Introduction

^^^^^^^^^^^^

The Dec. 2013 assembly of the human genome (GRCh38 Genome Reference Consortium

Human Reference 38), is called hg38 at UCSC. This directory contains the genome

as released by UCSC, selected annotation files and updates. The directory

"genes/" contains GTF/GFF files for the main gene transcript sets.

For more information about this assembly, see these NCBI resources:

http://www.ncbi.nlm.nih.gov/genome/51

http://www.ncbi.nlm.nih.gov/genome/assembly/883148

http://www.ncbi.nlm.nih.gov/bioproject/31257

These files are used by the UCSC Genome Browser for display and analysis. If you

want to do analysis and show it later on the browser, it is usually easiest to

run your analysis on the UCSC hg38 file. For most users, this will be the file

"latest/hg38.fa.gz" in this directory. However, if you need a genome file for

alignment or variant calling, please read the section "Analysis set" below.

The sequences of the main chromosomes are identical to the genome files distributed

by NCBI and the EBI, but the sequence names are different. For example, the

name of chromosome 1 is called "chr1" at UCSC, "NC\_000001.11" at NCBI, and "1"

at the EBI. Also, the lowercasing in the files is not exactly identical, as

UCSC, NCBI and EBI run Repeatmasker with slightly different settings.

The NCBI accession of the UCSC hg38 genome is GCA\_000001405.15. The version

that includes the updates for patch release 14 GRCh38.p14 has the NCBI

accession GCA\_000001405.29.

Analysis set

^^^^^^^^^^^^

The GRCh38 assembly contains more than just the chromosome sequences, but also

a mitochondrial genome, unplaced sequences, centromeric sequences

and alternates. To better capture variation in the human genome across the world

it contains more copies of some loci than hg19. Some of these additions, like

the EBV genome, are mostly relevant for genomic analysis, i.e. alignment.

For an overview of the different types and reasons for the additions see

https://software.broadinstitute.org/gatk/documentation/article?id=11010

This means that if you want to use the genome sequence for alignment and

especially for variant calling, you should use the optimal genome file for your

aligner. The genome file can make a big difference, especially for variant

calling. In most cases, the authors of your alignment program will provide

advice on which hg38 genome version to use and usually they recommend one of

the files in our analysisSet/ directory, like the GATK link above. These

special genome files sometimes remove the alternate sequences, sometimes they

add decoys or change single nucleotides towards the major allele, but they never

insert or delete sequences, so the annotation coordinates remain the same.

- for BWA see also https://lh3.github.io/2017/11/13/which-human-reference-genome-to-use

- for Novoalign see its manual at http://www.novocraft.com/userfiles/file/Novocraft.pdf

- For Bowtie, see the different versions of the human genome that the Bowtie authors

provide: http://bowtie-bio.sourceforge.net/index.shtml

Also see analysisSet/README.txt for further details

Patches

^^^^^^^

Like hg19, hg38 has been updated with patches since its release in 2013. GRC

patch releases do not change any previously existing sequences; they simply add

small, new sequences for fix patches or alternate haplotypes that correspond to

specific regions of the main chromosome sequences (see below). For most users,

the patches are unlikely to make a difference and may complicate the analysis

as they introduce more duplication. If you want a version of the genome

without these complexities, look at the analysisSet/ subdirectory.

The initial/ subdirectory contains files for the initial release of GRCh38,

which includes the original alternate sequences (261) and no fix sequences.

The p11/ subdirectory contains files for GRCh38.p11 (patch release 11).

The p12/ subdirectory contains files for GRCh38.p12 (patch release 12).

The p13/ subdirectory contains files for GRCh38.p13 (patch release 13).

The p14/ subdirectory contains files for GRCh38.p14 (patch release 14).

The "latest/" symbolic link points to the subdirectory for the most recent

patch version.

hg38.\* files in this directory are the same as files in the initial/

subdirectory, i.e. they are from the initial GRCh38 release and do not

include the patch sequences that are now included in the Genome Browser.

(The recently added hg38.gc5Base.\* files are an exception to the rule;

they do include patch sequences.)

Sequence names

^^^^^^^^^^^^^^

For historical reasons, what UCSC calls "chr1", Ensembl calls "1" and NCBI

calls "NC\_000067.6". The sequences are identical though. To map between UCSC,

Ensembl and NCBI names, use our table "chromAlias", available via our Table

Browser or as file:

https://hgdownload.cse.ucsc.edu/goldenPath/mm10/database/chromAlias.txt.gz We

also provide a Python command line tool to convert sequence names in the most

common genomics file formats:

http://hgdownload.cse.ucsc.edu/admin/exe/linux.x86\_64/chromToUcsc

During genome assembly, reads are assembled into "contigs" (a few kbp long),

which are then joined into longer "scaffolds" of a few hundred kbp. These are

finally placed, often manually e.g. with FISH assays, onto chromosomes.

As a result, the hg38 genome sequence files contains different types of sequences:

Chromosomes:

- made from scaffolds placed onto chromosome locations, 95% of the genome file

- format: chr{chromosome number or name}

- e.g. chr1 or chrX, chrM for the mitochondrial genome.

Unlocalized scaffolds:

- a sequence found in an assembly that is associated with a specific

chromosome but cannot be ordered or oriented on that chromosome.

- format: chr{chromosome number or name}\_{sequence\_accession}v{sequence\_version}\_random

- e.g. chr17\_GL000205v2\_random

Unplaced scaffolds:

- a sequence found in an assembly that is not associated with any chromosome.

- format: chrUn\_{sequence\_accession}v{sequence\_version}

- e.g. chrUn\_GL000220v1

Alternate loci scaffolds:

- a scaffold that provides an alternate representation of a locus found

in the primary assembly. These sequences do not represent a complete

chromosome sequence although there is no hard limit on the size of the

alternate locus; currently these are less than 1 Mb. These could either

be NOVEL patch sequences, added through patch releases, or present in the

initial assembly release.

- format: chr{chromosome number or name}\_{sequence\_accession}v{sequence\_version}\_alt

- e.g. chr6\_GL000250v2\_alt

Fix loci scaffolds:

- a patch that corrects sequence or reduces an assembly gap in a given

major release. FIX patch sequences are meant to be incorporated into

the primary or existing alt-loci assembly units at the next major

release.

- these sequences are not part of the files in the initial/ directory

- format: chr{chromosome number or name}\_{sequence\_accession}v{sequence\_version}\_fix

- e.g. chr2\_KN538362v1\_fix

Files

^^^^^

Files in this directory reflect the initial 2013 release of the genome,

the most current versions are in the "latest/" subdirectory:

hg38.fa.gz - "Soft-masked" assembly sequence in one file.

Repeats from RepeatMasker and Tandem Repeats Finder (with period of 12 or

less) are shown in lower case; non-repeating sequence is shown in upper

case. (again, the most current version of this file is latest/hg38.fa.gz)

hg38.2bit - contains the complete human/hg38 genome sequence

in the 2bit file format. Repeats from RepeatMasker and Tandem Repeats

Finder (with period of 12 or less) are shown in lower case; non-repeating

sequence is shown in upper case. The utility program, twoBitToFa (available

from the kent src tree), can be used to extract .fa file(s) from

this file. A pre-compiled version of the command line tool can be

found at:

http://hgdownload.cse.ucsc.edu/admin/exe/linux.x86\_64/

See also:

http://genome.ucsc.edu/admin/git.html

http://genome.ucsc.edu/admin/jk-install.html

hg38.agp.gz - Description of how the assembly was generated from

fragments.

hg38.chromFa.tar.gz - The assembly sequence in one file per chromosome.

Repeats from RepeatMasker and Tandem Repeats Finder (with period

of 12 or less) are shown in lower case; non-repeating sequence is

shown in upper case.

hg38.chromFaMasked.tar.gz - The assembly sequence in one file per chromosome.

Repeats are masked by capital Ns; non-repeating sequence is shown in

upper case.

hg38.fa.masked.gz - "Hard-masked" assembly sequence in one file.

Repeats are masked by capital Ns; non-repeating sequence is shown in

upper case.

hg38.fa.out.gz - RepeatMasker .out file. RepeatMasker was run with the

-s (sensitive) setting.

June 20 2013 (open-4-0-3) version of RepeatMasker

RepBase library: RELEASE 20130422

hg38.fa.align.gz - RepeatMasker .align file. RepeatMasker was run with the

-s (sensitive) setting.

June 20 2013 (open-4-0-3) version of RepeatMasker

RepBase library: RELEASE 20130422

hg38.trf.bed.gz - Tandem Repeats Finder locations, filtered to keep repeats

with period less than or equal to 12, and translated into UCSC's BED

format.

md5sum.txt - checksums of files in this directory

mrna.fa.gz - Human mRNA from GenBank. This sequence data is updated

regularly via automatic GenBank updates.

refMrna.fa.gz - RefSeq mRNA from the same species as the genome.

This sequence data is updated regularly via automatic GenBank

updates.

upstream1000.fa.gz - Sequences 1000 bases upstream of annotated

transcription starts of RefSeq genes with annotated 5' UTRs.

This file is updated regularly. It might be slightly out of sync with

the RefSeq data shown on the browser, as is it updated daily for most assemblies.

upstream2000.fa.gz - Same as upstream1000, but 2000 bases.

upstream5000.fa.gz - Same as upstream1000, but 5000 bases.

xenoMrna.fa.gz - GenBank mRNAs from species other than that of

the genome. This sequence data is updated regularly via

automatic GenBank updates.

hg38.chrom.sizes - Two-column tab-separated text file containing assembly

sequence names and sizes.

hg38.gc5Base.wigVarStep.gz - ascii data wiggle variable step values used

- to construct the GC Percent track

hg38.gc5Base.bw - binary bigWig data for the gc5Base track.

hg38.chromAlias.txt - sequence name alias file, one line

for each sequence name. First column is sequence name followed by

tab separated alias names.

hg38.chromAlias.bb - bigBed file for alias sequence names, one line

for each sequence name. The first three columns are the sequence in

BED format, followed by tab-separated alias names.

The .bb file is used by bedToBigBed as a URL to avoid having to download

the entire chromAlias.txt file. From the usage message:

-sizesIsChromAliasBb -- If set, then chrom.sizes file is assumed to be a

chromAlias bigBed file or a URL to a such a file (see above).

More documentation is found here:

https://genomewiki.ucsc.edu/index.php?title=Chrom\_Alias

Dropped in Genbank and Refseq official releases patch14 since these 2 old versions are obsolete and no longer needed.

This patch contains their v2 replacements.

chr11\_KQ759759v1\_fix

chr22\_KQ759762v1\_fix

Dropped in Refseq official release Patch14.

These 3 are contamination or obsolete.

chr10\_KI270825v1\_alt

chr22\_KI270734v1\_random

chr11\_KI270721v1\_random

Because of the difficulty of removing the old chroms chr11\_KQ759759v1\_fix and chr22\_KQ759762v1\_fix from all of the database tables and bigData files,

custom tracks, and hubs, we are not dropping them from the UCSC hg38 patch 14 .2bit and chromInfo.

However, we have dropped them from chromAlias to accord with the Genbank and Refseq official releases for patch14.

Dropped in Patch13 from Refseq

chrUn\_KI270752v1 HSCHRUN\_RANDOM\_CTG29 KI270752.1

KI270752.1 is no longer part of the RefSeq assembly because it is hamster sequence

derived from the human-hamster CHO cell line.

https://www.ncbi.nlm.nih.gov/grc/human/issues/HG-2587

------------------------------------------------------------------

How to Download

^^^^^^^^^^^^^^^

If you plan to download a large file or multiple files from this

directory, we recommend that you use ftp rather than downloading the

files via our website. To do so, ftp to hgdownload.cse.ucsc.edu

[username: anonymous, password: your email address], then cd to the

directory goldenPath/hg38/bigZips. To download multiple files, use

the "mget" command:

mget <filename1> <filename2> ...

- or -

mget -a (to download all the files in the directory)

Alternate methods to ftp access.

Using an rsync command to download the entire directory:

rsync -avzP rsync://hgdownload.cse.ucsc.edu/goldenPath/hg38/bigZips/ .

For a single file, e.g. chromFa.tar.gz

rsync -avzP

rsync://hgdownload.cse.ucsc.edu/goldenPath/hg38/bigZips/chromFa.tar.gz .

Or with wget, all files:

wget --timestamping

'ftp://hgdownload.cse.ucsc.edu/goldenPath/hg38/bigZips/\*'

With wget, a single file:

wget --timestamping

'ftp://hgdownload.cse.ucsc.edu/goldenPath/hg38/bigZips/chromFa.tar.gz'

-O chromFa.tar.gz

To unpack the \*.tar.gz files:

tar xvzf <file>.tar.gz

To uncompress the fa.gz files:

gunzip <file>.fa.gz

REDUCE SAMPLE

If the sample is too big and one can not compute the file. Try to reduce the file with the following scripts:

head -20000 SRR23019314\_Whole\_exome\_sequence\_of\_a\_patient\_with\_Harlequin\_Ichthyosis\_1.fastq > test\_1.fastq

gzip SRR23019314\_Whole\_exome\_sequence\_of\_a\_patient\_with\_Harlequin\_Ichthyosis\_1.fastq

DOCKERFILE

docker tag pipeline\_mock miguelestebancapdevila/pipeline\_mock:v1

PUSH DOCKER TO GCP

CHECK THE NUMBER OF LINES

wc -l SRR\*