# Fine-mapping regulatory variants with RASQUAL and ATAC-seq

Natsuhiko Kumasaka[†], Andrew Knights[†], Daniel Gaffney [*†]

## 1. Introduction

Association mapping of cellular traits is a powerful approach for understanding the function of genetic variation. Cellular traits that can be quantified by next generation sequencing (NGS) are particularly useful for association analysis because they provide highly quantitative information about the phenotype of interest and can easily be scaled genome-wide. Population scale studies using NGS-based cell phenotypes such as RNA-seq, ChIP-seq and DNaseI-seq have revealed an abundance QTLs for gene expression and isoform abundance [1–4], chromatin accessibility [5], histone modification, transcription factor binding (TF) [6–9] and DNA methylation [10], providing precise molecular information on the functions of human genetic variation at high resolution. However the effect sizes of many common variants are modest meaning that association analysis typically requires large sample sizes, which can be problematic when assays are labour intensive or cellular material is difficult to obtain. Furthermore, even well-powered studies can struggle to accurately fine-map causal variants.

Here we describe a novel statistical method, RASQUAL (Robust Allele Specific Quantitation and quality controL), that integrates population level changes, AS signals and technical biases on NGS-based cell phenotypes into a single, probabilistic framework for association mapping. RASQUAL can be applied to existing NGS data sets without requiring data filtering, masking or the creation of personalised reference genomes. When applied to RNA-seq, ChIP-seq and DNaseI-seq data sets, RASQUAL significantly outperformed existing methods, both in its ability to detect QTLs and to fine-map putatively causal variants. We used RASQUAL to generate the first map of chromatin accessibility QTLs in a European population using ATAC-seq [12]). Despite a modest sample size of 24 individuals, RASQUAL detected over 2700 independent chromatin accessibility QTLs (FDR 10%) providing a rich resource for the functional interpretation of human noncoding variation.

## 2. Statistical model

RASQUAL models each sequenced feature, such as ChIP-seq peak or the union of exons over an entire transcript, and considers all genotyped variants within a given distance of the feature (the *cis*-window). For simplicity, RASQUAL assumes a single *cis*-regulatory variant (rSNP) at each feature. Let $Y_i$ be the total fragment count at the feature and $(Y_{il}^R, Y_{il}^A)$ be the AS fragment counts at each feature variant $l$, where $R$ for reference and $A$ for alternative alleles respectively, for individual $i$ ($i = 1, \ldots, N$). The model contains two components: (i) population signals are captured by regressing the total fragment count $Y_i$ onto the number of alternative alleles $G_i$ ($G_i = 0, 1, 2$) at rSNP, assuming read counts follow a negative binomial distribution ($p_{\text{NB}}$) and (ii) AS signals are modelled assuming the alternative fragment count, $Y_{il}^A$, at a fSNP $l$ given the total reads overlapping that fSNP, $Y_{il}$, follows a beta binomial distribution ($p_{\text{BB}}$). The model components are connected by a single cis-regulatory effect parameter $\pi$ such that the expected total count is proportional to $\{2(1-\pi), 1, 2\pi\}$ for $G_i = 0, 1, 2$ and the expected allelic ratio in an individual heterozygous for the putative causal SNP becomes $\{\pi, 1-\pi\}$.

Although the parametrization of this type has been previously proposed [6, 11], RASQUAL improves over the methods via (i) robust estimation of read count overdispersion for between individual feature counts and within individual AS counts, (ii) substantial refinement of genome imputation by modeling uncertainty in genotype and haplotype configurations between rSNP and each fSNP and (iii) explicit modelling of a broad range of technical biases in AS data by using information from all individuals. A key novelty of our approach is the use of read counts at both heterozygous and homozygous fSNPs to significantly improve genotype error correction and the estimation of bias parameters. The likelihood is then given by

$$\begin{aligned}\mathcal{L}(\pi, \delta, \phi, \lambda, \theta) &\propto \prod_{\substack{i=1 \\ \text{sample}}}^{N} \sum_{G_i} \underbrace{p(G_i) p_{\text{NB}}(Y_i | G_i; \pi, \lambda, \theta)}_{\text{population signal}} \\ &\quad \times \prod_{\substack{l=1 \\ \text{fSNP}}}^{L} \sum_{D_{il}} \underbrace{p(D_{il}|G_i) p_{\text{BB}}(Y_{il}^A | Y_{il}, D_{il}; \pi, \delta, \phi, \theta)}_{\text{AS signal}},\end{aligned}$$

where $D_{il}$ denotes the diplotype configuration in individual $i$ between the putatively causal variant and the fSNP $l$, $p(G_i)$ and $p(D_{il}|G_i)$ denote prior probabilities of genotype and diplotype configuration (obtained from SNP phasing and imputation). In addition to the *cis*-regulatory effect ($\pi$), total read count depends on $\lambda$, a scaling parameter for absolute mean of the negative binomial distribution. The allelic ratio depends upon $\delta$, the probability that an individual read maps to an incorrect location in genome and $\phi$, the reference mapping bias (where $\phi$=0.5 corresponds to no reference bias). Overdispersion in both $Y_i$ and $Y_{il}^A$ is captured by a single shared parameter $\theta$.

Parameter estimation and genotypes are iteratively updated during model fitting by an expectation-maximisation (EM) algorithm [13] to arrive at the final QTL call for each sequenced feature. For statistical hy-

[*]Corresponding author: dg13@sanger.ac.uk
[†]Wellcome Trust Sanger Institute

pothesis testing of QTL, all five parameters for each SNP-feature combination in the *cis*-regulatory window are estimated independently to get the maximum likelihood under alternative hypotheses. Under the null hypothesis, all parameters except $\pi$ are estimated for each feature independently, while $\pi$ is set to 0.5 and we use a likelihood ratio test to compare the null and alternative hypotheses for each SNP-feature combination using the $\chi^2$ distribution with one degree of freedom (for $\pi$). We do not introduce any common parameters across individuals or features estimated *a priori*, but instead introduced prior distributions for all the parameters to increase the stability and usability of RASQUAL.

### 3. Mapping chromatin accessibility QTLs with RASQUAL and ATAC-seq

We next applied RASQUAL to address a specific biological problem: mapping chromatin accessibility QTLs (caQTLs) in a small sample. We generated genome-wide chromatin accessibility landscapes in 24 LCLs from the 1000 Genomes GBR population using ATAC-seq [12] (see Online Methods and Supplementary Methods for details). Despite the modest sample size RASQUAL detected 2,706 caQTLs at FDR=10% using a permutation test (see Online Methods). Lead SNPs detected by RASQUAL were very highly enriched within the ATAC peak itself (842 peaks; OR=47, $P < 10^{-16}$) (Figurea), with a smaller number in perfect LD with one or more SNPs within the peak (130 in perfect LD with a single fSNP, and 35 with 2 fSNPs). In the set of 1007 lead SNPs within a peak or in perfect LD with an fSNP, the majority (692) overlapped a known transcription factor binding motif that was disrupted by one of the SNP alleles (Figure b). An example caQTL where a putatively causal variant (r2886870) is located within both the ATAC peak and an NF$\kappa$B1 motif is shown in Figure c. This SNP is predicted to produce a large (85%) change in binding affinity, with the predicted change in binding corresponding with a change both in ATAC-seq peak height and in AI at flanking heterozygous fSNPs in individuals that are heterozygous at r2886870.

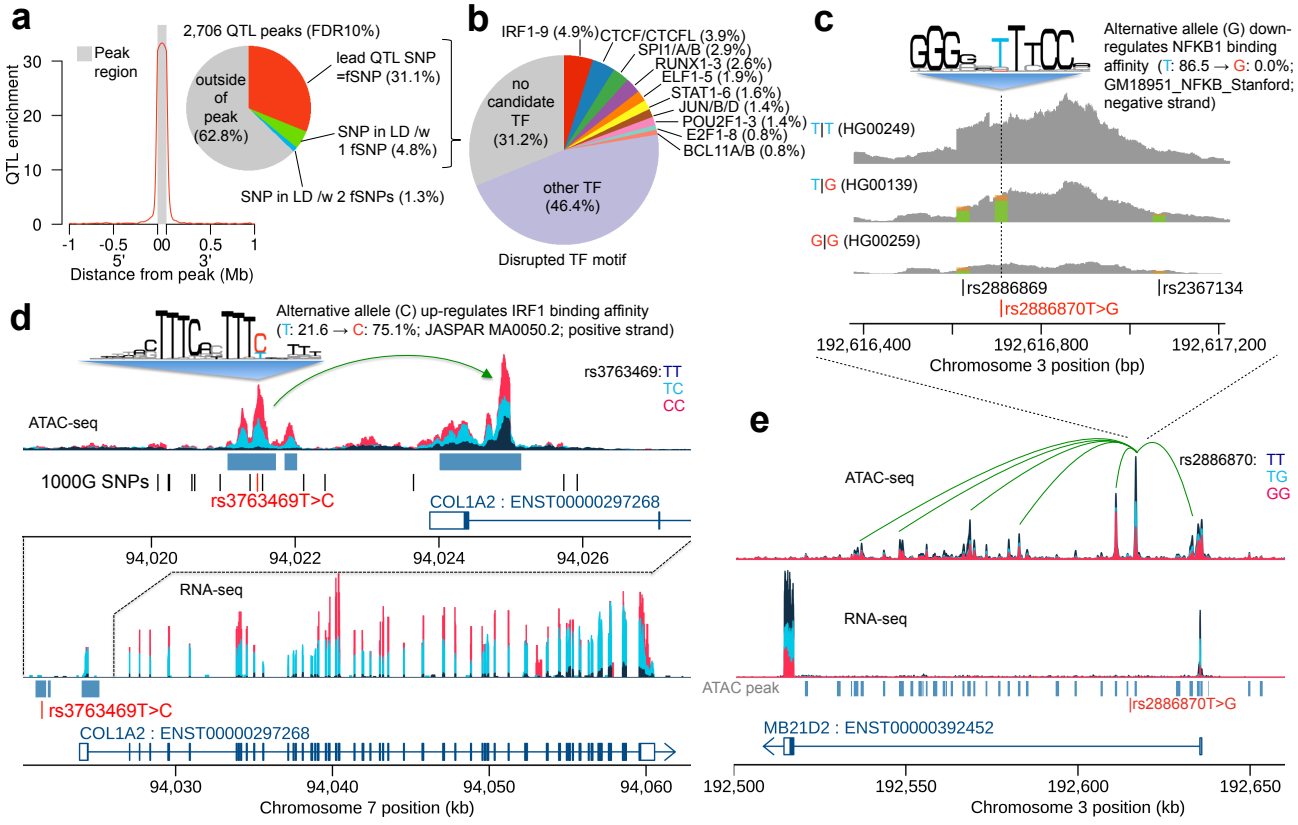


**Fig.** ATAC-QTL mapping with RASQUAL. (a) Location of ATAC-QTL lead SNP, relative to peak boundaries, averaged across all 1,798 FDR 10% significant associations detected; inset shows proportion of lead SNPs located inside, outside and in perfect LD ($r^2 > 0.99$) with a SNP inside the ATAC peak. (b) Proportion of lead SNPs, located inside the ATAC peak that overlapped an identifiable transcription factor binding motif. (c) An example of an NF$\kappa$B1 motif-disrupting QTL (d) Example of a "multi-peak" ATAC-QTL (rs3763469) that perturbs a putative enhancer-promoter interaction in the COL1A2, also driving variation in gene expression. (e) Example of a "multi-peak" QTL: the same genetic variant illustrated in panel c (rs2886870), drives associations at 6 peaks in the intron and promoter the MB21D2 gene.

Further analysis of our detected caQTLs revealed 154 "multipeak" QTLs, where a SNP was associated with variation in chromatin accessibility across multiple independent peak regions. In some cases, these long-range associations appeared to result from enhancer-promoter interactions that are perturbed by a genetic variant. For example, rs3763469 is the lead caQTL SNP for a region of open chromatin located approximately 2.5kb upstream of the promoter of the COL1A2 gene (Figure d) with the alternative allele predicted to increase binding affinity of the transcription factor IRF1. However, we observed that this SNP is also a QTL for the adjacent ATAC peak located over the promoter region of COL1A2 gene, for which no other common SNPs were annotated in the 1000 Genomes database. In other striking examples, we observed genetic associations spanning a large number of additional peaks spread over many tens of kilobases (Figure d). For example the lead caQTL SNP in Figure c also appeared as the lead SNP or SNP in perfect LD with the lead SNP at 5 other peaks in the intron and promoter regions of MB21D2 gene.

## 4. Discussion

We have developed a novel statistical model, RASQUAL, for mapping associations between genotype and NGS-based cellular phenotypes. A major difference between RASQUAL and the other methods we have tested is that RASQUAL handles bias and detection of genetic signals in a single statistical framework, using information from all individuals in the data set and without removing data. In contrast, other methods treat bias in NGS data as a data quality control (QC) issue and either ignore it or rely on a series of read realignment and data filtering steps to remove problematic regions *a priori*. Better handling of a range of biases in NGS data is likely to explain much of the differences in performance we observed. Furthermore, data filtering and QC steps as an alternative to handling bias may introduce a number of other issues. First, data QC involving read filtering may also inadvertently remove substantial signal from the data. A second issue is that it is often difficult to set sensible thresholds for data QC. In addition to boosting model sensitivity and specificity, we believe that minimising the amount of data pre-processing significantly improves the usability of RASQUAL. Users of RASQUAL are not required set arbitrary thresholds for data QC, nor are they required to perform disk and CPU-intensive read remapping or simulations to account for biases NGS data, but can instead run the model on existing data as is. In addition, RASQUAL output contains an informative set of parameters that can highlight genomic regions with problematic AS signals enabling more informed downstream interpretation of analysis results.

We also generated a novel ATAC-seq data set in LCLs from European individuals and illustrated how RASQUAL can be used to extract meaningful genetic signals from data sets of a modest size. Our analysis of ATAC-seq data demonstrates how genetic variation can be leveraged to connect distal regulatory elements with gene promoters or with other regulatory elements. A strength of this approach, compared with experimental techniques such as Hi-C or CHiAPET, is that these interactions are linked to specific genetic changes enabling characterisation of causal relationships between regulatory elements and their target genes. We expect that genetic analysis of long-range regulatory interactions will be a powerful complement to standard experimental techniques in future studies.

A potential limitation of RASQUAL is its reliance on high quality genotype imputation and phasing to compute genotype and diplotype likelihoods. Poor quality imputation or phasing information is likely to affect the RASQUAL's ability to detect QTLs, particularly in cases where the distance between the true rSNP and fSNP is large, due to the increased probability of haplotype switching errors. However our analysis illustrates that these problems can be readily overcome using imputation into the large, high quality panels now available from the 1000 Genomes and UK10K projects.

Finally, we believe that our results also highlight how RASQUAL's performance with modest sample sizes will enable researchers to collect and analyse multiple complementary NGS data sets, rather than focusing resources on maximising sample sizes for an individual phenotype. Combined with RASQUAL's improved ability to localise causal variants we suggest that a major future application of our model will be the fine-mapping of causal regulatory variants to understand the molecular mechanisms underlying phenotypic variation.

## References


[1] Pickrell JK, Marioni JC, Pai AA, Degner JF, Engelhardt BE, et al. (2010) Nature 464: 768-72.

[2] Montgomery SB, Sammeth M, Gutierrez-Arcelus M, Lach RP, Ingle C, et al. (2010) Nature 464: 773-7.

[3] Lappalainen T, Sammeth M, Friedländer MR, 't Hoen PA, Monlong J, et al. (2013) Nature 501: 506-11.

[4] Battle A, Mostafavi S, Zhu X, Potash JB, Weissman MM, et al. (2014) Genome Res 24: 14-24.

[5] Degner JF, Pai AA, Pique-Regi R, Veyrieras JB, Gaffney DJ, et al. (2012) Nature 482: 390-4.

[6] McVicker G, van de Geijn B, Degner JF, Cain CE, Banovich NE, et al. (2013) Science 342: 747-9.

[7] Kilpinen H, Waszak SM, Gschwind AR, Raghav SK, Witwicki RM, et al. (2013) Science 342: 744-7.

[8] Kasowski M, Kyriazopoulou-Panagiotopoulou S, Grubert F, Zaugg JB, Kundaje A, et al. (2013) Science 342: 750-2.

[9] Ding Z, Ni Y, Timmer SW, Lee BK, Battenhouse A, et al. (2014) PLoS Genet : in press.

[10] Banovich NE, Lan X, McVicker G, van de Geijn B, Degner JF, et al. (2014) PLoS Genet 10: e1004663.

[11] Sun W (2012) Biometrics 68: 1-11.

[12] Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ (2013) Nat Methods 10: 1213-8.

[13] Dempster AP, Laird NM, Rubin DB (1977) Journal of the Royal Statistical Society, Series B 39: 18.