SCRIPPS GENOMIC MEDICINE
A Collaboration of Scripps Health and The Scripps Research Institute

Scripps Translational Science Institute

Accurate genotyping of INDELS from population-scale short read sequence data

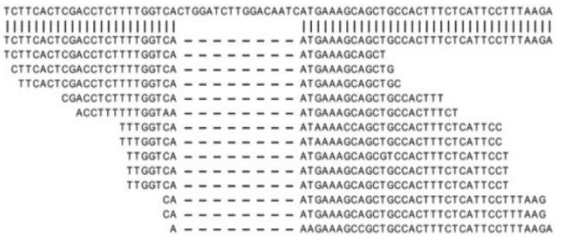
Vikas Bansal, Ph.D.

Scripps Genomic Medicine, Scripps Translational Science Institute

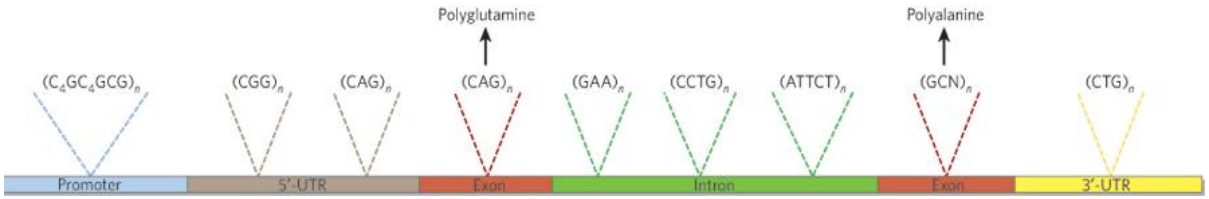
July 12, 2010

INDEL polymorphisms in the human genome

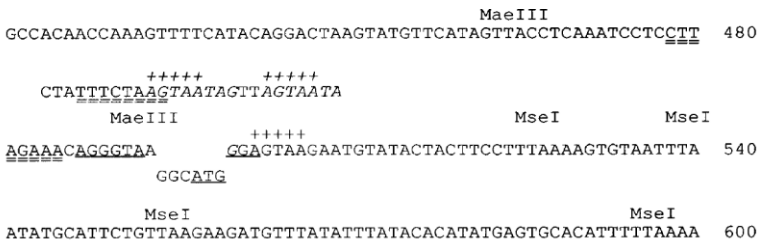
- Short insertion/deletions (1-100 base pairs)



- Micro-satellites or SSRs



- Complex insertion-deletions



- Long insertion/deletions (> 100 bp), copy number variants

(figures from Li et al. Gen. Res. 2008; Mirkin, Nat. 2007, Fernie & Hobart, Hum Gen 97)

What do we know about indels ?

- Short indels are the second most frequent form of variation in the human genome after SNPs
  - approximately 1 indel for every 10 SNPs*
- Indels can be used as markers in association studies (Bhangale et al. 2005)
- Indels are more likely to be functional especially in coding regions (no synonymous indels)

Number of entries in HGMD by type

Data type	Number of entries (public release for academic/non-profits only)	Nur
Mutation data	TOTAL (public release) **72414**	
Missense/nonsense	41295	
Splicing	7011	
Regulatory	1051	
Small deletions	11745	
Small insertions	4664	
Small indels	1065	
Repeat variations	178	
Gross insertions/duplications	744	
Complex rearrangements	555	
Gross deletions	4106	
Gene/sequence data		
Genes	2689	
cDNA reference sequences		

* estimate based on data from various genome sequencing projects

Feasible to identify indels from whole-genome sequencing using short reads

- ~ 400,000 indels (1-16 bp) identified in an African individual using 36 bp Illumina reads (Bentley et al. 2008)
- 135,262 indels in the genome of an Asian individual (Wang et al. 2008)
- ~ 230,000 indels detected using ABI SOLiD sequencing of the genome of an African individual (McKernan et al. 2009)

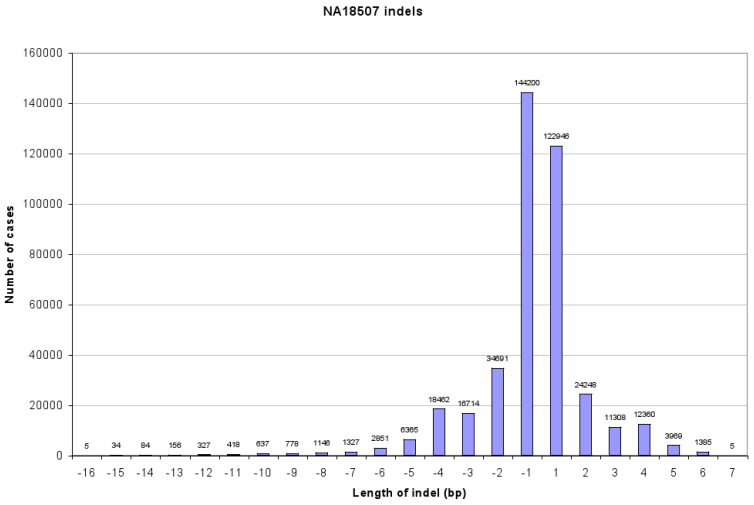


Figure S16. Analysis of short indel calls in human genome data for N18507. ***a.*** total number of calls and fraction that match previous entries in dbSNP. ***b.*** Distribution of size in the 404,416 indels. + and – values on the x axis correspond to presence or absence of bases in NA18507 relative to the reference sequence.

In contrast to SNPs, number of indels called in an individual genome dependent on read-length, sequencing platform and indel detection tool.

High false positive rate for indels...

Nature Jan 14, 2010

A comprehensive catalogue of somatic mutations from a human cancer genome

Erin D. Pleasance[1]*, R. Keira Cheetham[2]*, Philip J. Stephens[1], David J. McBride[1], Sean J. Humphray[2], Chris D. Greenman[1], Ignacio Varela[1], Meng-Lay Lin[1], Gonzalo R. Ordóñez[1], Graham R. Bignell[1], Kai Ye[3], Julie Alipaz[4], Markus J. Bauer[2], David Beare[1], Adam Butler[1], Richard J. Carter[2], Lina Chen[1], Anthony J. Cox[2], Sarah Edkins[1], Paula I. Kokko-Gonzales[2], Niall A. Gormley[2], Russell J. Grocock[2], Christian D. Haudenschild[5], Matthew M. Hims[2], Terena James[2], Mingming Jia[1], Zoya Kingsbury[2], Catherine Leroy[1], John Marshall[1], Andrew Menzies[1], Laura J. Mudie[1], Zemin Ning[1], Tom Royce[4], Ole B. Schulz-Trieglaff[2], Anastassia Spiridou[2], Lucy A. Stebbings[1], Lukasz Szajkowski[2], Jon Teague[1], David Williamson[5], Lynda Chin[6], Mark T. Ross[2], Peter J. Campbell[1], David R. Bentley[2], P. Andrew Futreal[1] & Michael R. Stratton[1,7]

"A total of 680 small deletions and 303 small insertions were predicted, of which 182 were evaluated and 66 (36%) confirmed. Thus the false-positive rate for insertions and deletions was higher than for substitutions."

Methods for detecting indels from short reads

- *Gapped alignment of reads (BWA, MAQ, Shrimp, Soap...)*
  - Length of indels limited by length of reads
- *Mapping of paired-end reads to a reference genome (Modil, BreakDancer)*
- *Split-read alignments (PINDEL)*
  - Can identify long deletions from paired-end reads
- *De novo assembly of short reads (Velvet, EULER, Abyss, etc)*
  - Can identify long insertions
  - Computationally challenging for large genomes, e.g. human

Indel detection and genotyping

- Several alignment tools (BWA, MAQ, ...) can identify indels
- SAMtools can call genotypes for indels and corrects for misaligned reads
- Existing methods do not leverage sequence data from a population of individuals to improve accuracy of indel detection and genotyping
- No easy way to incorporate additional indels identified by other approaches, e.g de novo assembly, split-read mapping

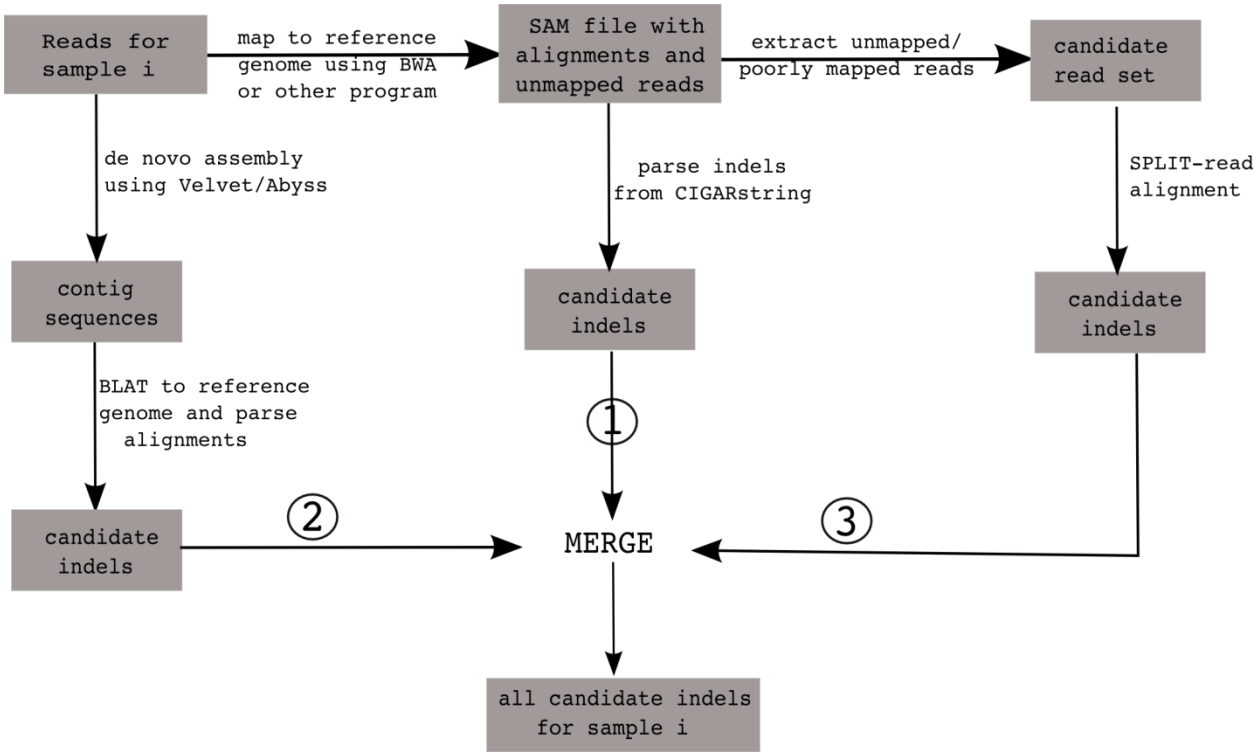
➔ A pipeline/tool for accurate detection and genotyping of INDELS from population sequence data would enable:

- Use of indels in sequencing-based association studies
- Use of indels as phylogenetic markers

A pipeline for comprehensive detection and genotyping of Indels from population-scale short read sequence data

- Use gapped alignment, split-read mapping and de novo assembly to enable comprehensive detection of indels
- Re-alignment of reads to indel consensus sequences to modify alignments of misaligned/unmapped reads and accurately determine allele counts for each indel
- Probabilistic method (MCMC) used to assign genotypes for each indel leveraging allele counts across all individuals

Step 1: identify candidate indels in each sample



- Gapped alignment can find short indels (1-10 bp)
- SPLIT-read alignment can detect long deletions and medium-sized insertions
- De novo assembly can detect long insertions and deletions

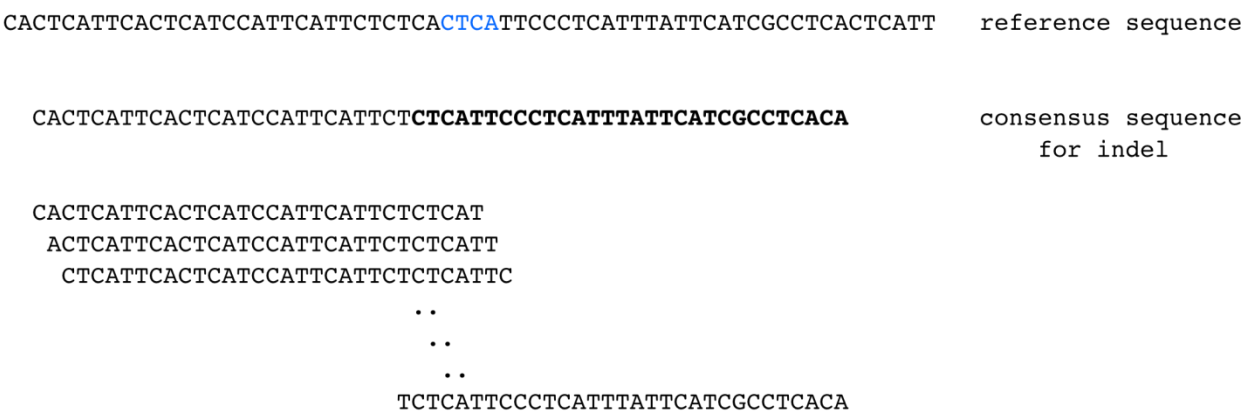
Step 2: merge candidate indels across all samples

- Common indels identified in multiple samples while rare ones in a few samples
- For split-read mapping, combined evidence from all samples to identify long deletions
- Apparently different indel calls can correspond to the same insertion/deletion event
- For each indel, we determine the 'leftmost' start position in the genome
- Indels with identical leftmost start position and set of bases are merged

```
CACTCATTCACTCATCCATTCATTCTCTCACTCATTCCCTCATTTATTCATCGCCTCACTCATT ref.sequence

CACTCATTCACTCATCCATTCATTCTCTCA----TTCCCTCATTTATTCATCGCCTCACTCATT  chr3 4654043 -CTCA

CACTCATTCACTCATCCATTCATTCT----CTCATTCCCTCATTTATTCATCGCCTCACTCATT  chr3 4654039 -CTCA

CACTCATTCACTCATCCATTCATTCTCT----CATTCCCTCATTTATTCATCGCCTCACTCATT  chr3 4654041 -CACT

CACTCATTCACTCATCCATTCATTCTCTC----ATTCCCTCATTTATTCATCGCCTCACTCATT  chr3 4654041 -ACTC
```

Step 3: Create consensus sequences for each indel

- For each indel, create a consensus sequence/super-read that contains all reads (of a given length) that support the alternate allele



- *Indel-sensitive reference sequence* = set of super-reads for each indel

Step 4: re-align reads to indel-sensitive reference sequence and modify SAM file

- All reads aligned (without gaps) to the new reference sequence
- Reads for which the new alignment is better than original alignment (SAM file) identified
  - Fixes misaligned reads that contain indel close to the 'ends' of the read
  - Aligns unmapped reads that correspond to long indels identified via denovo assembly/ split read analysis

Step 5: determine allele counts for each indel across population

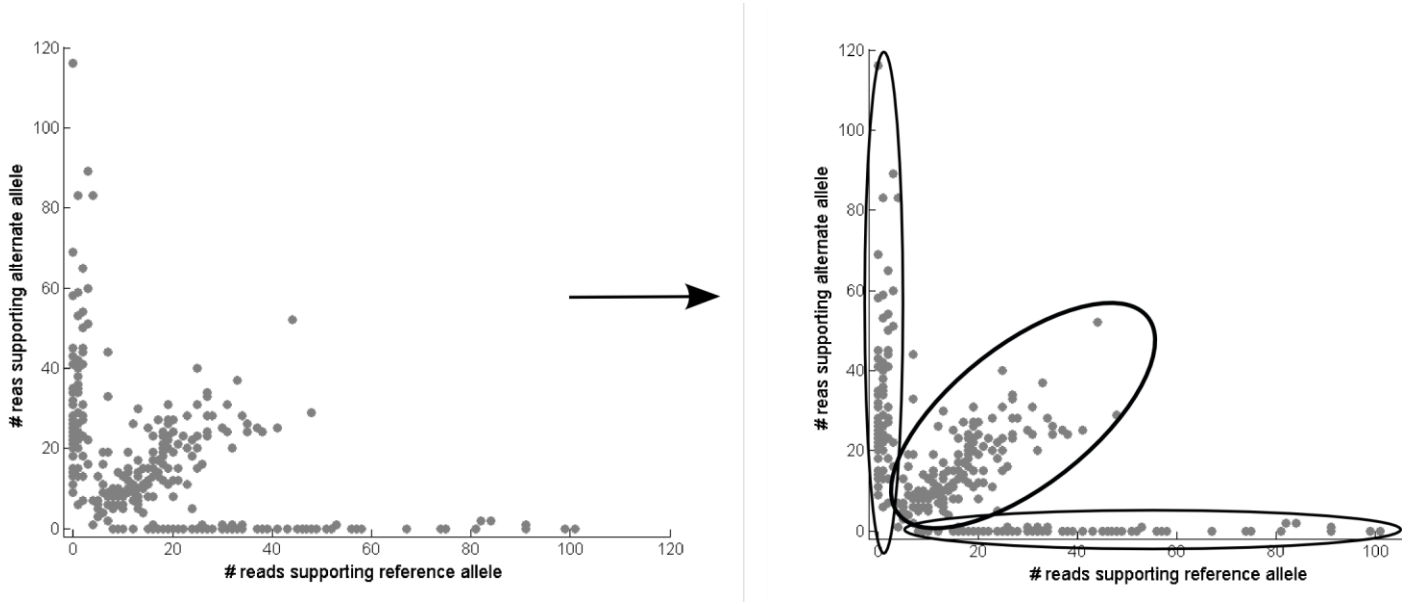
- For each indel: determine number of reads supporting the reference allele and the alternate allele

- Deletion chr3 4654039 -CTCA ref:alt 3:4

Step 6: Call population genotypes for each indel

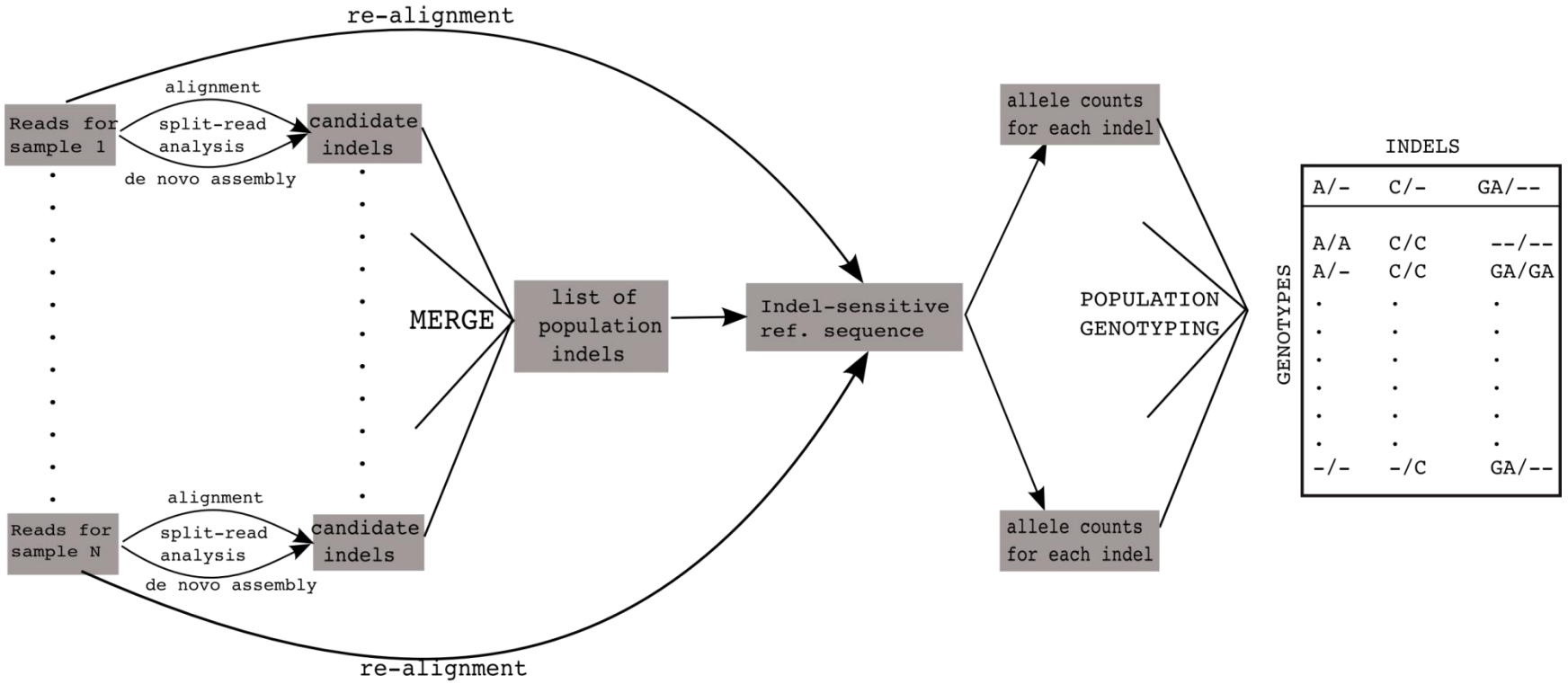
- bi-allelic indel with two alleles 'a' and 'b'
  - Three possible genotypes: <aa> <ab> <bb>
  - Let f(G) be the fraction of reads with the alternate allele for genotype 'G'
  - Ideal case: f(<aa>) = 0 f(<ab>) ~ 0.5 f(<bb>) = 1
- Given allele counts (a_i ,b_i) for N individuals
  - Determine most likely estimates of f(<aa>), f(<ab>) and f(<bb>) under a probabilistic model
  - Assign genotypes G_i to each individual
  - Filter out false indels

Step 6: Call population genotypes for each indel



- Use an MCMC algorithm to iteratively sample f(<aa>), f(<ab>), f(<bb>) and genotypes: $G_1, G_2 ..... G_N$
  - $Pr(a_i, b_i \mid G_i) = Binomial(a_i, a_i + b_i, f(G_i))$
  - $Pr(f(G) \mid G_1, G_2 .... G_N) =$ Beta distribution of allele counts
- Indels for which f(<aa>), f(<ab>) and f(<bb>) are not well-separated are removed

Indel detection and genotyping pipeline: Overview



Two applications of Indel pipeline

- Sequencing-based association studies to find rare and common disease-related variants
  - Sequencing of 2 candidate genes for obesity in 289 individuals of European ancestry (Harismendy, Bansal et al. submitted)

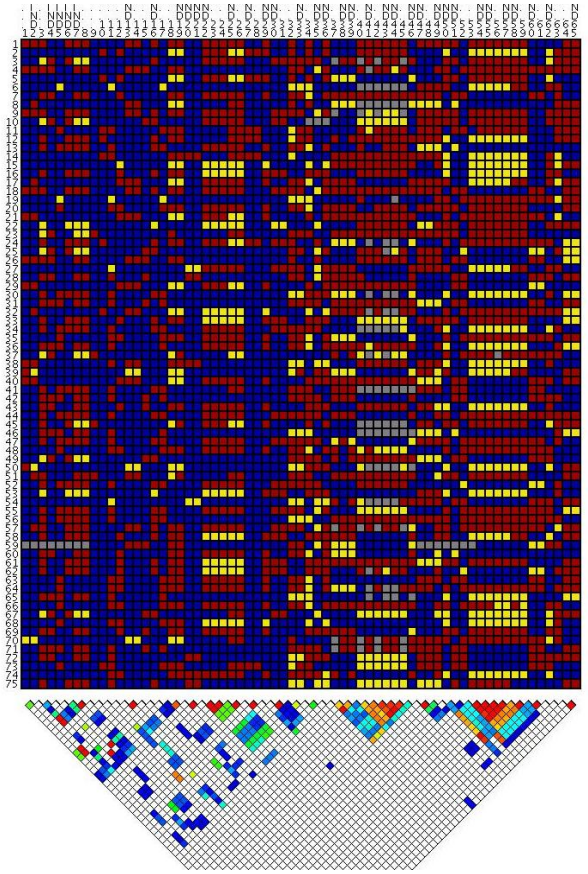
- Whole-genome sequencing of pathogens to study evolution and virulence
  - Sequencing of 39 Methycillin-resistant Staph. aureus (MRSA) strains

Population sequencing of 2 candidate genes for obesity

- 2 genes spanning ~ 188 kb sequenced in **289** individuals (143 with BMI > 40 and 146 with BMI < 30) using 36 bp PE Illumina reads
- ~ 1400 SNVs identified using MAQ alignment and SNP caller
- Initial set of ~ 1200 candidate indels identified using BWA, de novo assembly and split-read alignments
- Indel pipeline used to filter out false indels and assign population genotypes
- 100 indels (62 deletions and 38 insertions) with called genotypes for at least 75% samples
  - Gapped alignment (BWA) able to identify short indels (1-5 bp)
  - Split read mapping and de novo assembly able to identify several 8-30 bp indels and microsatellite polymorphisms

Population analysis of Indels in 289 individuals

- 54/100 indels with minor allele detected at least twice
- Only three indels failed Hardy-Weinberg equilibrium (p-value < 0.005)
- Most of the common indels in perfect or near-perfect Linkage Disequilibrium with another SNP or another indel



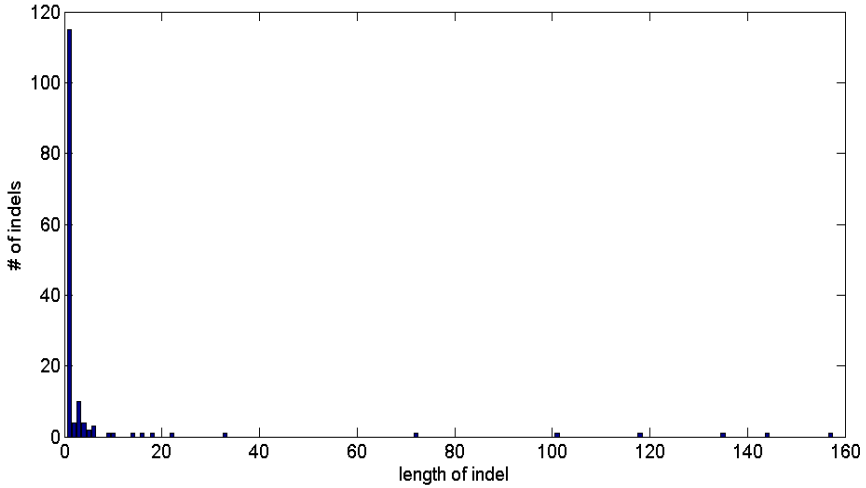
- **Comparison to 1000 Genomes indel calls from pilot 1 CEU population:**
  - 44/54 (81%) indels also called in 1000 genomes data
  - 2 indels with MAF >= 1% in 1000 genomes not detected

- **Compared indel calls using SAMtools for each individual with population-scale indel calls**
  - Slight under-calling of 'alternate homozygotes' using SAMtools indel caller
  - Lot of heterozygous indel calls in individual samples that appear to be false when looking at population data
  - Population scale analysis of indels makes it possible to filter out false variants that is impossible with individual sequence data

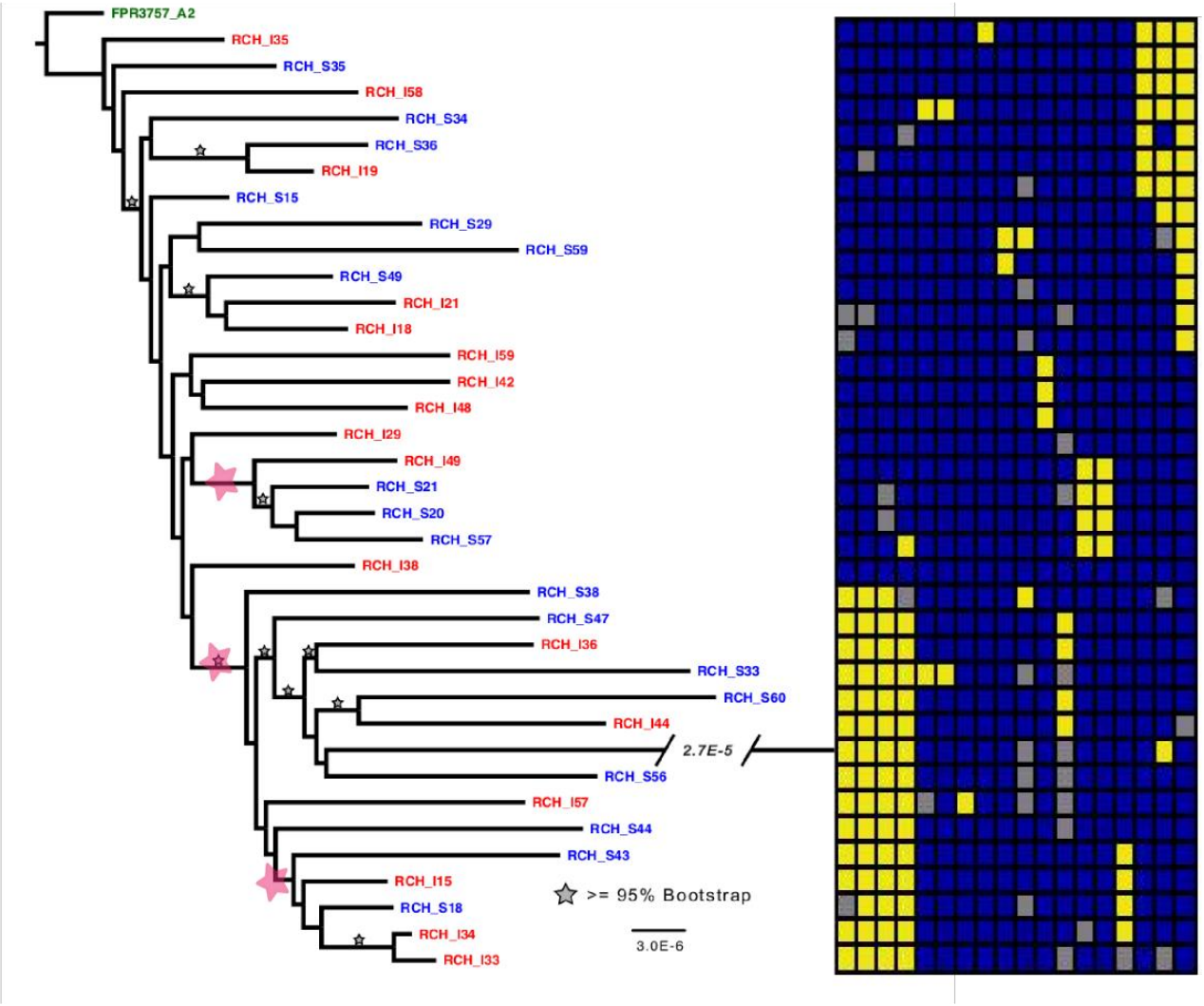
- **Features of "false" indels:**
  - 1-2 bp indels in homopolymer runs
  - Excess heterozygozity and fail Hardy-Weinberg equilibrium
  - Do not cluster into 2/3 well-separated clusters using MCMC algorithm

Sequencing of 39 MRSA strains

- Complete genomes of 39 CA-MRSA (Methicillin-resistant Staphylococcus aureus) strains sequenced using Illumina GA (Tewhey et al.)
- Reference strain USA300 also sequenced twice
- ~ 150 indels identified in the core genome (excluding the plasmids) using Indel-pipeline
- 116 1bp indels, several 10-200 bp indels and 13kb insertion
- Most of the indels are rare (minor allele observed more than once for 22 indels)



Indel genotypes consistent with SNP-based phylogeny



Summary

- *Automated pipeline for the accurate detection and genotyping of indels from short read population sequence data*
  - Use multiple methods for identifying candidate indels
  - Combine evidence from multiple samples to improve power to detect indels
  - Correct for misaligned reads leading to accurate allele counts
  - MCMC algorithm for population genotyping of indels and filtering out false indels
- *Enable the use of indels as markers for association mapping and evolutionary analyses*
- *Sequencing is the ultimate tool for genotyping all variants including indels*
- *Computational tools can enable maximum utilization of sequence data for variant discovery and genotyping*