Update on the Eukaryotic Genome Annotation Pipeline

Françoise Thibaud-Nissen

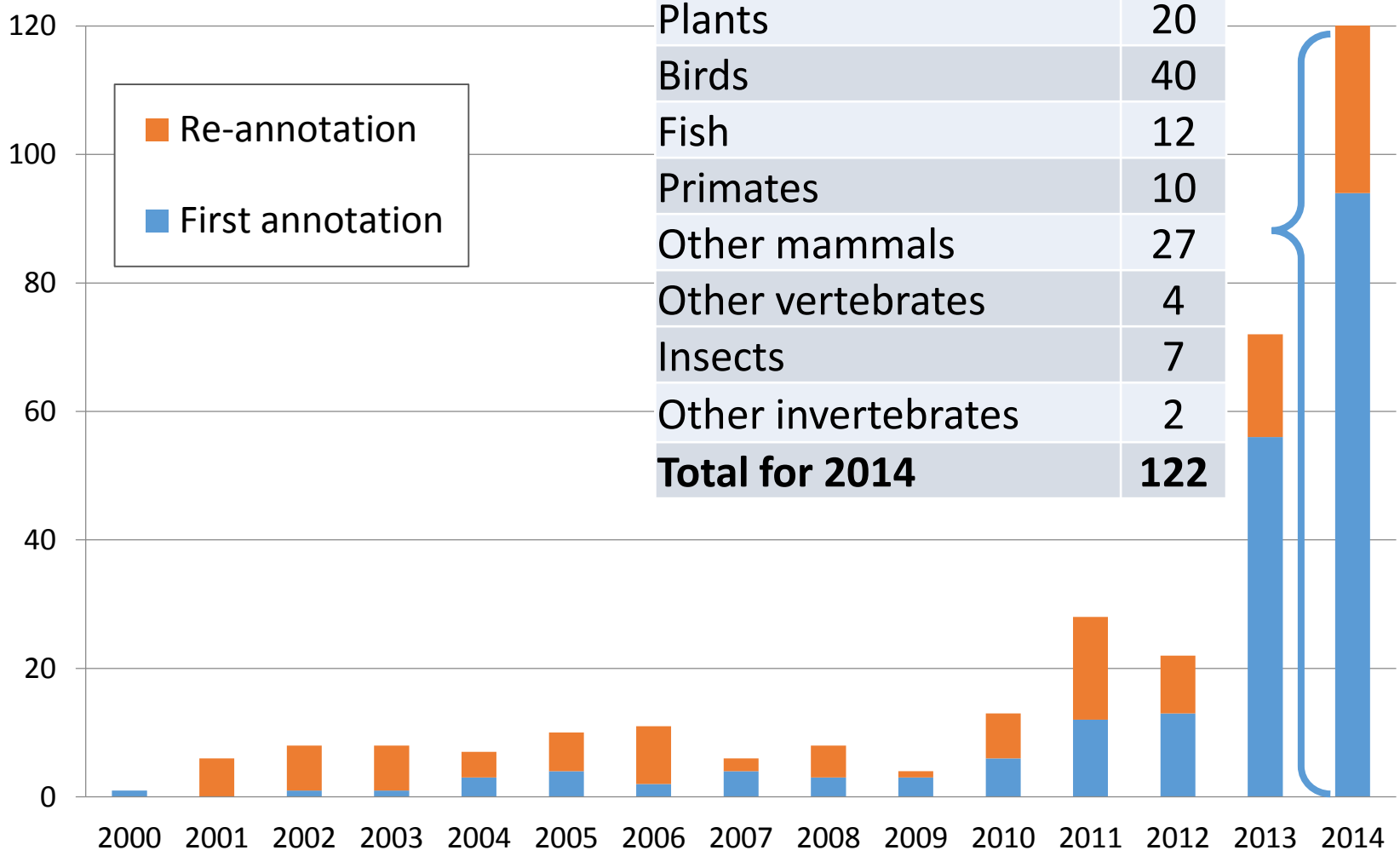
PAG XXIII January 10-14, 2015

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, DHHS

Over 200 organisms annotated

Taxonomic group	Count
Primates	15
Rodents	14
Other mammals	47
Birds	51
Fish	18
Other vertebrates	8
Invertebrates	26
Plants	26

Number of annotations released per year

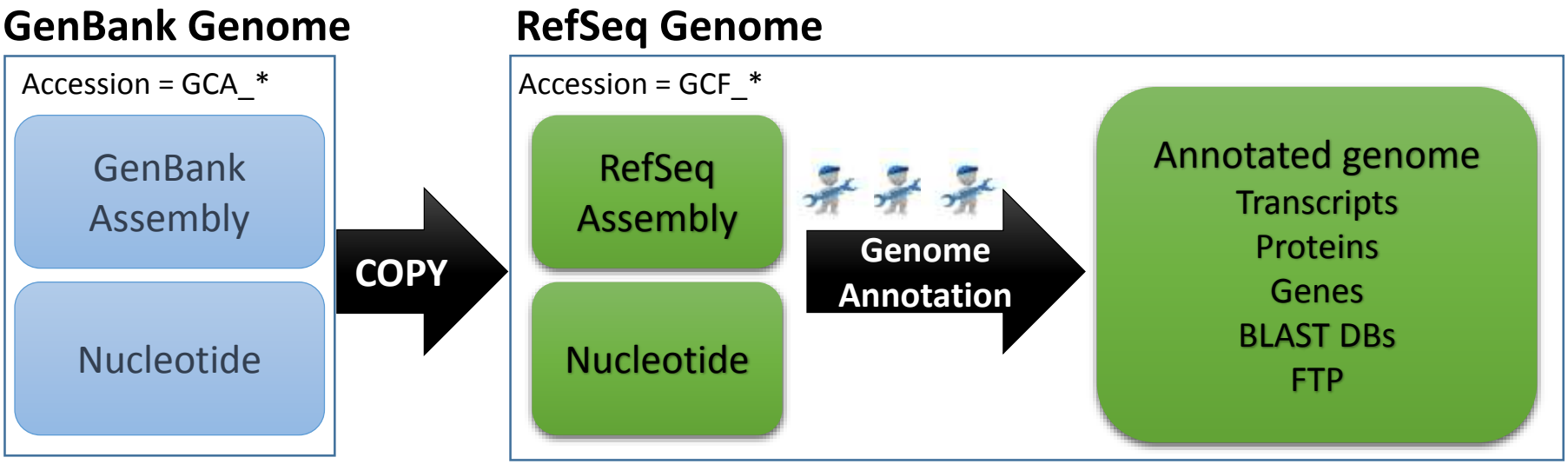


Plants	20
Birds	40
Fish	12
Primates	10
Other mammals	27
Other vertebrates	4
Insects	7
Other invertebrates	2
Total for 2014	**122**

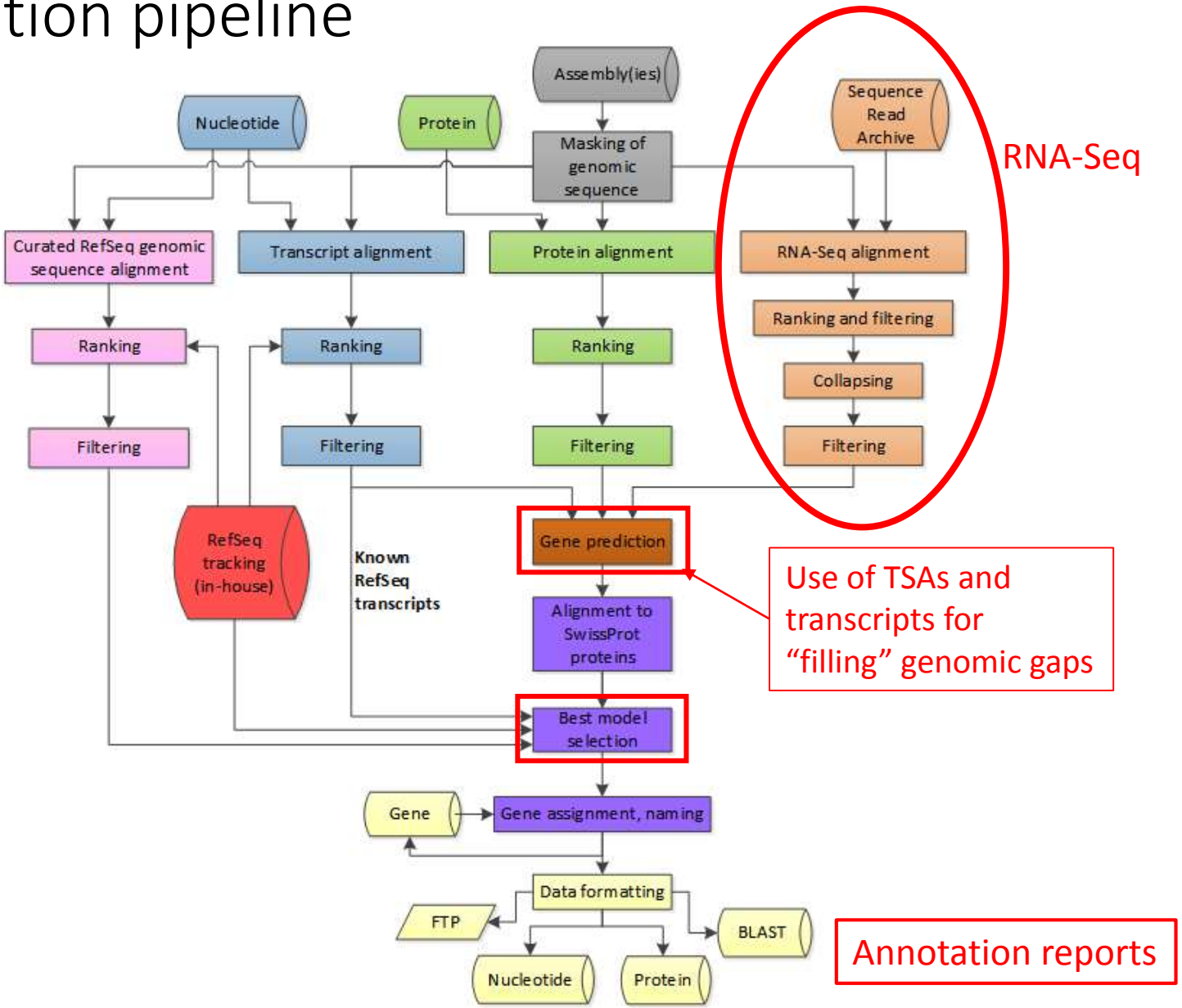
Inputs to the annotation pipeline

- Public data!
- Assemblies submitted to GenBank, ENA or DDBJ
- RNA-Seq available in SRA
- Transcripts, proteins, ESTs
- TSA (Transcript Shotgun Assemblies)

Background


GenBank Genome
Accession = GCA_*
GenBank Assembly
Nucleotide
COPY
RefSeq Genome
Accession = GCF_*
RefSeq Assembly
Nucleotide
Genome Annotation
Annotated genome
Transcripts
Proteins
Genes
BLAST DBs
FTP

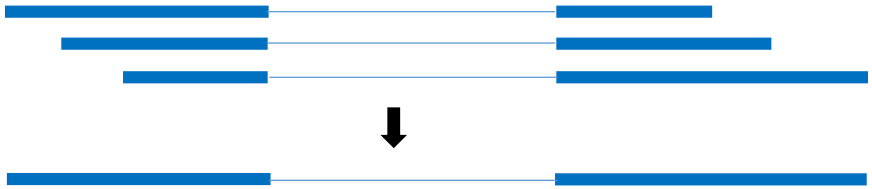
The annotation pipeline



RNA-Seq incorporation into model prediction - spring of 2013

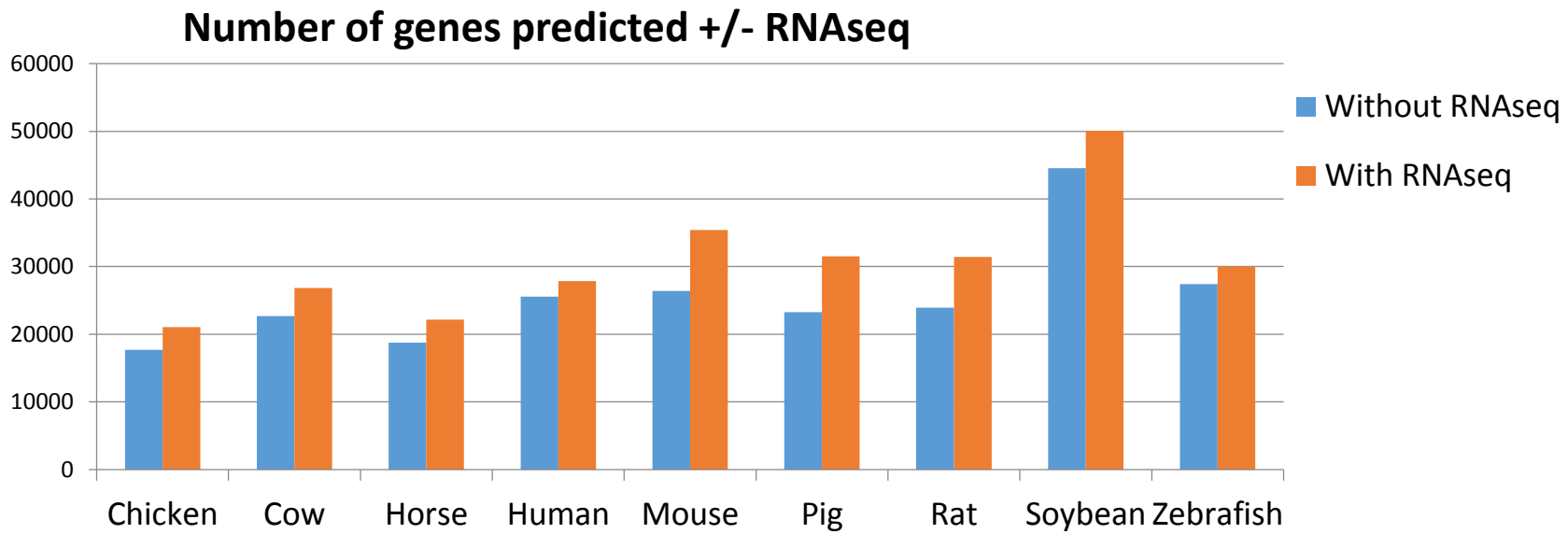
A computational challenge!

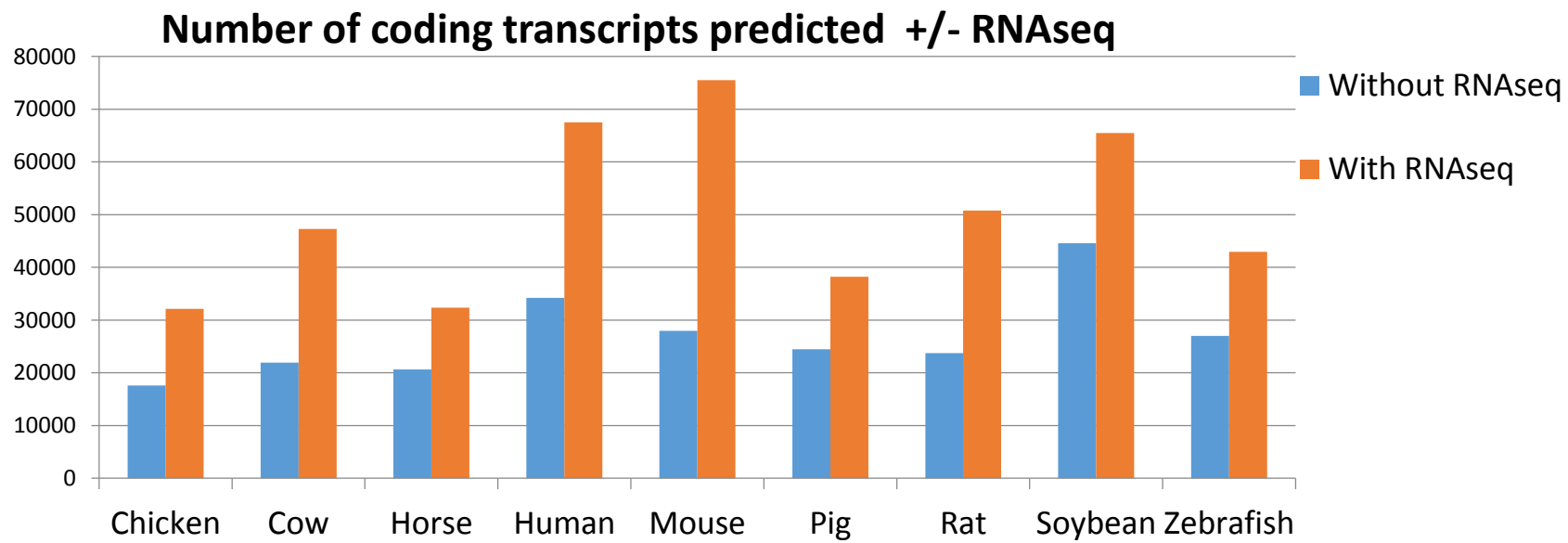
- Query of the data in SRA
- Uniq'ing of reads queried from SRA
- Alignment to the genome
- Collapsing of compatible alignments



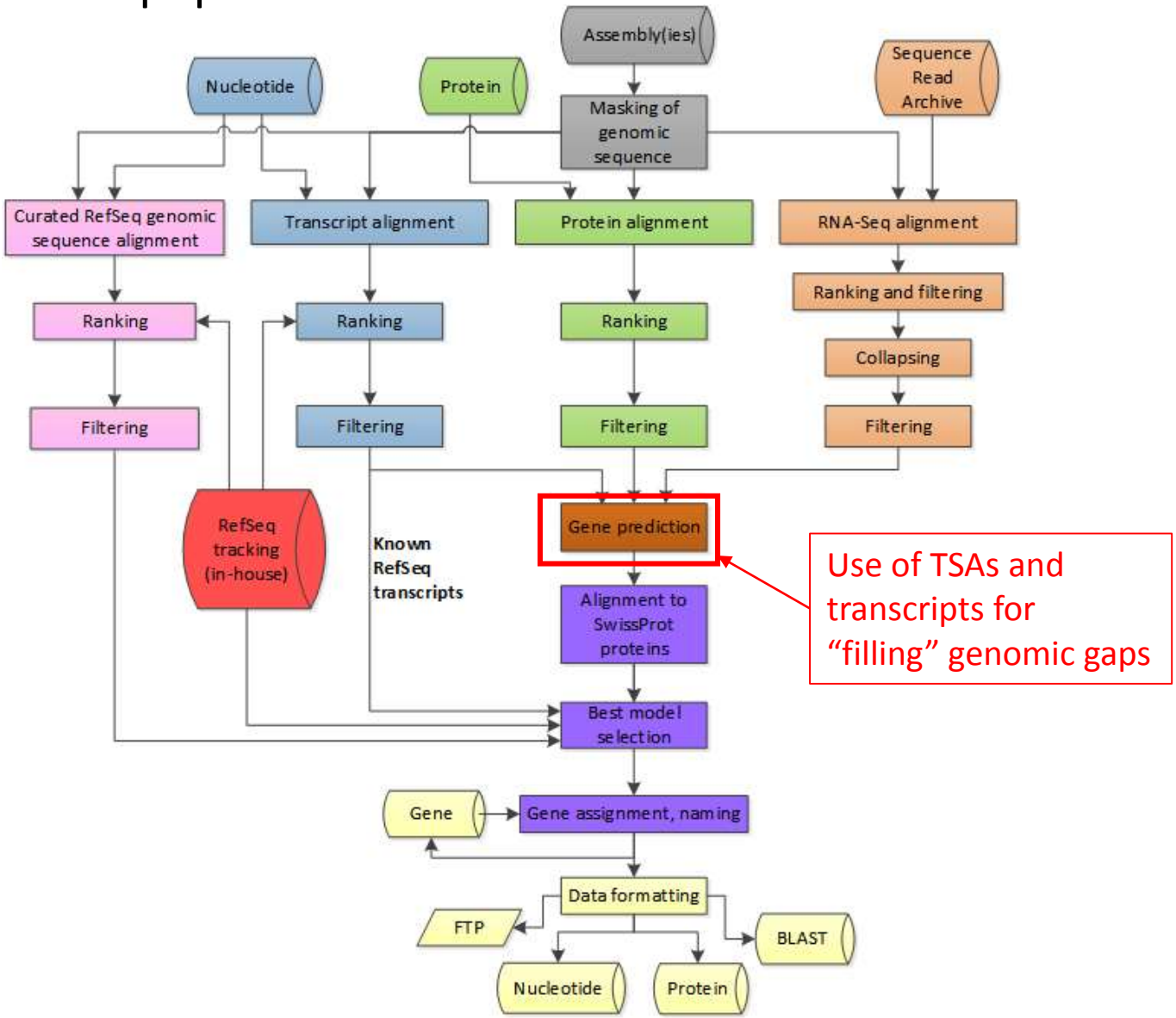
- Tracking of runs and samples associated with a model prediction for downstream evalutation of the models produced

Impact of RNA-Seq on gene prediction

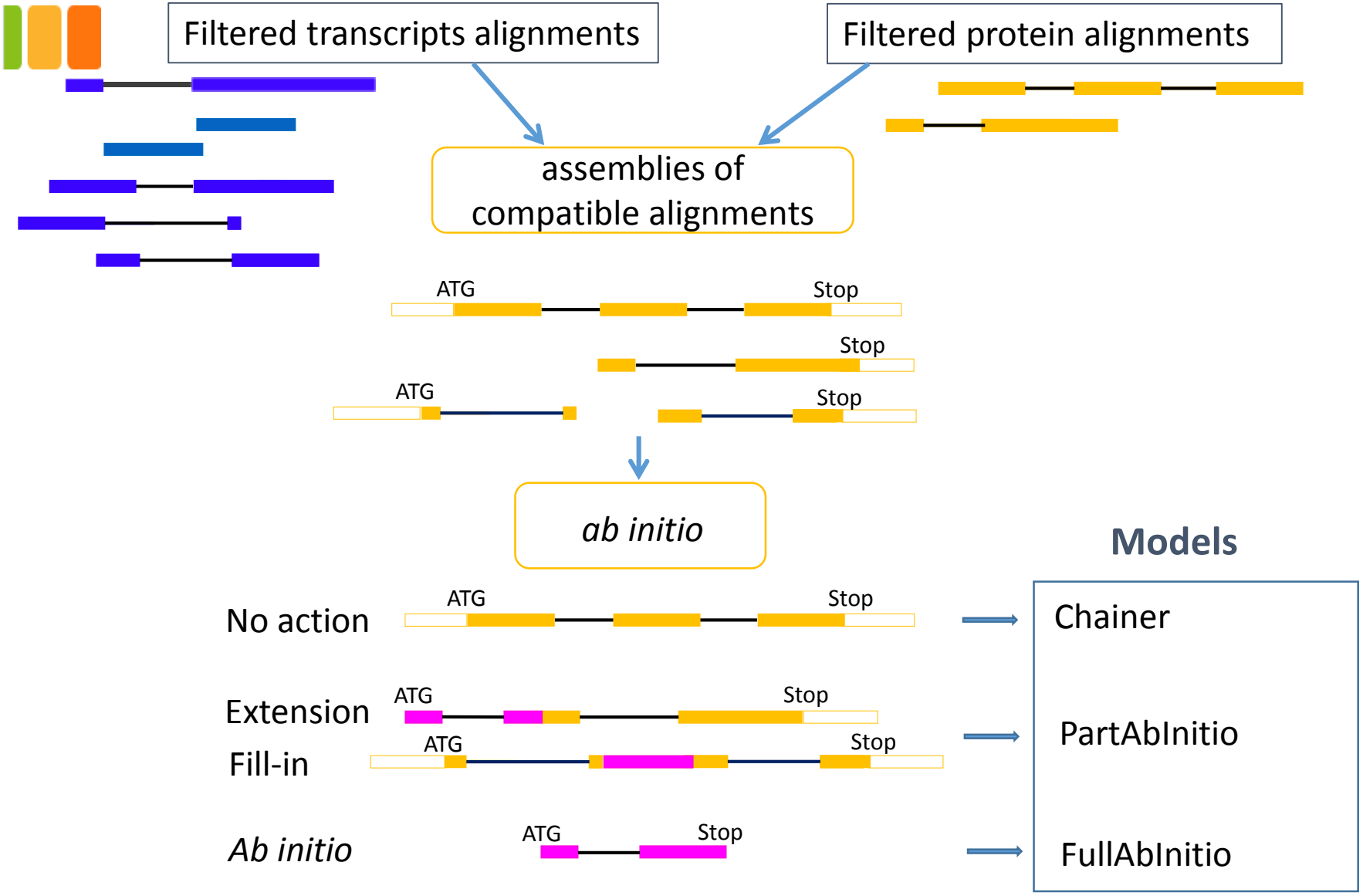




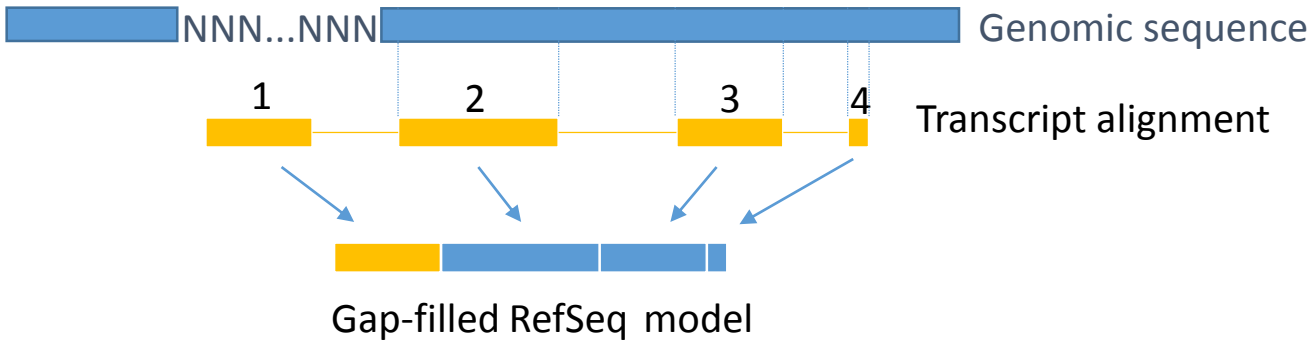
The annotation pipeline



Model prediction - Gnomon



Gap-filling: using transcripts to compensate for missing genomic sequence – April 2014



XM_007659754.1: Ornithorhynchus anatinus corticotropin releasing hormone binding protein (CRHBP), mRNA

PRIMARY	REFSEQ_SPAN	PRIMARY_IDENTIFIER	PRIMARY_SPAN	COMP
	1-523	JQ350810.1	1-523	
	524-672	AAPN01287557.1	7062-7210	c
	673-790	AAPN01287557.1	4039-4156	c
	791-1741	AAPN01287557.1	507-1457	c

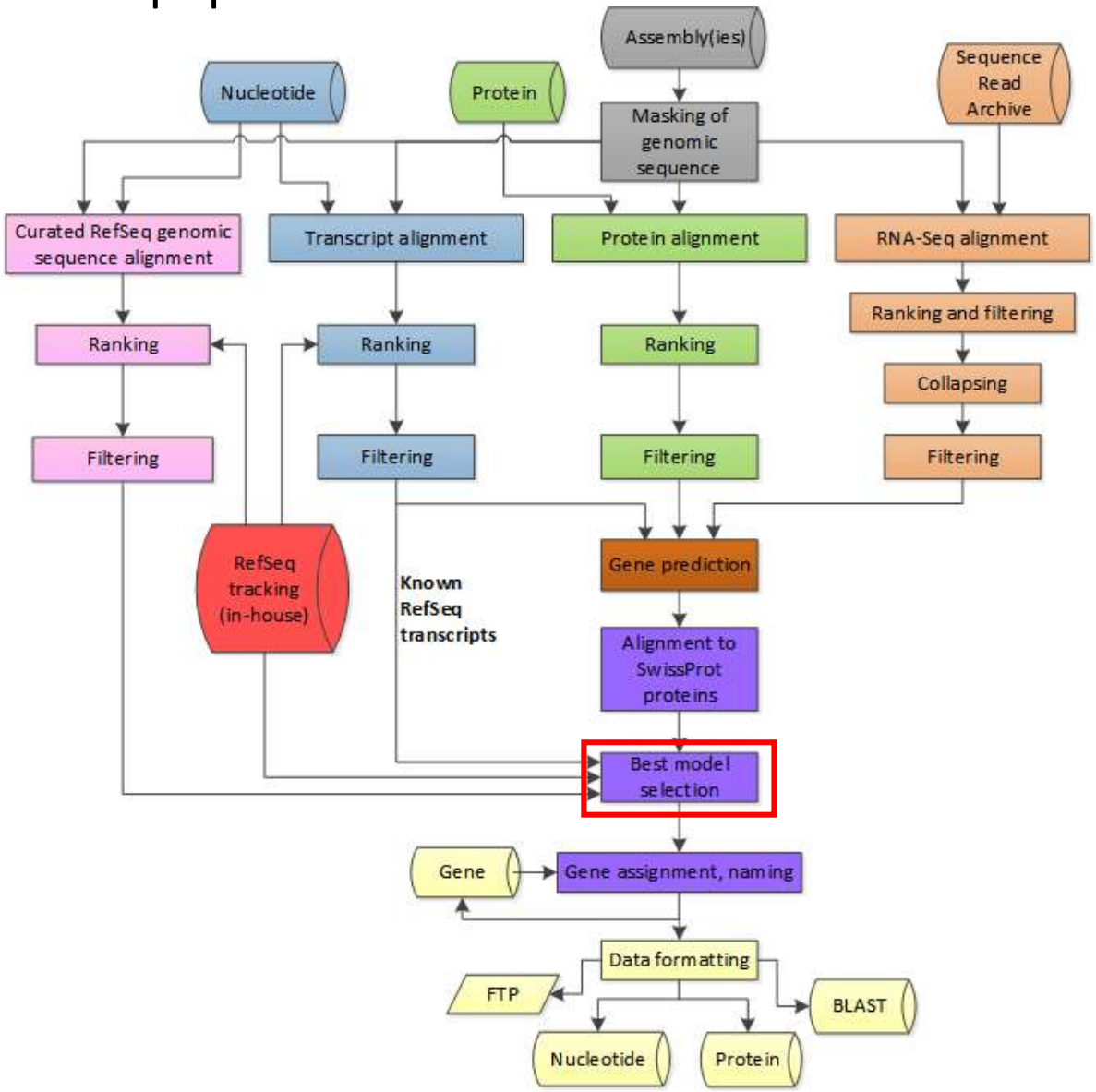
```
COMMENT     MODEL REFSEQ:  This record is predicted by automated computational
            analysis. This record is derived from a genomic sequence
            (NW_001594470.1) and transcript sequence (JQ350810.1) annotated
            using gene prediction method: Gnomon, supported by mRNA evidence.
```

```
##RefSeq-Attributes-START##
assembly gap :: added 523 transcript bases to patch genome assembly
                gap
##RefSeq-Attributes-END##
```

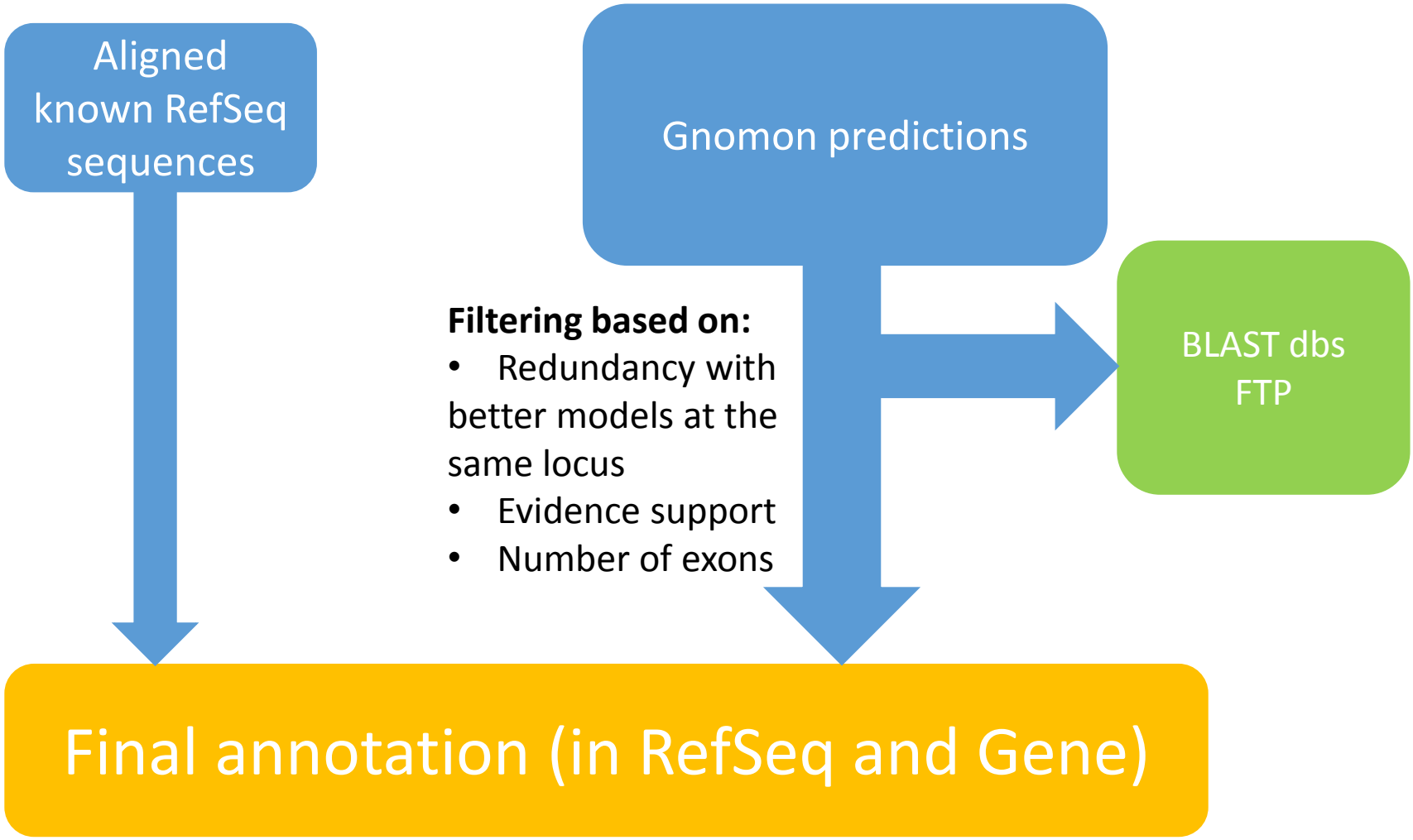
Impact of gap-filling on annotation

Species	Number of gap-filled gene	Number of TSA available
Green anole (Anolis carolinensis)	3,948	8,538,081
Rabbit (Oryctolagus cuniculus)	3,804	2,180,071
Elephant shark (Callorhinchus milii)	2,820	17,031
Marmoset (Callithrix jacchus)	2,385	210,006
Chimpanzee (Pan troglodytes)	1,654	119,002
Black rockcod (Esox lucius)	1,516	38,696
Chinese hamster (Cricetulus griseus)	1,256	63,953
Elaeis guineensis (African oil palm)	1,073	147,058

The annotation pipeline



Selection of the best models for RefSeq



Key features of the pipeline

- Higher weight given to known RefSeq sequences
- Extensive use of RNA-Seq data for gene prediction
- Annotation of splice variants
- Compensation of inaccuracies in genomic sequence in RefSeq models
  - Correction of indels
  - Correction of frameshifts
  - Gap-filling with transcript data
- Use of orthology calculation for protein naming
- Tracking of genes from annotation version to the next
- Possibility to annotate multiple assemblies for the same organism in parallel

Annotation in progress

http://www.ncbi.nlm.nih.gov/genome/annotation_euk/status

Annotation runs in progress

An annotation run is marked in progress until the data produced is available in the sequence databases, in Gene and on the FTP site.

Species	RefSeq assembly(ies)	Annotation Release	Freeze Date	Status
Camelus dromedarius (Arabian camel)	PRJNA234474_Ca_dromedarius_V1.0	100	2015-01-05	Automated processing in progress
Elaeis guineensis (African oil palm)	EG5 (GCF_000442705.1)	100	2015-01-05	In sequence databases
Populus euphratica (Euphrates poplar)	PopEup_1.0 (GCF_000495115.1)	100	2015-01-05	In sequence databases
Camelus bactrianus (Bactrian camel)	Ca_bactrianus_MBC_1.0 (GCF_000767855.1)	100	2015-01-02	In sequence databases

Recently completed annotation runs

Last 10 genomes annotated.

F - FTP, MV - Map Viewer, B - Organism-specific BLAST, AR - Annotation Report

Species	RefSeq assembly(ies)	Annotation Release	Freeze Date	Release Date	Links		
Esox lucius (northern pike)	EsoLuc1.0 (GCF_000721915.1)	100	2014-12-29	2014-12-31	F		AR
Notothenia coriiceps (black rockcod)	NC01 (GCF_000735185.1)	100	2014-12-18	2014-12-23	F		AR
Meleagris gallopavo (turkey)	Turkey_5.0 (GCF_000146605.2)	101	2014-12-10	2014-12-18	F	B	AR
Larimichthys crocea (large yellow croaker)	ASM74293v1 (GCF_000742935.1)	100	2014-12-15	2014-12-18	F	B	AR

Annotation reports – December 2013

Annotation Release information

Assemblies

The following assemblies were included in this annotation run:

Assembly name	Assembly accession	Submitter	Assembly date	Reference/Alternate
Bos_taurus_UMD_3.1.1	GCF_000003055.6	Center for Bioinformatics and Computational Biology, University of Maryland	11-25-2014	Reference
Btau_4.6.1	GCF_000003205.5	Cattle Genome Sequencing International Consortium	11-02-2011	Alternate

Annotation reports – December 2013

Annotation Release in

Assemblies

The following assemblies wer

Assembly name	As ac
Bos_taurus_UMD_3.1.1	GC
Btau_4.6.1	GC

Reference/Alternate
Reference
Alternate

Gene and feature statistics

Counts and length of annotated features are provided below for each assembly.

Feature counts

Feature	Bos_taurus_UMD_3.1.1	Btau_4.6.1
Genes and pseudogenes	34,263	36,534
protein-coding	21,577	22,923
non-coding	9,070	9,680
pseudogenes	3,616	3,931
genes with variants	11,864	4,905
mRNAs	51,907	31,598
fully-supported	49,611	28,871
with > 5% ab initio	1,010	1,120
partial	1,987	2,240
with filled gap(s)	164	79
known RefSeq (NM_)	13,144	13,195
model RefSeq (XM_)	38,763	18,403
Other RNAs	15,619	15,795

Annotation reports – alignemnts

Transcript alignments

Bos_taurus_UMD_3.1.1 - Primary Assembly

Source	Number of sequences retrieved from Entrez	Number (%) of sequences aligned by Splign	Number (%) of sequences passed to Gnomon	Average % identity	Average % coverage
Same-species known RefSeq (NM_/NR_)	14,136	14,124 (99.92%)	13,724 (97.09%)	99.47%	98.39%
Same-species Genbank	19,155	18,703 (97.64%)	17,285 (90.24%)	99.68%	97.36%
Same-species EST	1,559,351	1,439,281 (92.30%)	1,359,899 (87.21%)	97.67%	96.91%

Annotation reports – alignemnts

Transcript alignments

Bos_taurus_UMD_3.1.1 - Primary Assembly

Source	Number of sequences retrieved from Entrez	Number (%) of sequences aligned by Splign	Number (%) of sequences passed to Gnomon	Average % identity	Average % coverage

Short read transcript alignments

The following short reads (RNA-Seq) from the Sequence Read Archive were also used for gene prediction:

Hide alignments statistics, by sample (SAME, SAMN, SAMD, DRS)

Bos_taurus_UMD_3.1.1 - Primary Assembly

Sample Id	Track name	Number of reads	Number (%) of aligned reads	Number (%) spliced reads	Number of introns
All	Aggregate of all aligned samples	7,921,963,239	6,255,606,573 (78.97%)	1,348,684,976 (17.02%)	406,671
SAMD00000728	Cows (Japanese Black) (Bos taurus, SAMD00000728)	354,627,156	222,872,071 (62.85%)	31,485,845 (8.88%)	192,504
SAMN00120237	IVF embryos (Bos taurus, SAMN00120237)	54,352,331	27,722,770 (51.01%)	4,137,919 (7.61%)	89,792
SAMN00120238	IVF embryos (Bos taurus, SAMN00120238)	54,770,027	22,448,598 (40.99%)	2,375,125 (4.34%)	84,741
SAMN01084067	Bull Sperm Transcriptome (Bos taurus, SAMN01084067)	37,140,700	1,384,934 (3.73%)	228,029 (0.61%)	25,669
SAMN01093727	omental fat tissue (Bos taurus, 30 months, female, SAMN01093727)	31,844,692	26,709,287 (83.87%)	6,865,274 (21.56%)	160,869

Annotation reports – alignemnts

Transcript alignments

Bos_taurus_UMD_3.1.1 - Primary Assembly

Protein alignments

Bos_taurus_UMD_3.1.1 - Primary Assembly

Source	Number of sequences retrieved from Entrez	Number (%) of sequences aligned by ProSplign	Number (%) of sequences passed to Gnomon	Average % identity	Average % coverage
Homo sapiens known RefSeq (NP_)	38,283	37,404 (97.70%)	37,364 (97.60%)	77.19%	82.02%
Same-species GenBank	16,482	15,814 (95.95%)	15,804 (95.89%)	82.47%	87.45%
Same-species known RefSeq (NP_)	13,269	13,177 (99.31%)	13,164 (99.21%)	83.31%	86.69%

SAMD00000728	Cows (Japanese Black) (Bos taurus, SAMD00000728)	354,627,156	222,872,071 (62.85%)	31,485,845 (8.88%)	192,504
SAMN00120237	IVF embryos (Bos taurus, SAMN00120237)	54,352,331	27,722,770 (51.01%)	4,137,919 (7.61%)	89,792
SAMN00120238	IVF embryos (Bos taurus, SAMN00120238)	54,770,027	22,448,598 (40.99%)	2,375,125 (4.34%)	84,741
SAMN01084067	Bull Sperm Transcriptome (Bos taurus, SAMN01084067)	37,140,700	1,384,934 (3.73%)	228,029 (0.61%)	25,669
SAMN01093727	omental fat tissue (Bos taurus, 30 months, female, SAMN01093727)	31,844,692	26,709,287 (83.87%)	6,865,274 (21.56%)	160,869

Annotation reports – alignemnts

Transcript alignments

Bos_taurus_UMD_3.1.1 - Primary Assembly

Source	Number of sequences retrieved from Entrez	Number (%) of sequences aligned by Splign	Number (%) of sequences passed to Gnomon	Average % identity	Average % coverage
Same-species known RefSeq (NM_/NR_)	14,136	14,124 (99.92%)	13,724 (97.09%)	99.47%	98.39%
Same-species Genbank	19,155	18,703 (97.64%)	17,285 (90.24%)	99.68%	97.36%
Same-species EST	1,559,351	1,439,281 (92.30%)	1,359,899 (87.21%)	97.67%	96.91%

Annotation reports – alignemnts

Transcript alignments

Bos_taurus_UMD_3.1.1 - Primary Assembly

Source	Number of sequences retrieved from Entrez	Number (%) of sequences aligned by Splign	Number (%) of sequences passed to Gnomon	Average % identity	Average % coverage

Short read transcript alignments

The following short reads (RNA-Seq) from the Sequence Read Archive were also used for gene prediction:

Hide alignments statistics, by sample (SAME, SAMN, SAMD, DRS)

Bos_taurus_UMD_3.1.1 - Primary Assembly

Sample Id	Track name	Number of reads	Number (%) of aligned reads	Number (%) spliced reads	Number of introns
All	Aggregate of all aligned samples	7,921,963,239	6,255,606,573 (78.97%)	1,348,684,976 (17.02%)	406,671
SAMD00000728	Cows (Japanese Black) (Bos taurus, SAMD00000728)	354,627,156	222,872,071 (62.85%)	31,485,845 (8.88%)	192,504
SAMN00120237	IVF embryos (Bos taurus, SAMN00120237)	54,352,331	27,722,770 (51.01%)	4,137,919 (7.61%)	89,792
SAMN00120238	IVF embryos (Bos taurus, SAMN00120238)	54,770,027	22,448,598 (40.99%)	2,375,125 (4.34%)	84,741
SAMN01084067	Bull Sperm Transcriptome (Bos taurus, SAMN01084067)	37,140,700	1,384,934 (3.73%)	228,029 (0.61%)	25,669
SAMN01093727	omental fat tissue (Bos taurus, 30 months, female, SAMN01093727)	31,844,692	26,709,287 (83.87%)	6,865,274 (21.56%)	160,869

Annotation reports – alignemnts

Transcript alignments

Bos_taurus_UMD_3.1.1 - Primary Assembly

Protein alignments

Bos_taurus_UMD_3.1.1 - Primary Assembly

Source	Number of sequences retrieved from Entrez	Number (%) of sequences aligned by ProSplign	Number (%) of sequences passed to Gnomon	Average % identity	Average % coverage
Homo sapiens known RefSeq (NP_)	38,283	37,404 (97.70%)	37,364 (97.60%)	77.19%	82.02%
Same-species GenBank	16,482	15,814 (95.95%)	15,804 (95.89%)	82.47%	87.45%
Same-species known RefSeq (NP_)	13,269	13,177 (99.31%)	13,164 (99.21%)	83.31%	86.69%

SAMD00000728	Cows (Japanese Black) (Bos taurus, SAMD00000728)	354,627,156	222,872,071 (62.85%)	31,485,845 (8.88%)	192,504
SAMN00120237	IVF embryos (Bos taurus, SAMN00120237)	54,352,331	27,722,770 (51.01%)	4,137,919 (7.61%)	89,792
SAMN00120238	IVF embryos (Bos taurus, SAMN00120238)	54,770,027	22,448,598 (40.99%)	2,375,125 (4.34%)	84,741
SAMN01084067	Bull Sperm Transcriptome (Bos taurus, SAMN01084067)	37,140,700	1,384,934 (3.73%)	228,029 (0.61%)	25,669
SAMN01093727	omental fat tissue (Bos taurus, 30 months, female, SAMN01093727)	31,844,692	26,709,287 (83.87%)	6,865,274 (21.56%)	160,869

Annotation reports - quality

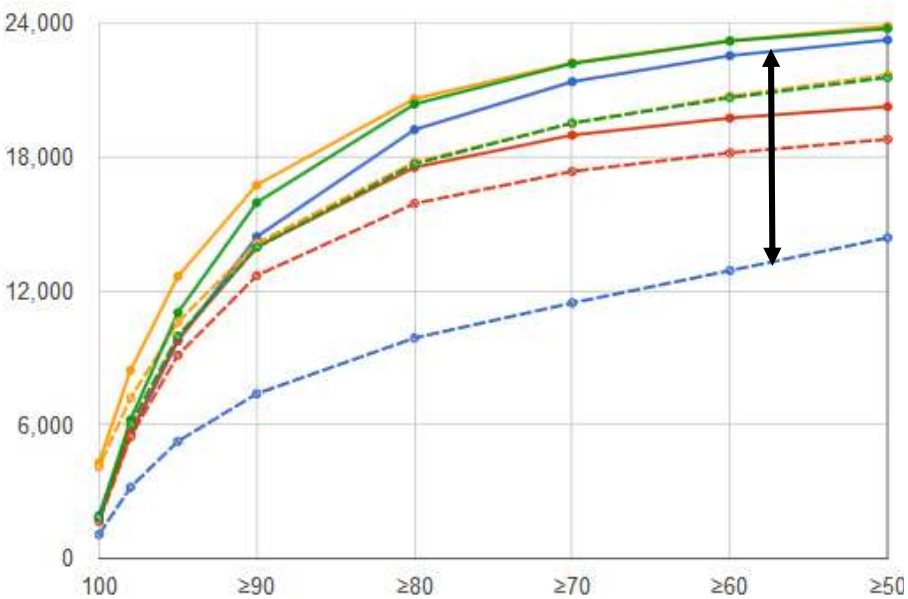
RefSeq transcript alignment quality report

The known RefSeq transcripts (NM_ and NR_ accessions) are a set of high-quality transcripts maintained by the RefSeq group at NCBI. Alignment statistics for this group of transcripts, such as percent and number of sequences not aligning at all, percent best alignments split between multiple scaffolds, and percent alignments not covering the full CDS are indicative of the genome quality and are provided below.

	Bos_taurus_UMD_3.1.1 Primary Assembly	Btau_4.6.1 Primary Assembly
Number of sequences retrieved from Entrez	14,136	14,136
Number (%) of sequences not aligning	12 (0.08%)	13 (0.09%)
Number (%) of sequences with multiple best alignments (split genes)	325 (2.31%)	295 (2.09%)
Number (%) of sequences with CDS coverage < 95%	820 (6.21%)	864 (6.52%)

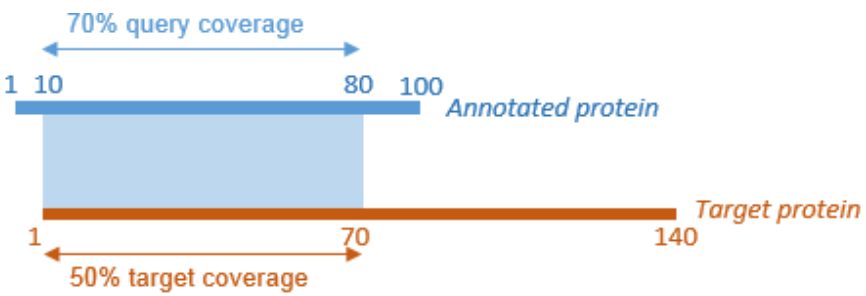
Alignments of annotated proteins to curated proteins – December 2014

Cumulative number of genes with an alignment to UniProtKB/Swiss-Prot curated proteins



Query (full) or target (dashed) percent coverage

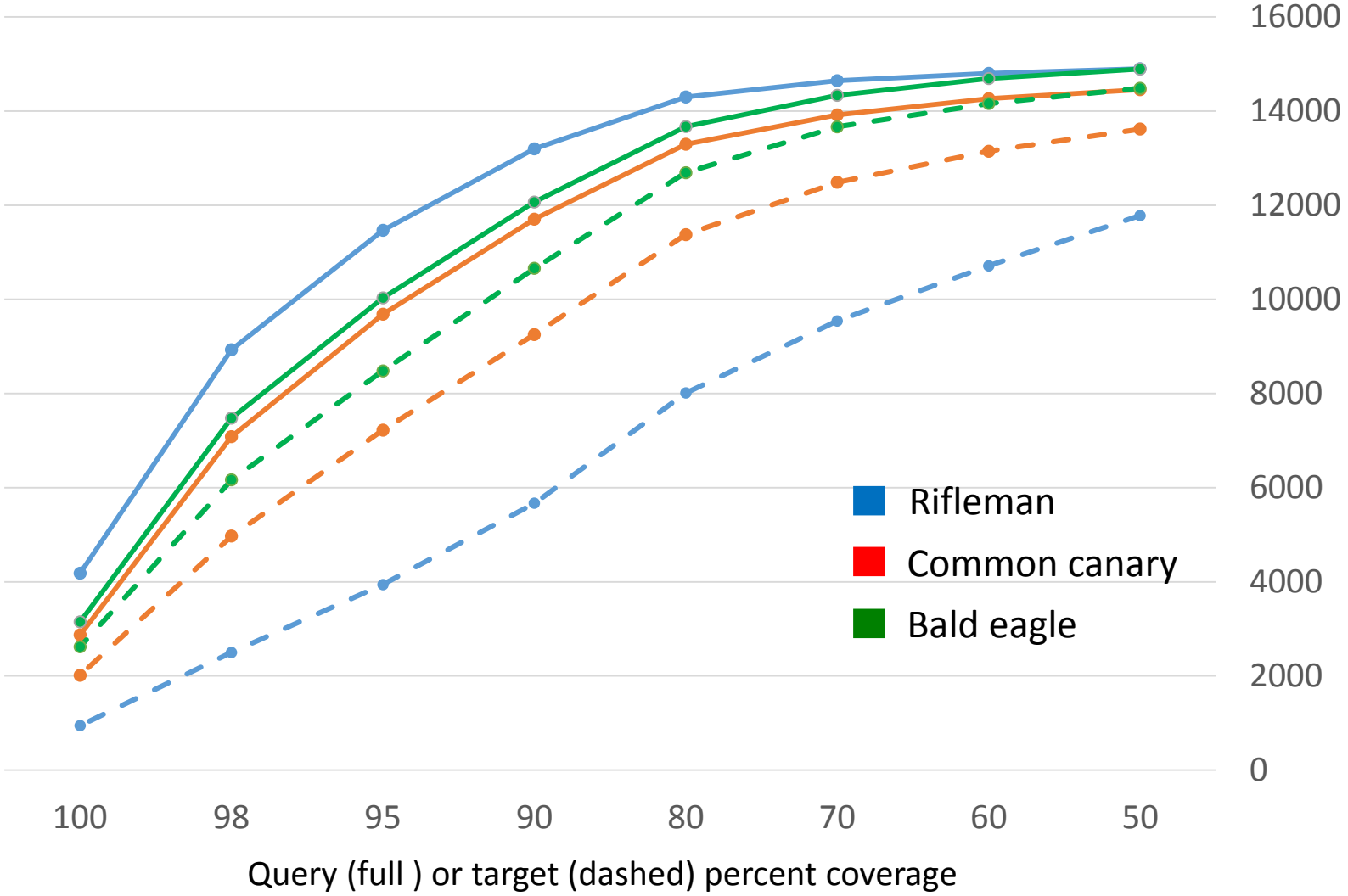
- Black rockcod – contig N50 = 17 Kb
- Tongue sole – contig N50 = 27 Kb
- Amazon molly – contig N50 = 57 Kb
- Zebrafish – contig N50 = 1,258 Kb



Birds – Why assembly quality matters

	rifleman Acanthisitta chloris	common canary Serinus canaria	bald eagle Haliaeetus leucocephalus
Assembly	ASM69581v1, GCF_000695815.1	SCA1, GCF_000534875.1	Haliaeetus_leucocephalus-4.0, GCF_000737465.1
Contig N50	20,602	53,884	105,493
Scaffold N50	64,469	17,815,079	9,145,499
Coding genes	15,064	14,775	15,227
Partial genes	8,312	1,171	629
Human orthologs	10,230	12,450	13,281

Case study - Number of genes with SwissProt hits



Request an annotation

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

NCBI Genome Resources Workshop

- 12:50– 1:15 Improving the Flow of Data to NCBI Sequence Repositories, SRA and GenBank
  Ilene Mizrachi
- 1:15– 1:40 Variation at NCBI: Resources, Tools and Submissions
  Jennifer Lee
- 1:40– 2:05 Update on the Eukaryotic Genome Annotation Pipeline
  Françoise Thibaud-Nissen
- 2:05– 2:30 From Genes to Genomes – New Features and Data Access
  Kim D. Pruitt
- 2:30– 2:50 Annual Report
  Tatiana Tatusova
- 2:50-3:00 Questions & Discussion

NCBI Booth 618
info@ncbi.nlm.nih.gov