

はじめに

現在、現生人類の世界人口は70億人を超えると推定されており、その地理的分布も世界のほぼすべての大陸で広範囲に及ぶ。しかしながら、遠い過去の人口はそれほど大きくなかった、あるいは大変に小さかったと考えられており、現在のような指数的な人口増加が見られるのは、比較的最近、特に産業革命以降であったことがわかっている。ヒトゲノムに残された多様性からは、過去の人類の集団サイズ（繁殖有効集団サイズ）は長期にわたって非常に小さく、1万人程度であったと推定されている。このような長期にわたる定常的な小集団の過去と、比較的最近になってからの爆発的な人口増加（しかも地理的偏りを持った）から、現在のヒト集団のゲノムには、背反する2つの遺伝的特性が付加されたと推察される。つまり、長期にわたる定常小集団としての歴史で遺伝的浮動が強く作用して構成された特徴と、最近の指数的人口増加による大量の突然変異の導入である。このように相反する特徴が、ヒトの健康問題にどのような効果を持つか、いまだ、はっきりしたことはわかっていない。この2つの特徴は、それぞれありふれた疾患に関するCDCV（common disease common variant）仮説とCDRV（common disease rare variant）仮説の理論的背景をなすものだが、その定量的割合は不明である。

過去には、主にコスト面からの折衷として、さまざまなヒト疾患や形質で、ヒト集団中に高頻度で存在するSNPを用いたGWAS研究が非常な成功を収めてきた。他方、WGS（whole genome sequencing）／WES（whole exome sequencing）の読み取りコストが低下し、現在の技術での下限に迫りつつある。また、読み取り精度も向上し、いくつかの国際的バイオバンクでは数万人レベルでの大規模なWGS／WESデータの蓄積が急速に行われつつある。このような現状で、ここ最近の方法論の発展と今後の展望を整理し、実用的なWGS／WESデータの利用ガイドを提供するために本書が企画された。第1章では、次世代シーケンシング（NGS）技術の基礎と、計算負荷の高いWGS／WESデータのスーパーコンピューター（スペコン）上での解析演習を取り上げた。第2章では、特に最近の進歩が著しい計算機上でのバリアントアノテーションについて解説した。第3章では、NGSの実用例として、希少難病の原因遺伝子探索と多因子疾患での柔軟なレアバリアント遺伝統計解析をテーマとしている。第4章では、既存データの活用や多因子疾患のリスク計算、関連バリアントの臨床的意義解釈でのAlphaFoldの利用などを扱った。そして、第5章では、来たるべき、さらなるNGS解析の普及に必要な新技術について取り上げている。また、各章のテーマに即した派生的で先端的な話題をコラムとして紹介した。これら的重要な方向性についてはいずれ別途まとめて行きたい。

大規模なNGS解析は、データ生産から遺伝統計解析まで、現状では限られた施設でしか実地経験をつむことができない。そのようななかで、本書が、初学者にとって、実際に自ら手を動かしてNGSデータ解析に必要なノウハウを習得する一助となることを願っている。

なお、本書の基礎となる遺伝学の基本部分については、拙著「ゲノム医学のための遺伝統計学」（共立出版、2015年）を参照されたい。また、SNPデータの扱いについては本書では取り上げていないが、「ゼロから実践する遺伝統計学セミナー」（岡田随象、羊土社、2020年）を参照されたい。

2023年7月

東北大学大学院医学系研究科／理化学研究所革新知能統合研究センター

田宮 元