

Genome Assembly and Annotation

Erik Arner
Omics Science Center, RIKEN
Yokohama, Japan
arner@gsc.riken.jp

NCBI Map Viewer

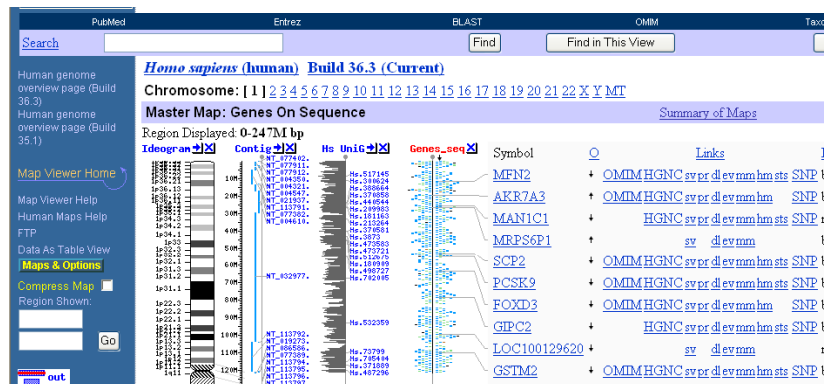
Map Viewer
Map Viewer Home
Map Viewer Help
Human Maps Help
Release Notes

NCBI Resources
Genome Project
TaxPlot
Consensus CoDing Sequence (CCDS)
Human Genome Resources

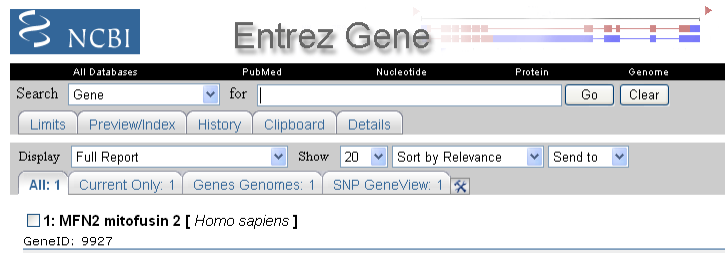
Homo sapiens (human) genome view
Build 36.3 statistics [Switch to previous build](#)

1 2 3 4 5 6 7 8 9 10 11 12 13
14 15 16 17 18 19 20 21 22 X Y III

NCBI Map Viewer



Entrez Gene



Entrez Gene

1: MFN2 mitofusin 2 [*Homo sapiens*]

GeneID: 9927

updated 25-Apr-2008

Summary

Official Symbol	MFN2	provided by HGNC
Official Full Name	mitofusin 2	provided by HGNC
Locus tag	RP5-107789.3	
See related	Ensembl: ENSG00000116688 ; HPRD:08495 ; MIM:608507	
Gene type	protein coding	
RefSeq status	Reviewed	
Organism	Homo sapiens	
Lineage	<i>Eukaryota</i> ; <i>Metazoa</i> ; <i>Chordata</i> ; <i>Craniata</i> ; <i>Vertebrata</i> ; <i>Euteleostomi</i> ; <i>Mammalia</i> ; <i>Etheria</i> ; <i>Euarchontoglires</i> ; <i>Primates</i> ; <i>Haplorrhini</i> ; <i>Catarrhini</i> ; <i>Hominidae</i> ; <i>Homo</i>	
Also known as	HSG; MARF; CMT2A; CPRP1; CMT2A2; KIAA0214	
Summary	This gene encodes a mitochondrial membrane protein that participates in mitochondrial fusion and contributes to the maintenance and operation of the mitochondrial network. This protein may play a role in the pathophysiology of obesity.	

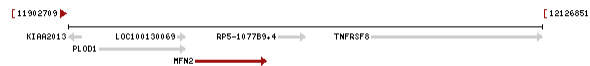
Entrez Gene

Genomic regions, transcripts, and products

Go to [reference sequence details](#)

Genomic context

chromosome: 1; Location: 1p36.22



Entrez Gene

Related Sequences

Nucleotide	Protein
Genomic AL096840.25	CAI19086.1 CAI19087.2 CAI19088.2
Genomic AL096840.25	CAI19086.1 CAI19087.2 CAI19088.2
Genomic CH471130.1	EAW71726.1 EAW71727.1
mRNA AF036536.2	AAD02058.2
mRNA AK001189.1	None
mRNA AK021947.1	None
mRNA AK022453.1	None
mRNA AK289828.1	BAF82517.1
mRNA AL122114.1	None
mRNA AL137666.1	CAB70866.2
mRNA AY028429.1	AAK18728.1
mRNA BC017061.1	AAH17061.1
mRNA BU178912.1	None
mRNA D86987.1	BAA34389.2

Aims of lecture

- How was the human genome put together
 - Limitations
- What is in the genome and where is it?
 - Genes, SNPs etc
- Databases
 - GenBank → RefSeq → Entrez Gene
- This lecture is focused on NCBI resources
 - UCSC, Ensembl etc use similar methods
- Genomics ↔ bioinformatics

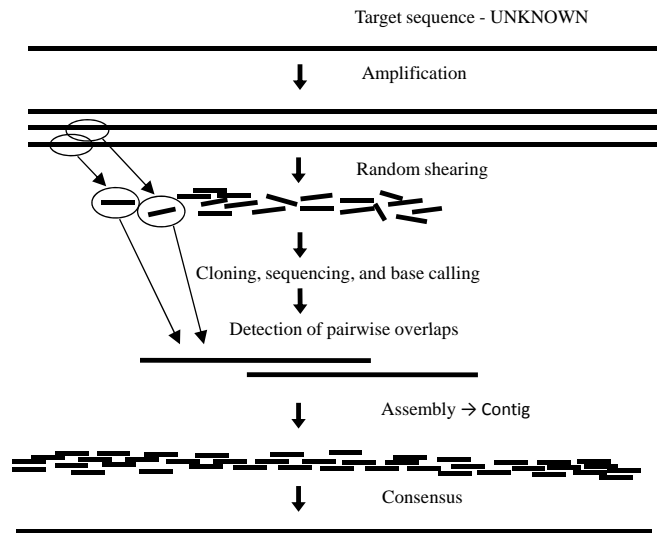
Outline

- Genome assembly
 - Basics of shotgun sequencing and assembly
 - Clone-by-clone
 - Whole Genome Shotgun
 - Current status of human reference genome
- Genome annotation
 - Placing genes on the genome
 - Known mRNAs
 - Predicted genes
 - Other features
 - SNPs

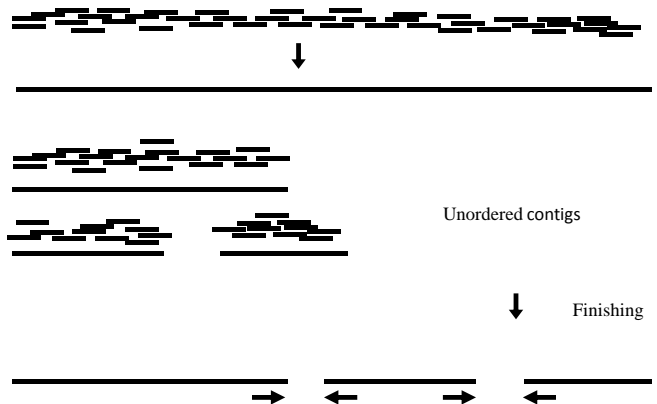
Basics of shotgun sequencing

- Problem: with current sequencing technology, the longest continuous sequence that can be reliably determined is around 800 – 1000 nt
 - Human genome contains ~ 3 billion bases
 - 50 – 250 Mb chromosomes
 - Novel technologies will produce longer sequences (several kb), however: we will not be able to sequence millions of bases anytime soon
- Solution: shotgun sequencing
 - Sanger 1980, still state of the art
 - Divide and conquer
 - Puzzle

Basics of shotgun sequencing



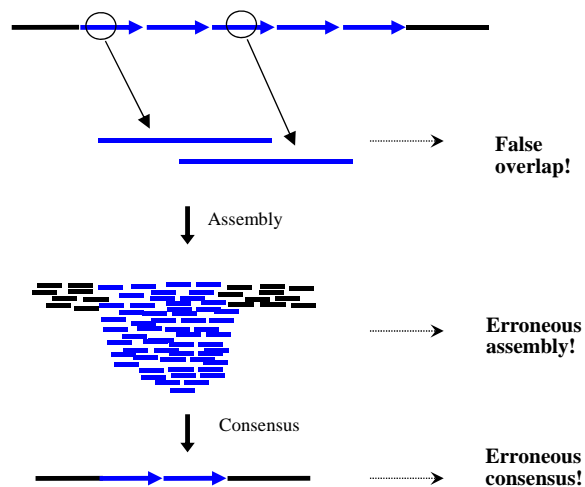
Basics of shotgun sequencing



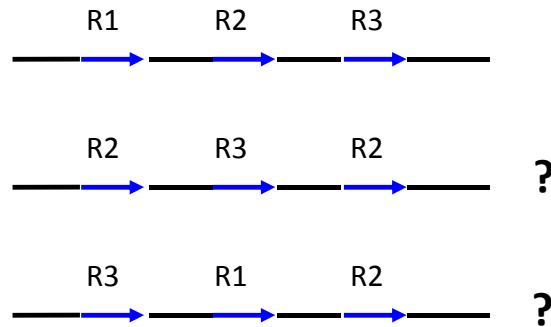
Basics of shotgun sequencing

- Main problems
 - Contamination
 - Vector sequences
 - Bacterial genome (from amplifying clones in bacteria)
 - Other projects being sequenced
 - Polymorphisms
 - Haplotype expansion (two alleles represented as separate loci)
 - Missing allele (unplaced contig)
 - Repeats
 - Component mis-assemblies
 - Path level mis-assemblies
 - Non-overlapping contigs appear to overlap

Basics of shotgun sequencing




Basics of shotgun sequencing



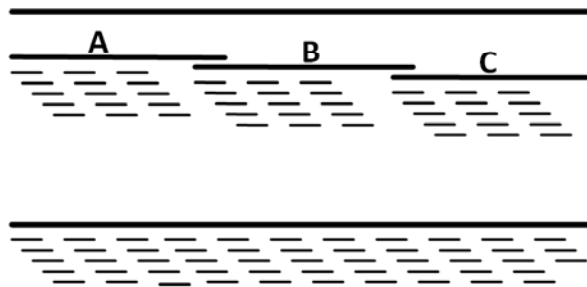
Basics of shotgun sequencing

- Finishing
 - Major part of sequencing project
 - Finishing of human genome still ongoing
 - 50% of time and money spent on finishing
 - Repeats major reason for long and expensive finishing process

Basics of shotgun sequencing

- Two major strategies for shotgun sequencing of whole genomes
 - Clone-by-clone 
 - Hierarchical
 - Minimal set of overlapping clones selected for sequencing
 - Overlapping segments sequenced separately, then merged
 - Whole Genome Shotgun (WGS)
 - Everything at once

Basics of shotgun sequencing



Basics of shotgun sequencing

- Clone-by-clone
 - Advantage: problems stay local
 - Disadvantage: difficult to coordinate, lots of pre-processing and infrastructure needed
- WGS
 - Advantage: pre-processing step eliminated
 - No clone mapping, coordination etc
 - Disadvantage: problems go global
- Public genome initiative used clone-by-clone strategy, private used WGS
- Merits heavily debated, combination might be preferable

Current status of human genome

- Essentially finished ...or?
 - 234 gaps remain in euchromatic part of genome
 - 17 million bases (0.5%)
 - Centromeres and telomeres not sequenced
 - Will require new technology
 - ~ 45% of the human genome consists of repeats interspersed with non-repetitive sequences
 - Transposon derived (LINEs, SINEs)
 - 3 – 4% segmental duplications (> 1kb, > 90% similar), ~ 40% believed to be misassembled
 - Multi-gene families
 - Large number of anonymous donors for HGP, but much of the DNA comes from one individual
 - Individual structural variants
 - Duplications, deletions, inversions, translocations...

NCBI Map Viewer

Map Viewer

Map Viewer Home

Map Viewer Help

Human Maps Help

Release Notes

NCBI Resources

Genome Project

TaxPlot

Consensus Coding Sequence (CCDS)

Human Genome Resources

Homo sapiens (human) genome view
[Build 36.3 statistics](#) [Switch to previous build](#)

NCBI Map Viewer

Homo sapiens (human) Build 36.3 (Current)

Chromosome: [1] 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X Y MT

Master Map: Genes On Sequence

Region Displayed: 0-247M bp

Gene Symbol	Links
MFN2	OMIM HGNC sv pr dl ev mm hm sts SNP
AKR7A3	OMIM HGNC sv pr dl ev mm hm sts SNP
MAN1C1	HGNC sv pr dl ev mm hm sts SNP
MRPS6P1	sv dl ev mm
SCP2	OMIM HGNC sv pr dl ev mm hm sts SNP
PCSK9	OMIM HGNC sv pr dl ev mm hm sts SNP
FOXD3	OMIM HGNC sv pr dl ev mm hm sts SNP
GIPC2	HGNC sv pr dl ev mm hm sts SNP
LOC100129620	sv dl ev mm
GSTM2	OMIM HGNC sv pr dl ev mm hm sts SNP

Genome annotation

- Placing the genes
 - Aligning RefSeqs
 - Prediction
 - Using additional transcript information (GenBank)
 - Ab initio
 - Gene assignment
- SNPs
- MUCH more not covered in this lecture
 - Predictions using bioinformatics
 - Alignment of other types of data

Detour: GenBank vs RefSeq

- GenBank:
 - Archive of primary sequence data
 - mRNA, genomic DNA (may include genes), non-coding RNA (rRNA, tRNA, miRNA...)
 - Not curated (basic quality checks)
 - Sequencing performed by submitter
 - Only submitter can update records (annotation etc)
 - Multiple entries for same locus
 - Data exchanged with EBI/EMBL and DDBJ
 - Data from more than 260 000 species (Jan 2008)

Detour: GenBank vs RefSeq

- RefSeq:
 - “GenBank de luxe”
 - Curated collection of DNA, RNA, and protein sequences
 - Built by NCBI
 - Only “major”/“model” organisms for which sufficient data is available
 - More than 5000 organisms (March 2008)
 - One example of each natural biological molecule
 - One good example from GenBank
 - Combination of GenBank entries
 - Imported from other curated databases
 - SGD (Yeast), FlyBase (Drosophila)

Detour: GenBank vs RefSeq

GenBank	RefSeq
Not curated	Curated
Author submits	NCBI creates from existing data
Only author can revise	NCBI revises as new data emerge
Multiple records for same loci common, records can contradict each other	Single records for each molecule of major organisms
No limit to species included	Limited to model organisms
Data exchanged among INSDC members	Exclusive NCBI database
Akin to primary literature	Akin to review articles

Genome annotation

- Aligning RefSeqs to genome
 - mRNAs, non-coding RNAs
 - SPLIGN
 - BLAST + dynamic programming
 - Heuristics (e.g. splice signals) included
 - Attempts to use as little *a priori* information as possible (codon usage, polyA signals)
 - Alternate RefSeq derived models sharing one or more exons on same strand are grouped under the same gene
 - Requirements for gene annotation
 - Defining RefSeq transcript alignment is $\geq 95\%$ identity
 - Aligned region covers $\geq 50\%$ of the length, or at least 1000 bases
 - RefSeqs aligning to multiple genomic locations
 - Best alignment selected and annotated
 - If they are of equal quality, both are annotated
 - Arbitrary parameters!
 - Gene **MODELS!**

Genome annotation

- Prediction
 - Gnomon
 - HMM based on transcript and protein alignment data, heuristics
 - Goal: to ensure that if a biological expert is presented with the same data, they could not produce an obviously improved gene model

Genome annotation

- Gnomon
 - Additional GenBank mRNAs and ESTs aligned to genome using SPLIGN
 - Combined with RefSeq alignments → clusters
 - Additional information used (e.g. codon usage)
 - BLAST (translated nucleotide against protein db) hits incorporated, including hits in other organisms
 - Frameshifts/stop codons → pseudogene candidates
 - HMM generated, used for ab initio prediction
 - Ab initio models blasted against protein db, HMM updated iteratively
 - Resulting models included in annotation
 - Only models not overlapping with RefSeq based models

Genome annotation

- SNPs
 - dbSNP
 - Mapped to genome using BLAST
 - High/low confidence
 - High: 95% of the flanking sequence aligns with 0-6 mismatches
 - Low: 75% of the flanking sequence aligns with < 3% mismatches
 - SNPs with ambiguous map positions are annotated with a warning

Genome annotation

- More stuff (not covered in this lecture)
 - Predictions using bioinformatics
 - miRNA target sites
 - TF (activator/repressor) binding sites
 - Other regulatory elements (insulators, enhancers etc)
 - Repeats
 - Alignment of other types of data
 - Regions defined by ChIP-chip/ChIP-seq
 - Transcription Start Sites (CAGE)
 - Homology to other species

Genome annotation

- Challenges
 - False positives
 - False negatives
 - Errors
 - Genes
 - Exons
 - UTRs
 - UTRs and non-coding genes hard to predict ab initio

Genome assembly and annotation summary

- Handle with care!
- Information mostly reliable
 - but be aware of limitations
- REFERENCE genome
- Gene MODELS
- Genome and annotation not static, they are subject to change