Compression and Integration of Genomic Variants Into Smart EHR Systems

Andrew Gritsevskiy and Adithya Vellal
Mentor: Dr. Gil Alterovitz

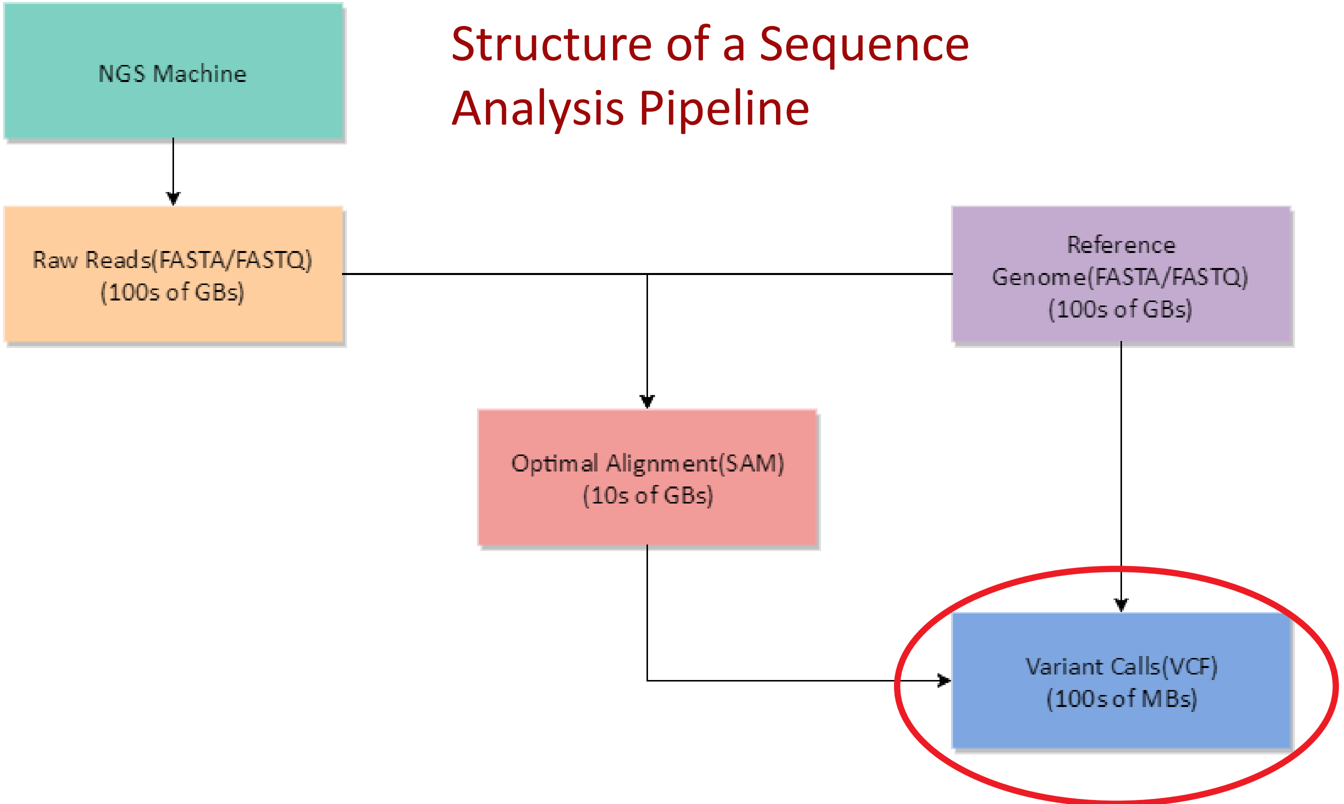
6th Annual PRIMES Conference
May 22 2016

An Introduction to Genomic Data

- Next Generation Sequencing (NGS) machines allow for simple, cheap human genomic data
- Human genomic **variants** are the key to precision medicine and personalized drug development
- However, genomic data is:
    - **Very large** (raw output from NGS machine ∼200 GB)
    - **Expensive** to store and maintain
    - **Computationally intensive** to process

Genomic Sequence Analysis Pipelines

- NGS machines do not actually output a total genomic sequence
- Three step process required to finally obtain variant data for analysis



Structure of a Sequence Analysis Pipeline

NGS Machine

Raw Reads(FASTA/FASTQ) (100s of GBs)

Reference Genome(FASTA/FASTQ) (100s of GBs)

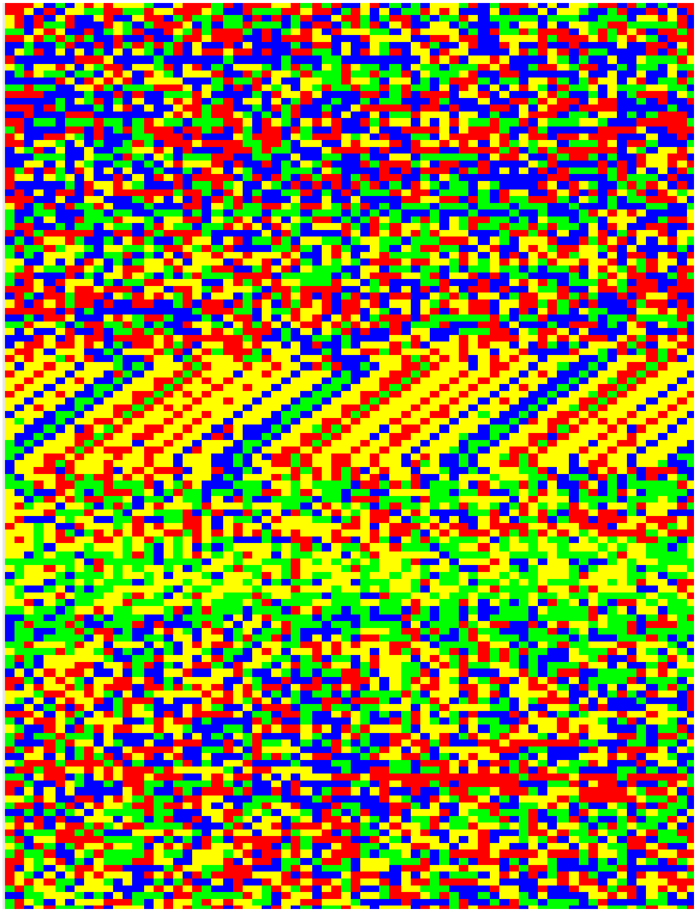
Optimal Alignment(SAM) (10s of GBs)

Variant Calls(VCF) (100s of MBs)

Compression of Genomic Data

- Compression will make it much more accessible
- Intrinsic biological patterns provide a unique opportunity for compression
- Understanding these features will enable improvements in precision medicine and genomic analysis

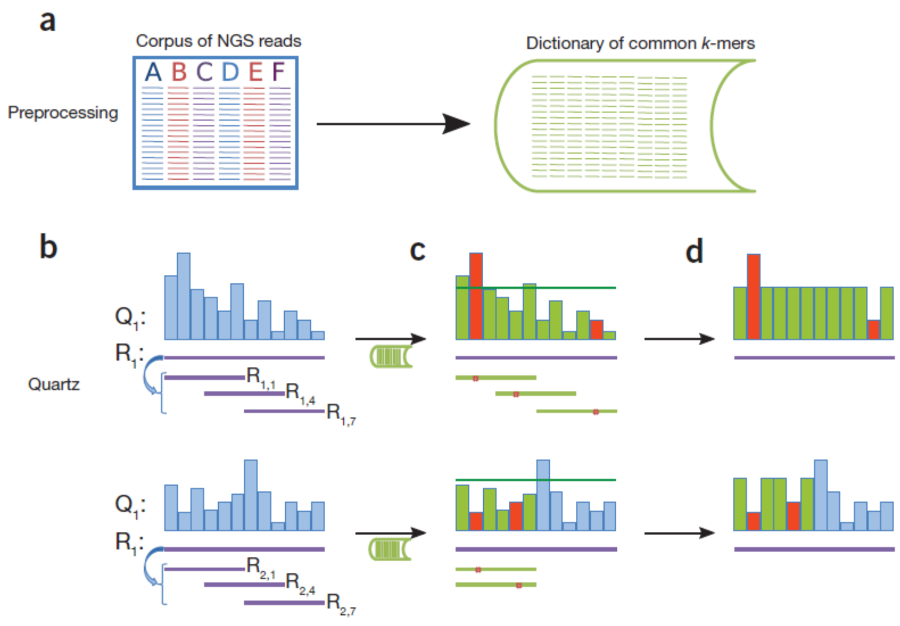
Pictorial Representation of Genomic Data



Ongoing Research Into Genomic Compression

- Quality Score Reduction at Terabyte Scale (QUARTZ)
    - Compression of raw data through standardized quality scores
    - *Lossy* compression
- Compressive Read Mapping Accelerator (CORA)
    - Uses redundancy of NGS output reads to speed up read mapping
- Both methods improve only **raw data**

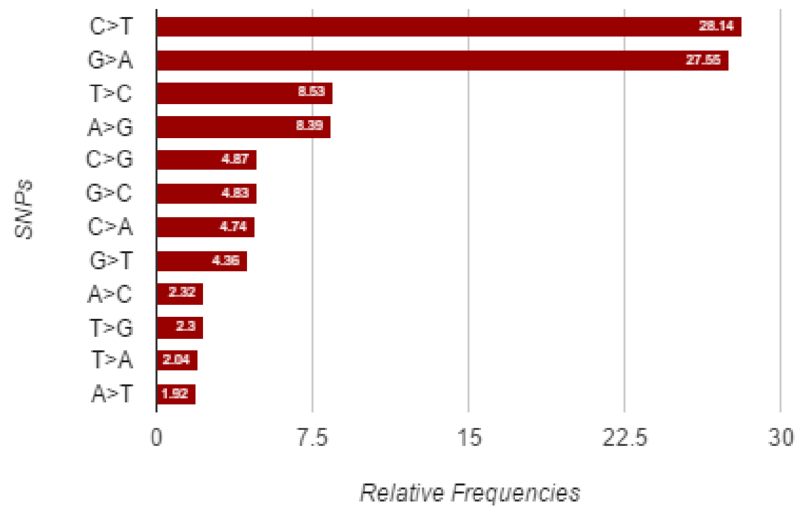
Process of Standardizing Quality Scores Using QUARTZ



Analysis of Variant Data

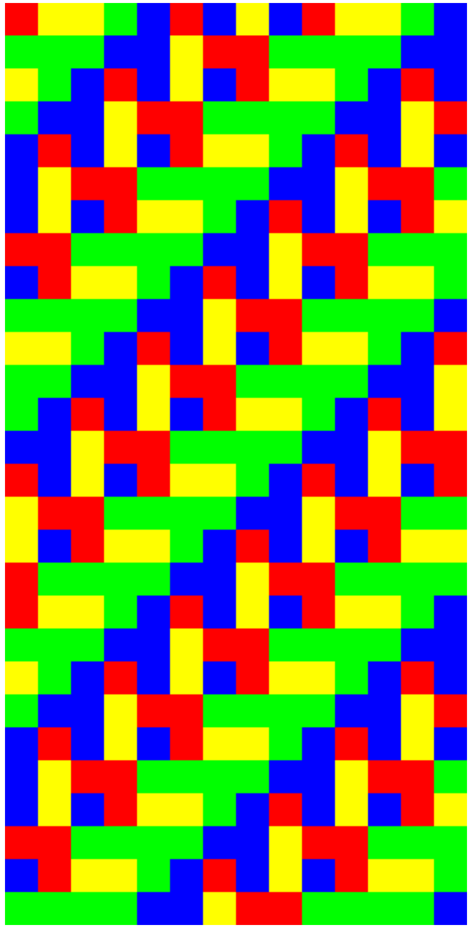
- No current method that exploits intrinsic patterns to compress variant data
- Focus on intrinsic patterns found in **Single Nucleotide Polymorphisms (SNPs)**
- Found that one direction occurred significantly more than the other in transitions **and** transversions

Relative Frequencies of SNPs



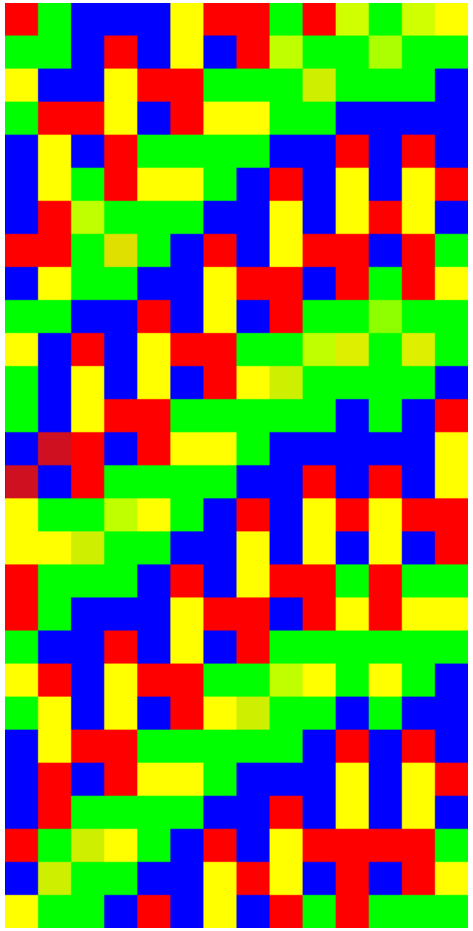
Ones Algorithm Analysis

- Used for analyzing intrinsic **SNP** patterns
- Gives an optimal set of rules describing any pattern
- Only works for **noiseless** data



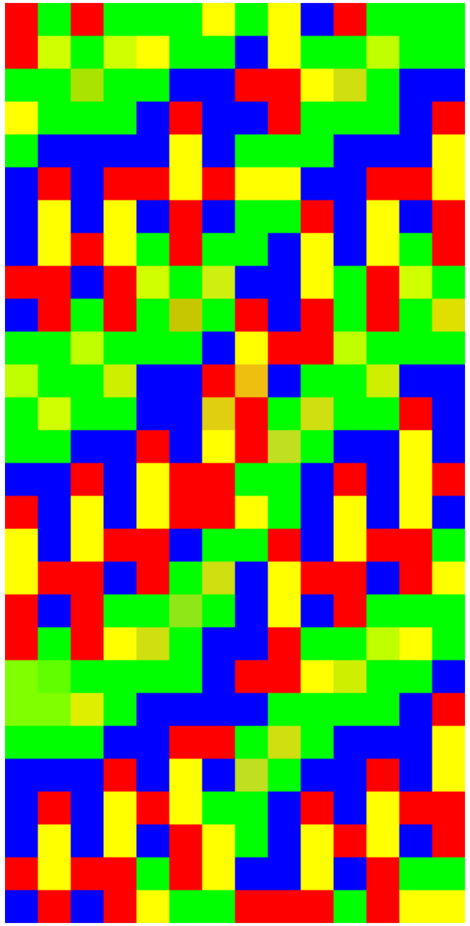
Ones Algorithm Analysis

- Used for analyzing intrinsic **SNP** patterns
- Gives an optimal set of rules describing any pattern
- Only works for **noiseless** data



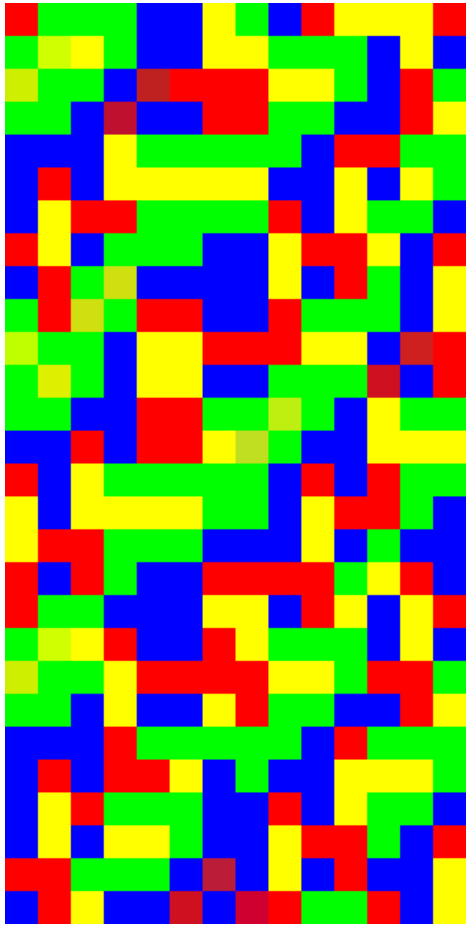
Ones Algorithm Analysis

- Used for analyzing intrinsic **SNP** patterns
- Gives an optimal set of rules describing any pattern
- Only works for **noiseless** data



Ones Algorithm Analysis

- Used for analyzing intrinsic **SNP** patterns
- Gives an optimal set of rules describing any pattern
- Only works for **noiseless** data



Initial Compression of Variant Calls

- Considering pairs of consecutive SNPs
    - 144 Possible Pairs
    - Gaps (# of bp) between the 2 SNPs
- Only encode pairs where SNPs are within 40 bp of each other
    - Frequency Based Encoding
    - *Lossless*
    - ∼80% of all pairs
- Compressed files **28%** of original file size(∼3.5 x compression)

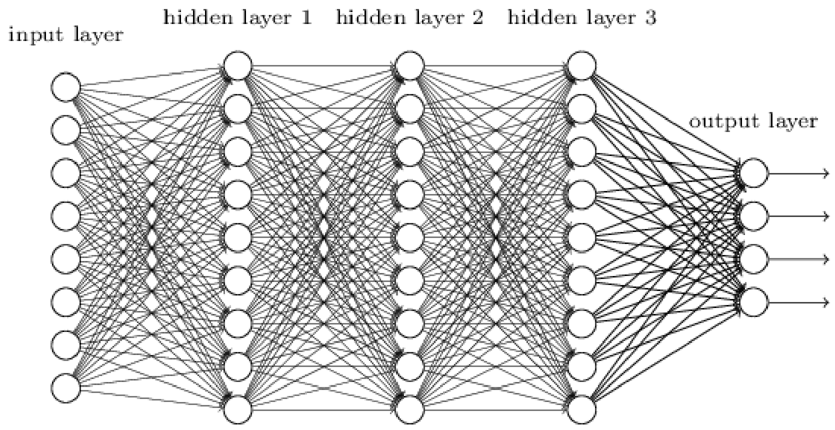
Improvements to Current Compression

- Look at groups of 3 consecutive SNPs and corresponding gaps
- Integrate Ones Algorithm results to represent SNP patterns in a simple manner
- Make **modifications to the reference genome** using analysis of a noise-considering Ones Algorithm
    - Ex: Approximately every 5th SNP is T>A

Applying Deep Learning to Variant Call Compression

- Initial analysis demonstrates huge potential for compression in variant call files
- Deep learning can be applied to learn intrinsic biological patterns
- Unsupervised learning(no labels)
- Opportunity to develop a **better understanding of biological variants**

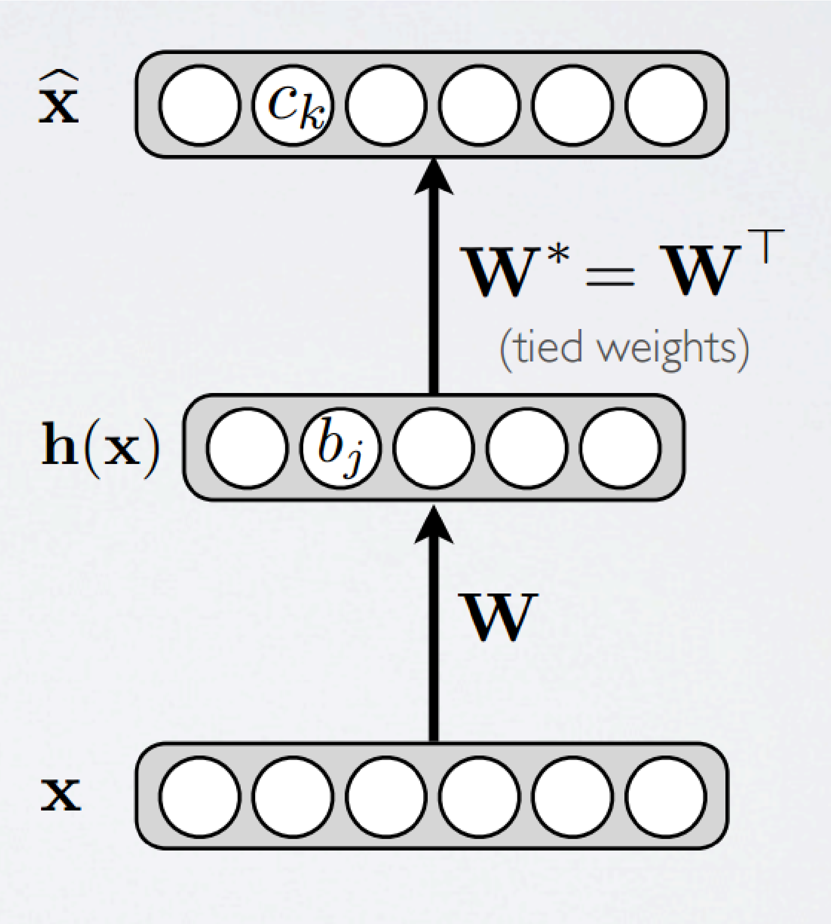
Structure of an Artificial Neural Network



Autoencoders for Variant Compression

- Goal: Output = Input
- Simple network with three layers
  - Input Layer
  - *Undercomplete* Hidden Layer
  - Output Layer
- *Lossy* compression
  - Can be made *lossless* with enough features/training iterations
  - Can adjust for determined *loss* to ensure *losslessness*

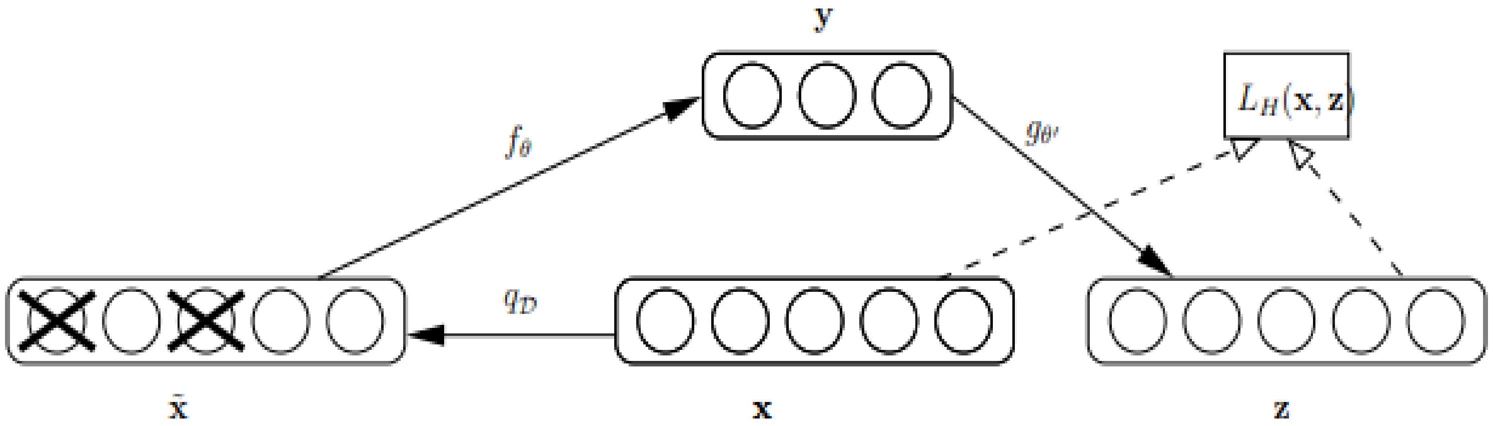
Autoencoder with Undercomplete Hidden Layer



Denoising Autoencoder

- Original input corrupted to ensure robust feature learning
- Does not necessarily need an *undercomplete* hidden layer
  - However will most likely result in best compression
- Will learn most important features
  - Forced to compress data in two different ways simultaneously

Basic Structure of a Denoising Autencoder



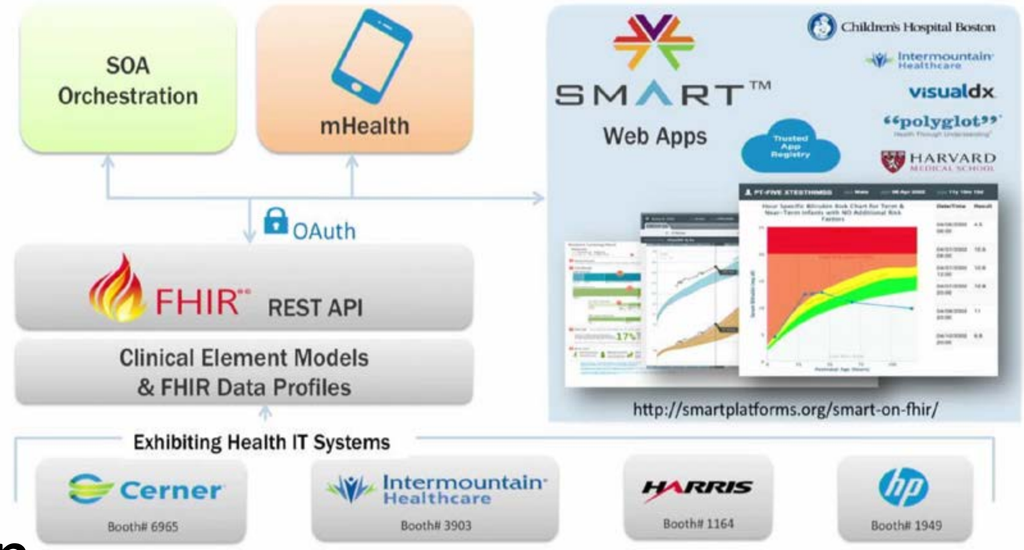
Electronic Health Record (EHR) Systems

- Increasing in popularity as paper medical records become obsolete
- Enable patient access to all medical data
- Development of personalized apps which utilize this data
- Opportunity to integrate genomic data with rest of medical information
    - Simplifies development of precision medicine

SMART on FHIR

- Fast Health Interoperability Resource
- Built on Health Level 7 (HL7) International Standards
- Defined set of resources for various patient data
  - Allows simple creation of apps
- Each resource is defined using a standard format such as **json**
- Key to making EHRs easily accessible in a standard format

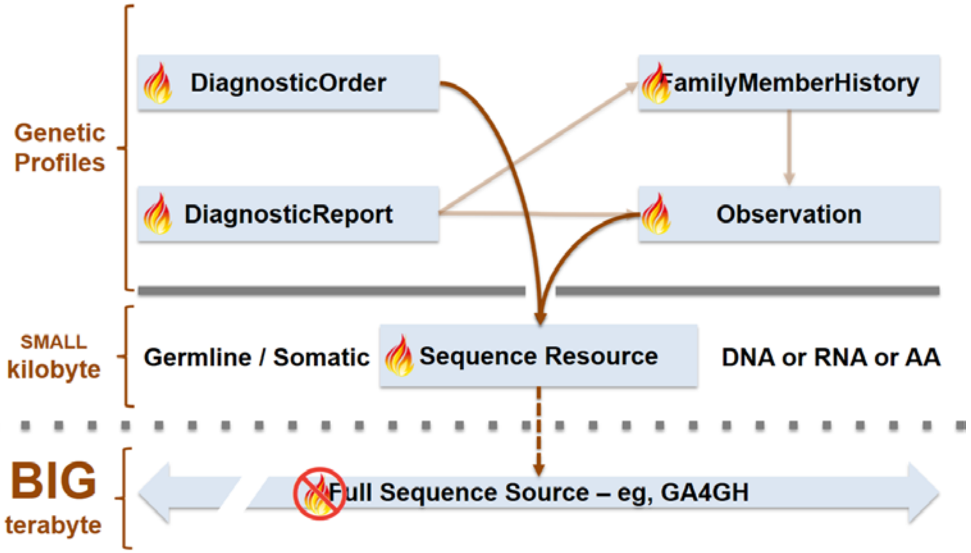
Overall Structure of SMART on FHIR



FHIR Genomics Sequence Resource

- Specific set of resources built to handle genomic information
- Focus on patient genomic **variant data**
- Looks at only small *windows* of genomes which contain useful variant data are stored
- Provides link to GA4GH repository to easily access full sequence data

Role of Sequence Resource



Examples of Variant Information in the Sequence Resource

SNP

```
{
  "species": {"text": "Homo
sapiens"},
  "id": "t10116",
  "type": "DNA",
  "variation": {
      "start": 86552205,
      "end": 86552206,
      "observedAllele": "G",
      "referenceAllele": "A"
  },
  "resourceType": "Sequence",
  "referenceSeq": {
      "genomebuild": "37",
      "windowStart":"86552200”,
      "chromosome": 22,
      "windowEnd": "86552210",
      "referenceSeqId": "GRCh"
  }
}
```

INSERTION

```
{
  "species": {"text": "Homo
sapiens"},
  "id": "t175",
  "type": "DNA",
  "variation": {
      "start": 712040,
      "end": 712047,
      "observedAllele": "CAGCTGT",
      "referenceAllele": "C"
  },
  "resourceType": "Sequence",
  "referenceSeq": {
      "genomebuild": "37",
      "windowStart": "712040",
      "chromosome": 22,
      "windowEnd": "712050",
      "referenceSeqId": "GRCh"
  }
}
```

Future Work

- Full implementation of the deep learning algorithm
  - Autoencoder with *Undercomplete* Hidden Layer
  - Denoising Autoencoder
- Analysis on which features of genomic variant data allow for compression
- Complete encoding and *lossless* decoding of VCF files using compression determined by deep learning
- Full integration of compressed files into FHIR Sequence Resource for use with EHRs

Conclusions

- Variant calls, the most important genomic data to medical and biological institutions, are expensive to store, maintain and process due to their size
- Initial analysis has proven that extensive compression is possible in this data due to intrinsic biological patterns and dependencies
- Deep learning provides a method to achieve far better compression while also learning new biology about genomic variants
- Integration into smart EHR systems such as FHIR will allow simple doctor and patient access to this data in the future

Acknowledgements

Thank you to:

- **MIT PRIMES** for providing this excellent and challenging research opportunity
- **Dr. Gil Alterovitz** for all his guidance and support
- **Other members of our lab** for providing assistance
- **Our parents** for their support

More Cool Pictures of Genomic Data