Identifying recurrent mutations in population-level sequencing data

Kelsey Johnson
SAGES
June 1st, 2018
What is a recurrent mutation?

acggaagctag
acggaagctag
acggaagctag
acggaagctag
acggaagctag
acggaagctag
acggacgctag
acggacgctag
acggacgctag
acggaagctag
acggaagctag
acggaagctag
What is a recurrent mutation?

Identical by descent (IBD):

= mutation event
What is a recurrent mutation?

Identical by descent (IBD):

Recurrent:

* = mutation event
Why care about recurrent mutations?
Recurrent mutations are a hallmark of some Mendelian diseases

<table>
<thead>
<tr>
<th>Gene</th>
<th>Disease</th>
</tr>
</thead>
<tbody>
<tr>
<td>CFTR</td>
<td>cystic fibrosis</td>
</tr>
<tr>
<td>SCN8A</td>
<td>epileptic encephalopathy</td>
</tr>
<tr>
<td>PKD1</td>
<td>polycystic kidney disease</td>
</tr>
<tr>
<td>FGFR1</td>
<td>Pfeiffer syndrome</td>
</tr>
<tr>
<td>FGFR3</td>
<td>achondroplasia</td>
</tr>
<tr>
<td>LMNA</td>
<td>Hutchinson–Gilford progeria syndrome</td>
</tr>
</tbody>
</table>
Recurrent mutations are used to identify genes associated with complex disease

ARTICLE
Received 15 Sep 2014 | Accepted 16 Oct 2014 | Published 24 Nov 2014
DOI: 10.1038/ncomms6595

Recurrent *de novo* mutations implicate novel genes underlying simplex autism risk


1 Department of Genome Sciences, University of Washington School of Medicine, Seattle, Washington 98195, USA.
2 Department of Psychiatry and Behavioral Sciences, University of Washington, Seattle, Washington 98195, USA.
3 Howard Hughes Medical Institute, University of Washington, Seattle, Washington 98195, USA.

w Present address: Department of Molecular and Medical Genetics, Oregon Health and Science University, Portland, Oregon 97239, USA. * These authors contributed equally to this work. Correspondence and requests for materials should be addressed to J.S. (email: shendure@uw.edu) or to E.E.E. (email: eee@gs.washington.edu).

NATURE COMMUNICATIONS | 5:5595 | DOI: 10.1038/ncomms6595 | www.nature.com/naturecommunications

© 2014 Macmillan Publishers Limited. All rights reserved.
These studies rely on family-based sequencing to identify recurrent mutations
Family-based study
Family-based study

Population-based study
What features can distinguish recurrent and IBD alleles?
Differences in $t_{\text{MRCA}}$ for IBD vs. recurrent alleles

Identical by descent (IBD):
Differences in $t_{MRCA}$ for IBD vs. recurrent alleles

Identical by descent (IBD):

Recurrent:
Population-level sequencing data with diploid genotypes
Mathieson & McVean, 2014
If the $t_{\text{MRCA}}$ of two alleles is known, the conditional probability distribution of the recombination distance is:

$$f(d_L \mid t_{\text{MRCA}})$$
If the $t_{\text{MRCA}}$ of two alleles is known, the conditional probability distribution of the recombination distance is:

$$f(d_L \mid t_{\text{MRCA}})$$

With the probability distribution of the $t_{\text{MRCA}}$ for recurrent and IBD alleles, we can calculated the probability of $d_L$:

$$f(d_L) = \int_{t_{\text{MRCA}}} f(d_L \mid t_{\text{MRCA}}) f(t_{\text{MRCA}}) \, dt_{\text{MRCA}}$$
Theory vs. data: recurrent mutations

UK10K multiallelic 8ton
Theoretical recurrent 8ton

Density

Recombination distance (cM)
Theory vs. data: IBD mutations

UK10K biallelic 8tons
Theoretical IBD 8tons
Recombination distances follow a predictable pattern:

- Short $t_{MRCA}$, long rec. dist.
- Long $t_{MRCA}$, short rec. dist.
Recombination distances follow a predictable pattern.

**short** $t_{\text{MRCA}}$, **long** rec. dist.  
**long** $t_{\text{MRCA}}$, **short** rec. dist.
Statistical approach

• Calculate likelihood of observed data under 2 scenarios (IBD or recurrent):
  – Recombination distances on right & left hand sides
Statistical approach

• Calculate likelihood of observed data under 2 scenarios (IBD or recurrent):
  – Recombination distances on right & left hand sides
  – Distance ranks on right & left hand sides
Statistical approach

• Calculate likelihood of observed data under 2 scenarios (IBD or recurrent):
  – Recombination distances on right & left hand sides
  – Distance ranks on right & left hand sides
• Compute test statistic of composite likelihood ratio
Statistic performance depends on allele count

![Graph showing TPR vs. FPR for different allele counts: 2, 6, and 10.](image)
Application to UK10K: CpG enrichment
Application to UK10K: CpG enrichment
What’s next?

• Application to empirical datasets (e.g. UK10K)
  – Updated measurement of SFS
What’s next?

- Application to empirical datasets (e.g. UK10K)
  - Updated measurement of SFS
  - Mutation rate variation
What’s next?

• Application to empirical datasets (e.g. UK10K)
  – Updated measurement of SFS
  – Mutation rate variation

• Rare variant burden tests
Thank you!

Voight Lab
Ben Voight
Paul Babb
Diana Cousminer
Kat Gawronski
Kim Lorenz
Katie Siewert
Chris Thom

Thesis Committee
Casey Brown
Maja Bucan
Struan Grant
Sarah Tishkoff

Funding
Genetics Training Grant
T32GM008216