トランスクリプトーム解析の現況
~マイクロアレイ vs. RNA-seq~

東京大学・大学院農学生命科学研究科

アグリバイオインフォマティクス教育研究プログラム

門田幸二（かどた　こうじ）

kadota@iu.a.u-tokyo.ac.jp

http://www.iu.a.u-tokyo.ac.jp/~kadota/

Google 門田 東大

スライドPDFはウェブから取得可能です

http://www.iu.a.u-tokyo.ac.jp/~kadota/ 門田 幸二のホームページ

門田 幸二のホームページ

名前 門田 幸二(かどた こうじ)

所属 東京
アグ

身分 特任

研究分野 バ

所属学会 日本
日本

研究テーマ(last

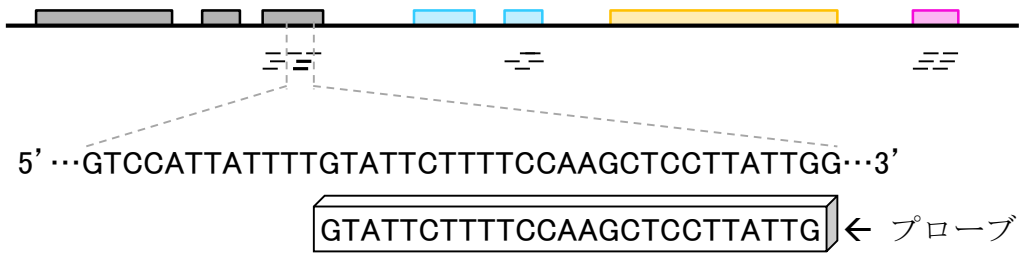
トランスクリプトーム解析
などによって得られる様
への応用を目指します。
す。これまでの主な研究
とめにできます。また、実
「(R で)塩基配列解析」上

講演など(上記講義以外)(last modified: 2013.10.30) NEW

26. 題目:「Rでゲノム・トランスクリプトーム解析」, HPCI チュートリアル・バイオインフォマティクス 実習コース, 生命情報工学研究センター(東京), 2014.03.07
25. 題目:「トランスクリプトーム解析の 現況2013(詳細版)」, 東京大学大学院農学生命科学研究科第124回アグリバイオインフォマティクス セミナー, 東京大学(東京), 2013.11.01
24. 題目:「トランスクリプトーム解析の 現況:マイクロアレイ vs. RNA-seq」, 生命医薬情報学連合大会「オミックス・計算そして創薬」・オミックス解析における実務者意見交換会, タワーホール船堀(東京), 2013.10.30
23. 題目:「食品機能解析研究とバイオインフォマティクス」, 日本農芸化学会2013年度大会・シンポジウム4SY08, 東北大学(宮城), 2013.03.27
22. 題目:「Rでトランスクリプトーム解析」, HPCI チュートリアル セミナー, 生命情報工学研究センター(東京), 2013.03.07
21. 題目:「Rでトランスクリプトーム解析」, HPCI チュートリアル セミナー, 生命情報工学研究センター(東京), 2012.03.09
20. 題目:「Rによるトランスクリプトーム解析 ~NGS由来塩基配列データを自在に解析する~」, Rでつなぐ 次世代オミックス情報統合解析研究会, 理化学研究所横浜研究所(神奈川), 2012.02.22
19. 題目:「RNA-Seqデータ解析リテラシー」, Illumina Webinar Series・RNAシーケンスを始めよう・ セッション3:データ解析, イルミナ株式会社(東京), 2011.11.17
18. 題目:「農業生物のトランスクリプトーム解析における情報処理」, 東京大学大学院農学生命科学研究科第91回アグリバイオインフォマティクス セミナー, 東京大学(東京), 2011.11.11
17. 題目:「RNA-Seqデータ解析における正規化法の 選択:RPKM値でサンプル間比較は危険?!」, イルミナ株式会社 バイオインフォマティクス講習会【中級】, 富士ソフトウェア アキバプラザ(東京), 2011.9.29

ステレオタイプなイメージ

- マイクロアレイの長所
  - 取り扱いやすいデータ量(~100Mb程度)
  - 長年の実績:解析手法がほぼ確立。(Windows Rのみで解析可能)
  - 検査用チップが利用可能(MammaPrintなど)
- マイクロアレイの短所
  - 解析可能範囲が搭載転写物に限定
  - プローブが3'末端に偏っている(3'発現解析用アレイ)
  - ダイナミックレンジが狭い



ステレオタイプなイメージ

- RNA-seqの短所
  - 取り扱いづらいデータ量（数百Gb?!）
  - Windows userは自力解析が困難（ほとんどがLinux用）
  - ダイナミックレンジが広いがために?!「変」な結果に遭遇。
  - ゼロカウントデータの取り扱い（本当は気にしなくてもいいのに…）

- RNA-seqの長所
  - （多少のoff-targetは含むが）全発現転写物の解析が可能
  - 解像度：遺伝子レベル → 転写物レベル
  - ダイナミックレンジが広い

マイクロアレイ

- （少なくとも）ニュートリゲノミクス（食品）系では主力
  - 解析対象生物種：マウスやラット
    - 特定の栄養素欠乏状態の遺伝子発現解析など
    - 納豆、バナナ、カテキン、乳酸菌、大豆、食物繊維、ポリフェノール、…
  - Gene Ontology解析やパスウェイ解析
    - 実績のある市販アレイに搭載されている遺伝子のみでも「この栄養素はこのパスウェイに効いている」的な新規知見が得られればよい、という思想
    - 「個別の遺伝子の変動解析」というよりは「遺伝子セットの変動解析」
  - 同一アレイを用いている限り全体的な情報量が豊富
    - 公共データベース（GEO, ArrayExpressなど）
    - 3’発現解析用アレイが未だに使われる所以
  - 異なるアレイであっても同一生物種であればマージ可能
    - virtualArray (Heider and Alt, *BMC Bioinformatics*, **14**:75, 2013)など

ダイナミックレンジ周辺の雑感

- 既知濃度のspike-inデータとシグナル強度との直線性
- Hekstra et al., *Nucleic Acids Res*., **31**: 1962-1968, 2003
  - マイクロアレイはシグナル強度が高発現側で飽和し、低発現側では実際の濃度よりも高めに見積もられる（Fig. 4B）

上記論文のFig. 4B

プローブレベルのハイブリダイゼーションはLangmuir-adsorption modelに従う

ダイナミックレンジ周辺の雑感

- Langmuir-adsorption modelによる直線性向上の取り組み
  - 非特異的結合（non-specific binding; NSB）の理解
  - 総説(Harrison et al., *Nucleic Acids Res*., **41**: 2779-2796, 2013)
    - Gが4つ以上連続するプローブは外れ値になりやすい(Upton et al., 2008)
    - 4G signatureを持つプローブ同士がGカルテットを形成(Langdon et al., 2009)
    - …
  - 方法
    - Hook法 (Binder et al., *Algorithms Mol. Biol*., **3**: 11, 2008)
    - Inverse Langmuir法(Mulders et al., *BMC Bioinformatics*, **10**: 64, 2009)
    - MSNS model (Furusawa et al., *Bioinformatics*, **25**: 36-41, 2009)

ダイナミックレンジ向上を目指した方法は存在する

ダイナミックレンジ周辺の雑感

- 既知濃度のspike-inデータとシグナル強度との直線性
- “昔の方法”で数値化したアレイデータとの比較が多い
  - Nookaew et al., *Nucleic Acids Res.*, **40**: 10084-10097, 2012
    - PLIER(2004年ごろ)とcubic spline法(Workman et al., 2002)
  - Xu et al., *BMC Bioinformatics*, **14 Suppl 9**: S1, 2013
    - RMA (Irizarry et al., *Biostatistics*, **4**: 249-264, 2003)
  - Raghavachari et al., *BMC Med. Genomics*, **5**: 28, 2012
    - RMA (Irizarry et al., *Biostatistics*, **4**: 249-264, 2003)
  - Mortazavi et al., *Nat. Methods*, **5**: 621-628, 2008
    - MAS5 (Hubbell et al., *Bioinformatics*, **18**: 1585-1592, 2002)

比較的最近の方法との評価をすべきではある

マイクロアレイ（前処理法 or 正規化法）

- よく使われてきた方法
  - RMA (Irizarry et al., *Biostatistics*, 2003)
    - **特徴**：データセット中の複数のアレイデータ情報を利用(multi-array basis)
    - **probe level正規化**：quantile normalization
    - **要約統計量**：median polish
  - MAS5 (Hubbell et al., *Bioinformatics*, 2002)
    - **特徴**：アレイごとに独立して前処理(正規化)を実行(per-array basis)
    - **probe level正規化**：なし
    - **要約統計量**：one-step Tukey's biweight

RMAがいいという評価がほぼ定着?!

マイクロアレイ（前処理法 or 正規化法）

- RMAの問題点
  - 本当はばらつきの大きいデータを過小評価
    - median polishを利用しているため、手続き的に必要以上にサンプル間で似た結果を返す(Giorgi et al., *BMC Bioinformatics*, **11**: 553, 2010)
  - サンプル数の増減のたびに、RMA再実行の必要性
    - quantile normalizationを利用しているため、リファレンス分布が変化。例えばサンプル数の増加の場合、元々存在していたサンプルの数値も変わってしまう。
- MAS5の問題点
  - 低発現領域でばらつきが大きい傾向
    - Absent callのデータをフィルタリング(すればいいのに)しないため(McClintick and Edenberg, *BMC Bioinformatics*, **7**: 49, 2006)

このあたりを認識できていないヒト、意外に多いのかも…

マイクロアレイ（前処理法 or 正規化法）

- fRMA(McCall et al., *Biostatistics*, **11**: 242-253, 2010)
  - RMAの改良版(サンプル数の増減の影響を受けない)
  - シグナル強度を得たいデータセット以外の多様なデータを用いて、正規化に必要な「リファレンス分布」と「プローブ効果の推定値」の情報を予め取得（パラメータをfrozenしておき、それを新規サンプルに独立に適用）

    → 目的データセット中のサンプルのシグナル強度を(データセット中の他のサンプルの影響を受けずに)得ることが可能
  - RMA → refRMA → fRMA。refRMA(Katz et al., 2006)では一定と仮定していたバッチ効果を考慮
  - 短所
    - パラメータ推定が大変らしく、Affymetrixチップの一部しか利用不可能
    - Affymetrix Exon array用のパラメータ提供が論文に…(McCall et al., 2012)

マイクロアレイ（前処理法 or 正規化法）

- IRON(Welsh et al., *BMC Bioinformatics*, **14**: 153, 2013)
  - 解析データセットの中からリファレンスサンプルを一つ選び、それと他のサンプルをペアワイズで正規化。リファレンスが固定されているので、サンプル数の増減の影響を受けない。要約統計量はTukey's biweight。
- RMX (Kohl and Deigner, *BMC Bioinfo*rmatics, **11**: 583, 2010)
  - MAS5と同じで、要約統計量の計算部分がrobust rmx estimatorに置き換わったもの。正真正銘per-array basisのものであるため、ややこしいことを考えなくてよく、使用感もよい(個人の感想です)。
- ベイズオンライン学習(Lahti et al., *Nucleic Acids Res*., **41**: e110, 2013)
  - Affymetrix以外の様々な市販アレイにも対応した拡張性の高いアルゴリズム

進展していますね。。。

マイクロアレイ（デバイス自体）

- 3’発現解析用アレイ → exon array → transcriptome array
  - Affymetrix Human Transcriptome Array (HTA 2.0)
  - Furney et al., *Cancer Discov*., **3**: 1122-1129, 2013.
  - GPL17585(exon level)
  - GPL17586(gene level)

転写物数は有限であるため、
RNA−seqによる網羅的な同定後は、
「トランスクリプトームアレイ」に移
行するほうがお手軽かもしれない

3’発現解析用アレイ、エクソンアレイ、
HTA2.0アレイのプローブの比較の図
（どこから得たか忘れました
…Affymetrixさんから直接もらったかも）

NGS(RNA-seq)…教える側は大変

- 大人数のハンズオン講義でコマンドライン系はアリエナイ
  - 基本的なコマンド：cd, pwd, ls
  - 「スペース」の概念：ファイル名中に普通に存在…
    - 「bowtie -k2 filename」→「bowtie-k2filename」
  - エラーの認識：ごく初期段階でダメになってるのに…
  - 「興味本位系受講生」と「本気で学びたい系受講生」で大きな差
  - 「かじったことがある系」と「本当の初心者系」でも…

今はLinuxコマンド抜きで一通り解析可能

- *SRAdb* (Zhuら, 2013)
  - 公共DBからのRNA-seqデータ（FASTQファイル）取得
- ***QuasR* (Lerchら, unpublished)**
  - **リファレンス配列（ゲノム or トランスクリプトーム）へのマッピング**
    - **Bowtie (Langmeadら, 2009) or SpliceMap (Auら, 2010)を選択可能**
    - **出力はBAM形式ファイル、QCレポートも**
  - **遺伝子アノテーション情報をもとにカウントデータ取得**
    - ***GenomicFeatures* (Lawrenceら, 2013)から得られるTranscriptDbオブジェクトを利用**
    - **UCSC known genesやEnsembl genesのカウントデータなど**
- *TCC* (Sunら, 2013)
  - 内部的に*edgeR* (Robinsonら, 2010)や*DESeq* (Anders, 2010)などを用いて頑健な発現変動解析を実行

アセンブル以外ならWindows（のR）上でどうにかなる時代がやってきました

解析例：SRP017142のデータ

- *SRAdb*を用いたgzip圧縮FASTQ形式ファイルのダウンロード
  - Neyret-Kahn et al., *Genome Res*., **23**: 1563-1579, 2013のデータ
  - 複製あり2群間比較用RNA-seqデータ(3 Ras vs. 3 Proliferative)

FileName	SampleName
SRR616151.fastq.gz	Pro_rep1
SRR616152.fastq.gz	Pro_rep2
SRR616153.fastq.gz	Pro_rep3
SRR616154.fastq.gz	Ras_rep1
SRR616155.fastq.gz	Ras_rep2
SRR616156.fastq.gz	Ras_rep3

計6GB程度。*QuasR*パッケージは圧縮ファイルのままでマッピング可能

- *QuasR* (Bowtie)を用いたヒトゲノムへのマッピング
  - *BSgenome.Hsapiens.UCSC.hg19*パッケージを利用
  - 18種類程度の生物種のゲノム配列がRパッケージとして利用可能

解析例：SRP017142のデータ

- *QuasR* (Bowtie)を用いたカウント情報取得

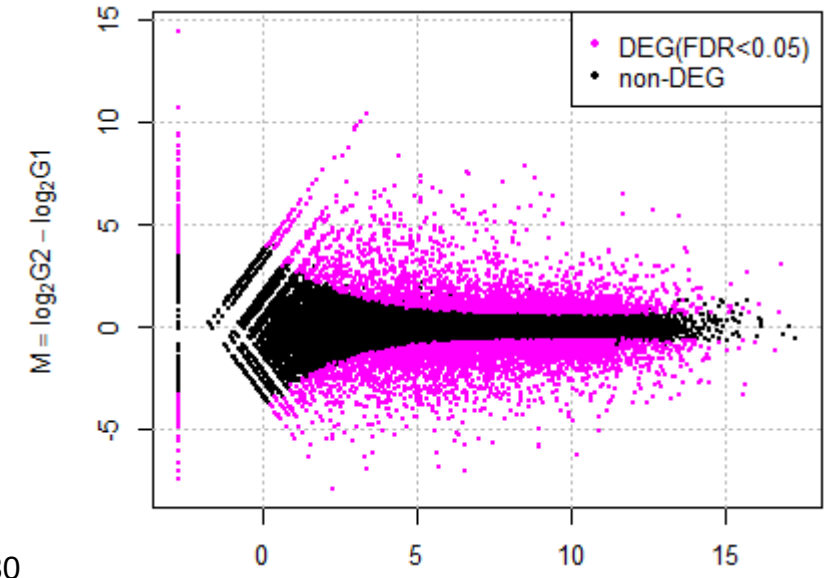
raw countデータ

	Pro_rep1	Pro_rep2	Pro_rep3	Ras_rep1	Ras_rep2	Ras_rep3
ENSG00000000003	480	513	366	124	271	366
ENSG00000000005	0	0	0	1	0	0
ENSG00000000419	282	354	208	165	301	209
ENSG00000000457	167	198	155	156	248	129
ENSG00000000460	114	112	101	55	81	59
…						

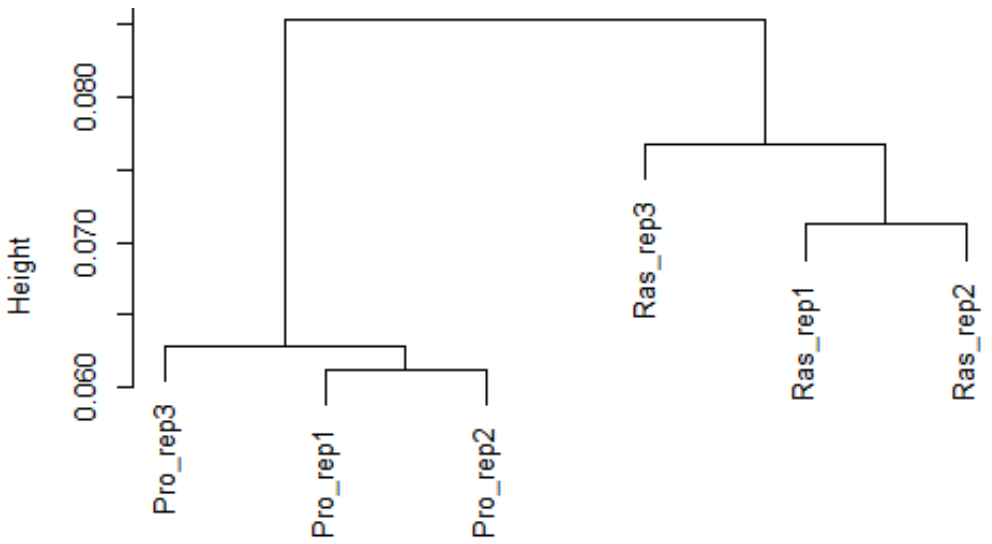
RPKMデータ

	Pro_rep1	Pro_rep2	Pro_rep3	Ras_rep1	Ras_rep2	Ras_rep3
ENSG00000000003	7.14	7.68	6.17	3.07	4.51	6.81
ENSG00000000005	0.00	0.00	0.00	0.05	0.00	0.00
ENSG00000000419	10.31	13.03	8.62	10.04	12.32	9.57
ENSG00000000457	1.92	2.29	2.02	2.98	3.19	1.85
ENSG00000000460	0.79	0.78	0.80	0.64	0.63	0.51
…						

*TCC*を用いた発現変動遺伝子同定



サンプル間クラスタリング



雑感（発現変動遺伝子DEG検出精度）

- 一般的な解析手順：データ正規化法 → DEG検出法
- マイクロアレイ
  - 前処理法がmulti-array basis のときはFold-change（FC）系
    - 例：RMA法 → Rank products法
  - 前処理法がper-array basisのときは*t*検定（モデル）系
    - 例：MAS5法 → SAM法
  - 最近の前処理法ではどうなのだろう？
    - fRMAやIRONのときはFC系で、RMXのときはモデル系が成立するか？

- RNA-seq
  - 提供されているRパッケージ（*edgeR*や*TCC*）はモデル系
    - 例：処理されていないカウントデータ → NBモデル
  - 各種補正後（配列長など）のデータでもDEG検出を行う枠組み?!
  - 統計的には有意かもしれないが…直感的に変な結果…。

(Rで)マイクロアレイデータ解析

(last modified 2013/10/17, since 2005)

What's new?

- 2013年10月30日13:30-15:00に開
  計算　そして創薬」のフォーカスト
  いてざっくり話す予定です。(2013
- (かなり先の話ですが...)平成26年
  析を行います。情報はかなりアッ
- 「(Rで)塩基配列解析」もリニュー
- どのブラウザからでもエラーなく見
- 2013年7月18日まで公開していた
  らダウンロード可能です(88MB程
- R3.0.1がリリースされていたので
- 遺伝子セット解析の一つであるG

- はじめに (last modified 2013/07/1
- Rのインストールと起動(last modi
- Rの昔のバージョンのインストール
- 使用例(初心者向け)(last modifie
- サンプルデータ (last modified 20
- イントロ | 発現データ取得 | 公共D
- イントロ | 発現データ取得 | inSilic
- イントロ | 発現データ取得 | Array
- イントロ | 発現データ取得 | GEOc

(Rで)塩基配列解析(主に次世代シーケンサーのデータ)

(last modified 2013/10/19, since 2010)

What's new?

- 一連の解析パイプライン(RNA-seqデータ取得 -> マッピング -> カウントデータやRPKMデータ取得 -> サンプル間クラスタリングや発現変動解析およびM-A plot描画まで)をアップデートしました。項目名の一番下のほうです。(2013/10/19) NEW
- 発現変動解析用Rパッケージ TCC (ver. 1.2.0; Sun et al., BMC Bioinformatics, 2013)がBioconductorよりリリースされました。最新版を利用したい方は、R (ver. 3.0.2)をインストールしたのち、Bioconductor (ver. 2.13)をインストールしてください。10/17以降にRのインストールと起動を参考にして、通常のインストール手順で行えばそのバージョンになるはずです。(2013/10/17) NEW
- BioMart周辺の情報をアップデートしました。(2013/09/26) NEW
- 3群間比較解析(single-factorのみ)を一通り掲載しました。(2013/09/16)
- 2013年10月30日13:30-15:00に開催される、バイオインフォマティクス系学会の合同年会生命医薬情報学連合大会「オミックス・計算　そして創薬」のフォーカストセッション オミックス解析における実務者意見交換会では、トランスクリプトーム解析周辺についてざっくり話す予定です。(2013/10/08) NEW
- (かなり先の話ですが...)平成26年3月7日に東京お台場にて、HPCIチュートリアルの一部としてRでゲノム・トランスクリプトーム解析を行います。情報はかなりアップデート予定です。興味ある方はどうぞ。(2013/09/11)
- どのブラウザからでもエラーなく見られる(W3C validation)ようにリニューアルしました。(2013/07/30)
- 「(Rで)マイクロアレイデータ解析」もリニューアルしました。(2013/07/19)
- 2013年7月29日まで公開していた以前の「(Rで)塩基配列解析」のウェブページや関連ファイルはRdeennki.zipからダウンロード可能です(110MB程度)。(2013/07/30)

- 2013年6月6日に開催されたNAIST植物グローバル教育プロジェクト・平成25年度ワークショップのときに利用した、R(ver. 3.0.1)とTCC(ver. 1.1.99)などのインストール方法はこちら(Windows用のみ;hoge.zipはおまけ)です。

- はじめに (last modified 2013/07/30)
- Rのインストールと起動(last modified 2013/09/27) NEW
- サンプルデータ(last modified 2013/10/17) NEW
- イントロ | 一般 | ランダムに行を抽出(last modified 2013/10/10) NEW
- イントロ | 一般 | 任意の文字列を行の最初に挿入(last modified 2013/10/10) NEW

今後の予定

2014年度も「アグリバイオインフォマティクス教育プログラム」は継続予定!!

- 2013年10月31日（木）10:30-12:00@タワーホール船堀 小ホール
  - JSBi年会DBCLSスポンサードセッション
  - オープンサイエンスアワード（DSW48）内で、5分程度「（Rで）…」のウェブページに関する話をする予定です。
- 2013年11月01日（金）15:00-17:00@東大・農学部図書館
  - アグリバイオインフォマティクスセミナー（**公開**）
  - 1時間、「TCC論文の内容」や「FASTQファイルからの一連の解析パイプラインについてのデモ」などを行う予定です。
- 2014年03月07日（金）10:30-17:00@産総研・CBRC（お台場）
  - HPCI講習会・バイオインフォマティクス実習コース（の一部）
  - CBRCの設備（PC）を用いて、大幅に更新した「（Rで）…」のウェブページを用いたハンズオンセミナーを丸一日かけて行います。
- 「Rでトランスクリプトーム解析」の本（仮題）出版予定
  - 来年度の講義に間に合うよう、平成25年度初めまでにはなんとか。。。（鋭意執筆中）
  - 徹底的にR！単なる使い方はウェブに任せて、概念の理解や、バイオインフォマティクス的なものの考え方、Rコード中身の理解など応用力重視の内容（を目指してます。。。）

謝辞

共同研究者

清水 謙多郎　先生（東京大学・大学院農学生命科学研究科）

西山 智明　先生（金沢大学・学際科学実験センター）

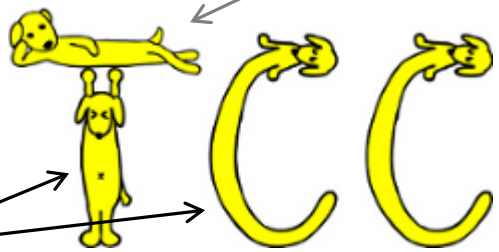
孫 建強　氏（東京大学・大学院農学生命科学研究科・大学院生）

グラント

- 基盤研究(C)（H24-26年度）:「シークエンスに基づく比較トランスクリプトーム解析のためのガイドライン構築」(代表)
- 新学術領域研究(研究領域提案型)(H22年度-):「非モデル生物におけるゲノム解析法の確立」(分担;研究代表者:西山智明)

（妻の）門田 雅世さま作

挿絵やTCCのロゴなど



（有能な秘書の）三浦 文さま作