De novo assembly and genotyping of variants using colored de Bruijn graphs [Zamin Iqbal et al.]

Damianos Melidis

ETH - University of Zurich

dmelidis@student.ethz.ch

December 2, 2013

Overview

Motivation

Goal:

- Detect genetic variation highly divergent from reference
- Identify variants between samples when reference is not available

Previous approach $\leftarrow$ Mapping

Mapping Limitations:

- Sample subsequences divergent from reference
- Incomplete reference for some genome regions
- Studied samples with no available reference

Solution

Cortex!

- Alignment free!
- Handling a population of samples, by extending de Bruijn graphs
- Based on a statistical model, trying to identify the variant class
- Based on a statistical model, show how to abjust method parameters for capturing a new experiment

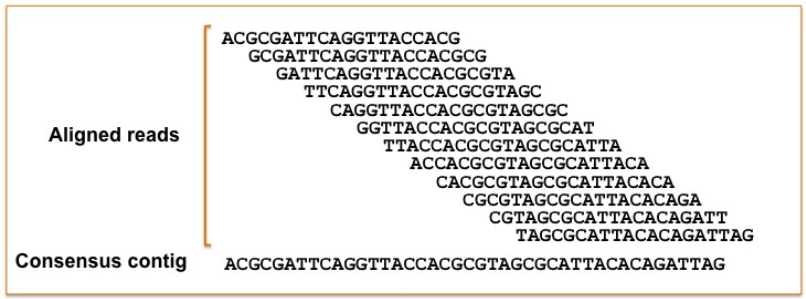
de Bruijn Graph (Introduction)

Sequence Assembly [1]:

Aligning and merging short fragments of DNA to reconstruct the original sequence

De Bruijn graph: [Philip E C Compeau et al.]

Contig:



[2]

[1] http://en.wikipedia.org/wiki/Sequence_assembly

[2] http://contig.files.wordpress.com/2010/02/alignment1.jpg

de Bruijn Graph (Proposed Model)

Starting from a de Bruijn (directed) graph:
Node: k-mer from DNA alphabet
In\out degree: number of edges going in/out of a node
Generalize to a multicolor graph with a different color for each allele of an individual and an unique color for the reference
Supernode: **maximal** length path with only the first and last node having in/out degree $\neq 1$
Bubble: A pair of supernodes with the same start and end nodes
Branch: Each supernode in a bubble
Tip: A short path ending in a node with out-degree 0
Confounded bubble: A non-callable variant because of overlapping with another part of the genome
Effective coverage: The true coverage of each read in this graph
Effective read-length: The true parsed bases of a read and the genome in this graph

Visual Examples (Proposed Model) :)

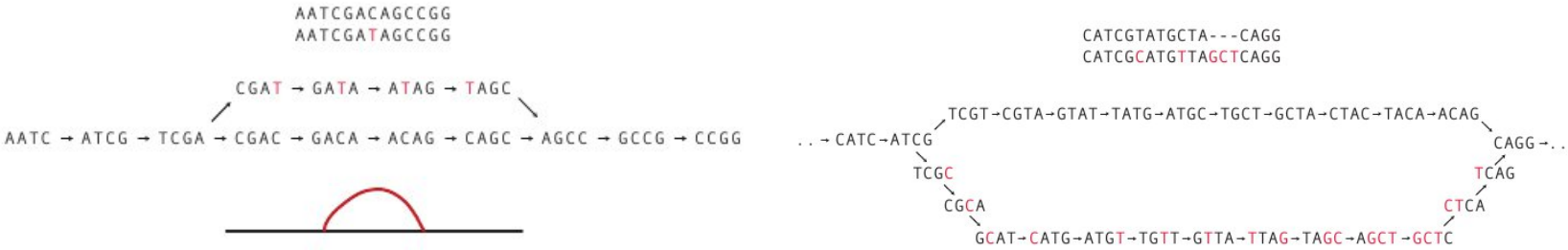


Figure : Simple SNPs(Left), Deletion with 2 SNPs(Right)

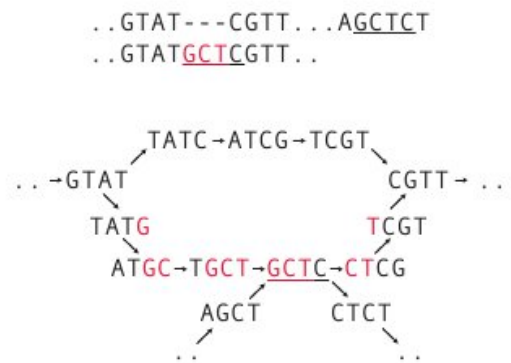


Figure : Confounded bubble

Genome Complexity

Estimating Genome Complexity

For simple SNPs:

- Take the human chromosome 1 from NCBI36, varying k-mer size:
- Create all possible simple SNPs
- If the resulted path forms a "clear" supernode
- and the reference allele creates a "clear" supernode
- then the variant is callable
- G(,k) = ♯ callable variants / ♯ all possible variants

For more complex polymorphisms:

- Defining G(t,) varying k-mer size:
- The probability to call a M bps polymorphism approximates the probability a whole M bps contig fits in a supernode

Poisson Model[3]

How to model the read sampling from a genome with length G?
Coverage: $D = \frac{N*R}{G}$
Let's think a queue with size the read length $L = R$ (average service time) and $\lambda = \frac{D}{R}$ reads per sequenced base (average arrival time)
Using Little's Law in our queue the average number of reads (customers): $C = \lambda * L = D$
In de Bruijn graph the *effective read length* $L_{eff} = R - k + 1$, so $D_{eff} = \lambda * L_{eff}$

[3]Help from Google search: "Chapter 5.1 Lander-Waterman Statistics for Shotgun Sequencing, Prof Tesler"

More Poisson

Defining the probability E that a variant with length t is present in graph: Preposition: If C is defined as the distance between the starts of the first and last reads in a contig, then the probability distribution of C is

$P(C \geq t) \cong (1 - e^{-\lambda L})e^{-\lambda e^{-\lambda L} t} Ind(t > 0) + e^{-\lambda L} Ind(t = 0)$

Corollary: The probability P that an allele of length d is present in the graph is approximated by:

$P = (1 - e^{-\lambda L})P(C \geq d)$

Resulting to $V(t) = (1 - e^{-\lambda L})P(C \geq t)$

Error Model

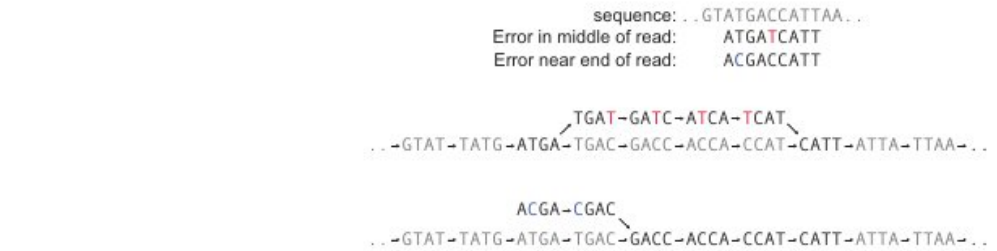
The primary effect of sequencing error reduces the arrival time by a 1-kε
2 errors types: slow + fast, on average $\hat{\varepsilon} = \sqrt{0.001 * 0.05}$
Tip clipping: Remove tips of length at most k+1
Guaranteed for k $> \frac{R}{2}$ and single error in a single read
But if the error occurs in the middle of the read forming a full bubble escapes tip clipping
P(escape tip clipping) $= \frac{80-k+1}{80}$
Removing low coverage supernodes: Remove supernodes with coverage $\leq$ 1 or 2 for every interior node
Repeated errors: at least 2 errors in the k-mer of a variant site create supernodes with some nodes coverage at least 2
Define the probability E that a site is callable despite all errors:

E = 1 - (k-1 + k-1) * $(\sum_{i \geq 2} dpois(i, D_{\text{eff}})) \sum_{n=2}^{i} binom(i, \frac{2\hat{e}}{3}, n)(1 - \frac{1}{2^{n-1}}) \sum_{j \geq 2} dpois(j, \lambda k)(1 - \varepsilon)^{\mu j}$ *

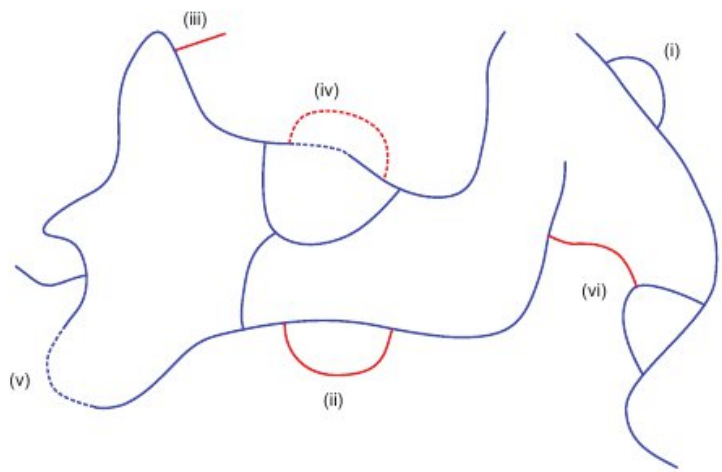
P(escape tip clipping)

Visual Examples of Error Model :)

a

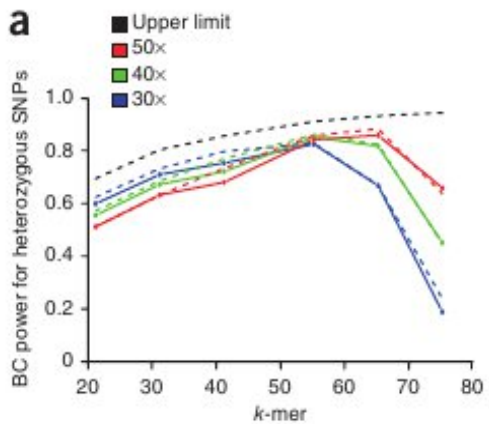


b



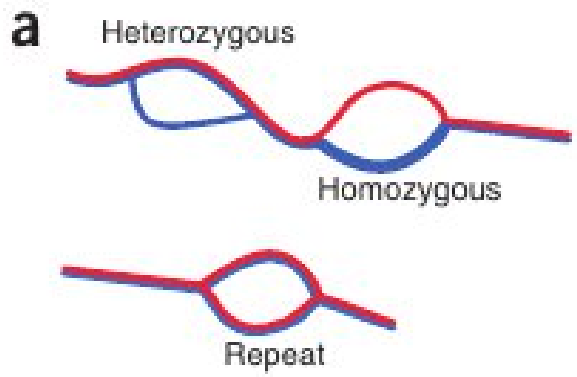
Power of Variant Discovery

For a given k-mer size, and variant size t, the probability of discovering a homo/heterozygous variant:
Power(t,k) = probability(Species genome enables to call it) * probability(Entire variant into one supernode) * probability(Variant is callable despite errors)
Incorporating previous analysis: Power(t,k) = G(t,k) * V(t) * E



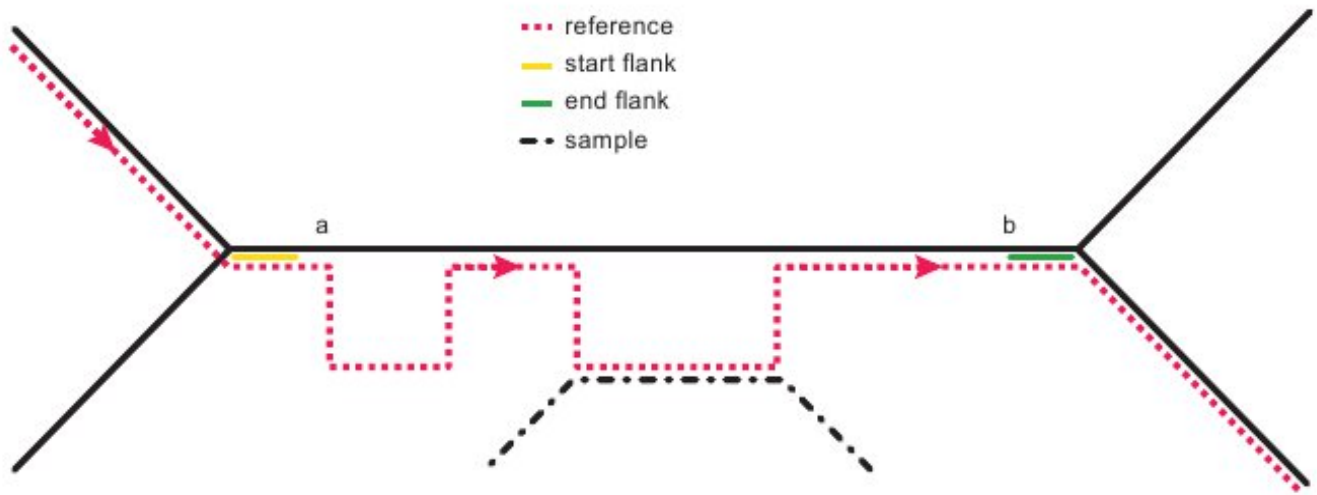
Bubble Caller

Pseudocode:
Start with all nodes unvisited for each node u in the graph:
if outdegree(u) is 2 and u is unvisited:
mark u as visited
get the supernodes S1 and S2 from u
mark the nodes in S1 and S2 as visited
if last_node(S1) equals last_node(S2) AND orentation_is_correct(S1,S2):
bubble_found = true



Path Divergence Caller

Goal: Discover more complex polymorphisms(like deletions)
Idea:
Follow the path of the joint reference-sample
Identify points where the reference breaks away from the sample and then return again on it
Restricting to: Both breakpoints appear in a single supernode of sample graph



Multiple Sample Analysis (Definition)

Using many individuals from a population classify if the bubble created by variation, error or repeat

Idea: Each bubble type is a probabilistic model

Priors:

Variation model: each individual has 0,1,2 copies of the former allele and 2,1,0 of latter allele

Prior = binomial(2,x), x population allele frequency

Repeat model: intermediate allele balance with no variation

Prior = symmetric beta(2,2)

Error model: the allele with error will have substantial low coverage

Prior = beta(100*ε,100)

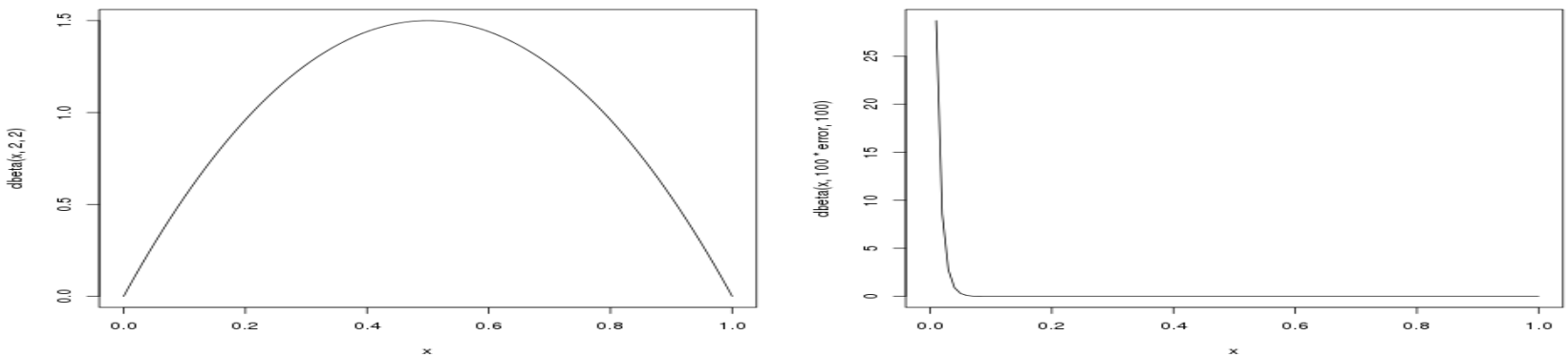


Figure : Prior for repeat and error (left and right respectively)

Multiple Sample Analysis (Model Selection)

Use Bayes factor K to select the best model:

$$K = \frac{Pr(Data|M_1)}{Pr(Data|M_2)} = \frac{\int Pr(\theta_1|M_1)Pr(Data|\theta_1, M_1)d\theta_1}{\int Pr(\theta_2|M_2)Pr(Data|\theta_2, M_2)d\theta_2}$$

Select model with log10(K) $\geq$ 10

Genotyping Algorithm

Given a multi-color de Bruijn graph with different color for each allele and an unique color for the reference

For each alleles(paths) γ_1, γ_2 do:

1.Ignore any segment in reference but not in sample

2.Decompose both paths into shared s_i and unique q_i and count the number of reads in each section

3.Count number of nodes in 2 contig (with length = length of γ_1 and γ_2) but not in shared and unique section

4.Following the Poisson model in a section with length l_i the average number of arriving reads r_i in this section is a Poisson with rate $\theta_i = \lambda * l_i$ and $\frac{\theta_i}{2}$ for homozygous and heterozygous alleles

The approximate likelihood for each possible genotype:

$$P(\text{genotype} = \gamma_1 \cup \gamma_2 \mid \text{Data}) = \prod_{s_i} \theta_i^{r_i} * \frac{e^{-\theta_i}}{r_i!} \prod_{u_i} (\frac{\theta_i}{2})^{r_i} * \frac{e^{\frac{-\theta_i}{2}}}{r_i!} S(n)$$

Implementation

Cortex!

Graph is implicitly represented as a hash table with:

hash-key the binary representation of a k-mer (and its alternative in DNA alphabet)

hash-value the a object representing a node (storing coverage and out-nodes for each color)

The time complexity scales linearly to the size of de Bruijn graph

The memory scales linearly for each sample, k-mer size and total number of graph nodes

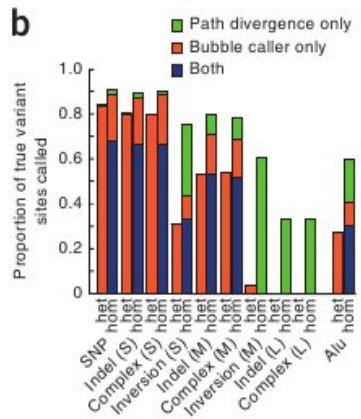
Features:

Cortex is the only assembler handling multiple eukaryote genomes

The proposed algorithms can be parallelized

Simulation 1: Single High-Coverage Diploid Genome

Simulate Bubble and Path Divergence Caller to 10-50x sequencing data from a diploid human sample
Observe that: Increasing k-mer size the propability of having an error in a contig is smaller but the probability of the k-mer having the error is increased.
For coverage and k-mer size boosting Bubble Caller, compare Bubble and Path Divergence Caller:



For small polymorphisms 80-90% (heterozygous-homozygous) is detectable
For polymorphisms with moderate size 50-75% (heterozygous-homozygous) can be identified
For large polymorphisms only 35% of homozygous can be detected

Simulation 2: Population-Based Variant Calling

Analyze data from chromosome 22 for 10 human individuals with error-free and error-containing reads

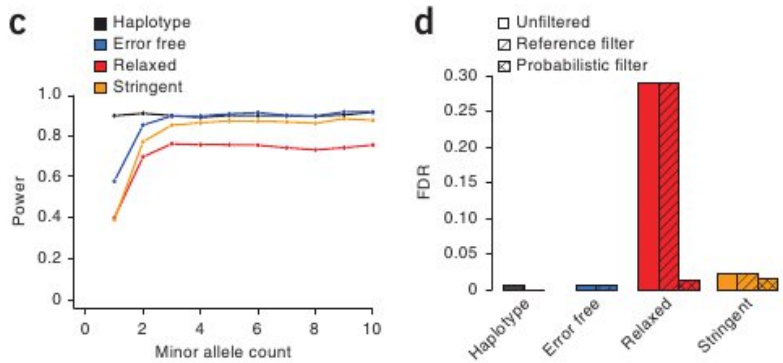


Figure : Power to detect SNP variants using BC(left), FDR for left sets for classifying a polymorphism as variant, repeat or error

From left figure →sufficient coverage for losing power only in rare variants

From right figure →probabilistic model selection minimizes the FDR

Case 1: Variant Calling in High-Coverage Genome

Analyze single human individual to compare Cortex variant calling with mapping-based approaches from 1000 Genomes Project

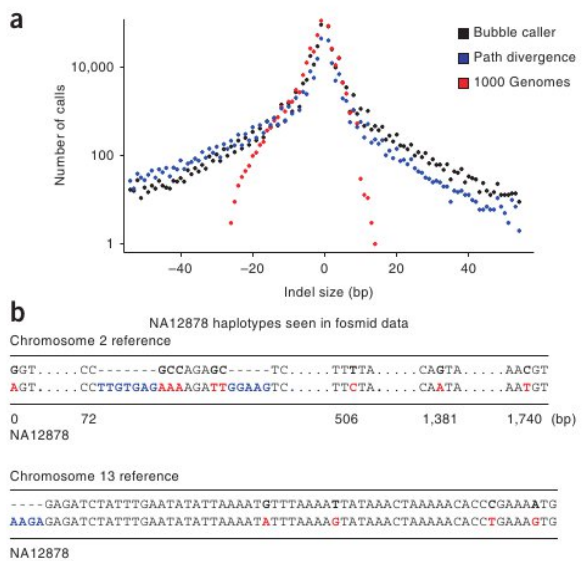
Table 1 Comparison of 1000 Genomes and Cortex calls to fosmid data

		Cortex		Bubble caller		Path divergence	
Variant type	1000 Genomes[a]	All	High confidence[b]	All	High confidence	All	High confidence
SNP (Hom.)	1,085 (0)	1,071 (4.0)	605 (0.5)	1,057 (3.9)	591 (0.5)	340 (8.5)	144 (1.4)
SNP (Het.)	2,350 (28)	1,155 (32)	1,029 (32)	1,155 (32)	1,029 (32)	0 (–)	0 (–)
Indel (Hom.)	64 (0)	96 (6.3)	20 (0.0)	79 (6.3)	16 (0.0)	37 (5.4)	5 (0.0)
Indel (Het.)	127 (29)	67 (40)	43 (30)	67 (40)	43 (30)	0 (–)	0 (–)
Complex (Hom.)	–	258 (1.9)	202 (1.5)	112 (2.7)	77 (1.3)	174 (1.7)	139 (2.2)
Complex (Het.)	–	161 (26)	137 (25)	161 (26)	137 (25)	0 (–)	0 (–)

Het., heterozygous; Hom., homozygous.

[a]Values reported are the number of each variant per genotype combination called, and those in parentheses are the percentage of cases in which only the reference allele was observed in the fosmid sequence data. [b]High-confidence call set requires $\log_{10}$ (Bayes factor) for the reported genotype to be at least 4.

The mapping-based approach are better in a smaller region of the true reference ,but,



The mapping-based approach detects only small variants, but BC and PD do detect longer-complex variants

Case 2: Detection of novel sequence from population graphs

Construction of three pooled human population (CEU, YRI, CHB) sequenced in low coverage from 1000 Genomes Project and add the reference as a fourth color

Cortex identified 21,000 novel contigs of $\geq$ 100 bp

Some sequences showed strong "preference" towards a single population

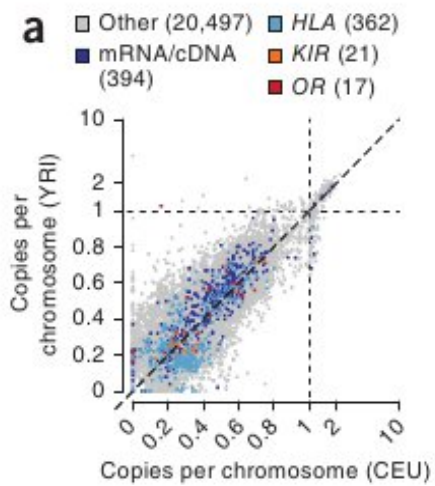
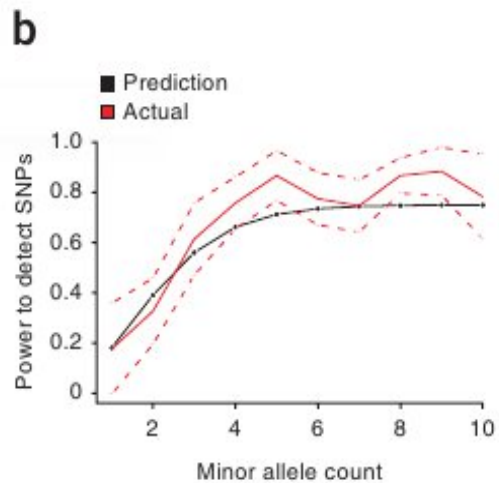


Figure : Estimates of mean copy for novel contigs for YRI and CEU populations

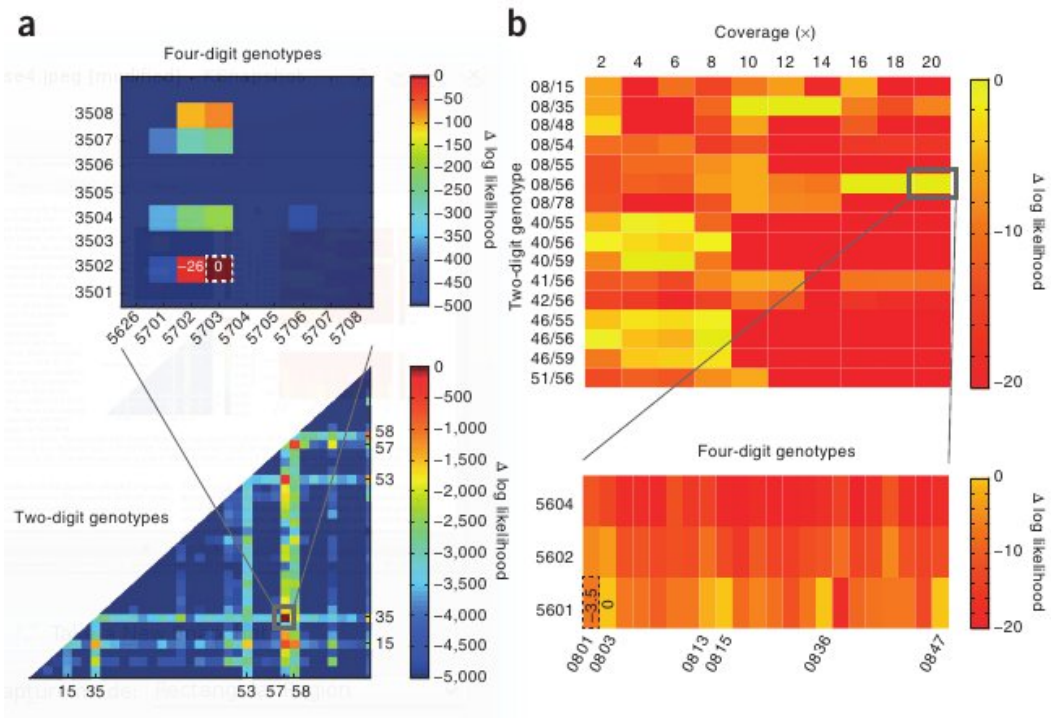
Case 3: Using Population to Classify Bubbles

Apply probabilistic classification for bubbles from 10 Western Chimpanzees
Use (relaxed) supernode cleaning
Identification of 3.5 million polymorphisms
Bubbles classified as variants have FDR 3.5%!



Case 4: Genotyping Simple and Complex Variants

Apply genotype algorithm for simple HapMap2 SNPs and complex variants HLA-B genotypes from two human individuals
Classical HLA genotypes are important in some areas of medical genetics
Compare well established approach (DNA sequencing) with HTS analysis
Construct the reference genome and all known HLA-B alleles, Cortex did find a genotype that agrees with DNA sequencing (for the first sample)
For the second sample Cortex is not verified by DNA sequencing



Limimitations

- Paired ends reads are not supported
- A better error correction as k increases
- Current implementation may lead to graph explosion adding more individuals

Questions?

Thanks for your attention :)

References

Zamin Iqbal et al.

De novo assembly and genotyping of variants using colored de Bruijn graphs

Nature Genetics 44, 226—232 (2012).

Philip E C Compeau et al.

How to apply de Bruijn graphs to genome assembly

Nature Biotechnology 29, 987—991 (2011).