GOENOMICS

World of biological sequences

HIGH QUALITY GENOME ANNOTATION

Boost your research with highest quality genome annotations. Save time and efforts, and select annotations from the genome-catalog or let us annotate your genome assembly.

FUNCTIONAL ANNOTATION

gene/protein name
enzyme number
domain architecture
molecular function
cellular location
biological process

SNP | SV

single nucleotide polymorphisms
structural variants

AMINO ACID SEQUENCE

here: of mutually exclusive exons

ALTERNATIVE SPLICING

allowing multiple alternative transcripts from one gene
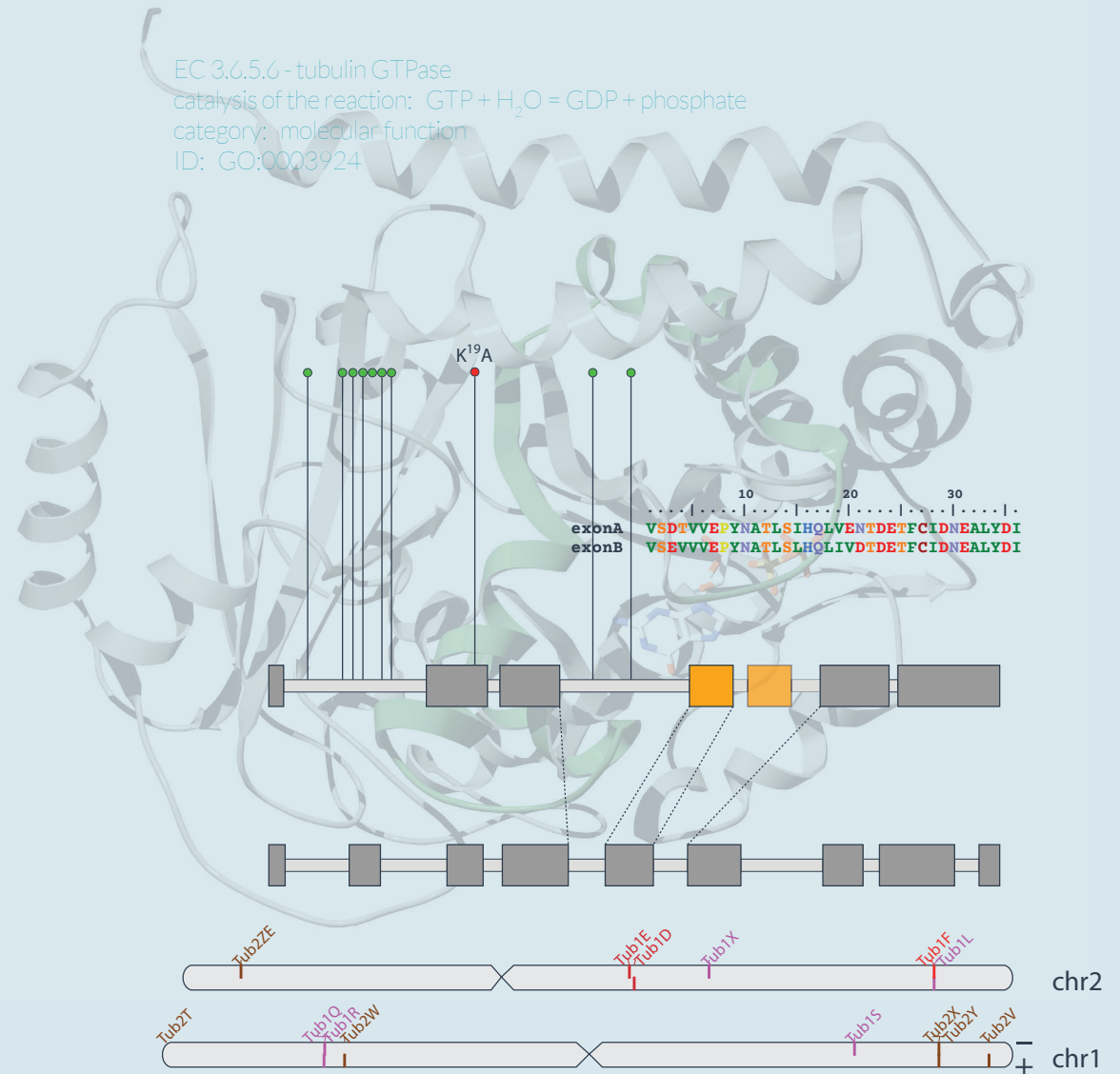
GENE STRUCTURE

characterized by exons (boxes) and introns (line)

GENE LOCUS

on contig or chromosome

EC 3.6.5.6 - tubulin GTPase
catalysis of the reaction: $GTP + H_2O = GDP + phosphate$
category: molecular function
ID: GO:0003924

$K^{19}A$

```
           10        20        30
           .........|.........|.........|...
exonA  VSDTVVEPYNATLSIHQLVENTDETFCIDNEALYDI
exonB  VSEVVVEPYNATLSLHQLIVDTDETFCIDNEALYDI
```

Tub2ZE

Tub1E
Tub1D   Tub1X   Tub1F
                              Tub1L   chr2

Tub2T   Tub1Q
         Tub1R
         Tub2W   Tub1S   Tub2X
                               Tub2Y
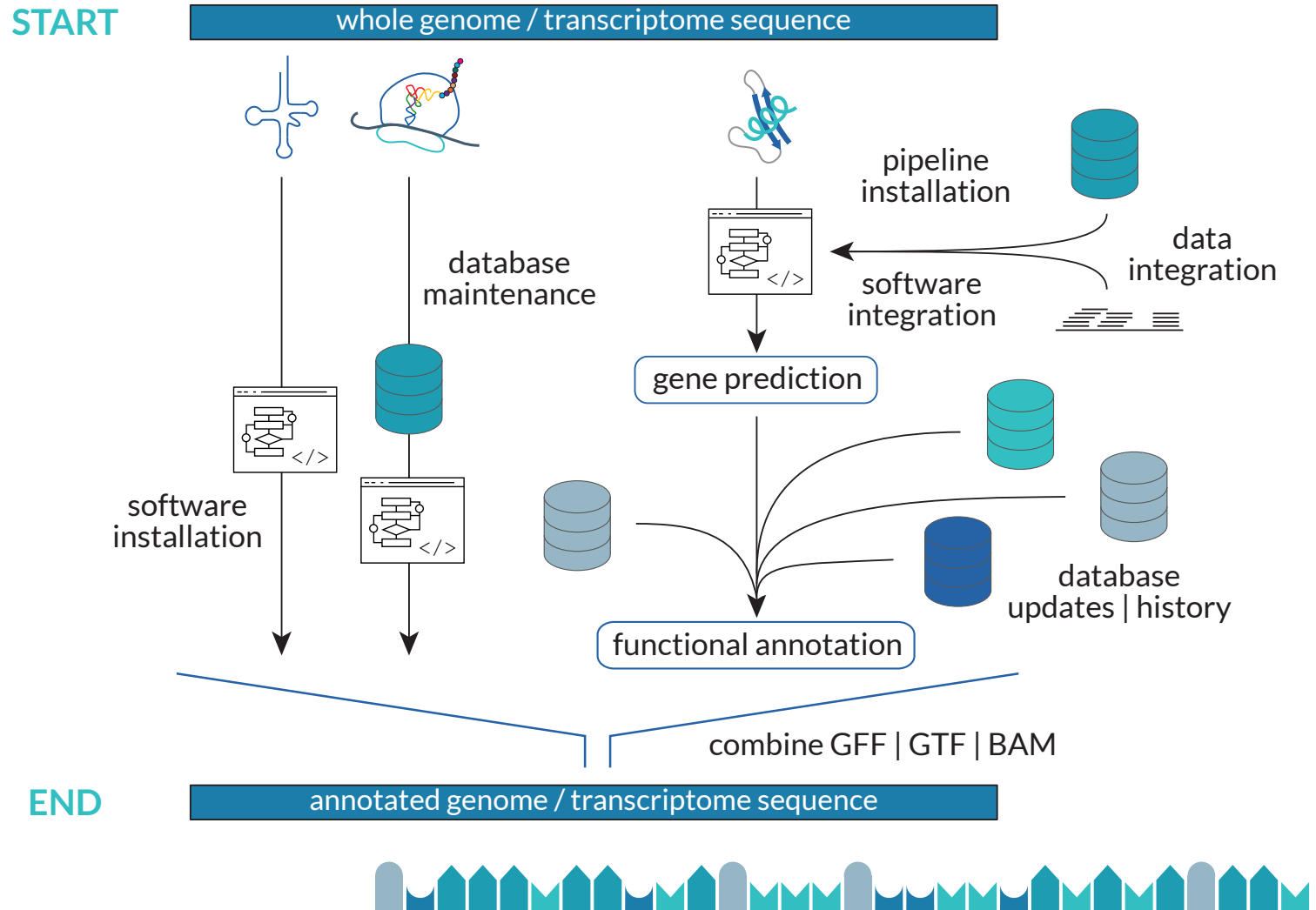   Tub2V   chr1

GOENOMICS developed proprietary data structures, software and databases to support your research with excellent genome annotations, be it for genetics, molecular biology, biotechnology, plant breeding, disease understanding, population genomics, proteogenomics or anyother research area in need of biological sequences.

ONE CLICK SOLUTION

GOENOMICS offers a full annotation of protein-coding and most common RNA-family genes. Annotation of UTRs and lncRNA genes might be possible. Let us take care about all the hassle with installation of software, pipelines and databases, their maintenance and updates, and choosing the best parameters for each genome. Each annotated feature is tracked over time with respect to structural and functional annotation.
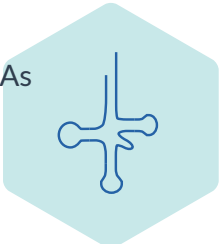

START — whole genome / transcriptome sequence

pipeline installation

data integration

software integration

database maintenance

gene prediction

software installation

functional annotation

database updates | history

combine GFF | GTF | BAM

END — annotated genome / transcriptome sequence

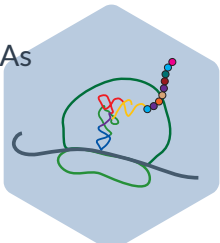# CONSISTENT ANNOTATION across the TREE of LIFE

Proteins

tRNAs

rRNAs

A HEXOKINASE IS A HEXOKINASE IS A HEXOKINASE

Profit from consistent naming and function assignment in our database. Whenever revisions become necessary, former annotations are available and searchable in the history. Multi-domain proteins get multiple annotations.

Hexokinase
EC 2.7.1.1
Hexokinase_1 (PF00349)
GO:0004396 hexokinase activity | molecular function
catalysis of the reaction:
ATP + D-hexose = ADP + D-hexose 6-phosphate
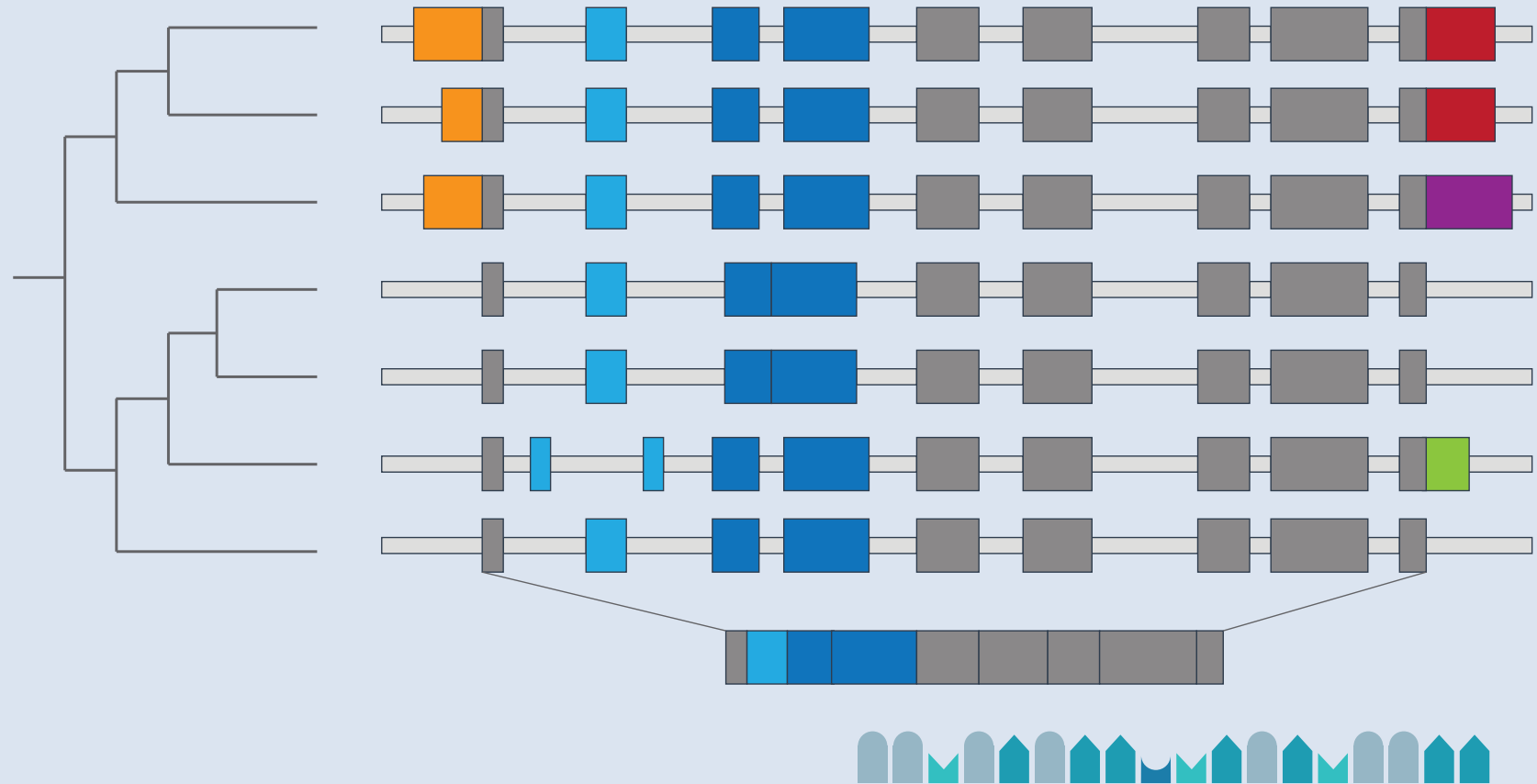
GOENOMICS

GENOMES and GENOME ASSEMBLIES

GOENOMICS examines all assemblies according to statistical criteria and performs quality checks. The data from the analyses will be presented in comparison to data from related/similar genomes.

ASSEMBLY GENERATION

Preferably, a genome assembly consists of as few contigs as possible (ideally just the number of chromosomes and organellar genomes) nevertheless also being as complete as possible. The genome assembly process is a balance between longer contigs versus higher per-base coverage.
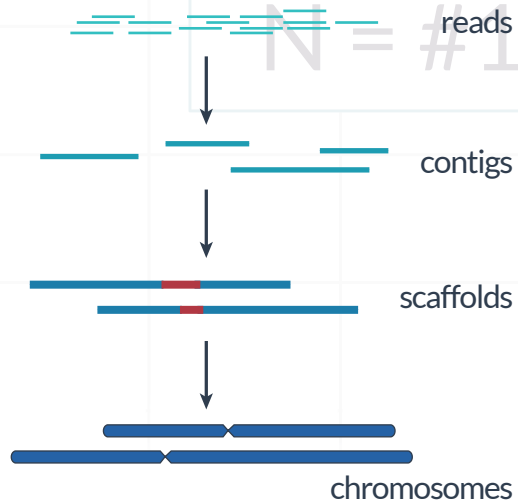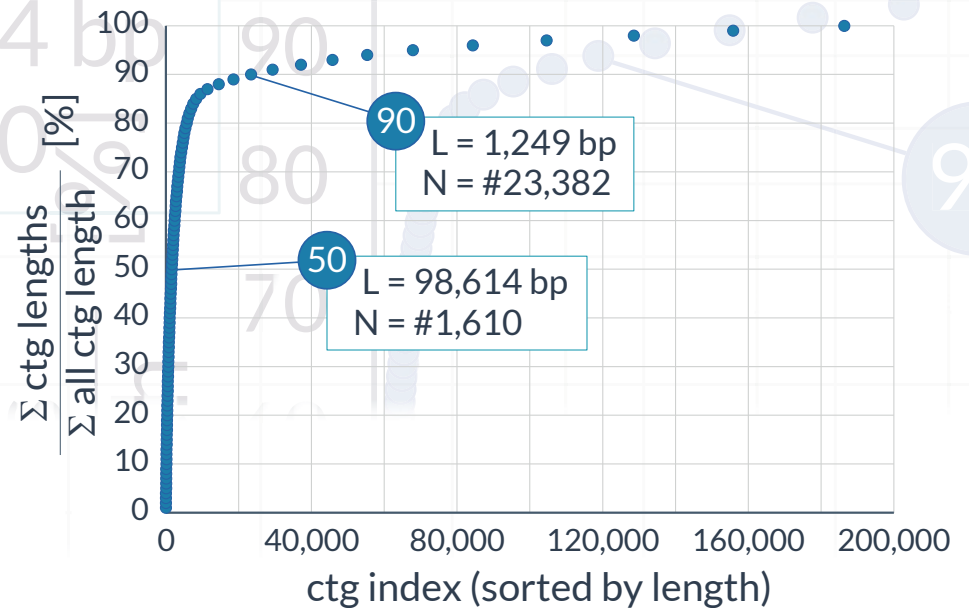
ASSEMBLY QUALITY EVALUATION

The N50/L50 are metrics to describe the quality of a genome assembly. Because these numbers can at least slightly be tweaked by applying minimum contig length cut-offs or by adjusting the accuracy bounding for aligning reads, it is best to inspect the trend of contig lengths and numbers over the entire genome assembly.

We are aware that the N50 and L50 are often switched, e.g. the N50 denoting the length of the contig and L50 referring to the number of contigs. Also, the descriptions to compute N/L50 vary between publications and webpages.

If the genome size is known, often used metrics are NG50 and LG50, which are identical to N50 and L50 except that the computation is not related to the total genome assembly lenght but the (real/estimated) genome size.

reads

contigs

scaffolds

chromosomes

L = 1,249 bp
N = #23,382

L = 98,614 bp
N = #1,610

Σ ctg lengths / Σ all ctg length [%]

ctg index (sorted by length)

READ MORE   goenomics.com/glossary.tml#assemblies

GOENOMICS

GENOME ANNOTATIONS WILL BE
PROVIDED IN FORMATS TAILORED TO YOUR NEEDS IN DOWNSTREAM ANALYSES/PROCESSES

RESULTS IN COMMON FILE FORMATS

FASTA

The FASTA-format is the most common file format for biological sequences.

The first line of each sequence starts with the character ">" immediately followed (no space) by a description of the sequence. The lines until the next sequence header or the end of the file contain the sequence.

GFF | GTF

GFF (General Feature Format) and GTF (Gene Transfer Format) are file formats that hold information about gene and genome annotation. The format has initially been designed to define gene structures with respect to sequences, which could either be included in the file or present in an accompanying file. Both formats in their current versions require data in nine tab-delimited columns (not space delimited).

GFF only | not allowed in GTF
comment specifying the GFF-version (mandatory), further comments are optional
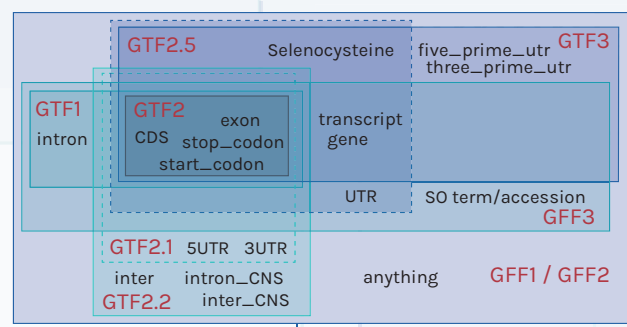
stop codon		
GTF3	excluded	
GFF3	included	
preGTF3	NS	
preGFF3	NS	

format (exactly, including spaces and semicolons)

GTF3	attribute_name "attribute_value"; attribute_name "attribute_value";
GFF3	attribute_name=attribute_value;attribute_name=attribute_value

allowed keys/values | tags/values

GTF3	Two mandatory attributes: gene_id, transcript_id; any additions allowed.
GFF3	Any. IDs mandatory for features that have children or span multiple lines.

```
                                        float/[.]    [+-.?]    [012.]

<seqname> <source> <feature> <start> <end> <score> <strand> <frame> [attribute]

##gff-version 3
contig1    software   mRNA    1750    8632    .       +       .    ID=mrna0001;Name=genexy
contig1    software   exon    1750    2500    .       +       .    ID=exon00001;Parent=mrna0001
contig1    software   exon    2550    3476    .       +       .    ID=exon00002;Parent=mrna0001
contig1    software   exon    4190    4980    .       +       .    ID=exon00003;Parent=mrna0001
contig1    software   exon    5612    5970    .       +       .    ID=exon00004;Parent=mrna0001
contig1    software   exon    7881    8632    .       +       .    ID=exon00005;Parent=mrna0001
```

GOENOMICS

World of biological sequences

CEO

Dr. Martin Kollmar
kollmar@goenomics.com

CTO

Dr. Dominic Simm
simm@goenomics.com

GOENOMICS GmbH

Benfeyweg 9
37075 Göttingen
Germany

@goenomics

www goenomics.com

HIGH QUALITY GENOME ANNOTATION