KIRAN V GARIMELLA

HYPOTHESIS-FREE DETECTION OF GENOME-CHA EVENTS IN PEDIGREE SEQUENCING

WELLCOME TRUST CENTRE FOR HUMAN GENETICS

PUBLISHED BY WELLCOME TRUST CENTRE FOR HUMAN GENETICS

WWW.WELL.OX.AC.UK

First printing, November 2014

Contents

List of Figures

List of Tables

Abstract

This is a working draft of my dissertation.

1 Introduction

1.1 Introduction

This is where the introduction goes.

2 Background

2.1 How genome changes

2.1.1 Cross-over

2.1.2 Gene conversion

2.1.3 Point mutations

2.1.4 Structural variants

2.1.4.1 Small (indels)

2.1.4.2 Large (fusions, NAHR)

2.1.4.3 Chromosomal changes

2.2 Rates

2.3 Factors influencing

2.3.1 Replication time

2.3.2 Mat/pat age effects

2.3.3 Biases/locality

2.4 Known events in species

2.4.1 P.f.

2.4.2 Human

2.4.3 Chimp

2.4.4 Others

3 Detection

3.1 Basic design

3.2 Current state of the art

3.2.1 Cross-over

3.2.2 Gene conversion

3.2.3 Point mutations

3.2.4 Structural variation

3.2.5 Chromosomal

3.3 Potential for de novo assembly to detect events

3.3.1 Why assembly should theoretically work

3.3.2 Haploids/perfectly assembled diploids

3.3.3 What do we look for

3.4 Limitations of read data

3.4.1 Sequencing errors

3.4.2 Read length / repeat length

3.4.3 Coverage fluctuations

3.4.4 Rare vs error

3.4.5 Algorithm error can mimic real biology

3.5 Outline of the work

4 *Methods*

4.1 Overview

4.1.1 Start with NGS data from mother, father, child

4.1.2 Need to identify relevant motifs within data

4.1.2.1 Discovery/exploration

4.1.2.2 Validation

4.1.2.3 Interpretation

4.2 Discovery/exploration

4.2.1 Assembly

4.2.2 Annotation of kmers and links

4.2.3 "Fishing"

4.2.4 Visualization

4.3 Validation

4.3.1 In silico

4.3.1.1 Contig decoration

4.3.1.2 Decision (trust / not trust)

4.3.1.3 Simulations

4.4 Empirical

4.4.1 Known AHRs

4.4.2 Known NAHRs

4.4.3 Comparison of 3D7 (Illumina) to 3D7 (ref), using 3D7 (PacBio) to adjudicate

4.5 Experimental

4.5.1 PacBio

4.5.2 Sanger

4.6 Interpretation

4.6.1 Align

4.6.2 Classify

4.6.3 Compare to existing events

5 Pf

5.1 Lit review

5.1.1 Review of Kong et al., 2002

Augustine Kong et al. discuss a new genetic map of recombination rates using genotyping information from 869 individuals in 146 Icelandic families. This is the first such map made after the sequencing of the human genome, and is thus able to leverage the new reference sequence in order to correctly order the genotyped markers. It is a substantially higher-resolution map than provided by the former gold-standard, the Marshfield map. The Marshfield map contained data on only 188 meioses, whereas the Kong et al. map contained data on 1, 257. The new map reveals marked differences in recombination rates between males and females (e.g. the recombination rate in female autosomes is a factor of 1.65 higher than that observed in males) for reasons beyond sequence features.

6 *Chimp*

6.1 *Lit review*

6.1.1 *Review of Kong et al., 2002*

Augustine Kong et al. discuss a new genetic map of recombination rates using genotyping information from 869 individuals in 146 Icelandic families. This is the first such map made after the sequencing of the human genome, and is thus able to leverage the new reference sequence in order to correctly order the genotyped markers. It is a substantially higher-resolution map than provided by the former gold-standard, the Marshfield map. The Marshfield map contained data on only 188 meioses, whereas the Kong et al. map contained data on $1,257$. The new map reveals marked differences in recombination rates between males and females (e.g. the recombination rate in female autosomes is a factor of 1.65 higher than that observed in males) for reasons beyond sequence features.

7 *Discussion*

7.1 *Lit review*

7.1.1 *Review of Kong et al., 2002*

Augustine Kong et al. discuss a new genetic map of recombination rates using genotyping information from 869 individuals in 146 Icelandic families. This is the first such map made after the sequencing of the human genome, and is thus able to leverage the new reference sequence in order to correctly order the genotyped markers. It is a substantially higher-resolution map than provided by the former gold-standard, the Marshfield map. The Marshfield map contained data on only 188 meioses, whereas the Kong et al. map contained data on $1,257$. The new map reveals marked differences in recombination rates between males and females (e.g. the recombination rate in female autosomes is a factor of 1.65 higher than that observed in males) for reasons beyond sequence features.

8 Bibliography