「COVID-19診断用プライマー交差性解析システム」の整備･公開

（Web公開版）

2020年5月29日

国立医薬品食品衛生研究所 遺伝子医薬部

概要

- ✓ COVID-19のPCR検査に用いるプライマーおよびプローブについては，常在菌や他の病原体の核酸配列との交差性をインシリコ解析により確認することが求められています．具体的には，プライマー/プローブの配列と他の生物種の核酸配列が相補性を有する領域において，相補結合する塩基の数の割合が80%未満であることを確認することが求められています．
- ✓ 信頼性の高いPCR診断系を迅速に構築するためには，上記の交差性を正確に，漏れなく，効率よく確認することが望まれます．しかし，PCRプライマー（20-25塩基長程度）のような短い塩基配列の検索においては，一般に広く用いられるBLAST等の検索アルゴリズムでは検索漏れが生じることが知られています．また，検索すべき生物種（COVID-19関連生物種）のゲノム配列を包括的に含んだデータベースは現状では存在しません．
- ✓ 以上の背景をふまえ，国立医薬品食品衛生研究所 遺伝子医薬部では，COVID-19関連生物種のゲノムデータベースを新規に構築し，高速塩基配列検索ソフトウェアGGGenomeによる検索サイトを公開しました．このゲノムデータベースとGGGenomeを組み合わせた「COVID-19診断用プライマー交差性解析システム」を用いることで，プライマーおよびプローブの交差性確認を正確・迅速・簡便に実施することが可能になりました．

背景

- ✓ COVID-19に関するWHO緊急使用ガイダンス（EUL文書）ならびにFDA緊急使用許可ガイダンス（EUA文書）においては，COVID-19のPCR検査における性能評価のひとつとして，COVID-19と関連する生物種との交差性（Cross-reactivity）を確認することが明示されている（資料1）．

- ✓ 具体的には，PCR検査に用いるプライマーおよびプローブについて，呼吸関連器官に存在する常在菌や病原微生物（以下，COVID-19関連生物種）の核酸配列との交差性をインシリコ解析により確認することが求められている．検索の対象となるCOVID-19関連生物種については，EUL文書およびEUA文書においてそれぞれ提示されている（資料2, 3）．

- ✓ 「交差性を有する」と判断する基準については，「SARS-CoV-2の配列に基づいて設計されたPCRプライマーあるいはプローブのいずれかについて，他の生物種の核酸配列との相補性（プライマー/プローブの結合領域において相補結合する塩基の割合）が80%以上の場合」とされている（資料1）．なお，インシリコ解析で交差性があった場合には，ウェット解析により該当する生物種の核酸配列を増幅しないことを実験的に確認することが求められる．

- ✓ 信頼性の高いPCR診断系を迅速に構築するためには，上記の交差性を正確に，漏れなく，効率よく確認することが望まれる．しかし，PCRプライマー（20-25塩基長程度）のような短い塩基配列の検索については，一般に広く用いられるBLAST等の検索アルゴリズムではパラメータ設定やデータの後処理が煩雑であり，検索漏れが生じることが知られている（資料13, 14）．また，資料2, 3に示したCOVID-19関連生物種について，包括的に検索することが可能なゲノムデータベースは現状では存在しない．

実施内容

- ✓ 以上の背景をふまえ，国立医薬品食品衛生研究所 遺伝子医薬部では，COVID-19の診断に用いるPCRプライマー/プローブの交差性解析を支援するため，「COVID-19診断用プライマー交差性解析システム」を新規に構築し，公開した．（http://www.nihs.go.jp/mtgt/covid-19info.html）.
- ✓ 具体的には，EUL文書およびEUA文書に提示されたCOVID-19関連生物種（資料2, 3）の標準ゲノム配列をRefSeqから選別・集約し（資料6），4種類のゲノムデータベースを構築した（資料4, 5：EUL掲載生物種群，EUA掲載生物種群，コロナウイルス7種，ベータコロナウイルス5種）.
- ✓ さらに，本データベースをライフサイエンス統合データベースセンター（DBCLS）が公開している高速塩基配列検索ソフトウェアGGGenome（資料7）において検索できる体制を構築した．（https://GGGenome.dbcls.jp/）
- ✓ 今回構築した「COVID-19診断用プライマー交差性解析システム」の使用手順の概要を資料8, 9に示した．また，検索結果の例を資料10-12に示した．さらに，一般的な検索法（既存のゲノムデータベースとBLASTを用いた手法）と本システム（新規に構築したゲノムデータベースとGGGenomeを用いた手法）の比較を資料14に示した．両者の比較の詳細については，補足資料（資料15-19）を御参照ください．

以上の取り組みは，GGGenomeの開発者である内藤雄樹氏（ライフサイエンス統合データベースセンター：DBCLS）の御協力の下，実施しました．この場を借りて深く感謝申し上げます．

WHO緊急使用ガイダンス（EUL文書）ならびにFDA緊急使用許可ガイダンス（EUA文書）におけるPCRプライマー等の交差性に関する記載

資料1

6.3.1.6 Analytical specificity

b) Cross reactivity

In silico analysis

- The analysis should include multiple representative strains from GenBank sequence database[7] for each organism.
- If *in silico* analysis reveals other potential cross-reactants (i.e., ≥80% homology between one of the primers or the probe to any of the sequences of listed potential cross reactants), carefully review the alignments and determine based on the positions of the homologous stretches and mismatches if additional cross-reactivity and/or interference (please refer to microbial interference studies) wet testing will be required to rule out cross-reactivity or interference of that organism that may affect the performance of your device.

EUL文書

“Instructions for Submission Requirements: In vitro diagnostics (IVDs) Detecting SARS-CoV-2 Nucleic Acid Emergency Use Listing of IVDs”

より転載

J. PERFORMANCE EVALUATION

3) Cross-reactivity (Analytical Specificity):

Cross-reactivity studies are performed to demonstrate that the test does not react with related pathogens, high prevalence disease agents and normal or pathogenic flora that are reasonably likely to be encountered in the clinical specimen. The recommended list of organisms to be analyzed in silico and by wet testing is provided in the table below. For wet testing, concentrations of 10^6 CFU/ml or higher for bacteria and 10^5 pfu/ml or higher for viruses is recommended. In silico analyses alone may be acceptable for organisms that are difficult to obtain. FDA defines in silico cross-reactivity as greater than 80% homology between one of the primers/probes and any sequence present in the targeted microorganism.

EUA文書

“Molecular Diagnostic Template for Manufacturers”

より転載

EUL文書に提示されている交差性解析の対象生物種

資料2

Table 2: Cross-Reactivity: List of Organisms to be wet-tested and/or analyzed *in silico*

Other high priority pathogens from the same virus family	Laboratory testing (wet-tested)	*In silico analysis*
Human coronavirus 229E	✓	✓
Human coronavirus OC43	✓	✓
Human coronavirus HKU1	✓	✓
Human coronavirus NL63	✓	✓
SARS-coronavirus	✓	✓
MERS-coronavirus	✓	✓
High priority organisms likely in the circulating area	✓	✓
Adenovirus (e.g. C1 Ad. 71)	✓	✓
Human Metapneumovirus (hMPV)	✓	✓
Parainfluenza virus 1-4	✓	✓
Influenza A	✓	✓
Influenza B	✓	✓
Enterovirus (e.g. EV68)	✓	✓
Respiratory syncytial virus	✓	✓
Rhinovirus	✓	✓
Chlamydia pneumonia	✓	✓
Haemophilus influenzae	✓	✓
Legionella pneumophila	✓	✓
Mycobacterium tuberculosis	✓	✓
Streptococus pneumonia	✓	✓
Streptococcus pyrogenes	✓	✓
Bordetella pertussis	✓	✓
Mycoplasma pneumoniae	✓	✓
Pneumocystis jirovecii (PJP)	✓	✓
Pooled human nasal wash - *to represent diverse microbial flora in the human respiratory tract*	✓	×
Influenza C	×	✓
Parechovirus	×	✓
Candida albicans	×	✓
Corynebacterium diphtheriae	×	✓
Legionella non-pneumophila	×	✓
Bacillus anthracosis (Anthrax)	×	✓
Moraxella cararrhalis	×	✓
Neisseria elongate and miningitidis	×	✓
Pseudomonas aeruginosa	×	✓
Staphylococcus epidermis	×	✓
Staphylococcus salivarius	×	✓
Leptospirosis	×	✓
Chlamydia psittaci	×	✓
Coxiella burneti (Q-Fever)	×	✓
Streptococcus aureus	×	✓

EUL文書

“Instructions for Submission Requirements: In vitro diagnostics (IVDs) Detecting SARS-CoV-2 Nucleic Acid Emergency Use Listing of IVDs”

より転載

EUA文書に提示されている交差性解析の対象生物種

資料3

Recommended List of Organisms to be Analyzed *in silico* and by Wet Testing

Other high priority pathogens from the same genetic family	High priority organisms likely in the circulating area
Human coronavirus 229E	Adenovirus (e.g. C1 Ad. 71)
Human coronavirus OC43	Human Metapneumovirus (hMPV)
Human coronavirus HKU1	Parainfluenza virus 1-4
Human coronavirus NL63	Influenza A & B
SARS-coronavirus	Enterovirus (e.g. EV68)
MERS-coronavirus	Respiratory syncytial virus
	Rhinovirus
	Chlamydia pneumoniae
	Haemophilus influenzae
	Legionella pneumophila
	Mycobacterium tuberculosis
	Streptococcus pneumoniae
	Streptococcus pyogenes
	Bordetella pertussis
	Mycoplasma pneumoniae
	Pneumocystis jirovecii (PJP)
	Pooled human nasal wash - *to represent diverse microbial flora in the human respiratory tract*
	Candida albicans
	Pseudomonas aeruginosa
	Staphylococcus epidermis
	Streptococcus salivarius

EUA文書 “Molecular Diagnostic Template for Manufacturers” より転載

COVID-19関連生物種ゲノムデータベースに含まれる生物種一覧

資料4

Organisms	division	Genus/ Species	Cross-reactivity check database for COVID-19 diagnostic primers, NIHS 2020/5/1			
			7 CoV + 32 organisms listed in WHO EUL	7 CoV + 21 organisms listed in U.S. EUA	7 CoV	5 beta-CoV
SARS-CoV-2	VRL		+	+	+	+
Human coronavirus 229E	VRL		+	+	+	
Human coronavirus OC43	VRL		+	+	+	+
Human coronavirus HKU1	VRL		+	+	+	+
Human coronavirus NL63	VRL		+	+	+	
SARS-coronavirus	VRL		+	+	+	+
MERS-coronavirus	VRL		+	+	+	+
Adenovirus	VRL		+	+		
Human Metapneumovirus (hMPV)	VRL		+	+		
Parainfluenza 1 - 4	VRL		+	+		
Influenza A	VRL		+	+		
Influenza B	VRL		+	+		
Enterovirus	VRL		+	+		
Respiratory Syncytial Virus	VRL		+	+		
Rhinovirus	VRL		+	+		
Chlamydophila pneumoniae	BCT	Species	+	+		
Haemophilus influenzae	BCT	Species	+	+		
Legionella pneumophila	BCT	Species	+	+		
Mycobacterium tuberculosis	BCT	Species	+	+		
Streptococcus pneumoniae	BCT	Species	+	+		
Streptococcus pyogenes	BCT	Species	+	+		
Bordetella pertussis	BCT	Species	+	+		
Mycoplasma pneumoniae	BCT	Species	+	+		
Pneumocystis jirovecii (PJP)	PLN		+	+		
Influenza C	VRL		+			
Parechovirus	VRL		+			
Candida albicans	PLN		+	+		
Corynebacterium diphtheriae	BCT	Species	+			
Legionella (non-pneumophila)	BCT	Genus	+			
Bacillus anthracis (Anthrax)	BCT	Species	+			
Moraxella catarrhalis	BCT	Species	+			
Neisseria elongata and Neisseria meningitidis	BCT	Species	+			
Pseudomonas aeruginosa	BCT	Species	+	+		
Staphylococcus epidermidis	BCT	Species	+	+		
Streptococcus salivarius	BCT	Species	+	+		
Leptospira sp.	BCT	Genus	+			
Chlamydophila psittaci	BCT	Species	+			
Coxiella burnetii (Q-Fever)	BCT	Species	+			
Staphylococcus aureus	BCT	Species	+			

COVID-19関連生物種ゲノムデータベースの概要

資料5

- ✓ COVID-19診断用PCRプライマー/プローブの交差性解析の対象とされている生物種（資料2, 3）について, NCBI RefSeqに登録されているゲノム配列データを取得し（資料6）, 下記のカテゴリ毎に集約した4種のデータベースを構築した.
- ✓ 各データベースには, 検索の際のポジティブコントロールとして, SARS-CoV-2のゲノム配列を含めている.
- ✓ 各データベースに収載されている生物種を資料4に示した. また, 収載されたゲノム配列のID等の詳細は別途エクセルファイルにとりまとめた（別表1）.

データベースの名称	登録されている生物種の分類	登録されている生物種の種類･数
Cross-reactivity check for COVID-19 diagnostic primers (7 CoV + 32 organisms in WHO EUL), NIHS 2020/5/1	EUL文書（WHOガイダンス）に提示された生物種	コロナウイルス7種 + 上記以外の生物種32種
Cross-reactivity check for COVID-19 diagnostic primers (7 CoV + 20 organisms in US EUA), NIHS 2020/5/1	EUA文書（FDAガイダンス）に提示された生物種	コロナウイルス7種 + 上記以外の生物種20種
Cross-reactivity check for COVID-19 diagnostic primers (7 CoV), NIHS 2020/5/1	コロナウイルス	コロナウイルス7種
Cross-reactivity check for COVID-19 diagnostic primers (5 beta-CoV), NIHS 2020/5/1	β コロナウイルス	β コロナウイルス5種

ゲノム配列の取得について

資料6

COVID-19関連生物種（交差性解析の対象生物種）のゲノム配列は，NCBI RefSeqから取得した．対象生物種はNCBIのデータベースにおいて，VRL（ウィルス），BCT（バクテリア），PLN（植物・真菌類など）の3つのカテゴリー（Division）に分類された．ゲノム配列は各カテゴリーごとに下記の方針で取得した．

VRL（ウィルス）に分類される生物種：
各ウィルスのNCBI Taxonomy ID（taxid：脚注参照）配下に含まれる全ての生物種について，RefSeqに登録されたゲノム配列を取得した．
（例：資料4の「Adenovirus」については，Human mastadenovirus A～Gのゲノム配列を取得した）

PLN（植物・真菌類など）に分類される生物種および
BCT（バクテリア）に分類される生物種のうち，種レベルまで指定されている生物種：
VRLと同じ手法では登録配列数が多すぎるため（例：資料4の「Pseudomonas aeruginosa」では約90万配列が登録されている），各生物種のtaxidに対応する生物種の参照ゲノム配列（reference genome）あるいはNCBI Genomeで定義されている代表株のゲノム配列（representative genome）を取得した．

BCT（バクテリア）に分類される生物種のうち，属レベルまでしか指定されていない生物種：
「属」は特定の種を規定しないので，ゲノム配列が存在しない．そこで，NCBI Taxonomyにおいて指定された属に含まれる種の一覧を取得し，各種の代表株のゲノム配列を選択した．

NCBI Taxonomy：主としてNCBIの塩基配列データベースに登録されている遺伝子が由来する生物種に関して，系統分類と学名に関する情報が収集されたデータベース．

Taxid：上記のNCBI Taxonomyにおいて，系統分類の各階層に振られているID．

資料7

GGGenomeの概要

超絶高速ゲノム配列検索
GGGenome

https://GGGenome.dbcls.jp/

- ✓ ライフサイエンス統合データベースセンター（DBCLS）の内藤雄樹氏が開発した高速塩基配列検索ソフトウェア．PCRプライマーやプローブ等の比較的短い塩基配列についても，見落としなく高速に検索することができる．
- ✓ 相補結合領域に存在するミスマッチや挿入・欠失を加味した検索が可能であり（資料9参照），かつ，検索漏れがない（BLASTなどの他の配列検索プログラムでは検索条件の設定が煩雑で，検索漏れが生じる可能性がある）．
- ✓ 国内における核酸医薬の承認審査において，オフターゲット候補遺伝子の検索ツールとして活用されている．
- ✓ GenBank/ENA/DDBJ国際塩基配列データベースに登録されたウイルス由来の全配列を横断検索できる．
- ✓ 各国から登録されている新型コロナウイルスゲノムを横断検索することも可能（2020年3月に整備）．2020年5月19日現在，3156配列*が登録されており，半自動的にアップデートされる．変異箇所の情報も最新情報を取得可能．（*各研究グループがGenBank/ENA/DDBJに登録したSARS-CoV-2ゲノム配列を取得）
- ✓ ウェブ版は誰でも無償で利用可能．検索する配列を秘匿したい場合には，パッケージ版を用いてオフラインで検索することが可能（https://gggenome.retrieva.jp/）．

GGGenomeの検索画面(操作の概要)

資料8

超絶高速ゲノム配列検索 GGGenome

Help | English

AAATTTTGGGGACCAGGAAC 検索

データベース：

CoV

SARS-CoV-2 complete genomes, GenBank 2020/5/19
Cross-reactivity check for COVID-19 diagnostic primers (7 CoV + 32 organisms in WHO EUL), NIHS 2020/5/1
Cross-reactivity check for COVID-19 diagnostic primers (7 CoV + 20 organisms in US EUA), NIHS 2020/5/1
Cross-reactivity check for COVID-19 diagnostic primers (7 CoV), NIHS 2020/5/1
Cross-reactivity check for COVID-19 diagnostic primers (5 beta-CoV), NIHS 2020/5/1

○ ミスマッチ/ギャップを許容 ◉ ミスマッチのみ許容 ： 4 塩基まで (検索する配列長の25%まで)
◉ 双方向を検索 ○ ＋方向のみ検索 ○ －方向のみ検索

① プライマー配列を入力.

② 生物種選択のメニューを「cov」等で検索することで, CoV関連データベースを選択可能.

③ 検索条件を入力.
（資料9を参照）

Results:

検索語に色がつきます（ミスマッチ・挿入欠失）。

NC_045512.2 Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, complete genome
position: 29125-29144 ▼29125
CGGCAGACGTGGTCCAGAACAAACCCAAGGAAATTTTGGGGACCAGGAACTAATCAGACAAGGAACTGATTACAAACATT

NC_004718.3 SARS coronavirus, complete genome
position: 28974-28993 ▼28974
TGGGAGACGTGGTCCAGAACAAACCCAAGGAAATTTCGGGGACCAAGACCTAATCAGACAAGGAACTGATTACAAACATT

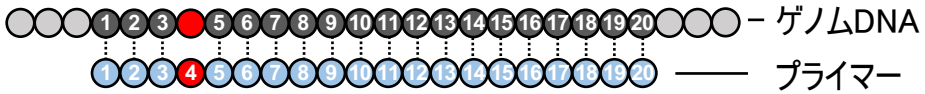
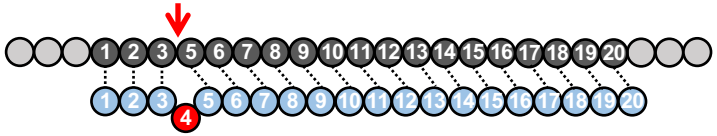
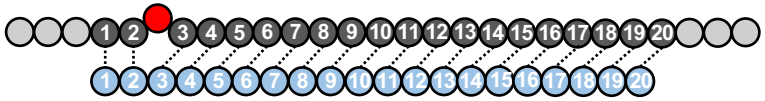
NC_002929.2 Bordetella pertussis Tohama I, complete genome
position: 374624-374643 ▼374624
TGCAGGGCGGATTGGGTCATGACTATTCGATTACTTTCGGGACCAGGAACAGGCCGTCGGCCGATTCGGGGGCGTTGGCC

NC_002929.2 Bordetella pertussis Tohama I, complete genome
position: 1861704-1861723 ▼1861704
GTTCTATTACATGCCGCTCGAACATGCCGAAGATTTGGCGGCCCAGGAACAGTGTGTAAGCCTGATGGCGCCGCTGCACG

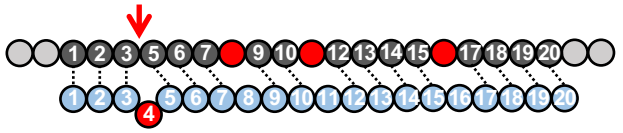
④ 結果が表示される.

相補結合領域における不適合性の分類

資料9

不適合性の種類	相補結合領域の概念図
ミスマッチ (mis)	
挿入 (ins)	
欠失 (del)	

不適合箇所の数の和 = d (distance)*



例：ミスマッチ3箇所 + 挿入1箇所 → d=4

*Yoshida T, Naito Y *et al, Genes to Cells*, 24,12, 827-835 (2019)

- ✓ GGGenomeの検索条件において，「ミスマッチのみ許容」を選択すると，プライマーとの相補結合領域にミスマッチを含む領域を検索することができる．
- ✓ 「ミスマッチ/ギャップを許容」を選択すると，相補結合領域のミスマッチに加え，ギャップ（挿入や欠失）を含む領域を検索することが可能．この条件で，例えば「4塩基まで（d≦4）」と指定すると，相補結合領域におけるミスマッチ部位，挿入部位ならびに欠失部位の総和（不適合箇所の総和：d値*）が「4」までの配列を検索することができる（右上に示したような配列がヒットする）．
- ✓ 20塩基長のプライマーについて，80%以上の相補性（=20%未満の不適合性）の有無を検証する場合には，不適合箇所が4以下のゲノム配列を検索することになるため（20×0.2=4塩基長以内），「4塩基まで」を許容した検索を実施するのが妥当と考えられる．
- ✓ ミスマッチに加え，ギャップ（挿入や欠失）を考慮した検索が必要であるかについては，今後検討を要する．
  （相補結合領域に挿入や欠失がある場合に，どの程度PCR反応に影響を及ぼすかを検証する必要がある）

GGGenomeを用いた検索例

資料10

米国，中国，フランスのCOVID−19診断用PCRプライマー（計12本）を用いて，EUL文書（WHO緊急使用ガイダンス）に提示された生物種のデータベース〔**Cross-reactivity check for COVID-19 diagnostic primers (7 CoV + 32 organisms in WHO EUL)**〕に対して，「ミスマッチのみ許容」，「4塩基まで（のミスマッチを許容）」で検索.

相補結合領域に含まれるミスマッチの数

開発国	プライマーの名称	プライマーの長さ	ヒット配列の数 0	1	2	3	4	ヒット配列の例（相補結合領域のアライメント）		ヒットした生物種 ミスマッチ数 交差性
米国	2019-nCoV_N1-F	20	1	0	1	85	744	GATCCCAAGATCAGGGAAAC || ||||| ||||| |||| GACCCCAAAATCAGCGAAAT	NC_002929.2 Primer(+鎖)	百日咳菌 4 80%
	2019-nCoV_N1 R	24	1	0	1	0	5	CAGATTCAACTGACAATAACCAGA |||||||||||| || |||||||| CAGATTCAACTGGCAGTAACCAGA	NC_004718.3 Primer(−鎖)	SARSコロナ 2 91.7%
	2019-nCoV_N2-F	20	2	1	7	93	938	TTACTAACATTGCCGGCAAA |||| ||||||| | ||||| TTACAAACATTGGCCGCAAA	NC_000912.1 Primer(+鎖)	マイコプラズマ 3 85%
	2019-nCoV_N2-R	18	1	0	14	325	4218	TTCTTTGGAATGTCACGC ||||| |||||||| ||| TTCTTCGGAATGTCGCGC	NC_004718.3 Primer(−鎖)	SARSコロナ 2 88.9%
	2019-nCoV_N3-F	22	1	1	0	1	38	GGGAGCCTTGAATACACCCAAA |||||||||||||||||| ||| GGGAGCCTTGAATACACCAAAA	NC_004718.3 Primer(+鎖)	SARSコロナ 1 95.5%
	2019-nCoV_N3-R	21	1	0	1	1	82	CAATGCTGCCACCGTGCTACA ||||||||| | ||||||||| CAATGCTGCAATCGTGCTACA	NC_004718.3 Primer(−鎖)	SARSコロナ 2 90.5%

上側の配列：ヒットしたCOVID−19関連生物種のゲノム配列
下側の配列：プライマーの配列（+鎖 or −鎖）

GGGenomeを用いた検索例（続き）

資料11

開発国	プライマーの名称	プライマーの長さ	ヒット配列の数 0	1	2	3	4	ヒット配列の例（相補結合領域のアライメント）		ヒットした生物種 ミスマッチ数 交差性
中国	CN-CDC_primer1	21	1	0	1	10	96	CCCAGTGGGTTTTACACTTAG ||| |||||||||||||| CCCTGTGGGTTTTACACTTAA	NC_004718.3 Primer(+鎖)	SARSコロナ 2 90.5%
	CN-CDC_primer2	19	1	0	7	144	1250	TCAGCTGAAGAACTATCGT |||||||| | || ||||| TCAGCTGATGCACAATCGT	NW_017264778.1 Primer(−鎖)	ニューモシスチス 3 84.2%
	CN-CDC_primer4	22	1	0	1	6	67	GGGGAAATTCTCCTGCTCGAAT |||||| ||||||||| |||| GGGGAACTTCTCCTGCTAGAAT	NC_004718.3 Primer(+鎖)	SARSコロナ 2 90.9%
	CN-CDC_primer5	22	1	0	1	2	23	CAGCTTGAGAGCAAAGTTCTG |||||||||||||| | |||| CAGCTTGAGAGCAAAATGTCTG	NC_004718.3 Primer(−鎖)	SARSコロナ 2 90.9%
フランス	FR-Pasteur nCoV_IP2-12669Fw	18	1	1	17	290	3969	ATGAGGTTATTCCAGTTG ||||| ||| ||| |||| ATGAGCTTAGTCCTGTTG	NW_017264787.1 Primer(+鎖)	ニューモシスチス 3 83.3%
	FR-Pasteur nCoV_IP2-12759Rv	18	1	2	26	464	5583	ACAACACAACCAAAGGAG |||||||||| || |||| ACAACACAACAAAGGGAG	NC_002017.1 Primer(−鎖)	インフルエンザ 2 88.9%
	FR-Pasteur nCoV_IP4-14059Fw	19	1	0	11	153	1685	GGTAAATGGTATGATTTTG ||||| ||||||||| | GGTAACTGGTATGATTTCG	NC_006213.1 Primer(+鎖)	コロナOC43 2 89.4%
	FR-Pasteur nCoV_IP4-14146Rv	20	1	0	2	15	313	CTTAAATTCACCTTTACCAG | || ||| ||||| ||||| CCTATATTAACCTTGACCAG	NC_000912.1 Primer(−鎖)	マイコプラズマ 4 80%

資料12参照

GGGenome：検索条件の比較

資料12

- ✓ GGGenomeの検索条件として，「ミスマッチのみ許容」と「ミスマッチ/ギャップを許容」のいずれかを選択することができる（資料9参照）．
- ✓ この検索条件の違いが結果にどのように反映されるかを例示するため，資料11に示したフランスのプライマー配列（FR-Pasteur nCoV_IP4-14059Fw：19mer,下から2段目）を用いて，資料10,11と同じデータベース〔**Cross-reactivity check for COVID-19 diagnostic primers (7 CoV + 32 organisms in WHO EUL)**〕に対して，「ミスマッチ/ギャップを許容」，「4塩基まで」の条件で検索した（下表）．
- ✓ 「ミスマッチのみ許容」，「4塩基まで」の条件で検索した結果（資料11，下から2段目）と比較すると，ギャップ（挿入・欠失）まで含めて検索した場合には，ヒット配列の数が大きく増大することが確認される（下表：青字）．以下に例示したようなヒット配列は「ミスマッチのみ許容」では得られないため，交差性解析を行う場合には，「ミスマッチ/ギャップを許容」を選択した方が，より正確な解析ができると考えらえる．

相補結合領域に含まれるミスマッチ/ギャップの数

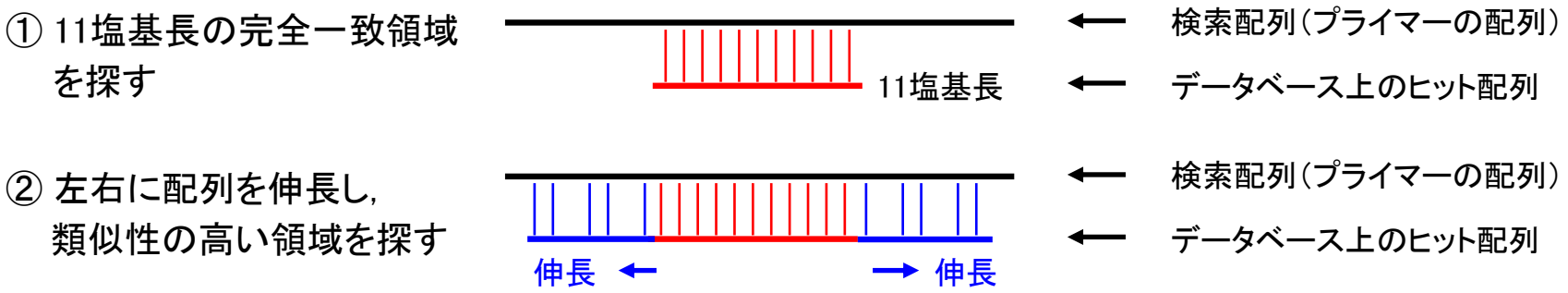
資料9を参照

プライマー情報（開発国・名称・長さ）	ヒット配列の数 0	1	2	3	4	ヒット配列の例（相補結合領域のアライメント）		ヒットした生物種 ミスマッチ/ギャップの数 交差性	d値
フランス FR-Pasteur nCoV_IP4-14059Fw (19mer)	1	0	30	726	13580	GGTAA-TGG-ATGATTTCG ||||| ||| |||||||| GGTAACTGGTATGATTTCG	NZ_CP004006.1 Primer(+鎖)	レジオネラ 挿入2 89.4%	2
						GGTAACAGGTA-GATTTCG |||||| |||| ||||||| GGTAACTGGTATGATTTCG	NZ_KL370759.1 Primer(+鎖)	レジオネラ ミスマッチ1+ 挿入1 89.4%	2
						GGTAACTAGTATCGATTTCG ||||||| |||| ||||||| GGTAACTGGTAT-GATTTCG	NZ_NPDX01000001.1 Primer(+鎖)	レプトスピラ ミスマッチ1 + 欠失1 90%	2

BLAST（Basic Local Alignment Search Tool）について

資料13

- ✓ 一般的に利用されている代表的な配列類似性検索プログラム．アミノ酸配列を検索する「BLASTp」と核酸配列を検索する「BLASTn」がある（以降では，BLASTnについて記載する）． NCBIが提供するBLASTには，広く汎用されているウェブ版（NCBI blast：一般用）とスタンドアローン版（NCBI blast+：習熟者用）がある．
- ✓ BLASTの原理は，①連続する11塩基長の完全一致領域をまず検索・特定し（下図：赤），②左右にアライメントを伸長しながら検索することで（下図：青），配列類似性の高い領域を特定する．「配列類似性の高い領域」の類似性の程度は，別途パラメータで調整可能．
- ✓ ①の完全一致領域の初期設定値は「11塩基長」であるが，「NCBI blast」では15塩基長あるいは7塩基長に変更することが可能．一方，「NCBI blast+」では4塩基長以上の任意の長さを選択可能．
- ✓ 完全一致領域の塩基長の値を小さくすると，膨大な数の検索結果（ヒット配列）が得られ、検索に要する時間も長くなる．（検索結果の解釈・整理に労力と時間を要する）
- ✓ 連続する11塩基長（あるいは，最小の4塩基長）の完全一致領域が存在しない配列はヒットしない．（プライマーのような短い配列では検索漏れが生じやすい）



BLASTとGGGenomeの比較（まとめ）

-プライマーの交差性解析プログラムとしての観点から-

資料14

項目	NCBI blast	NCBI blast+	GGGenome
COVID-19関連生物種のゲノムデータベース	個別に生物種を入力（一部指定不可）	独自に構築が必要	構築済み（今回の取り組み：国立衛研）
操作方法	平易（マウス操作で指定）	習熟が必要（コマンド入力）	平易（マウス操作で指定）
得られるアライメントの領域	プライマー配列の一致部分の領域のみ	プライマー配列の一致部分の領域のみ	プライマー配列の全体領域
交差性の評価	困難	困難	容易
検索に要する時間	やや長い（*数分）	長い（*1時間以上）	短い（*数十秒）
検索結果の数（ヒット件数）	多い（*約12万件）7塩基以上連続一致を含む配列のみ	非常に多い（*約4200万件）4塩基以上連続一致を含む配列のみ	適切（*約4万件）
検索漏れの可能性	有	有	無
検索結果の取り扱い	労力を要する（後処理に手間がかかる）	労力を要する（後処理に手間がかかる）	容易
指定（習熟）が必要なパラメータの数	多い	多い	少ない
結果に影響するプログラムの制限	完全一致塩基長：7以上（「不適合箇所」のパラメータはない）	完全一致塩基長：4以上（「不適合箇所」のパラメータはない）	不適合箇所の割合：25%まで（「完全一致塩基長」のパラメータはない）
最大ヒット件数	2万件	無制限	10万件
タブ区切り出力の方法	検索実行後にリンククリック	検索実行時にオプション指定	検索実行後にリンククリック
計算機性能への依存度	低い	高い	低い

* 例として，EUL掲載COVID-19関連生物種に対して「資料18，検索条件２」で検索した結果を提示

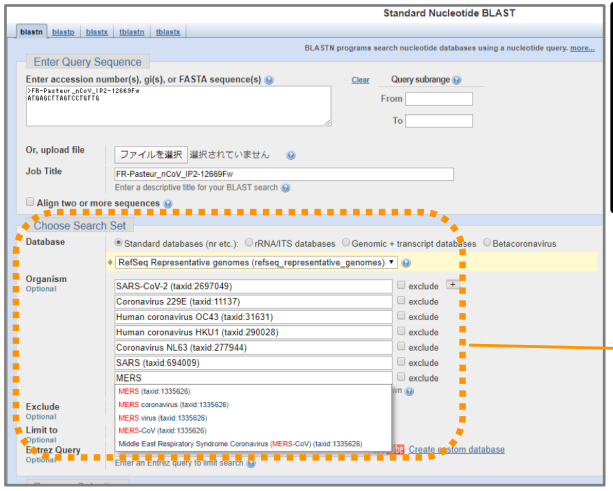
補足資料

プライマー交差性確認（類似配列検索）手順の比較
1：データベース

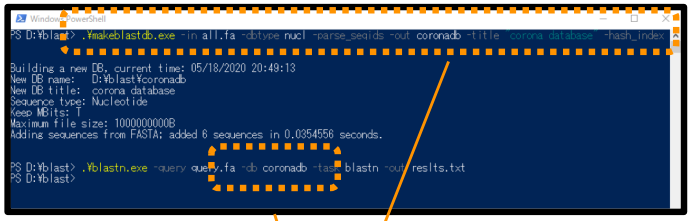
資料15

操作	NCBI blast	NCBI blast+	GGGenome
データベースの準備	不要： 既存のデータベースを活用. ただし, 配列の取捨選択が必要.	必要： 対象生物種の配列を収集し, コマンド入力によりデータベースを構築.	不要： 今回の取り組みで専用データベースを構築済み.
データベースの選択	ブラウザで既存のデータベースから対象生物種を配列を個別に選択.	構築したデータベースを検索時のオプションで指定.	ブラウザで構築したデータベースを選択.
特記事項	操作は平易だが, 数十の生物種の選択・指定に手間がかかる. リスト記載の生物種名のままでは検索できない生物種もある.	配列収集には核酸配列データベースに関する知識が必要. コマンド入力による操作が必要なため, 操作には習熟が必要.	任意の配列で新規にデータベースを構築する場合は, GGGenomeのパッケージ版（有償）を利用する必要がある

NCBI blast

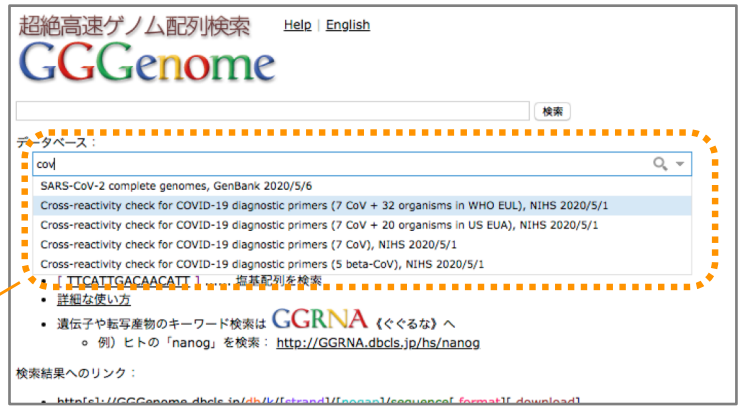


NCBI blast+



データベースを選択・指定する場所

GGGenome

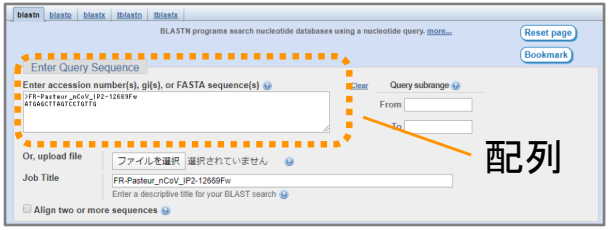


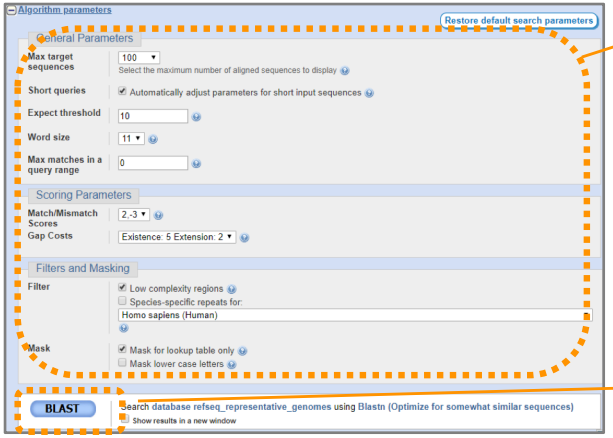
プライマー交差性確認（類似配列検索）手順の比較
2：検索パラメータ

資料16

操作	NCBI blast	NCBI blast+	GGGenome
配列の指定	配列をブラウザ上に張り付ける.	fasta形式ファイルを作成し，実行時にファイル名を指定.	配列をブラウザ上に張り付ける.
パラメータの指定	ブラウザ上で必要なパラメータを指定.	実行時のオプションで指定.	ブラウザ上で必要なパラメータを指定.
特記事項	操作は平易だが，アルゴリズムを理解していないと設定できないパラメータが多い.	コマンド入力による操作が必要なため，操作には習熟が必要. アルゴリズムを理解していないと設定できないパラメータが多い.	ウエット系の研究者が直感的に理解可能な，少数のパラメータのみを設定.

NCBI blast





NCBI blast+

配列（ファイル）　パラメータ

GGGenome

GGGenomeで指定するパラメータ

1. 「ミスマッチ/ギャップを許容」か「ミスマッチのみ許容」を選択.
2. 許容するミスマッチ/ギャップの塩基数の最大値を指定.
3. 検索するstrandを指定.（プライマーの場合は「双方向」の検索が必要）

プライマー交差性確認(類似配列検索)手順の比較
3:検索結果の表示

資料17

操作	NCBI blast	NCBI blast+	GGGenome
検索結果	ブラウザ上にグラフィカルに表示される.	指定ファイルに出力される.	ブラウザ上にグラフィカルに表示される.
タブ区切り出力	検索実行後にダウンロード	検索実行時に指定	検索実行後にダウンロード
特記事項	相補性が高い領域(連続一致領域と伸長できた範囲)のみが表示される. プライマー全長のアライメントは表示されない.	相補性が高い領域(連続一致領域と伸長できた範囲)のみが表示される. プライマー全長のアライメントは表示されない.	プライマー全長のアライメントが得られる. 末端にミスマッチがあっても表示される.

NCBI blast+:デフォルトの出力形式を例示

```
>NC_045512.2 Severe acute respiratory syndrome coronavirus 2 isolate
Wuhan-Hu-1, complete genome
Length=29903
Score = 33.7 bits (36),  Expect = 0.11,  Identities = 18/18 (100%),
Gaps = 0/18 (0%),  Strand=Plus/Plus
Query  1      ATGAGCTTAGTCCTGTTG  18
              ||||||||||||||||||
Sbjct  12690  ATGAGCTTAGTCCTGTTG  12707
```

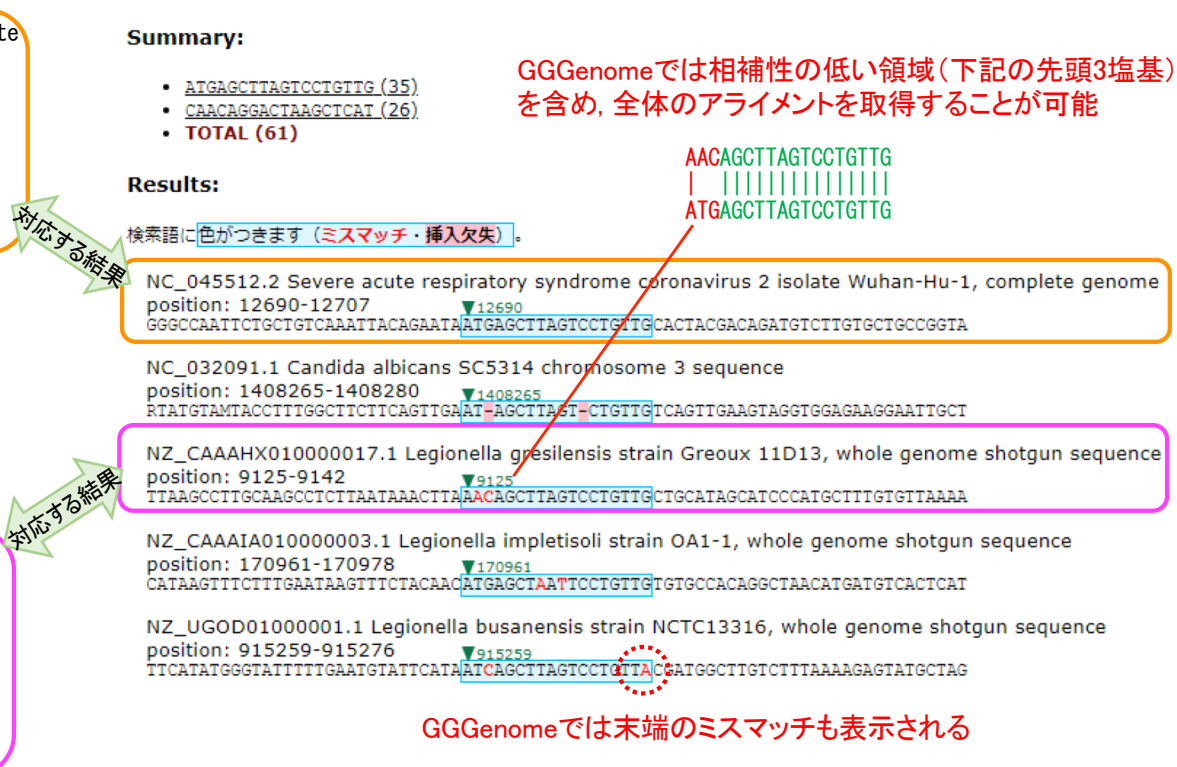
SARS-CoV-2そのもの(完全一致)

```
>NZ_UGOD01000001.1 Legionella busanensis strain NCTC13316, whole
genome shotgun sequence
Length=3414911
Score = 28.3 bits (30),  Expect = 4.8,  Identities = 15/15 (100%),
Gaps = 0/15 (0%),  Strand=Plus/Plus
Query  4        AGCTTAGTCCTGTTG  18
                |||||||||||||||
Sbjct  1107996  AGCTTAGTCCTGTTG  1108010
```

```
>NZ_CAAAHX010000017.1 Legionella gresilensis strain Greoux 11D13,
whole genome shotgun sequence
Length=41723
Score = 28.3 bits (30),  Expect = 4.8,  Identities = 15/15 (100%),
Gaps = 0/15 (0%),  Strand=Plus/Plus
Query  4      AGCTTAGTCCTGTTG  18
              |||||||||||||||
Sbjct  9128   AGCTTAGTCCTGTTG  9142
```

BLASTではアライメントの一部(緑:15塩基長)しか表示されない

GGGenome:デフォルトの出力形式を例示



GGGenomeでは相補性の低い領域(下記の先頭3塩基)を含め,全体のアライメントを取得することが可能

GGGenomeでは末端のミスマッチも表示される

プライマー交差性確認（類似配列検索）手順の比較
4：所要時間と検索結果数

資料18

	NCBI blast	NCBI blast+	GGGenome
検索条件1	✓ 検索配列：FR-Pasteur_nCoV_IP2-12669Fw（ATGAGCTTAGTCCTGTTG, 18mer） ✓ データベース：EUL掲載COVID-19関連生物種（NCBI blastは指定可能な生物種のみ） ✓ パラメータ：各プログラムのデフォルト値		
	最短完全一致塩基長：11 e-value:1000（実行時に自動調整）	最短完全一致塩基長：11 e-value:10	完全一致のみ検索 +/-の両strandを検索
検索時間/検索結果数	数十秒/254件	1秒未満/3件	1秒未満/1件
検索条件2	✓ 検索配列：FR-Pasteur_nCoV_IP2-12669Fw（ATGAGCTTAGTCCTGTTG, 18mer） ✓ データベース：EUL掲載COVID-19関連生物種（NCBI blastは指定可能な生物種のみ）		
	最短完全一致塩基長：7（最小値） e-value：100万	最短完全一致塩基長：4（最小値） e-value：500万	不適合箇所「4」まで許容 ミスマッチ/ギャップを許容 +/-の両strandを検索
検索時間/検索結果数	数分（状況による）/約12万件	約72分/約4200万件	数十秒/約4万件
検索条件3	✓ 検索配列：NIID_WH-1_R913（CTTTACCAGCACGTGCTAGAAGG, 23mer） ✓ データベース：EUL掲載COVID-19関連生物種（NCBI blastは指定可能な生物種のみ）		
	最短完全一致塩基長：7（最小値） e-value：100万	最短完全一致塩基長：4（最小値） e-value：500万	不適合箇所「5」まで許容 ミスマッチ/ギャップを許容 +/-の両strandを検索
検索時間/検索結果数	数分（状況による）/約33万件	約3時間/約6200万件	数十秒/約1500件
特記事項	• 大部分が短い部分一致配列（前後の配列は不明）. • 検索時間は利用者の環境に依存しない.サーバの混雑状況に依存.	• 大部分が短い部分一致配列（前後の配列は不明）. • 検索時間は利用者の実行環境に依存（本解析はCore i7-7700搭載PCで実施）.	• プライマー全長のアライメントが得られる. • いずれの条件も検索漏れがない. • 検索時間は利用者の環境に依存しない. サーバの混雑状況に依存.

プライマー交差性確認（類似配列検索）手順の比較
5：検索結果の後処置

資料19

	NCBI blast	NCBI blast+	GGGenome
プライマー全長のアライメントの取得	以下の手順例に従い，検索結果（表示されたアライメントの結果）をプライマー全長の領域に伸長する必要がある（プログラムによる処理が必要）． 1. ヒットした配列のIDとポジションの情報を元に，周辺を含む配列を取得． 2. ミスマッチ（必要に応じて挿入・欠失）を考慮して，プライマー全長のアライメントを計算し，アライメント結果を出力． 3. 検索結果の数分，上記の1,2を繰り返す．（結果によっては数千万回の処理が必要）		全ての検索結果がプライマー全長のアライメントを含むため，後処理が不要． 検索結果に，一致，ミスマッチ，挿入，欠失の塩基数が含まれるので，交差性（相補結合する塩基数の割合）を計算式で直接算出することが可能．

NCBI blast/NCBI blast+

```
>NZ_JHYC01000034.1 Legionella fairfieldensis ATCC 49588 T345DRAFT_scaffold00029.29_C,
whole genome shotgun sequence
Length=23907

 Score = 22.0 bits (23),  Expect = 713
 Identities = 13/14 (93%), Gaps = 0/14 (0%)
 Strand=Plus/Plus

Query  4      AGCTTAGTCCTGTT  17
              |||||| |||||||
Sbjct  18107  AGCTTAATCCTGTT  18120


 Score = 19.3 bits (20),  Expect = 2487
 Identities = 10/10 (100%), Gaps = 0/10 (0%)
 Strand=Plus/Plus

Query  8     TAGTCCTGTT  17
             ||||||||||
Sbjct  7000  TAGTCCTGTT  7009


 Score = 18.4 bits (19),  Expect = 8680
 Identities = 11/12 (92%), Gaps = 0/12 (0%)
 Strand=Plus/Plus

Query  2     TGAGCTTAGTCC  13
             ||||| ||||||
Sbjct  6115  TGAGCATAGTCC  6126


 Score = 17.5 bits (18),  Expect = 86
 Identities = 9/9 (100%), Gaps = 0/9
 Strand=Plus/Plus

Query  10     GTCCTGTTG  18
              |||||||||
Sbjct  17828  GTCCTGTTG  17836
```

最短完全一致塩基長（最小値4あるいは7）以上が一致している箇所が検索結果として得られる．いずれも前後のアライメントの状況は不明．このような結果が～数千万件出力される．　タブ区切り出力が可能．

GGGenome

Results:

検索語に色がつきます（ミスマッチ・挿入欠失）。

NC_045512.2 Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, complete genome
position: 12690-12707
▼12690
GGGCCAATTCTGCTGTCAAATTACAGAATAATGAGCTTAGTCCTGTTGCAGTACGACAGATGTCTTGTGCTGCCGGTA

NC_032091.1 Candida albicans SC5314 chromosome 3 sequence
position: 1408265-1408280
▼1408265
RTATGTAMTACCTTTGGCTTCTTCAGTTGAAT-AGCTTAGT-CTGTTGTCAGTTGAAGTAGGTGGAGAAGGAATTGCT

NZ_CAAAHX010000017.1 Legionella gresilensis strain Greoux 11D13, whole genome shotgun sequence
position: 9125-9142
▼9125
TTAAGCCTTGCAAGCCTCTTAATAAACTTAAACAGCTTAGTCCTGTTGCTGCTAGCATCCCATGCTTTGTGTTAAAA

NZ_CAAAIA010000003.1 Legionella impletisoli strain OA1-1, whole genome shotgun sequence
position: 170961-170978
▼170961
CATAAGTTTCTTTGAATAAGTTTCTACAACATGAGCTAATTCCTGTTGTGTGCCACAGGCTAACATGATGTCACTCAT

NZ_UGOD01000001.1 Legionella busanensis strain NCTC13316, whole genome shotgun sequence
position: 915259-915276
▼915259
TTCATATGGGTATTTTTGAATGTATTCATAATCAGCTTAGTCCTGTTACGATGGCTTGTCTTTAAAAGAGTATGCTAG

プライマー全長にわたるアライメントの結果が得られるため，そのまま一致率（交差性）を計算することができる．タブ区切り出力が可能．