

A Brief Intro to the Human Genome and FASTA

Aaron Quinlan

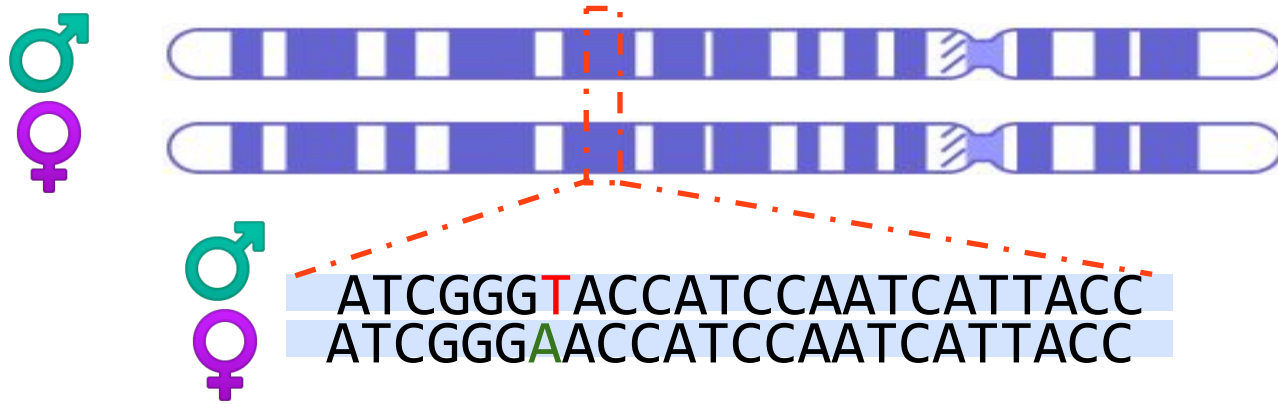
Departments of Human Genetics and Biomedical Informatics

USTAR Center for Genetic Discovery

University of Utah

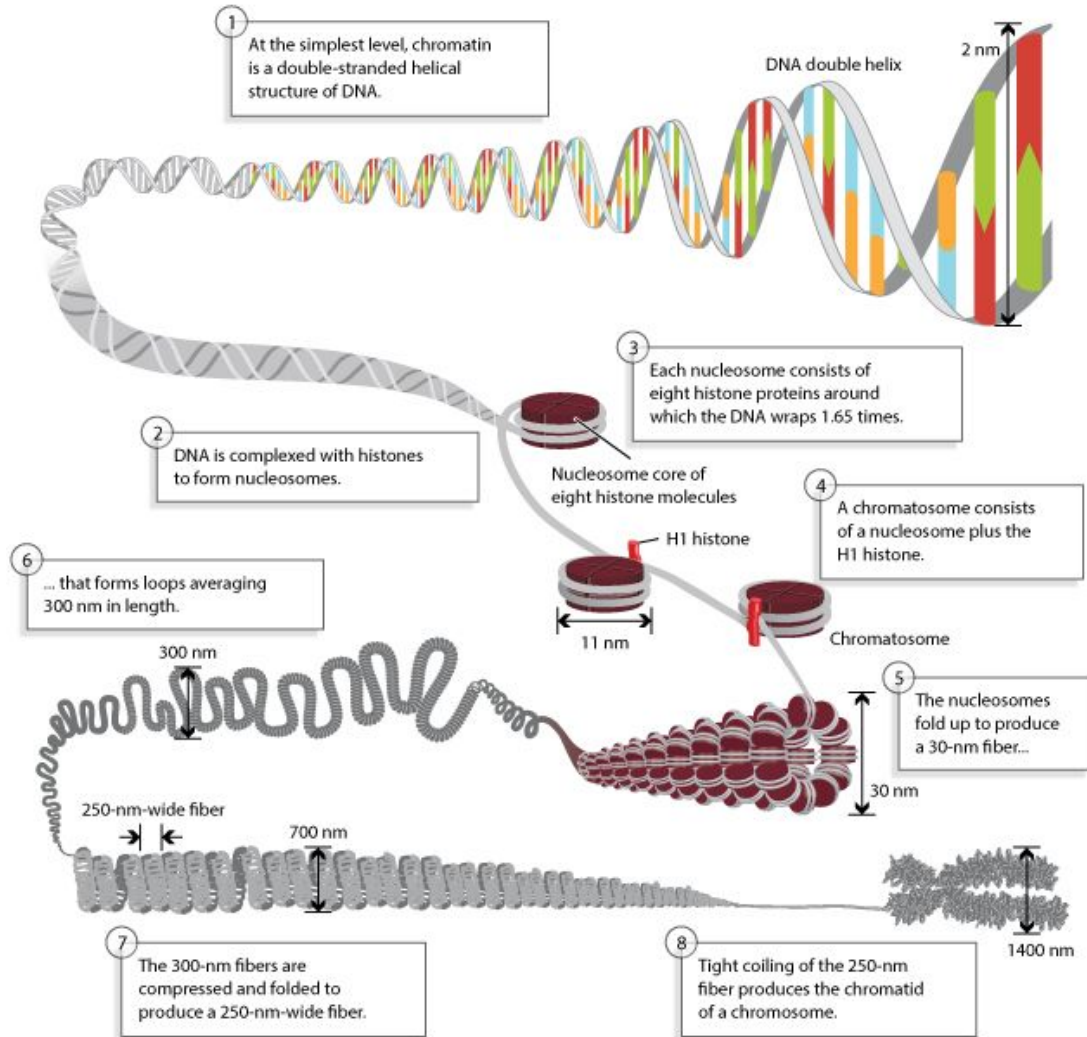
quinlanlab.org

Humans are diploid.



Our genome is comprised of a paternal and a maternal "haplotype". Together, they form our "genotype"

The human genome from a macro to micro scale



Our genome: mini quiz



How many *distinct* chromosomes in the nuclear human genome?

24: the autosomes (chromosome 1-22), sex chromosomes (X, Y)

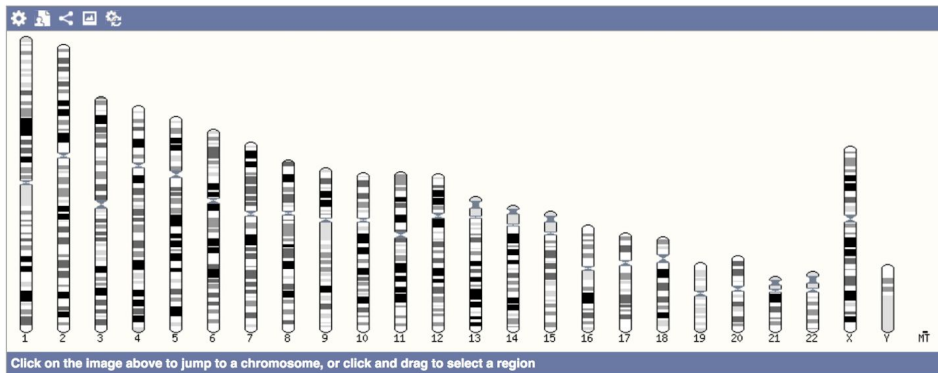
How many chromosomes exist in a (typical) haploid human genome ?

23: the autosomes (chromosome 1-22) and one sex chromosomes (X or Y)

How many chromosomes exist in a (typical) diploid human genome ?

46: two haploid genomes - one from mother and one from father

The human genome - basic stats



- 3.096 billion base pairs (haploid)
- 20,441 protein coding genes
- 198,002 coding transcripts
(isoforms of a gene that each encode a distinct protein product)

Summary

| | |
|--------------------------------|--|
| Assembly | GRCh38.p7 (Genome Reference Consortium Human Build 38), INSDC Assembly GCA_000001405.22 , Dec 2013 |
| Database version | 87.38 |
| Base Pairs | 3,547,762,741 |
| Golden Path Length | 3,096,649,726 |
| Genebuild by | Ensembl |
| Genebuild method | Full genebuild |
| Genebuild started | Jan 2014 |
| Genebuild released | Jul 2014 |
| Genebuild last updated/patched | Jun 2016 |
| Gencode version | GENCODE 25 |

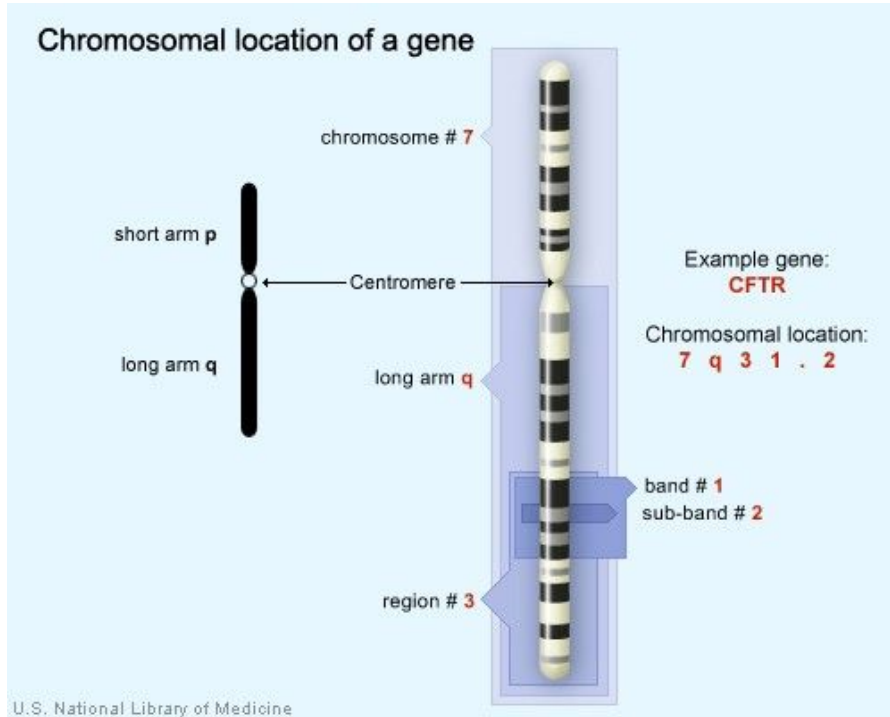
Gene counts (Primary assembly)

| | |
|------------------------|-------------------------------|
| Coding genes | 20,441 (incl 526 readthrough) |
| Non coding genes | 22,219 |
| Small non coding genes | 5,052 |
| Long non coding genes | 14,727 (incl 214 readthrough) |
| Misc non coding genes | 2,222 |
| Pseudogenes | 14,606 (incl 5 readthrough) |
| Gene transcripts | 198,002 |

http://uswest.ensembl.org/Homo_sapiens/Location/Genome



Chromosome Giemsa banding (G-banding)



- Heterochromatic regions, which tend to be rich with adenine and thymine (AT-rich) DNA and relatively gene-poor, **stain more darkly** with Giemsa and result in G-banding
- Less condensed ("open") chromatin, which tends to be (GC-rich) and more transcriptionally active, incorporates less Giemsa stain, resulting in **light bands in G-banding**.
- Cytogenetic bands are labeled p1, p2, p3, q1, q2, q3, etc., **counting from the centromere out toward the telomeres**. At higher resolutions, sub-bands can be seen within the bands.
- For example, the locus for the CFTR (cystic fibrosis) gene is **7q31.2**, which indicates it is on **chromosome 7, q arm, region 3, band 1, and sub-band 2**. (Say 7,q,3,1 dot 2)

A first map of the human genome

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.

A first map of the human genome ("build 1")

Table 8 Chromosome size estimates

Table 8

1 of 4

| Chromosome* | Sequenced bases† (Mb) | FCC gaps‡ | | SCC gaps§ | | Sequence gaps# | | Heterochromatin and short arm adjustments**(Mb) | Total estimated chromosome size (including chromosome artefactual duplication in draft genome sequence)†† (Mb) | Previously estimated chromosome size‡‡ (Mb) |
|-------------|-----------------------|-----------|---------------------------|-----------|---------------------------|----------------|---------------------------|---|--|---|
| | | Number | Total bases in gaps§ (Mb) | Number | Total bases in gaps¶ (Mb) | Number | Total bases in gaps* (Mb) | | | |
| All | 2,692.9 | 897 | 152.0 | 4,076 | 142.7 | 145,514 | 80.6 | 212 | 3,289 | 3,286 |
| 1 | 212.2 | 104 | 17.7 | 347 | 12.1 | 11,803 | 6.5 | 30 | 279 | 263 |
| 2 | 221.6 | 50 | 8.5 | 296 | 10.4 | 12,880 | 7.1 | 3 | 251 | 255 |
| 3 | 186.2 | 71 | 12.1 | 336 | 11.8 | 14,689 | 8.1 | 3 | 221 | 214 |
| 4 | 168.1 | 39 | 6.6 | 343 | 12.0 | 12,768 | 7.1 | 3 | 197 | 203 |
| 5 | 169.7 | 46 | 7.8 | 337 | 11.8 | 10,304 | 5.7 | 3 | 198 | 194 |
| 6 | 158.1 | 15 | 2.6 | 275 | 9.6 | 5,225 | 2.9 | 3 | 176 | 183 |
| 7 | 146.2 | 27 | 4.6 | 195 | 6.8 | 4,338 | 2.4 | 3 | 163 | 171 |
| 8 | 124.3 | 41 | 7.0 | 249 | 8.7 | 8,692 | 4.8 | 3 | 148 | 155 |
| 9 | 106.9 | 19 | 3.2 | 122 | 4.3 | 6,083 | 3.4 | 22 | 140 | 145 |
| 10 | 127.1 | 14 | 2.4 | 163 | 5.7 | 8,947 | 5.0 | 3 | 143 | 144 |
| 11 | 128.6 | 29 | 4.9 | 193 | 6.8 | 8,279 | 4.6 | 3 | 148 | 144 |
| 12 | 124.5 | 26 | 4.4 | 168 | 5.9 | 8,226 | 4.6 | 3 | 142 | 143 |
| 13 | 92.9 | 12 | 2.0 | 115 | 4.0 | 5,065 | 2.8 | 16 | 118 | 114 |
| 14 | 86.9 | 13 | 2.2 | 40 | 1.4 | 775 | 0.4 | 16 | 107 | 109 |
| 15 | 73.4 | 18 | 3.1 | 104 | 3.6 | 5,717 | 3.2 | 17 | 100 | 106 |
| 16 | 73.1 | 55 | 9.4 | 102 | 3.6 | 4,757 | 2.6 | 15 | 104 | 98 |
| 17 | 72.8 | 41 | 7.0 | 95 | 3.3 | 4,261 | 2.4 | 3 | 88 | 92 |
| 18 | 72.9 | 22 | 3.7 | 113 | 4.0 | 4,324 | 2.4 | 3 | 86 | 85 |
| 19 | 55.4 | 49 | 8.3 | 108 | 3.8 | 2,344 | 1.3 | 3 | 72 | 67 |
| 20 | 60.5 | 7 | 1.2 | 33 | 1.2 | 469 | 0.3 | 3 | 66 | 72 |
| 21 | 33.8 | 4 | 0.1 | 0 | 0.0 | 0 | 0.0 | 11 | 45 | 50 |
| 22 | 33.8 | 10 | 1.0 | 0 | 0.0 | 0 | 0.0 | 13 | 48 | 56 |
| X | 127.7 | 141 | 24.0 | 182 | 6.4 | 4,282 | 2.4 | 3 | 163 | 164 |
| Y | 21.8 | 6 | 1.0 | 19 | 0.7 | 113 | 0.1 | 27 | 51 | 59 |
| NA | 5.1 | 0 | 0 | 134 | 0.0 | 577 | 0.3 | 0 | 0 | 0 |
| UL | 9.3 | 38 | 0 | 7 | 0.0 | 566 | 0.3 | 0 | 0 | 0 |

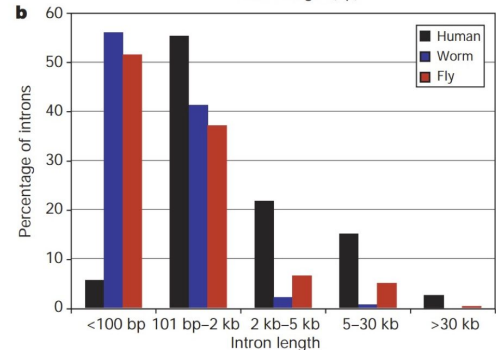
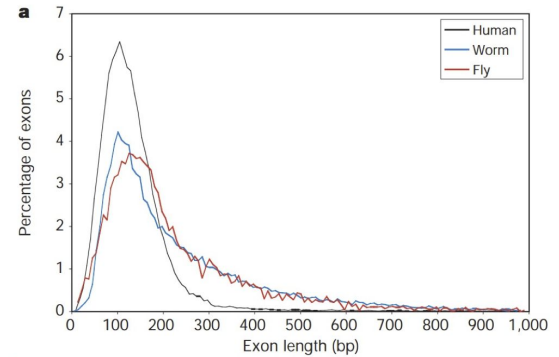
ACCACAACCAGGCATAGGGGAAAGATTGGAGGAAAGATGAGTGAGAGCATCAACTTCTCTCACAACTAGGCCAGTAAGTAGTGCTTGTGCT
ATCTCCTTGGCTGTGATACGTGGCCGGCCCTCGCTCCAGCAGCTGGACCCCTACCTGCCGTCTGCTGCCATCGGAGCCCAAAGCCGGGCTGT
ACTGCTCAGACCAGCCGGCTGGAGGGAGGGGCTCAGCAGGTCTGGCTTTGGCCCTGGGAGAGCAGGTGGAAGATCAGGCAGGCCATCGCTGC
ACAGAACCAGTGGATTGGCCTAGGTGGGATCTCTGAGCTCAACAAGCCCTCTCTGGGTGGTAGGTGCAGAGACGGGAGGGGCAGAGCCGCA
GCACAGCCAAGAGGGCTGAAGAAATGGTAGAACGGAGCAGCTGGTGATGTGTGGGCCACCGGCCCCAGGCTCCTGTCTCCCCCAGGTGTG
GGTGATGCCAGGCATGCCCTTCCCAGCATCAGGTCTCCAGAGCTGCAGAAGACGACGGCCGACTTGGATCACACTCTTGTGAGTGTCCCA
TGTTGCAGAGGTGAGAGGAGAGTAGACAGTGAGTGGGAGTGGCGTCGCCCTAGGGCTCTACGGGGCCGGCGTCTCCTGTCTCCTGGAGAGG
TTCGATGCCCCCTCCACACCCTCTTGATCTTCCCTGTGATGTCATCTGGAGCCCTGCTGCTTGGGTGGCCTATAAAGCCTCCTAGTCTGGCT
CAAGGCCTGGCAGAGTCTTCCAGGGAAAGCTACAGCAGCAAACAGTCTGCATGGGTGCATCCCCTTCACTCCAGCTCAGAGCCAGGCC
GGGGCCCCAAGAAAGGCTCTGGTGGAGAACCTGTGCATGAAGGCTGTCAACCAGTCCATAGGCAAGCCTGGCTGCCTCCAGCTGGGTGCAG
GACAGGGGCTGGAGAAGGGGAGAAGAGGAAAGTGAGGTTGCCCTGCCCTGTCTCCTACCTGAGGCTGAGGAAGGAGAAGGGGATGCACTGTTG
GGAGGCAGCTGTAACTCAAAGCCTTAGCCTCTGTTCCACGAAGGCAGGGCCATCAGGCACCAAAGGGATTCTGCCAGCATAGTGCTCCTGG
CCAGTGATACACCCGGCACCCCTGTCTGGACACGCTGTTGGCCTGGATCTGAGCCCTGGTGGAGGTCAAAGCCACCTTTGGTTCTGCCATTG
TGCTGTGTGGAAGTTCACCTCCTGCCTTTTCCCTTAGAGCCTCCACCACCCGAGATCACATTTCTCACTGCCTTTTGTCTGCCAGTT
CACCAGAAGTAGGCCTCTTCCCTGACAGGCAGCTGCACCACTGCCTGGCGCTGTGCCCTTCCCTTGTCTGCCCGCTGGAGACGGTGTTTGTC
TGGGCCTGGTCTGCAGGGATCCTGCTACAAAGGTGAAACCCAGGAGAGTGTGGAGTCCAGAGTGTGCCAGGACCCAGGCACAGGCATTAGT
CCCGTTGGAGAAAACAGGGGAATCCCGAAGAAATGGTGGGTCTGGCCATCCGTGAGATCTTCCAGGTGTGCCGTTTTCTCTGGAAGCCTC
TAAGAACACAGTGGCGCAGGCTGGGTGGAGCCGTCCCCCATGGAGCACAGGCAGACAGAAGTCCCCGCCCCAGCTGTGTGGCCTCAAGCCA
CCTTCCGCTCCTTGAAGCTGGTCTCCACACAGTGCTGGTTCCGTCACCCCCCTCCAAGGAAGTAGGTCTGAGCAGCTTGTCTGGCTGTGTC
ATGTCAGAGCAACGGCCAAAGTCTGGGTCTGGGGGGGAAGGTGTCATGGAGCCCCCTACGATTCCAGTCGTCTCGTCTCCTCTGCCTGT
GCTGCTGCCGTGGCGGCAGAGGAGGGATGGAGTCTGACACGCGGGCAAAGGCTCCTCCGGGCCCTCACCAGCCCCAGGTCCTTTCCAGAG
TGCCTGGAGGGAAAAGGCTGAGTGAGGGTGGTTGGTGGGAAACCCTGGTTCCCCCAGCCCCCGGAGACTTAAATACAGGAAGAAAAAGGCAG
ACAGAATTACAAGGTGCTGGCCAGGGCGGGCAGCGGCCCTGCCTCCTACCCTTGCGCCTCATGACCGGAGCCATAGCCAGGCAGGAGGGC
GAGGACCTCTGGTGGCGGCCAGGGCTTCCAGCATGTGCCCTAGGGGAAGCAGGGGCCAGCTGGCAAGAGCAGGGGGTGGGCAGAAAGCACC
GGTGGACTCAGGGCTGGAGGGGAGGAGCGATCTTGCCCAAGGCCCTCCGACTGCAAGCTCCAGGGCCCCTCACCTTGTCTCCTGCTCCTTC

Gene content

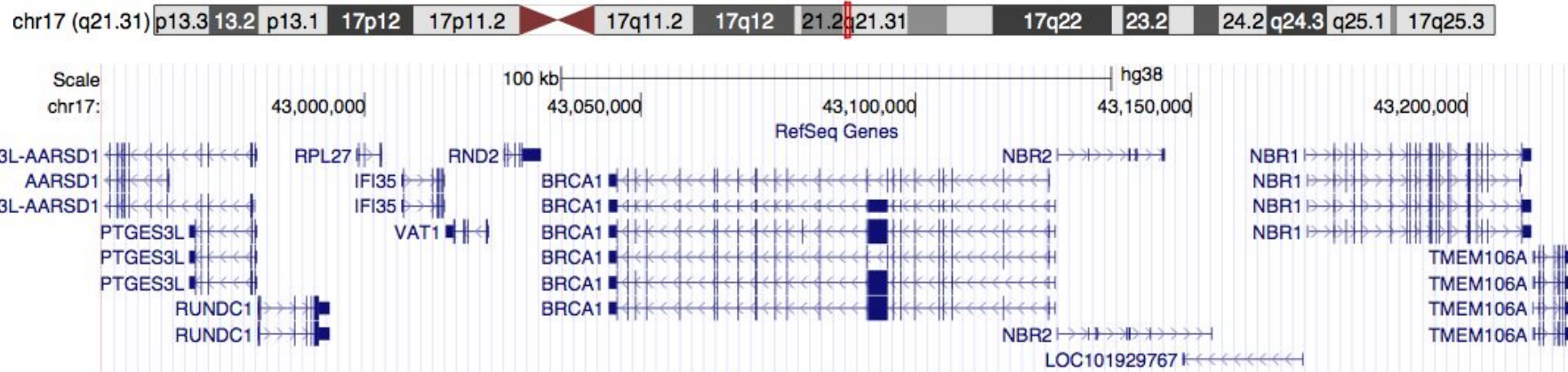
"There appear to be about 30,000-40,000 protein-coding genes in the human genome -- only about twice as many as in worm or fly. However, the genes are more complex, with more alternative splicing generating a larger number of protein products." (Over time this has evolved to an estimate of approximately 20,000 protein coding genes, which reflects roughly the number of genes in fly and worm)

Table 21 Characteristics of human genes

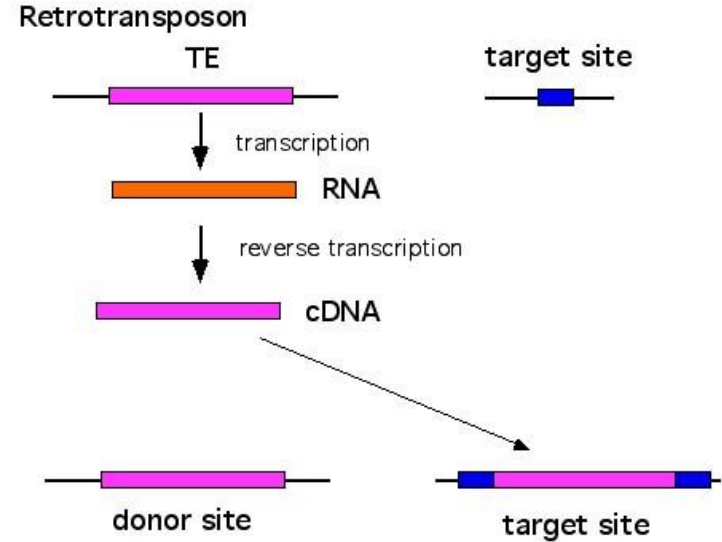
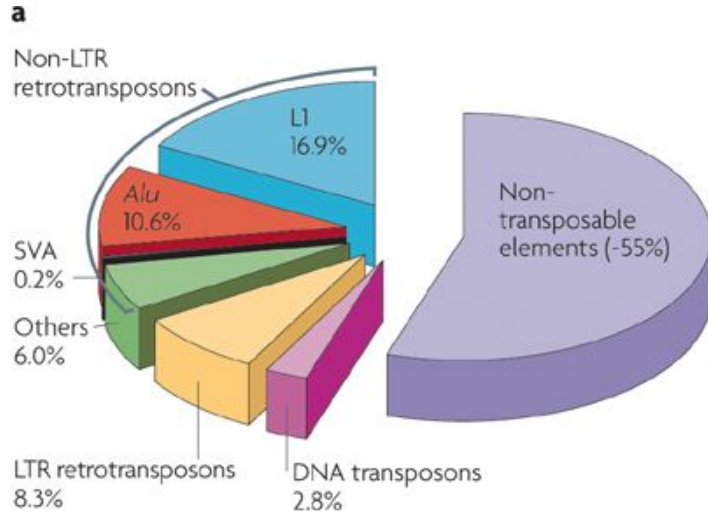
| | Median | Mean |
|-----------------------|----------|----------|
| Internal exon | 122 bp | 145 bp |
| Exon number | 7 | 8.8 |
| Introns | 1,023 bp | 3,365 bp |
| 3' UTR | 400 bp | 770 bp |
| 5' UTR | 240 bp | 300 bp |
| Coding sequence (CDS) | 1,100 bp | 1,340 bp |
| | 367 aa | 447 aa |
| Genomic extent | 14 kb | 27 kb |



Solely 2% of the human genome encodes proteins.



Half of the human genome is comprised of repeats



McClintock's
"jumping
genes" in maize

Retrotransposons use a "copy/paste" mechanism
DNA transposons use a "cut/paste" mechanism

Half of the human genome is comprised of repeats



The human reference genome continues to change.

- Ongoing efforts to fill "gaps" and properly/thoroughly represent complex structures and loci in the genome (e.g., Major Histocompatibility Complex)
- Each improvement leads to a new genome "build". Currently on build 38.
- Experimental and computational methods provide new genome annotations
 - New gene models, transcription factor binding sites, and loci where human individuals differ (i.e., polymorphisms)
- Therefore, the human reference genome is by no means "complete"!
- How does the same genome yield such phenotypic diversity across tissue types?
- How does the genome evolve within an individual (tissues) and among a population?

Searching for and counting patterns in genomes with `grep`

`~/workspace/dnaseq/references/all_sequences.fa`

What will this command do?

```
grep ">" ~/workspace/dnaseq/references/all_sequences.fa
```

How many adenosines are there?

```
grep -v ">"
```

```
~/workspace/dnaseq/references/all_sequences.fa | grep
```

```
-c "A"
```