Session 2: Analysis of single nucleotide variants and indels

Peter J Park
Department of Biomedical Informatics
Harvard Medical School
July 18-19, 2016
Some basics

- SAM (Sequence Alignment Map) files
- BAM: binary version of SAM files
Genome alignment

- Bowtie (typical ChIP-seq, RNA-seq, etc)
- BWA (if you care about indels, translocations, etc)
Alignment

- One can specify how many mismatches are to be tolerated
- This can also be quantified by accounting for quality scores
- A typical criterion might be 1-3 mismatches for 100bp reads
- From the raw sequences, ~80-90% of the reads are typically aligned to the genome
  - sequencing errors
  - multiple matches in the genome (how to handle this?)
  - deviations from the reference genome (SNPs, insertions, etc)
  - problems with the aligner / genome assembly
- The % of mapped reads is a good measure of data quality
- Often need to normalize using an “alignability map”
Example of somatic mutation in IGV (chr2:179611406-179611406; TTN stopgain mutation)
Recurrent insertions before poly-Q repeats

- Chr12:132547068-132547069, In_Frame_Ins, p.Arg2755X/c.8265*>+GCA
Read length and alignability

uniquely mappable genome reads with zero mismatches

% uniquely mappable reads

reads length (bp)

D.Miranda
hg18
mm9
m4
c_elegans
a_thaliana
s_cerevisiae
zebrafish
d_melanogaster_fb5_22_NOhet
d_melanogaster_fb5_22
Quality scores

• Each base position in a sequence comes with a “quality score”
• This measures the probability that a base is called incorrectly, by a phred-like algorithm similar to that originally developed for Sanger sequencing experiments.
• The quality score of a given base, Q, is defined by
  \[ Q = -10\log_{10}(e) \]
  where \( e \) is the estimated probability of the base call being wrong.
• A quality score of 20 represents an error rate of 1 in 100, with a corresponding call accuracy of 99%.
Quality scores
Every platform has problems

• Reads are short
  - difficult to assemble/map repetitive regions

• Not all sequences are equally likely to be sequenced
  - GC content
  - fragment length

• Amplification bias

• Sequencing errors
  - especially toward the end
Extreme GC regions are hard to sequence

- Some regions do not get a lot of reads
- Biggest reason is due to extreme GC content
- It hard to call SNVs and indels if you don’t have a lot of reads!
Targeted sequencing

- ‘Capture and release’ techniques
- Exome or a custom set of genes
- Platforms
  - Agilent In-solution capture
  - Nimblegen arrays
- Various other methods for validation
  - Droplet-based multiplex PCR (RainDance)
  - Custom methods
- Cost vs coverage
- Some regions are captured more effic
Performance comparison of exome DNA sequencing technologies

Michael J Clark¹,⁴, Rui Chen¹,⁴, Hugo Y K Lam¹, Konrad J Karczewski¹, Rong Chen², Ghia Euskirchen¹,³, Atul J Butte² & Michael Snyder¹,³
Exome capture platforms

![Graph showing percent of targeted bases vs depth of coverage for different platforms: Nimblegen, Agilent, Illumina.](image-url)
Whole-genome sequencing is more powerful than whole-exome sequencing for detecting exome variants

Aziz Belkadi\textsuperscript{a,b,1}, Alexandre Bolze\textsuperscript{c,1,2}, Yuval Itan\textsuperscript{c}, Aurélie Cobat\textsuperscript{a,b}, Quentin B. Vincent\textsuperscript{a,b}, Alexander Antipenko\textsuperscript{c}, Lei Shang\textsuperscript{c}, Bertrand Boisson\textsuperscript{c}, Jean-Laurent Casanova\textsuperscript{a,b,c,d,e,3,4}, and Laurent Abel\textsuperscript{a,b,c,3,4}

\textsuperscript{a}Laboratory of Human Genetics of Infectious Diseases, Necker Branch, INSERM U1163, 75015 Paris, France; \textsuperscript{b}Paris Descartes University, Imagine Institute, 75015 Paris, France; \textsuperscript{c}St. Giles Laboratory of Human Genetics of Infectious Diseases, Rockefeller Branch, The Rockefeller University, New York, NY 10065; \textsuperscript{d}Howard Hughes Medical Institute, New York, NY 10065; and \textsuperscript{e}Pediatric Hematology-Immunology Unit, Necker Hospital for Sick Children, 75015 Paris, France

- WES vs WGS in six individuals
- WES target region (81.5\% of the consensus coding genome):
  - WES: 84192 SNVs, 13325 indels; WGS: 84,968, 12702.
- A mean of 105 coding HQ SNVs and 32 indels was exclusive to WES; 692 HQ SNVs and 105 indels exclusive to WGS.
- Sanger: FP for WES was 78\%, for WGS 17\%, for SNVs, ~45\% for WES/WGS for indels
- The vast majority of SNVs and indels were identified by both techniques, but ~650 high-quality coding SNVs (~3\% of coding variants) were detected by WGS and missed by WES.
GATK (Genome Analysis Toolkit)
Indel realignment

• Indels (especially near the ends) can trick the mappers into mis-aligning with mismatches

• We can improve the original alignments of the reads based on multiple sequence (re-)alignment

• There are many other tools to remove noise in the GATK pipeline. Now version v3.0
All these variants are in dbSNP.
Several consecutive “SNPs” only found on reads ending on the left of the homopolymer.

7bp “T” homopolymer run.

Several consecutive “SNPs” only found on reads ending on the right of the homopolymer.
Local realignment uncovers the hidden indel in these reads and eliminates all the potential FP SNPs.
Calling somatic mutations

muTect

Default values assume some mutation rate

Cibulskis et al, Nat Biotech, 2013
Removing germline variants

- If a germline variant is missed (only) in the normal sample, it will be called as a somatic variant in the tumor sample.
- Even with high-coverage normal data, some variants will be missed.
- To avoid such false positive calls, compile a panel of normals and remove any variants that were seen in this set.
Sequencing matched normals?

- 815 tumor-normal paired samples, 15 tumor types
- WES at ~200X; panel (111 genes) at >1000X
- 140 somatic mutations per exome, 4.3 per panel
- **Tumor only: FPs comprising 65% for exomes and 31% for panels**
Sequencing matched normals?

- 669 Somatic candidate mutations from targeted analyses
  - 304 Germline
  - 148 Somatic
    - COSMIC: 32
    - 67 in actionable genes
  - 83 Germline
- Candidate somatic mutations after dbSNP filtering
- True somatic mutations
- Germline mutations

- 140,107 Somatic candidate mutations from whole-exome analyses
  - 101,924 Germline
  - 13,153 Somatic
    - COSMIC: 951
    - 77 in actionable genes
  - 24,869 Germline
- Candidate somatic mutations after dbSNP filtering
- True somatic mutations
- Germline mutations
How to remove those germline variants?

- Constructed a set of 931 samples from healthy, unrelated individuals from two sequencing platforms to serve as a virtual normal (VN).
- This removed (1) >96% of the germline variants also removed by the matched normal (MN) sample and (2) additional variants (2%–8%) not corrected for by the MN.
Using virtual normals

- MN+VN: up to ~30% improvement compared with MN
- The number of unrelated genomes needed to be as efficient as the MN:
  - ~200 for SVs, 400 for SNVs/indels
- Methods need to leverage information about the area surrounding variants.
Sensitivity and specificity

**muTect**

![Graphs showing sensitivity and specificity](image)

- **Graph a:** Sensitivity vs. False positive rate (Mb⁻¹)
  - Calculation (Q35)
  - MuTect STD
  - MuTect HC
  - $f = 0.2$

- **Graph b:** Sensitivity vs. Tumor sample sequencing depth
  - Calculation (Q35)
  - MuTect STD (virtual tumors)
  - MuTect HC (virtual tumors)
  - MuTect HC (downsampling)
  - MuTect HC + PON (downsampling)
  - $f = 0.1$
  - $f = 0.2$
  - $f = 0.4$
  - $f = 0.05$
Identifying somatic mutations

• Calling single-nucleotide mutations in a cancer genome is more difficult than expected

• Major sequencing centers have different call sets

• Structural variations are more difficult†

Analysis by Su Yeon Kim & Terry Speed
Somatic variant calling - TCGA Benchmark

31% of the variants are detected by all
detected by only one

Based on 16 exome pairs (lung)

Kim & Speed, *BMC Bioinformatics*, 2013
Comparison of variant callers

Why the discrepancies?

- Some algorithms are more conservative than others
- Different filtering criteria, e.g., strand bias
- More disagreements for variants with low allele frequency
- For these hard-to-detect variants, Sanger sequencing can give ambiguous results

- The Genome in a Bottle Consortium has provided high quality genotype calls for NA12878 using 14 datasets and 5 technologies (Zook et al, Nature Biotech, 2014)
Mutation Calling Challenge

ICGC-TCGA DREAM Mutation Calling Challenge

methods for identifying cancer rearrangements in whole data.

Human Tumour Data

Simulated Tumour Data

SVs
- Balanced accuracy across all 10 T/N pairs

SNVs
- Balanced accuracy across all 10 T/N pairs

Challenge 1A

Challenge 1B

SVs
- Tumour 1
- Tumour 2
- Tumour 3
- Tumour 4
- Tumour 5

SNVs
- Tumour 1
- Tumour 2
- Tumour 3
- Tumour 4
- Tumour 5

2A-1 to 2A-5

2B-1 to 2B-5
Validation of mutation calls

- Depending on the platform, variant caller, and threshold for statistical significance, mutation calls will be different.
- The three sequencing centers do not agree entirely—there is a working group among the sequencing centers to compare their calls.
- For major journals, additional validation will be needed.
- Sanger sequencing is one option, but there are often too many candidates/samples.
- A rapid, high-throughput, cost-effective validation pipeline will be important.
Mutation call comparisons in kidney cancer

- Across 10 samples
  - BI+BCM call 403 events
  - BCM uniquely calls +48 events (+12%)
  - BI uniquely calls +269 events (+67%)

<table>
<thead>
<tr>
<th>sample</th>
<th>bi-only</th>
<th>both</th>
<th>bcm-only</th>
</tr>
</thead>
<tbody>
<tr>
<td>BCM-KIRC-B2-3924</td>
<td>35</td>
<td>52</td>
<td>12</td>
</tr>
<tr>
<td>BCM-KIRC-B2-4098</td>
<td>32</td>
<td>33</td>
<td>9</td>
</tr>
<tr>
<td>BCM-KIRC-B2-4099</td>
<td>37</td>
<td>42</td>
<td>3</td>
</tr>
<tr>
<td>BCM-KIRC-B8-4620</td>
<td>18</td>
<td>31</td>
<td>9</td>
</tr>
<tr>
<td>KIRC-B2-3924-DN</td>
<td>22</td>
<td>56</td>
<td>3</td>
</tr>
<tr>
<td>KIRC-B2-4098-DN</td>
<td>24</td>
<td>43</td>
<td>0</td>
</tr>
<tr>
<td>KIRC-B2-4099-DN</td>
<td>26</td>
<td>45</td>
<td>0</td>
</tr>
<tr>
<td>KIRC-B2-4101-DN</td>
<td>27</td>
<td>25</td>
<td>4</td>
</tr>
<tr>
<td>KIRC-B8-4620</td>
<td>18</td>
<td>36</td>
<td>6</td>
</tr>
<tr>
<td>KIRC-B8-4622</td>
<td>30</td>
<td>40</td>
<td>2</td>
</tr>
</tbody>
</table>

Kristian Cibulskis (Broad)
An artifact in the GBM paper

• “We reported somatic mutations in the *ERBB2* gene in 7 of 91 cases analyzed by capillary DNA sequencing and validated by mass spectrometric genotyping.”

• “Further analysis of these cases has revealed that the mutations were only present in the whole genome amplified tumor DNA used for the study but not in the unamplified tumor DNA (see Greulich et al, PNAS, 2012 for more details).”
Sequencing depth?

- What are the variant allele frequencies (VAF) of important mutations?
- Given a platform, how many variants are we missing?
- How much more would we gain if we sequence deeper?
EGFR T790M - primarily associated with a mechanism of acquired resistance to EGFR kinase inhibitors erlotinib or gefitinib in lung tumors.
Variant allele frequencies at hotspots
Why are the allele frequencies so low?

- Tumor purity is often very low in practice (from biopsy?)
- Subclonality
- Acquired mutation

Data from Lee et al, Oncotarget, 2016

3832 lung cancers
From *in silico* dilution experiments...

“Limit of detection” (95% sensitivity):

- 20%: 49X
- 10%: 114X
- 5%: 315X
- 2%: 1039X
How much would you miss?
Finding driver mutations

- High rate of abnormalities, sometimes 10K mutations per cancer
  - Few are “driver” mutations
  - The rest are “passenger” mutations

- Is it possible to identify changes (“drivers”) in the tumor genomes that drive cancer progression?

- Are the drivers high frequency mutations?
- One could validate using RNA-seq data

- Exome vs whole-genome vs RNA-seq. All of them?
Estimating significance of mutations

- “MutSig” (Broad)

Overview of the algorithm

Mutations in each tumor

Mutation tally

Gene scores
Simplifying assumptions

• Background mutation rate is uniform
  - across sequence contexts (melanoma: C\rightarrow T, UV-induced)
  - across patients
  - across genes

• There could be \sim 10 fold difference in background mutation rates along the genome
  - For examples, highly expressed genes and early-replicating genes have lower mutation rates

• Need to model these aspects to derive more accurate estimates of significance
Background mutation rate?

Which mutations are important?

Davoli et al, *Cell*, 2013
LoF mutations: nonsense and frameshift mutations

Analysis across 8200 exomes revealed the large set of cancer-related genes (~300)

Davoli et al, *Cell*, 2013