ADMM Fused Lasso for Copy Number Variation Detection in Human Genomes

Yifei Chen and Jacob Biesinger

3 March 2011

Human Variation

Small differences in the human genome can have big effects

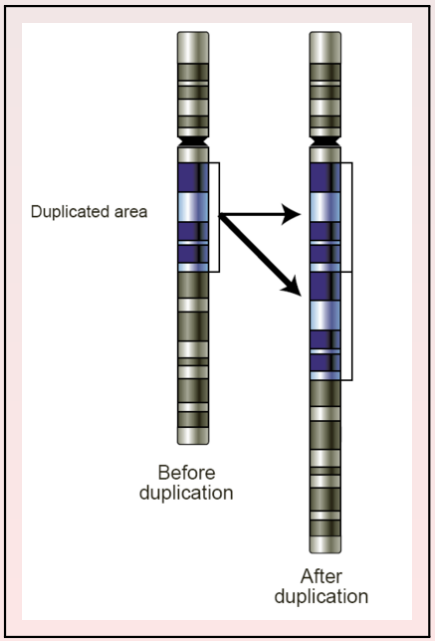
We're all 99.9% the same...right?

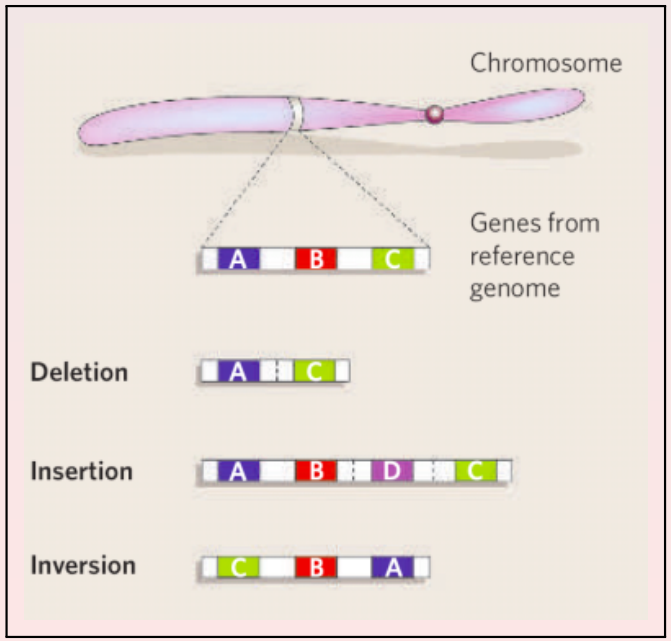
Most people have the exact same letters in their genome sequence.
We differ at only 0.1% of our 3.2 billion bases.
So what's responsible for the differences between us?

We're all 99.9% the same. . . right?

Most people have the exact same letters in their genome sequence.
We differ at only 0.1% of our 3.2 billion bases.
So what's responsible for the differences between us?

Perhaps it's **structural variation**





Structural Variation

- Thought to be much more common than single base mutations (12% of the genome, Redon et al. 2006)

[1] Figure from slides by Michael Snyder

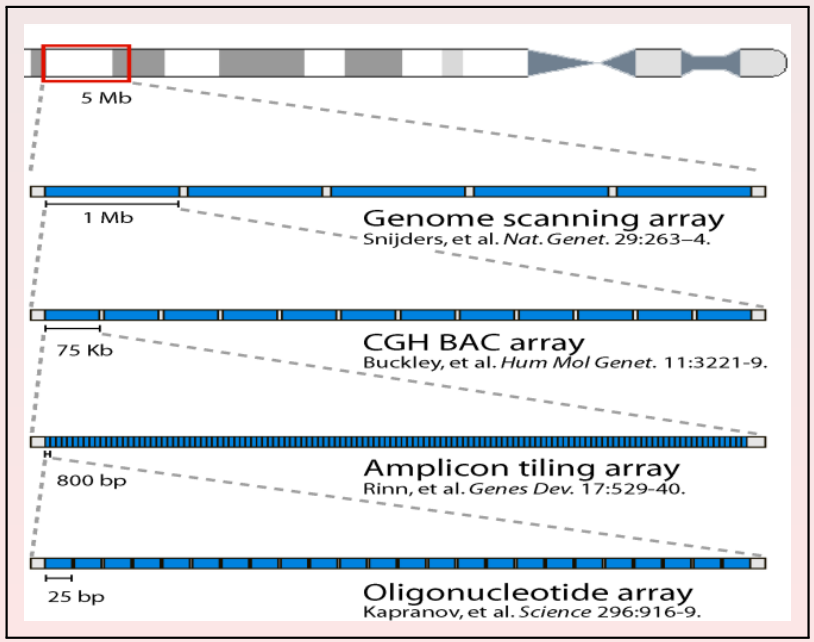
Structural Variation

- Thought to be much more common than single base mutations (12% of the genome, Redon et al. 2006)
- Difficult to study– Rearrangements can be complex, and may involve repetitive elements.

[1] Figure from slides by Michael Snyder

Structural Variation

- Thought to be much more common than single base mutations (12% of the genome, Redon et al. 2006)
- Difficult to study– Rearrangements can be complex, and may involve repetitive elements.
- Until recently, resolution was low

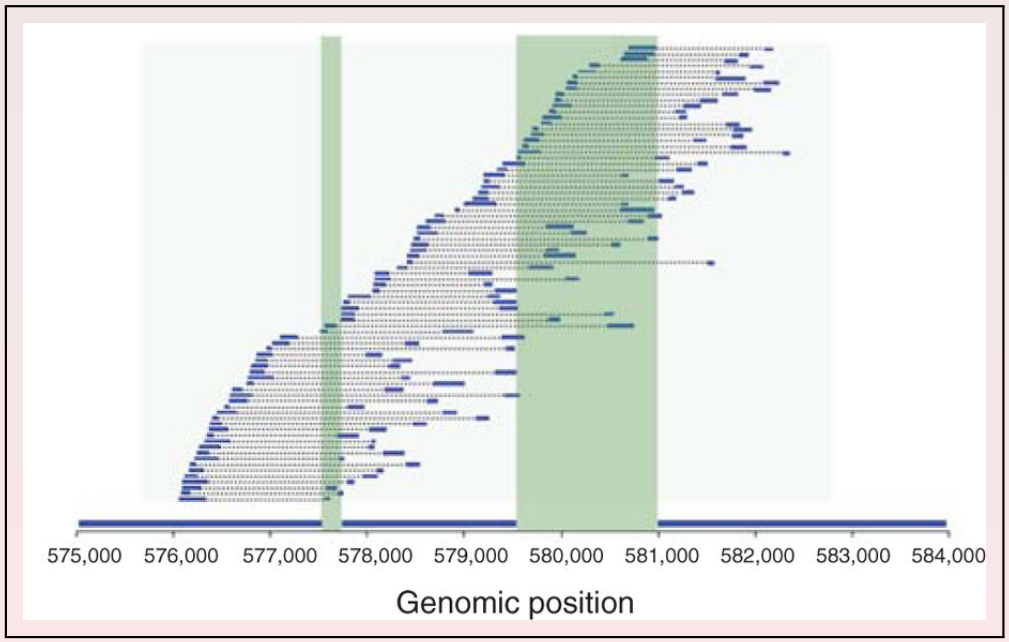


[1]

[1] Figure from slides by Michael Snyder

1000 Genomes Project

- Recent work to do medium coverage of 1000 human genomes (still in progress but we get to play with their data already!).
- Most of the data is "paired-end" reads– a pair of short strings (36 letters) from the genome.

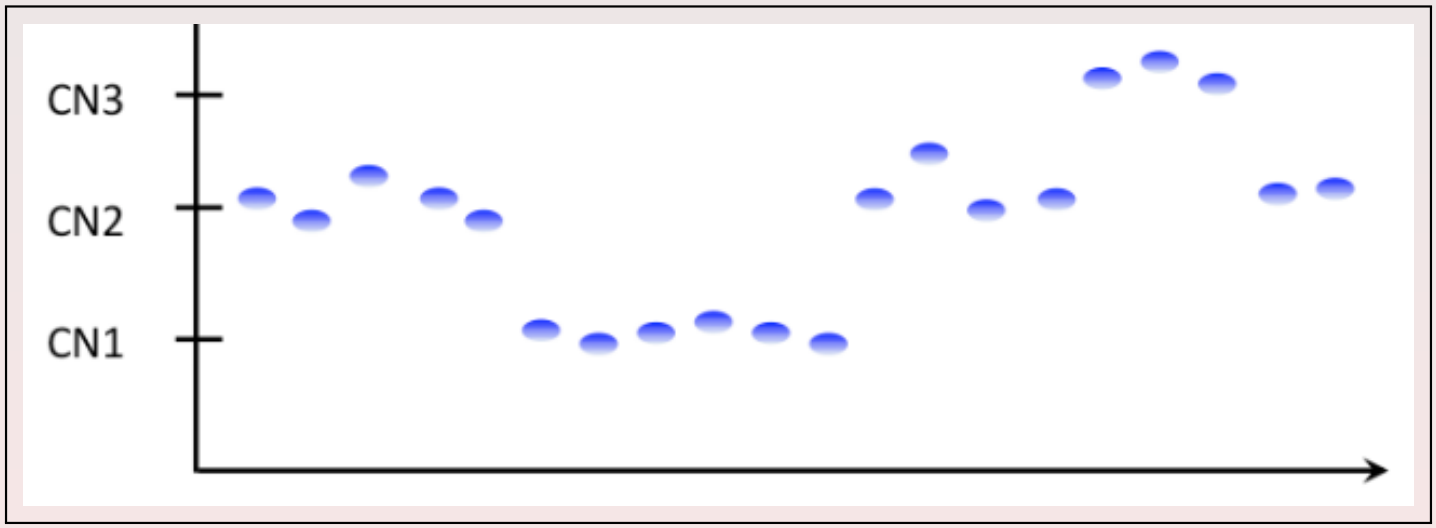


2

[2]Figure from "3K Long-Tag Paired End sequencing with the Genome Sequencer FLX System" Nature Methods 5, May 2008

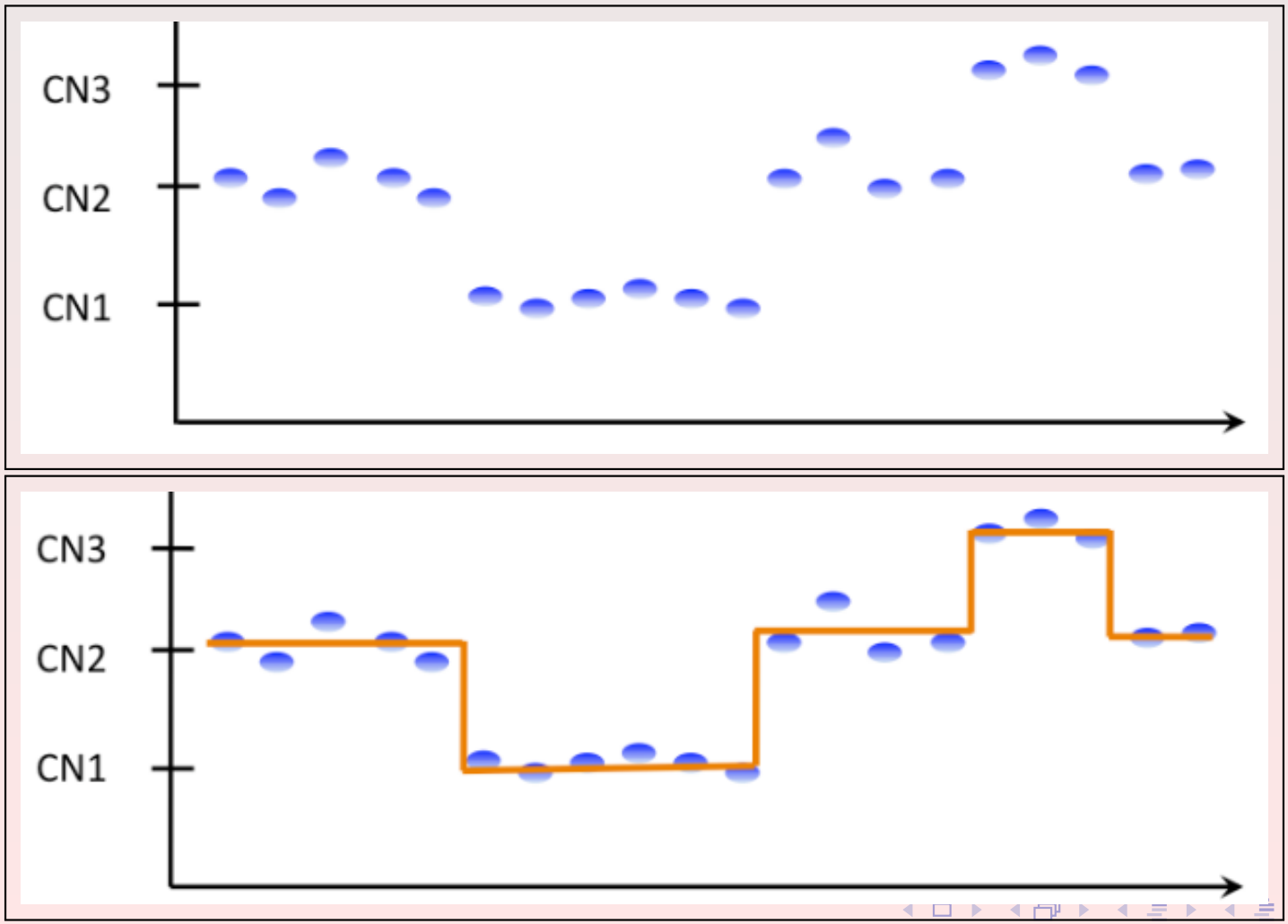
Detecting Copy Number Variation from short reads

Given noisy reads from a genome, can we determine how many copies there are of each gene?



Detecting Copy Number Variation from short reads

Given noisy reads from a genome, can we determine how many copies there are of each gene?



Fundamentals: Dual ascend method

Equality Constraint Problem:

$$\min_x f(x)$$
$$s.t.\ Ax = b$$

Lagrangian:

$$L(x, v) = f(x) + v^T(Ax - b)$$

Dual ascend framework:

$$x^{k+1} = \arg\min_x L\left(x, v^k\right)$$
$$v^{k+1} = v^k + \alpha^k\left(Ax^{k+1} - b\right)$$

Fundamentals: Two variations

Dual decomposition

$$f(x) = \sum_{i=1}^{N} f_i(x_i)$$

Augmented Lagrangian and method of multipliers

$$\min_x f(x) + \frac{\mu}{2}\|Ax - b\|_2^2$$
$$s.t.\ Ax = b$$

$$L_\mu(x, v) = f(x) + v^T(Ax - b) + \frac{\mu}{2}\|Ax - b\|_2^2$$

Alternating direction method of multipliers - ADMM

Problem Pattern:

$$\min_{x_1,x_2} f_1(x_1) + f_2(x_2)$$
$$s.t.\ A_1x_1 + A_2x_2 = b$$

Augmented Lagrangian:

$$L_\mu(x_1, x_2, v) = f_1(x_1) + f_2(x_2) + v^T(A_1x_1 + A_2x_2 - b) + \frac{\mu}{2}\|A_1x_1 + A_2x_2 - b\|_2^2$$

ADMM framework:

$$x_1^{k+1} = \arg\min_{x_1} L_\mu(x_1, x_2^k, v^k)$$
$$x_2^{k+1} = \arg\min_{x_2} L_\mu(x_1^{k+1}, x_2, v^k)$$
$$v^{k+1} = v^k + \mu(A_1x_1^{k+1} + A_2x_2^{k+1} - b)$$

Remark: decentralized optimization + method of multipliers

Fused Lasso Signal Approximator Problem -FLSA

Model:

$$\min_{\beta} \frac{1}{2}\|y-\beta\|_2^2 + \lambda_1\|\beta\|_1 + \lambda_2\|L\beta\|_1$$

Practical Intuition: Value sparsity high, favor small change in between signals.
Introducing auxiliary variable:

$$\begin{aligned} \min_{\beta} \ & \frac{1}{2}\|y-\beta\|_2^2 + \lambda_1\|a\|_1 + \lambda_2\|b\|_1 \\ s.t. \ a &= \beta \\ b &= L\beta \end{aligned}$$

What we propose is a generalized model of FLSA, and how to solve it with ADMM

Generalized FLSA

piece-wise constant signal

$$\min_{\beta} \frac{1}{2}\|y-\beta\|_2^2 + \lambda_1 \sum_{i=1}^{K} p_i \|\beta - c_i\|_1 + \lambda_2 \|L\beta\|$$

where c_i represent the set of constant signal, and p_i controls the weight among them.

The equivalent equality constraint problem is:

$$\min_{\beta,\{a_i\},b} = \frac{1}{2}\|y-\beta\|_2^2 + \lambda_1 \sum_{i=1}^{K} p_i \|a_i\|_1 + \lambda_2 \|b\|_1$$

$$s.t. a_1 = \beta - c_1$$

$$\vdots$$

$$a_K = \beta - c_K$$

$$b = L\beta$$

ADMM for Solving Generalized FLSA

The Lagrangian of generalized FLSA is:

$$L(\beta, \{a_i\}, b, \{u_i\}, v) = \frac{1}{2}\|y - \beta\|_2^2 + \lambda_1 \sum_{i=1}^{K} p_i \|a_i\|_1 + \lambda_2 \|b\|_1$$
$$+ \sum_{i=1}^{K} \langle u_i, \beta - a_i - c_i \rangle + \langle v, L\beta - b \rangle$$
$$+ \frac{\mu_1}{2} \sum_{i=1}^{K} \|\beta - a_i - c_i\|_2^2 + \frac{\mu_2}{2} \|L\beta - b\|_2^2$$

ADMM for Solving Generalized FLSA

Plug-in ADMM algorithm

$$\beta^{k+1} = \arg\min_{\beta} L_{\mu_1,\mu_2}(\beta, \{a_i^k\}, b^k, \{u_i^k\}, v^k)$$

$$a_i^{k+1} = \arg\min_{a_i} L_{\mu_1,\mu_2}(\beta^{k+1}, a_i, b^k, u^k, v^k), i = 1, 2, .., K$$

$$b^{k+1} = \arg\min_{b} L_{\mu_1,\mu_2}(\beta^{k+1}, \{a_i^{k+1}\}, b, u^k, v^k)$$

$$u_i^{k+1} = u_i^k + \mu_1(\beta^{k+1} - c_i - a_i^{k+1}), i = 1, 2, .., K$$

$$v^{k+1} = v^k + \mu_2(L\beta^{k+1} - b^{k+1})$$

ADMM for Solving Generalized FLSA

By taking (sub)gradient and set them to or contain 0, we get:

$$(K\mu_1 + 1)I + \mu_2 L^T L)\beta^{k+1} = y + \sum_{i=1}^{K}(\mu_1(a_i^k + c_i) - u_i^k) + L^T(\mu_2 b^k - v^k)$$

$$a_i^{k+1} = \Gamma_{\frac{\lambda_1 p_i}{\mu_1}}\left(\beta^{k+1} - c_i + \frac{u_i^k}{\mu_1}\right), i = 1, 2, ...K$$

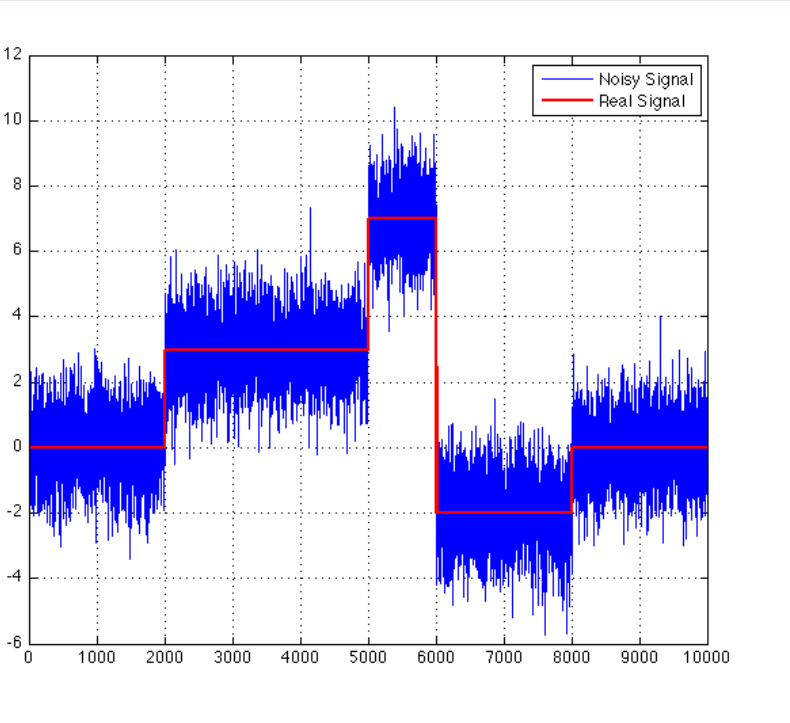
$$b^{k+1} = \Gamma_{\frac{\lambda_2}{\mu_2}}\left(L\beta^{k+1} + \frac{v^k}{\mu_2}\right)$$

$$u_i^{k+1} = u_i^k + \mu_1(\beta^{k+1} - a_i^{k+1} - c_i), i = 1, 2, ..., K$$

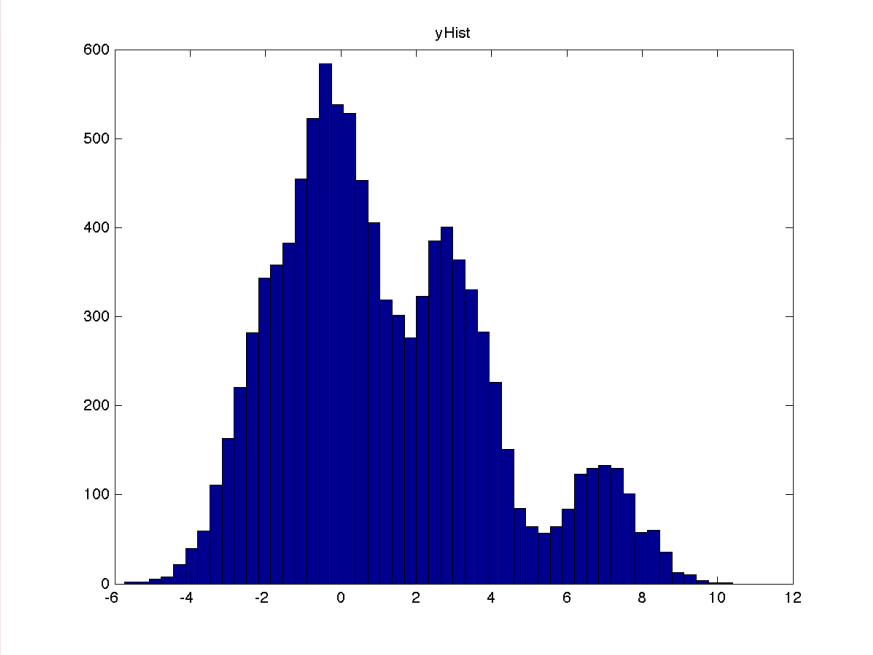
$$v^{k+1} = v_k + \mu_2(L\beta^{k+1} - b^{k+1})$$

Simulation

A 10-thousand dimensional sequence. It's piece-wise constant at value 0, 3, 7, -2, 0, with a Gaussian noise ($\sigma = 1$) added to it



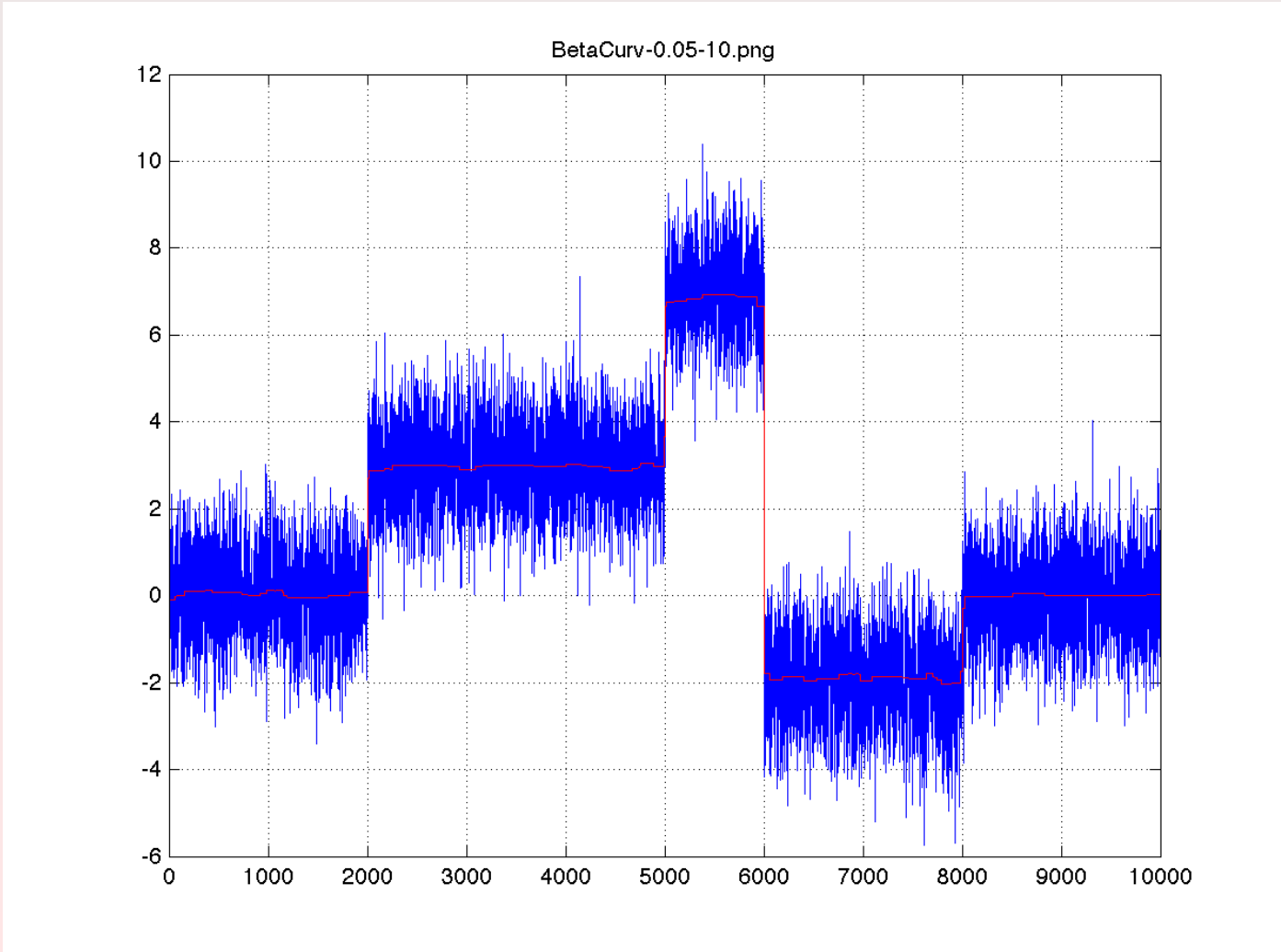
noisy signal and real pattern



histogram of noisy signal

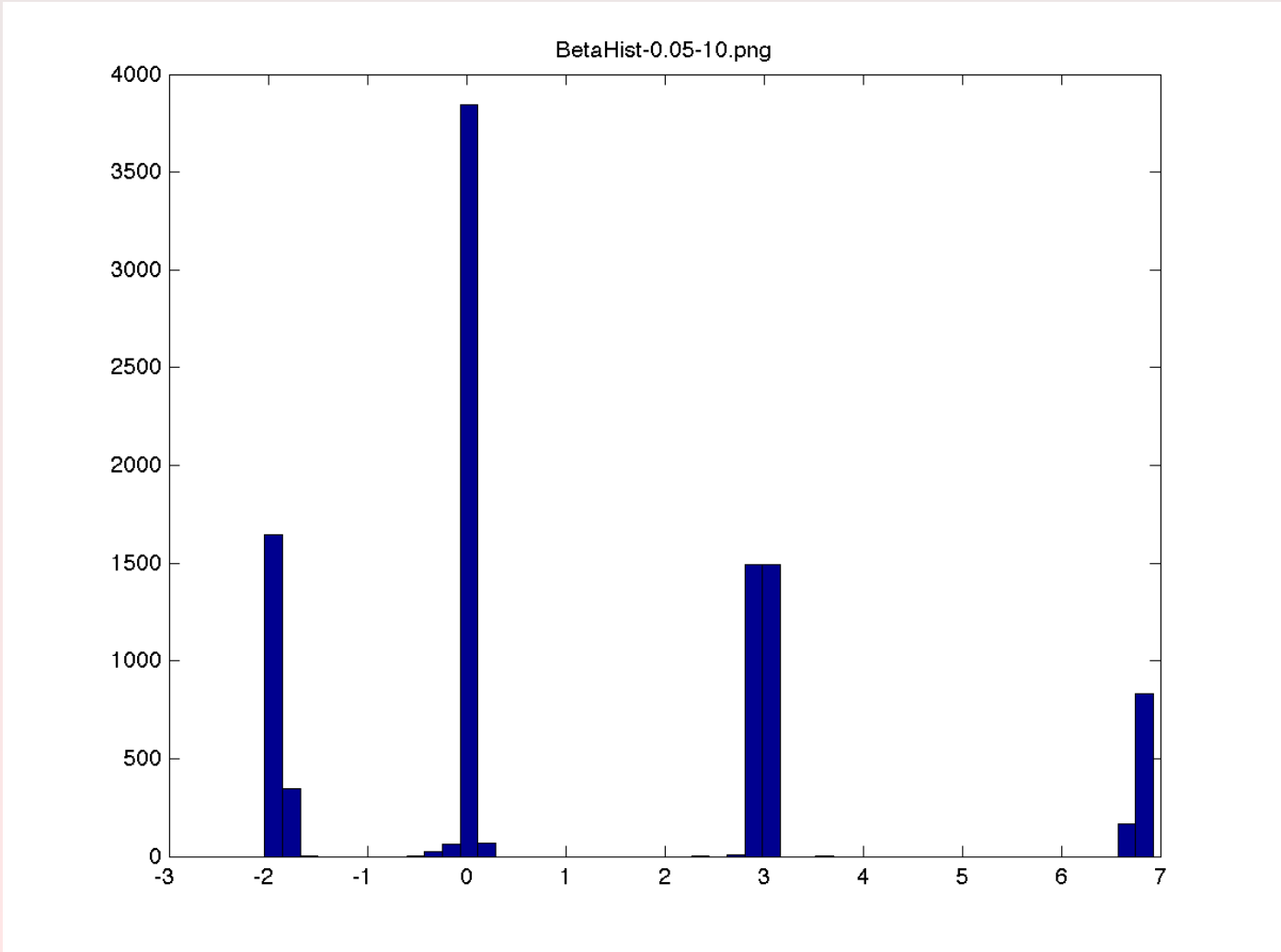
Simulation

We tune regularization parameter λ_1, λ_2 on a 2-dimensional grid of $\{0.05, 0.1, 0.5, 1, 5, 10\}$. A good result happens at $\{\lambda_1 = 0.05, \lambda_2 = 10\}$

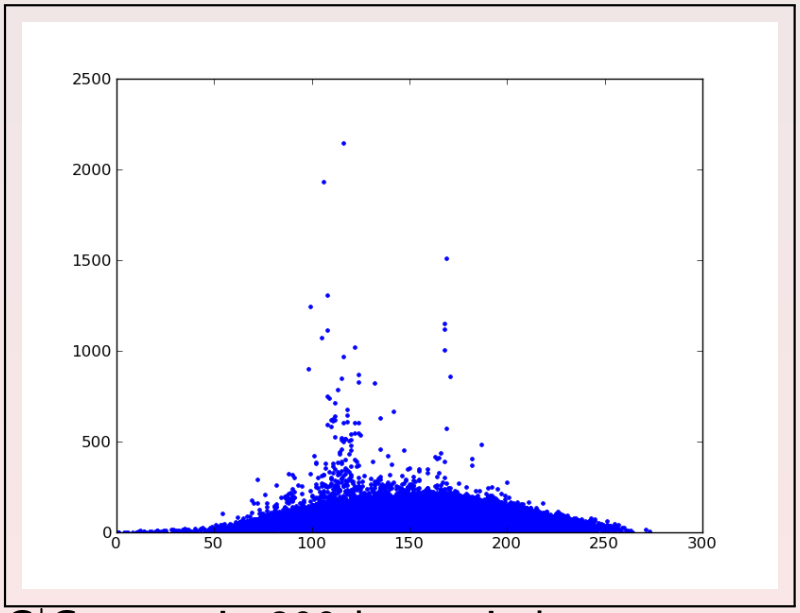


Simulation

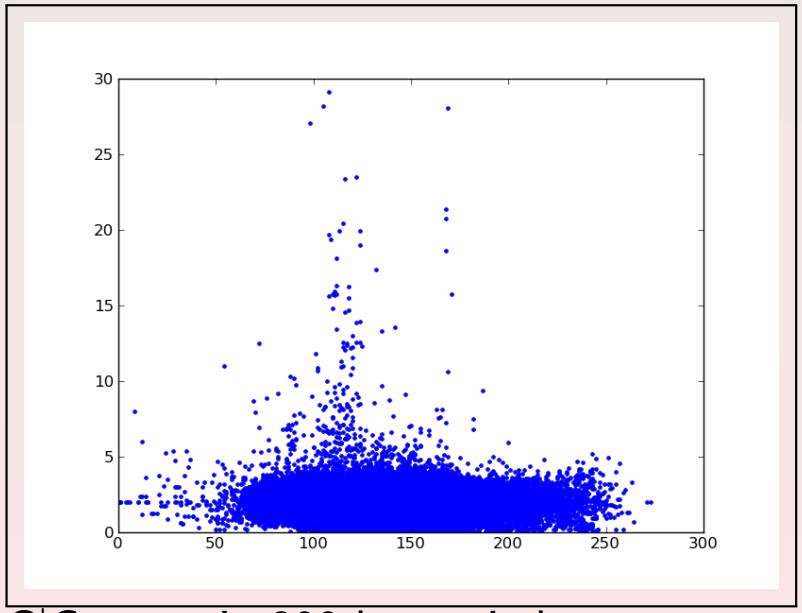
We tune regularization parameter λ_1, λ_2 on a 2-dimensional grid of $\{0.05, 0.1, 0.5, 1, 5, 10\}$. A good result happens at $\{\lambda_1 = 0.05, \lambda_2 = 10\}$



- For our real data, we took low-coverage data for chromosome 20 from a Yoruban female (NA18505)
- We corrected for increased reads in GC-rich areas by normalizing by the average reads in similar GC sites

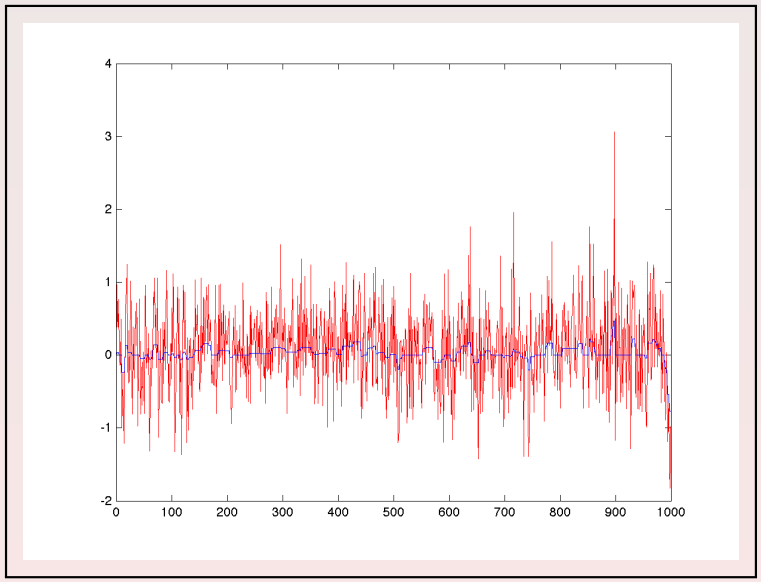


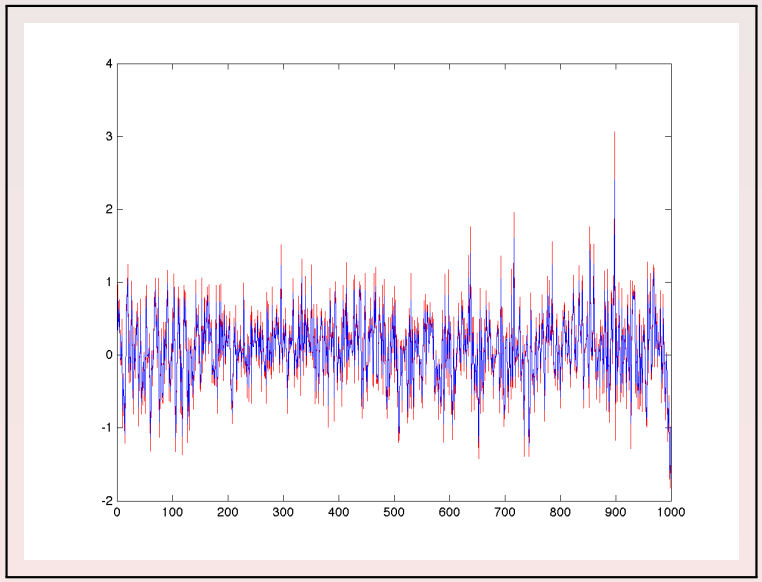
G|C count in 300 base window



G|C count in 300 base window

- Then we binned the raw reads, varying the bin width from 50bp to 5000bp.
- For each bin width, we explored the effect of varying the two parameters, λ_1 and λ_2





- We haven't yet completed validation on known copy number variants– this will require some postprocessing.

Conclusions

- Fused Lasso method looks promising for detecting the underlying copy number from Short Read data.
- The results are highly dependent on the choice of λ parameters. We're not sure what the best method for validation would be for us.
- It might also be useful to add additional constraints, attracting the β values towards integer values.
- Finally, in this framework, we could easily pool samples to detect common variants in a population or add additional data sources to improve power.