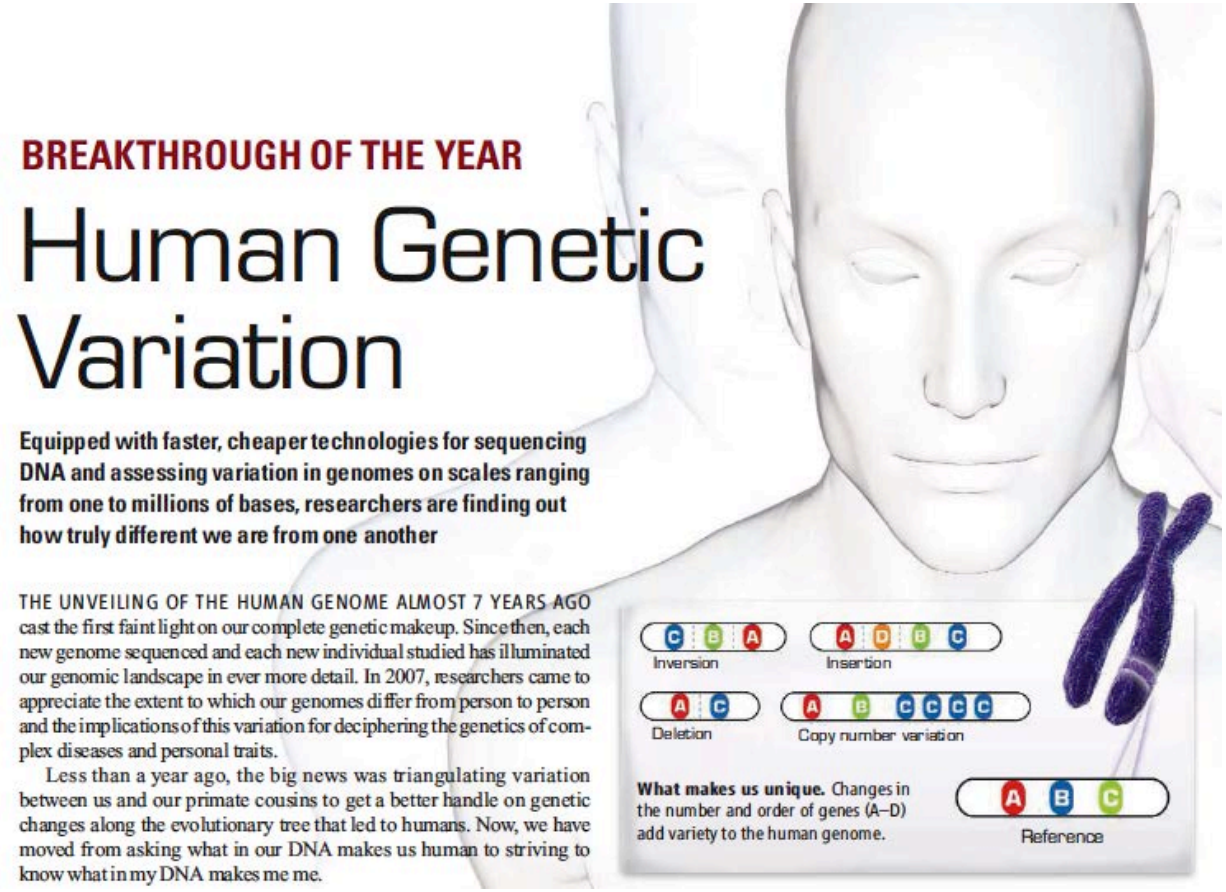


BREAKTHROUGH OF THE YEAR

Human Genetic Variation

Equipped with faster, cheaper technologies for sequencing DNA and assessing variation in genomes on scales ranging from one to millions of bases, researchers are finding out how truly different we are from one another

THE UNVEILING OF THE HUMAN GENOME ALMOST 7 YEARS AGO cast the first faint light on our complete genetic makeup. Since then, each new genome sequenced and each new individual studied has illuminated our genomic landscape in ever more detail. In 2007, researchers came to appreciate the extent to which our genomes differ from person to person and the implications of this variation for deciphering the genetics of complex diseases and personal traits.

Less than a year ago, the big news was triangulating variation between us and our primate cousins to get a better handle on genetic changes along the evolutionary tree that led to humans. Now, we have moved from asking what in our DNA makes us human to striving to know what in my DNA makes me me.

This article details some of the common sequence variations that have been found when comparing individual human genome sequences including single nucleotide polymorphisms (SNPs), gene copy number variations (CNVs) and other structural rearrangements

Breakthrough of the Year: Human Genetic Variation Science 318: 1842 Dec. 21, 2007
http://fire.biol.wwu.edu/trent/trent/humanvariation.pdf

Digital Memories, Piling Up, May Prove Fleeting

By KATIE HAFNER (NYT) 1712 words
Published: November 10, 2004

The nation's 115 million home computers are brimming over with personal treasures -- millions of photographs, music of every genre, college papers, the great American novel and, of course, mountains of e-mail messages.

Yet no one has figured out how to preserve these electronic materials for the next decade, much less for the ages. Like junk e-mail, the problem of digital archiving, which seems straightforward, confounds even the experts.

"To save a digital file for, let's say, a hundred years is going to take a lot of work," said Peter Hite, president of Media Management Services, a consulting firm in Houston. "Whereas to take a traditional photograph and just put it in a shoe box doesn't take any work." Already, half of all photographs are taken by digital cameras, with most of the shots never leaving a personal computer's hard drive.

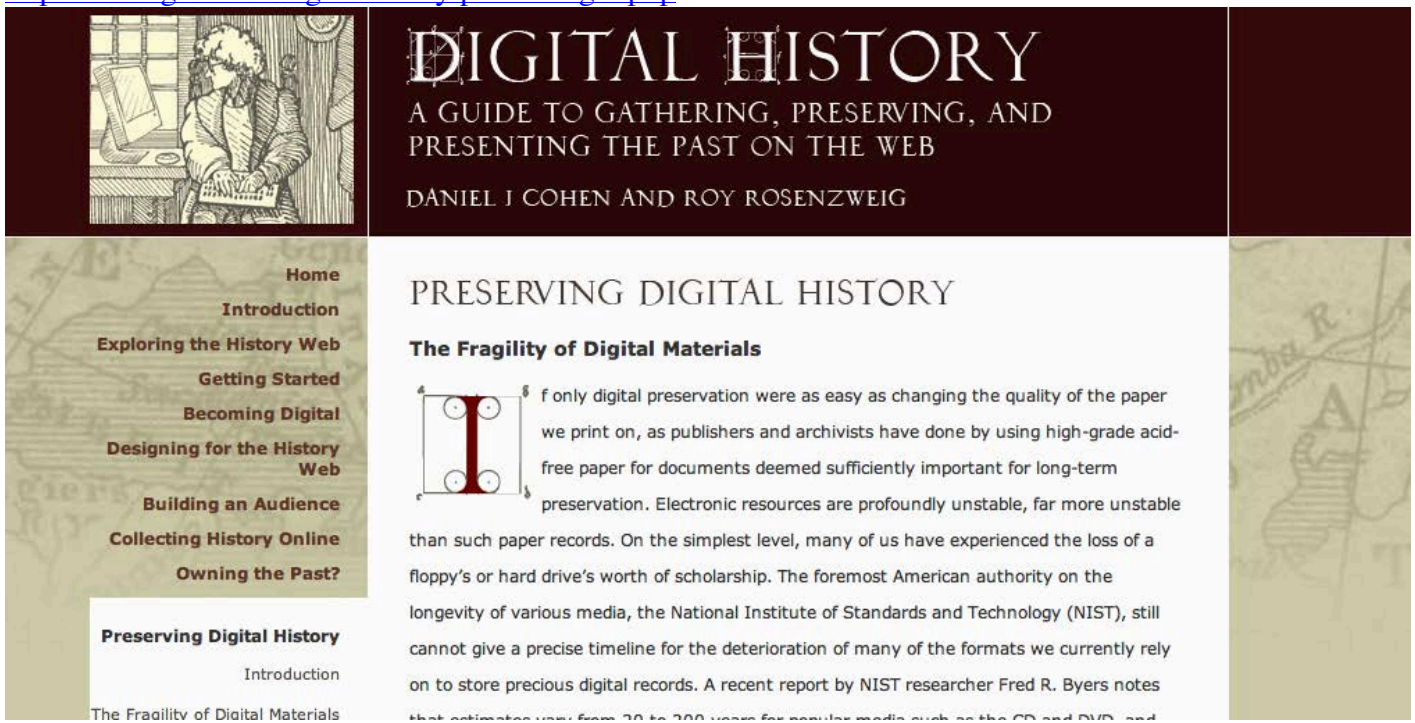
So dire and complex is the challenge of digital preservation in general that the Library of Congress has spent the last several years forming committees and issuing reports on the state of the nation's preparedness for digital preservation.

Jim Gallagher, director for information technology services at the Library of Congress, said the library, faced with "a deluge of digital information," had embarked on a multiyear, multimillion-dollar project, with an eye toward creating uniform standards for preserving digital material so that it can be read in the future regardless of the hardware or software being used. The assumption is that machines and software formats in use now will become obsolete sooner rather than later.

"It is a global problem for the biggest governments and the biggest corporations all the way down to individuals," said Ken Thibodeau, director for the electronic records archives program at the National Archives and Records Administration.

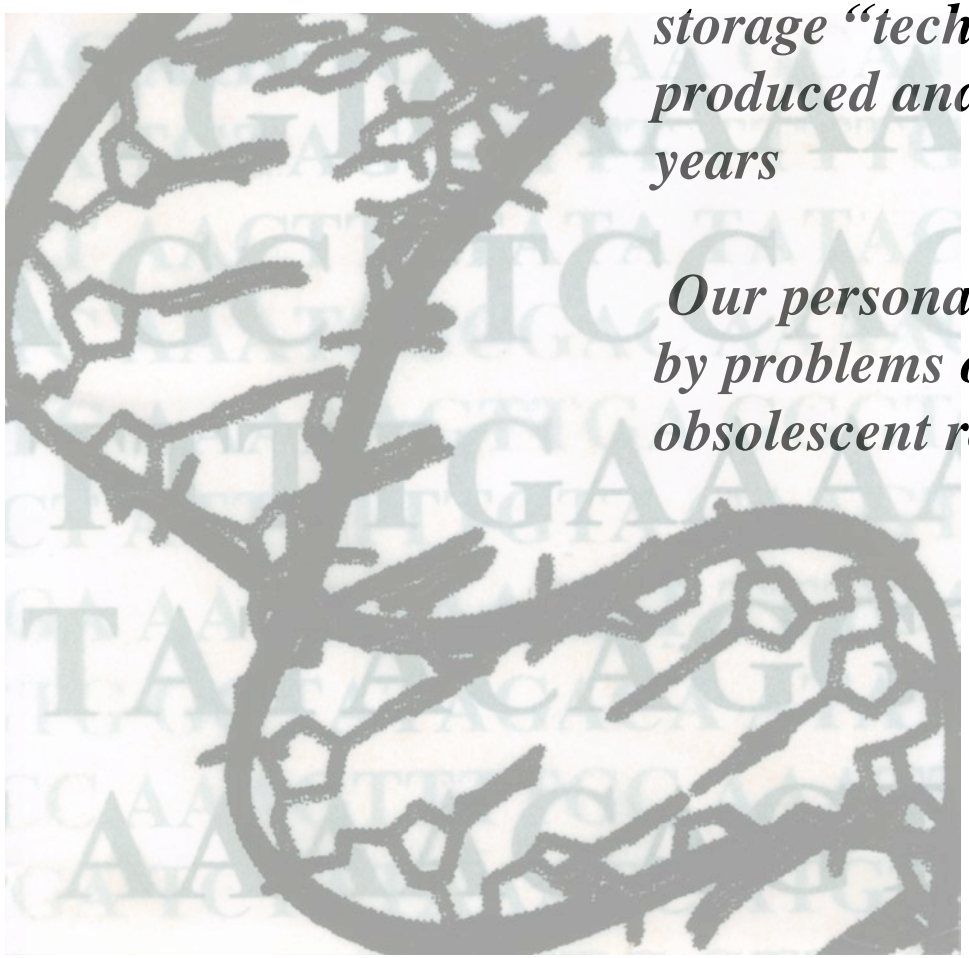
In the meantime, individual PC owners struggle in private. Desk drawers and den closets are filled with obsolete computers, stacks of Zip disks and 3 1/2-inch diskettes, even the larger 5 1/4-inch floppy disks from the 1980's. Short of a clear solution, experts recommend that people copy their materials, which were once on vinyl, film and paper, to CD's and other backup formats.

http://chnm.gmu.edu/digitalhistory/preserving/1.php



Data currently being stored in magnetic or optical media will probably become unrecoverable within a century or less (estimates vary). This will be due to the combined effects of

1. ***software obsolescence***
2. ***obsolescence of hardware for retrieval***
3. ***decay of the storage medium (aka material deterioration).***



This human-generated digital technology contrasts with the impressive information storage "technology" that evolution has produced and refined over the past 3+ billion years

Our personal digital archive is NOT plagued by problems of chemical fragility or by obsolescent retrieval systems

New approaches are required that will permit retrieval of information stored for centuries or even millennia

DNA has three properties that recommend it as a vehicle for long-term information storage:

- **First,** *DNA has stood the informational "test of time"* during the billions of years since life emerged. Non-replicating DNA molecules are quite robust.
  ➜ *good chemical stability*
- **Second,** because DNA is our genetic material, *methods for both storage and reading of DNA-encoded information is central to technological civilizations* and will undergo continual improvements.
  ➜ *DNA-R-US*
- **Third,** use of DNA as a storage medium *permits each segment of information to be stored in an enormous number of identical molecules. This extensive informational redundancy would strongly mitigate effects of any losses due to stochastic decay.*
  ➜ *easy to make copies via PCR*

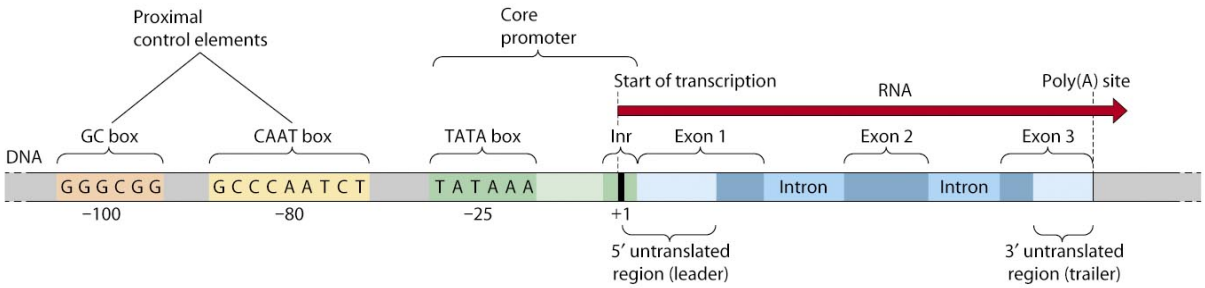
Bottom Line: Data retrieval of information stored in DNA should ideally require minimal prior knowledge beyond a familiarity with molecular biological techniques (DNA sequencing and PCR --polymerase chain reaction)

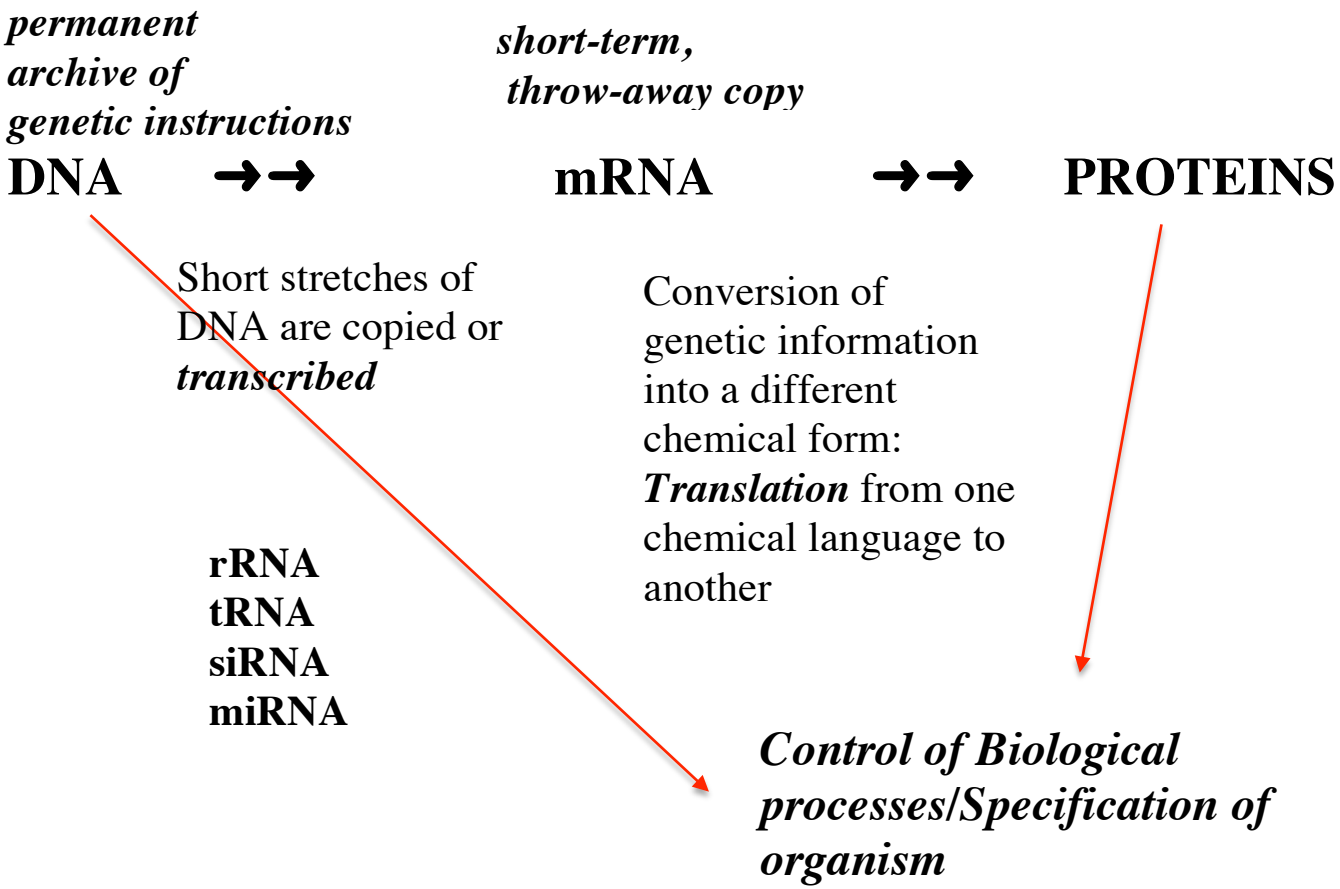
The changing definition of the gene:

Mendelian: the fundamental functional unit of heredity that carries information from one generation to the next

Biochemical: a unit of heredity that specifies the production of a polypetide

Molecular: a segment of DNA composed of a transcribed region and adjacent regulatory regions that control transcription




permanent archive of genetic instructions
DNA
short-term, throw-away copy
mRNA
PROTEINS
Short stretches of DNA are copied or transcribed
Conversion of genetic information into a different chemical form: Translation from one chemical language to another
rRNA
tRNA
siRNA
miRNA
Control of Biological processes/Specification of organism

Our focus with respect to molecular genetics will be on:

- *how mutations come about*
- *how mutations affect gene function*
- *techniques to directly assess genotype at the DNA level*

***Sex, Errors and the Genome* by Mark Ridley**
Natural History (6/2001)

"At **conception**, human embryos average about 200* copying errors and about 50% of the embryos have a botched number of chromosomes. "

WHO IS TO BLAME?

**recent estimates are lower – see below*

When a thirty year old man breeds with a 30 year old woman:

- his DNA (in his sperm cells) has been copied *430 times* against her *33 cell division* (in egg cells).
- with thirteen times as many errata in his DNA, about *185 of the 200** copying mistakes in each human conception may come from the sperm.
- however, a woman's eggs are more likely to carry serious errors in chromosome numbers, and these errors increase with maternal age.

**recent estimates are lower*

Variation in genome-wide mutation rates within and between human families
or
Reality is more Complex

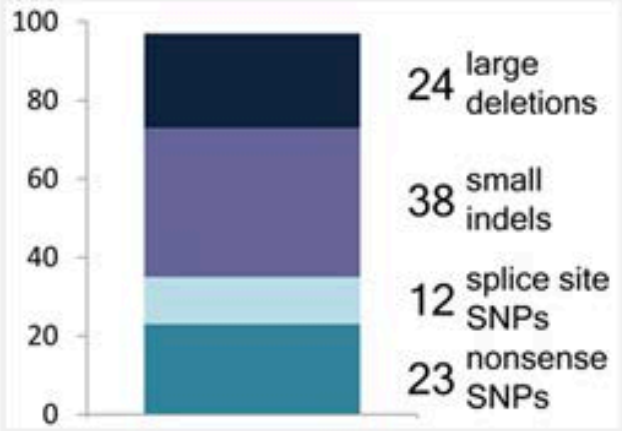
J.B.S. Haldane proposed in 1947 that the male germline may be more mutagenic than the female germline . Diverse studies have supported Haldane's contention of a higher average mutation rate in the male germline in a variety of mammals, including humans. ***Here we present, to our knowledge, the first direct comparative analysis of male and female germline mutation rates from the complete genome sequences of two parent-offspring trios***. Through extensive validation, we identified ***49 and 35 germline de novo*** **mutations (DNMs)** in two trio offspring, as well as 1,586 non-germline DNMs arising either somatically or in the cell lines from which the DNA was derived. Most strikingly, in one family, we observed ***that 92% of germline DNMs were from the paternal germline***, whereas, in contrast, in the other family, ***64% of DNMs were from the maternal germline.*** These observations suggest considerable variation in mutation rates within and between families.

VOLUME 43 | NUMBER 7 | JULY 2011 **Nature Genetics**

All genomes are dysfunctional: broken genes in healthy individuals

16/02/2012
Categories: Journal Club
Written by Daniel MacArthur

I don't normally blog here about my own research, but I'm making an exception for this paper. There are a few reasons to single this paper out: firstly, it's in *Science* (!); and secondly, no fewer than five Genomes Unzipped members (me, Luke, Joe, Don and Jeff) are co-authors. For me it also represents the culmination of a fantastic postdoc position at the Wellcome Trust Sanger Institute (for those who haven't heard on Twitter, I'll be starting up a new research group at Massachusetts General Hospital in Boston next month).



Breakdown of the number of loss-of-function variants in a "typical" genome

Readers who don't have a *Science* subscription can access a pre-formatted version of the manuscript here. In this post I wanted to give a brief overview of the study and then highlight what I see as some of the interesting messages that emerged from it.

http://www.genomesunzipped.org/2012/02/all-genomes-are-dysfunctional-broken-genes-in-healthy-individuals.php

Mutation Jargon

GENE MUTATION = POINT MUTATION

(scale of mutation is small and is localized to a specific region,
a single nucleotide or a few adjacent base pairs)

↓

at the DNA level:

- single base pair substitutions: *transitions & transversions*
- single (or a few) base pair addition or deletion: *indels*
- gene mutation by transposon insertion

at the level of gene expression:

promoter mutations
splicing mutations
regulatory mutations

at the protein level:

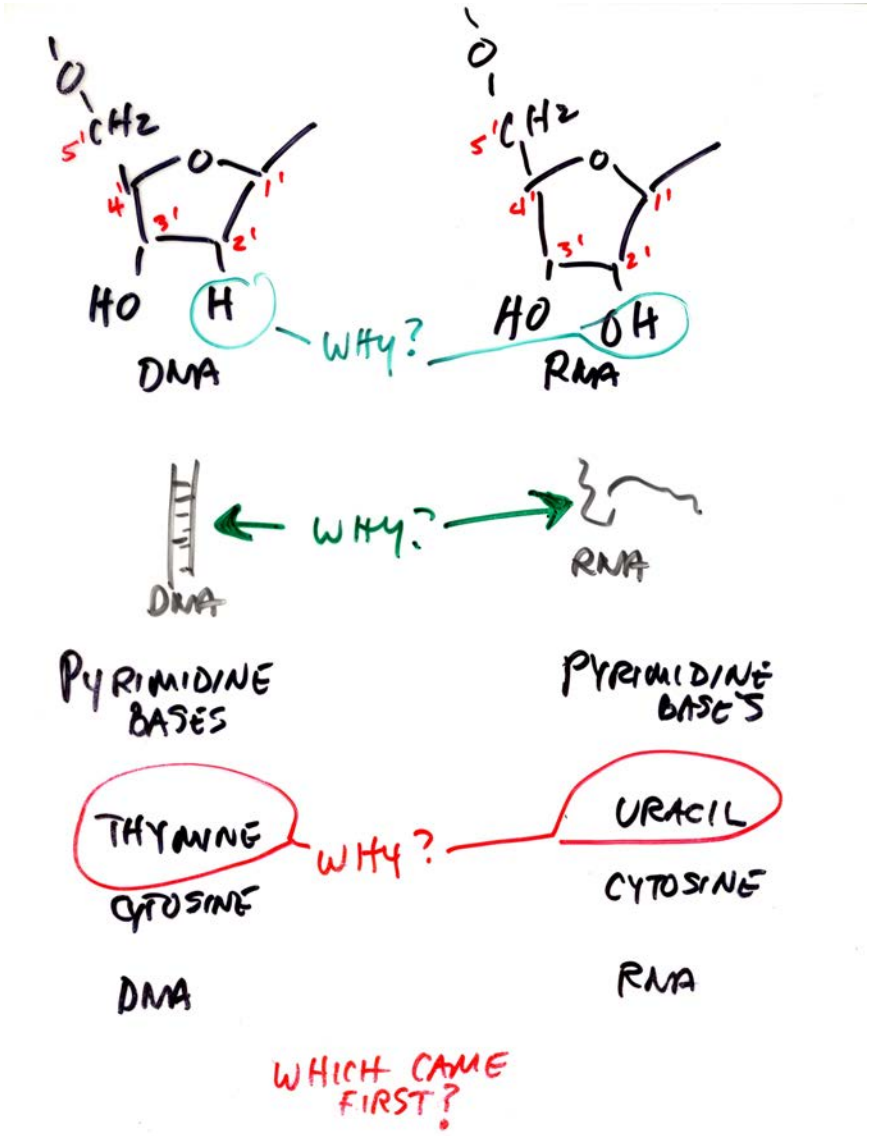
nonsense
missense
[neutral]
silent
frameshift

at the level of gene function:

loss-of-function
gain-of-function
[neutral]

CHROMOSOME MUTATION

- involves *segments of chromosomes* or *whole chromosomes* or *whole genomes*
- alterations in chromosome structure and number
- deletion, duplications, translocations and inversions
- CNVs: copy number variations



Review DNA structure and DNA replication

- general overview
- biochemistry of chain elongation
- features of DNA polymerases

5′ TTACCCATTCAGCCCATTCCCTGCAAACCAGTGGAGTATCCGCTGCAGCTGCTGCACAGCCCCCCTGCCCCAGTGGTGAAGAGGCC
TGGGGCCATGGCCACCCACCACCCCCTGCAGGAGCCCTCCCAGCCCCTGAACCTCACAGCCAAGCCCAAGGCCCCCGAGCTGCCCAACA
CCTCCAGCTCCCCAAGCCTGAAGATGAGCAGCTGTGTGCCCCGCCCCCCCAGCCATGGAGGCCCCACGCGGGACCTGCAGTCCAGCCCC
CCGAGCCTGCCTCTGGGCTTCCTTGGTGAAGGGGACGCTGTCACCAAAGCCATCCAGGATGCTCGGCAGCTGC……..
…….. etc, etc, etc, 3′

The coding information contained in DNA is

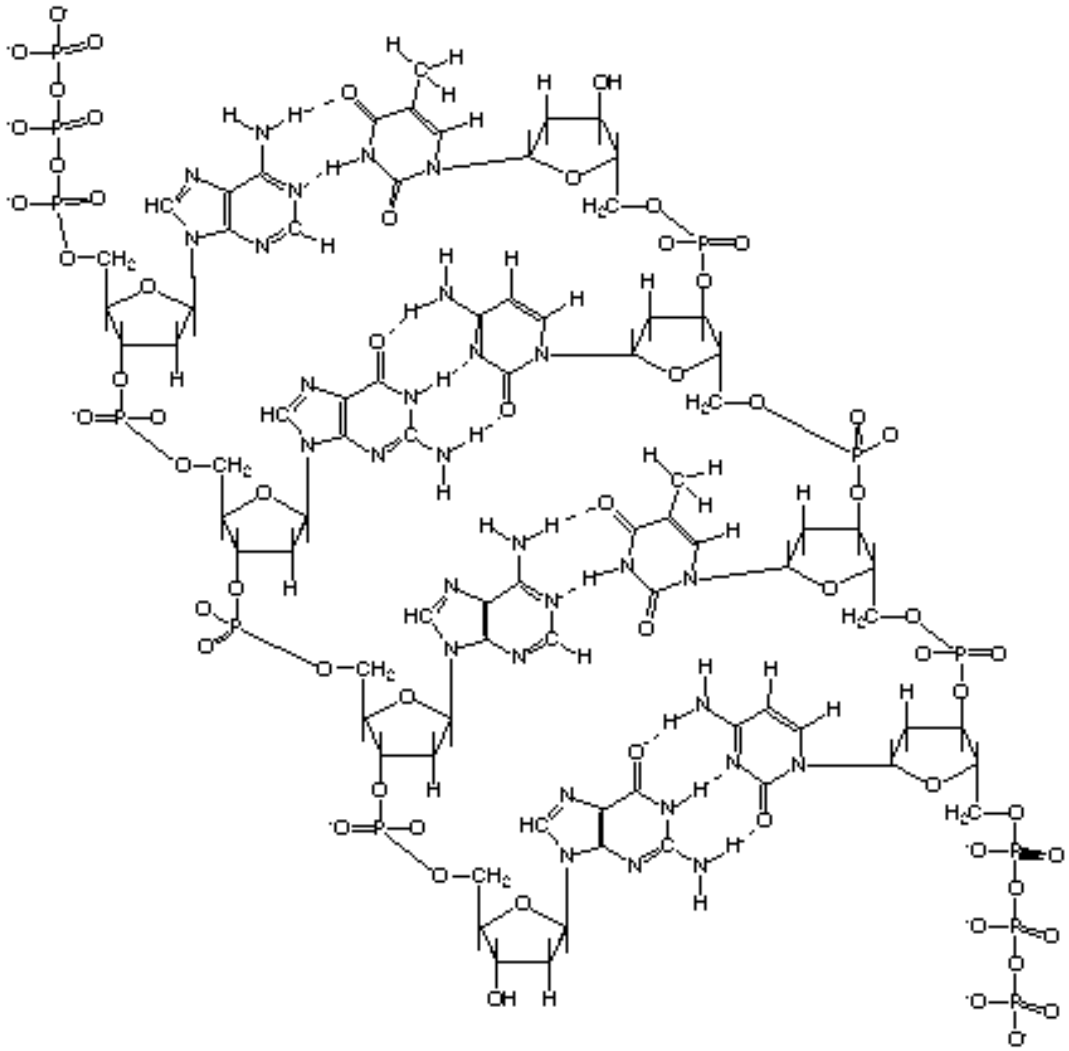
- **one dimensional:** encoded along the length of a molecule
- **digital:** because the basic information unit (the nucleotide or base) can exist in only one of four* discrete states abbreviated: G, A, T, C

Review DNA structure here:
http://users.rcn.com/jkimball.ma.ultranet/BiologyPages/D/DoubleHelix.html

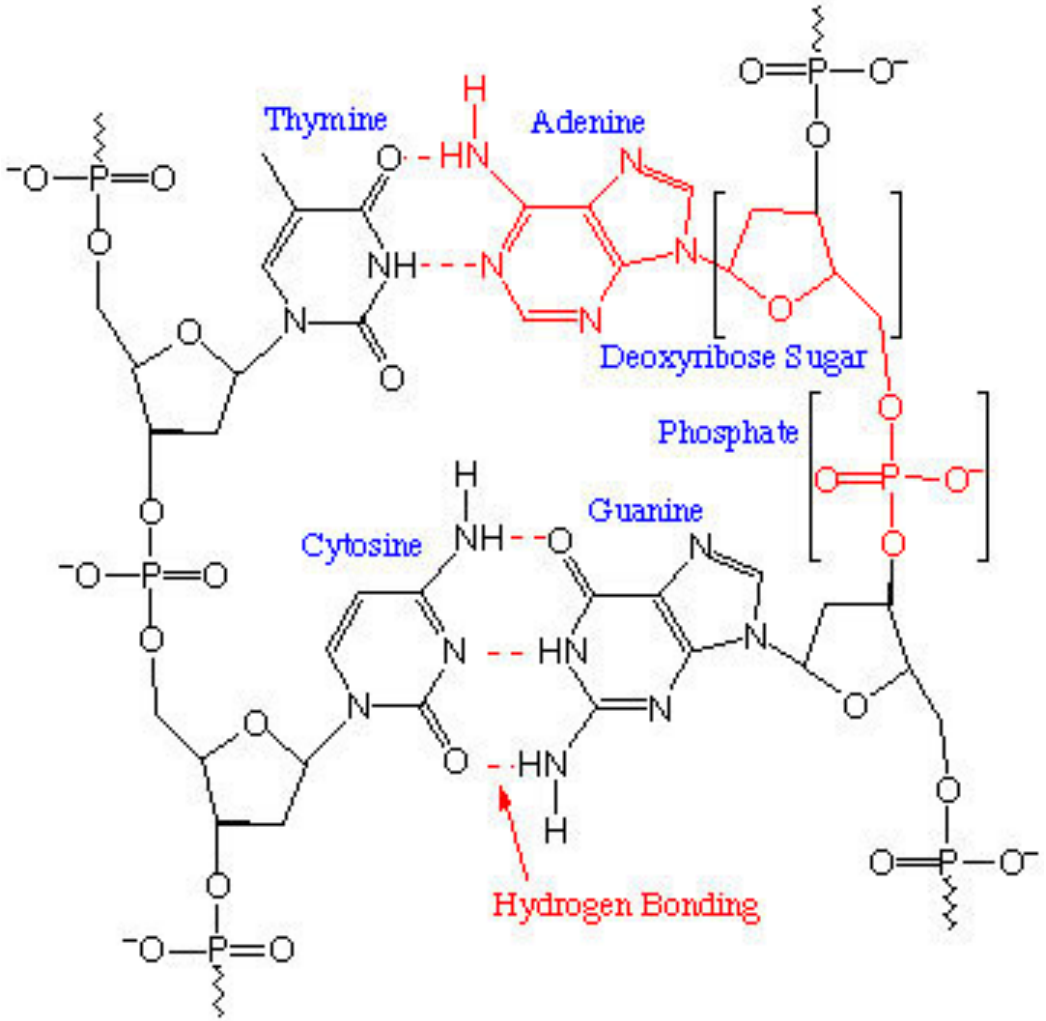
Cool stuff here
http://www.johnkyrk.com/DNAanatomy.html

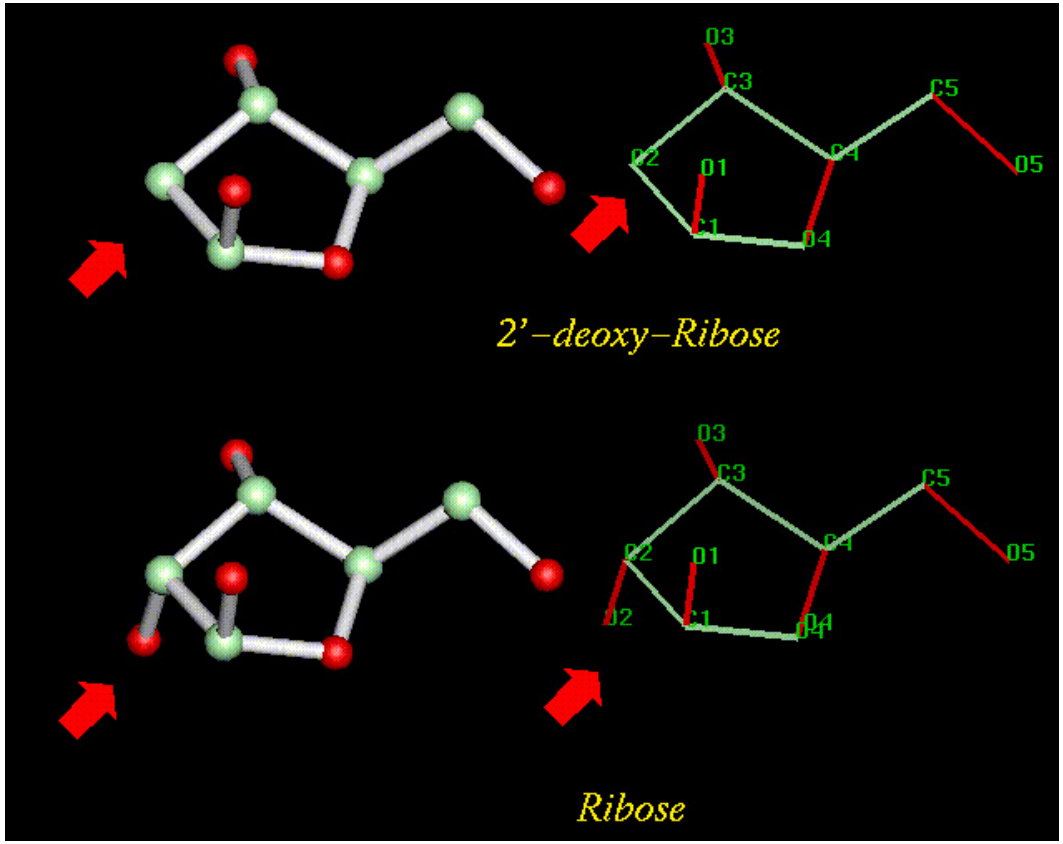
http://www.geneticengineering.org/chemis/Chemis-NucleicAcid/DNA.htm

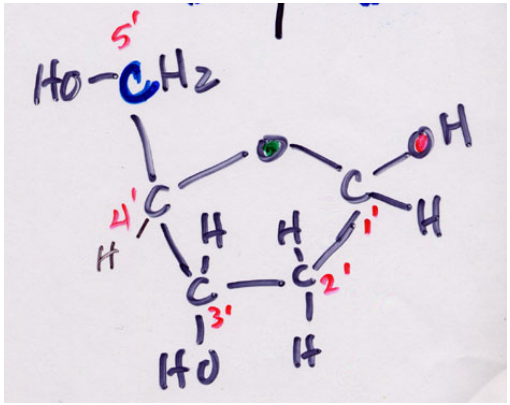
***DUH -- somebody has realized that more info could stored on a CD if each point (position) could be represented by more than two (0 or 1) possibilities .**



2-D look at DNA: note strands are *antiparallel*



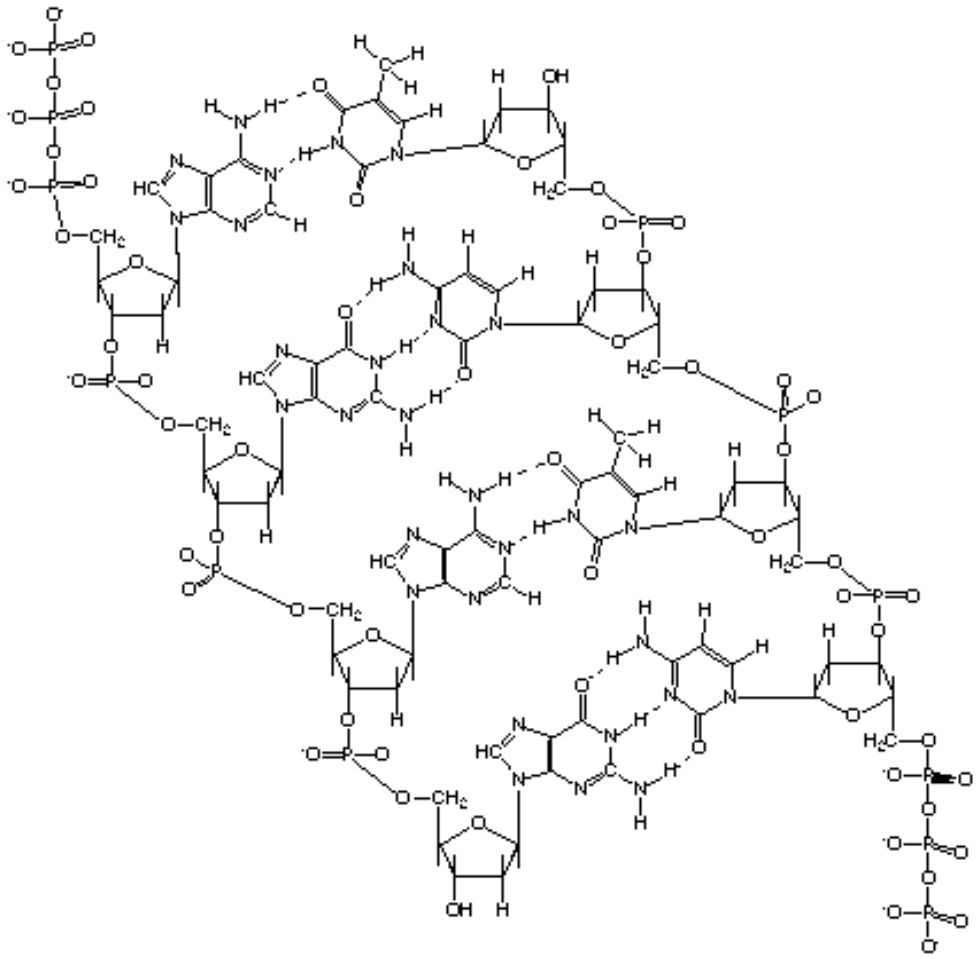


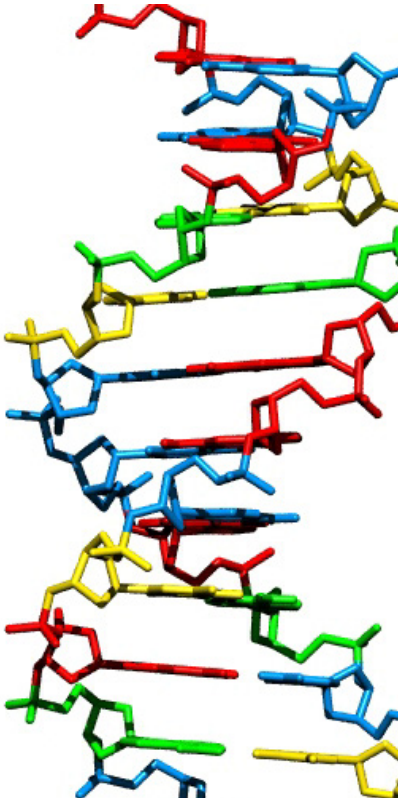


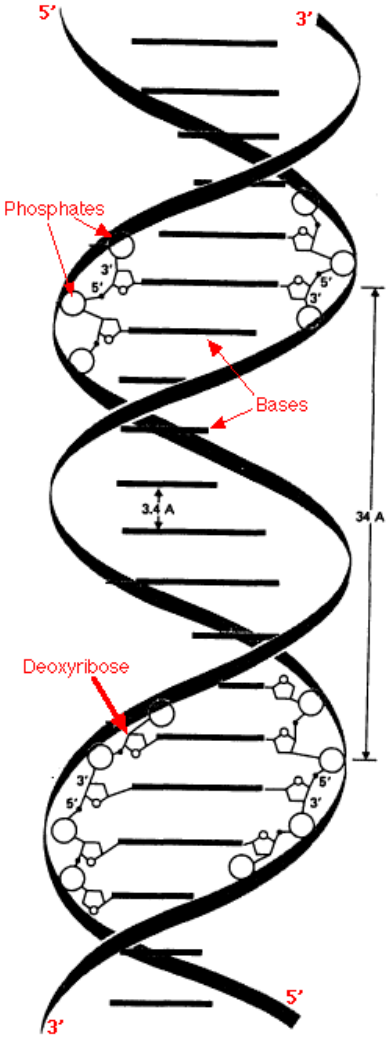
WHY 2' deoxy in DNA?

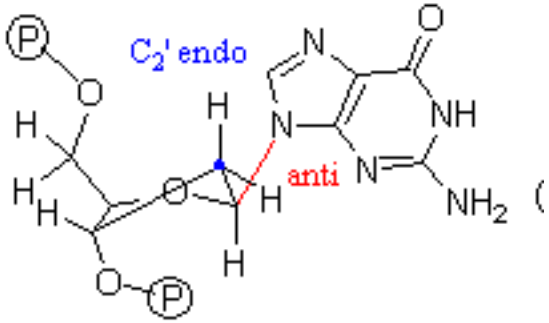
THE ENDS OF A DNA POLYMER CAN BE DISTINGUISHED BASED ON WHETHER THERE IS A FREE 5' OR 3' HYDROXYL

- **3' and 5' ends**
- **antiparallel**





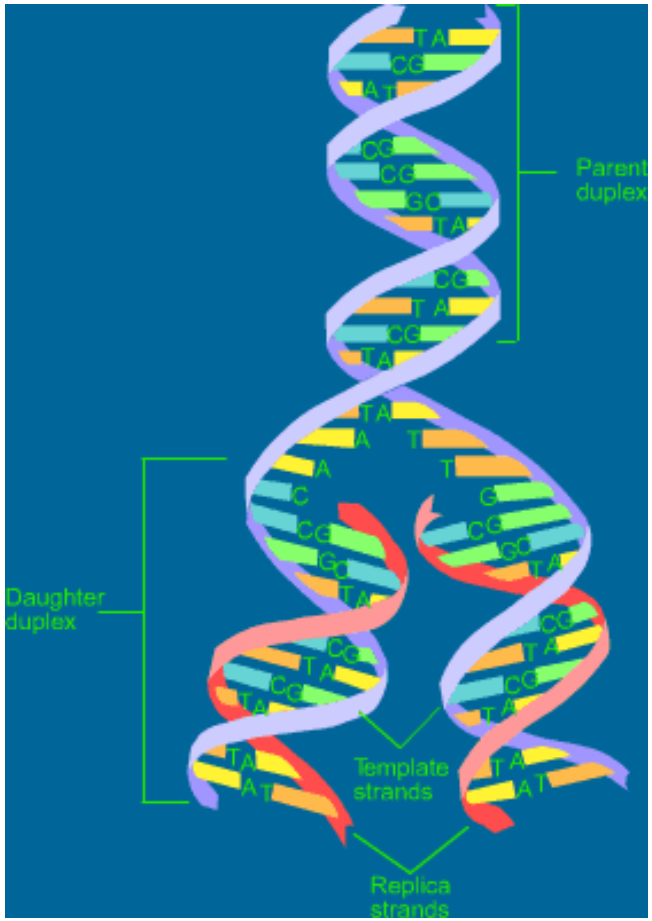

5′
3′
Phosphates
3′
5′
5′
3′
Bases
3.4 A
34 A
Deoxyribose
3′
5′
5′
3′
5′
3′


P
O
H
H
H
O
O
P
C2′ endo
H
N
N
O
NH
N
anti
NH2
H
H

What is Crick holding in his hand?

http://biocrs.biomed.brown.edu/Books/Chapters/Ch%208/DH-Paper.html

http://www.nature.com/genomics/human/watson-crick/



Semi-conservative DNA replication:

1. The parental strands of the DNA double helix separate
2. Each parental strand serves as template for the synthesis of a complementary copy
3. The nucleotide sequence of the newly synthesized daughter strand is determined by the
   - sequence of the parental template
   - pairing (hydrogen-bonding) specificities of the purine and pyrimidine bases

Nice animation:

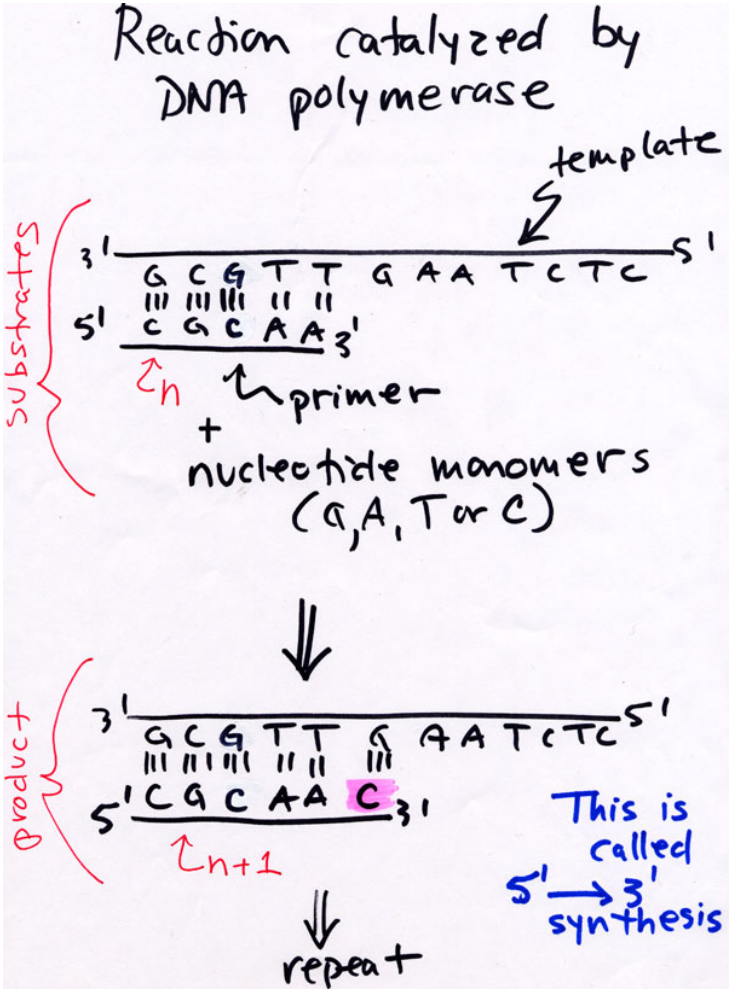
http://www.johnkyrk.com/DNAreplication.html

What do you know about the enzymology of this process?

What is the name of the enzyme that catalyzes the elongation of the nucleotide chain?

DNA polymerase: **synthesizes new strands of DNA by catalyzing the chain elongation reaction: one nucleotide is added to the existing DNA chain with the release of inorganic phosphate**

- **Synthesis of the polymer is always 5' ---> 3'**
- **A template is required**
- **A primer is required**



http://www.geneticengineering.org/chemis/Chemis-NucleicAcid/DNA.htm

Nice animation:

http://www.johnkyrk.com/DNAreplication.html

Figure shows features common to all DNA polymerases:

- it takes "instructions" from a template -- the parental strand of DNA
- it can only catalyze the addition of a nucleotide monomer to a 3' carbon of ribose
  - it cannot catalyze the addition of a nucleotide monomer to the 5' carbon of ribose
  - this is called 5' to 3' synthesis

DNA polymerase cannot lay down the first nucleotide of a DNA strand

- It requires a "bit" of polymer to add onto
- This short segment of polymer is called a *primer*
- It provides a 3' hydroxyl for the DNA polymerase to add onto

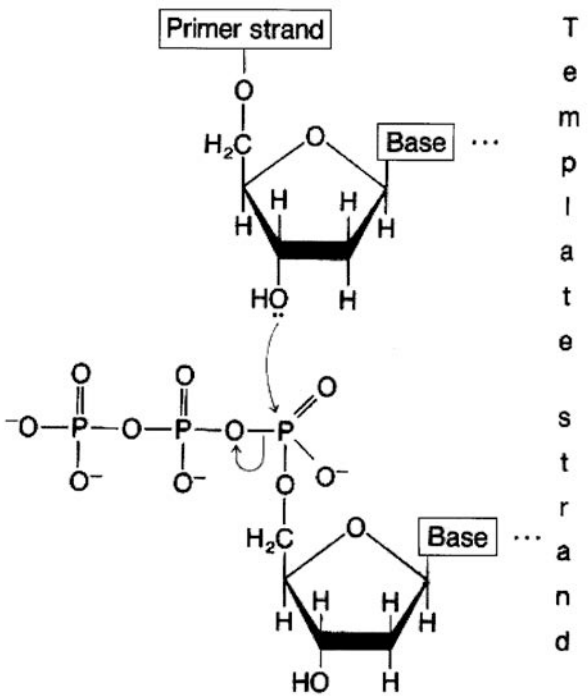
During DNA replication *in the cell (in vivo):*

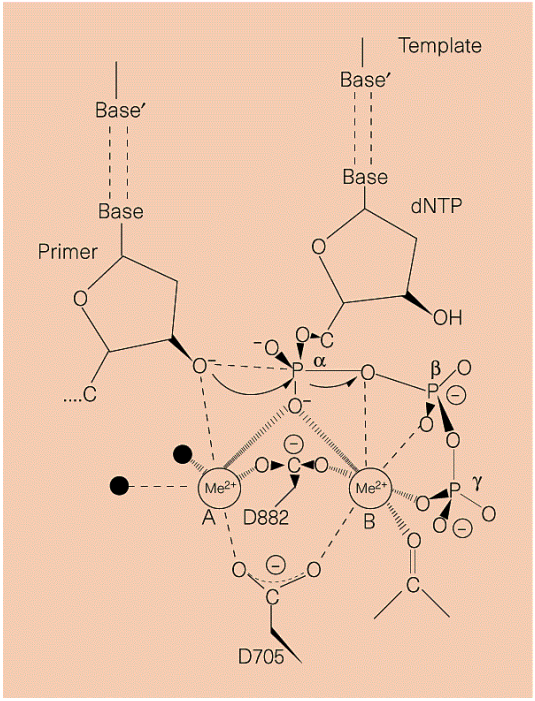
- new DNA chains are initiated with an RNA primer to which the newly synthesized DNA is attached
- then, the growing DNA chain acts as the primer

DNA synthesis *in the test tube (in vitro):*

- primers are short polymers of DNA (oligonucleotides)

What is the significance of the primer requirement?
What happens if there is no primer?





DNA synthesis occurs in the 5' to 3' direction

The monomer substrates are in the form of a dNTP
d= 2' deoxy N = A, C, G or T TP = triphosphate

dNTP's are chemically reactive monomer units

Note again that it is the 3' end of the primer chain that forms the bond

The Fidelity of DNA Replication

- ***A species genome is the record of instructions specifying the assembly and functioning of the organism***

- ***Propagation of the species requires the accurate copying over (replication) of this set of instructions***

- ***For organisms with large complex genomes, attaining sufficient accuracy is an impressive feat***

- ***Studies have shown that high complexity of a genetic program necessitates a correspondingly high accuracy of copying for it to be transmitted to offspring as a meaningful program.***

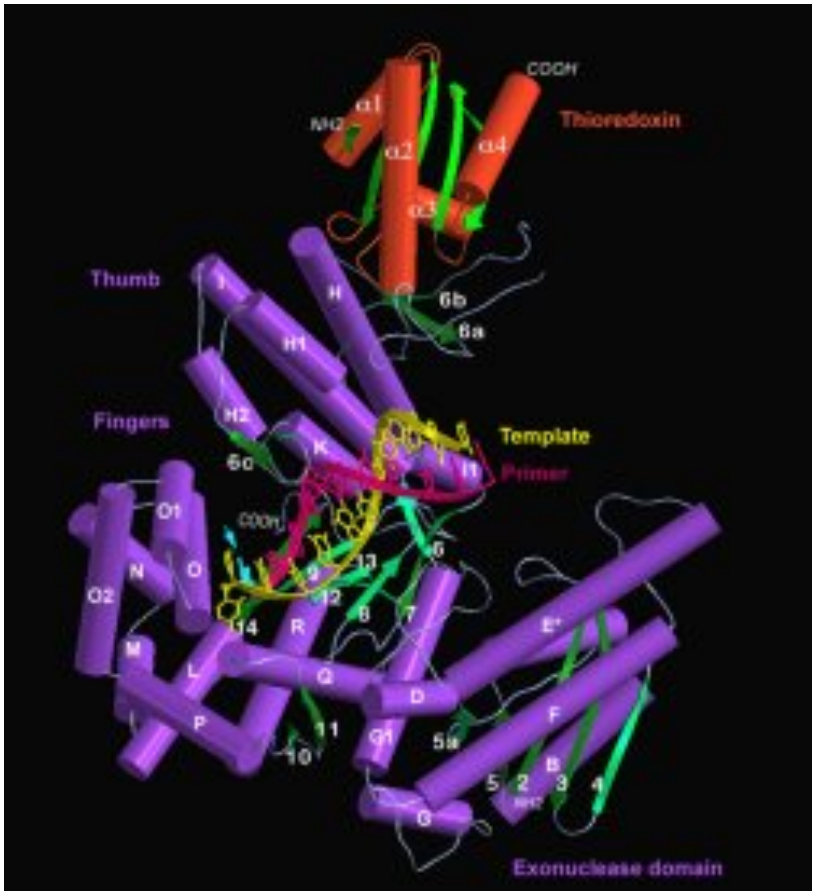
- ***For a given genomic complexity, there is a critical-copying fidelity below which information can not longer be maintained.***

Cell's strategy for getting a high-quality DNA replication product?

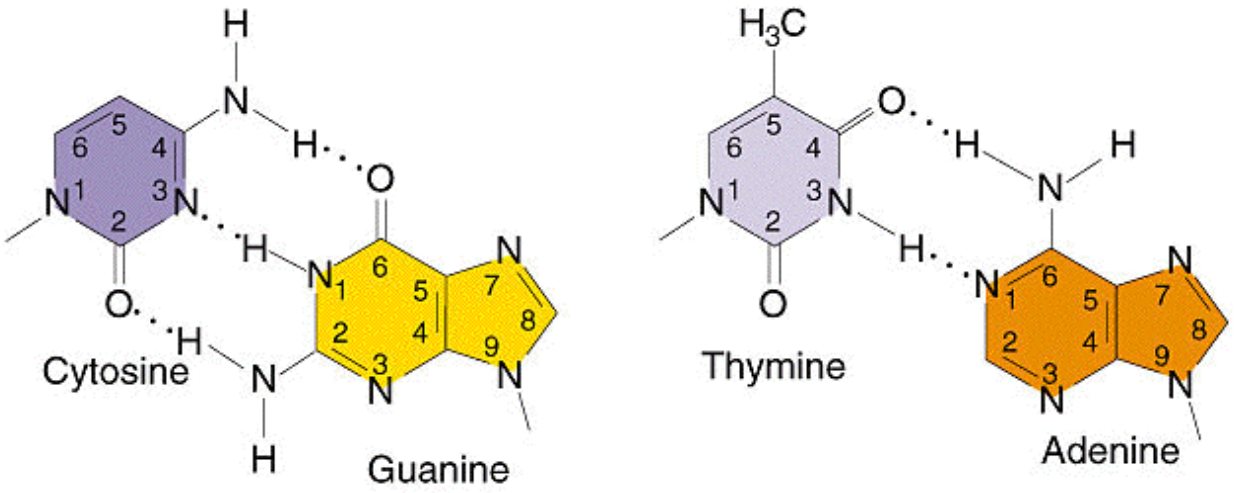
- ***Nucleotide Selection:*** ***do a good job to begin with*** **(some inherent limitations of the base-pairing specificities though)**

- ***Proofreading:*** ***perform an immediate double-check*** **of your work as you go along**

- ***Post-replication mismatch repair:*** ***go back and double-check your work again*** **against the original copy after you've completed the task**



Quality control mechanisms:
(1) ***Nucleotide selection:*** correct pairing is the most energetically favorable; that is, the complex of polymerase, template DNA and nucleotide triphosphate is the most stable when the nucleotide bound is complementary to the template nucleotide.



How specific is this interaction: GC and AT?

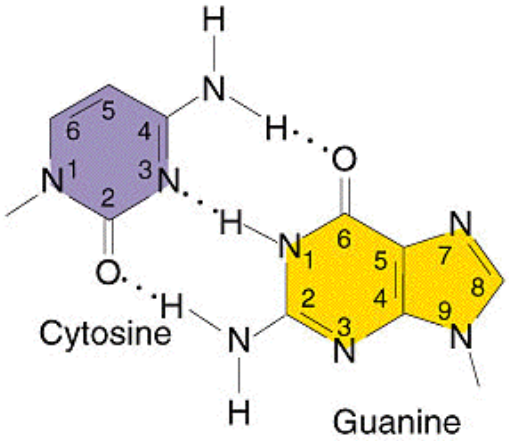
How many mistakes are made by DNA polymerase at this step?

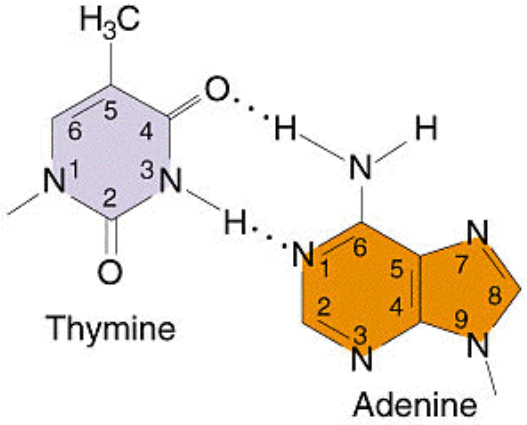
In a DNA molecular inside the cell, the bases are in an equilibrium between a stable form (shown in previous figure) and an unstable form (next figure)

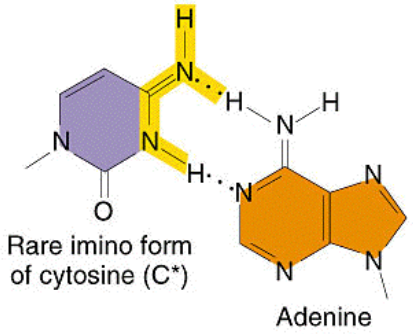
Common forms ⟵ ⟶ ***rare form**

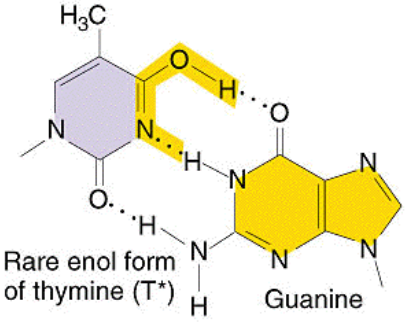
A-T G-C **G*-T G-T***

A-C* A*-C

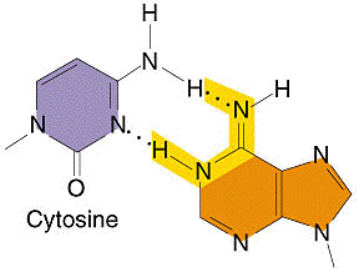

Cytosine
Guanine

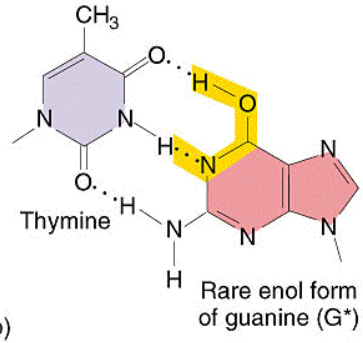

Thymine
Adenine


Rare imino form
of cytosine (C*)
Adenine


Rare enol form
of thymine (T*)
Guanine

(a)


Cytosine
Rare imino form
of adenine (A*)


Thymine
Rare enol form
of guanine (G*)

(b)

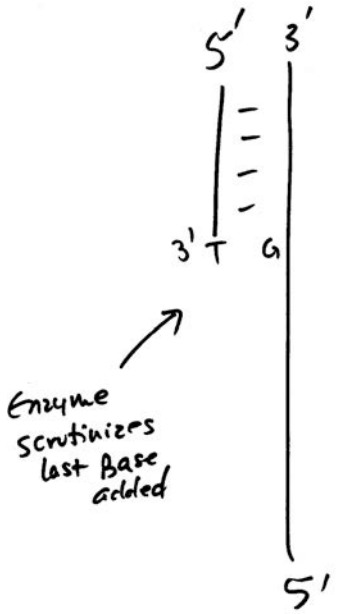
Mistake in nucleotide selection can occur when a base is in its rare form

Error rate at this initial polymerization step: ***1 in 10^4 - 10^6 nucleotides incorporated is non-complementary***

The error rate at nucleotide selection seems pretty good. ***Is it?***

(2) <u>Proofreading by the DNA polymerase</u>: the polymerase edits the DNA sequence by removing mispaired nucleotides that have been incorrectly inserted during the polymerization reaction.

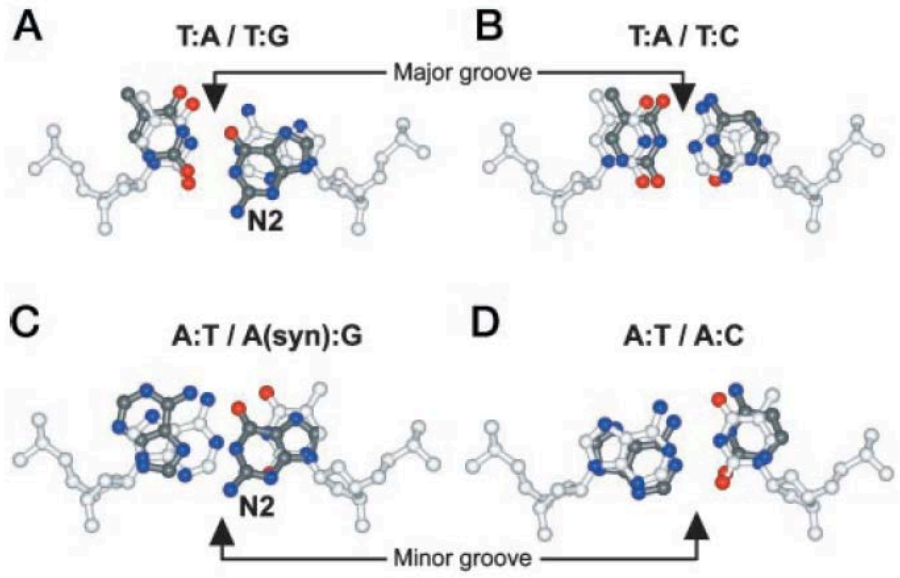
DNA polymerase has a 3'-5'exonuclease function that acts as a proofreader: **the enzyme "scrutinizes" the most recently added nucleotide and removes it if it is non-complementary**

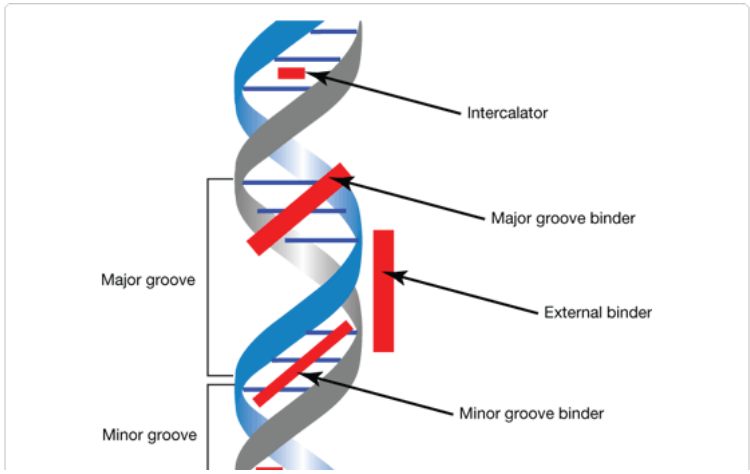


Mismatched bases are detected as they have weaker bonding interactions—the 'melting' temperature is lower—and this increases the chance of switching from the polymerase to the exonuclease active site

When DNA polymerase stalls at a mismatched base, the proofreading function kicks in

Superpositions of the structures of Watson-Crick basepairs and mispairs ***Base Substitution Specificity of DNA Polymerase β Depends on Interactions in the DNA Minor Groove:*** unfavorable interactions between an active site arginine side chain and mispair-specific atoms in the minor groove contribute to DNA polymerase specificity.





THE JOURNAL OF BIOLOGICAL CHEMISTRY Vol. 274, pp. 20749–20752, 1999

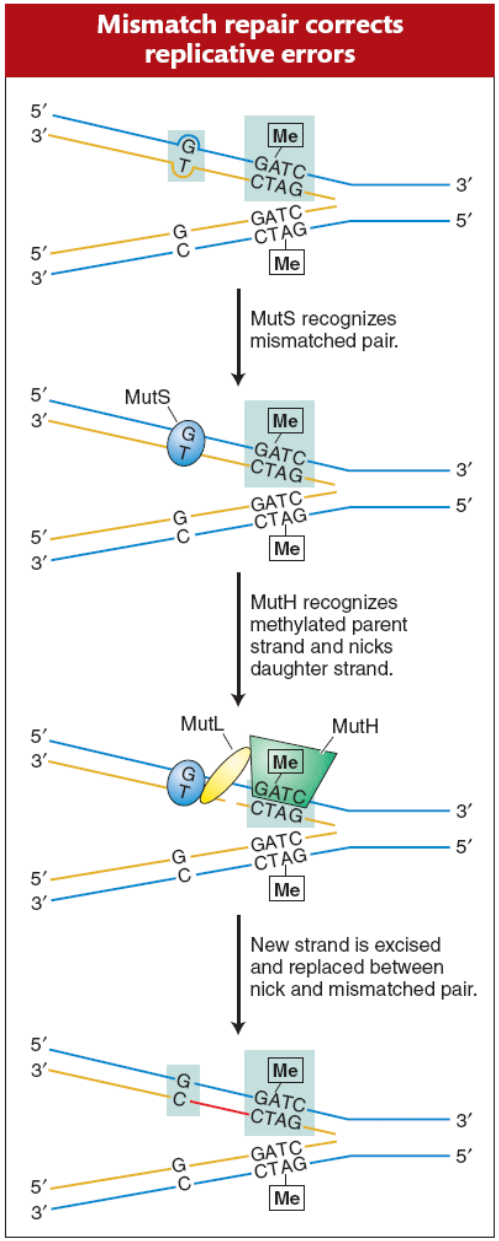
.\ Each panel shows a ball-and-stick representation of a correct Watson-Crick base pair superimposed with the mispair indicated (as described under "Experimental Procedures"). The Watson-Crick base pairs are shown in white, and the bases of the mispaired nucleotides are shown in gray. The oxygen atoms of the bases are depicted as red balls and the nitrogen atoms as blue balls. The major and minor grooves of the DNA helix are indicated. A, comparison of a correct T-A base pair with a T-G wobble base pair; B, the T-A base pair compared with a T-C mispair; C, comparison of a Watson-Crick A-T base pair with the A(syn)-G pair; D, the A-T base pair compared with an A-C mispair.

(3) Post-synthesis correction mechanisms: correction of errors that remain after proofreading. One example is mismatch repair which correct mismatches in newly synthesized DNA by

(a) detecting a mismatched base pair,
(b) removing the mismatched nucleotide,
(c) replacing it with the correct nucleotide.

How does the enzymatic machinery know which of the two mismatched bases to correct?

Figure 15-26 in 9th
Figure 16-23 in 10th



Post replication mismatch repair in E. coli

Mut = mutator

The components of mismatch repair systems are ***very highly conserved from bacteria to man***

Eukaryotes use different mechanisms to distinguish between the parental and daughter strands (but as of 2007 were yet undefined)

Inherited mutations in the mismatch repair system predispose an inidividual to colon cancer (HNPCC)

Net (final) error rate after post-synthesis correction is estimated to be: 1 in 10^9 - 10^{10}
(one mistake per one billion to ten billion nucleotides replicated)

Interestingly, the final error rate (# mistakes per nucleotide replicated) appears to be the same in E. coli and humans despite the very large difference in genome size

Analysis of Genetic Inheritance in a Family Quartet by Whole-Genome Sequencing

Jared C. Roach,[1*] Gustavo Glusman,[1*] Arian F. A. Smit,[1*] Chad D. Huff,[1,2*] Robert Hubley,[1] Paul T. Shannon,[1] Lee Rowen,[1] Krishna P. Pant,[3] Nathan Goodman,[1] Michael Bamshad,[4] Jay Shendure,[5] Radoje Drmanac,[3] Lynn B. Jorde,[2] Leroy Hood,[1†] David J. Galas[1†]

We analyzed the whole-genome sequences of a family of four, consisting of two siblings and their parents. Family-based sequencing allowed us to delineate recombination sites precisely, identify 70% of the sequencing errors (resulting in >99.999% accuracy), and identify very rare single-nucleotide polymorphisms. We also directly estimated a human intergeneration mutation rate of $\sim 1.1 \times 10^{-8}$ per position per haploid genome. Both offspring in this family have two recessive disorders: Miller syndrome, for which the gene was concurrently identified, and primary ciliary dyskinesia, for which causative genes have been previously identified. Family-based genome analysis enabled us to narrow the candidate genes for both of these Mendelian disorders to only four. Our results demonstrate the value of complete genome sequencing in families.

30 APRIL 2010 VOL 328 **SCIENCE** www.sciencemag.org

Is this frequency inconsistent with the error rates on the previous page?

Current Biology 2001, 11:R929–R931

Evolution: **Constantly avoiding mutation**

Mutation plays a crucial but paradoxical role in evolution: the genetic variation that it introduces is ultimately required for adaptation, but individual mutations are far more likely to be deleterious than to be beneficial. Every natural population thus labors upstream against a steady current of deleterious mutations, and natural selection favors decreased mutation rates under almost all circumstances. ***Realizing this, A.H. Sturtevant wondered in an early essay why mutation rates did not evolve all the way to zero.***

In answer to his own question, Sturtevant offered the surmise that "...the nature of genes does not permit such a reduction. In short, mutations are accidents, and accidents will happen...".

In a fundamental sense, mutations are indeed accidents. They originate through spontaneous genome damage and replication error and the rogue activities of transposable elements. An elaborate molecular machinery has evolved to reverse most of these genomic insults; what this machinery does not repair — or what it repairs incorrectly — passes through to the next generation as a mutation. Although certain types of sequence are more susceptible to damage and mutation than others, specific changes are not manufactured on demand, and beneficial mutations are therefore every bit as accidental as deleterious one

➡ *Some geneticists suggest that natural selection may have exhausted most the general ways of maintaining genetic fidelity -- in other words, this is a good as it gets given the inherent noise in any biochemical process*

The above error rate is that observed/estimated for replication of genomic DNA in eukaryotes and prokaryotes.

- *In contrast the error rate during replication of many viral genomes is much higher, probably due to the fact that their replication enzymes generally don't have error correction capabilities.*
- For example, the error rate for replication of the HIV (AIDS virus) genome is $1X10^{-4}$ or one mistake in every 10,000 bases copied.
- This high error rate certainly contributes to the high mutation rate of this virus.

Optional Reading Assignment

For your personal enrichment, see this thoughtful essay on biological processes and fidelity

Nature 413: 115 Sept. 13, 2001

http://fire.biol.wwu.edu/trent/trent/fidelity.pdf