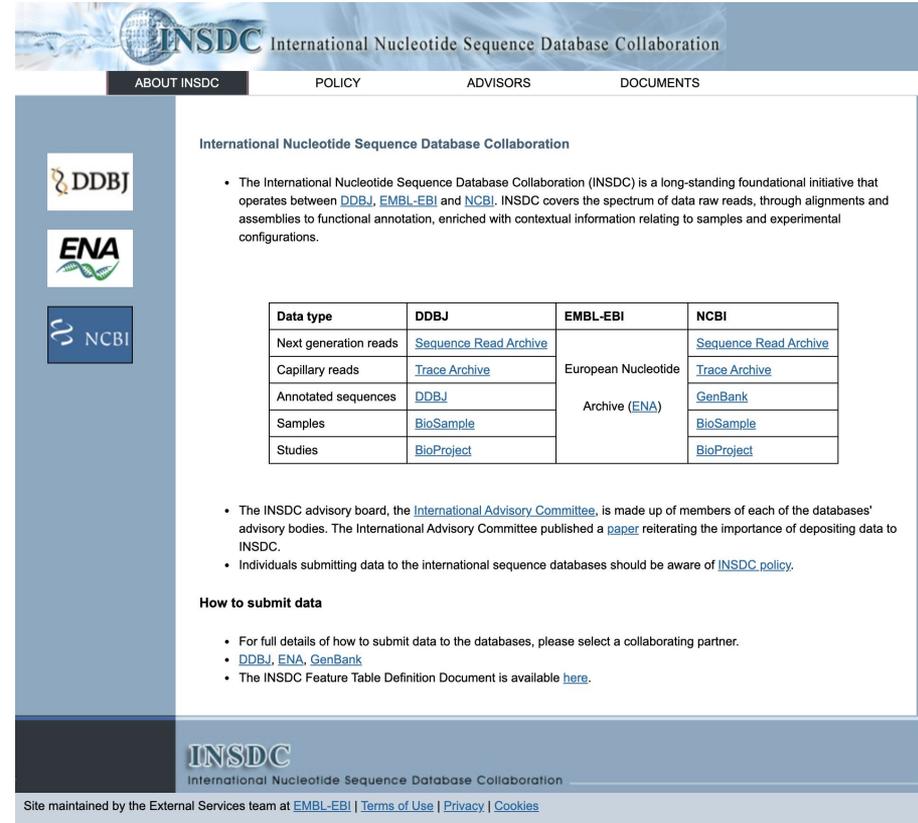


# Bases de données de séquences biologiques

*Jacques van Helden*

# International Nucleotide Sequence Database Consortium (INSDC)

- Avant de publier un article scientifique qui repose sur des séquences, les biologistes sont tenus de déposer ces séquences dans l'une des trois bases de données internationales de référence:
  - ❑ NCBI (Etats-Unis)
  - ❑ EMBL-EBI-ENA (Europe)
  - ❑ DDBJ (Japon)
- Ces bases de données se sont organisées en un consortium : International Nucleotide Sequence Database Consortium (INSDC).
- Les séquences soumises à chaque base de donnée sont automatiquement copiées dans les deux autres.



International Nucleotide Sequence Database Collaboration

ABOUT INSDC POLICY ADVISORS DOCUMENTS

International Nucleotide Sequence Database Collaboration

- The International Nucleotide Sequence Database Collaboration (INSDC) is a long-standing foundational initiative that operates between [DDBJ](#), [EMBL-EBI](#) and [NCBI](#). INSDC covers the spectrum of data raw reads, through alignments and assemblies to functional annotation, enriched with contextual information relating to samples and experimental configurations.

Data type	DDBJ	EMBL-EBI	NCBI
Next generation reads	<a href="#">Sequence Read Archive</a>	European Nucleotide Archive (ENA)	<a href="#">Sequence Read Archive</a>
Capillary reads	<a href="#">Trace Archive</a>		<a href="#">Trace Archive</a>
Annotated sequences	<a href="#">DDBJ</a>		<a href="#">GenBank</a>
Samples	<a href="#">BioSample</a>		<a href="#">BioSample</a>
Studies	<a href="#">BioProject</a>		<a href="#">BioProject</a>

- The INSDC advisory board, the [International Advisory Committee](#), is made up of members of each of the databases' advisory bodies. The International Advisory Committee published a [paper](#) reiterating the importance of depositing data to INSDC.
- Individuals submitting data to the international sequence databases should be aware of [INSDC policy](#).

**How to submit data**

- For full details of how to submit data to the databases, please select a collaborating partner.
- [DDBJ](#), [ENA](#), [GenBank](#)
- The INSDC Feature Table Definition Document is available [here](#).

INSDC  
International Nucleotide Sequence Database Collaboration

Site maintained by the External Services team at [EMBL-EBI](#) | [Terms of Use](#) | [Privacy](#) | [Cookies](#)

- Le National Center for Biotechnology Information (NCBI) est le plus grand centre international de référence pour les données biologiques.
- Via son site Web « Entrez », le NCBI donne accès à une série de bases de données pour différents types d'informations
  - ❑ Séquences nucléiques (ADN, ARN)
  - ❑ Génomes
  - ❑ Séquences protéiques
  - ❑ Taxonomie des espèces vivantes
  - ❑ Littérature biomédicale
  - ❑ ...

The screenshot shows the NCBI website homepage. At the top, there is a navigation bar with the NCBI logo, a search bar, and links for 'Resources' and 'How To'. Below the navigation bar, there is a banner for the 'UNITE' initiative, which aims to end structural racism and achieve racial equity in the biomedical research enterprise. The main content area is divided into several sections: 'Welcome to NCBI' with a brief description of the center's mission and links to 'About the NCBI', 'Mission', 'Organization', and 'NCBI News & Blog'; 'Popular Resources' listing various databases like PubMed, Bookshelf, and BLAST; 'Submit' (Deposit data or manuscripts into NCBI databases), 'Download' (Transfer NCBI data to your computer), and 'Learn' (Find help documents, attend a class or watch a tutorial); 'Develop' (Use NCBI APIs and code libraries to build applications), 'Analyze' (Identify an NCBI tool for your data analysis task), and 'Research' (Explore NCBI research and collaborative projects). A sidebar on the left contains a 'Resource List (A-Z)' with links to various categories like 'Chemicals & Bioassays', 'Data & Software', 'DNA & RNA', etc. At the bottom, there are news snippets about browser updates and a webinar.

- La section Genomes du NCBI permet d'accéder rapidement à l'information disponible pour les génomes complètement séquencés.

The screenshot displays the NCBI Genomes browser interface for the Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, complete genome. The browser shows the NCBI Reference Sequence (NC\_045512.2) and provides various tools and information.

**Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, complete genome**  
 NCBI Reference Sequence: NC\_045512.2  
[FASTA](#) [Graphics](#)

**LOCUS** NC\_045512 29903 bp ss-RNA linear VRL 18-JUL-2020  
**DEFINITION** Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, complete genome.  
**ACCESSION** NC\_045512  
**VERSION** NC\_045512.2  
**DBLINK** BioProject: [PRJNA485481](#)  
**KEYWORDS** RefSeq.  
**SOURCE** Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2)  
**ORGANISM** [Severe acute respiratory syndrome coronavirus 2](#)  
 Viruses; Riboviria; Orthornavirae; Pisuviricota; Pisoniviricetes; Nidovirales; Coronidovirineae; Coronaviridae; Orthocoronavirinae; Betacoronavirus; Sarbecovirus.

**REFERENCE** 1 (bases 1 to 29903)  
**AUTHORS** Wu, F., Zhao, S., Yu, B., Chen, Y. M., Wang, W., Song, Z. G., Hu, Y., Tao, Z. W., Tian, J. H., Pei, Y. Y., Yuan, M. L., Zhang, Y. L., Dai, F. H., Liu, Y., Wang, Q. M., Zheng, J. J., Xu, L., Holmes, E. C. and Zhang, Y. Z.  
**TITLE** A new coronavirus associated with human respiratory disease in China  
**JOURNAL** Nature 579 (7798), 265-269 (2020)  
**PUBMED** [32015508](#)

**REMARK** Erratum:[Nature. 2020 Apr;580(7803):E7. PMID: 32296181]  
**REFERENCE** 2 (bases 13476 to 13503)  
**AUTHORS** Baranov, P. V., Henderson, C. M., Anderson, C. B., Gesteland, R. F., Atkins, J. F. and Howard, M. T.  
**TITLE** Programmed ribosomal frameshifting in decoding the SARS-CoV genome  
**JOURNAL** Virology 332 (2), 498-510 (2005)  
**PUBMED** [15680415](#)

**REFERENCE** 3 (bases 29728 to 29768)  
**AUTHORS** Robertson, M. P., Igel, H., Baertsch, R., Haussler, D., Ares, M. Jr. and Scott, W. G.  
**TITLE** The structure of a rigorously conserved RNA element within the SARS virus genome  
**JOURNAL** PLoS Biol. 3 (1), e5 (2005)  
**PUBMED** [15630477](#)

**REFERENCE** 4 (bases 29609 to 29657)

**Right-hand side navigation:**  
 GenBank | Send to: | Change region shown | Customize view | Analyze this sequence | Run BLAST | Pick Primers | Highlight Sequence Features | Find in this Sequence | NCBI Virus | Retrieve, view, and download SARS-CoV-2 coronavirus genomic and sequences. | Related information | Assembly | BioProject | Protein | PubMed | Taxonomy | Full text in PMC | Gene | Genome | Identical GenBank Sequence | Mature Peptides | Other INSDC Genome Sequences | PubMed (Weighted)

- [www.uniprot.org](http://www.uniprot.org)
- KB: knowledge base
- UniProt KB vise à rassembler l'information sur toutes les séquences protéiques caractérisées par les biologistes.
- Swiss-Prot contient des séquences protéiques "annotées" par des biologistes. L'annotation consiste à associer à une séquence les connaissances résultant d'expérimentation.
  - ❑ Fonction de la protéine
  - ❑ Domaines structurels
  - ❑ Sites actifs (enzymes)
  - ❑ ...

The screenshot shows the UniProtKB interface for the entry **PODTC2 (SPIKE\_SARS2)**. The page title is "UniProtKB - PODTC2 (SPIKE\_SARS2)". The main content area displays the following information:

- Protein:** Spike glycoprotein
- Gene:** S
- Organism:** Severe acute respiratory syndrome coronavirus 2 (2019-nCoV) (SARS-CoV-2)
- Status:** Reviewed - Annotation score: ●●●●● (5 stars) - Experimental evidence at protein level<sup>i</sup>

Below the main information, there is a section titled "Function<sup>i</sup>" with the following text:

**Spike protein S1:** attaches the virion to the cell membrane by interacting with host receptor, initiating the infection. Binding to human ACE2 receptor and internalization of the virus into the endosomes of the host cell induces conformational changes in the Spike glycoprotein (PubMed:32142651, PubMed:32221306, PubMed:32075877, PubMed:32155444).

Binding to host NRP1 and NRP2 via C-terminal polybasic sequence enhances virion entry into host cell (PubMed:33082294, PubMed:33082293).

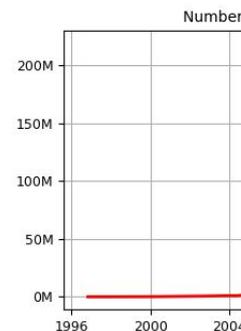
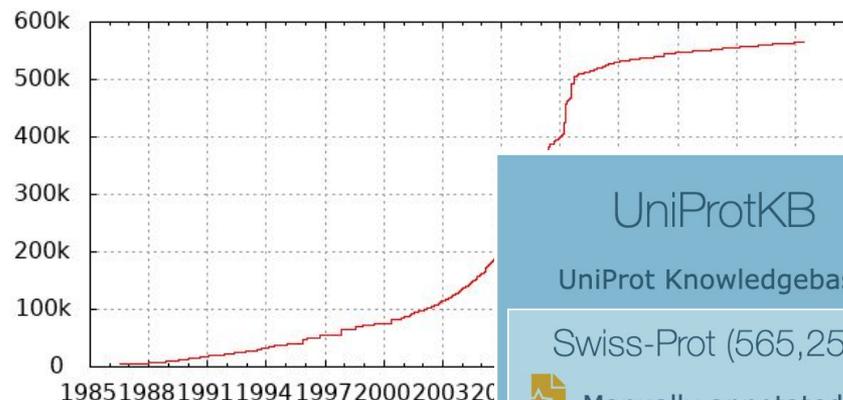
This interaction may explain virus tropism of human olfactory epithelium cells, which express high level of NRP1 and NRP2 but low level of ACE2

The left sidebar shows a list of features with checkboxes:

- Function
- Names & Taxonomy
- Subcell. location
- Pathol./Biotech
- PTM / Processing
- Expression
- Interaction
- Structure

- Deux limitations à l'annotation
  - Le nombre de publications augmente tellement qu'il n'est pas possible à l'équipe de Swiss-Prot de tout annoter
  - Le nombre de séquences augmente de façon tellement rapide qu'il est impossible de toutes les caractériser expérimentalement
- TREMBL
  - annotation automatique des séquences traduites de la base de données EMBL.
- UniProt = Swiss-Prot + TREMBL (3/11/21)
  - Swiss-Prot 565.254 séquences
  - TREMBL 219.174.961 !
- Conséquence: la vaste majorité des séquences sont annotées automatiquement, sans possibilité de les vérifier individuellement

Number of entries in UniProtKB/Swiss-Prot



## UniProtKB

UniProt Knowledgebase

Swiss-Prot (565,254)



Manually annotated and reviewed.

Records with information extracted from literature and curator-evaluated computational analysis.

TrEMBL (219,174,961)



Automatically annotated and not reviewed.

Records that await full manual annotation.