

SLIDER

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

BIOINFORMATICS, 2009

Marc Fiume

SLIDER in a slide

2

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

3

Motivation

The Big Picture

4

Alignment

why?

SNP (and other genetic variation) discovery

why?

Genetic Disease

Genetic Variation and Disease

5



Member Login | Physician Login | Request Information

(866) 522-1585 / +1 (650) 585-7743

What We Offer

Genetics & Health

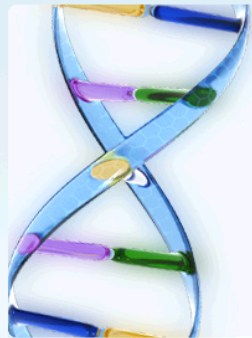
For Physicians

About Us

Order Now

There's DNA. And then there's what you do with it.

Navigenics is the leading personalized genetic testing company. We use the latest science and technology to give you a view into your DNA, revealing your genetic predispositions for important health conditions and empowering you with knowledge to help you take control of your health future.



Our genetic testing services: \$999

As science and technology improve, we are committed to bringing you premium personalized genetic insights, with a focus on privacy and security. When you make this important health investment, our service is often **eligible for reimbursement** from your flexible spending (FSA) or health savings (HSA) accounts.

Learn More



Success Stories



genetics just got personal.

Search 23andMe

Go

Log in

Claim Codes

Blog

Help

Your Cart

welcome

ancestry

health

how it works

store

Choose the DNA test that's right for you.



Fill in your family tree.

Ancestry Edition, \$399 [Learn more](#)

Buy Now



23andMe Help - How does 23andMe genotype my DNA? - Google Chrome

https://www.23andme.com/you/faq/in/chip/

How does 23andMe genotype my DNA?

snpl

1 of 15

The process by which 23andMe's contracted laboratory genotypes your DNA uses the latest in DNA technology.

Once the lab receives your sample, DNA is extracted from cheek cells in your saliva. Your DNA is then copied many times so that there is enough DNA to use for the genotyping step. Next, the DNA is cut into smaller, more manageable pieces. These DNA pieces are then applied to a DNA "chip." The DNA chip is a small glass slide with millions of microscopic beads on its surface. Attached to each bead are "probes"—bits of DNA complementary to sites in your genome where SNPs are located. There is a pair of probes for each SNP, corresponding to the two versions of each SNP. Because two complementary pieces of DNA stick together, your DNA sticks to whichever probes match your versions of a SNP.

To tell which versions you have, your DNA is extended in a way similar to the process of DNA replication inside your cells. But on the chip the process adds not just DNA but a fluorescent marker as well. By determining which beads are glowing we can tell which versions of a SNP you have.

The DNA chip that we use genotypes hundreds of thousands of SNPs at one time. It actually reads 550,000 SNPs that are spread across your entire genome. Although this is still only a fraction of the 10 million SNPs that are estimated to be in the human genome, these 550,000 SNPs are specially selected "tag SNPs." Because many SNPs are linked to one another, we can often learn about the genotype at many SNPs at a time just by looking at one SNP that "tags" its group. This maximizes the information we can get from every SNP we analyze, while keeping the cost low.

In addition, we have hand-picked tens of thousands of additional SNPs of particular interest from the scientific literature and added their corresponding probes to the DNA chip. As a result, we can provide you personal genetic information available only through 23andMe.

Copyright © 2010 23andMe, Inc. All rights reserved.

Find a disease or trait that we cover:

Select a Disease or Trait

Popular Topics:

- Type 2 Diabetes
- Rheumatoid Arthritis
- Psoriasis
- Breast Cancer
- Colorectal Cancer
- Prostate Cancer
- Celiac Disease
- Crohn's Disease
- Hemochromatosis
- Restless Legs Syndrome
- Age-related Macular Degeneration
- Parkinson's Disease
- Coumadin® / Warfarin Sensitivity
- Plavix® Efficacy

Browse all 136 health and traits topics

SLIDER

6

Alignment

SNP (and other genetic
variation) discovery

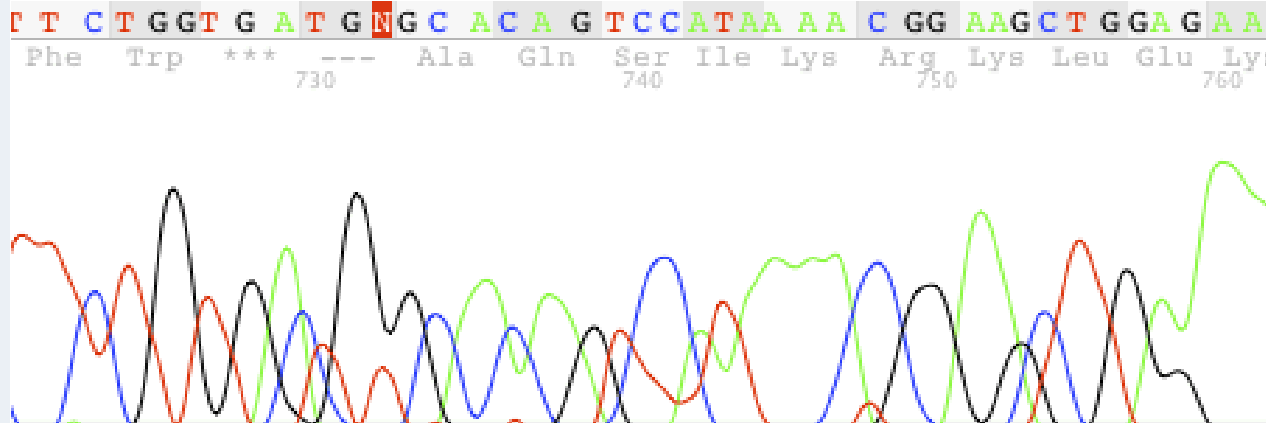
Genetic Disease



The Improvement

8

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

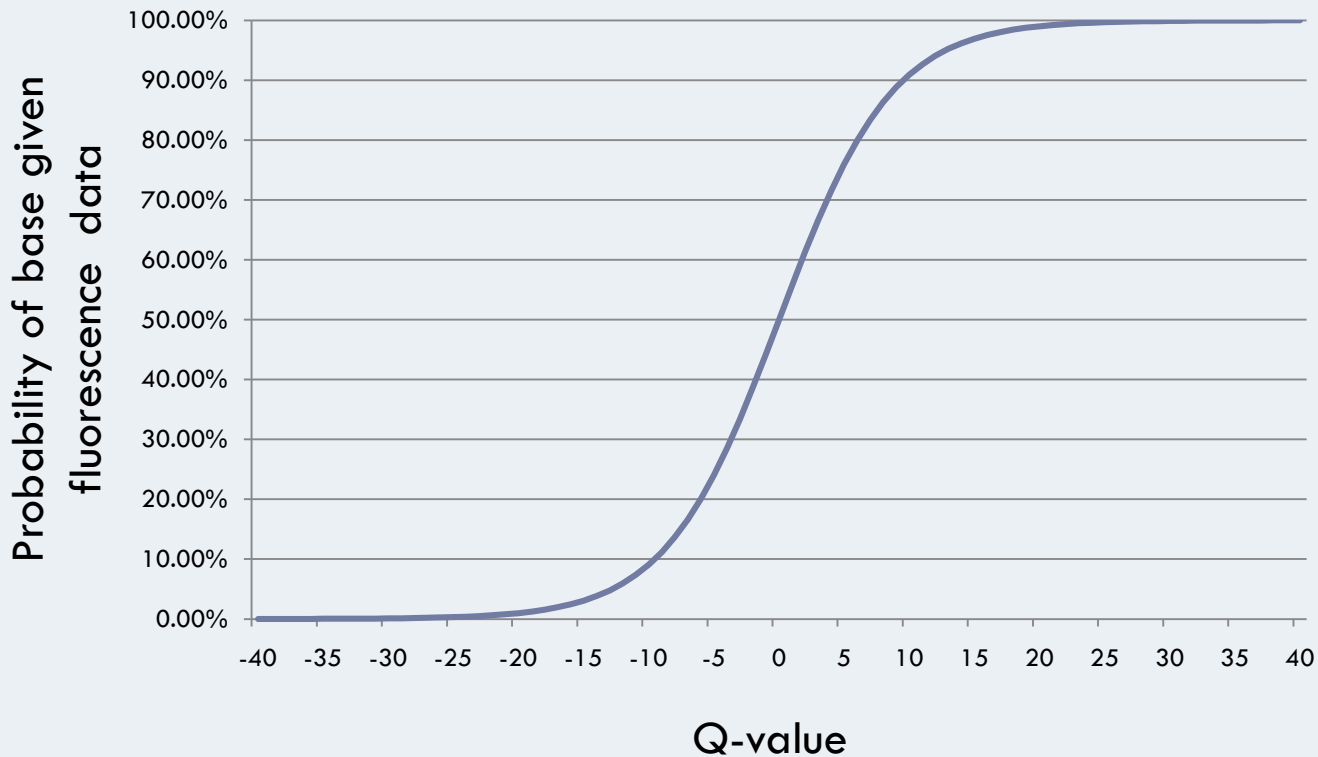


- most aligners **ignore confidence values**

Illumina's confidence values

9

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



12

SLIDER: Alignment

1. Create reads database

13

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

Base crispness

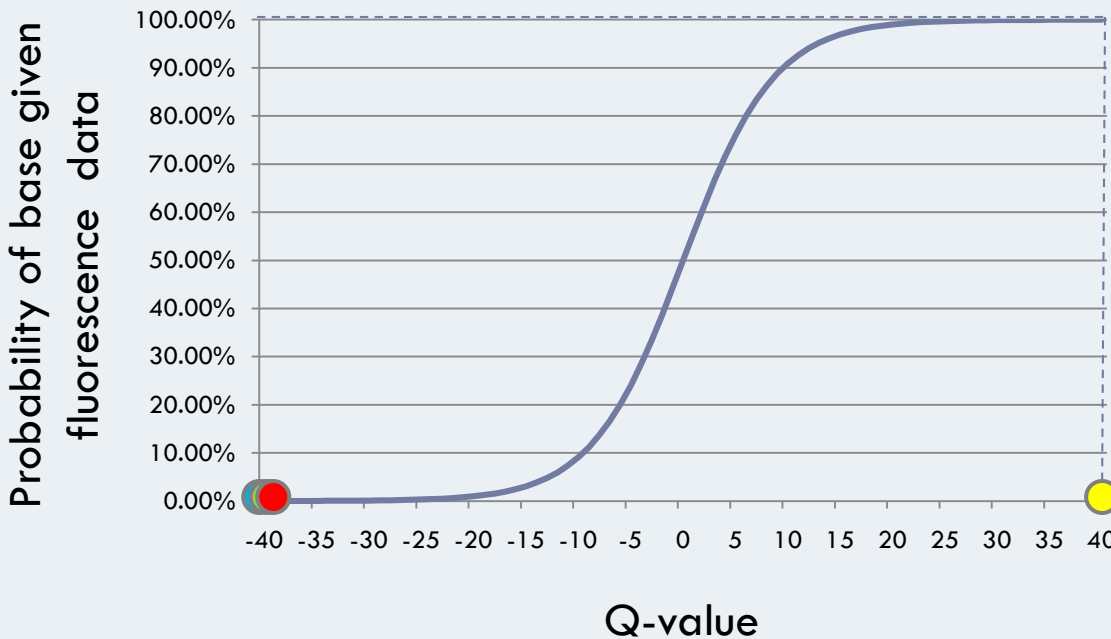
14

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

Base crispness

16

- **crisp base:** confidence $\sim 100\%$ for a particular base

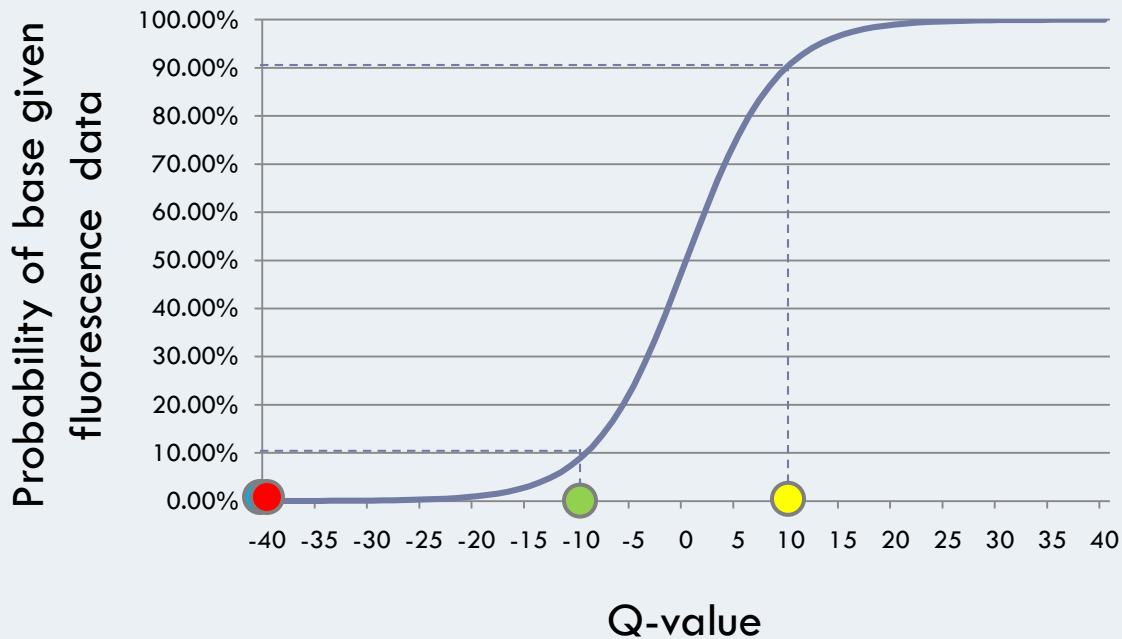


QA	40	100%
QC	-40	0%
QG	-40	0%
QT	-40	0%

Base crispness

17

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



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

1. Create reads database

18

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

19

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

20



Reads Database:

GCTATA A T	99* 90 *99*99*99*99* 3 *99	=	2.5%
GGTATA A T	99* 9 *99*99*99*99* 3 *99	=	0.3%
GCTATA G T	99* 90 *99*99*99*99* 96 *99	=	81.3%
GGTATA G T	99* 9 *99*99*99*99* 96 *99	=	8.1%

2. Create a reference database

21

- put all k-mers from reference into database



- sort database

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

22

- 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

23

SLIDER: SNP Detection

Support for SNPs

24

- consider positions of **mismatches**

ATTAGAT**A**GATCGAT
AT**C**GATCGACG... } reads
GATTAGAT**A**GATCGA
ATTAGAT**A**GATCGAT
...CGATTAGAT**C**GATCGATCG... reference

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

25

Results

RESULTS

26

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

RESULTS : Alignment Accuracy

27

- $P_{\text{mis}} \sim$ % mapped to **incorrect** reference
- $P_{\text{uq}} \sim$ % mapped to **correct** reference

Table 3. Alignment results

	27		32		36	
	$P_{\text{mis}}(\%)$	$P_{\text{uq}}(\%)$	$P_{\text{mis}}(\%)$	$P_{\text{uq}}(\%)$	$P_{\text{mis}}(\%)$	$P_{\text{uq}}(\%)$
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

28

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

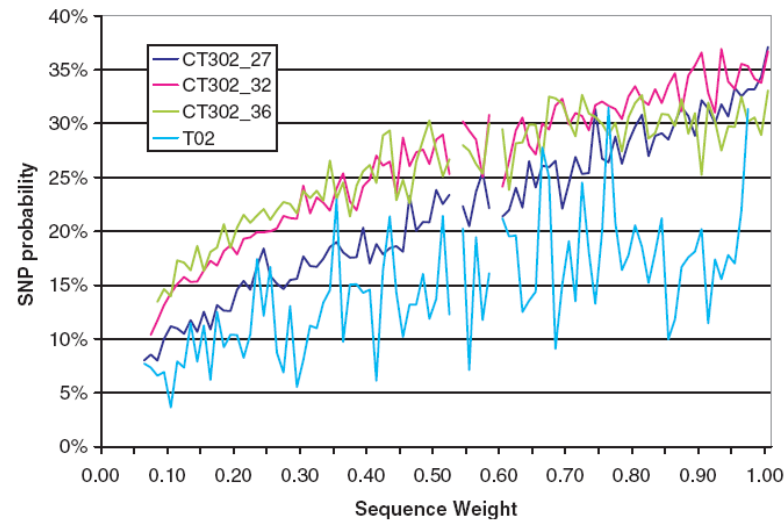


Fig. 2. Probability that a given base mismatch is a true SNP as a function of the read sequence weight.

29

Conclusions

Conclusions

30

- 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