Институт биоинформатики

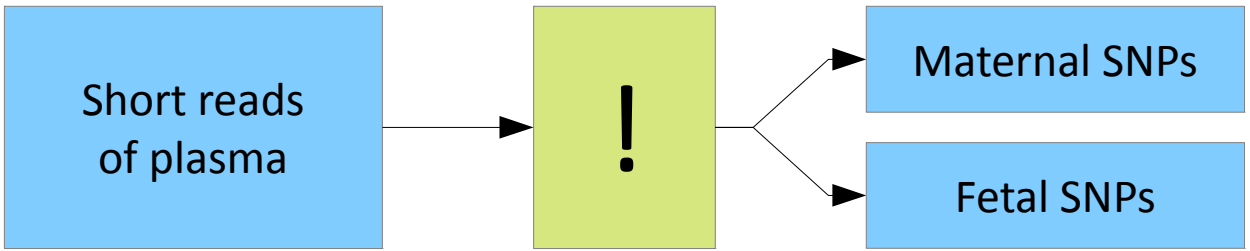
BINOM

The "Mother-Fetus" Project

Scientific advisor: Anton Afanasiev, iBinom
Students: Nadya Sitdykova, Kirill Grigorev

Project description

Task: develop an algorithm to separate simultaneously sequenced genomes of a mother and a fetus.



Base premise:

- the phenomenon of cell-free fetal DNA;
- research of Quake labs at Stanford (*Nature*[1]), University of Washington (*Science*[2]), BGI Shenzhen (*Genome Medicine*[3]).

1. doi:10.1038/nature11251
2. doi:10.1126/scitranslmed.3004323
3. doi:10.1186/gm422

Novelty

The cited researchers used more data than we do.

The described methods relied on the use of parental haplotypes.

Chen et al. **(BGI)**	Kitzman et al. **(UW)**	Fan et al. **(Stanford)**
Trio strategy with corresponding grandparents and CHS	Maternal: fosmid-based approach	Maternal: single-cell approach

This project's goal was to perform analyses based on a typical dataset (sequencing of one patient).
Thus there was a need for a new algorithm.

Data

This is yet to become the "hip" area of research, and of the data that exists a lot is behind restricted access.

The data we used:

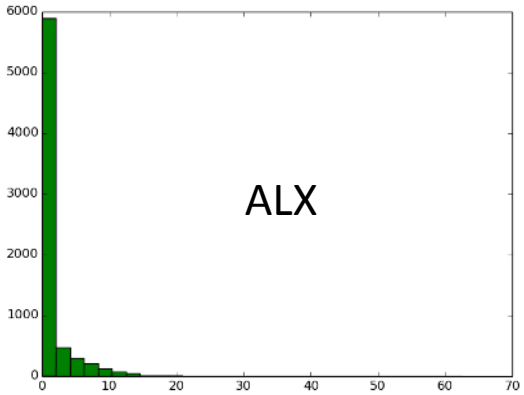
- Codename "ALX". Pregnant female's plasma + microarrays of the mother and the father.
- Data from the BGI Shenzhen study. Plasma, father, mother, offspring.
- Some requests pending.
- Simulated data.

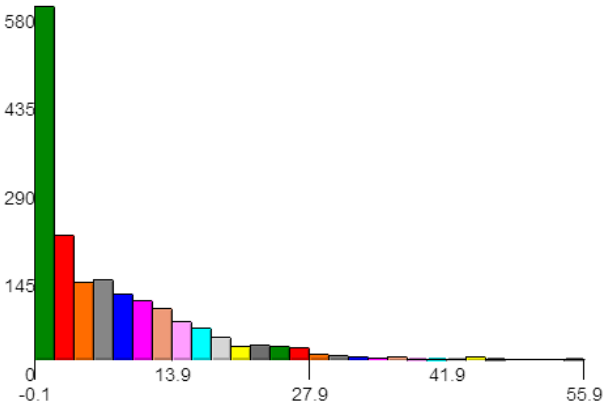
What we've found right off the bat:

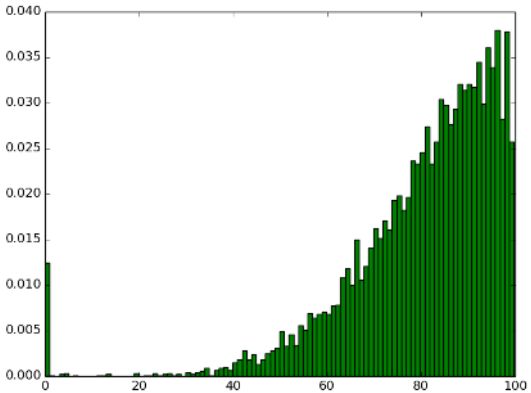
cffDNA degrades highly unevenly.

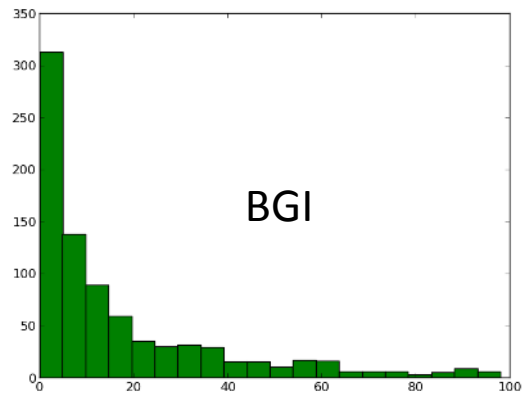
Global fetal DNA fraction isn't a meaningful estimate, but local is.

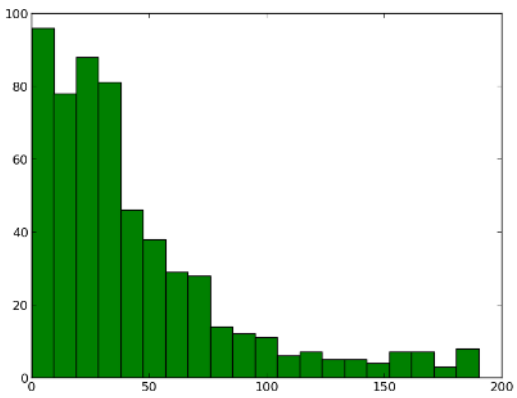
The real picture







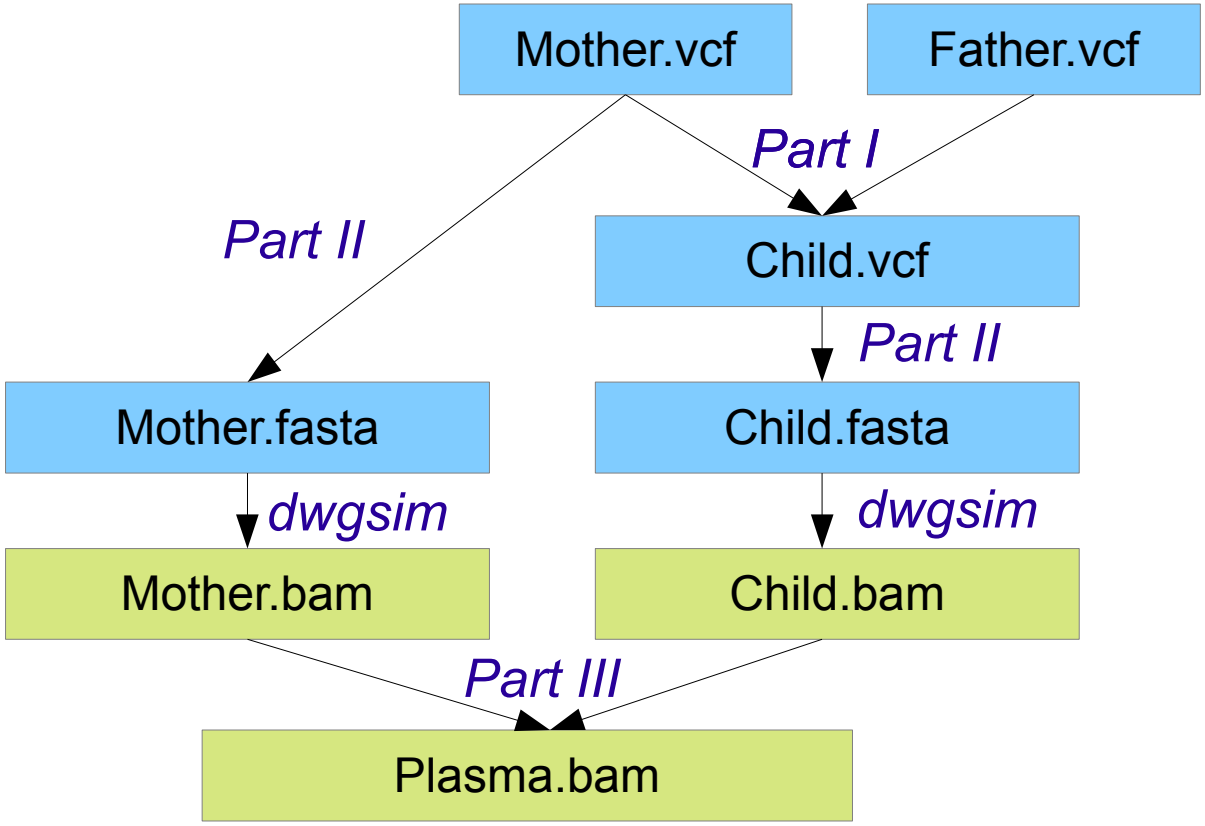




X axis:
minor allele fraction = 2*G/(A+G)

Y axis:
frequency

Data simulation



Part I

Simulate recombination

Part II

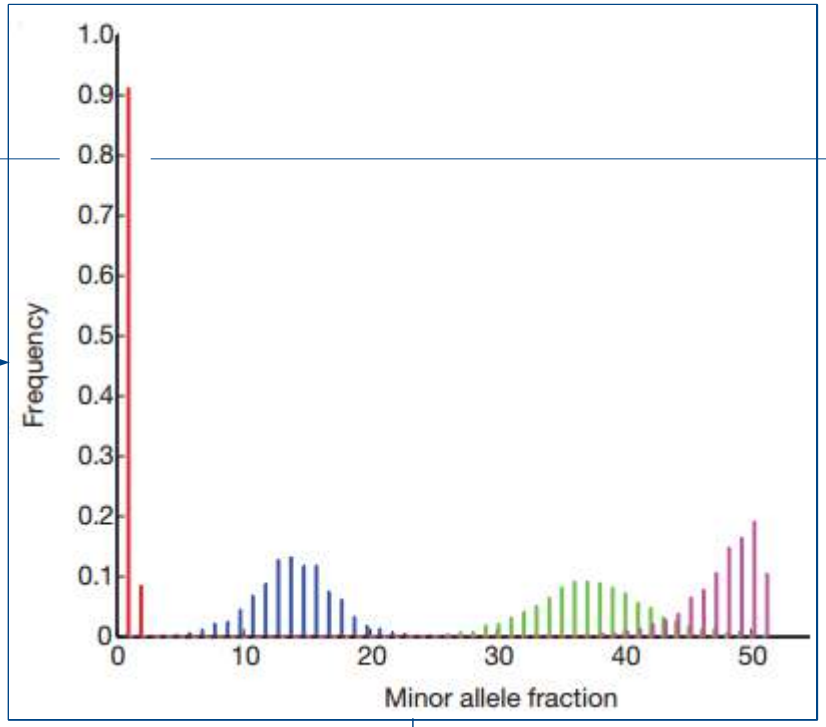
- Generate haploid reference
- Substitute SNPs in reference with SNPs from vcf

Part III

- Choose random positions
- Get all mother's reads and part of child's reads mapped on it

The idea

M	F	|G|
AA	AA	0
AA	AG	ε
AG	AA	0.5-ε
AG	AG	0.5



Beagle +
1000 genomes

Recover previously
missed SNPs
within haploblocks

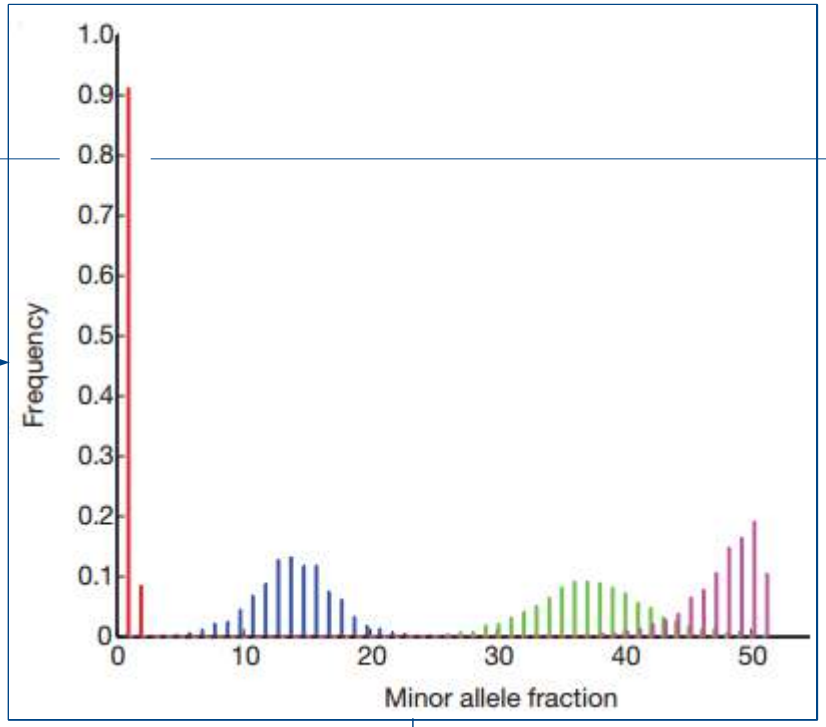
Haplotype
imputation

Maternal variants

Fetal variants

The idea

M	F	|G|	M(t)	M(u)	P(t)
AA	AA	0	A	A	A
AA	AG	ε	A	A	G
AG	AA	0.5-ε	A	G	A
AG	AG	0.5	?	?	?



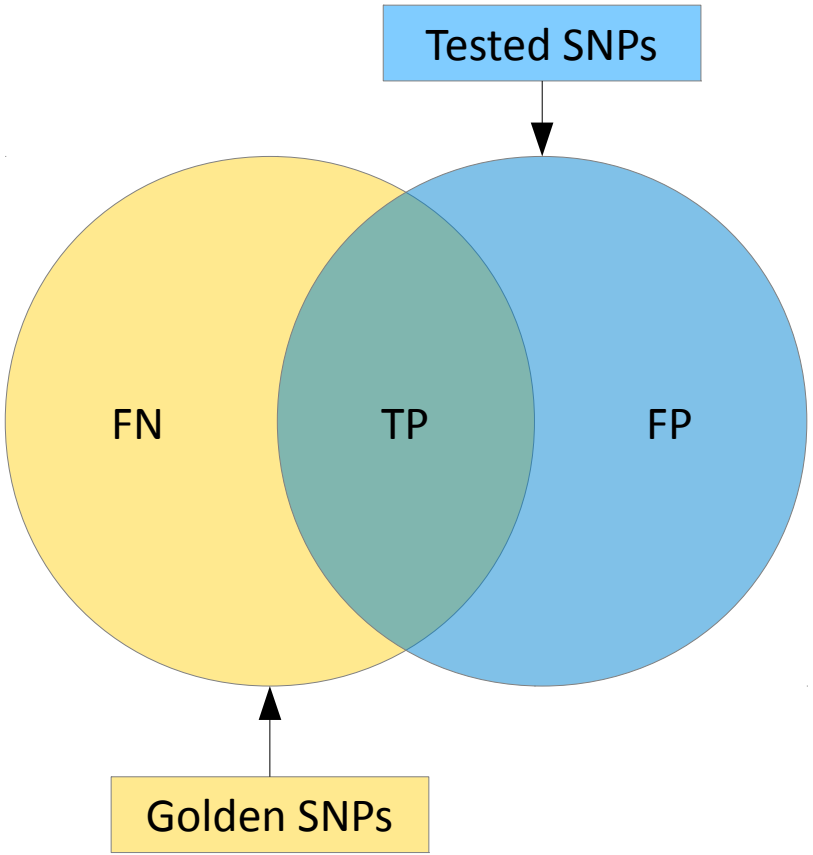
Recover previously missed SNPs within haploblocks

Haplotype imputation

Maternal variants

Fetal variants

Quality checker




```
TP  = True Positive
FP  = False Positive
FN  = False Negative

SNP = {position, alleles}

Precision = TP/(TP+FP)
Recall    = TP/(TP+FN)
F measure = 2PR/(P+R)
```

Results

Dataset		ALX	BGI	Simulated
Child SNP's		5655	6310	2040
Algorithm SNP's		133113	327606	51293
Full	Count	3713	2843	1378
	Precision	0.880	0.009	0.0027
	Recall	0.657	0.451	0.675
	F-measure	0.752	0.017	0.052
Positions	Count	4139	5339	1660
	Precision	0.981	0.016	0.0032
	Recall	0.732	0.846	0.814
	F-measure	0.838	0.032	0.062

Raw;
Min depth = 50

Filtering results: an example

Simulated, chr1	Before dbSNP validation	After dbSNP validation
Offspring SNPs	123	123
Algorithm SNPs	6517	1106
TP	92	92
FN	31	31
FP	6425	1014
Precision	0.014	0.084
Recall	0.748	0.750
F-measure	0.028	0.151

The end?