

Disclaimer

This material is only for internal use and is given to the students to prepare the final evaluation for the course of Applied Genomics of the Master degree course of Bioinformatics.

This file and its content are confidential and intended solely for the use of the individuals to whom they are given. If you have received this file it means that you are a student of the course of Applied Genomics of the master degree course in Bioinformatics, regularly enrolled for the academic year 2013-2014. If you are not a student of this course you should not disseminate, distribute or copy this file. Please notify the professor immediately by e-mail if you have received this file.

In any case, also students of this course are notified that disclosing, copying, distributing or taking any action in reliance on the contents of this information is strictly prohibited.

For the students: please note that the content of this file is not enough to pass the exam. The content of this file could contain a few errors as it has not been peer reviewed or edited after its preparation.

Applied Genomics

articles

Initial sequencing and analysis of the human genome

International Human Genome Sequencing Consortium*

** A partial list of authors appears on the opposite page. Affiliations are listed at the end of the paper.*

The human genome holds an extraordinary trove of information about human development, physiology, medicine and evolution. Here we report the results of an international collaboration to produce and make freely available a draft sequence of the human genome. We also present an initial analysis of the data, describing some of the insights that can be gleaned from the sequence.



The Sequence of the Human Genome

J. Craig Venter *et al.*

Science **291**, 1304 (2001);

DOI: 10.1126/science.1058040

There are two approaches for sequencing large repeat-rich genomes.

- 1) whole-genome shotgun sequencing approach,
- 2) hierarchical shotgun sequencing' approach also referred to as 'map-based', 'BAC-based' or 'clone-by-clone'.

This approach involves generating and organizing a set of large-insert clones (typically 100–200 kb each) covering the genome and separately performing shotgun sequencing on appropriately chosen clones.

Because the sequence information is local, the issue of long-range misassembly is eliminated and the risk of short-range misassembly is reduced.

One caveat is that some large-insert clones may suffer rearrangement, although this risk can be reduced by appropriate quality-control measures involving clone fingerprints.

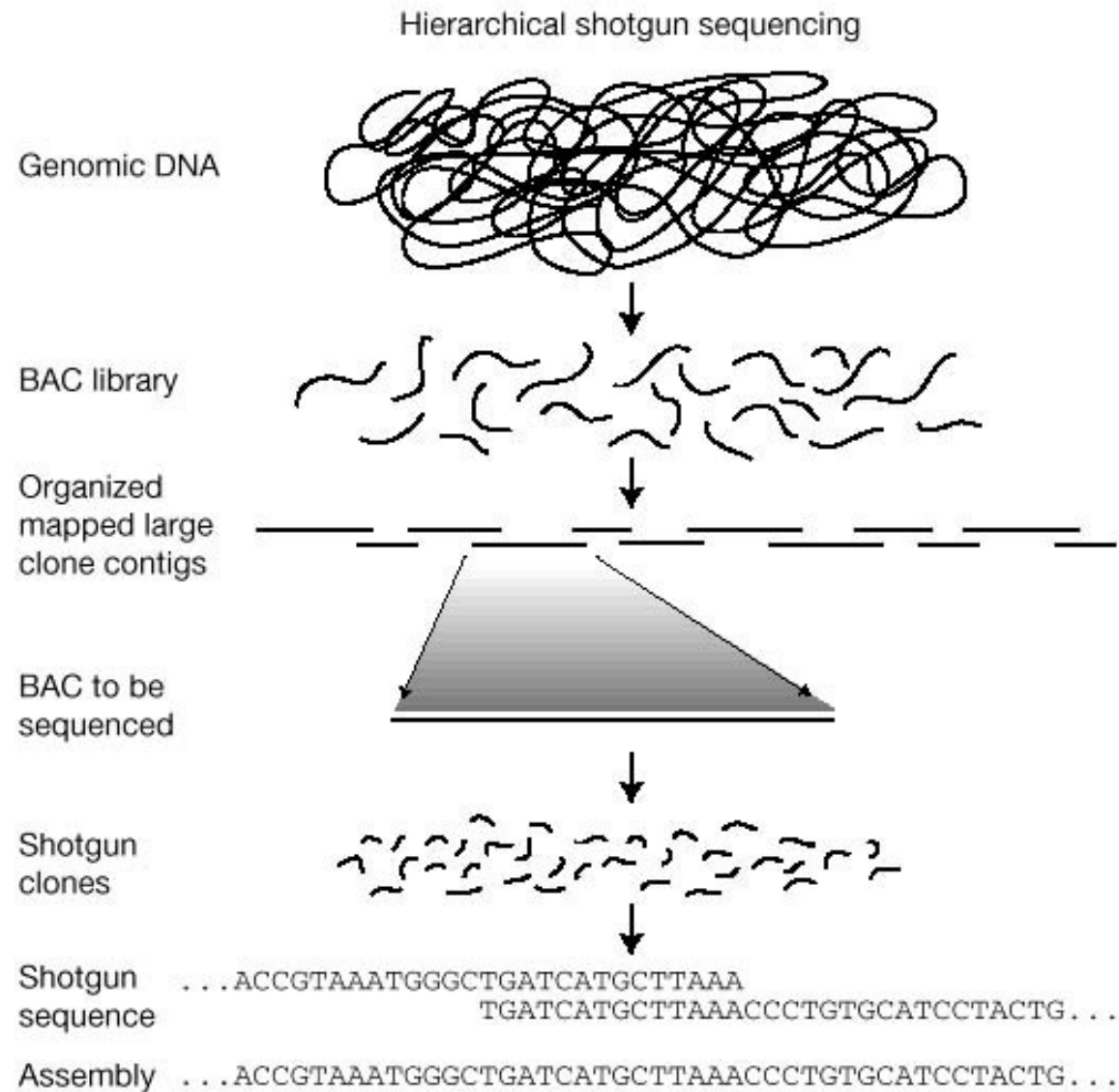
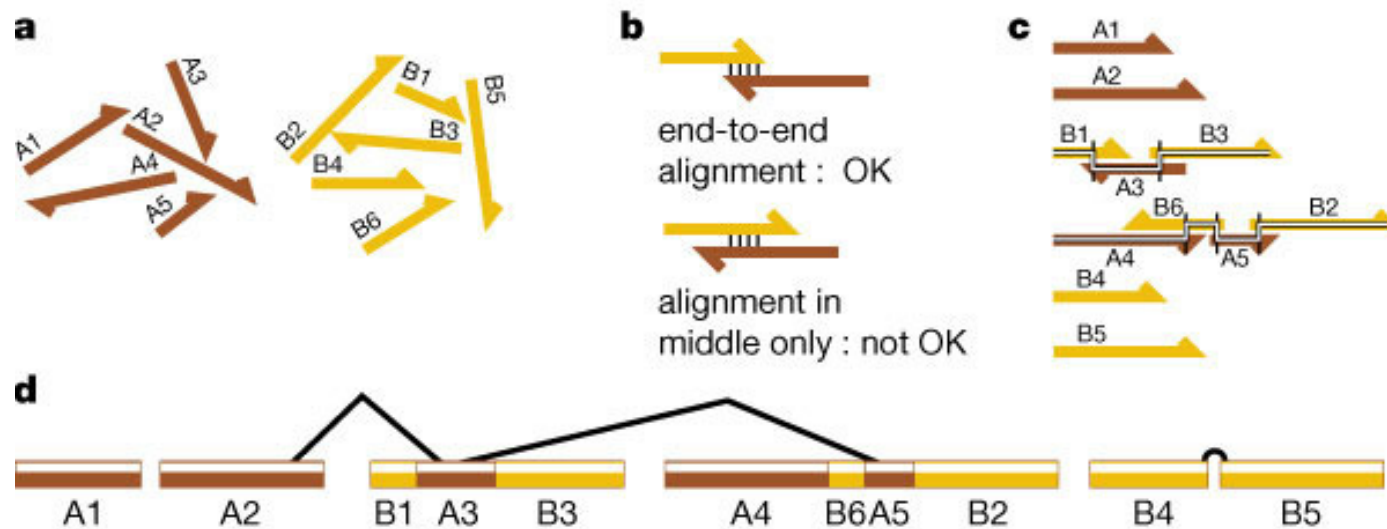


Figure 2 Idealized representation of the hierarchical shotgun sequencing strategy. A library is constructed by fragmenting the target genome and cloning it into a large-fragment cloning vector; here, BAC vectors are shown. The genomic DNA fragments represented in the library are then organized into a physical map and individual BAC clones are selected and sequenced by the random shotgun strategy. Finally, the clone sequences are assembled to reconstruct the sequence of the genome.



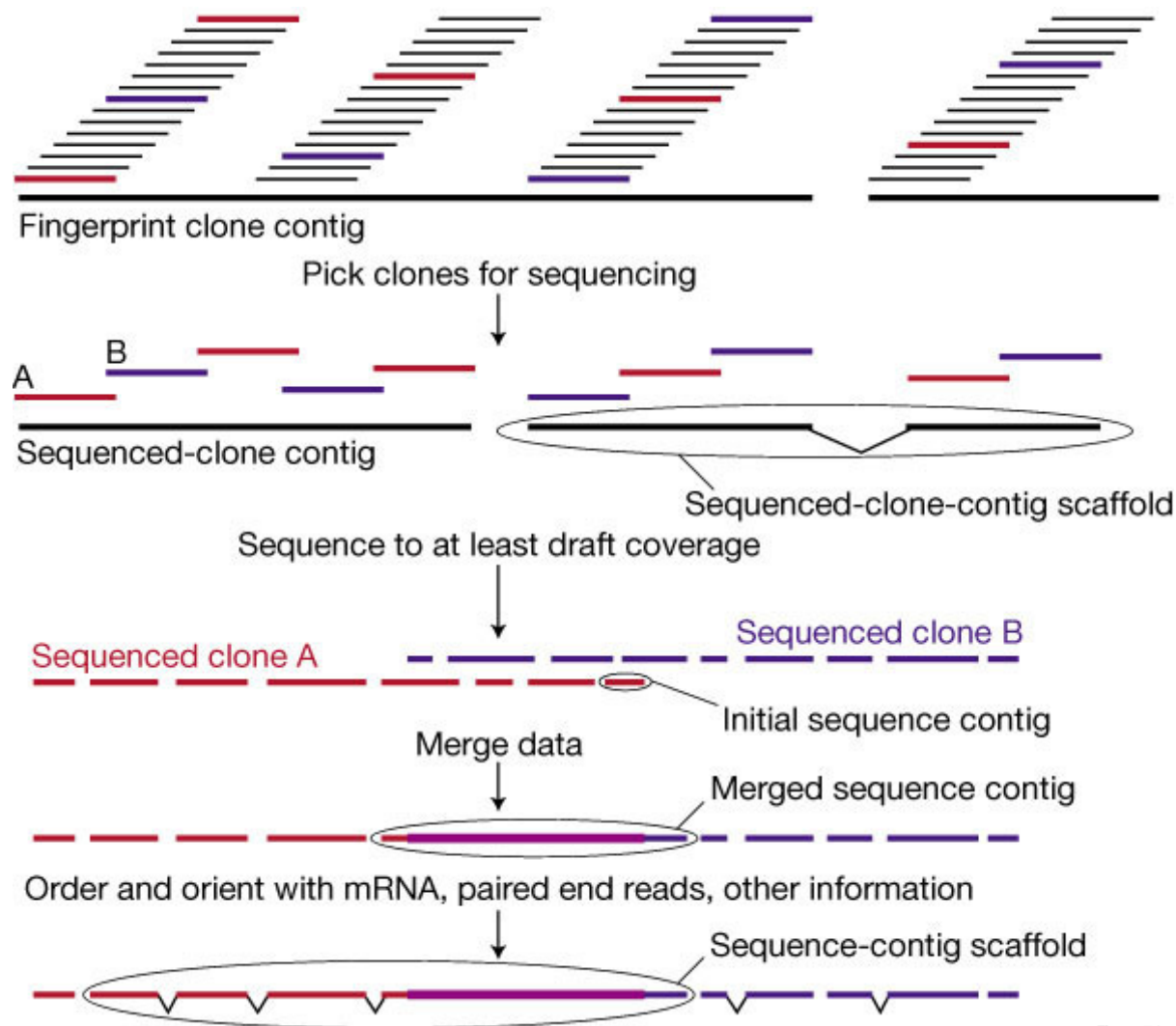
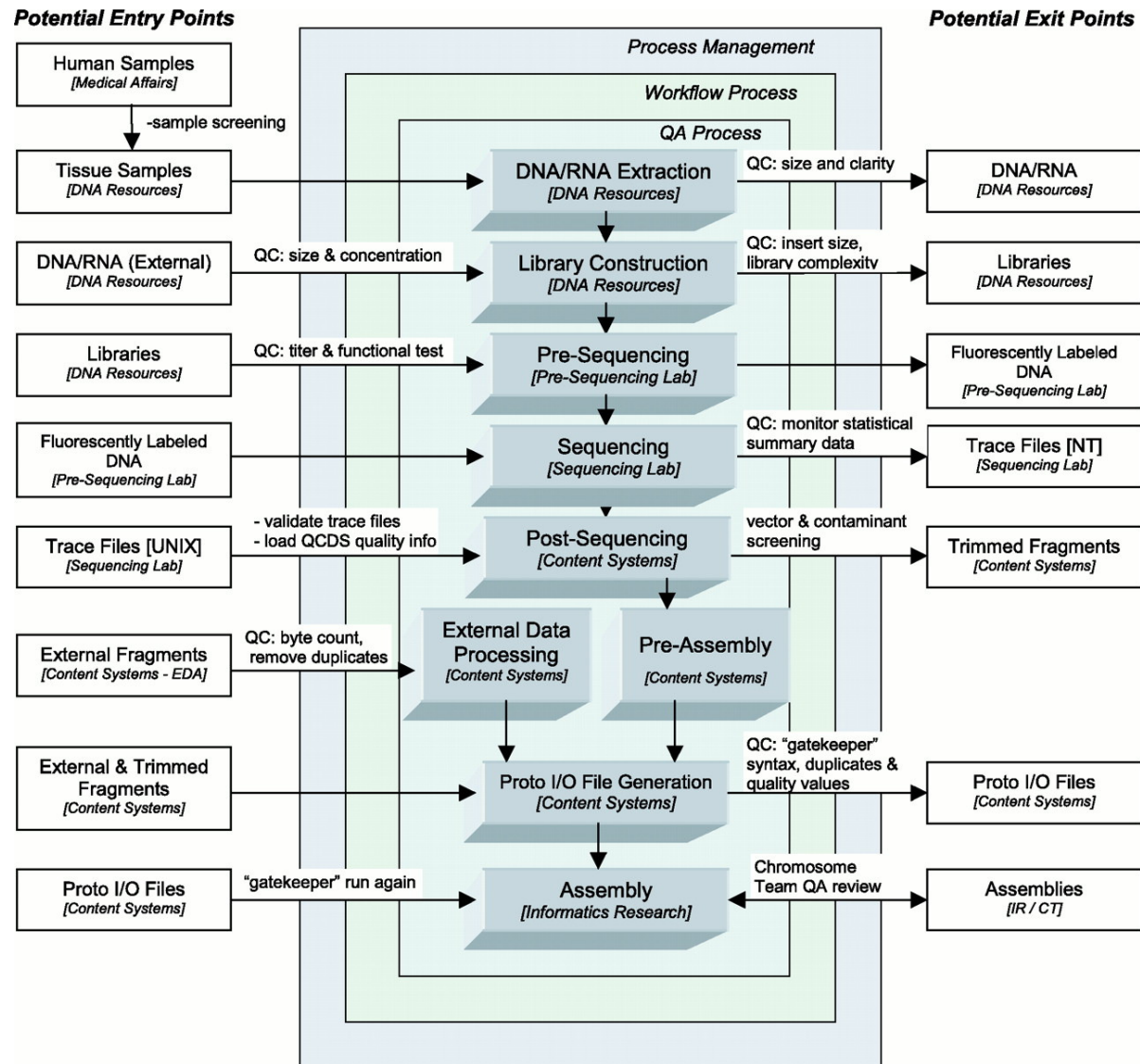
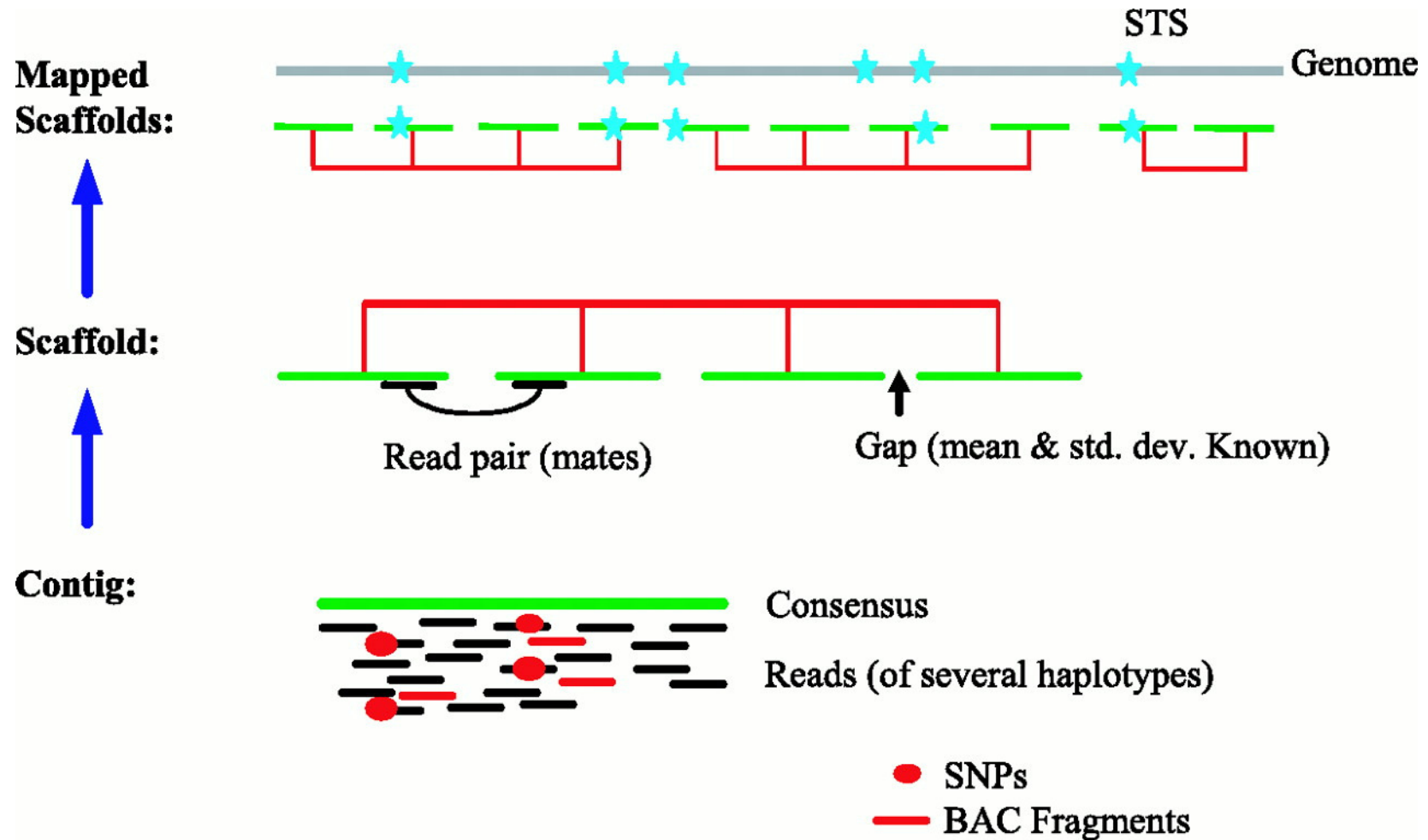


Figure 7 Levels of clone and sequence coverage. A 'fingerprint clone contig' is assembled by using the computer program FPC^{84,451} to analyse the restriction enzyme digestion patterns of many large-insert clones. Clones are then selected for sequencing to minimize overlap between adjacent clones. For a clone to be selected, all of its restriction enzyme fragments (except the two vector-insert junction fragments) must be shared with at least one of its neighbours on each side in the contig. Once these overlapping clones have been sequenced, the set is a 'sequenced-clone contig'. When all selected clones from a fingerprint clone contig have been sequenced, the sequenced-clone contig will be the same as the fingerprint clone contig. Until then, a fingerprint clone contig may contain several sequenced-clone contigs. After individual clones (for example, A and B) have been sequenced to draft coverage and the clones have been mapped, the data are analysed by GigAssembler (Fig. 6), producing merged sequence contigs from initial sequence contigs, and linking these to form sequence-contig scaffolds (see Box 1).

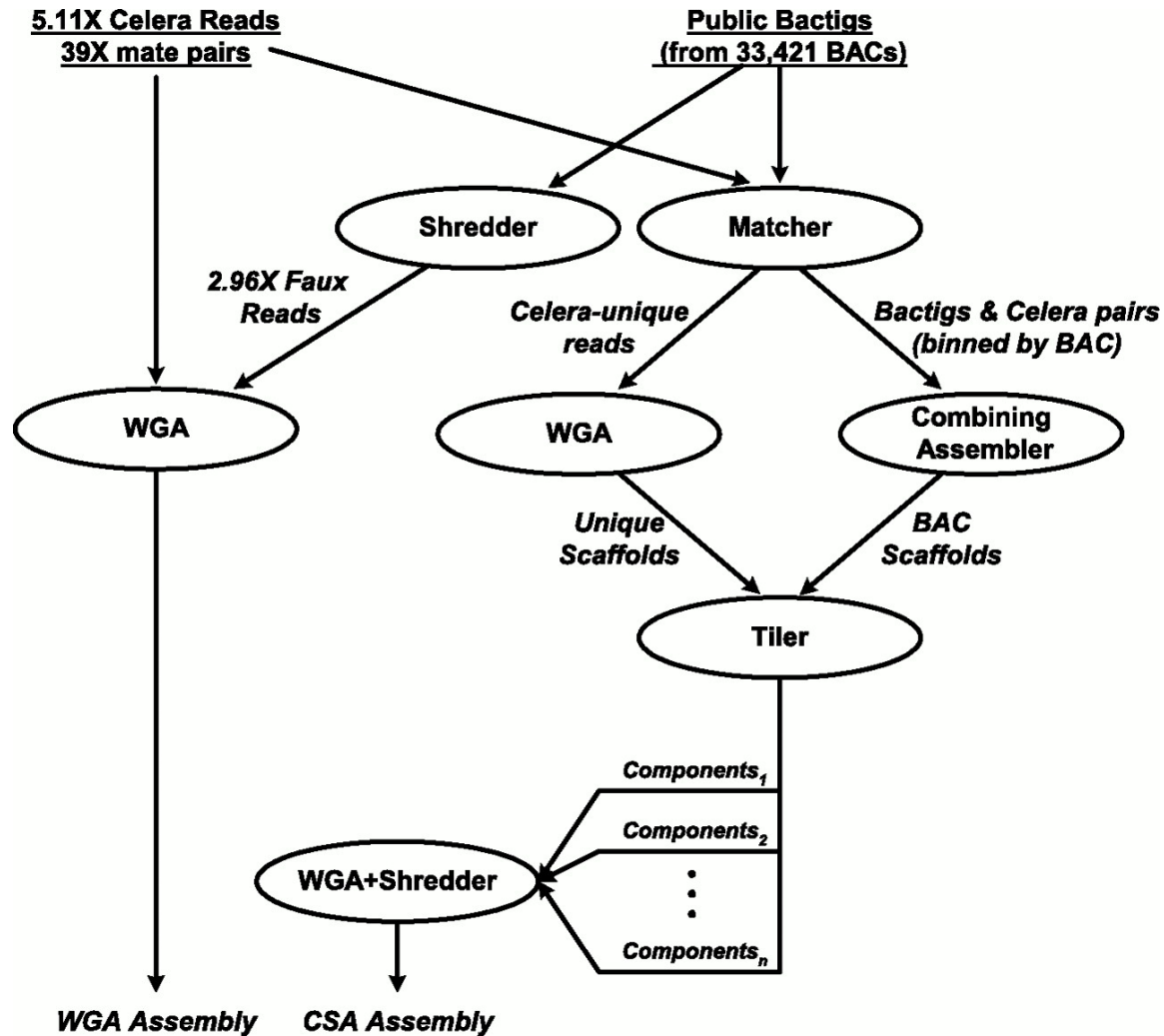
Flow diagram for sequencing pipeline.



Anatomy of whole-genome assembly.

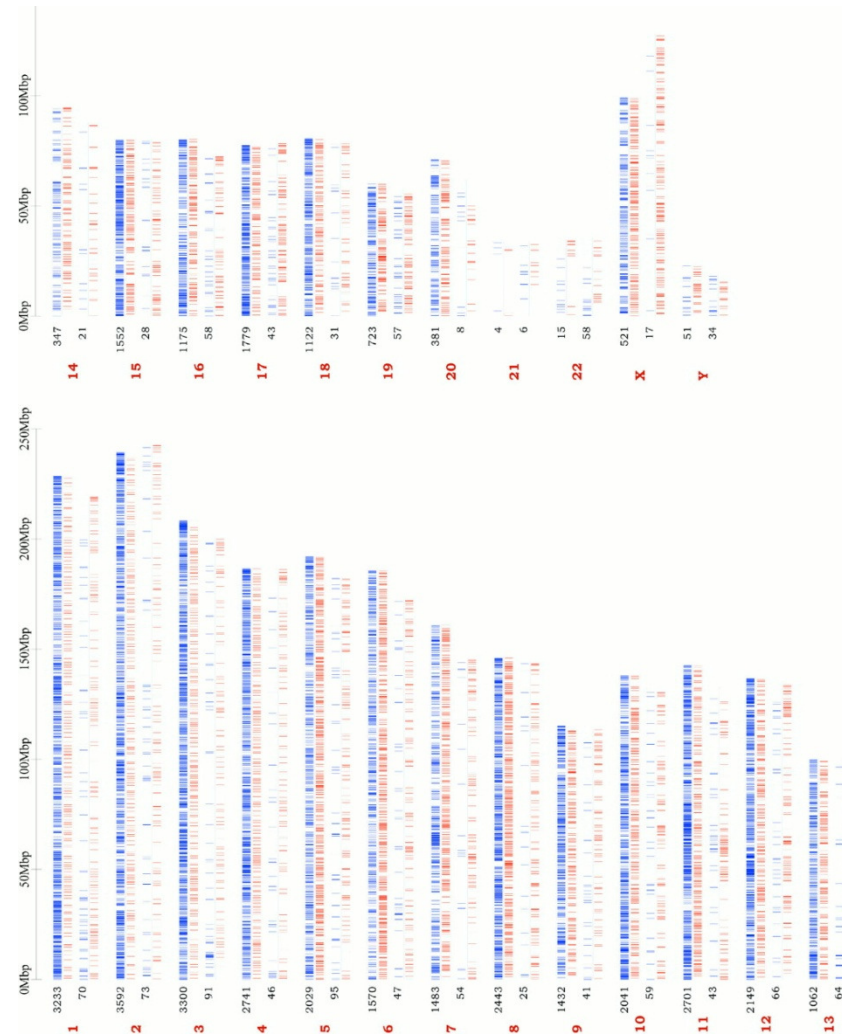


Architecture of Celera's two-pronged assembly strategy.



J C Venter et al. Science 2001;291:1304-1351

Schematic view of the distribution of breakpoints and large gaps on all chromosomes.



J C Venter et al. Science 2001;291:1304-1351

Genome glossary

Sequence

Raw sequence Individual unassembled sequence reads, produced by sequencing of clones containing DNA inserts.

Paired-end sequence Raw sequence obtained from both ends of a cloned insert in any vector, such as a plasmid or bacterial artificial chromosome.

Finished sequence Complete sequence of a clone or genome, with an accuracy of at least 99.99% and no gaps.

Coverage (or depth) The average number of times a nucleotide is represented by a high-quality base in a collection of random raw sequence. Operationally, a 'high-quality base' is defined as one with an accuracy of at least 99% (corresponding to a PHRED score of at least 20).

Full shotgun coverage The coverage in random raw sequence needed from a large-insert clone to ensure that it is ready for finishing; this varies among centres but is typically 8–10-fold. Clones with full shotgun coverage can usually be assembled with only a handful of gaps per 100 kb.

Half shotgun coverage Half the amount of full shotgun coverage (typically, 4–5-fold random coverage).

Clones

BAC clone Bacterial artificial chromosome vector carrying a genomic DNA insert, typically 100–200 kb. Most of the large-insert clones sequenced in the project were BAC clones.

Finished clone A large-insert clone that is entirely represented by finished sequence.

Full shotgun clone A large-insert clone for which full shotgun sequence has been produced.

Draft clone A large-insert clone for which roughly half-shotgun sequence has been produced. Operationally, the collection of draft clones produced by each centre was required to have an average coverage of fourfold for the entire set and a minimum coverage of threefold for each clone.

Predraft clone A large-insert clone for which some shotgun sequence is available, but which does not meet the standards for inclusion in the collection of draft clones.

Contigs and scaffolds

Contig The result of joining an overlapping collection of sequences or clones.

Scaffold The result of connecting contigs by linking information from paired-end reads from plasmids, paired-end reads from BACs, known messenger RNAs or other sources. The contigs in a scaffold are ordered and oriented with respect to one another.

Fingerprint clone contigs Contigs produced by joining clones inferred to overlap on the basis of their restriction digest fingerprints.

Sequenced-clone layout Assignment of sequenced clones to the physical map of fingerprint clone contigs.

Initial sequence contigs Contigs produced by merging overlapping sequence reads obtained from a single clone, in a process called sequence assembly.

Merged sequence contigs Contigs produced by taking the initial sequence contigs contained in overlapping clones and merging those found to overlap. These are also referred to simply as 'sequence contigs' where no confusion will result.

Sequence-contig scaffolds Scaffolds produced by connecting sequence contigs on the basis of linking information.

Sequenced-clone contigs Contigs produced by merging overlapping sequenced clones.

Sequenced-clone-contig scaffolds Scaffolds produced by joining sequenced-clone contigs on the basis of linking information.

Draft genome sequence The sequence produced by combining the information from the individual sequenced clones (by creating merged sequence contigs and then employing linking information to create scaffolds) and positioning the sequence along the physical map of the chromosomes.

N50 length A measure of the contig length (or scaffold length) containing a 'typical' nucleotide. Specifically, it is the maximum length L such that 50% of all nucleotides lie in contigs (or scaffolds) of size at least L .

Computer programs and databases

PHRED A widely used computer program that analyses raw sequence to produce a 'base call' with an associated 'quality score' for each position in the sequence. A PHRED quality score of X corresponds to an error probability of approximately $10^{-X/10}$. Thus, a PHRED quality score of 30 corresponds to 99.9% accuracy for the base call in the raw read.

PHRAP A widely used computer program that assembles raw sequence into sequence contigs and assigns to each position in the sequence an associated 'quality score', on the basis of the PHRED scores of the raw sequence reads. A PHRAP quality score of X corresponds to an error probability of approximately $10^{-X/10}$. Thus, a PHRAP quality score of 30 corresponds to 99.9% accuracy for a base in the assembled sequence.

GigAssembler A computer program developed during this project for merging the information from individual sequenced clones into a draft genome sequence.

Public sequence databases The three coordinated international sequence databases: GenBank, the EMBL data library and DDBJ.

Map features

STS Sequence tagged site, corresponding to a short (typically less than 500 bp) unique genomic locus for which a polymerase chain reaction assay has been developed.

EST Expressed sequence tag, obtained by performing a single raw sequence read from a random complementary DNA clone.

SSR Simple sequence repeat, a sequence consisting largely of a tandem repeat of a specific k -mer (such as $(CA)_n$). Many SSRs are polymorphic and have been widely used in genetic mapping.

SNP Single nucleotide polymorphism, or a single nucleotide position in the genome sequence for which two or more alternative alleles are present at appreciable frequency (traditionally, at least 1%) in the human population.

Genetic map A genome map in which polymorphic loci are positioned relative to one another on the basis of the frequency with which they recombine during meiosis. The unit of distance is centimorgans (cM), denoting a 1% chance of recombination.

Radiation hybrid (RH) map A genome map in which STSs are positioned relative to one another on the basis of the frequency with which they are separated by radiation-induced breaks. The frequency is assayed by analysing a panel of human–hamster hybrid cell lines, each produced by lethally irradiating human cells and fusing them with recipient hamster cells such that each carries a collection of human chromosomal fragments. The unit of distance is centirays (cR), denoting a 1% chance of a break occurring between two loci.

Classes of interspersed repeat in the human genome

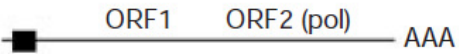
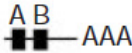
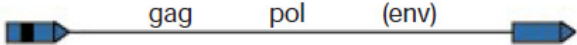



			Length	Copy number	Fraction of genome
LINEs	Autonomous		6–8 kb	850,000	21%
	Non-autonomous		100–300 bp		
Retrovirus-like elements	Autonomous		6–11 kb	450,000	8%
	Non-autonomous		1.5–3 kb		
DNA transposon fossils	Autonomous		2–3 kb	300,000	3%
	Non-autonomous		80–3,000 bp		

Table 11 Number of copies and fraction of genome for classes of interspersed repeat

	Number of copies (× 1,000)	Total number of bases in the draft genome sequence (Mb)	Fraction of the draft genome sequence (%)	Number of families (subfamilies)
SINEs	1,558	359.6	13.14	
Alu	1,090	290.1	10.60	1 (~20)
MIR	393	60.1	2.20	1 (1)
MIR3	75	9.3	0.34	1 (1)
LINEs	868	558.8	20.42	
LINE1	516	462.1	16.89	1 (~55)
LINE2	315	88.2	3.22	1 (2)
LINE3	37	8.4	0.31	1 (2)
LTR elements	443	227.0	8.29	
ERV-class I	112	79.2	2.89	72 (132)
ERV(K)-class II	8	8.5	0.31	10 (20)
ERV (L)-class III	83	39.5	1.44	21 (42)
MaLR	240	99.8	3.65	1 (31)
DNA elements	294	77.6	2.84	
hAT group				
MER1-Charlie	182	38.1	1.39	25 (50)
Zaphod	13	4.3	0.16	4 (10)
Tc-1 group				
MER2-Tigger	57	28.0	1.02	12 (28)
Tc2	4	0.9	0.03	1 (5)
Mariner	14	2.6	0.10	4 (5)
PiggyBac-like	2	0.5	0.02	10 (20)
Unclassified	22	3.2	0.12	7 (7)
Unclassified	3	3.8	0.14	3 (4)
Total interspersed repeats		1,226.8	44.83	

The number of copies and base pair contributions of the major classes and subclasses of transposable elements in the human genome. Data extracted from a RepeatMasker analysis of the draft genome sequence (RepeatMasker version 09092000, sensitive settings, using RepBase Update 5.08). In calculating percentages, RepeatMasker excluded the runs of Ns linking the contigs in the draft genome sequence. In the last column, separate consensus sequences in the repeat databases are considered subfamilies, rather than families, when the sequences are closely related or related through intermediate subfamilies.

Short interspersed repetitive elements: SINEs

- Example: Alu repeats
 - Most abundant repeated DNA in primates
 - Short, about 300 bp
 - About 1 million copies
 - Likely derived from the gene for 7SL RNA
 - Cause new mutations in humans
- They are **retrotransposons**
 - DNA segments that **move** via an **RNA intermediate**.
- MIRs: Mammalian interspersed repeats
 - SINES found in all mammals
- Analogous short retrotransposons found in genomes of all vertebrates.

Long interspersed repetitive elements: LINEs

- Moderately abundant, long repeats
 - LINE1 family: most abundant
 - Up to 7000 bp long
 - About 50,000 copies
- Retrotransposons
 - Encode reverse transcriptase and other enzymes required for transposition
 - No long terminal repeats (LTRs)
- Cause new mutations in humans
- Homologous repeats found in all mammals and many other animals

Other common interspersed repeated sequences in humans

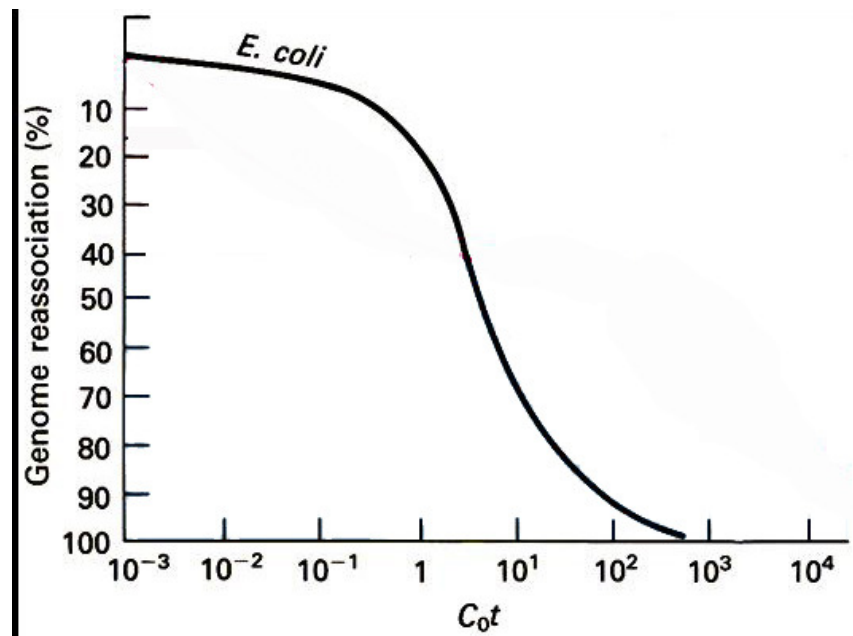
- LTR-containing retrotransposons
 - MaLR: mammalian, LTR retrotransposons
 - Endogenous retroviruses
 - MER4 (MEdium Reiterated repeat, family 4)
- Repeats that resemble DNA transposons
 - MER1 and MER2
 - Mariner repeats

Finding repeats

- Compare a sequence to a database of known repeat sequences from the organism of interest
- RepeatMasker
- <http://www.repeatmasker.org/>

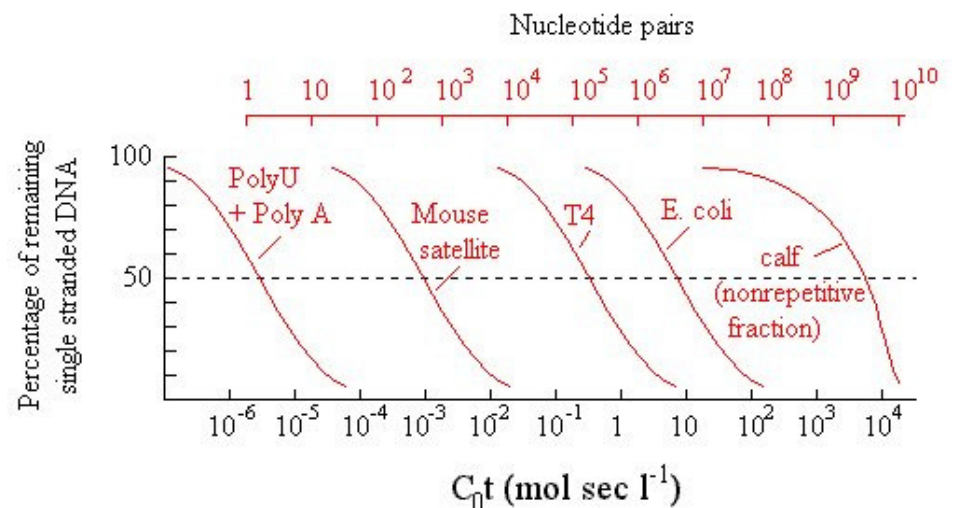
Cot Curves

- Cot curves are generated by shearing DNA to about 1000 bp length, then melting it, lowering the temp and allowing it to re-anneal, measuring the % still single stranded at various time points.
- The rate-limiting step is the collision of two complementary molecules, giving second-order reaction kinetics. The rate of collisions is proportional to initial concentration (C_0) times time (t), or C_0t .
- Whether a collision results in formation of a double stranded molecule depends on whether the two strands are complementary.
- Get a sigmoid curve which can be characterized by the **$C_0t_{1/2}$** value, the point where 1/2 of the DNA is still single stranded.



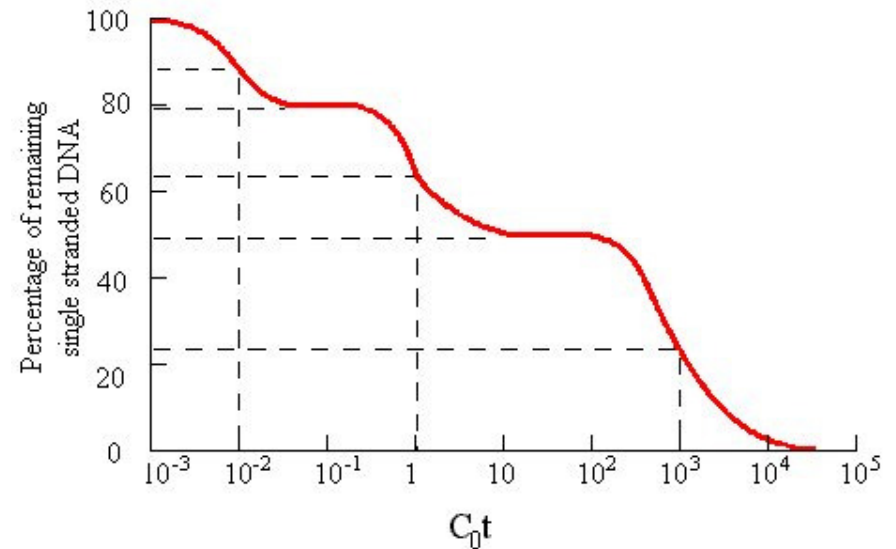
Cot Curves and Copy Number

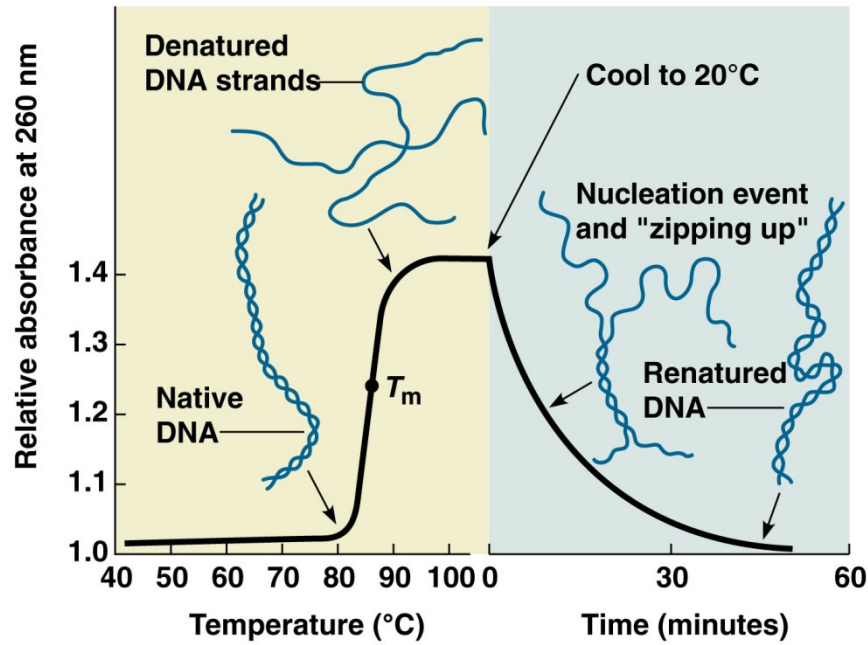
- Number of copies of each sequence determines the rate: how many collisions a given strand has to make before it finds a match.
- For example, if one strand is all U's (poly U) and the other strand is all A's (poly A), on the average only 2 collisions will occur before a strand finds a match.
- For 50 kb phage DNA cut into 1 kb lengths, only 1 collision in 100 will result in a match: $C_0t \frac{1}{2}$ is bigger.
- For 4 Mbp E. coli genome, one collision in 8000 will be productive.



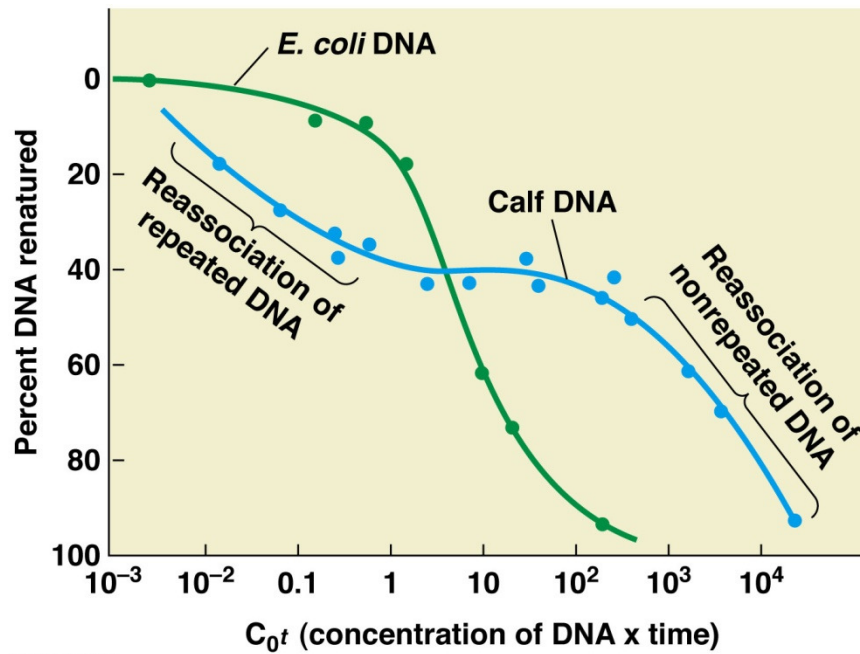
Complex Cot Curves

- For eukaryotic DNA, Cot curves are not simple sigmoid curves. Computer analysis generally resolves them into 3 sections:
 - highly repeated DNA: average of 50,000 copies per genome, about 10% of total DNA
 - moderately repeat DNA: average of 500 copies, a total of 30% of the genome
 - unique sequence DNA: up to 10 copies: about 60% of the genome.



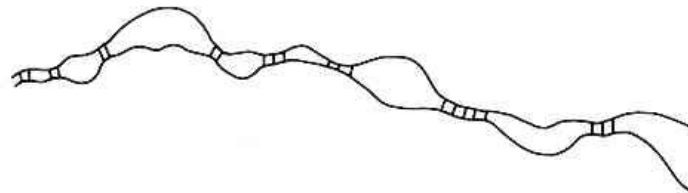


© 2012 Pearson Education, Inc.

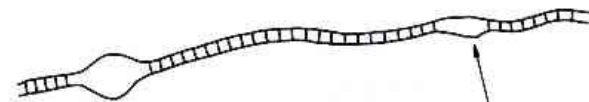


© 2012 Pearson Education, Inc.

(a) An unstable hybrid

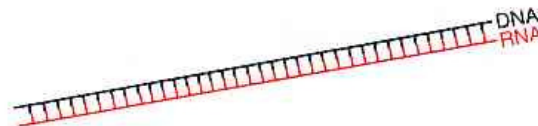


(b) A stable hybrid



Short non-complementary regions
do not affect overall stability

(c) A DNA-RNA hybrid

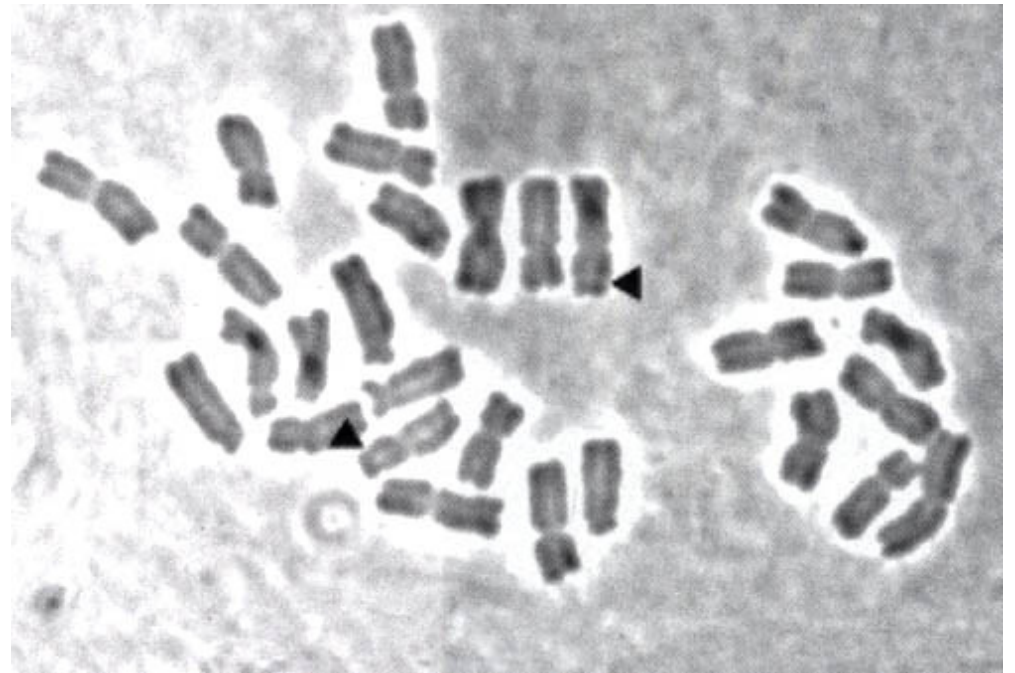


What Repeat Classes Represent

- Unique DNA:
 - highly conserved coding regions: 1.5%
 - other highly conserved regions: 3%
 - other non-conserved unique sequences: 44%
- Moderately repeated DNA
 - transposon-based repeats: 45%
 - large gene families
- Highly repeated DNA:
 - constitutive heterochromatin: 6.6%
 - microsatellites: 2%
 - a few highly repeated transposon families (Alu sequences)

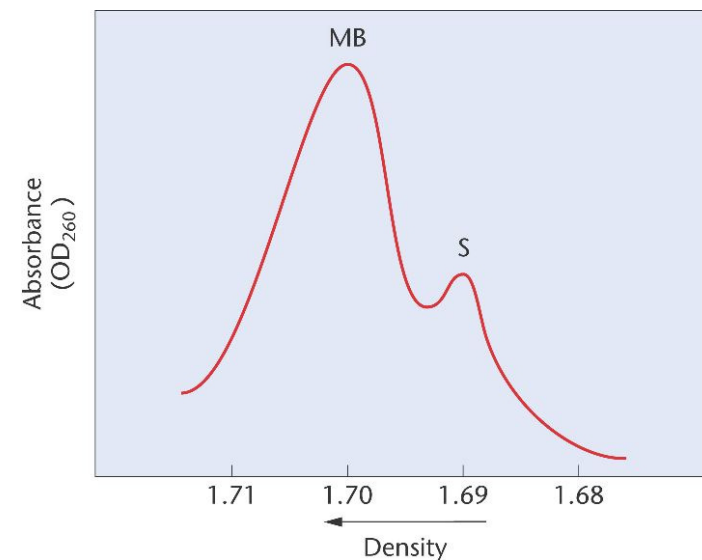
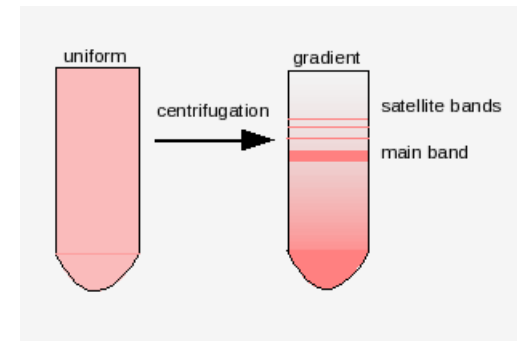
Highly Repeated Sequences

- Short sequences in long tandem arrays, mostly near centromeres or on the short arms of acrocentric chromosomes. Some are also on other chromosome arms, appearing as “secondary constrictions” in metaphase chromosomes under the microscope (centromere is the primary constriction).
- Constitutive heterochromatin is composed of highly repeated DNA. As seen in the microscope, it is densely staining and late replicating chromosomal material. It contains very few genes.
- These sequences are not normally transcribed.
- Subclasses of highly repeated sequences: satellite, minisatellites, microsatellites.



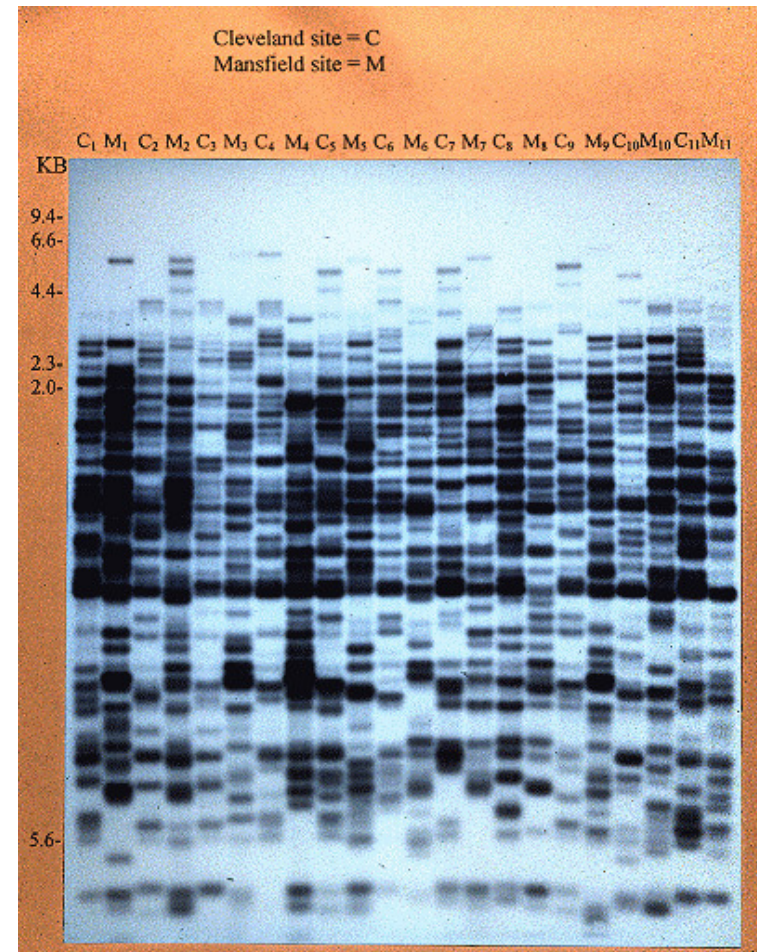
Satellite DNA

- Satellite DNA: based on DNA's behavior during density gradient (isopycnic) centrifugation. During centrifugation at $50,000 \times g$, a CsCl solution settles into a gradient of density: more dense near the bottom of the tube. Objects in the solution float to their neutral buoyancy point.
- The bulk of human DNA forms a band at a density of 1.55.
- However, short tandem repeats have a slightly different density because they don't have the same base composition as bulk DNA--they form density "satellites" in the centrifuge tube, bands of slightly different density above or below the main DNA band.
- Three density satellites for human DNA: I, II, and III. Found in centromere regions on all chromosomes.
- An example: "alpha" (or "alphoid") sequence is 171 bp repeat found at all centromeres in many copies. It apparently binds the kinetochore proteins (which anchor the spindle fibers). Lots of variation between chromosomes, and the variants seem to evolve rapidly.



Mini- and Microsatellite DNA

- Minisatellites are other short repeats, mostly 10-30 bp long, mostly found in and near the telomeres. Some of them are “hypervariable”, meaning that the number of copies of the repeat varies greatly among individuals.
- This property makes them useful for DNA fingerprinting: getting a unique DNA profile for individuals using a single probe. These hypervariable minisatellites are also called “variable number tandem repeats” (VNTRs).
- Microsatellites (SSRs) are much shorter, 2-5 bp repeats, and microsatellite arrays are found all over the genome.

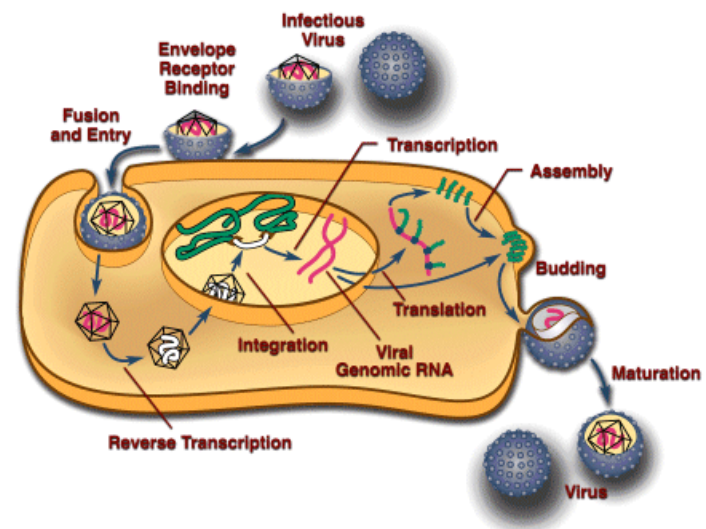
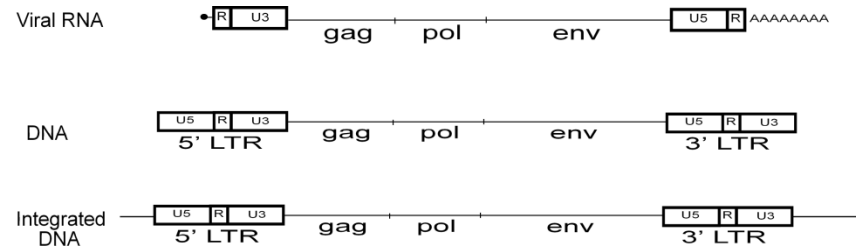


Moderately Repeated DNA

- Most of the moderately repeated DNA is derived from mobile DNA sequences (transposable elements, or transposons), which can move to new locations on occasion. This is sometimes called “selfish DNA”--subject to natural selection partly independent of the rest of the genome, it survives random mutational decay by replicating more frequently than other sequences, but not so frequently as to harm the individual.
- Two basic classes of transposon: RNA (retrotransposons) and DNA transposons.
- Retrotransposons replicate through an RNA intermediate: they are transcribed by RNA polymerase. The RNA intermediate is then reverse-transcribed back into DNA, which gets inserted at some random location in the genome.
 - Note that RNA transposons stay in place: a *copy* moves to a new location.
 - there are 3 important groups of retrotransposon: LINEs (long interspersed nuclear elements), SINEs (short interspersed nuclear elements), and LTR elements (LTR = long terminal repeat).
- DNA transposons move by cutting out the DNA sequence of the element and inserting it in a new location (usually).
- Another important distinction: autonomous transposons can move independently: they code for the enzymes necessary for transposition. Non-autonomous elements rely on enzymes produced by autonomous elements elsewhere in the genome.

Retroviruses

- The retrovirus genome is RNA. When it enters a cell, the RNA gets translated to form reverse transcriptase, which copies the viral RNA into DNA. This DNA then integrates into the genome: it becomes a provirus. The provirus DNA is transcribed to make more viral RNAs and proteins. The virus buds out through the cell membrane.
- Basic structure of retrovirus: 3 genes
 - gag: RNA-binding proteins (virus core)
 - pol: reverse transcriptase and other processing enzymes
 - env: outer coat protein
- LTR (long terminal repeat). The ends of the provirus are exact copies of each other. The viral RNA only has one copy of the LTR, split into 2 sections at the ends of the RNA. These sections are duplicated during reverse transcription.
- The 5' LTR acts as a promoter for the provirus.
- The transcribed viral RNA is spliced in several different ways to produce messenger RNAs for the various retroviral proteins.



Retrotransposons

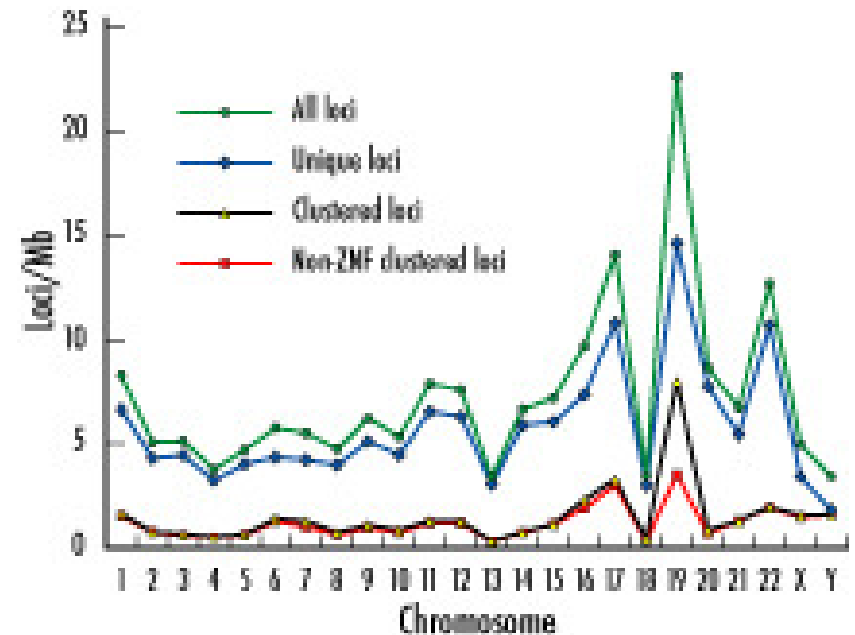
- LTR-containing retrotransposons are very similar to retroviruses. The difference is: retrotransposons lack the *env* gene, which produces the coat protein and allows movement outside of the cell. So, retrotransposons are strictly intracellular.
- In humans, LTR retrotransposons are also called endogenous retroviral sequences (ERV). Most copies are defective, with mutated or missing *gag* and *pol* genes. However, some are capable of transposition.
- LINEs (long interspersed nuclear elements) are autonomous transposable elements (or defective copies) that have a reverse transcriptase gene but don't have LTRs.
 - Promoter is within the 5' untranslated region: the promoter itself is transcribed into RNA.
 - Reverse transcription starts at the 3' end of the RNA, and often fails to reach the 5' end. So, defective copies are usually missing the 5' end. A full length active LINE1 (L1) element is 6.1 kb, but the average L1 element (including defective copies) is only 900 bp.
 - Three main human families: L1, L2, L3. Only L1 has active, autonomous copies.
 - The L1 reverse transcriptase also occasionally reverse-transcribes other RNAs in the cell.

More Transposons

- SINEs (short interspersed nuclear elements) are very small: 100-400 bp. They contain internal promoters for RNA polymerase 3. Several families, some originated as tRNA genes and others as 7SL RNA, the RNA involved in the signal recognition particle that guides secreted and membrane protein translation into the endoplasmic reticulum.
 - Most important SINE is the Alu sequence, which started as a 7SL RNA. Alu sequences make up 7% of genome, about 10^6 copies, about every 3 kb scattered throughout the genome. Can be used to clone or detect human DNA in mouse hybrid cells: there is nearly always an Alu sequence near any human gene (although not usually in the coding region: selection against mutant genes), but none are found in mouse DNA.
 - SINEs are transcribed by pol 3, but they need to be reverse-transcribed to re-integrate into the genome.
- DNA transposons are flanked by short inverted repeats (as opposed to LTRs, which are direct repeats).
 - They code for a transposase gene
 - Many families, mostly not active.
 - Unlike retrotransposons, DNA transposons usually excise themselves from the genome and re-insert themselves at a new location. However, sometimes they duplicate themselves.

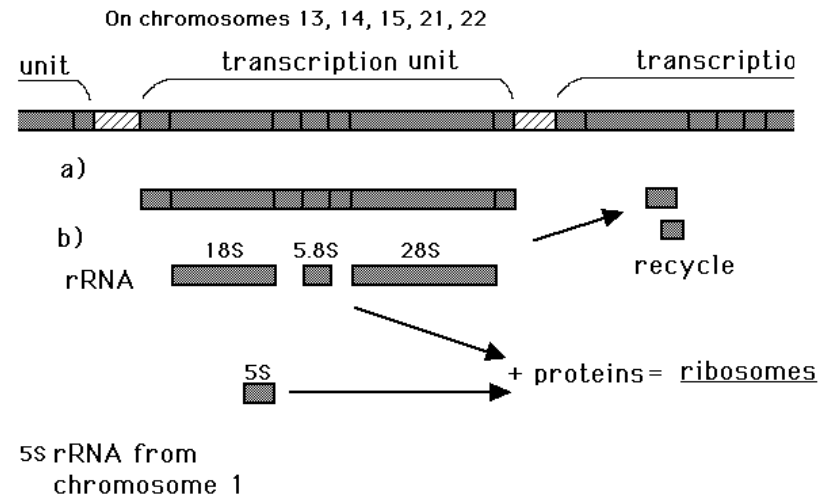
Genes

- Probably about 20,000 genes, not a particularly large number compared to other species.
- Gene density varies along the chromosomes: genes are mostly in euchromatin, not in the heterochromatin near the centromeres or on the short arms of acrocentric chromosomes.
- Most genes (90-95% probably) code for proteins. However, there are a significant number of RNA-only genes, and recent work has shown that RNA genes are far more important than previously thought.



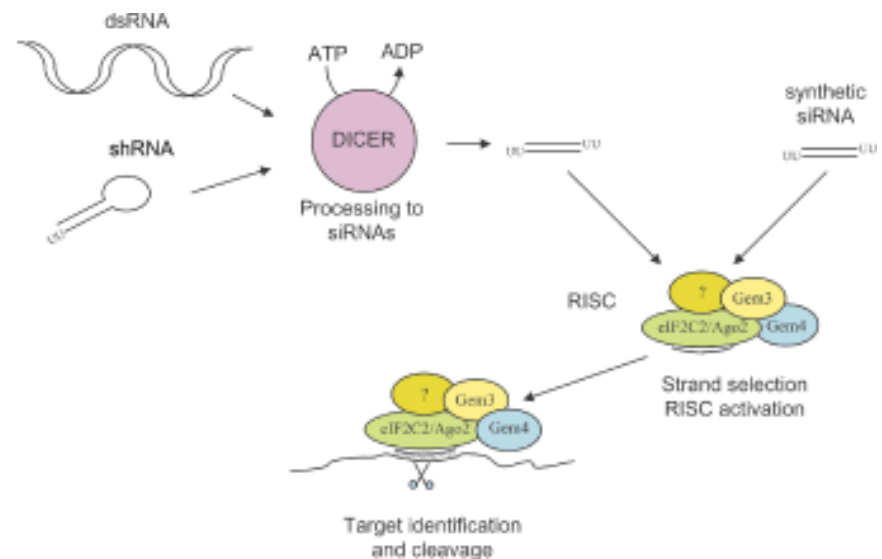
RNA Genes

- Protein-coding genes are transcribed by RNA polymerase 2 (pol2), while RNA genes are transcribed by pol1 or pol3.
- The best known RNA genes are ribosomal RNA and transfer RNA genes.
- Ribosomal RNA: 3 of the 4 rRNAs are transcribed from a single transcription unit. These transcription units form large arrays. The nucleolus sits on these genes, which are sometimes called nucleolus organizer regions. rRNA clusters exist on 5 of the chromosomes.
- The other ribosomal RNA, 5S RNA, is transcribed from large clusters elsewhere in the genome.
- Transfer RNA genes are dispersed throughout the genome, usually in small clusters. There are 49 families of tRNA genes: the third base of most codons is covered by one or two tRNAs: wobble.
 - selenocysteine, a 21st amino acid that contains selenium, is used in a few enzymes. Under certain conditions, a UGA stop codon is read by a special tRNA as selenocysteine.



Other RNA Genes

- Catalytic RNA molecules (ribozymes) are involved with RNA splicing and RNA base modification. The genes for these are small nuclear RNA (snRNA) and small nucleolar RNA (snoRNA) genes.
- Also genes for telomerase, signal recognition, X-chromosome inactivation, imprinting, and probably quite a bit else.
- MicroRNAs (miRNA) and small interfering RNAs (siRNA) regulate translation of specific mRNAs by binding to the mRNA: they are antisense RNAs, complementary to the “sense” strand of the mRNA.
 - miRNA seems to have a role in development. This is a very hot area of research at the moment.
 - siRNA is a basis for a popular technique called RNA interference, which allows specific genes to be inactivated.
- RNA interference starts with a double stranded RNA, which can either be artificially generated or the product of an siRNA gene that produces an RNA that folds into a hairpin loop.
 - An enzyme called “dicer” cuts out 20-25 bp regions of the double stranded RNA and combines them with proteins to form the RISC (RNA-induced silencing complex). RISC molecules then bind to mRNAs complementary to the siRNA and destroy them.



Protein-coding Genes

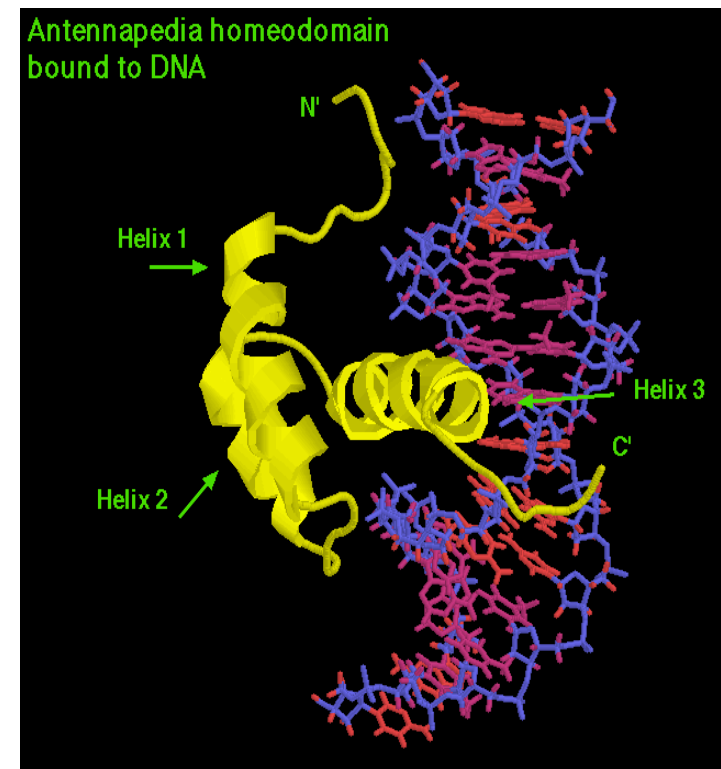
- Genes vary greatly in size and intron/exon organization.
- Some genes don't have any introns. Most common example is the histone genes. Histones are the proteins DNA gets wrapped around in the lowest unit of chromosomal organization, the nucleosome.
- Some genes are quite huge: dystrophin (associated with Duchenne muscular dystrophy) is 2.4 Mbp and takes 16 hours to transcribe. More than 99% of this gene is intron (total of 79 introns).
 - However, highly expressed genes usually have short introns
- Most exons are short: 200 bp on average. Intron size varies widely, from tens to millions of base pairs.

Gene Families

- Genes involved in the same biochemical pathway or functional unit are generally not clustered together. This also includes different subunits of the same protein: their genes are usually unlinked.
- However, genes that are related by having similar sequences (DNA sequence families) are very common. Possible causes:
 - conserved sequence domain or motif
 - segmental duplication/dispersed gene families
 - tandem duplication

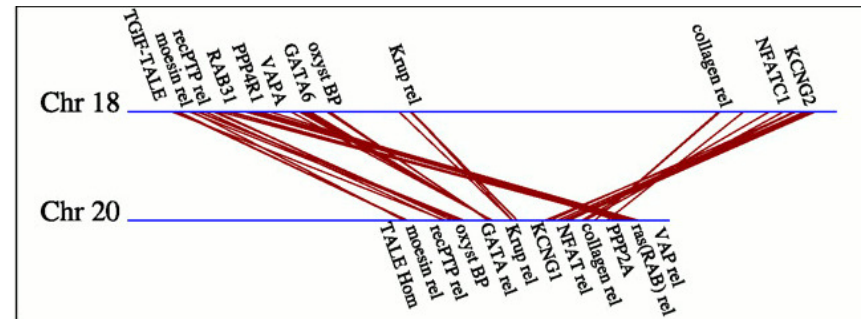
Conserved Domains and Motifs

- The appearance of truly novel functions is unusual. Most useful functions are re-used in many different proteins, which often show little sequence similarity with each other. This is the result of very ancient gene duplications and functional divergence, mostly long before we became human.
- Domain: a large region of amino acids on a protein that performs a specific function. A typical protein has one or a few domains. Often the three-dimensional structure of the protein shows the domains folded into separate units. The Hox proteins all share the homeobox domain, which is about 60 amino acids long. There is an ATP binding domain found in many proteins. Many examples, often found by X-ray crystallography.
- Motif: a short region of conserved amino acids that have a common function. E.g. the DEAD box (Asp-Glu-Ser-Asp) found in RNA helicases.



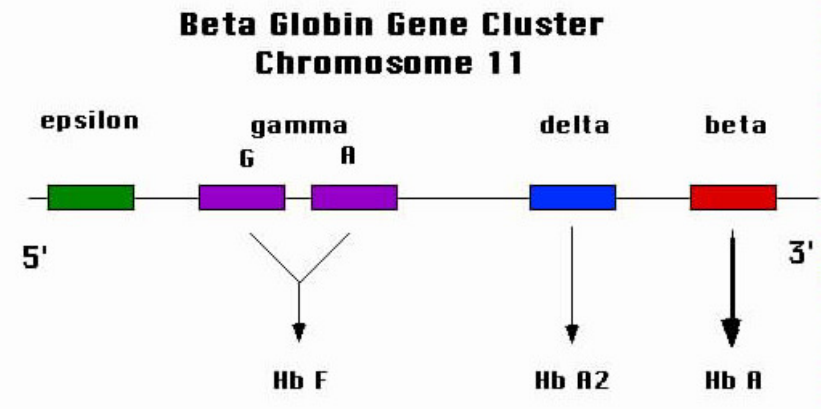
Segmental Duplications

- As much as 5% of the human genome consists of regions that have more than 90% identity with each other.
- Size range from 1 kb up to hundreds of kb.
- Most of them seem to have occurred since the divergence of the Great Apes from the monkeys, and about 1/3 of them have occurred since humans diverged from the chimpanzees.
- Can cause problems if crossing over occurs between duplicated regions on the same chromosome.
- Possibly catalyzed by transposable element movements: the ends of the duplicated regions are often transposable elements. Also, in plants many small segments of DNA are moved to new locations by DNA transposons.
- May be the cause of many dispersed gene families: very similar genes located far from each other, often on different chromosomes.
- Related phenomenon: pieces of the mitochondrial genome continue to invade the nucleus. Leads to multiple copies of genes that were originally in the mitochondria.

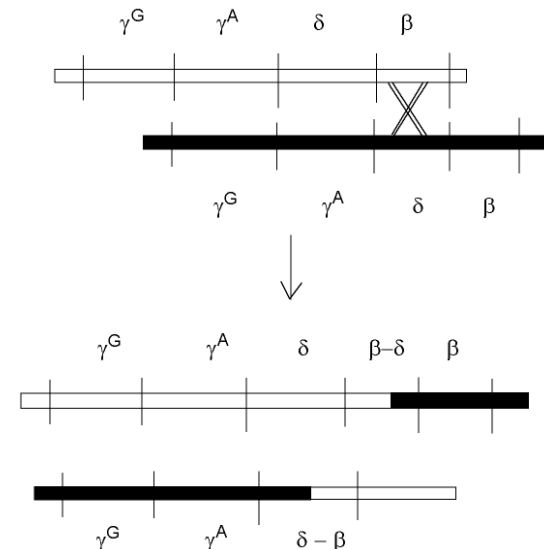


Tandem Duplications

- Many genes are found in small clusters of almost identical copies. The classic case is the beta-globin cluster, which contains 5 very similar genes. All play the “beta” role in hemoglobin molecules ($\alpha_2\beta_2$), but in different ways: beta is part of HbA, 99% of adult hemoglobin; delta is part of HbA2, 1% of adult hemoglobin; the two gamma genes (almost identical) are part of HbF, fetal hemoglobin; epsilon is part of embryonic hemoglobin.
- Sometimes a cluster of genes is regulated together (as in the beta globin genes. But usually the genes are completely independent of each other.
- The red-green color receptor genes on the X chromosome (cause of colorblindness) are another good example.
- Copy number changes through unequal crossing over during meiosis: the genes are so similar that the homologous recombination mechanism sometimes misaligns them, leading to increases or decreases in the number of copies in the array.
- New copies sometimes evolve new functions, but often they get inactivated by random mutations. This makes them pseudogenes, which quickly decay to undetectability.

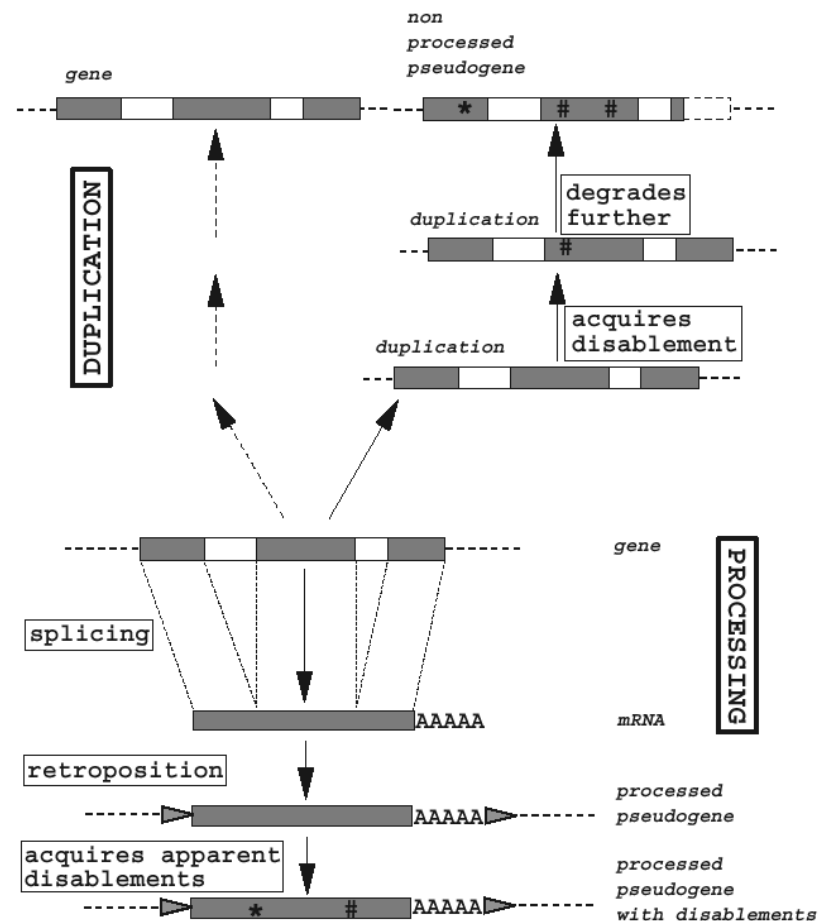


Unequal crossing over in the beta-globin region



Pseudogenes

- Pseudogenes are defective copies of genes. They contain most of the gene's sequence, but have stop codons or frameshifts in the middle, or they lack promoters, or are truncated or are just fragments of genes.
- Non-processed (duplicated) pseudogenes are the result of tandem gene duplication or transposable element movement. When a functional gene gets duplicated, one copy isn't necessary for life. Sometimes the copy will evolve a new function (as in the beta globin genes). Other times one copy will become inactivated by random mutation and become a pseudogene. Pseudogenes don't have a very long life span: once a region of DNA has no function it quickly picks up more mutations and eventually becomes unrecognizable.



Processed Pseudogenes

- Processed pseudogenes come from mRNA that has been reverse-transcribed and then randomly inserted into the genome. Processed pseudogenes lack introns because the mRNA was spliced. They also often have poly A tails and they lack promoters and other control regions.
 - Good example: the ribosomal proteins. There are 79 proteins encoded by 95 functional genes (a few duplications), but also 2090 processed pseudogenes,
 - Sometimes processed pseudogenes insert into a location that is transcribed. Leads to a new fusion protein or a intronless gene. These are sometimes called “expressed processed pseudogenes”. A whole group of them is expressed exclusively in the testes, with intron-containing homologues expressed in other tissues.
 - RNA genes are especially prone to becoming processed pseudogenes, because they often have internal promoters for pol3. That is, the retrotranscribed sequence contains its own promoter and doesn’t need to insert near another promoter. Alu sequences are an example of this: They are modified version of the signal recognition 7SL RNA .

Gene Oddities

- For the most part, genes are separated from each other by regions of non-conserved unique sequence DNA, which we believe is random junk being used as spacers. However, a few exceptions exist.
- Bidirectional gene organization. Cases where the 5' ends of two genes involved in the same functional unit are very close to each other. This probably results in common gene regulation. Several DNA repair genes are organized like this.
- Partially overlapping genes. The same DNA sequence used in two different reading frames for a few amino acids, usually transcribed from opposite DNA strands. This is very common in virus genomes (which need to be very compact). Causes problems because the overlap regions is intolerant of most mutations.
- Genes within genes. An intron contains another gene transcribed from the opposite strand. An example is the neurofibromatosis 1 (NF1) has a large intron that has 3 small genes (each of which has its own intron) within it.



Figure 1 Overlap between three human genes: *MUTH*, *FLJ13949*, and *TESK2*. Dark boxes represent coding sequence. Light boxes represent untranslated regions.

Courtesy of V. Veeramachaneni, 2004

