グラフゲノムブラウザ

―疾患研究を促進する遺伝情報の可視化―

1.　背景

ヒトを含めたあらゆる生物では、遺伝情報をゲノム、すなわち塩基配列の形で保持している。ゲノムには生物を形作るために必要な情報のほとんど全てが含まれていると考えられており、生物学・医学・薬学・農学など生物を扱うあらゆる学問分野においてゲノム研究は極めて重要である。同じ生物種でも個体ごとで塩基配列に違いがあるため、参照配列と呼ばれる種を代表する塩基配列が定められている。多くのゲノム研究では、遺伝子の座標区間や様々な個体の参照配列からの差分（多型情報）など、多様なゲノム由来の情報を参照配列に対応付けた上で様々な解析を行っている。「ゲノムブラウザ」はそうしたゲノムに関わる情報を統合し、インタラクティブに可視化するソフトウェアであり、情報の解釈を容易にし、新たな知見を得るために、主に研究者の間で広く用いられている。

DNAの読み取り技術はこの12年間で約100億倍の進歩を遂げ、これに伴って個体間のゲノム配列の違いがより明らかになった。例えば近年、配列間の欠失・挿入・逆位・重複・転座などの構造多型が網羅的に同定されている。これによって遺伝子が重複・欠損・融合し、それが遺伝性疾患やがんの原因となる例が知られていることから、構造多型の解析は重要である。しかし構造多型には、ただ1本の参照配列の上にあらゆる情報を対応付けることが難しいケースも多く存在するため、ゲノムブラウザによる可視化が不十分になり、その解析が困難となっている。このような問題に対して、ゲノムを1本の線ではなく、グラフ構造を用いて分岐やループなど様々なパターンを表現可能にする「ゲノムグラフ」を用いた解析が近年有望視されている。

2.　目的

本プロジェクトでは、ゲノムグラフの可視化を行うことのできるゲノムブラウザとして、Webアプリケーション「MoMIG: Modular Multi-scale Integrated Genome Graph Browser」を開発した。MoMIGによって、従来のゲノムブラウザでは困難であった構造多型の可視化を改善することが可能になった。

3.　開発の内容

本プロジェクトで開発した「MoMIG」は、全ゲノムレベル（Overall View）、遺伝子レベル（Graph View）、塩基配列レベル（Linear View）といった複数の視点の組み合わせで構成される「階層的ビュー」というコンセプトに基づき、対象とするスケールの異なる複数のコンポーネントをダッシュボード上に組み合わせることができる実装とした（図1）。これにより、対象とするゲノム、生物種に適したデザインを選択して可視化することが可能となる。ここでは、構造多型を可視化する場合に限定し、本ソフトウェアに組み込んだコンポーネントを抜粋して説明する。

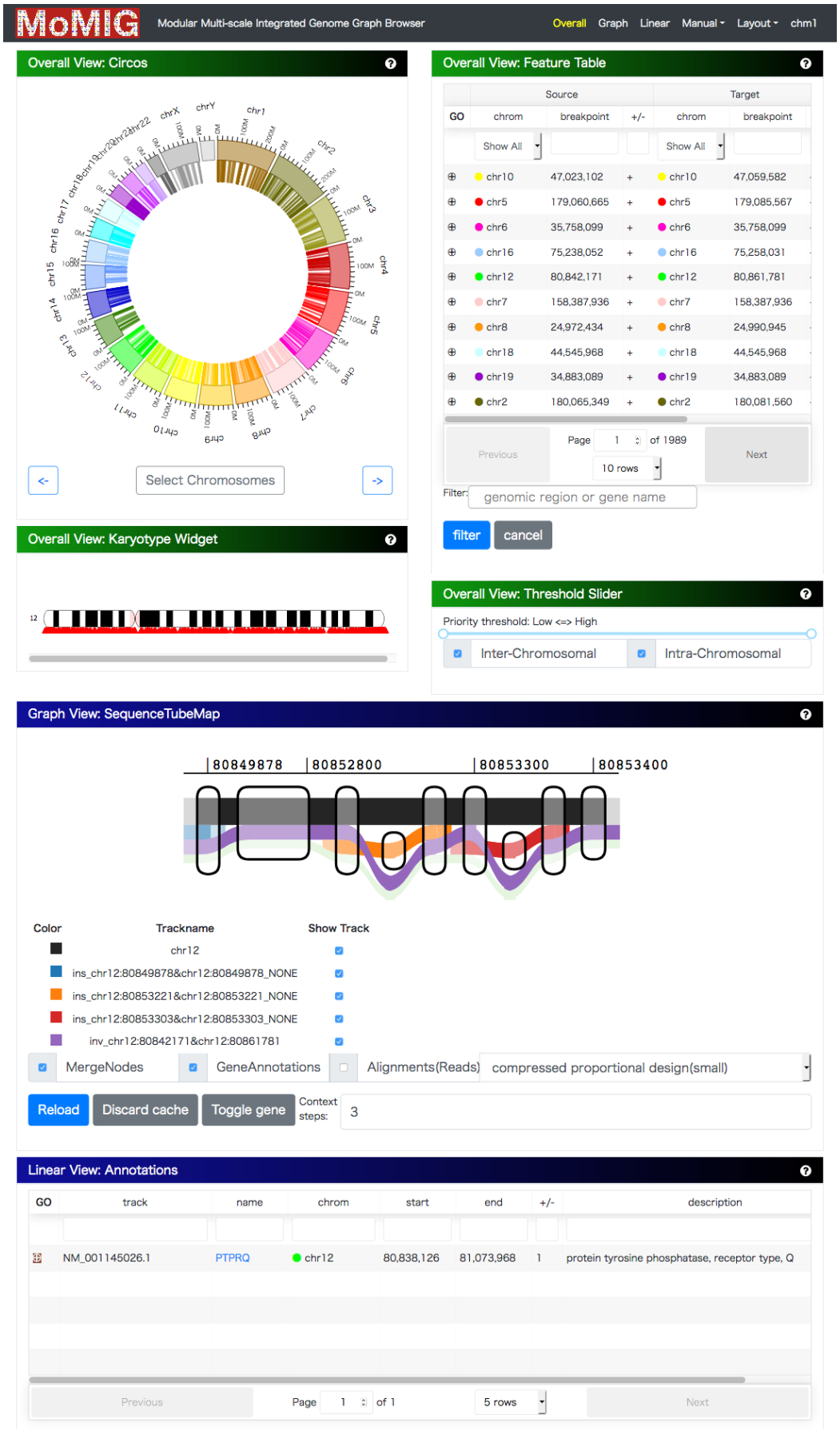

図1. 開発したソフトウェアのスクリーンショット

- **Overall View: Circos Plot**

Circosは、環状に染色体の意匠を配置し、染色体上の構造多型が存在する位置に対応する曲線を引くデザインであり、ゲノム科学でしばしば用いられている。このデザインによって、染色体上の構造多型の分布を可視化しているとともに、ナビゲーションとして、曲線をクリックすることで、可視化する対象の構造多型を選択できる。

- **Graph View: SequenceTubeMap**

SequenceTubeMap は、Overall Viewによって選択した構造多型に対応する部分グラフを描画している。これにより、構造多型が参照配列に対してどのような関係になっているかを、グラフ構造に基づいて確認することができる。本ソフトウェアの実装では、既存のライブラリを構造多型の性質にあわせて改変して統合した。

- **Linear View: Annotations**

Annotations では、Graph Viewで表示される参照配列に対応する遺伝子の情報を、データベースから取得している。これは構造多型と遺伝子の関連性を把握するのに必要である。

4.　従来の技術（または機能）との相違

ゲノムグラフの可視化ができるゲノムブラウザとして設計し、それによって構造多型の可視化を改善できることを示した点が新規である。

既存のゲノムブラウザは、ゲノムグラフの可視化をすることができない。また構造多型のような、参照配列上で遠く離れた座標における関係性を解釈することが困難であった。これに対して構造多型をゲノムグラフで表現し、その部分グラフを可視化することで、参照配列上の距離によらず、構造多型を一画面に収めた可視化を行うことが可能になった（図2 a）。

一方で従来の構造多型を可視化するためのソフトウェアは、複数の構造多型が入れ子になるような、参照配列のある区間に対応付けるのが困難な構造多型の可視化が不十分であったが、本ソフトウェアではゲノムグラフ表現を可視化に直接利用することで、複数の構造多型の関係性を含めて可視化することが可能になった（図2 b）。

(a) 遠く離れた座標間の構造多型　　(b) 入れ子になった配列の重複

図2. 複雑な構造多型の可視化例

また、ゲノムグラフの一種であるアセンブリグラフを可視化するためのツールは、ゲノムブラウザのようにゲノムに関連する様々な情報を統合するという点が不十分であった。そのため、構造多型の解釈など他の様々な用途には適していなかった。

本ソフトウェアによってゲノムグラフを、そのゲノムに関連する様々な情報と統合して可視化することが可能になったため、多数の構造多型の候補を絞り込み、可視化し、そして検証するという一連の解析ワークフローがGUI上で実現できる。そのため本ソフトウェアは、他の構造多型を可視化するためのソフトウェアと比べても、より有力な可視化手段を提供している。

5. 期待される効果

本ソフトウェアによって改善される構造多型の可視化は極めて重要である。例えば構造多型の一種として、参照配列に含まれていない約9,600箇所の塩基配列が、日本人が持つ配列として特定されている。このように、構造多型は個人差としてみられるほど普遍的な現象である。

更に、構造多型の中には疾患に関連しているものも存在する。構造多型と相関があることが知られる遺伝病が、全世界の約3,500万人に影響を及ぼしているほか、構造多型が遺伝子の領域に影響を及ぼすことによって生じる融合遺伝子は、肺腺がんや大腸がんをはじめ、様々な疾患の原因となっていることが明らかになっている。今後、より多くの疾患ゲノムの解読が進むにつれて、疾患と相関のある未知の構造多型が多数同定されることが想定されるため、可視化は必要不可欠である。

また、今後数年の間に数百万人規模のゲノムが解読されると考えられており、ゲノムグラフによって集団ゲノムを表現することが、計算量を削減し、統一した座標系の上で多数のゲノムを扱うために必要とされると考えられる。将来的にはゲノム解析がゲノムグラフ上で行われることも考えられ、ゲノムグラフという手法が広まるにつれ、グラフゲノムブラウザはゲノム科学者の研究を支えるインフラとなるだろう。

6. 普及（または活用）の見通し

一般の人々が自らの全ゲノムを解読するにはまだ困難も多いため、ゲノムブラウザ自体が広く一般に使われるようなソフトウェアではないが、ゲノム研究者にとってゲノムブラウザは欠かせないツールであり、主に研究者をターゲットとしてユーザの獲得を目論んでいる。

本プロジェクト期間終了後も、複数の研究会での発表が決まっており、普及活動に努めるとともに、コミュニティ活動も並行して行うことを予定している。

7. クリエータ名（所属）

横山 稔之（東京大学大学院）

（参考）関連URL

GitHub: https://github.com/MoMI-G/MoMI-G