SLIDER

MAXIMUM USE OF PROBABILITY INFORMATION FOR ALIGNMENT OF SHORT SEQUENCE READS AND SNP DETECTION

BIOINFORMATICS, 2009

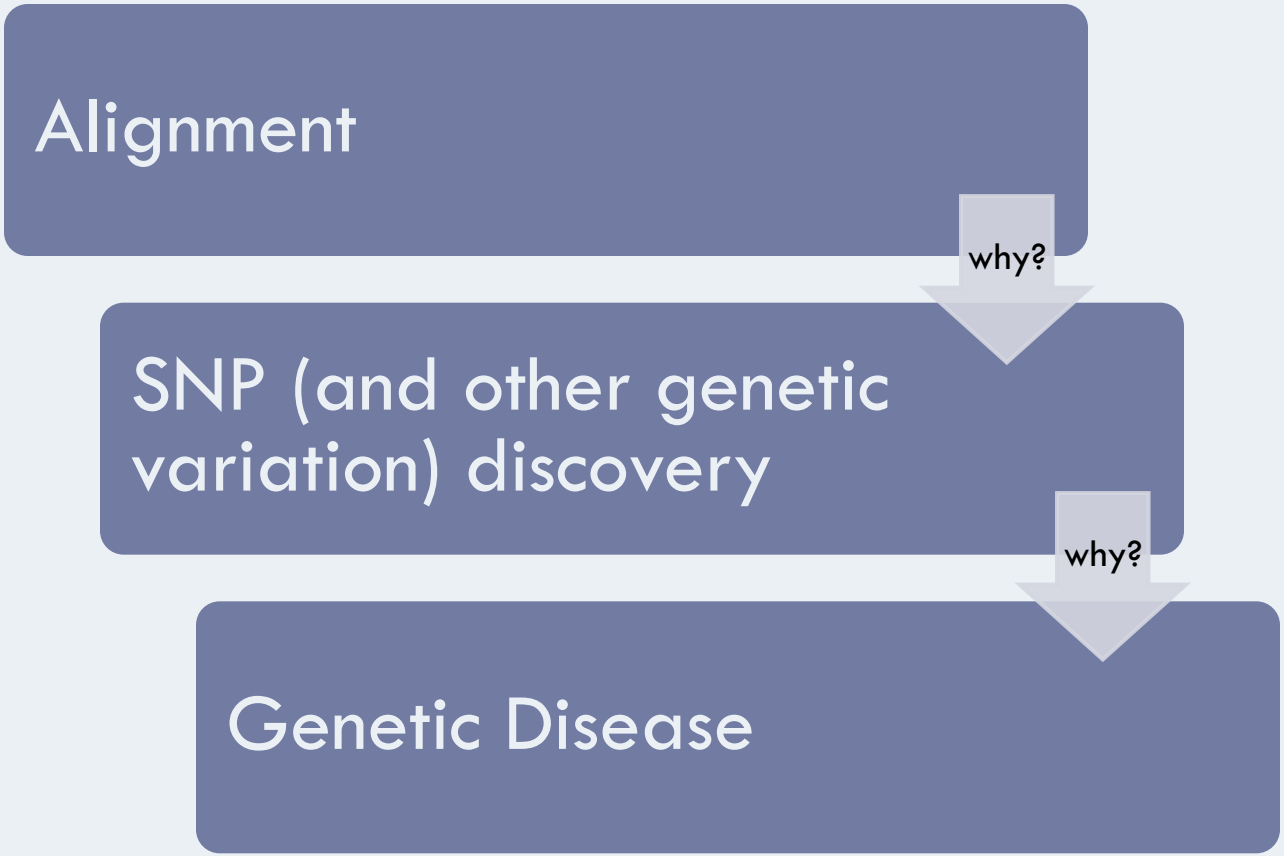
Marc Fiume

SLIDER in a slide

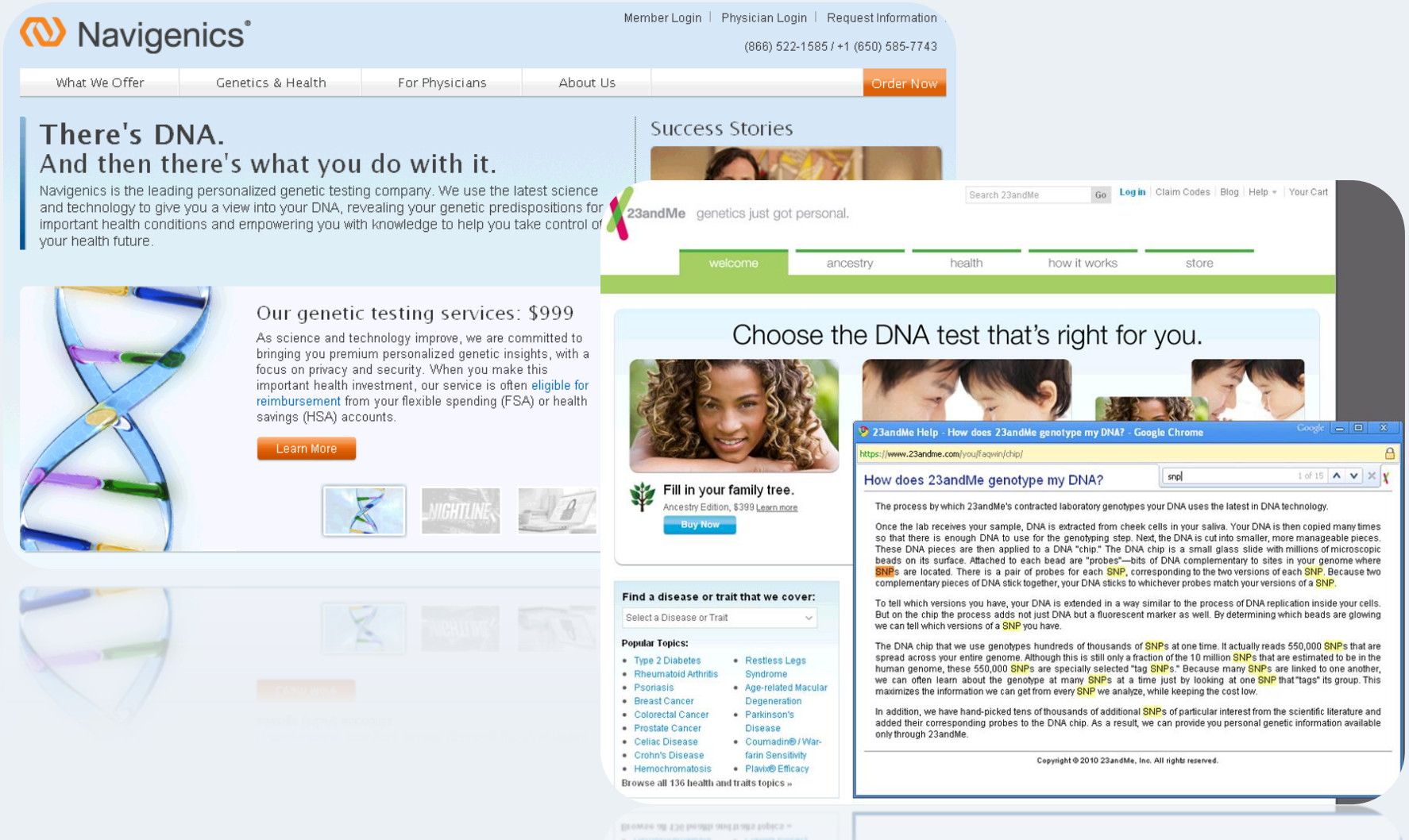
- SLIDER is…
  - not a very technical/algorithmic aligner
  - more of a proof-of-concept:
    - **can use confidence values to improve alignment**

Motivation

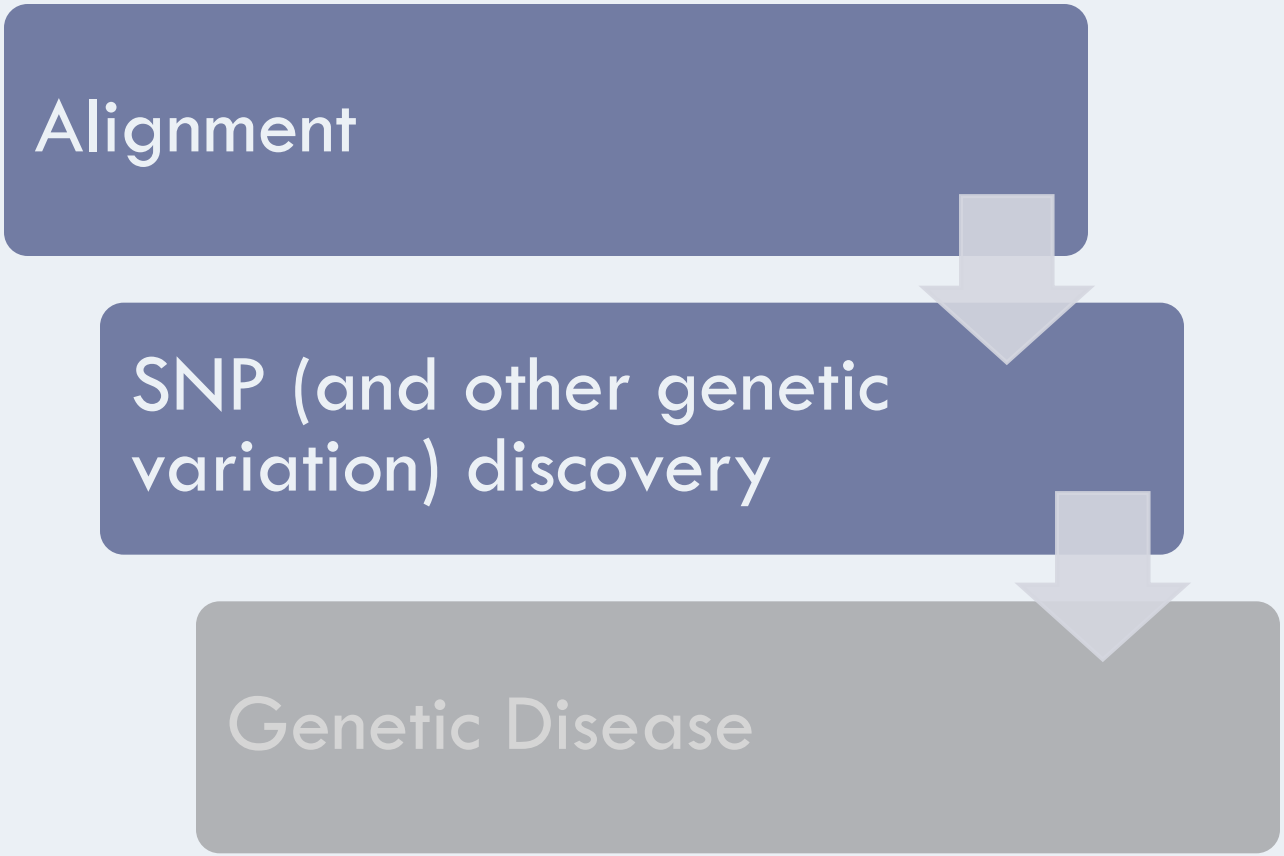
The Big Picture



Genetic Variation and Disease

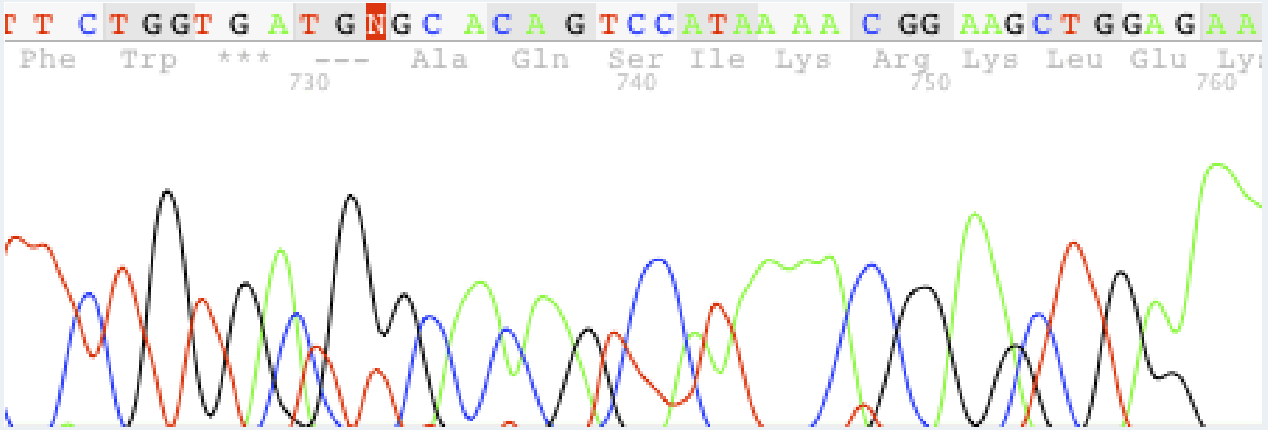


SLIDER


Alignment
SNP (and other genetic variation) discovery
Genetic Disease

The Improvement

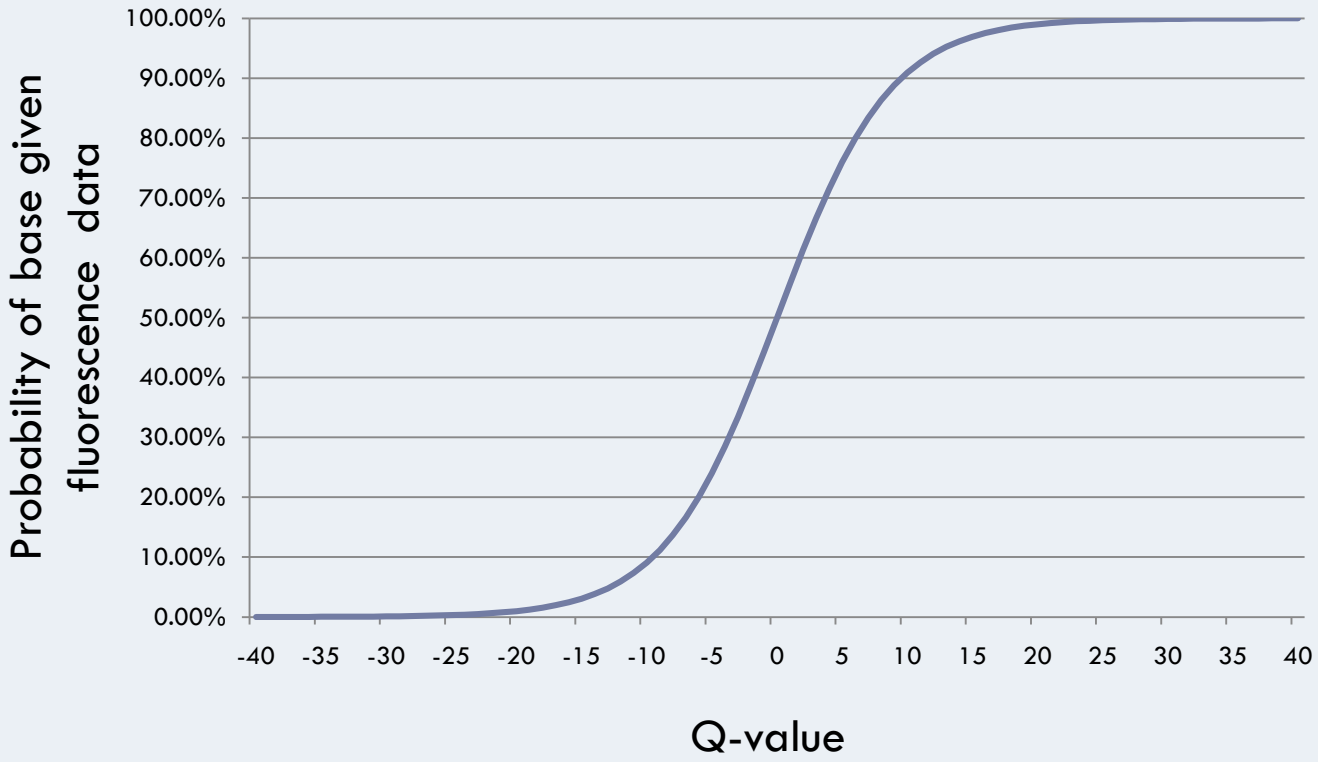
- Recall: base-caller has two outputs for each base:
  - **most likely nucleotide**
  - **confidence value**



- most aligners **ignore confidence values**

Illumina's confidence values

- stored in ***prb* files**
  - for each base called, 4 **Q-values** (Q_A, Q_C, Q_G, Q_T)



SLIDER: Alignment

1. Create **reads** database

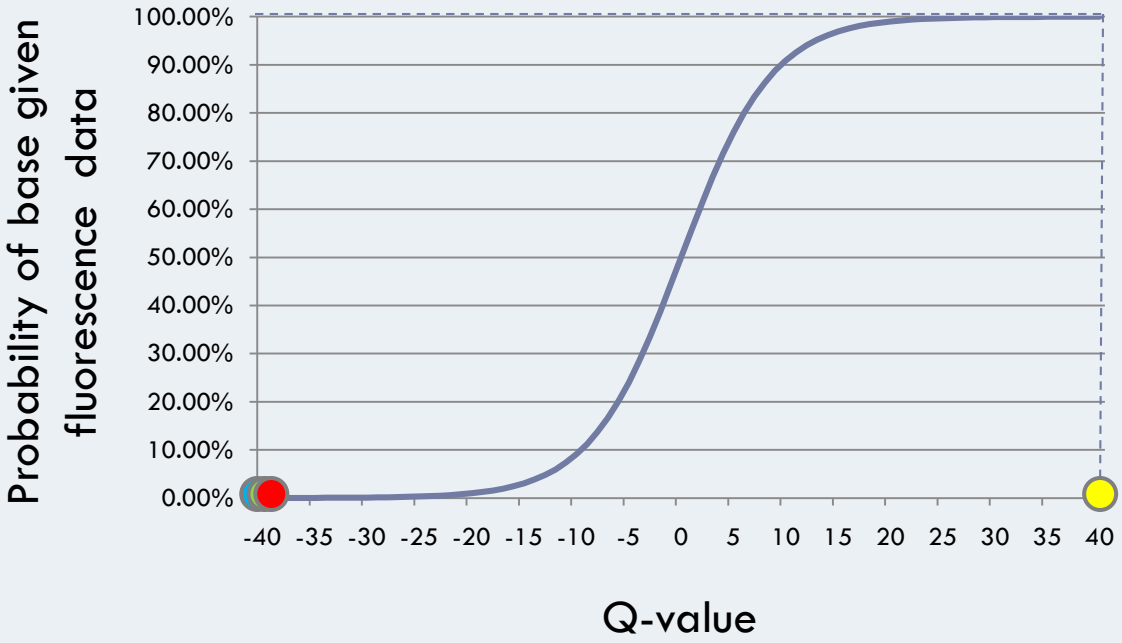
- produce all “probable” reads based on Q-values

Base crispness

- Each base in a read is classified as:
  - **crisp** base
  - **non-crisp** base

Base crispness

- **crisp base:** confidence ~100% for a particular base



QA	40	100%
QC	-40	0%
QG	-40	0%
QT	-40	0%

Base crispness

- **non-crisp base:** confidence <100% for any base

QA	**10**	**90%**
QC	-40	0%
QG	-10	10%
QT	-40	0%

1. Create **reads** database

read position

	1	2	3	4	5	6	7	8
QA	ε	ε	ε	99	ε	99	3	ε
QC	ε	90	ε	ε	ε	ε	ε	ε
QG	99	9	ε	ε	ε	ε	96	ε
QT	ε	ε	99	ε	99	ε	ε	99

1. Create **reads** database

read position

	1	2	3	4	5	6	7	8
QA				99		99	3	
QC		90						
QG	99	9					96	
QT			99		99			99

1. Create **reads** database

Reads Database:

G**C**TATA**A**T	99*__90__*99*99*99*99* __3__*99	= 2.5%
G**G**TATA**A**T	99* __9__*99*99*99*99* __3__*99	= 0.3%
G**C**TATA**G**T	99*__90__*99*99*99*99*__96__*99	= 81.3%
G**G**TATA**G**T	99* __9__*99*99*99*99*__96__*99	= 8.1%

2. Create a **reference** database

- put all k-mers from reference into database

- sort database

3. **Align** the reads DB to the reference DB

- every entry in reads DB is aligned with every entry in reference DB
  - **no match:**
    - read aligns to no location
  - **unique match:**
    - read aligns to one location
  - **multi-match:**
    - read aligns to multiple locations

SLIDER: SNP Detection

Support for SNPs

- consider positions of **mismatches**

Support for Event	SNP	Sequence Error / Misalignment
Coverage	**High**	Low
Percent coverage that is mismatched	**High**	Low
Sequence complexity	**High**	Low
Read Weight	**High**	Low

Results

RESULTS

- Align reads to
  - **correct** reference (RefBAC)
  - **incorrect** reference (RefEX)

RESULTS : Alignment Accuracy

- P_{mis} ~ % mapped to **incorrect** reference
- P_{uq} ~ % mapped to **correct** reference

Table 3. Alignment results

	27		32		36	
	$P_{mis}(\%)$	$P_{uq}(\%)$	$P_{mis}(\%)$	$P_{uq}(\%)$	$P_{mis}(\%)$	$P_{uq}(\%)$
Eland	2.791	76.65	3.002	79.47		
RMAP	2.828	76.69	3.002	79.45	3.520	81.68
Slider	1.169	77.08	1.172	80.19	1.302	83.16

Results of aligning sequences from CT302 to its reference RefBAC and the human genome excluding chromosome 6.

RESULTS: SNP Accuracy

- limited validation
- claim reasonably accurate SNP prediction at low coverage

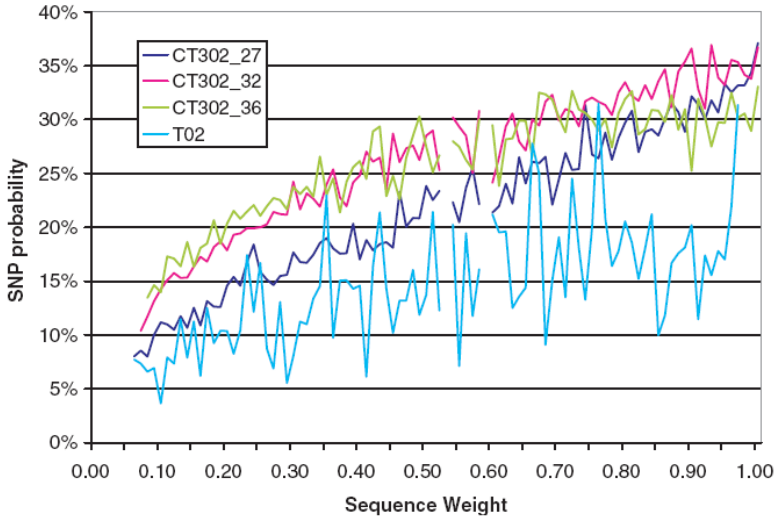


Fig. 2. Probability that a given base mismatch is a true SNP as a function of the read sequence weight.

Conclusions

Conclusions

- read alignment is an important first step in genetic variation discovery
- remember that read sequences are abstractions of noisy data
- can incorporate confidence values in alignment

END