concentration specified was used (see Methods). Modal size of the cDNAs after shearing was about 350 nt.

Complexity of Liver Poly(A)^+mRNA and Respective cDNA

As summarized in Table I, the sequence complexity of liver poly(A)^+mRNA and respective cDNAs, synthesized using oligo-p(dT) or random oligonucleotides as primers, is equivalent as determined by annealing scDNA to saturation. Thus, within the limits of the sensitivity of this measurement, all of the complex, infrequent class of mRNA appears to be represented in the cDNAs. Apparent saturation was achieved with all the driver nucleic acids at Cot 10^4 - 2x10^4 (Fig. 2). Although kinetic data for the cDNA driven reaction is incomplete, the cDNAs and poly(A)^+mRNA appeared to anneal scDNA at similar rates. This implies, since both reactions are pseudo-first order, that the "driver" nucleic acids probably had a similar sequence frequency distribution.

Estimation of Template RNA Complexity from the Kinetics of Hybridization

Poly(A)^+mRNA driven hybridization of cDNA occurs over 5 Cot decades indicating the presence of 3 general copy frequency classes of mRNA (Fig. 3). Kinetics obtained using tracer cDNA synthesized from either oligo-p(dT) or random oligonucleotide primed RNA were essentially the same. The reaction terminates at Cot 600-800. At termination 90-92% of the cDNA was resistant to S^1 nuclease (see Discussion).

We estimated the complexity of each frequency class by the method of Jacquet et al (16). EMC viral RNA driven hybridization of respective cDNA performed under identical conditions was used as a complexity standard (9.9 kb, 13). The complexity of mRNAs of the high, intermediate and infrequent copy frequency class is 13, 1300 and

Table I. Sequence Complexity of cDNA and Respective Template Poly(A)^+mRNA.

<table>
<thead>
<tr>
<th>Driver Nucleic Acid</th>
<th>% scDNA Hybridized</th>
<th>Complexity in Kilobases</th>
</tr>
</thead>
<tbody>
<tr>
<td>Poly(A)^+mRNA, Liver</td>
<td>2.0 ± 0.1 (4)</td>
<td>6.5 x 10^4</td>
</tr>
<tr>
<td>cDNA, Oligo-p(dT) Primed</td>
<td>2.0 ± 0.2 (6)</td>
<td>6.5 x 10^4</td>
</tr>
<tr>
<td>cDNA, Random Oligonucleotide Primed</td>
<td>2.1 ± 0.2 (6)</td>
<td>6.8 x 10^4</td>
</tr>
</tbody>
</table>

Number in parentheses is number of apparent saturation values averaged to yield value given. Complexity in kb is calculated on the basis that the haploid amount of single copy DNA in the mouse genome is 3.25 x 10^9 nucleotides ± one standard deviation.
Figure 2. Hybridization or annealing of scDNA. Solid circles indicate liver poly-(A)$^+$ mRNA driven hybridization; open circles indicate respective oligo-p(dT)-primed cDNA driven annealing. Curve is fitted to RNA driven hybridization.

57,000 kb respectively. The summed complexity of these classes is fairly similar to the complexity estimate from saturation hybridization of scDNA (Table I).

40% of the hybridized ($S_1$ nuclease resistant) cDNA mass is complementary to the

Figure 3. Liver poly(A)$^+$ mRNA driven hybridization of labeled cDNA. Solid circles: oligo-p(dT)-primed cDNA; open circles: random deoxynucleotide-primed cDNA.
infrequent, complex class of the mRNA population. Assuming 33 pg of total RNA per liver cell, of which 1% or $4 \times 10^5$ molecules 1.5 kb in size is poly(A)$^+$mRNA (17), the average copy frequency of the infrequent, complex class is calculated as: $(4.0 \times 10^5) (0.04/4 \times 10^4) = 4$, 1.5 kb molecules/cell where 0.04 is the fraction of mRNA comprised by the complex class and $4 \times 10^4$ is the number of different mRNA species, 1.5 kb in average size. If mouse liver poly(A)$^+$mRNA content is similar to that of the rat (0.13 pg/cell, Sippel et al (8), then the copy frequency of the complex class is 2.0 copies/cell.

DISCUSSION

These results show that qualitative and quantitative representation in cDNA of essentially all of the sequences in a complex population of mRNA molecules can be obtained. This was suggested by the observation that an approximate mass equivalent of the cDNA relative to template mRNA was obtained and established by cDNA driven annealing of scDNA. Because of the uncertainty as to whether all the sequence complexity of diverse mRNAs is represented in cDNA probes, mDNAs (scDNA complementary to messenger RNA) have been the probes of choice in making qualitative comparisons between mRNA populations of different cell types and tissues. While the use of mDNA tracers for comparing the extent of homology between mRNA populations is a direct and useful method, difficulties in preparing mDNA probes of complete hybridizability are apparent (18,19). Furthermore, when very little of the scDNA is complementary to the mRNA population in question, it is burdensome to recover the small fraction of the total genomic DNA as mDNA. Since our results show that the infrequent, complex class of mRNA is fully represented in respective cDNA, cDNA tracers can be used as probes for making valid qualitative comparisons between complex mRNA populations. Although such comparisons have been previously made using cDNA probes, conclusions were based on the assumption that most sequences in the RNA were represented, in proportion to their relative mass, in the cDNA.

Theoretically 100% of the cDNA should be hybridizable, but in practice this is usually not observed at least in terms of resistance to S$_1$ nuclease. It is possible that some end nibbling of the hybrids occurs during "exhaustive" treatment with S$_1$ nuclease. Also cDNA sequences are known to be present, which, despite adequate size, fail to hybridize with template RNA (Van Ness and Hahn, unpublished data).

It is possible that the presence of an oligo or poly(dT) sequence on all cDNA molecules synthesized using oligo-p(dT) primers might have an effect on hybridization kinetics if base pairing of poly(A) of the mRNA and oligo or poly(dT) of the cDNA increases stability of nucleation events. If this is the case, the effect must be small since the kinetics of poly(A)$^+$mRNA driven hybridization of cDNA synthesized using random oligonucleotides or oligo-p(dT) as primers are similar (Fig. 3).
The accuracy of complexity estimates on mRNA populations by the kinetic (cDNA) method depends upon the reliability of the complexity standard and upon the precision with which the rate constant, or Cot% for the component(s) representing the complex, infrequent mRNA species is determined (see 20 for discussion of copy frequency components in mRNA). It is desirable to determine the hybridization kinetics of the complexity standard under the same conditions (particularly salt and temperature) as the test system to avoid errors which may result when normalizing corrections are made (Van Ness and Hahn, in preparation; see also 3). If the cDNA tracer is not fully hybridizable, or if the complex class is inadequately represented in the cDNA, estimate of Cot at termination or of the Cot% pure of the slowest component cannot be accurately made. Earlier Hastie and Bishop (21) estimated the complexity of poly(A)$^{+}$ cytoplasmic RNA from mouse liver to be about 24,000 kb. A complexity estimate of about 15,000 kb for mouse liver poly(A)$^{+}$ cytoplasmic RNA is also apparent from the results of Ryffel and McCarthy (22), although these authors expressed substantial reservations regarding complexity estimates based on the cDNA method. These estimates are about 3 times lower than we report here. In each of the above studies only about 75% of the cDNA hybridized and mRNA species of the complex class appear not to have been adequately represented in the cDNA since the hybridization apparently terminated by about Cot 100. It is worth noting that the kinetics of hybridization of our cDNA probes, up to 70% hybridization, is similar to that presented in the reports cited above, but termination of the slowest reacting component occurs between Cot 600-800 when more than 90% of the cDNA has become resistant to S$	extsubscript{nuc}$ nuclease.

As shown here, the estimate of the complexity of liver poly(A)$^{+}$mRNA obtained by the cDNA method is similar to that based on complementarity to scDNA. This was also the case for complexity measurements made previously on poly(A)$^{+}$ and poly(A)$^{-}$ mRNA from brain (3,13) and for rat liver poly(A)$^{+}$mRNA (23). Agreement between these two methods indicates that overestimates do not result from the scDNA method which would likely occur if most sequences which code for mRNA are slightly repetitious or if scDNA probes were significantly enriched for code sequences due to preferential loss of non-coding scDNA during preparative procedures (see 24 for discussion).

Hybridization of cDNA shows that most of the sequence complexity of the mRNA is contained in about 40% of the mass. Since the cDNA represents qualitatively and quantitatively essentially all of the mRNA sequences, this estimation of the mass comprising most of the complexity is reliable. Thus about 40% of the mass of the mRNA is driver in the hybridization of most of the scDNA, or "driver" Cot is equal to about 0.4 of gross Cot. Relative to the EMC standard, and the
hybridization of cDNA representing the complex class of mRNA (Fig. 3), hybridization of scDNA occurs overall much slower than expected assuming 40% of the mass of the mRNA constitutes most of the complexity. The slower than expected rate and biphasic hybridization of scDNA shown in Fig. 2 is due, in large extent, to sequence discontinuity of many of the scDNA fragments with complementary mRNA driver. scDNA fragments (about 350 nt) are frequently a mixture of mRNA coding and intervening sequences. We have found that sequence discontinuity markedly retards the rate of hybridization. (Experimental details and data pertinent to this contention will be presented elsewhere, Van Ness and Hahn, in preparation.)

Although the mechanism whereby sequence discontinuity retards the rate of hybridization remains to be elucidated, it is apparent that the portion of the total mRNA mass comprised by the complex class species can be greatly underestimated from the rate of mRNA driven hybridization of scDNA as assayed by resistance to S1 nuclease. Due to lack of a single pseudo-first order component of hybridization of scDNA (Fig. 2), it is difficult to estimate the mass of the mRNA which drives the reaction but a rough estimate is in the range of 5-10%, in contrast to about 40% from hybridization of cDNA. The extent of discrepancy by the two methods may vary considerably depending upon the mRNA code sequence arrangement of the genomic DNA of the species in question, and the size of the scDNA fragments used as tracer. Recent studies on cloned sea urchin DNA probes representing infrequent mRNA species suggest that, collectively, species of the complex, infrequent class constitute about 40% of the total mRNA mass in sea urchin gastrulae (E.H. Davidson, personal communication), which is about 2-4 times greater than estimates based upon the kinetics of scDNA hybridization (15,25).

Acknowledgement
This work was supported by a grant from the National Institutes of Health.

References


