

# Supplement to “Modelling haplotypes with respect to reference cohort variation graphs”

Yohei Rosen, Jordan Eizenga and Benedict Paten  
UC Santa Cruz Genomics Institute,  
University of California Santa Cruz, Santa Cruz, CA 95064, USA

March 22, 2017

## 1 Appendix A: An $\mathcal{O}(n \cdot m)$ implementation of the rectangular decomposition construction

Suppose that we wish to find subhaplotypes embedded in the graph which are consistent with a query sequence  $h$  of nodes. In brief, in the gPBWT, indexing information for haplotypes is stored in such a manner that this can be achieved by calling a function `STARTSEARCHATNODE(Node)` on the first node of  $h$ , which returns a search interval  $gPBWTInt$  of a form analogous to the search interval of a Burrows–Wheeler Transform based index of sequences. This search interval is extended by calling an operation `EXTEND(gPBWTInt, Node)` to extend this search with each additional node in  $h$ . Finally, this search interval can be converted into a count of matching subhaplotypes using a function `COUNT(gPBWTInt)`. It is shown by Novak *et al.* (2016) that `STARTSEARCHATNODE`, `EXTEND` and `COUNT` all admit  $\mathcal{O}(1)$  implementations.

It is evident that this search process yields a function `COUNTHAPLOTYPEMATCHES(h)` which is  $\mathcal{O}(n)$  in the length  $|h|$  of  $h$  in nodes. Let  $h_1 h_2 h_3 \dots h_{|h|-1} h_{|h|}$  denote the node sequence of  $h$ . Using `COUNTHAPLOTYPEMATCHES` we can identify the set  $A$  of nodes in  $h$  such that either  $J_a^{a-1} \neq J_{a-1}^{a-1}$  or  $I_a^a \neq 0$  in  $\mathcal{O}(n)$  independent length-2 subhaplotype count queries:

---

**Algorithm 1** Identifying  $A$ , the set of “relevant” nodes

---

```
1: function BUILDA( $h, B[]$ )
2:    $A \leftarrow [1]$ 
3:    $ht_{prev} \leftarrow |B[h_1]|$ 
4:   for  $i = 2, \dots, |h|$  do
5:      $ht_{new} \leftarrow |B[h_i]|$ 
6:      $J_i^{i-1} \leftarrow \text{COUNTHAPLOTYPEMATCHES}(h_{i-1} h_i)$ 
7:     if  $J_i^{i-1} < ht_{new}$  or  $ht_{prev} > ht_{new}$  then
8:       APPEND( $A, i$ )
9:    $ht_{prev} \leftarrow ht_{new}$ 
```

---

Given that we have constructed  $A$ , we can determine the rest of the rectangular decomposition and all of the  $J$ -values according to the following algorithm:

---

**Algorithm 2** Building the  $J$ 's and  $A_{curr}$ 's

---

```
1: function BUILDJS( $h, B[]$ )
2:    $J_1^1 \leftarrow |B[h_1]|$ 
3:    $A_{curr}^1 \leftarrow 1$ 
4:   for  $i \in A$  do
5:      $A_{curr}^i \leftarrow []$ 
6:     if  $|B[h_i]| > J_i^i$  then
7:       APPEND( $A_{curr}^i, i$ )
8:        $S_i \leftarrow \text{STARTSEARCH}(h_i)$ 
9:       for  $j \in A_{curr}^{i-1}$  do
10:         $S_j \leftarrow \text{EXTEND}(S_j, h_i)$ 
11:        if COUNT( $S_j$ )  $\neq 0$  then
12:           $J_i^j \leftarrow \text{COUNT}(S_j)$ 
13:          APPEND( $A_{curr}^i, j$ )
14:       else
15:         break
```

---

## 2 Appendix B: Arithmetic for derivation of Equation 6

Here we lay out the arithmetic to derive Equation (6) of section 2.3, which is used in our iterative computation of likelihood of a haplotype  $h$  with respect to a population reference cohort  $H$  embedded in a variation graph  $G$ . The reasoning is straightforward but involves many subcases which require care.

### 2.1 Notation

**Definition 1.** A haplotype is a sequence of nodes  $n_1 \rightarrow \dots \rightarrow n_{|h|}$  in a variation graph. The base sequence of a haplotype is the sequence of DNA bases spelled by its node labels. A haplotype subinterval is a contiguous subsequence of a haplotype. A haplotype base sequence subinterval is analogously defined. Denote by  $|h|$  the length of a haplotype base sequence in base pairs.

**Definition 2.** Haplotypes  $h, h'$  are consistent if  $|h| = |h'|$  and  $n_i = n'_i \forall i$ .

**Definition 3.** A mosaic of haplotypes  $x$  consistent with  $h$  is a vector  $\langle x_{(i)} \rangle$  of subintervals of base sequences of haplotypes in  $H$  whose concatenation is consistent with the base sequence of  $h$ . The recombination count  $R(x)$  is one less than the number of elements in  $\langle x_{(i)} \rangle$ . NB: defining these in terms of base sequence rather than node subintervals permits recombination within nodes. Recall Figure 2 from the main text.

**Definition 4.**  $\chi(h)$  is the set of all mosaics  $x$  consistent with  $h$ .  $\chi(h)^R$  is the subset with  $R(x) = R$ .  $\chi(h)[,g]$  is the subset whose final subinterval is a subinterval of  $g$ .  $\chi(h)[g,]$  is that with initial subinterval a subinterval of  $g$ .  $|\chi(h)|$  is the number of elements in  $\chi(h)$ .

### 2.2 Arithmetic shortcuts

**Lemma 1.** *There exists a partition of  $h$  into subintervals  $h_1, h_2, \dots, h_n$  such that if a haplotype  $g \in H$  has a subinterval consistent with a subinterval of  $h_i$  then it has a subinterval consistent with all of  $h_i$ .*

*Proof.* It is straightforward to verify that the intervals between successive nodes in the set  $A$  described in the main text produce such a partition of  $h$ .  $\square$

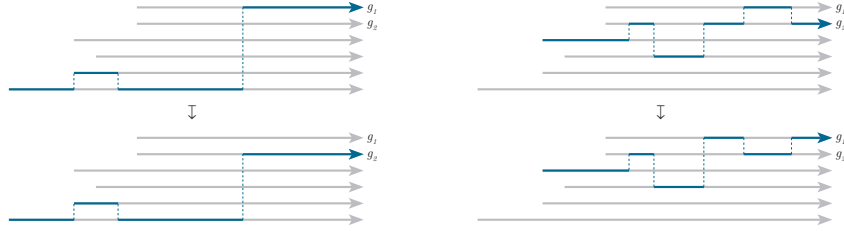
This is important because we will show that it is simple to calculate  $|\chi(h_i)|$  within any interval with this property.

The following is a more notationally precise statement of Lemma 1 from the main text:

**Lemma 2.** *For any  $b \in A, a \leq b$ , given that  $f$  and  $g$  are members of the same equivalence class  $S_b^a$  of haplotypes, the haplotype mosaics  $\chi^R(h_{[0,b)}[f]$  and  $\chi^R(h_{[0,b)}[g]$  consistent with the subinterval  $h_{[0,b)}$  and ending with subintervals of  $f$  and  $g$  are in bijective correspondence.*

*Proof.* We assume that  $g \neq f$  else this is trivial. Consider any mosaic  $x$  in  $\chi^R(h_{[0,b)}[f]$ . Given  $x = \langle x_1, x_2, \dots, x_{R+1} \rangle$ , let  $j = \max\{i \in 1, \dots, R \text{ such that } x_i \text{ is not a subinterval of } g \text{ or } f\}$ . We will construct a mosaic  $y = \langle y_1, y_2, \dots, y_{R+1} \rangle$  such that for all  $i \leq j$ ,  $y_i = x_i$ , and for all  $i > j$ ,  $y_i$  is the subinterval of the same length as  $x_i$  but derived from the opposite haplotype of the pair  $f, g$ .

The concatenation  $y_1 y_2 \dots y_{R+1}$  is consistent with  $h_{[0,b)}$  since given that both  $f, g \in S_b^a$ , the first node of  $y_{j+1}$  must be at or after  $a$ . Therefore clearly  $y_i \in \chi^R(h_{[0,b)}[g]$  since its final subinterval corresponds to  $g$ . The inherent invertibility of this transformation proves that it is a bijection.  $\square$



**Figure 1:** Visual proof of the above lemma by explicit construction of the bijection involved

**Lemma 3.** *Suppose that  $h_i$  is a subinterval of  $h$  such that if a haplotype  $g \in H$  has a subinterval consistent with a subinterval of  $h_i$  then it has a subinterval consistent with all of  $h_i$ . Then suppose that  $f_1, f_2, g \in H$ , and all have subintervals consistent with all of  $h_i$ . Then for all  $R < |h_i|$  there is a bijection between  $\chi(h_i)[f_1, g]$  and  $\chi(h_i)[f_2, g]$ .*

*Proof.* The proof imitates that of the previous lemma.  $\square$

### 2.3 The case of a single simple interval $h_i$

Suppose that  $h_i$  is an interval of the form in Lemma 2,  $\ell$  base pairs in length, and has subintervals of  $ht$  haplotypes of  $H$  consistent with it. Consider  $f, g \in H$  such that both have subintervals consistent with  $h_i$ . Suppose that we wish to calculate, for some  $R < \ell$ , the number  $|\chi^R(h)[g]|$  of mosaics consistent with  $h_i$  having  $R$  recombinations and ending with haplotype  $g$ . To calculate  $|\chi(h_i)|$  within an interval of the form above, we need only calculate

1.  $|\chi^R(h_i)[g, g]|$ , the number of paths both beginning on and ending on  $g$  and
2.  $|\chi^R(h_i)[f, g]|$ , the number of paths beginning on  $f \neq g$  and ending on  $g$ , which, by lemma 4, is the same for all such  $f$ .

Consider  $R = \ell - 1$ . It is clear that

$$\sum_j |\chi^R(h_i)[j, g]| = (ht - 1)^R \quad (1)$$

Lemma 4 tells us that all haplotypes  $f \neq g$  are equivalent for the purposes of enumeration, therefore we write  $\neg g$  to denote any arbitrary representative  $f \neq g$ . There are  $ht - 1$  such haplotypes.

$$|\chi^R(h_i)[g, g]| + (ht - 1)|\chi^R(h_i)[\neg g, g]| = (ht - 1)^R \quad (2)$$

We begin by calculating  $|\chi^R(h_i)[\neg g, g]|$ . Consider first  $\ell = R + 1 = 1$ , for which, given the lack of possible recombinations,  $|\chi^R(h_i)[\neg g, g]| = 0$ . For  $\ell = R + 1 = 2$ , any  $x \in \chi^R(h_i)[\neg g, g]$  must at its second node visit a haplotype which is neither  $g$  nor the  $\neg g$  under consideration, therefore  $|\chi^R(h_i)[\neg g, g]| = (ht - 2)$ . Suppose now that, for arbitrary  $\ell = R$ , we know  $|\chi^R(h_i)[\neg g, g]|$ . Then, counting the  $(ht - 1)$  possible haplotypes before finally recombining to  $g$  shows us that

$$|\chi^{R+1}(h_i)[g, g]| = (ht - 1)|\chi^R(h_i)[\neg g, g]| \quad (3)$$

By (2), we know that

$$|\chi^{R+1}(h_i)[\neg g, g]| = \frac{(ht - 1)^R - |\chi^{R+1}(h_i)[g, g]|}{(ht - 1)}$$

Which by (3) implies

$$\begin{aligned} |\chi^{R+1}(h_i)[\neg g, g]| &= \frac{(ht - 1)^R - (ht - 1)|\chi^R(h_i)[\neg g, g]|}{(ht - 1)} \\ \implies |\chi^{R+1}(h_i)[\neg g, g]| &= (ht - 1)^{R-1} - |\chi^R(h_i)[\neg g, g]| \end{aligned} \quad (4)$$

Using (4) as the induction step with base case  $\ell = R + 1 = 2$  we find that  $\forall \ell = R + 1 \geq 2$

$$|\chi^R(h_i)[\neg g, g]| = \frac{(ht - 1)^{R-1} + (-1)^R}{ht}$$

We now relax the restriction that  $R = \ell$ . For given  $R < \ell$  each subset of nodes at which recombinations happen will define an additional set of possible recombinations. Counting all possible such subsets

$$|\chi^R(h_i)[\neg g, g]| = \binom{\ell - 1}{R} \frac{(ht - 1)^{R-1} + (-1)^R}{ht}$$

and

$$|\chi^R(h_i)[g, g]| = (ht - 1)|\chi^R(h_i)[\neg g, g]|$$

## 2.4 Extending a computation for a prefix by a simple subinterval $h_i$

To extend our ability to calculate  $|\chi(h)|$  beyond the single interval  $h_i$ , suppose we have a partition  $\{h_1, h_2, \dots, h_n\}$  of  $h$  into subintervals of the form in Lemma 2. Let  $b \in A$  such that  $b$  is a node on the boundary of such an interval, let  $h_{[0, b-1]}$  be the prefix of  $h$  formed by concatenation of the subintervals preceding node  $b$ , and let  $h_{[b-1, b]}$  be the subinterval beginning with node  $b$ .

Suppose now that we have calculated each  $|\chi^R(h_{[0, b-1]}), f]|$  and now wish to calculate these values up to  $b$ , the node in  $A$  succeeding  $b - 1$ . By Lemma 2, the intervening sequence  $h_{[b-1, b]}$  is of the form for which we have just calculated  $|\chi^R(h)[g, g]|$  and  $|\chi^R(h)[\neg g, g]|$ . We divide this into cases.

*Case 1:* Suppose that  $f$  has no subinterval consistent with  $h_{[b, b+1]}$ , that is,  $f \in S_{b-1}^a$  for some  $a$  but  $f \notin S_b^a$ . Then any mosaic extending any mosaic in  $\chi^R(h_{[0, b]}), f]|$  must recombine. Since  $f \notin S_b^a$ , there are  $ht := J_b^b$  possible haplotypes to which this recombination at  $b - 1 \rightarrow b$  may occur. Let  $\ell$  be

the length (in base pairs) of the interval  $b-1$  to  $b$ , then  $\forall R' < \ell(b)$  we have previously calculated in (2) that

$$|\chi^{R'}(h_{[b-1,b]}), g| = \binom{\ell-1}{R'} (ht-1)^{R'-1}$$

and therefore, where we write  $\chi^R(h_{[0,b-1]}), f] \circ \chi^{R'}(h_{[b-1,b]}), g]$  for the set of mosaics formed by continuing mosaics in  $\chi^R(h_{[0,b-1]}), f]$  such that they recombine between  $h_{[0,b-1]}$  and  $h_{[b-1,b]}$  and end with a subinterval of  $g$ ,

$$|\chi^R(h_{[0,b-1]}), f] \circ \chi^{R'}(h_{[b-1,b]}), g| = |\chi^R(h_{[0,b-1]}), f| \binom{\ell-1}{R'} (ht-1)^{R'-1} \quad (5)$$

*Case 2:* Suppose now that we know  $|\chi^R(h_{[0,b-1]}), f|$ , and  $f \in S_b^a$  for some  $a$ , that is,  $f$  does have a subinterval consistent with  $h_{[b-1,b]}$

There are two subcases: either there is, or there is not a recombination between the last base in  $h_{[0,b-1]}$  and the subsequent base at the beginning of  $h_{[b-1,b]}$ . Suppose that there is not. In this case, where we write  $\chi^R(h_{[0,b-1]}), f] \ominus \chi^{R'}(h_{[b-1,b]}), g]$  for the set of mosaics formed by continuing mosaics in  $\chi^R(h_{[0,b-1]}), f]$  such that they do not recombine between  $h_{[0,b-1]}$  and  $h_{[b-1,b]}$  and such that they do end with a subinterval of  $g$ ,

$$|\chi^R(h_{[0,b-1]}), f] \ominus \chi^{R'}(h_{[b-1,b]}), g| = |\chi^R(h_{[0,b-1]}), f| |\chi^{R'}(h_{[b-1,b]}), g| \quad (6)$$

such that if  $f \neq g$

$$|\chi^R(h_{[0,b-1]}), f] \ominus \chi^{R'}(h_{[b-1,b]}), g| = |\chi^R(h_{[0,b-1]}), f| |\chi^{R'}(h_{[b-1,b]}), -g| \quad (7)$$

else

$$|\chi^R(h_{[0,b-1]}), f] \ominus \chi^{R'}(h_{[b-1,b]}), g| = |\chi^R(h_{[0,b-1]}), f| |\chi^{R'}(h_{[b-1,b]}), g| \quad (8)$$

The other subcase is that there is a recombination between the last base in  $h_{[0,b-1]}$  and the subsequent base at the beginning of  $h_{[b-1,b]}$ . In this case if  $f \neq g$ ,

$$\begin{aligned} |\chi^R(h_{[0,b-1]}), f] \circ \chi^{R'}(h_{[b-1,b]}), g| &= |\chi^R(h_{[0,b-1]}), f] \circ \chi^{R'}(h_{[b-1,b]}), g| + \\ &\quad \sum_{f' \neq f, g} |\chi^R(h_{[0,b-1]}), f] \circ \chi^{R'}(h_{[b-1,b]}), f', g| \quad (9) \\ &= |\chi^R(h_{[0,b-1]}), f| |\chi^{R'}(h_{[b-1,b]}), g| + \\ &\quad \underbrace{(ht-2) |\chi^R(h_{[0,b-1]}), f| |\chi^{R'}(h_{[b-1,b]}), -g|}_{\text{by Lemma 4}} \quad (10) \end{aligned}$$

else

$$|\chi^R(h_{[0,b-1]}), f] \circ \chi^{R'}(h_{[b-1,b]}), g| = (ht-1) |\chi^R(h_{[0,b-1]}), f| |\chi^{R'}(h_{[b-1,b]}), -g| \quad (11)$$

## 2.5 Deriving the Formula for $P(h|G, H)$

Suppose that we have calculated  $|\chi(h_{[0,b-1]}), f|$  for all  $f$  and now wish to calculate  $|\chi(h_{[0,b]}), g|$  for some  $g \in S_b^a$ , for some  $a \leq b$ .

Note that as defined in the main text,  $R_{b-1}(a) = |\chi(h_{[0,b-1]}), f|$  for the  $a$  such that  $f \in S_b^a$ . This means this calculation will in fact give us the formula with which to calculate  $\overrightarrow{R}_b$  given  $\overrightarrow{R}_{b-1}$ . Let us write  $R_b(f)$  for  $R_b(a)$  such that  $f \in S_b^a$ .

Accounting for all prefixes in  $\chi(h_{[0,b-1]})$  which can produce mosaics in  $\chi(h_{[0,b]}), g$ , then

$$\begin{aligned}
|\chi(h_{[0,b]}), g| &= \sum_{\substack{R_1 < |h_{[0,b-1]}| \\ R_2 < |h_{[b-1,b]}|}} \rho^{(R_1+R_2)} |\chi^{R_1}(h_{[0,b-1]}) \ominus \chi^{R_2}(h_{[b-1,b]}), g| + \\
&\quad \sum_{\substack{R_1 < |h_{[0,b-1]}| \\ R_2 < |h_{[b-1,b]}|}} \rho^{(R_1+R_2+1)} |\chi^{R_1}(h_{[0,b-1]}) \otimes \chi^{R_2}(h_{[b-1,b]}), g| \\
&= \sum_{\substack{R_1 < |h_{[0,b-1]}| \\ R_2 < |h_{[b-1,b]}|}} \left( \rho^{(R_1+R_2)} |\chi^{R_1}(h_{[0,b-1]}) \ominus \chi^{R_2}(h_{[b-1,b]}), g| + \right. \\
&\quad \left. \rho^{(R_1+R_2+1)} |\chi^{R_1}(h_{[0,b-1]}) \otimes \chi^{R_2}(h_{[b-1,b]}), g| \right) \\
&= \sum_{\substack{R_1 < |h_{[0,b-1]}| \\ R_2 < |h_{[b-1,b]}|}} \rho^{(R_1+R_2)} \left( \sum_{\substack{a < b \\ f \in S_b^a \\ f \neq g}} |\chi^{R_1}(h_{[0,b-1]}), f| \ominus \chi^{R_2}(h_{[b-1,b]}), f| \right. \\
&\quad + |\chi^{R_1}(h_{[0,b-1]}), g| \ominus \chi^{R_2}(h_{[b-1,b]}), g| \\
&\quad + \rho \left( |\chi^{R_1}(h_{[0,b-1]}), g| \otimes \chi^{R_2}(h_{[b-1,b]}), g| \right. \\
&\quad + \sum_{\substack{a < b \\ f \in S_b^a \\ f \neq g}} |\chi^{R_1}(h_{[0,b-1]}), f| \otimes \chi^{R_2}(h_{[b-1,b]}), g| \\
&\quad \left. \left. + \sum_{\substack{a < b \\ f \notin S_b^a}} |\chi^{R_1}(h_{[0,b-1]}), f| \otimes \chi^{R_2}(h_{[b-1,b]}), g| \right) \right) \\
&= \sum_{\substack{R_1 < |h_{[0,b-1]}| \\ R_2 < |h_{[b-1,b]}|}} \rho^{(R_1+R_2)} \left( \sum_{\substack{a < b \\ f \in S_b^a \\ f \neq g}} \underbrace{|\chi^{R_1}(h_{[0,b-1]}), f| |\chi^{R_2}(h_{[b-1,b]}), \neg g|}_{\text{by (7)}} \right. \\
&\quad + \underbrace{|\chi^{R_1}(h_{[0,b-1]}), g| |\chi^{R_2}(h_{[b-1,b]}), g|}_{\text{by (8)}} \\
&\quad + \rho \left( \underbrace{|\chi^{R_1}(h_{[0,b-1]}), g| (ht - 1) |\chi^{R_2}(h_{[b-1,b]}), \neg g|}_{\text{by (11)}} \right. \\
&\quad + \sum_{\substack{a < b \\ f \in S_b^a \\ f \neq g}} \underbrace{|\chi^{R_1}(h_{[0,b-1]}), f| (|\chi^{R_2}(h_{[b-1,b]}), g| + (ht - 2) |\chi^{R_2}(h_{[b-1,b]}), \neg g|)}_{\text{by (10)}} \\
&\quad \left. \left. + \sum_{\substack{a < b \\ f \notin S_b^a}} |\chi^{R_1}(h_{[0,b-1]}), f| |\chi^{R_2}(h_{[b-1,b]}), g| \right) \right)
\end{aligned}$$

$$\begin{aligned}
&= \sum_{\substack{R_1 < |h_{[0,b-1]}| \\ R_2 < |h_{[b-1,b]}|}} \rho^{(R_1+R_2)} \left( \sum_{a < b} \sum_{f \in S_b^g} |\chi^{R_1}(h_{[0,b-1]}), f| |\chi^{R_2}(h_{[b-1,b]}), [-g, g]| - |\chi^{R_1}(h_{[0,b-1]}), g| |\chi^{R_2}(h_{[b-1,b]}), [-g, g]| \right. \\
&\quad + |\chi^{R_1}(h_{[0,b-1]}), g| |\chi^{R_2}(h_{[b-1,b]}), g, g| \\
&\quad + \rho \left( |\chi^{R_1}(h_{[0,b-1]}), g| (|\chi^{R_2}(h_{[b-1,b]}), g| - |\chi^{R_2}(h_{[b-1,b]}), g, g|) \right. \\
&\quad \left. + \sum_{\substack{a < b \\ f \in S_b^g \\ f \neq g}} |\chi^{R_1}(h_{[0,b-1]}), f| (|\chi^{R_2}(h_{[b-1,b]}), g| - |\chi^{R_2}(h_{[b-1,b]}), [-g, g]|) \right. \\
&\quad \left. + \sum_{a < b} \sum_{f \notin S_b^g} |\chi^{R_1}(h_{[0,b-1]}), f| |\chi^{R_2}(h_{[b-1,b]}), g| \right) \Big) \\
&= \sum_{\substack{R_1 < |h_{[0,b-1]}| \\ R_2 < |h_{[b-1,b]}|}} \rho^{(R_1+R_2)} \left( \sum_{a < b} \sum_{f \in S_b^g} |\chi^{R_1}(h_{[0,b-1]}), f| |\chi^{R_2}(h_{[b-1,b]}), [-g, g]| - |\chi^{R_1}(h_{[0,b-1]}), g| |\chi^{R_2}(h_{[b-1,b]}), [-g, g]| \right. \\
&\quad + |\chi^{R_1}(h_{[0,b-1]}), g| |\chi^{R_2}(h_{[b-1,b]}), g, g| \\
&\quad + \rho \left( \sum_{a < b} \sum_{f \in S_b^g} |\chi^{R_1}(h_{[0,b-1]}), f| (|\chi^{R_2}(h_{[b-1,b]}), g| - |\chi^{R_2}(h_{[b-1,b]}), [-g, g]|) \right. \\
&\quad \left. + |\chi^{R_1}(h_{[0,b-1]}), g| |\chi^{R_2}(h_{[b-1,b]}), [-g, g]| - |\chi^{R_1}(h_{[0,b-1]}), g| |\chi^{R_2}(h_{[b-1,b]}), g, g| \right. \\
&\quad \left. + \sum_{a < b} \sum_{f \notin S_b^g} |\chi^{R_1}(h_{[0,b-1]}), f| |\chi^{R_2}(h_{[b-1,b]}), g| \right) \Big) \\
&= \sum_{\substack{R_1 < |h_{[0,b-1]}| \\ R_2 < |h_{[b-1,b]}|}} \rho^{(R_1+R_2)} \left( (1 - \rho) \left( \sum_{a < b} \sum_{f \in S_b^g} |\chi^{R_1}(h_{[0,b-1]}), f| |\chi^{R_2}(h_{[b-1,b]}), [-g, g]| \right. \right. \\
&\quad \left. \left. - |\chi^{R_1}(h_{[0,b-1]}), g| |\chi^{R_2}(h_{[b-1,b]}), [-g, g]| + |\chi^{R_1}(h_{[0,b-1]}), g| |\chi^{R_2}(h_{[b-1,b]}), g, g| \right) \right. \\
&\quad \left. + \rho \sum_{a < b} \sum_{f \in S_{b-1}^g} |\chi^{R_1}(h_{[0,b-1]}), f| |\chi^{R_2}(h_{[b-1,b]}), g| \right)
\end{aligned}$$

Letting

$$\begin{aligned}
RRSame &= \sum_{R_2 < |h_{[b-1,b]}|} \rho^{R_2} |\chi^{R_2}(h_{[b-1,b]}), g, g|, \\
RRDiff &= \sum_{R_2 < |h_{[b-1,b]}|} \rho^{R_2} |\chi^{R_2}(h_{[b-1,b]}), [-g, g]|
\end{aligned}$$

(And we note that  $RRSame$  and  $RRDiff$  do not actually depend on choice of  $g$ )

$$\begin{aligned}
= & \sum_{R_1 < |h_{[0, b-1]}|} \rho^{R_1} \left( (1 - \rho) \left( \sum_{a < b} \sum_{f \in S_b^a} |\chi^{R_1}(h_{[0, b-1]}), [f]| RRDiff \right. \right. \\
& \left. \left. - |\chi^{R_1}(h_{[0, b-1]}), [g]| RRDiff + |\chi^{R_1}(h_{[0, b-1]}), [g]| RRSame \right) \right. \\
& \left. + \sum_{R_2 < |h_{[b-1, b]}|} \rho^{(R_2+1)} \sum_{a < b} \sum_{f \in S_{b-1}^a} |\chi^{R_1}(h_{[0, b-1]}), [f]| \underbrace{\binom{|h_{[b-1, b]}| - 1}{R_2}}_{\text{by (1)}} (ht - 1)^{R_2} \right)
\end{aligned}$$

Noting that

$$\sum_{R_1 < |h_{[0, b-1]}|} \rho^{R_1} |\chi^{R_1}(h_{[0, b-1]}), [f]| = R_{b-1}(f)$$

Letting:

$$\begin{aligned}
S_1 & := \sum_{a < b} \sum_{f \in S_b^a} R_{b-1}(f) \\
S_2 & := \sum_{a < b} \sum_{f \notin S_b^a} R_{b-1}(f)
\end{aligned}$$

then the above is equal to

$$\begin{aligned}
(1 - \rho) \left( S_1 RRDiff - R_b(g) (RRDiff - RRSame) \right) \\
+ (S_1 + S_2) \sum_{R_2 < |h_{[b-1, b]}|} \rho^{(R_2+1)} \binom{\ell(b) - 1}{R_2} (ht - 1)^{R_2}
\end{aligned}$$

For  $g \in S_b^b$ , the calculation is similar:

$$\begin{aligned}
|\chi(h_{[0, b]}), [g]| &= \sum_{\substack{R_1 < |h_{[0, b-1]}| \\ R_2 < |h_{[b-1, b]}|}} \rho^{(R_1+R_2+1)} |\chi^{R_1}(h_{[0, b-1]}) \otimes \chi^{R_2}(h_{[b-1, b]}), [g]| \\
&= \sum_{\substack{R_1 < |h_{[0, b-1]}| \\ R_2 < |h_{[b-1, b]}|}} \rho^{(R_1+R_2+1)} \left( \sum_{f \in S_b^a} |\chi^{R_1}(h_{[0, b-1]}), [f]| |\chi^{R_2}(h_{[b-1, b]}), [g]| \right) \\
&= (S_1 + S_2) \sum_{R_2 < |h_{[b-1, b]}|} \rho^{(R_2+1)} \left( \binom{|h_{[b-1, b]}| - 1}{R_2} (ht - 1)^{R_2} \right)
\end{aligned}$$

We can simplify the sums above by writing

$$\begin{aligned}
RRS(ht, \ell) &:= \sum_{R_2 < \ell} \rho^{R_2} \left( \binom{\ell - 1}{R_2} (ht - 1)^{R_2} \right) \\
&= \left( 1 + (ht - 1)\rho \right)^{\ell - 1} \tag{12}
\end{aligned}$$

Given a second definition



$$RRT(\ell) := (1 - \rho)^{\ell-1} \tag{13}$$

we can actually write

$$\begin{aligned} RRSame - RRDiff &= RRT(|h_{[b-1,b]}|) \\ RRDiff &= \frac{RRS(ht, |h_{[b-1,b]}|) - RRT(|h_{[b-1,b]}|)}{ht} \end{aligned}$$

and so finally, we can write our formula for  $R_b(g)$  in a compact form as

$$R_b(g) = \begin{cases} (1 - \rho) \left( S_1 \frac{RRS(ht, |h_{[b-1,b]}|) - RRT(|h_{[b-1,b]}|)}{ht} + R_{b-1}(g) RRT(|h_{[b-1,b]}|) \right) & \text{if } g \notin S_b^b \\ \rho(S_1 + S_2) RRS(ht, |h_{[b-1,b]}|) & \text{if } g \in S_b^b \end{cases}$$

which gives us equation 6 of the main text.