Week 10, Lecture 20

István Albert

Biochemistry and Molecular Biology
and Bioinformatics Consulting Center

Penn State
“Holistic” Data Analysis

- Put together **EVERY STEP** of the analysis **BEFORE** optimizing any of the intermediate steps.

- Try to imagine what the end result needs to look like and work towards that goal.

- Think of an artist drawing portrait → it is a successive refinement of the entire image.
Origins of genetic variation 1

• A regular diploid human cell contains 46 chromosomes

• 23 pairs of **homologous** chromosomes = 46  
  (22 pairs + sex chromosomes XX(female) XY(male))

• One set of chromosomes inherited from each parent

Note that the reference genome is a “consensus” across all chromosomes of DNA pooled from multiple individuals
Origins of genetic variation 2

Meiosis → four genetically unique haploid gametes that each contain a unique mixture of the genetic code of the maternal and paternal chromosomes of the cell.

- Genetic diversity → phenotype → natural selection → adaptation → evolution
Origins of human genetic variation 3

• No two humans are genetically identical (not even monozygous twins that start out as such)

• About 30 new variations per generation.

• An allele is one of two or more forms of a gene or a genetic locus

• Both alleles are the same → homozygotes.
• If the alleles are different → heterozygotes.
**Single nucleotide polymorphisms: SNP**

- A single nucleotide — A, T, C or G — in the genome differs between members of a population or chromosome pairs.

- Originally defined as occurring at least in 1% of the population (these definitions may shift in time) → **SNV** (single nucleotide variant) if observed very rarely.

- SNP, SNV → may fall within coding sequences of genes, non-coding regions of genes, or in the intergenic regions.

- **DIP**: deletion/insertion polymorphism,

- **Single Nucleotide Polymorphism Database**[^1] (dbSNP)

- As of 26 June 2012, dbSNP listed 187,852,828 SNPs in humans.
SNP Calling

• Not nearly as well standardized as one might think

• The Archon Genomics X PRIZE will award $10 million to the first team to rapidly, accurately and economically sequence 100 whole human genomes to a level of accuracy never before achieved.
Organizers of the Archon Genomics X-Prize called off their $10-million competition Thursday, just two weeks before teams were set to begin work on 100 high-quality human genome sequences, in 30 days, at a cost of less than $10,000 per genome.
Updated comparison of variant detection methods: Ensemble, FreeBayes and minimal BAM preparation pipelines

### Variant evaluation

I previously discussed the comparison of reference calls produced from my *genome*, In this post, I will compare the results produced with the GATK and samtools pipelines. The comparisons used a tool that identifies concordant and discordant calls. If we use this automated validation to compare the two pipelines, we can achieve the following results:

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>concordant: total</td>
<td>160851</td>
</tr>
<tr>
<td>concordant: SNPs</td>
<td>136146</td>
</tr>
<tr>
<td>concordant: indels</td>
<td>24705</td>
</tr>
<tr>
<td>GATK discordant: total</td>
<td>13925</td>
</tr>
<tr>
<td>GATK discordant: SNPs</td>
<td>1315</td>
</tr>
<tr>
<td>GATK discordant: indels</td>
<td>12610</td>
</tr>
<tr>
<td>samtools discordant: total</td>
<td>25368</td>
</tr>
<tr>
<td>samtools discordant: SNPs</td>
<td>17247</td>
</tr>
<tr>
<td>samtools discordant: indels</td>
<td>8121</td>
</tr>
</tbody>
</table>

8% of GATK and 14% of samtools SNP calls are discordant!
SNP calling checklist

• Unique sample or pooled samples?
  – unique samples ➔ the expectation for each allele will be 50%

• External information ➔ SNPs tend to occur in clusters

• Coverage and quality filtering are very important
A large number of SNP callers have been published

• Each is good at some aspects (well publicized) – and not so good at others (less publicized)

• SNP calling is seems deceivingly simple – why can’t we just enumerate all the bases at a position?

• Greatest challenge: misalignments $\rightarrow$ incorrect SNP calls
Sequence Realignment/Quality Recalibration

• Correcting an alignment with respect of observed variations AND the other reads in the data.

• Rescaling quality measures based on the empirically observed sequencing errors.
10,000 exonic sites where the RNA does not match the DNA

All 12 possible categories of discordance have been observed
In total, we generated ~1.1 billion reads of 50 base pairs (bp) (~41 million reads and 2 Gb of

Next, we validated our findings experimentally by Sanger sequencing of both DNA and RNA

Proteomic evidence for RDD.

and gene density among chromosomes. RDD sites are significantly \( P < 10^{-10} \) enriched in genes
Comment on “Widespread RNA and DNA Sequence Differences in the Human Transcriptome”

Joseph K. Pickrell,1* Yoav Gilad,2 Jonathan K. Pritchard1,2

they attributed to previously unrecognized mechanisms of gene regulation. We found that at least 88% of these sequence mismatches can likely be explained by technical artifacts such as errors in mapping sequencing reads to a reference genome, sequencing errors, and genetic variation.

Comment on “Widespread RNA and DNA Sequence Differences in the Human Transcriptome”

Wei Lin,1* Robert Piskol,2* Meng How Tan,2 Jin Billy Li2†

12 possible mismatch types. Before accepting such a fundamental claim, a deeper analysis of the sequencing data is required to discern true differences between RNA and DNA from potential artifacts.
Questioning the evidence for non-canonical RNA editing in humans

15/03/2012
Categories: Journal Club
Written by Joe Pickrell

In May of last year, Li and colleagues reported that they had observed over 10,000 sequence mismatches between messenger RNA (mRNA) and DNA from the same individuals (RDD sites, for RNA–DNA differences) [1]. This week, Science has published three technical comments on this article (one that I wrote with Yoav Gilad and Jonathan Pritchard; one by Wei Lin, Robert Piskol, Meng How Tan, and Billy Li; and one by Claudia Kleinman and Jacek Majewski). We conclude that at least ~90% of the Li et al. RDD sites are technical artifacts [2,3,4]. A copy of the comment I was involved in is available here, and Li et al. have responded to these critiques [5].
Genome Analysis Toolkit

The Genome Analysis Toolkit or GATK is a software package developed at the Broad Institute to analyse next-generation resequencing data. The toolkit offers a wide variety of tools, with a primary focus on variant discovery and genotyping as well as strong emphasis on data quality assurance. Its robust architecture, powerful processing engine and high-performance computing features make it capable of taking on projects of any size.

Learn more »

About
Overview of the GATK and the people behind it

Guide
Detailed documentation, guidelines and tutorials

Community
Forum for questions and announcements

Events
Materials from live and online events

Latest version: 2.7-4
Release Notes
Download now
For-profit users: click here
GATK workflow

GATK is like a life-style
Homework 20

1. Using the default parameters for `wgsim` generate 1 million paired end reads from the yeast genome

2. Create an BAM alignment file from the paired end files

3. Call SNPS using `samtools` or any other variation caller and save them into a VCF file

4. Provide a screenshot that shows a region in IGV with both the alignment file and the.

5. Estimate how many of your original mutations generated with `wgsim` can be recovered from the VCF file.