MPG NGS workshop:

Discovery, genotyping, and analysis of SNPs, Indels, and CNVs

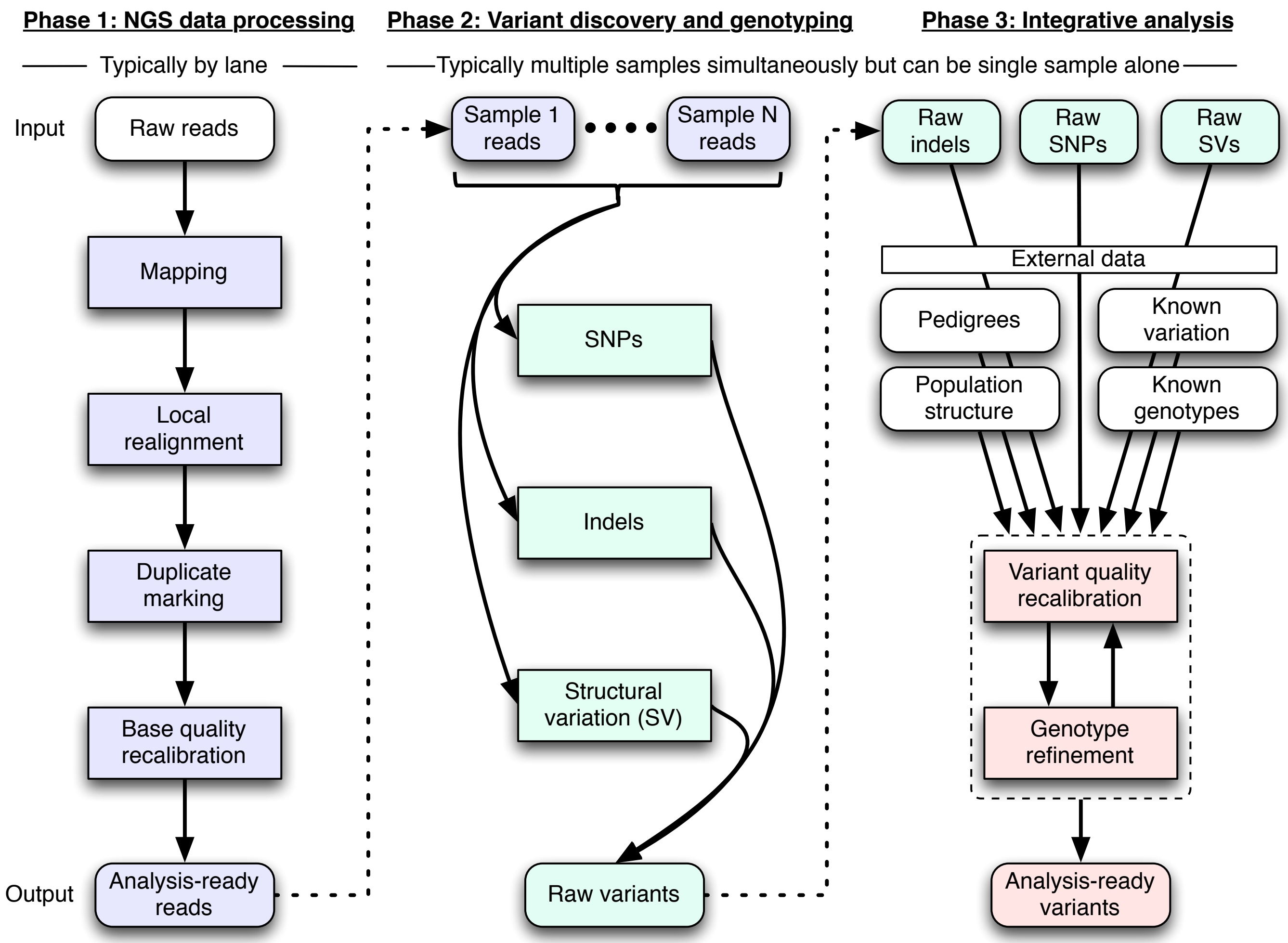
February 2011

Mark DePristo
Group Leader, Genome Sequencing and Analysis
Manager of Medical and Population Genetics Analysis

What will you learn today?

- A brief overview of the NGS workflow today
- Thoughts on future challenges for this community
- “Tutorial” of methods to discover and genotype SNPs, Indels, and CNVs from NGS data
  - From BAMs to VCFs
- Tools to analyze NGS variant calls and genotypes for association with disease
  - VCFs to insights

The paradigm today



Phase 1: NGS data processing

Typically by lane

Input

Raw reads

Mapping

Local realignment

Duplicate marking

Base quality recalibration

Output

Analysis-ready reads

Phase 2: Variant discovery and genotyping

Typically multiple samples simultaneously but can be single sample alone

Sample 1 reads

Sample N reads

SNPs

Indels

Structural variation (SV)

Raw variants

Phase 3: Integrative analysis

Raw indels

Raw SNPs

Raw SVs

External data

Pedigrees

Known variation

Population structure

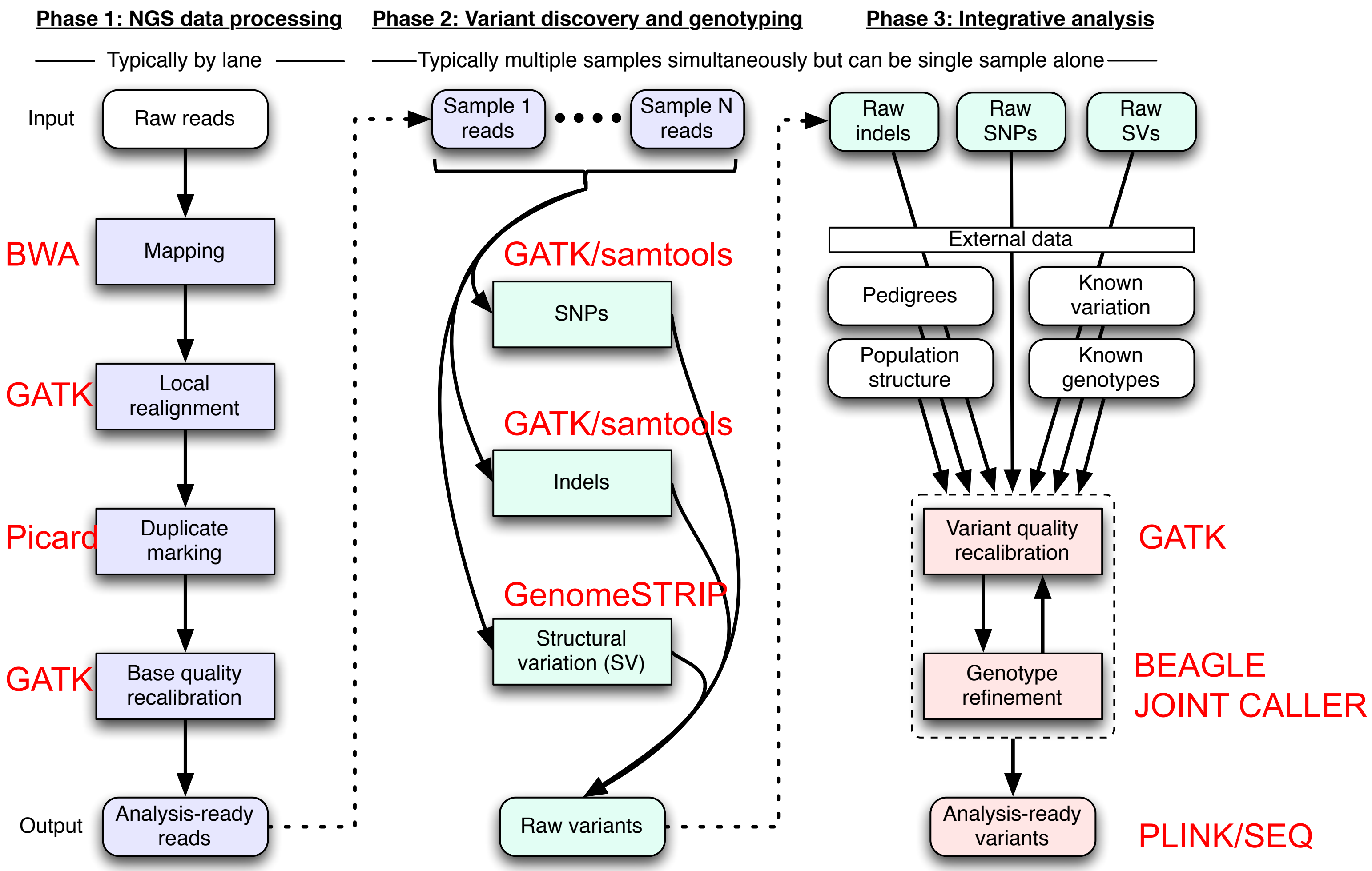
Known genotypes

Variant quality recalibration

Genotype refinement

Analysis-ready variants

The paradigm today



Phase 1: NGS data processing

Typically by lane

Input

Raw reads

BWA Mapping

GATK Local realignment

Picard Duplicate marking

GATK Base quality recalibration

Output

Analysis-ready reads

Phase 2: Variant discovery and genotyping

Typically multiple samples simultaneously but can be single sample alone

Sample 1 reads • • • • Sample N reads

GATK/samtools

SNPs

GATK/samtools

Indels

GenomeSTRIP

Structural variation (SV)

Raw variants

Phase 3: Integrative analysis

Raw indels

Raw SNPs

Raw SVs

External data

Pedigrees

Known variation

Population structure

Known genotypes

Variant quality recalibration

GATK

Genotype refinement

BEAGLE

JOINT CALLER

Analysis-ready variants

PLINK/SEQ

Capabilities at BI: today and tomorrow

		Capabilities as of today	Tomorrow (< 6 months)	Under active development?
SNPs				
	Exomes	◉	●	As a side-effect of WGS
	Low-pass WGS	◉	◉	Yes
	High-pass WGS	◉	◉	As side-effect to above
Indels				
	Exomes	◎	◉	Yes
	Low-pass WGS	○	◎	Yes
	High-pass WGS	◎	◉	Yes
CNVs				
	Exomes	◌	○	Many thinking about it
	Low-pass WGS	◎	◎	Yes
	High-pass WGS	◌	○	Maybe comes with above?

◌	No capabilities
○	Prototyped
◎	Analyzable
◉	Excellent
●	Complete

Some thoughts on the future

- Data production and processing challenges
  - New sequencing technologies
  - Real-time data generation and processing
  - How do we get to a perfect genome?
- Analytic challenges
  - Refocus on allele discovery?
  - Soft-called, genotype-free methods?
  - Hypothesis testing directly from the data?

New sequencing technologies look very promising

	PacBio amplicon sequencing data sets					
	phasing		200bp		2kbp	
	Hiseq	PacBio	Hiseq	PacBio	Hiseq	PacBio
SNP calls	547	531	119	65	591	543
Calls at HapMap	40	38	18	11	126	115
Ti/Tv ratio	1.99	2.03	1.43	1.71	2.23	2.31
Ti/Tv known	3.00	3.22	2.60	4.50	4.25	4.23
Ti/Tv novel	1.93	1.96	1.29	1.45	1.92	2.01

Mauricio Carneiro, Menachem Fromer, Pat Cahill, Chris Hartl, Carsten Russ, Niall Lennon

Real-time data generation

- How fast can I go from samples to sequence?
  - Today: ~3 months from idea to data at BI
- Could we do this in one business day?
  - Select regions?
  - Upfront sample prep?
  - Incremental data processing? Can we make a fast path?
  - MiSeq can do this already
- When sequencing itself is free, perhaps we can only process select parts of the data?
  - "Hybrid capture" without the capture
- Amortized deep sequencing
  - Data freezes in depths: 10x, 20x, 30x, controllable by sample

How do we get to a perfect genome?

- Practical goal:
  - 100% confidence in “good” areas of the genome
  - Lower confidence in the “bad” areas
  - And the wisdom to tell the difference

- Theoretical challenge:
  - Explain all reads under hypothesized genome sequence to the aggregate machine error rate
  - “No read left behind”

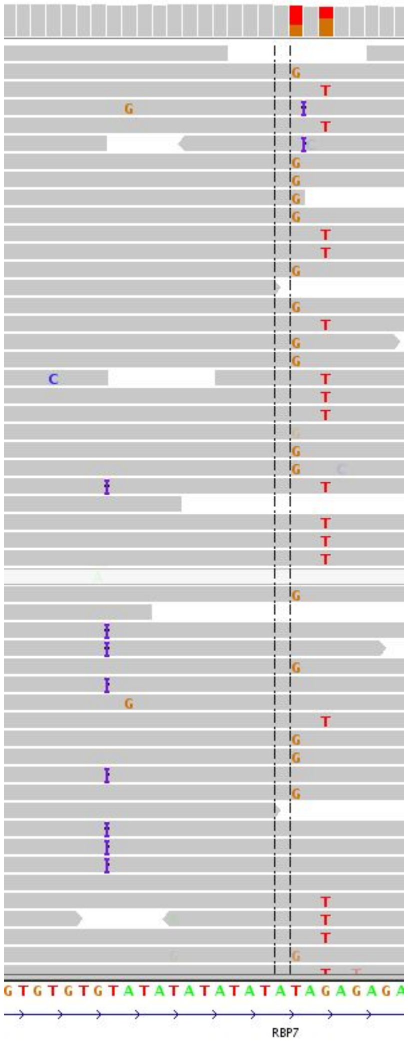
From reads to alleles: the first frontier

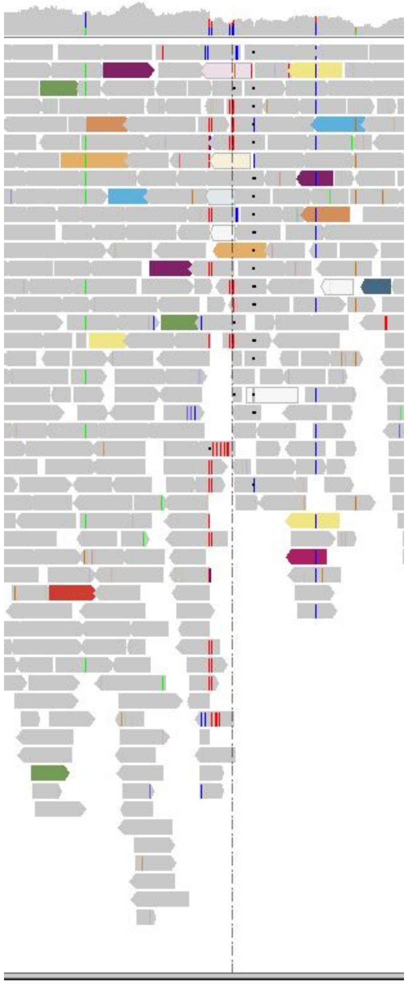
- Can’t calculate a likelihood for a hypothesis you don’t consider

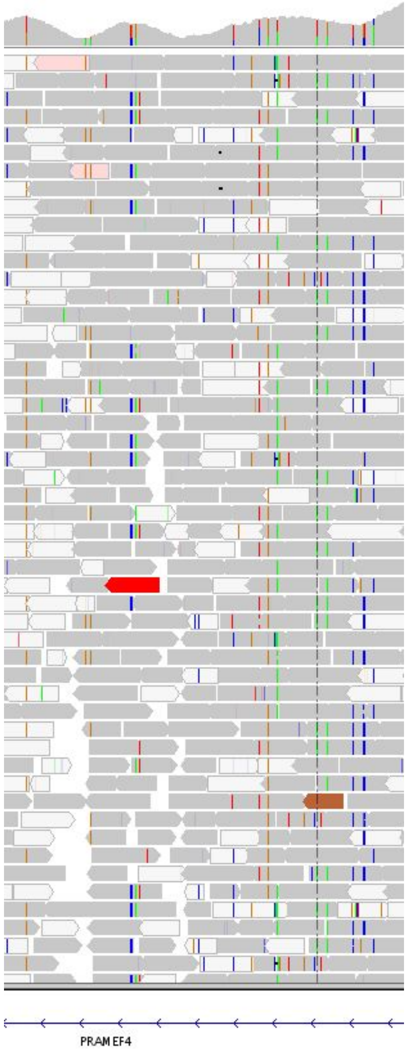
- How do I know what genetic variant I’m looking at, given the read data along?
  - A SNP, an INDEL, an SV, or something else?
  - Reference-free approaches?

- We’ve skipped over this problem by focusing on SNP calling
  - Enumerate 10 diploid genotypes (AA, AC, …, TT)
  - Not possible for indels or CNVs

Too systematic to be machine errors



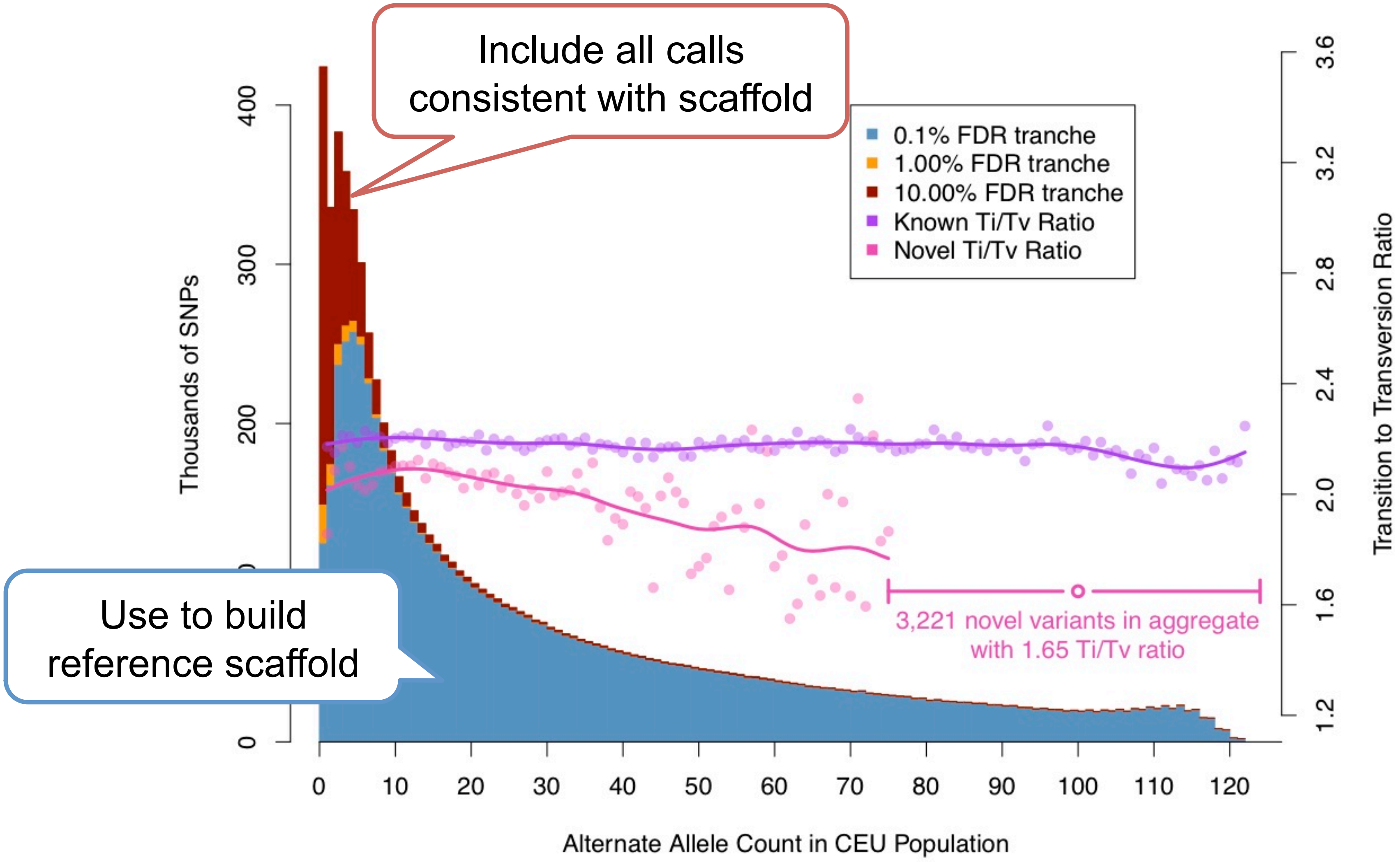




Soft-call, genotype-free methods

- We have uncertainty from all directions
  - Is this an indel here, or a SNP?
  - Is this a real generic difference, a systematic machine error, or a data processing artifact?
  - What's are the likelihoods of alleles A and B, whatever they are, in all my samples?
- Integrating in all sources of uncertainty may:
  - Help us avoid missing variants that are a bit odd but really interesting for my disease
  - Avoid over-interpreting the significance of uncertain events

Uncertainty from low-pass sequencing



Include all calls consistent with scaffold

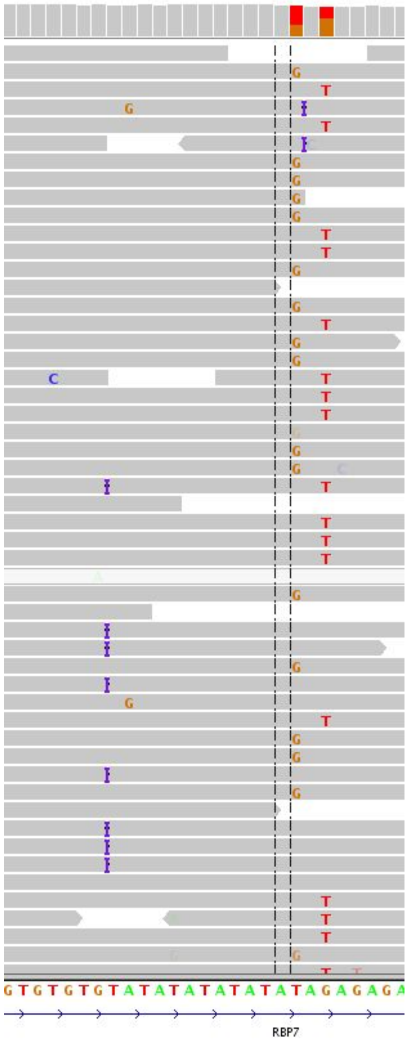
Use to build reference scaffold

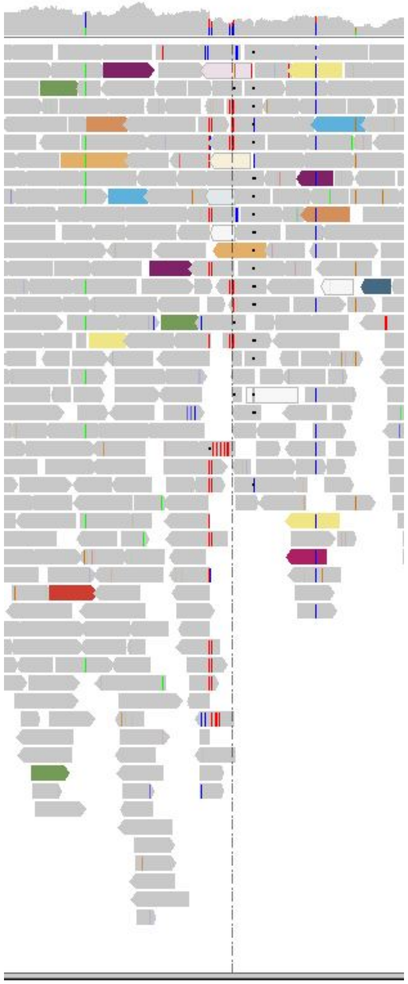
Ryan Poplin, Eric Banks

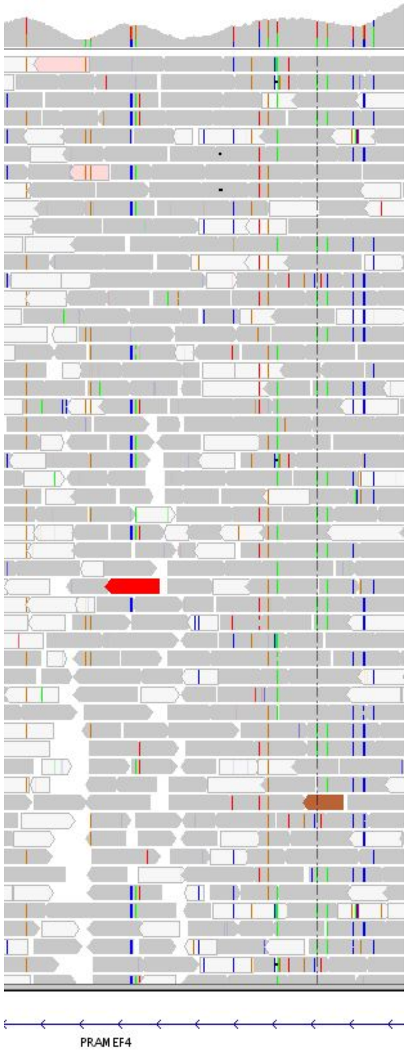
Hypothesis testing directly from the data

- Current paradigm to disease associated loci:
  - Call variants and genotypes from NGS data
  - Test association of sites with genotypes
- Works well when raw data bottoms out in genotypes, as with chips
- We can go one step further and directly test the association hypothesis in the reads themselves
  - Calling-free approach to look for features of the data that segregate with disease status
  - See next slide for examples

We could directly test these for association without known what they really are







Thank you all for attending

- I hope you find this workshop useful in your day to day work with NGS data

- All of the slides and video will be available online soon

- Thank in advance the upcoming speakers for their hard work in summarizing their work for us today