فیزینگ سریع دومرحله‌ای داده‌های توالی بزرگ‌مقیاس

پروژه نهایی ژنومیک محاسباتی

جواد راضی

استاد درس: دکتر مطهری

زمستان ۱۴۰۱

فهرست مطالب

- مقدمه
- روش‌ها
- نتایج
- جمع‌بندی

مقدمه

- داده‌های ژنوتایپ
  - عموما فیزنشده
  - نیاز به استفاده از متدهای آماری برای استنتاج توالی الل‌های والدین
- فِیزینگ هپلوتایپ
  - تخمین هپلوتایپ‌هایی که از هریک از والدین به ارث رسیده‌اند
  - تحلیل‌ها و مطالعات متعدد متکی به داده‌های فیزینگ هپلوتایپ

مقدمه

- دقت فیزینگ
  - رابطه مستقیم با اندازه نمونه
- جهش‌های اصلی در روش‌های فیزینگ
  - ابزارهای نوآور: HAPI-UR, SHAPEIT, EAGLE23
  - خطی‌بودن زمان محاسباتی و حافظه مصرفی با اندازه نمونه
  - امکان آنالیز دیتاست‌های بسیار بزرگ‌تر

روش‌ها

- پیاده‌سازی روش مقاله در نسخه ۵٫۲ ابزار Beagle
- مدل HMM برای فیزینگ
  - پنل مرجع: پنل مرجع مرکب هپلوتایپ
- الگوریتم فیزینگ تکرارشونده پیشرو
  - حالت «در حال پیشرفت» یا «پایان‌یافته» برای هر ژنوتایپ هتروزایگوس
  - تغییر حالت یک هتروزایگوس «در حال پیشرفت» به «پایان‌یافته» در پایان هر چرخه
  - تخصیص یک نسبت اطمینان به هر هتروزایگوس در حالت پیشرفت
    - انتخاب هتروزایگوس با بیشترین نسبت اطمینان برای تغییر حالت

روش‌ها

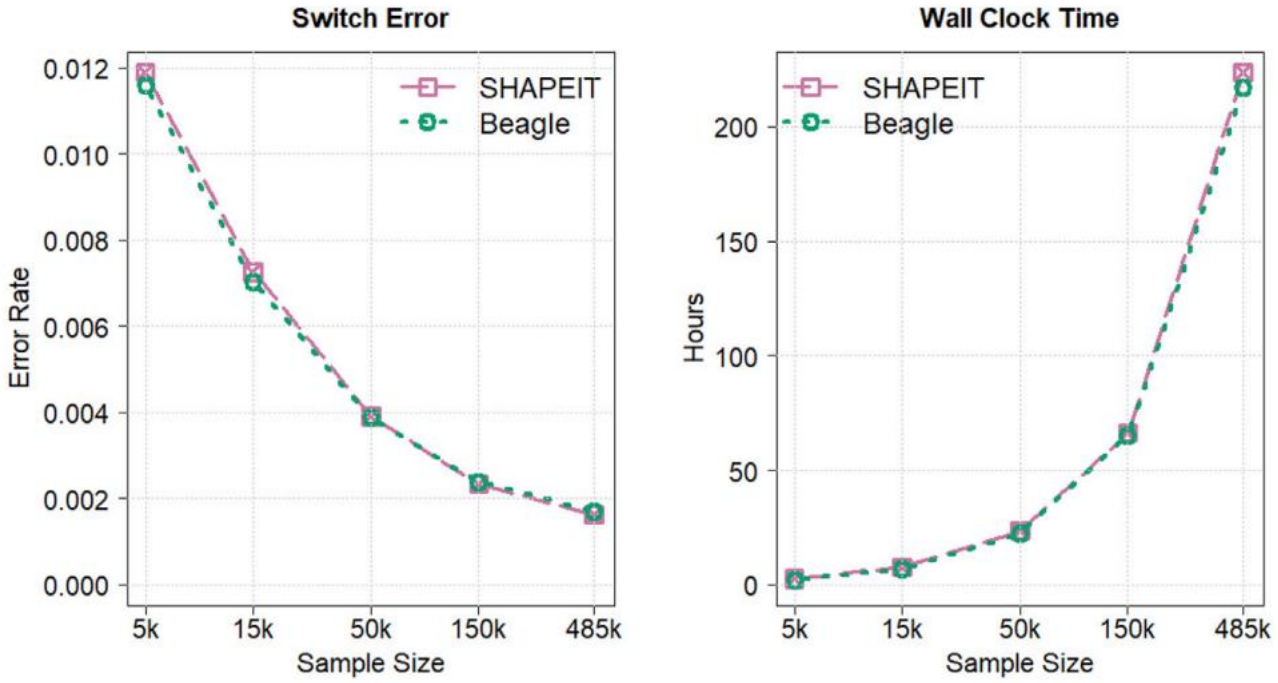
- الگوریتم دو مرحله‌ای برای دیتاست با درصد بالای هپلوتایپ‌های کم‌فرکانس
  - فیزینگ مارکرهای ژنتیکی پرفرکانس در مرحله اول
  - استفاده از هپلوتایپ‌های فیزشده برای جانهی (Imputation) ژنوتایپ‌ها
  - استنتاج هپلوتایپ مارکرهای ژنتیکی کم‌فرکانس با استفاده از الل‌های جانهی‌شده
- پنجره مارکرهای ژنتیکی
  - اندازه قابل کنترل
  - تقسیم توالی به بخش‌های کوچک‌تر
  - فیزینگ مستقل هریک با مدل HMM
  - ترکیب اطلاعات هپلوتایپ‌ها از پنجره‌های مجاور در مرحله دوم، برای افزایش دقت

نتایج

- دیتاست‌های مورد استفاده
  - داده آرایه‌های SNP دیتاست UK Biobank
  - داده توالی ژنوم دیتاست TOPMed
- ارزیابی مدل
  - ابزار مورد استفاده برای مقایسه عملکرد: SHAPEIT
  - از رایج‌ترین ابزارهای مورد استفاده برای فیزینگ
  - قابلیت فیزینگ دیتاست‌های بزرگ، و مقیاس‌پذیری
- سنجه نرخ خطا
  - نرخ خطای سوییچ (SER)
  - نتیجه جابجایی نادرست در اختصاص هپلوتایپ‌ها برای SNPهای مجاور هم
  - حاصل تقسیم تمام خطاهای سوییچ، بر تمام SNPهای فیزشده

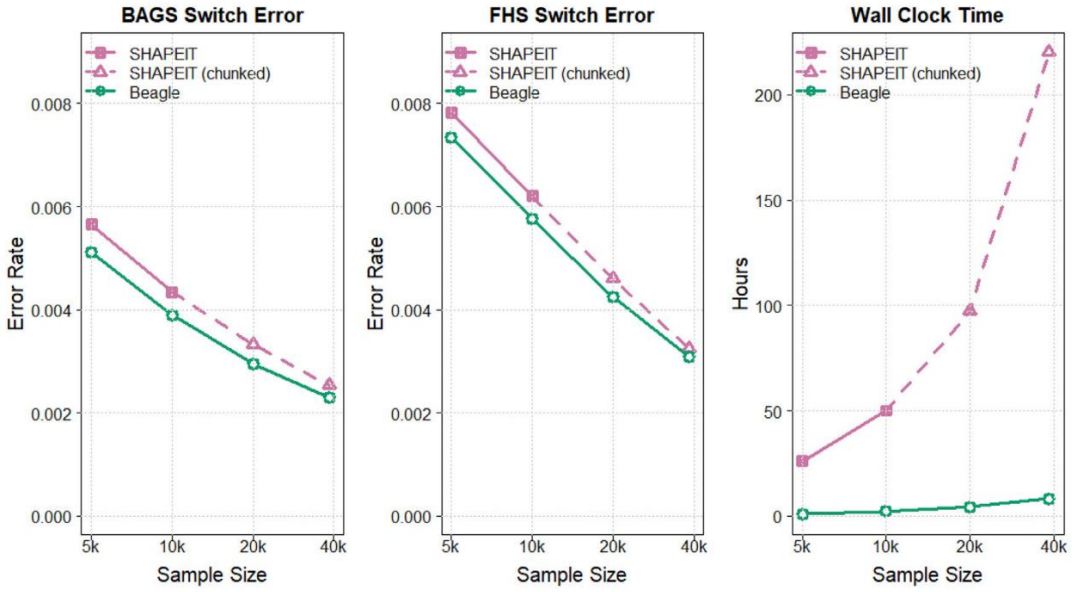
نتایج

- نتایج ارزیابی با دیتاست UK Biobank
  - عملکرد یکسان از لحاظ خطای SER و سرعت محاسباتی



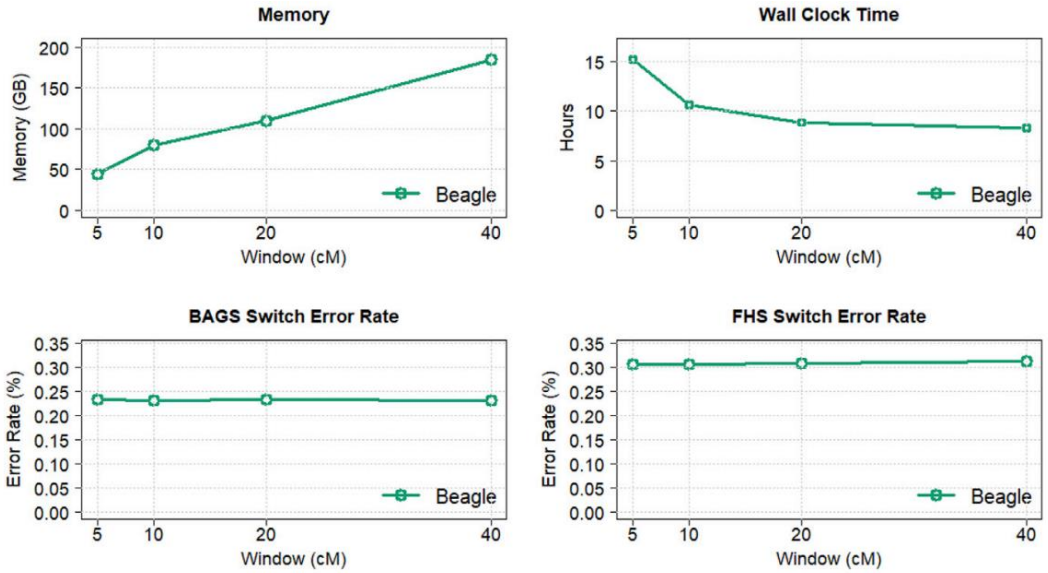
نتایج

- نتایج ارزیابی با دیتاست TOPMed
  - عملکرد یکسان از لحاظ خطا
  - سرعت بیش از ۲۰ برابری Beagle 5.2 نسبت به SHAPEIT 4.2.1 روی این دیتاست



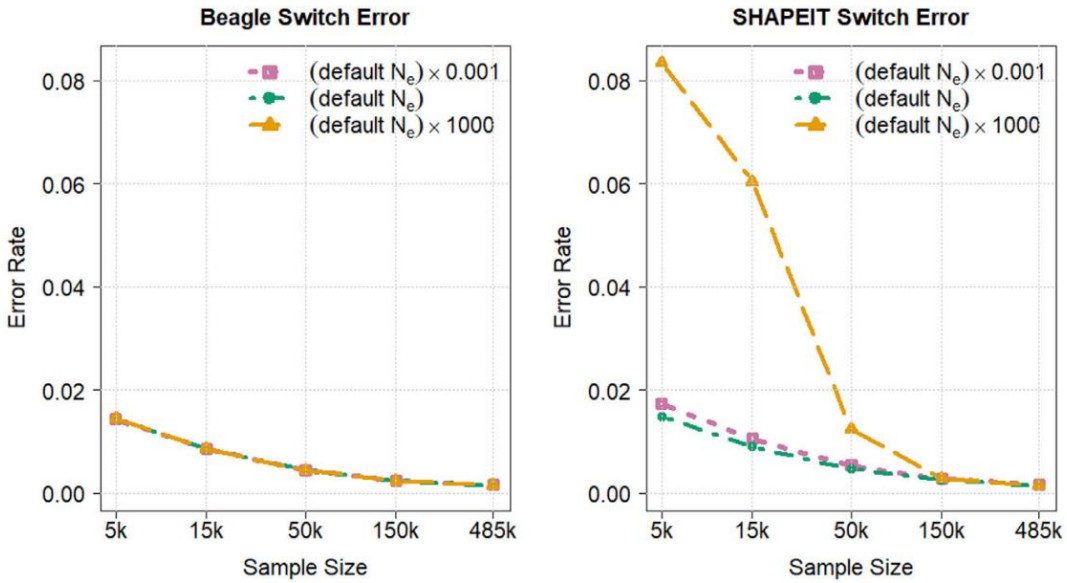
نتایج

- کنترل میزان حافظه مصرفی با تغییر اندازه پنجره مارکر
  - کوتاه کردن اندازه پنجره
  - امکان فیزینگ دیتاست‌های بزرگ‌تر با کاهش حافظه مصرفی
  - Trade-off: افزایش زمان محاسباتی
  - عدم وجود تاثیر محسوس تغییر اندازه پنجره در نرخ خطا



نتایج

- اندازه جمعیت موثر
  - تعریف «جمعیت موثر»: جمعیتی که از لحاظ گونه‌گونی ژنتیکی، مشابه جمعیت واقعی باشد.
  - پارامتر ورودی مدل
  - تاثیر در محاسبات احتمالات گذار در HMM
- تاثیر مقدار اولیه اندازه جمعیت موثر
  - مدل SHAPEIT
    - نوسانات در دقت هنگام دور شدن مقدار اولیه از مقدار واقعی
  - مدل Beagle
    - مستقل از مقدار اولیه جمعیت موثر
    - عدم تاثیر نادرستی پارامتر ورودی در دقت مدل



جمع‌بندی

- فیزینگ هپلوتایپ: تخمین هپلوتایپ‌ها از داده‌های ژنوتایپ
- مزیت‌های مدل ارائه‌شده
  - سریع، دقیق، استفاده بهینه از حافظه
  - مقیاس‌پذیری بالا و قابلیت فیزینگ دیتاست‌های بزرگ
- داده‌های مناسب مدل
  - داده‌های با تعداد بالای مارکرهای ژنتیکی کم‌فرکانس
  - استفاده از الگوریتم دومرحله‌ای فیزینگ
    - عدم فیزینگ داده‌های کم‌فرکانس در مرحله نخست
    - افزایش سرعت محاسباتی

جمع‌بندی

- پیاده‌سازی مدل
  - پیاده‌سازی در نسخه 5.2 نرم‌افزار Beagle
- ارزیابی مدل
  - مقایسه با ابزار SHAPEIT 4.2.1
  - دیتاست‌های مورد استفاده برای جمعیت نمونه
    - UK Biobank
    - TOPMed
  - سرعت محاسباتی بیش از ۲۰ برابر در فیزینگ با داده‌های TOPMed

مراجع

- Browning BL, Tian X, Zhou Y, Browning SR. Fast two-stage phasing of large-scale sequence data. Am J Hum Genet. 2021 Oct 7;108(10):1880-1890. doi: 10.1016/j.ajhg.2021.08.005. Epub 2021 Sep 2. PMID: 34478634; PMCID: PMC8551421

سپاس از توجه‌تان