UNIVERSITY OF CALIFORNIA
RIVERSIDE

Consensus Sequences and Variability Information for the Human and Chimpanzee rDNA Repeats

A Dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Evolution, Ecology, and Organismal Biology

by

Curtis S. Adams

June 2008

Dissertation Committee:
Dr. Leonard Nunney, Chairperson
Dr. Stefano Lonardi
Dr. David Reznick

Signature Page on file in the
Graduate Division

ABSTRACT OF THE DISSERTATION

Consensus Sequences and Variability Information
for the Human and Chimpanzee rDNA Repeats

by

Curtis S. Adams

Doctor of Philosophy, Graduate Program in Evolution, Ecology, and Organismal Biology
University of California, Riverside, June 2008
Leonard Nunney, Chairperson

In the past obtaining sequence information for the entire ribosomal DNA (rDNA) repeat has been difficult and rarely attempted in spite of its importance as a component of all cellular genomes, with potential involvement in aging and speciation. The limitations have persisted into the genomic era since current assembly methods cannot deal with highly repetitive regions such as the rDNA repeat arrays. Here I describe Iterative Stochastic Assembly, a new method for obtaining sequence information from highly repetitive regions based on an iterative approach, where partial or inaccurate sequence assemblies are repeatedly improved with available read information until a stable consensus sequence is found. I apply this method to the human and chimpanzee rDNA repeats, obtaining verifiable consensus sequences for both species. Comparison of the human consensus with the two available sample human rDNA sequences indicates typical sequences differ markedly from the consensus and thus a consensus approach is needed. Both species show high variability with 3.1% of human and 3.2% of chimpanzee base pairs demonstrably polymorphic in the functionally unconstrained intergenic spacer region. Comparison of the human and chimpanzee consensuses shows that the entire

repeat, apart from some functionally conserved regions, evolves rapidly, with mutations fixed approximately 3 times as fast as in euchromatic sequence. Polymorphisms become fixed much more rapidly than would be predicted by drift-based models, suggesting a role for selection in the evolution of rDNA repeat diversity. Unlike in *D. melanogaster* and *S. cerevisiae*, there is minimal transposable element activity in the rDNA repeat. The rapid creation and fixation of polymorphisms, however, may permit a window into subspecies structure to address questions such as the status of proposed chimpanzee subspecies and interbreeding between *Homo sapiens* and other hominids. The ISA technique can readily be applied to other species and other tandem repeats, allowing testing of open hypotheses on tandem repeat evolution and function and deepening our understanding of these poorly understood regions.

Table Of Contents

List of Tables

List of Figures

Introduction

Biology as a science has experienced tremendous progress in the recent past. Many deep questions such as the mechanisms of heredity and evolution, and the descent relationships of living things, have largely yielded to scientific inquiry. Still, some of the most important restrictions to human lives, including aging and cancer, remain incompletely explained. One fundamental question with aging and cancer is how two processes with substantial fitness costs can have such similar characteristics over a wide phylogenetic range in spite of the demonstrated ability to evolve lower rates of these costly processes.

Of these two processes, aging has been subject to much more extensive theoretical work. Two primary models have been suggested: mutation accumulation, in which mutations with late-life effects are weakly selected against, and antagonistic pleiotropy, in which mutations causing aging provide compensatory advantages. Either process could explain the wide prevalence of aging (Charlesworth 1994). However, neither process can explain the near-universality of the exponential increase in death rate with aging, referred to as the Gompertz curve (Gavrilov and Gavrilova 2001, Suematsu and Kohno, 1999). Models have been produced with auxiliary assumptions, but the assumptions are either implausible (deleterious alleles have a Poisson distribution, Gavrilov and Gavrilova 2001) or falsified (aging does not affect survival in the wild, Charlesworth 2001).

One proposed explanation for diseases of aging is that undescribed, or at least incompletely understood, disease processes cause them (Cochran *et al.*, 2000). The prime motivation for this is theoretical estimates of the difficulty of evolving reduced rates of

major aging-associated diseases, which indicates that the observed rates of such disease should be much lower than is actually observed. If disease processes cause aging, then preventing aging requires that the host species win an evolutionary race with the underlying pathogens, a more difficult process than developing countermeasures for a particular biochemical or mechanical process. An underlying disease process also provides a direct explanation for the Gompertz curve, as the spread of a disease typically follows exponential kinetics. Direct evidence for disease processes causing aging is absent, although atherosclerotic plaques in humans often have associated pathogenic bacteria and this finding has been cited as support for the model of aging as a disease process (Liu *et al.*, 2006).

However, direct experimentation has largely excluded external pathogens as a basic cause of aging. Animals raised in pathogen-free environments age at roughly the same rate as animals exposed to pathogens, although their life expectancies increase due to the less hazardous environment (Coleman *et al.,* 1977). Hence, if aging is a pathogen-mediated process almosts all animals would have to be born with them.

There are, indeed, pathogens virtually all animals are born with, endogenous transposable elements (Arkhipova and Meselson, 2000). These include a variety of different elements that are transmitted vertically from parent to child. Some, the endogenous retroviruses, are related to exogenous horizontally transmitted viruses (Löwer *et al.,* 1996), while others, including LINES and Alu sequences, are not thought to be related to viruses (Sun *et al.,* 2007).

Some transposable elements can shorten lifespan if active in somatic cells. The p element, a transposable element found in fruit flies, has transcriptional regulators that suppress its activity in somatic tissues. When *Drosophila melanogaster* is experimentally infected with p elements engineered to transpose in somatic tissue, they display shortened lifespans in comparison to *D. melanogaster* infected with unengineered p elements (Woodruff and Nikitin, 1995). Similarly, in *Caenorhabditis elegans*, the TC1 element, which is somatically active, is associated with aging (Egilmez *et al.,* 1994). In the microorganisms *Saccharomyces cerevisiae* and *Podospora anserina*, self-replicating circular episomes derived from the rDNA repeat and the mitochondrion, respectively are the primary cause of aging in wild-type strains (Hekimi and Guarente, 2003, Dufour *et al.,* 2000).

A possible model for general aging, then, would be somatic activity of transposable elements. Transposable elements have the components required of a disease-related cause of aging: they are ubiquitous in multicellular organisms (Löwer *et al.,* 1996); individuals are born with their transposable elements and a pathogen-free environment will provide no protection; and an Gompertzian exponentially increasing mortality rate would be a least a reasonable expectation.

There is some indirect evidence for an association of transposable elements with aging. Transposable elements would be expected to suppress their mobilization in somatic tissues, since replication there does not spread the element but does risk DNA damage to the host. Indeed, this is the case for some transposable elements (Borie et al., 2002, Woodruff and Nitikin, 1995). Applications of reverse transcriptase inhibitors,

which block transposition of transposable elements that transpose through RNA intermediates, were also shown to extend lifespan in *D. melanogaster* (Driver and Vogrig, 1994), while somatic activity of the mariner element is associated with reduced lifespan in *D. simulans*, although not in *D. melanogaster* (Nitikin and Woodruff, 1995). In addition, transposable element activity increases the rate of chromosomal rearrangement and double-strand breaks (Raskina *et al.,* 2004), both of which increase with aging in humans, mice, and yeast (Vorobtsva *et al.* 2001,Dollé and Vijg, 2001, McMurry and Gottschling 2003).

Many transposable elements are unaffected by aging or do not increase overall with aging (Lund et al., 2002, Gaubatz and Cutler, 1990, Filatov *et al.,* 1997). The difficulty with testing a transposable element-driven model of aging is that it is not necessary for all or even many elements to be involved in the process. A few elements in each lineage could potentially drive the aging process. However, testing the model requires looking at all transposable element activity, a daunting prospect with current technology.

Cancer is very widely spread throughout the higher vertebrate lineage although it is not commonly observed outside of higher vertebrates. The evolution of cancer has not been theoretically studied to the extent that aging has been. As a result, evolutionary models for cancer are less mathematically precise and not subject to tests like the Gompertz curve for aging, although cancer rates frequently increase rapidly with age in a quasi-exponential manner (Riggs, 1994). However, cancer rates, like aging rates, appear subject to evolution in that large animals have much lower rates of cancer per cell than smaller animals. Given the high cancer-induced mortality observed in outbred mice in

captivity (Anisomov *et al.*, 2004), this raises the question of why the smaller animals have not evolved more of the cancer defense of large animals since cancer causes a nontrivial fitness loss.

An association between cancer and transposable element activity is better demonstrated than with that between aging and transposable element activity. Although transposable element activity generally cannot be demonstrated in normal human somatic lineages, it is readily demonstrated in many cancer cell lineages (Wang-Johanning *et al.*, 2003, Yi *et al.*, 2004, Florl *et al.* 1999, Löwer *et al.*, 1996, Woodcock *et al.* 1996). Cancer cell lineages also almost always show active chromosomal rearrangement to a far greater extent than normal cell lineages (Tucker and Preston, 1996). Notably, the active chromosomal rearrangements often precede development of frank cancer (Daisonville and Bailly, 1998; Rennstan *et al.*, 2001, Domon-Dell *et al.*, 2003), suggesting a causative role for transposable elements. However, there is no good evidence addressing the presence or directionality of causation. Some viruses related to the transposons can directly induce cancers but this process does not explain the vast majority of cancers.

Somatic activity of transposable elements could potentially explain the limitation of cancer rates to low but not trivial rates in animal species. Somatic cancer would be a fitness cost for parasitic transposable elements as well as for their host. High rates of cancer would thus favor less active transposable elements, but nearly inactive elements would be disfavored due to slowed transposition in the germline.

While an attractive hypothesis, transposable element mediated limitations on somatic survival has been difficult to test. The problem is the combination as a large number of

candidate elements, plus with the possibility of even more undescribed ones, plus the large number of potential targets. If only a few of the transposable elements in a species had exponentially increasing somatic transposition with age, that would produce an aging phenotype even if most elements did not transpose somatically. Likewise cancer could be caused by cancer-promoting mutations in single cells. The ideal test would be multiple complete DNA sequences of multiple cell lineages within individuals but this is obviously impractical.

As an intermediate step, whole-genome sequencing information from highly repetitive regions could provide limited information on somatic transposable element activity. The multiple repetitive locations substitute for single-copy elements in multiple locations. Two pieces of information could be extracted from such data. First, active elements could be identified as sequences attached to parts of the repeat. Second, sequences subject to insertion/deletion cycles could be identified as regions more likely to vary within species or to changes between species.

Potentially active elements and target regions by themselves would not provide strong evidence for the involvement of transposable elements in aging. However, they would provide more specific hypotheses for more rigorous testing. As mentioned above, the greatest difficulty with testing for a causative association between transposable elements and deleterious processes like aging and cancer is the extremely large number of possibilities to test. If specific elements and specific target regions are identified, then testing for activity requires a much smaller set of sequencing efforts to determine where

in the genome the specific elements are located and which target regions are having insertion and deletion events.

Developing techniques for extracting the desired sequence information from whole-genome sequences requires the choice of an initial target. There are two choices: which repetitive region to sequence and which species to study. The choice needs to be made to increase the chance of an effective test of somatic DNA transposition models. Additionally, properly chosen targets may allow the test of other hypotheses about repetitive DNA other than somatic DNA transposition models, hypotheses which have not been well-tested due to the lack of available sequence information for these regions.

Which repetitive region should we obtain sequence information for? There are several highly repetitive regions that could be sequenced, including the ribosomal repeat, the centromeres, dispersed satellites, and the telomeric repeats. The most useful region for the purpose of testing for an association between transposable elements and cancer or aging would have several features:

1) Known somatic variability, indicating a possibility of somatic transposable element activity.

2) An association with aging or cancer in at least some model systems.

3) Wide phylogenetic conservation for interspecies comparisons.

The ribosomal repeat displays all of these features. Ribosomal repeat lengths are known to vary somatically in dogs (Strehler and Johnson, 1972) and flies (de Cicco and Glover, 1983) and to be subject to rearrangements in humans (Caburet *et al.*, 2005). A class of transposable element in arthropods is restricted to replication within the rDNA

repeat and reaches high frequencies in some species (Averbeck and Eickbush, 2005). Episomes in the ribosomal repeat are a causative agent of aging in the yeast *Saccharomyces cerevisiae* (Hekimi and Guarente, 2003). And finally, parts of the ribosomal repeat are conserved well enough to permit phylogenies spanning the entire tree of life (Woese, 2000).

Which species to target for sequencing? The proposed sequencing method requires the use of whole-genome shotgun (wgs) sequence, limiting the choice of target species to ones with wgs information available. In addition, comparing species consensus sequences to detect regions of sequence subject to transposable element targeting requires the sequences be aligned with minimal ambiguity, suggesting that the initial target species should be closely related. Fortunately, humans and common chimpanzees both have sequencing projects partly based on wgs information. This pairing has the additional potential benefit of medical applications from the human sequence.

The chosen ribosomal DNA repeat target, in the human and chimpanzee species, also permits testing a number of other scientific hypotheses. First, the intergenic spacers within the rDNA transcribed region are known to be hypervariable. However, it is not know to what extent the hypervariability results from high mutation rates and to what extent it results from extended sojourn times. In particular it is not well known to what extent rDNA variants can be shared between species.

Second, a recent paper by Ganley and Kobayashi (2007) found tandem rDNA arrays in certain fungal species had low variability within single arrays. However, the species they choose were selected for high inbreeding and single arrays and it is not known

whether their results generalize to other species. Humans and chimps, as outbred species with multiple arrays, allow this test.

Third, Caburet *et al.* (2005) recently found evidence of active human somatic rearrangements within the rDNA repeat. The evidence consisted of variable spacing and arrangements of fluorescently tagged sequences within the repeat. Sequence information allows a precise test of this finding.

Fourth, rDNA sequences are normally studied only with sample sequences. It is not known how well a single sample sequence will represent a population of rDNA repeats within a species. Multiple reads from different repeats may permit estimates of this variability.

Fifth, sequence information for two closely related species permits analysis of the basis of changes between the two species. The relative importance of mutation and recombination could possibly be determined and important determinants of mutational and evolutional rates identified.

This dissertation will accordingly derive information on the ribosomal rDNA repeat sequence from the human and chimpanzee sequencing projects, and use that information to attempt to identify transposable elements active within the ribosomal repeats and target sequences for transposable elements within the repeat.

The chapters in this dissertation are slightly adapted from individual papers. There are three chapters. The first describes the techniques used to determine consensus and variability information for the rDNA repeat from whole genome shotgun information. The second chapter describes the application of that procedure to the human ribosomal

repeat and analyzes the result to estimate ribosomal variability and validate the technique. The third chapter applies the techniques to the common chimpanzee ribosomal repeat and compares the two repeats to estimate lifespan of ribosomal variants and selective forces operating on the repeat.

Averback, K.T. and Eickbush, T.H. 2005. Monitoring the mode and tempo of concerted evolution in the *Drosophila melanogaster* rDNA locus 2005. *Genetics* 105:171:1837-1846

Arkhipova, Irina; Meselson, Matthew. 2000. Transposable elements in sexual and ancient asexual taxa. PNAS **97** (26) : 14473-14477

Borie, N, Maisonhaute, C, Sarrazin, S, Loevenbruck, c, and Biémont, c. 2002. Tissue-specificty of 412 retrotransposon expression in *Drosophila simulans* and *D. melanogaster*. Heredity **89**:247-252

Caburet, S., Conti, C., Schurra, C., Lebofsky, R.. Edlestein S.J., and Bnesimon, A. 2005. Human ribosomal RNA gene arrays display a broad range of palindromic structures *Genome Res.* **15**: 1079-1085

Charlesworth, Brian. 1994. *Evolution in Age-structured Populations.* Cambridge: Cambridge University Press.

Charlesworth, Brian. 2001 Patterns of Age-specific Means and Genetic Variances of Mortality Rates Predicted by the Mutation-Accumulation Theory of Ageing. J. Theor. Biol. **210**:47-65

Cochran, Gregory M.; Ewald, Paul W.; Cochran, Kyle D. 2001. Infectious causation of disease: An evolutionary perspective. Perspectives in Biology and Medicine **43** (3) : 406-44

Coleman G L; Barthold S W; Osbaldiston G W; Foster S J; Jonas A M. 1977 Pathological Changes During Aging In Barrier Reared Fischer 344 Male Rats Journals of Gerontology **32** (3) : 258-278

Dassonneville, Laurent; Bailly, Christian. 1998 Chromosomal translocations and secondary leukemias induced by topoisomerase II inhibitors. Bulletin du Cancer (Paris) **85** (3) : 254-261

de Cicco, D.V. and Glover, D.M. 1983. Amplification of rDNA and type I sequences in drosophila males deficient in rDNA. *Cell* **32**:1217-1225

Dollé, Martijn E. T. and Vijg, Jan. 2002. Genome Dynamics in Aging Mice. Genome Research **12**:1732-1738

Domon-Dell, Claire; Schneider, Anne; Moucadel, Virginie; Guerin, Eric; Guenot, Dominique; Aguillon, Sarah; Duluc, Isabelle; Martin, Elisabeth; Iovanna, Juan; Launay, Jean-Francois; Duclos, Bernard; Chenard, Marie-Pierre; Meyer, Christian; Oudet, Pierre; Kedinger, Michele; Gaub, Marie-Pierre; Freund, Jean-Noel. 2003. Cdx1 homeobox gene during human colon cancer progression. Oncogene **22** (39) : 7913-7921

Driver, Christopher J. I.; Vogrig, Darren J. 1994. Apparent retardation of aging in Drosophila melanogaster by inhibitors of reverse transcriptase. Annals of the New York Academy of Sciences; Pharmacology of aging processes: Methods of assessment and potential interventions : 189-197

Dufour, Eric, Boulay, Joceline, Rincheval, Vincent, and Sainsard-Chanet, Annie. 2000 A causal link between respiration and senescence in Podospra anserine. PNAS **97**(8):4138-4143

Egilmez, Nejat K. and Schookly-Reis, Robert J. 1994. Age-dependent somatic excision of transposable element Tc1 in *Caenorhabditis elegans*. Mutation Research **316**:17-24

Filatov, D.A., Morozova, T.V., and Pasyukova, E.G. 1998. Age dependence of the copia transposition rate is positively associated with copia transcript abundance in a *Drosophila melanogaster* isogenic line. Mol. Gen. Genet. **258**:646-654

Florl, A.R., Löwer R., Schmitz-Dräger, B.J., and Schulz W. A. 1999 DNA methylation and expression of LINE-1and HERV-K provirus sequences in urothelial and renal cell carcinomas. British J. of Cancer **80**(9) :1312-1321

Gaubatz, james W. and Cutler, Richard G. 1990 Mouse Satellite DNA Is Transcribed in Senescent Cardiac Muscle. J. Biol. Chem. **265**(29):17753-17758

Hekimi, Siegfried and Guarente, Leonard. 2003. Genetics and the Specificity of the Aging Process Science **299**:1351-1354

Liu, Ruiqin; Moroi, Masao; Yamamoto, Masato; Kubota, Tetsuya; Ono, Tsuyoshi; Funatsu, Atsushi; Komatsu, Hiroki; Tsuji, Takahiro; Hara, Hisao; Hara, Hidehiko; Nakamura, Masato; Hirai, Hironori; Yamaguchi, Tetsu. 2006. Presence and severity

of Chlamydia pneumoniae and Cytomegalovirus infection in coronary plaques are associated with acute coronary syndrome International Heart Journal **47** (4) : 511-519

Löwer, Roswitha, Löwer, Johannes, and Kurth, Rienhard. 1996. The viruses in all of us: Characteristics and biological significant of human endogenous retrovirus sequences. PNAS **93**:5177-5184

Lund, James, Tedesco Patricia, Duke, Kyle, Wang, John, Kim, Stuart K. and Johnson, Thomas E. 2002. Transcriptional Profile of Aging in *C. elegans.* Current Biology **12**:1566-1573

McMurray, Michael A. and Gottschling, Daniel E. 2003. An Age-Induced Switch to a Hyper-Recombinational State. Science **301**:1908-1911

Nikitin, A. G.; Woodruff, R. C. 1995 Somatic movement of the mariner transposable element and lifespan of Drosophila species, Mutation Research **338** (1-6) : 43-49

Raskina, Olga, Belyayev, Alexande, and Nevo, Eviatar. 2004. Activity of the *En/Spm*-like transposon in meiosis as a base for chromosome repatterning in a small, islated, peripheral population of *Aegilops speltoides* Tausch. Chromosome Research **12**:153-161

Rennstam, Karin; Baldetorp, Bo; Kytola, Soili; Tanner, Minna; Isola, Jorma. 2001. Chromosomal rearrangements and oncogene amplification precede aneuploidization in the genetic evolution of breast cancer. Cancer Research **61** (3) : 1214-1219

Riggs, Jack E. 1994. Carcinogenesis, Genetic Instability and Genomic Entropy: Insight Derived from Malignant Brain Tumor Age Specific mortality Rate Dynamics. J. Theor. Biol. **170**:331-338

Strehler, B and Johnson, R. 1972. 30 percent decrease in ribosomal DNA dosage during aging of dog brain. *Federation Proceedings* **31**:910

Suematsa, Kazumi and Kohno, Minoru. 1999. Age invariant of Gompertz Function and Exponential Decay of Populations Commensuration with CLOV Experiments. J. Theor. Bio. **201**:231-231

Sun, Feng-Jie; Fleurdepine, Sophie; Bousquet-Antonelli, Cecile; Caetano-Anolles, Gustavo; Deragon, Jean-Marc. 2007. Common evolutionary trends for SINE RNA structures. Trends in Genetics **23** (1) : 26-33

Tucker, J.D. and Preston, R.J. 1996. Chromosome aberrations, micronuclei, aneuploidy, sister chromatid exchanges, and cancer risk assessment. Mutat. Res. **385**:147-159

Vorobtsoba, Irena, Semenov, Alexey, Timofeyeva, Natalia, Kanayeva, All, and Zvereva,Irena. 2001. An investigation of the age-dependency of chromosome abnormalities in human population exposed to low-dose ionizing radiation. Mech. Aging Devel. **122**:1373-1383

Wang-Johanning, Feng, Frost, Andr R, Jian Bixi, Epp, Lidia, Lu Danielle W, and Johannin Gary L. 2003. Quantitation of HERV-K env gene expression and splicing in human breast cancer. Oncogene **22**: 1528-1535

Woese, C. 2000. Interpreting the universal phylogenetic tree. Proceedings of the National Academy of Sciences of the United States of America 97 (15): 8392-8396.

Woodcock, D. M., Willamson, M.R., and Doherty, J.P. 1996. A Sensitive RNase Protection Assay to Detect Transcrpts from Potentially Functional Human Endogenous L1 Retrotransposons. Biochem. Biophys. Res. Comm. **222**:460-465

Woodruff, R. C.; Nikitin, A. G. 1995 P DNA element movement in somatic cells reduces lifespan in Drosophila melanogaster: Evidence in support of the somatic mutation theory of aging. Mutation Research **338** (1-6) : 35-42

Yi, Joo-Mi, Kim, Tae-Hyun, Huh, Jae-Won, Park, Kyoung sun, jang, Se Bok, Kim, Hwan-Mook, Kim, Heui-Soo. 2004. Human endogenous retroviral elements belonging to the HERV-S familiar from human tissues, cancer cells, and primates: expression, structure, phylogeny and evolution Gene **342**:283-292

Chapter 1: Techniques to Obtain Sequence Information for Highly Repetitive Regions Directly from Whole-Genome Shotgun Reads

Abstract: Current genome sequencing projects do not attempt to produce sequences for highly repetitive regions. Although these regions can be of considerable biological importance, standard assembly techniques cannot deal with the ambiguities in such regions. Methods have been proposed to assemble such regions through graph networks but these methods are very computationally demanding and have not yet been applied to ongoing sequencing projects. Here I describe a new method for obtaining sequence information from highly repetitive regions based on an iterative approach. An initial draft, which can be low-fidelity and incomplete, is refined repeatedly until no further improvements can be made. Short sections with homology to known sequence or a single trace read will suffice for an initial draft. Specific techniques are used to close gaps and to resolve internal subrepeats. The technique has been successfully applied to the chimpanzee and human rDNA repeats; other rDNA repeats are ongoing.

Introduction

Genome sequencing has become one of the most important tools for biological research. Comparison to known sequences permits rapid formulation and testing of a wide variety of evolutionary and mechanistic hypotheses. Hundreds of species now have

completed or ongoing sequencing projects (Windsor and Mitchell-Olds 2006) and the number is growing rapidly with continuing decreases in sequencing costs (Hall 2007).

Current sequencing methodology, however, faces limitations. One of the more significant is being largely unable to assemble highly repetitive, or heterochromatic, DNA regions into complete, verified sequences (Hoskins et al., 2007). Assembly of the raw sequencing reads into finished sequences requires unambiguous alignments of the reads, which is practically, and sometimes even theoretically, impossible in highly repetitive regions as any particular read may align to any of many highly similar or even identical regions. A variety of specific finishing techniques can produce assembled sequences of some of these repetitive regions when they are small or more variable (International Human Genome Consortium, 2004; Hoskins *et al.*, 2007), but they are labor-intensive and not always well suited for large sequencing projects. Where practical and worthwhile, they are applied in the finishing stages of a sequencing project. Even so, an estimated 8% of the human genome, the most aggressively finished, remains unsequenced due to these limitations (International Human Genome Consortium, 2004).

Highly repetitive regions often include important biological information desirable for evolutionary and medical reasons (Shaday and Sinclair 2007; Zafiropulos *et al.* 2005; Badaeva *et al.*, 2004;O'Neill *et al.*,, 2004, Mateuca et al. 2006). Some of the most important, and intractable, repetitive regions include: satellite DNAs, where striking changes often associate with species boundaries (Choi *et al.*, 2003, Cuadrado and Jouve, 2002); centromeres, fundamental to cell division; telomeres, important in aging and cancer prevention, and the ribosomal DNA, the template for the most common RNA in

the biosphere. These regions have been ignored with non-genomic site sequencing as well since they are very difficult to sequence with site-specific approaches as well as with genomic approaches. Sequence information on these regions could allow us to address some old hypotheses such as heterochromatin change affecting speciation (White, 1978) and ribosomal variability affecting aging (Strehler and Chang, 1979, Strehler et al. 1979).

There are two broad techniques for basic genome sequencing: hierarchical shotgun and whole-genome shotgun (Waterston *et al.,* 2002). In hierarchical shotguns, a library of large clones (commonly BACs, typically 150-250 kb long) is generated, their relative positions mapped, and a subset of these clones chosen that covers as much of the genome as possible. Each of these clones is then sequenced and these subsequences tiled according to the predefined map. In whole-genome shotguns a library of comparatively small (4 -50 kb) clones is obtained and the ends of millions of these sequenced. Mathematical algorithms are then used to assemble this large set of sequences into linkage groups. In both cases the resulting base sequence has many gap and errors, which are then corrected, when possible, by finishing each particular problem individually. The choice of shotgun methodology has been quite contentious; in general, hierarchical shotguns generate better quality sequence, requiring less finishing, but require more time and greater expense. Currently whole genome shotguns are becoming more common mostly as they are easier to automate (Yu *et al.,* 2006).

Neither of the two primary genome sequencing methods can determine sequences in highly repetitive regions. Clones for hierarchical sequencing are often generated by restriction enzymes. In repetitive regions, restriction enzymes either cut at least once per

repeat or never at all and so either the resulting clones are to small to be useful or the DNA fragments are too large to clone. Even if clones are obtained, highly repetitive regions prevent unambiguous clone tiling. Clones from these regions are thus not sequenced so sequence information on highly repetitive regions is highly restricted, mostly to the edges where clones intrude into repetitive regions (Yu *et al*, 2006).

Whole-genome shotguns, by contrast, do generate raw sequencing reads in highly repetitive regions. However, correcting errors from individual sequencing reads requires unambiguous alignments of the reads, so that multiple sequences are available for comparison in each region. Likewise assembling the small individual sequencing reads into large complete contigs also requires unambiguous alignments. So current assembly techniques cannot produce long continuous regions of final sequence. Current assemblers consequently simply ignore these highly repetitive tandem regions (Yu *et al.,* 2006).

Theoretical modifications to assemblers that would allow a type of assembly in highly repetitive regions have been proposed (Myers, 2005). However, these ideas require enormous computing resources and even today are not yet practical for a full sequencing project.

However, although a full assembly is not possible, whole-genome sequencing information could be used in a manner analogous to that used for consensus/variability analyses of dispersed repetitive DNA (Khan et al, 2006). This would not produce a complete sequence of the entire repetitive region. However, it could provide a consensus sequence for a repeat unit, the approximate frequency of any particular variation from that consensus in the actual sequence, and possible disequibrium estimate for

polymorphisms. While not as useful as a complete assembled sequence, this would still provide extensive information useful for phylogenies, phylogenetic conservation, and for testing models of the development of repetitive DNA (Ganley and Kobayashi 2007).

Unlike the finished sequences used with the dispersed repetitive DNA, raw sequence reads have substantial error rates and small sizes and this produces a number of technical obstacles. These include: determining even a draft consensus sequence if the repeat region is larger than a typical read; identifying inserts or deletions large in comparison to a read; computationally identifying a best-fit consensus to tens of thousands of compared reads; and determining which read variations result from genuine DNA variation as opposed to sequencing error. Presumably these formidable technical obstacles are the reason such techniques have not been attempted for highly repeated regions.

Here I present iterative stochastic alignment (ISA), a set of techniques to obtain consensus and variability estimates for highly repetitive regions from shotgun reads even when proper assembly is impossible. Briefly, they are: an iterative alignment strategy to generate a consensus even from highly inaccurate or incomplete drafts; an extension process akin to chromosome walking for filling gaps; multiple window sizes for complexly varied regions; mismatch detection for improperly aligned sequences; and topological interpretations of mismatched read ends to identify long indels and terminations of tandem repeats.

Determining the consensus sequence for a set of reads is not directly feasible due to computational complexity. To address this, a hill-climbing technique was used. At any stage there is a draft sequence. Individual reads are compared to the draft to determine

how to refine the draft to better correspond to the reads. This more accurate draft is the refined again, and the process repeated until refinement produces the same sequence as the draft.

The goal of the iterative approach also had to be carefully defined. "Consensus" could have many possible meanings when applied to a large set of varied sequences. As an example, in a region with a varied number of repeats, the consensus number of repeats could be either the mean, median, or mode of the observed number of repeats in each sequence. For the purposes of this work, the consensus was defined as the sequence producing the highest overall alignment score when aligned against all the actual repeats.

The basic strategy of Iterative Stochastic Alignment is a series of processes to move from minimal information to the complete consensus and variability information. The steps are as follows:

1) Choose an initial draft sequence

2) Align all available reads against that sequence

3) Refine the draft sequence to more closely correspond to the reads

2 and 3 are repeated until the draft sequence cannot correspond any better to the reads.

4) Separating read errors from sequence variation

5) Identifying reads bridging the repeat and other sequence

Choosing an initial draft sequence

The initial draft sequence need not be complete, highly accurate, or even cover a large portion of the actual consensus. A 300 base region from a single human rDNA

sequence was a sufficient start for the entire 44 kb chimpanzee rDNA sequence. All that is required is a sequence similar enough to one region of the true consensus that only reads from the repeat will have reasonable alignment scores. Other possible starting seeds are individual reads not aligned to single or low-copy sequence and thus probably part of a repeat, sequence from related species as was done with the chimpanzee, or highly conserved sequence from distant species.

Aligning Reads

All available reads are scanned for meaningful homology to the draft and the higher-quality sections (phrap score above 20) of these reads aligned against the draft. The draft sequence is then broken into small windows, and refined by replacing each window with the sequence best matching all reads in that window. Since it aligns at least as well as the draft to each window, the refined draft should be a better match for the true consensus as well. The refined draft is then refined again, and the process repeated until additional refinements do not change the sequence

Alignments were performed in a two-step process for computational efficiency. Initially the reads are aligned using the same method as for a standard assembler, which is to assume that an exact match longer than a threshold indicated the read aligned to the consensus. The remaining regions were aligned using the detailed techniques below.

For the final sequence of refinements more consistent alignments were desired. First, all very long matches (over 30 bp) were evaluated for the highest score. Second, the resulting alignment was redone as with NW but using only the section of the consensus initially matched.

The primary alignment methodology used was a Needleman-Wunsch (NW) algorithm (Needleman and Wunsch, 1970) with modifications for extended gaps, unknown sequence terminations, and alignment of adjacent regions. NW alignments also produce a score indicating the quality of the alignment. The theoretical NW algorithm scores every possible alignment of two sequences with a bonus for every match and a deduction for every mismatch and the length of every gap and chooses the alignment with the highest score. This corresponds to choosing a maximum-likelihood series of mutations to transform one sequence into the other, with a simple mutation model. The payoff matrix used was a standard for closely related sequences with +1 for every match, -3 for every mismatch, and -5 for every gap. Conveniently, with this payoff matrix a score of n is evaluated as equivalent to an exact match on n bases.

In the basic NW algorithm, each base of a gap produces the same score deduction. This excessively penalizes multibase indels, which can be produced by a single mutation yet are costed as multiple mutations. A standard modifier is to make the cost of the first base of a gap (the "gap open" cost) higher than the cost for each subsequent gap (the "gap extension" cost), producing the gapped Needleman-Wunsch algorithm. For this work a gap extension payoff of -1 was used.

The number of all possible alignments goes up exponentially with sequence length so examining all possible alignments directly to determine the best score is computationally intractable for all but the shortest sequences. However, there is a standard recursive algorithm that makes the problem very tractable. A zero-indexed rectangular x+1 by y+1 matrix is created with x and y the lengths of the two sequences.

The entry in each location (a, b) is the best possible score for the first a bases of the first sequence (the query) aligned with the first b bases of the second sequence (the subject). This matrix can be filled out recursively: the best score for (a+1, b+1) is one of three possibilities: the score for (a,b) plus the match or mismatch payoff for aligning base a+1 of the query with base b+1 of the subject; the score for (a,b+1) plus the gap penalty; or the score for (a+1,b) plus the gap penalty. Once the matrix is filled out the entry at (x+1, y+1) gives the score and the alignment can be extracted by inspecting the matrix at each point to determine which step was made at that point to produce the score.

Algorithmically the gapped Needleman-Wunsch was implemented with 3 tables of pair scores, one for alignments where the last pair of bases was aligned, one for alignments ending with a gap in the query, and one for alignments ending with a gap in the subject. The score in the aligned table at (a+1,b+1) was the best score in any table for (a,b) plus the match/mismatch payoff for that base pairing. The score in the query gapping table was the best of: the best score at (a+1,b) in the aligned or subject gapping tables minus the gap open cost; or the score at (a+1,b) in the query gapping table minus the gap extension cost. The score in the subject gapping table was determined analogously to that for the query gapping table.

The standard NW algorithm is a global algorithm and assumes that the two aligned sequences should align to each other in their entirety. This was manifestly untrue in many alignments here, as any read necessarily matched only a small part of the entire repeat. Initial alignments with global NW often would miss good but imperfect homology in favor of pathological alignments with tiny bits of the read aligned to

scattered regions over the length of the draft consensus. Here I used a partially local algorithm, differing from the standard Smith-Waterman local alignment method. This method simply ignored gap extension costs beyond the end of either sequence. Algorithmically this was implemented by making the gap extension cost zero on the first row and column of the table if the left end of the read was free and zero on the last row and column if the right end was free, and zero on all edges if both ends were free.

The standard NW algorithm implicitly assumes the sequences are embedded in matching surrounding sequences. In the process of these alignments, we sometimes know the alignment of adjacent sequences and sometimes they do not match. If so, a gap extended to that end of those sequences is assigned a gap open penalty inappropriately, since there must be a gap there anyway. In these circumstances the modified algorithm suppresses the gap open penalty along the appropriate edges.

Finally, the standard NW algorithm assumes both sequences are homologous wherever an alignment can be generated. However, due to variable inserts and deletions as well as certain sequencing errors, sometimes read sequences will contain segments not homologous to the RDNA repeat. Observation indicated this was almost entirely restricted to read ends, so an ad hoc correction technique was employed. Each end of the read was scanned inward until a region of 50 bases in the read with at least 47 exactly matched bases was reached. Any bases closer to the end than this region were considered non-homologous "overhangs" and were removed from the alignment for the purposes of consensus determination. Overhangs were used to identify large indels in sequence variants.

Refining drafts

Refining has two different methods. One is a extension method to fill in gaps in the sequence. The other is a window-based method to determine the sequence most closely matching the reads in a set of small windows covering the draft sequence. Chronologically the extension process is often used first but the window process is described first as the extension process was based on the window process.

The idea of the window process is that in a small region of the draft it is relatively easy to determine what possible sequence best matches all the traces. Generally windows of one base were used for the refinement process. Each observed sequence in the window was aligned against all other sequences observed in the window and scored with its alignment score against each other sequence weighted by the other sequence's frequency. For single base windows, this was almost always equivalent to simply choosing the most common observed sequence.

To illustrate, suppose a draft of a piece of the sequence was the sequence ctcggcgcctctg. The sequence is aligned against a group of reads as follows

```
Original        ctcggcg-cctctg
Reads           ctcggcg-cctctg
                ctcggcgccctatg
                ctcggcgccctatg
                ctcggcgccctatg
                ctcggcgccctatg
                ctcggcg-cctctg

Result          ctcggcgccctatg
```

Most of the bases are perfectly matched to all the reads. The third g, however, matched to g twice and gc 4 times. Since gc was most common of all the matches, gc replaces the g in the next iteration. Likewise, a replaces the last c for the same reasons.

Single base windows were almost always adequate but can fail when faced by multiple alternatives, none of which are the majority of sequences. As an example, consider the following alignment:

```
Original        ggtttgg
Reads           ggcttgg
                ggcttgg
                ggtctgg
                ggtctgg
                ggttcgg
                ggttcgg
                ggttcgg


Result          ggtttgg
```

All 3 t's in the result match the majority of bases in the corresponding location in the reads, but the resulting ttt occurs in none of the reads. Here the best read match to ttt is ttc and the resulting refinement should be ggttcgg. This situation might seem pathological in a family of such closely related sequences, but did occur once in the chimpanzee rDNA repeat in a sequence of three successive variable microsatellites.

The solution is to match to sections longer than one base. With the chimpanzee sequence, 23 base windows were used. The last few remaining bases were put into a separate window. There are 23 possible frames depending on where the first window starts and each of these frames was used to generate a new refinement. These 23 refinements were then treated as 23 different sequences; each was aligned against the previous draft and then a new refinement produced with single base windows. Applied to the above example, this procedure results in the desired ggttcgg refinement.

Extending drafts

When draft sequences were incomplete, missing areas were filled before window refinements of the entire sequence began. This was a practical decision. Theoretically filling in missing areas could be done at the same time as refinement of existing sequence. However, refining a full draft into a final version proved a relatively short process, requiring only a few dozen iterations to complete. Filling in gaps, however, was rather slow and determining an entire 44 kb repeat from a small starting region required thousands of refinements. So, as a practical matter, a small region at the end of the known sequence was taken and extended outward. Once a stable extension was obtained, a region at the end of the now-extended sequence was taken and extended again. The process was repeated until known sequence was reached (since these are consensuses of tandem repeats, they behave much like circular DNA).

Extension was performed by using reads aligning to the piece at the end of the draft that extended beyond the end. The most common nucleotide one base from the end of the draft sequence was taken as the first extended base in the next refinement, the most

common second base was taken as the second extended base, etc. In effect, this was treating the missing bases as unknown and breaking the missing range into single-base windows. As an example, consider the following alignment:

```
Original        ctcggcg-cct
Reads           ctcggcg-cctctgccg
                ctcggcgccctatgcg
                ctcggcgccctatgccga
                ctcggcgccctatg
                ctcggcgccctatgc
                ctcggcg-cctctg

Result          ctcggcgccctatgccga
```

The result was then realigned against the same reads for another round of refinement, generally with many fewer overhangs, and the process repeated until the alignment was stable. The result of the process was a sequence which best matched the reads originally matched to the end of the alignment.

It was then necessary to determine what in the final refined region extended the original sequence and how much was accurate enough to add to the existing draft. Initially the final refined extended piece was simply aligned against the existing draft and the part extending beyond the existing draft considered the extension. However, in the presence of quasisatellite repeats this alignment could be ambiguous and this became a problem in the SalI-containing repeats after the end of the transcribed pre-rRNA sequence. To solve this, the positions of the original draft ends were tracked through the entire sequence of extending refinements. The sections beyond where the original ends mapped to were added to the draft for the next round of sequencing.

Since the read sequences contain many errors, particularly towards the ends, it was expected that this procedure would be slow and problematic. After only a few dozen

bases, most reads would have experienced an indel and so the nth base out in one read would not correspond to the nth base in most others. The consensus thus would not be meaningful and only the first few dozen bases of this extension would be useful. The refinement process would need many iterations, each adding only a few bases.

Surprisingly, the extension process proved very efficient, reaching an effective steady state after only 2 or 3 refinements. In almost all cases, there would be small regions with high concentrations of one nucleotide. This would produce a run of that nucleotide in the initial extension. In the next refinement, the appropriate section in each read would align to that run, causing the read extensions to be mostly correctly placed against each other over entire sequence. Refinement would then very quickly generate a stable consensus where possible..

The number of sequences used for a step in the extension process had to be chosen carefully. If the number of sequences used substantially exceeded the number of actual high-quality matches then the matches would include reads for different regions with related sequence. The final refinement could end up representing the other regions rather than the original search region.

On the other hand, high-quality usable regions of sequences are comparatively short, averaging less than 500 bp. When a sequence set is then aligned against the search sequence, the number of sequences drops off rapidly with distance from the end of the search sequence. Hence a large number of sequences would be expected in increase the quality of the alignment at a distance from the search sequence. Trial and error indicated

a limit of 100 sequences prevented this condition and allowed the extension to proceed accurately.

The trace database was expected to produce an ample number of sequences for these purposes. The chimpanzee genome has 300-400 copies of the rDNA repeat and with 8-fold coverage the naive prediction was that there would be 2400 – 3200 sequences covering any small part of the repeat. However, some regions had far fewer, most notably in the transcribed spacers, which all had regions with fewer than 100 sequences.

The length of additional sequence to extend also required careful selection. Short extensions required more extension steps, and steps were slow. However, as the extended region became longer fewer reads are available as fewer cover the entire desired extension. Further, the error rate increases since reads become less accurate towards the ends. In practice with the chimpanzee sequence 200 base extensions proved practical although this might have to change for other specific examples. In three regions with poor coverage 200 bases was too far. In these cases the ends were too inaccurate and further extensions failed because no reads had adequate homology. These regions were rerun with smaller extension lengths until successfully crossed. In one case, the chimpanzee second transcribed spacer region, the extension length had to be reduced to 25.

A potential concern with stepwise extension is that it could become confused in highly repetitive regions. The 4 giant microsatellite complexes present in the human sequence and (correctly) expected in the chimpanzee sequence provided a stringent test

of performance. All 4 were present, but somewhat surprisingly, the extension process went right through all 4 without difficulty.

Sequence confirmation

Two criteria were applied to confirm a sequence. First was that all topological alterations suggesting inserts or deletions had to be uncommon, less that 5% of reads at their location or only one read, or with the cause of the topological alteration identified as minority indel or bridge sequence. Second was that all bases in the sequence had to have multiple reads matching the consensus in a 40-base region including that base.

Identifying variability sources

A Poisson test distinguished read variations resulting from actual sequence variation from read sequencing errors. The cutoff parameters were a p value of 0.001 given a 0.5% rate of the specific sequence error, with the number of samples equal to the number of reads observed at that site. The sequencing error rate of reads is approximately 2% and this was arbitrarily divided into 4 categories: substitutions to any of the other 3 bases, or any form of indel. The rate of any particular error is therefore at most 0.5%. The p value was chosen as a compromise between type 1 and type II errors. Given that the read database is a limited sample of all sequences, type II errors are inevitable, especially for rarer variations. At the same time with over 43,000 base pairs even a moderate p value could produce many spurious variations. Hence the p value was set relatively low to minimize type I errors; one type I error is expected per 1000 bases. However, even at this level the type II error rate is respectable. A Monte Carlo simulation for the human

sequence indicated the procedure would accurately identify 47% of sequence variations present in only 1% of the sequences.

Identifying bridging sequences

At the ends of repeated regions or at large DNA inserts some reads will properly match to the consensus in some regions but will not be properly matched in others. A number of artifacts can also produce traces that only partially match to the consensus, notably vector artifacts. These mismatch regions needed to be removed from the alignments during refinements since they would imply spurious variation from the consensus sequence. In the final step they can be used to indicate the beginning or end of repetitive regions, inserts into the tandem repeats, or large rearrangements. During the extension process they were acceptable since variation was not analyzed at that stage and as long are they were a relatively small proportion of the sequences they would not interfere with the process.

The method used was to scan from each end of each aligned read until a region of 50 bases with at most 3 differences from the draft consensus was reached. Sequence beyond that region was removed from the alignment and indicated as an overhang at the consensus location.

As an illustration, consider the following hypothetical match, with the assumption that the match continues perfectly to the left:

```
ctgcgggcccgaggaggggcggctggcgtgtgggggagtgtgacccaccctcggtgagaagccttct
ctgcgggcccgaggaggggcggctggcgtgttgggggagtctcccccccccccggtgagaaggcttct
```

Scanning the alignment from the right reveals 47 exact matches in 50 bases to the left of the second leftmost mismatched c. The mismatched portion of the draft is excised from the alignment, leaving

```
ctgcgggcccgaggaggggcggctggcgtgtgggggagtgtg-------------------------
ctgcgggcccgaggaggggcggctggcgtgttggggagtctcgccccccccccggtgagaaggcttct
```

and indicating the read sequence after the tatg should be treated as an insert.

Topological analysis

The rDNA repeat contains large-scale duplications. In the presence of these, stepwise extension can potentially skip sections or go into spurious endless loops. To illustrate, consider a sequence with the structure ABACD. An extension rightward from A could go to either B or C legitimately. If extensions from the first A go to C, the sequence produced will be ACD, with the BA deleted. If extensions from the second A go to B, the extension falls into a loop of ABABAB etc. Such errors from the extension process must be recognized and corrected.

The alignment process cannot directly identify inserts longer than a typical read. Short inserts produce an unaligned regions flanked by aligned regions. If the insert is longer than the read there will be only one flanking aligned region. The insert in the read will either be misaligned to the consensus or identified as unaligned, depending on how different the sequences are.

Suppose by the process above we obtain ACD as a draft for an actual sequence of ABACD. To identify the insert, we obtain all sequences that match the draft and align them by standard procedures. Reads from the AB region will be included due to their

strong homology to A. When they are aligned, the A part of the read will align to A in ACD. However, the B part of the read will not normally match any part of the draft and will be misaligned. Since we are assuming B is long (otherwise it would be identified as an insert) the misalignment will extend to the end of the sequence. Using the misalignment procedure, this section will be tagged as an overhang.

A single overhang indicates little, since the high-error tails of most reads produce overhangs by this method. However, a large number of overhangs at the same base with very similar sequence suggests some kind of large-scale sequence alteration in at least a large fraction of the actual repeats. The overhang can then be matched against other regions in the consensus to identify large indels and reversals, or against the NR database to determine foreign sequence insertions. Since short sequences are difficult to identify unambiguously, overhangs shorter than 40 bases were ignored for topological analysis. In this work overhangs found in 5% of locally aligned reads and in least two reads, were considered to signal a possible variation.

Vector inserts proved a significant source of artifactual overhangs. Overhang sequences frequently aligned to known vector sequence. All identified overhangs were therefore blasted against the NCBI nr sequence database with default parameters (word size of 11 and chance expectation of 10 sequences). If vector sequences appeared with a p value below 0.01 the overhangs was ignored as spurious.

If the process indicates there was an insert omitted from the consensus, the missing insert must be determined and inserted. First we obtain a piece of the altered section by aligning all the overhangs as if they were overhangs for the first round of a sequence

extension. This piece is then extended in both directions until a long region (600 bp) is reached with an exact match to the existing draft. This indicates we have extended beyond the insert back onto known sequence. The insert variation is then examined manually to determine how it fits into the sequence. If the entire variation exactly matches known sequence we have either a deletion in a minority of sequences or an exact duplication in an unknown number of sequences. Match counts can be examined to determine whether an insert is universal or minority – doubling over the region indicates universal, little change indicates minority. Variations matching the sequence in reverse indicate a minority reversal. Variations with ends matching to nearby or reversed areas on the draft and a middle which does not align indicate a true insert, which can be pasted in with the exact matches as guides.

In the above ABACD/ ACD example, the mismatched region will be the first part of B. Extending that yields ABA, at which point extension terminates since both A regions match the draft. Aligning the ends to ACD, we will find the left end aligns to the right end of A and the right end aligns to the left end of A. This indicates an insert of BA and so we recover the true sequence, ABACD.

Conclusions and future directions

The techniques presented are readily applicable to rDNA sequences for almost any eukaryote with a whole genome shotgun sequencing project. The 18S and 28S RNAs have some regions so well conserved that a similarity search on any of those regions would identify reads in any eukaryote containing those reads. One single read is an

adequate start for the entire extension and refinement process leading to a complete consensus sequence with variability.

One result of the initial application to the great ape rDNA sequences has been to validate previous reports that the rDNA repeat contains hypervariable regions in combination with hyperconserved regions. The hypervariability is associated with strikingly high intragenomic variability present even within the hyperconserved areas. Incomplete lineage sorting, proposed as a cause of intragenomic variability, can be ruled out since intragenomic variability is nonetheless lower than intergenomic variability, even in nonconserved sequence. This suggests elevated mutational pressure on the ribosomal repeat. How sequence conservation can be maintained in the face of such high mutational pressure and the presence of high percentages of presumably partially defective sequences is an intriguing question.

The techniques could also be used to detect variability in satellite DNA sequence associated with speciation. The appearance of novel sequences in large numbers remains a puzzle. With quantitative variability information, the origin and spread of new satellite variants can be examined in unprecedented detail.

In conclusion, this new approach to sequence information permits a wide variety of promising approaches to old, difficult, and neglected problems in biology. ISA generated a stable consensus for the Pan troglodytes ribosomal RNA repeat, one of the most challenging repetitive DNAs due to its 44 kb length and high variability. The genome sequencing projects turn out to provide even more useful information than had previously been anticipated. Even when the original goal of the sequencing project cannot be

obtained – a verified chromosomal sequence – large amounts of valuable information can be extracted for a variety of scientific questions.

References

Badaeva, E.D.; Amosova, A.V.; Samatadze, T. E.; Zoshchuk, S. A; Shostak, N.G.; Chikida, N. N; Zelenin, A.V.; Raupp, W. J.; Friebe, B.; and B. S. Gill. 2004. Genome differentiation in Aegilops. 4. Evolution of the U-genome cluster *Plant Syst. Evol.* **246**: 45–76

Cuadrado A. and Jouve N. 2002. Evolutionary Trends of Different Repetitive DNA Sequences During Speciation in the Genus Secale *Journal of Heredity* **93**(5) 339-345

Hall N. 2007. Advanced sequencing technologies and their wider impact in microbiology. *Journal Of Experimental Biology* **210** (9): 1518-1525

Hoskins, R. A.; Carlson, J. W.; Kennedy, C; Acevedo, D; Evans-Holm, M.; Frise, E.; Wan, K.H.; Park, S.; Mendez-Lago, M.; Rossi, F.; Villasante, A.; Dimitri, P. Karpen, G. H.; and Celniker, S. E. 2007. Sequence Finishing and Mapping of Drosophila melanogaster Heterochromatin. *Science* **316**: 1625-1628

International Human Genome Sequencing Consortium. Finishing the Euchromatic Sequence of the Human Genome. 2004. *Nature* **431**: 931-945

Khan, Hameed; Smit; Arian; and Boissinot; Stéphane. 2006. Molecular Evoluation and Tempo of Amplification of Retrotransposons Since the Origin of Primates *Genome Res.* **16**:78-87

Mateuca, R;Lombaert, N; Aka, P.V; Decordier, I.; Kirsch-Volders M. Chromosomal changes: induction, detection methods and applicability in human biomonitoring. 2006. *Biochimie* **88**: 1515–1531

Myers, Gene. 2005. Building Fragment Assembly String Graphs. *Bioinformatics* p.1-7

O'Neill, R. J.; Eldridge, M. D. B. ; and Metcalfe, C. J. Centromere 2005. Dynamics and Chromosome Evolution in Marsupials *Journal of Heredity*:**95**(5):375–381 2004

Michan, Shaday and Sinclair, David. 2007. Sirtuins in mammals: insights into their biological function *Biochemical Journal* **404** (1):1-13

Needleman, S. B. and Wunsch, C. D. A. 1970. General Method Applicable to the Search for Similarities in the Amino-Acid Sequence of 2 Proteins. *J. Mol. Biol.* **48**:443-53

Strehler, B. L., & Chang, M. P. 1979. Loss of hybridizable ribosomal DNA from human post-mitotic tissues during aging. II.Age-dependent loss in human cerebral cortex—hippocampal and somatosensory cortex comparison. *Mechanisms of Ageing and Development*, **11**, 379–382.

Strehler, B. L., Chang, M. P., & Johnson, L. K. 1979. Loss of hybridizable ribosomal DNA from human post-mitotic tissues during aging. I. Age-dependent loss in human myocardium. *Mechanisms of Ageing and Development*, **11**, 371–378.

Waterston, Robert H.; Lander, Eric S.; Sulston, John E. 2002. On the sequencing of the human genome *PNAS* **99** (6):3712-3716

Windsor Aaron J and Mitchell-Olds, Thomas. 2006. Comparative genomics as a tool for gene discovery. *Current Opinion in Biotechnology* **17**:161-167

Yu, Jun; Ni, Peixiang; Wong, Gane Ka-Shu. 2006. Comparing the whole-genome-shotgun and map-based sequences of the rice genome. *Trends in Plant Science* **11** (8) : 387-391

Zafiropoulos, A.; Tsentelierou, E.; Linardakis, M., Kafatos, A. and Spandidos D.A., 2005. Preferential loss of 5S and 28S rDNA genes in human adipose tissue during ageing *International Journal of Biochemistry & Cell Biology* **37** (2) : 409-415

Chapter 2: A Consensus Sequence for the Human Ribosomal DNA repeat

Abstract: The ribosomal DNA (rDNA) repeat has been hypothesized to play roles in aging and speciation, and has potential applications in phylogenies of closely related organisms. These applications, however, require consensus and variability data among the rDNA repeats, which have so far been stringently limited by technical obstacles. Here I apply an iterative stochastic alignment methodology (ISA) to the shotgun trace database for the human genome to generate a consensus sequence for the human rDNA repeat, along with estimates of local variability. Sequence coverage was shown to vary over a hundredfold between regions, suggesting shotgun sequencing has strong biases against certain regions. The consensus repeat is 43,113 bp and differs from the only published sample rDNA repeat by 370 changes, and by 106 changes from a recently available sample sequence on chromosome 22. Polymorphism across repeats is high compared to population variation of previously sequenced regions of the human genome, with 3.1% of sites in the intergenic spacer demonstrated to be polymorphic. Indirect evidence indicates repeats within the five chromosomal tandem arrays vary and that recombination between repeats is present but rare. Similar techniques applied to other heterochromatic DNA may allow characterization of these common and biologically significant classes of DNA, currently refractory to genomic analysis.

The ribosomal DNA (rDNA) repeat is essential to life as we know it. In rapidly growing cells, ribosomes may comprise 25% of dry cell mass (*E. coli* Statistics, Institute for Biomolecular Design). Most of this mass is in the ribosomal RNAs (rRNA), which have essential structural and catalytic roles in the ribosome. There are four rRNAs in the ribosome, and in animals three are transcribed from the rDNA repeat. The three rRNAs are derived by cleavage from a single pre-mRNA, which is followed by a spacer region containing transcriptional signals (Mougey *et al.,* 1996) to form the repeat unit.

The rDNA is always found in tandem repeats (arrays) on one or more chromosomes. Tandem copies of the rDNA genes are necessary to produce the large quantities of rRNA needed by many cells. With a protein gene, one DNA sequence allows transcription of many RNAs, each of which allow translation of many proteins, and a single gene can provide the template for a substantial fraction of a cell's mass. However, since the rDNA gene codes for RNA, there is no translation phase for additional multiplication, and a single copy cannot produce the required quantities. Characteristically, in eukaryotes, the rDNA arrays are embedded in other repetitive DNA, commonly pericentromeric. The *Homo sapiens* genome contains approximately 300-400 copies distributed between five arrays on each of the short arms of the five acrocentric chromosomes: 13, 14, 15, 21, and 22, (Henderson *et al.*, 1972). The arrays are located at the edge of the pericentromeric region at the junction with the subtelomeric region (Gonzalez and Sylvester, 2001).

Figure 2.1 presents a schematic for the human rDNA repeat, based on the published Gonzalez and Sylvester sample sequence, appreviated GSss (Gonzalez and Sylvester,

1995). The human pre-mRNA transcript consists of the 18S, 5.8S, and 28S RNAs, plus transcribed spacer regions that are removed during processing. The human intergenic spacer mostly consists of Alu repeats, short microsatellites, and short stretches of sequence of unclear origin, all of which are routinely found in human non-coding DNA. It contains 5 slightly unusual features: 1 threefold imperfect tandem duplication at the end of the transcribed region and 4 large microsatellite complexes.

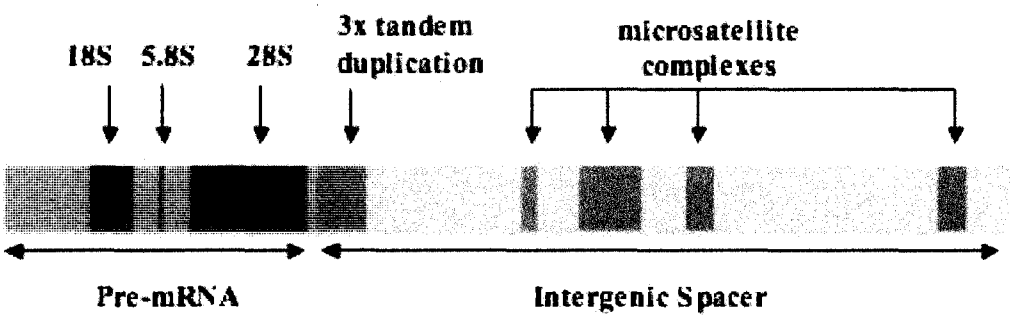

Figure 2.1: Human rDNA Repeat Schematic
Based on the published sample sequence of Gonzalez and Sylvester, 1995. This schematic is typical of mammalian repeats. In other eukaryotes the intergenic spacer is generally short and the composition and arrangement of RNAs in the pre-mRNA may change.

In spite of the rDNA repeat's importance and universality, sequence information for individual complete repeating units is limited, and consensus sequences and variability data almost completely absent. Sample sequences from individual repeats are available for some species, but these provide no variability information, and it is not possible to determine how closely they resemble typical repeats from that species. Sequence variability information is available in many species for the internal transcribed spacers, but these are only a few hundred bases long and may or may not be representative of the

entire repeat. The only overall sequence variability information is from 5 fungal species with single rDNA arrays; in those five species the arrays are essentially monomorphic within individuals (Ganley and Kobayashi, 2006). In species with longer intergenic spacers such as mammals, even sample sequences are often unavailable, possibly due to technical difficulties resulting from large microsatellites and complex interspersed repetitive DNA (Gonzalez and Sylvester, 1995, Alvarez *et al.,* 2002).

The lack of data results from the size and copy number of the repeat, making it impractical for traditional sequencing. The rDNA repeat is at least 10 kb long in most eukaryotes and in mammals usually even longer, over 40 kb (Gonzalez and Sylvester, 1995, Grozdanov *et al.*, 2003,). In combination with the high copy number, the amount of sequencing to produce a consensus is prohibitive – roughly 13- 17 megabases for *Homo sapiens.*

Genome sequencing might be expected to provide authoritative sequences but to date this has not proved practical. The combination of high copy number, high conservation in some regions, and location within heterochromatin makes assembly of rDNA sequences during genomic sequencing highly problematic. With a hierarchical shotgun approach, clones containing rDNA will normally not even be sequenced; clones are only sequenced if they can be unambiguously placed in a genomic map and clones from large repetitive regions can go in any of multiple locations (She *et al.*, 2004) . With a whole genome shotgun approach, sequencing reads will be made in the repetitive regions. However, current assembler methodology will ignore the reads since their assembly will be highly ambiguous due to multiple copies and the high error rate of individual reads (Venter *et*

al., 2001). As a final obstacle, rDNA arrays vary in length within a single individual at least in *Drosophila melanogaster* (de Cicco and Glover, 1983) and *Canis familiriaris* (Strehler and Johnson, 1972), so there may well not even be a stable sequence of repeats within arrays to assemble.

Consensus and variability information for the ribosomal repeat would allow tests of several hypotheses of the biological role of the ribosomal DNA repeat in aging and speciation. The leading cause of aging in *Saccharomyces cerevisiae* has been established as the proliferation of an episome that escapes from the rDNA repeat and eventually becomes present in such numbers that the cell is unable to divide (Sinclair and Guarente 1997). The anti-aging sirtuin genes impede this process (Kaberlein *et al.*, 1999). These genes also function to extend lifespan in *Drosophila melanogaster* and *Caenorhabdytes elegans* (Wood *et al.*, 2004), although the mechanism is not currently known. In mammals, knockout of a sirtuin gene produces a very aggressive rapid aging syndrome (Mostoslavsky et al., 2006) and aging in *Canis familiaris* also associated with changes in the length of the rDNA repeat (Strehler and Johnson, 1972). Human ribosomal arrays show evidence of dynamic somatic rearrangement (Caburet *et al.*, 2005) although not associated with aging.

Somatic mobile element activity within the rDNA is an attractive model for aging in species benefited by sirtuins, as well as a plausible cause of the observed somatic changes. Mobile elements are known to be frequently present in rDNA and to be associated with rapid sequence evolution (Averbeck and Eickbush, 2005, Guimond and Moss, 1999) At present, though, there is no systematic method to identify possible mobile

DNA agents within the rDNA array. Hence we cannot determine whether sirtuin genes suppress aging in animals the same way they do in yeast.

Changes in the rDNA repeat are also associated with speciation. *D. melanogaster* has active ribosomal arrays on the male Y chromosome while *D. simulans* lacks them. Ribosomal activity in rescued hybrids also indicates epistatic interactions lower ribosomal transcription in a manner predicted by Dobzhansky for speciation genes (Granadino *et al.*, 1996). Changes in repeated DNA, such as rDNA, are quite characteristic of species boundaries and are a major part of chromosomal differences between species (White, 1978; Vershinin *et al.* 1996; Pons and Gillespie 2003; Malik and Hendreki, 2002). Thus, the critical and highly expressed rDNA repeat could be involved in both chromosomal and genic speciation.

The transcribed spacers within the rDNA repeat have been used to create phylogenies of closely related species or even subspecies. Even though some parts of the rRNAs are so highly conserved they can serve for phylogenies spanning the entire tree of life, the transcribed spacers change more rapidly than most single-copy DNA and are thus useful for phylogenies of very closely related groups. However, accurate use of these sequences for phylogeny requires knowledge of intraspecies variability; rDNA sequences can vary markedly not only within a species but within a single individual (Kuo *et al*, 1996, Averbeck and Eickbush, 2005, Delaney 2000). In particular, the lifespan of a variant from introduction to fixation can profoundly influence the scale over which the sequences are phylogenetically useful. Evidence-based models for rDNA variability would thus help plan use of these hypervariable regions.

Because of the limited sequence information, the evolution of complete rDNA repeats is largely unknown. The intense conservation of some rRNA regions demonstrates strong purifying selection (Gerbi *et al.,* 1987), but it is unclear what mechanism could maintain consistent genetic information across 300+ copies. Can an organism tolerate some defective copies or must essentially all be functional? Deep evolutionary analysis indicates most RNA conservation occurs at the level of secondary structure but species sometimes differ in the exact nucleotides composing a conserved pairing. Can sequences containing a single base change without a compensating mutation become common or must the compensating mutation occur before a sequence variant can become common? Purification could occur via rapid homogenization, allowing selection to proceed on whole arrays. The homogenization of tandemly repeated DNA has some theoretical models – but testing these models require variability data not available from just sample sequences (Alkan *et al*, 2004) and so the models remain untested for rDNA.

In this work I produce consensus and variability data by extracting sequence information directly from whole genome shotgun reads. These reads are expected to provide an enormous amount of sequence information on repetitive regions such as ribosomal DNA. Although full assembly of the sequence of repeats within chromosomal arrays appears impossible by these methods, it is possible to align enough of the sequences to determine a consensus sequence, defined as the sequence best matched to the entire set of reads. Comparisons of read variations with the resulting consensus indicates how often the original sequences vary at each site, and large-scale rearrangements such as mobile DNA inserts or large deletions can be identified with

traces that partly align to one region of rDNA and partly align to either mobile DNA sequence or other regions, respectively.

This study addresses the following questions:

1) Can raw reads from genome sequencing projects be used to extract consensus sequence information without sample sequences? Current assembly methodology ignores highly repetitive regions due to mathematical and computational complexity. However, raw sequencing reads of rDNA are available for almost all sequenced genomes. If this information could be used to construct consensus sequences, then many of the questions raised above could be addressed with any additional sequencing.
2) Are rDNA arrays in *H. sapiens* essentially homogenized as in fungi with single rDNA arrays? As Ganley and Koboyashi point out, their finding of minimal variability within individual arrays was restricted to species with single arrays and highly inbred population structures. Variability information from the human repeat can determine whether an outbred species with multiple arrays displays the same pattern of homogenous arrays.
3) Is there sequence evidence of large-scale rearrangements between or within the human rDNA arrays? Caburet *et al.*, 2005, found repeat length variation and indications of reversals using a chromosome painting technique. However, there is no sequence data supporting their finding. Both length variation and reversals require that the repeat arrays have large-scale rearrangements; indels for length

variation and inversions for reversal. RDNA sequence variability allows us to directly look for the predicted DNA changes.

4) Do the arrays show evidence of recombination between divergent repeats? The standard model for tandemly repeated DNA evolution has it driven by illegitimate recombination. In addition, one model for concerted evolution of the rDNA arrays also works through illegitimate recombination between different arrays on different chromosomes. Both processes predict DNA recombination between divergent arrays will occur at crossover points. Sequence variability data will allow us to identify the presence of recombination and possibly hotspots or features associated with elevated recombination.
5) Can active mobile DNA be detected within the human rDNA array? Mobile DNA element activity is associated with aging in *S. cerevisiae* and also provides a mechanism for the indels and inversions predicted by Caburet *et al.* There are also active mobile DNA elements in the *D. melanogaster* arrays. However, it is not know if there are active mobile DNA elements in the human arrays nor is there an easy method to detect them. Variability information could identify occasional DNA insertions and thus produce direct candidates for active DNA elements, or it could show that such elements are at least comparatively rare.
6) How does ribosomal DNA repeat variability compare to euchromatic DNA? Because of the multiple copies, ribosomal DNA is expected to be more variable than single-copy DNA. The extent of the relative variability affects at what phylogenetic level rDNA will be useful. In the case of human, increased rDNA

exact subrepeat in the 328-base regions beginning at bases 21668 and 24163 (not shown in Figure 2.2). In the threefold subrepeat region the first repeat had 95.9% identity to the second repeat and 96.2% to the third. All three copies are also surrounded by ct/cttt rich microsatellite sequence.

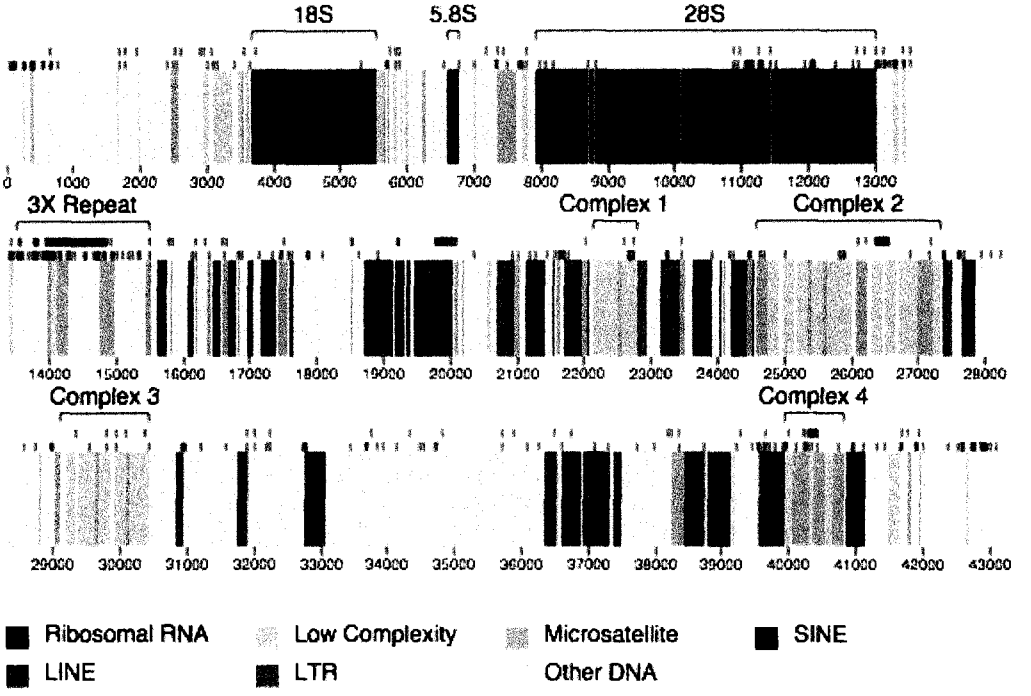

Figure 2.2: Schematic of the human ribosomal DNA repeat consensus. Tick marks immediately above the bars indicate differences from the Gonzalez and Sylvester sample sequence (Gonzalez and Sylvester, 1995); tick marks above those indicate differences from the chromosome 22 sample sequence. Solid bars instead of tick marks, indicate regions deleted in the sample sequence. The four microsatellite complexes are large regions containing almost exclusively C and T residues.

On the transcribed strand, the microsatellites and low complexity regions are overwhelmingly dominated by pyrimidines (cytosine and thymidine). The overall rDNA repeat was 58.4% pyrimidine but complex 3 was 79.4% pyrimidine and complexes 1, 2, and 4 were 87.6%, 87.6%, and 88.3%, respectively. All four microsatellite complexes

variability may mean more phylogenetic information from earlier human evolution survived the bottleneck approximately 250,000 years ago and provide a unique window into that period.

7) How typical of human rDNA are the known sample sequences? Sample rDNA sequences are often used as adequately representing typical rDNA repeats and part of the motivation for Ganley and Kobayashi's work was to determine if this was true. If all or most repeats are very similar to any sample sequence there is little need to perform more detailed sequencing or assembly to determine individual repeats. As an outbred species with multiple arrays, humans would be expected to have fairly variable arrays and so if a human rDNA sample adequately represents most human repeats then the current method of determining sample sequences only will suffice for most needs.

Results

51,203 sequencing reads matched to the final 43,113 bp sequence (Appendix C), which is presented schematically in Figure 2.2. The upper bar shows the transcribed region and is dominated by the small (18S) and large (28S) subunits. The intergenic spacer is split between the middle and lower bars, containing the 3-fold repeat and the four microsatellite complexes. All the SINE elements are Alus or Alu fragments. RepeatMasker identified the 24 regions as derived from 17 different Alu families and generally highly divergent, resembling typical Alu junk found throughout the sequenced human genome. The intergenic sequence contains two groups of large direct subrepeats: a threefold subrepeated region immediately after the end of the transcribed region and an

had only some repetitive guanosines in a few stretches and no repetitive adenosines whatsoever. Shorter microsatellites scattered through the repeat are also mostly polypyrimidine but sometimes are mixed or polypurine tracts.

GC proportion also varies within the overall repeat. The transcribed region is extremely GC-rich (72.5%), as universally observed in transcribed pre-mRNA. The intergenic spacer is somewhat enriched, with 53.6% GC, significantly higher than the human average of 41%. The first three microsatellite complexes are mostly CT repeats and accordingly have GC proportions of 47.7%, 48.3%, and 47%. The fourth complex has long stretches of CTTT repeats, resulting in a lowered GC content of 35.7%.

Consensus accuracy

With an indirectly determined sequence it is necessary to verify that at most only a few reads were included incorrectly, that the extracted consensus matches a large number of actual sequences at each site, and that few reads were missed determining the sequence at each site. As described in the methods, the selection process included several procedures to identify and exclude possible sources of intrusive reads from other sequences. 55 genomic segments were identified as potential sources of intrusive sequences and all reads from those sequences excluded.

Some sections of the rDNA repeat have homology to other regions of the genome, particularly the Alu repeats and the pyrimidine-rich microsatellite regions. Intrusive reads from homologous regions could potentially be inaccurately assigned to the rDNA repeat

and alter the consensus. However, many reads in the human genome sequencing project are associated with mate pairs, other reads known to be from nearby DNA due to the sequencing techniques. An intrusive read would almost never have a mate pair that also was incorrectly aligned to the rDNA repeat, and so excluding sequences without mate pairs also aligned to the repeat would exclude them. Excluding sequences without mate pairs did not change the consensus, indicating that the intrusive sequences, if any, did not alter the final consensus.

I identified that one genomic sequence on chromosome 22 (accession AL592188) producing reads aligning to the rDNA consensus contained a complete rDNA copy (base pairs 105422 to 149395). This copy was used as a sample sequence (chr22ss), supplementing the published sample sequence (GSss, Gonzalez and Sylvester, 1995). That source clone was not used to exclude reads. However, a 2 kb insert in microsatellite complex 1 region of chr22ss had weak homology to the consensus microsatellite complex 2 region. Reads from this region would have been misaligned to the complex 2 region, producing spurious variations, and so they were excluded from the analysis.

The consensus corresponded to actual sequences at all locations, except at one base pair junction where there was no aligned read. I compared each 40-base window in the consensus to aligned traces. The average percentage of reads with exact matches was 98%. The smallest percentage of exact matches was 18.3%, which occurred in a region containing three successive highly polymorphic repeats, but even this site had 31 exact matches.

Since trace reads were selected based on matches in a large region but large mismatched areas were allowed (see methods), improperly aligned reads intruding from other sequences should sometimes include mismatches at both ends (overhangs). Thus, all aligned sequences were examined for high phrap quality overhangs at both ends. No such sequences were found, demonstrating that intrusive sequences are at least comparatively rare.

If some individual repeats differed from the consensus too much in some regions, then regions would have a shortage of aligned reads. Read coverage was examined over the length of the repeat to determine whether this had occurred. The expectation was that coverage should be high in view of the 300-400 source sequences and relatively constant throughout the sequence. Coverage was actually found to vary markedly by site (Figure 2.3), which indicates that either the sequencing process has strong biases or that the set of reads used for the consensus is not properly representative of the entire set of repeats in low coverage regions. If some groups of repeats were too divergent from the overall sequence for reads from that region to be aligned, some reads would cover the point where the repeat begins to diverge from the consensus. These reads would be identified

as overhangs, with the part of the read from the nondivergent area aligned and the part from the divergent area unaligned. No such overhangs were found near the extreme shortfalls (<20 covered reads), indicating that the selection process did indeed capture virtually all appropriate reads. The shortfall thus comes from biases in sequencing which result in some regions being sequenced quite poorly.

Coverage was examined in more detail to determine possible causes of the shortfalls. Average coverage was 397 reads per nucleotide, with a minimum of 1, in the second internal transcribed spacer, to a maximum of 1318, in the intergenic spacer at base 33985 (Figure 2.3). Since the rDNA repeat is estimated to have 300-400 copies, 397 reads per site indicates approximately 1 to 1.3 per physical DNA segment. The lowest coverage rates mostly occurred in transcribed spacer regions; at the end of the 5' external spacer, in both internal spacers, and at the beginning of the 3' external spacer. The exception occurred in the large ribosomal unit. The most extreme shortfalls are thus associated with sequences subject to cleavage during transcription termination or post-translational processing.

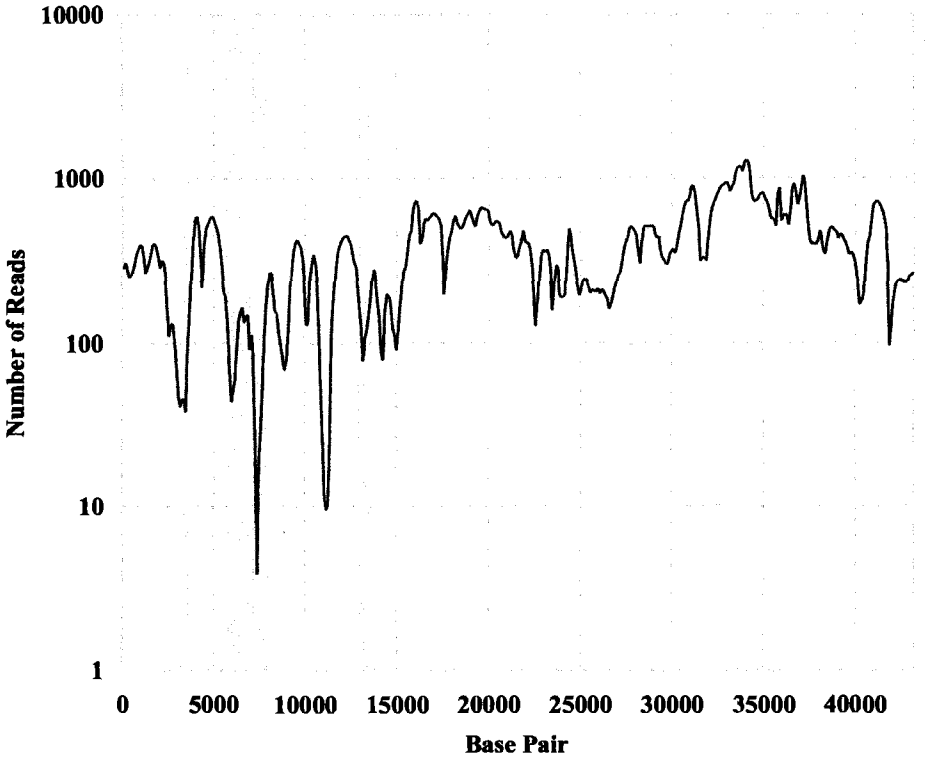

Figure 2.3: Read coverage along the consensus sequence of the ribosomal repeat. The number of reads aligning to each base was totaled and then averaged in blocks of 100 bp. Grey region indicate transcribed areas associated with cleavage sites.

Sequence variability

Variability by base was assessed to permit comparisons with other sequenced regions and to determine which variations resulted from sequencing error and which represented actual polymorphisms. Of 17,121,237 comparisons between a consensus base pair and an aligned read, in 84,698 (0.49%) the read differed from the consensus, including both length and point variations. In 59,722 cases (0.35%) the polymorphism was statistically supported, by rejecting the null hypothesis of sequencing error with a p value < 0.001

(see methods for details). This corresponds to 1075 sites having a statistically demonstrable polymorphism (2.5%).

The repeat was divided into three regions expected to differ in polymorphism rates: the transcribed region, the microsatellite complexes, and the remainder of the intergenic spacer. The transcribed region was expected to have a lower polymorphism rate due to sequence constraints, while microsatellite complexes were expected to have an increased rate due to elevated stutter and skip mutation rates (Hile *et al.,* 2000). Frequencies of polymorphic sites varied substantially by region ($X^2 = 381$, $p << 0.001$). The transcribed region had only 153 demonstrable polymorphisms in 13,347 bases (1.1%), against 822 polymorphisms in 29,766 bases in the intergenic spacer (3.1%). Within the intergenic spacer, the microsatellite complexes had 266 demonstrable polymorphisms in 5961 sites (4.4%.) compared to 656 polymorphisms in 23805 sites (2.7%) in the remainder of the intergenic spacer.

Figure 2.4 shows percentages of aligned reads matching the consensus versus the number of sites at which that percentage was observed. Only verified polymorphisms were included. The restriction to verified polymorphisms created a an ascertainment bias in that rare polymorphisms could be misidentified as unverified if there were not enough observations. Consequently, sites with a consensus proportion exceeding 97% were excluded from the chart. Most polymorphisms are uncommon, and there are no concentrations of frequencies suggestive of multiple fixed differences between chromosomal arrays.

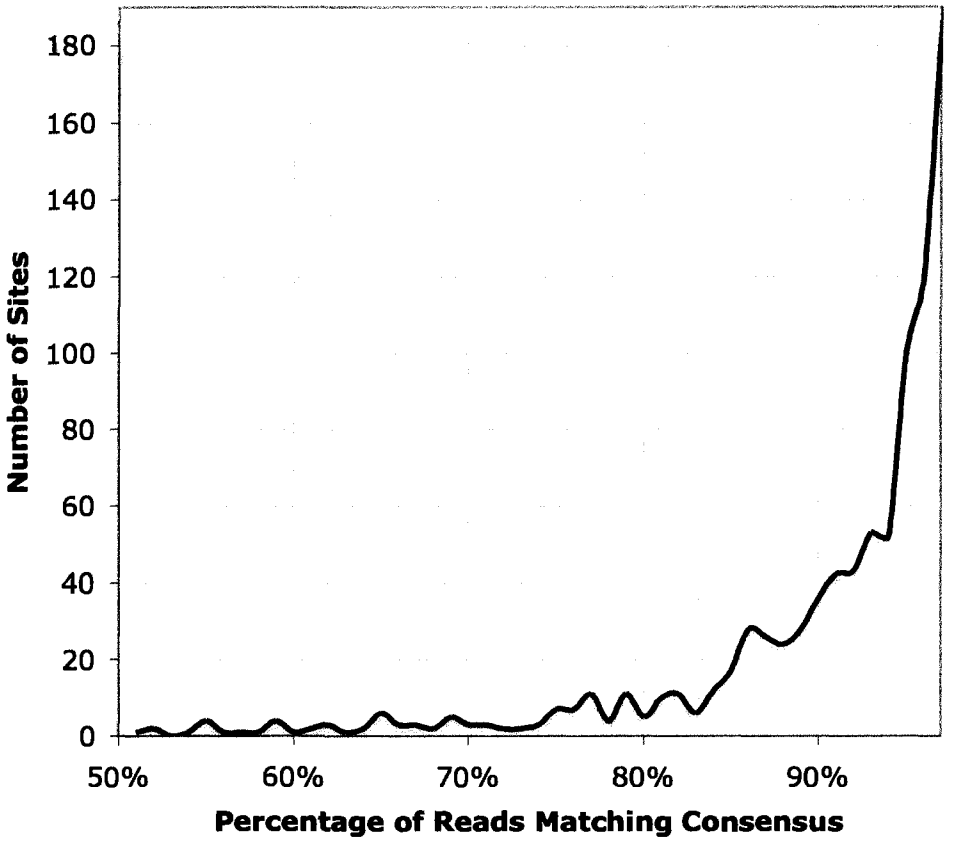

Figure 2.4: Frequencies of the Proportion of the Consensus Variant. The x axis indicates percentages and the y axis gives the number of variable sites at which that percent of reads matched the consensus. Fixed differences between chromosomal arrays would produce clumps of variable site frequencies.

Array monomorphism and Recombination

Pairs of biallic loci (4 possible allele combinations) were chosen which could be associated by shared presence on individual read or read mate pars. These pairs were classified as bivariate (2 combinations observed), trivariate (3 combinations), or quadrivariate (all combinations) to determine the frequency of recombination and whether individual arrays were monomorphic. High recombination predicts rare bivariate

and common quadrivariate pairings while monomorphic arrays predict common bivariate and rare quadrivariate pairings. Low coverage can produce spurious missing combinations due to undersampling so only pairings with at least 6 expected occurrences for each of the four alleles were examined.

1895 pairings of biallelic loci were found but in 414 no variation was observed at one or the other locus. These 414 degenerate pairings occur since with nearly half of the sequences pairing information is not available and so rare alleles can be lost in a subsample. Of the remaining 1481 pairings, 56 (4%) were bivariate, 895 (60%) were trivariate, and 530 (36%) were quadrivariate. Although bivariate and trivariate pairings were 64% of all non-degenerate pairings, only 514 (34.7%) of the non-degenerate pairings were statistically distinguishable from linkage equilibrium using a chi-square test (p value of 0.01) Because of the low number of pairings available in many case this number might underestimate the number of pairs in linkage disequilibrium.

Bivalent and trivariate pairs provide information about the associations of particular alleles at different loci. In a bivariate pair the presence of either allele at one locus implies the presence of the matching allele at the other locus. In a trivariate pair, one of the two alleles at each locus predicts the presence of a specific allele at the other locus. If these pairings are dense enough then the presence of the allele at the second locus predicts the presence of additional alleles at other loci, and the chain of implications can be pursued to produce haplotypes of variants. All bivariate and trivariate pairings statistically distinguishable from linkage equilibrium were used in an attempt to construct

haplotypes, but there were not enough such pairs to construct useful haplotypes. The largest groupings contained only 6 loci, all near microsatellite complex 4.

Rearrangements

Any large inserts or deletions in an array were expected to produce overhangs, in the form of reads spanning the junction which would partly align to the consensus and partly fail to align to the region immediately adjacent. Techniques to identify overhangs are detailed in the methods. 418 overhangs longer than 40 bp were found, averaging 114 bases long. 101 overhangs were excluded as unlikely to demonstrate actual rearrangements on various grounds: 29 overhangs were too short and repetitive to determine possible sources, 8 were chimeras to unidentifiable or nonhuman sequence, 21 were chimeras to nonrepetitive sequence not containing rDNA arrays or likely mobile elements, and 43 were rDNA variants misidentified as overhangs due to local rearrangements or 4 or more variations in a 50 bp region.

Only a few overhangs suggested rearrangements to non-rDNA sequence. 9 overhangs were chimeras to dispersed repetitive sequence: 7 to LINE sequences, 1 to an LTR/Alu mix, and one to a LINE/Alu mix. All were to incomplete elements. 5 reads identified the bridge to non-rDNA sequence in chromosome 22. No other bridges were identified.

The remaining 303 overhangs suggested rearrangements within the rDNA repeat based on strong homology to the rDNA repeat more than 100 base pairs from the area of the repeat they had been aligned to. They fell into 34 groups, with each group containing a single attachment point and extensive sequence similarity between overhangs. All groups had microsatellite repeats either in the overhang or in the sequence the overhang

was attached to. Alignments to particular places within the microsatellite repeats are ambiguous and therefore a modified extension method was attempted to verify the rearrangement with a minimum exact match distance of 80 nucleotides due to the highly repetitive sequence. The longest read from each group was used as a seed and the extension process continued forwards and backwards until regions of more than 300 bp homology to the consensus were found. When extended, 2 groups recovered the insert in the chromosome 22 sample sequence (chr22ss) microsatellite 1 complex. 1 group recovered the chr22ss microsatellite 4 complex. 1 more group recovered a sequence strongly homologous to the microsatellite 4 complex, but differing from both samples and the consensus. The groups recovering chr22ss complex 1, however, misidentified the attachment point as being in complex 2. None of the other 30 groups were successfully extended; extension either failed or simply recovered the consensus sequence from the region of the attachment. The failures to extend suggest that the large number of sequence matching the consensus outweighed those matching the seed and the iterative alignments thus just returned the consensus sequence.

When an overhang could be aligned to a region of the consensus sequence, the direction, the read then had two matched areas and these were used to determine whether the indicated rearrangement was a reversal. Each match area matched either the transcribed or the antitranscribed strand of the rDNA repeat. If both matched regions matched to the same strand then the rearrangement was in frame, while if they aligned to opposite strands then the suggested rearrangement was a reversal. All internal rearrangements were in frame, indicating none were reversals.

Sample Sequence Comparisons

GSss and chr22ss were aligned against the consensus to identify how each differed from the consensus. The trace database was examined for observations of all GSss and chr22ss variants. Gonzalez and Sylvester's published sequence (1995) had 112 deletions, 80 insertions, and 178 single-base replacements compared to the consensus. 10 additional "replacements" occurred where the sample sequence did not identify a specific base and were not considered further. Support for variations from the consensus were determined by comparison to the read database. Variations that were matched by no reads at all were categorized as "not observed". Variations that matched some traces but not enough to be considered statistically supported according to the tests described in the methods were categorized as "unsupported". Variations that were matched by enough reads to be considered supported by the statistical tests described in the methods were categorized as "supported". Results are presented in Table 2.1. Locations of the variants are presented in Figure 2.2.

Table 2.1
Statistical Support for Variations in the Published Sample rDNA Sequence

	Deletions	Insertions	Replacements
Supported	13	9	44
Unsupported	15	6	17
Not Observed	84	65	59
Total	112	80	178

The trace database supported the presence of 37% of replacements, 12% of deletions, and 11% of insertions. An additional 14% of replacements, 13% of deletions, and 8% of

replacements were observed at rates not statistically distinguishable from sequencing error (see methods for test). The remaining 49% of replacements, 75% of deletions, and 81% of insertions were never observed. When supported, variations were common – 71% occurred in more than 5% of reads, while when unsupported, they were almost always rare (less than 3 occurrences) even if observed. The procedure for testing support excluded most variations with a frequency above 7 but not necessarily variations occurring between 4 and 6 times. A large proportion of the replacements occurred as dinucleotide reversals (30 of the 178 replacements in 15 reversals). Monte Carlo simulations were used to estimate the likelihood of so many adjacent replacements occurring by change. 178 replacements were randomly placed on a 43,113 bp sequence and the number of adjacent replacements counted. Even without requiring that adjacent replacements be reversals, in 100,000 trials, the number of adjacent replacements was always less than 6, indicating this clustering was highly unusual. None of these reversals were supported, nor were any of their component changes as individual replacements, suggesting the dinucleotide reversals resulted from sequencing error in the original publication.

A comparatively high proportion of the supported variations occurred in the region containing the three-fold imperfect tandem repeat shortly after the transcription termination signal. In all, the region from 13500 to 15500 in the consensus contains 75 (20%) of all supported differences exhibited by the GS sample sequence (GSss). Since this repeat is imperfect, if GSss had a different arrangement from the consensus it would be identified as having a higher rate of variants. Likewise if either the GSss or the

consensus sequence was assembled incorrectly (plausible in view of both direct repeats and extensive ambiguous microsatellite sequence) this region could have an artifactually high variant rate.

The variations obtained for the published sample sequence were used for a power estimate of the ability of the variant detection system to identify actual sequence variations. For each varying site, the number of sequences necessary to identify a variant was calculated. The variant was assumed to be present in 1% of DNA sites and the chance of a sequencing error was assumed to be the rate of mismatches observed with the overall aligned read set. The chance of observing enough variants reads to classify a variant as supported was calculated with the binomial distribution, assuming the number of observed reads at the site and the rDNA-wide mismatch rate, and averaged over all 370 sites at which GSss differs from the consensus. The average was 47%, indicating the power of correctly identifying a variant with 1% penetrance is approximately 47%.

Chr22ss was 43,973 bp long, somewhat longer than the 43,113 bp consensus. Table 2.2 compares chr22ss with the consensus. Chr22ss had fewer differences from the consensus than GSss, with only 106 differences (0.25%) rather than 370. Sequence reads supported 41% of deletions, 55% of insertions, and 58% of replacements, and an additional 32% of deletions, 10% of insertions, and 18% of replacements were observed although not statistically supported. The remaining 27% of deletions, 35% of insertions, and 24% of replacements were never observed in the aligned traces. There were no dinucleotide reversals. However, some of the changes were much larger than those observed with the previously published sample sequence, most notably the 2020 base

insert into the first microsatellite complex consisting of CT-dominated microsatellite. This insert showed some homology to complex 2 (not to complex 1, where it was found) but the homology was rather poor (85.6% sequence identity). This increase in length was partially compensated for by large deletions in complex 2 and complex 4 (complex 3 was very similar to the consensus, with no indels and only 5 replacements). In the threefold subrepeat region, the 2nd repeat was also deleted with the region between 14021 and 14848 in the consensus missing and replaced with 29 base pairs of microsatellite. All deleted regions have low-complexity c and t rich regions at both ends. Chr22ss differs from GSss by 413 changes, more than either differs from the consensus.

Table 2.2
Statistical Support for Variations in the Chromosome 22 Sample rDNA Sequence

	Deletions	Insertions	Replacements
Not Observed	10	10	10
Unsupported	3	12	7
Supported	16	15	23
Total	29	37	40

Discussion

Iterative stochastic alignment (ISA) was shown to successfully obtain extensive sequence information for the ribosomal repeat, a biologically important but previously unassembled class of DNA. Iterative stochastic alignment (ISA) produced and verified a consensus sequence of 43,133 base pairs. All but one set of overlapping forty 40-bp windows of the consensus were supported by multiple exactly matched regions in reads, despite sequence polymorphism, read sequencing errors, and coverage shortfalls. The

lack of support in one group of 40 windows reflected only low sequencing coverage in one region. The large number of reads (51,203) gives high confidence in most regions of the consensus sequence. Large rearrangements or intrusions would have produced supported overhangs but none were found outside of large microsatellites. While a few reads not derived from the rDNA repeat might share enough sequence similarity to be incorrectly aligned, there would have to be very few, since there were only 14 overhangs involving non-rDNA sequence. Common variations were reliably identified; power analysis showed variations present in only 1% of rDNA repeats would be identified 47% of the time.

1) Can raw reads be used to extract de novo sequence information?

Iterative stochastic alignment (ISA) was shown able to determine consensus sequences directly from the trace databases. A local application of the ISA extension process obtained both the microsatellite 1 and microsatellite 4 alleles from chr22ss, starting solely from individual traces with large overhangs. The ISA extension process also obtained an alternative microsatellite 4 allele present in neither sample sequence nor the consensus, This succeeded in spite of the highly repetitive nature of these regions and the presence of very similar alternative sequences in the region due to reads resembling GSss. Attempts to recover other variants suggested by groups of overhangs only recovered the main consensus sequence. From this we may conclude the ISA process can reliably obtain consensus sequences even from complex repetitive regions with multiple common alleles, and in some cases can even distinguish alternatives to the consensus.

2) Are human rDNA arrays homogenized?

Ganley and Kobayashi found (2007) nearly perfectly homogenous arrays within 5 fungal species. The data from the human trace databases support the model that each human chromosomal array is polymorphic. If the repeats within an array are largely homogenized, then repeats within a chromosomal array with a shared variant at one site will generally share variants at another sites, but the number of such bivariate pairings is low. Indeed bivariate pairings were much rarer than degenerate pairings where one locus was momomorphic in the pairable reads due to sampling error, suggesting most bivariate observations result from undersampling. Second, most variants are uncommon, occurring in less than 10% of aligned reads. If variation were primarily due to differences between monomorphic arrays, variation frequencies should mostly be those of combinations of arrays, which would generally be above 20% since there are only 5 arrays. Third, both sample sequences differed from the consensus at numerous sites where no read in the entire sequencing project had the same sequence as the sample chromosome. Hence all these variants must be rare and I predict that most individual repeats will have several rare variant sites, which in turn means array monomorphism is impossible.

Hypothetically, the unsupported variants in the sample sequence could result from sequencing errors in producing the sample sequences, but estimated error rates for the sample sequence indicates this is implausible. Sequence errors are present in the Gonzalez and Sylvester sample sequence, i.e., the 10 undetermined bases and at least the majority of the 15 unusual and unsupported dinucleotide variants analyzed in the results section. However, for all the unobserved variations to be errors would require an error

rate of 208 nucleotides (0.5%), implausible given the careful quality control discussed in their paper. The chromosome 22 sample sequence was part of the human genome project and subject to strict quality control criteria, with an error rate estimated as less than 1 in 10,000.

The difference between fungal and human intrachromosomal variation is not unexpected. Ganley and Kobayashi observe that the fungi all have only one array and inbred population structure, predisposing them to low variation. To verify the differences between human and those particular fungal rDNA arrays are genuine and not an artifact of differing techniques, the techniques of this work were used to replicate Ganley and Kobayashi's result with *S. paradoxus*. The results were similar, with extremely limited variation (data not shown). Coverage was also more even than with the human sequence, varying from 82 to 306 reads aligning at each base pair, compared to 1 to 1317 with the human trace database.

3) Is there evidence of large-scale rearrangements?

The presence or absence of large-scale sequence rearrangements within individual repeats cannot be definitively determined. The 2 kb insert into the chr22ss microsatellite complex 1 and the deletion in the threefold subrepeat show large-scale rearrangements in the microsatellite complexes are possible but not that they are common. Apart from low-complexity polypyrimidine regions, there are no rearrangements in either sample sequence and not even any candidate rearrangements indicated by the traces. However, the average coverage for individual repeat only 1 to 1.3X, so there is a possibility of lacking reads across a particular rearrangement junction with enough high-quality

sequence on both sides of the junction to be identified as a candidate rearrangement. Hence a high frequency of rearrangement not involving polypyrimidines can be excluded but a low frequency of such rearrangements remains possible.

In polypyrimidine-dominated repeat regions, there are rearrangements. The chromosome 22 sample sequence contains a 2 kb insert in microsatellite complex 1 and an approximately 500 bp deletion in the threefold repeat, between two polypyrmidine sections. Four candidate rearrangements were identified; 2 matched the chr22ss insert and 2 more matched alternative versions of the microsatellite 4 complex. However, 30 candidate rearrangements could not be precisely identified. A definitive answer to the frequency and causes of rearrangements will require improvement to the techniques to handle the challenges of alignments in microsatellite complexes.

No rearrangements were found that would produce the palindromic (partially reversed) observations of Caburet *et al.,* (2005.) They painted chromosomes using tags for particular regions of the ribosomal repeat, and found the tag orders were sometimes reversed, suggesting those sections had been reversed. While there are many candidate rearrangements that cannot be precisely placed, they all involved rearrangements between polypyrimidine tracts. Rearrangments are identified by traces which do not align in their entirety to one region but which are composed of two sections that align to the consensus in different locations. Since the large polypyrimidine tracts all have the same orientation, overhangs from a rearrangement with a reversal should therefore connect an aligned polypyrimidine (forward) section with a polypurine (reverse) section, or a polypyrimidine section with another section reversed compared to the consensus. In all

the candidate rearrangements, both sections had the same orientation. A reversed polypurine section would be very similar to a nonreversed polypyrimidine section, but the 3 polypurine tracts were too short (< 90 bp). Overhangs too short to classify could come from reversals, but of all 303 identifiable rearrangements, none were reversed, and so reversals must be rare.

4) How common is recombination?

Recombination rates between repeats appear quite low. In a large majority (64%) of paired polymorphic sites there was no evidence of recombination since only a minority of paired biallelic polymorphisms was quadrivariate. Even these sites were usually far from linkage equilibrium. If recombination between differing sequences was common most paired polymorphisms would be quadrivariate. Since trivariate pairings were found to dominate (60% of observations), mutation is apparently a more significant force than recombination. Assuming homogenation occurs primarily by unequal exchange (Stephan and Cho, 1994) this suggests that homogenation occurs either in larger tracts (producing a large change in variant counts with each recombination) or that recombination is restricted mostly to highly similar sequences, allowing changes in variant frequencies that only occasionally produce identifiable recombinants. Future explicit simulations of rDNA evolution will allow investigation of parameters which can produce the observed variant spectra.

5) Is there active mobile DNA?

Active mobile DNA can be identified with chimeric traces that partly align to the consensus and partly to mobile DNA elements. Chimeras between rDNA and LINE

sequences suggest mobile DNA activity in the ribosomal repeats although the evidence is not conclusive. However, the total number of rDNA/LINE chimeric sequences is small and there were even more chimeras to nonrepetitive sequence. The chimeras to nonrepetitive sequence are presumably artifactual as no particular chimera was ever supported by multiple reads and almost all involved chromosomes not containing rDNA arrays.

With high coverage (approximately 4 fold or more per individual repeat), trace information could determine not only mobile DNA activity but whether it occurred in somatic or germline tissue. Germline rearrangements should affect all copies in an individual, and since coverage was expected to be high, should occur mostly in multiples. Lone rearrangements, by contrast, suggest somatic effects since they are unlikely to be present in all cells of an individual. However, the low coverage means that germline rearrangements would also be expected to be observed infrequently and often in one read, and so with the current data set germline and somatic mutations cannot be distinguished

6) How does rDNA variability compare to single-copy variability?

Polymorphism is quite high in comparison to reports of polymorphism in euchromatic regions. Based on the frequency of supported polymorphic variants, a typical rDNA repeat differs from the consensus at only 0.49% of sites, while single-copy euchromatic sequence is estimated to differ from the most common sequence at only 0.1-0.2% of sites (Tishkoff and Verrelli, 2003), yielding a ratio of 1:2.5 to 1:5.

Predicting the variability of the rDNA repeat is difficult. Variation will be lost due to exchange between different repeats but there are hundreds of repeats in 5 arrays and no

standard model for expected rates of exchange between two different regions. However, with a simple model assuming each individual's repeats are a random sample of their parent's repeats, the polymorphism for a neutral variant would be 300 times as high since the effect would be the same as a 300-fold increase in population size, effective only for the rDNA array (Clark 1997). More realistic models with limited exchange between physically separated repeats would further increase expected polymorphism. We can conclude there must be some mechanism resulting in the loss of variability beyond neutral drift.

One model for the relatively low polymorphism ratios is that a relatively small proportion of repeats in an individual contribute to the following generation, reducing the analogue to effective population size for individual repeats. This would suggest than only 2.5 to 5 of the repeats in any individual typically contribute to the next generation. While random amplification during life and gametogenesis is certainly possible, it seems unlikely that so few of the 300-400 typically produce "offspring" in the next generation; this would produce essentially monomorphic arrays and they were not observed.

An alternative model, based on Dover's molecular drive model (1986), is that at least part of the apparent deficiency in polymorphism results from selective sweeps in which some sequences proliferate more efficiently into the next generation (Stephan *et al.*, 1992). Although calculations of effective selection values would require guesses at unmeasured parameters including local recombination rate and mutation frequencies, only very modest selection coefficients would serve to drastically reduce effective population size. (Harpending *et al.*, 1993)

7) How typical is the current published sample sequence?

Both sample sequences varied in many locations from the consensus. More notably, a typical read varied from the consensus at 0.49% of sites, indicating that no single sample sequence would be expected to closely match the consensus. Chr22ss was closer to the consensus (0.25 % variations) than would be expected for an average read but even so differed at 106 sites. Although we have only two observations, it suggests that no sample sequence will accurately describe the rDNA repeat in *H. sapiens*, and possibly in other outbred species with multiple rDNA arrays.

The published rDNA repeat sample sequence (Gonzalez and Sylvester, 1995) was somewhat more variable than an expected typical rDNA variant. It differed from the consensus at 0.86% of sites while a typical read differed at only 0.49% and chr22ss differed at only 0.25%. Most of these variants (67%) were never observed in any read. These variants are apparently a mixture of inadequate sampling in the genome project and sequencing error in the original effort by Gonzalez and Sylvester, which was done when sequencing such a long region with complex internal structure was a great challenge. The complete absence of support for any of the dinucleotide reversals, plus the 10 sites where Gonzalez and Sylvester could not identify a base, indicate there were some sequencing errors in their original effort. However, power analysis indicates rarer variants can be missed in the genomic sequencing, and some variants were missing from the chromosome 22 sample sequence as well, although fewer (28%). The relative contribution of sampling and sequencing error to the high variation rate in GSss is difficult to determine.

Conclusions

Iterative stochastic alignment has approximated, for the first time, a population-wide ribosomal DNA repeat with non-monomorphic arrays. Variability data suggests mutation and changes in repeat frequencies play leading roles in the evolution of the rDNA repeat while recombination is present but rare. Homogenization, if it occurs through unequal crossover, must usually involve either large length alterations with each exchange or exchanges only between nearly identical repeats. Large indels or internal rearrangements preferentially affect CT-microsatellite sequence.

These data suggest the sequencing project produced far fewer reads for the rDNA repeat than would be expected from its performance on euchromatic sequence. This is supported by three pieces of evidence: overall coverage was only 1-1.3X, compared to the 5X expected from the Celera information alone (Venter *et al.*, 2001), the failure to identify more than 1 of the expected 10 bridges to non-rDNA sequence (2 for each chromosomal array), and the presence of known, essential sites with as few as 1 aligned read. This is strikingly different from Ganley and Kobayashi, who found very even coverage throughout the fungal rDNA repeats. This could result from chromatin structure on the mammalian repeat; chromatin structure can interfere with sequencing. Alterations to sequencing chemistry might be useful for future sequencing projects, to allow better coverage in the rDNA repeats.

The techniques used in this work may also serve as tools for several of the repetitive DNA classes that have been somewhat neglected in spite of demonstrated or potential biological importance. Sequence-based studies of centromeric DNA in particular have

been done but have been restricted to large clones and assembled sequence. The centromere varies in local variability, however, and these assembled regions may not be representative of the entire centromeric repeat. The techniques of this report could potentially produce information on typical, not just centromeric repeats amenable to current assembly techniques.

With modifications, these techniques may also help describe more complex heterochromatin such as that found in pericentromeric or subtelomeric regions. These region have proved challenging to assemble on a large scale due to complex and irregular repeated structures with variable lengths and ends (Horvath *et al*, 2003). Although some reads have been successfully assembled, many have not and the information from those reads is currently not used. ISA implicitly assumes sequences occur in tandem arrays or have common beginning and end points. This limitation could be overcome by treating the repeats as collection of sequences with differing start points. With that, it will be possible to assemble unused reads from these complex repeats into variants and determine to what extent the assembled pericentromeric and subtelomeric repeats represent the unassembled sections.

Future work can address other species, and use variability data to address both the population structure of the ribosomal repeat copies and the relation of intrapopulation variability to interpopulational shifts, a core element of the theory of evolution. The human chromosome sequencing projects, in particular, will permit detailed analysis of the homogenization and exchange processes between different chromosomal arrays.

Iterative stochastic alignment allows us to better understand the evolution and function of one of the most important and widespread types of DNA in the biosphere.

Methods

Data sources

Blast searches for extensions were conducted with NCBI Web Blast (National Center for Biological Information). Trace reads were downloaded from the ENSEMBL database (European Bioinformatics Institute and Wellcome Trust Sanger Institute).

Alignment methodology

Alignments were constructed by a combination of exact match and Needleman-Wunsch (NW) techniques (Needleman and Wunsch, 1970). The NW algorithm was performed with standard scoring with +1 for each match, -3 for each mismatch, + 5 for opening a gap and +1 for extending a gap. An initial draft consisting of Gonzalez and Sylvester's 1995 sample sequence was used to align reads and those with a NW score above that of a 100 bp exact match were saved to a file for further processing. 100 was chosen as large enough to minimize spurious alignments but still small in comparison to a typical usable read length of 500 bases. High error ends of reads were trimmed using the phrap score provided in the trace database. Trimming began at each end of each read and continued until a stretch of 10 bases with a phrap score exceeding 20 (estimated error rate < 1%) was reached. The trimmed reads in this file were then aligned against the initial draft sequence and the draft sequence refined to match the most common read sequence for each base pair. Refinements were repeated until the sequence was stable to

refinements with those sequences. Then the entire selection and refinement process was repeated until the sequence was stable throughout the entire process.

Sequence variability

Since individual sequence reads have a relatively high error rate, observed variations do not necessarily correspond to actual variations and statistical methods are required to distinguish actual sequence variants from sequencing error. Sequencing errors were modeled as a Poisson process with the number of aligned reads times the overall mismatch rate for all reads taken as the expected number of mismatches. A variant's existence was considered supported if the cumulative p value for the observed variant count or greater was below 0.001 given a sequencing error probability of 0.0015 (estimated from the data) with the number of aligned reads over the region.

Array Monomorphism and Recombination

Array monomorphism and recombination were analyzed by measuring linkage disequilibrium at paired biallelic loci with more than 25 copies of each allele at each site. Sites could be identified as paired when they occurred on the same read or when they occurred on two reads which were mate pairs, pairs of reads known to come from adjacent sequence (Venter *et al.* 2001). There are four possible arrangements in these regions. When the paired biallelic condition is created by mutation there will initially be only 3 combinations – each of the two alleles at the originally variable loci in combination with the ancestral allele at the other locus, and the new allele at the second locus in combination with one allele at the other locus. The state is referred to as a trivariate. If each of the 5 chromosomal arrays is largely monomorphic, each array can

carry only one allele, and thus at most 5 alleles can exist for extended periods. Frequently one of the 3 combinations will be lost and a bivariate (2 combination) state will result. Recombination, by contrast, will sometimes create the fourth combination and a quadrivariate state. A high frequency of bivariate pairings thus suggests array monomorphism, which a high frequency of quadrivariate states suggests high recombination.

Array monomorphism or polymorphism was verified using frequencies of variable reads at polymorphic sites. If polymorphism results from fixed interarray differences, the proportion of variant reads will be large, equal to at least the proportion of repeats on the individual chromosome (approximately 20% with only 5 arrays in *H. sapiens*). With polymorphic repeats, variations can occur at low frequencies as well.

Overhang detection

Candidate rearrangements were identified using reads that aligned closely to the consensus over only part of the high-quality region of the read. This indicates part of the consensus sequence that is not part of a continuing stretch of tandem repeats. Such reads are expected at the ends of tandem arrays, at breaks due to large rearrangements within array, and at large inserts or deletions within arrays. The section not closely aligning to the consensus is referred to as an overhang.

Overhangs were defined as sections of a homologous aligned read at the end that did not align to the rDNA. An alignment score between the read and the repeat at least equaling that of a 100 bp exact match demonstrated homology. Mismatches were identified as sections of reads in which each section of 50 bases contained at least 3

mismatches. Inserts and deletions were considered one mismatch regardless of how many bases were involved. Internal mismatches received no special treatment as they were captured by polymorphism analysis. Where multiple overhangs from the same site were nearly identical (one difference or less) they were considered as likely to be from the same rearrangement and the longest such overhang was considered as representative of the entire group.

While a read could include a relatively short section of overhang, short overhangs generally do not allow determining whether the overhang region is homologous to known DNA sequence. With such overhangs it is therefore impossible to determine which kind of rearrangement it was. Further, groups of variable sites could sometimes produce artifactual short overhangs and, due to the ambiguity of such short sequences, no method could be determined to separate such artifacts from genuine overhangs. Thus, if the mismatch in the quality-trimmed region was shorter than 40 bp, the overhang was not counted or analyzed.

Detecting Rearrangements within the rDNA sequence

All overhanging regions were realigned against the rDNA consensus. An alignment including an exact match exceeding 20 base pairs was considered to indicate a possible internal rearrangement. A rearrangement was considered confirmed if more than one read had an overhang with highly similar sequence arising at the identical site in the consensus. The target of the rearrangement was confirmed by extending the rearrangement in both directions until long regions of homology with the consensus (at least 300 bp) were reached. Extension was performed by obtaining all traces aligning

closely to the rearranged region, aligning them to the sequence, and then determining the sequence in the aligned traces in the next hundred bases past the end of the sequence.

Detecting External Rearrangements

All overhangs not identified as internal rearrangements were compared to published sequence data. Overhangs were blasted against the NCBI nr database with default alignment parameters and low-complexity and repetitive regions included. If vectors appeared among the top matches the overhang was considered an artifact. Otherwise the overhang was combined with 40 bp of anchoring sequence and blasted against both the NR and the human trace database for matches. Any hit crossing the junction between anchor and overhang in the NR was considered to confirm the rearrangement. At least 2 hits in the trace database were required since the test was to determine whether the first, known, read corresponded to valid sequence.

Intrusive reads

The ribosomal rDNA repeat contains numerous imperfect Alu repeats and extended microsatellites. Reads from similar regions could artifactually align to the rDNA repeat, altering variation estimates and possibly even slightly altering the overall consensus. One Alu9 repeat family was identified as a source of many artifacts and all reads aligning to this family's consensus more closely than to the rDNA consensus were excluded.

Overhangs were also examined for non-rDNA sequence. Reads not from rDNA repeats could potentially align to the rDNA repeat over part of the sequence, producing an overhang. Each read with an overhang was blasted to determine if there was a non-rDNA clone or genomic sequence with a nearly exact match to the read. Such sequences

could potentially be the source of other reads that aligned to the rDNA consensus even though they were not part of the actual repeat. As a result, such sequences were recorded and all reads were compared to all such potential sources of intrusive reads. Any read that aligned to such a sequence more closely than to the consensus was excluded.

As a final step, mate pairs were used to confirm the consensus. All reads that did not have their partner mate pair also matched to the consensus were excluded and the consensus redetermined with this restricted set.

References

Alkan C., Eichler E.E., Bailey, J.A., Sahinalp S.C., and Tüzün, E. 2004. The role of unequal crossover in alpha-satellite DNA evolution: A computational analysis, *J Comp Biology* **11**:933-944

Alvarez L.E., Polanco C., Brison, O., Coutinho, L.L., and Ruize, I.R.G. 2002. Molecular evolution of ribosomal intergenic spacers in *Odontophyrynus Americana* 2n and 4n (Amphibia: Anura) *Genome* **45**:71-81

Averback, K.T. and Eickbush, T.H. 2005. Monitoring the mode and tempo of concerted evolution in the *Drosophila melanogaster* rDNA locus 2005. *Genetics* 105:171:1837-1846

Caburet, S., Conti, C., Schurra, C., Lebofsky, R.. Edlestein S.J., and Bnesimon, A. 2005. Human ribosomal RNA gene arrays display a broad range of palindromic structures *Genome Res.* **15**: 1079-1085

Clark, A. 1997. Neutral behavior of shared polymorphism. *PNAS* **94**:7730-7734

de Cicco, D.V. and Glover, D.M. 1983. Amplification of rDNA and type I sequences in drosophila males deficient in rDNA. *Cell* **32**:1217-1225

Delany, M E. 2000. Patterns of ribosomal gene variation in elite commercial chicken pure line populations. *Animal Genetics* **31**:110-116

Dover G.A. 1986. Molecular drive in multigene families: How biological novelties arise, spread and are assimilated. *Trends Genet* **2**:159–165

European Bioinformatics Institute and Wellcome Trust Sanger Institute. Ensembl Trace database. http://trace.ensembl.org/

Ganley, A.R.D. and Kobayashi, T. 2007. Highly efficient concerted evolution in the ribosomal DNA repeats: Total rDNA repeat variation revealed by whole-genome shotgun sequence data. *Genome Res.* **17**: 184-191

Gerbi, S.A., Jeppesen, C., Stebbins-Boaz, B., Ares Jr., M. (1987) Evolution of eukaryotic rRNA: constraints imposed by RNA interactions. *CSH Symp. Quant. Biol.* **11**:709-719

Gonzalez, I.L., Sylvester J.E. 1995. Complete sequence of the 43-kb human ribosomal DNA repeat: analysis of the intergenic spacer. *Genomics* **27**:320-328

Gonzalez, I.L., Sulvester J.E. 2001. Human rDNA: Evolutionary patterns within the genes and tandem arrays derived from multiple chromosomes. Genomics **73**:255-263

Granadino,B., Penalva,L.O.F., and Sanches,L. 1996. Indirect evidence of alteration in the expression of the rDNA genes in interspecific hybrids between *Drosophila melanogaster* and *Drosophila simulans*. *Mol. Gen. Genet.* **250**: 89-96

Grozdanov, P., Georgiev, O., and Karagyozov, L. 2003. Complete sequence of the 45-kb mouse ribosomal repeat: analysis of the intergenic spacer *Genomics* **82**: 637-643

Guimond, A. and Moss, T. 1999. A ribosomal orphon sequence from *Xenopus laevis* flanked by novel low copy number repetitive elements. *Biol. Chem.* **380**:167-174

Harpending, H.C., Sherry, S.T., Rogers, A.R. and Stoneking, M. 1993. The genetic structure of ancient human populations. *Current Anthropology* **34**: 483-496.

Henderson A.S., Warburton D, and Atwood K.C. 1972. Location of ribosomal DNA in the human chromosome complement. *PNAS* **69**:3394-3398

Hile, S.E., Yan, G., Eckert, K.A. 2000. Somatic mutation rates and specificities at TC/AG and GT/CA microsatellite sequences in nontumorigenic human lymphoblastoid cells *Cancer Research* **60**: 1698-1703

Hillis, D.M., Moritz, C, Porter, C.A., and Baker, R.J. 1991.Evidence for biased gene conversion in concerted evolution of ribosomal DNA. *Science* **251**:308–310.

Horvath, J E *et al.* 2003. Using a Pericentromeric Interspersed Repeat to Recapitulate the Phylogeny and Expansion of Human Centromeric Segmental Duplications *Mol. Biol Evol.* **20**:1463-1479

Institute for Biomolecular Design. E. coli Statistics. http://redpoll.pharmacy.ualberta.ca/CCDB/cgi-bin/STAT_NEW.cgi

Kaeberlein, M., McVey, M., and Guarente, L. 1999. The SIR2/3/4 complex and SIR2 alone promote longevity in Saccharomyces cerevisiae by two different mechanisms. *Genes Dev.* **13**: 2570–2580.

Kuo, B.A., Gonzalez, I.L., Gillespie, D.A., and Sylvester, J.E. 1996. Human ribosomal RNA variants from a single individual and their expression in different tissues. *Nuc. Acids Res.* **24**:4817-4824

Malik, H.S. and Henikoff S. 2002. Conflict begets complexity: the evolution of centromeres. *Curr. Op Genet. Develop.* **12**:711-718

McVean, G., Spencer, C.C., and Chaix, R. 2005. Perspective on human genetic variation from the Hapmap project. *PLOS Genetics* **1**:e54

Mostoslavsky *et al.* 2006. Genomic instability and aging-like phenotype in the absence of mammalian SIRT6. *Cell* **124**:315-329.

Mougey, E.B.; Pape, L.K.; and Solber-Webb, B. 1996. Virtual the entire *Xenopus laevis* rDNA multikilobase spacer serves to stimulate polymerase I transcription. *J. Biol. Chem.* **271**: 27138-27145

National Center for Biological Information. BLAST http://www.ncbi.nlm.nih.gov/BLAST/Blast.cgi

Needleman, S. and Wunsch, C. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol Biol.* **48**:443-53

Pons, J. and Gillespie, R.G. 2003. Common origin of the satellite DNAs of the Hawaiian spiders of the genus *Tetragnatha*: evolutionary constraints on the length and nucleotide composition of the repeats. *Gene* **313**:169-177.

Schlotterer, C. and Tautz, D. 1994. Chromosomal homogeneity of Drosophila ribosomal DNA arrays suggests intrachromosomal exchanges drive concerted evolution. *Curr. Biol.* **4**:777–783.

She, Xinwei *et al.* 2004 The structure and evolution of centromeric transition regions within the human genome. *Nature* **430**: 857-864

Sinclair, D.A. and Guarente, L. 1997. Extrachromosomal rDNA cicles – a cause of aging in yeast. *Cell* **91**:1033-1042

Strehler, B and Johnson, R. 1972. 30 percent decrease in ribosomal DNA dosage during aging of dog brain. *Federation Proceedings* **31**:910

Stephan, W. and Cho, S. 1994. Possible role of natural selection in the formation of tandem-repetitive noncoding DNA. *Genetics* **136**:333-341

Stephan, W., Wiehe, T.H.E., and Lenz, M.W. 1992. The effect of strongly selected substitutions on neutral polymorphism – analytic results based on diffusion-theory. *Theo. Pop. Bio.* **41**:237-254

Tishkoff, S.A. and Verelli,G. 2003. Patterns of human Genetic Diversity: implications for Human Evolutionary History and Disease. *Annu. Rev. Genomic Hum. Genet.* **4**:293-340.

Ventner, Craig *et al*, 2001. "The Human Genome" *Science* **291**:1304-1351

Vershinin, A.V., Alkhimova, E. G., and Heslop-Harrison, J. S. 1996. Molecular diversification of tandemly organized DNA sequences and heterochromatic chromosome regions in some Triticeae species. *Chrom. Res.* **4**:517-525.

White, M.J.D. 1978. Modes of Speciation. W. H. Freeman and Company, San Francisco

Wood, J.G., Rogina, B., Lavu, S., Howitz, K., Helfand, S.L., Tatar,M., and Sinclair, D. 2004. Sirtuin activators mimic caloric restriction and delay ageing in metazoans. *Nature* **430**: 686–689.

Chapter 3: Evolution of the Ribosomal DNA repeat in Homo sapiens and Pan Troglodytes

Abstract: In the past obtaining sequence information for the entire ribosomal DNA (rDNA) repeat has been difficult and rarely attempted in spite of its importance as a component of all cellular genomes, with potential involvement in aging and speciation. As a result, the complex evolution of the multiple copies is poorly understood, including intraindividual variability and coalescence time. Here Iterative Stochastic Assembly (ISA) was used to obtain the rDNA consensus and a variability survey for the *Pan troglodytes*, the common chimpanzee. The consensus and variability were compared to those previously obtained for the human sequence. The rDNA repeat, apart from functionally constrained sections, was found to be one of the most rapidly changing classes of DNA, with mutations fixed at approximately three times the rate of euchromatic sequence. The repeat also showed intraspecific variability much higher than the rough estimates for euchromatic sequence. High variability was present even in some highly conserved regions. Variant lifespan was comparatively short, suggesting rDNA variability may be very useful for distinguishing subspecies. Coverage was more limited than expected, suggesting current shotgun sequencing techniques could be substantially improved for the rDNA repeat. No active mobile elements were identified.

Introduction

The ribosomal DNA (rDNA) repeat is essential to life as we know it and is responsible for more dry weight in most cells than any other gene. It codes for three of the four RNA components of the ribosome, with spacer regions that are transcribed but excised during maturation. All known cellular organisms have rDNA genes, and they are conserved so strongly that some sections of rDNA can be reliably aligned in all organisms. Nonetheless, the repeat remains poorly characterized. Consensus sequences for the entire repeat are available only for five fungal species with short repeats and minimal variability (Ganley and Kobayashi, 2007). In mammals even sample sequences are rare, available for only humans and mice (GenBank).

Technical difficulties with sequencing limit the characterization of the rDNA repeat. The primary limitation is its sheer size. The rDNA repeat is at least 15 kb long in most eukaryotes and in mammals usually even longer, over 40 kb (Grozdanov *et al.*, 2003, Gonzalez and Sylvester, 1995). Further, in all known organisms, rDNA occurs in tandem arrays; either a single array or multiple arrays on multiple chromosomes, with copies of the transcribed pre-RNA separated by spacer regions. The *Homo sapiens* genome contains approximately 300-400 copies on the short arms of the five acrocentric chromosomes: 13, 14, 15, 21, and 22, (Henderson *et al.*, 1972). In combination with the high copy number, the amount of sequencing to produce a consensus is prohibitive – roughly 13- 17 megabases for *Homo sapiens*. Even sample sequences have been skipped for most mammals due to technical difficulties resulting from the extremely large microsatellites and complex interspersed repetitive DNA found in the only two mammals

with sample sequences to date, *H. sapiens* (Gonzales and Sylvester, 1995) and *Mus domesticus* (Grozadnov *et. al.,* 2003).

Genome sequencing might be expected to fill this gap but to date this has not proved practical. The basic problem is that sequencing is a high-error process and multiple sequencing runs must be aligned against each other in order for the errors to be identified and corrected. Additionally, rDNA arrays vary in length within a single individual at least in *Drosophila melanogaster* (de Cicco and Glover, 1983) and *Canis familiaris* (Strehler and Johnson, 1972). In highly repetitive regions like the rDNA repeat, alignments become ambiguous and so there is no direct method to correct the errors. In some genome projects, clones containing rDNA are identified as not alignable and are not even sequenced (She et al., 2004) . In other projects, sequencing reads (traces) will be made but assemblers ignore the traces (Venter et al., 2001.)

The shortfall of sequence information for the rDNA repeat results from the practical difficulties of obtaining rDNA sequences, not from a lack of interest. The leading cause of aging in *Saccharomyces cerevisiae* as been established as the proliferation of an episome that escapes from the rDNA repeat and eventually becomes present in such numbers that the cell is unable to divide (Sinclair and Guarante 1997). The anti-aging sirtuin genes impede this process (Kaeberlein *et al.*, 1999, Mostoslavsky *et al*, 2006). These genes also function to extend lifespan in *Drosophila melanogaster* and *Caenorhabdytes elegans*, (Wood et al, 2004) although the mechanism is not currently known. In mammals, aging in *Canis familiaris* also associated with changes in the length of the rDNA repeat. At present, though, there is no systematic method to identify possible

mobile DNA agents within the rDNA array (or anywhere else). Hence the possibility that the sirtuin genes suppress other age-associated self-replicating DNA remains uninvestigated.

Changes in the rDNA repeat also associate with speciation. *D. melanogaster* has active ribosomal arrays on the male Y chromosome while the closely related species *D. simulans* lacks them. Ribosomal activity in rescued hybrids also indicates epistatic interactions suppress ribosomal transcription as predicted by Dobzhansky for speciation genes (Granadino *et al.*, 1996). Changes in repeated DNA are quite characteristic of species boundaries (White, 1978), and the critical and highly expressed rDNA repeat could provide a mechanism for interactions of chromosomal and genic speciation. However, in the absence of consensus sequences comparisons of different related rDNA sequence are unproductive for this purpose, as inter- and intra-species variability cannot be distinguished.

The rDNA repeat is also a unique model for concerted evolution. In none of the other examples of high-copy tandem repeats (centromeric, telomeric, and satellite DNA) is the repeated sequence strongly expressed, although recent evidence shows that in humans only some of the many rDNA repeats are expressed at one time in a cell (Huang *et al.*, 2005). Concerted evolution is thought to mostly occur through a birth-death model but it is unknown whether this is the case for the rDNA repeat, nor is it clear how concerted evolution interacts with the strong sequence conservation of the rDNA repeat. Models have been proposed for concerted evolution, but without intraspecific sequence

information on the consensus repeat and variability, it is not possible to test these models (Averback and Eickbush, 2005; Alkan *et al*, 2004, Hillis 1991).

Somewhat surprisingly in view of the intense conservation of some regions of the rDNA repeat, other regions change so rapidly they have great value in separating closely related species or even subspecies (Kuo *et al.,* 1996) Potentially this intraspecific variability could be very useful for subsecies identification as the high conservation of some rDNA areas would allow cloning high-variability sequences even in species with poorly characterized DNA. Again, however, without information on typical intraspecies variability or the typical lifespan of variant rDNAs this technique is not practical. Also, only the intergenic spacers, a small fraction of the rDNA repeats, has been analyzed for hypervariability (Averbeck and Eickbush, 2005, Delaney 2000). The much larger and potentially more informative intergenic region has very little sequence information, particularly in amniotes.

Chapter 2 of this work found a variety of unusual features in the human rDNA repeat. Among these were a strong pyrimidine bias in the direction of trasncription, a high frequency of indels in polypyrimidine stretches possibly associated with the intensity of the local polypyrimidine bias, a C+G bias even apart from that associated with the rRNA coding regions, evidence for polymorphic chromosomal arrays, and a high variability in comparison to single-copy sections of the genome. A chimpanzee sequence may help demonstrate whether any of these features are restricted to humans or whether they are more general characteristics of primate rDNA consensuses.

Another feature of potential import is the typical sojourn duration of a rDNA variant. Different sojourn durations can have significant implications or uses. If sojourn durations are long, then the elevated polymorphism rates in the human sequence may arise primarily from incomplete lineage sorting with the chimpanzee and have minimal use for identifying populations. However, if the elevated variability arises from an elevated mutation rate, repeat variants may help to identify recently separated populations. In the case of the chimpanzee, this could resolve unsettled issues of the proper characterization of certain chimpanzee group as populations, subspecies, of full species. Human are currently thought to descend almost entirely from one recent founding population (Harpending, 1993) but in this case rDNA variability might provide information on the population structure of the prehuman populations.

Here I built upon my work on the human rDNA repeats by determining the consensus repeat sequence for the common chimpanzee, *Pan troglodytes*, using Iterative Stochastic Assembly (ISA). As with the human genome, this produces a stable well-supported consensus and extensive information on variability within the ribosomal arrays. The comparison of the two consensus sequence provides a variety of information on recent evolution of the rDNA array within the African Apes.

This chapter will address the following questions:

1) Can iterative stochastic alignment successfully extract a sequence without any sample sequences? The prior application of this technique used a pre-existing sample sequence. Sample sequences are available for only a few species but starting solely from

homology to existing rDNA sequences would allow the technique to be used for any sequenced species.

2) Does the chimpanzee rDNA repeat occur in polymorphic chromosomal arrays like the human rDNA repeat? The only other species with variability information are certain fungal species showing virtually no variability, a finding at contrast with the high variability with humans. Which pattern prevails for the chimpanzee?

3) How does chimpanzee variability compare to human variability? The human population has a much smaller long-term populations size. A comparison of the two would help indicate effects of population size on variant sojurn duration.

4) Does the variability demonstrated in the human ribosomal repeat arise from a high mutation rate or a long duration of polymorphism? If the high variability results from a more rapid evolution then the repeat may well provide information on the structure of pre-human populations. On the other hand, if the high variability results from incomplete lineage sorting and coexistence of multiple deep lineages the repeat would be difficult to interpret in terms of past population history.

5) Is there detectable selection on repeat length in humans and chimpanzees? Human and mouse sequences are similar in length, even though the two species are well separated phylogenetically and most of the repeat is not conserved in sequence. The similarity in length suggests there might be a selection for particular repeat lengths, similar to what has been observed for centromeric repeats (Pons and Gillespie, 2003).

6) Is there differential selection for or against particular sequences in the human and chimpanzee lineages? The human sequence is variable, but less so than expected from

simply neutral models of its evolution (Clark, 1997). A very simple explanation for the existing human variability is selection for or against particular sequences, which could easiliy produce the observed pattern (Stephan *et al.*, 1992, Dover, 1986). However, there is no direct evidence for selection currently but a comparison of human and chimpanzee variation could identify selective effects.

7) Is there an association between pyrimidine content on the forward strand and mutation rates? In the human species the high-pyrimidine microsatellites are associated with higher variability, and possibly mutation, and variability is reduced in the one microsatellite region with lower pyrmidine content. Variability from more microsatellite regions, however, would be necessary to determine whether there is a relationship or whether this is simply coincidence.

Results

The entire chimpanzee repeat was successfully obtained starting with a 300 bp seed from the human sequence. A schematic of the 43,473 bp consensus P. troglodytes rDNA repeat is presented in figure 3.1. The complete sequence is provided in Appendix D.

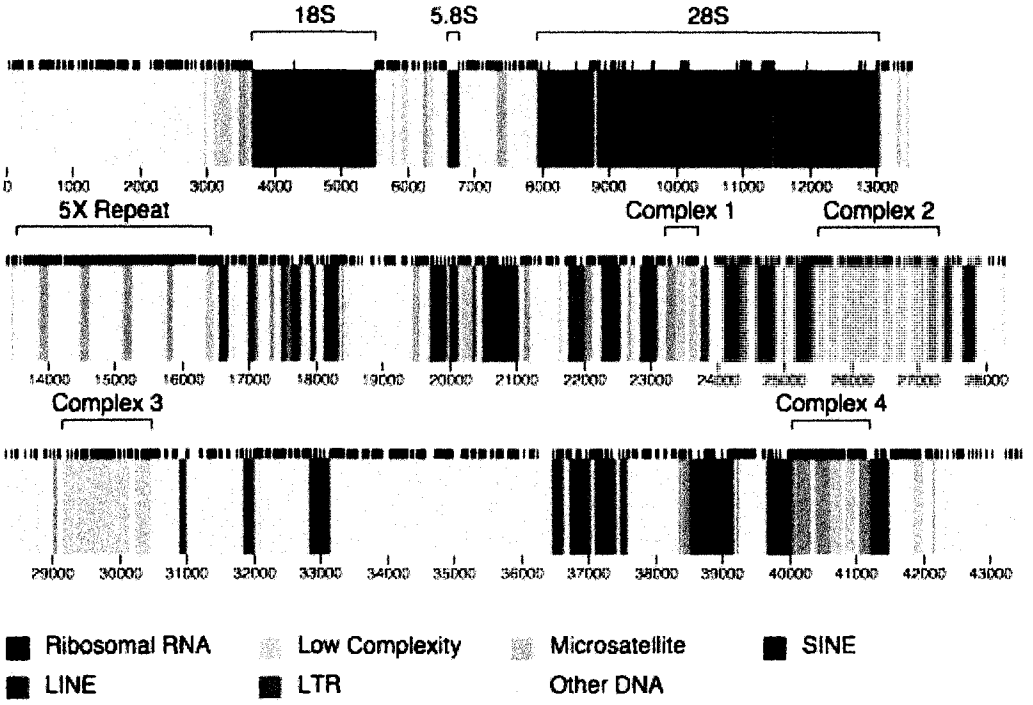

Figure 3.1. Schematic of the *Pan troglodytes* ribosomal DNA repeat consensus. Tick marks indicate differences from the human consensus. Solid bars instead of tick marks, indicate regions present in the chimpanzee but absent in the human. The four microsatellite complexes are large regions containing almost exclusively C and T residues. The upper bar shows the transcribed region and is dominated by the small (18S) and large (28S) subunits. The intergenic spacer is split between the middle and lower bars, containing the 3-fold repeat and the four microsatellite complexes. All the 17 SINE elements are Alus or Alu fragments. They do not derive from the same Alu family and generally are highly divergent, resembling typical Alu junk found throughout the sequenced chimpanzee genome. The intergenic sequence contains two groups of large direct repeats, a 5-fold repeat immediately after the transcribed region and a duplication in the region between complexes 1 and 2 (not diagrammed).

Comparison to human consensus

The P. troglodytes sequence has 283 insertions not present in the human sequence, 197 deletions from in the human consensus, and 1368 replacements. Sequence identity is 96.6% at aligned bases and 90.5% overall. The sequence changes between chimpanzee and human are spread broadly throughout the entire sequence, and not restricted

primarily to microsatellite repeats or the tandemly repeated region. However, sequence identity was higher in the transcribed region (97.8% at aligned and 94.9% overall) and lower in the microsatellite completes (95.8% at aligned and 84.8% overall). These differences were highly statistically significant ($X^2 = 208$, 2 d.f., $p << 0.001$).

Both pyrimidine and G+C proportions have similar patterns to the human repeat. The microsatellites and low complexity regions on the transcribed stratnd are overwhelmingly dominated by pyrimidines (cytosine and thymidine). The overall repeat is biased to pyrimidines on the forward strand, with 58.0% pyrimidine (versus 58.4% in the human) and the microsatellite complexes are mostly pyrimidine (86.9%, versus 85.4% in the human). However, it differs from the human sequence in that the microsatellite complex 3 has a very similar pyrimidine content to the other complexes, while in the human microsatellite 3 has a lower forward strand pyrimidine content of 79.4%. All four microsatellite complexes had only some repetitive guanosines in a few stretches and no repetitive adenosines whatsoever, as with the human. Shorter microsatellites scattered through the repeat are also mostly polypyrimidine but sometimes there are mixed or polypurine tracts.

GC proportions resemble those in the human repeat. The transcribed region is extremely gc-rich (72.0%, versus 72.5% in humans). The intergenic spacer is somewhat enriched, with 52.7% GC, as compared to 53.6% in humans. The first three microsatellite complexes have a gc proportions of 48.8%, 48.6%, and 49.4%, respectively, compared to 47.7%, 48.3%, and 47% in the human. The fourth complex has long stretches of CTTT

repeats, resulting in a lowered GC content of 37.8 % , although not as low as the 35.7% observed in humans.

Consensus accuracy

As with the human sequence, possible sources of intrusive reads were identified and reads possibly resulting from those segments excluded from the data. 14 genomic segments were identified as potential sources of intrusive sequences and all reads from those sequences excluded. No complete copies of the chimpanzee rDNA sequence were found.

To show the consensus corresponds to actual sequences at all locations, I compared each 40-base window in the consensus to aligned traces. The average percentage of reads with exact matches was 97%. The smallest percentage of exact matches was only 7.4%, which occurred in the intergenic spacer outside of the microsatellite repeats, but even this site had 24 exact matches.

Read coverage was analyzed to demonstrate the consensus was everywhere based on a representative sample of reads. Coverage is presented in Figure 3.2. Average coverage was 599 reads per nucleotide, with a minimum of 17, in a microsatellite repeat in the large ribosomal subunit, to a maximum of 2171, in the intergenic spacer at base 33009. Since the rDNA repeat is estimated to have 300-400 copies, 397 reads per site indicates a average coverage of approximately 1.5 to 2.0X. Most of the extreme shortfalls occurred in the same locations as in the human sequence, including all regions subject to cleavage during translational post-processing, supporting the model that incomplete sequencing in these regions produce shortfalls. However, an additional, if less extreme, shortfall

occurred at the start of the second microsatellite repeat (base pair 25848) where the human sequence had no particular shortfall. As with the other shortfalls, there were no overhangs on either side, indicating the shortfall did not result from failing to include reads. Strong secondary structure can potentially interfere with cloning, but this region did not have complex secondary structure.

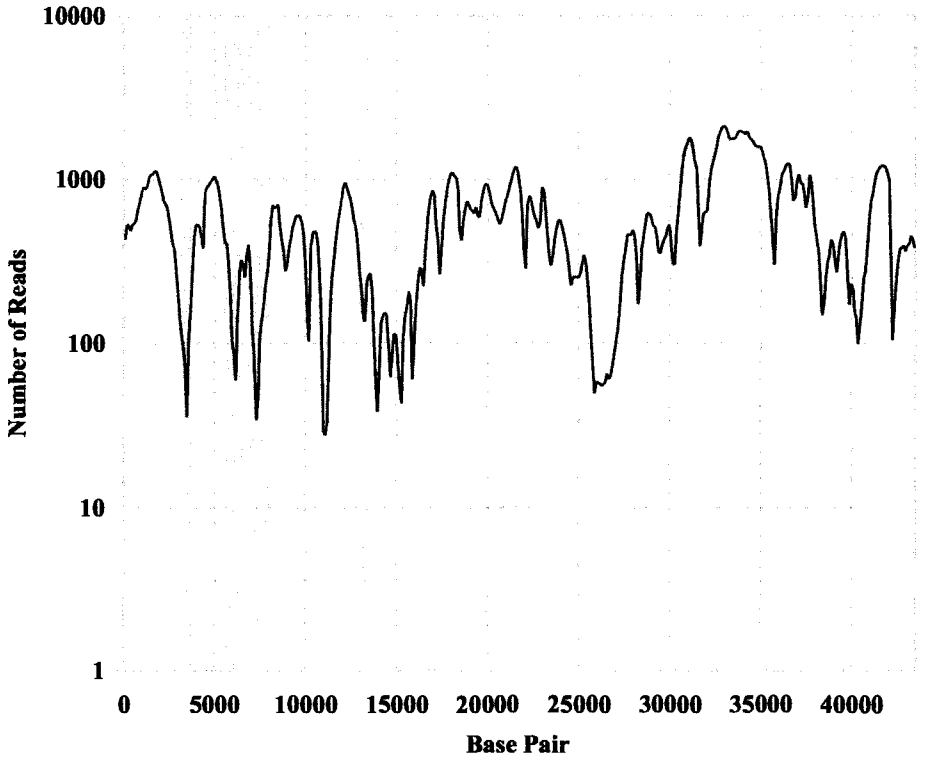

Figure 3.2: Read coverage by site. The number of reads aligning to each base was totaled and then averaged in blocks of 100 bp. Grey areas indicate regions subject to cleavage during pre-mRNA processing.

Sequence variability

Variability by base was assessed to identify statistical significant, or verified, polymorphisms for comparison with human variability data and to determine whether the source of variability was primarily incomplete lineage sorting or recent mutation. Of 26,052,543 comparisons between an aligned read and the consensus, in 165,498 (0.64%) the read differed from the consensus. 116,933 mismatches were verified (0.45% of aligned bases). In all, 1266 sites had a verified polymorphism.

The repeat was divided into three regions expected to differ in polymorphism rates: the transcribed region, the microsatellite complexes, and the remainder of the intergenic spacer. Based on results in the human sequence, the transcribed region was expected to have a lower polymorphism rate due to sequence constraints, and the microsatellite complexes were expected to have a higher polymorphism rate. Both findings were supported. The transcribed region had 300 demonstrably polymorphic sites in 13,369 bases (2.2%) while the microsatellite complexes had 263 sites in 4766 bases (5.5%). The remainder of the IGS had 703 sites in 25338 bases (2.7%). The differences were highly statistically significant ($X^2 = 137$ with 2 d.f., $p << 0.001$).

Sequence variability within the chimpanzee lineage displayed the same pattern as sequence changes between the human and chimpanzee lineages, with the transcribed region having low variability and fewer changes, and the microsatellite regions having higher variability and more changes. This suggests that similar processes might be driving both intralineage variability and interlineage changes. To test this model, the frequency with which the variant from the human lineage was observed in the

chimpanzee reads was determined. The model would predict that number of variable sites is proportional to the number of consensus changes in any region and therefore the proportion of changes consensus sites with observed variability should be similar in region under mild selection.

Of the 1848 changes between the human and the chimpanzee sequence only 219 (11.8%) were verifiably polymorphic. In 302 more changes (16.3%) variants were observed but not often enough to exclude sequencing error. In the remaining 68.9% the human variant was never observed. The microsatellite complexes did not differ from the remainder of IGS regions in the proportion of changed consensus sites with observed polymorphism (73%, and 69% never observed respectively; X^2=0.39 with 1 d.f., p = NS). In the transcribed region, however, fewer changes were observed as variations than in the IGS (78% vs. 70% X^2=17.7 with 1 d.f., p <.001) Thus, in the less-selected IGS regions, the differences in polymorphism rates can explain most of the variation in fixation rates. Since the transcribed region is known to be subject to strong selection, the differences there can be ascribed to a lower rate of neutral polymorphisms permitting fewer changes to exist as polymorphisms.

Selection and rDNA evolution

As observed in Chapter 2, polymorphism in the human rDNA repeat, although high, is much smaller than expected from simple models. Selection could easily explain the observed pattern but more direct evidence is needed to demonstrate or disprove a role for selection. The data available with the ISA technique permits testing two specific selective models: selection for length and selection on short DNA sequences.

Selection based on length.

Centromeric DNA repeats are selected to fall within a certain range of lengths, slightly longer than the length of DNA wrapped by a histone (Stephan and Cho, 1994). Potentially a similar mechanism might operate on the rDNA sequence although no mechanism is known. The human and chimpanzee repeats are quite similar in length in spite of large indel variations in both. Additionally, the mouse rDNA repeat is only moderately longer (45 kb) in spite of a considerable evolutionary difference. To test stabilizing selection for repeat length, indels between the chimp and human sequence were taken as representative of the distribution of potential length changes. This distribution of fixed mutation length changes was resampled with replacement to produce a distribution of net length changes expected between the chimpanzee and human sequences. The actual net change was tested on this distribution to see if it was closer to 0 than would be expected by chance, using a p value of 0.05. This technique was also used for changes between the two human sample sequences.

In 10000 resamplings, the difference between the chimpanzee and the chimpanzee with the resampled indels was equal or less than the observed difference of 274 bases 25.6% of the time. An equal or lesser change was observed in 32.3% of the resamplings for the chromosome 22 sample sequence, and 52.3% with the Gonzalez and Sylvester sequence. Thus there is no evidence for selection on repeat element length.

Selection based on DNA sequences

In *D. melanogaster* and *S. cervesiae*, the rDNA loci are known to carry mobile elements. These elements introduce a possibility of selection based on sequence motifs

within the rDNA repeat if the element have insertion preferences (de Cicco and Glover, 1983). Sites more subject to insertion will have a higher mutation rate and thus will be more likely to have fixed sequence changes over time. In addition, DNA repair can involve large-scale DNA exchange, which could result in the preferential loss of the targeted sequences.

Four-base sequences in the chimpanzee consensus were examined to determine if any particular sequences were more likely to differ between the human and chimpanzee sequence. Each occurrence of each the 256 possible sequences was compared to the aligned sequence in the human consensus and scored for the presence or absence of replacements. The four bases immediately preceding the window were also scored for the presence or absence of replacements to serve as a control for regional variations in mutation rates. Insertions and deletions were ignored due to the ambiguities of placing them in repetitive regions. Sequences were classed in those containing and lacking CG dimers based on the known increased rate of mutation in CG dimers. In each class the percentage of changes in the window as a percentage the total number of changes in the window plus the preceding control window was estimated. The probability of obtaining the observed proportion of changes in each tested regions was estimated with the bionomial distribution, assuming the expected proportion for test region changes for the group (cg-containing or non-cg-containing), with a correction for continuity. 26 of the 256 had a p-value of 0.05 or less. This is significantly more than the number of false positives expected by chance (12.8 expected, $p = .001$ by binomial dstibution) Sequences more likely to have a replacement in the human sequence are listed in table 3.1

Table 3.1: Four-base sequences with higher than expected rates of changes between the human and chimpanzee sequence. Observed is the number of occurrences within the chimpanzee sequences. Control Changes is the number of times the four-base sequence immediately preceding the window had a replacement in the aligned section of the human sequence. Test Changes is the number of changes observed in the four-base sequence contained the target sequence. Sequences were divided into to classes; those containing a CG dimer and those lacking a CG dimer, based on expected and observed differences in changes between the two sequences. The P value gives the probability of observing a that many changes or fewer, given then number of observed reads and a with each sequence having en expectation of a change equal to the average proportion of changes observed for its class of sequence.

Sequence	Observed	Control Changes	Test Changes	p value
tcta	77	7	14	0.952
ctca	134	10	18	0.953
tcat	79	10	18	0.953
gact	105	9	17	0.958
ttga	106	6	13	0.958
ctac	73	6	13	0.958
agaa	137	16	26	0.961
ccct	317	23	35	0.967
tggt	118	10	19	0.967
accc	159	7	15	0.967
tacg	52	5	16	0.970
tgaa	123	6	14	0.972
tccc	442	33	48	0.976
accg	141	10	26	0.977
cagt	86	7	16	0.978
cgct	183	13	32	0.981
gaac	101	4	12	0.982
tcca	115	10	21	0.984
cgtg	224	16	38	0.985
aacg	74	4	16	0.986
cacc	197	21	36	0.988
cgcc	399	26	57	0.991
gtta	42	2	10	0.992
gcac	107	10	23	0.993
tgac	119	7	19	0.994
acag	123	7	20	0.996

Intrachromosomal differentiation

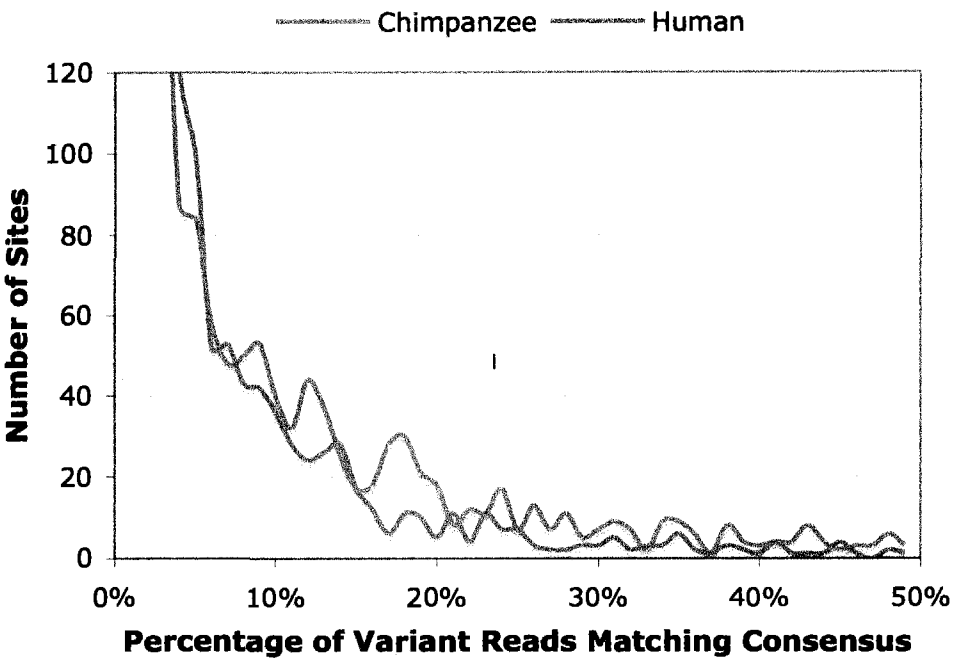

Figure 3.3: Comparison of Human and Chimpanzee variant frequencies. For each verified polymorphism, the frequency of that polymorphism was determined as a percentage and rounded off to the nearest whole number. The number of sites with a minority variant of each percentage was plotted against that percentage. Peaks indicate many different variants at different sites share frequencies and suggest groups of variations are all nearly fixed over the same sets of chromosomal arrays.

Potentially, the variability observed within the chimpanzee repeat could result from differences between the 5 known chromosomal arrays, with each array being relatively fixed. Since there are only five arrays, in this case individual variant frequencies should cluster around the frequencies of particular chromosomes. This is examined in figure 3.3 which displays the number of sites with a variant at a particular frequency.

The visible peaks in the graph of variant read frequencies suggested there was some clumping of frequencies and thus possibly fixed, or nearly fixed interchromosomal differences. A chi-square test against a smoothed frequency model was used to test the statistical significance of the peaks. Sites counts were log-transformed, on which scale they decreased linearly with percentage. The estimated log-transformed site counts were smoothed by averaging over a five percentage wide window. The estimate was then transformed into a expected counts, adjusted slightly so the total expected sites equal the total observed sites over the smoothable range, and the observed site counts were then compared using chi-square. Over the entire range, the results did not quite reach statistical significance ($p = 0.051$). When restricted to the range of 4% to 25%, where the visible peaks were, the results were no longer even close to statistical significance. Thus although there is weak evidence for clustering of frequencies, suggesting a possible population substructure, the peaks that suggested fixed interchromosomal difference were not statistically supported and were indistinguishable from chance variation.

Discussion

As with the human sequence, the iterative stochastic alignment process produced a well-supported sequence covering the entire repeat. Support was even better than for the human sequence, with every single 40-base window matched by multiple reads. This presumably resulted from the superior coverage in the chimpanzee sequence; there were more sequences overall and a substantially improved minimum coverage.

The human consensus sequence was derived from five individuals both to produce a more representative sample and to protect the anonymity of the DNA donors. By

contrast, the chimpanzee data derived entirely from one common chimpanzee (Chimpanzee Sequencing and Analysis Consortium, 2005). This makes exact comparisons difficult. However, the resulting biases are often unidirectional, biasing the chimpanzee sequence to lower polymorphism and less polymorphic chromosomal arrays. It is thus often possible to demonstrate the chimpanzee sequence shares certain characteristics with the human sequence, or has a stronger or weaker expression of the characteristics.

Using resulting sequence information, we are also now able to address the questions of the study:

1) Can iterative stochastic alignment successfully extract a sequence without any sample sequences?

The chimpanzee sequence was successfully derived from a single 300 bp region of the human sequence. Simple extension proved inadequate but the overhang methodology successfully identified and assembled those regions missed by the initial extension process. The final result proved very similar to the human sequence in spite of the extremely limited start seed and the partial intermediates with large deletions.

2) Does the chimpanzee rDNA repeat occur in polymorphic chromosomal arrays like the human rDNA repeat?

The chimpanzee rDNA repeat shares extensive rare minority variation and a lack of significant clustering with the human rDNA repeat. Both these pieces of evidence indicate the chimpanzee rDNA repeat has polymorphic chromosomal arrays. The chimpanzee has even fewer bivariate polymorphic pairs (pairs of polymorphic sites with

complete linkage disequilibrium) as well, although this may reflect better sampling rather than an actual difference.

3) How does chimpanzee variability compare to human variability?

Polymorphism was more frequent in the chimp data both in terms of sites affected and average read divergence. The individual chimpanzee reads varied from the consensus at 0.64% of bases and verifiable polymorphisms were identified at 1266 sites. By contrast, human reads varied from the consensus only at 0.49% of bases and there were only 1075 sites with verifiable polymorphisms (X^2 = 715,064 with 2 d.f., $p < 0.000001$ for percent variantion;X^2 = 14.4 with 2 d.f., p = .007 for number of variant sites). The data set was actually biased to find the reverse result, since the chimpanzee data came from only one individual, while the human data came from five. From this we can conclude the chimp lineage actually has greater variability than the human. This could plausibly result from the greater population size of the chimp, which might produce longer-lived variations.

4) Does the variability demonstrated in the human ribosomal repeat arise from a high mutation rate or a long polymorphism sojourn time?

If the median sojourn time of a polymorphism were similar to the species divergence times then approximately 50% of sequence changes would still be polymorphic within species. However, in only 12% of changes was a polymorphism demonstrable in the chimpanzee. This indicates the sojourn duration is relatively limited and substantially less than the time since the human-chimpanzee divergence. This is further supported by the relatively small average # of sites where individual reads differed from the consensus

(0.64% in chimpanzee and 0.49% in human) compared to the differences between the sequences (8.5%). We can conclude that the high polymorphism in both species is of relatively recent origin.

The sojourn time data allows us to address the cause of elevated polymorphism in the rDNA repeat. Assuming largely neutral evolution, higher polymorphism could be caused either by an elevated mutation rate or by an extended sojourn time for polymorphism due to the need for a new mutant to spread through all loci on the arrays on all chromosomes. The *Pan troglodytes* population size is already large enough that it shares alleles at nuclear loci with *P paniscus* (Deinard and Kidd, 2000.) Human-chimp divergence is estimated at 6 mya in contrast to 2 mya for the common and bonobo chimpanzee (Chen and Li, 2001), so a tripling of expected population size would result in incomplete lineage assortment between human and chimpanzee, at odds with this data. Hence at least part of the increased polymorphism must result from elevated mutation rates.

5) Is there detectable selection on repeat length in humans and chimpanzees?

The human and chimpanzee consensus sequence differ by only 360 bp in length spite of several large indels having occurred during their evolution, suggesting some stabilizing selection operating on sequence length. Nonetheless, resampling indicated the observed difference was not significantly less than expected from a random sample of the observed indel lengths. Similar results were obtained for both human variants. From this we find no evidence for selection of repeat lengths in the human and chimpanzee lineages.

6) Is there differential selection for or against particular sequences in the human and chimpanzee lineages?

There are significant differences between fixation rates with particular sequences in the human-chimpanzee divergence. Some of the reported divergences are false positives expected from the statistical test used but the number of excessively divergent sequences markedly exceeds the number expected by chance. However, difference in mutation rates between short sequence motifs are commonly observed (Hile *et al.*, 2000) and comparisons with polymorphism data or with a larger number of species will be necessary to demonstrate selection.

7) Is there an association between pyrimidine content on the forward strand and mutation rates?

As with the human rDNA repeat, large indels are associated with pyrimidine microsatellite complexes on the forwards strand. The chimpanzee microsatellite complex 3 has a ct content similar to the first two microsatellite complexes, and as predicted by the association, also has a comparable indel polymorphism rate. However, since within both humans and chimpanzees only one of the 8 microsatellite complexes (human complex 3) has a reduced forward strand pyrimidine content more data from other great ape species is need for a conclusive result.

Further Directions

The techniques could be applied to other species and potentially other repeat regions. Highly repetitive regions remain understudied due to the difficulty in getting sequence

information from them. This could extend the genomic era to cover all sequence, and not merely single-copy sequence.

Future applications of this technique to other species will permit much more detailed understanding of the forces driving the evolution of the rDNA repeat. With a large number of related species methods are available to correct for mutation rates while determining the presence of selection (Asthana *et al.*, 2007). A larger set will also provide more power for detecting selection on length.

This works may also be formalized, using the approach of Meyer et. al. proposed for determining sequences in moderately repetitive areas. An advantage of this approach will be a defined mathematical foundation for the results. A project to accomplish this is currently underway.

The sequence shortfall in the 5-fold repeat after the transcribed region is not completely explained. The shortfall could potentially result from excess copies in the consensus. Attempts to determine alternate paths through the tandem repeat region with the technique described here failed. More precise techniques are being developed to deal with these more recalcitrant regions.

In conclusion, Iterative Stochastic Alignment was shown to be a general procedure able to produce sequence information for highly repetitive regions in the absence of any initial sample sequences. The intergenic region of the chimpanzee rDNA consensus shares all the major landmarks of the human sequence, including the tandem repeat immediately after the transcribed region, and the 4 large microsatellite complexes. The conservation of these features suggests a possible functional or evolutionary role for

them. The high variability was shown to result primarily from an elevated mutational rate. Sojourn times are substantially shorter than either the divergence time for the two species or than predicted by a simple model of stochastic transfer of individual variants from parents to offspring. Similar variation rates between the two species, in spite of the marked differences in effective population size, suggest a role for selection in the short sojourn durations.

Methods

Methods were as described for Chapter 2 except that the nearly complete mate pair information allowed reads whose mate pairs did not align to the repeat to be excluded from variability analysis and identification of internal repeats. Such reads were used to exclude intrusive sequences, identify bridges, and indication potential intrusive sequences. The consensus was determined with the seqeuences excluded.

References

Alkan C., Eichler E.E., Bailey, J.A., Sahinalp S.C., and Tüzün, E. 2004. The role of unequal crossover in alpha-satellite DNA evolution: A computational analysis, *J Comp Biology* **11**:933-944

Asthana, Saurabh; Roytberg, Mikhail; Stamatoyannopoulos, John, Sunyaev, Shamil. 2007. Analysis of sequence conservation at nucleotide resolution. *PLOS* **3**:2559-2568

Averback, K.T. and Eickbush, T.H. 2005. Monitoring the mode and tempo of concerted evolution in the *Drosophila melanogaster* rDNA locus 2005. *Genetics* **105**:171:1837-1846

Chen F.C., Li W.H. 2001. Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees. *Am. J. Hum. Genet.* **68**:444–56

Chimpanzee Sequence and Analysis Consortium. 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437**:69–8

Clark, A. 1997. Neutral behavior of shared polymorphism. *PNAS* **94**:7730-7734

de Cicco, D.V. and Glover, D.M. 1983. Amplification of rDNA and type I sequences in drosophila males deficient in rDNA. *Cell* **32**:1217-1225

Delany, M E. 2000. Patterns of ribosomal gene variation in elite commercial chicken pure line populations. *Animal Genetics* **31**:110-116

Deinard AS and Kidd KK, Identifying conservation units within captive chimpanzee populations. Am. J. Phys. Anthropol. 111 (2000), pp. 25–44

Dover G.A. 1986. Molecular drive in multigene families: How biological novelties arise, spread and are assimilated. *Trends Genet* **2**:159–165

Ganley, A.R.D. and Kobayashi, T. 2007. Highly efficient concerted evolution in the ribosomal DNA repeats: Total rDNA repeat variation revealed by whole-genome shotgun sequence data. *Genome Res.* **17**: 184-191

Gonzalez, I.L., Sylvester J.E. 1995. Complete sequence of the 43-kb human ribosomal DNA repeat: analysis of the intergenic spacer. *Genomics* **27**:320-328

Granadino,B., Penalva,L.O.F., and Sanches,L. 1996. Indirect evidence of alteration in the expression of the rDNA genes in interspecific

hybrids between *Drosophila melanogaster* and *Drosophila simulans*. *Mol. Gen. Genet.* **250**: 89-96

Grozdanov, P., Georgiev, O., and Karagyozov, L. 2003. Complete sequence of the 45-kb mouse ribosomal repeat: analysis of the intergenic spacer *Genomics* **82**: 637-643

Harpending, H.C., Sherry, S.T., Rogers, A.R. and Stoneking, M. 1993. The genetic structure of ancient human populations. *Current Anthropology* **34**: 483-496.

Henderson A.S., Warburton D, and Atwood K.C. 1972. Location of ribosomal DNA in the human chromosome complement. *PNAS* **69**:3394-3398

Hile, S.E., Yan, G., Eckert, K.A. 2000. Somatic mutation rates and specificities at TC/AG and GT/CA microsatellite sequences in nontumorigenic human lymphoblastoid cells *Cancer Research* **60**: 1698-1703

Hillis, D.M., Moritz, C, Porter, C.A., and Baker, R.J. 1991.Evidence for biased gene conversion in concerted evolution of ribosomal DNA. *Science* **251**:308–310.

Kaeberlein, M., McVey, M., and Guarente, L. 1999. The SIR2/3/4 complex and SIR2 alone promote longevity in Saccharomyces cerevisiae by two different mechanisms. *Genes Dev.* **13**: 2570–2580.

Kuo, B.A., Gonzalez, I.L., Gillespie, D.A., and Sylvester, J.E. 1996. Human ribosomal RNA variants from a single individual and their expression in different tissues. *Nuc. Acids Res.* **24**:4817-4824

Mostoslavsky *et al.* 2006. Genomic instability and aging-like phenotype in the absence of mammalian SIRT6. *Cell* **124**:315-329.

NCBI BLAST http://www.ncbi.nlm.nih.gov/BLAST/Blast.cgi

Pons, J. and Gillespie, R.G. 2003. Common origin of the satellite DNAs of the Hawaiian spiders of the genus *Tetragnatha*: evolutionary constraints on the length and nucleotide composition of the repeats. *Gene* **313**:169-177.

Sinclair, D.A. and Guarente, L. 1997. Extrachromosomal rDNA cicles – a cause of aging in yeast. *Cell* **91**:1033-1042

Strehler, B and Johnson, R. 1972. 30 percent decrease in ribosomal DNA dosage during aging of dog brain. *Federation Proceedings* **31**:910

Stephan, W. and Cho, S. 1994. Possible role of natural selection in the formation of tendem-repetitive noncoding DNA. *Genetics* **136**:333-341

Stephan, W., Wiehe, T.H.E., and Lenz, M.W. 1992. The effect of strongly selected substitutions on neutral polymorphism – analytic results based on diffusion-theory. *Theo. Pop. Bio.* **41**:237-254

Venter, Craig *et al*, 2001. "The Human Genome" *Science* **291**:1304-1351

White, M.J.D. 1978. Modes of Speciation. W. H. Freeman and Company, San Francisco

Wood, J.G., Rogina, B., Lavu, S., Howitz, K., Helfand, S.L., Tatar,M., and Sinclair, D. 2004. Sirtuin activators mimic caloric restriction and delay ageing in metazoans. *Nature* **430**: 686–689.

Conclusions

The Iterative Stochastic Alignment technique was able to address the hypotheses raised in the introduction. Extensive and well-supported sequence information was determined for both species' rDNA repeats. The information disconfirmed the primary hypothesis of a role of somatic transposition involving the rDNA in chronic diseases, as discussed below under somatic transcription. Answers were derived for the secondary hypotheses mentioned in the introduction: source of variability, homogeneity of arrays, presence of frequent large-scale rearrangements, usefulness of sample sequences, and specific causes of mutation and evolution in the human-chimpanzee species pair.

Method functionality

The functionality of the Iterative Stochastic Alignment method is strongly supported by these results. Consensus sequences were found for both human and chimpanzee populations in spite of high variability and unexpected shortfalls in read coverage in certain region. The shortfalls occurred in the same places in both species, near transcribed areas subject to RNA processing, except for one shortfall only in the chimpanzee. Notably, the chimpanzee sequence, generated de novo from a small seed in the human sequence, had all the landmarks found in the human sequence. The extension method proved robust even in the microsatellite repeat complexes. It successfully navigated all 4 complexes in spite of lengths considerably exceeding the longest available read, low complexity (85%+ ct content) and highly similar composition in the first 3 repeats.

The overhang detection and topological analysis also successfully resolved errors resulting from single large-scale duplications. The current method, however, may not be

robust to multiple duplications with a total length exceeding a long read (approximately 500 bp). In the human sequence, the 3-fold repeat consensus occurred in a different order from either of the sample repeats. In the chimpanzee consensus, the homologous section had 5 repeats. These differences may reflect actual changes or may be artifacts of misassembly. Future assemblies with more exacting match requirements may help resolve this issue.

Effects of somatic transposition

The study indicated mobile DNA elements do not play a particularly important role in the evolution of the rDNA repeat. *S. cerevisiae* and *D. melanogaster* have rDNA repeats with active insertion/deletion events, and in *S. cerevisiae* these events at least partially cause aging. The ISA methodology would identify such events with chimeric reads containing partly sequence from the rDNA repeat sequence and partly sequence from dispersed repetitive DNA (indicating an insertion) or chimeras with sequence from two different regions of the rDNA repeat (indicating a deletion).

Only 9 reads in the human data contained were chimeras suggestive of insertions. None of the possible inserts were observed on more than one read, suggesting none were germline insertions. In addition, there were 29 chimeras to nonrepetitive or nonhuman sequence, suggesting that the observed chimeras might be artifactual. Insertions of repetitive elements into the rDNA array, whether somatic or germline must currently be rare.

There were 303 reads possibly suggestive of rearrangements within the rDNA repeat, possibly including insertions or deletions. Here multiple reads often suggested the same

rearrangement, in total indicating 34 rearrangements. However, these rearrangements do not appear characteristic of deleterious episome escape as in the *S. cerevisiae* deletions. All occurred in highly repetitive CT microsatellite sequence, to the point that identifying the junction was difficult. In 4 cases a modified ISA extension procedure identified the actual cause of the rearrangement and in all 4 case it was an alternate sequence in the microsatellite complex rather than an insertion/deletion event. The most plausible explanation for the data is thus multiple versions of the generally highly polymorphic microsatellite sequences in different germline repeats and not currently active transposition. This is backed up by the fact that the two available sample sequences have different versions of the microsatellite complex one repeat.

Thus, even though the samples were obtained from differentiated blood cells, minimal evidence of active DNA transposition was found. This indicates humans and chimpanzees differ from *S. cerevisiae* and *D. melanogaster* in having somatically quiescent rDNA repeats. The human and chimpanzee rDNA repeats are unlikely to provide the necessary mutational base for aging as with *S. cerevisiae*.

Source of Variability

As established in the discussion for chapter 3, the variability is not much higher for the chimpanzee than for the human lineage, in spite of the considerably larger historical effective population size for the chimpanzee. This suggests the main limit to variability is not the population size. Accordingly, positive selection for some lineages is the straightforward explanation for typical sojourn times of variant rDNA alleles. The selection need not be strong. With a interspecies divergence of 8.5% but a intraspecific

variability of only 0.49% in the human lineage, the average existing duration of a sojourn variant is estimated as 0.49%/(8.5%/2)*6 my = 690,000 years in the human lineage and 900,000 years in the chimpanzee lineage.

Although selection provides the simplest explanation for the sojourn time of an rDNA repeat, there is no direct evidence for selection in the available data. Strong stabilizing selection for repeat length is excluded. Differing fixation rates with short sequences is compatible with selection but does not demonstrate it as it can also result from differing mutation rates. Distinguishing selection from variable mutation rates can be done with analyses of minority frequency spectra but this would require exact estimates of variant frequencies, with effects of sequencing error eliminated. More extensive sampling of the rDNA repeat would provide the necessary replication to overcome this obstacle.

The estimated average duration of a human rDNA lineage, however, is considerably longer than the estimated duration of the human species (250,000 years). A geographic survey of human rDNA variation thus may test contentious theories of genetic exchange between *Homo sapiens* and other *Homo* species during the *sapiens* expansion. Notably, significant genetic exchange with a species with a differentiated geographic range, such as *H. neanderthalensis*, would be expected to leave divergent sequences with a similar geographic distribution. The large number of rDNA repeats also creates many opportunities for genetic exchange in hybrids. Geographic survey of rDNA variability may thus be useful in determining the genetic history of the human species.

Homogeneity of chromosomal arrays

Indirect evidence for both the human and chimpanzee rDNA arrays indicates they are highly polymorphic, with most individual repeats having few or no exact copies. This differs from the results of Ganley and Kobayashi, 2007, for certain fungal arrays, which were confirmed using ISA techniques for one of the species they examined, *S. paradoxus*. However, the polymorphisms in humans and chimpanzees is compatible with Ganley and Kobayashi's hypothesis that their results might be restricted to species like the ones they studiesd with highly inbred population structures and single rDNA arrays. Humans and chimpanzees, by contrast, are both highly outbred species with multiple arrays. The relative importance of inbred population structure and multiple arrays in rDNA diversity may be addressed soon by applying techniques like those used for this study to other species with various combinations of outbred population structure and multiple chromosomal arrays. The finding in this study that the rDNA arrays do not seem to cluster strongly into groups suggestive of chromosome-specific variants does suggest that outbred population structure may be more important than multiple arrays in producing rDNA diversity.

Presence of large-scale rearrangements

This work found, in contrast to prior work with chromosome painting, that there were no large-scale rearrangements in either the human or the chimpanzee repeats apart from the pyrimidine-associated indels. No rearrangements were found with a reversal of direction, and, where both ends of the insert could be identified none extended beyond a

single microsatellite complex or repeat. If any large rearrangements occur they are between pyrimidine microsatellites and do not include reversals.

Usefulness of individual sample sequences

Currently rDNA repeats are typically described with single sample sequences (Ganley and Kobayashi, 2007). This work indicates that practice is inadequate for at least some species. Both available human sample sequences markedly diverged from the human consensus. In addition, most local variants are rare; in combination with the high sequence variability this indicates most individual repeat sequences are rare. Hence no single sequence can be expected to closely represent the consensus. Multiple sample sequences will normally be required for a representative description of an rDNA repeat.

Specific causes of mutation and evolution

One of the unusual features of the human and chimpanzee rDNA repeats is the large pyrimidine-rich microsatellite complexes. They are very long – as much as 2 kilobases – and very long in comparison to pyrimidine repeats in the sequenced portion of the human genome. In addition, having four such repeats in proximity would not be expected given their rarity. This work shows those feature are found in the chimpanzee repeat as well as the human repeat. In addition, these repeats are hypermutable and subject to a particularly high frequency of large indels, much more so than other regions of the repeat. When they occur, large indels are always associated with a polypyrimidine repeat. Because of the low sequence complexity in the microsatellite repeats, determining exactly which complexes are involved is not generally possible, but in all the cases in which the endpoints of the rearrangement can be determined the indel is within a single repeat. The

cause of this unusual propensity for large indels deserves further investigation and could be related to the existence and strong directional biases of these microsatellites.

The high levels of linkage disequilibrium suggested by the low frequency of quadrivariant pairs in both the human and chimpanzee repeats suggests recombination between divergent repeats is a relatively rare occurrence. A model that can explain this finding is that the primary influences on rDNA evolution are mutation and sequence duplication/gene conversion involving only local sequences. The low frequency of most variants suggests that sequence duplication/loss is fairly infrequent in the germline, occurring at a rate that allows new mutations to occur at a rate similar to the effective duplication rate.

Summary

Iterative Stochastic Assembly, a novel group of techniques, successfully generated a consensus sequence for both the human and the chimpanzee rDNA repeats. The techniques require only whole-genome shotgun reads for the repeat regions and a draft consensus for a 300-bp region, which will be readily available in the highly conserved rRNA transcribed regions. Breakpoints and ends of rearrangements are readily identified, including transposable elements, and in some cases entire inserts can be assembled as well. The technique should be easily adopted for tandem repeats other than the rDNA repeat.

The rDNA repeat in the human-chimpanzee lineage was shown to have a number of unusual evolutionary features. Variation is relatively high and individual chromosomal arrays polymorphic. Effective recombination, although present, is fairly rare, with

mutation comparable in effect to recombination. Sojourn times of individual mutations are relatively short and the distribution suggests an effect of selection, although a specific source of selection cannot be identified. The long C+T-dominated microsatellites are subject to a high frequency of large indels indistinguishable from rDNA intergenic spacer sequence. The ISA techniques permit unique repeat-wide analyses of tandem repeat regions and can both answer existing question and produce new hypotheses about these tandem repeats.

Appendix A: Needleman-Wunsch example

Suppose we wish to align two sequences to obtain the alignment with the best possible score. Examining all possible alignments is not feasible except for the very shortest sequences. However, base-by-base scores determine Needleman-Wunsch alignments as so changing the alignment in one location does not influence the scoring at other locations. In particular, if you have an alignment of the first x bases of one sequence with the first y bases of the other no matter what alignment is done downstream that upstream section will have the same alignment score. We can make a n_1+1 by n_2+1 table (n_1 and n_2 being the lengths of the sequences) with the x,y cell containing the best score for the subalignment of the first x-1 base of one sequence with the first y-1 bases of the other.

The table can be filled out with a simple recursive rule. For any alignment of length x from sequence one and length y from sequence two, there are three possibilities for the last base. First is that the last two bases align; this gives a best score equal to the best score for (x-1,y-1), plus the match or mismatch penalty for aligning base x of sequence one with y of sequence two. Second, sequence one may be gapping, in which case the best score will that for (x,y-1) plus the gap penalty. Third, sequence two may be gapping, in which case the best score will be that for (x-1,y) plus the gap penalty. Viewing the table, then, the best score in cell x,y will be the highest of these three numbers:

1) (x-1,y-1) plus the match/mismatch score for the appropriate bases

2) (x,y-1) plus the gap penalty

3) (x-1,y) plus the gap penalty

The top left cell is for an alignment of nothing with nothing and thus scores zero. The rest of the cells can be filled out recursively. The actual alignment can be extracted by inspection. Start at the bottom right, which indicates where the ends of the sequences (after the last bases) align. Determine which of the three steps was made to get to that square. That tells you the last two bases of the alignment. Proceed to that source square and repeat until the upper left corner is reached.

To align the sequence actc to acgtc using a standard set of NW parameters for closely related sequences (+1 match, -3 mismatch, -5 gap), first you lay out a 5 by 6 matrix (the bases are added here as row and column headers)

Start with 0 in the upper right and fill out the table recursively. The value in (1,1), for example, is the highest of

(0,0) [0] plus the score for a vs. a [match: +1] :1

(1,0) [-5] plus a gap [-5] : -10

(0,1) [-5] plus a gap [-5] : -10

which is 1. The result is:

Table A.1 Sample Needleman-Wunsch Scoring

		a	c	g	t	c
	0	-5	-10	-15	-20	-25
a	-5	1	-4	-9	-14	-19
c	-10	-4	2	-3	-8	-13
t	-15	-9	-3	-1	-2	-7
c	-20	-14	-8	-6	-4	-1

so the alignment score is -1. The alignment is extracted by inspection: the last cell had to be reached by aligning the last two bases (-1= -2+1 > -7-5 or -4-5), which was reached by aligning those two bases, which was reached by a gap in the longer sequence (-3 = 2-5 > -4-3 or -9-5, which was reached by sequential alignments. So the alignment is:

acgtc

ac-tc

Appendix B: Needleman-Wunsch modifications

Gap Extension Penalty Reduction

A standard modification to the simple Needleman-Wunsch algorithm is to cost lengthening gaps less severely than opening gaps. This complicates the algorithm. Without the difference if an alignment is split into two parts each can be scored separately and the overall score is equal to the sum, and that feature is essential to the mechanism of the algorithm. With differing scores for extensions and gap openings the score of part of an alignment can be affected by adjacent regions in that there are now three different states (aligned, the first sequence gapping and the second sequence gapping) and the state at the junction alters the score.

To solve this the gapped Needleman-Wunsch was implemented with 3 tables of pair scores, one for each state. The score in the aligned table at (a+1,b+1) was the best score in any table for (a,b) plus the match/mismatch payoff for that base pairing. The score in each gapping table was the best of: the best score at (a+1,b) in the aligned or other gapping table minus the gap open cost; or the score at (a+1,b) in that gapping table minus the gap extension cost.

Top and left columns were altered for each of the table to exclude alignment beginnings incompatible with that table's state by adding an additional "impossible" penalty of -10000. The top left cell corresponds to the sequences starting aligned and was not altered for any table. The remainder of the top row corresponds to starts in which the top sequence is gapping and had the impossibility penalty added for the aligned and left

sequence gapping table. Likewise the remainder of the left column corresponds to alignment starts where the left sequence begins gapping and the impossibility penalty was added in the aligned and top sequence gapping table. Although the top left cell itself is technically impossible for either gapping table, in neither gapping table are the calculations altered by that value. Programmatically it is easier to set up the initial column and row if that does not receive an impossibility penalty in any table, so that score was not altered.

For an example, consider the alignment of gtgaacaaaaagtg with gtgacagtg. If gap extensions get the same cost and gap openings, the alignment will be:

```
gtgaacaaaaagtg
gtg-ac----agtg
```

However, since a single indel and a point mutation is more probable biologically than two separate indels, a more likely alignment is:

```
gtgaacaaaaagtg
gtg-----acagtg
```

The three alignment tables that produce this alignment are as follows. The bold number indicate the path taken to construct the alignment.

Table B.1 Sample Modified Needleman-Wunsch Aligned Scoring

		g	t	g	a	c	a	g	t	g
	0	-100000	-100001	-100002	-100003	-100004	-100005	-100006	-100007	-100008
g	-100000	**1**	-8	-5	-10	-11	-12	-9	-14	-11
t	-100001	-8	**2**	-7	-8	-9	-10	-11	-8	-13
g	-100002	-5	-7	**3**	-6	-7	-8	-5	-10	-7
a	-100003	-10	-8	-6	**4**	-5	-2	-7	-8	-9
a	-100004	-11	-9	-7	-1	1	0	-5	-6	-7
c	-100005	-12	-10	-8	-6	0	-2	-3	-8	-9
a	-100006	-13	-11	-9	-3	-5	1	-5	-6	-10
a	-100007	-14	-12	-10	-4	-6	-4	-2	-7	-8
a	-100008	-15	-13	-11	-5	-7	-5	-7	-5	-10
a	-100009	-16	-14	-12	-6	-8	-6	-8	-10	-8
a	-100010	-17	-15	-13	-7	-9	-7	-9	-11	-13
g	-100011	-14	-16	-10	-12	-10	-12	-6	-12	-10
t	-100012	-19	-13	-15	-13	-11	-13	-11	-5	-14
g	-100013	-16	-18	-12	-14	-12	-14	-8	-14	-4

Table B.2 Sample Modified Needleman-Wunsch Top Gapped Scoring

		g	t	g	a	c	a	g	t	g
	0	-5	-6	-7	-8	-9	-10	-11	-12	-13
g	-100000	-10	-4	-5	-6	-7	-8	-9	-10	-11
t	-100001	-11	-9	-3	-4	-5	-6	-7	-8	-9
g	-100002	-12	-10	-8	-2	-3	-4	-5	-6	-7
a	-100003	-13	-11	-9	-7	-1	-2	-3	-4	-5
a	-100004	-14	-12	-10	-8	-6	-4	-5	-6	-7
c	-100005	-15	-13	-11	-9	-7	-5	-6	-7	-8
a	-100006	-16	-14	-12	-10	-8	-9	-4	-5	-6
a	-100007	-17	-15	-13	-11	-9	-10	-9	-7	-8
a	-100008	-18	-16	-14	-12	-10	-11	-10	-11	-10
a	-100009	-19	-17	-15	-13	-11	-12	-11	-12	-13
a	-100010	-20	-18	-16	-14	-12	-13	-12	-13	-14
g	-100011	-21	-19	-17	-15	-13	-14	-13	-11	-12
t	-100012	-22	-20	-18	-16	-14	-15	-14	-15	-10
g	-100013	-23	-21	-19	-17	-15	-16	-15	-13	-14

Table B.2 Sample Modified Needleman-Wunsch Left Gapped Scoring

		g	t	g	a	c	a	g	t	g
	0	-100000	-100001	-100002	-100003	-100004	-100005	-100006	-100007	-100008
g	-5	-10	-11	-12	-13	-14	-15	-16	-17	-18
t	-6	-4	-9	-10	-11	-12	-13	-14	-15	-16
g	-7	-5	-3	-8	-9	-10	-11	-12	-13	-14
a	-8	-6	-4	-2	-7	-8	-9	-10	-11	-12
a	-9	-7	-5	-3	-1	-6	-7	-8	-9	-10
c	-10	-8	-6	-4	-2	-4	-5	-9	-10	-11
a	-11	-9	-7	-5	-3	-5	-6	-8	-11	-12
a	-12	-10	-8	-6	-4	-6	-4	-9	-10	-11
a	-13	-11	-9	-7	-5	-7	-5	-7	-11	-12
a	-14	-12	-10	-8	-6	-8	-6	-8	-10	-13
a	-15	-13	-11	-9	-7	-9	-7	-9	-11	-13
g	-16	-14	-12	-10	-8	-10	-8	-10	-12	-14
t	-17	-15	-13	-11	-9	-11	-9	-11	-13	-15
g	-18	-16	-14	-12	-10	-12	-10	-12	-10	-15

As with the simpler algorithm, the alignment is extracted by starting at the bottom right cell: the cell with the highest score of the three tables is used. The alignment is the determined by going back through the table to determine which prior cell each highest-scoring cell was derived from. This can, and usually does, involve switching between tables, which reflects that the alignment changes state between aligned and gapping as you back through it.

Aligning subregions

Sometimes the alignment algorithm is used in regions where the adjacent alignment is known. As an example, suppose we align the sequence g with the sequence gg. Either

g-
gg

or

-g
gg

would be equally plausible, and receive equal scores, a priori. However, if this region were embedded in a larger sequence one or the other would be more plausible if there were a gap already on that side. The score thus has to be adjusted for the effects of possible continuing gaps on either side.

To obtain the appropriate result, the state of the adjacent alignment is used to modify the scoring tables. If either sequence to the left of the region to be aligned is gapping then the initial score of the appropriate row or column is set to the gap extension penalty rather than the gap open penalty. If either sequence to the right of the region to be aligned is gapping then a gap open penalty is added to the bottom right cells in the aligned score table and the score table for the other sequence gapping, to reflect the gap penalty that will be imposed when the entire region is scored.

Alignment of partial sequences

In generating consensus sequences from raw read information, reads are aligned against drafts of the consensus. Due to size restrictions on sequencing, any given read is always a small portion of the entire sequence. This produces artifact on alignment since the "gaps" beyond the read are counted against the score and so maximization stretches the obtained sequence out in pathological ways. As an example, consider aligning cagtg to caaaaaaaaaaaagtg. Application of the standard NW algorithm produces

```
c------------agtg
caaaaaaaaaaaagtg
```

when based on knowledge of the sequence process we can conclude it is far more likely that the first c in the second sequence is most likely an error or variant and the correct alignment is:

```
------------cagtg
caaaaaaaaaaaaagtg
```

With longer sequences and high-error read ends, the results become extremely pathological, with small aligned areas of only a few bases scattered over the length of the consensus separated by implausibly long gaps. To correct this, ends of read sequences are identified as "loose ends", produced effectively by cleavage, and potentially aligning to any location within the other sequence rather than to the end. The implementation is that the gap extension cost is set to 0 at the edges of particular rows and columns of the gapping tables. If the left end of either sequence is loose, the gap extension cost is set to 0 in the upper row of the top sequence gapping table and the left column of the left sequence gapping table. If the right end of either sequence is loose, the gap extension cost is set to 0 in the lower row of the top sequence gapping table and the right column of the left sequence gapping table. This results in gaps of arbitrary length at the beginning or termination of sequence carrying no penalty when the start location at that end is not known.

Appendix C: Human Consensus Sequence

```
   1 gctgacacgctgtcctctggcgacctgtcgctggagaggttgggcctccggatgcgcgcggggctctggc
  71 ctaccggtgacccggctagccggccgcgctcctgcttgagccgcctgccggggcccgcgggcctgctgtt
 141 ctctcgcgcgtccgagcgtcccgactcccggtgccggcccgggtccgggtctctgacccacccgggggcg
 211 gcggggaaggcggcgagggccaccgtgcccccgtgcgctctccgctgcgggcgcccggggcggccgcgac
 281 aaccccaccccgctggctccgtgccgtgcgtgtcaggcgttctcgtctccgcggggttgtccgccgcccc
 351 ttccccggagtgggggttggccggagccgatcggctcgctggccggccggccggcctccgctcccgggg
 421 ggctcttcgtgatcgatgtggtgacgtcgtgctctcccgggccgggtccgagccgcgacgggcgaggggc
 491 ggacgttcgtggcgaacgggaccgtccttctcgctccgccccgcgggggtcccctcgtctctcctctccc
 561 cgcccgccggcggtgcgtgtgggaaggcgtggggtgcggaccccggcccgacctcgccgtcccgcccgcc
 631 gccttctgcgtcgcgggtgcgggccggcggggtcctctgacgcggcagacagccctcgctgtcgcctcca
 701 gtggttgtcgacttgcgggcggcccccctccgcggcggtgggggtgccgtcccgccggcccgtcgtgctg
 771 ccctctcgggggtttgcgcgagcgtcggctccgcctgggcccttgcggtgctcctggagcgctccgggt
 841 tgtccctcaggtgcccgaggccgaacggtggtgtgtcgttcccgcccccggcgcccccctcctccggtcgc
 911 cgccgcggtgtccgcgcgtgggtcctgagggagctcgtcggtgtggggttcgaggcggtttgagtgagac
 981 gagacgagacgcgcccctcccacgcggggaagggcgcccgcctgctctcggtgagcgcacgtcccgtgct
1051 cccctctggcgggtgcgcgcgggccgtgtgagcgatcgcggtgggttcgggccggtgtgacgcgtgcgcc
1121 ggccggccgccgaggggctgccgttctgcctccgaccggtcgtgtgtgggttgacttcggaggcgctctg
1191 cctcggaaggaaggaggtgggtggacgggggggcctggtggggttgcgcgcacgcgcgcaccggccgggc
1261 ccccgccctgaacgcgaacgctcgaggtggccgcgcgcaggtgtttcctcgtaccgcagggcccctccc
1331 ttccccaggcgtccctcggcgcctctgcgggcccgaggaggagcggctggcgggtgggggagtgtgacc
1401 caccctcggtgagaaaagccttctctagcgatctgagaggcgtgccttgggggtaccggatccccgggc
1471 cgccgcctctgtctctgcctccgttatggtagcgctgccgtagcgacccgctcgcagaggaccctcctcc
1541 gcttcccccctcgacggggttgggggggagaagcgagggttccgccggccaccgcggtggtggccgagtgc
1611 ggctcgtcgcctactgtggcccgcgcctcccccttccgagtcgggggaggatcccgccgggccgggcccg
1681 gcgtcccagcgggttgggacgcggcggccggcgggcggtgggtgtgcgcgcccggcgctctgtccggcgc
1751 gtgactccctccgccgcgagtcggctctccgcccgctcccgtgccgagtcgtgaccggtgccgacgaccg
1821 cgtttgcgtggcacggggtcgggcccgcctggccctgggaaagcgtcccacggtgggggcgcgccggtct
1891 cccggagcgggaccgggtcggaggatggacgagaatcacgagcgacggtggtggtggcgtgtcgggttcg
1961 tggctgcggtcgctccggggcccccggtggcggggccccggggctcgcgaggcggttctcggtgggggcc
2031 gagggccgtccggcgtcccaggcggggcgccgcgggaccgccctcgtgtctgtggcggtgggatcccgcg
2101 gccgtgttttcctggtggcccggccgtgcctgaggtttctccccgagccgccgcctctgcgggctcccgg
2171 gtgcccttgccctcgcggtccccggccctcgcccgtctgtgccctcttccccgcccgccgcccgccgatc
2241 ctcttcttcccccgagcggctcaccggcttcacgtccgttggtggccccgcctgggaccgaacccggca
2311 ccgcctcgtggggcgccgccgccggccactgatcggcccggcgtccgcgtccccggcgcgcgccttggg
2381 gaccgggtcggtggcgccccgcgtggggcccggtgggcttcccggagggttccgggggtcggcctgcggc
2451 gcgtgcgggggaggagacggttccgggggaccggccgcgactgcggcggcggtggtgggggagccgcgg
2521 ggatcgccgagggccggtcggccgccccgggtgccgcgcggtgccgccggcggcggtgaggccccgcgcg
2591 tgtgtcccggctgcggtcggccgcgctcgaggggtccccgtggcgtccccttccccgccggccgcctttc
2661 tcgcgccttccccgtcgccccggcctcgcccgtggtctctcgtcttctcccggcccgctcttccgaaccg
2731 ggtcggcgcgtcccccgggtgcgcctcgcttcccgggcctgccgcggcccttccccgaggcgtccgtccc
2801 gggcgtcggcgtcggggagagcccgtcctccccgcgtggcgtcgccccgttcggcgcgcgcgtgcgcccg
2871 agcgcggcccggtggtccctcccggacaggcgttcgtgcgacgtgtggcgtgggtcgacctccgccttgc
2941 cggtcgctcgccctctccccgggtcggggggtggggcccgggccggggcctcggccccggtcgcggtccc
3011 ccgtcccgggcgggggcgggcgcgccggccggcctcggtcgccctcccttggccgtcgtgtggcgtgtgc
3081 cacccctgcgcccgcgcccgccggcggggctcggagccgggcttcggccgggccccgggccctcgaccgg
3151 accggtgcgcgggcgctgcggccgcacggcgcgactgtccccgggccgggcaccgcggtccgcctctcgc
3221 tcgccgcccggacgtcggggccgccccgcggggcgggcggagcgccgtccccgcctcgccgccgcccgcg
3291 ggcgccggccgcgcgcgcgcgcgcgtggccgccggtccctcccggccgccgggcgcgggtcgggccgtcc
```

3361 gcctcctcgcgggcgggcgcgacgaagaagcgtcgcgggtctgtggcgcggggccccggtggtcgtgtcg
3431 cgtgggggggcgggtggttggggcgtccggttcgccgcgccccgccccggccccaccggtcccggccgccg
3501 cccccgcgcccgctcgctccctcccgtccgcccgtccgcggcccgtccgtccgtccgtccgtcgtcctcc
3571 tcgcttgcggggcgccgggcccgtcctcgcgaggcccccggccggccgtccggccgcgtcggggcctcg
3641 ccgcgctctaccttacctacctggttgatcctgccagtagcatatgcttgtctcaaagattaagccatgc
3711 atgtctaagtacgcacggccggtacagtgaaactgcgaatggctcattaaatcagttatggttcctttgg
3781 tcgctcgctcctctcctacttggataactgtggtaattctagagctaatacatgccgacgggcgctgacc
3851 cccttcgcgggggggatgcgtgcatttatcagatcaaaaccaacccggtcagcccctctccggccccggc
3921 cgggggcgggcgccggcggctttggtgactctagataacctcgggccgatcgcacgccccccgtggcgg
3991 cgacgacccattcgaacgtctgccctatcaactttcgatggtagtcgccgtgcctaccatggtgaccacg
4061 ggtgacggggaatcagggttcgattccggagagggagcctgagaaacggctaccacatccaaggaaggca
4131 gcaggcgcgcaaattacccactcccgacccggggaggtagtgacgaaaaataacaatacaggactctttc
4201 gaggccctgtaattggaatgagtccacttttaaatcctttaacgaggatccattggagggcaagtctggtg
4271 ccagcagccgcggtaattccagctccaatagcgtatattaaagttgctgcagttaaaaagctcgtagttg
4341 gatcttgggagcgggcgggcggtccgccgcgaggcgagccaccgcccgtccccgccccttgcctctcggc
4411 gccccctcgatgctcttagctgagtgtcccgcggggcccgaagcgtttactttgaaaaaattagagtgtt
4481 caaagcaggcccgagccgcctggataccgcagctaggaataatggaataggaccgcggttctattttgtt
4551 ggttttcggaactgaggccatgattaagagggacggccgggggcattcgtattgcgccgctagaggtgaa
4621 attcttggaccggcgcaagacggaccagagcgaaagcatttgccaagaatgttttcattaatcaagaacg
4691 aaagtcggaggttcgaagacgatcagataccgtcgtagttccgaccataaacgatgccgaccggcgatgc
4761 ggcggcgttattcccatgacccgccgggcagcttccgggaaaccaaagtctttgggttccgggggggagta
4831 tggttgcaaagctgaaacttaaaggaattgacggaagggcaccaccaggagtggagcctgcggcttaatt
4901 tgactcaacacgggaaacctcacccggcccggacacggacaggattgacagattgatagctctttctcga
4971 ttccgtgggtggtggtgcatggccgttcttagttggtggagcgatttgtctggttaattccgataacgaa
5041 cgagactctggcatgctaactagttacgcgacccccgagcggtcggcgtcccccaacttcttagagggac
5111 aagtggcgttcagccacccgagattgagcaataacaggtctgtgatgcccttagatgtccggggctgcac
5181 gcgcgctacactgactggctcagcgtgtgcctaccctacgccggcaggcgcgggtaacccgttgaacccc
5251 attcgtgatggggatcggggattgcaattattccccatgaacgaggaattcccagtaagtgcgggtcata
5321 agcttgcgttgattaagtccctgccctttgtacacaccgcccgtcgctactaccgattggatggtttagt
5391 gaggccctcggatcggccccgccggggtcggcccacggccctggcggagcgctgagaagacggtcgaact
5461 tgactatctagaggaagtaaaagtcgtaacaaggtttccgtaggtgaacctgcggaaggatcattaacgg
5531 agcccggagggcgaggcccgcggcggcgccgccgccgccgcgcgcttccctccgcacacccacccccca
5601 ccgcgacgcggcgcgtgcgcgggcggggcccgcgtgcccgttcgttcgctcgctcgttcgttcgccgccc
5671 ggccccgccggccgcgagagccggagaactcgggagggagacgggggagagagagagagagagagagaga
5741 gagaaagaagggcgtgtcgttggtgtgcgcgtgtcgtggggccggcgggcggcggggagcggtccccggc
5811 cgcggccccgacggcgtgggtgtcggcgggcgcgggggcggttctcggcggcgtcgcggcgggtctgggg
5881 ggtctcggtgccctcctcccgccggggcccgtcgtccggccccgccgcgccggctccccgtcttcgggg
5951 ccggccggattcccgtcgcctccgccgcgccgctccgcgccgccgggcacggccccgctcgctctccccg
6021 gccttcccgctagggcgtctcgagggtcgggggccggacgccggtcccctccccgcctcctcgtccgcc
6091 cccccgccgtccaggtacctagcgcgttccggcgcggaggtttaaagacccctttgggggatcgcccgtc
6161 cgcccgtgggtcgggggcggtggtgggcccgcgggggagtcccgtcgggaggggcccggcccctcccgcg
6231 cctccaccgcggactccgctccccggccggggccgcgccgccgccgccgcgcggcggccgtcgggtggg
6301 ggctttacccggcggccgtcgcgcgcctgccgcgcgtgtggcgtgcgccccgcgccgtgggggcgggaac
6371 ccccgggcgcctgtggggtggtgtccgcgctcgcccccgcgtgggcggcgcgcgcctccccgtggtgtga
6441 aaccttccgacccctctccggagtccggtcccgtttgctgtctcgtctggccggcctgaggcaaccccct
6511 ctcctcttgggcgggggggggggggacgtgccgcgccaggaagggcctcctcccggtgcgtcgtcgggag
6581 cgccctcgccaaatcgacctcgtacgactcttagcggtggatcactcggctcgtgcgtcgatgaagaacg
6651 cagctagctgcgagaattaatgtgaattgcaggacacattgatcatcgacacttcgaacgcacttgcggc
6721 cccgggttcctcccggggctacgcctgtctgagcgtcgcttgccgatcaatcgcccccgggggtgcctcc
6791 gggctcctcggggtgcgcggctgggggttccctcgcagggcccgccgggggccctccgtcccctaagcg

6861 cagacccggcggcgtccgccctcctcttgccgccgcgcccgccccttcccctccccccgcgggccctgc
6931 gtggtcacgcgtcgggtggcgggggggagaggggggcgcgcccggctgagagagacggggagggcggcgc
7001 cgccgccgcccgcgaagacggagagggaaagagagagccggctcgggccgagttcccgtggccgccgcct
7071 gcggtccgggttcctccctcgggggctccctcgcgccgcgcgcggctcggggttcggggttcgtcggcc
7141 ccggccgggtggaaggtcccgtgcccgtcgtcgtcgtcgtcgcgcgtcgtcggcggtgggggcgtgttgc
7211 gtgcggtgtggtggtgggggaggaggaaggcgggtccggaaggggaagggtgccggcggggagagagggt
7281 cgggggagcgcgtcccggtcgccgcggttcgccgcccgcccccggtggcggcccggcgtccggccgaccg
7351 ccgctcccctcctccccgccgcccctcctccgaggccccgcccgtcctcctcgccctccccgcgcgtacg
7421 cgcgcgcgcccgcccgcccggctcgcctcgcggcgcgtcggccggggccgggagcccgccccgcggcccg
7491 cccggccgcgcccgtggccgcggcgccggggttcgcgtgtccccggcggcgacccgcgggacgccgcggt
7561 gtcgtccgccgtcgcgcgcccgcctccggctcgcggccgcgccgcgccgcgccggggccccgtcccgagc
7631 ttccgcgtcggggcggggcggctccgccgccgcgtcctcggacccgtccccccgacctccgcgggggaga
7701 cgggtcggggcgtgcggcgcccgtcccgcccccggcccgtgcccctccctccggtcgtcccgctccggcg
7771 gggcggcgcgggggtgccgtcggccgcgcgctctctctcccgtcgcctctcccccctcgccgggcccgtct
7841 cccgacggagcgtcgggcgggcggtcgggccggcgcgattccgtccgtccgtccgccgagcggcccgtcc
7911 ccctccgagacgcgacctcagatcagacgtggcgacccgctgaatttaagcatattagtcagcggaggag
7981 aagaaactaaccaggattccctcagtaacggcgagtgaacagggaagagcccagcgccgaatccccgccc
8051 cgcggcggggcgcgggacatgtggcgtacggaagacccgctccccggcgccgctcgtgggggcccaagt
8121 ccttctgatcgaggcccagcccgtggacggtgtgaggccggtagcggcccccggcgcgccgggcccgggt
8191 cttcccggagtcgggttgcttgggaatgcagcccaaagcgggtggtaaactccatctaaggctaaatacc
8261 ggcacgagaccgatagtcaacaagtaccgtaagggaaagttgaaaagaactttgaagagagagttcaaga
8331 gggcgtgaaaccgttaagaggtaaacgggtggggtccgcgcagtccgcccggaggattcaacccggcggc
8401 gggtccggccgtgtcggcggcccggcggatctttcccgccccccgttcctcccgacccctccacccgccc
8471 tcccttcccccgccgcccctcctcctcctcccccggaggggcgggctccggcgggtgcgggggtgggcgg
8541 gcggggccgggggtggggtcggcgggggaccgtcccccgaccggcgaccggccgccgccgggcgcatttc
8611 caccgcggcggtgcgccgcgaccggctccgggacggctgggaaggcccggcggggaaggtggctcggggg
8681 gccccgtccgtccgtccgtcctcctcctccccgtctccgcccccggccccgcgtcctccctcgg
8751 gagggcgcgcgggtcggggcggcggcggcggcggtggcggcggcggcggcggcgggaccgaaac
8821 cccccccgagtgttacagcccccccggcagcagcactcgccgaatcccggggccgagggagcgagacccg
8891 tcgccgcgctctcccccctcccggcgcccaccccgcggggaatccccgcgagggggtctcccccgcg
8961 ggggcgcgccggcgtctcctcgtgggggggccgggccacccctcccacggcgcgaccgctctcccacccc
9031 tcctccccgcgccccccgcccggcgacgggggggtgccgcgcgcgggtcgggggcggggcggactgtc
9101 cccagtgcgccccgggcgggtcgcgccgtcgggcccgggggaggttctctcggggccacgcgcgcgtccc
9171 ccgaagaggggacggcggagcgagcgcacggggtcggcggcgacgtcggctacccacccgacccgtctt
9241 gaaacacggaccaaggagtctaacacgtgcgcgagtcgggggctcgcacgaaagccgccgtggcgcaatg
9311 aaggtgaaggccggcgcgctcgccggccgaggtgggatcccgaggcctctccagtccgccgagggcgcac
9381 caccggcccgtctcgcccgccgcgccggggaggtggagcacgagcgcacgtgttaggacccgaaagatgg
9451 tgaactatgcctgggcagggcgaagccagaggaaactctggtggaggtccgtagcggtcctgacgtgcaa
9521 atcggtcgtccgacctgggtataggggcgaaagactaatcgaaccatctagtagctggttccctccgaag
9591 tttccctcaggatagctggcgctctcgcagacccgacgcacccccgccacgcagttttatccggtaaagc
9661 gaatgattagaggtcttggggccgaaacgatctcaacctattctcaaactttaaatgggtaagaagcccg
9731 gctcgctggcgtggagccgggcgtggaatgcgagtgcctagtgggccacttttggtaagcagaactggcg
9801 ctgcgggatgaaccgaacgccgggttaaggcgcccgatgccgacgctcatcagaccccagaaaaggtgtt
9871 ggttgatatagacagcaggacggtggccatggaagtcggaatccgctaaggagtgtgtaacaactcacct
9941 gccgaatcaactagccctgaaaatggatggcgctggagcgtcgggcccatacccggccgtcgccggcagt
10011 cgagagtggacgggagcggcgggggcggcgcgcgcgcgcgcgtgtggtgtgcgtcggagggcggcggc
10081 ggcggcggcggcgggggtgtggggtccttcccccgccccccccccacgcctcctccccctcctcccgccc
10151 acgccccgctccccgccccggagccccgcggacgctacgccgcgacgagtaggagggccgctgcggtga
10221 gccttgaagcctagggcgcgggcccgggtggagccgccgcaggtgcagatcttggtggtagtagcaaata
10291 ttcaaacgagaactttgaaggccgaagtggagaagggttccatgtgaacagcagttgaacatgggtcagt

10361 cggtcctgagagatgggcgagcgccgttccgaagggacgggcgatggcctccgttgccctcggccgatcg
10431 aaagggagtcgggttcagatccccgaatccggagtggcggagatgggcgccgcgaggcgtccagtgcggt
10501 aacgcgaccgatcccggagaagccggcgggagccccggggagagttctcttttctttgtgaagggcaggg
10571 cgccctggaatgggttcgccccgagagaggggcccgtgccttggaaagcgtcgcggttccggcggcgtcc
10641 ggtgagctctcgctggcccttgaaaatccgggggagagggtgtaaatctcgcgccgggccgtacccatat
10711 ccgcagcaggtctccaaggtgaacagcctctggcatgttggaacaatgtaggtaagggaagtcggcaagc
10781 cggatccgtaacttcgggataaggattggctctaagggctgggtcggtcgggctggggcgcgaagcgggg
10851 ctgggcgcgcgccgcggctggacgaggcgccgccgccgcccccccacgcccggggcacccccctcgcgg
10921 ccctcccccgccccacccgcgcgcgccgctcgctccctccccacccgcgccctctctctctctctc
10991 ccccgctccccgtcctcccccctcccgggggagcgccgcgtgggggcggcggcgggggagaagggtcg
11061 gggcggcaggggccggcggcggcccgccgcggggccccggcggcgggggcacggtcccccgcgaggggg
11131 cccgggcacccgggggccggcggcggcggcgactctggacgcgagccgggcccttcccgtggatcgccc
11201 cagctgcggcgggcgtcgcggccgcccccggggagcccggcgggcgccggcgcgcccccccccacccca
11271 ccccacgtctcgtcgcgcgcgcgtccgctgggggcggggagcggtcgggcggcggcggtcggcgggcggc
11341 ggggcggggcggttcgtccccccgccctaccccccggccccgtccgcccccgttcccccctcctcctc
11411 ggcgcgcggcggcggcggcggcggcggcaggcggcggaggggccgcgggccggtcccccccgccgggtcc
11481 gcccccggggccgcggttccgcgcggcgcctcgcctcggccggcgcctagcagccgacttagaactggtg
11551 cggaccaggggaatccgactgtttaattaaaacaaagcatcgcgaaggcccgcggcgggtgttgacgcga
11621 tgtgatttctgcccagtgctctgaatgtcaaagtgaagaaattcaatgaagcgcgggtaaacggcgggag
11691 taactatgactctcttaaggtagccaaatgcctcgtcatctaattagtgacgcgcatgaatggatgaacg
11761 agattcccactgtccctacctactatccagcgaaaccacagccaagggaacgggcttggcggaatcagcg
11831 gggaaagaagaccctgttgagcttgactctagtctggcacggtgaagagacatgagaggtgtagaataag
11901 tgggaggcccccggcgcccccccggtgtccccgcgaggggcccggggcggggtccgccggccctgcgggc
11971 cgccggtgaaataccactactctgatcgtttttcactgacccggtgaggcgggggggcgagccccgagg
12041 ggctctcgcttctggcgccaagcgcccggccgcgcgccggccgggcgcgacccgctccggggacagtgcc
12111 aggtggggagtttgactggggcggtacacctgtcaaacggtaacgcaggtgtcctaaggcgagctcaggg
12181 aggacagaaacctcccgtggagcagaagggcaaaagctcgcttgatcttgattttcagtacgaatacaga
12251 ccgtgaaagcggggcctcacgatccttctgacctttgggttttaagcaggaggtgtcagaaaagttacc
12321 acagggataactggcttgtggcggccaagcgttcatagcgacgtcgcttttgatccttcgatgtcggct
12391 cttcctatcattgtgaagcagaattcaccaagcgttggattgttcacccactaatagggaacgtgagctg
12461 ggtttagaccgtcgtgagacaggttagttttaccctactgatgatgtgttgttgccatggtaatcctgct
12531 cagtacgagaggaaccgcaggttcagacatttggtgtatgtgcttggctgaggagccaatggggcgaagc
12601 taccatctgtgggattatgactgaacgcctctaagtcagaatcccgcccaggcggaacgatacggcagcg
12671 ccgcggagcctcggttggcctcggatagccggtcccccgcctgtccccgccggcgggccgccccccctc
12741 cacgcgccccgcgcgcgcgggagggcgcgtgccccgccgcgcgccgggaccggggtccggtgcggagtgc
12811 ccttcgtcctgggaaacggggcgcggccggaaaggcggccgccccctcgcccgtcacgcaccgcacgttc
12881 gtggggaacctggcgctaaaccattcgtagacgacctgcttctgggtcggggtttcgtacgtagcagagc
12951 agctccctcgctgcgatctattgaaagtcagccctcgacacaagggtttgtccgcgcgcgcgcgcgcgcg
13021 tgcgtgcgggggcccggcggggcgtgcgcgtccggcgccgtccgtccttccgttcgtcttcctccctcc
13091 cggcctctcccgccgaccgcgggcgtggtggtgggggtgggggggagggcgcgcgaccccggtcggcgcg
13161 ccccgcttcttcggttcccgcctcctccccgttcaccgccggggcggctcgtccgctccgggccgggacg
13231 gggtccggggagcgtggtttgggagccgcggaggcggccgcgccgagccgggcccgtggcccgccggtcc
13301 ccgtcccgggggttggccgcgcgggccccggtggggcggccacccggggtcccggccctcgcgcgtcctt
13371 cctcctcgctcctccgcacgggtcgaccagcagaccgcgggtggcgggcggcgggcggcgaggccccacg
13441 gggcgtccccgcacccggccgacctccgctcgcgacctctcctcggtcgggcctccggggtcgaccgcct
13511 gccgcccgcgggcgtgagactcagccggcgtctcgccgtgtcccgggtcgaccggcgggccttctccacc
13581 gagcggcgtgtaggagtgcccgtcgggacgaaccgcaaccggagcgtccccgtctcggtcggcacctccg
13651 gggtcgaccagctgccgcccgcgagctccggacttagccggcgcctgcacgtgtcccgggtcgaccagca
13721 ggcggccgccggacgctgcggcgcaccgacgcgagggcgtcgattcccgttcgcgcgcccgcgacctcca
13791 ccggcctcggcccgcggtggagctgggaccacgcggaactccctctctcacatttttttcagccccaccg

13861 cgagtttgcgtccgcgggacttttaagagggagtcactgctgccgtcagccagtaatgcttcctcctttt
13931 ttgcttttaggttttgtcttgccttttttttttttttttttttctttctttctttctttctttctttt
14001 ctttctttctttctttctttctcgctctcgctctctccctccctcgctcgttttctctctctctcgctct
14071 tgccctctcgctctctccctcgctcgttttctctctctctctctctctctctctgtctctcgctctcgcc
14141 ctctctctctctctgtctctctctctctctctctctctctctctctctctctctctgtcgctctcgccct
14211 ctcgctctctctctgtctctgtctgtgtctctctctctccctccctccctccctccctcccctccccttcc
14281 ttggtgccttctcggctcttgagacttagccgctgtctcgccgtgccccgggtcgaccggcgggccttct
14351 ccaccgagcggcgtgtaagagtgcccgtcgggacgagccggacccgccgcgtccccgtctcggtcggcac
14421 ctccggggtcgaccagctgccgcccgcgagctccggacttagccggcgtctgcacgtgtcccgggtcgac
14491 cagcaggcggccgccggacgctgcggcgcaccgacgcgagggcgctgattcccgttcacgcgcccgcgac
14561 ctccaccggcctcggcccgccgtggagctgggaccacgcggaactccctctcctacatttttttcagccc
14631 caccgcgagtttgcgtccgcgggacttttaagagggagtcactgctgccgtcagccagtaatgcttcctc
14701 ctttttgctttaggttttgtcttgccttttttttttttttttttttttctttctttctttctttc
14771 tttctttctttctttctttctttctttctcgctctcgctctctcgctctctccctcgctcgttttc
14841 tttctctttctctctctctctctctctctctctctctctgtctctcgctctcgccctctctctctctc
14911 tttctctctgtctctctctctctctctctctctctctctctctctctctctctccctcccctccctccctc
14981 tctcccctccttggtgccttctcggctcttgacacttagccgctgtctcgccgtgtcccgggtcgaccg
15051 gcgggccttctccaccgagcggcgtgtaagagtgcccgtcgggacgagccggacccgccgcgtccccgtc
15121 tcggtcggcacctccggggtcgaccagctgccgcccgcgagctccggacttagctggcgtctgcacgtgt
15191 cccgggtcgaccagcaggcggccgccggacgctgcggcgcactgacgcgagggcgtcgattccggttcac
15261 gcgccggcgacctccaccggcctcggcccgcggtggagctgggaccacgcggaactccctctcctacatt
15331 tttttcagccccactgcgagtttgcgtccgcgggacttttaagagggagtcactgctgccgtcagccagt
15401 aatgcttcctcctttttgctttttggttttgccttgcgttttctttctttctttctttctttctt
15471 tctttctttctttctttctctctctctctctctctctctctctctgtctctctcccctccctccctccttg
15541 gtgccttctcggctcgctgctgctgctgcctctgcctccacggttcaagcaaacagcaagttttctattt
15611 cgagtaaagacgtaatttcaccattttggccgggctggtctcgaactcccgacctagtgatccgcccgcc
15681 tcggcctcccaaagactgctgggagtacagatgtgagccaccatgcccggccgattccttccttttttca
15751 atcttattttctgaacgctgccgtgtatgaacatacatctacacacacacacacacacacacacacacac
15821 acacacacacacacacacacacccgtagtgataaaactatgtaaatgatatttccataattaatacgtt
15891 tatattatgttacttttaatggatgaatatgtatcgaagccccatttcatttacatacacgtgtatgtat
15961 atccttcctcccttccttcattcattatttattaataatttcgtttatttatttctttttttgggg
16031 ccggcccgcctggtcttctgtctctgcgctctggtgacctcagcctcccaaatagctgggactacaggga
16101 tctcttaagcccgggagggagaggttaacgtgggctgtgatcgcacacttccactccagcttacgtgggc
16171 tgcggtggggtggggtggggtggggtggggtggggtgcagagaaaacgattgattgcgatctcaattgcc
16241 ttttagcttcattcataccctgttatttgctcgtttattctcatgggttcttctgtgtcattgtcacgtt
16311 catcgtttgcttgcctgcttgcctgtttatttccttccttccttccttccttccttcccctccttcc
16381 ttccttccttcccctcccttactggcagggtcttcctctgtctctgccgcccaggatcaccccaacctcaa
16451 cgctttggaccgaccaaacggtcgttctgcctctgatccctcccatccccattacctgagactacaggcg
16521 cgcaccaccacaccggctgacttttatgttgtttctcatgttttccgtaggtaggtatgtgtgtgtgtgt
16591 gtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtatctatgtatgtacgtatgtatgtatgtatgt
16661 gagtgagatgggtttcggggttctatcatgttgcccacgctggtctcgaactcctgtcctcaagcaatcc
16731 gcctgcctgcctcggccgcccacactgctgctattacaggcgtgagacgctgcgcctggctccttctaca
16801 tttgcctgcctgcctgcctgcctgcctgcctgcctgcctgcctgcctgcctatcaatcgtcttcttttta
16871 gtacggatgtgctctcgctttattgtccatgctctgggcacacgtggtctcttttcaaacttctatgatt
16941 attattattgtaggcgtcatctcacgtgtcgaggtgatctcgaacttttaggctccagagatcctcccgc
17011 atcggcctcccggagtgctgtgatgacacgcgtgggcacggtacgctctggtcgtgtttgtcgtgggtcg
17081 gttctttccgtttttaatacggggactgcgaacgaagaaaatttccagacgcatctcaccgatccgcctt
17151 ttcgttctttctttttattctctttagacggagtttcactcttgtcgcccagggtggagtacgatggcgg
17221 ctctcggctcaccgcaccctccgcctcccaggttcaagtgattctcctgcctcagccttcccgagtagct
17291 ggaatgacagagatgagccatcgtgcccggctaatttttctatttttactacagatggggtttctccatc

17361 ttggtcaggctggtcttcaacttccgaccgttggagaatcttaactttcttggtggtggttgttttcctt
17431 tttctttttttttctttttctttttccttctcctccccccccaccccccttgtcgtcgtcctcctcct
17501 cctcctcctcctcctcctcctcctcctcctcctcctcctcctcctttcatttctttcagctgggctctcct
17571 acgtgtgttgctctgttgctcacgctggtctcaaactcctggccttgacgcttctcccgtcacatccgcc
17641 gtctggttgttgaaatgagcatctctcgtaaaatggaaaagatgaaagaaataaacacgaagacggaaag
17711 cacggtgtgaacgtttctcttgccgtctcccggggtgtaccttggacccggaaacacggagggagcttgg
17781 ctgagtgggttttcggtgccgaaacctcccgagggcctccttccctctcccccttgtccccgcttctccc
17851 ccagccgaggctcccaccgccgccctggcattttccataggagaggtatgggagaggactgacacgcctt
17921 ccagatctatatcctgccggacgtctctggctcggcgtgccccaccggctacctgccaccttccagggag
17991 ctctgaggcggatgcgacccccacccccccgtcacgtcccgctaccctccccggctggcctttgccggg
18061 cgaccccaggggaaccgcgttgatgctgccttcggatcctccggcgaagacttccaccggatgccccggg
18131 tgggccggttgggatcagactggaccaccccggaccgtgctgttcttgggggtgggttgacgtacagggt
18201 ggactggcagccccagcattgtaaagggtgcgtgggtatggaaatgtcacctaggatgccctccttccct
18271 tcggtctgccttcagctgcctcaggcgtgaagacaacttcccatcggaacctcttctcttccctttctcc
18341 agcacacagatgagacgcacgagagggagaaacagctcaatagataccgctgaccttcatttgtggaatc
18411 ctcagtcatcgacacacaagacaggtgactaggcagggacacagatcaaacactatttccgggtcctcgt
18481 ggtgggattggtctctctctctctctctctctctctctctctctctctctctcgcacgcgcacgcgcg
18551 cacacacacacacaatttccatatctagttcacagagcacactcacttccccttttcacagtacgcaggc
18621 tgagtaaaacccgccccacccctccacccgttggctgacgaaacccccttctctacaattgatgaaaaagat
18691 gatctgggccgggcacgctagctcacgcctgtcactccggcactttgggaggccgaggcgggtggatcgc
18761 ttggggccgggagttcgagaccaggctggccgacgtggcgaaaccccgtctctctgaaaaatagaacgat
18831 tagccgggcctggtggcgtgggcttggaatcacgaccgctcgggagactggggcgggcgacttgttccaa
18901 ccggggaggccgaggttgcgatgagctgagatcgtgccgtggcgatgcggcctggatgacggagcgagac
18971 cccgtctcgagagaatcatgatgttattataagatgagttgtgcgcggtgatggccgcctgtagtcgcgg
19041 ctactcgggaggctgagacgaggagaagatcacttgaggccccacaggtcgaggcttcggtcggccgtga
19111 cccactgtatcctgggcagtcaccggtcaaggagatatgccccttccccgtttgcttttctttttcttccc
19181 ttctctttttcttcttttgcttctctttttctttctttctttctttctttctttctttctttctttc
19251 tttttctttttctctcttcccctctttctttcctgccttcctgcctttcttctttttcttctttcctccct
19321 tcctcccttccttcttttcctcccgcctcagcctcccaaagtgctgggatgactggcgggaggcaccatgc
19391 ctgcttggcccaaagagaccctcttggaaagtgagacgcagagagcgccttccagtgatctcattgactg
19461 atttagagacggcatctcgctccgtcaccccggcagtggtgccgtcgtaactcactccctgcagcgtgga
19531 cgctcctggactcgagcgatccttccacctcagcctccagagtacagagcctgggaccgcgggcacgcgc
19601 cactgtgcccacaccgtttttaattgtttttttttccccgagacagagtttcactctcgtggcctagac
19671 tgcagtgcggtggcgcgatcttggctcaccgcaacctctgcctcccggtttcaagcgattctcctgcatc
19741 ggcctcctgagtagccgggattgcgggcatgcgctgccacgtctggctgatttcgtatttttagtggaga
19811 cggggcttctccatgtcgatcgggctggtttcgaactcccgacctcaggtgatccgccctccccggcctc
19881 cggaagtgctgggatgacaggcgtgagccaccgcgcccggccttcatttttaaatgttttcccacagacg
19951 gggtctcatcatttctttgcaaccctcctgcccggcgtctcaaagtgctggcgtgacgggcgtgagccac
20021 tgcgcctggactccggggaatgactcacgaccaccatcgctctactgatcctttctttctttctttctttt
20091 ctttctttctttctttctttctttctttctttctttctttcttgatgaattatcttatgatttatttgtgtact
20161 tattttcagacggagtctcgctctgggcggggcgaggcgaggcgaggcacagcgcatcgctttggaagcc
20231 gcggcaacgcctttcaaagccccattcgtatgcacagagccttattcccttcctggagttggagctgatg
20301 ccttccgtagccttgggcttctctccattcggaagctttgacaggcgcaaccccacccagaggctggctg
20371 cggctgaggattaggggtgtgttggggctgaaaactgggtcccctatttttgatacctcagccgacaca
20441 tcccccgaccgccatcgcttgctcgccctctgagatcccccgcctccaccgccttgcaggctcacctctt
20511 actttcatttcttcctttcttgcgtttgaggaggggtgcgggaatgagggtgtgtgtggggaggggtg
20581 cggggtggggacggaggggagcgtcctaagggtcgatttagtgtcatgcctctttcaccaccaccaccac
20651 caccgaagatgacagcaaggatcggctaaataccgcgtgttctcatctagaagtgggaacttacagatga
20721 cagttcttgcatgggcagaacgaggggaccggggacgcggaagtctgcttgagggaggaggggtggaag
20791 gagagacagcttcaggaagaaaacaaaacacgaatactgtcggacacagcactgactacccgggtgatga

20861 aatcatctgcacactgaacacccccgtcacaagtttacctatgtcacaatcttgcacatgtatgcttgaa
20931 cgacaaataaaagttaggggggagaagagaggagagagagagagagagagacagagagagacagagag
21001 agagagagaggagggagagagaaaacgaaacaccacctccttgacctgagtcaggggtttctggccttt
21071 tgggagaacgttcagcgacaatgcagtatttgggcccgttctttttttttcttcttctttttctttctttt
21141 ttttggactgagtctctctcgctctgtcacccaggctgcggtgcggtggcgctctctcggctcactgaa
21211 acctctgcttcccgggttccagtgattcttcttcggtagctgggattacaggcgcacaccatgacggccg
21281 gctcatattcctatttcagtagagacggggtttctccacgttggccacgctggtctcgaactcctgacc
21351 tcaaatgatccgccttcctgggcctcccaaagtgctggaaacgacaggcctgagccgccgggatttcagc
21421 ctttaaaagcgcgggccctgccacctttcgctgtggcccttacgctcagaatgacgtgtcctctctgccg
21491 taggttgactccttgagtcccctaggccattgcactgtagcctgggcagcaagagccaaactccgtcccc
21561 ccacctccccgcgcacataataactaactaacaaactaactaactaactaaactaactaaataaataaaa
21631 tctctacacgtcacctctaagtgtgtgttcccgtgaggagtgatttctaagaaatggcactgtacactga
21701 acgcagtggctcacgtctgtcatcccgaggtcaggagttcgagaccagcccggccaacgtggtgaaaccc
21771 ccgtctctactgaaaatacgaaatggagtcaggcgccgtggggcaggcacctgtaaccccagctactcgg
21841 gaggctggggtggaagaattgcttgaacctggcaggcggaggctgcagtgacccaagatcgcaccactgc
21911 actacagcctgggcgacagagtgagacccggtctccagataaatacgtacataaataaatacacacatac
21981 atacatacatacatacatacatacatacatacatccatgcatacagatatacaagaaagaaaaaaagaaa
22051 agaaaagaaagagaaaatgaaagaaaaggcactgtattgctactgggctagggccttctctctgtctgtt
22121 tctctctgttcgtctctgtctttctctctgtgtctctttctctgtctgtctgtctctttctttctctctg
22191 tctctgtctctgtctttgtctctctctctccctctctgcctgtctcactgtgtctgtcttctgtcttact
22261 ctctttctctccccgtctgtctctctctctctctctccctccctgtttgtttctctctctccctccctgt
22331 ctgtttctctctctctctttctgtctgtttctgtctctctctgtctgtctatgtctttctctgtctgtct
22401 ctttctctgtctgtctgcctctctctttctttttctgtgtctctctgtcggtctctctctctctgtctgt
22471 ctgtctgtctctctctctctctctctgtgcctatcttctgtcttactctctttctctgcctgtctgtctg
22541 tctctccctccctttctgtttctctctctctctctctctctccccctctccctgtctgtttctctccg
22611 tctctctctctttctgtctgtttctctctgtctctgtctttctgtgtgtctgtctgtctctctctctatt
22681 tgtctttctccctccctgtctgtttctctctctctctctctctctctctctctctctctgtctgtctgtt
22751 tctctctatctctcgctgtccatctctgtctttctatgtctgtctctttctctgtcagtctgtcagacac
22821 ccccgtgccgggtagggccctgccccttccacgaaagtgagaagcgcgtgcttcggtgcttagagaggcc
22891 gagaggaatctagacaggcgggccttgctgggcttccccactcggtgtatgatttcgggaggtcgaggcc
22961 gggtccccgcttggatgcgaggggcattttcagacttttctctcggtcacgtgtggcgtccgtacttctc
23031 ctatttccccgataagctcctcgacttcaacataaacggcgtcctaagggtcgatttagtgtcatgcctc
23101 tttcaccgccaccaccgaagatgaaagcaaagatcggctaaataccgcgtgttctcatctagaagtggga
23171 acttacagatgacagttcttgcatgggcagaacgaggggaccggggacgcggaagcctgcttgagggag
23241 gaggggtggaaggagagacagcttcaggaagaaaacaaaacacgaatactgtcggacacagcactgacta
23311 cccgggtgatgaaatcatctgcacactgaacacccccgtcacaagtttacctatgtcacagtcttgctca
23381 tgtatgcttgaacgacaaataaaagttcgggggggagaagagaggagagagagagagacggggagaga
23451 gggggagagggggggagagagagagagagagagagagagagagagagagagagaaagagaagtaaaa
23521 ccaaccaccacctccttgacctgagtcaggggtttctggccttttgggagaacgttcagcgacaatgca
23591 gtatttgggcccgttctttttttcttcttcttctttttctttctttttttttggactgagtctctctcgct
23661 ctgtcacccaggctgcggtgcggtggcgctctctcggctcactgaaacctctgcttcccgggttccagtg
23731 attcttcttcggtagctgggattacaggtgcgcaccatgacggccggctcatcgttctatttttagtaga
23801 gacggggtttctccacgttggccacgctggtctcgaactcctgaccacaaatgatccaccttcctgggcc
23871 tcccaaagtgctggaaacgacaggcctgagccgccgggatttcagcctttaaaagcgcgggccctgccac
23941 ctttcgctgcggcccttacgctcagaatgacgtgtcctctctgccataggttgactccttgagtccccta
24011 ggccattgcactgtagcctgggcagcaagagccaaactccgtccccccacctccccgcgcacataataac
24081 taactaactaactaactaactaaaatctctacacgtcacccataagtgtgtgttcccgtgaggagtgatt
24151 tctaagaaatggtactgtacactgaacgcagtggctcacgtctgtcatcccgaggtcaggagttcgagac
24221 cagcccggccaacgtggtgaaaccccgtctctactgaaaatacgaaatggagtcaggcgccgtggggca
24291 ggcacctgtaaccccagctactcgggaggctggggtggaagaattgcttgaacctggcaggcggaggctg

24361 cagtgacccaagatcgcaccactgcactacagcctgggcgacagagtgagacccggtctccagataaata
24431 cgtacataaataaatacacacatacatacatacatacatacatacatacatacatacagatataca
24501 agaaagaaaaaaagaaaagaaaagaaagagaaaatgaaagaaaaggcactgtattgctactgggctaggg
24571 ccttctctctgtctgtttctctctgttcgtctctgtctttctctctgtgtctctttctctgtctgtctgt
24641 ctgtctgtctgtctctttctttctttctgtctctgtctttgtccctctctctccctctctgcctgtctca
24711 ctgtgtctgtcttctatcttactctctttctctccccgtctgtctctctctcactccctccctgtctgtt
24781 tctctctctctctctttctgtctgtttctgtctctctctgtctgcctctctctttctctatctgtctctt
24851 tctctgtctgtctgcccctctctttctttttctgtgtctctctgtctgtctctctctctctctgtgccta
24921 tcttctgtcttactctctttctctgcctgtctgtctgtctctctctgtctctccctccctttctgcttct
24991 ctctctctctctctctctcccccctccctgtctgtttctctctgtctccctctctttctgtctgtttctc
25061 actgtctctctctgtctgtctgtttcattctctctgtctctgtctctgtctctctctctctctgtctctc
25131 cctctctgtgtgtatcttttgtcttactctccttctctgcctgtccgtctgtctgtctgtctctctctct
25201 ccctgtccctctctctttctgtctgtttctctctctctctctctctctctctctctgtctctgtctttct
25271 ctgtctgtccctttctctgtctgtctgcctctctctttctctttctgtgtctctctgtctctctctctgt
25341 gcctatcttctgtcttactctctttctctgcctgtctatctgtctgtctctctctgtctctctccctgcc
25411 tttctgtttctctctctctccctctctcgctctctctgtctttctctctttctctctgtttctctgtctc
25481 tctctgtccgtctctgtctttttctgtctgtctgtctctctctttctttctgtcgtctgtctctgtctct
25551 gtctctgtctctctctctctctctccttgtctctctcactgtgtctgtcttctgtcttactctccttc
25621 tctgcctgtccatctgtctgtctgtctctctctctctccctacctttctgtttctctctcgctagctc
25691 tctctctctctgcctgtttctctctttctctctctgtctttctctgtctgtctctttctctgtctgtctg
25761 tctctttctctctgtctctgtctctgtctctctctctctctctctctctctgcctctctcactgtgtctg
25831 tcttctgtcttattctctttctctctctgtctctctctctctctcctttcctgtctgtttttctctctct
25901 ctctctctttctgcctgtttctctctgtctgtctctgtctttctctgtctgtctgcctctctctttcttt
25971 ttctgcgtctctctgtctctctctctctctctctgttcctatcttctgtcttactctgtttccttgcctg
26041 cctgcctgtctgtgtgtctgtctctctctctctctctctctctctccctccctttctctttctctgtc
26111 tctctctctctttctgggtgtttctctctgtctctctgtccatctctgtctttctatgtctgtctctctc
26181 tttctctctgtctctgtctctgcctctctctctctctctctctctctctctctgtctgtctctctcactg
26251 tgtgtgtctgtcttctgtcttactctccttctctgcctgtccgtctgtctgtctgtctctccctctctct
26321 ccctccctttctgtttctctctctctctctttctgtctgtttctctctttctctctctgtctgtctctttt
26391 ctctgtctgtctgtctctctctttctttttctctgtctctctgtctctctctgtgtctgtctctctgtct
26461 gtgcctatcttctgtcttactctctttctctggctgtctgcctgtctctctctctctctctgtctgtctc
26531 cgtccctctctccctgtctgtctgtttctctctctgcctctctctctctctgtctgtctctttctctgtc
26601 tgtctgtctctctctttctttttctctgtctctctgtctctctctgtgtctgtctctctttctgtgccta
26671 tcttctgtcttactctctttctctggctgtctgcctgtctctctctctctgcctgtctccgtccctccct
26741 ccctgtctgtctgtttctctctctgtctctgtctctctgtccatctctgtctgtctctttctctttctct
26811 ctctctgtctctgtctctctctctctctgcctgtctctctcactgtgtctgtcttctgtcttactctctt
26881 tctctgcctgcctctctgtctgtctgtctctctccctccatgtctctctctctctctcactcactctctc
26951 tccgtctctctctctttctgtctgtttctctctctgtctgtctctctccctccatgtctctctctctctc
27021 tctcactcactctctctccgtctctctctctctttctgtctgtttctctctctgtctgtctctctccctc
27091 catgtctctctctctccctctcactcactctctctccgtctctctctctctttctgtctgtttctctgtc
27161 tgtctgtctgtctgtctgtctctctctctctctctctctctctctctctctgtttgtctttctccctccc
27231 tgtctgtctgtctgtctctctctctctgtctctgtctctgtctctctctctttctctttctgtctgtttc
27301 tctctatctctcgctgtccatctctgtctttctatgtctgtctctttctctgtcagtctgtcagacaccc
27371 ccgtgccgggtagggccctgccccttccacgagagtgagaagcgcgtgcttcggtgcttagagaggccga
27441 gaggaatctagacaggcgggccttgctgggcttccccactcggtgtacgatttcgggaggtcgaggccgg
27511 gtccccgcttggatgcgaggggcattttcagacttttctctcggtcacgtgtggcgtccgtacttctcct
27581 atttccccgataagctcctcgacttcaacataaactgttaaggccggacgcaacacggcgaaaccccgtc
27651 tctactaaaaatacaaagctgagtcgggagcggtggggcaggcccctgtaatgccagctcctcgggaggc
27721 tgaggcgggagaatcgcttgaaccagggaagcggaggctgcagggagccgagatcgcgccactgcactac
27791 ggcccaggctgtagagtgagtgagactcggtctctaaataaatacggaaattaattaattcattaattct

27861 tttccctgctgacggacatttgcaggcaggcatcggttgtcttcgggcatcacctagcggccactgttat
27931 tgaaagtcgacgtgacacggagggaggtctcgccgacttcaccgagcctggggcaacgggtttctctctc
28001 tcccttctggaggcccctccctctctccctcgttgcctagggaacctcgcctagggaacctccgccctgg
28071 cgggggccctattgttctttgatcggcgctttacttttctttgtgttttggcgcctagactcttctactt
28141 gggctttgggaagggtcagtttaattttcaagttgccccccggctccccccactacccacgtcccttcac
28211 cttaatttagtgagtcggttaggtgggtttcccccaaaccgcccccccccccccgcctcccaacaccctg
28281 cttggaaaccttccagagccaccccggtgtgcctccgtcttctctcccctttcccccacccccttgccggcg
28351 atctcattcttgccaggctgacatttgcatcggtgggcgtcaggcctcactcgggggccaccgtttttga
28421 agatgggggcggcacggtcccacttccccggaggcagcttgggccgatggcatagccccttgacccgcgt
28491 gggcaagcgggcgggtctgcagttgtgaggctttcccccccgctgcttcccgcctcaggcctccctccct
28561 aggaaagcttcaccctggctgggtctcggtcacctttttatcacgatgttttagtttctccgccctccggc
28631 cagcagagtttcacaatgcgaagggcgccacggctctagtctgggccttctcagtacttgcccaaaatag
28701 aaacgctttctgaaaactaataactttgctcacttaagatttccagggacggcgccttggcccgtgtttg
28771 ttggcttgttttgtttcgttctgttttgttttgttcgtgtttttccttttctcgtatgtctttctttttcag
28841 gtgaagtagaaatccccagttttcaggaagacgtctatttttccccaagacacgttagctgccgttttttc
28911 ctgttgtgaactagcgcttttgtgactctctcaacgtgcagtgagagccggttgatgtttactatccttc
28981 atcatgacatcttattttctagaaatccgtaggcgaatgctgctgctgctcttgttgctgttgttgttgt
29051 tgttgttgtcgtcgttgctgttgtcgttgtcgttgttgttgtcgttgtcgttgttttcaaagtataccccc
29121 ggccaccgtttatgggatcaaaagcattataaaatatgtgtgattatttcttgagcacgcccttcctccc
29191 cctctctctgtctctctgtctgtctctgtctctctctttctctgtctgtcttctctctctctctctctct
29261 gtgtctctctctctctgcctgtctgtttctctctctctgcctctctctctctctctctctgcctgtct
29331 ctctcactgtgtctgtcttctgtcttactccctttctctgtctgtctgtcggtctctctctctctctctc
29401 cctgtctgtatgtttctctctgtctctgtctctctctctctttctgtttctctctctccgtctctgtctt
29471 tctctgactgtctctctctttccttctctctgtctctctctgcctgtctctctcactctgtcttctgtct
29541 tactctctctctctgcctgcctgtctctctcactctctctctctgtgtgtctctctctctctttctgttt
29611 ctctctgtctctctgtccgtctctgtctttctctgtctgtctctttgtctgtctgtctttgtctttcctt
29681 ctctctgtctctgtctctctcactgtgtctgtcttctgtcttagtctctctctctctctccctgtctg
29751 tctgtctctctctctctctcccccctgtctgttctctctctctctctctctctctctctgtctttgtct
29821 ttctttctgtctctgtctctctctctctctctgtgtgtctgtcttctgtcttactgtctttctctgcctg
29891 tctgtctgtctgtctctctctgtctgtctctctctctctctcccccctgtcggctgtttctctgtctctg
29961 tctgtgtctctctttctgtctgtttctctctgtctgtctttctctctctgtctctttctctctgtctctc
30031 tgtctgtctctgtctctctctctgtctctctctctctgtgggggtgtgtgtgtgtgtgtgtatgtgtgtg
30101 tgtgtgtgtgtgtgtgtctgccttctgtcttactctctttctctgcctgtctgtctgcctgtctgtttgt
30171 ctctctctctctgcctgtctctctcccttcctgtctgtttctctctctttctgtttctctctgtctctgt
30241 ccatctctgtctttctccgtctgtctctttatctgtctctctccgtctgtctctttatctgtctctctct
30311 ctctttctgtctttctctctctgtgtatcgttgtctctctctgtctgtctctgtctctgtctctctgtct
30381 ctctctctctctctctctctctgtctgtctgtccgtctgtctgtctcggtctctggctctcgctatctcc
30451 cgccctctctttttttgcaaaagaagctcaagtacatctaatctaatcccttaccaaggcctgaattctt
30521 cacttctgacatcccagatttgatctccctacagaatgctgtacagaactggcgagttgatttctggact
30591 tggatacctcatagaaactacatatgaataaagatccaatcctaaaatctggggtggcttctccctcgac
30661 tgtctcgaaaaatcgtacctctgttcccctaggatgccggaagagttttctcaatgtgcatctgcccgtg
30731 tcctaagtgatctgtgaccgagccctgtccgtcctgtctcaaatatgtacgtgcaaacacttctctccat
30801 ttccacaactacccacggccccttgtggaaccactggctctttgaaaaaaatcccagaagtggttttggc
30871 tttttggctaggaggcctaagcctgctgagaactttcctgcccaggatcctgtgtgaccaaaagtgcctc
30941 tgctgggagctgggatcctcgggaccatgcttgctagcgctggatgagtctctggaaggacgcacgggac
31011 tccgcaaagctgacctgtcccaccgaggtcaaatggatacctctgcattggcccgaggcctccgaagtac
31081 atcaccgtcaccaaccgtcaccgtcagcatccttgtgagcctgcccaaggccccgcctccggggagactc
31151 ttgggagcccggccttcgtcggctaaagtccaaagggatggtgacttccacccacaaggtccccactgaa
31221 cggcgaagatgtggagcgtaggtcagagaggggaccaggaggggagacgtcccgacaggcgacgagttcc
31291 caaggctctggccaccccacccacgccccacgccccacgtcccgggcacccgcgggacaccgccgcttta

31361 tcccctcctctgtccacagccggccccaccccaccacgcaacccacgcacacacgctggaggttccaaaa
31431 ccacacggtgtgactagagcctgacggagcgagagcccatttcacgaggtgggaggggtgggggtggggt
31501 gggttgggggttgtggggtctgtggcgagcccgattctccctcttgggtggctacaggctagaaatgaat
31571 atcgcttcttgggcggaggggcttccttaggccatcaccgcttgcgggactacctctcaaaccctccctt
31641 gaggccacaaaatagattccaccccacccatcgacgtttcccccgggtgctggatgtatcctgtcaagag
31711 acctgagcctgacaccgtcgaattaaacaccttgactggctttgtgtgtttgtttgtttctgagatggag
31781 tcttgctctgtcccccaggctggagtgcagtggcgtgatctcagctcactggaacctctgcctcctgggt
31851 tcaagtgattctcctgtctcagcgccaccatggccggctcatttttttttttttttttttttttttttt
31921 ggtagacacggggtttcaccctctttcattggttttcactggagattctagattcgagccacacctcatt
31991 ccgtgccacagagagacttctttttttttttttttttaagcgcaacgcaacatgtctgccttatttgagt
32061 ggcttcctatatcattataattgtgttatagatgaagaaacggtattaaacactgtgctaatgatagtga
32131 aagtgaagacaaaagaaaggctatctattttgtggttagaataaagttgctcagtatttagagctaccta
32201 aatacgtcagcatttacactcttcctagtaaaagctggccaatctgaataatcctccttttaaacaaacac
32271 aatttttgatagggttaagatttttttaagaatgcgactcctgcaaaatagctgaacagacgatacacat
32341 ttaaaaaaataacaacacaaggatcaaccagacttgggaaaaaatcgaaaaccacacaagtcttatgaag
32411 aactgagttcttaaaataggacggagaacgtagctatcggaagagaaggcagtattggcaagttgattgt
32481 tacgttggtcagcagtagctggcactatctttttggccatctttcgggcaatgtaactactacagcaaaa
32551 tgagatatgatccattaaacaacatattcgcaaatcaaaaagtgtttcagtaatataatgcttcagattt
32621 agaagcaaatcaaatgatagaactccactgctgtaataagtcaccccaaagatcaccgtatctgacaaaa
32691 taactaccacagggttatgacttcagaatcatactttctcttgatatttacttatgtatgtatttatttt
32761 ttttaatttatttctcttgagacggcgtctcgctctgtcgcccaggctggagtgcgatggtgtgatctcg
32831 gctcactgcaaccgccacctccctgggttcaagcgattctcctgcctcagcctcccgagtagctgggact
32901 acaggtgcccgccaccacgcccagctaatctttatacttttaatagagacggggtttcaccgtgtcggcc
32971 cggatggtctcgatctcttgacctcgtgacccgcccgcctcggcctcccaaagtgctgggatgacaggcg
33041 tgagccactgagcccggccttctcttgacgtttaaactatgaagtcagtccagagaaacgcaataaatgt
33111 caacggtgaggatggtgttgaggcagaagtaggaccacacttttcctatcttattcagttgataacaat
33181 atgacctaggtagtaatttcctatgtgcctacttatacacgagtacaaaagagtaaaacagagagactgc
33251 taaattaaagggtacgtgaagttcttcatagtaactccgtaaactggaacactgtcaaaaagcagcagct
33321 agtgaattgtttccatgtatttttctattatccaataagtgaactatgctattcctttccagtctcccaa
33391 gcacttcttgtccccatcaccacttcggtgctcgaagaaaaagtaacaaatcaaggaacacaactaaaga
33461 aacacacacacaaaccaaagacaactacagcgtctgcaaaagtttgctagaagactgaaactgttgagta
33531 taaggatctggtattctacgatcatgagttcacttcagagtttgttcaagacatacgtttcgtaaggaaa
33601 catcttagttagaagttattcagcagtaggtaccatccctaagtatttttcaccaaattcgtgacaataa
33671 agagctatctaaccagaaaaattagcgagtaccggcaccatccatagggctttgtctttacgcttcatta
33741 gcacttaccatgccttacaatgtctaggattgaccctgatagcatttcgaaaacaagctaatgctttgtc
33811 cagttcttcagtgaagacaagctcacgccctaatgcgctataggcataagcatcatttggatccacttcg
33881 agagttctctggaagaattgaatcgcaatatcgtgttcccgttgcagaccgaaacagttccctgcagcac
33951 accaggcctctggctggcgaatttttatccatgtctgtgaagtctttggacagaactgaaagagcaacct
34021 ctttcggaggatgccaaagtgttgtagagtagatctccatgccttcgactctgtaattctcaatcctcct
34091 aacctctgagaattgtctttcagcttgcgtggactctgaaagtttacaataggcccttccgatttggcac
34161 agtacccaaccggtattgcagtggtgagaagctagatggctcaagatgctgatagcttctttgccgtggt
34231 aagaacacaaagctaaataacctttccccctttcacgaagaaggctcatcaagccttccgctgctgcttt
34301 ttgtagattaaaagcctgaatctgaggcgcgattgcggctattttcccttctgaaatgacggaagagtcc
34371 aattttgtcacttccaggctatcacttatgttcggtggagttattgctcctttattagtttactttgg
34441 ttcttctgtttgggatttttaggtggaaacttcatttttaattttctcctattctcctcggttgtggagct
34511 gtcactagtcaagagtcgtgaatttcttcgaggcggtgcatttgggggagatgccatagtggggctcaat
34581 acctgaggtgttgcccttgtcggcggaccagaactttgtgtttttgcaaggactggagttacctttcggc
34651 tctttcccctctgcgagaagacagacggtgttccggtttggccgattctggcaacaggctttttgaagg
34721 ggctccggtggatggcacgtcaatgacagacggtgtctcataccagtgcagttttgtcaatagggtccgt
34791 ctccgggacttggggtttctaatggcaaaatgccaacacttggggttaatggactaacagctgctggtcc

34861 tcctaataaacttcgaccagtttttggtttatgttgaacctgtttagatcatatggaagttcctgttccc
34931 agtgggacagtatcaggtgaaaggacagctgaatcgatagaagacactggggagtctgtattcaaggagt
35001 actttgaattggaagattctaaattccatccgtttcattcgacggtgtcctggggtgtttccgtaagaac
35071 ggtctcgggctgtctgtgacataaactaggacgaggtccaagtgttgtggcgcaacacttggacaggcag
35141 ttgctaaagctctctagagaggtgaatcaaaatgtttggtcaggatctggcttttccccctatttcaca
35211 tcatgattcaaagggacaccagaggaaaggatttcaacgaaggctcttttggtcacattctgatcctttg
35281 gtaagccgatctgtcttgcaatatacatgtcccgacgatggaaggggaaagcgagctgaatcaccaaact
35351 caggaacgataatatcatcgtggcttttctgcttatgaaacactccacccgataagatttgatccccttc
35421 tgcaagcttgctgagatcaacacaacatttcgcaagcaggcatttgcattgcggggtagtacaactgtgt
35491 cctttcaagagtctatatgttttataggcctttcctgagcggtaagaacaggtcgccagtaagaacaagg
35561 cttcttctgagtgtacttctgcataaaggcgttctgcgggggaaaccgcatctcggtaggcatagtggtt
35631 tagtgcttgccatatagcagcctggacgggtccctgcagcaccgccatcctcgaggctcaggcccacttt
35701 ctgcagtgccacaggcacccccccccccccccatagcggctccggcccggccagccccggctcatttaaag
35771 gcaccagccgccgttaccgggggatgggggagtccgagacagaatgacttctttatcctgctgactctgg
35841 aaagcccggcgccttgtgatccattgcaaaccgagagtcacctcgtgtttagaacacggatccactccca
35911 agttcagtgggggggatgtgaggggtgtggcaggtaggacgaaggactctcttccttctgattcggtctgc
35981 acagtggggcctagggctggagctctctccgtgcggaccgctgactccctctaccttgggttccctcggc
36051 cccaccctggaacgccgggccttggcagattctggcccttcctggcccttcagtcgctgtcagaaacccc
36121 atctcatgctcggatgccccgagtgactgtggctcgcacctctccggaaacattggaaatctctcctcta
36191 cgcgcggccacctgaaaccacaggagctcgggacacacgtgctttcgggagagaatgctgagagtctctt
36261 gccgactctctcttgacttgagttcttcgtgggtgcgtggttaagacgtagtgagaccagatgtattaac
36331 tcaggccgggtgctggtggctcacgcctgtaaccccaacactttgggaggccgaggccgtaggatccctc
36401 gaggaatcgcctaaccctggggaggttgaggttgcagtgagtgagccatagttgtgtcactgtgctccag
36471 tctgggcgaaagacagaatgaggccctgccacaggcaggcaggcaggcaggcaggcagaaagacaacagc
36541 tgtattatgttcttctcagggtaggaagcaaaaataacagaatacagcacttaattaatttttttttttc
36611 ccttcggacggagtttcactcttggtgcccacgctggagtgcagtggcaccatctcggctcaccgcaacc
36681 tccacctcccgcgttcaagcgattctcctgcctcagcctcctgagtagctgggattacagggaggagcca
36751 ccacacccagctgattttgtattgttagtagagacggcatttctccatgtgggtcaggctggtctcgaac
36821 tggcgaccccagtggatctgcccgccccggcctcccaaagtgctggggtgacaggcgtgagccatcgtga
36891 ctggccggctacgtttatttatttatttttttaattattttacttttttttagttttccattttaatcta
36961 tttatttatttacatttatttatttatttatttatttacttatttatttatttcgagacagactctcgc
37031 tctgctgcccaggctggagtgcagcggcgtgatctcggctcactgcaagctccgcctcccgggttcacgc
37101 cattctcctgcctcagcctcccaagtagctgggactacaggcgcccgccaccgtgcccggctaacttttt
37171 gtattttgagtagagatggggtttcactgtggtagccaggatggtctcgatctcctgaccccgtgatccg
37241 tccacctcggcctcccaaagtgctgggatgacaggcgtgagccaccgccccggcctatttatctattta
37311 ttaactttgagtccaggttatgaaaccagttagtttttgtaatttttttttttttttttttttgaga
37381 cgaggtttcaccgtgttgccaaggcttggaccgagggatccaccggccctcggcctcccaaaagtgcggg
37451 gatgacaggcgcgagcctaccgcgcccggacccccctttcccctccccgcttgtcttcccgacagac
37521 agtttcacggcagagcgtttggctggcgtgcttaaactcattctaaatagaaatttgggacgtcagcttc
37591 tggcctcacggactctgagccgaggagtcccctggtctgtctatcacaggaccgtacacgtaaggaggag
37661 aaaaatcgtaacgttcaaagtcagtcattttgtgatacagaaatacacggattcacccaaaacacagaaa
37731 gcagtcttttagaaatggccttagccctggtgtccgtgccagtgattcttttcggtttggaccttgactg
37801 agaggattcccagtcggtctctcgtctctggacggaagttccagatgatccgatgggtgggggacttagg
37871 ctgcgtccccccaggagccctggtcgattagttgtggggatcgccttggagggcgcggtgacccactgtg
37941 ctgtgggagcctccatccttccccccacccccctccccaggggatcccaattcattccgggctgacacgct
38011 cactggcaggcgtcgggcatcacctagcggtcactgttactctgaaaacggaggcctcacagaggaaggg
38081 agcaccaggccgcctgcgcacagcctggggcaactgtgtcttctccaccgcccccgcccccacctccaag
38151 ttcctccctcccttgttgcctaggaaatcgccactttgacgaccgggtctgattgacctttgatcaggca
38221 aaaacgaacaaacagataaataaataaaataacacaaaagtaactaactaaataaaataagtcaatacaa
38291 cccattacaatacaataagatacgatacgataggatgcgataggatacgataggatacaatacaatacga

38361 tacgatacaatacaatacaatacaatacaatacaatacaatacaatacaatacaatacaatacaatacgc
38431 cgggcgcggtggctcatgcctgtcatcccgtcactttgggatgccgaggtggacgcatcacctgaagtcg
38501 ggagttggagacaagcccgaccaacatggagaaatcccgtctcaattgaaaatacaaaactagccgggcg
38571 cggtggcacatgcctataatcccagctgctaggaaggctgaggcaggagaatcgcttgaacctgggaagc
38641 ggaggttgcagtgagccgagattgcgccatcgcactccagtctgagcaacaagagcgaaactccgtctca
38711 aaaataaatacataaataaatacatacatacatacatacatacatacatacatacatacatacatacatacaaa
38781 ataaaataaataaatgggccctgcgcggtggctcaagcctgtcatcccctcactttgggaggccaaggcc
38851 ggtggatcaagaggcggtcagaccaacagggccagtatggtgaaaccccgtctctactcacaatacacaa
38921 cattagccgggcgctgtgctgtgctgtactgtctgtaatcccagctactcgggaggccgagctgaggcag
38991 gagaatcgcttgaacctgggaggcggaggttgcagtgagccgagatcgcgccactgcaacccagcctggg
39061 cgacagagcgagactccgtctccaaaaaatgaaaatgaaaatgaaacgcaacaaaataattaaaaagtga
39131 gtttctggggaaaaagaagaaaagaaaaaagaaaaaaacaacaaaacagaacaaccccaccgtgacatac
39201 acgtacgcctctcgcctttcgaggcctcaaacacgttaggaattatgcgtgatttctttttttaacttca
39271 ttttatgttattatcatgattgatgtttcgagacggagtctcggaggcccgccctccctggttgcccaga
39341 caaccccgggagacagaccctggctgggcccgattgttcttctccttggtcaggggtttccttgtctttc
39411 ttcgtgtctttaacccgcgtggactcttccgctcgggtttgacagatggcagctccactttaggccttgt
39481 tgttgttggggacttttcctgattctccccagatgtagtgaaagcaggtagatttgccttgcctggacttg
39551 cctggccttgccttttctttctttctttctttattactttctcttttttcttcttcttcttcttcttcttc
39621 ttcttcttcttcttcttcttcttcttttttttttttgagacagagtttcactcttgttgcccaggctagag
39691 ggcaatggtgcgatctcggctcaccgcaccctccgcctcccaggttcaagcgattctcctgcctcagcct
39761 cctgattagctgggattacaggcatgggccaccgtgcctggctgatgtttgtacttttagtagagacggt
39831 gttttccatgttggtcaggctggtctcccactcccaacctcaggtggtccgcctgccttagcctcccaa
39901 agtgctgggatgacaggcgtgagccaccgcgcccagcctctctctctctctctctctctctctctctc
39971 tctctctctcgctcgcttgcttgcttgctttcgtgctttcttgctttcccgttttcttgctttctttctt
40041 tctttcgtttctttcatgcttgctttcttgcttgcttgcttgctttcgtgctttcttgctttcctgtttt
40111 cttttctttctttctttctttctttctttgtttctttcttgcttgctttcttgcttgcttgcttgctt
40181 tcgtgctttcttgctttcctgttttctttctttctttctttctttgtttctttcttgcttgctttcttg
40251 cttgcttgctttcgtgctttcttgttttctcgctttctttctttctttgtttctttcctgcttgctttc
40321 ttgcttgcttgctttcgtgctttcttgctttcctgttttctttctttctttctttctttgtttctttct
40391 tgcttgctttcttgcttgcttgctttcgtgctttcttgttttctcgatttctttctttctttgtttctt
40461 tcctgcttgctttcttgcttgattgctttcgtgctttcttgctttcttgttttctttctttctttgttt
40531 ctttctttcttgcttccttgttttcttgctttcttgcttgcttgctttcgtgctttcttgttttcttgct
40601 ttctttctttgtttctttcttgcttgctttcttgcttccttgttttcttgctttcttgcttgcttgctt
40671 tcgtgctttctttcttgctttctttttctttctttctttttctttttctttctttcttgctttctttttctttt
40741 cattcattcattctttctttctttccttttctttctttctttctttctatctttctttctttctttctttc
40811 tttctttctttctttctttctgtttcgtccttttgagacagagtttcactcttgtttccacggctagagt
40881 gcaatggcgcgatcttggctcaccgcaccttccgcctcccgggttcgagcgcttctcctgcctcagcctc
40951 ccgattagcggggattacagggaggcacccccacgcctggcttggctgatgtttgtgttttagtaggca
41021 cgccgtgtctctccatgttgctcaggctggtctccaactcccgacctcctgtgatgcgcccacctcggcc
41091 tctcgaagtgctgggatgacgggcgtgagccaccgtgcccggcctgttgactcatttcgctttttattt
41161 ctttcgtttccacgcgtttacttatatgtattaatgtaaacgtttctgtacgcttatatgcaaacaacga
41231 caacgtgtatctctgcattgaatactcttgcgtatggtaaatacgtatcggttgtatggaaatagacttc
41301 tgtatgatagatgtaggtgtttgtgttatacaaataaatacacatcgctctataaagaagggatcgtcga
41371 taaagacgtttattttacgtatgaaaagcgtcgtatttatgtgtgtaaatgaacgagcgtacgtagttat
41441 ctctgttttctttcttcctctccttcgtgtttttcttccttcctttcttcctttctctccttctttaggt
41511 ttttcttcctctcttccttccttctttctctctttctgtccttttttccttcgtgctttatttctctttt
41581 cgttccctgtgtttccttctttttttctttcctctctgtttctttttcccttctttccttcgtttctttcc
41651 tcattctttctctctttttcgtgtttctttccttttccgtctgtctttttaaaaaatggagtgtttcagaag
41721 tttactttgtgtatctacgttttctaaattgtctctctttttctccattttcttcctccctccctccctcc
41791 ctccctgctcccttcccctccctccttcccttttcgccatctgtctcttttccccactcccctcccccgtc

41861 tgtctctgcgtggattccggaagagcctacgcattctgcctctccgtgtgtctgcagcgacccgcgaccg
41931 agtccttgtgtgttctttctccctccctccctccctccctccctccctccctccctgcttccgagaggca
42001 tctccaaacacccacgcgccgtgggttgtcttctgactctgtcgcggtcgaggcagagacgcgttttggg
42071 caccgtttgtgtggggttggggcagaggggctgcgttttcggcctcgggaagagcttctcgactcacggt
42141 ttcgctttcgcggtccacgggccgccctgccagccggatctgtctcgctgacgtccgcggcggttgtcgg
42211 gctccatctggcggccgctttgagatcgtgctctcggcttccggagctgcggtggcagctgccgagggag
42281 gggaccgtccccgctgtgagctaggcagagctccggaaagcccgcggtcgtcagcccggctggcccggtg
42351 gcgccagagctgtggcgcgtcgcttgtgagtcacagctctggcgtgcaggtttatgtgggggagaggctg
42421 tcgctgcgcttctgggcccgcggcgggcgtggggctgcccgggccggtcgaccagcgcgccgtagctccc
42491 gaggcccgagccgcgacccgcggggacccgccgcgcgtggcgcgggaggctggggacgcccttcccggcc
42561 cggtcgcgggtccgcgctcatcctggccgtctgaggcggcggccgaattcgtttccgagtccccgtgggg
42631 agccggggaccgtcccgcccccgtcccccgggtgccggggagcggtccccgggccgggccgcggtccctc
42701 tgccgcgatcctttctggcgagtccccgtgcggagtcggagagcgctccctgagcgcgcgtgcggcccga
42771 gaggtcgcgcctggccggccttcggtccctcgtgtgtcccggtcgtaggaggggccggccgaaaatgctt
42841 ccggctcccgctctggagacacgggccggccccctgcgtgtggcacgggcggccgggagggcgtccccgg
42911 cccggcgctgctcccgcgtgtgtcctggggttgaccagagggccccgggcgctccgtgtgtggctgcgat
42981 ggtggcgtttttggggacaggtgtccgtgtcgcgcgtcgcctgggccggcggcgtggtcggtgacgcgac
43051 ctcccggccccgggggaggtatatctttcgctccgagtcggcattttgggccgccgggttatt

Appendix D: Pan troglodytes Consensus Sequence

```
   1 gctgacacgctgtcctctggcgacctgtcgctggagaggttgggactccggatgcgtgcggggctctggc
  71 ctaccggtgacccggctagccggccgtgctcctgcttgagccgcctgctggggcccgcgggcctgctgtc
 141 ctctcgcgcgcccgggcgtgtcccgactcccggtgccggcccgggtccgggtctctgacccgcctggggg
 211 cggcggggaaggcggcgagggccaccctgcccccgtgcgctctccgctgcgggcgcccggggcggccgcg
 281 acaaccccacccgctggctccgtgccgtgcgtgtcaggcgttctcgtctccgctgggttgtccgccgcc
 351 ccttccccggagcagggcgtcggccggccggccggccgacctccgctcccgggggctcttcgtgatcga
 421 tgtggtgacgtcgtgctctcccgggccgggtccgagccgcgacgggcgaggggcggacgttcgtggcgaa
 491 cgggaccgttcttctcgctccgccccgcggggtcccctcgtctctcctctcccgcctgcgggtggcgcg
 561 tgtgggggaggcgtgtggggtgcggagcccggcctgacctcgccgtcccgcccgctgccttctgcgtcgc
 631 ggggcgggtcggcggggtcctctgacgcggcaggcacccctcgctgtcgcctccagtggttgtggacttg
 701 cgggcggccccctccgcggcggtgggggtgccgtcccgccggcccgtcgtgctgccctctcgggggt
 771 gtgcgcgagcgtcggctccgcttgggcctttgcggtgctcctggagcgctccgggttgtccctcaggtgc
 841 ccgaggccgagcggtggtgtgtcgttcccgcccccagcgccccctcctccggtcgccgccgcggtgtccg
 911 cgcgtgggtcctgagggagctcgtcggtgtggggttcgaggcggtcgagggagacgagacgcgcccctcc
 981 cacgcggggaagggcgtccgcctggtccggcgagcgcacgtcccgtgctcccctctggcgggtgcgcgcg
1051 ggccgtgtgagcgatcgcggtgggttcgggccggtggcgcgtgcgccggccggccgccgaggggctgccg
1121 ttctgcctccgaccggtagtgtgtgggttgacttcgggggtgctttgcctcggaaggaaggaggcgggtg
1191 gacgggggggcctcgtggggttgcgcgcacgcgcgcaccggccgggcccccgccctgaccgcggacgctc
1261 aaggtggccgcacgcaggtgtttcctcgtactgcaggcccctccttccccaggcgtccctcggcgcct
1331 ctgcgggcccgaggaggggcggctggcgtgtgggggagtgtgacccaccctcggtgagaaagccttctct
1401 agcgatccgagaggcgtgccttggggtaccggatcccccggactgccgcctctgtctctgcctccgtgat
1471 gggagcgctgccgtagcgactcgctcgcagaggaccctcctcccctccccctcgacggggttcgcgggg
1541 gagagcgagggtttccgccggccaccgccgtggtggccgagcgcggctcgtcgcctactgtgtggcccgc
1611 gcctcccccttctgagtcgggggaggatcccgccgggccgggccctggcgtcccagcgggttgggacgtg
1681 gcggccggcgggcggtgggtgtgcgcgcgcggcgttccgtccggcgcgtgacccccctccgccgcgagtcg
1751 gctctccgcccgctcccgtgtgccgagtcgtgaccggtgcccacgaccgcgtttgcgtggcgcggggtcg
1821 ggccgcctggccctgggaaagcgtcccacggtgggggcgcgccggtctcccggagcgggaccgggccgga
1891 ggatggacgagagaatcacgagcgcggtggtggtgctggcgttgtcgggcttgtggctgcggtcgctttg
1961 gggcccccggtggcgggacccggggctcgcgaggcgggtctcggtgggggccgagggccgtccggcgtc
2031 ccaggcggggcgccgcgggaccgccctcgtgtctgtggcggtgggatcccgcggccgtgttttcctggtg
2101 gcccggccgtgcctgaggtttctccccgagccgccgcctctgcgggctcccgggtgcccttgcccgtctg
2171 tgccctcttccccgcccgccgcccgccgatcctctcttctccccgagcggctcaccggcttgacgtcggt
2241 tggtggccccgtctgggaccgaacccggcaccgccttgttgggggcgccgccgccggccactggttggcc
2311 cggtgtccgcgtccccggcgcgcgccttcgggaccgggtcggtggcgccccgcgtggggcccggtgggc
2381 ttcccggagggtttcgggggggtcggcctgtggcgcgcgtgcgagggaagagagggtgccgggggaccgg
2451 ccgcgactgcggcggtggtgggggggagccgcggggagcgccgagggccggtcggccgctccgggtgccg
2521 ctccgggtgccgcgcggggccgccttcccgctcggaggctgccggcggtgagacccccgcgtgtgtgtcc
2591 cggccgcggtcggctgcgcccgaggggtcccgtggcgtcccctccccgccggccgcctttctcgcgcct
2661 tccccgtcgcctcggcctcgcccgtggtccctcgtcttctcccggcccgctcttccgagccgggtcggcg
2731 cgtcccccgggtgcgcctcgcttcccgggcctgccgtggccctccctggaggcgtccgtcccgggcgtcg
2801 gcgtcggggagagcgcgtcctccccgcgtggcgttgccccgttcggcgcgcgcgtgcgcccgagcgcggc
2871 ccggtggtccctcccggacaggcgttcgtgcgatgtgtggcgcaggtcggcctccgcctttgcccggtcg
2941 ctcgccctttccccgggtcgggggtggggcccgggccggggcctcggccccggtcgcggtccctcgtcc
3011 cgggtgggggcgggcggttgccggccggctgcggtcgtcctcccttggccgtcgtgtggcgtgtgccacc
3081 cctgcgcccgcgcccgccggcggggctcggagccgggcttcggccgggctccgggccctcgaccggaccg
3151 gtgcgcgggcgctgcggcccgcacggcgcgactgtccccgggccgggcgccgcggtccgcctctcgctcg
3221 ccgcccgaacgtcggggccgccccgcggggcgggcggagcgccgtccccgcctcgccgccgcccacgggc
3291 gccggccgcgcgcgcgtgcgcgtggccgccggtccctcccgggcgcgggtcgggccgtccgcctcctcgc
```

3361 gggcgggcgcgacgagaaagggtctcgcgggtccgtggcgtgagcccggcgcgggggggtgggggcggt
3431 ggtcgtgtcgtgttgggggcgggtggtcggggcgtccggttcgccgcgccccgccccggcccccaccgt
3501 cccggccgccgccccgcgcccgctcgctccctcccgcccgcccgcccgtccgcggtccgtccgtccgtcg
3571 tcctcctcgcttccggggcgccgggcccgtcctcgcgaggcccccggccggccggccggccgcgtcgggg
3641 cctcgccgcgcttctacctacctggttgatcctgccagtagcatatgcttgtctcaaagattaagccatg
3711 catgtctaagtacgcacggccggtacagtgaaactgcgaatggctcattaaatcagttatggttcctttg
3781 gtcgctcgctcctctcctacttggataactgtggtaattctagagctaatacatgccgacgggcgctgac
3851 ccccttcgcgggggggatgcgtgcatttatcagatcaaaaccaacccggtcagcccctctccggccccgg
3921 ccgggggcgggcgccggcggctttggtgactctagataacctcgggccgatcgcacgccccccgtggcg
3991 gcgacgacccattcgaacgtctgccctatcaactttcgatggtagtcgccgtgcctaccatggtgaccac
4061 gggtgacggggaatcagggttcgattccggagagggagcctgagaaacggctaccacatccaaggaaggc
4131 agcaggcgcgcaaattacccactcccgacccggggaggtagtgacgaaaaataacaatacaggactcttt
4201 cgaggccctgtaattggaatgagtccactttaaatcctttaacgaggatccattggagggcaagtctggt
4271 gccagcagccgcggtaattccagctccaatagcgtatcttaaagttgctgcagttaaaaagctcgtagtt
4341 ggatcttgggagcgggcgggcggtccgccgcgaggcgagccaccgcccgtccccgcccccttgcctctcgg
4411 cgcccccctcgatgctcttagctgagtgtcccgcggggcccgaagcgtttactttgaaaaaattagagtgt
4481 tcaaagcaggcccgagccgcctggataccgcagctaggaataatggaataggaccgcggttctattttgt
4551 tggttttcggaactgaggccatgattaagagggacggccgggggcattcgtattgcgccgctagaggtga
4621 aattcttggaccggcgcaagacggaccagagcgaaagcatttgccaagaatgttttcattaatcaagaac
4691 gaaagtcggaggttcgaagacgatcagataccgtcgtagttccgaccataaacgatgccgaccggcgatg
4761 cggcggcgttattcccatgacccgccgggcagcttccgggaaaccaaagtctttgggttccgggggggagt
4831 atggttgcaaagctgaaacttaaaggaattgacggaagggcaccaccaggagtggagcctgcggcttaat
4901 ttgactcaacacgggaaacctcacccggcccggacacggacaggattgacagattgatagctctttctcg
4971 attccgtgggtggtggtgcatggccgttcttagttggtggagcgatttgtctggttaattccgataacga
5041 acgagactctggcatgctaactagttacgcgacccccgagcggtcggcgtcccccaacttcttagaggga
5111 caagtggcgttcagccacccgagattgagcaataacaggtctgtgatgcccttagatgtccggggctgca
5181 cgcgcgctacactgactggctcagcgtgtgcctaccctacgccggcaggcgcgggtaacccgttgaaccc
5251 cattcgtgatggggatcggggattgcaattattccccatgaacgaggaattcccagtaagtgcgggtcat
5321 aagcttgcgttgattaagtccctgccctttgtacacaccgcccgtcgctactaccgattggatggtttag
5391 tgaggccctcggatcggccccgccggggtcggcccacggccctggcggagcgctgagaagacggtcgaac
5461 ttgactatctagaggaagtaaaagtcgtaacaaggtttccgtaggtgaacctgcggaaggatcattaacg
5531 gagccgaaggggggcgcgaggccgcggcggcgccgccgcgcgcttccctcccccccacccccccgccacaa
5601 cgcggcgcgtgcgcgggcgggcccgtgtgccgttcgttcgttcgttcgctgcccggccccgccggccgc
5671 gagagccggaggactcgggagggcgacgtgggggggagagcgagagagacagaaagaagggggcgcgtgt
5741 tcgcgtgcgcgtgtcgcggggccggcgggccggtgggcggcggggagcggtccccggccgcggccccgac
5811 ggacgtgtgtgtcggcgggcgcgggggcggttctcggcggcgtcacggcgggtttgggggggggggtctcg
5881 gtgccctcctccccgccggggcccgtcgtccggccccgccgcgcgccggctccccgtcgtcggggccggg
5951 ccggattcccgtcgccgcctccgccgcgcgccgctccgcgccaccgggcacggccccgctcgctctcccc
6021 ggccttcccgctagggcgtctcgagggtcgggggccggacgccggtcccccctcctcgtccgcccccccc
6091 acccccgccgtccaggtacctagcgcgttccggcgcggaggtttaaagacccctgggggatcgcccgt
6161 ccgcccgcgggtcgggggcggtggtgggcccgcgggggagtcccgtcgggaggggcccggcccctcccgc
6231 gcctcccccgcggactccgccccccggccggggccgcgccgcctccgccggtgccgcggtcgcggcggcc
6301 gtcgggtgggggctttacccggcggccgtcgcgcgcgcgtgccgcgcgtgtggcgtgcgccccgcgccgt
6371 gggggcgggaacccccccgggcgcctgtggggtggtgtccgcgctcgcccctgcgtgggcggcgcgcgcct
6441 ccccgtggtgtgaaaccttccgacccctctccggagtccggtcccgtttttgctgtctctctggccggc
6511 ctgaggcaaccccctctcctctgggggggggacgtgccgcgccaggagggcctcccggtgtgtgtgtcgg
6581 gagcgccctcgccaaatcgacctcgtacgactcttagcggtggatcactcggctcgtgcgtcgatgaaga
6651 acgcagctagctgcgagaattaatgtgaattgcaggacacattgatcatcgacacttcgaacgcacttgc
6721 ggccccgggttcctcccggggctacgcctgtctgagcgtcgcttgccgatcaatcgccccccgggggtgc
6791 ctccgggctcctcggggtgcgcggctgggggttccctcgcagggcccgccggggccctccgtcccctaa

6861 gcgcagacccggcgacgtccgccctcctcttcccgccgcgcccgcccttccccctcccccccgcgggcc
6931 ctgcgtggtcacgcgtcgggtggaggggggggaggggggcgcccggtcgagagcgagagagagagagggag
7001 ggcggcgccgccgcccgcgaagacggagaggggaagagagagccggctcgggccgagttcccgtggccgc
7071 cacctgcggtccgggttcctccctcgggggctccctcgcgccgcacgcggctctcggggttcggggttc
7141 gtcggccccggccgggtggaaggtcccgtgcccgtcgccgccgtcgtcgtcggcggtgggggggcgtgtt
7211 gcgtgtggtgggggggggaggaaggcgggtccggagggaaggggtctggcggggagagagagggtgggg
7281 gagcgcgtcccggtcgccgcggttcgccgcccgcccctggtggcggcccggcgtccggccgaccgccgct
7351 cccgcgcccctcctcctcctccccgccgcccctcctccggggccccgtcctcctcgccctcccgcacgca
7421 cgcccgcccgcccggctcgcctcgcggcgcgtcgtccggggccggaagcccgccccgcggcccgcccggc
7491 cgcgcccgtggccgcggtcccggggttcgcgtgtccccggcggcgacccgcgggacgccgcggtgtcgtc
7561 cgccgtcgcgcgcccgccccggctcgcggccgcgccgcgccgtgccggggccccgtcccgagcttccgc
7631 gtcggggcggggcggctccgccgccgcgtcctcggacccgtccccgcgacctccgcgggggacgggtcgg
7701 ggtgtgtgcggtgcccgtcccgtccccggcccgtgcccctccctccggtcgtcccgctccggcggggcgg
7771 cgcgggggtgccgtcggccgcggctctctctctcccgtcgcctctcccctcgccgggcccgtctctcgac
7841 ggggcgtcgggcgggcggtcggtcgggccggcgcgatgccgtccgtccgcgtccgccgagcggcccgtcc
7911 ccgcccgccggccccgttcccctccgagacgcgacctcagatcagacgtggcgacccgctgaatttaag
7981 catattagtcagcggaggaaaagaaactaaccaggattccctcagtaacggcgagtgaacagggaagagc
8051 ccagcgccgaatccccgccccgcggcggggcgcgggacgtgtggcgtacggaagacccactccccggcgc
8121 cgctcgtgggggcccaagtccttctgatcgaggcccagcccgtggacggtgtgaggccggtagcggccc
8191 ccggcgcgccgggcccgggtcttcccggagtcgggttgcttgggaatgcagcccaaagcgggtggtaaac
8261 tccatctaaggctaaataccggcacgagaccgatagtcaacaagtaccgtaagggaaagttgaaaagaac
8331 tttgaagagagagttcaagagggcgtgaaaccgttaagaggtaaacgggtggggtccgcgcagtccgccc
8401 ggaggattcaacccggcggcgggtccggccgtgtcggcggcccggcggatctttcccgccccccgttcct
8471 cccgacccctccacccgccctccctttcccccgccgcccctcttcctcctcccccggaggggcgggctcc
8541 ggcgggtgcgggggtgggcgggcggggccgggggtggggtcggcgggggaccgtcccccgaccggcgacc
8611 ggccgccgccgggcgcatttccaccgcggcggtgcgccgcgaccggctccgggacggctgggaaggcccg
8681 gcggggaaggtggctcgggggccccgtccgtcctcctcctcctcccaccccgtctccgcccccggccc
8751 cgcgtcctcccccgggagggcgcgcgggtcggggcggcggtggcggcggcggcggcggcggtggcggcgg
8821 gaccgaaaccccccccgagtgttacagcccccccggcagcagcactcgccgaatcccggggccgagggagcg
8891 agacccgtcgccgcgctctcccccctcccggcgcccacccccgcggggatatcctccgcgaggggggtct
8961 cccccgcgggggcgcgccggcgtctcctcgtgggggggccgggccacccctcccacggcgcgaccgctct
9031 cccacccccctcctccccgcaacccccctctcccggcgacgggggagggccgcgcgcgggtcgggggcgg
9101 ggcggactgtccccagtgcgccccgggcgggtcgcgccgtcgggcccgggggggaggttctctcggggcc
9171 acgcgcgcgtccccgaagagggggacggcggagccgagcgcacggggtcggcggcgatgtcggccaccc
9241 acccgacccgtcttgaaacacggaccaaggagtctaacacgtgcgcgagtcgggggctcgcacgaaagcc
9311 gccgtggcgcaatgaaggtgaaggccggcgcgcgctcgccggccgaggtgggatcccgaggcctctccag
9381 tccgccgagggcgcaccaccggcccgtctcgcccgccgcgccggggaggtggagcacgagcgcacgtgtt
9451 aggacccgaaagatggtgaactatgcctgggcagggcgaagccagaggaaactctggtggaggtccgtag
9521 cggtcctgacgtgcaaatcggtcgtccgacctgggtataggggcgaaagactaatcgaaccatctagtag
9591 ctggttccctccgaagtttccctcaggatagctggcgctctcgcagacccgacgcacacccccccacgca
9661 gttttatccggtaaagcgaatgattagaggtcttggggccgaaacgatctcaacctattctcaaacttta
9731 aatgggtaagaagcccggctcgctggcgtggagccgggcgtggaatgcgagtgcctagtgggccactttt
9801 ggtaagcagaactggcgctgcgggatgaaccgaacgccgggttaaggcgcccgatgccgacgctcatcag
9871 accccagaaaaggtgttggttgatatagacagcaggacggtggccatggaagtcggaatccgctaaggag
9941 tgtgtaacaactcacctgccgaatcaactagccctgaaaatggatggcgctggagcgtcgggcccatacc
10011 cggccgtcgccggcagtcgagagtggacgggagcggcgggggcggcgcgggcgtgtgcgcgcgcgcgtgt
10081 gtgtgtgcgtgtgtgtcggagggcggtggcggcgggggtggggtcctccccctccccacgccgcctccc
10151 ctcctcccacccaccaccgccgccgccacccccgctccccgcccccggagccccgcggacgctacgccgc
10221 gacgagtaggagggccgctgcggtgagccttgaagcctagggcgcgggcccgggtggagccgccgcaggt
10291 gcagatcttggtggtagtagcaaatattcaaacgagaactttgaaggccgaagtggagaagggttccatg

10361 tgaacagcagttgaacatgggtcagtcggtcctgagagatgggcgagcgccgttccgaagggacgggcga
10431 tggcctccgttgccctcggccgatcgaaagggagtcgggttcagatccccgaatccggagtggcggagat
10501 gggcgccgcgaggcgtccagtgcggtaacgcgaccgatcccggagaagccggcgggagccccggggagag
10571 ttctcttttctttgtgaagggcagggcgccctggaatgggttcgccccgagagaggggcccgtgccttgg
10641 aaagcgtcgcggttccggcggcgtccggtgagctctcgctggcccttgaaaatccgggggagagggtgta
10711 aatctcgcgccgggccgtacccatatccgcagcaggtctccaaggtgaacagcctctggcatgttggaac
10781 aatgtaggtaagggaagtcggcaagccggatccgtaacttcgggataaggattggctctaagggctgggt
10851 cggtcgggctggggcgcgaagcggggctgggcgcgcgccgcggctggacgaggcgccgccgccccccca
10921 cgcccggggcacccccctcgcggccctccccgccccacccgcgcgcgccgctcgctccccgcccccc
10991 ccgcgccctctctctccctctctctcccgctccccgtcctcccccctccccgggggagcgccgcgtgggg
11061 gcggcgggtgtgggggagaagggtccgggacgggcgggggcgcggcggcggccgccgcggggcccggcg
11131 gcgggggcacggtcccccgcgaggggggcccgggcacccgggggggccggcggcggcggcgactctggacg
11201 cgagccgggcccttcccgtggatcgccccagctgcggcgggcgtcgcggccgcccccggggagcccggcg
11271 ggcgccggcgcgccccccctctcccgtgtcgtcgctccgtcgcgggggcggggagcggtcgggcggcggc
11341 ggcggcggtcggtgggcggcggggcggggcggttcgtccccccgccctcccccccggccccgtccgtccc
11411 ccgttcctcccccgtcctcgccgcgcggcggcggcggcggcggagttggggccgcgggccggtcccc
11481 cccgccgggtccgcccccggggccgcggttccgcgcggcgcctcgcctcggccggcgcctagcagccgac
11551 ttagaactggtgcggaccaggggaatccgactgtttaattaaaacaaagcatcgcgaaggcccgcggcgg
11621 gtgttgacgcgatgtgatttctgcccagtgctctgaatgtcaaagtgaagaaattcaatgaagcgcgggt
11691 aaacggcgggagtaactatgactctcttaaggtagccaaatgcctcgtcatctaattagtgacgcgcatg
11761 aatggatgaacgagattcccactgtccctacctactatccagcgaaaccacagccaagggaacgggcttg
11831 gcggaatcagcggggaaagaagaccctgttgagcttgactctagtctggcacggtgaagagacatgagag
11901 gtgtagaataagtgggaggcccccggcgcccccccggtgtcccgcgaggggttccggggcggggtccgc
11971 cggccctgcgggccgccggtgaaataccactactctgatcgttttttcactgacccggtgaggcgggggg
12041 gcgagccccgaggggctctcgcttctggcgccaagcgcccggccgcgcgccggccgggcgcgacccgctc
12111 cggggacagtgccaggtggggagtttgactggggcggtacacctgtcaaacggtaacgcaggtgtcctaa
12181 ggcgagctcagggaggacagaaacctcccgtggagcagaagggcaaaagctcgcttgatcttgattttca
12251 gtacgaatacagaccgtgaaagcggggcctcacgatccttctgacctttgggttttaagcaggaggtgt
12321 cagaaaagttaccacagggataactggcttgtggcggccaagcgttcatagcgacgtcgcttttgatcc
12391 ttcgatgtcggctcttcctatcattgtgaagcagaattcaccaagcgttggattgttcacccactaatag
12461 ggaacgtgagctgggtttagaccgtcgtgagacaggttagttttaccctactgatgatgtgttgttgcca
12531 tggtaatcctgctcagtacgagaggaaccgcaggttcagacatttggtgtatgtgcttggctgaggagcc
12601 aatggggcgaagctaccatctgtgggattatgactgaacgcctctaagtcagaatcccgcccaggcggaa
12671 cgatacggcagcgccgcggagcctcggttggcctcggatagccggtcccccgcctgtccccgccggcggg
12741 ccgcccccccccccccaccgcgccccgcgcgcggggaggaaggcgcgtgccccgccgcgcgccgggaccgg
12811 ggtccggtgcggagtgcccctcgtcctgggaaacggggcgcggccggaaaggcggccgccccctcgcccg
12881 tcacgcaccgcacgttcgtggggaacctggcgctaaaccattcgtagacgacctgcttctgggtcggggt
12951 ttcgtacgtagcagagcagctccctcgctgcgatctattgagagtcagccctcgacacaagggtttgtcc
13021 gcgcgcgcgcgcgtgcgcgcgggggcccggcggggcgtgcgcgtccggcgccgtccgtccgtccgttcg
13091 tcttccttcctcccctcccggcctctcccgccgaccacgggcgtgggagggaggggggagggagggcgcgc
13161 gtccccggtcggcgcgccccgcttcttcggttcccgcctcctccccgttcaccgccggggcggctcgtcc
13231 gctccgggccgggacggggcccgggggagcgcgtgggtgggaaccgcggaggcggccgcgccgagccgggc
13301 ccgtggcccgccggcccccgtcccgggggggggggttggccgcgcgggccccggtggggcggccacctcgg
13371 ggttccggccctcgcgcgtccttcctcctctctcctccgcacgggtcgaccagcagaccgcgggtggcgg
13441 gcggcgggcggcgaggcccccgggggcgtccccgcgcccggcccgcctccgctcgccacctctccccggt
13511 cggaacctccggggtcgaccagctgccgcccgcgagctccggacttagccgccgctgtctccacgtgtcc
13581 cgggtcgaccagcaggcggccgccggacgctgcggcgcggcgatgcgagggcgccgattcccgttcacga
13651 gccggcgacctccgccggcctcggccttcggtggagctgggacgacgcggaactccctgccccgcatttt
13721 tttcagccccactgcgagtttgcgtccgcggtacttttaagagggagtcactgttgccgtcagccggcaa
13791 tacttcctccctttttgctttttggttctgtctcccgtttttttgttttgttttgttttgttttccccccc

13861 ccccttttctctcgctcgctctctcgctctctccctcgctcttctgtctctgtctctctctgtctctctg
13931 tctctctgtctctctgtctctctctctctcctctctctctctctctctctctctctccttccttgg
14001 tgccttctcggctcgtgagacttagccgctgtctccccgggtcgaccggcgggccttctcgaccgagcgg
14071 cgtgtaggaatgcccgtcgggacgagtcggacccggggcgtccgcgtctcggtcggaacctccggggtcg
14141 accagctgccgcccgtgagctccggacttagccgccgctgtctccacgtgtcccgggtcgaccagcaggc
14211 ggccgccggacgctgcggcgcggcgatgcgagggcgccgattcccgttcacgagccggcgacctccgccg
14281 gcctcggccttcggtggagctgggacgacgcggaactccctgccccgcatttttttcagccccactgcga
14351 gtttgcgtccgcggtacttttaagagggagtcactgttgccgtcagccggcaatacttcctccctttttg
14421 ctttctggttctgtctcccgtttttttgttttgttttgttttgttttttcccccccccccctttttctctcgct
14491 cgctctctcgctctctccctcgctcttctgtctctgtctctctctgtctctctgtctctctgtctctctg
14561 tctctctctctctctctctctctctctctctctctctctctctctctctctctctccttccttggtgc
14631 cttctcggctcgtgagacttagccgctgtctccccgggtcgaccggcgggccttctcgaccgagcggcgt
14701 gtaggaatgcccgtcgggacgagtcggacccggggcgtccgcgtctcggtcggaacctccggggtcgacc
14771 agctgccgcccgtgagctccggacttagccgccgctgtctccacgtgtcccgggtcgaccagcaggcggc
14841 cgccggacgctgcggcgcggcgatgcgagggcgccgattcccgttcacgagccggcgacctccgccggcc
14911 tcggccttcggtggagctgggacgacgcggaactccctgccccgcatttttttcagccccactgcgagtt
14981 tgcgtccgcggtacttttaagagggagtcactgctgccgtcagtcggcaatacttcctccctttttgctt
15051 tttggttttgtctcccgtttttttgttttgttttgttttgttttttcccccccccccctttttctctcgctcg
15121 ctctctcgctctctccctcgctcttctgtctctgtctctctctgtctctctgtctctctgtctctctctc
15191 tctctctctctctctctctctctctctctctctctctctctctctctccttccttggtgccttctcggct
15261 cgtgagacttagccgctgtctccccgggtcgaccggcgggccttctcgaccgagcggcgtgtaggaatgc
15331 ccgtcgggacgagtcggacccggggcgtccgcgtctcggtcggaacctccggggtcgaccagctgccgcc
15401 cgtgagctccggacttagccgccgctgtctccacgtatcccgggtcgaccagcaggcggccgccggacgc
15471 tgcggcgcggcgatgcgagggcgccgattcccgttcacgagccggcgacctccgccggcctcggccttcg
15541 gtggagctgggacgacgcggaactccctgccccgcatttttttcagccccactgcgagtttgcgtccgcg
15611 gtacttttaagagggagtcaccgttgccgtcagccggcaatacttcctccctttttgcttttttggttctg
15681 tctcccgtttttttgttttgttttgttttgttttgttttttcccccccccctttttctctcgctcgctctct
15751 cgctctctccctcgctcttctgtctctgtctctctctgtctctctgtctctctgtctctctctctctctc
15821 tctctctctctctctctctctctctctctctccttccttggtgccttctcggctcgtgagacttagccgc
15891 tgtctccccgggtcgaccggcgggccttctcgaccgagcggcgtgtaggaatgcccgtcgggacgagtcg
15961 gacccggggcgtccgcgtctcggtcggaacctccggggtcgaccagctgccgcccgtgagctccggactt
16031 agccgccgctgtctccacgtatcccgggtcgaccagcaggcggccgccggacgctgcggcgcggcgatgc
16101 gagggcgccgattcccgttcacgagccggcgacctccgccggcctcggccttcggtggagctgggacgac
16171 gcggaactccctcccccgcatttttttcagccccactgcgagtttgcgtccgcggtacttttaagaggga
16241 gtcactgctgccgtcagtcggtaatacttcctccctttttgctttttggttttgtcttgcatttttttt
16311 ctttctttctttctttttctttctttctttccttctctttctctctctctctctctctctctctctct
16381 ctctctctctctctctctctccctccttctccctccctccctctccctctccctctccctctccctctcc
16451 ttctccttctccttctccttccttggtgccttctcggctcactgctaccgctgcctctgcctccacggat
16521 caagcaaacggcaggttttctacttcgaggaaagacggaatttcaccatgttggccgggccggtctcgaa
16591 ctcccgacctagtgatccgcccgcctcggcctcccaaagactgctgggagtacaggcgtgagccaccacg
16661 cctggccgattccttccttttttcaatcttattttctgaacgctgccgtgtatgaacgtacatttataca
16731 cacacacacacacacacacacacacacacacacacacgtagtgataaaactctgtaaatgatatttccataa
16801 ttaatacgtttatattatgttacttttaatggatgagtatgtattaaaaccccatttcatttacctacac
16871 gtgtatgtatatccttcctcccttccttccttcatttattatttattcataattttcgtgtatttatttt
16941 cttttcttttggggccggcccgcctggtcttctgactctgggctctggtgacctcagcctcccacgtagc
17011 tgggactacagggatctcttaagcccgggagggagaggctaacgtgggctgtgatcgcgcacttccactc
17081 cagcttacgtgggctgcggtggggtggggtggggtggggtgcggtggggtgcagagaaaacaatg
17151 gattgcgatcttaattgccttttagcttcattcataccctgttatttgcttgtttattctcatgggttat
17221 tctgtgtcactgtcatgttcctcgtttgcttgcctgcttgcttgtttatttccttccttccttccttcct
17291 tccttccttccttccttcccctcctcctcctcctcctcctcctcctcctcctcctcctcctcctcctcctc

17361 cctcccttactggcagggtcttcctctgtctctgccgcccaggatcaccccaacctcaatgctttggacc
17431 gaccaaacggtcgtcctgcctctgatccctcccatccccattacctgagaccacaggcacgcaccaccac
17501 accggctgacttctatgttgcttctcctgttttccgtaggtaggtatgtatgtgtgggtatgtatgtatg
17571 tacgtatgtatgtacgtatgtatgtatgtatgtgagtgagatgggtttcgggattctatcatgttgccca
17641 cgctggtctcgaactcctgtcctcaagcaatccgcctgcctgcctcggccgcccacactgctgctattac
17711 aggcgtgagacgctgcgcctggctccttctacgttgcctgcctgcctgcctgcctgcctgcctgcc
17781 tacctatcaatcgtcttctttttagtacggatgtgctctcgctttgttgtccatgctctgggcacacgtg
17851 atctctttttaaacttctatgattatgatgattgtaggcgtcatctcacgtgtcggggtgatctccaact
17921 tttaggctccagagatcctcccgcatcggcctcccggagtgctgtgatgacacgcgtgggcacggtacgc
17991 tctggtcatgtttgtcatgggtcggttctttccgtttttaatacggggactgcgaacagataaaatgttc
18061 acacgcatctcaccgatccgccttttcgttctttctttttttttctctctttagacggagtttcactctt
18131 gtcgcccagggtggaggacgatggcggatctcggctcaccgcaccctccgcctcccaggttcaagggatt
18201 ctcctgcctcagccttcccgagtagctgcgatgacagcgatgagccatcgtgcctggctaatttttctat
18271 tcctagtacagatggggtttctccatcttggttaggctggtcttcaacttccgaccgttggaggatctta
18341 actttcttggtggtggttgcttttctttttctttttttctttctttctttccttctccttcccccccc
18411 ccaacccccccttgtcgtcgtcgtcgtcctcctcctcctcctcctcctcctcctctttcatttctttcagctgg
18481 gctctcctacttgtgttgctctgttgctcacgctggtctcaaactcctggccttgacacttctcgcgtca
18551 catccgccgtctgcttgttgaaatgagcatctctcgtaaaatggaaaagatgaaagagataaacacgaag
18621 acggaaagcacggtgtgaacgtttctcttgccgtctcccggggtgtaccttggacgcggaaacacggagg
18691 gagcttggctgagtgggttttcggtgccgaaacctcccgagggcctccttccctctccccccttgtccccg
18761 cttctcccccagccgaggctcccaccgccgccctggcattttccacaggagaggtatagggagaggactg
18831 acacgccttccgcatctatatcctgccggacgtctctggctcggcgtgccccaccggctacctgccacct
18901 tccagggagctctgaggcggatgcgaccacacccccgtccccccccgtcacgtcccgctaccctcccccgg
18971 ctggcctttgccgggcgaccccaggggaaccgcgttgacgctgccctcggatcctccggcgaagacttcc
19041 accgcatgccccgggtgggccggttgggatgagactggaccaccccggaccgtgctgttcttgggggtgg
19111 gctgacgtacggggtggactggcggccccagcattgtaaagggtgcgtgggtatggaaatgtcacgtagg
19181 atgccctccttcccttcggtctgccttcagctgcctcaggcgtgaagacaacttcccatcggaacctctt
19251 ctcttccctttctccagcacacagatgagacgcgtgagagggagaaacagctcaatagatactgctgacc
19321 ttcatttgtggaatcctcagtcgtctacacacaagacaggtgactaggcagggacacagatcaaacacta
19391 tttccgggtcctcgtggtgggattggtgtctctcctctctctctctctctctctctctctctctct
19461 ctctctctctcacacacacacacgcgcgcgcgcgcgcacgcacgcatgcacacacgcacacacacacaca
19531 cacacacacacacatttccatatctagttcacaaagcacactcacttaacctttcacagtacgcaggct
19601 gagtaaaacccacccccaccttccacccgttggctgacgaaacccccttctctacaatggatgaaaaagatg
19671 atctgggccgggcacgctagctcacgcctgtcattccggcactttgggaggccgaggccggtggatcgct
19741 tggggccgggagttcgagaccaggctggccgacgtggcgaaaccccgtctctctgaaaaaatagaacgat
19811 tagccgggcctggtggcgtgggcttggaatcacgaccgctcgggagactggggcgggcgagttgttccaa
19881 ccggggaggccgaggttgcgatgagctgagatcgtgccgtggcgatccagcctggatgacggagcgagac
19951 cccgtctcgatacaatcatgatgtgattataagatgagttgtgcgcggtgatggccgcctgtagtcgcag
20021 ctactcgggaggctgagatgaggagaggatcacttgaggccccacaggtcgaggcttcggtcggccgtga
20091 cccactgtatcctgggcggtcaccggtcaaggagatatgccccctccccgtttgcttttcttttcttccc
20161 ttctcttttcttcttttttcttctctttctttttcttctttctttctttctttctttctttctttctttc
20231 tttctttctctctctctctcttcccctctttctttcctgccttcctgccttcctgccttcctgccttt
20301 cttcttttcttcttcctcccttcctcccttcctttctttcctcccgcctcagcctcccaaagtgctggga
20371 tgactggcgtgaggtaccatgcctgcttggcctaaagagaccctcttggaaagtgagacacagagagcgc
20441 cttccagtgatctcatggactgatttagagacggcatctcgctccgtcaccccggcagtggtgccatcgt
20511 aactcactccctgcagcgtggacgctcctggactcgagcgatccttccgcctcagcctccagagtacaga
20581 acctgggaccgcaggcacgcgccactgtgcccacaccgttttaaattttcttttcctcccccgcccgaga
20651 cagagtttcgctctggtggcctagactgcagtgcggtggcgcgatcttggctcgccgcaacctctgcctc
20721 ccggtttcaagcgattctcctgcatcggcctcctgagtagccgggattgcgggcgtgcgctgccacgtct
20791 ggctgattttgtatttttagtggagacggggcttctccatgtcgatcgggctggtctcgaactcccgacc

20861 tcaggtgatccgccctccccggcctccggaactgctgggatgacaggcgtgagccaccgcgcccggcctt
20931 cattttaaatgttttcccacagacggggtctcatcatttcgttgcaaccctcctgcccggcgtctcaaa
21001 gtgctggcgtgacgggcgtgagccactgcgcctggactccggggaatgattcacgaccacggtcgctgta
21071 ctgattctttctttctttctttctttctgtctgtctgtctgtctgtctgtctgtctgtctgtctttcttt
21141 ctttctttccttttctttccttttctttctttttctttctttctttcttatttattgataaattattttatga
21211 tttatttgtgtacttattttcagacggagtctcgctctgggcggggcggggcgaggcgaggcacagcgca
21281 tcgctttggaagccgcggcaacgcctttcaaagccccattcatatgcacaaagccctattcccttcctgg
21351 agttggagctgatgccttccgtagccttgggcttctctccattcggaagctttgacaggcgcaaccccac
21421 ccagaggctgcctgcggctgaggattaggggtgtgttgcggctgaaaagtgggtcccctagttttgata
21491 cctcagtggacacatcccccgaccgccatcgcttgctcgccctctgagatcccccgcctccaccgccttg
21561 caggctcaccgcttactttcatttctttcttcctttcttgcgtttcaggaggggtgcgggaatgagggt
21631 gtgtgtggggaggggtgcggggtggggacggaggggagcgtcctaagggtcgatttagtgtcatgcctc
21701 tttcaccgccaccaccgaagatgaaagcaacgatcggctacataccgcgtgttctcatctagaagtggga
21771 acttacagatgagagttcttgcatgggcagaacgagggggaccagggacgcggaagcctgcttgagggag
21841 gaggggtggaaggagagacagcttcaggaaaaaaacaaaacacgaatactgtcggacacagcactgacta
21911 cccgggtgatgaaatcatctgcacactgaacacccccgtcccaagtttacctatgtcacaatcttgcaca
21981 tgtatgcttgaacgacaaataaaagttaggggggagagaggagagagagcgagagagcgagagcgagagc
22051 gagagagagagcgagagcgagagagagagagagagcgagagagagagagagagagagagagagagaga
22121 gagagagacacaagtacaaccaaacaccacctccttgacctgagtcagggggtttctggccttttgggag
22191 aatgttcagcgacaatgcagtatttgggccggttcttttgtttcttcttcttcttctttttctttctttt
22261 tttggactgagtctctctcgctctgtcacccaggctgcggtgcagtggcgctctctcggctcactggaac
22331 ctctgcttcccgggctccagtgattcttcttcggtagctgggattacaggcacgcaccaggatggccggc
22401 tcatattcttatttttagtagagaggggttttctccacgttggccacgctggtatcgaactcttgacctc
22471 aaatgatgcgccttcctgggcctcccaaagtgctggaaacgaccggcctgagccgccgggatttcagcct
22541 ttaaaagcgcgggccctgccaactttcgctgcggcccttacgctcagaaggacgtgtcctctctgccata
22611 ggttgactccttgagtcccctaggccattgcactgtagcctgggcagcaagagccaaactccgtcccccc
22681 acctccccgcgcaaataataactaactaacgaacgaacgaactaactaagtaaataaaatctctacacgt
22751 cacctctaagtgtgtgttcccgtgaggagtgatttctaagaaatggcactgtacactgaacgcagtggct
22821 cacgtctgtcatcccgaggtcaggagttcgagaccagcccggccaacgtggtgaaaccccgtctctactg
22891 aaaacacgaaattgagtcaggtgccgtggggccggcacctgtcatcccagctactcgggaggctgaggcg
22961 gaagaattgcttgaacctggcaggcggaggttgcagtgacccaagatcgcaccactgcactacggcctgg
23031 gcgacagagtgagacccggtctccagataaatacataaatacatacatacgtacgtacgtacgtacatac
23101 atacgtacagatatacaagacagaaaacaagaaaagaaaggaaagagaaaatgaaagaaaaggcactgta
23171 tcgctactgggctaggaccgtctctctgtctgtttctctctgttcgtctctgtctttctctctgtgtctc
23241 tttctctgtcggtctgcctgtctgtctgtctgtctgtctctttctttctgtctgtctctgtctctgtctt
23311 tgtctctctctctcccctctctgcctgtctcactgtgtctgtcttctgtcttactctctttctctccccgt
23381 ctgtctctctctctctccctccctgtctgtttctctctctctttctgtctgtttctgtctctctctgt
23451 ctgtctatgtctttctctgtctgtctgcctctctctttctttttctgtgtctctctgtcggtctctctct
23521 ctctctgtgcctatcttctgtcttactctctctttctctgcctgtctgtctgtctctctctgtctctccc
23591 tccctgtccgtttctctctctctctctcccccctctccctgtctgtttctctccgtctctctctctttctg
23661 tctgtttctcactgtctctctctgtccatctctgtctttctatgtctgtctctttctctgtcagtctgtc
23731 agacacccccgtgccgggtagggccctgcccctccacgaaagtgagaagcgcgtgcttcggtgcttcga
23801 gaggccgagagaaatctagacaggcgggccttgctgggcttccccactcggtgtatgatttcgggaggtc
23871 gaggccgggtccccgcttggatgcgaggggcattttcagacttttctctcggtcacgtgtggcgtccgta
23941 cttctcctatttccctgataagctcctcgacttaaacataaacggcgtcctaagggtcgatttagtgtca
24011 cgcctctttcaccgccaccagcgaagatgaaagcaaagatcggctgaataccgcgcgttctcatctacaa
24081 gtgggaacttacagatgagagttcttgcatgggcagaacgaggggaccagggacgcggaagcctgcttg
24151 agggaggaggggtggaaggagagacagcttcaggaaaaaaacaaaacacgaatactcttggacacagcac
24221 tgactacccgggtgacgaaatcatctgcacactgaacacccccgtcacaagtttacctatgtcacaatct
24291 tgctcatgtatgcttgaacgacaaataaaagttcgggggggagaagagaggagagagagagagagacgg

24361 ggagagagagacggggagagagagaggggcggggagagagagagagagacagagacagagacagagagag
24431 agagagagagagagagaagtaaaaccaaacaccacctccttgacctgagtcagggggtttctggcctttt
24501 gggagaacgttcagcgacaatgcagtatttgggcccgttctttttttttttttttttttttcttttctt
24571 tctttttttggactgagtctctctcgctctgtcacccaggctgcggtgcagtggcgctctctcggctca
24641 ctgaaacctctgcttcccgggttccagtgattcttcttcggtagctgggattacaggcgcgcaccatgac
24711 ggccggctcgtagttctatttttagtagagatggggtttctccacgttggccacgctggtctcgaactcc
24781 tgacctcaaatgatccgccttcctgggcctcccacagtgctgggacgacaggcctgagccgccgggattt
24851 cagcctttaaaagcacgggccctgccaactttcgctgcggcccttacgctcagaaggacgtgtcctctct
24921 gccataggttgactccttgagtcccctaggccattgcactgtagcctgggcagcaagagccaaactccgt
24991 ccccccacctccccgtgcaaaaataactaactaactaacaactaactaaaatctctgcacgtcacccat
25061 aagtgtgtgttcccatgagtgatttctaagaaatggcactctacactgaacgcagtggctcacgtctgtc
25131 atcccgaggtcaggagttcgaggccagcctggccaacgtggtgaaaccccgtctctactgaaaatacgaa
25201 actgagtcaggcgccgtggggcaggcacctgtaaccccagctactcgggaggctgagacggaagaattgc
25271 ttgaacctggcaggcggaggttgcagtgacccaagatcgcgccactgcactacagcctgggagacagagt
25341 gagacccggtctccggatacatacatacatacatacatacatacatacatacatacatacatacataaaa
25411 gaaagacaaaagaaaagaaagagaaaatgaaagaaaaggcactgtatcgctacggggctaggaccttct
25481 ctctgtctgtttctctctgtctctctctgttcgtctctttctctctgtgtctctttctctgtctgtctgt
25551 ctgtctctttcttgctctctgtctctgtctttgtctctctctctccctctctgcctgtctcactgtgtct
25621 gtcttctgtcttactctctttctctccccgtctgtctctctctctctctctccctccctgtctgtttc
25691 tctctctctctctctctgtctctgtctttctctgtctgtccctttctctgtctgtctgcctctctctt
25761 tctctttctgtgtctctctgtctctctctctgtgcctatcttctctcttactctctttctctgcctatct
25831 gtctgtttctctgtgtctctctccctgcctttctgtttctctctctctctccctctctcgctctctctgt
25901 ctttctctctttctctctgtgtctctgtctccctctgtccatctctgtctttttctgtctgtctctctct
25971 ttctttctgtctgtctctgtctctctctctctctctctctctctctgcttgtctctctcactgtgtct
26041 gtcttctgtcttactctctttctctgcctgtccatctgtctgtctctctccctccctttctgtttctctc
26111 tctctcgctctctctgcctgtttctctcttgctctgtctgtctctgtctttctctgtctgtctctttctc
26181 tctgtctctgtctctgtctctctctctctctgcctgtctcactgtgtctgccttctgtcttattctcttt
26251 ctctctctgtctgtctctctctctctcccttcctgtctgttctctctctctctctctctctctctttc
26321 tgcctgtttctctctgtctgtctctgtctttgtctgtctgcctctctctttctttttctgcgtcgctctg
26391 tgtctctctctctctctctctgttcctatcttctgtcttactctgtttccttgcctgcctgtctgtctct
26461 ctgtctgtctgtctgtctctctctcccccctccctttctctctctctctctctctttctgtctgtttctct
26531 ctgtctctctgtccatctctgtctttctctgtctgtctctctctttctctctgtctctctctctgcctct
26601 ctctctctctctctgtctgtctctctcactgtgtgtgtctgtcttctgtcttactctccttctctgcctg
26671 tccgtctctccctctctctccctccctttctgtttctctctctctctctctttctgtctgtttctctc
26741 tttctctctctgtctgtctctttctctgtctgtctgtctctctctttctttttctctgtctctctctgtg
26811 tctgtctctctgtctgtgcctatcttctgtcttactctccttctctggctgtctgcctgtctctctctct
26881 ctctctgtctgtctccctccctccctccctccctgtctgtctgtttctctctctgtctctctctctctgt
26951 ccatctctgtctgtctctttctctttctctctctttctttctctctgtctctgtctctctctctctgcct
27021 gtctctctcactgtgtctgtcttctgtcttactctctttctctgcctgcctctctgtctgtctgtctctc
27091 tccctccatgtctctctctctctctctctctctctcactcactctctctccgtctctctctctctttctg
27161 tctgtgtctctgtctctgtctgtctgtctctctctctctctctgtttgtctttctccctcactgtctgtc
27231 tgtctctctctctccctctctctgtctctgtctctctctctttctctttctgtctgtttctctctatctc
27301 tcgctgtccatctctgtctttctatgtctgtctctttctctgtcagtctgtcagacacccccgtgccggg
27371 tagggccctgccccttccacgaaagtgagaagcgcgtgcttcggtgcttagagaggccgagagaaatcta
27441 gacaggcgggccttgctgggcttccccacttggtgtatgatttcgggaggtcgaggctgggtccccgctt
27511 ggatgcgaggggcattttcagacttttctctcggtcacgtgtggcgtccgtacttctcctatttccctga
27581 taagctcctcgacttacacataaactgttaaggccggacgcaacacggcgaaaccccgtctctactaaaa
27651 atacaaagctgagtcgggagcggtggggcaggcccctgtaatgccagctcctcgggaggctgaggcggga
27721 gaatcgcttgaacctgggaggcggaggcttgcagggagccgagatcgcgccactgcactacggcccaggc
27791 tgtagagtgagtgagactcggtctctaaataaatcaggaaattcattaattcattaattcttttccctgc

27861 tgacggacatttgcaggcaggcatcggttgtcttctggcatcacctagcggccactgttattgaaagtcg
27931 acgtgacacggagggaggtctcgccgacttcaccgagcctggggcaacgggtttctctctctcccttctg
28001 gaggcccctccctctctccctcgttgcctagggaacctccgccctggcgggggccctattgttctttgat
28071 cggcgctttagttttctttgtgtttttggcgcctagactcttctacttgggctttgggaagggtcagttt
28141 aattttcaagttgcccccggcttcccccactacccacgtcccttcaccttaatttagtgagtcggtta
28211 ggtgggtttcccccaaccccccccccccccacccccccgcccccaacaccctgcttggaaaccttccaga
28281 gccaccccggtgtgcctccgccttctctcccctttccccacccctttgccggcgatctcattcttgccagg
28351 ctgacatttgcatcggtgggcgtcaggcctcactcaggggccaccgtttttgaagatgggggcggcacgg
28421 tcccacttccccagaggcagcttgggccgatggcatagcccttgacccgcgtgggcaagcgggcgggtct
28491 gcagttgtgaggcttttccccccgctgcttcccgcctcaggcctccctccctaggaaagcttcaccctgg
28561 ctgggtctcggtcacctttaatcgtgatgttttagtttctccgccctccggccagcagagtttcacaatg
28631 cgaagggcgccacggctctagtctgggccttcttagtacttgcccaaaatagaaacgctttctgaaaact
28701 aataactttgctcacttcagatttccagggacggtgccttggcccgtgtttgttggcttgttttgttttg
28771 ttttgttttgttttgttcgtgtttttccttttctcgtatgtctttctttttcaggtgaagtagaaatcccca
28841 gttttcaggaagacgtctatttcccccaagacacgtcagctgctgtttttttcctgttgttaactagcgct
28911 tttgtgaatctctcaacgtgcagtgagagccggttgatgtttactatatacttcatcatgacatcttatt
28981 ttctagaaatccgtaggcgaatgctgctgctcttgttgctgttgttgttgttgtcgttgctgttgtc
29051 gttgtcgttgttgttgccgttgttttcaaagtatacccggccaccgtttatgtgatcaaaagcattata
29121 aaatatgtgtgattatttcttgagcacgcccttcctcccccctctctctgtctctctgtctgtctctgtct
29191 ctctctttctctgtctgtcttctctctctctctctctctctgtgtctctctctctctgcctgtctgtttc
29261 tctctctctgtctctctctctctgcctgtctctctcactgtgtctgtcttctgtcttactccctttctct
29331 gcctgtctgcctgtctgtctgtcggtctctctctctctctctcccccctgtctgtatgtttctctctgtct
29401 ctgtctctctctctctttctgtgtctctctctccgtctctgtctttctctgtctgtctgtctctctcttt
29471 ctttctctctgtctctctctgcctgtctctgtcactctgtctgtcttctgtcttactctctctctgcctg
29541 cctgtctctctcactctctctctctgcgtgtctctctctctctttctttctctctgtctctctctgcctg
29611 tctgtctcactctgtctgtcttctgtcttactctctctctgcctgcctgtctctctcactctctctctgc
29681 gtgtctctctctctctctctttctgtctgtttctctctgtctctctctgtccgtgtctgtctttcttt
29751 gtctgtctctctttgtctttccttctctctgtctctgtctctctcattgtgtctgtcttctgtcttagtc
29821 tctctgtctctctcccctgtctgtctgtttgtctctctctctctctctgtctttgtctttctttctctctg
29891 tctctgtctctctctgtctctctctctgtgtgtctgtcttctctcttactgtctttctctgcctgtctgtct
29961 gtctgtctctctctgtctctctctctctctctctctctctcccccccccccgtcggctgtttctctgtct
30031 ctgtctgtgtctctctttctgtctgtttctctctgtctttctctctctgtctctttctctctgtctctct
30101 gtctgtctctgtctctctctctctctctctctctctctctctctctgtgtgtgtgtgggggggggtgtg
30171 gggggtgtgtgtgtctgccttctgtcttactctctttctctgcctgtctgtctgcctgtctgtttgtct
30241 ctctctctctctgcctgtctctctcccttcctgtctctgtttctctctctttctgtttctctctgtctct
30311 gtccatctgtgtctttctccgtctgtctctttatctgtctctctctctctctctttttctgtctttccct
30381 ctttgtgtatcgttgtctctctctgtctgtctctgtctctctgtctctgtctctctgtctctctctctct
30451 ctctctctctctctctctctctgtctgtctgtctgtctcggtctctggctctggctctcgctatctcccg
30521 ccctctctttttttgcaaaataagctcaagtacatctaatctaatccctcaccaaggcccgaattcttaa
30591 cttctgacatcccagatttgatctccctacagaatgctgtacagaactggcgagttgatttctggacttg
30661 gatacctcatagaaactacataggaataaagatccaatcctaaaatctggtgtggcttctccctcgactg
30731 tctcgaaaaatcgtacctctgttcccctaggatgccggaagagttttctccatgtgcatctgcccgtgtc
30801 ctaagtgatctgtgaccgagccctgtccgttctgtctcaaatatgtatgtgcaaacacttctctccattt
30871 ccacaactacccacggccccttgtggaaccactggctctttgaaaaaaatcccagaagtggttttggctt
30941 tttggctaggaggcctaagcctgctgagaactttcctgcccaggatcctgtgtgaacaaaagtgcctctg
31011 ctgggagctgggatcgtcgggaccatgcttgctagcgctggatgagtctctggaaggacgcacgggactc
31081 cgcaaagctgacctgtcccaccgaggtcaaatggatacctctgcattggcccgaggcctccgaagtacat
31151 caccgtcaccaaccgtcaccgtcagcatccttgtgagcctgcccaagaccccgcctccggggagactctt
31221 gggagcccggccttcatcggctaaagtccaaagggatggtgacttccacccacaaggtccccactgaacg
31291 gcgaagatgtggagcgtaggtcagagaggggaccgggaggggagacgtcctgaccggcgatgagttccct

31361 aggctctggccaccccacccatgccccacgtcccacgtcccgggcacccgcgggacaccgccgcttgatc
31431 ccctcctctgtccacagccggccccaccccaccacgcaacccacgcacacacgctggaggttccaaaagc
31501 acacggtgtgaatagagcctgacggagcgagagcccatttcacgaggtgggaggggtgggggtggggtgg
31571 gttgggggttgtggggtctgtggagagcccgattctccctcttgggtggctacaggccagaaatgaatat
31641 cgcttcttgggcggaggggcttccttaggccatcaccgcttgcgggactacctctcaaaccctcccttga
31711 ggccacaaaatagattccaccccacccatcgacgtttccccccgggtgctggatgtatcctgtcaagaga
31781 cctgagcctgacaccatcgaattaaacacctttactggcttggtgggtttgtttgtttctgagatggagt
31851 cttgctctgtcccccaggctggagtgcagcggcgtgatctcagctcactggaacctctgcctcctgggtt
31921 caagtgattctcctgtctcagcgccaccatggccggctcatatttttttttttttttttttttttttt
31991 ttagtagacacggggtttcaccctctttcattggttttcactggagattccagattcgaaccacacctca
32061 ttctgtgccacagagagacttcttcttcttcttcttcttttttttaagtgcaacgcaacatgtctgc
32131 cttatttgagtggcttcctatatcattataattgtgttatagatgaagaaaaggtattaaacactgtgct
32201 aatgatagcgaaagtgaagacaaaagaaaggctatccattttgtggttagaataaagttgctcagtattt
32271 agagctacctaaatacgtcagcatttacactcttcctagtcaaagctggccgatctgaataatcctcctt
32341 gaaacacaatttttgatagggttaagatttttttaagaatgcgactcctgcaaaagagctgaacagacaa
32411 tacacatttaaaaacatcacaacacaaggatcaaccagacttgggaaaaaatcgaaaacgacacaagtct
32481 tatgaagaactgagttcttaaaataggacggagaacatagccatcggaagagaaggcagtattggcaagt
32551 tgattgtgacgttggtcagcagtagctggcactatctttttggccatctttcgggccacgtaactactac
32621 agcaaaatgagatatgatccattgaacaacatattcgcaaatcaaaaaatgtttccgtaatataatgctt
32691 cagatttagaagcaaatcaaatgatagaacgccactgctgtaataagtcaccccaaagaccaccgtatct
32761 gacaaaataactaccacagggttatgacttcagaattatactttctcttgatatttacttatgtatgtat
32831 ttatttttttaatgtatttctcttgagacggcgtctcgctctgtcgcccaggctggagtgcaatggtgtg
32901 atctcggctcactgcaaccgccacctctccgggttcaagcgattctcctgcctcagcctcccgagtagct
32971 gggactacaggtgcccaccaccacgcccagctaatctttatacttttaatagagacggggtttcaccgtg
33041 ttggcccggatggtctcgatctcttgacctcgtgacccgcccgcctcggcctcccaaagtgctgggatga
33111 caggcgtgagccactgcgcccggccttctcttgacgtttaaactatgaagtcagtccagaaaaatgcaat
33181 caatgtcaacggtgagtatgctgttgaggcagagggaggaccacactttctcatgtcttattcagttgat
33251 cacaataggacctaggtagtaatttcctatgtgcccacttatacacgagtacaaaagagtaaaacagaga
33321 gaccgctacatgaaagggtacgtgaagttcttcatagtaactccgtaaactggaacactgtcaaaaagca
33391 gcagctcgtgagttgtttccatctatttttctattctccaataagtgaactatgctattcctttccaatc
33461 tcccaagcgctccttgtccccatcaccacttcggtgctcgaagaaaaagtaacaaatcaagtcacacaac
33531 taaagaaacacatacgcgaaccaaagacaactacagcgtctgcaaaagtttgctagaagactgaaactgt
33601 tgagtataaggatctggtattctaggatcatgagttcatttcagagtttgttcaagacatacgtttcata
33671 aggaaacatcttagttagaagttattcagcactatgtaccatccctaagtatttttaaccaaattcgtga
33741 caataaagagctatctaaccagaaaaactagcgagtaccggcaccatccatagggctttgtctttacact
33811 tcattaccacttaccatgccttacaatgtctgggattgaccctgatagcatttcgaaaacaagctgatgc
33881 tttgtccaattcttcagtgaagacaagctcacgccctaatgcactataggcataagcatcatttggatcc
33951 acttcgagagttctctggaagaattgaatcgcaatatcgtgttcccgttgcagaccgaaacagttccctg
34021 cagcacaccaggcctctggctggcgaagttttatccctgtctgtgaagtctttggacagaactgaaagag
34091 caacatctttcggaggatgccaaagtgttgtagagcagatctccatgccttcgactctgtaattctcaat
34161 ccttctaacctctgagaattttctttcagcttgcgtggactctgaaagtttaaaataggcccttccaatt
34231 tggcacagtacccaaccggtattgcggtggtgagaagctagatggctcaagatactgactgacagcttct
34301 ttgccgtggtgagaacacaaagctaaataaccttcccccttcacgaggaaggctcatcaagccttccg
34371 ctgctgctgctttttgtagattaaaagcctgaatctgaggcgcgattgcggctattttcccttctggaat
34441 gatggaagagtccaattttgtcacttccatggctatcacttatgttcggtggagttattcctcctttatt
34511 agctttacttttgtttttctgtttgggattttaggtggaaacttccctttaattttctcctattctcc
34581 tcggttgtggagctgtcactagtcaagagtcgtgaatttcttcgaggcggtgcatttgggggagatgcca
34651 tagtggggctcaatacctgaggtgttgcccttgtcggcgcaccagaactttgtgtttttgcaagaattgg
34721 ggttacctttcggctctttcccctctgcgagaagacagactgtgttccggtttggccgattctggcaaca
34791 gacttttttgaaggggctccggtggatggcacgtcaacgacagaaggtgtctaataccagtgcagttttg

34861 tccataggatccgtctccgggacttggggcttctaatggcaaaatgccaacgcttggggctcatggacta
34931 actgctgctggtcctcctaatacacttcgaccagtttttggtttatgttgaacctgtttagatcatatgg
35001 aagttcctgttcccagtgggacagtatcaggtgaaaggacagctgaatcgatagaagacactggggagtc
35071 tgtattcaaggagtactttgaattggaagattctcaattcaatccgtttaattcaacggtgtcctggggt
35141 gtttccgtaagaacggtctcaggctgtctgtgacataaactaggacgaggtcccagtgttttggcgcaac
35211 acttggacaggcagttgctaaagctctctagagatgtgaatcaaaatgtttggtcaggatctggcttttc
35281 ccccctgtttcacatcatgattcaaagggacaccggaggaaaggatttcaacgaaggctcttttggtcac
35351 attctgatcctttggtaagccgatctgccttgcaatatacatgtcccaacgatggacggggaaagcgagc
35421 tgaatcagcaaactcaggaacaataatatcatcatggcttttctgcttatgaaacactccacccgataag
35491 atttgatccccttctgcaagcttgctgagatcaacacaacatttcgcaagcaggtatttgcattgcggtg
35561 tagtacaactgtgtcctttccagagtctatatgttttataggcctttcctgagcggtaagaagaggttgc
35631 cagtaaaaacaaggcttcttctgagtgtacttctgcatagaggcgttctgcgaggaaaaccgcatctcgg
35701 taggcatagtggcttagtgcttgccatatagcagcctggacgggtccctgcagcaccgccatcctcgagg
35771 ctcaggcccactttctgcagtgcctcaggcaccgccccccccacccccccacccccccacccccatagcggc
35841 tccggcccggccagccccggctcatttaaagtcaccagcgaccgttaccgaccgtttaccgtgggatggg
35911 ggagtccgagccagaatgacttctttatcctgccgactctggaaagcccggcccccttgtgatccattgca
35981 aaccgagagtcacctcgtgtttagaacacggatccactcccaagttcagtgggggatgtgagggatgtg
36051 gcaggtaggacgaaggactctcttccttctgattcggtctgcacagtggggcctagggctggagctctct
36121 ctgtgcggaccgctgactccctctaccttgggttccatcggccccaccctggaacgcgggccttggcaga
36191 ttctggcccttcctggcccttcagtcgctgtcagaaaccccatctcgtgctcggatgccccgagtgactg
36261 tggctcgcacctctccggaaacattggaaatctctcctctacgcgcggccacctgaaaccacaggagctc
36331 gggacacacgtgctttcgggagagaatgctgagagtctcttgccgactctctcttgacttgagttcttcg
36401 tgggtgcgtggttaagacgtagtgagaccagatgtattaactcaggccgggtgctggtggctcgcgcctg
36471 taaccccagcactttgggaggccgaggccgtaggatccctcgaggaatcgcctaaccctggggaggttga
36541 ggctgcagtgagtgagccataattgtgtcactgcgctccagtctgggcgaaagacagaatgaggccctgc
36611 cacaggcaggcaggcaggcaggcaggcaggcaggcagaaagacaacagctgtattatgttcttgtc
36681 agggtaggaagcaaaaataacagaatacagcacttatttttttttcccttcggacggagtttcactct
36751 tggtgcccacgctggagtgcagtggcaccatctcggctcaccgcaacctccacctctcgcgttcaaacga
36821 ttctcctgcctcagcctcctgagtagctgggattacaggcaggggccaccacacccggctgattttgtat
36891 tgttagtagagacggcatttctccatgtgggtcaggctggtctcgaactggcgaccccagtggatctgcc
36961 cgcctcgccctcccaaagtgctggggtgacaggcgtgagccatcgtgaccggccggctacgttttttttt
37031 ttttaaattattttacttttttttagttttcaattttaatctatttatttatttatttacatttatttat
37101 ttatgtatgtacttatttatttatttttcgagacagactctcgctctgttgcccagactggagtgcagcgg
37171 cgcgatttcggctcactgcaagctctgcctcccgggttcacgccattctcctgcctcagcctcccaagta
37241 gccgggactacaggcgcccaccactgtgcctggctgacttttgtattttgagtagagatggggtttcac
37311 tgtggtagccaggatggtctcgatctcctgaccctgtgatccgtccacctcggcctcccaaagtgctggg
37381 atgacaggcgtgagccaccgccccggcctatttatctatttattaactttgagtccaggttatgaaacc
37451 agttagtttttgtaatttttttttttttttttgagacgaggtttcaccgtgttgccaaggcttgggcc
37521 gagggatccaccggccctcggcctcccaaaagtgcggggatgacaggcgcgagcctacggcgcccggacc
37591 ccccttcccccctccccgcttgtcttcccgacagacagtttcacggcagagcgtttggctggcgtgct
37661 taaattcattctaaatagaaatttggggcgtcagcttctggcctcatggactctgagctgaggagtcccc
37731 tggtctgtctatcacaggaccgtacacgtaaggaggagaaaaatcgtaacgttcaaagtcagtcattttg
37801 tgatacagaaatacacagattcacccaaaacacagaaaccagtcttttagaaatggccttagtcctgctg
37871 tccgtgccagtgattcttttcagtttggaccttgactgagagaattcccagtcggtctctcgtctctgga
37941 cggaagttccagatgatccgatgggtgggggacttaggctgcgtccccccaggagccccggtcgattagt
38011 tgtggggatcgctttggagggcgcggtgacccactgtgctgtgggagcctccatccttcccccccacccc
38081 tccccaggggatcccaattcattccgggctgacactctcactggcagacgtcgggcatcacctagcggtc
38151 actgtgactctgaaaacggaggcctcacagaggaagggagcaccaggccgcctgcgcacagcctggggca
38221 actgtgtcttctccaccgcccccacccccacctccaagttcctccctcctttgttgcctaggaaatcgcc
38291 actttgacgactgggtctgattgacctttgatcaggcaaaaacgaacaaacaaataaataaataaaataa

38361 cacaaaagtaactaactgaataaaataagtcaatacaatccattgcaatgcaataaaataccacacgata
38431 cgataggatacaatacaatacaatacaatacaatacaatacaatacaatacaatacaatacaataggccg
38501 ggtgcggtggctcatgcctgtcatcccatcactttgggaggccgaggtggacgcatcacctgaagtcggg
38571 agttggagacaagcccgaccaacatggagaaatcccgtctcaattgaaaatacaaaactagccgggcgcg
38641 gtggcacatgcctctaatcccagctgctaggaaggctgaggcaggagaatcgcttgaacctgggaagcgg
38711 aggttgcggtgagccgagattgcgccatcgcactccagtctgagcaaaaacagcgaaactccgtctcaaa
38781 aataaataaataaataaataaataaatacacaaatacataaattaaaataaataaatacaatgcaataaa
38851 ataaatgggccctgcgcggtggctcaagcctgtcacccctcactttgggaggccaaggccggtggatca
38921 agaggcggtcagaccaacagggccagtatggtgaaaccccgtctctactcacaatacacagaattagccg
38991 ggcgccgtgctgtgctgtactgtctgtaatcccagctactcgggaggccgagctgaggcaggagaatcgc
39061 ttgaacctgggaggtggaggttgcagtgagccgagatcgtgccactgcaccccagcctgggcgacagagc
39131 gagactccgtctcaaaaaaatgaaaatgaaaatgaaatgcaacaaaataattaaaaagtgagtttccggg
39201 gaaaaagaaaaaacaacaacaacaacaacaacaacaacaacaacaacaaaaaacaaccccaccgtgacatcca
39271 cgtacgtacgcctctcgcctttcgaggcatcaaacacgttaggaattatgcgtgatttctttttttttaac
39341 ttacttttatttttattctcatgatttctgtttcgagacggagtctcggaggcccgccctccctccgtccc
39411 attccctggttgcccagacaacctcaggagacagaccctggctgggccagattgttcttctcctcggtcg
39481 atggtttccttgtctttcttcgtgtctttaacccgcgtggactcttccgctcgggtttgacagatggcag
39551 ctccactttaggccttgttgttgttggggactttcccgattctccccagatgtagtgaaagcaggtagat
39621 ttgccttgcctggccttgcctggccttgcctggccttgccttttctttctttctttctttctttctttct
39691 ttctttctttctttcttattactttctttttttcttcttcttcttcctttttttttttttttttttttt
39761 tttttttgagacagagtttcactcttgttgcccaggctagagggcaatggcgcgatctcggctcaccgc
39831 accctccgcctcccaggttcaagcgattctcctgcctcggcctcctgattagctgggattacaggcatgg
39901 gccaccgtgcctggctgatgtttgtacttttagtagagacggtgtttttccatgttggtcaggctggtct
39971 ccaactcccaacctcaggtggtccgcctgccttggcctcccaaagtgctgggatgacaggggtgagccac
40041 agcgcccagcctctctctctctctgtctctctctctctctctctctctctctctctctctctctctcg
40111 cttgcttgcttgcttgctttcgtgctttcttgcttccttgttttctttctttctttgtttctttcttgc
40181 ttgctttcttgcttgcttgttttcttgctttcttgctttctttttctttcttgcttgcttgctttctggct
40251 ttcctttctttttctttcttgctttcttgcttgcttgctttcgtgctgtcttgttttcttgctttcttgc
40321 ttgcttgcttgcttgcttgctttcagcgcccagcctctctctctctctctctcgcttgcttgcttgct
40391 tgcttgcttgctttcgtgctttcttgctttcttgttttctttctttctttgtttctttcttgcttgctt
40461 tcttgcttgcttgttttcttgctttcttgctttctttttctttcttgcttgcttgctttcttgctttcctt
40531 tctttttctttcttgctttcttgcttgcttgctttcgtgctgtcttgttttcttgcttgcttgcttgctt
40601 gcttgctttcgtgctttcttgctttcttgttttcttgctttctttctttctttgtttctttcttgcttt
40671 cttttctttcttgcttgctttcttgctttctttttctttttctttcttgctttcttgctttcttgctttct
40741 ggttttcttgctttcttgcttgctttcgtgctttcttgttttcttgctttctgtcttttgcttctttctt
40811 tcttgctctctcgcttgcttgcttgcttgctttcgtgctttcttgttttcttgctttctttctttctttt
40881 gtttctttcttgcttgctttcttgctttcttgttttcttgctttcttgctttctttcctgctttctttct
40951 tgctttgttttctttctagctttctttttctttcttgcttgctttcttgctttctttttctttttctttctt
41021 tctttcttgctttctttttttcttgctttcttgctttctttttctctttctttctttctttcctgctttttt
41091 gctttcttgcttgctttcttgcttgctttcttgcttgcttgctttctgtctttttctttctttctttctt
41161 ctttctttctttctttctttctttctgtttcgtccttttgagacagagtttcactcttgtttccacggct
41231 agagtgcaatggcgcgatcttggctcaccgcaccttccgcctcccgggttcgagggcttctcctgcctca
41301 gcctcctgattagcggggattacagggaggcacccccacgcctggcttggctgatgtttgcgtttttagt
41371 aggcacaccgtgtctctccgtgttggtcaggctggcctccgactcccgacctcatgtgatgcgcccacct
41441 cggcctctcgaagtgctgggatgacgggcgtgagccaccgtgcccggcctgttgactcatttcgtttttt
41511 ttatttctttcctttccatgtatttactgacgtttattaatgtgaacatttttgtacatttgtatgtaaa
41581 cgacaatgcgtatatctttgcatttactactctcgtgtgtagtaaacacgtaccgattgtatagaaacat
41651 ccttctatatgatagatgtaggtgtttctgtgatacaaataaatacacatcgctctacagagaagcgatc
41721 gtcgagaaatacgtttacgtttacgtacgaaaagtgttgtatttatgtttataaatgaacaagtgtacgt
41791 acttatctctgttttctttcttcctctcctttacgtttttcttccttcctttcttcctttctctccttct

41861 gtaggttttttcttcctctcttcctttccttctctctctctttctgatcttttctccttcgtgctttattt
41931 ctctttcgttccctgtctctccttctgttttctttcctctctgtttctttttcccttctttccttcgttt
42001 ctttcctcattctttctctctttttcatgtttctttccttttccgtctgtctttttaaaaatggagtgtttc
42071 aaaagttttctctgtgtatcgacgtttttttaaattgtctctctttttctccattgtcttcctccctccctc
42141 cctccctgctcccttccctccctccttcccttttcaccatctgtctctttttccccattccccgcccccccc
42211 ccccccccccgtctgtctctgcgtggattccggaagagcctatgcattttgcctctccgtgtgtctgca
42281 gcgacccgcgaccgagtccttgtgtgttctttctccctccctccctccctccctccctgcttccga
42351 gatgcatctccgaacacccacgcgccgtgggttgtcttctgactctgtcgcggtcgaggcagagacgcgt
42421 tttgggcaccgtgtgtgtggggttggggcagaggggctgcgttttcggcctcgggaagagcttctcggct
42491 cacggtttcgcttttgcggtccacgggccgccctgcctgccatccggatctgtctcgctgacgttcgcgg
42561 cggtcgtcgggctccatctggcggccgcttttagatcgtgctctcggcttccggagctgcggtggcagct
42631 gccgagggaggggaccgtccccgctgtgagctgtgcagagctccggaaagcccgcggtcgtcagcccggc
42701 tggcccggtggcgccagagctgtggcgcgtcgcttgtgagtcagagctctggcgtgcaggtttatgtggg
42771 ggagaggctgtcgctgcgcttctgggcccgagccgggcgtggggctgcccgggccggtcgaccagcgcgc
42841 cgcagctcccgaggcccgagccgcgacccgcggggacccgccgcgcgtggcgcgggaggctggggacgcc
42911 cttcccggcccggtcgcgggtccgtccgcgctcatcctggccgtctgaggcggcggccgaattcgtttcc
42981 gagtccccgtggggagccggggaccgtcctgcccccgtcccccgggtgccggggagcggtccccgggccg
43051 ggccgcggtccctctgccgcgatcctttctggcgagtccccgtgcggagtcggagagcgctccctgagcg
43121 cgcgtgcggcccgagaggtcgcgcctggccggccttcggtccctcgtgtgtcccggtcgtaggaggggcc
43191 ggccgaaaatgcttccggctcccgctctggggacacgggccggccccctgcgtgtggcacgggcggccgg
43261 gagggcgtccctggcccggcgctgctcccgcgtgtgtcccggggttgaccagagggccccgggcgctccg
43331 tgtgtggctgcgatggtggcgtttttggggacaggtgtccgtgtcgcgcgtttcctgggccggcggcgtg
43401 gtcggtgacacgacctcccggccccgggggaggtatatctttcactccgagtctgcatcttgggccaccg
43471 ggttatt