まえがき

がん細胞で多く見られる新たな遺伝子の発見！　新たな肥満抑制遺伝子の発見！　お米の数を増やす遺伝子の発見！のような生命科学分野の研究は**新規腫瘍マーカー，やせ薬?!，食糧問題解決**の可能性を秘めた重要な基礎研究成果である．特にヒトの病気に関しては，血液のみから各種がんを診断する技術開発などが盛んに行われている．トップランナーは乳がんに関するものであり，乳がんの予後を予測（高リスク群と低リスク群）する検査はすでに実用化されている．現在乳がん患者は，希望すれば（保険外診療にはなるが）がん再発リスクの程度を把握し，負担のかかる化学療法の必要性に関する，より信頼性の高い情報を得ることができる．本書の内容と密接に関連したこれらの研究成果の多くは，生体内に存在する数万遺伝子の働きの程度（遺伝子発現レベル）に関する数値データの取得，およびがん細胞と正常細胞のような比較するサンプル間で働きの程度が異なる（発現に差がある）候補遺伝子同定に基づいている．当然，候補遺伝子の良し悪しは**①遺伝子発現レベルの数値化，②サンプル間でのデータ正規化，③発現変動遺伝子同定**の各ステップにおける手法選択が重要である．そしてユーザが知りたいのは，**どのような手法が存在し，どのような思想のもとに開発されたのか，どう使うのか，どの組合せが最適か，結果をどのように解釈するのか**である．

本書のタイトルである**トランスクリプトーム (transcriptome) 解析**は，上記**遺伝子発現解析**と本質的に同じである．現在，細胞内で実際に働き（転写され）生命活動に重要な役割を担う実体が**転写物** (transcripts) であることは広く知られている．この転写物全体を指す言葉がトランスクリプトームであり，代表的な計測技術がマイクロアレイと RNA-seq である．2010 年頃より，本格的に計測手段がマイクロアレイから RNA-seq（次世代シーケンサ NGS を用いた転写物の網羅的解析技術）に移行しつつある．そのため，便宜上各章でマイクロアレイと RNA-seq に分けて説明しているが，RNA-seq に関する記述の大部分はマイクロアレイの知識を前提としている．

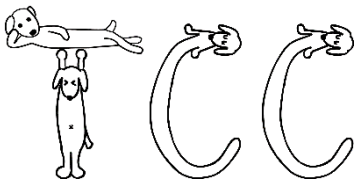
本書は，トランスクリプトーム解析を行うための一連のスクリプト集である著者の 2 つのウェブページ（**R で）マイクロアレイデータ解析**および（**R で）塩基配列解析**を体系的にまとめた初の書籍である．今日では，インターネット上の検索エンジンでキーワード検索すれば，個別の情報は簡単に得られる．しかし，本書のメインターゲットである生命科学分野の実験系研究者やこれからバイオインフォマティクスを学ぼうとする大学院生にとっては，特に統計関連の記述は難解であろ

う．巷に溢れている統計関連書籍の記述内容もまた，意味不明だという声をよく聞く．理由は簡単で，統計の専門家の多くは，最初から一般式を多用する説明の仕方ではわからないというヒトの気持ちがわかっていないためである．本書は，まず手元にある実際のデータやその解析結果を示し，解釈の仕方を述べてから一般論に導く記述形式を採用している．主な目的は，実データの解析結果を徹底的に眺めることで，統計的な感覚や数式の感覚を身につけることである．一般に，書籍中に記載されているRパッケージや関数は，実際にやろうとするとパッケージ自体がなくなっていたりオプションが変更されるなどすぐに陳腐化していく．本書においても，もちろん執筆時点において最新の解析手順やRの関数を利用しているが，それらの**賞味期限**は短いことが予想される．そのため，最新の利用手順は2つのウェブページを参考にされたい．

著者とRとの出会いは2005年である．この年，東京大学・大学院農学生命科学研究科に教員として着任し，主に実験（農学）系大学院生にバイオインフォマティクスの教育を自分が手軽に行う手段としてRを採用することにした．そして，Rを勉強し始めた著者のための備忘録として始まったのが2005年の初公開からすでに8年以上経過している（**Rで）マイクロアレイデータ解析**と2010年からスタートした（**Rで）塩基配列解析**である．これらのウェブページは，**プログラムなんてちゃんと動けばそれでよし！**という思想が根本にあること，そしてメインユーザである実験系研究者からの多数の要望に対してアドホックな対応を繰り返してきたため，非効率かつ不細工なRコードで満たされていた．今回，一念発起してウェブページ全体の大幅な加筆修正を行ったが，編集者である同志社大学の金 明哲氏からの原稿執筆のお誘いがなければおそらく全体を修正することはなかったであろう．金氏および共立出版の横田穂波氏の尽力に対し感謝したい．

著者の恩師でもある清水謙多郎教授（東京大学大学院農学生命科学研究科 応用生命工学専攻 生物情報工学研究室）は，著者の現所属でもあるアグリバイオインフォマティクス教育研究プログラム設立の立役者である．清水教授なくして本書の根幹をなす2つのウェブページは存在しない．本稿についても，入念なチェックおよび数多くの有益なコメントをいただいたことを記しておきたい．著者の塩基配列 (NGS) 解析との実質的な出会いは，東京大学大学院農学生命科学研究科 生産・環境生物学専攻 昆虫遺伝研究室の嶋田 透教授，勝間 進准教授，そして河岡慎平氏との共同研究がきっかけである．**コンティグ，アダプター配列，センス・アンチセンス鎖**など，聞きなれない用語に戸惑いつつ，NGSデータ解析分野に後発参入したのが2009〜2010年頃である．第4章で主に用いたトランスクリプトームデータ解析用パッケージ *TCC* は，金沢大学の西山智明氏との共同研究として，また各種助成金（課題番号：24500359，22128008）の成果として得られたものである．ベータ版（ver.0.4まで）開発や *TCC* の命名は，西山氏による．*TCC* ver. 1.2.0で追加された組織特異的発現遺伝子検出法ROKUは，上田太一郎氏によって提案された**複数外れ値の簡易検出法**を内部的に用いている．*TCC* への実装に際し，奔走していただいた米谷 学氏および関係者に感謝したい．また，今回の出版に合わせたウェブページの大幅なリニューアル作業（W3C validation，美しいコーディングなど），そして *TCC* 開発の実働部隊は孫 建強氏である．著者からの**要望や希望**は，彼の圧倒的なコーディング能力の高さ・迅速な仕事により速やかに満たされ，円滑な原稿執筆を可能にした．

2つのウェブページを含むサーバの安定的な維持管理は，寺田 透氏の尽力によるものである．多くのバグレポートや改善提案をいただいた中井雄治氏，NGS 解析の講義を含め長らく苦楽を共にした徐 泰健氏，クオリティの高いコーヒー提供から *TCC* ロゴ作成までそつなくこなす有能な秘書の三浦 文氏にも感謝したい．最後に，怪獣2人（友宏8歳，浩輝6歳）を連れ出して執筆環境整備に尽力してくれた妻（雅世）に感謝するとともに，近い将来，子供たちが本書を読んで勉強することを期待したい．下記のロゴは，妻（T の上）と秘書（T の下）の合作である．



2014 年 1 月　　門田 幸二